Sonderforschungsbereich 876

Verfügbarkeit von Information
durch Analyse unter Ressourcenbeschränkung

technische universität
dortmund

Finanzierungsantrag

2015 - 2018

Deutsche Forschungsgemeinschaft
- Sonderforschungsbereiche -
53170 Bonn

Antrag auf Finanzierung der zweiten Förderperiode des Sonderforschungsbereichs 876

Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

Providing Information by Resource-Constrained Data Analysis

für die Jahre	2015 - 2016 - 2017 - 2018
Sprecherhochschule:	Technische Universität Dortmund

Sprecherin des SFB:

Prof. Dr. Katharina Morik

Otto-Hahn-Straße 12
D-44227 Dortmund

Telefon: 0231/755-5100
Telefax: 0231/755-5105
E-Mail:
katharina.morik@tu-dortmund.de

Geschäftsführung des SFB:

Dr.-Ing. Stefan Michaelis

Otto-Hahn-Straße 12
D-44227 Dortmund

Telefon: 0231/755-4813
Telefax: 0231/755-5105
E-Mail:
stefan.michaelis@tu-dortmund.de

Dortmund, den 14.07.2014

Prof. Dr. Katharina Morik
(Sprecherin des SFB)

Dortmund, den 10.07.2014

Prof. Dr. Ursula Gather
(Rektorin)

Inhaltsverzeichnis

1. Allgemeine Angaben **7**
1.1. Kenndaten . . . 7
1.1.1. Vorstand und Gremien . . . 7
1.1.2. Beteiligte Teilprojektleiterinnen und Teilprojektleiter . . . 7
1.1.3. Beteiligte Einrichtungen . . . 8
1.1.4. Übersicht über die Projektbereiche und Teilprojekte . . . 9
1.2. Wissenschaftliches Profil des Sonderforschungsbereichs . . . 12
1.2.1. Zusammenfassung des Forschungsprogramms . . . 12
1.2.2. Ausführliche Darstellung des Forschungsprogramms . . . 13
1.2.3. Positionierung des Sonderforschungsbereichs in seinem weiteren Fachgebiet . . . 22
1.2.4. Inhaltliche Einbindung des Sonderforschungsbereichs am Standort . . . 24
1.2.5. Nationale und internationale Kooperationen und Vernetzung . . . 25
1.3. Strukturelle Einbindung und Auswirkungen des Sonderforschungsbereichs am Standort . . . 26
1.3.1. Personelle Situation . . . 27
1.3.2. Forschungsinfrastruktur . . . 28
1.3.3. Forschungsdaten . . . 29
1.3.4. Gleichstellung von Wissenschaftlerinnen und Wissenschaftlern und Vereinbarkeit von wissenschaftlicher Karriere und Familie . . . 32
1.3.5. Wissenschaftlicher Nachwuchs . . . 33
1.3.6. Erkenntnistransfer und Öffentlichkeitsarbeit . . . 38
1.4. Andere Förderungen der beteiligten Teilprojektleiterinnen und Teilprojektleiter . . . 39

2. Vorhandene und beantragte Mittel **47**
2.1. Vorhandene Mittel . . . 47
2.1.1. Übersicht über vorhandene Sachmittel . . . 47
2.1.2. Übersicht über das vorhandene Personal . . . 47
2.1.3. Liste der Geräte, die dem Sonderforschungsbereich zur Verfügung stehen . . . 48
2.2. Bisherige und beantragte Mittel . . . 49
2.2.1. Übersicht . . . 49
2.2.2. Übersicht über Personal, für das Mittel beantragt werden . . . 50
2.2.3. Liste der Investitionen, für die Mittel beantragt werden . . . 50
2.2.4. Angaben zur Versuchstierhaltung . . . 50

3. Darstellung des Programms nach Projektbereichen und Teilprojekten **51**
A1 (Morik/Spinczyk): Data Mining für ubiquitäre Systemsoftware . . . 53
A2 (Sohler/Teubner): Algorithmik von Lernverfahren in eingebetteten Systemen . . . 75
A3 (Marwedel/Rahnenführer): Methoden der Effizienten Ressourcennutzung in Algorithmen des Maschinellen Lernens . . . 95
A4 (Spinczyk/ten Hompel/Wietfeld): Ressourceneffiziente und verteilte Plattformen zur integrativen Datenanalyse . . . 117
A5 (Biskup/Kern-Isberner): Austausch und Fusion von Information unter Verfügbarkeits- und Vertraulichkeitsanforderungen in Multiagentensystemen . . . 143
A6 (Kersting/Mutzel/Sohler): Ressourceneffiziente Analyse von Graphen . . . 155
TB1 (Baumbach/Rahnenführer): Ressourcenbeschränkte Analyse von Spektrometriedaten . . . 179
B2 (Chen/Hergenröder/Müller): Ressourcenoptimierte Echtzeitanalyse stark Artefakt-behafteter Bildsequenzen zur Detektion von Nanoobjekten . . . 201
B3 (Deuse/Morik): Data Mining in Sensordaten automatisierter Prozesse . . . 223
B4 (Kersting/Schreckenberg/Wietfeld): Analyse und Kommunikation für die dynamische Verkehrsprognose . . . 245
C1 (Schramm/Lee/Rahmann): Merkmalsselektion in hochdimensionalen Daten am Beispiel der Risikoprognose in der Onkologie . . . 269

C3 (Morik/Rhode/Ruhe): Mehrstufige statistische Analyse von hochfrequenten räumlich-zeitlichen Prozessdaten . . . 291
C4 (Ickstadt/Sohler): Regressionsverfahren für sehr große, hochdimensionale Daten . . . 317
C5 (Spaan/Teubner): Echtzeitanalyse und Speicherung für hochvolumige Daten aus der Teilchenphysik . . . 337
MGK (Rhode): Integriertes Graduiertenkolleg . . . 359
Z (Morik): Zentrale Aufgaben des Sonderforschungsbereichs . . . 377

4. Ordnung des Sonderforschungsbereichs 389
I. Definition und Zweck des Sonderforschungsbereichs . . . 389
II. Organisatorischer Aufbau des Sonderforschungsbereichs . . . 389
III. Mitglieder . . . 389
IV. Mitgliederversammlung . . . 390
V. Vorstand und Sprecher/in . . . 390
VI. Gemeinsame Projekte und Einrichtungen . . . 391
VII. Schlussbestimmung . . . 391

5. Erklärung zur Raumsituation des Sonderforschungsbereichs 393

6. Erklärung zu den Publikationsverzeichnissen 395

7. Ethikvotum C1 397

8. Änderung des Kooperationsvertrags TB1 399

1. Allgemeine Angaben

1.1. Kenndaten

1.1.1. Vorstand und Gremien

Sprecherin:	Prof. Dr. Katharina Morik
Stellvertretender Sprecher:	Prof. Dr. Peter Marwedel
Vertreter für Projektbereich A:	Prof. Dr. Christian Sohler
Vertreter für Projektbereich B:	Prof. Dr. Christian Wietfeld
Vertreter für Projektbereich C:	Prof. Dr. Dr. Wolfgang Rhode

Über Mitgliederversammlung und Vorstand hinausgehend gibt es keine weiteren Gremien.

1.1.2. Beteiligte Teilprojektleiterinnen und Teilprojektleiter

Teilprojektleitende	Geburtsjahr	Jahr der Promotion	Institut, Ort	Teilprojekt
Baumbach, Prof. Dr., Jörg Ingo	1959	1990	B&S Analytik GmbH, Dortmund	TB1
Chen, Prof. Dr., Jian-Jia	1978	2006	LS 12, Fakultät für Informatik, TU Dortmund	B2
Deuse, Prof. Dr., Jochen	1967	1998	Lehrstuhl für Arbeits- und Produktionssysteme, Fakultät Maschinenbau, TU Dortmund	B3
Hergenröder, Dr., Roland	1961	1992	Leibniz-Institut für Analytische Wissenschaften (ISAS) e.V., Dortmund	B2
Ickstadt, Prof. Dr., Katja	1965	1994	Mathematische Statistik und Biometrische Anwendungen, Fakultät Statistik, TU Dortmund	C4
Kersting, Prof. Dr., Kristian	1973	2006	LS 8, Fakultät für Informatik, TU Dortmund	A6, B4
Lee, Dr., Sangkyun	1977	2011	LS 8, Fakultät für Informatik, TU Dortmund	C1
Marwedel, Prof. Dr., Peter	1949	1974	LS 12, Fakultät für Informatik, TU Dortmund	A3
Morik, Prof. Dr., Katharina	1954	1981	LS 8, Fakultät für Informatik, TU Dortmund	A1, B3, C3, Z
Müller, Prof. Dr., Heinrich	1953	1981	LS 7, Fakultät für Informatik, TU Dortmund	B2
Mutzel, Prof. Dr., Petra	1964	1994	LS 11, Fakultät für Informatik, TU Dortmund	A6
Rahmann, Prof. Dr., Sven	1974	2004	Lehrstuhl für Genominformatik, Fakultät für Medizin, Universität Duisburg-Essen	C1

Teilprojektleitende	Geburtsjahr	Jahr der Promotion	Institut, Ort	Teilprojekt
Rahnenführer, Prof. Dr., Jörg	1971	1999	Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, TU Dortmund	A3, TB1
Rhode, Prof. Dr. Dr., Wolfgang	1961	1990/1993	Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund	C3, MGK
Ruhe, Dr., Tim	1981	2013	Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund	C3
Schreckenberg, Prof. Dr., Michael	1956	1985	Physik von Transport und Verkehr, Fakultät Physik, Universität Duisburg-Essen	B4
Schramm, PD Dr., Alexander	1968	1999	Onkologisches Labor, Kinderklinik, Universitätsklinikum Essen	C1
Sohler, Prof. Dr., Christian	1973	2002	LS 2, Fakultät für Informatik, TU Dortmund	A2, A6, C4
Spaan, Prof. Dr., Bernhard	1960	1988	Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund	C5
Spinczyk, Prof. Dr., Olaf	1970	2002	LS 12, Fakultät für Informatik, TU Dortmund	A1, A4
ten Hompel, Prof. Dr., Michael	1958	1991	Lehrstuhl für Förder- und Lagerwesen, Fakultät Maschinenbau, TU Dortmund	A4
Teubner, Prof. Dr., Jens	1976	2006	LS 6, Fakultät für Informatik, TU Dortmund	A2, C5
Wietfeld, Prof. Dr., Christian	1966	1997	Lehrstuhl für Kommunikationsnetze, Fakultät für Elektro- und Informationstechnik, TU Dortmund	A4, B4

1.1.3. Beteiligte Einrichtungen

Am SFB 876 sind die folgenden Fakultäten der Technischen Universität Dortmund und der Universität Duisburg-Essen beteiligt:

TU Dortmund, Fakultät für Informatik

- Lehrstuhl 2 – Effiziente Algorithmen und Komplexitätstheorie
- Lehrstuhl 6 – Datenbanken und Informationssysteme
- Lehrstuhl 7 – Grafische Systeme
- Lehrstuhl 8 – Künstliche Intelligenz
- Lehrstuhl 11 – Algorithm Engineering
- Lehrstuhl 12 – Eingebettete Systeme

TU Dortmund, Fakultät für Elektro- und Informationstechnik

- Lehrstuhl für Kommunikationsnetze

TU Dortmund, Fakultät Maschinenbau

- Lehrstuhl für Arbeits- und Produktionssysteme
- Lehrstuhl für Förder- und Lagerwesen

TU Dortmund, Fakultät Physik

- Lehrstuhl Experimentelle Physik E5

TU Dortmund, Fakultät Statistik

- Statistische Methoden in der Genetik und Chemometrie
- Mathematische Statistik und Biometrische Anwendungen

Universität Duisburg-Essen, Fakultät für Medizin

- Lehrstuhl für Genominformatik

Universität Duisburg-Essen, Fakultät Physik

- Lehrstuhl Physik von Transport und Verkehr

Universität Duisburg-Essen, Universitätsklinikum

- Kinderheilkunde, Onkologie

Hochschule Reutlingen

- Fakultät Angewandte Chemie

Außeruniversitäre Einrichtungen

- B&S Analytik GmbH, eine Spin-off-Firma des Leibniz-Instituts für Analytische Wissenschaften
- Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., ein Institut an der TU Dortmund, Abteilung für Materialanalyse

1.1.4. Übersicht über die Projektbereiche und Teilprojekte

Projektbereich A: Übergreifende Methoden				
TP	Status	Titel	Fachgebiet und Arbeitsrichtung	Teilprojektleitende, Institut, Ort
A1		Data Mining für ubiquitäre Systemsoftware	Data Mining/ Maschinelles Lernen	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund
			Eingebettete Systeme	O. Spinczyk, LS 12, Fakultät für Informatik, TU Dortmund
A2		Algorithmik von Lernverfahren in eingebetteten Systemen	Algorithmische Komplexitätstheorie	C. Sohler, LS 2, Fakultät für Informatik, TU Dortmund
			Informationssysteme	J. Teubner, LS 6, Fakultät für Informatik, TU Dortmund
A3		Methoden der Effizienten Ressourcennutzung in Algorithmen des Maschinellen Lernens	Eingebettete Systeme	P. Marwedel, LS 12, Fakultät für Informatik, TU Dortmund
			Maschinelles Lernen	J. Rahnenführer, Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, TU Dortmund

A4		Ressourceneffiziente und verteilte Plattformen zur integrativen Datenanalyse	Eingebettete Systeme	O. Spinczyk, LS 12, Fakultät für Informatik, TU Dortmund
			Kommunikationsnetze	C. Wietfeld, Lehrstuhl für Kommunikationsnetze, Fakultät für Elektro- und Informationstechnik, TU Dortmund
			Logistik	M. ten Hompel, Lehrstuhl für Förder- und Lagerwesen, Fakultät Maschinenbau, TU Dortmund
A5	E	Austausch und Fusion von Information unter Verfügbarkeits- und Vertraulichkeitsanforderungen in Multiagentensystemen	Informationssysteme und Sicherheit	J. Biskup, LS 6, Fakultät für Informatik, TU Dortmund
			Information Engineering	G. Kern-Isberner, LS 1, Fakultät für Informatik, TU Dortmund
A6	N	Ressourceneffiziente Analyse von Graphen	Data Mining	K. Kersting, LS 8, Fakultät für Informatik, TU Dortmund
			Algorithm Engineering	P. Mutzel, LS 11, Fakultät für Informatik, TU Dortmund
			Algorithmische Komplexitätstheorie	C. Sohler, LS 2, Fakultät für Informatik, TU Dortmund

Projektbereich B: Technische Ressourcenbeschränkungen				
TP	Status	Titel	Fachgebiet und Arbeitsrichtung	Teilprojektleitende, Institut, Ort
TB1		Ressourcenbeschränkte Analyse von Spektrometriedaten	Statistische Methoden	J. Rahnenführer, Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, TU Dortmund
			Metabolomik	J. Baumbach, B&S Analytik GmbH, Dortmund
B2		Ressourcenoptimierte Echtzeitanalyse stark Artefakt-behafteter Bildsequenzen zur Detektion von Nanoobjekten	Visual Computing	H. Müller, LS 7, Fakultät für Informatik, TU Dortmund
			Biomedizinische Analytik	R. Hergenröder, ISAS e.V., Dortmund
			Eingebettete Systeme	J.-J. Chen, LS 12, Fakultät für Informatik, TU Dortmund
B3		Data Mining in Sensordaten automatisierter Prozesse	Produktionssysteme	J. Deuse, Lehrstuhl für Arbeits- und Produktionssysteme, Fakultät Maschinenbau, TU Dortmund
			Data Mining/ Maschinelles Lernen	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund
B4		Analyse und Kommunikation für die dynamische Verkehrsprognose	Verkehrsmodellierung	M. Schreckenberg, Physik von Transport und Verkehr, Fakultät Physik, Universität Duisburg-Essen
			Kommunikationsnetze	C. Wietfeld, Lehrstuhl für Kommunikationsnetze, Fakultät für Elektro- und Informationstechnik, TU Dortmund
			Data Mining	K. Kersting, LS 8, Fakultät für Informatik, TU Dortmund

Projektbereich C: Informationsgewinnung unter Ressourcenbeschränkung				
TP	Status	Titel	Fachgebiet und Arbeitsrichtung	Teilprojektleitende, Institut, Ort
C1		Merkmalsselektion in hochdimensionalen Daten am Beispiel der Risikoprognose in der Onkologie	Maschinelles Lernen	S. Lee, LS 8, Fakultät für Informatik, TU Dortmund
			Molekularmedizin	A. Schramm, Onkologisches Labor, Kinderklinik, Universitätsklinikum Essen
			Bioinformatik	S. Rahmann, Lehrstuhl für Genominformatik, Fakultät für Medizin, Universität Duisburg-Essen
C3		Mehrstufige statistische Analyse von hochfrequenten räumlich-zeitlichen Prozessdaten	Astroteilchenphysik	W. Rhode, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund
			Astroteilchenphysik	T. Ruhe, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund
			Data Mining/ Maschinelles Lernen	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund
C4		Regressionsverfahren für sehr große, hochdimensionale Daten	Algorithmische Komplexitätstheorie	C. Sohler, LS 2, Fakultät für Informatik, TU Dortmund
			Statistische Methoden	K. Ickstadt, Lehrstuhl Mathematische Statistik und Biometrische Anwendungen, Fakultät Statistik, TU Dortmund
C5	N	Echtzeitanalyse und Speicherung für hochvolumige Daten aus der Teilchenphysik	Teilchenphysik	B. Spaan, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund
			Informationssysteme	J. Teubner, LS 6, Fakultät für Informatik, TU Dortmund

Integriertes Graduiertenkolleg			
TP	Status	Titel	Teilprojektleiter, Institut, Ort
MGK		Integriertes Graduiertenkolleg	W. Rhode, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund

Zentrales Verwaltungsprojekt			
TP	Status	Titel	Teilprojektleitende, Institut, Ort
Z		Zentrale Aufgaben des Sonderforschungsbereichs	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund

1.2. Wissenschaftliches Profil des Sonderforschungsbereichs

1.2.1. Zusammenfassung des Forschungsprogramms

Der Sonderforschungsbereich 876 bringt die Forschungsgebiete der Datenanalyse (Data Mining, maschinelles Lernen, Statistik) und der eingebetteten Systeme (Cyber-Physical Systems) zusammen und erweitert sie derart, dass Information aus verteilten, dynamischen Datenmassen für Entscheidungsprozesse realzeitlich vor Ort verfügbar wird. Die Erfassung und Speicherung von Hochdurchsatzexperimenten der Biomedizin, von astrophysikalischen Teleskopen oder teilchenphysikalischen Experimenten übertrifft die Kapazität heutiger Rechner, sodass die Analyse bereits bei der Datenerfassung in die Auswahl zu speichernder Daten einfließen muss und neue Methoden der Speicherung nötig sind. Dicht besetzte Datenräume fordern die Skalierbarkeit von Analysealgorithmen heraus. Die Analyse von Datenströmen, verteilten Daten und unzuverlässig übermittelten Daten erschwert es weiter, aus der Datenflut nützliche Information zu gewinnen. Schließlich müssen auch komplexe Abhängigkeiten zwischen Variablen, beispielsweise durch Graphen ausgedrückt, berücksichtigt werden.

Die Ressourcenbeschränkungen von Rechnern sind stets in Relation zu den Daten zu sehen. Fallen sie bei kleinen, eingebetteten Systemen sofort auf, gelten sie bei besonders großen, hochdimensionalen, verteilten, dynamischen oder komplexen Daten sogar für große Rechenanlagen und -zentren:

- Beschränkung der Kommunikation,
- Beschränkung der Energie,
- Beschränkung von Rechenzeit,
- Beschränkung des Speichers.

Die Untersuchung von Algorithmen in Hinblick auf diese Ressourcenbeschränkungen wurde im SFB 876 erfolgreich begonnen. Damit gehört der SFB 876 zu den frühen Vertretern des Trends, der später unter dem Namen *Big Data* überaus populär geworden ist. Das Volumen der Daten, die Heterogenität und Verteiltheit von Daten sowie die Geschwindigkeit, mit der Signale eintreffen, sind Datencharakteristika, mit denen alle Schritte der Informationsgewinnung zu tun haben:

1. Datenerfassung und Speichern
2. Merkmalsextraktion, Filtern, Datenbereinigung
3. Integration von Informationen, Datenzusammenfassung, Merkmalsselektion
4. Datenanalyse zur Gruppierung, Modellierung, Vorhersage
5. Anwendung der Analyse-Ergebnisse und Evaluierung des Nutzens der Analyse

Während die Beachtung der Laufzeit im Datenanalyseschritt Standard ist, rückt der Forschungsbedarf bei darüber hinaus gehenden Ressourcen und über alle Schritte hinweg erst in jüngster Zeit in den Vordergrund. In beiderlei Hinsicht ist der SFB 876 Vorreiter. Schon in der ersten Projektphase gab es erste Ansätze zu allen Schritten, vorrangig wurde aber an Algorithmen zu den zentralen Schritten 2 – 4 gearbeitet. Bezogen auf die Ressourcenbeschränkungen werden verteilte, parallele und Datenstromalgorithmen für die Merkmalsextraktion, die Datenzusammenfassung und die Modellierung entwickelt. In der zweiten Phase sollen Lösungsansätze für die ressourceneffiziente Erfassung und Speicherung von Daten (1) sowie der Modellanwendung und -evaluierung (5) hinzu kommen.

Die Ressourcenbeschränkungen werden für bestimmte Programmiermodelle und Ausführungsplattformen betrachtet. Die quantitative Charakterisierung des Ressourcenbedarfs ist eine wichtige Voraussetzung für den Vergleich verschiedener Methoden. Deshalb wurden in der ersten Phase neuartige Messtechniken als Grundlage für die Modellierung der Ressourcen entwickelt. Auch Kostenmodelle für Programme tragen zur Quantifizierung von Ressourcen bei und ermöglichen so den Vergleich unterschiedlicher Programmiermodelle bezüglich ihres Ressourcenverbrauchs. Erste Ressourcenmodelle für Sensorknoten, Mobilfunkendgeräte und eingebettete Betriebssysteme wurden bereits in der ersten Projektphase entwickelt. In der zweiten Phase sollen Ressourcenmodelle auch automatisch erstellt in die Gewinnung von Ressourcen, z.B. das *energy harvesting*, einbezogen werden. Das Fernziel bleibt dabei, dass intelligente Datenanalyse in die Systeme eingebettet wird. Bisher verlief die Entwicklung von eingebetteten Systemen und Algorithmen der Datenanalyse weitgehend getrennt. Für viele Herausforderungen, die Lebenswissenschaften, Logistik, Maschinenbau oder Astrophysik an die Informatik herantragen, ist aber deren Integration gefordert. Langfristig profitieren alle Bereiche voneinander und werden gemeinsam neuartige Systeme entwickeln, die trotz beschränkter Ressourcen wertvolle Information liefern.

1.2.2. Ausführliche Darstellung des Forschungsprogramms

Das langfristige Ziel des Sonderforschungsbereichs 876 bleibt, dass aus dem allgemeinen Vorhandensein von sehr vielen Daten nutzbare Information wird, die zuverlässig und pünktlich vor Ort zur Verfügung steht. Die Vorteile großer Datensammlungen werden zunehmend erkannt. Aus Anwendungssicht steht im Vordergrund, dass durch die Datenanalyse aus ohnehin vorhandenen Daten nutzbare Information wird. Für eine datengestützte Wissenschaft ist wichtig, dass erst bei sehr große Datenmengen auch seltene Ereignisse vorkommen. Aus methodischer Sicht gibt es für die Analyse sowohl Vorteile als auch neue Herausforderungen. Die Vorteile sehr großer, nahezu vollständiger Datensammlungen sind beispielsweise, dass sie höherdimensionale Datenräume dichter besetzen können, als es bei kleineren Sammlungen möglich war, und dass die Häufigkeit von Beobachtungen sich ihrer Wahrscheinlichkeit annähert und Ausreißer leichter zu erkennen sind. Gleichzeitig sprengen sie aber den jahrelang als selbstverständlich betrachteten Rahmen der Datenanalyse in allen Schritten der Informationsgewinnung und erweitern die bisher üblichen Gütekriterien (Fehlerraten) um solche, die physikalische Ressourcen betreffen. Zentrales Thema des Vorhabens ist die Ressourcenbeschränkung als Herausforderung für Algorithmen der Datenanalyse. Wir definieren die Ressourcenbeschränkung als Verhältnis von technischer Ausführungskapazität einerseits und Dimensionalität und Masse der Daten andererseits. Daraus ergibt sich die Struktur des SFB 876 in die drei Bereiche A, B und C.

- Zum einen sind die Ressourcenbeschränkungen durch lokale, mobile Systeme gegeben. Diese technischen Ressourcenbeschränkungen werden im **Projektbereich B** in Hinblick auf die Ausführung von Algorithmen untersucht.

- Zum anderen ergeben sich die Ressourcenbeschränkungen aus der Schwierigkeit der Analyseaufgabe durch eine hohe Dimension, Dynamik und Volumen an Daten. Diese Ressourcenbeschränkungen für die Informationsgewinnung werden im **Projektbereich C** in Hinblick auf die Datenanalysealgorithmen untersucht.

- Die beiden Säulen werden durch die übergreifenden Projekte im **Projektbereich A** zusammengeführt.

Bevor wir die drei Bereich darstellen, gehen wir im Folgenden die Schritte der Informationsgewinnung durch und verweisen dabei auf die Forschungsfragen, die sich aus der ersten Phase für die nächste und darüber hinaus ergeben haben. Die Forschungsfragen werden von Projekten aus verschiedenen Projektbereichen gestellt, sodass sich eine enge Verzahnung ergibt, die durch Zusammenarbeiten im SFB realisiert wird.

Daten erfassen und speichern: Algorithmen zur ressourcenschonenden Datenerfassung sind insbesondere bei eingebetteten Systemen notwendig. Dort reichen die Ressourcen oft für ein umfangreiches Mitschreiben (*logging*) nicht aus. Die Erfassung würde die übliche Funktion des Geräts verhindern. Insbesondere bezüglich der Ressourcen Energie und Kommunikation besteht das bekannte Problem, durch die Erfassung das zu Messende zu ändern und so die Messung zu verfälschen. Diesem Problem stellen sich insbesondere die Teilprojekte A1 und A4. Sie arbeiten an Approximationen, die den Ressourcenverbrauch für die Datenerfassung bei möglichst geringem Präzisionsverlust reduzieren.

Eigene Methoden zur Datengewinnung entwickeln TB1 mit dem Atemluftsensor und B2 mit seiner Plasmonen-Mikroskopie. Hier kann die Geräteentwicklung direkt auf die Analyse abgestimmt werden. So sollen Verbesserungen des PAMONO-Sensors durch neuartige Bindungen der Moleküle an magnetische Teilchen und an die Sensorfläche in ihrer Auswirkung auf die Erkennung der Nanoteilchen in den Bildfolgen untersucht werden. Dem SFB ist es gelungen, für das Teilprojekt C1 die neuartige Datengewinnung *Next Generation Sequencing* für die Untersuchung des Neuroblastoms einzuwerben. Diese neue Technik erfasst auch Sequenzmutationen, die für die Untersuchung von fortschreitenden Tumorstadien wichtig sind, erhöht das Datenvolumen aber gleichzeitig in den Petabyte-Bereich. Sven Rahmann bringt mit seiner Professur für *Computational Biology* der Universitätsallianz Ruhr Spezialgeräte und Personalmittel in den SFB ein für die aufwändige Gewinnung der NGS-Daten, die zur Analyse dem SFB zur Verfügung gestellt werden.

Die Ablage und Verwaltung von Datenmengen im Petabyte-Bereich ist zu einem notwendigen Untersuchungsgegenstand geworden und kann in der nächsten Projektphase insbesondere von dem neuen Kollegen Jens Teubner systematisch angegangen werden. Das Datenvolumen überfordert die klassischen Speichermethoden. Im Umfeld großer Internetdienste sind deshalb Techniken entstanden, mit denen auch sehr große Datenmengen effizient, skalierbar und fehlertolerant gespeichert werden können. Sie bauen auf verteilten Speicher- und Rechenarchitekturen auf. Dabei ist die Verteilungsstrategie und die Indexierung zu untersuchen. Dies kann im Teilprojekt C5 in Kooperation mit dem *Worldwide LHC Computing Grid* geschehen.

Heterogene Rechnersysteme für verteilte Dateisysteme und massiv parallele Programmiermodelle sollen bezüglich des schnellen Zugriffs für die Datenanalyse, bezüglich der Kommunikation und des Energieverbrauchs untersucht werden.

Merkmalsextraktion, Filtern, Datenbereinigung: Oft werden die Daten bereits gefiltert oder zusammengefasst, bevor sie gespeichert werden. Bei den Datenströmen aus der Astroteilchenphysik ist es unabdingbar, dass nur bestimmte Daten überhaupt betrachtet werden. In der ersten Phase hat sich die Lambda-Architektur, die den *Batch Layer* mit dem Programmiermodell *Map Reduce* und den *Speed Layer* mit dem Datenstrommodell integriert, für eine Reihe von Projekten als geeignet herausgestellt. A1 hat eine abstrakte Modellierung von Prozessen auf Datenströmen erreicht, die realzeitfähig und für eingebettete Systeme geeignet ist. In Zusammenarbeit mit C3 wurden eingehende Teleskopdaten realzeitlich vorverarbeitet, gefiltert und visualisiert. Insbesondere lassen sich mit dem *Streams Framework* Merkmale aus eingehenden Datenströmen extrahieren. Damit ist eine Grundlage für Arbeiten in der nächste Phase geschaffen. Die Merkmalsextraktion war in den Teilprojekten TB1 und B2 so erfolgreich, dass diese Aufgabe in der nächsten Phase dort nicht mehr bearbeitet wird. Auch die neuen Hochdurchsatzdaten von C1 werden in der nächsten Phase bezüglich der Energieeffizienz untersucht werden. Die Datenbereinigung bei verteilten Daten ist im Teilprojekt B3 ein Thema. Die Anomaliedetektion ist dort bei vertikal verteilten Daten, wie sie häufig in Sensornetzwerken vorkommen (d.h. Sensorknoten erfassen verschiedene Merkmale derselben Beobachtung), erstmals in Hinblick auf die Kommunikationseffizienz untersucht worden. Es ist zu vermuten, dass der Energiebedarf in diesem Fall mit dem Volumen, das kommuniziert wird, korreliert.

Integration von Informationen, Datenzusammenfassung, Merkmalsselektion: Die Integration unterschiedlicher Messungen wird insbesondere in A1, B3 und B4 untersucht. Durch die Verteilung der MAGIC-Teleskope kann aber auch in C3 die Integration verschiedener Beobachtungen behandelt werden. Hier sind die Merkmale gleichartig, aber die Zugehörigkeit zu einem bestimmten Ereignis muss erst festgestellt werden. Algorithmen zur *Datenzusammenfassung*, die auf Datenströmen arbeiten, wurden in den Teilprojekten A1 und A2 betrachtet. Der enge Bezug zwischen Anomalieentdeckung und dem Prinzip der minimalen Beschreibungslänge wird in der nächsten Projektphase von A1 daraufhin untersucht, in welchem Verhältnis die Güte des gelernten Modells zum Ressourcenverbrauch steht. Dabei sollen verschiedene Algorithmen und sogar unterschiedliche Ausführungsplattformen betrachtet werden. Aufbauend auf dem Energiemodell, das Teilprojekt A4 in der ersten Phase entwickelt hat, wird A2 in der nächsten Phase die Betrachtung von Kernmengen für die Zusammenfassung in Hinblick auf den Energieverbrauch und Kommunikationsbedarf untersuchen. Die *Merkmalsselektion* wurde bereits in der ersten Phase von A1, TB1, B2, B3, C1, C3 und C4 erfolgreich bearbeitet. Damit ist das Thema jedoch in keiner Weise abgeschlossen. Vielmehr hat sich bestätigt, dass die Algorithmen zur Merkmalsselektion durch große Datenvolumina und sehr hohe Dimensionalität besonders herausgefordert sind, sodass die Arbeit an neuen, besseren Methoden in diesen Projekten fortgesetzt wird. In Teilprojekt C4 werden in der zweiten Phase speziell Merkmalsselektionsverfahren für Regressionsmodelle, in denen die Anzahl der Variablen sehr viel größer ist als die Anzahl der Beobachtungen, entwickelt. Unter dem allgemeineren Begriff der Modellselektion wird das Teilprojekt A3 in der nächsten Phase den Zusammenhang von Ressourcen und ausgewählten Merkmalen auf den verschiedenen Ebenen, vom Algorithmus bis zur Ausführungsplattform, charakterisieren. Die Optimierung der Parameter von Lernverfahren, die in A3 als zur Modellselektion gehörend systematisch erkundet wird, steht in B3 im Zusammenhang verteilter Datenanalyse. Über die automatische Parameteroptimierung von RapidMiner hinaus, die in C3 erfolgreich eingesetzt wurde, sollen in B3 kommunikationseffiziente Algorithmen zur Parameteroptimierung für vertikal verteilte Daten gefunden werden.

Datenanalyse zur Gruppierung, Modellierung, Vorhersage: Dieser Schritt der Datenanalyse bleibt für die Informationsgewinnung unter Ressourcenbeschränkung zentral. Er zieht sich durch alle Projekte. Die Datenanalyse wird für verschiedene Programmiermodelle, also parallele, verteilte oder Online-Algorithmen sowie *Map Reduce*, im Hinblick auf Speicher-, Energie- und Kommunikationseffizienz hin untersucht. Ebenso wird die Realzeit als Ressource betrachtet. Während im maschinellen Lernen die Gütemaße bisher lediglich bezogen auf einen Fehler (*Precision, Recall, Accuracy*) bzw. Fehlerraten (*ROC, AUC*) formuliert werden, wird hier der Ressourcenverbrauch als weiteres Kriterium eingeführt. Es können die unterschiedlichen Bewertungsmaße gegeneinander abgewogen werden. Das Ziel ist, für einen bestimmten Vorrat einer Ressource das passende Fehlerintervall angeben zu können, und umgekehrt, für eine vorgegebene Fehlertoleranz die nötigen Ressourcen zu bestimmen. Für diese neuen Bewertungskriterien unter dieser Betrachtung ergeben sich auch neue Lernalgorithmen und Approximationen. Tatsächlich müssen die Algorithmen für bestimmte

Lernprobleme wie etwa das der *max margin separation*, der *min sum of squares* oder der *max likelihood* wieder neu bewertet und verbessert werden. Im Lichte der Ressourceneffizienz werden bekannte Lösungsansätze wieder neu analysiert. Beispielsweise wurde als verteilter Algorithmus für Entscheidungsbäume ein auf Eigenwerten beruhender Algorithmus von Hillol Kargupta eingeführt. Das Teilprojekt A1 entwickelt und analysiert paradigmatisch neue Algorithmen für probabilistische graphische Modelle für unterschiedliche Ausführungsplattformen mit ihren spezifischen Ressourcenanforderungen. Neue Algorithmen werden die Ausführung dieser besonders anspruchsvollen Lernverfahren sogar auf stark beschränkten eingebetteten Systemen ermöglichen. Im Teilprojekt A2 steht nun der bekannte Lösungsansatz k-Means erneut zur Diskussion und wird in Bezug auf unterschiedliche Ressourcen und Programmiermodelle studiert. Entsprechend werden im Teilprojekt C4 verschiedene Regressionsverfahren wie verallgemeinerte lineare Regression und Bayessche Verfahren, besonders unter dem Aspekt der Ressourcenbeschränkung, neu untersucht.

Die Beschäftigung mit raumzeitlichen Modellen, von A1 und C1 angestoßen, hat sich zu einem zentralen Thema des SFB 876 entfaltet. Probabilistische graphische Modelle sind eine besondere Herausforderung sowohl bezüglich des Speicherbedarfs als auch bezüglich der Laufzeit und des Energiebedarfs. A1 arbeitet weiterhin an Algorithmen für massiv parallele Ausführungsplattformen. Darüber hinaus hat A1 sich zum Ziel gesetzt, Approximationsalgorithmen zu schaffen, die allgemeine Markov Random Fields unter sehr starken Ressourcenbeschränkungen (z.B. ohne Fließkommaarithmetik) ausführbar machen. Die Zusammenarbeit mit B4 zu probabilistischen graphischen Modellen in der Verkehrsmodellierung wird fortgesetzt. Das neue Projekt A6 betont diesen Bereich, seine Graphen sind aber die Eingabe in die Datenanalyse, nicht das gelernte (probabilistische) Modell. Die komplexen Eingabedaten, wie sie beispielsweise in sozialen Netzwerken entstehen, stellen eine Herausforderung an Algorithmen bezüglich Speicher und Laufzeit dar. In A6 werden statistische und graphtheoretische Ansätze verbunden, um ressourcenschonende Algorithmen für das *Graph Mining* zu entwerfen.

Anwendung der Analyseergebnisse und Evaluierung des Nutzens: Es wurde schon mehrfach unter den Wissenschaftlern des maschinellen Lernens kritisiert, dass die Bewertung gelernter Modelle sich oft auf die Angabe statistischer Gütemaße beschränkt. Wenn zudem die vielfältig vorbereiteten Datensätze der *UCI Library* verwendet werden, besteht die Gefahr, dass sich das Gebiet an diese vereinfachten Probleme überanpasst und der Nutzen der Datenanalyse gar nicht mehr betrachtet wird. Der Sonderforschungsbereich 876 ermöglicht zusätzlich zu den üblichen Gütemaßen auch Bewertungen bezüglich des Ressourcenverbrauchs. So sollen Atemluftdiagnosen (TB1) und Virenerkennung (B2) durch kleinere Geräte durchführbar werden. Anhand der Analyseergebnisse sollen automatisch Entscheidungsvorschläge für den medizinischen Alltag unterbreitet werden. Dazu werden in der Ausatemluft fragestellungsbezogen relevante Stoffwechselprodukte und in Zeitreihen kritische Entwicklungen identifiziert. Die Entwicklung entsprechender Cyber-Physical Systems ist Teil dieses Sonderforschungsbereichs. Insbesondere werden in TB1 Vorbereitungen für eine Integration in Kleidung und Telemedizinanbindung getroffen. Die Verbindung von *Cyber-Physical Systems* und *Big Data Analytics* ermöglicht obendrein den direkten Einsatz der gelernten Modelle zur Steuerung von Geräten und somit die Feststellung des Nutzens der Modelle für Geräte, die eine bestimmte Aufgabe erfüllen sollen. Wie insbesondere in A1 aufgeführt, spielt die Ressourcenschonung an zwei Stellen eine Rolle: einmal bei der Bewertung von Algorithmen – wie viel Energie, Speicher, Laufzeit, Kommunikationaufwand brauchen sie für welche Qualität ihrer Modelle? Zum anderen kann ein gelerntes Modell nun auf ein physikalisches möglicherweise ubiquitäres System angewendet werden, um dieses energiegewahr, speicher- und kommunikationseffizient zu machen. Diese übergeordnete Ressourcenschonung belegt den Nutzen der Datenanalyse. Es kann dann erstmals der Aufwand mit dem Nutzen in Bezug gesetzt werden. Nach dem Lernschritt folgt hier nicht mehr nur die Interpretation durch einen Experten, sondern die Steuerung von realweltlichen Prozessen beispielsweise in Logistik (A4), Produktion (B3), Verkehr (B4) und physikalischen Experimenten (C3, C5).

Förderperioden: Gerade haben wir die Forschungsfragen nach den Schritten der Informationsgewinnung gegliedert. Jetzt fassen wir sie für die Förderperioden des SFB 876 zusammen.

1. Phase: In den ersten Jahren hat der SFB 876 eine gemeinsame Arbeitsgrundlage für die interdisziplinäre Forschung zur Informationsgewinnung durch ressourcenbeschränkte Analyse geschaffen. Dabei ist hier die Interdisziplinarität nicht nur durch die Kooperation einer Methoden- mit einer Substanzwissenschaft entstanden. Vielmehr arbeiten auch verschiedene Methodenausrichtungen zusammen, die jeweils andere Grundlagen, Terminologien und Herangehensweisen aufweisen. In der ersten Phase haben maschinelles Lernen, Data Mining, Big Data Analytics, Statistik, Cyber-Physical Systems, Elektrotechnik, Physik und Bioinformatik sich gegenseitig ihre Methoden erklärt, gemeinsame Erkenntnisinteressen

präzisiert. Hier wollen wir zur griffigen Charakterisierung der ersten Förderperiode nur einige zentrale Themen anführen:

- Das Thema Energieeffizienz hat sich zu einem Schwerpunkt entwickelt, der weiter entfaltet wird. Zur Illustration führen wir hier nur einige in enger Zusammenarbeit entstandene Arbeiten an. Bei Mobilgeräten wurden in A1 zur Positionsbestimmung Methoden unterschiedlichen Energiebedarfs und entsprechend unterschiedlicher Präzision untersucht, um die Systemsoftware zu befähigen, die Lokalisierung dem Ressourcenverbrauch und Qualitätsanforderungen anzupassen. A4 entwickelte das *Context Aware Power Consumption Model* für LTE-basierte Mobilfunkendgeräte. Techniken des Teilprojektes A3 benutzen das Wissen über den Energieverbrauch von Assembleroperationen für energieoptimierende Compiler. Mit dem neu gewonnenen Kollegen Jian Jia Chen als Nachfolger von Peter Marwedel wird das Thema der Energie und Thermik im Projekt B2 noch weiter voran getrieben.
- Diskussionen im Arbeitskreis Datenanalyse zeigten das Potenzial, das die Beschäftigung mit graphischen Modellen im SFB hat. Die Entwicklung effizienter Algorithmen für probabilistische graphische Modelle wurde von A1 und C1 gemeinsam begonnen. Dabei geht es um *Markov Random Fields*, die gemäß des Ressourcenverbrauchs von Kommunikation, Speicherzugriff, Arithmetik bestimmter Ausführungsplattformen optimal sind. Die neue Methode der *Spatio-Temporal Random Fields* wurde für die Prognose des Energiebedarfs bei Smartphones entwickelt, eignet sich aber auch zur Modellierung von Verkehrsflüssen, wie sie in B4 betrachtet werden. Das neue Projekt A6 wird das Thema durch die Analyse von Daten vertiefen, die in Form von Graphen vorliegen.
- Für die Analyse von sehr großen und/oder hochdimensionalen Daten wurden zum einen Datenstromalgorithmen (in A1 und A2), zum anderen Methoden zur geeigneten Reduktion der Datenmengen beispielsweise durch Kernmengen (A2, C4) untersucht. Gerade die Physik bringt besonders große Datenmengen ein. Das Projekt C3 erlebte die besondere Würdigung der IceCube Kollaboration, der ihr Projektleiter Wolfgang Rhode angehört, als *Breakthrough of the Year* 2013 (Physics World). Hier führt sich – wie oben schon ausgeführt – das Speichern der Massendaten als Forschungsthema ein. Der neu gewonnene Kollege für Informationssysteme (Nachfolger von Joachim Biskup), Jens Teubner, wird in dem neuen Projekt C5 dieses Thema in den SFB einbringen.

2. Phase: In der zweiten Förderperiode werden die Fragestellungen untersucht, die schon bei dem Einrichtungsantrag für die zweite Phase angegeben waren, sich nun aber konkretisiert haben:

- Wie können Terabyte einströmender Daten gefiltert, bereinigt und gespeichert und so aggregiert werden, dass sie einerseits mit der jeweiligen Speicherbeschränkung auskommen und andererseits für die weitere Analyse gut geeignet sind?
- Wie kann man die Analyse sehr großer, hochdimensionaler, dynamischer und (nach Raum und Zeit) strukturierter Daten so dekomponieren, dass sie auf lokale Geräte und Großrechner verteilt wird, ohne die Kommunikationsbeschränkung zu verletzen?
- Wie kann man die Sensoren und Prozessoren vor Ort für die Durchführung von Analyseschritten verbessern, ohne dass viel Energie verbraucht wird?

3. Phase: In der dritten Förderperiode sollen die Erfahrungen mit parallelen, verteilten und Datenstromalgorithmen unter Betrachtung der Speicher-, Kommunikations-, Energiebeschränkung und Realzeitanforderungen zu nutzbaren Grundlagen für die universitäre Lehre und spezifische Forschung für Anwendungen in den Bereichen der Physik, der Logistik und der Biomedizin werden.

Hervorgehobene Publikationen und Ehrungen: Mehrere Mitglieder des Sonderforschungsbereichs wurden dadurch ausgezeichnet, dass sie in Anerkennung ihrer Leistungen Rufe erhalten haben.

- Alexander Skopalik war schon an der Vorbereitung des SFB beteiligt und vom 1.10.2011 bis 30.9.2012 im Teilprojekt A2 beschäftigt. Er hat dort sein Wissen über die algorithmische Spieltheorie in die Modellierung von eigennützigem Verhalten in verteilten Systemen einfließen lassen. Zum 1.10.2012 nahm er einen Ruf auf eine Juniorprofessur für algorithmische Spieltheorie an die Universität Paderborn an.
- Jan Vahrenhold hat im Teilprojekt A2 auf der Basis geometrischer Modellierung an Speichermodellen und einer speichereffizienten Darstellung für k-Means gearbeitet. Seine hervorragenden Leistungen wurden auch dadurch anerkannt, dass er auf eine W3-Professur der Universität Münster berufen wurde. Er nahm den Ruf an.

- Sven Rahmann erhielt einen Ruf an das Universitätsklinikum Essen, den er annahm. Eine strategische Entscheidung der Universitätsallianz Ruhr (UA Ruhr) zur Stärkung des SFB 876 führte zur ersten Professur der Mercator-Stiftung an der UA Ruhr, die der *Computational Biology* gewidmet wurde und Sven Rahmann übertragen wurde. Seit 2014 vertritt er sowohl an der TU Dortmund als auch am Klinikum Essen die Bioinformatik bzw. Genominformatik.

Der Sonderforschungsbereich 876 hat mit insgesamt 313 Publikationen seine Forschungsergebnisse bekannt gemacht.

- Das Lehrbuch von Peter Marwedel *Embedded System Design – Embedded Systems Fundation of Cyber-Physical Systems*[1] wird weltweit verwendet, es gibt sogar eine chinesische Ausgabe. Damit ist Peter Marwedel im Microsoft Ranking der Autoren für den Hardware- und Architekturbereich in Europa an 8. Stelle und weltweit auf Platz 88.[2]

- Eine aus einem Gastaufenthalt am SFB 876 entstandene Publikation hat ihren Weg in ein Grundlagenbuch gefunden: *Managing and Mining Sensor Data*, Charu Aggarwal (Hg.) 2013 enthält das Kapitel von Marco Stolpe und Kanishka Bhaduri *Distributed Data Mining in Sensor Networks.*

- Best Student Paper Award der wichtigsten europäischen Tagung zu maschinellem Lernen und Data Mining, ECML PKDD, 2013: Nico Piatkowski, Sangkyun Lee, Katharina Morik (A1, C1);

- Best Student Paper Award IEEE SCVT 2013 (Symposium on Communications and Vehicular Technology in the Benelux), 2013: Brian Niehöfer, Florian Schweikowski, Christian Wietfeld (B4);

- Best Paper Award der Gesellschaft für Klassifikation, 2013: Tobias Voigt, Roland Fried, Michael Backes (C3).

- Auch Abschlussarbeiten unserer Studenten wurden herausgehoben. Den Hans-Uhde-Preis erhielten am 30.03.11 Benedikt Konrad (B3) und am 1.3.2012 Christoph Borchert (A4).

Besondere Ehrungen haben einige Projektleiter erhalten: Michael Schreckenberg wurde am 5.3.2011 in die Europäische Akademie der Wissenschaften und Künste aufgenommen. Johannes Schulte (C1) wurde mit dem Friedrich-Linneweh-Preis am 8.10.2012 ausgezeichnet. Peter Marwedel (A3, A4, B2) erhielt den herausragenden *EDAA Lifetime Achievement Award* am 19.3.2013. Joachim Biskup (A5) wurde von der internationalen Informatikgruppe IFIP WG 11.3 mit dem *Outstanding Research Contributions Award* am 17.7.2013 geehrt. Christian Wietfeld (A4, B4) erhielt von der informationstechnischen Gesellschaft im VDE den ITG-Preis 2013.

Gliederung und beteiligte Teilprojekte

Die Struktur des SFB 876 bleibt gleich, denn die Gliederung in die drei Bereiche hat sich bewährt. Aufbauend auf Erfolgen der TU Dortmund und kollegialem Austausch, haben wir neue Möglichkeiten aufgegriffen. Die genaue Festlegung der Projekte für die zweite Förderperiode erfolgte in einem internen Prozess der Qualitätskontrolle. Es wurden fünf Antragsskizzen für zusätzliche Projekte bis zum Juni 2013 beim Vorstand eingereicht. In einem intensiven zweitägigen Workshop, 10./11.7.2013, wurden alle Teilprojekte, bestehende wie neu eingereichte Antragsskizzen, gemeinsam diskutiert. Die Kriterien waren: die Exzellenz der Idee, ihr Nutzen für den SFB 876 und die Kooperation mit anderen Teilprojekten, die längerfristige Perspektive, der bisherige Beitrag der Antragstellerinnen oder Antragsteller zum Forschungsstand, die Publikationssituation und die Durchführbarkeit an der TU Dortmund. Ausführliches Feedback wurde anonym per Webformular abgegeben. Daraus entstand eine Diskussion, die schließlich auf dem Workshop im Oktober 2013 Synergien herausarbeitete.

So wurden zwei bestehende Teilprojekte durch neue Projektleiter ergänzt und verstärkt. Kristian Kersting wurde als zusätzlicher Teilprojektleiter von B4 aufgenommen, um dort seine Analysekompetenz einzubringen. Michael ten Hompel verstärkt das Plattformprojekt A4, indem er die besonders ressourcenbeschränkten inBins und Fahrzeuge aus der Logistik einbringt.

Das erfolgreiche Teilprojekt A5 läuft aus. Es hat algorithmische Lösungen mit beweisbaren Sicherheitseigenschaften gefunden. Für das logikbasierte Szenario mit einem anfragenden und einem die Vertraulichkeit

[1] Springer 2nd Ed. 2011

[2] http://academic.research.microsoft.com/RankList?entitytype=2&topDomainID=2&subDomainID=3&last=0&start=1&end=100

verteidigendem Agenten wurden Inferenzen modelliert, die für Anfrage- und Änderungsoperationen auf Datenbanken die Vertraulichkeit wahren. Die Arbeit ist abgeschlossen. Die Ergebnisse stehen weiterhin zur Verfügung. Im weiteren Verlauf des Sonderforschungsbereichs wird an anderen Datenbankkonzepten gearbeitet. Die Szenarien werden die Inferenz betreffend einfacher, bezüglich der Datenmengen, die auch verteilt sein können, sowie der Realzeitanforderungen anspruchsvoller.

Die Teilprojektleiter wechseln bei einigen Teilprojekten, insbesondere um neue Errungenschaften aufzugreifen. So werden die neuen Möglichkeiten aufgegriffen, die durch die Ernennung von Sven Rahmann als Professor der UA Ruhr für das Teilprojekt C1 entstehen. Herr Rahmann wird Teilprojektleiter im Projekt C1, in das er eigene Stellen für Wissenschaftler und Geräte für die neuartige Sequenzierung einbringt. Dafür macht Katharina Morik ihre Projektleiterposition bei C1 frei für den Nachwuchswissenschaftler Sangkyun Lee, der sich mit seinen Kenntnissen in Bioinformatik, Statistik und Optimierung in der ersten Phase als idealer Kandidat dafür empfohlen hat. Die Datenanalyse für die Atemluft im Projekt B1 hatte Sven Rahmann in der ersten Phase bezogen auf die Merkmalsextraktion aus den Spektraldaten zu einem Zwischenstand gebracht, der die Übergabe der Analyseaufgaben in an Jörg Rahnenführer ermöglicht. Im Abschnitt *Personelle Situation* werden wir genauer auf die Projektleiterschaften, insbesondere auch einen weiteren Nachwuchswissenschaftler als Projektleiter eingehen.

Es gibt zwei neue Teilprojekte. Das Teilprojekt A6 ermöglicht Kristian Kersting und Petra Mutzel ihre Behandlung von Daten, die in Form von Graphen vorliegen, in Zusammenarbeit mit Christian Sohler in den SFB einzubringen. Auf diese Weise wird das Thema der Graphen verstärkt. Das Teilprojekt C5 hat das Angebot von Bernhard Spaan aufgegriffen, die enormen Datenmassen des CERN zu nutzen. Hier finden die Speicherverfahren von Jens Teubner herausfordernde Aufgaben, die eine Erprobung seiner Lösungsansätze ermöglichen. Damit findet nun das Thema der Speicherung sehr großer dynamischer Datenmassen seinen Platz im SFB.

Der Sonderforschungsbereich setzt nun ein Teilprojekt nicht fort und hat zwei neue Teilprojekte eingeführt, die die Themen der drei Bereiche verstärken und ergänzen. Auf der Mitgliederversammlung am 17.Oktober wurden die Teilprojekte für die zweite Förderperiode im Einklang mit dem Vorstand festgelegt.

Projektbereich A: Die Teilprojekte des A-Bereichs bilden gemeinsam das Fundament für die *Cyber-Physical Systems* des B-Bereichs und die C-Projekte zu *Big Data*. Es gibt zum einen die beiden Projekte A2 und A6, die theoretische Grundlagen der Datenanalyse für die Informationsgewinnung unter Ressourcenbeschränkung erarbeiten. A6 ist ein neues Teilprojekt, das die Analyse von Daten, die in Form von Graphen vorliegen, untersucht. Wir wollen dabei randomisierte Verfahren aus der Algorithmentheorie, Graphalgorithmen und Graph Mining zusammenbringen. Die in A2 erarbeiteten theoretischen Analysen sollen dort entsprechend erweitert werden. Die Beschäftigung mit probabilistischen graphischen Modellen für ressourcenbeschränkte Datenanalyse begann im Teilprojekt A1 in der ersten Förderperiode und wird in der zweiten dort fortgeführt. Neue Algorithmen für *Conditional Random Fields* und *Markov Random Fields* sollen nun durch approximative Algorithmen ergänzt werden. Ziel ist es hierbei, für ganz unterschiedliche Ressourcenbeschränkungen Berechnungen der maximalen *Likelihood* zu entwickeln und ihre Korrektheit und Vollständigkeit mit dem Ressourcenverbrauch in Bezug zu setzen. A1 verbindet das maschinelle Lernen mit Cyber-Physical Systems. Es ist in sich bereits interdisziplinär und untersucht alle oben aufgeführten Schritte. Daher steht es auch mit den beiden Projekten A4 und A3 im Zusammenhang, die Methoden der technischen Informatik für die Informationsgewinnung unter Ressourcenbeschränkung erarbeiten. Die Messung und Untersuchung des Ressourcenverbrauchs bei unterschiedlichen Ausführungsplattformen steht in A4 im Vordergrund. Dabei wird nun mit der Logistik ein besonders komplexes Szenario einbezogen, an dem sich auch kleine, verteilte, mobile Systeme untersuchen lassen. Der Sonderforschungsbereich nutzt dabei auch die Stärke des Fraunhofer Instituts für Materialfluss und Logistik (IML) und den weltweit einzigen LogistikCampus an der TU Dortmund, der am 22.4.2013 eröffnet wurde. Eine Versuchshalle für zellulare Fördertechnik, kleine Fahrzeuge für die Bewegung in Kleinteilelagern (Rack Racer) und Pakete mit stark beschränkten Prozessoren (inBins) stehen nun zur Verfügung, um Methoden der Merkmalsextraktion, Datenzusammenfassung und verteilten Datenanalyse in Bezug auf Kommunikation und Energie als beschränkte Ressourcen zu studieren. A3 ist in sich interdisziplinär. Es nutzt Methoden eingebetteter Systeme für die statistische Modellselektion so wie A1 Methoden der Datenanalyse für eingebettete Systeme nutzt. Die Verzahnung der Bereiche und das wechselseitige Verständnis der beteiligten Disziplinen hat in der ersten Förderperiode zugenommen und damit eine hervorragende Möglichkeit für die zweite Förderperiode geschaffen.

Projektbereich B: Der Bereich der B-Projekte ist den technischen Ressourcenbeschränkungen gewidmet. *Cyber-Physical Systems* werden in den Projekten TB1 und B2 sogar physikalisch hergestellt. Die Geräte für die Atemluftdiagnose (TB1) liefern Ionenmobilitätsspektren. Die Geräte wurden in der ersten Förderperi-

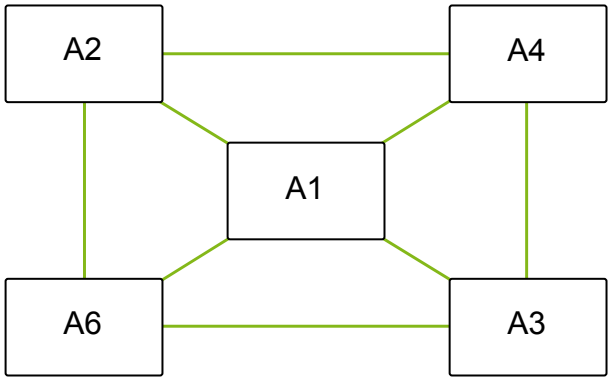


Abbildung 1.1.: Übersicht über den Bereich der A-Projekte: links die beiden Projekte zu theoretischen Grundlagen der Datenanalyse, rechts die beiden Projekte mit dem Fokus auf Cyber-Physical Systems, in der Mitte das Leitprojekt. Die Linien geben inhaltliche Verbindungen an.

ode bereits erfolgreich verkleinert. In der zweiten Förderperiode werden Vorbereitungen für eine Integration in Kleidung und Telemedizinanbindung getroffen. Dies schließt eine weitere Miniaturisierung aller Spektrometerkomponenten von der Probenahme bis zur modularen bedarfsbezogenen Hardwareausgestaltung unter Ressourcenbeschränkungen ein. Durch die Einbeziehung der Dortmunder Firma B&S, von der die Geräte produziert werden, und aufgrund der Fülle von Erfahrungen mit der Atemluftanalyse von Jörg Ingo Baumbach kann der SFB 876 nun nicht nur Daten analysieren, die von einem kleinen System geliefert werden, sondern es kann auch rückwirkend aus der Merkmalsextraktion und Datenanalyse heraus die Hardware verbessert werden. Die Spektroskopiedaten sind Wertereihen, bei denen zunächst die *peaks* als Merkmale extrahiert wurden. Grundlage für die Klassifikation ist die Merkmalsextraktion und -selektion, die auch in der zweiten Phase weiter untersucht wird. Gerade für die weitere Miniaturisierung und die Steuerung der Probenahme muss eine minimale Anzahl von Signalparametern bestimmt werden.

Das Teilprojekt B2 arbeitet mit einer neuartigen Technik, der Plasmonen-unterstützten Mikroskopie von Nanoobjekten (PAMONO). Das Leibniz-Institut für Analytische Wissenschaften an der TU Dortmund (ISAS e.V.) hat dieses Gerät entwickelt und kann seine Verbesserung auch physikalisch umsetzen. Deshalb kann nun anhand der Datenanalyse die Ausführungsumgebung angepasst werden, während gleichzeitig die Datenanalyse auf die Ausführungsplattform abgestimmt wird. Die Datenanalyse verwendet hier Methoden der Bildanalyse, insbesondere zur Merkmalsextraktion und -selektion. Ausreißererkennung und Methoden zu *Concept Shift* und *Concept Drift* sollen auf den Sensor bezogen werden, um Messfehler und den Bedarf der Kalibrierung festzustellen. Durch die enge Verbindung mit dem realen System kann auch die Ressourcenschonung besonders gut studiert werden. Die Verknüpfung von Programm- und Plattformebene wird so gestaltet, dass unterschiedliche Scheduler und Rechennetzwerke verglichen und bezüglich der Ressourcen bewertet werden können. Besonders hervorzuheben ist das dynamische Temperaturmanagement der mobilen Sensoreinheiten, das in B2 untersucht werden soll. Die Geräteentwicklung zielt vor allem auf die Mobilität und damit auf die Miniaturisierung ab.

B3 vertritt den unter dem Schlagwort *Industrie 4.0* in Deutschland bekannt gewordenen Forschungsbereich, bei dem die Datenanalyse in die physikalische Produktion integriert wird. Der Sonderforschungsbereich nutzt die Kontakte des Projektleiters Jochen Deuse. So ist es möglich, dass wir mit realen Daten aus den Edelstahlwerken arbeiten und in Zusammenarbeit mit ehemaligen Promovenden dort vor Ort auch Ergebnisse des SFB langfristig in die Produktion integriert und anhand der Erfahrungen verbessert werden können. Die verteilte Datenanalyse steht im Vordergrund. Qualitätsmängel sollen frühzeitig erkannt werden, damit nicht den Anforderungen entsprechende Stahlstäbe direkt ausgeschleust oder die Auftragsplanung geändert werden kann. In der ersten Förderperiode wurde verteiltes Clustering und die 1-Klassen SVM mit der *Core Vector Machine* für geringere Kommunikationskosten weiter entwickelt. Eine Besonderheit ist, dass es keine abgrenzbaren Objekte in den Daten gibt, weil flüssiger Stahl in Stäbe aufgeteilt wird, ohne dass man kennzeichnen könnte, woher genau ein Stab kam. In der zweiten Phase soll von der verteilten Datenanalyse aus auch eine Regelung und aktives Lernen für gezielte Datenerfassung entwickelt werden. Trade-Offs zwischen Laufzeit, Speicherverbrauch, Kommunikationskosten, Energieverbrauch und Prognosegüte werden für den Fall untersucht, dass Sensordaten dezentral gespeichert und analysiert werden.

B4 behandelt Prognoseprobleme des Straßenverkehrs. Es werden dynamische, mikroskopische Verkehrs-

modelle vorgesehen, die basierend auf gelernten Empfehlungen in Echtzeit neu parametrisierbar sind und damit die Heterogenität des Verkehrs in der innerstädtischen Netzstruktur besser abbilden können. Das langfristige Ziel ist dabei, die durchschnittlichen Fahrzeiten und den Energiebedarf zu reduzieren. Hierbei sollen relationale Modelle wie z.B. Markov Logic Networks untersucht werden, in denen die Vorhersage der Fahrzeit auf einem Straßenabschnitt die Vorhersage an benachbarten oder ähnlichen Segmenten beeinflusst. Die raumzeitlichen Modelle aus A1 können genutzt werden. Die methodische Weiterentwicklung wird aber durch die Zusammenführung heterogener Datenquellen herausgefordert. In der ersten Phase ging es vorrangig um Kommunikationsmodelle und ein mikroskopisches Verkehrsmodell. In der nächsten Förderperiode ist nun auch ein Projektleiter für die Datenanalyse beteiligt, Kristian Kersting, der gerade auf statistisch-relationale Verfahren spezialisiert ist (siehe auch den Abschnitt *Personelle Situation*). Die Datenzusammenfassung und -analyse soll in tatsächliche Handlungsempfehlungen für die Fahrzeuge münden.

Projektbereich C: Der Projektbereich C behandelt sehr große Datenvolumina. Der Bereich kann als Big Data bezeichnet werden, vor allem als *Big Data Analytics*. Es gilt, Algorithmen zur Datenspeicherung, -zusammenfassung und -analyse zu entwickeln, die mit sehr hoher Dimensionalität und sehr hoher Beispielzahl umgehen können. Das Teilprojekt C1 behandelt hochdimensionale Daten mit sehr wenigen Beobachtungen. In der ersten Förderperiode wurden stabile und robuste Verfahren zur Merkmalsselektion entwickelt und ein algorithmisch elegantes Plug-in für RapidMiner veröffentlicht. Dennoch konnte – wie auch an anderen Orten – eine zufriedenstellende Prognose der Überlebenszeit anhand von Exon-Expressionswerten bislang nicht erreicht werden. Eine weitere Verbesserung durch Anwendung anderer Lern- und Selektionsverfahren erscheint unwahrscheinlich. Stattdessen soll die Datenlage verbessert werden. Hier stehen nun durch den neuen Projektleiter, Sven Rahmann (siehe Abschnitt *Personelle Situation*) umfassendere genetische Daten zur Verfügung, die die Expressionsprofile der ersten Förderperiode an Aussagekraft übersteigen. Diese Resultate der Gensequenzierung liefern zwar jetzt noch mehr Daten, zeigen aber vielleicht die für die Krebstherapie entscheidende Information. Die Merkmalsextraktion muss auf großen Rechen-Clustern oder Multicore-Servern stattfinden. Wieviel Energie und Laufzeit sind nötig für die Merkmalsgewinnung aus *Next Generation Sequencing* Daten? Fasst man die Merkmale als Knoten und die Kanten als signifikante Beziehungen zwischen Merkmalen auf, ergibt sich ein Graph, in dem spärlich besetzte Subgraphen gefunden werden sollen. Probabilistische graphische Modelle, wie sie bereits in Zusammenarbeit mit A1 in der ersten Förderperiode entwickelt wurden, sollen hier weiter entwickelt werden. Die *Latent Dirichlet Allocation*, die gerade im Bereich der Big Data Analytics vielversprechend ist, soll hier zur zeitlichen Modellierung genutzt werden.

Das Teilprojekt C4 beschäftigt sich ebenfalls mit hochdimensionalen Daten, insbesondere bezüglich ressourcenbeschränkter Regression. In diesem Projekt soll die Theoriebildung interdisziplinär von Statistik und Informatik vorangetrieben werden. Zunächst steht die Datenzusammenfassung im Vordergrund, die hier vor allem mit Hilfe zufälliger linearer Projektionen betrieben wird. Effiziente Algorithmen für die Regression bieten in der interdisziplinären Zusammenarbeit von zwei Methodenwissenschaften eine Möglichkeit, zu neuen Einsichten über Bayessche Regression, Dimensionsreduktion und Merkmalsselektion zu gelangen. Diese Theoriebildung wird in der nächsten Förderperiode fortgesetzt und erste Ergebnisse in anderen Teilprojekten verwendet (TB1, C1, C3).

Das Teilprojekt C3 setzt sich mit hochfrequenten Teleskopdaten auseinander. Die Daten von Gamma-Teleskopen, wie MAGIC oder FACT, und die des IceCube-Neutrinoteleskops weisen die Nachteile sehr großer Datenmengen auf. Statt, wie es zum Beispiel bei Sprachdaten der Fall ist, die Muster deutlicher hervortreten zu lassen, verlangen sie nun, die Nadel im Heuhaufen zu suchen, da die uninteressanten Untergrundereignisse dominieren. In der ersten Förderperiode wurden sehr erfolgreich Datenströme direkt vom FACT-Teleskop aufgegriffen, visualisiert und mit einfachen Lernverfahren auf Datenströmen für eine Merkmalsextraktion zusammengefasst. Für IceCube wurde eine sorgfältige Evaluation der Analyse mit dem *Random-Forest*-Verfahren auf bereits dimensionsreduzierten Daten erreicht, mit der die Erkennungsrate von Neutrinos erheblich gesteigert werden und erstmals das Spektrum atmosphärischer Neutrinos bis zu einem Energieniveau von 1 Petaelektronenvolt rekonstruiert werden konnte. Da sämtliche Analysen nur unter Zuhilfenahme simulierter Ereignisse durchgeführt werden können und die Simulation dieser Ereignisse allein innerhalb der am SFB teilnehmenden Arbeitsgruppe etwa 5 Mio. CPU-Stunden pro Jahr erfordert, kommt der effizienten und ressourcenschonenden Produktion von Monte-Carlo-Ereignissen eine entscheidende Rolle zu. In der zweiten Förderperiode wird das Projekt umgestellt. Die Arbeit an Datenstromalgorithmen für das FACT-Teleskop, die in einer Zusammenarbeit mit dem Projekt A1 angefangen hat, kann nicht mehr nebenher vorangetrieben werden. Die gemeinsame Arbeit an bestimmten Entfaltungsmethoden des inzwischen promovierten Tim Ruhe ist so vielversprechend, dass er nunmehr die Aufgabe als Teilprojektleiter in C3 übernehmen soll. Die erfolgreichen Methoden zur Klassifikation in Untergrund und Signal, die von Roland Fried erarbeitet wurden, können genutzt werden. Ihre algorithmische Weiterentwicklung zur Skalierbarkeit für sehr große

Datenmengen muss jedoch in der Informatik erfolgen (siehe Abschnitt *Personelle Situation*).

C5 ist das Ergebnis von informellen Gesprächen zu datengestützter Wissenschaft. Durch die Besetzung der Nf. Biskup mit Jens Teubner, einem ausgewiesenen Experten für das effiziente Speichern von Big Data, ist die Möglichkeit gegeben, anhand der Daten des bekannten *Large Hadron Colliders* in Genf die schnelle Filterung und Zusammenfassung von Petabyte an Daten und den Nutzen von *MapReduce* und *Hadoop* zu studieren. Obendrein sollen Ausführungsplattformen systematisch bezüglich der verteilten Datenhaltung untersucht werden. Durch Datenhaltung, -zusammenfassung und -analyse sollen – wie schon im Teilprojekt C3 – Erkenntnisse der Physik gewonnen werden. Die Zusammenarbeit zwischen Physik und Informatik wird in den beiden Projekten C3 und C5 den Bereich der Cyber-Physical Systems verstärken. Während C3 auf Datenströme setzt, geht es in C5 um die persistente Speicherung. Beide Projekte können insofern voneinander profitieren, als die Zusammenfassungen der Datenströme von C3 auch ressourcenschonender gespeichert werden und die Trigger von C5 auch als Datenstromalgorithmen realisiert werden könnten.

Kooperation innerhalb des Sonderforschungsbereichs

Die Kooperation zwischen den Teilprojekten des SFBs wird in vielfältiger Weise befördert. Die übergreifende Zusammenarbeit ist in den Teilprojektanträgen konkret ausgearbeitet und zeigt die enge Vernetzung auf (siehe Tabelle 1.7). Es gibt bereits 31 Publikationen aus Zusammenarbeiten verschiedener Teilprojekte. Dies ist insofern viel, als bereits die Arbeit innerhalb eines Projektes meist interdisziplinär ist. Der SFB 876 organisiert die interne Zusammenarbeit in drei Arbeitskreisen. Die Arbeitskreise bearbeiten die Themenfelder Informationsgewinnung (INF), Ressourcenbeschränkungen (RES) und Datenanalyse (DAT). Sie treffen sich im Wechsel jeweils alle 3 Wochen. Jeder Wissenschaftler im SFB arbeitet mindestens in einem der Arbeitskreise aktiv mit. Durch die Bildung von Teilprojekten mit jeweils zwei oder teilweise sogar drei Teilprojektleitern ist gewährleistet, dass bereits innerhalb der Teilprojekte unterschiedliche Forschungsfelder des SFBs *interdisziplinär* zusammenarbeiten. Gleichzeitig bieten die Arbeitskreise die Gewähr dafür, dass der Diskurs *innerhalb* eines spezialisierten Forschungsfelds intensiv betrieben werden kann. So werden einerseits innovative Lösungsansätze über Grenzen der Fachgebiete hinweg erzielt (wie z.B. der erstmalige Einsatz moderner Datenanalysemethoden in der Astropartikelforschung) und anderseits auch die Wettbewerbsfähigkeit und Passung innerhalb der eigenen fachlichen Community der Wissenschaftler des SFBs gesteigert.

Teilprojekt	Benötigt/profitiert von Ergebnissen aus	Mitarbeit in Arbeitskreis
A1-UbiDM	A2, A3, A4, A6, B4, C1	INF, RES, DAT
A2-Theorie	A1, A4, A6, B2, B3, C3, C4	INF, DAT
A3-StatES	A1, A2, B3, C1, C4	INF, RES
A4-Plattform	A1, A2, A6, B2, B4	INF, RES
A6-Graph	A2, A4, B4, C1	INF, DAT
TB1-Spektren	A1, A3, A4, C1, C4	DAT, RES
B2-Nano	A1, A2, A3, A4, TB1, B2, C3	RES, DAT
B3-SensorDM	A1, A2, A3, A4, A6, B4, C3	DAT, INF, RES
B4-Mobil	A1, A2, A4, A6, B3, C4	INF, RES, DAT
C1-DimRed	A3, A6, C3, C4	INF, DAT
C3-RaumZeit	A1, A2, B2, C1	INF, DAT
C4-Regress	A3, TB1, B4, C1	INF, DAT
C5-Speicher	A1, A2, C3	INF, DAT

Tabelle 1.7.: Tabellarische Aufstellung der Zusammenarbeit zwischen den Teilprojekten des Sonderforschungsbereichs mit Angabe der Teilnahme in den Arbeitskreisen Informationsgewinnung (INF), Ressourcenbeschränkungen (RES) und Datenanalyse (DAT)

Die aktive Mitarbeit in den Arbeitskreisen zeigt sich darin, dass eigene Forschungsergebnisse zur Diskussion gestellt werden, von Besuchen internationaler Konferenzen berichtet wird oder an der Einladung von Gästen mitgewirkt wird, die neue Impulse von außen in den SFB einbringen. Die Arbeitskreise bieten weiterhin eine Plattform zur Vorstellung und Diskussion von innovativen Methoden und Werkzeugen, die

in mehreren Teilprojekten des SFBs zum Einsatz kommen können (z.B. Methoden zur Messung von sehr geringen Leistungsaufnahmen zur Bestimmung des Energieverbrauchs). Zu bestimmten Themenbereichen ist auch die Verknüpfung von Arbeitskreisen sinnvoll, z.B. durch gemeinsame Veranstaltungen der Arbeitskreise Ressourcenbeschränkungen und Datenanalyse zu Verfahren der Datenanalyse auf hochparallelen Systemen.

Neben den Arbeitskreisen bildet das Graduiertenkolleg eine weitere wichtige Plattform für die übergreifende Zusammenarbeit im SFB. Über Veranstaltungen zu Methoden (z.B. Präsentationstechniken) und die Mitarbeit an der Organisation der Sommerschulen sowie das gemeinsame intensive Studium bei der Sommerschule selbst findet eine teilprojektübergreifende Verzahnung der wissenschaftlichen Arbeit der Promovenden statt.

Zentrale Mittel: Bereits vor Beginn der ersten Phase des Sonderforschungsbereichs stand fest, dass die Nachfolgen für Peter Marwedel und Joachim Biskup während der Förderperiode neu ausgeschrieben und besetzt werden sollten. Hinzu kam die Ausschreibung zur Nachfolge der Professur von Sven Rahmann, die ebenfalls an der weiteren inhaltlichen Stärkung des SFB ausgerichtet war. Um den in der beantragten zweiten Phase neu am Sonderforschungsbereich zu beteiligenden Projektleitern eine effektive Integration in die Forschungsvorhaben zu ermöglichen, waren insbesondere die zentral verwalteten pauschalen Mittel sehr hilfreich. Damit konnten die neuen Professoren bereits vor Ende der ersten Förderperiode personell in ähnlichem Umfang wie die Projektleiter der laufenden Projekte unterstützt werden, während die TU Dortmund das entsprechende Maß für die personelle und sachliche Grundausstattung zur Verfügung stellte. Details zur Verwendung der Mittel sind im Projekt Z, bzw. für Gäste und Workshops auch im integrierten Graduiertenkolleg aufgelistet.

1.2.3. Positionierung des Sonderforschungsbereichs in seinem weiteren Fachgebiet

International ist das Schlagwort *Big Data* sehr populär, wie die Trend-Grafik von Google (28.6.2014) zeigt (Abbildung 1.2), und zwar ungefähr seit Beginn der ersten Förderperiode des SFB 876. Der Bundesverband

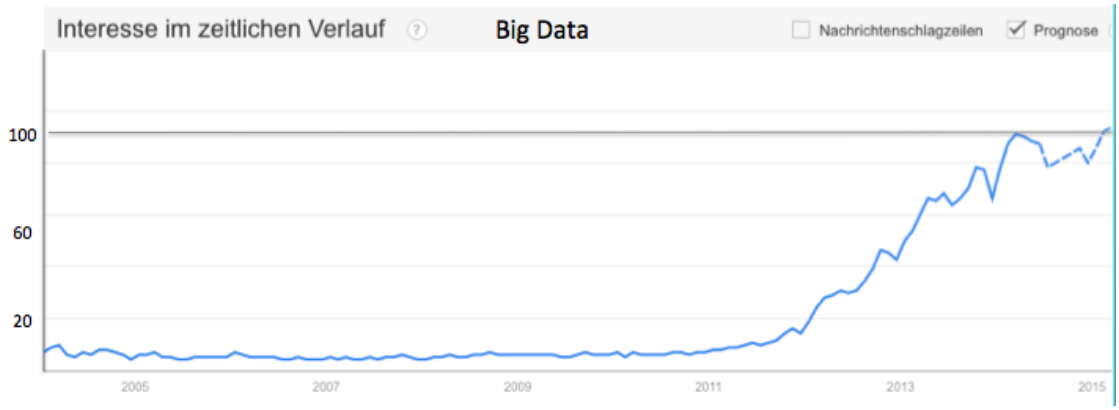


Abbildung 1.2.: Google Trends 28.6.2014: Die Zahlen in der Grafik geben an, wie oft im Vergleich zu allen im Laufe der Zeit auf Google gestellten Suchanfragen nach einem bestimmten Begriff gesucht wurde. Sie stellen keine absoluten Suchvolumenzahlen dar, da die Daten normalisiert sind und auf einer Skala von 0 bis 100 angezeigt werden. Jeder Punkt in der Grafik wird durch den Höchstwert oder durch 100 geteilt.

Informationswirtschaft, Telekommunikation und neue Medien e.V., BITKOM, beschrieb Daten als den Rohstoff der digitalen Welt, aus dem durch Big Data Analytics ein hoher Nutzen erzeugt werden kann[3]. Als Beispiele des Nutzens wird interessanterweise die Verkehrssteuerung und die Krebstherapie genannt – zwei Szenarien, die auch im SFB 876 vertreten sind. Die Fraunhofer-Allianz Big Data bündelt 25 Institute unter der Leitung von Prof. Dr. Stefan Wrobel. Ihre 2013 vorgestellte Studie zu Innovationspotenzialen untersucht Ziele und Branchen und stellt 55 internationale Big-Data-Anwendungsfälle mit den jeweils notwendigen Analysefähigkeiten vor. Als wichtige Aufgabe für die Universitäten wird die Ausbildung in den Datenwissenschaften herausgestellt. In zahlreichen Strategiepapieren wird auf die Bedeutung von Big Data und Big Data Analytics hingewiesen. An dem europäischen Strategiepapier zu Big and Open Data war Katharina

[3]Pressemitteilung vom 6.3.2013 http://www.bitkom.org/files/documents/BITKOM_Vortrag_PK_Big_Data.pdf

Morik beteiligt, insbesondere bezüglich der Analyse von Datenströmen[4]. Das Thema *Big Data* umfasst moderne Datenbanken, soziale Netzwerke, interaktive Exploration und Visualisierung, Datenanalyse und neue Programmiermodelle wie Datenstromalgorithmen und MapReduce sowie geeignete Ausführungsplattformen. Dank der neuen Projekte A6 für Netze und C5 für das Speichern von Datenvolumina im Petabyte-Bereich sind nun alle diese Themen auch im SFB 876 vertreten. Die Herkunftsdisziplinen Datenbanken, maschinelles Lernen und Data Mining beschäftigen sich mit Big Data, wenn sie sich besonderen Herausforderungen durch ein großes Datenvolumen, Heterogenität und Verteiltheit von Daten sowie die Geschwindigkeit, mit der Signale eintreffen, stellen. Wir können jetzt sagen: wenn sie zusätzlich die Ressourcenbeschränkungen betrachten. Algorithmen sollen nicht nur bezüglich ihrer Korrektheit und Vollständigkeit, sondern auch bezüglich ihres Ressourcenverbrauchs untersucht werden. Die Laufzeit und der Speicherbedarf sowie die Skalierbarkeit für große Datenmengen gehören im maschinellen Lernen und Data Mining zu den Kriterien, die für neue Methoden stets angegeben werden. Nun sollen aber zusätzlich physikalische Größen mit einbezogen werden: Energiebedarf und Kommunikationsbedarf. Auch für die Zeit- und Platzangaben sollen tatsächliche Zeiten mit Maßeinheit und physikalische Speichermodelle betrachtet werden.

Bei der Behandlung von Ressourcenbeschränkungen bauen wir auf dem Stand der Technik im Bereich der eingebetteten Systeme auf. Kurz vor der ersten Antragsphase wurde in den USA der Begriff der *Cyber-Physical Systems* (CPS) von der *National Science Foundation* (NSF) in den USA eingeführt. National hat acatech, die Deutsche Akademie der Technikwissenschaften, 2011 ein Positionspapier zu CPS herausgegeben[5] und hierbei die Unterstützung des Branchenverbandes BITKOM e.V. erfahren. CPS stellen eine Verallgemeinerung der eingebetteten Systeme dar, womit die entsprechenden Techniken ein zentraler Bestandteil sind. Der Begriff Cyber-Physical Systems hat weltweit zu zahlreichen Initiativen zur Entwicklung von Entwurfs- und Nutzungstechniken für CPS geführt. Forschungsfragen der eingebetteten Systeme bestehen damit aber weiterhin. Inhaltlich umfassen CPS eine Vielfalt von Systemen. Ein Schwergewicht liegt dabei vielfach im Bereich der Integration von Kommunikationstechnik mit eingebetteten Systemen. Vom Namen und von der Idee her sind CPS aber weiter gefasst. Sie beinhalten die Integration von jeder Form von Sensoren und Aktuatoren mit Informationstechnik, Medizintechnik, Verkehrswesen, physikalischen Experimenten sowie Produktion und Logistik bis hin zur Fahrzeugtechnik. Naturgemäß reicht es dabei nicht aus, die anfallenden Daten nur zu erfassen. Vielmehr muss die große Datenmenge zu sinnvoller Information führen. Hierbei beobachten wir eine zunehmende Zahl von Anwendungen der Datenanalyse im Bereich CPS. Allerdings sind dies typischerweise direkte Übertragungen von Standardtechniken der Datenanalyse durch Forscher aus dem CPS-Bereich, ohne dass Experten aus dem Bereich der Datenanalyse und der CPS angepasste Lösungen erzeugen, die beispielsweise ressourceneffizientes Lernen ermöglichen. Die jüngsten Entwicklungen haben gezeigt, dass der SFB 876 frühzeitig auf dem richtigen Weg und seiner Zeit einige Jahre voraus war. Mit der Integration der Kommunikationstechnik hat der SFB 876 die Entwicklung im CPS-Bereich vorweggenommen. Auch heute gestattet die enge Integration der beiden Bereiche Wechselwirkungen, die durch ein reines Anwenden der Techniken eines Bereichs auf den anderen nicht möglich wären. Der SFB 876 geht mit der Integration von medizinischen Themen, Logistik, Verkehrswesen und Produktion auch über eine reine Integration eingebetteter Systeme mit der Kommunikation hinaus. Er deckt die Breite von CPS umfangreich ab und integriert damit zwei der aktuellsten Themen, nämlich Cyber-Physical Systems und Big Data.

Sichtbar ist der SFB 876 international vor allem durch die Publikationen und die Diskussionen auf Tagungen. Alle Projektleiter sind bei den wichtigen Tagungen, Zeitschriften und Arbeitskreisen vertreten und auch die Promovenden sind in ihren jeweiligen Disziplinen gut eingeführt. Interessant ist, dass inzwischen auch die jeweils interdisziplinären Aspekte in die eigenen Disziplinen eingebracht werden. Als Maßnahme zur internationalen Präsenz hatte der SFB 876 sich vorgenommen, Open Source Software bereitzustellen. Der SFB876 versteht sich als Teil der internationalen Gemeinschaft der Forschenden und will daher seine Ergebnisse dort einbringen. Der Nutzen für die Community zeigt sich auch darin, dass unsere Entwicklungen von anderen aufgegriffen und verwendet werden. Auch entstehen so Kontakte und ein detaillierter Gedankenaustausch über Inhalte weit vor der Publikation, ganz im Sinne der wissenschaftlichen Werkstatt. Das Werkzeug, das verschiedene Entwurfsparadigmen und eine Fülle von Kriterien zu Relevanz, Redundanz und Stabilität der Merkmalsselektion zur Verfügung stellt, ist als Plug-in für RapidMiner entwickelt worden und wurde bereits mehr als 2700 mal herunter geladen. Ein anderes RapidMiner Plug-in bietet eine Werkbank für Optimierungsverfahren. Das *streams* Framework des Teilprojektes A1 wurde auf den internationalen RapidMiner-Tagungen (jetzt: RapidMiner World) 2012 und 2013 vorgestellt, der eingeladene Hauptvortrag über *Resource-Aware Data Mining* von Frau Morik stellte den Bezug zu eingebetteten Systemen und Industrie 4.0 her. Mit den Entwicklern von RapidMiner wurde intensiv über Erweiterungen von RapidMiner, neue

[4] http://nem-initiative.org/wp-content/uploads/2013/11/NEM-PP-016.pdf

[5] http://www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Material_fuer_Sonderseiten/Cyber-Physical-Systems/acatech_POSITION_CPS_web.pdf

Programmiermodelle und mögliche Scheduling-Algorithmen für die Ausführung von RapidMiner auf anderen Plattformen diskutiert. RapidMiner ist auch 2014 wieder das mit Abstand international beliebteste Werkzeug zur Datenanalyse[6]. Weitere Entwicklungen aus A1 und C3 wurden ebenfalls als Plug-ins veröffentlicht.

Das Werkzeug TraceR, entstanden in Kooperation mit dem Teilprojekt A3, das den Ressourcenverbrauch von Funktionen der Sprache R misst, wurde beim Bioconductor Developers Meeting 2013 und in der Bioinformatik-Arbeitsgruppe der Universität Ulm vorgestellt. R Pakete zu BatchJobs, BatchExperiments und zur Signaturselektion sind ebenfalls veröffentlicht.

Eine Bibliothek zur Ressourcenvorhersage bei Mobiltelefonen und drahtlosen Kommunikationsnetzen, MobiSIM, stellt A4 zur Verfügung. Einen Satellitensimulator hat das Teilprojekt B4 entwickelt. Er ist sogar Teil des OpenSource Paketes *INET Framework* für Kommunikationsnetzwerke geworden.

1.2.4. Inhaltliche Einbindung des Sonderforschungsbereichs am Standort

Der SFB 876 forscht in dem Profilschwerpunkt *Modellbildung, Simulation und Optimierung komplexer Prozesse und Systeme* der TU Dortmund, insbesondere im Sinne der datengestützten Wissenschaft, die von der TU Dortmund schon frühzeitig als eines ihrer langfristigen interdisziplinären Forschungsziele identifiziert worden ist. Es geht um den Paradigmenwechsel von der hypothesengesteuerten zur datengesteuerten Wissensentdeckung[7]. Gerade die Analyse sehr großer Datenmengen hat sich die Universität auf die Fahnen geschrieben. Der SFB 876 trägt zu dieser Strukturentscheidung bei.

Der SFB 876 forscht auf dem Gebiet der Informatik und bringt Fachrichtungen zusammen, die sich aus den traditionellen Bereichen – theoretische, praktische und technische Informatik – entwickelt haben und bislang selten verknüpft wurden. Der SFB verstärkt so die Zusammenarbeit innerhalb der Informatik. Die TU Dortmund unterstützt den SFB 876 in vielfältiger Weise. Die besonders großzügigen Stellenzuweisungen finden sich im Abschnitt *Personelle Situation*, die besonderen zusätzlichen Investitionen in im Abschnitt *Forschungsinfrastruktur*.

Im SFB 876 arbeiten methodisch Informatik, Elektrotechnik und Statistik zusammen. Die exemplarisch behandelten Erkenntnisinteressen, denen die Datenwissenschaften dienen sollen, stammen aus den Bereichen der Biomedizin, des Maschinenbaus, der Logistik und der Physik. In der zweiten Phase des SFB 876 möchten 11 Projektleiter aus der Informatik zusammen mit 12 Projektleitern aus anderen Gebieten (5 x Physik, 2 x Statistik, 1 x Maschinenbau, 1 x Logistik, 1 x Biomedizin, 1 x Elektrotechnik) gemeinsam die Datenanalyse unter Ressourcenbeschränkung erforschen. Die Teilprojekte sind dabei so angelegt, dass sich immer mindestens zwei Kollegen mit unterschiedlichem Hintergrund bei der Lösung der Teilfrage ergänzen. Durch diese Konstruktion trägt der SFB 876 einerseits substanziell zu der Vernetzung der beteiligten Fakultäten bei, andererseits stärkt er die vernetzend integrative und Synergie stiftende Rolle der Fakultät für Informatik an der TU Dortmund.

Die enge Zusammenarbeit der Ruhr-Universitäten in der Universitätsallianz Ruhr ermöglicht die Zusammenarbeit mit der Verkehrsphysik und dem Universitätsklinikum der Universität Duisburg-Essen. Die Verbindung wurde auch wegen des SFB 876 noch enger, indem seit 2014 Sven Rahmann die erste Professur der Mercator-Stiftung an der UA Ruhr sowohl an der Informatik der TU Dortmund als auch am Universitätsklinikum Essen innehat. Darüber hinaus ist Sven Rahmann an dem Forschungsbau der Ruhr-Universität Bochum für molekulare Protein-Diagnostik (ProDi) beteiligt, der im Juni 2014 bewilligt wurde. Hiermit ist auch der Kontakt zum SFB 642 zur Proteinforschung verstärkt.

Der jetzt zum SFB 876 hinzukommende Projektleiter Michael ten Hompel bringt auch das Fraunhofer-Institut für Materialfluss und Logistik (IML) und den LogistikCampus der TU Dortmund, ein Vorreiterprojekt auf seinem Gebiet, ein. Auch das An-Institut, das Leibniz-Institut für Analytische Wissenschaften, ISAS e.V., und die Firma B&S Analytik GmbH vom BioMedizinZentrum Dortmund fügen sich passgenau in den Sonderforschungsbereich ein. Der SFB 876 hat mit der Professur der UA Ruhr die Universitätsallianz verstärkt und sich in den lebendigen Campus der TU Dortmund eingebracht, der auch den LogistikCampus, das ISAS und das BioMedizinZentrum umfasst.

Schon bei der Ausarbeitung des Einrichtungsantrags bestanden enge Kontakte, die jetzt in die Beteiligung am SFB 876 mündeten. Petra Mutzel aus der Fakultät für Informatik wurde bereits angesprochen. Bernhard Spaan von der Teilchenphysik konnte nun gemeinsam mit Jens Teubner ein neues Teilprojekt ausarbeiten. Damit kann das Speichern sehr großer Datenmengen in den SFB 876 einbezogen werden. Das Thema der

[6] Das weltweit führende Portal für Data Mining, KDnuggets.com, hat in seiner internationalen Umfrage zu den verwendeten Werkzeugen Juni 2014 erneut RapidMiner als das am meisten verwendete ermittelt. Von über 3000 Abstimmenden – USA (39%), Europa (36%), Asien und Australien (15%), sonstige (10%) – stimmten 44,2% für RapidMiner. R liegt mit 38,5% an zweiter Stelle.

[7] So auch Farnam Jahanian bei der NSF der USA, CISE Directorate, 10.3.2014: *from hypothesis driven to data driven discovery.*

Datenspeicherung wurde zunächst von Herrn Vahrenhold vertreten, der den SFB aber bald verließ, als er einen Ruf an die Universität Münster annahm. Am Beispiel der CERN-Daten (*Large Hadron Collider*), die Herr Spaan einbringt, können die neuen Datenhaltungskonzepte von Herrn Teubner erprobt und weiter entwickelt werden.

Am ISAS e.V. beteiligte sich der Abteilungsleiter von Herrn Zybin (B2), *Roland Hergenröder*, bereits seit einiger Zeit an den Forschungsarbeiten. Da Herrn Zybin nun altersbedingt nur noch wenige Stunden pro Monat zur Verfügung steht, wird Herr Hergenröder ihn als Teilprojektleiter ersetzen. Kontakte mit Michael ten Hompel bestanden ebenfalls schon längere Zeit. Beispielsweise hat Heinrich Müller gemeinsam mit Herrn ten Hompel Masterarbeiten betreut. Einwerbung und Aufbau des LogistikCampus brauchte aber zu viel Zeit, um in der ersten Phase des SFB eine Beteiligung zu erlauben. Jetzt können wir den LogistikCampus für Experimente mit ressourcenbeschränkten Geräten nutzen, deren Kommunikationsverhalten maßgeblich Energieverbrauch und mögliche Energieaufnahme beeinflusst. Die Gespräche mit Kollegen am Standort haben sich für den SFB 876 gelohnt.

Mit den fachlich verwandten Sonderforschungsbereichen und Graduiertenkollegs an der TU Dortmund bestehen personell und fachlich enge Verknüpfungen. Zu dem SFB 823 über „Statistik nichtlinearer dynamischer Prozesse“ bestehen Gemeinsamkeiten bei der Erforschung Bayesscher Ansätze zur Datenanalyse und bei der Lösung von inversen Problemen. Eine personelle Überlappung spiegelt sich in der Mitgliedschaft der SFB-876-Mitglieder Katja Ickstadt und Wolfgang Rhode auch im SFB 823 wider. Im SFB 823 steht die Lösbarkeit von nichtlinearen Probleme aus dem Bereich der Wirtschaft und Ingenieurwissenschaft im Vordergrund, sodass fachlich keine relevanten Überschneidungen auftreten.

Die in den Teilprojekten forschenden Masterstudierenden und Promovenden erhalten, wie bei der Beschreibung des integrierten Graduiertenkollegs ausgeführt, eine fachlich übergreifende Ausbildung. Bei der Promovendenausbildung kooperiert der SFB 876 mit dem DFG-Graduiertenkolleg „Diskrete Optimierung technischer Systeme unter Unsicherheit (GRK 1855)“, an dem das SFB 876-Mitglied Petra Mutzel als betreuende Professorin aktiv ist. Der SFB 876 beschäftigt sich mit der Datenanalyse unter beschränkten Ressourcen wie Energie, Zeit oder Speicherkapazität. Themen aus dem Bereich der eingebetteten Systeme werden im GRK 1855 nicht behandelt, die Datenanalyse kaum. Bezüge zu den Forschungszielen des SFB 876 gibt es in einigen Bereichen, da Modellierungsansätze und Optimierungsmethoden auch im SFB 876 eine Rolle spielen.

Die Promotionsausbildung an der TU Dortmund hat sich mit Verabschiedung einer für alle Fakultäten gültigen Promotionsrahmenordnung mit klarem Bekenntnis zu einem strukturierten Promotionsstudium bekannt. Die vor diesen Aktivitäten der Universität begonnene Promotionsausbildung im SFB 876 hat in diesem Sinn an der TU eine Vorreiter- und Vorbildrolle übernommen, bei der einige aus der interdisziplinären Natur der bearbeiteten Fragestellungen herrührenden zusätzlichen Probleme gelöst wurden. Da die wissenschaftliche Vorbildung der Promovenden zu heterogen ist, um einen Besuch von identischen Vorlesungen von allen Promovenden zu verlangen, wurden von den beteiligten Lehrenden auf den SFB 876 ausgerichtete Vorlesungen definiert, die sowohl für Promovenden verschiedener Fakultäten als auch für fortgeschrittene Masterstudierende geeignet sind[8]. Ebenfalls wurden zur Hinführung zu unseren Themen wie auch für das Masterstudium spezifische Projektgruppen und Oberseminare entwickelt und angeboten[9]. Der interdisziplinäre Charakter des SFB 876 hat sich so mit großer Akzeptanz bei den Studierenden auch in der Lehre verankert. In der Informatik gibt es Bemühungen für einen Master-Studiengang *Cyber-Physical Systems*. Der SFB setzt sich für eine Überarbeitung des Bachelor-Studiengangs *Angewandte Informatik* im Hinblick auf Datenanalyse und Cyber-Physical Systems ein.

1.2.5. Nationale und internationale Kooperationen und Vernetzung

Die Teilnahme an größeren Forschungsverbünden ergänzt thematisch Arbeiten im SFB und bringt dessen Themen auch in die entsprechenden Kollegenkreise ein. So ist insbesondere das Projekt von Kristian Kersting im SPP 1527 der DFG „Autonomes Lernen“ und das von Petra Mutzel im SPP 1736 „Algorithms for Big Data“ zu nennen. Der SFB hat die Akquisition weiterer Forschungsprojekte erleichtert. Umgekehrt wird auch die Sichtbarkeit des SFB durch zusätzliche Projekte der Teilprojektleiter, national und international, wie sie in den Forschungsprofilen angegeben sind, erhöht. Teilweise profitieren internationale Projekte und der SFB auch direkt. So bringt Katharina Morik über das EU-Projekt INSIGHT (Koordinator: Dimitrios Gunopoulos, Universität Athen) Verkehrsdaten der Stadt Dublin in den SFB 876 ein. Umgekehrt wurde in

[8] Details siehe Beschreibung des Graduiertenkollegs.

[9] z.B.: PG SS 2012 CNI Open Source Satellite Simulation (B4 Wietfeld), Fachprojekt SS 2012 Datenstromalgorithmen zur Log-File Analyse (A1, Morik, Bockermann) PG SS 2012, WS 2012/13 Kooperatives Data Mining mit vernetzten Robotern (A1 Morik, B4 Wietfeld) Vorlesung SS 2013 Wissensentdeckung in Datenbanken (A1 Morik, gemeinsam mit der Statistik) Seminar WS 2013/2014 Fortgeschrittene Methoden der Datenanalyse (C3,C5 Rhode, Spaan) Fachprojekt SS 2013 Data Mining mit Graphikprozessoren (A1 Morik, Piatkowski) Spezialvorlesung SS 2014 Diskrete Optimierung (C1 Lee)

Zusammenarbeit mit den Projektpartnern vom Technion, Haifa, das in A1 entwickelte *Streams Framework* mit deren Ereigniserkennung verbunden und auf seine Skalierbarkeit hin getestet. Es konnten durchschnittlich 281 000 Ereignisse pro Sekunde erkannt werden, was einem Vielfachen der realen Rate entspricht. Die Anwendung TechniBall, bei der aus Sensordaten von Fußballern in Realzeit Spielereignisse erkannt werden mussten, gewann den Publikumspreis bei dem Wettbewerb der internationalen DEBS-Tagung 2013. Die Zusammenarbeit mit dem INSIGHT-Projekt betrifft auch das raumzeitliche Modell von A1, das für den Verkehr in Dublin eingesetzt wird. Während der Sonderforschungsbereich die Grundlagen erforscht, können seine Ergebnisse in dem EU-Projekt in praktischen Szenarien erprobt werden. Eine Zusammenarbeit von Olaf Spinczyk mit Henrik Blunck von der Universität Aarhus hat zur Integration des energieeffizienten GPS-Positionierungssystems EnTracked in die Android-Systemsoftware geführt.

In den Forschungsprofilen ist das Engagement der Projektleiterinnen und -leiter aufgeführt, hier verweisen wir nur auf Ämter in internationalen Foren, die auch dem SFB 876 nützen, seine Sicht bekannt zu machen und interdisziplinäre Bezüge herzustellen.

- Jochen Deuse ist Full Member of European Academy on Industrial Management (AIM).
- Katja Ickstadt war 2013 Vorsitzende des Savage Trust Committee, International Society for Bayesian Analysis.
- Peter Marwedel ist IEEE Fellow, Mitglied des europäischen ArtistDesign Netzwerkes und war im Organisationsteam der Cyber-Physical Systems Week 2014, die im April in Berlin stattfand.
- Katharina Morik ist Mitglied im europäischen ECML PKDD Council [10], Kristian Kersting war 2012 einer der Leiter des Programmkomitees und 2013 Leiter dieser Tagung.
- Heinrich Müller ist Chairman of the Eurographics Workshops and Working Group Board.
- Sven Rahmann ist im Management Committee der EU COST Action BM1006: Next Generation Data Analysis Network.
- Wolfgang Rhode ist Mitglied der großen internationalen Kollaborationen MAGIC und IceCube. IceCube lieferte 2013 laut *Physics World* den *Breakthrough of the Year*; die Bestätigung von extraterrestrischen Neutrinos wurde vom *Scientific American* zu einer der Top-10 wissenschaftlichen Nachrichten gekürt. Eine methodische Diskussion zur Datenanalyse wurde von Herrn Rhode in der Kollaboration angestoßen und auch durch eine Vorlesung von Frau Morik auf der *1. Mercur Winter School on Plasma Astroparticle Physics* weiter vorangetrieben.
- Bernhard Spaan ist *Chair of the LHCb Collaboration Board.*
- Michael ten Hompel ist im Steering Committee der European Technology Platform on Logistics (ALICE).

Der SFB 876 war auch Anstoß für Tagungen und Industriekooperationen. Zum Beispiel bietet der Industriearbeitskreis *Wissen aus Daten generieren*, den Jochen Deuse mit anderen gegründet hat, Möglichkeiten zum Dialog zwischen Industrie und Wissenschaft bezüglich Datenanalyse und digitaler Fabrik. Auch veranstaltete Jochen Deuse am 5. Juni 2014 die 5. Dortmunder Industrial Engineering Tagung unter dem Motto *Industrial Engineering – Erfolgsfaktor für Industrie 4.0.* Eine Zusammenarbeit, die Grundlagen aus dem SFB in eine Anwendung überführt, ist auch Katharina Moriks Kooperation mit der SMS SIEMAG AG, die Datenstromalgorithmen zur realzeitlichen Prognose im Stahlwerk einsetzt. Weitere Zusammenarbeiten in Projekten auch mit Partnern aus der Industrie sind in den Forschungsprofilen zu finden.

1.3. Strukturelle Einbindung und Auswirkungen des Sonderforschungsbereichs am Standort

Die Technische Universität Dortmund hat sich insgesamt zu dem Ziel der *Data Volume Driven Science* bekannt. Der SFB 876 unterstützt dieses Ziel, indem er die Datengewinnung durch Cyber-Physical Systems, die Datenspeicherung und die Datenanalyse unter Ressourcenbeschränkung behandelt. Wir haben oben dargestellt, wie der SFB im Standort eingebunden ist. Hier gehen wir nun auf die besonders guten Bedingungen

[10] Die Bedeutung der Tagung ECML PKDD kann damit illustriert werden, dass der Tagungsband der ECML PKDD 2008 (Herausgeber Walter Daelemans, Bart Goethals, Katharina Morik) mit knapp 53 000 Downloads zu den top 25% der meist heruntergeladenen *eBooks* des Springer Verlags gehört.

an der TU Dortmund für den SFB 876 ein. Wie gleich erklärt wird, ist die Anzahl der Wissenschaftler aus der Grundausstattung um ein Fünftel auf 60 Personen gestiegen. Hinzu kommt, dass einige Teilprojektleiter ihre Projektmitarbeiter aus anderen Quellen finanzieren, sodass sie keine Ergänzungsausstattung beantragen. Diese extern finanzierten Mitarbeiter sind in die Zahl nicht einbezogen. Die Steigerung der Grundausstattung bezogen auf Geräte fällt sogar noch deutlicher aus, wie im Abschnitt *Forschungsinfrastruktur* dargestellt wird.

1.3.1. Personelle Situation

Die TU Dortmund vernetzt die Fakultäten Informatik, Statistik, Elektro- und Informationstechnik, Physik und Maschinenbau inklusive Logistik durch drei Sonderforschungsbereiche (siehe *Inhaltliche Einbindung des Sonderforschungsbereichs am Standort*). Der Schwerpunkt der datengestützten Forschung wird von der Universität nachhaltig gestützt. Durch erfolgreiche, zügige Nachbesetzungen, zusätzliche Ausschreibungen und das innovative Konzept der Professur der UA Ruhr hat die Universität den SFB 876 gezielt verstärkt, dem es gelang, herausragende Kollegen für die Stellen zu gewinnen. Wir können deshalb für die geplanten Arbeiten aus dem Vollen schöpfen. In seinem Kerngebiet, der Informatik, hat es, wohl auch von dem SFB 876 beflügelt, personelle Veränderungen gegeben.

Wie oben schon angeführt, erhielt *Jan Vahrenhold* einen Ruf auf eine W3-Professur in Münster, den er annahm. In Absprache mit der DFG wurde das Teilprojekt A2 dann von Christian Sohler als einzigem Projektleiter weiter geführt.

Sven Rahmann erhielt einen Ruf an die Universität Duisburg-Essen und nahm ihn an. An der Universität Duisburg-Essen gibt es bereits zwei Teilprojektleiter, sodass kein neuer Standort einbezogen wird. Der SFB 876 hat sich erfolgreich dafür eingesetzt, dass Herr Rahmann die Lehrbefugnis für die Fakultät für Informatik der TU Dortmund behielt und von der Stiftung Mercator eine Ausstattung für die erste Professur der UA Ruhr erhielt. Für den SFB ergibt sich daraus zudem der Vorteil, dass wir die neuen Sequenziermethoden genetischer Daten in unser Projekt C1 einbeziehen können, für die die Geräte von der Universität Duisburg-Essen zur Verfügung gestellt werden. Deshalb ist Herr Rahmann nun Teilprojektleiter von C1, mit dem er ohnehin von Anfang an zusammen gearbeitet und bereits gemeinsame Publikationen hat. Langfristig wird der bewilligte Forschungsbau zur molekularen Proteindiagnostik in Bochum, an dem auch Herr Rahmann beteiligt ist, die enge Zusammenarbeit der Ruhruniversitäten intensivieren. Der Aufbau eines biomedizinischen Forschungsgebiets für die UA Ruhr verstärkt die Forschung des SFB.

Die Nachfolge Rahmann an der TU Dortmund wurde für *Data Mining* ausgeschrieben, um diesen für den SFB 876 entscheidenden Bereich zu verstärken. Bei sehr guter Bewerberlage konnten wir *Kristian Kersting* gewinnen, der am 1.9.2013 seinen Dienst am Lehrstuhl 8 der TU Dortmund antrat. Sein Forschungsschwerpunkt, statistische relationale Modelle, und seine internationalen Kooperationen passen hervorragend zum SFB und zum Profil der TU Dortmund insgesamt. Seine Arbeiten zum *Graph Mining*, Christian Sohlers Arbeiten zur Analyse sehr großer Graphen sowie die auf Daten in Form von Graphen arbeitenden Algorithmen von Petra Mutzel passen hervorragend zusammen. So entstand das neue Teilprojekt A6.

Wie im Einrichtungsantrag angekündigt, wurden die Nachfolge Biskup und die Nachfolge Marwedel schon während ihrer Amtszeit ausgeschrieben. Bei der Besetzung wurden die Themen des SFB 876 berücksichtigt. Bei hervorragender Bewerberlage konnte zügig eine Liste erfolgreicher Forscher des Gebiets der Informationssysteme mit *Jens Teubner* als Erstplatziertem erstellt werden. Herr Teubner nahm den Ruf an und übernahm am 1.4.2013 den Lehrstuhl 6 für Datenbanken. Er vertritt Datenbankkonzepte, die auf sehr große Datenmengen ausgelegt sind. Dabei untersucht er Bezüge von Ausführungsplattformen einerseits und Speicher- und Indexierungsalgorithmen andererseits. Die Sichtweise des SFB 876, für die Informationsgewinnung die Datenanalyse auf allen Ebenen im Zusammenhang zu betrachten, von der Aufgabenstellung bis hin zur Ausführung, und damit die Schranken zwischen Informationssystemen und technischen Systemen einzureißen, findet sich in den Arbeiten von Herrn Teubner deutlich wieder. Wir haben dadurch den Datenbankenbereich in die Richtung von *Big Data*-Konzepten im Sinne des SFB 876 modernisiert.

Die Besetzung der Nachfolge Marwedel ist international erfolgt. Bei wiederum hervorragender Bewerberlage konnte der Erstplatzierte, *Jian Jia Chen*, gewonnen werden. Er arbeitet an Methoden zur dynamischen Energieverwaltung sowie an Realzeitmethoden, insbesondere der Vorhersagbarkeit von Zeitverhalten. Obwohl er erst am 01.04.2014 sein Amt am Lehrstuhl 12 der TU Dortmund antrat, konnte er schon in ein bestehendes Teilprojekt integriert werden und ersetzt in B2 Herrn Marwedel. Langfristig verstärkt seine fachliche Ausrichtung den SFB 876 in den Gebieten der Energieeffizienz, Ressourcen- und Zeitbeschränkung sowie Zuverlässigkeit von Rechensystemen.

Peter Marwedel bleibt dem SFB 876 für die nächste Förderperiode noch im Teilprojekt A3 erhalten. Hier kann Herr Chen sich noch nicht direkt einbringen. Vielmehr wird Herr Marwedel die interdisziplinäre

Zusammenarbeit, wie die technische Informatik der Datenanalyse helfen kann, in der zweiten Förderperiode dahin bringen, dass dies in der dritten Förderperiode möglich ist.

Die Stellung als stellvertretender Sprecher wird Herr Marwedel in der zweiten Förderperiode abgeben. Dies ist vorbereitet und wird in der ersten Mitgliederversammlung besprochen werden, wenn und falls die nächste Phase bewilligt ist. Während der Antragstellung für die zweite Förderperiode und im Sinne eines sanften Übergangs garantiert Herr Marwedel die Kontinuität der erprobten Zusammenarbeit der an Cyber-Physical Systems arbeitenden Kollegen mit dem Vorstand.

Die Universität hat zur besonderen Stärkung der Informatik im Bereich des SFB 876 eine befristete W2-Professur ausgeschrieben. Die Berufung soll eines der folgenden Themen verstärken: Cyber-Physical Systems, Big Data Analytics oder Privacy/Security, am besten im Bezug auf einander. Der Universität ist besonders daran gelegen, hoch qualifizierte Professorinnen oder Professoren zu gewinnen. Da die Bewerberlage vorab schwer einzuschätzen war, wurde die Stelle mit einem *oder* ausgeschrieben. Tatsächlich gab es wieder eine sehr gute Bewerberlage und die Fakultät hat gerade Ende Juni der Liste der Berufungskommission zugestimmt.

Wie bereits oben aufgeführt, hat der SFB 876 am Standort weitere Kollegen als Projektleiter gewonnen. In der Informatik ist dies Petra Mutzel, deren Expertise das Thema der Algorithmen für Graphen im SFB, insbesondere im neuen Teilprojekt A6, verstärkt. Mit Roland Hergenröder ist nun der Arbeitsgruppenleiter im Bereich Materialanalytik des ISAS e.V. Teilprojektleiter von B2.

Die Förderung des Nachwuchs hat dazu geführt, dass jetzt zwei Nachwuchswissenschaftler Teilprojektleiter sind. Beide stammen aus dem SFB 876. *Tim Ruhe* hat im Teilprojekt C3 promoviert. *Sangkyun Lee* hat im Projekt C1 als PostDoc gearbeitet. Katharina Morik beendet ihre Projektleiterschaft in C1 und Herr Lee führt die Datenanalyse dort fort. Insgesamt hat sich der SFB 876 mit diesen personellen Veränderungen internationalisiert und von einem Durchschnittsalter von 54 Jahren auf eines von 49 Jahren verjüngt.

1.3.2. Forschungsinfrastruktur

Eine besondere Freude ist für den SFB das neue Gebäude für die Informatik, das im Juli 2014 bezogen wurde. Es ist mit dem alten verbunden, sodass erstmals alle Informatiker in dem selben Gebäude untergebracht sind. Dies erleichtert den alltäglichen Gedankenaustausch.

Auch bezüglich der Rechnerinfrastruktur hat die TU Dortmund erheblich in den SFB 876 investiert. Die Grundausstattung des SFB 876 ist dadurch stark erhöht und die Zukunftssicherheit der Arbeitsmöglichkeiten im Schnittbereich von Datenanalyse und technischer Informatik erreicht worden.

Parallel zu der Erforschung der Reduktion der benötigten Ressourcen ist an der TU Dortmund das Zentrum für wissenschaftliches Rechnen (DoWiR) gegründet worden, dessen Hauptaufgabe in der fachlichen Unterstützung bei der Konzeption, der Analyse und der Realisierung von Simulationsmethoden und Werkzeugen des wissenschaftlichen Rechnens für komplexe Prozesse und Anwendungsprobleme auf Hochleistungsrechnern besteht. Das Zentrum soll der zentrale Ansprechpartner innerhalb der TU Dortmund für die Entwicklung, Bereitstellung und Anwendung von flexiblen, effizienten und anwenderfreundlichen Verfahren und Softwarelösungen in Modellbildung und Simulation sein. Der SFB 876 ist personell (in dem fünfköpfigen DoWiR-Vorstand sind Katharina Morik und Wolfgang Rhode vertreten) und über die wissenschaftlichen Ziele eng mit DoWiR und über DoWiR mit allen wissenschaftlich rechnenden Gruppen der TU Dortmund verknüpft. Zur Zeit wird vom ITMC der TU Dortmund unter der Leitung von DoWiR mit LiDO3 ein modernisierter Ersatz für LiDOng und das GRID-Ressourcenzentrum organisiert.

Im Folgenden werden zunächst die bislang genutzten Farmen und dann die Nachfolge-Farm, deren Beschaffung auch für den SFB 876 zur Zeit vorangetrieben wird, vorgestellt:

- Der High-Performance-Rechencluster PhiDO wurde Anfang 2010 als zentraler Rechencluster von Arbeitsgruppen aus der Physik in Betrieb genommen. Die Forschungsbereiche, für die PhiDO genutzt wird, sind angesichts der verschiedenen Arbeitsgruppen breit gestreut, die überwiegende Nutzung von CPU und Speicherressourcen liegt bei den im SFB 876 vertretenen Gruppen in der Teilchenphysik (C5/LHCb) und Astroteilchenphysik (C3/FACT, MAGIC, CTA und IceCube). Nutzungsberechtigt sind neben den Gruppen aus der Physik auch alle kooperierenden Arbeitsgruppen aus dem SFB 876. Für die Dortmunder Analysen für das LHCb-Experiment zu Untersuchungen zur Materie-Antimaterie-Asymmetrie ist PhiDO unverzichtbar, wobei neben der Datenanalyse auch zahlreiche Simulationsrechnungen durchgeführt werden. Im Rahmen der Beteiligung an den Astroteilchenexperimenten nimmt PhiDO eine besondere Rolle ein, da hier ein großer Teil der von den Forschungsverbünden benötigten Monte-Carlo-Simulationen erzeugt wird. PhiDO besteht aus zwei Management-Servern, zwei Gateways, 11 Fileservern und 1184 CPU-Kernen. Der gesamte Hauptspeicher der Rechenknoten beträgt 4,6 TByte. Systemweit stehen ca. 230 TByte Festplattenplatz über das parallele Filesystem FhGFS zur

Verfügung. Der vergleichsweise große Festplattenplatz geht auf die Anforderungen der experimentellen Gruppen zurück, da die Daten für die lokalen Analysen sowie die in Dortmund erzeugten simulierten Ereignisse gespeichert werden müssen.

- Die Technische Universität Dortmund ist im deutschen DGRID-Verbund aktiv und betreibt eigene Grid-Ressourcen. Im Rahmen der zweiten bundesweiten Ausbauphase konnte die Technische Universität Dortmund 2008 einen Cluster mit insgesamt 2048 CPU-Kernen und über 100 TByte Plattenspeicherplatz aufstellen. Die Beschaffung diente dem Aufbau des „D-Grid Ressourcen-Zentrums Ruhr“ als Anbieter von Rechenleistung und Speicherkapazität im D-Grid für die Unversitätslandschaft im Ruhrgebiet im Rahmen der UA Ruhr. Nutzungsberechtigt sind alle am SFB 876 beteiligten Arbeitsgruppen.

- Die letzte Ausbaustufe der zentralen Computing Farm der TU Dortmund, der Linux-HPC-Cluster LiDOng wurde ebenfalls 2008 in Betrieb genommen[11]. LiDOng steht als Werkzeug zum wissenschaftlichen Rechnen allen Wissenschaftlern an der Technischen Universität Dortmund, der Fachhochschule Dortmund und (eingeschränkt) im Ressourcenverbund NRW zur Verfügung. Alle Teilprojekte des SFB 876 können LiDOng kostenfrei nutzen, dieser Service wird von den rechenzeitintensiven Teilprojekten mit einem Anteil von 40% in Anspruch genommen. Altersbedingt soll LiDOng 2015 durch den nachfolgenden HPC-Cluster LiDO3 ersetzt werden.

- Gegenwärtig wird vom Dortmunder Zentrum für wissenschaftliches Rechnen (DoWiR) unter besonderer Berücksichtigung und zum großen Teil für den SFB 876 die HPC-Farm LiDO3 geplant. Mit etwa 3500 Kernen sollen von LiDO3 30 Mio. Corestunden pro Jahr zur Verfügung gestellt werden, wobei verschiedene Vernetzungsarten und Computingkonzepte (Zahl der Kerne/CPU, Memory/Kern, GPU-Nutzung, etc.) in einer Maschine vereint werden. Der verfügbare Plattenspeicher soll 1200 TByte betragen. LiDO3 soll die lokalen Aufgaben von LiDOng und dem GRID-Rechner übernehmen.

- Nachteil der zentralen Cluster für wissenschaftliches Rechnen an der TU Dortmund sind die begrenzten Möglichkeiten zur Forschung auf den untersten Schichten. Für die Erforschung verschiedener Ausführungsplattformen und -algorithmen brauchen wir nicht so sehr die Ergebnisse der Berechnungen als vielmehr die Kennzahlen der Ausführung (Laufzeit, Speicherverbrauch, Kommunikations- und Energieaufwand) bei unterschiedlichen Ausführungen (z.B. Scheduling-Algorithmen). Deshalb befindet sich in Ergänzung zu den zentralen Systemen aktuell ein aus Grundaustattungsmitteln finanzierter Forschungscluster speziell für den SFB 876 in der Beschaffung. Dieser wird allen Mitgliedern des SFB für die freie Konfiguration zur Verfügung stehen. Geplant ist das System mit 12 Rechenknoten mit jeweils 20 Kernen und 256 GByte Hauptspeicher. Drei der Knoten werden zusätzlich als GPU-Recheneinheiten mit leistungsfähigen CUDA-Grafikkarten ausgestattet. Die Betreuung des Systems erfolgt durch die Mitarbeiter des Projekts Z, unterstützt räumlich durch das IT & Medien Centrum (ITMC) und administrativ durch die Informatikrechner-Betriebsgruppe (IRB). Ergänzt werden die Rechenknoten durch ein passendes Speichersystem, was sowohl für die kurzfristige Speicherung der Daten für die Verarbeitung als auch für die Langzeitarchivierung zum Einsatz kommen soll. Dieses wird im folgenden Abschnitt 1.3.3 näher beschrieben.

Für die Teilprojekte des SFBs werden jeweils auch spezifische Formen der Infrastruktur benötigt, die von der TU Dortmund bzw. der Universität Duisburg-Essen finanziert werden. In der Regel handelt es sich dabei um Laborrechner und Labore, in denen elektronische bzw. elektrotechnische Komponenten überprüft und erstellt werden können. Details dazu finden sich in den Teilprojektbeschreibungen. Die primär außerhalb der Informatik beheimateten Teilprojektleiter steuern – ggf. ebenfalls unterstützt durch die TU Dortmund – aus den Anwendungen bzw. Experimenten stammende Daten und zu optimierende eingebettete Systeme bei.

1.3.3. Forschungsdaten

In der ersten Phase lagen die Ziele der Arbeiten für den Umgang mit Forschungsdaten des SFB im einfachen Austauschen zwischen den Projekten, der langfristigen Archivierung von Rohdaten und der öffentlichen Bereitstellung ausgewählter Datensätze. Die dazu notwendige Infrastruktur wurde im Rahmen des Projekts Z eingerichtet, was sowohl den Betrieb der Datenbankserver als auch die Erstellung der Software für den Zugriff auf die Daten betrifft.

[11] 400 Rechenknoten mit insgesamt 832 Quadcore Prozessoren vom Typ Intel/Xeon, 7 TByte Hauptspeicher in den Rechenknoten und 250 TByte Festplattenspeicher

Austausch und Archivierung: Alle Mitglieder des SFB haben die Möglichkeit, selbstständig die Datensätze des Projekts auf dem Server abzulegen und Zugriffsrechte zu erteilen. Grundsätzlich stehen die Daten allen Projekten innerhalb des SFB offen. Das zu Beginn der ersten Phase angeschaffte System stellt 15 TByte effektiv nutzbaren Speicherplatz zur Verfügung, der durch einen RAID-Verbund entsprechend gesichert wird. Die langfristige Archivierung und Wartung des Systems wird mit Unterstützung der Informatik-Rechnerbetriebsgruppe (IRB), garantiert, die eine tägliche Sicherung der Daten durchführt und eine Wiederherstellung verschiedener Versionsstände bis maximal 6 Monate zurückliegend ermöglicht. Durch den Einsatz des aktuell ausgeschriebenen Rechen- und Speichersystems soll der verfügbare Speicherplatz Ende 2014 um 360 TByte effektiv (960 TByte brutto) erhöht werden.

Veröffentlichung und Nachnutzung: Die zu jedem Datensatz hinterlegten Metadaten erweitern das bekannte Modell des *Machine Learning Repository* der University of California at Irvine[12] (UCI) mit Beschreibungen des Datensatzes um die formelle Spezifikation von Feldern nach der DataCite Metadata Specification 3.0[13].

Die Erfassung von Metadaten nach diesem Schema ist Voraussetzung für die Vergabe von Document Object Identifiern (DOI) nach ISO-Standard 26324:2012. Der Sonderforschungsbereich 876 ist seit Juni 2014 als offizielles Datencenter zur Vergabe von DOIs bei der Technische Informationsbibliothek Hannover registriert. Damit besteht die Möglichkeit zur eigenverantwortlichen Vergabe zitierfähiger und öffentlich verfügbarer Datensätze, womit der wissenschaftliche Nutzwert über die eigenen Arbeiten des SFB hinaus gesteigert wird. Die Spezifikation unterstützt auch Formate, die über klassische Attribut-Werte-Repräsentation hinausgehen, wie es z.B. bei den PAMONO-Daten des Projekts B2 der Fall ist.

Wie im UCI-Repository besteht die Möglichkeit der Verknüpfung von Datensätzen mit Publikationen des SFB. Dies dient der Reproduzierbarkeit von veröffentlichten Ergebnissen und erlaubt anderen Forschern, ihre Verfahren mit denen des SFB auf derselben Datenbasis zu vergleichen. Zitiert werden können dann entweder die assoziierte Veröffentlichung oder der Datensatz bei Verfügbarkeit einer DOI.

Bei der Veröffentlichung von Datensätzen müssen diese mit einer entsprechenden Nutzungslizenz versehen werden. Die vom SFB genutzten Lizenzen basieren auf den Definitionen des Open Data Commons-Projekts[14]. Die betreffenden Lizenzen definieren verschiedene Stufen der Nutzung. Die großzügigste ist die Open Data Commons Public Domain Dedication and License (PDDL), die sämtliche Formen der Nutzung und Weiterverbreitung erlaubt. Abgestufte Formen davon sind die Open Data Commons Attribution License, die das Verteilen, Verändern und Verarbeiten unter der Bedingung der Referenzierung der Quelle erlaubt, bzw. die Open Data Commons Open Database License (ODbL), die zusätzlich noch die gleichen Bedingungen zur Offenheit und Weitergabe fordert. Die ODbL ist die vom SFB bevorzugte Lizenz, da sie die Verfügbarkeit der veröffentlichten Datensätze auch in verarbeiteter Form garantiert.

Datensätze lassen sich damit im jeweiligen Zustand einfrieren, bei Änderungen und Erweiterungen der Datensammlung kann je nach Umfang eine neue DOI vergeben oder eine neue Version des Datensatzes erzeugt werden. Selbstverständlich ist die Veröffentlichung der Daten nicht auf das Portal des Sonderforschungsbereichs begrenzt, sondern eine zusätzliche Veröffentlichung bei anderen Datenbanken kann die Reichweite noch deutlich erhöhen.

Für Daten, die in einer Attribut-Werte-Repräsentation vorliegen, können Metadaten über das aus RapidMiner bekannte AML Attribute Description File Format mit den Daten selbst verknüpft werden. Dies geht über die Möglichkeiten des DataCite-Schemas hinaus, indem hier attributspezifische Informationen wie Umfang, Wertebereiche und Statistiken zu Maximal-/Minimalwerten usw. angegeben werden. AML-Dateien können direkt aus RapidMiner erzeugt und auf den Datenbankserver des SFB hochgeladen werden. Das auf XML basierende Datenformat wird zusätzlich auf dem Webserver in eine lesbare Online-Darstellung überführt, sodass Interessenten bereits vor dem Download von Daten die wichtigsten Informationen einsehen können. Diese Metadaten können dann ebenfalls direkt als Attributdefinitionen in Werkzeugen des maschinellen Lernens wie RapidMiner genutzt werden. Ist keine AML-Metadatendatei vorhanden, kann der Server diese direkt aus den bereitgestellten Datensätzen erzeugen. Dazu werden die Typen der Attribute aus den Daten geschätzt und die Statistiken generiert. Fehlerhaft interpretierte Datentypen können manuell korrigiert werden.

Datensätze, die nicht auf der Webseite des SFB archiviert und veröffentlicht werden können, z.B. aus rechtlichen Gründen, können aber trotzdem mit dem SFB verknüpft werden, da zu jeder Registrierung eines Datensatzes auch ein Verweis auf andere Verzeichnisse angegeben werden kann.

[12] http://archive.ics.uci.edu/ml/datasets.html
[13] http://schema.datacite.org/
[14] http://opendatacommons.org/

Prinzipiell können diese Mechanismen nicht nur für die Archivierung der Forschungsdaten genutzt werden, sondern das gleiche Modell lässt sich auf die Archivierung von Software ausweiten. Soll damit zu einer Publikation die jeweilige Software im Zuge der Nachhaltigkeit archiviert werden, kann das entsprechende Archiv über die Webseite zugänglich gemacht werden.

Publikationen des SFB werden ebenfalls auf den Servern archiviert. Unterstützt wird der SFB dabei noch durch die Hochschulbibliothek, die mit ihrem Dokumentserver Eldorado[15] die langfristige Archivierung und nötigenfalls Konvertierung in neuere Datenformate garantiert.

Forschungsdaten innerhalb der Projektbereiche: In den einzelnen Projekten kommen verschiedene externe oder SFB-interne Datenquellen zum Einsatz. Ziel ist es, bei selbst erhobenen Daten eine öffentliche Nutzungsmöglichkeit über den Sonderforschungsbereich zu schaffen, oder, falls rechtliche Aspekte dem entgegenstehen, die Archivierung der Daten sicher zu stellen. Innerhalb des SFB stehen die Daten in der Regel allen Projekten zur Verfügung. Darüber hinaus werden für den internationalen Vergleich der in den Projekten entwickelten Methoden oft Datensätze des UCI-Machine-Learning-Repositorys genutzt, die bereits öffentlich verfügbar sind.

Projekt A1 sammelt darüber hinaus durch das im Sonderforschungsbereich entwickelte *MobiDAC* detaillierte Daten auf Ebene das Betriebssystems und kann diese auf Smartphones mit lokalen Sensordaten kombinieren. So wurde bereits für die Sommerschule 2012 eine umfassende Datensammlung von Smartphone-Nutzungsdaten durchgeführt, die den Teilnehmern auch für die Nachnutzung zur Verfügung gestellt wurde. Die Daten sind dabei anonymisiert und von Ortsinformationen befreit.

In Projekt A4 werden Messdaten rund um den Energieverbrauch eingebetteter und mobile Systeme über das eigens entwickelte *MIMOSA* erhoben. Die Messungen berücksichtigen ebenfalls verschiedene Szenarien der Kommunikation über den Mobilfunkstandard LTE durch den Betrieb eines eigenen Basisstationsemulators. Erweitert wird dies in der 2. Phase durch Einbeziehung lokaler Netze, von WLAN bis RFID.

Im Projekt zur Analyse der Atemluft, TB1, stehen Datensätze sowohl aus eigenen Messungen mittels der neu entwickelten Spektrometer als auch aus der von der B&S Analytik aufgebauten Datenbank *DBreath2012* zur Verfügung, was auch die langfristige Archivierung der Primärdaten beinhaltet.

Die Daten des B2-Projektes zum Erkennen von Nanoobjekten in Plasmonen-Mikroskopie sind ebenfalls in mehrere Datensätze überführt worden. Diese beinhalten die Zeitreihen der Objekterkennung aus Bildsequenzen des PAMONO-Sensors.

Die Messdaten des Teilprojektes B3 bestehen aus realen Messungen im Stahlwerksbetrieb und repräsentieren damit unternehmensspezifisches Expertenwissen. Diese unterliegen der Vertraulichkeit, sind aber innerhalb des SFB für die Nutzung und Archivierung verfügbar.

Im Teilprojekt B4 werden digitale Fahrzeugdaten sowohl simulativ als auch durch reale Messungen erhoben. Diese Daten liegen in Form von Messreihen vor und werden in Datenbanken für den Zugriff beider Lehrstühle abgelegt. Darüber hinaus erfolgten in der Laufzeit der ersten Phase Aufzeichnungen des CAN-Bus innerhalb von PKWs mit sämtlichen Fahrparametern, um diese mit den makroskopischen Verkehrsdaten für die Prognose kombinieren zu können. Am Lehrstuhl für Kommunikationsnetze wurde eine GPS-Datenbank aufgebaut[16], die als Webservice eine Schnittstelle für den Abruf der gesammelten Funkfeldmessungen darstellt. Auf Antrag können diese auch außerhalb des SFB zu Forschungszwecken zugänglich gemacht werden.

Im Teilprojekt C1 liegen die erfassten Expressionsdaten für jede Probe als Liste von reellwertigen Zahlen vor. Während der ersten Phase wurden diese Daten mit verfügbaren Daten des Gene Expression Omnibus (GEO)[17] und der Datenbank R2 vom AMC Amsterdam[18] verbunden und zu diesen Datenbanken beigetragen. Für die zweite Phase werden diese Daten noch um Next-Generation-Sequencing (NGS)-Daten erweitert.

Im Teilprojekt C3 wird mit Daten gearbeitet, die von den internationalen Experimenten MAGIC und IceCube erhoben werden. Die Daten werden von den internationalen Kollaborationen in elektronischer Form (Tape und Festplatten) so lange wie möglich, mindestens jedoch zehn Jahre, verteilt gespeichert und katalogisiert. Die von Teilprojekt C3 genutzten Daten bilden nur einen sehr kleinen Bruchteil dieser Menge.

Teilprojekt C4 verwendet die Datensätze anderer Teilprojekte, die dort nachhaltig gespeichert werden. Darüber hinaus werden in Teilprojekt C4 Daten aus einer Harnblasenkrebsstudie des IfADo (Leibniz-Institut für Arbeitsforschung an der TU Dortmund) verwendet.

Letztendlich ist einer der Forschungsschwerpunkte von Teilprojekt C5 die Verbesserung von Datenhaltung, -zugriff und -archivierung. Die am *Large Hadron Collider* anfallenden Datenmengen können aktuell nicht vollständig erfasst und gespeichert werden, was durch das Projekt nachhaltig verbessert werden soll. Aktuell

[15] https://eldorado.tu-dortmund.de/
[16] http://www.kn.e-technik.tu-dortmund.de/GNSSdatabase
[17] http://www.ncbi.nlm.nih.gov/geo/
[18] http://hgserver1.amc.nl/cgi-bin/r2/main.cgi

wird der Großteil der verbleibenden Daten auf Bandlaufwerken archiviert, interessante oder häufig genutzte Daten auf Festplatten zwischengespeichert.

1.3.4. Gleichstellung von Wissenschaftlerinnen und Wissenschaftlern und Vereinbarkeit von wissenschaftlicher Karriere und Familie

Die forschungsorientierten Gleichstellungsstandards der DFG definieren als eine der Grundlagen der Qualitätssicherung in der Forschung eine (1) durchgängige, (2) transparente, (3) wettbewerbsfähige und zukunftsorientierte, (4) und kompetente Sicherung der Gleichstellung. Nach diesen Standards ist die Technische Universität Dortmund in der besten Qualitätsstufe eingeordnet.

Vereinbarkeit von Karriere und Familie: Bestätigt wird die gelebte Familienfreundlichkeit an der Technischen Universität Dortmund bereits seit 2008 durch das Zertifikat *audit familiengerechte hochschule* der gemeinnützigen Hertie-Stiftung. Im August 2012 konnte dieses Audit wieder erfolgreich erneuert werden. Das Audit dokumentiert transparent und öffentlich den Zielerreichungsgrad und setzt neue Ziele für die daran anschließende Phase bis zur nächsten Zertifizierung. Viele der bei der Antragsstellung der ersten Phase geplanten Ziele sind inzwischen von der Universität erfolgreich umgesetzt.

Im Juni 2014 wurde ergänzend die Charta *Familie an der Hochschule* durch die Leitung der TU Dortmund unterzeichnet. Diese Charta thematisiert die Bereiche Führung und Betreuung, Forschung, Arbeits- und Studienbedingungen, Gesundheitsförderung, Infrastruktur und Vernetzung und demonstriert damit öffentlich das Bewusstsein, sowohl Bildungs-, als auch Lebensort zu sein.

Als zentrale Plattform zur Information über alle Maßnahmen für Gleichstellung und Familie ist kurz nach der ersten Begehung für die Einrichtung des SFB 876 das Familienportal der TU Dortmund auf der Startseite online gegangen. Von hier ausgehend sind alle Details der aktuellen Projekte abrufbar.

Das Angebot für Kinderbetreuung an der TU Dortmund wurde stark erweitert. Im März 2011 öffnete die Kindertagesstätte HoKiDo für bis zu 60 Kinder in altersgemischten Gruppen ihre Pforten. Besonders wichtig im Sinne der Flexibilität ist die Kurzzeitkinderbetreuung KuKi. Die Spielgruppe für Kinder von 1 bis 4 Jahren erlaubt regelmäßige Betreuung, aber auch Betreuung bei Engpässen bedingt durch Meetings und Sprechstunden oder bietet Unterstützung bei krankheitsbedingten Ausfällen. Zur Eingewöhnung in Vorbereitung auf Engpässe können die Kinder vorher bereits an der Spielgruppe teilnehmen.

Da Eltern in den Schulferien besonders häufig unter einem Betreuungsengpass leiden, wurden Ferienbetreuungsgruppen für Kinder von 6 bis 13 Jahren eingerichtet. Jeweils während der Oster- und Herbstferien, sowie in den ersten beiden Wochen der Sommerferien, können Kinder von Studierenden und Beschäftigten an dem Programm mit Ausflügen, Besichtigungen und Spielen teilnehmen.

Aufbauend auf den etablierten Maßnahmen hat die TU Dortmund noch weitere für die Zeit seit 2012 umgesetzt, zum Beispiel die MensaCard für kostenlose Kindermahlzeiten. Andere Maßnahmen konzentrieren sich auf den Ausbau von Beratungs- und Informationsmöglichkeiten, Väternetzwerke oder familienfreundliche Meetingzeiten, wie sie im SFB schon seit Beginn gepflegt werden.

Ähnlich umfassende Maßnahmen sind auch an den anderen am SFB beteiligten universitären Einrichtungen umgesetzt. Sowohl an der Universität Duisburg-Essen (UDE) als auch am Universitätsklinikum Essen gibt es ebenfalls Einrichtungen in Form einer Betriebskindertagesstätte und eine Kinderferienbetreuung. Seit 2012 wird an der UDE dieses Angebot durch eine Kurzzeitbetreuung ergänzt, die Lücken in den üblichen Betreuungszeiten insbesondere am Nachmittag oder Abend schließt. Für die Betreuung vor Ort während Weiterbildungsveranstaltungen, Tagungen und Konferenzen an der Universität vermittelt die UDE Betreuungskräfte, die über das Elternservicebüro gebucht werden können.

Zentrale Gleichstellungsmaßnahmen der TU Dortmund und anderer Einrichtungen: Das Mentoring-Programm[19] der UA Ruhr, getragen von TU Dortmund, Universität Duisburg-Essen und Ruhr-Universität Bochum, ist einer der Eckpfeiler für die überfachliche Karriereförderung von Wissenschaftlerinnen. Im *mentoring*[3] starten im jährlichen Wechsel jeweils eine Programmlinie für Doktorandinnen oder Postdoktorandinnen. Während der zweijährigen Laufzeit des Programms profitieren die Wissenschaftlerinnen von einer Mentoring-Beziehung mit einer erfahrungsälteren Person. In der persönlichen Beratung werden Potenziale und Kompetenzen weiter ausgebaut. Den Mentees wird wichtiges Erfahrungswissen über Strukturen und Prozesse im Wissenschafts- oder Industriebetrieb zugänglich gemacht. Ergänzt wird das Mentoring durch Seminare, die auf übergreifende Kompetenzen ausgerichtet sind, wie z.B. Führungsanforderungen, Akquise, Bewerbungs- und Berufungstraining oder Teammanagement.

[19] http://www.scn-ruhr.de/mentoring/index.html

An Studentinnen und Wissenschaftlerinnen dagegen, die eine weitere Karriere in der Wissenschaft ausschließen, richtet sich das Angebot *female.2.enterprises*[20]. Diese Veranstaltungsreihe unterstützt Frauen in der Kontaktaufnahme zu Unternehmen der Region. Im Mittelpunkt stehen dabei Formate, die individuelle Gespräche mit Vertreter/-innen von Unternehmen ermöglichen. Sie werden ergänzt durch ein Angebot an Coachings und karrierelevanten Workshops.

Darüber hinausgehend sollten die ersten Fördermaßnahmen auch schon vor der Entscheidung für ein Studiengang und Fachrichtung starten. Um weiblichen Nachwuchs für die Universität und auch explizit die Informatik zu begeistern, wird jährlich TU-weit der Girl's Day und von der Fakultät für Informatik die Schnupperuni mit einem expliziten Teil für Schülerinnen ausgerichtet, den der SFB 876 mitgestaltet hat.

Die letzte Evaluation der Fakultät für Informatik für den Berichtszeitraum 2010-2012 hat die Wirksamkeit der Maßnahmen bestätigt. Die Zahl der Studentinnen konnte im fachwissenschaftlichen Studium auf 20% gesteigert werden. Im auslaufenden Diplomstudiengang liegt der Frauenanteil unter den Absolventen und Absolventinnen noch bei etwas mehr als 10%. Für die Zukunft wird dieser Wert noch weiter steigen, liegt doch der Anteil an Studentinnen im ersten Semester bei 36% im BA Informatik.

Im Berichtszeitraum lag der Frauenanteil unter den Promotionen bei 13,5%. Das ist zwar geringer als ihr aktueller Anteil unter den Studierenden mit der großen Zahl an Frauen in den ersten Fachsemestern, aber höher als ihr Anteil in den beiden Diplomstudiengängen Kern- und angewandte Informatik.

Ergänzende Maßnahmen des Sonderforschungsbereichs: Die zentralen Maßnahmen der TU Dortmund und UA Ruhr ergänzt der Sonderforschungsbereich durch spezifische Maßnahmen für seine Mitarbeiterinnen und Mitarbeiter. So konnte im mentoring-Programm eine eigene Gruppe für sechs Wissenschaftlerinnen des SFB 876 gebildet werden, für die ein spezielles Qualifizierungsprogramm entwickelt und die durch die Koordinatorin des Programms betreut wurde. Insgesamt beträgt die Laufzeit der SFB-Mentoring-Gruppe zwei Jahre, geplantes Ende ist Januar 2015. In der zweiten Förderperiode soll dieses Programm ebenfalls wieder den Wissenschaftlerinnen im SFB angeboten werden.

Als weitere Maßnahme zur Karriereförderung wurde den Wissenschaftlerinnen des SFB auch die Teilnahmemöglichkeit am Seminar *Das Arroganz-Prinzip* ermöglicht. Dies schloss explizit die Projektleiterinnen des SFB ein. Trotz des plakativen Namens bietet dieses Seminar eine praxisorientierte Erprobung des Umgangs mit männlichen Kollegen, Mitarbeitern und Vorgesetzten. Mittels eines persönlichen Sparringspartners je Teilnehmerin werden in Übungen unter anderem die Themen angemessene Selbstdarstellung, Raumverhalten als Machtfaktor oder Wahrnehmung der Leistung von Frauen durch Männer trainiert.

Zur besseren Vereinbarkeit von Familie und Beruf wurde im September 2011 unter Führung des Dortmunder Sonderforschungsbereichs 823 eine Tagespflegestelle für Kinder von Beschäftigten in DFG-Projekten, die *9 × klugen Zwerge*, eingerichtet. Dieses Angebot erweitert die existierenden öffentlichen Betreuungsangebote. Die Groß-Tagespflegestelle, die mit dem Mütterzentrum Dortmund zusammenarbeitet, ist ein Zusammenschluss von drei Tagesmüttern, die jeweils drei Kinder in diesen Räumlichkeiten betreuen. Aktuell befinden sich zwei Kinder von Promovendinnen und Promovenden in der Einrichtung. Die Kosten werden auf Basis eines Dienstleistungsvertrags mit dem Mütterzentrum zu ca. 30% aus Gleichstellungsmitteln getragen, der Rest aus Mitteln der Grundausstattung.

Als Maßnahme zur Verbesserung der *Sicherheit auf Dienstreisen* kann, wie bei Antragsstellung der ersten Phase geplant, ein Tagungshotel auch dann als Unterbringung gewählt werden, wenn dieses teurer als die festgesetzten Maximalsätze sind und es kann nachts ein Taxi genommen werden.

Diese Aktionen sollen auch in der zweiten Phase des Sonderforschungsbereichs fortgeführt werden, wozu im Teilprojekt Z wieder Mittel für Chancengleichheitsmaßnahmen beantragt werden.

1.3.5. Wissenschaftlicher Nachwuchs

Als übergreifendes Karriereentwicklungsprogramm existiert das ScienceCareerNet Ruhr[21] (SCNR) seit 2009. Diese Initiative der UA Ruhr richtet sich an Promotionsinteressierte, Promovendinnen und Promovenden oder Postdoktorandinnen und Postdoktoranden. Das in Abschnitt 1.3.4 beschriebene mentoring[3] ist einer der Bausteine aus diesem Programm. Weitere Möglichkeiten zur Fortbildung sind das Karriereforum, Coaching-PLUS und die Frühjahrs-/Herbstakademie. Geeignet sind die Programme für alle Fachrichtungen, wobei diese aufgrund der unterschiedlichen Anforderungen teilweise auf Ingenieur-, Natur- sowie Geistes-, Wirtschafts-, Gesellschafts- und Bildungswissenschaften zugeschnitten sind. Abgerundet werden die Veranstaltungen des SCNR durch die Frühjahrs- und Herbstakademien, sowie den Sommerabend. Die Akademien bieten Leitfäden und Information einerseits für Promotionsinteressierte, frisch mit der Arbeit begonnene (Frühjahr) als auch

[20] http://www.tu-dortmund.de/uni/Wirtschaft/Recruiting/female_2_enterprises/
[21] http://www.scn-ruhr.de/

bereits in der Endphase befindliche Promovendinnen und Promovenden (Herbst). Der Sommerabend ist eine Vortragsveranstaltung mit dem Fokus auf der Vernetzung von Postdoktorandinnen und Postdoktoranden.

Neben den Promotionsarbeiten der Wissenschaftlerinnen und Wissenschaftler erfüllen auch Abschlussarbeiten Studierender eine wichtige Funktion innerhalb des Sonderforschungsbereichs. Eine davon ist der eigentliche Inhalt, ebenso wichtig sind aber die frühzeitige Qualifikation und Einbindung vielversprechender Kandidaten für eine Position im SFB. Den Promovenden des SFB wiederum ermöglicht die Betreuung von Abschlussarbeiten eine verantwortungsvolle Rolle in der Ausbildung. Bis zur Antragstellung für die zweite Phase wurden ca. 70 Abschlussarbeiten zu den Themen des Sonderforschungsbereichs erfolgreich abgeschlossen.

Immer wieder unterstützen die Mitarbeiterinnen und Mitarbeiter des Sonderforschungsbereichs auch Veranstaltungen für die Förderung junger Talente und Forscher. Dies betrifft z.B. Kurse für Teilnehmergruppen des *Bundeswettbewerbs Informatik* zur Einführung in das maschinelle Lernen oder die Mitgliedschaft in der Jury von *Jugend forscht*. Die in Dortmund bestehenden Programme zur Graduiertenausbildung werden im Einzelnen bei der Beschreibung des integrierten Graduiertenkollegs diskutiert.

Zu den obersten Zielen, die der SFB 876 mit der Ausbildung von Promovierenden verbindet, gehört eine frühe Heranführung an wissenschaftlich selbständige Forschung und nach Möglichkeit auch Lehre. Mit besonderer Freude berichten wir deshalb, dass ein Wissenschaftler, der schon bei der Vorbereitung des SFBs mitgewirkt einen Ruf erhalten hat: *Alexander Skopalik* war vom 1.10.2011 bis 30.9.2012 im Teilprojekt A2 des SFBs beschäftigt. Er hat dort sein Wissen über die algorithmische Spieltheorie in die Modellierung von eigennützigem Verhalten in verteilten Systemen einfließen lassen. Zum 1.10.2012 nahm er einen Ruf auf eine Juniorprofessur für algorithmische Spieltheorie an die Universität Paderborn an.

Zudem haben sich zwei Mitarbeiter des SFB nun als Teilprojektleiter qualifiziert. So ist unter den Doktoranden *Tim Ruhe* in der ersten Projektphase besonders hervorgetreten. Als Diplom-Physiker hatte Herr Ruhe die Aufgabe übernommen, Methoden des maschinellen Lernens in der im SFB 876 erforschten Form zur Signal-Untergrund-Trennung in der Neutrino-Astronomie zu etablieren. Vor Beginn der Forschungsarbeit von Herrn Ruhe wurde die Signal-Untergrund-Trennung in dem exemplarisch für die Studien genutzten Großteleskop IceCube mit einfachen statistischen Methoden durchgeführt, was zu einer eher nur qualitativen Übereinstimmung der selektierten Signalereignisse mit der Monte Carlo-Erwartung führte. In seiner Dissertation hat Herr Ruhe diese Übereinstimmung bis auf Fehler im Bruchbereich von Promille verbessert und dabei die Rate an Signalereignissen verdreifacht. Da sich das Neutrinoteleskop IceCube am Südpol befindet, liegen die mit der Verbesserung der Analyse verbundenen Einsparungen an Energie-, Speicher- und Datenübertragungsressourcen auf der Hand. Die methodische Forschung im Rahmen des SFB 876 hat somit maßgeblich zu der Entdeckung des extraterrestrischen Neutrino-Signals durch IceCube beigetragen. Parallel dazu hat Herr Ruhe bereits als Doktorand bei der Ausbildung von Co-Promovierenden, in Seminaren und Vorlesungsvertretungen zur Lehre sowie zu administrativen Aufgaben bei der Drittmittelverwaltung beigetragen und somit das Aufgabenspektrum eines Postdocs erfüllt. Seit dem WS 2013 ist Herr Ruhe als Postdoc und in der Lehre an der Fakultät Physik tätig. Unabhängig von der Hauptlinie seiner Dissertation hat Herr Ruhe damit begonnen, ein auf maschinelles Lernen gestütztes Verfahren zur Lösung von mehrdimensionalen inversen Problemen zu entwickeln. Erste Erfolge des Konzeptes wurden auf der größten Cosmic Ray Konferenz (ICRC) 2013 vorgestellt. Der SFB 876 freut sich, Herrn Ruhe als Teilprojektleiter für das Projekt C3 gewonnen zu haben.

Im August 2011 konnte *Sangkyun Lee* als Postdoktorand für die Arbeiten im Projekt C1 gewonnen werden. Seine Dissertation hatte er direkt zuvor als Wissenschaftler an der Universität Wisconsin-Madison, USA, mit dem Thema *Optimization Methods for Regularized Convex Formulations in Machine Learning* abgeschlossen. Der erste Kontakt zu Herrn Lee fand in 2009 über Katharina Morik auf der ECML PKDD-Konferenz statt. Zu dieser Zeit befand sich der Sonderforschungsbereich in der Entwurfsphase für den Antrag zur ersten Förderperiode, wobei bereits erkennbar war, wie die Expertise von Herrn Lee den SFB bereichern könnte, zumal er auch über einen Abschluss in Bioinformatik und in Statistik verfügt. Der Abschluss seiner Dissertation erfolgte passend, um ihn zeitnah in den neu gegründeten Sonderforschungsbereich zu integrieren. Mit Herrn Lee stärkt der SFB den Bereich der Bioinformatik. Mit ihm ist der SFB 876 auch etwas internationaler geworden. Mit seiner Expertise im Bereich der Optimierung und statistischen Modellierung konnte Sangkyun Lee das Projekt C1 überaus erfolgreich vorantreiben. Kernbeiträge im Projekt waren z.B. die Entwicklung einer neuen Methode zur Identifikation biologisch relevanter Attribute unter statistischen Garantien und der effiziente Einsatz von Support Vector Machines für sehr hochdimensionale Daten. Er bringt seine herausragende Kompetenz in diskreter Optimierung in unterschiedliche Verfahren ein. Die meisten Lernverfahren lösen im Kern ein Optimierungsproblem und können daher von verbesserten Optimierungsalgorithmen und Regularisierungen profitieren. Die Aufgabe als Teilprojektleiter im Projekt C1 für die zweite Förderphase ist die logische Folge der wissenschaftlichen Arbeiten von Herrn Lee, die er dann eigenverantwortlich weiter

voran treiben kann. In Vorbereitung der zweiten Phase arbeitet er bereits an Veröffentlichungen zu den Themen der Optimierung graphischer Modelle. Im Rahmen des SFB hat Herr Lee Studierende betreut und eine sehr gut aufgenommene Vorlesung mit anspruchsvollen Übungen zur diskreter Optimierung entwickelt und durchgeführt. Der SFB 876 erwartet, dass Herr Lee seine akademische Karriere erfolgreich fortsetzen wird.

Laufende oder während der Förderperiode abgeschlossene Dissertationen

TP	Name, Vorname	Finanz.	Thema	Beginn ggf. Abschluss
A1	Streicher, Jochen	EA	Data-driven Adaptation in Ubiquitous System Software	01.2012
A1	Piatkowski, Nico	EA	Probabilistic Graphical Models in Resource Constrained Environments	01.2012
A1	Pölitz, Christian	GA	Hypothesis-driven mappings to suitable example spaces for high-volume linguistic data	04.2014
A1	Jungermann, Felix	GA	About the Exploration of Data Mining Techniques using Structured Features for Information Extraction	01.2008 - 06.2012
A2	Krivošija, Amer	GA	Kernmengen für Clusteringprobleme	02.2012
A2	Schwiegelshohn, Chris	EA	Kernmengenbasierte Datenstromverfahren für Lernprobleme	02.2011
A3	Rempel, Eugen	GA	Statistische Analyse von genomweiten toxikologischen Daten	01.2010
A3	Lang, Michel	EA	Model selection for high-dimensional survival analysis	01.2011
A3	Kotthaus, Helena	EA	Optimizations for Efficient Resource Utilization in biostatistical Applications described in the dynamic Programming Language R	04.2012
A3	Cordes, Daniel	GA	Automatic Parallelization for Embedded Multi-Core Systems using High-Level Cost Models	03.2009 - 11.2013
A3	Plazar, Sascha	GA	Memory-based Optimization Techniques for Real-Time Systems	07.2007 - 06.2012
A3	Kammers, Kai	GA	Survival models with gene groups as covariates	04.2007 - 06.2012
A4	Böcker, Stefan	GA	Leistungsbewertung von Nahbereichsmobilfunksystemen zur Steuerung induktiver Ladesysteme	01.2013
A4	Borchert, Christoph	GA	Aspektorientierte Methoden der Fehlertoleranz in eingebetteten Betriebssystemen	01.2011
A4	Kaulbars, Dennis	EA	Optimierung der Ressourceneffizienz drahtloser Kommunikation in eingebetteten Systemen	10.2013
A4	Buschhoff, Markus	EA	Resource-Aware Operating System Design	11.2011
A4	Dusza, Björn	EA	Context-Aware Power Consumption Modeling for Energy Efficient Mobile Communication Services	01.2011 - 02.2014

A4	Putzke, Markus	GA	Selbstorganisierende Minimierung der Interferenz von Femtozellen in heterogenen Netzen durch zufällige Frequenzsprungverfahren	01.2011
A4	Masoudinejad, Mojtaba	GA	Modelling and Prediction of Hybrid Indoor Energy-Harvesting Applications in Facility Logistics	01.2013
A4	Roidl, Moritz	GA	Bewertung der Leistungsverfügbarkeit von Logistiksystemen ressourcenbeschränkter, verteilter Ladungsträger	01.2013
A4	Emmerich, Jan S.	EA	Entwurfsraumexploration ressourcenbeschränkter, cyberphysischer Systemplattformen für komplexe Logistiksysteme	01.2013
A5	Krümpelmann, Patrick	EA	Inferenz- und Änderungsoperationen nicht-monotoner Logiken für die Anwendung in Multiagenten Systemen unter Berücksichtigung von Vertraulichkeitsanforderungen.	04.2011
A5	Tadros, Cornelia	EA	Belief Change Operations under Confidentiality Requirements in Multiagent Systems	02.2011 - 04.2014
A5	Preuß, Marcel	GA	Inference-proof materialized views	02.2011
A6	Zey, Bernd	GA	Optimierung von stochastischen Netzwerkproblemen (bis 12.2014 in Elternzeit)	03.2010
A6, B4	Molina Ramirez, Alejandro	GA	Mining Relational Contingency Tables	01.2014
A6	Kriege, Nils	GA	Algorithm Engineering für Probleme aus der Cheminformatik	11.2010
TB1	Martin, Marcel	GA	Algorithms and Tools for the Analysis of High-Throughput DNA Sequencing Data	10.2007 - 11.2013
TB1	Horsch, Salome	EA	Statistische Methoden für die Analyse von IMS-Daten	08.2014 (vor.)
TB1	Kopczinski, Dominic	EA	Ressourcenbschränkte Analyse von Spektrometriedaten	01.2011
TB1	D'Addario, Marianna	EA	Ressourcenbeschränkte Analyse von Spektraldaten	01.2011
B2	Timm, Constantin	GA	Resource-Efficient Processing and Communication in Sensor/Actuator Environments	09.2006 - 10.2012
B2	Fiedler, David	GA	Beiträge zur Analyse, Modellierung und Kalibrierung von Kameras und 3D-Tiefensensoren	01.2008 - 03.2014
B3	Stolpe, Marco	EA	Distributed Real-Time Analysis of Vertically Partitioned Sensor Measurements	01.2009
B3	Blom, Hendrik	GA	Realzeitliche Steuerung durch Sprachkonzepte auf Datenströmen	06.2011

B3	Konrad, Benedikt	GA	Eine Methodik zur Produktfamilien-basierten Austaktung von Variantenfliesslinien	11.2010
B3	Bohnen, Fabian	GA	Eine Methodik zur Produktionsnivellierung auf der Basis von Fertigungsfamilien	07.2009 - 9.2013
B4	Habel, Lars	EA	Simulation of Highway Traffic with Cellular Automata	03.2012
B4	Ide, Christoph	EA	Efficient LTE Machine-Type Communication in Vehicular Environments	01.2011
B4	Schweikowski, Florian	GA	Hochgenaue Lokalisierung in In- und Outdoor Szenarien	10.2013
C1	Timm, Henning	GA	Ressourcen-effiziente DNA Sequenzanalyse mit Gapped q-Grammen auf moderner Hardware	10.2013
C1	Schwermer, Melanie	EA	Funktionelle Validierung von alternativ genutzten Transkripten in aggressiven Neuroblastomen	05.2011
C1	Fielitz, Kathrin	EA	Modellierung der MYCN-getriebenen Tumorgenese in vitro und in transgenen Modellen	03.2011
C3	Thieler, Anita Monika	GA	Robuste Verfahren zur Periodendetektion in ungleichmäßig beobachteten Lichtkurven	04.2009 - 01.2014
C3	Milke, Natalie	GA	Unfolding of the atmospheric neutrino flux spectrum with the new program TRUEE and IceCube	01.2009 - 06.2012
C3	Backes, Michael	EA	Long-term observations of the TeV blazar 1ES 1959+650	01.2009 - 12.2011
C3	Köhne, Jan-Hendrick	GA	Der Leptonpropagator PROPOSAL	01.2011 - 12.2013
C3	Overkemping, Ann-Kristin	EA	Entwicklung von Methoden zur Analyse von Daten der Gamma-Astronomie (MAGIC) mit RapidMiner	05.2012
C3	Scheriau, Florian	GA	Optimierung von Data Mining-Methoden in der Neutron-Astronomie	01.2012
C3	Ruhe, Tim	GA	Data-Mining on the Rocks: A measurement of the atmospheric muon neutrino flux using IceCube in the 59-string configuration and a novel data mining based approach to unfolding	01.2010 - 05.2013
C3	Fuchs, Tomazs	Merc.	Optimierung des Monte Carlo Lepton-Propagators PROPOSAL zur Analyse atmosphärischer Myonen mit IceCube	10.2012
C3	Schmitz, Martin	GA	Entwicklung vom Methoden zur Entfaltung von Messdaten der Neutron-Astronomie	10.2011

C3	Thaele, Julia	EA	Entwicklung von Methoden zur Analyse von Daten der Gamma-Astronomie (FACT) mit RapidMiner	05.2012
C3	Temme, Fabian	HAP	Entwicklung von Analyse-Software für das Cherenkov-Teleskop FACT	02.2013
C3	Buss, Jens	GA	Entwicklung einer Online-Analyse für die Gamma-Astronomie auf Grundlage von Datenstromalgorithmen	10.2013
C3	Bockermann, Christian	EA	The streams framework – modeling processes on streaming data	01.2011
C4	Munteanu, Alexander	EA	Effiziente Bayessche Regressionsverfahren	02.2011
C4	Geppert, Leo	EA	Effiziente Bayessche Regressionsverfahren	01.2011
C4	König, Helena	GA	Copy Number Estimation Using Nonparametric Monotonic Bayesian Regression	04.2010

1.3.6. Erkenntnistransfer und Öffentlichkeitsarbeit

Wichtiges Instrument für die Außendarstellung des Sonderforschungsbereichs und erste Anlaufstelle für Informationsmaterial ist der Internetauftritt[22]. Informationen über die Projekte sind hier ebenso abrufbar wie die letzten Publikationen, technischen Berichte oder Aktionen des Graduiertenkollegs wie die Sommerschulen und Topical Seminars. Frei verfügbare Publikationen sind hier ebenfalls direkt zum Download zu finden. Neben den eigenen Mitarbeiterinnen und Mitarbeitern liegt die primäre Zielgruppe dabei auf einem Fachpublikum, das bereits Vorwissen im Forschungsbereich des SFB hat.

Für den Transfer der Bedeutung der Forschung für die allgemeine Öffentlichkeit sind Artikel in regionalen und überregionalen Medien zu nennen:

- Westfälische Rundschau, 08.09.2012, *80 Teilnehmer aus aller Welt* zur ersten Sommerschule.
- Westdeutsche Allgemeine Zeitung, 20.09.2012, *Forscher der TU Dortmund sortiert Trefferliste auf Google.*
- Der Westen, 08.10.2012, *Suche nach dem Muster für Relevanz.*
- Westdeutsche Allgemeine Zeitung, 09.10.2012, *Wie die Nadel im Heuhaufen.*
- Ruhrnachrichten, 26.10.2012, *Sensor erkennt Vireninfektion beim Menschen.*
- Laborpraxis, 10.12.2013, *Nanoobjekte vor Ort aufspüren.*
- Laborjournal, 01.06.2013, *Analyse-Limit.*
- Süddeutsche Zeitung, 23.10.2013, *Schnüffel-Diagnose.*
- Die Glocke, 22.03.2014, *Wie lange der Akku hält, hängt auch vom Netz ab.*
- ARD, 24.07.2014, *Die große Show der Naturwunder*, Beitrag zum PAMONO-Virensensor.

Die Wissenschaftlerinnen und Wissenschaftler konnten auch auf vielen öffentlichen Vorträgen außerhalb der Universität die Bedeutung der Themen des Sonderforschungsbereichs darstellen, zum Beispiel hielt Peter Marwedel auf dem jährlichen Treffen der NSF-Projektleiter im Oktober 2012 in Washington auf Einladung der NSF einen Vortrag zur europäischen Sicht auf die Ausbildung in diesem Bereich und stellte das Konzept eines Dortmunder CPS-Studiengangs auf einem Ausbildungsworkshop während der CPS-Week 2013 in

[22] sfb876.tu-dortmund.de

Philadelphia vor. Er hat in den letzten drei Jahren insgesamt zehn Einladungen zu Vorträgen im Bereich CPS erhalten. Auf dem 2. Westfalenkongress am 18.9.2013 in Dortmund hielt Katharina Morik einen Vortrag, in dem der SFB 876 unter dem Thema *Der Wert der Daten und wie man ihn ausschöpft* vorgestellt wurde. Selbstverständlich beteiligen sich die Mitglieder des SFB auch bei den öffentlichen Veranstaltungen der TU Dortmund, zuletzt auf der 40-Jahrfeier der Fakultät Informatik oder dem Tag der offenen Tür mit Experimenten und Vorführungen zu den Projekten A1, B3 oder B2.

Um die Inhalte des Sonderforschungsbereichs bei Veranstaltungen auch einer breiteren Öffentlichkeit anschaulich vermitteln zu können, wurden im ersten Schritt prototypisch von den Projekten A4, B2, C1 und C3 Leporellos mit ausgewählten Zielen des Projekts erstellt. Im nächsten Schritt entstand daraus eine Broschüre über den gesamten Sonderforschungsbereich, in der unter dem Leitmotiv *Große Daten – Kleine Geräte* sowohl die übergreifende Zielsetzung und langfristige Perspektive dargestellt wurden, als auch in Faltblättern die Möglichkeit zur Einzeldarstellung der Projekte gegeben wurde.

Das Themenspektrum der Faltblätter greift dazu nicht den gesamten Inhalt des jeweiligen Projekts auf, sondern stellt jeweils exemplarisch die Auswirkungen auf die Lebenswelt dar. Die Titel reichen von *Smarter Phones – lernende Telefone müssen seltener an die Steckdose* über *Stauvermeidung 2.0 – Vom Wissen Vieler zum Gewinn für Alle* bis zu *Was hat Krebserkennung mit Astrophysik zu tun?*

Der Sonderforschungsbereich war im November 2013 auf der *Wissenswerte*[23] in Bremen mit einem eigenen Stand vertreten. Diese Konferenz ist die größte deutsche Veranstaltung für Wissenschaftsjournalisten, bei der diese sich über Zukunftstrends in der Wissenschaft informieren. Vertreten war der SFB vor Ort mit den Projekten A4, B2, C1 und C3 mit Demonstratoren und Anschauungsobjekten. Auch hier wurden unter dem Leitmotiv *Große Daten – Kleine Geräte* die Gemeinsamkeiten der zunächst augenscheinlich vollkommen verschiedenen Projekte dargestellt, die Ressourcenbeschränkung. Von den kleinen Geräten in A4 und B2 über die hochdimensionalen Daten in C1 bis hin zu den großvolumigen Daten in C3 zeigen diese das Gesamtkonzept des Sonderforschungsbereichs beispielhaft für alle Projekte auf. Während des Formats *Meet the experts* standen Katharina Morik und Alexander Schramm für detaillierte Fragen zur Verfügung. Hier kam der Kontakt zu der Journalistin zustande, der zu der Präsentation des Teilprojektes B2 in der ARD-Sendung *Die große Show der Naturwunder* führte.

Bei der Öffentlichkeitsarbeit wird der SFB unterstützt durch das Referat Hochschulkommunikation, das Referat Hochschulmarketing sowie das IT & Medien-Centrum, z.B. hinsichtlich Gestaltung des Messestandes.

1.4. Andere Förderungen der beteiligten Teilprojektleiterinnen und Teilprojektleiter

Beteiligte/r TP-LeiterIn	TP	Forschungsprojekt	Beantragter/ Bewilligter Zeitraum	Zuwendungsgeber / Aktenzeichen
Baumbach	TB1	YeastScent	2014 - 2017	BMBF 031 A301 F
Baumbach	TB1	SpiCos	2014 - 2017	BMBF bortnikv-043
Chen	B2	Generating and Executing Dependable Application Software on UnReliable Embedded Systems	2013 - 2015	DFG SPP 1500, CH985/4-1
Chen	B2	Parametrische Analyse und Optimierung der maximalen Laufzeit für eingebettete Echtzeitsysteme (PARTES)	2014 - 2017	DFG (in Beantragung seit 03.2014)
Chen	B2	Erstellen und Ausführen von zuverlässiger Anwendungssoftware auf unzuverlässigen eingebetteten Systemen (Get-SURE) - II	2015 - 2017	DFG (in Beantragung seit 03.2014)
Deuse	B3	CSC	2013 - 2016	BMBF 02PJ2655

[23] www.wissenswerte-bremen.de

Deuse	B3	CSR	2012 - 2014	BMAS E015-NW-015
Deuse	B3	Ergokom	2012 - 2014	AiF 440 ZN
Deuse	B3	IndivA	2014 - 2016	BMBF 16SV6252
Deuse	B3	NED	2013 - 2015	34-78/V51A/19010/Ta
Deuse	B3	NeFIna	2012 - 2014	HBS S-2012-543-1
Deuse	B3	Pro Mondi	2012 - 2015	BMBF 02PJ1110
Deuse	B3	ReiMaFlu	2012 - 2014	AiF 17495 N/1
Deuse	B3	variANa	2012 - 2014	AiF 17531 N
Deuse	B3	LCA-Plan (Planungssystematik für hybride Montagesysteme unter Berücksichtigung von Low-Cost-Automation-Ansätzen)	2014 - 2017	BMWi/AiF (N04227/14, in Beantragung)
Deuse	B3	MITHRAS (Multi-human-robot Assembly for Collaborative Automation)	2015 - 2018	EU (in Beantragung seit 04.2014)
Hergenröder	B2	Tailoring photo-induced interfacial electron transport processes for artificial photo-synthesis	2015 - 2019	EU, SEP-20142582 (in Beantragung)
Hergenröder	B2	In situ Untersuchung der Interaktion zwischen Wasser und kleinen Peptiden an Au(110) und Cu(110) Oberflächen mittels Umgebungsdruck-Photo-Elektronen-Spektroskopie (NAP-XPS) und Reflexions-Anisotropie-Spektroskopie (RAS)	2015 - 2018	DFG HE 1941/23-1 (in Beantragung)
Ickstadt	C4	Statistische Methoden für Schädigungsprozesse unter zyklischer Beanspruchung	2013 - 2017	SFB 823, Teilprojekt B5
Ickstadt	C4	Messung und statistische Analyse von Clusterphänomenen bei Signalproteinen in der Plasmamembran	2011 - 2014	DFG-Projekt IC 5/3-1
Kersting	A6, B4	Relational exploration, learning and inference	2012 - 2015	DFG KE 1686/2-1 im Rahmen des SPP 1527: Autonomes Lernen
Kersting	A6, B4	Efficient Inference for Probabilistic Relational Models using Symmetries and Linear Programming Relaxations	2012 - 2015	GIF 1180
Kersting	A6, B4	Counts are What Counts: From Counting to Modeling Counts in Relational Worlds	1 Jahr	Google (in Beantragung seit 04.2014)

Kersting	A6, B4	COgnitive MoDEls and Methods for Autonomous Robots	2015 - 2017	EU (in Beantragung seit 04.2014)
Marwedel	A3	FEHLER	2011 - 2016	DFG MA 943/10-1/2/3
Marwedel	A3	RTools	2013 - 2014	Oracle
Marwedel	A3	TACLe	2012 - 2016	EU COST IC1202
Morik	A1, B3, C3, Z	Vista-TV	2012 - 2014	EU STREP 296126
Morik	A1, B3, C3, Z	Insight	2012 - 2015	EU 318225
Morik	A1, B3, C3, Z	Kobra	2012 - 2015	BMBF 01UG12445A
Morik	A1, B3, C3, Z	Intelligente Prozess-Steuerung auf Basis von Prozessmerkmalen	2011 - 2014	SMS Siemag
Morik	A1, B3, C3, Z	Data Driven Material Development	2012 - 2014	Mercator Stiftung, Merkur Pr-2012-0023
Morik	A1, B3, C3, Z	Networked platform for the remote monitoring of motor and cognitive status, the personalization of treatment and self-management of Movement Disorders	36 Monate	EU (in Beantragung seit 04.2014)
Morik	A1, B3, C3, Z	Training Future European Data Scientists	36 Monate	EU (in Beantragung seit 04.2014)
Mutzel	A6	GraBaDrug: Graph-basierte Methoden für das rationale Wirkstoffdesign	2014 - 2017	DFG MU 1129/10-1 im Rahmen des SPP 1736: Algorithms for Big Data
Müller	B2	Geometrieverarbeitung für die virtuelle Realisierung produktionstechnischer Prozesse	2011 - 2014	SFB 708, TP B1, B4 (Verlängerung 2015 - 2018 beantragt)
Rahmann	C1	UA Ruhr-Professur Computational Biology	2014 - 2016	MERCUR-Pe-2013-12
Rahmann	C1	Flexibel bis es bricht: Dynamik von Genen & Merkmalen, Dichte & Diversität in Gesellschaften unter sich veränderndern Umweltbedingungen	2014 - 2017	DFG RA 1898/1-1

Rahmann	C1	Data Driven Materials Design	2013 - 2015	Mercator Research Center Ruhr – MERCUR
Rahmann	C1	Mouse modeling and cross-species analysis to optimize identification and targeting of MYCN co-drivers in neuroblastoma - High Throughput data analysis and integration with data from human tumors	2014 - 2018	BMBF eMed
Rahnenführer	A3, TB1	Gen-Verteilungen	2011 - 2016	DFG RA 870/4-1
Rahnenführer	A3, TB1	Differentielle Netzwerke	2011 - 2015	DFG RA 870/5-1
Rahnenführer	A3, TB1	eTop: LivSys	2013 - 2015	BMBF 031A270B
Rahnenführer	A3, TB1	eTop: SysDT	2014 - 2015	BMBF 031A272E
Rahnenführer	A3, TB1	Modelle der Krankheitsprogression	2011 - 2016	DFG RA 870/6-1
Rahnenführer	A3, TB1	SAW	2014 - 2017	Leibniz, SAW-2014-ISAS-2
Rhode	C3, MGK	Verbundprojekt MAGIC: MAGIC High Quantum Efficiency (Teilprojekt 2)	2014 - 2017	BMBF, Deutsches Elektronensynchrotron (DESY), Projektträger für Astroteilchenphysik, BMBF grant 05A14PE1
Rhode	C3, MGK	Verbundprojekt CTA: Softwareentwicklung für das CTA-Projekt	2011 - 2014	BMBF, Deutsches Elektronensynchrotron (DESY), Projektträger für Astroteilchenphysik, BMBF grant 05A11PEA
Rhode	C3, MGK	Theoretische Untersuchung von Signaturen der kosmischen Strahlung in astrophysikalischen Hochenergieflares	2012 - 2014	MERCUR, Vorhaben Pr-2012-0008
Rhode	C3, MGK	Helmholtz Alliance for Astroparticle Physics (HAP) , Teilprojekt Dortmund	2011 - 2016	Helmholtz-Gemeinschaft Deutscher Forschungszentren
Rhode	C3, MGK	Rekonstruktion zeitvariabler Verteilungen in statistischen inversen Problemen	2013 - 2017	Sonderforschungsbereich Statistik (SFB 823), Statistical modelling of nonlinear dynamic processes
Schramm	C1	Role of the tyrosine kinase TrkA and TrkB in checkpoint activation and DSB repair	2012 - 2015	DFG GRK 1739

Schramm	C1	Role of tumor- and host-derived Galectin-1 in neuroblastoma	2012 - 2015	Deutsche Krebshilfe
Schramm	C1	Modelling primary tumor metabolism in neuroblastoma to identify central nodes for therapeutic intervention	2014 - 2018	BMBF eMed
Schreckenberg	B4	BAB-NRW	2001 - laufend	Straßen.NRW 02100.150/4.30.15.01/ OLSIM
Schreckenberg	B4	ESAM	2011 - 2014	DFG SCHR 527/5-1
Schreckenberg	B4	UR:BAN	2012 - 2015	BMWi 19P11007Q
Sohler	A2, A6, A4	Sublinear Algorithms for the Analysis of Very Large Graphs	2012 - 2017	ERC-StG 307696
Sohler	A2, A6, A4	Computational-intensive methods for the robust statistical analysis of non-standard data	4 Jahre	EU oc-2013-1-15257 (in Beantragung)
Spaan	C5	LHCb: Upgrade, Detektorbetrieb und Datenanalyse	2012 - 2015	BMBF: 05H12PE2
Spinczyk	A1, A4	LavA	2013 - 2015	DFG, SP 968/4-2
Spinczyk	A1, A4	DanceOS	2013 - 2014	DFG, SP 968/5-2 (Verlängerung 2015 - 2016 beantragt)
Spinczyk	A1, A4	Fehlertolerante und echtzeitfähige System- und Kommunikationsinfrastrukturen für die Schutz- und Leittechnik	2012 - 2014	DFG, SP 968/6-1 (Verlängerung 2015 - 2017 beantragt)
ten Hompel	A4	Realtime Logistics - Echzeitfähige Logistik	2012 - 2015	DFG, HO 2403/9-1
ten Hompel	A4	Analytische Methoden für die Berechnung der Leistungsverfügbarkeit komplexer Materialflusssysteme	2012 - 2015	DFG, HO 2403/8-1
ten Hompel	A4	DyCoNet	2010 - 2014	01MA09057
ten Hompel	A4	smaRTI	2010 - 2014	01|C10L10A
ten Hompel	A4	Smart Face-Smart Micro Factory für Fahrzeuge mit schlanker Produktionsplanung	2014 - 2016	BMBF 01MA13007D
ten Hompel	A4	Inventairy	2014 - 2016	BMBF 01MA13006A
ten Hompel	A4	EBS Fraunhofer Supply Chain School	2011 - 2015	BMBF 01/C10L08D
ten Hompel	A4	Hub2Move	2013 - 2015	BMBF 01/C12L02A
ten Hompel	A4	InKoRisk	2014 - 2016	BMBF 19G11003D
ten Hompel	A4	PlaSkoPro	2013 - 2016	BMBF 01MA13007D
ten Hompel	A4	Prestige	2014 - 2017	BMBF 16KIS0086
ten Hompel	A4	VILOMA Visual Logistics Management	2013 - 2016	BMBF 19G13003D

ten Hompel	A4	APEM Automatische Paletten Etikettier-Maschine	2013 - 2015	BMBF 02PK2375
ten Hompel	A4	CACTUS	2012 - 2015	BMVI 03EMEN06B
ten Hompel	A4	ModuLOG	2012 - 2015	NRW Bank 300 225 202
ten Hompel	A4	LOG2020	2012 - 2014	EU 527700-LLP-1-2012-1-NL-ERASMUS-EMCR
ten Hompel	A4	SmARPro	—	BMBF (in Beantragung seit 02.2013)
ten Hompel	A4	ICL Innovation City Logistik Bottrop	—	VDI (in Beantragung seit 05.2014)
ten Hompel	A4	Erstellung eines Leitfadens zur Umsetzung der VDI-Richtlinie 4486 Leistungsverfügbarkeit bei Abnahmen von Intralogistiksystemen	2014 - 2016	BMWi AiF BVL, Antragsnr. 522
Wietfeld	A4, B4	TIE-IN	2011 - 2014	Ziel 2, 64.65.69-EM-1022A
Wietfeld	A4, B4	openECOSPhERE	2011 - 2014	BMWi 01ME12091
Wietfeld	A4, B4	e-DASH	2011 - 2014	EU-FP7 Grant Agreement No 285586
Wietfeld	A4, B4	metropol-E	2012 - 2014	BMVBS 03EM0605D
Wietfeld	A4, B4	SmartC2Net	2012 - 2015	EU-FP7 Grant Agreement No 318023
Wietfeld	A4, B4	AIRBEAM	2012 - 2015	EU-FP7 Grant Agreement No 261769
Wietfeld	A4, B4	Anchors	2012 - 2015	BMBF 13N12204
Wietfeld	A4, B4	eNterop	2012 - 2014	BMWi 01MX12060D
Wietfeld	A4, B4	DFG-Forschergruppe 1511	2012 - 2014	Wi 3751/1-1, Wi 3751/2-1
Wietfeld	A4, B4	PlanGridEV	2013 - 2016	EU 608957
Wietfeld	A4, B4	SecInCoRe	2014 - 2017	EU-FP7 Grant Agreement No 607832
Wietfeld	A4, B4	TaMIS	2014 - 2016	BMBF (in Beantragung seit 02.2014)
Wietfeld	A4, B4	AUTOMAT	2015 - 2017	EU (in Beantragung seit 04.2014)
Wietfeld	A4, B4	L4SER	2015 - 2017	EU (in Beantragung seit 04.2014)
Wietfeld	A4, B4	Entfernter Eigenverbrauch mit Elektrofahrzeugen	2015 - 2017	BMVI (in Beantragung seit 07.2014)

Wietfeld	A4, B4	ARCODIN	2015 - 2017	EU (in Beantragung seit 04.2014)
Wietfeld	A4, B4	BERCOM	2015 - 2017	BMBF (in Beantragung seit 05.2014)
Wietfeld	A4, B4	PSA-NG	2014 - 2017	BMBF (in Beantragung seit 04.2014)

2. Vorhandene und beantragte Mittel

2.1. Vorhandene Mittel

2.1.1. Übersicht über vorhandene Sachmittel

Haushaltsjahr	Grundausstattung der antragstellenden Einrichtung(en)	Grundausstattung der weiteren beteiligten Einrichtungen	Weitere Mittel	Gesamtsumme
2011	87.600	31.000	0	118.600
2012	78.600	28.000	0	106.600
2013	84.200	28.000	0	112.200
2014	81.800	28.000	0	109.800
Endende Förderperiode	332.200	115.000	0	447.200
2015	103.500	31.000	0	134.500
2016	90.500	31.000	0	121.500
2017	90.500	31.000	0	121.500
2018	88.500	31.000	0	119.500
Neue Förderperiode	373.000	124.000	0	497.000
Summe	705.200	239.000	0	944.200

2.1.2. Übersicht über das vorhandene Personal

Kategorie	Anzahl der beteiligten Personen
Professuren	19
Nachwuchsgruppenleiterinnen und -leiter	2
Postdoktorandinnen und Postdoktoranden	6
Doktorandinnen und Doktoranden	33
Sonstige wissenschaftliche Mitarbeiterinnen und Mitarbeiter	0
Nichtwissenschaftliches Personal	18
Studentische und Wissenschaftliche Hilfskräfte	16

2.1.3. Liste der Geräte, die dem Sonderforschungsbereich zur Verfügung stehen

TP	Bezeichnung des Gerätes	Anschaffungs-jahr	Anschaffungs-preis	Art der Finanzierung
Alle TP	Computer-Farmen LiDOng, PhiDO, DGrid-RR; anteilig	2007 - 2010	500.000	§91b, Bund, Hochschule
A1, B3, C1, C3	Labor „Analyse großer Datenmengen und verteilter Datenströme“	2014	47.900	Investitionsmittel der Fakultät für Informatik
A1, A4	Labor „Systemsoftware für Cyber-physische Systeme“	2014	66.000	Investitionsmittel der Fakultät für Informatik
A4, B4	LTE-Basisstation und Core-Emulator	2013	78.000	BMBF Projekt ANCHORS (ab ca. Mai 2015 für den SFB verfügbar)
A4	Geräteausstattung Energy-Harvesting Labor: Rohde & Schwarz Network-Vektor-Analysator NVB-4 und Zubehör	2013	41.000	Fraunhofer IML
A4	Elektronik Prototypenfertigung	2013	92.000	Fraunhofer IML
A4, B4	FSVR - Real Time Spectrum Analyzer	2011	118.500	EU/ZIEL2 Projekt AVIGLE (Projekt beendet, in die Grundausstattung übergagangen)
A4, B4	Propsim C8 Funkkanalemulator	2011	229.100	BMBF Projekt SPIDER (Projekt beendet, in die Grundausstattung übergagangen)
A4, B4	CMW 500 - LTE Tester und Basisstationsemulator	2011	188.200	BMBF Projekt SPIDER (Projekt beendet, in die Grundausstattung übergagangen)
A6	Data Mining Lab	2014	33.800	Strukturfonds NRW
B1	IMS Bioscout (2x)	2010	154.000	B&S Analytik GmbH
B2	Nanosight-Gerät als Referenz zur Plasmonendetektion	2013	25.000	ISAS
B2	Quantax	2008	100.000	ISAS
B2	Near-Ambient Pressure XPS	2009	800.000	ISAS
C1	Luminiszenzimager (Fa. Turner Biosystems)	2008	20.000	Haushalt
C1	QIAXcel (Fa. Qiagen) zur DNA und RNA Qualitätskontrolle	2010	20.000	Haushalt
C1	Xcelligence (Fa. Roche) zur Echtzeitdetektion von Zellproliferation	2008	57.000	BMBF
C1	Zwei Multicore-Server und Storage-Array (ca. 100 TB)	2011-2012	80.000	Berufungsmittel/Haushalt
Z, alle TP	SFB-Forschungscluster (in Beschaffung, geplanter Betrieb ab Ende 2014)	2014	390.000	Strukturfonds NRW, Haushalt

2.2. Bisherige und beantragte Mittel

2.2.1. Übersicht

Haushaltsjahr/ Förderperiode	Personalmittel	Sachmittel	Investitionsmittel	Stipendienmittel	Globale Mittel	Gesamt Summe
2011	1.538.400	144.900	0	13.200	40.000	1.736.500
2012	1.562.100	115.200	0	13.200	50.000	1.740.500
2013	1.630.500	115.200	0	13.200	130.000	1.888.900
2014	1.670.100	115.200	0	13.200	340.000	2.138.500
Endende Förderperiode	6.401.100	490.500	0	52.800	560.000	7.504.400
2015	2.091.600	149.300	0	14.400	77.200	2.332.500
2016	2.091.600	141.700	0	14.400	77.200	2.324.900
2017	2.091.600	141.700	0	14.400	77.200	2.324.900
2018	2.091.600	141.700	0	14.400	77.200	2.324.900
Neue Förderperiode	8.366.400	574.400	0	57.600	308.800	9.307.200

2.2.2. Übersicht über Personal, für das Mittel beantragt werden

TP	2015				2016				2017				2018			
	Postdocs	Promovierende	Nichtwiss. Personal	Hilfskräfte	Postdocs	Promovierende	Nichtwiss. Personal	Hilfskräfte	Postdocs	Promovierende	Nichtwiss. Personal	Hilfskräfte	Postdocs	Promovierende	Nichtwiss. Personal	Hilfskräfte
A1	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
A2	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
A3	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
A4	0	3	0	0	0	3	0	0	0	3	0	0	0	3	0	0
A6	0	3	0	0	0	3	0	0	0	3	0	0	0	3	0	0
B1	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0
B2	1	2	0	0	1	2	0	0	1	2	0	0	1	2	0	0
B3	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
B4	0	3	0	0	0	3	0	0	0	3	0	0	0	3	0	0
C1	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
C3	0	3	0	0	0	3	0	0	0	3	0	0	0	3	0	0
C4	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
C5	0	2	0	0	0	2	0	0	0	2	0	0	0	2	0	0
Z	2	0	1	22	2	0	1	22	2	0	1	22	2	0	1	22
Summe	3	29	1	22	3	29	1	22	3	29	1	22	3	29	1	22

Für die Promovierenden im geplanten Sonderforschungsbereich werden 100%-Stellen vorgesehen und in den Teilprojekten entsprechende Mittel beantragt.

2.2.3. Liste der Investitionen, für die Mittel beantragt werden

Es werden keine Investitionsmittel beantragt.

2.2.4. Angaben zur Versuchstierhaltung

Es werden keine Versuchstiere gehalten.

3. Darstellung des Programms nach Projektbereichen und Teilprojekten

3.1 Allgemeine Angaben zum Teilprojekt A1

3.1.1 Titel:

Data Mining für ubiquitäre Systemsoftware

3.1.2 Fachklassifizierung:

Data Mining/Maschinelles Lernen, Eingebettete Systeme

3.1.3 Projektleitung

Prof. Dr. Morik, Katharina, 14.10.1954

LS 8, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 12
44227 Dortmund

Tel.: 0231-755-5100
Fax : 0231-755-5105
E-Mail: `katharina.morik@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Spinczyk, Olaf, 06.04.1970

LS 12, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 16
44227 Dortmund

Tel.: 0231-755-6322
Fax : 0231-755-6116
E-Mail: `olaf.spinczyk@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Das Teilprojekt A1 schlägt die Brücke zwischen eingebetteten Systemen und Datenanalyse. Es entwickelt Algorithmen zu allen Schritten der Informationsgewinnung, von der Datenerfassung bis zur Evaluation des Nutzens der Analyse für die Verringerung des Ressourcenbedarfs eingebetteter Systeme. Die Datenanalyse soll den Ressourcenverbrauch eingebetteter Systeme senken. Gleichzeitig sollen die Algorithmen der Datenanalyse selbst möglichst wenig Ressourcen verbrauchen. In diesem Spannungsverhältnis werden Verfahren nicht nur gemäß dem Verhältnis ihres eigenen Ressourcenverbrauchs zur Qualität ihres Ergebnisses bewertet, sondern auch bezogen auf den Nutzen für die Ressourcenschonung anderer Schritte. Das Spannungsverhältnis kann am Beispiel der *Smartphones* als ubiquitäre Ausführungsplattform verdeutlicht werden. Dort sind die Beschränkungen der Ressourcen Laufzeit, Energie, Speicher und Kommunikation anschaulich gegeben. Gleichzeitig bieten sie eine komplexe Systemsoftware, anhand derer die Instrumentierung und die Systemanpassung untersucht werden kann. Der mögliche Nutzen gelernter Modelle ist dort leicht zu vermitteln: Reduktion des Energieverbrauchs, kürzere Antwort- und *Start-Up*-Zeiten oder ein verbesserter Netzwerkdurchsatz.

Die Datenerhebung bei gleichzeitigem Gebrauch des Gerätes stellt eine besondere Herausforderung dar: die Messung des Ressourcenverbrauchs darf diesen nicht verändern und den Gebrauch nicht erschweren. Das in der ersten Förderperiode entwickelte generische *Framework* für die Datenerhebung, MobiDAC, unterstützt verschiedene Mechanismen der Instrumentierung, so dass wir in der zweiten Phase Methoden untersuchen können, mit denen im Datenstrom gezielt Kontextinformationen abgefragt werden können. Auch kann dann an Verfahren gearbeitet werden, die den Energieverbrauch der Datenerhebung reduzieren.

Als Lernproblem bietet die Parameterschätzung probabilistischer graphischer Modelle gemäß *Maximum-Likelihood* die Möglichkeit, sowohl diskriminative *Conditional Random Fields* (CRF) als auch generative *Markov Random Fields* (MRF) zu untersuchen. Algorithmen werden für unterschiedliche Ausführungsplattformen (GPU, ARM) entwickelt und auf ihren Ressourcenverbrauch hin untersucht. Neben parallelen Lösungen wird auch eine Approximation graphischer Modelle entwickelt, welche ausschließlich auf ganzzahligen Berechnungen basiert und daher praktisch weniger CPU-Taktzyklen und damit weniger Energie verbraucht. Diese Studien zu architekturspezifischen Algorithmen für Lernverfahren sollen systematisch vorangetrieben werden.

Ausgehend von Daten zur Nutzung von *Apps* wurde ein effizienter MRF-Algorithmus für raumzeitliche Daten entwickelt. Dabei konnte erstmals eine Sparsifizierung der MRF-Modellparameter durchgeführt werden, ohne die bedingte Abhängigkeitsstruktur des graphischen Modells zu verändern. Künftige Untersuchungen sollen die empirischen Ergebnisse verallgemeinern und Einflussfaktoren auf das Verhältnis von Güte und Kompression feststellen. Über die Ressourcenschonung bei der Datenanalyse hinaus soll der Nutzen des raumzeitlichen MRF (*Spatio-Temporal Random Fields*, STRF) für die Ressourcenschonung bei *Smartphones* systematisch untersucht werden. Die Ressourcenbeschränkung soll ganzheitlich untersucht werden. Diese Einsicht leitet die folgenden Arbeiten, indem wir stets zusätzlich zu Ressourcenverbrauch und Güte des einzelnen Schrittes auch den Nutzen für ein verbessertes Verhältnis von Ressourcenverbrauch und der Güte der anderen Schritte betrachten.

Die Systemanpassung anhand von (gelernten) Nutzungsmustern wird systematisch untersucht. Sie soll die Ausnutzung gelernter Modelle im Querschnitt durch alle Komponenten und Abstraktionsebenen des Betriebssystems ermöglichen. Unterschiedliche Maße für den Nutzen des Gesamtsystems müssen operationalisiert werden. Aufbauend auf den Energiemessungen, die bereits vorgenommen wurden, und der Modellierung im Teilprojekt A4, soll die Einsparung von Energie bei ubiquitären Systemen unterstützt werden. Bisher fehlen für die komplexen raumzeitlichem Abhängigkeiten, wie sie bei Nutzerverhalten auftreten, öffentlich verfügbare Benchmarks. Diese Lücke will A1 durch eine Entwicklung von Benchmarks und Simulationsmodellen für die Evaluation schließen.

3.3 Bisherige Entwicklung des Teilprojekts

Im Folgenden werden die Fortschritte des Projekts A1 in der ersten Förderperiode erläutert und mit den Entwicklungen beim Stand der Forschung in Beziehung gesetzt. Die Struktur orientiert sich an den Arbeitspaketen des Erstantrags.

3.3.1 Bericht und Stand der Forschung

Datenerhebung (AP1/AP2): Im Gegensatz zu „statischen“ Datensätzen, die nach der Erhebung händisch gepflegt und für die Datenanalyse vorbereitet werden, wie dies z.B. bei den Datensätzen des *UCI Machine Learning Repository* der Fall ist, werden Daten im Projekt A1 im laufenden Betrieb ubiquitärer Systeme

aus deren Systemsoftware erhoben, woran sich direkt die Analyse und die Anwendung der Analyseergebnisse anschließen kann. Es ergeben sich daraus ähnliche Fragestellungen wie beim *Profiling* von Software. Die Instrumentierung darf den Betrieb nicht beeinflussen und der dafür benötigte Ressourcenbedarf hat direkten Einfluss auf die Nützlichkeit des Ergebnisses. Daher sind systemsoftwareseitige Datenquellen sorgfältig auszuwählen und zu konfigurieren, weshalb es einer flexiblen Infrastruktur zur dynamisch rekonfigurierbaren Datensammlung bedarf. Eine solche Infrastruktur ist das im Projekt A1 entwickelte MobiDAC, womit im Rahmen eines Experiments ein frei verfügbarer Datensatz von zahlreichen Android-Telefonen erhoben wurde [A1/56]. Wir haben uns dabei auf *Opportunistic Sensing* beschränkt.

Größere Datensammlungen zur Nutzungs- und Mobilitätsanalyse von Mobiltelefonen wurden auch anderenorts durchgeführt, z.B. *Reality Mining* am MIT oder bei der *Lausanne Data Collection Campaign* [28]. Zu einer laufenden Kampagne gehört der *Cambridge Device Analyzer*[1]. Das neuere *Funf* [2] erlaubt eingeschränkt die, auch entfernte, Konfiguration instrumentierter Geräte: Auf der Granularität einzelner Sensoren und Datenquellen der Android-API können Periode und Dauer konfiguriert werden.

MobiDAC ist dabei wesentlich flexibler und unterstützt verschiedene Mechanismen der Instrumentierung, womit Daten sowohl auf Kernelebene als auch auf höheren Abstraktionsebenen, wie z.B. der Android-API, erhoben werden können. Über einen zentralen Dienst können die Smartphones dabei jederzeit mittels entsprechender Instrumentierungsmodule für verschiedenste Kampagnen rekrutiert werden, was aber unter konfigurierbarer Kontrolle der Nutzer steht. Die flexiblen Module können Daten nicht nur sammeln, sondern auch vorverarbeiten und aggregieren. Um Daten aus dem Betriebssystemkern zu erheben, verwendet MobiDAC *SystemTap*, womit wir solche Daten auch auf Mobilgeräten erheben konnten, wie zum Beispiel Sequenzen dateisystembezogener Systemaufrufe [A1/48].

Bei der Arbeit an der Systemanpassung (AP7) hat sich gezeigt, dass wir neben reinen Software-Sensordaten (z.B. aus dem Betriebssystemkern) auch externe, physikalische Sensoren (Beschleunigung, GPS, ...) als Datenquellen berücksichtigen müssen, da der externe Kontext des Gerätes und seines Benutzers für viele der angestrebten Optimierungen essenziell ist. Im Gegensatz zur Datenerhebung mittels Softwaresensoren (z.B. mit SystemTap), ist die Erhebung softwareunabhängiger Kontextdaten über physikalische Sensoren auch in den (überwiegenden) Zeitabschnitten von Interesse, in denen das Gerät nicht bedient wird. Wir haben daher in Kooperation mit Henrik Blunck von der Universität Aarhus an der Integration des energieeffizienten GPS-Positionierungssystems *EnTracked* [29] in die Android-Systemsoftware gearbeitet. Zum teilweise hohen Energiebedarf der entsprechenden Hardware (GPS) kommt hinzu, dass die CPU ausschließlich zum Erfassen der Sensordaten wiederum regelmäßig aufgeweckt werden muss. Beispielsweise setzt EnTracked eine auf dem Beschleunigungssensor basierende Bewegungsdetektion ein. Während diese auf der ursprünglichen Plattform (Nokia N95/N97) nur wenige Milliwatt zusätzlicher Leistungsaufnahme erforderte, musste die CPU bei unserer Android-Variante explizit geweckt und wachgehalten werden[2]. Die in der Zwischenzeit entstandenen Ansätze, die Erfassung und Verarbeitung von Sensordaten auf energiesparende Coprozessoren auszulagern [34], wollen wir in der zweiten Phase verstärkt berücksichtigen.

Allgemein kommt der energieeffizienten Positionsbestimmung von Mobilgeräten große Bedeutung zu. Methoden, die sich an der gewünschten Präzision orientieren, sind dabei die Verwendung alternativer Positionierungsmethoden (anstatt GPS) über mobile Netzwerke [45] und das Scheduling der Positionierungshardware mittels Duty Cycling [29, 38], was um die temporäre Abschätzung der Position mittels Koppelnavigation [33] oder sogar Trägheitsnavigation [39] erweitert werden kann. Ebenfalls eingesetzt wird die Vorhersage der Position durch Mobilitätsanalyse [11, 39]. Diesen Methoden gemein ist die Notwendigkeit einer Möglichkeit zur Abschätzung des Positionierungsfehlers. Weitere Methoden sind hierarchisches Sensorscheduling mittels Bewegungsdetektion über weniger energieintensive Sensoren [39, 27, 29, 38] und die Verwendung von Straßenkarten zur Präzisierung der Position. Je nach Systemschnittstelle können diese Methoden in der Applikation selbst angewandt werden. Synergien zwischen mehreren kontextsensitiven Diensten können aber nur in der Systemsoftware selbst ausgenutzt werden. Um dies zu erreichen, ist jedoch detailliertes Wissen darüber nötig, wie die Anforderungen jeder Applikation bezüglich der Qualität der Positionierung aussehen. Existierende APIs erlauben dies nur in sehr groben Zügen, was unsere künftigen Arbeiten in diesem Themenbereich motiviert.

Datenstromalgorithmen zur Datenaggregation (AP3): Die regelmäßige Erfassung verschiedener Daten ubiquitärer Systeme erzeugt eine Menge von Datenströmen. Diese Datenströme werden entweder direkt von einem Lernverfahren verarbeitet oder für eine spätere Batch-Verarbeitung gespeichert. In beiden Fällen sollen die Daten komprimiert werden, um einerseits Speicher zu sparen und andererseits mit geringem Aufwand

[1] http://deviceanalyzer.cl.cam.ac.uk/

[2] Um Energie zu sparen wechselt Android in einen Ruhezustand („suspend to RAM“), sobald das Gerät nicht mehr aktiv genutzt (Display deaktiviert) wird und dies softwareseitig nicht explizit (über sogenannte *Wakelocks*) verhindert wird.

übertragen und verarbeitet zu werden. In der ersten Förderperiode wurden daher Datenstromalgorithmen zur Datenaggregation hinsichtlich der Qualität der resultierenden Aggregation sowie ihres Ressourcenverbrauchs untersucht und erweitert.

Aggregation: Aggregationen einströmender Daten werden danach beurteilt, wie stark sie komprimieren, und welchen Nutzen sie für ein nachfolgendes Lernverfahren haben. Zur Aggregation von Sequenzen wurden zunächst *Hierarchical Heavy Hitters* (HHH) [13] betrachtet. Ausgehend von den Arbeiten von Graham Cormode zu HHH haben wir untersucht, ob sich ein Strom von Linux-Systemaufrufen auf einem ubiquitären System so komprimieren lässt, dass ein k-Nearest Neighbor (kNN) Verfahren direkt auf dem ubiquitären Gerät Nutzungsmuster klassifizieren kann. Da HHH hierarchische Daten behandelt, konnten die Systemaufrufe durch drei Taxonomien als Merkmale repräsentiert werden. Anhand von 1,8 GB gesammelten *Logs* von Systemaufrufen wurden die Laufzeit und der Speicherbedarf von HHH und die Güte im Vergleich zur exakten Lösung für unterschiedliche Parameter untersucht. Die kompakten Muster (häufig zusammen vorkommende Tupel) wurden als Klassen von Nutzerverhalten interpretiert. Der Lernschritt war dann die Klassifikation von Log-Daten in diese Nutzerverhalten. Tatsächlich erwiesen sich die Kompression (im Kilobyte-Bereich) und die Laufzeit (bei *Full Ancestry* maximal 14 Millisekunden) als hervorragend und bestätigten die Eignung von HHH für die Zusammenfassung, wenn die Daten durch Taxonomien beschrieben sind [A1/48].

Clustering bietet eine Möglichkeit, Daten zusammenzufassen. Zum einen wurde ein Algorithmus entwickelt, der mittels multikriterieller evolutionärer Optimierung aus der großen Menge häufiger Mengen eine Pareto-Front erstellt, deren Lösungen Clusterings auf der Grundlage von häufigen Mengen darstellen [A1/51]. Clustering wird hier als multikriterielles Optimierungsproblem aufgefasst, das häufige Mengen in hierarchische Cluster überführt. Damit wird die Redundanz der häufigen Mengen vermieden und eine Struktur großer Kollektionen von gemeinsam auftretenden Daten erzeugt. Ein anderes Clustering-Verfahren wurde anhand der Problemstellung des Teilprojektes C1, hochdimensionale Gendaten in Gruppen zu partitionieren, entwickelt. Darin haben wir untersucht, wie die Diskretisierung der Variablen und ein darauf folgendes *Subspace-Clustering* parallelisiert werden können. Das Subspace-Clustering, wie es mit dem CLIQUE-Algorithmus [1] eingeführt wurde, ist auf nominale Wertebereiche beschränkt. Im Gegensatz zu den gängigen Verfahren berücksichtigen wir die ordinale Ordnung der Intervalle. Durch Parallelisierung ist das neue Subspace-Clustering-Verfahren effizient gelöst. Die Algorithmen sind sowohl für Compute-Cluster als auch eine Datenstrom-Umgebung geeignet.

Daneben wurde die Aggregation von Zeitreihen mittels Clustering und dem *Dynamic Time Warping* (DTW) Distanzmaß [25] evaluiert. Obwohl die asymptotische Laufzeit von DTW quadratisch in der Länge der Zeitreihen ist, zeigen empirische Versuche, dass die reale Laufzeit durch eine Kaskade verschiedener Abbruchkriterien für die Distanzberechnung mit der vom euklidischen Distanzmaß vergleichbar ist. Es konnte gezeigt werden, dass k-means mit DTW als Abstandsmaß bei Fixpunktiteration für die Berechnung der Zentroiden zu einem (lokalen) Optimum konvergiert.

Datenstromverarbeitung: Die Lambda-Architektur für *Big Data* verbindet Datenströme (*Speed Layer*) mit sehr großen Mengen gespeicherter Daten (*Batch Layer*) für bestimmte Dienste (*Service Layer*) [37]. Angestrebt wird eine abstrakte Schicht, die sich in bekannte Realisierungen unterschiedlicher Programmiermodelle und Architekturen übersetzen lässt. Das Konzept der Entwicklung von Prozessen für die Lambda-Architektur wurde mit dem *Streams-Framework* (kurz: Streams) [A1/57] realisiert. Es stellt eine abstrakte Schicht bereit, die die Verarbeitung hochdimensionaler Datenströme erlaubt. Vorhandene Lernverfahren auf Datenströmen können eingebettet werden und Streams kann leicht an verschiedene Datenquellen, auch auf ubiquitären Systemen, angebunden werden. Die Verfügbarkeit auf unterschiedlichen Architekturen erleichtert die gezielte Untersuchung von Analysealgorithmen bezogen auf den Ressourcenverbrauch: welcher Algorithmus ist bei welchem Programmiermodell besser geeignet? In einer Kollaboration mit Albert Bifet wurde die *Massive Online Analysis* (MOA) Programmbibliothek[3] an Streams angebunden, wodurch die dort implementierten Algorithmen auf ubiquitären Systemen ausgeführt werden können. Ebenso wurde Streams an das Datensammlungstool MobiDAC (aus Arbeitspaket 2) angebunden, so dass Datenstromalgorithmen direkt auf ubiquitären Geräten evaluiert werden können. Die Realzeitfähigkeit von Streams wurde in anderen Projekten demonstriert, einmal zur Analyse von Fernseh-Daten im europäischen Projekt *Vista-TV*[4], zum anderen in dem europäischen Projekt *Insight*[5], das realzeitlich heterogene, raumzeitliche strömende Daten analysiert. In diesen Anwendungen wird Streams mit einer Ereignisverarbeitung[6] verknüpft. Die Realzeit-

[3] Massive Online Analysis: http://moa.cms.waikato.ac.nz [6]
[4] Video Stream Analysis: http://vista-tv.eu
[5] Intelligent Synthesis and Real-time Response using Massive Streaming of Heterogeneous Data: http://www.insight-ict.eu
[6] Esper for Complex Event Processing: http://esper.codehaus.org

fähigkeit wurde in einem Wettbewerb auf der internationalen Tagung zu *Distributed Event-Based Systems* eindrucksvoll gezeigt und der Beitrag erhielt den Publikumspreis.

Strukturelle Modelle (AP4): Als Lernproblem bietet die Parameterschätzung probabilistischer graphischer Modelle gemäß *Maximum-Likelihood* die Möglichkeit, sowohl diskriminative *Conditional Random Fields* (CRF) als auch generative *Markov Random Fields* (MRF) zu untersuchen. Wir werden zunächst die Notation probabilistischer graphischer Modelle einführen, damit wir dann auf deren Approximation und Reparametrisierung eingehen können.

In diesem Arbeitspaket wurde ebenfalls die Eignung verschiedener graphischer Modelle für die laufenden Systemanpassung (AP6) evaluiert, darunter diskriminative Modelle für Sequenzielle Daten (Conditional Random Fields), generative Modelle für Sequenzielle Daten (Markov Random Fields), sowie generative raumzeitliche Modelle für Sequenzen räumlicher Strukturen (Spatio-Temporal Random Fields).

Probabilistische graphische Modelle: Zu den probabilistischen graphischen Modellen zählen wir alle Wahrscheinlichkeits(dichte)funktionen multivariater Zufallsvariablen $\boldsymbol{X}$ der Form

$$p_{\boldsymbol{\theta},G}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z_G(\boldsymbol{\theta})}\psi_{\boldsymbol{\theta},G}(\boldsymbol{x})$$

mit dem Parametervektor $\boldsymbol{\theta} \in \mathbb{R}^d$ sowie dem Graph $G = (V, E)$ mit Knotenmenge V und Kantenmenge E. Die Knoten des Graphen entsprechen einzelnen Komponenten von $\boldsymbol{X}$ und die Kanten repräsentieren bedingte Unabhängigkeitsrestriktionen. Die Funktion $\psi_{\boldsymbol{\theta},G}$ weist jeder möglichen Realisation $\boldsymbol{x}$ in Abhängigkeit von G und $\boldsymbol{\theta}$ einen positiven reelen Wert zu. Die *Normalisierung* $Z_G(\boldsymbol{\theta})$ ist durch die Bedingung $\sum_{\boldsymbol{x}} p_{\boldsymbol{\theta},G}(\boldsymbol{X} = \boldsymbol{x}) = 1$ eindeutig bestimmt. Ihre Berechnung ist im Allgemeinen #**P**-schwer [10]. Die gängigste Form dieser Modellklasse ist die Exponentialfamilie

$$p_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})}\exp(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle)$$

mit suffizienter Statistik ϕ. Aus Gründen der Übersichtlichkeit wird die explizite Abhängigkeit vom Graph G nicht notiert. Die Parameter $\boldsymbol{\theta}$ werden mittels Maximum-Likelihood Schätzung bestimmt. Bei der Anwendung des Modells sind die Werte einiger Zufallsvariablen $\boldsymbol{x'}$ bekannt, die Werte der anderen Zufallsvariablen $\boldsymbol{y}$ sind unbekannt. Die Vorhersage des Modells ist dann der wahrscheinlichste Wert der multivariaten Zufallsvariable $\boldsymbol{y}$, also $\boldsymbol{y}^* := \max_{\boldsymbol{y}\in\mathcal{Y}} p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x'})$. Die Berechnung von $\boldsymbol{y}^*$ heißt maximum-a-posteriori (MAP) Vorhersage und ist im Allgemeinen **NP**-schwer. Für das Lernen der Modellparameter ist die wiederholte Berechnung des Gradienten der Normalisierung erforderlich. Diese Berechnung ist im Allgemeinen ebenfalls #**P**-schwer. Die Komplexität dieser Berechnungen hängt von der Baumweite von G ab. Ist der Graph G des Modells ein Baum, so lassen sich die o.g. Probleme in Polynomialzeit lösen. Der gängigste Ansatz zur Reduktion der Komplexität bei nicht-kreisfreien Graphen ist die paarweise Approximation. Dabei werden nur die Knoten und Kanten eines Graphen in dem Modell betrachtet und keine Cliquen höherer Ordnung. Diese Vorgehensweise wurde auch von uns eingesetzt. Der *Belief-Propagation* (BP) Algorithmus [32] findet eine Approximation des Gradienten der Normalisierung sowie die wahrscheinlichste Konfiguration (MAP-Vorhersage).

Die raumzeitliche Parametrisierung probabilistischer graphischer Modelle: Ausgehend von Daten zur Nutzung von *Apps* wurde ein effizienter MRF-Algorithmus für raumzeitliche Daten entwickelt. Dabei konnte erstmals eine Sparsifizierung der MRF Modellparameter durchgeführt werden, ohne die bedingte Abhängigkeitsstruktur des graphischen Modells zu verändern. Um vorherzusagen, mit welcher Wahrscheinlichkeit ein System zum Zeitpunkt t am Ort v einen Zustand x_v annimmt, können räumliche und zeitliche Abhängigkeiten von einem generativen probabilistischen graphischen Modell zur Vorhersage genutzt werden. Ist das Verhalten eines Systems über einen gewissen Zeitraum konstant, d.h. die Wahrscheinlichkeiten für den Zustand x am Ort v sind zu den Zeitpunkten t und $t+1$ sehr ähnlich, entstehen Redundanzen in den Modellparametern. Gewöhnliche Modelle sind nicht in der Lage, diese Redundanzen zu erkennen. In der ersten Förderperiode wurde eine Klasse graphischer Modelle, die *Spatio-Temporal Random Fields* (STRF), entwickelt [A1/C1/53], die temporale Redundanzen komprimieren. Dabei wird das ursprüngliche Modell wie folgt reparameterisiert:

$$\boldsymbol{\theta}_v^t(\boldsymbol{\Delta}) = \sum_{i=1}^{t} \frac{1}{t-i+1}\boldsymbol{\Delta}_{vi}.$$

Die neuen Parameter $\boldsymbol{\Delta}$ entsprechen der Änderung der klassischen Parameter $\boldsymbol{\theta}$ über der Zeit. Beim Lernen der neuen Parameter kann nun eine Regularisierung durchgeführt werden, um kleine Änderungen, d.h. kleine Werte $\boldsymbol{\Delta}_{vi}$, auf 0 zu setzen und somit eine spärliche Parametrisierung zu erhalten. Empirische Versuche zeigen, dass die Anzahl der Parameter, die zur Repräsentation des Modells benötigt werden, mit Hilfe einer ℓ_1- sowie ℓ_2-Regularisierung um den Faktor 7 reduziert wird, während die Güte im Vergleich zum unkomprimierten Modell unverändert bleibt. Auf der internationalen Fachtagung *ECML PKDD* 2013 haben wir den *Best-Student-Paper Award* für unsere Arbeiten an raumzeitlichen graphischen Modellen erhalten.

Graphische Modelle in der laufenden Systemanpassung: Mobile Kommunikation erfolgt über drahtlose Netze, die in *Funkzellen* aufgeteilt sind. Eine ressourceneffiziente Methode zur approximativen Bestimmung des Ortes sind die Identifikationsnummern von Mobilfunkzellen. Es wurden zwei alternative Ansätze für die Vorhersage von Funkzellen verglichen [A1/50]. Aus zeitlichen Sequenzen von Mobilfunkzellenidentifikatoren wurden dort Vorhersagemodelle für zukünftig besuchte Mobilfunkzellen eines Nutzers gelernt. Für die Stützvektormethode wurde die raumzeitliche Struktur, d.h. die jeweils k zuletzt besuchten Zellen, mittels Indikatorfunktionen (Ist Zelle i und Position j in der Sequenz) auf einen binären Merkmalsraum abgebildet. Es wurden verschiedene Kerne evaluiert, wobei mit radialen Basisfunktionen (RBF) die höchste Vorhersagegenauigkeit erzielt wurde. Als Alternative wurden generative graphische Modelle untersucht, wobei die zeitliche Struktur mittels eines Graphen abgebildet wird und die verschiedenen Mobilfunkzellen durch Zustände der Knoten $\mathcal{X}_v, v \in V$ repräsentiert werden. Wenn auch die Vorhersagegenauigkeit des generativen Modells in diesem Fall unter der der Stützvektormethode lag, ist der Speicherverbrauch unabhängig von der Anzahl der Trainingsinstanzen, während bei der Stützvektormethode und anderen Kernelmaschinen zunächst der vollständige Datensatz für das Training vorliegen muss. Für die Vorhersage der besuchten Mobilfunkzellen, benötigt das MRF ca. 950 Tsd. Trainingsinstanzen weniger Speicher als die SVM.

Abhängigkeitsstrukturen bei Sequenzen von Systemaufrufen wurden mit CRF erfolgreich modelliert. Verschiedene Ansätze zum Sparen von Ressourcen durch Anpassung von Android-Systemen aufgrund des gelernten Modells, wie intelligenteres *Prefetching* bei Dateizugriffen und Paketbündelung beim UMTS-Datenverkehr, wurden konzipiert und evaluiert. Gerade an diesem Beispiel wurde die Bedeutung des Nutzens als zusätzlichem Gütemaß deutlich: die hervorragende Güte des gelernten Modells führte nicht zu einem ebenso hervorragenden Nutzen für das System.

Hochparallele und verteilte Algorithmen (AP5/AP6): Zur Erhöhung der Energieeffizienz und zur Reduktion der Laufzeit wurden massiv parallele Algorithmen zur Berechnung graphischer Modelle entwickelt und für die Ausführung der Modelle auf programmierbaren Grafikprozessoren (GPU), wie man sie in modernen ubiquitären Systemen vorfindet, implementiert. Die Datenparallelität und Datenlokalität von Lernaufgaben ermöglichen die Ausführung in MapReduce-Umgebungen, die sehr große Datenmengen verteilt ausführen, in verteilten graphbasierten Berechnungsmodellen wie GraphLab [36], oder programmierbaren Grafikprozessoren, die über einen sehr hohen Grad an Parallelität verfügen. Aktuelle GPU können beispielsweise 2880 arithmetische Instruktionen gleichzeitig ausführen. Bekkerman und Kollegen betrachten Parallelität als einen Schlüssel zur Skalierung maschineller Lernverfahren [5].

GPU-beschleunigte Belief Propagation: Die Belief Propagation ist ein wesentlicher Bestandteil der Parameterschätzung graphischer Modelle, da sie für die Berechnung des Gradienten der Normalisierung erforderlich ist. Der Algorithmus besteht aus einer Fixpunktiteration, wobei in jeder Iteration so genannte *Nachrichten* zwischen den Endpunkten einer jeden Kante ausgetauscht werden. Ist $(v, u) \in E$ eine Kante sowie $\mathcal{X}_v$ und $\mathcal{X}_u$ die (diskreten) Domänen der Knoten v und v, so lässt sich die Nachricht von v an u für einen Zustand x_u berechnen als

$$m_{v\to u}(x_u) = \sum_{x_v \in \mathcal{X}_v} \exp(\boldsymbol{\theta}_{v=x_v} + \boldsymbol{\theta}_{vu=x_v x_u}) \prod_{w \in \mathcal{N}_v \setminus \{u\}} m_{w\to v}(x_v), \tag{3.1}$$

wobei $\mathcal{N}_v$ die Menge der Nachbarknoten von v ist und die Nachrichten für jeden möglichen Zustand des Empfängers berechnet werden müssen. In einer Iteration des BP-Algorithmus' werden also $h := \sum_{(v,u)\in E} |\mathcal{X}_v| + |\mathcal{X}_u|$ neue Nachrichten auf Basis der h alten Nachrichten berechnet. Der Algorithmus terminiert, sobald die Berechnung der Nachrichten einen Fixpunkt erreicht hat. Da bei der Berechnung der h neuen Nachrichten keinerlei Schreib- oder Lesekonflikte auftreten, können die h Nachrichten parallel berechnet werden, d.h. BP weist eine hohe Datenparallelität auf. Die Datenlokalität ist dabei durch den Graphen G des Modells gegeben, da nur die Nachrichten der Nachbarknoten sowie die eigenen Parameter zur Berechnung der jeweiligen neuen Nachricht vorliegen müssen. In der ersten Förderperiode haben wir verschiedene Parallelisierungen

der Belief Propagation evaluiert. Bestehende GPU-Algorithmen für Belief Propagation aus dem Bereich der Computer Vision verwenden eine fest einkodierte graphische Struktur [19] und sind daher nicht für allgemeine Lernaufgaben nutzbar.

Eine asynchrone Berechnung der Nachrichten ist auf aktuellen GPUs nicht möglich, da eine feste Menge an Recheneinheiten stets die gleiche Instruktion ausführen muss. Bei der synchronen Belief Propagation kann die Berechnung neuer Nachrichten erst beginnen, sobald die Berechnungen aller Nachrichten der vorherigen Iteration abgeschlossen sind. Ist der Zustandsraum einzelner Knoten groß, so kann dies den Übergang zur nächsten Interation stark verzögern. Ausgehend von einem Algorithmus, der alle Nachrichten einer Iteration parallel berechnet, wurde eine weitere Ebene der Parallelität eingeführt, um dieses Problem zu lösen. Dazu wird die Summe in Gleichung (3.1) von $\frac{1}{2}|\mathcal{X}_v|$ Threads in $\log|\mathcal{X}_v|$ Schritten parallel berechnet [A1/52]. In empirischen Versuchen kann eine Verringerung der Laufzeit um den Faktor 10 im Vergleich zur CPU-basierten Belief Propagation gezeigt werden.

Bei diskriminativen Modellen wie Conditional Random Fields (CRF) gibt es eine zusätzliche Ebene an Parallelität, da die Belief Propagation für jede Trainingsinstanz einzeln durchgeführt werden muss. Der entsprechende GPU-Algorithmus führt daher BP für einen Batch von b Trainingsinstanzen gleichzeitig aus. Die optimale Wahl von b hängt dabei von der tatsächlichen Anzahl vorhandener Rechenkerne sowie von der graphischen Struktur ab. Für sequenzielle Daten wurde eine spezielle Variante der Belief Propagation, der *Forward-Backward Algorithmus*, ebenfalls für die GPU-Ausführung modifiziert. Da die graphische Struktur in diesem Fall fest in den Algorithmus einkodiert ist, führt dies zu einer Reduktion der Speicherzugriffe sowie zu einer geringeren Divergenz der Ausführungspfade einzelner Recheneinheiten.

Verteilte Schätzung der Parameter raumzeitlicher Modelle: Beim verteilten Lernen von Modellparametern wird das Optimierungsproblem in lokale und globale Teilprobleme partitioniert, wie beispielsweise bei der *Alternating Direction Method of Multipliers* (ADMM) für das verteilte Training von Stützvektormaschinen [18]. In TP A1 wurde ein approximativer Optimierungsalgorithmus zweiter Ordnung für verteilte graphische Modelle entwickelt [A1/C1/53], der vollkommen ohne globale Berechnungen auskommt. Der Austausch von Daten findet dabei lediglich zwischen benachbarten Knoten zur Berechnung der Belief Propagation statt. Im Unterschied zu einer Parameterschätzung mit gewöhnlichem Gradientenabstieg, erfolgt eine Schätzung der zweiten Ableitung, sowie eine Schrittweitenadaption. Existierende Optimierungsalgorithmen zweiter Ordnung wie BFGS benötigen Zugriff auf den vollständigen Gradienten, wobei unser Ansatz lediglich lokale Informationen benötigt. Das Training verteilter graphischer Modelle mit dem von uns entwickelten Algorithmus konvergiert in empirischen Versuchen 5 mal schneller als mit gewöhnlichem Gradientenabstieg.

Laufende Systemanpassung (AP7/AP8): Unsere initialen Experimente mit kernelseitig implementierten Betriebssystemstrategien zeigten, dass *Recall* und *Precision* als Bewertungsmaß hier nicht ausreichen, und immer der Nutzen im realen System betrachtet werden muss. Beispielsweise wurde von uns ein gelernter Klassifikator für sequentielle Dateizugriffe verwendet, um die *Read-Ahead*-Strategie des Linux-Kernels zu verbessern [A1/48]. Trotz der hohen Qualität des Klassifikators brachte eine Ersetzung der existierenden adaptiven Heuristik keine, und eine Anpassung dieser mit Hilfe der Vorhersagen nur sehr kleine Verbesserungen bezüglich der I/O-Wartezeit.

Bei den von uns betrachteten Systemen werden Ressourcen stoßweise genutzt, wobei für den Benutzer weniger der Durchsatz als vielmehr Reaktivität und Energieeffizienz wichtig sind. Die größten Energiekonsumenten sind dabei neben dem Display die mobile Datenkommunikation und die Erhebung und Inferenz von Kontext, dabei insbesondere die Positionierungshardware (GPS). Zur Minimierung des Energieverbrauchs ist es dabei erstrebenswert, belegte Ressourcen möglichst auszunutzen. Hierzu gehören beispielsweise zugewiesene Übertragungskanäle und -zeitscheiben in mobilen Netzwerken, deren Leistungsaufnahme einen fixen, von der tatsächlichen Datenübertragung unabhängigen Anteil besitzt, der auch anfällt, wenn die vorhandene Übertragungskapazität nicht komplett genutzt wird. Hierzu gehören insbesondere die *Tail-States* bei mobiler Kommunikation. Durch systemsoftwareseitige Bündelung von Applikationsanfragen kann daher Energie gespart werden. Dies gilt für die Kommunikation [4, 35, 43], für die CPU (*Timer-Alignment* bei periodischen Aufgaben [31]), aber auch für die Sensorik wie z.B. Positionierung (*Piggybacking* von GPS-Fixes [45]). Wenn Systemdienste zu einem späteren Zeitpunkt energieeffizienter ausgeführt werden können, kann es sinnvoll sein diese zu verzögern. Bei der Kommunikation entscheiden die jeweils verfügbaren Netze, der momentane Pfadverlust zur Basisstation und die verfügbare Datenrate über den Energieverbrauch. Es existieren Ansätze, die lediglich anhand von Mittelwert und Varianz dieser Werte über eine Verzögerung der Kommunikation [41] entscheiden, wobei die aktuelle Verzögerung als Korrekturterm einfließt, aber auch solche, die mittels Vorhersage der zu erwartenden Signalstärken und unter garantierter Einhaltung von Deadlines [42] den Energieverbrauch optimieren.

Eine Abschlussarbeit zur Verzögerung und Bündelung von Netzwerkpaketen auf Android-basierten Smartphones wurde von uns durchgeführt. Momentan kennen wir jedoch noch keine Methode, die *Verzögerungstoleranz* einzelner Pakete ohne Anwendungswissen zuverlässig zu bestimmen.

Zu dem aus einem früheren Projekt entstandenen aspektorientierten Betriebssystem CiAO [A1/A4/49] haben wir in Kooperation mit dem Projekt A4 einen hochkonfigurierbaren IP-Stack [A1/A4/46] entwickelt, der uns weiterführende und gezieltere Experimente auch auf sehr kleinen Systemen ermöglichen wird. Interessant ist dabei die Anwendung maschineller Lernverfahren auf typische Fragestellungen hochkonfigurierbarer Infrastruktursoftware-Produktlinien, wie die automatische Auswahl und die Vorhersage nichtfunktionaler Eigenschaften einer bestimmten Variante.

Bei der Positionsbestimmung haben wir das Sensorscheduling in einer Python-VM gekapselt. Die Paketbündelung haben wir als *Queuing Discipline* implementiert. Dies ist ein bereits vorgesehener Mechanismus des Linux-Kernels zur modularen Integration von Paketschedulern. An generischen Ansätzen zur sicheren Adaption ubiquitärer Betriebssysteme anhand unserer Lernergebnisse arbeiten wir momentan. In der kommenden Phase wollen wir Datensammlung, Analyse und die Anwendung der Lernergebnisse in einer gemeinsamen Infrastruktur zusammenführen.

Experimente auf echter Hardware stellen für die Evaluation des Energieverbrauchs eine Herausforderung dar, da Kontext (Position, Signalstärke) und Applikationsverhalten (Ressourcennutzung) vom Nutzerverhalten abhängen. Bei stationären Messungen mit entsprechender Mess-Hardware sind diese schwer zu reproduzieren. Der übliche Ansatz hier ist das Aufzeichnen von für die Änderung der Leistungsaufnahme relevanten Ereignissen, und deren anschließende Auswertung mithilfe entsprechender Energiemodelle. Am Beispiel des Entwickler-Smartphones *Nexus One* haben wir die Eignung batterieinterner Messungen untersucht und konnten nach einer Anpassung eines Treibers den relativen Fehler auf 1% beschränken. Als Problem bleibt nun noch, dass hierfür die CPU wiederum periodisch aufgeweckt werden musste. Ein anderer Ansatz ist das Aufzeichnen relevanter Systemereignisse für die Auswertung mit entsprechenden Energiemodellen. Objektive und vergleichbare Aussagen über Systemverbesserungen sind mittels derartiger Feldstudien jedoch schwer und nur unter großem Aufwand zu treffen, da unterschiedliche Gerätehardware und -software eine besonders breite Nutzerbasis erfordern.

Übliche Ansätze, diese Probleme zu umgehen und automatisiert zu evaluieren, sind daher entweder Simulationen oder die Verwendung standardisierter Benchmarks. Simulationen haben den Vorteil der Reproduzierbarkeit, können durch ihren hohen Abstraktionsgrad jedoch zu verfälschten Ergebnissen führen, da die Reaktion des Benutzers, der Anwendungen oder auch der Kommunikationspartner des Mobilgerätes (z.B. ein Server im Internet) nicht modelliert wird. Benchmarks für Desktop- und Serversysteme sind nur von geringer Aussagekraft für die von uns betrachteten Geräte. *MiBench* [20] wurde für eingebettete Systeme entwickelt, bewertet im Wesentlichen aber nur die Leistungsfähigkeit der Hardware. AndroStep [23] bezieht auch die Systemsoftware ein, untersucht aber nur die Performanz des nichtflüchtigen Speichers. Insbesondere die Interaktion mit dem Nutzer wird für Benchmarks bisher noch nicht betrachtet.

Ebenfalls in Kooperation mit A4 haben wir den Embedded-System-Simulator MobiSIM [A4/A1/47] entwickelt, der auch die Modellierung von Anwendungen und Nutzern ermöglicht. Diese Modelle automatisiert zu erstellen ist noch ein offenes Problem, welches in der zweiten Phase angegangen werden soll.

Bereitstellung von Algorithmen (AP9): Die Bereitstellung von Lernalgorithmen und -umgebungen für andere Teilprojekte war sehr erfolgreich und soll in der geplanten zweiten Phase beibehalten werden. Im AK Datenanalyse wurden RapidMiner und die Datenstromumgebung vorgestellt und für die Projekte C1, B3 und C3 genutzt. Es wurde ein allgemeines *Term Set Clustering* mit evolutionärer Optimierung gefundener häufiger Muster als Plug-in für RapidMiner publiziert [A1/51]. Das in A1 entwickelte Werkzeug MobiDac dient der flexiblen Sammlung von Daten in Smartphone-Netzwerken. Eine im Rahmen der SFB Sommerschule mit den Teilnehmern durchgeführte Datensammlung führte zu einem Datensatz, der beispielsweise auch im Projekt A4 genutzt wird. Im Rahmen des AK Datenanalyse, aber auch in vielen Einzelgesprächen hat das Teilprojekt Promovenden bezüglich ubiquitärer Geräte und maschineller Lernverfahren beraten. Innerhalb der Laufzeit wurden die oben ausführlicher besprochenen Algorithmen publiziert. Die Arbeit an und mit dem in mehreren internationalen Benutzerumfragen führenden[7] und in der Gartner-Studie 2014 im Quadranten der besten und zukunftsorientiertesten Werkzeuge eingestuften Werkzeug zur Datenanalyse, RapidMiner, wurde ebenfalls weiter vorangetrieben. So wurden auf den internationalen Treffen der Entwickler, RComm 2012 und RComm 2013, Vorträge von Christian Bockermann und Hendrik Blom sehr gut aufgenommen. Es ging um die Anbindung an das Streams-Framework, um die Anbindung an Java Beans und um eine neuartige

[7] 2014 wurde RapidMiner erneut bei der internationalen Umfrage unter 3000 Forschern und Anwendern des Data Mining als das bei 44,2% beliebteste Werkzeug ermittelt

Benutzerschnittstelle, bei der die Parallelität oder Abfolge von Prozessen gezeichnet wird. Katharina Morik hielt 2012 den eingeladenen Hauptvortrag über „Mining Machine to Machine Data“.

Literatur

[1] R. Agrawal, J. Gehrke, D. Gunopulos und P. Raghavan. „Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications“. In: *SIGMOD Record (ACM Special Interest Group on Management of Data)* 27.2 (1998), S. 94–105.

[2] N. Aharony, W. Pan, C. Ip, I. Khayal und A. Pentland. „Social fMRI: Investigating and Shaping Social Mechanisms in the Real World“. In: *Pervasive and Mobile Computing* 7.6 (2011), S. 643 –659.

[3] B. Ahmadi, K. Kersting, M. Mladenov und S. Natarajan. „Exploiting Symmetries for Scaling Loopy Belief Propagation and Relational Training“. In: *Machine Learning Journal* 92.1 (2013), S. 91–132.

[4] N. Balasubramanian, A. Balasubramanian und A. Venkataramani. „Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications“. In: *Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference*. ACM. 2009, S. 280–293.

[5] R. Bekkerman, M. Bilenko und J. Langford. *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2011.

[6] A. Bifet, J. Read, B. Pfahringer, G. Holmes und I. Zliobaite. „CD-MOA: Change Detection Framework for Massive Online Analysis“. In: *Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013*. Hrsg. von A. Tucker, F. Höppner, A. Siebes und S. Swift. Springer, 2013, S. 92–103.

[7] D. M. Blei, A. Y. Ng und M. I. Jordan. „Latent Dirichlet Allocation“. In: *Journal of Machine Learning Research* 3 (2003), S. 993–1022.

[8] C. Bucila, R. Caruana und A. Niculescu-Mizil. „Model Compression“. In: *Conference on Knowledge Discovery and Data Mining (KDD)*. Hrsg. von T. Eliassi-Rad, L. H. Ungar, M. Craven und D. Gunopulos. ACM, 2006, S. 535–541.

[9] H. H. Bui, T. N. Huynh und S. Riedel. „Automorphism Groups of Graphical Models and Lifted Variational Inference“. In: *29th Conference on Uncertainty in Artificial Intelligence (UAI)*. Hrsg. von A. Nicholson und P. Smyth. 2013.

[10] A. Bulatov und M. Grohe. „The Complexity of Partition Functions“. In: *Automata, Languages and Programming*. Hrsg. von J. Díaz, J. Karhumäki, A. Lepistö und D. Sannella. Bd. 3142. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, S. 294–306.

[11] Y. Chon, E. Talipov, H. Shin und H. Cha. „SmartDC: Mobility Prediction-Based Adaptive Duty Cycling for Everyday Location Monitoring“. In: *IEEE Transactions on Mobile Computing* 13.3 (2014), S. 512–525.

[12] D. Chu, N. D. Lane, T. T.-T. Lai, C. Pang, X. Meng, Q. Guo, F. Li und F. Zhao. „Balancing Energy, Latency and Accuracy for Mobile Sensor Data Classification“. In: *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. SenSys ’11. ACM. Seattle, Washington: ACM, 2011, S. 54–67.

[13] G. Cormode, F. Korn, S. Muthukrishnan und D. Srivastava. „Finding Hierarchical Heavy Hitters in Streaming Data“. In: *ACM Transactions on Knowledge Discovery from Data* 1.4 (2008), S. 1–48.

[14] W. Duivesteijn. „Exceptional Model Mining“. Diss. Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University, 2013.

[15] Ú. Erlingsson, M. Peinado, S. Peter, M. Budiu und G. Mainar-Ruiz. „Fay: Extensible Distributed Tracing from Kernels to Clusters“. In: *ACM Transactions on Computer Systems (TOCS)* 30.4 (2012), S. 13.

[16] S. Ermon, C. P. Gomes, A. Sabharwal und B. Selman. „Taming the Curse of Dimensionality: Discrete Integration by Hashing and Optimization“. In: *International Conference on Machine Learning*. Bd. 28. JMLR Proceedings. 2013, S. 334–342.

[17] N. Esfahani, A. Elkhodary und S. Malek. „A Learning-Based Framework for Engineering Feature-Oriented Self-Adaptive Software Systems“. In: *Software Engineering, IEEE Transactions on* 39.11 (Nov. 2013), S. 1467–1493.

[18] P. A. Forero, A. Cano und G. B. Giannakis. „Consensus-Based Distributed Support Vector Machines“. In: *The Journal of Machine Learning Research* 99 (2010), S. 1663–1707.

[19] S. Grauer-Gray und J. Cavazos. „Optimizing and Auto-tuning Belief Propagation on the GPU“. In: *Languages and Compilers for Parallel Computing*. Hrsg. von K. Cooper, J. Mellor-Crummey und V. Sarkar. Bd. 6548. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, S. 121–135.

[20] M. Guthaus, J. Ringenberg, D. Ernst, T. Austin, T. Mudge und T. Brown. „MiBench: A Free, Commercially Representative Embedded Benchmark Suite“. In: *WWC'04: Proc. of the 4th IEEE International Workshop on Workload Characteristics*. Austin, Texas, USA, 2001, S. 3–14.

[21] Y. Han, J.-M. Kang, S.-s. Seo, A. Mehaoua und J. W.-K. Hong. „An Energy Efficient User Context Collection Method for Smartphones“. In: *Network Operations and Management Symposium (APNOMS), 2013 15th Asia-Pacific*. IEEE. 2013, S. 1–6.

[22] T. Hazan und T. Jaakkola. „On the Partition Function and Random Maximum A-Posteriori Perturbations“. In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Hrsg. von J. Langford und J. Pineau. ICML '12. New York, NY, USA: Omnipress, 2012, S. 991–998.

[23] S. Jeong, K. Lee, J. Hwang, S. Lee und Y. Won. „Framework for Analyzing Android I/O Stack Behavior: From Generating the Workload to Analyzing the Trace“. In: *Future Internet* 5.4 (2013), S. 591–610.

[24] A. Kansal, S. Saponas, A. Brush, K. S. McKinley, T. Mytkowicz und R. Ziola. „The Latency, Accuracy, and Battery (LAB) Abstraction: Programmer Productivity and Energy Efficiency for Continuous Mobile Context Sensing“. In: *Proceedings of the 2013 ACM SIGPLAN International Conference on Object Oriented Programming Systems Languages & Applications*. Indianapolis, Indiana, USA: ACM, 2013, S. 661–676.

[25] E. Keogh und M. Pazzani. „Scaling up Dynamic Time Warping for Datamining Applications“. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 2000, S. 285–289.

[A6/26] K. Kersting, M. Mladenov, R. Garnett und M. Grohe. „Power Iterated Color Refinement“. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*. Hrsg. von C. Brodley und P. Stone. AAAI Press, 2014.

[27] D. H. Kim, Y. Kim, D. Estrin und M. B. Srivastava. „Sensloc: Sensing Everyday Places and Paths using Less Energy“. In: *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM. 2010, S. 43–56.

[28] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez und J. Laurila. „Towards Rich Mobile Phone Datasets: Lausanne Data Collection Campaign“. In: *Proc. ICPS, Berlin* (2010).

[29] M. B. Kjærgaard, J. Langdal, T. Godsk und T. Toftkjær. „Entracked: Energy-Efficient Robust Position Tracking for Mobile Devices“. In: *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*. ACM. 2009, S. 221–234.

[30] V. Kolmogorov und M. Wainwright. „On the Optimality of Tree-Reweighted Max-Product Message-Passing“. In: *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*. Arlington, Virginia: AUAI Press, 2005, S. 316–323.

[31] V Kononen und P Paakkonen. „Optimizing Power Consumption of Always-on Applications based on Timer Alignment“. In: *Communication Systems and Networks (COMSNETS), 2011 Third International Conference on*. IEEE. 2011, S. 1–8.

[32] F. R. Kschischang, B. J. Frey und H.-A. Loeliger. „Factor Graphs and the Sum-Product Algorithm“. In: *IEEE Transactions on Information Theory* 47.2 (2001), S. 498–519.

[33] Q. Ladetto, V. Gabaglio und B. Merminod. „Combining Gyroscopes, Magnetic Compass and GPS for Pedestrian Navigation“. In: *Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics, and Navigation*. 2001, S. 205–213.

[34] F. X. Lin, Z. Wang, R. LiKamWa und L. Zhong. „Reflex: Using Low-power Processors in Smartphones Without Knowing Them“. In: *SIGARCH Comput. Archit. News* 40.1 (März 2012), S. 13–24.

[35] H. Liu, Y. Zhang und Y. Zhou. „TailTheft: Leveraging the Wasted Time for Saving Energy in Cellular Communications“. In: *Proceedings of the 6th Intern. Workshop on MobiArch*. ACM. 2011, S. 31–36.

[36] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola und J. M. Hellerstein. „Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud“. In: *Proceedings of the VLDB Endowment* 5.8 (2012), S. 716–727.

[37] N. Marz und J. Warren. *Big Data – Principles and Best Practices of Scalable Realtime Data Systems*. Maning Publications, 2013.

[38] J. Paek, J. Kim und R. Govindan. „Energy-Efficient Rate-Adaptive GPS-based Positioning for Smartphones“. In: *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*. ACM. 2010, S. 299–314.

[39] M. Papandrea und S. Giordano. „Enhanced Localization Solution“. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE. 2012, S. 241–246.

[40] A. Pereira, G. Franco, L. Silva, W. Meira Jr. und W. Santos. „The USAR Characterization Model“. In: *Proceedings of the 7th IEEE Intern. Workshop on Workload Characterization (WWC-7)*. IEEE. 2004, S. 63–70.

[41] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger und M. J. Neely. „Energy-Delay Tradeoffs in Smartphone Applications“. In: *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*. ACM. 2010, S. 255–270.

[42] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain und V. N. Padmanabhan. „Bartendr: A Practical Approach to Energy-Aware Cellular Data Scheduling“. In: *Proceedings of the sixteenth Annual International Conference on Mobile Computing and Networking*. ACM. 2010, S. 85–96.

[43] E. J. Vergara, J. Sanjuan und S. Nadjm-Tehrani. „Kernel Level Energy-Efficient 3G Background Traffic Shaper for Android Smartphones“. In: *Proceedings of the 9th International Wireless Communications and Mobile Computing Conference (IWCMC), IEEE*. 2013.

[44] J. Yarkony, C. Fowlkes und A. Ihler. „Covering Trees and Lower-bounds on Quadratic Assignment“. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010.

[45] Z. Zhuang, K.-H. Kim und J. P. Singh. „Improving Energy Efficiency of Location Sensing on Smartphones“. In: *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*. ACM. 2010, S. 315–330.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[A1/A4/46] C. Borchert, D. Lohmann und O. Spinczyk. „CiAO/IP: A Highly Configurable Aspect-Oriented IP Stack“. In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. New York, NY, USA: ACM, Juni 2012, S. 435–448.

[A4/A1/47] M. Buschhoff, J. Streicher, B. Dusza, C. Wietfeld und O. Spinczyk. „MobiSIM: A Simulation Library for Resource Prediction of Smartphones and Wireless Sensor Networks“. In: *Proceedings of the 46th Annual Simulation Symposium*. ANSS ’13. Society for Computer Simulation International, 2013.

[A1/48] P. Fricke, F. Jungermann, K. Morik, N. Piatkowski, O. Spinczyk, M. Stolpe und J. Streicher. „Towards Adjusting Mobile Devices To User’s Behaviour“. In: *Analysis of Social Media and Ubiquitous Data*. Hrsg. von M. Atzmueller, A. Hotho, M. Strohmaier und A. Chin. Bd. 6904. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2011, S. 99–118.

[A1/A4/49] D. Lohmann, O. Spinczyk, W. Hofer und W. Schröder-Preikschat. „The Aspect-Aware Design and Implementation of the CiAO Operating-System Family“. In: *Transactions on AOSD IX*. Hrsg. von G. T. Leavens, S. Chiba, M. Haupt, K. Ostermann und E. Wohlstadter. Lecture Notes in Computer Science 7271. Springer-Verlag, 2012, S. 168–215.

[A1/50] S. Michaelis, N. Piatkowski und K. Morik. „Predicting Next Network Cell IDs for Moving Users with Discriminative and Generative Models“. In: *Mobile Data Challenge by Nokia Workshop in conjunction with Int. Conf. on Pervasive Computing*. Newcastle, UK, 2012.

[A1/51] K. Morik, A. Kaspari, M. Wurst und M. Skirzynski. „Multi-Objective Frequent Termset Clustering“. In: *Knowledge and Information Systems* 30.3 (2012), S. 715–738.

[A1/52] N. Piatkowski. „Parallel Algorithms for GPU Accelerated Probabilistic Inference“. In: *International Workshop on Big Learning, Neural Information Processing Systems (NIPS)*. 2011.

[A1/C1/53] N. Piatkowski, S. Lee und K. Morik. „Spatio-Temporal Random Fields: Compressible Representation and Distributed Estimation“. In: *Machine Learning Journal* 93.1 (2013). Hrsg. von H. Blockeel, K. Kersting, S. Nijssen und F. Zelezny, S. 115–139.

[A1/C1/54] N. Piatkowski, S. Lee und K. Morik. „The Integer Approximation of Undirected Graphical Models“. In: *3rd International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. Hrsg. von M. De Marsico, A. Tabbone und A. Fred. SciTePress, 2014.

[A1/55] J. Streicher, O. Nassour und O. Spinczyk. „System Support for Privacy-preserving and Energy-efficient Data Gathering in the Global Smartphone Network – Opportunities and Challenges“. In: *Proceedings of the 3rd International Conference on Pervasive and Embedded Computing and Communication Systems (PECCS '13)*. SciTePress, 2013, S. 80–85.

[A1/56] J. Streicher, N. Piatkowski, K. Morik und O. Spinczyk. „Open Smartphone Data for Mobility and Utilization Analysis in Ubiquitous Environments“. In: *Proceedings of the 4th International Workshop on Mining Ubiquitous and Social Environments (MUSE)*. Hrsg. von M. Atzmüller und C. Scholz. 2013.

Andere Veröffentlichungen

[A1/57] C. Bockermann und H. Blom. *The Streams Framework*. Techn. Ber. 5. TU Dortmund, Dez. 2012.

3.4 Planung des Teilprojekts

Ziele

Das Teilprojekt A1 schlägt weiterhin die Brücke zwischen eingebetteten Systemen und Datenanalyse. Es entwickelt Algorithmen zu *allen* Schritten der Informationsgewinnung, von der Datenerfassung bis zur Evaluation des Nutzens der Analyse für die Verringerung des Ressourcenbedarfs eingebetteter Systeme und ist diesbezüglich auch Ansprechpartner für andere Teilprojekte.

Im Vordergrund aller Arbeiten in A1 steht der *Nutzen* der Datenanalyse für das Gesamtsystem, welcher sich durch ein komplexes Zusammenspiel von Datenerhebung, Datenanalyse und Maßnahmen zur Systemanpassung ergibt (siehe Abbildung 3.1). Um eine positive Bilanz des Nutzens zu erreichen, erforscht A1 nicht nur ressourcenschonende Methoden bzw. Algorithmen zur Datenerhebung und Datenanalyse, sondern zielt darüber hinaus ab auf eine Reduktion des Ressourcenverbrauchs dieser Schritte durch Ausnutzung von Wissen über die konkreten Systemanpassungsmaßnahmen und die damit verbundenen Anforderungen an die Analyse.

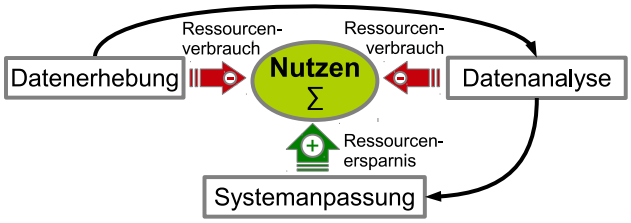


Abbildung 3.1.: Das Spannungsverhältnis von Datenerhebung, -analyse und Systemanpassung.

In der ersten Förderperiode wurden Datenerhebung und Systemanpassung vor allem anhand von *Smartphones* als ubiquitäre Ausführungsplattform behandelt, während Datenzusammenfassung und -analyse für unterschiedliche Plattformen untersucht wurde. Die Arbeit an *Android*-basierten Systemen wird fortgesetzt, weil an ihnen wiederverwendbare Techniken zur Systemanalyse und -anpassung sowie neue Ansätze zur Strukturierung eingebetteter Systemsoftware erforscht werden können.

Ressourcenschonende Datenerhebung im laufenden System: Die Erfassung von Kontextdaten über physikalische Sensoren und auch die potentiell hochfrequente Erfassung von Systemzuständen oder -ereignissen führt zunächst nur zu einer Erhöhung des Ressourcenverbrauchs. Damit diese Kosten den Nutzen nicht übersteigen, In der kommenden Förderperiode soll untersucht werden, wie sich dieser Mechanismus besser

ausnutzen lässt. Dazu sollen Synergien gleichzeitig bearbeiteter Datenerhebungsaufgaben mit Hilfe einer expliziten Spezifikation zeitlicher und qualitativer Anforderungen automatisch erkannt und für eine ressourcenschonende Auswahl und Konfiguration der Sensorik verwendet werden.

Datenanalyse unter Ressourcenbeschränkung: Neben parallelen Lösungen wird eine Approximation graphischer Modelle entwickelt, welche ausschließlich auf ganzzahligen Berechnungen basiert und daher praktisch weniger CPU Taktzyklen und damit weniger Energie verbraucht. Diese Studien zu architekturspezifischen Algorithmen für Lernverfahren sollen systematisch vorangetrieben werden. Es sollen nun auch Algorithmen für Datenströme und MapReduce untersucht werden, sowie Algorithmen für besonders stark beschränkte Systeme. Ausgehend von unseren Arbeiten zur Sparsifizierung von MRF-Modellparametern, bei denen die Reparametrisierung als Funktion der *zeitlichen* Struktur betrachtet wurde, sollen in der zweiten Förderperiode auch Reparametrisierungen als Funktionen anderer latenter Strukturen, wie zum Beispiel der *räumlichen* Struktur, betrachtet werden. Dabei kommen sowohl Muster in der graphischen Struktur als auch eine Zerlegung der graphischen Struktur in Frage. Künftige Untersuchungen sollen die empirischen Ergebnisse verallgemeinern und Einflussfaktoren auf das Verhältnis von Güte und Kompression feststellen. Über die Ressourcenschonung bei der Datenanalyse hinaus soll der Nutzen des STRF für die Ressourcenschonung bei *Smartphones* systematisch untersucht werden.

Systemanpassung: In der ersten Förderperiode wurden einzelne Aspekte von Android-Systemen bezüglich der Verbesserung durch maschinelle Lernverfahren isoliert voneinander betrachtet. Außerdem waren das Lernen eines Vorhersagemodells und die Nutzung des Modells zeitlich getrennte Phasen. In Zukunft soll untersucht werden, wie eine zentrale, schichtübergreifende Lernkomponente in ubiquitäre Systemsoftware integriert werden könnte, die es ermöglicht, über geeignete Schnittstellen Lernaufgaben klassischer Systemkomponenten im laufenden Betrieb entgegenzunehmen, zu analysieren und unter Ausnutzung von Synergien passend parametrisierte Datenanalysen und Datenerhebungen zu verknüpfen. Die so gelernten Modelle können von mehreren Systemkomponenten gleichzeitig genutzt werden, was zu einer effizienteren Ressourcennutzung führt. Es soll untersucht werden, ob die durch diese Infrastruktur ermöglichte umfassende Systemanpassung die Qualität unsicherer Entscheidungen in ubiquitärer Systemsoftware verbessern und damit den Ressourcenverbrauch senken kann.

Bewertung des Nutzens: Unsere Erfahrungen mit konkreten Systemanpassung haben gezeigt, dass ein gelerntes Modells trotz hervorragender Güte nicht unbedingt zu einem ebenso hervorragenden Nutzen für das System führt. Gleichzeitig haben einige Modelle von erheblichem Nutzen nicht notwendigerweise herausragende Werte bei den üblichen Gütemaßen. Es muss daher neben den üblichen Gütemaßen wie *Accuracy*, *Precision* oder *Recall* auch der tatsächliche Ressourcenverbrauch (Laufzeit, Energie, Kommunikation, Speicher) erfasst und erklärt werden. Weil die Konstanten in der asymptotischen Komplexität einen erheblichen Einfluss auf den tatsächlichen Nutzen haben, müssen sie genauer betrachtet werden. So kann gezielt nach Algorithmen gesucht werden, die zwar die gleiche Komplexität aber höheren Nutzen aufweisen.

Die Evaluation des Nutzens der durchgeführten Systemanpassungen geht über die üblichen Bewertungen hinaus. Weder Performanztests, wie sie bei Betriebssystemen üblich sind, noch die Kreuzvalidierung gemäß statistischer Maße, wie sie in der Datenanalyse üblich ist, reichen aus. Obendrein ist bei graphischen Modellen die Kreuzvalidierung insofern nicht anwendbar, als die Rollen von Merkmalen und Labeln in generativen Modellen nicht festgelegt sind. In der zweiten Förderperiode sollen Benchmark-Datensätze und -generatoren entwickelt werden für die doppelte Bewertung, einerseits nach dem eigenen Verhältnis von Güte und Ressourcenverbrauch, andererseits nach dem Nutzen für weitere Schritte. Wir streben die Entwicklung von Benchmarks an, die auf Simulationsmodellen repräsentativen Nutzerverhaltens basieren und dadurch eine zuverlässige und automatisierte Evaluation ressourcenschonender Systemoptimierungen ermöglichen.

Arbeitsprogramm

Die ressourcenschonende Datenerhebung (AP1) erstreckt sich über den Betriebssystemkern, Dienstprozesse und die Anwendungsschicht. Die Zentralisierung dieser Infrastruktur erlaubt es, Synergien zu nutzen und so besonders ressourcenschonend vorzugehen. Die Daten werden in Form von Datenströmen entweder lokal oder im Netzwerk/in der Cloud zusammengefasst, auf Muster hin untersucht, bereinigt und komprimiert. Dabei spielen Datenstromalgorithmen eine entscheidende Rolle (AP2). Weiterhin müssen Algorithmen erforscht werden, die es erlauben, Modelle mit vielfältigen Abhängigkeiten bei akzeptablem Ressourcenverbrauch zu lernen (AP3). Um den Ressourcenverbrauch darüber hinaus zu senken, sollen architekturspezifische Lernverfahren (AP4) entwickelt werden, die mit speziellen Hardware-Eigenschaften wie dem Fehlen einer FPU

oder dem Vorhandensein eines FPGAs effizient umgehen. Gelernte Modelle können dann für eine umfassende Systemanpassung (AP5) genutzt werden. Einzelne Systemkomponenten greifen entweder direkt oder indirekt (durch automatisierte Komponentenkonfigurierung) auf Lernergebnisse zurück, um ressourcenschonend zu agieren. In AP6 werden neuartige Techniken für die Evaluation entwickelt. Dies ist notwendig, da bisherige Ansätze keine zuverlässige und automatisierte Evaluation gemäß des Ressourcenverbrauchs gestatten. Wie in der ersten Förderperiode sollen auch jetzt wieder andere Teilprojekte bzgl. der Nutzung von Datenanalyse- und Lernverfahren beraten werden (AP7).

Arbeitspaket 1. Ressourcenschonende Datenerhebung: Für die sensorbasierte Kontexterfassung existieren, wie bereits berichtet, zahlreiche Möglichkeiten, den Energieverbrauch bei nur minimalem Präzisionsverlust zu reduzieren. Wir wollen von diesen Möglichkeiten auf Ebene der Systemsoftware Gebrauch machen, um dabei Synergien zwischen verschiedenen Datenanfragen sowohl von Applikationen als auch Systemkomponenten und Lernverfahren ausnutzen zu können. Hierfür müssen sich solche Anfragen am erwarteten Nutzen orientieren und Informationen zu Zeit- und Genauigkeitsanforderungen enthalten. Für die Positionierung ist dies beispielsweise bei der *Latency, Accuracy, and Battery Abstraction* [24] der Fall, welche die Spezifikation kontextabhängiger Genauigkeitsanforderungen für die Positionsbestimmung erlaubt, wie es z.B. bei *Geofencing* benötigt wird.

Wir streben die Repräsentation solcher Anfragen als Datenstrom an [A1/55]. Eine Anfrage für Geofencing wäre dann beispielsweise die Verknüpfung des Positionsdatenstroms mit standardisierten Operatoren, so dass der resultierende Datenstrom nur noch die Information darüber enthält, ob man sich in der Nähe eines bestimmten Ortes befindet. Der Ansatz ist damit vergleichbar mit [21], wobei wir uns jedoch nicht auf binäre Kontexte beschränken wollen.

Die verschiedenen Datenstromanforderungen werden kombiniert und transformiert um einen ressourcenschonenden Ausführungsplan für die Datenerhebung zu erhalten. Dabei werden gleichzeitig bearbeitete Anfragen intern vereinigt, so dass Synergien genutzt werden können. Genauigkeitsanforderungen für Sensordaten werden so nahe wie möglich an der Quelle ausgenutzt, indem z.B. *Duty Cycling* für Sensoren verwendet wird, oder diese komplett durch Schätzungen ersetzt werden. Dafür notwendige Berechnungen wollen wir dynamisch günstigen Ausführungsplattformen zuordnen (im Kernel, in einem Service, zentral, z.B. in einer Cloud-Infrastruktur oder auf einem stromsparenden Koprozessor), um so den Ressourcenverbrauch geschickt zu verteilen und Kontextwechsel einzusparen. Datenströme aus dem Smartphone werden entsprechend der zeitlichen Anforderungen gebündelt, womit Kosten für die Kommunikation eingespart werden. Für die Arbeiten nutzen wir Ressourcenmodelle aus dem Projekt A4.

Arbeitspaket 2. Datenströme: Algorithmen für Datenströme zeichnen sich durch geringen Speicherbedarf aus. Insofern sind sie besonders gut für ubiquitäre Geräte geeignet. Datenanalyse wird daher auf ubiquitären Systemen oft durch Lernverfahren realisiert, die Daten online und mit stark begrenztem temporärem Arbeitsspeicher verarbeiten. Die Analyse kann dabei von häufigen Mengen und Clustering über Regression und Klassifikation bis hin zu raumzeitlichen oder allgemein graphischen Modellen reichen. Die unüberwachten Lernverfahren eignen sich darüber hinaus für die Datenzusammenfassung und das Finden von Motiven. Im Bereich des *compressed sensing* und der *data summaries* werden Methoden verwendet, die lokal an einem Sensorknoten oder allgemein an einem ubiquitären System Muster finden, die als Merkmale einem nachfolgenden Lernverfahren zur Verfügung gestellt werden. Vielversprechend ist die Steuerung des unüberwachten Lernens häufiger Mengen durch das Prinzip der *Minimum Description Length*. Hier soll weiter gearbeitet werden. Fortgesetzt wird auch die Arbeit am Clustering von Zeitreihen. Es sollen nun auch räumliche Beziehungen einbezogen werden, so dass raumzeitliche Cluster gefunden werden. Für ubiquitäre Systeme sollen die Daten lokal zusammengefasst werden. Dies stellt besondere Anforderungen an die Energieeffizienz, den Speicherbedarf und die Laufzeit der Datenstromalgorithmen, die von den bisherigen Ansätzen noch nicht erfüllt werden. Die Subgruppenentdeckung ist für die Bestimmung von Nutzerprofilen eine interessante Lernaufgabe, die sowohl als erster Schritt der Datenzusammenfassung als auch als Lernverfahren zu untersuchen ist. Gerade das Entdecken von Ausnahmen (Exceptional Model Mining) [14] aggregiert Information. Die Zuordnung von Daten zu einem EMM-Muster ist für ressourcenbeschränkte Systeme geeignet. Die Entdeckung solcher Ausnahmen hingegen noch nicht. Dies soll in der beantragten zweiten Phase untersucht werden.

Die Arbeit an Datenströmen betrifft zusätzlich die Ausnutzung der Streams-Umgebung, so dass bekannte Algorithmen aus der MOA-Bibliothek [6] und im SFB 876 entwickelte Algorithmen miteinander verglichen werden können. Dabei wollen wir die üblichen Vergleiche nun erstmals um die Güte bezüglich ihres Ressourcenverbrauchs erweitern.

Arbeitspaket 3. Graphische Modelle unter Ressourcenbeschränkung: Die vielfältigen Abhängigkeiten zwischen den in AP1 erhobenen Daten erfordern eine ressourceneffiziente Repräsentation der Modelle [8]. In Anlehnung an unsere Arbeiten zur Kompression raumzeitlicher Modelle [A1/C1/53] werden nun weitere Aspekte graphischer Modelle untersucht. Ausgehend von graphischen Modellen mit latenten Variablen wie die Latente Dirichlet Allokation (LDA) [7] sollen Muster, bzw. *Topics*, im Parameterraum von ungerichteten graphischen Modellen gefunden werden. Das Modell $\boldsymbol{\theta}$ wird komprimiert, indem funktionale Zusammenhänge f_i zwischen den Parametern eines Topics T_j genutzt werden. Auf diese Weise sollen spärliche Reparametrisierungen automatisch gefunden und regularisiert werden.

Zur Verringerung des Energieverbrauchs der Parameterschätzung und der MAP-Vorhersage wird die graphische Struktur zerlegt. Der bekannteste dekompositionelle Ansatz ist die *Tree-Reweighted* Belief-Propagation (TRW-BP) von Kolmogorov und Wainwright [30], wobei ein allgemeiner Graph in die Menge seiner Spannbäume zerlegt wird. Es kann gezeigt werden, dass dieses Vorgehen zu einer oberen Schranke für die Wahrscheinlichkeit der MAP Vorhersage führt, allerdings ist der reale Ressourcenverbrauch höher als bei einfacher BP, da zusätzlich Wahrscheinlichkeiten für die Existenz einer jeden Kante gespeichert werden müssen. Yarkony und Kollegen [44] stellen eine Vereinfachung von TRW-BP vor, wobei der ursprüngliche Graph in nur einen Baum, den so genannten *Covering Tree*, transformiert wird. Der Nutzen der Parameterschätzung und der Modellanwendung mit *Covering Trees* soll evaluiert und mit der Belief Propagation auf nicht-kreisfreien Graphen verglichen werden.

Symmetrien innerhalb der graphischen Struktur können ebenfalls zur Reduktion der Komplexität verwendet werden [9, 3]. Der Graph wird dabei auf eine einfachere graphische Struktur abgebildet, die Ergebnisse werden auf der einfachen Struktur berechnet und mit Hilfe der entsprechenden Umkehrabbildung in den Ursprungsraum übertragen. Die Einsparung an Ressourcen hängt dabei von Symmetrien innerhalb der graphischen Struktur sowie innerhalb der Modellparameter ab. Wir werden die Frage untersuchen, ob die Struktur der Daten zur (automatischen) Einsparung von Ressourcen führt.

Die einzige zur Zeit bekannte Approximation mit garantierter Güte ist von Ermon und Kollegen [16] und beruht darauf, dass alle Probleme in $\#\mathbf{P}$ durch eine polynomielle Anzahl nicht-adaptiver **NP**-Orakelanfragen approximiert werden können. Wenn auch die Güte der Approximation in diesem Fall sehr gut ist, liegt der Ressourcenverbrauch des entsprechenden randomisierten Algorithmus aktuell weit oberhalb der Grenzen ubiquitärer Systeme, da das MAP-Problem mehrfach exakt gelöst werden muss, um die Güte zu garantieren. Hier sollen Covering-Trees eingesetzt werden, um die MAP-Lösungen zu approximieren und so eine Approximation des Gradienten zu erhalten. Der Ressourcenverbrauch dieser Schätzung wird mit dem von BP verglichen. Dabei ist es offensichtlich, dass eine einzelne Iteration der Maximum-Likelihood Schätzung mehr Ressourcen benötigen wird als bei der Nutzung von BP, allerdings kann eine genauere Schätzung des Gradienten zu einer beschleunigten Konvergenz der Optimierung und damit zu einem insgesamt geringeren Ressourcenverbrauch führen. In der Methode von Ermon und Kollegen [16] werden Nebenbedingungen bei der MAP-Vorhersage eingeführt, um den Raum der möglichen Lösungen einzuschränken. Alternativ kann dies durch Perturbationen des Parametervektors erreicht werden [22]. Da sich die Varianten durch Laufzeit, Güte sowie Speicherverbrauch unterscheiden, wird hier untersucht, welches Vorgehen zu einem größeren Nutzen für das ubiquitäre System führt.

Schließlich sollen auch die Arbeiten an der strukturellen Variante der SVM fortgesetzt werden, weil sie einen Vergleich zwischen generativen und diskriminativen Modellen erlaubt. Können aus den Methoden, mit denen Ressourcen bei probabilistischen graphischen Modellen eingespart werden, auch Methoden für die Ressourcenschonung von strukturellen SVM gewonnen werden? Damit streben wir allgemeine Ergebnisse zu strukturellen Modellen und ihrem Ressourcenbedarf an.

Arbeitspaket 4. Architekturspezifische Lernverfahren: Als Fortführung des bisherigen AP5 werden hier, basierend auf neuen Erkentnissen über approximative graphische Strukturen [44], bestehende parallele Algorithmen dahingehend verbessert, dass die graphische Struktur so zerlegt wird, dass die Größe der einzelnen Zusammenhangskomponenten an die verfügbare Architektur angepasst wird. Allerdings sollen auch andere Eigenschaften von Berechnungsarchitekturen für das ressourceneffiziente maschinelle Lernen genutzt werden. In konkreten Rechnerarchitekturen benötigen ganzzahlige Operationen in einigen Fällen weniger als die Hälfte der Taktzyklen der entsprechenden Gleitkommaoperationen und damit auch weniger Energie. Darüber hinaus verfügen einige Rechnerarchitekturen über keinen nativen Gleitkommaprozessor, was den Einsatz der meisten maschinellen Lernverfahren verhindert. Basierend auf einer ersten Arbeit über ganzzahlige graphische Modelle [A1/C1/54] sollen zum einen weitere Methoden des maschinellen Lernens daraufhin untersucht werden, ob die Vorhersagegüte ganzzahliger Approximationen mit der von Gleitkommaalgorithmen vergleichbar ist und ob Schranken für den Approximationsfehler gefunden werden können.

Es wird eine spezielle Variante der Belief Propagation entwickelt, die lediglich ganzzahlige Operationen für

die approximative Berechnung des Gradienten sowie der MAP-Vorhersage verwendet. Dies ist bislang nur auf Baumstrukturen möglich. Die Lösungsidee ist, dass die Proportionalität der Wahrscheinlichkeiten erhalten bleibt, solange alle Zwischenergebnisse die richtige Bitlänge aufweisen. Dieser Ansatz ist vielversprechend und soll weiter untersucht und verallgemeinert werden. Gibt es auch ganzzahlige Varianten der Stützvektormaschine sowie von Entscheidungsbäumen und Clusteringverfahren und wie gut approximieren sie? Zum anderen soll neben ganzzahligen Approximationen und parallelen GPU-Algorithmen überprüft werden, ob architekturspezifische Eigenschaften in ARM- und FPGA-Architekturen [5] zu maschinellen Lernalgorithmen mit besonders niedrigem Taktzyklusverbrauch führen können.

Mit der im Projekt LavA (Olaf Spinczyk) entwickelten Entwurfsmethodik für *On-Chip*-Multiprozessorsysteme (MPSoCs) sollen die Datenstromalgorithmen aus Arbeitspaket 3 auf FPGA-Systeme abgebildet und untersucht werden. Neben der Synthese heterogener MPSoCs (CPU-Kerne und spezifische Beschleunigerkomponenten) erlaubt das LavA-Framework auch die Erzeugung von *Many-Core*-Systemen auf FPGAs mit mehr als 100 homogenen CPU-Kernen und *On-Chip*-Verbindungsnetzwerk. Die so erzeugten Prototypen künftiger eingebetteter *Many-Core*-Systeme, schlagen die Brücke zwischen ubiquitären Systemen und verteilten Berechnungsarchitekturen wie MapReduce, Gather-Apply-Scatter [36] sowie der Lambda-Architektur [37], bei der Modelle sowohl auf statischen Daten, als auch auf realzeitlichen Datenströmen gelernt und angewendet werden. Hier soll nun untersucht werden, welchen Einfluss die Ausführung der Datenstromalgorithmen auf den verschiedenen Architekturen auf den Ressourcenverbrauch hat oder wie die Algorithmen für verschiedene Ausführungen verändert werden müssen, um zu einer Verringerung der Ressourcen zu führen.

Arbeitspaket 5. Umfassende Systemanpassung: Das Ziel, Systemsoftware durch Analyse von Nutzungsmustern, Kontext und Systemzustand zu verbessern, haben wir bisher punktuell anhand von Fallstudien verfolgt. Nun wollen wir Datenerhebung, Lernen und Adaption als integralen Bestandteil der Systemsoftware betrachten. Dafür wird eine zentrale Systemkomponente konzipiert, die entsprechende Dienste bereitstellt. Durch die Zentralisierung wird es möglich, Synergien auszunutzen. Andere Systemkomponenten, die diese Dienste in Anspruch nehmen, können dies in unterschiedlicher Weise tun. Zwei mögliche Ansätze sollen verfolgt werden.

Zunächst untersuchen wir wie bisher Systemkomponenten, welche auf bestimmte Vorhersagen oder Klassifikationen angewiesen sind, für welche ansonsten Heuristiken verwendet würden. Was gelernt wird, kann dabei, wie beim *Kobe*-System [12] ein expliziter Auftrag an den zentralen Dienst sein. Anstatt Trainingsdaten vorzugeben, können Samplingverfahren angegeben werden. Wir untersuchen dabei schrittweise, wie das Wissen über maschinelles Lernen aus den Subsystemen in eine zentrale Systemkomponente verlagert werden kann. Für die Parameteroptimierung für Lernaufgaben gibt es bereits in RapidMiner und bei Kobe Methoden, die auch in A3 betrachtet werden. Wir wollen dabei zur Betrachtung globaler Zielfunktionen, wie z.B. der Batterieentladung, übergehen.

Der zweite Ansatz basiert unter anderem auf Arbeiten zu merkmalsorientierter selbstadaptierender Software [17]. Gelernt wird dabei eine lineare Beziehung zwischen Merkmalkombinationen und entsprechenden Änderungen der Performanzmetriken. Um Änderungen dieser Metriken zu bewirken, wird anhand dieser Beziehung eine optimale Auswahl der Konfigurationsmerkmale ermittelt, wobei Merkmale, die keine relevante Änderung bewirken, nicht berücksichtigt werden. Hierbei sind sich die Systemkomponenten nicht darüber bewusst, dass gelernt wird. Stattdessen erfolgt die Systemverbesserung von außen durch gelerntes Wissen über die optimale Komponentenkonfiguration. Während [17] ausschließlich die Konfiguration als ursächlich für die Performanz, und kontextabhängige Einflüsse als *concept drift* betrachtet werden, wollen wir diese explizit einschließen. Ein möglicher Ansatz wäre dabei die Betrachtung des Kontextes als fixen Teil der Konfiguration. Da wir es somit mit einer größeren Zahl an auch nicht-binären Merkmalen zu tun haben, müssen wir Alternativen zur beschriebenen linearen Beziehung untersuchen. Diesen Ansatz wollen wir auch für die Konfiguration von Infrastruktur-Softwareproduktlinien für sehr kleine eingebettete Systeme verfolgen. Bei diesen Systemen geht es weniger um Adaptivität als um eine optimale statische Maßschneiderung der Systemsoftware. Als Fallstudie verwenden wir dabei das im Teilprojekt A4 entwickelte energiegewahre Betriebssystem und dessen Netzwerkprotokoll-Stapel.

Bei beiden Ansätzen betrachten wir natürlich die Möglichkeit der Auslagerung sowohl des Lernens als auch der Berechnungen auf den gelernten Modellen auf externe Hardware, um Ressourcen zu sparen, sofern die dabei entstehenden Latenzen toleriert werden können. Fallstudienartig soll wieder gezeigt werden, wie die Infrastruktur hilft, auf Heuristiken zu verzichten und den Ressourcenverbrauch zu verbessern. Neben der lokalen Datenanalyse soll weiter erforscht werden, wie kollektiv im Netz gesammelte Daten zur Systemverbesserung beitragen können.

Arbeitspaket 6. Benchmarking, Evaluation und Simulation: Seit 1987 stellt die Datensammlung *UCI Machine Learning Repository*, eingerichtet von David Aha, unterstützt von der National Science Foundation der USA, Datensätze für Klassifikation, Regression, Clustering und einige relationale Lernprobleme zur Verfügung. Die Datensätze sind gut aufbereitet und in Hinblick auf bestimmte Eigenschaften maßgeschneidert. Sie sind der Standard, an dem bis heute die meisten neuen Algorithmen getestet werden, um sie vorzustellen. Dabei gibt es durchaus Kritik an dieser Sammlung. Oft werden die Datensätze als „realweltlich" bezeichnet, weil sie von einer realen Anwendung abstammen. Sie sind aber keinesfalls so, wie man Daten in realen Anwendungen vorfindet. Daher wird der Aufwand der Vorverarbeitung leicht unterschätzt, weil sie bei den publizierten Datensätzen bereits stattgefunden hat. Sie sind bereits für ein bestimmtes Lernproblem aufbereitet, obwohl der ursprüngliche Datensatz die Untersuchung unterschiedlicher Lernprobleme erlaubt hätte. Sie illustrieren auf eine bestimmte Weise Phänomene, die sich auch anders ausprägen könnten. Bleibt man bei den Datensätzen der UCI, verengt sich der Blick auf die Phänomene. Auf Konferenzen wird deshalb schon davon gesprochen, dass sich das Fachgebiet so stark an diesen Datensätzen ausrichtet, dass man auf wissenschaftstheoretischer Ebene von *overfitting* sprechen kann.

In diesem Arbeitspaket wollen wir bei der Evaluation unserer Methoden weitergehen, indem wir das im AP1 entwickelte Werkzeug zur Erzeugung tatsächlich realweltlicher Daten bereit stellen. Dabei sollen diese Daten für unterschiedliche Lernprobleme und Vorverarbeitungsschritte nutzbar sein. Zum anderen wollen wir den Ansatz der Datengeneratoren verfolgen. Datengeneratoren haben den Vorteil, dass sie eine sehr kompakte Darstellung besitzen und sich für unterschiedliche Untersuchungsziele einstellen lassen.

Für graphische Modelle sind diese Ansätze nicht geeignet. Während die Lernprobleme, für die das UCI Repository oder die meisten Datengeneratoren Testfälle liefern, durch eine einzige Fragestellung mit einem wahren Wert je Beispiel charakterisiert sind, brauchen die graphischen Modelle erst einmal Fragestellungen, zu denen der wahre Wert gesucht wird. Als probabilistische Modelle liefern sie obendrein nicht nur einen Wert, sondern dazu eine Wahrscheinlichkeit. Der einfache Abgleich wie bei Testdatensätzen sonst üblich ist hier also nicht möglich. Perturbierte MAP-Vorhersagen können als abgekürzte Markovketten Monte-Carlo Simulation interpretiert werden [22]. Es wird untersucht, inwiefern solche Methoden des maschinellen Lernens für eine simulationsbasierte Evaluation einsetzbar sind. Eine realitätsnahe Simulation von echtem Netzwerkverkehr wäre zur Evaluation unserer aktuellen Arbeiten im Bereich der Funkkommunikation von großem Interesse. Schließlich ist einem Datensatz üblicherweise ein Gütekriterium im Sinne von wahren Zielwerten beigegeben. Der Ressourcenverbrauch ist aber nicht dazu angegeben. Langfristig sollten Datengeneratoren aber auch Kostenmodelle für Ressourcen berücksichtigen können, so dass auch der Trade-off zwischen Güte des Modells und Ressourcenverbrauch mit generierten Daten systematisch untersucht werden kann.

Da die Systemanpassung indirekt wieder das Nutzerverhalten beeinflussen kann, sind Simulationsmodelle notwendig, die auch auf externe Ereignisse (z.B. Netzwerkverkehr) reagieren und miteinander interagieren können. Dafür reicht die (isolierte) Generierung von Daten nicht aus. Die erstellten Modelle setzen wir dabei nicht nur in Simulationswerkzeugen (z.B. MobiSIM) ein, sondern untersuchen auch die Umsetzung im Gerät, z.B. durch die Injektion von Ereignissen an der Nutzerschnittstelle oder durch dedizierte Benchmarks.

Arbeitspaket 7. Bereitstellung von Algorithmen: Die im Teilprojekt neu entwickelten Algorithmen sollen in eine solche Form gebracht werden, dass sie in bekannte Werkzeuge integriert sind, so dass sie inner- und außerhalb des SFB leicht verwendet werden können. Außerdem werden weiterhin Promovenden aus anderen Teilprojekten bezüglich der Datenanalyse beraten und in der Nutzung vorhandener Methoden unterstützt.

Zeitplan

Arbeitspaket / Quartal	2015				2016				2017				2018			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Ressourcenschonende Datenerhebung:	▬	▬	▬	▬	▬	▬										
2. Datenströme:					▬	▬	▬	▬								
3. Graphische Modelle unter Ressourcenbeschränkung:	▬	▬	▬	▬												
4. Architekturspezifische Lernverfahren:									▬	▬	▬	▬				
5. Umfassende Systemanpassung:							▬	▬	▬	▬	▬	▬				
6. Benchmarking, Evaluation und Simulation:													▬	▬	▬	▬
7. Bereitstellung von Algorithmen:	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬	▬

3.5 Stellung innerhalb des Sonderforschungsbereichs

Das Teilprojekt A1 nimmt eine zentrale Stellung im Sonderforschungsbereich ein. Es untersucht Methoden der Analyse von sehr großen, strömenden, heterogenen Daten, die Realzeit-Antworten erfordern, in Bezug auf ihre Eignung, den Ressourcenverbrauch von ubiquitären Systemen zu verringern und selbst so wenig Ressourcen zu verbrauchen, dass sie auf diesen Architekturen ablaufen können. Insofern arbeitet das Projekt sowohl im Arbeitskreis *Datenanalyse* mit den anderen Projekten eng zusammen, die an Lernverfahren arbeiten, beispielsweise mit A2 und in der nächsten Periode mit A6, als auch im Arbeitskreis *Ressourcen* mit den Projekten, die sich um den Ressourcenbedarf und die Plattformen kümmern, beispielsweise mit A3 und A4. Im Einzelnen gibt es konkrete Themen für Zusammenarbeiten wie etwa die Untersuchung des Energieverbrauchs von Datenbankanfragen und bestimmten Datenanalysealgorithmen, die im TP A2 untersucht werden. Auch A1 untersucht den Energiebedarf bestimmter Algorithmen für Verfahrensklassen, so dass wir gemeinsam zu einem breiten Überblick gelangen können. Bezüglich der Auslagerung von Analysen in die Cloud kooperieren wir mit Teilprojekt A3, welches das entsprechende Problem der Ablaufplanung behandelt. Die Betrachtung der Betriebssysteme ubiquitärer Systeme führt A1 mit Teilprojekt A4 bezüglich der automatischen Konfiguration von Softwareproduktlinien zusammen. Bei uns wird die Beziehung von Kontext, Konfiguration und Ressourcenverbauch top-down gelernt, wobei zunächst ein nutzerindividuelles Modell entsteht. Aus den Daten mehrerer Nutzer können dabei allgemeine Modelle abgeleitet werden. A4 geht umgekehrt vor und modelliert bottom-up, wobei allgemeine Plattformmodelle die Basis für die Modellierung des individuellen Ressourcenverbrauchs dienen. Die in A4 entwickelten Energiemodelle werden von uns für die energiesparende Sensorik als auch für die Evaluation (AP 6) verwendet.

Aber auch über den A-Bereich hinaus gibt es erfolgreiche Zusammenarbeiten, die fortgesetzt werden. So arbeiten wir mit C1 an Algorithmen für den vielen Lernmethoden inhärenten Optimierungsschritt zusammen. Eine Verbindung zu Teilprojekt B4 besteht bezüglich der energieeffizienten Datenerhebung unter Berücksichtigung von Qualitätsparametern wie Genauigkeit und Latenz.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

Vista-TV

(Morik,Förderkennzeichen EU STREP 296126) (Förderzeitraum: 2012 - 2014)

Aufgabe des Projektes ist die Verbindung bestehender Verfahren zu einem neuartigen System für Linked Open Data. Das Projekt annotiert Internetfernsehprogramme automatisch und empfiehlt Benutzern darauf aufbauend Sendungen. Das in A1 entwickelte *Streams-Framework* wurde erfolgreich im Projekt angewandt, um aus Videodaten Merkmale zu extrahieren.

Insight

(Morik,Förderkennzeichen EU 318225) (Förderzeitraum: 2012 - 2015)

Das Projekt behandelt Big Data in Bezug auf Volumen, Heterogenität und Realzeitanforderung. Wir arbeiten an raumzeitlichen Prozessen zur Vorhersage und Steuerung der Verkehrsflüsse in Dublin. Aus diesem Projekt wollen wir die Daten für das Teilprojekt B4 nehmen. Das Streams-Framework wurde hier zur Ereigniserkennung eingesetzt.

Kobra

(Morik,Förderkennzeichen BMBF 01UG12445A) (Förderzeitraum: 2012 - 2015)

In diesem Projekt geht es um einen möglichst umfassenden Korpus des Deutschen und dessen Analyse. Ein Zusammenhang mit den Arbeiten in A1 kann bei den *Topic Models* entstehen.

Intelligente Prozess-Steuerung auf Basis von Prozessmerkmalen

(Morik,Förderkennzeichen SMS Siemag) (Förderzeitraum: 2011 - 2014)

Das Projekt wird direkt von SMS Siemag gefördert. Es wurde eine Anwendung des maschinellen Lernens für die Bestimmung des Prozessendes bei einem BOF im Stahlwerk Dillinger Hütte sowie eine Infrastruktur für das „Stahlwerk im Büro“ entwickelt. Hierbei wurde auch das Streams-Frameworks eingesetzt.

Data Driven Material Development

(Morik,Förderkennzeichen Mercator Stiftung, Merkur Pr-2012-0023) (Förderzeitraum: 2012 - 2014)

Die Daten zur Untersuchung verschiedener Mischungen chemischer Substanzen in Bezug auf bestimmte Materialeigenschaften sollen effizienter genutzt werden, so dass aus wenigen Messpunkten die Eigenschaften an allen möglichen Messpunkten vorhergesagt werden können. Dies hat keinen direkten Bezug zum SFB 876.

LavA

(Spinczyk,Förderkennzeichen DFG, SP 968/4-2) (Förderzeitraum: 2013 - 2015)

Durchgängige anwendungsspezifische Maßschneiderung von Betriebssystem und Rechnerstruktur in konfigurierbaren eingebetteten Systemen. Im Arbeitspaket „Umfassende Systemanpassung“ wollen wir die Erkenntnisse und Werkzeuge aus dem Projekt LavA nutzen. Datenanalyse spielt bei LavA keine Rolle.

DanceOS

(Spinczyk,Förderkennzeichen DFG, SP 968/5-2) (Förderzeitraum: 2013 - 2014)

Erforschung neuer softwaretechnischer Ansätze zur Realisierung softwarebasierter Fehlertoleranzmaßnahmen mit Hilfe so genannter Aspekte. Fehlertoleranz ist für den vorliegenden Antrag nicht relevant.

Fehlertolerante und echtzeitfähige System- und Kommunikationsinfrastrukturen für die Schutz- und Leittechnik

(Spinczyk,Förderkennzeichen DFG, SP 968/6-1) (Förderzeitraum: 2012 - 2014)

Konzeption einer auf Virtualisierung basierenden Systemarchitektur für Cyber-Physical Systems. Es gibt keine Überschneidungen zu dem beantragten Vorhaben.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	117.600	—	117.600	—	117.600	—	117.600
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	117.600		117.600		117.600		117.600	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (A1)							
Wissen-schaftliches Personal	1	Morik, Katharina, Prof. Dr., Prof	Data Mi-ning/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	4,0	—	Grund-ausstattung
	2	Spinczyk, Olaf, Prof. Dr., Prof	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	4,0	—	Grund-ausstattung
	3	Duivesteijn, Wouter, Dr., PostDoc	Data Mi-ning/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	29,85	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	4	Greve, Andreas, Techn. Ang.	—	LS 8, Fak. Inf, TU Do	9,95	—	Grund-ausstattung
	5	Klingspor, Sabine, Verw. Ang.	—	LS 8, Fak. Inf, TU Do	4,97	—	Grund-ausstattung
Beantragte Personalmittel (A1)							
Wissen-schaftliches Personal	6	Piatkowski, Nico, Dipl.-Inf. WiMi	Data Mi-ning/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	—	Doktorand/in	—
	7	Streicher, Jochen, Dipl.-Inf. WiMi	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Morik, Katharina**
   Projektleitung. Schwerpunkt Algorithmen maschinellen Lernens. Mitarbeit in den Arbeitspaketen 2,3,4 und 6. Beratung anderer Teilprojekte bezüglich maschinellen Lernens, Kooperationen und Darstellung des SFB im nationalen und internationalen Rahmen.

2. **Spinczyk, Olaf**
   Projektleitung. Schwerpunkt ubiquitäre Systemsoftware. Mitarbeit in den Arbeitspaketen 1,4,5 und 6.

3. **Duivesteijn, Wouter**
   Datenströme (AP3) sowie Umfassende Systemanpassung (AP5) mit dem Schwerpunkt Exceptional Model Mining. Mitarbeit in AP1 und AP6.

4. **Greve, Andreas**
   Unterstützung der technischen Infrastruktur zum Datenaustausch sowie bei der Installation von Anwendungsprogrammen.

5. **Klingspor, Sabine**
   Dokument- und Terminverwaltung sowie ggf. Reiseplanung und- abrechnung.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

6. **Piatkowski, Nico**
   Graphische Modelle unter Ressourcenbeschränkung (AP2), Datenstromalgorithmen (AP3) sowie Architekturspezifische Lernverfahren (AP4). Insbesondere ressourceneffiziente Algorithmen für probabilistische Inferenz und Parameterschätzung. Benchmarking, Evaluation und Simulation (AP6) für probabilistische Modelle.

7. **Streicher, Jochen**
   Ressourcenschonende Datenerhebung (AP1) sowie umfassende Systemanpassung (AP5). Insbesondere die Integration von Lernkomponenten in die Systemsoftware. Benchmarking, Evaluation und Simulation (AP6) von und auf ubiquitären Systemen.

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	7.000	7.000	7.000	7.000
Summe der vorhandenen Mittel	7.000	7.000	7.000	7.000
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt A2

3.1.1 Titel:

Algorithmik von Lernverfahren in eingebetteten Systemen

3.1.2 Fachklassifizierung:

Algorithmische Komplexitätstheorie, Informationssysteme

3.1.3 Projektleitung

Prof. Dr. Sohler, Christian, 19.02.1973

LS 2, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 14
44227 Dortmund

Tel.: 0231-755-6940
Fax : 0231-755-2047
E-Mail: `christian.sohler@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Teubner, Jens, 22.03.1976

LS 6, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 20
44227 Dortmund

Tel.: 0231-755-6481
Fax : 0231-755-2405
E-Mail: `jens.teubner@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Ziel dieses Projekts ist die Entwicklung von Entwurfsparadigmen für Lernalgorithmen in eingebetteten Systemen. Dazu wollen wir einzelne wichtige Merkmale eingebetteter Systeme identifizieren, diese theoretisch modellieren, in den entwickelten Modellen Algorithmen entwerfen und analysieren und dann empirisch die Performanz dieser Algorithmen auswerten, um Rückschlüsse auf die Effizienz der Algorithmen und die Aussagekraft der Modelle ziehen zu können. Diese Rückschlüsse sollen, wenn nötig, wieder in die Modellierung einfließen. Aus den auf diese Weise erlangten Erkenntnissen wollen wir dann allgemeine algorithmische Entwurfsparadigmen ableiten.

Die Aspekte eingebetteter Systeme, die wir in der zweiten Antragsphase betrachten wollen, sind *Energiebedarf, Kommunikations- und Rechnerarchitektur, Datenströme* und *verteilte Systeme.* Je nach Stand der Forschung in den einzelnen Bereichen werden wir uns zunächst der Entwicklung von Rechenmodellen oder sofort der Entwicklung neuer Algorithmen und Entwurfsmethoden widmen.

3.3 Bisherige Entwicklung des Teilprojekts

In der ersten Förderphase hat Projektleiter Jan Vahrenhold einen Ruf an die Universität Münster erhalten und hat diesen zum 01.03.2012 angenommen. Da seine Expertise insbesondere in der Modellierung und bei den energieeffizienten Lernalgorithmen fehlte, haben wir uns dazu entschlossen, diese Bereiche erst in der zweiten Projektphase intensiver anzugehen und uns in der ersten Förderphase auf die Themenbereiche *verteilte Lernalgorithmen, Datenstromalgorithmen* und *Empirische Analyse* zu konzentrieren. In der zweiten Projektphase soll nun Jens Teubner Projektleiter werden. Er bringt insbesondere Wissen zur hardwarenahen Implementierung von Algorithmen und zur Architektur moderner Rechensysteme in das Projekt ein. Im nachfolgenden Bericht wird die "wir"-Form sowohl bei der Beschreibung von Ergebnissen von Jan Vahrenhold und Christian Sohler als auch von Jens Teubner benutzt.

3.3.1 Bericht und Stand der Forschung

Wir beginnen damit, zwei für die neue Projektphase relevante Modelle vorzustellen, die parallele bzw. verteilte Systeme modellieren. Das erste Modell sind die sogenannten Nimble Algorithms (flinke Algorithmen) [74]. Diese modellieren verteilte Algorithmen, die auf s Servern arbeiten. Jeder Server hat zur Lösung eines Problems polynomiell viel Speicher und Laufzeit zur Verfügung und die Daten sind beliebig auf den Servern verteilt. Bewertet wird der Algorithmus durch die Anzahl der Bits, die für den Kommunikationsaustausch zwischen den Servern notwendig sind.

Beim zweiten Modell handelt es sich um das Multi-BSP-Modell [89], das ein mehrstufiges Rechnersystem beschreibt, wobei in jeder Stufe mehrere Komponenten parallel arbeiten und es in jeder Stufe unterschiedlich viel Speicherplatz, unterschiedliche Kommunikations- und Synchronisationskosten zu der darunter liegenden Stufe gibt. Algorithmen werden nun nach Laufzeit, den Kommunikations- und den Synchronisationskosten bewertet. Es ist möglich, Algorithmen zu entwerfen, die ohne Kenntnisse der Architekturparameter bis auf konstante Faktoren bezüglich aller drei Bewertungsmaße optimales Verhalten haben. Solche Algorithmen heißen *parameterunabhängig.* Bekanntes Beispiel für ein Modell parameterunabhängiger Algorithmen sind die *cache-oblivious* Algorithmen [68]; damit können Algorithmen entworfen werden, die effizient sind für jede mögliche Cachekonfiguration. Beide Modelle dienen uns als Startpunkt unserer Modellierungsaufgaben.

Im Folgenden werden wir auf die in der ersten Projektphase erzielten Ergebnisse eingehen. Dabei gliedern wir nach den Arbeitspaketen aus der ersten Antragsphase und gehen auch auf die Weiterentwicklung des Stands der Forschung während der vergangenen Projektphase ein.

Modellierung: Ein vereinendes Charakteristikum von eingebetteten Systemen ist die Ressourcenbeschränkung. Die genaue Ausprägung der Ressourcen ist dabei vielfältig, als Beispiele lassen sich Energie, Speicher, Bandbreite oder Laufzeit des eingebetteten Systems betrachten. In einem verteilten System konkurrieren die einzelnen Teilnehmer unter Umständen um diese Ressourcen. Dabei handeln die Nutzer des Systems häufig eigennützig (z.B. werden Geräte bei zu geringer zur Verfügung stehender Bandbreite abgeschaltet, um einen Reset der Verbindung durchzuführen). Ein solches Szenario kann durch die algorithmische Spieltheorie modelliert werden. Jeder der n Spieler entspricht dabei einem kleinen Gerät und hat das Ziel, seine Berechnungen auf den zur Verfügung stehenden Ressourcen mit maximalem Profit (Geschwindigkeit, Energievebrauch, Speicherplatz, Bandbreite, etc.) durchzuführen. Zunächst ließ sich die Möglichkeit einer Ressourcenallokation untersuchen, bei der kein Spieler seinen Profit verbessern kann. Allerdings ist es bereits sehr schwierig

zu entscheiden, ob eine solche Allokation existiert. Wir haben die Frage nach einer Ressourcenallokation gestellt, bei der kein Spieler seinen Profit um mehr als einen vorgegebenen Faktor β verbessern kann. Es ist uns gelungen für ein konstantes β in polynomieller Zeit eine Allokation erstmals zu berechnen [A2/93].

Ferner haben wir dieses Problem auch in einem dynamischen Fall untersucht. Hier wählt jeder Spieler alle β Schritte seine Ressourcen in Abhängigkeit von den übrigen Spielern, wobei $\beta \in O(1)$. Wir konnten zeigen, dass nach einer Gesamtzahl von $T \cdot \lceil \log \log n \rceil$ Schritten eine Allokation berechnet wird, bei der sich kein Spieler um mehr als einen Faktor $O(\beta)$ verbessert [A2/95].

Energieeffiziente Lernverfahren: Ein aus theoretischer Sicht nicht gut quantifizierbarer und deswegen häufig vernachlässigter Faktor bei der Beurteilung eines Algorithmus ist dessen Energieverbrauch. Es ist möglich, für die Durchführung eines Programms den Energieverbrauch auf Systemebene zu minimieren, indem beispielsweise Rechenkomponenten nach Bedarf hoch und runter gefahren werden (siehe auch Teilprojekt A3). Eine Übersicht dieser Ansätze findet sich im Artikel von Albers [59]. Wir sind hingegen an Entwurfsmustern für Algorithmen interessiert, die bereits energieeffiziente Eigenschaften haben. Der Energieverbrauch eines Algorithmus wächst unter der Annahme von konstantem Energieverbrauch bei elementaren Operationen proportional mit der Laufzeit, allerdings trifft diese Annahme für keine in der Praxis eingesetzte Rechenmaschine zu. Bei einem herkömmlichen Rechner, von dem wir vereinfacht annehmen, dass Daten sich entweder im Hauptspeicher oder auf der Festplatte befinden, führt jeder Zugriff auf die Festplatte zu einem erhöhten Energieverbrauch. Um die Festplattenzugriffe zu minimieren, verfolgten wir in der ersten Phase zwei Zeile:

- *In-place*-Algorithmen, für deren Durchführung wenig Platz im Hauptspeicher erforderlich ist.
- *Cacheunabhängige* Algorithmen (*cache-oblivious algorithms*), die die Anzahl der Zugriffe auf den langsameren Speicher reduzieren, ohne dabei von der Größe des Hauptspeichers und der Transferkapazität eines Festplattenzugriffs abhängig zu sein.

Wir merken an, dass beide Modelle sich auch für Hierarchien mit mehr als nur zwei Arten von Speichern anwenden lassen.

Die Berechnung der k Eckpunkte der konvexen Hülle einer gegebenen n-elementigen Punktmenge stellt eine zentrale Aufgabe in der algorithmischen Geometrie beziehungsweise der geometriebasierten Datenanalyse dar. Wir konnten in [A2/103] nachweisen, dass es möglich ist den Algorithmus von Kirkpatrick und Seidel [76] mit konstantem zusätzlichem Speicheraufwand so zu implementieren, dass alle weiteren theoretischen Eigenschaften erhalten bleiben. Dieser Algorithmus ist somit der einzige bekannte In-Place-Algorithmus, der neben einer ausgabesensitiven Laufzeit von $O(n \log k)$ auch Instanzoptimalität aufweist, also für jede Eingabe nur konstant mehr Speicherplatz nutzt, ohne dabei Randomisierung nutzen zu müssen.

Eine ϵ-semiseparierte Paardekomposition einer Menge P von n Punkten im d-dimensionalen euklidischen Raum sind Teilmengen $A_i, B_i \subset P$, wobei für jedes i die kleinsten umschließenden Bälle $D(A_i)$ von A_i und $D(B_i)$ von B_i mindestens Abstand $\epsilon \cdot \min(\text{radius } (D(A_i)), \text{radius } (D(B_i)))$ für einen Parameter $s > 1$ haben und für jedes Paar an Punkten $p, q \in P$ es genau ein i gibt mit $p \in A_i$ und $q \in B_i$ oder umgekehrt. Eine solche Dekomposition ist oft Grundlage für weitere Algorithmen zur Lösung von Abstandsproblemen. Es ist uns gelungen, eine solche Dekomposition cacheunabhängig mit erwartet $O(\text{sort}(\epsilon^{-d} n) \log_2 n)$ Speichertransfers durchzuführen [A2/101], wobei $\text{sort}(x)$ die Anzahl der benötigten Speichertransfers zum Sortieren von x Zahlen ist.

Verteilte Lernverfahren: Wir haben uns im Bereich verteilte Lernverfahren schwerpunktmäßig mit der Entwicklung von kernmengenbasierten Clustering-Verfahren beschäftigt. Unter zentrumsbasiertem Clustering fassen wir Algorithmen zusammen, die für eine gegebene Menge an Punkten X im d-dimensionalen euklidischen Raum k Zentren C finden, sodass die Zielfunktion

$$Z(X, C) := \sum_{x \in X} \min_{c \in C} d(x, c)$$

minimiert wird. Mit $d(x, c)$ bezeichnen wir hier je nach Problem den euklidischen oder quadrierten euklidischen Abstand zweier Punkte. Die zugehörigen Probleme heißen *k-Median-Problem* und *k-Means-Problem*. Ist die Anzahl der Clusterzentren nicht auf k beschränkt sondern beliebig, wobei jedes benutzte Clusterzentrum einen uniformen Kostenbeitrag f zur Zielfunktion hat, so sprechen wir vom *uniformen Facility-Location-Problem*. Letzteres kann man auch als Clustering mit einem Regularisierungsterm auffassen, der die Anzahl der Zentren kontrolliert. Ferner müssen die Zentren nicht aus Punkten im Raum bestehen, sondern können ihrerseits ebenfalls Unterräume der Dimension j sein. Im speziellen Fall, dass als Zentrum

genau ein linearer Unterraum in Frage kommt und die Daten so normiert sind, dass der Erwartungswert 0 ist, entspricht das Problem der *Hauptkomponentenanalyse* (*PCA* - principle component analysis).

Ein effektiver Einsatz von Clusteringalgorithmen ist besonders relevant für große Datenmengen. Hier sind nicht nur die Qualität der produzierten Lösung und die Rechenzeit wichtig, sondern insbesondere der Speicherbedarf eines Algorithmus, der den Arbeitsspeicher von aktuellen Rechnern möglicherweise übersteigt. Tritt dieser Fall ein, so stehen zwei Alternativen zur Auswahl:

1. Parallele Verarbeitung von Daten auf mehreren Rechnern
2. Online vorgenommene Kompression eines Datenstroms auf einem Rechner

Algorithmen zur parallelen Verarbeitung von Daten finden in vielen Gebieten Anwendung. So können die Daten bereits verteilt vorliegen, wie es beispielsweise in manchen eingebetteten Systemen wie Sensornetzwerken der Fall ist. Die Daten können dann, wenn sie in nicht allzu großer Anzahl vorliegen, einzeln zu einem zentralen Server geschickt werden, auf dem die Analyse vorgenommen wird, aber für große Datenmengen ist der Kommunikationsaufwand nicht praktikabel.

Sowohl bei der parallelen Verarbeitung wie bei einer online vorgenommenen Kompression ist ein Grundbaustein von Algorithmen eine kompakte Repräsentation eines Datensatzes. Für den Repräsentationsgrad nutzen wir einen Parameter $\epsilon > 0$. Es gibt eine Vielzahl von theoretischen und praktischen Aggregationsverfahren mit einer ebenso großen Anzahl an Eigenschaften. Im Rahmen dieses Teilprojekts wurden *Kernmengen* untersucht, welche aus theoretischer Sicht besonders starke Repräsentationseigenschaften besitzen. Allgemein lassen sich Kernmengen als kleine Punktmengen verstehen, die sich bezogen auf das untersuchte Problem nahezu identisch zur Ausgangspunktmenge verhalten. Speziell für Clustering ist eine Kernmenge eine (möglicherweise gewichtete) Punktmenge K, sodass für jede Menge an k Zentren C gilt

$$|Z(K,C) - Z(X,C)| \leq \epsilon \cdot Z(X,C).$$

Kernmengen, die die obige Definition erfüllen, sind im Allgemeinen unter Vereinigung abgeschlossen; haben wir zwei Kernmengen K_1 und K_2, die die disjunkten Punktmengen P_1 und P_2 repräsentieren, dann ist $K = K_1 \cup K_2$ eine Kernmenge für $P_1 \cup P_2$. Da gewichtete Punkte sich auch als mehrfache Punkte interpretieren lassen, kann man die Kernmenge einer Kernmenge K für P bilden und erhält eine Kernmenge K', die ebenfalls eine Kernmenge für P ist, aber im Allgemeinen eine geringere Güte besitzt. Kernmengen eignen sich sehr gut, um Lernverfahren in verteilten Systemen zu implementieren, bei denen die Daten auf mehrere Rechner verteilt sind. Man rechnet zunächst für jeden Rechner lokal eine Kernmenge aus und sammelt dann alle Kernmengen auf einem zentralen Rechner. Da die Kernmengen klein sind, verursacht dies relativ wenig Kommunikationsaufwand. Ist der Kommunikationsaufwand trotzdem zu hoch, weil alle Kernmengen zum selben Rechner geschickt werden, so kann man diesen reduzieren, indem man die Kernmengen zunächst lokal vereinigt und für jede Vereinigung von Kernmengen eine neue Kernmenge berechnet. Auf diese Weise kann man die Daten mit einem baumartig verzweigten Prozess an einem Rechner sammeln und den Kommunikationsaufwand pro Rechner auf eine konstante Anzahl von Kernmengen reduzieren. Daneben kann man Kernmengen, die die oben erwähnten Eigenschaften erfüllen, mit Hilfe des sogenannten Merge-and-Reduce-Prinzips für die Entwicklung von Datenstromalgorithmen einsetzen.

Für k-Median und k-Means kann die Existenz von Kernmengen mit $O(d^2 \cdot k^3/\epsilon^2)$ Punkten garantiert werden [78]. Feldman und Langberg zeigen sogar $O(k \cdot d/\epsilon^2)$ für das k-Median-Problem [66]. Für allgemeine Clusteringprobleme mit k Zentren, die aus j-dimensionalen Unterräumen bestehen, existieren Kernmengenkonstruktionen der Größe $O(kjd \log^{2f(j,k)}(dn)/\epsilon^2)$, wenn die Eingabe aus Punkten auf ganzzahligen, polynomiell großen Koordinaten besteht [90]. Im Allgemeinen sind Kernmengen mit deutlich geringerem Speicherbedarf als die Größe der Ausgangsmenge für dieses Problem jedoch nicht möglich [71].

In [A2/96] wurden Kernmengen mit reduzierter Varianz eingeführt: Eine Kernmenge ist hier ein Paar aus einer gewichteten Punktmenge K und einem Wert Δ, sodass für jede Wahl einer Menge C von k Zentren gilt:

$$|Z(K,C) + \Delta - Z(X,C)| \leq \epsilon \cdot Z(X,C).$$

Letztendlich bedeutet dies, dass ein Clustering der Kernmenge immer um einen Term Δ geringere Kosten aufweist. Dies hat schon für sich einige positive Effekte. Wenn wir z.B. einen $(1+\epsilon)$-Approximationsalgorithmus auf der Kernmenge einsetzen, so ist der maximale Fehler, den dieser Algorithmus macht, um ca. $\epsilon\Delta$ kleiner als bei der alten Kernmengendefinition. Unsere Hauptmotivation war jedoch, dass wir für diese Kernmengendefinition zeigen konnten, dass es Kernmengen für k-Means Clustering gibt, deren Größe unabhängig von d und n ist. Dazu haben wir gezeigt, dass, wenn man die Eingabepunktmenge durch ihre Projektion auf die ersten $O(k/\epsilon^2)$ Hauptkomponenten ersetzt, die Kosten im Sinne der zweiten Kernmengendefinition

beschränkt werden können. Dies ist zunächst nur eine Dimensionsreduktion, aber daraus folgt, dass eine optimale k-Means Lösung für die projizierten Punkte eine $(1+\epsilon)$-Approximation für die Ausgangsmenge ist. Nun kann man existierende Kernmengenkonstruktionen für die projizierten Punkte verwenden und erhält eine Kernmenge der Größe $(k/\epsilon)^{O(1)}$. Ähnlich kann man auch für das Clusteringproblem für quadrierte euklidische Distanzen mit k j-dimensionalen Unterräumen vorgehen und unter der Voraussetzung, dass die Eingabepunkte ganzzahlig sind mit maximaler Koordinate $n^{O(1)}$, erhält man ebenfalls Kernmengen, deren Größe nur logarithmisch von n und d abhängt[1].

Verfahren für die Approximation einer Matrix $A \in \mathbb{R}^{n,d}$ vom Rang $\min(n,d)$ durch eine Matrix $A_j \in \mathbb{R}^{n,d}$ von Rang j wurden durch Clarkson und Woodruff [64] untersucht, die eine untere Schranke der Größe $\Omega(j/\epsilon \cdot (n+d)\log(nd))$ für die Speicherung von A_j bewiesen haben (hier ist die Schranke in Bytes, während wir bei der Kernmengengröße immer die Anzahl der Punkte angeben). Dieses Problem ist verwandt mit der Unterraumapproximation und der Hauptkomponentenanalyse, allerdings ist die Ausgabegröße bei letzterem Problem geringer und es gibt Algorithmen mit deutlich geringerem Speicherverbrauch. Für die Hauptkomponentenanalyse wurde in [A2/96] eine Kernmengenkonstruktion der Größe $O(j/\epsilon \cdot \log(nd))$ angegeben. Ferner ist die Kernmengenkonstruktion deterministisch und benötigt im Gegensatz zu [64] keinen zusätzlichen Speicher, um die Fehlerwahrscheinlichkeit zu beschränken.

Die meisten der von uns entwickelten Kernmengenverfahren lassen sich auch für die Entwicklung von Datenstromalgorithmen einsetzen.

Datenstromalgorithmen: Es gibt drei gängige Ansätze zur Lösung von Clusteringproblemen im Datenstrom. Im ersten Ansatz werden die Zentren online für jeden gelesenen Punkt aktualisiert, wie dies beispielsweise der Algorithmus von MacQueen tut [80]. Solche Algorithmen nutzen nur wenig Speicher, sind einfach zu implementieren und bei geeigneter Implementierung sehr schnell, liefern jedoch aus theoretischer und (häufig auch) aus praktischer Sicht schlechte Ergebnisse.

Die anderen Ansätze fassen den Datensatz beispielsweise in Form von Kernmengen zusammen, um anschließend jeden gewünschten Algorithmus auf dem aggregierten Datensatz laufen zu lassen. Hier unterscheiden wir nochmal nach zwei Vorgehensweisen:

1. Die Daten werden gemäß des *Merge-and-Reduce*-Paradigmas blockweise eingelesen. Für jeden Block wird eine Kernmenge berechnet, welche anschließend hierarchisch aggregiert werden.

2. Die Kernmenge wird online für jeden eingegebenen Punkt aktualisiert.

Wie bereits im Bereich der verteilten Algorithmen erwähnt, lassen sich viele Kernmengenkonstruktionen mit dem ersten Ansatz kombinieren, allerdings erhalten wir diese Flexibilität nicht ohne Nachteile. Bei jeder neuen Berechnung einer Kernmenge (nach Vereinigung zweier Kernmengen) verschlechtert sich die Güte um einen Faktor $(1+\epsilon)$. Die Güte der endgültigen Kernmenge ist dann entweder von der Länge des Datenstroms abhängig oder das ϵ muss mit einem zusätzlichen Faktor von $\log n$ umskaliert werden, was zu einer Vergrößerung der Kernmenge führt. Ferner müssen die Kernmengen für Teile des Datenstroms zwischengespeichert werden und der Berechnungsaufwand steigt.

Kernmengen, die online aufrecht erhalten werden können, umgehen diese Probleme, allerdings haben deren Größen bislang alle eine exponentielle Abhängigkeit von der Dimension, siehe [67]. Beispiele für praktisch eingesetzte Algorithmen der blockweisen Verarbeitung sind StreamLS [70] und StreamKM++ [58], eine populäre online betriebene Aggregation ist der Algorithmus BIRCH [92]. Eine vollständige Übersicht über eingesetzte Datenstromalgorithmen für Clusteringprobleme findet sich in [85].

Wir haben den Datenstromalgorithmus BICO für das k-Means-Problem entwickelt [A2/97], der nachweislich Kernmengen der Größe $O(k \log n/\epsilon^{d+2})$ berechnet und dies ohne Qualitätsverlust für eine online verarbeitete Eingabe tut. Gleichzeitig wurde der Algorithmus mit kleinem zur Verfügung stehenden Speicher gegen die in [85] referenzierten Algorithmen BIRCH, StreamLS und StreamKM++ evaluiert. Grundlage von BICO sind das Updateverfahren von BIRCH kombiniert mit einer Variation der Kernmengenkonstruktion aus [67], sodass die Clusterqualität von Kernmengen gut erhalten bleibt und diesbezügliche Probleme von BIRCH umgangen werden, aber gleichzeitig von der hohen Geschwindigkeit von BIRCH profitiert wird. BICO liefert auf nahezu allen untersuchten Datensätzen die besten Ergebnisse in Bezug auf Qualität und Geschwindigkeit im Vergleich mit den Referenzalgorithmen bei gleichem zur Verfügung stehenden Speicher.

[1] Die Größe der sich ergebenen Kernmengen für k-Means und projektives Clustering ist geringfügig größer als von uns in [A2/96] angegeben. Dies werden wir in der vollständigen Version der Publikation korrigieren.

Datenstromalgorithmen für Facility Location: Die Untersuchung von Facility Location in Datenströmen unterscheidet sich von den oben genannten Clusteringproblemen darin, dass die Ausgabe des Problems aus den geöffneten Zentren besteht und genauso groß wie die Eingabe sein kann (zum Beispiel bei sehr kleinen Öffnungskosten, wo genau n Facilities eröffnet werden) und für die Größe von Kernmengen eine natürliche untere Schranke in der Beschreibungsgröße der Ausgabe besteht. Für die Entwicklung von Datenstromalgorithmen, die eine höchstens polylogarithmische Abhängigkeit von n anstreben, kommt somit die Berechnung einer Kernmenge nicht in Frage. Stattdessen haben wir das Problem einer guten Approximation des Zielfunktionswerts einer optimalen Auswahl an Zentren betrachtet. Wir haben angenommen, dass die Punkte sich auf den Eckpunkten eines 2-dimensionalen Gitters der Größe Δ befinden. Indyk [73] hat als erster einen Algorithmus angegeben, der mit $\operatorname{poly}\log\Delta$ Speicher eine $O(\log^2\Delta)$ Approximation des optimalen Zielfunktionswertes auf Datenströmen berechnet. Das bislang beste Ergebnis durch Lammersen und Sohler [77] berechnete mit $\operatorname{poly}\log\Delta$ Speicher eine konstante Approximation.

Für dieses Problem ist es uns in [A2/94] erstmals gelungen eine $(1+\epsilon)$ Approximation im Datenstrom zu berechnen, die mit $\operatorname{poly}\left(\frac{\log\Delta}{\epsilon}\right)$ Speicherplatz auskommt. Grundlage dieses Algorithmus ist eine schnelle, einfache Aufteilung des des Raums. Durch diese Aufteilung wurde ferner erstmals ein $(1+\epsilon)$-Approximationsalgorithmus mit fast linearer Laufzeit angegeben, der ohne die Arora-Partionierung zur Approximation von Problemen im Euklidischen Raum auskommt, siehe [61].

Experimente: Die experimentelle Auswertung wurde zum Teil als Unterpunkt in den oben aufgeführten Arbeiten betrieben, siehe zum Beispiel [A2/96] und [A2/97] und teilweise eigenständig in Form von Abschlussarbeiten durchgeführt. Die untersuchten Datensätze stammen aus dem UCI Machine Learning Repository [88] und aus Simulationen mit astrophysischer Anwendung zum FACT Experiment, welches insbesondere Forschungsgegenstand von Teilprojekt C3 ist. In der Arbeit von Bulinski [A2/C3/104] wurden Kernmengen für das k-Means Problem parallel mittels MapReduce [65] berechnet. Untersucht wurde die Austauschbeziehung zwischen Kommunikationsaufwand, Lösungsqualität und Rechengeschwindigkeit.

Außerdem haben wir einen Clusteringalgorithmus entwickelt [A2/99], der Robustheit bei fehlerbehafteten Eingabedaten aufweist. Motiviert wurde die Arbeit durch eine vorzunehmende Datenanalyse auf einem eingebetteten System in einem Imaging-Satelliten. Auf Grund kosmischer Strahlungen ist damit zu rechnen, dass im Speicher Bits unkontrolliert flippen, was wiederum Clusteringergebnisse massiv beeinflussen kann. Im Rahmen dieser Arbeit haben wir eine fehlertolerante Version der kd-Baum-Datenstruktur entwickelt, die sich gegenüber einer beliebig vorgebbaren Anzahl δ möglicher Bitfehler robust verhält, wobei diese Fehler im Unterschied zur klassischen Verwendung von ECC-RAM-Bausteinen beliebig auf den Speicher verteilt sein können. Unter Verwendung von Techniken aus dem Bereich speichereffizienter Algorithmen ergibt sich eine Konstruktionszeit für diese Datenstruktur von $O(n\log^2 n+\delta^2)$. Diese Datenstruktur wird als zentraler Baustein für einen fehlertoleranten k-means-Algorithmus verwendet. Im Rahmen einer umfangreichen Evaluation wurde nachgewiesen, dass dieses Verfahren in der Praxis sogar dann anwendbar ist, wenn eine deutlich höhere als die angenommene Häufigkeit von Fehlern zu verarbeiten ist, diese aber im Unterschied zu den worst-case-Annahmen an zufälligen Stellen auftreten.

Schließlich haben wir uns mit der Analyse von Trajektorien und Bewegungsdaten beschäftigt. Dies findet breite Anwendungen bei Sportlern, Tieren oder meteorologischen Phänomenen. Wir haben einen Algorithmus zum Auffinden von Clustern solcher (Teil-)Trajektorien entwickelt [A2/100], der den worst-case Speicherbedarf von $O(n^2)$ auf $O(n)$ reduziert und zudem auf Grund seines modularen Designs eine hohe Flexibilität bei der Parametrisierung bis hin zum teilweisen Neustart der Verarbeitungspipeline ermöglicht. Diese Flexibilität wurde durch Experimente mit Daten aus zwei unterschiedlichen Anwendungsdomänen in Fußball und Meteorologie bestätigt.

Energiemodelle für (Eingebettete) Systeme: Praktisch alle heutigen Computersysteme sind energiebegrenzt. Wichtiges Kriterium bei eingebetteten Systemen ist die Batterielaufzeit; in großen Rechenzentren wird die Kühlung/Klimatisierung zum zentralen Problem, insbesondere auch aus Kostengründen; bei "Standardprozessoren", etwa in Desktopsystemen, besteht die Gefahr eines Hardwareschadens durch Überhitzung, weshalb typischerweise bereits beim Design der Hardware Kompromisse in der Rechenleistung zugunsten eines niedrigeren Energieverbrauchs gemacht werden [63]. Dabei entfällt ein immer größer werdender Anteil des Energieverbrauchs auf den Prozessor des Systems. Abbildung 3.1 zeigt die Entwicklung dieses Energieverbrauchs wie sie vom ITRS-Konsortium für die kommenden Jahre und Consumer-Prozessoren (Desktops, Server) prognostiziert wird. Bereits heute entfällt die Mehrheit des Energieverbrauchs auf den dynamischen Anteil.

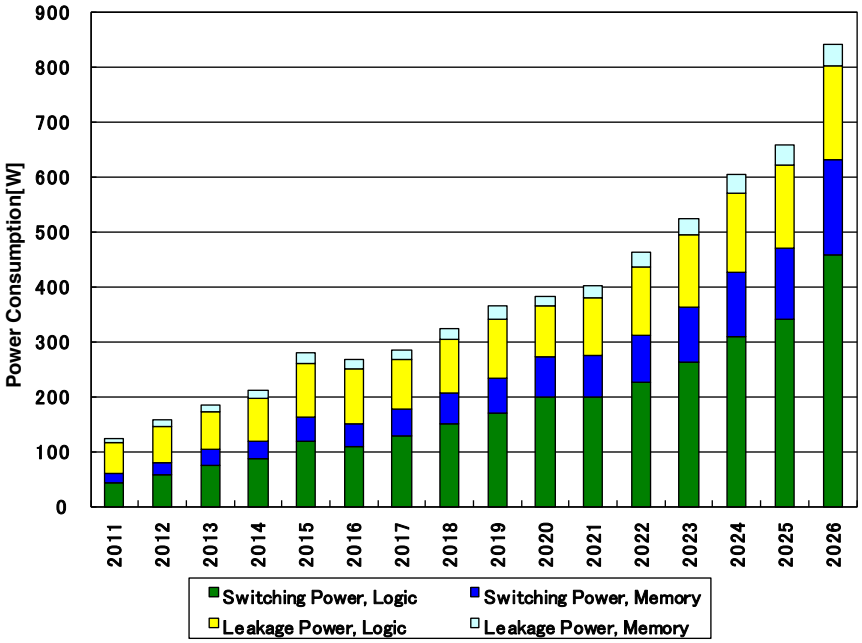


Abbildung 3.1.: Prognostizierter Energieverbrauch pro Chip, aufgeschlüsselt in statischen und dynamischen Energieverbrauch sowie die Komponenten Logik (also Rechenleistung) und Speicher. Quelle: International Technology Roadmap for Semiconductors (ITRS), 2011 Edition.

Bislang wurde diesem Trend hauptsächlich durch zwei Herangehensweisen begegnet: *(a)* auf der Hardwareseite kann mit verschiedenen Ansätzen der Energieverbrauch optimiert werden, z. B. durch noch weiteres Absenken der Versorgungsspannung oder durch Optimierung der verwendeten Dielektrika; *(b)* auf der Softwareseite versucht moderne Betriebssystemsoftware, durch geeignete Scheduling-Verfahren und durch *dynamic voltage and frequency scaling* (also einem Absenken der Taktfrequenz) den Energieverbrauch der laufenden Software zu optimieren. Der Einfluss des Scheduling wurde bereits vor fast 20 Jahren gezeigt [91]. Einen Überblick über die entstandenen Scheduling-Strategien gibt z. B. die Arbeit von Grunwald et al. [69]. Eine weitere Herangehensweise stellen energieoptimierende Compiler dar, wie etwa der in Dortmund entwickelte *encc*[2] oder die in Teilprojekt A3 entwickelten Techniken. Sie basieren auf Modellen zum Energieverbrauch einzelner Assemblerinstruktionen, wie etwa dem von Tiwari et al. [87]. Wichtige Erkenntnis solcher Modelle ist, dass bei gleicher Laufzeit der Energieverbrauch eines Algorithmus (also die Leistungs- bzw. Stromaufnahme) auch davon abhängt, welche Operationen er verwendet bzw. durchführt.

Unterschiede im Energieverbrauch verschiedener Rechenoperationen können sich aus verschiedenen Gründen ergeben. Zum Beispiel ist das Bewegen von Daten aus physikalischen Gründen inhärent mit einem Energieaufwand verbunden. Borkar und Chien [63] zeigen, dass der Energieaufwand, der durch Datenaustausch innerhalb des Prozessors entsteht, zunehmend zum Problem in modernen Prozessorarchitekturen wird. Moderne Prozessordesigns übersetzen den Rechenaufwand einzelner „teurer“ Operationen vielfach in eine erhöhte Zahl von Transistoren („chip space“), um die Aufgabe effizient zu erfüllen. Die zusätzlichen Transistoren führen in der Regel auch zu höherem Energieverbrauch, was die zugehörigen Assemblerinstruktionen besonders energieintensiv macht.

Die genannten Ansätze sind gut geeignet, um einen gegebenen Algorithmus möglichst energieoptimal auf einem gegebenen System zu implementieren bzw. in entsprechenden Maschinencode zu übersetzen. Sie können allerdings an dieser Stelle keine algorithmischen Entscheidungen mehr kompensieren, die durchaus signifikanten Einfluss auf die Energieeffizienz einer Problemlösung haben können. Gleichzeitig existieren keine Modelle, mit denen der Energieverbrauch eines Algorithmus bereits auf höherer Ebene charakterisiert werden könnte.

Kommunikations- und Rechnerarchitektur: Die Begrenzung durch den Energieverbrauch zeigt sich auch in der Art und Weise, wie sich Hardwaresysteme weiterentwickeln. Die Energieaufnahme eines einzelnen Prozessors steigt superlinear mit seiner Rechenleistung. Daher erreichen parallele Systeme nominell eine

[2] http://ls12-www.cs.tu-dortmund.de/daes/de/forschung/energy-aware-c-compiler.html

wesentlich höhere Energieeffizienz und sind deshalb selbst im Bereich eingebetteter Systeme bereits weit verbreitet.

Der Trend zur Parallelität unterstreicht allerdings gleichzeitig auch die Bedeutung der Kommunikation von Recheneinheiten untereinander. Dies führt in Mehrprozessorsystemen zu *NUMA*-Effekten (*non-uniform memory access*); in Prototypen wie Intels "Single-Chip Cloud Computer" [72] oder Tileras TILE-Gx [86] geht der Trend sogar so weit, dass dort die einzelnen Einheiten nur entlang eines 2D-Gitters direkt miteinander verbunden sind und nur Einheiten am Rand des Gitters direkt auf den vorhandenen Hauptspeicher zugreifen können. Multi-Core-Architekturen werden dadurch zu einer Art verteiltem System, dessen Leistung zunehmend durch die Kommunikation zwischen Recheneinheiten sowie mit dem Hauptspeicher begrenzt werden.

Im Bereich der Datenbankforschung wurde in einzelnen Arbeiten bereits der Fall der NUMA-Architekturen untersucht. Albutiu et al. stellten die These auf, dass NUMA-Effekten vollständig Rechenschaft getragen werden kann, wenn der entsprechende Algorithmus sich auf bestimmte Zugriffsmuster (sequentiell/zufällig; lesend/schreibend; etc.) beschränkt [60]. Li et al. konnten zeigen, dass sich Permutationsoperationen im Hauptspeicher deutlich beschleunigen lassen, wenn dabei die Topologie der Kern-zu-Kern-Verbindungen berücksichtigt wird [79]. Kiefer et al. konnten zeigen, wie sich durch eine NUMA-optimierte Datenpartitionierung die Anfrageperformanz um bis zu 75 % steigern lässt [75]. Porobic et al. zeigten mit Untersuchungen an transaktionslastigen Datenbankworkloads und dem Shore-MT-System, dass die NUMA-Topologie eines Systems deutliche Auswirkungen auf das Laufzeitverhalten haben kann [83]. Allerdings gibt es keine Platzierungs-/Auswertungsstrategie für Shore-MT, die für alle Workloads und alle NUMA-Topologien optimal wäre.

Aufgrund von NUMA- und Kommunikationseffekten zeigen daher selbst vermeintliche Einzelsysteme Ähnlichkeiten zu verteilten Systemen. Diese Beziehung von Parallelität in Multi-Core-Systemen zu verteilten Systemen wurde im Multimed-Projekt an der ETH Zürich sehr direkt umgesetzt. Folgend dem Modell des verteilten Ganymed-Systems [81, 82] wurden in Multimed die einzelnen Hardware-Prozessoren als eigenständige Knoten eines verteilten und replizierten Datenbanksystems betrachtet [84]. Dadurch wurde ein Leistungs- und Skalierungsverhalten erreicht, das sonst nur durch aufwändige Veränderungen im Datenbankkern erreichbar gewesen wäre. Im ähnlichen Stil betrachtet das Forschungs-Betriebssystem "Barrelfish" [62] Mehrprozessorsysteme als verteiltes System, in dem nur durch expliziten Nachrichtenaustausch kommuniziert werden kann.

Bestehende Systeme mit NUMA-Unterstützung wurden stets für eine *bestimmte* NUMA-Topologie untersucht und vorgestellt. In der Gruppe von Jens Teubner sind jedoch auch Algorithmen vertreten, die prinzipiell geeignet sind mit *jeder* (oder zumindest fast jeder) NUMA-Topologie effizient umzugehen. Ein Beispiel dafür ist "Handshake Join", bei dem Datenströme als Fluss entlang einer linearen Kette von Prozessoren modelliert werden [A2/102]. Das dadurch erhaltene Kommunikationsmuster beschränkt sich ausschließlich auf Kommunikation mit unmittelbaren Nachbarn und lässt sich auf alle gängigen NUMA-Architekturen direkt und einfach abbilden. Handshake Join ist "Cyclo-Join" nachempfunden, das Joins in verteilten Systemen besonders kommunikations-effizient implementiert [A2/98]. Einer ähnlichen Struktur folgen auch die vorgeschlagenen Lösungen für das "Frequent Item"-Problem mittels programmierbarer Hardware, bzw. mittels moderner Multi-Core-Prozessoren, sowie für das "Skyline"-Problem auf FPGAs.

Literatur

[58] M. R. Ackermann, M. Märtens, C. Raupach, K. Swierkot, C. Lammersen und C. Sohler. „StreamKM++: A clustering algorithm for data streams". In: *ACM Journal of Experimental Algorithmics* 17.1 (2012).

[59] S. Albers. „Energy-efficient algorithms". In: *Communications of the ACM* 53.5 (2010), S. 86–96.

[60] M.-C. Albutiu, A. Kemper und T. Neumann. „Massively Parallel Sort-Merge Joins in Main Memory Multi-Core Database Systems". In: *Proceedings of the VLDB Endowment* 5.10 (Juni 2012), S. 1064–1075.

[61] S. Arora. „Polynomial Time Approximation Schemes for Euclidean Traveling Salesman and other Geometric Problems". In: *Journal of the ACM.* 1996, S. 2–11.

[62] A. Baumann u. a. „The Multikernel: A New OS Architecture for Scalable Multicore Systems". In: *Proc. of the 22nd ACM Symposium on Operating Systems Principles (SOSP).* Bigsky, MT, USA, Okt. 2009, S. 29–44.

[63] S. Borkar und A. A. Chien. „The Future of Microprocessors". In: *Communications of the ACM* 54.5 (Mai 2011), S. 67–77.

[64] K. L. Clarkson und D. P. Woodruff. „Numerical linear algebra in the streaming model“. In: *Proceedings of the fourty-first ACM Symposium on Theory of Computing (STOC)*. 2009, S. 205–214.

[65] J. Dean und S. Ghemawat. „MapReduce: a flexible data processing tool“. In: *Communications of the ACM* 53.1 (2010), S. 72–77.

[66] D. Feldman und M. Langberg. „A unified framework for approximating and clustering data“. In: *Proceedings of the fourty-third ACM Symposium on Theory of Computing (STOC)*. 2011, S. 569–578.

[67] G. Frahling und C. Sohler. „Coresets in dynamic geometric data streams“. In: *Proceedings of the thirty-seventh ACM Symposium on Theory of Computing*. 2005, S. 209–217.

[68] M. Frigo, C. Leiserson, H. Prokop und S. Ramachandran. „Cache-Oblivious Algorithms“. In: *Proc. of the 40th IEEE Symp. on Foundations of Computer Science (FOCS)*. New York, NY, USA, Okt. 1999, S. 285–297.

[69] D. Grunwald, P. Levis, K. I. Farkas, C. B. M. III und M. Neufeld. „Policies for Dynamic Clock Scheduling“. In: *Proc. of the 4th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. San Diego, CA, USA, Okt. 2000, S. 73–86.

[70] S. Guha, A. Meyerson, N. Mishra, R. Motwani und L. O'Callaghan. „Clustering Data Streams: Theory and Practice“. In: *IEEE Transactions on Knowledge and Data Engineering* 15.3 (2003), S. 515–528.

[71] S. Har-Peled. „No, Coreset, No Cry“. In: *Proceedings of the twenty-fourth International Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*. 2004, S. 324–335.

[72] J. Howard u. a. „A 48-Core IA-32 Message-Passing Processor with DVFS in 45nm CMOS“. In: *2010 IEEE International Solid-State Circuits Conference*. San Francisco, CA, USA, Feb. 2010.

[73] P. Indyk. „Algorithms for Dynamic Geometric Problems over Data Streams“. In: *Proceedings of the thirty-sixth ACM Symposium on Theory of Computing (STOC)*. 2004, S. 373–380.

[74] R. Kannan, S. Vempala und D. Woodruff. „Nimble Algorithms for Cloud Computing“. In: *Computing Research Repository (CoRR)* abs/1304.3162 (2013).

[75] T. Kiefer, B. Schlegel und W. Lehner. „Experimental Evaluation of NUMA Effects on Database Management Systems“. In: *Tagungsband der 15. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW)*. Magdeburg, Germany, März 2013, S. 185–204.

[76] D. G. Kirkpatrick und R. Seidel. „The Ultimate Planar Convex Hull Algorithm“. In: *SIAM Journal of Computing* 15.1 (Feb. 1986), S. 287–299.

[77] C. Lammersen und C. Sohler. „Facility Location in Dynamic Geometric Data Streams“. In: *Proceedings of the sixteenth European Symposion on Algorithms (ESA)*. 2008, S. 660–671.

[78] M. Langberg und L. J. Schulman. „Universal epsilon-approximators for Integrals“. In: *Proceedings of the twenty-first ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2010, S. 598–607.

[79] Y. Li, I. Pandis, R. Müller, V. Raman und G. M. Lohman. „NUMA-Aware Algorithms: the Case of Data Shuffling“. In: *Proc. of the 6th Biennial Conference on Innovative Data Systems Research (CIDR)*. Asilomar, CA, USA, Jan. 2013.

[80] J. B. MacQueen. „Some Methods for Classification and Analysis of MultiVariate Observations“. In: *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967, S. 281–297.

[81] C. Plattner und G. Alonso. „Ganymed: Scalable and Flexible Replication“. In: *IEEE Data Engineering Bulletin* 27.2 (Juni 2004), S. 27–34.

[82] C. Plattner, G. Alonso und M. T. Özsu. „Extending DBMSs with satellite databases“. In: *The VLDB Journal* 17.4 (Juli 2008), S. 657–682.

[83] D. Porobic, I. Pandis, M. Branco, P. Tözün und A. Ailamaki. „OLTP on Hardware Islands“. In: *Proceedings of the VLDB Endowment* 5.11 (Juli 2012), S. 1447–1458.

[84] T.-I. Salomie, I. E. Subasu, J. Giceva und G. Alonso. „Database Engines on Multicores, Why Parallelize When You Can Distribute?“ In: *Proc. of the 6th European Conference on Computer Systems (EuroSys)*. Salzburg, Austria, Apr. 2011, S. 17–30.

[85] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho und J. a. Gama. „Data Stream Clustering: A Survey“. In: *ACM Computing Surveys* 46.1 (Juli 2013), 13:1–13:31.

[86] Tilera Corp. *TILE-Gx Processor Family.* URL: http://www.tilera.com/products/processors/TILE-Gx_Family.

[87] V. Tiwari, S. Malik und A. Wolfe. „Power Analysis of Embedded Software: A First Step Towards Software Power Minimization“. In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 2.4 (Dez. 1994).

[88] *UCI Machine Learning Repository.* 2014. URL: http://archive.ics.uci.edu/ml.

[89] L. G. Valiant. „A bridging model for multi-core computing“. In: *Journal of Computer and System Sciences* 77.1 (2011), S. 154–166.

[90] K. Varadarajan und X. Xiao. „A Near-linear Algorithm for Projective Clustering Integer Points“. In: *Proceedings of twenty-third ACM-SIAM Symposium on Discrete Algorithms (SODA).* 2012, S. 1329–1342.

[91] M. Weiser, B. B. Welch, A. J. Demers und S. Shenker. „Scheduling for Reduced CPU Energy“. In: *Proc. of the 1st USENIX Symposium on Operating Systems Design and Implementation (OSDI).* Monterey, CA, USA, Nov. 1994, S. 13–23.

[92] T. Zhang, R. Ramakrishnan und M. Livny. „BIRCH: A New Data Clustering Algorithm and Its Applications“. In: *Data Mining and Knowledge Discovery* 1.2 (1997), S. 141–182.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[A2/93] I. Caragiannis, A. Fanelli, N. Gravin und A. Skopalik. „Approximate pure Nash equilibria in weighted congestion games: existence, efficient computation, and structure“. In: *Proceedings of the 13th ACM Conference on Electronic Commerce.* Hrsg. von B. Rovan, V. Sassone und P. Widmayer. EC '12. New York, NY, USA: ACM, 2012, S. 284–301.

[A2/94] A. Czumaj, C. Lammersen, M. Monemizadeh und C. Sohler. „$(1+\varepsilon)$-Approximation for Facility Location in Data Streams“. In: *Proceedings of the 24th Symposium on Discrete Algorithms (SODA).* Hrsg. von S. Khanna. SIAM, 2013, S. 1710–1728.

[A2/95] A. Fanelli, L. Moscardelli und A. Skopalik. „On the Impact of Fair Best Response Dynamics“. In: *Proceedings of the 37th International Symposium on Mathematical Foundations of Computer Science.* Hrsg. von B. Faltings, K. Leyton-Brown und P. Ipeirotis. Springer, 2012.

[A2/96] D. Feldman, M. Schmidt und C. Sohler. „Turning Big Data into Tiny Data: Constant-size Coresets for k-means, PCA and Projective Clustering“. In: *Proceedings of the 24th Symposium on Discrete Algorithms (SODA).* Hrsg. von S. Khanna. SIAM, 2013, S. 1434–1453.

[A2/97] H. Fichtenberger, M. Gillé, M. Schmidt, C. Schwiegelshohn und C. Sohler. „BICO: Birch meets Coresets for k-means“. In: *Proceedings of the 21st European Symposium on Algorithms (ESA).* Hrsg. von H. L. Bodlaender und G. F. Italiano. Springer, 2013.

[A2/98] P. W. Frey, R. Goncalves, M. L. Kersten und J. Teubner. „A Spinning Join That Does Not Get Dizzy“. In: *Proc. of the 2010 Int'l Conference on Distributed Computing Systems (ICDCS).* Genova, Italy, Juni 2010, S. 283–292.

[A2/99] F. Gieseke, G. Moruz und J. Vahrenhold. „Resilient K-d Trees: K-Means in Space Revisited“. In: *Frontiers of Computer Science* 6.2 (2012), S. 166–178.

[A2/100] J. Gudmundsson, A. Thom und J. Vahrenhold. „Of Motifs and Goals: Mining Trajectory Data“. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL).* Hrsg. von E. Tanin, P. Kröger und P. Widmayer. ACM, 2012, S. 129–138.

[A2/101] S. Temme und J. Vahrenhold. „Revisiting the Construction of SSPDs in the Presence of Memory Hierarchies“. In: *Proceedings of the 28th European Workshop on Computational Geometry.* Hrsg. von W. Didimo und G. Liotta. 2012, S. 57–60.

[A2/102] J. Teubner und R. Müller. „How Soccer Players Would Do Stream Joins“. In: *Proc. of the 2011 ACM SIGMOD Conference on Management of Data.* Athens, Greece, Juni 2011, S. 625–636.

[A2/103] J. Vahrenhold. „On the Space-Efficiency of the Ultimate Planar Convex Hull Algorithm“. In: *Proceedings of the 24th Canadian Conference on Computational Geometry (CCCG).* Hrsg. von G. Aloupis und D. Bremner. 2012, S. 139–144.

Andere Veröffentlichungen

[A2/C3/104] M. Bulinski. „Parallelisierung von Clusteralgorithmen“. Bachelorarbeit. TU Dortmund, 2012.

3.4 Planung des Teilprojekts

Ziele

Ziel dieses Projekts ist die Entwicklung von Entwurfsparadigmen für Lernalgorithmen in eingebetteten Systemen. Dazu wollen wir einzelne wichtige Merkmale eingebetteter Systeme identifizieren, diese theoretisch modellieren, in den entwickelten Modellen Algorithmen entwerfen und analysieren und dann empirisch die Performanz dieser Algorithmen auswerten, um Rückschlüsse auf die Effizienz der Algorithmen und die Aussagekraft der Modelle ziehen zu können. Diese Rückschlüsse sollen, wenn nötig, wieder in die Modellierung einfließen. Aus den auf diese Weise erlangten Erkenntnissen wollen wir dann allgemeine algorithmische Entwurfsparadigmen für Lernalgorithmen in eingebetteten Systemen ableiten.

Konkret wollen wir Algorithmen unter den vier Teilaspekten *Energieverbrauch, Kommunikations- und Rechnerarchitektur, Datenströme* und *verteilte Systeme* betrachten, die in eingebetteten Systemen häufig von großer Relevanz sind und die wir im Folgenden detaillierter diskutieren. Die von uns geplante Herangehensweise setzt eine enge Kooperation zwischen Algorithmentheorie (Christian Sohler) und hardwarenaher Algorithmenforschung (Jens Teubner) voraus. Jan Vahrenhold, der einen Ruf an die Universität Münster angenommen hat, wird in der zweiten Projektphase durch Jens Teubner, der 2013 an die TU Dortmund berufen wurde, als Projektleiter ersetzt. Jens Teubner ist durch etliche Veröffentlichungen zu hardwarenaher Implementierung von Algorithmen bestens für die geplante Arbeit in diesem Projekt ausgewiesen.

Energieverbrauch: Über viele Jahre hinweg waren Miniaturisierung und eine Steigerung der Rechenleistung die zentralen Kriterien für die Weiterentwicklung von Rechnersystemen. Zunehmend sind Rechnersysteme jedoch begrenzt durch ihren Energieverbrauch bzw. durch die begrenzten Möglichkeiten, entstehende Wärme effizient abzuführen. Gerade in eingebetteten Systemen wird diese Problematik besonders deutlich. Hier ist die Lebensdauer von Akkus oder Batterien entscheidend. Außerdem ist die mögliche Wärmeabfuhr bei vielen Geräten (z. B. bei einem Mobiltelefon in der Hosentasche) deutlich begrenzt bzw. ein zu starkes Aufwärmen des Geräts unerwünscht.

In der Vergangenheit wurde nahezu ausschließlich versucht, das Energieproblem außerhalb des eigentlichen Anwendungsprogramms zu lösen, insbesondere durch energieoptimierte Hardware oder durch Schedulingtechniken. Lediglich Ansätze von energieoptimierenden Übersetzern nutzen die Tatsache aus, dass die Leistungsaufnahme eines Programms auch wesentlich vom konkreten Programmcode abhängt. Es wurde gezeigt, dass die Leistungsaufnahme von unterschiedlichen Assemblerinstruktionen sich um bis zu knapp einem Faktor zwei unterscheiden kann [87].

Existierende Modelle für die Leistungsaufnahme von Algorithmen sind jedoch nicht geeignet für den Entwurf von Algorithmen in der Algorithmentheorie. Energieangaben zu einzelnen Assembleranweisungen, wie etwa die von Tiwari et al. [87], sind zu feingranular, um damit Strategien auf algorithmischer Ebene zu bewerten. Als wichtiges Teilziel im Bereich der Energieeffizienz wollen wir daher Modelle herleiten, mit denen verwertbare Aussagen über den Energieverbrauch eines Algorithmus möglich sind. Am Beispiel der Lernalgorithmen wollen wir diese Modelle anschließend verifizieren und anwenden. Unser Ziel ist insbesondere, herauszufinden, ob bestimmte Algorithmen generell energieeffizienter sind als andere. Dies steht im Gegensatz zu Ansätzen (wie z.B. [59]), die über eine geeignete Systemsteuerung z.B. mittels Geschwindigkeitsskalierung oder über geeignete Lastverteilungsverfahren den Energieverbrauch eines Rechners senken. Außerdem wollen wir, insbesondere basierend auf den Ergebnissen der ersten Projektphase, energieeffiziente Algorithmen für ausgewählte Lernprobleme entwickeln. Dabei werden wir uns zunächst auf Algorithmen für das k-Means-Problem konzentrieren.

Kommunikations- und Rechnerarchitektur: Der zweite Aspekt eingebetteter Systeme, den wir modellieren wollen, ist die Kommunikations- und Rechnerarchitektur. Die von uns angestrebte Modellierung soll eine genauer auf die jeweilige Architektur abgestimmte Algorithmenentwicklung ermöglichen, um so Algorithmen zu erhalten, die auch unter starken Ressourcenbeschränkungen effizient sind. Primärziel ist dabei die Laufzeitreduktion. Bisher hat die moderne Rechnerarchitektur in der Algorithmentheorie vor allem durch die Modellierung von Speicherhierarchien (I/O-effiziente Algorithmen, Cache-oblivious Algorithmen) und die Entwicklung entsprechender Algorithmen Einzug gefunden. Durch die zunehmende Parallelisierung von CPUs gewinnen auch alte Modelle wie PRAM wieder an Bedeutung. Wir sind allerdings der Meinung,

dass fast alle bisherigen Modelle einen wichtigen Aspekt nicht beachten, der häufig für eine gute Performanz entscheidend ist: Die Kommunikation innerhalb des Rechners. Die Performanz von Algorithmen lässt sich unserer Meinung nach genauer modellieren, indem man eine Kombination von Speicher- und Kommunikationshierarchien betrachtet. Ein Rechenzentrum besteht beispielsweise aus Racks, die untereinander vernetzt sind (z. B. via Ethernet); in jedem Rack sind Rechner verbunden (ebenfalls via Ethernet); innerhalb eines Rechners sind Prozessoren durch Punkt-zu-Punkt-Verbindungen miteinander verbunden (z. B. via *HyperTransport* oder *QPI*); in jedem Prozessor sind mehrere Rechenkerne durch ein Netzwerk innerhalb des Chips verbunden. Diese Selbstähnlichkeit soll unser Ausgangspunkt sein, um Modelle für Algorithmen zu entwickeln, die vorhandene Kommunikationsstrukturen besonders effizient ausnutzen. Ähnlich wie beim *cache-oblivious*-Modell erhoffen wir uns, dass daraus entstehende Algorithmen mit (nahezu) *jeder* Kommunikationsarchitektur effizient umgehen können, also unabhängig von der Systemebene (Rechenzentrum, ..., Rechenkerne) und möglichst unabhängig von Parametern (etwa von Details der Topologie) sind. Daher wollen wir die Kommunikationsstruktur eines Rechners rekursiv durch ein Netzwerk beschreiben, wobei jeder Knoten eine Speichereinheit repräsentiert und gleichzeitig selbst ein Kommunikationsnetzwerk darstellt, das den *Datenfluss* innerhalb des Rechners modelliert. Auf der untersten Ebene besteht ein Netzwerk dann aus einem einzelnen Knoten und die Rechnung läuft innerhalb des Knotens ab.

Multi-BSP wurde von Valiant entwickelt und analysiert [89] und ist ein erstes eingeschränktes Modell in dieser Richtung. Wir wollen dieses Modell mit dem Ziel verallgemeinern, möglichst gut an die Rechnerarchitektur angepasste Algorithmen zu entwickeln. Dabei wollen wir zunächst das oben skizzierte Modell so präzisieren, dass die Entwicklung und Bewertung von Algorithmen möglich ist. Dann wollen wir zunächst für grundlegende Probleme untersuchen, wie effiziente Algorithmen für eine vorgegebene Architektur aussehen müssen. Idealerweise wollen wir, ähnlich wie bei den cache-oblivious Algorithmen, möglichst von der konkreten Rechnerarchitektur abstrahieren. Die von uns entwickelten Algorithmen werden dann experimentell evaluiert, um zu sehen, ob die von uns gewählte Modellierung geeignet ist. Ist dies der Fall, so werden wir dann bestehende Lernverfahren in diesem Modell analysieren und auch neue Verfahren entwickeln. Andernfalls werden wir die gewonnenen Erkenntnisse in eine bessere Modellierung fließen lassen.

Datenströme: Ein weiterer wichtiger Teilaspekt verteilter Systeme ist das Auftreten von Datenströmen. Darunter versteht man ein Szenario, in dem sehr große Mengen von Daten sequentiell gelesen werden und die Größe der Datenmenge eine klassische Verarbeitung im Hauptspeicher unmöglich macht. Datenströme müssen daher sequentiell und mit viel weniger Speicher als die Größe des Datensatzes verarbeitet werden. Sie treten in vielen Bereichen eingebetteter Systeme auf; in diesem SFB z.B. bei der Messung von Verkehrsdaten (B4), der Analyse von Betriebssystemprozessen (A1) oder der Auswertung physikalischer oder statistischer Daten (C3, C4 und C5). In der Algorithmik sind unterschiedliche Modelle für Datenstromalgorithmen etabliert, die verschiedene Szenarien modellieren. Typischerweise nimmt man an, dass der zur Verfügung stehende Speicherplatz polylogarithmisch in der Größe der zu verarbeitenden Daten ist. Es gibt aber andere Modelle, die z.B. mehrere Durchläufe über die Daten zulassen oder im Speicherbedarf weniger restriktiv sind. Im Rahmen des Projekts wollen wir auch die bereits etablierten Modelle aufgreifen und neue Datenstromalgorithmen für Lernprobleme entwickeln, sowie existierende Algorithmen verbessern.

In Zusammenarbeit mit Teilprojekt C4 hat sich in der ersten Projektphase gezeigt, dass für viele überwachte Klassifikationsprobleme kein Approximationsalgorithmus im klassischen Datenstrommodell existiert, da es einfache (pathetische) Beispiele gibt, die zeigen, dass man linearen Speicherplatz benötigt. Wir sind daher dazu übergegangen, die klassische Worst-Case-Annahme zu relaxieren und konnten auf diese Weise erste Datenstromalgorithmen für lineare Klassifikation entwickeln. Allerdings haben unsere bisherigen Algorithmen alle eine exponentielle Abhängigkeit von der Dimension des Merkmalsraums. Ein Teilziel der neuen Projektphase ist es nun, diese Dimensionsabhängigkeit durch eine Weiterentwicklung des Algorithmus polynomiell zu beschränken.

Verteilte Systeme: Im Bereich verteilter Algorithmen wollen wir uns auf einfache Rechenmodelle konzentrieren. Wir werden typischerweise von der Netzwerkstruktur abstrahieren und ein vollständiges Kommunikationsnetzwerk annehmen. Die Schwierigkeit liegt darin, dass die Daten auf die Knoten des Netzwerks (die Rechner) verteilt sind. Wir nehmen an, dass der Engpass des Systems die Kommunikation zwischen den Rechnern darstellt. Insbesondere werden wir auch annehmen, dass rundenweise und synchron gerechnet wird. Die geforderten Eigenschaften können in den meisten Szenarien mit Standardmethoden zur Verfügung gestellt werden. Die algorithmische Herausforderung liegt darin, dass kein (günstiger) globaler Zugriff auf die Daten möglich ist.

Für einige Probleme haben sich Kernmengen als eine effiziente Herangehensweise für die Entwicklung von Lernalgorithmen in einem solchen Modell gezeigt. Diese können genutzt werden, um lokal eine Zu-

sammenfassung der Daten zu erhalten, diese dann zu einem zentralen Rechner zu übertragen und dann dort die eigentliche Fragestellung zu lösen. Um einen Engpass bei der Übertragung durch die All-to-One-Kommunikation zu vermeiden, kann man die Kernmenge auch Schritt für Schritt vereinigen und wieder neu berechnen. Ein Schwerpunkt der zweiten Projektphase wird auch weiterhin die Entwicklung von Kernmengen für nicht-überwachte Lernprobleme sein. Wir wollen existierende Verfahren weiter verbessern und so noch effizientere Algorithmen erhalten.

Die oben skizzierte Herangehensweise, Kernmengen für verteilte Algorithmen zu nutzen, stellt allerdings gewisse Anforderungen an die berechneten Kernmengen. Beispielsweise muss die Vereinigung zweier Kernmengen wieder eine Kernmenge sein. Für die Vermeidung des Engpasses muss die Kernmengenkonstruktion außerdem mehrmals anwendbar sein. Es gibt eine Reihe von Kernmengenkonstruktionen, die diese Eigenschaften nicht oder nur eingeschränkt besitzen (z.B. die dimensionsunabhängigen Kernmengen für das Problem des kleinsten einschließenden Balls). Hier stellt sich die Frage, wieviel zusätzlicher Kommunikationsaufwand notwendig ist, um solche Kernmengen im Verteilten zu nutzen. Dies wollen wir in der nächsten Projektphase untersuchen.

Wir wollen außerdem untersuchen, inwieweit sich sogenannte LP-type-Probleme in einem verteilten Modell effizienter berechnen lassen. Dabei handelt es sich um eine geometrisch/kombinatorische Formulierung von Problemen, die eine ähnliche Struktur haben wie lineare Programme fester Dimension. Sie beinhalten beispielsweise lineare Programmierung, kleinster einschließender Ball oder die Berechnung der Maximum-Margin-Hyperebene, die den Abstand zu zwei unterschiedlich gefärbten Punktmengen maximiert. LP-type-Probleme können in Linearzeit gelöst werden, wobei die Dimension als konstant angenommen wird (und die Abhängigkeit der Laufzeit von der Dimension nicht polynomiell ist).

Arbeitsprogramm

Im Folgenden beschreiben wir die Arbeitspakete, die wir nach den Projektzielen gliedern.

Arbeitspaket 1. Energieverbrauch: Durch die hohe Komplexität moderner Rechnersysteme ist bislang sehr wenig darüber bekannt, wieviel Energie einzelne Teile einer Anwendungslösung tatsächlich verbrauchen. Chiphersteller etwa bleiben gar so vage, dass sie oft nur die Thermal Design Power ihrer Komponenten angeben, sie also nur den Grad an externer Kühlung angeben, den sie für den Betrieb etwa eines Prozessors für ausreichend halten.

Eine Schwierigkeit beim Aufstellen entsprechender Angaben war schon immer der hohe experimentelle Aufwand, um die aufgenommene Energie auch tatsächlich zu messen und somit aufgestellte Modelle zu validieren (Carroll und Heiser etwa leisteten diesen Aufwand für zwei Smartphone-Modelle). Diese Problematik hat sich deutlich vereinfacht, seit moderne Prozessoren direkt mit Sensoren ausgerüstet sind, mit denen Wärmeentwicklung und Energieverbrauch mit feiner Granularität direkt im Chip gemessen werden können. Durch eine Instrumentierung von Programmcode können die Informationen der Hardware ausgelesen werden und gegebenenfalls einzelnen Programmteilen zugeordnet werden. Ein entsprechendes System wurde in der Arbeitsgruppe von Jens Teubner angeschafft und steht für Messungen im Teilprojekt zur Verfügung. Ähnliche Messungen können auch mit der *MIMOSA*-Messtechnik aus Teilprojekt A4 durchgeführt werden. Allerdings bewertet *MIMOSA* nicht primär den Energieverbrauch von Algorithmen, sondern ist auf Systemkomponenten für Kommunikation oder Sensorik ausgerichtet.

Mit Hilfe der Funktionalität der Hardware kann somit der Energieverbrauch für verschiedene Algorithmenteile oder für verschiedene algorithmische Herangehensweisen mit vertretbarem Aufwand nachgewiesen und vermessen werden. Solche Messungen mit dem vorhandenen System sollen die Grundlage für die aufgestellten Modelle sein. Geplant ist zunächst, verschiedene Algorithmen und ihre Implementierungen detailliert bezüglich ihres Energieverbrauchs zu vermessen. Daraus sollen dann Modelle abgeleitet werden, um den grundsätzlichen Energieverbrauch eines Algorithmus zu bewerten.

Konkret sollen zunächst bereits vorhandene Datenanalysealgorithmen vermessen werden. Dazu stehen zum einen die bereits entwickelten Algorithmen aus der ersten Phase des SFB bereit. Zum anderen wurden in der Gruppe von Jens Teubner hardwareoptimierte Algorithmen aus dem Datenbankbereich entwickelt. Der Vorteil dieser existierenden Algorithmen ist, dass ihr Verhalten bereits sehr gut verstanden ist und vorgenommene Energiemessungen sich gut mit diesem Verhalten korrelieren lassen. Nach Möglichkeit sollen darüber hinaus weitere Algorithmen aus anderen Teilprojekten des SFB bewertet werden, um somit möglichst aussagekräftige Energiemodelle zu entwickeln.

Energiemessungen wollen wir zunächst an den in der ersten SFB-Phase entwickelten Algorithmen zum k-Means-Problem durchführen. Neben den Lösungen aus dem SFB (StreamKM++, BICO) wurden in der Literatur zahlreiche weitere Verfahren vorgeschlagen (z. B. k-means, k-means++, StreamLS, BIRCH), zu

denen auch teilweise fertige Implementierungen der Arbeitsgruppe vorliegen. Die vorhandenen Lösungen verfolgen teilweise deutlich unterschiedliche algorithmische Herangehensweisen. Durch Messungen wollen wir untersuchen, wie sich das auf den Energieverbrauch der Algorithmen auswirkt.

Aus den Messungen wollen wir Modelle ableiten, mit deren Hilfe sich der Energieverbrauch eines Algorithmus beschreiben lässt. Mit ihrer Hilfe wollen wir dann auf eine energieoptimierte Lösung des Problems hinarbeiten.

Um die Energieeffekte besser verstehen zu können, wird es voraussichtlich notwendig sein, die verschiedenen Algorithmen jeweils in ihre Teile zu zerlegen. Wir gehen davon aus, dass Messungen an einfachen algorithmischen Teilaspekten der Lösungen am besten geeignet sind, um die relevanten Zusammenhänge und Energieeffekte tatsächlich zu verstehen. Entsprechende Teilaspekte der genannten Algorithmen könnten z. B. Suchstrukturen (*kd*-Bäume) oder Datenzugriffsmuster sein.

In ersten Vorarbeiten zu diesem Teilprojekt konnten wir bereits zeigen, dass sich etwa die Leistungsaufnahme zweier Implementierungsstrategien (*Hash Join* und *Sort Merge Join*) um 20 % oder mehr unterscheiden kann. Beide Strategien lösen das selbe Datenbankproblem, zeigen jedoch aufgrund ihrer algorithmischen Struktur eine deutlich unterschiedliche Leistungsaufnahme. Auch hier gehen wir davon aus, dass sich durch eine Zerlegung der Algorithmen am besten Aussagen über die Ursachen und Zusammenhänge des Energieverbrauchs machen lassen. In früheren Arbeiten von Jens Teubner wurde bereits das Cache- und Laufzeitverhalten der Algorithmen untersucht. Die Ergebnisse aus diesen früheren Arbeiten eröffnen uns die Möglichkeit, Energiemessungen mit dem Wissen über das Speicherzugriffsverhalten zu korrelieren und eventuell vorhandene Zusammenhänge zu erkennen und in das Energiemodell zu integrieren.

Die entwickelten Modelle wollen wir anschließend sukzessive validieren und verfeinern durch die Betrachtung weiterer Algorithmen, insbesondere auch aus anderen Teilprojekten des SFB. Mögliche Kandidaten dafür wären aus unserer Sicht die Algorithmen aus den Teilprojekten B1 und B2 sowie evtl. die Triggeralgorithmen aus C5.

Arbeitspaket 2. Kommunikations- und Rechnerarchitektur: Wir gehen davon aus, dass durch den zunehmenden Grad an Parallelisierung die Kommunikation von Recheneinheiten einen immer wichtigeren Einfluß auf die Performanz haben wird. Die Übertragung von Daten ist nicht zuletzt sehr energieintensiv und hängt damit auch zusammen mit dem vorgehend genannten Problem der Energieeffizienz. Gelingt es, den Kommunikationsbedarf eines Algorithmus auf passende Weise einzuschränken, kann dadurch unmittelbar seine Effizienz gesteigert (und möglicherweise sein Energieverbrauch reduziert) werden. Längerfristig könnte der konsequente Einsatz kommunikationseffizienter Algorithmen sogar von Hardwareherstellern ausgenutzt werden, um wertvolle Chip- und Energieressourcen einzusparen und die Effizienz noch weiter zu steigern.

Um dieses Ziel zu erreichen, ist es zunächst notwendig, den durch Kommunikation entstehenden Aufwand entsprechend zu beschreiben. Nur so können Modelle für das Design von Algorithmen aufgestellt werden, die im Hinblick auf ihr Kommunikationsverhalten optimiert sind. Wichtig ist dabei auch ein geeigneter Detail- und Abstraktionsgrad. So existieren Modelle auf der Ebene von Maschinencode; sie sind jedoch ungeeignet für algorithmische Untersuchungen. Umgekehrt nehmen grobgranulare Modelle meist an, dass jede Recheneinheit direkt mit jeder anderen kommunizieren kann (z. B. die nimble algorithms von Kannan [74] und teilweise auch das Multi-BSP-Modell von Valiant [89]). Tatsächlich aber bilden bereits heute die Komponenten paralleler Systeme meist kein vollständig verbundenes Netzwerk, und wir gehen davon aus, dass dies in zukünftigen Architekturen noch deutlicher ausgeprägt sein wird.

Eine Schwierigkeit beim Aufstellen von Modellen ist die Komplexität realer Systeme. Zunächst resultiert ein Kommunikationsmechanismus grundsätzlich in zwei Effekten: Bandbreitenbeschränkung und Zugriffslatenz. Bereits hier ist nicht klar, welcher der beiden Effekte in der Praxis überwiegt oder inwieweit ein Algorithmus bezüglich beider Einschränkungen gleichzeitig optimiert werden kann. Entwicklungen auf der Hardwareseite deuten darauf hin, dass reale Systeme zunehmend bandbreitenbegrenzt sind. Etwa ist die lokale Speicherbandbreite von Intel-Mehrprozessorsystemen während der letzten Generationen schneller gewachsen als die Bandbreite des *Quick Path Interconnect*, mit dem die Prozessoren untereinander verbunden sind.

Eine weitere Schwierigkeit wird sein, dass Kommunikationseffekte inhärent auch mit anderen Effekten interagieren, die bislang nur isoliert betrachtet wurden. Etwa lässt sich der Zugriff auf eine Speicherhierarchie ebenfalls als Kommunikation zwischen Speicher- und Rechenkomponenten verstehen. Die Synchronisation paralleler Prozesse stellt ebenfalls eine Form der Kommunikation dar. Schließlich tritt ein wichtiger Kommunikationseffekt, die limitierte Bandbreite, typischerweise erst in Verbindung mit paralleler Verarbeitung auf.

Um ein Modell zu erhalten, das für die Algorithmenforschung praktikabel ist, wollen wir uns bei der Modellierung zunächst auf den *Datenfluss* eines Algorithmus konzentrieren, weil sich damit sowohl die ver-

wendete Bandbreite eines Algorithmus quantifizieren lässt als auch die Latenz (z. B. in Form der Anzahl von *Hops* durch das Netzwerk). Ein Ansatz könnte sein, Algorithmen als Rechenschritte zu betrachten, die sich zusammen mit den Daten entlang der Kommunikationskanäle durch das Netzwerk bewegen. Idealerweise bräuchte dann die Parallelität des Systems nicht mehr explizit modelliert werden, was die Anwendung des Modells für die Algorithmenforschung deutlich vereinfachen würde.

Gleichzeitig wollen wir bestehende Algorithmen näher betrachten, deren Effizienz in Bezug auf ihr Kommunikationsverhalten bereits in der Praxis gezeigt wurde. Hier wollen wir untersuchen, wie (bzw. ob) sich die individuellen Vorgehensweisen dieser Algorithmen generalisieren lassen und ob sich daraus Modelle für allgemeingültige Aussagen ableiten lassen. Teil dieser Arbeit werden auch experimentelle Untersuchungen der bestehenden Ansätze sein, u. a. auch, um die Effizienz dieser Ansätze zu verifizieren (so wurde z. B. bereits durch Li et al. [79] gezeigt, dass die Modelle von Albutiu et al. [60] zu einfach sind, um daraus hilfreiche Aussagen über die NUMA-Effizienz von Datenbankalgorithmen zu machen).

Arbeitspaket 3. Verteilte Algorithmen: Ein Schwerpunkt dieses Arbeitspakets, das die Arbeiten aus der ersten Projektphase fortsetzen soll, ist auch weiterhin die Entwicklung von effizienten Methoden, die auf der Berechnung von Kernmengen basieren. In der ersten Projektphase haben wir Kernmengen für eine Reihe von Problemen aus dem unüberwachten Lernen entwickelt wie z.B. k-means, subspace k-clustering oder PCA. Mit Hilfe dieser Kernmengen lassen sich einfache verteilte Algorithmen entwickeln. Außerdem kann man kernmengenbasierte Verfahren relativ einfach in Map-Reduce verwenden. Aufbauend auf den Arbeiten aus der ersten Projektphase wollen wir versuchen, die Kernmengengrößen für diese Probleme weiter zu reduzieren. Dabei stellt sich z.B. die Frage, ob man durch eine direkte Kombination einer Dimensionsreduktion über Hauptkomponentenanalyse mit Kernmengenkonstruktionen für k-Means Clustering bessere Ergebnisse erzielen kann als durch dieselbe Kombination, wenn die Kernmengenkonstruktion als Black Box betrachtet wird. Die von uns in der ersten Projektphase entwickelten Methoden zur Dimensionsreduktion im Kontext von Kernmengen basieren stark auf Eigenschaften des quadrierten euklidischen Abstands. Es stellt sich die Frage, ob man ähnliche Ergebnisse auch für andere Abstandmaße wie z.B. die Summe der euklidischen Abstände erzielen kann. Darüber hinaus wollen wir versuchen, das Kernmengenparadigma auch auf andere Lernprobleme zu übertragen wie z.B. Entscheidungsbaumlernen.

In der Vergangenheit hat sich gezeigt, dass kernmengenbasierte Verfahren insbesondere für Datenstromalgorithmen in der Praxis sehr effizient sind (BICO, StreamKM++, etc.). Diese lassen aufgrund ihrer Architektur vermuten, dass man in der Praxis ebenfalls sehr effiziente *verteilte* Algorithmen mit Hilfe von Kernmengen entwickeln kann. Wir haben dazu bereits einige Untersuchungen in der ersten Projektphase durchgeführt, haben dabei aber bisher noch nicht völlig überzeugende Ergebnisse erhalten. Bei Verfahren wie Map-Reduce (Hadoop) scheinen die Vorteile von kernmengenbasierten Verfahren aufgrund des relativ großen Overheads nicht gut zur Geltung zu kommen. Aufgrund der einfachen Struktur kernmengenbasierter Verfahren könnte man hier aber auch über eine direkte verteilte Implementierung nachdenken und diese mit Map-Reduce und anderen Architekturen vergleichen. Dies soll in der zweiten Projektphase geschehen.

Eine weitere interessante Frage ist die Approximation der Kernmengengröße. Anstatt - wie zur Zeit üblich - zu zeigen, dass eine bestimmte Kernmengengröße für jede beliebige Eingabe erreicht werden kann, wäre es interessant zu untersuchen, ob man für einen gegebenen Güteparameter eine optimale, d.h. möglichst kleine Kernmenge effizient bestimmen kann, oder ob es zumindest möglich ist eine Approximation zu erhalten; also eine Kernmenge, die die Güteanforderung erfüllt, aber höchstens um den Approximationsfaktor größer ist als die bestmögliche Kernmenge. Ebenso kann man sich für eine gegebene Beschränkung des Speichers fragen, ob man die bestmögliche Kernmenge bezüglich ihrer Güte finden kann. Insbesondere die letztere Fragestellung ist aus praktischer Sicht interessant, denn typischerweise ist der lokale Speicherplatz beschränkt und die Kernmenge muss dort abgespeichert werden können.

Einige Kernmengenkonstruktionen sind nicht abgeschlossen unter Vereinigung, was ihre Einsatzmöglichkeiten in verteilten Systemen und Datenströmen zumindest sehr einschränkt. Andererseits sind diese Kernmengen häufig viel kleiner. Beispielsweise gibt es für den kleinsten einschließenden Ball eine Kernmenge, deren Größe unabhängig von der Dimension ist, während die beste bekannte Kernmenge, die abgeschlossen unter Vereinigung ist, exponentielle Größe in der Dimension hat. In einem verteilten System stellt sich daher die Frage, ob man nicht unter Vereinigung abgeschlossene Kernmengen effizient berechnen kann, wenn die Daten auf unterschiedliche Rechner verteilt sind.

Außerdem wollen wir verteilte Algorithmen zum Lösen von LP-type-Problemen entwickeln. Als erstes werden wir dazu die geometrischen Ansätze zum Lösen niedrigdimensionaler linearer Programme auf ihren Einsatz in verteilten Systemen untersuchen. Beispielsweise betrachtet der Algorithmus von Seidel die Bedingungen des linearen Programms in zufälliger Reihenfolge und aktualisiert die Lösung. Dabei muss im Erwartungswert nur in $O(\log n)$ Fällen rekursiv ein $(d-1)$-dimensionales lineares Programm gelöst werden.

In allen anderen Fällen ändert sich nichts an der aktuellen Lösung. Dies sollte man in verteilten Systemen ausnutzen können, denn letztendlich bedeutet dies, dass man nur sehr selten eine neue Lösung durch das Netzwerk propagieren muss. Die meiste Zeit können die einzelnen Rechner einfach überprüfen, ob die nächste einzufügende Bedingung die aktuelle Lösung verletzt.

Arbeitspaket 4. Datenstromalgorithmen: Das Datenstrommodell dient zur Modellierung von Algorithmen für sehr große Datenmengen, die in Form von Datenströmen, d.h. Sequenzen von Daten, auftreten. Dementsprechend wird angenommen, dass die Eingabe eine (sehr lange) Sequenz von Daten ist, die vom Algorithmus von links nach rechts gelesen wird. Im Fall von unüberwachten Lernproblemen sind die Daten typischerweise Merkmalsvektoren bzw. Punkte. Bei überwachten Lernproblemen erhalten die Punkte zusätzlich ihre Klassenzugehörigkeit. Da es darum geht, sehr große Datenmengen zu verarbeiten, wird zusätzlich verlangt, dass der Speicherbedarf des Algorithmus höchstens $\log^{O(1)} n$ ist, wobei n die Anzahl der Punkte im Datenstrom bezeichnet.

Da wir im Datenstrommodell nicht die gesamten Daten im Hauptspeicher halten können, ist es normalerweise nicht möglich, nichttriviale Probleme exakt zu lösen. Es scheint, dass man bestenfalls einen $(1+\epsilon)$-Approximationsalgorithmus erhalten kann. Kennt man die Länge des Datenstroms, so kann man ϵ z.B. auf $1/\log n$ setzten und so (polynomiellen Speicherbedarf in $1/\epsilon$ vorausgesetzt) einen Algorithmus mit $\log^{O(1)} n$ Speicherbedarf erhalten, der für großes n eine immer bessere Approximationsgüte erhält. In vielen Szenarien ist n allerdings nicht a priori bekannt und es stellt sich die Frage, ob man das oben beschriebene Verhalten auch ohne Vorwissen über die Länge des Datenstroms erhalten kann. Interessanterweise scheint dies für einige Probleme möglich zu sein (wie z.B. Frequency Moments oder k-Means Clustering), für andere jedoch nicht (wie z.B. Ausdehnungsapproximation von Punktmengen). Damit würde dann ein bestmöglicher Algorithmus nicht eine $(1+\epsilon)$-Approximation liefern, sondern für n gegen ∞ gegen eine optimale Lösung konvergieren. Letztere Algorithmen nennen wir *asymptotisch optimal*. Wir wollen in der zweiten Antragsphase genauer untersuchen, für welche Probleme es asymptotisch optimale Datenstromalgorithmen gibt.

In der ersten Projektphase haben wir in Kooperation mit C4 Datenstromalgorithmen für lineare Klassifikationsverfahren entwickelt. Diese basieren auf einer Kombination von Subsampling und Algorithmen zur Approximation der Ausdehnung von Punktmengen. Unter einer relaxierten worst-case-Annahme konnten wir eine Approximationsgüte von $(1+\epsilon)$ beweisen. Leider haben die von uns entwickelten Algorithmen eine exponentielle Abhängigkeit von der Dimension und sind somit z.B. für kernbasierte Verfahren nicht geeignet. In der zweiten Phase wollen wir versuchen, die Dimensionsabhängigkeit auf polynomiell zu reduzieren oder sogar nicht abhängig von der Dimension zu sein. Unter Umständen wird es dazu notwendig sein, die worst-case-Annahmen weiter zu relaxieren. Eine interessante Alternative könnte auch die *Smoothed Analysis* liefern, bei der man die Punkte einer worst-case-Eingabe mit zufälligen gaußverteilten Rauschen versieht und die Komplexität eines Algorithmus in Abhängigkeit der Varianz dieses Rauschens analysiert. In unserem Fall würde sich das Rauschen allerdings auf die Approximationsgüte auswirken und hier stellt sich zunächst die Frage nach der Definition von Approximationsgüte unter Smoothed Analysis. Man könnte natürlich einfach die erwartete Approximationsgüte betrachten, aber dies schließt Probleme aus, bei denen auch nur mit winzig kleiner Wahrscheinlichkeit keine Approximationsgüte erreicht werden kann.

In vielen Szenarien im Lernen ist es teuer, Label für die Beobachtungen zu erhalten. Daher stellt sich die Frage nach Lernalgorithmen, die das Vorhandensein von Beobachtungen ohne Label ausnutzen können. Dazu hat sich aus einer Diskussion mit Albert Bifet im Rahmen des Topical Seminars folgende Modifikation des k-Means-Algorithmus ergeben, die wir gemeinsam im Hinblick auf den Einsatz für Nächste-Nachbar-Verfahren untersuchen wollen: Der Algorithmus erhält zwei Parameter r und b, wobei r die Anzahl der roten Cluster ist und b die Anzahl der blauen. Die Farben rot und blau entsprechen dabei den Klassenlabeln. Der Algorithmus startet nun genau wie k-Means mit einer Menge von $k = r + b$ Zentren, von denen genau r rot und b blau sind. Die roten Eingabepunkte werden dem nächstgelegenen roten Zentrum zugewiesen und die blauen dem nächstgelegenen blauen Zentrum. Nicht gefärbte Punkte werden dem nächstgelegenen Zentrum zugewiesen. Ziel ist eine Menge von Zentren zu berechnen, um diese für die Nächste-Nachbar-Klassifikation einzusetzen. Weiterhin soll das Verfahren auf mehrere Klassenlabel und mit Hilfe von Kernmengen auch auf den Einsatz in Datenströmen erweitert werden. Ferner sollen theoretische Performanzgarantien hergeleitet werden, sowie die Performanz unserer Methoden sowohl in Bezug auf Klassifikationsgüte als auch Ressourceneffizienz empirisch untersucht werden.

Ein weiterer großer Arbeitsbereich wird die praktische Umsetzung der von uns entwickelten theoretischen Verfahren in effiziente Datenstromalgorithmen sein. Der von uns in der ersten Phase entwickelte Algorithmus BICO basiert im Gegensatz zu den meisten kernmengenbasierten Datenstromalgorithmen nicht auf dem Merge-and-Reduce-Prinzip, sondern auf Techniken aus dem BIRCH-Algorithmus. BICO ist deutlich effizienter als z.B. StreamKM++ (letzterer basiert auf Merge-and-Reduce). Da wir jedoch zwei völlig un-

terschiedliche Algorithmen miteinander verglichen haben, können wir daraus natürlich nicht schließen, dass Merge-and-Reduce generell langsamer ist als andere Techniken. Wir wollen in dieser Phase genauer untersuchen, welche algorithmischen Techniken für die Umsetzung eines Datenstromalgorithmus bei hardwarenaher Implementierung besonders gut geeignet sind. Insbesondere planen wir auch eine möglichst effiziente Implementierung des Merge-and-Reduce-Verfahrens, da diese dann als Black Box für neue Kernmengenkonstruktionen eingesetzt werden kann.

Zeitplan

	2015				2016				2017				2018			
Arbeitspaket Quartal	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Energieverbrauch:																
2. Kommunikations- und Rechnerarchitektur:																
3. Verteilte Algorithmen:																
4. Datenstromalgorithmen:																

3.5 Stellung innerhalb des Sonderforschungsbereichs

Projekt A2 soll die theoretischen Grundlagen für die Entwicklung von Lernalgorithmen in eingebetteten Systemen schaffen. Insbesondere ist das Ziel, allgemeine algorithmische Entwurfsparadigmen herzuleiten, mit denen Lernverfahren in eingebetteten Systemen entwickelt werden können. Damit ergänzt Teilprojekt A2 die Arbeit aus A1 um die Algorithmentheorie. Während in A1 Problemlösungen entwickelt werden, die alle Ressourcenbeschränkungen aus der konkreten Fragestellung berücksichtigen müssen, betrachten wir unterschiedliche Beschränkungen zunächst einzeln, um so einen besseren Blick auf die Konsequenzen für die Algorithmenentwicklung zu haben. Dafür nehmen wir in Kauf, dass die von uns entwickelten Algorithmen nicht immer direkt in eingebetteten Systemen einsetzbar sind.

Teilprojekt A4 beschäftigt sich ebenfalls mit der Modellierung von Energieverbrauch. In A4 bezieht sich die Modellierung auf den Verbrauch einzelner Systemkomponenten (etwa für Kommunikation oder Sensorik). Im Gegensatz dazu wollen wir in Arbeitspaket 1 untersuchen, wie sich verschiedene Algorithmen auf den Energieverbrauch auswirken. Beide Teilprojekte ergänzen sich damit. Etwa beabsichtigen wir, durch die Verwendung der MIMOSA-Plattform aus A4 ein möglichst breites Spektrum an Hardwareplattformen abzudecken. Bei der Bewertung von verteilten Algorithmen könnten Modelle für Kommunikation oder Sensorik aus A4 in unser Teilprojekt einfließen.

Energieverbrauch ist auch ein wichtiges Kriterium in Teilprojekt A1. Allerdings wird dort angenommen, dass sich der Energieverbrauch primär aus der Laufzeit (bzw. der verbrauchten Taktzyklen) ergibt. Teilprojekt A1 ist damit komplementär zu unserem Arbeitspaket 2, in dem wir untersuchen wollen, wie sich einzelne algorithmische Strukturen auf den Energieverbrauch auswirken. Wir hoffen, so gemeinsam einen guten Überblick zu erhalten.

Zusammen mit Teilprojekt A6 sollen Kernmengen zur Approximation des Centrality Index von Graphen eingesetzt werden. Um diesen zu berechnen, benötigt man für jeden Knoten die Summe der Abstände zu allen anderen Knoten. Diese können mit Hilfe von Kernmengen für die Kürzeste-Wege-Metrik effizient approximiert werden. Auch für andere Optimierungsprobleme wie z.B. Varianten des Steinerbaumproblems ist eine gemeinsame Untersuchung von kernmengenbasierten Ansätzen geplant.

Gemeinsam mit Teilprojekt B2 haben wir bereits einen Algorithmus zur Signal/Untergrund-Klassifikation von Zeitreihen entwickelt und diesen erfolgreich im Kontext der Erkennung (biologischer) Viren angewendet. Die Klassifikation erfolgt dabei mit Hilfe des Nächste-Nachbar-Verfahrens auf den Zentren eines k-Means-Clusterings, das mit Hilfe des in A2 entwickelten BICO Algorithmus berechnet wurde. In der zweiten Förderperiode wollen wir diese Kooperation fortsetzen.

Weiterhin wollen wir in A2 entwickelte Datenstromalgorithmen einsetzen, um Sensordaten aus Teilprojekt B3 zu analysieren.

Gemeinsam mit Teilprojekt C3 wurde in der ersten Förderperiode ein k-Means-Algorithmus erfolgreich für die Analyse des FACT-Datensatzes eingesetzt. Diese Zusammenarbeit wollen wir in der nächsten Phase fortführen.

Die Kooperation mit Teilprojekt C4 bei der Entwicklung und Analyse von Datenstromalgorithmen für überwachte Lernprobleme soll ebenfalls in der nächsten Förderperiode fortgesetzt werden.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

Sublinear Algorithms for the Analysis of Very Large Graphs
(Sohler,Förderkennzeichen ERC-StG 307696) (Förderzeitraum: 12/2012 - 11/2017)

Das Projekt beschäftigt sich mit der Entwicklungen und Analyse von Zufallsprozesses zur Analyse sehr großer Graphen im Rahmen des Property Testings. Es besteht kein inhaltlicher Bezug zu Teilprojekt A2.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	117.600	—	117.600	—	117.600	—	117.600
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	117.600		117.600		117.600		117.600	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (A2)							
Wissen-schaftliches Personal	1	Sohler, Christian, Prof. Dr., Prof	Algorithmische Komplexitäts-theorie	LS 2, Fak. Inf, TU Do	5,0	—	Grund-ausstattung
	2	Teubner, Jens, Prof. Dr., Prof	Informationssys-teme	LS 6, Fak. Inf, TU Do	8,0	—	Grund-ausstattung
	3	Krivošija, Amer, M. Sc. WiMi	Algorithmische Komplexitäts-theorie	LS 2, Fak. Inf, TU Do	27,86	—	Grund-ausstattung
	4	N.N. , WiMi	Informationssys-teme	LS 6, Fak. Inf, TU Do	27,86	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	5	Stankjawitschene, Alla, Verw. Ang.	—	LS 2, Fak. Inf, TU Do	1,99	—	Grund-ausstattung
	6	Stankjawitschene, Alla, Verw. Ang.	—	LS 6, Fak. Inf, TU Do	1,99	—	Grund-ausstattung
Beantragte Personalmittel (A2)							
Wissen-schaftliches Personal	7	Schwiegelshohn, Chris, Dipl.-Inf. WiMi	Algorithmische Komplexitäts-theorie	LS 2, Fak. Inf, TU Do	—	Doktorand/in	—
	8	N., N., WiMi	Informationssys-teme	LS 6, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Sohler, Christian**
   Projektleitung. Schwerpunkt Algorithmik von Lernverfahren. Mitarbeit in allen Arbeitspaketen.

2. **Teubner, Jens**
   Projektleitung. Schwerpunkt Datenbanksysteme. Mitarbeit primär in den Arbeitspaketen 1 und 2, Unterstützung der anderen Arbeitspakete insbesondere in Bezug auf experimentelle Auswertungen.

3. **Krivošija, Amer**
   Mitarbeit in Arbeitspaketen 3 und 4 (insbesondere die Weiterentwicklung von kernmengenbasierten Clusteringverfahren); sowie bei der Entwicklung von Modellen in AP1.

4. **N.N.,**
   Energieeffiziente Algorithmen, AP1.

5. **Stankjawitschene, Alla**
   Sekretariat.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

7. **Schwiegelshohn, Chris**
   Mitarbeit in den Arbeitspaketen 2-4

8. **N., N.**
   Kommunikations- und Rechnerarchitektur, AP2

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	6.500	6.500	6.500	6.500
Summe der vorhandenen Mittel	6.500	6.500	6.500	6.500
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt A3

3.1.1 Titel:

Methoden der Effizienten Ressourcennutzung in Algorithmen des Maschinellen Lernens

3.1.2 Fachklassifizierung:

Eingebettete Systeme, Maschinelles Lernen

3.1.3 Projektleitung

Prof. Dr. Marwedel, Peter, 30.04.1949

LS 12, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 16
44227 Dortmund

Tel.: 0231-755-6111
Fax : 0231-755-6116
E-Mail: `peter.marwedel@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Rahnenführer, Jörg, 19.05.1971

Fachgebiet Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik Technische Universität Dortmund
Vogelpothsweg 87
44227 Dortmund

Tel.: 0231-755-3121
Fax : 0231-755-5303
E-Mail: `joerg.rahnenfuehrer@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Ziel der zweiten Phase des Teilprojekts ist die Entwicklung von Methoden zur Modellwahl unter Ressourcenbeschränkungen. Für Probleme mit einer sehr großen Anzahl an Beobachtungen oder Variablen soll aus einer umfangreichen Menge von rechenintensiven Lernverfahren dasjenige mit der besten Vorhersageleistung ausgewählt werden. Die Realisierung der Lernverfahren und der Methoden zur Modellwahl soll in der Sprache R erfolgen. Für die Sprache R soll durch Ausnutzung von Information über verfügbare Ressourcen eine Ausführung der Lernverfahren erreicht werden, welche eine Lösung der betrachteten komplexen Probleme erlaubt. Die Optimierung der Ausführung soll dabei gezielt für Lernverfahren erfolgen und bei der Entwicklung von neuen Methoden der Modellwahl sollen gezielt Ressourcenbeschränkungen beachtet werden.

Solche Methoden zur Modellwahl sind wichtig für die personalisierte Medizin, deren Ziel die Bestimmung einer wirksamen Therapie basierend auf individuellen genetischen Mustern ist. Dazu werden hochdimensionale genetische Messungen zur Vorhersage und Modellierung von Überlebenszeiten verwendet. Methoden der Modellwahl sollten daher in diesem Bereich zwei zentrale Punkte berücksichtigen, nämlich die Entwicklung von Modellen für Teilmengen von Beobachtungen und die bestmögliche Auswahl von Modellen mit Hilfe von vielen konkurrierenden Lernverfahren mit verschiedenen zu optimierenden Parametereinstellungen. In vielen Fällen, wie z.B. bei einer Krebstherapie, erweisen sich Vorhersage-Modelle nur innerhalb von Patiententeilgruppen als hinreichend genau. Ein vielversprechender Ansatz besteht darin, zu untersuchen, welche Patienten aus anderen Teilgruppen zusätzliche nützliche Informationen beisteuern können. Solche Patienten werden dann bei der Modellbildung gewichtet berücksichtigt. Dieses Vorgehen und die ohnehin große Komplexität der Daten führen zu einer riesigen Anzahl an möglichen Modellen, für deren Konstruktion verschiedenste konkurrierende Lernverfahren existieren.

Die automatische Modellwahl unter Ressourcenbeschränkung ist hier dazu geeignet, in Datensituationen aus der großen Anzahl an Lernverfahren mit oft verschiedenen möglichen Parametereinstellungen diejenigen mit größter Güte, zum Beispiel höchster Vorhersagegenauigkeit, zu finden. Der entsprechende Algorithmus für die Modellwahl bewegt sich anhand der Analyse und Bewertung vorheriger Auswertungen von Modellen schnell und effizient in die Regionen der interessantesten Modelle und zugehörigen Parametereinstellungen. Durch die dennoch große Anzahl auszuwertender Modelle handelt es sich dabei um eine laufzeitintensive Prozedur, welche eine Optimierung der Ausführung erfordert. Dazu betrachten wir sowohl die Optimierung von einzelnen Programmabläufen als auch die optimierte Ausführung einer großen Menge von Programmabläufen auf parallelen Hardwaresystemen, wie z.B. Multiprozessor- und Rechenclustersystemen. Hierzu findet eine Analyse des Ressourcenbedarfs der Programmabläufe in Abhängigkeit von Modell- und Hardwareparametern statt, welche die Basis für Offline- bzw. Online-Entscheidungen für die Ablaufplanung darstellt. Hierbei soll als Grundlage für die optimierte Ablaufplanung ein Rückkopplungsverfahren entworfen werden. Dieses verwendet Methoden des maschinellen Lernens, um Parameter des Gesamtsystems auf wechselnde Umgebungsbedingungen bei Hardware- sowie Modelleigenschaften anpassen zu können. Dabei werden als zu optimierende Ressourcen nicht nur die aufzuwendende Rechenzeit, sondern darüber hinaus auch der Speicherbedarf berücksichtigt. Somit wird auf Techniken der eingebetteten Systeme in diesem Bereich aufgebaut. Dabei spielen Verfahren der Ablaufplanung für Realzeitsysteme eine wesentliche Rolle. Die Größenordnungen der Zeiten sind in unserem Fall andere als bei Realzeitsystemen üblich, dennoch können deren Methoden adaptiert werden. Wesentliche Erweiterungen gegenüber dem Stand der Technik sind aufgrund der mehrkriteriellen Bewertung heterogener Ausführungsplattformen erforderlich. Im Laufe der zweiten Phase wird im Teilprojekt eine Entwicklung von der Compiler-orientierten Optimierung von einzelnen Threads hin zu einer intensiven Betrachtung der Ausführung von Aufgaben auf parallelen Systemen stattfinden.

3.3 Bisherige Entwicklung des Teilprojekts

Das Ziel von A3 ist die Auswahl geeigneter Lernverfahren zur Vorhersageoptimierung auf großen biomedizinischen Datensätzen. Dabei treten Ressourcenbeschränkungen auf, sowohl in Bezug auf einzelne rechenintensive Lernverfahren als auch auf Grund des Vergleichs von sehr vielen verschiedenen Lernverfahren für ein spezifisches Problem. Ziel ist daher, Lernverfahren bzgl. ihres Ressourcenbedarfs zu evaluieren und zu optimieren. Die Anwendungsdomäne richtet sich gezielt auf Lernverfahren, welche in der Programmiersprache R implementiert sind und zur Analyse hochdimensionaler biomedizinischer Datensätze dienen. Die Optimierung dieser Verfahren geschieht im Hinblick auf die Reduktion ihres Ressourcenbedarfs.

Die Hauptziele der ersten Phase waren die Bestimmung des Umfangs der Ressourcennutzung von R, die Analyse des Optimierungspotenzials von Lernverfahren in Bezug auf Ressourcennutzung, die Entwicklung von effizienten Lernverfahren, bei denen der Ressourcenbedarf explizit modelliert bzw. berücksichtigt wird,

und die Evaluation des Einsparpotenzials.

Hauptaugenmerk in der zweiten Phase ist, neben der Optimierung einzelner Lernverfahren, eine ressourcengewahre Modellwahl. Die Hauptziele der zweiten Phase sind die Konstruktion und der Vergleich von Algorithmen zur automatisierten Modellwahl auf großen Datensätzen, die Entwicklung von Ablaufplanungsverfahren, die die Algorithmen zur Modellwahl bezüglich ihrer Ressourcennutzung optimieren, und die Erarbeitung von Methoden zur Modellwahl für klinisch oder genetisch vordefinierte Untergruppen von Patienten.

3.3.1 Bericht und Stand der Forschung

Bestimmung des Umfangs der Ressourcennutzung von R: Die Programmiersprache R hat sich als Standard in der Statistik durchgesetzt. R ist eine Sprache mit dynamischen Eigenschaften. Diese fördern die effiziente Entwicklung von statistischen Verfahren, jedoch wirkt sich diese Flexibilität negativ auf die Laufzeit und den Speicherverbrauch aus, was besonders kritische Aspekte sind, da bzgl. der Ressourcenbeschränkungen aktuell keine effiziente Ausführungsumgebung für R existiert. R Programme werden über einen in C implementierten Interpreter ausgeführt. Um die Ressourcennutzung von Lernverfahren, vor allem bei der Verarbeitung von großen Datenmengen, zu optimieren, war es zunächst notwendig ihre charakteristischen Eigenschaften zu analysieren, um aufzuzeigen, an welchen Stellen Optimierungspotenzial besteht.

Die R-Ausführungsumgebung bietet bereits Werkzeuge zur Analyse von R-Programmen, z.B. Rprof [130]. Diese Werkzeuge sind jedoch nur zur oberflächlichen Analyse des Programmverhaltens geeignet. So ist es z.B. nicht möglich, interne Funktionen, die zur Ausführung eines R-Programms benötigt werden (wie z.B. den Ressourcenbedarf des Speichermanagements), zu analysieren. Daher wurde in der ersten Phase des Projekts A3 die Entwicklung eines Werkzeugs namens *TraceR*[1] zur Analyse des Ressourcenverbrauchs von R spezifiziert, begleitet und evaluiert. Als Basis diente ein Prototyp, welcher von der Purdue University entwickelt wurde[2]. Dadurch konnten wir mit anerkannten R-Experten wie Jan Vitek (Purdue University) Kooperationen aufbauen. Für eine aktive Zusammenarbeit wurden und werden immer noch in regelmäßigen Arbeitstreffen aktuelle Forschungsergebnisse ausgetauscht, wie z.B. beim Dynamic Languages for Scalable Data Analytics Workshop (DALI)[3], zudem finden wöchentliche Telefonkonferenzen statt. Mit TraceR ist es möglich, auch interne Funktionen des R-Interpreters zu analysieren. Darüber hinaus integriert TraceR die Messungen in die Codebasis der Ausführungsumgebung, was die Präzision von Messungen stark erhöht. TraceR kann sowohl zur Analyse des Laufzeitverhaltens genutzt werden, um Informationen wie Aufrufhäufigkeiten pro Funktion oder die Art und Anzahl der Übergabeparameter zu bestimmen, als auch zur Analyse des Speicherverbrauchs. Die Analysen sind detailliert aufgebaut und reichen vom Profiling des gesamten Speicherverbrauchs bis zur Analyse des Ressourcenbedarfs für einzelne Datenstrukturen. TraceR wurde erfolgreich eingesetzt und Schwachstellen auf Ebene der R-Programmiersprache konnten aufgedeckt werden [A3/151]. Darüber hinaus hat TraceR bei den R-Entwicklern für großes Interesse gesorgt und wurde z.B. beim Bioconductor Developers' Meeting 2013[4] und in der Bioinformatik-Arbeitsgruppe der Universität Ulm vorgestellt.

Zusätzlich zur Analyse des Ressourcenbedarfs auf Anwendungsebene wurden in einem gemeinsamen Arbeitspaket mit dem Projekt A4 auch Analysen auf Ebene der Ausführungsplattform durchgeführt und speichergewahre Compileroptimierungen für C-Programme entwickelt [A3/148, A3/156]. So wurde in diesem Bereich eine Dissertation abgeschlossen [A3/156], in der demonstriert wurde, wie eine *Worst-Case Execution Time* (WCET)-gewahre Kompilierung auch dann eingesetzt werden kann, wenn kein WCET-gewahres Compiler-Backend vorhanden ist. Diese Techniken sind teilweise auch für die Optimierung von R anwendbar.

Die Bestimmung des Umfangs der Ressourcennutzung von R-Programmen wurde ergänzt durch die Bestimmung des Umfangs der Ressourcennutzung von C-Programmen, wie sie als Teil des Teilprojekts A4 beschrieben und in der Dissertation von D. Cordes [A3/154] ausführlich erläutert wird. Mit einem ebenfalls Profiling-basierten Ansatz wurden hier Laufzeit- und Kommunikationsverhalten von einzelnen C-Programmabschnitten bestimmt. Es konnte gezeigt werden, dass eine Parallelisierung unter Ausnutzung der verfügbaren Informationen über die Prozessorplattform detaillierte Wechselwirkungen zwischen den Bewertungen auf verschiedenen Kenngrößen liefern kann. Die Ergebnisse der Arbeit und auch der daraus entstandenen Publikationen [A3/146, A3/147] können für unsere zweite Phase, in der wir die parallele Ausführung von Lernverfahren bzgl. ihres Ressourcenbedarfs optimieren wollen, genutzt werden.

Analyse des Optimierungspotenzials des Ressourcenbedarfs: Eine allgemeine Evaluation der Sprache R wurde bereits in [127] durchgeführt. Als Programmeingaben wurden überwiegend synthetisch erzeugte Daten

[1] TraceR: https://github.com/allr/tracer
[2] Reactor: http://r.cs.purdue.edu
[3] DALI Workshop: http://www.dynali.org/
[4] Bioconductor Developers' Meeting: https://sites.google.com/site/eurobioc2013/

verwendet. Bei unserer Analyse haben wir uns hingegen explizit auf Lernverfahren in Kombination mit Eingabedatensätzen [88] fokussiert, welche eine große Anzahl an Beobachtungen oder Variablen enthalten. Da die Eingabedatensätze den Ressourcenbedarf eines Programms stark beeinflussen, konnte somit sichergestellt werden, dass aussagekräftige Ergebnisse erzeugt wurden. Sowohl die Lernverfahren als auch die verwendeten Eingabedatensätze und weitere Benchmarks wurden öffentlich verfügbar gemacht[5]. Durch den Austausch dieser Sammlung von statistischen Anwendungen, welche einen hohen Speicher- und Laufzeitbedarf aufweisen, mit anderen Wissenschaftlern, konnten weitere Kooperationen mit internationalen Partnern geschlossen werden, wie z.B. mit Alex Bertram, dem Entwickler von Renjin [108], einer Java-basierten Variante des R-Interpreters. Neben einigen Arbeitstreffen wurde Herr Bertram ebenfalls für einen Fachvortrag eingeladen.

Das Laufzeitverhalten und der Speicherverbrauch der R-Benchmarks wurde auf Basis von TraceR analysiert und bzgl. des Optimierungspotenzials in verschiedene Kategorien unterteilt [A3/151]. Die hierzu verwendeten maschinellen Klassifikationsverfahren sind ebenfalls öffentlich verfügbar[6]. Bei der Analyse des Laufzeitverhaltens besteht die Kategorie mit dem geringsten Optimierungspotenzial aus externen Codeteilen. In R ist es z.B. möglich, C- oder Fortran-Bibliotheken anzubinden. Da diese außerhalb des R-Interpreters ausgeführt werden und insgesamt bereits einen geringen Ressourcenverbrauch aufweisen, stellen sie kein lohnendes Optimierungspotenzial dar. Die Kategorie mit dem höchsten Optimierungspotenzial stellen Interpreter-interne Funktionen dar, wie z.B. das Speichermanagement oder der Prozeduraufruf. Bei der Analyse des Speicherverbrauchs wird das Optimierungspotenzial über die bei der Ausführung des Programms verwendeten Datenstrukturen bestimmt. Hier belegen die Vektor-Datenstrukturen, die bei den analysierten Programmen über die Eingabedaten erzeugt wurden, über 50% des insgesamt angeforderten Speichers. In R wird jeder Wert durch eine Vektordatenstruktur repräsentiert. Durch unsere Analysen konnten wir zeigen, dass über die Hälfte der erzeugten Vektordatenstrukturen nur aus einem einzelnen Wert bestehen und durch die Einführung von skalaren Datentypen ein hohes Optimierungspotenzial besteht. Aber auch auf der Seite der Interpreter-internen Datenstrukturen besteht ein hohes Optimierungspotenzial. Diese Datenstrukturen werden für die Ausführung des Programms benötigt und sind für fast die Hälfte aller Speicheranforderungen verantwortlich. Dies zeigt eine Schwachstelle im R-Ausführungsmodell, da in einer effizienten Ausführungsumgebung der Großteil des Speichers für Eingabedaten und deren Verarbeitung dienen sollte. Dieser Overhead kann durch die Einführung von ressourcensparenden Datenstrukturen oder auch durch neue Ausführungsumgebungen der Sprache optimiert werden. In [143] wurde gezeigt, dass durch die Einführung von skalaren Datenstrukturen Speicher eingespart werden kann. Hier wurden jedoch nicht explizit Lernverfahren in Situationen mit hohen Ressourcenanforderungen betrachtet.

Bei dynamisch typisierten Sprachen wie R sind Informationen über verwendete Datentypen erst bei der Ausführung vorhanden und R-Programme können sich zur Laufzeit selbst modifizieren. Deshalb wird hier die dynamische Kompilierung bevorzugt eingesetzt, z.B. in Form von *Just-in-Time*-Compilern [105, 113]. Für R wurde von uns auf der UseR! Konferenz eine solche dynamische Kompilierungsstrategie vorgestellt [A3/150], die auf der Java Virtual Machine (JVM) basiert. Ein JVM-basierter Ansatz bietet hierbei den Vorteil, dass verschiedene dynamische Compiler zur Konfiguration zur Verfügung stehen. Da R-spezifische Optimierungen durchgeführt werden sollen, wie z.B. die Transformation von Vektoroperationen in hocheffizienten Maschinencode, muss der verwendete dynamische Compiler anpassbar sein. Ein besonders einfach anpassbarer, in Java implementierter dynamischer Compiler ist Graal[7]. Da unsere Ergebnisse zur Analyse des Optimierungspotenzials des Ressourcenbedarfs von Lernverfahren nicht nur konkrete Schwachstellen aufdecken, sondern darüber hinaus die Entwicklung von neuen alternativen R-Ausführungsumgebungen unterstützen, konnten wir ebenfalls Kooperationen zu Entwicklern von alternativen Plattformen für R aufbauen, wie zu Thomas Würthinger, einem der Entwickler von Graal, von der Johannes Kepler-Universität (JKU) in Linz. Zum Austausch von aktuellen Forschungsergebnissen wurden wir zu einem gemeinsamen Arbeitstreffen an die JKU eingeladen. Zudem findet hier ein regelmäßiger Austausch von Forschungsergebnissen statt.

Ressourceneffizientes Lernen in der Medizin: Unser Ziel hier ist die Entwicklung und Untersuchung von Methoden zur Modellwahl speziell im Hinblick auf medizinische Fragestellungen unter Ressourcenbeschränkungen. Für Probleme mit einer sehr großen Anzahl an Beobachtungen oder Variablen soll aus einer umfangreichen Menge von rechenintensiven Lernverfahren dasjenige mit der besten Vorhersageleistung ausgewählt werden.

Hierfür sind üblicherweise Methoden des *Meta-Learning* sehr geeignet. Diese generieren zunächst Wissen darüber, welche Güte bestimmte Lernverfahren auf verschiedenen Lernproblemen besitzen. Dann wird ein Zusammenhang zwischen den Charakteristika der Daten eines Lernproblems und den Charakteristika der

[5] R-Benchmark-Sammlung: https://github.com/allr/benchR
[6] UCI Eingabedaten und Lernverfahren: https://github.com/allr/benchR/tree/master/MachineLearningAlg/mlrUCI
[7] Graal Compiler: http://openjdk.java.net/projects/graal/

Lernverfahren unterstellt. Mit den Verfahren des Meta-Learning wird dieser Zusammenhang gelernt bzw. modelliert. Für ein neues Lernproblem werden zuerst die Charakteristika der Daten gemessen, dann wird ein geeignetes Lernverfahren vorgeschlagen. *Stacked Generalisation* erweitert dieses Vorgehen noch um die Kombination verschiedener geeigneter Lernverfahren zu einem Gesamtverfahren. Ein Beispiel für die Einbindung vorheriger Analysen ist die automatische Verbesserung von gesamten Prozessen der Wissensentdeckung in der RapidMiner-Extension IDA[8]. Hier werden für ein Vorhersageproblem auf einem gegebenen Datensatz Analyseprozesse vorgeschlagen, die auf ähnlichen Vorhersageproblemen eine hohe Güte erreicht haben. All diese Methoden sind jedoch nur erfolgreich, wenn für einen neuen Datensatz genügend ähnliche Datensätze zur Verfügung stehen, für die das beste Lernverfahren bereits identifiziert ist. Wir betrachten Fragestellungen, in denen dies nicht der Fall ist.

Ein wichtiges Einsatzgebiet von maschinellen Lernverfahren in der Medizin ist die Entwicklung von Prognosemodellen und von Klassifikationsmethoden, welche die Therapiewahl für erkrankte Menschen verbessern können. Unter dem Schlagwort „personalisierte Medizin“ werden moderne Methoden subsumiert, die Patientinnen und Patienten anhand genetischer Merkmale hinsichtlich Diagnose und Therapie in Untergruppen einteilen. Beispielsweise soll ein bestimmtes Medikament bzw. eine Therapie gezielt Patienten gegeben werden, bei denen mit großer Wahrscheinlichkeit eine positive Wirkung und geringe Nebenwirkungen auftreten. Das Ansprechen auf die Behandlung lässt sich dabei häufig nicht binär messen, sondern wird durch die Zeit bis zum Auftreten eines definierten Ereignisses ausgedrückt. Dadurch treten unvollständige Beobachtungen in der Zielvariable auf, was durch Methoden der Überlebenszeitanalyse berücksichtigt wird [122].

In der Molekularbiologie ist in den letzten zehn Jahren die Menge an zur Verfügung stehenden genetischen und genomischen Daten stark gestiegen, zuallererst durch neue molekularbiologische Techniken, angefangen von Microarray-Experimenten Ende der 90er Jahre hin zu RNAseq-Daten und proteomischen genomweiten Messungen heutzutage. Für Krebspatienten ist eine der wichtigsten Fragen die Schätzung der Überlebenszeitfunktion in Abhängigkeit von individueller Genetik und Therapieentscheidung. Zur Analyse hochdimensionaler Überlebenszeitdaten stehen in R eine Vielzahl an Algorithmen und Modellen bereit. Die Task View zur Überlebenszeitanalyse[9] des Comprehensive R Archive Network (CRAN) listet dazu fast 200 Pakete. Darüber hinaus stellt das BioConductor-Projekt [118] zusätzliche ca. 750 Bioinformatik-Pakete zur Verarbeitung, Annotation und Analyse von Hochdurchsatzdaten bereit.

Damit ist eine einzigartige Fülle an Möglichkeiten in R geboten, aber die Auswahl eines Modells auf diesen Daten gestaltet sich als sehr herausfordernd. Besonders auf Grund der hohen Dimensionalität – je nach Technologie zwischen $25k$ und $500k$ genetischen Einzeldaten – können einzelne Analysestrategien, bestehend aus Vorverarbeitung der Rohdaten, eventueller Dimensionsreduktion, Modellanpassung und Modellvalidierung, mehrere Stunden Rechenzeit in Anspruch nehmen. Eine saubere Analyse erfordert zusätzlich die Trennung in Trainings- und Testmengen, um unverzerrte Ergebnisse zu erhalten. Da diese Aufteilung stochastisch ist, ist zudem eine mehrfache Wiederholung dieser Prozedur nötig, um die Varianz der Vorhersagegüte zu reduzieren. Weiterhin muss diese Analysekette für jede Aufteilung zur Vermeidung von Überoptimismus komplett neu durchgeführt werden. Wenn mehrere Optionen für Vorverarbeitung, Filter und konkurrierende Modelle in Betracht gezogen und verglichen werden, entsteht leicht eine sequentielle Rechenzeit von mehreren Jahren.

Für solche Berechnungen kann die Unabhängigkeit der Trainings-Aufteilung zur Parallelisierung ausgenutzt werden. In der ersten Phase wurde für den Fall einer naiven Parallelisierung ein Paket für die Ausführung auf Hochperformanz-Clustern erstellt, da die Parallelisierung auf Mehrprozessormaschinen oder MPI-Clustern, wie durch das Paket *parallel* unterstützt, den Anforderungen an Laufzeit und Speicherverbrauch nicht gerecht wurde. Im Teilprojekt A3 wurden deshalb die Pakete *BatchJobs* und *BatchExperiments* [A3/153] erstellt. Ersteres stellt die Infrastruktur zur Parallelisierung auf verwalteten Hochperformanz-Clustern durch eine Abstraktionsebene für verschiedene Scheduler bereit. Das Interface zum Definieren von Jobs und zum Einsammeln von Ergebnissen orientiert sich dabei an dem populären Map-Reduce Paradigma. Das zweite Paket, BatchExperiments, erweitert BatchJobs um eine Abstraktionsschicht zum einfachen, strukturierten und reproduzierbaren Definieren statistischer Experimente und Vergleichsstudien.

Werden die Parameter der Vorverarbeitung, des Filterns und der Modelle mittels einer einfachen Gittersuche kreuzvalidiert eingestellt, müssen oft schon für kleinere Studien mehr als 10^5 Modelle angepasst werden. Aufgrund der zahlreichen Modellierungsoptionen, sowie deren Ressourcenanforderungen, ist ein erschöpfender Vergleich von Methoden alleine mit Hilfe von naiver Parallelisierung daher nicht möglich. Zusätzlich ist eine sinnvolle Einschränkung auf eine Teilmenge an Methoden erforderlich. In solchen Situationen ist eine automatisierte Modellwahl von großer Hilfe, bei der für einen neuen Datensatz ein vorher festgelegtes Gütemaß optimiert wird, zum Beispiel die Vorhersagegüte bei Prognosemodellen. Solch ein Verfahren liefert eine Menge von interessanten Modellen, aus welchen in einem anschließenden Schritt Informationen extrahiert

[8] Intelligent Discovery Assistant: http://www.e-lico.eu/ida-extension.html
[9] CRAN Task View Survival Analysis: http://cran.r-project.org/web/views/Survival.html

werden können, zum Beispiel kann ein einziges Modell selektiert oder konstruiert werden. Ein frühes Beispiel für die effiziente Auswahl der n besten Modelle aus einer gegebenen Anzahl von insgesamt $m > n$ Modellen ist der *Sequential Sampling*-Algorithmus von Scheffer und Wrobel [135]. Unter der Annahme, dass mit steigender Anzahl an Auswertungen M die Güte der untersuchten Lernverfahren beliebig genau bestimmt werden kann, wird ein Algorithmus angegeben, der schnell für vorgegebene Schranken für die statistische Sicherheit die besten Modelle bestimmt. Auch Methoden, die unter den Begriff *Multi-Armed Bandit* fallen, sind verwandt [107]. Da hier jedoch verschiedene Parametrisierungen von statistischen Modellen berücksichtigt werden sollen, müssen unendlich viele Modelle betrachtet werden, so dass sich beide Ansätze nicht einfach übertragen lassen.

Eine Alternative ist durch das *Iterated F-Racing* [126] gegeben, bei dem zunächst viele Verfahren auf wenigen Trainingsaufteilungen evaluiert werden. Die Exploration des Modellraums erfolgt mittels statistischer Verteilungen, welche schrittweise an guten Verfahren im Raum zentriert werden, während die *Exploitation* durch eine Minderung der Verteilungsvarianz bewerkstelligt wird. Dieses Verfahren wurde in der ersten Phase zur Modellierung von Überlebenszeiten für vier Lungenkrebsdatensätze angewendet [A3/152]. Dabei ließen sich durch die Analyse des Optimierungspfades nicht nur wichtige und neue Erkenntnisse über die verwendeten Filter und Modelle in einer konkreten Datensituation ableiten, auch über die Datenqualität und die hohe Heterogenität in Bezug auf verschiedene Datensätze zu derselben medizinischen Fragestellung konnten richtungsweisende Informationen gesammelt werden. Insbesondere konnte eine pro Datensatz individuelle Modellwahl die Vorhersagequalität relevant verbessern.

Für die Risikobewertung von Einflussfaktoren auf Überlebenszeiten ist das Cox-Regressionsmodell als Referenzmodell anerkannt. Einzelne Variablen tragen hier multiplikativ zur Risikobewertung eines Patienten bei. Für Modelle mit hochdimensionalen genetischen Variablen ist die Variablenselektion zentraler Bestandteil der Modellbildung, entweder explizit etwa durch vorgeschaltete Filtermethoden oder implizit wie bei der Lasso-Cox-Regression als einem populären Vertreter der penalisierten Likelihood-Verfahren [134, 119].

Eine univariate Selektion als Filtermethode für hochdimensionale Daten haben wir in [A3/145] untersucht. Bei Vorliegen von zwei Patientengruppen (z.B. krank vs. gesund oder Subtypen einer Krankheit) wurden Maßzahlen in ihrer Fähigkeit verglichen, um relevante Teilgruppen von Patienten mit abweichenden Werten in einzelnen Variablen zu identifizieren. Eine Verzerrung der Ergebnisse durch die Vorschaltung der Filtermethode wird vermieden, indem die Zielgröße bei diesem Schritt nicht berücksichtigt wird.

Es sind zahlreiche, teilweise neue Methoden für die multivariate Modellierung von Überlebenszeiten verfügbar. Ein erster Vergleich für ausgewählte populäre Methoden zur Prognose für Brustkrebspatientinnen wurde in [114] erstellt. Dieser wurde von uns in [A3/149] unter anderem um die Anreicherung der Eingabedaten durch biologische Annotationen erweitert. Eine weitere vielversprechende neue Methode zur Variablenselektion wurde in Zusammenarbeit mit Teilprojekt C1 entwickelt. Hierbei zeichnet sich das entwickelte Modell durch Stabilität der ausgewählten Genmengen bei verschiedenen Trainings-Aufteilungen aus, was die Interpretierbarkeit des Modells für die medizinische Anwendung deutlich erhöht. Ermöglicht wurde diese Arbeit durch Verwendung von BatchExperiments. Eine weitere Konsequenz unserer Arbeiten war, dass sich eine kombinierte Betrachtung von Güte und Laufzeit zur Modellwahl bei gegebener Varianz als schwierig gestaltet. Der Fokus sollte also zuerst auf der Modellwahl liegen, denn nur in einer Klasse von Modellen mit überzeugender Güte ergibt eine Berücksichtigung der Laufzeit während der Modellierung Sinn.

Die Modellwahl für eine spezielle Untergruppe von Patienten fügt zwei weitere Ebenen der Komplexität hinzu, durch die Bestimmung der Untergruppe und durch die Art der Berücksichtigung anderer Untergruppen bei der Modellbildung. Untergruppen werden dabei auf verschiedene Arten gebildet. In einer Dissertation haben wir Lungenkrebspatienten nach klinischen Merkmalen wie Tumorart oder Raucherstatus gruppiert [A3/155], in einer Masterarbeit nach der Klinik, in der sie behandelt wurden. Die zu Grunde liegende Idee der gewichteten Likelihood stammt von [109] und wurde von uns auf den Kontext der Modellierung von Überlebenszeiten übertragen. Zur Bestimmung der Gewichte muss ein zusätzliches hochdimensionales Klassifikationsproblem gelöst werden, was die Laufzeiten der Algorithmen wiederum stark erhöht. In der zweiten Phase soll nun zudem die Einteilung der Patienten auch nach genetischen Gesichtspunkten erfolgen.

Evaluation des Einsparpotenzials: Zur effizienteren Ausführung von R-Programmen wurden verschiedene Ansätze entwickelt, welche u.a. auf dem DALI-Workshop[3] präsentiert wurden, bei dem unser Projekt A3 ebenfalls vertreten war. Diese unterscheiden sich vorwiegend in der Wahl der Ausführungsplattform. So gibt es Ansätze, welche den R-Interpreter erweitern, indem sie kommerzielle, hocheffiziente Bibliotheken anbinden[10] oder bestehende Teile des Interpreters optimieren [143, 128, 140, 129] oder Ansätze, die eine komplett neue Ausführungsumgebung wählen [108, 117, 121, 137, 144, 131].

[10]Revolution Analytics R Enterprise: http://www.revolutionanalytics.com/revolution-r-enterprise

Die Ansätze, welche auf einer anderen Ausführungsplattform als der des originalen R-Interpreters aufbauen, haben starke Kompatibilitätsprobleme, weil viele der bestehenden R-Programme und R-Pakete noch nicht auf ihnen ausgeführt werden können. Dabei hängen R-Programme sehr stark von den öffentlich verfügbaren R-Paketen und Bibliotheken ab[11]. Bei der Wahl einer neuen Ausführungsplattform ist auch das im R-Interpreter vorhandene Ausführungsmodell und damit die Ausführung der Programmiersprache R erneut in der Sprache der Ausführungsumgebung zu implementieren. Dieser Ansatz erfordert einen extrem hohen Programmieraufwand, da die Eigenschaften von R vollständig zu verstehen und zu implementieren sind und aktuell die Spezifikation der Sprache nur aus ihrer Originalimplementierung in C besteht. Aus diesen Gründen ist die Entwicklung einiger der genannten Ansätze bereits nach ca. einem Entwicklungsjahr gestoppt worden [121, 131].

Die Ansätze, welche auf dem originalen R-Interpreter aufbauen, haben die oben genannten Kompatibilitätsprobleme nicht. Einige dieser Ansätze konzentrieren sich vorwiegend auf die Optimierung von Schleifenstrukturen [140, 143, 129]. Da es jedoch in R höchst ineffizient ist, Schleifen einzusetzen, enthalten auch unsere Lernverfahren nur wenige R-Schleifen. Das Optimierungspotenzial für diese Ansätze ist daher eher gering und geht z.B. für [140, 128] bis zu einem Faktor von ca. 1,5. Für den Ansatz aus [143] war uns ein Vergleich nicht möglich, da die Software nicht öffentlich zur Verfügung steht, hier wird jedoch von einem Faktor 3,5 der Laufzeitbeschleunigung für schleifenbasierte Programme im Vergleich zu [140] berichtet.

Zur Zeit der Planung der ersten Antragsphase erschien es lohnenswert, R-Programme mittels einer Übersetzung nach C zu beschleunigen. Mit diesem Ziel wurden inzwischen für R viele in C geschriebene Bibliotheken manuell erstellt. Die so erreichten Beschleunigungen bestätigen die Ideen der Planungen, lassen aber gleichzeitig den Weg einer automatischen Übersetzung weniger lohnend erscheinen. Es ließen sich lediglich noch die Grenzen der Leistungsfähigkeit einer automatischen Übersetzung ausloten. Vorteile für die von uns untersuchten Lernverfahren können aber nicht mehr erwartet werden. Zusätzlich ist zu berücksichtigen, dass während der ersten Förderperiode die Bedeutung von Multicore Prozessoren dramatisch zugenommen hat. Damit erhält die parallele Programmierung ein erheblich höheres Gewicht. Unter den Möglichkeiten, parallele Programmierung zu unterstützen, fokussieren wir uns auf die konkreten Formen der Parallelität, wie sie bei den betrachteten Lernverfahren möglich sind. Es hat sich gezeigt, dass eine solche Unterstützung generalisiert und auch in anderen Teilprojekten des SFBs genutzt werden kann.

Zur allgemeinen parallelen Ausführung von R-Programmen gibt es bereits im CRAN Task View „High-Performance and Parallel Computing“ einige Pakete[12]. Diese Pakete unterscheiden sich bzgl. der Parallelisierung nach *low-level* und *high-level* Parallelisierungstechniken. Low-level Parallelisierungstechniken arbeiten über das Einbinden von Bibliotheken wie z.B. der BLAS-Bibliothek, welche effiziente Vektorinstruktionen enthält. Auf der high-level Ebene unterscheidet man zwischen spezifischen Algorithmen, die parallel implementiert wurden, und R-Paketen, bei denen der Benutzer die Daten und Programmteile manuell über das Aufrufen von spezifischen Funktionen zur Parallelisierung einteilen muss. Hierunter fallen auch die von uns entwickelten R-Pakete BatchJobs und BatchExperiments. Die verwendeten Technologien basieren dabei auf Bibliotheken wie OpenMP, Pthreads oder MPI [136].

Darüber hinaus gibt es auch Arbeiten zur automatischen Parallelisierung von R-Schleifenstrukturen [129]. Da unsere Anwendungen aus Effizienzgründen jedoch kaum R-Schleifen beinhalten, sind diese Verfahren weniger geeignet. Das RapidMiner-Framework besitzt ebenfalls Verfahren zur parallelen Ausführung von Lernverfahren, hier wurden z.B. die Kreuzvalidierungsschritte und Parameteroptimierungen parallelisiert. Eine weitere Arbeit zu parallel ausgeführten Statistikanwendungen optimiert durch das Einführen von verteilten Array-Datenstrukturen die Laufzeit. Jedoch eignet sich diese Art der Optimierung hauptsächlich für Anwendungen mit dünn besetzten Matrizen. Hier muss zudem der Benutzer den Inhalt seiner Datenstrukturen manuell aufteilen [142]. In [125] werden Techniken aus dem Bereich des maschinellen Lernens eingesetzt, um die Verteilung von R-Programmen im Rechenclustern zu optimieren. Hierbei wird ein Neuronales Netz zur Abschätzung der Kosten von Funktionen eingesetzt. Der Einsatz von Profiling-basierten Kostenmodellen zur Optimierung von Klassifikationsverfahren spielt ebenfalls im Bereich der eingebetteten Systeme eine wichtige Rolle. So wurden u.a. in [12] die Wechselwirkungen von Energiebedarf, Zeitbedarf und Ergebnisqualität für Klassifikationsverfahren zur Auswertung von Sensordaten untersucht.

Keines der genannten Verfahren betrachtet die Optimierung des Ressourcenbedarfs von parallel ausgeführten R-Programmen unter Berücksichtigung von detaillierten Hardwarekostenmodellen, auf die wir uns in der zweiten Phase konzentrieren werden. Darüber hinaus werden im Vergleich zu den oben genannten Verfahren in A3 auch die Parameter der Anwendung zur Modellwahl bei der Ressourcenanalyse berücksichtigt, um so die Basis für eine effiziente Ablaufplanung zu schaffen.

[11] CRAN: http://cran.r-project.org
[12] Task View Parallel Computing: http://cran.r-project.org/web/views/HighPerformanceComputing.html

Literatur

[105] G. Almasi und D. A. Padua. „MaJIC: A Matlab Just-In-Time Compiler“. In: *Languages and Compilers for Parallel Computing*. 2001, S. 68–81.

[106] D. P. Anderson. „BOINC: A System for Public-Resource Computing and Storage“. In: *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*. 2004, S. 4–10.

[107] P. Auer, N. Cesa-Bianchi und P. Fischer. „Finite-time Analysis of the Multiarmed Bandit Problem“. In: *Machine Learning* 47.2-3 (2002), S. 235–256.

[108] A. Bertram. *Renjin: JVM-based Interpreter for the R Language for Statistical Computing*. 2014. URL: http://www.renjin.org.

[109] S. Bickel. „Learning under Differing Training and Test Distributions“. Diss. Universität Potsdam, 2009.

[110] E. Bilal u. a. „Improving Breast Cancer Survival Analysis through Competition-Based Multidimensional Modeling“. In: *PLoS Computational Biology* 9.5 (2013).

[111] H. Binder, A. Allignol, M. Schumacher und J. Beyersmann. „Boosting for High-Dimensional Time-To-Event Data with Competing Risks“. In: *Bioinformatics* 25.7 (Feb. 2009), S. 890–896.

[112] R. D. Blumofe und C. E. Leiserson. „Scheduling Multithreaded Computations by Work Stealing“. In: *J. ACM* 46.5 (Sep. 1999), S. 720–748.

[113] C. F. Bolz, A. Cuni, M. Fijalkowski und A. Rigo. „Tracing the Meta-Level: PyPy's Tracing JIT Compiler“. In: *Proceedings of the 4th workshop on the Implementation, Compilation, Optimization of Object-Oriented Languages and Programming Systems*. Genova, Italy, 2009, S. 18–25.

[114] H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, O. Borgan, A. Frigessi und O. C. Lingjærde. „Predicting Survival from Microarray Data - A Comparative Study“. In: *Bioinformatics* 23.16 (2007), S. 2080–2087.

[115] T. D. Braun u. a. „A Comparison of Eleven Static Heuristics for Mapping a Class of Independent Tasks onto Heterogeneous Distributed Computing Systems“. In: *Journal of Parallel and Distributed computing* 61.6 (2001), S. 810–837.

[12] D. Chu, N. D. Lane, T. T.-T. Lai, C. Pang, X. Meng, Q. Guo, F. Li und F. Zhao. „Balancing Energy, Latency and Accuracy for Mobile Sensor Data Classification“. In: *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. SenSys '11. ACM. Seattle, Washington: ACM, 2011, S. 54–67.

[116] R. I. Davis und A. Burns. „A Survey of Hard Real-time Scheduling for Multiprocessor Systems“. In: *ACM Comput. Surv.* 43.4 (Okt. 2011), 35:1–35:44.

[117] *FastR*. Purdue University, 2014. URL: https://github.com/allr/fastr.

[118] R. C. Gentleman u. a. „Bioconductor: Open Software Development for Computational Biology and Bioinformatics“. In: *Genome Biology* 5.10 (Sep. 2004), R80.

[119] I. Guyon und A. Elisseeff. „An Introduction to Variable and Feature Selection“. In: *Journal of Machine Learning Research (JMLR), Special Issue on Variable and Feature Selection* 3 (2003), S. 1157 –1182.

[120] D. R. Jones, M. Schonlau und W. J. Welch. „Efficient Global Optimization of Expensive Black-Box Functions“. In: *Journal of Global Optimization* 13.4 (1998), 455–492.

[121] M. J. Kane und J. W. Emerson. *Not Quite R for the Parrot VM*. Yale University, 2011. URL: https://github.com/NQRCore.

[122] J. P. Klein und M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. Statistics for Biology and Health. Springer, 2003.

[123] R. Kumar, D. M. Tullsen, P. Ranganathan, N. P. Jouppi und K. I. Farkas. „Single-ISA Heterogeneous Multi-Core Architectures for Multithreaded Workload Performance“. In: *SIGARCH Comput. Archit. News* 32.2 (März 2004), 64pp.

[124] V. Kumar, D. Frampton, S. M. Blackburn, D. Grove und O. Tardieu. „Work-stealing Without the Baggage“. In: *Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications*. Tucson, Arizona, USA, 2012, S. 297–314.

[125] J. Li, X. Ma, K. Singh, M. Schulz, B. De Supinski und S. McKee. „Machine Learning Based Online Performance Prediction for Runtime Parallelization and Task Scheduling“. In: *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on.* Apr. 2009, S. 89–100.

[126] M. López-Ibánez, J. Dubois-Lacoste, T. Stützle und M. Birattari. „The irace Package, Iterated Race for Automatic Algorithm Configuration“. In: *IRIDIA, Université Libre de Bruxelles, Belgium, Tech. Rep. TR/IRIDIA/2011-004* (2011).

[127] F. Morandat, B. Hill, L. Osvald und J. Vitek. „Evaluating the Design of the R Language: Objects and Functions for Data Analysis“. In: *Proceedings of the 26th European conference on Object-Oriented Programming.* Beijing, China, 2012, S. 104–131.

[128] R. Neal. *pqR - A Pretty Quick Version of R.* University of Toronto, 2014. URL: https://github.com/radfordneal/pqR.

[129] F. Padberg und M. Mirold. „An Experimentation Platform for the Automatic Parallelization of R Programs“. In: *Software Engineering Conference (APSEC), 2012 19th Asia-Pacific.* 2012.

[130] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. 2014. URL: http://www.R-project.org.

[131] *Reactor - An R Virtual Machine Built on Top of VMKIT.* 2012. URL: http://r.cs.purdue.edu/hg/reactor.

[132] M. Reif, F. Shafait und A. Dengel. „Prediction of Classifier Training Time Including Parameter Optimization“. In: *KI 2011: Advances in Artificial Intelligence.* Springer, 2011, S. 260–271.

[133] J. N. v. Rijn u. a. „OpenML: A Collaborative Science Platform“. In: *Machine Learning and Knowledge Discovery in Databases.* Hrsg. von H. Blockeel, K. Kersting, S. Nijssen und F. Železný. Lecture Notes in Computer Science 8190. Springer Berlin Heidelberg, 2013, S. 645–649.

[134] Y. Saeys, I. Inza und P. Larrañaga. „A Review of Feature Selection Techniques in Bioinformatics“. In: *Bioinformatics* 23.19 (2007), S. 2507–2517.

[135] T. Scheffer und S. Wrobel. „Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling“. In: *Journal of Machine Learning Research* 3 (2002), S. 833–862.

[136] M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, L. Tierney und U. Mansmann. „State-of-the-art in Parallel Computing with R“. In: *Journal of Statistical Software* 47.1 (2009).

[137] J. Talbot, Z. DeVito und P. Hanrahan. „Riposte: A Trace-driven Compiler and Parallel VM for Vector Code in R“. In: *Proceedings of the 21st International Conference on Parallel Architectures and Compilation Techniques.* Minneapolis, Minnesota, USA: ACM, 2012, S. 43–52.

[138] C. Thornton, F. Hutter, H. H. Hoos und K. Leyton-Brown. „Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms“. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2013, 847–855.

[139] C. Tian, Y. Wang, F. Qi und B. Yin. „Decision Model for Provisioning Virtual Resources in Amazon EC2“. In: *Proceedings of the 8th International Conference on Network and Service Management.* Las Vegas, Nevada, 2013, S. 159–163.

[140] L. Tierney. *A Byte Code Compiler for R. Technical Report.* Department of Statistics und Actuarial Science University of Iowa, 2012.

[88] *UCI Machine Learning Repository.* 2014. URL: http://archive.ics.uci.edu/ml.

[141] J. Vanschoren und H. Blockeel. „A Community-based Platform for Machine Learning Experimentation“. In: *Machine Learning and Knowledge Discovery in Databases.* Springer, 2009, S. 750–754.

[142] S. Venkataraman, E. Bodzsar, I. Roy, A. AuYoung und R. S. Schreiber. „Presto: Distributed Machine Learning and Graph Processing with Sparse Matrices“. In: *Proceedings of the 8th ACM European Conference on Computer Systems.* Prague, Czech Republic: ACM, 2013.

[143] H. Wang, P. Wu und D. Padua. „Optimizing R VM: Allocation Removal and Path Length Reduction via Interpreter-level Specialization“. In: *Proceedings of the International Symposium on Code Generation and Optimization (CGO).* Orlando, Florida, 2014.

[144] T. Würthinger u. a. „One VM to Rule Them All“. In: *Proceedings of the 2013 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming & Software.* Indianapolis, Indiana, USA: ACM, 2013.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[A3/145] M. Ahrens u. a. „Detection of Patient Subgroups with Differential Expression in Omics Data: A Comprehensive Comparison of Univariate Measures“. In: *PLoS ONE* 8.11 (2013), e79380.

[A3/146] D. Cordes, M. Engel, O. Neugebauer und P. Marwedel. „Automatic Extraction of Pipeline Parallelism for Embedded Heterogeneous Multi-Core Platforms“. In: *Proceedings of the Sixteenth International Conference on Compilers, Architectures, and Synthesis for Embedded Systems (CASES 2013)*. Montreal, Canada, Okt. 2013.

[A3/147] D. Cordes, M. Engel, O. Neugebauer und P. Marwedel. „Automatic Extraction of Task-Level Parallelism for Heterogeneous MPSoC“. In: *Proceedings of the Fourth International Workshop on Parallel Software Tools and Tool Infrastructures (PSTI 2013)*. Lyon, France, Okt. 2013.

[A3/148] H. Falk und H. Kotthaus. „WCET-driven Cache-aware Code Positioning“. In: *Proceedings of the International Conference on Compilers, Architectures and Synthesis for Embedded Systems (CASES)*. Taipei, Taiwan, Okt. 2011, S. 145–154.

[A3/149] K. Kammers, M. Lang, J. G. Hengstler, M. Schmidt und J. Rahnenführer. „Survival Models with Preclustered Gene Groups as Covariates“. In: *BMC Bioinformatics* 12.1 (2011), S. 478.

[A3/150] H. Kotthaus, S. Plazar und P. Marwedel. „A JVM-based Compiler Strategy for the R Language“. In: *Abstract Booklet of the 8th International R User Conference (UseR!)* Nashville, Tennessee, USA, Juni 2012, S. 68.

[A3/151] H. Kotthaus, I. Korb, M. Lang, B. Bischl, J. Rahnenführer und P. Marwedel. „Runtime and Memory Consumption Analyses for Machine Learning R Programs“. In: *Journal of Statistical Computation and Simulation*. 2014.

[A3/152] M. Lang, H. Kotthaus, P. Marwedel, C. Weihs, J. Rahnenführer und B. Bischl. „Automatic model selection for high-dimensional survival analysis“. In: *Journal of Statistical Computation and Simulation* ahead-of-print (2014), 1–15.

Andere Veröffentlichungen

[A3/153] B. Bischl, M. Lang, O. Mersmann, J. Rahnenführer und C. Weihs. *Computing on High Performance Clusters with R: Packages BatchJobs and BatchExperiments*. Techn. Ber. 1. TU Dortmund, 2012.

[A3/154] D. A. Cordes. „Automatic Parallelization for Embedded Multi-Core Systems using High-Level Cost Models“. Diss. TU Dortmund, 2013.

[A3/155] C. Netzer. „Vorhersage der Überlebenswahrscheinlichkeit für Patientenuntergruppen mit hochdimensionalen Daten am Beispiel zweier Lungenkrebskohorten“. Diss. TU Dortmund, 2013.

[A3/156] S. Plazar. „Memory-based Optimization Techniques for Real-Time Systems“. Diss. TU Dortmund, Juli 2012.

3.4 Planung des Teilprojekts

Ziele

Ziel des Projekts ist die Entwicklung von Methoden zur Modellwahl unter Ressourcenbeschränkungen. Es werden Vorhersageprobleme mit sehr großen Anzahlen an Variablen betrachtet, für die viele Lernverfahren zur Verfügung stehen, die Modelle mit einer potenziell kompetitiven Vorhersageleistung generieren. Die Ressourcenbeschränkungen ergeben sich sowohl aus der Dimension der Daten als auch aus der großen Anzahl zu vergleichender Verfahren, die oft bereits einzeln laufzeitintensive Prozeduren darstellen. Zusätzlich müssen für viele Lernverfahren noch Parametereinstellungen optimiert werden.

Der Parameterraum für die Modellwahl besteht aus den verschiedenen Lernverfahren und den jeweils zugehörigen Parametern dieser Lernverfahren. Um in diesem Parameterraum eine geeignete Auswahl zu treffen, werden in diesem Projekt Methoden der automatisierten Modellwahl verwendet. Hierbei wird ein funktionaler Zusammenhang zwischen Parameterraum und Qualitätsbewertung unterstellt, der zum Beispiel durch ein Regressionsmodell beschrieben werden kann. Ein guter Algorithmus für die Modellwahl bewegt sich dann anhand der Analyse und Bewertung vorheriger Auswertungen von Modellen möglichst schnell

und effizient in die Regionen der interessantesten Modelle und zugehörigen Parametereinstellungen. Ein solcher Algorithmus ist durch die Anzahl auszuwertender Modelle eine laufzeitintensive Prozedur, welche eine Optimierung der Ausführung erfordert. Das Projekt hat drei Hauptziele:

- Es sollen Algorithmen zur automatisierten Modellwahl entwickelt und verglichen werden. Kandidaten für solche Algorithmen sind Methoden der modellbasierten Optimierung und des Racing. Ein Schwerpunkt wird die Verbesserung der Vorhersage von Überlebenszeiten von Patienten mit hochdimensionalen genetischen Messungen sein. Die Verfahren sollen auf öffentlich zugänglichen Datensätzen evaluiert werden.

- Mit Methoden der eingebetteten Systeme sollen Ablaufplanungsverfahren entwickelt werden, die die Algorithmen zur Modellwahl bezüglich ihrer Ressourcennutzung optimieren. Hinsichtlich der ressourceneffizienten Ausführung von Lernverfahren liegt ein Schwergewicht auf der Nutzung von Multicore-, Multiprozessor- und HPC-Clustersystemen. Die Ausführungsumgebung ist dabei charakterisiert durch eine Anzahl von Rechenknoten mit unterschiedlichen Kosten-, Performanz- und Ressourceneigenschaften. Für die Ausführung gibt es (weiche) Zeitschranken, da die Ergebnisse der Lernverfahren in einer vertretbaren Zeit verfügbar sein müssen. Ziel ist ein geeigneter Kompromiss zwischen der Qualität der ausgewählten Modelle und der Ressourcennutzung.

- Es sollen Methoden zur Modellwahl für klinisch oder genetisch vordefinierte Untergruppen von Patienten entwickelt werden. Bedingt durch die Heterogenität von Patientenkollektiven, insbesondere in der Krebsforschung, sind oft bestimmte Diagnose- und Therapieverfahren spezifisch für solche Untergruppen geeignet. Für eine Untergruppe wird zuerst die ressourcenoptimierte automatisierte Modellwahl einmal auf dem Gesamt-Patientenkollektiv und einmal nur auf der Untergruppe durchgeführt und die Vorhersagegüte verglichen. Dann wird mit statistischen Methoden für die anderen Patienten bestimmt, wie sie bei der Modellbildung zusätzlich gewinnbringend berücksichtigt werden können. Schließlich wird für die Modellwahl für eine Untergruppe der Parameterraum der Lernverfahren direkt um die Art der Integration der anderen Patienten erweitert. Dies erhöht die Anzahl der zu evaluierenden Verfahren deutlich, so dass die Ressourceneffizienz bei der Modellwahl weiter verbessert werden muss.

Das Zusammenspiel der geplanten Arbeitspakete ist in Abbildung 3.1 schematisch dargestellt. In AP1 werden die Verfahren zur automatischen Modellwahl entwickelt. Der Ablauf der Modellwahl wird hinsichtlich möglicher Flaschenhälse in der parallelen Ausführung durch AP2 genau analysiert. Die hier gewonnenen Erkenntnisse fließen direkt in die Offline-Ablaufplanung in AP3 ein, welches durch die Online-Ablaufplanung in AP4 erweitert wird. Die Modellwahl für Subgruppen in AP5 kann sowohl von der Offline- als auch Online-Ablaufplanung profitieren. In AP6 wird in regelmäßigen Abständen mit Werkzeugen aus der ersten Phase der sequentielle Ablauf der entwickelten R-Programme kontrolliert und gegebenenfalls hinsichtlich des Ressourcenbedarfs optimiert.

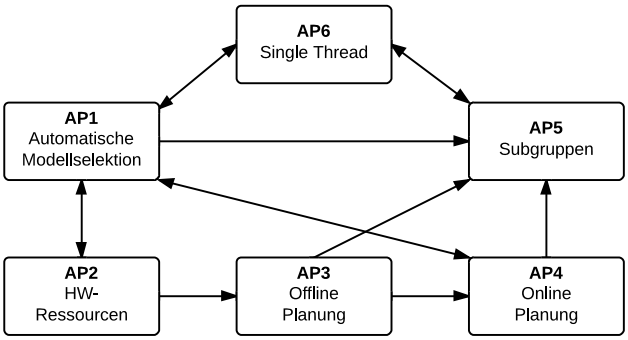


Abbildung 3.1.: Zusammenhang der Arbeitspakete

Arbeitsprogramm

Arbeitspaket 1. Automatische Modellwahl: Ziel dieses Arbeitspakets ist die Entwicklung von Methoden zur automatisierten Modellwahl unter Ressourcenbeschränkungen. Dabei werden Vorhersageprobleme für

sehr große Datensätze betrachtet, für deren Modellbildung viele mögliche Lernverfahren verwendet werden können.

Eine Auswirkung von ausführlichen Methodenvergleichen ist, besonders bei der Modellierung mit hochdimensionalen Daten, das Auftreten großer Laufzeitprobleme. Das eigentliche Problem stellen dabei zumeist nicht die Methoden, sondern deren Parameter dar. Über die Sensitivität der zu Grunde liegenden Verfahren gegenüber einer Parameteränderung ist mehrheitlich nichts bekannt. Verschiedenste Parametereinstellungen müssen somit validiert werden, um einen fairen Vergleich zu ermöglichen. Jedoch ist der aufgespannte Parameterraum in der Regel sowohl teilweise unbeschränkt als auch gemischt-skaliert mit stetigen Komponenten. In diesen Fällen stehen unendlich viele Parametereinstellungen zur Auswahl. Die Abhängigkeit zwischen Parametern, etwa dass gewisse Einstellungen nur kombiniert Sinn machen, erschwert die Modellwahl durch Einführung hierarchischer Strukturen im Parameterraum.

Wird nun jede Verfahrenswahl der Analysekette ebenfalls als Parameter betrachtet, lässt sich die Suche einer kompetitiven Analysestrategie als Minimierung einer einzelnen Zielfunktion formulieren. Dazu wird zunächst ein Algorithmus $\mathcal{A}$ definiert, welcher je nach Parametereinstellungen $\theta \in \Theta$ verschiedene statistische Verfahren anwendet. Eine Auswertung des Algorithmus findet auf zufälligen Probleminstanzen $\omega \in \Omega$ statt, typischerweise Teilmengen bzw. Trainingsmengen des Gesamtdatensatzes. Mittels einer Kostenfunktion $c_{\mathcal{A}} : \Omega \times \Theta \to \mathbb{R}$ wird die Güte der Lösung quantifiziert. In der Modellwahl entspricht dies der Validierung der Modellvorhersagen auf einer unabhängigen Testmenge mittels einer geeigneten Gütefunktion. Die Suche nach der besten Analysestrategie lässt sich schließlich als Suche nach dem θ formulieren, welches die erwarteten Kosten über stochastische Probleminstanzen minimiert:

$$\theta^* = \arg\min_{\theta \in \Theta} \; \mathrm{E}\left[c_{\mathcal{A}}(\omega, \theta)\right].$$

Eine geschlossene Form der Zielfunktion oder gar deren mathematische Ableitung sind nicht möglich, daher wird von *Black-Box Optimierung* [120] gesprochen. Da die Auswertungen einzelner Konfigurationen θ in dem Modellwahl-Szenario mitunter mehrere Stunden Laufzeit in Anspruch nehmen kann, werden Verfahren benötigt, welche mit möglichst wenigen Auswertungen kompetitive Konfigurationen bestimmen können.

Für derartige Probleme lassen sich zwei populäre Verfahren der diskreten Optimierung zur Anwendung auf komplex-strukturierte und unbeschränkte Parameterräume adaptieren, das *Iterated F-Racing* und die *Modellbasierte Optimierung (MBO)*. Iterated F-Racing wurde von uns bereits erfolgreich auf vier Lungenkrebsdatensätze angewandt [A3/152] und somit erstmalig in der Überlebenszeitanalyse verwendet. MBO ist ein flexibleres Optimierungsverfahren, bei dem im Parameterraum an die Werte der Zielfunktion ein Regressionsmodell, das Surrogatmodell, angepasst wird. Neue vielversprechende Kandidaten werden iterativ anhand der *erwarteten Verbesserung* vorgeschlagen, um das Surrogatmodell mit den neuen Auswertungen entsprechend zu aktualisieren. Die erwartete Verbesserung sorgt sowohl für die Erforschung interessanter Täler der Zielfunktion als auch für die Erschließung unerforschter Gebiete, indiziert durch eine hohe Unsicherheit bzw. Varianz des Surrogatmodells. In Abgrenzung zu evolutionären Strategien, welche wie in RapidMiner zum Parametertuning oder zur allgemeinen Optimierung in diversen R-Paketen implementiert sind, wird der Suchraum wesentlich weniger zufällig und damit effizienter durchlaufen.

Durchführung: Die Ergebnisse der Vorarbeiten auf den Lungenkrebsdaten motivieren einen umfassenderen Vergleich, sowohl beruhend auf Iterated F-Racing als auch unter Verwendung von MBO. Dafür wurde eine Implementierung von MBO als R-Paket *mlrMBO*[13] bereits begonnen. Dies entspricht einer Art „Auto-R“ als Pendant zu Auto-WEKA [138], jedoch erweitert auf allgemeine Zielfunktionen, um somit auch Regressions- und Überlebenszeitmodelle zu unterstützen. Das MBO-Paket soll an vielen Stellen durch austauschbare Module möglichst generisch realisiert werden, so dass Modifizierungen des Optimierungsablaufs zur Adjustierung an die Anwendung oder zur Berücksichtigung weiterer Nebenbedingungen einfach zu bewerkstelligen sind. So sind etwa die Anforderungen an das Surrogatmodell durch die komplex-strukturierten Parameterräume sehr restriktiv. Da durch die Hierarchie des Parameterraums stets fehlende Werte entstehen, muss zudem zu jeder Wahl eines Surrogatmodells eine passende Imputierungsstrategie gefunden werden, was in diesem AP ausgiebig analysiert werden soll.

Weiterhin werden momentan Optimierungsverfahren anhand ihrer Lösung nach dem Verbrauch eines vorgegeben Budgets, was hier Funktionsauswertungen bzw. Modellanpassungen entspricht, verglichen. Da in der Praxis jedoch die tatsächliche Laufzeit relevant ist, ergibt eine Berücksichtigung der Laufzeiten verschiedener Konfigurationen Sinn. Anstatt also ein sehr teures Modell mit leicht geänderten Parametern erneut zu validieren, könnte die mehrfache Auswertung günstigerer Alternativen, bei gleichen Kosten, bessere Ergebnisse

[13] Quellcode verfügbar unter https://github.com/berndbischl/mlrMBO

erzielen. Dazu sollen die benötigten Ressourcen mit den in AP2 entwickelten Techniken aufgezeichnet werden, um anschließend den Bedarf neuer Konfigurationen zu modellieren und die prognostizierten Ressourcen bei der Kandidatengenerierung zu berücksichtigen. Die modellierten Ressourcenanforderungen dienen ebenso als Eingabe bei der Offline- und Online-Ablaufplanung zur parallelen Ausführung in AP3 und AP4.

Neben der ressourcengewahren Parallelisierung des MBO-Pakets ist die Anwendung und Validierung auf weiteren Datensätzen aus der Überlebenszeitanalyse geplant. Dazu sollen Wettbewerbsdatensätze wie etwa [110] genutzt werden, um die Ergebnisse der automatischen Modellwahl mit Modellen zu vergleichen, welche von Experten manuell und aufwendig eingestellt wurden. Weiterhin läuft derzeit in Kooperation mit J. Vanschoren eine Masterarbeit zur Integration von Überlebenszeitdaten und -modellen in die Experiment-Datenbank OpenML [141, 133], wodurch langfristig umfassende Vergleiche sowohl von Methoden der Überlebenszeitanalyse als auch verschiedener Optimierungsstrategien ermöglicht werden.

Arbeitspaket 2. Anwendungsflaschenhälse & Hardware-Ressourcenanalyse: Ziel dieses Arbeitspaketes ist es, Verfahren zu entwickeln, welche es ermöglichen, Engpässe und den Ressourcenbedarf von parallel ablaufenden Lernverfahren bzgl. Laufzeit und Speicherbedarf zu analysieren, um somit die notwendigen Informationen über die Ausführungsplattform für eine, wie in AP1 beschriebene, ressourcengewahre Parallelisierung zur Verfügung zu stellen. Dazu gehören insbesondere Informationen über den Typ und die Anzahl verfügbarer Prozessoren, deren Speicherhierarchie, deren Nutzungskosten sowie deren Vernetzung. Damit wird die Grundlage für die Ablaufplanungsverfahren geschaffen, welche in AP3 und AP4 entwickelt werden. Darüber hinaus sollen Verfahren erstellt werden, die eine feingranulare Betrachtung des Ressourcenbedarfs von sequentiell ablaufenden Lernverfahren ermöglichen. Diese Daten können dann z.B. zur optimierten Ausnutzung von Speicherhierarchien verwendet werden, um so auch die Entwicklung von Optimierungen für sequentiell ablaufende Lernverfahren in AP6 zu unterstützen. Für die Ausführung der Lernverfahren kommen insbesondere aktuelle Multicore-, Multiprozessor-, und Rechenclustersysteme zum Einsatz, ggf. unter der Verwendung von vorhandener GPGPUs. Dabei kann jede dieser Plattformen eine unterschiedliche Performanz aufweisen oder unterschiedliche Ressourcenbeschränkungen (z.B. für Cache- und Hauptspeicher) und Kosten (monetär oder in Form von Energiebedarf) verursachen. Dabei können wir auf Erfahrungen bzgl. Energiemodellen, wie sie z.B. in der ersten Förderphase im Teilprojekt A4 betrachtet wurden, aufbauen. Insbesondere bei Hochleistungsrechnersystemen spielt die Leistungsaufnahme eine wachsende Rolle. Insgesamt sind die Ausführungsplattformen aufgrund des oben Gesagten heterogen.

Durchführung: Für die Analyse des Ressourcenbedarfs von parallel und sequentiell ablaufenden Lernverfahren soll das in Phase 1 verwendete TraceR-Werkzeug als Basis dienen. Hierbei ist nicht nur die Konfigurierbarkeit der bestehenden Analysetechniken auf parallel ablaufende Anwendungen zu erweitern, sondern es sollen auch die Voraussetzungen für feingranulare Analysen des Ressourcenbedarfs geschaffen werden.

Dabei ist es wichtig, die Analyse für die vorhandenen Systemparameter, welche den Ressourcenbedarf beeinflussen, zu ermöglichen. So kann z.B. die Überschreitung der Hauptspeichergröße der Ausführungsplattform zu einer extremen Verlangsamung des Programms bis hin zum Programmabbruch führen. Ein Abbruch würde dann eine erneute Berechnung auf einer anderen Ausführungsplattform erzwingen und damit eine Verschwendung von Ressourcen darstellen. Ein solcher Abbruch kann vermieden werden, wenn der Speicherverbrauch der Anwendung sowie die Systemparameter vor dem Start des Programms bekannt sind. Neben der Analyse der Systemparameter ist ebenfalls der Ressourcenverbrauch einer Anwendung in Abhängigkeit von ihrer Ausführungszeit zu analysieren. Dies ermöglicht die Erstellung von sogenannten Arbeitsmengen, welche dann z.B. die Anzahl der Zugriffe auf den Speicher in Abhängigkeit der Programmlaufzeit enthalten oder die Frage beantworten, welche Datenstrukturen zu welcher Zeit im Speicher angefordert wurden. So können später in AP6 auf Basis dieser Daten neue Datenstrukturen, mit unterschiedlichen Elementgrößen, in die R-Ausführungsumgebung integriert werden, um eine bessere Ausnutzung der Speicherhierarchie zu unterstützen.

Zur Realisierung sollen verfügbare Laufzeitumgebungen eingesetzt werden, um die Entwicklung von umfangreicher Basissoftware zu vermeiden. Eine geeignete Laufzeitumgebung soll daher in einer frühen Projektphase ausgewählt werden. Hierbei ist die Verwendung von virtuellen Maschinen geplant. Bei einer feingranularen Analyse des Ressourcenbedarfs sind ggf. Hardwarecounter zu nutzen, welche z.B. die Anzahl von Cache-Misses im Speicher zurückgeben. Hardwarecounter sind jedoch evtl. in virtualisierter Form nicht verfügbar. Daher soll eine Analyse durchgeführt werden, die abschätzt, für welche Art von Optimierung bzw. Ablaufplanung welches Detaillevel von Informationen über den Hardwareressourcenbedarf benötigt wird.

Arbeitspaket 3. Offline-Ablaufplanung für Lernverfahren: Für die in AP1 und AP5 beschriebenen Lernverfahren liegen als Ausgangsbasis Daten vor. Der Erfolg der Modellierung hängt sowohl von der Wahl als

auch von der Parametrisierung der Lernverfahren ab. Der Umfang des Parameterraums in Kombination mit teuren Modellauswertungen verlangt, wie in AP1 beschrieben, nach neuen methodischen Ansätzen.

Auch wenn in der Regel für diese Art der Modellierung keine harte Echtzeitschranke vorliegt, so ist für die Wartezeit bis zum Abschluss der Optimierung oft eine weiche Schranke gegeben. Als Beispiel sei die Modellwahl auf mehreren Lungenkrebsdatensätzen, durchgeführt in [A3/152], angeführt. Die dort aufgeführten Ergebnisse der Modellwahl weisen auf eine deutliche Unterlegenheit einer domänenweiten Parametrisierung gegenüber einer Datensatz-spezifischen Parametrisierung hinsichtlich der prognostischen Güte hin. Entsprechend muss für jede Kohorte eine eigene Optimierung durchgeführt und, unter Berücksichtigung neuer Patienten, regelmäßig aktualisiert werden. Dabei sollte die benötigte Rechenzeit eine anstehende Therapieentscheidung natürlich nicht merklich verzögern. Damit ergibt sich ein Kompromiss zwischen dem Einsatz an Hardwareressourcen und der Wartezeit bis zur Verfügbarkeit des Ergebnisses, eine aus eingebetteten Realzeitsystemen durchaus bekannte Situation. Jedoch dürften in diesem Bereich zwar üblicherweise mehr Ressourcen bereitstehen, als dies bei eingebetteten Systemen der Fall ist. Dennoch ergibt sich dieselbe Struktur der zu lösenden Optimierungsaufgaben, nur bei anderen Größen der jeweiligen Ressourcen.

Durch eine spezialisierte Ausführungsplanung kann bis zum Erreichen einer Echtzeitschranke der Parameterraum besser exploriert werden und die erzielten Modelle können dadurch verbessert werden. Für dieses Arbeitspaket gehen wir davon aus, dass mögliche Rechenressourcen aus einem Pool (z.B. dem lokalen Rechencluster) weitgehend bekannt sind und deren Verwendung *offline* (d.h. vor dem Start der nächste Runde von Berechnungen) geplant werden kann. Aufgrund teilweise großer Rechenzeitanforderungen erwarten wir Vorteile davon, zunächst die Berechnungen zu planen, statt sie einfach zu starten.

Durchführung: Für eine parallele Fassung des in AP1 beschriebenen MBO-Algorithmus planen wir einen Ablauf gemäß der nachfolgend beschriebenen Iterationen, wobei jede Iteration i die Aktionen 1–4 enthält:

1: Surrogat-Regressionsmodell: In der Iteration i wird ein neues Surrogatmodell R^i auf der Basis eines bereits vorhandenen Modells R^{i-1} und der bereits ausgewerteten Punkte des Parameterraums angepasst. Für $i - 1 = 0$ wird eine geeignete Initialisierung benötigt, etwa durch Evaluation von Punkten eines grobmaschigen Gitters im Parameterraum.

2: Punktemenge: Aus dem Modell werden Vorhersagen über noch nicht besuchte Bereiche generiert und anhand der „erwarteten Verbesserung“ wird ein neuer Punkt vorgeschlagen und evaluiert. Erweiterte Ansätze erlauben das gleichzeitige Vorschlagen einer Vielzahl von Punkten P^i während jedes Iterationsschrittes, welche parallel evaluiert werden können.

3: Ressourcenschätzung: Analog zu dem Surrogatmodell können im selben Parameterraum in jedem Schritt weitere Regressionsmodelle angepasst werden. Wird der Regressand der Zielfunktionsauswertungen durch aufgezeichnete Ressourcen ersetzt, können an Punkten P^i benötigte Ressourcen zur effizienten parallelen Ausführung prognostiziert werden. Anders als etwa in [132] wird hier direkt im Parameterraum modelliert, nicht in einem neuen Raum konstruierter Parameter. Dabei erlaubt die Modellierung des Speicherbedarfs eine Ablösung der sonst üblichen konservativen Speicherreservierung, wodurch eine höhere Parallelität erreicht werden kann.

4: Zuteilung von Ressourcen: Nach der Auswahl der Punkte P^i und der Schätzung der zugehörigen Anforderungen müssen die Berechnungen den vorhandenen Ressourcen zugeteilt werden. Dafür sollte die Ausführungsumgebung eine gewisse Kontrolle der Zuteilung erlauben.

Integration von 2–4: Bei der Wahl von P^i können Ressourcenanforderungen explizit berücksichtigt werden. Dies ermöglicht eine dynamische Adaption an Struktur und Auslastung der Ausführungsumgebung. Beispielsweise wird man bei hohen Anforderungen an einen gemeinsamen Speicher weniger Punkte parallel evaluieren als bei geringen Anforderungen, um so ein Speicher-*Thrashing* zu vermeiden. Durch einen Kompromiss aus Modellgüte und Laufzeit können günstige Unterräume des Parameterraums bevorzugt exploriert werden, was zu besseren Ergebnissen bei gleichbleibender Echtzeitschranke führen kann. Zudem kann das Modell nach interessanten P^i unter Nebenbedingungen, etwa geringem Speicherverbrauch, gefragt werden. Dies ist insbesondere zur Auslastung heterogener Systeme wichtig.

Übergang $i \to i + 1$: Dieses Vorgehen wird iterativ wiederholt, bis ein vorgegebenes Budget erschöpft ist. Ein Ziel ist hier, eine (weiche) Zeitschranke Δ vorgeben zu können, innerhalb derer die Rechnungen abgeschlossen werden sollen. Auf diese Weise kann erreicht werden, dass innerhalb der verfügbaren Zeit Ergebnisse zur Verfügung stehen. Wir grenzen uns von *best-effort*-Ansätzen ab, die evtl. auch nach längerer Zeit keine Ergebnisse liefern. Die Offline-Ablaufplanung erlaubt uns hier, ggf. rechtzeitig erkennen zu können, dass zusätzliche Rechenkapazität extern (beispielsweise durch Anmietung) bereitgestellt werden muss, um innerhalb der Zeitschranke die gewünschten Ergebnisse zu erreichen. Hier sollen Techniken aus

eingebetteten Realzeitsystemen zum Einsatz kommen, auch wenn die üblichen Größenordnungen der konkreten Zeitschranken Δ unterschiedlich sind. Auch können Wechselwirkungen zwischen Ressourceneinsatz, Zeitbedarf und Ergebnisqualität untersucht werden.

Die Ressourcenzuordnung stellt eine übliche Aufgabe der Entwicklung von eingebetteten Systemen dar. Die in diesem Arbeitspaket zu entwickelnden Techniken bilden eine Verallgemeinerung von Schedulingverfahren. Während die in jüngster Zeit zunehmend betrachteten Multiprozessor-Schedulingverfahren v.a. homogene Multiprozessorsysteme betrachten (siehe [116] für eine Übersicht), ist hier eine Unterstützung von heterogenen Multiprozessor- und Rechenclustersystemen notwendig, die zunehmend weitere Verbreitung finden. Das Scheduling für solche Systeme ist derzeit noch nicht weit entwickelt. Generell gibt es derzeit noch eine geringe Anzahl von Schedulingverfahren für derartige Systeme, auch wenn sie zunehmend interessant werden [123]. [115] enthält einen Vergleich von bekannten, heuristischen Verfahren. In unserem Fall gibt es keine einfache skalare Zielfunktion, die im Rahmen des Schedulings zu minimieren wäre, sondern eine multikriterielle Zielfunktion. Schließlich ist sogar die Menge der Tasks nicht fest, sondern kann aufgrund der Struktur des Lösungsraums flexibel gestaltet werden. Aus diesem Kontext heraus ist klar, dass neue Ablaufplanungsmethoden entwickelt werden müssen.

Arbeitspaket 4. Online-Ablaufplanung: Die in AP3 entworfene Offline-Ablaufplanung eignet sich besonders für Umgebungen, in denen die Verfügbarkeit von Rechenressourcen weitgehend statisch ist und sich diese potenziell nur in Hinblick auf die von dem jeweiligen Knoten zur Verfügung gestellten Ressourcen unterscheiden. Die im Rahmen dieses Teilprojekts entworfenen Methoden und Algorithmen sollen in Zukunft eine weitreichende Verbreitung in medizinischen Diagnose- und Analyseprozessen erreichen. Die dazu erforderliche Rechenleistung in Form von dedizierten Rechenclustern, deren Nutzung durch die Ergebnisse aus AP3 und AP6 optimiert wird, ist jedoch in vielen Fällen nicht uneingeschränkt exklusiv verfügbar.

In diesem Arbeitspaket soll nun zwei Entwicklungen Rechnung getragen werden. Zum einen ist in vielen Umgebungen nicht davon auszugehen, dass ein Rechencluster exklusiv zur Verfügung steht. Mit der stetig steigenden Verbreitung von Methoden der personalisierten Medizin wird die Anzahl der Nutzer eines Rechenclusters steigen. Die gleichzeitige Nutzung von Rechenclustern, die durch Techniken wie Virtualisierung realisiert wird, führt dazu, dass die Verfügbarkeit von Ressourcen (Prozessoranzahl, Hauptspeicher, Anzahl nutzbarer Rechenknoten) während der Ausführung lang laufender Berechnungsvorgänge starken Schwankungen unterworfen sein kann. Eine rein statische Ablaufplanung, wie in AP3 beschrieben, ist nicht dazu in der Lage, diese Änderungen zu berücksichtigen. Erstes Ziel dieses Arbeitspakets ist es damit, verteilte Anwendungen auf solchen Systemumgebungen effizient ausführbar zu machen. Zum anderen ist es mittlerweile möglich, Rechenzeit und zugehörige Speicher- und Kommunikationskontingente auf Abruf, z.B. über Angebote wie Amazon EC2 [139], anzumieten. Dies erlaubt es Nutzern auch in Überlastsituationen und in gerätetechnisch schlechter ausgestatteten Umgebungen, komplexe Berechnungen mit minimalen Investitionen an Fixkosten durchzuführen. Dabei erfolgt die Abrechnung nach gebuchter Systemkategorie sowie nach Nutzungsdauer. Die einzelnen Kategorien unterscheiden sich in Ressourcenverfügbarkeit, z.B. Prozessoranzahl und -geschwindigkeit, Hauptspeichergröße und Kommunikationsanbindung. Ein weiteres Ziel ist damit die Reduzierung dieser variablen Kosten durch eine geschickte adaptive Ablaufplanung, die auf den Analysen von AP3 aufbaut.

Ziel dieses Arbeitspakets ist es also, eine effiziente und adaptive Online-Ablaufplanung, insbesondere für die in R implementierten Verfahren aus AP1 und AP5 zu entwickeln, zu entwerfen und zu realisieren, die diesen variablen und heterogenen Anforderungen an die Ausführungsumgebung Rechnung trägt. Hierbei soll u.a. untersucht werden, wie sich zentrale wie dezentrale Ansätze zur Verteilung von Rechenlast, wie z.B. *work stealing*, mit den von AP2 zur Verfügung gestellten Informationen über Ressourcen sowie den in AP1 durchgeführten Modellselektionen kombinieren lassen.

Durchführung: Als mögliche Ansätze zur Ablaufplanung sollen in diesem AP zentrale und dezentrale Methoden auf ihre Eignung zur Integration mit der Modellwahl aus AP1 sowie den Ressourcenanalysen aus AP2 untersucht werden. Hierzu erfolgt zunächst eine Selektion vielversprechender Ansätze anhand von Kriterien wie Adaptivität auf heterogene und wechselnde Umgebungsbedingungen und Ressourcengewahrheit. Die Untersuchung der selektierten Ablaufplanungsmechanismen auf ihre Eignung für die spezifische Umgebung dieses Teilprojekts wird im Folgenden exemplarisch anhand von work stealing [112] beschrieben.

Beim work stealing-Ansatz bemüht sich jeder Prozessor in einem verteilten System darum, bei einer anstehenden *Idle*-Phase nach Ende einer Berechnung eine neue Aufgabe zu erhalten. Dies erfolgt unter Berücksichtigung der lokal verfügbaren Rechenleistung und Speichergröße, die durch die Methoden zur Analyse von Hardwareressourcen aus AP2 ermittelt werden können, sowie der Ressourcenanforderungen der selektierten Modelle des MBO-Algorithmus aus AP1. Dabei erfolgt die zu Grunde liegende Ausgangsverteilung

der Rechenaufträge auf die einzelnen Knoten des verteilten Systems anhand des Algorithmus aus AP3. In der skizzierten Umgebung, die in diesem AP zu Grunde gelegt wird, ist jedoch damit zu rechnen, dass die Vorhersage von Ausführungszeiten der einzelnen Tasks nur unpräzise erfolgen kann. Im Vergleich zum Algorithmus aus AP3 ist bei work stealing der zur Laufzeit zu betreibende Aufwand höher. Für die zu untersuchenden Ansätze muss daher sichergestellt werden, dass der zusätzliche Aufwand durch die Online-Ablaufplanung nicht höher ist als die Einsparungen durch Nutzung von Idle-Zeiten und -Ressourcen. Ein aktueller Vergleich von work stealing mit alternativen Ablaufplanungsverfahren, wie z.B. *Fork-Join*, zeigt [124], dass eine effiziente Realisierung des work stealing für eine Reihe untersuchter Benchmarks möglich ist. Allerdings weisen die untersuchten Benchmarks vergleichsweise geringe Ressourcenanforderungen auf. Hier ist zu untersuchen, inwiefern sich der gewählte Ansatz – hier also work stealing – für Aufgaben mit langen Laufzeiten und hohen Anforderungen an Prozessorzeit und Hauptspeicher, wie sie im Falle der zu untersuchenden R-Anwendungen vorliegen, eignet bzw. adaptierbar ist.

Ein weiterer Vorteil der Online-Ablaufplanung ist, dass sie Adaptivität durch die Ausführung spekulativer Berechnungen ermöglicht. Die jeweils in AP1 selektierten Modelle sind durch das verwendete Surrogatmodell mit einer Bewertung ihrer Relevanz für die Untersuchung des Problemraums versehen. Hierbei existieren insbesondere auch Punkte im Problemraum, über die bisher keinerlei Aussagen getroffen wurden. Für eine Reihe dieser Punkte ist es daher von Interesse, diese auszuwerten. Hier kann nun bei Verfügbarkeit von nicht belegten Ressourcen eine größere Menge dieser Punkte analysiert werden. Die Ergebnisse dieser Berechnungen sind nicht unbedingt für die Ermittlung des Ergebnisses der Gesamtberechnung erforderlich, haben aber das Potenzial, die Genauigkeit des verwendeten Surrogatmodells zu erhöhen.

Durch die Ausnutzung von verfügbaren Systemparametern, Zeitschranken sowie den Kenntnissen aus AP3 wollen wir uns von den üblichen Best-Effort-Projekten zur verteilten Ausführung von Berechnungen auf großen Datenmengen, wie z.B. BOINC [106], abgrenzen. Die Berücksichtigung unzuverlässiger Hardware stellt eine mögliche Erweiterung dar.

Arbeitspaket 5. Modellwahl für Subgruppen: In diesem Arbeitspaket sollen die in den Arbeitspaketen AP1-AP4 entwickelten Methoden auf eine noch schwierigere Problemstellung erweitert werden. Es sollen Methoden der Modellwahl entwickelt werden, die eine Untergruppenstruktur von Patientenkollektiven berücksichtigen.

Die häufige Heterogenität von Patientenkollektiven, besonders in der Krebsforschung, kann zur Folge haben, dass ein statistisches Modell, das für eine bestimmte Untergruppe von Patienten gebildet wurde, für andere Patientenuntergruppen keine guten Vorhersagen liefert. In diesem Fall ist es zunächst plausibel, für jede Untergruppe unabhängig voneinander die Modellwahl durchzuführen. Falls die Anzahl der Patienten in einer Untergruppe aber vergleichsweise klein ist und in anderen Untergruppen Patienten mit ähnlichen klinischen oder genetischen Eigenschaften existieren, liegt es nahe, diese bei der Konstruktion von geeigneten Vorhersagemodellen zu berücksichtigen.

Unser spezielles Interesse liegt auf der Vorhersage von Überlebenszeiten für Patientengruppen, die entweder bezüglich klinischer und demographischer Variablen wie Geschlecht, Alter, Tumorhistologie oder Tumorstadium oder beruhend auf genetischen Variablen gebildet werden. Im Kontext der Überlebenszeitanalyse mit hochdimensionalen genetischen Kovariablen sind die untersuchten Modelle meist Erweiterungen des populären Cox-Modells. Neben penalisierten Likelihoodverfahren wie der Lasso-Cox-Regression finden laufzeitintensive Methoden, z.B. beruhend auf Boosting-Schritten [111], Anwendung.

Wir haben in ersten Untersuchungen spezifische Vorhersagemodelle für klinisch definierte Untergruppen von Lungenkrebspatienten bestimmt. Hierzu wurden zunächst für alle Untergruppen separat Modelle erstellt. Für kleine Untergruppen wurden dann beruhend auf dem Konzept der gewichteten Likelihood aus [109] in einem vorgeschalteten Schritt Gewichte für einzelne Beobachtungen bestimmt, die besagen, wie gut diese zur Untergruppe passen [A3/155]. Die Bestimmung der Gewichte beruht auf einem Klassifikationsproblem mit hochdimensionalen Kovariablen, so dass ein noch größeres Laufzeitproblem bei der Modellwahl entsteht. Mit Hilfe des in der ersten Phase entwickelten Parallelisierungs-Frameworks aus dem R-Paket *BatchExperiments* [A3/153] konnte die Modellwahl prinzipiell durchgeführt werden, allerdings auf Grund der Ressourcenprobleme insbesondere bezüglich der Laufzeiten nur für einen eingeschränkten Raum von Modellen.

Bei unserer Zielstellung stehen unzählige Kombinationen von statistischen Methoden und Lernverfahren zur Verfügung, angefangen von der Art der Vorverarbeitung der Rohdaten über eine eventuell zusätzliche Dimensionsreduktion bis hin zur eigentlichen Modellanpassung. Hinzu kommt nun noch die Bestimmung von Gewichten zur Modellierung für Untergruppen. Eine geeignete Modellwahl sollte alle diese Aspekte beachten.

Durchführung: Die in AP1-AP4 entwickelten Methoden der ressourcenoptimierten automatisierten Modellwahl werden schrittweise bezüglich der zusätzlich zu berücksichtigenden Untergruppenstruktur erweitert. Sei G eine Untergruppe von Patienten eines Gesamtkollektivs K, z.B. Patienten mit einem Adenokarzinom als

Untergruppe von Lungenkrebspatienten. Als Startpunkt wird zunächst das gesamte Patientenkollektiv K zur Modellbildung verwendet und dann nur die Untergruppe G. Beide Vorgehensweisen liefern Modelle, deren Vorhersagegüte für die Untergruppe G geschätzt werden kann. Ziel ist nun die Güte zu verbessern, indem für jede Beobachtung aus K geschätzt wird, wie geeignet sie für die Modellbildung für G ist.

Hierbei werden zwei verschiedene Vorgehensweisen evaluiert. Im ersten Ansatz werden die Lernverfahren, die auf dem Gesamtkollektiv K die beste Güte erreichten, vorausgewählt, und die Parameter der Verfahren werden für die Untergruppe neu angepasst, bevor mit dem Prinzip der gewichteten Likelihood die Gewichte einzelner Beobachtungen für die Modellbildung bestimmt werden. In einem zweiten Ansatz werden diese beiden Schritte zusammengefasst, d.h. die automatisierte Modellwahl wird mit einem einzigen Algorithmus durchgeführt, der das Beste oder eine Menge von kompetitiven Lernverfahren bestimmt. Hierbei wird der Parameterraum für die Modellwahl, der in AP1 (und damit auch in AP3 und AP4) aus den Lernverfahren und deren Parametern bestand, direkt um die Art der Integration der Patienten aus den anderen Untergruppen erweitert.

Ziel ist eine umfassendere Analyse der Qualität von Überlebenszeitvorhersagen für Krebspatienten aufgrund von hochdimensionalen genetischen Messungen. In der personalisierten Medizin ist es üblich und inhaltlich gut begründet, die Einteilung von Patienten in Untergruppen nicht nur nach klinischen, sondern auch nach genetischen Aspekten vorzunehmen. In einem weiteren Schritt werden in AP5 deshalb die Patienten nach genetischen Merkmalen in Gruppen eingeteilt, z.B. basierend auf Mutationen von bestimmten Genen oder extremen Expressionswerten von Genen. Da prinzipiell Zehntausende von Genen für die Einteilung zur Verfügung stehen, wird zunächst eine Vorauswahl anhand von biologisch-medizinischem Wissen erfolgen. Dennoch wird sich die Anzahl der Modellanpassungen durch dieses Vorgehen noch einmal deutlich erhöhen. Zusätzlich soll als alternativer Ansatz zur inhaltlich motivierten Einteilung auch diese Vorauswahl von Genen noch mit in den Prozess der automatisierten Modellwahl eingebaut werden.

Da erste untergruppenbasierende Überlebenszeitvorhersagen von uns gerade erst entwickelt wurden, ist auch bezüglich der Methoden in der Überlebenszeitanalyse noch weitere Arbeit zu leisten. Beispielsweise stehen verschiedene Gütekriterien für die Bewertung von Modellen zur Verfügung (C-Index, Brier-Score) und es müssen vielfach noch geeignete Funktionen in R programmiert werden. Die Güte eines Prognoseverfahrens hängt auch stark von der Einteilung von Patienten in Trainings- und Testmenge ab. Die meisten Analysen müssen daher für verschiedene solcher Aufteilungen wiederholt werden.

Insgesamt wird AP5 auf AP1-AP4 aufbauen und von diesen profitieren.

Arbeitspaket 6. Single Thread-Optimierung: Ziel dieses Arbeitspaketes ist es, den Ressourcenbedarf von auf einem Kern ausgeführten Instanzen von Lernverfahren zu optimieren und so die effiziente parallele Ausführung der Lernverfahren, welche in AP1 und AP5 entwickelt werden, zu unterstützen. Als Grundlage für die Realisierung von Optimierungen für einzelne Tasks dienen die Ergebnisse, welche in der ersten Phase von A3 durch die Analyse des Optimierungspotenzials von R mit Hilfe von TraceR gewonnen wurden, sowie die in AP2 entwickelten neuen feingranularen Analyseverfahren für den Ressourcenbedarf. Die Realisierung der Ansätze soll dabei direkt im originalen Interpretercode stattfinden, um Kompatibilitätsprobleme zu verwendeten R-Paketen zu vermeiden.

Durchführung: Wir fokussieren uns auf Bereiche, in denen die Ergebnisse in der ersten Projektphase Einsparpotenzial ergeben haben. Erstens versprechen wir uns Einsparungen durch die Einführung von skalaren Datentypen. Zweitens bietet die Einführung von Datenstrukturen, die per Referenz übergeben werden können, ebenfalls ein Optimierungspotenzial, das wir nutzen wollen. Drittens sollen (auf Basis der Informationen über Cache-Zugriffe aus AP2) Datenstrukturen entwickelt werden bzw. die Elementgrößen innerhalb von bestehenden Vektordatenstrukturen so an die jeweils vorhandene Speicherhierarchie angepasst werden, dass diese effizient ausgenutzt wird. Viertens planen wir zur Reduktion des Aufwands für die Typüberprüfung bei einem Funktionsaufruf die Nutzung der Funktionsspezialisierung, wie sie bereits in anderen dynamischen Sprachen wie Python, JavaSkript oder Matlab mit Erfolg eingesetzt wurde [113, 105]. Fünftens wollen wir die Erzeugung von sogenannten *Promises* in den Fällen vermeiden, in denen dies möglich ist. Weitere Möglichkeiten können sich im Laufe des Projekts ergeben.

Zeitplan

Arbeitspaket / Quartal	2015				2016				2017				2018			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Automatische Modellwahl:	■	■	■	■	■	■	■	■	■	■						
2. Anwendungsflaschenhälse & Hardware-Ressourcenanalyse:	■	■														
3. Offline-Ablaufplanung für Lernverfahren:			■	■	■	■	■	■								
4. Online-Ablaufplanung:									■	■	■	■	■	■	■	■
5. Modellwahl für Subgruppen:											■	■	■	■	■	■
6. Single Thread-Optimierung:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

3.5 Stellung innerhalb des Sonderforschungsbereichs

Es bestehen enge Zusammenhänge zu anderen Teilprojekten. In den Teilprojekten A1, A2, B3, C1 und C4 werden Algorithmen verändert oder das Lernziel wird approximiert, um Ressourceneffizienz zu erreichen. Von all diesen Ansätzen kann das Teilprojekt A3 profitieren, der Fokus liegt hier jedoch auf einer ressourcengewahren Modellwahl mit effizienten Methoden und bestmöglicher Berücksichtigung vorheriger Analysen.

In Teilprojekt A1 werden Algorithmen der Datenanalyse entwickelt, die den Ressourcenverbrauch eingebetteter Systeme senken und dabei selbst möglichst wenig Ressourcen verbrauchen. Dabei wird erforscht, wie in ubiquitären Systemen erhobene Daten über Nutzungsmuster, Kontext und Systemzustand ressourcenschonend lokal oder zentral analysiert werden können. Das so gewonnene Wissen soll für eine laufende Systemanpassung eingesetzt werden. Das Teilprojekt A3 verfolgt ein ähnliches Ziel, aber mit dem Schwerpunkt auf einer ressourcenschonenden Modellwahl und einer reduzierten Bedeutung der Erfassung von Nutzungsmustern. Naturgemäß liegt ein reger Austausch hinsichtlich der eingesetzten Methoden einschließlich der Ablaufplanungsverfahren nahe.

Im Teilprojekt A2 ist ebenso wie im Teilprojekt A3 eine Planung unter Ressourcenbeschränkung wichtig.

Mit dem Teilprojekt A4 existieren zahlreiche Berührungspunkte. Dazu gehören Modelle von Ausführungsplattformen einschließlich der Modellierung von Ressourcen. Auch gibt es in beiden Projekten den Gedanken der automatischen Anpassung des Betriebszustandes, wenngleich wiederum für unterschiedliche Anwendungsszenarien. Das Teilprojekt A4 wird Hilfestellung bei der Bewertung von Lösungen des Ablaufplanungsproblems und der dynamischen Anpassung bieten.

Das Teilprojekt B1 profitiert ebenfalls von den ressourcenoptimierten Methoden der Modellwahl aus Teilprojekt A3. In Teilprojekt B1 werden viele verschiedene Klassifikationsverfahren für multivariate Proteindaten verglichen. So ist ein Einsatz der entwickelten Methoden zur automatischen Modellwahl vielversprechend.

Eine weitere Kooperation mit Teilprojekt B2, das sich mit Ressourcenoptimierungen für komplexe mobile Sensorsysteme für ambulante medizinische Untersuchungen befasst, liefert Informationen über medizinische Analysen auf großen Datenmengen, die mit anderen Programmiersprachen als R entwickelt wurden. Das Teilprojekt B2 selbst kann von den in diesem Teilprojekt entwickelten Feedback-Methoden und der Integration einer regressionsbasierten Analyse zur Bestimmung des Gesamtberechnungsverlaufs wie auch von den Ablaufplanungsverfahren profitieren.

Es besteht eine enge Zusammenarbeit mit Teilprojekt C1 hinsichtlich der Modellierung von hochdimensionalen genetischen Daten, insbesondere in Bezug auf die Variablenselektion. In der ersten Phase wurden gemeinsam stabile Modelle zur Merkmalsauswahl für Überlebenszeitprognosen anhand von Expressionsdaten entwickelt. Diese Methoden fließen in die Modellvergleiche in A3 ein und profitieren ihrerseits von den in A3 entwickelten Laufzeitverbesserungen. In der zweiten Phase wird diese Kooperation fortgesetzt, und die in A3 zu entwickelnden Methoden werden zur automatisierten Modellwahl für die Analyse der NGS-Daten verwendet.

Teilprojekt C4 beschäftigt sich mit der Entwicklung von hocheffizienten Regressionsverfahren für sehr große Datenmengen. Die dort entwickelten Methoden können die Konstruktion des Surrogat-Regressionsmodells verbessern, welches in A3 den Zusammenhang zwischen Parameterraum und Qualitätsbewertung modelliert.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

Gen-Verteilungen
(Rahnenführer,Förderkennzeichen DFG RA 870/4-1) (Förderzeitraum: 2011 - 2016)
Das Projekt „Verbesserte prognostische Signaturen aus Microarray-Studien durch Auswahl von Genen mit charakteristischen Verteilungen“ erstellt auch Vorhersagen für Überlebenszeiten aus hochdimensionalen genetischen Daten. Der Fokus liegt aber auf der Vorauswahl von Genen anhand ihrer charakteristischen Verteilungen. Das Projekt kann von den in A3 entwickelten Methoden zur Laufzeitoptimierung stark profitieren.

Differentielle Netzwerke
(Rahnenführer,Förderkennzeichen DFG RA 870/5-1) (Förderzeitraum: 2011 - 2015)
Das Projekt „Klinische Prognosen auf Grundlage von genregulatorischen Netzwerken“ beschäftigt sich mit der Schätzung differentieller Netzwerke aus Expressionsdaten von vorausgewählten Genen. Laufzeiten und Modellvergleiche sind nicht Thema des Projekts.

eTop: LivSys
(Rahnenführer,Förderkennzeichen BMBF 031A270B) (Förderzeitraum: 2013 - 2015)
Das Projekt behandelt hochdimensionale Expressionsdaten für die Reaktion von Leberzellen auf toxikologische Stoffe. Die Modellwahl ist in diesem Projekt durch die Einbeziehungen von biologischen Annotationen getrieben.

eTop: SysDT
(Rahnenführer,Förderkennzeichen BMBF 031A272E) (Förderzeitraum: 2014 - 2015)
In dem Projekt werden mit statistischen und systembiologischen Methoden Zeitreihen-Messungen von Stammzellen analysiert.

Modelle der Krankheitsprogression
(Rahnenführer,Förderkennzeichen DFG RA 870/6-1) (Förderzeitraum: 2011 - 2016)
Das Projekt „Vergleich von Modellen der Krankheitsprogression bei Krebs und HIV und Entwicklung von Bewertungsmaßen zur statistischen Modellwahl“ beschäftigt sich mit speziellen Progressionsmodellen, die an Hand von vorgegebenen binären Variablen typische Abläufe beschreiben. Die Anzahl an Variablen ist dabei sehr klein.

SAW
(Rahnenführer,Förderkennzeichen Leibniz, SAW-2014-ISAS-2) (Förderzeitraum: 2014 - 2017)
In dem Verbundprojekt „(Reverse) Proteomics as novel tool for biodiversity research“ (Ausführende Stelle: Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.) werden die Gebiete Biodiversität und Proteomik verknüpft. In dem statistischen Teilprojekt werden Methoden entwickelt zum Vergleich von Spektren und Peptidmerkmalen mit dem Ziel der Klassifikation von Organismen.

FEHLER
(Marwedel,Förderkennzeichen DFG MA 943/10-1/2/3) (Förderzeitraum: 2011 - 2016)
Im Projekt “Flexible Error Handling for Embedded Real-Time Systems” (FEHLER) betrachten wir Wechselwirkungen zwischen dem Versuch, Realzeitschranken einzuhalten und dem Versuch, funktional korrekte Rechenergebnisse abzuliefern. FEHLER und das hier beantragte Projekt betrachten Realzeitschranken und nutzen Compilertechniken. Weitergehende Überlappungen gibt es nicht. Die Beantragung der 3. Förderphase (2015–2016) überlappt sich mit der Beantragung des vorliegenden Antrags.

RTools
(Marwedel,Förderkennzeichen Oracle) (Förderzeitraum: 2013 - 2014)
Im Rahmen dieses entwicklungsorientierten Projekts ergänzen wir die Programmierumgebung der Sprache R. Ein Schwergewicht liegt hier bei spezifischen Profiling-Werkzeugen für R. Die Ergebnisse werden als Open-Source-Software bereitgestellt. Die hier beschriebenen Ablaufplanungstechniken sind nicht Teil des Projekts RTools.

TACLe: Timing Analysis on Code-Level
(Marwedel,Förderkennzeichen EU COST IC1202) (Förderzeitraum: 2012 - 2016)

Im Rahmen des TACLe-Projektes erfolgt eine Abstimmung und Koordination der Forschung zur Analyse der Ausführungszeit von Rechenprogrammen auf der Codeebene in Europa. Die Förderung ist auf Reisekosten zu Projektreffen begrenzt.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	117.600	—	117.600	—	117.600	—	117.600
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	117.600		117.600		117.600		117.600	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (A3)							
Wissen-schaftliches Personal	1	Rahnenführer, Jörg, Prof. Dr. Prof	Maschinelles Lernen	Statistische Methoden in der Genetik und Chemometrie , Fak. Statistik, TU Do	4,0	—	Grund-ausstattung
	2	Marwedel, Peter, Prof. Dr. Prof	Eingebettete Systeme	LS 12, Fak. für Inf, TU Do	2,0	—	Grund-ausstattung
	3	Neugebauer, Olaf, Dipl.-Inf. WiMi	Eingebettete Systeme	LS 12, Fak. für Inf, TU Do	13,93	—	Grund-ausstattung
	4	Rempel, Eugen, Dipl.-Math. WiMi	Maschinelles Lernen	Statistische Methoden in der Genetik und Chemometrie , Fak. Statistik, TU Do	19,9	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	5	Graute, Claudia, Verw. Ang.	—	LS 12, Fak. für Inf, TU Do	3,98	—	Grund-ausstattung
Beantragte Personalmittel (A3)							
Wissen-schaftliches Personal	6	Lang, Michel, Dipl.-Stat. WiMi	Maschinelles Lernen	Statistische Methoden in der Genetik und Chemometrie , Fak. Statistik, TU Do	—	Doktorand/in	—
	7	Kotthaus, Helena, M.Sc. WiMi	Eingebettete Systeme	LS 12, Fak. für Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Rahnenführer, Jörg**
   Projektleitung. Schwerpunkt statistische Modellselektion. Mitarbeit vor allem in den Arbeitspaketen AP1 und AP5.

2. **Marwedel, Peter**
   Projektleitung. Schwerpunkt eingebettete Systeme. Mitarbeit in den Arbeitspaketen AP2-4, AP6.

3. **Neugebauer, Olaf**
   Effizienten Realisierung von Anwendungen auf Multiprozessorsystemen unter Berücksichtigung heterogener Plattformen mit unterschiedlich performanten Prozessoren.

4. **Rempel, Eugen**
   Klassifikation bei hochdimensionalen Daten und Analyseprozesse für toxikologische genomweite Messungen.

5. **Graute, Claudia**
   Frau Graute betreut das Projekt administrativ.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

6. **Lang, Michel**
   Statistische Methoden zur automatisierten Modellwahl (AP1) und zur Modellwahl für Subgruppen (AP5), Methoden der Ablaufplanungsverfahren (AP3, AP4) und Input für AP6 (Single Thread Optimierung).

7. **Kotthaus, Helena**
   Analyse von Ressourcenengpässen und die Erfassung der Ausführungsumgebung (AP2) und Ablaufplanungsverfahren (AP3, AP4) sowie Single Thread Optimierung (AP6).

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	7.000	4.000	4.000	2.000
Summe der vorhandenen Mittel	7.000	4.000	4.000	2.000
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt A4

3.1.1 Titel:

Ressourceneffiziente und verteilte Plattformen zur integrativen Datenanalyse

3.1.2 Fachklassifizierung:

Eingebettete Systeme, Kommunikationsnetze, Logistik

3.1.3 Projektleitung

Prof. Dr. Spinczyk, Olaf, 06.04.1970

Lehrstuhl 12, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 16
44227 Dortmund

Tel.: 0231-755-6322
Fax : 0231-755-6116
E-Mail: `olaf.spinczyk@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. ten Hompel, Michael, 19.01.1958

Lehrstuhl für Förder- und Lagerwesen, Fakultät Maschinenbau, Technische Universität Dortmund
Jospeh-von-Fraunhofer-Straße 2-4
44227 Dortmund

Tel.: 0231-755-2793
Fax : 0231-755-4768
E-Mail: `michael.tenhompel@flw.mb.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Wietfeld, Christian, 21.01.1966

Lehrstuhl für Kommunikationsnetze, Fakultät für Elektrotechnik und Informationstechnik, Technische Universität Dortmund
Otto-Hahn-Straße 6
44227 Dortmund

Tel.: 0231-755-4515
Fax : 0231-755-6136
E-Mail: `christian.wietfeld@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Die quantitative Charakterisierung des Ressourcenbedarfs eingebetteter Systemplattformen ist eine wichtige Voraussetzung für den Entwurf und die effiziente Ausführung von Anwendungen zur *Informationsgewinnung unter Ressourcenbeschränkungen*. Nur durch entsprechende Ressourcenmodelle, welche Einflussmöglichkeiten sowohl zum Entwurfszeitpunkt durch den Entwickler, als auch zur Laufzeit durch die Anwendung unter Beachtung von Kontextabhängigkeiten im laufenden Betrieb berücksichtigen, können verteilte Analyseverfahren ressourcengewahr entworfen werden, beziehungsweise agieren. Das langfristige Ziel des Projekts ist daher die Entwicklung einer Methodik, welche es Entwicklern eingebetteter Systemplattformen unter vertretbarem Aufwand ermöglicht, sowohl den Ressourcenverbrauch einer Systemplattform inklusive aller Hardware- und Systemsoftwarekomponenten zu erfassen, als auch die dazugehörigen Ressourcenmodelle zu entwickeln und diese dem Anwendungsentwickler beziehungsweise der Anwendung in geeigneter Form zur Verfügung zu stellen. Herausforderungen bestehen dabei in der Komplexität und Variantenvielfalt selbst kleiner eingebetteter Systeme, aber auch im mangelnden Verständnis einer adäquaten Modellierung aller Zusammenhänge.

In der ersten Projektphase sind, basierend auf komplexer und teils neuartiger Messtechnik, deterministische und stochastische Ressourcenmodelle für die Hardware von Sensorknoten und Mobilfunkendgeräten mit Long Term Evolution (LTE), sowie erste Modelle zu einem eingebetteten Betriebssystem inkl. TCP/IP-Netzwerkprotokollstapel entstanden. Diese Ressourcenmodelle wurden mit Hilfe eines neu entwickelten Simulatorframeworks validiert und in Fallstudien zur Optimierung von Ressourcenverbräuchen auf Anwendungsebene genutzt.

In der zweiten Phase werden diese Arbeiten konsequent fortgeführt und erweitert. Wichtige neue Forschungsfragen in diesem Kontext sind die *Automatisierung* der Erstellung von Ressourcenmodellen, die Einbeziehung der *Produktion* von Ressourcen, z.B. durch *Energy Harvesting*, sowie die stärkere Berücksichtigung von *Realzeitanforderungen*. Zudem soll die Betrachtungsebene von der Einzelplattform auf komplexere *verteilte, hochgradig vernetzte Gesamtsysteme* angehoben werden. Hierbei werden die Wechselwirkungen zwischen Netz- und Endgeräteressourcen im Indoor- und Outdoorbereich, sowie die aktive *Steuerung des Systemkontexts* intensiv untersucht. Neben der Bereitstellung von Systemplattformen und Methoden für andere SFB-Teilprojekte (z.B. A1, A6, B2 und B4) befasst sich A4 mit den Herausforderungen bei der Modellierung komplexer, verteilter Systeme der *Logistik*. Durch die konsequente Modellierung ressourcenbeschränkter Plattformen, deren Spektrum von intelligenten Briefmarken bis zu großen Transportcontainern reicht, soll eine Verbesserung logistischer Planungsmethoden erreicht werden.

3.3 Bisherige Entwicklung des Teilprojekts

3.3.1 Bericht und Stand der Forschung

Die Arbeiten der ersten Projektphase gliedern sich in die Bereiche *Modellierung, Konzeption und Validierung*. Im Folgenden wird der geleistete Beitrag zum Stand der Forschung in diesen Bereichen beschrieben. Zu beachten ist das neue Themengebiet der Logistik, welches durch den neuen Teilprojektleiter Michael

ten Hompel bearbeitet werden soll. Die Motivation für die Einbeziehung dieses Themengebiets, sowie für eine durch eine Anschubfinanzierung ermöglichte Testumgebung für eingebettete Systeme, werden in zwei separaten Abschnitten vorgestellt.

Modellierung (APs 1-3 der 1. Förderperiode): Der Bereich Modellierung wurde in die Pakete *Messung*, *Hardware-Modelle* und *Software-Modelle* gegliedert. Das Arbeitspaket *Messung* wurde hier eingegliedert, da die Messtechnik eine wichtige Grundlage für das Erstellen von Modellen aufgrund empirischer Daten bildet. Insbesondere die Messung von Energiebedarfen stellte im Bereich von stark ressourcenbeschränkten Systemen, wie z.B. *Cyber-Physical Systems* (CPS), eine große Herausforderung dar. Hier werden häufig Microcontroller und Sensoren verwendet, deren Stromaufnahme im Micro- und Nanoamperebereich liegt. Durch ihre starken Energierestriktionen sind *Wireless Sensor Networks (WSNs)* ein wichtiger Ausgangspunkt für aktuelle Forschungsvorhaben im Bereich der Energiemodellierung. Damit in diesem Themengebiet ein Beitrag geleistet werden konnte, musste Messtechnik untersucht werden, die in dem genannten Messbereich noch verwertbare Ergebnisse liefert.

Zur Energiemessung in eingebetteten Systemen steht eine Vielzahl technischer Lösungen zur Verfügung, die in [165] zusammengefasst vorgestellt werden. Die in diesem Zusammenhang verbreitetsten Ansätze basieren auf der indirekten Strommessung mittels eines *Shunt-Widerstandes*. Aktuelle Arbeiten dazu verwenden eine klassische messtechnische Anordnung aus Shunt-Widerstand und Messverstärker [187], welche in der Praxis jedoch zahlreiche Probleme aufweist. Sowohl die Messgenauigkeit, als auch die zeitliche Auflösung, können hier nicht als zufriedenstellend erachtet werden. Andere Verfahren verwenden eine induktive Ankopplung an das System [160], um anhand des elektrischen Feldes eines stromdurchflossenen Leiters Rückschlüsse auf den fließenden Strom zu erlauben. Alternativ dazu werden auch Verfahren zur Umwandlung von Ladung in abzählbare Signale [167, 168] verwendet. All diese Verfahren zeigen Schwächen, die sie für den Einsatz von zeitlich akkuraten Messungen im WSN-Bereich unzulänglich machen. Aus diesem Grund entwickelten wir im Rahmen von AP 1 die *MIMOSA*-Messtechnik [A4/189], welche die Spannungsversorgung eines Gerätes ersetzt und durch schnelle und präzise Spannungsregler einen deutlich größeren Spannungsabfall am Mess-Shunt ermöglicht. Dadurch werden hochempfindliche Messverstärker unnötig. Um auch kürzeste Spitzen im Stromfluss noch korrekt zu ermitteln, verwendet MIMOSA zudem analoge Integratoren. Diese sind in der Lage, auch zwischen zwei Samplezeitpunkten der *Analog-Digital-Converter* (ADC) noch hochfrequente Spitzen zu erkennen und energetisch korrekt an die ADCs weiterzugeben. Die Forschung zu MIMOSA wurde in der ersten Projektphase abgeschlossen. Die Plattform wird jedoch weiterhin durch studentische Arbeiten unterstützt und weiterentwickelt.

Für die Messung von Energieverbräuchen im Bereich aktueller LTE-Hardware wurden Methoden und Laboraufbauten zur hochauflösenden Erfassung des Ressourcenverbrauchs von LTE-basierten Kommunikationsplattformen entwickelt [A4/B4/192]. Hierzu wurde ein spezialisierter Messaufbau, basierend auf einem LTE-Basisstationsemulator, verwendet. Dieses Gerät ermöglicht eine vollständige Kontrolle aller Systemparameter, die von einer LTE-Basisstation gesteuert werden, insbesondere die Uplink-Sendeleistung.

Basierend auf den so gewonnenen Kennlinien wurden im Rahmen von AP 2 geschlossene analytische Modelle entwickelt, mit deren Hilfe sich die Leistungsaufnahme in Abhängigkeit spezifischer Gerätecharakteristika, der Trägerfrequenz, sowie der benötigten Sendeleistung, bestimmen lässt [A4/B4/192]. Vergleichbare Untersuchungen zur empirischen Modellierung der Leistungsaufnahme von LTE Endgeräten wurden parallel zu unseren eigenen Arbeiten auch an der Aalborg University durchgeführt. Die bei der Modellierung berücksichtigten Parameter unterscheiden sich jedoch signifikant zwischen den beiden Ansätzen. Während Lauridsen et al. eine Vielzahl von Parametern mit eher geringem Einfluss auf die Leistungsaufnahme berücksichtigen, wird dort mit der Trägerfrequenz ein erwiesenermaßen entscheidender Parameter vernachlässigt [171, 173].

Für die Bewertung des Ressourcenverbrauches von Mobilfunksystemen ist, neben der Leistungsaufnahme während des Sendevorganges, auch die Datenrate und die damit einhergehende Übertragungsdauer von entscheidender Bedeutung. Übliche Ansätze zur Bestimmung des Zusammenhanges zwischen Kontextparametern, wie den Funkkanalbedingungen und dem erreichbaren Durchsatz, basieren entweder auf Simulationen [157] oder Feldmessungen [169]. Zur Kombination der spezifischen Vorteile beider Ansätze wurden im Rahmen der ersten Phase umfangreiche Messungen der funkkanalabhängigen erreichbaren Datenrate mit Hilfe eines LTE-Funkkanalemulators im Labor durchgeführt [A4/B4/193]. Während man hier, ähnlich einer Simulation, volle Kontrolle über alle System- sowie Kanalparameter hat, müssen, wie bei einer Feldmessung, keinerlei Annahmen bezüglich Hardwareeigenschaften, wie beispielsweise der Empfängerempfindlichkeit, gemacht werden.

Ein wichtiges Thema im Bereich der Mobilfunktechnik und der WSNs ist die Abschätzung der zu erwartenden Batterielaufzeit. Diese Abschätzungen basieren bei Mobilfunkendgeräten heute noch vielfach auf der Annahme, dass die Leistungsaufnahme im Zustand „Senden“ konstant und unabhängig vom jeweiligen

Gerätetyp oder von Systemparametern, wie zum Beispiel der Sendeleistung, ist [161, 183]. Sofern für eine konkrete Technologie gar keine messtechnisch erfassten Werte vorliegen, wird zudem häufig die Leistungsaufnahme in unterschiedlichen Systemzuständen geschätzt [181] oder mit Werten von Vorgängertechnologien, z.B. Universal Mobile Telecommunication System (UMTS) statt LTE [170], verwendet. Die Tatsache, dass die Sendeleistung nicht nur die Leistungsaufnahme beeinflusst, sondern sich aufgrund einer dynamischen Sendeleistungsregelung auch funkkanalabhängig ändert, wird unserer Kenntnis nach bisher in keiner anderen Arbeit berücksichtigt. Aus diesem Grund wurde im Verlauf der ersten Phase CoPoMo (Context Aware Power Consumption Model) entwickelt. CoPoMo ist das erste umfassende Energiemodell für LTE-basierte Mobilfunkendgeräte, das neben einer Vielzahl an Systemparametern (z.B. Frequenz, Anzahl an Funkressoucen etc.) auch den Kontext eines Nutzers (z.B. Zellumgebung, Teilnehmergeschwindigkeit, Funkkanal etc.) in Form der Zellumgebung und des Datenverkehrsaufkommens berücksichtigt [A4/B4/192]. Es basiert auf gerätespezifischen Leistungskennlinien sowie messtechnisch im Labor erfassten, funkkanalabhängigen Datendurchsatzwerten und ist aufgrund seines modularen Aufbaus einfach erweiterbar und somit auch für zukünftige Mobilfunkgenerationen, wie beispielsweise LTE-Advanced, geeignet [A4/191]. CoPoMo bestimmt die durchschnittliche Leistungsaufnahme P_Σ, indem der Energiebedarf eines Endgerätes durch unterschiedliche Systemzuständen abgebildet wird. Die Leistungsaufnahme in den spezifischen Zuständen wird dabei durch $\bar{P}_i$ beschreiben und mit den dazugehörigen Aufenthaltswahrscheinlichkeiten der einzelnen Zustände p_i gewichtet: $P_\Sigma = \sum_i \bar{P}_i \cdot p_i$. Die Werte p_i werden mit Hilfe eines stochastischen Markov-Modells, welches den kontextabhängigen Aufenthalt des Endgeräts in den unterschiedlichen Zuständen beschreibt, berechnet. Abgesehen von diesen eigenen Arbeiten sind quantitative Untersuchungen des Einflusses zukünftiger Mobilfunktechnologien auf die Leistungsaufnahme nur für den LTE-Downlink verfügbar [172]. Die starke Varianz der Leistungsaufnahme von LTE-A-basierten Endgeräten, insbesondere im Uplink, wird jedoch dabei, trotz ihrer großen Bedeutung für die zu erwartende Batterielaufzeit, nicht betrachtet.

Neben den Modellen für Hardware wurde das Thema Software-Kostenmodelle (AP 3) in der Dissertation „Automatic Parallelization Extraction for Embedded Multi-Core Systems using High-Level Cost Models“ von Daniel Cordes bearbeitet. In der Arbeit werden Optimierungen zur Reduktion der Laufzeit paralleler Programme betrachtet. Die Bestimmung von Laufzeit, Energie und Übertragungsvolumen erfolgt hierbei simulativ. Die Arbeit in diesem Paket erfolgte, wie bereits im Antrag zur ersten Phase beschrieben, in enger Zusammenarbeit mit dem Teilprojekt A3.

Zum Ende der ersten Phase wird ein Konzept für Gerätetreiber entwickelt, welches die Analyse von Kosten, insbesondere Energie, durch streng geregelte Zugriffe auf die Hardware erleichtert. Schubert et al. stellen in [184] ein System vor, welches bereits existierende Energiemodelle für CPU und Peripherie den Betriebssystemfunktionen zuordnet. Hierdurch wird es möglich, ressourcenhungrige Systemaufrufe zu identifizieren und eine Abschätzung des Energiebedarfs von Software durchzuführen. Die Aufrufe von System- und Treiberfunktionen werden durch Profiling bestimmt und auf der Treiberebene bereits existierenden Energiemodellen zugeordnet. Im Gegensatz zu dieser stochastischen Vorgehensweise werden in dem von uns geplanten Ansatz die kostenverursachenden Ereignisse analytisch ermittelt.

Konzeption (APs 4-6 der 1. Förderperiode): Im Rahmen der Konzeption wurde die Verwendung von Ressourcenmodellen für die Konfigurierung von Systemen betrachtet. Der Schwerpunkt lag hier auf der Analyse des Einflusses hochkonfigurierbarer Systemeigenschaften auf den Ressourcenverbauch. Zudem sollte die Verschiebung von Entscheidungen von der Entwurfszeit (statisch, offline) zur Laufzeit (dynamisch, online) betrachtet werden. Zur Untersuchung der Konzepte wurde mit der Entwicklung einer Experimentierumgebung begonnen, welche Konfigurierung unter Einbeziehung von Ressourcenmodellen auf statischer und dynamischer Ebene zulässt.

Zur Betrachtung von statischer Konfigurierbarkeit („Produktlinienmodelle“) wurde exemplarisch der Netzwerkprotokollstack eines Betriebssystems als Ausschnitt einer typischen Systemplattform gewählt, da die Möglichkeit, die hier vorhandenen Trade-Offs durch hohe Konfigurierbarkeit auszuschöpfen, sehr erfolgversprechend war. Im Rahmen der Konzeption wird in [A1/A4/46] die Implementierung eines hochkonfigurierbaren TCP/IP-Stacks beschrieben und bestehenden Implementierungen gegenübergestellt. Insbesondere die hohe Zahl der optionalen Features bei TCP/IP prädestiniert ein solches System für den Einsatz von Konfigurationswerkzeugen. Beispiele für solche Optionen können die variable Paketgröße, die Größe der Paketpuffer, die Unterstützung von sog. *sliding Windows* zur Durchsatzoptimierung, die Adresskonfiguration oder der Umfang der Unterstützung von Hilfsprotokollen zur Flusskontrolle und Adressabbildung sein, um nur wenige zu nennen. Es zeigte sich, dass sowohl für die Codegröße, wie auch für den Durchsatz und den Energiebedarf, bessere Ergebnisse erzielt werden, als bei den betrachteten und stark verbreiteten Implementierungen *Micro-IP* (uIP) und *Lightweight-IP* (lwIP). Bei der Messung des Einflusses der möglichen Features

auf Speicher, Energieverbrauch und Laufzeit wurde hier der hohe Aufwand deutlich, der insbesondere bei der manuellen Energiemessung anfällt. Hieraus leitet sich eines der Ziele der zweiten Phase, nämlich die Automatisierung der Messungen nichtfunktionaler Eigenschaften, ab.

Den Übergang von statischer zu dynamischer Konfigurierung haben wir in [A4/188] untersucht. Die Beschreibung von Ressourcenbedarf geschieht häufig über die Verwendung von *Finite-State-Machine*-Modellen (FSM), da sowohl Hardware wie auch Funkprotokolle zustandsbasiert betrachtet werden können und sich auf diese Weise gut analysieren lassen. Gängige FSM-Modelle, wie die *Priced Timed Automata* (PTA), erlauben zudem, das Zeitverhalten von Hardware zu modellieren. Sie ermöglichen so eine sehr genaue Beschreibung und Simulation von Ressourcenverbräuchen. Dies geschieht üblicherweise *offline*, also zur Entwurfszeit des Systems, da die Simulation mit einem Rechen- und Speicheraufwand einher geht, der von tief eingebetteten Systemen nicht bewältigt werden kann. Wir beschrieben eine Methode, welche die Ergebnisse von Vorabsimulationen nutzt, um auch *online*, also zur Laufzeit des Systems, noch effizient Entscheidungen über die Peripherienutzung treffen zu können. Basis hierfür sind Annahmen über den Kontext, welche durch Szenariosimulationen nachgestellt wurden. Die Ergebnisse der Szenariosimulationen können sehr kompakt dargestellt werden. Die laufende Anwendung kann anschließend anhand dieses vereinfachten Modells den zu erwartenden Energieverbrauch bestimmen. Formale Methoden zur Modellierung von Ressourcen für eingebettete Systeme werden in [186] umfassend betrachtet. Hier liegt der Schwerpunkt jedoch stark auf den Modellierungswerkzeugen und Beschreibungssprachen für Ressourcenmodelle und Systemverhalten. Im Teilprojekt A4 wird dagegen die Generierung, Verwendung und Einbettung von Ressourcenmodellen zur Unterstützung bei Entwurfs- und Laufzeitentscheidungen untersucht.

Bei der Optimierung der Batterielaufzeit von LTE-basierten Endgeräten spielt die Wahl der Systemparameter, wie beispielsweise des Schedulers, eine entscheidende Rolle. Durch Anwendung von CoPoMo konnte gezeigt werden, dass durch die geschickte Umverteilung einer festen Menge an Funkressourcen die Batterielaufzeit um bis zu 350% verbessert werden kann. Ähnliche Untersuchungen wurden auch in [170] durchgeführt, wobei hier lediglich der Einfluss zusätzlicher Ressourcen (d.h. einer Erhöhung der Datenrate) auf die Übertragungsenergie bestimmt wurde. Darüber hinaus konnte von uns in [A4/B4/190] gezeigt werden, wie im Fall von Echtzeitanwendungen die Batterielaufzeit durch Bereitstellung zusätzlicher Funkressourcen deutlich erhöht werden kann. Alle im Rahmen der ersten Phase durchgeführten Arbeiten zur Modellierung des Ressourcenverbrauchs von Kommunikation sowie zur Verbesserung der Batterielaufzeit durch geeignete Systemparameterisierung wurden zudem in der Dissertation von Björn Dusza mit dem Titel „Context-Aware Power Consumption Modeling for Energy Efficient Mobile Communication Services“ veröffentlicht. Im Rahmen dieser Dissertation sind neben den Veröffentlichungen [A4/B4/190, A4/191, A4/B4/192, A4/B4/193] noch vier weitere Publikationen (u.a. IEEE INFOCOM, IEEE VTC) entstanden, die hier aus Platzgründen nicht aufgeführt werden können.

In der zweiten Föderperiode soll das Konzept der Modellierung von Ressourcen, insbesondere Energie, und deren Verwendung erweitert werden. Bisher wurden Methoden betrachtet, welche konkrete Modelle erstellen und verwenden. Zukünftig sollen Konzepte erforscht werden, die die automatische Erstellung von Modellen untersuchen. Aktuelle Ansätze hierzu beziehen sich unseres Wissens nach immer auf die Erarbeitung und Schärfung von Energiemodellen von Endgeräten (z.B. Smartphones) im laufenden Betrieb, wie z.B. bei [180]. Hierbei wird der Batteriestand des Geräts ständig über die bereits vorhandene Hardware ermittelt und Ereignissen im System zugeordnet. Diese Systeme erstellen somit stark auf ein einzelnes Endgerät und Applikationsverhalten ausgeprägte Energiemodelle und können deshalb nicht zu Entwurfsentscheidungen für eine ganze Softwareproduktlinie beitragen.

Neben den in der ersten Phase fokussiert betrachteten zellularen Funknetzen, spielt in der zweiten Phase des SFB auch der Betrieb von hochskalierten, energieautarken drahtlosen Sensornetzen (WSN) eine entscheidende Rolle. Trotz der vielfältigen Anwendungen von WSNs und der damit einhergehenden umfangreichen Forschung in diesem Gebiet, gibt es gerade im von diesem SFB adressierten Anwendungsszenario eine Reihe offener Forschungsfragen.

Dies betrifft insbesondere zwei Punkte: Erstens sind im Umfeld logistischer Anwendungen regelmäßig Netze mit bis zu mehreren tausend Knoten zu betrachten. Dies stellt insbesondere unter der Randbedingung eines dynamischen Systems, in dem sich Kommunikationsrouten kontinuierlich ändern, eine große Herausforderung dar [175]. Darüber hinaus bringt zweitens die extreme Beschränkung der Energieressourcen besondere Herausforderungen mit sich, die durch ein intelligentes und auf den speziellen Anwendungsfall hin optimiertes Routing- und Collection-Protokoll adressiert werden müssen. Einen in dieser Hinsicht viel versprechenden Ansatz verspricht das *Collection Tree Protocol* (CTP) [164], welches sich durch Energieeffizienz ebenso auszeichnet wie durch Robustheit, Zuverlässigkeit und die Unterstützung von hochdynamischen Szenarien. Ein alternativer in diesem Kontext zu betrachtender Ansatz ist das Verfahren *Smart Boundary Yao Gabriel Graph* (SBYaoGG) [159], welches ebenfalls eine energieeffiziente, verteilte Organisation der Topologie von

WSN bereitstellt. Ein auf Energy Harvesting basierender Ansatz zur Energieversorgung drahtloser Sensornetze ist darüber hinaus in [176] vorgestellt.

Validierung (APs 7-9 der 1. Förderperiode): Um die oben beschriebenen Modelle in komplexen Szenarien zu validieren, wurden existierende Energie- und Systemsimulatoren für ressourcenbeschränkte Systeme betrachtet [158]. Da die hier gezeigten Simulatoren entweder sehr speziellen Anforderungen dienen (z.B. der Simulation eines bestimmten Betriebssystems) oder die Integration komplexer Modelle nichtfunktionaler Eigenschaften verhindern, wurde in der ersten Phase *MOBISIM* [A4/A1/47] entwickelt. MOBISIM setzt einen Schwerpunkt auf die Modellierung nichtfunktionaler Eigenschaften, wie Energieverbrauch oder Verwendung von Arbeitsspeicher. MOBISIM verzichtet explizit auf eine hochdetaillierte Modellierung funktionaler Eigenschaften, z.B. der Simulation der Befehlsarchitektur eines spezifischen Prozessors, zugunsten von detaillierten Ressourcenmodellen der verwendeten Hardware. MOBISIM erlaubt dadurch die Beschreibung von Applikationen auf einer abstrakten Ebene, wobei aus verschiedenen Modellen für Plattformhardware und Funktechnik gewählt werden kann. Das Anwendungsverhalten wird in C++ formuliert, wobei ein *Application Programming Interface* (API) bereit steht, welches stark vereinfachte Systemaufrufe ermöglicht. Als Beispiel hierfür kann eine Funktion zum Senden von Daten dienen. Für die Bewertung der nichtfunktionalen Eigenschaften weist der eigentliche Inhalt eines Datenpakets keine Relevanz auf. Stattdessen ist rein die Größe des Datenpakets für die Bestimmung des Energieverbrauchs von Belang. Dies spiegelt sich in einem entsprechend vereinfachten API wieder. Das Energiemodell, das den Sendevorgang abbildet, ist hierbei sehr detailliert: Neben den aus der Energiemessung entstandenen Verbrauchsmodellen fließt auch das Zeitverhalten aus der LTE-Spezifikation in das Modell ein. So können technische Parameter, wie die Dauer von sogenannten *Tail States*, die erforderliche Empfangsfeldstärke an der Basisstation, der erwartete Pfadverlust, etc. festgelegt werden. Die vorhandenen Modelle sind weitestgehend parametrisch gehalten, sodass diese schnell an neue Bedingungen (z.B. andere Endgeräte) angepasst werden können. MOBISIM verfolgt mit dieser Vorgehensweise zwei Kernziele: Zum einen dient es als Plattform zur Validierung neuer Modelle. Diese werden in MOBISIM implementiert und können dann mit neuen Anwendungsszenarien simuliert und der Realität, z.B. in Form von empirischen MIMOSA-Messungen, gegenübergestellt werden. Die Validierung folgt aus der Erkenntnis, dass die Simulation des Gesamtsystems unter Einbeziehung der Einflüsse bereits implementierter Komponentenmodelle dem gemessenen Realverhalten auch in Fällen entspricht, die nicht zur Modellbildung selbst herangezogen wurden. Weiterhin erlaubt MOBISIM das Testen und Evaluieren von Anwendungsstrategien, welche durch das leichte API schnell prototypisch realisiert und betrachtet werden können. So wurden in Kooperation mit dem Teilprojekt A1 in [A4/A1/47] Applikationsszenarien zur verteilten Datenerfassung, sowie diverse Benchmarks, erarbeitet und mit realen Implementierungen verglichen.

Zur zuverlässigen Validierung von CoPoMo wurde parallel zu MOBISIM eine Monte-Carlo-Simulation entwickelt, die neben den relevanten Systemparametern auch den Kontext eines Mobilfunknutzers berücksichtigt. Hierzu wird ein realitätsnahes Mobilitätsverhalten eines Nutzers durch eine mittels Ray-Tracing-Simulationen charakterisierte Funkzelle simuliert. Während der Bewegung durch die Zelle werden einer Zufallsfunktion folgend Daten übertragen, wobei der hierzu benötigte Energieverbrauch von der aktuellen Position in der Zelle abhängt [A4/B4/192]. Darüber hinaus wurde die durch CoPoMo geschätzte Batterielaufzeit durch kontextgewahre Laufzeitmessungen unter Laborbedingungen validiert. Sowohl die simulative als auch die messtechnische Validierung von CoPoMo führen zu dem Ergebnis, dass durch das Modell eine äußerst zuverlässige Vorhersage der zu erwartenden durchschnittlichen Leistungsaufnahme bereitgestellt wird.

Im weiteren Verlauf des Projekts wird die Simulation als Evaluationsmethode eine zentrale Rolle spielen. Um umfassende Aussagen über das Systemverhalten zu treffen, wird nicht nur das System selbst simuliert, sondern auch seine Umgebung. Geplant ist dazu eine Cosimulation mit Simulatoren der Informatik, Kommunikationstechnik und der Logistik. Isasa et al. stellen in [166] Cosimulationen aus kontinuierlichen und diskreten Simulationen vor, um so das Energieverhalten eingebetteter Systeme zu evaluieren.

Ein anderer Ansatz wird in [162] verfolgt. Hier besteht die Cosimulation aus Simulationen der Hardware und der Software. Etwas weiter gehen Molina et al. in [178]. Hier wird ein SystemC-Framework vorgestellt, welches Simulationsmodelle für die elektrotechnische Ebene, das Netzwerk, Sensorsignale und Signalausbreitung zusammenführt. Am Lehrstuhl von Prof. Wietfeld wurde zudem im Rahmen der DFG-Forschergruppe 1511 erfolgreich eine HLA-basierte Simulatorkopplung zwischen Simulatoren der Energietechnik und der Kommunikation umgesetzt. Eine entsprechende Publikation von Hanno Georg et al. wurde auf der IEEE SmartGridCom 2013 mit einem Best Paper Award ausgezeichnet. Den genannten Ansätzen gemein ist, dass sie immer nur das konkrete System betrachten, und so Optimierungspotentiale, die sich aus der Umgebung des Systems ergeben, außer Acht lassen.

Ein Beispiel hierfür liefert der Kontext der Logistik: Bewegungen von eingebetteten Systemen durch ein logistisches Gesamtsystem (z.B. eine Produktionsstätte oder ein Lager) sind durchaus planbar und somit

auch auf mehreren Ebenen optimierbar, da hierbei sowohl die Umgebung, als auch das einzelne System geplant und angepasst werden kann.

Einbindung eines komplexen Anwendungskontextes am Beispiel aktueller Forschung in der Logistik: Bislang wird die Leistungsfähigkeit von Plattformen für eingebettete Systeme üblicherweise durch Benchmarks bewertet. Dies hat einerseits den Vorteil, dass die Vergleichbarkeit zwischen verschiedenen Systemen objektiv bewertet werden kann, andererseits den Nachteil, dass der Anwendungsbezug verloren gehen kann, solange er nicht in den einzelnen Benchmarktests verankert ist. In hinreichend komplexen, realen Anwendungskontexten ist eine Bewertung mit bisherigen Methoden der Informatik nur begrenzt möglich [179]. Mit der zunehmenden Bedeutung von CPS, in denen vernetzte, eingebettete Systeme die Steuerung physischer Entitäten übernehmen, bedarf es neuartiger Bewertungsmethoden, welche die starke Verknüpfung von Rechnerelementen und physischem Kontext berücksichtigen.

Aktuelle Forschungen im Bereich der Logistik untersuchen die massive Verwendung von solchen verteilten, eingebetteten Systemen als neue Grundlage moderner Produktions-, Transport- und Distributionsstrategien [A4/197, 174]. Diese Bemühungen werden häufig unter dem Schlagwort *Industrie 4.0* zusammengefasst [182, 185]. Auf Basis des Konzepts eines CPS übernimmt ein mit dem Transportgut eng verknüpftes, eingebettetes System autark die steuernde Rolle und nutzt Transportdienste sowie weitere Dienste von Förderanlagen entlang der Supply Chain. Mit dem intelligenten Behälter (inBin) wurde ein solches System vom Lehrstuhl für Förder- und Lagerwesen entwickelt (vgl. [A4/194], siehe Abbildung 3.1). Dabei spielen neben dem Umgebungskontext auch nichtfunktionale Anforderungen, wie Energieeffienz, eine wichtige Rolle [A4/196]. In aktuellen Forschungsarbeiten dazu wird der Begriff der Leistungsverfügbarkeit verwendet, welcher, unabhängig von der technischen Verfügbarkeit, die Einhaltung weicher Echtzeitanforderungen bezüglich der geforderten Systemleistung beschreibt [177, A4/195]. Für solche Systeme existiert derzeit keine Embedded-Plattform, die über geeignete Modelle oder Validierungstrategien verfügt. Dies hat zur Folge, dass auch keine geeigneten Planungsmethodiken für derartige komplexe, logistische Systeme existieren. Die gemeinsamen Arbeiten in diesem Teilprojekt leisten, neben der Validierung der Plattform in komplexen Anwendungskontexten, einen Forschungsbeitrag für die Planung von Materialflusssystemen, die sich anwendungsunabhängig auf die effiziente Verwaltung und Bewegung physischer Ressourcen und Objekte beziehen. Dies erlaubt die Übertragbarkeit aus der Logistik heraus auf andere Anwendungsgebiete.

Abbildung 3.1.: Versuchshalle für zellulare Fördertechnik mit intelligentem Behälter (inBin) vorne links

Versuchsfeld für ressourcenbeschränkte und verteilte eingebettete Systeme: Aktuell entsteht ein Aufbau eines Versuchsfelds für ressourcenbeschränkte, intelligente Warenträger. Dieses Versuchsfeld soll in die bestehende 1.020 m^2 große Versuchshalle für zellulare Fördertechnik (ZFT) am Fraunhofer IML integriert

werden (siehe Abbildung 3.1), dessen geschäftsführender Institutsleiter Herr ten Hompel ist. Der Aufbau wird mit Hilfe einer Anschubfinanzierung durch den neu hinzukommenden Lehrstuhl für Förder- und Lagerwesen durchgeführt. Es werden 350 Hardwareplattformen (z.T. mit Display) entwickelt und gefertigt, die als interaktive Versuchsplattformen in einen realistischen Logistikkontext platziert werden und das Vermessen und Kartographieren der Umgebungsbedingungen für Energy Harvesting ermöglichen. Zudem können Versuche zur Skalierbarkeit der Funkkommunikation durchgeführt werden. Alle Hardwareknoten werden hierbei mittels Software gesteuert, die auf den im TP A4 erarbeiteten Konzepten beruht. Eine Besonderheit dieses Testbeds ist die Möglichkeit, Sensorknoten über die vorhandene zellulare Fördertechnik gezielt im Raum zu bewegen. Im Gegensatz zu existierenden Wireless Sensor Network Testbeds, wie sie in [163] beschrieben werden, soll das neue Testbed Energy-Harvesting-Methoden im realitätsnahen, industriellen Innenbereich thematisch untersuchen. Existierende logistische Testbeds wie der METRO real,- Future Store oder das SAP Future Retail Center legen ihren Fokus eher auf den Konsumentenbereich im Handel.

Literatur

[157] A. Ameen, E. Mellios, A. Doufexi, N. Dahnoun und A. Nix. „LTE-advanced downlink throughput evaluation in the 3G and TV white space bands“. In: *Proc. of the 24th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*. 2013, S. 771–775.

[158] R. Cheour, M. Jmal, A. Lay-Ekuakille, F. Derbel, O. Kanoun und M. Abid. „Choice of efficient simulator tool for wireless sensor networks“. In: *Proc. of the IEEE International Workshop on Measurements and Networking Proceedings (M&N)*. 2013, S. 210–213.

[159] T. Chiwewe und G. Hancke. „A Distributed Topology Control Technique for Low Interference and Energy Efficiency in Wireless Sensor Networks“. In: *IEEE Transactions on Industrial Informatics* 8.1 (2012), S. 11–19.

[160] C. Cirstea, M. Cernaianu und A. Gontean. „An inductive system for measuring microampere currents“. In: *Proc. of the 18th IEEE International Symposium for Design and Technology in Electronic Packaging (SIITME)*. 2012, S. 197–200.

[161] K. Dufková, J.-Y. Le Boudec, M. Popović, M. Bjelica, R. Khalili und L. Kencl. „Energy Consumption Comparison Between Macro-Micro and Public Femto Deployment in a Plausible LTE Network“. In: *Proc. of the 2nd International Conference on Energy-Efficieny Computing and Networking (e-Energy)*. ACM, 2011, S. 67–76.

[162] P. Ehrlich und S. Radke. „Energy-aware software development for embedded systems in HW/SW codesign“. In: *Proc. of the 16th IEEE International Symposium on Design and Diagnostics of Electronic Circuits Systems (DDECS)*. 2013, S. 232–235.

[163] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton und T. Razafindralambo. „A survey on facilities for experimental internet of things research“. In: *Communications Magazine, IEEE* 49.11 (2011), S. 58–67.

[164] O. Gnawali, R. Fonseca, K. Jamieson, M. Kazandjieva, D. Moss und P. Levis. „CTP: An Efficient, Robust, and Reliable Collection Tree Protocol for Wireless Sensor Networks“. In: *ACM Transactions on Sensor Networks* 10.1 (2013), S. 1–49.

[165] A. Hergenröder und J. Furthmüller. „On Energy Measurement Methods in Wireless Networks“. In: *Proc. of the IEEE International Conference on Communications (ICC)*. 2012, S. 7836–7840.

[166] J. A. E. Isasa, F. O. Hansen und P. G. Larsen. „Embedded systems energy consumption analysis through co-modelling and simulation“. In: *Proc. of the International Conference on Modeling and Simulation*. 2013.

[167] X. Jiang, P. Dutta, D. Culler und I. Stoica. „Micro power meter for energy monitoring of wireless sensor networks at scale“. In: *Proceedings of the 6th international conference on Information processing in sensor networks*. IPSN '07. Cambridge, Massachusetts, USA: ACM, 2007, S. 186–195.

[168] V. Konstantakos, K. Kosmatopoulos, S. Nikolaidis und T. Laopoulos. „Measurement of Power Consumption in Digital Systems“. In: *IEEE Transactions on Instrumentation and Measurement* 55.5 (2006), S. 1662–1670.

[169] J.-B. Landre, Z. El Rawas, R. Visoz und S. Bouguermouh. „Realistic Performance of LTE: In a Macro-Cell Environment“. In: *Proc. of the 75th IEEE Vehicular Technology Conference (VTC Spring)*. 2012, S. 1–5.

[170] M. Lauridsen, A.-R. Jensen und P. Mogensen. „Reducing lte uplink transmission energy by allocating resources“. In: *Proc. of the IEEE Vehicular Technology Conference (VTC Fall)*. IEEE. 2011, S. 1–5.

[171] M. Lauridsen, P. Mogensen und L. Noel. „Empirical LTE Smartphone Power Model with DRX Operation for System Level Simulations“. In: *Proc. of the 78th IEEE Vehicular Technology Conference (VTC Fall)*. 2013, S. 1–6.

[172] M. Lauridsen, H. Wang und P. Mogensen. „LTE UE Energy Saving by Applying Carrier Aggregation in a HetNet Scenario“. In: *Proc. of the 77th IEEE Vehicular Technology Conference (VTC Spring)*. 2013, S. 1–5.

[173] M. Lauridsen, L. Noël, T. B. Sørensen und P. Mogensen. „An Empirical LTE Smartphone Power Model with a View to Energy Efficiency Evolution“. In: *Intel Technology Journal* 18.1 (2014), S. 172–193.

[174] M. Lewandowski, M. Gath, D. Werthmann und M. Lawo. „Agent-based Control for Material Handling Systems in In-House Logistics-Towards Cyber-Physical Systems in In-House-Logistics Utilizing Real Size“. In: *Proc. of 2013 European Conference on Smart Objects, Systems and Technologies (SmartSysTech),* VDE. 2013, S. 1–5.

[175] Y. Liu, Y. He, M. Li, J. Wang, K. Liu und X. Li. „Does Wireless Sensor Network Scale? A Measurement Study on GreenOrbs“. In: *IEEE Transactions on Parallel Distributed Systems* 24.10 (2013), S. 1983–1993.

[176] M. Magno, N. Jackson, A. Mathewson, L. Benini und E. Popovici. „Combination of hybrid energy harvesters with MEMS piezoelectric and nano-Watt radio wake up to extend lifetime of system for wireless sensor nodes“. In: *Proc. of 26th International Conference on Architecture of Computing Systems (ARCS)*. 2013, S. 1–6.

[177] M. Maier. „Praxisgerechte Abnahmeprozeduren für intralogistische Systeme unter Berücksichtigung der Zuverlässigkeits-und Verfügbarkeitstheorie“. Diss. Technische Universität Ilmenau, 2012.

[178] J. M. Molina, M. Damm, J. Haase, E. Holleis und C. Grimm. „Model Based Design of Distributed Embedded Cyber Physical Systems“. English. In: *Models, Methods, and Tools for Complex Chip Design*. Hrsg. von J. Haase. Bd. 265. Lecture Notes in Electrical Engineering. Springer International Publishing, 2014, S. 127–143.

[179] T. Niemueller, D. Ewert, S. Reuter, U. Karras, A. Ferrein, S. Jeschke und G. Lakemeyer. „Towards Benchmarking Cyber-Physical Systems in Factory Automation Scenarios“. In: *KI 2013: Advances in Artificial Intelligence*. Springer, 2013, S. 296–299.

[180] A. Pathak, Y. C. Hu und M. Zhang. „Where is the Energy Spent Inside My App?: Fine Grained Energy Accounting on Smartphones with Eprof“. In: *Proc. of the 7th ACM European Conference on Computer Systems*. EuroSys ’12. Bern, Switzerland: ACM, 2012, S. 29–42.

[181] J. Puttonen, E. Virtej, I. Keskitalo und E. Malkamaki. „On LTE performance trade-off between connected and idle states with always-on type applications“. In: *Proc. of the 23rd IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*. 2012, S. 981–985.

[182] S. Russwurm, A. Huber, G. Baum, H. Borcherding, M. Broy und M. Eigner. *Industrie 4.0: Beherrschung der industriellen Komplexität mit SysLM*. Springer-Verlag New York Incorporated, 2013.

[183] A. Saleh, O. Bulakci, S. Redana, B. Raaf und J. Hamalainen. „Evaluating the Energy Eficiency of LTE-Advanced Relay and Picocell Deployments“. In: *Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*. 2012, S. 2335 –2340.

[184] S. Schubert, D. Kostic, W. Zwaenepoel und K. Shin. „Profiling Software for Energy Consumption“. In: *Proc. of the IEEE International Conference on Green Computing and Communications (GreenCom)*. 2012, S. 515–522.

[185] B. Vogel-Heuser. *Engineering von der Anforderung bis zum Betrieb*. Bd. 3. Kassel University Press GmbH, 2013.

[186] A. Vulgarakis. „A Resource-Aware Framework for Designing Predictable Component-Based Embedded Systems“. Diss. Mälardalen University, School of Innovation, Design und Engineering, 2012.

[187] N. Zhu und I. O’Connor. „Energy Measurements and Evaluations on High Data Rate and Ultra Low Power WSN Node“. In: *Proc. of the 10th IEEE International Conference on Networking, Sensing and Control (ICNSC)*. 2013, S. 232–236.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[A1/A4/46] C. Borchert, D. Lohmann und O. Spinczyk. „CiAO/IP: A Highly Configurable Aspect-Oriented IP Stack“. In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. New York, NY, USA: ACM, Juni 2012, S. 435–448.

[A4/188] M. Buschhoff, C. Günter und O. Spinczyk. „A unified approach for online and offline estimation of sensor platform energy consumption“. In: *Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International*. IEEE, Aug. 2012, S. 1154 –1158.

[A4/189] M. Buschhoff, C. Günter und O. Spinczyk. „MIMOSA, a Highly Sensitive and Accurate Power Measurement Technique for Low-Power Systems“. In: *Real-World Wireless Sensor Networks*. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013.

[A4/A1/47] M. Buschhoff, J. Streicher, B. Dusza, C. Wietfeld und O. Spinczyk. „MobiSIM: A Simulation Library for Resource Prediction of Smartphones and Wireless Sensor Networks“. In: *Proceedings of the 46th Annual Simulation Symposium*. ANSS '13. Society for Computer Simulation International, 2013.

[A4/B4/190] B. Dusza, C. Ide und C. Wietfeld. „Utilizing Unused Network Capacity for Battery Lifetime Extension of LTE Devices“. In: *Proc. of the IEEE International Conference on Communications (ICC)*. Ottawa, Canada: IEEE, Juni 2012.

[A4/191] B. Dusza, P. Marwedel, O. Spinczyk und C. Wietfeld. „A Context-Aware Battery Lifetime Model for Carrier Aggregation Enabled LTE-A Systems“. In: *Proc. of the IEEE Consumer Communications and Networking Conference*. Las Vegas, USA, Jan. 2014.

[A4/B4/192] B. Dusza, C. Ide, L. Cheng und C. Wietfeld. „CoPoMo: A Context-Aware Power Consumption Model for LTE User Equipment“. In: *Transactions on Emerging Telecommunications Technologies (ETT), Wiley* 24.6 (Okt. 2013), S. 615–632.

[A4/B4/193] B. Dusza, C. Ide, P.-B. Boek und C. Wietfeld. „Optimized Cross-Layer Protocol Choices for LTE in High-Speed Vehicular Environments“. In: *Proc. of the 9th IEEE International Wireless Communications and Mobile Computing Conference (IWCMC 2013)*. Cagliari, Italy, Juli 2013.

[A4/194] J. Emmerich, M. Roidl, T. Bich und M. ten Hompel. „Entwicklung von energieautarken, intelligenten Ladehilfsmitteln am Beispiel des inBin“. In: *Logistics Journal* 2012.9 (2012).

[A4/195] E.-N. Jung und M. ten Hompel. „Analytische Stauprognose in Stetigfördersystemen im Rahmen der Systemplanung“. In: *Logistics Journal* 2013.10 (2013).

[A4/196] C. Prasse, A. Kamagaew, S. Gruber, K. Kalischewski, S. Soter und M. ten Hompel. „Survey on energy efficiency measurements in heterogenous facility logistics systems“. In: *Proc. of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE. 2011, S. 1140–1144.

[A4/197] D. Rüdiger, M. Roidl und M. ten Hompel. „Towards Agile and Flexible Air Cargo Processes with Localization Based on RFID and Complex Event Processing“. In: *Dynamics in Logistics*. Hrsg. von H.-J. Kreowski, B. Scholz-Reiter und K.-D. Thoben. Lecture Notes in Logistics. Springer Berlin Heidelberg, 2013, S. 235–246.

3.4 Planung des Teilprojekts

Ziele

Das Ziel des A4-Projekts bleibt weiterhin das Erarbeiten von Methoden, Ressourcenverbräuche zu modellieren und diese sowohl online, als auch offline, für die Plattform bereit zu stellen. Hierbei steht die Idee im Vordergrund, dass die Ressourcenmodelle auf zwei Wegen Einfluss auf das Plattformverhalten nehmen können: Zur statischen Offlineoptimierung existieren Policymodelle, also Regeln zur Instanziierung bestimmter Hard- und Softwarefeatures, welche sich auf die Anwendungsanforderungen und die erarbeiteten Ressourcenmodelle stützen. Zur dynamischen Verwendung zur Laufzeit der Plattform („online“) sollen die Modelle in die Betriebssystemsoftware einfließen und der Applikation so Entscheidungen ermöglichen, die ohne Kenntnis des Ressourcenverbrauchs nicht optimal getroffen werden könnten.

Die in der ersten Phase entwickelten Methoden wurden exemplarisch durchgeführt und aus den Ergebnissen Beschreibungen generischer Modellen, z.B. für den LTE-Mobilfunk, extrahiert. Diese wurden in einer eigens entwickelten Simulationsplattform (MOBISIM) validiert und dienten als Grundlage für Konzepte zur kompakten Darstellung der Modelle für die Onlinenutzung in tief eingebetteten Systemen.

Hierbei zeigte sich, dass für eine vollständige Beschreibung einer Plattform eine große Anzahl an Modellen erstellt und verdichtet werden muss, da eine Plattformproduktlinie, wie sie im A4-Projekt angestrebt ist, über eine extrem hohe Zahl an Varianten verfügt. Neben Architektur- und Kommunikationsoptionen existiert eine große Bandbreite an Sensoren und Peripherie, welche durch unterschiedliche Implementierungen des Betriebssystems gestützt werden. Als Beispiele hierfür können verschiedene Schedulingverfahren für Tasks genannt werden, unterschiedliche Kommunikationsprotokolle oder verschiedene Optionen zur Ansteuerung der Peripherie.

In der zweiten Phase sollen nun Methoden betrachtet werden, welche Plattformfamilien vollständig bezüglich ihres Ressourcenbedarfs modellieren. Hierbei steht insbesondere die Automatisierung der in der ersten Phase erarbeiteten Verfahren im Vordergrund. Es sollen Methoden entwickelt und evaluiert werden, die das automatisierte messtechnische Erfassen von Ressourcenbedarfen der Plattformperipherie ermöglichen, um daraus Kostenmodelle abzuleiten. Dazu sollen zuerst allgemeine PTA-Modelle (*Priced Timed Automata*) für die Peripherie erstellt werden, wie sie bereits in der ersten Phase betrachtet wurden. Aus diesen Modellen soll automatisiert Software abgeleitet werden, welche den betreffenden Peripheriebaustein und die Messtechnik instrumentiert, um das gegebene Modell um die ermittelten Kosten zu ergänzen. In einem weiteren Schritt sollen Methoden untersucht werden, die die Verdichtung der so entstandenen Modelle automatisieren. Während die Verdichtung in der ersten Phase prototypisch erfolgte, ist für die automatisierte Umsetzung insbesondere die algorithmische Analyse der ebenfalls zu erarbeitenden Kontextmodelle erforderlich, um die zur Laufzeit notwendigen Entscheidungsregeln bereits im Vorfeld einzugrenzen. Aus Ressourcenmodell, Kontextmodell und Anwendungsbeschreibung soll letzlich ein Entscheidungsmodell extrahiert werden, welches die Anwendung zur Laufzeit unterstützt und selbst möglichst gering zum Ressourcenverbrauch beiträgt. Dieses Entscheidungsmodell erlaubt die Untersuchung des Entwurfsraumes, um die Plattform entsprechend ihrer Anforderungen zu optimieren. Die Arbeiten im Bereich der automatischen Modellgenerierung und -verdichtung werden in AP 6 durchgeführt. Das AP 7 beschäftigt sich mit den Methoden zur Optimierung durch Entwurfsraumexploration.

Als weiterer Schritt zur Optimierung der Zielplattform wird die Integration der Plattform in ihre physische Umgebung (Kontext) betrachtet. Am Beispiel aktueller Forschung der Logistik („Industrie 4.0“) soll gezeigt werden, wie CPS so weit in ihren Kontext eingebettet werden können, dass diese sogar energieautark agieren. Dies wird nur möglich sein, wenn die Systeme ihre Umgebung genau kennen und auch gezielt beeinflussen können (AP 2). So sind intelligente Warenträger denkbar, die sich gezielt zu Lichtquellen transportieren lassen, um für zukünftige Aufgaben genug Energie durch Solartechnik zu schöpfen. Berücksichtigt man dies bereits bei der Planung logistischer Systeme, entsteht hier eine „Planungsrückkopplung“, welche große Optimierungspotentiale verspricht (AP 1). Die Zielsetzung ist die Anpassung und Erweiterung bestehender logistischer Planungsmethoden. Hierzu müssen die bisherigen Plattformmodelle um Modelle für das Energy Harvesting erweitert werden (AP 5). Weiterhin müssen Konzepte der Logistik in die Systemplanung einfließen. Hierzu werden neben einer Anforderungsanalyse auch neue Evaluationswerkzeuge nötig sein. Diese Konzepte zu erforschen und Werkzeuge zu erstellen wird die Aufgabe in AP 9 sein.

Da es sich bei den hier betrachteten logistischen Systemen um stark verteilte und tief eingebettete Systeme handelt (Beispiel intelligenter Warenträger „inBin“), müssen diese mit effizienten Systemen zur Kommunikation miteinander und mit ihrer Umwelt ausgestattet sein. Bei kleinen Warenträgern in Form von einfachen Behältern muss die Funktechnik vor allem energieeffizient arbeiten. Andere Behälter, z.B. Flugfrachtcontainer (Projekt „DyCoNet“), stellen hohe Anforderungen an die Erreichbarkeit. Während in der ersten Phase der Schwerpunkt der Untersuchungen auf der LTE-Technologie lag, wird im Rahmen der zweiten Phase das Spektrum an detailliert betrachteten Kommunikationstechnologien deshalb stark erweitert. Dies umfasst sowohl Nachfolgetechnologien von LTE (LTE-Advanced, 5G) als auch alternative Systeme wie Wireless Local Area Networks (WLAN), Bluetooth und Low-Energy Radio (AP 4). Darüber hinaus werden Funktechnologien und Erweiterungen im Detail betrachtet, die speziell für die Kommunikation *Machine to Machine* (M2M) entworfen und optimiert sind (AP 3). Neben einer Charakterisierung und Modellierung der Leistungsaufnahme verschiedener Technologien steht die Entwicklung von Verfahren und Protokollen zur dynamischem Auswahl des am besten geeigneten Funkverfahrens im Vordergrund der Untersuchung. Hierbei ist zu beachten, dass beim Umschalten zwischen verschiedenen Technologien (*Vertical Handover*) keine für die Applikation relevante Unterbrechung des Datenstroms auftritt.

Die Akzeptanz der gewonnenen Ergebnisse bei exzellenten Konferenzen und Journals in den hier tragenden Fachbereichen erfordert neben den soliden theoretischen Konzepten in jedem Fall auch eine experimentelle

Evaluierung. In der ersten Phase wurden hierzu verschiedene Systeme herangezogen. So wurden neben realer Hardware (TI ez430 Chronos, diverse LTE-Hardware) auch Simulatoren entwickelt und benutzt, um die erarbeiteten Methoden zu stützen. Jedoch ist der Erkenntnisgewinn beim alleinigen Einsatz von Simulation begrenzt, da neue Modelle erst gegen die Realität validiert werden müssen, um die Aussagekraft einer Simulation zu stärken.

Daher soll mindestens eine konkrete Systemplattform prototypisch in Hard- und Software realisiert und auch in enger Zusammenarbeit mit anderen Teilprojekten des SFBs (insbesondere A1, B2, B4, aber auch weiteren, interessierten Teilprojekten) untersucht werden. Neben der erhöhten Akzeptanz von Veröffentlichungen erwarten wir hier einen deutlichen Erkenntnisgewinn, da unseres Wissens nach bisher keine Plattform existiert, welche zum einen dem Stand der Technik im Bereich der Ressourceneffizienz entspricht und zum anderen offen genug strukturiert wurde, um Forschungsarbeiten in diesem Kontext zu ermöglichen und aussagekräftige Modelle abzuleiten. Die Systemsoftware der Plattform soll Forschungen im Bereich der Energie- und Kontextmodellierung unterstützen und erarbeitete Konzepte hierzu umsetzen. Die Arbeiten hierzu finden im Rahmen von AP 10 (Hardware) und AP 8 (Software) statt.

Arbeitsprogramm

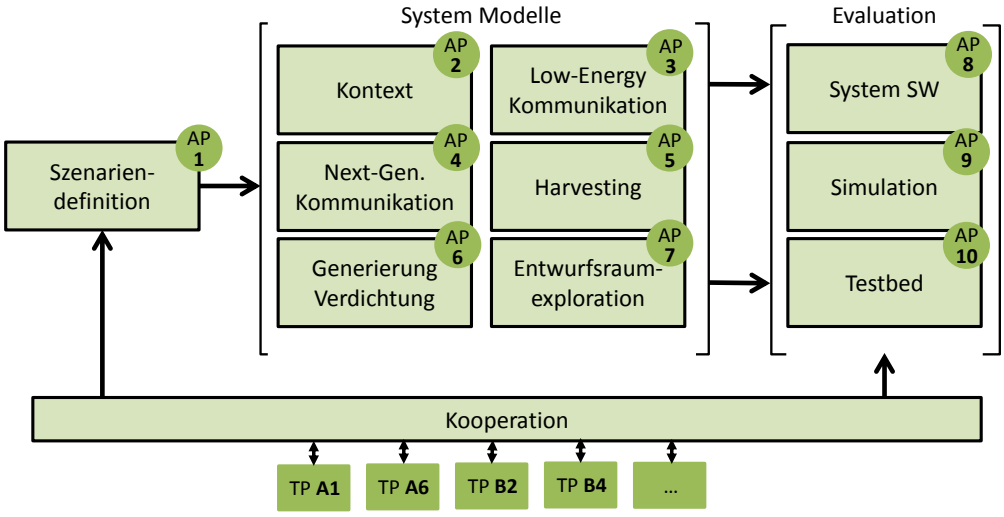


Abbildung 3.2.: Arbeitspaketstruktur des Teilprojektes

Die in der zweiten Förderperiode adressierten Arbeiten werden in insgesamt zehn Arbeitspaketen geleistet. Nach der Definition der Referenzszenarien in AP 1, die sich sowohl aus den Anforderungen der Logistik als auch aus denen anderer, kooperierender Teilprojekte ergeben, werden in insgesamt sechs, teils parallel durchgeführten, Arbeitspaketen (AP 2 bis AP 7) verschiedenen Systemaspekte modelliert mit dem Ziel, den Ressourcenverbrauch vorhersagen und durch geeignete Maßnahmen verringern zu können. Anschließend werden die erarbeiteten Modelle mit Hilfe unterschiedlicher Methoden und unter Zugrundelegung der unterschiedlichen Anwendungsszenarien validiert (AP 8 bis AP 10). Abbildung 3.2 veranschaulicht den Arbeitsplan.

Arbeitspaket 1. Szenariendefinition: Am Beispiel unterschiedlicher, dezentraler Logistikanwendungen werden in diesem Arbeitspaket verschiedene Einsatzszenarien entwickelt, die als Kernbestandteil eine große Zahl von Embedded-System-Plattformen enthalten, welche essentieller Bestandteil der logistischen Leistungserbringung sind. Die Szenarien orientieren sich an realen Logistikproblemstellungen, für die Daten und Kennwerte existieren. Die Beschreibung umfasst den physischen Teil des Systems, bestehend aus Mechanik, physischer Kommunikationsinfrastruktur und Umgebungskontext, sowie die logische Steuerung in Form von anwendungsbezogenen Prozessbeschreibungen. Die zeitlichen Leistungsanforderungen werden in Form von realitätsnahen, generierten Auftragsdaten bereitgestellt. Die wesentlichen zu analysierenden Szenarien sind:

- Szenario aus dem Luftfrachtbereich mit LTE-Anbindung
- Szenario aus dem Intralogistikbereich mit WLAN, Bluetooth und Low-Energy-Funk
- Detailliertes Szenario aus dem Bereich der Testbed-Umgebung

Es erfolgt eine Sammlung und Strukturierung der Anforderungen, sodass hier eine Basis für die spätere Validierung geschaffen wird. Des weiteren werden neue Einflussgrößen identifiziert und Methoden der Evaluation für die Szenarien entwickelt. Das Ergebnis dieses Arbeitspakets sind Szenarienbeschreibungen, die einerseits sehr unterschiedliche Anforderungen und Kontextumgebungen an die Embedded-Plattform bereitstellen und andererseits aus logistischer Sicht ein hohes Einsparungspotential durch den Einsatz von CPS bieten.

Arbeitspaket 2. Kontext: In der ersten Phase des Teilprojektes konnte gezeigt werden, wie sich neben Systemparametern auch der Kontext eines (Kommunikations-) Systems auf dessen Leistungsaufnahme und somit Ressourceneffizienz auswirkt. Die bereits als relevant identifizierten Kontextparameter sind hierbei unter anderem das Nutzerverhalten, sowie die Funkkanalbedingungen.

Ziel dieses Arbeitspaketes ist es, den Kontext eines Systems in die Ressourcenmodellierung einzubeziehen, um hierdurch auf sich dynamisch verändernde Rahmenbedingungen adäquat, durch Anpassung der Policies, reagieren zu können. Die hierfür erforderliche Modellierung lässt sich in zwei Schritte unterteilen: In einem ersten Schritt ist das Ziel, den Kontext selbst zu charakterisieren und zu modellieren. Dies geschieht beispielsweise durch die Identifikation repräsentativer Umgebungen und die sich daran anschließende Modellierung der spezifischen, stochastischen Eigenschaften (Funkkanalbedingungen, Solarpotential etc.). Ebenso ist das applikationsspezifische charakteristische Nutzerverhalten in Form von Verkehrsprofilen zu bestimmen, deren statistische Eigenschaften mathematisch zu beschreiben sind.

Nachdem die Beschreibung der für einen speziellen Kontext relevanten Zufallsvariablen vorliegt, wird in einem zweiten Schritt der Einfluss eines solchen Kontexts auf die Leistungsparameter eines konkreten Systems modelliert. Dies umfasst beispielsweise die positions- und funkkanalabhängig erreichbare Datenrate des verwendeten Kommunikationssystems.

Vom Kontext des Systems stark beeinflusst ist neben den Kommunikationskomponenten auch das Energy Harvesting. Im Rahmen dieses Arbeitspaketes wird daher untersucht, wie sich ein konkreter Kontext zur Dimensionierung und Auslegung des Energy-Harvesting-Moduls modellieren lässt. Ähnlich der Modellierung von positionsabhängigen Funkkanalcharakteristiken werden zu diesem Zweck umfangreiche Messungen und Simulationen des Energiepotentials (z.B. Lichtstärke bei Verwendung von Solarzellen) für charakteristische Umgebungen durchgeführt. Dies hat zum Ziel, hieraus mathematische Modelle der stochastischen Lichtverteilung zu erstellen. Diese Modelle können anschließend beispielsweise zur Dimensionierung des Solarmoduls verwendet werden. Im Rahmen der Systemsoftware können sie zudem genutzt werden, um die Bewegung mobiler Sensorknoten anhand des vorhandenen Energiepotentials zu steuern (siehe AP 8).

Alle in diesem Arbeitspaket erarbeiteten Modelle des Kontextes, sowie dessen Einfluss auf unterschiedliche Systemkomponenten, werden im Rahmen der simulativen Validierung des Gesamtsystems zur Anwendung gebracht. Hierdurch wird sichergestellt, dass dem Validierungsszenario ein konkreter Kontext in Form empirisch parameterisierter, stochastischer Modelle zugrundeliegt.

Arbeitspaket 3. Low-Energy-Kommunikation: Für die im Logistikkontext eingesetzten Kommunikationsknoten existiert häufig eine extreme Beschränkung der Energieressourcen bei hoher Integrationsdichte der Komponenten. Um auch diesen, im Vergleich zu zellularer Mobilfunktechnologie sehr unterschiedlichen, Rahmenbedingungen Rechnung zu tragen, muss eine realitätsgetreue Modellierung durch eine angepasste Messtechnik (MIMOSA) sowie spezialisierte Modellierungsansätze erreicht werden. Hierbei sind die spezifischen Freiheitsgrade der Parametrisierung, wie beispielsweise eine geringere Anzahl von Quantisierungsstufen der Sendeleistung, zu berücksichtigen.

Neben der Modellierung der spezifischen Charakteristika von Low-Energy-Kommunikation ist es Ziel dieses Arbeitspaketes, durch speziell entwickelte, energieoptimierte Kommunikationsprotokolle den Energieverbrauch signifikant zu senken. Hierbei sind insbesondere ein effizienter Kanalzugriff sowie eine effiziente Ausnutzung der zur Verfügung stehenden Funkressourcen von entscheidender Bedeutung, um notwendige Wiederholungen der Datenübertragung auf ein Minimum zu begrenzen. In diesem Kontext sind die besonderen Rahmenbedingungen und Herausforderungen zu beachten, die sich aus der beispielhaften Anwendung eines Kommunikationssystems im Logistikbereich ergeben. Durch die extrem hohe Anzahl gleichzeitig aktiver Funkknoten, sowie die gleichzeitig vorliegende starke Ressourcenbeschränkung, die ein kontinuierliches Abhören des Funkkanals unmöglich macht, sind hochspezialisierte Protokolle und Verfahren zur Sicherstellung einer zuverlässigen Kommunikation notwending. Um diese Herausforderungen zu adressieren, wird im Rahmen des Arbeitspaketes das Collection Tree Protocol untersucht werden. Auf dieser Basis werden dann Methoden zur Anpassung von Protokollen an Applikationen betrachtet. Darüber hinaus werden bereits in der Literatur beschriebene Verfahren zur energieautarken Kommunikation basierend auf Energy Harvesting und Nanowatt Wake-Up Funkempfängern evaluiert und für die betrachteten konkreten Szenarien angepasst.

Um für jeden Knoten des Sensornetzes die individuell optimale Sendeleistung zu bestimmen, werden im Rahmen dieses Arbeitspaketes szenario- bzw. kontextabhängige Abbildungen der Empfangsfeldstärke (*Received Signal Strength Indicator*, RSSI) auf den *Packet Reception Ratio* (PRR) bestimmt. Basierend auf den so hergeleiteten Zusammenhängen lässt sich die notwendige Sendeleistung zur Erreichung einer gewünschten Dienstgüte (QoS) präzise steuern.

Arbeitspaket 4. Next-Generation-Kommunikation: Während in der ersten Phase des Sonderforschungsbereichs die LTE-Technologie als exemplarisches Funksystem ausgewählt wurde, um umfangreiche Untersuchungen der kontext- und systemparameterabhängigen Leistungsaufnahme durchzuführen, liegt der Fokus dieses Arbeitspaketes auf der deutlichen Erweiterung des Spektrums betrachteter Funklösungen. Das primäre Ziel ist es, hierbei zum einen auch zukünftige Entwicklungen geeignet zu berücksichtigen, zum anderen jedoch auch eine Menge von Entscheidungsmöglichkeiten zwischen verschiedenen, von Endgeräten unterstützten Funktechnologien zur Verfügung zu stellen. Nur wenn alle auf einer Plattform bereitstehenden Funkschnittstellen detailliert bezüglich ihrer system- und kontextparameterabhängigen Leistungsaufnahme modelliert sind, können fundierte Entscheidungen über ggf. sinnvolle vertikale Handover, also dem Umschalten zwischen unterschiedlichen Netzen, getroffen werden.

Zur Erreichung dieses Ziels wird die in der ersten Phase entwickelte Messmethodik zur empirischen Modellierung der parameterabhängigen Leistungsaufnahme verschiedener Endgeräte konsequent weiterentwickelt. Sowohl bei der messtechnischen Erfassung, als auch bei der daran anschließenden Modellierung der Leistungsaufnahme müssen hierbei technologiespezifische Aspekte berücksichtigt und das CoPoMo-Modellierungsframework entsprechend erweitert werden. Im Fall der LTE-Advanced-Technologie, die aufgrund der zu erwartenden großen Verbreitung von besonderem Interesse ist, sind insbesondere die spezifischen Einflüsse der heterogenen Netztopologie, der großen, durch Carrier Aggregation erreichten Bandbreiten und die möglichen Einflüsse von miteinander kooperierenden Basisstationen detailliert zu berücksichtigen.

Neben den bis dato fokussierten Zellularfunktechnologien werden in der zweiten Phase auch alternative Datenverbindungen via WLAN und (Low Energy) Bluetooth sowie RFID detailliert untersucht und modelliert. Ebenso wie bei den verschiedenen nutzbaren Zellularfunknetzen steht auch hier die Frage im Vordergrund, unter welchen Rahmenbedingungen (Kontext) ein vertikaler Handover zu einer dieser Technologien aus Sicht der Energieeffizienz sinnvoll ist.

Arbeitspaket 5. Harvesting: Basierend auf den in AP2 entwickelten Modellen sollen kontextabhängige Techniken des Energy Harvestings und des Ressourcenmanagements entwickelt werden. Durch empirische Studien sollen Methoden entwickelt werden, mit denen sich Vorhersagen bei wechselnden Umgebungsbedingung machen lassen und die sich zum Design einer Plattform wie auch als Onlineenergiemodell im Betriebssystem eignen.

Einen Schwerpunkt bildet die Untersuchung von Innenbereichs-Photovoltaik im Industrieumfeld, bei der in logistischen Anwendungsgebieten der Einfluss des Umgebungskontextes ein maßgeblicher Faktor darstellt. Die in Industriehallen verwendeten Kunstlichtspektren moderner, extrem sparsamer Industriebeleuchtung werden hinsichtlich ihrer Auswirkung auf heute gängigen amorphen sowie neuen Solartechnologien bei unterschiedlichen Lichtverhältnissen untersucht. Weiterhin wird der Einfluss von durch die Betriebsstrategie, und somit vom Betriebssystem, veränderlichen Umgebungsparametern, wie dem Einstrahlungswinkel auf den Wirkungsgrad bestimmt. Dabei soll auch der Einfluss von optischen Oberflächenstrukturen sowie der Einfluss der Umgebungstemperatur für diesen Anwendungsfall getestet werden.

Ein weiteres Forschungsfeld bilden Methoden des Energiemanagements basierend auf Modellen empirischer Daten. Dazu werden neue Energiespeichertechnologien für Langzeitspeicher sowie Kurzzeitpuffer evaluiert und auf ihre Kontextabhängigkeit untersucht. Dabei werden die höchst unterschiedlichen Anforderungen aus den Energieprofilen der Verbraucher berücksichtigt. Weiterhin sollen gängige Ladetechnologien beim herkömmlichen Energy Harvesting auf ihre Übertragbarkeit beim Micro Energy Harvesting überprüft werden. Es werden vergleichende Messungen für mögliche Methoden der Spannungsbereitstellung durchgeführt, die bisher nicht beim Design von Ultra-Low-Power-Elektronik berücksichtigt werden, aber zu relevanten Einspareffekten führen können. Ziel ist es, den Wirkungsgrad des Energiemanagements zu steigern, ohne die Leerlaufverluste zu erhöhen. Weiterhin sollen Möglichkeiten untersucht werden, die dem Betriebssystem relevante Leistungsdaten des Energiemanagements zur Verfügung zu stellen, ohne den dadurch geschaffenen Nutzen durch den Energiebedarf der Messungen zu nivellieren. Dazu wird ein sparsames Messverfahren für Sekundärzellen entwickelt, das auch eine Onlinemessung ermöglicht.

Arbeitspaket 6. Generierung / Verdichtung: Die Größe des Entwurfsraumes selbst kleiner eingebetteter Systeme erlaubt es nicht, eine vollständige manuelle Messung und Beschreibung von Ressourcenverbrauch

und -gewinnung jedes Features durchzuführen. Dementsprechend kommen Hardwareplattformen und Softwareproduktlinien nicht ohne eine automatisierte Gewinnung von Ressourcenmodellen aus. In diesem Arbeitspaket sollen Methoden evaluiert werden, die dieses ermöglichen. Maßgeblich sollen Systemplattform- und Treiberarchitekturen untersucht werden, welche in einem selbststeuernden Messaufbau die Hardware gezielt in jeden möglichen Arbeitszustand versetzen, um dabei die verbrauchte oder erzeugte Energie systematisch zu messen. Hierzu wird es notwendig sein, ein Treibermodell zu definieren, welches gezielt und bewusst mit dem Automatenmodell der anzusteuernden Hardware und dessen Zuständen interagiert. Da hierbei eine starke Normierung der Treiberarchitektur angestrebt wird, profitieren auch Verfahren zur Codeanalyse davon.

Da diese Modelle beliebig komplex sein können, ermöglichen sie lediglich die statische Optimierung der Zielplattform. Um auch eine dynamische Optimierung der Plattform, also eine Optimierung zur Laufzeit, durchzuführen, müssen diese Modelle auf der Zielplattform zur Verfügung stehen. Dem im Weg stehen jedoch die Beschränkungen von Speicher und Rechenkapazität von tief eingebetteten Systemen. Von daher müssen Ressourcenmodelle kompakt („verdichtet“) dargestellt werden. Eine Vorberechnung von erwarteten Kontextinstanzen (z.B. Energieverbrauch unter „starker“, „mittlerer“ oder „kleiner“ Last) kann helfen, das Modell auch für Kleinstsysteme handhabbar zu machen, gehen aber mit einem Verlust an Vorhersagegenauigkeit einher. Die Verdichtung von Modellen kann dabei bereits beim Systementwurf erfolgen, ist aber auch bei der Anpassung des Systems an einen individuellen Kontext interessant.

In diesem Arbeitspaket sollen also Methoden entwickelt und evaluiert werden, Ressourcenmodelle möglichst automatisiert zu erzeugen und zu verdichten. Hierbei soll insbesondere der Grad der Vorhersagegenauigkeit und der resultierende Verbrauch an Speicher und Rechenleistung steuerbar werden.

Arbeitspaket 7. Entwurfsraumexploration: In diesem Arbeitspaket sollen Algorithmen vorbereitet und umgesetzt werden, die der automatischen Auswahl von Hard- und Softwarekomponenten dienen. Es deckt damit eines der Kernziele des A4-Projekts ab, da hier erstmalig die erarbeiteten Modelle für eine automatisierte Entscheidungsfindung zur Anwendung kommen.

Dem Bereich der Vorbereitung der Algorithmen fällt hierbei insbesondere die prototypische Spezifikation maschinenlesbarer Formate und Sprachen zur Darstellung von nichtfunktionalen Eigenschaften und Entscheidungsmodellen. Während bisher keine einheitliche Darstellung der Modelle existierte, ist diese zur automatisierten Auswertung unumgänglich. Konkret sollen deshalb also Beschreibungssprachen in das Produktlinienmodell eingebracht werden, welche die Darstellung der erarbeiteten Ressourcen- und Kontextmodelle und die Darstellung von Anwendungsanforderungen ermöglichen. Im Anschluss daran sind Algorithmen zur *Entwurfsraumexploration* zu erforschen und entwickeln, um die automatisierte Auswahl von Features und Hardwarekomponenten gemäß eines vorgegebenen Optimierungsziels zu ermöglichen.

Arbeitspaket 8. Systemsoftware: Dieses Arbeitspaket dient der Erforschung von Systemarchitekturen, welche die Planung von Ressourcenverbräuchen ermöglichen und zudem mit den speziellen Eigenschaften von Energy-Harvesting-Systemen effizient umgehen können. Nach den in der ersten Phase entwickelten Konzepten zur Darstellung von Ressourcen auf der Ebene der Systemsoftware sollen hier nun Konzepte betrachtet werden, welche sich mit der Integration von Planungskomponenten für Ressourcenverbräuche beschäftigen. Hierzu dienen Onlineressourcenmodelle und eine dazugehörige System-API aus der bisherigen Arbeit als Grundlage. Durch die Möglichkeit einer solchen API, nicht nur aktuelle Ressourcenwerte, sondern auch hypothetische Situationen zu betrachten, wird es nun möglich, den Ressourcenbedarf aktiv zu planen. Im Zusammenspiel mit dem Kontextmodell kann eine Planungskomponente also auf die aktuelle Ist-Situation schließen und durch Analyse des Systemzustands auf mögliche zukünftige Szenarien frühzeitig reagieren. Ein Harvestingsystem zum Beispiel, welches so erkennt, dass in naher Zukunft eine große Wahrscheinlichkeit für den Ausfall der Energiequelle (Solarenergie, Radiowellen) bevorsteht, wäre in der Lage zu entscheiden, dass sekundäre Energieverbraucher abzuschalten oder zu reduzieren sind, um dadurch ein Aufladen von Kondensatoren zur Überbrückung der Ausfallzeit zu ermöglichen. In Zusammenarbeit mit Teilprojekt A1 sollen hier zudem Lernverfahren zum Einsatz kommen, um die Prognose der zukünftigen Ereignisse zu schärfen.

Weiterhin werden in Phase 2 erstmals Energy Harvesting-Techniken betrachtet. Diese stellen jedoch besondere Herausforderungen an die Systemsoftware, da der Programmfluss aufgrund mangelnder Energiequellen jederzeit unterbrochen werden kann („brown-out“, „black-out“). Bestehende und neu erarbeitete Konzepte müssen also dahingehend untersucht werden, dass diese in Situationen mit hoher Ausfallwahrscheinlichkeit der Energiequelle noch tragen. Zudem sollen Hardware- und Softwaremethoden erforscht werden, den Systemzustand eines eingebetteten Systems auch durch eine Black-Out-Phase hindurch zu erhalten. In Frage kommen hier Techniken wie *brown-out detection*, also das frühzeitige Erkennen der einbrechenden Versorgungsspannung, gekoppelt mit softwareseitigen Absicherungs- und Fehlertoleranzmechnismen und der Ver-

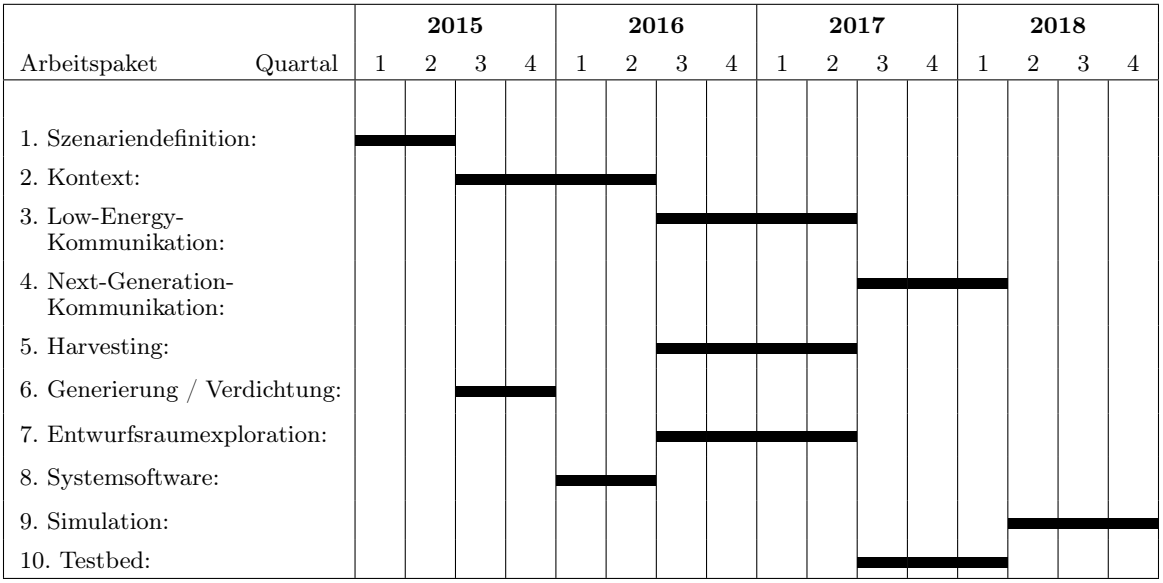
wendung von nichtflüchtigem, energieeffizientem Speicher, wie z.B. *Ferroelectric Random Access Memory* (FRAM).

Arbeitspaket 9. Simulation: Zur integrierten Bewertung der Arbeiten in den Teilbereichen Logistik, Kommunikationstechnik und Systemsoftware ist es notwendig, eine gemeinsame Plattform zur Evaluierung von Experimenten sowie zur Validierung von Modellen zu schaffen. Hierzu ist eine Kopplung der für den jeweiligen Bereich verwendeten Simulatoren mittels *High Level Architecture* (HLA) vorgesehen. Dadurch wird es möglich, Experimentmodelle aus verschiedenen Perspektiven der realen Welt zu betrachten: Während die Simulation im Bereich der Logistik Bewegungen von intelligenten Objekten unter einer gemeinsamen Zielsetzung abbildet, werden für diese Objekte in den entsprechenden Simulatoren der Kommunikationstechnik die Kanaleigenschaften, die Kommunikationsprotokolle und die Funkwellenausbreitung simuliert. Die Informatik steuert Simulationen der Hardware- und Softwareplattformen bei, welche anhand der zuvor bestimmten Ressourcenmodelle valide Aussagen über den Gesamtverbrauch an Ressourcen liefern. Der besondere Vorteil einer derartigen Integration ist der extrem hohe erreichbare Detaillierungsgrad, der sich aus den jeweils ausgereiften und spezialisierten Einzelsimulatoren ergibt. Somit wird es möglich, dedizierte Abhängigkeiten zwischen Materialflussprozessen, Kommunikation und Software auf den Knoten zu modellieren und zu optimieren. So wird es beispielsweise möglich sein, den Einfluss eines einzelnen Parameters des Kommunikationssystems auf den Materialfluss qualitativ sowie quantitativ zu bewerten. Darüber hinaus erlaubt die integrierte Hybridsimulation eine nahezu beliebige Skalierung der simulierten Komponenten und ermöglicht dadurch valide Aussagen über die Grenzen der Skalierbarkeit eines tatsächlichen Systems.

Zusammengefasst wird also ein simulatorübergreifendes Daten- und Ereignismodell erstellt und in eine HLA-Schnittstelle integriert. Anschließend wird die Simulatorumgebung durch gegenseitige Validierung der einzelnen Simulatoren im Versuchsfeld überprüft. Um die Grenzen und Leistungsfähigkeit der Simulationsumgebung zu bestimmen, wird ihre Skalierbarkeit analysiert. Zudem werden typische Referenzszenarien simuliert, um gültige Aussagen über die Genauigkeit der Simulation treffen zu können.

Arbeitspaket 10. Testbed: Zur experimentellen Evaluierung der erarbeiteten Konzepte soll mindestens eine Systemplattform prototypisch in Hard- und Software realisiert werden. Dabei dient die auf Seite 123 beschriebene Testumgebung als Hardwaregrundlage, welche um Komponenten für die Forschung im Bereich Energy-Harvesting und Mobilfunk erweitert wird. Auf dieser Hardware erfolgt dann die Implementierung einer ressourcengewahren Betriebssystemumgebung. Hierbei sollen die in AP 7 erarbeiteten Konzepte zur automatischen Featureauswahl evaluiert werden. Grundlage hierfür soll ein Betriebssystem bilden, dass die in AP 6 beschriebenen Treiberarchitekturen umsetzt. Zudem soll das in AP 8 erarbeitete API zur Prognose von Ressourcenbedarfen genutzt werden, um dann in Kooperation mit anderen Teilprojekten mögliche Prognoseverfahren zu untersuchen und zu bewerten. Um Energy-Harvesting-Ansätze und Kommunikationssysteme in einer logistischen Gesamtumgebung zu evaluieren, werden entsprechende Hard- und Softwareschnittstellen vorgesehen.

Zeitplan



3.5 Stellung innerhalb des Sonderforschungsbereichs

Alle Teilprojekte, bei denen Algorithmen in Hinblick auf nichtfunktionale Systemeigenschaften statisch oder dynamisch optimiert werden sollen, benötigen Informationen über die jeweilige, zu Grunde liegende Ausführungsplattform. Somit ergeben sich Bezüge zu einer großen Zahl von anderen anderen Teilprojekten. Die Kooperationen umfassen hierbei sowohl die Berücksichtigung der individuellen Anforderungen anderer Projekte im SFB, als auch die Bereitstellung von Ressourcenmodellen für andere Teilprojekte. Im Folgenden werden drei bereits in der ersten Phase intensiv verfolgte Kooperationen näher vorgestellt, die auch im Verlauf der zweiten Phase eine große Rolle für die Kooperation im SFB spielen werden.

Teilprojekt A1 entwickelt ressourcengewahre Algorithmen des Data Mining und Lernverfahren und beschäftigt sich dabei unter anderem mit der Generierung von Ressourcenmodellen. Hierbei liegt der Fokus in der Erzeugung von Modellen auf Basis maschinellen Lernens. So kann z.B. durch Lernen des Benutzerverhaltens eine Applikation hinsichtlich ihres Ressourcenbedarfs optimiert werden. Das Benutzerverhalten fließt somit in ein individuelles, dynamisches Ressourcenmodell ein. Im Gegensatz dazu beschäftigt sich das A4-Projekt mit allgemeinen Plattformmodellen, die sowohl bei Entwurfsentscheidungen, wie auch bei Laufzeitentscheidungen zum Tragen kommen. Durch die offensichtlich gegebene Schnittstelle in den Interessen dieser Projekte kam es zu umfangreichen Kooperationen. So wurde z.B. MOBISIM im A4-Projekt entwickelt und gemeinsam mit dem A1-Projekt evaluiert. Das A1-Projekt steuerte wichtige Simulationsdaten bei und nutzte MOBISIM intensiv als Werkzeug für die Auswertung eigener Experimente.

In Teilprojekt A6 werden Lernverfahren zur Bestimmung optimaler Graphen untersucht. Die hieraus resultierenden Ergebnisse werden im Verlauf der zweiten Förderperiode im A4-Projekt dazu verwendet werden, optimale Routen der Datenpakete durch das Wireless Sensor Network zu finden. Hierbei liegt die besondere Herausforderung in der hohen Dynamik des Systems sowie in der großen Anzahl zu betrachtender Knoten.

Teilprojekt B2 konzipiert und realisiert ein eingebettetes Analysesystem mit Echtzeitanforderungen. In diesem Zusammenhang werden im Verlauf der zweiten Phase kontextsensitive Modelle des Energieverbrauchs von A4 genutzt. Ziel ist es hierbei, zur Laufzeit fundierte Entscheidungen darüber zu treffen, inwieweit eine lokale Datenverarbeitung auf einem in den Sensor eingebetteten System einem zentralistischen Ansatz vorzuziehen ist. Dabei ist zu berücksichtigen, dass keine allgemeingültige, optimale Antwort auf diese Fragestellung zu erwarten ist. Vielmehr ist davon auszugehen, dass abhängig von z.B. Funkkanalbedingungen und Zellauslastung verschiedene Stufen der lokalen Vorverarbeitung ausgeführt werden müssen.

Teilprojekt B4 befasst sich in der zweiten Phase des SFB mit Algorithmen zur effizienten Auswertung von Sensordaten mobiler Systeme in urbanen Umgebungen. Dabei liefert B4 die Anforderungen an die Plattform zur Weitverkehrskommunikation und verwendet darüber hinaus die dynamischen Ressourcenmodelle aus A4. Durch den speziellen Fokus auf innerstädtische Umgebungen, und die damit einhergehenden, stark schwankenden Funkkanalbedingungen, kommt insbesondere der vom A4-Projekt adressierten Modellierung des Kontextes eine besondere Bedeutung zu. Darüber hinaus ergeben sich aus dem B4-Projekt Anforderungen an die Modellierung von z.B. der Verfügbarkeit verschiedener LTE(-A)-Funkfrequenzen im Projekt A4. Aufgrund der stark ausgeprägten Synergien zwischen den Teilprojekten, beispielsweise im Bezug auf die Modellierung von funkkanalabhängigen Datenraten, konnten bereits im Verlauf der ersten Phase zahlreiche teilprojektübergreifende Veröffentlichungen erarbeitet werden. Zur Abgrenzung wird im Folgenden zudem noch auf Teilprojekt A2 eingegangen. Wie A4 befasst sich auch das Teilprojekt A2 mit der Modellierung der Ressource Energie. Dabei liegt der Fokus jedoch auf Algorithmen zur Datenanalyse und deren CPU- und Hauptspeichernutzung. A4 adressiert dagegen die Systemsoftware kleinerer eingebetteter Systeme und die Kosten der Sensorik und Kommunikation. A2 kann auf den Ressourcenmodellen von A4, insbesondere für Kommunikationskosten im Kontext verteilter Algorithmen, aufbauen. Darüber hinaus ist geplant, A2 die in A4 entwickelte MIMOSA-Messtechnik zur Verfügung zu stellen, um deren Ergebnisse auch auf kleineren eingebetteten Systemen zu validieren.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

LavA

(Spinczyk,Förderkennzeichen DFG, SP 968/4-2) (Förderzeitraum: 2013 - 2015)

Durchgängige anwendungsspezifische Maßschneiderung von Betriebssystem und Rechnerstruktur in konfigurierbaren eingebetteten Systemen. Der in LavA verfolgte Ansatz ist komplementär zu dem in diesem Antrag verfolgten Ansatz: Während in TP A4 von einer Hardware mit messbarem Ressourcenverbrauch ausgegangen wird, geht es bei LavA um die Synthese von Hardwarestrukturen.

DanceOS

(Spinczyk,Förderkennzeichen DFG, SP 968/5-2) (Förderzeitraum: 2013 - 2014)

Erforschung neuer softwaretechnischer Ansätze zur Realisierung softwarebasierter Fehlertoleranzmaßnahmen mit Hilfe sogenannter Aspekte. Dabei liegt der Fokus auf der Härtung von Betriebssystemcode gegen das temporäre oder permanente Versagen von Hardware. Es gibt Anknüpfungspunkte zu Arbeitspaket 8, da hier Fehlertoleranz für Black-Out-Überbrückung ein Rolle spielt.

Fehlertolerante und echtzeitfähige System- und Kommunikationsinfrastrukturen für die Schutz- und Leittechnik

(Spinczyk,Förderkennzeichen DFG, SP 968/6-1) (Förderzeitraum: 2012 - 2014)

Konzeption einer auf Virtualisierung basierenden Systemarchitektur, die Fehlertoleranz und Echtzeitfähigkeit bietet und damit im Bereich der CPS zur Hardwarekonsolidierung eingesetzt werden kann. Es gibt keine Überschneidungen zu dem beantragten Vorhaben.

TIE-IN

(Wietfeld,Förderkennzeichen Ziel 2, 64.65.69-EM-1022A) (Förderzeitraum: 2011 - 2014)

Entwurf einer modellbasierten Test- und Prüfumgebung für Kommunikationsprotokolle für eine PLC-Übertragungstrecke (*Power Line Communications*) im Bereich der Elektromobilität.

openECOSPhERE

(Wietfeld,Förderkennzeichen BMWi 01ME12091) (Förderzeitraum: 2011 - 2014)

Entwurf, Analyse und Bewertung IKT-basierter Strategien zur Erbringung von Systemdienstleistungen durch Elektrofahrzeuge vor dem Hintergrund der gezielten Nutzung fluktuierender erneuerbarer Energieträger zur Ladung von Elektrofahrzeugen. Im Fokus steht dabei die Steuerung von Ladepunkten mittels mittel PLC, Mobilfunk und dem *Open Charge Point Protocol* (OCPP).

e-DASH

(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 285586) (Förderzeitraum: 2011 - 2014)

Erforschung eines funkbasierten Kommunikationsdienstes zur Steuerung von Fahrzeugflotten zur Erbringung von Energiedienstleistungen mit Fokus auf Backendsysteme.

metropol-E

(Wietfeld,Förderkennzeichen BMVBS 03EM0605D) (Förderzeitraum: 2012 - 2014)

Erforschung der Kommunikationsdienste für ein neuartiges Informationssystem für mobile Dienste in der Elektromobilität mit dem Fokus aus der Planung und Steuerung einer kommunalen Fahrzeugflotte unter Nutzung von Strom aus regenerativer Eigenerzeugung.

SmartC2Net

(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 318023) (Förderzeitraum: 2012 - 2015)

Entwurf und Validierung einer integrierten Systemarchitektur für einen sicheren und zuverlässigen Smart-Grid-Betrieb mit besonderem Fokus auf neuartige, miteinander verknüpfte energie- und kommunikationstechnischen Regelkreise.

AIRBEAM

(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 261769) (Förderzeitraum: 2012 - 2015)

Entwurf und Validierung von Kommunikationsprotokollen für heterogene, flugfähige, vernetzte Sensorplattformen für den Einsatz im Rettungswesen und Katastrophenschutz.

Anchors

(Wietfeld,Förderkennzeichen BMBF 13N12204) (Förderzeitraum: 2012 - 2015)

Erforschung von neuartigen drahtlosen Kommunikationsdiensten zur redundanten Vernetzung autonomer, unbemannter Robotiksysteme in der Luft und am Boden mit dem Ziel der Fernerkundung bei Störfällen in Industrieanlagen.

eNterop

(Wietfeld,Förderkennzeichen BMWi 01MX12060D) (Förderzeitraum: 2012 - 2014)

Entwurf einer Referenzimplementierung interoperabler Kommunikation zwischen Elektrofahrzeug und Ladesäule sowie Erforschung einer automatisierten TTCN-3-basierten (*Testing and Test Control Notation*) Konformitätstestumgebung in direkter Kooperation mit der Automobilindustrie und Infrastrukturanbietern.

DFG-Forschergruppe 1511
(Wietfeld,Förderkennzeichen Wi 3751/1-1, Wi 3751/2-1) (Förderzeitraum: 2012 - 2014)

Weiträumige Systemüberwachung und Schutz von elektrischen Energiesystemen. Erforschung einer neuen, hybriden Simulationsplattform zur gemeinsamen Untersuchung von Energie- und IKT-Systemen. Konzeptionierung und Untersuchung neuartiger, echtzeitfähiger Kommunikationsinfrastrukturen unter Einbeziehung von *Software-Defined-Networking*-Konzepten.

PlanGridEV
(Wietfeld,Förderkennzeichen EU 608957) (Förderzeitraum: 2013 - 2016)

Entwicklung neuartiger Planungsregeln und operativer Prinzipien für die optimale Integration von Elektrofahrzeugen in zukunftsorientierte Energienetze.

SecInCoRe
(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 607832) (Förderzeitraum: 2014 - 2017)

Verbesserung der grenzüberschreitenden Zusammenarbeit von Polizei, Rettungsdiensten, Feuerwehren und sonstigen Rettungsorganisationen auf Basis eines cloud-basierten gemeinsamen Informationsraums.

Realtime Logistics - Echzeitfähige Logistik
(ten Hompel,Förderkennzeichen DFG, HO 2403/9-1) (Förderzeitraum: 2012 - 2015)

Entwicklung und Analyse echtzeitfähiger Logistiksysteme auf Basis agentenbasierter Steuerungen.

Analytische Methoden für die Berechnung der Leistungsverfügbarkeit komplexer Materialflusssysteme
(ten Hompel,Förderkennzeichen DFG, HO 2403/8-1) (Förderzeitraum: 2012 - 2015)

Entwicklung eines integrierten Modells der Leistung und Verfügbarkeit herkömmlicher intralogistischer Systeme, das im Rahmen der Systemplanung eine Prognose der Leistungsverfügbarkeit ermöglicht.

DyCoNet
(ten Hompel,Förderkennzeichen 01MA09057) (Förderzeitraum: 2010 - 2014)

Hard- und Softwareentwicklung für Full-Composite Lightweight Luftfrachtcontainer, welche in bestehende Flughafen-IT-Systeme integriert werden. Keine Betriebssystementwicklung.

smaRTI
(ten Hompel,Förderkennzeichen 01|C10L10A) (Förderzeitraum: 2010 - 2014)

Entwicklung intelligenter Ladungsträger und Logistikkonzepte zur Gestaltung von Prozessen, Technologien und Software, die der Identifizierung, Lokalisierung und Steuerung logistischer Systeme dienen. Keine Hardware- oder Betriebssystementwicklung.

Smart Face-Smart Micro Factory für Fahrzeuge mit schlanker Produktionsplanung
(ten Hompel,Förderkennzeichen BMBF 01MA13007D) (Förderzeitraum: 2014 - 2016)

Anpassung von Fertigungsstrukturen an die Anforderungen einer Kleinserienfertigung.

Inventairy
(ten Hompel,Förderkennzeichen BMBF 01MA13006A) (Förderzeitraum: 2014 - 2016)

InventAIRy ist in den Bereich der Cyber-Physical Systeme einzuordnen und verfolgt das Ziel der Entwicklung eines autonomen Flugrobotersystems auf Basis des Internets der Dinge.

EBS Fraunhofer Supply Chain School
(ten Hompel,Förderkennzeichen BMBF 01/C10L08D) (Förderzeitraum: 2011 - 2015)

Entwicklung eines serviceorientierten Instrumentariums zur Gestaltung von Logistiksystemen in globalen Produktionsnetzwerken.

Hub2Move
(ten Hompel,Förderkennzeichen BMBF 01/C12L02A) (Förderzeitraum: 2013 - 2015)

Entwickelt wird ein umzugsfähiges Umschlagszentrum, das sich einfach an veränderte Anforderungen anpassen lässt. Handhabungs-, Förder- und Lagertechnik im Hub können in wenigen Tagen anpassbar sein, ein kompletter Hub kann an einem neuen Standort innerhalb weniger Wochen lauffähig werden.

InKoRisk

(ten Hompel,Förderkennzeichen BMBF 19G11003D) (Förderzeitraum: 2014 - 2016)

Integrierte Terminierung und Transportplanung unterstützt durch kollaboratives Risikomanagement in der Automobilindustrie.

PlaSkoPro

(ten Hompel,Förderkennzeichen BMBF 01MA13007D) (Förderzeitraum: 2013 - 2016)

Planung von Steuerungskonzepten für die schlanke Produktionsplanung.

Prestige

(ten Hompel,Förderkennzeichen BMBF 16KIS0086) (Förderzeitraum: 2014 - 2017)

Privacyerhaltende Methoden und Werkzeuge für cloudbasierte Geschäftsprozesse.

VILOMA Visual Logistics Management

(ten Hompel,Förderkennzeichen BMBF 19G13003D) (Förderzeitraum: 2013 - 2016)

Das Projekt verfolgt das Ziel, die logistischen Prozesse in komplexen Wertschöpfungsnetzwerken besser plan- und steuerbar zu machen. Um dies zu erreichen, sollen vorhandene Visualisierungsansätze aus anderen Forschungsrichtungen aufgegriffen und neue Methoden entwickelt, Informationen verdichtet und zielgruppenspezifisch aufbereitet werden.

APEM Automatische Paletten Etikettier-Maschine

(ten Hompel,Förderkennzeichen BMBF 02PK2375) (Förderzeitraum: 2013 - 2015)

Entwicklung von Produktionsmaschinen zur vollautomatischen Applizierung von RFID-Transpondern auf Euro-Holzpaletten.

CACTUS

(ten Hompel,Förderkennzeichen BMVI 03EMEN06B) (Förderzeitraum: 2012 - 2015)

Modelle und Methoden für die Evaluierung und die optimale Anwendung von Batterielade- und wechseltechnologien für Elektrobusse

ModuLOG

(ten Hompel,Förderkennzeichen NRW Bank 300 225 202) (Förderzeitraum: 2012 - 2015)

Masterplan Strategien für Logistik & Supply Chain in der modularen Chemie- und Pharmaproduktion.

LOG2020

(ten Hompel,Förderkennzeichen EU 527700-LLP-1-2012-1-NL-ERASMUS-EMCR) (Förderzeitraum: 2012 - 2014)

Logistics and Supply Chain Management 2020: training for the future. Ziel ist die Entwicklung eines Masterprogramms für Rollen und Berufe in der Logistik.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	3	176.400	3	176.400	3	176.400	3	176.400
Summe:	—	176.400	—	176.400	—	176.400	—	176.400
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	176.400		176.400		176.400		176.400	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (A4)							
Wissen-schaftliches Personal	1	ten Hompel, Michael, Prof. Dr., Prof	Logistik	LS Förder- und Lagerwesen, Fak. Maschinenbau, TU Do	4,0	—	Grund-ausstattung
	2	Spinczyk, Olaf, Prof. Dr., Prof	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	4,0	—	Grund-ausstattung
	3	Wietfeld, Christian, Prof. Dr., Prof	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	4,0	—	Grund-ausstattung
	4	Masoudinejad, Mojtaba, M.Sc. WiMi	Logistik	LS Förder- und Lagerwesen, Fak. Maschinenbau, TU Do	11,94	—	Grund-ausstattung
	5	Roidl, Moritz, Dipl.-Inf. WiMi	Logistik	LS Förder- und Lagerwesen, Fak. Maschinenbau, TU Do	15,92	—	Grund-ausstattung
	6	Böcker, Stefan, Dipl.-Ing. WiMi	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	15,92	—	Grund-ausstattung
	7	Borchert, Christoph, Dipl.-Inf. WiMi	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	11,94	—	Grund-ausstattung
	8	N.N. , WiMi	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	11,94	—	Grund-ausstattung

	9	N.N. , SHK	Logistik	LS Förder- und Lagerwesen, Fak. Maschinenbau, TU Do	10,0	—	Grund-ausstattung
	10	N.N. , SHK	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	10,0	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	11	Sondhof, Uwe, Techn. Ang.	—	LS Förder- und Lagerwesen, Fak. Maschinenbau, TU Do	5,97	—	Grund-ausstattung
	12	Foese, Matthias, Techn. Ang.	—	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	5,97	—	Grund-ausstattung
Beantragte Personalmittel (A4)							
Wissen-schaftliches Personal	13	Emmerich, Jan Sören, Dipl.-Inf. WiMi	Logistik	LS Förder- und Lagerwesen, Fak. Maschinenbau, TU Do	—	Doktorand/in	—
	14	Kaulbars, Dennis, M.Sc. WiMi	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	—	Doktorand/in	—
	15	Buschhoff, Markus, Dipl.-Inf. WiMi	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. ten Hompel, Michael
Projektleitung. Schwerpunkt Logistik. Mitarbeit in den Arbeitspaketen 1, 2, 5, 9 und 10.

2. Spinczyk, Olaf
Projektleitung. Schwerpunkt Rechnerarchitekturen und eingebettete Systeme. Mitarbeit in den Arbeitspaketen 1, 6, 7, 8, 9 und 10.

3. Wietfeld, Christian
Projektleitung. Schwerpunkt Kommunikationsnetze. Mitarbeit in den Arbeitspaketen 1, 2, 3, 4 und 9.

4. Masoudinejad, Mojtaba
Mitarbeit in den Arbeitspaketen 1, 2, 5, 9 und 10, Energy Harvesting, Simulation und Testbed.

5. Roidl, Moritz
Mitarbeit in den Arbeitspaketen 1, 2, 5, 9 und 10, Szenariendefinition, Modellierung des Kontextes sowie der Simulationskopplung.

6. Böcker, Stefan
Mitarbeit in den Arbeitspaketen 1, 2, 3, 4 und 9, Modellierung des Kontextes sowie von Kommunikationssystemen.

7. Borchert, Christoph
Mitarbeit in den Arbeitspaketen 1, 6, 7, 8, 9 und 10, Entwurf von Systemsoftware.

8. N.N.,
Unterstützung im AP 1, 2, 3, 4 und 9.

9. N.N.,
Unterstützung der wissenschaftlichen Mitarbeiter in den relevanten APs.

10. N.N.,
Unterstützung der wissenschaftlichen Mitarbeiter in den relevanten APs.

11. Sondhof, Uwe
Unterstützung der technischen Infrastruktur für Simulatorkopplung und Testbed.

12. Foese, Matthias
Unterstützung der technischen Infrastruktur für Datenbankzugriffe zwischen den Instituten.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

13. Emmerich, Jan Sören
Mitarbeit in den Arbeitspaketen 1, 2, 5, 9 und 10. Hierbei insbesondere bei der Modellierung des Kontextes, Energy Harvesting und Testbed

14. Kaulbars, Dennis
Mitarbeit in den Arbeitspaketen 1, 2, 3, 4 und 9, hierbei insbesondere bei der Modellierung des Kontextes sowie von Kommunikationssystemen.

15. Buschhoff, Markus
Mitarbeit in den Arbeitspaketen 1, 6, 7, 8, 9 und 10.

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	9.000	9.000	9.000	9.000
Summe der vorhandenen Mittel	9.000	9.000	9.000	9.000
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt A5

3.1.1 Titel:

Austausch und Fusion von Information unter Verfügbarkeits- und Vertraulichkeitsanforderungen in Multiagentensystemen

3.1.2 Projektleitung

Prof. Dr. Biskup, Joachim, 11.07.1947

Lehrstuhl 6 ISSI, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 14
44227 Dortmund

Tel.: 0231-755-2569
Fax : 0231-755-2405
E-Mail: `joachim.biskup@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

( ) nein (x) ja, befristet bis zum 30.04.2015

Prof. Dr. Kern-Isberner, Gabriele, 22.01.1956

LS 1 IE, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 12
44227 Dortmund

Tel.: 0231-755-2045
Fax : 0231-755-2405
E-Mail: `gabriele.kern-isberner@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.2 Entwicklung des Teilprojekts

3.2.1 Bericht

Aufgabenstellung und Ausgangslage
Gemäß der Antragsstellung aus dem Jahr 2010 sollten unsere Arbeiten dem technischen Fernziel dienen, ein neuartiges Rahmenwerk für die Durchsetzung von Verfügbarkeit und Vertraulichkeit in ressourcenbeschränkten Multiagentensystemen zu entwickeln, das insbesondere für Anwendungen im Bereich Data Mining in Embedded Systems einsetzbar ist. Dafür sollten in der Berichtsphase drei aufeinander bezogene Gesichtspunkte behandelt werden:

1. Architektur, vorrangig hinsichtlich interner Operationen eines Agenten zur Entscheidungsfindung und Interaktionen zwischen Agenten;
2. Modellbildung, vorrangig hinsichtlich Wissensrepräsentation und Inferenzverhalten von Agenten;
3. Ressourcenbeschränkung der Agenten.

Die vorgesehenen Untersuchungen sollten sich auf die unterschiedlichen Vorarbeiten der Arbeitsgruppen Biskup beziehungsweise Kern-Isberner stützen. Einerseits verfügte die Arbeitsgruppe Biskup über reiche Erfahrungen im Gebiet des Vertraulichkeitsschutzes durch Inferenzkontrolle in logikorientierten Client-Server Informationssystemen, wobei insbesondere ein objektiver, deklarativer Vertraulichkeitsbegriff verwendet wird, der neben den lokalen Zuständen der Beteiligten auch das globale Systemverhalten umfassend berücksichtigt. Andererseits verfügte die Arbeitsgruppe Kern-Isberner über reiche Erfahrungen im Gebiet der Repräsentation und Verarbeitung von unsicherem und unvollständigem Wissen für autonome Agenten in

Multiagentensystemen, wobei insbesondere ein subjektiver Zugang zum Verhaltens eines einzelnen Agenten im Hinblick auf seine Entscheidungsfindung und in diesem Sinne das Ziel einer subjektiven Vertraulichkeit aus der lokalen Sicht des Agenten verfolgt wird.

Übersicht, Einordnung und Bewertung der Ergebnisse
Die bisherigen Arbeiten in der Berichtsphase lieferten wichtige Fortschritte zum allgemeinen Fernziel und den genannten drei weitgefassten Gesichtspunkten. Im Folgenden beschreiben wir diese Fortschritte unter fünf stärker eingegrenzten Gesichtspunkten:

1. Operationen/Interaktionen (als Eingrenzung von Architektur),
2. Inferenzverhalten (als Eingrenzung von Modellbildung) und
3. Effizienz (als Eingrenzung von Ressourcenbeschränkung),

sowie teilweise über den Antrag hinausgehend und übergreifend

4. Vereinheitlichung (bezogen auf das Fernziel) und
5. Modellbildung (im Sinne von vertiefter Konzeptbildung).

Vorweg und kurz zusammengefasst belegen die erzielten Fortschritte insbesondere Dreierlei: für wegen ihrer semantischen Reichhaltigkeit höchst anspruchsvolle Anforderungen hinsichtlich Verfügbarkeit und Vertraulichkeit können algorithmische Lösungen mit beweisbaren Sicherheitseigenschaften tatsächlich gefunden werden; inhaltliche Beherrschbarkeit der Anforderungen mit ihren algorithmischen Lösungen bedürfen einer weitergehenden Vereinheitlichung und Parametrisierung; und für den praktischen Einsatz sind im Allgemeinen abgeschwächte Anforderungen, insbesondere hinsichtlich optimaler Verfügbarkeit notwendig.

Unsere Beschreibungen beziehen sich im Allgemeinen auf ein fokussiertes Szenario von zwei intelligenten, autonomen und kooperierenden Agenten. Dabei betrachten wir einen der Agenten in der Rolle eines verteidigenden Agenten, der Vertraulichkeit gegenüber seinem Interaktionspartner durchsetzen will. Dieser Interaktionspartner wird daher im Sinne des Vertraulichkeitsschutzes in der Rolle eines angreifenden Agenten gesehen, der seine Mittel vollständig ausschöpft, um Vertrauliches aufzudecken. Das Inferenzverhalten des verteidigenden Agenten ist durch die anwendungsorientierten Anforderungen zum Einsatz eines wissensbasierten Agenten motiviert. Ebenso anwendungsorientiert werden die Interaktionsmöglichkeiten aller Agenten zum Zweck des Informationstausches und mit dem Fernziel der Kooperation bestimmt. Darüber hinausgehend wird unter dem Aspekt der Vertraulichkeit dem angreifenden Agenten ein entsprechend den Annahmen ausgewähltes Inferenzverhalten zugesprochen, das dieser als Mittel einsetzt, um Vertrauliches zu folgern. Darauf basierend wird dem verteidigenden Agenten ein weitergehendes Inferenzverhalten beigestellt, das dieser als Mittel der Entscheidungsfindung zur Vertraulichkeitswahrung einsetzen kann, um das Inferenzverhalten des angreifenden Agenten nachzubilden.

Bezüglich der Operationen/Interaktionen können neben der Beantwortung von Anfragen nunmehr auch Auswertungen von Änderungs- und Revisionsanforderungen [A5/209, A5/216] unter objektiven Vertraulichkeitsanforderungen eingesetzt werden, wobei insbesondere die jeweiligen Invarianten bezüglich der Integrität der behandelten Information einerseits und der durchzusetzenden Vertraulichkeit andererseits mitsamt deren Konflikten berücksichtigt werden. Darüber hinaus werden fortlaufend auch degenerierte Interaktionen behandelt: effiziente Fragmentierungen von Datenbankrelationen [A5/207], Veröffentlichungen von Data Mining Ergebnissen von relationalen Datenbanken unter Abwehr von sowohl rein logischen als auch kombinatorischen Inferenzen, sowie effizienten Veröffentlichungen von für die Vertraulichkeit abgeschwächten XML-Dokumenten [A5/206] und Datenbankrelationen in einer laufenden Arbeit.

Weiterhin wurden bezüglich Operationen/Interaktionen mit dem Ziel der subjektiven Vertraulichkeit basierend auf einer erweiterten Modellbildung für wissensbasierte autonome Agenten notwendige Eigenschaften der Änderungs-, Deliberations- und *Means-ends-reasoning*-Operationen eines derart vertraulichkeitswahrenden Agenten definiert und Realisierungen entsprechender Operationen erarbeitet [A5/213, A5/210]. Ferner wurden die Änderungsoperationen von Agenten im Speziellen betrachtet und und es wurden für nichtmonotone Logiken (insbesondere für die Antwortmengenprogrammierung) Revisionsformalismen für Wissensbasen vorgestellt [A5/212], da die Repräsentationsform der Wissensbasis zum einen besser für die Antwortmengenprogrammierung geeignet ist und zum anderen Basisrevision im Hinblick auf Einsparung von Ressourcen erfolgversprechend ist. Weitere Arbeiten, [A5/214] und eine Masterarbeit, verwenden Argumentationsansätze zur Unterstützung von Revisionsverfahren.

Grundsätzlich kann man mit den in diesem Zusammenhang entwickelten Methoden die Ausdrucksfähigkeiten hinsichtlich der Interaktionen mit zusätzlichen Einzelfällen immer weiter steigern, wodurch man immer realistischer würde bezüglich tatsächlicher Verhältnisse von Anwendungen. Um die sich ergebende Vielfalt allerdings beherrschbar zu halten, wird zukünftig der bislang nur einführend betrachtete Gesichtspunkt Vereinheitlichung immer bedeutsamer.

Bezüglich des Inferenzverhaltens können nunmehr neben klassischer Aussagenlogik und Fragmenten der Prädikatenlogik auch ausgewählte nichtmonotone Logiken zur Beschreibung von subjektivem Wissen (Glauben) [A5/212, A5/213, A5/209, A5/210, A5/216] eingesetzt werden. Dabei wurden zunächst zwei typische, in unterschiedlichen semantischen Umgebungen angesiedelte Formalismen für das Inferenzverhalten eines seine Vertraulichkeitsziele verteidigenden Agenten betrachtet. Diese sind ordinale konditionale Funktionen (*ordinal conditional functions*) als Vertreter der semi-quantitativen, semantischen Formalismen in [A5/209, A5/216] sowie die Antwortmengenprogrammierung als Vertreter syntaxbasierter Default-Logiken in [A5/213]. Ergänzend zu den Beiträgen in [A5/209, A5/216] zur objektiven Vertraulichkeitswahrung bei komplexen Interaktionsfolgen und ausgehend von der Modellbildung zur subjektiven Vertraulichkeitswahrung [A5/213] wurde in [A5/210] das Folgern eines Agenten über das Einhalten von Vertraulichkeitseinschränkungen (*secrecy constraints*) unter einer einzelnen Informationsaktion betrachtet. Insbesondere wurde in [A5/209, A5/216] unter einer entscheidenden Vollständigkeitsannahme, die dem verteidigenden Agenten vollständige Kenntnis der Mittel des angreifenden Agenten zuspricht, die komplexe Inferenz des angreifenden Agenten betrachtet, die implizite Informationsflüsse bei der Interaktion mit dem verteidigenden Agent ausnutzt. Ausgehend von sorgfältig erörteten Vereinfachungen wurde in [A5/210] ein solches Inferenzverhalten nicht betrachtet, während die Vollständigkeitsannahme jedoch wesentlich abgeschwächt wurde. Als Folge der nunmehr teilweisen Unkenntnis über den angreifenden Agent muss der verteidigende Agent begründete Entscheidungen angesichts potenzieller Verletzungen von Vertraulichkeitseinschränkungen treffen, genauso wie ein Agent bei normativem Folgern (normative reasoning) potenzielle Verletzungen einer Norm in seinen Entscheidungsprozess einbeziehen muss.

Wiederum kann man grundsätzlich auch bei der Betrachtung des Inferenzverhaltens der eingesetzten Methoden die Ausdrucksfähigkeiten mit zusätzlichen Einzelfällen immer weiter steigern, wodurch man immer realistischer würde, aber auch den bislang nur einführend betrachteten Gesichtspunkt Vereinheitlichung vertieft beachten müsste.

Darüber hinaus betrachteten wir auch die Herausforderungen für den Umgang mit Bewegungsdaten in einem Beitrag zu MDM 2013 und für integrierte Informationssysteme [A5/208].

Bezüglich der Effizienz verschärften sich durch die erweiterten Möglichkeiten für sowohl Operationen/Interaktionen als auch Inferenzen die bekannten Probleme hinsichtlich Laufzeiteffizienz und Skalierbarkeit. Diese Probleme wurden durch Beiträge zu zwei Ansätzen bearbeitet, die insbesondere praktische Effizienz mit Reaktionszeiten bei Interaktionen höchstens im Sekundenbereich anstreben. Diese Ansätze behandeln einerseits statische Kontrollen in Form der degenerierten Interaktionsformen Fragmentierung (*data fragmentation*) [A5/207] und Veröffentlichung (*data publishing*) in [A5/206] sowie in laufenden Arbeiten und andererseits Laufzeitoptimierungen für interessante Spezialfälle [A5/205, A5/211], die auch in studentischen Abschlussarbeiten vertiefend behandelt wurden. Zukünftig wird es aber wohl immer bedeutsamer, nach effizienten Annäherungen mit immer noch beweisbaren Vertraulichkeitszusicherungen unter Preisgabe (zu) hoher Verfügbarkeitsanforderungen zu suchen.

Zur Realisierung von Vertraulichkeit auf subjektiver Ebene ist die adäquate Berücksichtigung von unsicherem Wissen in der Form nichtmonotoner Schlussfolgerungsmechanismen unabdingbar. Einer der erfolgreichsten Formalismen insbesondere im Hinblick auf Laufzeiteffizienz und Skalierbarkeit ist hier die Antwortmengenprogrammierung, die in einigen Bereichen bereits sehr erfolgreich eingesetzt worden ist [203]. So wurde die Antwortmengenprogrammierung für die konkrete Ausprägung von Agentenarchitekturen für die subjektive Vertraulichkeitswahrung auf Ebene der Inferenz [A5/213], der Änderungsoperationen [A5/212] und der Durchsetzung von subjektiver Vertraulichkeit (in einer Diplomarbeit) betrachtet. Die betrachteten Änderungsoperationen [A5/212, A5/214] setzten zudem auf der Theorie der Basisrevision auf, welche große Vorzüge in Bezug auf die Effizienz im Vergleich mit anderen Revisionsmethoden aufweist.

Im Zuge der weiteren Arbeiten zur Vereinheitlichung verfolgten wir auch über den Antrag hinausgehend zwei Vorgehensweisen. Zum einen wurden die Modellbildungen zur objektiven Vertraulichkeitswahrung unter Interaktionsfolgen als grundsätzlich parameterisierbar angelegt. Diese Vorgehensweise wurde in zusammenfassenden Beiträgen verdeutlicht, die Parameter der Inferenzkontrolle aus Vorgängerarbeiten unter verschiedenen Gesichtspunkten kategorisieren und vereinheitlichen. Beispielhaft für den Parameter Inferenzverhalten der Agenten wurde zusätzlich ein erster Ansatz zur Vereinheitlichung erarbeitet, indem Klassen von Inferenzverhalten betrachtet wurden, einerseits durch hinreichende Eigenschaften von Axiomatisierungen solcher Klassen in [A5/209, A5/216] für objektive Vertraulichkeit und andererseits durch notwendige

Eigenschaften von Wissensoperationen in [A5/210] zur subjektiven Bewertung potentieller Verletzungen von Vertraulichkeitseinschränkungen.

Zum anderen wurden Vertraulichkeitskontrollen für komplexe Interaktionen im Einzelfall dadurch entwickelt, dass die in den Einzelschritten enthaltenen impliziten Anfragen bestimmt und dann deren Zusammenwirken mit den Kontrollstrukturen der Interaktionen analysiert wurden [A5/209, A5/216]. Eine genauere Weiterentwicklung dieser Vorgehensweisen hinsichtlich Verallgemeinerungen wird dann schnell auf die bekanntermaßen sehr schwierigen Probleme der Eigenschaftserhaltung unter Verfeinerungen (refinement problem) und Kompositionen (compositionality problem) stoßen.

Die Arbeiten zu Operationen unter subjektiver Vertraulichkeitswahrung dienten neben der Vereinheitlichung vorrangig der Modellbildung [A5/213, A5/210, A5/215]. In [A5/213] und einer Vorarbeit wurden die relevanten Aspekte eines Agentenmodels für die Vertraulichkeitswahrung herausgearbeitet. Hierbei wurde ein generelles Agentenmodel, welches bekannte Architekturen wie das Beliefs-Desires-Intentions (BDI) Modell erweitert, entwickelt und in diesem Eigenschaften von Komponenten des Agenten für die Vertraulichkeitswahrung definiert. Basierend darauf wurde eine Klasse von Agentenmodellen spezifiziert, die im Wesentlichen ausgehend von den subjektiven Vertraulichkeitsinteressen eines Agenten in seinem epistemischen Zustand Eigenschaften der zustandsverändernden Operationen in Bezug auf die Wahrung dieser Interessen definiert. Hieraus ergibt sich die Definition eines (subjektiv) vertraulichkeitswahrenden Agenten.

Auf Grundlage dieser erweiterten Agentenmodelle wurde das *Angerona* Rahmenwerk zur Konzipierung und Implementierung von Multiagentensystemen entwickelt [A5/215]. Dieses Plug-in basierte *JAVA* Rahmenwerk setzte die Modellbildung auf Implementierungsebene fort, wobei seine modulare Architektur die Umsetzung verschiedenster Ansätze bezüglich Agentenoperationen, -interaktionen und -inferenzen ermöglichte. Unter anderem wurden die Beiträge aus [A5/213, A5/209, A5/216, A5/212, A5/214] und dreier studentischer Abschlussarbeiten in dem Rahmenwerk umgesetzt. In diesem Rahmenwerk konnten so unter pragmatischeren Anforderungen verschiedene Arbeitsstränge zusammengeführt werden. Insbesondere wurden Komponenten zur Interaktionskontrolle und Agentensimulation, die für ein vereinfachtes Agentenmodell, aber umfassenderes Interaktionsmodell in [A5/209, A5/216] als hinreichend für die objektive Vertraulichkeit nachgewiesen wurden, in das erweiterte BDI-Modell des Rahmenwerks integriert.

Einzelarbeiten zu weiteren Interaktionen und Anwendungen von Inferenzen gemäß Anforderungen aus anderen Teilprojekten konnten bislang nur ansatzweise realisiert werden. So wurde in Kooperation mit dem Teilprojekt B4 ein auf Multiagenten basierender Ansatz für die Durchsetzung von Vertraulichkeits- und Verfügbarkeitsanforderungen bei der Datenerhebung zur Verkehrsprognose erstellt und im Rahmen einer Diplomarbeit implementiert. Allerdings ergaben sich in anderen Zusammenarbeiten bislang noch keine überzeugenden Modellierungen der in den jeweiligen Teilprojekten behandelten Anwendungen als Multiagentensysteme.

In einem größeren Zusammenhang gesehen, fügen sich unsere Ergebnisse in die folgenden aktuellen Forschungslinien ein, wobei eine Besonderheit unseres Vorhabens in der übergreifenden Zusammenführung der in diesen Linien verfolgten Ansätze besteht:

- Informationsflusskontrolle (*information flow control*) [199] – insbesondere Spezifikation von nicht erlaubtem Wissen,
- Sicherheitsautomaten (*security automata*) [200] – insbesondere Entwurf von Kontrollmechanismen für Sicherheitspolitiken,
- Vertraulichkeit bewahrende Datenveröffentlichungen (*privay preserving data publishing*) [202],
- normatives Verhalten (*normative behavior*) [198] – insbesondere normativer Status von Sicherheitspolitiken,
- Theorien der Wissensänderung [201] und
- Interaktionen in Multiagentensystemen [204].

Externe Zusammenarbeiten konnten wie folgt durchgeführt werden:

- mit Sebastian Link (University of Auckland)/Sven Hartmann (TU Clausthal) in [A5/211] und in einer grundlegenden Arbeit zur konditionalen Unabhängigkeit,
- mit Piero Bonatti (Universita di Napoli Frederico II) in einer laufenden Arbeit zu komplexitätstheoretischen Untersuchungen der Inferenzkontrolle,

- mit Lan Li (Shanghai Jiao Tong University) in [A5/206],
- mit Chris Clifton (Purdue University) und anderen in einem Beitrag zu Bewegungsdaten,
- mit Guillermo Simari (Universidad National del Sur, Argentinien) und Matthias Thimm (jetzt Uni Koblenz/Landau) in einer Arbeit zur nicht priorisierten Wissensrevision von Wissensbasen [A5/214].

Zusammenfassungen von Einzelarbeiten

Im Folgenden beschreiben wir die in den wichtigsten Schriften enthaltenen Ergebnisse im Einzelnen.

In unserem Konferenzbeitrag für FoIKS 2012 und dem darauf aufbauenden Zeitschriftenartikel in AMAI [A5/209] behandelten wir aufeinander bezogene Gesichtspunkte bezüglich Interaktionen und Inferenzverhalten unter objektiver Vertraulichkeitswahrung. Neben kontrollierten Anfragen wurde erstmals die Kontrolle der Interaktion der Revision in einer nichtmonotonen Logik betrachtet. Dabei beschränkten wir uns auf Revisionsanforderungen, die entweder ganz zurückgewiesen oder ganz durchgeführt werden sollen; weitergehende Betrachtungen für Revisionen mit zusätzlichen Änderungen und Anpassungen haben wir bislang nur ansatzweise in Bachelorarbeiten betrachten lassen. Als Grundlage für die algorithmische Kontrolle von Sequenzen von Anfragen und Revisionen führten wir ein Modell des kontrollierenden Agenten ein, dessen Kern aus einer Simulation des angreifenden Agenten unter beidseitigen Unsicherheiten über die Verhältnisse beim jeweiligen Interaktionspartner besteht. Dabei beschränkten wir uns bislang darauf, den angreifenden Agenten als skeptisch anzusehen. Während im Konferenzbeitrag noch beispielhaft nur ordinale konditionale Funktionen (OCFs) für die nichtmonotone Wissensverarbeitung in rein semantischer Form zugrunde gelegt wurden, konnten wir im Zeitschriftenartikel geeignete Eigenschaften eines Axiomensystems für nichtmonotone Logiken herausarbeiten. Auf diese Weise kann nicht nur ein Axiomensystem für nichtmonotone Inferenz mit Hilfe von OCFs behandelt werden, sondern vereinheitlichend für alle Klassen nichtmonotoner Inferenz, die den spezifizierten Eigenschaften genügen. Ferner ermöglicht ein solches Axiomensystem wenigstens grundsätzlich eine volle Algorithmisierung der Simulation der skeptischen Inferenz des angreifenden Agenten. Dieser Ansatz wurde beispielhaft für ordinale konditionale Funktionen in dem Rahmenwerk [A5/215] implementiert. Im Rahmen der Modellierung des kontrollierenden Agenten wurden insbesondere die Annahmen genau herausgearbeitet, unter denen der Beweis der Inferenzsicherheit der von uns entwickelten Kontrolle dann formal geführt werden konnte. Das Protokoll zur Kontrolle mittels Verweigerung beruht auf einer genauen Analyse des Zusammenspiels der in den Einzelschritten übermittelten Information mit den indirekten und impliziten Informationsflüssen durch die programmiersprachliche Kontrollstruktur des Protokolls. Dabei werden insbesondere solche Informationsflüsse erfasst, die durch eine Benachrichtigung über eine Verweigerung ausgelöst werden können (und somit sogenannte Metainferenzen ermöglichen).

In einer Vorgängerarbeit konnten wir als Workshopbeitrag für DPM 2011 bereits in einer klassischen Logik zur Wissensverarbeitung unter Vollständigkeit neben Anfragen auch Änderungen betrachten. Die erstmalige Analyse der Informationsflüsse des Protokolls für eine kontrollierte Änderungsoperation mittels Verweigerung, insbesondere des inhärenten Konflikts zwischen Integrität, Vertraulichkeit und Verfügbarkeit von Information, war grundlegend für die Beiträge in FoIKS 2012 und [A5/209].

In der Dissertation von Cornelia Tadros [A5/216] wurden dann unter anderem die Ergebnisse aus den Beiträgen in DPM 2011 und [A5/209] vereinheitlichend zusammengeführt und in vielen Einzelheiten vertieft. Eine zusätzliche Vereinheitlichung ging von einer weiteren Vorgängerarbeit als Konferenzbeitrag zu IWSEC 2010 aus, die bereits während der Antragsstellung entstand. In dieser Vorgängerarbeit ordneten wir die vorherigen Arbeiten zur kontrollierten Anfrageauswertung in den Ansatz Runs & Systems von Halpern/O'Neill zur Secrecy in Multiagent Systems genau ein. Insbesondere reicherten wir dazu den Ansatz von Halpern/O'Neill durch die Einführung von Vertraulichkeitspolitiken an. In der Dissertation [A5/216] wurde dann der Ansatz Runs & Systems mit Politiken vertieft ausgearbeitet und durchgehend für die vereinheitlichende Modellierung genutzt. Aufbauend auf die Vorgängerarbeit und den weitergehenden Ergebnissen aus der Dissertation [A5/216] bereiten wir derzeit eine umfangreichere Veröffentlichung über „Notions of Secrecy in Runs & Systems Based on Possibility Policies“ vor.

Hauptsächlich unter dem Gesichtspunkt Inferenzverhalten und teilweise unter dem Gesichtspunkt Vereinheitlichung wurde für FOIKS 2014 in [A5/210] das Folgern eines Agenten über das Einhalten von Vertraulichkeitseinschränkungen (*secrecy constraints*) durch deklarative Prinzipien und durch eine entsprechende, ansatzweise Algorithmisierung beschrieben. Dabei wurde eine einzelne Entscheidung des verteidigenden Agenten fokussiert, einen angreifenden Agenten über die Wahrheit einer Aussage zu informieren. Unter diesem Fokus konnten weitere bedeutsame Gesichtpunkte wie implizite Informationsflüsse und sequentielle Entscheidungen des verteidigenden Agenten begründet nicht behandelt werden. Unter diesen besonderen Voraussetzungen fielen dann die Sichtweisen der objektiven und der subjektiven Vertraulichkeitswahrung

weitgehend zusammen. Die deklarativen Prinzipien geben dem verteidigenden Agenten Richtlinien zur Klassifikation der betrachteten Informationsaktionen, wobei eine hohe Klassifikation ein Abraten von der Aktion ausdrückt. Diese Richtlinien beziehen sich einerseits darauf, wie der verteidigende Agent mit seiner teilweisen Unkenntnis über den angreifenden Agenten sinnvoll umgehen soll. Andererseits geben diese Richtlinien durch eine entsprechende Klassifikation Handlungsanweisungen im Falle einer Aktion des verteidigenden Agenten entgegen seiner Verpflichtung zur Vertraulichkeit, die durch andere Interessen des Agenten motiviert sein kann. Eine solche Form der Handlungsanweisung ist im Bereich des normativen Folgerns als *contrary to duty*-Verpflichtung bekannt.

Eine bislang noch unveröffentlichte Arbeit behandelt ausgehend von einer vorangegangenen Masterarbeit ebenfalls die Gesichtspunkte Interaktionen und Inferenzverhalten: Als Sonderfall einer Interaktion wird die Veröffentlichung von Data Mining-Ergebnissen über Datenbankrelationen betrachtet und neben rein logischen Inferenzen sind auch kombinatorische Inferenzen zu berücksichtigen. Das Besondere unseres Beitrags besteht im Wesentlichen darin, dass wir anstelle des binären Transaktionsmodells das ausdrucksstärkere relationale Datenmodell zugrunde legen und ansatzweise mögliches Vorwissen beim Empfänger berücksichtigen. In diesem Rahmen untersuchen wir zunächst intra-dimensionale Muster, die Beziehungen zwischen Werten eines einzelnen Attributs beschreiben. Aufgrund der Ähnlichkeiten zum binären Transaktionsmodell gelingt es, ein dort bekanntes Verfahren zum Entfernen aller sogenannten intra-dimensionalen k-Inferenzkanäle auf den relationalen Fall zu übertragen. Weiterhin untersuchen wir inter-dimensionale Muster, die Beziehungen zwischen Werten verschiedener Attribute beschreiben und damit die relationale Struktur der vorliegenden Daten weitergehend ausnutzen. Wir belegen die sich daraus ergebenden erhöhten Vertraulichkeitsprobleme und beweisen, dass der Kern dieser Probleme co-NP schwer ist. Ferner geben wir ein neues Verfahren zum Entfernen von inter-dimensionalen k-Inferenzkanälen an, das korrekt, aber nicht vollständig ist, und diskutieren, wie eine (notwendigerweise algorithmisch komplexe) Vollständigkeit erreicht werden könnte.

Im Rahmen unserer Arbeiten zum Gesichtspunkt Effizienz betrachteten wir in unseren Konferenzbeiträgen zu ISC 2011 und dann weiter vertiefend zu DBSEC 2013 [A5/207] den von anderen Autoren vorgeschlagenen Ansatz der Fragmentierung von Datenbankrelationen, der im Allgemeinen effizient durchführbar ist. Solche Fragmentierungen zielen vorrangig auf die Durchsetzung von Vertraulichkeitsanforderungen eines Anwenders ab, der seine Daten zu externen Dienstleistern auslagert: unter unterschiedlichen Modellierungen und Annahmen sollen die Dienstleister keine vertrauliche Information sehen oder rekombinieren können. In unseren Betrachtungen betteten wir die bislang nur grob auf Schemaebene spezifizierten Vertraulichkeitsanforderungen, die durch als sensitiv angesehene Attributmengen angegeben werden, in die in unserem Ansatz untersuchten feingranularen Vertraulichkeitspolitiken ein, die logikorientiert auf der Ebene von Aussagen ausgedrückt werden. Darüber hinaus untersuchten wir den Einfluss von Vorwissen auf den erreichbaren Schutz der Vertraulichkeit. Als Vorwissen beschränkten wir uns weitgehend auf semantische Bedingungen in Form von Gleichheits- und Tupel-erzeugende Abhängigkeiten für Datenbankrelationen. Die Hauptergebnisse von [A5/207] lassen sich kurz wie folgt zusammenfassen: Wir geben hinreichende, aber ausdrucksstarke syntaktische Bedingungen dafür an, dass eine gegebene Fragmentierung tatsächlich inferenzsicher in unserem Sinne der objektiven Vertraulichkeitswahrung ist, und wir schlagen ein praktisch effizientes Verfahren vor, eine inferenzsichere Fragmentierung zu finden.

Unser Beitrag zu DBSEC 11 und erweitert ausgearbeitet unser Zeitschriftenartikel in JCS [A5/205] beinhalten jeweils eine Fallstudie zum Gesichtspunkt der Effizienz: Es wurde untersucht, wie die notwendige Zustandsabhängigkeit der dynamisch vorgenommenen Inferenzkontrolle nicht nur mit Hilfe eines Logfiles vorgenommen werden kann, sondern alternativ durch schrittweises Anpassen der ursprünglichen Vertraulichkeitspolitik an das bislang beim Benutzer angefallene Wissen. Die Grundidee dazu besteht darin, jeweils alle bedeutsamen Schritte formaler Inferenzen von vertraulich zu haltenden Daten im Voraus zu berechnen. Zum Zeitpunkt einer Interaktion müssen dann nicht mehr alle solchen Beweise erzeugt werden, sondern es muss nur noch geprüft werden, ob durch zurückgelieferte Information bedeutsame Schritte derart überdeckt werden, dass ein kritischer Beweis vervollständigt werden kann.

Unser Konferenzbeitrag zu DBSEC 2012 [A5/211] sowie damit verbunden eine Diplom- und eine Masterarbeit beinhalten eine Ausprägung des Ansatzes der Politikanpassung für einen Sonderfall bezüglich relationaler Daten. In diesem Sonderfall betrachteten wir relationale Daten unter einem Relationenschema mit funktionalen Abhängigkeiten und Verbundabhängigkeiten, das als dem Benutzer bekanntes Vorwissen angenommen wurde. Zudem wurden für Anfragen und Elemente einer Vertraulichkeitspolitik nur Selektion-Projektion-Ausdrücke zugelassen. Wir entwickelten dann ein Kontrollverfahren, das ähnlich wie Intrusion Detection Systeme sogenannte Signaturen verwendet: die Signaturen werden vorweg durch eine Analyse aller möglichen Inferenzen gewonnen; zur Laufzeit werden die Signaturen dann nur noch durch die getriggerten Auswertungen einer geeigneten SQL-Anfrage überprüft und gegebenenfalls angepasst. Dieses Kontrollverfahren wurde innerhalb eines Oracle/SQL Systems implementiert und erwies sich als hocheffizient und skalierbar.

Die Antwortzeiten für als typisch angesehene Anfragen lagen auch bei großen Datenmengen unterhalb einer Sekunde.

Ein oft beschrittener Ausweg aus der im Allgemeinen unvermeidbar hohen algorithmischen Komplexität von Inferenzkontrolle liegt darin, nur degenerierte Interaktionenen in Form von Zugriffen auf vorweg inferenzsicher gemachte und dann veröffentlichte Daten zu erlauben (bekannt als privacy preserving data publishing). Die Art und der Grad der Inferenzsicherheit werden dann natürlich bestimmt durch das zugrunde gelegte Datenmodell, das betrachtete Vorwissen der Benutzer und die Art von Informationseinheiten, die man vertraulich halten kann. Diesen Ausweg habe wir exemplarisch für XML-Daten in unserem Zeitschriftenbeitrag für TDSC [A5/206] behandelt. Dabei entwickelten wir ein Verfahren, ein gegebenes XML-Dokument und die zugehörige DTD derart abzuschwächen, dass ein Benutzer mit Vorwissen in Form einer Menge von Wenn-dann-Regeln für qualifizierte Pfade (im Wesentlichen Teilbäume mit einem ausgezeichneten Pfad) keine als vertraulich spezifizierte Information, wiederum dargestellt als qualifizierte Pfade, mehr inferieren kann. Das Verfahren verwendet pragmatisch gerechtfertigte Heuristiken, um auch große XML-Dokumente noch hinreichend effizient bearbeiten zu können. Experimente mit einer prototypische Implementierung belegten die angestrebte Effizienz und Skalierbarkeit.

In Bezug auf die Betrachtung von vollständigen Agentenmodellen haben wir in unserem Konferenzbeitrag [A5/213] die vorläufigen Ergebnisse aus einem Workshopbeitrag zu relevanten Aspekten von Agentenmodellen für die subjektive Vertraulichkeitswahrung erweitert und konkretisiert. Insbesondere betrachteten wir in dem Konferenzbeitrag die etablierte BDI-Agentenarchitektur nunmehr unter Vertraulichkeitsanforderungen. In diesem Kontext erweiterten wir die Wissenskomponente von Agenten um eine explizite Repräsentation von Vertraulichkeitsinteressen und einer Sicht auf die den angreifenden Agenten zur Verfügung stehenden Informationen. Darauf aufbauend definierten wir den Begriff von (subjektiv) vertraulichkeitsbewahrenden Agenten und erarbeiteten relevante Eigenschaften von Änderungsoperationen für die erweiterte Wissenskomponente eines Agenten. Für dieses erweiterte Agentenmodell betrachteten wir Aspekte der Aktionsauswahl für die Vertraulichkeitswahrung als Teil der *Means-ends-reasoning* Operationen eines Agenten und instanziierten dieses Model beispielhaft unter Verwendung der Antwortmengenprogrammierung zur Wissensrepräsentation. In dem technischen Bericht [A5/215] präsentierten wir einen, in Bezug auf [A5/213] um die Betrachtung komplexerer *Means-ends-reasoning* Operationen erweiterten, Blick auf die Vertraulichkeitsaspekte in Agentenmodellen und stellten das in unserem Projekt entwickelte Rahmenwerk *Angerona* für Multiagentensysteme vor. In diesem Rahmenwerk können Agenten unter Verwendung verschiedener Agentenmodellen und Wissensrepräsentationsformalismen erstellt und darauf aufbauende Interaktionen in Multiagentensystemen simuliert werden. Als erweiterte Betrachtung der *Means-ends-reasoning* Operationen eines Agenten wurde in diesem Rahmen als Teil einer Diplomarbeit in dem zugrunde liegenden Agentenmodel aus [A5/213] ein Verfahren zum Planen von Kommunikation unter Bewahrung von subjektiver Vertraulichkeit entwickelt und implementiert.

Basierend auf Vorarbeiten ist das Buchkapitel [A5/214] entstanden, in welchem wir argumentationsbasierte Operationen der selektiven Revision vorstellten. Selektive Revisionsoperationen sind nicht priorisierte Wissensänderungsoperationen basierend auf der etablierten Theorie der Basisrevision, welche neue Information evaluieren und diese gegebenenfalls nur in Teilen für die nachfolgende Revision des bestehenden Wissens akzeptieren. Für diese Evaluation verwendeten wir Argumentationsformalismen und betrachteten Erweiterungen, welche die Glaubwürdigkeit bzw. Zuverlässigkeit von Agenten in diesem Prozess berücksichtigen. Wir zeigten formal, dass unsere neuen Operationen viele der wünschenswerten Eigenschaften erfüllen.

In der Arbeit [A5/212] erweiterten wir die Theorie der Basisrevision für nichtmonotone Logiken am Beispiel der Antwortmengenprogrammierung. Entsprechende Operationen werden in einem ressourcenbeschränkten, nichtmonoton schließenden Agenten benötigt. Hierzu adaptierten und erweiterten wir die etablierten Postulate für die deklarative Definition der Operationen. Des Weiteren stellten wir konstruktive Verfahren für die Realisierung der Operationen vor und zeigten in einem Repräsentationstheorem die Entsprechung zur deklarativen Beschreibung. In einer Masterarbeit wurden die neuen Operationen der Basisrevision von Antwortmengenprogrammen um eine argumentative Selektionsfunktion erweitert und somit die Ergebnisse aus den Arbeiten [A5/212] und [A5/214] verknüpft und erweitert.

Alle Operationen wurden in dem Rahmenwerk *Angerona* [A5/215] für die Benutzung in vertraulichkeitsbewahrenden Agentenmodellen implementiert.

Literatur

[198] G. Andrighetto und C. Castelfranchi. „Norm Compliance. The Prescriptive Power of Normative Actions“. In: *Paradigmi* 2 (2013), S. 120–135.

[199] M. Balliu, M. Dam und G. Le Guernic. „ENCoVer: Symbolic Exploration for Information Flow Security“. In: *25th IEEE Computer Security Foundations Symposium*. Hrsg. von S. Chong. IEEE, 2012, S. 30–44.

[200] D. A. Basin, V. Juge, F. Klaedtke und E. Zalinescu. „Enforceable Security Policies Revisited“. In: *ACM Trans. Inf. Syst. Secur.* 1 (2013), S. 3.

[201] E. Ferme und S. Hansson. „AGM 25 Years - Twenty-Five Years of Research in Belief Change.“ In: *Journal of Philosophical Logic* 40 (2011), S. 295–331.

[202] B. C. M. Fung, K. Wang, A. W.-C. Fu und P. S. Yu. *Introduction to Privacy-Preserving Data Publishing – Concepts and Techniques*. Chapman & Hall, 2011.

[203] M. Gebser, R. Kaminski, B. Kaufmann und T. Schaub. *Answer Set Solving in Practice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan und Claypool Publishers, 2012.

[204] M. J. Wooldridge. *An Introduction to MultiAgent Systems*. 2. Aufl. Wiley, 2009.

3.2.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[A5/205] J. Biskup. „Dynamic policy adaptation for inference control of queries to a propositional information system“. In: *Journal of Computer Security* 20.5 (2012), S. 509–546.

[A5/206] J. Biskup und L. Li. „On Inference-Proof View Processing of XML Documents.“ In: *IEEE Transactions on Dependable and Secure Computing* 10.2 (2013), S. 99–113.

[A5/207] J. Biskup und M. Preuß. „Database Fragmentation with Encryption: Under Which Semantic Constraints and A Priori Knowledge Can Two Keep a Secret?“ In: *Proceedings of the Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy XXVII (DBSec 2013)*. Hrsg. von L. Wang und B. Shafiq. Bd. 7964. Lecture Notes in Computer Science. Springer, 2013, S. 17–32.

[A5/208] J. Biskup und C. Tadros. „Idea: Towards a Vision of Engineering Controlled Interaction Execution for Information Services“. In: *Engineering Secure Software and Systems (ESSoS 2014)*. Hrsg. von J. Jürjens, F. Piessens und N. Bielova. Bd. 8364. Lecture Notes in Computer Science. Springer, 2014, S. 35–44.

[A5/209] J. Biskup und C. Tadros. „Preserving confidentiality while reacting on iterated queries and belief revisions“. In: *Annals of Mathematics and Artificial Intelligence* (2013), S. 1–49.

[A5/210] J. Biskup, G. Kern-Isberner, P. Krümpelmann und C. Tadros. „Reasoning on Secrecy Constraints under Uncertainty to Classify Possible Actions“. In: *Proceedings of the 8th International Symposium on Foundations of Information and Knowledge Systems (FoIKS2014)*. Hrsg. von C. Beierle und C. Meghini. Bd. 8367. Lecture Notes in Computer Science. Springer, 2014, S. 97–116.

[A5/211] J. Biskup, S. Hartmann, S. Link, J.-H. Lochner und T. Schlotmann. „Signature-Based Inference-Usability Confinement for Relational Databases under Functional and Join Dependencies“. In: *Data and Applications Security and Privacy XXVI - 26th Annual IFIP WG 11.3 Conference*. Hrsg. von N. Cuppens-Boulahia, F. Cuppens und J. García-Alfaro. Bd. 7371. Lecture Notes in Computer Science. IFIP WG 11.3. Springer, 2012, S. 56–73.

[A5/212] P. Krümpelmann und G. Kern-Isberner. „Belief Base Change Operations for Answer Set Programming“. In: *Proceedings of the 13th European Conference on Logics in Artificial Intelligence (JELIA'12)*. Bd. 7519. Lecture Notes in Artificial Intelligence. Springer, 2012, S. 294–306.

[A5/213] P. Krümpelmann und G. Kern-Isberner. „Secrecy preserving BDI Agents based on Answerset Programming“. In: *Proceedings of the 11th German Conference on Multi-Agent System Technologies (MATES'13)*. Hrsg. von M. Klusch, M. Thimm und M. Paprzycki. Bd. 8076. Lecture Notes in Computer Science. Springer, 2013, S. 124–137.

[A5/214] L. H. Tamargo, M. Thimm, P. Krümpelmann, A. J. García, M. A. Falappa, G. R. Simari und G. Kern-Isberner. „Credibility-based Selective Revision by Deductive Argumentation in Multi-agent Systems“. In: *Trends in Belief Revision and Argumentation Dynamics*. Hrsg. von E. L. Ferme, D. M. Gabbay und G. R. Simari. College Publications, 2013.

Andere Veröffentlichungen

[A5/215] P. Krümpelmann, T. Janus und G. Kern-Isberner. *Angerona - A Multiagent Framework for Logic Based Agents with Application to Secrecy Preservation.* Techn. Ber. 3. TU Dortmund, 2014.

[A5/216] C. Tadros. „Belief Change Operations under Confidentiality Requirements in Multiagent Systems“. Diss. TU Dortmund, 2014.

3.3 Ausstattung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert. Es wird mit Ablauf der zu Ende gehenden Förderperiode beendet.

3.3.1 Personal im Teilprojekt während der zu Ende gehenden Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (A5)							
Wissen-schaftliches Personal	1	Kern-Isberner, Gabriele, Prof. Dr., Prof	Information Engineering	LS 1, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	2	Biskup, Joachim, Prof. Dr., Prof	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	16,0	—	Grund-ausstattung
	3	Preuß, Marcel, Dipl.-Inf. WiMi	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	19,9	—	Grund-ausstattung
	4	Zarouali, Jaouad, SHK	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	5	Bring, Martin, SHK	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	6	Benet Gonzalez, Lorenzo, SHK	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	8,0	—	Grund-ausstattung
	7	Schalge, Dirk, SHK	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	8	Diekmann, Katharina, SHK	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	9	Rötner, Stefan, SHK	Information Engineering	LS 1, Fak. Inf, TU Do	7,0	—	Grund-ausstattung
	10	Wierzoch, Pia, SHK	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
Mit bewilligten Mitteln finanziertes Personal (A5)							
Wissen-schaftliches Personal	11	Krümpelmann, Patrick, Dipl.-Inf. WiMi	Information Engineering	LS 1, Fak. Inf, TU Do	—	Dokto-rand/in	—
	12	Tadros, Cornelia, Dr. PostDoc	Informationssysteme und Sicherheit	LS 6, Fak. Inf, TU Do	—	Dokto-rand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter (vorhandene Mittel):

1. **Kern-Isberner, Gabriele**
   Projektleitung. Schwerpunkt Multiagentensysteme plausibles Schlussfolgern, Wissensrevision.

2. **Biskup, Joachim**
   Projektleitung. Schwerpunkt Inferenzkontrolle und dynamische Adaption von Vertraulichkeitspolitiken.

3. **Preuß, Marcel**
   Fragmentierung und Veröffentlichung von relationalen Datenbanken durch Abschwächen von Information zum Vertraulichkeitsschutz.

4. **Zarouali, Jaouad**
   Optimierte Auswahl eines Zensors und Implementierung und Experimente zum Laufzeitverhalten der signaturbasierten Inferenzkontrolle .

5. **Bring, Martin**
   Implementierung und Experimente zum Laufzeitverhalten der signaturbasierten Inferenzkontrolle.

6. **Benet Gonzalez, Lorenzo**
   Implementierung eines Zensors für Änderungsoperationen mit Lügenansatz.

7. **Schalge, Dirk**
   Entwurf/Implementierung von Änderungsoperationen als Front-end des Oracle DBMS.

8. **Diekmann, Katharina**
   Entwurf/Implementierung von Änderungsoperationen als Front-end des Oracle DBMS.

9. **Rötner, Stefan**
   Implementierung/Erweiterungen von Prototypen.

10. **Wierzoch, Pia**
    Implementierung/Erweiterung im Angerona Rahmenwerk.

Aufgaben der Mitarbeiterinnen und Mitarbeiter (bewilligte Mittel):

11. **Krümpelmann, Patrick**
    Modellbildung für BDI-Agenten und Eigenschaften der Änderungs-, Deliberations- und Means-ends-reasoning Operationen sowie entsprechende Operatoren bei Verwendung der Antwortmengenprogrammierung. Simulation von vertraulichkeitsbewahrenden Agenten.

12. **Tadros, Cornelia**
    Verfahren zur Informationsflusskontrolle bei Änderungs- und Revisionsanforderungen die skeptische Inferenz des angreifenden Agenten im Rahmenwerk zur Spezifikation von Vertraulichkeitspolitiken. Simulation der Inferenz zur Kontrolle von Interaktionen unter nichtmonotonem Inferenzverhalten des verteidigenden Agenten.

3.1 Allgemeine Angaben zum Teilprojekt A6

3.1.1 Titel:

Ressourceneffiziente Analyse von Graphen

3.1.2 Fachklassifizierung:

Data Mining, Algorithm Engineering, Algorithmische Komplexitätstheorie

3.1.3 Projektleitung

Prof. Dr. Kersting, Kristian, 28.11.1973

LS 8, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 14
44227 Dortmund

Tel.: 0231-755-7876
Fax : 0231-755-5105
E-Mail: `kristian.kersting@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Mutzel, Petra, 13.06.1964

LS 11, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 14
44227 Dortmund

Tel.: 0231-755-7700
Fax : 0231-755-7740
E-Mail: `petra.mutzel@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Sohler, Christian, 19.02.1973

LS 2, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 14
44227 Dortmund

Tel.: 0231-755-6940
Fax : 0231-755-2047
E-Mail: `christian.sohler@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Verlinkte Daten und Netzwerke treten immer häufiger im Zusammenhang mit eingebetteten Systemen auf. Durch Sensoren, RFID-Chips, Kameras etc. produzieren Alltagsgegenstände kontinuierlich Daten und kommunizieren sowohl untereinander als auch mit dem Nutzer. Projekte wie Google Glass und Google Tango zeigen, dass Rechner, eingebettete Systeme und soziale Netzwerke in vielen Lebensbereichen noch bedeutender werden. Im Rahmen des Sonderforschungsbereichs treten netzwerkbasierte Daten beispielsweise in den Teilprojekten A4 (Kommunikationsnetzwerke), B4 (Verkehrsnetze und Car2X-Kommunikation) und auch mit Einschränkungen in A1 (Aufrufgraphen von Betriebssystemsfunktionen), A2 (Datenflüsse) und C1 (Interaktionsgraphen) auf. Eine natürliche Repräsentation von verlinkten Daten bilden Graphen, die eine Menge von Objekten (die Knoten) zusammen mit den zwischen diesen Objekten bestehenden Querbezügen (den Kanten) darstellen. Knoten und Kanten enthalten dabei typischerweise Attribute. Im Vergleich zur üblichen Darstellung von Objekten (den Datenpunkten) mittels Merkmal-Wert-Vektoren erfasst die Struktur dieser Graphen die komplexen, vielfältigen Beziehungen zwischen den Attributen und den Objekten. Diese zusätzliche Information ist für eine Datenanalyse typischerweise sehr hilfreich. Im Rahmen dieses Teilprojekts wollen wir daher neue Ansätze und Algorithmen zur Analyse von Graphen, dem so genannten *Graph Mining*, mit dem Ziel der Klassifikation unter besonderer Berücksichtigung von Ressourcenbeschränkungen entwickeln. Insbesondere wollen wir Klassifikationsverfahren zur Analyse von Graphen im Hinblick auf Speicherbedarf, Laufzeit, I/Os und Bandbreite (in verteilten Systemen) weiterentwickeln. Dies beinhaltet auch adaptive Verfahren, die auf dynamische Veränderungen der Graphdaten reagieren, wie sie z.B. im Zusammenhang mit sich schnell ändernden Kommunikationsnetzwerken in eingebetteten Systemen relevant sind. Dazu sollen randomisierte Verfahren aus der Algorithmentheorie mit Lernverfahren für Graphen, insbesondere Graphkernen, und Methoden aus dem Algorithm Engineering verbunden werden, um besonders ressourceneffiziente Methoden zu erhalten. Diese sollen praktisch realisiert, anhand von Benchmarkdaten evaluiert und im Rahmen des Sonderforschungsbereichs in den Teilprojekten A4 und B4 eingesetzt werden. Gemeinsam mit Teilprojekt A1 soll auch die Forschung zu graphbasierten Lernverfahren für die Energievorhersage in eingebetteten Systemen vorangetrieben werden. Mit Teilprojekt A2 soll beim Einsatz von Kernmengen für die effiziente Approximation von Graphstrukturen (kürzeste-Wege, Steinerbaumvarianten) kooperiert werden. Verteilungsbasierte Graphmerkmale sollen gemeinsam mit Teilprojekt C1 auf medizinischen Daten untersucht werden.

3.3 Ausgangssituation des Teilprojekts

Das „Internet der Dinge" trägt wesentlich zum Anstieg des Datenvolumens bei. Mehr und mehr sind Infrastrukturen, Maschinen, Fahrzeuge und Objekte des Alltags wie Verpackungen, Fernseher, Smartphones und Stromanschlüsse mit intelligenten Funktionen ausgestattet und insbesondere vernetzt. Durch Sensoren, RFID-Chips, Kameras etc. werden Alltagsgegenstände zu „Smart Artifacts", die kontinuierlich Daten produzieren und untereinander oder mit dem Nutzer kommunizieren. Verlinkte Daten und Netzwerke treten daher immer häufiger im Zusammenhang mit eingebetteten Systemen auf. Um dieser Vernetzung in der

Datenanalyse Rechnung zu tragen, kann man existierende eingebettete Systeme als Graphen darstellen und sie, je nachdem, ob sie eine Eigenschaft wie z.B. geringen Energieverbrauch erfüllen, mit „ja“ oder „nein“ beschriften. Mit diesen Informationen können dann Verfahren des Graph Minings strukturelle Besonderheiten der eingebetteten Systeme finden, die einen geringen Energieverbrauch nahelegen. Oder man denke an fahrzeugbezogene Datenkommunikation. Bei zentral auf einem Server durchgeführten Analysen fallen sofort die riesigen Kommunikationskosten auf. In den Fahrzeugen eingebettete Verfahren hingegen würden es erlauben, die aktuelle Situation im Fahrzeug und im Verkehrsnetzwerk lokal zu analysieren [272, 254]; Kommunikationskosten können dramatisch reduziert werden. Allerdings sind die beobachteten Netzwerke oft zu komplex bzw. werden mit der Zeit zu zahlreich, um sie mit herkömmlichen Verfahren eingebettet im Fahrzeug zu analysieren. Neben der reinen Menge an Graphen (Tausende oder gar Millionen von Graphen) und ihrer Hochdimensionalität kann dabei auch die exponentielle Explosion der möglichen Muster zu Skalierungsproblemen führen [234] und somit einer Anwendung in eingebetteten Systemen im Wege stehen.

In den letzten Jahren hat sich Graph Mining als ein eigenes Teilgebiet des Data Mining entwickelt. Es umfasst ein breites Spektrum an Aufgaben, von der Klassifikation über die Cluster- und Assoziationsanalysen bis hin zur Regression. Man unterscheidet zwischen dem Transactional Graph Mining, bei dem viele verschiedene Graphen gegeben sind, und dem *Single Graph Mining*, bei dem es um einen einzigen (großen) Graphen geht. Allerdings sind viele der in diesem Gebiet entwickelten Methoden aufgrund ihres Speicherbedarfs und ihrer Laufzeit insbesondere bei einer hohen Anzahl an Knoten und Kanten beziehungsweise vielen Graphen nur bedingt geeignet für einen Einsatz in Problemstellungen von eingebetteten Systemen. Daher wollen wir in diesem Projekt effiziente Methoden zur Klassifikation vor allem im Transactional Graph Mining entwickeln, die auch sehr große und/oder sehr viele Graphen ressourceneffizient verarbeiten können. Wir wollen insbesondere versuchen, mit Hilfe von Samplingverfahren, wie sie etwa im Bereich des Property Testing in der Algorithmentheorie entwickelt wurden, sehr schnell und speichereffizient die relevanten Informationen aus den Graphen zu extrahieren. Solche Verfahren (insbesondere mit beweisbaren Garantien) sind bisher im Bereich des Graph Mining bisher noch nicht ausreichend untersucht worden.

3.3.1 Stand der Forschung und eigene Vorarbeiten

Wir werden im Folgenden den Stand der Forschung im Bereich der effizienten Analyse von Graphen beschreiben und dann eigene Vorarbeiten angeben. Da das Gebiet zu groß für eine vollständige Beschreibung ist, werden wir uns auf die für die Durchführung des Projektes wichtigsten relevanten Arbeiten beschränken, insbesondere auf so genannte kernbasierte Verfahren zur Graphklassifikation. Für eine gegebene Menge von Trainingsbeispielen $(G_1, Y_1), \ldots, (G_k, Y_k)$ wollen wir also für einen Graphen G, den wir bisher noch nicht gesehen haben, die Klasse Y vorhersagen. Um dieses Lernproblem zu lösen, muss man einen Bezug von (G, Y) zur Trainingsmenge herstellen. Dazu hat sich folgende Intuition bewährt: ähnliche Graphen sollen ähnlichen Klassen zugeordnet werden. Die Ähnlichkeit in der Ausgabe (den Klassen) wird dabei typischerweise mittels einer Verlustfunktion beschrieben. Es verbleibt die Aufgabe, Ähnlichkeiten in einer Menge von Graphen zu finden. Dies geschieht durch die Definition von Ähnlichkeitsmaßen oder auch Distanzmaßen auf Graphen. Das stärkste ist sicherlich die Graphisomorphie, die eine absolute Übereinstimmung der zu vergleichenden Graphen fordert. Da bis heute kein Polynomialzeit-Algorithmus für das Graphisomorphieproblem bekannt ist und dieses Maß für die meisten Anwendungen auch zu streng ist, bildet man Graphen oft in einen hochdimensionalen Merkmalsraum ab, um Vektorrepräsentationen zu erhalten. Dies geschieht entweder explizit durch Merkmalsextraktion oder implizit durch die Verwendung von Graphkernen.

Graphkerne zur Klassifikation: Eine häufig genutzte Herangehensweise zur Klassifikation von Objekten, die durch Ausprägungen von Merkmalsvariablen $X_1, X_2, \ldots, X_n$ beschrieben werden, ist die Trennung der Klassen durch eine lineare Funktion oder — geometrisch — eine Hyperebene (in dem Falle von zwei Klassen). Natürlich sind viele Objektmengen insbesondere in niedrigen Dimensionen nicht durch eine lineare Funktion separierbar. So muss z.B. in zwei Dimensionen vielleicht eine kompliziertere, nicht-lineare Trennfunktion herangezogen werden, um zwei Klassen voneinander zu trennen. In vielen Fällen kann man komplexere Trennfunktionen darstellen, indem man den Merkmalsraum mit Hilfe einer Abbildung Φ in einen hochdimensionalen Raum abbildet und dort eine lineare Trennfunktion benutzt. Dabei kann die Berechnung der Abbildung Φ in vielen Fällen mittels einer Kernfunktion $k(x, y)$ umgangen werden, die man effizient berechnen kann. Die Existenz der Abbildung Φ kann dann implizit mit $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ nachgewiesen werden (dazu muss nach dem Satz von Mercer nur gezeigt werden, dass $k(x, y)$ positiv definit ist). Dann kann jeder Lernalgorithmus, den man über die Berechnung innerer Produkte beschreiben kann, mit Hilfe der Kernfunktion ausgeführt werden. Ein klassisches Beispiel sind z.B. Stützvektormaschinen.

Als *Graphkern* bezeichnet man nun eine Kernfunktion, die auf Graphen definiert ist. Ausgehend von David Hausslers Konvolutionskernen [249] werden dabei aktuell drei Herangehensweisen zur Konstruktion von Graphkernen verfolgt (wir verweisen auf [267] für entsprechende Referenzen): (A) Analyse von Pfaden und der stationären Verteilung von Zufallsbewegungen — also einer (zufälligen) Folge von Knoten, in der jeweils zwei aufeinander folgende Knoten durch eine Kante verbunden sind — innerhalb des Graphen; (B) Analyse der Häufigkeit von Untergraphen (oft beschränkter Größe) sowie (C) Propagieren von Knotenbeschriftungen innerhalb des Graphen. Aktuell viel Beachtung erhält der von Shervashidze *et al.* [267] eingeführte Weisfeiler-Lehman-Kern, den wir hier beispielhaft illustrieren wollen. Dieser Graphkern wird iterativ berechnet. Anfangs färbt man alle Knoten ein, z.B. mit ihrem Grad. In jeder Runde werden dann Knoten neu eingefärbt, die bisher dieselbe Farbe hatten, aber für eine Farbe c unterschiedlich viele Nachbarn der Farbe c haben. Die Iteration stoppt, wenn die so entstehende Farbpartitionierung der Knoten sich nicht mehr ändert. Aus den so in jeder Iteration entstehenden Farbhistogrammen werden Merkmale extrahiert, die zur Berechnung von Kernfunktionen zwischen Graphen herangezogen werden. Das kann auf Graphen mit (teilweise) beobachteten (diskreten und kontinuierlichen) Attributwerten erweitert werden, wie z.B. in unseren Vorarbeiten [A6/283, A6/282] gezeigt. Alternativen sind Subgraph-Matching Verfahren, wie z.B. in anderen Vorarbeiten von uns vorgeschlagen [A6/280], oder Ansätze, die ähnliche Teilpfade miteinander vergleichen [242]. Letztere stellen zur Zeit den State-of-the-Art in der Klassifikation von Graphen mit kontinuierlichen Knotenattributen dar, wie sie häufig in eingebetteten Systemen vorkommen. Allerdings skalieren sie mit $\mathcal{O}(n^2(m + \Delta^2))$, wobei n die Zahl der Knoten, m die Zahl der Kanten und Δ die Länge des längsten kürzesten Weges im Graphen bezeichnen, was für den Einsatz in eingebetteten Systemen deutlich zu langsam ist.

Graphmerkmale: Unabhängig von Graphkernen sind viele andere, oftmals sehr viel einfachere Merkmale zur statistischen Analyse von (ungewichteten) Graphen vorgeschlagen und untersucht worden, siehe z.B. [226] für einen Überblick und weitere Referenzen. Ein klassisches und einfaches Maß ist der *Knotengrad*, der die Anzahl der direkten Verbindungen (Kanten) zu anderen Knoten angibt. Viele soziale, biologische und Kommunikationsnetzwerke weisen nicht-triviale Gradverteilungen auf. Andere klassische Beispiele sind die Nähe (*Closeness Centrality*) — das Inverse der Summe aller Längen der kürzesten Wege zu den anderen Knoten — und die *Betweenness Centrality* — die Zahl der kürzesten Wege, die durch einen Knoten führen. Die Histogramme dieser und anderer Merkmale sind intensiv empirisch untersucht worden, siehe z.B. [261] und weiterführende Referenzen dort. Alle zeigen, dass Graphen komplexer Systeme weder sehr regelmäßige Strukturen aufweisen noch rein zufällig sind. Daher ist es nicht überraschend, dass analytische Untersuchungen der Verteilungen zu einem sehr heterogenen Gesamteindruck führen. So folgt die Gradverteilung in skalenfreien Netzwerken per Definitionem einem Potenzgesetz, die Verteilung von Erdös-Renyi-Zufallsgraphen — bei denen eine Kante zwischen zwei Knoten mit Wahrscheinlichkeit p existiert — allerdings im Grenzwert einer Poissonverteilung. Für andere Merkmale können sich ganz andere Verteilungen ergeben. Für das Graph Mining ist diese Heterogenität ein gravierender Nachteil, weil jedes Graphmodell (skalenfrei, Erdös-Renyi, etc.), nach dem ein vorliegender Graph generiert wurde, eine andere Verteilung für ein Merkmal benötigt. Darüber hinaus kennt man das zugrundeliegende Graphmodell meistens nicht, sodass ein einfacher verteilungsbasierter Klassifikationsansatz schwierig wenn nicht unmöglich erscheint. Zur Zeit zeichnet sich aber die *Verteilung der kürzesten Weglängen* in Graphen als ein allgemeines Maß ab. Sie spiegelt wichtige Eigenschaften der Knoten und des gesamten Graphen wider, die z.B. bei der Performanz von Kommunikationsnetzwerken eine Rolle spielen, und steht in enger Beziehung zu den schon angesprochenen Centrality-Maßen [223]. In eigenen Vorarbeiten [A6/274] konnten wir zeigen, dass die Verteilung der kürzesten Weglängen in Erdös-Renyi-Graphen Weibull verteilt ist und die Weibull-Verteilung empirisch eine gute Trennung zwischen unterschiedlichen Graphmodellen erlaubt; Graphen unterschiedlicher Graphmodelle wie z.B. skalenfreie und Erdös-Renyi-Graphen tendieren dazu, zu unterschiedlichen Parametern der Weibull-Verteilung zu führen. Hier steht man aber erst am Anfang, und das Teilprojekt soll dies weiter untersuchen.

Approximative Graphmerkmale: Die Berechnung der Länge der kürzesten Wege zwischen allen Paaren von Knoten in ungewichteten Graphen ist zeit- und ressourcenaufwändig. Selbst mit aktuellen Indexing-Ansätzen zur exakten Berechnung, siehe z.B. [217], würde die Berechnung in Graphen mit Millionen von Knoten noch Stunden oder Tage dauern (geschätzt aus den in den Papieren angegebenen Laufzeiten) und vor allem sehr speicherintensiv sein. Das schließt eine Anwendung in eingebetteten Systemen klar aus. Als attraktiver Ansatz zur Beherrschung der Berechnungskomplexität bietet sich daher approximative Graphmerkmale an. Ziel dieser Forschungsrichtung ist es, sehr schnelle Verfahren zu entwickeln und dafür etwas an Genauigkeit zu verlieren. So approximieren All-Distances-Sketches (ADS) das Histogramm der Verteilung der Längen der kürzesten Wege [233]. Der ADS eines Knotens v_i enthält eine zufällige Stichprobe von anderen Knoten v_j, wobei die Wahrscheinlichkeit einen Knoten v_j aufzunehmen mit seiner Distanz abnimmt. Für jeden aufge-

nommenen Knoten enthält der ADS die Distanz von v_j zu v_i. Für Centrality-Indizes, die auf kürzesten Wegen beruhen, liegt die Laufzeit der besten exakten Verfahren zwischen $\mathcal{O}(n^2)$ und $\mathcal{O}(n^3)$. Durch Approximation kann das auf eine fast lineare Laufzeit reduziert werden [240]. Die Berechnung des Centrality-Index ist auch eng mit sogenannten Kernmengen für das metrische 1-Median-Problem verwandt, weil die inverse Centrality eines Knotens die Summe der Abstände zu allen anderen Knoten bezüglich der kürzeste-Wege-Metrik ist. Dabei ist eine Kernmenge für das 1-Median-Problem eine kleine gewichtete Teilmenge der Punkte des zu Grunde liegenden Raums, die für jeden Anfragepunkt die Summe der Abstände zu den anderen Punkten bis auf den Faktor $(1+\epsilon)$ approximiert. Somit liefert eine Kernmenge zusammen mit den Abständen der Kernmengenpunkte zu den anderen Knoten des Graphen eine relative $(1+\epsilon)$-Approximation des Centrality-Index, d.h. für jeden Knoten wird die Summe der Abstände zu den restlichen Knoten bis auf den Faktor $(1+\epsilon)$ approximiert (im Gegensatz zu z.B. [240], die i.A. nur eine additive Approximation erhalten). Kernmengen für metrische Räume wurden z.B. in [229] entwickelt.

Generell wird die Approximation von Eigenschaften großer Graphen mit Hilfe von zufälligen Stichproben im Rahmen des Property Testings untersucht [244]. Dabei nimmt man einen Orakelzugriff auf einen Graphen an und soll mit möglichst wenig Stichproben mit konstanter Wahrscheinlichkeit unterscheiden, ob der Graph eine vorgegebene Eigenschaft (z.B. Bipartitheit) besitzt oder sich in einem ϵ-Bruchteil seiner Repräsentation von jedem Graph mit dieser Eigenschaft unterscheidet. Eine Grapheigenschaft heißt *testbar*, wenn es eine Funktion f gibt, sodass für jede Wahl von n (Zahl der Knoten) und ϵ ein Property Tester existiert, der nur $f(\epsilon)$ Anfragen an das Orakel stellt. Dicht besetzte Graphen werden durch Adjazenzmatrizen repräsentiert. In diesem Fall ist eine Grapheigenschaft genau dann testbar, wenn sie sich auf das Testen einer endlichen Menge sogenannter regulärer Partitionen reduzieren läßt [218]. Bei einer regulären Partition handelt es sich um eine Zerlegung der Knotenmenge des Graphen, sodass fast alle Paare dieser Teilmengen sogenannte ϵ-reguläre Paare sind, d.h. pseudozufällig bzgl. der Kantendichte sind. In gradbeschränkten Graphen haben wir in eigener Vorarbeit gezeigt, dass jede hyperfinite Grapheigenschaft testbar ist [A6/284]. Außerdem sind Eigenschaften wie Expansion [253] und k-Zusammenhang testbar [246]. Für Graphen mit kleinem Durchschnittsgrad aber ohne Schranke bezüglich des Knotengrads sind nur wenige Ergebnisse bekannt: Eigenschaften wie Zusammenhang [246] oder Bipartitheit für die Klasse der planaren Graphen sind testbar [A6/277]. Die zum Erreichen dieser Ergebnisse eingesetzten Stichprobenverfahren beinhalten unter anderem gleichverteilte Stichproben (in dicht besetzten Graphen), Random Walks und die Approximation der Verteilung von k-Scheiben, d.h. der durch die Knoten mit maximaler Distanz k induzierten Untergraphen (in gradbeschränkten Graphen).

Euklidische Einbettungen und Graph-Layout: Eine Möglichkeit Graphkerne zu erstellen, ist (eine Menge von) Graphen in einen euklidischen Raum der Dimensionalität d so abzubilden, dass die Abstände zwischen den Knoten bzw. Graphen möglichst exakt erhobenen Ähnlichkeiten entsprechen. Dort bildet man dann mittels der euklidischen Metrik $||\cdot||$ Kernfunktionen wie z.B. den Gauß-Kern $k(\mathbf{x},\mathbf{x}') = \exp(-\gamma||\mathbf{x}-\mathbf{x}'||)$, $\gamma > 0$. Da Graphen aber nicht vektorieller Natur sind, müssen dazu spezielle Ähnlichkeitsmaße herangezogen werden, wie z.B. Graph-Editierdistanzen, siehe [243] und Referenzen dort, die auch aus einer Menge von gegebenen Graphmatchings — Paare von Graphen, für die jedem Knoten bzw. jeder Kante in dem einen Graphen ein Knoten bzw. Kante in dem anderen Graphen zugeordnet ist — gelernt werden können [225]. Die Distanzen können auch zur Einbettung in den Euklidischen Raum und damit zur Visualisierung der Graphen herangezogen werden [266]. Soll der menschliche Betrachter schnell interessante Zusammenhänge wahrnehmen, ergeben sich aber weitere Einschränkungen. Man muss sich z.B. auf wenige Dimensionen (z.B. $d = 2, 3$) beschränken und die Visualisierung durch z.B. möglichst gleichförmige Verteilung der Knoten und möglichst kurze Kantenverläufe mit wenigen Kantenkreuzungen verständlich halten. Genau damit beschäftigt sich das Gebiet des Graphzeichnens, siehe z.B. [268, 252] für einen Überblick.

Obwohl das Graphzeichnen wie das Graph Mining die Extraktion der wichtigen Merkmale von Graphen betrachtet, gibt es bisher kaum Übertragungen der Ergebnisse. Das ist erstaunlich, weil die zusätzlichen Einschränkungen der Visualisierungen zu einer Informationsverdichtung [269] und die wiederum zu einer verbesserten Klassifikationsperformanz führen kann. Erste Anknüpfungspunkte gibt es z.B. in [A6/278], in dem ein Verfahren zur Visualisierung von Graphen als topographische Landkarte vorgestellt wird, das auf einem hierarchischen Clustering beruht, das mittels der Betweenness-Centrality erstellt wurde, und die Struktur eines Ähnlichkeitsgraphen beachtet und visualisiert. Außerdem sind viele der im Graphzeichnen entwickelten Algorithmen und Datenstrukturen (z.B. Einbettungsalgorithmen und SPQR-Bäume) z.B. in der Open Source-Softwarebibliothek *Open Graph Drawing Framework* (OGDF) [232] auch für Graphmerkmale und -kerne von Interesse. OGDF enthält u.a. eine effiziente Implementierung des auf Dijkstra basierenden kürzeste-Wege-Algorithmus, die in der Lage ist, die Closeness Centrality eines Graphen (hier: Twitter) mit bis zu 2,4 Millionen Kanten (auf einem zweifach Hexacore Server) innerhalb vertretbarer Rechenzeit (1 Stunde)

zu berechnen. Weitere erste Berührungspunkte zwischen Graph Mining und Graph Drawing sind von Rodrigues Jr. *et al.* [265] aufgewiesen worden. Daher wollen wir in diesem Projekt untersuchen, ob und inwieweit die Berücksichtigung der zusätzlichen Randbedingungen des Graphzeichnens, wie z.B. planare Einbettung, zu besserer Klassifikationsperformanz führen kann. Darüber hinaus sind die zusätzlichen Beschränkungen interessant um ressourceneffiziente Kernmethoden und Visualisierungen zu entwickeln.

Graphvergleich: Das klassische *Graphisomorphieproblem* erhält als Eingabe zwei Graphen G und H und fragt nach der Existenz einer isomorphen Abbildung, die Knoten von G auf die Knoten von H abbildet und dabei in beide Richtungen kantenerhaltend ist. Etwas allgemeiner ist das *Subgraphisomorphieproblem*, bei dem man fragt, ob ein gegebener Graph H Teilgraph eines anderen Graphen G ist. Während die Komplexität des Graphisomorphieproblems bis heute ungelöst ist, weiß man, dass das Subgraphisomorphieproblem NP-vollständig ist, dies sogar noch für zusammenhängende außenplanare Graphen. In partiellen k-Bäumen hingegen, die k-zusammenhängend sind oder beschränkten Knotengrad besitzen, kann das Problem in polynomieller Zeit gelöst werden. Sobald jedoch H (der kleinere Graph) nicht k-zusammenhängend ist oder mehr als k Knoten mit unbeschränktem Knotengrad besitzt, wird das Problem auch auf partiellen k-Bäumen wieder NP-vollständig. Praktisch können Subgraphisomorphieprobleme auf kleinen allgemeinen Graphen durch *Backtracking*-Algorithmen gelöst werden. In der Arbeit [A6/279] haben wir einen Backtracking-Algorithmus für das Subgraphisomorphieproblem entwickelt, der die speziellen Eigenschaften von Molekülgraphen ausnutzt und damit relativ effektiv arbeitet.

Beim Graphvergleich spielen oft auch die Ähnlichkeiten zwischen den wichtigen Komponenten eine Rolle. Immer wieder taucht dabei die Fragestellung auf, die zusammenhängenden Komponenten (Communities) zu finden, deren Knoten eine wichtige Rolle spielen. Formal wird das Problem z.B. in [265] als *Connection Subgraph Maximization* im Rahmen des CEPS-Konzepts (*Center-Piece Subgraph*) eingeführt. Dabei sind Gewichte für alle Knoten gegeben und das Ziel ist es, den maximalen zusammenhängenden Subgraph zu berechnen, wobei oft eine Größenbeschränkung vorgegeben ist. Formal entspricht dies genau dem *Maximum Connected Subgraph Problem*, für das in [219] ein exakter Branch-and-Cut Algorithmus angegeben wird, der auf einer neuen Formulierung des Problems als ganzzahliges lineares Programm basiert. Bezieht man die Kantengewichte mit ein, dann entspricht das Problem genau dem *Prize-Collecting Steinerbaumproblem* (PCST), das wir in [A6/281] studiert haben. Es gibt eine Vielzahl ähnlicher Probleme, die z.B. im Kompendium der Steinerbaumprobleme [248] kategorisiert sind. Darunter fallen auch verschiedene Typen von Broadcast Problemen, z.B. das *Minimum Energy Broadcast Problem*, das auch in Teilprojekt A4 auftaucht. Das Steinerbaumproblem selbst, bei dem es darum geht, eine Teilmenge der Knotenmenge, die Terminale, möglichst kostengünstig in einem Graphen zu verbinden, gehört zu den prominentesten kombinatorischen Optimierungsproblemen und ist NP-schwierig. Es sind konzeptuell einfache polynomielle 2-Approximationsverfahren bekannt sowie relativ komplexe polynomielle Verfahren, die zu 1,39 Approximationen führen [224].

Dynamische Algorithmen: Kommunikationsnetzwerke sind ständigen Änderungen unterworfen: Knoten oder Kanten fallen weg oder kommen neu hinzu, Knoten- oder Kantenattribute ändern sich. Für solche Fälle bieten sich *dynamische Graphalgorithmen* an, deren Laufzeit nicht bezüglich der Graphgröße gemessen werden, sondern bezüglich der notwendigen Änderungen. Ramalingam und Reps [263] führten dazu den Term |CHANGED| ein, der als Maß die Änderungen in Ein- und Ausgabe des dynamischen Problems darstellt plus zusätzliche Elemente, die im Wesentlichen unvermeidbar sind. Im Fall des kürzeste-Wege-Problems mit einem Startknoten entsprechen der Menge |CHANGED| genau diejenigen Knoten, die durch die Graphmodifikation neue Distanzwerte erhalten und deren Nachbarkanten. Dynamische Algorithmen, wie z.B. der in [263] für das kürzeste-Wege-Problem, mit einer Laufzeit, die polynomiell bezüglich der Menge |CHANGED| ist, heißen *beschränkt*. Während für das kürzeste-Wege-Problem sowie für Zusammenhangs- und Planaritätsfragen positive Resultate bekannt sind, gibt es für das Subgraphisomorphieproblem hingegen negative Resultate. In [241] wurde gezeigt, dass das dynamische Subgraphisomorphieproblem *unbeschränkt* ist. Ein dynamischer Algorithmus anderer Art wurde für das *All-Pairs-Shortest-Path* Problem in [238] angegeben. Dieser sogenannte *voll-dynamische* Algorithmus liefert für eine Sequenz von k Graphmodifikationen für einen Graphen $G = (V, E)$ die Laufzeit $kO(|V|^2 \log |V|^2)$. Allerdings hat der Ansatz einen sehr hohen Speicherbedarf von $O(|V||E| \log |V|)$, weil er eine große Menge an lokal kürzesten Wegen vorhält. Es wurden auch bereits dynamische Approximationsalgorithmen für kürzeste-Wege-Probleme entwickelt (siehe z.B. [250] und Referenzen dort). Dynamische Algorithmen heißen *Batch*-Algorithmen, wenn sie nicht nur eine, sondern mehrere Graphmodifikationen in einem Durchlauf verarbeiten können. Die Arbeit [221] zeigt, dass sich Modifikationen an den wichtigsten dynamischen Algorithmen für kürzeste-Wege-Probleme mit einem Startpunkt lohnen, denn experimentell konnte dort eine 15-fache Verbesserung der Rechenzeit nachgewiesen werden.

Parallelisiertes und verteiltes Graph Mining: Aufgrund immer größerer Datenmengen und der Tatsache, dass Daten oft nur noch verteilt durch eingebettete Systeme geliefert werden, und den damit einhergehenden Ressourcenbeschränkungen Platz, Zeit bzw. Bandbreite versucht man natürlich auch die Verfahren im Graph Mining zu parallelisieren bzw. auf verteilten Systemen zu realisieren. Für die klassischen Verfahren des maschinellen Lernens, die nicht Graphen analysieren, gibt es dafür eine Vielzahl an Ansätzen, die Multicore-Rechner, GPUs und Rechner-Cluster im Allgemeinen ausnutzen, siehe z.B. [273, 257] und Referenzen dort. Im Wesentlichen beruhen alle diese Verfahren auf Onlinealgorithmen, die nacheinander immer nur einen kleinen Teil der vorliegenden Daten verarbeiten. Diese Ansätze sind vielversprechend, wurden allerdings bisher fast nicht zur Klassifikation von Graphen und vernetzten Daten untersucht und verfolgt. (Mengen von) Graphen stellen eine besondere Schwierigkeit dar, weil die Datenpunkte selbst nicht unabhängig sind; sie sind vernetzt. In unseren eigenen Vorarbeiten konnten wir empirisch zeigen, dass sich solche Onlineverfahren zum Lernen von Wahrscheinlichkeitsmodellen für komplexvernetzte Daten heranziehen lassen, wenn man die Graphen in viele lokale Graphen aufbricht. Für exakte Berechnungen geht das allerdings nicht. Das liegt daran, dass Ansätze für verteilte Graphklassifikation nicht datenparallel, sondern graphparallel sein müssen. Erste Schritte zur Graphparallelisierung sind Pregel, Dryad und GraphLab. Für Graphen mit Millionen von Knoten und Kanten zeigen sie Vorteile gegenüber einer rein datenparallelen Analyse von Graphen mittels z.B. MapReduce und GPUs. Insbesondere GraphLab [259] erlaubt es, verteilte Graphalgorithmen nicht nur mittels MapReduce und GPUs sondern auch kommunikationseffizienter mittels Nachrichtenübertragung (Message Passing) zu realisieren: Programme beschreiben Runden, in denen jeder Knoten Nachrichten aus der vorherigen Runde erhalten und neue Nachrichten senden kann. Auch kann jeder Knoten seine lokale Topologie in jeder Runde verändern. Dies eignet sich besonders gut für Graphkerne mittels Propagierung (siehe oben). Für unsere Arbeiten in AP3 nehmen wir u.a. das *Message Passing* Modell an, das unter dem Namen CONGEST [262] bekannt ist. Die n Knoten haben zu Beginn keine Informationen über das Netzwerk, sie kennen jedoch ihre Nachbarknoten. Dabei wird jeder Knoten durch $O(\log n)$ Bits repräsentiert. Das synchrone Kommunikationsmodell nimmt an, dass jeder Knoten B Bits über seine inzidenten Kanten in einer Synchronisationsrunde senden kann. Typischerweise wird $B = O(\log n)$ gesetzt, was bedeutet, dass jeder Knoten eine konstante Anzahl an Knoten bzw. Kanten IDs per Nachricht übermitteln kann. In diesem Modell wird nur die Anzahl der notwendigen Synchronisationsrunden analysiert, denn es wird davon ausgegangen, dass diese die lokalen Berechnungen dominieren.

Eine weitere Herangehensweise sind plattenbasierte, also I/O-effiziente Ansätze auf einem einzelnen Rechner [256], die direkt in Systeme eingebettet werden können. Externspeicheralgorithmen zielen genau darauf ab, die Anzahl der I/Os (d.h. der Lese- und Schreibzugriffe) zwischen dem schnellen internen Speicher und dem langsamen externen Speicher zu reduzieren. Sie wurden ursprünglich entwickelt, um den Speicherhierarchien in den modernen Rechnerarchitekturen gerecht zu werden. Zur Analyse wird z.B. das Modell von Vitter und Shriver (1994) (*External-Memory-Modell*) verwendet, bei dem bei jedem I/O-Zugriff jeweils ein Block von B Daten geliefert wird. Externspeicheralgorithmen versuchen nun zeitliche und örtliche Lokalität bei den Daten und den Zugriffen zu erhalten, um so möglichst wenige I/O Zugriffe durchführen zu müssen. Für einen Überblick zu Externspeicheralgorithmen verweisen wir auf [270]. Graphenalgorithmen werden bereits seit längerem erfolgreich im External-Memory-Modell entwickelt, aber es gibt noch immer grundsätzliche ungelöste Fragen (z.B. externe Tiefensuche).

Eigene Vorarbeiten: Für die Durchführung des Projekts ist die Expertise aller drei beteiligten Arbeitsgruppen (AG) notwendig. Kristian Kersting ist im Forschungsgebiet effiziente Wissensentdeckung in komplexen und großen Datenmengen ausgewiesen, Petra Mutzel in Graphenalgorithmen und Algorithm Engineering, und Christian Sohler in der Algorithmentheorie mit Schwerpunkt auf großen Datenmengen.

Die AG Kersting bringt in das geplante Projekt Wissen zum Lernen mit Graphen und zum statistisch-relationalen Lernen ein. Konkret werden die bereits angesprochenen Ergebnisse zur analytischen Darstellung von Verteilungen von kürzesten Wegen in Graphen [A6/274], zur Berechnung von Graphkernen mittels Informationspropagierung auf Graphen [A6/283, A6/282] und zur Charakterisierung des Weisfeiler-Lehman-Algorithmus als nicht-lineare, kontinuierliche Optimierung [A6/26] eingebracht. Die AG Kersting hat auch Beiträge zum verteilten Schätzen der Parameter von relationalen Wahrscheinlichkeitsmodellen geleistet.

Die AG Mutzel bringt in das geplante Projekt Graphalgorithmen und kombinatorische Optimierungsalgorithmen ein. Konkret werden die bereits angesprochenen Ergebnisse zur effizienten Substruktursuche in großen Molekülgraphdatenbanken [A6/279], zu Graphkernen über Graphen mit kontinuierlichen Attributen mittels Teilgraph Matching [A6/280] sowie zur Visualisierung von Graphen [A6/278, 252] eingebracht. Auch ihre langjährigen Erfahrungen auf dem Gebiet der kombinatorischen Optimierung NP-vollständiger Probleme mittels ganzzahligen linearen Programmen sind für die Teilprojekte des SFB nützlich. So wurden neue gerichtete Formulierungen als ganzzahlige lineare Programme angegeben, die bisherige Formulierun-

gen theoretisch und praktisch dominieren und zu neuen exakten Branch-and-Cut Algorithmen für Probleme im Umfeld des Steinerbaumproblems führen [A6/281, 231, 219]. Die AG Mutzel hat Erfahrungen in der Entwicklung und Implementierung von Algorithmen auf parallelen Architekturen [239, 260].

Die AG Sohler bringt in das geplante Projekt Wissen zur Algorithmentheorie und des Property Testings ein. Konkret werden insbesondere Ergebnisse zur Approximation von Eigenschaften eines Graphen mit Hilfe von Zufallsprozessen eingebracht. Wie bereits angesprochen, wurde in [A6/275] zum ersten Mal gezeigt, dass eine große Klasse von Grapheigenschaften unter anderem für planare Graphen testbar ist. Aufbauend auf dieser Arbeit und anderen Nachfolgearbeiten [222, 247] wurde in [A6/284] gezeigt, dass jede hyperfinite Grapheigenschaft in konstanter Zeit approximativ entschieden werden kann. In [A6/276] konnten wir zeigen, dass man α-Expandergraphen in $\tilde{O}(\sqrt{n})$ Zeit von Graphen unterscheiden kann, die ϵ-entfernt von jedem α^*-Expander sind, wobei $\alpha^* = \Theta(D^2\alpha^2/log(n/\epsilon))$ und wir in der Laufzeit ϵ, D und α als konstant annehmen. In einer noch nicht veröffentlichten Arbeit konnten wir den Tester für Expandergraphen auf k-Clusterings erweitern. Dabei ist eine k-Clusterung ein Graph der in maximal k gute Expander zerlegt werden kann, die jeweils durch kleine Schnitte verbunden sind. Darüber hinaus soll Wissen zum Thema Kernmengen und metrische Einbettungen in das Projekt einfließen.

Literatur

[217] T. Akiba, Y. Iwata und Y. Yoshida. „Fast exact shortest-path distance queries on large networks by pruned landmark labeling“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 2013, S. 349–360.

[218] N. Alon, E. Fischer, I. Newman und A. Shapira. „A Combinatorial Characterization of the Testable Graph Properties: It's All About Regularity“. In: *SIAM J. Comput.* 39.1 (2009), S. 143–167.

[219] E. Alvarez-Miranda, I. Ljubic und P. Mutzel. „The Maximum Weight Connected Subgraph Problem“. In: *Facets of Combinatorial Optimization*. Hrsg. von M. Jünger und G. Reinelt. Heidelberg: Springer, 2013, S. 245–270.

[220] S. Arora und S. Kale. „A combinatorial, primal-dual approach to semidefinite programs“. In: *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*. 2007, S. 227–236.

[221] R. Bauer und D. Wagner. „Batch Dynamic Single-Source Shortest-Path Algorithms: An Experimental Study“. In: *Experimental Algorithms*. Hrsg. von J. Vahrenhold. Bd. 5526. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, S. 51–62.

[222] I. Benjamini, O. Schramm und A. Shapira. „Every Minor-Closed Property of Sparse Graphs is Testable“. In: *Advances in Mathematics* 223 (2010), S. 2200–2218.

[223] P. Boldi, M. Rosa und S. Vigna. „HyperANF: approximating the neighbourhood function of very large graphs on a budget“. In: *Proceedings of the 20th International Conference on World Wide Web (WWW)*. 2011, S. 625–634.

[224] J. Byrka, F. Grandoni, T. Rothvoß und L. Sanità. „Steiner Tree Approximation via Iterative Randomized Rounding“. In: *J. ACM* 60.1 (2013), 6:1–6:33.

[225] T. Caetano, J. McAuley, L. Cheng, Q. Le und A. Smola. „Learning Graph Matching“. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 31.6 (2009), S. 1048–1058.

[226] D. Chakrabarti und C. Faloutsos. *Graph Mining: Laws, Tools, and Case Studies*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2012.

[227] P. Chalermsook und J. Fakcharoenphol. „Simple Distributed Algorithms for Approximating Minimum Steiner Trees“. In: *Proceedings of the 11th Annual International Conference on Computing and Combinatorics*. COCOON'05. Kunming, China: Springer-Verlag, 2005, S. 380–389.

[228] B. Chazelle, R. Rubinfeld und L. Trevisan. „Approximating the Minimum Spanning Tree Weight in Sublinear Time“. In: *SIAM J. Comput.* 34.6 (2005), S. 1370–1379.

[229] K. Chen. „On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications“. In: *SIAM J. Comput.* 39.3 (2009), S. 923–947.

[230] M. Chimani, P. Hungerländer, M. Jünger und P. Mutzel. „An SDP approach to multi-level crossing minimization“. In: *ACM Journal of Experimental Algorithmics* 17.1 (2011).

[231] M. Chimani, M. Kandyba, I. Ljubić und P. Mutzel. „Obtaining Optimal k-cardinality Trees Fast“. In: *Journal of Experimental Algorithmics* 14 (2010), 5:2.5–5:2.23.

[232] M. Chimani, C. Gutwenger, M. Jünger, G. W. Klau, K. Klein und P. Mutzel. „The Open Graph Drawing Framework (OGDF)“. In: *Handbook of Graph Drawing and Visualization*. Hrsg. von R. Tamassia. CRC Press, 2013. Kap. 17, S. 543–569.

[233] E. Cohen. „All-Distances Sketches, Revisited: HIP Estimators for Massive Graphs Analysis“. In: *Proceedings of the Symposium on Principles of Database Systems (PODS)*. 2014.

[234] R. Cordeiro, C. Faloutsos und C. Traina Jr. *Data Mining in Large Sets of Complex Data*. Springer Briefs in Computer Science. Springer, 2013.

[235] A. Czumaj und C. Sohler. „Estimating the Weight of Metric Minimum Spanning Trees in Sublinear Time“. In: *SIAM J. Comput.* 39.3 (2009), S. 904–922.

[236] G. Damiand, C. Higuera, J.-C. Janodet, A. Samuel und C. Solnon. „A Polynomial Algorithm for Submap Isomorphism“. In: *Graph-Based Representations in Pattern Recognition*. Bd. 5534. LNCS. Springer Berlin Heidelberg, 2009, S. 102–112.

[237] T. De Bie und N. Cristianini. „Fast SDP Relaxations of Graph Cut Clustering, Transduction, and Other Combinatorial Problem“. In: *Journal of Machine Learning Research* 7 (2006), S. 1409–1436.

[238] C. Demetrescu und G. Italiano. „Fully dynamic all pairs shortest paths with real edge weights“. In: *J. Comput. Syst. Sci.* 72.5 (2006), S. 813–837.

[239] D. van Dyk, M. Geveler, S. Mallach, D. Ribbrock, D. Göddeke und C. Gutwenger. „HONEI: A collection of libraries for numerical computations targeting multiple processor architectures“. In: *Computer Physics Communications* 180.12 (2009), S. 2534–2543.

[240] D. Eppstein und J. Wang. „Fast Approximation of Centrality“. In: *J. Graph Algorithms Appl.* 8 (2004), S. 39–45.

[241] W. Fan, X. Wang und Y. Wu. „Incremental graph pattern matching“. In: *ACM Trans. Database Syst.* 38.3 (2013), S. 18.

[242] A. Feragen, N. Kasenburg, J. Petersen, M. de Bruijne und K. Borgwardt. „Scalable kernels for graphs with continuous attributes“. In: *27th Annual Conference on Neural Information Processing Systems (NIPS)*. 2013, S. 216–224.

[243] X. Gao, B. Xiao, D. Tao und X. Li. „A survey of graph edit distance“. In: *Pattern Analysis and Applications* 13.1 (2010), S. 113–129.

[244] O. Goldreich, S. Goldwasser und D. Ron. „Property Testing and its Connection to Learning and Approximation“. In: *J. ACM* 45.4 (1998), S. 653–750.

[245] O. Goldreich und D. Ron. „On Testing Expansion in Bounded-Degree Graphs“. In: *Electronic Colloquium on Computational Complexity (ECCC)* 7.20 (2000).

[246] O. Goldreich und D. Ron. „Property Testing in Bounded Degree Graphs“. In: *Algorithmica* 32.2 (2002), S. 302–343.

[247] A. Hassidim, J. A. Kelner, H. N. Nguyen und K. Onak. „Local Graph Partitions for Approximation and Testing“. In: *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*. 2009, S. 22–31.

[248] M. Hauptmann und M. Karpinski. *A Compendium on Steiner Tree Problems*. Dez. 2013. URL: http://theory.cs.uni-bonn.de/info5/steinerkompendium/.

[249] D. Haussler. *Convolution Kernels on Discrete Structures*. Techn. Ber. ICSC-CRL-99-10. University of California at Santa Cruz, Department of Computer Science, 1999.

[250] M. Henzinger, S. Krinninger und D. Nanongkai. „A Subquadratic-Time Algorithm for Decremental Single-Source Shortest Paths“. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2014, S. 1053–1072.

[251] S. Holzer und R. Wattenhofer. „Optimal distributed all pairs shortest paths and applications“. In: *ACM Symposium on Principles of Distributed Computing (PODC)*. ACM, 2012, S. 355–364.

[252] M. Jünger und P. Mutzel. *Graph Drawing Software*. Springer, 2004.

[253] S. Kale und C. Seshadhri. „An Expansion Tester for Bounded Degree Graphs“. In: *SIAM J. Comput.* 40.3 (2011), S. 709–720.

[254] H. Kargupta, K. Sarkar und M. Gilligan. „MineFleet®: an overview of a widely adopted distributed vehicle performance data mining system“. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD)*. 2010, S. 37–46.

[255] B. Kulis, A. Surendran und J. Platt. „Fast Low-Rank Semidefinite Programming for Embedding and Clustering“. In: *Proc. Eleventh Internat. Conf. Artificial Intelligence and Statistics (AISTATS)*. 2007, S. 235–242.

[256] A. Kyrola, G. Blelloch und C. Guestrin. „GraphChi: Large-Scale Graph Computation on Just a PC“. In: *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 2012.

[257] Q. Le, T. Sarlós und A. Smola. „Fastfood - Computing Hilbert Space Expansions in loglinear time“. In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Bd. 28. JMLR Proceedings. 2013, S. 244–252.

[258] J. Leskovec, K. Lang, A. Dasgupta und M. W. Mahoney. „Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters“. In: *Internet Mathematics* 6.1 (2009), S. 29–123.

[259] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin und J. Hellerstein. „Distributed GraphLab: A Framework for Machine Learning in the Cloud“. In: *PVLDB* 5.8 (2012), S. 716–727.

[260] S. Mallach und C. Gutwenger. „Improved Scalability by Using Hardware-Aware Thread Affinities“. In: *Facing the Multicore-Challenge*. Hrsg. von R. Keller, D. Kramer und J.-P. Weiss. Bd. 6310. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, S. 29–41.

[261] M. McGlohon, L. Akoglu und C. Faloutsos. „Statistical Properties of Social Networks“. In: *Social Network Data Analytics*. 2011, S. 17–42.

[262] D. Peleg. *Distributed Computing: A Locality-sensitive Approach*. Philadelphia, PA, USA: Society for Industrial und Applied Mathematics, 2000.

[263] G. Ramalingam und T. Reps. „An Incremental Algorithm for a Generalization of the Shortest-path Problem“. In: *J. Algorithms* 21.2 (Sep. 1996), S. 267–305.

[264] M. Ramana, E. Scheinerman und D. Ullman. „Fractional isomorphism of graphs“. In: *Discrete Mathematics* 132 (1994), S. 247–265.

[265] J. Rodrigues Jr., H. Tong, J.-Y. Pan, A. Traina, C. Traina Jr. und C. Faloutsos. „Large Graph Analysis in the GMine System“. In: *IEEE Trans. Knowl. Data Eng.* 25.1 (2013), S. 106–118.

[266] B. Shaw und T. Jebara. „Structure preserving embedding“. In: *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*. 2009, 118 ff.

[267] N. Shervashidze, P. Schweitzer, E. van Leeuwen, K. Mehlhorn und K. Borgwardt. „Weisfeiler–Lehman Graph Kernels“. In: *Journal of Machine Learning Research* 12 (2011), S. 2539–2561.

[268] R. Tamassia. *Handbook of Graph Drawing and Visualization*. Discrete Mathematics and Its Applications. Boca Raton: Chapman und Hall/CRC, 2013.

[269] N. Tishby, F. Pereira und W. Bialek. „The Information Bottleneck method“. In: *The 37th annual Allerton Conference on Communication, Control, and Computing)*. 1999, S. 368–377.

[270] J. S. Vitter. *Algorithms and Data Structures for External Memory*. Hanover, MA, USA: Now Publishers Inc., 2008.

[271] M. Wainwright, T. Jaakkola und A. Willsky. „MAP estimation via agreement on trees: message-passing and linear programming“. In: *IEEE Transactions on Information Theory* 51.11 (2005), S. 3697–3717.

[272] H. Zhang und Z. Li. „Weighted ego network for forming hierarchical structure of road networks“. In: *International Journal of Geographical Information Science* 25.2 (2011), S. 255–272.

[273] M. Zinkevich, M. Weimer, A. J. Smola und L. Li. „Parallelized Stochastic Gradient Descent“. In: *24th Annual Conference on Neural Information Processing Systems (NIPS)*. 2010, S. 2595–2603.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[A6/274] C. Bauckhage, K. Kersting und B. Rastegarpanah. „The Weibull as a Model of Shortest Path Distributions in Random Networks“. In: *Working Notes of the International Workshop on Mining and Learning with Graphs (MLG 2013) at KDD 2013*. Chicago, IL, USA, Aug. 2013.

[A6/275] A. Czumaj, A. Shapira und C. Sohler. „Testing Hereditary Properties of Nonexpanding Bounded-Degree Graphs“. In: *SIAM J. Comput.* 38.6 (2009), S. 2499–2510.

[A6/276] A. Czumaj und C. Sohler. „Testing Expansion in Bounded-Degree Graphs“. In: *Combinatorics, Probability & Computing* 19.5-6 (2010), S. 693–709.

[A6/277] A. Czumaj, M. Monemizadeh, K. Onak und C. Sohler. „Planar Graphs: Random Walks and Bipartiteness Testing“. In: *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*. 2011, S. 423–432.

[A6/278] M. Gronemann, M. Jünger, P. Mutzel und N. Kriege. „MolMap – Visualizing Molecule Libraries as Topographic Maps“. In: *Proc. Int. Conf. Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications (GRAPP & IVAPP)*. 2013.

[A6/26] K. Kersting, M. Mladenov, R. Garnett und M. Grohe. „Power Iterated Color Refinement“. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*. Hrsg. von C. Brodley und P. Stone. AAAI Press, 2014.

[A6/279] K. Klein, N. Kriege und P. Mutzel. „CT-index: Fingerprint-based Graph Indexing Combining Cycles and Trees“. In: *IEEE 27th Int. Conf. on Data Engineering (ICDE)*. Apr. 2011, S. 1115 –1126.

[A6/280] N. Kriege und P. Mutzel. „Subgraph Matching Kernels for Attributed Graphs“. In: *Proc. of the 29th Int. Conf. on Machine Learning (ICML)*. Omnipress, 2012.

[A6/281] I. Ljubić, R. Weiskircher, U. Pferschy, G. Klau, P. Mutzel und M. Fischetti. „An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem“. In: *Mathematical Programming* Series B 105 (2006), S. 427–449.

[A6/282] M. Neumann, R. Garnett und K. Kersting. „Coinciding Walk Kernels: Parallel Absorbing Random Walks for Learning with Graphs and Few Labels“. In: *Asian Conference on Machine Learning (ACML)*. Bd. 29. JMLR Proceedings. 2013, S. 357–372.

[A6/283] M. Neumann, N. Patricia, R. Garnett und K. Kersting. „Efficient Graph Kernels by Randomization“. In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Part 1*. 2012, S. 378–393.

[A6/284] I. Newman und C. Sohler. „Every Property of Hyperfinite Graphs Is Testable“. In: *SIAM J. Comput.* 42.3 (2013), S. 1095–1112.

3.4 Planung des Teilprojekts

Ziele

Im Rahmen dieses Teilprojekts sollen neue Ansätze und Algorithmen zur Analyse von Graphen (Graph Mining) mit dem Ziel der Klassifikation unter besonderer Berücksichtigung von Ressourcenbeschränkungen entwickelt werden. Die von uns betrachteten Ressourcenbeschränkungen sind Laufzeit, I/O-Verhalten, benötigter Speicherplatz und Bandbreite (bei verteilten Anwendungen). Dazu wollen wir neue randomisierte Ansätze aus der Algorithmentheorie mit aktuellen Graphkernverfahren aus dem maschinellen Lernen sowie mit Methoden des Algorithm Engineering zur Spezialisierung von Graphalgorithmen verbinden, um ressourceneffiziente Algorithmen mit theoretischer Fundierung zu erhalten.

Um unser Ziel zu erreichen, wollen wir zwei Hauptherangehensweisen verfolgen:

- Die *effiziente Merkmalsextraktion* aus Graphen mit Hilfe von randomisierten, approximativen Verfahren. Diese Merkmale können dann mit klassischen Lernverfahren für vektorielle Daten verarbeitet werden.
- Die Entwicklung und Verbesserung von *Graphkernen*. Auch hier wollen wir Approximation und Randomisierung zur Reduktion des Ressourcenbedarfs einsetzen.

Beide Verfahren hängen natürlich eng zusammen. Durch eine Kombination mit bestehenden kernbasierten Verfahren (SVMs, etc.) kann man aus einer Merkmalsextraktion einen Graphkern machen und umgekehrt definiert jeder Graphkern über die implizite Abbildung Φ in einen inneren Produktraum auch Merkmalsvektoren. Wir betrachten daher Verfahren, die Merkmale explizit (und Kerne implizit) definieren unter Merkmalsextraktion, und Verfahren, die Kerne explizit und die Merkmale implizit definieren, unter Graphkernen.

In beiden Bereichen ist unser Ziel, hochperformante Verfahren für die Graphklassifikation zu entwickeln, die besonders ressourcenschonend sind und auch bei Bedarf einfach in eingebetteten verteilten oder parallelen

Systemen umgesetzt werden können, sowie idealerweise auch schnell auf dynamische Veränderungen im Netzwerk reagieren können. Um dieses Ziel zu erreichen, wollen wir insbesondere Methoden entwickeln, die mit Hilfe von randomisierten Verfahren wie zufälligen Stichproben schon aus theoretischer Sicht sehr speicher- und laufzeiteffizient implementiert werden können, und die bei Bedarf dann auch gut auf verteilte oder parallele Architekturen portiert werden können. Dies können z.B. Verfahren sein, die auf Kollisionsstatistiken von Random Walks, Verteilungen von Untergraphen, oder Verteilungen von klassischen Graphmerkmalen, wie den Längen der kürzesten Wege, basieren. Diese haben sich teilweise bereits in der Vergangenheit als für die Klassifikation gut geeignet erwiesen und sind prinzipiell gut für den Einsatz von Stichprobenverfahren geeignet. Unser Ansatz unterscheidet sich von bestehenden Verfahren insbesondere auch dadurch, dass wir bei der Auswertung der Merkmale bzw. des Graphkerns Zufall einsetzen anstelle der deterministisch zu bestimmenden Verteilung des Zufallsprozesses. Dies hat den großen Vorteil, dass durch den Einsatz der Stichprobenverfahren die algorithmische Komplexität des Problems erheblich reduziert werden kann und damit das Potenzial für große Performanzgewinne gegeben ist.

Arbeitsprogramm

Um unsere Herangehensweise erfolgreich umzusetzen, gliedern wir das Projekt in mehrere Teilziele, die jeweils einem Arbeitspaket entsprechen: Ressourcenbeschränkte Merkmalsextraktion aus Graphen, effiziente Graphkerne, verteilte Graphen als Eingabe, dynamisches und adaptives Graph Mining sowie Algorithm Engineering. Im Folgenden beschreiben wir diese Arbeitspakete und Teilziele.

Arbeitspaket 1. Ressourcenbeschränkte Merkmalsextraktion aus Graphen: Hier ist unser Ziel die Entwicklung ausdrucksstarker, zufallsbasierter und ressourceneffizienter Merkmalsextraktionsverfahren. Dabei soll die Ressourceneffizienz (Zeit, Speicher, Anfragen) und Ausdrucksstärke (Klassifikationsgüte) der Verfahren theoretisch und/oder empirisch nachgewiesen werden. Dazu wollen wir neue Methoden zur Merkmalsextraktion aus Graphen entwickeln, die zum einen auf *verteilungsbasierten Merkmalen* und zum anderen auf *randomisierten Prozessen* basieren. Mittels der extrahierten Merkmale können dann Graphkerne konstruiert werden.

Verteilungsbasierte Graphmerkmale. Wir halten insbesondere Merkmale, die auf Verteilungen klassischer Grapheigenschaften basieren, für ausdrucksstark, d.h. sie erlauben eine gute Klassifikation. Dies sind z.B. die Verteilung der kürzesten Weglängen zwischen allen Knotenpaaren, der Closeness und Betweenness Centrality-Index eines Graphen oder die Verteilung der Kantenlängen in minimalen Spannbäumen oder Steinerbäumen. Mittels Skalarprodukten zwischen den Verteilungen o.ä. können Kernfunktionen zur Klassifikation gebildet werden. Erste praktische und theoretische Ergebnisse auf Zufallsgraphen sind vielversprechend, um z.B. das Graphmodell, das den Graphen erzeugt hat, zu bestimmen. Es konnte gezeigt werden, dass die Längen der kürzesten Wege in Erdös-Rény Zufallsgraphen Weibull verteilt sind (siehe auch eigene Vorarbeiten). Eine Einbettung von mehreren tausend Zufallsgraphen verschiedener Modelle (Erdös-Rény, skalenfrei, etc.) in den 2-D Raum der Skalen- und Formparameter der Weibull-Verteilung zeigt ein erstaunliches Ergebnis: Graphen eines Graphmodelles gruppieren sich. Wir können also eine Menge von Graphen mittels der Längen der kürzesten Wege in verschiedene Graphmodelle klassifizieren. Diesen Ansatz wollen wir weiterverfolgen und untersuchen, ob es eine analytische Darstellung von Verteilungen von kürzesten Wegen in anderen Graphen als Erdös-Rény Graphen gibt, was bisher unklar ist.

Die Bedeutung des *Centrality-Index* bei der Analyse sozialer Netzwerke sowie die der *minimalen Spannbäume* im Clustering (der minimale Spannbaum beschreibt agglomeratives Clustering mit Complete Linkage Maß) und ähnlicher Interpretationsmöglichkeiten von *Steinerbäumen* legen ebenfalls nahe, dass diese bzw. ihre zugehörigen Verteilungen aussagekräftige Merkmale für die Klassifikation darstellen. Daher wollen wir randomisierte und/oder approximative Verfahren entwickeln, die Verteilungen von Kantenlängen in den gerade genannten klassischen Graphstrukturen (minimale Spannbäume, Steinerbäume) approximieren. Idealerweise muss dazu ein einfacheres Graphproblem gelöst werden als bei einer exakten Berechnung oder es genügt sogar, nur zufällige Teilgraphen zu betrachten. So erscheint es plausibel, dass man die Verteilung der Längen der kürzesten Wege in einem Graph bereits durch das Lösen weniger kürzeste-Wege-Probleme für eine kleine, zufällige Menge von Startknoten approximieren kann. Dadurch würde man die Komplexität von einem All-Pairs-Shortest-Paths Problem auf wenige Single-Source-Shortest-Path Probleme reduzieren. Zusammen mit spezialisierten Kürzesten-Wege-Algorithmen aus AP5 werden daher Laufzeit und Speicherbedarf erheblich reduziert, sodass sich die Verfahren besser unter starken Ressourcenbeschränkungen verwenden lassen als ihre exakten Gegenstücke. Um die Verteilung der Kantenlängen eines minimalen Spannbaums zu approximieren, wollen wir auf Verfahren aufbauen, die die Kosten eines minimalen Spannbaums mit Hilfe von Zufallsprozessen approximieren [228, 235] und die deutlich effizienter sind als die Berechnung des

Spannbaums. Hier stellt sich die Frage, wie man diese Prozesse so verändern bzw. weiterentwickeln kann, dass man nicht nur die Kosten sondern sogar die Verteilung der Kantengewichte approximiert. Eine ähnliche Herangehensweise kann wahrscheinlich auch für die Approximation der Kantenverteilung in Steinerbäumen – eine Verallgemeinerung des Spannbaums, bei dem nur eine Teilmenge der Knoten verbunden werden muss – erfolgreich sein. Da die Struktur von Steinerbäumen ungleich komplexer ist als die von minimalen Spannbäumen, ist dies (unserer Meinung nach) vielleicht die größte Herausforderung in diesem Arbeitspaket. Wir wollen außerdem untersuchen, ob sich Kernmengenmethoden auf Varianten des Steinerbaumproblems anwenden lassen, um so eine höhere Effizienz zu erreichen. Wenn man die Definition der in Teilprojekt A2 betrachteten Kernmengen von euklidischen auf metrische Räume erweitert, so kann man Kernmengen auch für die Approximation des Centrality-Index eines Graphen verwenden. Dazu wendet man Kernmengenkonstruktionen für das 1-Median-Problem für metrische Räume wie z.B. [229] auf die kürzeste-Wege-Metrik eines Graphen an, und kann dann mit Hilfe weniger kürzeste-Wege Berechnungen den Centrality-Index mit einem relativen Fehler von $(1 + \epsilon)$ approximieren. Wir wollen diesen Ansatz verfolgen (hier gilt in erster Linie zu klären, wie man die Kernmenge effizient berechnen kann) und mit spezialisierten Algorithmen für das kürzeste-Wege-Problem aus AP5 kombinieren, um den Centrality-Index effizient zu berechnen.

Zufallsprozessbasierte Merkmale. Eine weitere interessante Möglichkeit ist die Berechnung von Merkmalen, die auf Zufallsprozessen basieren. Beispielsweise kann eine Kollisionsstatistik von Endpunkten von Random-Walks von einer Menge von Startknoten, die ebenfalls zufällig gewählt wurden (von jedem dieser Knoten werden aber mehrere Walks gestartet), interessante Informationen über die globale Graphstruktur liefern. Wir wollen die bereits im Vorarbeitenteil angesprochenen Zusammenhänge zum Gebiet des Property Testings ausnutzen, um sehr effiziente (geringer Zeitaufwand) und aussagekräftige Merkmale (keine wesentlichen Einbußen in der Klassifikationgüte) zu erhalten. Dabei ist die generelle Schwierigkeit, dass die meisten Verfahren aus dem Property Testing zwar theoretisch effizient, aus praktischer Sicht die involvierten Konstanten aber typischerweise viel zu groß sind. Ein gutes Beispiel ist hier die Theorie zur Testbarkeit in gradbeschränkten Graphen. Hier ist bekannt, dass die Approximation der Verteilung von sogenannten k-Scheiben (dabei ist eine k-Scheibe um einen Knoten v der durch alle Knoten mit Distanz maximal k induzierte Untergraph) jeden gradbeschränkten hyperfiniten (z.B. planaren) Graph bis auf ϵn Kanten bestimmt (dabei ist k konstant, hängt aber von anderen Konstanten wie ϵ und dem Maximalgrad des Graphen ab). Die Konstante k ist jedoch für den praktischen Einsatz viel zu groß. Allerdings ist unklar, inwieweit bereits die Verteilung deutlich kleinerer Untergraphen ebenfalls aussagekräftig ist. Wir wollen daher untersuchen, ob Verteilungsvektoren kleinerer Graphen ebenfalls aussagekräftig sind, und ob man hier effizienter werden kann, indem man die in einer Stichprobe gezogenen lokalen Untergraphen nur auf bestimmte Eigenschaften untersucht oder z.B. mit Hilfe von Hashingverfahren auf eine kleinere Repräsentation komprimiert. Bei der Approximation des Verteilungsvektors und beim Hashing muss dabei aber insbesondere aus einem Graph eine eindeutige Nummer generiert werden, was mindestens genauso schwierig wie das Graphisomorphieproblem ist. Aufgrund der kleinen Größe der Untergraphen sollte das aber trotzdem umsetzbar sein, und gerade für hyperfinite Graphen vermutlich auch in polynomieller Zeit möglich sein. Wir wollen auch die bereits erwähnten All-Distances-Sketches zur Berechnung von Graphmerkmalen untersuchen.

Arbeitspaket 2. Effiziente Graphkerne: Hier ist unser Ziel die Entwicklung effizienter, zufallsbasierter und ressourceneffizienter Graphkerne. Auch hier soll die Ressourceneffizienz und Ausdrucksstärke (Güte) für die Graphklassifikation theoretisch und/oder empirisch nachgewiesen werden.

Verallgemeinerung des Weisfeiler-Lehman-Kerns. Der Weisfeiler-Lehman-Graphkern ist wahrscheinlich der zur Zeit vielversprechendste Graphkern. Auf mehreren Benchmarks schlägt er andere State-of-the-Art Graphkerne sowohl in Klassifikationsgüte als auch Laufzeit. Hier werden Informationen über Knoten und Struktur iterativ durch den Graphen propagiert. In jeder Iteration werden die aktuellen Verteilungen von Merkmalen zur Konstruktion von Kernfunktionen extrahiert. Wir wollen diese grundlegende Idee erweitern und untersuchen, ob beliebige Diffusionsprozesse auf Graphen zur Konstruktion von Graphkernen herangezogen werden können. Nehmen wir also an, dass allgemeine Informationen über Knoten und Kanten an benachbarte Knoten und Kanten weitergegeben werden und so durch den Graphen diffundieren. Damit würde nicht nur die Topologie der Graphen sondern auch der Diffusionsprozess in die Berechnungskomplexität der Graphkerne eingehen. Wir erwarten, dass das zu verbesserten Laufzeiten führt. Als Illustration betrachte man einen $n \times n$ Grid-Graphen, wie er häufig in Computer Vision betrachtet wird. Wäre der Diffusionsprozess durch eine Faltung ausdrückbar, wie zum Beispiel das gewichtete Mittel aller Knoten in einer Nachbarschaft, dann reduziert sich der Aufwand pro Iteration von $\Theta(n^2)$ auf $O(n \log(n))$, da Faltungen schnell mittels einer Fouriertransformation berechnet werden können. Auch ergibt sich die Möglichkeit, effiziente probabilistische Inferenzverfahren wie z.B. Loopy Belief Propagation zur Graphklassifikation heranzuziehen.

Unser zweiter Ansatzpunkt zur Erweiterung des Weisfeiler-Lehman-Kerns ist seine Darstellung als konti-

nuierliches Optimierungsproblem. Bisher wurde der Weisfeiler-Lehman-Algorithmus, der den entsprechenden Kernen zu Grunde liegt, rein kombinatorisch betrachtet. Das ist erstaunlich, weil Ramana *et al.* [264] gezeigt haben, dass der Weisfeiler-Lehman-Algorithmus als ein lineares Programm (LP) formuliert werden kann. Die besten LP Solver beruhen aber auf kontinuierlicher Optimierung. Wir werden Formulierungen des Weisfeiler-Lehman-Algorithmus als kontinuierliche Optimierungsprobleme untersuchen. So konnten wir im Zuge der Projektvorbereitung bereits zeigen, dass der Weisfeiler-Lehman-Algorithmus einem Frank-Wolfe Algorithmus zum Lösen des LPs entspricht, der implizit eine Symmetrie-Regularisierung vornimmt [A6/26]. Dieser konnte dann in einen Ansatz für Graphclustering basierend auf Zufallswegen umformuliert werden. Allerdings wäre eine explizite Formulierung der Regularisierung in der Zielfunktion besser, da dann ein beliebiger Solver benutzt werden könnte. Dafür erscheinen semi-definite Programme vielversprechend, mit deren Hilfe schon einige andere kombinatorische Optimierungsprobleme (auch im Graph Drawing) gelöst wurden [230] und die auch im maschinellen Lernen zum Einsatz kommen, siehe z.B. [237, 255]. Auf diese Weise können wir einen ersten Brückenschlag zu zentralen Ansätzen des maschinellen Lernens z.B. für Clustering und das teilweise überwachte Lernen bilden und diese auf Graphkerne übertragen. Damit bildet es die Grundlage in AP5 zur Übertragung von I/O-effizienten Ansätzen des maschinellen Lernens und somit zur einfachen Einbettung von Graphanalyse in existierende Systeme.

Graphkerne durch Zerlegung in einfachere Graphen. Eine Alternative zur Verallgemeinerung von Weisfeiler-Lehman-Graphkernen für komplexe Graphen (deren Knoten Attribute haben oder deren Kanten gewichtet sein können) ist der folgende Ansatz: man reduziert komplexe Graphen auf eine Folge bzw. Menge von einfachen Graphen. Als ein Beispiel betrachte man einen gewichteten Graphen. Wir normalisieren z.B. die Gewichte auf das Intervall $[0, 1]$ und lassen den Schwellenwert ϵ von 0 bis 1 laufen z.B. mit Schrittweite $0,01$. Auf die sich ergebende Sequenz von Graphen können wir jeglichen Graphkern anwenden und die resultierenden Kernfunktionen zu einer einzigen Kernfunktion z.B. mittels einer Summe zusammenfassen. Analog kann man sich vorstellen, Graphen mit kontinuierlichen Knotenattributen in eine Menge von einfachen Graphen zu überführen. Um mit hochdimensionalen Knoten- und Kantenattributen umgehen zu können, sollen angelehnt an AP1 Zufallsprozesse oder -projektionen benutzt werden. Das heisst, wir ziehen eine Menge von Subgraphen aus einem Graphen je nachdem, ob der Attributvektor eines Knotens in einem zufällig generierten Halbraum liegt. Die Hoffnung ist, dass sich wichtige Merkmale des Originalgraphen in der Struktur vieler der generierten Graphen widerspiegeln. Damit sollten sie aber auch prominent in den entsprechenden Graphmerkmalen (auch der in AP1 entwickelten) sein.

Sparsity. Die lokale Struktur eines Graphen kann über die Verteilung von sogenannten k-Scheiben beschrieben werden (siehe oben). Eine interessante Fragestellung ist, ob wir ausnutzen können, wenn der Verteilungsvektor der k-Scheiben dünn besetzt ist, d.h. nur wenige große Einträge hat (ein einfaches Beispiel für solche Graphen sind Gitter). Große Einträge im Verteilungsvektor der k-Scheiben sind häufig vorkommende Untergraphen, die auch als Graph- oder Netzwerkmotive oder Graphlets bezeichnet werden. Wir bezeichnen einen Graph mit n Knoten als (r, ℓ, k)-sparse, wenn es nur ℓ Einträge im Verteilungsvektor der k-Scheiben gibt, deren Wert größer als $1/n^r$ ist. Wir wollen untersuchen, ob man für solche (r, ℓ, k)-sparse Graphen die Spärlichkeit ausnutzen kann, um Graphkerne schnell zu erzeugen, ohne die Klassifikationsperformanz zu beschneiden. Auch wollen wir untersuchen, wie sich der Weisfeiler-Lehman-Kern auf solchen Graphen verhält.

Visualisierung und Graphkerne. Wir wollen ebenfalls untersuchen, inwieweit Techniken zum Zeichnen von Graphen hilfreich bei der Konstruktion von Graphkernen sein können. Nach der Einbettung in die Ebene können einerseits übliche Kernfunktionen angewendet werden und andererseits spezielle Kernfunktionen entwickelt werden, die gewisse Merkmale der Topologie oder Geometrie der Layouts mitberücksichtigen. Wir wollen systematisch untersuchen, ob die Randbedingung, dass die Visualisierungen durch Menschen einfach zu verstehen sind, zu einer verbesserten Klassifikationsperformanz führt. Darüber hinaus sollen interaktive Änderungen der Visualisierung durch den Benutzer zu einer Rückkopplung mit den Graphkernen führen, diese also dynamisch angepasst werden. Das erlaubt es, das Expertenwissen direkt in die Graphkerne einfließen zu lassen.

Arbeitspaket 3. Verteilte Graphen als Eingabe: In den bisherigen APs sind wir implizit davon ausgegangen, dass die Graphen zentral auf einem Rechner vorliegen. Gerade im Kontext von eingebetteten Systemen liegen die Daten oft verteilt vor und die Rechner sind vernetzt; man hat also häufig ein Szenario, in dem verteilte Graphen wichtig sind. In diesem AP wollen wir daher die verteilte Graphklassifikation untersuchen. Aufbauend auf den Ergebnissen der bisherigen APs werden wir Methoden entwickeln, die mit der Einschränkung umgehen können, dass jeder Prozessor nur einen Teil eines Graphen kennt. Wir erwarten, dass mit speziell für solche Situationen entwickelten Methoden weitere Effizienzsteigerungen zu erzielen sind. Dabei werden wir zwei Szenarien betrachten, die unterschiedliche Aspekte der verteilten Graphklassifikation betonen: ein

Knoten pro Prozessor und ein Teilgraph pro Prozessor.

Szenario 1: Ein Knoten pro Prozessor. Dieses Modell, bei dem jeder Knoten des Graphen eindeutig zu einer Recheneinheit (Prozessor) korrespondiert, taucht typischerweise bei Kommunikationsnetzwerken auf. Es stellt sich die Frage, ob wir Graphklassifikatoren entwickeln können, die deutlich weniger Kommunikationskosten verursachen als notwendig wäre, um die Daten zentral zu sammeln (letzteres hätte auch den Nachteil eines hohen Speicherbedarfs). Dazu werden wir neue Algorithmen im *Message Passing* Modell CONGEST [262] entwerfen und analysieren (s. oben). Verteilte Algorithmen für kürzeste-Wege-Probleme wurden bereits seit langer Zeit studiert. Umso erstaunlicher war es, dass erst kürzlich in [251] gezeigt wurde, dass das ungewichtete APSP-Problem in diesem Modell mit Hilfe eines einfachen Algorithmus in $O(n)$ Runden gelöst werden kann, was bis auf einen logarithmischen Faktor optimal ist. Weiterhin wird ein Algorithmus zur Lösung des SSSP-Problems in $O(D)$ (D ist der Durchmesser des Graphen) Runden dazu benutzt, den Durchmesser, den Radius, und die zentralen Knoten des Netzwerks zu approximieren. Dies geschieht mittels der Berechnung einer kleinen k-dominierten Knotenmenge und anschließenden SSSP-Berechnungen ausgehend von dieser kleinen Menge. Wir wollen versuchen, unsere Ansätze zur Berechnung des Centrality-Indexes (siehe zufallsbasierte Merkmalsextraktion aus AP1) auf dieses Setting und insbesondere die einfachen Algorithmen aus [251] zu übertragen. Unser Ziel ist hierbei, neue approximative verteilte Verfahren zur Berechnung kürzester-Wege-basierter Merkmale zu entwickeln.

Eine andere Linie, die wir hier verfolgen wollen, ist die Entwicklung und Analyse von verteilten Algorithmen für Steinernetzwerke. Insbesondere wollen wir die bisher ([227] und auch in eigenen Vorarbeiten) erzielte 2-Approximationsgüte im oben genannten verteilten Modell auf verwandte Probleme, wie z.B. das Prize-Collecting Steinerbaumproblem, übertragen. Die bestehenden verteilten Verfahren (z.B. [227]) gehen jeweils davon aus, dass die Steinernetzwerke vollkommen neu berechnet werden. Wir werden die Frage untersuchen, wieviel man an Rechenzeit (bzw. Runden) gewinnen kann, wenn man bereits ein Steinernetzwerk berechnet hat und sich nun einzelne dynamische Veränderungen des Netzwerks ergeben. Die hier erzielten Ergebnisse bilden die Grundlage für Lernverfahren, die mittels Graphkernen klassifizieren. Ein Beispiel ist die Berechnung von Weisfeiler-Lehman-Graphkernen, die entscheiden sollen, in welche Klasse von Netzwerktopologie das aktuelle Netzwerk (gegeben durch die Recheneinheiten und deren Verbindungen) fällt. Hierzu muss jeder Knoten z.B. lokal entscheiden, ob seine Umgebung zu einem vorgegebenen Netzwerk isomorph oder zumindest ähnlich ist. Auch dieses bisher noch nie betrachtete Problem wollen wir studieren.

Szenario 2: Ein Graph verteilt auf viele Prozessoren. In diesem Szenario kennt jede Recheneinheit jeweils nur einen kleineren Teil des riesigen Netzwerks, wie es z.B. bei den Car2X-Kommunikationsnetzen der Fall ist (jedes Auto kennt einen Teil der Strecke). Hier beschränken nicht die Kommunikationskosten das System (wie in Szenario 1) sondern der Speicher und die Laufzeit. Es beinhaltet insbesondere den Fall, dass wir einen oder mehrere große Graphen mit Hilfe von vielen Rechnern (auf verteilten Systemen) analysieren wollen. Wir werden zunächst annehmen, dass die Knotenmenge allen Rechnern bekannt ist und nur die Kantenmenge auf die Rechner verteilt ist. Je nach Modellvorgabe kann man entweder annehmen, dass die Kantenmengen dabei disjunkt sind oder dass dieselbe Kante auf mehreren Rechnern abgespeichert ist. Letzteres ist z.B. in Straßennetzwerken der Fall: Jedes Auto kennt jeweils nur die Verkehrssituation in seinem lokalen Bereich, in dem es gerade unterwegs ist. Außerdem soll der Platzbedarf jedes Rechners auf $\widetilde{O}(n)$ beschränkt sein. Da wir hier erstmal nur einen Graph haben, stellt sich vor allem die Frage, ob man die Merkmale aus dem ersten Arbeitspaket in diesem Szenario effizient berechnen kann. Diese Fragestellung weist Verbindungen zum sogenannten Semi-Streaming-Modell auf, in dem man $\widetilde{O}(n)$ Speicherplatz hat und die Kanten nur einmal sequentiell lesen kann. Der Unterschied liegt darin, dass wir zwar freien Zugriff auf die Kanten an jedem Rechner haben, aber die Kommunikation zwischen den Rechnern beschränkt ist. Nun stellt sich die Frage, wie viel Kommunikation (Runden, Volumen, etc.) man benötigt, um z.B. kürzeste Wege oder minimale Spannbäume oder nur die Verteilungen der Kantenlängen in diesen Strukturen zu berechnen.

Des Weiteren wollen wir untersuchen, wie man die Graphkerne aus AP2 auf riesige verteilte Graphen mit mehreren Millionen von Knoten und Kanten anwenden kann. Hierfür wollen wir gängige parallele Systemen untersuchen. Insbesondere prüfen wir die Verwendung von HADOOP, die Open Source-Version von MapReduce, sowie von GraphLab [259], das mittels des Message Passing Interface (MPI) die Parallelisierung von maschinellen Lernverfahren auf und mit Graphen erlaubt. Die entsprechenden Algorithmen zur Graphklassifikation aus AP2 werden auf die Umsetzung in diesen Plattformen untersucht und — sofern möglich — angepasst. Besonders spannend finden wir hier auch die Umsetzung der Algorithmen aus AP2 speziell für die Berechnung direkt von auf Festplatten gespeicherten Graphen. Hier planen wir, eventuell auf das Open Source-Framework GraphChi [256] aufzubauen, das speziell zum Lösen großer Lernprobleme auf kleinen oder eingebetteten Geräten konzipiert wurde.

Arbeitspaket 4. Dynamisches und adaptives Graph Mining: Kommunikationsnetzwerke sind ständigen Änderungen unterworfen: Knoten oder Kanten fallen weg oder kommen neu hinzu, Knoten- oder Kantenattribute ändern sich. Adaptive Verfahren passen sich diesen Änderungen an. Treten die Änderungen nur lokal auf und haben nur lokale Auswirkungen für die berechneten Werte (wie z.B. bei kürzesten Wegen die Distanzwerte), sollen sogenannte dynamische Graphenalgorithmen und -kerne untersucht werden. Das Ziel ist dabei, vom aktuellen Status auszugehen und nur *soviel wie notwendig* zu berechnen statt alles wieder neu zu berechnen.

Dynamische Merkmalsextraktion. Wir wollen effiziente dynamische Verfahren zur Berechnung verschiedener Graphmerkmale für z.B. Kommunikationsnetzwerke entwickeln und hierfür die in AP1 gewonnenen Erkenntnisse benutzen. Weiterhin werden wir untersuchen, wie in Batches kommende Netzwerkänderungen ausgenutzt werden können.Bei der Zerlegung komplexer Graphen in einfachere Graphen, wie oben vorgeschlagen, hat man es mit dem Spezialfall inkrementeller Graphen zu tun (es wird oft monoton entfernt oder addiert). Wir untersuchen, ob dies zu schnelleren Berechnungen führen kann. Der Beweis in [241], der zeigt, dass das dynamische Subgraphisomorphieproblem bereits für eine einzige Graphmodifikation unbeschränkt ist, basiert auf einem Baum und einem Wald. Wir hingegen wollen die bekannten Spezialfälle untersuchen, für die das Subgraphisomorphieproblem in polynomieller Zeit lösbar ist, wie z.B. für zwei gegebene partielle k-Bäume, die beide k-zusammenhängend sind oder beschränkten Knotengrad besitzen. Dabei gehen wir davon aus, dass die Graphmodifikation diese Eigenschaft beibehält. Wenn wir also z.B. zwei Bäume als Eingabe für das Subgraphisomorphieproblem betrachten, dann erlauben wir nur solche Graphmodifikationen, die baumerhaltend sind. Dies ist gerade für inkrementelle Modifikationen, in denen nur hinzugefügt wird, sinnvoll und wird – falls es zu beschränkten dynamischen Algorithmen führt – zu deutlichen Verbesserungen der Laufzeit führen.

Adaptive Graphkerne. Die Frage, inwiefern man Graphkerne schnell neu berechnen kann, wenn sich nur einzelne Daten ändern, wird untersucht. Wir studieren mögliche Vorberechnungen, die zusätzliche Informationen beinhalten, und mit deren Hilfe z.B. der Weisfeiler-Lehman-Graphkern beschränkt dynamisch berechnet werden kann. Das heisst, dass nur solche Daten während der Berechnung angefasst werden müssen, die benachbart zu denjenigen Daten sind, die im Laufe der Berechnung andere Werte erhalten. Unter anderem wollen wir dazu auch die in AP2 entwickelten Formulierungen des Weisfeiler-Lehman-Algorithmus als kontinuierliches Optimierungsproblem verwenden. Wir wollen untersuchen, ob sie zu adaptiven Graphkernen führen. Wenn sich der Graph nur lokal ändert, so sollte der neue Kern nicht weit weg vom ursprünglichen Kern liegen. Eine Initialisierung des Solvers mit dem alten Kern im Rahmen einer Sensitivitätsanalyse sollte zu einer Beschleunigung führen. Im Fall dieser Beschleunigung wollen wir untersuchen, ob der kontinuierliche Ansatz wieder in einen kombinatorischen überführt werden kann [220].

Arbeitspaket 5. Algorithm Engineering: Unser abschließendes AP und Teilziel ist eine weitere Effizienzsteigerung der entwickelten Algorithmen zur Graphklassifikation durch Einbindung von Wissen über Eingabestrukturen oder -verteilungen. In systematischen Experimenten soll untersucht werden, welche Eigenschaften der Graphkerne und Graphmerkmale effiziente Lösungen unter Ressourcenbeschränkungen ermöglichen. Dazu werden wir uns auf zwei Kernansätze aus AP1 und AP 2 konzentrieren: die Berechnung von kürzesten Wegen und das Testen von (Subgraph-) Isomorphie. Dieses wird eng mit der empirischen Evaluierung aller Methoden der vorangegangen APs verknüpft erfolgen.

Kürzeste-Wege Berechnungen. Sowohl verteilungsbasierte Graphmerkmale aus AP1 und der Weisfeiler-Lehman-Graphkern aus AP2 stehen in enger Verbindung zu kürzeste-Wege-Algorithmen. Während die Einbindung von Eingabestrukturen vor allem bei (s, t)-Abfragen in Straßennetzwerken zu neuen Erkenntnissen geführt haben, haben wir es hier mit normalen single-source-shortest-path (SSSP) oder sogar all-pairs-shortest-path (APSP) Problemen zu tun. Wir wollen die Eigenschaften bestimmter großer Graphen (z.B. der zur Verfügung stehenden Webgraphen) studieren, um diese für schnellere SSSP-Berechnungen auszunutzen. So ist z.B. bekannt, dass manche dieser großen Netzwerke nur relativ wenige Knoten mit hohem Knotengrad besitzen (Hubs) oder nur einen großen Kern und viele Randbezirke besitzen [258]. Bei der approximativen Centrality-Berechnung in Kernmengen (s. AP 1) berechnet man eine Folge von SSSP Berechnungen, die aufgrund der Auswahl der Kernmengen eine gewisse Struktur besitzen. Wir wollen daher versuchen diese speziellen Strukturen für die Berechnungen auszunutzen. Eventuell kann man auch die Auswahl der Kernmengenpunkte auf spezielle, schneller zu berechnende SSSP Folgen, angleichen. Wir wollen dabei auch auf die Ressourcenbeschränkungen Speicherplatz und I/Os achten.

Subgraphisomorphie. Das Ausnutzen von Eingabestrukturen für Subgraphisomorphie wird die Berechnung von Graphkernen (AP2) beschleunigen. Bei Anwendungen, wie z.B. der Bildklassifizierung, ist es sinnvoll, die topologische Einbettung der Graphen zu erhalten. Diese Version des *Submap-Isomorphie* Problems kann unter gewissen Randbedingungen in polynomieller Zeit gelöst werden [236]. Wir wollen untersuchen, ob wir

diese Randbedingungen reduzieren können. Weiterhin wollen wir untersuchen, wie die Integration weiterer geometrischer Informationen, z.B. Winkel zwischen den eingebetteten Kanten, erfolgreich in die Subgraphisomorphiesuche gelingen kann. Die Topologie von Kommunikationsnetzwerken hängt sehr stark von den verwendeten Algorithmen ab. Möchte man die Energie minimieren, dann basieren die Berechnungen auf einem minimalen Spannbaum, der relativ dünn ist; möchte man hingegen die Inferenzen oder auch die Routingtabellen klein halten, dann nimmt man eher Delaunay-Graphen, die jedoch deutlich dichter sind. Wir untersuchen Topologien von Kommunikationsnetzwerken, die mit Hilfe geometrischer Spannergraphen, wie z.B. Delaunay Graph, Gabriel Graph, Yao Graph, erstellt werden auf Eigenschaften, die das Subgraphisomorphieproblem polynomiell berechenbar machen. Hierfür eignet sich z.B. die Eigenschaft *partieller k-Baum mit beschränktem Knotengrad.*

Experimentelle Evaluierung. Wir werden unsere entwickelten Algorithmen kontinuierlich umsetzen und experimentell evaluieren. Die Ressourceneffizienz unserer neuen Verfahren wird empirisch nachgewiesen. Für die Evaluierung verwenden wir jeweils State-of-the-Art Benchmark Libraries sowohl aus der Graphenalgorithmik (wie z.B. von DIMACS Computation Challenges) als auch aus der Graph Mining Community (wie z.B. aus dem Stanford Network Analysis Project, der UF Sparse Matrix Collection, den üblichen Molekül-datenbanken und anderen). Weiterhin werden wir in enger Kooperation mit den anderen für uns relevanten Teilprojekten vielversprechende Verfahren in die jeweiligen Systeme umsetzen und praktisch evaluieren (s. Stellung innerhalb des Sonderforschungsbereichs).

Zeitplan

	2015				2016				2017				2018			
Arbeitspaket / Quartal	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Ressourcenbeschränkte Merkmalsextraktion aus Graphen:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2. Effiziente Graphkerne:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
3. Verteilte Graphen als Eingabe:									■	■	■	■	■	■	■	■
4. Dynamisches und adaptives Graph Mining:					■	■	■	■	■	■	■	■	■	■		
5. Algorithm Engineering:			■	■	■	■	■	■	■	■	■	■	■	■	■	■

3.5 Stellung innerhalb des Sonderforschungsbereichs

Netzwerke und verlinkte Daten treten in vielen Teilprojekten des Sonderforschungsbereichs auf. A6 widmet sich der Entwicklung ressourceneffizienter Algorithmen zur Klassifikation von Graphen mit theoretischer Fundierung. Im Gegensatz zu A1, in dem probabilistische Verfahren zur Datenanalyse entwickelt und untersucht werden, sind in A6 nicht die Modelle sondern die Daten verlinkt. Graphen sind nicht vektorieller Natur und können daher nicht ohne Weiteres mittels A1-, A2- bzw. A3-Methoden modelliert werden. Die verteilten Berechnungsaufgaben sind nicht unabhängig sondern miteinander verbunden, weil schon die Datenpunkte (die Knoten der Graphen) miteinander verbunden sind. Zusammen mit B3 soll in AP3 untersucht werden, wie sich die dort entwickelten verteilten Analysemethoden auf Graphen übertragen lassen. Die Erfahrungen mit Kernmengenmethoden aus A2 fließen direkt in AP1 und AP5 ein. Zusammen mit A1, das graphische Wahrscheinlichkeitsmodelle untersucht, soll ein Austausch an Methoden etabliert werden und so zur Effizienzsteigerung der im jeweiligen Teilprojekt untersuchten Methoden führen (z.B. AP2 in A6). Auch sollen Erfahrungen mit verteilten Plattformen wie z.B. GPUs und GraphLab ausgetauscht und neue Methoden innerhalb der Plattformen entwickelt werden. Weiterhin soll gemeinsam mit A1 untersucht werden, ob sich Erweiterungen von den in A1 entwickelten Baumkernen, um Aufrufgraphen von Betriebssystemfunktionen zu klassifizieren, auf Graphen auszahlen. Diese Herangehensweise ist auch eng mit effizienten Inferenzverfahren in graphischen Wahrscheinlichkeitsmodellen verwandt. Hier ist bekannt, dass die wahren *Marginals* durch intelligentes Mischen von entsprechenden Verfahren auf Spannbäumen des dem Wahrscheinlichkeitsmodells zugrunde liegenden Graphen approximiert werden können, siehe z.B. [271].

In Teilprojekt A4 wird untersucht, wie in dicht besetzten Multi-Agenten-Systemen, wie in der Logistik oft anzutreffen, energieeffizient kommuniziert werden kann. Dabei sollen insbesondere auch unnötige Kanalzugriffe vermieden und die Interferenz minimiert werden. Die in A4 verwendeten Kommunikationsknoten haben eine beschränkte maximale Kommunikationsreichweite, aus der sich bei einer vorgegebenen Verteilung ein maximaler Kommunikationsgraph ergibt. Diesen will man aber natürlich typischerweise nicht zur Kommunikation verwenden. Stattdessen benutzt man dünn besetzte und einfach lokal zu berechnende Kommunikationsgraphen wie z.B. den Gabriel Graph, Yao Graph, Θ Graph oder Kombinationen dieser Graphen wie den Smart Boundary Yao Gabriel Graphen. Wir wollen nun gemeinsam mit A4 durch eine Simulation für Beispiele maximaler Kommunikationsgraphen die besten Kommunikationsgraphen berechnen und dieses Mehrklassenkonzept mit Hilfe der von uns entwickelten Kernmethoden lernen und dann im Anwendungskontext auswerten, siehe auch AP5. Unsere dynamischen kürzeste-Wege-Algorithmen finden direkten Eingang in die Routing-Methoden von A4.

In Teilprojekt B4 sollen Flüsse in Verkehrsnetzwerken untersucht werden. Hier ist die Aufgabe aus historischen und simulierten Verkehrsdaten (die natürlich auch als Netzwerke dargestellt werden können) Wahrscheinlichkeitsmodelle für den Verkehrsfluss zu lernen. Mittels Skalarprodukten zwischen diesen Verteilungen können diese in Zusammenarbeit mit A6 in Graphkerne überführt werden, sodass Verkehrssituationen automatisch in verschiedene Kategorien eingestuft werden können. Alternativ können hierfür die in AP1 und AP2 entwickelten Merkmale und Graphkerne (insbesondere die für Graphen mit kontinuierlichen Attributen) direkt zum Einsatz kommen. Damit kann man dann z.B. (un)typische Verkehrssituationen mittels sogenannter Ein-Klassen-SVMs (Stützvektormaschinen) bestimmen. Dafür wollen wir auf die Expertise von B3 zurückgreifen, das Ausreißer mittels Ein-Klassen-SVMs identifiziert.

Zusammen mit C1 sollen die in AP1 entwickelten verteilungsbasierten Graphmerkmale für die Analyse von biologischen und medizinischen Daten untersucht werden. Hierfür soll eine Posterior-Regularisierung von Graph-Lasso Verfahren mittels der Verteilung der Längen der kürzesten Wege im resultierenden Graph durchgeführt werden. Da gewisse Parametercluster der Weibull-Verteilung anscheinend zu gewissen Graphmodellen wie z.B. skalenfreien Graphen korrespondieren, soll so direkt nach z.B. skalenfreien Gen-Interaktionsgraphen mittels Graph-Lasso gesucht werden.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

Relational exploration, learning and inference

(Kersting,Förderkennzeichen DFG KE 1686/2-1 im Rahmen des SPP 1527: Autonomes Lernen) (Förderzeitraum: 2012 - 2015)

Das Ziel dieses Projekts im Rahmen des DFG Schwerpunktprogramms 1527 „Autonomous Learning“ ist die Erforschung von Exploration, Lernen und Inferenz auf der Grundlage sogenannter relationaler Repräsentationen. Solche Repräsentationen implizieren starke A-priori-Annahmen über die Weltstruktur und ermöglichen es, aus unsicherer Erfahrung kompakte und über einzelne Gegenstände verallgemeinernde Modelle über die Wirkungsweise von Aktionen zu lernen. Dies erlaubt eine konzeptionell neuartige Art von Exploration, die bewusst Objekte in Situationen untersucht, auf welche das derzeitige Weltverständnis nicht verallgemeinert. Graphkerne sind nicht Gegenstand dieses Projekts.

Efficient Inference for Probabilistic Relational Models using Symmetries and Linear Programming Relaxations

(Kersting,Förderkennzeichen GIF 1180) (Förderzeitraum: 2012 - 2015)

Das Ziel dieses Projekts im Rahmen der German Israeli Foundation for Scientific Researchs and Development (GIF) ist die Erforschung von effizienten Schlussfolgerungsmethoden für probabilistische, relationale Modelle. Es fokussiert sich auf Wahrscheinlichkeitsmodelle über endlichen, diskreten Zufallsvariablen und insbesondere auf Relaxierungen des Schlussfolgerns mittels linearer Programme, wenn viele Symmetrien vorhanden sind. Ziel ist ein besseres theoretisches Verständnis für Symmetrien und ihre Ausnutzung zur effizienten Inferenz. Graphkerne sind nicht Gegenstand dieses Projekts.

GraBaDrug: Graph-basierte Methoden für das rationale Wirkstoffdesign

(Mutzel,Förderkennzeichen DFG MU 1129/10-1 im Rahmen des SPP 1736: Algorithms for Big Data) (Förderzeitraum: 2014 - 2017)

Das Ziel des Projekts ist die Entwicklung neuer und effizienter Algorithmen zur Analyse von riesigen Moleküldatenbanken mit bis zu einer Milliarde Moleküle. Auch dort spielt die Ähnlichkeitssuche von Graphen eine wichtige Rolle. Wir betrachten dort jedoch im Gegensatz zu dem Projekt hier das Maximum Similar Subgraph Paradigma das auf dem Maximum Common Subgraph Problem mit erlaubten

Abweichungen basiert und versuchen die Graphen auf dieser Basis zu clustern. Merkmalsextraktion und Graphkerne spielen dabei keine Rolle.

Sublinear Algorithms for the Analysis of Very Large Graphs (Sohler,Förderkennzeichen ERC-StG 307696) (Förderzeitraum: 12/2012 - 11/2017)

Das Projekt beschäftigt sich mit der Entwicklung und Analyse von Zufallsprozessen zur Analyse sehr großer Graphen im Rahmen des Property Testings. Die dort durchgeführten Untersuchungen zielen auf die Bestimmung von Klassen von Grapheigenschaften ab, die mit Hilfe von Zufallsprozessen getestet werden können. Auch wenn die Stichprobenkomplexitäten der Verfahren oft konstant sind, so sind die involvierten Konstanten typischerweise viel zu groß für den Einsatz in der Praxis. Daher können und sollen die dort erzielten Ergebnisse als Startpunkt für unsere Arbeit dienen und in unserem Projekt in aus praktischer Sicht effiziente Verfahren umgesetzt werden. Dies soll im vorgestellten Projekt geschehen.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird bisher nicht gefördert und es wurde kein Antrag auf Förderung gestellt.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	3	176.400	3	176.400	3	176.400	3	176.400
Summe:	—	176.400	—	176.400	—	176.400	—	176.400
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	176.400		176.400		176.400		176.400	

3.7.3 Personal

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (A6)							
Wissen-schaftliches Personal	1	Sohler, Christian, Prof. Dr., Prof	Algorithmische Komplexitäts-theorie	LS 2, Fak. Inf, TU Do	5,0	—	Grund-ausstattung
	2	Kersting, Kristian, Prof. Dr., Prof	Data Mining	LS 8, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	3	Mutzel, Petra, Prof. Dr., Prof	Algorithm Engineering	LS 11, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	4	Molina Ramirez, Alejandro, M.Sc. WiMi	Data Mining	LS 8, Fak. Inf, TU Do	9,95	—	Grund-ausstattung
	5	Zey, Bernd, Dipl.-Inf. WiMi	Algorithm Engineering	LS 11, Fak. Inf, TU Do	7,96	—	Grund-ausstattung
	6	Kriege, Nils, Dipl.-Inf. WiMi	Algorithm Engineering	LS 11, Fak. Inf, TU Do	11,94	—	Grund-ausstattung
	7	N.N. , WiMi	Algorithm Engineering	LS 11, Fak. Inf, TU Do	11,94	—	Grund-ausstattung
	8	N.N. , SHK	Algorithmische Komplexitäts-theorie	LS 2, Fak. Inf, TU Do	5,0	—	Grund-ausstattung
	9	N.N, , SHK	Data Mining	LS 8, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	10	N.N. , SHK	Algorithm Engineering	LS 11, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	11	Jankord, Gundel, Verw. Ang.	—	LS 11, Fak. Inf, TU Do	1,99	—	Grund-ausstattung

	12	Henning, Helmut, Techn. Ang.	—	LS 11, Fak. Inf, TU Do	3,98	—	Grund-ausstattung
Beantragte Personalmittel (A6)							
Wissen-schaftliches Personal	13	N.N. , WiMi	Algorithmische Komplexitäts-theorie	LS 2, Fak. Inf, TU Do	—	Doktorand/in	—
	14	N.N. , WiMi	Algorithm Engineering	LS 11, Fak. Inf, TU Do	—	Doktorand/in	—
	15	N.N. , WiMi	Data Mining	LS 8, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Sohler, Christian**
   Projektleitung. Schwerpunkt algorithmische Komplexitätstheorie. Mitarbeit bei der Entwicklung und Analyse zufallsbasierter Graphkerne und Merkmalsextraktionsalgorithmen.

2. **Kersting, Kristian**
   Projektleitung. Schwerpunkt Graph Mining. Mitarbeit in den Arbeitspaketen 1, 2, 3, 4, 5.

3. **Mutzel, Petra**
   Projektleitung. Schwerpunkt Algorithm Engineering. Mitarbeit in den Arbeitspaketen 2, 3, 4, 5.

4. **Molina Ramirez, Alejandro**
   Mitarbeit an AP3, AP5, AP1.

5. **Zey, Bernd**
   Mitarbeit bei den Steinerbaumvarianten in AP1 und AP3

6. **Kriege, Nils**
   Mitarbeit bei der Entwicklung der Graphkerne in AP2 und bei den Subgraphisomorphieproblemen insbesondere in AP4 und AP5

7. **N.N.,**
   Mitarbeit bei den dynamischen Algorithmen insbesondere Subgraphisomorphie in AP4 und AP5

8. **N.N.,**
   Experimentelle Auswertung von Algorithmen zur Berechnung von Graphmerkmalen

9. **N.N.,**
   Beteiligung bei den Programmierarbeiten in AP5

10. **N.N.,**
    Beteiligung an den Programmierarbeiten in AP5

11. **Jankord, Gundel**
    Verwaltungsunterstützung, wie z.B. Abrechnungen

12. **Henning, Helmut**
    Unterstützung in Hinblick auf die technische Infrastruktur

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

13. **N.N.,**
    Mitarbeit bei der Entwicklung und Analyse zufallsprozessbasierter Merkmale und Graphkerne einschließlich verteilter Algorithmen und Algorithmen Engineering.

14. **N.N.,**
    Mitarbeit bei der Entwicklung der Graphkerne in AP2, den verteilten Algorithmen in AP3, dem dynamischen Graph Mining in AP4 und der Algorithm Engineering Ansätze in AP5.

15. **N.N.,**
    Effiziente Graphkerne (AP2), Verteilte Graphen als Eingabe (AP3), sowie Dynamisches und Adaptives Graph Mining (AP4). Insbesondere Verallgemeinerungen von WL-Kernen basierend auf kontinuierlicher Optimierung und Informationspropagierung. Benchmarking und Evaluation (AP5) auf Basis der neuen Graphkerne.

3.7.4 Sachmittel

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	11.500	11.500	11.500	11.500
Summe der vorhandenen Mittel	11.500	11.500	11.500	11.500
Beantragte Mittel	0	0	0	0

3.1 Allgemeine Angaben zum Transferprojekt TB1

3.1.1 Titel:

Ressourcenbeschränkte Analyse von Spektrometriedaten

3.1.2 Fachklassifizierung:

Statistische Methoden, Metabolomik/Analytik

3.1.3 Projektleitung

Prof. Dr. Baumbach, Jörg Ingo, 02.07.1959

Hochschule Reutlingen, Alteburgstraße 150, 72762 Reutlingen
B & S Analytik GmbH, Otto-Hahn-Straße 15, 44227 Dortmund

Tel.: 07121-271-2043
Fax : 07121-271-90-2043
E-Mail: `joerg.baumbach@reutlingen-university.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Rahnenführer, Jörg, 19.05.1971

Fachgebiet Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, Technische Universität Dortmund
Vogelpothsweg 87
44227 Dortmund

Tel.: 0231-755-3121
Fax : 0231-755-5303
E-Mail: `joerg.rahnenfuehrer@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Anwendungspartner:

B & S Analytik GmbH

Ansprechpartner:
Sprave, Gabriele

Geschäftsführerin
Otto-Hahn-Straße 15
44227 Dortmund

Tel.: 0231-9742-6412
Fax : 0231-9743-6413
E-Mail: `sprave@bs-analytik.de`

B & S Analytik GmbH, gegründet am 16.1.2009. Firmeninhaber: Prof. Dr. Jörg Ingo Baumbach. Branche: Medizinprodukte, analytische Messgeräte. Gegenstand des Unternehmens: die Entwicklung, der Bau und die Vermarktung sowie der Vertrieb von technischen Mess- und Analyseinstrumenten, sowie die Entwicklung und der Vertrieb von Software und Datenbanken. Zahl der Beschäftigten: 7. Jahresumsatz: 500.000 EUR, keine Auslandsbeteiligungen.

3.1.5 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Ziel des Projekts ist die Entwicklung ressourcenbewusster Algorithmen zur Analyse von Spektrometriedaten. Im Mittelpunkt steht die Verarbeitung von Daten, die mittels Ionenmobilitätsspektrometrie (IMS) erzeugt wurden. IMS ist eine Technologie, die kostengünstig die Detektion flüchtiger organischer Metabolite in der Ausatemluft erlaubt. Von dem Anwendungspartner im Transferprojekt stehen Geräte zur Messung von IMS-Daten und die Datenbank *DBreath2012* zur Verfügung. Die Datenbank enthält unzählige Datensätze mit Messungen von IMS-Daten für verschiedene klinische Anwendungen (darunter Krebs, Entzündungskrankheiten und Alzheimer).

Ziel der zweiten Phase ist die Konzeption und Realisierung eines patienten- und bettseitig einsetzbaren Atemluftspektrometers. Basierend auf IMS-Daten sollen Entscheidungsvorschläge für den medizinischen Alltag unterbreitet werden. Dazu werden in der Ausatemluft fragestellungsbezogen relevante Stoffwechselprodukte und in Zeitreihen kritische Entwicklungen identifiziert. Es werden Vorbereitungen für eine Integration in Kleidung und Telemedizinanbindung getroffen. Dies schließt eine weitere Miniaturisierung aller Spektrometerkomponenten von der Probenahme bis zur modularen bedarfsbezogenen Hardwareausgestaltung unter Ressourcenbeschränkungen ein.

In der ersten Phase wurden neue effiziente Methoden zur Vorverarbeitung von Spektrometriedaten entwickelt, speziell zur automatischen Peak-Extraktion aus Rohdaten von IMS-Messungen. In der zweiten Phase werden Gesamtalgorithmen entworfen, die Vorverarbeitung und Methoden zur klinischen Entscheidungsfindung ressourcenbewusst kombinieren. Dazu müssen geeignete statistische Lernverfahren und Datenanalyseprozesse entworfen werden.

Zuerst werden Klassifikationsverfahren für IMS-Daten erstellt, die zwei phänotypisch unterschiedliche Patientengruppen unterscheiden. Hierbei werden sowohl univariate Klassifizierer zur Identifikation von relevanten Metaboliten als auch multivariate Klassifizierer zur Fehlerminimierung bei der Unterscheidung von Gruppen basierend auf verschiedenen Lernverfahren wie zum Beispiel Random Forests oder SVMs betrachtet. Die Modellwahl ist dabei ein wichtiger Teil des Optimierungsprozesses.

Danach werden Methoden für Zeitreihen untersucht und entwickelt, die kritische Wechselpunkte (*change points*) in Verläufen von Metaboliten-Konzentrationen frühestmöglich identifizieren. Mit Glättungsverfahren sollen zunächst typische Verläufe für verschiedene Krankheitsbedingungen bzw. Patientenkohorten geschätzt werden. In einem zweiten Schritt werden dann multivariate Verfahren zur Detektion unterschiedlicher Verläufe zwischen den Gruppen erarbeitet. Ein Beispiel ist eine beginnende Sepsis, deren frühzeitige Erkennung entscheidend für den Therapieerfolg für ein Individuum ist. Hierfür stehen Zeitreihendaten der Universitätsklinik Homburg aus Rattenmodellen an intubierten Tieren zur Verfügung, die zur Modellierung der Entwicklung polymikrobieller wie auch monomikrobieller Sepsis auf Basis der Ausatemluftkonzentration bestimmter Metabolite geeignet sind.

3.3 Bisherige Entwicklung des Transferprojekts

3.3.1 Bericht und Stand der Forschung

Wir berichten zunächst zu den Ergebnissen der ersten Phase des Transferprojekts, die von Jörg Ingo Baumbach (weiterhin in TB1) und Sven Rahmann (in der zweiten Phase Projektleiter in Teilprojekt C1) geleitet wurde. Im Anschluss stellen wir Vorarbeiten des neuen Projektleiters Jörg Rahnenführer dar. Die Änderung bei der Projektleitung resultiert aus einer Änderung des Aufgabenspektrums in der zweiten Phase. Nachdem in der ersten Phase die bioinformatisch-algorithmische Vorverarbeitung der Spektrometriedaten im Vordergrund stand, liegt der Fokus in der zweiten Phase auf der Entwicklung von Gesamtalgorithmen für die medizinische Entscheidungsfindung. Dies beinhaltet statistische Methoden zur Verarbeitung der vorverarbeiteten Daten und Methoden zur optimalen Kombination von Vorverarbeitungs- und Auswertungsalgorithmen. Ziel für die zweite Phase ist die konkrete Konstruktion von funktionsfähigen Gesamtanalyseprozessen für verschiedene Anwendungsszenarien.

In der ersten Phase haben wir uns grundlegend vor allem der Definition und Extraktion von Merkmalen (Peakparametern) aus komplexen Messdaten gewidmet. Während die Peakextraktion bei klassischen massenspektrometrischen Messungen dank klar definierter Peaks unproblematisch ist, gibt es andere Technologien, die breitere, verschwommene Peaks erzeugen, die sich überlappen können und erst identifiziert und durch Modellierung getrennt werden müssen. Die Methoden wurden am Beispiel von Ionenmobilitätsspektrometriemessungen (IM-Spektrometrie-Messungen, IMS-Messungen), die durch Multikapillarsäulen (*multi capillary columns*, MCCs) vorgetrennt wurden, getestet und evaluiert. MCC/IMS ist eine Technologie, die die Detektion flüchtiger organischer Metabolite in der Ausatemluft oder Raumluft bei Umgebungsdruck und -temperatur erlaubt und somit wesentlich einfacher und kostengünstiger als Gaschromatographie/Massenspektrometrie (GC/MS) ist. Entsprechende Geräte wurden durch den Anwendungspartner im Transferprojekt bereitgestellt. Eine beschichtete Säule (MCC) trennt dabei Stoffe anhand physikalischer Eigenschaften. Diese Stoffe könnten ansonsten im Spektrometer nicht unterschieden werden.

Eine wesentliche Herausforderung ist die möglichst genaue Bestimmung der Lage und Intensität der Peaks unter starken zeitlichen Beschränkungen. Dies ist insbesondere im Hinblick auf den Einsatz unserer Methoden in mobilen, gegebenenfalls später in Kleidung integrierten, Ionenmobilitätsspektrometern wichtig, die durch den Anwendungspartner entwickelt werden. Solche miniaturisierten mobilen Spektrometer haben keine Möglichkeit (oder nur bei sehr hohem Energieaufwand), eine vollständige komplexe Messung zu speichern. Sie müssen stattdessen bei laufender Messung alle potenziell für eine Klassifikation wichtigen Merkmale extrahieren.

Stand der Forschung - IMS-Technologie: Die Analyse flüchtiger Metabolite ist herausfordernd. Oft wird Gaschromatographie in Verbindung mit modernen und teuren Massenspektrometern genutzt. Die Ionenmobilitätsspektrometrie ist eine alternative Technik zu dieser Analytik. Ionenmobilitätsspektrometer (IMS) sind kommerziell als handtragbare Version (aber ohne jegliche Kopplung zu gaschromatographischen Vortrenneinheiten) verfügbar und werden meist zur Detektion von Sprengstoffen, chemischen Kampfstoffen oder illegalen Drogen eingesetzt. Seit 2002 werden IMS auch für komplexere Gemische entwickelt, wie zum Beispiel für die Untersuchung der menschlichen Ausatemluft. Um eine große Anzahl verschiedener Analyte trennen zu können, werden Multikapillarsäulen (multicapillary columns, MCCs) eingesetzt. Diese werden vor allem verwendet, weil sie bei Feuchtigkeitsgehalten bis 100% betrieben werden können. Der Stand der Entwicklungen von IMS ist in Übersichtsartikeln von Borsdorf et al. [286] und von Baumbach [285] zusammengefasst.

Generell gibt es eine Reihe unterschiedlicher Techniken, welche für die wissenschaftliche Untersuchung flüchtiger Metaboliten im Spurenbereich eingesetzt werden können. Dies schließt Kopplungen von Gaschromatographie mit Massenspektrometrie (GC/MS), verschiedene massenspektrometrische Methoden (basierend auf Protonen-Transfer-Reaktionen, Ionen-Molekül-Reaktionen oder ausgewählten Ionenflüssen) und laserspektrometrische Methoden (einschließlich *infrared cavity leak-out spectroscopy* oder *tunable absorption spectroscopy*), teilweise in Kombination mit Quantenkaskadenlasern und verschiedenen Arten von Sensoren für die Einzelmoleküldetektion, ein. Einige dieser Methoden verwenden spezielle Beutel zum Sammeln größerer Probenmengen oder benutzen Voranreicherungstechniken etwa auf Basis von SPME (*solid-phase microextraction*). Die IMS bieten eine Reihe praktischer Vorteile im Vergleich zu MS und GC insbesondere bezüglich niedrigerem Gewicht, geringerer Größe, geringeren Betriebskosten und weniger Gasverbrauch; ferner ist für IMS kein Hochvakuum erforderlich. Im Vergleich zu Fourier-Transformations-Infrarotspektrometern (FTIR) haben die IMS niedrigere Nachweisgrenzen. FTIR haben demgegenüber Probleme mit optischen Interferenzen und optischen Fenstern, insbesondere in feuchter Luft. Die Hauptnachteile von spezielleren Sensortechnologien als spektrometrischem Verfahren liegen in deren Querempfindlichkeiten.

Mit Blick auf flüchtige Stoffwechselprodukte wurden MCC/IMS wegen der direkten Probenahme vor Ort, der Möglichkeit der Echtzeitanalyse mittels neuer Algorithmen und wegen der nicht erforderlichen Probenvorbereitung oder Anreicherungsschritte ausgewählt. Seit kurzem werden MCC/IMS bei verschiedenen prozessanalytischen Anwendungen eingesetzt. So wurden erste Anwendungen für die Detektion von Metaboliten bei verschiedenen Krankheiten berichtet [TB1/304]. Zusätzlich wurden flüchtige Stoffwechselprodukte von Bakterien detektiert (*bacterial scent*). Die Visualisierung der Spektren, die Peaksuche, die Datenauswertung sowie die Referenzierung der Peaks wurden in Kooperation der Projektpartner Jörg Ingo Baumbach und Sven Rahmann realisiert; in einer anderen Zusammenarbeit von Herrn Baumbach wurde die Detektion volatiler Stoffwechselprodukte von Bakterienkulturen in flüssigen Medien erfolgreich durchgeführt. Eine Zusammenfassung des Standes der Technik bezüglich der Anwendung von IMS für medizinische Fragestellungen findet sich bei Fink [TB1/302].

Begriffe und Notation: Eine Messung mit einem Ionenmobilitätsspektrometer (IM-Spektrometer), das mit einer Multikapillarsäule (MCC) zur Vortrennung gekoppelt ist, erzeugt eine Matrix von Messwerten (Signalstärken) $S = (S(r,t)) = (S_{r,t})$ mit $r \in R$ (Retentionszeitpunkte), $t \in T$ (Driftzeitpunkte) und $S_{r,t} \in \mathbb{Z}$ (Signalstärke, diskretisiert gemessen). Retentionszeitpunkte sind Zeitpunkte, zu denen ein IM-Spektrum aufgenommen wird. Dies geschieht etwa alle 100 ms. In einem Intervall zwischen zwei Retentionszeitpunkten hat eine bestimmte Stoffmenge die MCC verlassen und wird mit dem nächsten Spektrum analysiert. Im IMS werden die Moleküle ionisiert und durch ein elektrisches Feld beschleunigt. Die Driftzeit ist antiproportional zur Mobilität, die u.a. von der Masse und dem Querschnitt der Ionen abhängt. Es wird zu diskreten Driftzeitpunkten T (bis zu 25 mal pro Millisekunde) die Spannungsänderung auf einer Faraday-Platte gemessen. Dieses Signal, welches mit $S_{r,t}$ bezeichnet wird, wird in einen 12-Bit-Digitalwert konvertiert.

Da die Messzeitpunkte bei aktuell verfügbarer Hardware äquidistant sind, kann man Einheiten geeignet wählen, sodass r und t einfach als Indizes der Matrix S aufgefasst werden können. Tatsächlich ist aber beim Vergleich verschiedener Messungen unter verschiedenen Bedingungen die physikalische Realität zu beachten: Zusätzlich verwendet man statt der Driftzeit eine proportionale normalisierte Größe, die reduzierte inverse Mobilität $1/K_0$, gemessen in Voltsekunden pro Quadratzentimeter, die äußere Einflüsse wie Druck und Temperatur herausnormalisiert [299].

Die aus einer Messung resultierende Datenmatrix S bezeichnen wir als *IM-Spektrum-Chromatogramm* (IMSC). Eine Zeile (fester Retentionszeitpunkt r) heißt *Spektrum*, eine Spalte (feste Driftzeit t) heißt *Chromatogramm*. Da wir (r,t) als Indexpaar auffassen, liegen zusätzlich zu S der Vektor $R = (R_r)$ der Retentionszeiten und der Vektor $T = (T_t)$ der inversen reduzierten Mobilitäten vor. Ein Peak in einem IMSC ist eine Region erhöhter Intensitäten. Eine formale Peakdefinition ist schwierig und war Gegenstand unserer Forschung, siehe nachfolgende Paragraphen.

Eine Menge von N (inhaltlich zusammengehörender) Messungen (S^i, R^i, T^i), $i = 1, \ldots, N$, heißt IMS-*Datensatz*. Dabei kann der Laufindex i eine zeitliche Bedeutung haben. Wir sprechen dann von einer Zeitreihe. Zu den Messungen kann zusätzlich ein Klassenlabel C^i gegeben sein.

Automatisierte Peakerkennung und -extraktion aus IMS-Datensätzen: Ziel war es, Lage und Intensität (Höhe oder Volumen) aller Peaks zu erkennen und parametrisch zu beschreiben. Zu Beginn des Projekts existierten keine Methoden, die eine automatisierte Analyse von IMS-Datensätzen ermöglichten, nicht einmal offline (nach Abschluss aller Messungen). Die existierende Software der im Transferprojekt kooperierenden Firma B&S-Analytik erfordert manuelle Interaktion über eine graphische Benutzeroberfläche und Bewertung visualisierter Messungen durch Experten, sodass eine vollautomatische Analyse, auch offline, nicht möglich war. Möglichkeiten zur Peakdetektion bei laufender Messung fehlten bislang gänzlich. Die semimanuelle Analyse eines IMSC oder eines IMS-Datensatzes durch eine Fachkraft gilt nach wie vor als der Goldstandard.

Da es sich bei der Atemluftdiagnostik um Konzentrationen der Metaboliten bis hinab in den Pikogramm-pro-Liter-Bereich handelt, kommt der Erkennung von Peaks im umliegenden Rauschen besondere Bedeutung zu. Daher ist die erste Frage gewesen, welche Prozeduren das automatische Auffinden eines Peaks ermöglichen. Wir haben daher zunächst die Grundlagen einer automatisierten, interaktionsfreien Peakerkennung und Peakmodellierung erarbeitet, wobei verschiedene effiziente Methoden für jeden Teilschritt gefunden und mit dem Goldstandard verglichen wurden.

Im Laufe dieser Arbeiten ist das Software-Framework PEAX entstanden [TB1/301], mit dem eine Messung vollautomatisch in mehreren Schritten analysiert wird (siehe Abb. 3.1). Die Analyse gliedert sich in die folgenden Schritte:

Vorverarbeitung: Die Rohdaten sind durch Spannungsfluktuationen verrauscht und enthalten zudem das konstante Signal der Ionen des Driftgases (*reactant ion peak*, RIP), das sich als Basislinie bemerkbar

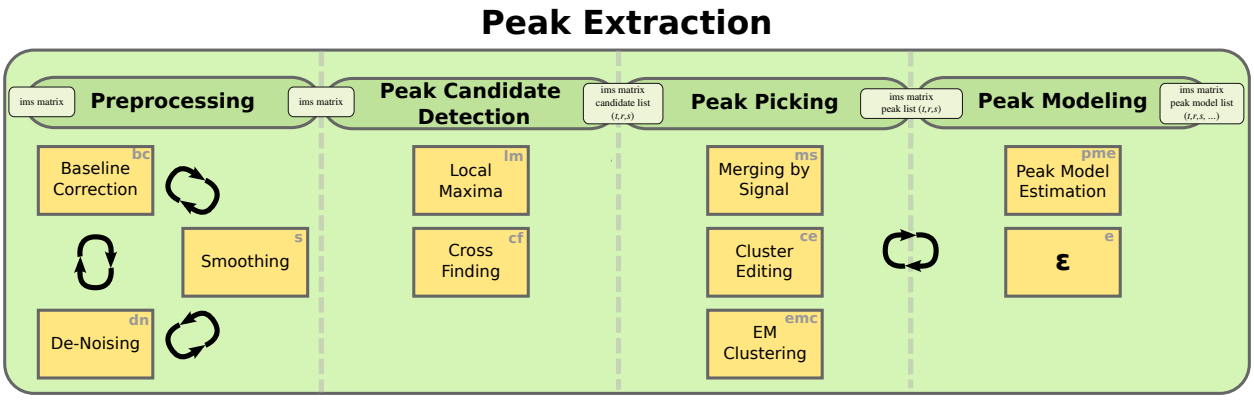


Abbildung 3.1.: Übersicht der Schritte des ersten automatisierten MCC/IMS Analyseprozesses aus unserer Arbeit [TB1/301].

macht. Ziel der Vorverarbeitung ist, diese Einflüsse zu reduzieren, um die folgende Peakerkennung zu vereinfachen. Hierzu haben wir mehrere Module entwickelt [TB1/301].

Detektion von Peak-Kandidaten: Die Detektion von Peaks erfolgt zweischrittig. Im ersten Schritt werden mögliche Ortskoordinaten von Peaks erkannt. Diese werden im zweiten Schritt gefiltert, um die realen Peaks zu erhalten. Eine formale Peakdefinition ist problematisch gerade im Bereich sehr geringer Stoffkonzentrationen (pg/l), die sich kaum aus dem Rauschen abheben und auch für den menschlichen Experten nur mit Vorwissen zu erkennen sind. Die von uns entwickelten Methoden verwenden die Signalhöhe, die Ableitungen (finite Differenzen) in R- und T-Richtung, die Krümmung (finite Differenzen zweiter Ordnung) sowie deren Approximation durch lokale Polynome, z.B. mittels Savitzky-Golay-Filtern. Zusätzlich zu den in [TB1/301] beschriebenen Verfahren beinhaltet diese Aufzählung Methoden, die gemeinsam mit der PG 572 ("Ressourcenbeschränkte Analyse von Ionenmobilitätsspektren mit dem Raspberry Pi"), einer studentischen Projektgruppe, entworfen wurden.

Filtern von Peaks (Peak Picking): Die Peak-Kandidatenliste beinhaltet häufig noch falsch positive Peaks, z.B. aufgrund von verbleibendem Rauschen oder hohen Plateaus in den Signalen. Nahe beieinander liegende Peak-Kandidaten stellen vermutlich dasselbe Molekül dar und werden daher mit Clusteringverfahren verschmolzen.

Modellierung von Peaks: Die Peakdetektionsverfahren liefern nur den Ort (r, t) eines Peaks, aber keine weiteren Parameter. Diese werden in einem Modellierungsschritt bestimmt. Im einfachsten Fall kann man die Signalstärke $S_{r,t}$ als einzigen weiteren Peakparameter heranziehen. Dies ist allerdings eine grobe Approximation und nur in stark ressourcenbeschränkten Systemen nötig. In [TB1/305] haben wir den Ansatz entwickelt, einen Peak unabhängig von zur Modellierung benutzter Verteilungsfunktionen durch universelle Parameter wie Mittelwert, Standardabweichung und Modus (Ort maximaler Intensität) jeweils sowohl auf der R- als auch auf der T-Achse zu charakterisieren. Parameter der Modellfunktionen und universelle Parameter können dabei ineinander umgerechnet werden. Wir konnten auch zeigen, dass ein Distanz- bzw. Ähnlichkeitsmaß zwischen Peaks auf der Basis der universellen Parameter sinnvoller ist als auf Basis der Modellfunktionsparameter. Details zu unserem Vorgehen bei der parametrischen Peakmodellierung werden im nächsten Abschnitt beschrieben.

Jedes Modul der einzelnen Analyseschritte verfügt über eigene Parameter, die vom Benutzer eingestellt werden können und mit (gerätespezifischen) sinnvollen Standardwerten versehen sind. Um Messungen mit verschiedenen Parameterkombinationen effizient analysieren zu können, wurde von uns, gemeinsam mit Teilprojekt C1, eine Programmiersprache für Workflows, *Snakemake*, entwickelt. Diese stellt eine Kreuzung aus GNU Make und Python dar, ist formal eine Erweiterung von Python und erlaubt die regelbasierte Beschreibung von Workflows und deren automatisierte Abarbeitung [TB1/306].

Bei der Evaluation der einzelnen Kombinationen von Modulen (108 verschiedene Kombinationen) gegenüber der manuellen Peakannotation durch eine Fachkraft zeigt sich, dass konsistent ein Fowlkes-Mallows-Index (geometrisches Mittel aus Sensitivität und Positivem Prädiktionswert PPV) von ≥ 0.7 erreicht wird, im besten Fall 0.74. Dabei ist die manuelle Annotation nicht unbedingt richtiger als die automatische. Es wird lediglich eine Aussage über den Grad der Übereinstimmung gemacht. Es zeigt sich eine starke Abhängigkeit vom zu Grunde liegenden Intensitätsschwellenwert, der das Rauschen von schwachen Signalen trennt.

Peakmodellierung: Ausgangspunkt waren vorangegangene Diplomarbeiten seit 2007 (Bödeker, Kopczynski) am Lehrstuhl XI der Fakultät für Informatik. Eine besondere Herausforderung sind zahlreiche niedrige und überlappende Peaks zu frühen Retentionszeiten.

Unser Ansatz, der detailliert in einer auf der EUSPICO 2012 erschienenen Arbeit [TB1/305] beschrieben ist, verwendet zur Trennung und Modellierung der Einzelpeaks Varianten des EM-Algorithmus (*Expectation Maximization*). Grundidee ist, die beobachtete Messung ($S_{r,t}$) als beobachtete Stichprobe einer Mischverteilung aus gleichverteiltem Hintergrund und mehrerer spezifischer Peakverteilungen zu modellieren. Ein Peak wird dabei als Produkt (in R- und T-Achse) zweier verschobener inverser Gaußverteilungen modelliert. Die Modelldichte lautet somit

$$f_{\omega,\theta}(r,t) := \frac{\omega_0}{|R|\,|T|} + \sum_{j=1}^{C} \omega_j\, P_{\theta_j}(r,t),$$

wobei $\omega = (\omega_0, \omega_1, \ldots, \omega_C)$ die Gewichte des Hintergrundmodells (ω_0) und der C verschiedenen Peaks sind, und $\theta = (\theta_1, \ldots, \theta_C)$ der Vektor der jeweiligen Peakparameter ist. Weiter ist $P_\theta(r,t)$ eine zweidimensionale achsenparallele diskretisierte verschobene inverse Gaußverteilung mit

$$P_\theta(r,t) := g_{\mu_\mathrm{R},\lambda_\mathrm{R},o_\mathrm{R}}(r) \cdot g_{\mu_\mathrm{T},\lambda_\mathrm{T},o_\mathrm{T}}(t) \;/\; Z_\theta \quad \text{mit } \theta = (\mu_\mathrm{R}, \lambda_\mathrm{R}, o_\mathrm{R}, \mu_\mathrm{T}, \lambda_\mathrm{T}, o_\mathrm{T}) \text{ und}$$

$$g_{\mu,\lambda,o}(x) := [x > o] \cdot \sqrt{\frac{\lambda}{2\pi(x-o)^3}} \cdot \exp\Big(-\frac{\lambda\big((x-o)-\mu\big)^2}{2\mu^2(x-o)}\Big) \quad \text{(verschobene inverse Gaußverteilung).}$$

Dabei ist Z_θ eine Diskretisierungs- und Normalisierungskonstante, sodass $\sum_{r\in R, t\in T} P_\theta(r,t) = 1$ gilt.

Die Parametrisierung durch verschobene inverse Gaußverteilungen wurde aufgrund ihrer Flexibilität und der guten Übereinstimmung mit den vorliegenden empirischen Daten gewählt [TB1/305].

Ziel ist, die Modellparameter gemäß dem *Maximum-Likelihood*-Prinzip so zu wählen, dass die Wahrscheinlichkeit der beobachteten Daten maximal wird. Wir nutzen Varianten des EM-Algorithmus, um die relativen Gewichte des gleichverteilten Hintergrunds (ω_0), der einzelnen Peaks ($\omega_1, \ldots, \omega_C$), (versteckte) Zuordnungen jedes Messpunkts (r,t) zu allen Peaks und die jeweiligen Peakparameter θ_j, $j = 1, \ldots, C$ zu schätzen. Durch die gute Lokalisierung der Peaks in einem beschränkten Rechteck und dank der dimensionsweisen Separierbarkeit des Peakmodells lassen sich erhebliche Laufzeiteinsparungen erreichen.

Im Rahmen einer durch den Projektleiter Sven Rahmann und Mitarbeiter Dominik Kopczynski betreuten Bachelorarbeit (Henning Funke, LS XI) wurde dieser Ansatz auch auf einer GPGPU implementiert.

Online-Merkmalsgewinnung: Im Zuge der Entwicklung miniaturisierter stark ressourcenbeschränkter IM-Spektrometer war es unser Ziel, Peaks mit ihren Parametern direkt während der Messung zu erhalten und die Roh-IMSCs nicht speichern zu müssen, weil die entsprechenden Geräte nicht über einen großen Speicher verfügen. Dies verbietet insbesondere die Verwendung von Methoden, die Eigenschaften der gesamten Messung oder auch nur über mehrere Spektren hinweg verwenden. Beispiele, die wir im offline-Szenario nutzen konnten, sind Intensitätshistogramme zur Bestimmung des Rauschniveaus oder Ableitungen in Retentionszeit-Richtung. Ein (realistischer) Kompromiss besteht im Zwischenspeichern einer kleinen Anzahl von 5–10 Spektren.

Die für das offline-Szenario entwickelten Methoden, insbesondere die der PG 572, verwenden nur lokale Eigenschaften der Messung, sodass eine Adaption (sofern eine konstante Anzahl Spektren zwischengespeichert wird) elementar ist. Problematisch ist jedoch die Vorverarbeitung, insbesondere die Schätzung des globalen Rauschniveaus und der Basislinie, wozu jeweils die gesamte Messung verwendet wird. Insbesondere enthalten die frühen Spektren viele Peaks, während die späten Spektren gut zur Bestimmung des Rauschniveaus und der Basislinie herangezogen werden könnten. Eine abschließende Lösung muss noch gefunden werden. Pragmatisch ist die im Bericht der PG 572 beschriebene Herangehensweise, mit einer leeren Luftmessung zu starten, daraus die benötigten Parameter abzuleiten und diese in den folgenden Messungen zu verwenden bzw. kontinuierlich zu aktualisieren [TB1/311].

Alignmentmethoden für parametrisierte Spektren: Sowohl bei der Online-Peakdetektion als auch beim Vergleich mehrerer Messungen stellt sich die Frage, wann zwei Peaks dem gleichen Molekül (volatile organic compound, VOC) zuzuordnen sind. Im ersten Fall hat man zu einer festen Retentionszeit r^0 Peaklokationen $t_1^0, \ldots t_n^0$ bestimmt und tut dies analog für die nächste Retentionszeit $r_1 := r_0 + 1$ mit $t_1^1, \ldots, t_m^1$, wobei $m \neq n$ sein kann. Im zweiten Fall hat man Peakkoordinaten $(r_1^0, t_1^0), \ldots, (r_n^0, t_n^0)$ aus der einen Messung und $(r_1^1, t_1^1), \ldots, (r_m^1, t_m^1)$ aus der anderen Messung. Gesucht ist eine optimale Zuordnung der Peaklokationen zwischen den Spektren bzw. Messungen.

Eine physikalisch motivierte Lösung besteht in der Anwendung einer technologiespezifischen Fängerfunktion, die besagt, dass Peaks zu demselben Stoff gehören können, wenn sich deren Retentionszeiten r^0, r^1 um weniger als $3\text{s} + 0.1 \min\{r^0, r^1\}$ unterscheiden und deren inverse reduzierte Mobilitäten t^0, t^1 um weniger als 0.003Vs/cm^2 unterscheiden. Mehrdeutigkeiten werden dabei durch das Nearest-Neighbor-Prinzip aufgelöst [TB1/311]. Obwohl es sich nur um eine Heuristik handelt, führt dieses Vorgehen zu guten Ergebnissen. Eine Formulierung als *Maximum-Matching-Problem* brachte bislang keine erkennbaren Vorteile. Durch Mutterschutz- und Elternzeit ruhen zurzeit die Arbeiten an dieser Formulierung, sie werden 2015 fortgesetzt.

Die ursprünglich erwartete Problematik, dass sich durch Dekalibrierung Driftzeit- und Retentionsachse stark verschieben oder anders skaliert sind, trat in keinem Datensatz während der letzten vier Jahre auf, sodass wir darauf verzichtet haben, diese Problematik näher zu untersuchen.

Evaluierung von Merkmalsgewinnungsmethoden: Die zur Klassifikation einer MCC/IMS-Messung genutzten Merkmale sind grundsätzlich ihre Peaks und deren Parameter. In einem Datensatz kann jedoch die Anzahl der Peaks pro Messung variieren, während viele Klassifikationsalgorithmen Merkmalsvektoren konstanter Länge erwarten. Eine Peakliste pro Messung muss also zunächst in einen Merkmalsvektor konstanter Länge transformiert werden, bevor Merkmale zur Klassifikation eingesetzt werden können.

Bisher wurde die Messung dazu in ein festes Raster eingeteilt und die Gesamtintensität pro Rasterelement als Merkmal genommen. Als Nachteil ergeben sich viele Merkmale, von denen die meisten irrelevant sind, da sie in jeder Messung nur Rauschen enthalten. Einige davon können sich durch Zufall zwischen den Klassen geringfügig aber systematisch unterscheiden. Eine statistische Kontrolle dieses Fehlers führt zu reduzierter Trennschärfe, d.h. es werden weniger relevante Merkmale gefunden.

Unser Ansatz findet die relevanten Merkmale durch *Clustering* aller Peaks des Trainingsdatensatzes: Alle Peaks aller Messungen eines Datensatzes werden anhand ihrer Lage und entweder der bekannten Fängerfunktion oder eines EM-Algorithmus mit achsenparallelen Gauß'schen Komponenten in Cluster eingeteilt. Die Zentren der Cluster werden als Repräsentanten ausgewählt. Die Intensitäten der zugehörigen Peaks in den einzelnen Messungen bilden dann die Merkmalsvektoren, wobei die Intensität eines nicht vorhandenen Peaks auf Null gesetzt wird. Entsprechende Verfahren werden in [TB1/311] und in einem eingereichten Artikel beschrieben.

Die erhaltenen Merkmale und damit auch die Leistung eines darauf basierenden Klassifikators hängen von den zuvor eingesetzten Peakerkennungsmethoden und von den Merkmalsexktraktionsmethoden ab. Um die Leistung eines Gesamtanalyseprozesses vollständig zu messen, haben wir in einer ersten Analyse als Maß die Klassifikationsleistung des aus einem solchen Prozess resultierenden Klassifikators gemessen und verschiedene Kombinationen von Verfahren gegenübergestellt [TB1/303]. Hier konnten zunächst keine grundlegenden Unterschiede zwischen einer automatischen oder manuellen Analyse festgestellt werden. Auch gegenüber Störungen in den Eingabedaten verhielten sich alle untersuchten Methoden ähnlich robust.

Klinische Diagnostik durch Identifikation von metabolischen Biomarkern: Durch das PEAX-Software-Framework wurde es erstmals möglich, Peaks in IMS-Datensätzen automatisch zu extrahieren und in einem geeigneten Format bereitzustellen, das eine Visualisierung erlaubt. So wurde ein direkter Vergleich der Ergebnisse automatischer Prozeduren mit jenen händischen Auffindens von Peaks ermöglicht.

Es wurden folgende Faktoren untersucht, die teilweise noch in anderen zusätzlichen hier nicht angegebenen eigenen Publikationen beschrieben sind:

- Einfluss der Umgebung auf die Ergebnisse - Einatemluft, Materialemissionen, Klimaanlagen, Mundpflegeprodukte [TB1/310],
- Stabilität der Ergebnisse eines Spektrometers (Fluktuationen) und Vergleichbarkeit der IMS-Chromatogramme unterschiedlicher Spektrometer, d.h. der Peaklagen und Peakintensitäten untereinander [TB1/310],
- Bildung und Vergleich eines Referenzgemisches aus sieben flüchtigen organischen Verbindungen [TB1/305, TB1/308],
- Bildung und Vergleich von Referenzgasen wie synthetischer Luft, trocken und befeuchtet [TB1/305],
- Bildung, Zusammensetzung und Messprogramm einer Kontrollgruppe [TB1/305],
- Zeitreihen für einen bekannten Analyten (Propofol) über einen Messtag bzw. während einer Operation [TB1/305],

- Vergleich von gut charakterisierten Patientengruppen mit meist nicht ebenso gut charakterisierten Kontrollgruppen [TB1/309],
- Untersuchung des Einflusses unterschiedlicher Messzeitpunkte (z.B. Kontrollgruppe vormittags, Patientengruppe nachmittags, Patientengruppe über das Jahr verteilt, Kontrollgruppe kampagnenartig in einem Monat fokussiert) [TB1/309, TB1/310],
- Vergleich der Tagesgänge der Raumluft unterschiedlicher Krankenhäuser [TB1/310]
- Vorbereitung des direkten visuellen Vergleiches von Daten aus MCC/IMS- mit jenen aus GC/MSD-Messungen eines Patienten mittels Integration der Chromatogramme (GC) in die Darstellung von IMS-Chromatogrammen anstatt der Höhenlinien - Bereitstellung einer SOP [TB1/307],
- Anbindung der Peakauswertung aus GC/MSD an NIST-Datenbanken für den späteren Vergleich von MCC/IMS- und GC/MSD-Messreihen [TB1/308],
- Extraktion von relevanten Unterschieden (mit Signifikanzwerten von 95%, 99% und 99,9%) bei verschiedenen klinischen Fragestellungen (Sepsis, Asbestose, Neutropenie, Alzheimer, pathogene Bakterien, Lungenkrebs, COPD etc.) inklusive Berücksichtigung der Peakanzahl, welche in der Regel größer war als die Probandenanzahl, und Anwendung statistischer Methoden nach Bonferroni zur Qualitätskontrolle [TB1/300, TB1/307, TB1/309].

Referenzspektren: Referenzspektren im Sinne von Referenz-IMSCs liegen für Raumluftmessungen, feuchtes Referenzgas (Verunreinigungen im Trägergas) und das Kalibrationsgemisch (7er und 16er Gemisch) als Zeitreihen vor. Zusätzlich wurden im Vergleich händisch ausgewertete IMS-Chromatogramme von Messungen zu verschiedenen Fragestellungen betrachtet. Schwerpunkt war der Versuch der Transformation, welche Signale in MCC/IMS-Chromatogrammen so beschreibbar sind, dass sie mittels automatischer Methoden aufgefunden werden können und wie hoch die auf eine händische Auswertung bezogenen Raten von falsch Positiven bzw. falsch Negativen sind. Hierfür wurde das Softwarepaket *multi-peax* entwickelt, welches erlaubt automatisch schrittweitengesteuert Parametervariationen zu realisieren [TB1/301, TB1/305].

Ausgehend von der systematischen Parametervariation sollen nun Gütefaktoren ermittelt werden, sodass die für die jeweilige Peakfindung passenden vorgeschlagen und festgelegt werden können. Dabei soll auch berücksichtigt werden, dass die optimale Parametereinstellung möglicherweise retentionszeitabhängig ist. Dafür soll aus Verdünnungsreihen am Referenzgemisch R7 der Zeitpunkt des Verschwindens eines Peaks aus dem Gemisch gesondert betrachtet werden, da hieraus zusätzliche Abbruchkriterien erwartet werden.

Weiterentwicklung miniaturisierter IMS-Spektrometer: Ausgehend von einer organisch gewachsenen Funktionsmusterlösung, welche konstruktiv auf der Grundlage erprobter Dichtungskonzepte für IMS basierte, wurde hardwareseitig eine Lösung gefunden, mit der das Volumen des Spektrometers auf 1/3 reduziert werden konnte. Die bisher einzeln verbundenen Komponenten Driftröhre mit Ionisationsraum, Ventilblock, MCC-Vortrennung, Pumpe und Massenflusskontrolleinheiten wurden innerhalb eines einzelnen Blockes fixiert. Die elektronisch nebeneinander stehenden Einheiten wurden versorgungspannungsseitig zusammengefasst und mikroprozessorgesteuert in ein System integriert. Hierfür wurde eine Plattform entwickelt, welche eine Ansteuerung aus einem einzigen Bildschirm sowohl für spontanatmende als auch für beatmete Patienten und über Referenzsystemen erlaubt. Auf diese Weise sind aus zwölf unabhängigen Platinen jetzt drei geworden.

Weiterentwicklung integrierter IMS-Softwaresysteme: Zusätzlich konnte der Systemcharakter dadurch ausgebaut werden, dass externe Einheiten im OP wie EEG-Monitore und TCI-Pumpen (*Target Controlled Infusion*) in das Anzeigesystem integriert wurden, sodass alle klinisch relevanten Parameter und Signalentwicklungen auf einem einzigen Bildschirm dargestellt und für spätere Modellbildungen gespeichert werden. Hierfür wurden in der Software *VisualNow* eine direkte Anbindung an VOCan geschaffen, sodass ausgewählte Peaks minimalretardiert maximal 10 Sekunden nach dem Ende der Messungen grafisch darstellt und signalseitig ausgewertet werden können. Sofern für die jeweiligen Analyte Kalibrationskurven existieren und hinterlegt sind, können auch quantitative Aussagen erfolgen. Eine Softwarelösung mit automatischer Zuordnung der Peaks nach Retentions- und Driftzeit zu einer Referenzdatenbank nach Position und Überlappungsgrad konnte erfolgreich getestet werden.

Verschobene Arbeiten: Die folgenden ursprünglich geplanten Arbeitspakete wurden verschoben.

Bei der Alignierung von Peaks zwischen Messungen (oder zwischen aufeinanderfolgenden Spektren) sind die Arbeiten an der Formulierung des entsprechenden Matching-Problems, insbesondere der Modellierung einer passenden Bewertungsfunktion, noch nicht abgeschlossen und werden nach der Elternzeit der Projektmitarbeiterin 2015 fortgeführt.

Die Idee, aus der chemischen Strukturformel die Lage des entsprechenden Peaks in einem IMSC vorherzusagen, bzw. entsprechende Modelle mittels maschineller Lernverfahren zu trainieren, musste aufgrund mangelnder Trainingsdaten zurückgestellt werden. Zurzeit liegen nur 7er- bzw. 16er-Gemische bestimmter Moleküle vor, sodass einzelne Moleküle nicht eindeutig identifiziert werden können.

Klassifikationsalgorithmen: Die klinische Datenbank *DBreath* mit IMS-Messungen enthält mehr als 300.000 Datensätze und über 60.000 Messungen aus Zeitreihen. Diese beziehen sich auf sehr unterschiedliche Anwendungen, da die Messung von Metaboliten in der Atemluft für verschiedenste Zwecke nutzbar gemacht werden kann. Eine der wichtigsten und häufigsten Fragestellungen ist die der Klassifikation von Beobachtungen in verschiedenen Klassen. Bei Krebserkrankungen oder neurodegenerativen Krankheiten wie Alzheimer kann dies zum Beispiel diagnostische Fragestellungen betreffen, in der Krebsforschung können es Patienten mit und ohne Ansprechen auf eine Therapie sein, bei entzündlichen Erkrankungen Patienten mit und ohne schwerwiegender Entzündungsreaktion. In der Anästhesieüberwachung mit minimaler Verzögerung ist das Ziel, schwerwiegende Krankheitsereignisse frühzeitig zu entdecken. Aktuelle Beispiele aus der Datenbank sind die Neutropenie von Kindern, die polymikrobielle Sepsis bei Ratten, die Asbestose und die Silikose bei Patienten und Peritonitis-Signale vor Abschluss einer Operation.

Alle diese Beispiele haben gemeinsam, dass ein Klassifikationsproblem zu lösen ist. Die einzelnen Messungen enthalten oft mehr Variablen als Beobachtungen. Im Vergleich zu hochdimensionalen Daten, wie sie etwa in den Teilprojekten A3, C1, C3 und C4 untersucht werden, ist die Variablenanzahl hierbei nicht in den Zehntausenden oder noch höher; es können aber einige Hundert Metabolite gemessen werden. Für die Klassifikation von Daten mit mehr Variablen als Beobachtungen stehen verschiedene Verfahren zur Verfügung. Eine Übersicht findet sich in [298]. Standardansätze aus der Statistik sind regularisierte Versionen von logistischer Regression und linearer Diskriminanzanalyse [293], bei denen auf die Log-Likelihood ein Strafterm für die Parameterschätzer addiert wird, bevor dieser maximiert wird. Eine beliebte Alternative ist das k-Nearest-Neighbor-Verfahren, bei dem nicht Prototypen für Klassen bestimmt werden, sondern die multivariat ähnlichsten Beobachtungen der Trainingsmenge für die Vorhersage herangezogen werden. Support Vektor Maschinen [287, 290] modellieren explizit die lineare Hyperebene, die jeweils zwei Klassen trennt und sind oft erfolgreich angewendet worden. Schließlich werden baumbasierte Methoden wie Klassifikationsbäume [289] eingesetzt. Sie haben den Vorteil einfacher Interpretierbarkeit, sind aber oft unstabil. Die Erweiterung zu *Random Forests* [288] ist ein *Ensemble*-Verfahren, bei dem ähnlich wie bei *Bagging* und *Boosting* [295] verschiedene einfache Klassifizierer zu einem Verfahren mit höherer Klassifikationsgüte und größerer Stabilität kombiniert werden.

Zusätzlich ist eine vorgeschaltete Variablenselektion zu beachten, bei der vor der Konstruktion des Klassifizierers einzelne Variablen eliminiert werden. All dies muss in einem geeigneten Kontext der Modellwahl und Modellvalidierung geschehen, was im Standardfall mit Hilfe der Kreuzvalidierung oder bei gleichzeitigem Parametertuning mit geschachtelter Kreuzvalidierung geschieht. Für IMS-Daten existieren noch keine ausführlichen Vergleichsstudien all dieser Klassifikationsansätze. Vorarbeiten dazu sind in [TB1/303] zu finden. Ferner wurde an der Fakultät Statistik der TU Dortmund unter Betreuung von Jörg Rahnenführer von Stephanie Mau eine Bachelorarbeit geschrieben (April 2014), die Methoden der linearen Diskriminanzanalyse und k-Nearest-Neighbor-Verfahren kreuzvalidiert vergleicht.

Zeitreihenanalysen: Besonders bei der Online-Detektion von krankheitsrelevanten Ereignissen kommt der statistischen Analyse von Zeitreihen von IMS-Messungen eine große Bedeutung zu. Hierbei ist das Ziel, einen Wechselpunkt (*change point*) zuverlässig und so früh wie möglich zu erkennen. Dafür stehen im Eindimensionalen Standardverfahren zur Verfügung. Für die hier betrachtete Situation kommen zwei Erschwerungen hinzu. Einmal repräsentiert nicht notwendig nur ein einzelner Metabolit den Wechselpunkt, es müssen wiederum multivariate Methoden betrachtet werden. Zum Zweiten müssen robuste Verfahren angewendet werden, da zu einzelnen Zeitpunkten durch Fehlmessungen oder schwankende Umweltbedingungen Abweichungen vom generellen Verlauf auftreten können.

Für die Erkennung von Wechselpunkten stehen noch keine für die Eigenschaften von IMS-Messungen angepassten bzw. sogar optimierten Verfahren zur Verfügung. Als konkrete Vorarbeit wurde bereits an der Fakultät Statistik eine Masterarbeit durchgeführt (Salome Horsch, Analyse zeitabhängiger IMS-Messungen,

Juni 2014), in der einfache Glättungsverfahren für die Zeitreihen mit einer multivariaten Diskriminanzanalyse für Klassifikationsaufgaben gekoppelt werden.

Literatur

[285] J. Baumbach. „Ion mobility spectrometry in scientific literature and in the International Journal for Ion Mobility Spectrometry (1998–2007)“. In: *International Journal for Ion Mobility Spectrometry* 11.1-4 (2008), S. 3–11.

[A3/153] B. Bischl, M. Lang, O. Mersmann, J. Rahnenführer und C. Weihs. *Computing on High Performance Clusters with R: Packages BatchJobs and BatchExperiments*. Techn. Ber. 1. TU Dortmund, 2012.

[286] H. Borsdorf, T. Mayer, M. Zarejousheghani und G. A. Eiceman. „Recent Developments in Ion Mobility Spectrometry“. In: *Applied Spectroscopy Reviews* 46.6 (2011), S. 472–521.

[287] B. E. Boser, I. M. Guyon und V. N. Vapnik. „A training algorithm for optimal margin classifiers“. In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory (COLT)*. Hrsg. von D. Haussler. 1992, S. 144–152.

[288] L. Breiman. „Random Forests“. In: *Machine Learning* 45.1 (2001), S. 5–32.

[289] L. Breiman, J. H. Friedman, R. A. Olshen und C. J. Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[290] C. Cortes und V. N. Vapnik. „Support–Vector Networks“. In: *Machine Learning Journal* 20 (1995), S. 273–297.

[291] R. Fried und H. Dehling. „Robust nonparametric tests for the two-sample location problem“. In: *Statistical Methods and Applications* (2011).

[292] R. Fried, J. Einbeck und U. Gather. „Weighted repeated median smoothing and filtering“. In: *Journal of the American Statistical Association* 480 (2007), S. 1300–1308.

[293] T. Hastie, R. Tibshirani und J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Statistics. Springer, 2009.

[294] K. Michalak und H. Kwaśnicka. „Correlation-based feature selection strategy in classification problems“. In: *International Journal of Applied Mathematics and Computer Science* 16.4 (2006), S. 503–511.

[295] R. E. Schapire, Y. Freund, P. Bartlett und S. Lee. „Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods“. In: *Annals of Statistics* (1998), S. 1651–1686.

[296] K. Schettlinger, R. Fried und U. Gather. „Real Time Signal Processing by Adaptive Repeated Median Filters“. In: *International Journal of Adaptive Control and Signal Processing* 24.5 (2010), S. 346–362.

[297] K. Schettlinger, R. Fried und U. Gather. „Robust filters for intensive care monitoring: beyond the running median/Robuste Filter für intensivmedizinisches Monitoring: mehr als ein gleitender Median“. In: *Biomedizinische Technik* 51.2 (2006), S. 49–56.

[298] H. Schwender, K. Ickstadt und J. Rahnenführer. „Classification with High-Dimensional Genetic Data: Assigning Patients and Genetic Features to Known Classes“. In: *Biometrical Journal* 50.6 (2008), S. 911–926.

[299] J. Stach und J. Baumbach. „Ion Mobility Spectrometry - Basic Elements and Applications“. In: *Int. J. Ion Mobility Spectrom.* 5 (2002), S. 1–21.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[TB1/300] J. Baumbach u. a. „Significant different volatile biomarker during bronchoscopic ion mobility spectrometry investigation of patients suffering lung carcinoma“. In: *International Journal for Ion Mobility Spectrometry* 14 (2011), S. 159–166.

[TB1/301] M. D'Addario, D. Kopczynski, J. I. Baumbach und S. Rahmann. „A modular computational framework for automated peak extraction from ion mobility spectra“. In: *BMC Bioinformatics* 15.1 (Jan. 2014), S. 25.

[TB1/302] T. Fink, J. I. Baumbach und S. Kreuer. „Ion mobility spectrometry in breath research“. In: *Journal of Breath Research* 8.2 (2014), S. 027104.

[TB1/303] A.-C. Hauschild, D. Kopczynski, M. D'Addario, J. I. Baumbach, S. Rahmann und J. Baumbach. „Peak Detection Method Evaluation for Ion Mobility Spectrometry by Using Machine Learning Approaches“. In: *Metabolites* 3.2 (2013), S. 277–293.

[TB1/304] M. Jünger, W. Vautz, M. Kuhns, L. Hofmann, S. Ulbricht, J. I. Baumbach, M. Quintel und T. Perl. „Ion mobility spectrometry for microbial volatile organic compounds: A new pathogen identification tool by smelling human pathogenic bacteria“. In: *Applied Microbiology and Biotechnology* 93.6 (2012), S. 2603–2614.

[TB1/305] D. Kopczynski, J. I. Baumbach und S. Rahmann. „Peak modeling for Ion mobility spectrometry measurements“. In: *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE. 2012, S. 1801–1805.

[TB1/306] J. Köster und S. Rahmann. „Snakemake: a scalable bioinformatics workflow engine“. In: *Bioinformatics* 28.19 (2012), S. 2520–2522.

[TB1/307] S. Maddula, T. Rabis, U. Sommerwerck, O. Anhenn, K. Darwiche, L. Freitag, H. Teschler und J. I. Baumbach. „Correlation analysis on data sets to detect infectious agents in the airways by ion mobility spectrometry of exhaled breath“. In: *International Journal for Ion Mobility Spectrometry* 14.4 (2011), S. 197–206.

[TB1/308] T. Schneider, A.-C. Hauschild, J. I. Baumbach und J. Baumbach. „An Integrative Clinical Database and Diagnostics Platform for Biomarker Identification and Analysis in Ion Mobility Spectra of Human Exhaled Air“. In: *Journal of Integrative Bioinformatics* 10.2 (2013).

[TB1/309] M. Westhoff, P. Litterst, S. Maddula, B. Bödeker und J. I. Baumbach. „Statistical and bioinformatical methods to differentiate chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control by breath analysis using ion mobility spectrometry“. In: *International Journal for Ion Mobility Spectrometry* (2011).

Andere Veröffentlichungen

[TB1/310] R. Cumeras, K. Rupp, T. Schneider, P. Favrod, S. Maddula und J. I. Baumbach. *Influence of Room Air and Operational Background Emissions on Breath Analysis Using MCC/IMS*. 20th Annual conference on Ion Mobility Spectrometry, Edinburgh, Scotland; 24.07.2011 - 29.07.2011. 2011.

[TB1/311] A. Egorov u. a. *Ressourcenbeschränkte Analyse von Ionenmobilitätsspektren mit dem Raspberry Pi*. Abschlussbericht der Projektgruppe 572 der Fakultät für Informatik. Technischer Bericht 5. TU Dortmund, Mai 2014.

3.3.3 Dokumentation weiterer Aktivitäten

Eine Projektgruppe (PG) ist eine Lehrveranstaltungsform der Fakultät für Informatik der TU Dortmund, die insgesamt 16 SWS verteilt auf zwei Semester umfasst und von 8–12 studentischen Teilnehmern gemeinsam belegt wird. In der PG 572 ("Ressourcenbeschränkte Analyse von Ionenmobilitätsspektren mit dem Raspberry Pi", kurz "BreathBerry") im Sommersemester 2013 und Wintersemester 2013/14 wurden Aspekte des Arbeitspakets 1 (Online-Algorithmen zur Peakdetektion) von elf Studierenden im Master-Studium untersucht, die von den aus der Ergänzungsausstattung geförderten Projektmitarbeitern Dominik Kopczynski und Marianna D'Addario betreut wurden. Die vorläufigen Ergebnisse daraus sind oben unter Arbeitspaket 1 beschrieben. Die PG hat einen Teil ihrer Ergebnisse auf der *German Conference on Bioinformatics* (GCB) 2013 in Göttingen mit einer Posterpräsentation vorgestellt. Dies wurde durch die Fakultät für Informatik mit Exkursionsmitteln finanziell gefördert. Der Abschlussbericht der PG 572 ist unter https://eldorado.tu-dortmund.de/handle/2003/33015 abrufbar.

Zwei Bachelorarbeiten (Robert Kirberich: Online Analyse von zweidimensionalen Spektrometriedaten; Henning Funke: Accelerating modeling of MCC/IMS data using GPGPU programming) sind direkt dem Projekt zuzuordnen. Im Rahmen des DAAD RISE Programms war Benjamin Yip aus Kanada acht Wochen lang zu Gast an der Fakultät für Informatik und hat die Arbeiten von Robert Kirberich weitergeführt.

Vom 9.-12. Juni 2013 fand in Saarbrücken-Wallerfangen das *Breath Analysis Summit 2013* unter Schirmherrschaft der IABR (International Association for Breath Research), der TU Dortmund (SFB 876) und des Universitätsklinikums des Saarlands mit 180 Teilnehmern statt.

3.4 Planung des Transferprojekts

Ziele

Ziel ist die Entwicklung ressourcenbewusster Algorithmen zur Analyse von Spektrometriedaten. Die Algorithmen werden für Messungen von Metaboliten in der Ausatemluft entwickelt und evaluiert, welche mittels Ionenmobilitätsspektrometrie (IMS) erzeugt wurden. In der ersten Phase des Teilprojekts TB1 wurden Methoden zur Vorverarbeitung von IMS-Daten entwickelt, insbesondere zur automatischen Peakextraktion aus den Rohdaten. Ziel der zweiten Phase ist die Bereitstellung von bett- und patientenseitig einsetzbaren und in Textilien/Kleidung integrierten Atemluftspektrometern für Risikogruppen und die Entwicklung von Gesamtanalyseprozessen, welche die Methoden der Vorverarbeitung und weitere statistische Lernverfahren auf den Daten in verschiedenen anwendungsspezifischen Szenarien ressourcenschonend kombinieren.

Erstes algorithmisches Ziel ist die optimale Klassifikation von Patientengruppen mit Hilfe der Spektrometriedaten. Die Datenbank DBreath2012 enthält unzählige Datensätze mit Messungen von IMS-Daten verschiedenster klinischer Anwendungen wie beispielsweise Entzündungen, Krebs oder Alzheimer. Zuerst werden Klassifikationsverfahren entwickelt und verglichen, die zwei oder mehr phänotypisch unterschiedliche Patientengruppen diskriminieren. Hierbei werden sowohl univariate Klassifizierer zur Suche von einzelnen relevanten Metaboliten als auch multivariate Klassifizierer zur Fehlerminimierung bei Vorhersagen betrachtet, basierend z.B. auf Random Forests oder SVMs. Kombinationen von Klassifizierern mit vorgeschalteten Methoden der Variablenselektion werden ebenfalls evaluiert.

Neben einzelnen Messungen am Patienten rücken Zeitreihen stärker in den Fokus, z.B. zur Erkennung einer Sepsis. Mit Glättungsverfahren sollen zunächst typische Verläufe für verschiedene Krankheitsbedingungen bzw. Patientenkohorten geschätzt werden. In einem zweiten Schritt werden dann multivariate Verfahren zur Detektion unterschiedlicher Verläufe zwischen den Gruppen erarbeitet. Ein Beispiel ist eine beginnende Sepsis. Die frühzeitige Erkennung der Sepsis ist entscheidend für den individuellen Therapieerfolg. Im Rahmen von Rattenmodellen an intubierten Tieren wurden an der Universitätsklinik Homburg Zeitreihen der Ausatemluftkonzentration bestimmter Metabolite gemessen, die zur Modellierung der Entwicklung polymikrobieller und monomikrobieller Sepsis geeignet sind. Anhand dieser Daten aus den Experimenten an Ratten werden exemplarisch statistische Methoden entwickelt, die den Umschlagpunkt (*change point*) in Verläufen von Metabolitenkonzentrationen frühestmöglich identifizieren.

Im letzten Schritt werden dann sowohl für die Klassifikation als auch für die Zeitreihenanalyse algorithmische Gesamtanalyseprozesse erarbeitet, die die Algorithmen für die Vorverarbeitung von Spektrometriedaten und die statistischen Verfahren für die Klassifikation oder für die Erkennung eines Umschlagpunkts optimal miteinander verbinden. Diese Prozesse werden an verschiedenen konkreten Beispielen entwickelt und evaluiert.

Darstellung von Kriterien, mit denen der Erfolg des Transferprojektes bewertet werden kann:

- Können wir einen Algorithmus entwickeln, der außer dem aktuellen Driftzeitspektrum nur fünf vergangene speichert, ein Chromatogramm parametrisch in Peaks zerlegt und dabei höchstens um 5% von der genauesten Parametrisierung abweicht, die bei Vorlage des Gesamtchromatogramms erreichbar ist?
- Können wir unsere Verfahren auf einem Miniatur-IMS-Gerät mit wenig Speicher und ohne Laptopanschluss ausführen?
- Können wir mit den genannten Verfahren und nachgeschalteter Datenanalyse (Peak-Alignment, Klassifikation, Zeitreihenanalyse) medizinische diagnostische Entscheidungen besser oder schneller als nach dem aktuellen Goldstandard treffen?

Arbeitsprogramm

Arbeitspaket 1. Technische Weiterentwicklung des IMS: Zielstellung ist die Bereitstellung von bett- und patientenseitig einsetzbaren und in Textilien/Kleidung integrierter Atemluftspektrometer für Risikogruppen. Hierfür werden bettseitig einsetzbare biomedizinische Spektrometer inklusive zugehöriger elektronischer Schaltkreise entwickelt. Diese ermöglichen die Messung flüchtiger Verbindungen im Atem, auf der Haut oder in Körperflüssigkeiten. Die miniaturisierten Spektrometerlösungen werden in Textilien/Kleidung integriert.

Ausgehend von der Miniaturisierung der Hardware des IMS wird es zwei Entwicklungslinien geben. Die erste Linie bezieht sich auf die frühzeitige Erkennung einer Sepsis. Die Diagnostik auf Grund der Ausatemluft von Patienten soll dabei einen frühzeitigeren Übergang zu einer gezielten Antibiotika-Strategie ermöglichen.

Die zweite Entwicklungslinie bezieht sich auf die Detektion von Infektionen. Hierfür wird eine situationsbezogene Sofortanalytik von Metaboliten am Körper benötigt, bei der die Bedarfsermittlung und die Erfolgskontrolle als individueller Regelkreis erfolgen.

Hierfür sind folgende zwei Anwendungsszenarien geplant:

- Bettseitiger Einsatz und minimalretardierte Signal-Verlaufs-Analyse durch Verfolgen weniger Signale im OP unter Einbettung in ein Beatmungssystem und Einschluss weiterer Sensoren wie z.B. EEG-Messungen: Auswertung und Vorhersage von Zeitverhalten (Zeitreihen), Ziehen von Rückschlüssen in Bezug auf laufende Aktion (Verbindung von Diagnose und Therapie), Handlungsvorschläge für die medizinische Behandlung.

- Patientenseitiger Einsatz mit Schwerpunkt Diagnostik bei atemluftvermittelten Krankheiten, Entzündungen, Krebs und Infektionen.

Dabei sollen für die diagnostischen und therapeutischen Entscheidungen nicht nur einzelne, singuläre Messungen der Konzentrationen der in der Ausatemluft nachweisbaren Stoffwechselprodukte verwendet werden, im Sinne einer Biomarkerdetektion, sondern multivariate Messungen in Zeitreihen bei einem Patienten betrachtet werden. Auf diese Weise sollen äußere Einflüsse minimiert und die Reaktion des individuellen Patienten personalisiert in die Entscheidungsvorschläge einbezogen werden. Die Integration erfolgt zunächst am Körper von außen und danach am Körper durch Integration in Kleidung. Für Patienten erfolgt die Integration in Kleidung oder in patientennahe Funktionstextilien.

Für eine Integration am Körper werden zunächst die Zuleitungen sowie der Spektrometerkörper (Ionisationsraum, Driftraum) in textile Gewebe eingebracht, inklusive der zugehörigen elektronischen Schaltungen und der Bedienfelder (Human-Machine-Interface). Die Energieversorgung wird extern am Gürtel realisiert. Hierfür werden sowohl die geometrischen Abmessungen der Spektrometerhardware verringert als auch die Integrationsdichte der Schaltungen im Rahmen biegsamer Leiterplatten erhöht. Periphere Elemente der Probenahme wie Pumpen und Gasreinigungskreisläufe (Driftgas) werden zunächst separat in Textilien eingebracht und danach in einer Schultereinheit integriert.

Arbeitspaket 2. Klassifikationsalgorithmen für IMS-Daten: In diesem Arbeitspaket sollen Klassifikationsalgorithmen entwickelt werden, die an die speziellen Eigenschaften von IMS-Daten angepasst sind. Da in dem Projekt eine sehr große Zahl an Datensätzen zur Verfügung steht, können generelle Eigenschaften von IMS-Daten untersucht werden, bevor anwendungsspezifische Klassifizierer konstruiert werden. Das bedeutet, dass in einem ersten Schritt univariate und multivariate Analysen an Daten aus der Datenbank erstellt werden, um allgemeingültige Verteilungseigenschaften und Korrelationen zwischen Metaboliten zu identifizieren.

Eine Erkenntnis aus diesen Voranalysen wird die Korrelationsstruktur der Messungen für die Metabolite sein. Neben der klassischen (unbedingten) Pearson-Korrelation werden auch partielle Korrelationen untersucht, die die Abhängigkeit von Variablenpaaren nach Bedingen auf jeweils alle anderen Variablen messen. Zusätzlich werden nicht-lineare Korrelationsmaße verglichen und ausgewertet, die nützlich sein können, wenn die entsprechenden IMS-Messungen nicht approximativ normalverteilt sind. Die Kenntnis der Korrelationsstruktur kann im Sinne einer korrelationsbasierten Variablenselektion [294] für Klassifikationsalgorithmen hilfreich sein.

Danach werden Klassifikatoren konstruiert, wobei das gesamte existierende Methodenspektrum berücksichtigt wird, von regularisierter logistischer Regression und linearer Diskriminanzanalyse über k-Nearest-Neighbor-Verfahren, Support Vektor Maschinen und Klassifikationsbäume bis hin zu Ensemble-Verfahren wie Random Forests oder Boosting-Ansätzen. Hinzu kommen Filtermethoden für die Variablenselektion. Dies bedeutet, dass problemspezifisch jeweils eine extrem große Anzahl an Algorithmen zur Auswahl steht. Zudem muss für eine faire Modellvalidierung jeweils eine Überanpassung (*overfitting*) vermieden werden. Üblicherweise geschieht dies mit Hilfe von Methoden der Kreuzvalidierung. Oft werden dann zur Reduktion der Varianz der geschätzten Vorhersagegüte zusätzlich verschiedene Aufteilungen von Trainings- und Testmenge evaluiert.

Hierfür wird in einem ersten Schritt das Framework aus dem R-Paket `BatchExperiments` [A3/153] verwendet, das in Teilprojekt A3 in der ersten Phase entwickelt und implementiert wurde. Weiterhin werden in den ersten beiden Jahren der Phase II des SFBs in Teilprojekt A3 Methoden der automatisierten Modellwahl entwickelt. Solche Algorithmen werden dann in Teilprojekt TB1 für IMS-Daten verwendet werden. Schließlich sollen auch andere Methoden des *Meta-Learning* in Betracht gezogen werden, um Zusammenhänge zwischen Charakteristiken von IMS-Datensätzen und Charakteristiken von Lernverfahren mit guter Prognosegüte zu modellieren.

Ein erstes konkretes Beispiel für die Klassifikation anhand von IMS-Daten wird die Unterscheidung von Ratten mit und ohne induzierter mikrobieller Sepsis auf Basis der Ausatemluftkonzentration sein. Die Messungen an Ratten werden dabei ausdrücklich nicht für die Arbeiten im Teilprojekt TB1 des SFB durchgeführt, sondern liegen bereits vor oder werden im Rahmen anderer Projekte z.B. an der Universitätsklinik Homburg generiert, für die die jeweils notwendigen Voten von Ethikkommissionen vorliegen.

Initiale Untersuchungen im Rahmen einer Bachelorarbeit an der Fakultät Statistik der TU Dortmund verwenden die lineare Diskriminanzanalyse und ein k-Nearest-Neighbor-Verfahren, wobei eine einfache Variablenselektion durchgeführt wird, indem lediglich ein t-Test pro Variable als Filter angewendet wird. Eine geeignete Vornormalisierung der multivariaten Daten muss ebenfalls noch erarbeitet werden, sowohl beim Vergleich von Messungen an verschiedenen Individuen als auch in Bezug auf anwendungsspezifische Normalbedingungen. Ferner wären robuste Tests für den Vergleich zweier Gruppen eine sinnvolle Alternative [291].

Schließlich soll bei Vorliegen eines Klassifizierers mit hoher Güte im Hinblick auf das biologische Verständnis der ausgewählten Metabolite noch der Nutzen einer differenziellen Netzwerkanalyse geprüft werden. Ein Netzwerk ist dabei ein Graph, bei dem die Knoten die Metabolite repräsentieren und die Kanten die partiellen Korrelationen zwischen den Metaboliten. Es werden also für zwei zu unterscheidende Gruppen jeweils erst getrennt Netzwerke geschätzt und dann an Hand von Maßzahlen verglichen, welche von den Netzwerken abgeleitet werden. Beispiele für solche Maßzahlen sind die maximale Differenz der Korrelationen über alle Kanten des Netzwerks hinweg oder die durchschnittliche Differenz der Korrelation im gesamten Graph. Die Signifikanz von Unterschieden kann mit Permutationstests überprüft werden.

Im Sepsis-Beispiel bedeutet dies, dass unabhängig voneinander für Ratten mit und ohne induzierter Sepsis jeweils ein Netzwerk der Zusammenhänge der Metaboliten geschätzt wird. Die Unterschiede der Zusammenhänge bieten dann die Möglichkeit einer funktionellen Interpretation des Klassifizierers.

Zu beachten ist ferner, dass für die Bewertung der Klassifizierer verschiedene Maßzahlen in Betracht gezogen werden müssen. Neben den etablierten Maßen Klassifikationsfehler (bzw. *accuracy*), Sensitivität und Spezifität sind im klinisch-diagnostischen Bereich der PPV (positiv prädiktiver Wert) und der NPV (negativ prädiktiver Wert) von zentraler Bedeutung, da sie die Wahrscheinlichkeit der Richtigkeit einer positiven oder negativen Diagnose bewerten. Abhängig von der Zielsetzung kann dann der PPV oder der NPV jeweils höher gewichtet werden.

Arbeitspaket 3. Zeitreihenanalyse und Früherkennung von Krankheiten: Zeitreihen von IMS-Messungen können zur Online-Überwachung von Patienten und zur frühzeitigen Erkennung von kritischen Ereignissen verwendet werden. Für einen Patienten im Krankenhaus etwa ist die frühzeitige Erkennung einer beginnenden Sepsis entscheidend für den Therapieerfolg. Die Überlebenschance sinkt dabei abhängig vom Erkennungszeitpunkt innerhalb der ersten zwölf Stunden ungefähr von 80% auf 20%. Als erstes Beispiel werden Daten der Universitätsklinik Homburg aus Experimenten an Ratten untersucht, bei denen unter kontrollierten Bedingungen eine Sepsis ausgelöst wurde. Von Interesse ist dann der Vergleich mit einer gleich behandelten Referenzgruppe ohne induzierte Sepsis. Abbildung 3.2 zeigt erste Ergebnisse aus der Bachelorarbeit.

Die Abbildung zeigt zunächst die Verläufe von Konzentrationen eines ausgewählten Metaboliten für acht behandelte Ratten (mit induzierter Sepsis, rote Linien) und für acht unbehandelte Ratten (ohne induzierte Sepsis, blaue Linien). Insgesamt liegen für diese Analyse 37 Messungen vor, jeweils im Abstand von 20 Minuten (0 Minuten bis 720 Minuten) mit 38 gleichzeitig bestimmten Metaboliten-Konzentrationen. Einige Linien enden vorzeitig (vor Messpunkt 37), wenn die entsprechenden Ratten verstorben sind.

Ab etwa 300 Minuten sinkt die Konzentration für die behandelten Tiere, mit Ausnahme eines unbehandelten Tieres, wobei Ausreißer nach unten klar zu erkennen sind. Die zweite Grafik zeigt, dass ein Zweistichprobentest, unabhängig ausgewertet für jeden Zeitpunkt, ab etwa 300 Minuten zu einem sehr kleinen und somit signifikanten p-Wert führt. Eigentliches Ziel ist aber die Vorhersage der Sepsis, konkret die Prognose einer Klasse. Der ausgewählte Metabolit würde eine gute Prognose liefern, wurde aber datengetrieben ausgewählt, was zu Überoptimismus führt. Die anderen beiden Grafiken zeigen, dass die korrekt kreuzvalidierte Fehlerrate einer multivariaten Klassifikation beruhend auf linearer Diskriminanzanalyse und 1-Nearest-Neighbor-Klassifikation unbefriedigende Ergebnisse liefert.

Das weitere Vorgehen in Arbeitspaket AP3 ist deshalb in drei Schritte unterteilt. Zunächst sollen Glättungsverfahren auf die Zeitreihen angewendet werden, um die typischen Verläufe für verschiedene Krankheitsbedingungen bzw. Patientenkohorten von Rauschen zu befreien. Dies geschieht mit Hilfe eines gleitenden Mittelwerts oder moderner und robuster Alternativen wie z.B. adaptiven wiederholten Medianfiltern [297, 296, 292]. Dann werden die multivariaten Verfahren aus Arbeitspaket AP2 angewendet, sowohl ohne als auch mit Variablenselektion. Schließlich wird versucht, den funktionellen Verlauf der Zeitreihen besser zu charakterisieren und geeignete Maßzahlen abzuleiten, die den Umschlagpunkt in Verläufen der Metaboliten-Konzentration frühestmöglich identifizieren sollen.

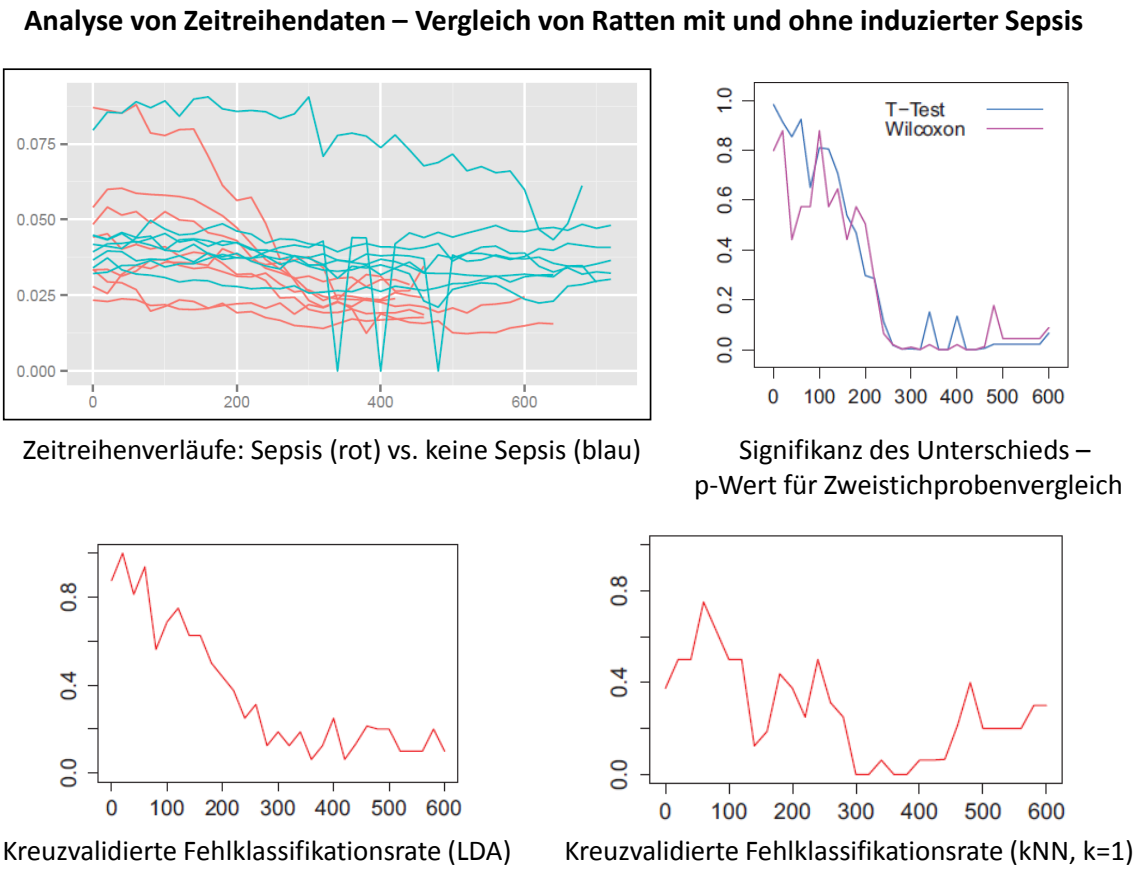


Abbildung 3.2.: Auswertungen von Zeitreihendaten der Universitätsklinik Homburg von IMS-Messungen aus kontrollierten Rattenexperimenten mit Experimentalgruppe (induzierte Sepsis) und Kontrollgruppe.

Die Abbildungen 3.3 und 3.4 zeigen erste Ergebnisse aus einer gerade abgeschlossenen Masterarbeit (Salome Horsch), bei der bereits einfache Glättungsmethoden offline auf die Zeitreihenmessungen angewendet wurden. Die Grafik in Abbildung 3.3 enthält in den Zeilen Zeitpunkte, absteigend geordnet, und in den Spalten einzelne Metabolit-Messungen. Der Farbcode gibt an, ob die Konzentration des entsprechenden Metaboliten zu dem entsprechenden Zeitpunkt in den beiden Klassen im Mittel statistisch signifikant unterschiedlich ist. Auf die ersten beiden Metabolite beispielsweise trifft dies ab einem mittleren Zeitpunkt zu, auf die nächsten beiden zu keinem Zeitpunkt. Die Grafik unterstreicht die Bedeutung der Merkmalsselektion, die von dem hier vorliegenden univariaten Fall noch auf den multivariaten erweitert werden muss. Zusätzlich müssen Online-Verfahren zur Wechselpunkterkennung eingesetzt werden, die keine zukünftigen Zeitpunkte in die Glättung der Zeitreihen miteinbeziehen.

Die Grafik in Abbildung 3.4 repräsentiert in den Zeilen einzelne Ratten und in den Spalten chronologisch geordnete Zeitpunkte. Der Farbcode gibt hier an, ob eine Ratte zu dem entsprechenden Zeitpunkt richtig klassifiziert wurde (induzierte Sepsis oder nicht). Man sieht, dass zwar ab einem mittleren Zeitpunkt passable Vorhersageraten erzielt werden, aber beispielsweise die ersten beiden Ratten fast immer falsch klassifiziert werden. Eine erste Analyse der konkreten Werte zeigte, dass hier voraussichtlich zusätzliche Normierungsschritte in der Vorverarbeitung benötigt werden.

In einem weiteren Schritt soll dann das Spektrum der Zeitreihenanalyse noch weiter ausgeschöpft werden, indem Methoden direkt für Diskriminanzanalysen für multivariate Zeitreihen entwickelt werden, anstatt die Glättung und die Klassifikation in zwei Schritte zu unterteilen.

Arbeitspaket 4. Entwicklung und Bewertung von Gesamtanalyseprozessen: In der ersten Phase sind in Teilprojekt TB1 Vorverarbeitungsschritte für Spektroskopie-Daten entwickelt worden, insbesondere zur automatischen Peakerkennung und Peakextraktion. Diese beruhen auf der Anpassung verschiedener Parameter,

Merkmalauswahl für Zeitreihendaten – Vergleich von Ratten mit und ohne induzierter Sepsis

Metaboliten mit univariat signifikanten Differenzen zwischen Ratten mit und ohne induzierter Sepsis (**grau: nie signifikant**, **gelb: selten signifikant**, **rot: sehr oft signifikant**)

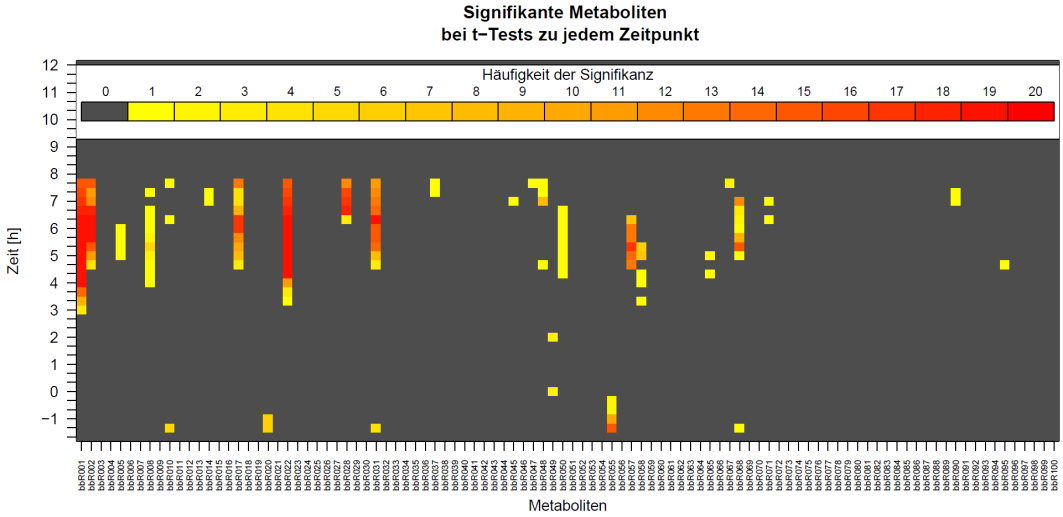


Abbildung 3.3.: Merkmalsauswahl für Zeitreihendaten von IMS-Messungen aus kontrollierten Rattenexperimenten mit Experimentalgruppe (induzierte Sepsis) und Kontrollgruppe.

die abhängig von der Anwendung bzw. Fragestellung und von den Umweltbedingungen bei der Messung gewählt werden müssen. Die Einstellung der Parameter kann das Gesamtergebnis eines anschließenden statistischen Lernverfahrens entscheidend beeinflussen. Die Implementierung eines Gesamtanalyseprozesses von Vorverarbeitung und Lernverfahren auf einem eingebetteten System (Arbeitspaket AP5) erfordert die Exploration eines noch größeren Parameterraums als in Arbeitspaket AP3. Daher soll auch hierfür eine für IMS-Daten optimierte automatisierte Modellwahl erfolgen, auch mit Hilfe der Methoden aus Teilprojekt A3 des SFBs und in Zusammenarbeit mit Teilprojekt A1.

Arbeitspaket 5. Implementierung auf eingebettetem System: Dieses Arbeitspaket baut auf den Ergebnissen aus den Arbeitspaketen AP2-AP4 auf, indem die Verfahren auf einem eingebetteten System implementiert werden. Neben der patientenseitigen Realisierung der Hardware in Kleidung in einer Weise, sodass die Sensorik vom Patienten im Idealfall überhaupt nicht wahrgenommen wird, gilt gleiches auch für die Informationsverarbeitung, die so in textile Strukturen eingebettet wird, dass sie wenigstens nicht störend wahrgenommen wird. Dies betrifft hardwareseitig die Integration der Probenahme in textile Strukturen, die Einbettung von Pumpe und Driftröhre in Schulterstücke, die des Betriebes (inkl. der Steuerung der Probenahme) und die der Auswertung (GPS inkl. externer Signalweiterleitung). Hierbei sollen zuerst die gesamten IMS-Chromatogramme, danach nur die Peaklisten und schließlich nur die problembezogen notwendigen Signalparameter übermittelt werden. Alle steuerungsseitigen Elemente für die MCC/IMS werden inklusive der Signalleitungen in textile Strukturen eingebettet sein.

Hierfür werden elektronische Komponenten (Pumpen- und Probenahmesteuerung, Ventilsteuerung, Probenschleifensteuerung, Probengabe, Reinigungskreislaufsteuerung) und das Human-Machine-Interface definiert und die Signalvorverarbeitung (automatische Probegabeerkennung, Start Probenahme, Reinigungskreisläufe, Signalsubstraktion (Einatem- bzw. Raumluft), Signalbewertung (ready to run) und Signalweiterleitung (Alarmierungsfunktion) sowie Positionierung (GPS) und Verbindung zur Intensivmedizin (Infektion, Sepsis) und zum Hausarzt realisiert.

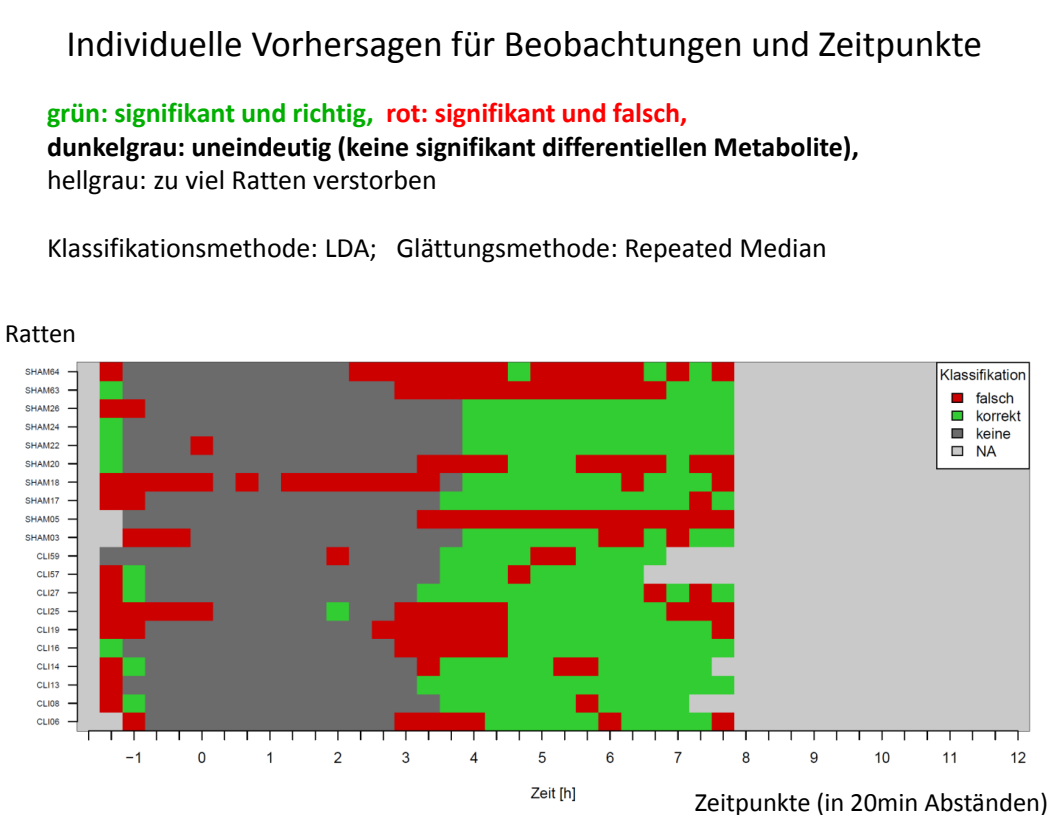


Abbildung 3.4.: Klassifikationsergebnisse für Zeitreihendaten von IMS-Messungen aus kontrollierten Rattenexperimenten mit Experimentalgruppe (induzierte Sepsis) und Kontrollgruppe.

Zeitplan

	2015				2016				2017				2018			
Arbeitspaket / Quartal	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Technische Weiterentwicklung des IMS:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2. Klassifikationsalgorithmen für IMS-Daten:	■	■	■	■												
3. Zeitreihenanalyse und Früherkennung von Krankheiten:					■	■	■	■	■	■						
4. Entwicklung und Bewertung von Gesamtanalyseprozessen:											■	■	■	■	■	■
5. Implementierung auf eingebettetem System:					■	■	■	■	■	■	■	■				

3.5 Stellung innerhalb des Sonderforschungsbereichs

In Teilprojekt TB1 werden die Datenanalyse und ein eingebettetes System unmittelbar miteinander verknüpft. In Phase I lag der Fokus auf der Entwicklung neuer Algorithmen für die Vorverarbeitung von Spektrometriedaten unter Ressourcenbeschränkungen. Es wurden Methoden zur automatischen Peakerkennung aus IMS-Messungen entwickelt, die online angewendet werden können. In Phase II wird nun ein Gesamtanalyseprozess entwickelt, der die technische Atemluftmessung, die automatische Vorverarbeitung der Rohdaten

und die anwendungsspezifische Entscheidungsfindung bzgl. einer Diagnose oder Therapiewahl miteinander verbindet.

Eine enge Zusammenarbeit ist in Phase II mit Teilprojekt A3 geplant, in dem Methoden zur automatischen Modellwahl entwickelt werden. Ein typischer Datenanalyseprozess in TB1 beinhaltet viele Entscheidungen, von der Wahl von Parametern bei der Vorverarbeitung über die Merkmalsauswahl bis hin zur konkreten Auswahl des Lernverfahrens bei Klassifikationsaufgaben, sodass TB1 von den neuen Methoden aus A3 profitieren kann. Ebenfalls ist eine Kooperation mit Teilprojekt A1 geplant, da die Konstruktion von Gesamtanalyseprozessen dort etabliert ist und in diesem Rahmen auch ein RapidMiner Plug-in für Zeitreihenverfahren existiert.

Mit den Teilprojekten C1 und C4 wird es einen Austausch bzgl. der Variablenselektionsmethoden geben, TB1 kann von den Erkenntnissen in diesen Teilprojekten profitieren. Insbesondere wird in Teilprojekt TB1 in Phase II auf Korrelationsaspekte zwischen den Messungen einzelner Metabolite eingegangen werden. Die Berücksichtigung von Korrelationen steht bei hochdimensionalen Daten wie im Teilprojekt C1 auf Grund der zu hohen Anzahl an Variablen normalerweise eher im Hintergrund.

Die Software für IMS-Geräte ist bereits im Antrag zu Phase I als Anwendungsbeispiel für die Arbeiten an der Plattform in Teilprojekt A4 benannt worden. Für Phase II ist nun geplant, ein Energiemodell für ein IMS-Gerät zu entwickeln, sodass danach ein geringerer Energieverbrauch als Kriterium für die Softwareentwicklung verwendet werden kann.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

YeastScent
(Baumbach,Förderkennzeichen BMBF 031 A301 F) (Förderzeitraum: 2014 - 2017)

Entwicklung einer Prozesssteuerung für Fermenter zur Hefeproduktion auf der Grundlage der Steuerung metabolischer Prozesse.

SpiCos
(Baumbach,Förderkennzeichen BMBF bortnikv-043) (Förderzeitraum: 2014 - 2017)

Entwicklung einer nichtradioaktiven Ionisationsquelle für Ionenbeweglichkeitsspektrometer.

Gen-Verteilungen
(Rahnenführer,Förderkennzeichen DFG RA 870/4-1) (Förderzeitraum: 2011 - 2016)

Das Projekt „Verbesserte prognostische Signaturen aus Microarray-Studien durch Auswahl von Genen mit charakteristischen Verteilungen“ erstellt Vorhersagen für Überlebenszeiten aus hochdimensionalen genetischen Daten. Der Fokus liegt auf der Vorauswahl von Genen anhand ihrer charakteristischen Verteilungen.

Differentielle Netzwerke
(Rahnenführer,Förderkennzeichen DFG RA 870/5-1) (Förderzeitraum: 2011 - 2015)

Das Projekt „Klinische Prognosen auf Grundlage von genregulatorischen Netzwerken“ beschäftigt sich mit der Schätzung differentieller Netzwerke aus Expressionsdaten von vorausgewählten Genen.

eTop: LivSys
(Rahnenführer,Förderkennzeichen BMBF 031A270B) (Förderzeitraum: 2013 - 2015)

Das Projekt behandelt hochdimensionale Expressionsdaten für die Reaktion von Leberzellen auf toxikologische Stoffe. Die Modellwahl ist in diesem Projekt durch die Einbeziehungen von biologischen Annotationen getrieben.

eTop: SysDT
(Rahnenführer,Förderkennzeichen BMBF 031A272E) (Förderzeitraum: 2014 - 2015)

In dem Projekt werden mit statistischen und systembiologischen Methoden Zeitreihenmessungen von Stammzellen analysiert. Die Zeitreihen bestehen aus sehr wenigen Zeitpunkten und für die Expressionsdaten stehen biologische Annotationen zur Verfügung, die es in dieser Form für die Metabolite nicht gibt. Das Projekt hat daher mit Teilprojekt TB1 nur den temporalen Aspekt gemeinsam.

3.7 Ausstattung des Transferprojekt

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Leistungen des Anwendungspartners für die neue Förderperiode

	2015	2016	2017	2018
Personal (in Wochenstunden)	83.6	83.6	83.6	83.6
Sachausgaben	10.000	10.000	10.000	10.000
Investitionen	0	0	0	0

3.7.3 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	1	58.800	1	58.800	1	58.800	1	58.800
Summe:	—	58.800	—	58.800	—	58.800	—	58.800
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	58.800		58.800		58.800		58.800	

3.7.4 Personal für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (B1)							
Wissen-schaftliches Personal	1	Rahnenführer, Jörg, Prof. Dr. Prof	Statistische Methoden	Statistische Methoden in der Genetik und Chemometrie , Fak. Statistik, TU Do	4,0	—	Grund-ausstattung
Anwendungspartner (B1)							
Wissen-schaftliches Personal	2	Baumbach, Jörg-Ingo, Prof. Dr. Prof	Metabolo-mik/Analytik	B&S Analytik GmbH, Dortmund	4,0	—	Grund-ausstattung
	3	Metrailler, Laurence, B.Sc. WiMi	Analytik	B&S Analytik GmbH, Dortmund	19,9	—	Grund-ausstattung
	4	Hauschild, Ann-Cristin, M.Sc. WiMi	Analytik	Hochschule Reutlingen	19,9	—	Grund-ausstattung
	5	N.N. , WiMi	Analytik	Hochschule Reutlingen	19,9	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	6	Kaiser, Peter, Techn. Ang.	—	B&S Analytik GmbH	19,9	—	Grund-ausstattung
Beantragte Personalmittel (B1)							
Wissen-schaftliches Personal	7	Horsch, Salome, M.Sc. WiMi	Statistische Methoden	Statistische Methoden in der Genetik und Chemometrie , Fak. Statistik, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Rahnenführer, Jörg**
   Projektleitung. Schwerpunkt Statistik. Mitarbeit an den Arbeitspaketen AP2 , AP3, AP4.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (Anwendungspartner):

2. **Baumbach, Jörg-Ingo**
   Projektleitung. Schwerpunkt Metabolomik, Klinische Diagnostik. Mitarbeit an AP1, AP5.
3. **Metrailler, Laurence**
   Mitarbeit an AP1 und AP5, Hard-Soft-Firmware.
4. **Hauschild, Ann-Cristin**
   Mitarbeit an AP1 und AP5, Dimensionsreduktion, Clustering.
5. **N.N.,**
   Die Mitarbeit an AP1 und AP5, Datengewinnung und Referenzierung.
6. **Kaiser, Peter**
   Mitarbeit an AP1 und AP5, Hard-Soft-Firmware.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

7. **Horsch, Salome**
   Statistischen Analysen in den Arbeitspaketen 2, 3, 4 und Schnittstelle zur Datenbank und zu den Analysen zur Vorverarbeitung.

3.7.5 Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	2.000	2.000	2.000	2.000
Anwendungspartner	10.000	10.000	10.000	10.000
Summe der vorhandenen Mittel	12.000	12.000	12.000	12.000
beantragte Mittel	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt B2

3.1.1 Titel:

Ressourcenoptimierte Echtzeitanalyse stark Artefakt-behafteter Bildsequenzen zur Detektion von Nanoobjekten

3.1.2 Fachklassifizierung:

Visual Computing, Eingebettete Systeme, Biomedizinische Analytik

3.1.3 Projektleitung

Prof. Dr. Chen, Jian-Jia, 09.05.1978

LS 12, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 16
44227 Dortmund

Tel.: 0231-755-6111
Fax : 0231-755-6116
E-Mail: `jia.chen@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Dr. Hergenröder, Roland, 06.08.1961

Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.
Bunsen-Kirchhoff-Straße 11
44227 Dortmund

Tel.: 0231-1392-178
Fax : 0231-1392-120
E-Mail: `roland.hergenroeder@isas.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Müller, Heinrich, 30.07.1953

LS 7, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 16
44227 Dortmund

Tel.: 0231-755-6324
Fax : 0231-755-6321
E-Mail: `heinrich.mueller@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Das Projekt befasst sich mit echtzeitkonformen Bildanalyseverfahren zur Detektion von spezifischen Nanoobjekten in stark artefaktbehafteten Bildsequenzen bei limitierten Ressourcen, insbesondere der Energie. Das wesentliche Ziel ist eine Optimierung des Trade-off zwischen Hardwareeffizienz, Zeiteffizienz und Detektionsqualität geeigneter Bild- und Datenanalyseverfahren. Als anwendungsrelevante Evaluationsstudie dient die neuartige PAMONO-Technik (engl. *Plasmon assisted Microscopy of Nano Objects*) zur bildbasierten Erkennung von biologischen Viren (allg. Nanoobjekten).

Das Projekt hatte bisher Verfahren zur Echtzeitanalyse von Bildsequenzen aus einem Einzelsensor zum Gegenstand, die auf einer mit dem Sensor verknüpften Berechnungsplattform ausgeführt werden. Dabei wurde angenommen, dass nur nach einem einzigen Typ von Nanoobjekt gesucht wird, welcher nach dem Eintritt ortsständig ist und sich die Eigenschaften der Bildfolgen bei verschiedenen Messungen nicht wesentlich unterscheiden. In der zweiten Förderperiode sollen mehrere verschiedene zeitlich und örtlich voneinander unabhängige Sensoren betrachtet werden, die über drahtlose Datenkommunikation untereinander und zu Servern verfügen. Zudem sollen mehrere Nanoobjekttypen gleichzeitig sowie nicht ortsständige Nanoobjekte betrachtet werden, welche sich aus Weiterentwicklungen des PAMONO-Sensors innerhalb des Projektes ergeben.

Mehrere Nanoobjekttypen und nicht ortsständige Nanoobjekte erfordern die Entwicklung neuer algorithmischer Bildanalyselösungen – insbesondere für die Verfolgung kleiner, bewegter Objekte in verrauschten Umgebungen. Die Ausführung vieler Instanzen der Bildanalyseverfahren in einer verteilten Sensorumgebung eröffnet zudem die Möglichkeit des Sammelns von Daten über deren Ausführungsverhalten. Es sollten Verfahren zur automatischen Fehlfunktionserkennung und Kalibrierung der individuellen Sensoren und Instanzen auf Grundlage der Analyse entsprechender Daten entwickelt werden, um auf diese Weise eine funktionelle Leistungssteigerung der Bildanalyse zu erzielen.

Zum Einsatz der verteilten Systemumgebung sollen Paradigmen auf Algorithmen- und Ausführungsebene ausgearbeitet werden. Ein ausführungskontextsensitives Algorithmen- und Programmierparadigma soll die Abbildung von Algorithmen und Programmen auf verteilte Berechnungsplattformen unter Berücksichtigung des Ressourcenverbrauchs effektiver und effizienter machen, indem für den Algorithmenentwurf und die Programmierung entsprechende Ausdrucksmöglichkeiten geschaffen werden. Für die Ausführungsebene soll einem kollaborativem Ausführungsparadigma gefolgt werden, in dessen Mittelpunkt die Realisierung des Konzepts des Offloading für Berechnungen mit Echtzeitanforderung auf ressourcenbeschränkten, verteilten, eingebetteten Systemen steht.

Zum verteilten, mobilen Einsatz des PAMONO-Sensors sollen Schritte zu dessen Miniaturisierung unternommen werden. Ferner soll die PAMONO-Methode auf andere Anwendungsfelder, insbesondere die simultane Erkennung mehrerer Typen von Nanoojekten, die Detektion von Partikeln in Luft und die Detektion von nicht ortsständigen Nanoobjekten (z. B. Grippeviren) erweitert werden.

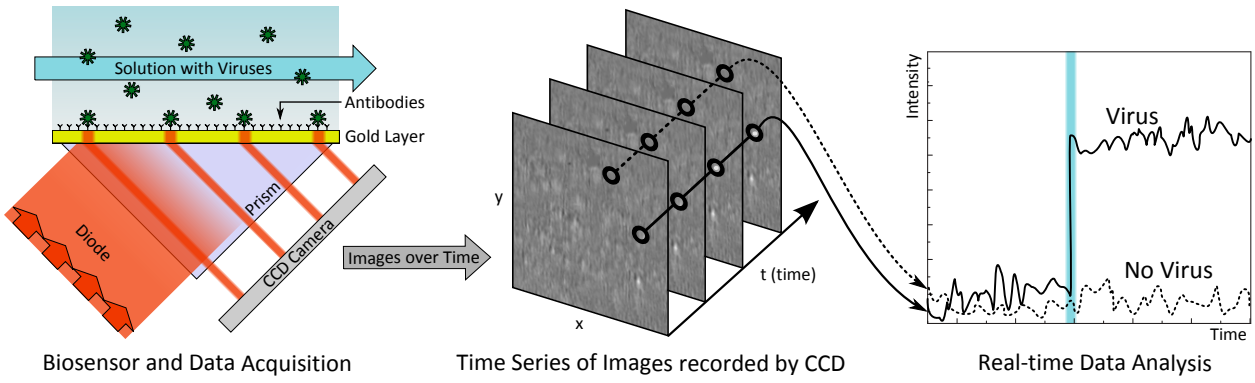


Abbildung 3.1.: Funktionsprinzip des PAMONO-Sensors

3.3 Bisherige Entwicklung des Teilprojekts

Projektleiter der ersten Förderperiode sind Prof. Dr. Peter Marwedel, Prof. Dr. Heinrich Müller und Dr. Alexander Zybin. Durch den Eintritt in den Ruhestand scheiden Herr Marwedel und Herr Zybin mit Ablauf der Förderperiode als Projektleiter aus. In der zweiten Förderperiode treten stattdessen Prof. Dr. Jian-Jia Chen und Dr. Roland Hergenröder mit in die Projektleitung ein. Herr Chen ist der Nachfolger von Herrn Marwedel an der Fakultät für Informatik und Herr Hergenröder gehört der gleichen Abteilung des ISAS wie Herr Zybin an.

3.3.1 Bericht und Stand der Forschung

Das globale Ziel des Projekts ist die Erkennung spezifischer Nanoobjekte in Bildfolgen, die durch mehrere erschwerende Randbedingungen gekennzeichnet sind. Gegenstand sind sensorerfasste, datenintensive Bildfolgen natürlicher Prozesse, die bedingt durch den erfassenden Sensor stark verrauscht sind und zu detektierende Ereignisse zeigen, die relativ zu dem durch die Bilder erfassten Ortsausschnitt klein sind und deren Eintrittszeit kurz ist. Trotzdem soll die Verarbeitung in Echtzeit geschehen und sollen die eingesetzten Sensor- und Berechnungsressourcen gering gehalten werden.

Als Anwendungsfallstudie dient die Automatisierung der bis dahin manuellen Bildsequenzauswertung der neuartigen „Plasmonen-unterstützten Mikroskopie von Nanoobjekten“ (PAMONO). Abb. 3.1, links, zeigt das Funktionsprinzip. Auf der Oberseite einer Goldschicht, die mit Antikörpern versehen ist, strömt Flüssigkeit ein, welche die zu detektierenden Nanoobjekte, z. B. Viren, enthält. Auf der Unterseite sind eine Laserdiode und eine Kamera angebracht. Die Kamera empfängt das von der Goldschicht reflektierte Licht und gibt das Reflexionsergebnis als Bildfolge wieder (Abb. 3.1, Mitte). Die auf der Oberfläche auftreffenden Nanoobjekte hinterlassen eine örtliche und zeitliche Veränderung in der Bildfolge. Für eine ausführliche Darstellung und Einordnung der PAMONO-Technik in den Stand der Forschung wird auf den Erstantrag verwiesen. Eine wesentliche Schwierigkeit ist, dass die örtliche Veränderung aufgrund des auftretenden Bildrauschens recht schwer von anderen Artefakten zu unterscheiden ist. Die technischen Voraussetzungen der PAMONO-Technik sollen mit dem Ziel verbessert werden, die Erkennungsleistung und die Ressourceneffizienz des Systems zusätzlich zu den optimierten Algorithmen zu steigern. Es soll erreicht werden, dass das PAMONO-Verfahren vor Ort einsetzbar wird und dort unmittelbar Ergebnisse liefert.

Die im Berichtszeitraum geleisteten Arbeiten zu dieser Zielsetzung wurden in vier Teilbereiche gegliedert:

A. Echtzeitfähige Analyse stark artefaktbehafteter Bildsequenzen (AP 1, AP 2, AP 3, vorrangig vom Lehrstuhl Informatik 7 (Müller) bearbeitet)

B. Plattformentwurf unter Optimierung des Ressourcenverbrauchs (AP 4, AP 5, vorrangig vom Lehrstuhl Informatik 12 (Marwedel) übernommen)

C. Entwicklung hochdifferenzierender Biosensoren auf Basis der Plasmonenunterstützten Mikroskopie zur Detektion von Nanoobjekten (AP 6, AP 7, vorrangig vom ISAS (Zybin) bearbeitet)

D. Optimierung des Trade-off zwischen Erkennungsleistung, Verarbeitungszeit und Ressourcenbedarf und Validierung der Resultate (AP 8, von allen gemeinsam bearbeitet)

Im Folgenden wird zu jedem der Teilbereiche eine Darstellung der bisherigen Entwicklung gegeben, an die sich Ausführungen zum Stand der Forschung und zu den neuen Problemstellungen anschließen.

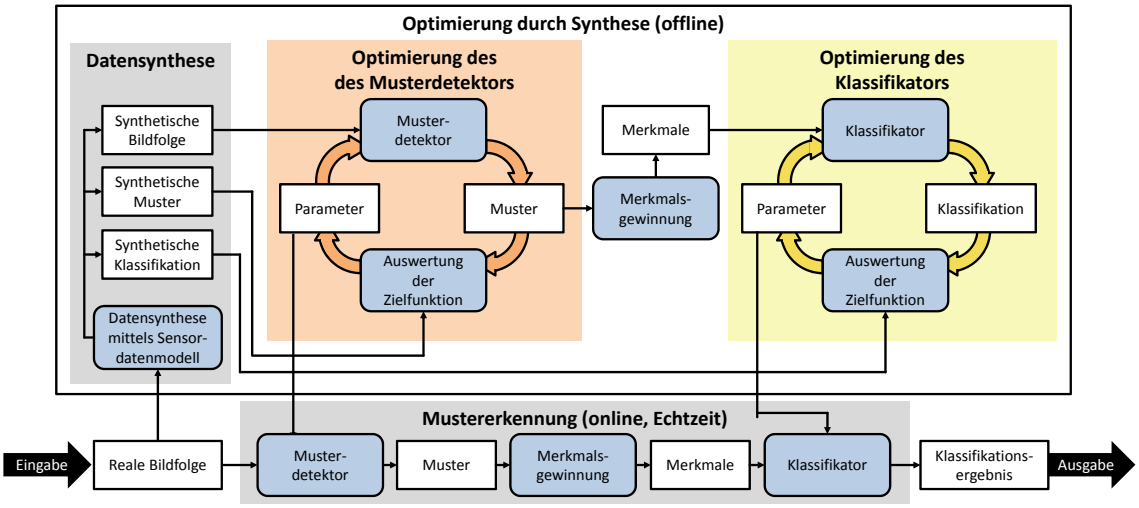


Abbildung 3.2.: Das Prinzip der Optimierung durch Synthese (im oberen Teil), angewandt auf die Verarbeitungskette der Erkennung von Nanoobjekten (im unteren Teil)

A. Echtzeitfähige Analyse stark artefaktbehafteter Bildsequenzen

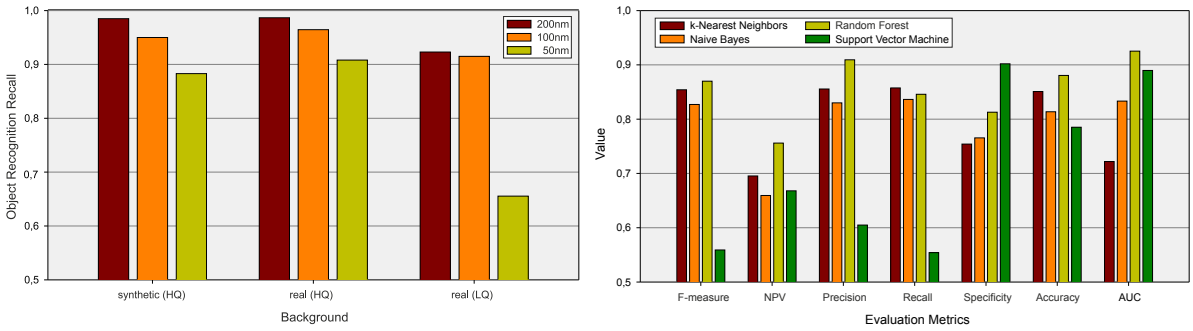
Die Erkennung der Nanoobjekte soll verschiedenen Zwecken dienen (siehe Erstantrag): Nachweis der Existenz von Nanoobjekten eines speziellen Typus (Aufgabe 1 - Existenz), Auffinden möglichst aller Nanoobjekte eines speziellen Typus (Aufgabe 2 - Auffinden eines Typus), Bestimmung der örtlichen und zeitlichen Konzentration der Nanoobjekte eines speziellen Typus (Aufgabe 3 - Konzentration) und Auffinden von Nanoobjekten endlich vieler vorgegebener, unterschiedlicher Typen (Aufgabe 4 - Auffinden mehrerer Typen). In der laufenden Förderperiode lag der Schwerpunkt auf den ersten drei Aufgaben. Die Bearbeitung der vierten Aufgabe war für die zweite Förderperiode vorgesehen und wird in diesem Antrag aufgegriffen.

Das wesentliche Resultat dieses Teilbereichs ist ein allgemeines Verfahren zur optimierten Analyse durch Synthese zur Erkennung von kleinen Objekten vor verrauschtem Hintergrund, siehe Abb. 3.2. Der Vorteil des Verfahrens liegt in einem geringen manuellen Aufwand für das überwachte Lernen.

Das Verfahren wurde schwerpunktmäßig für den PAMONO-Sensor umgesetzt. Gemäß der Abb. 3.2, unten, konzentriert es sich auf die Aufgabe 2 (Existenz). Die dazu verwendete Verarbeitungskette folgt der klassischen Vorgehensweise. Ihr erster Teil, der Musterdetektor, besteht aus einem Bildverbesserer, auf den ein Musterextrator folgt. Aufgrund der verrauschten Daten und zusätzlich festgestellter Artefakte wie Bildverschiebungen und Bildverzerrungen kommt der Bildverbesserung besondere Bedeutung zu. Dazu wurden Verfahren zur Hintergrundbereinigung mit statischem oder gleitendem Hintergrund modular mit ortsfrequenzbasierten Verfahren zur Rauschreduzierung (Denoising) kombiniert, wobei translations-invariante Wavelets [314] eingesetzt und Fuzzy-Regeln zur Rauschreduktion (2D+t) entwickelt wurden [B2/338]. Ferner wurde in Anlehnung an das *Video Block Matching* [315] untersucht, ob fehlerhafte Virussignale durch die nicht-lokale, kollaborative Filterung innerhalb der Rauschreduzierung gemindert werden können. Des Weiteren wurden Methoden der Variationsrechnung [320] zur Bildstabilisierung eingesetzt, welche die zeitliche Kohärenz der Ortskoordinaten berücksichtigen.

Die Musterextraktion beruht auf der Beobachtung, dass sich die Grauwerte der Pixel, an denen ein neues Nanoobjekt auftrifft, in typischer Weise über die Zeit in Form einer Sprungfunktion verändern (Abb. 3.1, rechts). Benachbarte Pixel mit dieser Eigenschaft werden zu Gebieten zusammengefasst und bilden so Muster. In Kooperation mit TP C3 wurde eine Masterarbeit [C3/312] durchgeführt, welche sich mit robusten statistischen Methoden zur Strukturbrucherkennung in Zeitreihen befasst.

Die Merkmalsgewinnung (s. Abb. 3.2) ordnet den Gebieten zeitliche, örtliche und raumzeitliche Merkmale zu. Beispiele für eingesetzte örtliche Merkmale sind Formmerkmale und Intensitätskrümmungsmerkmale auf Basis lokaler Schätzungen der Hesse-Matrix von Gebieten. Für den *Klassifikator* (s. Abb. 3.2) wurden weit verbreitete Klassifikationsverfahren aufgegriffen: k-Nearest Neighbors (k-NN) [293], Random Forest [288], Support Vector Machine (SVM) [331] und Naive Bayes [293]. Verfahren des Musterdetektors und des Klassifikators sind kombinierbar und automatisch adaptiv parametrisierbar[B2/344, B2/340, B2/339, B2/342]. Sie wurden in OpenCL mit GPU-Hardware als Zielsystem implementiert und erlauben die Bearbeitung von PAMONO-Bildfolgen in Echtzeit. Dies ist eine Grundlage der Optimierung in den Teilbereichen der Plattformoptimierung (Abschnitt B) und der Optimierung des Trade-off zwischen Erkennung, Verarbeitungszeit



(a) (b)

Abbildung 3.3.: Evaluierung des Verfahrens zur Erkennung von Nanoobjekten für den PAMONO-Sensor: (**a**) Recall-Wert zur Erkennung für drei Typen von Datensätzen mit jeweils unterschiedlichen Größen von Nanoobjekten und (**b**) Erkennungsleistung für vier Klassifikationsverfahren gemäß verschiedener Qualitätsmaße, gemittelt über alle Datensätze.

und Ressourcenbedarf (Abschnitt D).

Die Optimierung durch Synthese (Abb. 3.2, oben) dient der Anpassung von Steuerungsparametern des Musterdetektors und des Klassifikators. Die Anpassung geschieht durch Optimierung von Zielfunktionen, die deren Ergebnisse mit den Sollergebnissen vergleichen. Die Optimierung des Musterdetektors erfolgt aufgrund der höheren Anzahl an Freiheitsgraden und der Flexibilität bezüglich Verfahren des Musterdetektors evolutionär (JGAP, [326]; SPEA2, [335]; NSGA-II, [316]) und die der Klassifikatoren durch *Grid Search*.

Die Bildsequenzen werden inklusive der Sollergebnisse anhand von Realdaten durch Berechnung synthetisiert (Abb. 3.2, oben links). Die Synthese basiert auf einem Sensordatenmodell, in das auch eine kleine Anzahl von Bildbestandteilen aus den realen Daten eingeht, die von Hand ermittelt werden. Auf die verwendeten Realdaten wird später eingegangen. Das Sensordatenmodell, publiziert in [B2/343], besteht aus einer Hintergrundkomponente, multiplikativ moduliert mit dem Nano-Partikelsignal und additiv überlagert durch ein Sensorrauschen. Der Hintergrund wiederum wird als Produkt eines zeitlich konstanten Bildes der Sensoroberfläche mit einem Modulationsterm der störenden Artefakte modelliert.

Abb. 3.3 stellt hierzu einige Resultate der umfangreichen Evaluierung des Verfahrens kompakt zusammen. Sie umfassen reale Sensoraufnahmen mit Goldbeschichtungen hoher und geringer Qualität sowie synthetische Aufnahmen mit Partikelgrößen von 50, 100 und 200 nm (siehe auch Abschnitt C). Basierend auf derartigen Datensätzen wurde auch die Erfassungsleistung des Sensors analysiert. Es ergab sich ein linearer Zusammenhang zwischen der Partikelkonzentration in der Probe und der Anzahl der auf dem Sensor anhaftenden und detektierten Partikel. Dies ermöglicht eine kontinuierliche Ableitung der Konzentration und damit die Bestimmung der örtlichen und zeitlichen Konzentration der Nanoobjekte eines speziellen Typus (Aufgabe 3 - Konzentration). Als weitere Anwendung wurden Astrodaten aus dem TP C3 aufgegriffen, die ähnliche Charakteristika aufweisen – entsprechende Analysen befinden sich zum Zeitpunkt des Berichts noch in der Bearbeitung.

Mit dem beschriebenen Verfahren wurde ein Lösungsansatz geschaffen, der tragfähig für die Anpassung und Fortentwicklung für weitere Anwendungsgebiete in zukünftigen Forschungsarbeiten ist. Interessant sind Änderungen der räumlichen Ausdehnung, der Peakintensität, der Form und der zeitlichen Charakteristik der Detektionsereignisse. Damit ist etwa im Zusammenhang mit dem PAMONO-Sensor bei dessen Weiterentwicklung zur Erkennung von Nanopartikeln in Luft zu rechnen (siehe Abschnitt C).

Eine zusätzliche Herausforderung ist die Erkennung von Nanoobjekten von verschiedenen vorgegebenen Typen (Aufgabe 4). Im Grundsatz kann entsprechend wie bisher vorgegangen werden. Allerdings ist die Aufgabe strukturell komplexer und erfordert mehr Rechenleistung. Die Herausforderung ist, den Ressourcenbedarf bei Echtzeitanforderung gering zu halten.

Weitergehend ist die Betrachtung beweglicher Nanoobjekte in den Bildfolgen, im Unterschied zu den bisher als ortständig angenommen Objekten. Dies führt auf eine neue Aufgabe, die Erkennung und die Verfolgung bewegter Nanoobjekte in verrauschten Bildfolgen. Sie wird im Folgenden als Aufgabe 5 - Objektverfolgung bezeichnet. Bei der Objektverfolgung (Object Tracking) handelt es sich um ein intensiv behandeltes Problem des Computersehens mit zahlreichen Anwendungen, das nach wie vor Gegenstand großen Forschungsinteresses ist [322, 332, 334]. Die Objektverfolgung in verrauschten Bildsequenzen wurde bisher wenig behandelt. Bekannte Arbeiten in diesem Zusammenhang betreffen eher große verfolgte Objekte [322]. Zur Verfolgung

kleiner Objekte gibt es Arbeiten in Bezug auf die Zell- und Partikelverfolgung [327]. Hier ist ein Defizit bezüglich verrauschter Umgebungen festzustellen. Die für den PAMONO-Sensor relevante Anwendung liegt in dessen Erweiterung zur Erfassung von „rollenden“ Viren (siehe Abschnitt C).

Weitere Problemstellungen ergeben sich aus dem potenziell unterschiedlichen technischen Verhalten baugleicher Sensoren, welches vom Fertigungsprozess und der Anwendungsumgebung beeinflusst sein kann. Beim PAMONO-Sensor sind dies die untersuchte Probe und die Umwelt (Wärme). Dies erfordert Kompensations-, Kalibrierungs- und Qualitätskontrollmaßnahmen, welche für die Aufgabenstellungen des Projekts bisher noch nicht bearbeitet wurden. Eine verwandte Aufgabe, aus der sich Anregungen für Lösungsansätze ergeben können, ist das Tracking von Personen über mehrere Kameras mit unterschiedlichem, auch potenziell disjunktem Sichtfeld [319]. Die aufgenommenen Bildfolgen unterschiedlicher Kameras können sich etwa durch unterschiedliche Beleuchtungsverhältnisse merklich unterscheiden. Ansätze zur Kompensation sind die Angleichung der Bilder durch Intensitätstransferfunktionen [323] sowie die Verwendung beleuchtungsunempfindlicher Merkmale. Die Maßnahmen, insbesondere die Qualitätskontrolle auf Grundlage von Verhaltensdaten einer größeren Anzahl von Sensoren, erfordert zusätzliche Serverleistung außerhalb des Einzelsensors, wodurch sich die Anforderung an die Kommunikationsfähigkeit der Sensoren ergibt. Als Beispiel für den Einfluss von Wärme ist eine Untersuchung des 3D-Sensors *Microsoft Kinect* zu nennen [318].

B. Plattformentwurf unter Optimierung des Ressourcenverbrauchs

Das Ziel des Plattformentwurfs ist die Minimierung des Ressourcenverbrauchs bei der Ausführung eines vorgegebenen Algorithmus oder Algorithmentyps. Freiheitsgrade des Plattformentwurfs sind die Abbildung eines vorgegebenen Algorithmus auf die Plattform sowie die Variation der Struktur der Plattform. In diesem Projekt wurde diese Problemstellung vorrangig für PCs mit GPU-Karte als Coprozessor betrachtet. Dabei wurden zwei Fälle unterschieden: Optimierung für eine feste Plattform und Optimierung über Plattformvarianten.

Zur Erfassung des Ressourcenverbrauchs wurden zwei methodische Ansätze verfolgt und realisiert: ein messtechnikbasierter und ein simulationsbasierter Ansatz. Dabei lag der Schwerpunkt auf den Ressourcen „Ausführungszeit“ und „Energie“. Für den messtechnikbasierten Ansatz wurde eine rechnergesteuerte Messplattform zur Erfassung des Energieverbrauchs während einer Programmausführung auf realer Hardware aufgebaut. Der simulationsbasierte Ansatz verwendete den Simulator *GPGPU-Sim* [313] und das Energiemodell *GPU-Wattch* [325].

Die Optimierung erfolgte im Wesentlichen mit mehrkriteriellen evolutionären Algorithmen in Form der Implementierungen SPEA2 [335] und NSGA-II [316] unter Verwendung von Benchmark-Programmen sowie der PAMONO-Objekterkennung aus Abschnitt A. Eine entsprechende Abbildung der Algorithmen auf die Plattform geschah über deren Implementierung in OpenCL. Zur Optimierung wurden im Wesentlichen zwei Freiheitsgrade eingesetzt, das *Instruction-Scheduling* und die Abbildung von *OpenCL-Workgroups*. Basierend auf diesen Methoden wurden für den Fall eines fest gegebenen GPU-Coprozessors, unter Betrachtung mehrerer Systeme, die folgenden konkreten Untersuchungen mit der messtechnischen Methode durchgeführt:

1. Mehrkriterelles *Instruction-Scheduling* innerhalb eines Basisblocks in OpenCL („MOBLIS“, Kap. 5.3, Dissertation Timm [B2/347]). Für anwendungsbezogene Benchmark-Programme wurde eine Verbesserung des Energieverbrauchs um bis zu 9% und der Rechenzeit um bis zu 12% auf einer Nvidia 8400GS und bis zu 6.67% bzw. 8.5% auf einer Nvidia GTS 250 erzielt.

2. Mehrkriterielles, speicherschonendes, globales *Instruction-Scheduling* zwischen Basisblöcken in OpenCL (FALIS, Kap. 5.4 der Dissertation Timm [B2/347]). Es hat sich gezeigt, dass durch die Optimierung ca. 10% des Energieverbrauchs eingespart und etwa 13% Rechenzeitreduktion für anwendungsbezogene Benchmark-Programme auf verschiedenen GPU-Karten erzielt werden können.

3. Optimiertes Abbilden von *Working-Groups* in OpenCL (Kap. 5.5, Dissertation Timm [B2/347]). Es konnte festgestellt werden, dass der Lastausgleich und die Skalierbarkeit des Anwendungsprogramms eine direkte Auswirkung auf die Energieeffizienz haben.

Die Inhalte der Dissertation von Herrn Timm wurden in mehreren Konferenz- und Workshop-Beiträgen international publiziert. Die Dissertation enthält eine Übersicht über diese Publikationen.

Für den Fall variabler Plattformen erfolgte die Optimierung simulationsbasiert, wobei aufgrund des hohen Rechnerzeitbedarfs auf einer Cloud von Rechnern gerechnet wurde. Die Optimierung geschah über mehrere GPU-Kartenarchitekturen als Freiheitsgrad. Dazu gehörten Desktop- und mobile GPU-Karten. Hier wurde bis zum Zeitpunkt der Berichtserstellung eine größere Anzahl von Benchmark-Programmen und der PAMONO-Erkenner (siehe Abschnitt A) optimiert. Die Optimierungsziele waren der Energieverbrauch, die

Leistungsaufnahme sowie simultan der Energieverbrauch und die Anzahl der Zyklen. Dabei zeigte sich, dass unter einer Auswahl von Grafikkarten der Kepler-Architektur solche mit GDDR5-Speicher energieeffizienter arbeiten als die mit GDDR3-Speicher. Ferner erwies sich eine größere Anzahl an Streaming-Multiprozessoren von Vorteil. Jedoch war die maximal mögliche Anzahl Streaming-Multiprozessoren nicht für alle Programme am besten für den Energieverbrauch. GPU-Karten mit wenigen Streaming-Multiprozessoren sind gut für die Leistungsaufnahme, aber schlecht für den Energieverbrauch. Erste Resultate wurden in [B2/337] publiziert.

Zur Optimierung des Ressourcenbedarfs wurden im laufenden Projekt Freiheitsgrade der Ausführung der Algorithmen bzw. der sie realisierenden Programme sowie der Struktur der Hardware-Plattformen genutzt. Der wesentliche Freiheitsgrad der Algorithmen- bzw. Programmausführung besteht in dem Ort und der Reihenfolge der Ausführung von unabhängigen Teilen. Dies ist die Aufgabe des Scheduling. In diesem Projekt fand das Scheduling im Wesentlichen auf Instruktionsebene und nur wenig auf Funktionsebene statt. Im OpenCL-Programmiermodell stehen Mechanismen zur Formulierung der Abhängigkeit und damit auch der Unabhängigkeit von Programmteilen auf Funktionsebene zur Verfügung, wohingegen die Abhängigkeit auf Instruktionsebene, wie in üblichen Programmiersprachen, nur implizit ausgedrückt werden kann und in der Regel nicht explizit genutzt wird. Letzteres ist aus praktischen Gründen nachvollziehbar und die automatische Optimierung durch Compiler hinreichend.

Auf Funktionsebene sind jedoch weitergehende Mechanismen denkbar. Zu nennen ist hier die Programmiersprache *HALIDE* [330] und das zugrunde liegende Konzept der Trennung von Funktionsdeklaration und Scheduling bei der Programmierung. Eine weitere Möglichkeit ist die Bereitstellung alternativer Formulierungen von Funktionen, beispielsweise im Hinblick auf die Wiederverwendung von Ergebnissen durch Speicherung oder die jeweilige Neuberechnung. HALIDE ist speziell für Bildverarbeitungsalgorithmen und im Hinblick auf die Minimierung der Rechenzeit entworfen. Die Erweiterung des Konzepts im Hinblick auf andere Ressourcen und andere Anwendungsdomänen sind interessante offene Fragestellungen.

Im bisherigen Projekt standen Plattformen mit GPU-Coprozessoren im Vordergrund. Sie erwiesen sich bei der PAMONO-Anwendung als hinreichend leistungsfähig zur Echtzeitklassifikation von Nanoobjekten. Die angestrebten rechenaufwändigeren Aufgaben lassen sich grundsätzlich auf dieser Basis bei entsprechend erweiterter Rechenleistung der Plattform angehen. Damit ist jedoch ein Energiebedarf verbunden, welcher der Miniaturisierung für den mobilen Einsatz, beispielsweise von Sensoreinheiten wie dem PAMONO-Sensor, aus Gründen des Akku-Betriebs und der potenziellen Auswirkung der Wärmeentwicklung auf die Datenerfassung entgegensteht. Eine Möglichkeit, dieses Problem anzugehen, besteht im Auslagern (Offloading) von Teilen der Berechnung auf andere Berechnungsressourcen über drahtlose Netzverbindungen. Für diese Vorgehensweise gibt es zahlreiche andere Anwendungen im Zusammenhang mit Cyber-Physical Systems. Basierend auf Beschreibungen der Plattform und der Anwendung kann ein geeignetes Scheduling stattfinden.

Innerhalb der letzten zehn Jahre wurden einige Studien durchgeführt, die sich mit dem *Remote Computing* in eingebetteten Systemen beschäftigt haben. So wurde in [333, 329] das Auslagern von Berechnungen für mobile Endgeräte verwendet. Dabei wurde die mobile Anwendung als gerichteter Graph modelliert, um zu entscheiden, wie die Anwendung für die entfernte Berechnung aufgeteilt werden kann.

Die meisten existierenden Ansätze zum Auslagern von Berechnungen verwenden keine weichen oder harten Echtzeiteigenschaften. So liegt der Fokus in [333, 329, 324] auf der Beschleunigung der Berechnung und einer Verbesserung des Energieverbrauchs, allerdings ohne das Zeitverhalten zu berücksichtigen. Von Nimmagadda et al. wird in [328] ein Framework für mobile Roboter vorgestellt, mit dem sich bewegende Objekte in Echtzeit erkannt und verfolgt werden. Des Weiteren untersuchen Ferreira et al. in [317], wie das Auslagern die Qualität des Services in Echtzeitsystemen verbessern kann.

Die erwähnten Ansätze haben aber zumeist eine oder mehrere Einschränkungen: (a) Die lokale CPU wird nicht ausgelastet, während auf die entfernt berechneten Ergebnisse gewartet wird. (b) Die aktuelle Serverauslastung wird nicht berücksichtigt. (c) Das Echtzeitverhalten wird nicht berücksichtigt. In der Praxis steht ein Server nicht nur dem einen mobilen Endgerät zur Verfügung, sondern es teilen sich mehrere Dienste den Server, zumal ein mobiles Endgerät den Server alleine kaum auslasten könnte. Um die Echtzeitfähigkeit zu gewährleisten, sollte der Server bestimmte Programm-Laufzeiten garantieren können. Dies umfasst sowohl Garantien für die Antwortzeiten, als auch solche für den Ressourcenverbrauch. Aus diesem Grund wird auch auf Seiten des Servers eine Virtualisierung und Reservierung von Ressourcen benötigt.

Um die Leistung von eingebetteten Systemen durch das Auslagern von Berechnungen zu verbessern, wurden bereits mehrere Vorarbeiten geleistet [B2/345, B2/346]. Insbesondere wurden Offloading-Schemata für Echtzeitaufgaben untersucht, wobei die anstehenden Aufgaben zur gleichen Zeit beginnen und die gleichen harten Schranken für die Ausführungszeit haben. Die Studie [B2/346] zeigt ein Offloading-Schema unter der Annahme, dass der oder die Server eine Zeitgarantie für die auszulagernde Aufgabe geben können. In [B2/345] können einzelne mobile Geräte auch auf ein Offloading verzichten, woraufhin der Server Reservierungen freigeben kann und sich dadurch die Garantien für andere Aufgaben verbessern. Die Garantien

für die Antwortzeit hängen dabei vom dynamischen Verhalten ab, also wie genau die Aufgaben auf den Server ausgelagert werden. Diese ersten Ergebnisse [B2/345, B2/346] betreffen sehr einfache Gegebenheiten, in denen alle Aufgaben dieselbe Ankunftszeit und dieselbe Frist haben, und in denen die Parameter und Konfigurationen der Server und der Kommunikationen fest und im Voraus bekannt sind.

Zusammengefasst sind die größten Herausforderungen für das Auslagern von Berechnungen in ressourcenbeschränkten eingebetteten Systemen die Umsetzung der Ressourcenreservierung auf hoch performanten Serverumgebungen, die Konfiguration der Kommunikationskanäle zur Bereitstellung der Daten und der Kommunikation in Echtzeit sowie die Umsetzung von reaktiven und proaktiven Ausführungsschemata auf ressourcenbeschränkten Endgeräten.

C. Entwicklung hochdifferenzierender Biosensoren auf Basis der Plasmonen-unterstützten Mikroskopie zur Detektion von Nanoobjekten

Die hierzu ausgeführten Arbeiten umfassten die Untersuchung von Einflussgrößen des PAMONO-Sensorprinzips auf die Erkennungsleistung und deren Beeinflussung zur Erhöhung der Erkennungsleistung, die Beschreibung des zugrunde liegenden physikalischen Effekts sowie die Erzeugung von Testdatensätzen und deren Einsatz in Experimenten zur Evaluation des PAMONO-Sensors, s. Abschnitt A.

Die Erkennungsleistung wird wesentlich durch die Auswirkung einer auftretenden Speckle-Struktur auf die laterale Auflösung der optischen Abbildung der mit Laserlicht illuminierten, goldbeschichteten Fläche beeinträchtigt. Die untersuchten, potenziellen Einflussgrößen betrafen die Qualität der Goldschicht, die Größe der zur Detektion nutzbaren Sensorfläche, die Beleuchtung und das Kamerasystem.

Durch theoretische Überlegungen und umfassende praktische Untersuchungen mit maschineller Auswertung an Goldschichten hoher und geringer Qualität konnte gezeigt werden, dass die thermische Bedampfung mit Gold, die bisher als Methode bei der Oberflächenplasmonenresonanz etabliert ist, die erreichbare Qualität der Sensorschicht begrenzt. Ferner konnte nachgewiesen werden, dass eine grundsätzliche Limitierung im Hinblick auf die Detektion von Partikeln unterhalb einer Größe von 50 nm gegeben ist [B2/343]. Für einen neu entwickelten Ansatz, der ein Magnetron Sputtering zur Erzeugung der Goldschicht beinhaltet, konnte eine wesentlich geringere Rauigkeit und eine höhere Homogenität der Oberfläche erreicht werden, die eine Verdoppelung des SNR (Signal-Rausch-Verhältnis) zur Folge hatte. Das SNR ist ein wesentlicher limitierender Faktor bei der Detektion von Nanopartikeln.

Es wurde herausgefunden, dass die analytische Nachweisstärke von Nanoobjekten von der Größe der Detektionsfläche abhängig ist. Zur Minderung dieser Problematik konnte eine Verbesserung des nutzbaren Detektionsflächenanteils erzielt werden. Da die abzubildende Fläche schräg zu der optischen Achse positioniert ist, kann bei der klassischen Abbildungsoptik nur ein Fragment der Fläche von ca. 0.2 mm^2 scharf abgebildet werden. Über ein neu entwickeltes Abbildungssystem mit einem spezifischen Beugungsgitter konnte die relevante Flächengröße auf ca. 10 mm^2 gesteigert werden.

Durch Analyse der Wirkung der Beleuchtung auf die Speckle-Struktur wurde festgestellt, dass es durch Hochfrequenzmodulation (500 MHz) des Stroms der Laserdiode möglich ist, die Speckle-Struktur ebenfalls auf technischer Ebene zu reduzieren [B2/341]. Ferner wurden alternative Lichtquellen untersucht, z. B. Superlumineszenz-Dioden, welche zwar ein geringes SNR aufweisen, aber auch eine niedrige Leistung und schlechtere Strahlqualität – dieser Aspekt ist im Zusammenhang mit den Trade-offs zu berücksichtigen. Durch Untersuchung unterschiedlicher Kamerasysteme unter Berücksichtigung der Zielkriterien eines optimalen SNR sowie akzeptabler lateraler und zeitlicher Auflösung konnte gezeigt werden, dass durch Auswahl eines passenden Kamerasystems das Schrotrauschen um den Faktor 3 bis 4 reduziert und gleichzeitig die zeitliche Auflösung verdoppelt werden kann [B2/336]. Dies ist bedeutungsvoll, da erkannt wurde, dass das SNR primär durch Schrotrauschen begrenzt ist und eine Akkumulation von Photonen bei der Bildaufnahme dies zumindest reduziert.

Für den physikalischen Effekt, welcher dem PAMONO-Sensor zugrunde liegt, wurde eine quantitative Beschreibung hergeleitet. Eine zugehörige Publikation befindet sich in Vorbereitung. Aufgrund der neuen theoretischen Basis kann jetzt eine differenzierte Optimierung von Messungen effektiv durchgeführt werden. Diese erlaubt auch, das entwickelte Sensormodell [B2/343] weiter den realen Gegebenheiten anzupassen.

Zur experimentellen Analyse, die insbesondere die Evaluierung des PAMONO-Sensors einschloss, deren Resultate in Abschnitt A skizziert werden, wurden reale Testdatensätze mit unterschiedlichen Gegebenheiten hergestellt, u. a. Proben unter Verwendung von 280 nm- und 200 nm-Polystyren sowie Virus-Like-Particles (VLPs). Mittels eines PAMONO-Sensors wurden daraus Bildsequenzen erfasst. In den Bildsequenzen wurde schließlich eine manuelle Annotierung von relevanten Nanoteilchen in den Bildsequenzen durchgeführt [B2/344, B2/338]. Durch die Verbesserungsmaßnahmen wurde ein wesentlicher Beitrag zur automatischen Erkennbarkeit von Nanoobjekten erzielt.

Durch das Zusammenspiel der dargestellten Verbesserung des Verhaltens und der effizienten Analyse konnte die Brauchbarkeit des PAMONO-Konzepts für Nanoobjekte in Flüssigkeiten nachgewiesen werden. Ein Themenkreis für die zukünftige Forschung betrifft die Einsetzbarkeit des Detektionsprinzips für andere Gegebenheiten. Hier ist zunächst die Erkennung von mehr als einem vorgegebenen Typ von Nanoobjekten zu nennen (Aufgabe 4). Dies kann durch Aufteilung der Sensorfläche in unterschiedlich sensitive Gebiete erzielt werden. Aufgrund der dargestellten Erkenntnis, dass die Größe der sensitiven Fläche für die Bedeutungsleistung wesentlich ist, stellen sich hier neue Herausforderungen an den Aufbau des Sensors. Eine weitere Fragestellung ist das Erkennen und die Konzentrationsmessung (Aufgabe 3) von Nanoteilchen in anderen Medien. Ein anwendungsrelevantes Beispiel sind Feinstaubpartikel in der Luft. Dies eröffnet zahlreiche neue Anwendungsgebiete, wie Umweltschutz und Abgasanalyse bei Dieselmotoren. Neue Aufgaben ergeben sich hier in anderen Formen der Nanoobjekte in den Bildern und einer anderen zeitlichen Charakteristik. Eine dritte Fragestellung betrifft nicht ortsständige Nanoobjekte auf der Sensoroberfläche. Das bedeutet, dass die Objekte beweglich, etwa „rollend“, auf der Sensoroberfläche wahrnehmbar sind. Das ist beispielsweise bei Grippeviren der Fall. In diesem Fall ist die Bewegungstrajektorie ein Merkmal der Erkennung. Hier stellen sich etwa Fragen zum Transportverhalten der Nanoobjekte, der Beleuchtung und des Verständnisses der auftretenden katalytischen Prozesse.

Ein zweiter Themenkreis ist die Steigerung der Mobilität des Sensors, wobei die Miniaturisierung ein wesentlicher Aspekt ist. Ein Anwendungsszenario ist der Betrieb mehrerer drahtlos kommunizierender Sensoren an Orten wie Flughäfen oder Grenzstationen, die zusätzlich mit einem Server (oder mehreren Servern) kommunizieren können. Ein Aufgabenkreis in diesem Zusammenhang ist die Verbesserung der sensorischen Eigenschaften im Hinblick auf die zeitliche Auflösung und den Bindungsverlauf. Ein weiterer Aspekt ist die Bestimmung von Nachweisgrenzen.

Ein dritter Themenkreis ist die Qualitätssicherung der Sensorfunktion. Dies betrifft die Kalibrierung des Sensors, die Probenvor- und -aufbereitung sowie die Temperaturabhängigkeit des Sensors. Hier ist eine Beziehung zum etablierten Gebiet des Condition-Monitoring gegeben, in dem insbesondere Datenanalysemethoden zur Maschinenüberwachung Anwendung finden [321]. Neue Aspekte im Rahmen dieses Projektes liegen in der vorliegenden, anders gelagerten Datencharakteristik sowie die Realisierung unter Beachtung der Ressourceneffizienz.

D. Optimierung des Trade-off zwischen Erkennungsleistung, Verarbeitungszeit und Ressourcenbedarf und Validierung der Resultate

Dieser Themenkreis befindet sich gemäß Zeitplan zum Zeitpunkt dieser Berichterstellung noch in der Bearbeitung. Er befasst sich mit Trade-offs zwischen Erkennungsleistung und Ressourcenverbrauch. Zur Bewertung und Optimierung des Trade-off von Erkennungsleistung und Energieverbrauch wurden die Algorithmen der Plattformoptimierung mit Algorithmen aus AP 1 kombiniert. Dabei werden zur Beurteilung der Erkennungsrate die üblichen Gütemaße aufgegriffen: *F1-Measure*, *Accuracy*, *Precision*, *Recall* und *Matthews Correlation Coefficient*. Die Optimierung geschieht mit evolutionären Algorithmen.

Zum Zeitpunkt der Berichtserstellung wurden bereits kleinere Auswertungen durchgeführt, in denen die Laufzeit über Algorithmen und Plattformparameter optimiert wurde. Die eigentliche Auswertung steht noch aus. Der nächste Schritt ist dann die mehrkriterielle Optimierung des Energieverbrauchs und der Erkennungsrate. Hier gehen sowohl Plattformparameter in Form von unterschiedlichen GPUs als auch Algorithmenparameter ein, die Erkennungsrate und Laufzeit beeinflussen. Dadurch sollen die angestrebten Erkenntnisse zum Trade-off zwischen Energieverbrauch und Virus-Erkennungsrate erzielt werden.

Die gewählte Vorgehensweise der Optimierung soll für die neuen Gegebenheiten bezüglich Algorithmen und Plattformen weiter verwendet werden.

Literatur

[C3/312] S. Abbas. „Detektion von Nanoobjekten in Graustufenbildern und Bildsequenzen mittels robuster Zeitreihenmethoden zur Strukturbrucherkennung“. Masterarbeit. TU Dortmund, 2013.

[313] A. Bakhoda, G. Yuan, W. Fung, H. Wong und T. Aamodt. „Analyzing CUDA Workloads using a detailed GPU Simulator“. In: *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on.* 2009, S. 163–174.

[288] L. Breiman. „Random Forests“. In: *Machine Learning* 45.1 (2001), S. 5–32.

[314] R. R. Coifman und D. L. Donoho. „Wavelets and Statistics“. In: Hrsg. von A. Antoniadis und G. Oppenheim. Springer New York, 1995. Kap. Translation-invariant De-noising, S. 125–150.

[315] K. Dabov, A. Foi und K. Egiazarian. „Video Denoising by Sparse 3D Transform-Domain Collaborative Filtering“. In: *15th European Signal Processing Conference*. Bd. 1. 2. 2007.

[316] K. Deb, S. Agrawal, A. Pratap und T. Meyarivan. „A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II“. In: *Parallel Problem Solving from Nature – PPSN VI*. Berlin: Springer, 2000, S. 849–858.

[317] L. L. Ferreira, G. Silva und L. M. Pinho. „Service Offloading in Adaptive Real-Time Systems“. In: *Emerging Technologies & Factory Automation (ETFA), IEEE 16th Conf. on*. IEEE. 2011, S. 1–6.

[318] D. Fiedler und H. Müller. „Impact of Thermal and Environmental Conditions on the Kinect Sensor“. In: *Advances in Depth Image Analysis and Applications*. Bd. 7854. LNCS. 2013, S. 21–31.

[319] S. Gong, M. Cristani, S. Yan und C. Loy. *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition. Springer London, 2014.

[320] E. Haber und J. Modersitzki. „Numerical Methods for Volume preserving Image Registration“. In: *Inverse Problems* 20 (2004), S. 1621–1638.

[321] Y. Han und Y. H. Song. „Condition Monitoring Techniques for electrical Equipment - a Literature Survey“. In: *IEEE Transactions on Power Delivery* 18.1 (2003), S. 4–13.

[293] T. Hastie, R. Tibshirani und J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Statistics. Springer, 2009.

[322] A. S. Jalal und V. Singh. „The State-of-the-Art in Visual Object Tracking.“ In: *Informatica (Slovenia)* 36.3 (2012), S. 227–248.

[323] O. Javed. „Appearance Modeling for Tracking in Multiple Non-overlapping Cameras“. In: *In IEEE International Conference on Computer Vision and Pattern Recognition*. 2005, S. 26–33.

[324] D. Kovachev, T. Yu und R. Klamma. „Adaptive Computation Offloading from Mobile Devices into the Cloud“. In: *Parallel and Distributed Processing with Applications (ISPA), 2012 IEEE 10th International Symposium on*. IEEE. 2012, S. 784–791.

[325] J. Leng, T. Hetherington, A. ElTantawy, S. Gilani, N. S. Kim, T. M. Aamodt und V. J. Reddi. „GPUWattch: Enabling Energy Optimizations in GPGPUs“. In: *International Symposium on Computer Architecture* (2013).

[326] K. Meffert, N. Rotstan, C. Knowles und U. Sangiorgi. „JGAP - Java Genetic Algorithms and Genetic Programming Package“. In: *URL: http://jgap.sf.net* (2011).

[327] E. Meijering, O. Dzyubachyk und I. Smal. „Methods for Cell and Particle Tracking“. In: *Imaging and Spectroscopic Analysis of Living Cells* 504 (2012), S. 183–200.

[328] Y. Nimmagadda, K. Kumar, Y.-H. Lu und C. G. Lee. „Real-Time moving Object Recognition and Tracking using Computation Offloading“. In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE. 2010, S. 2449–2455.

[329] S. Ou, K. Yang und A. Liotta. „An adaptive Multi-Constraint Partitioning Algorithm for Offloading in Pervasive Systems“. In: *Pervasive Computing and Communications, 2006. PerCom 2006. Fourth Annual IEEE International Conference on*. IEEE. 2006, 10pp.

[330] J. Ragan-Kelley, A. Adams, S. Paris, M. Levoy, S. P. Amarasinghe und F. Durand. „Decoupling Algorithms from Schedules for easy Optimization of Image Processing Pipelines“. In: *ACM Trans. Graph.* 31.4 (2012), S. 32.

[331] B. Schölkopf, C. J. C. Burges und A. J. Smola. *Advances in Kernel Methods: Support Vector Learning*. MIT Press Cambridge, 1999.

[332] A Smeulders, D Chu, R. Cucchiara, S. Calderara, A. Dehghan und M. Shah. „Visual Tracking: an Experimental Survey“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013).

[333] K. Yang, S. Ou und H.-H. Chen. „On Effective Offloading Services for Resource-Constrained Mobile Devices Running Heavier Mobile Internet Applications“. In: *IEEE Comm. Magazine* 46.1 (2008), S. 56–63.

[334] A. Yilmaz, O. Javed und M. Shah. „Object Tracking: A Survey“. In: *ACM Comput. Surv.* 38.4 (2006), S. 13.

[335] E. Zitzler u. a. „SPEA2: Improving the Strength Pareto Evolutionary Algorithm For Multiobjective Optimization“. In: *Proc. Int. Conf. on Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems*. Athens, Greece, 2001.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[B2/336] E. L. Gurevich, V. Temchura, K. Überla und A. Zybin. „Analytical Features of Particle Counting Sensor based on Plasmon assisted Microscopy of Nano Objects“. In: *Sensors and Actuators B: Chemical* B 160.1 (Sep. 2011), S. 1210 –1215.

[B2/337] P. Libuschewski, D. Siedhoff und F. Weichert. „Energy-aware Design Space Exploration for GPGPUs“. In: *Computer Science - Research and Development* (2013), S. 1–6.

[B2/338] P. Libuschewski, C. Timm und F. Weichert. „Parameteroptimierte und GPGPU-basierte Detektion viraler Strukturen innerhalb Plasmonen-unterstützter Mikroskopiedaten“. In: *Bildverarbeitung für die Medizin.* Lecture Notes in Informatics. Springer-Verlag, 2012, S. 237–242.

[B2/339] P. Libuschewski, D. Siedhoff, C. Timm, A. Gelenberg und F. Weichert. „Fuzzy-enhanced, Real-time capable Detection of Biological Viruses Using a Portable Biosensor“. In: *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSIGNALS).* 2013, S. 169–174.

[B2/340] P. Libuschewski, D. Siedhoff, C. Timm und F. Weichert. „Mobile Detektion viraler Pathogene durch echtzeitfähige GPGPU-Fuzzy-Segmentierung“. In: *Bildverarbeitung für die Medizin 2013* (2013), S. 326–331.

[B2/341] V. Shpacovitch. „Application of Surface Plasmon Resonance (SPR) for the detection of single viruses and single biological nano-objects“. In: *Journal of Bacteriology and Parasitology* 3.7 (2012), S. 1–3.

[B2/342] D. Siedhoff, F. Weichert, P. Libuschewski und C. Timm. „Detection and Classification of Nano-Objects in Biosensor Data“. In: *Microscopic Image Analysis with Applications in Biology* (2011).

[B2/343] D. Siedhoff, P. Libuschewski, F. Weichert, A. Zybin, P. Marwedel und H. Müller. „Modellierung und Optimierung eines Biosensors zur Detektion viraler Strukturen“. In: *Bildverarbeitung für die Medizin. Lecture Notes in Informatics. Springer-Verlag* (2014), S. 108–113.

[B2/344] C. Timm, P. Libuschewski, D. Siedhoff, F. Weichert, H. Müller und P. Marwedel. „Improving Nanoobject Detection in Optical Biosensor Data“. In: *Proceedings of the 15th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2011.* Hrsg. von N. e. a. Callaos. Bd. II. Orlando, Florida, USA: IIIS, 2011, S. 236–240.

[B2/345] A. Toma und J.-J. Chen. „Computation Offloading for Frame-Based Real-Time Tasks with Resource Reservation Servers“. In: *Real-Time Systems (ECRTS), 2013 25th Euromicro Conference on.* IEEE. 2013, S. 103–112.

[B2/346] A. Toma und J.-J. Chen. „Computation Offloading for Real-Time Systems“. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing.* ACM. 2013, S. 1650–1651.

Andere Veröffentlichungen

[B2/347] C. Timm. „Resource Efficient Processing and Communication in Sensor/Actuator Environments“. Dissertation. TU Dortmund, 2012.

3.4 Planung des Teilprojekts

Ziele

Das globale Ziel des Vorhabens ist die Erkennung spezifischer Nanoobjekte in Bildfolgen, die durch mehrere erschwerende Randbedingungen gekennzeichnet sind. Gegenstand sind sensorerfasste, datenintensive Bildfolgen natürlicher Prozesse, die bedingt durch den erfassenden Sensor stark verrauscht sind und zu detektierende Ereignisse zeigen, die relativ zu dem durch die Bilder erfassten Ortsausschnitt klein sind und deren Eintrittszeit kurz ist. Trotzdem soll die Verarbeitung in Echtzeit geschehen und sollen die eingesetzten Sensor- und Berechnungsressourcen gering gehalten werden. Hierzu sind methodische Ansätze der Bildfolgenanalyse für die bildbasierte Erfassung des Auftretens von Nanoobjekten unter Berücksichtigung der gegebenen Randbedingungen zu identifizieren, die einen Trade-off zwischen Erkennungsleistung, Verarbeitungszeit und Ressourcenbedarf aufweisen und eine Online-Verarbeitung erlauben. Darauf aufbauend sollen Methoden zur mehrkriteriellen Optimierung des Systems aus Bildfolgenerfassung, Bildfolgenanalyse

und Berechnungsplattform in Bezug auf Erkennungsleistung, Echtzeitverarbeitung und Ressourcenbedarf gefunden werden.

Mit diesen Methoden soll als Anwendungsfallstudie eine zuverlässige Bildsequenzauswertung der plasmonen-unterstützten Mikroskopie von Nanoobjekten (PAMONO) auf portabler ressourcenoptimierter Echtzeitanalysehardware ermöglicht werden. Die dafür notwendigen technischen Voraussetzungen der PAMONO-Technik sollen mit dem Ziel verbessert werden, die Erkennungsleistung und die Ressourceneffizienz des Systems zusätzlich zu den optimierten Algorithmen zu steigern. Ferner sollen neue Anwendungsfelder der PAMONO-Technik erschlossen werden.

Innerhalb dieser globalen Zielsetzung, die auch schon Gegenstand der ersten Förderphase war, werden hinsichtlich der Bildanalyse, der Rechensystemumgebung und der PAMONO-Technik neue Teilziele verfolgt. Bei der Bildanalyse beziehen sie sich auf neue Aufgabenvarianten und die Berücksichtigung des Anwendungskontextes zur Steigerung der funktionalen Leistungsfähigkeit:

TZ 1 - Mehrobjektdetektion und Objektverfolgung: Das Ziel sind Algorithmen zur Erkennung von kleinen Objekten in verrauschten Bildfolgen mit Schwerpunkten auf der Detektion von mehreren Nanoobjekttypen und nicht ortsständigen Nanoobjekten, sowie der Bewegungsverfolgung dieser. Die Algorithmenentwicklung soll nach dem Paradigma von TZ 3 erfolgen. Ferner sollen die Algorithmen freie Parameter für eine übergreifende Optimierung bereitstellen.

TZ 2 - Global gesteuerte Bildanalyse: Das Ziel ist die Schaffung zentraler, übergeordneter Qualitätskontroll-, Kalibrierungs- und Kompensationsverfahren für sich dezentral in der Anwendung befindliche Bilderkennungsplattformen durch Sammlung und Analyse von Information über deren Verhalten.

Bei der Rechensystemumgebung ergeben sich die Teilziele aus der Fokussierung auf verteilte Systeme, wozu Paradigmen auf Algorithmen- und Ausführungsebene ausgearbeitet werden:

TZ 3 - Ausführungskontextsensitives Algorithmen- und Programmierparadigma: Das Paradigma soll die Abbildung von Algorithmen und Programmen auf verteilte Berechnungsplattformen unter Berücksichtigung des Ressourcenverbrauchs effektiver und effizienter machen, indem für den Algorithmenentwurf und der Programmierung entsprechende Ausdrucksmöglichkeiten geschaffen werden.

TZ 4 - Kollaboratives Ausführungsparadigma: Im Zentrum steht die Realisierung des Konzepts des Offloading für Berechnungen mit Echtzeitanforderung auf ressourcenbeschränkten, verteilten, eingebetteten Systemen mit dem Ziel der Verbesserung der Dienstgüte durch geeignete Verfahren der Ressourcenreservierung auf Servern, durch die Konfiguration von Kommunikationskanälen, um zeitgerechte Datenbereitstellung und Datenkommunikation zu gewährleisten, und durch reaktive und proaktive Ausführungsschemata auf ressourcenbeschränkten Klienten. Das Offloading soll freie Parameter für eine übergreifende Optimierung bereitstellen. Neben dem Offloading sollen Kollaborationsmechanismen geschaffen werden, die die Realisierung der global gesteuerten Bildanalyse aus TZ 2 ermöglichen.

Bezüglich der PAMONO-Methode ergibt sich folgende konkrete Zielsetzung:

TZ 5 - Erweiterung und Miniaturisierung des PAMONO-Sensors: Die PAMONO-Methode soll auf andere Anwendungsfelder, insbesondere die simultane Erkennung mehrerer Typen von Nanoobjekten, die Detektion von Partikeln in der Luft und die Detektion nicht ortsständiger Nanoobjekte erweitert werden. Ferner soll der Sensor zur Erhöhung der Portabilität miniaturisiert werden.

Arbeitsprogramm

Das Arbeitsprogramm gliedert sich in 10 Arbeitspakete (Abb. 3.4). Sie werden aufgegliedert in die Teilbereiche A bis D, die aus der ersten Förderperiode übernommen werden: Teilbereich A bestehend aus AP 1 bis AP 4, wird vorrangig vom Lehrstuhl Informatik 7 (Müller) bearbeitet; Teilbereich B mit AP 5 bis AP 7, wird vorrangig vom Lehrstuhl Informatik 12 (Chen) bearbeitet; Teilbereich C mit AP 8 und AP 9, wird vorrangig vom ISAS (Hergenröder) bearbeitet, und Teilbereich D mit AP 10, der von allen gemeinsam bearbeitet wird.

Arbeitspaket 1. Ausführungskontextsensitive Algorithmen: Das Arbeitspaket ist dem TZ 3 zugeordnet. Es befasst sich mit der Entwicklung von Konzepten zur Formulierung von Freiheitsgraden von Algorithmen, die eine optimierte Ausführung auf einer Berechnungsplattform erlauben. Die besondere Anforderung dabei ist, dass die Abbildung auf eine konkrete Ausführungsplattform auf einfache Weise möglich sein soll. Diese Abbildung soll dann ausgehend von dem Konzept, das in diesem Arbeitspaket entsteht, in AP 5 geleistet werden.

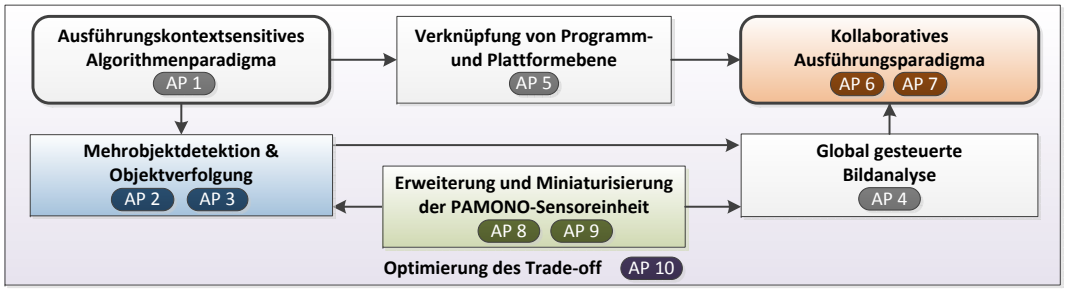


Abbildung 3.4.: Schematische Darstellung zum konzeptionellen Verbund der Arbeitspakete

Die erste Aufgabe dieses Arbeitspakets ist die grundsätzliche Entwicklung des Konzepts. Dies soll auf Funktionenebene angesiedelt und damit benutzungsfreundlich sein. Eine Funktion ist beispielsweise die Repräsentation durch Kernel-Programme in OpenCL. Es müssen nützliche Freiheitsgrade identifiziert werden, z. B. für die parallele Ausführbarkeit von Funktionen, die über Synchronisationsmechanismen gesteuert werden kann, so wie dies auch in OpenCL möglich ist. Ein Beispiel für eine weitergehende Möglichkeit ist die Angabe von algorithmischen Alternativen zur Realisierung der gegebenen Eingabe-Ausgabe-Funktionalität einer Funktion. Diese Möglichkeit kann dann bei der Abbildung auf eine konkrete Plattform im Hinblick auf die Optimierung der Ressourcenausnutzung und die Verfügbarkeit von Ressourcen genutzt werden. Solche Funktionsalternativen können von berechnungslastig bis speicherlastig reichen. Ein weiterer Aspekt ist der Granularitätsgrad der Funktionen, der alternativ gewählt werden kann.

Die zweite Aufgabe besteht in der softwaretechnischen Formulierung der Freiheitsgrade. Die Alternativen hierzu reichen von der Bereitstellung einer speziell entworfenen Programmiersprache (Vorbild HALIDE) bis zur Bereitstellung einer speziellen Programmbibliothek (z. B. für C++ und OpenCL), so wie etwa OpenCL eine Programmbibliothek für massiv-parallele Berechnungen für gängige Programmiersprachen wie etwa C++ darstellt.

Die dritte Aufgabe ist die prototypische Implementierung der gewählten Alternative, die eine Schnittstelle zur Realisierung der Abbildung auf konkrete Plattformen anbieten soll, die dann in AP 5 genutzt wird.

Arbeitspaket 2. Erkennung verschiedener Typen von Nanoobjekten: Dieses Arbeitspaket ist dem TZ 1 zugeordnet. Als Basismethoden sollen der GPU-basierte Erkennungs- und Optimierungsansatz aus der ersten Förderperiode sowie das Konzept der ausführungssensitiven Algorithmen aus AP 1 aufgegriffen werden.

Die betrachteten Anwendungen ergeben sich aus Erweiterungen des PAMONO-Sensors in AP 8: Nanopartikel im Medium Luft (Aufgabe 2 – Erkennung, Aufgabe 3 – Konzentration) sowie Erkennung von Nanoobjekten von verschiedenen, vorgegebenen Typen (Aufgabe 4 – unterschiedliche Partikeltypen).

Die erste Aufgabe ist die Beschaffung von Daten der neuen PAMONO-Sensorvarianten in Kooperation mit AP 8. Darauf aufbauend sind dann klassifizierte Testsequenzen zu realisieren.

Die zweite Aufgabe ist die Ausarbeitung von Lösungsverfahren für die beide Anwendungsaufgaben in Verbindung mit AP 8. AP 8 strebt dabei verbesserte Sensordaten an, während in diesem Arbeitspaket Kriterien für die Qualität der Sensordaten erarbeitet werden.

Die dritte Aufgabe betrifft die algorithmische Umsetzung der beiden Basismethoden mit dem Ziel der Echtzeitverarbeitung unter Beachtung des Ressourcenverbrauchs für die beiden Anwendungen. Dabei folgt die Analyse und Optimierung des Ressourcenverbrauchs bei der GPU-basierten Implementierung dem Ansatz der ersten Förderperiode. Für das Konzept der ausführungskontextsensitiven Lösung ist eine Implementierung durchzuführen. Diese bildet die Grundlage für die Analyse und Optimierung auf Grundlage der Offloading-Umgebung aus AP 7. Die Methoden erscheinen auch für andere Anwendungen der Bildanalyse relevant. Im SFB betrifft dies etwa astrophysikalische Daten (TP C3). Solche anderen Anwendungen sollen analysiert und bei Eignung selektiv bearbeitet werden.

Arbeitspaket 3. Tracking nicht ortsständiger Nanoobjekte: Dieses Arbeitspaket ist ebenfalls dem TZ 1 zugeordnet. Gegenstand ist die Bearbeitung der neuen Aufgabe 5 – Auffinden und Verfolgen der Trajektorien von sich bewegenden Nanoobjekten.

Objekt-Tracking-Verfahren stellen im Kontext des Projekts einen neuen Verfahrenstyp dar, sodass die existierenden Verfahren zunächst grundsätzlich in Bezug auf die relevanten Leistungsaspekte (Rechenzeit, Speicherbedarf, Energiebedarf) analysiert werden sollen. Objekt-Tracking-Verfahren bestehen in der Regel aus zwei Teilen: dem Auffinden von neu auftretenden Objekten und dem Verfolgen von vorhandenen Ob-

jekten. Das Verfolgen wiederum lässt sich in die Teile „Objektzuordnung“ und „Prädiktion“ trennen. Bei der Objektzuordnung lassen sich bildbasierte und merkmalsbasierte Ansätze unterscheiden. Die Prädiktion geschieht üblicherweise mit stochastischen, Bayes-basierten Verfahren, in Ausprägungen wie dem Kalman-Filter und seinen Erweiterungen sowie Partikelfiltern. Aufgrund der Artefakte der betrachteten Bildsequenzen sind die merkmalsbasierten Ansätze der Objektzuordnung sowie die stochastischen Prädiktionsverfahren von besonderem Interesse. Der nächste Schritt ist die Definition einer generischen Verarbeitungskette. In deren Rahmen sind ausführungskontextsensitive Algorithmen im Sinn von AP 1 zu formulieren und Steuerparameter für die Optimierung der Algorithmen zu identifizieren.

Im Zusammenhang mit der Objekterkennung ist die Trajektorienbestimmung als ein Teil der Musterextraktion zu verstehen, da diese als weiteres Merkmal für die Klassifikation der Nanoobjekte verwendet werden können. Basierend auf diesem Ansatz sind Klassifikatoren auszuwählen, für die Optimierung zu konfigurieren und im Sinne des Konzepts der ausführungskontextsensitiven Algorithmen zu implementieren.

Für die PAMONO-Anwendung der nicht ortsständigen Partikel sind in Kooperation mit AP 7 reale Testsequenzen und darauf aufbauend synthetische Testsequenzen zu erstellen. Letzteres erfordert ein neues Sensordatenmodell. Die erstellten Implementierungen bilden die Grundlage für die Analyse und Optimierung basierend auf der Offloading-Umgebung aus AP 7.

Arbeitspaket 4. Global gesteuerte Bildanalyse: Dieses Arbeitspaket ist dem TZ 2 zugeordnet. Die Qualitätskontrolle soll auf dem Sammeln von Daten über das Verhalten von sich im Betrieb befindlichen Sensoren auf einem Serversystem beruhen, auf dem auch die Verfahren der Qualitätskontrolle ausgeführt werden. Angestrebte Zwecke der Qualitätskontrolle sind die Erkennung von Funktionsfehlern des Sensors sowie die Gewährleistung und Optimierung der Funktionsfähigkeit durch Steuerung von Kompensations- und Kalibrierungsverfahren. Ein wesentliches zu lösendes Problem ist die Festlegung der relevanten Daten unter Berücksichtigung der Anforderung der späteren Verwendung und des Zusatzaufwands zur Datensammlung insbesondere auf den Sensoreinheiten. Dabei besteht ein Konflikt zwischen dem Wunsch nach viel Information einerseits und geringem Zusatzaufwand andererseits, für den Lösungen zu finden sind. Diese ist anwendungsabhängig und soll an den genannten drei Zwecken der Qualitätskontrolle betrachtet werden:

Erkennung von Funktionsfehlern des Sensors: Der verfolgte Ansatz besteht darin, statistische Ausreißer in den gesammelten Daten zu finden.

Kompensationsverfahren: Zunächst wird bei der Entwicklung von Erkennungsalgorithmen grundsätzlich angestrebt, den Kompensationsbedarf klein zu halten. Nicht vermeidbarer Kompensationsbedarf soll auf drei Ebenen angegangen werden: bei den Bilddaten, bei den Merkmalen und beim Klassifikator. Es sollen generische Kompensationsverfahren geschaffen werden, die über Parameter konfigurierbar sind. Die Konfiguration soll anhand der gesammelten Daten aus Anwendungen gelernt werden.

Kalibrierverfahren: Es sollen zwei Aufgaben betrachtet werden. Die erste Aufgabe ist das Feststellen der Notwendigkeit der Kalibrierung. Dies soll einerseits anhand von statistischen Funktionsfehlern und andererseits mit speziellen Kalibriermustern, die für den Fall des PAMONO-Sensors im AP 9 zu realisieren sind, geleistet werden. Die zweite Aufgabe ist die Durchführung der Kalibrierung. Hierzu sollen generische Kalibrierverfahren entwickelt werden, deren Parameter anhand der gesammelten Daten aus Anwendungen gelernt werden.

Arbeitspaket 5. Verknüpfung von Programm- und Plattformebene: Das Arbeitspaket ist dem TZ 3 zugeordnet. Die wesentliche Aufgabe besteht in der Verknüpfung des Konzepts der ausführungskontextsensitiven Algorithmen mit dem Scheduling, das die Ausführung auf Berechnungsplattformen organisiert. Dazu ist ein flexibler Scheduler zu konzipieren und zu implementieren, der die Schnittstelle zur Programmierumgebung aus AP 1 bedient. Dabei sollen zwei Wege eingeschlagen werden, die von den Gegebenheiten auf der Zielplattform abhängig sind. Die erste Gegebenheit ist die Zugänglichkeit des Schedulers der vorhandenen Plattform, um eigene Veränderungen anzubringen. Die zweite Möglichkeit ist, einen eigenen Scheduler über den vorhandenen Scheduler zu setzen, der diesen dann zu der gewünschten Reaktion bewegt. Bei der Wahl von OpenCL als Systemumgebung ist der zweite Weg einzuschlagen.

Arbeitspaket 6. Client/Server-Konfiguration und kollaborative Berechnung: Das Arbeitspaket befasst sich mit der Erarbeitung von Konzepten zur kooperativen und kollaborativen Berechnung bei gleichzeitiger Berücksichtigung von Ressourcenschranken sowie variablen Client/Server-Architekturen und Netzwerk-Konfigurationen. Es ist dem TZ 4 zugeordnet.

Die Sicherstellung von zeitnahen Antworten für die verteilte Berechnung, u. a. über das Prinzip des Offloadings, stellt für wandelbare Konfigurationen von mehreren (mobilen) Systemen (z. B. Biosenoreinheiten, s. AP 9), welche (wahlweise) sowohl als Client und Server agieren können, besondere Herausforderungen dar. Im Zusammenhang mit dem Offloading ist die Verarbeitungszeit, insbesondere die Einhaltung harter Zeitschranken, bedeutsam. Untersucht werden sollen Offload-Mechanismen zur Identifikation tragfähiger Kompromisse zwischen lokaler und verteilter Berechnung und Scheduling-Mechanismen mit statischer und dynamischer Priorisierung bei entsprechenden Zeitschranken. Hierbei sind Konzepte zur Kompensation bei der Nichteinhaltung von Antwortzeitvorgaben bedeutsam. Auch sind Konfigurationen mit expliziten Servern zu berücksichtigen, die über eminente Speicher- und Berechnungsressourcen verfügen. Im Sinne der Nachhaltigkeit sollen verschiedene Ausprägungen von Servern (z. B. GPGPU-Farmen und Hoch-Durchsatz Server) beachtet werden.
Für diese Fragestellung sind Konzepte zur Ressourcenreservierung und zu Virtualisierungsschematas zu entwerfen, speziell auch für entsprechende GPU-gestützte Server und Hoch-Durchsatz Server. Zur Untersuchung der damit assoziierten unterschiedlichen Ausführungscharakteristiken im Trade-off mit den zu beachtenden Ressourcen (z. B. Energieverbrauch) sollen Methoden zum kollaborativen Ressourcenmanagement und zur kollaborativen Datenanalyse (s. a. AP 4) entworfen und im Hinblick auf das AP 7 und AP 10 optimiert werden. In Anbetracht der zu entwickelnden (autonomen) mobilen Biosensoreinheiten (AP 9) mit entsprechenden eingebetteten Architekturen ist auch die faire Verteilung von Ressourcen und von Methoden zur Sicherstellung dieser relevant. Die Entscheidung über die Auslagerung ist zudem vereinheitlicht mit der Kommunikationskonfiguration – niedrige (*WirelessHART*), mittlere (WiFi) und hohe Bandbreite (SFP+, QSFP+, 10Gb Netzwerk) und allgemein der Netzwerkkonfiguration zu berücksichtigen.

Arbeitspaket 7. Optimierung der Ressourcenverwendung: Ausgehend von denen im AP 6 entwickelten Mechanismen zur kollaborativen Berechnung bei skalierbaren Ressourcen werden in diesem Arbeitspaket Konzepte zur mehrkriteriellen Optimierung unter Beachtung wandelbarer Netzwerk- und Systemkonfigurationen erarbeitet – das Arbeitspaket ordnet sich damit in das TZ 4 ein.
Unter Beachtung entsprechender eingebetteter Architekturen und dem Offloading von Berechnungen, sollen neue Methoden zur optimierten Zuweisung von Ressourcen erarbeitet werden, um gleichzeitig den Energieverbrauch, die Erkennungsleistung, die Verarbeitungszeit von Echtzeitanwendungen (AP 2) auf ressourcenbeschränkten eingebetteten Systemen (AP 9) zu verbessern. Dies beinhaltet konzeptionell auch die multikriterielle Optimierung des Plattformentwurfs im Trade-off zwischen multiplen (eventuell verteilten) Ressourcen. Dabei soll sowohl aufseiten der Clients als auch aufseiten des Servers (eventuell auch mehrerer Server) der Energieverbrauch, mit der Nebenbedingung der Verarbeitungszeit, optimiert werden. Das umfasst eine dynamische Verwaltung der Leistungsaufnahme, um zu entscheiden, wann Komponenten an- oder abgeschaltet werden sollen und eine Energieverwaltung mit dynamischer Spannungs- und Frequenzskalierung. Speziell bei mobilen Sensoreinheiten ist mit der Temperatur eine weitere Ressource zu beachten, da eingebettetes System und Sensor eine Funktionskomponente bilden und komplexe Berechnungen auch hohe Temperaturen innerhalb der Berechnungseinheit verursachen – dies kann die Genauigkeit des Sensors beeinflussen. Daher soll ein dynamisches Temperaturmanagement realisiert sowie untersucht werden, in welcher Form sich kollaborative Berechnungen auf die Temperatur auswirken. Unter dieser Sichtweise sind auch die im AP 7 identifizierten Schedulingmethoden zur parallelen Auslagerung unter vereinheitlichter Beachtung von Berechnungs- und Kommunikationsressourcen sowie Echtzeitvorgaben zu optimieren. Zur Optimierung sollen Simulations-/Emulations-zentrierte Methoden für adaptive und skalierbare reale, realkonforme und synthetische (zukünftige) Netzwerkkonfigurationen und Client-Server-Architekturen identifiziert bzw. entwickelt werden. Auf einer ersten Ebene sind ausgehend von der Untersuchung mobiler Systeme, u. a. mit einer Big.Little-Architektur und den Betriebssystemen Linux oder Android, auch GPU-gestützte Server und Hoch-Durchsatz Server bedeutsam, welche die limitierten Berechnungsressourcen eingebetteter Systeme unterstützen können. Auf einer zweiten Ebene sollen, zur Sicherstellung der Verfügbarkeit einer größeren Variationsvielfalt eingebetteter Systeme, modellbasierte Simulationen (u. a. mittels entsprechender MATLAB-Toolboxen) und Emulationen (z. B. angepasste *Hardware-in-the-Loop*-Verfahren zur Effektivierung der Berechnungszeit) genutzt werden. Dabei ist sowohl die Modellentwicklung selbst als auch die Adaptierung und Integration bestehender Modelle der Teilprojekte A3 und A4 ein Teilaspekt. Die konzeptionellen Ansätze bilden zudem eine Basis für die übergreifende Optimierung und Validierung innerhalb des AP 10.

Arbeitspaket 8. Generalisierung der Detektionseigenschaften: Das Arbeitspaket ordnet sich mit der Erweiterung der PAMONO-Methode im Hinblick auf die gleichzeitige Detektion unterschiedlicher Nanoobjekte, die Verwendung verschiedener Medien (Flüssigkeiten, Luft) und die Analyse nicht ortsständiger Nanoobjekte in das TZ 5 ein.

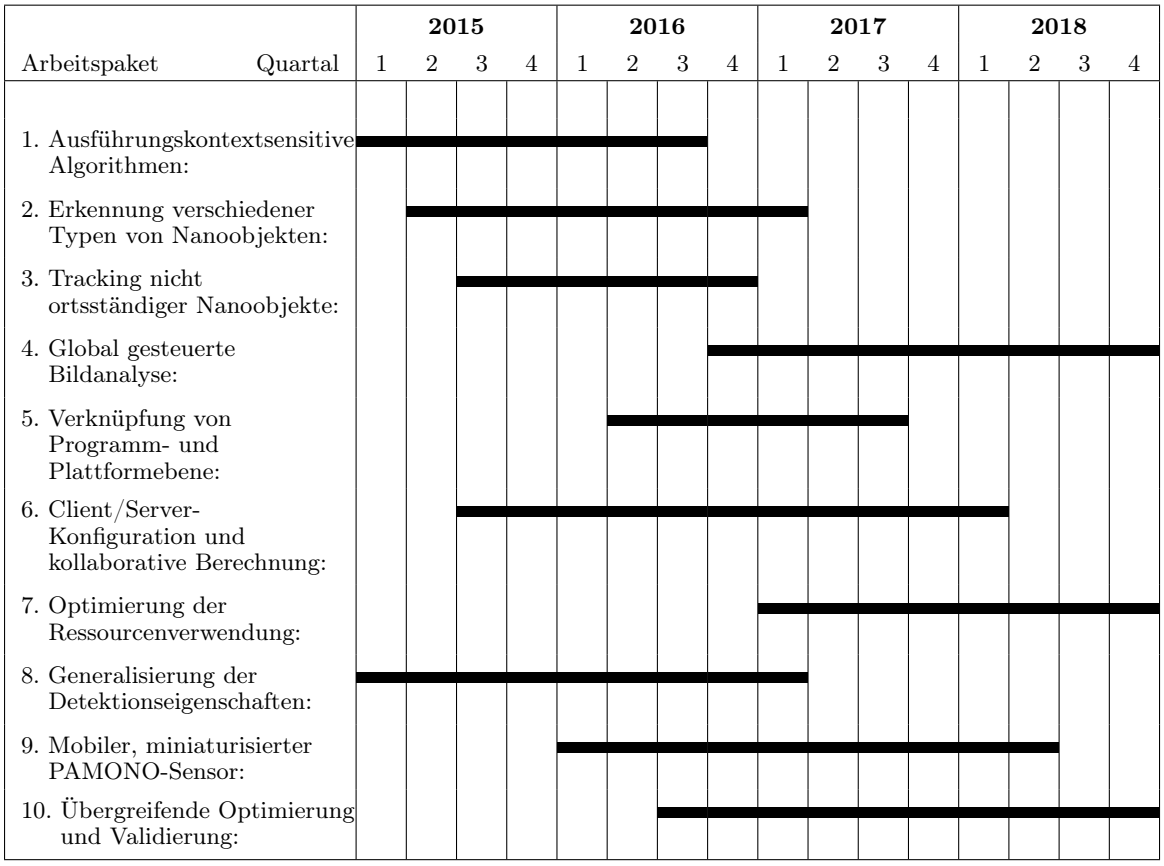
Zum Nachweis verschiedener Nanoobjekte innerhalb einer Probe müssen Sensorflächen mit Multispezifität zur Anwendung gebracht werden. Entsprechend verschiedene Antikörper werden hierzu als Array mit definierter Anordnung und Position auf der Sensoroberfläche angeordnet. Über in AP 4 zu entwickelnde Registrierungsverfahren und auf der Sensoroberfläche aufgebrachte Referenzmarkern ist eine exakte Zuordnung zwischen Sensoroberfläche und zugehöriger Bildebene zu erzielen. Zu untersuchen ist hierbei, welches Quantum an Multispezifität in Anbetracht der lateralen Auflösungsbegrenzung proklamiert werden kann. Gemäß Wissensstand der Antragsteller ist es aktuell technisch möglich, bis zu 30 Spots mit verschiedenen Antikörpern gleichzeitig zu verwenden, wobei aber eine entsprechende um den Faktor 30 schlechtere Konzentrationsempfindlichkeit vorliegt. Dies wird zu evaluieren und sowohl der technische Aufbau des Sensors (u. a. Verbesserung der Optik) als auch die bisher genutzten Immobilisierungsmethoden entsprechend zu modifizieren sein. Zudem ist aufseiten der Bildverarbeitung zu berücksichtigen, dass sich abweichende Antikörper auch in einer divergenten Charakteristik in den Sensordaten manifestieren (u. a. bezogen auf die räumliche Ausdehnung, Form, Signalintensität). Insbesondere bringt der Trennbereich zwischen den Spots (ca. 20 Nanometer) zusätzliche Artefakte (z. B. Unschärfe) in die Bilddaten ein. In Anbetracht von Nanoobjekten unterhalb von 50 Nanometern und bei einer gleichzeitigen Erweiterung auf verschiedene Medien stellt dies einen Sachverhalt dar, der bei der Bildanalyse besonders zu beachten ist (AP 2). Im Falle von Luft als Medium (z. B. zur Detektion von Rußpartikeln) ändert sich zusätzlich die zeitliche Charakteristik der Detektionsereignisse, welche sich nur noch durch Peaks ausprägen. Die Herausforderung ist, die korrekten Peaks aufgrund der globalen (Anhaftungs-)Ausprägung zu identifizieren.
Ergänzend zur bisher thematisierten analytischen bzw. diagnostischen Anwendung der plasmonenunterstützten Mikroskopie ist sie auch ideal für grundlegende Untersuchungen zum Transportverhalten von Viren, allgemein Nanoteilchen an Grenzflächen, geeignet. Dieser Aspekt ist aus Sicht der Medizin relevant, da bestimmte Viren (z. B. Influenzaviren) die Zellmembrane nur an definierten Stellen durchdringen können und die Viren zu diesen Membranpositionen „rollen“. Von Bedeutung sind Untersuchungen, die diesen Effekt genauer erklären und damit eine Beeinflussung der Durchdringung ermöglichen. Zur automatischen Detektion werden Verfahren zum bildbasierten Tracking der Nanoobjekte (auch unter Beachtung zeitweiser Verdeckung) und zur Ermittlung sowie Analyse der Bewegungsfelder (Trajektorien) benötigt. Diese werden in AP 3 entwickelt. Hierfür wird der Einsatz von Hoch-Apertur-Objektiven untersucht, welche zwar eine um 3-4 Größenordnungen schlechtere Nachweisstärke bei Konzentrationsmessungen aufweisen, aber für Untersuchungen von (Bewegungs-)Prozessen einzelner Nanoobjekte prädestiniert sind.

Arbeitspaket 9. Mobiler, miniaturisierter PAMONO-Sensor: Das Arbeitspaket befasst sich im Hinblick auf das TZ 5 mit der Weiterentwicklung des bisher stationären PAMONO-Sensors in Richtung eines mobilen Biosensors. Neben der grundlegenden Miniaturisierung des Sensors, sind Konzepte zur Kalibrierung und zum Monitoring gemäß AP 4 zu operationalisieren, aber auch zur Beschleunigung des Nachweisprozesses zu erarbeiten, um eine nachhaltige Einsatzfähigkeit in öffentlichen Räumen und bei Echtzeitanforderungen zu gewährleisten.
Zur Steigerung der Effektivität des Messvorgangs für die Miniaturisierung sollen im Unterschied zur ersten Förderperiode magnetische Nanoteilchen eingesetzt werden, um den aktiven Transport der untersuchten Nanoobjekte zur Sensorfläche zu ermöglichen. Eine aktive Bindung für diesen Sensortyp ist völlig neu und ermöglicht zudem die Messbarkeit um mehrere Größenordnungen niedrigere Konzentrationen. Hierzu ist die Probenaufbereitung in der Form abzuwandeln, dass die zu untersuchenden Objekte (Proteinmoleküle, Viren, aber auch Exosomen) zwei unterschiedliche Bindungsstellen besitzen – die erste Bindungsstelle dient zur Bindung des Objekts an das magnetische Teilchen, die zweite an die Sensorfläche. Der Bindungsprozess erfolgt um mehrere Größenordnungen schneller als bisher und die charakteristischen Muster der zur Detektion verwendeten Zeitreihen ändern sich aufgrund der abweichenden Geschwindigkeiten. Ferner tritt eine magnetisch durchgeführte Ablösung der ungebundenen Magnetteilchen im Anschluss an die aktive Bindung ein, wobei nur die durch spezifische Bindung gebundenen Teilchen an der Sensoroberfläche verbleiben und detektiert werden können. Dadurch ergibt sich eine zweistufige Ausprägung des Prozesses aus aktiver Bindung und selektiver Entbindung. Die Auswirkung dieser Effekte auf die Bildfolgen ist zu untersuchen und im Hinblick auf die automatische Erkennung geeignet zu beeinflussen. So sind differenzierte Untersuchungen zum Trade-off zwischen magnetischer Innervierung und echtzeitkonformer Detektion durchzuführen. Ferner sind die resultierenden Anforderungen, für die Bildanalyseverfahren von AP 2, zu spezifizieren.
Im Hinblick auf die Miniaturisierung sind gleichzeitig Abwandlungen der optischen und mechanischen Eigenschaften durchzuführen. Für die global gesteuerte Bildanalyse (AP 4) sind Konzepte zur Kalibrierung und zum Online-Monitoring in den Messvorgang zu integrieren. Es müssen hierzu proaktive Voting-Methoden entwickelt werden, welche trotz der ausgeprägten Artefakte den Grad der Konformität bestimmen. Untersucht werden soll u. a. der Einsatz standardisierter Proben (definierte Nanoobjekte). Es sind zudem Kon-

zepte des Online-Condition-Monitoring zu nivellieren, welche durchgehend die Korrektheit des Gerätes, die Handhabung und Messung selber überprüfen und eventuelle Retardierungen und Läsionen durch z. B. Temperaturschwankungen oder Erschütterungen innerhalb der Messungen als Artefakte identifizieren können. Ferner ist zu verifizieren, ob die Trägerplättchen noch eine ausreichende Bindung erlauben.

Arbeitspaket 10. Übergreifende Optimierung und Validierung: Aufbauend auf den Realisierungen der algorithmischen Konzepte in AP 2 und 3, der Systemkonzepte in AP 4 bis 7 sowie von Sensorvarianten in AP 8 und AP 9 wird in diesem Arbeitspaket die funktionelle und ressourcenbezogene Leistungsfähigkeit des Gesamtsystems analysiert und optimiert. Die Optimierung erfolgt bezüglich des Trade-off im Hinblick auf Kriterien der Erkennungsleistung, der Verarbeitungszeit und des Ressourcenbedarfs. Bezüglich der Optimierung soll die methodische Vorgehensweise des entsprechenden Arbeitspakets der ersten Förderperiode weiter verfolgt werden. Diese ist simulationsbasiert. Ein Herausforderung stellt der potenziell große Suchraum dar. Es soll untersucht werden, ob mittels hierarchischer Optimierungsansätze, etwa die Trennung in eine Grob- und Feinoptimierung, eine Effizienzsteigerung erzielt werden kann. Neben der simulationsbasierten Optimierung soll ferner an einem aufgebauten Prototypen eine messbasierte Optimierung erfolgen.

Zeitplan



3.5 Stellung innerhalb des Sonderforschungsbereichs

Bezüglich Systemsoftware-bezogener Aspekte des Offloading sowie zur Auslotung des Konzepts des ausführungskontextsensitiven Algorithmenentwurfs für Datenanalysealgorithmen soll mit TP A1 kooperiert werden. Mit TP A2 soll in Bezug auf grundlegende Konzepte des Algorithmenentwurfs, die den Ausführungskontext einschließen, zusammengearbeitet werden. Mit TP A3 ergibt sich, ausgehend von der Ausrichtung von TP B2 in der ersten Förderperiode, eine Zusammenarbeit in Bezug auf die Abbildung von Algorithmen und Programmen auf Multiprozessoren sowie das Online-Monitoring von Ressourcen. Für TP B2 sind die Arbeiten von TP A4 zu Ressourcenmodellen für die Hardware von Sensorknoten, zu Methoden zur Handhabung realistischer Plattformszenarien, zum Ausschöpfen von Optimierungspotenzialen durch schichtenübergreifende

Betrachtung der Plattform, zur kontextberücksichtigenden Planung und zur Prognose von Strategien und Systemverhalten und zur kontextsensitiven Optimierung von Systemparametern von Interesse, wohingegen TP B2 durch die Abbildung von Algorithmen auf Plattformen und zum Echtzeit-Offloading zu den Arbeiten von TP A4 nutzbringend beitragen kann. TP A6 kann zur Bildsegmentierung im Zusammenhang mit dem dortigen Einsatz von Graphikkernen beraten werden. Mit TP B1 ist ein Austausch zur Verkleinerung von Biosensoren hilfreich. Die bestehende Zusammenarbeit mit TP C3 zu stark artefaktbehafteten Bildfolgen in der Astrophysik soll fortgesetzt werden.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

Generating and Executing Dependable Application Software on UnReliable Embedded Systems (Chen,Förderkennzeichen DFG SPP 1500, CH985/4-1) (Förderzeitraum: 03.2013 - 02.2015)

Dieses Projekt untersucht, wie die Zuverlässigkeit durch Softwaretechniken, die in Verknüpfung mit Hardwaretechniken angewandt werden können, mit minimalem Zusatzaufwand zu verbessern ist. Damit adressiert das Projekt ein Thema, das keinen Bezug zum TP B2 hat.

Geometrieverarbeitung für die virtuelle Realisierung produktionstechnischer Prozesse (Müller,Förderkennzeichen SFB 708, TP B1, B4) (Förderzeitraum: 01.2011 - 12.2014)

Der Schwerpunkt der TP B1 und B4 des SFB 708 „3D-Surface Engineering für Werkzeugsysteme der Blechformteilefertigung – Erzeugung, Modellierung, Bearbeitung“ liegt auf der Modellierung und Simulation von Prozessen in der Fertigungstechnik, die keinen Bezug zu diesem Vorhaben haben.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Postdoktoranden, 100%	1	63.600	1	63.600	1	63.600	1	63.600
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	181.200	—	181.200	—	181.200	—	181.200
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Geräte, Verbrauchsmaterial:	8.800		1.200		1.200		1.200	
Summe:	8.800		1.200		1.200		1.200	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	190.000		182.400		182.400		182.400	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (B2)							
Wissen-schaftliches Personal	1	Müller, Heinrich, Prof. Dr., Prof	Visual Computing	LS 7, Fak. Inf, TU Do	4,0	—	Grund-ausstattung
	2	Chen, Jian-Jia, Prof. Dr., Prof	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	4,0	—	Grund-ausstattung
	3	Hergenröder, Roland, Dr., Grp.-Leiter	Biomedizinische Analytik	ISAS, Dortmund	3,99	—	Grund-ausstattung
	4	Zybin, Alexander, Dr., PostDoc	Biomedizinische Analytik	ISAS, Dortmund	19,98	—	Grund-ausstattung
	5	Weichert, Frank, Dr., PostDoc	Visual Computing	LS 7, Fak. Inf, TU Do	5,99	—	Grund-ausstattung
	6	Kehrt, Thomas, Dipl.-Ing. WiMi	Visual Computing	LS 7, Fak. Inf, TU Do	7,99	—	Grund-ausstattung
	7	N.N. , WiMi	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	17,98	—	Grund-ausstattung
	8	Görtz , Alexander, SHK	Visual Computing	LS 7, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	9	N.N. , SHK	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	10	Rapp, Heike, Verw. Ang.	—	LS 7, Fak. Inf, TU Do	0,99	—	Grund-ausstattung
	11	Graute, Claudia, Verw. Ang.	—	LS 12, Fak. Inf, TU Do	0,99	—	Grund-ausstattung

Beantragte Personalmittel (B2)							
Wissen-schaftliches Personal	12	Shpacovitch, Victoria, Dr., PostDoc	Biomedizinische Analytik	ISAS, Dortmund	—	Postdoktorand/in	—
	13	Siedhoff, Dominic, Dipl.-Inf. WiMi	Visual Computing	LS 7, Fak. Inf, TU Do	—	Doktorand/in	—
	14	Huang, Kevin Wen-Hung, M. Eng. WiMi	Eingebettete Systeme	LS 12, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Müller, Heinrich**
   Projektleitung. Schwerpunkt Visual Computing. Mitarbeit in den Arbeitspaketen 1, 2, 3, 4, 10.

2. **Chen, Jian-Jia**
   Projektleitung. Schwerpunkt Eingebettete Systeme. Mitarbeit in den Arbeitspaketen 5, 6, 7, 10.

3. **Hergenröder, Roland**
   Projektleitung. Schwerpunkt Biomedizinische Analytik. Mitarbeit in den Arbeitspaketen 8, 9, 10.

4. **Zybin, Alexander**
   Herr Zybin ist in der laufenden Förderperiode einer der Leiter des Projekts. Nach dem bevorstehenden Eintritt in den Ruhestand wird er in der zweiten Förderperiode im Rahmen eines Zeitarbeitsvertrags beratend an der Weiterentwicklung des PAMONO-Sensors mitwirken.

5. **Weichert, Frank**
   Mitarbeit in allen Bereichen des Projekts.

6. **Kehrt, Thomas**
   Herr Kehrt unterstützt bei der Erarbeitung und Realisierung der Sensor-assoziierten Algorithmen.

7. **N.N.,**
   Mitarbeit bei den plattformbezogenen APs 5, 6, 7.

8. **Görtz , Alexander**
   Implementierung von Bildverarbeitungsalgorithmen.

9. **N.N.,**
   Unterstützung bei der Erstellung und Wartung der Systemsoftware der eingesetzten Rechnersysteme.

10. **Rapp, Heike**
    Dokument- und Terminverwaltung sowie ggf. Reiseplanung und -abrechnung.

11. **Graute, Claudia**
    Dokument- und Terminverwaltung sowie ggf. Reiseplanung und -abrechnung.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

12. **Shpacovitch, Victoria**
    Frau Shpacovitch führt ihre Arbeiten der ersten Förderphase an dem PAMONO Sensor fort in den Arbeitspaketen 8 und 9.

13. **Siedhoff, Dominic**
    Algorithmen zur Objekterkennung und Arbeitspakete 1 bis 4.

14. **Huang, Kevin Wen-Hung**
    Echtzeitfähiges Offloading, Arbeitspakete 5, 6 und 7.

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	6.500	6.500	6.500	6.500
ISAS: vorhandene Mittel aus Haushalt	6.000	6.000	6.000	6.000
Summe der vorhandenen Mittel	12.500	12.500	12.500	12.500
beantragte Ergänzungsausstattung	8.800	1.200	1.200	1.200

Geräte bis 10.000 EUR, Software und Verbrauchsmaterial	**für 2015**
Biologische Reaktive 2x Biotin-thiol (je 850 Euro): 1.700 Euro; 8x Biotin-konjugierter primärer Antikörper (je 350 Euro): 2.800 Euro	EUR 4.500
Optik, Elektronik, Mechanik zum Aufbau eines neu konzipierten Virensensors Stromkontroller für Diodenlaser, LDC 205C: ca. 900 Euro; Temperaturkontroller für Diodenlaser, TED 200C: ca. 900 Euro; CCD-Kamera: ca. 700 Euro; Linsen, Spiegel, Optik: ca. 200 Euro; Drehtisch: ca. 400 Euro	EUR 3.100

Geräte bis 10.000 EUR, Software und Verbrauchsmaterial	**für 2015 – 2018**
Beschichtete Glasplättchen Die mit Gold beschichteten Glasplättchen sind für Experimente von Nanoteilchen und speziellen VLPs notwendig.	EUR je Jahr 1.200

3.1 Allgemeine Angaben zum Teilprojekt B3

3.1.1 Titel:

Data Mining in Sensordaten automatisierter Prozesse

3.1.2 Fachklassifizierung:

Data Mining/Maschinelles Lernen, Produktionssysteme

3.1.3 Projektleitung

Prof. Dr.-Ing. Deuse, Jochen, 29.12.1967

Institut für Produktionssysteme, Professur Arbeits- und Produktionssysteme, Fakultät Maschinenbau, Technische Universität Dortmund
Leonhard-Euler-Straße 5
44227 Dortmund

Tel.: 0231-755-2652
Fax : 0231-755-2649
E-Mail: `jochen.deuse@ips.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Morik, Katharina, 14.10.1954

LS 8, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 12
44227 Dortmund

Tel.: 0231-755-5100
Fax : 0231-755-5105
E-Mail: `katharina.morik@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Die individuelle Qualitätsprognose von Erzeugnissen ermöglicht die Vermeidung interner Fehlerkosten, wenn sie realzeitlich im laufenden Produktionsprozess geschieht. Eine solche Prognose basiert auf Sensordaten an verschiedenen Punkten des Produktionsprozesses. Für die Entscheidung und sich daraus ableitenden Maßnahmen stehen in starr verketteten Prozessen oft nur wenige Sekunden zur Verfügung. Im Rahmen des Teilprojekts B3 wird untersucht, wie dezentrales Data Mining unter der Ressourcenbeschränkung Zeit zur Prognose der Produktqualität genutzt und in die Produktion eingebunden werden kann.

Die Qualitätssicherung in Produktionsprozessen beruht darauf, Zustände zu erkennen, die innerhalb akzeptabler Toleranzgrenzen liegen, so dass das Endprodukt eine vorgegebene Spezifikation erfüllt. Anhand von Qualitätsprognosen, die auf Daten von an ausgewählten Stellen im Produktionsablauf platzierten Sensoren basieren, sollen frühestmöglich Entscheidungen zur Steuerung und Anpassung des Prozessablaufs getroffen werden. Ein Beispiel ist die Ausschleusung fehlerhafter Erzeugnisse im laufenden Produktionsprozess. Erst die maschinelle Verarbeitung von Sensordaten mit Methoden der Datenanalyse kann solche Vorgänge automatisieren. Die Prognosen, Entscheidungen und die Steuerung von Produktionsanlagen unterliegen dabei zeitlichen Restriktionen im Bereich von wenigen Millisekunden bis Sekunden. Insbesondere erfordern starr verkettete Produktionsprozesse Reaktionen auf Fehlerzustände in Realzeit. Als Alternative zur zentralen Speicherung und Verarbeitung aller Daten untersucht das Teilprojekt, ob und wie sich durch dezentrale Datenanalyse Reaktionszeiten verkürzen, das zu übertragene Datenvolumen vermindern und Produktionssysteme selbstadaptiver gestalten lassen. Die dezentrale Qualitätsprognose wirft dabei neue Probleme auf, die im traditionellen Data Mining so nicht behandelt werden.

In der ersten Förderperiode wurden qualitätsrelevante Prozessschritte in einem realen Produktionsprozess zum Warmwalzen von Stahl identifiziert, Sensoren installiert und Daten für Analysen aufgezeichnet. Es wurde ein Verfahren für das neuartige Lernproblem entwickelt, aus aggregierten Werten einer Zielgröße (dem sog. *Label*), die für Gruppen von Beobachtungen gegebenen sind (hier Qualitätsinformationen zu Aufträgen von Produkten) ein Prognosemodell für individuelle Beobachtungen abzuleiten. Die Beobachtungen sind dabei durch sog. *Merkmale* repräsentiert, die charakteristische Eigenschaften der Prozesse in Form diskreter oder numerischer Werte ausdrücken. Für die Aggregation von Werten und die Ableitung von Merkmalen aus den ursprünglich gegebenen Messreihen wurden generische sowie individuelle Prozesse in der Data Mining Software RapidMiner definiert. Anhand der extrahierten Merkmale konnten mittels Clustering wesentliche Produktionsmodi identifiziert werden, deren numerische Charakterisierung zur Normalisierung der Messreihen und zur Definition von Normabweichungen genutzt werden soll. Des Weiteren wurde die 1-class-Variante der Core Vector Machine (CVM) so modifiziert, dass sie mit geringen Kommunikationskosten auf verteilten Sensordaten arbeiten kann. Für die Integration der Methoden in Produktionsprozesse wurde das Konzept der Intelligent Manufacturing Control entwickelt.

Das langfristige Ziel des Projekts besteht in der gezielten Anpassung von Steuerungsparametern, um Prozessfehler rechtzeitig zu verhindern. Eingriffe in die Prozesssteuerung beschränkten sich in der ersten Förderperiode auf binäre Entscheidungen und Prognosen auf einen einzelnen Wert zur Beschreibung der Qualität. Für gezielte Eingriffe in die Steuerung sollen die Arbeiten aus der ersten Förderperiode um eine detailliertere Fehlerprognose mittels Simulation gewonnener Daten erweitert werden. Steuerungsentscheidungen sollen neben detaillierteren Fehlerprognosen außerdem auch Kundenspezifikationen einbeziehen. Die Arbeiten zur verteilten Datenanalyse werden fortgesetzt und um Untersuchungen zur verteilten Merkmalsextraktion, -selektion und Parameteroptimierung ergänzt.

3.3 Bisherige Entwicklung des Teilprojekts

3.3.1 Bericht und Stand der Forschung

Im Bereich der Serienfertigung werden nach dem Fließprinzip arbeitende Fertigungssysteme eingesetzt, die oftmals einen hohen Automatisierungsgrad aufweisen. Beispiele für derartige Fertigungssysteme sind u. a. in der Stahlproduktion oder der Elektronikfertigung zu finden.

Bei der automatisierten Fließfertigung führen Fehler, wenn sie erst am Ende der Wertschöpfungskette entdeckt werden, schnell zu großen Mengen an Ausschussteilen, welche aufwendig und kostenintensiv nachbearbeitet oder direkt verschrottet werden müssen. Zur Vermeidung solcher Fehler werden die Produktionsprozesse und die Qualität der bearbeiteten Zwischenprodukte fortlaufend überwacht [351]. So setzt die statistische Prozesskontrolle kontinuierlich erfasste Prozesscharakteristika und -parameter in Bezug zur beobachteten Qualität. Bei Abweichungen werden Prozessparameter entsprechend korrigiert [373]. Voraussetzung ist jedoch einerseits die Produktion einer größeren Anzahl gleicher oder ähnlicher Produkte, so

dass Abweichungen erkannt werden können. Andererseits gilt die implizite Annahme, dass qualitätsrelevante Eigenschaften am Produkt ohne Verzögerung und aufwendige Prozesse überprüft werden können. Viele Bearbeitungsprozesse, wie z. B. das Warmwalzen von Stahlriegeln, ermöglichen ein solches Vorgehen aufgrund der Charakteristika der Bearbeitungs- und Prüfprozesse nicht. Darüber hinaus resultiert die Qualität von Stahlprodukten insbesondere aus ihrer inneren (Gefüge-)Struktur. Bearbeitungstemperaturen von über 600°C lassen erst eine Prüfung am Ende der Prozesskette zu. Der Ansatz der statistischen Prozesskontrolle ist somit nicht zu realisieren.

Für die Überprüfung der Oberflächeneigenschaften wurden anwendungsfallübergreifend auf dem Gebiet der automatisierten optischen Kontrolle (engl.: automated optical inspection, AOI) unterschiedliche Lösungsansätze entwickelt [385, 391]. Dabei wird i. A. mit Hilfe von optischen Kamerasystemen oder Lasermesseinrichtungen die Oberfläche der Produkte erfasst und mittels Algorithmen der Bildverarbeitung hinsichtlich Rissen, Poren o. ä. analysiert. So können bereits während der Produktion geeignete Maßnahmen zur Korrektur der Qualitätseigenschaften der Oberfläche eingeleitet werden. Eine Beurteilung der inneren Gefügestruktur als zentrales Qualitätskriterium ist jedoch nicht möglich. Diese Problematik besteht ebenfalls für die in der Prozessindustrie entwickelte Model Predictive Control (MPC). MPC-Ansätze lösen für kontinuierliche Prozesse in jedem Kontrollzyklus ein mathematisches Optimierungsproblem, das basierend auf dem beobachteten Prozessergebnis und den aktuellen Prozessparametern die Anpassungskosten minimiert [360]. Wie der zuvor vorgestellte Ansatz der AOI erfordern auch die Verfahren der MPC eine Kenntnis und Bewertung der Qualität der hergestellten Produkte, um steuernd bzw. regelnd in den Prozess einzugreifen. Darüber hinaus müssen die nacheinander bearbeiteten Produkte vergleichbar sein. Sonst können die in einem Zyklus erfassten Qualitätseigenschaften des fertiggestellten Produktes nicht für die Anpassung der Bearbeitungsparameter des nächsten Produktes verwendet werden. Somit eignet sich keiner der zuvor genannten Ansätze für eine kontinuierliche Qualitätsüberwachung und -regelung.

Zentrale Anforderung an eine fortwährende Qualitätskontrolle ist die vollständige Erfassung der Produktqualität in jedem Stadium des Herstellungsprozesses, insbesondere in solchen Phasen, in denen eine Qualitätsprüfung am Produkt selbst nicht möglich ist. An dieser Stelle wurde bereits bei der Beantragung der ersten Förderperiode die Idee der in der Folge entwickelten Initiative „Industrie 4.0“ antizipiert. Die im Zuge von Industrie 4.0 verfolgte Vernetzung von Anlagen durch Informations- und Kommunikationstechnologie zu cyber-physischen Systemen mit dem Ziel der Überwachung, Steuerung und Optimierung komplexer Produktionssysteme beruht auf echtzeitfähigen Ansätzen, welche die in den cyber-physischen Systemen erhobenen Informationen auswerten und in Prozesswissen überführen. Die Arbeiten der ersten Förderperiode zielen auf genau solche Ansätze. Durch die Installation und Vernetzung von Sensoren zur Erfassung der Prozessparameter unterschiedlicher Anlagen sowie deren Auswertung sollen sekundäre Informationen über den Prozessverlauf expliziert werden. Die Qualitätseigenschaften der Produkte entstehen dabei aufgrund der Wechselwirkung vieler unterschiedlicher Prozessparameter, deren exakte Ausprägungen erst während des laufenden Prozesses ermittelt werden können. Der rechtzeitige Eingriff in die Steuerung nachfolgender Prozessschritte erfordert daher die Erfassung und Analyse der Sensordaten sowie die Identifikation kritischer Prozessmuster in Realzeit [B3/405]. Methoden der Datenanalyse werden zur Fehlerdiagnose zwar bereits eingesetzt [366], jedoch ohne realzeitliche Schranken zu berücksichtigen. Auch das in [361] beschriebene prozessintegrierte Qualitätsregelungssystem für stabile Fertigungsprozesse berücksichtigt keine Ressourcenbeschränkungen, insbesondere keine Realzeitanforderungen.

Für die erste Förderperiode wurden zehn Arbeitspakete definiert, die sich in die Problembereiche der Datenerhebung, dem Lernen aus unsicheren Labeln, der Datenaggregation und Merkmalsgewinnung, der Qualitätsprognose an einzelnen und mehreren Stationen und der Integration in Produktionsprozesse gliedern lassen. Die folgenden Abschnitte beschreiben zusammenfassend wesentliche Ergebnisse dieser Arbeitspakete und setzen sie in Bezug zu anderen aktuellen Ansätzen.

Datenerhebung (AP 1, AP 2 und AP 4): Es wurde die wesentliche Erkenntnis gewonnen, dass die Datenerhebung keinen vom Prozess der Datenanalyse unabhängigen Teil darstellt. Stattdessen sollten die geeignete Auswahl von Entscheidungspunkten, zu installierender Sensoren und die Repräsentation von Rohdaten für die Analyse als eigenes Forschungsproblem betrachtet werden. Es wurde daher eine Methodik vorgeschlagen [B3/400], die den Schritt der Datenerhebung ausdrücklich in Standardprozessen für die Datenanalyse berücksichtigt.

Für das Fallbeispiel eines starr verketteten Bearbeitungsprozesses, das Warmwalzen von Stahl, wurden mittels der entwickelten Methodik vier potentiell qualitätsrelevante Fertigungsschritte identifiziert. Zunächst werden Stahlriegel in einem aus fünf Zonen bestehenden Drehherdofen erwärmt. Danach erfolgt die Bearbeitung durch eine sog. Blockwalze und zwei Fertigwalzen. Am Ende werden die Riegel in Stäbe zerteilt und deren Qualität über eine Ultraschallprüfung bestimmt. Diese erlaubt Rückschlüsse auf Fehlerarten und

deren Ausprägung, nicht aber auf ihre exakte Position. Die Prognose der finalen Qualität und Steuerungsentscheidungen sollen nach jedem der genannten Fertigungsschritte erfolgen, und zwar anhand der in Realzeit erfassten Werte der Prozessparameter.

Qualitätsrelevante Prozessparameter wurden ausgehend von den physikalischen bzw. technologischen Gesetzmäßigkeiten des Walzprozesses bestimmt. Als Datenquellen dienen die Anlagensteuerungen, an denen die ex-ante definierten Prozessabläufe, die sog. Stichpläne, erfasst werden. Diese beinhalten neben der Anzahl der Umformschritte auch deren vorgegebene Steuerungsparameter. Neben den Soll-Größen können dabei auch Ist-Größen aufgezeichnet werden. Für den Zugriff auf die Daten aller relevanten Bearbeitungsprozesse wurde eine Schnittstelle zu den Maschinensteuerungen geschaffen. Dies erlaubt die Erfassung von Prozessparametern wie der Walzenanstellung, Drehzahl und verwendeten Kaliber in Echtzeit. Für die Erfassung der Werkstücktemperatur wurden zusätzliche Pyrometer an den Walzgerüsten angebracht. Die Erfassung der Umformkräfte erfolgt über Kraftmessdosen, die an den Walzgerüsten installiert wurden. Die Aufwärmprofile von Stahlriegeln im Drehherdofen lassen sich aus der computerbasierten Prozesssteuerung auslesen.

Zur Gewährleistung der Reproduzierbarkeit von Analysen werden die Prozessparameter nicht nur im laufenden Produktionsbetrieb erfasst, sondern zeitgleich in einer Datenbank abgelegt. Diese berücksichtigt dabei bereits Anforderungen an die nachfolgende dezentrale Verarbeitung der Daten, indem die Messwerte unterschiedlicher Prozessschritte getrennt voneinander ablegt werden. Bis heute sind so seit dem ersten Quartal des Jahres 2012 über eine Milliarde Messwerte gesammelt und im Rahmen des Projekts verfügbar gemacht worden. Die Daten lassen sich exportieren und mit in RapidMiner definierten Analyseprozessen weiter verarbeiten. Auf aufwändige Methoden zur Bereinigung der Daten konnte verzichtet werden, da bereits die einfache Aggregation durch Mittelwertbildung oder die Verwendung von Vorgängerwerten ausreicht, um fehlende Werte oder Ausreißer in den Messreihen zu ersetzen.

Das Ziel, einen Datensatz für Analysen zu gewinnen, wurde erreicht. Allerdings ist es bisher nicht gelungen, die Daten so zu anonymisieren, dass sie öffentlich zugänglich gemacht werden können. Für die entwickelten Datenanalyseverfahren stehen allerdings Datengeneratoren zur Verfügung. Die Sammlung von Daten soll in der zweiten Förderperiode fortgesetzt werden, bildet jedoch kein eigenes Arbeitspaket mehr. Zusätzliche Fehlerinformationen sollen stattdessen im Rahmen des zukünftigen Arbeitspakets zur detaillierten Fehlerprognose (AP 2) gewonnen werden.

Lernen aus unsicheren Labeln (AP 3): Für Stahlriegel mit eindeutiger Zuordnung zu den aus ihnen geschnittenen Stäben wurde eine Abbildung definiert [B3/398], die Fehler je nach Ausprägung unterschiedlich gewichtet und zu einem reellwertigen Wert pro Riegel aufsummiert. Diese Zielgröße kann mit gewissen Einschränkungen als prozentualer Anteil des fehlerhaften Materials an einem Riegel interpretiert werden. Aus technischen Gründen kann jedoch die zuvor genannte Zuordnung verloren gehen. Dann sind lediglich Prüfinformationen über die Stäbe eines gesamten Auftrags gegeben. Die daraus resultierende Lernaufgabe entspricht keinem üblichen Szenario. Überwachte Lernverfahren erwarten individuell gelabelte Beispiele. Unüberwachte Lernverfahren suchen nach Strukturen in einer Menge von Beobachtungen, für die keinerlei Label verfügbar sind. Semi-überwachte Lernverfahren treffen die Annahme, dass wenige individuell gelabelte Beobachtungen und viele ungelabelte Beobachtungen vorliegen. Am Ende von Produktionsprozessen hingegen liegen häufig nur statistische Informationen über die Label vor, z. B. die prozentualen Anteile eines Fehlers pro Auftrag. Ähnliche Zuordnungsprobleme treten auch in anderen Bereichen auf, etwa bei der Charakterisierung von Wählertypen anhand gegebener Wahlergebnisse. Zu dem neuartigen Lernproblem, bei dem Label-Proportionen für Gruppen von Beobachtungen aggregiert vorliegen, gibt es bisher nur einige wenige Ansätze.

Der im Rahmen des Projekts entwickelte Algorithmus *Learning from Label Proportions* (LLP) [B3/403] clustert erst die Beobachtungen mittels k-Means und weist den Clustern danach Label so zu, dass die mittlere quadratische Abweichung zu den gegebenen Label-Proportionen minimiert wird. Mit einer evolutionären Attributgewichtung erreichte das Verfahren auf Standarddatensätzen eine signifikant bessere Klassifikationsgenauigkeit als die besten bis dahin bekannten Verfahren (*AOC* [354], *MeanMap* [376] und *Inverse Calibration* [382], bei deutlich geringerer Trainingszeit. Zuvor nicht gesehene Beobachtungen können außerdem sehr schnell den durch k-Means bestimmten Zentroiden zugeordnet werden. Zudem können gelabelte Beispiele und auch mehr als zwei Klassen sehr einfach berücksichtigt werden. Der Ansatz zur evolutionären Attributgewichtung mit k-Means konnte erfolgreich auch auf andere Bereiche der Produktion, nämlich die Teilfamilienbildung, übertragen werden [B3/395, B3/396, B3/397]. Ein Vergleich von LLP mit den neuesten Arbeiten [364, 392], steht noch aus und soll bis zum Ende der ersten Förderperiode vorgenommen werden.

Die Idee der Verwendung von semi-überwachten transduktiven Ansätzen zum Transferlernen, wie der in [348] behandelte, wurde verworfen, da die Lernaufgabe für die in Produktionsprozessen erhobenen Daten

eher untypisch ist. Entweder ist es technisch möglich, Produkten Label zuzuordnen (dann sind alle Beobachtungen gelabelt) oder nicht (dann ist keine Beobachtung gelabelt).

Alle evaluierten Verfahren weisen eine hohe Varianz bezüglich unterschiedlicher Auswahlen von Gruppengrößen, Trainings- und Testmengen auf. Diese Instabilität ist problembedingt, da nicht genügend Informationen zur Bestimmung einer eindeutigen Lösung gegeben sind. Aus diesem Grund wird das Arbeitspaket zum Lernen aus unsicheren Labeln nicht fortgeführt. Stattdessen soll die Menge der bereits gelabelten Beispiele auf intelligente Weise mittels Ansätzen des Active Learnings und vorhandenen Simulationen erweitert werden (zukünftiges AP 3).

Bereinigung, Aggregation und Merkmalsgewinnung (AP 5, AP 6, AP 7): Ein einzelnes Lernbeispiel für die Qualitätsprognose besteht prinzipiell aus der Menge aller während eines Produktionsprozesses aufgezeichneter Wertereihen und dazugehörigen statischen Daten. Für jeden Schritt im Produktionsprozess können dabei mehrere Wertereihen erhoben werden, die unterschiedliche Skalenniveaus, Längen und Anzahlen semantischer Einheiten (z. B. Walzvorgänge) aufweisen. Es bieten sich prinzipiell viele verschiedene Möglichkeiten an, Zeitreihendaten der genannten Art zu repräsentieren.

Zunächst wurde der Ansatz verfolgt, eine Abbildung der Rohdaten auf Merkmalsvektoren fester Länge zu definieren und sie auf diese Weise einer Vielzahl unterschiedlicher Datenanalyseverfahren zugänglich zu machen. Vor dem Hintergrund der gegebenen zeitlichen Ressourcenbeschränkungen soll die Abbildung zum einen den von den Lernverfahren zu bewältigenden Datenumfang reduzieren, auf der anderen Seite aber wesentliche Eigenschaften der Wertereihen erhalten. Ein im Rahmen des Projekts entwickeltes Verfahren [B3/401] folgt dabei ähnlich wie Verfahren zur Klassifikation von Bilddaten der Idee, semantisch zusammenhängende Einheiten der Wertereihen zu identifizieren (hier z. B. Walzstiche) und diese durch Merkmale zu beschreiben (hier längeninvariante Merkmale wie beispielsweise Mittelwert, Varianz, Minimum und Maximum). Diese werden dann auf zuvor definierte Bereiche eines Vektors fester Länge abgebildet. Die Segmentierung erfolgt in diesem Fall anhand von Signalen, welche den Beginn eines neuen Bearbeitungsschrittes anzeigen. Über die abgeleiteten Merkmale können wiederum Merkmale gebildet werden, wodurch sich unterschiedliche Granularitätsstufen der Wertereihen gleichzeitig im selben Merkmalsvektor repräsentieren und auf ihren Zusammenhang mit der Zielgröße hin untersuchen lassen.

Das als modularer Prozess in RapidMiner realisierte Verfahren konnte die insgesamt bis zu 60.000 für einen Stahlriegel aufgezeichneten Messwerte auf 213 Merkmale reduzieren. Ein Clustering mittels selbstorganisierender Karten und k-Means liefert trotz der starken Aggregation sinnvolle Ergebnisse. So ist es gelungen, wesentliche Betriebsmodi des Produktionsprozesses zu identifizieren und quantitativ zu beschreiben. Die Ergebnisse konnten von Domänenexperten verifiziert werden. Das in Teilprojekt A1 entwickelte k-Means Verfahren zum Clustering von Zeitreihen mittels Dynamic-Time-Warping-Distanz (DTW) liefert qualitativ ähnliche Ergebnisse, benötigt aufgrund der Berücksichtigung aller Rohdaten jedoch deutlich mehr Zeit. Aus diesem Grund sollen andere auf den Rohdaten arbeitende Ansätz, etwa spezielle Kernfunktionen für Zeitreihen oder die Modellierung von zeitlichen Abhängigkeiten durch Graphkerne, aus der weiteren Betrachtung zunächst ausgeschlossen werden.

Auf den extrahierten Merkmalen trainierte Klassifikationsverfahren (k-NN, Entscheidungsbäume, Naïve Bayes, lineare SVM und SVM mit Radial-Basis-Kernfunktion) zur Vorhersage der Qualität des Endprodukts ergaben bisher keine signifikant höhere Genauigkeit als die Vorhersage der Mehrheitsklasse. Die häufigen Muster der identifizierten Betriebsmodi überlagen dabei möglicherweise lokale Muster, die für die Qualität relevant sein könnten. Auch geeignete Transformationen der Wertereihen automatisch mit dem Ansatz der *Methodenbäume* [372] evolutionär zu lernen oder der direkte Vergleich der Rohdaten mit von der Länge unabhängigen Distanzmaßen, wie DTW oder der Kreuzkorrelation, brachte keine Verbesserung.

Ohne eine entsprechende Normalisierung der Messreihen anhand der identifizierten globalen Muster ist nicht zu erwarten, dass erweiterte Ansätze wie die Symbolisierung von Zeitreihen und Bestimmung häufiger Wörter mittels *SAX* [383], das Mining *sequentieller Muster* [375], das Auffinden von *Motifs* [371], *Shapelets* [379], *modellbasierte Beschreibungen* [368] oder die Repräsentation von Zeitreihen durch *Codetabellen* [389] die Prognoseergebnisse wesentlich verbessern könnten. Bis zum Ende der ersten Förderperiode sollen die Wertereihen daher normalisiert und nur noch die Unterschiede zu den globalen Mustern in Merkmalen codiert werden. Zudem sollen die Arbeiten zur Aggregation und Merkmalsgewinnung in der zweiten Förderperiode im geplanten AP 1 fortgeführt und um verteilte Ansätze erweitert werden.

Qualitätsprognose an einzelnen und mehreren Stationen (AP 8): Das Lernen jeweils eines Prognosemodells anhand von Sensordaten der aktuellen Fertigungsstation und all ihrer Vorgänger kann als Lernproblem auf verteilten Daten aufgefasst werden. Die realzeitliche Analyse erfordert kommunikationseffiziente verteilte Algorithmen, welche die Zeit von Datenübertragungen reduzieren und damit schnellere Reaktionen im

laufenden Produktionsprozess ermöglichen können. In der Trainingsphase sollen insbesondere weniger Daten übertragen werden als bei der Übermittlung des gesamten Datensatzes an einen zentralen Knoten. Kommunikationseffizienz ist außerdem häufig ein entscheidender Faktor bei der Reduktion des Energieverbrauchs von Datenanalysealgorithmen in drahtlosen Sensornetzwerken [B3/394]. Derartige Netzwerke sind im Produktionsumfeld im Zuge von Industrie 4.0 beispielsweise für die Kommunikation zwischen Werkstücken bzw. Werkstückträgern und Produktionsanlagen oder auch in der Produktionslogistik (vgl. Teilprojekt A4) von Bedeutung.

Unterschieden wird zwischen der *horizontalen Verteilung*, bei der Beobachtungen mit ihren Merkmalen über mehrere Knoten verteilt sind, und der *vertikalen Verteilung*, bei der stattdessen die Merkmale der Beobachtungen auf unterschiedlichen Knoten vorliegen. In Produktionsprozessen sind Daten häufig vertikal verteilt, da an jeder Fertigungsstation nur Teile des gesamten Prozessverlaufs beobachtet werden können. Während für das horizontale Szenario verschiedene Datenanalysealgorithmen entwickelt wurden (beispielsweise [352], [18] oder [370]), existieren für das Szenario der vertikalen Verteilung nur einige wenige. Zunächst ähnelt das Szenario dem des *Multi-View Learnings*, bei dem heterogene Informationen aus unterschiedlichen Quellen (Views) vorliegen. Ansätze wie z. B. das *Co-Training* setzen allerdings die bedingte Unabhängigkeit der Views von der Zielgröße voraus. Diese Unabhängigkeit ist insbesondere bei Produktionsprozessen häufig nicht gegeben. Vertikal verteilte Ansätze im Bereich des *Privacy Preserving Data Mining*, wie zur SVM [393], oder vertikal verteilt arbeitende iterative Optimierungsverfahren, wie die *Alternating Direction Method of Multipliers* (ADMM) [353] oder der *Frank-Wolfe-Algorithmus* [352], übertragen unter Umständen mehr Daten als den ursprünglichen Gesamtdatensatz. Dies gilt in der Regel auch für die Modellierung von Sensornetzwerken durch *graphische Modelle* [350]. Vielversprechendere Ansätze könnten solche sein, die wie die in [356] beschriebene *1-class SVM* zur vertikal verteilten Anomalieentdeckung Daten sampeln.

Im Rahmen des Teilprojekts wurden zwei Verfahren für das Szenario vertikal verteilter Daten entwickelt. Die in Zusammenarbeit mit dem Teilprojekt C1 entwickelte verteilte SVM [B3/C1/399] löst auf jedem Knoten unabhängig voneinander ein lokales SVM-Problem. Die Gesamtvorhersage ergibt sich als gewichtete Summe der Vorhersagen aller lokalen Modelle. Während des Trainings werden keine Daten übertragen und bei der Anwendung pro Beispiel nur ein Wert pro Knoten. Ein Nachteil der Methode kann darin bestehen, dass sie die bedingte Unabhängigkeit der lokalen Merkmale von der Zielgröße voraussetzt. Die *Vertically Distributed Core Vector Machine* (VDCVM) [B3/404] hingegen entdeckt Anomalien in vertikal verteilten Daten. Der Algorithmus arbeitet (ähnlich wie [370]) lediglich auf einer repräsentativen Teilmenge der Trainingsbeispiele, dem sog. *Core Set*. Summenberechnungen der Kernfunktionen werden verteilt unmittelbar auf den lokalen Knoten ausgeführt. Dadurch muss pro Element des Core Sets nur ein einzelner Wert statt aller Attributwerte an den zentralen Knoten übertragen werden. Trotz eines bisweilen um den Faktor zehn verminderten Kommunikationsbedarfs erreicht das Verfahren auf kontrollierten wie realen Datensätzen teils eine ähnlich hohe Klassifikationsgüte wie das in [356] beschriebene Verfahren. Die VDCVM soll noch im Rahmen der ersten Förderperiode gemäß [387] zur verteilten 2-class-SVM mit linearem Kern erweitert und mit den zuvor erwähnten bestehenden Ansätzen für die verteilte Optimierung verglichen werden.

Bisher ungelöste Probleme stellen die verteilte Optimierung von Hyperparametern und die verteilte Gewinnung sinnvoller Merkmale dar. Die Arbeiten zur verteilten Analyse von Sensordaten sollen in der zweiten Förderperiode im AP 1 dementsprechend fortgesetzt und erweitert werden.

Ansätze zur Einbindung der Qualitätsprognose in Produktionsprozesse (AP 9): Mit der Intelligent Manufacturing Process Control (IMPC) [B3/398] wurde ein Ansatz entwickelt, der die realzeitlich gelieferten Prognosen in die Prozesssteuerung integriert. Im Sinne eines modularen Konzepts integriert die IMPC alle Funktionen, die in den übrigen Arbeitspaketen erarbeitet wurden und die für eine realzeitliche Qualitätsprognose in starr verketteten Produktionsprozessen notwendig sind. Die insgesamt vier Module gliedern sich in die Module zur Datenerfassung und -sicherung, zur Datenüberwachung, zur Qualitätsprognose sowie zur Generierung der Steuerungsentscheidung. In dem Modul der Datenerfassung werden sämtliche Schnittstellen der IMPC zur Sensorik in den Fertigungsprozessen gebündelt. Neben der Sensorik beinhaltet dieses Modul auch die notwendige Software zur Vorauswahl und Auswertung der erfassten Signale sowie zur Konvertierung in ein im Weiteren verarbeitbares Datenformat. Die Module zur Datenüberwachung und Qualitätsprognose umfassen alle Schritte von der Vorverarbeitung der Daten bis hin zur Ableitung der Qualitätsprognose für ein Zwischenprodukt. Basierend auf den Qualitätsprognosen und den Informationen über den Fortschritt des Fertigungsprozesses werden in dem letzten Modul die Steuerungsentscheidungen abgeleitet.

Für die Interpretation der Qualitätsprognose zur Ableitung dieser Steuerungsentscheidungen im Fertigungsprozess wurde u. a. in [B3/398] ein auf Schwellwerten basierender Ansatz vorgestellt. Nach jedem Fertigungsschritt wird die prognostizierte Produktqualität mit dem für diesen Produktionsprozess festgelegten Schwellwert verglichen. Für die weiteren Bearbeitungsschritte werden nur solche Produkte zugelassen,

deren Qualitätsprognose den Schwellwert mindestens erfüllt. Über alle Produktionsprozesse hinweg weisen die Schwellwerte einen asymptotischen Verlauf auf. Grundgedanke dabei ist, dass eventuelle Qualitätsabweichungen in frühen Prozessen noch leichter durch die folgenden Prozesse ausgeglichen werden können, während in den finalen Prozessen die Kundenanforderungen erfüllt werden müssen. Zur Abstufung der Schwellwerte werden zwei unterschiedliche Vorgehen verfolgt. Einerseits wird das Wissen der jeweiligen Domänenexperten (z. B. spezifische Prozess- und Betriebsmitteleigenschaften) in Kombination mit prozessbezogenen physikalischen Wirkzusammenhängen (z. B. allgemeine Umformgrade, Walzgeschwindigkeiten und deren Auswirkungen) ausgenutzt. Andererseits werden Informationen aus multivariaten Analysen verwendet. Dazu werden die Varianzaufklärungen der qualitätsrelevanten Prozessparameter bzw. der daraus abgeleiteten Merkmale für jeden Bearbeitungsprozess ausgewertet. Die in den Modellen enthaltene Varianzaufklärung definiert dabei die relative Höhe des Schwellwertes des betrachteten Prozesses im Vergleich zu den Vorgänger- bzw. Nachfolgerprozessen.

Der modulare Aufbau der IMPC erlaubt eine leichte Adaption an den jeweiligen Anwendungsfall und unterstützt somit direkt eine leichtere Integration der erarbeiteten Ansätze in der Praxis. Des Weiteren trägt das Konzept der Modularisierung vor dem Hintergrund der langfristigen Zielstellung des SFB 876 bereits in dieser frühen Phase des Forschungsvorhabens zu einer Abstraktion der Lösungsansätze von konkreten Anwendungsfällen bei.

Neben der Entwicklung der IMPC wurden in Zusammenarbeit mit dem Projekt B4 grundlegende Untersuchungen zur Übertragbarkeit der Modellierungsansätze der Verkehrsphysik in die Produktionstechnik durchgeführt. In diesem Rahmen wurden u. a. zelluläre Automaten, insbes. das Nagel-Schreckenberg-Modell, hinsichtlich ihrer Anwendbarkeit zur Beschreibung von verketteten Fertigungsprozessen und zur Analyse des Systemverhaltens untersucht [B3/B4/402]. Aufgrund der identifizierten Anknüpfungspunkte befindet sich ein gemeinsames Forschungsvorhaben in der Vorbereitung, mit dem Ziel, neue Methoden der Produktionsmodellierung basierend auf den Modellen der Verkehrsphysik zu entwickeln.

Prototypische Integration der Qualitätsprognose in den Produktionsprozess (AP 10): In diesem Arbeitspaket sollen die in den übrigen Arbeitspaketen erarbeiteten Methoden und Konzepte in den Anwendungsfall integriert und ihre Potenziale evaluiert werden. Als letztes Arbeitspaket der ersten Förderperiode wurde mit der Bearbeitung begonnen, jedoch liegen zu diesem Zeitpunkt hier noch keine konkreten Ergebnisse vor.

Literatur

[348] M. T. Bahadori, Y. Liu und D. Zhang. „A general framework for scalable transductive transfer learning“. In: *Knowledge and Information Systems* 38.1 (2014), S. 61–83.

[349] M.-F. Balcan, S. Hanneke und J. W. Vaughan. „The true sample complexity of active learning“. In: *Machine Learning* 80.2–3 (2010), S. 111–139.

[350] S. Barbarossa, S. Sardellitti und P. D. Lorenzo. „Distributed Detection and Estimation in Wireless Sensor Networks“. In: *CoRR* abs/1307.1448 (2013).

[351] H. Becker. *Phänomen Toyota: Erfolgsfaktor Ethik*. Springer, 2006.

[352] A. Bellet, Y. Liang, A. B. Garakani, M.-F. Balcan und F. Sha. „Distributed Frank-Wolfe Algorithm: A Unified Framework for Communication-Efficient Sparse Learning“. In: *CoRR* abs/1404.2644 (2014).

[353] S. Boyd, N. Parikh, E. Chu, B. Peleato und J. Eckstein. „Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers“. In: *Foundations and Trends in Machine Learning* 3.Januar (Jan. 2011), S. 1–122.

[354] S. Chen, B. Liu, M. Qian und C. Zhang. „Kernel k-Means Based Framework For Aggregate Outputs Classification“. In: *Proceedings of the International Conference on Data Mining Workshops (ICDMW)*. 2009, S. 356–361.

[355] Y.-W. Chen, Y.-Z. Lu und G.-K. Yang. „Hybrid evolutionary algorithm with marriage of genetic algorithm and extremal optimization for production scheduling“. In: *International Journal of Advanced Manufacturing Technology* 36 (2008), S. 959–968.

[356] K. Das, K. Bhaduri und P. Votava. „Distributed anomaly detection using 1-class SVM for vertically partitioned data“. In: *Stat. Anal. Data Min.* 4.4 (2011), S. 393–406.

[357] K. Das, K. Bhaduri und H. Kargupta. „A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to-peer networks“. In: *Knowledge and Information Systems* 24.3 (2010), S. 341–367.

[358] B. Demir und L. Bruzzone. „A multiple criteria active learning method for support vector regression“. In: *Pattern Recognition* 47.7 (2014), S. 2558–2567.

[359] J. Deuse, B. Konrad und F. Bohnen. „Reduzierung von Variabilität - Einsatz von Data Mining in Produktionsystemen“. In: *Produzieren in Deutschland - Wettbewerbsfähigkeit im 21. Jahrhundert.* Hrsg. von A. Friedewald und H. Lödding. Gito, 2013, S. 299–316.

[360] B. Ding, M. T. Cychowski, Y. Xi, W. Cai und B. Huang. „Model Predictive Control“. In: *Journal of Control Science and Engineering* 2012 (2012), S. 1–2.

[361] H. W. Dörmann Osuna. „Ansatz für ein prozessintegriertes Qualitätsregelungssystem für nicht stabile Prozesse“. Diss. Fakultät Maschinenbau, TU Ilmenau, Jan. 2008.

[18] P. A. Forero, A. Cano und G. B. Giannakis. „Consensus-Based Distributed Support Vector Machines“. In: *The Journal of Machine Learning Research* 99 (2010), S. 1663–1707.

[362] O. Gusikhin, R. Caprihan und K. E. Stecke. „Least in-sequence probability heuristic for mixed-volume production lines“. In: *International Journal of Production Research* 46.3 (2008), S. 647–673.

[363] A. Heidari und M. R. Forouzan. „Optimization of cold rolling process parameters in order to increasing rolling speed limited by chatter vibrations“. In: *Journal of Advanced Research* 4 (2013), S. 27–34.

[364] J. Hernández-González, I. Inza und J. A. Lozano. „Learning Bayesian network classifiers from label proportions“. In: *Pattern Recognition* 46.12 (2013), S. 3425–3440.

[365] K. L. Hoffman, M. Padberg und G. Rinaldi. „Traveling salesman problem“. In: *Encyclopedia of Operations Research and Management Science.* Springer, 2013, S. 1573–1578.

[366] R. Isermann. *Fault-diagnosis systems: an introduction from fault detection to fault tolerance.* Springer, 2006.

[367] B. Korte und J. Vygen. „Das Traveling-Salesman-Problem“. In: *Kombinatorische Optimierung.* Springer-Lehrbuch Masterclass. Springer Berlin Heidelberg, 2012, S. 589–627.

[368] H.-P. Kriegel, P. Kröger und A. Zimek. „Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering“. In: *ACM Transactions on Knowledge Discovery from Data* 3.1 (März 2009), 1:1–1:58.

[369] D. D. Lewis und W. A. Gale. „A Sequential Algorithm for Training Text Classifiers“. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York, NY, USA: Springer-Verlag New York, Inc., 1994, S. 3–12.

[370] S. Lodi, R. Ñanculef und C. Sartori. „Single-Pass Distributed Learning of Multi-class SVMs Using Core-Sets“. In: *Proc. of SDM'10.* 2010, S. 257–268.

[371] A. Mcgovern, D. H. Rosendahl, R. A. Brown und K. K. Droegemeier. „Identifying Predictive Multi-dimensional Time Series Motifs: An Application to Severe Weather Prediction“. In: *Data Mining and Knowlegde Discovery* 22.1-2 (Jan. 2011), S. 232–258.

[372] I. Mierswa und K. Morik. „Method trees: building blocks for self-organizable representations of value series: how to evolve representations for classifying audio data“. In: *Proceedings of the Genetic and Evolutionary Computation Conference GECCO 2005, Workshop on Self-Organization In Representations For Evolutionary Algorithms: Building complexity from simplicity.* New York, NY, USA: ACM, 2005, S. 293–300.

[373] A. Mitra. *Fundamentals of Quality Control and Improvement.* Wiley, 2012.

[374] E. Montañés, R. Senge, J. Barranquero, J. R. Quevedo, J. J. del Coz und E. Hüllermeier. „Dependent binary relevance models for multi-label classification“. In: *Pattern Recognition* 47.3 (2014), S. 1494–1508.

[375] C. H. Mooney und J. F. Roddick. „Sequential Pattern Mining – Approaches and Algorithms“. In: *ACM Computing Survey* 45.2 (März 2013), 19:1–19:39.

[376] D. R. Musicant, J. M. Christensen und J. F. Olson. „Supervised Learning by Training on Aggregate Outputs“. In: *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (ICDM).* Washington, DC, USA: IEEE Computer Society, 2007, S. 252–261.

[377] B. Naderi und A. Arshadi. „Scheduling a variant of flowshop problems to minimize total tardiness“. In: *Technical Journal of Engineering and Applied Sciences* 3.22 (2013), S. 3142–3149.

[378] R. U. Pedersen. „Worst Case Execution time Analysis of Sparse Kernel Learners“. In: *International Workshop on Knowledge Discovery from Data Streams at ECML.* 2006.

[379] T. Rakthanmanon und E. J. Keogh. „Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets“. In: *SIAM International Conference on Data Mining (SDM)*. 2013, S. 668–676.

[380] J. Read, B. Pfahringer, G. Holmes und E. Frank. „Classifier Chains for Multi-label Classification“. In: *Machine Learning* 85.3 (Dez. 2011), S. 333–359.

[381] D. Roth und K. Small. „Margin-Based Active Learning for Structured Output Spaces“. In: *Proceedings of the 17th European Conference on Machine Learning (ECML)*. Berlin, Heidelberg: Springer-Verlag, 2006, S. 413–424.

[382] S. Rüping. „SVM Classifier Estimation From Group Probabilities“. In: *Proceedings of the 27th International Conference on Machine Learning (ICML)*. 2010, S. 911–918.

[383] P. Senin und S. Malinchik. „SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model“. In: *IEEE 13th International Conference on Data Mining (ICDM)*. Dez. 2013, S. 1175–1180.

[384] K. Sun, G. Yang und C. Pan. „Population-based Extremal Optimization Algorithm for Hot Rolling Scheduling Problem“. In: *Journal of Computational Information Systems* 6.7 (2010), S. 2415–2422.

[385] B. R. Suresh, R. A. Fundakowski, T. S. Levitt und J. E. Overland. „A Real-Time Automated Visual Inspection System for Hot Steel Slabs“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5.6 (Nov. 1983), S. 563 –572.

[386] L. Tan und X. Wang. „A Scatter Search Algorithm for a Multistage Production Scheduling Problem With Blocking and Semi-Continuous Batching Machine“. In: *IEEE Transactions on Control Systems Technology* 19.5 (2011), S. 976–989.

[387] I. Tsang, J.-Y. Kwok und J. Zurada. „Generalized Core Vector Machines“. In: *IEEE Transactions on Neural Networks* 17.5 (Sep. 2006), S. 1126–1140.

[388] I. Tsochantaridis, T. Joachims, T. Hofmann und Y. Altun. „Large Margin Methods for Structured and Interdependent Output Variables“. In: *Journal of Machine Learning Research* 6 (2005), S. 1453–1484.

[389] J. Vreeken, M. Leeuwen und A. Siebes. „Krimp: mining itemsets that compress“. In: *Data Mining and Knowledge Discovery* 23.1 (2011), S. 169–214.

[390] Z. Wang und J. Ye. „Querying Discriminative and Representative Samples for Batch Mode Active Learning“. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. New York, NY, USA: ACM, 2013, S. 158–166.

[391] M. A. Younes, M. Shahtout und M. Damir. „A parameters design approach to improve product quality and equipment performance in hot rolling“. In: *Journal of Materials Processing Technology* 171 (2006), S. 83–92.

[392] F. X. Yu, D. Liu, S. Kumar, T. Jebara und S.-F. Chang. „$\propto$-SVM for learning with label proportions“. In: *CoRR* abs/1306.0886 (2013).

[393] H. Yunhong, F. Liang und H. Guoping. „Privacy-Preserving SVM Classification on Vertically Partitioned Data without Secure Multi-party Computation“. In: *5th International Conference on Natural Computation (ICNC)*. Bd. 1. Aug. 2009, S. 543–546.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[B3/394] K. Bhaduri und M. Stolpe. „Distributed Data Mining in Sensor Networks“. In: *Managing and Mining Sensor Data*. Hrsg. von C. C. Aggarwal. Berlin, Heidelberg: Springer, 2013. Kap. 8.

[B3/395] F. Bohnen, M. Buhl und J. Deuse. „Systematic Procedure for leveling of low volume and high mix production“. In: *CIRP Journal od Manufacturing Science and Technology* 6.1 (2013), S. 53–58.

[B3/396] F. Bohnen, M. Stolpe, J. Deuse und K. Morik. „Using a Clustering Approach with Evolutionary Optimized Attribute Weights to Form Product Families for Production Leveling“. In: *Robust Manufacturing Control*. Hrsg. von K. Windt. Lecture Notes in Production Engineering. Berlin, Heidelberg: Springer-Verlag, 2013, S. 189–202.

[B3/397] J. Deuse, B. Konrad und F. Bohnen. „Renaissance of Group Technology: Reducing Variability to Match Lean Production Prerequisites“. In: *Manufacturing Modelling, Management, and Control*. Hrsg. von N. Bakhtadze, K. Chernyshov, A. Dolgui und V. Lototsky. Bd. 7. 2013, S. 998–1003.

[B3/398] B. Konrad, D. Lieber und J. Deuse. „Striving for Zero Defect Production: Intelligent Manufacturing Control through Data Mining in Continuous Rolling Mill Processes“. In: *Robust Manufacturing Control*. Hrsg. von K. Windt. Lecture Notes in Production Engineering. CIRP. Berlin, Heidelberg: Springer-Verlag, 2013, S. 215–229.

[B3/C1/399] S. Lee, M. Stolpe und K. Morik. „Separable Approximate Optimization of Support Vector Machines for Distributed Sensing“. In: *Machine Learning and Knowledge Discovery in Databases*. Hrsg. von P. Flach, T. D. Bie und N. Cristianini. Bd. 7524. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2012, S. 387–402.

[B3/400] D. Lieber, O. Erohin und J. Deuse. „Wissensentdeckung im industriellen Kontext - Herausforderungen und Anwendungsbeispiele“. In: *Zeitschrift für wirtschaftlichen Fabrikbetrieb (ZWF)* (2013).

[B3/401] D. Lieber, M. Stolpe, B. Konrad, J. Deuse und K. Morik. „Quality Prediction in Interlinked Manufacturing Processes based on Supervised & Unsupervised Machine Learning“. In: *Procedia CIRP - 46th CIRP Conf. on Manufacturing Systems*. Bd. 7. Elsevier, 2013, S. 193–198.

[B3/B4/402] T. Maschek, B. Konrad, J. Deuse, G. Hermanns, D. Weber und M. Schreckenberg. „Verkehrsforschung in der Produktionsflussanalyse - Übertragung von Modellen der statistischen Physik auf die Analyse von Produktionssystemen“. In: *ZWF - Zeitung für wirtschaftlichen Fabrikbetrieb* 106.11 (2011), S. 833–837.

[B3/403] M. Stolpe und K. Morik. „Learning from Label Proportions by Optimizing Cluster Model Selection“. In: *Machine Learning and Knowledge Discovery in Databases*. Hrsg. von D. Gunopulos, T. Hofmann, D. Malerba und M. Vazirgiannis. Bd. 6913. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2011, S. 349–364.

[B3/404] M. Stolpe, K. Bhaduri, K. Das und K. Morik. „Anomaly Detection in Vertically Partitioned Data by Distributed Core Vector Machines“. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013*. Hrsg. von H. Blockeel, K. Kersting, S. Nijssen und F. Zelezný. Springer, 2013, S. 321–336.

[B3/405] M. Stolpe, K. Morik, B. Konrad, D. Lieber und J. Deuse. „Challenges for Data Mining on Sensor Data of Interlinked Processes“. In: *Proceedings of the Next Generation Data Mining Summit (NGDM) 2011*. 2011.

3.4 Planung des Teilprojekts

Eine Übersicht über das Arbeitsprogramm für die zweite Förderperiode gibt Abb. 3.1. Die in der ersten Förderperiode betrachteten Fragestellungen werden in dem Arbeitsprogramm der zweiten Periode zu großen Teilen fortgeführt und durch neue Aspekte ergänzt.

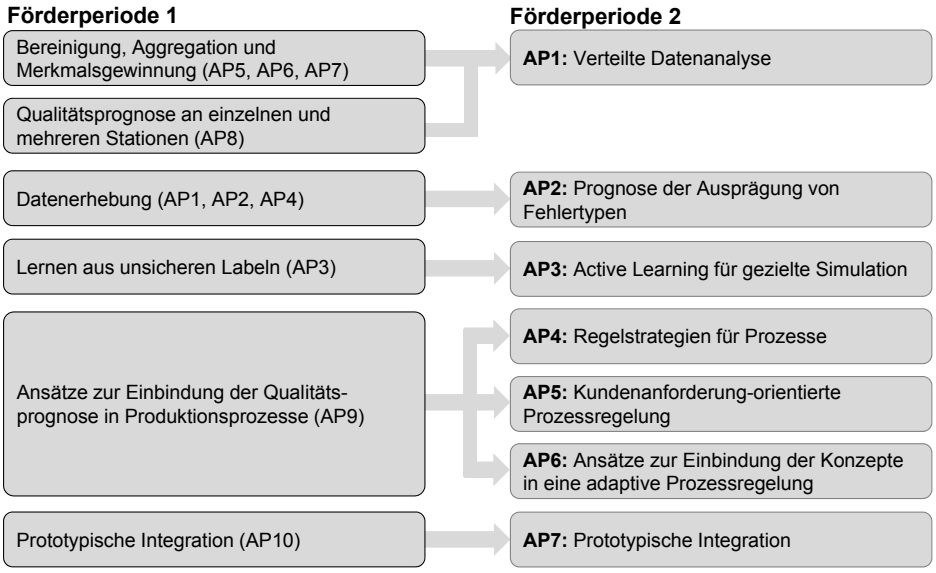


Abbildung 3.1.: Übersicht Arbeitsprogramm Teilprojekt B3

Ziele

Langfristiges Ziel des Teilprojektes ist die Entwicklung von Verfahren zur realzeitlichen Qualitätsprognose und Steuerung von Produktionsprozessen. Es sollen Trade-Offs zwischen Laufzeit, Speicherverbrauch, Kommunikationskosten, Energieverbrauch und Prognosegüte untersucht werden, wenn Sensordaten nicht mehr zentral, sondern dezentral gespeichert und analysiert werden.

Die erste Förderperiode konzentrierte sich dabei auf Verfahren zur Vorhersage einer einzelnen Zielgröße, der Qualität des Endprodukts, und der binären Entscheidung, ob ein Produkt ausgeschleust werden soll oder nicht. Das Vorgehen reduziert einerseits Fehlerkosten, die aus der Weiterbearbeitung fehlerhafter Produkte entstehen, andererseits führt der Abbruch der Bearbeitung unweigerlich zu Kosten für das Verschrotten bzw. Recycling der fehlerhaften Produkte.

Ein zentrales Ziel der zweiten Förderperiode stellt daher die Erweiterung der Prozesssteuerung dar: Bei prognostizierten Qualitätsabweichungen soll die Bearbeitung derart angepasst werden können, dass die vom Kunden geforderten Qualitätseigenschaften ohne einen Abbruch der Bearbeitung erreicht werden. Dazu sollen zwei Ansätze untersucht werden: die Anpassung der Prozessparameter zur Sicherstellung der Spezifikationserfüllung und die Neuzuordnung von Kundenaufträgen zu Zwischenprodukten in Abhängigkeit der prognostizierten Qualitätseigenschaften der Produkte. Letzterer Ansatz ist dann relevant, wenn Qualitätsabweichungen auftreten, die bei einem Kundenauftrag das Produkt als Ausschuss kennzeichnen würden, während diese Abweichungen bei einem anderen Kundenauftrag noch akzeptiert würden. Die aus der Qualitätsprognose ableitbaren Handlungsoptionen sollen dementsprechend erweitert und es soll aktiv und in Realzeit in die Steuerung nachfolgender Prozessschritte eingegriffen werden.

Um erweiterte Steuerungsoptionen ableiten zu können, müssen die Qualitätsprognosen hinsichtlich des Detaillierungsgrades wie auch der Reaktionszeit bei der Erstellung weiterentwickelt werden. Genügten bisher binäre Qualitätsprognosen, benötigt insbesondere die Adaption der Bearbeitungsparameter differenziertere Aussagen über die Qualität, etwa Angaben zur erwarteten Position und Ausprägung von Fehlern innerhalb des Endprodukts. Parallel dazu soll die Fortführung der Forschungsarbeiten auf dem Gebiet der verteilten Datenanalyse die Reaktionszeit der Verfahren zur Qualitätsprognose senken. Für die Elemente der Qualitätsprognose und Steuerung soll dann versucht werden, harte Realzeitschranken abzuleiten, etwa ihre *Worst Case Execution Time* (WCET), wie sie in der Analyse von eingebetteten Systemen üblich ist.

Arbeitsprogramm

Arbeitspaket 1. Verteilte Datenanalyse: Die Arbeiten zur verteilten Datenanalyse sollen in der zweiten Förderperiode fortgesetzt und erweitert werden. Iterative verteilte Verfahren wie [350], [352] oder [353] übertragen während des Trainings von Prognosemodellen unter Umständen sogar mehr Daten als den ursprünglichen Datensatz. Im Vergleich dazu reduziert der in [370] beschriebene Ansatz für horizontal verteilte Daten das zu übertragende Datenvolumen durch die Verwendung von Core Sets. Einen ähnlichen Ansatz verfolgt die in der ersten Förderperiode entwickelte VDCVM [B3/404] zur Anomaliedetektion in vertikal verteilten Daten und garantiert, dass im schlechtesten Fall nicht mehr als die vorhandenen Trainingsbeispiele übertragen werden.

Selbst wenn das Lernverfahren nicht iterativ arbeitet und das zu übertragende Datenvolumen mittels intelligentem Sampling reduziert werden kann, können vorangehende oder übergeordnete Schritte dennoch iterativ arbeiten. So werden etwa die Hyperparameter der SVM und Kernelparameter oft über sog. *Wrapper-Ansätze* optimiert, die Parameterkombinationen mittels unterschiedlicher Suchstrategien testen. Dabei wird das Lernverfahren für jede getestete Kombination mindestens einmal ausgeführt. Ohne eine entsprechende Anpassung an das verteilte Szenario würden also bei jedem Aufruf unter Umständen erneut Daten übertragen, und damit evtl. mehr als der Gesamtdatensatz. Ähnliche Probleme entstehen bei Wrapper-Ansätzen für die Merkmalsextraktion und -selektion und der Evaluation von Lernverfahren. Die Wahl optimaler Hyperparameter oder Mengen von Merkmalen wird in sehr vielen Arbeiten zu verteilten Verfahren für die Datenanalyse entweder gar nicht oder nur sehr eingeschränkt behandelt. Im Rahmen des Arbeitspakets sollen daher Verfahren für die kommunikationseffiziente verteilte Parameteroptimierung und Merkmalsselektion, insbesondere für den Fall vertikal verteilter Daten, entwickelt werden.

Für die Parameteroptimierung besteht ein möglicher Ansatz darin, auf das iterative Testen verschiedener Kombinationen ganz zu verzichten und diese heuristisch abzuleiten. Bei der SVM mit RBF-Kern basiert eine mögliche Heuristik für die Wahl des γ-Parameters etwa auf der Bestimmung von Quantilen der paarweisen Distanzen aller Datenpunkte. Im Rahmen des Arbeitspakets sollen dementsprechend kommunikationseffiziente verteilte heuristische Verfahren für die Ableitung von Hyperparametern entwickelt werden. Des Weiteren

soll untersucht werden, inwieweit sich durch verschiedene Parametrisierungen entstehende Lösungen überlappen und Redundanzen in den Modellen zur Reduktion von Kommunikation ausnutzen lassen. Dabei soll Wissen über Strukturen der Modelle ausgenutzt werden, im Gegensatz zu *multi-armed bandit* Ansätzen oder dem in Teilprojekt A3 untersuchten *iterated F-Racing* bzw. der *modellbasierten Optimierung* (MBO), welche die zu testenden Modelle als Black-Box betrachten. Eine andere Frage besteht darin, ob sich Parameter bei iterativ oder inkrementell arbeitenden Verfahren (wie der VDCVM) schon während des Lernvorgangs anpassen lassen, also ohne eine übergeordnete iterative Suche.

Für die Merkmalsselektion stellt die vertikale Verteilung eine besondere Herausforderung dar. Während für das horizontale Szenario erste Ansätze entwickelt wurden [357], reicht eine lokale Selektion im vertikalen Szenario nicht aus, sofern die Zielgröße von den Merkmalen unterschiedlicher Knoten gemeinsam abhängt. Kombinationen über mehrere Knoten hinweg zu testen erfordert zwangsläufig Kommunikation. Deren Kosten könnten durch eine vorgängige Aggregation an den lokalen Knoten verringert werden. Es soll untersucht werden, wie stark Messreihen aggregiert werden dürfen, um noch immer sinnvolle Merkmale über mehrere Knoten hinweg auswählen zu können. Dabei soll auch überprüft werden, ob Informationen von anderen Knoten und über die Zielgröße so berücksichtigt werden können, dass die Gesamtauswahl von Merkmalen verbessert wird. Die finale Merkmalsselektion soll nach der Aggregation erfolgen. Hier sollen die Kommunikationskosten herkömmlicher Ansätze zur Merkmalsselektion mit solchen verglichen werden, die Merkmale in einem einzigen Lauf über die Daten selektieren. Des Weiteren bietet sich ein Vergleich der ausgewählten Merkmale mit durch Lernverfahren bestimmten Merkmalsgewichten an. Kann das verteilte Lernverfahren selbst sinnvolle Merkmale bestimmen, ließe sich der vorangehende Schritt der Merkmalsselektion (und damit unter Umständen Kommunikation) einsparen.

Arbeitspaket 2. Prognose der Ausprägung von Fehlertypen: Bisher wurden Informationen über im Endprodukt auftretende Fehler zu einem einzigen numerischen Label aggregiert. Die daraus abgeleiteten Prognosemodelle unterstützen einfache Entscheidungen, etwa ob ein Produkt aus dem laufenden Prozess ausgeschleust werden soll. Gezieltere Eingriffe in die Steuerung benötigen detailliertere Prognosen, welche die Positionen von Fehlern und den Grad ihrer Ausprägung beinhalten. Objekten müssen hier also mehrere, teils reellwertige Label zugeordnet werden. Dabei kann es Abhängigkeiten geben, etwa wenn Fehlbearbeitungen aus frühen Bearbeitungsschritten zu weiteren Fehlern in nachfolgenden Schritten führen würden. Im Bereich der Datenanalyse werden solche Prognosen auch als *strukturierte Ausgaben* bezeichnet. Als Trainingsbeispiele können weiterhin die bereits erfassten Prozessparameter dienen, während Angaben zu Positionen und Graden von Fehlern aus Spektren der Ultraschallprüfung oder mittels existierender Umformsimulationen gewonnen werden können.

Einfache Ansätze für das Lernen mit strukturierten Ausgaben ignorieren mögliche Abhängigkeiten zwischen den Labeln und trainieren unabhängige binäre Klassifikatoren für jedes Label (*binary relevance learning*). Der erweiterte Ansatz der *classifier chains* [380] berücksichtigt Abhängigkeiten, indem jeweils die Vorhersagen des vorhergehenden Klassifikators in einer Kette in die Prognose einfließen. Während dieser Ansatz nur jeweils eine mögliche Anordnung wählt, schätzen *probabilistic classifier chains* [374] stattdessen vollständig die gemeinsame Verteilung von Beispielen und Labeln. Die *strukturelle SVM* [388] hingegen maximiert den Margin zwischen der besten und allen anderen in Frage kommenden Labelungen. Bei nur geringen Abhängigkeiten zwischen den Labeln können auch einfache Verfahren hohe Prognosegüten erreichen. Sonst schneiden die erweiterten Methoden in der Regel besser ab, benötigen dafür aber auch länger für das Training. Im Rahmen des Arbeitspakets sollen unterschiedliche Ansätze für die detaillierte Fehlerprognose daher zunächst auf ihre Eignung hin untersucht und miteinander verglichen werden. Im Fokus steht dabei die Ermittlung von Trade-Offs zwischen Laufzeit und zu erzielender Prognosegüte. Die Ansätze sollen außerdem mit den in den Teilprojekten A1 und A6 entwickelten Methoden zur Prognose mittels grafischer Modelle und Graphkernen verglichen werden.

Alle Methoden für das Lernen strukturierter Ausgaben haben gemein, dass Training und Prognose in der Regel aufwendiger sind als das Lernen eines einzelnen Labels. Bereits bei der einfachen Methode des *binary relevance learning* muss z. B. für jedes Label ein eigener Klassifikator trainiert und später auch angewandt werden. Im Falle der strukturierten SVM muss je nach Lernproblem ggf. auch bei der Anwendung des Modells ein Optimierungsproblem gelöst werden. Vor dem Hintergrund der zuvor ermittelten Trade-Offs soll daher untersucht werden, inwieweit sich Verfahren mit hoher Prognosegüte auch für eine realzeitliche Prognose in laufenden Produktionsprozessen eignen und ob sich ihre Laufzeit ggf. verringern lässt, etwa durch Anwendung der im Rahmen von AP 1 entwickelten Methoden für die verteilte Datenanalyse.

Arbeitspaket 3. Active Learning für gezielte Simulation: Für die Produktion kann die genaue Charakterisierung der gelernten Entscheidungsgrenze und von Beispielen in deren Nähe eine besondere Bedeutung

haben, da sich aus ihnen möglicherweise qualitätsrelevante und kritische Prozessparameterbereiche ableiten lassen. Dabei gibt es im Produktionsbereich häufig ein starkes Ungleichgewicht zwischen positiven und negativen Beispielen, da ein fehlerhafter Prozessablauf eine Ausnahmesituation darstellt. Zur Verbesserung der Entscheidungsgrenze wäre es daher wünschenswert, negative Beispiele gezielt erzeugen zu können. Da dies im realen System mit hohen wirtschaftlichen Aufwänden verbunden wäre, können potentiell zu Fehlern führende Prozessverläufe lediglich durch bestehende Simulationsmodelle der Bearbeitungsprozesse, beispielsweise Umformprozesse, nachgestellt werden. Da solche Simulationen bis zu mehrere Stunden pro simuliertem Bearbeitungsschritt in Anspruch nehmen, können jedoch auch hier nur wenige Prozessparameterkombinationen untersucht werden. Das Ziel dieses Arbeitspakets besteht somit in der Entwicklung eines Verfahrens, welches die Anzahl der zu testenden Parameter und damit die Zeit für benötigte Simulationen verringert.

Einen möglichen Ansatz stellen Arbeiten im Bereich des *Active Learnings* dar. Mit ihnen lässt sich die Anzahl benötigter Trainingsbeispiele unter bestimmten Umstände verringern, indem noch ungelabelte Beispiele zum Labeln zielgerichtet und nicht rein zufällig ausgewählt werden [349]. Zur Auswahl wurden unterschiedliche Kriterien entwickelt, etwa Maße für die Unsicherheit der Klassenzugehörigkeit oder die Nähe der Beispiele zur Entscheidungsgrenze. Werden Beispiele allein basierend auf dem bis dahin gelernten Modell ausgewählt, so kann allerdings die ursprüngliche Verteilung der Beispiele verzerrt werden. In neueren Arbeiten wie [358] werden daher teils mehrere Kriterien kombiniert und auch die ursprüngliche Verteilung berücksichtigt, etwa über ein Clustering der Trainingsmenge. Im Rahmen des Arbeitspakets sollen die bestehenden Ansätze und Kriterien auf ihre Eignung für die aktive Auswahl repräsentativer Wertereihen hin untersucht und ggf. weiterentwickelt werden. Da bestehende Simulationen in der Regel offline außerhalb des laufenden Produktionsprozesses im Batchbetrieb ausgeführt werden und nicht verteilt arbeiten, soll an dieser Stelle auf die Betrachtung verteilter Ansätze verzichtet werden.

Viele Ansätze zum Active Learning wählen bereits vorhandene Beobachtungen nacheinander zur unmittelbaren Labelung durch menschliche Personen aus. Die automatische Generierung von Beispielen zum Labeln kann daran scheitern, dass die Beispiele von Menschen nur schlecht interpretiert werden können. Dieses Problem besteht hier nicht, da die Beispiele automatisch durch eine Simulation beurteilt werden. Physikalisch oder steuerungstechnisch nicht realisierbare Prozessverläufe sollen allerdings durch die Einbeziehung von Hintergrundwissen, etwa die Eingrenzung sinnvoller Parameterbereiche, ausgeschlossen werden. Schließlich soll untersucht werden, ob es sinnvoller ist, Beispiele einzeln, nacheinander oder eine Beispielmenge als Batch [390] labeln zu lassen.

Das zu entwickelnde Verfahren sollte, so weit es geht, unabhängig von der jeweils gegebenen Simulation zu verwenden sein. Dazu sind geeignete Formalismen zu entwickeln, um Hintergrundwissen von der konkreten Domäne unabhängig zu beschreiben. Geeignet formalisiertes Wissen kann möglicherweise auch in die Kriterien einfließen, anhand derer für die Simulation geeignete Prozessverläufe bestimmt werden. Hier ergeben sich Bezüge zum Lernen strukturierter Ausgaben in AP 2, da es bisweilen möglich ist, Wissen über die Abhängigkeiten von Strukturen im Ausgaberaum zur Reduktion des Aufwandes für das Labeln auszunutzen [381]. Im Vergleich zu bestehenden Verfahren des Active Learnings, die einzelne Instanzen labeln lassen, könnte es im Zusammenhang mit Simulationen auch Sinn machen, die Grenzen relevanter eingegrenzter *Bereiche* des Instanzenraums vorzugeben. Bei allen vorgenannten Arbeiten ist zu berücksichtigen, dass der Raum der Prozessverläufe unter Umständen nicht unmittelbar dem Parameterraum der Simulation entspricht, weshalb erst eine entsprechende Abbildung zwischen den Räumen definiert oder ggf. gelernt werden muss.

Arbeitspaket 4. Regelstrategien für Prozesse: Für die Einbindung der Informationen aus den Qualitätsprognosen in die Steuerung der Produktionsprozesse wurde bereits in der ersten Förderperiode ein auf multivariaten Analysen basierender Ansatz entwickelt und in dem Konzept der intelligenten Prozesssteuerung implementiert [359, B3/398]. Aus den Qualitätsprognosen wurde anhand von Qualitätsschwellwerten die Steuerungsentscheidung abgeleitet, ob die Bearbeitung fortgesetzt oder abgebrochen wird. Je weiter fortgeschritten der Bearbeitungsprozess, desto höher der Schwellwert, den die Qualitätsprognosen erfüllen müssen, da die prozessbedingten Freiheitsgrade mit dem Prozessfortschritt abnehmen. In der Fortführung der Arbeiten in AP 3 ist es das Ziel dieses Arbeitspakets, die in der ersten Förderperiode entwickelten Ansätze der Prozesssteuerung hin zu einer Prozessregelstrategie zu erweitern.

In der Literatur bestehende Ansätze zur Ableitung von Regelungsentscheidungen bei Walzprozessen basieren auf der bestmöglichen Kombination der Prozessparameter und haben zumeist die Optimierung des Schwingungsverhaltens zum Ziel. Hierzu verwenden beispielsweise Heidari und Forouzan [363] für den Fall des Kaltwalzens Walzgeschwindigkeit, Abwalzgrad, Oberflächenspannung oder Stabdicke als Prozessparameter für die Vermeidung von Rattermarken. Younes et al. [391] bilden zur Betrachtung von Rattermarken beim Warmwalzen Klassen von Prozess- (Abwalzgrad, Temperatur des Stabes, Bearbeitungsgeschwindigkeit), Qualitäts- (Profil und Dicke des Stabes, Abstand zwischen den Rattermarken) und Performanzparametern

(Walzkraft-, Vibrations- und Abnutzungsgrad). Die Entscheidung über einen möglichen Steuerungseingriff bei den Bearbeitungsparametern basiert hier auf drei separaten Regressionsmodellen, welche die betrachteten Input- und Outputparameter der Walzprozesse zueinander in Beziehung setzen. Diese und vergleichbare Ansätze, die auf der Prozessmodellierung mittels funktionaler Zusammenhänge basieren, können jedoch nicht auf den hier verfolgten Ansatz der Qualitätsprognose übertragen werden.

Stattdessen muss die Prozessregelstrategie die Ergebnisse der Qualitätsprognosen auswerten können und basierend darauf festlegen, ob eine weitere Bearbeitung des Werkstücks erlaubt wird bzw. welche Anpassungen im nächsten Bearbeitungsschritt vorzunehmen sind. Die Anpassungen können einerseits die Parameter im aktuellen Prozess oder in den Folgeprozessen betreffen, andererseits kann es, je nach Anwendungsfall, auch möglich bzw. nötig sein, die Prozessfolge selbst anzupassen. Im Gegensatz zu dem in der ersten Förderperiode entwickelten Konzept muss dabei neben dem Prozessfortschritt auch berücksichtigt werden, ob Anpassungen der Prozessparameter oder Prozessabfolge möglich sind, so dass die prognostizierten Qualitätsabweichungen ausgeglichen werden können. Die Entscheidung über die Akzeptanz von Parameteranpassungen kann nur für bereits beobachtete Prozessspezifikationen erfolgen oder, wenn die Prozessresultate für solche Prozessspezifikationen bereits durch die in AP 3 entwickelte Methodik identifiziert und durch Umformsimulationen verifiziert wurden.

Ein möglicher Ansatz zur Entwicklung einer Lösung für diese Fragestellung kann die Detaillierung des bestehenden Schwellwertverfahrens darstellen. Durch die Untersuchung der Einflüsse einzelner Prozessschritte auf das Prozessergebnis - beispielsweise durch die Auswertung von Domänenwissen und die Durchführung multivariater Analysen zur Auswertung des Freiheitsgrades zur Qualitätsoptimierung an einzelnen Anlagen - können dann Intervalle für die prognostizierte Qualität der Zwischenprodukte definiert werden, innerhalb derer entweder die Produktion ohne Anpassungen fortgesetzt, eine Parameter- oder Prozessschrittanpassung akzeptiert oder die Bearbeitung vollständig abgebrochen wird.

Arbeitspaket 5. Kundenanforderung-orientierte Prozessregelung: Eine Grundannahme, die den bisherigen Arbeiten zur Erstellung der Qualitätsprognosen und der Verwendung dieser zur Steuerung der Produktionsprozesse implizit zugrunde liegt, ist, dass jedes Produkt am Ende der Prozesskette vollkommen fehlerfrei sein muss. Gilt diese Zielstellung beispielsweise bei Montagelinien hinsichtlich der Funktionalität des montierten Produktes, muss diese Grundannahme für andere Produkte und Qualitätseigenschaften hinterfragt werden. Beispielsweise deckt eine ultraschallbasierte Qualitätskontrolle, wie sie bei der Warmwalzbearbeitung eingesetzt wird, Materialfehler auf detailliertester Ebene auf, wobei viele dieser Fehler jedoch keine Verringerung der Funktionalität des hergestellten Produktes darstellen. Ausschließlich der Abgleich der gemessenen Qualitätseigenschaften mit den vom Kunden festgeschriebenen Qualitätsanforderungen definiert die Grenze, ab welcher festgestellte Qualitätsabweichungen tatsächlich die Funktionalität des Produktes in den Augen des Kunden beeinträchtigen. Werden geringfügige Qualitätsabweichungen so als unkritisch identifiziert, können Steuerungseingriffe unterlassen werden. Liegen jedoch relevante Qualitätsabweichungen vor, erlaubt der in der ersten Förderperiode entwickelte Ansatz nur den Abbruch der Bearbeitung. Die in der zweiten Förderperiode, AP 4, angestrebte Parameteradaption geht bereits einen Schritt weiter und zielt auf eine Anpassung der Prozessparameter zum Ausgleich der Qualitätsabweichungen. Beide Ansätze berücksichtigen jedoch nicht die expliziten Kundenanforderungen, sondern streben ein vollständig fehlerfreies Produkt an. Wird das Wissen über die vorliegenden Kundenaufträge und -anforderungen mit in die Prozessregelung eingebunden, ermöglicht dies darüber hinaus eine weitere Entscheidung: Die Zuordnung des aktuellen Zwischenprodukts zu einem anderen Kundenauftrag aus dem Auftragsbestand, dessen Qualitätsanforderungen trotz der prognostizierten Qualitätsabweichungen erfüllt werden können. Ist ein solcher Auftrag vorhanden, ist der Eingriff in die Bearbeitungsprozesse nicht nötig.

Vorbild für ein solches Vorgehen sind Prozesssteuerungsverfahren in der Automobilindustrie zur Optimierung der Fertigungs- und Montageprozessreihenfolge. Bei diesen erfolgt die Zuordnung von Kundenaufträgen zu Fahrzeugen jedoch nicht mit dem Ziel, die Qualitätsanforderungen des Kunden zu erfüllen. Vielmehr wird mit Blick auf den zugesagten Liefertermin versucht, die Auslastung der Mitarbeiter in den Montageprozessen zu optimieren. Unter anderem wird in [362] deshalb die Zuordnung von Kundenaufträgen zu Rohkarossen als ein virtueller Resortierpuffer verwendet. An mehreren Punkten in dem Herstellungsprozess, beispielsweise nach der Lackierung und vor dem Beginn der Endmontage, können so die Kundenaufträge neu zugeordnet und die Termintreue verbessert werden. Mit Blick auf die erzielten Qualitätsmerkmale von Produkten kann ein solcher Ansatz helfen, die Ausschussrate und die benötigten Eingriffe in die Prozesse zu reduzieren.

Das Ziel dieses Arbeitspakets ist die Entwicklung eines solchen Ansatzes, der unter dem Gesichtspunkt der Produktqualität eine Neuzuordnung von Kundenaufträgen zu Zwischenprodukten ermöglicht. Dies entspricht der Problemstellung der Reihenfolgebildung bzw. -anpassung. Da die hier betrachteten automatisierten Produktionssysteme von den Aufträgen nach dem FIFO-Prinzip (first-in, first-out, d. h. Aufträge können sich

nicht überholen) durchlaufen werden, erfolgt diese Planung nur für das Gesamtsystem und nicht für die einzelnen Anlagen. Es wird für jeden Takt festgelegt, welcher Produktionsauftrag in das System eingelastet wird, so dass für den betrachteten Planungshorizont die Produktionsreihenfolge definiert wird.

Dazu müssen zunächst die Zielkriterien definiert werden, die zu einer Entscheidung über die neue Zuordnung von Kundenaufträgen aus dem Auftragsbestand zu einem Zwischenprodukt führen. Mögliche Kriterien können u. a. die Minimierung der Steuerungseingriffe und der Abweichung von den Kundenanforderungen sowie dem zugesagten Liefertermin sein. In diesem Zuge sind ebenfalls die zu betrachtenden Restriktionen zu erarbeiten, beispielsweise müssen der aktuelle Rüstzustand der Anlage und die bereits in der Produktion befindlichen Aufträge berücksichtigt werden. Die Zielstellung der Neuzuordnung von Zwischenprodukten zu Kundenaufträgen kann somit als multikriterielles Optimierungsproblem aufgefasst und durch ein mathematisches Programm formalisiert werden.

Bestehende Arbeiten zur Optimierung der Produktionsreihenfolge in Warmwalzwerken betrachten als Zielstellungen u. a. die benötigte Produktionszeit [355, 386], Rüstkosten und Rückstellkosten [384] sowie Termintreue [377]. Als Restriktionen werden unter anderem Rüstzeiten, Maschinenkapazitäten, frühest-/spätestmögliche Starttermine oder Zwangsreihenfolgen der Aufträge berücksichtigt, jedoch nicht der aktuelle Rüstzustand der Anlagen. Für den Fall einer sich zyklisch wiederholenden Rüstreihenfolge soll untersucht werden, ob das Optimierungsproblem auf das Rundreiseproblem (Travelling Salesman Problem - TSP) zurückgeführt werden kann, wenn die Distanzen im Anwendungsfall die Kosten für erforderliches Rüsten, nötige Steuerungseingriffe oder resultierende Lieferterminverschiebungen abbilden. Da das resultierende Optimierungsproblem, insbesondere auch das TSP, aufgrund der Vielzahl zu berücksichtigender Zielgrößen, Restriktionen sowie Aufträgen zur Klasse NP-schwerer Probleme zählt (u. a. [355, 365, 367]), erfolgt die Lösung mittels Heuristiken, wie variabler Nachbarschaftssuche, genetischen Algorithmen oder Simulated Annealing.

Die verschiedenen Lösungsansätze für das Optimierungsproblem sollen insbesondere hinsichtlich der benötigten Laufzeit und Lösungsqualität untersucht werden. Insbesondere die zulässige Laufzeit ist stark begrenzt, da die Abstände zwischen aufeinanderfolgenden Bearbeitungsschritten in automatisierten Produktionssystemen i. d. R. nur wenige Sekunden betragen. Um dennoch eine Lösung zu generieren, sollen hier Ansätze entwickelt und hinsichtlich ihrer Lösungsqualität untersucht werden, die das Optimierungsproblem nur für eine Teilmenge der Kundenaufträge untersuchen, beispielsweise Onlinelösungsansätze.

Um die Komplexität der Problemstellung der Auftragszuordnung weiter zu reduzieren, sollen die individuellen Kundenanforderungen zu Klassen ähnlicher Anforderungen zusammengefasst werden. Dazu können unüberwacht lernende Verfahren des Data Mining wie beispielsweise clusteranalytische Verfahren eingesetzt werden. Repräsentiert werden die einzelnen Klassen schließlich durch die Minimal- und Maximalausprägungen hinsichtlich der betrachteten Produkteigenschaften. Für das zu entwickelnde Zuordnungsproblem sind dabei insbesondere die unteren Grenzen dieser Eigenschaften relevant. Als Ergebnis dieses Arbeitspakets steht eine Methodik zur Verfügung, die basierend auf den erstellten Qualitätsprognosen die Möglichkeit bietet, die aktuell bearbeiteten Produkte flexibel solchen Kundenaufträgen zuzuordnen, deren Qualitätsanforderungen die Produkte erfüllen. So können die Eingriffe in die Prozessparameter minimiert werden, ohne dass dies zu Ausschuss oder Nacharbeit führen würde.

Arbeitspaket 6. Ansätze zur Einbindung der Konzepte in eine adaptive Prozessregelung: Durch die Arbeiten im Rahmen der ersten Förderperiode sowie den Inhalten der Arbeitspakete 4 und 5 stehen drei unterschiedliche Ansätze zur Verfügung, die dazu eingesetzt werden können, Ausschuss und Nacharbeit in verketteten Produktionsprozessen zu minimieren. Ziel dieses Arbeitspakets ist es, diese Ansätze zu einer Prozessregelungsmethodik zusammenzuführen, die unter Berücksichtigung aller drei Optionen (Zuordnung zu Kundenaufträgen, Parameteranpassung und Ausschleusen), die bestmögliche Steuerungsentscheidung ableitet. Die Arbeit in diesem Arbeitspaket ist somit stark mit den Inhalten der Arbeitspakete 4 und 5 verknüpft, um sicherzustellen, dass die dort entwickelten Teillösungen zu einem Gesamtvorgehen zusammengefügt werden können.

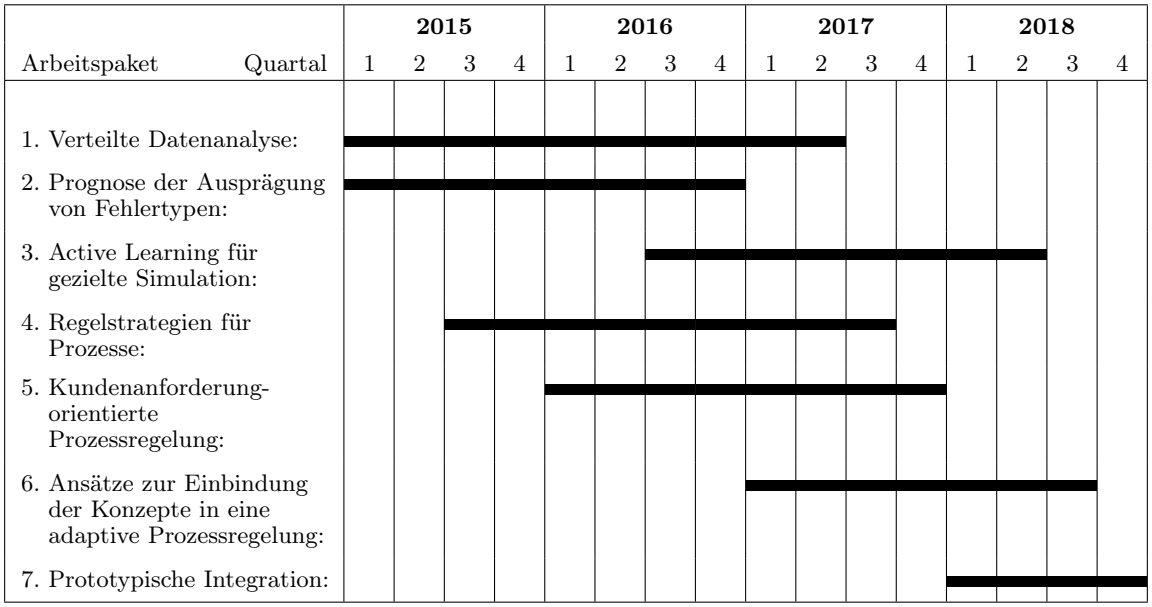
Grundgedanke dieser realzeitlichen, adaptiven Regelung ist der Soll-Ist-Vergleich von Qualitätsprognosen und Kundenanforderungen für das jeweilige Produkt, auf Basis dessen relevante Qualitätsabweichungen erkannt und geeignete Steuerungsentscheidungen abgeleitet werden. Die Ableitung der Steuerungsentscheidung muss als wesentliche Restriktionen sowohl die Freiheitsgrade hinsichtlich der Ausgleichung der Qualitätsabweichungen als auch die Ressourcenbeschränkung Zeit berücksichtigen. Letztere ergibt sich daraus, dass zur Ableitung der drei Steuerungsoptionen unterschiedliche Berechnungszeiten erforderlich sind.

Diesen Berechnungszeiten stehen wiederum unterschiedliche Zeitkontingente gegenüber, innerhalb derer die Qualitätsprognose durchgeführt und eine darauf basierende Entscheidung getroffen werden muss. Das zur Verfügung stehende Kontingent hängt im Wesentlichen davon ab, an welcher Stelle im Bearbeitungs-

prozess eine Entscheidung abgeleitet werden soll, und lässt sich aufgrund von Hintergrundwissen über den Bearbeitungsprozess bestimmen. Die Methodik muss Optionen ausschließen können, die das zur Verfügung stehende Zeitkontingent überschreiten würden. Dazu ist es erforderlich, für die Datenanalyse und nachfolgende Steuerungsentscheidungen benötigte Zeiten exakt abzuschätzen oder zu garantieren. Im Rahmen des Arbeitspakets soll daher versucht werden, die WCET für die verschiedenen an der Steuerungsentscheidung beteiligten Module für eine konkrete Hardwareplattform abzuleiten.

Arbeitspaket 7. Prototypische Integration: Bereits in der ersten Förderperiode wurden Ansätze zur Integration des Jidoka-Konzepts entwickelt und erprobt. Diese Konzepte sollen im Rahmen dieses Arbeitspakets weiterentwickelt und um die in der zweiten Förderperiode entwickelten Konzepte zur Prozessregelung erweitert und schließlich praktisch erprobt werden. Dabei ist besonders die Bedeutung der Integration der unterschiedlichen Datenquellen in die Prozessregelung hervorzuheben. Dazu sind Konzepte zu entwickeln, wie Daten über Kundenaufträge und Anforderungen so zusammengefasst und laufend aktualisiert werden können, dass diese für die Regelung nutzbar gemacht werden können. Gleiches gilt für die Daten der Prozesssimulationen, die zur Prozessparameteradaption herangezogen werden. Insbesondere sind mögliche Ansätze zur dezentralen Einbettung der einzelnen Verfahren in die Prozesse der betrachteten Fallstudie zu entwickeln und zu überprüfen.

Zeitplan



3.5 Stellung innerhalb des Sonderforschungsbereichs

Das Teilprojekt B3 arbeitet bei der Analyse von Sensordaten teils an ähnlichen Fragestellungen wie das Teilprojekt A1. In beiden Teilprojekten liegen Zeitreihendaten vor, in denen interessante oder für eine bestimmte Zielgröße relevante Muster entdeckt werden sollen. Auch spielt in beiden Fällen die Ressourcenbeschränkung Energie eine Rolle. Eine in A1 entwickelte Methode zum Clustering von Zeitreihen mittels DTW wurde bereits erfolgreich in B3 zur Identifikation unterschiedlicher Betriebsmodi angewandt. Auch das *streams framework* wurde schon zur Vorverarbeitung der Zeitreihendaten in B3 eingesetzt. Die in A1 oder auch dem Teilprojekt A2 entwickelten Streamingansätze könnten zudem bei der realzeitlichen Analyse von Sensordaten hilfreich sein. Umgekehrt lassen sich möglicherweise in B3 entwickelte Ansätze zur Merkmalsgewinnung auf die in A1 erhobenen Daten übertragen. Ein wesentlicher Unterschied zu A1 besteht darin, dass die Daten in B3 an verschiedenen Sensoren und Fertigungsstationen verteilt erhoben werden, während die Daten in A1 lokal im gleichen Gerät erfasst und auch verarbeitet werden.

Teilprojekt A3 arbeitet an parallelen Verfahren zur Modellselektion und Optimierung von Hyperparametern. Die Suchstrategien betrachten die auszuführenden Lernverfahren dabei als Black Box, während die in B3 zu entwickelnden Verfahren u. a. Wissen über die Form gelernter Modelle ausnutzen sollen. Zudem

liegt der Fokus von A3 auf den Ressourcen Laufzeit und Speicher, während B3 auch die durch die Verteilung entstehenden Kommunikationskosten senken möchte. Diese spielen in A3 eine wesentlich geringere Rolle, da die Instanzen eines parallelen Algorithmus entweder Zugriff auf einen gemeinsamen Speicher haben oder über Leitungen mit hoher Bandbreite kommunizieren. Dies ist bei den im Kontext von Industrie 4.0 oftmals erwähnten drahtlosen Sensornetzwerken nicht zwangsläufig der Fall. Letztlich entspricht die Aufteilung von Trainings- und Testmengen in A3 außerdem einer horizontalen Verteilung, während die Daten bei Produktionsprozessen in der Regel vertikal verteilt vorliegen.

Eine wichtige Rolle für die Analyse der WCET und die Integration der in B3 entwickelten Methoden in reale eingebettete Systeme könnten die in Teilprojekt A4 entwickelten Methoden spielen. Während A4 den Energieverbrauch in verteilten drahtlosen Sensornetzwerken aus technischer Sicht modellieren und ggf. reduzieren möchte, konzentriert sich B3 auf verteilte Algorithmen zur Datenanalyse. Diese könnten möglicherweise die in A4 entwickelte allgemeine Schnittstelle zur Abfrage des aktuellen Energieverbrauchs oder der Restladung von Akkus nutzen.

Die zeitlichen Abhängigkeiten der in B3 erfassten Wertereihen und ihre Zusammenhänge mit dem Label lassen sich ggf. auch durch Graphen modellieren bzw. mittels Graphkernen analysieren. Somit könnten sich Zusammenhänge mit in Teilprojekt A6 untersuchten Fragestellungen ergeben. Hierbei ist allerdings immer die in B3 gegebene zeitliche Ressourcenbeschränkung zu berücksichtigen.

Ein weiteres Projekt, das Sensordaten erfasst, ist das Teilprojekt B4. Es ist zu untersuchen, inwiefern die in B3 entwickelten Methoden zum Lernen von Prognosemodellen für vertikal verteilte Daten auf die in B4 gegebenen Fragestellungen übertragbar sind. Umgekehrt könnten sich die in B4 untersuchten relationalen Modelle als relevant für die Modellierung der in B3 gegebenen Abhängigkeiten zwischen Wertereihen erweisen.

Der in B3 entwickelte Ansatz zur verteilten Anomaliedetektion, insbesondere die Techniken der CVM, eignet sich prinzipiell für große Datenmengen und Daten, bei denen die Beobachtungen einer Klasse entweder sehr häufig oder nur äußerst selten vorkommen. Eine ähnliche Art der Verteilung ist in Teilprojekt C3 gegeben, bei dem äußerst selten vorkommende Neutrinoereignisse identifiziert werden sollen. Es soll überprüft werden, ob sich die 1-class-Variante der CVM zur Analyse der in C3 aufgezeichneten Daten eignet.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

CSC

(Deuse,Förderkennzeichen BMBF 02PJ2655) (Förderzeitraum: 2013 - 2016)

Im Rahmen des Projektes „Cyber System Connector - Maschinendokumentation intelligent erstellen und nutzen“ soll eine aktuelle technische Dokumentation durch ein virtuelles Abbild der Anlagen über den gesamten Produktentstehungsprozess gewährleistet werden, um den enormen Aufwand der technischen Dokumentation für heutige Anlagen zu reduzieren. Die Professur für Arbeits- und Produktionssysteme ist Konsortialmitglied.

CSR

(Deuse,Förderkennzeichen BMAS E015-NW-015) (Förderzeitraum: 2012 - 2014)

Das Ziel des Projektes „Change pro Corporate Social Responsibility“ ist die systematische und stufenweise Verankerung der Werte von CSR in der Unternehmensführung und -kultur kleiner und mittelständischer Unternehmen aus dem Bereich Metall-, Maschinenbau und Industrieelektronik des Netzwerks Industrie RuhrOst (NIRO) und die damit einhergehende Stärkung der Wettbewerbsfähigkeit. Die Professur für Arbeits- und Produktionssysteme ist Projektpartner.

Ergokom

(Deuse,Förderkennzeichen AiF 440 ZN) (Förderzeitraum: 2012 - 2014)

Das Projekt „Entwicklung und technische Integration einer Bewertungsmethodik zur Ermittlung von Mitarbeiterbelastungen in Kommissioniersystemen“ zielt darauf ab, eine spezifizierte und aufwandsarme Methodik zur Analyse, Bewertung und Visualisierung auftretender körperlicher Belastungen in Kommissioniersystemen unter Berücksichtigung anerkannter Arbeitsanalyseverfahren zu entwickeln. Die Professur für Arbeits- und Produktionssysteme ist ausführende Forschungsstelle.

IndivA

(Deuse,Förderkennzeichen BMBF 16SV6252) (Förderzeitraum: 2014 - 2016)

Das Ziel des Projektes „Individualisierte soziotechnische Arbeitsplatzassistenz für die Produktion“ ist die Entwicklung sowie der Aufbau eines hochflexiblen hybriden Montagesystems mit einer fähigkeitsorientierten und situationsabhängigen Arbeitsteilung zwischen Mensch und Maschine. Wesentliche Bestandteile

des Projektes sind die Ermittlung der individuellen anthropometrischen und biomechanischen Eigenschaften zur Analyse der physischen Leistungsfähigkeit des Beschäftigten sowie die Weiterentwicklung der Parametrisierung digitaler Menschmodelle. Die Professur für Arbeits- und Produktionssysteme ist Konsortialmitglied.

NED

(Deuse,Förderkennzeichen 34-78/V51A/19010/Ta) (Förderzeitraum: 2013 - 2015)

Das Forschungsvorhaben „NED - Netzwerk-Exzellenz-Demografie“ zielt auf einen netzwerkübergreifenden Veränderungsprozess im Bereich des demografischen Wandels. Dazu werden Leitbilder entwickelt, anhand derer Unternehmensprozesse derart gestaltet werden können, dass diese den Anforderungen des demografischen Wandels Rechnung tragen. Die entwickelten Ansätze werden in den beteiligten Unternehmen implementiert und evaluiert. Die Professur für Arbeits- und Produktionssysteme ist Konsortialmitglied.

NeFIna

(Deuse,Förderkennzeichen HBS S-2012-543-1) (Förderzeitraum: 2012 - 2014)

Das Ziel des Projektes „Neue Formen von Industriearbeit - Das Beispiel Siemens-Produktionssystem“ ist es, die Voraussetzungen, Einführungsprozesse und Konsequenzen der Einführung des Konzeptes ganzheitlicher Produktionssysteme (GPS) in Bezug auf sozial- und arbeitswissenschaftliche Aspekte zu untersuchen. Die Professur für Arbeits- und Produktionssysteme ist Projektpartner.

Pro Mondi

(Deuse,Förderkennzeichen BMBF 02PJ1110) (Förderzeitraum: 2012 - 2015)

Im Rahmen des Projektes „Prospektive Ermittlung von Montagearbeitsinhalten in der digitalen Fabrik“ soll eine Methode zur vorausschauenden Ermittlung von Montagearbeitsinhalten in einer frühen Phase des Produktentstehungsprozesses (PEP) entwickelt werden. Die Professur für Arbeits- und Produktionssysteme ist Konsortialmitglied und Projektkoordinator.

ReiMaFlu

(Deuse,Förderkennzeichen AiF 17495 N/1) (Förderzeitraum: 2012 - 2014)

Das Projekt „Reinheitsgerechte Materialflusssysteme - Entwicklung eines Leitfadens zur reinheitsgerechten Gestaltung von Materialflüssen entlang des Wertstroms“ hat das Ziel der Entwicklung von Konzepten zur Gewährleistung der Bauteilsauberkeit bei innerbetrieblichen Materialflussoperationen. Die Professur für Arbeits- und Produktionssysteme ist ausführende Forschungsstelle.

variANa

(Deuse,Förderkennzeichen AiF 17531 N) (Förderzeitraum: 2012 - 2014)

Im Rahmen des Projektes „Multivariante Analyseverfahren zur Strategiewahl in der Nachserie“ soll die vereinfachte und optimierte Strategieauswahl für Elektronikbaugruppen durch Bildung von Ersatzteilklassen mittels strukturabbildender Verfahren erreicht werden. Die Professur für Arbeits- und Produktionssysteme ist ausführende Forschungsstelle.

Vista-TV

(Morik,Förderkennzeichen EU STREP 296126) (Förderzeitraum: 2012 - 2014)

Aufgabe des Projektes ist die Verbindung bestehender Verfahren zu einem neuartigen System für Open Linked Data. Das Projekt annotiert Internetfernsehprogramme automatisch und empfiehlt Benutzern darauf aufbauend Sendungen. Die Projektarbeit konzentrierte sich auf die Entwicklung einer Infrastruktur für die Verbesserung der Daten bei Zattoo und BBC und die Anwendung vorhandener Lernverfahren. Das in A1 entwickelte streams framework wurde erfolgreich im Projekt angewandt, um aus Videodaten Merkmale zu extrahieren.

Insight

(Morik,Förderkennzeichen EU 318225) (Förderzeitraum: 2012 - 2015)

Das Projekt behandelt Big Data in Bezug auf Volumen, Heterogenität und Realzeitanforderung. Wir arbeiten an raumzeitlichen Prozessen zur Vorhersage und Steuerung der Verkehrsflüsse in Dublin. Aus diesem Projekt wollen wir die Daten für das Teilprojekt B4 nehmen. Das streams framework wurde hier zur Ereigniserkennung eingesetzt.

Kobra

(Morik,Förderkennzeichen BMBF 01UG12445A) (Förderzeitraum: 2012 - 2015)

In diesem Projekt geht es um Sprachdaten und deren Analyse. Es hängt mit den Arbeiten im SFB 876 kaum zusammen.

Intelligente Prozess-Steuerung auf Basis von Prozessmerkmalen (Morik,Förderkennzeichen SMS Siemag) (Förderzeitraum: 2011 - 2014)

Das Projekt wird direkt von SMS Siemag gefördert. Es wurde im ersten Jahr eine Anwendung des maschinellen Lernens für die Bestimmung des Blasendes bei einem BOF im Stahlwerk Dillinger Hütte entwickelt. Nach einer Pause wurde in einer Fortsetzung eine Infrastruktur zur Erfassung und Speicherung der Daten für das Lernen entwickelt. Hierbei wurde auch die Anwendung des streams frameworks vorgesehen. Im Gegensatz zum Teilprojekt B3, bei dem der Fokus auf der Entwicklung neuer Verfahren für die Analyse verteilter Sensordaten liegt, wird hier nur ein einzelner Schritt der Produktion betrachtet. Die verteilte Anordnung von Sensoren spielt keine Rolle.

Data Driven Material Development (Morik,Förderkennzeichen Mercator Stiftung, Merkur Pr-2012-0023) (Förderzeitraum: 2012 - 2014)

Die Daten zur Untersuchung verschiedener Mischungen chemischer Substanzen in Bezug auf bestimmte Materialeigenschaften sollen effizienter genutzt werden, so dass aus wenigen Messpunkten die Eigenschaften an allen möglichen Messpunkten vorhergesagt werden können.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	117.600	—	117.600	—	117.600	—	117.600
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	117.600		117.600		117.600		117.600	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (B3)							
Wissen-schaftliches Personal	1	Morik, Katharina, Prof. Dr., Prof	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	4,0	—	Grund-ausstattung
	2	Deuse, Jochen, Prof. Dr., Prof	Produktionssysteme	LS APS, Fak. Maschinenbau, TU Do	4,0	—	Grund-ausstattung
	3	Blom, Hendrik, Dipl.-Inf. WiMi	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	29,85	—	Grund-ausstattung
	4	Konrad, Benedikt, Dipl.-Wirt.-Ing. MSIE (USA) WiMi	Produktionssysteme	LS APS, Fak. Maschinenbau, TU Do	29,85	—	Grund-ausstattung
	5	Buschjäger, Sebastian, SHK	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	8,0	—	Grund-ausstattung
	6	N.N. , SHK	Produktionssysteme	LS APS, Fak. Maschinenbau, TU Do	8,0	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	7	Kotala, Eugen, Techn. Ang.	—	LS APS, Fak. Maschinenbau, TU Do	9,95	—	Grund-ausstattung
Beantragte Personalmittel (B3)							
Wissen-schaftliches Personal	8	Wiegand, Mario, M.Sc. WiMi	Produktionssysteme	LS APS, Fak. Maschinenbau, TU Do	—	Dokto-rand/in	—
	9	Stolpe, Marco, Dipl.-Inf. WiMi	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	—	Dokto-rand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Morik, Katharina**
   Projektleitung. Schwerpunkt (verteilte) Datenanalyse. Mitarbeit in AP 1 - 3 sowie AP 7.
2. **Deuse, Jochen**
   Projektleitung. Schwerpunkt Produktionssysteme. Mitarbeit in den AP 4 - 7.
3. **Blom, Hendrik**
   Mitarbeit an den AP 2, 3 und 7.
4. **Konrad, Benedikt**
   Mitarbeit in den AP 3 - 7.
5. **Buschjäger, Sebastian**
   Unterstützung bei der Implementierung von Algorithmen zur Datenanalyse.
6. **N.N.,**
   Mitarbeit in den AP 3 - 7 sowie Unterstützung bei der Dokumentation durchgeführter Arbeiten.
7. **Kotala, Eugen**
   Unterstützung bei der Sammlung und Aufbereitung relevanter Daten und bei der Bereitstellung der technischen Infrastruktur.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

8. **Wiegand, Mario**
   Mitarbeit in den Arbeitspaketen 3 - 7 mit Bezug auf die Produktionssysteme und Operations Research.
9. **Stolpe, Marco**
   Mitarbeit in den Arbeitspaketen 1, 3 und 7 mit Bezug auf verteilte Datenanalyse und Active Learning.

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	10.000	10.000	10.000	10.000
Summe der vorhandenen Mittel	10.000	10.000	10.000	10.000
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt B4

3.1.1 Titel:

Analyse und Kommunikation für die dynamische Verkehrsprognose

3.1.2 Fachklassifizierung:

Data Mining, Kommunikationsnetze, Verkehrsmodellierung

3.1.3 Projektleitung

Prof. Dr. Kersting, Kristian, 28.11.1973

LS 8, Fakultät für Informatik, Technische Universität Dortmund
Joseph-von-Fraunhofer-Straße 23
44227 Dortmund

Tel.: 0231-755-7876
Fax : 0231-755-5105
E-Mail: `kristian.kersting@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Schreckenberg, Michael, 28.09.1956

Physik von Transport und Verkehr, Fakultät für Physik, Universität Duisburg-Essen
Lotharstraße 1
47048 Duisburg

Tel.: 0203-379-3552
Fax : 0203-379-6564
E-Mail: `michael.schreckenberg@uni-due.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr.-Ing. Wietfeld, Christian, 21.01.1966

Lehrstuhl für Kommunikationsnetze, Fakultät für Elektrotechnik und Informationstechnik, Technische Universität Dortmund
Otto-Hahn-Straße 6
44227 Dortmund

Tel.: 0231-755-4515
Fax : 0231-755-6136
E-Mail: `christian.wietfeld@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Das Ziel dieses Teilprojekts ist die Erforschung hochpräziser Prognoseverfahren für das dynamische Verhalten des Straßenverkehrs basierend auf ressourceneffizienter, mobilfunkgestützter Übertragung von extended Floating Car Data (xFCD) und weiteren Datenquellen.

In der ersten Förderperiode ist die Optimierung von Verkehrsströmen mit Simulationen auf abgegrenzten, mehrspurigen Einzelstrecken, wie z.B. Autobahnen, betrachtet worden. In der zweiten Förderperiode soll die Optimierung in komplexen Netzinfrastrukturen, insbesondere im Innenstadtbereich, im Fokus stehen. Durch die Hinzunahme von Kristian Kersting als drittem Projektleiter gewinnt das Teilprojekt Expertise im Bereich des relationalen Data Minings. Diese wollen wir nutzen, um neue Methoden des Data Minings für die Anwendung in Verkehrssimulationen und Kommunikationssystemen verfügbar zu machen.

Als Datenquellen werden u.a. umfangreiche Verkehrsdaten aus Dortmund, Düsseldorf und Dublin zur Verfügung stehen. Die Auslöser für Störungen des Fahrzeugflusses sollen durch die von den Fahrzeugen gesammelten Daten frühzeitig erkannt und daraus in Echtzeit Gegenmaßnahmen eingeleitet werden. Bisher bekannt sind Ansätze, die das mikroskopische Fahrzeugverhalten beeinflussen, z.B. durch Geschwindigkeitsanpassungen für eine „Grüne Welle". Ein neuer Ansatz dieses Teilprojekts basiert hingegen auf einer dynamischen Anpassung der Flusseigenschaften des Straßennetzes. Hierfür sollen verschiedene Strategien zur dynamischen Anpassung der Netzstruktur analysiert werden, z.B. Flussrichtungsänderung oder Sperrung, Spurbindung oder spurungebundenes Fahren. Darüber hinaus ist die Entwicklung dynamischer, mikroskopischer Verkehrsmodelle vorgesehen, die basierend auf Empfehlungen des Data Minings in Echtzeit neu parametrisierbar sind und damit die Heterogenität des Verkehrs in der innerstädtischen Netzstruktur besser abbilden können. Globalziele sind dabei, die durchschnittlichen Fahrzeiten und/oder den Energiebedarf zu reduzieren. Ebenfalls von Interesse sind hier die Verteilungsfunktionen z.B. der Fahrzeiten.

Ein wesentliches Ziel ist es, diese Verteilungsfunktionen automatisch aus ganzzahligen Messungen zu bestimmen. Hierbei sollen relationale Modelle wie z.B. Markov Logic Networks untersucht werden, in denen die Vorhersage der Fahrzeit auf einem Straßenabschnitt die Vorhersage an benachbarten oder ähnlichen Segmenten beeinflusst. Insbesondere sollen Methoden für relationale Generalisierte Lineare Modelle (GLMs) entwickelt werden - Modellierung, Inferenz und Lernen - und der Nachweis erbracht werden, dass sie bessere Vorhersagen z.B. der Fahrzeiten erlauben als existierende Ansätze.

Zur Umsetzung der Strategien ist in der zweiten Förderperiode die Betrachtung bidirektionaler Echtzeit-Kommunikation notwendig, um neben der fortdauernden Erfassung von Sensordaten in den Fahrzeugen nun auch die aktive Steuerung der Fahrzeuge und der dynamischen Netzinfrastruktur zu ermöglichen. Um die Wechselwirkung dieser *Machine-Type Communication (MTC)* mit dem durch Menschen veranlassten Kommunikationsverkehr zu steuern, soll eine Vorhersagekomponente zur erwarteten Dauer eines Kanalzustands integriert werden, um den Sendezeitpunkt für Sensor- oder Steuerdaten eines fahrenden Fahrzeugs z.B. je nach örtlicher Netzbelastung beeinflussen zu können. Dies ermöglicht eine ressourceneffiziente Nutzung des Kommunikationsnetzes. Für die Kanalprädiktion wird eine langfristige Mobilitätsvorhersage benötigt, die mit Verfahren des maschinellen Lernens erfolgen soll. Die zu entwickelnden Verfahren sollen auch moderne

Technologien wie LTE-A, insbesondere *Carrier Aggregation*, energieeffiziente MTC, Übertragung in heterogenen Netzen und *CoMP (Coordinated Multipoint)* unterstützen.

Die Verfahren zur satellitengestützten spurgenauen Lokalisierung sollen in der zweiten Förderperiode experimentell im Innenstadtumfeld validiert und insbesondere über Verfahren zur kombinierten Nutzung von unterschiedlichen Satellitensystemen (Galileo, GPS, GLONASS, etc.) weiterentwickelt werden. Hierbei soll die Ortungsgenauigkeit durch Methoden der Mustererkennung weiter erhöht werden.

3.3 Bisherige Entwicklung des Teilprojekts

3.3.1 Bericht und Stand der Forschung

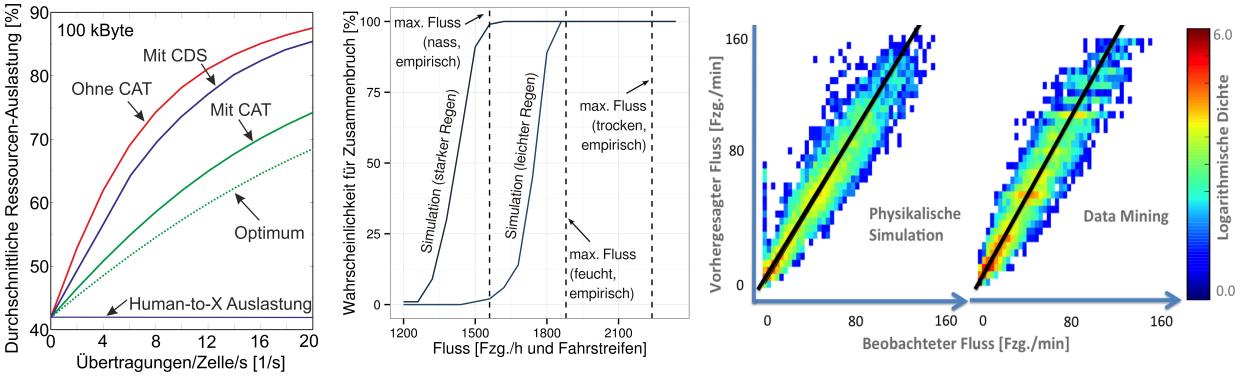
In der ersten Förderperiode stehen drei Forschungsschwerpunkte im Vordergrund: Die ressourceneffiziente Bereitstellung von Sensorinformationen, die zuverlässige Positionserkennung von Fahrzeugen und die verbesserte Verkehrsprognose.

Die fahrzeugbezogene Datenkommunikation ist ein aktuelles Forschungsthema bei der Entwicklung von modernen Kommunikationssystemen und insbesondere Mobilfunknetzen [407]. Dies gilt auch für die Bereiche der Verkehrsbeeinflussung [433] und des energieeffizienten Fahrens [406]. Im Gegensatz zu anderen Ansätzen (z.B. eCall [418]), bei denen die Fahrzeugkommunikation auf Global System for Mobile Communications (GSM) basiert, liegt der Fokus bei den in diesem Teilprojekt erarbeiteten Analysen auf dem deutlich zukunftsorientierten Long Term Evolution (LTE) Standard. Im Hinblick auf die notwendige ressourceneffiziente, mobilfunkgestützte Bereitstellung von *extended Floating Car Data (xFCD)* können folgende Ergebnisse hervorgehoben werden:

Zur quantitativen Modellierung und Analyse der Wechselwirkung zwischen der Übertragung von xFCD und endkunden-orientierten Mobilfunkdiensten (*Human-to-X*) wurde ein *Multi-class Erlang Loss Model* entwickelt. Dieses beschreibt die Auslastung von LTE-Mobilfunkzellen unter heterogenem Kommunikationsverkehr. Im Gegensatz zu einer Modellierung von LTE-Unterträgern [442], arbeitet das Modell realitätsnäher auf Basis von *Resource Blocks (RBs)*. Die Wechselwirkung der unterschiedlichen Serviceklassen wird dabei in Form einer Blockierungswahrscheinlichkeit analytisch berechnet. Zur Gewährleistung von aussagekräftigen Ergebnissen wurde eine realitätsnahe Parametrisierung des Modells durchgeführt. Diese basiert auf Strahlenverfolgungsberechnungen (Ray-Tracing-Analysen), um die Kanaleigenschaften zwischen Mobilfunkbasisstation und im Fahrzeug installiertem Endgerät für ein konkretes Szenario besonders exakt modellieren zu können, und auf LTE-Datenratenmessungen im Labor (Zusammenarbeit mit Teilprojekt A4). Basierend auf den kanalabhängigen Datenraten werden die Bedienprozesse für unterschiedliche Kanaleigenschaften des *Multi-class Erlang Loss Models* angepasst. Dieses Vorgehen ermöglicht eine besonders genaue Kosten-Nutzen-Analyse zwischen der Belastung des öffentlichen Mobilfunknetzes und der Menge an bereitgestellten Daten für die Verkehrsprognose, welche in [B4/460] veröffentlicht wurde.

Zur ressourcenoptimierten Übertragung der xFCD-Daten wurde darüber hinaus ein funkkanalsensitives Kommunikationsverfahren entwickelt: Channel-Aware Transmission (CAT) [B4/A4/459]. Es konnte beispielsweise gezeigt werden, dass mit Hilfe von CAT die Anzahl an freien Verbindungen bei konstanter Kommunikationsdauer um 75 %, bei gleichzeitiger Reduktion der Leistungsaufnahme des Endgeräts um bis zu 35 % (Kooperation mit A4), gesteigert werden kann. Im Vergleich zu anderen kanalsensitiven Kommunikationsverfahren (z.B. Channel-Dependent Scheduling [408], [444]), welche zentralisiert sind und somit eine Anmeldung des Teilnehmers an die Basisstation voraussetzen, arbeitet CAT besonders ressourceneffizient durch eine dezentrale Funktionsweise. Dabei wird die Sendeentscheidung, anders als bei anderen dezentralen Algorithmen zur Bestimmung von verbesserten Sendezeitpunkten, welche auf Schwellwertfunktionen basieren (siehe z.B. [451]), durch einen stochastischen Prozess, welcher die Anwendung, die Kanaleigenschaften und die Fahrzeuggeschwindigkeit berücksichtigt, gesteuert. Dies gewährleistet eine besonders ressourcenschonende Übertragung von relevanten xFCD (siehe Abb. 3.1a), z.B. von langsam fahrenden Fahrzeugen auf der Autobahn unter Berücksichtigung von tolerierbaren Verzögerungen (siehe [B4/461]). Außerdem wurde mit Hilfe eines Verfahrens zur Koalitionsbildung der Informationsgehalt der Objektflussdaten gesteigert. Es konnte mit Hilfe einer gekoppelten Verkehrs- und Kommunikationssimulation gezeigt werden, dass das xFCD-Verkehrsaufkommen um bis zu 50 % durch Fahrzeug-Fahrzeug-Kommunikation reduziert werden kann.

In der zweiten Förderperiode soll das CAT-Verfahren zur ressourceneffizienten Übertragung von Sensordaten (für das Mobilfunknetz sind dieses Hintergrunddaten, ohne direkte menschliche Interaktion) um eine Vorhersagekomponente erweitert werden (*predictiveCAT*). Erste Vorarbeiten zu dieser Erweiterung haben wir bereits in [B4/A4/468] veröffentlicht. In Gegensatz zu anderen Prädiktionsverfahren von Kommunikationsgüten, welche insbesondere auf Wireless Local Area Network (WLAN) und Universal Mobile Telecommunications System (UMTS) beruhen (z.B. [441]), soll in diesem Teilprojekt der aktuellste Mobilfunkstandard



(a) Reduzierung der LTE-Ressourcenauslastung durch CAT mit Vergleich zu Channel-Dependent Scheduling (CDS) (vgl. [B4/A4/459]).

(b) Wahrscheinlichkeit eines wetterinduzierten *Traffic Breakdowns* für verschiedene Stärken des Wettereinflusses in der Simulation.

(c) Verkehrsflussmodellierung auf dem Kölner Ring mittels physikalischer Simulation (links) und Poisson Dependency Networks (rechts). Der Scatterplot zeigt das Verhältnis von vorhergesagten zu beobachteten Flüssen an. Je dichter die Punktwolke sich an die Diagonale schmiegt, desto besser.

Abbildung 3.1.: Ergebnisse der ersten (a,b) und Vorarbeiten für die zweite Förderperiode (c).

LTE-Advanced verwendet werden. Dabei werden Verfahren von LTE-Advanced, wie z.B. Carrier Aggregation [449] und Beamforming [425], insbesondere unter Einfluss von Mobilität [455], berücksichtigt.

Die Zuverlässigkeit der Prädiktion von Objektströmen, wie zum Beispiel dem Straßenverkehr, steht im direkten Zusammenhang mit der örtlichen wie auch zeitlichen Genauigkeit der Detektion von verkehrsbeeinflussenden Ereignissen. Deshalb wurde eine Möglichkeit entwickelt, Flussobjekte im Straßenverkehr satellitenbasiert möglichst genau und unabhängig von den jeweiligen Gegebenheiten zu positionieren und somit die örtliche und zeitliche Auflösung der Datenquellen zu erhöhen.

Herausfordernd dabei ist jedoch die vor allem in herausfordernden Umgebungen zu hohe Ungenauigkeit von > 15 m bei handelsüblichen *GPS*-Empfängern [453]. Existierende Verbesserungsmethoden und -ansätze zur Fehlervermeidung wie [421] zeigen ohne kostenintensive Zusatzsensorik keine hinreichende Genauigkeitssteigerung um innovative Applikationen im Straßenverkehr zu ermöglichen.

Daher wurde im Zuge der ersten Förderperiode zuerst die Performanz von Multi-GNSS-Systemen analysiert, die resultierende Genauigkeitssteigerung als messbar, jedoch für die geforderten QoS-Parameter als zu gering klassifiziert [B4/466]. Entgegen der meisten existierenden Ansätze wurde im weiteren Verlauf eine *post-operative* Verbesserungsstrategie gewählt, bei der von einem Restfehler ausgegangen und anschließend versucht wird, diesen zu kompensieren.

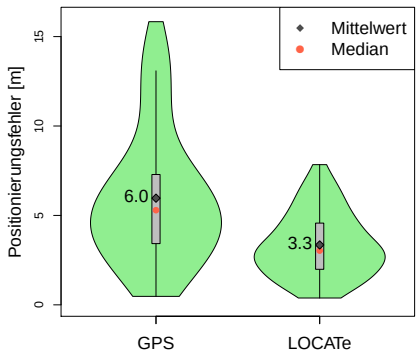


Abbildung 3.2.: Verringerung des Positionsfehlers mittels LOCATe.

Die entwickelte *LOcal interference CompensATion (LOCATe)* [B4/467] wird dabei genutzt, um die entstehenden Fehlereinflüsse für eine spezifische Ort/Zeit-Kombination zu prädizieren, den Gesamtfehler zu quantifizieren und somit im Nachgang zu kompensieren. Im Gegensatz zum *Intelligent Urban Localization Mechanism* in [422] oder der Spurdetektion in [457] sind dafür keine genaue Initialpositionierung, keine Exklusion von NLOS-Satelliten, keine vorgegebenen Satellitenflughöhen und auch keine Kenntnisse über Laternenmasten bei Autobahnfahrten notwendig. Dies ermöglicht vor allem in herausfordernden Umgebungen eine adäquate und verlässliche Positionierung mit einer signifikanten Fehlerreduktion bei der Verwendung von handelsüblichen GPS-Empfängern.

Die Machbarkeit dieses Ansatzes wurde durch eine prototypische Umsetzung auf einem *Advanced Software-Defined GNSS-Receiver (ASDR)* nachgewiesen, mit dem die Genauigkeit ohne Hinzunehmen von zusätzlichen Sensor- oder Ortsinformationen auf ein spurgenaues Level ($> 1{,}5$ m) in bis zu 60 % aller Fälle erhöht wurde. Dabei konnten die maximal auftretenden Beträge der Fehlpositionierungen allein durch LOCATe und ohne Zusatzinformationen oder -hardware von teilweise > 15 m auf ($< 7{,}2$ m) und der mittlere Fehler von 6 m auf 3,3 m verringert werden, was einer signifikanten Fehlerreduktion von 45 % entspricht (siehe Abb. 3.2) [B4/467].

In der zweiten Förderperiode sollen die Ergebnisse weiter analysiert und die Verfahren verbessert werden.

Dabei sollen sowohl die Ortungsgenauigkeit durch Methoden der Mustererkennung als auch die Modellgeneration für LOCATe selbst verbessert werden. In [435] wird ein Weg beschritten, der 3D-Laser-Scanning vom Dach eines Gebäudes zur Aufnahme von Punktwolken der umliegenden Objekte durchführt, um daraus hochgenaue Modelle für weitere Simulationsschritte zu entwickeln. Für LOCATe wurde dieses Verfahren in vereinfachter Form exemplarisch angewandt, doch aufgrund steigender Rechen- und Speicherkapazitäten soll eine vollständige Echtzeitverarbeitung direkt im Fahrzeug möglich werden. Durch kontinuierlich wachsende Datenbanken mit höher aufgelösten Karten- und Umgebungsdaten rücken auch die Plausibilitätsbetrachtungen weiter in den Fokus und werden mit den Methoden der Mustererkennung verknüpft. Durch Sensordatenfusion sollen zudem die im Fahrzeug selbst vorhandenen Daten dazu genutzt werden, Fehlpositionierungen zu erkennen und zu vermeiden. In [426] und [436] wurde bereits gezeigt, dass ein solches Verfahren an Orten mit schwachem GPS-Signal gut funktioniert, um das Fahrzeug auch dort weiter verfolgen zu können. Im Rahmen des vorliegenden Projekts soll das Konzept hingegen so aussehen, dass sämtliche Sensordaten in LOCATe einfließen, um die Genauigkeit des Systems weiter deutlich zu verbessern.

Im Bereich der Verkehrsprognose wurde zunächst eine quantitative Abschätzung der notwendigen Durchdringung von mit FCD-Sendern ausgerüsteten Fahrzeugen vorgenommen, ab der präzise Reisezeitschätzungen auf Basis ausschließlich dieser Daten vorgenommen werden können [B4/460]. Hierzu wurden Simulationen mit einem mikroskopischen Verkehrsmodell auf Basis des *Nagel-Schreckenberg-Modells* mit unterschiedlichen Anteilen von FCD-ausgerüsteten Fahrzeugen durchgeführt. Im Modell wurde eine Straße mit mehreren Engstellen nachgebildet, auf denen sich mit anderen Fahrzeugen 5 %, 1 % oder 0,2 % FCD-sendende Fahrzeuge bewegen. Es konnte gezeigt werden, dass bereits 1 % ausgerüstete Fahrzeuge ausreichen, um eine 94 %-genaue Messung der Reisezeit aller Fahrzeuge zu ermöglichen. Bei 5 % Ausstattungsrate steigt die Genauigkeit der Reisezeitmessung auf fast 97 %. Auf Basis dieser Zahlen wurde ebenfalls in [B4/460] eine Abschätzung mit dem *Multi-class Erlang Loss Model* vorgenommen, wie stark die Kommunikation der Fahrzeuge die bestehenden LTE-Ressourcen zusätzlich belastet. Hier wurde gezeigt, dass bei den genannten Ausstattungsraten nur eine geringe Beeinträchtigung der nicht-fahrzeugbezogenen Mobilfunknutzer zu erwarten ist: In einem Szenario einer Stadtautobahn mit hoher Verkehrsdichte ist die Zahl der möglichen Nutzer bei 5 % Ausstattungsrate um 17 %, bei 1 % jedoch nur um 4 % geringer als ohne Fahrzeugkommunikation, sofern *Channel-Aware Transmission (CAT)* [B4/A4/459] verwendet wird.

Parallel zu diesen Untersuchungen wurde eine Abschätzung möglicher Potentiale von xFCD für die Verbesserung mikroskopischer Verkehrsmodelle vorgenommen. Während FCD nur Basiswerte wie Zeit, Position und Geschwindigkeit der Fahrzeuge umfassen, sind bei Hinzunahme von xFCD-Quellen auch Messwerte aus dem fahrzeuginternen CAN-Datenbus enthalten, die Rückschlüsse auf den Fahrzeugbetrieb (z.B. Kraftstoffverbrauch), die Verkehrslage (z.B. Häufigkeit des Bremsens) oder die Umgebungsbedingungen (z.B. Tätigkeit von Regen- und Lichtsensor, Häufigkeit der ABS-Eingriffe) ermöglichen. Hierzu wurden Testfahrten von mit Datenloggern ausgerüsteten Fahrzeugen durchgeführt, um Livedaten aus dem CAN-Bus aufnehmen und analysieren zu können. Durch die Auswertung der möglichen Datenquellen des CAN-Bus konnte eine Kategorisierung der Eingangsdaten vorgenommen werden, die Grundlage einer Priorisierung von Sendeentscheidungen sein kann, je nach dem, wie die Nützlichkeit der Information für die Verkehrsprognose eingeschätzt wird [B4/461].

Die Verbesserung der Verkehrsprognose durch Hinzunahme zusätzlicher Datenquellen konnte exemplarisch für sog. Umfelddaten gezeigt werden. Da die Verfügbarkeit von in großer Zahl erhobenen empirischen xFCD durch den bisher unzureichenden Ausstattungsgrad der Fahrzeuge sehr schlecht ist, wurde zur weiteren Analyse auf stationär erhobene Daten zurückgegriffen. Hierzu wurden Daten, die das Wetter und den Straßenzustand an lokalen Punkten des Autobahnnetzes beschreiben, aus dem Datenverteiler des Landes Nordrhein-Westfalens erschlossen. Hierbei handelt es sich um minutengenaue Daten, die laufend erhoben werden. Diese Daten werden permanent in eine eingerichtete Datenbank übertragen, sodass mittlerweile ein Datenbestand von mehreren Jahren für die Analyse zur Verfügung steht. Im Vergleich zu früheren Untersuchungen (siehe [430, 450] für eine Übersicht) ist es mit den beschriebenen Datenquellen möglich, viel präzisere Einflüsse zu erkennen und darzustellen, da in den bisherigen Studien keine Echtzeitinformationen über das Wetter zur Verfügung standen, sondern oft nur stundengenaue Werte. Ebenso findet sich dort auch keine Untersuchung, die fahrstreifen- und fahrzeugtypspezifische Auswertungen bereitstellt.

Beispielhaft konnte aus den Umfelddaten der Einfluss von Regen, Dunkelheit und Außentemperaturen auf das Geschwindigkeitsverhalten auf Autobahnen und damit die Fahrzeiten situationsspezifisch quantifiziert werden. Bei diesen drei Beobachtungsgrößen ist ein Rückschluss auf die Äquivalenz des Einflusses der Werte bei Messung in einem Fahrzeug einfach möglich. So konnte auf Basis der Minutenwerte gezeigt werden, dass die Freiflussgeschwindigkeiten bei starkem Regen (in Form eines hohen Wasserfilms auf der Straße) um bis zu − 23 % sinken. Regen bei Dunkelheit führt zu einer Zunahme der Fahrzeit um + 15 %. Ein Vergleich der Korrelation zwischen Geschwindigkeit und Wasserfilm im gesamten Datensatz zeigt, dass die Effekte im

Berufsverkehr doppelt so hoch sind wie außerhalb. Zusätzlich wird deutlich, dass die Vorhersagbarkeit eines Geschwindigkeitsmesswertes mit zunehmendem Regen stark nachlässt, während zwischen den Geschwindigkeiten auf den Fahrstreifen eine Synchronisierung stattfindet.

Die empirischen Auswirkungen wurden verwendet, um ein mikroskopisches Verkehrsmodell zu entwickeln, das Wetterinformationen verarbeiten kann. Als Grundlage wurde ausgehend von früheren Entwicklungen ein komplexes Modell [B4/458] entwickelt, dass die Fahrdynamik wesentlich detaillierter abbilden kann als einfache Zellularautomatenmodelle [447], weil es z.B. unterschiedliche Bremsbeschleunigungen je nach Fahrzeug unterstützt sowie realistische Fahrstreifenwechselregeln enthält. Mit den entwickelten Ergänzungen der Modellregeln für Wetterereignisse ist es nun möglich, wetterbedingte Verkehrszusammenbrüche in Verkehrsprognosesystemen zu berücksichtigen (siehe Abb. 3.1b). Derzeit wird untersucht, wie andere Zellularautomatenmodelle in ähnlicher Weise erweitert werden können. Mit dem Ziel, in Echtzeit erhobene Wetterdaten in die Simulation einfließen zu lassen, wurde die am Lehrstuhl von Prof. Schreckenberg entwickelte und im Produktiveinsatz stehende Simulationsumgebung OLSIMv4 zur Implementierung des Modells verwendet, die auch in der zweiten Förderperiode zum Einsatz kommen soll.

In der zweiten Förderperiode wird der Schwerpunkt der Arbeiten auf Verkehrsmodellierung und -steuerung im Bereich von innerstädtischen Straßennetzen liegen. Aufgrund häufig unzureichender Datenlage und der generell erhöhten Komplexität [412] ist der Stadtverkehr bislang im Vergleich zu Autobahnverkehr nur wenig untersucht worden. Dies gilt sowohl für empirische Untersuchungen als auch für die Entwicklung realitätsnaher Simulationsmodelle. Die Beschreibung von Stadtverkehr erfolgt deshalb auch heute noch oft anhand von Theorien, die mitunter einige Jahrzehnte alt sind (siehe [415, 419] für eine Übersicht). Neueste Forschungen u.a. in der Arbeitsgruppe von Michael Schreckenberg haben tiefergehendes Verständnis des Stadtverkehrs ermöglicht. So konnte beispielsweise synchronisierter Verkehr im innerstädtischen Bereich empirisch nachgewiesen werden [B4/462] – eine Erkenntnis, die den bisherigen Theorien widerspricht. Um weitere Erkenntnisse zu ermöglichen, soll den Herausforderungen der Datenlage und Komplexität in der zweiten Förderperiode durch ein neuartiges Zusammenspiel zwischen Verkehrsphysik und Data Mining begegnet werden. Insbesondere die geplante Verbindung zwischen Data Mining und mikroskopischen Verkehrssimulationen zur Analyse und Adaption in Echtzeit ist bislang nicht durchgeführt worden, es finden sich lediglich Belege für durch Data Mining verbesserte Multiagentensysteme im Kontext von Internet-Applikationen [409]. Zellularautomaten sind bislang ebenfalls nur für die reine Implementierung von Verfahren des Data Minings verwendet worden [417].

Durch die zunehmende Verfügbarkeit mobiler Kommunikationsnetze ist die Forschung an Konzepten für den Einsatz kommunizierender Fahrzeuge (Car2X) [443] weit verbreitet. Im innerstädtischen Bereich wird hierbei sehr oft die lokale Kommunikation mit "intelligenten" Lichtsignalanlagen untersucht. Hierfür und für allgemeine Verkehrsanalysen wird häufig die Open-Source-Simulationsumgebung SUMO [432] verwendet, mit der die Projektbeteiligten bereits Erfahrungen gesammelt haben. Neuartige netzweite Strategien zur Vermeidung von Verkehrszusammenbrüchen, wie das *breakdown minimization principle* (BMP) [428], sind bislang im Bereich der Fahrzeugkommunikation für die Verkehrssteuerung noch nicht untersucht worden. Dies ist in der zweiten Förderperiode vorgesehen. Unter den klassischen Ansätzen zur Verkehrssteuerung stellen die Prinzipien von Wardrop [452] sicherlich die bekanntesten dar, die die Minimierung von Fahrzeiten nach dem Nutzergleichgewicht oder dem Systemoptimum fordern. Beide Prinzipien sind spieltheoretisch bereits häufig untersucht worden (z.B. [416, 431]), allerdings werden die Optimierungsziele empirisch häufig nicht erreicht [413]. Das BMP unterscheidet sich von Wardrops Prinzipien durch die Zielfunktion, die erstmals empirische Eigenschaften von Verkehrszusammenbrüchen an Engstellen berücksichtigt (z.B. den Übergang von Freifluss zu synchronisiertem Verkehr als stochastischem Prozess) und die Minimierung der Wahrscheinlichkeit für derartige Zusammenbrüche anstrebt. Die Modellierung dieses Verkehrsverhaltens ist mit mikroskopischen Verkehrsmodellen auf Basis von Zellularautomaten sehr gut möglich [427] und wegen der Komplexität und u.U. Nichtlinearität der zu modellierenden Effekte bei einfacher Implementierung aus unserer Sicht anderen Ansätzen vorzuziehen. Erste Simulationen dieser Art haben gezeigt, dass die Verkehrssteuerung mit dem BMP sogar höhere Verkehrsflüsse erlaubt als die klassischen Ansätze [428], sodass netzweite Simulationen mit der Einbeziehung von Kommunikation zur Steuerung der Fahrzeuge vielversprechend erscheinen.

Den bereits beschriebenen Herausforderungen für die Datenanalyse wird in der zweiten Förderperiode durch die Beteiligung eines ausgewiesenen Wissenschaftlers in diesem Bereich (Kristian Kersting) als drittem Projektleiter Rechnung getragen. Die Arbeitsgruppe befasst sich mit relationalem Data Mining und konnte es schon erfolgreich auf Verkehrsprobleme anwenden. So zeigt [B4/465] beispielsweise, dass das Schätzen der latenten Wechselwirkungen zwischen z.B. dem PKW-Aufkommen und der Anzahl von Reisenden im öffentlichen Personennahverkehr zu einer verbesserten Vorhersage der Einzelmodalitäten mittels relationalen Gaußschen Prozessen führen kann. In anderen Arbeiten wurde das Abbilden von GPS-Positionsmessungen

auf ein Straßennetz als eine graphkernbasierte Einbettung formuliert und es konnte empirisch gezeigt werden, dass das Rauschen besser reduziert werden kann als in existierenden Ansätzen. Beiden Ansätzen liegen Ideen des statistisch-relationalen Lernens zugrunde [420, 414]. Hier bricht man mit der Tradition klassischer Ansätze des maschinelles Lernens und des Data Mining, die annehmen, dass Datenpunkte (Merkmal-Wert-Vektoren als Zeilen in einer Datenmatrix) statistisch unabhängig sind. In der heutigen vernetzen Welt müssen Sensoren nicht nur als "statistische Datenpunkte" modelliert werden, sondern insbesondere als "soziale Datenpunkte" mit komplexen Wechselwirkungen. Statistische Abhängigkeiten können nicht nur zwischen Merkmalen (den Spalten einer Datenmatrix) des selben Datenpunktes oder zwischen zwei Datenpunkten auftreten, sondern zwischen verschiedenen, miteinander in Beziehung stehenden Datenpunkten (den Zeilen und Spalten mehrerer Datenmatrizen). Das statistisch-relationale Lernen beschäftigt sich mit solchen "sozialen" Datenpunkten. Dieses jungen Forschungsgebiet untersucht Verfahren, die multi-relationale Daten (die gerade in mehreren Datenmatrizen unterschiedlicher Dimensionen vorliegen), logische Repräsentationen und probabilistische Inferenz- und Lernverfahren in Einklang bringen und dadurch den Umgang mit strukturierten Daten aller Art ermöglichen. Einerseits erlauben die logischen Grundlagen die einfache Repräsentation komplexer Objekte (z.B. Straßennetzwerke) mit allgemeinen Regeln (Wissen wird durch die Abstraktion und Unifikation für ganze Mengen an Objekten wie z.B. Straßenkreuzungen formuliert), andererseits wird durch die Hinzunahme statistischer Lernverfahren der Lernprozess auf eine solide theoretische Grundlage gestellt, die es erlaubt mit Unsicherheiten umzugehen. Insbesondere wird durch die Abstraktion von individuellen Objekten das Lernen oftmals beschleunigt, weil das, was man an einer Straßenkreuzung beobachtet auch für andere Straßenkreuzungen gilt.

Prominente Beispiele sind sogenannte Markov Logic Networks [446] und Relational Dependency Networks [440]. Markov Logic Networks stellen relational Verteilungen durch gewichtete logische Regeln wie z.B.

$$2{,}5\ \texttt{verbunden}(\texttt{Seg}_1, \texttt{Seg}_2) \rightarrow \big(\texttt{stau}(\texttt{Seg}_1) \leftrightarrow \texttt{stau}(\texttt{Seg}_2)\big)$$

dar. Diese Regel besagt zum Beispiel, dass auf zwei benachbarten Straßensegementen ($\texttt{Seg}_1$ und $\texttt{Seg}_2$ sind beides Platzhalter für konkrete Straßensegemente) entweder Stau oder kein Stau ist. Natürlich stimmt diese Regel nicht immer und daher hat sie ein Gewicht von 2,5 und nicht von ∞. Relational Dependency Networks stellen solche Verhältnisse über sich gegenseitig beeinflussenden Zufallsvariablen dar, die mittels logischer Prädikate parametrisiert sind. In den letzten Jahren wurde in einer Vielzahl an Publikationen aufgezeigt, dass solche und ähnliche Ansätze klassische Ansätze signifikant in der Performanz übertreffen können: Entitätenresolution, Informationsverbreitungen, Marktanalysen, Objekttracking, Event-Prädiktion uva. [448, 410, 438, 429]. Um solche hochkompakten statistischen Modelle zu lernen, haben sich Boosting-Verfahren als bisher bester Ansatz durchgesetzt [B4/464, B4/463]. Die Idee ist, dass man additive Modelle lernt, die iterativ aus "schwachen" relationalen Modelle zusammengestellt werden, die sich auf das fokussieren, was das bisherige (additive) Modell nicht gut modelliert.

Insgesamt ist das Gebiet zu groß, um eine vollständige Beschreibung des aktuellen Forschungsstandes zu geben, wichtig ist nur die Feststellung, dass sich das Forschungsgebiet hauptsächlich mit multinomialen Verteilungen befasst hat. Erst in den letzten Jahren wurden die ersten Ansätze beschrieben, bei denen Gaußsche Verteilungen für relationale Strukturen modelliert werden, siehe z.B. [411]. Zwar wurden relationale Modelle zur Verkehrsmodellierung herangezogen [434], allerdings sind Verkehrdaten oftmals weder multinomial noch gaußverteilt. Sie sind Zählverteilungen und so z.B. poissonverteilt. Folglich ist eine weitere Verbesserung — neben der Ausnutzung von Querbezügen zwischen Zufallsvariablen — durch die relationale Modellierung von Zählverteilungen zu erwarten. Zählverteilungen haben eine lange Tradition in der Statistik. Vielleicht am bekanntestes sind Generalisierte Linear Modelle (GLMs) [439, 437]. Sie verallgemeinern die klassische lineare Regression, in der die Zielvariable eine Verteilung aus der exponentiellen Familie besitzt. Diese Verteilungsklasse beinhaltet neben der Normalverteilung auch die Binomial-, Poisson-, Gamma- und inverse Gaußverteilung. GLMs wurden mit Erfolg in der Ökonomie, der Biologie, der Medizin uvm. angewandt. Es ist daher erstaunlich, dass es noch keine relationalen Varianten gibt. Zwar gibt es Ansätze zu Netzwerken von Zählverteilungen [424, 423, 445, 456], die zu verbesserten Analysen und Entdeckung von sozialen sowie biologischen Netzwerken geführt haben, allerdings werden keine allgemeingültige logischen Regeln verwendet. Gerade die Abstraktion und Unifikation mittels Logik sind es allerdings, die statistische-relationale Ansätze so erfolgreich machen. Einzig Xian und Neville [454] haben sogenannte relationale Gaußsche Copula-Verteilungen vorgeschlagen und zur Analyse von sozialen Netzwerken benutzt. Allerdings ist bisher kein Ansatz zur automatischen Modellselektion bekannt und sie wurden nicht zur Modellierung von Zählverteilungen herangezogen. Zusammenfassend lässt sich sagen, dass statistisch-relationale Modelle sich zwar großer Beliebtheit erfreuen, sie aber bisher nicht für Zählverteilungen untersucht wurden. Das Schließen dieser Lücke ist eines der Ziele des Projekts. Das wiederum würde den Weg zu neuen, relationalen Statistiken führen. So können multi-relationale Kontingenztabellen analysiert werden, also multi-relationale,

statistische Tests entwickelt werden, die Querbezüge zwischen mehreren Kontingenztabellen berücksichtigen. Das wäre ein wesentlicher Schritt auf dem Weg zu einer Antwort auf eine der grundlegendsten offenen Fragen im relationalen Lernen: Wie evaluiert man statistisch-valide multi-relationale Modelle, die Vorhersagen für verschiedene Variablen treffen?

In einem ersten Vorversuch haben alle drei Projektpartner zusammen bereits neue Modelle zur Verkehrsflussvorhersage entwickelt und untersucht. Dabei sind wir dem weiter unten beschriebenen Dependency Network Ansatz gefolgt, allerdings nur in einer propositionalen und nicht in einer relationalen Variante; es werden also keine abstrakten Regel gelernt, die über Straßensegmente hinweg generalisieren. Dazu wurden für ca. 45 Messquerschnitte auf dem Kölner Ring Messdaten in einer Zeit-Ort-Matrix mit minütlichen Werten einer Woche zusammengestellt. Es wurde die Summe des Verkehrsflusses in Fahrzeuge pro Minute betrachtet. Über diesen Daten wurde ein Poisson Dependency Network mittels Boosting gelernt und über eine "leave-one-out" Kreuzvalidierung evaluiert. Die in Abb. 3.1c gezeigte Scatterplots der "leave-one-out" Kreuzvalidierung zeigt eine hohe Korrelation der vorhergesagten mit den beobachteten Werten für das aus den Beobachtungen gelernte Model. Es erscheint kompetitiv zur bisher üblichen physikalischen Simulation, insbesondere für kleinere Verkehrsflüsse. Allerdings scheint die Simulation etwas besser größere Verkehrsflüsse zu modellieren. Insgesamt ist das Ergebnis als vielversprechend für die vollausgebauten Modelle und Inferenz- bzw. Lernverfahren der zweiten Förderperiode zu bewerten.

Literatur

[406] M. Alsabaan, W. Alasmary, A. Albasir und K. Naik. „Vehicular Networks for a Greener Environment: A Survey". In: *Commun. Surveys Tut.* 15.3 (2013), S. 1372–1388.

[407] G. Araniti, C. Campolo, M. Condoluci, A. Iera und A. Molinaro. „LTE for Vehicular Networking: A Survey". In: *IEEE Commun. Mag.* 51.5 (2013), S. 148–157.

[408] F. Calabrese, P. Michaelsen, C. Rosa, M. Anas, C. Castellanos, D. Villa, K. Pedersen und P. Mogensen. „Search-Tree Based Uplink Channel Aware Packet Scheduling for UTRAN LTE". In: *Proc. of IEEE 67th Vehicular Technology Conference (VTC-Spring).* IEEE, 2008, S. 1949–1953.

[409] L. Cao. *Data Mining and Multi-agent Integration.* Springer, 2009.

[410] J. Choi, A. Guzmán-Rivera und E. Amir. „Lifted Relational Kalman Filtering". In: *Proc. of 22nd International Joint Conference on Artificial Intelligence (IJCAI).* AAAI Press, 2011, S. 2092–2099.

[411] J. Choi, D. Hill und E. Amir. „Lifted Inference for Relational Continuous Models". In: *Proc. of 26th Conference on Uncertainty in Artificial Intelligence (UAI).* AUAI Press, 2010, S. 126–134.

[412] A. H. Chow, A. Santacreu, I. Tsapakis, G. Tanasaranond und T. Cheng. „Empirical assessment of urban traffic congestion". In: *J. Adv. Transp.* (2013).

[413] L. C. Davis. „Realizing Wardrop equilibria with real-time traffic information". In: *Physica A* 388.20 (2009), S. 4459–4474.

[414] L. De Raedt. *Logical and Relational Learning.* Springer, 2008.

[415] F. Dion, H. Rakha und Y.-S. Kang. „Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections". In: *Transport. Res. B* 38.2 (2004), S. 99–122.

[416] D. Easley und J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.* Cambridge University Press, 2010.

[417] T. Fawcett. „Data Mining with Cellular Automata". In: *SIGKDD Explor. Newsl.* 10.1 (2008), S. 32–39.

[418] R. Filjar, K. Vidovic, P. Britvic und M. Rimac. „eCall: Automatic Notification of a Road Traffic Accident". In: *Proc. of 34th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).* IEEE, 2011, S. 600–605.

[419] N. H. Gartner und C. Stamatiadis. „Traffic Networks, Optimization and Control of Urban". In: *Encyclopedia of Complexity and Systems Science.* Hrsg. von R. A. Meyers. Springer, 2009, S. 9470–9500.

[420] L. Getoor und B. Taskar, Hrsg. *Introduction to Statistical Relational Learning.* The MIT Press, 2007.

[421] P. D. Groves. „Shadow matching: A new GNSS positioning technique for urban canyons". In: *J. Navigation* 64.3 (2011), S. 417–430.

[422] P. D. Groves, Z. Jiang, L. Wang und M. Ziebart. „Intelligent Urban Positioning using Multi-Constellation GNSS with 3D Mapping and NLOS Signal Detection“. In: *Proc. of 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS)*. ION, 2012, S. 458–472.

[423] P. D. Hoff. „Random effects models for network data“. In: *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. Hrsg. von R. Breiger, K. Carley und P. Pattison. The National Academies Press, 2003, S. 303–312.

[424] P. D. Hoff, A. E. Raftery und M. S. Handcock. „Latent Space Approaches to Social Network Analysis“. In: *J. Amer. Statist. Assoc.* 97.460 (2002), S. 1090–1098.

[425] H. Je, H. Lee, K. Kwak, K. Choi, Y.-J. Hong und B. Clerckx. „Long-Term Channel Information-Based CoMP Beamforming in LTE-Advanced Systems“. In: *Proc. of IEEE Global Telecommunications Conference (GLOBECOM)*. IEEE, 2011, S. 1–6.

[426] K. Jo, K. Chu und M. Sunwoo. „Interacting Multiple Model Filter-Based Sensor Fusion of GPS With In-Vehicle Sensors for Real-Time Vehicle Positioning“. In: *IEEE Trans. Intell. Transp. Syst.* 13.1 (2012), S. 329–343.

[427] B. S. Kerner. *Introduction to Modern Traffic Flow Theory and Control: The Long Road to Three-Phase Traffic Theory*. Springer, 2009.

[428] B. S. Kerner. „Optimum Principle for a Vehicular Traffic Network: Minimum Probability of Congestion“. In: *J. Phys. A: Math. Theor.* 44, 092001 (2011).

[429] C. Kiddon und P. Domingos. „Coarse-to-Fine Inference and Learning for First-Order Probabilistic Models“. In: *Proc. of 25th AAAI Conference on Artificial Intelligence*. AAAI Press, 2011, S. 1049–1056.

[430] M. J. Koetse und P. Rietveld. „The Impact of Climate Change and Weather on Transport: An Overview of Empirical Findings“. In: *Transport. Res. D* 14 (2009), S. 205–221.

[431] E. Köhler, R. H. Möhring und M. Skutella. „Traffic Networks and Flows over Time“. In: *Algorithmics of Large and Complex Networks*. Hrsg. von J. Lerner, D. Wagner und K. A. Zweig. Bd. 5515. LNCS. Springer, S. 166–196.

[432] D. Krajzewicz. „Traffic Simulation with SUMO – Simulation of Urban Mobility“. In: *Fundamentals of Traffic Simulation*. Hrsg. von J. Barceló. Bd. 145. International Series in Operations Research & Management Science. Springer, 2010, S. 269–293.

[433] L. Li, D. Wen und D. Yao. „A Survey of Traffic Control With Vehicular Communications“. In: *IEEE Trans. Intell. Transp. Syst.* 15.1 (2014), S. 425–432.

[434] M. Lippi, M. Bertini und P. Frasconi. „Collective Traffic Forecasting“. In: *Machine Learning and Knowledge Discovery in Databases*. Hrsg. von J. L. Balcázar, F. Bonchi, A. Gionis und M. Sebag. Bd. 6322. LNCS. Springer, 2010, S. 259–273.

[435] M. Liso Nicolas, M. Jacob, M. Smyrnaios, S. Schon und T. Kurner. „Basic Concepts for the Modeling and Correction of GNSS Multipath Effects using Ray Tracing and Software Receivers“. In: *Proc. of IEEE-APS Topical Conference on Antennas and Propagation in Wireless Communications (APWC)*. IEEE, 2011, S. 890–893.

[436] E. D. Martí, D. Martín, J. García, A. de la Escalera, J. M. Molina und J. M. Armingol. „Context-Aided Sensor Fusion for Enhanced Urban Navigation“. In: *Sensors* 12.12 (2012), S. 16802–16837.

[437] C. E. McCulloch. „Maximum Likelihood Algorithms for Generalized Linear Mixed Models“. In: *J. Amer. Statist. Assoc.* 92.437 (1997), S. 162–170.

[438] A. Nath und P. Domingos. „Efficient Lifting for Online Probabilistic Inference“. In: *Proc. of 24th AAAI Conference on Artificial Intelligence*. AAAI Press, 2010, S. 1193–1198.

[439] J. Nelder und R. Wedderburn. „Generalized Linear Models“. In: *J. R. Stat. Soc. Ser. A Gen.* 135.3 (1972), S. 370–384.

[440] J. Neville und D. Jensen. „Relational Dependency Networks“. In: *J. Mach. Learn. Res.* 8 (2007), S. 653–692.

[441] A. J. Nicholson und B. D. Noble. „BreadCrumbs: Forecasting Mobile Connectivity“. In: *Proc. of 14th ACM International Conference on Mobile Computing and Networking (MobiCom)*. ACM, 2008, S. 46–57.

[442] V. Pla, J. Martinez-Bauset und V. Casares-giner. „Comments on "Call Blocking Probability and Bandwidth Utilization of OFDM Subcarrier Allocation in Next-generation Wireless Networks"“. In: *IEEE Commun. Lett.* 12.5 (2008), S. 349–349.

[443] R. Popescu-Zeletin, I. Radusch und M. A. Rigani. *Vehicular-2-X Communication*. Springer, 2010.

[444] N. Prasad, H. Zhang, H. Zhu und S. Rangarajan. „Multiuser Scheduling in the 3GPP LTE Cellular Uplink“. In: *IEEE Trans. Mobile Comput.* 13.1 (2014), S. 130–145.

[445] S. Rajaram, T. Graepel und R. Herbrich. „Poisson-Networks: A Model for Structured Point Processes“. In: *Proc. of 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*. Society for Artificial Intelligence und Statistics, 2005, S. 277–284.

[446] M. Richardson und P. Domingos. „Markov Logic Networks“. In: *Mach. Learn.* 62.1-2 (2006), S. 107–136.

[447] A. Schadschneider, D. Chowdhury und K. Nishinari. *Stochastic Transport in Complex Systems: From Molecules to Vehicles*. Elsevier, 2010.

[448] P. Sen, A. Deshpande und L. Getoor. „Exploiting shared correlations in probabilistic databases“. In: *PVLDB* 1.1 (2008), S. 809–820.

[449] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger und F. Xu. „Overview of 3GPP LTE-advanced Carrier Aggregation for 4G wireless Communications“. In: *IEEE Commun. Mag.* 50.2 (2012), S. 122–130.

[450] I. Tsapakis, T. Cheng und A. Bolbol. „Impact of Weather Conditions on Macroscopic Urban Travel Times“. In: *J. Transp. Geogr.* 28 (2013), S. 204–211.

[451] S.-H. Wang, A.-D. Lin und Y.-W. Hong. „Channel-Aware Transmission Control for Cooperative Random Access Networks“. In: *Proc. of IEEE 71th Vehicular Technology Conference (VTC-Spring)*. IEEE, 2010, S. 1–5.

[452] J. G. Wardrop und J. I. Whitehead. „Correspondence. Some theoretical aspects of road traffic research“. In: *ICE Proceedings: Engineering Divisions* 1.5 (1952), S. 767–768.

[453] M. G. Wing, A. Eklund und L. D. Kellogg. „Consumer-grade global positioning system (GPS) accuracy and reliability“. In: *J. Forest.* 103.4 (2005), S. 169–173.

[454] R. Xiang und J. Neville. „Collective Inference for Network Data with Copula Latent Markov Networks“. In: *Proc. of 6th ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2013, S. 647–656.

[455] K. Yagyu, T. Nakamori, H. Ishii, M. Iwamura, N. Miki, T. Asai und J. Hagiwara. „Investigation on Mobility Management for Carrier Aggregation in LTE-Advanced“. In: *Proc. of IEEE 74th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2011, S. 1–5.

[456] E. Yang, P. Ravikumar, G. Allen und Z. Liu. „On Poisson Graphical Models“. In: *Advances in Neural Information Processing Systems*. Hrsg. von C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani und K. Q. Weinberger. Bd. 26. Curran Associates, 2013, S. 1718–1726.

[457] R. Yozevitch, B Ben Moshe und H. Levy. „Breaking the 1 meter accuracy bound in commercial GNSS devices“. In: *Proc. of IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*. IEEE, 2012, S. 1–5.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[B4/458] L. Habel und M. Schreckenberg. „Asymmetric Lane Change Rules for a Microscopic Highway Traffic Model“. In: *accepted for presentation at 11th International Conference on Cellular Automata for Research and Industry (ACRI)*. Krakow, Poland, 2014.

[B4/A4/459] C. Ide, B. Dusza, M. Putzke und C. Wietfeld. „Channel Sensitive Transmission Scheme for V2I-based Floating Car Data Collection via LTE“. In: *Proc. of IEEE International Conference on Communications (ICC)*. Ottawa, Canada: IEEE, 2012, S. 7151–7156.

[B4/460] C. Ide, T. Knaup, B. Niehoefer, D. Weber, L. Habel, M. Schreckenberg und C. Wietfeld. „Efficient Floating Car Data Transmission via LTE for Travel Time Estimation of Vehicles“. In: *Proc. of IEEE 76th Vehicular Technology Conference (VTC-Fall)*. Québec City, Canada: IEEE, 2012, S. 1–5.

[B4/461] C. Ide, L. Habel, T. Knaup, M. Schreckenberg und C. Wietfeld. „Interaction between Machine-Type Communication and H2H LTE Traffic in Vehicular Environments“. In: *Proc. of IEEE 79th Vehicular Technology Conference (VTC-Spring)*. Seoul, Korea, 2014.

[B4/462] B. S. Kerner, S. L. Klenov, G. Hermanns, P. Hemmerle, H. Rehborn und M. Schreckenberg. „Synchronized flow in oversaturated city traffic“. In: *Phys. Rev. E* 88 (5 2013), S. 054801.

[B4/463] T. Khot, S. Natarajan, K. Kersting und J. Shavlik. „Learning Markov Logic Networks via Functional Gradient Boosting“. In: *Proc. of IEEE 11th International Conference on Data Mining (ICDM)*. Vancouver, Canada: IEEE, 2011, S. 320–329.

[B3/B4/402] T. Maschek, B. Konrad, J. Deuse, G. Hermanns, D. Weber und M. Schreckenberg. „Verkehrsforschung in der Produktionsflussanalyse - Übertragung von Modellen der statistischen Physik auf die Analyse von Produktionssystemen“. In: *ZWF - Zeitung für wirtschaftlichen Fabrikbetrieb* 106.11 (2011), S. 833–837.

[B4/464] S. Natarajan, T. Khot, K. Kersting, B. Gutmann und J. Shavlik. „Gradient–based Boosting for Statistical Relational Learning: The Relational Dependency Network Case“. In: *Mach. Learn.* 86.1 (2012), S. 25–56.

[B4/465] M. Neumann, K. Kersting, Z. Xu und D. Schulz. „Stacked Gaussian Process Learning“. In: *Proc. of IEEE 9th International Conference on Data Mining (ICDM)*. Miami, USA: IEEE, 2009, S. 387–396.

[B4/466] B. Niehoefer, S. Lehnhausen und C. Wietfeld. „Optimizing-Strategies for a Time-Efficient Evaluation of Position-Specific Communication Aspects in Disaster Relief Scenarios“. In: *Proc. of IEEE 1st AESS European Conference on European Satellite Telecommunications (ESTEL)*. Rome, Italy: IEEE, 2012, S. 1–6.

[B4/467] B. Niehoefer, F. Schweikowski, S. Lehnhausen und C. Wietfeld. „Cloud-aided SDR Solution for Lane-Specific Vehicle Positioning via Local Interference Compensation“. In: *Proc. of IEEE Aerospace Conference (AEROSPACE)*. Big Sky, USA, 2014.

[B4/A4/468] C. Wietfeld, C. Ide und B. Dusza. „Resource Efficient Mobile Communications for Crowd-Sensing“. In: *Proc. of 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*. San Francisco, USA, 2014.

3.4 Planung des Teilprojekts

Ziele

Das Teilprojekt B4 verfolgt das übergreifende Ziel, mittels ressourceneffizienter Informationserhebung und Datenanalyse neuartige Prognose- und Beeinflussungsverfahren für dynamische Verkehrssysteme zu entwickeln, welche in Echtzeit gegenüber dem heutigen Stand der Forschung deutlich verbesserte Ergebnisse erzielen.

In der ersten Förderperiode standen hierbei die spurgenaue Lokalisierung von Fahrzeugen, eine ressourceneffiziente Übertragung von *extended Floating Car Data* (xFCD) via LTE sowie die Nutzung dieser Daten für die Verbesserung der Verkehrsprognose auf Autobahnen im Vordergrund.

In der zweiten Förderperiode wird das Einsatzfeld der entwickelten Verfahren auf innerstädtische Straßeninfrastrukturen ausgedehnt sowie um Aspekte der Verkehrssteuerung erweitert. Diese erweiterte Ausrichtung bringt vor allem durch die Heterogenität und Komplexität des Stadtverkehrs z.B. durch Kreuzungen, Parkverhalten und Interaktionen unterschiedlicher Verkehrsmittel neue Herausforderungen, welche innovative Lösungsansätze in vielen unterschiedlichen Bereichen benötigen, mit sich. Daraus resultieren hohe Anforderungen an die Datenanalyse, die in der zweiten Förderperiode durch die Hinzunahme von Kristian Kersting vertieft durchgeführt werden kann. Hierzu werden innerstädtische Verkehrsdaten aus Dortmund, Düsseldorf und Dublin zur Verfügung stehen. Innerstädtische Verkehrsmessungen werden weniger als eine Menge statistischer Datenpunkte modelliert, sondern eher als Datenpunkte mit komplexen vielfältigen Beziehungen, Attributen und unterschiedlichsten Vernetzungen wie in sozialen Netzen betrachtet werden können. Dazu werden vor allem unterschiedliche Algorithmen zur relationalen Datenanalyse und der Modelschätzung aus Zählungen von Fahrzeugen in gewissen Zeitintervallen untersucht und entwickelt. Relationale Ansätze können den komplexen Interaktionen zwischen Messungen Rechnung tragen und diese kompakt darstellen. Durch das Anwendungsszenario der Vorhersage auch im Fahrzeug selbst wird sichergestellt, dass die entwickelten Algorithmen auch unter Ressourcenbeschränkungen eingesetzt werden können.

Auf Basis der Datenanalyse werden mikroskopische Verkehrsmodelle entwickelt, die in Echtzeit analysebasiert parametrisierbar sind, um die komplexen Phänomene im Innenstadtbereich durch Simulationen besser abbilden zu können. Um diese Abbildung weiter zu präzisieren, werden Verfahren entwickelt, mit denen Deckungslücken der Datenversorgung geschlossen werden können. Mit diesen Modellen und Verfahren sowie einer Koppelung der Simulation mit der Datenanalyse können dann komplexe Fragestellungen der Verkehrssteuerung durch adaptierbare, ressourceneffiziente Infrastruktur untersucht werden.

Darüber hinaus wird die Verkehrssteuerung auf der Ebene einzelner Fahrzeuge zur Vermeidung von Zusammenbrüchen in Verkehrsnetzwerken und die Optimierung von Verkehrsflüssen in Bezug auf Fahrzeiten und Kraftstoffverbrauch untersucht. Hierzu werden ressourceneffiziente, bidirektionale Echtzeitkommunikationsverfahren für die Übertragung von Sensordaten und Steuerbefehlen entwickelt. Dabei sollen insbesondere die neuen technischen Möglichkeiten von LTE-A berücksichtigt werden. Darüber hinaus ist ein Ziel des Teilprojekts, Verfahren zur Vorhersage der Kommunikationsgüte zu entwickeln, welche eine besonders effiziente und vorausschauende Kommunikation sicherstellen. Die Verfahren zur spurgenauen Lokalisierung werden in der zweiten Förderperiode auf Innenstadtszenarien erweitert und es ist eine umfangreiche Validierung im Feld vorgesehen.

Arbeitsprogramm

Das Arbeitsprogramm ist wie in Abbildung 3.3 gegliedert und orientiert sich an den zuvor benannten Zielen.

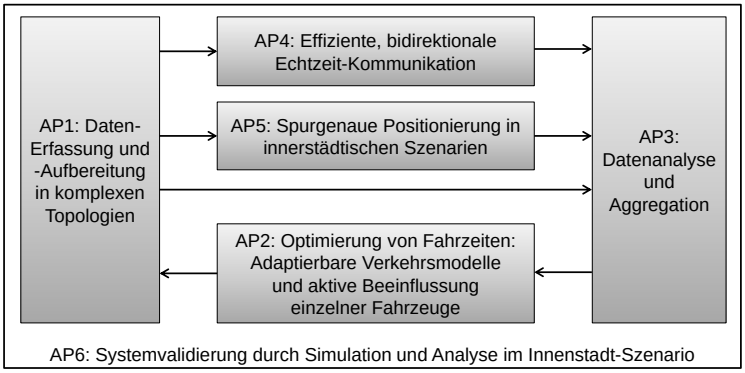


Abbildung 3.3.: Arbeitspaketstruktur

Arbeitspaket 1. Daten-Erfassung und -Aufbereitung in komplexen Topologien: Für die beantragte Förderperiode werden Datenquellen benötigt, die Informationen aus innerstädtischen Umgebungen bereitstellen. Dazu können die folgenden Daten erschlossen werden:

- Stationär gemessene Verkehrsdaten der Stadt Düsseldorf: Ca. 150 Detektoren und weitere Ereignismesser an Lichtsignalanlagen auf innerstädtischen Straßen, sowohl historische Daten als auch Live-Daten aus dem MobilitätsDatenMarktplatz (MDM)
- Live-Daten der Parkplatzauslastung der Stadt Düsseldorf aus dem MDM
- Floating Car Daten der Stadt Dortmund: Elektroautoflotte mit 10 Fahrzeugen, ca. 3 Millionen Datensätze (Datum, Uhrzeit, GPS-Position, Geschwindigkeit, UMTS-Kanalgüte und Smartphone-Beschleunigungssensor) zum Großteil aus dem Stadtgebiet Dortmund, von Dezember 2012 bis Juni 2014
- Daten aus Dublin: Insight Projekt von Prof. Morik

Ergänzend stehen wie in der ersten Förderperiode Verkehrsdaten aus dem Autobahnnetz inklusive autobahnähnlicher Straßen (z.B. Dortmund, B1) zur Vervollständigung der innerstädtischen Datenbasis (aus dem Datenverteiler NRW) zur Verfügung.

Im Rahmen dieses Arbeitspakets werden vorbereitende Arbeiten durchgeführt, um die verschiedenen Datenquellen für die folgenden Arbeitspakete verfügbar zu machen. Hierzu wird zunächst eine gemeinsame Datenbankstruktur spezifiziert, in der die Daten aus den genannten Quellen nach Aufbereitung und Prüfung auf Plausibilität zusammengeführt werden.

Arbeitspaket 2. Optimierung von Fahrzeiten: Adaptierbare Verkehrsmodelle und aktive Beeinflussung einzelner Fahrzeuge: Innerstädtische Straßennetze stellen eine große Herausforderung für die Verkehrslenkung dar, da aufgrund der baulichen Gegebenheiten ein weiterer Ausbau des Netzes nicht möglich ist. Deshalb müssen intelligente, dynamische Verfahren zur effizienten Auslastung der bestehenden Ressourcen gefunden werden. Zwei vielversprechende Möglichkeiten sind zum einen eine dynamische Modifikation des vorhandenen Straßennetzes als auch ein situationsangepasstes Routing der Verkehrsteilnehmer.

- **Entwicklung analysebasierter dynamischer Verkehrsmodelle:** Die bisher verwendeten mikroskopischen Verkehrsmodelle auf Basis von Zellularautomaten [427, 447, B4/458] unterliegen statischen Regelsätzen. Erfahrungen bei der Simulation komplexer innerstädtischer Szenarien haben gezeigt, dass dieser Ansatz die Verkehrsabläufe nicht optimal beschreibt, weil das Verhalten einzelner Verkehrsteilnehmer stark von den lokalen Gegebenheiten abhängt. Diese sind im Stadtbereich wesentlich heterogener und komplexer als auf den überörtlichen Straßen. Als Beispiele seien komplexe Vorfahrtregeln, Kreisverkehre sowie die Interaktion der Verkehrsteilnehmer mit Langsamverkehr oder ÖPNV genannt.

  In diesem Arbeitspaket sollen zur Abbildung der Heterogenität des Stadtverkehrs lokale Echtzeitinformationen in die Verkehrsmodelle einfließen. Hierzu werden zunächst die Verkehrsmodelle so erweitert, dass sie besser parametrisierbar sind. In Zusammenarbeit mit AP3 werden dann Verfahren entwickelt, die eine Umparametrisierung der Modellparameter in Echtzeit anhand der Erkenntnisse aus AP3 erlauben. Als Ergebnis sollen einerseits Modellierungsfehler durch verbesserte Datenlage minimiert werden und andererseits die Robustheit der Modellparameter durch Methoden des maschinellen Lernens erhöht werden. Die erweiterten Verkehrsmodelle werden in AP6 in einer komplexen Simulation evaluiert.

- **Beeinflussung einzelner Fahrzeuge:** Das von B. Kerner entwickelte *breakdown minimisation principle* [428] dient zur Ermittlung der optimalen Verteilung der Verkehrsflüsse in komplexen Straßennetzen. Zur Umsetzung dieser Verteilung muss die Information an die entsprechenden Verkehrsteilnehmer übermittelt werden. Im Rahmen dieses Arbeitspakets soll für verschiedene Verfahren untersucht werden, wie dies effizient geschehen kann. Für die Beurteilung der Effizienz sind folgende Punkte zu beachten: Welche Durchdringungsraten von kommunizierenden Fahrzeugen sind für den Erfolg des Verfahrens notwendig und wie anfällig ist das Verfahren, wenn sich einige Fahrzeuge nicht an die Empfehlungen halten.

  Im Zusammenhang mit der Beeinflussung einzelner Fahrzeuge soll untersucht werden, ob sich die bidirektionale Kommunikation zwischen den Fahrzeugen nutzen lässt, um so einen Mehrwert für den gesamten Verkehrsfluss oder mindestens für die kommunizierenden Fahrzeuge zu realisieren. Neben einer reinen Fahrzeitersparnis kann hier z.B. auch ein reduzierter Kraftstoffverbrauch als möglicher Vorteil gesehen werden. Letzterer lässt sich z.B. durch die Bildung von festen Fahrzeugkolonnen erreichen, die ihre Bewegung durch den Austausch von Nachrichten miteinander koordinieren. Im innerstädtischen Bereich ist die Kommunikation durch die Bebauung oft beeinträchtigt. Dieser Aspekt wird in AP4 vertieft.

Zur Untersuchung dieser neuartigen Verfahren zur Optimierung der Reisezeit werden die in der ersten Förderperiode entwickelten Verkehrsflussmodelle für die Erfordernisse der Simulation des Stadtverkehrs angepasst. Eine besondere Herausforderung ist hierbei die Schaffung von Schnittstellen zur Übernahme der in AP1 erfassten Daten, um sie in der Simulation adäquat zu berücksichtigen. Um die Echtzeitfähigkeit der Simulation zu gewährleisten, muss bei der Implementierung der zu untersuchenden Verfahren besonderes Augenmerk auf performantes Softwaredesign gelegt werden.

Arbeitspaket 3. Datenanalyse und Aggregation: In diesem Arbeitspaket sollen aktuelle Methoden zur Analyse der in AP1 erhobenen Daten zur dynamischen Verkehrsprognose und -beeinflussung untersucht und entwickelt werden. In Zukunft werden Sensoren und Systeme in diesem Gebiet weniger als "statistische Datenpunkte" modelliert werden, sondern als "soziale Datenpunkte" mit komplexen vielfältigen Beziehungen und Attributen und unterschiedlichsten Vernetzungen. Ein Stau an einer Kreuzung wirkt sich direkt auf benachbarte Straßensegemente aus. Daher bieten sich relationale Lernverfahren — die Datenpunkte und ihre Querbezüge betrachten — an. Insbesondere müssen dafür aber relationale Modelle für Zählverteilungen entwickelt und analysiert werden. Diese dienen sowohl der direkten Datenanalyse als auch deren model-basierter Aggregation. Wir gehen zunächst davon aus, dass alle für den Algorithmus relevanten Daten lokal vorhanden sind oder aber — beispielsweise durch Zugriff auf sekundäre oder verbundene Speichermedien — verfügbar gemacht werden können. Ausgangspunkt für unsere Arbeiten ist die folgende Idee, die im Projektverlauf konkretisiert werden soll. Zur Illustration beschränken wir uns hier auf Poissonverteilungen. Allgemein sollen

aber Relationale Generalisierte Lineare Modelle auch anderer Verteilungen entwickelt werden. Wir nehmen also eine Parametrisierung der Verbundwahrscheinlichkeit einer Menge von poisson-verteilten Zufallsvariablen $\mathbf{X} = X_1, X_2, \ldots, X_n$ mittels $P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|\mathbf{X}_{-i})$ mit $\mathbf{X}_{-i} := \mathbf{X} - \{X_i\}$ an. Weiterhin nehmen wir an, dass $P(X_i|\mathbf{X}_{-i})$ einer Poissonverteilung folgt, die mittels eines GLMs dargestellt werden kann. Dieses Modell kann nun in einem ersten Schritt mittels Poisson-Regression pro Variable X_i gelernt werden. Allerdings hat ein solcher Ansatz ein aus Sicht des Teilprojekts und seiner Ziele einen entscheidenen Nachteil. Er betrachtet einen potentiell sehr hochdimensionalen Parameterraum und kann — wenn überhaupt — nur schlecht mit partiell beobachteten Daten oder sogar versteckte Variablen umgehen. Daher sollen Inferenzmethoden (wie z.B. Gibbs-Sampling, Variationsrechnungen u.ä.) auf ihre Übertragung auf Poisson Dependency Networks untersucht und geklärt werden, wie sie in den Lernalgorithmus eingebunden werden können. Hierfür sehen wir sogenannte Boosting-Verfahren als vielversprechend an, die die Verbundwahrscheinlichkeit als eine Summe von Regressionsmodellen darstellen. Diese Regressionsmodelle werden iterativ gelernt. Der erhoffte Vorteil, der gerade bei ressourcenbeschränkten Systemen eine Rolle spielt, ist, dass der Lernalgorithmus den hochdimensionalen Parameterraum nie explizit betrachten muss; Merkmale werden on-the-fly in die Modelle geboostet. Die Skalierbarkeit kann noch weiter gesteigert werden, indem man z.B. Hoeffding-Bäume lernt. Außerdem ermöglicht dies die Untersuchung und Einbettung von Online-Lernverfahren. Um mit versteckten — also niemals beobachteten — Variablen umgehen zu können, kann die Inferenz in einen stochastischen Expectation-Maximization-Ansatz eingebettet werden, der zwischen Schätzen der versteckten Variablen und dem Lernen der Modelle alterniert. Ein weiterer Vorteil im Hinblick auf die Ziele dieses Projekts ist, dass Boosting im Prinzip das Lernen von relationalen Modellen ermöglicht. Man ersetzt die Variable $\mathbf{X} = X_1, X_2, \ldots, X_n$ durch logisch parametrisierte Variablen, also Prädikate, und ersetzt das Regressionslernen durch ein relationes Regressionsverfahren. Die sich ergebenen relationalen Lernverfahren sollten mit normalen relationalen Poisson-Regressionsbäumen verglichen werden — die parallel im Projekt entwickelt werden sollen — um den Vorteil des Boostings nachzuweisen. Der Übergang zu relationalen Beschreibungen hat aber auch den Vorteil, dass Zeit als ein weiteres Argument der betrachteten Prädikate gesehen werden kann, sodass – im Hinblick auf die Projektziele — auch die Dynamik von Verkehrssituationen untersucht werden kann. Generell erwarten wir als Ergebnisse neue Data-Mining-Modelle für Zählverteilungen, die zu einer Verbesserung der Verkehrsmodelle in AP2 und der Kommunikationsmodelle in AP4 führen, und eine gründliche Analyse hinsichtlich ihrer Ressourcenschonung sowie des Grads der Kopplung der physikalischen Modelle mit denen des Data Minings.

Allerdings stellt sich aus wahrscheinlichkeitstheoretischer Sicht die Herausforderung, dass die bisher vorgeschlagenen Poisson Dependency Networks nicht garantiert zu konvergenten Inferenzverfahren führen. In ressourcenbeschränkten Systemen, wie in diesem Projekt untersucht, kann eine solche Garantie aber notwendig sein. In einem ersten Schritt sollen "beschnittene" Poisson-Verteilungen untersuchen werden, deren Wahrscheinlichkeitsmasse für $X > R$ einfach zu $X = R$ bewegt wird. Das führt dann zu relationalen Netzwerken beschnittener Poission-Verteilungen. Hier wird $P(X_i'|\mathbf{X}_{-i}')$ mittels einer Boltzmann-Verteilung "normalisiert", $P(X_i'|\mathbf{X}_{-i}') = \exp\big(P(X_i|\mathbf{X}_{-i})\big)/\big(\sum_j \exp\big(P(X_j|\mathbf{X}_{-j})\big)\big)$. Das heißt, wir führen im Wesentlichen eine Reparametrisierung durch. Das führt dann zu Poisson Markov Logic Networks, da wir jetzt $P(X_i|\mathbf{X}_{-i})$ als einen relationen Faktor in einem graphischen Modell interpretieren können. Auch hier werden wieder Inferenz- und Lernverfahren entwickelt und untersucht.

Unser Ziel sind aber nicht nur Poisson-Verteilungen. Die Familie der Generalisierten Linearen Modelle soll insgesamt zur Modellierung von relationalen Situationen erweitert werden. Damit würde eine Vielzahl wichtiger Verteilungen wie z.B. Binomial-, Poisson-, Gamma- und inverse Gaußverteilung dem relationalen Lernen zugreifbar gemacht werden. Alle entwickelten Modelle und Algorithmen werden systematisch mit Daten aus inner- und außerstädtischen Infrastrukturen untersucht und getestet. Die Ergebnisse sollen mit der Simulation rückgekoppelt werden und in andere Techniken des Data Minings einfließen wie z.B. der Clusteranalyse, um z.B. prototypische Verkehrssituationen zu entdecken, und zum Lernen von relationalen Varianten von Gravity-Models herangezogen werden, wie sie in der Verkehrsmodellierung bekannt sind. Insbesondere soll die Anwendbarkeit der Techniken zur statistischen Analyse der Vorhersagegüte von vernetzten, eingebetteten Systemen empirisch untersucht werden. Wenn in einem System eine Klassifikation (z.B. zur Kanalgüteklassifikation) durchgeführt wird, die an benachbarte Systeme kommuniziert und dort zur selben Klassifikation herangezogen wird (kollektive Klassifikation genannt) so sollte eine potentielle Korrelation zwischen den Vorhersagen berücksichtigt werden, damit die statistischen Ergebnisse nicht übermäßig optimistisch ausfallen.

Arbeitspaket 4. Effiziente, bidirektionale Echtzeit-Kommunikation: In diesem Arbeitspaket werden die neuen Herausforderungen der Kommunikation adressiert. Die benötigten Kommunikationsressourcen stehen dabei immer in direktem Zusammenhang mit der Güte der optimierten Fahrzeiten, welche in AP2 erforscht

werden, da diese Güte stark von der Aktualität und Menge der bereitgestellten Sensordaten abhängt. Neben der Quantifizierung der Kommunikationsressourcen unter Berücksichtigung der Daten-Erfassung und -Selektion (vgl. AP1), ist eine Reduzierung des Kommunikationsaufwands, vor allem mit Fokus auf Stadtgebiete, Ziel dieses Arbeitspakets. Dies geschieht zum einen durch geeignete dezentrale Voranalyse der relevanten Daten in Zusammenarbeit mit AP3 und zum anderen mit Hilfe von intelligenten Sendeverfahren, die eine möglichst geringe Belastung anderer Kommunikationsteilnehmer ermöglichen. Diese Algorithmen sollen auch Informationen der Verkehrsprognosesysteme berücksichtigen. Beispielsweise könnte eine Datenübertragung vor der Einfahrt in ein Gebiet mit erhöhtem Verkehrsaufkommen (z.B. entstehender Stau oder aktuell stark befahrender Innenstadtring) zur Entlastung der Netze in diesen Gebieten sinnvoll sein. In Abgrenzung zur ersten Förderperiode ergeben sich folgende neue Herausforderungen im Hinblick auf eine ressourcenschonende Kommunikation:

Bidirektionale Kommunikation: Neben der Übertragung der Sensordaten, welche den Uplink der Kommunikationsinfrastruktur belasten, benötigt eine Verkehrsbeeinflussung auch eine Downlink-Kommunikation in Form von Steuerbefehlen. Die in AP2 entwickelten Methoden zur Optimierung von komplexen Netzinfrastrukturen definieren die Anforderungen an die benötigte Kommunikation insbesondere im Hinblick auf die zulässige Verzögerung der Steuerbefehle, sodass z.B. mit Hilfe von Priorisierungsalgorithmen gearbeitet werden muss. Die Modellierung der Kommunikationsnetze unter Berücksichtigung von realitätsnahem Innenstadtverkehr wird mit Hilfe einer gekoppelten Verkehrs- und Kommunikationssimulation (vgl. AP6) erreicht.

Innerstädtische Infrastruktur: Eine komplexe Straßeninfrastruktur hat einen signifikanten Einfluss auf die Belastung der ressourcenbeschränkten Kommunikationsnetze, über welche relevante Sensordaten übermittelt werden sollen. Es wurde in der ersten Förderperiode die benötigte Durchdringungsrate für Autobahn-Szenarien analysiert (siehe [B4/460]). Diese ist signifikant abhängig vom Verkehrsfluss. Deshalb werden in der Innenstadt deutlich höhere Durchdringungsraten benötigt, da in einem Detektionszeitraum deutlich weniger Fahrzeuge die Straße passieren, sodass sich eine neue, besondere Herausforderung an die Kommunikationsnetze ergibt. Die innerstädtische Infrastruktur hat zudem bedingt durch starke Bebauung einen starken Einfluss auf die Funkkanalbedingungen zwischen Endgerät im Auto und Basisstation. Dies hat zur Folge, dass sich die Kanalbedingungen innerhalb kürzester Zeit (z.B. bei der Einfahrt in eine Häuserschlucht) stark verändert können, sodass eine zuverlässige Abschätzung der Konnektivität besonders anspruchsvoll ist.

Zur Adressierung dieser neuen Herausforderungen soll das CAT-Verfahren weiterentwickelt werden. Dies beinhaltet zum einen eine Vorhersage der Konnektivität (predictiveCAT) auf Basis von historischen und aktuellen Informationen über die lokal und zeitlich abhängigen Eigenschaften der Mobilfunknetze. Die Vorhersage wird mit Hilfe von Verfahren des maschinellen Lernens (siehe AP3) durchgeführt werden. Darüber hinaus werden die neuen Features von LTE-Advanced (LTE-A) berücksichtigt werden. Dies beinhaltet beispielsweise die Erweiterung des in der ersten Förderperiode entwickelten *Multi-class Erlang Loss Models* auf größere Zell-Bandbreiten, welche durch *Carrier Aggregation* erreicht werden können. Es sollen effiziente LTE-A basierte Kommunikationsverfahren entwickelt werden, welche z.B. eine Übertragung in sehr robusten, kontextabhängigen Kommunikationssituation sicherstellen. Solche Gegebenheiten kennzeichnen sich z.B. durch

- eine gleichzeitige Verbindung eines Teilnehmers mit mehreren Basisstationen basierend auf *Coordinated Multipoint*,
- eine Übertragung der Daten im robusten 800 MHz-Band im Fall von *Carrier Aggregation* oder
- ein gutes Signal-zu-Rausch-Verhältnis durch *Beamforming*.

Im innerstädtischen Bereich kann außerdem eine heterogene Netzinfrastruktur (z.B. WLAN-Hotspots, LTE-Makro- und Femtozellen) effizienter genutzt werden als auf Autobahnen, sodass multikriterielle Netzauswahlverfahren benötigt werden.

Zur Validierung der in diesem Arbeitspaket entwickelten Verfahren steht dem Sonderforschungsbereich zum einen ein LTE-Basisstationsemulator sowie ein Funkkanalemulator für Messungen im Laborumfeld zur Verfügung. Zum anderen betreibt der Lehrstuhl für Kommunikationsnetze eine eigene LTE-Basisstation mit Core Network Emulator.

Dieser Feldaufbau ermöglicht es mit Zuhilfenahme des am Lehrstuhl entwickelten *Smart Traffic Generator (STG)*, eine LTE-Zelle durch örtlich verteilte Lasten gezielt unter Datenverkehr zu setzen, sodass die Wechselwirkung zwischen unterschiedlichen Datenverkehrsklassen (z.B. xFCD-Übertragung und konventionellem Internet-Verkehr) unter kontrollieren Bedingungen analysiert werden kann.

Die Ergebnisse aus diesem Arbeitspaket in Verbindung mit den Analysen aus AP2 ermöglichen eine Kosten-Nutzen-Abschätzung zwischen der zur Verfügung gestellten Datenmenge an Sensordaten, und somit Kosten bzw. Ressourcen der Mobilfunkkommunikation und der durch die Daten ermöglichten Prognosegenauigkeit.

Arbeitspaket 5. Umfassende spurgenaue Positionierung in innerstädtischen Szenarien mittels heterogenem Systemansatz: Die Zuverlässigkeit der Prädikation von Verkehrsflüssen (siehe AP2) steht im direkten Zusammenhang mit der örtlichen wie auch zeitlichen Genauigkeit der Detektion von verkehrsbeeinflussenden Ereignissen. Daher ist die spurgenaue Lokalisierung ein maßgeblicher Leistungsindikator für die Möglichkeiten der verbesserten Verkehrsprognose.

Die im Zuge der ersten Förderperiode entwickelte Möglichkeit der post-simulativen Positionierungsverbesserung durch eine Kompensation der lokalen Einflüsse auf die Satellitensignale (*LOcal interference CompensATion, LOCATe*), zeigte bereits die deutliche Leistungssteigerung und die Möglichkeit auf, Fahrzeuge auf Überlandstrecken oder Autobahnen mit hohen Zuverlässigkeit und Durchdringung spurgenau zu lokalisieren. Um dieses Niveau auch innerstädtisch nicht nur weiterhin garantieren zu können, sondern auch zu steigern, sollen im Verlauf der zweiten Förderperiode auf die durch LOCATe bereits verbesserten Werte Methoden der Mustererkennung und Sensordatenfusion angewandt werden.

Methoden der Mustererkennung können dabei zur Ermittlung baulicher Einflüsse auf die Satellitenpositionierung genutzt werden. Dabei geht es vor allem um Abschattungs- und Mehrwegeeffekte durch meist künstlich geschaffene Objekte im Verbindungspfad vom Satelliten zum Empfänger. Da diese auf alle zeitlich benachbarten Flussobjekte gleichermaßen Einfluss nehmen, können Methoden aus AP3 dabei speziell auf die hier entstandene Problemstellung angewandt werden und ermöglichen dadurch eine Kompensation bzw. eine weitere Positionierungsverbesserung im Nachgang.

Methoden der Sensordatenfusion ermöglichen neue Verfahren zur Detektion und Vermeidung von Fehlpositionierungen. Moderne Kraftfahrzeuge können dabei als fahrende Sensorplattformen genutzt werden. Durch das Auslesen einzelner Sensoren, wie beispielsweise dem Lenkwinkelsensor, könnten starke und plötzlich auftretende Bewegungsänderungen laut GPS somit validiert werden. Durch die Fusion vieler Sensoren, eventuell von mehreren Fahrzeugen, könnten desweiteren vollständige Bewegungsprofile erstellt werden, die nach Abgleich mit den entsprechenden GPS-Messungen zur Generierung von ortsspezifischen Differenzdaten genutzt werden könnten, um diese zeitlich folgenden Flussobjekten zur Verfügung zu stellen.

Zusammenfassend können aus den so gewonnen Informationen folgenden Dienste direkt abgeleitet werden:

- Serverseitige Positionierungsverbesserungen von Positionsdaten mobiler Knoten
- Differenzdatenerstellung im Embedded Device
- Unmittelbare Positionierungsverbesserung im Embedded Device
- Systemunabhängiger Integritätsdienst für satellitengestützte Positionierungen

Arbeitspaket 6. Systemvalidierung durch Simulation und Analyse im Innenstadt-Szenario: In diesem Arbeitspaket werden Methoden und Ergebnisse der vorangegangenen Arbeitspakete zusammengeführt und in Simulationen evaluiert. Ziel ist außerdem die Abbildung eines innerstädtischen Straßennetzes, das die Analyse komplexer netzweiter Effekte ermöglicht. Dies erlaubt die Entwicklung von Strategien zur Verkehrssteuerung.

Konkret sind hierbei vorgesehen:

- **Erschließung von Deckungslücken in Topologie und Verkehrsdaten** Der Aufbau komplexer Verkehrssimulationen erfordert detaillierte Informationen über die zugrundeliegende Straßentopologie. Häufig sind die notwendigen Daten jedoch nicht vorhanden oder unpräzise. Dies ist insbesondere in städtischen Verkehrsnetzen aufgrund der höheren Komplexität der Fall. Gleiches gilt für die Detektion der Fahrzeuge im Straßennetz. Im Rahmen dieses Arbeitspakets sollen deshalb mit den Methoden aus AP3 Verfahren entwickelt werden, die diese Deckungslücken minimieren können.

  Für die Simulation sind insbesondere Verfahren wesentlich, die das Einfügen der real vorfindbaren Fahrzeuganzahl in die Simulation ermöglichen. Hierzu ist die Entwicklung virtueller Detektoren erforderlich, die Verkehrsdaten auf Basis der Verfahren aus AP3 auch an Orten bereitstellen, an denen keine reale Detektion vorhanden ist. Mit diesen Verfahren können auch Detektionslücken überbrückt werden, die durch temporäre Ausfälle von Detektoren entstanden sind.

  In Autobahnnetzen findet man in der Regel Detektoren an den Auf- und Abfahrten. Derartige Zwangspunkte gibt es in einem innerstädtischen Straßennetz nicht. Der Zufluss der Fahrzeuge zum Netz kann

jederzeit vom Straßenrand oder einem Privatgrundstück erfolgen, gleiches gilt für den Abfluss. In Zusammenarbeit mit AP3 soll deshalb ein Einfügeverfahren entwickelt und evaluiert werden, das über den punktförmigen Ansatz auf Autobahnen hinausgeht und flächenbasiert z.B. mit Informationen eines Straßenzuges oder Ortsteils arbeitet.

- **Analyse adaptierbarer Infrastruktur zur Verkehrssteuerung** Die dynamische Modifikation des Straßennetzes erfordert eine Koordination durch eine zentrale Stelle. Hierbei ist zu untersuchen, wie sich die Verkehrsströme nach einer solchen Modifikation (z.B. Umdrehen der Fahrtrichtung einer Einbahnstraße, Freigabe von Seitenstreifen) neu ordnen. Insbesondere gilt es, kritische Knoten des Straßennetzes zu identifizieren, an denen Eingriffe bevorzugt vorzunehmen sind.

  Hierzu soll eine gekoppelte Verkehrs- und Kommunikationssimulation für komplexe innerstädtische Szenarien erstellt werden, um die Genauigkeit der entwickelten Prognoseverfahren in Abhängigkeit der bereit gestellten Sensorinformationen validieren zu können. Dies beinhaltet z.B. die realitätsnahe Modellierung von Kreuzungen mit Lichtsignalanlagen, welche über Mobilfunk gesteuert werden können und steuerbare Einbahnstraßen.

  Die Modifikation des Straßennetzes ist besonders bei Großereignissen sinnvoll, weil in diesen Fällen das Verkehrsaufkommen zu dem (Messe, Konzert) oder von dem (Großschadenslage) Ereignisort stark erhöht ist. Deshalb soll die Analyse anhand ausgewählter Szenarien in diesem Kontext erfolgen.

Nach Validierung der entwickelten Strategien erfolgt eine Integration der Teilkomponenten in ein Gesamtsystem. Mit diesem Gesamtsystem ist es möglich, konkrete Handlungsempfehlungen für die städtische Verkehrsplanung und -lenkung zu geben.

Zeitplan

	2015				2016				2017				2018			
Arbeitspaket Quartal	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Daten-Erfassung und -Aufbereitung in komplexen Topologien:	■	■	■	■												
2. Optimierung von Fahrzeiten: Adaptierbare Verkehrsmodelle und aktive Beeinflussung einzelner Fahrzeuge:	■	■	■	■	■	■	■	■	■	■	■	■				
3. Datenanalyse und Aggregation:	■	■	■	■	■	■	■	■	■	■	■	■	■	■		
4. Effiziente, bidirektionale Echtzeit-Kommunikation:					■	■	■	■	■	■	■	■	■	■	■	■
5. Umfassende spurgenaue Positionierung in innerstädtischen Szenarien mittels heterogenem Systemansatz:							■	■	■	■	■	■	■	■	■	■
6. Systemvalidierung durch Simulation und Analyse im Innenstadt-Szenario:											■	■	■	■	■	■

3.5 Stellung innerhalb des Sonderforschungsbereichs

Das Teilprojekt B4 bietet viele Anknüpfungspunkte zu anderen Teilprojekten im Bereich der Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung.

Eine enge Kooperation ist mit Teilprojekt A1 geplant. Gemeinsame Schnittpunkte sind die ressourceneffiziente Sensordatenerhebung, welche sowohl im AP4/B4 als auch im AP1/A1 im Vordergrund steht, Genauigkeitsanforderungen für Sensordaten im AP5/B4 und AP1/A1 als auch maschinelles Lernen auf Mobil-

funkdaten (AP3/B4 und AP4/A1). Insbesondere werden in A1 auch graphische Wahrscheinlichkeitsmodelle entwickelt. Hier bietet sich ein Austausch von Methoden an.

Darüber hinaus werden Ressourcen- und Kostenmodelle für das Kommunikationsmedium benötigt, um die Datenaggregation für eine optimale Ausnutzung der beschränkten Energieressourcen adaptiv auszuführen. Hierfür definiert A4 geeignete Modelle, welche z.B. LTE-A berücksichtigen. Diese Zusammenarbeit ermöglicht beispielsweise die Bewertung der Energieeffizienz der in B4 entwickelten effizienten Kommunikationsverfahren. Durch die starken Synergien zwischen den Teilprojekten A4 und B4 ergaben sich in der ersten Förderperiode schon ausgeprägte Zusammenarbeiten, welche in zahlreichen gemeinsamen Veröffentlichungen dokumentiert wurden.

Netzwerke und verlinkte Daten treten in vielen Teilprojekten des Sonderforschungsbereichs auf und die Entwicklung von Analyseverfahren ist eine grundlegende Aufgabe. So werden in Teilprojekt A6, in dem Kristian Kersting auch ein Teilprojektleiter ist, ressourceneffiziente Graphkerne entwickelt, die zur Analyse von Verkehrssituationen herangeführt werden sollen. So kann man z.B. (un)typische Verkehrssituation mittels sogenannter Ein-Klassen-SVMs — die auf Kernfunktionen beruhen — bestimmen. Gleichzeitig können mittels Kernmethoden vorhergesagte Klassen als weitere Merkmale Eingang in die Wahrscheinlichkeitsmodelle von B4 finden. Ausserdem können die in B4 entwickelten Wahrscheinlichkeitsmodelle z.B. mittels Skalarprodukten zwischen den Verteilungen in Zusammenarbeit mit A6 in Graphkerne überführt werden, sodass Verkehrssituationen automatisch in verschiedene Kategorien eingestuft werden können. Vielversprechend ist auch die Zerlegung großer Straßennetzwerke in kleinere Netzwerke, sodass Vorhersagemodelle effiziente gelernt werden können. Da B4 verschiedene wissenschaftliche Disziplinen vereint, ist auch die Visualisierung von vernetzen Daten in A6 von großer Bedeutung. Man muss über Modelle und ihr Verhalten fächerübergreifend reden können.

Das Teilprojekt B3 strebt mit der Entwicklung von Verfahren zur realzeitlichen Qualitätsprognose und Prozesssteuerung ähnliche übergeordnete Ziele an, wie sie im Bereich der Verkehrsprognose und -lenkung in diesem Teilprojekt vorliegen. Hier bietet sich an, den in der ersten Förderperiode begonnenen Austausch über Methoden [B3/B4/402] fortzusetzen.

Ein weiterer wichtiger Schnittpunkt ergibt sich mit den Teilprojekten A2 und C4 bezüglich der Verarbeitung von Datenströmen. Die in Teilprojekt A2 entwickelten Datenstromalgorithmen zum Clustering sollen auf die in B4 erhobenen Sensordaten angewandt werden, um beispielsweise den benötigten Speicherbedarf in B4 zu reduzieren. Insbesondere das Teilprojekt C4 beschäftigt sich mit Datenstromalgorithmen für verallgemeinerte lineare Regression. Daher wird mit C4 eine Zusammenarbeit auf methodischer Ebene angestrebt, wobei sich B4 im Gegensatz zu C4 auf relationale Domänen und Modelle fokussieren wird. Ferner können eventuell Ergebnisse aus A2 zum Lernen von Entscheidungsbäumen mittels Kernmengen auf das Schätzen von additiven Modellen zur relationalen, verallgemeinerten linearen Regression in B4 erweitert werden.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

BAB-NRW

(Schreckenberg,Förderkennzeichen Straßen.NRW 02100.150/4.30.15.01/OLSIM) (Förderzeitraum: 2001 - laufend)

Die Erfahrungen aus der Entwicklung und dem Betrieb des Verkehrsinformationssystems *autobahn.nrw.de* stellen einen der Ausgangspunkte dieses Teilprojekts dar, da die Daten der Autobahn-Messstellen in diesem Vorhaben genutzt werden. Umgekehrt liefern die Ergebnisse des Teilprojekts wichtige Erkenntnisse für eine Weiterentwicklung dieser und anderer Verkehrsinformationsplattformen.

ESAM

(Schreckenberg,Förderkennzeichen DFG SCHR 527/5-1) (Förderzeitraum: 2011 - 2014)

Bei dem Projekt ESAM handelt es sich um ein DFG-Projekt zur experimentellen und simulativen Untersuchung des Einflusses von Anreizsystemen (z. B. Maut) auf das Verhalten von Verkehrsteilnehmern. Unter einem Anreizsystem bezogen auf ein Verkehrssystem versteht man eine Gesamtheit von Eigenschaften und Beziehungen, die durch das Verkehrssystem und dessen Umgebung induziert wird und auf Reaktionen sowie Entscheidungsverhalten des Verkehrsteilnehmers einwirkt.

UR:BAN

(Schreckenberg,Förderkennzeichen BMWi 19P11007Q) (Förderzeitraum: 2012 - 2015)

Im Projekt "UR:BAN - Urbaner Raum: Benutzergerechte Assistenzsysteme und Netzmanagement" werden empirische (anonymisierte) TomTom-Fahrzeugdaten in städtischem Verkehr in einer mikroskopischen

Verkehrssimulation mit dem Kerner-Klenov-Modell der Drei-Phasen-Verkehrstheorie simuliert. Es wird eine Datenbank mit typischen Verkehrsmustern für jede Strecke einer Stadt am Beispiel der Landeshauptstadt Düsseldorf aufgebaut. Die Datenbank dient als Grundlage zur Berechnung von relativen Mehrverbräuchen pro Streckenabschnitt in Abhängigkeit von der aktuellen Verkehrslage. Ziel ist die Berechnung von energieeffizienten Routen in Abhängigkeit von der Antriebsart des Fahrzeugs.

TIE-IN
(Wietfeld,Förderkennzeichen Ziel 2, 64.65.69-EM-1022A) (Förderzeitraum: 2011 - 2014)

Entwurf einer modellbasierten Test- und Prüfumgebung für Kommunikationsprotokolle für eine PLC-Übertragungstrecke (*Power Line Communications*) im Bereich der Elektromobilität.

openECOSPhERE
(Wietfeld,Förderkennzeichen BMWi 01ME12091) (Förderzeitraum: 2011 - 2014)

Entwurf, Analyse und Bewertung IKT-basierter Strategien zur Erbringung von Systemdienstleistungen durch Elektrofahrzeuge vor dem Hintergrund der gezielten Nutzung fluktuierender erneuerbarer Energieträger zur Ladung von Elektrofahrzeugen. Im Fokus steht dabei die Steuerung von Ladepunkten mittels mittel PLC, Mobilfunk und dem *Open Charge Point Protocol* (OCPP).

e-DASH
(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 285586) (Förderzeitraum: 2011 - 2014)

Erforschung eines funkbasierten Kommunikationsdienstes zur Steuerung von Fahrzeugflotten zur Erbringung von Energiedienstleistungen mit Fokus auf Backendsysteme.

metropol-E
(Wietfeld,Förderkennzeichen BMVBS 03EM0605D) (Förderzeitraum: 2012 - 2014)

Erforschung der Kommunikationsdienste für ein neuartiges Informationssystem für mobile Dienste in der Elektromobilität mit dem Fokus aus der Planung und Steuerung einer kommunalen Fahrzeugflotte unter Nutzung von Strom aus regenerativer Eigenerzeugung. Die vom Informationssystem erhobenen FCD können im SFB verwendet werden.

SmartC2Net
(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 318023) (Förderzeitraum: 2012 - 2015)

Entwurf und Validierung einer integrierten Systemarchitektur für einen sicheren und zuverlässigen Smart-Grid-Betrieb mit besonderem Fokus auf neuartige, miteinander verknüpfte energie- und kommunikationstechnischen Regelkreise.

AIRBEAM
(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 261769) (Förderzeitraum: 2012 - 2015)

Entwurf und Validierung von Kommunikationsprotokollen für heterogene, flugfähige, vernetzte Sensorplattformen für den Einsatz im Rettungswesen und Katastrophenschutz.

Anchors
(Wietfeld,Förderkennzeichen BMBF 13N12204) (Förderzeitraum: 2012 - 2015)

Erforschung von neuartigen drahtlosen Kommunikationsdiensten zur redundanten Vernetzung autonomer, unbemannter Robotiksysteme in der Luft und am Boden mit dem Ziel der Fernerkundung bei Störfällen in Industrieanlagen.

eNterop
(Wietfeld,Förderkennzeichen BMWi 01MX12060D) (Förderzeitraum: 2012 - 2014)

Entwurf einer Referenzimplementierung interoperabler Kommunikation zwischen Elektrofahrzeug und Ladesäule sowie Erforschung einer automatisierten TTCN-3-basierten (*Testing and Test Control Notation*) Konformitätstestumgebung in direkter Kooperation mit der Automobilindustrie und Infrastrukturanbietern.

DFG-Forschergruppe 1511
(Wietfeld,Förderkennzeichen Wi 3751/1-1, Wi 3751/2-1) (Förderzeitraum: 2012 - 2014)

Weiträumige Systemüberwachung und Schutz von elektrischen Energiesystemen. Erforschung einer neuen, hybriden Simulationsplattform zur gemeinsamen Untersuchung von Energie- und IKT-Systemen. Konzeptionierung und Untersuchung neuartiger, echtzeitfähiger Kommunikationsinfrastrukturen unter Einbeziehung von *Software-Defined-Networking*-Konzepten.

PlanGridEV

(Wietfeld,Förderkennzeichen EU 608957) (Förderzeitraum: 2013 - 2016)

Entwicklung neuartiger Planungsregeln und operativer Prinzipien für die optimale Integration von Elektrofahrzeugen in zukunftsorientierte Energienetze.

SecInCoRe

(Wietfeld,Förderkennzeichen EU-FP7 Grant Agreement No 607832) (Förderzeitraum: 2014 - 2017)

Verbesserung der grenzüberschreitenden Zusammenarbeit von Polizei, Rettungsdiensten, Feuerwehren und sonstigen Rettungsorganisationen auf Basis eines cloud-basierten gemeinsamen Informationsraums.

Relational exploration, learning and inference

(Kersting,Förderkennzeichen DFG KE 1686/2-1) (Förderzeitraum: 2012 - 2015)

Das Ziel dieses Projekts im Rahmen des DFG Schwerpunktprogramms 1527 „Autonomous Learning“ ist die Erforschung von Exploration, Lernen und Inferenz auf der Grundlage sogenannter relationaler Repräsentationen. Solche Repräsentationen implizieren starke A-priori-Annahmen über die Weltstruktur und ermöglichen es, aus unsicherer Erfahrung kompakte und über einzelne Gegenstände verallgemeinernde Modelle von der Wirkungsweise von Aktionen zu lernen. Dies erlaubt eine konzeptionell neuartige Art von Exploration, die bewusst Objekte in Situationen untersucht, auf welche das derzeitige Weltverständnis nicht verallgemeinert. Verkehrsdaten und die Erweiterung von generalisierte linear Modelle im Allgemeinen auf relationale Repräsentationen sind nicht Gegenstand dieses Projektes.

Efficient Inference for Probabilistic Relational Models using Symmetries and Linear Programming Relaxations

(Kersting,Förderkennzeichen GIF 1180) (Förderzeitraum: 2012 - 2015)

Das Ziel dieses Projekts im Rahmen des German Israeli Foundation for Scientific Researchs and Development (GIF) ist die Erforschung von effizienten Schlussfolgerungsmethoden für probabilistische, relationale Modelle. Es fokussiert sich auf Wahrscheinlichkeitsmodelle über endlichen , diskreten Zufallsvariablen und insbesondere auf Relaxierungen des Schlussfolgerns mittels linearer Programme, wenn viele Symmetrien vorhanden sind. Ziel ist ein besseres theoretisches Verständnis für Symmetrien und ihrer Ausnutzung zur effizienten Inferenz. Verkehrsdaten und Zählverteilungen sind nicht Gegenstand dieses Projektes.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	3	176.400	3	176.400	3	176.400	3	176.400
Summe:	—	176.400	—	176.400	—	176.400	—	176.400
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	176.400		176.400		176.400		176.400	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (B4)							
Wissen-schaftliches Personal	1	Kersting, Kristian, Prof. Dr., Prof	Data Mining	LS 8, Fak. Inf, TU Do	10,0	—	Grund-ausstattung
	2	Schreckenberg, Michael, Prof. Dr., Prof	Verkehrsmodellie-rung	Physik von Transport und Verkehr, Uni Duisburg-Essen	10,0	—	Grund-ausstattung
	3	Wietfeld, Christian, Prof. Dr., Prof	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	10,0	—	Grund-ausstattung
	4	Molina Ramirez, Alejandro, M.Sc. WiMi	Data Mining	LS 8, Fak. Inf, TU Do	19,9	—	Grund-ausstattung
	5	Schweikowski, Florian, M.Sc. WiMi	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	19,9	—	Grund-ausstattung
	6	N.N. , WiMi	Verkehrsmodellie-rung	Physik von Transport und Verkehr, Uni Duisburg-Essen	29,85	—	Grund-ausstattung
	7	N.N. , SHK	Verkehrsmodellie-rung	Physik von Transport und Verkehr, Uni Duisburg-Essen	10,0	—	Grund-ausstattung
	8	N.N. , SHK	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	10,0	—	Grund-ausstattung

Nichtwissen-schaftliches Personal	9	Hoffmann, Jutta, Verw. Ang.	—	Physik von Transport und Verkehr, Uni Duisburg-Essen	3,98	—	Grund-ausstattung
	10	Foese, Matthias, Techn. Ang.	—	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	3,98	—	Grund-ausstattung
Beantragte Personalmittel (B4)							
Wissen-schaftliches Personal	11	Habel, Lars, Dipl.-Phys. WiMi	Verkehrsmodellie-rung	Physik von Transport und Verkehr, Uni Duisburg-Essen	—	Doktorand/in	—
	12	Ide, Christoph, Dipl.-Ing. WiMi	Kommunikations-netze	LS Kommunikati-onsnetze, Fak. ET/IT, TU Do	—	Doktorand/in	—
	13	N.N. , WiMi	Data Mining	LS 8, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Kersting, Kristian**
   Projektleitung. Schwerpunkt Data Mining. Mitarbeit in den Arbeitspaketen 1, 3, 5, 6.

2. **Schreckenberg, Michael**
   Projektleitung. Schwerpunkt Daten der Verkehrsmodellierung und -prognose. Mitarbeit in den Arbeitspaketen 1, 2, 6.

3. **Wietfeld, Christian**
   Projektleitung. Schwerpunkt Kommunikationsnetze. Mitarbeit in den Arbeitspaketen 1, 4, 5, 6.

4. **Molina Ramirez, Alejandro**
   Mitarbeit im Arbeitspaket 3.

5. **Schweikowski, Florian**
   Mitarbeit in den Arbeitspaketen 1, 3 und 5.

6. **N.N.,**
   Mitarbeit in den Arbeitspaketen 1, 3 und 6.

7. **N.N.,**
   Unterstützung der wissenschaftlichen Mitarbeiter in den relevanten APs.

8. **N.N.,**
   Unterstützung der wissenschaftlichen Mitarbeiter in den relevanten APs.

9. **Hoffmann, Jutta**
   Dokument- und Terminverwaltung sowie ggf. Reiseplanung und -abrechnung.

10. **Foese, Matthias**
    Unterstützung der technischen Infrastruktur für Datenbankzugriffe zwischen den Instituten.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

11. **Habel, Lars**
    Mitarbeit in den Arbeitspaketen 1, 2 und 3, insbesondere bei der Modellierung und Simulation von Fahrzeugverkehr.

12. **Ide, Christoph**
    Mitarbeit in den Arbeitspaketen 1, 4 und 6, insbesondere bei der Modellierung, Simulation sowie Erforschung von Verfahren zur Entlastung von ressourcenbeschränkten Kommunikationskanälen.

13. **N.N.,**
    Mitarbeit in den Arbeitspaketen 1, 3, 5 und 6, insbesondere bei der Entwicklung und Evaluierung von Data Mining Verfahren für die ressourcen-efficienten Verkehrsflussplanung.

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	8.000	8.000	8.000	8.000
Uni DUE: vorhandene Mittel aus Haushalt	4.000	4.000	4.000	4.000
Summe der vorhandenen Mittel	12.000	12.000	12.000	12.000
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt C1

3.1.1 Titel:

Merkmalsselektion in hochdimensionalen Daten am Beispiel der Risikoprognose in der Onkologie

3.1.2 Fachklassifizierung:

Maschinelles Lernen, Bioinformatik, Molekularmedizin

3.1.3 Projektleitung

PD Dr. Schramm, Alexander, 21.09.1968

Onkologisches Labor, Kinderklinik, Universitätsklinikum Essen
Hufelandstraße 55
45122 Essen

Tel.: 0201-723-2506
Fax : 0201-723-5750
E-Mail: `alexander.schramm@uni-due.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Dr. Lee, Sangkyun, 15.02.1977

LS 8, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 12
44227 Dortmund

Tel.: 0231-755-6490
Fax : 0231-755-5105
E-Mail: `sangkyun.lee@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

( ) nein (x) ja, befristet bis zum 31.12.2014, eine weitere Beschäftigung ist vorgesehen bis zum 31.12.2018

Prof. Dr. Rahmann, Sven, 01.08.1974

Lehrstuhl für Genominformatik, Institut für Humangenetik, Universitätsklinikum Essen, Universität Duisburg-Essen
Hufelandstraße 55
45147 Essen

Tel.: 0201-723-4048
Fax : 0201-723-5826
E-Mail: `Sven.Rahmann@uni-due.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	(x) ja	( ) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	(x) ja	( ) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	(x) ja	( ) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Ziel des Teilprojekts ist die Konstruktion zuverlässiger Prädiktionsmodelle aus sehr hochdimensionalen Daten zur präzisen Risikoabschätzung in der Onkologie am Beispiel des Neuroblastoms, eines häufigen soliden Tumors des Kindesalters.

Technologisch stehen DNA-Microarrays zur Verfügung, die als Merkmale Expressionswerte (eine quantitative Maßzahl der Aktivität eines Gens) in Tumor- und Kontrollgewebe von mehreren zehntausend Genen oder mehreren hunderttausend Exons messen. Neuerdings gibt es zudem hochparallele, als *next generation sequencing* (NGS) bezeichnete, DNA- und RNA-Sequenziertechnologien, die neben der quantitativen Erfassung derselben Merkmale auch Mutationen in der DNA und RNA als zusätzliche Merkmale liefern. Das Problem hierbei ist, dass die Anzahl der Merkmale viel größer ist als die Anzahl der Patienten; die Ressource der verfügbaren Stichproben ist also relativ gegenüber der Zahl möglicher Merkmale sehr stark beschränkt. Ein wichtiger Schritt ist daher eine stabile und interpretierbare Merkmalsauswahl.

In der ersten Projektphase wurden hierzu grundlegende Merkmalsselektionsverfahren (zum Beispiel Predictive Analysis for Microarrays (PAM), LARS, SVM-basierte Verfahren, LASSO und viele andere mehr) als Erweiterung von RapidMiner implementiert. Neben der Robustheit wurde auch die Stabilität der Auswahl modelliert. Diese Erweiterung wurde bereits über 2700 mal heruntergeladen. Ein weiteres Ensemble-Merkmalsselektionsverfahren basierend auf präkonditioniertem LASSO wurde als R-Paket `rsig` implementiert. Mit Hilfe von Merkmalsauswahlverfahren, die auf maschinellen Lernverfahren wie der Stützvektormethode (SVM), PAM, präkonditioniertem LASSO und Ensemblemethoden basieren, haben wir für aggressive Tumore charakteristische Merkmale identifiziert, zum Beispiel die Expression von JARID1C und dem NTRK1-Gen. Die biologische Bedeutung dieser Merkmale wird in Laborexperimenten weiter verfolgt. Dennoch konnte, auch an anderen Orten, eine exakte Risikoabschätzung anhand von Exon-Expressionswerten bislang nicht erreicht werden. Eine weitere Verbesserung durch Anwendung anderer Lern- und Selektionsverfahren erscheint unwahrscheinlich.

Daher ist es sinnvoll, weitere Daten außer Genexpressionswerten einzubeziehen, nämlich Mutationen in den Gensequenzen, und wir werden in der zweiten Projektphase zusätzlich NGS-Daten nutzen. Sven Rahmann, der an Algorithmen auf diesem Gebiet forscht, wurde daher als neuer Projektleiter hinzugezogen. Ein typischer NGS-Datensatz umfasst 50 Millionen DNA-Fragmente (*reads*) zu je 100 Nukleotiden (5 GB Rohdaten). Diese werden durch Sequenzvergleich gegen das Referenzgenom (3 GB) und statistische Tests zu Merkmalen (Exon-Expressionswerte, Sequenzmutationen) verdichtet. Der Prozess des *read mappings* ist sehr speicher- und rechenaufwändig. Er wird weltweit auf Tausenden von Rechenclustern durchgeführt; dabei werden sehr viele CPU-Stunden verbraucht. Wir werden daher erstmalig den Energiebedarf dieser Merkmalsgewinnung systematisch untersuchen und dann optimieren, indem wir auf aktuelle Hardware (GPUs, SSDs) angepasste Algorithmen erforschen und verwenden.

Auf der Modellierungsseite widmen wir uns verstärkt probabilistischen graphischen Modellen, um strukturelle Beziehungen zwischen den Merkmalen (zum Beispiel Geninteraktionen oder Korrelationen zwischen Genmutationen und Expressionswerten anderer Gene) aus den Daten abzuleiten und von Menschen interpretierbare Prognosemodelle zu erstellen.

Wir werden zusätzlich Aspekte der Tumor-Evolution betrachten, die für die Tumorbiologie und insbesondere für die Therapie entscheidend sind. Wie fast alle Tumorentitäten ist das Neuroblastom initial gut therapierbar. In der Rückfallsituation jedoch erweisen sich diese Tumoren als resistent gegen alle gängigen medizinischen Verfahren. Methoden zur räumlich-zeitlich aufgelösten Merkmalsselektion sollen basierend auf probabilistischen graphischen Modellen wie *Latent Dirichlet Allocation* (LDA) unter Nutzung der Daten aus den Hochdurchsatzverfahren von Patienten vor und nach Therapie entwickelt werden. Hierdurch wollen wir molekulare Selektionsprozesse im Tumor nachvollziehen und ein Bild der Resistenzentstehung entwickeln, welches möglicherweise einen Weg zur Behandlung bisher nicht therapierbarer Tumorrezidive weist.

3.3 Bisherige Entwicklung des Teilprojekts

3.3.1 Bericht und Stand der Forschung

Wir berichten zunächst zum Stand der jeweiligen Arbeitspakete aus der ersten Phase und stellen dann den Stand der Forschung bei der Mutations- und Genexpressionsanalyse von Hochdurchsatzsequenzdaten dar, auf dem der neue Projektleiter Sven Rahmann besonders ausgewiesen ist und über gemeinsame Vorarbeiten mit Alexander Schramm verfügt.

Probenpräparation und Qualitätskontrolle (Arbeitspaket 1): Begleitend zur ersten Phase des SFB wurden durch BMBF-Mittel im Rahmen des Nationalen Genomforschungsnetzes (NGFN) Expressionsprofile von Patienten mit Neuroblastom unter Einsatz von Affymetrix Human Exon STv1.0 Arrays generiert. Im Rahmen von bestehenden Zusammenarbeiten mit der nationalen Neuroblastom-Studie und europäischen Kooperationspartnern (AMC Amsterdam, Holland; CMGG Ghent, Belgien; Royal College of Surgeons, Dublin, Irland) wurden hierzu auch auf der gleichen technischen Plattform erhobene Daten aus nationalen und internationalen Studien eingeschlossen. Die Datenspeicherung erfolgte anonymisiert; der Zugang zu den Daten erfolgte über Nutzername / Password in der auch von großen EU-geförderten Konsortien wie ENCCA (http://www.encca.eu) genutzten Datenbank R2 (http://www.r2.amc.nl). Alle Array-Experimente an den verschiedenen Standorten wurden nach Standard-Prozeduren (SOPs) durchgeführt; die resultierenden Dateien mit den Rohexpressionswerten (CEL-Dateien) wurden dem Teilprojekt zur Verfügung gestellt.

Die Sonden auf den Affymetrix-Arrays liefern ca. 6 Millionen Merkmalswerte, die entweder auf der Ebene von Genen oder Exons zusammengefasst werden können. Um die Daten verschiedener Patienten vergleichbar zu machen, sind Normalisierungsverfahren notwendig. Ein hierfür häufig angewendetes Verfahren ist der RMA-Algorithmus (robust multi-array analyses [475]). RMA erhält als Eingabe eine Matrix der Dimension Merkmale mal Experimente mit unnormalisierten Sonden-Expressionswerten (Rohdaten), um daraus normalisierte Expressionswerte abzuleiten und zu Exon- oder Genexpressionwerten zusammenzufassen, die zwischen verschiedenen Experimenten verglichen werden können. Die Auswahl der sogenannten *core*-Sondenmengen liefert ca. 300 000 Exon-Expressionswerte.

Dieses Vorgehen ist für unsere Zwecke aus mehreren Gründen ungünstig: Zum einen hängen die Ergebnisse der Normalisierung von der Eingabemenge der Experimente ab, und daher wird es von Batch zu Batch zu unterschiedlichen Resultaten kommen. Dies wird insbesondere beim Partitionieren von Datensätzen in Test- und Trainingsdaten problematisch, da separat RMA-normalisierte Daten nicht notwendigerweise vergleichbar sind. Somit ist eine Grundvoraussetzung für statistische Lernverfahren nicht gegeben. Zum anderen muss RMA bei der Integration neuer Datensätze stets neu durchgeführt werden, was nicht nur rechenintensiv und ineffizient ist, sondern auch zu Problemen bei der Definition generalisierbarer Prädiktoren führt, da sich die Datenbasis, auf der die Prädiktoren berechnet wurden, und die auf RMA normalisierten Daten beruht, ändert. Als Alternative haben wir die fRMA-Methode (*frozen RMA* [482]) identifiziert, die eine globale Referenz aus über 1000 Micorarrays zur Normalisierung nutzt. Dadurch werden die oben angeführten Einschränkungen von RMA überwunden.

Zur Qualitätskontrolle werden vom Array-Hersteller Affymetrix Spike-In Kontrollsonden und Referenzpunkte zur Grid-Orientierung angeboten. Als plattformunabhängige Qualitätskontrolle diente der GNUSE-Score (global normalised unscaled standard error; [483]), der im Gegensatz zu Verfahren für individuelle Microarrays genutzt werden kann. Durch den kombinierten Einsatz von fRMA und GNUSE war es möglich, die Daten aus den verschiedenen europäischen Partnerlaboren vergleichbar zu machen, was einen wichtigen Schritt auf dem Weg zu einem einheitlichen Prädiktor darstellt. Die Details hierzu haben wir im Rahmen eines technischen Reports veröffentlicht [C1/498].

Die mit fRMA prozessierten und mit GNUSE gefilterten Daten ($n = 176$ nach Filtern) wurden in APs 3 und 6 eingesetzt, wobei Daten aus verschiedenen Standorten zum Einsatz kamen (Amsterdam, Dublin,

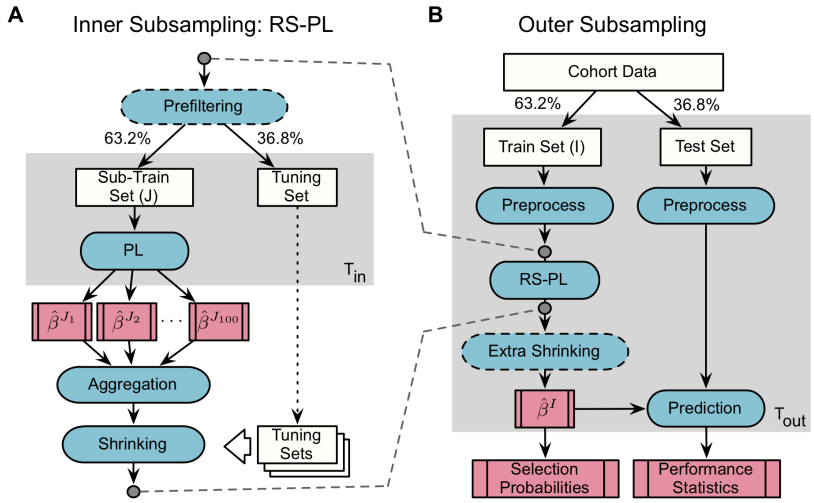


Abbildung 3.1.: **Ensemblemethoden zur Merkmalsselektion (Arbeitspaket 3).** A: RS-PL-Algorithmus (*Robust Selection with Preconditioned LASSO*) B: Äußeres Subsampling.

Essen, Ghent, Lund). Obwohl diese Daten standortabhängige Zufallseffekte enthalten, die durch technische Unterschiede bedingt sein können, ermöglichen unsere Präprozessierungsschritte, sie als einen Datensatz zu betrachten.

Dateninspektion und technische Validierung (Arbeitspaket 2): Gefundene Genexpressionsmuster wurden sowohl mit bioinformatischen Methoden als auch laborseitig auf ihre Funktion hin analysiert. Auf bioinformatischer Seite wurde eine Gene Set Enrichment Analyse (GSEA) annotierter funktioneller Terme mit Hilfe der Gene Ontology (GO) durchgeführt. Biologisch wurde die Bedeutung von Targetgenen für die Aufrechterhaltung des Phänotyps von Tumorzellen untersucht. Neben bereits bekannten Zusammenhängen wurde u.a. gezeigt, dass für die Neuroblastom-Biologie relevante Faktoren auch die unterschiedlichen Exon-Gebräuche regulieren [C1/TB1/495]. In Kooperation mit einer holländischen Arbeitsgruppe konnten wir ein neues Neuroblastom-verursachendes Gen identifizieren [C1/492]. Weiterhin konnten wir zeigen, dass eine Klasse von neuen tumorspefizischen Wachstumsinhibitoren mit klinischer Relevanz auch gegen aggressive Neuroblastome wirksam sein könnte [C1/496].

Ensemblemethoden zur Merkmalsselektion (Arbeitspaket 3): Wir haben zwei neue Methoden zur hybriden Merkmalsselektion entwickelt: eine, die auf Schnell-Ensemble von MRMR (minimum redundancy maximum relevance) beruht und eine, die auf regularisierten Cox-Regressionsmodellen beruht.

Unsere erste Vorgehensweise [C1/494] teilt innerhalb einer Kreuzvalidierung eine Beispielmenge in m Teilmengen auf und sucht in jeder der m Mengen nach relevanten Merkmalen gemäß des Kriteriums, das an den Algorithmus übergeben wird. Die Methode nutzt die Dekomponierbarkeit der Maße für Redundanz und Relevanz in unabhängige Summen aus. Der Algorithmus skaliert sehr gut, weil er Zwischenergebnisse speichert und nur einmal über die Beispiele gehen muss für jede Kombination von Merkmalen oder Merkmal und Klasse. Die Merkmalsselektionshäufigkeiten werden also nicht jedes Mal erneut berechnet, wie es bei einem reinen Wrapper-Ansatz geschieht. Als Kriterium kann die Kombination von maximaler Relevanz und minimaler Redundanz gewählt werden. Ansätze, die ausschließlich auf Relevanz und Redundanz basieren, führen häufig zu statistisch instabilen Merkmalsauswahlen. Durch die Verwendung von Ensemblemethoden kann die Stabilität erhöht werden und durch die Umordnung von Berechnungen kann unsere Fast-Ensemble-Methode Zwischenergebnisse wiederverwenden, wodurch der praktische Einsatz Ensemble-basierter, stabiler Merkmalsauswahlen ermöglicht wird. Die Software ist als *Feature Selection Extension Plug-in* von RapidMiner verfügbar und wird erfolgreich im Teilprojekt C3 wie überhaupt international genutzt[1].

Unsere zweite Vorgehensweise ist in Abbildung 3.1 skizziert. Eine entsprechende Veröffentlichung im Journal PLoS One wurde unter der Maßgabe kleinerer Änderungen bereits akzeptiert, aber noch nicht endgültig angenommen. Die Gesamtmenge der Datensätze wird in eine Trainingsmenge I (63.2% der Datensätze) und eine Testmenge (Rest, 36.8%), analog zum Bootstrapping [472], aufgeteilt und getrennt präprozessiert. Die Trainingsmenge I wird verwendet, um einen Klassifikator $\hat{\beta}^I$ zu trainieren, und die Testmenge, um

[1] http://sourceforge.net/projects/rm-featselext

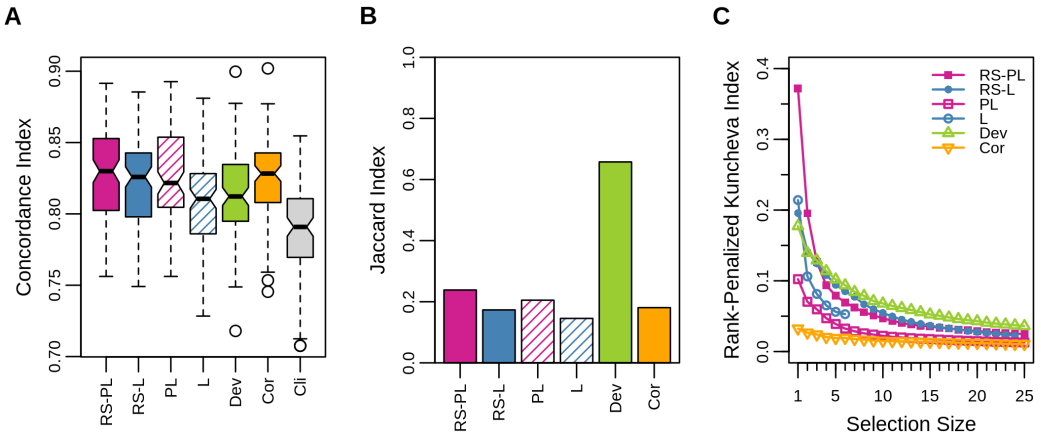


Abbildung 3.2.: **Ensemblemethoden zur Merkmalsselektion (Arbeitspaket 3): Performance.** A: Vorhersageperformanz. B,C: Robustheit der Merkmalsselektion. (A,B,C: höher ist besser). Methoden: RS-PL: robuste Selektion mit präkonditioniertem LASSO (unsere Methode); RS-L: robuste Selektion mit LASSO; PL: präkonditioniertes LASSO; L: LASSO; Dev: Filterung nach Merkmalsvarianz; Cor: Filterung nach Korrelation zwischen Merkmalsgröße und korrekter Klassifizierung; Cli: Ausschließliche Nutzung klinischer Daten (keine Microarrays).

Performance-Maßzahlen dieses Klassifikators zu erhalten. Der Klassifikator $\hat{\beta}^I$ ist dabei ein dünnbesetzter Vektor der Dimension des Merkmalsraums; mit ihm wird eine Linearkombination der Merkmale x gebildet, die dann in einem Cox-Modell [471] zur Risikoabschätzung verwendet wird:

$$h(t \,|\, x) = h_0(t)\, \exp(x^T \hat{\beta}^I),$$

Dabei ist $h_0(t)$ eine Basis-Risikofunktion, die aus Überlebensdaten geschätzt werden kann.

Die Prozedur wird $T_{\text{out}} = 100$ mal wiederholt. In jeder Iteration wird der Klassifikator $\hat{\beta}^I$ durch Aggregation und Shrinking mehrerer Sub-Klassifikatoren $\hat{\beta}^{J_k}$ gebildet, $k = 1, \ldots, T_{\text{in}} = 100$. Dazu wird die Trainingsmenge zufällig weiter aufgeteilt in eine Untertrainingsmenge J_k und eine Tuningmenge $I \setminus J_k$ mit denselben Verhältnissen wie oben. Auf jedes J_k wird der *Preconditioned LASSO* (PL) Algorithmus angewendet [484], der jeweils die Koeffizienten $\hat{\beta}^{J_k}$ eines Modells liefert. Aus diesen Modellen wird das Modell $\hat{\beta}^I$ durch Aggregation gebildet. In diesem Fall ist dies durch Durchschnittsbildung der Koeffizienten möglich, da es sich um generalisierte lineare Modelle handelt. Der Durchschnittsvektor wird durch Anwendung eines geeigneten Schwellenwerts geschrumpft (d.h. kleine Koeffizienten auf Null gesetzt), um den Klassifikator von weniger Merkmalen abhängig sein zu lassen. Jede solche Iteration (für eine feste Trainingsmenge I) nennen wir RS-PL (*Robust Selection with Preconditioned LASSO*).

Wegen des sehr hohen Rechenbedarfs wurde das in Teilprojekt A3 entwickelte R-Paket `BatchExperiments` zur Parallelisierung genutzt. Unser Verfahren steht als R-Paket `rsig` in CRAN zur Verfügung (http://cran.r-project.org/package=rsig).

In Abbildung 3.2 ist die Klassifikationsgüte des resultierenden RS-PL-Prädiktors im Vergleich zu anderen Verfahren als Konkordanzindex (A), als Jaccard-Index (B) und als Kuncheva-Index (C) dargestellt. Der Konkordanzindex [473] spiegelt dabei den Anteil aller Patientenpaare (überlebend vs. nicht-überlebend) wider, bei denen der Patient mit dem höheren Risikofaktor eine kürzere Überlebenszeit hat. Dies kann auch als generalisierter AUC-Wert (*area under the ROC curve*) aufgefasst werden. Der Jaccard-Index misst die durchschnittliche relative Größe der Schnittmenge zwischen gewählten Merkmalen in verschiedenen Iterationen des RS-PL-Algorithmus. Der *rank-penalized Kuncheva index* zwischen zwei Merkmalsmengen ist eine von uns neu eingeführte Größe, die als Quotient des *Kuncheva-Index* und der Canberra-Distanz definiert ist. Der Kuncheva-Index ist ein Maß für die durchschnittliche relative Größe der Schnittmengen zwischen Merkmalsmengen; die Canberra-Distanz ermittelt den durchschnittlichen Grad der Rang-Unterschiede zwischen (sortierten) Merkmalsgruppen. Hinsichtlich der kombinierten Performanz und Robustheit schneidet unser Verfahren RS-PL am besten von allen getesteten Methoden ab (Abb 3.2). Lediglich bei der Robustheit schneidet eine Methode (Dev) besser ab als RS-PL; dies wird jedoch durch die verringerte Performanz von Dev erkauft.

Gruppierung von Merkmalen zur Merkmalsselektion (Arbeitspaket 4): In diesem Arbeitspaket wird die Identifizierung korrelierter Merkmale wie z.B. ko-regulierter Exons adressiert, um Redundanz in der Merkmalsauswahl aufzuspüren und zu vermeiden. Die Identifikation von solchen Merkmalsclustern kann für die Merkmalsauswahl nützlich sein und zu einer verbesserten Performanz und Stabilität führen. Die Ausrichtung dieses Arbeitspaketes musste jedoch angepasst werden, da die statistisch verlässliche Identifikation von Merkmalsgruppen im hochdimensionalen Raum bei geringem Stichprobenumfang schwierig ist und sich die Zusammensetzung der Merkmalsgruppen durch Variation der Trainingsdatensätze ändern kann.

In unserem modifizierten Ansatz verfolgten wir die Idee, implizite Merkmalsgruppen in Hauptkomponenten (PCs, principal components) dergestalt zu finden, dass die Projektion der Merkmale auf den durch die Hauptkomponenten aufgespannten Unterraum einen maximalen Konkordanzindex erreicht. Dieses Verfahren ist als Hauptkomponenten-Regression beschrieben worden [293, Abschnitt 3.5.1]. Dieses Prinzip wurde auch in Arbeitspaket 3 genutzt, bei dem die implizite Identifizierung von Merkmalsgruppen in die präkonditionierenden Schritte von RS-PL integriert wurde. Dies trägt dazu bei, dass RS-PL signifikant besser als RS-L (LASSO ohne Präkonditionierung) in Bezug auf Performanz und Stabilität abschneidet (Abbildung 3.2).

Meta-Studien zum Einsatz verschiedener Plattformen (Arbeitspaket 5): Zur Validierung der Exon-Array-Studien mit anderen technologischen Plattformen wurden RNA-Sequenzierungen und quantitative PCR an primären Neuroblastomen eingesetzt. Hier konnten wir zunächst eine hohe Korrelation der eingesetzten Technologien untereinander nachweisen [C1/TB1/493]. Beim Vergleich von RNA-Sequenzierung und Exon-Arrays gibt es durch die unterschiedliche Abdeckung der genomischen Bereiche intrinsische Limitationen der Vergleichbarkeit. Wir konnten dennoch gemeinsame Signalwege identifizieren, die anhand der Ergebnisse von Array- und Sequenziertechnologie typisch für aggressive Tumore sind [C1/496]. Besonders interessant sind die übereinstimmend gefundenen Veränderungen primärer Stoffwechselwege. Diese Ergebnisse werden im Rahmen eines vom BMBF positiv begutachteten Systembiologieverbunds (SYSMED NB, BMBF-Ausschreibung „e-med“) in Zukunft weiter verfolgt.

Integration fremder Datensätze (Arbeitspaket 6): Dieses Arbeitspaket war durch die Beobachtung motiviert, dass sich in Abhängigkeit des Standorts (Amsterdam, Dublin, Essen, Ghent, Italien und Lund), an dem Microarrays präpariert werden, Unterschiede in der Verteilung der Intensitätswerte ergeben. Dies kann am ehesten durch technische Unterschiede und nicht durch Unterschiede in den Patientenkollektiven erklärt werden [C1/498]. Während die in Essen präparierten Arrays insgesamt eher schwächere Signale lieferten, zeigten sich in Arrays aus anderen Standorten typische Artefakte. Zur Beseitigung solcher technischer Unterschiede wurden Methoden zur Entfernung von Zufallseffekten getestet, u.a. COMBAT [476] und SVA [479]. Die folgende Aufstellung zeigt die Anzahl der verwendbaren Microarray-Datensätze von den verschiedenen Standorten, nachdem sie mit der GNUSE-Score (Schwellenwert 1.0) gefiltert wurden. Die GNUSE-Filtermethode wurde bereits in AP 1 vorgestellt.

Standort	Amsterdam	Italien	Dublin	Essen	Lund	Gesamt
Anzahl	87	29	37	16	7	176

Für diese Datensätze haben wir Transferlernen mit verschiedenen Quell- und Zielstandorten versucht, basierend auf dem Ansatz von Helleputte und Dupont [474]. Dabei wurde variiert, wie viele Merkmale des Quellstandortes (1%, 5%, 10%) verwendet werden; alternativ wurde kein Transferlernen durchgeführt („T Only”, nur Zieldaten; „All”, kombinierte Daten aller Standorte).

Abbildung 3.3 zeigt die Ergebnisse für zwei verschiedene Auswahlen von Quelle und Ziel. Dabei wurden 100 Versuche mit zufälligem Subsampling (Training/Test = 70%/30%) durchgeführt. In den Transferlern-Ansätzen wurden die 1% (bzw. 5%, 10%) der Merkmale mit dem höchsten Konkordanzindex in den Quelldaten gewählt. Die gewählten Quellmerkmale wurden weniger (50%) bestraft, wenn LASSO [485] für die endgültige Merksmalsauswahl auf den Zieldaten angewendet wurde. LASSO wurde auch bei „T Only” und „All” verwendet. Alle Parameter wurden durch 10-fache Kreuzvalidierung bestimmt. Wie in der Abbildung ersichtlich, brachte das Transferlernen keinen signifikanten Vorteil (hier gezeigt anhand des Konkordanzindex) gegenüber der Nutzung aller Daten („All”), p-Werte > 0.1. Die Verwendung von Transferlernern erhöhte ebenso die Varianz der Prädiktionsgüte, was wir als Nachteil werten.

Aus den genannten Gründen haben wir uns entschlossen, die Daten der verschiedenen Standorte als einen großen Datensatz zu betrachten und zu analysieren, mit dem Vorteil, dass standortübergreifende, transferierbare Signaturen erhalten blieben (AP 3).

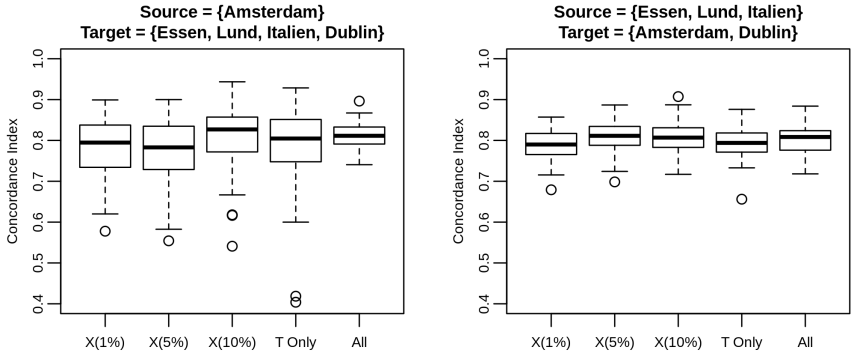


Abbildung 3.3.: **Vergleich des Transferlernens mit der Nutzung kombinierter Daten.** Transferlern-Ansätze („X") mit verschiedenen Merkmalsanteilen der Quelldaten wurden mit Ansätzen ohne Transfer veglichen: nur Zieldaten („T Only") oder alle Daten kombiniert („All"). Zur Merkmalsauswahl und Prädiktor-Schätzung wurde Cox-Regression mit LASSO-Strafterm (100 Versuche mit zufälligem Subsampling) wurde verwendet.

Kombination von Merkmalssätzen (Arbeitspaket 7): Eine Kombination aus Gen/Exon-Expressionsdaten und klinisch-molekularen Parametern wie dem MYCN-Status und der Patientengeschichte wurde in unserer Analyse berücksichtigt. In diesem Kontext zeigte sich, dass die Expression einzelner Gene mindestens genauso informativ war wie der derzeit stärkste prognostische Parameter, die MYCN-Kopienzahl (Auswahlhäufigkeit 0.67); insbesondere die Gene NTRK1 und SCN7A können als sehr robuste Merkmale herangezogen werden (relative Auswahlhäufigkeit 1.00 bzw. 0.98). Andere häufig ausgewählte Merkmale sind u.a. Gene für Natrium- und Kaliumkanalproteine. Die Hinzunahme der Information über den MYCN-Status brachte keinen signifikanten Zugewinn bei der Performanz des Gen-basierten Prädiktors. Für die Zukunft halten wir eine Kombination von Gen-Expressionsdaten und DNA-Sequenzmutationen, die sich aus sogenannten *next generation sequencing*-Daten (NGS-Daten) ergeben, für vielversprechend. Da die Datenbasis hierfür gerade geschaffen wird, sind Aussagen über die Performanz derzeit noch nicht möglich und sollen in der nächsten Phase untersucht werden.

Analyse von Hochdurchsatz-Sequenzdaten (neuer PL Rahmann): Hochdurchsatz-Sequenzierung, auch als *next generation sequencing* (NGS) bezeichnet, umfasst eine Reihe von Technologien (zurzeit z.B. Illumina, Roche/454, Pacific Biosciences, Ion Torrent), die seit etwa 2005 die klassische DNA-Sequenzierung nach Sanger ablösen. Große Mengen von DNA oder RNA können hiermit kostengünstig sequenziert werden. RNA-Sequenzierung (RNA-seq) wird eingesetzt, um Genexpressionswerte analog zur Microarray-Technologie, die in der ersten Phase verwendet wurde, zu bestimmen. RNA-seq hat dabei den Vorteil, dass sie bei geeigneter Sequenziertiefe eine bedeutend feinere Auflösung als die (indirekten) Microarray-Signale liefern kann, da direkt die RNA-Fragmente, die die einzelen Gen-Exons repräsentieren, identifiziert und gezählt werden. Ferner ist es mit RNA-seq möglich, Mutationen in den transkribierten Genen zu identifizieren. Bei der DNA-Sequenzierung erhält man statt der Sequenzen der aktiven Gene die Sequenz des Gesamtgenoms oder eines ausgewählten Teils, z.B. des Exoms. Unter dem Exom versteht man den Bereich des Genoms, der grundsätzlich in RNA übersetzt werden könnte; dies sind ca. 2% des Gesamtgenoms. Der Vorteil der DNA-Sequenzierung gegenüber der RNA-Sequenzierung ist, dass die Sequenziertiefe unabhängig von der Aktivität (dem Expressionswert) ist, so dass auch Mutationen in nicht oder niedrig exprimierten Genen gefunden werden können. Erfahrungen mit der Analyse von NGS-Daten liegen am Lehrstuhl für Genominformatik (Rahmann) einschlägig vor, insbesondere in Kooperation mit dem Projektleiter A. Schramm [C1/496, C1/TB1/495, C1/TB1/493]; siehe Forschungsprofile.

DNA-Mutations-Analyse. Mit Hilfe von Exom-Sequenzierung und des von uns entwickelten Exomate-Systems haben wir krankheitsursächliche Genenmutationen des Aderhautmelanoms sowie verschiedener seltener Syndrome entdeckt.

Die Mutationsanalyse besteht aus zahlreichen Schritten (Qualitätskontrolle, Adaptertrimmung, *read mapping*, Qualitätsrekalibrierung, Re-Alignierung, statistische Tests auf Mutationen gegen Sequenzierfehler). Dabei wird das *Genome Analysis Toolkit* (GATK) unterstützend verwendet, und es werden die vom GATK vorgeschlagenen *Best Practice* Empfehlungen eingehalten [470].

Zur Adaptertrimmung hat sich als Standard-Werkzeug `cutadapt` aus der Arbeitsgruppe von Sven Rah-

mann durchgesetzt, das von dem ehemaligen SFB-Doktoranden Marcel Martin geschrieben wurde; Herr Martin wurde für seine Doktorarbeit von der TU Dortmund für den GI-Dissertationspreis vorgeschlagen.

Ein *read mapper* wird verwendet, um die DNA-Fragemente aus der Sequenzierung im Referenzgenom zu finden und anschließend mittels Sequenzvergleich und statistischen Tests an jeder Position zu entscheiden, ob eine Mutation vorliegt. Typischerweise wird einer der read mapper BWA [480] oder bowtie2 [477] verwendet; diese sind auf wenig Arbeitsspeicherverbrauch optimiert, aber CPU-intensiv, und nutzen effektiv Multicore-Prozessoren aus. Wir haben vor kurzem einen GPU-basierten *read mapper* („PEANUT") fertiggestellt, der für die rechenintensiven Schritte erstmalig eine *GPU occupancy* nahe 1.0 erreicht. Erste Tests haben ergeben, dass PEANUT sowohl schneller als auch sensitiver (Mutations- und Sequenzierfehler-toleranter) als die oben genannten read mapper sind. Nach dem *read mapping* erfolgt eine Bereinigung der Daten durch Entfernen von PCR-Duplikaten.

Gefundene Mutationen werden in einer Datenbank gesammelt. Diese verwaltet neben den Mutationen auch Proben, Patienten, Krankheiten, Phänotypen, Sequenziereinheiten, technische Parameter, und Qualitätswerte. Die Mutationen werden mit Hilfe öffentlich vorhandener Genomannotation (ENSEMBL) bewertet: Mutiert beispielsweise ein Aminosäure-Codon zu einem Stopp-Codon, wird die Funktion des Proteins dabei i.d.R. beeinträchtigt oder zerstört, und die Mutation wird als „gravierend" eingestuft. Verschiedene Filter, die über ein Web-Interface dynamisch erstellt werden können, erlauben es Medizinern, die Mutationen interaktiv im Browser nach Relevanz zu sortieren und zu visualisieren. Die dynamische Einstellbarkeit der Filter ist bislang ein Alleinstellungsmerkmal von Exomate.

RNA-seq-Analyse. Im Bereich der RNA-Sequenz-Analyse haben wir substanzielle Erfahrung im Bereich der nichtkodierenden regulatorischen RNAs, insbesondere mikro-RNAs (miRNAs), gesammelt und konnten bereits miRNAs identifizieren, die in verschiedenen Neuroblastom-Stadien differenziell exprimiert sind [C1/TB1/493]. In dieser Arbeit werden auch Normalisierungsverfahren vorgestellt, die bei wenig aktiven Genen (bzw. miRNAs) robust funktionieren.

Zur Bestimmung der Expressionswerte werden für jedes Exon bzw. jedes Gen diejenigen sequenzierten Fragmente, die auf den entsprechenden Abschnit *mappen*, gezählt. Dabei existieren jedoch zahlreiche Verzerrungsfaktoren (*biases*: Exonlänge, Sequenziertiefe, Art der RNA-Präparation), die mit Hilfe statistischer Modelle, wie sie in R-Paketen wie DEseq oder EdgeR implementiert sind, herausgerechnet werden. Diese Pakete nutzen dazu die Modellierung der absoluten Anzahl mittels Zufallsvariablen aus negativen Binomialverteilungen (auch als *overdispersed Poisson* bezeichnet) [469]. Die Effekte, die durch die Art der RNA-Präparation entstehen, etwa eine ungleichmäßige Verteilung der Fragment-Startpunkte abhängig vom Sequenzkontext, werden jedoch von diesen Modellen bisher nicht systematisch erfasst.

Daten. In einem von der deutschen Krebshilfe geförderten Projekt sind unter unserer Beteiligung (Alexander Schramm) ca. 200 Exome von Patienten mit Neuroblastom erstellt worden; diese können wie im Arbeitsprogramm beschrieben ausgewertet werden. Von etwa 150 Patienten werden Sequenzdaten von nichtkodierender RNA, speziell microRNAs vorliegen. Weitere Daten (u.a. RNA-seq) werden von nationalen und internationalen Konsortien, an denen die Projektleiter Alexander Schramm und Sven Rahmann beteiligt sind (z.B. im Rahmen von BMBF-eMed), während der Projektlaufzeit generiert und nach initialer Publikation und Verwertung im Konsortium dem SFB-Projekt zur Verfügung gestellt.

Literatur

[469] S. Anders, D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber und M. D. Robinson. „Count-based differential expression analysis of RNA sequencing data using R and Bioconductor". In: *Nature Protocols* 8.9 (2013), S. 1765–1786.

[470] G. A. Van der Auwera u. a. „From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline". In: *Current Protocols in Bioinformatics.* Bd. 43. John Wiley & Sons, Inc., 2013. Kap. 11.10.1–11.10.33.

[7] D. M. Blei, A. Y. Ng und M. I. Jordan. „Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), S. 993–1022.

[471] D. R. Cox. „Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), S. 187–220.

[472] B. Efron. „Bootstrap Methods: Another Look at the Jackknife". In: *Annals of Statistics* 7.1 (1979), S. 1–26.

[473] F. E. Harrell, K. L. Lee und D. B. Mark. „Multivariable Prognostic Models: Issues In Developing Models, Evaluating Assumptions And Adequacy, And Measuring And Reducing Errors“. In: *Statistics in medicine* 15.4 (1996), S. 361–387.

[293] T. Hastie, R. Tibshirani und J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd. Statistics. Springer, 2009.

[474] T. Helleputte und P. Dupont. „Feature Selection by Transfer Learning with Linear Regularized Models“. In: *European Conference on Machine Learning.* Springer, 2009.

[475] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs und T. P. Speed. „Summaries of Affymetrix GeneChip probe level data“. In: *Nucleic acids research* 31.4 (2003), e15.

[476] W. E. Johnson, C. Li und A. Rabinovic. „Adjusting batch effects in microarray expression data using empirical Bayes methods“. In: *Biostatistics* 8.1 (2007), S. 118–127.

[477] B. Langmead und S. L. Salzberg. „Fast gapped-read alignment with Bowtie 2“. In: *Nature Methods* 9.4 (2012), S. 357–359.

[478] S. L. Lauritzen. *Graphical Models.* Oxford University Press, 1996.

[479] J. T. Leek und J. D. Storey. „Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis“. In: *PLoS Genetics* 3.9 (2007), e161.

[480] H. Li und R. Durbin. „Fast and accurate short read alignment with Burrows-Wheeler transform“. In: *Bioinformatics* 25.14 (2009), S. 1754–1760.

[481] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville und G. M. Church. „RNA-Guided Human Genome Engineering via Cas9“. In: *Science* 339.6121 (2013), S. 823–826.

[482] M. N. McCall, B. M. Bolstad und R. A. Irizarry. „Frozen robust multiarray analysis (fRMA)“. In: *Biostatistics* 11.2 (2010), S. 242–253.

[483] M. McCall, P. Murakami, M. Lukk, W. Huber und R. Irizarry. „Assessing affymetrix GeneChip microarray quality“. In: *BMC Bioinformatics* 12.1 (2011), S. 137.

[484] D. Paul, E. Bair, T. Hastie und R. Tibshirani. „“Preconditioning” for feature selection and regression in high-dimensional problems“. In: *Annals of Statistics* 36.4 (2008), S. 1595–1618.

[485] R. Tibshirani. „Regression Shrinkage and Selection via the Lasso“. In: *Journal of the Royal Statistical Society (Series B)* 58 (1996), S. 267–288.

[486] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu und K. Knight. „Sparsity and smoothness via the fused lasso“. In: *Journal Of The Royal Statistical Society Series B* 67.1 (2005), S. 91–108.

[487] C. Wang, A. C. Potter und T. Senthil. „Classification of Interacting Electronic Topological Insulators in Three Dimensions“. In: *Science* 343.6171 (2014), S. 629–631.

[488] M. Yuan und Y. Lin. „Model selection and estimation in the Gaussian graphical model“. In: *Biometrika* 94.1 (2007), S. 19–35.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[B3/C1/399] S. Lee, M. Stolpe und K. Morik. „Separable Approximate Optimization of Support Vector Machines for Distributed Sensing“. In: *Machine Learning and Knowledge Discovery in Databases.* Hrsg. von P. Flach, T. D. Bie und N. Cristianini. Bd. 7524. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2012, S. 387–402.

[C1/489] S. Lee und S. J. Wright. „ASSET: Approximate Stochastic Subgradient Estimation Training for Support Vector Machines“. In: *International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012).* 2012, S. 223–228.

[C1/490] S. Lee und S. J. Wright. „Manifold Identification in Dual Averaging Methods for Regularized Stochastic Online Learning“. In: *Journal of Machine Learning Research* 13 (2012), S. 1705–1744.

[C1/491] S. Lee und S. J. Wright. „Stochastic Subgradient Estimation Training for Support Vector Machines“. In: *Mathematical Methodologies in Pattern Recognition and Machine Learning.* Hrsg. von P. Latorre Carmona, J. S. Sánchez und A. L. Fred. Bd. 30. Springer Proceedings in Mathematics & Statistics. Springer New York, 2013, S. 67–82.

[C1/492] J. J. Molenaar u. a. „LIN28B induces neuroblastoma and enhances MYCN levels via let-7 suppression“. In: *Nature Genetics* 44.11 (2012), S. 1199–1206.

[A1/C1/53] N. Piatkowski, S. Lee und K. Morik. „Spatio-Temporal Random Fields: Compressible Representation and Distributed Estimation“. In: *Machine Learning Journal* 93.1 (2013). Hrsg. von H. Blockeel, K. Kersting, S. Nijssen und F. Zelezny, S. 115–139.

[C1/TB1/493] S. Rahmann, M. Martin, J. H. Schulte, J. Köster, T. Marschall und A. Schramm. „Identifying Transcriptional miRNA Biomarkers by Integrating High-Throughput Sequencing and Real-Time PCR Data“. In: *Methods* 59.1 (Jan. 2013), S. 154–163.

[C1/494] B. Schowe und K. Morik. „Fast-Ensembles of Minimum Redundancy Feature Selection“. In: *Ensembles in Machine Learning Applications*. Hrsg. von O. Okun, G. Valentini und M. Re. Studies in Computational Intelligence. Springer, 2011, S. 75–95.

[C1/TB1/495] A. Schramm u. a. „Exon-level expression analyses identify MYCN and NTRK1 as major determinants of alternative exon usage and robustly predict primary neuroblastoma outcome“. In: *British Journal of Cancer* 107.8 (2012), S. 1409–1417.

[C1/496] A. Schramm u. a. „Next-generation RNA sequencing reveals differential expression of MYCN target genes and suggests the mTOR pathway as a promising therapy target in MYCN-amplified neuroblastoma“. In: *International Journal of Cancer* 132.3 (2013), S. 154–163.

[C1/497] J. H. Schulte u. a. „MYCN and ALKF1174L are sufficient to drive neuroblastoma development from neural crest progenitor cells“. In: *Oncogene* 32.8 (2013), S. 1059–1065.

Andere Veröffentlichungen

[C1/498] S. Lee und A. Schramm. *Preprocessing of Affymetrix Exon Expression Arrays*. Techn. Ber. 3. TU Dortmund, 2013.

3.4 Planung des Teilprojekts

Ziele

In der molekularen Medizin werden Hochdurchsatztechnologien zur simultanen Messung von zehntausenden bis hunderttausenden Expressionswerten und hunderttausenden Sequenzmutationen in internationalen Großprojekten immer häufiger eingesetzt. Diese Technologien, die als *next generation sequencing* (NGS) Technologien bezeichnet werden, erlauben zusätzlich zu Expressionswerten auch Mutationen im Genom und in den Gentranskripten der Tumorproben zu detektieren. Mit den bislang genutzten DNA-Microarrays konnten solche Sequenzvarianten nicht explizit erkannt werden, sehr wohl aber die Messergebnise verfälschen. Durch die basenpaar-genaue Bestimmung der Sequenz wird die Analyse sehr viel präziser. Das Ressourcenproblem besteht hierbei in der vergleichsweise geringen auswertbaren Fallzahl (Stichprobenzahl) von wenigen hundert Patienten gegenüber der sehr hohen potenziellen Merkmalsanzahl. Gegenüber der ersten Phase *erhöht* sich die zu betrachtende Merkmalsanzahl durch die Hinzunahme von Sequenzmutationen sogar noch bis in den Millionenbereich. Nach unseren bisherigen Erkenntnissen sind Informationen über Sequenzmutationen notwendig für das Verständnis der Zusammenhänge komplexer biologischer Systeme in fortschreitenden Tumorstadien. Das erste Ziel besteht also nach wie vor in der Auswahl relevanter Merkmale für die Risikoprognose. Technisch sind hier beim Merkmalsvergleich Ressourcenbeschränkungen von Laufzeit und Speicher zu beachten; die explizite Berechnung und Darstellung einer $p \times p$ Korrelationsmatrix zwischen Merkmalen muss zum Beispiel vermieden werden.

Da sowohl Ärzte als auch Patienten einer Diagnose aus einem abstrakten maschinellen Lernverfahren kritisch gegenüberstehen werden, ist zudem wichtig, dass die gewählten Merkmale, etwa bestimmte Genmutationen und Kombinationen von Genexpressionwerten, biologisch interpretierbar und erklärbar sind. Für die Entstehung eines Neuroblastoms gibt es vermutlich keine isolierte genetische Ursache. Interessant ist daher die Frage, ob es gemeinsam zugrunde liegende Ereignisse gibt, welche die Aggressivität eines Tumors bestimmen und damit mildere oder intensivere Behandlungen erforderlich machen. Wir untersuchen dazu die raum-zeitliche Plastizität des Tumorgenoms, indem wir entsprechende Proben in unterschiedlichen zeitlichen Stadien, vor der Behandlung und nach Behandlung in der Rückfallsituation miteinander vergleichen. Wir untersuchen auch die intratumorale Heterogenität mit Hilfe von Tumorproben aus unterschiedlichen Geweben desselben Patienten. Dabei setzen wir probabilistische graphische Modelle zur Merkmalsauswahl ein, die verschiedenartige Nebenbedingungen strukturell modellieren können. Statt irgendeiner Lösung mit

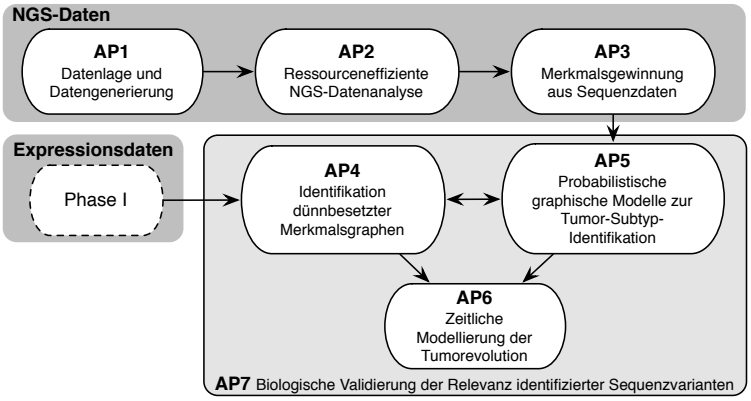


Abbildung 3.4.: Beziehungen zwischen den Arbeitspaketen in Phase II.

guter Prädiktion wird also eine Lösung mit zusätzlichen strukturellen Eigenschaften gefordert. Unsere Hypothese ist, dass diese Modellierung sowohl den Lösungsraum beschränkt als auch die Interpretierbarkeit der resultierenden Merkmale verbessert. Zusätzlich kann der Effekt einer Behandlung auf das Tumorgenom im Modell charakterisiert werden.

Durch die Hinzunahme der neuen kategoriellen Merkmale (Sequenzmutationen) wächst die Menge der Rohdaten deutlich: Statt einer numerischen Matrix von 300.000 Expressisonswerten pro Patient wie bei den Microarraydaten liegen in der zweiten Phase zunächst etwa 5 GB an Sequenzdaten vor, die zu den Mutationsmerkmalen verdichtet werden müssen. Ähnliche Fragestellungen werden zurzeit weltweit in sehr vielen Projekten untersucht; allein die deutsche Max-Planck-Gesellschaft hat mittlerweile mehr als ein Petabyte an Sequenzdaten produziert. Die entsprechenden Sequenzanalysen, bei denen es stets um die Bestimmung von Merkmalen wie Genexpressionwerten oder Mutationen aus Sequenz-Rohdaten geht, finden zurzeit meist auf großen Cluster-Architekturen oder auf Multicore-Servern statt und sind in jeder Hinsicht ressourcenhungrig. Dies betrifft sowohl die Anschaffungskosten der nötigen Hardware als auch den Verbrauch an CPU-Stunden und RAM-Stunden. Die nötigen Ressourcen zur Sequenzdatenanalyse können grundsätzlich über den Energieverbrauch der (exklusiv) damit beschäftigten Systeme bestimmt werden. Die bisher unbeantwortete Frage ist: Welcher Ressourceneinsatz ist für eine hinreichend genaue Bestimmung der Merkmale zur Risikoprognose denn überhaupt notwendig? Unser Ziel ist eine systematische Untersuchung des minimal notwendigen Energie- und Zeitbedarfs der Merkmalsgewinnung aus NGS-Daten, sowie ihre Optimierung mit Hilfe neuer Algorithmen im Zusammenspiel mit moderner Hardware wie GPUs und SSDs.

Arbeitsprogramm

Wir planen sieben Arbeitspakete (APs), deren Beziehungen in Abbildung 3.4 veranschaulicht werden. Dabei erforschen wir zwei verschiedene Typen von Hochdurchsatzdaten: Microarray-Expressionsdaten und NGS-Daten, wobei der Fokus in Phase II auf den NGS-Daten liegt. AP 1 widmet sich der Datenakquise der NGS-Daten, AP 2 der ressourceneffizienten DNA- und RNA-Sequenzanalyse und AP 3 der Merkmalsextraktion aus den NGS-Daten (Expressionswerte und Sequenzmutationen). AP 5 verwendet die Merkmale aus AP 3 als Eingabe, um darauf die Merkmalsauswahl für die Risikoprognose durchzuführen. AP 4 verwendet parallel dazu die Microarray-Daten aus Phase I zum Lernen struktureller Beziehungen zwischen den Merkmalen. Beide Datenquellen (Microarrays und NGS) können anhand der Ergebnisse von AP 4 und 5 verglichen werden. In AP 6 erweitern wir die Modelle zur Merkmalsauswahl auf Merkmale, die für die Tumorevolution über die Zeit wichtig sind. Die Ergebnisse der APs 4–6 werden in AP 7 durch Laborexperimente validiert.

Arbeitspaket 1. Datenlage und Datengenerierung: Für das Teilprojekt liegen 200 fertig sequenzierte Exome von Neuroblastomen sowie dem korrespondierenden Normalgewebe vor. Ferner wurden für 30 dieser 200 Fälle auch das Gesamt-Tumor-Genom (mit niedriger, i.e. 1–3-facher Abdeckung) sequenziert; das sind 3–9 Gbp Sequenzdaten pro Tumorgenom. Beide Sequenzierungen wurden durch die Deutsche Krebshilfe gefördert.

Wir werden unter Aspekten der Datenqualität und Sequenzabdeckung 120 Exome von Tumor/Normalpaaren auswählen. Nur solche Tumore, bei denen mindestens 90% des Exoms mit einer Sequenziertiefe von mindestens 20 repräsentiert ist und die mindestens 2.5 Gbp Gesamtsequenz aufweisen, werden in die Auswertung eingeschlossen. In der Interaktion mit AP2 werden wir ermitteln, wie man ressourcenbewusst auch die genomweiten Sequenzdaten geringer Tiefe hinzufügen kann. Eine Hauptaufgabe von AP1 besteht

somit darin, die Daten für eine ressourceneffiziente Analyse aufzubereiten und insbesondere Metadaten sowie klinische Parameter in maschinenlesbaren Formaten bereitzustellen. Zusätzlich zu den bislang verwendeten Excel-Tabellen kommt hierbei das strukturierte HDF5-Format zum Tragen.

Da alle bisherigen Studien darauf hindeuten, dass es keine isolierte genetische Ursache für die Entstehung eines Neuroblastoms gibt, wollen wir untersuchen, ob es gemeinsam zugrunde liegende Ereignisse gibt, welche die Aggressivität eines Tumors bestimmen und damit mildere oder intensivere Behandlungen erforderlich machen. Um uns dieser Frage anzunehmen, stehen uns neben dem Vergleich der Tumor-Normal-Proben zusätzlich Exom-Paare für 19 Patienten nach Erstdiagnose und in der Rückfallsituation zur Verfügung, um mit Metastasierung und aggressivem Tumorwachstum assoziierte genetische Vorgänge zu identifizieren. Für dieses kleinere Patientenkollektiv stehen auch korrespondierende RNA-seq- und Methylierungsdaten (Illumina 450k Array) zur Verfügung. Neben diesen Verlaufsproben, an denen wir die zeitliche Entwicklung des Tumorgenoms untersuchen möchten, gibt es von 11 anderen Patienten multiple Tumor-Proben des gleichen Entnahmezeitpunkts, mit deren Untersuchung wir ein besseres Bild der intratumoralen Heterogenität erhalten wollen. Mit Hilfe der in AP 6 beschriebenen Verfahren wollen wir so ein Modell der Neuroblastom-Entwicklung im Raum und über die Zeit entwickeln. Im Rahmen des APs werden auch die Daten dieser 19 bzw. 11 Patienten aufbereitet.

Wir erwarten, dass durch weitere zukünftige Konsortiums-Beteiligungen des Projektleiters Alexander Schramm während der Laufzeit des Projekts weitere Datensätze verfügbar werden. In diesem Fall werden diese, wann immer möglich, dem SFB zugänglich gemacht und die Metadaten in gleicher Weise aufbereitet. Durch ggf. andere laborseitige DNA- oder RNA-Aufbereitung müssen die resultierenden Sequenzen dann zunächst auf Vergleichbarkeit geprüft bzw. gesonderten Anpassungen unterworfen werden.

Arbeitspaket 2. Ressourceneffiziente NGS-Datenanalyse: Die APs 2 und 3 erzeugen gemeinsam aus NGS-Rohdaten hochdimensionale Merkmalsvektoren für die Prognosemodelle der folgenden APs. Das AP 2 widmet sich der ressourceneffizienten grundlegenden Analyse der Rohdaten.

Exom-Sequenzierung. Ein Exom-Datensatz umfasst ca. 50 Millionen DNA-Sequenzfragmente der Länge 100, die als FASTQ-Datei vorliegen. Jeder der 219 vorliegenden Exom-Datensätze (200 Tumor/Normal-Paare, 19 Normal/Tumor/Rezidiv-Trios) wird zunächst mit unserem Exomate-Analyse-Verfahren (siehe Vorarbeiten) bearbeitet, um frühzeitig erste Ergebnisse zu erhalten.

Jedes einzelne Fragment wird mit einem read mapper (typischerweise BWA in der Variante BWA-MEM oder bowtie2) im humanen Referenzgenom lokalisiert und mit der öffentlichen Referenzsequenz hg19 („Normalgenom“) verglichen. Zusätzlich werden wir den von uns neu entwickelten GPGPU-basierten read mapper PEANUT einsetzen, der den Index nicht auf dem Genom, sondern on the fly auf den reads erstellt (siehe Vorarbeiten). Wir werden systematisch untersuchen und vergleichen, welche Relationen zwischen dem Energiebedarf der Methoden (gemessen durch Leistungsaufnahme des Gesamtsystems während des read mapping Prozesses) und der Qualität der Ergebnisse zwischen den verschiedenen read mapping Strategien bestehen und hierzu die erste Evaluation erstellen.

Alle reads, die eine bestimmte Stelle im Referenzgenom überlappen, werden auf Sequenzebene mit dieser Stelle verglichen. Mit Hilfe statistischer Tests (z.B. Fisher's exact test auf *strand bias*) wird entschieden, ob eine Einzelnukleotidvariante, Insertion oder Deletion in Abgrenzung zu einem technisch bedingten Sequenzierfehler vorliegt. Hierbei werden die Qualitätswerte der Rohdaten berücksichtigt, und in mehreren Qualitäts-Rekalibrierungs-Schritten und Re-Alignierungsschritten die Entscheidung verfeinert. Dies wird durch Werkzeuge des *Genome Analysis Toolkit* (GATK) realisiert. Dieser Prozess ist sehr zeit- und rechenaufwändig. Wir haben vor, diese Prozedur durch eine ressourcenschonendere zu ersetzen; jedoch leidet darunter bisher entweder die Qualität der Alignments oder die Abdeckung (Anzahl betrachteter *reads*). Unsere Idee ist, eine nur leicht geringere Abdeckung in Kauf zu nehmen, indem die problematischen *reads*, in denen die betrachtete Position nahe an einem Ende liegt, ignoriert werden. Im Gegenzug entwickeln wir eine Art Algebra für Sequenzvarianten, um Evidenz aus verschiedenen Datensätzen für eine mögliche Mutation an dieser Stelle zu kombinieren. Abhängig von der Anzahl (hochzuverlässiger) *reads* lassen sich p-Werte für die Nullhypothese „keine Mutation” aus den absoluten Zahlen der *reads* in Vorwärts- und Rückwärtsrichtung, mit und ohne Mutation durch einfache algebraische Operationen gewinnen, wobei die relevanten Kennzahlen durch Analyse der gemappten *reads* aller Datensätze vorberechnet werden. Hiervon versprechen wir uns eine erhebliche Reduzierung des Ressourcenbedarfs der DNA-Exom-Analyse.

RNA-seq. Bei der RNA-Analyse der 19 vorliegenden RNA-seq-Datensätze genügt für die Bestimmung der Expressionswerte grundsätzlich eine ungefähre Lokalisierung; erst für die Bestimmung von Mutationen ist wiederum ein basenpaar-genauer Sequenzvergleich (Alignment) erforderlich. Die Standard-Tools trennen die Schritte der Lokalisierung und des Alignments nicht; unser *read mapper* PEANUT ist jedoch in der Lage, nur

die Lokalisierung durchzuführen, was erhebliche CPU- oder GPU-Ressourcen spart. Roh-Expressionswerte werden durch Zählen der *reads*, die auf jedes Exon *mappen*, berechnet. Diese werden in AP 3 mit Hilfe statistischer Modelle in normalisierte Expressionswerte umgerechnet.

Energieeffizienz vs. Laufzeit. Dank eines vom Mercator Research Center Ruhr (MERCUR) geförderten Projekts („UA Ruhr-Professur Computational Biology") stehen uns mehrere Möglichkeiten zum *read mapping* zur Verfügung, die verschiedene Hardware (multicore-CPU, GPU, SSD) unterschiedlich nutzen und verschiedene Laufzeiten und Energiebedarfe aufweisen. In Kombination mit den in diesem Arbeitspaket zu entwickelnden Vorberechnungsmethoden ergeben sich zahlreiche methodische Kombinationen für die NGS-Datenanalyse. Diese werden systematisch bezüglich ihrer Ressourcenbedarfe (Gesamtzeit, RAM-Bedarf, Externspeicherbedarf, Energieverbrauch) charakterisiert, um die Frage zu beantworten, welche Ressourcen minimal notwendig für eine zuverlässige Analyse sind.

Arbeitspaket 3. Merkmalsgewinnung aus Sequenzdaten: Dieses Arbeitspaket dient der Generierung von Merkmalsvektoren aus identifizierten Mutationen und Expressionswerten.

Merkmale aus Mutationen. Identifizierte DNA- und RNA-Mutationen werden als Sequenzvarianten in das von uns entwickelte Exomate-System importiert. Die Variantentabelle in einer Exomate-Instanz umfasst alle in mindestens einem Datensatz (aus beliebig vielen Experimenten) entdeckten Varianten. Durch Subtraktion zwischen Paaren von Tumor- und Normalgewebe-Varianten bleiben Tumor-spezifische Mutationen übrig, und wir können Schnitt- und Vereinigungsmengen über mehrere Patienten bilden.

„Interessante" (z.B. Tumor-spezifische) Varianten oder veränderte Gene können über (komplexe) SQL-Abfragen mit verschiedenen Filterparametern (zu Abdeckung, Variantenqualität, Variantenbewertung, u.a.) gefunden werden. Ein Webfrontend in Exomate erlaubt die automatisierte dynamische Generierung typischer SQL-Anfragen, wobei alle relevanten Filterparameter eingestellt werden können. Auf diese Weise werden „triviale" Varianten, etwa solche, die bekanntermaßen charakteristisch für eine Bevölkerungsgruppe sind, gar nicht erst betrachtet. Die resultierenden (dennoch zahlreichen) Varianten werden in AP 1 inspiziert und auf Plausibilität überprüft und ggf. herausgefiltert, wenn es sich um vermutete falsch positive Identifikationen handelt. Durch wiederholte Interaktion zwischen AP 1 und diesem AP wird das Exomate-System kontinuierlich verbessert.

Aus der Variantentabelle V und der Relation zwischen Datensätzen werden binäre Merkmalsvektoren der Länge $|V|$ für jeden Datensatz erstellt, die sehr dünn besetzt sind.

Merkmale aus Expressionswerten. Um vergleichbare Expressionswerte aus absoluten *read*-Zahlen zu erhalten, werden die unter „Stand der Forschung" genannten *overdispersed-Poisson*-Modelle von uns um Effekte, die durch RNA-Präparation entstehen, erweitert. Bisherige Arbeiten legen nahe, dass der lokale Sequenzkontext vor und hinter dem *read*-Startpunkt einen Einfluss auf die Wahrscheinlichkeit hat, dass ein *read* an genau dieser Stelle beginnt. Die Größe des relevanten Kontexts beträgt vermutlich nur wenige Nukleotide, so dass ein Modell zuverlässig aus jedem Experiment einzeln geschätzt werden kann. Experimente mit der gleichen RNA-Präparationsmethode sollten sehr ähnliche Modelle liefern; dies werden wir testen.

Ein Modell besteht dabei aus einem Präferenzvektor $\rho = (\rho_c)$, wobei c die verschiedenen Kontexte indiziert. Bei einer angenommenen Kontextlänge von 6 Nukleotiden gibt es $4^6 = 4096$ mögliche Kontexte. Für Kontext c gibt ρ_c an, mit welcher Wahrscheinlichkeit ein zufällig gezogener *read* mit Kontext c beginnt. Verschiedene Kontextgrößen werden miteinander verglichen und ein optimaler Trade-off zwischen Komplexität (Kontextlänge) und Effekt bestimmt.

Durch Berücksichtigung der Kontexte können Expressionswerte zwischen verschiedenen Exons bzw. Genen besser verglichen werden; die in dem Review [469] genannten Ansätze fokussieren vor allem auf differenzielle Expression, d.h. den Vergleich von Expressionswerten desselben Exons/Gens zwischen Experimenten. Der hier vorgeschlagene Ansatz ist ein Schritt zu absoluten Expressionswerten, die besser mit denen der Microarrays vergleichbar sein sollten. Solche Vergleiche werden wir zur Evaluierung mit den aus Phase I vorliegenden Microarray-Daten vornehmen.

Arbeitspaket 4. Identifikation dünnbesetzter Merkmalsgraphen: Dieses Arbeitspaket widmet sich der Identifikation kleiner Subgraphen des Graphen, deren Knoten die Merkmale darstellen und deren Kanten eine signifikante Beziehung zwischen den Merkmalen anzeigen. Graph LASSO [488] ist ein effizienter Algorithmus zur Identifizierung solcher Subgraphen. Für jeweils p-dimensionale Merkmalsvektoren von n Patienten, $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n$, die wir als Stichproben einer p-dimensionalen Gaussverteilung $\mathcal{N}(\mathbf{0}, \Sigma)$ auffassen, löst Graph

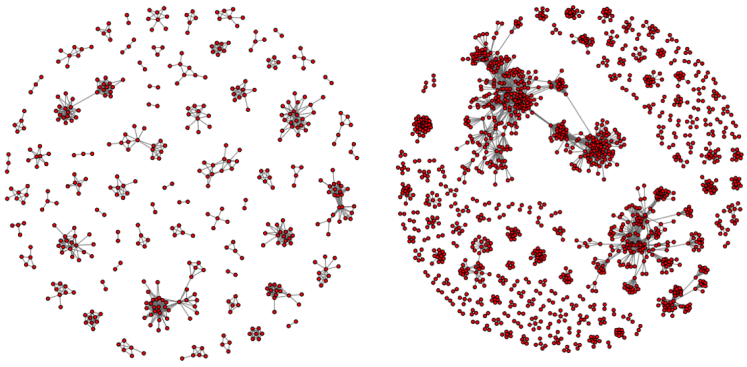


Abbildung 3.5.: **Dünnbesetzte Merkmalsgraphen, die durch Graph-LASSO auf Exon-Microarray-Daten identifiziert wurden (AP 4).** Der Grad der Dünnbesetztheit wird durch einen Parameter λ kontrolliert. Links: $\lambda = 7.177$; rechts: $\lambda = 5.741$.

LASSO das folgende konvexe Optimierungsproblem:

$$\Theta^* = \underset{\Theta \in \Re^{p \times p},\, \Theta \succeq \mathbf{0},\, \Theta = \Theta^T}{\operatorname{argmin}} - \log\det \Theta + \operatorname{tr}(S\Theta) + \lambda \|\Theta\|_1.$$

Dabei ist $S := \frac{1}{n} X X^T$ die Stichprobenkovarianzmatrix und X die $p \times n$ Matrix mit den n Merkmalsvektoren als Spalten. Die symmetrische positiv semidefinite Lösung Θ^* hat zwei Eigenschaften: Sie ist dünnbesetzt (d.h. nur wenige Einträge sind von Null verschieden, wobei der Grad der Dünnbesetztheit durch den Parameter λ des ℓ_1-Regularisierungsterms bestimmt wird), und es ist $\Theta^* \approx \Sigma^{-1}$. Von Null verschiedene Nichtdiagonal-Einträge von Σ^{-1} kennzeichnen Merkmale, die bedingt unabhängig sind, d.h., wenn $\Sigma^{-1}_{ij} \neq 0$ für $i \neq j$, dann sind die durch i und j indizierten Merkmale unabhängig, gegeben alle anderen Merkmale. Daher wird durch Θ^* ein dünnbesetzter Graph definiert, der die bedingte Unabhängigkeit zwischen Merkmalen approximiert, wenn man eine Kante zwischen Merkmalen i und j zieht, wann immer $\Theta^*_{ij} \neq 0$ für alle $i, j \in \{1, 2, \ldots, p\}$, $i \neq j$. Abbildung 3.5 zeigt zwei Beispiele mit verschiedenen Werten für λ.

Obwohl Optimierungsalgorithmen für das konvexe Graph-LASSO Problem existieren, bleiben Herausforderungen zu lösen. Erstens: Die existierenden Algorithmen skalieren nicht gut mit der Anzahl p der Merkmale (die bei uns mehrere hunderttausende sind); problematisch ist etwa die explizite Speicherung der aktuellen Lösung Θ der Größe $p \times p$. Daher ist die Entwicklung speichereffizienter Algorithmen notwendig, etwa durch implizite Repräsentation von Θ. Zweitens: Wir möchten mittels statistischer Tests diejenigen Subgraphen auswählen, die für das klinische Resultat (z.B. die Überlebenszeit) relevant sind. Drittens: Wir benötigen eine systematische Methode, um den Parameter λ, der die Dünnbesetztheit des Graphen kontrolliert, optimal zu wählen. Graphentheoretische Charakteristika wie die Verteilung der Knotengrade oder die Verteilung der Längen kürzester Pfade zwischen den Knoten können benutzt werden, um den Wert von λ zu bestimmen; hierfür ist eine enge Zusammenarbeit mit TP A6 vorgesehen.

Wir werden auch Erweiterungen des Graph-LASSO-Problems betrachten, indem wir Gene, die zu einer Gruppe gehören, verstärkt zusammengruppieren. Eine Gruppe kann dabei etwa durch einen gemeinsamen Gene Ontology (GO) Term definiert sein.

Arbeitspaket 5. Probabilistische graphische Modelle zur Tumor-Subtyp-Identifikation: Die Sequenzdaten wurden durch die APs 2 und 3 in eine Liste von binären Indikatorvariablen umgewandelt, die Präsenz oder Absenz von jeweils einer spezifischen Mutation (Einzelnukleotid-Varianten, kurzen Insertionen und Deletionen von Sequenzabschnitten) anzeigen. In diesem AP sollen zusammengehörende Gruppen solcher genomischen Profile identifiziert werden, die verschiedenen Subtypen des Neuroblastoms entsprechen; jeder Subtyp wird dabei durch eine (kleine) Teilmenge von Mutations-Merkmalen charakterisiert.

Wir werden die Subtypen mit *latent Dirichlet allocation* (LDA) modellieren [7]. LDA ist ein probabilistisches graphisches Modell, das ursprünglich vorgeschlagen wurde, um Themen von Dokumenten zu identifizieren, und das seitdem viele Anwendungen erfahren hat. Sei ein Patient durch $\mathbf{x} = (x_1, x_2, \ldots, x_{|V|})$ gegeben, wobei V das Universum der möglichen Varianten darstellt; ein Patient entspricht also einem sehr dünn besetzten Vektor $\mathbf{x}$ (einige hundert Eins-Einträge von mehreren hunderttausenden). Eine Kohorte von N Patienten ist dann durch $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ gegeben.

LDA ist ein generatives Modell, in dem Patienten als zufällige Mischungen (Konvexkombinationen) latenter Subtypen modelliert werden. Jeder Subtyp entspricht dabei einer Wahrscheinlichkeitsverteilung über

Mutationen. Im einfachsten Modell wird angenommen, dass die Anzahl K der Subtypen bekannt und fest ist. Das generative Modell hinter LDA erzeugt dann ein Patientenprofil $\mathbf{x}$ wie folgt:

1. Ziehe den K-dimensionalen Parametervektor θ (entspricht der Subtyp-Mischung) aus einer Dirichlet-Verteilung mit dem K-dimensionalen Parametervektor α (ein Hyperparameter, der die Breite der Verteilung der Mischungskoeffizienten steuert)

2. Für jede der $|V|$ möglichen Sequenzvarianten (Mutationen):
   a) Ziehe einen Subtyp $(1, \ldots, K)$ aus der K-Dimensionalen Multinomialverteilung mit Parameter θ
   b) Ziehe die Indikatorvariable der Variante aus einer Binomialverteilung mit Erfolgswahrscheinlichkeit $\beta_{z,i}$. Dabei ist β also eine $K \times |V|$-Matrix, die für jeden Subtyp und jedes Merkmal angibt, mit welcher Wahrscheinlichkeit das Merkmal in dem Subtyp auftritt.

Das entsprechende probabilistische graphische Modell ist in Abbildung 3.6(A) veranschaulicht. Unter diesem Modell ist die Wahrscheinlichkeit, eine Patientenkohorte $\mathcal{X}$ der Größe N zu generieren, gegeben durch

$$\mathrm{Prob}(\mathcal{X}|\alpha,\beta) = \prod_{n=1}^{N} \int \mathrm{Prob}(\theta_n \mid \alpha) \left(\prod_{i=1}^{|V|} \sum_{z=1}^{K} \mathrm{Prob}(z \mid \theta_n)\, \mathrm{Prob}(x_{n,i}|\beta_{z,i}) \right) \mathrm{d}\theta_n.$$

Die Hyperparameter α und β werden nach dem Maximum-Likelihood-Prinzip geschätzt, entweder über Sampling-Ansätze oder Variations-Optimierung. Die geschätzten Parameter, insbesondere $\beta = (\beta_{z,i})$, dienen der Merkmalsauswahl wie folgt: Merkmale, die eine hohe Wahrscheinlichkeit in einem Subtyp aufweisen, sind charakteristisch für den Subtyp. Des Weiteren können für jeden Subtyp über die Patienten (und ihre geschätzten Subtyp-Mischungskoeffizienten) durch die Korrelation mit klinischen Variablen passende Schwellenwerte identifiziert werden. Wir werden LDA so erweitern, dass wir auch klinische Variablen berücksichtigen. Auf diese Weise können wir neue Beiträge zur Merkmalsauswahl für die Subtypauswahl und Intra-Tumor-Heterogenität liefern.

Arbeitspaket 6. Zeitliche Modellierung der Tumorevolution: Die Flexibilität graphischer Modelle erlaubt die Erfassung struktureller Änderungen im Merkmalsraum anhand von Beobachtungen. Dies ist wichtig für die Modellierung der Tumorevolution im zeitlichen Verlauf bei Neuroblastom-Patienten. Wir untersuchen zwei Modelle, um Veränderungen in (1) der bedingten Unabhängigkeitsstruktur der Merkmale und (2) der Subtyp-Struktur zu identifizieren. Dazu verwenden wir Daten von zwei Zeitpunkten: vor und nach der initialen Behandlung. Damit werden die in APs 4 und 5 vorgeschlagenen Modelle um eine zeitliche Dynamik erweitert.

Eine raum-zeitliche Erweiterung des (räumlichen) Graph-LASSO (AP 4) erhalten wir beispielsweise dadurch, dass wir den Knoten jedes Merkmals für jeden Zeitpunkt replizieren. (Die Annahme dabei ist, dass

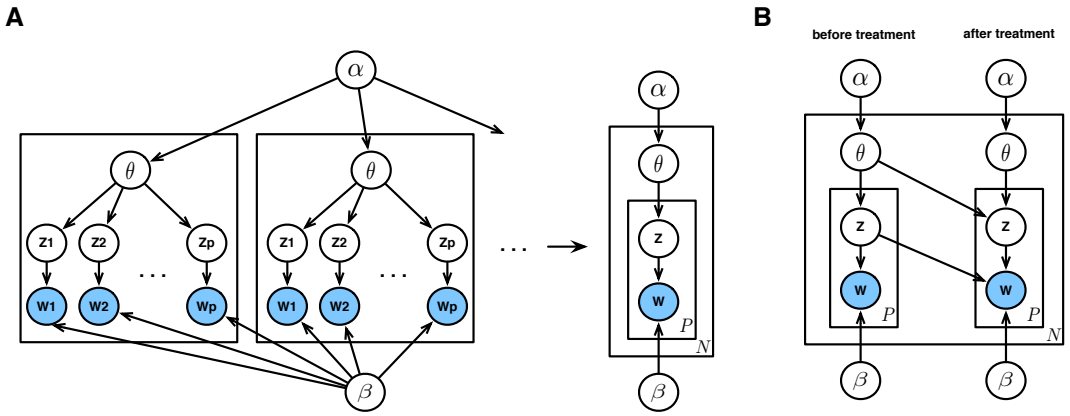


Abbildung 3.6.: **Latent Dirichlet Allocation (LDA) Modelle.** (A): Reduktion der Darstellung eines probabilistischen graphischen LDA-Modells zu einer flachen Darstellung (zu AP 5). (B): Erweiterung von LDA, um zeitliche Veränderungen, etwa zwischen den Zeitpunkten vor und nach einer Behandlung eines Patienten, zu modellieren.

jedes Merkmal zu jedem Zeitpunkt auch gemessen wurde.) Das Optimierungsproblem kann dann so formuliert werden:

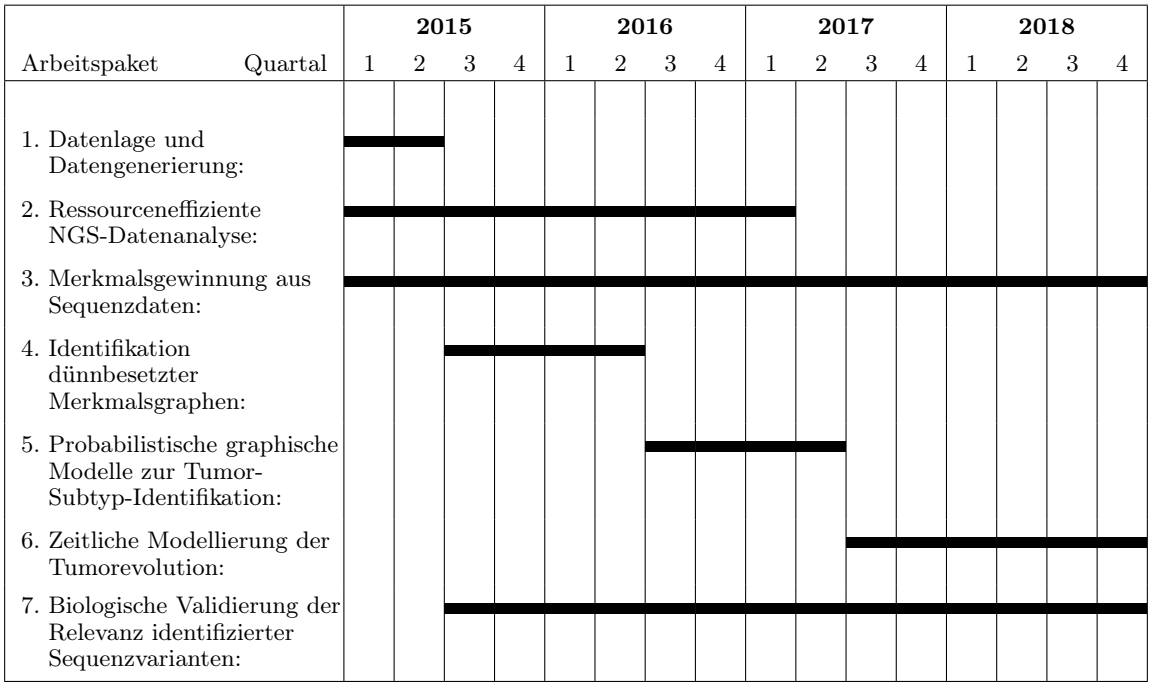
$$\Theta^* = \underset{\Theta \in \Re^{2p \times 2p}, \Theta \succ \mathbf{0}, \Theta = \Theta^T}{\operatorname{argmin}} - \log\det\Theta + \operatorname{tr}(S\Theta) + \lambda_1 \|\Theta\|_1 + \lambda_2 \|\Theta_{1:p,1:p} - \Theta_{(p+1):2p,(p+1):2p}\|_{\mathrm{F}}^2,$$

wobei $\|\Theta\|_{\mathrm{F}}^2 = \sum_{ij} \Theta_{ij}^2$ das Quadrat der Frobeniusnorm von Θ ist. Dieser zweite Regularisierungsterm in der Zielfunktion, der durch $\lambda_2 > 0$ parametrisiert wird, bestraft große Unterschiede zwischen Merkmalsgraphen vor und nach der Behandlung. Ähnliche Ansätze zur Regularisierung wurden auch im Rahmen von *fused LASSO* [486] verwendet. Für diskrete Merkmale können ähnliche Ansätze betrachtet werden, basierend beispielsweise auf einer Zussamenarbeit mit Teilprojekt A1 [A1/C1/53].

Zeitliche Erweiterungen von LDA (AP 5) können genutzt werden, um Hypothesen über Veränderungen in Subtypen, Subtyp-Mischungen in Patienten und Merkmalsverteilungen in Subtypen zu formulieren. Solche Hypothesen entsprechen dem Hinzufügen von Kanten zwischen zwei Instanzen eines probabilistischen graphischen LDA-Modells, wie in Abbildung 3.6(B) illustriert wird. Dazu muss das Inferenzverfahren für LDA angepasst werden.

Arbeitspaket 7. Biologische Validierung der Relevanz identifizierter Sequenzvarianten: In der Arbeitsgruppe des Antragstellers Alexander Schramm sind gut charakterisierte Zellmodelle des Neuroblastoms zur experimentellen Überprüfung funktioneller Relevanz der in AP 5 und 6 identifizierten Sequenzvarianten etabliert. Zur Validierung einer biologischen Bedeutung von rezidiv- oder verlaufsspezifischen Mutationen werden diese in den Zellkultursystemen mittels CRISPR/Cas9-vermitteltem genomischen Editing eingebracht. CRISPR/Cas ist ein ursprünglich von Bakterien entwickeltes System zur Erkennung und zur Abwehr von Fremd-DNA. Mit dem für humane Zellen seit kurzem verfügbaren modifizierten CRISPR/Cas9-System ist es möglich, Einzelbasenaustausche gezielt und ohne DNA-Vektor-basierte Transfektionen in Zellen einzubringen [481, 487]. Erster Schritt dieses APs ist es daher, das CRISPR/Cas9 System auf kultivierte Neuroblastomzellen zu adaptieren. Wir wollen dieses System dann nutzen, um die funktionelle Bedeutung von rezidiv- und verlaufsspezifischen Varianten in den kultivierten Neuroblastomzellen zu überprüfen. Für diese Zelllinien, die bereits aus Neuroblastomtumoren etabliert sind, liegen uns die Daten bezüglich prä-existierender genomischer Veränderungen bereits vor, da hierzu Exom-Analysen (vgl. AP 1) durchgeführt wurden. Der Hauptteil der Arbeit in diesem AP wird also darin bestehen, gezielt den Einfluss von Rezidiv-oder Verlaufsspezifischen Mutationen in der Zellkultur zu prüfen, indem wir Parameter wie das Zellwachstum in 2D und 3D, das Zellwanderungspotenzial und die Sensibilität gegenüber Tumortherapie in Abhängigkeit der CRISPR/Cas9 vermittelten Mutationen testen. Dies wird nicht nur zur Verfeinerung und Validierung der in AP 5 und AP 6 generierten Modelle führen, sondern auch neue Einblicke in die Modulation der Tumorbiologie und ggf. des Therapieansprechens durch die hier identifizierten Sequenzvarianten liefern.

Zeitplan



3.5 Stellung innerhalb des Sonderforschungsbereichs

Die molekularbiologischen Daten, die in Teilprojekt C1 untersucht und zur individuellen Risikoprognose herangezogen werden, sind extrem hochdimensional, bis in den Millionenbereich. Bei dem Ziel einer robusten und von Menschen interpretierbaren Merkmalsauswahl treten die folgenden Ressourcenbeschränkungen zutage:

- Die Anzahl der verfügbaren unabhängigen Stichproben ist begrenzt auf maximal wenige hundert.
- Die Interpretierbarkeit und Vermittelbarkeit (im Sinne einer durch Ärzte erklärbaren Diagnose) muss gegeben sein; dies ist eine „weiche“ Ressourcenbeschränkung.
- Die Darstellung der Daten und Lösungen der in APs 4–6 genannten Optimierungsprobleme erfordert naiv $O(p^2)$ Speicher (konkret ca. 10^{12} Worte); auch auf großzügig bemessenen Systemen ist dieser aber begrenzt.
- Die Rohdaten aus der DNA- und RNA-Sequenzierung sind sehr umfangreich; es ist bisher offen, mit welchen minimalen Ressourcen hieraus Merkmale generiert werden können. Wir betrachten insbesondere den Gesamtaufwand, gemessen am Energieverbrauch, im Vergleich zur benötigten Zeit.

Um die hohe Dimension im Vergleich zur Stichprobenzahl zu berücksichtigen und relevante Merkmale auszuwählen, haben wir bereits in der ersten Phase Ensemble-Methoden verwendet. Diese erlauben eine effiziente Parallelisierung. Bei der Erforschung praktikabler Methoden zur Parallelisierung und zur Verteilung der Subprobleme haben wir mit den Projektleitern Morik, Rahmann und Rahnenführer zusammengearbeitet; dies werden wir weiter vertiefen. Eine weitere Zusammenarbeit mit Teilprojekt A3 ist bei Methoden zur Modellwahl geplant, insbesondere in den APs 4–6. Zur Variablenselektion gibt es aus der ersten Phase bereits (eingereichte) gemeinsame Vorarbeiten mit TP A3 (Lee, Rahnenführer), sowie die Zusammenarbeit mit Teilprojekt C3 zur schnellen Variablenselektion gemäß maximaler Relevanz und minimaler Redundanz. Weitere Analysen der Methoden und die beweisbare Abschätzung der Güte unserer Methoden erfolgt in Zusammenarbeit mit Teilprojekt C4.

In AP 4 muss ein Parameter λ, der die Dünnbesetztheit des resultierenden strukturellen Gen- bzw. Exonmodells kontrolliert, geschätzt werden. Um eine geeignete Wahl zu treffen, können Eigenschaften des resultierenden Graphen, wie die Verteilung der Knotengrade oder typische Längen kürzester Pfade, herangezogen werden. Diese Kennzahlen können auch dazu dienen, verschiedene Untergruppen von Patienten zu vergleichen. Methodisch (Berechnung der Kennzahlen in großen Graphen, Testen von Subgraph-Isomorphie) besteht eine enge Verwandtschaft zu den Methoden in Teilprojekt A6, so dass hier eine enge Kooperation und ein Austausch von Methoden geplant ist.

Bei der Konzeption probabilistischer graphischer Modelle, wie sie hier in APs 4–6 geplant ist, gibt es bereits aus der ersten Phase eine Zusammenarbeit mit Teilprojekt A1, die in einer gemeinsamen Publikation zu *spatio-temporal random fields* (STRFs) resultierte; diese Zusammenarbeit soll in Bezug auf die Skalierbarkeit und Energieffizienz der in C1 zu entwickelnden Modelle fortgesetzt werden.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

UA Ruhr-Professur Computational Biology
(Rahmann,Förderkennzeichen MERCUR-Pe-2013-12) (Förderzeitraum: 2014 - 2016)

Im Rahmen des vom Mercator Research Center Ruhr (MERCUR) geförderten Projekts wird die Professur von Herrn Rahmann zu einer UA Ruhr-Professur erweitert, so dass er zugleich den Lehrstuhl für Genominformatik an der Universität Duisburg-Essen und eine Arbeitsgruppe Bioinformatik an der TU Dortmund leitet. In der Förderung sind auch Personalmittel zum ressourceneffizienten *read mapping* vorgesehen. Hierbei werden Verfahren erforscht, die aktuelle Hardware, insbesondere GPUs und SSDs optimal für NGS-Datenanalysen ausnutzen sollen. Die Resultate des MERCUR-Projekts werden dem Sonderforschungsbereich zur Verfügung gestellt und in das TP C1 integriert. Im MERCUR-Projekt werden jedoch keine weiterführenden Analysen wie das Detektieren von Sequenzmutationen oder die Bewertung des Effekts einzelner Mutationen oder die Generierung von Merkmalen aus Mutationsdaten erforscht. Die Forschungsfragen von TP C1 gehen also deutlich über die im MERCUR-Projekt zu leistende Basisarbeit hinaus.

Flexibel bis es bricht: Dynamik von Genen & Merkmalen, Dichte & Diversität in Gesellschaften unter sich veränderndern Umweltbedingungen
(Rahmann,Förderkennzeichen DFG RA 1898/1-1) (Förderzeitraum: 2014 - 2017)

Dieses Projekt, das Teil des SPP 1704 (Flexibility matters: Interplay between trait diversity and ecological dynamics using aquatic communities as model systems – DynaTrait) ist, muss noch vom DFG Hauptausschuss bewilligt werden, wurde aber bereits positiv begutachtet. Hierbei sollen metagenomische und metatranskriptomische Analysen durchgeführt werden, wobei Organismen in Ufernähe von Gewässern und aktive Genfunktionen in der entsprechenden Lebensgemeinschaft identifiziert werden sollen. In diesem Projekt werden auch DNA- und RNA-Sequenzanalysemethoden verwendet. Ressourcenbeschränkungen spielen jedoch keine Rolle, und die Fragestellungen hinter das Analyse sind grundlegend anders. Ziel hierbei ist, numerische Daten für die Modellierung des Ökosystems mit Hilfe eines Differenzialgleichungssystems bereitzustellen.

Data Driven Materials Design
(Rahmann,Förderkennzeichen Mercator Research Center Ruhr – MERCUR) (Förderzeitraum: 2013 - 2015)

In diesem Projekt sollen Motiverkennungsalgorithmen aus der Bioinformatik auf das Problem des Designs neuer Materialien (Legierungen) übertragen werden. Aus gemessenen Materialeigenschaften sollen nach Diskretisierung Muster bestimmt werden, die andere Materialeigenschaften voraussagen. Es gibt keine inhaltliche Nähe zu Teilprojekt C1.

Mouse modeling and cross-species analysis to optimize identification and targeting of MYCN co-drivers in neuroblastoma - High Throughput data analysis and integration with data from human tumors
(Rahmann,Förderkennzeichen BMBF eMed) (Förderzeitraum: 2014 - 2018)

Dies ist ein Teilprojekt des BMBF-Verbunds „Systems Medicine for Neuroblastoma“ (SYSMED-NB, Koordinatorin Prof. Angelika Eggert, Charité Universitätsmedizin, Berlin), in dem die AG Rahmann eine Produktiv-Pipeline zur Bearbeitung von Exom- und RNA-seq-Daten aus Maustumoren etablieren und im Rahmen einer Dienstleistung für die anderen Verbundpartner anbieten wird. Mittelfristig werden Forschungsergebnisse des SFB-Teilprojekts C1 in das BMBF-Konsortium eingehen, um die Pipeline zu verbessern. Eine Forschung an neuen Bioinformatik-Methoden findet in diesem Konsortium nicht statt.

Role of the tyrosine kinase TrkA and TrkB in checkpoint activation and DSB repair
(Schramm,Förderkennzeichen DFG GRK 1739) (Förderzeitraum: 2012 - 2015)

Dieses Projekt ist im Rahmen des Graduiertenkollegs 1739 „Molekulare Determinanten der zellulären Strahlenantwort und ihre Bedeutung für die Modulation der Strahlensensitivität“ eingeordnet, bei dem wir Strahlenempfindlichkeit von Tumorzellen in Abhängigkeit von Rezeptor-Expression messen. Eine inhaltliche Nähe zum Projekt C1 besteht nicht.

Role of tumor- and host-derived Galectin-1 in neuroblastoma
(Schramm,Förderkennzeichen Deutsche Krebshilfe) (Förderzeitraum: 2012 - 2015)

In diesem Projekt untersuchen wir an transgenen Mausmodellen die Bedeutung eines möglichen Therapieangriffsziels, Galectin-1 für die Erhaltung des Tumorphänotyps. Inhaltlich ergibt sich keine Überschneidung zum TP C1.

Modelling primary tumor metabolism in neuroblastoma to identify central nodes for therapeutic intervention
(Schramm,Förderkennzeichen BMBF eMed) (Förderzeitraum: 2014 - 2018)

Dies ist ein Teilprojekt des BMBF-Verbunds „Systems Medicine for Neuroblastoma“ (Koordinatorin Prof. Angelika Eggert, Charité Universitätsmedizin, Berlin), in dem die AG Schramm zusammen mit Biochemikern und Systembiologen enzymatische Messungen an Tumorzellen durchführt, um therapeutische Interventionspunkte für die gerichtete Tumorbehandlung zu ermitteln. Eine inhaltliche Überschneidung zum Projekt C1 besteht nicht.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	117.600	—	117.600	—	117.600	—	117.600
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Geräte, Verbrauchsmaterial:	20.000		20.000		20.000		20.000	
Summe:	20.000		20.000		20.000		20.000	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	137.600		137.600		137.600		137.600	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (C1)							
Wissen-schaftliches Personal	1	Rahmann, Sven, Prof. Dr., Prof	Bioinformatik	Lehrstuhl für Genominformatik, Fak. Medizin, Uni DUE	10,0	—	Grund-ausstattung
	2	Schramm, Alexander, PD Dr., Grp.-Leiter	Molekularmedizin	Onkologisches Labor, Kinderklinik, Klinikum Essen	9,95	—	Grund-ausstattung
	3	Lee, Sangkyun, Dr., PostDoc	Maschinelles Lernen	LS 8, Fak. Inf, TU Do	29,85	—	Grund-ausstattung
	4	Timm, Henning, Dipl.-Inf. WiMi	Bioinformatik	Lehrstuhl für Genominformatik, Fak. Medizin, Uni DUE	39,8	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	5	Wolf, Claudia, Verw. Ang.	—	Lehrstuhl für Genominformatik, Fak. Medizin, Uni DUE	9,95	—	Grund-ausstattung
	6	Mahlow, Ellen, Techn. Ang.	—	Onkologisches Labor, Kinderklinik, Klinikum Essen	19,9	—	Grund-ausstattung
Beantragte Personalmittel (C1)							
Wissen-schaftliches Personal	7	N.N. , WiMi	Molekularmedizin	Onkologisches Labor, Kinderklinik, Klinikum Essen	—	Doktorand/in	—
	8	N.N. , WiMi	Maschinelles Lernen	LS 8, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Rahmann, Sven**
   Projektleitung. Schwerpunkt Bioinformatik. Mitarbeit in den Arbeitspaketen 1, 2, 3, vor allem 1. Koordinierung der Arbeiten im Teilprojekt mit denen des MERCUR-Projekts.

2. **Schramm, Alexander**
   Projektleitung. Schwerpunkt Molekularmedizin. Mitarbeit in den Arbeitspaketen 1, 7.

3. **Lee, Sangkyun**
   Projektleitung. Schwerpunkt Data Mining. Mitarbeit in den Arbeitspaketen 4,5,6.

4. **Timm, Henning**
   Bearbeitung der Arbeitspakete 2 und 3.

5. **Wolf, Claudia**
   Unterstützung des PL Rahmann bei Verwaltungsaufgaben.

6. **Mahlow, Ellen**
   AP 7, technische Unterstützung bei der laborseitigen Validierung der Befunde.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

7. **N.N.,**
   AP7, insbesondere die Etablierung des CRISPR/Cas-Systems in Tumorzellen und die Laborseitige Validierung der Befunde zu potentiell relevanten Mutationen.

8. **N.N.,**
   AP 4,5,6, Entwicklung effizienter Algorithmen und ihre Evaluation für verschiedene biologische Aufgabenstellung.

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
Uni DUE: vorhandene Mittel aus Haushalt	5.000	5.000	5.000	5.000
Uniklinikum Essen: vorhandene Mittel aus Haushalt	6.000	6.000	6.000	6.000
Summe der vorhandenen Mittel	11.000	11.000	11.000	11.000
beantragte Ergänzungsausstattung	20.000	20.000	20.000	20.000

Geräte bis 10.000 EUR, Software und Verbrauchsmaterial **für 2015 – 2018**

Verbrauchsmaterial Zellkultur, PCR, Antikörper, Phänotypisierungsassays (Durchflusszytometrie, ELISAs), EUR je Jahr 20.000
benötigt für Datenerhebung und technische Validierung

3.1 Allgemeine Angaben zum Teilprojekt C3

3.1.1 Titel:

Mehrstufige statistische Analyse von hochfrequenten räumlich-zeitlichen Prozessdaten

3.1.2 Fachklassifizierung:

Data Mining/Maschinelles Lernen, Astroteilchenphysik

3.1.3 Projektleitung

Prof. Dr. Morik, Katharina, 14.10.1954

LS 8, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 12
44227 Dortmund

Tel.: 0231-755-5100
Fax : 0231-755-5105
E-Mail: `katharina.morik@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Dr. Rhode, Wolfgang, 17.10.1961

Experimentelle Physik E5, Fakultät Physik, Technische Universität Dortmund
Otto-Hahn-Straße 4
44227 Dortmund

Tel.: 0231-755-3550
Fax : 0231-755-4547
E-Mail: `wolfgang.rhode@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Dr. Ruhe, Tim, 08.09.1981

Experimentelle Physik E5, Fakultät Physik, Technische Universität Dortmund
Otto-Hahn-Straße 4
44227 Dortmund

Tel.: 0231-755-6166
Fax : 0231-755-4547
E-Mail: `tim.ruhe@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

( ) nein (x) ja, befristet bis zum 30.09.2016, eine weitere Beschäftigung ist vorgesehen bis zum 30.09.2019

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Die aufsehenerregende Entdeckung von Neutrinos aus astrophysikalischen Quellen [527] durch das IceCube Neutrinoteleskop [508] eröffnet ein völlig neues physikalisches Teilgebiet an der Schnittstelle von Astroteilchenphysik und Astronomie: die Neutrinoastronomie. Eine detaillierte Analyse dieser extraterrestrischen Neutrinos ermöglicht umfassende neue Erkenntnisse astrophysikalischer Phänomene wie Aktiver Galaktischer Kerne (AGN) [514] oder Gamma Ray Bursts (GRB) [531]. Diese Erkenntnisse sind insbesondere dann von Bedeutung, wenn sie mit den Daten von Gamma-Teleskopen, wie MAGIC [510] oder FACT [511], kombiniert und verglichen werden können. Die Beschleunigung der kosmischen Strahlung ist ein seit über 100 Jahren ungeklärtes physikalisches Phänomen (siehe z.B. [524] und die darin enthaltenen Referenzen). Eine Kombination von Erkenntnissen der Gamma- und der Neutrinoastronomie ermöglicht unter Umständen bahnbrechende Erkenntnisse in diesem Bereich. Dies erfordert, sowohl in der Gamma- als auch in der Neutrinoastronomie, eine hochpräzise Analyse der Daten.

In der Astroteilchenphysik erfolgt eine Analyse der hochfrequenten und hochdimensionalen Teleskopdaten in mehreren Arbeitsschritten. Den ersten Schritt bildet dabei die Auswahl interessanter Ereignisse für die Weiterverarbeitung in Echtzeit. In einem zweiten Schritt müssen dann die immer noch sehr hochfrequenten aber physikalisch uninteressanten Untergrundereignisse verworfen werden [509, C3/A1/540]. Die Rekonstruktion interessanter Teilcheneigenschaften, wie zum Beispiel des Energiespektrums, durch Entfaltung erfolgt in einem dritten Schritt [C3/546]. Alle drei Arbeitsschritte erfolgen unter Zuhilfenahme von simulierten Ereignissen, welche mit größtmöglicher Genauigkeit und in großer Zahl erzeugt werden müssen [500, 530]. Ressourcenbeschränkungen ergeben sich dabei einerseits durch eine Begrenzung an Speicherplatz und CPU-Zeit, welche ein Verwerfen von Ereignissen notwendig machen, und andererseits durch die abgeschiedene Lage der Teleskope (Südpol, La Palma), woraus sich wiederum eine begrenzte Möglichkeit zur Kommunikation, zum Beispiel durch Satelliten, ergibt.

Sämtliche der oben genannten Arbeitsschritte wurden innerhalb der ersten Förderperiode auf ihre Ressourcenbeschränkung untersucht. Der Fokus lag dabei vor allem auf der Entwicklung robuster und generalisierbarer Techniken und Methoden. Dabei konnte exemplarisch gezeigt werden, dass sich durch die konsequente Nutzung von Methoden aus dem Bereich des maschinellen Lernens signifikante Verbesserungen im Bereich der Datenanalyse erzielen lassen [C3/538]. Darüber hinaus konnte gezeigt werden, dass sich durch den Einsatz von GPU-Rechnungen in der Produktion simulierter Ereignisse ein Geschwindigkeitsgewinn von einem Faktor 10 bei einer gleichzeitig um einen Faktor 10 verringerten Leistungsaufnahme erreichen lässt (siehe Masterarbeit Tomasz Fuchs). Weitere signifikante Verbesserungen wurden durch den Einsatz von Streaming-Algorithmen erreicht [A1/517].

Das Potential der in der ersten Antragsphase erreichten Verbesserungen soll in der zweiten Antragsphase voll ausgeschöpft und wenn möglich konsequent ausgebaut werden. Die am Beispiel des IceCube-Teleskops durchgeführte und äußerst erfolgreiche Klassifikation astrophysikalischer Ereignisse [C3/A1/540] soll dabei auf die Gammaastronomie übertragen werden. Durch den konsequenten Einsatz von Streaming-Algorithmen soll die Ressourcenbeschränkung durch limitierte Satellitenkommunikation und begrenzten Speicherplatz am Ort der Experimente weitgehend aufgehoben werden. Der Propagations-Algorithmus PROPOSAL [C3/536]

soll vollständig auf GPUs übertragen werden, um simulierte Ereignisse effizient und ressourcenschonend zu produzieren. Der auf maschinellem Lernen basierende Entfaltungsalgorithmus DSEA/SpectEst [C3/542] soll in der zweiten Phase erstmals auf experimentelle Daten der verschiedenen Teleskope angewandt werden.

3.3 Bisherige Entwicklung des Teilprojekts

3.3.1 Bericht und Stand der Forschung

Datenaufzeichnung (AP 1): In der Astroteilchenphysik werden Signalsequenzen in einigen tausend Sensoren mit sehr hohen Sampling-Raten (2 GHz) registriert [510, 528]. Aufgrund der daraus resultierenden großen Datenmengen und der begrenzten Speicherkapazitäten müssen noch während des Messprozesses geeignete Regeln (Trigger) zur Selektion möglicherweise relevanter Signalereignisse angewandt werden [512, 507]. Für das FACT-Teleskop hat die Arbeitsgruppe Astroteilchenphysik vollen Zugriff auf die Datenausleseelektronik der Kamera. Der Trigger, der zu einer Speicherung der Daten führt, wurde über eine FPGA-Programmierung im Vorlauf des SFB 876 in das Datennahmesystem integriert (FACT, technische Veröffentlichung). Da die optimalen Triggerbedingungen auch eine Funktion der nachfolgenden Selektions- und Analysealgorithmen sind, wurden diese in der ersten Phase des SFB 876 noch nicht variiert. Jedoch wurde die zeitliche und räumliche Ausbreitung des Signals in der Kamera (Bachelorarbeit Fligg), das elektronische Rauschen der einzelnen Sensoren sowie das ungetriggerte bzw. zufällig getriggerte Signal der gesamten Kamera intensiv studiert (siehe Diplomarbeit Jens Buss). Ebenso wurde die Homogenität der räumlichen Signalverteilung einer eingehenden Untersuchung unterzogen (Dissertation in Vorbereitung, Dominik Neise). Alle diese Punkte sind Grundvoraussetzungen für eine erfolgreiche Optimierung des Triggers.

Die Analyse der FACT-Daten enthält eine Vielzahl von Schritten, von denen jeder noch ein großes Optimierungspotential bietet. Um dieses Potential vollständig auszuschöpfen wurde basierend auf dem *streams framework* aus Projekt A1 [A1/517] eine Experimentierumgebung geschaffen, die es erlaubt durch *rapid prototyping* neue Analyse-Schritte auszuprobieren und so die Datenanalyse der FACT-Daten experimentell zu untersuchen und zu erweitern. Eine Sammlung verschiedener Verarbeitungsfunktionen ist so in der Bibliothek *FACT Tools* entstanden. Mit Hilfe der *FACT Tools* lässt sich so die gesamte Verarbeitung beginnend mit den aufgezeichneten Rohdaten bis zur Kalibrierung und Merkmalsextraktion abbilden. Die Datenverarbeitung erfolgt dabei vollständig im Strom, sodass eine kontinuierliche und realzeitliche Verarbeitung auch vor Ort erfolgen kann.

Aus Sicht der Informatik ergeben sich dabei insbesondere Fragen nach der Art der Modellierung der zu implementierenden Datenverarbeitung:

- Welcher Abstraktionsgrad der Software ist für Endanwender hier sinnvoll und erforderlich?
- Wie lässt sich eine hohe Wiederverwendbarkeit von Software-Modulen sicherstellen?
- Wie lassen sich Prozesse benutzerverständlich repräsentieren und ermöglichen zudem eine verlässliche Reproduzierbarkeit von Ergebnissen?

Mit dem in *streams* gewählten Abstraktionsmodell werden die Analyseschritte operationalisiert und können dank der abstrakten API in verschiedenen Ausführungsmodellen wiederverwendet werden. So ist neben der Datenstromverarbeitung u.A. auf Kleinstrechnern wie dem Raspberry-PI auch die Abbildung auf verteilte Clusterumgebungen auf Basis des MapReduce-Frameworks (Apache Hadoop) möglich und in ersten Teilen erfolgreich untersucht worden. Weitergehend ist auch eine graphische Modellierung der Datenflüsse auf dieser operationalisierten Ebene möglich [A1/516].

Weiter wurden etablierte und für die Analyse relevante Bildparameter in *FACT tools* implementiert. Diese Parameter stehen somit auch schon für den Ereignisfilter zur Verfügung. Bisher wurden mehr als 30 Stream-Prozessoren zur Datenanalyse und Merkmalsextraktion umgesetzt, darunter auch Neuentwicklungen, die den zeitlichen Verlauf der Signale berücksichtigen (Masterarbeit in Vorbereitung, Jan Freiwald). Mit den FACT tools entsteht somit ein Rahmen zur Datenanalyse, der portabel, schnell erweiterbar und ressourcenschonend ist. Damit ist eine Grundlage geschaffen, um vorhandene Bibliotheken und eigene Entwicklungen zu Datenstromlernverfahren für das Teleskop einzusetzen. Durch das Streaming kann der erste Analyseschritt um einen Faktor fünf bis acht mal schneller durchgeführt werden als mit der herkömmlich eingesetzten Software. Wenn die Streaming Software soweit entwickelt und überprüft wurde, dass eine Speicherung der Ereignisse zwischen Datennahme und Streaming entfällt, kann die Triggerrate und somit die Effizienz des Teleskops soweit gesteigert werden, bis beide nur noch durch die Datenverarbeitungsrate im Streaming-Filter begrenzt sind.

Signal-Untergrund-Separation (AP2): In allen Experimenten der Astroteilchenphysik werden die physikalisch interessanten und relevanten Ereignisse von Untergrundereignissen überlagert. Die Anzahl dieser Untergrundereignisse ist je nach Fragestellung und Experiment um viele Größenordnungen größer als die Anzahl der Signalereignisse [509, 523]. Jede physikalische Analyse muss sich daher auf allen Stufen immer wieder mit der Separation von Signal und Untergrundereignissen beschäftigen. Im Berichtszeitraum wurde dabei vorrangig an der Optimierung der bis dahin erratischen Abfolge und Auswahl von Analyseschritten gearbeitet. Insbesondere wurden die folgenden unterschiedlichen Ansätze im Hinblick auf Effizienz und Robustheit untersucht:

Die verteilungsbezogene Merkmalskonstruktion erstellt neue Attribute für die Trennung von Ereignisklassen. Diese Methode wurde auf Cherenkov-Teleskop-Daten zur Trennung von hochenergetischen Gammas und Hadronen angewandt [C3/544]. Neue Merkmale konnten durch die Maximierung der Hellingerdistanz zwischen Signal- und Untergrundverteilungen erhalten werden. Die Anpassung einer elliptisch-symmetrischen Verteilung wie der multivariaten Normalverteilung an die Daten enthält dabei die Informationen über die zugrunde liegenden elliptischen Konturen. Um weitere Informationen in den Daten auszunutzen, wurden Methoden betrachtet, in deren Zentrum Hypothesentests stehen, um so das Wissen über das unterschiedliche Symmetrieverhalten der gamma- und hadroninduzierten Schauer miteinzubeziehen. Zusammen mit der Zeitinformation der Schauerentwicklung kann die Signalhypothese, die die für Signalereignisse festliegende Ausrichtung der Schauer in der Kamera nutzt, getestet werden. Hierfür notwendige Berechnungen können in FACT tools durchgeführt werden (s.o.). Die Abstände zwischen Signal- und Untergrundverteilungen sind vielversprechende Merkmale für die Klassifikation und bewirken drastische Verbesserungen der Fehlklassifikationsrate und der Schätzung der Anzahl der Signalereignisse.

Boosting bei sehr großen Datenmengen beruht darauf, dass diese zwar einerseits eine genaue Klassifikation auch bei verzerrten Verteilungen erleichtern – andererseits wird aber die Lerndauer bei Verfahren mit Komplexität größer $O(n)$ mit n Beobachtungen zu groß. Ein Ansatz ist es, aus dem riesigen Datenvolumen jeweils Stichproben zu ziehen, aus denen gelernt wird. Das Boosting ist ein Meta-Lernverfahren, das sich besonders bei stark verzerrten Verteilungen bewährt, weil die Verteilung durch Gewichtung der Beispiele stratifiziert, also ausbalanciert wird. Die Modelle der einzelnen Stichproben werden als Basismodelle für die Klassifikation genutzt. Das Verfahren *Ada²Boost* von Martin Scholz [532] baut auf *AdaBoost* [295] auf. Die Verteilung der Klassen in der Trainingsmenge wird bei jeder Iteration des Algorithmus verändert. Während dies bei *AdaBoost* anhand der durchschnittlichen Güte für alle Stichproben geschieht, verschiebt *Ada²Boost* die Verteilung für jede Stichprobe einzeln. *Ada²Boost* verwendet ausschließlich harte Entscheidungsfunktionen. Das bedeutet, dass lediglich das Vorzeichen der Entscheidungsfunktion genutzt wird, selbst wenn das Basismodell Konfidenzen liefert. In einer Zusammenarbeit von A1 und C3 hat Marius Helf *Ada²Boost* verbessert, sodass es Konfidenzen ausgibt. Die wichtigere Weiterentwicklung war aber die Erhöhung der Skalierbarkeit. Sein *Ada²Boost for Very Large Data Sets*, AdaVLDS, verwirft die aktuelle Trainingsmenge alle R Iterationen und zieht ein neues Sample aus der Datenbank. Bevor das nächste Basismodell trainiert wird, werden die Beispielgewichte der neuen Trainingsmenge nach den bekannten Umgewichtungsregeln neu berechnet. Dadurch, dass nicht nur Beispiele umgewichtet und so die Verteilungen angepasst werden, sondern den Basislernern auch beliebig häufig neue Beispiele zugeführt werden, lassen sich auch stärkere Klassifikatoren wie etwa die SVM *boosten*, die normalerweise nicht oder nur gering vom Boosting profitieren. In der Anwendung des Verfahrens auf MAGIC-Daten profitiert die SVM mit ihrer quadratischen Laufzeit bei größeren Trainingsmengen sehr stark von AdaVDSL. Es zeigen sich (neben der Steigerung der Klassifikationsgüte) auch deutliche Laufzeitverbesserungen bei der Verwendung von J48-Entscheidungsbäumen mit AdaVLDS. Wie auch Datenstromalgorithmen verarbeitet AdaVDSL große Datenmengen, indem iterativ kleinere Untermengen betrachtet werden. Im Gegensatz zu Datenstromalgorithmen braucht AdaVLDS jedoch wahlfreien Zugriff auf die gesamte Datenmenge.

Das Programmpaket ROOT [518, 513] ist das Standardwerkzeug für einen Großteil der Analysen in der Teilchen- und Astroteilchenphysik. Als Grundlage für die Signal-Untergrundtrennung in der Astroteilchenphysik wurde daher eine Schnittstelle zwischen ROOT und einer SQL-Datenbank implementiert, welche die Zugriffsfunktionen von ROOT nutzt. Alle Lernverfahren wurden als Operatoren in RapidMiner implementiert. Ein Prozess, der Skripte für die Anwendung von RapidMiner auf einem Rechencluster, hier: PhiDo, generiert, wurde implementiert. Auf diese Weise wurde zum ersten Mal eine Schnittstelle zwischen ROOT und *advanced learning* geschaffen. In der zweiten Phase kann auf diese Entwicklungen aufgebaut werden.

Für Merkmalsselektion und Anwendung eines Random Forest in IceCube stehen bis zu 2000 Merkmale für die Trennung von Signal- und Untergrundereignissen zur Verfügung. Die hohe Dimensionalität der Fragestellung im Zusammenspiel mit der wenig balancierten Verteilung der Klassen stellt eine der größten Herausforderungen bei der Bearbeitung dieser Aufgabe dar. Die Anwendung von Schnitten auf anhand von Kenntnissen über den Detektor ausgewählten Merkmalen stößt vor allem dann an ihre Grenzen, wenn einzel-

ne Ereignismerkmale, wie zum Beispiel Winkelinformationen, von mehreren unterschiedlichen Algorithmen rekonstruiert werden. Es wurde ein aus Merkmalsauswahl und Random Forest bestehender Analyseprozess entwickelt und sehr erfolgreich auf Daten einer geometrischen Konfiguration des IceCube-Neutrinodetektors angewendet. Die Rate der separierten Signalereignisse, die für weitere physikalische Analysen genutzt werden, konnte so verdreifacht werden [C3/538].

Da in IceCube ein Teil der Merkmale redundante oder für die Separation irrelevante Informationen enthält, ist die Reduktion des Merkmalsraumes ein entscheidender Schritt für eine hochperformante Trennung. Ziel ist es dabei immer, eine für die Separation geeignete Repräsentation der Daten zu finden. Zusätzlich kann durch eine geeignete Auswahl an Attributen die Laufzeit der nach der Merkmalsselektion ansetzenden Lernalgorithmen überschaubar gehalten werden. Für ein optimales Separationsergebnis darf möglichst keine relevante Information verloren gehen. Eine geeignete, auf Kenntnissen über den Detektor basierende Vorauswahl an Merkmalen zeigte hier einen erheblichen Geschwindigkeits- und Genauigkeitsgewinn für die eigentliche algorithmenbasierte Merkmalsauswahl. Dazu müssen erstens Attribute mit einer zu großen Anzahl von fehlenden Werten ausgeschlossen werden. Derartige Attribute sind in den Daten vorhanden, da nicht alle Attribute für jedes Ereignis erfolgreich rekonstruiert werden können. Merkmale mit mehr als 70% an vollständig rekonstruierbaren Beispielen sollten jedoch berücksichtigt werden, um möglichst viel Information zur Verfügung zu haben (Diplomarbeit Max Wornowizki). Weiter müssen Attribute ausgeschlossen werden, die aufgrund der enthaltenen Information entweder nutzlos sind oder die Gefahr einer potentiellen Fehlinterpretation bergen [C3/546]. Beide Schritte reduzieren den Merkmalsraum um bis zu 80%. Daraus resultiert ein erheblicher Geschwindigkeitsgewinn für die algorithmenbasierte Merkmalsauswahl als dritten und letzten Schritt der Kette.

Im Rahmen der Arbeit von Ruhe et al. [A1/C3/541, C3/A1/540, C3/546] konnte gezeigt werden, dass der Einsatz des Minimum-Redundancy-Maximum-Relevance-Algorithmus (MRMR) für den Einsatz im Bereich der Astroteilchenphysik besonders geeignet ist. Die effiziente Implementierung von Benjamin Schowe (Teilprojekt C1) ermöglicht auch bei hochdimensionalen Daten einen *Wrapper-Ansatz* [533]. Insbesondere die redundanten und hochkorrelierten Attribute können so adäquat berücksichtigt werden. Innerhalb des gesamten Analyseprozesses konnte in [C3/546] nachgewiesen werden, dass MRMR gegenüber anderen Verfahren (Forward Selection, Backward Elimination) eine signifikant höhere Stabilität aufweist. Im direkten Anschluss an die Merkmalsauswahl erfolgt eine Trennung von Signal und Untergrund mit einem Random Forest, der vor allem aufgrund der hohen Stabilität (Ensemblemethode) und aufgrund der guten Interpretierbarkeit baumbasierter Algorithmen geeignet ist. Innerhalb des Berichtszeitraums wurde der Einsatz verschiedener Lerner (Naive Bayes, k-NN, SVM mit verschiedensten Kernelfunktionen) in IceCube untersucht. Dabei konnte empirisch gezeigt werden, dass der Random Forest einen optimalen *trade-off* zwischen Laufzeit und Performanz darstellt [C3/546].

Eine Vorbereinigung der Daten durch sogenannte *Precuts* hat das Ergebnis des Random Forest positiv beeinflusst, da so bereits ein großer Anteil an unerwünschten Ereignissignaturen ausgeschlossen werden kann. Die im Rahmen des Trainings des Random Forests auf simulierten Ereignissen gewonnenen Erkenntnisse bezüglich der Performanz des Forests und der zu erwartenden Reinheit des Samples konnten auf gemessene Daten übertragen werden. Durch die konsequente Anwendung des im Rahmen der ersten Förderperiode erarbeiteten Analyseschemas wurde routinemäßig eine Reduktion des Untergrundes um sechs bis sieben Größenordnungen erreicht.

Das entwickelte Analyseschema wird derzeit von Martin Schmitz und Florian Scheriau auf andere Detektorgeometrien übertragen. Dabei zeigt sich bereits, dass das erarbeitete Analyseschema ohne Schwierigkeiten auch auf andere Detektorgeometrien übertragen werden kann. Erste Tests zeigen zusätzlich, dass sich durch eine systematische Betrachtung der Verteilung von Untergrundereignissen in energieabhängigen Variablen eine erhebliche Verbesserung des Recalls erreichen lässt. Auf diese Weise kann die Kontamination von niedrig populierten Bins mit Untergrundereignissen effektiv vermieden werden, was zu wiederum zu einer drastischen Reduktion der benötigten simulierten Ereignisse führt.

Zeitreihen-/Lichtkurvenanalyse (AP 3): Zeitreihen wurden in diesem Arbeitspaket auf zwei unterschiedlichen Skalen analysiert: zum einen auf der über Monate und Jahre laufenden Skala der Beobachtung von astrophysikalischen Objekten, zum anderen auf der Skala von etwa 100 ns, innerhalb der sich das Abbild eines Schauers in der Kamera ausbreitet.

Ist mit den diskutierten Methoden das Signal einer astrophysikalischen Quelle identifiziert worden, stellt sich die Frage nach deren Zeitabhängigkeit (Lichtkurve). Problembedingt können die Messungen nicht in äquidistanten Abständen durchgeführt werden. In Arbeitspaket 3 wurde daher mit der Frage nach der Detektierbarkeit von periodischen Schwankungen bei nicht äquidistanten Messungen die Messplanung in den Vordergrund gestellt. Obgleich der Schritt der eigentlichen Lichtkurvenanalyse in den Analyseketten weit

hinten angesiedelt ist und selbst keine nennenswerten Ressourcen benötigt, stellt die Optimierung von Messzeiten mit allen dazu benötigten Ressourcen einen entscheidenden Schritt zur Ergebnisverwertung dar. In dem Arbeitspaket wurde daher die im Folgenden beschriebene Methode zur Identifikation zeitlicher Variabilitäten entwickelt und getestet. Diese auf dem R-Package RobPer basierende Methode muss zur Versuchsplanung auf die individuelle physikalische Problemsituation wie folgt angepasst werden: Das R-Package RobPer stellt Funktionen zur Berechnung von Periodogrammen bei Vorliegen von unregelmäßigen Beobachtungszeitpunkten und zusätzlicher Angabe von Messgenauigkeiten zur Verfügung und erlaubt es, aufällige Periodogrammbalken mit Hilfe von Methoden der Ausreißererkennung zu identifizieren. Diese Methode ist auch für die Analyse von Lichtkurvendaten nützlich, zumal ein Generator für die Erzeugung künstlicher Lichtkurven in dem Paket enthalten ist. Ein Periodogramm einer Zeitreihe, das von der R-Funktion RobPer berechnet wird, besteht aus Periodogrammbalken, die jeweils einem Koeffizienten entsprechen, der sich bei der Anpassung einer periodischen Funktion mit unterschiedlichen Periodenlängen an die Zeitreihe ergibt. Periodische Stufenfunktionen, Fourier-Reihen bis zum einschließlich vierten Grad sowie periodische Spline-Funktionen werden dabei in der Software unterstützt. Neben der klassischen Methode der kleinsten Quadrate bietet RobPer auch andere Regressionstechniken wie die Methode der kleinsten absoluten Abweichung, *least trimmed squares* (LTS), M-Regression, S-Regression und τ-Regression. Die Effizienz einiger robuster Regressionstechniken wie des *repeated median*, der LTS oder der *least weighted squares* Regression zur Anpassung von Sinusfunktionen wurden im Rahmen einer Abschlussarbeit untersucht. Zusätzlich zur Berechnung vieler Periodogramme aus der Literatur erlaubt RobPer die Konstruktion von neuen Periodogrammen, die zuvor noch nicht betrachtet worden sind. Um auffällige Perioden zu erfassen, kann mit Hilfe robuster Anpassung einer Beta-Verteilung nach Ausreißern im Periodogramm gesucht werden, anstatt mit festen kritischen Werten zu arbeiten, die nur im Falle der Kleinste-Quadrate-Regression bei unabhängigen Periodogrammbalken und einer Null-Hypothese mit normalverteiltem weißen Rauschen gerechtfertigt sind. Die jeweiligen Vorteile der unterschiedlichen Methoden zur Periodogrammberechnung in RobPer werden in einer gründlichen Simulationsstudie analysiert. Für weitere Informationen sei auf die Doktorarbeit von Anita Thieler und ihren Technical Report verwiesen. Eine Vorabversion der Implementierung wurde bereits in [C3/543] verwendet.

Zeitreihen anderer Art ergeben sich, wenn die zeitliche Ausbreitung und Entwicklung von Abbildern der gamma- oder hadron-induzierten Luftschauer z.B. in FACT analysiert werden. Auch diese auf Ergebnissen des Projektes B2 beruhende und mit B2 durchgeführte Analyse soll letztlich zur Signal-Untergrund-Trennung beitragen. In Zusammenarbeit mit Projekt B2 wurde ein Verfahren zur Analyse von Graustufenbildern entwickelt, um die zeitliche Entwicklung von Abbildern von Nano-Objekten in einer Probenflüssigkeit zu beschreiben (Masterarbeit Sermad Abbas). Dieses Verfahren wird in C3 genutzt, um die sich zeitlich ändernde Kontur des Luftschauerabbildes (den Sprung zwischen der hellen Bildfläche des Schauers und der nur mit Rauschsignalen gefüllten Fläche) zu identifizieren. Die Bildpunkte der Kontur entsprechen den Koordinaten mindestens eines erfassten Sprungs. Das vorgeschlagene Verfahren basiert auf Zwei-Stichproben-Tests zur Erkennung von Pixelkoordinaten mit positiven Sprüngen. Die identifizierten Koordinaten werden zur Konturbestimmung bezüglich zeitlicher und räumlicher Kriterien zusammengefasst. Das Verfahren wird auf Grundlage simulierter Datensätze bewertet und auf Messungen angewandt. In der Bachelorarbeit von Dominik Baack wurden auf der Basis von Methoden zur aktiven Konturverfolgung erstmalig der räumliche und zeitliche Verlauf der Abbildung eines Luftschauers beschrieben.

Spektral-Rekonstruktion (AP 4): In Experimenten ist in der Regel die physikalisch interessante Größe nicht direkt beobachtbar, sodass sie aus Messungen korrelierter Variablen rekonstruiert werden muss (siehe z.B. [519]). In der Astroteilchenphysik gilt dies z.B. für die Energie von Teilchenschauern in der Atmosphäre, die zuvor als raum-zeitliche Muster selektiert wurden. Anstatt einer direkten Messung der Energie des Schauers werden energieabhängige Variablen bestimmt, mit Hilfe derer die Energie eines Schauers abgeleitet werden kann. In einem realen Detektor wird die Messung solcher energieabhängiger Variablen durch verschiedene Gesichtspunkte erschwert. Hierzu zählen insbesondere die Transformation der Variablen, die nicht notwendigerweise linear sein muss, sowie die begrenzte Akzeptanz des Detektors. Für das IceCube-Experiment kommt hinzu, dass die gesuchte Energie des Neutrinos nur sehr indirekt über Cherenkovlicht, das durch Energieverluste eines sekundär produzierten Myons erzeugt wurde, bestimmt werden kann. Im Experiment IceCube werden Maximum-Likelihood-Methoden verwendet, in denen die Energie abhängig von der Wahrscheinlichkeitsverteilung der Zeitverzögerung im Ansprechverhalten der Lichtsensoren oder aber deren Signalamplitude geschätzt wird [504, 499]. Alternativ können Regressionen mit statistischen Verfahren wie neuronalen Netzen oder Random Forrests, wie im MAGIC-Experiment, benutzt werden [509]. Um die Transferfunktion des Detektors realistisch zu berücksichtigen, müssen jedoch immer Entfaltungsmethoden verwendet werden.

Mathematisch kann ein solches inverses Problem durch die Fredhom'sche Integralgleichung zweiter Art beschrieben werden:

$$g(y) = \int A(x,y)f(x)dx \tag{3.1}$$

In Gleichung (3.1) wird die Verteilung der gemessenen Größe durch $g(y)$ beschrieben, wohingegen die Verteilung der gesuchten Größe durch $f(x)$ gegeben ist. Beide Verteilungen werden durch die Transferfunktion $A(x,y)$ des Detektors miteinander verknüpft. Dabei werden in $A(x,y)$ neben der eigentlichen Antwort des Detektors auf ein bestimmtes Signal auch weitere Effekte wie die Erzeugung von Sekundärteilchen und deren Propagation durch das entsprechende Medium zum Detektor berücksichtigt. Dies ist notwendig, da alle diese Effekte stochastischer Natur sind. Die Lösung eines solchen inversen Problems erfolgt in der Regel durch Überführen der Integralgleichung in eine Matrixgleichung der Form:

$$\vec{g}(y) = A(x,y)\vec{f}(x). \tag{3.2}$$

Die histogrammierten Verteilungen $\vec{g}(y)$ und $\vec{f}(x)$ fungieren dabei als Näherungen der tatsächlichen Verteilungen $g(y)$ und $f(x)$. Die Lösung von Gleichung (3.2) durch eine simple Inversion der Matrix wird durch verschiedene Aspekte erschwert. Zum Einen ist die Matrix $A(x,y)$ nicht zwingend quadratisch, sodass die Inverse A^{-1} nicht notwendigerweise existiert. Zum Anderen führen kleine Eigenwerte der Matrix zu Fluktuationen der Lösung. Durch ein solches unphysikalisches Verhalten der Lösung können Strukturen in $f(x)$ vorgetäuscht werden, die in der Realität nicht existieren.

Zur Lösung von Gleichung (3.1) wurde in enger Zusammenarbeit der Sonderforschungsforschungsbereiche SFB 823 und SFB 876 der Entfaltungsalgorithmus TRUEE (siehe [C3/537] und Dissertation Natalie Milke) entwickelt. Um die für inverse Probleme typische Fluktuation der Lösungen zu unterdrücken, wird in TRUEE ein zusätzlicher dämpfender Regularisierungsterm (Tikhonov-Regularisierung) eingeführt. Die Funktionsfähigkeit von TRUEE konnte innerhalb der Dortmunder Arbeitsgruppe bereits in Analysen für das IceCube-Neutrinoteleskop [C3/538] sowie für das MAGIC-Teleskop bestätigt werden [C3/545].

Trotz der hohen Funktionalität und der hervorragenden numerischen Genauigkeit von TRUEE besitzt diese Methode der regularisierten Entfaltung einige Nachteile, die durch den im SFB 876 in Entwicklung befindlichen Algorithmus DSEA/SpectEst überwunden werden. Im Folgenden sollen zunächst diese Nachteile dargelegt werden:

1. Die Anzahl der Eingangsvariablen in TRUEE ist auf drei begrenzt, weil es sich bei der Bestimmung der Transfer-Matrix um die Erstellung eines dichtebasierten Modells handelt und die Anzahl der für die Erstellung solch eines Modells benötigten Beispiele exponentiell mit der Anzahl der betrachteten Merkmale ansteigt. Aus der begrenzten Anzahl an Merkmalen, die für die Entfaltung verwendet werden können, ergibt sich somit ein Verlust an Information.

2. Im Rahmen der TRUEE-Entfaltung wird nur die gesuchte Verteilung $f(x)$ rekonstruiert. Es ist nicht möglich zu beurteilen, welche individuellen Ereignisse bei welchen x-Werten wie stark beigetragen haben. Auch Informationen über das gleichzeitige Verhalten anderer physikalisch interessanter Variablen (Zeit, Richtung) sind nicht verfügbar. Mehrdimensionale Verteilungen können nicht entfaltet werden.

3. Durch die Bestimmung einer mittleren Transfer-Matrix, die anschließend für alle Ereignisse verwendet wird, wird die Inhomogenität des Detektors ignoriert. Dies führt bei großen inhomogen instrumentierten Detektoren, wie IceCube, zu Problemen. Es wird erwartet, dass eine Berücksichtigung geometrischer Information bei dem Entfaltungsprozess zu einer Verbesserung des Ergebnisses führt.

Um die bestehenden Beschränkungen der regularisierten Entfaltung aufzuheben, wurde in [C3/542, C3/546] in Kooperation mit A1 die Methode DSEA/SpectEst entwickelt. DSEA/SpectEst ist ein auf Methoden des maschinellen Lernens aufbauender Ansatz zur Lösung inverser Probleme, der gegenüber den aktuell verwendeten Methoden zwei wesentliche Vorteile liefert: Er ist in der Anzahl der Eingangsgrößen nicht beschränkt, sodass beispielsweise geometrische Informationen über Detektor und Detektionsmedium verwendet werden können. Ebenso können mehrere x-abhängige Merkmale in die Entfaltung einbezogen werden. Darüber hinaus kann der Effekt einzelner Beobachtungen leicht beurteilt werden, was zu einer starken Reduzierung der Rechenzeit führt, wenn eine Veränderung des Spektrums über die Zeit oder ein anderes Attribut untersucht werden soll.

Im Gegensatz zu den in [520] und [521] vorgeschlagenen Ansätzen wird das inverse Problem (3.2) hier als multinominale Klassifikation betrachtet, in der bekannte Klassifikatoren angewendet werden. Der ebenfalls auf Data Mining-Methoden basierende Ansatz in [525] überführt die Entfaltung durch eine Reihe von Transformationen in ein binominales Klasssifikationsproblem, welches mit Hilfe eines neuronalen Netzes

gelöst wird. Durch die Behandlung der Entfaltung als multinominales Klassifikationsproblem werden die Transformationen jedoch obsolet. Darüber hinaus kann mit DSEA/SpectEst ein auf das jeweilige Problem zugeschnittener Lernalgorithmus verwendet werden. Unsere Studien, in denen ein einfacher Random Forest als Klassifikator benutzt wurde, zeigen, dass erhebliche Verbesserungen erreicht werden können, wenn die individuellen Vorhersagen der Bäume anstatt der kombinierten Vorhersage des Waldes benutzt werden.

Datenmodellierung und Simulation (AP 5): Die Entwicklung aller in diesem Teilprojekt untersuchten Analyseverfahren beruht auf der Existenz einer vollständigen Simulation der Physik und des Detektors mit Monte-Carlo-Verfahren. Zur Berechnung der Detektorantwort auf die im Laufe allein eines Jahres einfallenden und untersuchten Primärteilchen bei IceCube werden einige tausend CPU-Jahre benötigt [502]. Diese werden durch den permanenten Einsatz mehrerer großer CPU-Farmen weltweit mit jeweils etwa 100 Rechenknoten realisiert. Auch der zur Speicherung der Monte-Carlo-Informationen benötigte Plattenplatz übersteigt den für die gemessenen Daten um das zehnfache und liegt in der Größenordnung von hunderten von Terabyte. Die sich daraus ergebenden Probleme bei der Nutzung der Ressourcen CPU-Zeit und Plattenplatz ebenso wie bei den Energiekosten liegen auf der Hand.

Im Zentrum dieses Arbeitspakets steht die Frage, wie die Geschwindigkeit der Monte-Carlo-Produktion bei der geforderten hohen numerischen Qualität der Simulation ressourcensparend erhöht werden kann. Da in der Summe der in der Astroteilchenphysik für IceCube und MAGIC/FACT verwandten Monte-Carlo-Algorithmen hunderte von Mann-Jahren an Arbeit stecken, wird in diesem Arbeitspaket exemplarisch an der Optimierung eines der Monte-Carlo-Algorithmen gearbeitet.

Innerhalb des Berichtszeitraums wurde der Leptonpropagator PROPOSAL (PRopagator with Optimal Precision and Optimized Speed for All Leptons) im Rahmen der Dissertation von Jan-Hendrik Köhne sowie von begleitenden Master- und Bachelorarbeiten entwickelt (siehe [C3/536] und Dissertation J.H. Köhne). Dieser spezielle Monte-Carlo-Algorithmus hat die Aufgabe, die Energieabgabe von leptonischen Teilchen (z.B. Myonen) bei ihrem Weg durch Materie (z.B. Eis) mit hoher Präzision zu berechnen. Die Entwicklung erfolgte in drei Stufen. Zunächst wurde ein früher in der physikalischen Arbeitsgruppe erstellter Code in Bezug auf die Physik überarbeitet. Dabei wurde sichergestellt, dass der Code die Wechselwirkungen mit einer numerischen Genauigkeit von 10^{-5} so genau berechnet, dass diese Unsicherheit gegenüber den systematischen Fehlern der verschiedenen Wirkungsquerschnitte, welche aus deren Berechnung in der theoretischen Physik folgen, vernachlässigt werden kann. Dieser Code wurde in Kooperation mit dem Projekt A1 in einem zweiten Schritt systematisch effizienz- und laufzeitoptimiert, sodass er in der Version, die auf klassischen CPUs rechnet, etwa einen Faktor drei schneller ist als seine Vorgänger. Von PROPOSAL werden damit Energieverluste durch Ionisation, Bremsstrahlung, Paarproduktion und die photonukleare Wechselwirkung berechnet. Darüber hinaus sind für die Bremsstrahlung und die photonukleare Wechselwirkung jeweils unterschiedliche Berechnungen aus der theoretischen Physik implementiert. Zusätzliche Effekte, wie Vielfachstreuung oder der LPM-Effekt, können optional ebenfalls einbezogen werden. PROPOSAL wurde erfolgreich in die IceCube Simulationskette eingefügt und wird seit 2013 für alle Simulationen benutzt. Aufgrund der objektorientierten Programmstruktur kann der Code ohne weiteres auch außerhalb dieser Kette und für andere Experimente eingesetzt werden.

Parallel zu der Entwicklung von Techniken zur Ressourceneinsparung wurden flexible statistische Methoden zur Quantifizierung des Maßes an Übereinstimmung zwischen Monte-Carlo-generierten und gemessenen Verteilungen entwickelt. Hierzu werden f-Divergenzen genutzt, die dichte-basierte Distanzmaße zwischen Verteilungen sind. Divergenzmaße spezifizieren sich nicht auf bestimmte Momente. Daher sind entsprechende Tests in der Lage alle Arten von Abweichungen zwischen den zugrunde liegenden Verteilungen festzustellen und sind nicht nur auf Mittelwerts- und Varianzunterschiede beschränkt. Eindimensionale stetige Zufallsvariablen wurden betrachtet, um eine neue nichtparametrische Methode zur Schätzung der Maße vorzuschlagen, die Kerndichteschätzung, Spline-Glättung und numerische Integration beinhaltet. Simulationsexperimente zeigen, dass der Spezialfall der Hellingerdivergenz leichter zu schätzen ist als der der Kullback-Leibler-Divergenz, was möglicherweise mit der Beschränktheit des ersten Maßes auf dem Intervall (0,1) erklärbar ist. Der vorgeschlagene numerische Schätzer zeigt einen kleinen mittleren quadratischen Fehler und ist stabil. Er ist schlechter als einige semiparametrische Methoden, solange diese mit dem korrekten Modell für das Dichteverhältnis arbeiten. Permutationsmethoden ermöglichen es, das Problem der Zwei-Stichproben-Homogenität basierend auf beliebigen Divergenzschätzern anzugehen. Sie stellen keine Annahmen an zugrundeliegende Verteilungen und sind somit allgemein anwendbar. Das dabei entstehende Permutationsverfahren wird für verschiedene Divergenzschätzer, Verteilungen, Stichprobengrößen und H_1-Alternativen angewendet und mit einigen parametrischen Tests, einem asymptotischen divergenzbasierten Test mit einem semiparametrischen Divergenzschätzer und den nichtparametrischen Kolmogorov-Smirnov- und Anderson-Darling-Tests verglichen. Die Ergebnisse zeigen, dass die Wahl des Divergenzmaßes nicht so viel Einfluss auf die Ergebnisse hat

wie die Schätztechnik. Der Test basierend auf dem numerischen Schätzer ist abermals stabil und sehr gut in der Klasse der Permutationstests. Im Vergleich zu klassischen nichtparametrischen Verfahren erkennen die Permutationsmethoden Lokationsalternativen weniger oft, decken jedoch deutlich häufiger unterschiedliche Varianzen oder Schiefen der Verteilungen auf. Das asymptotische Testverfahren führt zu besseren Ergebnissen als die Permutationstests, solange die zugrunde liegenden Annahmen erfüllt sind, ist aber nicht mehr verwendbar, wenn diese verletzt werden. Die parametrischen Tests schneiden am besten für eine geeignete Alternative ab, sind aber verteilungsabhängig. Andernfalls ist er den Modellen jedoch überlegen. Die Methode wird angewandt, um interessante Variablen in IceCube-Daten zu identifizieren. Für mehr Details sei auf den technischen Bericht von Max Wornowizki verwiesen.

In Zusammenarbeit mit Teilprojekt C4 wurde ein Algorithmus entwickelt, der in dem Fall, dass Simulation und Realdaten sich signifikant unterscheiden, eine Korrekturverteilung berechnet. Dazu wird die zu simulierende Verteilung $F = s \cdot G + (1 - s) \cdot H, s \in [0, 1]$ als nichtparametrische Mischung aus der bisherigen Simulation G und der zu bestimmenden unbekannten Korrekturverteilung H modelliert. Letztere wird nun durch Maximierung des Mischungsparameters s so berechnet, dass die resultierende Mischung im gewünschten Konfidenzintervall um die Verteilung der realen Daten liegt. Die Korrekturverteilung kann zur Verbesserung der Simulationen verwendet werden und domänenspezifische Merkmale der untersuchten Verteilungen hervorheben.

Projektumstellung: Wie aus dem vorliegenden Bericht über das Teilprojekt deutlich wird, war das Projekt in der ersten Phase des SFB sehr erfolgreich. Bei der Analyse der Daten der verschiedenen Teleskope wurden deutliche Fortschritte sowohl in Hinblick auf die Einsparung von Ressourcen als auch in Hinblick auf hier nicht diskutierte physikalische Ergebnisse erzielt. Gleichzeitig wurden neue statistische Methoden entwickelt und publiziert, die wichtige Voraussetzungen der Monte-Carlo-Generation von Daten und der Datenanalyse klären. Für die zweite Phase des SFBs ergeben sich aufbauend auf die bisherigen Ergebnisse drei zentrale Problemstellungen, um die sich die Arbeitspakete ranken.

1. In der realzeitlichen Datenanalyse bildet das Streaming in Form der FACT tools einen völlig neuartigen Ansatz, die ersten Stufen der Datenanalyse in das Messinstrument zu integrieren. Zwar sind die betrachteten Teleskope räumlich deutlich größer als andere in diesem SFB betrachtete Systeme, funktional werden Teleskope jedoch durch eine Anpassung der Streaming-Analyse an den Trigger zu eingebetteten Systemen, die eine Reduktion des limitierenden Ressourcenbedarfs um eine Größenordnung ermöglichen.

2. Nach der Online-Analyse müssen die Daten gespeichert werden, um ggf. eine Rekalibrierung der Daten mit über längere Zeit ermittelten Konstanten, eine witterungsabhängige Datenselektion oder ähnliches durchzuführen. Zu der nun folgenden Offline-Analyse wurden wie berichtet sehr erfolgreich Methoden der künstlichen Intelligenz eingesetzt. Es wurde gezeigt, dass eine Integration von bisher konsekutiv ausgeführten Selektionsmethoden und Entfaltungsschritten zu deutlich besseren und effizienteren Ergebnissen führt. In diese Abfolge kann auch der neu entwickelte DSEA/SpectEst eingebunden werden, mit dem man – wenn die Entwicklung abgeschlossen ist – Event-Informations-erhaltend auch mehrdimensional entfalten kann.

3. Für die Offline-Analyse werden Monte-Carlo-Datensätze benötigt, für deren Berechnung weltweit ein erheblicher Ressoucenaufwand an großen Computing-Farmen betrieben werden muss, der den für die reine Verarbeitung der gemessenen Daten bei weitem übersteigt. Es wurde gezeigt, dass in dem exemplarisch optimierten Lepton-Propagations-Monte-Carlo PROPOSAL die benötigten Ressourcen durch den Einsatz von GPU-Computing erheblich reduziert werden können.

In jeder dieser Projektlinien ist, wie jeweils gezeigt und im Sinn der Ziele des SFB 876, eine erhebliche Einsparung von CPU- bzw. Speicherressourcen möglich. Alle drei Projektlinien wurden in Zusammenarbeit mit dem Projekt A1 bzw. dem Lehrstuhl Katharina Morik erarbeitet und getestet. Um ein optimales Resultat in der zweiten Phase gewährleisten zu können, haben wir uns entschieden, die Arbeiten in direkter Kooperation mit der Arbeitsgruppe von Frau Morik weiterzuführen. Die bisher bestehende Kooperation mit der Arbeitsgruppe von Roland Fried, die sehr erfolgreich Methoden zum Vergleich und der Klassifizierung entwickelte, wird gleichzeitig eingestellt, weil nicht absehbar ist, dass auf diesem Weg in Zukunft eine ähnlich signifikante Einsparung von Ressourcen erreicht werden kann. Dem Teilprojekt stehen die Daten der Gamma-Astronomie und der Neutrino-Astronomie zur Verfügung. Bereits jetzt zeichnet sich ab, dass die erarbeiteten Konzepte allgemein und übertragbar sein werden. Allerdings ist es für eine echte Optimierung unabdingbar, sehr viele Eigenschaften der zu analysierenden Datenströme zu berücksichtigen. Diese Eigenschaften unterscheiden sich zwischen den Gamma-Teleskopen und Neutrino-Teleskopen in einer Vielzahl von

relevanten Details. Hieraus begründet sich die zweite Änderung in der Projektstruktur. Die Optimierung der Analyse der Gamma-Datenströme soll schwerpunktmäßig von den Antragstellern Katharina Morik und Wolfgang Rhode verfolgt werden. Die Analyse der Neutrinodaten wird durch Frau Morik gemeinsam mit dem Juniorprojektleiter Tim Ruhe erfolgen. Herr Ruhe hat in dem Projekt (und über die Auswertung von IceCube Daten) promoviert und ist als Erfinder und Entwickler von DSEA/SpectEst für eine Fertigentwicklung und Optimierung der kombinierten KI-Analyse der Daten prädestiniert.

Aus den folgenden Erläuterungen von Zielen und Arbeitsprogramm geht hervor, wie die Vernetzung dieser Projektlinien in den Arbeitspaketen, die in ihrer Ausrichtung bestehen bleiben, geplant ist.

Literatur

[499] M. G. Aartsen u. a. „Energy reconstruction methods in the IceCube neutrino telescope“. In: *Journal of Instrumentation* 9.3 (2014), P03009.

[500] M. G. Aartsen u. a. „IceCube search for dark matter annihilation in nearby galaxies and galaxy clusters“. In: *Physical Review D* 88.12 (2013), S. 122001.

[501] M. G. Aartsen u. a. „The IceCube Neutrino Observatory Part II: Atmospheric and Diffuse UHE Neutrino Searches of All Flavors“. In: *ArXiv e-prints* (2013).

[502] M. G. Aartsen u. a. „The IceProd Framework: Distributed Data Processing for the IceCube Neutrino Observatory“. In: *ArXiv e-prints* (2013).

[503] M. Aartsen u. a. „Measurement of the Atmospheric ν_e Flux in IceCube“. In: *Physical Review Letters* 110.15 (2013), S. 151105.

[504] R. Abbasi u. a. „An improved method for measuring muon energy using the truncated mean of dE/dx“. In: *Nuclear Instruments and Methods in Physics Research A* 703 (2013), S. 190–198.

[505] R. Abbasi u. a. „IceTop: The surface component of IceCube.“ In: *Nuclear Instruments and Methods in Physics Research A* 700 (2013), S. 188–220.

[506] R. Abbasi u. a. „Search for ultrahigh-energy tau neutrinos with IceCube“. In: *Physical Review D* 86.2 (2012), S. 022005+.

[507] R. Abbasi u. a. „The IceCube data acquisition system: Signal capture, digitization, and timestamping“. In: *Nuclear Instruments and Methods in Physics Research A* 601 (2009), S. 294–316.

[508] A. Achterberg, W. Rhode u. a. „First year performance of the IceCube neutrino telescope“. In: *Astroparticle Physics* 26 (2006), S. 155–173.

[509] J. Albert u. a. „Implementation of the Random Forest method for the Imaging Atmospheric Cherenkov Telescope MAGIC“. In: *Nuclear Instruments and Methods in Physics Research A* 588 (2008), S. 424–432.

[510] J. Aleksic u. a. „Performance of the MAGIC stereo system obtained with Crab Nebula data“. In: *Astroparticle Physics* 35.7 (2012), S. 435–448.

[511] H. Anderhub u. a. „Design and operation of FACT - the first G-APD Cherenkov telescope“. In: *Journal of Instrumentation* 8.06 (2013), P06008.

[512] H. Anderhub u. a. „Electronics for the camera of the First G-APD Cherenkov Telescope (FACT) for ground based gamma-ray astronomy“. In: *Journal of Instrumentation* 7 (2012), S. C01073.

[513] I. Antcheva u. a. „ROOT - A C++ framework for petabyte data storage, statistical analysis and visualization“. In: *Computer Physics Communications* 180 (2009), S. 2499–2512.

[514] R. Antonucci. „Unified models for active galactic nuclei and quasars“. In: *Annual Review of Astronomy and Astrophysics* 31 (1993), S. 473–521.

[515] J. Behringer und et al. (Particle Data Group). „2013 Review of Particle Physics“. In: *Physical Review D* 86 (2012).

[6] A. Bifet, J. Read, B. Pfahringer, G. Holmes und I. Zliobaite. „CD-MOA: Change Detection Framework for Massive Online Analysis“. In: *Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013*. Hrsg. von A. Tucker, F. Höppner, A. Siebes und S. Swift. Springer, 2013, S. 92–103.

[A1/516] C. Bockermann. „A Visual Programming Approach to Big Data Analytics“. In: *Proc. HCI International 2014*. 2014.

[A1/517] C. Bockermann und H. Blom. „Processing Data Streams with the RapidMiner Streams-Plugin“. In: *Proceedings of the 3rd RapidMiner Community Meeting and Conference*. 2012.

[518] R. Brun und F. Rademakers. „ROOT - An object oriented data analysis framework“. In: *Nuclear Instruments and Methods in Physics Research A* 389 (1997), S. 81–86.

[519] G. Cowan. „A survey of unfolding methods for particle physics“. In: *Conf.Proc.* C0203181 (2002), S. 248–257.

[520] G. D'Agostini. „A multidimensional unfolding method based on Bayes' theorem“. In: *Nuclear Instruments and Methods in Physics Research A* 362 (1995), S. 487–498.

[521] G. D'Agostini. „Improved iterative Bayesian unfolding“. In: *ArXiv e-prints* (2010).

[522] K. Daum u. a. „Determination of the atmospheric neutrino spectra with the Frejus detector“. In: *Zeitschrift fur Physik C Particles and Fields* 66 (1995), S. 417–428.

[523] T. Deyoung. „Neutrino Astronomy with Icecube“. In: *Modern Physics Letters A* 24 (2009), S. 1543–1557.

[524] B. Falkenburg und W. Rhode. *From Ultra Rays to Astroparticles: A Historical Introduction to Astroparticle Physics*. SpringerLink : Bücher. Springer, 2012.

[525] M. Feindt. „A Neural Bayesian Estimator for Conditional Probability Densities“. In: *ArXiv Physics e-prints* (2004).

[526] M. Honda, T. Kajita, K. Kasahara, S. Midorikawa und T. Sanuki. „Calculation of atmospheric neutrino flux using the interaction model calibrated with atmospheric muon data“. In: *Physical Review D* 75.4 (2007), S. 043006.

[527] IceCube Collaboration. „Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector“. In: *Science* 342 (2013).

[528] A. f. Karle. „IceCube - status and recent results“. In: *ArXiv e-prints* (2014).

[529] D. Kifer, S. Ben-David und J. Gehrke. „Detecting Change in Data Streams“. In: *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*. VLDB '04. VLDB Endowment, 2004, S. 180–191.

[530] P. Majumdar, A. Moralejo, C. Bigongiari, O. Blanch und D. Sobczynska. „Monte Carlo simulation for the MAGIC telescope“. In: *International Cosmic Ray Conference* 5 (2005), S. 203.

[531] T. Piran. „The physics of gamma-ray bursts“. In: *Reviews of Modern Physics* 76 (2004), S. 1143–1210.

[295] R. E. Schapire, Y. Freund, P. Bartlett und S. Lee. „Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods“. In: *Annals of Statistics* (1998), S. 1651–1686.

[532] M. Scholz. „Scalable and Accurate Knowledge Discovery in Real-World Databases“. Diss. Department of Computer Science, University of Dortmund, 2007.

[533] B. Schowe. „Feature Selection for high-dimensional data in RapidMiner“. In: *Proceedings of the 2nd RapidMiner Community Meeting And Conference (RCOMM 2011)*. Hrsg. von S. Fischer und I. Mierswa. Aachen: Shaker Verlag, 2011.

[534] B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[C3/535] R. Abbasi, W. Rhode u. a. „The Energy Spectrum of Atmospheric Neutrinos between 2 and 200 TeV with the AMANDA-II Detector“. In: *Astroparticle Physics* 34 (2010), S. 48–58.

[C3/536] J.-H. Koehne, K. Frantzen, M. Schmitz, T. Fuchs, W. Rhode, D. Chirkin und J. Becker Tjus. „PROPOSAL: A tool for propagation of charged leptons“. In: *Computer Physics Communications* 184 (2013), S. 2070–2090.

[C3/537] N. Milke, M. Doert, S. Klepser, D. Mazin, V. Blobel und W. Rhode. „Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics“. In: *Nuclear Instruments and Methods in Physics Research A* 697 (2013), S. 133–147.

[C3/538] T. Ruhe u. a. „Measurement of the atmospheric ν_μ spectrum with IceCube 59“. In: *Proceedings of the 33rd International Cosmic Ray Conference*. 2013.

[C3/A1/539] T. Ruhe, K. Morik und W. Rhode. „Application of RapidMiner in Neutrino Astronomy“. In: *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Hrsg. von M. Hofmann und R. Klinkenberg. CRCPress Book, 2013. Kap. Feature Selection and Classification in Astroparticle Physics and in Medical Domains.

[C3/A1/540] T. Ruhe, K. Morik und W. Rhode. „Data Mining Ice Cubes“. In: *Astronomical Data Analysis Software and Systems*. Hrsg. von C. e. a. Gabriel. 2011.

[A1/C3/541] T. Ruhe, K. Morik und B. Schowe. „Data Mining on Ice“. In: *Procs. of the Workshop on Astrostatistics and Data Mining in Large Astronomical Databases*. Hrsg. von L. Sarro, C. Bailer-Jones, L. Eyer, W. O'Mullane und J. de Ridder. 2011.

[C3/542] T. Ruhe, M. Schmitz, T. Voigt und M. Wornowizki. „DSEA: A data mining approach to unfolding“. In: *Proceedings of the 33rd International Cosmic Ray Conference*. 2013.

[C3/543] A. M. Thieler, M. Backes, R. Fried und W. Rhode. „Periodicity detection in irregularly sampled light curves by robust regression and outlier detection“. In: *Statistical Analysis and Data Mining* 6 (2013), S. 73–89.

[C3/544] T. Voigt und R. Fried. „Distance based Feature Construction in a setting of astronomy“. In: *Proceedings of the ECDA 2013*. Hrsg. von B. Lausen, S. Krolak-Schwerdt und M. Boehmer. Springer, 2014.

Andere Veröffentlichungen

[C3/545] M. Doert. „The talkative AGN next door“. Diss. TU Dortmund, 2013.

[C3/546] T. Ruhe. „Data Mining on the Rocks - A measurement of the atmospheric neutrino flux using IceCube in the 59-string configuration and a novel data mining based approach to unfolding“. Diss. Dortmund, Germany: TU Dortmund, 2013.

3.4 Planung des Teilprojekts

Ziele

In der ersten Phase des SFB 876 stand die Analyse der grundlegenden Probleme der Datenanalyse unter Ressourcenbeschränkung und die Entwicklung von einzelnen *speziellen* Lösungen für Großexperimente der Astroteilchenphysik im Vordergrund. Mit solchen speziellen Lösungen für einzelne Stufen der Datenanalyse wurden, wie gezeigt, bereits sichtbare Erfolge in der Datenanalyse für die untersuchten physikalischen Experimente erzielt. In der zweiten Phase sollen nun exemplarisch für die zwei Typen der Gamma- und Neutrinoteleskope (MAGIC/FACT bzw. IceCube) optimierte Analyseprozesse aufgestellt werden, in die die bisher erzielten Erkenntnisse integriert und für die weitere neue Methoden entwickelt werden. Obgleich die Abfolge der Analyseschritte in der Gamma- und Neutrinoastronomie grundsätzlich dieselbe ist, bestehen zwischen den beiden behandelten Experimenten grundlegende Unterschiede.

Das Gamma-Teleskop FACT ist ein von wenigen physikalischen Arbeitsgruppen (darunter die antragstellende Arbeitsgruppe der TU Dortmund) aufgebauter und für Eingriffe in die Hardware zugänglicher Prototyp, bei dem auch auf hardware-nahem Niveau die Datennahme und Analysekette variiert werden kann, wenn die physikalischen Messungen dadurch nicht beeinträchtigt werden. Hier können z.B. Trigger-Optimierungen vorgenommen werden. Für dieses Experiment wurden weiter in der ersten Phase Streaming-Algorithmen entwickelt und deren Potenzial demonstriert. Hier sollen auch neue, in der Gamma-Astronomie bisher noch nicht genutzte Attribute definiert und in der Analyse genutzt werden. Diese Attribute müssen daher auch zunächst physikalisch verstanden und in die Monte-Carlo-Simulation einbezogen werden. Daher soll der Analyseprozess auf dieser Grundlage beginnend mit hardware-nahen Algorithmen optimiert werden. Die Arbeiten am Beispiel der Gamma-Teleskope sollen von den Antragstellern Wolfgang Rhode und Katharina Morik mit geeigneten Mitarbeitern durchgeführt werden.

Das Neutrino-Teleskop IceCube wird von einer großen internationalen Kollaboration am Südpol betrieben. Während der ersten Phase wurde von IceCube – auch unter Verwendung von Ergebnissen dieses SFBs – das seit mehr als 25 Jahren gesuchte Signal von hochenergetischen extraterrestrischen Neutrinos gefunden. Die Aufzeichnung geeigneter Daten zur Klärung der Frage nach dem Ursprung der Neutrinos wird ohne Verzug fortgesetzt. Abgesehen von der Unzugänglichkeit des Experimentes wäre zur Zeit eine Neukonzeption

der Datennahme nicht vermittelbar. Die Monte-Carlo-Simulation beschreibt die Daten des Experimentes bereits auf einem hohen, wenn auch noch nicht perfektem Niveau. In den Arbeitspaketen, die auf der Basis von IceCube-Daten durchgeführt werden, geht es um die Datenanalyse auf höherem Abstraktions- und Perfektionsniveau. Eine möglichst hohe Zahl von Attributen soll zur Modellierung der Daten in mehreren Dimensionen genutzt werden. Die erforderlichen sehr umfangreichen Rechnungen werden in Rechenclustern wie den Dortmunder Farmen LiDo oder PhiDo durchgeführt. Sowohl bezüglich der CPU-Nutzung als auch bezüglich der Nutzung des Plattenplatzes sind die hierfür notwendigen Rechnungen ressourcenbeschränkt. Die Arbeiten für diese Problemklasse soll von den Antragstellern Tim Ruhe und Katharina Morik mit geeigneten Mitarbeitern durchgeführt werden.

Beiden Problemfelder gemeinsam ist die Frage nach den Einsparungsmöglichkeiten durch GPU-Computing. An den entsprechenden Untersuchungen sind alle Projektleiter beteiligt.

Arbeitsprogramm

Basierend auf der Datenaufzeichnung (AP1) ergibt sich eine enge Verflechtung der einzelnen Arbeitspakete. Ergebnisse und Verbesserungen in einem Bereich des Teilprojektes wirken sich somit sofort auch positiv in anderen Bereichen aus. Trotz der engen Verzahnung der Arbeitspakete können diese aufgrund der bestehenden Infrastruktur (Teleskope und Monte-Carlo-Simulationen) dennoch weitgehend unabhängig voneinander bearbeitet werden. Die einzelnen Arbeitspakete sollen im Folgenden detailliert vorgestellt werden.

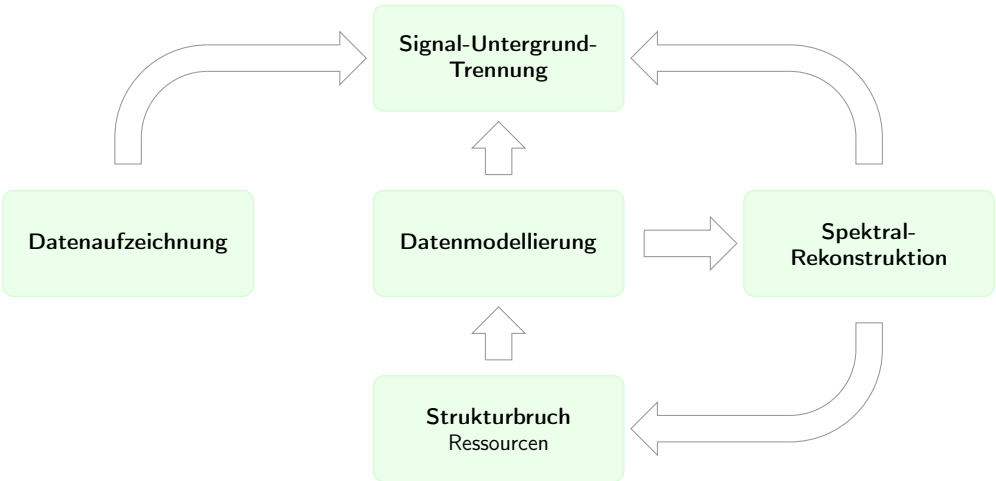


Abbildung 3.1.: Arbeitspakete.

Arbeitspaket 1. Datenaufzeichnung: Die Aufzeichnung astrophysikalischer Ereignisse mit verschiedenartigen Teleskopen (MAGIC, FACT, IceCube) bildet die Grundlage des gesamten Teilprojektes. Innerhalb der Datennahmeelektronik des betrachteten Teleskops wird innerhalb von Mikrosekunden eine Triggerentscheidung darüber gefällt, ob das letzte detektierte Ereignis in den Analyseprozess eingespeist oder verworfen werden soll. Diese erste Entscheidung wird mit Hilfe konfigurierbarer eingebetteter Rechnersysteme (FPGAs oder ASICs) getroffen. Ziel dieser Entscheidung ist es eine maximale Anzahl physikalisch relevanter Ereignisse für weitere Analysen zurückzubehalten. Die Art des verwendeten Triggers hängt dabei von zwei entscheidenden Faktoren ab. Der erste Faktor ist die Geschwindigkeit, mit der die Ereignisse nach der ersten Triggerentscheidung weiterverarbeitet werden können. Ist diese Zeit zu lang, müssen nachfolgende – unter Umständen interessantere – Ereignisse verworfen werden, da keine Rechnerkapazität zur Verarbeitung dieser Ereignisse zu Verfügung steht. Der zweite Faktor ist die Art der angestrebten Analyse, da sich die Anforderungen an die Ereignisse zwischen den einzelnen Analysen mitunter stark unterscheiden. Daraus ergeben sich zwei für dieses Arbeitspaket relevante Arbeitsschritte, welche signifikant zur Steigerung der Sensitivität der jeweiligen Teleskope beitragen können. Zum einen eine Verbesserung der Geschwindigkeit der Datenverarbeitung und zum anderen eine adaptive Anpassung der Triggerbedingungen an die jeweilige Datenanalyse. Für die Bearbeitung beider Teilaspekte kann auf Ergebnisse aus der ersten Förderperiode zurückgegriffen werden.

Um die Geschwindigkeit der Datenverarbeitung zu verbessern, sollen Streaming-Algorithmen systematisch eingesetzt und weiterentwickelt werden. Mit diesen Algorithmen kann eine hocheffiziente Verarbeitung der

Ereignisse bereits am Ort der Teleskope erfolgen. Bereits in der ersten Förderperiode konnte beispielhaft gezeigt werden, dass die für die Datenverarbeitung typischerweise zur Verfügung stehende Zeit von 10 ms um einen Faktor fünf bis acht verringert werden kann [A1/517]. Auf diese Weise können die zur Datenprozessierung notwendigen Ressourcen auf ein Minimum reduziert werden. So könnte im Rahmen des FACT-Teleskops beispielsweise auf den Betrieb einer kleinen Computer-Farm am Ort des Teleskops verzichtet werden.

Um Streaming-Algorithmen möglichst hochperformant am Ort der Teleskope einsetzen zu können, sollen, parallel zur Verbesserung des Streamings selbst, in der ersten Phase erprobte Lernalgorithmen in eine Streaming-Umgebung übertragen werden. Die Verwendung von Lernalgorithmen zur Triggerentscheidung ermöglicht eine verbesserte Erkennung physikalisch relevanter Ereignisse, woraus sich wiederum eine erhöhte Effizienz der Teleskope ergibt. Darüber hinaus kann auf diese Weise das Datenvolumen, welches zum Teil per Satellit an die jeweiligen Institute weltweit kommuniziert werden muss, drastisch verringert werden. Dieser Aspekt ist vor allem vor dem Hintergrund der Datennahme an entlegenen Orten, wie dem Südpol, wo die Datenübertragung durch Übertragungskapazität und Kosten limitiert ist, relevant. Vor diesem Hintergrund soll ebenfalls untersucht werden, bis zu welchem Anteil Untergrundereignisse bereits frühzeitig in der Analysekette verworfen werden können ohne Einbußen in der Sensitivität zu erleiden.

Neben der eigentlichen Übertragung der Algorithmen in eine Streaming-Umgebung erfordert dieser Arbeitsschritt auch die akkurate Auswahl und Validierung eines geeigneten Lernverfahrens. In der zweiten Phase sollen daher systematisch Lernverfahren im Hinblick auf ihre Performanz auf teleskopnahen Daten untersucht werden. Hier gilt es insbesondere die für das Lernen relevanten Attribute zu finden und ggf. zu konstruieren. Dieser Arbeitsschritt ist eng mit Arbeitspaket 2 verzahnt, dennoch ergeben sich Unterschiede: Neben der Performanz des Algorithmus steht bei der Anwendbarkeit auf detektornahen Daten vor allem die Laufzeit im Vordergrund. An dieser Stelle soll eine optimale Vereinbarkeit von Performanz und Laufzeit geschaffen werden um die gesamte Datenstromverarbeitung möglichst effizient zu gestalten.

Ist das Ziel einer bestimmten Messung, zum Beispiel die Rekonstruktion eines Energiespektrums, von vornherein bekannt, kann die gesamte detektornahe Verarbeitung der Daten an das Ziel dieser Analyse angepasst werden. Dazu sollen neben den in den Streaming-Prozessen variierbaren Parametern auch die hardwarenahen Schwellwerte, auf denen die Triggerentscheidung üblicherweise basiert, angepasst werden. Ziel ist es zu untersuchen, in welchem Maße die Sensitivität einer Analyse durch eine optimale Wahl der detektornahen Datenverarbeitung verbessert werden kann. Dabei soll auch überprüft werden, ob und wie eine Umstellung der Datenverarbeitung im laufenden Betrieb der Teleskope integriert werden kann. Die Bearbeitung dieses Teilaspektes baut eng auf Erkenntnissen aus den Arbeitspaketen 2 und 4 auf. Auf Erkenntnisse aus der ersten Förderperiode kann ebenfalls aufgebaut werden.

Da andere Arbeitspakete direkt von den in der Datenaufzeichnung erreichten Verbesserungen profitieren, soll unverzüglich mit der Bearbeitung derselben begonnen werden. Hierbei muss allerdings angemerkt werden, dass mögliche Verzögerungen innerhalb dieses Arbeitspaketes nur geringe Auswirkungen auf die anderen Arbeitspakete haben werden, da alle drei Teleskope auch ohne die hier angestrebte Verbesserung der Datenaufzeichnung weiter kontinuierlich Daten nehmen werden. Neue Erkenntnisse aus anderen Arbeitspaketen, insbesondere aus dem Bereich der Signal-Untergrund-Separation und der Strukturbrucherkennung wirken sich direkt auf die Datennahme aus und sollen kontinuierlich mit einbezogen werden.

Arbeitspaket 2. Signal-Untergrund-Separation: Die Arbeit an der Signal-Untergrund-Separation auf hohem Abstraktionsniveau der Analyse soll unter Einbeziehung der in der ersten Phase des SFB gewonnenen hervorragenden Ergebnisse konsequent fortgeführt werden. Insbesondere sollen die in der ersten Phase auf *High-Level*-Daten erarbeiteten Methoden für detektornahe *Low-Level*-Daten erweitert werden. In der ersten Phase des SFB durchgeführte Studien zeigen, dass die Anzahl der detektierten Signalereignisse durch den frühzeitigen und detektornahen Einsatz maschineller Lernmethoden noch einmal um einen Faktor 3 verbessert werden kann [C3/538]. Dies bedeutet weiterhin, dass das Signal-zu-Rausch-Verhältnis sich von bisher 10^{-3} auf etwa 10^{-6} verringert.

Die Trennung physikalisch relevanter Ereignisse vom Rauschen soll dabei mit den in der ersten Phase erfolgreich erprobten baumbasierten Methoden durchgeführt werden. Da in der ersten Phase ebenfalls gezeigt werden konnte, dass eine erfolgreiche Trennung wesentlich von einer qualitativ hochwertigen Merkmalsselektion abhängt, stellt die Merkmalsselektion einen weiteren zentralen Punkt des Arbeitspaketes dar. Auf der Basis des bereits erprobten MRMR-Algorithmus zur Auswahl relevanter Trennparameter sollen in der zweiten Phase Selektionsmethoden für detektornahe (und in der Regel weniger gut simulierte) *Low-Level*-Daten entwickelt werden. Bei allen zur Separation und Merkmalsauswahl angewandten Methoden muss berücksichtigt werden, dass die Anforderungen an die neuen Methoden auch aufgrund des signifikant schlechteren Signal-zu-Rausch-Verhältnisses ungleich höher sind.

Die für IceCube entwickelte Methodik soll in der zweiten Förderperiode auch auf das Cherenkovtele-

skop FACT übertragen werden. Erste Tests und Vergleiche zu einer Separation mit von Hand angepassten Schnitten zeigen, dass eine deutliche Erhöhung in der Rate der Signalereignisse erwartet werden kann. An der Verbesserung der Monte-Carlo-Beschreibung und der Implementierung der relevanten Parameter in die FACT tools wird intensiv gearbeitet. Die im Rahmen der Signal-Untergrund-Separation zurückbehaltenen Ereignisse sollen in einem zweiten Arbeitsschritt zur Rekonstruktion von Energiespektren verwendet werden. Hier wird aufgrund der größeren Ereignisanzahl eine deutliche Verbesserung der Signifikanz erwartet.

Für alle drei beispielhaft betrachteten Teleskope (MAGIC, FACT, IceCube) gilt, dass die für *Low-Level*-Daten zu nutzenden Merkmale weder notwendigerweise bekannt noch vollständig rekonstruiert sind. An dieser Stelle soll angesetzt werden, um die Separation durch die Konstruktion geeigneter Merkmale weiter zu verbessern.

Neben der konsequenten Fortführung der Signal-Untergrundseparation als Zweiklassenproblem (Myonneutrino/Untergrund, Gamma/Untergrund), soll der in [C3/546] entwickelte Analyseprozess auch für die Detektion von Elektronneutrinos eingesetzt werden. Aufgrund ihrer physikalischen Eigenschaften wird die Erkennung von Elektronneutrino-Ereignissen nicht nur durch ein hochfrequentes Rauschen, sondern insbesondere auch durch Myonneutrinos aus Wechselwirkungen mit sog. neutralen Strömen erschwert [503]. Schlecht rekonstruierte Myonneutrinos aus Wechselwirkungen mit sog. geladenen Strömen bilden einen weiteren Untergrundkanal. Das in der ersten Förderperiode untersuchte und sehr gut verstandene Zweiklassenproblem wird somit im Falle der Elektronneutrinos zu einem Mehrklassenproblem. Mittelfristiges Ziel dieses Teilaspektes im Arbeitspaket ist die erfolgreiche Separation von Myonneutrinos aus verschiedenen Untergrundklassen sowie die Entfaltung des Energiespektrums der Elektronneutrinos bei mittleren und hohen Energien.

Ein weiterer möglicher Detektionskanal in IceCube sind Tauneutrinos. Tauneutrinos konnten bisher mit IceCube noch nicht nachgewiesen werden [506]. Es wird jedoch erwartet, dass sie aufgrund ihrer einzigartigen Signatur eindeutig als solche identifiziert werden können. In der zweiten Förderperiode soll untersucht werden, ob bei der Suche nach Tauneutrinos mit IceCube durch den Einsatz maschineller Lernmethoden eine Verbesserung der Sensitivität erreicht werden kann. Diese Studie soll zunächst mit Hilfe simulierter Ereignisse durchgeführt und anschließend auf experimentellen Daten verifiziert werden. Dabei soll untersucht werden, ob Tauneutrinos insbesondere durch eine Outlier-Detection nachgewiesen werden können. Die Anwendung einer Outlier-Detection auf IceCube-Daten bietet neben der Möglichkeit der Detektion von Tauneutrinos auch die Möglichkeit exotische Teilchen (WIMP-Kandidaten, SUSY-Teilchen) zu detektieren. Dies betrifft bei einer *Outlier- Detection* insbesondere auch nicht simulierte Ereignisse. Die Detektion nicht simulierter Ereignisse in IceCube impliziert möglicherweise bahnbrechende Erkenntnisse für Astrophysik, Teilchenphysik und Kosmologie.

Neben der In-Ice-Komponente des Detektors enthält IceCube auch eine als IceTop bezeichnete Oberflächenkomponente [505]. Das Detektionsprinzip von IceTop entspricht dem der In-Ice-Komponente, jedoch wird IceTop zur Detektion von Luftschauern in der Erdatmosphäre benutzt. Anhand dieser Schauer kann dann das Energiespektrum der kosmischen Strahlung sowohl global, als auch für die einzelnen Komponenten (Protonen, leichte Kerne, schwere Kerne) bestimmt werden. Da im Falle von IceTop ausreichend Statistik vorhanden ist, liegt die Herausforderung vor allem in der korrekten Zuordnung einzelner Ereignisse zu bestimmten Gruppen von Atomkernen. Bei der Analyse mit herkömmlichen Methoden (Schnitte) ist diese Zuordnung mit großen Unsicherheiten behaftet, was wiederum zu großen Unsicherheiten für die Zusammensetzung der kosmischen Strahlung führt (siehe z.B. [515]). Im Rahmen der zweiten Förderperiode soll die Klassifikation von Ereignissen der kosmischen Strahlung durch den gezielten Einsatz maschineller Lernmethoden verbessert werden. Bei der Lösung dieses Multiklassenproblems kann erneut auf die Erfahrungen bei der Klassifikation von Zweiklassenproblemen zurückgegriffen werden. Es wird erwartet, dass durch die Klassifikation von Ereignissen der kosmischen Strahlung insbesondere in Kombination mit einer Rekonstruktion des Spektrums mit DSEA/SpectEst eine signifikante Verbesserung im Hinblick auf die Zusammensetzung der kosmischen Strahlung erzielt werden kann.

Das zweite Arbeitspaket behandelt Analyseschritte, die in der Abfolge der Analyse auf die im ersten Paket diskutierten Schritte (Trigger, Streaming) folgen. Logisch hängt die Frage, welche Ereignisse aufgezeichnet werden sollen, davon ab, welche Ereignisse weiter analysierbar sind. Wegen der Bearbeitung der Arbeitspakete am Beispiel unterschiedlicher physikalischer Experimente können sie dennoch weitgehend parallel bearbeitet werden.

Arbeitspaket 3. Strukturbrucherkennung unter Ressourcenbeschränkung: Für das physikalische Verständnis astrophysikalischer Quellen ist ein variables spektrales Verhalten derselben von besonderem Interesse. Variabilitäten deuten auf physikalische Grenzfälle hin, deren Analyse zu einem verbesserten Verständnis der Struktur der Quellen führen kann. Das Auffinden und die Analyse von Quellvariabilitäten in den Daten ist somit von entscheidender Bedeutung. Dies betrifft sowohl die Datenanalyse, welche sich mit dem Auffinden von

Variabilitäten beschäftigt, als auch die Hardwareseite, von der aus auf mögliche irreguläre Daten innerhalb von Sekunden reagiert werden muss. Die Ressourcen der Analysen sind im Falle der Gammaastronomie vor allem durch die limitierte Beobachtungszeit pro Quelle beschränkt. Aus dieser witterungsbedingt limitierten Beobachtungszeit ergeben sich wiederum nicht-äquidistante Beobachtungszeitpunkte, welche zusammen mit den sehr unterschiedlichen Skalen der Variationen (Minuten bis Wochen) eine zusätzliche Herausforderung für diese Analysen bilden. Im Rahmen der Neutrinoastronomie findet zwar eine kontinuierliche und äquidistante Beobachtung des gesamten Himmels statt, jedoch sind die Analysen durch die limitierte Anzahl an Ereignissen in ihren Möglichkeiten beschränkt.

Zeitliche Variationen in astrophysikalischen Quellen reichen von plötzlichen Flares (einige Minuten) bis hin zu Änderungen des Flusses, welche sich über einen Zeitraum von Wochen oder gar Monaten erstrecken können. Wurden in der ersten Phase vor allem langfristige Änderungen astrophysikalischer Flüsse und deren Auffinden bei witterungsbedingt nicht äquidistanten Daten untersucht, so sollen im Rahmen der zweiten Förderperiode auch kurzfristige Änderungen der Flüsse betrachtet werden. Im Bereich der Datenanalyse wurde die Erkennung von Änderungen in den Datenströmen in [529] erstmals formalisiert. Darüber hinaus sind verschiedene Methoden zur *Change Detection* bzw. der Erkennung von *Concept Drifts/Concept Shifts* entwickelt worden [6]. Zur Untersuchung der zeitlichen Variationen in den physikalischen Quellen sollen verschiedene Algorithmen aus dem Bereich des Concept Drift/Concept Shift in die FACT tools integriert und auf ihre Performanz und Eignung bezüglich der Teleskop-Daten untersucht werden. Ziel ist es dabei, innerhalb kürzester Zeit auf verschiedenste Variabilitäten angemessen reagieren zu können. Dies beinhaltet sowohl eine Steigerung der Datennahmerate im Falle interessanter astrophysikalischer Ereignisse (z.B. Flares) als auch eine mögliche Notabschaltung des Teleskops.

Neben Neutrinos und Gammas aus astrophysikalischen Quellen ist auch der atmosphärische Neutrinofluss für die Strukturbrucherkennung interessant. Zwar existiert eine Vielzahl von theoretischen Modellen, welche den atmosphärischen Neutrinofluss als Funktion des Zenitwinkels beschreiben (siehe z.B. [526]), jedoch fehlen aufgrund der limitierten Ereignisanzahl belastbare experimentelle Daten. Die Messung des atmosphärischen Neutrinoflusses ist eine der Kernkompetenzen des Teilprojektes (siehe dazu auch [522] und [C3/535]). Diese soll in der zweiten Förderperiode weiter ausgebaut werden. Zu diesem Zweck soll der atmosphärische Neutrinofluss gleichzeitig als Funktion der Energie *und* des Zenitwinkels gemessen und auf Strukturbrüche untersucht werden. Mit der Entwicklung einer hochperformanten Ereignisselektion (AP2) und des Entfaltungsalgorithmus DSEA/SpectEst (AP4) wurden innerhalb der ersten Förderperiode die entscheidenden Voraussetzungen für Analysen dieser Art geschaffen. Es wird erwartet, dass die gewonnenen Erkenntnisse zu einer erheblichen Verbesserung der theoretischen Neutrinoflussmodelle führen.

Neben der räumlichen Variabilität ist auch die zeitliche Variabilität des atmosphärischen Neutrinoflusses von Interesse. Aufgrund der begrenzten Anzahl an Ereignissen konnte bislang lediglich die Änderung der Ereignisrate als Funktion der Zeit, nicht aber Änderungen im Spektrum selbst untersucht werden [501]. Unter Zuhilfenahme der in der ersten Phase des SFB entwickelten Methode DSEA/SpectEst wird nun auch die Form des Spektrums als Funktion der Zeit untersucht werden. Auch hier sollen Methoden aus dem Bereich des Concept Shift/Concept Drift eingesetzt werden. Dabei gilt es unter anderem auch die Zeitskalen der Änderungen zu untersuchen. Hierbei ist vor allem von Interesse, ob der atmosphärische Neutrinofluss einer kontinuierlichen Änderung im Laufe der Jahreszeiten unterworfen ist, oder ob die Änderungen plötzlich und mit dem Einsetzen bestimmter atmosphärischer Bedingungen auftreten. Zu diesem Zweck sollen bestimmte Zeitfenster mit Hilfe von Ähnlichkeitsmaßen wie der Kullback-Leibler-Divergenz untersucht werden.

Da die jahreszeitlich wechselnden atmosphärischen Bedingungen ebenfalls mit dem Zeitpunkt der Beobachtung korrelieren, ermöglichen derartige Untersuchungen Rückschlüsse auf das Zusammenspiel von Erdatmosphäre und kosmischer Strahlung. Diese sind sowohl für die Neutrinophysik als auch für die Gammaastronomie (Verständnis des Untergrundes) von Interesse und liefern wichtigen *Input* für die Simulationsalgorithmen.

Arbeitspaket 4. Spektral-Rekonstruktion: Die Rekonstruktion von Verteilungen (z.B. Energiespektren) ist für die Extraktion physikalischer Erkenntnisse aus den vorgegebenen Daten von entscheidender Bedeutung, da auf diese Weise Rückschlüsse auf die Physik der Quellen ermöglicht werden. Zu diesem Zweck soll das in der ersten Phase des SFB entwickelte Verfahren zur Spektral-Rekonstruktion (DSEA/SpectEst) angewandt und weiter verbessert werden. Studien zeigen, dass die Anzahl der benötigten Monte-Carlo-Ereignisse durch den Einsatz von DSEA/SpectEst bei gleichbleibender Auflösung um einen Faktor fünf verringert werden kann [C3/542]. Ferner wird durch den Einsatz von DSEA/SpectEst erreicht, dass bei einer Analyse nur noch eine einzelne Entfaltung durchgeführt werden muss (auch wenn verschiedene Verteilungen rekonstruiert werden sollen), da durch die andersartige Herangehensweise sämtliche Informationen über die einzelnen Ereignisse erhalten bleiben. Auf diese Weise kann die für eine Spektralanalyse benötigte Arbeitszeit drastisch reduziert werden.

Durch den Einsatz dieser auf maschinellem Lernen basierenden Methode kann die Ressourcenbeschränkung bestehender Algorithmen weitgehend aufgehoben werden. Zusätzlich können durch den Einsatz von DSEA/SpectEst auch wenige oder sogar einzelne Ereignisse entfaltet werden. Dies bietet vor allem vor dem Hintergrund der aktuell durch IceCube entdeckten astrophysikalischen Neutrinos die Möglichkeit zur detaillierten Untersuchung derselben sowie zur detaillierten spektralen Untersuchung potentieller Neutrinoquellen. Die Anwendung von DSEA/SpectEst ermöglicht eine einfache Untersuchung periodischer Änderungen in Energiespektren und leistet somit einen signifikanten Beitrag bei der Erkennung von Strukturbrüchen. Dieses Potential soll sowohl im Bereich der Gammaastronomie als auch im Bereich der Neutrinopunktquellen genutzt werden.

In ersten Simulationen führt der Algorithmus zu Ergebnissen, die denen von TRUEE vergleichbar sind. Das Verfahren muss jedoch iterativ verwendet werden, um eine missspezifizierte Trainingsverteilung auszugleichen. Dies ermöglicht dem Benutzer jedoch, die Analyse mit einer weniger den Daten angepassten Trainingsverteilung zu beginnen. Die Abhängigkeit des Algorithmus von der im Training des Lerners verwendeten Verteilung soll in der zweiten Antragsphase im Detail untersucht werden.

Alternative Klassifikationsalgorithmen können in das neue Framework ohne Schwierigkeiten integriert werden. Es muss untersucht werden, welche Fluktuationen nach der Iteration verbleiben und wie sie unterdrückt werden können (Regularisierung). Schließlich werden konsistente Schätzer für die Varianz der Resultate entwickelt, um asymptotische Konfidenzintervalle zu erhalten. Die Anwendung unterschiedlicher Klassifikationsalgorithmen soll in der zweiten Phase systematisch untersucht werden, um DSEA/SpectEst möglichst ressourcenschonend einsetzen zu können.

Durch die Anwendung von DSEA/SpectEst auf simulierte Untergrundereignisse sollen ergebnisrelevante Strukturen in der Verteilung des Untergrundes ermittelt werden. Dadurch wird eine Subtraktion des Untergrundes in einem Sample, das eine Summe aus Signal und Untergrundereignissen enthält, möglich. Auf diese Weise können die bislang äußerst strikten Kriterien bei der Trennung von Ereignissen gelockert werden, wodurch wiederum eine größere Anzahl an physikalisch relevanten Ereignissen zur Verfügung steht. Ferner soll mit Hilfe von DSEA/SpectEst erforscht werden, bis zu welchem Grad eine Lockerung der Selektionskriterien möglich ist ohne Einbußen in Bezug auf die Genauigkeit der rekonstruierten Verteilung zu erhalten.

Neben der beispielhaften Anwendung der Methode auf Daten der Cherenkovteleskope MAGIC und FACT sowie auf Daten des Neutrinoteleskops IceCube ist die kontinuierliche Verbesserung der Methodik selbst ein Kernpunkt dieses Arbeitspaketes. Hierbei soll vor allem der Einfluss der drei zentralen Elemente von DSEA/SpectEst (Merkmalsselektion, Diskretisierung, Lerner) auf die Qualität des entfalteten Spektrums untersucht werden. Darüber hinaus soll erforscht werden, welche methodischen Verbesserungen durch die gezielte Konstruktion von Merkmalen aus Rohdaten erreicht werden können.

Durch Nutzung von Streaming-Algorithmen (vergl. Arbeitspaket 1) soll darüber hinaus eine Entfaltung von Daten online und am Ort der Teleskope durchgeführt werden. Neben den Eigenschaften eines Energiespektrums für hochenergetische Gammas und Neutrinos als solches ist auch die Änderung des Spektrums mit einer zweiten, unter Umständen ebenfalls experimentell nicht direkt zugänglichen Größe physikalisch interessant. Die simultane spektrale Rekonstruktion von zwei oder mehr Größen ist mit den bisherigen Methoden nicht möglich. Der vielversprechende Ansatz, ein inverses Problem als Multiklassenproblem zu behandeln, soll in der zweiten Phase des SFB auf mehrdimensionale Spektren erweitert werden, um auf diese Weise die Untersuchung der Spektren auf eventuelle Veränderungen oder Brüche hin untersuchen zu können. Dabei kann prinzipiell der gleiche Ansatz verfolgt werden wie bei der Rekonstruktion eindimensionaler Spektren. An dieser Stelle muss jedoch betont werden, dass die Anforderungen an die eingesetzten Algorithmen im Vergleich zum eindimensionalen Fall ungleich höher sind, da die Anzahl der Klassen exponentiell mit der Dimension des inversen Problems ansteigt. Im Rahmen der zweiten Förderperiode soll daher auch untersucht werden, inwiefern bestehende und für die eindimensionale Rekonstruktion des Spektrums hervorragend geeignete Algorithmen auch für die mehrdimensional Spektralrekonstruktion geeignet sind.

Arbeitspaket 5. Datenmodellierung und -simulation: Sämtliche im Teilprojekt C3 beschriebenen Arbeitspakete sind auf eine Monte-Carlo-Beschreibung der Daten angewiesen. Hierbei muss jedoch beachtet werden, dass die Generierungsrate simulierter Ereignisse etwa um einen Faktor 100 bis 1000 kleiner ist als die Rate der tatsächlich aufgenommenen Daten.

Im Rahmen der ersten Förderperiode wurde exemplarisch das Simulationsprogramm PROPOSAL (Propagator with Optimized Speed and Optimized Precision for all Leptons) entwickelt. Durch die Entwicklung neuer Methoden wurde das Programm darüber hinaus laufzeitoptimiert und erfolgreich in die bestehende IceCube-Simulationskette eingefügt. Zum gegenwärtigen Zeitpunkt ist PROPOSAL das Standardprogramm für die Propagation von Teilchen in IceCube. Innerhalb des Programms werden die von den Primärteilchen erzeugten Sekundärteilchen durch das Detektionsmedium propagiert. Die dabei für die jeweiligen Sekundär-

teilchen auszuführenden Rechenschritte ähneln sich stark. Derartige Algorithmen sind für den Einsatz im Bereich des GPGPU-Programming prädestiniert.

PROPOSAL soll daher im Rahmen der zweiten Förderperiode vollständig auf GPGPUs programmiert werden. In der ersten Förderperiode wurde bereits ein effektiver Zufallszahlengenerator mit einer langen Periodendauer für die Nutzung auf GPGPUs erfolgreich implementiert und getestet. Auf diese Entwicklung kann im Rahmen der zweiten Förderperiode aufgebaut werden. Es wird erwartet, dass auf diese Weise die Leistungsaufnahme und somit auch die Betriebskosten um bis zu 98% reduziert werden können (Masterarbeit Tomasz Fuchs). Dieses Einsparpotential kann insbesondere bei der numerischen Integration und bei der parallelen Propagation der Teilchen genutzt werden. Die im Bereich der GPGPU-Programmierung bestehende Zusammenarbeit mit Teilprojekt A1 soll zu diesem Zweck beibehalten und ausgebaut werden. Die auf GPGPUs implementierte Version von PROPOSAL soll ebenfalls in die IceCube-Simulationskette eingefügt werden.

Einen weiteren Ansatzpunkt für den ressourcenschonenden Einsatz von Monte-Carlo-Simulationen bieten adaptive Lernalgorithmen. Innerhalb dieser Lernalgorithmen werden simulierte Ereignisse mit bestimmten Merkmalen bei Bedarf und nach den Vorgaben des Lerners simuliert. Dies bietet den Vorteil, dass Ereignisse nur noch dann produziert werden, wenn diese auch tatsächlich vom Lerner benötigt werden. Die ressourcenintensive Produktion einer Vielzahl von Ereignissen aus einem für die jeweilige Analyse irrelevanten Energiebereich kann auf diese Weise vermieden werden. Derartige Ansätze wurden bereits im Bereich des *Active Learning* untersucht, um einem hohen Kostenaufwand bei der Erzeugung annotierter Beispiele entgegen zu wirken [534]. Innerhalb der zweiten Förderperiode soll daher untersucht werden, inwiefern die zugrunde liegenden Probleme aus der Astroteilchenphysik für den Einsatz von adaptiven Lernverfahren geeignet sind. Weiterhin soll geprüft werden, wie adaptive Lernverfahren in die Simulationsketten der einzelnen Experimente eingefügt werden können.

Das von Dominik Neise entwickelte Simulationsprogramm JAFS konnte innerhalb der ersten Förderperiode bereits erfolgreich für Tests mit DSEA/SpectEst eingesetzt werden. Da JAFS die Eigenschaften eines atmosphärischen Cherenkovteleskops mit hinreichender Genauigkeit, aber mit signifikant verkürzter Laufzeit simuliert, eignet es sich hervorragend zum Rapid Prototyping. Auf diese Weise kann das Ansprechverhalten des Teleskops auf geänderte Bedingungen schnell und ressourcenschonend untersucht werden. Das in JAFS exemplarisch umgesetzte Konzept des *Rapid Prototyping* soll im Rahmen der zweiten Förderperiode auch auf IceCube übertragen werden. Ziel ist es, systematische Studien schneller und mit einem erheblich geringeren Ressourcenaufwand betreiben zu können. Im Bereich des adaptiven Lernens soll JAFS als Orakel verwendet werden.

Ziel dieses Arbeitspaketes ist es, die Simulation von Ereignissen bei gleichbleibender oder besserer Qualität der Resultate zeit- und speichereffizienter zu gestalten. Die Bearbeitung des Arbeitspaketes kann parallel zu den anderen Arbeitspaketen erfolgen.

Zeitplan

	2015				2016				2017				2018			
Arbeitspaket / Quartal	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Datenaufzeichnung:	■	■	■	■	■	■	■	■	■	■						
2. Signal-Untergrund-Separation:			■	■	■	■	■	■	■	■	■	■	■	■	■	■
3. Strukturbrucherkennung unter Ressourcenbeschränkung:					■	■	■	■	■	■						
4. Spektral-Rekonstruktion:			■	■	■	■	■	■	■	■	■	■	■	■	■	■
5. Datenmodellierung und -simulation:											■	■	■	■	■	■

3.5 Stellung innerhalb des Sonderforschungsbereichs

Innerhalb dieses Teilprojektes steht Information im Vordergrund, die auf Messsystemen (Teleskopen) mit einer hohen Datennahmerate erzeugt wird. Diese Information wird in der Regel bereits am Ort der Daten-

nahme durch z.T. integrierte Rechner digitalisiert, weiterverarbeitet und kommuniziert. Die Beschränkung der Ressourcen besteht dabei direkt in der verfügbaren Energie und in der Bandbreite, mit welcher die Daten kommuniziert werden können. Indirekt wirkt sich die Beschränkung somit auch auf den verfügbaren Speicher und die zur Verfügung stehende CPU-Zeit aus.

Das Teilprojekt C3 ist einer der Lieferanten von *Big Data* innerhalb des Sonderforschungsbereiches und wurde in verschiedenen Arbeitsgruppensitzungen der Arbeitskreise *Datenanalyse* und *Ressourcen* als Kristallisationspunkt unterschiedlicher Methoden besprochen. Aufgrund der Stellung als Lieferant großer Datenmengen bietet sich im Rahmen des Teilprojektes die Möglichkeit zur Kooperation mit verschiedensten anderen Teilprojekten, sowohl im Rahmen der verwendeten Analyseverfahren als auch auf methodischer Ebene.

Die im Teilprojekt A1 entwickelten Lösungsansätze konnten bereits in der ersten Förderperiode außerordentlich erfolgreich für die in C3 bestehenden Herausforderungen angepasst und angewendet werden. Durch die einfache Übertragbarkeit der entwickelten Analyseprozesse auf einzelne astrophysikalische Fragestellungen konnte die Generalisierbarkeit der in A1 entwickelten Lösungsansätze gezeigt werden. An dieser Stelle sind insbesondere die FACT tools, DSEA/SpectEst und der gesamte Bereich der Signal-Untergrund-Separation zu nennen.

Die außerordentlich erfolgreiche und produktive Zusammenarbeit mit dem Teilprojekt A1 soll daher folgerichtig in der zweiten Phase des SFB fortgesetzt und ausgebaut werden. Innerhalb dieser Zusammenarbeit soll der erfolgreiche interdisziplinäre Austausch auf dem Gebiet der Signal-Untergrund-Separation fortgeführt und um die Komponente der Low-Level-Selektion von Ereignissen erweitert werden. Ebenso sollen in enger Kooperation der beiden Teilprojekte Algorithmen für die Suche nach extrem seltenen Ereignissen entwickelt und am Beispiel der Astroteilchenphysik erprobt werden. Ebenso soll die Entwicklung von DSEA/SpectEst konsequent fortgeführt werden. Darüber hinaus soll der Algorithmus selbst an einer Vielzahl astrophysikalischer Fragestellungen überprüft und validiert werden. Das in A1 entwickelte *streams framework* soll, ähnlich wie im Falle der FACT tools auch auf das IceCube-Neutrinoteleskop übertragen werden. Es bildet an dieser Stelle den zentralen technischen Ansatz bei der Entwicklung eines vollständigen *online*-Analyseprozesses. Die Beteiligung von Katharina Morik am Teilprojekt C3 bietet eine hervorragende Gelegenheit zu einer solchen intensivierten Zusammenarbeit. Es wird erwartet, dass beide Teilprojekte auch weiterhin erheblich von einem aktiven Austausch profitieren werden.

Im Rahmen einer engen Zusammenarbeit mit dem Teilprojekt A2 wurde der *k-means*-Algorithmus erfolgreich auf Daten des Neutrinoteleskops IceCube angewendet. Die Anwendung geschah dabei sowohl auf simulierte als auch auf echte Ereignisse (Bachelorarbeit Michael Bulinski). Die auf diese Weise identifizierten Cluster wurden sowohl in Bezug auf ihre Reinheit als auch in Bezug auf ihre Energieabhängigkeit untersucht. Ebenso wie die Zusammenarbeit mit Teilprojekt A1 soll auch der Austausch mit Teilprojekt A2 in der nächsten Förderperiode fortgesetzt werden. Dabei soll der *k-means*-Algorithmus auch auf andere Datensätze aus dem Bereich der Astroteilchenphysik (Daten der Cherenkov-Teleskope MAGIC und FACT) angewendet werden. Ebenso soll die Anwendbarkeit des Algorithmus für die Suche nach extrem seltenen, möglicherweise bisher unbekannten und somit auch nicht simulierten Ereignissen überprüft werden. Für die Zusammenarbeit mit A2 wird auch weiterhin ein hohes Maß an für beide Seiten gewinnbringenden Erkenntnissen erwartet.

Im Rahmen der Bachelorarbeit von Dominik Baack wurde in Zusammenarbeit mit dem Teilprojekt B2 ein aktives Konturverfahren zur Bilderkennung bei Luftschauern in FACT entwickelt. Durch diese Anwendung der in B2 entwickelten Techniken kann die bislang bestehende Beschränkung auf ein elliptisches Schauerabbild weitgehend aufgehoben werden, wodurch sich eine deutliche generellere und somit auch akkuratere Beschreibung der Luftschauer ergibt. Im Rahmen der zweiten Förderperiode soll – ebenfalls in Zusammenarbeit mit B2 – überprüft werden, inwieweit sich durch zusätzliche Verwendung der Zeitinformationen eine Verbesserung der Signal-Untergrund-Separation erreichen lässt. Im Rahmen einer Kooperation der drei Teilprojekte A1, B2 und C3 soll der von Herrn Baack entwickelte Algorithmus in die FACT tools integriert werden. Es wird erwartet, dass auch die Kooperation mit B2 für beide Teilprojekte weiterhin außerordentlich erfolgreich verläuft und einen hohen Erkenntnisgewinn für beide Teilprojekte liefert.

Ein enger Austausch ergab sich auch mit Teilprojekt C1, welcher vor allem durch die hohe Dimensionalität und die daraus folgende Notwendigkeit einer akkuraten Merkmalsauswahl begründet ist. In beiden Teilprojekten konnte erfolgreich auf eine in A1 erfolgte Implementierung des MRMR-Algorithmus zurückgegriffen werden. Der produktive interdisziplinäre Austausch mit C1 und die praktische Anwendung geeigneter Algorithmen auf hochdimensionale Fragestellungen soll auch in der zweiten Förderperiode fortgesetzt werden. Es wird erwartet, dass beide Teilprojekte auch weiterhin erheblich von der Zusammenarbeit profitieren werden.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

Vista-TV

(Morik,Förderkennzeichen EU STREP 296126) (Förderzeitraum: 2012 - 2014)

Aufgabe des Projektes ist die Verbindung bestehender Verfahren zu einem neuartigen System für Open Linked Data. Das Projekt annotiert Internetfernsehprogramme automatisch und empfiehlt Benutzern darauf aufbauend Sendungen. Die Projektarbeit konzentrierte sich auf die Entwicklung einer Infrastruktur für die Verbesserung der Daten bei Zattoo und BBC und die Anwendung vorhandener Lernverfahren. Das in A1 entwickelte *streams framework* wurde erfolgreich im Projekt angewandt, um aus Videodaten Merkmale zu extrahieren.

Insight

(Morik,Förderkennzeichen EU 318225) (Förderzeitraum: 2012 - 2015)

Das Projekt behandelt Big Data in Bezug auf Volumen, Heterogenität und Realzeitanforderung. Wir arbeiten an raumzeitlichen Prozessen zur Vorhersage und Steuerung der Verkehrsflüsse in Dublin. Aus diesem Projekt wollen wir die Daten für das Teilprojekt B4 nehmen. Das streams framework wurde hier zur Ereigniserkennung eingesetzt.

Kobra

(Morik,Förderkennzeichen BMBF 01UG12445A) (Förderzeitraum: 2012 - 2015)

In diesem Projekt geht es um Sprachdaten und deren Analyse. Es hängt mit den Arbeiten im SFB 876 kaum zusammen.

Intelligente Prozess-Steuerung auf Basis von Prozessmerkmalen

(Morik,Förderkennzeichen SMS Siemag) (Förderzeitraum: 2011 - 2014)

Das Projekt wird direkt von SMS Siemag gefördert. Es wurde im ersten Jahr eine Anwendung des maschinellen Lernens für die Bestimmung des Blasendes bei einem BOF im Stahlwerk Dillinger Hütte entwickelt. Nach einer Pause wurde in einer Fortsetzung eine Infrastruktur zur Erfassung und Speicherung der Daten für das Lernen entwickelt. Hierbei wurde auch die Anwendung des streams frameworks vorgesehen.

Data Driven Material Development

(Morik,Förderkennzeichen Mercator Stiftung, Merkur Pr-2012-0023) (Förderzeitraum: 2012 - 2014)

Die Daten zur Untersuchung verschiedener Mischungen chemischer Substanzen in Bezug auf bestimmte Materialeigenschaften sollen effizienter genutzt werden, sodass aus wenigen Messpunkten die Eigenschaften an allen möglichen Messpunkten vorhergesagt werden können.

Verbundprojekt ICECUBE (Teilprojekt 5)

(Rhode,Förderkennzeichen BMBF, Deutsches Elektronensynchrotron (DESY), Projektträger für Astroteilchenphysik, BMBF grant 05A11PE1) (Förderzeitraum: 2011 - 2014)

Durch das Projekt wird der Aufbau sowie der Betrieb des Neutrino-Teleskops IceCube sowie die Entwicklung physikalischer Nachweismethoden auf dem Gebiet der Neutrinoastronomie gefördert. Von IceCube in diesem Zeitraum aufgezeichnete Daten werden im SFB 876 als Datensätze genutzt. Das Antrag auf das Nachfolgeprojekt (2014 - 2017) befindet sich zur Zeit in der Begutachtung.

Verbundprojekt MAGIC: MAGIC High Quantum Efficiency (Teilprojekt 2)

(Rhode,Förderkennzeichen BMBF, Deutsches Elektronensynchrotron (DESY), Projektträger für Astroteilchenphysik, BMBF grant 05A14PE1) (Förderzeitraum: 2014 - 2017)

Durch das Projekt wird der Aufbau sowie der Betrieb des Gamma-Teleskops MAGIC/FACT sowie die Entwicklung physikalischer Nachweismethoden auf dem Gebiet der Gamma-Astronomie gefördert. Von MAGIC/FACT in diesem Zeitraum aufgezeichnete Daten werden im SFB 876 als Datensätze genutzt. Das Antrag auf das Nachfolgeprojekt (2014 - 2017) befindet sich zur Zeit in der Begutachtung.

Verbundprojekt CTA: Softwareentwicklung für das CTA-Projekt

(Rhode,Förderkennzeichen BMBF, Deutsches Elektronensynchrotron (DESY), Projektträger für Astroteilchenphysik, BMBF grant 05A11PEA) (Förderzeitraum: 2011 - 2014)

Durch das Projekt wird die Durchführung von Monte-Carlo-Studien und die Entwicklung von Algorithmen für die Analyse simulierter Daten für das in Planung befindliche Cherenkov-Teleskop-Array CTA auf dem Gebiet der Gamma-Astronomie gefördert. Das Antrag auf das Nachfolgeprojekt (2014 - 2017) befindet sich zur Zeit in der Begutachtung.

Theoretische Untersuchung von Signaturen der kosmischen Strahlung in astrophysikalischen Hochenergieflares

(Rhode,Förderkennzeichen MERCUR, Vorhaben Pr-2012-0008) (Förderzeitraum: 2012 - 2014)

In dem Projekt wird die Berechnung von transienten astrophysikalischen Teilchenflüssen mit Methoden der theoretischen Physik gefördert.

Helmholtz Alliance for Astroparticle Physics (HAP) , Teilprojekt Dortmund

(Rhode,Förderkennzeichen Helmholtz-Gemeinschaft Deutscher Forschungszentren) (Förderzeitraum: 2011 - 2016)

Von der HAP-Allianz werden experimentübergreifende Techniken und Berechnungen vorangetrieben. In dem Projekt werden für Dortmund die numerische Berechnung von Teilchenpropagation in verschiedenen Medien mit Methoden der theoretischen Physik sowie Hardwareentwicklungen gefördert.

Rekonstruktion zeitvariabler Verteilungen in statistischen inversen Problemen

(Rhode,Förderkennzeichen Sonderforschungsbereich Statistik (SFB 823), Statistical modelling of nonlinear dynamic processes) (Förderzeitraum: 2013 - 2017)

Im Rahmen dieses Projektes wurde von der Arbeitsgruppe ein auf Tikhonov-Regularisierung basierendes Programm zur Lösung inverser Probleme (TRUEE) entwickelt. Dieses Programm wurde in dem SFB 876 zur Entfaltung der Energiespektren in der Gamma- und Neutrinoastronomie eingesetzt. Es soll im SFB 876 in der beantragten Förderperiode durch das im SFB 876 hier in Entwicklung befindliche Programm DSEA/SpectEst ersetzt werden.

Rekonstruktion zeitvariabler Verteilungen in statistischen inversen Problemen

(Rhode,Förderkennzeichen Sonderforschungsbereich Statistik (SFB 823), Statistical modelling of nonlinear dynamic processes) (Förderzeitraum: 2013 - 2017)

In dieser Förderperiode soll das Programm TRUEE weiter verbessert und zur Lösung verschiedener Probleme aus dem Gebiet der Ingenieurwissenschaften eingesetzt werden. Als Referenz steht es bei der DSEA/SpectEst-Entwicklung damit weiter zur Verfügung.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	3	176.400	3	176.400	3	176.400	3	176.400
Summe:	—	176.400	—	176.400	—	176.400	—	176.400
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	176.400		176.400		176.400		176.400	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (C3)							
Wissen-schaftliches Personal	1	Morik, Katharina, Prof. Dr., Prof	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	6,0	—	Grund-ausstattung
	2	Rhode, Wolfgang, Prof. Dr. Dr., Prof	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	10,0	—	Grund-ausstattung
	3	Ruhe, Tim, Dr., PostDoc	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	19,9	—	Grund-ausstattung
	4	Scheriau, Florian, Dipl.-Phys., WiMi	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	19,9	—	Grund-ausstattung
	5	Fuchs, Tomazs, M.Sc., WiMi	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	19,9	—	Mercator
	6	Schmitz, Martin, Dipl.-Phys., WiMi	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	19,9	—	Grund-ausstattung
	7	Temme, Fabian, Dipl.-Phys., WiMi	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	19,9	—	Helmholtz-Allianz
	8	Buss, Jens, Dipl.-Phys., WiMi	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	19,9	—	Grund-ausstattung
	9	Brügge, Kai, SHK	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	5,0	—	Grund-ausstattung

	10	Lotze, Moritz, B.Sc., SHK	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	10,0	—	Grundausstattung
	11	Menne, Thorben, B.Sc., SHK	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	5,0	—	Grundausstattung
Nichtwissenschaftliches Personal	12	Klingspor, Sabine, Verw. Ang.	—	LS 8, Fak. Inf, TU Do	4,97	—	Grundausstattung
	13	Greve, Andreas, Techn. Ang.	—	LS 8, Fak. Inf, TU Do	7,96	—	Grundausstattung
	14	Stickel , Britta, Verw. Ang.	—	LS Experimentelle Physik 5, Fakultät Physik, TU Do	7,96	—	Grundausstattung
	15	Warda, Kai, Techn. Ang.	—	LS Experimentelle Physik 5, Fakultät Physik, TU Do	7,96	—	Grundausstattung
	16	Domke, Matthias, Techn. Ang.	—	LS Experimentelle Physik 5, Fakultät Physik, TU Do	7,96	—	Grundausstattung
Beantragte Personalmittel (C3)							
Wissenschaftliches Personal	17	Overkemping, Ann-Kristin, Dipl.-Phys., WiMi	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	—	Doktorand/in	—
	18	Bockermann, Christian, Dipl.-Inf. WiMi	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	—	Doktorand/in	—
	19	N.N. , WiMi	Astroteilchenphysik	LS Experimentelle Physik 5, Fakultät Physik, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Morik, Katharina**
   Projektleitung. Schwerpunkt Data Mining. Mitarbeit allen Arbeitspaketen.

2. **Rhode, Wolfgang**
   Projektleitung. Schwerpunkt Astrophysik. Mitarbeit allen Arbeitspaketen sowie Kooperation mit den internationalen physikalischen Kollaborationen.

3. **Ruhe, Tim**
   Projektleitung. Schwerpunkt Neutrinoforschung. Mitarbeit in allen Arbeitspaketen.

4. **Scheriau, Florian**
   Mitarbeit im Arbeitspaket 2.

5. **Fuchs, Tomazs**
   Mitarbeit im Arbeitspaket 5.

6. **Schmitz, Martin**
   Mitarbeit im Arbeitspaket 3.

7. **Temme, Fabian**
   Mitarbeit im Arbeitspaket 1.

8. **Buss, Jens**
   Mitarbeit im Arbeitspaket 1.

9. **Brügge, Kai**
   Kai Brügge studiert sowohl Informatik als auch Physik und wird in beiden Disziplinen gemeinsam betreut bei der Software-Entwicklung für MAGIC und FACT.

10. **Lotze, Moritz**
    Mitarbeit im Arbeitspaket 2.

11. **Menne, Thorben**
    Mitarbeit im Arbeitspaket 5.

12. **Klingspor, Sabine**
    Dokument- und Terminverwaltung sowie ggf. Reiseplanung und -abrechnung.

13. **Greve, Andreas**
    Unterstützung der technischen Infrastruktur zum Datenaustausch sowie bei der Installation von Anwendungsprogrammen.

14. **Stickel , Britta**
    Dokument- und Terminverwaltung sowie Reiseplanung und -abrechnung.

15. **Warda, Kai**
    Mitarbeit im Arbeitspaket 1 sowie Unterstützung der technischen Infrastruktur zum Datenaustausch.

16. **Domke, Matthias**
    Unterstützung der technischen Infrastruktur zum Datenaustausch sowie bei der Installation von Anwendungsprogrammen.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

17. **Overkemping, Ann-Kristin**
    Arbeitspaket Signal-Untergrund-Separation (2) sowie Arbeitspaket Strukturbrucherkennung (3).

18. **Bockermann, Christian**
    Arbeitspakete bzgl. der Datenströme, Integration von Datenzusammenfassungs- und Lernverfahren in die FACT tools.

19. N.N.,
Mitarbeit an den Arbeitspaketen Signal-Untergrund-Separation (2) und Spektralrekonstruktion (4).

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	10.000	10.000	10.000	10.000
Summe der vorhandenen Mittel	10.000	10.000	10.000	10.000
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt C4

3.1.1 Titel:

Regressionsverfahren für sehr große, hochdimensionale Daten

3.1.2 Fachklassifizierung:

Algorithmische Komplexitätstheorie, Statistische Methoden

3.1.3 Projektleitung

Prof. Dr. Ickstadt, Katja, 18.01.1965

Lehrstuhl für Mathematische Statistik und Biometrische Anwendungen, Fakultät Statistik, Technische Universität Dortmund
Vogelpothsweg 87
44227 Dortmund

Tel.: 0231-755-3111
Fax : 0231-755-5303
E-Mail: `katja.ickstadt@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Sohler, Christian, 19.02.1973

LS 2, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 14
44227 Dortmund

Tel.: 0231-755-6940
Fax : 0231-755-2047
E-Mail: `christian.sohler@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Regression ist ein wichtiges Werkzeug der Statistik, um Zusammenhänge und Abhängigkeiten zwischen Variablen zu untersuchen. Fortschritte in der Leistungsfähigkeit von modernen Rechnern ermöglichen die Analyse unter immer komplexeren statistischen Modellannahmen. Bei einer sehr großen Anzahl von Beobachtungen oder Variablen stoßen klassische Algorithmen dennoch an Grenzen der Durchführbarkeit. Dies gilt sowohl für den Servereinsatz auf Big Data als auch für den Einsatz in weniger leistungsfähigen eingebetteten Systemen mit stark eingeschränkten Ressourcen. Die Problematik tritt hier bereits bei moderaten Datenmengen auf. Das Ziel dieses Teilprojekts ist daher die Entwicklung von *hocheffizienten* Regressionsverfahren. Um den Einsatz in eingebetteten Systemen zu ermöglichen, wird dazu versucht, algorithmische Techniken zu verwenden, welche effizient im Datenstrommodell und in verteilten Szenarien umsetzbar sind. Beides ist dabei besonders wichtig, da sowohl Datenströme als auch eine verteilte Datenlage bei eingebetteten Systemen häufig vorkommen. Die Algorithmik von Problemen aus der Datenanalyse unter Ressourcenbeschränkungen und speziell in diesen Szenarien ist ein junges Forschungsgebiet und bietet viele offene Fragestellungen grundlegender theoretischer wie auch praktischer Natur. Der bisherige Schwerpunkt lag auf der Entwicklung von Verfahren zur Reduktion der Anzahl von Beobachtungen durch geeignete Zusammenfassung und somit bei der Bewältigung von Speicher- und Kommunikationsbeschränkungen. Für die Fortführung des Teilprojekts verfolgen wir das Ziel, die Verfahren weiterzuentwickeln sowie verstärkt Reduktionsverfahren für die Anzahl der Variablen, d.h. Merkmalsselektion, zu betrachten.

3.3 Bisherige Entwicklung des Teilprojekts

Der bisherige Schwerpunkt lag auf der Entwicklung von Verfahren zur Zusammenfassung von Beobachtungen zu einer niedrigeren Anzahl bei Erhaltung der statistischen Aussagekraft. Somit wurde direkt die Bewältigung von Speicher- und Kommunikationsbeschränkungen adressiert. Dabei wurde darauf Wert gelegt, dass die Algorithmen sowohl im Datenstrommodell als auch im verteilten Szenario effizient operieren können. Dies ist durch die hohe und grundlegende Bedeutung dieser Modelle für den Entwurf von Algorithmen für eingebettete Systeme begründet.

Technisch bieten sich zur Datenreduktion vor allem zufällige lineare Projektionen in niedrigdimensionale Unterräume sowie Samplingverfahren und das Konzept von Kernmengen an. Zufällige Projektionen können aufgrund ihrer Linearität dynamisch verwaltet werden, das heißt, es können sowohl Daten eingefügt als auch wieder gelöscht werden. Zudem können Vektoren zeilenweise, spaltenweise, in ganzen Blöcken oder auch eintragsweise aktualisiert werden. Dadurch ist ein Einsatz in Datenströmen besonders flexibel möglich. Die Linearität erweist sich auch in verteilten Systemen als vorteilhaft. So können einzelne Knoten eines Netzwerks unabhängig voneinander Daten zusammenfassen. Die kleinen Zusammenfassungen können dann effizient kommuniziert und zu einer Zusammenfassung der gesamten Datenmenge summiert werden.

Kernmengen sind kleine, ggf. gewichtete Zusammenfassungen von Datenpunkten, deren Funktionswert bezüglich der Zielfunktion des betrachteten Problems nah an dem der eigentlichen Daten liegt. Häufig sind Kernmengenkonstruktionen statisch, d.h. sie können nicht für jeden einzelnen Eingabepunkt effizient aktualisiert werden. Um diesem Problem entgegenzuwirken und dennoch eine effiziente Behandlung von Datenströmen zu erlauben, kann das sogenannte *Merge-and-Reduce*-Prinzip verwendet werden. Hierbei handelt es sich um ein Schema, welches erlaubt, statische Datenstrukturen dynamisch zu machen und dabei nur einen logarithmischen Faktor für den Speicherbedarf zu opfern. Dazu müssen die beteiligten Kernmengen eine effiziente und verlustfreie Vereinigung zulassen. Ist dies gegeben, so ist auch der Einsatz für verteilte Daten möglich. Wieder können Kernmengen verteilt berechnet werden und kommunikationseffizient zu einer gesamten Zusammenfassung aller Daten vereinigt werden.

Auf den resultierenden reduzierten Datenmengen können klassische Analyseverfahren dann wieder effizient arbeiten. Die statistische Aussagekraft der gewonnenen Modelle bleibt dabei mit geringen Verlusten approximativ erhalten.

Als grundlegende Verallgemeinerungen linearer Regressionsmodelle wurden hierbei vor allem die Bayes-Regression sowie die logistische Regression betrachtet. In enger interner Zusammenarbeit wurde die angesprochene Güte der Modelle theoretisch analysiert und empirisch evaluiert. Dafür wurden unter anderem reale Daten aus den Teilprojekten TB1 und C3 verwendet. Weitere Aspekte, die zur effizienten Lösung von Regressionsproblemen untersucht wurden, sind Strukturannahmen wie Unimodalität und diskrete Graphstrukturen.

Im Folgenden findet sich eine Übersicht der in den Arbeitspaketen aus der ersten Phase erzielten Ergebnisse.

3.3.1 Bericht und Stand der Forschung

Arbeitspaket 1. Datenstromalgorithmen für lineare Regression: Unsere wichtigste Entwicklung in diesem Arbeitspaket ist die erste lineare Unterraumeinbettung für ℓ_1-Räume [C4/604], welche bis auf deren Größe unabhängig von der betrachteten Eingabematrix konstruiert werden kann. Mit einer solchen Einbettung ist es möglich, die ℓ_1-Norm von beliebigen Vektoren aus einem Unterraum zu approximieren. Während die bekannte Johnson-Lindenstrauss-Transformation [583, 564] dazu verwendet werden kann, eine $(1 \pm \varepsilon)$-Approximation der euklidischen ℓ_2-Normen von beliebigen n-elementigen Mengen von Vektoren mit logarithmischem Speicherplatz zu erhalten, ist dies für ℓ_1-Normen nicht ohne weitere Annahmen möglich. Stammen die Vektoren allerdings aus einem linearen Unterraum vom Rang d, so kann diese spezielle Struktur dazu ausgenutzt werden, die umgebende Dimension von n auf $O(d \log d)$ zu reduzieren und die dabei entstehende Verzerrung gering zu halten. Genauer ausgedrückt wurde die Existenz einer Verteilung über linearen Abbildungen $R : \ell_1^n \to \ell_1^{O(d \log d)}$ nachgewiesen, welche mit beliebig hoher konstanter Wahrscheinlichkeit für einen beliebigen aber festen d-dimensionalen linearen Unterraum L des ℓ_1^n garantiert, dass für alle $x \in L$ für ein $D = O(d \log d)$ gilt $||x||_1 \leq ||Rx||_1 \leq D||x||_1$. Dies eröffnete die Möglichkeit, für ℓ_1-basierte Probleme effizientere Approximationsalgorithmen zu entwickeln. So gelang es uns beispielsweise, sogenannte *wohl-konditionierte Basen* schneller zu berechnen als dies zuvor möglich war. Dabei handelt es sich um Basen, welche die Normen von Vektoren nicht zu sehr verzerren. Im euklidischen Fall sind diese bekanntermaßen unter Rotation invariant, was aber für ℓ_1- und allgemeiner für ℓ_p-Räume nicht gilt. Bildet man nun Vektoren in eine solche Basis ab, dann entsprechen ihre Normen weitgehend ihrer *Wichtigkeit* für Probleme, die auf der Struktur von Unterräumen beruhen. Darunter zu nennen sind z.B. die in unserer Arbeit behandelten Probleme *least absolute deviation regression* sowie ℓ_1-*best fit hyperplane*, die ℓ_1-Gegenstücke zur Kleinste-Quadrate-Regression und der Hauptkomponentenanalyse. Für diese Probleme haben wir jeweils $(1 + \varepsilon)$-Approximationsalgorithmen entwickelt, deren Laufzeit und Speicherkomplexität im Vergleich zu vorherigen Algorithmen verbessert wurden. Die lineare Struktur der verwendeten Abbildungen ermöglicht es, die beschriebenen Einbettungen im verteilten wie auch im sequentiellen Datenstrommodell speicher- und kommunikationseffizient einzusetzen, wie einige verwandte Arbeiten zu Unterraumeinbettungen, Niedrigrangapproximation und Regression für große Datenmengen belegen [74, 64, 596].

Auf unserer Arbeit aufbauend wurden auch außerhalb des SFB 876 einige offene Problemstellungen gelöst, wie z.B. die effiziente Approximation der sogenannten *statistical leverage scores* in Zeit $o(nd^2)$ [586]. Diese ermöglichen samplingbasierte ℓ_2-Regression [567]. Weiterhin wurden Verallgemeinerungen auch für ℓ_p-Räume $p \in [1, \infty)$ entwickelt [562, 561]. Bezüglich der Einbettungsgeschwindigkeit wurden ebenfalls Optimierungen entwickelt. Seit den im SFB 876 Streaming Workshop vorgestellten Arbeiten gibt es Einbettungen, deren Laufzeit im Wesentlichen proportional zur Anzahl der Nicht-Null-Einträge in der Eingabematrix ist [560, 596]. Dies ist offensichtlich optimal, da diese Zeit allein für das Lesen der Eingabe nötig ist.

Im Rahmen von Abschlussarbeiten wurden ebenfalls Datenstromalgorithmen für lineare Regression entwickelt und evaluiert. Die Abschlussarbeit von Andreas Rüppert beschäftigt sich mit der LASSO-Regression für große Datenmengen speziell im Datenstrommodell. Hier werden zufällige lineare Projektionen verwendet, um den bekannten LARS-Algorithmus zu beschleunigen und dabei beweisbar eine $(1+\varepsilon)$-Approximation zu erhalten. In der empirischen Auswertung wird speziell ein Augenmerk auf die Abhängigkeit der Zieldimension von dem Approximationsparameter ε gelegt. Eine quadratische Abhängigkeit ist von theoretischer Seite immer ausreichend, während eine lineare Abhängigkeit wie beim Algorithmus von Woodruff und Clarkson für unpenalisierte Regression [64] zumindest in bestimmten Fällen nicht auszuschließen ist. Aufgrund der Lücke zwischen der quadratischen oberen und der linearen unteren Schranke aus [64] ist die Komplexität bis auf einen Faktor von ε^{-1} asymptotisch exakt bestimmt.

Die Abschlussarbeit von Jens Quedenfeld ist grundlegender Natur für Regressionsprobleme im Datenstrommodell. Sie beschäftigt sich mit der experimentellen Analyse von klassischen wie auch neueren ℓ_2-Einbettungen speziell vor dem Hintergrund dünn besetzter Eingabedaten. So zeigte sich, dass klassische Verfahren in der Praxis sehr langsam sind, während die neueren Ansätze bei sehr dünn besetzten Daten problematisch sind und eine höhere Einbettungsdimension erfordern, was das schnellste der untersuchten Verfahren [560, 588] für die Einbettung beliebiger Mengen von Vektoren unbrauchbar macht. Dieses ist nur unter Einschränkungen an die *sparsity* oder für die Einbettung ganzer linearer Unterräume bzw. Orthonormalbasen geeignet, wie auch in [560] bemerkt. Das um einen logarithmischen Term langsamere Verfahren von [585] eignet sich hingegen sehr gut für die Einbettung auch dünner Eingabedaten und hat eine moderate Einbettungsdimension.

Arbeitspaket 2: Verallgemeinerte Regressionsmodelle und diskrete Strukturen: In Zusammenarbeit mit Teilprojekt A2 haben wir einen Datenstromalgorithmus entwickelt, welcher unter relaxierten Bedingungen

an die Eingabedaten eine $(1+\varepsilon)$-Approximation für lineare Klassifikationsprobleme unter diversen Kostenfunktionen $Z(\cdot,\cdot)$ berechnet [A2/C4/608]. Wichtige Spezialfälle umfassen die sogenannte *Perceptron-Loss-Function* sowie die *Logistic-Error-Function* welche bei der logistischen Regression verwendet wird. Ziel ist es, für eine gegebene Menge von Punkten A eine Datenstruktur im Datenstrom so aufrecht zu erhalten, dass für jeden Lösungsvektor v der Zielfunktionswert $Z(v,A)$ bis auf einen $(1\pm\epsilon)$-Faktor genau angefragt werden kann.

Es wurde gezeigt, dass im Datenstrommodell die Konstruktion einer solchen Datenstruktur nur möglich ist, wenn im Wesentlichen alle Punkte der Instanz abgespeichert werden. Damit kann aber insbesondere ein Algorithmus, der dieses Problem mit polylogarithmischem Speicherplatz löst, nicht existieren.

Um dennoch Aussagen über die Kompressionsfähigkeit eines Datensatzes für die oben genannten Klassifikationsprobleme treffen zu können, wurde eine Relaxierung des Problems betrachtet. Wir definierten ein (asymmetrisches) Abstandsmaß für n-elementige Punktmengen A und B, bei der jeder Punkt von A auf einen Punkt von B abgebildet wird. Für jeden Richtungsvektor w und jeden Punkt $x\in A$ existiert dann ein Punkt $y\in B$, so dass $\left(w^T(x-y)\right)^2 \le \gamma\cdot\max_{x\in A}\sum_{a\in A}(w^T(x-a))^2$. B heißt dann γ-nah zu A. Anschaulich werden Punkte proportional zu ihrem Beitrag zur Varianz des Datensatzes entlang der Richtung w verschoben. Wir versuchen nun nicht notwendigerweise den Zielfunktionswert auf dem ursprünglichen Datensatz A anzufragen, sondern auf einem Datensatz, der mindestens γ-nah zu A ist. Wir konnten einen Algorithmus angeben, der im Datenstrom Schätzer aufrecht erhält, die für einen γ-nahen Datensatz B den Zielfunktionswert für jeden Vektor v bis auf einen $(1\pm\epsilon)$-Faktor genau schätzen können. Der benötigte Speicherplatz des Algorithmus ist dabei durch γ und ϵ parametrisiert und kann geeignet variiert werden. Der Algorithmus zieht mehrere Stichproben, mit deren Hilfe eine Approximation der konvexen Hülle, siehe [549], berechnet wird. Die Anzahl der Punkte in einem Halbraum lässt sich dann mittels der Anzahl der Schnitte einer konvexen Hülle mit dem Halbraum schätzen.

Zur Darstellung und Analyse diskreter Strukturen haben wir einen anderen Ansatz gewählt. In diesem Kontext eignen sich häufig Netzwerke, die bedingte Unabhängigkeiten modellieren. Eine oft verwendete Modellklasse stellen dabei Gaußsche Bayessche Netzwerke dar, die jedoch zum einen multivariate Normalverteilungsannahmen treffen und zum anderen nur lineare Abhängigkeiten modellieren. Deshalb wurde in [574] ein Mischungsmodell Gaußscher Bayesscher Netzwerke entwickelt, das in unserer Arbeit [C4/599] in Kooperation mit dem MPI für Molekulare Physiologie zu einem nichtparametrischen Bayesschen Netzwerkmodell verallgemeinert wurde, dem ein Mischungsmodell mit prinzipiell unendlich vielen Komponenten zugrunde liegt. Diese Modellklasse kommt im Allgemeinen mit wenigen Mischungskomponenten (*sparsity*) aus und hat durch den Einsatz latenter Variablen zur Analyse den Vorteil, dass Teilnetzwerke modelliert werden können, die Subgruppen von Beobachtungen beschreiben.

Arbeitspaket 3: Effiziente Bayessche Regressionsverfahren: Bei der Bayesschen Regression wird im Gegensatz zur frequentistischen Betrachtungsweise (vgl. AP1, AP2) nicht von einer festen optimalen Regressionslösung ausgegangen, sondern eine Verteilung über Regressionsparameter eingeführt. Dazu wird eine a priori-Verteilung $p_{\text{pre}}(\beta)$ über dem Parameterraum definiert, welche problemspezifisches Vorwissen modellieren kann. In vielen Anwendungen wird diese jedoch nichtinformativ gewählt, wenn kein tatsächliches Vorwissen vorhanden ist, um den Lösungsraum nicht unwissend einzuschränken. Weiterhin wird die *Likelihood* $\mathcal{L}(X,Y|\beta)$ der Daten unter der Unabhängigkeitsannahme als Produkt über die Verteilungen, die sich aus Einzelbeobachtungen ergeben, definiert. Die gewünschte a posteriori-Verteilung, welche das Wissen nach Beobachtung der Daten und unter Berücksichtigung der Vorinformation beschreibt, ergibt sich damit als:

$$p_{\text{post}}(\beta|X,Y) \propto \mathcal{L}(X,Y|\beta)\cdot p_{\text{pre}}(\beta).$$

Die Bayes-Regression als Verallgemeinerung der klassischen frequentistischen Ansätze zur linearen penalisierten Regression kann ebenfalls mittels zufälliger linearer Projektionen behandelt werden. Die Idee ist, eine Unterraumeinbettung Π zu verwenden, um die Datenmatrix $[X,Y]\in\mathbb{R}^{n\times(d+1)}$ auf eine *Skizze* $[\Pi X,\Pi Y]\in\mathbb{R}^{k\times(d+1)}$ für ein $k\in O(\text{poly}(d)/\varepsilon^2)$ abzubilden. Dabei entspricht das Ergebnis der Bayesschen Regressionsanalyse auf der Skizze beweisbar bis auf eine geringe, von ε abhängige Abweichung demjenigen, welches man bei Analyse des Originaldatensatzes erhält:

$$\begin{array}{ccc} [X,Y] & \xrightarrow{\Pi} & [\Pi X,\Pi Y] \\ \downarrow & & \downarrow \\ \mathcal{L}(X,Y|\beta)\cdot p_{\text{pre}}(\beta) & \approx_\varepsilon & \mathcal{L}(\Pi X,\Pi Y|\beta)\cdot p_{\text{pre}}(\beta). \end{array}$$

Dazu entstand eine teilprojektinterne gemeinsame Arbeit [C4/605]. Untersucht werden hierzu eingabeunabhängige Unterraumeinbettungen, welche auf den klassischen Einbettungsverfahren von [552, 548, 64] beruhen und eine lineare Einbettungsdimension haben. Die sogenannte *Subsampled-Randomized-Hadamard-Transform* (SRHT) [551] ist eine schnellere, aber um einen logarithmischen Faktor speicheraufwändigere [594] Variante der schnellen Johnson-Lindenstrauss-Transformation [550] und kann ebenfalls dazu verwendet werden, eine Unterraumeinbettung zu erreichen. Die schnellste untersuchte Unterraumeinbettung stellt die Methode von Clarkson und Woodruff [560] dar. Diese kann Matrizen in Laufzeit proportional zur Anzahl ihrer Nicht-Null-Einträge einbetten. Man beachte, dass diese Zeit allein zum Lesen der Eingabe notwendig ist. Die Einbettungsdimension ist quadratisch in der Dimension des Unterraums und somit optimal für solche Einbettungen, siehe [588, 589]. Es besteht also in unserem Kontext ein Trade-off zwischen Einbettungsgeschwindigkeit und Kompression der Daten. Die Abhängigkeit von dem Approximationsparameter ε ist bei allen erwähnten Unterraumeinbettungen quadratisch.

Wir haben von theoretischer Seite gezeigt, dass eine ℓ_2-basierte Likelihood bzgl. der *Wassersteinmetrik* [595] unter der Einbettung nur leichte Verzerrungen erfährt. Diese Metrik bezieht im Gegensatz zu den vorherigen Betrachtungen, vgl. z.B. [591, 64], sowohl den Lageparameter, also den Erwartungswert einer Normalverteilung, sowie auch deren Kovarianzen in das Distanzmaß ein. Unter *milden* Annahmen kann eine multiplikative $(1 + \varepsilon)$-Approximationsgüte bzgl. der Wassersteinmetrik gezeigt werden. Dazu wurde das Konzept des *Wassersteingewichts* eingeführt, welches die Distanz zur Dirac-Delta-Funktion bezeichnet. Diese konzentriert sämtliche Wahrscheinlichkeitsmasse auf dem geometrischen Nullpunkt und besitzt sonst einen Wert von Null. Die Verwendbarkeit von nichtinformativen a priori-Verteilungen, wobei auch unzulässige Verteilungen wie die Gleichverteilung auf den reellen Zahlen eingeschlossen sind, sowie beliebigen Normalverteilungen folgt direkt aus den bewiesenen Theoremen. Eine Verallgemeinerung hin zu einer deutlich größeren Klasse von a priori-Verteilungen, die gewisse Regularitätsbedingungen erfüllen, wird weiterhin erarbeitet. Besonders aussichtsreich erscheint in diesem Kontext zunächst die Betrachtung von Verteilungen, deren Definitionen über ℓ_p-Normen gegeben sind. Diese sind algebraisch handhabbar. Neben der bereits erwähnten Normalverteilung mit $p = 2$ entspricht beispielsweise die Verwendung einer Laplaceverteilung mit $p = 1$ dem Bayes-Pendant zur LASSO-Regression.

Zu den theoretischen Ergebnissen haben wir den Einsatz von zufälligen linearen Projektionen für die Bayes-Regression auch empirisch anhand von simulierten Datensätzen und Datensätzen, die von Teilprojekt C3 zur Verfügung gestellt wurden, analysiert. Im Gegenzug wurden die gewonnenen Regressionsmodelle dazu verwendet, die Energie von astrophysikalischen Beobachtungen aus Teilprojekt C3 vorherzusagen. Die empirischen Ergebnisse bestätigen unsere theoretischen Resultate bezüglich der algorithmischen Performanz unserer Methode. Wir verwenden zur vergleichenden Analyse *Markov Chain Monte Carlo* (MCMC)-Methoden. Diese sind in der Bayes-Statistik Standard; ihre Konvergenz ist überprüfbar und sie skalieren im Vergleich zu anderen Methoden gut mit hoher Dimensionalität. Dennoch ist ihr Einsatz bei sehr großen Datenmengen nicht mehr praktikabel. Diesem Problem begegnen wir mit unserem Ansatz. Durch die reduzierte Datengröße ist es auch für sehr große Datensätze möglich, MCMC-Methoden anzuwenden.

Die Bayesschen Regressionsmodelle, jeweils basierend auf dem Originaldatensatz und dem in einen Unterraum projizierten Datensatz, unterscheiden sich nur geringfügig voneinander. Die Mittelwerte der a posteriori-Verteilungen der Parameter sind praktisch identisch. Bei weiteren Kennzahlen wie Varianz, Median und Quartilen sind kleine Abweichungen festzustellen. Die empirischen Fehler zwischen den jeweiligen Verteilungen liegen jedoch meist deutlich unter den theoretisch errechneten Größen. Im Allgemeinen weisen die a posteriori-Verteilungen der Parameter basierend auf dem eingebetteten Datensatz eine leicht höhere Streuung auf. Die Einsparung in der Gesamtlaufzeit der Analyse inklusive Einlesen und ggf. Skizzieren der Daten sowie einer Analyse per MCMC-Methoden beträgt schon für einen vergleichsweise kleinen Originaldatensatz mit 50 000 Beobachtungen und 50 Variablen etwa 80%. Bei größeren Originaldatensätzen ist der relative Anteil der Laufzeiteinsparung entsprechend höher.

Ein weiterer Ansatz, den wir untersucht haben, ist eine Übertragung des *Merge-and-Reduce*-Prinzips [554, 549] auf die Bayessche Regression. *Merge-and-Reduce* liest einen Datensatz blockweise ein und führt die Regression auf jedem Block einzeln durch. Die Modelle werden entlang einer binären Baumstruktur vereinigt, sobald zwei Modelle auf derselben Hierarchieebene vorliegen. Dadurch ist der Speicherbedarf für die Modelle logarithmisch in der Anzahl der Beobachtungen. Alle Einzelmodelle einer Ebene sollen mit demselben Gewicht in das finale Modell einfließen. Dies kann durch die Baumstruktur und Gewichtung proportional zur zugrundeliegenden Anzahl an Beobachtungen erreicht werden. Bei der Vereinigung darf die Komplexität der Modelle nicht steigen. Wir betrachten und speichern für jede einzelne Bayessche Regression Kennzahlen über die interessierenden a posteriori-Verteilungen. Dies können Lageparameter wie das arithmetische Mittel oder der Median, Streuungsparameter wie die Varianz, Quantile oder weitere Kennzahlen sein. Bei jeder Vereinigung bilden wir das gewichtete arithmetische Mittel der beiden einzelnen Modelle. Am Ende können

wir damit die a posteriori-Verteilung(en) mithilfe der gewählten Kennzahlen charakterisieren. Die Werte der Kennzahlen weichen nur in geringem Maße von denjenigen ab, die wir über eine Bayes-Regression auf dem kompletten Datensatz erhalten. Die Laufzeit der Analyse ist bei der Verwendung von *Merge-and-Reduce* nach unseren empirischen Untersuchungen in der Mehrheit der Fälle erheblich schneller. Der Hauptvorteil im Vergleich zu einer Bayes-Regression auf dem kompletten Datensatz liegt jedoch darin, dass Datenmengen mit beliebiger Anzahl n an Beobachtungen analysiert werden können. Dabei steigt die benötigte Anzahl von Modellen, die zugleich im Speicher gehalten werden muss, nur logarithmisch in n. Die Größe der Blöcke kann so gewählt werden, dass der freie Speicher der Ausführungsplattform ausgenutzt wird.

Arbeitspaket 4: Strukturannahmen zur Dimensionsreduktion: Auch zusätzliche Strukturannahmen wie Monotonie oder Unimodalität können zur Dimensionsreduktion verwendet werden. Dies ist bereits bei einer niedrigen Anzahl an Variablen anspruchsvoll. Zur univariaten Regressionsanalyse unter der Strukturannahme der Unimodalität kann beispielsweise der Ansatz von [570] genutzt werden. Da eine unimodale Funktion mit bekanntem Modus durch zwei monotone Funktionen beschrieben werden kann, ist hierbei die Idee, über alle möglichen Modi zu iterieren und die beste Anpassung auszuwählen. Die Zusammensetzung zweier monotoner Funktionen ist jedoch insofern problematisch, als dass die daraus resultierende Funktion am Modus nicht differenzierbar ist, vielleicht sogar unstetig sein kann.

In unserer Arbeit [C4/600] zeigen wir, dass dieses Problem mithilfe von Splineregression gelöst werden kann. Bei der Benutzung der B-Spline-Basis kann durch Optimierung unter Nebenbedingung eine Splinefunktion mit festgelegtem Modus angepasst werden. Dann wird wiederum durch Iteration über alle möglichen Modi die beste unimodale Anpassung an die Daten gefunden. Alternativ zu diesem frequentistischen Vorgehen können die Regressionsfunktion und auch der Modus mit einem Bayesschen Verfahren bei geeigneter Wahl der a priori-Verteilungen durch a posteriori-Mittelwerte geschätzt werden.

Dieses Verfahren wurde in Zusammenarbeit mit Teilprojekt TB1 zur Modellierung von Daten aus der Ionen-Mobilitäts-Spektrometrie eingesetzt. Weiterhin wurde das Verfahren erfolgreich zur Analyse von Daten des Virensensors aus Teilprojekt B2 und von Ladekurven des FACT-Teleskops aus Teilprojekt C3 angewandt. Diese Ergebnisse wurden in Kooperation mit den entsprechenden Teilprojekten im Rahmen der Bachelorarbeit von Steffen Müller erzielt. Ein Vergleich dieser Ergebnisse mit denen einer nichtlinearen Regression bei unterschiedlichen Funktionsannahmen wurde in der Bachelorarbeit von Lena Okroy vorgenommen.

Es ist geplant, die Eigenschaften mehrdimensionaler Splinefunktionen ebenfalls entsprechend zu nutzen, um dieses Vorgehen auch auf multiple Regressionsmodelle zu übertragen.

In Zusammenarbeit mit Teilprojekt C3 ist außerdem eine gemeinsame Arbeit entstanden [C4/C3/607]. Motiviert durch die Bemühungen in der Physik, Monte-Carlo-Simulationen für physikalische Prozesse zu entwerfen, wurde eine neuartige statistische Methode entwickelt, welche dazu verwendet werden kann, eine vorhandene Simulation anhand von beobachteten und simulierten Werten zu evaluieren. Bisherige klassische Testverfahren wie der beliebte und vielseitig eingesetzte Kolmogorov-Smirnov-Test geben bei Vorgabe zweier Stichproben und eines Konfidenzniveaus Aufschluss darüber, ob die zugrundeliegenden Verteilungen gleich sind oder nicht. Im Falle der Gleichheit ist dies eine zufriedenstellende Information, während die reine Ablehnung der Homogenitätshypothese keinen ausreichenden Aufschluss darüber gibt, wie stark die Verteilungen voneinander abweichen. Im oben skizzierten Szenario sind wir darüber hinaus an weiteren Informationen interessiert, welche es uns ermöglichen, die vorhandenen Simulationen zu verbessern, wenn die Homogenitätshypothese abgelehnt wird. In der Arbeit [C4/C3/607] haben wir ein solches Verfahren entwickelt, dessen Ziel es ist, sowohl ein Maß anzugeben, wie stark reale Daten und Simulation übereinstimmen, als auch eine zusätzliche Komponente hinzuzufügen, welche einer nötigen Korrekturverteilung entspricht. Das mathematische Modell ist eine Mischung aus nichtparametrischen Verteilungsfunktionen $F = s \cdot G + (1 - s) \cdot H$. Hierbei bezeichnet $s \in [0, 1]$ das Mischungsverhältnis der Verteilungen. Die Verteilung der realen Daten wird mit F bezeichnet, die der Simulation mit G und die Korrekturverteilung ist die Komponente H. Für solche nichtparametrischen Mischungen ist es im Allgemeinen nicht möglich, die Komponenten zu identifizieren und zu trennen [578]. In unserer Modellierung ist allerdings die Komponente der Simulation bereits bekannt. Um nun eine Lösung zu finden, welche einen möglichst hohen Anteil der bisherigen Simulation mit einer möglichst geringen Korrekturkomponente kombiniert, kann nach dem Parameter s optimiert werden. Dies ist, wie in unserer Arbeit [C4/C3/607] gezeigt, mit einer festen additiven Genauigkeit $\varepsilon > 0$ effizient in Zeit $O(n \log(\frac{1}{\varepsilon}))$ möglich. Die Hinzunahme von heuristischen Verbesserungsschritten hat in der empirischen Auswertung eine weitere Verbesserung der Laufzeit um einen konstanten Faktor ergeben. Der entwickelte Algorithmus wurde weiterhin anhand von realen Daten evaluiert, welche aus der Kooperation mit der Physik innerhalb des Teilprojekts C3, sowie aus einer Kooperation mit den Mitarbeitern der Bioinformatik aus Teilprojekt TB1 stammen [TB1/301]. Die von unserem Algorithmus vorgeschlagenen Korrekturen tragen zur Verbesserung der Modellierung von Spektrometriedaten in Teilprojekt TB1 bei. Der Algorithmus wird

ebenfalls zur Validierung der verbesserten Modelle eingesetzt. Weiterhin werden die Ergebnisse verwendet, um interessante Variablen und domänenspezifische Details in den hochdimensionalen Daten des Teilprojekts C3 zu identifizieren.

Arbeitspaket 5: Merkmalsselektion: In Kooperation mit Teilprojekt TB1 haben wir ein Verfahren zur kompakten Darstellung einer Menge von Strings N entwickelt [TB1/C4/598]. Jeder String über einem Alphabet Σ entspricht einer Aneinanderkettung von Symbolen aus Σ. Gesucht war für eine gegebene Kostenfunktion eine Menge an Substrings $S \subseteq \Sigma^*$ mit minimalen Kosten, die jeden Eingabestring $s \in N$ durch Konkatenation erzeugt, also $s \in S^*$. Der entworfene Algorithmus basiert auf einer Formulierung des Problems mittels eines ganzzahligen linearen Programms und einer einfachen, aber in der Praxis effizient durchführbaren Methode, um untere Schranken für ein *Branch-and-Bound*-Verfahren auszurechnen. Es ist neben der erfolgreichen Anwendung auf zufällig erzeugten Daten gelungen, mit Hilfe dieses Algorithmus die Zusammensetzung des Textes „Alice im Wunderland“ von Lewis Carroll zu analysieren.

Auch für die lineare Bayessche Regression haben wir uns in der Abschlussarbeit von Verena Jabs [C4/606] vergleichend mit Mermalsselektionsmethoden beschäftigt. Diese werden vor der eigentlichen Regressionsanalyse durchgeführt, um die Dimension der Daten zu reduzieren. Dabei haben wir Hauptkomponentenanalyse, LASSO-Regression und Faktorenanalyse als vorgeschaltete Verfahren betrachtet. In der Arbeit zeigte sich Bayessche Regression basierend auf den wichtigsten Hauptkomponenten gegenüber den anderen Verfahren als überlegen, sowohl bezogen auf die Modellanpassung als auch auf die nötige Rechenzeit. Über die Anzahl der Hauptkomponenten, die für die weitere Analyse verwendet werden sollen, ist es möglich, eine bessere Anpassung auf Kosten einer höheren Rechenzeit zu erhalten oder auch umgekehrt.

Zudem sind im Bereich der Selektion einflussreicher Merkmale für die Erklärung bzw. Vorhersage der abhängigen Variablen innerhalb der ersten Förderphase insbesondere für genetische SNP-Studien einige Ergebnisse entstanden. So haben wir in [C4/597] die *cluster-localized regression* entwickelt, bei der basierend auf einem lokalen logistischen Regressionsmodell unterschiedliche SNPs für unterschiedliche Patientengruppen gefunden werden und die Variablenselektion direkt durch gewichtetes komponentenweises Boosting erfolgt. Die Methode wird zum Auffinden von SNPs für Harnblasenkrebs untersucht. In Schwender et al. [C4/603] haben wir ebenfalls die einflussreichsten SNPs und SNP-Kombinationen für Harnblasenkrebs selektiert und zwar stratifiziert nach Rauchern und Nichtrauchern. Die Stabilität der gefundenen Merkmale wurde mit Boosting überprüft.

Als weitere Methode zur Merkmalsselektion für Regressionsmodelle basierend auf diskreten Strukturen haben Schwender und Ickstadt für logicFS, einer Weiterentwicklung der logischen Regression, *Variable Importance Measures* (VIMs) zur Bestimmung des Einflusses von Interaktionen binärer Variablen auf die Vorhersage einer binären Zielvariablen formuliert. Diese haben wir in Schwender et al. [C4/602] eingesetzt, um einflussreiche SNP-Mengen z. B. in Pathways oder Genen zu finden, die mit der Erkrankung assoziiert sind. Um die Merkmale auf Signifikanz zu überprüfen, schlagen wir eine Test-Prozedur vor, bei der standardisierte VIMs als Test-Statistik verwendet werden.

Verwandt mit dem Problem der Merkmalsselektion in hochdimensionalen Modellen ist die Validierung gefundener Merkmale auf weiteren Datensätzen. Hier schlagen wir in der Zusammenarbeit [A3/C4/601] eine sequentielle Validierungsstrategie für hochdimensionale genetische Daten wie SNP- oder Genexpressionsdaten vor, bei der Merkmale, die in einer Studie gefunden werden, als Kandidaten in die folgende Studie eingehen. Durch Simulationen wird gezeigt, dass die optimale Studienanordnung die Studien nach absteigender Qualität verwendet.

Literatur

[547] M. J. Abásolo, A. D. Giusti und J. Blat. „A Hierarchical Triangulation for Multiresolution Terrain Models“. In: *The Journal of Computer Science and Technology, JCS&T* 3 (2000), S. 1–13.

[548] D. Achlioptas. „Database-friendly random projections: Johnson-Lindenstrauss with binary coins“. In: *J. Comput. Syst. Sci.* 66.4 (2003), S. 671–687.

[549] P. K. Agarwal, S. Har-Peled und K. R. Varadarajan. „Approximating extent measures of points“. In: *J. ACM* 51.4 (2004), S. 606–635.

[550] N. Ailon und B. Chazelle. „The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors“. In: *SIAM J. Comput.* 39.1 (2009), S. 302–322.

[551] N. Ailon und E. Liberty. „Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes“. In: *Discrete & Computational Geometry* 42.4 (2009), S. 615–630.

[552] N. Alon, Y. Matias und M. Szegedy. „The Space Complexity of Approximating the Frequency Moments“. In: *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, STOC*. 1996, S. 20–29.

[553] S. Arora und R. Kannan. „Learning mixtures of separated nonspherical Gaussians“. In: *The Annals of Applied Probability* 15.1A (2005), S. 69–92.

[554] J. L. Bentley und J. B. Saxe. „Decomposable Searching Problems I: Static-to-Dynamic Transformation“. In: *J. Algorithms* 1.4 (1980), S. 301–358.

[555] C. Boutsidis, P. Drineas und M. Magdon-Ismail. „Near Optimal Column-Based Matrix Reconstruction“. In: *Proceedings of the IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011*. 2011, S. 305–314.

[556] C. Boutsidis, P. Drineas und M. Magdon-Ismail. „Near-Optimal Coresets for Least-Squares Regression“. In: *IEEE Transactions on Information Theory* 59.10 (2013), S. 6880–6892.

[557] C. Boutsidis, P. Drineas und M. Magdon-Ismail. „Sparse Features for PCA-Like Linear Regression“. In: *25th Annual Conference on Neural Information Processing Systems 2011 NIPS*. 2011, S. 2285–2293.

[558] S. Chan, I. Diakonikolas, R. A. Servedio und X. Sun. „Efficient Density Estimation via Piecewise Polynomial Approximation“. In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing STOC*. 2014.

[559] D. Z. Chen und H. Wang. „Approximating Points by a Piecewise Linear Function“. In: *Algorithmica* 66.3 (2013), S. 682–713.

[560] K. L. Clarkson und D. P. Woodruff. „Low rank approximation and regression in input sparsity time“. In: *Symposium on Theory of Computing Conference, STOC*. 2013, S. 81–90.

[64] K. L. Clarkson und D. P. Woodruff. „Numerical linear algebra in the streaming model“. In: *Proceedings of the fourty-first ACM Symposium on Theory of Computing (STOC)*. 2009, S. 205–214.

[561] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng und D. P. Woodruff. „The Fast Cauchy Transform and Faster Robust Linear Regression“. In: *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013*. 2013, S. 466–477.

[TB1/301] M. D'Addario, D. Kopczynski, J. I. Baumbach und S. Rahmann. „A modular computational framework for automated peak extraction from ion mobility spectra“. In: *BMC Bioinformatics* 15.1 (Jan. 2014), S. 25.

[562] A. Dasgupta, P. Drineas, B. Harb, R. Kumar und M. W. Mahoney. „Sampling Algorithms and Coresets for ℓ_p Regression“. In: *SIAM J. Comput.* 38.5 (2009), S. 2060–2078.

[563] S. Dasgupta. „Learning Mixtures of Gaussians“. In: *Proceedings of the 40th Annual Symposium on Foundations of Computer Science, FOCS '99*. 1999, S. 634–644.

[564] S. Dasgupta und A. Gupta. „An elementary proof of a theorem of Johnson and Lindenstrauss“. In: *Random Struct. Algorithms* 22.1 (2002), S. 60–65.

[565] L. De Floriani und E. Puppo. „Hierarchical Triangulation for Multiresolution Surface Description“. In: *ACM Trans. Graph.* 14.4 (1995), S. 363–411.

[566] D. Donjerkovic, Y. Ioannidis und R. Ramakrishnan. „Dynamic Histograms: Capturing Evolving Data Sets“. In: *Data Engineering, 2000. Proceedings. 16th International Conference on*. 2000, S. 86–86.

[567] P. Drineas, M. W. Mahoney und S. Muthukrishnan. „Sampling algorithms for ℓ_2 regression and applications“. In: *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006*. 2006, S. 1127–1136.

[568] D. Feldman, M. Faulkner und A. Krause. „Scalable Training of Mixture Models via Coresets“. In: *Proceedings of the 25th Annual Conference on Neural Information Processing Systems NIPS 2011*. 2011, S. 2142–2150.

[569] W. D. Fisher. „On Grouping for Maximum Homogeneity“. In: *Journal of the American Statistical Association* 53.284 (1958).

[570] M. Frisén. „Unimodal regression“. In: *The Statistician* 35.4 (1986), S. 479–85.

[571] M. Garland und P. S. Heckbert. *Fast Polygonal Approximation of Terrains and Height Fields*. 1995.

[572] M. Garland und P. S. Heckbert. „Surface simplification using quadric error metrics“. In: *SIGGRAPH*. 1997, S. 209–216.

[573] P. Goos und B. Jones. *Optimal Design of Experiments: A Case Study Approach*. 1. Aufl. Wiley, Aug. 2011.

[574] M. Grzegorczyk, D. Husmeier, K. D. Edwards, P. Ghazal und A. J. Millar. „Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler.“ In: *Bioinformatics* 24.18 (2008), S. 2071–2078.

[575] S. Guha. „Space Efficiency in Synopsis Construction Algorithms“. In: *Proceedings of the 31st International Conference on Very Large Data Bases, VLDB'05*. 2005, S. 409–420.

[576] S. Guha, N. Koudas und K. Shim. „Approximation and streaming algorithms for histogram construction problems“. In: *ACM Trans. Database Syst.* 31.1 (2006), S. 396–438.

[577] A. S. Hadi und R. F. Ling. „Some Cautionary Notes on the Use of Principal Components Regression“. In: *The American Statistician* 52.1 (1998), S. 15–19.

[578] P. Hall, A. Neeman, R. Pakyari und R. Elmore. „Nonparametric inference in multivariate mixtures“. In: *Biometrika* 92.3 (2005), S. 667–678.

[579] J. Hardwick und Q. F. Stout. „Optimal Reduced Isotonic Regression“. In: *Proceedings of Interface 2012: Future of Statistical Computing*. 2012.

[293] T. Hastie, R. Tibshirani und J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Statistics. Springer, 2009.

[580] D. M. Hawkins. „On the Investigation of Alternative Regressions by Principal Component Analysis“. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 22.3 (1973), S. 275–286.

[581] H. Hoppe. „Progressive Meshes“. In: *SIGGRAPH*. 1996, S. 99–108.

[582] H. V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K. C. Sevcik und T. Suel. „Optimal Histograms with Quality Guarantees“. In: *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases*. 1998, S. 275–286.

[583] W. Johnson und J. Lindenstrauss. „Extensions of Lipschitz Mappings into a Hilbert Space“. In: *Conference in Modern Analysis and Probability (New Haven, Conn., 1982)*. Bd. 26. Contemporary Mathematics. American Mathematical Society, 1984, S. 189–206.

[584] I. T. Jolliffe. „A Note on the Use of Principal Components in Regression“. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31.3 (1982), S. 300–303.

[585] D. M. Kane und J. Nelson. „Sparser Johnson-Lindenstrauss transforms“. In: *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*. 2012, S. 1195–1206.

[74] R. Kannan, S. Vempala und D. Woodruff. „Nimble Algorithms for Cloud Computing“. In: *Computing Research Repository (CoRR)* abs/1304.3162 (2013).

[586] M. W. Mahoney, P. Drineas, M. Magdon-Ismail und D. P. Woodruff. „Fast approximation of matrix coherence and statistical leverage“. In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. 2012.

[587] Y. Matias, J. S. Vitter und M. Wang. „Wavelet-Based Histograms for Selectivity Estimation“. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 1998, S. 448–459.

[588] J. Nelson und H. L. Nguyen. „OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings“. In: *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS*. 2013, S. 117–126.

[589] J. Nelson und H. L. Nguyen. „Sparsity lower bounds for dimensionality reducing maps“. In: *STOC*. 2013, S. 101–110.

[590] V. Poosala, Y. E. Ioannidis, P. J. Haas und E. J. Shekita. „Improved Histograms for Selectivity Estimation of Range Predicates“. In: *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*. 1996, S. 294–305.

[591] T. Sarlós. „Improved Approximation Algorithms for Large Matrices via Random Projections“. In: *47th Annual IEEE Symposium on Foundations of Computer Science FOCS*. 2006, S. 143–152.

[592] Q. F. Stout. „Isotonic Regression via Partitioning“. In: *Algorithmica* 66.1 (2013), S. 93–112.

[593] M. Thorup. „Quick k-Median, k-Center, and Facility Location for Sparse Graphs“. In: *SIAM J. Comput.* 34.2 (2005), S. 405–432.

[594] J. A. Tropp. „Improved Analysis of the subsampled Randomized Hadamard Transform“. In: *Advances in Adaptive Data Analysis* 3.1-2 (2011), S. 115–126.

[595] C. Villani. *Optimal transport, old and new.* Grundlehren der mathematischen Wissenschaften. Berlin: Springer, 2009.

[596] D. P. Woodruff und Q. Zhang. „Subspace Embeddings and ℓ_p-Regression Using Exponential Random Variables“. In: *Proceedings of COLT 2013 - The 26th Annual Conference on Learning Theory.* 2013, S. 546–567.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[C4/597] H. Binder, T. Müller, H. Schwender, K. Golka, M. Steffens, J. G. Hengstler, K. Ickstadt und M. Schumacher. „Cluster-localized sparse logistic regression for SNP data“. In: *Statistical Applications in Genetics and Molecular Biology* 11.4 (2012).

[TB1/C4/598] S. Canzar, T. Marschall, S. Rahmann und C. Schwiegelshohn. „Solving The Minimum String Cover Problem“. In: *Proceedings of the SIAM Meeting on Algorithm Engineering and Experiments, ALENEX 2012.* 2012, S. 75–83.

[C4/599] K. Ickstadt, B. Bornkamp, M. Grzegorczyk, J. Wieczorek, M. Sheriff, H. E. Grecco und E. Zamir. „Nonparametric Bayesian Networks (mit Diskussion)“. In: *Bayesian Statistics.* Hrsg. von J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith und M. West. Bd. 9. 2011, S. 283–316.

[C4/600] C. Köllmann, B. Bornkamp und K. Ickstadt. „Unimodal regression using Bernstein-Schoenberg-splines and penalties“. In: *Biometrics* (erscheint demnächst) (2014).

[A3/C4/601] M. Lohr, C. Köllmann, E. Freis, B. Hellwig, J. G. Hengstler, K. Ickstadt und J. Rahnenführer. „Optimal strategies for sequential validation of significant features from high-dimensional genomic data“. In: *Journal of Toxicology and Environmental Health, Part A* 75.8-10 (2012), S. 447–460.

[C4/602] H. Schwender, I. Ruczinski und K. Ickstadt. „Testing SNPs and sets of SNPs for importance in association studies“. In: *Biostatistics* 12.1 (2011), S. 18–32.

[C4/603] H. Schwender, S. Selinski, M. Blaszkewicz, R. Marchan, K. Ickstadt, K. Golka und J. G. Hengstler. „Distinct SNP combinations confer susceptibility to urinary bladder cancer in smokers and non-smokers“. In: *Plos One* 7.12 (2012).

[C4/604] C. Sohler und D. P. Woodruff. „Subspace embeddings for the ℓ_1-norm with applications“. In: *Proceedings of the 43rd ACM Symposium on Theory of Computing STOC 2011.* 2011, S. 755–764.

Andere Veröffentlichungen

[C4/605] L. Geppert, K. Ickstadt, A. Munteanu und C. Sohler. *Random projections for Bayesian regression.* Techn. Ber. 4. TU Dortmund, 2014.

[C4/606] V. Jabs. „Vergleich von Methoden zur Dimensionsreduktion unter Berücksichtigung der Rechenzeit und des Speicherbedarfs“. Bachelorarbeit. TU Dortmund, 2012.

[C4/C3/607] A. Munteanu und M. Wornowizki. *Demixing empirical distribution functions.* Techn. Ber. 2. TU Dortmund, 2014.

[A2/C4/608] C. Schwiegelshohn und C. Sohler. *Logistic Regression for Datastreams.* Techn. Ber. 1. TU Dortmund, 2014.

3.4 Planung des Teilprojekts

Ziele

Für die Fortführung des Teilprojekts verfolgen wir mehrere übergeordnete Ziele. In Phase 2 wollen wir Beschränkungen der in Phase 1 entwickelten Methoden zu linearer Bayes-Regression sowie zu logistischer

Regression überwinden. Weitere verallgemeinerte Modelle wie die Poisson-Regression stellen eine neue Herausforderung dar. Als wesentlichen Zweig der modernen Forschung nehmen wir auch nichtparametrische Regressionsmodelle auf, bei denen wir den Weg zur Hochdimensionalität vorbereiten wollen. Eine neue Perspektive auf die entwickelten Reduktionsverfahren greifen wir auf, indem wir sie auf Situationen übertragen, in denen zunächst keine Informationen über die Zielvariable vorliegen und wir sie somit als Auswahlverfahren für interessante Beobachtungen einsetzen. Neben Datensätzen mit vielen Beobachtungen wollen wir auch Datensätze mit vielen Variablen betrachten, bei denen wir Verfahren zur Merkmalsselektion weiter- oder neu entwickeln, wobei wir auch hier die algebraische Struktur der Datensätze erhalten wollen.

Im frequentistischen Fall wollen wir in Phase 2 bei logistischer Regression die starke Abhängigkeit von der Dimension reduzieren und die Datenstromalgorithmen auf weitere verallgemeinerte Regressionsmodelle wie Poisson-Regression übertragen. Im Bayesschen Fall sind unsere Methoden aktuell auf normalverteilte Likelihood sowie normalverteilte oder nichtinformative a priori-Verteilungen beschränkt. Hier werden wir zuerst an einer größeren Klasse von a priori-Verteilungen arbeiten. Dazu gehört auch die Betrachtung von hierarchischen Modellen, bei denen mehrere Ebenen an Vorinformationen zur Verfügung stehen. Schließlich wollen wir auch Verallgemeinerungen der Normalverteilungsannahme bei der Likelihood betrachten. Insbesondere Mischungen von Normalverteilungen sind sehr interessant, allerdings ist unklar, ob sich die Einbettungen darauf übertragen lassen.

Auch auf den hochdimensionalen Fall, in dem mehr Variablen als Beobachtungen vorliegen, werden wir in der zweiten Phase weiter eingehen. Hier ist die Untersuchung von Verfahren zur Merkmalsselektion vorgesehen. In der ersten Phase haben wir außerplanmäßig mit der Untersuchung von Hauptkomponentenregression für den Fall $n > d$ begonnen. Dies wollen wir in der zweiten Phase auf den Fall $n < d$ übertragen. Darüber hinaus bieten sich Samplingverfahren zur Merkmalsselektion an, wofür bereits bestärkende Vorarbeiten existieren. Weiterhin wollen wir die logische Regression, die speziell zum Auffinden von Interaktionen für binäre Variablen bis zu einer Größenordnung von $d \approx 1000$ geeignet ist, auf deutlich mehr Variablen übertragen.

Die Betrachtung diskreter Strukturen und anderer Struktureinschränkungen zur effizienten Berechnung von Regressionsmodellen wird in der zweiten Phase vertieft und auf nichtparametrische Modelle übertragen. Für viele solcher Einschränkungen ist die Entwicklung effizienter Algorithmen insbesondere für *Big Data* und im Datenstrommodell ein offenes Forschungsproblem.

Neu hinzu kommt die Untersuchung von Regression bei unvollständigen Angaben zur Zielvariablen. Dazu zählen Fälle, in denen die Erhebung der Y-Werte mit hohen Kosten verbunden ist.

Das Arbeitsprogramm ist in fünf Pakete gegliedert, in denen jeweils diese Ziele verfolgt werden:

AP1: Die Entwicklung von Datenstromalgorithmen für verallgemeinerte lineare Regression. In Phase 1 lag der Fokus auf der Entwicklung von Datenstromalgorithmen für lineare Regression. Nach Abschluss dieses Arbeitspakets wollen wir in Phase 2 auf verallgemeinerte Regressionsmodelle eingehen.

AP2: Das Fortführen der in Phase 1 (AP3) entwickelten effizienten Bayesschen Regressionsmodelle. Dazu gehört zunächst die Erweiterung der Klasse der a priori-Verteilungen bei Beibehaltung der normalverteilten Likelihood, später auch weitere Klassen von Likelihood-Verteilungen.

AP3: Die Untersuchung von nichtparametrischer Regression unter Strukturannahmen. In Phase 1 haben wir im Rahmen von Arbeitspaket 4 bereits unimodale Regression untersucht. Dies wird fortgeführt und zu isotoner Regression verallgemeinert werden. Neu hinzu kommt die Übertragung von Methoden aus der algorithmischen Geometrie auf die nichtparametrische Regression.

AP4: Regressionsanalyse bei unvollständigen Angaben über die Zielvariable. In dieses neue Arbeitspaket fällt lineare Regression, bei der die Erhebung der Zielvariablen mit hohen Kosten verbunden ist.

AP5: Selektion der wichtigsten Merkmale für den hochdimensionalen Fall. Hier sollen Ansätze zu Hauptkomponentenregression aus der ersten Phase fortgeführt werden, aber auch neue Ansätze zur Verbesserung von logischer Regression verfolgt werden.

Insbesondere in den Arbeitspaketen 1 und 2 werden wir Algorithmen und Methoden entwickeln, die auch im Datenstrommodell verwendet werden können. Die Unterraumeinbettungen, die ebenfalls hauptsächlich in diesen beiden Arbeitspaketen eingesetzt werden, lassen sich verteilt und im Datenstrommodell berechnen. So sind sowohl die Fälle von sehr großen Datenmengen als auch von verteilter Berechnung unter Ressourcenbeschränkung abgedeckt.

Arbeitsprogramm

Im Folgenden werden die fünf Arbeitspakete ausführlicher beschrieben.

Arbeitspaket 1. Datenstromalgorithmen für verallgemeinerte lineare Regression: In der ersten Phase haben wir uns mit Datenstromalgorithmen für lineare Regression beschäftigt. In der zweiten Phase wollen wir diese Ergebnisse generalisieren, indem wir uns verallgemeinerten Regressionsmodellen widmen. Bereits in Phase 1 haben wir uns im Rahmen des alten Arbeitspakets 2 mit der logistischen Regression als einem wichtigen Spezialfall der generalisierten linearen Regression beschäftigt. In dem für die logistische Regression entwickelten Relaxierungsmodell [A2/C4/608] gibt es noch einige offene Fragen. Es ist zwar zur Zeit möglich, den Zielfunktionswert $Z(v, A)$ für jeden Vektor v approximativ anzufragen, aber es ist noch nicht möglich einen Vektor v mit kleinem Zielfunktionswert aus der im Datenstrom aufrecht erhaltenen Datenstruktur effizient zu extrahieren. Es stellt sich also unmittelbar die Frage nach einem Algorithmus, welcher den gesuchten Vektor v direkt auf den bereits vorhanden Datenstrukturen berechnet. Zudem wäre die Kernmengeneigenschaft wünschenswert, welche ebenfalls eine Approximation des Zielfunktionswertes aller Vektoren erfordert, die benutzten Schätzer und Datenstrukturen aber im Wesentlichen auf gewichtete Punktmengen beschränkt. Die effiziente Berechnung von v auf einer solchen Punktmenge wäre bereits mit vorhandenen Algorithmen möglich. Ferner haben die Datenstrukturen zur Zeit eine Größe mit exponentieller Abhängigkeit von der Dimension, welche aus der Approximation der konvexen Hülle resultiert. Es stellt sich die Frage, ob es möglich ist, einen Algorithmus mit nur polynomieller Abhängigkeit von der Dimension anzugeben. Besonders gut ließen sich hierfür beispielsweise die Eigenvektoren der Datenmatrix anstelle der konvexen Hülle verwenden, da es für erstere effiziente Approximationsalgorithmen gibt. Es ist interessant zu untersuchen, ob die Eigenvektoren genügend Aufschluss über die Struktur der Punkte geben, um geeignete Relaxierungsmodelle zu definieren. Ferner lässt sich der Fragestellung nachgehen, inwiefern sich logistische Regression mit mehreren (konstant vielen) Durchläufen über den Datensatz approximieren lässt. Ab wann eine Relaxierung nicht mehr notwendig wird, ist hier noch offen.

Zu diesem Thema entwickelte Algorithmen werden wir beispielsweise an einer Fall-Kontrollstudie zum Harnblasenkrebs für wenige SNPs untersuchen, die für jeweils mehr als 2000 Fälle und Kontrollen vorliegen. Diese Analyse wird in Kooperation mit Jan Hengstler und Klaus Golka vom Leibniz-Institut für Arbeitsforschung an der Technischen Universität Dortmund (IfADo) erfolgen, die uns zu diesem Zweck die anonymisierten Daten dieser Studie zur Verfügung stellen werden. Zur Verifikation der in diesem Arbeitspaket entwickelten Verfahren und um die Reproduzierbarkeit von Ergebnissen zu gewährleisten, werden wir zudem simulierte Daten generieren.

Als weitere Verallgemeinerung des linearen Falls werden wir Datenstromalgorithmen für Poisson-Regression analysieren. Poisson-Regression wird in der Statistik häufig zur Analyse von Zählvariablen verwendet. Inwieweit eine Übertragung der Techniken, die für lineare Regression entwickelt wurden, möglich ist, wird zu untersuchen sein. Neben der gewöhnlichen Poisson-Regression bietet sich auch die Untersuchung von darauf aufbauenden Modellen wie der *zero-inflated* Poisson-Regression oder dem Poisson-Hurdle-Modell an. Diese werden verwendet, wenn die Zielvariable systematische Nullen aufweist, die das einfache Poisson-Modell nicht erklären kann.

Arbeitspaket 2. Effiziente Bayessche Regressionsverfahren: Das Ziel dieses Arbeitspakets ist, die in der ersten Projektphase entwickelten Algorithmen für Bayes-Regression weiter zu generalisieren. Wir definieren die Likelihood $\mathcal{L}(X, Y|\beta)$ unter der Unabhängigkeitsannahme als Produkt über die Verteilungen, die sich aus den Einzelbeobachtungen ergeben. Weiterhin legen wir eine a priori-Verteilung $p_{\text{pre}}(\beta)$ über dem Parameterraum fest, um problemspezifisches Vorwissen modellieren zu können. Die gesuchte a posteriori-Verteilung ergibt sich damit als:

$$p_{\text{post}}(\beta|X, Y) \propto \mathcal{L}(X, Y|\beta) \cdot p_{\text{pre}}(\beta).$$

Während unsere bisherige Forschungsarbeit [C4/605] von einer normalverteilten Likelihood und einer normalverteilten a priori-Verteilung ausgeht, haben sich bereits erste Anzeichen für Bedingungen ergeben, unter denen sich die betrachteten algorithmischen Techniken auf allgemeinere Klassen von Verteilungen generalisieren lassen. Besonders aussichtsreich erscheint zunächst die Betrachtung von Verteilungen, deren Definitionen sich der ℓ_p-Normen bedienen. Spezialfälle umfassen neben der Normalverteilung mit $p = 2$ die Laplaceverteilung mit $p = 1$.

Zunächst soll die Likelihood weiterhin als Normalverteilung modelliert werden. Die a priori-Verteilung hingegen soll aus einer möglichst allgemeinen Klasse von Verteilungen wählbar sein. Unter leichten Regularitätsbedingungen erscheint eine Verallgemeinerung sogar zu nahezu beliebigen Verteilungen möglich. Weiterhin werden wir hierarchische a priori-Verteilungen untersuchen. Das folgende Beispiel veranschaulicht

die Wirkung einer möglichen Modellierung:

$$||X\beta - Y||_2^2 + ||U\xi - \beta||_2^2 + ||\xi||_2^2.$$

Der erste Term modelliert, gegeben einen Parametervektor $\beta \in \mathbb{R}^d$, bis auf Konstanten die klassische negative Log-Likelihood der Daten $\mathcal{L}(X, Y|\beta)$, also die übliche Kostenfunktion eines ℓ_2-Regressionsproblems. Die zweite Komponente bedient sich einer Orthonormalbasis $U \in \mathbb{R}^{d\times k}$ für ein $k < d$ und spannt somit einen k-dimensionalen Unterraum des $\mathbb{R}^d$ auf. Damit wird auf dieser Hierarchieebene sozusagen *mit hoher Wahrscheinlichkeit* erzwungen, dass der Parametervektor β aus einem niedrigdimensionalen Unterraum stammt. Diese Tatsache kann dann wiederum ausgenutzt werden, um eine Einbettung auf der ersten Hierarchieebene zu erreichen, deren Einbettungsdimension nicht mehr von d sondern nur noch von k abhängt. Für den Fall $n \gg d \gg k$ könnte möglicherweise sogar eine weitere Einbettung auf zweiter Hierarchieebene eingesetzt werden, um die Orthonormalbasis zu skizzieren. Auf der dritten Ebene wird, wie sonst auf der zweiten schon üblich, ein Penalisierungsterm eingesetzt, um die Varianz von ξ und somit implizit auch von β zu senken.

Unter Verwendung neuerer Techniken zur verzerrungsarmen Unterraumeinbettung für ℓ_p-Räume, $p \in [1, \infty)$, erscheint eine Generalisierung der theoretischen Resultate der ersten Phase erfolgversprechend, aber technisch herausfordernd und insbesondere nicht direkt übertragbar.

Eine Verallgemeinerung der Likelihood hin zu Mischungen von Normalverteilungen wäre wünschenswert, da es mit diesen möglich ist, jede stetige Verteilung beliebig genau zu approximieren. Es ist bislang nicht klar, ob dazu ebenfalls ℓ_2-Unterraumeinbettungen verwendet werden können. Dies soll im weiteren Verlauf des Projekts untersucht werden. Im positiven Fall ist eine hohe Abhängigkeit von der Anzahl der an einer Mischung beteiligten Komponenten zu erwarten. Sollte sich die Technik nicht übertragen lassen, können adaptive nicht-uniforme Samplingverfahren und das Konzept von Kernmengen hilfreich sein, eine gute Approximation zu erhalten [568]. Bisherige Ansätze gehen von einer nahezu isotropen Struktur der beteiligten ℓ_2-Räume aus, um eine beweisbare Güte zu erzielen, oder es wird eine deutliche Separierung der Komponenten gefordert [553]. So wird die Kondition $\sigma_{\max}/\sigma_{\min}$ der beteiligten linearen Abbildungen beispielsweise beschränkt [563] oder stärker $\varepsilon \leq \sigma_i \leq 1/\varepsilon$ für die Singulärwerte gefordert [568]. Wir wollen eine Approximation unter möglichst schwachen Annahmen erreichen.

Die Effizienz der entwickelten Regressionsverfahren wird weiterhin vergleichend mit MCMC-Methoden analysiert, die bei Bayes-Regression standardmäßig eingesetzt werden.

Arbeitspaket 3. Nichtparametrische Regression unter Strukturannahmen: In diesem Arbeitspaket werden wir uns mit nichtparametrischer Regression unter Strukturannahmen beschäftigen. Wir werden insbesondere solche strukturellen Nebenbedingungen untersuchen, welche es erlauben, die erhobenen Daten unter Speicherbeschränkungen effizient darzustellen.

Das allgemeine Problem lässt sich wie folgt formal beschreiben: Es seien Beobachtungen $y \in \mathbb{R}^n$ und eine Gewichtsmatrix $G \in \mathbb{R}^{n\times n}$ gegeben. Weiterhin legen wir als Kostenfunktion die ℓ_p-Vektornorm $||\cdot||_p$ für $p \in [1, \infty)$ zugrunde. Zu minimieren ist nun die gewichtete Zielfunktion

$$\min_{\zeta\in\mathcal{N}} ||G(\zeta - y)||_p$$

unter Nebenbedingungen $\mathcal{N} \subset \mathbb{R}^n$ an ζ. Ohne diese Nebenbedingungen wäre jede Instanz trivial durch die Wahl $\zeta = y$ gelöst.

Bekannte Beispiele für Nebenbedingungen umfassen Monotonie, also $\mathcal{N} = \{\, \zeta \in \mathbb{R}^n \mid \zeta_1 \leq \ldots \leq \zeta_n \,\}$, sowie Unimodalität, welche durch $\mathcal{N} = \{\, \zeta \in \mathbb{R}^n \mid \exists i \in [n] : \zeta_1 \leq \ldots \leq \zeta_i \geq \zeta_{i+1} \geq \ldots \geq \zeta_n \,\}$ ausgedrückt werden kann. Beide Beispiele lassen sich als Spezialfälle der sogenannten *isotonen Regression* auffassen, bei denen die Variablen ζ_i einer beliebigen partiellen Ordnung unterliegen [592]. In den oben genannten Fällen ist die gewünschte Ausgabe, also ein Vektor ζ, der die Kostenfunktion minimiert, im Allgemeinen von linearer Beschreibungskomplexität. Um die beschriebenen Ressourcenbeschränkungen zu modellieren, werden wir fordern, dass die Repräsentation der Daten in höchstens B Teilstücke gegliedert ist. Ein solches Szenario wird beispielsweise in [576] verwendet, um sogenannte *Histogram Construction* Probleme zu lösen. Eine gültige Lösung ist dabei eine stückweise konstante Funktion, wobei höchstens B Unstetigkeitsstellen erlaubt sind. Das Problem lässt sich mittels dynamischer Programmierung in quadratischer Zeit und auf linearem Speicherplatz exakt lösen [569, 582, 575]. Durch approximative Auswertung der Kostenfunktion ist es möglich, eine $(1+\varepsilon)$-Approximation in Linearzeit und mit linearem Speicher zu berechnen. In Datenströmen ist es damit nur im Zeitfenstermodell für ein polylogarithmisch großes Fenster einsetzbar. Zum Preis eines logarithmischen Faktors in der Laufzeit kann eine $(1+\varepsilon)$-Approximation auch mit logarithmischer Speicherbeschränkung berechnet werden. Das Konzept lässt sich auf die stückweise Beschreibung mittels linearer Funktionen [559, 576] oder sogar auf gradbeschränkte Polynome verallgemeinern [576]. Ebenfalls

ist die Betrachtung verschiedener Kostenfunktionen möglich. So wird abgesehen von der ℓ_2-Norm auch die Summe der absoluten Abweichungen ℓ_1 in [590, 587] vorgeschlagen und in [566] die χ^2-Kostenfunktion.

In der Arbeit [558] wird die *total variation distance* als Distanzmaß zugrunde gelegt. Hier wird die Dichte einer unbekannten univariaten Verteilung stückweise durch Polynome vom Grad d approximiert. Der Algorithmus zieht dazu aus der Verteilung eine Anzahl von $\tilde{O}(Bd/\varepsilon^2)$ Punkten. Aus diesen kann in polynomieller Zeit eine Approximation der Dichtefunktion berechnet werden, die stückweise aus Polynomen besteht und eine *total variation distance* von $O(\tau) + \varepsilon$ aufweist, wenn die beste Approximation aus dieser Klasse von Funktionen eine *total variation distance* von τ zu der unbekannten Verteilung aufweist. Die Stichprobenkomplexität ist bis auf polylogarithmische Faktoren optimal. Die in der beschriebenen Arbeit erzielten Ergebnisse liefern für viele relevante Klassen von Verteilungen die besten bekannten Resultate. Diese umfassen beispielsweise die Klasse der log-konkaven Verteilungen, Mischungen von Normalverteilungen und viele mehr.

Eine weitere Möglichkeit, die Menge der erlaubten Lösungen einzuschränken, ergibt sich durch eine hierarchische Partition der Domäne. Sind beispielsweise die Punkte $x_i \in \mathbb{R}^2$, an denen die Werte y_i erhoben wurden, in einem umschließenden Quadrat in der Ebene gegeben, so lässt sich dieses mit einem Quad-Tree partitionieren. Diese Datenstruktur erlaubt es, eine hierarchische, nicht uniforme Zerlegung des Raumes in immer feinere Zellen zu berechnen. Ist die Approximation in einer Zelle noch nicht gut genug, so kann diese weiter zerlegt werden, bis die Zerlegung die Datenpunkte adäquat approximiert. Diese Methode lässt sich auf beliebige Dimensionen verallgemeinern. Der Speicheraufwand für n Punkte beträgt $\Theta(n)$.

Um nun eine Ressourcenbeschränkung zu modellieren, stellen wir die zusätzliche Bedingung, dass der Raum in höchstens B Zellen unterteilt wird. Das Ziel ist es, unter allen derart eingeschränkten Quad-Trees denjenigen zu finden, der die Daten am besten repräsentiert, wobei die Kosten eines Baumes definiert sind als die Summe der minimalen Fehler über alle Zellen der durch ihn beschriebenen Partition. Wir werden wie in [576] die Methode der dynamischen Programmierung verwenden, um eine optimale Lösung zu finden. Unter Ausnutzung der Monotonieeigenschaften der Zielfunktion kann diese in Abhängigkeit von der Anzahl von Unterteilungen und der Anzahl von Punkten für die einzelnen Bereiche effizient approximiert werden, um die Laufzeit und den Speicherbedarf für die Berechnungen zu senken.

Darauf aufbauend werden wir die beschriebenen Approximationstechniken verwenden, um eine $(1 + \varepsilon)$-Approximation für die beste Lösung idealerweise in Linearzeit und mit sublinearem Speicherbedarf zu berechnen. Eine Umsetzung im Datenstrommodell, bei dem die Daten nur einmal sequenziell eingelesen werden dürfen, ist darüber hinaus erstrebenswert und stellt ein wichtiges Ziel für uns dar, da Datenströme und Speicherbeschränkungen in eingebetteten Systemen häufig vorkommen und somit übergeordnete Ziele des SFB 876 verfolgt werden.

Eine weitere Möglichkeit, nichtparametrische Regressionsmodelle für niedrig-dimensionale Datensätze zu approximieren, stellen sogenannte Delaunay-Triangulationen dar. Dabei werden die Datenpunkte als Stützpunkte verwendet, um Höhenlinien zu interpolieren, indem die Bereiche zwischen den Punkten in Form von Dreiecken gegliedert werden und die Höhen der Eckpunkte entsprechend ihres Regressionswertes gewählt werden. In einer Dimension betrachtet, erscheint diese Modellierung der stückweise linearen Modellierung sehr ähnlich. Sie hat aber den Vorteil, dass die resultierende Triangulation stetig ist, während bei der stückweise linearen Modellierung Unstetigkeitsstellen an den Übergängen zwischen benachbarten Abschnitten auftreten können. Gleichzeitig erscheint deren Optimierung unter Ressourceneinschränkungen schwieriger, da sich die Fehler zwar immer noch unabhängig in den einzelnen Unterteilungen berechnen lassen, aber die Optimierung des Modells innerhalb eines Bereiches Einfluss auf die unmittelbar benachbarten Regionen hat. Dennoch sind diese Abhängigkeiten nur lokal gegeben, was vermuten lässt, dass das Problem einen ähnlichen Schwierigkeitsgrad besitzt.

Die Delaunay-Triangulation einer Punktmenge in der Ebene besitzt eine lineare Beschreibungskomplexität. Somit ist eine perfekte Anpassung an die Daten möglich, welche einen hohen Speicheraufwand von $\Theta(n)$ erfordert. Beides ist nicht wünschenswert, da zum einen eine Überanpassung an die Daten geschieht und zum anderen eine ressourcenbeschränkte Repräsentation nicht gegeben ist. Daher werden wir zusätzlich fordern, dass die Triangulation über höchstens B Punkte der Eingabe berechnet wird. Das Ziel ist es dann, die B-elementige Teilmenge S so zu wählen, dass die Delaunay-Triangulation von S die Kostenfunktion über alle Eingabepunkte minimiert. Zur Lösung des somit definierten Optimierungsproblems könnte die Lokalität der Abhängigkeiten zwischen benachbarten Regionen ausgenutzt werden oder eine weitere Einschränkung des Lösungsraums durch eine hierarchische Aufteilung der Ebene. Eine solche kann beispielsweise durch eine verschachtelte Triangulation erreicht werden [565]. Eine weitere Möglichkeit ist es, Quad-Trees zur hierarchischen Aufgliederung zu verwenden [547] oder auch sogenannte k-d-Trees, die eine flexiblere Aufteilung des Raumes erlauben. Somit ist eine Kombination der bereits erwähnten Techniken ebenfalls denkbar.

In der Literatur finden sich darüber hinaus weitere Techniken der algorithmischen Geometrie und der

Computergrafik, welche auf das statistische Umfeld übertragen werden können. So wird mit dem Begriff *Terrain Approximation* die direkte Konstruktion solcher approximativer Triangulationen beschrieben. Bei den in [571] beschriebenen Algorithmen handelt es sich um Greedy-Verfahren ohne theoretisch bewiesene Gütegarantien. Hier ist also eine Lücke, die wir in diesem Projekt schließen wollen. Weiterhin existieren auch Algorithmen, welche mit einem feinmaschigen Gitter beginnen und dieses sukzessive vergröbern. Diese Techniken, bei denen beispielsweise durch Kantenkontraktionen oder das Löschen von Knoten eine ausreichend grobe Darstellung berechnet wird [572], sind unter dem Begriff *Terrain Simplification* bekannt. Im Bereich der Computergrafik sind solche Techniken, die eine Vergröberung, aber auch wieder Verfeinerung zulassen und somit eine effiziente verlustfreie Datenkompression erlauben, auch als *progressive Meshes* [581] bekannt. Wir werden versuchen, die dazu verwendeten Techniken auf das statistische Problem zu übertragen, um damit eine ressourcenbeschränkte Regression der Daten zu erreichen.

Bei der eingangs erwähnten isotonen Regression sind Einschränkungen an die Beschreibungskomplexität der Ausgabe im eindimensionalen Fall untersucht worden [579], während viele Problemvarianten ohne Ressourcenbeschränkungen auch in höheren Dimensionen bereits optimal gelöst sind. Ein Überblick findet sich in [592]. Daher stellt auch die Betrachtung dieses Problems eine Herausforderung dar. So erscheint die Übertragung der Algorithmik von Regressionsproblemen unter Strukturannahmen in das Datenstrommodell vor allem durch die zusätzlich geforderten starken Restriktionen an die Speicherressourcen vielfach nicht-trivial und bietet ein weites Gebiet für die Forschung im Teilprojekt C4 in der zweiten Förderphase.

Arbeitspaket 4. Unvollständige Zielvariablen: In diesem Arbeitspaket werden wir uns mit der Situation beschäftigen, in der die Erhebung von Werten der Zielvariable Y einer (frequentistischen oder Bayesschen) linearen Regression mit sehr hohen Kosten verbunden ist. Wir nehmen an, dass zunächst keinerlei Informationen über die Zielvariable bekannt sind. Es soll aber möglich sein, den Wert zu jeder der Beobachtungen zu erheben, wodurch jedoch hohe Kosten pro Wert entstehen. Die Bereitstellung von Information stellt somit eine knappe oder teure Ressource dar, deren Verfügbarkeit durch ein Budget beschränkt ist. Aus diesem Grund wollen wir mit möglichst wenigen Werten von Y eine gute Approximation der unbekannten optimalen Lösung finden. Eine nicht-uniforme Gewichtung der Kosten ist hierbei auch denkbar. Zur Lösung solcher Probleme könnte das Konzept von Kernmengen nützlich sein. So existieren für ℓ_p-Regression samplingbasierte Kernmengenkonstruktionen, welche wenige Zeilen der Eingabe unabhängig von den Ausprägungen des Zielvektors auswählen [562]. Die verwendeten Wahrscheinlichkeiten richten sich dabei nach den statistischen *leverage scores* der Beobachtungen [586]. Die Lösung der Regression auf der Kernmenge ist mit hoher Wahrscheinlichkeit bereits eine $O(1)$-Approximation, welche Hinweise darüber liefern kann, auf welche Weise und wie viele weitere Zeilen adaptiv zu wählen sind, um die Approximationsgüte zu verbessern. Für den Spezialfall $p = 2$ ist der Regressionsfehler sogar mit hoher Wahrscheinlichkeit eine $(1 + \varepsilon)$-Approximation. Das Konzept von Kernmengen verfolgt ebenfalls, wie in unserer skizzierten Fragestellung, die Minimierung ihrer Größe bei beweisbarer Einhaltung von Gütegarantien und wird beim Entwurf und der Analyse geeigneter Algorithmen hilfreich sein. Wir werden daher untersuchen, wie gut ein solcher Ansatz als Ausgangspunkt für unser Problem geeignet ist. Ebenfalls können untere Schranken aus diesem Bereich der Forschung problematische Fälle aufzeigen und Aufschluss darüber geben, wie viele Anfragen mindestens notwendig sind [556]. Betrachtet man den Fall der ℓ_1-Norm, welche als *robustes* Fehlermaß für lineare Regression bekannt ist, so sind Ähnlichkeiten zum 1-Median Clusteringproblem zu erkennen, welches ebenfalls robuster gegen Ausreißer ist. Für das 1-Median Problem genügt eine kleine zufällige Stichprobe bereits, um eine approximative Lösung zu finden [652, 593]. Wir werden untersuchen, inwieweit sich solche Ansätze auf das allgemeine Problem unvollständiger Zielvariablen übertragen lassen. Zusätzlich bietet sich ein Vergleich mit oder auch eine Kombination dieser Methode mit Ansätzen aus der Versuchsplanung an. Hier eignen sich insbesondere algorithmische Versuchspläne, mit deren Hilfe wenige wichtige Punkte aus einem diskreten Raum bestimmt werden können [573].

Arbeitspaket 5. Merkmalsselektion: Ein typischer Ansatz zur Dimensionsreduktion ist die Hauptkomponentenanalyse (PCA). Dabei werden die Daten auf ihre k *wichtigsten* Hauptkomponenten projiziert. Dieses Verfahren minimiert die Varianz, die durch die Projektion verloren geht, unter allen Möglichkeiten, die Daten auf einen niedrig-dimensionalen Unterraum abzubilden. Hauptkomponentenregression bezeichnet ein Verfahren, welches zunächst die Hauptkomponenten der Datenmatrix berechnet und anschließend eine Regression nur noch auf den k wichtigsten Hauptkomponenten durchführt. Wir werden in diesem Arbeitspaket hauptsächlich die Hauptkomponentenanalyse insbesondere für Bayes-Regression untersuchen. Dazu hatten wir in der ersten Berichtsperiode erste Untersuchungen im Rahmen einer Abschlussarbeit [C4/606] durchgeführt, in der sich Hauptkomponentenregression gegenüber anderen dimensionsreduzierenden Verfahren bewährt hat. Zur Auswahl der Hauptkomponenten wird bei diesem Ansatz nur die Varianz der unabhängigen Variablen

in Betracht gezogen, nicht aber deren Einfluss auf die abhängige Variable. Dies kann dazu führen, dass Komponenten ausgewählt werden, welche für die Regression keine Rolle spielen, während wichtige Komponenten verworfen werden. In der Literatur wird mehrfach auf dieses Problem hingewiesen [584, 580, 577]. *Partial Least Squares* bezeichnet ein Verfahren, welches heuristisch versucht, die Korrelation zwischen den unabhängigen und der abhängigen Variable zu berücksichtigen. Im Extremfall einer perfekten Korrelation ignoriert *Partial Least Squares* die Varianz der anderen Komponenten, welche für Bayes-Regression, im Gegensatz zur frequentistischen Regression, wichtig ist. Im Normalfall tendieren *Partial Least Squares* und Hauptkomponentenregression dagegen zu ähnlichem Verhalten [293]. Um nun speziell für die Bayes-Regression eine bessere Auswahl der Komponenten zu treffen, kann die Projektion des Zielvektors auf die Hauptkomponenten mit in Betracht gezogen werden. Im frequentistischen Fall bedarf dieses Vorgehen bereits einer Laufzeit, in der es möglich ist, das Regressionsproblem exakt zu lösen. Im Bayesschen Kontext ergibt sich dennoch ein Vorteil, da die eigentlich rechenzeitaufwendige Ausführung von MCMC-Algorithmen von der geminderten Dimension profitieren kann. Ziel ist es also, eine verbesserte Auswahlregel für die Hauptkomponenten zu entwickeln, welche ein möglichst gutes Regressionsmodell ermöglicht. Dazu kann die Wassersteinmetrik ein Maß liefern, welches quantifiziert, wie gut die Verteilung der Parameter durch eine k-elementige Teilmenge der Hauptkomponenten approximiert wird. Weiterhin bestehen Ansätze, welche durch eine deterministische Auswahl oder Sampling von wenigen Spalten der Datenmatrix neue Komponenten bilden [555, 557]. Diese Verfahren haben den Vorteil, dass man eine *sparse*, d.h. dünn besetzte Repräsentation der Unterräume, erhält. Die Algorithmen haben allerdings eine Güte, welche relativ zur Güte der PCA-Lösung ist. Ist also eine Projektion auf die ersten k Hauptkomponenten schlecht für die Regression, wie oben beschrieben, so ist auch diese Gütegarantie entsprechend schwach. Wir wollen auch für solche Verfahren untersuchen, ob und wie sich die Spalten der Daten speziell für die Regression besser auswählen lassen, um eine bessere Approximationsgüte garantieren zu können.

Außerdem werden wir Merkmalsselektion mit der Methode der logischen Regression für die Zusammenarbeit mit dem Teilprojekt C1 weiterentwickeln, um dort anhand von (dummy-kodierten) SNP-Daten den Krankheitsstatus von Patienten zu erklären. Das Ziel einer logischen Regression ist das Auffinden von Interaktionen höherer Ordnung zwischen binären Variablen. Das Verfahren sucht nach Booleschen Verknüpfungen von logischen Variablen, mit denen die Zielvariable erklärt werden kann. Die logischen Variablen werden dabei aus den ursprünglichen Daten gebildet. Die Booleschen Formeln können durch logische Bäume repräsentiert werden.

Um einen logischen Baum L zu finden, wird eine Variable zufällig ausgewählt und als Blatt verwendet. Ausgehend von einem aktuellen Baum wird ein neuer Baum aus einer beschränkten Nachbarschaft vorgeschlagen. Der neue Baum wird, einem *Simulated Annealing*-Schema folgend, mit einer gewissen Wahrscheinlichkeit akzeptiert oder der alte Baum wird beibehalten. Die beiden Schritte, einen neuen Baum vorzuschlagen und ihn mit einer gewissen Wahrscheinlichkeit zu akzeptieren, werden k-mal wiederholt. Der so gefundene logische Baum wird abschließend in eine disjunktive Form gebracht. Logische Regression kann sowohl zur Klassifikation (ein Baum) als auch zur Regression verwendet werden. Für die Regression lässt man T logische Bäume wachsen und kombiniert sie mit einem generalisierten linearen Modell zu

$$f(E(Y)) = \beta_0 + \sum_{t=1}^{T} \beta_t L_t.$$

Durch eine geeignete Wahl des generalisierten linearen Modells kann logische Regression neben binären Daten auch auf Überlebenszeitdaten und quantitative Zielvariablen angewendet werden.

Logische Regression ist aktuell auf einige tausend Variablen begrenzt. Viele moderne SNP-Analysen wie genomweite Assoziationsstudien weisen eine deutlich höhere Anzahl an Variablen auf. Die Einsatzmöglichkeiten von logischer Regression insbesondere im genetischen Kontext sind damit beschränkt. Um hier Abhilfe zu schaffen, wollen wir ein zweistufiges Verfahren untersuchen, bei dem im ersten Schritt nur eine beschränkte Höhe der Interaktionen untersucht wird. In diesem Schritt soll die optimale Lösung nur approximiert werden. Im einem zweiten Schritt wird die gefundene Lösung dann verbessert, indem Teilmengen mit schlechter Anpassung genauer betrachtet und Interaktionen höherer Ordnung wieder einbezogen werden können.

Zeitplan

	2015				2016				2017				2018			
Arbeitspaket Quartal	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Datenstromalgorithmen für verallgemeinerte lineare Regression:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2. Effiziente Bayessche Regressionsverfahren:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
3. Nichtparametrische Regression unter Strukturannahmen:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
4. Unvollständige Zielvariablen:					■	■	■	■	■	■	■	■				
5. Merkmalsselektion:	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Die Arbeitspakete 1, 2, 3 und 5 sind für das Teilprojekt zentral; daher sollen sie kontinuierlich während der zweiten Phase bearbeitet werden. Arbeitspaket 4 wird im angegebenen begrenzten Zeitrahmen durchgeführt.

3.5 Stellung innerhalb des Sonderforschungsbereichs

Das Teilprojekt C4 beschäftigt sich mit Regressionsverfahren in verschiedenen Ausprägungen wie lineare (Bayes-)Regression, verallgemeinerte lineare Regression, nichtparametrische Regression und Regressionsmethoden zur Merkmalsselektion. Solche Regressionsverfahren sind für viele der Forschungsfragen in diesem SFB ein zentrales Werkzeug. In den Teilprojekten A3, TB1, B4 und C1 werden Regressionsmethoden eingesetzt und methodisch weiterentwickelt. Schwerpunkt der Arbeit in C4 liegt auf der Entwicklung von Verfahren, deren Güte sowohl aus mathematisch-statistischer Sicht als auch aus Sicht der Algorithmentheorie analysiert werden. Dabei sind Verfahren für sehr große oder hochdimensionale Datenmengen im Fokus des Teilprojekts.

Insbesondere besteht eine enge Zusammenarbeit mit Teilprojekt A2 bei der Entwicklung und Analyse von Datenstromalgorithmen, da sich die im Bereich der logistischen Regression entwickelten Methoden auch für die Anwendung auf andere Klassifikationsprobleme eignen. Diese Zusammenarbeit hat sich bereits in der ersten Förderperiode als sehr gewinnbringend herausgestellt und soll deshalb fortgesetzt werden. In Teilprojekt A3 liegt der Fokus auf der Modellauswahl aus vielen Modellen, wobei auch die in C4 entwickelten Regressionsverfahren verwendet werden können. Eine konkrete Schnittstelle ergibt sich durch den Einsatz der effizienten Regressionsverfahren aus C4 für das Surrogat-Regressionsmodell, das den Zusammenhang zwischen Parameterraum der Lernverfahren und der Modellgüte beschreibt.

Mit Teilprojekt TB1 soll die erfolgreiche Zusammenarbeit fortgesetzt werden. In der zweiten Phase werden wir uns vorwiegend über Methoden zur Merkmalsselektion für biologische Daten austauschen. Wir erhoffen uns einen Erkenntnisgewinn für beide Teilprojekte. Teilprojekt B4 setzt in der zweiten Phase generalisierte lineare Modelle wie Poisson-Regression zur Datenanalyse ein und kann von den Algorithmen und Methoden aus unserem Arbeitspaket 1 profitieren.

Mit Teilprojekt C1 besteht eine Kooperation bei der Analyse von genetischen Daten, die in der zweiten Förderphase weiter ausgebaut werden soll. Vor allem unsere Methoden zur Merkmalsselektion und Identifizierung von Interaktionen können einen wichtigen Beitrag zur Analyse der Datensätze in Teilprojekt C1 leisten. Darunter fallen sowohl die logische Regression als auch unsere geplante Verbesserung der Hauptkomponentenregression, die mit den Ergebnissen der in Teilprojekt C1 eingesetzten Hauptkomponentenregression verglichen werden soll.

Die Kooperation mit Teilprojekt C3 soll weiterhin gepflegt werden. So können unsere Regressionsmethoden dazu verwendet werden, die Energiespektren von physikalischen Teilchen zu beschreiben und vorherzusagen und auch zwischen verschiedenen Teilchen zu unterscheiden. Dies kann mit unseren Methoden nun auch auf großen Datenmengen erfolgen, die bisher nicht handhabbar waren. Im Gegenzug bietet sich für uns der Zugang zu wirklich großen Datenmengen zur Evaluation unserer Methoden, sodass sich für beide Projekte ein Mehrwert in der Zusammenarbeit ergibt.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

Sublinear Algorithms for the Analysis of Very Large Graphs (Sohler,Förderkennzeichen ERC-StG 307696) (Förderzeitraum: 12/2012 - 11/2017)

Das Projekt beschäftigt sich mit der Entwicklung und Analyse von Zufallsprozessen zur Analyse sehr großer Graphen im Rahmen des Property Testings. Es besteht kein Bezug zu Teilprojekt C4.

Statistische Methoden für Schädigungsprozesse unter zyklischer Beanspruchung (Ickstadt,Förderkennzeichen SFB 823, Teilprojekt B5) (Förderzeitraum: 07/2013 - 06/2017)

In diesem Teilprojekt steht die Prozessmodellierung und -analyse für Fragestellungen aus dem Bauingenieurwesen im Vordergrund. Diese Arbeiten überschneiden sich inhaltlich nicht mit dem Teilprojekt C4.

Messung und statistische Analyse von Clusterphänomenen bei Signalproteinen in der Plasmamembran (Ickstadt,Förderkennzeichen DFG-Projekt IC 5/3-1) (Förderzeitraum: 11/2011 - 10/2014)

In diesem Projekt werden räumliche und zeitliche Clustermethoden zur Beschreibung von Clusterphänomenen bei Proteinen entwickelt. Regressionsverfahren werden nicht eingesetzt; somit unterscheiden sich dieses Projekt und das beantragte Teilprojekt methodisch.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	117.600	—	117.600	—	117.600	—	117.600
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	117.600		117.600		117.600		117.600	

3.7.3 Beantragte Personalmittel für die neue Förderperiode

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (C4)							
Wissen-schaftliches Personal	1	Sohler, Christian, Prof. Dr., Prof	Algorithmische Komplexitätstheo-rie	LS 2, Fak. Inf, TU Do	5,0	—	Grund-ausstattung
	2	Ickstadt, Katja, Prof. Dr., Prof	Statistische Methoden	LS Mathematische Statistik und Biometrische Anwendungen, Fak. Statistik, TU Do	8,0	—	Grund-ausstattung
	3	Köllmann, Claudia, Dipl.-Stat., WiMi	Statistische Methoden	LS Mathematische Statistik und Biometrische Anwendungen, Fak. Statistik, TU Do	22,28	—	Grund-ausstattung
	4	N.N. , SHK	Algorithmische Komplexitätstheo-rie	LS 2, Fak. Inf, TU Do	5,0	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	5	Stankjawitschene, Alla, Verw. Ang.	—	LS 2, Fak. Inf, TU Do	0,99	—	Grund-ausstattung
	6	Kupiec, Brigitte, Verw. Ang.	—	LS Mathematische Statistik und Biometrische Anwendungen, Fak. Statistik, TU Do	0,99	—	Grund-ausstattung
Beantragte Personalmittel (C4)							
Wissen-schaftliches Personal	7	Munteanu, Alexander, Dipl.-Inf., WiMi	Algorithmische Komplexitätstheo-rie	LS 2, Fak. Inf, TU Do	—	Doktorand/in	—
	8	Geppert, Leo, Dipl.-Stat., WiMi	Statistische Methoden	LS Mathematische Statistik und Biometrische Anwendungen, Fak. Statistik, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Sohler, Christian**
   Projektleitung. Schwerpunkt Effiziente Algorithmen. Mitarbeit in allen Arbeitspaketen.

2. **Ickstadt, Katja**
   Projektleitung. Schwerpunkt Bayessche Statistik. Mitarbeit in allen Arbeitspaketen.

3. **Köllmann, Claudia**
   Arbeitspakete über nichtparametrische Regression unter Strukturannahmen sowie Merkmalsselektion.

4. **N.N.,**
   Implementierung der entwickelten Algorithmen

5. **Stankjawitschene, Alla**
   Sekretariat

6. **Kupiec, Brigitte**
   Sekretariat

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

7. **Munteanu, Alexander**
   Arbeitspakete über Bayessche Regressionsverfahren, nichtparametrische Regression unter Strukturannahmen, Regression bei unvollständiger Zielvariable und Merkmalsselektion seitens der Informatik.

8. **Geppert, Leo**
   Arbeitspakete über Datenstromalgorithmen für verallgemeinerte lineare Regression, Bayessche Regressionsverfahren, Regression bei unvollständiger Zielvariable und Merkmalsselektion seitens der Statistik.

3.7.4 Beantragte Sachmittel für die neue Förderperiode

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	5.000	5.000	5.000	5.000
Summe der vorhandenen Mittel	5.000	5.000	5.000	5.000
beantragte Ergänzungsausstattung	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt C5

3.1.1 Titel:

Echtzeitanalyse und Speicherung für hochvolumige Daten aus der Teilchenphysik

3.1.2 Fachklassifizierung:

Informationssysteme, Teilchenphysik

3.1.3 Projektleitung

Prof. Dr. Spaan, Bernhard, 25.04.1960

Experimentelle Physik 5, Fakultät Physik, Technische Universität Dortmund
Otto-Hahn-Straße 4
44227 Dortmund

Tel.: 0231-755-3662
Fax : 0231-755-4547
E-Mail: `Bernhard.Spaan@tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

Prof. Dr. Teubner, Jens, 22.03.1976

LS 6, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 14
44227 Dortmund

Tel.: 0231-755-6481
Fax : 0231-755-2405
E-Mail: `jens.teubner@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.1.4 Angaben zu forschungsrechtlichen Belangen:

– Untersuchungen am Menschen oder am menschlichen Material	( ) ja	(x) nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei	( ) ja	(x) nein
– klinische Studien	( ) ja	(x) nein
Ein Beleg der Registrierung der Studien liegt dem Antrag in Kopie bei	( ) ja	(x) nein
– Tierversuche	( ) ja	(x) nein
– gentechnische Untersuchungen	( ) ja	(x) nein
– Untersuchungen an humanen embryonalen Stammzellen	( ) ja	(x) nein
Die gesetzliche Genehmigung liegt vor	( ) ja	(x) nein
– Untersuchungen betreffend Übereinkommen über die biologische Vielfalt	( ) ja	(x) nein

3.2 Zusammenfassung

Das *LHCb*-Experiment am *Large Hadron Collider (LHC)* in Genf dient dazu, weitere Informationen zur Asymmetrie von Materie und Antimaterie zu gewinnen. Durch Präzisionsmessungen soll nach Hinweisen auf physikalische Prozesse gesucht werden, die im Standardmodell der Teilchenphysik nicht beschrieben werden können. Im Experiment werden die Ereignisse aus Kollisionen hochenergetischer Protonen aufgenommen. Dabei fallen Datenmengen in der Größenordnung von mehreren Terabytes pro Sekunde an. Konkret werden die Teilchendetektoren in der kommenden Ausbaustufe des LHC pro Sekunde 40 Millionen Kollisionen beobachten, zu denen jeweils etwa 25.000 Nachweispunkte aufgenommen werden.

Aufgrund der Menge können diese Daten nicht vollständig gespeichert werden, sondern müssen in Echtzeit auf einer Rechnerfarm analysiert und gefiltert werden. Ein Triggersystem (*„High Level Trigger“, HLT*) selektiert über 99,95 % aller Ereignisse bereits bei der Entstehung aus. Begrenzender Faktor in der Filtereffizienz sind jedoch die vorhandenen Rechenressourcen. Im Rahmen des Teilprojekts sollen daher Methoden entwickelt werden, um diese Effizienz nachhaltig zu steigern. Dazu sollen Ansätze des *Hardware/Software Co-Design*, insbesondere unter Verwendung von *Field-Programmable Gate Arrays (FPGAs)* und *Grafikprozessoren (GPGPUs)*, eingesetzt werden. Daraus erwarten wir insbesondere wichtige Erkenntnisse über den Aufbau und Einsatz heterogener Hardwaresysteme für datenintensive Anwendungen.

Am Ausgang des Triggers wird noch immer ein Datenstrom verbleiben, der sich auf etwa 20 PB pro Jahr summieren wird. Diese verbleibenden Daten müssen dauerhaft gespeichert werden und sollen idealerweise jederzeit für (möglichst interaktive) Analysen verfügbar bleiben. Während die bestehenden Strategien zur Datenhaltung in den Naturwissenschaften mit diesen Anforderungen überfordert sind, haben in den vergangenen Jahren verschiedene Internet- oder Cloudplattformen gezeigt, wie sich durch verteilte Speicher- und Analyseansätze eine Skalierbarkeit bis in die relevanten Größenordnungen realisieren lässt (insbesondere *MapReduce* und *Hadoop*; zahlreiche Datenbanken oder datenbankähnliche Systeme verfolgen bereits entsprechende Ansätze).

Im Teilprojekt wollen wir solche Ansätze für den Einsatz für extrem datenintensive Problemstellungen (wie denen in der Teilchenphysik) untersuchen. Anders als in zahlreichen Internetsystemen genügt es hier nicht, die Anfrageauswertung nur auf einem festen Satz an Metadaten auszuführen. Vielmehr müssen je nach physikalischer Fragestellung unterschiedliche Ereignisparameter berücksichtigt werden. Im Teilprojekt wollen wir daher die Daten bei der Verteilung so organisieren, dass sie nach den Anforderungen der verschiedenen Analysen effizient zugreifbar bleiben. Dies schließt eine genaue Betrachtung der gegebenen physikalischen Fragestellungen und ihrer Interaktion mit dem Speicher-Backend ein. Durch eine geeignete Abstimmung dieser beiden Systembestandteile lassen sich wertvolle CPU- und Datenzugriffsressourcen einsparen.

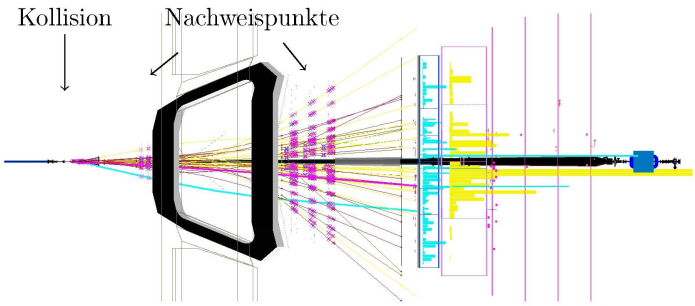
3.3 Ausgangssituation des Teilprojekts

Im folgenden werden wir zunächst das LHCb-Experiment am CERN vorstelen. Daraus ergeben sich wichtige Fragestellungen, die das Experiment bereits heute substanziell begrenzen, jedoch mit den bestehenden Methoden der Informatik nicht effizient gelöst werden können.

Large Hadron Collider (LHC) und das LHCb-Experiment: Der Teilchenbeschleuniger am CERN befindet sich derzeit (bis 2015) in der ersten *Long Shutdown*-Phase, in der der Beschleuniger selbst, aber auch die beteiligten Experimente ausgebaut werden. Ein Abschluss des Ausbaus des LHCb-Experiments ist ab dem Jahr 2019 geplant. Dabei soll das Experiment weitgehend neu gestaltet werden, insbesondere das Spurdetektorsystem, die Ausleseelektronik und die Rechnerfarm zur Vorfilterung der Ereignisse. Ab 2020 soll das Experiment dann mit wesentlich größeren Datenraten als bisher wieder in Betrieb gehen. Das beantragte SFB-Teilprojekt wird wichtige Beiträge zu diesem Ausbau und zur Bewältigung der anfallenden Datenmengen leisten.

Charakteristisch für die Experimente der Teilchenphysik ist die sehr geringe Wahrscheinlichkeit für das Auftreten relevanter Zerfallsprozesse. So wurde in 2013 ein Zerfallskanal von höchster Bedeutung gefunden, der allerdings in 10 Milliarden Ereignissen nur etwa dreimal zu finden ist [609]. Es ist daher auf der einen Seite erstrebenswert, das Experiment mit besonders hoher Luminosität zu betreiben, also die Anzahl der Kollisionsvorgänge pro Zeiteinheit zu erhöhen. Dadurch wächst allerdings das Problem mit den verfügbaren Ressourcen die gemessenen Ereignisse auch analysieren und interpretieren zu können.

Experiment: Das Anwendungsszenario ist in der Abbildung rechts dargestellt. Auf der linken Seite der Abbildung werden in einem Takt von 40 MHz Teilchen zur Kollision gebracht. Dadurch entstehen neue Teilchen, die in verschiedenen Bahnen durch einen Magneten (in schwarz dargestellt) fliegen und anschließend in einer Reihe von Spurdetektoren nachgewiesen werden. Die Messungen erfolgen im System der B-Hadronen, d. h. Hadronen, die ein b-Quark enthalten und zum Teil auch im System der Hadronen, die ein c-Quark enthalten.



LHCb Event Display.

Die Spurdetektoren des Experiments verfügen über eine sehr hohe Granularität mit einer entsprechend hohen Zahl von mehr als $4 \cdot 10^7$ Nachweiskanälen. Von diesen sprechen pro Ereignis etwa 25.000 an. Die einzelnen Nachweispunkte müssen dann den Teilchenspuren zugewiesen werden, um die Spurparameter am Ursprung und aufgrund der Spurkrümmung im Magnetfeld auch Teilchenimpulse bestimmen zu können (Spurrekonstruktion). Weiterhin durchqueren die Teilchen zwei *Ring Imaging Cherenkov*-Detektoren (RICH), durch die eine Identifizierung erfolgen kann sowie – in Verbindung mit der Impulsmessung – auch ihre Energie bestimmt werden kann. In den RICH-Detektoren wird Licht in einem geschwindigkeitsabhängigen Winkel zur Teilchenspur emittiert, das durch eine Optik dann auf einen Ring abgebildet wird. Für jeden dieser Ringe können dann etwa 10 Photonen nachgewiesen werden. Aufgrund der Vielzahl der Spuren können sich die Ringe überlappen bzw. gemeinsame Nachweiskanäle teilen. Das Finden der Ringe und deren Zuordnung zu Teilchenspuren stellen eine weitere Herausforderung dar.

Das LHCb-Experiment wurde von einer Kollaboration aus nunmehr über 1000 Mitgliedern aus etwa weltweit 60 Instituten konzipiert, entwickelt und gebaut und wird seit Beginn der Datennahme in 2010 von der Kollaboration betrieben. Mehr als 160 Publikationen in referierten Journalen sind seitdem erschienen. Die Arbeitsgruppe von Bernhard Spaan hat zum Bau des Experiments durch die Off-Detektor-Elektronik und durch das sogenannte Beam-Conditions-Monitor-System, das eine wichtige Schutzfunktion für das Experiment übernimmt, beigetragen. Die Arbeitsgruppe ist auch in der Datenanalyse sichtbar, wie zahlreiche Vorträge auf internationalen Konferenzen, die durch das Speakers Bureau der Kollaboration zugewiesen wurden, belegen. Die Dortmunder Arbeitsgruppe ist auch am weiteren Ausbau des Experiments beteiligt. So wurden in Dortmund die Grundlagen für den Bau eines Spurdetektors auf Basis szintillierender Fasern (SciFi-Tracker) gelegt, der nunmehr als Ersatz für das bestehende äußere Spurdetektorsystem, das vom Wechselwirkungspunkt aus gesehen hinter dem Magneten liegt, vorgesehen ist. Mitglieder der Arbeitsgruppe nehmen auch Koordinationsaufgaben in der Kollaboration wahr. So ist zur Zeit Bernhard Spaan der Vorsitzende des Collaboration Boards, d. h. der Institutsversammlung, in der die Institutsleiter vertreten sind.

Triggersystem: Ein *Triggersystem* soll den anfallenden Datenstrom bereits bei der Messung analysieren und vorfiltern. Das Triggersystem muss mindestens 99,95 % aller Rohdaten direkt verwerfen, um am Ausgang eine Datenmenge zu erhalten, die überhaupt mit existierenden Methoden und Ressourcen übertragen und gespeichert werden kann. Die Präzision des Triggersystems trägt dabei entscheidend zum Erkenntnisgewinn bei; nur mit einem präzisen Triggersystem ist die Steigerung der Luminosität überhaupt sinnvoll. Die Voranalyse der Daten im Triggersystem schließt eine Rekonstruktion der entstandenen Teilchenspuren ein – eine extrem rechenaufwändige Aufgabe, mit der bestehende Methoden trotz der verfügbaren Rechnerfarm mit 30.000 Prozessoren überfordert sind. Eine Steigerung der Effizienz lässt sich daher direkt übertragen in eine Steigerung der Präzision des Triggers und damit in einen verbesserten Erkenntnisgewinn in der Physik.

Ein wichtiger Teilbereich des Teilprojekts wird es sein, Methoden zu entwickeln, um die vorhandenen Rechnerressourcen für die Probleme der Teilchenphysik besser ausnutzen zu können. Dazu sollen sowohl Methoden entwickelt und angewandt werden, um *moderne Standardhardware* besser auszunutzen, als auch der Einsatz von *Hardware/Software Co-Design* (insbesondere in Form von FPGA- und GPGPGPU-basierten Koprozessoren) untersucht und evaluiert werden.

Speicherung von Messdaten: Datensätze, die nicht durch das Triggersystem unmittelbar ausselektiert werden, müssen dauerhaft gespeichert werden. Sie sind die Datenbasis, auf der Wissenschaftler regelmäßig neue Theorien oder Analysemodelle und -methoden evaluieren. Nach aktuellem Planungsstand der LHCb-Kollaboration werden nach dem Ausbau des Experiments am Triggerausgang etwa 20.000 Kollisionsdatensätze pro Sekunde verbleiben. Bei einer Datensatzgröße von ca. 100 kB pro Kollision und einer Messzeit von etwa

100 Tagen pro Jahr entspricht das einem Datenvolumen von über 20 Petabyte pro Jahr. Nach Möglichkeit sollen diese Daten so zugreifbar bleiben, dass Wissenschaftler nach wechselnden Kriterien Teile der Daten anfordern können, entsprechend den Kollisionsvorgängen, die für eine konkrete Analyse überhaupt relevant sind. Die Daten sollen von bis zu 1000 Mitgliedern der LHCb-Kollaboration weltweit genutzt werden.

Bestehende Ansätze sind mit diesem Datenvolumen und den Anforderungen an einen effizienten Zugriff überfordert. Bislang werden die Messdaten daher größtenteils nur auf Band weggeschrieben. Sie sind dann für konkrete Analysen nur noch schwer verfügbar. Für Analysen auf den eigentlichen Messdaten müssen sehr lange Wartezeiten in Kauf genommen werden, weil große Datenmengen erst von Band gelesen und nach dem Auftreten bestimmter Hadronen vorselektiert werden, bevor sie für die Analyse verwendet werden können. Als Workaround werden von den Datensätzen bislang nur Informationen zu markierten Hadronen auf Festplatten vorgehalten, was die Qualität und Möglichkeiten der Analysen jedoch einschränkt.

Der zweite Arbeitsbereich des Teilprojekts soll sich daher mit dem Einsatz hochskalierbarer, verteilter Speicher- und Rechenarchitekturen beschäftigen. Aus dem Umfeld großer Internetdienste und getrieben durch das *Cloud Computing* sind in jüngerer Vergangenheit Techniken entstanden, um auch sehr große Datenmengen effizient, skalierbar und fehlertolerant verarbeiten zu können. Im Teilprojekt soll der Einsatz solcher Techniken für die Analyse großer Datenmengen in der Teilchenphysik untersucht werden.

Die besonderen Herausforderungen – insbesondere auch mit Blick auf die anfallenden Datenmengen – liegen dabei insbesondere in zwei Bereichen:

(i) Für eine konkrete Analyse ist jeweils nur ein Teil des gesamten Datenbestandes überhaupt relevant. Die notwendige „Suche nach der Nadel im Heuhaufen“, das Vorselektieren der Datensätze gemäß gegebener Kriterien, ist dabei eine klassische Datenbankaufgabe. Die Aufgabe wird dadurch erschwert, dass die Selektionskriterien oft komplex sind und daher adäquate Indizierungs- und Optimierungstechniken benötigen. Im Unterschied zu den meisten anderen Anwendungen im Petabyte-Bereich ist dabei der gesamte Datenbestand relevant für die Auswertung von Anfragen, nicht nur z. B. die Metadaten dazu.

(ii) Nicht alle Selektionskriterien lassen sich auf übliche Bereichsprädikate zurückführen. Daher werden wir zunächst Formalismen entwerfen müssen, um die Kriterien überhaupt zwischen Speicher- und Analyseebene auszutauschen. Bei vielen Analysen ist es möglich, einzelne Ereignisse graduell zu untersuchen. Kann ein Ereignis während der Analyse früh verworfen werden, brauchen die verbleibenden Informationen gar nicht mehr gelesen werden. Mit einem Mechanismus, um ein solches Analysemuster auszudrücken, könnte ein signifikanter Anteil des Aufwands und Ressourcenverbrauchs einer Analyseoperation reduziert werden.

Physikalische Analysen: Der eigentliche Zweck des LHCb-Experiment ist es, aufgestellte Modelle zur Teilchenphysik mittels der gemessenen Daten zu überprüfen. Die wesentliche Arbeit der Physiker besteht – neben dem Aufbau und der Wartung des Experiments – darin, Modelle zu entwickeln und zu verfeinern. Die entwickelten Modelle werden als Softwareprogramme beschrieben und auf großen Rechnerfarmen evaluiert.

Zur Evaluation steht den Wissenschaftlern das Worldwide LHC Computing Grid (WLCG) zur Verfügung, an dem sich neben dem CERN eine Vielzahl von Institutionen aus der ganzen Welt beteiligen. Im WLCG-Verbund werden sowohl Rechenressourcen als auch die vorhandenen Messdaten ausgetauscht und koordiniert; anfallende Analyseaufgaben können prinzipiell auf alle Rechenzentren des WLCG verteilt werden. Aufgrund der anfallenden Datenmengen ist die Koordination der Messdaten eine zentrale Herausforderung. Aktuell wird von allen Messdaten mindestens eine Kopie am CERN auf Band gespeichert. Außerdem werden Replikate der Messdaten anteilig an zwölf Tier 1-Rechenzentren an internationalen Standorten verteilt (u.a. nach Karlsruhe). An diese Tier 1-Rechenzentren sind wiederum etwa 170 Tier 2-Rechenzentren angeschlossen, an die ebenfalls Teile der Daten weitergeleitet werden (entsprechend ihrer Aufnahmekapazität). Schließlich können sich Universitäten und andere Institutionen auch mit ihren eigenen Rechenressourcen beteiligen (Tier 3; u.a. auch die Dortmunder Physik) und Anteile der Daten erhalten. Tier 3-Cluster werden allerdings nicht mehr durch das WLCG zentral koordiniert.

Die WLCG-Ressourcen können durch Wissenschaftler aller beteiligten Institutionen genutzt werden. Dazu senden sie ihre Analyseprogramme, zusammen mit einer Beschreibungsdatei, an das Scheduling-System des WLCG. Dort wird versucht, ein Rechenzentrum zu finden, in dem *(a)* genügend Rechnerkapazitäten gemäß den Anforderungen in der Beschreibungsdatei verfügbar sind und in dem *(b)* möglichst viele der benötigten Daten bereits lokal vorhanden sind, um somit größere Netzwerktransfers zu vermeiden. Kann das Programm dort nicht ausgeführt werden (weil nicht genügend Ressourcen vorhanden sind oder weil bei der Ausführung ein Fehler aufgetreten ist), wird zunächst versucht, die Aufgabe an andere Rechenzentren weiterzureichen. Im wiederholten Fehlerfall wird eine *abort*-Meldung an den aufrufenden Benutzer zurückgeliefert.

Charakteristika physikalischer Analysen: Experimente in der Teilchenphysik zeichnen sich oft durch eine sehr hohe Datenintensität aus. Interessante Ereignisse treten mit Wahrscheinlichkeiten von nur 10^{-12} und weniger auf. Von allen aufgetretenen Ereignissen nehmen die Detektoren am CERN grundsätzlich nur Spuren innerhalb eines engen Raumwinkels auf, wodurch wiederum nur ein Teil der interessanten Ereignisse überhaupt beobachtet werden kann. Die für die Physik besonders interessanten Ereignisse sind gleichzeitig besonders aufwändig zu detektieren und zu interpretieren.

Um Aussagen mit hinreichender statistischer Signifikanz treffen zu können, müssen daher möglichst viele Kollisionsereignisse untersucht werden. Das ist einer der Hintergründe für den geplanten Ausbau des Experiments, mit dem insbesondere die Kollisionsraten deutlich nach oben gefahren werden sollen. Gleichzeitig steigen damit natürlich die Anforderungen, die vielen gemessenen Ereignisse überhaupt analysieren zu können. Spurrekonstruktion und die Zuordnung von Messdaten der RICH-Detektoren beinhalten kombinatorische Aufgaben, deren Schwierigkeit mit der Anzahl der gemessenen Spuren exponentiell zunimmt. Besonders das Triggersystem steht hier vor einer großen Herausforderung, schließlich sind an dieser Stelle fälschlich verworfene Ereignisse unwiederbringlich verloren.

Die mit dem LHCb-Experiment verbundenen physikalischen Fragestellungen betreffen spezielle Zerfallskanäle und deren Eigenschaften, die zum Teil äußerst selten auftreten. So wurde in 2013 ein Zerfallskanal von höchster Bedeutung gefunden, der im Mittel in nur etwa 3 mal in 10 Milliarden Ereignissen mit B-Hadronen zu finden ist. Ein weiteres Ziel der Analysen ist, weitere Informationen zur Materie- und Antimaterie-Asymmetrie im Universum zu gewinnen und durch Präzisionsmessungen Hinweise auf physikalische Prozesse zu finden, die im Standardmodell der Teilchenphysik nicht beschrieben werden können. Eine Vielzahl relativ seltener Prozesse muss daher mit maximaler Effizienz selektiert werden. Dabei muss gewährleistet sein, dass für die späteren Analysen diese erste Selektionseffizienz mit höchster Genauigkeit bekannt ist. Von besonderer Bedeutung ist dabei die Spurfindung und Spurrekonstruktion.

3.3.1 Stand der Forschung und eigene Vorarbeiten

Heterogenität / Nicht-Standardkomponenten: Die Ursache der Leistungsbegrenzung im Triggersystem ist im Kern eines der Grundprobleme der Informationstechnologie. Zwar lässt sich aus *Moore's Law* (die Anzahl der Transistoren pro Chip wächst exponentiell mit der Zeit) eine nominelle Leistungssteigerung von Computersystemen ableiten, die auch für die kommenden Jahre weiter so erwartet wird. Jedoch wird diese Steigerung begrenzt durch verschiedene Ressourceneinschränkungen. Insbesondere der Energieverbrauch ist heute eine zentrale Beschränkung in der Entwicklung von Rechensystemen [613].

Als erfolgversprechendster Ausweg aus dieser Problematik wird gemeinhin das *Hardware/Software Co-Design* betrachtet. Das bedeutet, dass Hard- und Software gezielt aufeinander abgestimmt werden müssen, um das wahre Leistungspotenzial der Hardware auch tatsächlich auszunutzen. Diese Notwendigkeit wird bereits bei aktueller Standardhardware offenbar. Ihre volle Leistung kann meist nur genutzt werden, wenn Algorithmen gezielt auf die Begebenheiten der Hardware angepasst werden. Beispiele dafür finden sich insbesondere in den Arbeiten von Jens Teubner, wobei cache-optimierte [C5/642] und SIMD-optimierte [C5/644] Algorithmen zur Lösung von Datenbankproblemen vorgeschlagen wurden, die bestehenden Methoden oft um ein Vielfaches überlegen waren.

Eine noch höhere Ressourceneffizienz lässt sich durch den gezielten Einsatz von Nicht-Standardkomponenten erreichen. Das wohl bekannteste Beispiel hierfür sind Grafikprozessoren (GPGPUs), die besonders geeignet sind, massiv datenparallele Aufgaben zu realisieren. Einen wesentlich höheren Grad an Spezialisierung kann man mit Hilfe von programmierbaren Logikbausteinen (*Field-Programmable Gate Arrays, FPGAs*) erreichen. Sie eignen sich besonders für eine strombasierte Verarbeitung sehr großer Datenmengen. In beiden Fällen kann ein Performanzvorteil allerdings nur erreicht werden, wenn das zugrunde liegende Problem entsprechend repräsentiert und parallelisiert werden kann. Bestehende Arbeiten aus der Datenbanktechnologie zeigen, wie das für einzelne Probleme bewerkstelligt werden kann.

Grafikprozessoren (GPGPUs). Grafikprozessoren zeichnen sich durch sehr hohe Parallelitätsgrade und eine sehr hohe Rechendichte aus. Bereits sehr früh wurde insbesondere im Umfeld von Datenbanksystemen gezeigt, wie sich die Architekturvorteile für datenintensive Probleme ausnutzen lassen. So nutzten Govindaraju et al. [622] frühe Grafikprozessoren zur Auswertung von Selektionsprädikaten in Datenbanken und zur Aggregation. In [623] wurde ein Sortieralgorithmus für Grafikkarten vorgestellt. He et al. [626] wendeten Grafikkarten zur Berechnung von Joins, einer Kernaufgabe von Datenbanksystemen, an und stellten später *GDB* vor [627], einen Prototyp für ein Datenbanksystem mit GPGPU-basiertem Co-Prozessor. Das Zusammentreffen verschiedener Features in Grafikkarten und Standardprozessoren stellt noch immer Herausforderungen an die Erstellung geeigneter Algorithmen und deren Aufteilung auf die vorhandenen Recheneinheiten (z. B. [614]). Ein von Heimel et al. [628] vorgeschlagener Lösungsansatz ist das Ausnutzen von plattformunabhängigen

Algorithmus-Repräsentationen wie etwa OpenCL.

In der experimentellen Teilchenphysik werden Ansätze verfolgt, GPGPUs in die Rechnerfarmen und den Trigger einzubinden, um die Qualität der Trigger-Entscheidungen zu steigern [625]. Insbesondere hat sich das Tracking, d. h. die Spurfindung und Bestimmung der Spureigenschaften als lohnendes Feld erwiesen [619]. So verfügt das ALICE-Experiment am LHC bereits über ein GPGPU-basiertes Online-Tracking für den zentralen Spurdetektor (TPC) [621]. Dort wurde eine Performanzsteigerung um einen Faktor von ca. 2,5 gegenüber einem CPU-basierten Tracking beobachtet. Die dortige Lösung lässt sich allerdings nicht unmittelbar auf das LHCb-Experiment übertragen; anders als in LHCb liegt die Komplexität des Problems weniger in der produzierten Ereignisrate, sondern vielmehr in der Komplexität jedes einzelnen Ereignisdatensatzes. Andere Experimente wie ATLAS und CMS verfügen über Hardwaretrigger, mit denen die Ereignisrate in ersten Schritten drastisch reduziert wird, sodass das Tracking erst bei deutlich verringerten Ereignisraten relevant wird.

Field-Programmable Gate Arrays (FPGAs). FPGAs lassen im Vergleich wesentlich höhere Freiheitsgrade zu; mit ihnen lassen sich elektronische Schaltungen realisieren, die ein gegebenes Problem direkt in Hardware lösen können. Insbesondere sind FPGAs nicht begrenzt auf reinen Datenparallelismus (wie etwa GPGPUs), sondern lassen beliebige Kombinationen z. B. mit Pipeline-Parallelismus zu. Sie sind damit besonders gut geeignet für Anwendungen, deren Schwierigkeit in einem sehr hohen Durchsatz bei nur mäßiger Problemkomplexität liegt.

Aus der Gruppe von Jens Teubner sind mehrere wichtige Arbeiten zu FPGA-basierten Lösungen mit und ohne Zusammenspiel mit Standard-CPUs entstanden. Etwa wurde gezeigt, wie sich Datenstromanalysen systematisch und effizient in FPGA-Schaltungen übersetzen lassen [C5/643, C5/647], wie sich XML-Datenströme mittels FPGAs in Echtzeit analysieren lassen [C5/645], oder wie sich Datenanalyse-Aufgaben auch vollständig im FPGA lösen lassen [C5/646]. Ähnliche Ansätze wurden etwa von Moussalli et al. [633] oder Najafi et al. [634] vorgeschlagen.

In der Arbeitsgruppe von Bernhard Spaan wurde für den bestehenden LHCb-Detektor bereits eine FPGA-basierte Datenvorverarbeitung für das äußere Spurkammersystem auf Basis von ALTERA Stratix II-FPGAs realisiert [C5/649, C5/641]. Das System ist verantwortlich für den Empfang der Daten von verschiedenen Detektormodulen mit einer Rate von 10^6 Ereignissen pro Sekunde, für die Datenvorverarbeitung und Weiterleitung der Daten an das Triggersystem. Die Arbeitsgruppe hat darüber hinaus für das LHCb-Experiment den so genannten Beam Conditions Monitor (BCM) entwickelt und gebaut [C5/648]. Dies ist ein weiteres FPGA-basiertes System mit höchsten Anforderungen an Zuverlässigkeit, bei dem Sensoren in der Nähe des Kollisionspunktes den Teilchenstrom vermessen. Übersteigt der Strom bestimmte Grenzwerte, die vom experimentellen Setup abhängen, löst das BCM-System einen Beam-Dump, einen Abort beider Protonenstrahlen des Large Hadron Colliders, zum Schutz des LHCb-Experiments aus.

Grenzen heterogener Systemansätze. Die bestehenden Arbeiten zeigen jeweils eindrucksvoll, wie einzelne (Teil-)Probleme auf Basis von Nicht-Standardkomponenten effizient gelöst werden können. Es ist aber gleichzeitig noch sehr wenig verstanden, wie das Zusammenspiel verschiedener Arten von Recheneinheiten – also ein heterogenes System – möglichst gewinnbringend und ressourceneffizient zur Problemlösung ausgenutzt werden kann. Dies trifft erst recht zu, sobald Problemstellungen verallgemeinert werden sollen oder auf konkrete Anwendungsanforderungen anwendbar sein sollen.

Verteilte Datenhaltung: Im Experiment müssen Datenmengen mindestens im zweistelligen Petabyte-Bereich dauerhaft abgespeichert werden und dabei für Analysen der Wissenschaftler effizient zugreifbar bleiben.

Als erstes Experiment der Teilchenphysik verfügte das BABAR-Experiment, das 1999 die Datennahme aufnahm, über eine Datenmenge von mehr als 1 Petabyte. Zum Management dieser Datenmenge wurde ein Datenbanksystem eingesetzt, wobei sehr innovative Ansätze in der Datenspeicherung entwickelt wurden. Es zeigte sich aber, dass dieses Datenbanksystem den Anforderungen, die sich aus einer Zahl von mehr als 500 Usern ergaben, nur bedingt gewachsen war [612], sodass das BABAR-Experiment für die Datenspeicherung auf ROOT [615] gewechselt ist. Da im Rahmen des LHCb-Experiments ROOT nunmehr auch an Grenzen stößt, soll ein neues Datenspeicherungs- und Zugriffsmodel entwickelt werden, das auch zum Teil die innovativen Ideen des BABAR-Experiments wieder aufgreift.

Vergleichbare Datenvolumina treten sonst fast nur im Kontext sehr großer Internetdienste (Cloud Computing) auf. Dort werden die Daten in großem Stil verteilt. Allerdings entwickeln sich auch weitere Disziplinen wie die Medizin zügig in Richtung Big Data, sodass die Arbeiten im Teilprojekt in einem deutlich größeren Umfeld zum Tragen kommen könnten. Durch die enge Zusammenarbeit von Yandex mit der LHCb-Kollaboration sehen wir auch ein großes Potential für den weiteren Einsatz unserer Entwicklungen im Bereich der Internet-Dienste.

Die Zielsetzungen der bestehenden Arbeiten sind dabei höchst unterschiedlich. Systeme wie das *Google File System* [620], *Ceph* [637] oder das *Hadoop Distributed File System (HDFS)* [636] orientieren sich primär an Dateisystemen. Die Architektur des Google File System etwa unterstützt besonders gut Schreiboperationen mit *append-only*-Charakteristik; Leseoperationen können sowohl kleine als auch beliebig große (Teil-)Daten effizient lesen. Bei Ceph werden nicht nur die eigentlichen Daten verteilt abgelegt. Vielmehr werden auch Metadaten verteilt verarbeitet, wodurch das System Metadatenoperationen mit POSIX-ähnlicher Semantik effizient bearbeiten kann. *HDFS* ist Teil des *Hadoop*-Pakets und wird dort als Basis für verteilte Rechenoperationen nach dem *MapReduce*-Prinzip [618] verwendet.

Andere Ansätze orientieren sich stärker an Datenbanksystemen. *BigTable* [616] stellt auf Basis des Google File Systems ein einfaches tabellarisches Datenmodell zur Verfügung. *HBase* [610] stellt ähnliche Funktionalität auf Basis von Hadoop bereit. Systeme wie Amazons *Dynamo* [617] oder *Cassandra* [631] schließlich werden auch als *key/value stores* bezeichnet und erlauben sehr hohe Transaktionsraten (auf Kosten eines gegenüber klassischen Datenbanksystemen abgeschwächten Konsistenzmodells).

Den vorgenannten Ansätzen ist gemein, dass sie zur Datenverteilung Hashfunktionen verwenden (z. B. mittels *consistent hashing* [630]). Das hat den Vorteil, dass auch Daten mit unbekannter Struktur und unbekannter Werteverteilung gleichmäßig auf die vorhandenen Netzwerkressourcen verteilt werden können. Viele Zugriffe müssen dann als parallele Scans implementiert werden. Das hält zwar die Übersetzung von Anfragen/Zugriffen einfach, doch skaliert die Strategie nur begrenzt auf sehr großen Datenmengen (über 1000 Knoten verteilt bedeutet ein Scan über 1 PB Daten noch immer ein Scanvolumen von 1 TB pro Knoten).

Klassisch werden in verteilten Datenbanken wertebasierte Partitionierungsstrategien – unter Ausnutzung von Kenntnissen über die Daten und ihre Werteverteilung – verwendet. Damit wird versucht, spätere Anfragen möglichst lokal nur auf wenigen Knoten des Systems ausführen zu können. Prinzipiell sind auch im LHCb-Experiment viele Charakteristika der Messdaten bekannt, ebenso wie typische Anfragekriterien für den späteren Zugriff. Es ist unklar, inwieweit das auch ausgenutzt werden könnte, um auch bei einem Cloud Computing-basierten Speicheransatz für LHCb einen möglichst hohen Lokalitätsgrad zu erreichen.

Verteilte und skalierbare Analysen: Die passende Ergänzung zu verteilten Datenhaltungssystemen sind Plattformen, die eine hochgradig verteilte Ausführung einer Rechenaufgabe, z. B. einer Datenanalyse, erlauben. Ebenfalls aus der Welt der Internetsysteme haben sich hier in den vergangenen Jahre Modelle nach dem Vorbild des MapReduce-Ansatzes von Google [618] etabliert. *Hadoop* entstand als Open Source-Alternative zum proprietären MapReduce-System. Die *Dryad*-Plattform von Microsoft [629] bietet ähnliche Funktionalität.

Aufbauend auf der MapReduce-Idee wird in verschiedenen Ansätzen und Systemen versucht, die Ausdrucksmächtigkeit und den Programmierkomfort zu erhöhen, ohne die positiven Eigenschaften (Skalierbarkeit und Fehlertoleranz) des Modells zu verlieren. Hierzu zählen Dryad [629], dessen Programmiermodell aus einem Graphen von Rechenschritten besteht, Stratosphere/PACT [611], das sogenannte *parallelization contracts* einführt, aber auch Apache Pig oder Hive, beides Beschreibungssprachen, die SQL-ähnliche Konstrukte in skalierbare MapReduce/Hadoop-Programme übersetzen. Im resultierenden Grenzgebiet zwischen Programmiersprachen, verteilten Systemen und Datenbanktechnologien ergeben sich besonders interessante Möglichkeiten, auch komplexe Analyseprobleme in effizient ausführbare Programme übersetzen zu lassen. Es ist nicht überraschend, dass einige Ansätze diese Zusammenhänge sehr explizit machen, etwa LINQ [632] bzw. DryadLINQ [638] oder Ferry [624]. Frühere Arbeiten von Jens Teubner (im Kontext des XQuery-Übersetzers *Pathfinder*) leisteten ebenfalls wichtige Beiträge zu diesem Gebiet.

Viele der Aufgaben, die diese Systeme (zumindest in ihren jeweiligen Anwendungsdomänen) bereits größtenteils automatisch lösen können, werden in den Analyseanwendungen der Teilchenphysik überwiegend von Hand oder mit Hilfe einfacher Werkzeuge bearbeitet. Die Einführung effizienter Ansätze zum skalierbaren Umgang mit großen Datenmengen könnte hier wertvolle Personal- und Sachressourcen einsparen. Aus ähnlichen Gründen orientieren sich auch andere datenintensive Anwendungsfelder in Richtung der oben skizzierten Programmiermodelle. Etwa schlagen Schatz und Langmead [635] vor, MapReduce-ähnliche Modelle auch in der DNA-Sequenzierung zu verwenden.

Die Arbeitsgruppe von Bernhard Spaan hat in der Vergangenheit intensiv im Bereich des verteilten Rechnens im DGRID-Projekt und vielen Abschlussarbeiten Untersuchungen durchgeführt.

Analysen in der Teilchenphysik (Flavourphysik): Die Flavourphysik ist ein zentrales Forschungsfeld im Bereich der Teilchenphysik. Das Feld eignet sich bestens für sensitive Tests des Standardmodells, das zwar die beobachtbaren Prozesse der Teilchenphysik gut beschreibt, aber keine Erklärung für das Vorhandensein von Dunkler Materie oder für die Materie-Antimaterie-Asymmetrie im Universum liefert. Daher wird nach Effekten von sogenannter *Neuer Physik* gesucht, die nicht Teil des Standardmodells der Teilchenphysik sind.

Da der Großteil aller Parameter des Standardmodells im Bereich der Flavourphysik liegen, ist die Hoffnung besonders groß, hier durch Präzisionsmessungen Abweichungen von den Vorhersagen des Standardmodells zu finden. Die schweren b-Quarks sind dabei besonders vielversprechende Kandidaten. Teilchen, die b-Quarks enthalten, zeigen besonders große Effekte der CP-Verletzung, d.h. Teilchen zeigen ein anderes Verhalten als ihre Antiteilchen. Darüber hinaus zeigte die Beobachtung von Teilchen-Antiteilchen-Oszillationen bei neutralen B-Mesonen ($B\bar{B}$-Oszillationen), die durch die deutlich schwereren top-quarks hervorgerufen werden, dass es hier berechenbare Effekte gibt, die auf Massenskalen weit jenseits der Masse eines b-Quarks hinweisen. Die Flavourphysik mit b-Quarks hat in den letzten Jahren sehr viel Aufmerksamkeit erfahren – nicht zuletzt durch die Erfolge der sogenannten B-Fabriken BABAR und Belle, die aber keine Abweichungen vom Standardmodell gefunden haben. Daher können sensitive Tests des Standardmodells nur durch Messungen mit deutlich höherer Präzision oder Messungen in Bereichen, die den B-Fabriken nicht zugänglich waren, erfolgen. Das LHCb-Experiment erfüllt beide Voraussetzungen durch deutlich höhere Ereignisraten und durch Zugang zu den sogenannten B_s-Mesonen, d.h. Mesonen, die neben dem b-Quark noch ein s-Quark enthalten. Wesentliche Grundlage für die Präzisionstests ist, dass die Genauigkeit der Vorhersagen mit entsprechend geringen theoretischen Unsicherheiten behaftet ist.

Schwerpunkte des Tests des Standardmodells mit B-Hadronen liegen auf dem Gebiet der Messung zeitabhängiger CP-Asymmetrien sowie auf Untersuchungen zu seltenen Zerfällen.

Im Rahmen der Arbeitsgruppe Spaan wurde bereits in der Vergangenheit am BABAR-Experiment (SLAC, Stanford, USA) intensiv nach seltenen B-Zerfällen gesucht. Auch sind Mitglieder der Arbeitsgruppe an Messungen zum Kanal $B_s \to \mu^+\mu^-$ mit dem LHCb-Experiment beteiligt. Die Untersuchung seltener und gut berechenbarer Prozesse stellen sensitive Tests dar, da hier bereits kleinere Beiträge neuer Physik sichtbar werden können, die bei häufig auftretenden Zerfällen im Rahmen der statistischen Unsicherheiten nicht beobachtbar sind.

Die Messungen von zeitabhängigen CP-Asymmetrien sind Schlüsselanalysen im zentralen Physikprogramm der LHCb-Kollaboration, da der Präzisionstest des Standardmodells hier auch über die Bestimmung von Phasendifferenzen erfolgt, für die es bereits häufig gute indirekte Vorhersagen im Rahmen des Standardmodells gibt. Aufgrund der $B\bar{B}$-Oszillationen muss bestimmt werden, ob das B-Meson zum Zeitpunkt seiner Erzeugung ein b oder $\bar{b}$-Quark enthalten hat, wozu weitere Informationen in dem Ereignis ausgewertet werden müssen (Tagging). Da die Qualität der Messungen zur CP-Asymmetrie direkt von der Genauigkeit des Taggings abhängen, wird permanent an der Verbesseung der Qualität des Taggings gearbeitet.

Die Arbeitsgruppe Spaan hat dazu bereits zahlreiche Vorarbeiten durchgeführt. Dies betrifft z.B. Beiträge zum Tagging bzw. die Kalibrierung des Taggings, Messung der Oszillationen im B_d-System [C5/639] und zu Präzisionsmessungen zu zeitabhängigen CP-Asymmetrien mit dem LHCb-Experiment [C5/640].

Die Messungen werden fortgeführt und auf das B_s-System ausgeweitet, wobei auch seltenere Zerfälle wie $B_s \to \phi\phi$ in Betracht gezogen werden, wofür auch bereits erhebliche Vorarbeiten geleistet wurden [C5/650].

Literatur

[609] R. Aaij und andere (inkl. Spaan, B.) „First evidence for the decay $B_s^0 \to \mu^+\mu^-$“. In: *Phys. Rev. Lett.* 110 (2013), S. 021801.

[610] *Apache HBase*. 2014. URL: http://hbase.apache.org/.

[611] D. Battré, S. Ewen, F. Hueske, O. Kao, V. Markl und D. Warneke. „Nephele/PACTs: A Programming Model and Execution Framework for Web-Scale Analytical Processing“. In: *Proc. of the 1st ACM Symposium on Cloud Computing (SoCC)*. Indianapolis, IN, USA, Juni 2010, S. 119–130.

[612] J. Becla und D. L. Wang. „Lessons learned from managing a petabyte“. In: *Proc. of the 22nd Biennial Conference on Innovative Data Systems (CIDR 2005)*. Asilomar, CA, USA, Jan. 2005.

[613] S. Borkar und A. A. Chien. „The Future of Microprocessors“. In: *Communications of the ACM* 54.5 (Mai 2011), S. 67–77.

[614] D. Broneske, S. Breß, M. Heimel und G. Saake. „Toward Hardware-Sensitive Database Operations“. In: *Proc. of the 17th Int'l Conference on Extending Database Technology (EDBT)*. Athens, Greece, März 2014, S. 229–234.

[615] R. Brun und F. Rademakers. „ROOT: An object oriented data analysis framework“. In: *Nucl.Instrum.Meth.* A389 (1997), S. 81–86.

[616] F. Chang u. a. „Bigtable: A Distributed Storage System for Structured Data“. In: *ACM Transactions Computer Systems (TOCS)* 26.2 (Juni 2008).

[617] G. DeCandia u. a. „Dynamo: Amazon's Highly Available Key-Value Store“. In: *Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP)*. Stevenson, WA, USA, Okt. 2007, S. 205–220.

[618] J. Dean und S. Ghemawat. „MapReduce: A Flexible Data Processing Tool“. In: *Communications of the ACM* 53.1 (Jan. 2010), S. 72–77.

[619] D. Emeliyanov und J. Howard. „GPU-based tracking algorithms for the ATLAS high-level trigger“. In: *J.Phys.Conf.Ser.* 396 (2012), S. 012018.

[620] S. Ghemawat, H. Gobioff und S.-T. Leung. „The Google File System“. In: *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP)*. Bolton Landing, NY, USA, Okt. 2003, S. 29–43.

[621] S. Gorbunov u. a. „ALICE HLT high speed tracking on GPU“. In: *IEEE Trans.Nucl.Sci.* 58 (2011), S. 1845–1851.

[622] N. K. Govindaraju, B. Lloyd, W. Wang, M. C. Lin und D. Manocha. „Fast Computation of Database Operations using Graphics Processors“. In: *Proc. of the 2004 ACM SIGMOD Int'l Conference on Management of Data*. Paris, France, Juni 2004, S. 215–226.

[623] N. K. Govindaraju, J. Gray, R. Kumar und D. Manocha. „GPUTeraSort: High Performance Graphics Co-Processor Sorting for Large Database Management“. In: *Proc. of the 2006 ACM SIGMOD Int'l Conference on Management of Data*. Chicago, IL, USA, Juni 2006, S. 325–336.

[624] T. Grust, M. Mayr, J. Rittinger und T. Schreiber. „FERRY: Database-Supported Program Execution“. In: *Proc. of the 2009 ACM SIGMOD Conference on Management of Data*. Providence, RI, USA, Juni 2009, S. 1063–1066.

[625] V. Halyo, A. Hunt, P. Jindal, P. LeGresley und P. Lujan. „GPU Enhancement of the Trigger to Extend Physics Reach at the LHC“. In: *JINST* 1310 (2013), P10005.

[626] B. He, K. Yang, R. Fang, M. Lu, N. K. Govindaraju, Q. Luo und P. V. Sander. „Relational Joins on Graphics Processors“. In: *Proc. of the 2008 ACM SIGMOD Int'l Conference on Management of Data*. Vancouver, BC, Canada, Juni 2008, S. 511–524.

[627] B. He, M. Lu, K. Yang, R. Fang, N. K. Govindaraju, Q. Luo und P. V. Sander. „Relational Query Coprocessing on Graphics Processors“. In: *ACM Transactions on Database Systems (TODS)* 34.4 (Dez. 2009).

[628] M. Heimel, M. Saecker, H. Pirk, S. Manegold und V. Markl. „Hardware-Oblivious Parallelism for In-Memory Column-Stores“. In: *Proceedings of the VLDB Endowment* 6.9 (2013), S. 709–720.

[629] M. Isard, M. Budiu, Y. Yu, A. Birrell und D. Fetterly. „Dryad: Distributed Data-Parallel Programs From Sequential Building Blocks“. In: *Proc. of the 2007 EuroSys Conference*. Lisbon, Portugal, März 2007, S. 59–72.

[630] D. R. Karger, E. Lehman, F. T. Leighton, R. Panigrahy, M. S. Levine und D. Lewin. „Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web“. In: *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing (STOC)*. El Paso, TX, USA, Mai 1997, S. 654–663.

[631] A. Lakshman und P. Malik. „Cassandra: A Decentralized Structured Storage System“. In: *Operating Systems Review* 44.2 (2010), S. 35–40.

[632] E. Meijer. „The World According to LINQ“. In: *Communications of the ACM* 54.10 (Okt. 2011), S. 45–51.

[633] R. Moussalli, M. Salloum, W. A. Najjar und V. J. Tsotras. „Massively Parallel XML Twig Filtering Using Dynamic Programming on FPGAs“. In: *Proc. of the 27th IEEE Int'l Conference on Data Engineering (ICDE)*. Hanover, Germany, Apr. 2011, S. 948–959.

[634] M. Najafi, M. Sadoghi und H.-A. Jacobsen. „Flexible Query Processor on FPGAs“. In: *Proceedings of the VLDB Endowment* 6.12 (2013), S. 1310–1313.

[635] M. C. Schatz und B. Langmead. „The DNA Data Deluge“. In: *IEEE Spectrum* (Juni 2013).

[636] K. Shvachko, H. Kuang, S. Radia und R. Chansler. „The Hadoop Distributed File System“. In: *Proc. of the 26th IEEE Symposium on Mass Storage Systems and Technologies (MSST)*. Lake Tahoe, NV, USA, Mai 2010, S. 1–10.

[637] S. A. Weil, S. A. Brandt, E. L. Miller, D. D. E. Long und C. Maltzahn. „Ceph: A Scalable, High-Performance Distributed File System“. In: *Proceeding of the 7th Symposium on Operating Systems Design and Implementation (OSDI)*. Seattle, WA, USA, Nov. 2006, S. 307–320.

[638] Y. Yu, M. Isard, D. Fetterly, M. Budiu, Ú. Erlingsson, P. K. Gunda und J. Currey. „DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language“. In: *Proc. of the 8th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. San Diego, CA, USA, Dez. 2008, S. 1–14.

3.3.2 Projektrelevante eigene Publikationen

Begutachtete Publikationen

[C5/639] R Aaij und andere (inkl. Spaan, B.) „Measurement of the B^0–$\bar{B}^0$ oscillation frequency Δm_d with the decays $B^0 \to D^-\pi^+$ and $B^0 \to J\ \psi K^{*0}$“. In: *Phys.Lett.* B719 (2013), S. 318–325.

[C5/640] R Aaij und andere (inkl. Spaan, B.) „Measurement of the time-dependent CP asymmetry in $B^0 \to J/\psi K^0_S$ decays“. In: *Phys.Lett.* B721 (2013), S. 24–31.

[C5/641] A. A. Alves Jr. und andere (inkl. Spaan, B.) „The LHCb detector at the LHC“. In: *Journal of Instrumentation (JINST)* 3 (2008), S08005.

[C5/642] C. Balkesen, J. Teubner, G. Alonso und M. T. Özsu. „Main-Memory Hash Joins on Multi-Core CPUs: Tuning to the Underlying Hardware“. In: *Proc. of the 29th IEEE Int'l Conference on Data Engineering (ICDE)*. Brisbane, Australia, Apr. 2013, S. 362–373.

[C5/643] R. Müller, J. Teubner und G. Alonso. „Streams on Wires—A Query Compiler for FPGAs“. In: *Proc. of the VLDB Endowment* 2.1 (Aug. 2009), S. 229–240.

[C5/644] P. Roy, J. Teubner und G. Alonso. „Efficient Frequent Item Counting in Multi-Core Hardware“. In: *Proc. of the 18th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*. Beijing, China, Aug. 2012, S. 1451–1459.

[C5/645] J. Teubner, L. Woods und C. Nie. „*XLynx* - An FPGA-Based XML Filter for Hybrid XQuery Processing“. In: *ACM Transactions on Database Systems (TODS)* 38.4 (Dez. 2013).

[C5/646] L. Woods, G. Alonso und J. Teubner. „Parallel Computation of Skyline Queries“. In: *21st IEEE Annual Int'l Symposium on Field-Programmable Custom Computing Machines (FCCM)*. Seattle, WA, USA, Apr. 2013, S. 1–8.

[C5/647] L. Woods, J. Teubner und G. Alonso. „Complex Event Detection at Wire Speed with FPGAs“. In: *Proc. of the VLDB Endowment (PVLDB)* 3.1 (Sep. 2010).

Andere Veröffentlichungen

[C5/648] C. Ilgner, M. Lieng, M. Domke, M. Nedos, J. Sauerbrey und andere (inkl. Spaan, B.) *The Beam Conditions Monitor of the LHCb Experiment*. 2010.

[C5/649] M. Nedos. „Entwicklung, Implementierung and Test eines FPGA-Designs für die Level-1-Frontend-Elektronik des Äusseren Spurkammersystems im LHCb-Detektor“. Diss. TU Dortmund, 2008.

[C5/650] S. Schleich. „First measurement of the inclusive ϕ meson production cross section in pp collisions at $\sqrt{s} = 7$ TeV and search for CP-violation in the $B_s \to \phi\phi$ decay at LHCb“. Diss. TU Dortmund, 2011.

3.4 Planung des Teilprojekts

Ziele

Das LHCb-Experiment mit Datenraten von mehreren Terabyte pro Sekunde ist ein eindrucksvolles Beispiel für die Datenintensität von Anwendungen aus den Naturwissenschaften. Die aktuell verwendeten Systeme sind mit diesen Datenmengen überfordert. Gleichzeitig wurden im Bereich der Datenbankforschung in den letzten Jahren wichtige Techniken für den Einsatz von heterogenen und verteilten Systemen für datenintensive Aufgaben entwickelt.

Zentrale Fragestellung des Teilprojekts ist es, wie solche Techniken auch auf andere Problemstellungen übertragen werden können. Diese Fragestellung wollen wir am Beispiel des LHCb-Experiments untersuchen.

Am Ende sollen dabei Methoden entstehen, die für ein möglichst breites Feld an datenintensiven Disziplinen anwendbar sind und dabei die Effizienz im Hinblick auf Laufzeit, Ressourcenverbrauch und Entwicklungsaufwand steigern. Aus Sicht der Physik wird eine deutliche Leistungssteigerung des konkreten Systems am CERN erwartet. Bessere Performanz ist hier nicht nur ein nice-to-have, sondern unbedingt notwendig, da diese datengetriebenen Analysen und damit die wissenschaftlichen Erkenntnisse notorisch durch vorhandene Rechenkapazität begrenzt sind.

Die Bewältigung des Datenvolumens wollen wir von zwei Seiten angehen. Zum Einen wollen wir dem laufenden Trend zu heterogenen Rechnersystemen Rechnung tragen und die Idee des Hardware/Software Co-Designs prototypisch für das LHCb-Experiment umsetzen (mit dem Einsatz von FPGAs und GPGPUs). Zum Anderen wollen wir die in den vergangenen Jahren entwickelten Ansätze für verteilte und hochskalierbare Speicher- und Auswertesysteme auf das Analyseproblem im LHCb-Experiment übertragen, untersuchen und weiterentwickeln. Für konkrete Analyseaufgaben aus der Physik wollen wir dabei die Vorteile von verteilten Dateisystemen und massiv parallelen Programmiermodellen demonstrieren. Damit kann in der dritten Förderphase des SFBs die Übertragung auf LHCb angegangen werden, um so die Eignung für das „Data Mining at the Petascale“ demonstrieren zu können. Die große Datenmenge und die damit verbundenen komplexen Analyseaufgaben haben auch das Interesse von Internetfirmen wie Yandex geweckt. So ist Yandex bereits eine Kooperation mit der LHCb-Kollaboration eingegangen[1].

Entwickelte Lösungen, Algorithmen und Systeme werden wir als Open Source-Software zur Verfügung stellen. Insbesondere streben wir natürlich auch einen regen Austausch mit den anderen Teilprojekten des SFBs an. Ziel wird es sein, die entwickelten Algorithmen und Methoden so aufzuarbeiten, dass sie sowohl in der Teilchenphysik als auch darüber hinaus genutzt werden können.

Arbeitsprogramm

Arbeitspaket 1. HLT-Performanzsteigerung: Im Bereich des HLT gilt es die Performanz des Systems zu optimieren, da hier ein Großteil aller Ereignisse verworfen wird. Daher müssen die Ereignisse von Interesse mit höchster Effizienz erkannt werden. Da Ereignisse, in denen B-Hadronen erzeugt werden, um ein Vielfaches häufiger auftreten als die Bandbreite der Trigger-Akzeptanz zulässt, werden komplexere Algorithmen benötigt, um innerhalb der Ereignisse mit B-Hadronen diejenigen zu identifizieren, in denen die interessanten B-Zerfälle auftreten. Diese Ressourcenbeschränkung im HLT führt daher zu auch zu Trigger-Ineffizienzen.

Profiling. Die Optimierung der Rechner- und Analysesysteme am CERN entlang der Ideen des Hardware/Software Co-Designs und des Einsatzes von Cloud-Technologien wurde bislang allenfalls rudimentär betrachtet. Im Rahmen eines Profiling wollen wir daher zunächst eine solide Datenbasis schaffen, um später die Entwicklungen des Teilprojekts zu validieren.

Dazu wollen wir zunächst zwei charakteristische B-Zerfallskanäle, die selektiert und untersucht werden sollen, herausgreifen, um daran das Verhalten von Hard- und Software experimentell zu untersuchen. Diese Untersuchungen sollen in Arbeitspaket 4 wieder aufgegriffen werden, um den Fortschritt durch das Teilprojekt zu belegen. Die beiden Fallbeispiele sind so gewählt, dass sie einmal eine eher einfache Aufgabe auf einer großen Datenmenge bedingen und einmal eine deutlich komplexere Analyse der Ereignisse erfordern. Am Lehrstuhl von Jens Teubner ist bereits eine Testinstallation vorhanden, um das Verhalten im LHCb-Experiment lokal und unabhängig zu analysieren. Anhand dieser Testinstallation soll der Ressourcenbedarf einzelner Systemkomponenten sowie das Datenzugriffsverhalten analysiert werden. Daraus wollen wir Strategien ableiten, um Aufgaben auf FPGAs oder GPGPUs auszulagern (siehe unten) sowie Daten verteilt zu speichern und zu analysieren (Arbeitspakete 2 und 3).

FPGA-basierte Datenvorverarbeitung. Die Auslese der Daten des LHCb-Experiments erfolgt über das Senden der Daten an eine Off-Detektor-Elektronik, die auf FPGA-Basis die Daten weiterverarbeitet. Durch eine Datenvorverarbeitung mit Hilfe von FPGAs im Eingang können Ressourcen auf der HLT-Farm dann für andere Aufgaben zur Verfügung stehen. Die Datenakquisition des LHCb-Experiments sieht bereits eine minimale Datenvorverarbeitung auf FPGA-Basis vor. Diese ist momentan aber beschränkt auf die Synchronisation verschiedener Detektorkomponenten, Tests auf Fehler oder die Zusammenführung der Daten verschiedener Detektorkomponenten. Eine weitergehende Datenvorverarbeitung findet nicht statt. Im Rahmen des Upgrades des LHCb-Experiments wird auch die gesamte Ausleseelektronik neu entwickelt. Dabei zeichnet sich jetzt schon ab, dass die FPGAs nicht vollständig mit den bestehenden Aufgaben ausgelastet sein werden. Im Bereich der Spurdetektoren werden beispielsweise wohl ca. 60% der Kapazität der FPGAs (ALTERA STratix IV) belegt sein, sodass hier Ressourcen, wenn auch stark limitiert, für eine weitere Bearbeitung der

[1] „Yandex launches Search Tool for LHC events at CERN“, Pressemitteilung Yandex, 10.4.2012, http://company.yandex.com/press_center/press_releases/2012/2012-04-10.xml

Daten vorliegen. Dabei liegt der Fokus auf der Clusterfindung bei den Spurdetektoren, da sich hier durch Clusterfindung eine deutliche Volumenreduzierung erzielen ließe, wodurch Ressourcen auf der Trigger-Farm frei werden, die dann für andere Aufgaben zur Verfügung stehen können. Bereits eine relativ simple Operation wie die Clusterfindung erfordert aufgrund der großen Datenmenge sehr große CPU-Ressourcen auf der Triggerfarm. Daher bieten sich die FPGAs zur Datenvorverarbeitung an. Die Bildung von Clustern und deren Bedeutung für die Spurfindung sollen nun am Beispiel des SciFi-Trackers illustriert werden. Im SciFi-Tracker werden Lagen szintillierender Fasern in einer Ebene senkrecht zur Strahlrichtung angebracht, wobei innerhalb einer Ebene alle Fasern dieselbe Orientierung aufweisen. Vom Wechselwirkungspunkt aus gesehen, sind mehrere dieser Ebenen hintereinander angeordnet, die über unterschiedliche Orientierungen verfügen, sodass aus der Kombination der Messwerte 3D-Raumpunkte für den Teilchendurchgang gebildet werden. Eine Herausforderung ist es daher, Cluster modulübergreifend in einer Ebene zu finden und diese Cluster denen in den anderen Ebenen zuzuordnen. Als Vorteil wird gesehen, dass die Ebenen sich in einem geringen räumlichen Abstand voneinander befinden, wodurch gewisse Richtungsinformationen gewonnen werden können. Verschiedene FPGA-basierte Clusteringalgorithmen sind bereits entwickelt worden, die aber für dieses Problem nur bedingt relevant sind, da die Daten aus verschiedenen Modulen stammen und so nicht direkt logisch verknüpft werden können. Daher müssen Algorithmen entwickelt werden, die diese Heterogenität der Daten berücksichtigen.

Im Rahmen des Arbeitspaketes stellt damit die genaue Bestimmung der Entscheidungskriterien eine besondere Herausforderung dar. Dies soll auf Basis simulierter Daten und nach Fertigstellung und Einbau des SciFi-Trackers von Ereignissen, die nach positivem Triggerentscheid bereits vollständig rekonstruiert wurden, geschehen, da hier die Spurinformationen bekannt sind. Auch soll die Zuordnung von Detektormodulen zu FPGAs optimiert werden, um eine maximale Zahl an 3D-Raumpunkten finden zu können.

Die Daten der anderen Spurdetektoren lassen sich auf ähnliche Weise behandeln, sodass für jeden Spurdetektor Cluster bereits 3D-Spurelemente und gegebenenfalls grobe Richtungsinformationen, die für die Spurfindung von Bedeutung sind, vorliegen. Da nicht alle 3D-Informationen aufgrund der limitierten Anzahl von Detektorkanälen, die auf ein FPGA geführt werden können, vorliegen, müssen die restlichen Punkte durch das nachfolgende Triggersystem ermittelt werden.

Aufgrund der begrenzten Kapazität der FPGAs müssen die Algorithmen besonders effizient sein.

GPGPU-basierte Datenverarbeitung. Die für die HLT-Farm vorgesehenen Rechner verfügen über die Möglichkeit, zusätzliche Grafikkarten zu installieren. Ein Ziel des Teilprojekts ist es, mit Hilfe von GPGPUs Ressourcen auf der Farm in einem so großen Umfang verfügbar zu machen, dass Spurinformationen samt Impulsbestimmung sowie gegebenenfalls Informationen der RICH-Subdetektoren, die für die Identifizierung geladener Teilchen unverzichtbar sind, nun auch auf der Triggerfarm Verwendung finden können.

Im Rahmen des Arbeitspakets sollen daher zunächst die Trigger-Algorithmen identifiziert werden, die sich besonders gut für die Verarbeitung durch GPGPUs eignen. In einem systematischen Vergleich der Performanz von CPU- und GPGPU-Verarbeitung soll dann eine Strategie entwickelt werden, wie das Zusammenspiel zwischen CPUs und GPGPUs optimiert werden kann, wobei auch der Overhead bei einem Übergang von CPU auf GPGPU und zurück berücksichtigt werden muss.

Die Trigger-Entscheidung sollte im Idealfall nicht nur auf den gefundenen Spuren basieren, da diese unabhängig von der Art des Teilchens (z. B. Pion, Kaon, Proton oder Elektron und Myon) sind. Eine Trigger-Entscheidung, die auf der Kenntnis der Zerfallsdynamik von B-Hadronen beruht, wird zum großen Teil auf einem Vergleich der rekonstruierten invarianten Masse mit dem Weltmittelwert für die Masse des B-Hadrons basieren. Dazu ist aber eine vollständige Spurrekonstruktion verbunden mit einer Impulsbestimmung notwendig. Diese Impulsbestimmung lässt sich nur durch einen Spurfit, z.B. auf Basis eines Kalman-Filters, mit der für eine Analyse erforderlichen Genauigkeit bestimmen. Im Rahmen des Triggers sind größere Ungenauigkeiten zulässig - wenn damit nicht unnötige systematische Unsicherheiten ins Spiel kommen. Dessen ungeachtet muss die Teilchenidentität bekannt sein, da die invariante Masse der unterschiedlichen Teilchen (z. B. Pion: $m_\pi \approx 140$ MeV/c^2, Kaon: $m_K \approx 500$ MeV/c^2, Proton: $m_p \approx 938$ MeV/c^2) großen Einfluss auf die rekonstruierte Masse des B-Hadrons hat. Da Elektronen mit Hilfe des Kalorimeters und Myonen mittels des Myon-Systems gut identifiziert werden können, müssen Pionen, Kaonen und Protonen mit einem dedizierten Teilchenidentifizierungssystem nachgewiesen werden. Im Falle von LHCb sind dies die Ring Imaging Cherenkov-Detektoren (RICH). Im Rahmen des Arbeitspakets sollen neben dem Tracking insbesondere die Algorithmen zur Ring-Rekonstruktion und Spurzuordnung auf eine Eignung für die Verarbeitung mit GPGPUs hin untersucht werden. Auch hier werden für CPUs optimierte RICH-Algorithmen entwickelt und mit den GPGPU-Resultaten verglichen, um eine maximale Steigerung der Trigger-Performanz zu erhalten.

Arbeitspaket 2. Datenspeicherung und -organisation: Trotz Vorselektion durch das Triggersystem verbleiben Messdaten im Umfang von etwa 20 PB pro Jahr, die dauerhaft gespeichert und für spätere physikalische

Analysen verfügbar bleiben müssen. Wie dargelegt, wird nur ein Teil der Daten, in denen bestimmte Teilchenzerfälle markiert worden sind, auf Festplatte vorgehalten. Die gesamte Datenmenge wird auf Bändern gespeichert. Ein gezielter Zugriff, zum Beispiel auf Grundlage konkreter, aber noch nicht bereits markierter Zerfallskanäle, ist unter solchen Bedingungen jedoch langwierig und mit hohem Aufwand verbunden. Im Rahmen des Arbeitspakets soll dieser Aufwand reduziert werden; visionäres Ziel wäre ein interaktives Arbeiten der Wissenschaftler mit den gespeicherten Datenbeständen.

Grundlage des Arbeitspakets sollen zunächst verteilte Speichertechniken aus dem Cloud Computing sein. Insbesondere die Speicher-Backends von MapReduce (Google File System) und Hadoop (Hadoop Distributed File System) bringen grundlegende Eigenschaften mit, wie sie für die Datenmengen am LHC geeignet erscheinen (effizienter Lesezugriff auf Teildaten beliebiger Größe; sequenzielles Schreiben mit hoher Bandbreite durch möglicherweise parallel laufende Schreibtransaktionen). Erstes Teilpaket soll die Bewertung dieser bestehenden Ansätze für die Datenmengen am LHC sein. Dabei wollen wir ein besonderes Augenmerk darauf richten, ob die resultierenden Datenverteilungen und Zugriffsmuster adäquat sind für die Analyseprogramme des LHCb-Experiments bzw. ob sie die vorhandene Hardware effektiv ausnutzen können.

Verteilungsstrategie. In klassischen verteilten Datenbanklösungen werden zur Datenverteilung meist Strategien angewandt, die Inhalt und Werte von Datenfeldern berücksichtigen (z. B. Range Partitioning). Dadurch kann bei späteren Anfragen Arbeit eingespart werden, weil diese u. U. nur noch lokal auf relevanten Knoten ausgeführt werden müssen. Umgekehrt ist es oft schwierig, solche Strategien auf einer großen Anzahl an Knoten zu skalieren, weil *(a)* die Daten nicht genügend fein-granular charakterisiert werden können, als dass jedem vorhandenen Knoten überhaupt ein Fragment zugeordnet werden könnte. Außerdem ist *(b)* die Strategie sehr anfällig für Über-/Unterlastsituationen. Eine falsche Auswahl von Fragmentgrenzen kann zu großen Ungleichgewichten in der resultierenden Datenverteilung führen.

Die erwähnten Cloud-Systeme setzen im Gegenzug ausschließlich auf Hash Partitioning oder andere Zufallsverfahren. Zum Beispiel werden künstliche Schlüssel durch eine Hashfunktion auf einen Knoten im Netzwerk abgebildet. Das führt nahezu trivial zu einer gleichmäßigen Auslastung der Knoten im System. In den genannten Cloud-Systemen ist solches Vorgehen oft schon deshalb notwendig, weil zu wenig über die Charakteristik und Verteilung der Daten bekannt ist, als dass solche Informationen für die Lokalisierung eingesetzt werden könnten. Umgekehrt gehen aber auch Beziehungen zwischen Dateninhalten und der Lokalisierung im System verloren, die später gut geeignet sein könnten, um etwa bei der Analyse der Daten die Übertragung großer Datenmengen über das Netzwerk möglichst einzuschränken.

Im Teilprojekt soll daher untersucht werden, wie sich vorhandenes Wissen über die gespeicherten Daten (in diesem Fall das Wissen der Physiker) ausnutzen lässt, um die Platzierung der Daten im Netzwerk zu optimieren. Konkrete Ansätze hierzu könnten Indexstrukturen sein, aus deren Struktur man eine Knotenzuordnung für die Datenelemente ableiten könnte. Außerdem könnten Hashfunktionen attraktiv sein, deren Abbildungsfunktion nicht völlig zufällig ist, sondern Werteähnlichkeiten in räumliche Nähe übersetzt (etwa *order-preserving hash functions*).

Indexstrukturen. Um die anfallenden Datenmengen überhaupt handhaben zu können, wird der anfallende Datenstrom am CERN zunächst voranalysiert. Dabei werden Replikate der Originalmessdaten erzeugt, von denen jedoch Teile stark aggregiert, komprimiert, oder gar entfernt werden. Diese Replikate werden nach festen und groben Kriterien kategorisiert. Replikate wichtiger Kategorien werden auf Festplatten mit kurzer Zugriffszeit gespeichert. Alle anderen Daten, einschließlich der Originalmessdaten, werden aus Platzgründen nur auf Band gespeichert. Ein Zugriff darauf wäre anschließend mit sehr hohem Aufwand verbunden. In langen Zeitabständen (etwa ein- bis zweimal pro Jahr) wird der Voranalysevorgang und die Kategorisierung der auf Band gespeicherten Daten wiederholt. Erst im Rahmen einer Wiederholung können veränderte Analyse- oder Kategorisierungskriterien für einen Datensatz berücksichtigt werden.

Das bisherige Vorgehen führt damit zu einem Paradoxon. Um der großen Datenmenge überhaupt Herr zu werden, werden Replikate mit verringerter Datenmenge erstellt. Dadurch ergibt sich aber eine *Vergrößerung* des Gesamtdatenvolumens. Der Effekt wird noch verstärkt dadurch, dass einzelne Ereignisse auch mehreren Kategorien angehören können, sodass ihr Replikat mehrfach abgespeichert wird. Die eigentlichen Messdaten können am Ende nur noch auf Band-Medien mit schlechter Verfügbarkeit gespeichert werden.

Im Teilprojekt wollen wir diesem Dilemma mit Hilfe von Indexstrukturen aus der Datenbankwelt begegnen. Anders als bislang sollen grundsätzlich die Originaldaten auf verfügbaren Medien vorgehalten werden. Indexstrukturen sollen dann ermöglichen, aus den vorgehaltenen Originaldaten effizient die relevanten Teildaten herauszusuchen, ohne dass diese Teil-Daten vorher explizit materialisiert werden müssen. Dadurch können nicht nur Redundanzen vermieden werden und mehr Informationen verfügbar gehalten werden. Vielmehr versprechen wir uns auch, dass durch die entwickelten Indexstrukturen wesentlich selektivere Zugriffe möglich werden. Das entlastet die vorhandenen Speicherressourcen und spart gleichzeitig wichtige Rechenzeit im eigentlichen Analyseprogramm des Benutzers.

Existierende Indexstrukturen werden den Anforderungen aus der Physik allerdings nicht unmittelbar gerecht werden, weil sich Zugriffskriterien hier nicht – wie typischerweise für Datenbankindizes angenommen – einfach durch Bereichs-Prädikate spezifizieren lassen. Im aktuellen System spezifizieren die Wissenschaftler nur sehr grob (in Form einer Kategorie-Bezeichnung) die für sie aktuell interessanten Datensätze. Weitere Kriterien für die Relevanz eines Datensatzes werden erst in der Analysesoftware selbst bewertet; „false hits“, also erst an dieser Stelle erkannte Nicht-Relevanzen, belasten die ohnehin knappen Analyseressourcen. Mit geeigneten Indexstrukturen könnten Relevanz-Kriterien hingegen bereits im Speichersystem berücksichtigt werden. Es ist allerdings keineswegs klar, wie die komplexen und aktuell in Form von Softwareprogrammen existierenden Kriterien effizient ausgedrückt und so an das Speichersystem übergeben werden können. Umgekehrt bestimmt die Art der Kriterien maßgeblich die Effektivität einer gewählten Indexstruktur. Charakterisierungen für Zugriffskriterien und passende Indexstrukturen sollen daher Hand in Hand zwischen den beteiligten Fächern entwickelt werden.

Ein erster Startpunkt für diese Arbeit aus der Sicht der Datenbankwelt sollen mehrdimensionale Indexstrukturen sein. Die reinen Messdaten werden als Dimensionen solcher Indexstrukturen allerdings nur wenig hilfreich sein. Vielmehr charakterisieren abgeleitete Parameter wesentlich besser, welche Ereignisse für eine gegebene Problemstellung relevant sein könnten. Hier wollen wir versuchen, solche abgeleiteten Parameter sukzessive aus den bisherigen Modellformulierungen oder Softwareprogrammen zu extrahieren.

Verteilungsstrategie ↔ Indexstrukturen. Die Aspekte der Verteilungsstrategie und der Indexstrukturen greifen eng ineinander. Wir gehen davon aus, dass etwa die gewählten Indexstrukturen und Ereignisparameter wichtige Anhaltspunkte für eine geeignete Speicherung sein können. Umgekehrt wird die Verteilung der Daten auch relevant sein, welche Indizierungstechniken besonders effektiv sind.

Arbeitspaket 3. Programmiermodelle zur Datenanalyse: Wie bereits mehrfach erwähnt, setzt die anfallende Datenmenge und die mitunter große Rechenintensität von Analyseanwendungen eine enorme Rechenleistung für die anzustellenden wissenschaftlichen Analysen voraus. Viele Institutionen weltweit stellen daher Rechnerressourcen für das Worldwide LHC Computing Grid (WLCG) bereit und machen es zum wohl größten Rechnerverbund überhaupt. Die Koordinierung und effiziente Nutzung wird damit zu einer zentralen Herausforderung.

Das WLCG ist nach dem Grid-Modell realisiert. Es stellt eine Plattform bereit, auf der die Wissenschaftler Analyseprogramme in eine Warteschlange einstellen können. Dabei wird jedem Programm eine Beschreibungsdatei beigefügt, in der der voraussichtliche CPU- und Speicherverbrauch des Programms deklariert wird sowie die Namen der Eingabedateien. Das Scheduling-System des WLCG versucht dann, jedes Programm auf einer Maschine laufen zu lassen, die den Anforderungen genügt und die außerdem möglichst nahe an den Eingabedateien liegt. Sollte das Programm in einen Fehler laufen, wird möglicherweise erneut versucht, das Programm zu starten und ansonsten die Fehlermeldung an den aufrufenden Benutzer zurückgeliefert. Wichtige Problemstellungen, die bei parallelen Analyseanwendungen anfallen, werden dabei nicht vom WLCG gelöst, sondern müssen vom aufrufenden Benutzer explizit berücksichtigt werden.[2] Dazu zählen insbesondere *(a)* das Aufbrechen von Rechenaufgaben in parallele Teilaufgaben, *(b)* die Synchronisation und Kommunikation zwischen solchen parallelen Aufgaben und *(c)* die Behandlung von auftretenden Fehlern. Aufgrund der wachsenden Komplexität der angewandten Analysen und der Größe der zugrundeliegenden Daten sind diese Problemstellungen ohne einen systematischen Lösungsansatz praktisch nicht mehr handhabbar, oder sie führen aufgrund wachsender Fehlerwahrscheinlichkeiten in großen Netzen zu hohen Effizienzverlusten.

Cloud Computing für Anwendungen der Teilchenphysik. In der Informatik wurden solche systematischen Lösungsansätze in den vergangenen Jahren entwickelt und etabliert. Bekannteste Vertreter sind MapReduce oder Alternativen verschiedener Anbieter, wie Hadoop, Dryad oder Stratosphere. Im Kern bestehen diese Systeme aus einem Programmiermodell und einer Laufzeitumgebung. Ein Einhalten des Programmiermodells garantiert, dass das formulierte Programm auf definierte Art und Weise parallelisiert werden kann und Fehlerfälle automatisch aufgelöst und korrigiert werden können. Die Laufzeitumgebung nutzt diese Eigenschaften, um Rechenaufgaben vollautomatisch im vorhandenen Maschinenpool zu lösen. Dabei wird außerdem versucht, Gesichtspunkte wie Ressourcenverbrauch oder Datenlokalität bei der Ausführung automatisch zu berücksichtigen.

Im Arbeitspaket wollen wir untersuchen, wie die Konzepte dieser Systeme eingesetzt werden können, um wissenschaftliche Analysen aus der Teilchenphysik auf massiven Datenmengen durchzuführen. Konkret wollen wir auf das Open-Source-System Hadoop aufsetzen und damit beispielhaft typische Analysen aus dem LHCb-Experiment realisieren. Die wichtigste Herausforderung auf der Systemseite wird dabei sein, die in Arbeitspaket 2 entwickelten Speicherkonzepte adäquat zu unterstützen. In Anbetracht der großen Daten-

[2] Die Nutzer greifen hierfür häufig auf eine Zahl von Skripten, Werkzeugen und Bibliotheken zurück, die im Laufe der Zeit für das WLCG entstanden sind.

mengen des Experiments werden dabei Kommunikationsressourcen besondere Bedeutung haben. Für die Anwendungsseite, physikalische Analysen in der Teilchenphysik, stellen Programmiermodelle wie MapReduce Neuland dar. Hier gilt es zu untersuchen, ob und wie sich die Problemstellungen der Physik in solchen Programmiermodellen darstellen lassen.
Interaktion Analyse ↔ Speichersystem. Nicht nur bei der Aufnahme von Messpunkten kann oft bereits in einer frühen Phase eines Analyseschritts entschieden werden, ob ein Ereignis für die konkrete Fragestellung nicht relevant ist und daher verworfen werden kann. Dies legt nahe, Analyseprogramme so zu formulieren, dass Datenwerte erst nach und nach vom Speichersystem angefragt werden. Bei Ereignissen, die bereits früh verworfen werden können, lassen sich so teure und möglicherweise bandbreitenintensive Zugriffsoperationen vermeiden. In der Praxis können Analysen z. B. hierarchisch ablaufen: zunächst werden nur grobe Informationen zum Ereignis geladen; immer weitere Details werden nachgeladen, bis schließlich der volle Messdatensatz herangezogen wird.

Für viele realistische Analysen ergibt sich durch ein solches Vorgehen ein sehr hohes Einsparpotenzial. Es ist jedoch keineswegs klar wie *(a)* die Daten so abgelegt werden können, dass das stückweise Nachladen überhaupt eine Ressourcenersparnis ergibt, und wie *(b)* die Interaktion zwischen Analysesoftware und Speichersystem so konzipiert werden kann, dass das Zugriffsmuster auf verwertbare Art an das Speichersystem weitergegeben werden kann; ansonsten könnte der Zugriff auf Detailinformationen jeweils ein erneutes Laden der Grobstruktur auslösen, was zu offensichtlich unnötigen Kosten führen würde.

Wir wollen daher untersuchen, wie sich Wissen aus der Analyseaufgabe im gewählten Programmiermodell dergestalt formulieren lässt, dass im Speichersystem auch ein Vorteil gezogen werden kann, und wie dieser Vorteil sich dann auch tatsächlich realisieren lässt. Dieser Teil des Arbeitspakets steht damit in engem Zusammenhang mit Arbeitspaket 2. Beide Arbeitspakete sollen daher im späteren Verlauf des Teilprojekts in Kombination bearbeitet werden.

Arbeitspaket 4. Evaluation und Anwendung: Beide Arbeitspakete, *Datenspeicherung und -organisation* und *Programmiermodelle zur Datenanalyse*, basieren auf Anwendungsszenarien aus der Physik. Hier stehen für das Teilprojekt mehrere Analysefälle zur Verfügung. Dabei wählen wir die Analyseszenarien so aus, dass sie ein möglichst breites Spektrum aus dem Anwendungsfeld abdecken, etwa solche, die eine sehr große Datenmenge auch tatsächlich lesen müssen und solche, bei denen sich sehr selektive Auswahlkriterien beim Datenzugriff formulieren lassen. Im Hinblick auf das Problem der Interaktion zwischen dem Analyseprogramm und der Speicherebene wollen wir auch gezielt Anwendungsfälle betrachten, die diese Eigenschaft aufweisen und ausnutzen können.

Damit wird das erste große Ziel sein, das neue Datenspeicher- und Programmiermodell auf einem Testcluster zu realisieren. In einem zweiten Schritt soll dieses Modell dann einem rigorosen Test unterzogen werden, um eventuelle Schwachstellen zu identifizieren. Mit Hilfe der geplanten physikalischen Analysen soll das Modell dann in Hinblick auf den Ressourcenverbrauch untersucht und optimiert werden. Als Daten werden zunächst vorselektierte Daten eingepflegt, um mittels physikalischer Analysen mit unterschiedlich hohen Anforderungen bzw. unterschiedlich hohem Komplexitätsgrad auf Basis des bestehenden und des neuen Speicher- und Zugriffsmodells einen direkten Vergleich mit den bisherigen Verfahren zu ermöglichen. Erweist sich das System als so performant wie erwartet, sind dann die Grundlagen für die Hinzunahme zusätzlicher Informationen und Daten geschaffen, die physikalische Analysen mit einen neuen Qualität ermöglichen sollen.

Für eine Entwicklungs- und Testumgebung des neuen Speicher- und Zugriffsmodells steht in Dortmund ein nach §91 beschaffter Cluster (Antragsteller Wolfgang Rhode, Bernhard Spaan (federführend) und andere) mit über 148 Workernodes und 1184 Cores und 230 TB zentralem Speicherplatz sowie lokalem Plattenplatz, d.h. pro Computenode, von 300 GB zur Verfügung. Damit lassen sich auch komplexe Analysen auf einem Teil der Daten wie z.B. vorselektierten Daten durchführen.

Die geplanten Analysen sollen in direkter Verbindung mit den Analyseschwerpunkten der Arbeitsgruppe Spaan stehen, sodass bereits erhebliche Erfahrungen auf dem Gebiet vorhanden sind und so ein zügiger Analysefortschritt zu erwarten ist. Folgende Analysen sind geplant:

- Untersuchungen zu einem seltenen B-Zerfallskanal (z.B. $B_s \to \rho^0\phi$ oder $B_d \to \rho^0\mu^+\mu^-$). Diese Analysen mit hoher Sensitivität auf neue Physik erfordern in der Regel nur Zugriff auf Zerfallszeit- und kinematische Informationen des markierten B-Hadrons. Vorstudien zum Zerfall $B_s \to \rho^0\phi$ mit den nachfolgenden Zerfällen $\phi \to K^+K^-$ und $\rho^0 \to \pi^+\pi^-$ werden momentan im Rahmen einer Masterarbeit (T. Schmelzer) durchgeführt. Dieser Zerfall ist von Interesse, da er ein ideales Testobjekt für Z-Pinguine darstellt. Im Rahmen des Standardmodells wird ein Verzweigungsverhältnis von etwa $4.4 \cdot 10^{-7}$ erwartet, wohingegen Prozesse neuer Physik das Verzweigungsverhältnis deutlich (Faktor 5)

erhöhen können.

- Präzisionsmessung zu zeitabhängigen CP-Asymmetrien z.B. in $B_s \to \phi\phi$ oder $B_d \to \phi K_S^0$. Diese Analysen erfordern den Zugriff auf komplexere Informationen im Ereignis wie beispielsweise das Tagging.

  Die Messung zeitabhängiger CP-Asymmetrien erfolgt sowohl im B_d als auch im $\bar{B}_s$-System. Hier muss unter anderem bestimmt werden, ob das B-Meson zum Zeitpunkt seiner Erzeugung ein b oder $\bar{b}$-Quark enthalten hat (Tagging). Dabei wird ausgenutzt, dass im Ereignis b-Quarks immer paarweise ($b\bar{b}$) erzeugt werden, sodass Zerfallsprodukte des anderen B-Hadrons im Ereignis dazu analysiert werden müssen. Da die Qualität der Resultate direkt von der Genauigkeit des Taggings abhängt, werden auch in der Arbeitsgruppe permanent große Anstrengungen unternommen, die Qualität des Taggings zu verbessern. Aufbauend auf diesen Erfahrung und Messungen zu $B\bar{B}$-Oszillationen sowie zu Präzisionsmessungen zu CP-Asymmetrien [C5/640] sollen nun Messungen zu CP-Asymmetrien in selteneren Zerfallskanälen durchgeführt werden. So kommt nun beispielsweise der Zerfall $B_s \to \phi\phi$ in Betracht, in dem ebenfalls eine hohe Sensitivität auf neue Physik vorliegt. In der Arbeitsgruppe wurden zu diesem Kanal bereits erhebliche Vorarbeiten geleistet [C5/650]. Der Kanal $B_d \to \phi K_S^0$ ist das entsprechende Pendant im B_d-Sektor und wird über einen ähnlichen Pinguin-Prozess erzeugt, bei dem Teilchen neuer Physik zusätzliche Beiträge liefern können.

Sollten die Zugriffszeiten im neuen Modell für die Analyse soweit minimiert worden sein, sollen Verfahren implementiert werden, die eine Speicherung von sogenannten N-Tupeln vermeiden. Dazu müssen die aus dem Ereignis ermittelten Zusatzinformationen wie die Tagging-Informationen in einem privaten Teil der Datenbank abgespeichert werden. Bei weiteren zukünftigen Analyseschritten stehen diese Daten dann für jedes relevante Ereignis zur Verfügung. Um eine Weiterverarbeitung mit z.B. dem Programmpaket ROOT [615] zu ermöglichen, soll ein entsprechendes Interface dazu entwickelt werden.

Eine Vision für die Zukunft wäre, ein erfolgreiches Speicher- und Zugriffsmodell der LHCb-Kollaboration vorzustellen, um daraufhin eine Installation an einem Tier-1-Zentrum oder im Erfolgsfall auch an allen Tier-1-Zentren zu ermöglichen.

Die Analysen im Arbeitspaket werden für die Selektion bewährte multivariate Methoden wie Boosted Decision Trees einsetzen, die im Rahmen des TMVA-Frameworks zur Verfügung gestellt werden. Durch eine Zusammenarbeit im SFB sehen wir auch die Chance, neuartige multivariate Methoden bzw. Techniken des maschinellen Lernens evaluieren und einsetzen zu können, um die Selektionsperformanz weiter zu erhöhen.

Zeitplan

	2015				**2016**				**2017**				**2018**			
Arbeitspaket / Quartal	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. HLT-Performanzsteigerung:																
2. Datenspeicherung und -organisation:																
3. Programmiermodelle zur Datenanalyse:																
4. Evaluation und Anwendung:																

Wir betrachten die meisten Arbeitspakete als langfristige Entwicklungsaufgaben, sodass sie mit variierender Intensität über die gesamte Förderperiode ablaufen werden. Arbeitspaket 1 besteht aus drei etwa gleichwertigen Bereichen (Profiling, FPGA, GPGPU), die auch nach Fertigstellung der eigentlichen Methoden eine Zusammenarbeit mit der LHCb-Kollaboration erfordern. Arbeitspaket 2 gliedert sich in die Bereiche Verteilungsstrategien, Indexstrukturen sowie Zusammenspiel der beiden. Die drei Bereiche bauen jeweils aufeinander auf. Arbeitspaket 3 erfordert erste Vorarbeiten aus Arbeitspaket 2. Da es in Zusammenarbeit mit Arbeitspaket 3 steht, muss es auch bis zum Ende der Förderperiode laufen. Arbeitspaket 4 erfordert erhebliche Vorarbeiten, um die Evaluierung zeitnah abschließen zu können. So müssen die Physikanalysen zunächst mit den bestehenden Systemen entwickelt, optimiert und getestet werden.

3.5 Stellung innerhalb des Sonderforschungsbereichs

Im Mittelpunkt dieses Teilprojekts stehen Informationen, die mit Experimenten der Teilchenphysik an Speicherringen mit einer hohen Datennahmerate erzeugt werden und zu einem sehr großen Speicherbedarf führen. Die Informationen liegen in Form von diskreten Ereignissen als Folge von Kollisionen von Teilchenpaketen vor. Betrachtet wird im Detail das LHCb-Experiment am Large Hadron Collider des CERN, das aber als exemplarisch für andere Experimente am LHC oder zukünftige Beschleuniger mit höchster Kollisionsrate zu sehen ist.

Aufgrund der Stellung als Lieferant großer Datenmengen mit hohem Komplexitätsgrad bieten sich vielfältige Kooperationsmöglichkeiten innerhalb des Sonderforschungsbereichs an, wobei der Arbeitskreis *Datenanalyse* von besonderem Interesse ist. Weiterhin wird der Kontakt zum Arbeitskreis *Ressourcenbeschränkung* als sehr wertvoll für die Arbeit in diesem Teilprojekt erachtet.

Konkret versprechen wir uns viel von einer Zusammenarbeit mit den Teilprojekten A1, A2 und C3. Durch das Teilprojekt A1 werden die Grundlagen für eine Optimierung der Ereignisselektion mit Hilfe moderner Data Mining-Techniken bzw. Techniken des maschinellen Lernens zur Verfügung gestellt. Im Rahmen des Teilprojekts C3 ist eine darauf basierende Anwendung bereits entwickelt worden. Hier profitieren wir von den Erfahrungen, die bei der Bereitstellung von Daten für die Verarbeitung mit RapidMiner gewonnen wurden. Obwohl in C3 ebenfalls große Datenmengen bereit gestellt werden, ist die Überlappung zwischen C3 und C5 inhaltlich relativ gering, sodass sich die Ressourcenbeschränkungen deutlich unterschiedlich bemerkbar machen. Darüber hinaus sollen auch im Teilprojekt C3 GPGPUs Verwendung finden, so z.B. bei Monte-Carlo-Rechnungen. Dazu wird es einen intensiven Erfahrungsaustausch geben. Interessante Kooperationsmöglichkeiten ergeben sich auch mit den anderen Teilprojekten, in denen GPGPUs (A1, A3, B2, C1, C3) und FPGAs (A1, A2, C3) eine Rolle spielen.

Ein Schwerpunkt von A2 sind Datenstromalgorithmen. Viele der physikalischen Analysen (u.a. auch das Triggering) lassen sich auf eine Datenstromformulierung zurückführen, sodass wir davon ausgehen, dass beide Teilprojekte gegenseitig voneinander profitieren können.

3.6 Abgrenzung gegenüber anderen geförderten Projekten

LHCb: Upgrade, Detektorbetrieb und Datenanalyse
(Spaan,Förderkennzeichen BMBF: 05H12PE2) (Förderzeitraum: 2012 - 2015)

Dieses vom BMBF geförderte Verbundprojekt stellt die Grundlage für die Teilnahme am Experimentierbetrieb mit dem LHCb-Experiment dar. An LHCb sind von deutscher Seite beteiligt: Universität Heidelberg, Universität Rostock, TU Dortmund, und das MPI für Kernphysik Heidelberg. Seit Dezember 2013 ist auch die RWTH Aachen an LHCb beteiligt und wird zusammen mit den anderen deutschen Universitäten einen Förderantrag für die Periode 2015 bis 2018 stellen. Das BMBF hat gegenüber der Kollaboration und CERN entsprechende Memoranda of Understanding unterzeichnet, wodurch die deutsche Teilnahme am Projekt langfristig gesichert ist. So wird die LHCb-Beteiligung der Arbeitsgruppe von Bernhard Spaan seit 2004 gefördert. Durch die BMBF-Förderung wird sicher gestellt, dass die beteiligten Universitäten ihre Verpflichtungen gegenüber dem Experiment erfüllen können. Dies betrifft insbesondere Betrieb und Wartung des Experiments. Darüber hinaus hat das BMBF auch den deutschen Beitrag zum Bau des Experiments gefördert. Die Verhandlungen mit dem BMBF über eine Förderung des geplanten Upgrades des Experiments haben begonnen.

Der Zugang zu den Daten des Experiments ist ebenfalls langfristig gesichert.

3.7 Ausstattung des Teilprojekts

3.7.1 Bisherige Förderung

Das Teilprojekt wird bisher nicht gefördert und es wurde kein Antrag auf Förderung gestellt.

3.7.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Doktoranden, 100%	2	117.600	2	117.600	2	117.600	2	117.600
Summe:	—	117.600	—	117.600	—	117.600	—	117.600
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Investitionsmittel:	Betrag		Betrag		Betrag		Betrag	
Summe:	0		0		0		0	
Summe insgesamt:	117.600		117.600		117.600		117.600	

3.7.3 Personal

	Nr.	Name, akademischer Grad, Dienststellung	engere Fachzugehörigkeit	Institut der Hochschule oder der außeruniversitären Einrichtung	Projekt-mitarbeit in Wochen-stunden	Kategorie	Finanzierungs-quelle
Vorhandenes Personal (C5)							
Wissen-schaftliches Personal	1	Teubner, Jens, Prof. Dr., Prof	Informationssys-teme	LS 6, Fak. Inf, TU Do	8,0	—	Grund-ausstattung
	2	Spaan, Bernhard, Prof. Dr., Prof	Teilchenphysik	Experimentelle Physik 5, Fak. Physik, TU Do	8,0	—	Grund-ausstattung
	3	Wishahi, Julian, Dr., PostDoc	Teilchenphysik	Experimentelle Physik 5, Fak. Physik, TU Do	3,98	—	Grund-ausstattung
	4	Niet, Ramon, MSc. Physik, WiMi	Teilchenphysik	Experimentelle Physik 5, Fak. Physik, TU Do	9,95	—	Grund-ausstattung
	5	Eitschberger, Ulrich, WiMi	Teilchenphysik	Experimentelle Physik 5, Fak. Physik, TU Do	9,95	—	Grund-ausstattung
	6	Meier, Frank, WiMi	Teilchenphysik	Experimentelle Physik 5, Fak. Physik, TU Do	7,96	—	Grund-ausstattung
	7	N.N. , WiMi	Informationssys-teme	LS 6, Fak. Inf, TU Do	27,86	—	Grund-ausstattung
	8	N.N. , SHK	Teilchenphysik	Experimentelle Physik 5, Fak. Physik, TU Do	10,0	—	Grund-ausstattung
Nichtwissen-schaftliches Personal	9	Stickel, Britta, Verw. Ang.	—	Experimentelle Physik 5, Fak. Physik, TU Do	3,98	—	Grund-ausstattung
	10	Stankjawitschene, Alla, Verw. Ang.	—	LS 6, Fak. Inf, TU Do	1,99	—	Grund-ausstattung

	11	Domke, Matthias, Techn. Ang.	—	Experimentelle Physik 5, Fak. Physik, TU Do	3,98	—	Grund-ausstattung
	12	Warda, Kai, Techn. Ang.	—	Experimentelle Physik 5, Fak. Physik, TU Do	7,96	—	Grund-ausstattung
Beantragte Personalmittel (C5)							
Wissen-schaftliches Personal	13	N.N. , WiMi	Teilchenphysik	Experimentelle Physik 5, Fak. Physik, TU Do	—	Doktorand/in	—
	14	N.N. , WiMi	Informationssys-teme	LS 6, Fak. Inf, TU Do	—	Doktorand/in	—

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Teubner, Jens**
   Projektleitung. Schwerpunkt Datenbanksysteme und Verteiltes Rechnen. Mitarbeit in allen Arbeitspaketen.

2. **Spaan, Bernhard**
   Projektleitung. Schwerpunkt Teilchenphysik. Koordination der Arbeiten mit LHCb.

3. **Wishahi, Julian**
   Unterstützung bei Simulationsrechnungen zum SciFi-Tracker und Trigger.

4. **Niet, Ramon**
   Unterstützung der Datenanalyse, Evaluation von Analysewerkzeugen.

5. **Eitschberger, Ulrich**
   Unterstützung der Arbeitspakete 1-3 in Hinblick auf LHCb-Software.

6. **Meier, Frank**
   Unterstützung der Arbeitspakete 1-3 in Hinblick auf LHCb-Software.

7. **N.N.,**
   Arbeitspakete 1 und 2.

8. **N.N.,**
   Begleitende Arbeiten.

9. **Stickel, Britta**
   Administrative Unterstützung, Unterstützung bei Organisation von Workshops und Dienstreisen etc.

10. **Stankjawitschene, Alla**
    Sekretariat.

11. **Domke, Matthias**
    Cluster Maintenance, GPU-Konfiguration.

12. **Warda, Kai**
    Unterstützung der FPGA- und GPU-Arbeiten.

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

13. **N.N.,**
    FPGA- und GPU- basierte Datenvorverarbeitung, gegen Projektende Mitarbeit in AP4.

14. **N.N.,**
    Arbeitspaket Programmiermodelle; außerdem Mitwirkung beim Profiling (während Anfangsphase, AP1 sowie gegen Teilprojektende, AP4).

3.7.4 Sachmittel

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	8.000	8.000	8.000	8.000
Summe der vorhandenen Mittel	8.000	8.000	8.000	8.000
Beantragte Mittel	0	0	0	0

3.1 Allgemeine Angaben zum Teilprojekt MGK

3.1.1 Titel: Integriertes Graduiertenkolleg

3.1.2 Projektleitung

Prof. Dr. Dr. Rhode, Wolfgang, 17.10.1961

Experimentelle Physik 5, Fakultät Physik, Technische Universität Dortmund
Otto-Hahn-Straße 4
44227 Dortmund

Tel.: 0231-755-3550
Fax : 0231-755-4547
E-Mail: `rhode@physik.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.2 Zusammenfassung

Das integrierte Graduiertenkolleg soll Doktorandinnen und Doktoranden sowie Post-Doktorandinnen und Post-Doktoranden des Sonderforschungsbereichs in stetiger Zusammenarbeit miteinander zu einem international herausragenden Qualifikationsprofil verhelfen, in dem Kompetenzen in Datenanalyse und eingebetteten Systemen erstmals zusammengeführt werden. Die jungen GK-Mitglieder sollen in der internationalen Forschung sichtbar werden. Über eine Teilnahme an Konferenzen hinaus sollen sie an eine aktive Gestaltung von Workshops und eine Mitgliedschaft in Programmkomitees herangeführt werden.

Das Konzept des Graduiertenkollegs des SFB 876 folgt im Wesentlichen dem Konzept des Einrichtungsantrages, das sich in der ersten Förderperiode bewährt hat. Zugang zu dem in den SFB integrierten Graduiertenkolleg erhalten exzellente, über internationale Ausschreibungen gefundene, junge Wissenschaftlerinnen und Wissenschaftler. Erwartet wird, dass sie bereits einen (Studien-)Schwerpunkt auf einem der Felder besitzen, die für ihr spezifisches Forschungsprojekt relevant sind. Bei einem Studium in Dortmund wird diese Qualifikation durch Teilnahme an ausgewählten, auf Themen des SFB bezogenen, Vorlesungen und ggf. einer Masterarbeit auf dem Feld des SFB erworben. In dem Graduiertenkolleg wird dieses Vorwissen in strukturierter Form zu einem umfassenden kombinierten Wissensprofil in Datenanalyse und eingebetteten Systemen komplettiert.

Die Basis der Ausbildung im GK bildet ein auf die Probleme des Sonderforschungsbereichs abgestimmter Kanon von kolloquialen Lehrveranstaltungen, die von Mitgliedern des SFB gehalten werden. In ihnen werden die Promovenden schnell, effektiv und in voller Breite an das internationale Forschungsniveau herangeführt.

Höchste Priorität nimmt die Ausbildung zur eigenständigen wissenschaftlichen Forschung ein. Die Promovenden stellen in den *Topical Seminars* der Arbeitskreise des SFB (Datenanalyse, Ressourcenbeschränkung und Informationsgewinnung) unter der Leitung von Teilprojektleiterinnen und -leitern eigene Arbeiten vor und diskutieren aktuelle Forschungsfragen mit eingeladenen Sprechern und Sprecherinnen. Da beabsichtigt ist, frühzeitig sehr gute Studierende als Doktorandinnen und Doktoranden zu gewinnen, sollen an den Lehrveranstaltungen auch Studierende teilnehmen, die ein Teilprojekt des SFB als Thema ihrer Bachelor- oder Masterarbeit gewählt haben. Einer begrenzten Zahl von Promovenden soll die Einarbeitung in ihr Promotionsthema noch in Gegenwart von scheidenden Projektmitarbeitern durch Stipendien ermöglicht werden.

In *Technical Skill*-Kursen werden Fähigkeiten für dem Umgang mit relevanten Softwarewerkzeugen wie RapidMiner, R, root u.a. vermittelt. Die Ausbildung wird abgerundet durch Kurse, in denen *Soft Skills* wie die kommunikative Kompetenz der Promovenden ausgebildet werden. Für letztere Kurse steht den Promovenden vorrangig das gesamte Programm der Mitarbeiterweiterbildung der TU Dortmund zur Verfügung.

Im jährlichen Wechsel finden Workshops und Sommerschulen statt. Hier präsentieren die Promovenden ihre wissenschaftlichen Ergebnisse und stellen ihre erworbenen Organisationsfähigkeiten unter Beweis.

3.3 Tabellarischer Bericht

Die Mittel des SFB 876 wurden, wie in der Tabelle 3.1 dargelegt, zur Förderung von Promovenden und Stipendiaten verwandt.

	Promovenden	Stipendiaten
Anzahl	44	3
Förderungsdauer, minimal	3 Monate	6 Monate
Förderungsdauer, durchschnittlich	28,91 Monate	6 Monate
Förderungsdauer, maximal	48 Monate	6 Monate
Förderungshöhe, minimal	25 % E13	1000 Euro/Monat plus 103 Euro Sachkosten
Förderungshöhe, durchschnittlich	84,7 % E13	1000 Euro/Monat plus 103 Euro Sachkosten
Förderungshöhe, maximal	100 % E13	1000 Euro/Monat plus 103 Euro Sachkosten

Tabelle 3.1.: Tabellarischer Bericht über die zur Förderung der Promovenden verplanten Mittel im gesamten Förderzeitraum. Kurze Förderzeiten oder geringe prozentuale Stellenanteile entstehen durch Wechsel von der SFB-Finanzierung in die Grundfinanzierung bzw. durch anteilige Förderung. Die maximale Förderdauer betrifft immer die geplante Förderungszeit.

Liste der abgeschlossenen Dissertationen (2011—2014, inkl. Arbeiten aus der Grundausstattung):

- Anita M. Thieler, Robuste Verfahren zur Periodendetektion in ungleichmäßig beobachteten Lichtkurven (2014)
- Cornelia Tadros, Belief Change Operations under Confidentiality Requirements in Multiagent Systems (2014)
- Bjoern Dusza, Context-Aware Power Consumption Modeling for Energy Efficient Mobile Communication Services (2014)
- Jan-Hendrik Koehne, Der Leptonpropagator PROPOSAL (2013)
- Marcel Martin, Algorithms and Tools for the Analysis of High-Throughput DNA Sequencing Data (2013)
- Daniel A. Cordes, Automatic Parallelization for Embedded Multi-Core Systems using High-Level Cost Models (2013)
- Tim Ruhe, Data Mining on the Rocks – A measurement of the atmospheric neutrino flux using IceCube in the 59-string configuration and a novel data mining based approach to unfolding (2013)
- Fabian Bohnen, Eine Methodik zur Produktionsnivellierung auf der Basis von Fertigungsfamilien (2013)
- Sascha Plazar, Memory-based Optimization Techniques for Real-Time Systems (2012)
- Constantin Timm, Resource Efficient Processing and Communication in Sensor/Actuator Environments (2012)
- Natalie Milke, Unfolding of the atmospheric neutrino flux spectrum with the new program TRUEE and IceCube (2012)
- Kai Kammers, Survival models with gene groups as covariates (2012)
- Felix Jungermann, About the Exploration of Data Mining Techniques using Structured Features for Information Extraction (2012)
- Michael Backes, Long-term observations of the TeV blazar 1ES 1959+650 (2011)

Im Rahmen der Promotionsvorbereitung betreute Stipendiaten:

- Venkatesh Umaashankar (betreut durch Katharina Morik, C1): SFB-Report zu einem RapidMiner-Plugin für Optimierungsverfahren.
- Amer Krivosija (betreut durch Christian Sohler, A2): aufgenommen als Promovend der Grundausstattung.
- Orwa Nassour (betreut durch Olaf Spinczyk, A1): der Kandidat verfügte über ein eigenes Stipendium, Report: [A1/55].
- Tamara Talakesh (betreut durch Wolfgang Rhode, C3): Stipendium läuft zur Zeit der Antragstellung.

Zusätzlich wurden von Promovenden des SFB 876 RISE-Projekte definiert und ausgeschrieben. Je nach Schwerpunkt des ausgeschriebenen Projektes erfolgte die Finanzierung über den SFB 876 (in drei Fällen) oder direkt über den DAAD.

3.4 Qualifizierungskonzept

Wir setzen voraus, dass die Promovierenden im integrierten Graduiertenkolleg über einen qualifizierten Studienabschluss in einer der am SFB beteiligten Disziplinen verfügen.

Um sehr gute und entsprechend qualifizierte Promovenden anwerben zu können, werden zwei Wege eingeschlagen. Internationale Kandidat/inn/en werden über eine permanente Ausschreibung und wo immer möglich über persönliche Kontakte der Teilgruppenleiter angesprochen. Ihre Eignung und ihr Interesse für eine Promotion im SFB kann ggf. über ein Stipendium für einige Monate getestet werden. Wegen der fachlich neuen Ausrichtung des SFB ist es jedoch nicht einfach, geeignete internationale Kandidaten anzuwerben. Dortmunder Kandidat/inn/en werden in Masterarbeiten und in darüber hinausgehenden Qualifizierungsmaßnahmen (u.a. Teilnahmen an den GK-Veranstaltungen) auf eine potentielle Promotion im SFB vorbereitet.

Unser Ziel ist, die Absolventen nach ihrer Promotion als kompetente und nach Möglichkeit auch international bekannte Nachwuchswissenschaftler aus dem SFB zu entlassen. Die auf eine Promotion folgenden Schritte richten sich dann entweder direkt auf eine berufliche Tätigkeit in der Industrie oder ggf. über die Annahme einer Forschungsstelle im Ausland oder den Aufbau einer Nachwuchsgruppe auf einen Verbleib im System der mit öffentlichen Geldern geförderten Forschung und Lehre aus. In beiden Szenarien wird von den Promovierten erwartet, selbständig und kreativ zu forschen, diese Forschung zu organisieren und nach außen zu kommunizieren und in der Lehre zu vertreten.

Dass die Promovierenden in der Lage sind, fachliche Inhalte in Vorlesungen zu erlernen, haben sie im Laufe ihres Studiums bewiesen. Der SFB ist mit Projekten mit Kernanforderungen aus den Gebieten der Informatik, Statistik, Physik, Maschinenbau, Medizin, Elektrotechnik und Biologie so interdisziplinär angelegt, dass jede/r Promovierende Gebiete findet, deren Inhalt er/sie sich zur Bewältigung der anstehenden Forschungsaufgaben aneignen muss. Da diese Inhalte sich von Promovend/in zu Promovend/in unterscheiden, haben wir für sie – mit nachdrücklichster Unterstützung der Promovierenden – bewusst keinen verpflichtenden[1] Kanon an zu hörenden Vorlesungen definiert. Stattdessen verpflichten sich die Promovierenden in einer Promotionsvereinbarung dazu, diese Wissenslücken durch eigenverantwortlichen Besuch von Lehrveranstaltungen zu füllen. Auf eine Vergabe von *Credit Points* für Unterschriften auf Anwesenheitslisten bei dem Besuch von Veranstaltungen wurde von Anfang an zugunsten der Eigenverantwortlichkeit der Promovierenden bewusst verzichtet. Wie die lebhafte interdisziplinäre Diskussionsatmosphäre, die sich unter den Promovierenden und in den GK-Veranstaltungen etabliert hat, zeigt, war diese Entscheidung richtig (siehe. auch Abschnitt 3.7).

Vorlesungen und Themengebiete, die den Promovierenden insbesondere zur Ergänzung ihres Wissens in Hinblick auf die Kernfragen des SFB 876 empfohlen werden, sind z.B.:

- Real-Time Systems (Chen, Informatik)
- Probabilistische graphische Modelle (Kersting, Informatik)
- Numerical Optimization (Lee, Informatik)
- Eingebettete Systeme (Marwedel oder Chen, Informatik)
- Cyber-Physical System Fundamentals (Marwedel, Chen, Informatik)

[1] Eine Anwesenheitspflicht für Vorlesungen darf aus rechtlichen Gründen an der TU Dortmund nicht ausgesprochen werden.

- Wissensentdeckung in Datenbanken (Morik oder Kersting, Informatik, und Weihs oder Ligges, Statistik)
- Maschinelles Lernen (Morik, Informatik)
- Computational Omics (Rahmann, Informatik)
- Software ubiquitärer Systeme (Spinczyk, Informatik)
- Data Processing on Modern Hardware (Teubner, Informatik)
- Einführung in die Kern- und Elementarteilchenphysik (Spaan, Physik)
- Astroteilchenphysik (Rhode, Physik)
- Statistische Methoden der Datenanalyse (Rhode, Physik)

- Modellierung und Dimensionierung von Kommunikationssystemen (Wietfeld, Elektrotechnik)
- Mobilfunknetze (Wietfeld, Elektrotechnik)

- Klassifikationsverfahren (Ickstadt, Statistik)
- Bayes-Statistik (Ickstadt, Statistik)
- Generalisierte lineare Modelle (Ickstadt, Statistik)
- Statistik in der Genetik (Ickstadt, Statistik)

Der Schwerpunkt der Ausbildung im Graduiertenkolleg liegt in der Ausbildung der Fähigkeiten zur selbständigen Forschung auf internationalem Niveau. Durch internationale Ausschreibung der Promovenden- und Wissenschaftlerstellen bemühen wir uns, eine internationale Atmosphäre zu schaffen. Die *Topical Seminar* genannten Vortrags- und Diskussionsveranstaltungen wie auch die Projektseminare werden in der Regel in englischer Sprache abgehalten. Auf Teilnahmen der Promovierenden an internationalen Tagungen wird höchster Wert gelegt.

Die Promovierenden übernehmen im Qualifizierungskonzept eine aktive Rolle im SFB. Sie kommunizieren untereinander und leiten Studierende während deren Bachelor- und Masterarbeiten an, sodass ihre Arbeit von Anfang an durch Kooperation geprägt ist. Sie laden Gastwissenschaftler/innen ein, organisieren Intensivkurse (z.B. über RapidMiner, R und Präsentationstechniken) und planen und leiten die konferenzähnlichen jährlichen SFB-Workshops. In erster Linie werden sie dabei von dem/r für ihre Ausbildung in formalen Dingen vorgesehenen Wissenschaftler/in unterstützt, weiter aber auch von dem gesamten Sonderforschungsbereich, insbesondere vom Vorstand. So lernen sie, selbst Workshops und Schulen zu planen, Forscher anzuschreiben und eine Zusammenarbeit mit ihnen zu gestalten. Sie lernen, wie Forschung in der Praxis funktioniert, indem sie sich engagieren, sind aber abgesichert durch die Begleitung seitens des Sonderforschungsbereichs. Neben dem Lerneffekt, den eigenständiges Arbeiten mit sich bringt, und der Motivation, die durch den Freiraum für eigene Vorstellungen eintritt, werden sie auf diese Weise auch außerhalb des Sonderforschungsbereichs bekannt. Dass solch ein Qualifikationsprofil auch für den Beruf höchst interessant ist, zeigt sich u.a. daran, dass keinerlei Schwierigkeiten von Promovierten bei der Stellensuche bekannt sind und zudem selbst Promovierenden bereits attraktive Angebote unterbreitet wurden.

Das Studienprogramm des Integrierten Graduiertenkollegs

Das **Studienprogramm** ist so gestaltet, dass den Promovierenden ein größtmöglicher Freiraum für die Forschung eingeräumt wird. Die minimale Belastung liegt im Semester bei 4 SWS. Diese Zeit verteilt sich auf zwei doppelstündige Veranstaltungen:

- In das *Topical Seminar*, in dem nach Vorträgen vor, von und mit eingeladenen Gästen und Lehrenden des GK fachliche Inhalte diskutiert werden. Die Veranstaltungen der *Topical Seminars* werden von den Promovierenden in den Arbeitskreisen organisiert, in denen die Gäste vorgeschlagen und eingeladen werden, die Gäste werden nach Möglichkeit auch von den Promovierenden betreut. Die *Topical Seminars* unterscheiden sich von klassischen Kolloquien durch einen hohen Anteil an Diskussionen – bevorzugt mit den Promovierenden.

- In ebenfalls doppelstündige Projektseminaren, in denen die Beteiligten und insbesondere die Promovierenden ihre aktuelle Forschung in Vorträgen vorstellen und aktuelle Probleme diskutieren.

Zu diesen Veranstaltungen kommt die Verpflichtung der Promovierenden, ggf. fehlende fachliche Kenntnisse über Spezialvorlesungen zu erwerben.

Überwiegend in der vorlesungsfreien Zeit werden Blockkurse zu den *Soft-* und *Technical Skills* angeboten. Hier steht den Promovierenden prioritär das gesamte Mitarbeiterbildungsprogramm der TU Dortmund zur Verfügung.[2] Eine praktische Anwendung findet die Ausbildung in *Soft Skills* in der Beteiligung der Promovierenden an der Lehre und Ausbildung von Studierenden in der Qualifikationsphase.

Die Promovierenden präsentieren ihre Ergebnisse auf den jährlich in Dortmund von ihnen organisierten Workshops in schriftlicher und mündlicher Form. Geeignete Ergebnisse werden ebenfalls auf (internationalen) Konferenzen präsentiert. In jedem zweiten Jahr (2012, 2014) organisieren die Promovierenden mit Unterstützung durch den Vorstand eine Sommerschule mit auswärtigen Lehrenden.

Sommerschulen

- In der ersten Sommerschule 2012 zu dem Schwerpunkt *Resource-aware Machine Learning*[3] wurden Unterrichtseinheiten zu folgenden Themen abgehalten:
  - Introduction to Machine Learning, Michèle Sébag (LRI Univ. Paris-Sud, Frankreich)
  - Numerical Optimization in Data Analysis, Sangkyun Lee
  - Data Mining with RapidMiner, Tim Ruhe
  - Data Mining from Ubiquitous Data Streams, Joao Gama (Univ. Porto, Portugal)
  - Statistical Methods for Model Selection, Jörg Rahnenführer, Michel Lang
  - Exploitation of Memory Hierarchies, Peter Marwedel
  - Battery Capacity Models, Peter Marwedel
  - Towards Self-Powered Systems, Jan Madsen (TU Denmark Lyngby, Dänemark)
  - From Data Taking to Rapid Mining, Wolfgang Rhode
  - Use of the Power of your GPU: Massively Parallel Programming with CUDA, Nico Piatkowski

  Es sei darauf hingewiesen, dass neben den internationalen Lehrenden und den Projektleitern aus dem SFB mit den Sprechern Lee, Ruhe, Lang und Piatkowski auch jüngere Wissenschaftler aus dem SFB bzw. GK zu den Lehrenden gehörten und jeweils ihre eigenen Kurse organisiert und durchgeführt haben.

  Neben den Vorlesungen wurde auf der Sommerschule zur praktischen Ausbildung ein Wettbewerb zur Datenanalyse durchgeführt. Organisiert durch Katharina Morik und Olaf Spinczyk bestand die Aufgabe in der Analyse von Smartphonedaten, die von der Sensorik (Beschleunigung, Lage, Helligkeit, etc.) über Ortsinformationen (Sichtbare Mobilfunkzellen, GPS-Koordinaten, etc.) bis hin zu Systemwerten (Ladestand des Akkus, Speicherausnutzung) reichen. Für die nötige umfassende Sammlung von neuen Nutzungs- und Sensordaten wurde im Rahmen des Projekts A1 eine Applikation für Android-Smartphones entwickelt, mit der in den Wochen vor und während der Sommerschule Daten gesammelt wurden. Informationen dieser Art können beispielsweise für eine dynamische Lastverteilung genutzt werden. Die Privatsphäre aller Datensammler wurde neben der vollständigen Anonymisierung sämtlicher personenbezogener Daten auch dadurch geschützt, dass GPS-Koordinaten entfernt wurden. Gleichzeitig hat dies die Lernaufgaben weiter erschwert. Neben den Daten hatten alle Teilnehmer einen RapidMiner-Beispielprozess zur Verfügung, der Platzhalter für die Schritte Vorverarbeitung und Merkmalsauswahl, Modelltraining und Vorhersage enthielt. Insbesondere durch die große Anzahl verfügbarer Merkmale war der erste Schritt der geeigneten Merkmalsauswahl von Bedeutung für einen Lernerfolg.

  Mit Blick auf die Fußballmeisterschaft berechneten Promovierende des SFB um Tim Ruhe öffentlichkeitswirksam mit Methoden des maschinellen Lernens Vorhersagen für den Ausgang der Spiele der Fußball-Europameisterschaft 2012 und waren mit ihrem Modell insofern erfolgreich, dass sie gegen Oddset in einer hypothetischen Wette insgesamt gewonnen hätten.

  Die Sommerschule richtete sich an alle Promovierenden oder weit fortgeschrittene Masterstudierenden, die sich mit maschinellem Lernen beschäftigen und dabei insbesondere auf Beschränkungen von Rechenleistung, Kommunikation oder Energie achten. Die Schule wurde international geöffnet und beworben,

[2] http://www.zhb.tu-dortmund.de/wb/de/home/TU_interne_Weiterbildung/
[3] http://sfb876.tu-dortmund.de/SummerSchool2012/index.html

sodass außer 50 Mitgliedern des SFB auch 30 externe Teilnehmer aus zwölf Nationen teilgenommen haben.

- In der zweiten Sommerschule 2014[4] sind Unterrichtseinheiten zu folgenden Themen geplant:
  - Introduction to Machine Learning, Céline Robardet (INSA, Univ. Lyon, Frankreich)
  - Coresets for k-means Clustering, Melanie Schmidt
  - From k-means clustering to DESICOM: Matrix Factorization for Data Analysis, Christian Bauckhage (Univ. Bonn)
  - Cyber-Physical Systems: Opportunities, Challenges and (some) Solutions, Peter Marwedel
  - Computation offloading for performance and energy improvements, Jian-Jia Chen
  - Privacy Aware Learning, John Duchi (Univ. Berkeley, USA)
  - Model Compression, Rich Caruana (Cornell Univ., USA)
  - Processing Data Streams: From Small Devices to Big Data, Christian Bockermann
  - Introduction to Astroparticle Physics Data Processing, Marlene Doert
  - Astroparticle Detectors as Embedded System, Tim Ruhe

  Auch hier gehören mit Christian Bockermann, Marlene Doert und Tim Ruhe wieder eine junge Wissenschaftlerin und zwei junge Wissenschaftler aus dem SFB zu den Lehrenden. Als Wettbewerbsaufgabe sollen in diesem Jahr Rauschanalysen von Kamerabildern, die bei völliger Abdunkelung der Kamera aufgezeichnet werden, durchgeführt werden. In dem Rauschen sollen die Muster identifiziert werden, die in dem Sensor-Chip durch Elementarteilchen aus der kosmischen Strahlung oder natürlichen Radioaktivität ausgelöst werden. Die von den Teilchen ausgelösten Defekte im Kamerabild visualisieren unvermeidbare Bit-Switch-Muster und -Raten, die bei der Miniaturisierung von Speicher- und CPU-Chips sowie bei der Optimierung von Programmcode berücksichtigt werden müssen.

- Workshop on Data Streams: Vom 23. bis 27. Juli 2012 fand, organisiert von Christian Sohler und seinen Mitarbeitern, an der TU Dortmund der „Workshop on Algorithms for Data Streams“ statt. In ca. 30 Vorträgen trugen weltweit führende Forscher im Bereich der Theorie der Datenstromalgorithmen über aktuelle Entwicklungen des Feldes vor. Unter anderem wurden die neuesten Ergebnisse zu den Themen lineare Einbettungen, Matrixapproximation, Graphströme, Compressive Sensing, Sampling aus Datenströmen sowie untere Schranken für die Komplexität von Datenstromalgorithmen diskutiert. Viele der auf dem Workshop präsentierten Arbeiten sind später auf renommierten Tagungen erschienen und unsere Graduierten hatten die Möglichkeit, sich bereits lange vor dem Erscheinen der entsprechenden Tagungsbände über die neuesten Entwicklungen in diesem Bereich zu informieren. Außerdem haben einige Mitglieder des SFB die Möglichkeit genutzt, eigene Arbeiten auf dem Workshop zu präsentieren[5].

Regeln guter wissenschaftlicher Arbeit

Die Aufklärung der Promovierenden über die Regeln guter wissenschaftlicher Arbeit sowie projektspezifischer Aspekte der Ethik oder Vertraulichkeit erfolgt in den Teilprojekten.

Programmkomponenten für die der Promotionsphase vorangehende Qualifizierungsphase

Die Qualifizierungsphase besteht für zukünftige Promovierende des SFB 876 im Wesentlichen im letzten Jahr ihres Masterstudiums, ggf. kann auch die Anfertigung der Bachelorarbeit oder der Forschungsarbeit von RISE-Stipendiaten zu dieser Phase gerechnet werden. Im Vordergrund steht hier, dass die Studierenden unter Anleitung der Promovierenden und anderer projektbeteiligter Wissenschaftlerinnen und Wissenschaftler an die Methodik des wissenschaftlichen Forschens herangeführt werden. Dazu gehört ggf. ebenso die in den Zeitplan der Abschlussarbeiten passende Vermittlung der relevanten Programmiertechniken wie die Möglichkeit zur Teilnahme an den Veranstaltungen des SFB. Nach Möglichkeit werden Studierende in dieser Phase über die Teilprojekte als studentische Hilfskräfte eingestellt.

Gastwissenschaftlerinnen und Gastwissenschaftler im Integrierten Graduiertenkolleg

Gastwissenschaftlerinnen und Gastwissenschaftler stellen sich und ihre Forschung in der Regel kurz nach ihrer Ankunft durch einen Vortrag im *Topical Seminar* vor. Sie nehmen in der Folgezeit an den Veranstaltungen der Projektseminare und der *Topical Seminars* teil. So kann und soll zwischen ihnen und den Promovierenden auch über den engeren Kreis der eigenen Forschung hinaus kommuniziert und diskutiert

[4] http://sfb876.tu-dortmund.de/SummerSchool2014/index.html
[5] http://ls2-www.cs.uni-dortmund.de/streamingWS2012/

werden. Über diese Veranstaltungen hinausgehende Kontakte sind selbstredend völlig im Sinn des SFB. Entsprechendes gilt für Gastwissenschaftlerinnen und Gastwissenschaftler, die zu den Sommerschulen eingeladen werden.

Maßnahmen zur wissenschaftlichen bzw. berufsbezogenen Qualifizierung

- Die TU Dortmund hat sich in besonderem Maße der Frauenförderung durch ein Mentoring-Programm[6] mit Schwerpunkt in den Ingenieurwissenschaften verschrieben: *„mentoring³“*. Seit 2005 besteht dieses Gemeinschaftsprogramm der Technischen Universität Dortmund, der Ruhr-Universität Bochum und der Universität Duisburg-Essen. Es wird vom Ministerium für Innovation, Wissenschaft und Forschung in NRW sowie vom Verband Deutscher Ingenieure unterstützt. Es bringt Doktorandinnen mit Mentoren (Professoren und Professorinnen, habilitierten Wissenschaftlern und Wissenschaftlerinnen, Führungskräfte der Industrie) zu einem informellen Erfahrungsaustausch zusammen, bei dem es nicht um die Betreuung der Doktorarbeit, sondern um Karriereplanung, -entwicklung und Beratung beim Umgang mit den wissenschaftlichen Institutionen wie etwa Tagungen und Zeitschriften geht. Sechs Promovendinnen aus dem SFB 876 haben an dem Programm teilgenommen.

- An der TU Dortmund wird der Prozess hin zu einem international orientierten Studium auch für deutsche Studierende konsequent vorangetrieben. Dazu gehört das Zertifikat Studium International, das das vielfältige kulturelle Potenzial der TU Dortmund nutzt, indem es den Austausch zwischen deutschen und ausländischen Studierenden auch außerhalb der Universität fördert und individuelles Engagement im internationalen Bereich belohnt. Auch die Promovierenden des SFB haben so die Möglichkeit, sich neben ihrem Studium für ihre spätere Karriere zu profilieren, indem sie wichtige Schlüsselkompetenzen z.B. in den Bereichen Kommunikation oder kulturelle Beziehungen erwerben. Zum Erwerb des Zertifikates müssen alle Teilnehmerinnen und Teilnehmer an einem Trainingsprogramm zur interkulturellen Kompetenz teilnehmen und zusätzlich zwei von folgenden drei Bedingungen erfüllen: (1) ein mindestens einmonatiger Forschungsaufenthalt im Ausland, (2) Absolvierung eines Sprachkurses am Sprachenzentrum der TU Dortmund oder (3) Engagement bei dem Internationalen Campus der TU Dortmund. Dieses Zertifikat wird von dem Akademischen Auslandsamt der TU Dortmund vergeben.[7]

- In der Endphase des Projektes soll geeigneten Promovierenden die Möglichkeit gegeben werden, am Leibniz-Zentrum für Informatik, Schloss Dagstuhl, Klausurtagungen zum gemeinsamen Verfassen von Büchern zum Thema des SFB abzuhalten.

Verteilung der Lehrveranstaltungen auf die Universitäten der UA Ruhr

An dem SFB 876 sind die beiden Standorte der Universitätsallianz Ruhr (UA Ruhr) Duisburg-Essen und Dortmund beteiligt, an denen geeignete Lehrveranstaltungen von Studierenden bereits im normalen Studium wahlweise auch an einem anderen Standort gehört werden (können). Der räumliche Abstand spielt hier keine relevante Rolle.

3.5 Organisations- und Betreuungskonzept

Anwerbung von Promovierenden

Wir möchten im SFB 876 Absolventen mit einem sehr gut geeigneten Qualifikationsprofil auf das Niveau internationaler Forschung führen und sie als bereits bekannte Nachwuchswissenschaftler entlassen. Sofern die Absolventen vorher bereits in Dortmund studiert haben, hatten die Mitglieder des SFB die Chance, Studierende in Lehrveranstaltungen für das Programm des SFB zu begeistern und insbesondere guten Studierenden Bachelor- und Masterthemen anzubieten. In den beteiligten Fachbereichen stellen die Lehrstühle – traditionell in etwas unterschiedlicher Kultur – den Studierenden mögliche Themen für Abschlussarbeiten vor. Die Promovendensprecher/innen des SFB organisieren zur Zeit eine einheitliche Vorstellung des SFB für Studierende in Form eines eintägigen Workshops (siehe Abschnitt 3.7).

Alle im SFB 876 angebotenen Stellen sind über eine permanente Ausschreibung über die Webseiten des SFB 876 grundsätzlich immer auch international ausgeschrieben.

Das Ausschreibungsverfahren wird von den Leitern des Teilprojektes, dem die Stelle zugeordnet ist, durchgeführt. Neben einem qualifizierten Abschluss wird von den Kandidat/inn/en durch Vorlesungen oder Seminare nachweislich vertieftes Wissen mindestens in einem der Schwerpunktsbereiche *moderne Datenbanksysteme*, *Rechnerarchitektur* und *Datenanalyse* erwartet.

[6] http://www.mentoring-hoch3.de
[7] http://www.aaa.uni-dortmund.de/cms/de/Internationale_Studierende/PROFIN/

Geeignete Kandidat/inn/en werden ausgewählt und zu einem Vorstellungsgespräch mit einem thematisch passenden Vortrag vor den Beteiligten des Teilprojektes eingeladen. Der/die Kandidat/in, dem/der die Stelle angeboten wird, wird von den Teilprojektleitern unter Berücksichtigung des Votums der Projektmitarbeiter ausgewählt. Der Vorschlag wird dem Leiter des Graduiertenkollegs vorgelegt, der die Einhaltung der Qualifikationskriterien kontrolliert und dem SFB-Vorstand sowie den Vertretern der Promovierenden berichtet.

An Themen der Teilprojekte arbeitende Promovierende, die *aus der Grundausstattung* bezahlt werden, werden in das Graduiertenkolleg aufgenommen, wenn sie entsprechenden Qualifikationskriterien genügen. Auch solch ein Vorschlag wird dem Leiter des Graduiertenkollegs vorgelegt, der die Einhaltung der Qualifikationskriterien kontrolliert und dem SFB-Vorstand sowie den Vertretern der Promovierenden berichtet. Mit dem Promovierenden wird eine Promotionsvereinbarung abgeschlossen (Siehe Abb. 3.1). Die Aufnahme in das Graduiertenkolleg berechtigt zur Teilnahme an dem strukturierten Ausbildungsprogramm.

Um Promovierenden – insbesondere aus dem Ausland – eine optimale Einarbeitung in ihr Promotionsthema zu ermöglichen, wurden in der Laufzeit der ersten Förderperiode des SFB insgesamt 48 Personenmonate für Kurzzeitstipendien beantragt und bewilligt. Ein Kurzzeitstipendium für einen Promovierenden darf die Dauer von acht Monaten nicht überschreiten, kann aber auch für einen kürzeren Zeitraum vergeben werden. Die Dauer des Kurzzeitstipendiums wird auf die dreijährige Promotionsdauer nicht angerechnet. Kandidat/inn/en können dem SFB-Vorstand von den Projektleitern vorgeschlagen werden. Der Vorstand bemüht sich, die Mittel für inhaltlich und zeitlich geeignete Projekte zur Verfügung zu stellen.

In der ersten Förderperiode zeigte sich, dass die SFB-Themen für Dortmunder Studierende so interessant waren, dass etliche Studierende sich schon von der Bachelorarbeit systematisch auf eine Promotion vorbereitet haben. Im internationalen Vergleich scheint das Ausbildungsniveau in Dortmund so hoch zu liegen, dass nur einzelne Bewerbungen von außerhalb qualitativ geeignet waren. Andererseits verfügten hochqualifizierte Bewerber/innen häufig über finanziell deutlich besser ausgestattete Vergleichsangebote, sodass nicht leicht internationale qualifizierte Kandidat/inn/en gefunden werden konnten, die fachliches Interesse pekuniären Aspekten überordnen wollten.

Zuordnung der Promovierenden zu den Betreuer/inn/en

Die Kandidatinnen und Kandidaten bewerben sich für ein Projekt, das zumindest die Wahl eine/r/s Betreuer/in suggeriert. In den Teilprojekten des SFB 876 kooperieren jeweils immer mindestens zwei Projektleiter mit unterschiedlichem fachwissenschaftlichem Hintergrund. Aus diesen mit der speziellen Problematik der Dissertation besonders gut vertrauten Wissenschaftlerinnen bzw. Wissenschaftler wurden in der Regel zwei in der Promotionsvereinbarung als Betreuende gewählt.

technische universität dortmund

Integriertes Graduiertenkolleg des Sonderforschungsbereich 876

PROMOTIONSVEREINBARUNG[1]

zwischen **der/dem Promovierenden**

Frau / Herrn _____,

der /dem ersten **Betreuer/in**

Frau / Herrn _____,

der / dem zweiten Betreuer/in

Frau / Herrn _____,

und dem integrierten Graduiertenkolleg des SFB 876 vertreten durch [dessen Leiter]

Frau / Herrn _____.

§ 1 Thema der Dissertation

Der/die Promovierende erstellt beginnend am _____ eine Dissertation zum Thema:

_____

Das Promotionsvorhaben zum Teilprojekt ___ wurde im Exposé vom _____ beschrieben und von den Betreuer/in/en angenommen.

§ 2 Zeit- und Arbeitsplan
Zu oben genanntem Promotionsvorhaben wurde ein Zeit- und Arbeitsplan erstellt. Die Durchführung des Promotionsvorhabens ist so zu gestalten, dass die Promotion innerhalb von drei Jahren abgeschlossen werden kann. Die Betreuenden und das Graduiertenkolleg werden die Einhaltung dieses Zeitplanes nach ihren Möglichkeiten unterstützen. Eine Änderung dieses Zeitplanes bedarf des gegenseitigen Einvernehmens.

§ 3 Aufgaben und Pflichten der Betreuenden
(1) Die Betreuenden verpflichten sich zur regelmäßigen fachlichen Beratung der/des Promovierenden sowie zu regelmäßigen Gesprächen über den Fortgang der Arbeit und die Einhaltung des Zeit- und Arbeitsplanes. Die Betreuenden unterstützen die wissenschaftliche Selbständigkeit des/der Promovierenden und geeignete Maßnahmen zur Erhöhung der internationalen Sichtbarkeit der/des Promovende/i/n.
(2) Der Betreuer/ die Betreuerin verpflichtet sich zur Betreuung - unabhängig von der Dauer der Finanzierung - bis zum Abschluss der Promotion.

§ 4 Aufgaben und Pflichten der/des Promovierenden
(1) Der/die Promovierende verpflichtet sich zu einer regelmäßigen Berichterstattung über inhaltliche Teilergebnisse der Dissertation sowie die Einhaltung des Zeit- und Arbeitsplanes. Über den Fortschritt des Promotionsvorhabens wird in Einzelgesprächen mit den Betreuenden gesprochen, er wird weiter in dem Projektseminar vorgestellt. Ein schriftlicher inhaltlicher Bericht der auch die Nennung der besuchten Lehrveranstaltungen und der besuchten Workshops und Konferenzen beinhaltet ist zu den Workshops einzureichen. Diesem Bericht sind die Veröffentlichungen und technischen Noten des/der Promovende/i/n beizulegen.

[1] Diese Muster-Promotionsvereinbarung orientiert sich an den Empfehlungen der DFG (DFG-Vordruck 1.90 – 7/08).

(2) Der/die Promovierende nimmt eigenverantwortlich am Studienprogramm des Graduiertenkollegs bestehend aus den *topical seminars*, den Projektseminaren, den *skill*-Veranstaltungen (in der vorlesungsfreien Zeit und in der Vorlesungszeit) sowie den Workshops und Sommerschulen teil.

§ 5 Integration in die Teilprojekte und das Graduiertenprogramm
Das Promotionsvorhaben wird innerhalb des Teilprojektes ___ des Sonderforschungsbereiches 876 durchgeführt. Der/die Promovend/in ist in die Arbeitsgruppen der Betreuenden integriert.

§ 6 Arbeitsbedingungen der/des Promovierenden
Dem/der Promovierenden ist als Studierende/r der TU Dortmund eingeschrieben. Über die daraus folgenden Privilegien hinaus werden folgende Ressourcen zur Verfügung gestellt (bitte ggf. ergänzen): Arbeitsplatz, Rechnerzugang, Laborzugang, _____

§ 7 Einhaltung gute Wissenschaftliche Praxis
Der/die Promovierende und der/die Betreuende verpflichten sich zur Einhaltung der Grundsätze guter wissenschaftlicher Praxis. Diese Regeln sind z.B. spezifiziert unter:
http://www.tu-dortmund.de/uni/Forschung/Ressourcen/regeln_guter_wissenschaftlicher_praxis

§ 8 Schlichtung von Konflikten
Bei Konflikten zwischen dem/der Promovierenden und der/dem Betreuenden können sich die Betroffenen an den Vorstand des SFB 876 wenden.

§ 9 Vereinbarkeit von Familie und Wissenschaft
Die Vereinbarkeit von Familie und wissenschaftlicher Tätigkeit wird besonders unterstützt. Spezielle Fördermaßnahmen werden nach Bedarf vereinbart.

Ort, Datum, Unterschrift

Promovend/in 1. Betreuer/in 2. Betreuer/in

Leiter des Graduiertenkollegs

2

Abbildung 3.1.: Zu individualisierender Entwurf der Promotionsvereinbarung

Organisation der Betreuung

Es erfolgt immer zumindest eine Doppelbetreuung der Arbeit. Angeregt durch die Diskussionen in den Arbeitskreisen haben sich zusätzlich Betreuungen einzelner Aspekte der Dissertationen von dritten Projektleitern bzw. Projektleiterinnen ergeben. Die Promovierenden diskutieren ihre Arbeiten somit grundsätzlich mit mehreren Fachwissenschaftlern. Das Diskussionsklima im SFB wird von Promovierenden und Betreuenden als sehr angenehm empfunden und die durch die interdisziplinäre Betreuung erreichte Qualität der Arbeiten ist sehr gut.

Durch die Doppelbetreuung und die breite Unterstützung, die in besonderem Maße dem interdisziplinären Charakter der Ausbildung im GK gerecht werden, wird die wissenschaftliche Selbstständigkeit der Promovierenden gefördert. Aus der Beschreibung der Promotionsvorhaben und Arbeitspakete der Teilprojekte ergibt sich, dass die Promovierenden auch untereinander engen fachlichen Kontakt halten müssen um die Probleme zu lösen. Teamarbeit ist wesentliches Merkmal der Arbeiten an den Dissertationsprojekten. Das Netz der Kontakte auch zu den Betreuenden intensiviert sich weiter, wenn Promovierende ggf. gemeinsam die Ko-Betreuung eines Studierenden während dessen Bachelor- oder Masterarbeit übernehmen.

Förderung der wissenschaftlichen Selbstständigkeit der Promovierenden

Aus der Betreuungsstruktur ergibt sich eine weitgehend selbständige Position der Promovierenden. Wissenschaftlich beheimatet in ihrem interdisziplinären Promotionsprojekt diskutieren sie auf den verschiedenen vom SFB dazu geschaffenen Podien fachlich mit den SFB-Mitgliedern. Sie haben die Freiheit, geeignete Gäste und mögliche wissenschaftliche Partner für die Seminarveranstaltungen über die Promovendensprecher/innen einladen zu lassen. Durch möglichst weitgehende Beteiligung an der Organisation der konferenzartigen Workshops und der Sommerschulen können und sollen sie administrative Verantwortung übernehmen. Nach finanziellen Möglichkeiten werden Reisemittel des SFB vorrangig für die Präsentation von Ergebnissen der Promovierenden auf nationalen und internationalen Workshops und Konferenzen aufgewendet. Dort eingereichte Papiere werden natürlich zuvor im SFB diskutiert. Auch Arbeitsaufenthalte in externen Gruppen können über den SFB finanziert werden. Ziel dieser Förderung ist der Aufbau eines Netzwerkes von Kontakten der Promovierenden und eine Erhöhung ihrer internationalen Sichtbarkeit. Die Möglichkeit zur Übernahme von Betreuungsverantwortung für Bachelor- und Masterstudierende hat sich als weiterer Anreiz zur selbständigen Forschung erwiesen.

Betreuungsvereinbarung

Zwischen den Betreuenden und den Promovierenden wird eine Promotionsvereinbarung (gem. DFG-Vordruck 1.90) geschlossen (siehe Abb. 3.1)

Fortschrittskontrollen

Die in der Betreuungsvereinbarung vorgesehenen **Erfolgskontrollen** finden auf den Workshops des SFB statt, auf denen die Promovierenden den Stand ihrer Arbeit in Vorträgen und in schriftlicher Form vorstellen. Die zu diesem Anlass schriftlich eingereichten vierseitigen Berichte werden online publiziert.[8] Die Berichte werden auf den Workshops und in den Teilprojekten diskutiert, sodass die Promovierenden außer von ihren Betreuern auch von anderen Seiten Anregung und Kritik erfahren. Über eine erfolgreiche Promotion im integrierten Graduiertenkolleg des SFB 876 wird ein Zertifikat ausgestellt.

Organisationsstruktur des GK und Gestaltungsmöglichkeiten der Promovierenden

Das Graduiertenkolleg wird von den Projektleiter/inne/n des SFB gemeinsam getragen und nach außen von der Sprecherin des SFB vertreten. Der Leiter des GK führt die laufenden Geschäfte und gehört dem Vorstand des SFB an, in dem er insbesondere die Interessen der Ausbildung und der Promovierenden vertritt. Als Betreuende gehören dem GK alle Teilprojektleitende des SFB an, deren Promovierende unabhängig von der Quelle der Bezahlung, jedoch bei entsprechender fachlicher Ausrichtung und Qualifikation, als Promovierende in das GK aufgenommen worden sind. Die Promovierenden wählen einen Rat aus drei Sprecherinnen bzw. Sprechern, wobei mindestens eine Sprecherin gefunden werden muss. Die Promovendensprecher/innen können jederzeit Vorschläge machen und für Verbesserungen sorgen. Insbesondere nehmen sie an der Organisation des Gästeprogramms des SFB intensiven Anteil. Bei der Entfaltung der Aktivitäten (Einladungen, Organisation von Veranstaltungen) steht den Promovierenden der/die Geschäftsführer/in des SFB zur Seite. Generelle Beschlüsse werden auf der Versammlung aller SFB-Projektleiter/innen gefasst. Einzelfallentscheidungen werden vom SFB-Vorstand gefällt. Das Graduiertenkolleg ist immer über seinen Sprecher und in der Regel über die Promovendensprecher/innen im Vorstand des SFB vertreten.

Die Promovendensprecher/innen sind an Entscheidungen über die Ausrichtung der Forschung im SFB, die ja auch ihre Betreuer betreffen könnte, nicht beteiligt. Ebenfalls sind sie nicht an direkten Personalentscheidungen beteiligt. Sitzungen des Vorstandes finden in der Vorlesungszeit monatlich und fakultativ in der vorlesungsfreien Zeit statt. Gemeinsame Treffen aller Betreuenden und Promovierenden finden auf den

[8] http://sfb876.tu-dortmund.de/FORSCHUNG/techreports.html

Workshops statt.

Die Promovierenden sind somit direkt oder über ihre Sprecher an allen Entscheidungen über die Lehre und Ausbildung beteiligt.

3.6 Umfeld des Graduiertenkollegs

Positionierung des Integrierten Graduiertenkollegs gegenüber weiteren Programmen der Nachwuchsförderung und strukturierten Promotionsprogrammen

Angebote der Nachwuchsförderung:

- Die am SFB 876 beteiligten Hochschulen legen besonderen Wert auf die Förderung des wissenschaftlichen Nachwuchses. Die TU Dortmund und die Universität Duisburg-Essen tragen gemeinsam mit der Ruhr-Universität Bochum das Science Career Net Ruhr (SCNR). Das Netzwerk bietet das Mentoring-Programm für Doktorandinnen und promovierte Wissenschaftlerinnen (*mentoring*[3]) an. Eine jährliche Herbstakademie widmet sich Fragen zur Promotion und Weiterqualifizierung. Im August 2009 startete zudem das Programm CoachingPLUS, ein speziell auf Fachkulturen zugeschnittenes Seminarprogramm für den fortgeschrittenen wissenschaftlichen Nachwuchs. Zudem bot SCNR im Januar 2010 erstmals ein Karriereforum in Dortmund an.[9]

- Den Nachwuchswissenschaftler/inne/n stehen die Weiterbildungsangebote des Zentrums für Hochschulbildung der TU Dortmund offen, darunter Seminare zur Gestaltung der Lehre, zum Projektmanagement oder zur Lebensplanung.[10]

- Die Grundausstattungsmittel des Graduiertenkollegs für 2015 sind eingeplant für fachspezifische Kurse zum Beginn der 2. Phase, die nicht von der TU selbst angeboten werden können, z.B. ein RapidMiner-Hadoop-Kurs.

- An der UDE gibt es seit Frühjahr 2010 das Promovierendenforum (ProFor[11]), eine Koordinationsstelle für alle Promovierenden.

- An der TU Dortmund gibt es derzeit die acht unten genannten strukturierten Graduiertenprogramme, darunter ein DFG-Graduiertenkolleg und drei NRW-Forschungsschulen. Daneben gibt es Netzwerke, die fakultätsintern (z.B. Raumplanung) oder fakultätsübergreifend (Fakultäten 12—16) organisiert sind. Das Umfeld des GK des SFB 876 besteht hier aus folgenden Nachwuchsförderungsprogrammen:
  - DFG – Graduiertenkolleg Diskrete Optimierung technischer Systeme unter Unsicherheit (GRK 1855), Fakultät für Informatik, Sprecher: Univ.-Prof. Dr. Peter Buchholz, Lehrstuhl für Praktische Informatik. Vom GRK 1855 wurden in der Vergangenheit Kompaktkurse zu verschiedenen Optimierungsmethoden und der Schnittstelle Mensch-Maschine angeboten. Die Promovierenden beider Graduiertenkollegs werden ermutigt, bei Interesse an den Kompaktkursen des jeweils anderen GKs teilzunehmen.
  - NRW-Forschungsschule: Energieeffiziente Produktion und Logistik, Ruhr-Universität Bochum, TU Dortmund, hier: Standortsprecher Univ.-Prof. Dr. Claus Weihs, Fakultät Statistik. Das Thema der Energieeffizienz betrifft auch den SFB. Die Graduiertenschule ist eine Initiative der Engineering Unit Ruhr, der Kooperationsplattform der Maschinenbaufakultäten der Ruhr-Universität Bochum und der TU Dortmund, unter Beteiligung von Prof. Jochen Deuse (TU Dortmund, Sprecher) und ist somit auch personell mit dem SFB 876 verknüpft. Promovierenden der Forschungsschule und Promovierenden unseres Graduiertenkollegs stehen die fachlichen Lehrveranstaltungen der anderen Einrichtung offen.
  - NRW-Forschungsschule: Ruhr Graduate School in Economics RGS Econ Ruhr-Universität Bochum, Universität Duisburg-Essen, RWI Rheinisch-Westfälisches-Institut für Wirtschaftsforschung, TU Dortmund, hier: Standortsprecher Univ.-Prof. Dr. Wolfgang Leininger, Wirtschafts- und Sozialwissenschaftliche Fakultät
  - Graduiertenschule Land NRW: Graduiertencluster für Industrielle Biotechnologie TU Dortmund, Universität Bielefeld, Heinrich-Heine-Universität Düsseldorf, Sprecher: Univ.-Prof. Dr.-Ing. Gerhard Schembecker, Fakultät Bio- und Chemieingenieurwesen

[9] http://www.scn-ruhr.de/uamr.html
[10] http://www.zhb.tu-dortmund.de
[11] http://www.uni-due.de/profor/

  - Graduiertenschule Land NRW: Fachdidaktische Entwicklungsforschung zu diagnosegeleiteten Lehr- und Lernprozessen (FUNKEN), Sprecher: Univ.-Prof. Dr. Stephan Hußmann, Dortmunder Kompetenzzentrum für Lehrerbildung und Lehr-/Lernforschung (DoKoLL)
  - Privatwirtschaft: Graduate School of Logistics TU Dortmund, Universität Duisburg-Essen, Universität Paderborn, WWU Münster, Sprecher: Univ.-Prof. Dr.-Ing. Axel Kuhn, Fakultät Maschinenbau
  - Max-Planck-Gesellschaft, International Max Planck Research School in Chemical Biology, Max-Planck-Institut für molekulare Physiologie, TU Dortmund (Fakultät Chemie und Chemische Biologie), Ruhr-Universität Bochum, Sprecher: Univ.-Prof. Dr. Martin Engelhard, Max-Planck-Institut für molekulare Physiologie
  - Fortschrittskolleg Energieeffizienz im Quartier – clever versorgen.umbauen.aktivieren. Die Fakultäten Raumplanung, Elektrotechnik und Informationstechnik sowie die Wirtschafts- und Sozialwissenschaftliche Fakultät der TU Dortmund beteiligen sich ab Juli 2014 am Kolleg, das neben technischen auch baulich-räumliche, wirtschaftliche, juristische und soziale Fragestellungen für die Energieeffizienz im Quartier untersucht.

- Die TU Dortmund, die Universität Duisburg-Essen und die Ruhr-Universität Bochum haben sich im März 2007 zur Universitätsallianz Ruhr (UA Ruhr) zusammengeschlossen. Die Universitäten bündeln ihre Stärken in Forschung und Lehre, ohne dabei ihre Eigenständigkeit aufzugeben.

- Im Jahr 2009 öffneten die Hochschulen der UA Ruhr ihr Lehrangebot für Studierende der jeweils anderen Universitäten und starteten das e-learning-Projekt Ruhr Campus Online.[12]

Insbesondere die fachlichen Veranstaltungen, die in diesem Rahmen von den Fakultäten Informatik, Statistik und Logistik angeboten werden, sind von potentiellem Interesse für Promovierende aus Projekten des SFB 876 mit entsprechendem Schwerpunkt. Alle fachlichen Veranstaltungen können natürlich von Promovierenden besucht werden. Wegen des unterschiedlichen fachlichen Hintergrundes der Promovierenden des SFB 876 kann eine Teilnahme nur freigestellt und empfohlen werden.
Leistung der Hochschule: Die beteiligten Universitäten haben sich sowohl über die Grundausstattung zu materieller und finanzieller Unterstützung verpflichtet, als auch über die Honorierung der für das GK zu erbringenden Lehre. Natürlich stehen Räume und sachliche Grundausstattung für die Promovierenden zur Verfügung. Sollte nach Auslaufen der maximalen Förderungsdauer ein Promotionsverfahren noch nicht abgeschlossen sein, so werden die beteiligten Fakultäten die Abschlussfinanzierung des Projektes sicherstellen.
Förderung der Promotionen: Über die Maßnahmen zur Förderung von Familien mit Kindern ist im Abschnitt 1.5.4 des Antrages berichtet worden. Die Maßnahmen gelten insbesondere auch für Promovierende.
Entlastung der am GK beteiligten Hochschullehrer: Die im Graduiertenkolleg geleistete Lehre, als deren Bestandteil auch die Betreuungsseminare gelten, wird auf das Lehrkontingent angerechnet. Der Leiter des GK wird von der Fakultät Physik von 2 SWS Lehre freigestellt.
Außeruniversitäre Forschungseinrichtungen: Im Rahmen des Teilprojektes TB1 findet eine Kooperation mit dem Leibniz-Institut für Analytische Wissenschaften (ISAS) Dortmund sowie dem ISAS-Spin-Off-Unternehmen der B & S Analytik GmbH statt. Durch diese Kooperation wird den beteiligten Doktorand/inn/en Zugriff auf die Daten zur Atemluftdiagnostik und das vorhandene Fachwissen zu deren Analyse gegeben. Für Doktorand/inn/en aus dem Teilprojekt C4 ist eine Kooperation mit dem An-Institut der TU Dortmund IfADo[13] vorgesehen. In Forschungsfragen aus dem Bereich eingebetteter Systeme erfolgen Kooperationen mit dem ICD (Informatik Centrum Dortmund e.V.[14]) und dem IMEC[15], Leuven, Belgien.

3.7 Zwischenbilanz und weitere Planung

Im Dezember 2013 wurde das Integrierte Graduiertenkolleg von den Promovierenden evaluiert. Die Ergebnisse dieser Evaluation werden im Folgenden zunächst zusammengefasst:

Als besonders positiv wurde von den Promovierenden die interdisziplinäre Zusammenarbeit empfunden, die es einerseits erlaubt, die eigene Arbeit fachlich über den Tellerrand der eigenen Forschung schauend mit Promovierenden anderer Disziplinen zu diskutieren und die andererseits gemeinsame Publikationen verschiedener

[12] http://www.uamr.de/aims/index.htm
[13] http://www.ifado.de
[14] http://www.icd.de
[15] http://www.imec.be

Teilprojekte ermöglicht. Insbesondere positiv gesehen wurde die Interdisziplinarität vor dem Hintergrund der Vereinigung von Expertise in den Bereichen der Datenanalyse und der eingebetteten Systeme.

Auch die Vermittlung der technischen Fähigkeiten in den *Technical Skill*-Seminaren zu R, RapidMiner und LaTeXsowie die technische Kurse auf der Sommerschule wurden ebenso gelobt wie die Vermittlung der *Soft Skills* wie z.B. der Präsentationskurs oder der Kurs zum Buch *Das Arroganz-Prinzip* für weibliche Mitglieder des SFB. Die Sommerschule als Ganzes wurde sehr positiv aufgenommen. Die Vorträge wurden von den Promovierenden als sehr interessant empfunden und auch hier wurde die interdisziplinäre und internationale Ausrichtung und der Austausch mit externen Studierenden gelobt.

Die Möglichkeit, Verantwortung zu übernehmen und die Mitbestimmungsmöglichkeiten der Promovierenden wurden mit Freude angenommen. Dies schließt die Organisation der Workshops ebenso ein wie die Möglichkeit, auf den Workshops und der Sommerschule eigene Lehrveranstaltungen abzuhalten. Auch hier wurde die Tatsache, dass die Veranstaltungen offen sind für Mitglieder aller Arbeitskreise sowie deren Studierende (Diplomand/inn/en, Bachelor/Master-Kandidat/inn/en) und Postdocs, als sehr angenehm empfunden.

Insbesondere positiv beurteilt wurden die *Topical Seminars* wegen der Möglichkeit, internationale Gastwissenschaftler einzuladen, so – vor oder nach dem Gastvortrag und auch z.B. beim gemeinsamen Abendessen – Kontakte zu Forschern von internationalem Format zu knüpfen und vielleicht einen Weg zu einer gemeinsamen Publikation mit diesen Gastwissenschaftlern zu betreten.

Schließlich wird das Modell, im Rahmen der Promotionsvereinbarung eigenverantwortlich über die Teilnahme an Lehrveranstaltungen im GK entscheiden zu können, von den Promovierenden unterstützt und für sehr gut befunden.

Um alle Promovierenden über bestehende Möglichkeiten optimal informieren und beraten zu können und (u. a.) das Gewicht der Promovierenden auch im Hinblick auf von der Universität geplante *Soft Skill*-Kurse besser nutzen zu können, wollen die Promovendensprecher/innen in regelmäßigen Abständen organisatorische Treffen der Promovierenden organisieren. Eine Reihe von Ideen für künftige Arbeiten, z.B. die Einladung von jüngeren Gästen oder die Einladung von Gastwissenschaftlern gemeinsam mit deren Mitarbeitern, werden besprochen. Wie können die Promovenden als Gutachterinnen und Gutachter für internationale Programmkomitees tätig werden? Weiter möchten sie aktiv die Werbeaktivitäten für den SFB in Dortmund für Studierende in der Qualifikationsphase verstärken, auch um ihren Anteil an betreuten Bachelor- und Masterarbeiten zu erhöhen. Die Promovendensprecher/innen haben daher damit begonnen, einen Vorstellungsworkshop zu organisieren, an dem die Promovierenden die in den Teilprojekten möglichen Bachelor- und Masterarbeitsthemen für Studierende vorstellen.

Gespräche mit den Promovenden zeigen, dass das Betreuungskonzept mit seinen Schwerpunkten in der interdisziplinären Kooperation, der wissenschaftlichen Sichtbarkeit und der Eigenverantwortung der Promovierenden gut aufgenommen wurde. Auch aus Sicht der Betreuenden war das GK in der ersten Förderperiode in der Ausbildung einer fachlich heterogenen, aber erfolgreich miteinander forschenden Gruppe von Promovierenden überaus erfolgreich. Nach einer Aufbauphase verfügen wir nun über eine Struktur von jungen Forschenden zwischen dem Bachelor und dem Postdoc, bei Verlängerung des SFB sogar bis zu jungen Teilprojektleitern, innerhalb der Wissen effektiv selbstorganisiert weitergegeben wird.

Das im SFB 876 angebotene strukturierte Promotionsprogramm entspricht mustergültig den Anforderungen der Promotionsrahmenordnung der TU Dortmund. Der Aspekt der Internationalität in Vorträgen und Programmkomitees, der in der Anlaufphase des SFB mit durchweg noch sehr jungen Promovierenden erst langsam relevant geworden ist, soll in der zweiten Förderperiode konsequent in den Vordergrund gestellt werden.

3.8 Ausstattung des Teilprojekts

3.8.1 Bisherige Förderung

Das Teilprojekt wird seit Januar 2011 im Sonderforschungsbereich gefördert

3.8.2 Vorhandene Mittel für die neue Förderperiode

	Nr.	Name, akad. Grad, Dienststellung	Fachzugehörigkeit	Institut der Hochschule	Projektmitarbeit in WS	Kategorie	Finanzierungsquelle
Vorhandenes Personal (MGK)							
Wissenschaftliches Personal	1	Rhode, Wolfgang, Prof. Dr. Dr., Prof	Astroteilchenphysik	Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund	4,0	—	Grundausstattung

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Rhode, Wolfgang**
   Leitung des MGK

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	10.000	0	0	0
Summe der vorhandenen Mittel	10.000	0	0	0
Beantragte Sachmittel	51.400	51.400	51.400	51.400

3.8.3 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Summe:	—	0	—	0	—	0	—	0
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Gäste	27.000		27.000		27.000		27.000	
Summe:	27.000		27.000		27.000		27.000	
Stipendienmittel:	Betrag		Betrag		Betrag		Betrag	
Doktoranden-stipendien	13.200		13.200		13.200		13.200	
Sachkostenzuschuss	1.200		1.200		1.200		1.200	
Summe:	14.400		14.400		14.400		14.400	
Globale Mittel:	Betrag		Betrag		Betrag		Betrag	
Projektspezifischer Workshop	10.000		10.000		10.000		10.000	
Summe:	10.000		10.000		10.000		10.000	
Summe insgesamt:	51.400		51.400		51.400		51.400	

3.8.4 Beantragte Personalmittel

Es werden keine Personalmittel beantragt

3.8.5 Beantragte Sachmittel

Gäste **für 2015 – 2018**

Gäste des Topical Seminars EUR je Jahr 27.000

Die in den Vorlesungszeiten wöchentlich stattfindenden Topical Seminars des Graduiertenkollegs sind eine wichtige Maßnahme für die Vernetzung der Graduierten in der Forschung. Wir gehen von 30 Gästen pro Jahr aus, so dass etwa in jeder Woche der Vorlesungszeit ein Gast bei uns ist. Bei einer Mischung von Inlands-, europäischen und transatlantischen Reisen unserer Gäste setzen wir im Schnitt 900,-EUR pro Person an.

Für das Gästeprogramm im Rahmen des Topical Seminars sind bereits Gäste für das Wintersemester 2014/2015 und daran anschließend eingeplant. Einen Auszug enthält die folgenden Liste; teilweise sind die Gäste bereits angefragt.

Gunnar Rätsch: Memorial Sloan Kettering Cancer Center, USA: Schwerpunkt der Arbeiten ist die Forschung an großen biomedizinischen Datenmengen unter Einsatz maschinellen Lernens.

Dino Pedreschi: University of Pisa, Italien: Forschungsschwerpunkt ist Data Mining in Datenbanken, insbesondere das Zusammenspiel zwischen Datenanalyse und -abfrage.

Fosca Giannotti: CNRS Pisa, Data Mining Group, Italien: Fosca Giannotti leitete ein europäisches Projekt zu *Privacy Preserving Data Mining*. Sie untersuchte insbesondere den Schutz der Privatheit bei zeitlich-räumlichen Daten, z.B. von Mobiltelefonen.

Monika Henzinger: Universität Wien, Österreich: Frau Henzinger forscht im Bereich großer graphischer Netze und der Optimierung durch Auktionsmodelle.

Nicolai Meinshausen: ETH Zürich, Schweiz: Er beschäftigt sich mit Computational Statistics, High-dimensional Data, Regularization, Lasso-type Estimators, Sparsity, und Multiple Testing.

Andrew Gelman: Columbia University, USA: Mitentwickler von Stan und Experte auf dem Gebiet der hierarchischen Bayes Modellierung.

Alexander Lemken: Fraunhofer Institut für Techno- und Wirtschaftsmathematik ITWM, Kaiserslautern: Er arbeitet auf Conditional Random Fields and Graphen.

Osmar Zaiane: University of Alberta, Kanada: Analyse sozialer Netze.

Michael Jordan: University of California, Berkeley, USA: Probabilistische graphische Modelle, Kernel-Methoden.

Jure Leskovic: Stanford University, USA: Graph Mining auf großen Graphen sozialer und informationstechnischer Netze.

Preben Morgensen: Aalborg University, Dänemark: Energieeffizienz in 5G-Mobilfunknetzen.

Jesús Alonso-Zárate: Centre Tecnològic de Telecomunicacions de Catalunya, Spanien: Machine-To-Machine-Kommunikation, insbesondere zur Energieeinsparung durch angepasste Warteschlangenmodelle.

Yung-Hsiang Lu: Purdue University, USA, High-Efficiency, Low-Power System Group: Prof. Lu forscht im Bereich des Ende-zu-Ende-Energiemanagements und damit auch Ressourcenbeschränkungen auf den eingebetteten Systemen und der Kommunikationsinfrastruktur.

Kai Groenbaek: University of Aarhus, Dänemark: Verteilte Messungen und Sensoren für die makroskopische Analyse zur Verbesserung von Energieverbräuchen.

Achim Streit: Karlsruhe Institute of Technology, Steinbuch Centre for Computing: Speicherreduzierende verteilte Berechnungen und Map-/Reduce-Verfahren.

Marco Sozzi: University of Pisa, Italien: Berechnungen auf Graphikprozessoren bei hochvolumigen Datenmengen der Physik.

Gabor Melli: VigLink, San Francisco, USA: Analyse großer Datenmengen für die Optimierung von Anzeigen.

Francesco Bonchi: Yahoo Research, Barcelona, Spanien, ist bekannt für seine Arbeiten zu *Privacy Preserving Data Mining*.

Leon van der Torre: University of Luxembourg, Luxemburg: Normatives Reasoning und Vertraulichkeit.

Luigi Sauro: University of Naples, Italien: Forscht im Bereich der Komplexität zu Garantien der Vertraulichkeit.

Nataliia Bielova: Inria Rennes, Frankreich: Analyse zur Computersicherheit, insbesondere zur Überwachung während der Laufzeit von Sitzungen in Webbrowsern.

3.8.6 Beantragte Stipendienmittel

Kurzzeitstipendien **für 2015 – 2018**

Stipendienmittel EUR je Jahr 13.200

Beantragt werden Stipendienmittel für Promovierende. Zielgruppe sind insbesondere ausländische Absolventen, denen mittels des Stipendiums die Möglichkeit zum Einstieg in die Themen des SFB gegeben werden soll. Nach Erwerb der entsprechenden Qualifikation soll bei Verfügbarkeit die Option bestehen, eine Stelle als Promovend auf einer Stelle der Ergänzungs- oder Grundausstattung zu erhalten. Beantragt werden Mittel für zwei Stipendien je Jahr mit einer Laufzeit von sechs Monaten. Die Höhe soll monatlich jeweils 1.100 EUR pro Stipendium betragen.

Sachkostenzuschuss EUR je Jahr 1.200

Sachkostenzuschuss in Höhe von 100 EUR pro Stipendium je Monat.

3.8.7 Beantragte Globale Mittel

Projektspezifischer Workshop **für 2015 – 2018**

Workshop EUR je Jahr 10.000

Für das erste und dritte Jahr ist ein Workshop mit auswärtiger Beteiligung geplant, den der SFB ausrichtet, für das zweite und vierte Jahr eine internationale Sommerschule. Dies sind hervorragende Möglichkeiten für alle Mitglieder, den SFB bekannt zu machen, und für die Promovenden, die Organisation einer Tagung selbst mitzuerleben und auch mitzugestalten. Die Mittel sind ursprünglich anhand von zwei Sommerschulen geschätzt, die die Sprecherin durchgeführt hat, was sich bei der ersten Sommerschule und den Workshops des SFB bestätigt hat.

3.1 Allgemeine Angaben zum Teilprojekt Z

3.1.1 Titel:

Zentrale Aufgaben des Sonderforschungsbereichs

3.1.2 Projektleitung

Prof. Dr. Morik, Katharina, 14.10.1954

LS 8, Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Straße 12
44227 Dortmund

Tel.: 0231-755-5100
Fax : 0231-755-5105
E-Mail: `katharina.morik@cs.tu-dortmund.de`
Ist die Stelle des Leiters/der Leiterin des Projekts befristet?

(x) nein ( ) ja, befristet bis zum

3.2 Dokumentation der zentralen Aktivitäten des Sonderforschungsbereichs

Die folgenden Abschnitte und Listen dokumentieren die Verwendung der von der DFG zentral bewilligten Mittel.

Verwendung pauschaler Mittel

Bereits zur Antragsstellung der ersten Phase waren die Neuberufungen der Professuren von Peter Marwedel und Joachim Biskup und die Integration der Nachfolger in den Sonderforschungsbereich eingeplant. Der größte Anteil der eingesetzten pauschalen Mittel entfällt dementsprechend auf die *Anschubfinanzierung und Integration* der neuen Projektleiter vor Beginn der zweiten Phase.

Neu berufen wurden Jens Teubner (Projekte A2, C5) und Jian-Jia Chen (Projekt B2). Die Ausstattung sollte dem Umfang entsprechen, wie er für die übrigen Projektleiter für die erste Phase genehmigt wurde. Aus den pauschalen Mitteln wurden jeweils Mittel für einen wissenschaftlichen Mitarbeiter der Ergänzungsausstattung, sowie Mittel für eine SHK oder Spezialhardware (FPGA-Entwicklungsboards) zur Verfügung gestellt. Weitere Sachmittel, z.B. für die Einrichtung von Labors mit Servern, wurden aus Mitteln das Haushalts der Fakultät oder dem SFB für die Grundausstattung zur Verfügung gestellten Mitteln finanziert. Die Projektleiter bringen ebenso wie bei den anderen Projekten des SFB auch üblich in gleichem Umfang wie aus pauschalen Mitteln finanziert Mitarbeiter ihrer Grundausstattung ein.

Neben den geplanten Berufungen wurde durch die Berufung von Sven Rahmann an die Universität Duisburg-Essen die Nachfolge an Kristian Kersting (Projekte A6, B4) mit der Professur für Data Mining vergeben. Herrn Kersting steht ebenfalls ein Mitarbeiter sowie durch Grundausstattung finanzierte Rechnerhardware zur Verfügung.

Michael ten Hompel wird in der zweite Phase das Projekt A4 verstärken. Die Integration bereits während der laufenden ersten Phase erfordert den Aufbau eines Versuchsfelds bestehend aus 50 Hardwareplattformen und 300 Sensorknoten. Die Anschubfinanzierung betrifft auch für dieses Vorhaben einen wissenschaftlichen Mitarbeiter, eine halbe SHK-Stelle und die projektspezifischen Hardwarekosten. Ein Mitarbeiter der Grundausstattung des Lehrstuhls sowie die notwendige Ausrüstung für Fertigung und Betrieb der Hardware werden von Herrn ten Hompel mit eingebracht.

Zusätzlich zu den Anschubfinanzierungen neuer Projektleiter und Vorhaben, wurden pauschale Mittel ebenfalls für *Öffentlichkeitsarbeit* (Details siehe unten) und *Publikationskosten* von Open Access-Publikationen verwandt. In der Regel wurden Publikationskosten allerdings zunächst durch eingesparte Projektmittel gedeckt, in Ausnahmefällen wie z.B. bei [A4/B4/192] dann durch die pauschalen Mittel.

Letztendlich wurden noch für den außerplanmäßigen Einsatz 10 Smartphones als Messplattform für Projekt A1 aus pauschalen Mitteln finanziert. Die Smartphones werden an wechselnde Mitglieder des Sonderforschungsbereichs ausgegeben und erfassen sämtliche Parameter der mobilen Nutzung. Die gewonnenen Daten stehen Projekt A1 für die Analyse zur Verfügung, was auch für Veröffentlichungen (siehe [A1/56]) genutzt wird. Die Smartphones waren ebenfalls wesentlicher Bestandteil der Messkampagne im Rahmen der Sommerschule 2012, mit denen die Teilnehmer sich sowohl an den Messungen beteiligen als auch die Datenanalyse von Mobildaten trainieren konnten. Es ist geplant, die Smartphones ebenfalls wieder in der Sommerschule 2014 einzusetzen.

Mittel für Chancengleichheitsmaßnahmen

Mentoring^3: Um die Karriere von Wissenschaftlerinnen zu fördern, existiert im Verbund Universitätsallianz Ruhr des Programm *mentoring*^{3}. Für den Sonderforschungsbereich konnte für 6 Mentees eine eigene Gruppe eingerichtet werden. Das Programm besteht aus den Komponenten des Einzelmentorings mittels persönlicher Mentoren, Seminaren für die Vermittlung von überfachlichen Kompetenzen, Netzwerkbildung und Workshops. Das Programm läuft noch über die erste Phase des SFB hinausgehend bis Anfang 2015.

Durch Mittel für Chancengleichheitsmaßnahmen wurde die Gruppe personell über eine wissenschaftliche Hilfskraft für die Betreuung unterstützt. Ebenfalls eingeladen wurden Gäste zu den Themen Präsentationstechniken (technisch) und überzeugende Präsentation als Wissenschaftlerin (Körper, Sprache und Stimme).

Kinderbetreuung: Initiiert durch den Dortmunder Sonderforschungsbereich 823 wurde eine U3-Kinderbetreuungseinrichtung für die Kinder von Mitarbeiterinnen und Mitarbeiter DFG-geförderter Projekte geschaffen.

Aktuell befinden sich zwei Kinder von Mitarbeitern des SFB 876 in der Betreuung. Personalkosten werden aus Gleichstellungsmitteln, die übrigen Sachkosten (Nebenkosten, Mieten, Versicherungen, ca. 70% der Gesamtkosten) aus Grundausstattungsmitteln des SFB finanziert.

Arroganztraining: Trotz des überdurchschnittlichen Frauenanteils innerhalb des SFB 876 sehen sich die Mitarbeiterinnen in ein weitestgehend männliches Arbeitsumfeld integriert. Das Arroganztraining schult den Umgang mit Kollegen und Vorgesetzten mittels praktischer Übungen und Sparringspartnern. Die Teilnahme am Training stand sowohl den Promovendinnen des Graduiertenkollegs als auch den Teilprojektleiterinnen offen.

Zentrale Veranstaltungen

Zentral koordiniert und mit SFB-Mitteln unterstützt werden die jährlich stattfindenden Workshops der Graduierten, in denen über die Fortschritte im Projekt und der eigenen Arbeiten berichtet wird.

Workshop 19.10.-21.10.2011: Vorträge der Graduierten aus den Projekten, sowie Kurse und Tutorials von Projektleitern, Graduierten und Gästen.

Streaming Workshop Projekt A2 23.07.-27.07.2012: Explizit mit der Theorie und Anwendung von Datenstromverfahren hat sich der von Christian Sohler, Projekt A2, organisierte einwöchige Workshop befasst. Dem internationalen Publikum gegenüber hat sich der SFB mit eigenen Vorträgen präsentiert. Im Gegenzug konnten die Mitglieder des SFB vom Fachwissen der externen Wissenschaftler profitieren. Die Liste eingeladener Gäste befindet sich im unten im entsprechenden Abschnitt auf Seite 382.

Workshop 03.09.2012: Berichte aus den Projekten durch die Graduierten.

Sommerschule 04.09.-07.09.2012: Die Sommerschule für *Resource-aware Machine Learning* hat mit Themen von Streammining über Modelselektion und Speicherhierarchien bis zu praktischen Übungen des Data Minings ein internationales Publikum nach Dortmund gebracht. Externe Lehrende der Sommerschule siehe unten bei der Auflistung eingeladener Gäste.

Workshop 10./11.07.2013: Ausblick der Projektleiter auf die kommenden Arbeitspakete im Projekt und Vorstellung von Ideen für Erweiterungen des Sonderforschungsbereichs.

Workshop 16./17.10.2013: Fortschrittsberichte der Graduierten.

Workshop 08.09.2014: Präsentation der Ergebnisse der Graduierten und Ausblick auf die geplante zweite Phase.

Sommerschule 29.09.-02.10.2014: 2. Sommerschule des Graduiertenkollegs mit Kursen zu Core Sets, Faktorisierung gigantischer Matrizen und praktischen Übungen zum Detektieren von Partikeln mittels eingebetteter Systeme.

Maßnahmen zur Öffentlichkeitsarbeit

Die Themen des Sonderforschungsbereichs einem breiteren Publikum näher zu bringen ist eine wichtige Ergänzung der Forschungsaufgaben. Presseartikel der Projektleiterinnen und Projektleiter gehören dazu genauso wie die Teilnahme an Veranstaltungen der Universität für die Öffentlichkeit (Tag der offenen Tür, 40-Jahrfeier der Fakultät für Informatik). Diese sind näher in Abschnitt 1.3.6, Erkenntnistransfer und Öffentlichkeitsarbeit, auf Seite 38 erläutert.

Die folgenden zentralen Maßnahmen wurden dabei zusätzlich durch den SFB unter Einsatz pauschaler Mittel gefördert:

Leporellos mit Projektvorstellungen: Die Themen einzelner Projekte und die Auswirkungen auf die Gesellschaft anschaulich darzustellen war die Aufgabe bei der Erstellung von Leporellos zu einzelnen Projekte. Prototypisch haben sich die Projekte A4 (ChocoLaTE fondue-to-go, Wie sich die Akkulaufzeit von LTE-Smartphones verlängern lässt), B2 (Ein Virenscanner für biologische Viren) und C1/C3 gemeinsam (Was hat Krebserkennung mit Astrophysik zu tun?) dargestellt.

Broschüre zum SFB: Über alle Projekte ergibt der Sonderforschungsbereich ein Gesamtbild. Unter dem anschaulichen Motto *Große Daten – Kleine Geräte* erfolgte aufbauend auf den Leporellos einzelner Projekte die Erstellung einer Broschüre mit dem Konzept des SFB insgesamt. Über die Themen großer Datenmengen, Cyber-Physical Systems und der Integration von Informationsverarbeitung und Gegenständen werden die Besonderheiten des SFB dargestellt. Zusätzlich hat jedes Projekt die Möglichkeit zur Einzeldarstellung des jeweiligen Fokus.

Wissenswerte 2013: Die jeden November stattfindende Wissenswerte ist die größte Konferenz für Wissenschaftsjournalisten in Deutschland. Hier informieren sich die Journalisten über zukünftige Trends in Wissenschaft und Forschung. Der Sonderforschungsbereich hat sich 2013 mit einem eigenen Stand dargestellt, um ein Bewusstsein für über die Verarbeitung großer Datenmengen hinausgehende Informationsgewinnung zu schaffen. Während die Analyse riesiger Datenmengen — nicht nur positiv — als Thema in der Öffentlichkeit angekommen zu sein scheint, fehlt oft noch das Wissen über die zusätzlichen Herausforderungen. Beschränkte Ressourcen, die durch Geschwindigkeit der Daten, kleine Geräte oder Kommunikation zusätzlich limitiert werden, sind ein wichtiger Treiber für die Forschung in der Informationsgewinnung, was mit den am Stand vertretenen Projekten (A4, B2, C1, C3) vermittelt wurde.

Übersicht eingeladener Gäste

Die folgenden Listen benennen die im Sonderforschungsbereich eingeladenen Gäste, gegliedert nach der Art ihres Besuchs. Die inhaltliche Betreuung der Gäste erfolgte durch die Mitarbeiterinnen und Mitarbeiter des einladenden Projekts mit organisatorischer Unterstützung durch die Mitarbeiterinnen und Mitarbeiter des Verwaltungsprojekts. Das Geschlecht der Gäste ist jeweils mit M/W abgekürzt angegeben.

Forschungsaufenthalte von Gästen: Einige Gäste wurden bereits mit spezifischen Aufgaben für die Zusammenarbeit für einen längeren Aufenthalt an den Sonderforschungsbereich eingeladen. Zusätzlich zur Arbeit in den jeweiligen Projekten haben die Gäste auch öffentlich im Topical Seminar des Graduiertenkollegs vorgetragen. Zusätzlich zur Aufenthaltsdauer ist bei den Gästen entsprechend das Datum des Vortrags angegeben.

Kamalika Das: W, NASA Ames Research, USA, 14.07.2011: Regression Algorithms for Large Scale Earth Science Data. Aufenthaltsdauer 3 Wochen.

Kanishka Bhaduri: M, NASA Ames Research, USA, 14.07.2011: Strategies for Scaling Data Mining Algorithms. Aufenthaltsdauer 3 Wochen. Es gab eine intensive Zusammenarbeit mit Kanishka Bhaduri und Kamalika Das zum Themenbereich der verteilten Datenanalyse. Im Rahmen von Projekt B3 wurde gemeinsam ein kommunikationseffizienter Algorithmus zur verteilten Anomalie-Detektion entwickelt und im weiteren Verlauf nach ihrem Besuch veröffentlicht (siehe [B3/394] im Literaturverzeichnis von B3 auf Seite 231). Dabei sind nicht die Beispiele zum Lernen über mehrere Knoten im Netzwerk verteilt, sondern die sie beschreibenden Attribute.

Yurii Kuritsyn: M, Russian Academy of Sciences, Russland, 24.11.2011: Examining of possible approaches to the signal quantification for PAMONO-method. Aufenthaltsdauer 2 Monate. Die PAMONO-Methode ermöglicht die Visualisierung der einzelnen Nanoteilchen, die sich an die Sensorfläche binden. Die gebundenen Teilchen produzieren Spots an der Abbildung der Sensorfläche. Dr. Kuritsyn hat untersucht wie sich die Form des Spots an unterschiedlichen Orten der Sensorfläche modifiziert. Die Kenntnisse sind notwendig für eine sichere Erkennung der Teilchen. Es wurde gezeigt, dass man die Erkennungsgrade deutlich steigern kann, indem man eine Faltung des Bildes mit der zu erwartenden Form des Spots einsetzt. Die schnelle Fourier-Transformation wurde genutzt um die Erkennung deutlich zu beschleunigen und den Rauschabstand zu verbessern.

Liang Cheng: M, Lehigh University, USA, 12.04.2012: Network Design and In-network Data Analysis for Energy-efficient Wireless Sensor Networks of Bridge-Monitoring Applications. Aufenthaltsdauer 4 Monate. Während seines Forschungsaufenthaltes an der TU Dortmund arbeitete Prof. Liang Cheng intensiv an der Modellierung der Leistungsaufnahme von LTE Endgeräten unter Berücksichtigung verschiedener System- und Kontextparameter mit. Hierbei entstand neben einem Posterbeitrag für die renommierte IEEE INFOCOM auch der Beitrag *CoPoMo: A Context-Aware Power Consumption Model for LTE User Equipment* in den Transactions on Emerging Telecommunications Technologies (ETT), siehe [A4/B4/192] im Literaturverzeichnis von A4 auf Seite 126. Die gemeinsamen Arbeiten zeigen, dass neben der Parameterisierung des LTE Netzes auch unbeeinflussbare Kontextparameter wie die Zellumgebung und das Applikationsverhalten einen signifikanten Einfluss auf die zu erwartende Batterielaufzeit von LTE Endgeräten haben. Darüber hinaus lässt sich das Modell im Kontext drahtloser Sensornetze dazu verwenden, die für eine gewünschte Laufzeit benötigte Batterie zu dimensionieren.

Gäste der Sommerschule: Die alle zwei Jahre stattfindenden Sommerschulen werden unterstützt durch die Einladung externer Lehrender mit in den Kontext des SFB passenden Vorlesungen. Angegeben ist bei den Lehrenden jeweils das Thema der Vorlesung und das Datum des ersten Vorlesungstages.

Michèle Sébag: W, Université Paris-Sud, Frankreich, 04.09.2012: Introduction to machine learning

João Gama: M, University of Porto, Portugal, 05.09.2012: Data Mining from Ubiquitous Data Streams

Jan Madsen: M, Technical University of Denmark, Dänemark, 06.09.2012: Towards Self-Powered Systems

Gäste des Topical Seminars des integrierten Graduiertenkollegs: Ein wichtiger Bestandteil des integrierten Graduiertenkollegs ist die Einladung von Gästen durch die Promovendinnen und Promovenden. Die folgende Liste gibt die im Rahmen des Seminars eingeladenen Gäste in zeitlich sortierter Reihenfolge an. Bei jedem Gast angegeben ist das Datum des Vortrags. Die Aufenthaltsdauern je Gast lagen in der Regel bei 2 Tagen, in denen die einladenden und andere interessierte Projektmitarbeiter Zeit für Fachgespräche nutzen konnten.

Johannes Gehrke: M, Cornell University, USA, 20.01.2011: Monitoring the Data Tsunami

Luca Benini: M, University of Bologna, Italien, 20.01.2011: Model predictive control and self-learning of thermal models for thermal control for multi-core platforms

Stefan Manegold: M, CWI Amsterdam, Niederlande, 17.02.2011: MonetDB: Open-source Columnar Database Technology Beyond Textbooks

Elke Rundensteiner: W, Worcester Polytechnic Institute, USA, 07.04.2011: Multiple Query Optimization for Complex Pattern Mining Requests over Data Streams

Henrik Blunck: M, University of Aarhus, Dänemark, 05.05.2011: Pervasive and Energy-Efficient Positioning and Trajectory Tracking for Mobile Devices

Frank Fitzek: M, University of Aalborg, Dänemark, 12.05.2011: Network coding for resource-efficient operation of mobile clouds

Preeti Ranjan Panda: M, Indian Institute of Technology, Delhi, Indien, 30.06.2011: Energy Efficiency in Graphics Rendering

Carlo Alberto Fabbretti: M, University of Pisa, Italien, 06.07.2011: Equivalence theorem about classification algorithms encoding

Gerhard Brewka: M, Universität Leipzig, 07.07.2011: Multi-Context Systems: A Flexible Approach for Integrating Heterogeneous Knowledge Sources

Corrado Loglisci: M, University of Bari Aldo Moro, Italien, 10.08.2011: A Data Mining Approach to the problem of Temporal Projection on Longitudinal Data

Kristian Kersting: M, Fraunhofer IAIS, 20.10.2011: Can machines feed a hungry world? Exploring massive data using simplex distributions

Tajana Rosing: W, University of California, USA, 25.10.2011: Energy efficient computing at the edge: from sensors to servers

Jan Baumbach: M, Max Planck Institut, 27.10.2011: Key Pathway Mining - Network-based data mining in biological data

Bart Goethals: M, University of Antwerpen, Belgien, 03.11.2011: Cartification: from Similarities to Itemset Frequencies

Fadi Kurdahi: M, University of California, USA, 04.11.2011: The Cross-Layer Multi-Dimensional Design Space of Power, Reliability, Temperature and Voltage in Highly Scaled Geometries

Michael Haupt: M, Oracle Labs, 24.11.2011: The Maxine Research Virtual Machine

Simon Peter: M, ETH Zürich, Schweiz, 01.12.2011: Fay: Extensible Distributed Software Tracing from OS Kernels to Clusters

Piero Bonatti: M, University of Naples, Italien, 08.12.2011: Confidentiality policies on the semantic web: Logic programming vs. Description logics

Alexander Bertram: M, Bedatadriven, Niederlande, 19.01.2012: Renjin: JVM-based Interpreter for the R Language for Statistical Computing

Andreas Krause: M, ETH Zürich, Schweiz, 05.04.2012: Optimizing Sensing: Theory and Applications

Axel Mosig: M, Ruhr-Universität-Bochum, 26.04.2012: Algorithmic Tools for Spectral Image Annotation and Registration

Stefan Funke: M, Universität Stuttgart, 10.05.2012: Route Planning: Energy-efficient, Constraint-respecting, and fast!

Stephan Holzer: M, ETH Zürich, Schweiz, 15.05.2012: New Lower Bounds and Algorithms in Distributed Computing

Ulf Brefeld: M, Technische Universität Darmstadt, 24.05.2012: Modeling User Navigation on the Web

Hans van Ditmarsch: M, University of Sevilla, 31.05.2012: The Ditmarsch Tale of Wonders - the dynamics of lying

Yücel Saygin: M, Sabanci Universit, Türkei, 14.06.2012: Privacy Preserving Publishing of Spatio-temporal Data Sets

T.S. Jayram: M, IBM Research, USA, 19.07.2012: The Johnson-Lindenstrauss Transform and Applications to Dimensionality Reduction

Lena Wiese: W, Universität Hildesheim, 15.11.2012: Deductive Generalization for Flexible Query Answering

Alexander Pretschner: M, Technische Universität München, 06.12.2012: Towards Quantitative Distributed Data Usage Control

Jens Teubner: M, ETH Zürich, Schweiz, 24.01.2013: Database Joins on Modern Hardware

Kresimir Mihic: M, Oracle Labs, USA, 14.02.2013: Solutions to optimization problems in resource constrained systems

Wouter Duivesteijn: M, Leiden University, Niederlande, 14.03.2013: Exceptional Model Mining - Identifying Deviations in Data

Hubert Rehborn: M, Daimler AG, 19.03.2013: Anwendungen der Drei-Phasen Verkehrstheorie zur intelligenten Verkehrssteuerung

Matthijs van Leeuwen: M, KU Leuven, Belgien, 28.03.2013: Patterns that Matter – MDL for Pattern Mining

Jan Koster: M, AMC Amsterdam, Niederlande, 11.04.2013: R2: Biologist friendly web-based genomics analysis & visualization platform

Pinar Tüzön: W, Ecole polytechnique fédérale de Lausanne, Schweiz, 18.04.2013: Transactions chasing Instruction Locality on multicores

Stefan Wrobel: M, Fraunhofer IAIS, 02.05.2013: Algorithms and Systems for Analyzing Graph-Structured Data

Stefan Kramer: M, Universität Mainz, 06.06.2013: New Algorithms for Graphs and Small Molecules

Michael Richter: M, University of Calgary, Kanada, 20.06.2013: Case-Based Reasoning: Was ist es und wie kann man es gebrauchen?

Frank Padberg: M, Karlsruhe Institute of Technology, 10.10.2013: Eine Experimentierplattform für die automatische Parallelisierung von R-Programmen

Albert Bifet: M, Yahoo Research, Spanien, 24.10.2013: Mining Big Data in Real Time

Ulrich Mansmann: M, LMU München, 13.11.2013: Indirect Comparison of Interaction Graphs

Osvaldo Anacleto: M, University of Edinburgh, Großbritannien, 14.11.2013: Using dynamic chain graphs to model high-dimensional time series: an application to real-time traffic flow forecasting

Arno Siebes: M, Utrecht University, Niederlande, 05.12.2013: MDL for Pattern Mining

Eduardo Feo: M, 16.01.2014, Dalle Molle Institute for Artificial Intelligence, Schweiz: Supervised learning of link quality estimates in wireless networks

Joaquin Vanschoren: M, 30.01.2014, Eindhoven University of Technology: OpenML: Open science in machine learning

Eike Kiltz: M, 20.02.2014, Ruhruniversität Bochum: Efficient Cryptography with Provable Security

Francois Schnitzler: M, 15.05.2014, Technion – Israel Institute of Technology: Gauss-Markov modeling and online crowdsensing for spatio-temporal processes

Andreas Hotho: M, 12.06.2014, Universität Würzburg: From Web 2.0 to the Ubiquitous Web

Gäste des Streamingworkshops: Der durch das Projekt A2 organisierte Workshop on Data Streams hat die Möglichkeit geboten, weitere international renommierte Forscher gebündelt nach Dortmund einzuladen. Die Liste gibt bei jedem Gast auch Datum und Titel des Vortrags im Workshop an, wobei in der Regel die Gäste die gesamte Woche am Workshop teilgenommen haben.

Sudipto Guha: M, University of Pennsylvania, USA, 23.07.2012: Streaming certificates and near linear time algorithms

Christiane Lammersen: W, Simon Fraser University, Kanada, 23.07.2012: $(1+\varepsilon)$-Approximation for facility location in data streams

Dan Feldman: M, Massachusetts Institute of Technology, USA, 23.07.2012: Google your life: From big GPS data to searchable diary

Andrew McGregor: M, University of Massachusetts, USA, 23.07.2012: Analyzing graphs via linear sketching

Rina Panigrahy: M, Microsoft Research, USA, 23.07.2012: Efficient distance measures for social networks

David Woodruff: M, IBM Research, USA, 23.07.2012: Low rank approximation and regression in input sparsity time

Ely Porat: M, Bar-Ilan University, Israel, 24.07.2012: Homomorphic signature

Anna Gilbert: W, University of Michigan, USA, 24.07.2012: Sublinear fourier sampling off the grid

Martin Strauss: M, University of Michigan, USA, 24.07.2012: Sublinear time, measurement-optimal sparse recovery for all

Amit Chakrabarti: M, Dartmouth College, USA, 25.07.2012: Why data stream researchers should learn Talagrand's inequality

Mert Saglam: M, University of Washington, USA, 25.07.2012: L_p Sampling from streams

Mikhail Kapralov: M, Stanford University, USA, 25.07.2012: On the communication complexity of maximum bipartite matching

Xiaoming Sun: M, Tsinghua University, China, 25.07.2012: Space bounded communication complexity

Krzysztof Onak: M, IBM Research, USA, 26.07.2012: Lower bounds for shortest paths and matchings

Jelani Nelson: M, Harvard School of Engineering and Applied Sciences, USA, 26.07.2012: On deterministic sketching and streaming for sparse recovery and norm estimation

Suresh Venkatasubramanian: M, University of Utah, USA, 26.07.2012: Distributed protocols for classifcation and optimization

Hossein Jowhari: M, Simon Fraser University, Kanada, 26.07.2012: Fast protocols for edit distance through locally consistent parsing

3.3 Ausstattung des Teilprojekts

3.3.1 Vorhandene Mittel für die neue Förderperiode

	Nr.	Name, akad. Grad, Dienststellung	Fachzugehörigkeit	Institut der Hochschule	Projektmitarbeit in WS	Kategorie	Finanzierungsquelle
Vorhandenes Personal (Z)							
Wissenschaftliches Personal	1	Morik, Katharina, Prof. Dr., Prof	Data Mining/Maschinelles Lernen	LS 8, Fak. Inf, TU Do	4,0	—	Grundausstattung
Nichtwissenschaftliches Personal	2	Greve, Andreas, Techn. Ang.	—	LS 8, Fak. Inf, TU Do	9,95	—	Grundausstattung

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (vorhandene Mittel):

1. **Morik, Katharina**
   Sprecherin des SFB 876, der die Gesamtorganisation und die Vertretung des Sonderforschungsbereichs nach außen obliegt. Sie sorgt für die SFB-interne Kooperation und initiiert internationale und nationale Kontakte.

2. **Greve, Andreas**
   Unterstützung der technischen Infrastruktur zum Datenaustausch sowie bei der Installation von Anwendungsprogrammen.

	2015	2016	2017	2018
TU Dortmund: vorhandene Mittel aus Haushalt	3.000	3.000	3.000	3.000
Summe der vorhandenen Mittel	3.000	3.000	3.000	3.000
Beantragte Sachmittel	93.500	93.500	93.500	93.500

3.3.2 Beantragte Mittel

	2015		2016		2017		2018	
Personalmittel:	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag	Anz.	Betrag
Koordination, 100%	1	75.600	1	75.600	1	75.600	1	75.600
Koordination, 50%	1	21.900	1	21.900	1	21.900	1	21.900
Hilfskräfte (für alle Projekte)	18.5	166.500	18.5	166.500	18.5	166.500	18.5	166.500
Hilfskräfte (Vergeben von MGK)	3	27.000	3	27.000	3	27.000	3	27.000
Vertretung	0.5	31.800	0.5	31.800	0.5	31.800	0.5	31.800
Summe:	—	322.800	—	322.800	—	322.800	—	322.800
Sachmittel:	Betrag		Betrag		Betrag		Betrag	
Gäste	6.000		6.000		6.000		6.000	
Reisen	87.500		87.500		87.500		87.500	
Summe:	93.500		93.500		93.500		93.500	
Globale Mittel	Betrag		Betrag		Betrag		Betrag	
Koordinierung	7.200		7.200		7.200		7.200	
Chancen-gleichheits-maßnahmen	30.000		30.000		30.000		30.000	
Pauschale Mittel	30.000		30.000		30.000		30.000	
Summe:	67.200		67.200		67.200		67.200	
Summe insgesamt:	483.500		483.500		483.500		483.500	

3.3.3 Beantragte Personalmittel

Aufgaben der Mitarbeiterinnen und Mitarbeiter für die beantragte Förderperiode (beantragte Mittel):

1. **Michaelis, Stefan**
   Geschäftsführer, der die Kommunikation zwischen den Teilprojekten und dem integrierten Graduiertenkolleg sowie die Koordination der wissenschaftlichen Aktivitäten unterstützt. Er achtet darauf, dass Berichte termingerecht abgegeben werden, organisiert die Lehrveranstaltungen und Seminare im Graduiertenkolleg und überwacht die Internetpräsenz. Insbesondere betreut er die informationstechnischen Einrichtungen zur Datenhaltung, Bibliographien, Softwarerepositories und internen Kollaborationswerkzeuge.
   Er unterstützt die Arbeitskreise mit den Bezügen zu Teilprojekten und dem Graduiertenkolleg und bereitet wissenschaftliche Tagungen mit vor.
   Er ist verantwortlich für das Controlling der Finanzen und erstellt mit den Kolleginnen und Kollegen der Administration die Berichte und Nachweise.
   Herr Michaelis hat bereits vor der ersten Phase des SFB im Themenumfeld promoviert und bringt seine Erfahrung aus vorherigen Tätigkeiten in der Industrie mit ein. Dadurch wurde Herr Michaelis auf der Ebene eines Gruppenleiters eingestuft.

2. Klingspor, Sabine, Dieckmann, Anja
Die Finanzverwaltung (Sach- und Personalmittel, Buchung von Rechnungen, Erstellung von Mittelanforderungenund Verwendungsnachweisen) und die allgemeine Administration (Durchführung der Organisation von Kolloquien, Workshops und Sommerschulen bezüglich der Buchung von Räumen, der Abrechnung mit Gästen, Schriftverkehr per eMail und direkt im Internet, überwiegend in englischer Sprache) wird von den beiden Sekretärinnen erledigt, die schon in mehreren Teilprojekten im Rahmen der Grundausstattung mitarbeiten. Ihre beiden halben Stellen werden in der GA auf E8 angehoben und sie teilen sich die halbe EA-Stelle.

3. Hilfskräfte (für alle Projekte, MGK und Z),
Studentische Hilfskräfte unterstützen die praktischen Arbeiten auch in den Projekten, z.B in der programmiertechnischen Umsetzung der Methoden. Gleichzeitig hilft eine Tätigkeit als SHK bei der Qualifizierung der Studierenden.
Dazu werden hier je Projekt und Projektleiterin/Projektleiter (ausgenommen der Anwendungspartner B&S Analytik und die Hochschule Reutlingen in Projekt TB1) Mittel für eine SHK mit 7 Wochenstunden entsprechend einer halben SHK-Stelle beantragt. Zusätzlich erhalten die Projekte Z und MGK jeweils Mittel für administrative Aufgaben in Höhe von 24 Wochenstunden, entsprechend 1,5 SHK-Stellen.

4. Studentische Hilfskräfte (zu vergeben durch MGK),
Um Projekte zu stärken, die besonderes Potenzial für die frühe Einbindung von Studierenden als Nachwuchs haben, werden drei volle SHK-Stellen mit 17 Wochenstunden beantragt. Diese werden jährlich auf Antrag und in Abstimmung mit den Promovendensprechern des Graduiertenkollegs an die Projekte vergeben.

5. Vertretungskosten,
Eine Teilprojektleiterin oder ein Teilprojektleiter kann beim Vorstand ein Forschungssemester beantragen, wobei die Genehmigung der Fakultät, ein konkreter Arbeitsplan und eine Liste möglicher Vertretungen vorzulegen ist. Die Vertretung des Teilprojektleiters während seines Forschungssemester kann durch einen auswärtigen Kollegen, eine auswärtige Kollegin wahrgenommen werden, so dass die Arbeit im SFB bereichert wird. Sie kann auch durch Postdoktoranden/Postdoktorandinnen, die auf diesem Wege ihre Qualifikation für eine Professur nachweisen können, wahrgenommen werden. Damit fördern wir den wissenschaftlichen Nachwuchs und erweitern die Perspektiven nach der Promotion. Eine Vertretung für 1 Semester pro Jahr wird nach E14 bezahlt.

3.3.4 Beantragte Sachmittel

Gäste **für 2015 – 2018**

Gastwissenschaftler EUR je Jahr 6.000

Die Möglichkeit, dass Kollegen ihr Sabbatical, in dem sie kein Gehalt von ihrer Universität erhalten, am SFB verbringen, soll zusätzlich zum Topical Seminar ermöglicht werden. Für einen längeren Aufenthalt eines amerikanischen Kollegen während seines Forschungssemester gehen wir von 4.000,- EUR pro Monat aus.

Geplante Gäste für einen längeren Forschungsaufenthalt im SFB

Lui Sha: Als erster Gast für die beantragte zweite Phase des Sonderforschungsbereichs im Frühjahr 2015 ist Lui Sha bereits angefragt.

Er ist Inhaber des Donald B. Gillis-Lehrstuhls an der Fakultät für Informatik der University of Illinois, Urbana-Champaign (UIUC). Lui Sha ist bekannt durch seine Forschungsarbeiten im Bereich Echtzeit- und cyberphysikalische Systeme. Die aktuellen Interessen liegen auf Multicore-Systemen und der sicheren Konfiguration medizinischer Geräte. Seine Hauptzusammenarbeit mit dem SFB soll in Projekt B2 stattfinden.

Claire Cardie: Ebenfalls angefragt für einen längeren Forschungsaufenthalt und Zusammenarbeit mit Katharina Morik wurde Claire Cardie von der Cornell University, USA. Ihre Interessen in der Forschung liegen im Bereich „Natural Language Processing“ mit dem Schwerpunkt auf der Informationsextraktion in Onlinetexten. Ihre Arbeiten zu Datenströmen in sozialen Netzen können die Methoden des SFB hervorragend ergänzen.

Reisen **für 2015 – 2018**

Reisen EUR je Jahr 87.500

Die Teilnahme an Tagungen auch im Ausland ist für alle Mitglieder des SFB von großer Bedeutung. Beispielsweise sind die folgenden Tagungen wichtig, um die eigenen Arbeiten der internationalen Kritik zu unterwerfen und sich mit Kolleginnen und Kollegen auszutauschen: ICDM, KDD, ICML, ECML PKDD, DATE, VLSI Design, HIPEAC, HICSS, WADS, VISAPP, WABI, COMPSTAT. Hinzu kommen Workshops, die gerade für die Promovenden interessant sind. Betrachten wir nur die Personen, die aus der Ergänzungsausstattung finanziert werden, so ergibt sich aus einer Mischung von europäischen und amerikanischen bzw. asiatischen Reisen inklusive der Tagungsgebühren die angeführte Summe.

3.3.5 Beantragte Globale Mittel

Chancengleichheitsmaßnahmen **für 2015 – 2018**

Gleichstellungsmittel EUR je Jahr 30.000

Zur Ergänzung der universitären Anstrengungen werden zusätzliche Kinderbetreuungsmaßnahmen bei Workshops oder ähnlichen Veranstaltungen des SFB, bei Dienstreisen von Wissenschaftlerinnen/Wissenschaftlern des SFB, bei Abendveranstaltungen des SFB organisiert und finanziert. Spezielle Schulungen zur wissenschaftlichen Karriere für Frauen sollen Wissenschaftlerinnen unterstützen. Ein dritter Aspekt betrifft die Sicherheit bei Dienstreisen, für die auch erhöhte Reisekosten durch Taxibenutzung oder Buchung des Tagungshotels nötig werden können.

Pauschale Mittel **für 2015 – 2018**

Pauschale Mittel EUR je Jahr 30.000

Die pauschalen Mittel werden für bei Antragsstellung nicht vorhersehbare Veränderungen eingeplant. Dies können Neuberufungen oder die Integration von Nachwuchswissenschaftler sein, aber auch Geräte, die durch den technischen Fortschritt wichtig für die Projekte geworden sind.

Koordinierung **für 2015 – 2018**

Verbrauchsmaterial EUR je Jahr 7.200

Für Bürobedarf und Druckkosten rechnen wir je 200,-EUR je Projekt und Projektleiterin/Projektleiter jährlich.

4. Ordnung des Sonderforschungsbereichs

I. Definition und Zweck des Sonderforschungsbereichs

Der Sonderforschungsbereich (SFB) 876 „Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkungen“ ist eine Einrichtung der Technischen Universität Dortmund. Er gliedert sich in Projektbereiche und Teilprojekte und schließt mehrere durch diese Geschäftsordnung gebundene Forschungsgruppen zusammen, die jeweils gemeinsam und im Verbund mit den anderen Gruppen Themen aus den Gebieten „Datenanalyse“ und „Ressourcenbeschränkung“ und „Informationsgewinnung“ bearbeiten. Er steht weiteren Interessenten an den Standorten Dortmund und Duisburg-Essen offen, die ein einschlägiges Arbeitsprogramm gemäß den Richtlinien des Sonderforschungsbereichs verfolgen. Der SFB 876 verfolgt seine Ziele u.a. durch

- gegenseitigen Erfahrungs- und Wissensaustausch und gemeinsame Erarbeitung von Problemlösungen in verschiedenen Arbeitsgruppen.
- Organisation gemeinsamer Veranstaltungen zum Erfahrungsaustausch und zur Präsentation der in den einzelnen Gruppen gewonnenen Forschungsergebnisse auf nationaler und internationaler Ebene.
- Einbindung ausländischer Wissenschaftler und Wissenschaftlerinnen in das Programm, etwa durch Gastaufenthalte oder länderübergreifende Koautorschaft.
- Förderung der Vereinbarkeit von wissenschaftlicher Karriere und Familie.
- Motivierung und Einbindung des wissenschaftlichen Nachwuchses, etwa durch möglichst frühe Einbindung als Teilprojektleiter und die dadurch verbesserten Forschungsmöglichkeiten und Berufungsaussichten. Weiterhin sollen bereits Master-Arbeiten im Rahmen des SFBs ermöglicht werden.

II. Organisatorischer Aufbau des Sonderforschungsbereichs

1. Der SFB hat folgende Organe:
   a) Mitgliederversammlung
   b) Vorstand
   c) Sprecher(in)
2. Teilprojektleiter bzw. Teilprojektleiterin soll stets der- /diejenige sein, der oder die das Forschungsprojekt maßgeblich konzipiert hat. Mehrere Wissenschaftler bzw. Wissenschaftlerinnen leiten ein Teilprojekt gemeinsam, wenn sie je auf ihrem Fachgebiet und in Absprache miteinander das Vorhaben konzipiert haben.
3. Die Vertretung der Promovenden im Integrierten Graduiertenkolleg berät den Vorstand. Sie wird aus dem Kreis der Promovenden von ihnen mit einfacher Mehrheit gewählt.

III. Mitglieder

a) Mitglieder des Sonderforschungsbereichs sind die für die Teilprojekte verantwortlichen Wissenschaftler und Wissenschaftlerinnen („Teilprojektleiter“), auch mehrere je Projekt. Die Mitglieder entscheiden im Rahmen der einschlägigen Richtlinien der DFG eigenverantwortlich über die ihnen zugewiesenen Mittel. Sie verpflichten sich zur gegenseitigen Beratung und Unterstützung und erwähnen in allen aus dem Sonderforschungsbereich hervorgehenden Publikationen die Finanzierung durch die DFG.

   Am Ende jeder Förderperiode oder bei ihrem Ausscheiden aus dem Sonderforschungsbereich legen die Mitglieder einen schriftlichen Bericht über ihre Projektarbeiten vor. Die Arbeitsberichte am Ende jeder Förderperiode sind von den verantwortlichen Wissenschaftlern bzw. Wissenschaftlerinnen eines Teilprojekts gemeinsam zu verfassen.

b) Die Mitgliedschaft endet:

- auf eigenen Wunsch eines Mitglieds durch schriftliche Anzeige bei dem/der Sprecher(in) des SFBs.
- ein Jahr nach Einstellung aller Forschungsprojekte des Sonderforschungsbereichs, an denen das Mitglied beteiligt war.
- auf Antrag von mindestens fünf Mitgliedern und Beschluss der Mitgliederversammlung mit 2/3-Mehrheit aller Mitglieder.

c) Wissenschaftler bzw. Wissenschaftlerinnen, die an einem Informationsaustausch oder an einer Mitarbeit im Sonderforschungsbereich interessiert sind, können auf Antrag "korrespondierendes Mitglied" des Sonderforschungsbereichs (ohne Antrags- und Stimmrecht) werden. Über den Antrag entscheidet der Vorstand.

d) Korrespondierende Mitglieder können neue Teilprojekte vorschlagen. Bei deren Bewilligung werden sie zu Mitgliedern.

e) Die Mitglieder sind zur Zusammenarbeit, gegenseitiger Beratung und Unterstützung verpflichtet. Gemeinsame Einrichtungen sowie die Mittel des Sonderforschungsbereiches können von allen Mitgliedern innerhalb der vorhandenen Möglichkeiten in Anspruch genommen werden.

f) Die Mitglieder sind verpflichtet, an der konzeptuellen Arbeit, den Maßnahmen zur Erhöhung der Chancengleichheit in der Wissenschaft und zur guten Vereinbarkeit von Familie und Karriere, der Förderung des wissenschaftlichen Nachwuchses sowie an der Verwaltung des SFB 876 mitzuwirken.

IV. Mitgliederversammlung

Die Mitglieder des Sonderforschungsbereichs bilden die Mitgliederversammlung. Diese tritt mindestens einmal pro Jahr zusammen. Sie ist darüber hinaus einzuberufen, wenn drei oder mehr Mitglieder dies verlangen. Dies gilt bereits für die Phase der Ausarbeitung des Vollantrages, also vor der Bewilligung/Förderung.

Die Mitgliederversammlung wählt aus ihrer Mitte die Mitglieder des Vorstandes mit absoluter Mehrheit der Anwesenden, wobei die Rolle der Sprecherin (des Sprechers) und des Stellvertreters (der Stellvertreterin) angegeben wird. Die Amtszeit beträgt eine Förderperiode. Wiederwahl ist zulässig.

Die Mitgliederversammlung befindet über den Gesamtfinanzierungsantrag mit einfacher sowie über die Ordnung des SFB 876 mit absoluter Mehrheit. Die Mitgliederversammlung nimmt einmal jährlich den Tätigkeitsbericht des Sprechers (der Sprecherin) entgegen.

V. Vorstand und Sprecher/in

Der Vorstand besteht aus dem Sprecher bzw. der Sprecherin, dem stellvertretenden Sprecher bzw. der stellvertretenden Sprecherin sowie bis zu drei weiteren Mitgliedern des Sonderforschungsbereichs. Sprecher bzw. Sprecherin soll stets der- /diejenige sein, der oder die das Forschungsvorhaben maßgeblich konzipiert hat. Jeder Projektbereich des Sonderforschungsbereichs ist mit mindestens einem Mitglied im Vorstand vertreten. Der Geschäftsführer bzw. die Geschäftsführerin wird von der Sprecherin bzw. dem Sprecher des SFBs bestimmt und gehört qua Amt ebenfalls dem Vorstand an.

Der Vorstand entscheidet mit einfacher Mehrheit aller Abstimmenden über alle Fragen, die die Arbeit und die Organisation und Tätigkeit des Sonderforschungsbereichs betreffen und die nicht der Mitgliederversammlung zugeordnet sind. Abstimmungen über Telefon oder Internet sind zulässig. Bei Stimmengleichheit zählt die Stimme der Sprecherin (des Sprechers) doppelt. Alle Entscheidungen sind zu protokollieren.

Der Vorstand kontrolliert die ordnungsgemäße Verwendung der Sach- und Personalmittel, organisiert gemeinsame Veranstaltungen und entscheidet über die Aufnahme von Projektanträgen (auch Fortsetzungsanträgen). Außerdem entscheidet er über die Aufnahme von Neumitgliedern.

Der Vorstand ist gegenüber der Mitgliederversammlung rechenschaftspflichtig und berichtet jährlich über seine Tätigkeiten und Beschlüsse.

Die mit dem Sprecheramt betraute Person leitet das Verwaltungsprojekt und vertritt den Sonderforschungsbereich nach außen. Sie beruft und leitet die Sitzungen des Vorstandes und die Mitgliederversammlung.

VI. Gemeinsame Projekte und Einrichtungen

Einrichtungen, die dem Sonderforschungsbereich als Ganzes zur Verfügung stehen, werden durch Entscheidung des Vorstandes einem Projektleiter oder einer Projektleiterin des Sonderforschungsbereichs verantwortlich zugeordnet.

Beim Ausscheiden einer Projektleiterin oder eines Projektleiters aus dem Sonderforschungsbereich können die aus Mitteln des Sonderforschungsbereichs erworbenen Geräte, Materialien und andere Forschungshilfen durch Vorstandsentscheid einem anderen Teilprojekt des Sonderforschungsbereichs übergeben werden. Wenn es mit den Interessen des Sonderforschungsbereichs vereinbar ist, kann der Vorstand im Einzelfall beschließen, die Geräte, Materialien und andere Forschungshilfen dem früheren Mitglied befristet zu belassen.

VII. Schlussbestimmung

Diese Ordnung tritt mit der Konstituierung des Sonderforschungsbereichs am 1.Januar 2011 in Kraft. Änderungen der Ordnung bedürfen einer Zweidrittelmehrheit der Mitgliederversammlung.

5. Erklärung zur Raumsituation des Sonderforschungsbereichs

Decken die vorhandenen Räume den Bedarf des Sonderforschungsbereichs zum Antragszeitpunkt?	Ja
Wird der Raumbedarf einschließlich einer ggf. geplanten Erweiterung des Sonderforschungsbereichs gedeckt in den Haushaltsjahren ...	
... 2015?	Ja
... 2016?	Ja
... 2017?	Ja
... 2018?	Ja
Falls eine der obigen Fragen mit „Nein“ beantwortet wird: Sollen die benötigten Räume beschafft werden durch ...	
... Umverteilung vorhandener Räume?	<Ja/Nein>
... Neu- bzw. Umbau?	<Ja/Nein>
... Anmietung?	<Ja/Nein>
Sollen die genannten Maßnahmen finanziert werden aus ...	
... Haushaltsmitteln der Hochschule?	<Ja/Nein>
... sonstigen der Hochschule zur Verfügung stehenden Mitteln?	<Ja/Nein>
Falls Arbeitsgruppen des Sonderforschungsbereichs im laufenden Jahr oder im Antragszeitraum neu erbaute oder umgebaute Räume beziehen geben Sie bitte hier den Zeitpunkt des Umzugs an:	
Lehrstuhl 1, Neubau Fakultät Informatik, Otto-Hahn-Str. 12	05.06.2014
Lehrstuhl 6, Gebäude Fakultät Informatik, Otto-Hahn-Str. 14	16.06.2014
Lehrstuhl 8, Gebäude Fakultät Informatik, Otto-Hahn-Str. 14 und Neubau Fakultät Informatik, Otto-Hahn-Str. 12	02.07.2014

Dortmund, den 2/7/14

Albrecht Ehlers

6. Erklärung zu den Publikationsverzeichnissen

Wir haben darauf geachtet und geprüft, dass bei der Antragstellung die Regelungen der DFG zu Publikationsverzeichnissen beachtet wurden. Dies gilt auch für den beiliegenden Band mit Forschungsprofilen der designierten Teilprojektleitenden.

Dortmund, den 14.07.2014

Prof. Dr. Katharina Morik

(Sprecherin bzw. Sprecher des Sonderforschungsbereichs)

Dortmund, den 10.07.2014

Prof. Dr. Ursula Gather

(Rektorin bzw. Rektor oder Präsidentin bzw. Präsident der antragstellenden Hochschule)

7. Ethikvotum C1

MEDIZINISCHE FAKULTÄT
DER UNIVERSITÄT DUISBURG-ESSEN

Ethik-Kommission
Robert-Koch-Str. 9-11 45147 Essen
Fax: +49 (0) 201/723-5837
ethikkommission@uk-essen.de
www.uni-due.de/ethikkommission

Vorsitzender: Prof. Dr. W. Havers
☎ +49 (0) 201 / 723-1847

Leiterin der Geschäftsstelle: Dr. S. Bein
☎ +49 (0) 201 / 723-3448
sabine.bein@uk-essen.de

Sekretariat: M. Oniangue-Ndza
☎ +49 (0) 201 / 723-3637
monika.oniangue-ndza@uk-essen.de

Wiss. Mitarbeiterin: Dr. K. Appel
☎ +49 (0) 201 / 723-3113
katrin.appel@uk-essen.de

Ethik-Kommission Medizinische Fakultät der Universität Duisburg- Essen

Herrn
PD Dr. Alexander Schramm
Klinik für Kinderheilkunde III
- im Hause -

Essen, den 26. März 14 / KA

Studientitel: **Merkmalsselektion in hochdimensionalen Daten am Beispiel der Risikoprognose in der Onkologie (Teilprojekt C1), im Rahmen des SFB 876 „Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung“**
Antragsteller: **PD Dr. Schramm, Klinik für Kinderheilkunde III**
Unser Zeichen: **08-3670**

Sehr geehrter Herr Dr. Schramm,

hiermit bestätigen wir Eingang und Kenntnisnahme Ihres Schreibens vom 11.03.2014.

Als Vorsitzender der Ethik-Kommission kann ich Ihnen bestätigen, dass skizzierte Vorhaben mit den vorhandenen Informationen und Einwilligungserklärungen abgedeckt ist und keine ethischen und rechtlichen Bedenken gegen die weitere Durchführung des o. g. Teilprojektes bestehen.

Mit freundlichen Grüßen

Prof. Dr. W. Havers

Universitätsklinikum Essen

8. Änderung des Kooperationsvertrags TB1

Änderungsvertrag zum Kooperationsvertrag vom 03.03.2010/05.03.2010 und 18.03.2010

Zwischen der **Technischen Universität Dortmund**
August-Schmidt-Straße 4
44227 Dortmund

vertreten durch die Rektorin, Frau Prof. Dr. Ursula Gather

ausführende Stelle: Frau Prof. Dr. Katharina Morik, Lehrstuhl VIII, Fakultät für Informatik

- nachstehend „Universität“ -

und **B & S Analytik GmbH**
Otto-Hahn-Str. 15
44227 Dortmund

vertreten durch die Geschäftsführerin, Frau Gabriele Sprave

und **Korea Institute of Science and Technology (KIST)**
Europe Forschungsgesellschaft mbH
Universität des Saarlandes
Campus E 7 1
66123 Saarbrücken

vertreten durch den Forschungsdirektor, Prof. Dr. Andreas Manz

- beide nachstehend "Unternehmen" -

und **Hochschule Reutlingen**
Alteburgstraße 150
72762 Reutlingen

vertreten durch die Kanzlerin, Frau Paula Mattes

- nachstehend „Hochschule“ -

wird der folgende Vertrag geschlossen.

Präambel

Die Universität und die Unternehmen führen seit dem 01.11.2011 gemeinsam das Forschungsprojekt (Kooperationsprojekt) mit der Bezeichnung

SFB 876 Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

durch und haben zu diesem Zweck am 03.03.2010/05.03.2010 und 18.03.2010 einen Kooperationsvertrag geschlossen.
Da der am KIST für das Projekt zuständige Projektleiter, Herr Dr. Baumbach, an die Hochschule Reutlingen gewechselt hat, ist es Wunsch aller in diesem Änderungsvertrag aufgeführten Parteien, den Kooperationsvertrag vom 03.03.2010/05.03.2010 und 18.03.2010 entsprechend anzupassen.

§ 1 Ausscheiden von KIST

Das KIST scheidet zum 31.08.2014 aus dem in der Präambel genannten Projekt aus.
Das KIST wird weiterhin die bis dahin vom KIST erarbeiteten Ergebnisse den Projektpartnern nach

Maßgabe des § 5 des Kooperationsvertrages vom 03.03.2010/05.03.2010 und 18.03.2010 zur Verfügung stellen. Darüber hinaus wird das KIST die Regelungen der §§ 6 (Vertraulichkeit) und 7 (Veröffentlichungen) des Kooperationsvertrags vom 03.03.2010/05.03.2010 und 18.03.201 bezüglich der bis zum Zeitpunkt seines Ausscheidens erarbeiteten bzw. zugänglich gewordenen Ergebnisse beachten.

§ 2 Einstieg Hochschule Reutlingen

Mit Wirkung vom 01.09.2014 wird die Hochschule Projektpartner im o.g. Projekt.
Die Hochschule übernimmt ab dem 01.09.2014 die für das KIST vorgesehenen Aufgaben.
Die Hochschule akzeptiert den genannten Kooperationsvertrag, der diesem Änderungsvertrag als Anlage beigefügt ist. Soweit der Kooperationsvertrag von „Universität" spricht, gelten diese Regelungen ebenfalls für die Hochschule.

§ 3 Sonstige Bestimmungen

Im Übrigen bleibt der Kooperationsvertrag vom 03.03.2010/05.03.2010 und 18.03.2010 unverändert.

Sollten einzelne Bestimmungen dieses Änderungsvertrages unwirksam sein oder werden, so bleibt die Gültigkeit der übrigen Bestimmungen davon unberührt. Die Vertragspartner werden sich in einem solchen Fall in gegenseitigem Einvernehmen um eine Vertragsergänzung im Sinne des ursprünglich Gewollten bemühen.

Änderungen und Ergänzungen dieses Vertrages bedürfen zu ihrer Wirksamkeit der Schriftform sowie der vorherigen Zustimmung der DFG. Dies gilt auch für die Änderung der Schriftformklausel.

Dortmund, den 12.06.2014

Technische Universität Dortmund
Für die Rektorin
Der Kanzler
Im Auftrag

Brigitte Trimpe

Für die ausführende Stelle

Prof. Dr. Katharina Morik

Dortmund, den 04.07.2014

B & S Analytik GmbH

Gabriele Sprave
B & S Analytik GmbH

Saarbrücken, den 04.07.2014

Korea Institute of Science and Technology (KIST)

Prof. Dr. Andreas Mänz

Reutlingen, den 10. Juli 2014

Hochschule Reutlingen

Paula Mattes
Hochschule Reutlingen

2