

Collaborative Research Centre 876



# Providing Information by Resource-Constrained Data Analysis

**tu** technische universität  
dortmund

Funding Proposal

2019–2022



**Proposal for the Third Funding Period of  
Collaborative Research Centre 876**

**Providing Information**  
by Resource-Constrained Data Analysis

funded since

1st of January 2011

for

2019 – 2020 – 2021 – 2022

Coordinating university:

Technische Universität Dortmund

---

**Spokesperson of the Collaborative Research Centre:**

Prof. Dr. Katharina Morik

Otto-Hahn-Straße 12  
D-44221 Dortmund

Telefon: 0231/755-5101  
Telefax: 0231/755-5105  
E-Mail: katharina.morik@tu-dortmund.de

**Management or office of the Collaborative Research Centre:**

Dr.-Ing. Stefan Michaelis

Otto-Hahn-Straße 12  
D-44221 Dortmund

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

Dortmund, 12.07.2018

*Katharina Morik*

---

Prof. Dr. Katharina Morik  
(Spokesperson of Collaborative Research Centre)

Dortmund, 12.07.2018

*Ursula Gather*

---

Prof. Dr. Dr. h.c. Ursula Gather  
(Rector of applicant university)



# Table of Contents

<b>1 General information</b>	<b>7</b>
1.1 Key data . . . . .	7
1.1.1 Governing bodies of the Collaborative Research Centre . . . . .	7
1.1.2 Principal investigators and project leaders . . . . .	7
1.1.3 Participating institutions . . . . .	8
1.1.4 Project groups and projects . . . . .	10
1.2 Research profile of the Collaborative Research Centre . . . . .	13
1.2.1 Summary of the research programme . . . . .	13
1.2.2 Detailed presentation of the research programme . . . . .	14
1.2.3 Positioning of the Collaborative Research Centre within its general research area . . . . .	22
1.2.4 National and international cooperation and networking . . . . .	27
1.3 Research profile of the applicant university/universities . . . . .	30
1.3.1 Strategy and planning . . . . .	30
1.3.2 Staff situation . . . . .	32
1.3.3 Research infrastructure . . . . .	33
1.4 Support structures . . . . .	35
1.4.1 Early career support . . . . .	35
1.4.2 Gender equality and family-friendly policies . . . . .	43
1.4.3 Management of research data and knowledge . . . . .	46
1.4.4 Knowledge transfer and public relations . . . . .	49
1.5 Other sources of third-party funding for principal investigators . . . . .	52
<b>2 Existing funds and requested funds</b>	<b>57</b>
2.1 Existing funds . . . . .	57
2.1.1 Overview of existing funds for direct costs . . . . .	57
2.1.2 Overview of existing staff . . . . .	57
2.1.3 List of existing instrumentation . . . . .	58
2.2 Previous and requested funds . . . . .	61
2.2.1 Overview . . . . .	61
2.2.2 Overview of funds requested for staff . . . . .	62
2.2.3 Overview of funds requested for instrumentation . . . . .	62
2.3 Upkeep of laboratory animals . . . . .	62
<b>3 Project details</b>	<b>63</b>
A1 (Morik/Chen): Data Mining for Ubiquitous System Software . . . . .	63
A2 (Sohler/Teubner): Algorithmic aspects of learning methods in embedded systems . . . . .	85
A3 (Chen/Rahnenführer): Methods for Efficient Resource Utilization in Machine Learning Algorithms . . . . .	103
A4 (ten Hompel/Wietfeld): Resource efficient and distributed platforms for integrative data analysis . . . . .	127
A6 (Kriege/Mutzel/Weichert): Resource-efficient Graph Mining . . . . .	153
TB1 (Baumbach/Rahnenführer): Analysis of Spectrometry Data with Restricted Resources	179
B2 (Hergenröder/Weichert): Resource-aware real-time analysis of artefact afflicted image sequences for the detection of nano-objects . . . . .	191

B3 (Deuse/Morik/Wiederkehr (née Kersting)): Data Mining on Sensor Data of Automated Processes . . . . .	217
B4 (Liebig/Schreckenberg/Wietfeld): Analysis and Communication for Dynamic Traffic Prognosis . . . . .	243
C1 (Rahmann/Schramm): Feature selection in high dimensional data for risk prognosis in oncology . . . . .	269
C3 (Morik/Rhode/Ruhe): Multi-level statistical analysis of high-frequency spatio-temporal process data . . . . .	293
C4 (Ickstadt/Sohler): Regression approaches for large-scale high-dimensional data . . . . .	315
C5 (Spaan/Teubner): Real-Time Analysis and Storage of High-Volume Data in Particle Physics . . . . .	335
MGK (Rhode): Integrated Research Training Group . . . . .	359
Z (Morik): Central Tasks of the Collaborative Research Centre . . . . .	381
<b>4 Bylaws of the Collaborative Research Centre</b>	<b>393</b>
<b>5 Declaration on working space for the Collaborative Research Centre</b>	<b>397</b>
<b>6 Declaration on lists of publications</b>	<b>399</b>

# 1 General information

## 1.1 Key data

### 1.1.1 Governing bodies of the Collaborative Research Centre

Spokesperson: Prof. Dr. Katharina Morik  
Surrogate spokesperson/project group B: Prof. Dr. Christian Wietfeld  
Representative for project group A: Prof. Dr. Christian Sohler  
Representative for project group C: Prof. Dr. Wolfgang Rhode

The only governing bodies are the general meeting and the board of management.

### 1.1.2 Principal investigators and project leaders

Principal investigators/Project leaders	Year of birth	Doctorate obtained in	Home institution, location	Project
Chen, Prof. Dr., Jian-Jia	1978	2006	LS 12, Fakultät für Informatik, TU Dortmund	A1, A3
Deuse, Prof. Dr., Jochen	1967	1998	Lehrstuhl für Arbeits- und Produktionsysteme, 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
Kriege, Dr., Nils	1983	2015	LS 11, Fakultät für Informatik, TU Dortmund	A6
Liebig, Dr., Thomas	1980	2013	LS 8, Fakultät für Informatik, TU Dortmund	B4
Morik, Prof. Dr., Katharina	1954	1981	LS 8, Fakultät für Informatik, TU Dortmund	A1, B3, C3, Z
Mutzel, Prof. Dr., Petra	1964	1994	LS 11, Fakultät für Informatik, TU Dortmund	A6
Rahmann, Prof. Dr., Sven	1974	2004	Genominformatik, Institut für Humangenetik, UK Essen, Universität Duisburg-Essen	C1
Rahmenführer, Prof. Dr., Jörg	1971	1999	Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, TU Dortmund	A3

<b>Principal investigators/Project leaders</b>	<b>Year of birth</b>	<b>Doctorate obtained in</b>	<b>Home institution, location</b>	<b>Project</b>
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
Schramm, PD Dr., Alexander	1968	1999	Molekulare Onkologie, Westdeutsches Tumorzentrum, Innere Klinik (Tumorforschung), UK Essen	C1
Schreckenberg, Prof. Dr., Michael	1956	1985	Physik von Transport und Verkehr, Fakultät Physik, Universität Duisburg-Essen	B4
Sohler, Prof. Dr., Christian	1973	2002	LS 2, Fakultät für Informatik, TU Dortmund	A2, C4
Spaan, Prof. Dr., Bernhard	1960	1988	Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund	C5
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
Weichert, Dr., Frank	1967	2010	LS 7, Fakultät für Informatik, Technische Universität Dortmund	A6, B2
Wiederkehr (née Kersting), Prof. Dr.-Ing., Petra	1980	2010	LS 14, Fakultät für Informatik, TU Dortmund	B3
Wietfeld, Prof. Dr., Christian	1966	1997	Lehrstuhl für Kommunikationsnetze, Fakultät für Elektro- und Informationstechnik, TU Dortmund	A4, B4

### 1.1.3 Participating institutions

#### 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
- Lehrstuhl 14 – Software Engineering

**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, UK Essen, Institut für Humangenetik**

- Lehrstuhl für Genominformatik

**Universität Duisburg-Essen, Fakultät Physik**

- Lehrstuhl Physik von Transport und Verkehr

**Universität Duisburg-Essen, UK Essen**

- Institut für Humangenetik, Lehrstuhl für Genominformatik
- Westdeutsches Tumorzentrum, Innere Klinik (Tumorforschung), Molekulare Onkologie

## Außeruniversitäre Einrichtungen

- Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., ein Institut an der TU Dortmund, Abteilung für Materialanalyse
- ARTES Biotechnology GmbH
- Paul-Ehrlich-Institut, Bundesinstitut für Impfstoffe und biomedizinische Arzneimittel

### 1.1.4 Project groups and projects

Project group A: Unifying Data Analysis				
PRJ	Status	Title	Research area	Principal investigator(s), institution(s), location(s)
A1		Data mining for Ubiquitous System Software	Interactive and Intelligent Systems	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund
			Embedded Systems	J.-J. Chen, LS 12, Fakultät für Informatik, TU Dortmund
A2		Algorithmic aspects of learning methods in Embedded systems	Theoretical Computer Science	C. Sohler, LS 2, Fakultät für Informatik, TU Dortmund
			Information Systems	J. Teubner, LS 6, Fakultät für Informatik, TU Dortmund
A3		Methods for Efficient Resource Utilization in Machine learning Algorithms	Embedded Systems	J.-J. Chen, LS 12, Fakultät für Informatik, TU Dortmund
			Interactive and Intelligent Systems	J. Rahnenführer, Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, TU Dortmund
A4		Resource efficient and distributed platforms for integrative data analysis	Communications, High-Frequency and Network Technology	C. Wietfeld, Lehrstuhl für Kommunikationsnetze, Fakultät für Elektro- und Informationstechnik, TU Dortmund
			Logistics	M. ten Hompel, Lehrstuhl für Förder- und Lagerwesen, Fakultät Maschinenbau, TU Dortmund
A6		Resource-efficient Graph Mining	Theoretical Computer Science (Algorithm Engineering)	N. Kriege, LS 11, Fakultät für Informatik, TU Dortmund
			Theoretical Computer Science (Algorithm Engineering)	P. Mutzel, LS 11, Fakultät für Informatik, TU Dortmund
			Process and Knowledge Management	F. Weichert, LS 7, Fakultät für Informatik, TU Dortmund

Project group B: Embedded systems				
PRJ	Status	Title	Research area	Principal investigator(s), institution(s), location(s)
TB1	E	Analysis of Spectrometry Data with Restricted Resources	Mathematics (Statistics)	J. Rahnenführer, Statistische Methoden in der Genetik und Chemometrie, Fakultät Statistik, TU Dortmund
			Medical Biometry, Medical Informatics	J. Baumbach, B&S Analytik GmbH, Dortmund
B2		Resource-aware real-time analysis of artefact afflicted image sequences for the detection of nano-objects	Medical Biometry, Medical Informatics	R. Hergenröder, ISAS e.V., Dortmund
			Image and Language Processing	F. Weichert, LS 7, Fakultät für Informatik, TU Dortmund
B3		Data mining on Sensor Data of Automated Processes	Production Automation	J. Deuse, Lehrstuhl für Arbeits- und Produktionssysteme, Fakultät Maschinenbau, TU Dortmund
			Interactive and Intelligent Systems	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund
			Production Automation (Virtual Machining)	P. Wiederkehr, LS 14, Fakultät für Informatik, TU Dortmund
B4		Analysis and Communication for Dynamic Traffic Prognosis	Traffic and Transport Systems	M. Schreckenberg, Physik von Transport und Verkehr, Fakultät Physik, Universität Duisburg-Essen
			Communications, High-Frequency and Network Technology	C. Wietfeld, Lehrstuhl für Kommunikationsnetze, Fakultät für Elektro- und Informationstechnik, TU Dortmund
			Intelligent and Automated Traffic	T. Liebig, LS 8, Fakultät für Informatik, TU Dortmund

Project group C: Data Volume				
PRJ	Status	Title	Research area	Principal investigator(s), institution(s), location(s)
C1		Feature selection in high dimensional data for risk prognosis in oncology	Oncology	A. Schramm, Molekulare Onkologie, Westdeutsches Tumorzentrum, Universitätsklinikum Essen
			Bioinformatics	S. Rahmann, Genominformatik, Institut für Humangenetik, Medizinische Fakultät, Universität Duisburg-Essen

<b>PRJ</b>	<b>Status</b>	<b>Title</b>	<b>Research area</b>	<b>Principal investigator(s), institution(s), location(s)</b>
C3		Multi-level statistical analysis of high-frequency spatio-temporal process data	Astrophysics	W. Rhode, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund
			Astrophysics	T. Ruhe, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund
			Massively Parallel and Data-Intensive Systems	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund
C4		Regression approaches for large-scale high-dimensional data	Massively Parallel and Data-Intensive Systems, Theoretical Computer Science	C. Sohler, LS 2, Fakultät für Informatik, TU Dortmund
			Mathematics (Statistics)	K. Ickstadt, Lehrstuhl Mathematische Statistik und Biometrische Anwendungen, Fakultät Statistik, TU Dortmund
C5		Real-Time Analysis and Storage of High-Volume Data in Particle Physics	Particles	B. Spaan, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund
			Information Systems	J. Teubner, LS 6, Fakultät für Informatik, TU Dortmund

<b>Integrated Research Training Group</b>				
<b>PRJ</b>	<b>Status</b>	<b>Title</b>	<b>Principal investigator(s), institution(s), location(s)</b>	
MGK		Integrated Research Training Group	W. Rhode, Lehrstuhl Experimentelle Physik 5, Fakultät Physik, TU Dortmund	

<b>Central administrative project</b>				
<b>PRJ</b>	<b>Status</b>	<b>Title</b>	<b>Principal investigator(s), institution(s), location(s)</b>	
Z		Central Tasks of the Collaborative Research Centre	K. Morik, LS 8, Fakultät für Informatik, TU Dortmund	

## 1.2 Research profile of the Collaborative Research Centre

### 1.2.1 Summary of the research programme

The long-term goal of the Collaborative Research Centre (CRC) 876 remains the same for its third phase: providing information by resource constrained data analysis.

The CRC 876 brings together the research areas of data analysis (data mining, machine learning, statistics) and embedded systems (cyber-physical systems) and expands them such that information from distributed, dynamic data masses becomes available for decision processes in real-time, on site. The acquisition and storage of high-throughput experiments in biomedicine, astrophysical telescopes or particle-physical experiments exceeds the capacity of today's computers, so that the analysis must be pushed to the edge, i.e., to the sensor. Also, the analysis of production processes that leads to direct interventions and the prediction of traffic for a better mobility management would benefit from at least a partial analysis at the data sources. The analysis of distributed, streaming data requires novel algorithms and models, that take into account the resource restrictions. Already at the start of our CRC 876, we pointed out that data analysis should inspect more closely its execution on diverse platforms. We defined resource constraints as the relation between the demands of analysing big data and the technical capabilities of a platform or device.

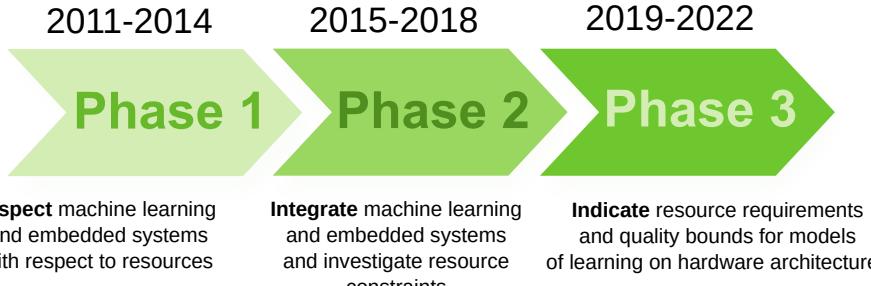


Figure 1.1: The funding phases: from inspecting resource requirements to indicating trade-offs

In the first phase of funding, the interdisciplinary collaboration within computer science was set up. Each discipline inspected the resource constraints. Resource restrictions of computers are always to be seen in relation to the data. While they are immediately noticeable in small, embedded systems, they also apply to large computer systems and centres for particularly large, high-dimensional, distributed, dynamic or complex data. This put the restrictions of energy, computing time, memory, and communication to the front of research.

In the second funding phase of the CRC 876, work on data analysis and on embedded systems have become integrated because of the common focus on resources. Energy measurements and energy harvesting have been pushed to ultra-low power-devices. At the same time, energy consumption of main memory processes and machine learning has been investigated. A common basis of understanding between the disciplines has been established.

The third phase now aims at indicating resource requirements and quality bounds for models of learning on hardware architectures. The goal is to develop models of learning with clear characteristics of resource demands and quality guarantees so that most resource-efficient combinations of machine learning and embedded systems together with algorithmic enhancements and well tailored data storage are indicated.

After the CRC 876, users may easily select learning methods and compose workflows for diverse hardware – balancing energy, memory consumption and communication requirements on the one hand and prediction, prescription and performance on the other hand.

## 1.2.2 Detailed presentation of the research programme

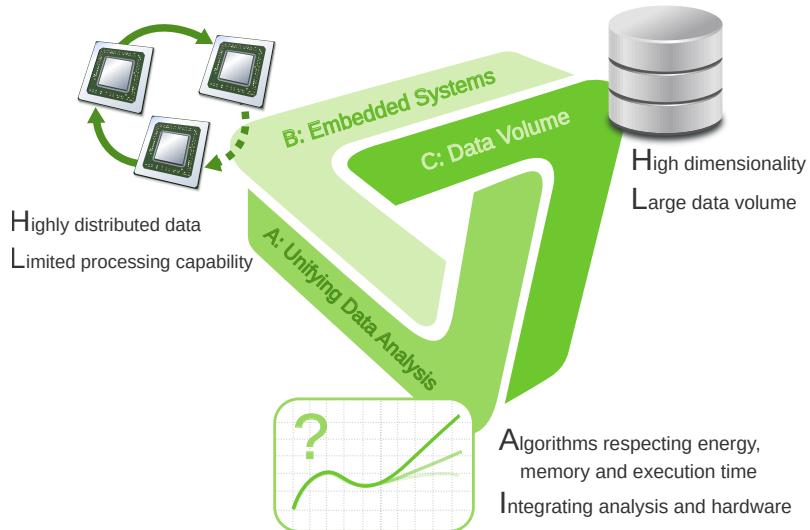


Figure 1.2: Structure of the CRC 876

According to the 2017 white paper of Seagate and IDC on the Data Age 2025<sup>1</sup> data will reach 163 Zettabytes by 2025. This is also due to the trend from centralised to embedded systems. The number of embedded system devices that feed the data centre per person globally is estimated to grow from 1 in 2017 to more than 4 in 2025. The number of interactions per day per person is estimated to become 4800 in 2025. The majority of the data does not need to be stored, but only be used and then discarded. This means that data must be analysed on the fly. It requires intelligent decision making, which data to store and in which aggregated way for further use. Moreover, the trend to real-time data and their use in life-critical applications such as autonomous cars or remote medical patient devices, challenges data analysis even further.

The novel data ecosystems no longer allow to take the von Neumann architecture for granted where only compilers or application systems had to address hardware issues. Already at the start of our CRC, we have pointed out that data analysis is necessary for coping with big data and that we should inspect more closely its execution on diverse platforms. We defined resource constraints as the relation between the demands of analysing data of high volume and high velocity and the technical capabilities of a platform or device. We expressed this view by our logo as is shown in figure 1.2 and used it as the structure of the CRC.

- The large (stored) data with either a huge amount of observations or a very high dimensionality demand scalable and robust analysis algorithms. Project area **C** consists of projects that analyse truly big data.
- Resource constraints often result from local, mobile, small devices. Project area **B** looks into sampling, aggregating and analysing data from distributed sources or even learning on the (distributed) small devices.
- Project area **A** brings these areas together and works on an integrated theory of data analysis and hardware architecture for the new data age.

We are continuously progressing along the path we designed in the beginning. It turns out that the path fits well into the international landscape, because we already planned for a state that now starts to be recognised. Now, let us look at the CRC in a more technical perspective, summarising

<sup>1</sup><https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>

the current state of its research and naming the tasks for the next phase of funding. We structure the CRC contributions along the following research topics:

- The issue which is pursued by all projects is the inspection and investigation of **resource constraints**: energy, memory, real-time, and communication. A short overview is given for all of them, first highlighting CRC contributions and then naming the tasks for the proposed next phase of funding.
- The trend towards the Internet of Things (IoT) and the many data producing devices has led to the programming paradigms of **distributed analysis** and federated analysis with data summaries or sketches as hot research topics.
- The integration of **machine learning and modern hardware** has recently started to raise international awareness and research efforts. Again, the CRC work in the previous phases and the proposed work in the next few years fits nicely the global scientific discussions.
- Solutions to the problem of providing information by resource constrained data analysis are important for the economy and for science. In a **data science** approach, the CRC interdisciplinarily works with biomedicine, engineering and physics. One transfer project is accomplished by the second funding phase, another projects is proposed to become a transfer project in the third phase.

### Resource constraints

Today, resource restrictions are of utmost importance. Energy consumption in particular receives considerable attention. Machine learning is put to good use in order to save energy. Google considers the application of DeepMind's machine learning to its data centres to be its most important application. The energy used for cooling could be reduced by up to 40 percent through machine learning<sup>2</sup>. Machine learning algorithms themselves are enhanced for low energy demands. One of the invited talks at the International Conference on Machine learning (ICML) 2018 is by Max Welling and has the title *Intelligence per Kilowatt-hour*. His statement supports our approach to join embedded systems and machine learning research: "The next battleground in AI might well be a race for the most energy efficient combination of hardware and algorithms." CRC 876 is already participating in exactly this race. Ultimately, for the ease of model selection and the understandability and trustworthiness of the models, there could be a kind of wash paper with characteristics of resource demands, quality and privacy guarantees. Then, users may configure their learning systems like we select our household appliances according to their brochures.

**Energy** The topic of energy efficiency starts with measuring energy consumption. Measuring the true energy consumption directly is difficult because of noise sampling and the need to use minimal resources for the sensing itself. The hardware-integrated sensing instruments are often not precise enough for determining the energy consumed by software running on an embedded system. Hence, an easy-to-use system for direct energy sensing and an energy model for ARM processors based on linear regression have been developed in the platform project A4 [A4/131, A4/130]. The Internet of Things (IoT) and communication networks in general connect small devices. Extremely restricted are the ultra-low-power devices that are used in logistics, e.g., devices attached to a container. Based on reliable measurements, energy harvesting of devices with photovoltaic elements can be realised such that the operating time is enhanced. In a collaboration with Luca Benini, the project succeeded in the even more demanding task of energy measurement for ultra-low power-devices [A4/134]. The analysis of energy consumption of main memory processes in the algorithm project

---

<sup>2</sup><https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/>

A2 [A2/61] and that of high-performance computing in the particle physics project C5 [C5/401] complete the energy modelling.

Minimising the *energy of machine learning processes* requires the analysis of the algorithms and the models of learning so that statistical guarantees can be given. Exponential families are a model of learning that covers many learning tasks, e.g., the estimation of probability density as it is used by, e.g., topic models, or the prediction of the maximum likely state as it is used by, e.g., naive Bayes or conditional random fields. A detailed analysis of the resource demands of exponential families has been achieved in the data analysis for embedded systems project A1<sup>3</sup>. Integer Markov random fields are almost as expressible as are real-valued ones, but can be executed on an ultra-low-power device [A1/C1/29]. Work on decision trees and feed-forward networks has already started.

- The platform project A4 will continue its work on energy harvesting in order to widen the power supply options of the system, cover heterogeneous devices, and enable parallel harvesting.
- Analysing models of learning and algorithms with respect to energy and other constraints will be continued in the next funding phase. It is one of the central issues of the CRC 876 regarded by all projects. Such models offer users the criteria to combine resource-efficiently learning methods and hardware.

**Memory** Reducing the memory demand of storage and analysis of data has been investigated by the algorithm project A2 and the particle physics project C5. In particular for genomic data, a task-specific compression could successfully reduce the storage need of a database [C5/400]. On the level of programming languages and operating systems, the resource utilisation project A3 reduces the memory footprint [A3/92]. Moreover, the dynamic sharing of memory has been optimised [A3/95]. Regarding models of learning, the reduction of the memory demand has been investigated for the exponential families. The key is the compression through regularisation and reparametrisation which exploits redundancies in the true parameters. For exponential families, A1 has investigated several ways of decreasing memory demands. This work will be continued in the next funding phase.

- The analysis of memory use and memory architecture will be enhanced in the next phase of the CRC 876. In addition to the work on exponential families, project A1 will investigate memory hierarchies for random forests and, more generally, how non-volatile memories can be used by embedded systems and machine learning.
- The graph project A6 aims at attentional layers for graphs and a dynamic pooling to reduce the memory consumption of deep learning.
- The oncology project C1 will inspect the balance between lookup speed and memory requirements and design a fast data structure for storing DNA  $k$ -mers.

**Real-time** Our CRC 876 takes into account the execution of learning programs and learned models not only regarding time complexity, but also regarding real-time. Since many embedded systems are integrated into large products that interact with the physical world, timeliness is an important issue. If the results are delivered too late, they may have become useless. As a result, real-time guarantees are needed for such systems. To efficiently utilise the available resources, e.g., processing power, memory, and accelerators, with respect to response time, energy consumption, and power dissipation, different scheduling algorithms and resource management strategies have been developed in different projects. Our CRC 876 has made significant contributions towards

---

<sup>3</sup>The PhD thesis of Nico Piatkowski 2018 *Exponential Families on Resource-Constrained Systems* shows the resource constraints of probabilistic graphical models.

realistic scheduling models and analyses for advanced embedded systems. Some of our results have answered several long-standing open problems. The data analysis project A1 has started to look at deadline misses in real-time systems. [A1/26]. The algorithm project A2 explores the impact of different scheduling strategies for processing queries in databases and energy efficiency of different resource management strategies in main-memory databases. The resource utilisation project A3 optimises the parallel execution of R programs with scheduling for heterogeneous runtimes as they are found both in servers and mobile embedded systems. The optimisation framework creates a scheduling plan based on runtime predictions of jobs. The platform project A4 studies scheduling strategies to improve the quality of transmission and the energy efficiency of Long Term Evolution (LTE) devices. The nano-object detection project B2 focuses on real-time scheduling, discussing partitioned scheduling in multiprocessor systems [B2/21] and task partitioning for self-suspending tasks [B2/205].

- In the proposed next phase of funding, scheduling will be investigated for distributed, asynchronous machine learning in the data analysis project A1 in order to obtain unified results for models of learning under different scheduling models.
- The resource utilisation project A3 will consider CPUs as knapsacks and analyse the trade-off between overhead and performance of solving the knapsack problem of task migration between CPUs.
- The platform project A4 will explore resource management strategies to achieve cooperative networking to overcome the resource constraints of ultra-dense environments.
- The nano-object detection project B2 will optimise resource usages by applying adaptive strategies in feature learning under changing scenarios.

**Communication** The last resource to be presented here is communication. This is particularly important for all processes on mobile devices used in the IoT, logistics as well as traffic forecasting and management. Hence, project platform A4 and traffic B4 were intensively investigating and modeling the communication demands of various scenarios. In his PhD thesis, Christoph Ide has investigated resource-efficient communication in vehicular environments. The view of cars as mobile sensors in upcoming 5G networks makes up for context-aware communication [B4/A4/290]. Communication-efficient machine learning from sensor data has been investigated in the industrial production project B3. The method was general enough to be applied to traffic data, as well.

- Communication-efficient methods will remain in the scope of the CRC 876, because they are a basis for IoT applications including logistics and also because distributed analyses demand communication. In particular, A4 examines 5G mm Wave approaches for ultra-dense IoT systems.
- In a vehicle to vehicle communication, project B4 will develop context-predictive communication with network interfaces to different communication technologies.

## Distributed analysis

Distributed data mining has been pushed more than 10 years ago. The terms peer-to-peer data analysis and ubiquitous knowledge discovery in data were coined the next generation of data mining in 2008<sup>4</sup>. The IoT now puts the subject again on the agenda. Google engineers in India describe *federated learning* as the collaboration of local models with training data remaining on the

---

<sup>4</sup>The famous book *Next Generation in Data Mining* 2008 listed ubiquitous, distributed, and high-performance data mining as one of the five emerging areas.

device and not transmitted to the cloud<sup>5</sup>. With the new architecture, the mobile device downloads the current model and improves it by learning from data on the device. The device generates a summary of what it has learned and sends it to the cloud where it becomes aggregated with other user summaries to refine the shared model. This renaissance fits to the IoT related machine learning in CRC 876 [B3/242, B3/243]. In the industrial production project B3, several distributed learning algorithms have been used for learning from distributed sensors in factories [B3/245] or in traffic systems [B3/246] with special consideration of privacy demands. The resource to be saved is communication between the nodes. Efficient communication has also been explored in the platform project A4 with respect to large-scale IoT warehouse systems.

Data summary or aggregation is necessary in order to learn from distributed sensor streams. The algorithm project A2 succeeded in theoretically well-based sketching or sampling for clustering data streams [A2/57, A2/58]. Its coresets are used to study distributed data analysis with respect to resource constraints [A2/C4/64]. Summaries with a fixed memory size are developed in the data analysis project A1 and synopses from CERN data are studied in the particle physics project C5. Distributed analysis needs to take care of real-time and communication constraints. Its particular paradigm of federated learning supports the achievements of the current funding phase and will be extended.

- The distributed analysis with data summaries, model merging, and model update will become a topic in the proposed next phase for A1, A2 and C4.
- A4 shifts to heterogeneous IoT networks for distributed logistic systems using distributed multi-radio access points. Similarly, C3 moves to an array of telescopes in project C3 and, hence, needs distributed data analysis.
- Parallel distributed systems are investigated by A6 in the course of developing graph kernels and by C5 for distributed storage of massive data streams.
- Project B4 will integrate reinforcements learning for routing and signalling into the distributed vehicle to vehicle communication.

### **Machine learning and modern hardware**

At the 2017 International Symposium on Physical Design, machine learning and artificial intelligence in general were granted a session in which Pradeep Dubey (Intel) advocated a “quest for the ultimate learning machine”. At the International Conference on Computer Design in the same year, FPGAs were claimed to accelerate predictions and, at the same time, require the redesign of the overall prediction pipeline. High throughput and low latency are more important for prediction than for training. For the execution of learned models, particularly for inference, modern hardware (such as an FPGA) has become decisive for business applications, because the energy and communication savings add up to enormous sums of money, when the model is massively applied day in, day out.

Fast training that uses less computation and communication motivates the use of FPGAs within the machine learning community. Fast and resource-restricted inference on FPGAs is also in demand in edge computing settings. While research combining machine learning and hardware was rare at the beginning of our Collaborative Research Centre, it has become multifaceted, receives more attention now, and is still developing.

It is the natural task of the platform project A4 to measure and analyse the performance of various architectures. Now, the project has even produced a cyber-physical node, the PhyNode. It is a

---

<sup>5</sup><https://analyticsindiamag.com/googles-federated-learning-architecture-can-enhance-privacy-while-ending-the-centralised-dataset/>

slave board with memory, radio communication, some sensors, and a solar cell, which is almost energy-neutral due to energy harvesting and energy efficiency [A4/137]. Through a master board, the PhyNode can be managed. In the future, this opportunity will be used for machine learning and inference.

In the particle physics project C5, the acceleration of inference through a combination of FPGA and GPU hardware is important, because the trigger for storing events needs to be very fast. Combinations of different architectures and programming frameworks are being tried. For instance, a cluster of ARM processors for streaming data is combined with Hadoop-based analysis.

In project A1, work on the real-time and low-energy execution of Gaussian processes and decision trees on FPGAs has begun, motivated by the astrophysical telescope array (project C3), where sensor data have to be filtered immediately so that telescopes may react to a perceived source [A1/C3/31]. Enhancing algorithms for storage and execution of decision tree ensembles such as random forests is ongoing work.

- In the proposed next phase of the CRC, A1 wants to develop a model of learning for random forests and deep feed-forward networks such that FPGAs become capable of tailoring compute hardware on demand.
- Learning tasks need to be analysed such that they can be executed on massively parallelised cores without sacrificing the guarantees of the learning model. This is an issue in several projects, especially in A1, C3, and C5.
- Modern hardware will remain a hot topic, especially but not exclusively for the projects A1, A3, B2, and C5.

## Data science

Using data to gain scientific insight has long been a key aim of statistics and knowledge acquisition and discovery. The current interest in data science is motivated by the mass of data and the high dimensions of data that no longer allow a scientist to inspect and explore the data with standard procedures. The international Conference on Data Science and Advanced Analytics started in 2014 and receives growing interest. In 2017, Katharina Morik gave the invited keynote reporting on the results of C3. As its title already indicates, the CRC 876 aims at providing information based on data. Projects in bioinformatics and physics demonstrate the impact of resource constrained data analysis for science.

**Biomedicine** Life sciences have experienced a tremendous upswing due to modern techniques that deliver detailed data. Cancer patients are better diagnosed and receive a more personalised therapy due to genomic data on the order of about 100 gigabytes per patient. Yet, detecting biomarkers in the high-dimensional data is still challenging. The oncology project C1 analyses tumour development on the example of neuroblastoma. The question whether mutations distinguish between primary and relapse neuroblastoma was successfully investigated, and a paper on the results was published in *Nature Genetics* [C1/321]. Mapping DNA fragments against a reference genome is computationally demanding. C1 has developed a new approach to read mapping that uses hashing to distinguish between true read and random matches [C1/319]. Recent nanopore sequencing may turn DNA sequencing into a commodity, but at the same time demands new data analysis methods. The raw data here is a large-volume high-frequency signal of ion currents, which need to be translated into a DNA sequence. Identifying biomarkers requires either better methods for DNA base calling from ion currents or a different representation of the biomarkers. C1 will investigate both alternatives.

The transfer project TB1 has developed methods for the analysis of breath measurements, i.e., ion mobility spectrometry (IMS). The CRC 876 together with the Center of Breath Research at the University of the Saarland, Reutlingen University and *B & S Analytik* organised an international symposium on *Metabolites in Process Exhaust Air and Breath Air* in 2015. The company B & S Analytik offers the products Edmon (measures exhaled drugs), BioScout, BreathDiscovery-Animal and BreathDiscovery-Bacteria. From the resource-aware automatic preprocessing of raw IMS data until the decision support for disease diagnosis and treatment selection, state of the art statistical classification algorithms have been identified and tested on prototypical cases. A publication in PLOS ONE summarizes the results [TB1/C1/186]. Hence, the transfer project ends successfully with the second phase.

Project B2 on detecting biological nano-objects such as, e.g., viruses or vesicles, uses the Plasmon Assisted Microscopy Of Nano-sized Objects (PAMONO). The journal *Sensors* had PAMONO on its cover and the paper on nanoparticle size distribution as the leading article in March 2017 [B2/C1/211]. A challenge is that the processing of the image sequences should be done in (soft) real-time while minimising resource consumption, e.g., that of energy and memory. Efficient and effective algorithms, partly in the framework of deep learning, have been developed. It turned out, that the PAMONO sensor can be extremely useful in relevant fields of pharmaceutical quality control and in-process control of biological materials that may contain viruses (blood products) or virus-like particles (vaccines). Hence, for a third funding phase, project B2 is proposed to be continued as a transfer project. Application partners are ARTES Biotechnology GmbH and Paul-Ehrlich-Institute. The PAMONO technology will be extended to become an adaptive bio-sensor/actuator unit that can be put to good use in medicinal product quality control.

**Engineering** Understanding production processes and developing models of controlling them are now supported by data-driven prognosis. Predictive maintenance, quality prediction, and support of model predictive control have found their way into international engineering conferences. The theme of the 2018 Conference on Manufacturing Systems is *Smart Manufacturing* and two keynote speakers address the use of data, called the “fourth industrial revolution” or “industrial digitalisation”, and the “disruptive technologies” such as, among others, big data analytics. The CRC 876 project B3 on data mining in automated processes contributed a paper on optimising the milling process through machine learning [B3/240]. Machine learning for production often suffers from too few data and these are extremely unbalanced with respect to *ok* and *not-ok*. Simulation contributes additional data, but is often very slow. A way out of this dilemma is active learning, which orders specific simulations for the enhancement of the current learned model. Combining simulation and active learning is a promising approach that also has an impact on another project, namely the Mercur project on tunnelling processes with the Ruhr University Bochum.

In addition to gain insight into production processes, quality prediction and predictive monitoring based on distributed sensor measurements are a key to practical applications. Project B3 together with the company RapidMiner organised the *Industrial Data Science Day* in 2017. Use cases on outlier detection, quality prediction and reduced testing efforts exists, that will be used for feedback on our methods. However, further research on real-time learning and management of many models is still urgently needed.

The book on *Industrie 4.0*, edited by Michael ten Hompel and colleagues<sup>6</sup> has 2.16 millions downloads, being number one in the field of engineering<sup>7</sup>. These two examples already show the visibility of Dortmund’s research in the field.

---

<sup>6</sup>Bauernhansl, T., ten Hompel, M., Vogel-Heuser, B (eds.): *Industrie 4.0 in Produktion, Automatisierung und Logistik*. Springer Berlin 2014, 2. Überarbeitet Auflage 2017; ins Chinesische übersetzt Publishing House of Electronics Industry, April 2015

<sup>7</sup><http://www.bookmetrix.com/detail/book/d1f86f3c-727e-492a-962e-e6a7704212b3>

Traffic prognosis and vehicle routing are important for sustainable transport systems, for private routing, routing of fleets of carrier vehicles, and for public transportation. After early studies of traffic flow, the more individual aspect of multimodal route planning has been investigated based on graph theory and algorithm engineering<sup>8</sup>. Mobility patterns of citizens have been analysed in geographic knowledge discovery. Now, automated vehicles have raised many questions about traffic control, routing, signalling, and communication between the vehicles and how autonomous cars and human drivers will coordinate their driving. Project B4 integrates the traffic data analysis and the information gathering from vehicles and road infrastructure through communication technologies, e.g., 5G networks, in order to model the hybrid traffic.

The European project VaVel already used spatio-temporal random fields of the analysis project A1 on streaming data [A1/C3/293] and the distributed label proportion approach of the production project B3 for privacy-preserving traffic prediction [B3/246]. Christian Wietfeld contributes results from the CRC 876 and his other projects to the 5G-initiative for Germany and collaborates with car manufacturers on optimal transfer times for car data. In this way, basic research of the CRC 876 is already being transferred.

**Physics** The role of experiments in physics has become more and more data oriented, because the events of interest demand sophisticated indirect sensing that produces petabytes of data. The finding of the 17th of August 2017, the merger of two supernovae, illustrates the many senses of astrophysics. 130 million light years away, two stars went supernova, i.e., super-dense neutron stars, circling around each other for most of the history of the universe. They merged in a collision and gamma ray bursts from twin jet streams, outflow of neutrinos, and ejected material were measured by the SWOPE telescope in Chile. Moreover, the gravitational wave observatories in the USA and Italy detected ripples in space and time<sup>9</sup>. Such evidence for hypotheses is only possible, if extremely rare events can be captured. The instruments need the application of data management and analysis supporting the way to a more fundamental understanding of the universe. The CRC 876 takes part in the endeavour with two projects, namely astrophysics project C3 and particle project C5.

The C3 project developed a complete pipeline from calibration and data cleaning through feature extraction and selection to signal separation and energy estimation for the FACT telescope, the FACT tools, built on the **streams-framework** in the second funding phase of the CRC 876 [C3/335]. A student group for one year, PG 594, used the compute cluster of the CRC 876 in order to implement the processes of the workflow in a parallel and distributed programming paradigm that combines MapReduce and Spark or Storm, testing various machine learning alternatives for each step in the pipeline. The next generation of large telescopes, as well the upcoming Cherenkov Telescope Array (CTA) and IceCube-Gen2 demand even more resource-efficient analytics. A visit of C3 member Kai Brügge at the French Alternative Energies and Atomic Energy Commission (CEA) in spring 2017 has ported the concepts and algorithms of the FACT tools to CTA. Work on real-time tools for CTA will be continued. Also, the work on spectral reconstruction and Deep Convolution Networks will be continued in order to exploit the telescopes' full potential.

The particle physics project C5 started in the second phase of the CRC 876. It is based on a collaboration with the Large Hadron Collider (LHCb) at CERN and aims at the development of compact and low-energy streaming algorithms that are accelerated using GPUs and FPGAs. As in C3, the learning task is of the type of searching a needle in the haystack. C5 focuses on memory and storage of the data and then uses the MapReduce paradigm for selecting the extremely rare events. The goal for the proposed funding phase is to connect the results from the current phase, yielding a pipeline that scales up to the coming upgrade of the LHCb project.

---

<sup>8</sup>The survey paper on *Route Planning in Transportation Networks* in the book *Algorithm Engineering – Selected Results and Surveys* 2016 gives an excellent overview of the research.

<sup>9</sup>The telescope FACT, being on the northern hemisphere, could not perceive the merger, neither did IceCube.

The physics projects combine hardware aspects of the measuring instruments and the platforms for analysis with the models of learning and algorithms. In addition to delivering tools and publishing collaborative papers at conferences, both in physics and in machine learning, the transfer into the physical scientific community has recently stepped up via the working group “Physics, Information Technology, Artificial Intelligence” within the German Physics Society (DPG), where Wolfgang Rhode and Katharina Morik are on the steering committee. Moreover, the data sets from our CRC will be used for machine learning training, since the data do not have privacy or competition restrictions. It is planned to deploy data sets to the *Competence Center on Machine Learning Rhein Ruhr* for its education transfer activities.

### 1.2.3 Positioning of the Collaborative Research Centre within its general research area

We explain the evolution and trends of the research disciplines that are brought together in the CRC 876 and indicate briefly for each of them, how the CRC contributions to the field become visible. Overall, the main means of scientific visibility is publication. Around 60% of the more than 100 Core-ranked CRC publications of the second phase are ranked with A\* (32%) and A (28%). Importantly, due to the interdisciplinary character of the CRC, many of the overall more than 340 CRC publications are not covered by the Core-ranking, as those publications address research communities beyond the core computer science community, e.g., physics, medicine, transportation, and electrical and mechanical engineering.

**Machine Learning** According to the AI index of the Stanford University, international publications on topics of Artificial Intelligence are increasing since 1996 much faster (by a factor of nearly 10) than those in computer science (factor 6) and these are growing more than the number from all fields (factor little more than 2). The conference attendance in the field of vision and pattern recognition and in the field of machine learning<sup>10</sup> rose to a level of around 3000 participants, and NIPS jumped to almost 6000. The hype has been raised by impressive results of Deep Learning in image recognition, so representation learning and sampling led to publicity for NIPS as well. Projects B2 and C3 already apply Deep Learning. Looking into the details of Deep Learning, as excellently summarised in the 2017 book<sup>11</sup>, we see that the underlying research questions on regularisation, convolution functions, structured probabilistic models of learning, likelihood estimation and maximum likelihood inference are general machine learning topics, which are also addressed within the CRC 876, particularly in A1’s work on probabilistic graphical models.

The fields of machine learning cover a wide range. A bird’s-eye view sees different approaches: geometric (e.g., decision trees, SVMs), probabilistic (e.g., probabilistic graphical models, Bayesian models), combinatoric (k-means, frequent sets), logic (e.g., inductive logic programming), and reinforcement models (e.g., bandit models). At a more technical level, we see learning tasks that specify the formal basis of machine learning methods, defining what is learned (classifier, regression, probability density, cluster model), from what it is learned (real-valued vectors, time series, categorical data, count data), under which constraints (quality criteria, streaming/online, distributed). As usual in statistics, the term “model” is used not only for the class of possible learning results given the types of input, output and quality criteria, but also for a particular instance, the learning result, as well. Combining approaches and learning tasks, we see the areas of machine learning. All of them are growing. Several algorithms have been developed within these areas. Many of them use algorithms for underlying inner procedures, such as, e.g., matrix factorisation, optimisation, regularisation, or sampling. The resource bounds on memory, runtime,

---

<sup>10</sup>The AI index only counts the ICML, not the KDD or ICDM conferences, probably subsuming the latter under database research and considering it disjoint from AI.

<sup>11</sup>Deep Learning by Ian Goodfellow, Yoshua Bengio, Aaron Courville, 2017, MIT Press

energy and sometimes communication receive more attention as a larger variety of platforms is used and discussed with respect to learning algorithms.

Investigating machine learning at all levels, from the models to hardware architectures, is the particular profile of our research centre, which contributes to machine learning through its publications and its active role on the program committees of the relevant conferences. Keynote addresses of the spokesperson at, e.g., the European Data Forum, the Int. Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), the SIAM Data Mining conference (SDM) and the Conference on Data Science and Advanced Analytics (DSAA) presented the approach of the CRC to production, probabilistic graphical models, and astrophysical science.

**Algorithmics** In the areas of algorithmics and complexity theory, we traditionally consider the resource constraints time and space. Algorithms are analysed in a computational model such as a Turing machine or the RAM model. Besides time and space, there is also a plethora of other resource constraints to be considered. For example, understanding the effects of constraints on communication in distributed systems is a fundamental research direction that has been studied for many years in the (sub)areas of distributed algorithms and communication complexity. In online algorithms we study the effect of not knowing the future (input) at the time the algorithm has to make a decision by comparing the algorithm's output to that of an algorithm that knows the whole future. The resource constraints in the area of streaming algorithms are a certain combination of restricted information (somewhat similar to online algorithms) and space. The areas discussed above are well-established in algorithmics and/or complexity theory and considered to be within the core of the leading conferences in this area such as STOC, FOCS, SODA, ICALP and ESA. In addition, numerous other, less established computational models have been introduced to study other resource constraints. The aspect of energy consumption, which is important for embedded systems, has not received much attention, when it comes to a general study similar to runtime analysis. However, particularly in the areas of online and scheduling algorithms there are many papers that consider scenarios where algorithms are used to reduce energy consumption of a system. Also, in distributed systems (in particular in wireless computation) the study of energy consumption has received a lot of attention. The results of the research centre in the areas of algorithmics and complexity theory will be mainly disseminated through publications in the proceedings of the leading conferences in the area. The workshop on *Algorithmic Challenges of Big Data* at the TU Dortmund in 2015 was organised by the A2 group of Christian Sohler. Petra Mutzel from project A6 brought the meeting of the DFG-SPP 1736 for *Algorithms for Big Data* to the TU Dortmund and the CRC 876 was represented by a keynote and a talk.

**Cyber-physical systems** Since the start of the Collaborative Research Centre (CRC), the research on Cyber-Physical Systems (CPS) has evolved considerably: Having been introduced ten years ago, CPS have become an established mainstream research area, yet receiving continuously growing attention by being a foundation for the Internet of Things (IoT) and a driver for future 5G networks. The broad acceptance and high relevance are indicated by numerous recent and future special issues of journals and magazines as well as by special sessions/panels at high-profile conferences and even dedicated conference series, to which members of the research centre actively promote our specific vision on CPS as editors, chairs, or keynote speakers (e.g., CPS Week, Special Issue on Cross-layer Design of Cyber-Physical Systems in IEEE Design and Test, CPS.HUB/NRW network, Wireless Days, SASIMI workshop). During the early discussions on CPS, the focus was on either the integration of the cyber world and the physical world, or on adding networking to embedded systems. From our point of view, this discussion provided distraction from the fact that almost any sensor will generate raw data samples that are essentially useless unless they are analysed. Even early descriptions of application areas (for example in the acatech report on CPS) were stressing the impacts of CPS in medical applications, traffic, logistics, and manufacturing. Clearly, data analysis is required in these cases. The same applies to other sectors affected by the introduction of CPS. Our research centre focused on the integration of CPS and data analysis very early,

long before the need was recognised in the CPS community and even before the term CPS was coined. As a result, the CRC 876 has paved the way for this integration and it can be considered a pioneer for this. Currently, we observe that many researchers around the world have started to work on this integration as well. As evidence of this fact, we mention special sessions on this topic at well-established conferences (for example, a special session on “AI for CPS: Machine Learning for Intelligent and Secure Cyber-Physical Systems” at ICCAD on Nov. 14th, 2017), special tracks at international conferences (for example, Artificial Intelligence in Cyber-Physical and Distributed Embedded Systems (AICPDES) at Porto on Sept 5-8, 2017) and specialised workshops (for example, Machine Learning for Cyber Physical Systems and Industry 4.0 (ML4CPS) at Karlsruhe on Oct 23-24, 2018). Peter Marwedel’s book with more than 10,000 reads (ResearchGate), especially its enhanced third edition (2018) also makes ideas of the CRC 876 popular. By the more than 14,000 views of his channel *cyphystems* on YouTube, the Dortmund profile is internationally well recognised.

**Physics** The profound impact of the CRC 876 on the results of world-leading experiments can be illustrated using the publication “Multi-messenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A”, which will probably appear in Science on June 21st, 2018. For a long time, high-energy neutrinos of astrophysical origin had been searched for with different experiments, when the IceCube collaboration in 2013 (also using novel methods developed in the CRC 876) managed to find unequivocal evidence of their existence in form of a clear excess over atmospheric backgrounds. This groundbreaking result opened the hunt for high-energy neutrino point sources. On September 22th, 2017, a follow-up observation with the Cherenkov-telescope system MAGIC on the detection of a highly energetic probably astrophysical neutrino detected by IceCube revealed a significant gamma ray flare of an active galactic nucleus positionally coincident with the neutrino event. The active galactic nucleus in question had been known as a source of lower-energy gamma rays beforehand, but was not a previously detected emitter of very-high-energy (VHE) photons, thus greatly underlining the crucial interplay of the experiments IceCube and MAGIC, both of which use methods developed within CRC 876. The coincident detection of a neutrino and of very high energy photons can be seen as a strong indication for the first detection of an individual, discrete source of high energy neutrinos, and, more generally, as potentially humankind’s first direct glimpse on an accelerator of the highest energy Cosmic Rays. Via project C3, we are collaborating intensely with IceCube and MAGIC and have contributed to many topics of the simulation-software and -procedure in both experiments, to the data mining leading to the signal-background separation (neutrino-background separation in IceCube, respectively gamma-hadron separation in MAGIC), to the directional and energy reconstruction of the signal events in both experiments and a multitude of other topics. With special respect to the groundbreaking combined neutrino/gamma ray study described above, members of our working groups provided the Monte Carlo simulations for both experiments, and took on an important role in the coordination of the multi-messenger campaign and in the crucially important rapid response to the neutrino alert and the analysis and treatment of systematic effects.

**Awards** The numerous best paper awards presented for publications of the Collaborative Research Centre 876 at various workshops, summits and conferences speak for its success and impact.

- Wen-Hung Huang and Jian-Jia Chen received the best paper award of IEEE Real-Time Computing Systems and Applications 2015 (RTCSA) in Hong Kong. The awarded paper is “Techniques for Schedulability Analysis in Mode Change Systems under Fixed-Priority Scheduling”.
- The paper written in the C3-project “Online Analysis of High-Volume Data Streams in Astroparticle Physics” by Christian Bockermann, Kai Brügge, Jens Buß, Alexey Egorov, Katharina Morik, Wolfgang Rhode and Tim Ruhe won the best industrial paper award of the ECML-PKDD 2015.

- In July 2016 Jian-Jia Chen received the outstanding paper award 2016 of the ECRTS for the publication “Partitioned Multiprocessor Fixed-Priority Scheduling of Sporadic Real-Time Tasks”.
- In December 2016, the outstanding paper award of the IEEE RTSS Symposium was awarded to Wen-Hung Huang, Maolin Yang and Jian-Jia Chen for the publication “Resource-Oriented Partitioned Scheduling in Multiprocessor Systems: How to Partition and How to Share?”.
- At “Datenbanksysteme für Business, Technologie und Web” (BTW 2017) in Stuttgart, the joint work of Sebastian Dorok, Sebastian Brefk, Jens Teubner, Horstfried Läpple, Gunter Saake and Volker Markl, “Efficient Storage and Analysis of Genome Data in Databases”, received the best paper award.
- On the same day, Stefan Noll, a master’s student of Jens Teubner, received the Best Student Paper Award at BTW 2017 in Stuttgart. His contribution “Energy Efficiency in Main Memory Databases” reports on the key results of his master’s thesis. The thesis was prepared within the DBIS Group and in the context of the Collaborative Research Centre 876.
- Janis Tiemann was awarded the so called best of the best papers award for “Scalable and Precise Multi-UAV Indoor Navigation using TDOA-based UWB Localization” at the International Conference on Indoor Positioning and Indoor Navigation (IPIN) in Sapporo, Japan in 2017.
- The joint work “Unsupervised Data Analysis for Virus Detection with a Surface Plasmon Resonance Sensor” of Dominic Siedhoff, Martin Strauch, Victoria Shpacovitch and Dorit Merhof received the best paper award of the IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA) 2017. The approach was developed in cooperation with the department of image processing, RWTH Aachen University.
- The joint work “On Avoiding Traffic Jams with Dynamic Self-Organizing Trip Planning” of Thomas Liebig and Maurice Sotzny received the best paper award of the International Conference on Spatial Information Theory (COSIT) 2017. It can be considered a preparatory work for Thomas Liebig becoming PI in B4.
- The joint work “LIMoSim: A Lightweight and Integrated Approach for Simulating Vehicular Mobility with OMNeT++” of Benjamin Sliwa, Johannes Pillmann, Fabian Eckermann and Christian Wietfeld received the best contribution award of the OMNeT++ Community Summit 2017.
- The article “Analysis of min-hashing for variant tolerant DNA read mapping” by Jens Quedenfeld (now at TU Munich) and Sven Rahmann has received the best paper award at the Workshop of Algorithms in Bioinformatics (WABI) 2017, held in Cambridge, MA, USA in 2017.
- Petra Mutzel received the best paper award of 28th International Symposium on Algorithms and Computation (ISAAC 2017) for the article “Crossing Number for Graphs with Bounded Pathwidth”. Authors: Therese Biedl (University of Waterloo), Markus Chimani (Universität Osnabrück), Martin Derka (University of Waterloo), and Petra Mutzel (TU Dortmund).
- The joint work “Real-Time Low SNR Signal Processing for Nanoparticle Analysis with Deep Neural Networks” of Jan Eric Lenssen, Anas Toma, Albert Seibold, Victoria Shpacovitch, Pascal Libuschewski, Frank Weichert, Jian-Jia Chen and Roland Hergenröder received the best paper award of the BIOSIGNALS 2018.
- Benjamin Sliwa, Thomas Liebig, Robert Falkenberg, Johannes Pillmann and Christian Wietfeld received the best student paper award for their publication “Efficient Machine-type

Communication using Multi-metric Context-awareness for Cars used as Mobile Sensors in Upcoming 5G Networks” at the IEEE Vehicular Technology Conference (VTC-Spring 2018).

Many individual researchers as well as research groups within the CRC 876 were honoured for their work or academic performance, in general.

- The Bernhard-Walke-Award, which is endowed with 1500 euros, was given to Björn Dusza for his PhD thesis with the title “Context-Aware Battery Lifetime Modeling for Next Generation Wireless Networks” in 2015.
- Christian Wietfeld and his group won the IPIN 2016 competition “Indoor mobile robot positioning” and were invited to present their approach at IPIN in Madrid, Spain.
- Benjamin Sliwa was awarded the ABB-award for finishing at the top of his class for his master’s degree in electrical engineering in 2016.
- Katharina Morik was appointed as a member of the *National academy of science and engineering* (acatech) in 2016. With this appointment, the academy recognises her research profile, her achievements as speaker of the CRC 876, her international reputation and innovative research in machine learning.
- Christoph Ide, alumnus of the CRC 876, received an award for his dissertation “Resource-Efficient LTE Machine-Type Communication in Vehicular Environments” by the Informationstechnische Gesellschaft (ITG). The highly competitive award was presented in a ceremony at the *Berlin-Brandenburgischen Akademie der Wissenschaften* in Berlin in November 2016.
- The Fritz-Lampert award of the TRANSAID-foundation for cancer-suffering children of the year 2016 has been awarded to Alexander Schramm, head of the pediatric-oncologic research lab at the University Hospital Essen. The German-Russian research award recognises excellent researchers and their work in the field of pediatric hematology and oncology for fundamental and clinical research. The award was presented at the semi-annual meeting of the *Gesellschaft für Pädiatrische Onkologie und Hämatologie* (GPOH) in Frankfurt on the 8th of November 2016.
- Katharina Morik became also a member of the *Nordrhein-Westfälische Akademie der Wissenschaften und der Künste* in 2016. The academy connects leading researchers in North Rhine-Westphalia and advises regional policy-makers on questions related to science and its impact on society.
- Michael Schreckenberg became a member of the *Nordrhein-Westfälische Akademie der Wissenschaften und der Künste* in 2017.
- Michael ten Hompel received the honorary doctorate by Hungarian University of Miskolc in 2017.
- Daniel Friesel received an award for finishing best in class of the computer science master’s programme in 2017. His master’s thesis “Automatisierte Verfeinerung von Energiemodellen für eingebettete Systeme” was prepared in the context of project A4 of the Collaborative Research Centre 876.
- Lukas Pfahler was awarded the Hans-Uhde Award 2017 for his master’s thesis “Explicit and Implicit Feature Maps for Structured Output Prediction”.
- Andrea Bommert received an award for finishing best in class for her master’s degree in statistics in 2017.

- The B2-Project publication “Application of the PAMONO-sensor for Quantification of Microvesicles and Determination of Nano-particle Size Distribution” by Victoria Shpacovitch et al. was selected by the journal *Sensors* as the lead article for February 2017.
- Claudia Köllmann received the PhD award of the TU Dortmund University for her outstanding contribution with the title “Unimodal Spline Regression and Its Use in Various Applications with Single or Multiple Modes” in 2017.
- Katharina Morik was selected to lead the working group *technological pioneers* of the platform of learning systems. The aim of the platform, initiated by the Federal Ministry of Education and Research, is to promote the shaping of learning systems for the benefit of individuals, society, and the economy. The working group examines the technological principles and enablers of self-learning systems.
- Nils Kriege became a member of the “Global Young Faculty V” founded by the Stiftung Mercator and the Mercator Research Centre Ruhr. As part of the program, he uses graph based methods developed in the project A6 for the analysis of social network data related to news articles.

#### 1.2.4 National and international cooperation and networking

**National cooperations** A collaboration with the CRC 837 on *Interaction Modeling in Mechanized Tunneling* at the Ruhr University Bochum is realised through a project, in which – based on results from project B3 – new soil settlement prediction methods are being developed for enhanced monitoring of tunnelling. The CRC 837 gives B3 access to its data in order to evaluate some new methods from B3 on real-world data. Information exchange on active learning and the integration of simulation and learning have taken place and will be continued.

As a member of the National academy of science and engineering (acatech), Katharina Morik participated in the project on modern mobility and exchanges information from her EU projects Insight and VaVel as well as results from CRC 876 project B4 with the other participants, who are mostly from transportation or automobile industries. Being as well a member in the cyber-physical hub NRW, she organised in May 2018 an international workshop in Dortmund on modern mobility with the international expert on privacy and mobility, Fosca Giannotti, the coordinator of Insight and VaVel, Dimitrios Gunopoulos, and Thomas Liebig, a member of Insight and VaVel and proposed project leader for B4 in the next funding phase. As always, consideration was given to the graduate school, and PhD students of CRC 876 attended the workshop.

As has already been mentioned above, the CRC 876 with project C3 has become visible in the German physics society (DPG) through the working group *Physics, Information Technology, Artificial Intelligence*. The scientific discussion is not restricted to C3, but involves also A1, e.g., Nico Piatkowski (A1) gave a talk at the astrophysics centre Garching.

In fall 2015, the *Ruhr Astroparticle and Plasma Physics Centre* (RAPP centre) was founded in order to combine research efforts within the fields of plasma- and particle-astrophysics in the Ruhr area. The three universities Ruhr-Universität Bochum, Technische Universität Dortmund and Universität Duisburg/Essen are located in a radius of 20 kilometres, enabling close collaboration between the universities. The founding PIs include Wolfgang Rhode and Bernhard Spaan, who are also project leaders of the CRC projects C3 and C5, respectively. During the inauguration workshop Katharina Morik gave an invited talk on the research impact of data mining for astroparticle physics. In the RAPP centre, about 80 researchers, from master’s level up to staff members, join forces to investigate fundamental physics questions and to break new ground by combining knowledge from the fields of plasma-, particle- and astrophysics.

The federal minister of transport and digital infrastructure started an initiative for a strategy of the next mobile phone generation 5G in Germany. Christian Wietfeld regularly reports to the initiative on the use of machine learning methods for mobile radio (CRC 876), and also on results from other projects, namely networked driving (InVerSiV), rescue robotics (LARUS) and wireless communication for critical infrastructures (TaMIS and BERCOM). The discussions in the 5G initiative also stimulate research in the CRC projects.

As already mentioned, the Federal minister of education and research has funded a platform for learning systems in which Katharina Morik together with Volker Markl (TU Berlin) chairs the working group 1 on technological pioneers. The working group has 12 members from universities in addition to the leaders and 9 members from industries plus representatives from the ministry. This helps to spread the word on CRC 876 and, at the same time, conveys the expert discussion to the CRC 876.

The governmental initiative to push machine learning resulted in the call for proposals on centres of competence. Four proposals were selected, among them the *Competence Center on Machine Learning Rhein Ruhr*, a cooperation between TU Dortmund (lead), Fraunhofer IML Dortmund, Fraunhofer IAIS St. Augustin, and University Bonn, which considers the CRC 876 a basic research reference point. The competence centre combines research excellence with the endeavour of technology transfer into practice, particularly into small and medium companies. The competence centre will start in 2018 and will cooperate with CRC 876, if it is funded in its third phase, offering excellent transfer opportunities for CRC 876.

**International cooperations** The topical seminar each Thursday offers the opportunity to invite colleagues. Some visits have led to more intensive collaborations. Also, having external reviewers with high reputation for our PhD students is a side effect of the topical seminar, where scientific interaction may start. The visit of Albert Bifet (Paris-Saclay), one of the developers of the popular open-source framework for machine learning from data streams *moa*, has led to him becoming a reviewer of the PhD thesis of Christian Bockermann on *Mining Data Streams for Multiple Concepts* (with distinction 2015).

The collaboration with Kanishka Bhaduri and Kamalika Das (both at NASA at that time) started in the first funding phase by their visit in 2011 and led to 4 joint papers, one in the journal of Data Mining and Knowledge Discovery in 2012, one in ECML PKDD in 2013, another one in the text book *Distributed Data Mining in Sensor Networks* 2013, and one in the book *Solving large scale learning tasks – challenges and algorithms* in 2016.

Nico Piatkowski (A1) gave 90-minute talks at NASA Ames Research, Netflix (where Kanishka Bhaduri worked at that time), and Google on his trip to California (October - November 2016) on integer undirected models, spatio-temporal random fields, and reparametrisation, respectively. He met Alejandro Perdomo-Ortiz, member of the NASA Quantum Artificial Intelligence Laboratory (QuAIL). They discussed probabilistic inference on the D-Wave quantum annealer and Nico Piatkowski got access to NASA's quantum annealer to run his own experiments. He stayed at Stanford University as a guest of Stefano Ermon, who became one of his PhD reviewers (with distinction 2018).

Martin Mladenov (A6/B4) got an internship at Google, Mountain View in 2016. He was one of the 32 candidates who were selected from 1,600 applications. He has moved to California.

Nils Kriege (A6) stayed as a visiting researcher at the computer vision and pattern recognition group of the University of York led by Edwin Hancock and Richard Wilson for four months (October to December 2015 and March 2016). Together with Pierre-Louis Giscard, he investigated graph kernels derived from optimal assignments, combinatorial sieves for counting small subgraphs, and the spectral characterisation of the Weisfeiler-Lehman partition. The collaboration resulted in two publications, one in NIPS 2016, the other in the Journal Algorithmica 2018.

Following the topical seminar visit by Luca Benini, Mojtaba Masoudinejad (A4) was able to visit his lab at ETH Zurich complementing the CRC research on energy-efficient systems and energy harvesting. The first stay for four weeks in December 2016 was so fruitful that a second stay in November 2017 followed. Luca Benini will become a reviewer of Mojtaba Masoudinejad's dissertation.

Christopher Morris (A6) stayed at Stanford University as a guest of Jure Leskovec in Stanford's Infolab, where he mainly collaborated on graph representation learning (January to March 2018). Together with other PhD students from the Infolab, he designed an expressive, differentiable pooling layer to lift end-to-end methods for node classification to the graph classification setting. Moreover, he worked on end-to-end trainable graph classification methods that are able to encode higher-order properties of graphs. The resulting work is currently under review for NIPS 2018.

Aswin Karthik Ramachandran Venkatapathy (A4) has been invited by Joe Paradiso at MediaLab, MIT, Cambridge, MA, USA, for a period of three months (June to September 2018). Given the PhyNetLab IoT test bed developed by project A4, the research exchange will work on a human centric IoT model with a specific focus on heterogeneous systems communication. The goal is to extend the work done at the CRC 876 International Summer School 2017 with a lecture on resource-aware machine learning and exercises on ultra-low-power learning. In particular, the platform should enable machine learning in ultra-low-power scenarios.

Christian Sohler has accepted an offer to work as a visiting researcher at Google Zürich from 6 August 2018 to 9 August 2019. One of the reasons for this collaboration, which was started in 2017 with a visit and a workshop participation in Zürich, is certainly his work in the CRC 876, in particular the work on coresets in the algorithm project A2. Silvio Lattanzi from Google Zürich visited Dortmund and gave a talk at the CRC on May 24th. The CRC 876 looks forward to achieving a more intensive collaboration with Google through Christian Sohler's stay in Zürich and his report when he returns in 2019.

Many project leaders of the CRC 876 are area chairs in international conferences and therefore acknowledged by the scientific community. In addition, Katharina Morik is a member of the ECML PKDD Council, which organises the yearly conference, discusses the field with the community, and is now becoming a professional society.

As has been mentioned, the CRC 876 is a member in the astrophysical collaborations MAGIC and IceCube, collaborates with the collaborations CTA and STA through project C3, and interacts with CERN Large Hadron Collider b through project C5. Membership in the collaborations MAGIC for 2018 and IceCube for 2015 and 2016 has been financed by university funds of the CRC 876.

**Cooperations for transfer** The transfer project B1, which successfully ends with the second funding period, cooperated with B&S Analytik. The company produces equipment for measuring and analysing ion mobility spectrometry data.

The project B2 on the detection of nanoparticles is now ready to become a transfer project. ARTES Biotechnology GmbH in Langenfeld and the Paul-Ehrlich Institute (PEI), the Federal institute for vaccines and biomedicines, are the application partners. They are involved in the methodological development and especially, are responsible for the subsequent commercial development of the PAMONO sensor in the fields of quality control in the production of vaccines (ARTES) and quality control of blood donations and blood products (PEI). The research challenge of CRC 876 is to adapt to altering environmental conditions and thus to increasingly diverse characteristics of images. Solutions for an automatic adjustment of the involved actuators and automatic monitoring of the behaviour of PAMONO unit sensor are still needed.

For the verification of applicability and suitability for real-world applications of its results, project B3 had a constant exchange of ideas and the use case of milling with the steel company Deutsche Edelstahlwerke (DEW) in the previous funding periods. For the proposed next phase, we intend to utilise multiple use cases from different industries such as the automobile, electronics and process industries to evaluate and validate our research results. An informal cooperation with BMW on a welding process has started and meetings and data supply have taken place. This use case of quality prediction will be contributed to the project B3. A use case from the electronics industry, to predict the final quality of printed circuit boards, has also been acquired. An application exists in the automotive industry made available process and test data from injector production. In addition to a binary quality label, a continuous label and measured quality test values are also given for all examples. Applications with other companies in production and logistics are being discussed and pursued at the level of bachelor's or master's theses.

## 1.3 Research profile of the applicant university/universities

### 1.3.1 Strategy and planning

Artificial intelligence and data science form a cornerstone of TU Dortmund University's institutional strategy, interconnecting to a range of large-scale research endeavours. Therefore, data analysis, modelling and simulation form a profile area of TU Dortmund University and accordingly receive long-range institutional support in strategic planning, staff appointment and institutional infrastructure.

TU Dortmund University's focus on artificial intelligence (AI) draws on its unique strengths in both basic research and application. Strategic university-wide dialogues under direct participation of the rectorate guide both of these approaches:

The dialogue on *next-generation data science* combines the expertise of the faculties of computer science, statistics and physics and the Max Planck Institute for Molecular Physiology to explore new methods of data analysis and interaction and how these advance scientific analyses. This interdisciplinary effort is also emphasised by two of the four prestigious, university-bridging UA Ruhr professorships appointed in bioinformatics and virtual machining.

Similarly, the *roundtable on industry 4.0* joins Dortmund's capabilities in AI with leading expertise in logistics research and social research, leveraging Dortmund's eminent position in research on labour and social innovation. This interdisciplinary approach develops answers and concepts for the advancing digitisation of labour and production processes in order to shape the ongoing transformation of economy and society.

Structurally, CRC 876 is ideally positioned to guide and catalyse these aims as PIs are members of both strategic dialogues. Furthermore, both UA Ruhr professorships in the data sciences (bioinformatics and virtual machining) are members of both the faculty of computer science and the CRC 876.

The upcoming competence centre on machine learning Rhine Ruhr has the role of coordinating the four competence centres on machine learning. The discussions and networking with the other competence centres will additionally strengthen TU Dortmund University's focus on data science, data analysis, and industrial or scientific applications.

**Embedding into teaching** The TU Dortmund University already offers a wide range of lectures around the specific single topics in the CRC, from cyber-physical system (CPS) to machine learning (ML). Therefore, these courses are not only directly impacted by the CRC research, i.e., teaching newly developed methods, but typically broaden these courses. Due to the consistent dual or triple leadership per project from different disciplines, each method integrated into a lecture can be complemented with an outlook on accompanying methods from the other PIs. A few specific examples where results of the CRC have been integrated into teaching directly:

Several people of the CRC (PIs Jens Teubner, Katja Ickstadt, Jörg Rahnenführer, graduates: Leo Geppert, former CRC postdocs: Michel Lang) are involved in the 9 months certification programme on *Data Science and Big Data*, an initiative of the faculty statistics with strong support of the faculty of computer science and the centre for higher education. The yearly programme started in 2017 and transfers novel methods developed in the CRC 876 (e.g., in project C4) directly to applications.

In project C4, the methods of work package 4 (Incomplete Dependent Variable) were developed within an advanced case studies course. The lecture *Computational Omics* introduces students of computer science to current methods of biomolecular data analysis, as they are also being researched in C1, with a focus on genome data. As part of the preparation for integrating Petra Wiederkehr as a new PI in B3, the application of ML methods for optimising machining processes is discussed – among others – during the proseminar *Computer science meets mechanical engineering*. Project A1 offers every year a so-called technical project for one semester in the bachelor programme under the title *Data mining and data analysis* with varying topics from the CRC.

A fundamental concept in all computer science degree programmes are project groups. These groups for up to 12 students target specific topics in a single projects for the course of one year. Students learn to use and implement novel technology to solve problems in different domains, similar to the dual/triple PI setup of the CRC projects. Several project groups have been created by members of the CRC and taught directly methods from their projects.

PG 594 *Big Data*, as mentioned in section 1.3.3, worked in the context of C3 and besides analysing this project's data also taught fundamental techniques via the CRC compute cluster. The PG 595 *Solar doorplate – Energy saving system software for ubiquitous systems* utilised parts of the platform developed in A4 to teach low-power programming. PG 608 *Manipulation – Development and critical analysis of a framework for extraction and transmission of semantic information between videos* used the same deep learning concepts and frameworks developed in A6 and B2 for altering videos to contain whole new statements. Both project groups 595 and 608 won the yearly award for best project group of the faculty in 2016 and 2017.

The PG 603 *Green Cluster Computing – Processing of high-volume data under resource restriction with modern low-power hardware* worked in the context of C5 and additionally published their results in [C5/401]. The PG 605 *PanGeA* has developed a data structure for pangenomes (collections of very similar genomes of individuals of the same species) and presented them as posters at the German Conference on Bioinformatics (GCB) 2017 in Tübingen.

And finally, for the first time, the CRC by itself offered an official course as part of the programme *Studium Fundamentale*. Courses of this type are mandatory for students of computer science, bio- and chemical engineering, electrical engineering and information technology, city and regional planning, and journalism. These shall offer a broadened view to the students beyond the typical scope of the degree programme. After the successful evaluation of the proposal for the second phase starting in 2015, the first workshop with presentations of the project's results was opened to students as part of the Studium Fundamentale. This had the combined effect of not only integrating new members of the CRC into the continuing research programme, but also providing a view into the research for students.

The course was officially accepted by the computer science faculty's commission on teaching to award 3 ECTS points and granted the authority of examination to Katharina Morik and Stefan Michaelis. As an exception, even students of computer science were entitled for this course due to the high interdisciplinary setup of the CRC's projects, enabling teaching of the scope beyond computer science alone.

The TU Dortmund University invests into CRC related teaching by establishing new professorships, as described in the next section.

### 1.3.2 Staff situation

The Collaborative Research Centre 876 connects the departments of computer science, statistics, electrical engineering, physics and mechanical engineering (including logistics) and is of central importance to the university, demonstrated by a number of permanent additions to academic staff (detailed below). Furthermore, the department of computer science happily extended the position of Katharina Morik so that she will be able to continue to lead the CRC until the end of the proposed third funding period.

Peter Marwedel has agreed to continue to serve on the board of the Collaborative Research Centre offering his invaluable advice and scientific experience.

Also promoted by the high visibility of the research centre, a number of principle investigators received offers from other universities. Christian Sohler received offers for a W3 position in algorithm engineering from Humboldt University Berlin and for theoretical computer science by the University of Hamburg. Due to the excellent conditions for research and teaching and, particularly, the exciting and stimulating environment offered by the CRC 876, he accepted an offer for a W3 position at TU Dortmund.

Sven Rahmann received an W3-offer from the University of Tübingen in May 2016 and led corresponding negotiations for his stay at the University Hospital Essen.

Kristian Kersting has accepted an offer for a full professor position at TU Darmstadt. The position has been reannounced and TU Dortmund is currently negotiating with potential candidates.

Olaf Spinczyk has accepted an offer from University of Osnabrück. His position will be reannounced. Jian-Jia Chen represents the embedded systems part in project A1 and the collaboration has already achieved publications.

Additionally, the department was successful in hiring Petra Wiederkehr as distinguished UA Ruhr professor for virtual machining. She has already started to participate in the CRC and will bring her experience into project B3.

In order to establish a long term cooperation with the Google team, Christian Sohler will be working from August 6th 2018 until August 9th 2019 as a visiting researcher at Google Zürich. The invaluable experiences gained during this visit will lead to inspiring new research ideas.

In addition to existing and new faculty, we have successfully integrated excellent young academics into the research centre as PIs. In particular, Thomas Liebig will follow Kristian Kersting in project B4 and Nils Kriege will replace Christian Sohler in A6.

TU Dortmund University is strengthening the faculty of computer science with four full and two junior professorships, including the UA Ruhr professorship for virtual machining (Petra Wiederkehr). Supporting the envisaged third phase of CRC 876, another four full and two junior professorships will be appointed or extended. These include:

- extending Katharina Morik's tenure as C4-professor for AI until the successful completion of CRC 876's third phase, together with the option to already announce the succession of Katharina Morik so that two AI professors offer lectures and courses,
- a W3 professorship for interactive data science,
- a professorship on statistical methods for big data at the faculty statistics,
- an endowed chair on machine learning for industrial applications in cooperation with logistics
- a new junior professorship with tenure track on applied data science,
- a junior professorship on smart city science (applications are currently processed by the hiring committee).

Over the next few years, TU Dortmund University will significantly expand the career path of junior professorship with a tenure track. To this end, promising research areas have been identified in each faculty and strengthened by the establishment of tenure track professorships. Thanks to the success of this concept in the *Bund-Länder-Programm zur Förderung des wissenschaftlichen Nachwuchses*, TU Dortmund University will receive 14 million euros in the coming years to support 15 tenure track positions. Since all faculties have also agreed to fill another professorship in the tenure track procedure by participating in the programme, TU Dortmund University will be creating options for at least 32 young scientists on this career path by 2028.

### 1.3.3 Research infrastructure

The CRC 876 brings together many research disciplines, and each one has its own needs and requirements for infrastructure and equipment. All of those are provided for by TU Dortmund University: The CRC's researchers have access to a broad and modern infrastructure and effective scientific service facilities, ranging from general services that provide modern computer systems available on demand, to highly specialised laboratory equipment. Below we detail some highlights.

To best assist its academic staff, TU Dortmund aims at further improving the efficiency and suitability of its support structures and of the university administration, including research support services and general administrative procedures. Research support services provide researchers with professional and comprehensive advice on issues of national and international research funding. In addition, they bundle and coordinate existing university-wide services for early-career researchers.

**Computer Infrastructure** Dortmund's most powerful computer is working at the TU Dortmund: In March 2018, LiDO3, the new high-performance computer, went into operation. This hardware cluster for scientific computing is primarily available to researchers and research groups of the TU Dortmund. The IT & Media Center (ITMC) of the TU Dortmund operates LiDO3, which consists of 366 individual compute nodes with a total of 8160 CPU cores and 30 TB of main memory. Access is established via two powerful access computers, so-called gateway servers. The scientific tasks to be calculated are prioritised with a workload manager system and distributed to the actual compute nodes in the cluster. LiDO3 offers users a well-equipped work environment, including software development tools (compilers, debuggers and profilers), free and commercial scientific software packages, and a very large hard disk space per user. The generous equipment, with 40 GPGPUs, also offers the possibility to calculate complex tasks efficiently and quickly in the future.

The CRC 876 invested in its own compute cluster comprising 12 compute nodes, each equipped with 24 CPU cores, 512 GB main memory, 480 GB SSD and 7 TB HDD memory. In addition, three

nodes offer two Nvidia Tesla K40 GPGPUs each, which amounts to 2880 GPU cores and 12 GB GDDR5 memory per node. Recently, the cluster was extended by four nodes powered by Xeon Phi, which has 272 virtual CPU cores (68 real CPU cores), making it possible to deploy massively parallelised algorithms. The compute cluster is running the lightweight operating-system-level virtualisation environment Docker. In contrast to the LiDo3 cluster, this allows each user to specify exactly which software he/she wants to use and run it without restrictions in so-called containers. In particular, every user is the administrator or root user in his/her started containers and can operate them freely, which includes installing any software. Long-term archiving of data and maintenance of the system is guaranteed with the support of the IT Computer Operations Group (IRB), which backs up the data on a daily basis and allows various versions to be restored up to a maximum of six months ago.

The compute cluster makes it possible to train students in managing big data and scheduling jobs of streaming data analysis. Studying computer science includes a year-long project group in the master's curriculum. The project group PG 594 consisted of 12 computer science students and worked in the context of project C3 of the CRC, supervised by Christian Bockermann and Katharina Morik. The group developed an Apache Spark compute environment installed on the CRC's own compute cluster.

The Apache Spark environment can be used to execute large-scale distributed analyses specified by using either the methods supported by Spark natively, e.g., Python scripts or Java programs, or by using the streams-framework developed in C3 during the last phase of the CRC, which allows users to specify analysis pipelines in an XML format. In particular, astrophysical data have been analysed in this parallel environment.

**Logistics** Innovations in logistics and human-technology interaction can be comprehensively researched and evaluated in realistic industrial applications in two halls with the most modern equipment at the Fraunhofer IML in Dortmund. In these warehouses, numerous autonomous actors can be combined into temporary formations that are controlled on an ad-hoc and decentralised basis in order to jointly provide logistical services. In addition to the PhyNodes, which were developed in the second phase of the CRC 876, several strategic devices are available in the research centre, which can be used in the next phase of the CRC 876. An optical, state of the art tracking system with passive infrared tracking and high-tech cameras allows to track three-dimensional movements of people, machines and objects in real time and with millimetre accuracy. Furthermore, a distributed radio-frequency measurement system can measure three-dimensional behaviour in wireless communication technologies. In addition, 30 drones, five mobile robots and more than 300 sensor tags integrated in the ground are available as nodes. These nodes allow researchers to quickly and easily build different networks of heterogeneous objects. An intelligent lighting system in this centre provides an infrastructure for evaluating photovoltaic energy generation indoors.

**Communication Networks** The communication networks laboratory at TU Dortmund University is equipped with a 5G communication network system, a 28 GHz mmWave Transceiver System by National Instruments, which supports new radio (NR) communication at a bandwidth of 800 MHz and a peak data rate of 2.8 Gb/s. This equipment enables researchers of CRC 876 to conduct field validations of beam tracking performance, tracking algorithms and compound localisation and communication systems under development with respect to energy-efficiency, computational complexity, and scalability.

In addition to stationary laboratory investigations, a recently established mobile 5G laboratory enables performance evaluations in a wide variety of field test environments, not limited to fixed laboratory locations. It is based on an extensively modified transporter (VW Transporter T5 with full extension of the interior) and enables the modular integration of previously stationary mobile radio equipment. Depending on individual requirements, the mobile 5G laboratory can be used,

for network operator-independent measurement campaigns and experiments, operation of all-in-one 4G/5G mobile radio networks including the core network as well as the air interface or, for example, as an operating centre for controlling and monitoring outdoor vehicular experiments.

**Biology** At Essen University Hospital, CRC 876's researchers working on the analysis of genome data have access to next-generation MINion DNA sequencers manufactured by Oxford Nanopore. These new sequencers demand new data analysis methods, because the raw data they produce is a large-volume high-frequency signal of ion currents, which must be translated into DNA sequences.

**Physics** Astrophysicists of TU Dortmund and CRC 876 are involved in the international telescope IceCube, located at the South Pole for the detection of neutrinos. This telescope consists of a volume of about one cubic kilometre of ice below the surface at the South Pole, in which highly sensitive light amplifiers search for Cherenkov light that could only have been produced by interaction products of neutrinos that have previously crossed the earth from north to south.

Furthermore, physicists of CRC 876 participate in the development and operation of the First G-APD Cherenkov Telescope (FACT), an Imaging Air Cherenkov Telescope (IACT) located on the Canary Island of La Palma. At the time of its construction in 2011 it was the first IACT to use silicon photomultipliers instead of the conventional photomultiplier tubes. On clear, dark nights, atmospheric Cherenkov light is recorded by the particles produced directly or indirectly by high-energy gamma interactions. The telescope collects up to 1TB of raw data per night, which poses new research challenges for CRC 876.

Physicists as well as computer scientists of CRC 876 collaborate on research at the Large Hadron Collider (LHC) in Geneva, the world's largest and most powerful particle accelerator. The LHC consists of a 27-kilometre ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way. The LHCb experiment at LHC tries to find further information about the asymmetry of matter and anti-matter. The generated masses of data are in the range of several terabytes per second. Because of this rate, the data can be stored only partially and needs to be filtered and analysed in real time, posing great challenges for database system and data analysis systems alike.

## 1.4 Support structures

### 1.4.1 Early career support

TU Dortmund University takes pride in its comprehensive support of early career researchers. Various centres within the university have provided training in transferable skills, career orientation and advice regarding funding opportunities and strategic research planning for years. Strengthening these proven structures, TU Dortmund University has recently created the Graduate Center TU Dortmund University as a university-wide service platform for early career researchers. The Graduate Center (GC) provides central visibility and easy access to all existing formats of support and provides a central contact point for all questions regarding academic careers. The GC is part of the recently established Research Academy Ruhr, the all-encompassing platform for support of early career researchers within the UA Ruhr.

As a joint endeavour of the universities of Bochum, Dortmund and Duisburg-Essen, Research Academy Ruhr supports young researchers and prepares them for careers inside and outside of academia by providing qualification programmes and career guidance for 10,000 young researchers in the Ruhr area. Research Academy Ruhr integrates the know-how of more than ten years of

innovative support of young researchers, including the Graduate Center TU Dortmund University, RUB Research School, Graduate Center plus at the University of Duisburg-Essen, ScienceCareerNet Ruhr and the large number of disciplinary graduate programs within UA Ruhr.

Aside from sheer size and the ensuing synergies, key innovations of Graduate Center and Research Academy Ruhr are:

- A comprehensive understanding of early research careers as ranging from the late master's phase all through to junior faculty,
- raising awareness for career paths outside academic research and providing respective career templates,
- providing tailored qualifications and programs for the respective career phases,
- easing transitions between career phases by providing information infrastructure and informal networks across career phases.

Graduate Center and Research Academy Ruhr are integral elements in the guidance and qualification of the newly created junior professorships tenure track, one of which will be located in the department of computer science.

The CRC 876 participates in career-building events held under the roof of the UA Ruhr targeted specifically for young researchers. For instance, Katharina Morik gave talks at the career forum of the ScienceCareerNet Ruhr in 2015 and 2017 under the title *Publish or Perish* and at the summer evening symposium for postdocs of the ScienceCareerNet Ruhr in 2016, this time on the importance of networking for research careers.

**Principal investigators – Young academics** Early leadership and development of an individual research profile are important milestones in the advancing career of young academics. The CRC supports these high-profile researchers with increased responsibility as principal investigators. Started in the second funding phase of the CRC with two young academics as PIs, the current proposal for the third funding phase further increases the integration with four young scientists:

- Nils Kriege received his PhD with distinction from the TU Dortmund University in 2015 for the thesis *Comparing Graphs: Algorithms & Applications*. He then stayed as a visiting researcher at the University of York for four months working on graph based methods for pattern recognition. Kriege has been involved in project A6 since its beginning and stepped in as principal investigator in 2017 after Kristian Kersting moved to the University of Darmstadt. His main research interests are data mining and machine learning with graphs, in particular, based on graph similarity measures. In project A6 Nils Kriege has been responsible for the development of graph kernels and has taken an active part in the collaboration with C1 and B2. His research has been published in top-tier (Core-Ranking A\*) machine learning and data mining conferences (IJCAI 2018, ICDM 2016, NIPS 2016, ICDM 2014, ICML 2012, ICDE 2011) and renowned algorithmic conferences and journals (Algorithmica, EJC, MFCS 2016, ISAAC 2016, MFCS 2014). Nils Kriege received several awards including the best-of-the-year award of the Department of Computer Science, TU Dortmund University, and two best paper awards (IVAPP). He is a member of the *Global Young Faculty V* of the Stiftung Mercator.
- Thomas Liebig received his PhD from the University of Bonn in 2013 for the thesis *Pedestrian Mobility Mining with Movement Patterns*. During his post-doc at the artificial intelligence unit in Dortmund, he participated in the CRC project A1 and collaborated (while leading the data mining work packages in European big data projects INSIGHT and VaVeL) with

B4, B3, B2, C1, and A1. Joint successes are the best student paper award at VTC spring 2018 and the best paper award at COSIT 2017. Due to his scientific service, more than 50 diverse publications on mobility prediction and analysis, and practical experience from various projects, Thomas Liebig is a renowned scientist in the field of spatial data mining and traffic modelling and a perfect successor as principal investigator in B4.

- Tim Ruhe received his PhD in 2013 from TU Dortmund University at that time already being a highly respected and valuable member of project C3. In his interdisciplinary thesis entitled “Data Mining on the Rocks” he utilised and adapted state of the art methods from machine learning to the search for neutrinos with the IceCube telescope. During his thesis he established a close collaboration with project C3 and initiated an outreach activity, aiming at predicting soccer matches of the 2012 European championship. After obtaining his PhD Tim Ruhe became a PI in project C3, continuing his work on data mining and deconvolution in the context of neutrino astronomy. Within C3, Tim Ruhe is fully responsible for all research carried out in the context of neutrino astronomy. The close collaboration with project A1 was continued, and other cooperations (e.g., B3 and C5) could be established. As a member of the IceCube collaboration and the recently founded AKPIK (Arbeitskreis Physik und moderne Informationstechnologie und künstliche Intelligenz) he has become a representative of the CRC in physics, always aiming at communicating the various goals and achievements of the CRC in a physical context. Tim Ruhe is internationally renowned for his interdisciplinary research.
- Frank Weichert received his PhD in computer science in 2010 from the University of Dortmund, Germany. Since 2010, he has been working as a postdoc at the computer graphics chair of the TU Dortmund. His current research interests include intelligent sensors in co-operative and collaborative sensor-actuator networks, medical image and signal processing and pattern recognition, and the intelligent analysis of large volumes of data with the aid of graphical data processing methods. He is author and co-author of well over 100 publications, some of which have received best paper awards. In addition, Frank Weichert has independently acquired more than 50 national and international research and development projects (including AiF, BMBF, BMWi, DFG, ZIM) and led them as principal investigator or project coordinator. Thereby new methods and theoretical concepts were developed and successfully applied in real-world scenarios, which also provide a profound basis for the projects B2 and A6. With regard to project B2 (sensor-based analysis), for example, the cross-sectional project *Intelligent Sensors* of the BMWi and with regard to the project A6 (geometric deep learning), the BMWi project *Multi-objective optimization-based planning of power-line grid expansions* could be mentioned. Frank Weichert is also an established researcher in cyber-physical systems. An example in this context is the Autonomik 4.0 collaborative project *Smart Face*. Moreover, he has participated in joint preparatory work and cooperation with other principal investigators inside and outside the CRC, e.g., Michael ten Hompel and Petra Mutzel. In the first and second project phase, Frank Weichert was already an employee in the project B2 and involved in the project coordination.

**Scholarships and integration of students** The graduate school of Collaborative Research Centre 876 grants scholarships to foreign students. This way, the members of the CRC 876 have the opportunity not only to work with and learn from talented researchers, but also to spread the results of CRC 876 to an international audience. In addition, these scholarships allow the CRC 876 to explore the possibilities of future international collaborations. During the last funding period, scholarships were granted to the following researchers:

- Liu Wei (November 2015 – April 2016) spent his time at the CRC 876 working with Jian-Jia Chen. He worked towards developing an OpenCL-based remote offloading framework designed for multiple computing units and efficient resource management schemes for these accelerators that also consider the energy efficiency of the embedded devices.

- Junjie Shi (July 2017 – December 2017) worked with Jian-Jia Chen and Jörg Rahnenführer in the A3 project with a focus on concept drift and model-based optimisation. He has also worked on probabilistic schedulability analysis to ensure timely and reliable communication quality, that will be applied in a 5G system. Junjie Shi continued and became a project member in the A3 project.
- Amal Saadallah (April 2017 – September 2017) worked with Katharina Morik in project B3, where she investigated the use of active learning methods in learning tasks that rely on training data generated in cost-intensive finite-element simulations. She distinguished herself by demonstrating her scientific talent and outstanding capabilities in machine learning for industrial processes. After her scholarship ended, she started working in B3 full time, filling the vacancy left by Marco Stolpe’s move to industry.

In CRC 876, students have the chance to participate in research. This is illustrated by the number of publications written by students, often publishing results of their respective bachelor’s or master’s thesis with their respective supervisors. For instance, Maurice Sotzny published results of his bachelor’s thesis with Thomas Liebig, Lennart Downar with Wouter Duivesteijn and Katharina Morik, and Stefan Noll with Jens Teubner, Junjie Shi presented results from work with Jian-Jia Chen.

Members of CRC 876 frequently participate in, or organise events targeted at even younger scientists such as high-school students. For instance, a challenge for the nationwide computer science competition *Bundeswettbewerb Informatik* was prepared and a group of students worked on applying methods of machine learning for object detection in video streams. A similar challenge was held in the context of the DoCampIng programme, which presents various fields of study in engineering to high-school students interested in starting their studies. Other occasions include the *Schülerstag Informatik* and the *Schnupper-Uni*, where students are granted an inside-look into the university.

PRJ	Surname, first name	Type of funding	Topic	Duration
A1	Pölitz, Christian	Existing funds	Automatic methods to extract latent meanings in large text corpora	01.10.2012–24.10.2016
A1	Piatkowski, Nico	DFG SFB876/A1	Exponential Families on Resource-Constrained Systems	01.01.2012–23.04.2018
A1	Buschjäger, Sebastian	Existing funds	Ensemble learning on cyber-physical systems	01.04.2017–21.01.2021
A2	Schwiegelshohn, Chris	DFG SFB876/A2	On Algorithms for Large-Scale Graph and Clustering Problems	01.02.2011–25.08.2017
A2	Krivošija, Amer	Existing funds	Datenstromalgorithmen für Lernverfahren	15.02.2012–31.12.2019
A3	Lang, Michel	Existing funds	Automatische Modellselektion in der Überlebenszeitanalyse	01.10.2011–26.03.2015
A3	Neugebauer, Olaf	Existing funds	Efficient Implementation of Resource-Constrained Cyber-Physical Systems Using Multi-Core Parallelism	01.09.2012–11.06.2017
A3	Kotthaus, Helena	Existing funds	Methods for Efficient Resource Utilization in Statistical Machine Learning Algorithms	01.03.2011–01.06.2018

(Completed or in progress dissertations by CRC staff members)

<b>PRJ</b>	<b>Surname, first name</b>	<b>Type of funding</b>	<b>Topic</b>	<b>Duration</b>
A3	Rempel, Eugen	Existing funds	Statistische Analyse von hochdimensionalen toxikologischen Expressionsdaten	01.04.2009–23.09.2016
A4, B4	Heimann, Karsten	Existing funds	Reliable Control Links for the Internet of Things in 5G Networks	15.01.2017–31.01.2020
A4	Borchert, Christoph	Existing funds	Aspect-Oriented Technology for Dependable Operating Systems	01.01.2011–04.05.2017
A4	Meier, Matthias	Existing funds	Co-Konfiguration von Hardware- und Systemsoftware-Produktlinien	01.06.2009–15.03.2017
A4	Falkenberg, Robert	DFG SFB876/A4	Ressourceneffiziente Kommunikation für Industrie- und Logistikumgebungen	15.12.2014–31.12.2019
A4	Ramachandran Venkatapathy, Aswin Karthik	DFG SFB876/A4	Developing a hybrid communication network for Industry 4.0 systems	01.12.2015–31.12.2019
A4	Putzke, Markus	Existing funds	Selbstorganisierende Minimierung der Interferenz von Femtozellen in heterogenen Netzen durch zufällige Frequenzsprungverfahren	01.09.2008–31.07.2014
A5	Marcel Preuß	Existing funds	Inference-Proof Materialized Views	01.02.2011–24.08.2016
A6	Morris, Christopher	DFG SFB876/A6	Graph Algorithms for Big Data	01.12.2015–31.05.2019
A6	Krieger, Nils	Existing funds	Comparing Graphs: Algorithms & Applications	01.03.2010–28.09.2015
A6	Droschinsky, Andre	Existing funds	Graph Similarity Problems	01.04.2015–31.03.2019
A6	Fey, Matthias	DFG SFB876/A6	Geometric Deep Learning: Algorithms & Applications	01.05.2018–30.04.2021
B2	Huang, Wen-Hung	Existing funds	Scheduling algorithms and timing analysis for hard real-time systems	01.04.2013–18.04.2017
B2	Libuschewski, Pascal	DFG SFB876/B2, Existing Funds	Exploration of cyber-physical systems for GPGPU computer vision-based detection of biological viruses	01.03.2011–22.03.2017
B2	Holzkamp, Olivera	Existing funds	MemoryAware mapping strategies for heterogeneous MPSoC systems	01.08.2010–16.03.2017
B2	Lenssen, Jan Eric	DFG SFB876/B2	Semantic Image Representations with Deep Neural Network Architectures	01.12.2017–31.12.2019

(Completed or in progress dissertations by CRC staff members)

PRJ	Surname, first name	Type of funding	Topic	Duration
B2	Siedhoff, Dominic	DFG SFB876/B2, Existing Funds	A parameter-optimizing model-based approach to the analysis of low-SNR image sequences for biological virus detection	02.01.2011–15.09.2016
B3	Finkeldey, Felix	DFG SFB876/B3	Optimierung von Fräsprozessen durch die Kombination von Simulationstechniken und Machinellem Lernen	01.11.2015–31.12.2020
B3	Erohin, Olga	Existing funds	Knowledge acquisition through data analysis for prospective time determination	01.06.2013–22.06.2016
B3	Konrad, Benedikt	Existing funds	A methodology for family-based balancing of variant flow lines	01.01.2014
B3	Stolpe, Marco	DFG SFB876/B3	Distributed analysis of vertically partitioned sensor measurements under communication constraints	01.01.2011–31.01.2017
B4	Niehöfer, Brian	Existing funds	Modellbasierte Interferenzkompensation für die satellitengestützte Ortung in urbanen Szenarien	01.10.2008–30.04.2016
B4	Sliwa, Benjamin	DFG SFB876/B4	Mobility-aware Vehicular Communication	01.06.2016–31.05.2021
B4	Ide, Christoph	Existing funds	Resource-Efficient LTE Machine-Type Communication in Vehicular Environments	01.10.2010–31.03.2016
C1	Köster, Johannes	Existing funds	Parallelization, Scalability, and Reproducibility in Next Generation Sequencing Analysis	01.01.2011–03.09.2015
C1	Hess, Sibylle	DFG SFB876/C1	Matrix Factorization with Binary Constraints	01.06.2015–31.12.2018
C1	Schulte, Marc	Existing funds	Funktionelle Validierung von Rezidiv-spezifischen Mutationen bei Neuroblastomen	01.03.2015–31.12.2018
C1	Timm, Henning	Existing funds	Error-tolerant hash functions for rapid DNA read mapping and ddRAD sequencing	01.01.2015–31.12.2019
C1	Ramke, Marianna	DFG SFB876/A6, Existing funds	Alignment of multiple ion mobility spectra under time and memory constraints	01.11.2011–01.06.2020
C3	Nöthe, Maximilian	Existing funds	Source investigations with FACT	01.07.2016–31.08.2019

(Completed or in progress dissertations by CRC staff members)

<b>PRJ</b>	<b>Surname, first name</b>	<b>Type of funding</b>	<b>Topic</b>	<b>Duration</b>
C3	Werthebach, Johannes	Existing funds	Multi Muon Measurements with IceCube	01.01.2017–31.12.2020
C3	Baack, Dominik	Existing funds	GPU Computation and Optimisation of Atmospheric Air-Shower Simulations	01.01.2017–31.12.2019
C3	Brügge, Kai	DFG SFB876/C3	Machine Learning VHE Gamma Ray Astronomy	01.08.2017–31.07.2019
C3	Buß, Jens	DFG SFB876/C3	Influence of bright light conditions to the First G-APD Cherenkov Telescope	01.10.2013–31.12.2018
C3	Börner, Mathis	Existing funds	Bestimmung des Energiespektrums von atmosphärischen Myonneutrino mit 3 Jahren Daten des IceCube-Detektors	01.10.2014–31.05.2018
C3	Menne, Thorben	Existing funds	Stacking Point Source Search for Lower Energy Contribution at HESE Event Positions with IceCube Data	15.10.2014–30.06.2018
C3	Linhoff, Lena	Existing funds	Multiwavelength Point Source Analyses	01.07.2017–30.06.2020
C3	Sandrock, Alexander	Existing funds	Radiative corrections to the energy loss cross sections of high-energy muons	15.04.2015–30.06.2018
C3	Overkemping, Ann-Kristin	Existing funds	Messages from a Black Hole - A long-term analysis of the Active Galactic Nucleus Markarian 421 in the light of gamma-rays measured by MAGIC-I	01.04.2012–31.03.2015
C3	Meier, Maximilian	Existing funds	Suche nach astrophysikalischen Tau-Neutrinos mit dem IceCube-Detektors	01.09.2015–30.09.2018
C3	Fuchs, Tomasz	Existing funds	Charmante Myonen im Eis - Messung des hochenergetischen atmosphärischen Myon-Energiespektrums mit IceCube in der Detektorkonfiguration IC86-I	01.11.2012–30.06.2016
C3	Temme, Fabian	Existing funds	On the Hunt for Photons - Analysis of Crab Nebula Data obtained by the first G-APD Cherenkov Telescope	01.06.2013–30.09.2016
C3	Einecke, Sabrina	Existing funds	The Data Mining Guide to the Galaxy - Active Galactic Nuclei in a Multi-Wavelength Context	01.06.2013–31.10.2017

(Completed or in progress dissertations by CRC staff members)

PRJ	Surname, first name	Type of funding	Topic	Duration
C3	Schenetten, Kai	Existing funds	Röntgen-Photoelektronenspektroskopie mit Silizium-Photomultipliern	01.10.2013–30.04.2016
C3	Bockermann, Christian	DFG SFB876/C3	Mining Big Data Streams for Multiple Concepts	01.10.2007–23.01.2015
C3	Frantzen, Katharina	Existing funds	Von der Monte-Carlo-Produktion zur Datenanalyse - Eine Analyse der 2012 genommenen Daten des Aktiven Galaktischen Kerns Mrk 421	01.01.2012–30.06.2015
C4	Köllmann, Claudia	Existing funds	Unimodal spline regression and its use in various applications with single or multiple modes	08.12.2010–09.09.2016
C4	Geppert, Leo N.	DFG SFB876/C4	Bayesian and Frequentist Regression Approaches for Very Large Data Sets	01.01.2010–30.09.2018
C4	Munteanu, Alexander	DFG SFB876/C4	Large Scale Statistical Data Analysis	01.01.2010–30.09.2018
C5	Eitschberger, Ulrich	Existing funds	Flavour-tagged Measurement of CP Observables in $B_s^0 \rightarrow D_s^\mp K^\pm$ Decays with the LHCb Experiment	01.01.2013–28.05.2018
C5	Breß, Sebastian	Existing funds	Efficient Query Processing in Co-Processor-accelerated Databases	01.04.2014–30.10.2015
C5	Meier, Frank	Existing funds	Measurement of $\sin 2\beta$ using charmonium and open charm decays at LHCb	01.01.2013–13.12.2016
C5	Niet, Ramon	Existing funds	Measurements of CP Violation in $B^0 \rightarrow [c\bar{c}] K_S^0$ Transitions at LHCb Experiment	01.01.2013–10.09.2018
C5	Stevens, Holger	DFG SFB876/C5	GPU based Tracking for the SciFi-Tracker of the LHCb Experiment	01.01.2016–31.12.2019
C5	Heinicke, Kevin	Existing funds	Flavour Tagging and CP Violation Measurements at LHCb	01.10.2016–20.09.2020
C5	Schellenberg, Margarete	DFG SFB876/C5	Measurements of CP Violation in $B$ Decays to open charm with the LHCb Experiment	15.10.2015–30.09.2019

(Completed or in progress dissertations by CRC staff members)

Duration of contract	Number of contracts for doctoral researchers		Number of contracts for postdoctoral researchers		Number of researchers in total
	male	female	male	female	
up to 12 months	26	7	3	5	28
up to 24 months	9	4	0	1	13
up to 36 months	11	4	0	0	15
up to 48 months	3	1	4	0	8

Typically, contracts should be scheduled accordingly equal to the pursued qualification, which is a minimum of three years for targeting a PhD. In case of funding, contracts should have a duration equal to the funding period. Major reason for shorter contracts therefore is the change between PhD finalisation of researchers and the new researchers starting in the second phase.

#### 1.4.2 Gender equality and family-friendly policies

To remain attractive in the competition for future talents, TU Dortmund will increase its already broad offer of accompanying measures in the fields of diversity, equality of opportunity, and family-friendliness. Family-friendliness will be further improved to allow a maximum of flexibility for parents. New offers will include emergency short-term care and kids room.

TU Dortmund University has been successful in establishing strategies for pursuing gender equality and has been awarded for these efforts. The audit *familiengerechte hochschule* was granted in 2008, and in 2014 the TU Dortmund signed the Charta *Familie an der Hochschule*. The university has successfully participated in the *Professorinnenprogramm des Bundes und der Länder* (2008, 2013) with one granted professorship for a female scientist so far. For further improvement of equal opportunities in research, the university has committed itself to implement the DFG's Research-Oriented Standards on Gender Equality and has been classified within the best category since 2011. The strategic goal of being a family-friendly university is incorporated in the quality control processes for studies and research.

In 2011, the TU Dortmund University was one of eight universities in Germany selected by the Stifterverband für die deutsche Wissenschaft (Donors' Association for German Science) to develop the *Shaping Diversity* audit. The university successfully completed the auditing process in 2012. On 8th of October 2018, it will face the re-auditing and present the development to date in the diversity field of action and strategy of the TU Dortmund University.

At present, approximately 23% of the professorships are filled by women, while the CRC has a slightly higher percentage of 25%. The TU Dortmund University will use every opportunity to increase this quota. The current target rates agreed for 2019 are 28% on average across the university. The university's appointment regulations, appointment management and appointment portal contribute to equality through gender equality and transparency in appointment procedures. A new service to be implemented in 2018 will be emergency day-care for the children of CRC's scientists.

To ensure fair appointment procedures, TU Dortmund enforces equal gender representations on appointment committees and ensures active participation of equal opportunities commissioners in all steps of the appointment procedure. Appointment procedure rules including the focus on gender equality aspects are stated in a guideline and a checklist (Berufungsordnung and Berufungsleitfaden, both updated and adjusted accordingly in 2012), which allow a transparent and flawless procedure.

To attract more applications from female scientists, recruitment with emphasis on gender equality measures and family-friendly working environment is pursued. In this context, a dual career programme has been developed to support the newly appointed professor's partner in finding a suitable position according to their previous career.

Dortmund offers the programme mentoring<sup>3</sup>, supporting female PhD students and postdocs. It is a joint initiative of the University Alliance Ruhr (UA Ruhr) and part of Research Academy Ruhr. The participants receive coaching by mentors who are researchers at all career stages. By them mentees obtain qualifications required for persons in leading positions. The programme consists of workshops and lectures, introducing topics like external funds acquisition, scientific networking, and career policies, providing essential information for a successful start into an academic career. TU Dortmund also supports DFG proposals for research grants for females at the postdoc level by coaching the applicants.

As a certified family-friendly institution, TU Dortmund University offers to their members multiple child care facilities and multi-faceted individual support when looking for child care through their family offices. The services comprise short time child care, child care during vacations, and child care in a kindergarten on campus. Further activities concern the improvement of the conditions for students with children for example by possibilities for part-time studies. Additional offers concern teleworking opportunities and flexible working times.

Child care during easter, summer, and fall vacations: Child care is offered during the vacations with an entertainment programme that tries to raise the interest of children in science. Child short time care KuKi: In the short time care facilities, children of university members starting from age one are taken care of by nurses. Also, emergency care possibilities are offered. Regular child care facilities: The kindergarten HoKiDo on campus provides multiple offers for 60 children. Further child care places are offered by the kindergarten *4 Jahreszeiten* with 120 care places. The family office cooperates with the Dortmund city administration for communal child-care institutions. Another on-campus kindergarten 9xkluge Zwerge offers a flexible child-care for children whose parents work in DFG financed projects.

Family office: The university operates a family office which helps to find child care services such as baby sitters and day nannies. The office also gives advice how to combine and organise studies with family. A measure to improve study conditions are individual examination dates, independent of official examination periods.

Availability of sufficient child-care facilities is important for both female and male young scientists in order to combine a career in science with family and will be taken care of.

#### A. Research Staff

	1 <sup>st</sup> Funding Period	2 <sup>nd</sup> Funding Period				3 <sup>rd</sup> Funding Period	
		Targeted proportion of women [%] <sup>12</sup>	Current number of men / women <sup>14</sup>	Current proportion of women [%] <sup>14</sup>			
				m	w		
Doctoral researchers	16	—	41	8	16	22	
Postdoctoral researchers	29	—	4	1	20	20	

<sup>12</sup>Reference date: proposal for 2<sup>nd</sup> funding period

<sup>13</sup>According to proposal for 2<sup>nd</sup> funding period

<sup>14</sup>Reference date: proposal for 3<sup>rd</sup> funding period

## B. Principal Investigators

Position	1 <sup>st</sup> Funding Period Proportion of women [%] <sup>12</sup>	2 <sup>nd</sup> Funding Period				3 <sup>rd</sup> Funding Period		
		Proportion of women [%] <sup>13</sup>	Current number of men / women <sup>14</sup>		Current proportion of women [%] <sup>14</sup>	Number of men / women		Proportion of women [%]
			m	w		m	w	
Postdoctoral researchers	0	0	2	0	0	4	0	0
Group leaders... <sup>15</sup>	0	0	1	0	0	1	0	0
Professors C3/W2	0	0	4	0	0	3	1	25
Professors C4/W3	27	21	12	3	20	9	3	25
Total	16	13	19	3	14	17	4	19

The targeted percentage of women in positions doctoral researchers is based on parity staffing of the currently open positions in this proposal, which would further increase the already higher than average number of researchers on all levels. The PI of the CRC are encouraged to try to fill the remaining positions accordingly.

**CRC-specific gender equality and family-friendliness activities** The central measures of the TU Dortmund and UA Ruhr are supplemented by the Collaborative Research Centre with specific measures for its employees.

At the beginning of the second funding phase a survey, supported by the equal opportunities, family and diversity unit of the TU Dortmund University, was conducted. This survey included conversations with female members of the CRC on all experience levels and tried to find answers if the CRC specifically is different from other research opportunities.

Hardly surprising, the feedback praised the overall excellent research environment the CRC offers for the participating women. This is not different from overall feedback of CRC members regardless of gender, who frequently mention the degrees of freedom of research, the networking opportunities, locally inside the CRC as well as internationally, and the potential impact they can achieve due to the interdisciplinary topics.

But two points occurred several times as feedback: Women do not want to be singled out, but naturally be seen simply as equal researchers, which was the case for the interviewed researchers. Additionally, the role of encouragement to pursue a PhD was mentioned. Even for extraordinary female students, simple direct encouragement to pursue a PhD by a master thesis supervisor or lecturer may provide the tipping point for the decision to follow this career path. The PIs as well as thesis supervising PhD graduates can improve the number of women by identifying promising candidates and informing them personally about the process of obtaining a PhD. Once being a research member of the CRC, the University programmes as well as these depicted below further strengthen the career planning activities.

<sup>15</sup>Research group leaders, junior research group leaders, junior professors

In the mentoring<sup>3</sup> programme, a separate group was formed for female scientists of the CRC in cooperation with the graduate school of logistics in Dortmund, for whom a special qualification programme was developed and which was supervised by the coordinator of the programme. These mentoring groups have a total duration of two years and the current groups is scheduled to end in November 2018. Due to the positive feedback ranging back to the first group in the first funding phase, a new group shall be offered again for the final phase.

As a further measure for career promotion, the female CRC scientists were again given the opportunity to participate in the *Das Arroganz-Prinzip* seminar. This explicitly included the CRC project leaders. Despite its catching name, this seminar offers a practice-oriented test of dealing with male colleagues, employees and superiors. By means of a personal sparring partner per participant, the topics of appropriate self-portrayal, spatial behaviour as a power factor or the perception of women's performance by men were trained in exercises.

Additionally, a proven valuable and ongoing activity for family friendliness is the day care for children of employees in DFG projects, *9 × kluge Zwerge*, under supervision of the collaborative research centre 823. This offer expands the existing public care services. The day care unit, which cooperates with the mothers' centre Dortmund, is an association of three childminders, each of whom looks after three children in these premises. A varying number of children of CRC researchers and PIs have been cared for during the last years. The costs are borne on the basis of a service contract with the mother's centre to approx. 30% by means of equality funds, the remainder by means of the basic equipment.

#### **1.4.3 Management of research data and knowledge**

TU Dortmund University places strong emphasis on research data management (RDM), currently developing a university-wide process and technical infrastructure applicable for all disciplines. The university receives external funding from the Federal Ministry of Education and Research and the Mercator Research Center Ruhr supporting its efforts and is also currently creating a permanent scientific position for implementing and supporting the central RDM infrastructure.

Research Data Management ensures the permanent backup, availability and reuse of research data in all scientific disciplines. RDM is understood to mean all steps that are taken from the processing of the data to its storage and archiving and possibly publication and exchange with other scientists. RDM does not mean mere technical storage, but an overall process, which describes the structured data storage and documentation, the usability, comprehensibility and long-term storage up to sharing and publication including legal protection of the data along its life cycle. In addition to the creation, processing and analysis of data, RDM covers dissemination, archiving and in particular the subsequent use of one's own data, whether for one's own follow-up work or that of other researchers. The value of data management is becoming more and more globally accepted and even triggers its own publication paths like data journals. Research data management obviously is a part of good scientific practice.

Due to the growing importance, the TU Dortmund University started the process of implementing RDM as a university-wide service. Surveys to assess pre-existing RDM practice and demands for a university-wide service began in 2016 with the CRC and its individual PIs being important pilot partners. In 2017, the university successfully acquired funding by the Federal Ministry for Education and Research to develop and implement a respective RDM process infrastructure for all research cultures (as a possible best-practice template) until 2019. In this project, potential solutions for software support, which needs to find the balance between standardisation and flexibility for various scientific disciplines, are identified. Full roll-out is planned to commence in 2021, enabling integration in and usage by the CRC during the third funding phase.

The process specification of RDM will cover the complete life cycle: Planning of data acquisition; data validation (including necessary anonymisation); analysis; sharing and access control (including ban period management); migration (including archiving); subsequent use (including documentation, accessibility and metadata).

Complementing the development of RDM processes, TU Dortmund University cooperates closely with its partners in the University Alliance Ruhr and those of Aachen and Cologne to implement joint storage infrastructure, as exemplified – but not limited to – a joint application for distributed storage system in the Major Research Instrumentation Programme.

Until full implementation of the technical RDM infrastructure, a central RDM group combining expertise from Research Support Services, University Library and the IT and Media Center advises toward interim solutions and general good practice in research data management. This service includes a joint UA Ruhr-wide instance of the Research Data Management Organizer Software (RDMO) to be implemented with external funding by the Mercator Research Center Ruhr.

**CRC-integrated research data management** The global research data management strategy of the university is complemented by the ongoing development of services offered centrally as part of the CRC itself. Concepts are tested inside the CRC and matched against the ongoing specification process of the university research data management group.

Local storage and archives for CRC-owned data have already been established for the second phase. The official registration as a data centre at the DataCite organisation enabled the CRC to issue (and freeze) citable data sets with a digital object identifier (DOI). This capability by now has been transferred to the university library for roll-out as a university-wide global service.

The CRC-exclusive cluster does not only provide computational facilities to the projects as described in section 1.3.3, but also large scale data storage for the experiments. While the initial concept concentrated on providing easy access to the necessary storage for the experiments, the current work broadens the focus on strong reproducibility of research.

Reproducibility and repeatability are of major importance for transparency of research results as well as comparability of future developments. For the CRC 876 this affects *software* as well as *data*. As current version control systems provide the tools to tag versions for later reference, these methods are not sufficient. Due to changes in operating systems and dependencies like libraries, the full ecosystem needs to be archived as well. The choice of *docker* as the underlying container environment enables capturing the full stack of dependencies. Beside providing the researchers with the same execution system during implementation of methods on smaller local machines as for the execution inside the cluster, docker permits tagging for images like version control systems. While this can not cover every scenario, e.g., due to changing processor architecture or operating system kernels, containerising software is a huge step forward for reproducible experiments.

Docker images may be stored, tagged and frozen internally inside the CRC compute cluster registry storage. Work in progress includes meta-data extensions to mark images and containers as *permanent* to select them for archiving and prevent accidental deletion.

Freezing complete software setups not only enhances the evolution of software implementations inside the CRC, but future publishing of containerised images is an easy way to redistribute and share software implementations of the novel algorithm implementations. Researchers outside the CRC may test and evaluate software without complicated installation instructions, which may even collide with the local operating system specificities. Publication of images using a DOI will also be evaluated during the third phase.

In addition, the process of capturing data is rarely completed. Data sets change and grow, errors are fixed, new attributes are added, etc. Reproducibility demands data sets to be frozen once they

are used for research, and that these be associated with publications and analysis software. Issuing a DOI for public data sets already provides the basic means. But as speed of research increases with fast evaluation cycles on growing data, more flexible solutions especially for internal data sets are needed. Work in progress includes the evaluation of different methods for providing version control system functionality to data. A roll-out for all members of the CRC is planned for the third phase. After completion of a working solution, this will be extended for full data analysis pipelines, integrating data versioning with software versioning for end-to-end repeatable research results.

**Project-specific data sets and data management** The projects in the CRC analyse and in several cases provide data from a multitude of sources and in numerous formats. Mainly, two distinctive types of sources separate the data sets available to the CRC: Self-generated data through experimental setups or simulations, and externally provided data sets via open repositories and industrial or scientific collaborations. For the former, typically a publication under an open source data licence is aimed for, while for the latter, access to normally restricted data provides enormous possibilities for algorithmic research on this data. The projects C3 and C5 in particular provide early access to data generated in the huge physical collaborations at IceCube, MAGIC, FACT, CTA and CERN. The following descriptions provide an overview on some of these different data sets and their availability.

In 2018, A1 published a new data set with Android load profile data under the Open Data Commons Attribution Licence. In more than 200,000 records the data contains system call traces during typical application usage profiles like web browsing, email, route planning, music listening and video playback. Data is registered and can be accessed via <http://dx.doi.org/10.17877/fcq6fwzegw>.

Project A4 was in charge of creating the PhyNodes as the large-scale embedded testing platform. The nodes feature a Texas Instruments ultra-low-power IC. To enable averaging modelling of energy harvesting converter ICs several data sets have been captured measuring voltage and current levels. These data sets are publicly made available under the ODbL licence, i.a. at <http://dx.doi.org/10.17877/fd72o6mn0g> and evaluated in the upcoming publication at the international Symposium on Power Electronics, Electrical Drives, Automation and Motion [A4/134].

Project A6 maintains a publicly available collection of benchmark data sets for graph classification at <http://graphkernels.cs.tu-dortmund.de>. The repository contains 53 graph data sets with class labels, which are frequently used for the comparison of graph kernels. The data sets come from different domains such as chem- and bioinformatics, social network analysis and pattern recognition or were synthetically generated. The repository also includes two data sets published by the project A6. A synthetic data set of graphs with continuous annotations was generated for the article [A6/176]. Another data set consists of graph based representations of cuneiform signs and was created for the paper [A6/B2/180].

For the proposed third phase, project B3 gained access to several data sets typically not available to the general public. The German multinational car manufacturer BMW provided project B3 with two time series data sets from the quality testing of a welding process. The sizes of these sets are respectively 40 and 104 MB. In addition, the engineering and electronics company Bosch provided project B3 with data from two use cases. The first use case consists of an injector quality prediction. The data set for this case is a time series of different production parameters and quality results on different testing points in the injectors. The data size is approximately 138 MB. The second case study is components status prediction using ECU signals. The data set for the second case is a collection of a number of input parameters mapped to the final status of the component represented by a categorical variable. The data set is split into three files from different profiles and its total size is 350 MB. Provided data from both companies is confidential, but can be used by B3 members.

A collaboration with the CRC 837 given by the Mercur project *Synthese von maschinellem Lernen und numerischer Simulation zur Echtzeitsteuerung von Tunnelvortriebsprozessen* investigates the use of machine learning for real-time prediction. There the FE-based simulation data for surface points settlements calculation in a mechanised tunnelling process is used to evaluate one of the active learning frameworks of B3. The same data is used to evaluate the work that will be developed for real-time learning and model adaptation and online management of many models. The data is a collection of 160 simulation scenarios (417.4 MB) reporting a time series of surface points settlements records and the simulation time series inputs. The data can be used for publications and for students thesis work after prior agreement with our collaborators from the Institute for Structural Mechanics of the Ruhr University Bochum.

Many aspects in milling process optimisation are considered such as process stability prediction, real-time cutting forces prediction in long running NC-milling processes, tool wear prediction, etc. B3 already started collecting both simulation and real process data from the virtual machining group of the institute for machining (ISF)-TU Dortmund. Both simulation and sensor data are time series data reporting the milling process characteristics. The data is collected with sample frequency of 20KHz, and an amount of 95 GB of data is acquired. The data sets are confidential and can only be used by B3 members for research work and joint publications.

In project C3 data from the astroparticle telescopes IceCube, MAGIC, FACT and CTA are used. These data are owned by the international collaborations operating the telescopes. The data (on the order of petabytes) are first stored by the collaborations on different levels of processing and filtering and published on different time scales only after finishing the collaboration-internal physical analyses. Via the Astroparticle Group the project C3 has the data access rights of a collaborator.

Primary sources of data for C5 stem from the LHCb experiment. The data consist of raw data which contain information from the sensors of the individual subdetectors as well as information reconstructed from the raw data such as track momenta or other properties of reconstructed particles. In addition, simulated data are used which are similar to the experimental data but also contain information on the simulated ground truth. LHCb intends to make the analysis data public, following a two-stage approach. Five years after data have been taken, 50% of analysis level data will become public: the other 50% will follow five years later. The data will be made public via the Open Data portal <http://opendata.cern.ch/>.

#### 1.4.4 Knowledge transfer and public relations

TU Dortmund University is a vital centre for the transfer of knowledge in the widest sense of the word. Dortmund's university campus is home to one of Europe's largest technology parks, and a sizable number of start-ups and companies – including RapidMiner, the leading provider of machine learning solutions – have originated with researchers from TU Dortmund University. Fostering entrepreneurship and transfer are a strategic goal of the university and its Center for Entrepreneurship and Transfer. The transfer of knowledge into industry is successful as befitting a technical university. Finally, TU Dortmund University takes informing and engaging the public to heart, maintaining an external exposition space at the heart of the city of Dortmund and hosting a number of formats to bring research to the public.

**Centre for Entrepreneurship & Transfer CET** The TU Dortmund University initiates, promotes, and honours business start-ups and the transfer of knowledge and technology from science. Since 2017, all start-up and transfer activities have been pooled in the central university service facility Centrum für Entrepreneurship & Transfer CET. The CET actively supports students, graduates and employees in the development and implementation of business ideas, as well as in the evaluation

and marketing of technical inventions and patents. The CET also coordinates the official job portal “Stellenwerk” of the TU Dortmund University, the “StartUp.InnoLab - Westphalian Ruhr Area” and is the office of the board of trustees of the tu>startup STIFTUNG in the Dortmund Foundation. The TU Dortmund is also a partner in the patent exploitation agency of the NRW universities PROVendis and the Technology Centre Dortmund. At the end of 2017, TU Concept Projekt- und Beteiligungs-GmbH was also founded as a wholly-owned subsidiary with which TU Dortmund University can also participate financially in innovative spin-offs for the first time. The overall aim is to promote the progress of the TU Dortmund University, the city and the region through research, qualification and knowledge and technology transfer.

**Transfer of Knowledge into Industry** CRC 876 presents research results to representatives of potential users at many events.

For the communication of B3 results and those of a former collaboration with SMS Siemag on the BOF converter, Katharina Morik gave talks in practice-oriented conferences: 6th and 7th *gwi (Gas- und Wärme International)-Praxistagung, Industrie 4.0* of the German engineering federation (Verband Deutscher Maschinen- und Anlagenbau), two workshops at Bayer AG in Leverkusen for several companies, the ABB research centre, and the Audi training colloquium, the Voestalpine Digitalization Day, all with a number of participants between 30 and 90.

On the one hand, B4 is in contact with suppliers for cars concerning the optimal point in time for transferring data gathered in the car. On the other hand, Christian Wietfeld was moderating regional conferences in Köln and Paderborn on energy, transportation and communication for around 80 industrial participants. This introduces CRC results to interested industries. In particular, his keynotes at the 21st and the 22nd broadband forum, organised by CPS.HUB NRW communicated knowledge from CRC 876 to a larger audience. At the “Münchner Kreis” conference “Neue Produkte in der digitalen Welt” in 2016 he gave a presentation on cyber-physical systems for logistics and at the Acatech symposium “iCity and Intelligent Logistics” he gave the keynote address titled “From Smart Cities to Intelligent City Logistics”.

In 2017 Christian Wietfeld also addressed students in a lecture series of the CampusLab Essen and gave a presentation on 5th generation mobile devices as a foundation for the internet of things.

Project C1 informed relevant audiences about results of the CRC 876 in talks by Alexander Schramm: At the “Klinikkongress Ruhr 2018” in Essen, he spoke on the potential of artificial intelligence in medicine for early diagnosis as well as treatments. Furthermore, he gave a presentation at the SmartHealthData.NRW conference in Dusseldorf in 2018 on artificial intelligence in general, and machine learning in particular, as a technological enabler, highlighting its possible future benefits for personalised medicine.

**Informing the Public** Artificial intelligence and data as a valuable resource increasingly became topics of public interest and public debate during the last phase. The CRC 876 participated in these discussions in different media and events.

Katharina Morik, together with Walter Krämer, edited the book “Daten – wem gehören sie, wer speichert sie, wer darf auf sie zugreifen?” published by the Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste. She authored a chapter under the same title, in which she discusses the questions of who owns data and who is allowed to store or access data.

Members of the CRC participated in panel discussions and interviews on the topic. For instance

- Between November 2015 and April 2016, Katharina Morik, Kristian Kersting, Christian Sohler and Michael TenHompel gave interviews on big data for NRWision in an interview

series called “THINK BIG”. Videos of these interviews are available online (<https://www.nrwvision.de/mediathek/suche/?query=thinkbig>).

- Katharina Morik discussed big data and the tension between individual safety and continuous surveillance at a public panel discussion in Moers organised by Pedro Marron (University of Duisburg-Essen) in November 2016.
- Peter Marwedel presented the CRC topics at the U.S.-German Workshop on the Internet of Things (IoT)/Cyber-Physical-System (CPS). The aim of the workshop was to prepare for intensified German-American cooperation in the thematic area of this workshop. Peter Marwedel was invited as one of three German university representatives.<sup>16</sup>
- A panel discussion on Big Data - Small devices has been held in New York in March 2016, targeting a broader audience. The collaborative research centre CRC 876 has been represented by Katharina Morik (Resource-Aware Data Science), Wolfgang Rhode (Science for Science) and Kristian Kersting (Not so Fast: Driving into (Mobile) Traffic Jams), while as a local presenter Claudia Perlich (Dstillery) gave her view on big data analysis. The discussion itself was moderated by Tina Eliassi-Rad (Northeastern-University/Rutgers University).<sup>17</sup> The event has been organised by the New York German Center for Research and Innovation and was co-sponsored by Deutsche Forschungsgemeinschaft and University Alliance UA Ruhr.
- Katharina Morik gave an interview for *gwi gaswärme international* in 2016 on the importance of the “data revolution” for the German economy.
- In 2017, she also spoke at the *Deutscher Gewerkschaftsbund* about the potential of big data and machine learning for improving work and working conditions in industry.

Another topic that has received public interest is the internet of things with its ever-increasing number of cyber physical systems connected to the internet – often using mobile communication. Christian Wietfeld und Michael ten Hompel addressed the challenges posed by this development in a joint interview in “Best Practice” in March 2017. Christian Wietfeld analysed the importance of modern mobile communication standards for interconnected cars with “Elektronik Praxis” in 2015.

The CRC 876 also frequently appeared in mass media, for instance with appearances of Katharina Morik on television *ZDF-Volle Kanne* in 2017 or by Christian Wietfeld who, in 2016, explained GPS navigation in the regional newspaper *Ruhrnachrichten* in the article “Die W-Frage: Wie funktioniert ein Navigationsgerät?”.

Further appearances, amongst others, included B4 traffic prognosis, Michael Schreckenberg, (UDE erforscht den Stau in Innenstädten, Rheinische Post); B2, Roland Hergenröder, about fast virus detection (Der schnelle Viren-Test. Forschung aus Dortmund: Mobiler Sensor erkennt gefährliche Erreger in Minuten, Ruhrnachrichten); A4 Michael ten Hompel/Christian Wietfeld real time communications (Echt Zeit für Echtzeit, T-Systems Best Practise).

Another important aspect of informing the public is education in digital classrooms. In 2017, Kristian Kersting and Katharina Morik gave lectures on machine learning as part of a massive open online course (MOOC) initiated by acatech.

---

<sup>16</sup><https://www.youtube.com/watch?v=vzNVUQp96Ic>

<sup>17</sup><https://vimeo.com/158275232>

## 1.5 Other sources of third-party funding for principal investigators

Principal investigator	PRJ	Project title	Funding period	Funding agency
Chen	A1, A3	Suspension-Aware Designs and Analyses for Real-Time Embedded Systems (Sus-Aware)	2011–2018	DFG CH 985/12-1
Chen	A1, A3	Design and optimisation of Non-Volatile One-Memory Architecture (NVM-OMA)	TBD	DFG CH 985/13-1
Chen	A1, A3	Partitioning, Scheduling, Spinning, Suspension, synchronisation, and Locking Protocols in Real-Time Systems (PS4Lock)	TBD	DFG CH 985/14-1
Deuse	B3	AIM	2016–2019	BMBF 02L14A160
Deuse	B3	InDaS	2017–2019	BMBF 01IS17063A
Deuse	B3	KoMPI	2017–2019	BMBF 02P15A066
Deuse	B3	PHASE	2017–2019	ZF4101110LF7
Deuse	B3	ROBOTOP	2017–2020	01MA17009H
Deuse	B3	Sysmag	2017–2019	AIF 19185 N/1
Deuse	B3	VariPro	2017–2019	IGF-Vorhaben 19683 N
Hergenröder	B2	AntiThromb	2018–2020	BMBF: KMU-NetC 03VNE2091F
Hergenröder	B2	NanoFilter	2018–2018	DAAD: 57403573
Hergenröder	B2	MRT-Filter	2017–2020	DFG: LA 1134/8-1
Ickstadt	C4	Statistical Methods for Damage Processes Under Cyclic Load	07/2013–06/2021	CRC 823, Project B5
Ickstadt	C4	Polymorphic Uncertainty Modelling for Stability Quantification of Fluid Saturated Soils and Earth Structures	01/2017–06/2020	DFG Project IC 5/6-1 in Priority Programme 1886
Liebig	B4	VAVEL	2016–2018	EU-H2020 Grant Agreement No 688128

Principal investigator	PRJ	Project title	Funding period	Funding agency
Morik	A1, B3, C3, Z	Variety, Veracity, Value: Handling the Multiplicity of Urban Sensors (VaVel)	2016–2020	Horizon2020-688380
Morik	A1, B3, C3, Z	Modellierung von Themen und Strukturen religiöser online-Kommunikation	2016–2018	MERCUR, PR-2015-0046
Morik	A1, B3, C3, Z	Synthese von maschinellem Lernen und numerischer Simulation zur Echtzeitsteuerung	2017–2019	MERCUR, PR-2016-0039
Morik	A1, B3, C3, Z	Kompetenzzentrum maschinelles Lernen Rhein Ruhr – ML2R	2018–2022	BMBF
Mutzel	A6	GraBaDrug	2014–2019	DFG: MU 1129/10-2
Mutzel	A6	GRK 1855	2013–2018	DFG: Member of the GRK 1855/1
Rahmann	C1	UA Ruhr Professorship Computational Biology	2014–2019	MERCUR Pe-2013-0012
Rahmann	C1	OsteoSys	2016–2019	EFRE-0800427
Rahnenführer	A3	E:Top Translationsphase: LivSys	2016–2019	BMBF 031L0119B
Rahnenführer	A3	E:Top Translationsphase SysDT	2016–2019	BMBF 031L0117E
Rahnenführer	A3	StemNet	2017–2019	BMBF 01EK1604C
Rahnenführer	A3	Identification of survival models that are prognostic across cohorts and stable regarding variable selection with methods of model-based optimization	2016–2019	RA 870/7-1
Rhode	C3, MGK	Berechnung und experimentelle Analyse der Myon-Wirkungsquerschnitte	2017–2020	DFG, RH 35/9-1
Rhode	C3, MGK	Verbundprojekt IceCube	2017–2020	BMBF, 05A17PEA
Schreckenberg	B4	MEC-View	2017–2019	19A16010H
Schramm	C1	Role of the tyrosine kinase TrkA and TrkB in checkpoint activation and DSB repair	2018–2021	DFG GRK 1739

Principal investigator	PRJ	Project title	Funding period	Funding agency
Schramm	C1	Modelling primary tumor metabolism in neuroblastoma to identify central nodes for therapeutic intervention	2017–2019	BMBF 01ZX1307C
Schramm	C1	Interaction of TrkA and MYCN in neuroblastoma	2017–2019	Sander-Stiftung 2016.119.1
Schramm	C1	Chemo-genetic interference with Survivin functions in embryonal tumors	2018–2018	Wegener Stiftung Project Nr. 29
Sohler	A2, C4	ERC Starting Grant	2012–2018	SUBLINEAR 307696
Sohler	A2, C4	MERCUR Project: “LPN-Krypt: Das LPN-Problem in der Kryptografie”	2017–2019	Pr-2016-0045
Spaan	C5	LHCb: Quark-Flavor-Physik am LHC: Flavorsignaturen in Theorie und Experiment - LHCb: Run 2 und Upgrade	2015–2018	05 H15PE15CL1
ten Hompel	A4	Innovations Lab (Innovationslabor)	2016–2019	KIT PTKA 02P16Z200
ten Hompel	A4	ConnectedFactories	2016–2019	EU 723777
ten Hompel	A4	Safelog	2016–2019	EU 688117
ten Hompel	A4	SENSE	2017–2020	EU 769967
ten Hompel	A4	Clusters 2.0	2017–2020	EU 723265
ten Hompel	A4	LEGOLAS	2017–2020	PT ETN IT-1-2-014a
ten Hompel	A4	InDaSpacePlus	2017–2020	BMBF 01IS17031
ten Hompel	A4	L4MS	2017–2021	EU 767642
Teubner	A2, C5	DFG Priority Program 2037 · MxKernel	2017–2020	TE 111/2-1
Weichert	A6, B2	ADJUTANT	2018–2020	AiF: ZF4119002DB7
Weichert	A6, B2	InÜDosS	2018–2020	FOSTA: P 1326/17/2018
Weichert	A6, B2	CuKa	2018–2021	DFG: WE 5036/4-1
Wiederkehr	B3	UA Ruhr-Professur “Virtual Machining”	2017–2022	MERCATOR Pe-2016-0024

Principal investigator	PRJ	Project title	Funding period	Funding agency
Wiederkehr	B3	Modelling and Simulation of the NC Grinding Process for the Controlled Generation of Workpiece Surfaces Under Consideration of Tool Topography and Wear	2017–2019	DFG WI 4762/5-1
Wiederkehr	B3	Adaption Intelligence of Factories in a Dynamic and Complex Environment (Spokesperson: Prof. Dr. Jakob Rehof)	2016–2020	DFG GRK 2193
Wiederkehr	B3	Stochastic Modeling of the Interaction of Tool Wear and the Machining Affected Zone in Nickel-Based Superalloys, and Application in Dynamic Stability	2018–2021	DFG WI 4762/7-1
Wietfeld	A4, B4	DFG-Forschergruppe 1511	2014–2018	Wi 3751/1-1, Wi 3751/2-1
Wietfeld	A4, B4	BERCOM	2015–2018	BMBF 13N13741
Wietfeld	A4, B4	OPUS	2017–2020	EFRE-0800885
Wietfeld	A4, B4	LARUS	2017–2019	BMBF 13N14133
Wietfeld	A4, B4	IDEAL	2016–2019	BMWi 03ET7557A
Wietfeld	A4, B4	CPS.HUB/NRW	2015–2018	EFRE-0400008
Wietfeld	A4, B4	InVerSiV	2016–2019	EFRE-0800422
Wietfeld	A4, B4	AutoMat	2015–2018	H2020 644657



## 2 Existing funds and requested funds

### 2.1 Existing funds

#### 2.1.1 Overview of existing funds for direct costs

Funding period	Core support provided by applicant university/universities	Core support provided by other participating institutions	Other funds	Total
2015	103.5	31.0	0	134.5
2016	90.5	31.0	0	121.5
2017	90.5	31.0	0	121.5
2018	88.5	31.0	0	119.5
Ending funding period	373.0	124.0	0	497.0
2019	73.5	37.9	0	111.4
2020	73.5	26.9	0	100.4
2021	73.5	26.9	0	100.4
2022	73.5	26.9	0	100.4
New funding period	294.0	118.6	0	412.6

(All figures in thousands of euros)

#### 2.1.2 Overview of existing staff

Category	Number of persons	
	at the applicant university/universities	at other participating institutions
Professors	13	3
Junior research group leaders	1	0
Postdoctoral researchers	5	5
Doctoral researchers	23	4
Other research staff	0	0
Non-research staff	14	4
Student and graduate assistants	12	1

### 2.1.3 List of existing instrumentation

PRJ	Description of instrumentation	Year of purchase	Cost of purchase	Source of funding	Location
A1	Hardware in the loop emulator	2014	43.0	University	TU Dortmund
A1	Dell Power Edge R430 server	2017	11.0	University	TU Dortmund
A1	Prototyping multi-core system	2014	23.5	University	TU Dortmund
A4	AIrena	2018	20.0	University	TU Dortmund
A4	Spectrum Analyser (R&S)	2017	40.9	University	TU Dortmund
A4	VR Studio	2017	70.0	University	TU Dortmund
A4	8x N210 Software Defined Radios	2016	19.4	BMBF Project TAMIS (project completed, now property of institute)	TU Dortmund
A4	Electric car (E-Smart)	2013	23.2	EU Project Open ECOSPhERE (project completed, now property of institute)	TU Dortmund
A4	Laser Projection system	2017	80.0	University	TU Dortmund
A4	6x N210 Software Defined Radios	2017	36.0	InnovationsLabor	TU Dortmund
A4	Vector Network Analyzer NVB-4 and peripherals	2013	41.0	Fraunhofer IML	TU Dortmund
A4	Drone swarm	2017	35.0	University	TU Dortmund
A4	Sensor floor (530x CC1350)	2017	15.4	University	TU Dortmund
A4	Production line for circuit board prototyping	2013	92.0	Fraunhofer IML	TU Dortmund
A4	PhyNetLab	2016	70.0	SFB876 and Fraunhofer IML	TU Dortmund
A4	Current-Analyzer	2017	58.9	University	TU Dortmund
A4	Vicon motion tracking in an industrial facility	2017	244.0	InnovationsLabor	TU Dortmund
A4	Robots	2017	119.5	University	TU Dortmund
A4, B4	FSVR - Real Time Spectrum Analyzer	2011	118.5	EU/ZIEL2 Projekt AVIGLE (project completed, now property of institute)	TU Dortmund
A4, B4	AnokiWave AWMP-0129 5G Pencil Beam Antenna	2017	50.5	University	TU Dortmund
A4, B4	Propsim C8 radio channel emulator	2011	229.1	BMBF Project SPIDER (project completed, now property of institute)	TU Dortmund

(All figures in thousands of euros)

PRJ	Description of instrumentation	Year of purchase	Cost of purchase	Source of funding	Location
A4, B4	Measuring vehicle for a transportable mobile communication lab	2018	59.7	University	TU Dortmund
A4, B4	NI PXI System - Software Defined Radio System	2017	74.9	University and EFRE Project InVerSiV	TU Dortmund
A4, B4	OptiTrack Motion Capture System	2016	14.1	University	TU Dortmund
A4, B4	CMW500 Upgrades for LTE Release 12 and eMTC	2016	256.0	EFRE Project CPS.HUB/NRW and University	TU Dortmund
A4, B4	LTE base station and core emulator	2013	78.0	BMBF Project ANCHORS (project completed, now property of institute)	TU Dortmund
A4, B4	mmWave Transceiver System	2017	190.2	University	TU Dortmund
A4, B4	CMW500 - Wideband Radio Communication Tester	2011	188.2	BMBF Project SPIDER (project completed, now property of institute)	TU Dortmund
B2	Beckman, Optima L-90K   Ultracentrifuge for VLP purification/characterisation	2011	25.9	ARTES GmbH, inventory	ARTES Biotechnology GmbH
B2	Malvern LM10 Instrument with pump	2013	25.0	ISAS	Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.
B2	BioRad, ChemDoc MP   gel/blot imaging and evaluation system for VLP characterisation	2016	28.7	ARTES GmbH, inventory	ARTES Biotechnology GmbH
B2	GE Health Care, Äkta Purifier   chromatography system for VLP purification	2003	25.2	ARTES GmbH, inventory	ARTES Biotechnology GmbH
B2	Shimadsu, LC20   HPLC system (SEC, UV detection) for VLP characterisation	2008	21.2	ARTES GmbH, inventory	ARTES Biotechnology GmbH
B2	Beckman, DelsaMax Core   light scattering system for nanoparticle characterisation	2014	26.1	ARTES GmbH, inventory	ARTES Biotechnology GmbH
B2	Tecan, Genios   microtiter plate ELISA reader for VLP characterisation	2003	16.7	ARTES GmbH, inventory	ARTES Biotechnology GmbH

(All figures in thousands of euros)

PRJ	Description of instrumentation	Year of purchase	Cost of purchase	Source of funding	Location
B2	Bio-Lab   Handling of any biological samples, permits to perform works at safety level 1	2017	28.0	ISAS	Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.
B2	Newport Optical Table with accessories	2017	10.0	ISAS	Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.
B3	Deckel Maho DMU50 eVolution 5-Axis Machining Centre	2004	250.0	BMBF (Institute of Machining Technology, Prof. D. Biermann)	TU Dortmund
B3	Deckel Maho HSC75 Linear 5-Axis Machining Centre	2010	395.0	University (Institute of Machining Technology, Prof. D. Biermann)	TU Dortmund
B3	Kistler 9255B Triaxial Force Dynamometer	2008	38.9	University (Institute of Machining Technology, Prof. D. Biermann)	TU Dortmund
B3	Brüel & Kjær Pulsanalyser IDA DIE Type 3035/Type 7539A	2005	28.8	University (Institute of Machining Technology, Prof. D. Biermann)	TU Dortmund
B3	Alicona InfiniteFocus G5 Optical Form and Surface Roughness Measurement Device	2016	230.0	DFG (Institute of Machining Technology, Prof. D. Biermann)	TU Dortmund
C1	PowerVault Storage System 460 TB	2014	38.0	University	University Hospital Essen
C1	Safety-level S2 laboratory equipment (devices for PCR and real-time PCR, flow cytometry, safety hoods, cell sorter, established MinION platform)	2018	300.0	University	University Hospital Essen
C1	PowerEdge R715 Compute Server	2014	11.0	Mercator Research Center Ruhr	University Hospital Essen

(All figures in thousands of euros)

## 2.2 Previous and requested funds

### 2.2.1 Overview

Financial year/funding period	Funding for					Total
	Staff	Direct costs	Instrumentation	Fellowships	Global funds	
2015	1,874.1	144.3	0.0	14.4	47.2	2,080.0
2016	1,908.8	136.7	0.0	14.4	67.2	2,127.1
2017	1,952.6	136.7	0.0	14.4	127.2	2,230.9
2018	1,987.7	136.7	0.0	14.4	347.2	2,486.0
Ending funding period	7,723.2	554.4	0.0	57.6	588.8	8,924.0
2019	1,994.0	163.4	0.0	14.4	147.2	2,319.0
2020	2,069.2	144.6	0.0	14.4	147.2	2,375.4
2021	2,069.2	143.7	0.0	14.4	147.2	2,374.5
2022	2,069.2	142.5	0.0	14.4	147.2	2,373.3
New funding period	8,201.6	594.2	0.0	57.6	588.8	9,442.2

(All figures in thousands of euros)

## 2.2.2 Overview of funds requested for staff

PRJ	2019					2020					2021					2022				
	Postdocs	Doctoral researchers	Other research staff	Non-research staff	Student assistants	Postdocs	Doctoral researchers	Other research staff	Non-research staff	Student assistants	Postdocs	Doctoral researchers	Other research staff	Non-research staff	Student assistants	Postdocs	Doctoral researchers	Other research staff	Non-research staff	Student assistants
A1	0	2	0	0	0	2	0	0	0	0	0	2	0	0	0	0	2	0	0	0
A2	0	2	0	0	0	2	0	0	0	0	0	2	0	0	0	0	2	0	0	0
A3	0	2	0	0	0	2	0	0	0	0	0	2	0	0	0	0	2	0	0	0
A4	0	2	0	0	0	2	0	0	0	0	0	2	0	0	0	0	2	0	0	0
A6	0	2	0	0	0	2	0	0	0	0	0	2	0	0	0	0	2	0	0	0
B2	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0
B3	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0
B4	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0
C1	0	2	0	1	0	0	2	0	1	0	0	2	0	1	0	0	2	0	1	0
C3	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0	0	3	0	0	0
C4	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0
C5	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0	0	2	0	0	0
MGK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Z	2	0	0	1	16	2	0	0	1	16	2	0	0	1	16	2	0	0	1	16
Total	3	26	0	2	16	3	26	0	2	16	3	26	0	2	16	3	26	0	2	16

In the individual projects, funding for doctoral researchers is requested for 100% positions. Single exception is project C1, where the standard 65% for doctoral researchers in clinical positions and a 100% technician is requested.

## 2.2.3 Overview of funds requested for instrumentation

The Collaborative Research Centre 876 does not request any funding for major research instrumentation.

## 2.3 Upkeep of laboratory animals

The Collaborative Research Centre 876 does not keep any laboratory animals.

### 3.1 General information about Project A1

### 3.1.1 Project title:

Data Mining for Ubiquitous System Software

### 3.1.2 Research area(s):

409-05 (Interactive and Intelligent Systems), 409-07 (Embedded Systems)

### **3.1.3 Principal investigator(s)**

Morik, Katharina, Prof. Dr., 14.10.1954, German

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

Phone: 0231-755-5100  
E-mail: katharina.morik@tu-dortmund.de

Chen, Jian-Jia, Prof. Dr., 09.05.1978, Taiwanese

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

Phone: 0231-755-6078  
E-mail: jian-jia.chen@tu-dortmund.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes	(x) no
2.	clinical trials	( ) yes	(x) no
3.	experiments involving vertebrates.	( ) yes	(x) no
4.	experiments involving recombinant DNA.	( ) yes	(x) no
5.	research involving human embryonic stem cells.	( ) yes	(x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes	(x) no

## 3.2 Summary

Project A1 develops data analysis methods for embedded systems. Learning methods are to consume resources minimally so that they can be executed by embedded systems. A1 contributes to these overall goals by developing the theoretical basis of machine learning under resource constraints. In particular, it investigates *machine learning models* that formalise implicit functional relations within the data. More precisely, a model, e.g., from the class of linear models, exponential families, decision trees, or convolutional networks, specifies the general formal basis of any machine learning method. It determines implicitly what is learned (e.g., classifier, probability density, clustering model), and from what it is learned (e.g., real-valued vectors, categorical data, count data). Instances of a particular model are assessed by a quality measure (e.g., root mean square error,  $F_\beta$ -score, correct classification rate, or the likelihood). In A1, we extend this notion of quality by adding simplicity of model representation and simplicity of inference, where simplicity is relative to the underlying hardware architecture. Thus, the quality of a model may comprise its memory consumption, numerical precision, parameter integrality, or approximability. Since learning itself is often expressed by one or more optimisation problems, resource constraints are added to the quality measure via regularisation. Several algorithms exist that can find (approximate) solutions to optimisation problems. Finally, the algorithms are executed on a specific hardware platform. Each of these layers, i.e., platforms, algorithms, models, is studied in the research centre. Project A3 starts with the execution on a certain platform, project A2 moves on to algorithms, and project A1 investigates the models of learning.

There is a plentitude of heuristics that algorithmically save resources. However, heuristics most often come without guarantees. In the A1 project, we specify restrictions at the model level and analyse the impact of the restriction on the quality. Moreover, a model of learning with proven guarantees makes explicit the link from the restrictions of the model to those of constrained hardware. It is a long way to achieve such models with guarantees. For the class of learning within the exponential family, we succeeded in developing spatiotemporal random fields, integer Markov random fields, and a novel quadrature based on first principles, each focusing on specific constraints. Proven properties of the novel models conclude the basic research on exponential families. At the same time, applications in other projects have already shown their practical usefulness. For top-down decision trees, the journey has just begun, and first publications show the potential of linking tree models of learning to FPGAs.

Olaf Spinczyk has moved to the University of Osnabrück and Jian-Jia Chen joins the A1 project. First publications already show results of the collaboration. Memory architectures and their impact on machine learning algorithms will be a focus of the investigations. Sampling and aggregation of data are closely related to distributed data analysis. A resource management that builds on statistical guarantees of learning algorithms concludes the work on real-time distributed data analysis in the third funding phase of A1.

## 3.3 Project progress to date

Work in the second phase is structured along the lines of a machine learning work flow, starting with data acquisition and ending with the evaluation. The work on probabilistic graphical models that started in the first phase has matured in the second phase to deliver sound theoretical results. Work on Random Forest model application and FPGAs has started.

### 3.3.1 Report and current state of research

**Resource-saving data acquisition (WP1):** Data acquisition for machine learning in the context of operating system (OS) kernels or in OS user-level services must be performed with minimal overhead – especially on resource constrained mobile platforms. We created a novel stream-oriented cross-address-space data acquisition and near-source data processing algorithm and implemented it as a research prototype named kCQL. The key features of kCQL are a safe, declarative, high-level query language for OS data, the ability to capture OS state as well as runtime events, database-like query processing close to the data sources, and a cross-address-space transport mechanism to combine data from different OS components.

Previously existing solutions fall into two categories. Low-level mechanisms, such as SystemTap or DTrace require a lot of manual programming effort and are thus error-prone. The alternative high-level mechanism of PicoQL cannot deal with events. Both kinds of tools are unable to combine data across address space boundaries. By the combination of kCQL’s unique features, data acquisition in OS kernels became much simpler and very efficient. For example, statistics on incoming network packets in GBit ethernet can now be created on the fly in the kernel address space. Only the aggregated data needs to be transported to the user space for further processing.

Another means of decreasing the data acquisition efforts is domain adaptation. Since annotating data with a label is a time-consuming task, it is relevant to adapt learning results from one data set A with labels to another one B without labels. In order to succeed, we need to find a common ground for A and B. A matrix that projects the distributions of the data sets onto a subspace forms (together with an inner product) a Stiefel manifold, i.e., a set of orthonormal vectors. Its nice properties are the local linearity and a metric on each manifold that measures the distance between two points on the manifold. The optimisation problem is to minimise the maximum mean discrepancy (MMD) between the distribution of two domains. We run the stochastic gradient descent (SGD) directly on the matrix manifold. For text classification, we have shown that the SGD steps compel the solution to stay on the Stiefel manifold. This manifold encompasses projection matrices of word vectors onto low-dimensional latent feature representations, which allows us to interpret the results: the rotation magnitude of the word vector projection for a given word corresponds to the importance of that word towards making the adaptation. Beyond this interpretability benefit, experiments could show that the Stiefel manifold method performs better than state of the art methods [A1/25].

**Data streams (WP2):** Project A1 has developed an environment for the analysis of streaming data, the **streams** framework. It has become widely used, e.g., in the European projects Vista-tv<sup>1</sup> and Insight<sup>2</sup>. It also became the basis of the FACT tools, which analyse Cherenkov telescope data in project C3. Physicists continue to use and enhance the software. Student projects used the **streams** framework in order to store, clean, and analyse streaming data according to the lambda-architecture which combines map-reduce programming (batch layer) and direct execution on data streams (speed layer). Hence, the Collaborative Research Centre 876 also had an impact on teaching modern big data methods<sup>3</sup>. Christian Bockermann’s dissertation (with distinction) presents concepts and applications of **streams**.

Sensor data often arrive with a high velocity so that they have to be aggregated, summarised, or sampled before storage. The stored summary should enable human data exploration as well as machine learning. Approaches that select observations for further processing are investigated in projects A1 and A2. A2 investigates diverse coresnet constructions and analyses the lower bounds of space complexity and impossibility results for, e.g., logistic regression, showing the importance of coresnet construction for the design of streaming algorithms. In A1, we enhance a particular algorithm that summarises data streams: the Sieve-Streaming algorithm [2] efficiently computes

---

<sup>1</sup><http://vista-tv.eu/things-to-read/publications/index.html>

<sup>2</sup><http://www.insight-ict.eu/sites/default/files/deliverables/D2-3.pdf>

<sup>3</sup>Project group 594 *Big Data* 2016 <http://www-ai.cs.uni-dortmund.de/LEHRE/PG/PG594/pg594-endbericht.pdf>

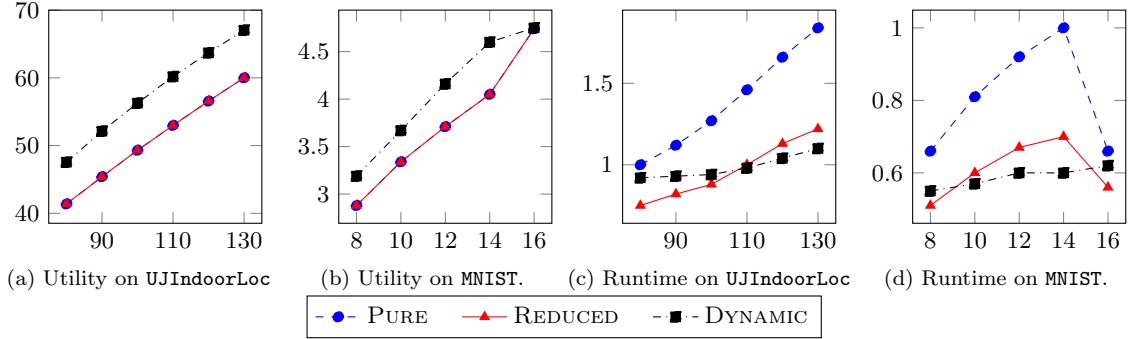


Figure 3.1: Experiments on UJIndoor location data and MNIST for different summary sizes. The first two plots depict the maximum utility value (higher is better) and the second two displays show the runtime (in milliseconds) per element (smaller is better).

summaries of streaming data by using a submodular function as a quality measure. The main idea of Sieve-Streaming is to store multiple candidate summaries and only include those items from the stream that seem to add significant gain to the quality of a summary. Threshold values for the quality gain bound the error of the summary. Submodular functions formalise diminishing returns: Adding a new element to a smaller set will increase the utility value more than adding the same element to a larger set. More formally, for  $S \subseteq V$  the gain of  $e \in V$  is defined by [15]

$$\Delta_f(e|S) = f(S \cup \{e\}) - f(S)$$

Furthermore, we call  $f$  submodular iff for all  $A \subseteq B \subseteq V$  and  $e \in V \setminus B$  it holds that

$$\Delta_f(e|A) \geq \Delta_f(e|B)$$

Both the entropy and the information gain of Gaussian processes are submodular functions that capture the notion of small, but expressive summaries in the context of kernel methods [19]. We have improved the bounds by directly exploiting the aforementioned quality functions. The reduced Sieve-Streaming reduces the number of candidate summaries significantly. This, in turn, leads to a smaller memory footprint as well as faster execution times. Another variant, dynamic Sieve-Streaming adjusts threshold values on the fly during execution, which often results in larger utility values. We have applied our enhanced Sieve-Streaming to telescope data of C3. Exemplary results of utility and runtime for the three algorithms on the standard MNIST data set are shown in figure 3.1.

**Probabilistic Graphical Models (WP3):** The analysis of resource consumption by data analytics requires more than empirical tests of diverse algorithms – we aim at theory formation. On the one hand, we want to generalise from algorithms to their subsuming models. On the other hand, we want to investigate the model’s parts, in detail. We demonstrate the theory formation for probabilistic graphical models (PGMs) and their essential components, formalised by the exponential family of densities.

$$\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}) = \exp(\langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle - \ln Z(\boldsymbol{\theta})) \quad (3.1)$$

That is, our results are valid for any machine learning model that can be written in the form of (3.1). This includes Markov random fields, conditional random fields, logistic regression, Bayesian networks, and deep Boltzmann machines. The terms of (3.1) have been analysed and grouped by their primary resource consumption. A careful review of methods that reduce the required resources shows how to adapt the essential components of the exponential family to specific constraints. At the same time, the quality of graphical models must be kept at high standards. For the first time, a set of model properties could be stated that expresses the desired balance of quality and

resource efficiency. (1) Any probabilistic model encodes a conditional independence structure – our methods must keep this structure intact. (2) Exponential family models are theoretically well-founded. Instead of merely proposing new algorithms, our extensions and their effects must be formalised within the framework of exponential families. (3) Exponential family models are derived from first principles – our techniques must not introduce new assumptions that are incompatible with the formal derivation of the base model. (4) Our extensions should not rely on properties of particular high-level applications or tasks – this does not exclude low-level characteristics like discrete random variables or time-series data.

A new view of a classic result by Pitman [17] allowed us to show that the exponential family can be derived with *any* base. Following this result, we present the base-2 exponential family, i.e.,  $\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{x}) = 2^{\langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle - \log_2 Z(\boldsymbol{\theta})}$ , which opens the path for integer-valued PGMs. By restricting the parameter space to the positive integers, we show that the exponential function (which inherently requires real-valued arithmetic) can be replaced by a simple bit-shift operation. This reduces the required computing facilities and energy consumption considerably: In message-passing-based inference, propagating only the *bitlength* of messages

$$b_{v \rightarrow u}(x_u) = \text{bitLength} \sum_{x_v \in \mathcal{X}_v} 2^{\boldsymbol{\theta}_{(v,u)} = (x_v, x_u) + \sum_{w \in \mathcal{N}(v) \setminus \{u\}} b_{w \rightarrow v}(x_v)},$$

suffices. Here,  $b_{v \rightarrow u}(x_u)$  is the message from  $v$  to  $u$  about  $x_u$ ,  $\mathcal{N}(v)$  is the neighbourhood of vertex  $v$ , and  $\text{bitLength}(n)$  returns the position of the most significant bit in a base-2 representation of  $n$  if  $n > 0$  and 1 otherwise. We could derive a theorem that explains how the conditional independence structure influences the error of our bitlength approximation [A1/C1/29]. The theorem was improved in the PhD thesis of Nico Piatkowski. Illustrative results are shown in the two rightmost plots in figure 3.2. Numerical experiments on MSP430-based 16-bit microcontroller units show that our inference method is 250 times faster than ordinary belief propagation with double-precision arithmetic. Inference in exponential families has an exponential worst case complexity. Existing approximate inference methods reduce the computational resources by simplifying the conditional independence structure. However, this may lead to wrong conclusions drawn from the model’s prediction. Exploiting numerical approximation theory [6], we could derive a near-minimax optimal polynomial approximation to the potential function  $\exp(\langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle)$  – the resulting approximate partition function is then

$$\hat{Z}_{\boldsymbol{\zeta}}^k(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in \mathcal{X}} \hat{\exp}_{\boldsymbol{\zeta}}^k(\langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle)$$

where  $\boldsymbol{\zeta}$  is the coefficient vector of a degree- $k$  Chebyshev approximation to  $\exp$  on the interval  $[-\|\boldsymbol{\theta}\|_1, \|\boldsymbol{\theta}\|_1]$  with approximation error  $\varepsilon$ . While this idea is well known for the approximation of intractable integrals, it was unknown in the context of probabilistic inference and the exponential family. To reduce the computational complexity, we derive the *tuple density*  $\mathbb{P}_{\boldsymbol{\zeta}, \phi}(\mathbf{J} = \mathbf{j}, I = i)$  of random index tuples  $\mathbf{J}$  with random dimension  $I$ . With this, we could prove a fundamental theorem, asserting that the random variable  $\hat{Z}_{\mathbf{J}, I}^k(\boldsymbol{\theta}) = \tau \operatorname{sgn}(\boldsymbol{\zeta}_I) \prod_{l=0}^I \boldsymbol{\theta}_{\mathbf{j}(l)}$  is an unbiased estimator for  $\hat{Z}_{\boldsymbol{\zeta}}^k(\boldsymbol{\theta})$ . More precisely:

$$\begin{aligned} \mathbb{E} [\hat{Z}_{\mathbf{J}, I}^k(\boldsymbol{\theta})] &= \sum_{i=0}^k \sum_{\mathbf{j} \in [d]^i} \mathbb{P}_{\boldsymbol{\zeta}, \phi}(\mathbf{J} = \mathbf{j}, I = i) \tau \operatorname{sgn}(\boldsymbol{\zeta}_i) \prod_{l=0}^i \boldsymbol{\theta}_{\mathbf{j}(l)} \\ &= \sum_{i=0}^k \boldsymbol{\zeta}_i \sum_{\mathbf{j} \in [d]^i} \left( \prod_{l=0}^i \boldsymbol{\theta}_{\mathbf{j}(l)} \right) \sum_{\mathbf{x} \in \mathcal{X}} \left( \prod_{l=0}^i \phi(\mathbf{x})_{\mathbf{j}(l)} \right) \\ &= \sum_{\mathbf{x} \in \mathcal{X}} \sum_{i=0}^k \boldsymbol{\zeta}_i \langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle^i = \hat{Z}_{\boldsymbol{\zeta}}^k(\boldsymbol{\theta}) \end{aligned}$$

where the constant  $\tau$  does not depend on  $\boldsymbol{\theta}$ . The partition function may hence be approximated by drawing  $N$  samples from  $\mathbb{P}_{\boldsymbol{\zeta}, \phi}(\mathbf{J} = \mathbf{j}, I = i)$  – we denote this estimator by  $\hat{Z}_{\boldsymbol{\zeta}}^{N, k}(\boldsymbol{\theta})$ . A new

theorem on the approximation error tells us how the number of samples and the polynomial degree influence the approximation error. Let  $\delta \in (0, 1]$ ,  $\epsilon > 0$ ,  $N = (\log \frac{2}{\delta})\tau^2 2\|\boldsymbol{\theta}\|_\infty^{2k'} \epsilon^{-2} |\mathcal{X}|^{-2}$ , with  $(k-1)k! \geq 8 \exp(2\|\boldsymbol{\theta}\|_1)/(\pi\epsilon)$ , and  $k' = 1$  if  $\|\boldsymbol{\theta}\|_\infty < 1$  or otherwise  $k' = k$ . Then,

$$\mathbb{P}[|\hat{Z}_\zeta^{N,k}(\boldsymbol{\theta}) - Z(\boldsymbol{\theta})| < \epsilon Z(\boldsymbol{\theta})] \geq 1 - \delta.$$

The norm of the parameter vector  $\|\boldsymbol{\theta}\|_1$  influences the error of our approximation to  $\exp(\langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle)$ . If we keep the norm small, a small error can be guaranteed with a small polynomial degree. Thus, we discovered a new connection between regularisation and the resource consumption of the model [A1/28].

For the proof, we discovered a new property of sufficient statistics that allows us to devise a fast Monte Carlo sampling procedure for the tuple density  $\mathbb{P}_{\zeta, \phi}(\mathbf{J} = \mathbf{j}, I = i)$  [A1/27]. We could show that the sufficient statistics of any *discrete* state Markov random field has this property, and precisely indicate under which circumstances *continuous* sufficient statistics also exhibit this property.

**Architecture-specific learning methods (WP4):** Let us now look at resource constrained devices more closely. Microcontroller units (as those used in project A4) are highly restricted in terms of their arithmetic capabilities – state of the art ultra-low-power devices do not even contain a floating point coprocessor. In addition, the main memory is restricted to only a few kilobytes. In case of field programmable gate arrays (FPGA), the limited chip size restricts the complexity and size of the machine learning model.

First, we investigated the parameter estimation in machine learning models. In most cases, model parameters are learned via numerical methods. While second-order numerical optimisation methods provide faster convergence, their memory consumption is prohibitive in the context of ultra-low-power devices. This is due to the (approximation of the) Hessian, which is needed for a Newton step in the optimisation. Instead, we focus on first-order numerical optimisation methods like, e.g., accelerated gradient descent or proximal point methods. These methods normally use floating point arithmetic, but ultra-low-power devices need integer computation. Hence, we extended our work on Boolean matrix factorisation [A1/C1/32] and introduced a new integer regularisation that penalises models with real-valued parameters:  $\rho_{\text{int}}(\boldsymbol{\theta}_i) = 1 - |1 - 2(\lceil \boldsymbol{\theta}_i \rceil - \boldsymbol{\theta}_i)|$ . Although our proposed regularisation is non-smooth and non-convex, we proved convergence to critical points of the objective function, justified by the Kurdyka-Łojasiewicz property [12]. To this end, we employ first-order proximal minimisation [16, 3]. Proximal methods handle non-smoothness by solving a sub-problem in each iteration of gradient descent. These sub-problems require efficient closed-form solutions to make proximal minimisation feasible. We could derive a theorem on the closed-form for the proximal operator of  $\rho_{\text{int}}(\boldsymbol{\theta}_i)$ , that reveals a close connection to numerical rounding:

$$\text{prox}_{\lambda R_{\text{int}}}(\boldsymbol{\theta})_i := \begin{cases} \text{round}(\boldsymbol{\theta}_i) & , \text{if } |\omega - \boldsymbol{\theta}_i| \leq 2\lambda \\ \boldsymbol{\theta}_i + 2\lambda & , \text{else if } \omega > \boldsymbol{\theta}_i \\ \boldsymbol{\theta}_i - 2\lambda & , \text{else if } \omega < \boldsymbol{\theta}_i \end{cases}$$

with  $\omega := \text{argmin}_{u \in \mathbb{N}} |u - \boldsymbol{\theta}_i|$ . Beside our derivation of integer models, our results can be used to interpret binary deep learning techniques [7, 10] as stochastic proximal minimisation procedures. Thus, we provide new theoretical insights in an area of machine learning where empirical results dominate.

Moreover, we succeeded in showing that specific choices of the regularisation weight  $\lambda$  do *always* lead to integer solutions. The resulting integer gradient descent method can be applied to all learning problems with a Lipschitz continuous gradient whose regularised objective is a Kurdyka-Łojasiewicz function. As an example, we proved that the exponential family of densities satisfies this property, which covers a large variety of machine learning models. The component-wise regulariser  $\rho_{\text{int}}(\boldsymbol{\theta}_i)$  and the corresponding proximal map are shown in the two leftmost plots in

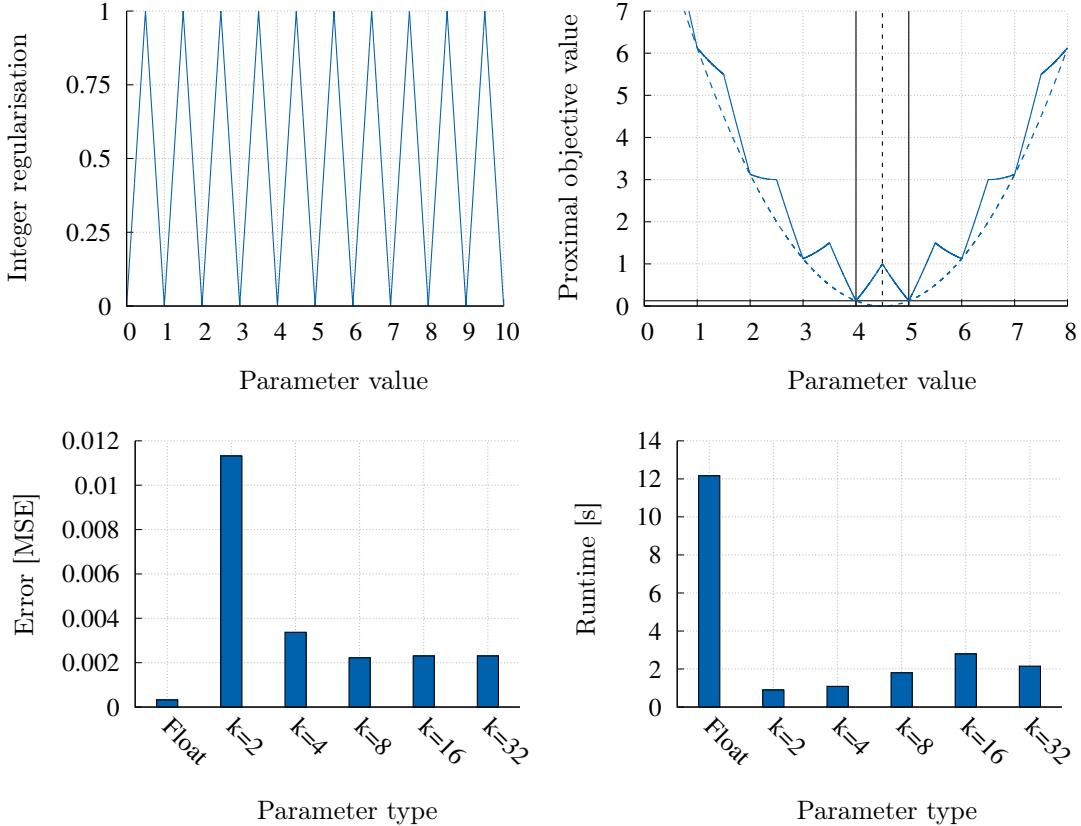


Figure 3.2: **(NW):** Plot of the component-wise integer regularisation  $\rho_{\text{int}}(\boldsymbol{\theta}_i)$  on  $[0; 10]$ . Any  $\theta_i \notin \mathbb{Z}$  is penalized proportionally to its absolute distance to the nearest integer. **(NE):** Proximal problem for  $\rho_{\text{int}}(\boldsymbol{\theta}_i)$  with  $\theta_i = 3.1$  and  $\lambda = 1/(4\kappa)$ . The solid curve is the objective function value for  $\gamma_i \in [0; 8]$ , and its minimum is indicated by the solid vertical line. The dashed curve shows the value of the quadratic term  $(\theta_i - \gamma_i)^2$ , and the dashed vertical line indicates the position of its minimum. The horizontal line shows the function value of the optimal  $\gamma_i$  – two different  $\gamma_i$  values are globally optimal, namely  $\gamma_i^* \in \{4, 5\}$ . **(SW)** and **(SE):** Exemplary result in terms of mean-squared-error (MSE) in empirical marginals and average runtime per training iteration as a function of the largest allowed integer parameter value  $k$  (x-axis).

figure 3.2. Together with our integer-valued PGMs from WP3, integer regularisation opens the way for machine learning without the need for floating-point hardware.

Second, we investigated the evaluation of learned models on restricted devices. Here, we looked at Random Forests (RFs) as they are used in project C3, where real-time execution is needed directly at the sensing node [A1/C3/31]. RFs need to compute tree traversal, which is fast in theory, but somewhat difficult from a computer architecture perspective. Each node of the tree involves branching, which reduces pipelining performance and caching performance. Three different ways to implement a decision tree are inspected with respect to their impact on making good use of data or instruction caching. First, we may store tree nodes using an array and traverse this array with a *while*-loop. Second, we can decompose the tree directly into its *if – else*-structure. Third, we may consider the binary decisions of a tree as a Boolean function and thus directly implement the conjunctive normal form of a tree. We theoretically compared these implementation schemes with respect to a probabilistic model of decision tree evaluation. We applied the fast RF execution to the classification of sensor measurements of project C3 – which measurement is to be kept, which

one to let go. A small microcontroller clocked with 1MHz is enough to filter roughly 12% of the measurements in real time. In addition to this analysis, we also looked at the implementation of Random Forests on FPGAs. We estimated the number of logic blocks necessary for different trees before executing the costly (logic) synthesis operation.

**System adaptation (WP5):** System software can benefit from machine learning in many different ways. As a showcase we are working on lock analysis in multi-core OS kernels. During the last decade it turned out that fine-grained synchronisation of concurrent control flows within the OS kernel is crucial for scalability in multi-core systems. In the Linux kernel, for instance, this led to a huge number of different lock objects of different kinds (blocking, non-blocking, reader/writer, etc.) and lock-free data structures. This makes it hard for kernel developers to tell which locks must be acquired in which order. Our approach to improve this situation is to trace the execution of kernel code and learn locking rules for accessed member variables automatically. Given a sufficient confidence on the correctness of the learned rule, we can check the existing sparse documentation and even generate new documentation for so far undocumented data structures. It could even become possible – if the majority of operations is correct – to detect operations that access member variables without holding the necessary lock(s).

Moreover, for real-time embedded systems, locking protocols in multiprocessor systems have been widely used to ensure both timing correctness and the mutual exclusion property of shared resources. Many locking protocols for real-time systems have been designed, but a principled and theoretical study for the joint considerations of task partitioning and priority assignment for locking protocols has been missing until our recent discovery. We introduced the concept of *resource-oriented partitioned* (ROP) scheduling in [B2/21]. The key novelty of the ROP scheduling is the change of the view angle. Instead of focusing on the computing tasks (for non-critical sections) as the standard approaches do, the ROP scheduling considers the shared resources most important and first assigns these resources to designated synchronisation processors. ROP is the first result with a constant resource augmentation factor even in cases where there is only one critical section per task.

For soft real-time systems, occasional deadline misses are acceptable, as long as the deadline misses can be quantified and bounded. Several convolution-based approaches have been developed in the literature. However, since the time complexity of (native) convolution-based approaches is exponential with respect to the number of jobs in the interval of interest, they are not scalable and are sensitive to the configurations of the task set. In [A1/26], we provided a novel approach based on the multinomial distribution that allows to calculate the deadline miss probability with much better analysis runtime and without precision loss, compared to the traditional convolution-based approach. The improvement is enhanced by a state-pruning technique that significantly improves the runtime and scalability of our analysis with a bounded loss of precision. In addition, we also provided analytical bounds based on the Hoeffding's inequality, Bernstein's inequality, and Chernoff bounds. Our approach is applicable for significantly larger task sets than the previously known convolution-based approaches.

**Benchmarking, Evaluation, and Simulation (WP6):** Benchmarks for mobile devices, such as Android-based smartphones, are typically intended to evaluate the *maximum* performance of components such as the CPU, GPU, or network interface. For the evaluation of system software improvements, as being developed in this project, these benchmarks are inadequate, because we are interested in resource savings in *average* use cases. In order to create such benchmarks in a flexible way, we developed the concept of trace-based benchmark composition [A1/24]. This allows users to record the resource usage of real-world Android apps (CPU, network, file system, Android wake lock, GPS, ...) and to replay the trace on arbitrary other devices at any time without installation of the traced app. With respect to the system software, this benchmark will behave in the same way as the original app. In order to produce more realistic workloads, we have studied how app traces are structured and how traces can be combined to a more complex benchmark, which represents a mix of apps. The benchmark can then be executed on newly developed hardware

platforms, which consist of smartphone hardware but support energy measurements for multiple hardware components separately.

Based on the benchmark data, we construct generative probabilistic graphical models. PGM-based simulation of usage data allows us to generate diverse scenarios for the evaluation of resource saving techniques. Probabilistic queries generate user data that obey specific characteristics to gain precise insights about the interplay of usage and resource consumption. Samples are generated via perturbation-based Monte Carlo sampling in order to avoid costly MCMC methods. The PGM is trained with an adversarial loss function to enforce the generation of samples that are representative but not too close to the mean usage.

**Providing Machine Learning Algorithms (WP7):** Together with project C1, we derived a binary regularisation method that led to new algorithms for Boolean matrix factorisation [A1/C1/32]. Moreover, we established probabilistic error bounds for the trustworthiness of pattern mining based on matrix factorisation [A1/C1/33]. Collaborating with the European project *VaVel – Variety, Veracity, Value* and Thomas Liebig from project B4, we successfully applied our spatiotemporal random field to estimate travel times and integrated the results in routing algorithms. Together with project C3, we applied hardware optimised decision trees for classification tasks in astroparticle physics. Together with projects A4 and B4, we derived a model for uplink transmission power in LTE networks based on downlink indicators, GPS data, and other channel-related features [A4/B4/A1/30]. On the one hand, such models can be used to build simulations of resource consumption in LTE networks. On the other hand, such models can be implemented into the handset itself to decide if a transmission should be postponed to prevent interference-based bottlenecks. Also with B4, we deploy resource constrained learning methods within reflector posts to build a distributed vehicle classification system. Together with project B2 and C3, we provide a python-based tree generator, which automatically applies architecture-aware code optimisations while taking the instruction and/or data cache into consideration. The source code is publicly available at <https://bitbucket.org/sbuschjaeger/arch-forest>.

## Bibliography

- [1] N. Asadi, J. Lin, and A. P. De Vries. “Runtime optimizations for tree-based machine learning models”. In: *IEEE Transactions on Knowledge and Data Engineering* 26.9 (2014), pp. 2281–2292 (cit. on p. 75).
- [2] A. Badanidiyuru, B. Mirzasoleiman, A. Karbasi, and A. Krause. “Streaming submodular maximization: Massive data summarization on the fly”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 671–680 (cit. on p. 65).
- [3] J. Bolte, S. Sabach, and M. Teboulle. “Proximal alternating linearized minimization for nonconvex and nonsmooth problems”. In: *Mathematical Programming* 146.1-2 (2014), pp. 459–494 (cit. on pp. 68, 274).
- [4] L. Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32 (cit. on p. 75).
- [5] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, et al. “Neuromorphic computing using non-volatile memory”. In: *Advances in Physics: X* 2.1 (2017), pp. 89–124 (cit. on p. 76).
- [6] C. W. Clenshaw and A. R. Curtis. “A method for numerical integration on an automatic computer”. In: *Numerische Mathematik* 2.1 (1960), pp. 197–205 (cit. on p. 67).

- [7] M. Courbariaux, Y. Bengio, and J.-P. David. “BinaryConnect: Training Deep Neural Networks with binary weights during propagations”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 3123–3131 (cit. on p. 68).
- [8] D. Davis, B. Edmunds, and M. Udell. “The Sound of APALM Clapping: Faster Nonsmooth Nonconvex Optimization with Stochastic Asynchronous PALM”. In: *Advances in Neural Information Processing Systems* 29 (2016) (cit. on p. 78).
- [9] C. De Sa, C. Zhang, K. Olukotun, and C. Re. “Taming the Wild: A Unified Analysis of Hogwild-Style Algorithms”. In: *Advances in Neural Information Processing Systems* 28 (2015), pp. 2674–2682 (cit. on p. 78).
- [10] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. “Binarized Neural Networks”. In: *Proceedings of the NIPS*. 2016, pp. 4107–4115 (cit. on p. 68).
- [11] M. Kamp, M. Boley, O. Missura, and T. Gärtner. “Radon machines: effective parallelisation for machine learning”. In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on p. 78).
- [12] K. Kurdyka. “On gradients of functions definable in o-minimal structures”. In: *Annales de l'institut Fourier* 48.3 (1998), pp. 769–783 (cit. on p. 68).
- [13] C. Lucchese, F. M. Nradini, S. Orlando, R. Perego, N. Tonellootto, and R. Venturini. “QuickScorer: Efficient Traversal of Large Ensembles of Decision Trees”. In: *Procs. ECML PKDD 2017*. Springer, 2017, pp. 383–387 (cit. on p. 75).
- [14] R. H. Möhring, M. Skutella, and F. Stork. “Scheduling with AND/OR Precedence Constraints”. In: *SIAM Journal on Computing* 33.2 (2004), pp. 393–415 (cit. on p. 78).
- [15] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. “An analysis of approximations for maximizing submodular set functions—I”. In: *Mathematical Programming* 14.1 (1978), pp. 265–294 (cit. on p. 66).
- [16] N. Parikh and S. Boyd. “Proximal Algorithms”. In: *Found. Trends Optim.* 1.3 (Jan. 2014), pp. 127–239 (cit. on p. 68).
- [17] E. J. G. Pitman. “Sufficient statistics and intrinsic accuracy”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 32 (1936), pp. 567–579 (cit. on p. 67).
- [18] R. Prenger, B. Chen, T. Marlatt, and D. Merl. *Fast map search for compact additive tree ensembles (cate)*. Tech. rep. Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2013 (cit. on p. 75).
- [19] B. Schölkopf and A. J. Smola. “Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond”. In: Cambridge, MA, USA: MIT Press, 2001. Chap. Kernels (cit. on p. 66).
- [C1/20] **S. Hess** and **K. Morik**. “C-SALT: Mining Class-Specific ALTerations in Boolean Matrix Factorization”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2017*. Springer, 2017 (cit. on pp. 80, 275, 287).
- [B2/21] **W.-H. Huang**, M. Yang, and **J.-J. Chen**. “Resource-Oriented Partitioned Scheduling in Multiprocessor Systems: How to Partition and How to Share?” In: *Real-Time Systems Symposium (RTSS)*. Porto, Portugal, Dec. 2016 (cit. on pp. 17, 70, 78, 107).

- [22] B. Van Essen, C. Macaraeg, M. Gokhale, and R. Prenger. “Accelerating a random forest classifier: Multi-core, GP-GPU, or FPGA?” In: *Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on*. IEEE. 2012, pp. 232–239 (cit. on p. 75).
- [23] M. Zhang, L. Zhang, L. Jiang, Z. Liu, and F. T. Chong. “Balancing Performance and Lifetime of MLC PCM by Using a Region Retention Monitor”. In: *International Symposium on High Performance Computer Architecture (HPCA)*, pp. 385–396 (cit. on p. 76).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [A1/24] **A. Lochmann**, F. Bruckner, and **O. Spinczyk**. “Reproducible Load Tests for Android Systems with Trace-based Benchmarks”. In: *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion*. ICPE ’17 Companion. New York, NY, USA: ACM Press, 2017, pp. 73–76 (cit. on p. 70).
- [A1/25] **C. Pöltz**, **W. Duivesteijn**, and **K. Morik**. “Interpretable Domain Adaptation via Optimization over the Stiefel Manifold”. In: *Machine Learning* 104.2-3 (2016), pp. 315–336 (cit. on pp. 65, 80).
- [A1/26] **G. von der Brüggen**, **N. Piatkowski**, **K.-H. Chen**, **J.-J. Chen**, and **K. Morik**. “Efficiently Approximating the Probability of Deadline Misses in Real-Time Systems”. In: *30th Euromicro Conference on Real-Time Systems, ECRTS 2018, July 3-6, 2018, Barcelona, Spain*. LIPIcs, 2018 (cit. on pp. 17, 70, 78).
- [A1/27] **N. Piatkowski** and **K. Morik**. “Fast Stochastic Quadrature for Approximate Maximum-Likelihood Estimation”. In: *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, California, USA, August 6-10, 2018*. 2018 (cit. on p. 68).
- [A1/28] **N. Piatkowski** and **K. Morik**. “Stochastic Discrete Clenshaw-Curtis Quadrature”. In: *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York, USA, 19-24 June 2016*. JMLR: W&CP. JMLR.org, June 2016 (cit. on p. 68).
- [A1/C1/29] **N. Piatkowski**, **S. Lee**, and **K. Morik**. “Integer undirected graphical models for resource-constrained systems”. In: *Neurocomputing* 173.1 (Jan. 2016), pp. 9–23 (cit. on pp. 16, 67).
- [A4/B4/A1/30] **R. Falkenberg**, **B. Sliwa**, **N. Piatkowski**, and **C. Wietfeld**. “Machine Learning Based Uplink Transmission Power Prediction for LTE and Upcoming 5G Networks using Passive Downlink Indicators”. In: *2018 IEEE 88th IEEE Vehicular Technology Conference (VTC-Fall)*. Chicago, USA, Aug. 2018 (cit. on p. 71).
- [A1/C3/31] **S. Buschjäger** and **K. Morik**. “Decision Tree and Random Forest Implementations for Fast Filtering of Sensor Data”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 65-I.1 (Jan. 2018), pp. 209–222 (cit. on pp. 19, 69, 75, 296, 303).
- [A1/C1/32] **S. Hess**, **K. Morik**, and **N. Piatkowski**. “The PRIMPING routine—Tiling through proximal alternating linearized minimization”. In: *Data Mining and Knowledge Discovery* 31.4 (July 2017), pp. 1090–1131 (cit. on pp. 68, 71, 80, 274, 287).

- [A1/C1/33] **S. Hess, N. Piatkowski, and K. Morik.** “The Trustworthy Pal: Controlling the False Discovery Rate in Boolean Matrix Factorization”. In: *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA*. SIAM. 2018, pp. 405–413 (cit. on pp. 71, 80, 274, 287).

### 3.4 Project plan

Research on cyber-physical systems and research on machine learning have moved towards each other in the first and second phase of funding. New models, inference algorithms, learning procedures, and data collection methods for resource constrained systems were developed. Our research on system-wise data collection is considered complete, but new challenges arise with large heterogeneous networks of small computational devices, e.g., in the context of the Internet of Things. To meet the requirements of this scenario, we investigate the structure of distributed learning of deep and probabilistic models. We study the resource consumption of learning algorithms in the context of emerging computation and memory technologies, while keeping the focus on the *models of learning*. Hence, in the third phase of the project A1, learning models and the underlying architectures will be optimised and considered jointly.

#### Goals

The ubiquity of small devices challenges the understanding of all the distributed data. Where project B3 already worked on distributed data analysis under communication constraints, here we continue the work of the previous phases of A1 on data aggregation and sampling as well as that on probabilistic graphical models (PGM) facing memory and real-time constraints, in particular. Our goal is to develop a sound theoretical basis for specifying the resource needs of distributed learning and in-network computing.

We do not want to restrict ourselves to the current memory architectures, but already investigate non-volatile memories. Today, it has not been investigated how they can be effectively and efficiently be exploited for machine learning. Since we already have analysed the memory requirements for standard architectures, we are now ready to inspect whether and how machine learning might benefit from non-volatile memories and non-von-Neumann architectures.

If analysis results are to act on data streams and not only to be delivered to a human who receives some insight, the execution of learning results themselves needs to perform in real-time. The goal is to develop a model of learning for Random Forests and deep feed-forward networks such that FPGAs become capable of tailoring computing hardware on demand. Moreover, learning tasks need to be analysed such that they can be executed on massively parallelised cores without sacrificing the guarantees of the learning model.

Our three goals of extending our analysis to distributed learning, emerging new memory models, and dynamic and adaptive execution of learning results are mapped onto objectives that are achieved by the work packages.

#### Work schedule

The overview of the A1 work packages is as follows:

- We co-design hardware and machine learning models that can be more suitable to perform data analysis in *real-time* in work package 1.

- We investigate the impact of emerging *memory architectures* in embedded systems on machine learning models and algorithms in work package 2.
- We improve machine learning algorithms in resource constrained distributed systems, targeting data *sampling and aggregation* in work package 3 and *distributed learning models* for joint optimisation of computation and data communication in work package 4.
- We design resource management and *scheduling strategies* to execute distributed/parallel machine learning algorithms to trade the statistical guarantees and resource usage in real-time applications in work package 5.
- The dissemination of our insights and methods within the research centre and within the research community is embodied by work package 6.

**Work package 1. Hardware/Software Co-Design for Machine Learning** Due to the Internet of Things, real-time executions of learned models have recently become an important topic. [A1/C3/31] consider the application of Random Forest (RF) models on embedded hardware and FPGAs. The analysis of their behaviour with respect to memory hierarchies such as caches or scratchpad memory is important, because compulsory cache misses as well as capacity cache misses differ for different implementations. Investigating memory use for decision trees as single entities is not enough, because they are often used by ensemble models. Based on the analysis of decision tree execution, the application of overall forests should be enhanced, possibly by storing and executing frequently used parts of an ensemble on acceleration hardware, such as FPGAs, so that we can improve overall performance. Constraints on the trees may also help fast execution of Random Forests. Current approaches fix the heights of trees to a certain limit[22, 18, 1, 13]. These approaches usually work well in boosting settings, where we often limit the expressiveness of base models to avoid overfitting. For Random Forests, however, the base-models must be complex in order to effectively reduce the bias error[4]. It is an open question whether there are other constraints on trees that are well suited for hardware acceleration while maintaining a low bias error, e.g., by introducing regularisation onto the tree structure during training.

In a more general setting, we investigate the resource consumption of (flash-based) FPGAs for probabilistic models and deep feed-forward networks. Dynamic configuration of FPGA allows us to provide tailored compute hardware on demand. Nevertheless, we have to trade off the cost of reconfiguration and the expected benefit of the underlying algorithm whenever we invoke dynamic configuration. It will hence be examined how far reinforcement-learning techniques, like bandit models, can be used to trigger dynamic reconfiguration while maximising the expected reward. However, the same trade-offs exist on other architectures as well. Besides microcontroller units, FPGAs, and traditional many-core hardware, we will hence investigate our strategies w.r.t. designated machine learning hardware accelerators (like Intel's Loihi, Google's Tensor Processing Unit). Tuning such trade-offs to leverage the performance of different hardware components and accelerators, e.g., for a tensor flow application or a Random Forest, will involve hyperparameter optimisation.

**Work package 2. Machine Learning Models and Algorithms for Emerging Memory** The emerging byte-addressable non-volatile memories (NVMs), such as Phase Change Memory (PCM), Spin-Transfer Torque RAM (usually abbreviated as STT-MRAM) and Resistive RAM (ReRAM), feature low leakage power, high density, and low unit costs. Due to their byte-addressability and (almost) negligible idle power, an NVM module can be used as main memory and/or storage to build up extra low-power computing systems. An example of such NVM devices is 3D XPoint available on the open market since October 2017 (or Optane memory DIMM, announced in May 2017 and

to be available in 2018)<sup>4</sup>, both developed by Intel. NVMs are hence popular alternatives to replace DRAM as main memory or replace hard disks and NAND flash as storage, especially for energy-constrained embedded systems.

Since complex machine learning methods require extensive computational capabilities and data, the separation of memory and computing units in traditional von Neumann architectures introduces significant waste of energy and time. The bottleneck can be potentially handled by co-locating memory and computation in one device, e.g., in neuromorphic computing systems. Specifically, the IBM TrueNorth<sup>5</sup> system is an example, which implements artificial synapses by using conventional CMOS technologies. To achieve low-power operation and massive parallelism, the scaling of dense non-volatile memory (NVM) crossbar arrays has played an important role [5].

Although such promising memory architectures will emerge very soon, how to use them efficiently and effectively from the perspectives of machine learning methods and embedded systems remain wide-open. In this work package, we will explore machine learning models and algorithms for such emerging memory architectures, especially for embedded systems. Specifically, we will explore the estimation of probabilistic graphical models, deep feed-forward networks, and RFs under such emerging memory architectures. Moreover, we aim at finding model-independent characteristics that allow us to identify if a machine learning model will benefit from non-volatile memories and non-von-Neumann architectures.

Although NVMs can keep the data without electricity, they are not perfectly non-volatile. The bit(s) stored in a cell can be corrupted if the cell is not charged again after a while, which is referred to as the *data retention problem*. Depending on the volatility requirement of the data, one can program a cell in an NVM to have different data retention times. Specifically, Zhang et al. [23] and some other earlier research results showed that there can be (at least) two writing modes to program a multiple-level-cell (MLC) PCM: *fast write* leaves the cells in a non-volatile state with a short data retention time, and *slow write* leaves the cells in a non-volatile state with a long data retention time. As an example, consider a first-order parameter update of a probabilistic graphical model in exponential family form. The parameters  $\theta$  are updated iteratively according to  $\theta^{(i+1)} = \theta^{(i)} - \eta^{(i)}(\hat{\mu} - \tilde{\mu})$ . Here,  $\tilde{\mu} = (1/N) \sum_{i=1}^N \phi(x)$  is an average feature vector (sufficient statistic) that is computed and written only once – the same value must be retained during the whole training procedure. On the other hand,  $\hat{\mu}$ , the model’s expectation of  $\phi(X)$  given  $\theta^{(i)}$ , as well as the parameter  $\theta^{(i)}$  itself are recomputed and rewritten every single iteration. Thus, it suffices to store them in cells whose data retention time is short, e.g., about the time required to recompute  $\hat{\mu}$ . These and other frequently updated data structures can be configured with “fast writes” and can significantly improve the performance and lower the resource consumption. We will identify such situations in machine learning algorithms and provide estimates for the saved resources.

**Work package 3. Resource Constrained Data Aggregation and Sampling** Sampling is one of the most fundamental methods in machine learning and can be used to increase a classifier’s performance as well as to reduce the resource consumption of a learning algorithm. For single models, sampling may increase performance for skewed distributed training data, e.g., as found in TP C3. For ensemble methods, sampling enables us to effectively reduce the variance of a learner leading to smaller generalisation errors. Additionally, sampling can be used to introduce parallelism into sequential learning algorithms to better suit these to different computer architectures.

On small, resource-scarce devices we cannot store large amounts of data and thus usually consider data streams. Therefore, sampling must be performed on streaming data. On a broader perspective, we want to tackle the question if we can find sampling algorithms with small memory

---

<sup>4</sup><https://techreport.com/news/31932/optane-dimms-and-companion-cpus-will-arrive-in-2018>

<sup>5</sup><http://www.research.ibm.com/articles/brain-chip.shtml>

requirements that offer guarantees for different quality criteria such as memory consumption, model performance, or communication costs in distributed systems. More specifically, we have already considered a sampling algorithm on streaming data with respect to small, expressive summaries using the Sieve-Streaming algorithm. Specifically for Sieve-Streaming we notice two drawbacks: First, Sieve-Streaming is not able to remove items from a summary, even though better representatives may be present. Recent work in the context of privacy-preserving computations allows for this deletion operation when it is triggered from outside (the user removes an item from a database due to privacy concerns), but it has not yet been studied to improve the quality of a summary. Second, Sieve-Streaming is restricted to submodular functions, which only include clustering or entropy-based quality functions at the moment.

For cluster analysis, subsampling and aggregation have already been done in the context of coresets in project A2. We wish to extend this approach in joint work with A2 to include additional constraints. For now, coreset constructions consider solely a single quality measure – the loss in objective function value. However, coresets consider multiple goals for the construction, e.g., integrality, amenability for specific approximation procedures or data acquisition techniques. The latter would correspond to a subset of data that guarantees a certain approximation error and, at the same time, is easy to collect.

**Work package 4. Distributed Machine Learning and In-Network Computing** Machine learning in large heterogeneous networks requires an efficient communication and a fault-tolerant distributed representation of the estimated model.

Probabilistic graphical models and deep neural networks (DNNs) rely on an inherent logical structure that defines how data pass through the model. Moreover, inference and learning sub-routines, like gradient computations, rely on function evaluations whose interactions are also described by the logical structure. In the context of distributed learning, the structure implies how data and intermediate results are transferred within the physical network. For DNNs, the structure is most often (hand-)tuned to achieve a low empirical misclassification rate. The conditional independence structure of probabilistic graphical models arises from fundamental properties of the underlying probability measure. If the structure is unknown, a consistent estimation via numerical optimisation methods is carried out. Existing research regarding the structure estimation in PGMs and DNNs does not consider the communication complexity of distributed learning. In this work package, we investigate the resource-aware structure estimation, including the quantification of the consumed resources and the structure estimation itself – constraints on the communication complexity are integrated into the learning problem, e.g., via regularisation.

Such distributed learning may be significantly constrained by limited bandwidth. The data on a node usually contain redundant information. Thus it is important to extract features from the raw data, which can significantly reduce the amount of exchanged data without much loss of the key information. As the complexity of the feature extraction system grows, the representation of information will be more concise and thus the communication load is reduced. The key information extraction will be studied by extending WP3. We will explore the scheduling of data communication to optimise the performance of communication together with computation.

At least two types of distributed architectures will be studied. One is based on Networks on Chips (NoCs), in which several cores/processors are interconnected by using switches. Another type is heterogeneous networks, in which some nodes are faster than others. In both cases, we will investigate scheduling and routing strategies to maximise the usage of switches to fit the distributed learning model. Moreover, locking mechanisms that synchronise the computation of distributed nodes will result in a waste of resources, since the faster nodes must wait for the slower nodes. In such scenarios, being determinate is very important to ensure that the output does not depend on the execution of the program but only on the model of computation. We therefore will

study locking mechanisms or alternatives to be *efficient and determinate* at the same time, or have bounded side-effects if one of them has to be slightly sacrificed. We plan to develop balancing strategies to distribute work non-uniformly among the heterogeneous nodes, such that the idle-time for the fast nodes is minimised. Moreover, state of the art synchronisation mechanisms, e.g., lock-free numerical optimisation methods [9, 8] and ROP [B2/21], and parallelisation techniques, e.g., Radon Machines [11], are considered to facilitate the consistent estimation of communication-efficient structures.

**Work package 5. Representation, Execution, and Dependencies of Learning Tasks** The main objective of this work package is the bidirectional exchange between machine learning and resource management to develop abstract execution models and handle non-trivial dependencies among various learning tasks. On the one hand, we bring insights and techniques from scheduling theory to the area of distributed machine learning in order to improve asynchronous learning, e.g., via [9, 8]. On the other hand, we leverage probabilistic graphical models to integrate dependencies among asynchronous tasks into scheduling and resource management. Our methods should not be tailored to specific machine learning algorithms or scheduling strategies. Instead, we aim at deriving unified results for classes of machine learning models (like feed-forward networks or exponential families) under different scheduling models.

For a machine learning task, the precedence constraints between threads are defined to ensure that the input data of a thread are ready before it starts. Dropping certain data-dependency or threads may sacrifice the statistical guarantee of the machine learning task slightly but can significantly improve the "makespan" (the finish time of the last thread) of the machine learning task. With such flexibility in mind, when applying machine learning algorithms for real-time applications, it is important to consider the trade-offs between the makespan and the statistical guarantee of the machine learning tasks. Here is an example: Suppose that we have  $M$  identical processors, the length  $L$  of the longest path in the dependency constraint is 20, and the total work is  $10M$ . This means that the makespan of this task is at least 20, and 50% of the processors are idle during the execution of this task. If we can drop certain dependency so that  $L$  becomes 10, then, we can potentially execute the work completely in parallel and the makespan can be ideally 10, where all the processors are 100% utilised. On one hand, the makespan is important for responding in real time. On the other hand, the statistical guarantee should still be provided to ensure the stability of the results.

We will investigate two orthogonal approaches. One approach starts from the dependency graph with the best learning output (with respect to the statistical performance). Some of the dependency constraints (or threads) will then be removed during the schedule design. Another approach starts from the dependency graph with the minimum required learning output (with respect to the statistical performance). Some of the dependency constraints will then be added during the schedule design. The former approach is easier from the machine learning perspective, because the loss of statistical performance can be specified in the edge. The latter approach is more difficult from the machine learning perspective, because there are potentially many dependency graphs with the same statistical performance and only one of them is specified. Due to multiprocessor anomaly, both approaches are not easy to handle. That is, removing a precedence constraint may make an originally good schedule worse, and vice versa. We will jointly consider the representation and execution of precedence-constrained learning tasks so that both approaches can be applied. Scheduling algorithms that are applicable for more flexible precedence constraints, e.g., AND/OR constraints [14] and imprecise computation, will be explored together with the construction of flexible dependency constraints of machine learning algorithms.

Furthermore, the analysis of the probability of deadline misses of our preliminary result in [A1/26] will be further enhanced to handle the dependency of the random variables that represent the execution time of real-time tasks. The probabilistic graphical models will be extended to integrate

such dependencies. Such enhancement would be needed to ensure probabilistic guarantees of both timing behaviour (with respect to the probability of deadline misses) and statistical performance (with respect to the quality of the machine learning algorithms). We will seek for methods with different trade-offs between the accuracy and the time/space complexity of the analysis so that they can be used for different learning scenarios.

**Work package 6. Provision of Data Analysis** In addition to our theoretical guarantees on quality and resource requirements, resource constrained machine learning techniques be made available and ready to use within the Collaborative Research Centre and beyond. On the one hand, our methods should be disseminated in well-known data analysis tools (like R or RapidMiner) to provide researchers with the right analysis techniques for resource constrained systems. On the other hand, projects and also PhD students within the research centre need expertise when it comes to the foundations of resource constrained data analysis. We will hence give advice to projects regarding our new techniques and will support projects in the use of existing methods.

### Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Hardware/Software Co-Design for Machine Learning																	
2. Machine Learning Models and Algorithms for Emerging Memory																	
3. Resource Constrained Data Aggregation and Sampling																	
4. Distributed Machine Learning and In-Network Computing																	
5. Representation, Execution, and Dependencies of Learning Tasks																	
6. Provision of Data Analysis																	

### 3.5 Role within the Collaborative Research Centre

Project A1 studies machine learning on cyber-physical systems with a focus on models for abstract learning tasks. With this focus, A1 forms one of the basic building blocks of the collaborative research centre. Thus, there are a range of opportunities inside the CRC to use the results from A1 to enhance and help other projects. On the other hand, A1 also directly or indirectly benefits from theoretical and practical insights from other projects.

Project A2 focuses on developing algorithm design paradigms for learning methods in CPS, e.g., by using coresets. Project A1 on the other hand focuses on machine learning models, e.g., by extracting

## Project A1

relevant points for learning. In the context of submodular function maximisation and coresets both approaches overlap. Thus it seems natural to study the relationship between submodular function maximisation and coresets together. Additionally, we want to investigate different quality functions for these methods e.g., by the means of multi-objective optimisation.

A3 focuses on model-based optimisation (MBO), which can be applied for hyperparameter optimisation in our WP1 for hardware/software co-design and WP4 for tuning PGMs and DNNs. We wish to collaborate with A3 to explore the applicability of MBO in the above work packages.

Together with Kristian Kersting (A6), we have worked on estimating distributions from text data along the lines of our work on transfer learning [A1/25].

Projects A4 and B4 study prediction methods for the resource consumption of hardware platforms. So far, we have derived a model for uplink transmission power in LTE networks based on channel-related features, which can be used on hand-held devices. We wish to extend our cooperation with projects A4 and B4 by providing less resource-demanding and more accurate machine learning models. We have collaborated with B4 on traffic prediction and dynamic route planning. In addition, we deploy resource constrained learning methods within reflector posts to build a distributed vehicle classification system. We wish to build on these results and further enhance our proposed methods in the context of project B4.

B2 focuses on the machine learning models to detect nanoparticles for medical applications. We plan to apply our hardware/software co-design and distributed machine learning for training and executing DNNs for detecting nanoparticles with B2. Moreover, part of the current analysis pipeline in B2 uses Random Forest models. We wish to use our aforementioned code generator to execute the RF efficiently to further reduce hardware costs and increase classification speed.

B3 studies active learning with respect to simulations. In particular, a common interest is how to handle data with class imbalance.

Collaboration with C1 is based on matrix factorisation methods [A1/C1/32, C1/20, A1/C1/33].

Project C3 studies celestial objects, in which one of the most fundamental tasks is background-noise separation of telescope measurements. Somewhat surprisingly, Random Forest (RF) models still outperform other state of the art machine learning models in this task. RFs are slow application-wise since each tree must be traversed to classify new measurements. This observation ultimately led to the aforementioned code generator to speed-up RF models for real-time evaluation in A1. We wish to extend this work in cooperation with C3 to bring machine learning closer to the telescope itself.

Load balancing for enhanced scheduling is a topic of Jian-Jia Chen, which will be further investigated in collaboration with C5.

## 3.6 Differentiation from other funded projects

### **Suspension-Aware Designs and Analyses for Real-Time Embedded Systems (Sus-Aware) (Chen, Reference number DFG CH 985/12-1) (Funding period: 2018–2011)**

This project focuses on robust and solid fundamental algorithms and analyses to carefully mitigate and analyse the impact of self-suspending behaviour in modern real-time embedded systems. Machine learning does not play any role in this project.

### **Design and optimisation of Non-Volatile One-Memory Architecture (NVM-OMA) (Chen, Reference number DFG CH 985/13-1) (Funding period: N.N.–N.N.)**

This project aims to enable the effectiveness of one-memory architectures, in which a NVM is used both for the storage and main memory. We plan to perform design-space exploration

in hardware and software designs and integrate analytical and optimised resource management in operating systems. The proposal was submitted in December 2017 and is under review. Machine learning does not play any role in this project.

**Partitioning, Scheduling, Spinning, Suspension, synchronisation, and Locking Protocols in Real-Time Systems (PS4Lock)**

**(Chen, Reference number DFG CH 985/14-1) (Funding period: N.N.–N.N.)**

The project will design practical and solid fundamental algorithms and analyses to handle shared resources based on lock mechanisms in multiprocessor embedded systems. Our project intends to find the break-even and dominating scenarios for task partitioning, task spinning, task suspension, and resource synchronisation. The proposal was submitted in Feb. 2018 and is under review. Machine learning does not play any role in this project.

**Variety, Veracity, Value: Handling the Multiplicity of Urban Sensors (VaVel)**

**(Morik, Reference number Horizon2020-688380) (Funding period: 2016–2020)**

The goal of the VaVeL project is to advance our ability to use urban data in applications that can identify and address citizen needs and improve urban life. The motivation comes from problems in urban transportation. Basic research from A1 has been applied within this European project. This is one way of transferring basic research to real-world applications. At the same time, the data from the city of Dublin and the city of Warsaw have been used by our learning methods. In particular, spatiotemporal random fields have been successfully applied within VaVel.

**Modellierung von Themen und Strukturen religiöser online-Kommunikation**

**(Morik, Reference number MERCUR, PR-2015-0046) (Funding period: 2016–2018)**

The project addresses two main questions: What are the structures of religious communication in online contexts, and how do religious topics spread across these structures? The study is based on computer-mediated communication (online forums, social media) of neo-conservative Christian and Muslim groups, e.g., Evangelical and Salafi communities. Text data in social media challenge the efficiency of learning. We have developed a more efficient learning of low-rank representations based on convex optimisation. Instead of explicitly learning low-dimensional features, we compute a low-rank representation implicitly by regularising full-dimensional solutions. Lukas Pfahler, the PhD student from this expiring project, has become a member of the CRC 876, who is financed by the university.

**Synthese von maschinellem Lernen und numerischer Simulation zur Echtzeitsteuerung**

**(Morik, Reference number MERCUR, PR-2016-0039) (Funding period: 2017–2019)**

As a collaboration with the CRC 837, the MERCUR project investigates the use of machine learning for real-time prediction. The physical relationships obtained from a simulation model and the knowledge gained through process-accompanying data analysis from monitoring and measurement data are merged in order to significantly improve process control in mechanical tunnel construction. The project is related to our work in B3, and meetings between the groups have taken place. The particular work on the tunnel data is on time series abstractions through clustering. The active learning framework from B3 could also be applied to data from tunnel processes.

**Kompetenzzentrum maschinelles Lernen Rhein Ruhr – ML2R**

**(Morik, Reference number BMBF) (Funding period: 2018–2022)**

The German Federal government has accepted four competence centres for machine learning that have the double function of achieving scientific excellence and transferring results into practice. Of course, stimulating discussions between members of the CRC 876 who work on machine learning and members of ML2R will be possible. This may strengthen Dortmund and attract excellent scientists.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	2	129,000	2	129,000	2	129,000	2	129,000
Total	—	129,000	—	129,000	—	129,000	—	129,000
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	129,000		129,000		129,000		129,000	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Jian-Jia Chen, Prof. Dr., professor	Embedded systems	TU Dortmund	9	—	Existing funds
	2	Katharina Morik, Prof. Dr., professor	Data mining	TU Dortmund	4	—	Existing funds
	3	Lukas Pfahler, M.Sc., doctoral researcher	Data mining	TU Dortmund	19.92	—	Existing funds
	4	Georg von der Brueggen, Dipl.-Inf., doctoral researcher	Embedded systems	TU Dortmund	19.92	—	Existing funds
	5	N.N., student assistant	Embedded systems	TU Dortmund	8	—	Existing funds
	6	N.N., student assistant	Data mining	TU Dortmund	8	—	Existing funds
Non-research staff	7	Claudia Graute, secretary	—	TU Dortmund	1	—	Existing funds
<b>Requested staff</b>							
Research staff	8	N.N., doctoral researcher	Embedded systems	TU Dortmund	—	Doctoral researcher	—
	9	Sebastian Buschjäger, M.Sc., doctoral researcher	Data mining	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

**1. Chen, Jian-Jia**

Project management. Focus on resource management and hardware and software integration models.

**2. Morik, Katharina**

Project management. Focus on machine learning models.

**3. Pfahler, Lukas**

Focus on the development of machine learning models that are suitable and useful for embedded system applications. Cooperation in the WPs 1, 4, 5 and 6.

**4. von der Brueggen, Georg**

Focus on the development of hardware and software interfaces for the studied machine learning models in embedded systems. Cooperation in the WPs 1, 2 and 5.

**5. N.N.**

Implementation and assistance in evaluation of algorithms.

**6. N.N.**

Implementation and assistance in evaluation of algorithms.

**7. Graute, Claudia**

Secretary.

**Job descriptions of staff for the proposed funding period (requested funds):**

**8. N.N.**

Focus on the development of hardware and software interfaces for the studied machine learning models in embedded systems. Cooperation in the WPs 1, 2, 4 and 5.

**9. Buschjäger, Sebastian**

Focus on the development of machine learning models that are suitable and useful for embedded system applications. Cooperation in the WPs 1, 2, 3, and 6.

### **3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from University	5,000	5,000	5,000	5,000
Sum of existing funds	5,000	5,000	5,000	5,000
Sum of requested funds	0	0	0	0

(All figures in euros)

### **3.7.5 Requested funding for instrumentation for the new funding period**

This project does not request any funding for major research instrumentation.

### 3.1 General information about Project A2

### 3.1.1 Project title:

Algorithmic aspects of learning methods in embedded systems

### 3.1.2 Research area(s):

409-01 (Theoretical Computer Science), 409-06 (Information Systems)

### 3.1.3 Principal investigator(s)

Sohler, Christian, Prof. Dr., 19.02.1973, German

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

Phone: 0231-755-6940  
E-mail: christian.sohler@tu-dortmund.de

Teubner, Jens, Prof. Dr., 22.03.1976, German

LS 6, Fakultät für Informatik, Technische Universität Dortmund  
Otto-Hahn-Straße 14  
44227 Dortmund

Phone: 0231-755-6481  
E-mail: jens.teubner@tu-dortmund.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

no       yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes	(x) no
2.	clinical trials	( ) yes	(x) no
3.	experiments involving vertebrates.	( ) yes	(x) no
4.	experiments involving recombinant DNA.	( ) yes	(x) no
5.	research involving human embryonic stem cells.	( ) yes	(x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes	(x) no

## 3.2 Summary

The main objective of this project is to develop design paradigms for learning algorithms in embedded systems. To this end, we will consider computational models that reflect important aspects of embedded systems, develop and analyse algorithms in these models and empirically evaluate their performance to make conclusions about their efficiency as well as the prediction power of the considered computational models. Then we would like to use the insights gained this way to derive general algorithmic design paradigms for embedded systems.

In this third phase of the project we will continue our study of the aspects energy consumption, communication and computation architecture, data streams and distributed systems. We know from the first two phases of the project as well as the findings of other projects (for example, A1, A6, B3, B4 and C4) that small summaries of the data, such as sketches or coresets, are a suitable design paradigm to obtain data-streaming and distributed algorithms. In the current phase we will continue in this direction and try to unite our findings in different computational models to design a single approach that fits different aspects of embedded systems. For this purpose, we want to continue to deepen our understanding of the individual approaches. Our focus will lie on energy consumption, which we will study both theoretically and empirically.

Christian Sohler will work as a visiting researcher at Google Zürich until 9th August 2019. During that time the main focus of the project will be on the study of energy consumption and communication and computation architecture.

## 3.3 Project progress to date

### 3.3.1 Report and current state of research

The goal of the second phase was to gain more knowledge about the influence of different aspects of embedded systems on algorithm design for learning problems. We have focused on data streams, distributed settings, energy consumption and computation and communication architecture. In the first work package, our focus has been on data reduction as an algorithmic technique and, in particular, its use in the context of streaming algorithms. Consequently, we have developed a number of new techniques for specific problems. Overall, the main insight is that data reduction in the form of coresets and sketching is a main algorithmic paradigm for developing streaming algorithms for learning problems. In the second work package, we have studied distributed learning algorithms. Again, our insight is that data reduction in the form of coresets, sketches, and also other sampling processes is a powerful design paradigm. Furthermore, it turns out that we can address the challenges of both data streams and distributed algorithms using a very similar approach.

Although energy consumption is increasingly relevant in the design of computing systems, the interaction between algorithmic properties and the resulting energy consumption is still largely unknown. This is what we have studied in a third work package by example of database micro-benchmarks. We found that compute-bound algorithms require a different treatment than memory-bound algorithms. Finally, in a fourth work package, we have developed design paradigms for data-intensive algorithms that explicitly consider a system's communication architecture, e.g., in the form of cache allocation within multi-core systems or in the form of optimised code for GPU accelerators.

**Streaming and Clustering Algorithms.** In the second project phase we have continued to study streaming algorithms for clustering problems in different computational models to further under-

stand the algorithmic foundations of such algorithms. Our first main result is a new algorithm for  $k$ -median clustering for dynamic data streams. In dynamic data streams we assume that the stream consists of update operations, i.e., insertions or deletions of data points, and we like to maintain a set of  $k$  centers that approximates the  $k$ -median objective function within a factor of  $(1 + \epsilon)$ . All previous results in this model required space exponential in the dimension [46, 47, 42]. We have developed a new coresset construction to obtain the first algorithm with polynomial dependence on the dimension of the input space [A2/57]. The high level idea of our construction is to combine the approach of Chen [40] with the grid based approach of Frahling and Sohler [42]. The main challenge is to show that only a polynomial number of grid cells are relevant. In order to achieve this we first randomly shift the grid. This way, with reasonable probability we know that the cluster centers are not extremely close to the boundary of the grid cells. Following ideas from Frahling and Sohler we conclude that we can focus on highly populated cells, where a cell is highly populated if the number of points inside a cell times the cell diagonal is at least  $\delta$  times the cost of an optimal solution for some appropriately chosen value of  $\delta$ . The reasoning behind this approach is to equally spread the error of replacing the points inside a cell by one or only few representatives. The difference to the construction of Frahling and Sohler lies in the final decomposition, the choice of  $\delta$ , and, most important, the fact that we take a uniform sample from the highly populated cells (rather than taking a single point as representative). We then argue how the above construction can be implemented in the setting of dynamic data streams. While this approach leads to polynomial dependence on the dimension, the polynomial factors (also for the other relevant parameters) involved are rather large. We therefore are developing further techniques to improve the space requirements of the algorithm.

Another problem we studied in dynamic data streams is the development of scalable locality sensitive hashing [A2/58]. We show that given a stream of insertions and deletions to two sets  $A$  and  $B$  from a ground set  $\Delta$  one can obtain an approximation to the Jaccard similarity of  $A$  and  $B$  with additive error  $\epsilon$  using space  $O(\log d/\epsilon^2)$ , where  $d$  is the size of the ground set. The result is obtained by a reduction to the case of estimating the number of distinct items in a data stream, i.e.,  $F_0$  estimation. The second result in this work is an algorithm for dynamic data streams that provides a locality sensitive hashing scheme to filter out pairs of low similarity. Our experiments show that the algorithms also perform well in practice.

Furthermore, we have studied the problem of  $k$ -centre and  $k$ -median clustering of time series of real values with respect to the Fréchet distance [A2/63]. Our input set consists of  $n$  time series each with at most  $m$  values. Since an optimal centre to the 1-median problem with respect to the Fréchet distance may be a time series of significantly more than  $m$  values, we add an additional restriction to the problem: We require that our cluster centres consist of time series of at most  $\ell$  values. This prevents overfitting and allows us to parametrise the runtime in terms of  $\ell$ . One of the challenges in clustering time series is that similar signals might be subsampled at different rates, so that the vectors representing the time series look very different at first sight. To handle this problem, we introduce the concept of  $\delta$ -signatures, which are subsampled representations of time series that satisfy certain approximation parameters. In our case, we can reduce time series of length  $m$  to time series of length  $\ell$  with minimal simplification error (using  $\delta$ -signatures). For applications with long time series and small  $\ell$ , this significantly reduces the space requirements. We then use the signatures to develop a  $(1 + \epsilon)$ -approximation algorithm for  $k$ -centre and  $k$ -median clustering (for our setting of centres with restricted complexity) in time  $O(nm \log m)$  for constant  $\epsilon, k$  and  $\ell$ .

We study  $k$ -median and  $k$ -means clustering in a setting that goes beyond worst case analysis in [A2/59]. We assume that if the input satisfies certain clustering stability conditions (distribution stability, spectral stability and perturbation resilience) then a solution computed by a standard local search algorithm has cost close to that of an optimal solution.

**Distributed Algorithms.** In a first approach to design a distributed algorithm for logistic regression together with project C4 we have developed a coresnet construction for logistic regression [A2/C4/66]. We have shown that if the input data satisfies a certain statistically well-motivated niceness assumption, then there is an algorithm to compute a coresnet of size  $O((\mu d \log n)/\epsilon)^{O(1)}$ , where  $\mu$  depends on the niceness of the data set. The algorithm samples points with probability proportional to the leverage scores (squared Euclidean norms of the rows of matrix  $U$  in a singular value decomposition of the input matrix  $A = U\Sigma V^T$ ) of the input data matrix and uses a variant [39] of the sensitivity framework of Feldman and Langberg [41] to analyse the quality of the construction. Then the construction is applied recursively to compute our final coresnet. We also show that without any niceness assumption on the data no coresnet of sublinear size exists. We have implemented our approach and compared it to uniform sampling and the  $k$ -means based coresnet construction for Bayesian regression by [45]. Since the computation of the leverage scores is rather time-consuming, we used a sketch-based approach to approximate them. While our construction is slower on most of the tested data sets, it yields solutions of significantly higher quality. Our experimental results can be seen in figure 3.1. While it is not immediately clear how we can generalise

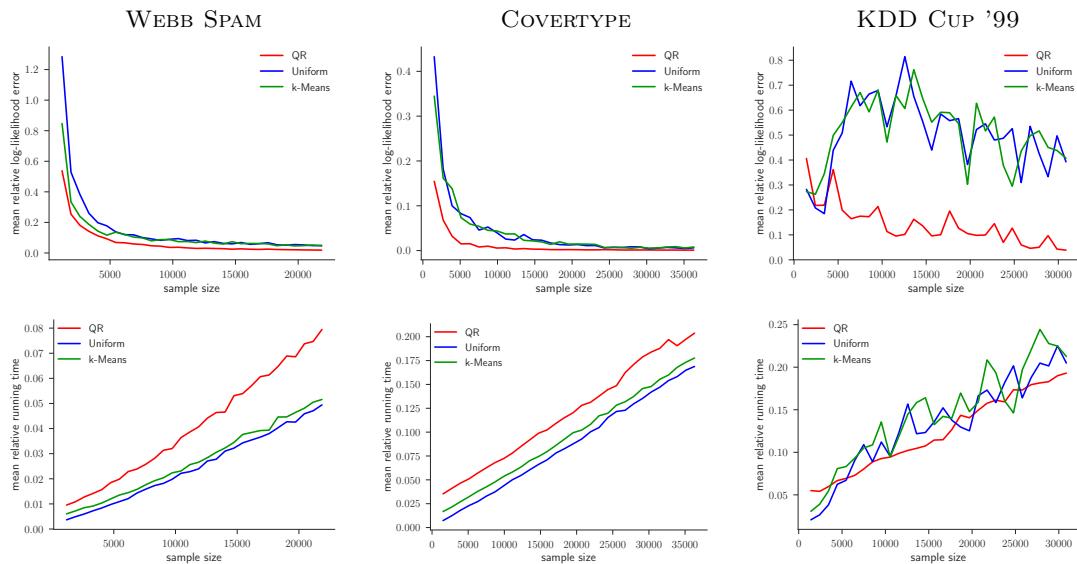


Figure 3.1: Each column shows the results for one data set comprising thirty different coresnet sizes (depending on the individual size of the data sets). The plotted values are standard deviations taken over twenty independent repetitions of each experiment. The plots show the standard deviations of the log-likelihood errors (upper row) relative to that of an optimal solution on the whole data set and standard deviations of the relative runtimes (lower row) of the three subsampling distributions, uniform sampling (blue), our leverage score based distribution, derived using QR decomposition (red), and the  $k$ -means-based distribution (green). All values are relative to the corresponding runtimes respectively optimal log-likelihood values of the optimisation task on the full data set. (Lower is better.)

our approach to distributed systems, we believe our algorithm is a first approach in this direction to obtain a coresnet based distributed algorithm with provable guarantees. The main remaining challenge is handling situations where the whole data set satisfies our niceness condition but the subsets on each server do not.

In another paper [A2/60] we examine algorithms that continuously monitor network features of an externally controlled network with the help of a second, internally controlled network; i.e., we can change the structure of the internally controlled network according to our needs. In our model, the

node set  $V$  is static and the undirected edge set changes over time. We further assume that time proceeds in synchronous rounds. In the  $i$ -th round the edges of the (external) network are given by a set  $E_i$  and the corresponding graph is called  $G_i = (V, E_i)$ . We assume that the node degree of the external network is polylogarithmic. Furthermore, we have an edge set  $D_i$  that contains the edges of the internal network. As soon as a node knows the identifier of another vertex, there is a corresponding edge in  $D_i$ . Initially, the set  $D_0$  is empty and the graph  $G_0$  is connected. Each node can send and receive at most polylogarithmically many bits to/from its neighbours in each communication round. In this setting we would like to compute properties of the external network. Since the external network is dynamic, its features change over time. We therefore say that an algorithm monitors a network feature with setup time  $t_0$  and delay  $\delta$ , if for every round  $i \geq t_0$  the algorithm reports the feature value of round  $i$  in round  $i + \delta$ . In the above model we develop scalable algorithms for computing the number of edges, the average node degree, the clustering coefficient, bipartiteness and the weight of a minimum spanning tree. Our algorithms have setup time at most  $O(\log^2 n)$  and delay  $O(\log n / \log \log n)$  except for bipartiteness, which has delay  $O(\log^2 n)$  and the computation and approximations of the minimum spanning tree weight, which assumes integer weights that are at least 1. In this case, the number of round depends on the maximum weight.

**Energy Consumption.** The interest of research and industry in the energy consumption characteristics of algorithms and systems keeps growing. During Phase 2 of the CRC, we performed a series of experiments to better understand the interactions between hardware and software with respect to energy consumption. It is easy to understand that the clock rate of a system influences a system's power consumption, i.e., its use of energy per time unit: higher clock frequencies result in a higher power consumption, approximately following the law

$$P_{dyn} \propto C_L V_{dd}^2 f ,$$

where  $f$  denotes the clock frequency. The necessary supply voltage  $V_{dd}$  has to be increased with the clock frequency, too, resulting in a super-linear power consumption. Although also more work can be performed per time unit with a higher clock rate, a low frequency results in a low energy consumption (for the overall task) under this model.

The discussion above only affects the dynamic power aspect of hardware, i.e., the energy needed to, e.g., flip the state of a hardware register. Tighter integration densities, however, result in higher leakage currents in modern microprocessor designs, also referred to as static power  $P_{stat}$ . This power consumption component is independent of the clock rate and can be responsible for up to 50 % of the total power consumption of a modern processor. The static power consumption component suggests using a high clock frequency, so as to finish a task early, then powering the CPU off – a strategy also referred to as race to idle.

The strong contribution of the static power component has made race to idle the folklore strategy that, e.g., most operating systems apply by default. The usefulness of that strategy, however, hinges on the assumption that programs run faster if clocked higher (proportionally, in an idealised world). That assumption is typically met by compute-intensive programs – which traditionally serve as easy benchmarks. The runtime of data-intensive programs, by contrast, is more often dominated by the time the program waits for memory (also called the memory wall). Making a CPU “faster” than memory only increases its power consumption, but does not improve the application speed.

As part of the work package “Energy Consumption” of the project proposal, we set up the necessary experimentation frameworks to verify and quantify the interactions between energy consumption on one end and the compute or data intensity of a task on the other. We studied the energy characteristics of typical database algorithms (including joins, hashing, and table scans) as well as of other data analysis algorithms (such as frequent item counting). A number of theses resulted from this work.

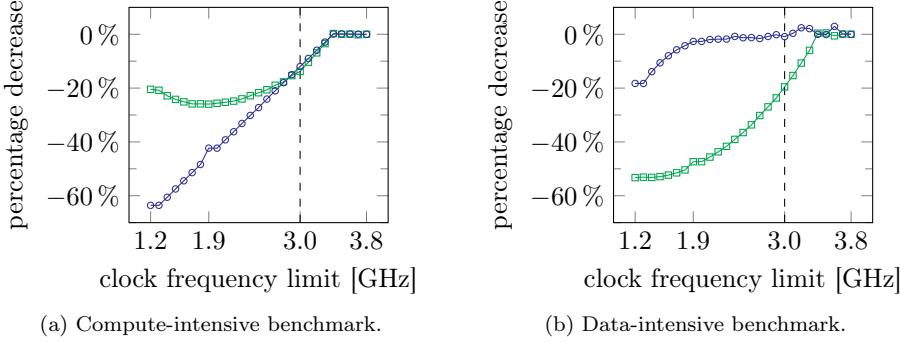


Figure 3.2: Percentage decrease of the energy consumption —□— and the runtime performance —○— of compute- and data-intensive benchmarks at different clock frequencies.

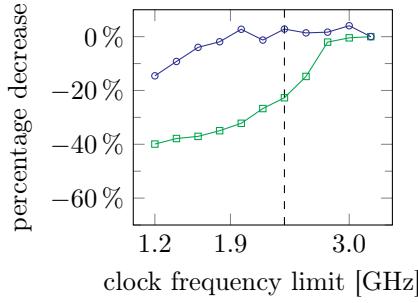


Figure 3.3: Percentage decrease of the energy consumption —□— and the runtime performance —○— when computing sketches for Bayesian regression.

In [A2/61], we experimentally studied energy characteristics for different workload classes. Figure 3.2 shows how runtime performance degrades much faster than any savings in energy for compute-intensive workloads (figure (a)). For data-intensive workloads (figure (b)), significant energy savings can be realised by tuning the clock frequency, while barely harming the programs runtime performance.

To assess whether this mechanism to reduce the energy consumption of data-intensive algorithms carries over beyond pure benchmarks, we studied the energy/performance characteristics of an algorithm to generate matrix sketches for Bayesian regression (along the sketching technique developed in Project C4 [C4/53]). The result is shown in figure 3.3 for a configuration where a large ( $\approx 17$  GB) matrix is sketched to a representation with about 500 MB. As can be seen in the figure, we can reduce the energy consumption of sketching by about 30 % with (almost) no impact on execution speed.

Understanding the different behaviours that algorithms exhibit is an important step in making data analysis systems more energy efficient. Actual database query execution plans, for instance, will run in phases with high compute demand and phases with high data demand. Deliberately balancing the two in conjunction with a dynamic and workload-aware frequency adjustment could allow a system to self-tune for the best energy-performance trade-off. On the conceptual side, our results prepare the basis for developing better models for energy consumption from an algorithm analysis perspective. In turn, such models may allow to design algorithms for energy efficiency from the ground up.

**Computation and Communication Architecture.** With hardware trending toward increased parallelism and heterogeneity, modern systems often look and behave like a “distributed system in a box.” This makes shared resources and communication channels between system components potential bottlenecks.

To understand and address the former class of bottlenecks, in the passing CRC phase we studied methods and algorithms to use and schedule access to shared caches. It is known that shared access to caches may benefit performance when parallel units collaborate (e.g., [56]), but also may seriously hurt performance because of cache pollution (e.g., [49]). By deliberately allocating cache space according to each database algorithm’s need, we showed in [A2/62] how the destructive pollution can be significantly reduced. The beauty of our technique is that it can easily be applied even to existing systems and improves performance for a meaningful class of database queries while at the same time not slowing down others. We evaluated our cache allocation strategy using the commercial SAP HANA platform and demonstrated that our results are not limited to idealised scenarios but provide an actual improvement in end-to-end performance for real-world applications.

Our work in [A2/65] addresses the latter class of bottlenecks, the communication channels between individual components of a heterogeneous system architecture. Specifically, we looked at bandwidth effects in a CPU/GPU heterogeneous system. In such configurations, the PCI Express link between host system and GPU is typically perceived as the critical limiter to performance. As we showed in [A2/65], that problem can actually be overcome with the use of batch-oriented processing.

Harder to eliminate is the bandwidth bottleneck behind the PCIe link. For data-intensive algorithms, the massive compute capacity available in modern graphics processors makes the bandwidth of the on-board GPU memory quickly the limiting factor, e.g., in GPU-accelerated database engines such as CoGaDB (for which the DBIS Group is an important contributor).

In [A2/65], we have developed a novel method to compile database query (sub)plans into native GPU code. Because of the GPU architecture model, this is a prerequisite for keeping shares of the data in GPU scratchpad memory, thus avoiding access to the bandwidth-limited on-board memory. Existing compilation strategies are at odds with the massively parallel, kernel-based execution model of graphics processors. Nevertheless, our novel query compiler HorseQC can compile entire query (sub)plans into fully pipelineable GPU code. As we showed in [A2/65], this can improve system performance for typical benchmarks by almost an order of magnitude.

## Bibliography

- [34] S. Albers, F. Müller, and S. Schmelzer. “Speed scaling on parallel processors”. In: *Algorithmica* 68.2 (2014), pp. 404–425 (cit. on p. 96).
- [35] A. Andoni and H. L. Nguyen. “Width of points in the streaming model”. In: *ACM Trans. Algorithms* 12.1 (2016), 5:1–5:10 (cit. on p. 95).
- [36] C. Balkesen, J. Teubner, G. Alonso, and M. T. Özu. “Multi-Core, Main-Memory Joins: Sort vs. Hash Revisited”. In: *Proceedings of the VLDB Endowment* 7.1 (Sept. 2013), pp. 85–96 (cit. on p. 97).
- [37] N. Bansal, K. Pruhs, and C. Stein. “Speed scaling for weighted flow time”. In: *SIAM Journal on Computing* 39.4 (2009), pp. 1294–1308 (cit. on pp. 96, 97).
- [38] B. D. Bingham and M. R. Greenstreet. “Computation with energy-time trade-offs: Models, algorithms and lower-bounds”. In: *Parallel and Distributed Processing with Applications, 2008. ISPA’08. International Symposium on*. IEEE. 2008, pp. 143–152 (cit. on p. 96).

- [39] V. Braverman, D. Feldman, and H. Lang. “New Frameworks for Offline and Streaming Coreset Constructions”. In: *arXiv preprint CoRR* abs/1612.00889 (2016) (cit. on p. 88).
- [40] K. Chen. “On Coresets for  $k$ -Median and  $k$ -Means Clustering in Metric and Euclidean Spaces and Their Applications”. In: *SIAM Journal on Computing* 39.3 (Aug. 2009), pp. 923–947 (cit. on p. 87).
- [41] D. Feldman and M. Langberg. “A unified framework for approximating and clustering data”. In: *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*. Ed. by L. Fortnow and S. P. Vadhan. ACM, 2011, pp. 569–578 (cit. on pp. 88, 94).
- [42] G. Frahling and C. Sohler. “Coresets in dynamic geometric data streams”. In: *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*. Ed. by H. N. Gabow and R. Fagin. ACM, 2005, pp. 209–217 (cit. on p. 87).
- [43] R. Gonzalez and M. Horowitz. “Energy dissipation in general purpose microprocessors”. In: *IEEE Journal of solid-state circuits* 31.9 (1996), pp. 1277–1284 (cit. on p. 96).
- [44] S. Har-Peled, D. Roth, and D. Zimak. “Maximum Margin Coresets for Active and Noise Tolerant Learning”. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. 2007, pp. 836–841 (cit. on p. 96).
- [45] J. H. Huggins, T. Campbell, and T. Broderick. “Coresets for Scalable Bayesian Logistic Regression”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2016, pp. 4080–4088 (cit. on pp. 88, 317, 327).
- [46] P. Indyk. “Algorithms for dynamic geometric problems over data streams”. In: *Proceedings of the thirty-sixth annual ACM Symposium on Theory of Computing*. Ed. by L. Babai. ACM, 2004, pp. 373–380 (cit. on p. 87).
- [47] P. Indyk and E. Price. “K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance”. In: *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*. Ed. by L. Fortnow and S. P. Vadhan. ACM, 2011, pp. 627–636 (cit. on p. 87).
- [48] S. Jana, J. Schuchart, and B. Chapman. “Analysis of energy and performance of pgas-based data access patterns”. In: *Proceedings of the 8th International Conference on Partitioned Global Address Space Programming Models*. ACM. 2014, p. 15 (cit. on p. 96).
- [49] R. Lee, X. Ding, F. Chen, Q. Lu, and X. Zhang. “MCC-DB: Minimizing Cache Conflicts in Multi-core Processors for Databases”. In: *Proc. of the VLDB Endowment* 2.1 (2009), pp. 373–384 (cit. on p. 91).
- [50] T. Neumann. “Efficiently Compiling Efficient Query Plans for Modern Hardware”. In: *Proceedings of the VLDB Endowment* 4.9 (2011), pp. 539–550 (cit. on p. 98).
- [51] I. Paul, V. Ravi, S. Manne, M. Arora, and S. Yalamanchili. “Coordinated energy management in heterogeneous processors”. In: *Scientific Programming* 22.2 (2014), pp. 93–108 (cit. on p. 96).
- [C5/52] P. Roy, **J. Teubner**, and 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, pp. 1451–1459 (cit. on p. 97).
- [C4/53] **L. N. Geppert, K. Ickstadt, A. Munteanu**, J. Quedenfeld, and **C. Sohler**. “Random projections for Bayesian regression”. In: *Statistics and Computing* 27.1 (2017), pp. 79–101 (cit. on pp. 90, 318, 319, 329).

- [54] D. P. Woodruff. "Sketching as a Tool for Numerical Linear Algebra". In: *Foundations and Trends in Theoretical Computer Science* 10.1-2 (2014), pp. 1–157 (cit. on p. 94).
- [55] F. Yao, A. Demers, and S. Shenker. "A scheduling model for reduced CPU energy". In: *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*. IEEE. 1995, pp. 374–382 (cit. on p. 96).
- [56] J. Zhou, J. Cieslewicz, K. A. Ross, and M. Shah. "Improving Database Performance on Simultaneous Multithreading Processors". In: *Proc. of the 31st Int'l Conference on Very Large Data Bases (VLDB)*. Trondheim, Norway, Aug. 2005, pp. 49–60 (cit. on p. 91).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [A2/57] V. Braverman, G. Frahling, H. Lang, **C. Sohler**, and L. F. Yang. "Clustering High Dimensional Dynamic Data Streams". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, Australia, August 6-11*. Ed. by D. Precup and Y. W. Teh. 2017 (cit. on pp. 18, 87, 94).
- [A2/58] M. Bury, **C. Schwiegelshohn**, and M. Sorella. "Sketch 'Em All: Approximate Similarity Search for Dynamic Data Streams". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, (WSDM) 2018, Los Angeles, CA, USA, February 6-8, 2018*. Ed. by Y. Maarek and Y. Liu. ACM, 2018 (cit. on pp. 18, 87).
- [A2/59] V. Cohen-Addad and **C. Schwiegelshohn**. "On the Local Structure of Stable Clustering Instances". In: *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2017), Berkeley, CA, October 15-17*. Ed. by C. Umans. 2017 (cit. on p. 87).
- [A2/60] R. Gmyr, K. Hinenthal, C. Scheideler, and **C. Sohler**. "Distributed Monitoring of Network Properties: The Power of Hybrid Networks". In: *44th International Colloquium, ICALP 2017, Warsaw, Poland, July 10-14. Proceedings*. Ed. by P. Indyk, F. Kuhn, and A. Muscholl. 2017 (cit. on p. 88).
- [A2/61] S. Noll, **H. Funke**, and **J. Teubner**. "Energy Efficiency in Main-Memory Databases". In: *Datenbank-Spektrum* (July 2017) (cit. on pp. 16, 90, 97).
- [A2/62] S. Noll, **J. Teubner**, N. May, and A. Böhm. "Accelerating Concurrent Workloads with CPU Cache Partitioning". In: *Proc. of the 34th IEEE Int'l Conference on Data Engineering (ICDE)*. Paris, France, Apr. 2018 (cit. on p. 91).
- [A2/63] **A. Driemel**, **A. Krivošija**, and **C. Sohler**. "Clustering time series under the Fréchet distance". In: *Proceedings of the 27th Symposium on Discrete Algorithms (SODA)*. Ed. by R. Krauthgamer. SIAM, 2016, pp. 766–785 (cit. on p. 87).
- [A2/C4/64] **A. Munteanu** and **C. Schwiegelshohn**. "Coresets - Methods and History: A Theoreticians Design Pattern for Approximation and Streaming Algorithms". In: *KI - Künstliche Intelligenz* 32.1 (2018), pp. 37–53 (cit. on pp. 18, 94, 316).
- [A2/65] **H. Funke**, **S. Breß**, S. Noll, V. Markl, and **J. Teubner**. "Pipelined Query Processing on Coprocessor Environments". In: *Proceedings of the 2018 ACM SIGMOD Conference on Management of Data*. June 2018 (cit. on pp. 91, 97, 98).

### b) Other publications

- [A2/C4/66] **A. Munteanu, C. Schwiegelshohn, C. Sohler**, and D. P. Woodruff. *On Coresets for Logistic Regression*. Tech. rep. arXiv:1805.08571 [cs.DS], 2018 (cit. on pp. 88, 316, 317, 326).

## 3.4 Project plan

### Goals

The main objective of this project is to develop design paradigms for learning algorithms in embedded systems. To this end, we will consider computational models that reflect important aspects of embedded systems, develop and analyse algorithms in these models, and empirically evaluate their performance to make conclusions about their efficiency, as well as the prediction power of the considered computational models. The insights gained in this way will then be used to derive general algorithmic design paradigms for embedded systems.

In order to achieve this goal we will continue to focus on data streaming algorithms, distributed computing, energy consumption and efficient use of hardware, which will also continue to be the structure of our work packages.

In the third phase we will concentrate on unifying the algorithmic findings and design paradigms obtained for our considered algorithmic aspects of embedded systems. We know from the projects' first two phases as well as the findings of other projects and researchers (see, for example, [41, A2/57, 54, A2/C4/64]) that small mergeable summaries of the data, such as sketches or coresets, are a suitable design paradigm for data streaming and distributed algorithms. Although common schemes exist, the design of mergeable summaries remains largely problem-dependent. Once we determine how mergeable summary is maintainable, general techniques, that turn the dependent algorithm to maintain such a summary into a streaming algorithm, can be applied. In the distributed setting, we maintain summaries locally and then aggregate them at a single computer. Interestingly, both processes share numerous similarities as they are typically derived from a tree-based data-aggregation process. In the streaming setting the merge & reduce approach reads a batch of data points and then computes a summary (which will become a leaf of the aggregation tree). Then the algorithm reads the next batch of points and computes a summary, which will become a neighbouring leaf. As long as there are neighbouring nodes in the aggregation tree, we merge them and compute a summary of the summaries and move it to the parent node. In the distributed setting a straightforward method to aggregate data is to merge them in a tree-like fashion as well.

We will therefore study this simple data-aggregation process as a prototypical example. Different aspects, such as its use in the streaming and distributed setting, as well as its energy consumption and implementations on modern hardware will be analysed in the project's work packages. The reasoning behind this approach is that the computation of the mergeable summaries is a process that typically exploits locality. For instance, in a distributed environment the summaries will be maintained at each computer. In many cases aggregating these summaries is the most costly operation (consider, for example, sensors connected by a wireless network). Even on a single machine it may still be beneficial to exploit locality in implementations. For example, if we have many cores, we can maintain summaries at each core's local cache and the costly operation will be to merge them. While this may not be evident at first sight, consider, for example, a machine with 32 cores running a sketch-based algorithm where the sketch size is 1% of the data. A reasonable scheme will first distribute the incoming stream on the 32 cores, so that each of them receives roughly 3% of the data. Since sketch dimension and data size are independent, each processor will

reduce this to 1% of the data, i.e., by a factor of 3. Thus, the overall size of the sketches to be aggregated is still about 1/3 of the input data and this requires non-local processing.

## Work schedule

Christian Sohler will work as a visiting researcher at Google Zürich until 9th of August 2019. During this time we will focus on WPs 3 and 4, whose input is also partially required for WP1. WP1 and WP 2 will start as soon as Christian Sohler is back in Dortmund. For the period from January to the end of July 2019 we are only applying for one additional staff position and two positions for the remainder of the project.

**Work package 1. Streaming Algorithms** In the streaming setting, we will study our aggregation process in greater detail to obtain a highly efficient, multi-threaded implementation of the merge & reduce technique. Findings from WP 3 and 4, on the energy consumption of merge & reduce trees and the efficient use of caches (with and without locks), will be integrated in our design. As a starting point, the load will be spread across cores by splitting the input stream in a round-robin fashion. We will pick a coresset size that is up to 50% of the size of the processor cache, so that two coressets fit into the cache (which will be ideal for a merge step). The first coressets will consist of points from the input stream. Then we compute a coresset on each core. Afterwards, we need to efficiently merge the coresets.

Meanwhile, our second direction of research aims to deepen our understanding of small summary computation. In the first two project phases, we obtained a good understanding of how to design streaming algorithms in the insertion-only setting, i.e., those cases in which a stream of feature vectors is available. Some open questions remain for problems with worst case lower bounds that rule out streaming algorithms, i.e., algorithms that process the data sequentially with polylogarithmic space in the input size. An example for such a problem is logistic regression. For cases such as these, we want to consider whether it is possible to find reasonable statistical assumptions on the input distribution and/or ordering of the points in the stream that allow us to obtain a streaming algorithm. In particular, we aim to continue our cooperation with project C4 on (logistic) regression. Although we have developed a coresset construction for this problem in the second phase of the project, it is not clear how to obtain a streaming algorithm, if we only have a statistical guarantee for the whole data set.

Another pertinent question that we will approach together with project A1 is the study of coresets under additional constraints like integrality that arise in embedded systems. If we require that the coreset points are integral, we get an additional assumption that is satisfied by some but not all coreset constructions. On the positive side, if we get the promise that solution and input points are integral, we can potentially exploit this to get smaller coresets. We will also consider coresets that approximate more than one objective function and/or constraint.

In collaboration with A6, we will investigate the use of coresets to reduce the size of spatiotemporal graphs. Such graphs appear, for example, in the context of path planning of unmanned aerial vehicles and will be relevant for project B2 as well.

Further topics open for research concern dynamic data streams. In this model, we obtain a stream of insertions and deletions from a universe  $\{1, \dots, \Delta\}^d$ . A particularly relevant challenge in this context is the creation of an algorithm for linear subspace approximation. One approach towards an algorithm is to first get a solution in the low-dimensional setting (possibly building on the work of Andoni on the width of points [35]) and combine it with a dimensionality-reduction technique. Somewhat surprisingly, dimensionality reductions using Johnson-Lindenstrauss (or similar) embedding have not been used as dimensionality reduction techniques in this model. One possible

reason might be that the resulting grid is slightly distorted, which results in algorithmic problems. A promising approach along these lines is the combination of a low dimensional approach with such an embedding result. Similar questions also arise for the  $k$ -median problem when one wants to reduce the space dependence on the dimension.

Furthermore, it is advantageous to develop a sketch-based algorithm for logistic regression, which would enable the handling of dynamic data streams. We also plan to consider other linear classifiers such as support vector machines (SVMs) in dynamic data streams. For SVMs a coresset construction is known [44], but it is unclear how it can be maintained during insertions and deletions under space constraints. Moreover, it is not too hard to see that without additional assumptions on the input, the problem does not admit a dynamic streaming algorithm with sublinear space. Identifying reasonable assumptions, under which dynamic streaming algorithms can be obtained, is also of significance. Similar questions can also be formulated for the non-separable case, which is technically closer to the case of logistic regression.

**Work package 2. Distributed Algorithms** For analysing energy and time efficiency of distributed systems, we employ the computational model suggested in [38]. In this model, computation and communication speed can be adjusted by the algorithm at the cost of increased energy consumption. In order to achieve a speed up factor  $s > 1$ , the energy consumption has to be increased by a factor  $s^\alpha$  with a constant  $\alpha > 1$  determined by the system. Convex energy costs of computation were introduced in [55] and often used in the scheduling community; see [34, 37] and the references therein. To evaluate and compare systems in terms of both energy and time, the *energy delay product* ( $\alpha = 1$ ) and *energy delay squared* ( $\alpha = 2$ ) are common measures (see [43, 48, 51]). The overall cost  $C$  of a system in this model is given as  $C = \mathcal{E}\mathcal{T}^\alpha$ , where  $\mathcal{E}$  is the systems energy consumption over the course of computation and  $\mathcal{T}$  is its required time. A task consisting of  $f$  elementary steps, each executed sequentially on a single processor and in unit time, has costs of  $f^{\alpha+1}$ .

We plan to consider our prototypical data-aggregation process (merge & reduce) in a distributed setting and to minimise the energy delay product. We will assume synchronous communication rounds, and we assume that we can transfer the aggregated data in one communication round to any node in the network (while paying an increased cost, if the node is not a direct neighbour). A different view of this problem is that the cost of communicating with another node in the network is given as the  $\alpha$ -power of a shortest-path metric of an unweighted graph (later on, we may also consider weighted graphs). Our goal is to find a rooted low-degree spanning tree that minimises the energy delay product (where the delay is simply the depth of the tree). In order to solve this problem, we plan to exploit connections to metric embeddings into distributions over tree metrics. This entails approximating the shortest path metric by a distribution of tree metrics (possibly adapting existing approaches to our problem) and then solving the problem of finding the low-degree spanning tree.

Together with project C4, we intend to continue the study of coresets for logistic regression. Here, our focus will be on the distributed computation of coresets. Similarly to streaming, coressets cannot simply be computed and aggregated straightforwardly. This is due to assumptions on the data set, that may not be satisfied for subsets stored at distributed network nodes. We will evaluate, if coressets can be computed via a distributed algorithm by only making assumptions about the distribution of the complete data set.

**Work package 3. Energy Consumption** In the ongoing project phase, we studied energy consumption empirically, mainly in microbenchmark settings. Our results show that the algorithm's characteristics, in particular their use of memory and bandwidth, can have a significant impact on

its energy efficiency. In recent [A2/65] and earlier [36, C5/52] work, we have developed techniques to understand and engineer the bandwidth need of data-centric algorithms. In phase 3 of the CRC, we want to (*a*) combine such results with the goal of understanding and further improving energy efficiency and (*b*) extend our assessment towards real-world settings.

With regard to (*a*), we expect that the observed effects will interact. Taking the example of our prototypical hierarchical merge & reduce operation, our current results suggest the following strategy: Run the lower levels of the hierarchy at a low frequency (to conserve power), but use a higher frequency for the upper levels (to improve performance (a similar strategy was also proposed by [37]). In earlier work, we studied the characteristics of merge sort implementations [36]; here, it may be beneficial to run compute-intensive in-cache operations at high clock frequencies and bandwidth-intensive out-of-cache merges at lower frequencies. Similar patterns can be observed in further data analysis settings, where algorithm phases with different bandwidth/compute characteristics take turns. In the third phase of the CRC, we want to verify, both empirically and theoretically, the effectiveness of strategies that exploit knowledge of an algorithm's bandwidth/compute behaviour. Combinations with, e.g., cache-partitioning strategies may be of interest, too, with respect to the resulting energy efficiency.

It is not entirely clear how our microbenchmark results translate into observable energy savings in end-to-end settings. On the one hand, additional complexity may diminish some of the energy saving effects. On the other hand, the composition of building blocks (to which we can apply our microbenchmark results) into larger algorithms opens up the possibility to combine these building blocks in different ways; our existing results indicate that balancing, e.g., compute- and bandwidth-intensive building blocks in a proper way could improve energy efficiency further. To illustrate, it may be possible to schedule out-of-cache and in-cache phases of merge sort to run concurrently, rather than sequentially, to improve balancing. Likewise, it may be beneficial to start computing  $k$ -means over a coresset in parallel to the coresset construction. Such a strategy could take advantage of unused compute resources during coresset construction and support a faster convergence of the  $k$ -means computation on the final coresset.

Related to that could be strategies to trade performance for energy efficiency, which we also expect to develop in this context (for instance, an application goal could be to minimise energy consumption while meeting a set performance requirement).

**Work package 4. Computation and Communication Architecture** In one of our most recent results [A2/61], we leveraged Intel's *Cache Allocation Technology (CAT)* to improve robustness and performance in a data-processing system.

CAT is a novel feature that allows to assign cache partitions to individual cores in a multi-core environment. CAT gives the programmer a mechanism to control cache allocation, which otherwise is fully at the hardware's discretion. The poster child use case for explicit cache allocation is to isolate parallel execution streams from one another. In particular, CAT can avoid cache-pollution effects when parallel threads access the same cache. In [A2/61], we quantified these pollution effects for a specific situation that arises in database systems: Scan operators that run concurrent to queries that are known to be cache-sensitive. And we devised a simple mechanism, based on heuristics, to mostly avoid the pollution effect and improve performance.

During Phase 3, we would like to generalise the mechanism we have developed (e.g., to further workload scenarios) and also put it onto solid foundations (rather than only using ad hoc heuristics).

Still inside the database world, real-world situations will exhibit arbitrary combinations of co-running queries and/or operators. To allocate cache space using CAT, the system must be able to predict each query/operator's cache demand. To this end, we need more detailed cache access

## Project A2

models for different operators. We plan to derive (and experimentally validate) these for relevant database and data analysis operators. Based on cache-access models, we can then develop strategies to allocate cache partitions for parallel operators at runtime. Optimal cache allocation – a variant of the bin-packing problem – most likely will be infeasible, and we plan to develop meaningful heuristics to decide cache allocation online.

Beyond the database domain, the cache interaction of many analysis and/or learning tasks resembles that of database operations. For instance, the construction of sketches has many similarities with the population of database hash tables. Therefore, in Phase 3 we plan to carry our insights to such tasks.

We expect that cache allocation will interact with the balancing of workload types (cf. WP 3) in a constructive way. Specifically, compute-intensive components of an algorithm (e.g., coreset construction or in-cache sorting) may depend on data structures to be resident within caches. A co-running bandwidth-intensive task, however, will likely pollute caches and both tasks may interfere. With explicit control over cache allocation, such interference could be avoided, therefore helping to harvest the benefits of workload balancing.

The Intel CAT feature specifically addresses the L3 cache of modern Intel CPUs. Looking at the functionality from a further distance, its mechanisms could well be useful in different classes of hardware or at different levels of the memory hierarchy. The latter view could become relevant, e.g., when looking at distributed settings, where local RAM could be treated like a cache to remote memories – and organised similarly to the way we successfully demonstrated L3 cache allocation.

From a different end, we addressed a system’s “communication architecture” by means of query compilation in heterogeneous CPU/GPU architectures. Query compilation to scalar CPU code seems well understood by now (following the ground-breaking work of Neumann [50]). The compilation problem becomes significantly more difficult in settings where the compiler has to address a parallel architecture. Our results in [A2/65] show a solution to target GPU architectures. In Phase 3, we plan to extend these results to target, e.g., SIMD instruction sets of modern processors, OpenMP-style parallelism, or environments with hybrid CPU/GPU processing.

## Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Streaming Algorithms																	
2. Distributed Algorithms																	
3. Energy Consumption																	
4. Computation and Communication Architecture																	

### 3.5 Role within the Collaborative Research Centre

Similarly to project A1, this project studies learning methods in embedded systems. A1 mainly focuses on studying models for abstract learning tasks, as well as predicting their resource requirements. In contrast project A2 develops algorithm design paradigms for learning methods in embedded systems.

Together with A1 we will study coresets under additional constraints like integrality that arise due to computational restrictions in embedded systems. This aspect concerns models as well as algorithms and therefore constitutes a commonality of both projects. Algorithmic techniques studied in A2, such as coresets, have been applied in many other projects, for example in A1, A6, B4, or C4. We plan, in particular, to continue our cooperation with C4 on coresets for regression problems.

The more general concept of a compact data structure to accelerate the handling of very large data also proved instrumental in project C5 (in the form of the *DeLorean* system), a connection that we plan to follow also in the coming project phase.

Together with A6 we plan to extend coresets to spatiotemporal graphs. These results will also be relevant for project B2 in the context of local path planning for unmanned aerial vehicles.

### 3.6 Differentiation from other funded projects

#### **ERC Starting Grant**

**(Sohler, Reference number SUBLINEAR 307696) (Funding period: 2012–2018)**

The project deals with the analysis of property testing algorithms for sparse graphs and is not related to this proposal.

#### **MERCUR Project: “LPN-Krypt: Das LPN-Problem in der Kryptografie”**

**(Sohler, Reference number Pr-2016-0045) (Funding period: 2017–2019)**

The project studies the learning-parity-with-noise (LPN) problem and its variants from the perspective of (post-quantum) cryptography. It is not related to this proposal.

#### **DFG Priority Program 2037 · MxKernel**

**(Teubner, Reference number TE 111/2-1) (Funding period: 2017–2020)**

Jens Teubner is a co-initiator of the DFG-funded priority program “Scalable Data Management for Future Hardware” and PI in the project “MxKernel: A Bare-Metal Runtime System for Database Operations on Heterogeneous Many-Core Hardware.” The project targets the co-design of database and operating system software and is not related to this proposal.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	2	91,400	2	129,000	2	129,000	2	129,000
Total	—	91,400	—	129,000	—	129,000	—	129,000
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	91,400		129,000		129,000		129,000	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Christian Sohler, Prof. Dr., professor	Algorithmic complexity theory	TU Dortmund	4	—	Existing funds
	2	Jens Teubner, Prof. Dr., professor	Information systems	TU Dortmund	6	—	Existing funds
	3	Henning Funke, M.Sc., doctoral researcher	Information systems	TU Dortmund	29.88	—	Existing funds
	4	David Mezlaf, M.Sc., doctoral researcher	Algorithmic complexity theory	TU Dortmund	29.88	—	Existing funds
Non-research staff	5	Alla Stankjawitschene, secretary	—	TU Dortmund	2	—	Existing funds
<b>Requested staff</b>							
Research staff	6	N.N., doctoral researcher	Information systems	TU Dortmund	—	Doctoral researcher	—
	7	Amer Krivošija, M.Sc., doctoral researcher	Algorithmic complexity theory	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

**1. Sohler, Christian**

Project management with focus on the algorithmic aspects of the project. Until 9th of August 2019 Christian Sohler will work as a visiting researcher at Google Zürich. His contribution to the project will start after that date.

**2. Teubner, Jens**

Project coordination. Jens Teubner will focus on experimental, energy and architecture/communication, aspects.

**3. Funke, Henning**

Henning Funke will work on energy measurements, communication-aware algorithms, and code compilation for communication efficiency (SIMD, CPU/GPU processing, etc.).

**4. Mezlafl, David**

David Mezlafl will work on questions in WPs 1 and 2 with focus on energy consumption of algorithms.

**5. Stankjawitschene, Alla**

Administrative support.

**Job descriptions of staff for the proposed funding period (requested funds):**

**6. N.N.**

Will work on energy measurements (together with Henning Funke); cache- and communication-aware algorithms; CAT.

**7. Krivošija, Amer**

Amer Krivošija will work on problems in WPs 1 and 2 with focus on streaming algorithms.

### **3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from university	7,500	7,500	7,500	7,500
Sum of existing funds	7,500	7,500	7,500	7,500
Sum of requested funds	0	0	0	0

(All figures in euros)

### **3.7.5 Requested funding for instrumentation for the new funding period**

This project does not request any funding for major research instrumentation.

### 3.1 General information about Project A3

### 3.1.1 Project title:

## Methods for Efficient Resource Utilization in Machine Learning Algorithms

### 3.1.2 Research area(s):

409-07 (Embedded systems), 409-05 (Interactive and Intelligent Systems)

### **3.1.3 Principal investigator(s)**

Chen, Jian-Jia, Prof. Dr., 09.05.1978, Taiwanese

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

Phone: 0231-755-6078  
E-mail: jian-jia.chen@tu-dortmund.de

Rahnenführer, Jörg, Prof. Dr., 19.05.1971, German

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

Phone: 0231-755-3121  
E-mail: joerg.rahnenfuehrer@tu-dortmund.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

( ) no (x) yes

Do any of the above mentioned persons hold fixed-term positions?

(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes	(x) no
2.	clinical trials	( ) yes	(x) no
3.	experiments involving vertebrates.	( ) yes	(x) no
4.	experiments involving recombinant DNA.	( ) yes	(x) no
5.	research involving human embryonic stem cells.	( ) yes	(x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes	(x) no

## 3.2 Summary

The goal of this project is the development of methods for algorithm selection and configuration under resource constraints. Especially of interest are scenarios where a single evaluation of an algorithm is expensive. We address the case where many competing candidate algorithms are available and each algorithm has specific hyperparameters that have to be tuned to obtain the best possible outcome. It is not possible to exhaustively search through this space, because the number of configurations that can be evaluated during the optimisation is heavily limited due to their long runtimes. In a typical situation the algorithm is a machine learning method that is applied on a regression or classification data set. The optimisation goal is to find a machine learning method from a set of candidates and configure its hyperparameters to achieve the best possible prediction quality.

Model-based optimisation (MBO) addresses this challenge. It uses a regression model as a surrogate to approximate the objective function. For example, the prediction quality of a machine learning algorithm on a given task is predicted by a Gaussian process regression. The predictions obtained from the surrogate model help to move quickly to regions in the search space with promising prediction quality. This reduces the number of expensive evaluations during the optimisation. Due to the enormous number of possible configurations, the overall MBO wall-clock runtime still can be unreasonably long. The use of parallel computing systems and efficient resource utilisation becomes essential.

To address these challenges, we have developed the framework *resource-aware model-based optimization* (RAMBO) with scheduling for heterogeneous runtimes. It extends MBO to work on parallel systems while maximising resource utilisation. Thus, methods applied for the optimisation to achieve high efficiency of embedded systems can be adapted for improving the parallel execution of machine learning tasks in modern computing systems.

In the third phase, we want to investigate scenarios with additional challenges, such as time-varying objective functions, insufficient prediction quality due to small sample sizes, and data streams. Efficient solutions for these challenges can be obtained by extending the RAMBO framework and by further improving the scheduling strategies. Time-varying objective functions are common in the real world, i.e., the best-performing algorithm configuration changes over time, which is generally called concept drift. For small sample sizes, we will investigate the combination of real data and additional simulated data. For data streams the machine learning method will adapt to changes in order to optimise the prediction quality. These extensions of RAMBO make it usable for a wide range of applications, including data rate prediction for mobile phones as well as traffic analysis and prediction.

To evaluate the new methods, we use established benchmarks that provide a controlled environment to draw objective conclusions. In a second step we will also verify our proposed approach with real-world data. In addition to the developed methods themselves, a major outcome of this project are self-contained and well-documented open-source software packages, assuring the reproducibility of the experiments and future usability for other researchers around the world.

## 3.3 Project progress to date

### 3.3.1 Report and current state of research

The goal of the project is the development of methods for algorithm configuration under resource constraints. Of particular interest are scenarios where a single evaluation of an algorithm is time

consuming. Typically many competing candidate algorithms are available, and an exhaustive evaluation of all of them is not feasible. Here, we follow two main strategies: the use of model-based optimisation for an efficient selection of the best algorithm configurations, and resource management to efficiently evaluate as many algorithms at the same time as possible without wasting resources.

The main tasks in the second phase were the development and comparison of methods for automated algorithm configuration, the development of scheduling strategies that optimise these methods with respect to resource management, and the development of methods for an optimal selection of prediction algorithms for clinically or genetically predefined subgroups of patients in the situation of large patient heterogeneity. First, algorithm development and implementation for automated algorithm configuration are summarised, and then our results regarding general optimisation of resource efficiency are described. Afterwards, our new framework for *resource-aware model-based optimization* (RAMBO) and corresponding results are described. Finally, the first results for medical predictions in the situation of heterogeneous subcohorts and methods for stable variable selection are presented.

**Automatic algorithm selection and configuration** The wide array of available methods to solve machine learning problems makes it increasingly difficult to find the best-performing method for a given data situation. There are no general, reliable rules that lead to a suitable choice of a specific method based on characteristics of the problem. Conventionally, an algorithm is selected through an exhaustive search or human domain knowledge. Once a well-performing algorithm is found, its settings (the so called hyperparameters) can be tuned by trying out different configurations in order to decide for the setting that gives the best prediction quality. This approach has some major drawbacks: First, it is expensive for the user to perform all steps manually and decide for one or multiple methods. The user has to invest time to apply those methods to the problem and finally has to invest more time to try different hyperparameter settings. Second, it is not guaranteed that the user does not overlook a promising method. Finally, in the process the user might have dismissed method A in favour of method B, as B initially performed better, although there is the possibility that with tuned hyperparameters A outperforms B.

To overcome most of these problems, an easily accessible automatic algorithm configuration tool is essential. For a given problem it can be used to find the machine learning algorithm and automatically configures its hyperparameters reliably to yield the best prediction quality. To that end, various methods with different hyperparameters have to be automatically evaluated on the data. However, with increasing data size, these evaluations become very time-consuming and it becomes impossible to search for the best method in an exhaustive way. The evaluation of a machine learning algorithm with a specific setting on a given data set can be thought as a black box. The choice of the method and its hyperparameters are the input, and the prediction quality, which is evaluated with a cross-validation or a similar method, is the output. The relation between the input and output is typically unknown and can only be derived by evaluating the black box. Sequential **model-based optimisation (MBO)** [76] is a state of the art [75, 87, 86] technique for such expensive black box optimisation problems. In comparison to other black box optimisation methods, like Genetic Algorithms or Simulated Annealing, MBO is favourable when evaluating a configuration takes a lot of time and resources.

Formally, the space of possible algorithms and their respective hyperparameters form the search space  $\Theta$  for the optimisation problem:

$$\theta^* := \operatorname{argmin}_{\theta \in \Theta} f(\theta),$$

where  $f(\theta)$  denotes the evaluation of the black box with the input  $\theta \in \Theta$ . To reduce the number of evaluations on  $f$  the key idea of MBO is to only evaluate values of  $\theta$  that are expected to lead to a small value of  $f(\theta)$ . The estimate  $\hat{f}(\theta)$  is generated by a so-called surrogate model. Typically this is a regression model that predicts the outcome of  $f$  based on previous evaluations

### Project A3

of  $f$ . Therefore, an initial design  $\theta$  of already evaluated configurations is needed. Iteratively the MBO algorithm fits the surrogate on the previous evaluations, proposes a configuration  $\theta$  and evaluates it on  $f$ . An infill criterion guides the proposal of new configurations  $\theta$  based on  $\hat{f}$ . It balances between exploration of not yet evaluated regions in  $\Theta$  and exploitation, i.e., the search on regions that promise the best outcomes. These steps are repeated until a budget, typically a limited wall-clock time, is exhausted. The setting  $\theta$  that leads to the best outcome is returned as the result. In this research project the result typically is a suggested machine learning algorithm with a specific hyperparameter configuration. We note that the MBO can also be applied for optimisation problems.

The building blocks of the MBO process are implemented in self-contained software packages. This allows for great flexibility and use for further applications. It also assures the reproducibility of the experiments, as the software is published as open source and well documented so it is easily usable by other researchers. The programming language R and its package system serve as a suitable environment. Software development and the integration of new methods in up-to-date software packages are very important and also part of the achievements of this project. The packages described in the following were also further developed within the project. The package `mlr` provides a consistent interface to a wide range of machine learning methods including classification, regression, survival analysis, and others. Additionally, preprocessing and visualising functions are included. It serves as a foundation for all machine learning experiments in our research. The continuous development together with researchers around the globe ensures that state of the art methods are included and can be extensively benchmarked. The framework of all our model-based optimisation methods is provided by `mlrMBO`. In its modular fashion, it allows virtually all components of the model-based optimisation algorithm to be changed. This makes it easy to re-implement published variants of MBO as well as to develop new algorithms, allowing a comprehensive and fair comparison as well as in-depth analysis of optimisation runs. `mlrMBO` also supports multi-criteria optimisation and was successfully applied to optimise the energy consumption, the runtime, and the detection rate of the PAMONO virus detection sensor from project B2.

One major challenge in automatic algorithm configuration is the selection of the candidate algorithms and the definition of the ranges of the hyperparameters that have to be optimised. Too many candidates increase the cost of the optimisation and too few might lead to the neglect of the best-performing ones. The same holds for the size and the number of hyperparameters for each method. The right choice of the search space can be crucial for the performance, making expert knowledge still essential. Therefore, this knowledge should be easily accessible. To make this possible `mlrHyperopt` (<https://jakob-r.github.io/mlrHyperopt>) was developed. The R-package connects to a web service that allows storing, sharing, and discussion of these search spaces. Furthermore, it interfaces `mlrMBO` for automatic model-based optimisation of machine learning methods based on the search spaces from the web service.

A popular example of an automatic algorithm configuration implementation is the GUI-based Auto-WEKA [89], which is restricted to the learners implemented in WEKA and not capable of survival analysis, in contrast to `mlrMBO`. Also, its capabilities to handle computation-intensive machine learning problems are very limited, due to the design of the software and because many evaluations are required to find the best learner and hyperparameter configuration.

The size of the data sets, for which the algorithms should be optimised, keeps increasing. A single evaluation of one hyperparameter setting can take up to several hours. Thus it becomes important to efficiently parallelise the optimisation process. However, our research has shown that pure parallelisation does not necessarily lead to efficient algorithms. Therefore, we focus on resource efficiency, which is discussed in the following paragraph.

**Resource efficiency** For resource efficiency, performance and memory consumption are critical aspects that can lead to extremely long execution time for single-threaded algorithms as well as for parallel multi-threaded algorithms. To efficiently utilise the available resources, e.g., processing power, memory, and accelerators, with respect to response time, energy consumption, and power dissipation, effective resource management strategies have been developed. To reduce the memory overhead of machine learning algorithms, a new optimisation based on dynamic sharing of memory was developed [A3/92]. This optimisation avoids duplication of page contents for large data structures and optimises the copy-on-write mechanism of the R language, which results in about 35% memory usage reduction with almost negligible runtime overhead. In addition to this new memory optimisation, parallelisation of the execution is used to speed up computation, which poses new resource utilisation challenges. For hyperparameter optimisation, the analysis with TraceR (developed in the first phase of the project) shows that a high runtime variance in the configuration space can cause inefficient resource utilisation due to different completion times of evaluations running in parallel (results were published in a technical report). In the yearly RIOT (R Implementation, Optimisation and Tooling) workshop we contributed talks on TraceR (2015) and on resource-aware scheduling strategies (2016/2017) and participated in the programme committee. The results obtained with TraceR motivated the development of new resource-aware scheduling methods for MBO.

For multiprocessor synchronisation, we have developed a *resource-oriented partitioned* (ROP) scheduling scheme [B2/21]. Since the shared resources are usually the bottlenecks, ROP changes the view angle and focuses on the *shared resources* instead of the *computing tasks*. The spirit behind ROP is to first assign each shared resource to one designated *synchronisation* processor, and the non-critical sections will be executed on other *application* processors, decoupled from the critical sections. Similar flipped views of resource management in modern multi-core systems have been recently utilised to handle bus contention [B2/97], GPU acceleration, power management, reliability enhancement, etc., where Jian-Jia Chen's group at TU Dortmund has made significant contributions in the recently years.

To optimise parallel algorithms not only for homogeneous architectures, we proposed a resource-aware scheduling strategy for heterogeneous architectures, like those commonly found in mobile embedded systems. Such devices typically consist of different processors with different frequencies and memory sizes. However, the parallel package in R distribution does not support heterogeneous architectures. To support heterogeneous systems within R, we extended the parallel package. The Core-R Team invited us to present our results at the DSC (directions of statistical computing, <https://www.r-project.org/dsc/2017/>) which is an invitation-only conference, and our improvements were integrated into the R distribution. Furthermore, we were invited to join the program committee of the UseR! 2017 conference.

When MBO is conducted on heterogeneous systems, the execution time of an evaluation of a configuration can vary heavily not only depending on the configuration but also on the underlying architecture. Key to our approach is a regression model that estimates the execution time of a configuration for each available processor type. In cooperation with the project B2, we demonstrated the effectiveness of our approach targeting the ARM big.LITTLE architecture, commonly found in mobile phones, e.g., Odroid-XU3. Within an additional cooperation with the project B2, the development of resource-aware scheduling strategies for heterogeneous systems was supplemented with the development of strategies for evaluating the efficiency of approximation techniques that are used to reduce the resource demands in embedded systems [A3/B2/96].

**Resource-aware model-based optimisation** The *resource-aware model-based optimization* framework (RAMBO) overcomes problems that occurred when parallelised model-based optimisation was carried out on problems with heterogeneous runtimes. For computer experiments, parallelisation has become of increasing interest to reduce the overall computation time, but it is not always feasible to parallelise the black box function itself. Originally, the MBO algorithm sequentially

proposes one point to be evaluated after another. To allow for parallelisation, techniques have been suggested that propose multiple configurations in each iteration. Multiple workers (e.g., CPUs) can be assigned to execute different configurations. After all evaluations are finished, the surrogate model is updated with the results and new candidates are generated. This leads to inefficient usage of resources when the runtime of the evaluation of the experiments is heterogeneous. Some evaluations will end earlier and the respective workers will idle.

To overcome this problem we enhanced the MBO framework with an additional regression that models the runtime of the evaluations in dependency of the hyperparameters of the configurations and combines the runtime predictions with scheduling, as outlined in figure 3.1. This way the

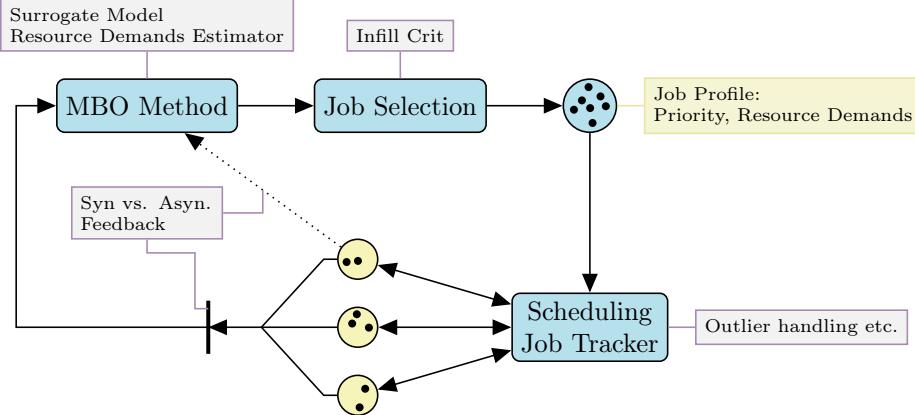


Figure 3.1: RAMBO Framework: Each job consists of one black box configuration. The resource predictions and the priority are used to generate a scheduling plan. In the asynchronous case the results are fed back into the surrogate as soon as they are available. In the synchronous case all results are aggregated before reiterating.

runtime for newly proposed configurations  $\theta$  can be estimated. The estimated runtimes are then used to select a set of jobs that can be scheduled such that idling times are avoided or minimised. The available time is used more efficiently, and the optimisation progresses faster. A first heuristic was presented in [A3/95]. To create multiple candidates the qLCB criterion [74] is used, by optimising multiple differently parameterised lower confidence bounds. A high parameter value leads to an infill function that favours exploration, whereas a lower value leads to the evaluation of points closer to the expected minimum, thus to more exploitation. The priority of a candidate is defined such that it balances exploration and exploitation of the surrogate. In a greedy first-fit heuristic, these candidates were mapped to the respective workers under the restriction that no worker runs longer than the candidate with the highest priority. This yields an improved hyperparameter optimisation result for an SVM within a limited wall-clock time budget. By being able to do more evaluations the confidence in the results can be increased.

Since the scheduling problem cannot be directly solved with conventional scheduling algorithms, as neither the set of points that has to be evaluated nor the wall-time is fixed, we developed an advanced heuristic in [A3/94]. First, a set of candidates is generated using an established multi-point proposal method. A priority, derived from the infill criterion function, is assigned to each candidate. To avoid simultaneous evaluation of similar configurations, the candidates are *clustered* hierarchically by their distances in the search space  $\Theta$ . Therefore, we assign a loss-value to each candidate accordingly. We start with  $k = 1$  and separate the set of candidates into  $k$  clusters. Clusters that already contain candidates with a loss-value are neglected. Then from the remaining candidates the one with the highest priority is selected, and the loss value  $k$  is assigned to this candidate. Finally, we set  $k = k + 1$  and reiterate until all candidates have loss values. Afterwards, we use knapsack scheduling to select the best subset of jobs that minimises the loss under the restrictions that no worker should take more time than the predicted runtime of the job with the

highest priority and that each job should only be run once. This way we achieve a balance between the execution of many jobs to avoid idling and the execution of only a few jobs to ensure that only configurations are evaluated that are purposeful for the search of the optimal configuration.

For the benchmark, we additionally implemented an asynchronous parallel model-based optimisation as presented in [72] (`asyn.eei`) and a faster approximation `asyn.ei.bel`. As can be seen in Figure 3.2 the advanced parallel methods drastically improved the utilisation in comparison to the simple multi-point proposals (`ei.bel` and `qLCB` [74]). On setups with accurate runtime

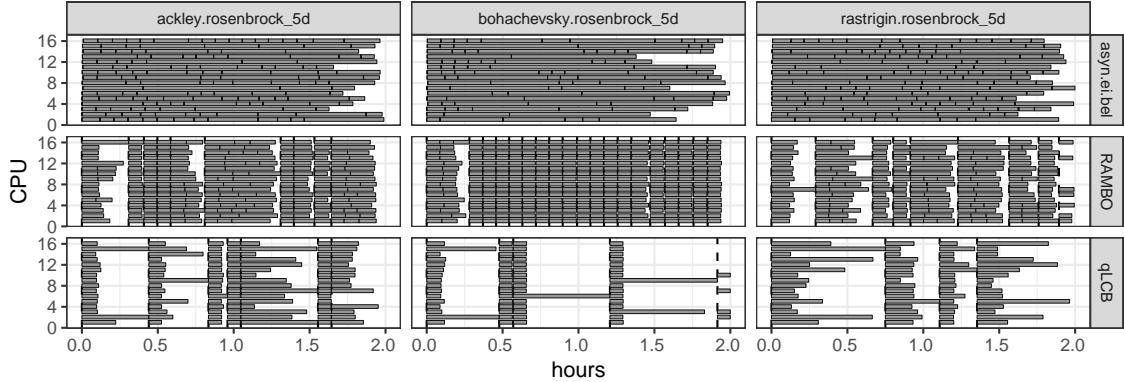


Figure 3.2: Scheduling of MBO algorithms. Time on  $x$ -axis and mapping of candidates to  $m = 16$  CPUs on  $y$ -axis. Each gray box represents a job. The gaps represent CPU idle time.

estimation quality, RAMBO converged faster to the optima than the competing MBO algorithms on average. This indicates, that the resource utilisation obtained by our new approach improves MBO, especially, when the number of available CPUs increases and for higher dimensional problems.

**Optimisation for patient cohorts with large numbers of covariates** One further specific application scenario of MBO is to find the best classification or regression method for patient cohorts with very large numbers of covariates (features), especially in the context of modern medical studies where thousands of genomic variables can be quantified simultaneously for each sample. An additional frequent problem in medical research, for example when constructing prediction methods for diagnosis or therapy choice of cancer patients, is the heterogeneity of patient cohorts. This scenario requires a computationally expensive evaluation of statistical learning algorithms, since the subgroup structure of the patient cohorts should be taken into account. This yields an additional level of complexity. The typical situation is that a statistical model created or selected for a particular subset of patients does not provide high-quality predictions for other patient subgroups. In this case, a common and plausible procedure is to perform the model choice for each subgroup independently. But if the number of patients in a subgroup is comparatively small and if other subgroups contain patients with similar clinical or genetic properties, then this information should also be exploited. Our special interest lies in predicting survival for patient subpopulations that are defined either by clinical and demographic variables such as gender, age, tumour histology, tumour stage, or by genetic variables. In the context of survival analysis with high-dimensional genetic covariates, the models studied are mostly extensions of the popular Cox model, e.g., a penalised likelihood procedure such as ridge or lasso Cox regression.

For lung cancer patients, we have analysed the additional value of the information contained in other subcohorts with respect to improving the prediction quality regarding a specific subgroup. As reference, models created separately for each subgroup were used. Then, for a specific subgroup G, a weighted likelihood approach was considered. For every other subgroup, an individual weight  $w \in [0, 1]$  determines the strength with which the observations of this subgroup enter into the likelihood-based optimisation of the model parameters for subgroup G. A weight close to 0

indicates that a subgroup should be discarded, and a weight close to 1 indicates that basically the different data sets are pooled, neglecting subgroup membership. Within the MBO framework we additionally optimised - besides the model configuration - the individual values of the subgroup weights. Interestingly, this results in situations where a compromise (values for  $w$  in the interval  $(0, 1)$ ) leads to an improved predictability of the resulting models, compared to subgroup analysis ( $w = 0$ ) and pooled analysis ( $w = 1$ ), as published in a technical report in 2018.

In the presence of high-dimensional genetic covariates for medical prediction models, an important point is the stability of the feature selection. This is due to the desired interpretability of the resulting model. At the same time high predictive accuracy and a small number of selected features are required. In a comparison study we thus have used three criteria when fitting a predictive model to a medical high-dimensional data set: the classification accuracy, the stability of the feature selection, and the number of chosen features [A3/93]. We analysed Pareto fronts and concluded that it is possible to find models with a stable selection of few features without losing much predictive accuracy. Two important aspects are the selection of the measure for assessing the stability of the feature selection and the choice of a potential filter method for variable selection as a pre-step. Comparing a variety of stability measures, we found that a main factor is if a measure contains a correction term for large numbers of chosen features. Overall, we found that the Pearson correlation for indicator variables of feature inclusion has the best theoretical and empirical properties. In an ongoing large comparison study, we aim at the identification of the best filter methods for variable selection. Based on our discoveries regarding stability assessment and filter methods, we will use an adaptation of the MBO framework towards multi-criteria optimisation [73] to identify the most suitable machine learning methods for a given medical prediction task with high-dimensional covariates.

**State of the Art of Concept Drift** In the presented research the optimisation problem is assumed not to change over time. However, time-varying functions and therefore time-varying optimisation problems are common in the real world. This scenario is widely known as concept drift. These changes of concepts over time may have different forms. In figure 3.3, different kinds of drifts are sketched, where the optimum  $\theta^*$  of the objective function (the concept) changes over time. Drifts may happen *abruptly* by directly switching from one concept to another, or *incrementally* with many intermediate concepts in between. Also, the situation may occur that the concept returns to the original one after drifting, which is called *recurring*. Besides these concept drifts, noise can also exist in the real world. Noise does not reflect the drift but refers to a random deviation, which is an additional challenge for concept drift handling.

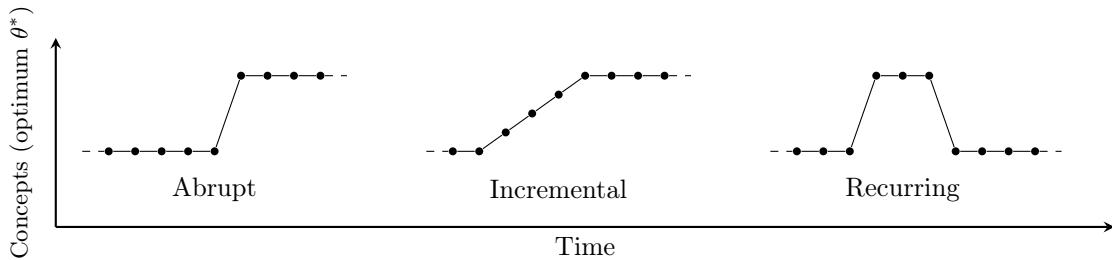


Figure 3.3: Patterns of different concept drifts, figure adapted from [70].

To keep high prediction accuracy in the case of a concept drift, the predictive models should have the ability to detect a drift and adapt to it. Thus, adaptive learning algorithms are applied in such a dynamically changing environment. Adaptive algorithms can update predictive models online during their operations. Furthermore, the costs of both time and computing resources should be taken into consideration while handling concept drifts. As concluded in [70], predictive models are required to:

1. detect and adapt concept drifts as soon as possible;
2. distinguish drift from noise, be sensitive to concept drift but stable to noise;
3. finish the operations (detect and adapt) within the sampling time (samples arrival time), and with cost-reasonable resources.

To fulfill these requirements, the following techniques are applied in most of the algorithms that handle concept drifts:

1. keep only a window of latest samples (the size of the window may be fixed or flexible);
2. include time as a new covariate in the predictive model (samples are associated with a weight that reflects their age);
3. maintain a set of models, reflecting different concepts (predictions are combined using voting, or weighted voting, or the most relevant model is selected).

**Related Work:** In most of the adaptive learning algorithms, in order to deal with potential concept drift, learning is based on most recent data, either a single sample or multiple samples. *WINNOW* [82] adopted the single sample learning strategy, where only one sample is learned at a time. When a new sample becomes available (as a set of features), a prediction is made. Once the true label (outcome) is available, then if the prediction is wrong, the model may be updated. Similarly, systems like *STAGGER* [83], *VFDT* [69], *DWM* [79] apply single-instance memory strategies to handle concept drifts.

Another strategy maintains a set of recent samples to update the predictive model. The popular *FLORA* series [90] includes a set of systems that considers multiple samples in the learning strategy. The original *FLORA* system maintains a window with a fixed length, which contains a set of most recent samples in FIFO order. In each iteration, the predictive model is retrained using the samples in that window. However, it is difficult to determine a suitable size for the window, as its choice balances a trade-off between stability (large window) and sensitivity (small window). Thus, to tolerate noise, methods with a flexible window size have been developed. The size of the window decreases when a drift is detected and increases when the concept is stable. For *recurring* drifts, the previous concept description is stored, to prevent recalculations. Competing approaches later also adopted the principle of a variable window size [71, 77, 67, 81].

Alternatively, an aging process could be applied to all samples for model training instead of a window-based approach. Each sample has a weight that reflects its age. Thus the newest sample has the highest weight and the oldest sample has the lowest weight. Some approaches use linear decay to determine the weights [80], whereas others use exponential decay [77].

Another way of dealing with concept drifts is ensemble learning. A set of models is maintained in memory, and their predictions are combined according to the current status of the concept. Adaptive ensembles are often motivated by the assumption that during a change, data is generated from a mixture distribution, which can be treated as a weighted combination of distributions characterising the target concepts; each individual model generates one distribution [85]. The weighted average of the individual predictions produces the final prediction, where the weight reflects the performance of the individual models on the most recent data. The weights change over time. Different kinds of ensemble learning algorithms have been proposed, including the SEA algorithm [88], Dynamic Weighted Majority algorithm (DWM) [79, 78], Diversity for Dealing with Drifts (DDD) [84], and Accuracy Updated Ensemble algorithm (AUE2) [68].

**Bibliography**

- [67] A. Bifet and R. Gavaldà. "Learning from time-changing data with adaptive windowing". In: *In SIAM International Conference on Data Mining*. 2007 (cit. on pp. 111, 228).
- [68] D. Brzezinski and J. Stefanowski. "Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm". In: *Neural Networks and Learning Systems, IEEE Transactions on* 25.1 (Jan. 2014), pp. 81–94 (cit. on p. 111).
- [69] P. Domingos and G. Hulten. "Mining High Speed Data Streams". In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*. New York, NY, USA: ACM, 2000, pp. 71–80 (cit. on p. 111).
- [70] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. "A Survey on Concept Drift Adaptation". In: *ACM Comput. Surv.* 46.4 (Mar. 2014), 44:1–44:37 (cit. on pp. 110, 228).
- [71] J. Gama, P. Medas, G. Castillo, and P. Rodrigues. "Learning with Drift Detection". In: *Advances in Artificial Intelligence - SBIA 2004*. Ed. by A. Bazzan and S. Labidi. Vol. 3171. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, pp. 286–295 (cit. on pp. 111, 228).
- [72] D. Ginsbourger, J. Janusevskis, and R. Le Riche. *Dealing with asynchronicity in parallel Gaussian Process based global optimization*. Tech. rep. 2011 (cit. on p. 109).
- [73] D. Horn and B. Bischl. "Multi-objective parameter configuration of machine learning algorithms using model-based optimization". In: *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. Dec. 2016, pp. 1–8 (cit. on p. 110).
- [74] F. Hutter, H. H. Hoos, and K. Leyton-Brown. "Parallel Algorithm Configuration". In: *Learning and Intelligent Optimization*. Ed. by Y. Hamadi and M. Schoenauer. Lecture Notes in Computer Science 7219. Springer Berlin Heidelberg, 2012, pp. 55–70 (cit. on pp. 108, 109).
- [75] F. Hutter, H. H. Hoos, and K. Leyton-Brown. "Sequential Model-Based Optimization for General Algorithm Configuration". In: *Learning and Intelligent Optimization*. Ed. by C. A. C. Coello. Lecture Notes in Computer Science 6683. Springer Berlin Heidelberg, 2011, pp. 507–523 (cit. on pp. 105, 114).
- [76] D. R. Jones, M. Schonlau, and W. J. Welch. "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13.4 (1998), 455–492 (cit. on pp. 105, 114).
- [77] R. Klinkenberg. "Learning Drifting Concepts: Example Selection vs. Example Weighting". In: *Intelligent Data Analysis (IDA), Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift* 8.3 (May 2004), pp. 281–300 (cit. on p. 111).
- [78] J. Z. Kolter and M. A. Maloof. "Dynamic weighted majority: An ensemble method for drifting concepts". In: *Journal of Machine Learning Research* 8.Dec (2007), pp. 2755–2790 (cit. on p. 111).
- [79] J. Z. Kolter and M. A. Maloof. "Dynamic Weighted Majority: A New Ensemble Method for Tracking Concept Drift". In: *Proceedings of the 3rd International IEEE Conference on Data Mining (ICDM-2003)*. Los Alamitos, CA, USA: IEEE Press, 2003, pp. 123–130 (cit. on p. 111).
- [80] I. Koychev. "Gradual forgetting for adaptation to concept drift". In: Proceedings of ECAI 2000 Workshop on Current Issues in Spatio-Temporal Reasoning. 2000, pp. 1–6 (cit. on p. 111).

- [81] L. I. Kuncheva and I. Žliobaitė. “On the window size for classification in changing environments”. In: *Intelligent Data Analysis* 13.6 (2009), pp. 861–872 (cit. on p. 111).
- [82] N. Littlestone. “Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm”. In: *Machine Learning* 2.4 (1988), pp. 285–318 (cit. on p. 111).
- [83] M. A. Maloof and R. S. Michalski. “A method for partial-memory incremental learning and its application to computer intrusion detection”. In: *Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on*. IEEE. 1995, pp. 392–397 (cit. on p. 111).
- [84] L. L. Minku and X. Yao. “DDD: A New Ensemble Approach for Dealing with Concept Drift”. In: *IEEE Transactions on Knowledge and Data Engineering* 24.4 (2012), pp. 619–633 (cit. on p. 111).
- [85] M. Scholz and R. Klinkenberg. *Boosting Classifiers for Drifting Concepts*. Tech. rep. 6/06. Dortmund, Germany: Collaborative Research Center on the Reduction of Complexity for Multivariate Data Structures (SFB 475), University of Dortmund, Jan. 2006 (cit. on pp. 111, 228).
- [86] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. “Taking the Human Out of the Loop: A Review of Bayesian Optimization”. In: *Proceedings of the IEEE* 104.1 (Jan. 2016), pp. 148–175 (cit. on pp. 105, 114).
- [87] J. Snoek, H. Larochelle, and R. P. Adams. “Practical Bayesian Optimization of Machine Learning Algorithms”. In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 2951–2959 (cit. on pp. 105, 114, 118, 120).
- [88] W. N. Street and Y. Kim. “A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification”. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*. 2001, pp. 377–382 (cit. on p. 111).
- [89] C. Thornton, F. Hutter, H. H. Hoos, and 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 (cit. on p. 106).
- [90] G. Widmer and M. Kubat. “Learning in the Presence of Concept Drift and Hidden Contexts”. In: *Machine Learning* 23.2 (1996), pp. 69–101 (cit. on p. 111).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [A3/91] B. Bischl, **M. Lang**, L. Kotthoff, J. Schiffner, **J. Richter**, E. Studerus, G. Casalicchio, and Z. M. Jones. “mlr: Machine Learning in R”. In: *Journal of Machine Learning Research* 17.170 (2016), pp. 1–5.
- [A3/92] I. Korb, **H. Kotthaus**, and **P. Marwedel**. “mmapcopy: Efficient Memory Footprint Reduction using Application-Knowledge”. In: *SAC 2016 31st ACM Symposium on Applied Computing*. 2016 (cit. on pp. 16, 107).
- [A3/93] **A. Bommert**, **J. Rahnenführer**, and **M. Lang**. “A multi-criteria approach to find predictive and sparse models with stable feature selection for high-dimensional data”. In: *Computational and Mathematical Methods in Medicine* 2017 (2017), pp. 1–18 (cit. on p. 110).

- [A3/94] **H. Kotthaus, J. Richter**, A. Lang, J. Thomas, B. Bischl, **P. Marwedel, J. Rahnenführer**, and **M. Lang**. “RAMBO: Resource-Aware Model-Based Optimization with Scheduling for Heterogeneous Runtimes and a Comparison with Asynchronous Model-Based Optimization”. In: *Proceedings of the 11th International Conference: Learning and Intelligent Optimization (LION 11)*. Lecture Notes in Computer Science. 2017, pp. 180–195 (cit. on pp. 108, 117).
- [A3/95] **J. Richter, H. Kotthaus**, B. Bischl, **P. Marwedel, J. Rahnenführer**, and **M. Lang**. “Faster Model-Based Optimization through Resource-Aware Scheduling Strategies”. In: *Proceedings of the 10th International Conference: Learning and Intelligent Optimization (LION 10)*. Vol. 10079. Lecture Notes in Computer Science (LNCS). Springer International Publishing, 2016, pp. 267–273 (cit. on pp. 16, 108).
- [A3/B2/96] **O. Neugebauer, P. Marwedel**, R. Kühn, and M. Engel. “Quality Evaluation Strategies for Approximate Computing in Embedded Systems”. In: *Technological Innovation for Smart Systems: 8th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2017*. Costa de Caparica, Portugal, 2017, pp. 203–210 (cit. on p. 107).
- [B2/97] **W.-H. Huang, J.-J. Chen**, and J. Reineke. “MIRROR: Symmetric Timing Analysis for Real-Time Tasks on Multicore Platforms with Shared Resources”. In: *Design Automation Conference (DAC)*. Austin, TX, USA: ACM, June 2016 (cit. on p. 107).
- [B2/21] **W.-H. Huang**, M. Yang, and **J.-J. Chen**. “Resource-Oriented Partitioned Scheduling in Multiprocessor Systems: How to Partition and How to Share?” In: *Real-Time Systems Symposium (RTSS)*. Porto, Portugal, Dec. 2016 (cit. on pp. 17, 70, 78, 107).

#### b) Other publications

- [A3/98] B. Bischl, **J. Richter**, J. Bossek, D. Horn, J. Thomas, and **M. Lang**. *mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions*. 2017. URL: <http://arxiv.org/abs/1703.03373>.
- [A3/99] **H. Kotthaus**. “Methods for Efficient Resource Utilization in Statistical Machine Learning Algorithms”. Diss. 2018.

### 3.4 Project plan

As described in Section 3.3.1, model-based optimisation (MBO) [76] is a state of the art [75, 87, 86] technique for global optimisation of computation-intensive black box functions. Such an optimisation black box can be applied to hyperparameter tuning to achieve the best possible prediction quality, e.g., for configuring the hyperparameters of a neural network or a TensorFlow application.

In the related recent research, including our own, it is typically assumed that the objective function does not change over time. For MBO, this means that the relationship between the input (algorithm and hyperparameters) and the target is constant over time. In the real world, however, often the relationship depends on the time, a scenario known as concept drift; see Section 3.3.1 (State of the Art of Concept Drift). A function that changes its optimum over time presents a new challenge for the MBO framework. A prominent example is the optimisation of running production processes. The quality of the production depends on controllable hyperparameters, which are usually subject

of the optimisation. Additionally, uncontrollable external parameters like temperature, humidity, etc. that influence the production quality. Such external parameters can be directly measurable or only latently measurable.

In online machine learning, the relationship between the features and the label changes over time, and the predictive model has to adapt to that change. On the one hand, online machine learning algorithms use their own methods to adapt to those changes. On the other hand, as with offline machine learning, hyperparameter tuning is promising to achieve the best possible predictive quality. Along with the change of the data, the optimal hyperparameters of the online machine learning algorithm also change.

## Goals

Overall, most of the research on concept drift refers to adaptations inside specific supervised online learning algorithms. To the best of our knowledge, taking concept drift into consideration while tuning hyperparameters has been a mostly neglected topic so far. Towards MBO with concept drift (CD), the project has the following goals in the third phase:

- We will investigate potential MBO-CD scenarios and build the fundamental cornerstone of the black box MBO methods under different concept drift scenarios (see **work package 1**).
- We will develop more efficient, flexible and adaptive RAMBO in **work package 2** and integrate MBO-CD and RAMBO as RAMBO-CD that can efficiently execute MBO-CD with resource awareness in **work package 3**.
- We will design methodologies to deal with insufficient samples (see **work package 4**), and the concept drifts due to increasing sample sizes of data streams (see **work package 5**).
- We will improve RAMBO-CD with active-learning strategies, interactively selecting and labelling samples, by considering the concept drift with respect to the quality of the machine learning methods (see **work package 6**).

## Work schedule

### Work package 1. MBO with the concept drift (MBO-CD)

**Problem:** Model-based optimisation (MBO) is a state of the art method for expensive black box problems. Here, the objective function is assumed to be constant over time. Let  $f(\theta): \Theta \rightarrow \mathbb{R}$  be an expensive black box function with a  $d$ -dimensional input domain  $\Theta \subset \mathbb{R}^d$ . The goal is to find  $\theta^* = \operatorname{argmin}_{\theta \in \Theta} f(\theta)$ . MBO initialises with  $k$  evaluations and iteratively repeats the following steps until the budget is exhausted: A Gaussian process regression is fitted to all past evaluations, serving as a surrogate to estimate  $f$  globally, the *Expected Improvement* (EI) is optimised to determine the most promising point  $\hat{\theta}$ , and  $\nu = f(\hat{\theta})$  is evaluated. The final optimisation result  $\hat{\theta}^*$  is the input that led to the minimal observed objective value.

In many domains time-varying functions are found and represent a new challenge for this framework. For example in cooperation with project A4 we plan to optimise the prediction of channel quality and capacity in 5G networks under heterogeneous and changing environments. We aim for a general approach where any underlying optimisation problem can be recalibrated by a concept-drift-aware optimiser. So now we assume that  $f(\theta)$  and  $\theta^*$  change over time  $t$ , denoted as  $f_t(\theta)$ , so that we are searching for

$$\theta_t^* = \operatorname{argmin}_{\theta \in \Theta} f_t(\theta).$$

**Proposed Solutions and Challenges:** We will investigate the usefulness of different approaches for addressing this situation. Most approaches are extensions of the MBO framework and they are inspired by the general ideas for dealing with concept drifts as described in the paragraph on related work above.

*Naïve approach:* When a concept drift is detected: Discard all previous evaluations, and start over with fitting the surrogate model.

*Historical approach:* When a concept drift is detected: Select the evaluated parameter settings that were best-performing before the concept drift, fit the surrogate function for these settings, and then continue as usual.

*Fixed Window approach:* In each iteration, consider only evaluations observed in the last  $t_\Delta$  time units.

*Weighted window approach:* In each iteration, consider only evaluations observed in the last  $t_\Delta$  time units, where newer evaluations are given higher weights in the surrogate fitting process than older evaluations.

*Flexible window approach:* Use the window approach, but with  $t_\Delta$  variable, depending on the estimated strength of the concept drift.

*Time-as-covariate:* Include time as an additional dimension  $t$  in the surrogate model. The infill criterion (e.g., EI) is optimised with  $t$  fixed to the current time.

*Window and time-as-covariate:* Combine the window approach and the time-as-covariate approach.

A potential advantage of the time-as-covariate approach is that interactions between  $\theta$  and  $t$  can be identified and the concept drift can be modelled by the surrogate function.

**Simplified Scenarios and Preliminary Solutions:** As proof of concept, we investigated our optimisation problem for the situations of an abrupt and of an incremental concept drift; see figure 3.3. Optimisation results must be provided in an online fashion in which the actual state of  $f_t$  is considered. Therefore,  $\hat{\theta}_t^*$  is obtained by selecting the input value from the set of evaluations where the surrogate predicts the smallest outcome. Accordingly, the performance is assessed by comparing  $f(\hat{\theta}_t^*)$  with the optimum  $f(\theta_t^*)$  at time  $t$ . We use the mean absolute error (MAE) between these two points, averaging over all times  $t$ . Figure 3.4 shows that, on a set of six exemplary functions, these approaches already represent an improvement over the normal MBO if a concept drift occurs, while at the same time they do not lead to a significant loss of performance if no drift occurs.

In this benchmark, the exemplary functions are designed such that the concept drift and its characteristics are fully known. In such a controlled scenario the best method can be objectively evaluated. Following this approach, we will use established benchmark functions and treat one dimension as the time as well as specifically designed functions that emulate abrupt and incremental concept drifts. In a second step we will confirm our findings from benchmark studies on real-life data from our cooperations.

## Work package 2. RAMBO with flexible scheduling

**Problem:** In order to speed up the computation in model-based optimisation, it is often indispensable to apply parallelisation and/or acceleration. Resource management is required to increase the effectiveness of parallel optimisation in MBO. Thus one essential research goal for achieving resource-aware MBO is to use the available resources adaptively and efficiently. In addition to the computation-bounded resource management, we will also carefully optimise for communication and memory accesses between different MBO jobs that may apply distributed computing, multi-core platforms, GPU accelerators, or designated machine learning hardware accelerators.

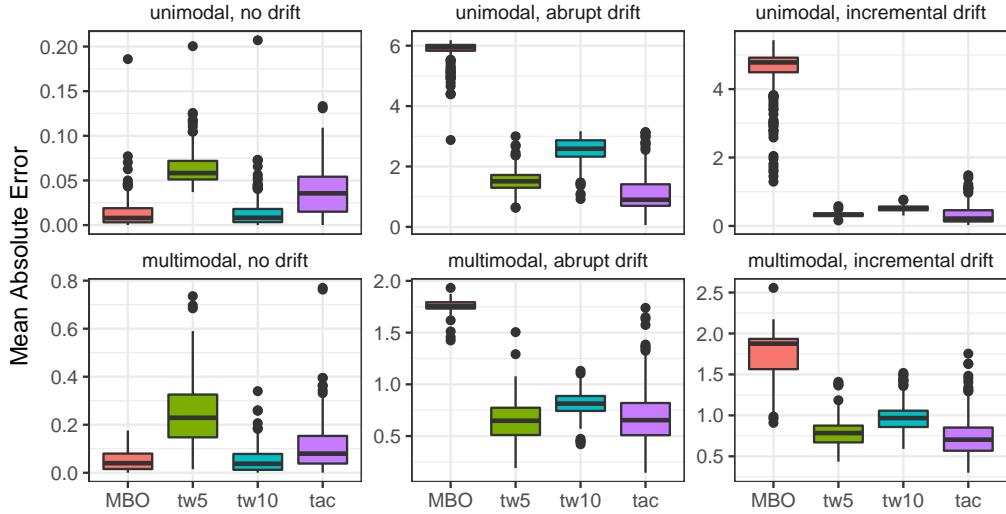


Figure 3.4: Mean absolute error over the whole period for MBO without consideration of concept drift (MBO), with the time-window approach ( $\text{tw}$  with  $t_\Delta = 5, 10$ ) and with the time-as-covariate approach ( $\text{tac}$ ). Unimodal stands for a function with only one local minimum and multimodal stands for a function with multiple local minima.

**Proposed Solutions and Challenges:** As part of the research results in the second phase of the project, we have developed the *resource-aware model-based optimization* (RAMBO), e.g., in [A3/94], to execute R programs with scheduling for heterogeneous runtimes. The backbone of RAMBO is a knapsack-based scheduling approach, developed for the purpose of increasing the utilisation of the computing resource. All similar points are clustered, and, for each iteration, only one point within a cluster is evaluated, so penalties incurred by evaluating similar points can be avoided. To decrease the CPU idle time, all CPUs are treated as knapsacks, where the length of the iteration is the volume of those knapsacks. Evaluations of selected points are treated as jobs, and each job has its *expected improvement* (profit) and *expected execution time* (weight). Then, the points which will be evaluated in the next iteration are selected by solving an instance of the knapsack problem.

Although such an approach improves the execution efficiency of MBO by utilising the available CPUs in parallel, there is still a large space for improvement. First, jobs in MBO are designed to be executed only on one CPU and different jobs can be executed on different CPUs, due to the current limitations in the R packages, and therefore CPUs are treated as knapsacks separately. If we can allow tasks to be migrated between different CPUs, then all CPUs can be combined to one large knapsack, so that the resource utilisation can be further improved. Second, to avoid the penalties from selecting similar configuration points to be evaluated within one iteration, the current practice is to cluster points that are *similar*. By assigning low loss values to points in different clusters, points from different clusters are preferred. However, this method is heuristic, which cannot guarantee the performance all the time. Moreover, the exploration in the second phase was limited to CPUs. GPU accelerators and designated machine learning hardware accelerators are also important resources that should be also utilised carefully in the MBO framework.

**Simplified Scenarios and Preliminary Solutions:** In this work package, a clear relationship between the distance of two points and the cancellation (CEL) of *Expected Improvement* (EI) will be explored, so that the total EI can be calculated if two points are evaluated within one iteration, e.g.,  $EI\{\theta_1, \theta_2\} = EI(\theta_1) + EI(\theta_2) - CEL$  (w.r.t.  $dis(\theta_1, \theta_2)$ ). To maximise the total EI, one specific approach is to deal with the above relationship by mapping the problem to the well-known quadratic knapsack problem (QKP). Towards this, we will analyse the trade-off between overhead

and performance of solving such a QKP. Apart from this, the job migration mechanism between CPUs in R programs will be established. With the migration mechanism, the jobs are allowed to be executed in different CPUs during the optimisation process, which can guarantee an efficient resource utilisation.

An example is shown in figure 3.5. In this case, the length of iteration is 3 time units, and the execution time of each task is 2 time units. If there is no migration mechanism, only two tasks can be executed within one iteration, and one time unit is wasted for each CPU. In contrast, if task migration is possible, then three tasks can be finished. Moreover, with an increasing number of CPUs integrated in one system, a process migration mechanism gives developers more freedom when implementing algorithms in R with different purposes (e.g., balance the workload in heterogeneous distributed systems).

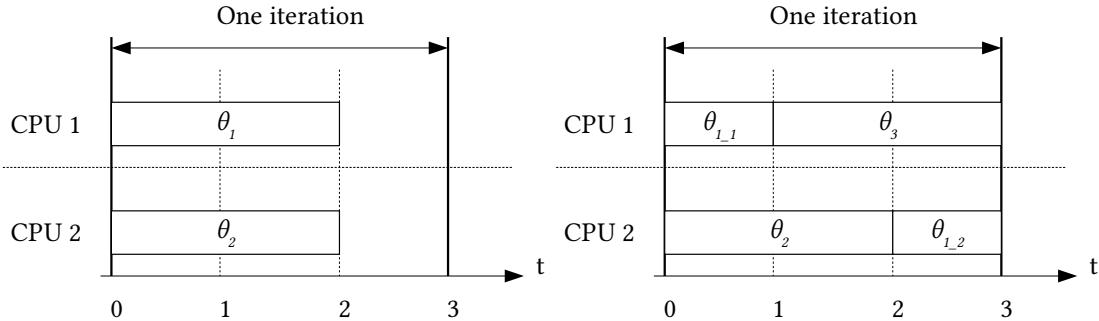


Figure 3.5: *Left:* Scheduling without migration mechanism *Right:* Scheduling with migration mechanism.

### Work package 3. RAMBO with concept drift (RAMBO-CD)

**Problem:** In case of a concept drift, the resource management policy can also change over time. Therefore, adaptive and flexible resource management is needed to react according to the concept drift. In this work package, we will investigate RAMBO-CD that combines MBO-CD (work package 1) and RAMBO (work package 2). Due to concept drift, the metric *Expected Improvement* (EI) in RAMBO becomes dynamic over time. Moreover, if a concept drift takes place at time  $t$ , one major resource management issue is to decide whether an unfinished evaluation of a point should be aborted or continued so that the resource can be utilised to evaluate more important points in the new concept.

**Proposed Solutions and Challenges:** The RAMBO approaches developed in the second phase and work package 2 will be extended to deal with time-dependent Expected Improvement (EI). An extension to the EI per second [87] is possible to focus on time-saving configurations. Updating the time-dependent EI (of all the points) can be costly if it is not carefully handled. Therefore, the updates will only be triggered when a certain threshold is exceeded. The approaches that are developed in work package 1 will be explored here to approximate the strength of the concept drift to reduce the complexity for updating the EI and the corresponding estimations. Therefore, the knapsack problem and its variants that have been used in RAMBO have to be configured to be time-dependent as well to react according to the concept drift.

When a concept drift takes place at time  $t$ , the previously selected points that are still under evaluation at time  $t$  may remain useful or become useless. Since the concept drift may be detected earlier/later than the actual concept drift, the resource management has to be done carefully to avoid aborting evaluations of the points consecutively. Continuing an unfinished evaluation at time  $t$  is a good option in case 1) the EI remains good and the evaluation is going to be finished

very soon anyway or 2) the concept drift detection is false. Aborting an unfinished evaluation at time  $t$  is a good option if the EI becomes bad or the evaluation is still going to take a long while. Moreover, for some evaluation methods, it is possible to interrupt an evaluation in which partial results are returned, and the evaluation can be continued later on if necessary. We will explore all the above options and perform fundamental explorations from both theoretical and practical perspectives.

#### **Work package 4. RAMBO with augmented samples in case of small sample size**

**Problem:** It is well known that small sample sizes can compromise the prediction quality of a model. One possible consequence is that findings cannot be extrapolated or generalised. We will investigate methods for combining real data with additional simulated data. Clearly, the underlying distribution of the real and the simulated data should be comparable. The difference between these distributions will determine the effectiveness and the validity of this enhancement. A specific application scenario is traffic analysis and prediction (project B4), e.g., travelling times, estimated car density, and traffic flow. Here, both real data and simulated data are available. An established simulation environment can be used to generate additional samples and evaluate the impact of the increased sample size.

The overall goal in this project is to improve the prediction accuracy for a small sample size. Reasons for small sample sizes can be that it is impossible to obtain more labeled samples or new samples are obtained only at an unreasonable cost. In some situations, well-established simulators can be applied to augment the sample data by applying simulations that try to imitate the relationship between the features and the labels. If those simulations are cheap and reflect the real relationship they could be used to improve the predictions of true labels.

**Proposed Solutions and Challenges:** When high accuracy is required, simulations can be very slow or alternatively not very accurate if not every detail is simulated. Therefore, simulations can be very expensive in the evaluation, and it is not proven if they indeed reflect the reality. Typically the simulation has parameters that steer its behaviour and therefore determine how realistically the simulator works. In this work package, we plan to enrich the real data with simulated data, which is then used to train the machine learning method. The prediction quality of the machine learning method, however, will only be evaluated on the real data, in order to avoid skewing the results towards the simulated data. Then, RAMBO will be extended to optimise not only the hyperparameters of the machine learning method, but at the same time the parameters of the simulated data. These parameters include both the number of simulations and the parameters of the simulator itself. As a result we expect that RAMBO will find the settings for the simulator such that it can generate samples that improve the accuracy of the machine learning method.

#### **Work package 5. RAMBO for data streams**

**Problem:** This work package will explore MBO for machine learning problems with variability due to increasing sample sizes. The problem is twofold. In the beginning the number of samples might be critically small and only increases over time. Here a minimal number of required samples to effectively start MBO for the machine learning method selection and configuration has to be determined. In the beginning the data set will be small and evaluations will be cheap, making it feasible to do an exhaustive search instead of applying MBO. In the beginning relatively simple methods will yield the best prediction performance. With increasing sample size, more sophisticated machine learning methods will outperform the initially best ones. This is the first form of concept drift.

During the process an effective number of samples has to be identified that make re-training necessary. In a typical scenario a facility uses a machine learning model to obtain predictions during the operation time. At fixed intervals (time difference or number of samples), the true labels are obtained and the model is updated. These updates are necessary to account for changes in the data regarding the true relationship between the variables and the labels. This is the second form of concept drift.

**Proposed Solutions and Challenges:** For the first form of concept drift, related to the starting problem, we will combine an exhaustive search in the beginning that benefits from the cheap evaluation costs. The surrogate in MBO will benefit from these early evaluations.

For the second form of concept drift, we will use the surrogate model with the time as a covariate to estimate the intensity of the drift. This enables us to prevent unnecessary model updates if no drift is present and to set the update interval according to the drift speed. Furthermore we will optimise the hyperparameters of the machine learning method at each scheduled update to ensure the highest possible prediction quality. This should only happen if a concept drift makes it necessary; otherwise resources would be wasted.

To determine whether a reoptimisation of the hyperparameters will be purposeful, we will use the EI. We plan to extend the EI per second [87], discussed in work package 3. If a threshold is exceeded a reoptimisation will be triggered. Including the time as a covariate in the surrogate will lead to an increasing uncertainty in the predictions over time. This leads to an increased EI value, which will eventually exceed the threshold. If the performance varies heavily over time, the uncertainty will also be higher, leading to an increased rate of reoptimisations and adaptations to the drift.

A challenging question is how representative the optimal algorithm configuration obtained at time  $t_i$  for the data at future time points is, since the performance can only be evaluated for data in the past. Therefore, if we optimise the performance of the machine learning method, we would optimise its hyperparameters for a past concept. To overcome this problem, we will put a higher weight on new observations. Additionally we will investigate possibilities of using the surrogate model to directly predict the best hyperparameters for future concepts. Further, we have to take into account, that the optimisation run itself only returns an optimisation result after considerable runtime.

## Work package 6. RAMBO-CD with interactive selection of samples

**Problem:** In this work package we will study how we can use RAMBO-CD to augment *Active Learning* with an automatic algorithm configuration. The machine learning problem consists of a data set where the true label is known only for a subset of samples. Determining the true label of a sample is connected with a significant cost. The goal is that the machine learning method can label all samples with the highest possible accuracy. Naturally, there will be samples where the machine learning method is pretty confident in its prediction (e.g., they are located in a neighbourhood where all samples belong to the same class) and samples that are only uncertainly labeled (e.g., because they are between two neighbourhoods). In active learning, to improve the overall accuracy, a proper budget is invested into labelling those samples that are in uncertainty regions.

In the context of RAMBO-CD with active learning, the RAMBO-CD algorithm decides how to invest the budget, i.e., which machine learning methods to run.

**Proposed Solutions and Challenges:** Assuming that the relationship between the features and the label of the data does not change, the increasing sample size, which is a result of the active learning strategy, might introduce a form of concept drift nonetheless, due to the changed data set.

We assume that the active learning strategy is not fixed but can be applied to a set of machine learning methods with hyperparameters. The goal is to find the machine learning method with the highest prediction quality. The quality is estimated by the surrogate model.

The simplest approach is to evaluate the active learning proposal only of the momentarily best machine learning method and then to use RAMBO-CD for selecting the best method after a sample has been labelled. This approach could lead to favouring only one specific machine learning method, for which iteratively more and more points are labelled.

However, in each step RAMBO-CD provides the assessment of the prediction quality of multiple methods. Thus, in an extension to the aforementioned simple approach, methods with higher prediction quality can be selected with higher probability as a basis for the active learning strategy to select the next sample.

### Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. MBO with the concept drift (MBO-CD)																	
2. RAMBO with flexible scheduling																	
3. RAMBO with concept drift (RAMBO-CD)																	
4. RAMBO with augmented samples in case of small sample size																	
5. RAMBO for data streams																	
6. RAMBO-CD with interactive selection of samples																	

### 3.5 Role within the Collaborative Research Centre

The methods developed in this project are generic methods that can be applied to solve various optimisation problems that fit the black box paradigm.

All our methods are implemented in self-contained and well-documented software packages, assuring the reproducibility of the experiments and future usability for other researchers. Integration of new methods in up-to-date software packages is very important.

In cooperation with project B2, MBO has been used for optimising the analysis pipeline for virus detection with the PAMONO sensor, with optimisation regarding energy consumption and runtime. This collaboration with B2 will continue in the third phase with a focus on tuning deep neural networks for classification. The high evaluation costs of neural networks make an efficient resource-aware tuning necessary.

## Project A3

In project A4 manually tuned neural networks are used to predict the data rate of mobile devices in cellular networks. Applying automated machine learning method selection and using MBO for tuning can further improve the predictive performance of the models.

We plan to apply RAMBO-CD to steer the hyperparameter optimisation of the machine learning methods that use different hardware components for acceleration in project A1.

Due to changing concepts of the machine learning problems that occur in industrial production processes after deployment, the topic of concept drift has to be taken into account. Therefore, we plan to integrate RAMBO-CD into the prediction-based quality gate design for process chains in project B3. Here, we plan to continuously optimise the hyperparameters of the applied machine learning methods that are used to predict the final product quality.

In project A6 graph synthesis with deep generative architectures also suffers from the small sample size problem; the method for combining real data with additional simulated data can be applied in that situation as well.

In cooperation with project B4 we plan to apply MBO-CD to guide the selection of the best method for prediction of travel times. Concept drifts appear due to changes in the environment, e.g., regarding weather conditions, time of day, or accidents.

## 3.6 Differentiation from other funded projects

### **Suspension-Aware Designs and Analyses for Real-Time Embedded Systems (Sus-Aware) (Chen, Reference number DFG CH 985/12-1) (Funding period: 2018–2021)**

This project focuses on robust and solid fundamental algorithms and analyses to carefully mitigate and analyse the impact of self-suspending behaviour in modern real-time embedded system. Machine learning does not play any role in this project.

### **Design and Optimization of Non-Volatile One-Memory Architecture (NVM-OMA) (Chen, Reference number DFG CH 985/13-1) (Funding period: N.N.–N.N.)**

This project aims to enable the effectiveness of one-memory architectures, in which a NVM is used both for the storage and main memory. We plan to perform design-space exploration in hardware and software designs and integrate analytical and optimised resource management in operating systems. The proposal was submitted in December 2017 and is under review. Machine learning does not play any role in this project.

### **Partitioning, Scheduling, Spinning, Suspension, Synchronization, and Locking Protocols in Real-Time Systems (PS4Lock)**

#### **(Chen, Reference number DFG CH 985/14-1) (Funding period: N.N.–N.N.)**

The project will design practical and solid fundamental algorithms and analyses to handle shared resources based on lock mechanisms in multiprocessor embedded systems. Our project intends to find the break-even and dominating scenarios for task partitioning, task spinning, task suspension, and resource synchronisation. The proposal was submitted in Feb. 2018 and is under review. Machine learning does not play any role in this project.

### **E: Top Translationsphase: LivSys**

#### **(Rahnenführer, Reference number BMBF 031L0119B) (Funding period: 2016–2019)**

In this project gene expression measurements are the basis for model stress responses of cultivated human hepatocytes to hepatotoxic compounds with genome-wide expression measurements. Most models for dose-response relationships in this project have only few parameters to be estimated; hyperparameter tuning as in project A3 is not an issue.

**E:Top Translationsphase SysDT**

**(Rahnenführer, Reference number BMBF 031L0117E) (Funding period: 2016–2019)**

The goal of this project is an in vitro system that replaces animal experiments for developmental toxicity. Statistical methods for the determination of compound-specific lowest concentrations with positive results and for the prediction of toxicity-indicating events are developed. This involves model selection for a classification task, but without resource efficiency.

**StemNet**

**(Rahnenführer, Reference number BMBF 01EK1604C) (Funding period: 2017–2019)**

The project aims at improving currently available differentiation protocols for HiPSC (human induced pluripotent stem cells)-derived HPCs (hepatocyte like cells). A genome-wide network analysis is used to rationally develop interventions. Cluster analyses and molecular enrichment analyses are performed. Model selection and hyperparameter tuning are not in the focus.

**Identification of survival models that are prognostic across cohorts and stable regarding variable selection with methods of model-based optimization**

**(Rahnenführer, Reference number RA 870/7-1) (Funding period: 2016–2019)**

The goal of this project is to develop new statistical methods tailored to survival analysis with genetic high dimensional data sets. Specific tasks are the identification of models with high prediction quality on independent cohorts and with stable variable selection. This project benefits from project A3, since MBO is used for selecting algorithms, but no method development for MBO is performed in this project, and resource efficiency is not considered.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	2	129,000	2	129,000	2	129,000	2	129,000
Total	—	129,000	—	129,000	—	129,000	—	129,000
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	129,000		129,000		129,000		129,000	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Jian-Jia Chen, Prof. Dr., professor	Embedded systems	TU Dortmund	8	—	Existing funds
	2	Jörg Rahnenführer, Prof. Dr., professor	Statistical methods in genetics and chemometrics	TU Dortmund	4	—	Existing funds
	3	Andrea Bommert, M.Sc., doctoral researcher	Statistical methods in genetics and chemometrics	TU Dortmund	19.92	—	Existing funds
	4	Lea Schönberger, M.Sc., doctoral researcher	Embedded systems	TU Dortmund	19.92	—	Existing funds
	5	N.N., student assistant	Embedded systems	TU Dortmund	8	—	Existing funds
Non-research staff	6	Eva Brune, secretary	—	TU Dortmund	1	—	Existing funds
	7	Claudia Graute, secretary	—	TU Dortmund	1	—	Existing funds
<b>Requested staff</b>							
Research staff	8	Jakob Richter, M.Sc., doctoral researcher	Statistical methods in genetics and chemometrics	TU Dortmund	—	Doctoral researcher	—
	9	Junjie Shi, M.Sc., doctoral researcher	Embedded systems	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

- 1. Chen, Jian-Jia**  
Project management. Focus on scheduling and resource management. Cooperation in all WPs.
- 2. Rahnenführer, Jörg**  
Project management. Focus on statistical learning. Cooperation in all WPs.
- 3. Bommert, Andrea**  
Research and development with a focus on statistical learning.
- 4. Schönberger, Lea**  
Research and development with a focus on scheduling.
- 5. N.N.**  
Implementation and assistance in evaluation of algorithms.
- 6. Brune, Eva**  
Secretary.
- 7. Graute, Claudia**  
Secretary.

**Job descriptions of staff for the proposed funding period (requested funds):**

- 8. Richter, Jakob**  
Research and development in all WPs with a focus on statistical learning. The researcher will develop the methodological extensions of MBO with respect to concept drift (WP1). In addition, he will develop statistical methods for small sample sizes, data streams and interactive selection of samples (WP3-6).
- 9. Shi, Junjie**  
Research and development in all WPs with a focus on scheduling and resource management. The researcher will develop the flexible scheduling algorithms to improve the resource management of MBO (WP2). In addition, he will work on the scheduling aspects for extending MBO towards small sample sizes, data streams and interactive selection of samples (WP3-6).

### **3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from university	5,000	5,000	5,000	5,000
Sum of existing funds	5,000	5,000	5,000	5,000
Sum of requested funds	0	0	0	0

(All figures in euros)

### **3.7.5 Requested funding for instrumentation for the new funding period**

This project does not request any funding for major research instrumentation.

### 3.1 General information about Project A4

### 3.1.1 Project title:

## Resource efficient and distributed platforms for integrative data analysis

### 3.1.2 Research area(s):

408-02 (Communications, High-Frequency and Network Technology), 407-04 (Logistics)

### 3.1.3 Principal investigator(s)

ten Hompel, Michael, Prof. Dr., Dr.h.c., 19.01.1958, German

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

Phone: 0231-755-2793

E-mail: michael.tenhompel@tu-dortmund.de

Wietfeld, Christian, Prof. Dr., 22.01.1966, German

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

Phone: 0231-755-4515

E-mail: christian.wietfeld@tu-dortmund.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes	(x) no
2.	clinical trials	( ) yes	(x) no
3.	experiments involving vertebrates.	( ) yes	(x) no
4.	experiments involving recombinant DNA.	( ) yes	(x) no
5.	research involving human embryonic stem cells.	( ) yes	(x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes	(x) no

## 3.2 Summary

The long-term vision of this project is to provide methods that enable the development of resource-aware applications for data analysis in potentially distributed cyber-physical systems (CPSs). For that purpose, the project systematically captures the resource demands of distributed and compound platforms to develop resource models that provide this information during design or runtime to the developer, or to the respective data analysis application of the system software. In addition, the project's achievements serve other CRC projects in the following two ways: First, application of state of the art machine learning approaches for synthesis of such models and second, the provision of these models for efficient data analysis on resource constrained devices.

The third project phase shifts the focus of the research from homogeneous system platforms to the topical subject of large-scale, heterogeneous and energy-constrained internet of things (IoT)-networks. Important and new research topics address a resource-optimised segmentation of network-wide data analysis tasks according to available energy resources and individual platform features, as well as a guaranteed power supply by parallel energy harvesting in dynamically changing environments, e.g., distributed logistic systems. Our achievements of indoor photovoltaic harvesting will be extended to different harvesting principles. In addition, hardware and methods for optimised multi-source energy scavenging on a single device have to be developed and modelled. The research will focus on methods that support data analysis application developers during design and development of such systems with mandatory model-knowledge, algorithms for segmentation of tasks, and the corresponding platform services. The project will also face the emerging challenge of overcoming the capacity limits of the involved massive machine type communication (mMTC) in ultra dense environments. For this purpose, we will extend and validate methods of the newest 4G/5G technologies, e.g., NB-IoT and mmWave pencil beams, by our novel methods for data-analysis-based resource prediction and positioning. It will be combined with the elaboration of multimodal indoor positioning and localisation systems together with methodologies for reference matching and data aggregation in graph based representations of network entities. Simulations and predictions of resource-demands in distributed scenarios should reveal the validity and feasibility of the considered approaches and solutions.

The process of modelling and verification in the entire project is supported by the continuous transfer of knowledge into the industrial environment of the logistics testbed *PhyNetLab*, which was developed and evaluated in the previous phase of the project. Although Olaf Spinczyk is leaving the project to pursue a W3 professorship at the University of Osnabrück, the joint outcome of phases 1 and 2 will be fully available to the project for phase 3 to ensure a seamless further evolution of the project.

## 3.3 Project progress to date

### 3.3.1 Report and current state of research

The design process of distributed energy-constrained systems demands precise models that enable online and offline estimation of the given context, environment and resources. This comprises models for energy generation, e.g., energy harvesting, radio resource estimation and prediction, localisation, and methods for automated energy model synthesis for hardware components by the system software. The following section describes the contributions of the second project phase and shows how they are tightly related to the current state of research in this area.

**Scenario Definition and Testbed (WP1, WP10):** Starting with concepts and requirements for close-to-reality validation, phase 2 included the design and development of the logistics testbed

*PhyNetLab* [A4/128] as an IoT-reference implementation for industry 4.0 use cases with intelligent containers. These containers with features for interaction, data acquisition and other actuation capabilities are used to model innovative processes in logistics or materials handling systems. The experimental hardware entity of the PhyNetLab is the *PhyNode* (cf. Fig. 3.1), which is controlled by an ultra-low-power microcontroller unit (MCU) and includes sensors to provide environmental data such as light level, temperature, and acceleration. It is equipped with a low-power 868 MHz radio transceiver for communication. Due to the typical scale of thousands of containers in such facilities, the nodes were designed as energy-neutral platforms.



Figure 3.1: Our *PhyNode* hardware in the CRC logistics testbed *PhyNetLab* [A4/128].

Such a testbed combines an actual application context with non-functional requirements, such as energy-efficiency, efficient communication and operating system support, into one platform. Other requirements that are unique to this testbed are to provide fast testing time using over-the-air updates for all nodes. This enables rapid iterations in the large-scale system for executing business processes, simulating large-scale networks, and also to experimenting with various versions of wireless communication protocols with respect to energy-efficient strategies. The nodes also have the capability to choose between an external rechargeable battery, supercapacitors, or thin-film batteries. Multiple integrated sensors not only allow us to perform application-based experiments, but they also provide an analytical basis for indoor energy harvesting using photovoltaics. Furthermore, the *PhyNetLab* mimics industrial processes with interacting elements in real-time which include the dynamics of wireless communication and indoor energy harvesting.

In contrast, the TidMarsh project from ResEnv, MIT, Cambridge [115], is a virtual observatory that deploys outdoor sensors. A comparison can be drawn to understand the nature of these two deployments, where *PhyNetLab* deploys 150 nodes in a  $500\text{ m}^2$  indoor hall with other logistics equipment whereas TidMarsh deploys 150 sensors in a  $2.4\text{ km}^2$  property of wetlands. Since the deployment area is large and the duty cycle of the communication together with energy requirements is considerably less compared to communicating industry 4.0 scenarios, this system has no demands for energy-neutral operation and large-scale communication. Therefore, the industrial scenario differentiates itself from the TidMarsh project by modelling the resource constraint nature of the nodes. Furthermore, *PhyNetLab* is a multi-layered architecture, which is interconnected by hierarchical communication between the energy-neutral wireless sensor network (WSN) of *PhyNodes* at its bottom, distributed multi-radio access points (APs) in the middle, and a connection to the Internet on the top.

**Context-Aware and Resource-Efficient Communication (WP2–WP4):** Particular attention has to be paid to the wireless communication system for IoT networks, since it is indispensable for functionality, while requiring a substantial amount of scarce resources. Versatile emerging communication technologies challenge system designers and developers to make the best decision with respect to costs, power consumption, data rate, latency, and scalability. In addition, these optimisation objectives depend not only on the technology itself but also on its configuration and dynamic effects, which are a consequence of the network density, activity, and type of radio-resource management (e.g., centrally organised like cellular networks or distributed like WiFi and Bluetooth). To perform both, offline decisions for systems design and qualified online decisions based on the momentary context of particular platforms, detailed models of communication systems and their influencing factors are required. A prominent outcome throughout both phases of this CRC is the *context-aware power consumption model (CoPoMo)*, which provides fine-grained, highly accurate and flexible power models for cellular mobile equipment. Although in phase 1 the model covered the essential context parameters, such as transmission size, arrival rate, cell environment and mobility, this phase addressed the inclusion of an additional – increasingly important – context parameter: the *degree of competition* or congestion of the shared radio medium, which is issued by other network participants. This required numerous preparatory works, which were essential for the generation and validation of the novel communication models and extensions, e.g., distributed smart traffic generators for controlled field evaluations and extensions of open-source software-defined radio (SDR) implementations, such as OpenAirInterface [120] and srsLTE [108], which give a deep insight through all layers of the protocol stack.

Recent advances in cellular network technologies, in terms of capacity and data rate, open up new opportunities for saving energy resources. The aggregation of spectral resources in long term evolution (LTE)-advanced (LTE-A) is one of those approaches. However, powering multiple transceiver chains, one for each component carrier, at first increases the power consumption of the devices. To quantify this, we extended *CoPoMo* to cover the power consumption of such approaches. By doing this, we provided a method to determine situations, where the initial investment of additional power resources results in an overall reduction of power consumption by quick transmissions and larger dwell time in low power modes [A4/135]. Based on characteristics of real user equipment, several of our case studies showed that the relative energy savings grow larger with both, an increasing amount of transmitted data and the data-rate improvement represented by the *carrier aggregation boost factor  $a_{CA}$*  [A4/135]. The positive effect on the power consumption further increases for higher frequencies since such transmissions generally require a higher transmission power to compensate the increased propagation loss. This mainly influences the minimum required boost factor to cross the break-over point from which the investment of powering additional transceiver chains becomes beneficial compared to a single-carrier transmission. Our extension was recently used by other researchers to evaluate the energy-efficiency of smartphones in public networks based on real trace data [117]. In a joint collaboration with projects A1 and B4 we leveraged machine learning and real data traces of public LTE networks to derive prediction models for the transmission power of mobile handsets based on passive connectivity indicators. Subsequently, the approach was used to elaborate the positive impact of channel-predictive transmissions (project B4) on the energy consumption. In addition, *CoPoMo* has been integrated by project B2 to perform energy considerations for offloading decisions of plasmon assisted microscopy of nano-sized objects (PAMONO) sensor data via LTE networks [B2/124].

Sticking to cellular wide area networks, anticipatory networking has gained attention recently [122, 103]. It is motivated by the increasing demands on latency and throughput with the parallel rise of network load up to spectral capacity limits, which is caused by an increasing number of mobile stations and emerging tendencies for large data transfers, e.g., multimedia streaming. While the main target is optimising the network performance based on predictions, this prevalent challenge is being tackled by numerous researchers on a wide range of system levels and with different approaches from computer science. Generally, these approaches can be classified by

the three attributes context, prediction, and optimisation [103]. Context describes the source of knowledge the prediction is based on, which might be location information, traffic types, or social influences. The prediction itself may be based on regression, classification, or time series, while the subsequent optimisation techniques include, e.g., convex optimisation, game theory, and heuristics [103, 126]. Most of these approaches are applied on the infrastructure side of the networks, with optional support of client data. On this topic, our contribution is *BaLAnce* (Battery Lifetime-Aware LTE Switching-Off Strategy in Green Network Infrastructures), which combines our estimation model CoPoMo to a method for trade-off analysis of power consumption between battery-powered handsets and power-intense base station infrastructure. Hence, we contribute to the hot topic of collaborative [106], learning-based [111] and dynamic [112] optimisation for green networks [107], which play a growing role in ultra-dense 5G-and-beyond deployments. Most of these works, however, focus only on the energy costs for the network operator or on the involved emission of  $CO_2$  while maintaining a predefined service level. In contrast, we also incorporate the impacts on battery lifetime of the attached mobile devices. By doing this, and in conjunction with our detailed energy models, we provided solutions for context-aware switch-off strategies for base stations during periods of low network activity to reduce the energy costs of the operator with a reasonable impact on the battery lifetime of the mobile handsets.

In addition, client-side approaches also offer a large potential for increasing the network-efficiency by acting cooperatively, according to the current network situation. However, this requires additional information, which first needs to be made available to the clients. Recent works like MobileInsight [116] extract low-level data, e.g., detailed traces of internal state transitions, directly from the transceiver hardware such as an LTE chipset and provide it to the application layer.

Although this gives information about the system itself, the information lacks the greater scope, which is the concurrent competition for radio resources by other network participants. With piStream [125] researchers tried to close this gap by tracking the spectral utilisation with additional equipment, such as real-time spectrum analysers, and adapting an application-layer process, e.g., the resolution of a video stream, according to this information. While this approach gives a coarse estimation of the instantaneous spectral utilisation, the information lacks detail about the degree of competition on the medium. However, this information is essential in case of a full spectral usage, since the fraction of resources assigned to an additional participant is greatly influenced by the number of claiming competitors. Moreover, this approach is capable of observing only the downlink activity (from the base station to the mobile device), because not all mobile devices in the cell might be perceivable by the observer due to shadowing or large distances.

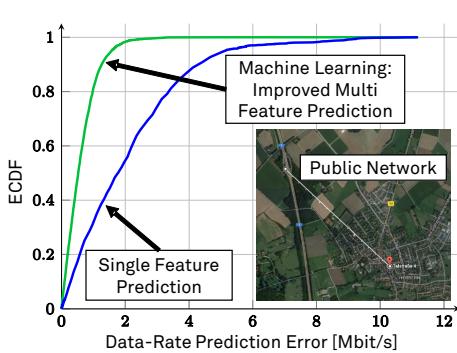


Figure 3.2: Accuracy of our passive data-rate prediction in a public cellular network [A4/136].

Our research closes this gap and provides a method for in-depth analysis of cellular radio resource utilisation in both directions, by performing a real-time analysis of control channels, such as the physical downlink control channel (PDCCH) in LTE, to uncover every single spectral allocation of every participant/competitor in the current cell. In contrast to LTEye's [113] offline approach, we provided the first real-time capable method for client-based control-channel analysis for connectivity estimation (C3ACE) and demonstrated the benefits of the provided information for passive resource predictions. By using additional features in conjunction with cross-validated neural networks, we extended C3ACE (*E-C3ACE*) to reliably predict the data rate of a terminal device in both, laboratory and public networks [A4/136], as shown in Fig. 3.2.

In collaboration with projects A1 and A3 we are cur-

rently pursuing this research to include mobility in the model by utilising methods of deep learning and model-based optimisation (MBO), respectively. At the same time, with OWL [102] other researchers have presented a comparable method for control channel analysis, although with the objective of network analysis. It is based on LTEye's method, which re-encodes decoded data for validation at bit level at the cost of the integrated forward error correction. However, under non-ideal radio conditions both approaches are sensitive to channel disruptions. In contrast, C3ACE performs a validation in a statistical manner, which is independent of the underlying radio conditions and is even applicable at the cell edges.

Detailed knowledge of the current resource utilisation provides novel and valuable metrics to client systems, which can be exploited for improvement of both performance and resource-efficiency. This challenge will be tackled in the next phase, in particular by utilising resource-efficient methods of machine learning from project A1 for resource-predictive floating-car data exchange in project B4. Furthermore, predicted resources, data rates, and delays are essential for profound offloading decisions in edge computing and context-aware mobile network operator selection, which are both applicable in project B2. Finally, tracing the traffic behaviour of real mobile devices facilitates the data-driven generation of representative behaviour models, which are required for simulative network analysis.

Further advances in Radio Frequency (RF) technology yield additional information, such as highly precise arrival time measurement that enables resource-efficient position-based services in future CPSs. In contrast to signal strength-based methods, the quality and reliability of the novel time-based ultra-wideband (UWB) position estimation exceed a technology barrier and enable control-grade feedback for a variety of applications. We were able to show the potential of our UWB technology by enabling unmanned aerial vehicle (UAV) navigation in global navigation satellite system (GNSS) denied environments. In the course of our recent work, we addressed three important aspects for positioning applications. Through a novel time-difference-of-arrival system topology, we increased the energy efficiency at the mobile node and enabled the use of battery-powered devices. We also optimised the scalability by reducing the communication overhead in the real-time localisation system by orders of magnitude. Furthermore, we participated in and won the EvaAL Competition at the International Conference on Indoor Positioning and Indoor Navigation (IPIN) in 2016 and demonstrated the location quality of the proposed approach. Most recently, we integrated the individual components in a multi-UAV testbed and received the Best Paper Award at the 2017 International Conference on Indoor Positioning and Indoor Navigation [A4/129]. Furthermore, five additional publications originated in the course of this work (e.g., IEEE VTC, IPIN), which could not be listed here due to spatial constraints.

**Automated Generation of Hardware Energy Models (WP6, WP8):** On the system platform layer, we researched methods for automated measurement and analysis of hardware energy models. As transiently powered platforms impose the requirement to handle power failures, power management within the system software must be aware of the current system state regarding power income, energy consumption, and planned activities. It also has to predict and detect power failures as soon as possible to react by limiting resources, persisting the system state, and informing communication partners. Therefore, the device is required to keep track of its instant energy balance. However, the complex functionality of existing modelling approaches, which address larger battery-driven platforms like smartphones [121, 110, 104], is not applicable to ultra-low power devices. Here, both the limitations of computational power and the lack of precise in-situ measurements inhibit expensive online regression. Thus, a mix of automated offline measurements combined with online benchmarking or simulation looks promising. Hence, Zhu et al. presented EMrise to answer the question of how many system states and parameters are required to also model the energy consumption of smaller devices [127]. Martinez et al. propose a model that includes energy harvesting, conversion and system functionality for IoT networks [118]. However, this model cannot easily be used for components on stand-alone systems.

To bridge the gap between automatic online modelling for larger, battery-driven platforms and the laborious creation of offline models for strongly resource constrained systems, we created an automatic measurement, analysis and modelling approach for the latter [A4/130]. Here, we were able to successfully combine former achievements of the project. A priced-timed automata (PTA)-based modelling scheme [A4/131] is used to support online energy accounting, and thus became the foundation of an approach to automatically generate energy cost models. These automata allow for deriving a list of software function calls that drive all components through all of their states and becomes the input for benchmark synthesis, which is run on the target device. This benchmark application is also in control of an externally connected energy measurement toolchain MIMOSA, which was developed in phase 1. MIMOSA logs the energy consumption of the corresponding model states and transitions of multiple benchmark runs. The measurement results are automatically collected by an analysis module, which uses a broad set of regression algorithms and a cross-validation algorithm to derive valid energy models for the parameter ranges of the observed system function calls. The results are then annotated as costs to the original PTA representation. The loop is closed by re-synthesising the PTA-based (energy accounting) system drivers and further model refinement. We were invited to contribute a book chapter about this topic in “ICT - Energy Concepts for Energy Efficiency and Sustainability”, in cooperation with the ICT Energy<sup>1</sup> group.

**Energy Harvesting (WP5):** Since an energy harvester is the first step in the power supply chain to a resource constrained system, it has to be evaluated and modelled precisely. A power supply module for a CPS consists of different sub-components that have to be modelled both separately and as a whole. According to the application field and power requirements, the number of these modules may differ. However, a three-compartment system is considered, consisting of energy harvesting, storage, and converter. Since multiple models for batteries [105], supercapacitors [119] and capacitors are already available, this project contributes to harvester and converter, providing an overall system model.

Among different possible harvesting principles, photovoltaic (PV) harvesting had proven best performance and has been focused on. Current knowledge of PV harvesting is mostly focused on the outdoor applications under AM1.5 condition with an arbitrary maximum terrestrial intensity, one sun spectra and  $1000 \text{ W/m}^2$  intensity. However, most industrial applications of current CPS are performed indoors. Therefore, available models for the outdoor application have to be re-evaluated for those cases. For this purpose, a special measurement platform has been developed [A4/133], replicating different indoor lighting conditions in a reproducible way. Using measurement data collected from this platform, a model for indoor PV harvesters has been proposed in [A4/132]. This model is specifically focused on the low light condition of light intensities lower than  $10 \text{ W/m}^2$ . However, this model does not include temperature aspects of the PV cells, because the indoor temperature can be considered as a constant value.

The last common component in an energy harvesting system and model is a converter connecting the harvester to the storage device. Energy harvesters and storage mostly show nonlinear behaviour. According to this nonlinearity, there are special operational conditions that lead to optimal energy scavenging. If the harvester was connected directly to the storage device, its voltage would be clipped to the value of the storage. It is very likely that this value is not the optimal operating point of the harvester known as maximum power point (MPP). Consequently, an intermediate device is required that keeps the harvester near its optimal operational point, matches its voltage to the storage and avoids overcharging it. We contributed multiple solutions for this middleware [A4/134] according to the specific requirements of each platform. However, the most reliable off-the-shelf components that are commonly used in multiple applications [A4/134] are two devices developed by Texas Instruments, which are used in PhyNode as well. Modelling of these chips using intensive deductive and inductive effort with extensive data collection has been done. A machine learning-based model of these chips is published in [A4/134] as the first available overall model.

---

<sup>1</sup>ICT Energy is a Coordination Activity of the European Union: <http://www.ict-energy.eu>

In contrast to that, we also conducted research in methods for cost-effective adaptation to transient power sources and limit the number of those extra components as much as possible. The approach adapts the workload of the device according to the momentary energy income, thus achieving energy neutrality. However, this requires *in situ* measurements of the energy income. To reduce the hardware costs for these measurements, we are currently investigating low-cost measurement methods and their combination with energy-neutral adaptation strategies in software. Here, we try to exploit parasitic properties of the components already in use. For example, the flank time of an MCU input pin is a result of the parasitic gate capacity of the attached internal transistor and follows the laws of capacitor charging. With a single external resistor and a programmable, internal pull-down resistor, flank-time measurements within the MCU have proven to be a reliable method for estimating energy income in our lab experiments. We achieve energy neutrality by reducing access to peripheral devices according to the energy income. By doing that, we move the power supply management from the hardware layer to software. The whole setup comes at additional expenses of less than 1 euro per unit in production costs and shows an energy efficiency of 70 %, which is close to the state of the art [109]. As a trade-off, our approach has higher minimum energy requirements than approaches using hardware management components, because it requires a powered MCU to drive power-management decisions. When powering the MCU fails, e.g., in low-light situations, inefficiencies occur.

To the best of our knowledge, other research conducted in this field does not focus on cost optimisation. For example, Brunelli et al. [101] calculated additional costs of approximately 50 euros for their approach in 2009, and more recent approaches [100, 114] still use expensive external components such as comparators and reference-voltage sources.

**Design-Space Exploration and Simulation (WP7, WP9):** Finally, in [A4/137] we merged previous methods, models, hardware, and context definitions and analysed the impact of radio channel access procedure and channel congestions on the energy consumption of network nodes in the PhyNetLab. The energy model for the radio transceiver was automatically derived from the embedded device radio driver as described previously. With this collaborative setup we evaluated the performance of the radio channel access scheme, which is defined for this band by the standard ETSI EN 300 220-1. The network was challenged by the typical logistics use case of querying for specific products in the warehouse. In terms of the underlying communication systems, this represents synchronous replies to broadcast messages by those nodes that match the requested goods. The experiments in the testbed provide quantitative results that show, that using this scheme with an increasing amount of competition on the radio channel raises the overall energy consumption the network nodes significantly due to collisions, repeated wake-up events, and larger energetic effort for clear channel assessment (CCA). Subsequent work addressed the improvement of the channel access procedure in dense networks and provided an OMNeT++-based and open-source simulation framework for design-space exploration and trade-off analysis to optimise the communication systems for large-scale IoT deployments such as PhyNetLab. The framework is designed to be extensible in the following two directions, which can be combined: System-in-the-loop simulations provide optimal system parameters for test runs in the testbed. In turn, the test runs provide model knowledge and updates for subsequent simulation runs. On the other hand, the simulation provides an adaptation layer for integration of mobility and the high-level logistics process, which is based on the actor model (*Akka*) and is capable of being deployed as simulation, testbed, or hybrid.

Additional evaluations of the overall developed platform in PhyNetLab have been done during the CRC876 summer school in 2017. Here, the main task was the localisation of PhyNodes inside a warehouse using environmental data. A replica of a warehouse as shown in Fig. 3.1 has been built in a small-scale setup, representing one aisle with two sides. Three APs enabled the communication in between and provided received signal strength indication (RSSI) values, which can be used for localisation by means of triangulation.

Different environmental data are stored within different experiments over a longer period of multiple days. These data were used during the summer school by a large group of attendants with different backgrounds and levels of knowledge about hardware and software for resource constrained systems. Each group developed and implemented an algorithm based on machine learning to localise their PhyNode using the training data. In spite of the limitation on available energy and limited memory space, four groups were able to develop a program and run it on the PhyNode.

This process not only proved the functionality of the whole system including PhyNodes and APs, but also demonstrated the maturity of the toolchain and experimental environment developed. It showed that even beginners with limited knowledge of the PhyNetLab are able to program and integrate their applications within a very short time period. Detailed information about the system structure, data collection procedure, developed models and their performance as well as constraints resulted in a joint publication with project A1, which is omitted due to spatial constraints.

## Bibliography

- [100] D. Balsamo, A. Das, A. S. Weddell, D. Brunelli, B. M. Al-Hashimi, G. V. Merrett, and L. Benini. “Graceful Performance Modulation for Power-Neutral Transient Computing Systems”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 35.5 (May 2016), pp. 738–749 (cit. on p. 134).
- [101] D. Brunelli, C. Moser, L. Thiele, and L. Benini. “Design of a Solar-Harvesting Circuit for Batteryless Embedded Systems”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 56.11 (Nov. 2009), pp. 2519–2528 (cit. on p. 134).
- [102] N. Bui and J. Widmer. “OWL: A Reliable Online Watcher for LTE Control Channel Measurements”. In: *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*. ATC ’16. New York, NY, USA: ACM, 2016, pp. 25–30 (cit. on p. 132).
- [103] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer. “A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques”. In: *IEEE Communications Surveys & Tutorials* 19.3 (2017), pp. 1790–1821 (cit. on pp. 130, 131, 143, 248).
- [104] M. Dong and L. Zhong. “Self-constructive high-rate system energy modeling for battery-powered mobile systems”. In: *Proceedings of the 9th international conference on Mobile systems, applications, and services*. ACM. 2011, pp. 335–348 (cit. on p. 132).
- [105] N. El Ghossein, J. P. Salameh, N. Karami, M. El Hassan, and M. B. Najjar. “Survey on electrical modeling methods applied on different battery types”. In: *2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAECE)*. Apr. 2015, pp. 39–44 (cit. on p. 133).
- [106] M. J. Farooq, H. Ghazzai, E. Yaacoub, A. Kadri, and M. S. Alouini. “Green Virtualization for Multiple Collaborative Cellular Operators”. In: *IEEE Transactions on Cognitive Communications and Networking* 3.3 (Sept. 2017), pp. 420–434 (cit. on p. 131).
- [107] H. Ghazzai, M. J. Farooq, A. Alsharoa, E. Yaacoub, A. Kadri, and M. S. Alouini. “Green Networking in Cellular HetNets: A Unified Radio Resource Management Framework With Base Station ON/OFF Switching”. In: *IEEE Transactions on Vehicular Technology* 66.7 (July 2017), pp. 5879–5893 (cit. on p. 131).

- [108] I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, and D. J. Leith. “srsLTE: An Open-source Platform for LTE Evolution and Experimentation”. In: *Proceedings of the Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*. WiNTECH ’16. New York, NY, USA: ACM, 2016, pp. 25–32 (cit. on p. 130).
- [109] A. Gomez, L. Sigrist, M. Magno, L. Benini, and L. Thiele. “Dynamic Energy Burst Scaling for Transiently Powered Systems”. In: *Proceedings of the 2016 Conference on Design, Automation & Test in Europe*. DATE ’16. San Jose, CA, USA: EDA Consortium, 2016, pp. 349–354 (cit. on p. 134).
- [110] M. B. Kjærgaard and H. Blunck. “Unsupervised Power Profiling for Mobile Devices”. In: *Mobile and Ubiquitous Systems: Computing, Networking, and Services. MobiQuitous 2011*. Ed. by A. Puiatti and T. Gu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 138–149 (cit. on p. 132).
- [111] S. Krishnasamy, P. T. Akhil, A. Arapostathis, S. Shakkottai, and R. Sundaresan. “Augmenting max-weight with explicit learning for wireless scheduling with switching costs”. In: *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. May 2017, pp. 1–9 (cit. on p. 131).
- [112] S. S. Kumar and A. Kumar. “Energy efficient rate coverage with base station switching and load sharing in cellular networks”. In: *2016 8th International Conference on Communication Systems and Networks (COMSNETS)*. Jan. 2016, pp. 1–6 (cit. on p. 131).
- [113] S. Kumar, E. Hamed, D. Katabi, and L. Erran Li. “LTE Radio Analytics Made Easy and Accessible”. In: *Proceedings of the 2014 ACM Conference on SIGCOMM*. New York, NY, USA: ACM, 2014, pp. 211–222 (cit. on p. 131).
- [114] H. G. Lee and N. Chang. “Powering the IoT: Storage-less and converter-less energy harvesting”. In: *The 20th Asia and South Pacific Design Automation Conference*. Jan. 2015, pp. 124–129 (cit. on p. 134).
- [115] Q. Li, G. Dublon, B. Mayton, and J. A. Paradiso. “Marshvis: Visualizing real-time and historical ecological data from a wireless sensor network”. In: *IEEE VIS Arts Program (VISAP)* (2015) (cit. on p. 129).
- [116] Y. Li, C. Peng, Z. Yuan, J. Li, H. Deng, and T. Wang. “MobileInsight: Extracting and Analyzing Cellular Network Information on Smartphones”. In: *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking*. MobiCom ’16. New York, NY, USA: ACM, 2016, pp. 202–215 (cit. on p. 131).
- [117] N. Ludant, N. Bui, A. G. Armada, and J. Widmer. “Data-Driven Performance Evaluation of Carrier Aggregation in LTE-Advanced”. In: *The 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC 2017)*. Montreal, Canada: IEEE, Oct. 2017, pp. 1–6 (cit. on p. 130).
- [118] B. Martinez, M. Montón, I. Vilajosana, and J. D. Prades. “The Power of Models: Modeling Power Consumption for IoT Devices”. In: *IEEE Sensors Journal* 15.10 (Oct. 2015), pp. 5777–5789 (cit. on p. 132).
- [119] A. Namisnyk and J Zhu. “A survey of electrochemical super-capacitor technology”. In: *Australian Universities Power Engineering Conference*. University of Canterbury, New Zealand. 2003 (cit. on p. 133).
- [120] OpenAirInterface Software Alliance. *Open Air Interface*. 2018 (cit. on p. 130).
- [121] A. Pathak, Y. C. Hu, M. Zhang, P. Bahl, and Y.-M. Wang. “Fine-grained power modeling for smartphones using system call tracing”. In: *Proceedings of the sixth conference on Computer systems*. ACM. 2011, pp. 153–168 (cit. on p. 132).

- [122] V. Pejovic and M. Musolesi. "Anticipatory Mobile Computing: A Survey of the State of the Art and Research Challenges". In: *ACM Comput. Surv.* 47.3 (Apr. 2015), pp. 47:1–47:29 (cit. on p. 130).
- [B4/123] **B. Sliwa**, J. Pillmann, F. Eckermann, **L. Habel**, **M. Schreckenberg**, and **C. Wietfeld**. "Lightweight joint simulation of vehicular mobility and communication with LIMoSim". In: *IEEE Vehicular Networking Conference (VNC)*. Torino, Italy, Nov. 2017 (cit. on pp. 141, 251).
- [B2/124] **P. Libuschewski**. "Exploration of Cyber-Physical Systems for GPGPU Computer Vision-Based Detection of Biological Viruses". Diss. Dortmund, Germany: TU Dortmund, 2017 (cit. on pp. 130, 195–199).
- [125] X. Xie, X. Zhang, S. Kumar, and L. E. Li. "piStream: Physical Layer Informed Adaptive Video Streaming over LTE". In: *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. MobiCom '15. New York, NY, USA: ACM, 2015, pp. 413–425 (cit. on p. 131).
- [126] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang. "Big data-driven optimization for mobile networks toward 5G". In: *IEEE Network* 30.1 (Jan. 2016), pp. 44–51 (cit. on p. 131).
- [127] N. Zhu and A. V. Vasilakos. "A generic framework for energy evaluation on wireless sensor networks". In: *Wireless Networks* 22.4 (2016), pp. 1199–1220 (cit. on p. 132).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [A4/128] **A. K. Ramachandran Venkatapathy**, **M. Roidl**, A. Riesner, **J. Emmerich**, and **M. ten Hompel**. "PhyNetLab: Architecture design of ultra-low power Wireless Sensor Network testbed". In: *IEEE 16th International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. June 2015, pp. 1–6 (cit. on pp. 129, 138).
- [A4/129] **J. Tiemann** and **C. Wietfeld**. "Scalable and Precise Multi-UAV Indoor Navigation using TDOA-based UWB Localization". In: *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. Sapporo, Japan, Sept. 2017, pp. 1–7 (cit. on pp. 132, 142).
- [A4/130] **M. Buschhoff**, **D. Friesel**, and **O. Spinczyk**. "Energy Models in the Loop". In: *Procedia Computer Science* 130 (2018), pp. 1063 –1068 (cit. on pp. 15, 133).
- [A4/131] **M. Buschhoff**, **R. Falkenberg**, and **O. Spinczyk**. "Energy-Aware Device Drivers for Embedded Operating Systems". In: *SIGBED Review* (2018) (cit. on pp. 15, 133).
- [A4/132] **M. Masoudinejad**, M. Kamat, **J. Emmerich**, and **M. ten Hompel**. "A Gray Box Modeling of a Photovoltaic Cell under Low Illumination in Materials Handling Application". In: *IEEE International Renewable and Sustainable Energy Conference (IRSEC)*. Dec. 2015, pp. 1–6 (cit. on p. 133).
- [A4/133] **M. Masoudinejad**, **J. Emmerich**, D. Kossmann, A. Riesner, **M. Roidl**, and **M. ten Hompel**. "A Measurement Platform for Photovoltaic Performance Analysis in Environments with Ultra-Low Energy Harvesting Potential". In: *Sustainable Cities and Society* 25 (2016), pp. 74 –81 (cit. on p. 133).

- [A4/134] **M. Masoudinejad**, M. Magno, L. Benini, and **M. ten Hompel**. "Average Modelling of State-of-the-Art Ultra-low Power Energy Harvesting Converter IC". In: *International Symposium on Power Electronics, Electrical Drives, Automation and Motion (accepted for presentation)*. Aug. 2018 (cit. on pp. 15, 48, 133).
- [A4/135] **R. Falkenberg**, **B. Sliwa**, and **C. Wietfeld**. "Rushing Full Speed with LTE-Advanced is Economical - A Power Consumption Analysis". In: *IEEE Vehicular Technology Conference (VTC-Spring)*. June 2017, pp. 1–7 (cit. on p. 130).
- [A4/136] **R. Falkenberg**, **K. Heimann**, and **C. Wietfeld**. "Discover Your Competition in LTE: Client-Based Passive Data Rate Prediction by Machine Learning". In: *IEEE Globecom*. Singapore, Dec. 2017, pp. 1–7 (cit. on pp. 131, 141).
- [A4/137] **R. Falkenberg**, **M. Masoudinejad**, **M. Buschhoff**, **A. K. Ramachandran Venkatapathy**, **D. Friesel**, **M. ten Hompel**, **O. Spinczyk**, and **C. Wietfeld**. "PhyNetLab: An IoT-based warehouse testbed". In: *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*. Sept. 2017 (cit. on pp. 19, 134).

## 3.4 Project plan

### Goals

The project's main goal in the third phase of the CRC is to expand the scope of model generation and information mining to the next level to not only cover distinct resource constrained platforms but also whole IoT-systems. Consequently, this requires modelling heterogeneous platforms including their individual capabilities and map tasks among them. The vision of the A4 project is illustrated in Fig. 3.3. Achievements in previous phases have enclosed the circle of system context, resource models, integration into applications, and corresponding control models around configurable uniform platform families. In the upcoming phase, the platform itself will be scaled and diversified at first, followed by modelling the whole systems. This also involves extensions of previous models, e.g., communication systems, as well as entirely new models, such as for mobility and localisation systems. As a consequence of up-scaling the density of interacting entities, new methods for resource-efficient communication have to be found to ensure the reliable interaction and operation of single entities and the whole system, respectively. Furthermore, this requires a concept for interoperability of such heterogeneous platforms, which has to be developed as well. Besides ultra-low-power platforms, such as the PhyNode in PhyNetLab [A4/128], devices with larger power requirements will also be part of the project's considerations. To ensure their operation and increase the reliability of these transiently powered devices, methods of parallel harvesting will also be part of the project.

Three main objectives are defined for the third phase, which build on each other and include collaborations with other projects. First, the project will provide fundamental concepts for building large-scale heterogeneous IoT systems. Based on this, the second objective comprises the provision of required models and methods that adequately describe the underlying entities' resources, capabilities, communication links and external influences. The third objective targets the integration and evaluation of previous results in order to prove the system's functionality as a whole. It includes the provision of system simulations, simulations of subsystems with attached hardware components, and exemplary large-scale deployments.

Before scaling a system and diversifying its entities, semantics and a common understanding are required to build harmony between these subsystems. System description and synthesis, coupling points, and inter-domain interaction are some aspects of heterogeneous system conception. This is not limited to human readable (understandable) language but also includes attributes for ma-

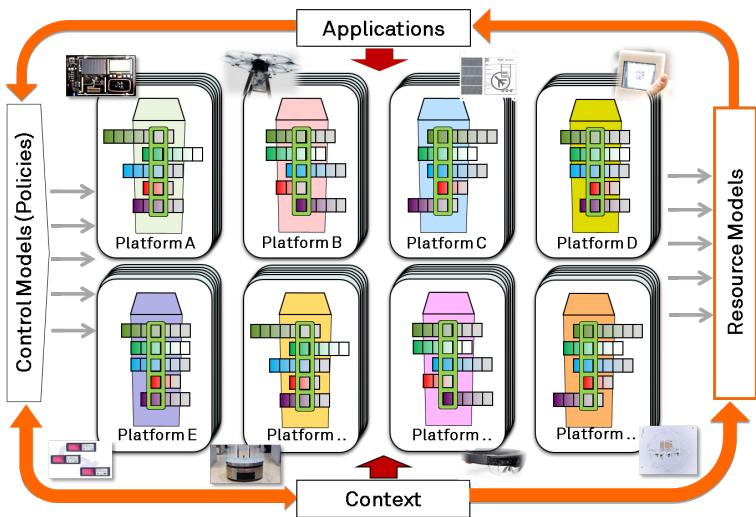


Figure 3.3: Project vision of providing methods for resource-aware application development and data analysis on large-scale heterogeneous and resource constrained IoT platforms.

chine integration and massive machine type communication (mMTC). For instance, a transport operation can be performed by a mobile robot, a drone, a human being, or a combination of them depending on available resources. This transport operation can be propagated to the system by heterogeneous entities such as a human, an IoT device, or a materials handling system. Realisation of this transport task has to be represented using a certain set of attributes in a specific semantic common for all participating entities.

However, scaling and diversifying systems drive current approaches to their resource limits. In particular this applies to wireless communication, localisation, and methods of reliable energy harvesting. Therefore, this project will provide models and methods addressing these aspects.

With the ongoing evolution of cellular networks towards 5G, these networks become attractive options for connecting IoT-devices in industrial facilities without the need to operate a dedicated network instance. However, the previous project phase pointed out the large impact of dynamic network load on the availability of spectral resources and energy-efficiency of transmissions. In particular, these network dynamics are also affected by auxiliary network participants, which are not part of the industrial process itself but share the same radio resources. Hence, models for heterogeneous IoT environments have to be developed to consider their impacts during system design and integrate them into the system itself. In order to evaluate and adapt emerging approaches to increase the spectral efficiency, such as pencil beam mmWave transmissions, appropriate channel models for industrial facilities have to be provided. These also have to be combined with mobility models of moving entities on the ground, e.g., driverless transport systems (DTSs), and in the air, e.g., UAVs.

Moreover, industrial processes, e.g., transport robots for material handling, demand efficient indoor localisation techniques. Some entity types use their own specific localisation techniques while some rely on other entities. Generalised models have to be generated to enable multimodal indoor localisation using available data from existing methods in addition to the other developed methods for indoor localisation.

Every entity will have its own specification on size, weight, and cost limiting system design. In a heterogeneous system, multiple harvesting methods have to be evaluated for the proper selection of energy-harvesting principles. Moreover, methods for parallel energy harvesting will be developed

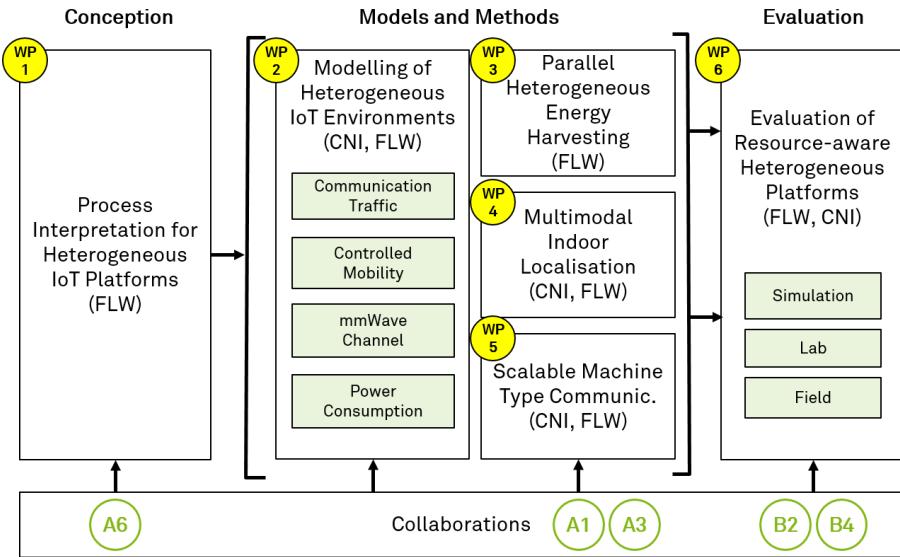


Figure 3.4: Work package overview of the A4 project.

to fulfil all tasks at all time instances for energy-intensive platforms.

Applications of developed methods and models will be shown by integrating them into scenarios using specific concepts for interaction of heterogeneous platforms, networks, or systems. In order to prove the applicability of the approaches developed through all three phases of the project, multiple exemplary large-scale systems will be realised. This will not be limited to logistics scenarios but will also be transferred to the planned collaborations with other projects, e.g., B2, B4.

### Work schedule

The work in the third funding period of this project is structured in six work packages with tight collaborations with A1, A3, A6, B2, and B4. Concepts for process interpretation of heterogeneous IoT platforms are developed in WP1. Afterwards, this project focuses on providing models and methods required for the realisation of the application scenarios (WP2–WP5). Finally, WP6 includes the validation at subsystems level as well as evaluation of the entire system as a high-density large-scale deployment. The WPs are either assigned to one research group, the communication networks institute (CNI) or the chair of material handling and warehousing, German *Lehrstuhl für Förder- und Lagerwesen* (FLW), or are processed jointly by both groups.

**Work package 1. Process Interpretation for Heterogeneous IoT platforms** In phase 2, homogeneous platforms enabled a direct interpretation of IoT system tasks to process operations. However, integration of diverse entities into an industrial system adds an extra degree of complexity to model tasks and operations. For instance, a transport operation from a logistics system can be performed by a mobile robot, a drone, a human operator, or a combination of them. Since direct interpretation of semantics in diverse systems is not possible, a common representation of tasks and operations is required. To enable automated interpretation of tasks and operations, a method needs to be developed, which includes a software tool, a domain specific language, and the definition of semantics in the communication. This method enables the generation of feature sets of both tasks and operations. During the design phase each task has to be matched into at least one set of operations. Depending on the task availability of the systems, operations demand can

be modelled during runtime. In addition to this method, machine-understandable definitions of operations and tasks need to be developed. The three important aspects of the developed model are tasks from IoT platforms, operations from processes, and the resources that will be analysed in the next part of this project. The general system representation developed in this WP has to be extended or revised if the system characteristics change as an outcome of the system modelling or introduction of new methods. In collaboration with project A6 a knowledge exchange will be established, where resource-aware platforms can be represented in a graph model. This representation provides an interpretation of processes and available systems. Therefore, it enables realisation of process collaboration for heterogeneous systems.

**Work package 2. Modelling of Heterogeneous IoT Environments** The process of scaling heterogeneous IoT networks requires detailed knowledge of the environment in which they should operate. Therefore, this work package will provide models to represent controlled mobility, communication traffic, and radio channels for novel 5G mmWave transmissions, as well as models to represent the power consumption involved for each component.

The mobility models will represent the movement of entities in industrial facilities, such as robots, connected containers, and even human workers within the planned system simulations. These models are tightly bound with models for communication traffic since these entities will be interconnected by wireless networks. As phase 2 already showed [A4/136], it is important to incorporate auxiliary network users (beyond the scope of the considered IoT system) into the modelling process to cover their influences on the communication system appropriately, e.g., congestions and latencies. Therefore, this work package will not only provide models for communication traffic of various system platforms but also models of externals, such as floating car data in collaboration with project B4 [B4/123]. Due to the heterogeneous manner of participants and devices (humans, wearables, robots, machines), analytical modelling is not practical. Therefore, data-driven approaches will utilise machine learning on empirical data to derive the models of such environments from real network traffic. The necessary data collection will be performed by extended methods, which have been developed in the previous phase of the project, e.g., C3ACE [A4/136].

This work package will also investigate 5G mmWave for the application in industrial environments, because of its potential for high spectral efficiency. Based on our novel 5G mmWave system with beamforming and pencil-beam antennas, a preliminary work already demonstrates the potential and applicability of air-to-ground communication with mobile UAVs together with highly directive antennas at 28 GHz. Due to the increased amount of reflections on metal surfaces as well as shadowing by such obstructions, additional channel models are required, which can be combined with traffic and mobility models of this package. These models will also be combined with approaches from previous CRC phases, such as CoPoMo, to provide detailed power consumption models of the IoT system platforms in those heterogeneous environments. Furthermore, marker-less tracking with RGB cameras in a closed space will be used for motion prediction, which in turn enables the development of mobility models of humans inside the facility.

**Work package 3. Parallel Heterogeneous Energy Harvesting** PV energy harvesting, especially in indoor environments, has been analysed and modelled in the last phase of the project. However, there are applications where the power demand of a system is extremely dynamic. A common logistics use case is a container equipped with PhyNode. As long as the system is in a normal situation, a typical operation and energy pattern occurs. However, when an order picking is initiated, containers communicate a much larger amount of data packets within a short time period. Although advancement from the current platform will assure operation of the system, this introduces a hard limitation on the system and its performance. To conquer this challenge and extend the power supply to the system, other harvesting options have to be analysed as well. For

instance, in the PhyNode example, a vibration-based harvester can be used in parallel to the PV harvester. While the container will be moving during the intensive demand phase of order picking, this secondary harvester provides more energy to the system.

In addition, the diversity of device structures, especially in a heterogeneous platform, requires an expansion of the harvesting principle as well. Therefore, other possible energy harvesting techniques such as (but not limited to) radio-frequency, thermal, and vibration harvesters will be evaluated. This includes an examination of pros and cons of each technique and matching them for specific available applications.

The design of converter and energy management has to be revised to enable parallel harvesting. In particular, parallel versions of maximum power point tracking (MPPT) have to be developed to assure continuous best performance of all harvesters. Moreover, harvester balancing has to be developed according to the models of different harvesters. Harvester balancing is motivated by the fact that in some cases switching from one harvester to the other has better performance than direct parallel harvesting according to the overhead energy required for running MPPT and converter.

**Work package 4. Multimodal Indoor Localisation** Although we have already achieved internationally competitive location quality [A4/129], our goal is to further improve those results in terms of accuracy, energy-efficiency, and latency. Current approaches to indoor localisation do not leverage detailed, fine-grained channel response information to improve the location accuracy due to unavailability or the significant processing overhead. We aim to leverage channel response data inherently generated at the receiver side that is currently unused due to the significant processing overhead. In the course of the next phase, the cross-project CRC machine learning competence will help to process the massive data required for the analysis. A two-step approach will help to improve the location accuracy in two scenarios. In a generic setup without a known environment, the accuracy will be increased through a generalised approach. Further detail and accuracy will be achieved by training advanced models for specific environments where a novel channel response-based fingerprinting is expected to significantly increase the location estimation.

A heterogeneous platform-based approach will help to develop the next generation of wireless localisation. The PhyNodes and UWB localisation developed in phase 2 will be incorporated in a multi-physical-layer localisation testbed. Here, novel and high-performance (2 GHz bandwidth, 28 GHz carrier frequency) SDR communication technology will be utilised to develop the future of wireless localisation. A multimodal 5G-integrated wireless localisation testbed has not been analysed yet due to unavailability and significant cost of the required 5G new radio (NR) hardware. The greater bandwidths and beam-steering antennas of those future 5G technologies will yield revolutionary location accuracy through precise time-of-arrival estimation. With our experience in existing time-based location estimation, modern beam-steering antenna systems, and our cross-project machine learning experience, we aim to raise the bar in wireless localisation through novel algorithms.

The research hall as shown in Fig. 3.5 has environmental sensors embedded in the floor for measuring sound and vibration. This data from the sensors can be communicated using a multi-physical-layer transceiver to a data sink to monitor the environmental changes in the hall. From the measurement of the wireless parameters, a phase-synchronised SDR deployment is used. Through the SDR measurements it will be possible to understand the channel state information and channel properties of the communicated data. Using this deployment of a sensor floor and SDRs, active wireless indoor localisation of communicating nodes can be developed for ultra-low-power devices. Additionally, other entities such as humans and robots can be localised passively by aggregating sensor information as well as channel state information.

To experimentally evaluate and demonstrate the achieved results, a multi-system testbed in the logistics localisation test centre will be developed utilising a high-performance motion capture system for ground-truth measurements. A practical deployment of different heterogeneous systems will be combined and fused using multi-stage coordinate transformations to obtain a common frame of reference among the individual systems.

**Work package 5. Scalable Machine Type Communication** Based on the models from WP2, WP3, and WP4, this package will bring together the developed models to provide methods for massively scalable communication systems with special emphasis on ultra-dense industrial IoT systems. Although 5G cellular networks, as well as narrow band IoT (NB-IoT) and enhanced machine type communication (eMTC), claim to provide mature solutions for these use cases, the massive densification will quickly drive even those systems into a saturated state. A promising approach to overcome the resource limitation on the radio spectrum is the utilisation of massive multiple input multiple output (MIMO) antenna arrays in conjunction with mmWave transmissions. Such antennas, together with short wavelengths, concentrate the emitted signal to a pencil beam, which can be steered almost instantly in arbitrary directions. Since spectral resources are only allocated along a single transmission beam, the resources can be simultaneously and multiply reused in narrowest space without interference. Therefore, this package will continue the examination of 5G mmWave approaches and identify suitable methods to achieve the highest scalability at given constraints such as availability of energy, computational capabilities, mobility, and current channel conditions. Besides typical performance indicators such as transmission time, latency, and coverage, our power-consumption models for all system layers will provide valuable additional quality criteria for system design.

Furthermore, the determinism and predictability of machine type communication (MTC) open up a large potential for cooperative networking to overcome the resource constraints of ultra-dense environments. Therefore, this WP will provide methods for anticipative and altruistic platform behaviour to save resources and increase the capacity of the entire system. In contrast to centralised approaches [103], distributed and client-based methods require those platforms to perceive the momentary state of their communication systems. To achieve this, we will develop prediction methods to provide information about the instant spectral utilisation in dense environments, which are based on passive indicators. Hence, the prediction does not stress the communication system by active probing or failing transmission attempts. Besides readily available indicators like received signal strength indication (RSSI), which are typically coarse-grained, we will provide detailed supplemental information through all protocol layers by using software-defined radios. The versatile information sources will be combined by resource-efficient machine learning methods from project A1 to derive compact prediction models. Furthermore, in collaboration with project A3 we will leverage online learning and Model-Based Optimisation with Concept Drift (MBO-CD) to obtain a continuously adapting prediction model in changing environments. The anticipative approach will also incorporate combinations of technologies in two ways: first, exploitation of side-effects to cover two use cases by one technology, e.g., combining communication with localisation (see WP4); second, multi-radio or multi-technology approaches in conjunction with resource-predictive link selections.

However, gaining the required information and integrating the full models might still overtax the resources of deeply embedded devices and foil any potential merit of this approach. To avoid this, parts of those models, as well as the retrieval of information, need to be offloaded and bundled into the infrastructure to provide condensed knowledge to the resource constrained devices. For example, access points might periodically broadcast network-load information for multiple channels, which is used by nodes as needed, i.e., just before a pending transmission.

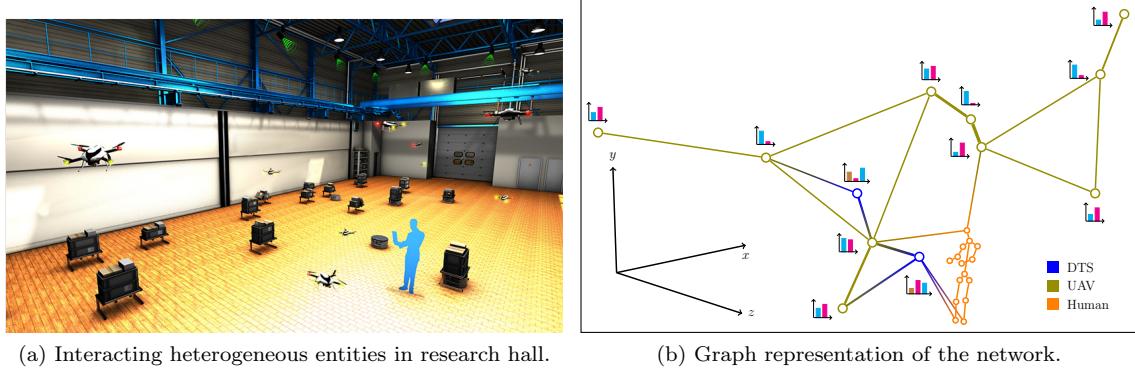


Figure 3.5: Overview of the envisioned innovation laboratory of interacting and heterogeneous network entities (e.g., DTS, UAVs, humans), which safely and reliably fulfil distributed tasks in a logistics scenario. Jointly with A6, graph models will represent entities as vertices, spatio-temporal relationships as edges, and self-descriptions (e.g., energy, transmission speed) as vertex features in order to provide a first approach on predictive collision-avoidance algorithms.

**Work package 6. Evaluation of Resource-aware Heterogeneous Platforms** This work package brings together the developed models and methods in order to perform large-scale and system-wide evaluations. However, due to the complexity, the validation must be performed in a multi-methodical approach that includes simulations at full-scale, trial runs in the laboratory, and field deployments. The first part of this package includes the development of a simulator that covers the project achievements throughout all three phases of the CRC. It will include models of logistic processes, the involved mobility, their requirements for communication, and models for the communication system itself as well as models of other involved resources, such as energy consumption, memory, and computational capabilities. These simulations will be supported by system-in-the-loop extensions, where parts of the simulation are interfaced to real hardware that runs in the laboratory. Finally, validation of the simulation results will be proven using the full field-deployments in the extended PhyNetLab.

This includes integration of the developed models and approaches from previous work packages into heterogeneous platforms which on one hand will be used in other projects of the CRC (e.g., B4 and B2). On the other hand, it will leverage machine learning solutions developed within the CRC (such as A1, A6). Different developed concepts, methods, and simulations will be validated by the realisation in multiple scenarios such as:

- A4 main use case (cooperating with A1, A6): The research hall for the innovation laboratory as shown in Fig. 3.5, will be used as a heterogeneous platform with various benchmarking systems for the evaluation of models and methods developed as a part of the system. This platform has heterogeneous systems such as driverless transport systems (DTSs), heterogeneous IoT devices, and unmanned aerial vehicles (UAVs). The models and methods for resource awareness are integrated into the heterogeneous platform in context of the process. This integration is based on WP1, where the process is interpreted as resource aware entities that fulfil various operations. System interpretability and interoperability are the two mammoth tasks that are focused on this work package during the transfer phase of this CRC. In collaboration with A6, graph models will enable the realisation of scenarios that require self-organising systems, interoperability of the heterogeneous systems, and interpretability for resource-aware process representation. These points are required for massively-scalable robot-assisted (ground and air) indoor logistics and include interoperability of systems for communication, localisation, and energy harvesting. This scenario also leverages machine learning algorithms from other projects, such as embedded streaming from A1.

- B4: Realisation of traffic scenarios, which incorporate a heterogeneous mix of vehicles (from fully manually controlled to fully autonomous vehicles) and 5G NR beam-assisted communications for machine learning improved traffic efficiency. This will be realised and validated jointly with B4.
- B2: Contribution of methods for reliable and efficient communication to application cases in project B2. Here, active pharmaceutical ingredients need to undergo a quality check by PAMONO sensors after shipment to developing countries. Due to limited (network-) infrastructure in such regions, the task of large-scale quality control is only feasible with efficient transmission of the sensor data.
- B2: Finally, the production lines of active pharmaceutical ingredients will also undergo the evolution towards industry 4.0, where reliable wireless communication enables fast and dynamic scaling of the production, e.g., in case of a pandemic.

Eventually, further scenarios have to be identified as the models and methods are developed in various contexts of the resource aware platforms.

### Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Process Interpretation for Heterogeneous IoT platforms																	
2. Modelling of Heterogeneous IoT Environments																	
3. Parallel Heterogeneous Energy Harvesting																	
4. Multimodal Indoor Localisation																	
5. Scalable Machine Type Communication																	
6. Evaluation of Resource-aware Heterogeneous Platforms																	

### 3.5 Role within the Collaborative Research Centre

Due to the tight relation of the A4 project to resource constrained platforms, it has many contact points to other projects in terms of sharing our achievements with them, and benefitting from methods and models provided by them. On the one hand, the project provides versatile resource models for resource constrained platforms from mobile robots and smartphones down to deeply embedded and energy-harvesting systems. This includes methods for the generation and application of these models. They serve other projects as tools or additional metrics to optimise non-functional system properties, e.g., expanding battery lifetime and scalability. On the other hand, this project also utilises methods of machine learning and data mining to support model generation. This

## Project A4

involves methods for model compression, because of the spatial and computational constraints of embedded platforms, optimisation of hyperparameters, and identification of subsets of the most important features, which still enable sufficiently accurate models.

In the next funding phase, we will continue the successful collaboration with B4, which repeatedly applies the resource models provided by A4 to vehicular communication approaches. Besides resource and cost models for the communication medium, A4 provides an evaluation platform, which is leveraged by B4 to improve the resource-efficiency of predictive scheduling approaches for vehicular data exchanges. While this already resulted in plentiful joint publications in the previous phases of the CRC, future methods for resource prediction, as stated in WP2 and WP6, will also flow directly into the toolkit of B4. Furthermore, approaches for reliable communication in highly mobile network topologies, which are major topics of project B4 for the outdoor case, will be adapted and applied to logistical indoor environments in A4. This concerns in particular UAVs and other mobile entities within industrial facilities.

In addition, methods for anticipatory networking and resource prediction, such as C3ACE, will be leveraged jointly with project B2 to enable large-scale deployments of PAMONO sensors for quality checks in developing countries with limited network infrastructure. However, high-tech production lines of active ingredients will also benefit from improved scalability in case of short-term expansions due to pandemics.

Addressing the opposite direction, project A1 will continue to provide topical resource-efficient methods of machine learning for a data-driven model synthesis, which is required for anticipatory and altruistic wireless communication. Especially in WP4 and WP5, multimodal information must be combined into efficient models for positioning, load prediction, and selection of the communication path. A key challenge is to find a minimum subset of input features to satisfy resource constraints and model accuracy in online applications.

In collaboration with A3 we will extend our recent achievements in machine learning-based data rate and resource prediction for cellular communication systems to dynamically changing environments. This will introduce online learning into the training process to dynamically adapt to the current network coverage and its particular configuration. As the relationship of features and targets may change drastically over time, which is typically known as concept drift, we will incorporate Model-Based Optimisation with Concept Drift (MBO-CD) into the learning process to obtain a continuously optimised predictive model.

Furthermore, in cooperation with A6, the models and systems developed by A4 are used in a cooperative evaluation of the spatio-temporal graph methods. One step of the evaluation is to create a graph model for the mobile entities in an industrial facility to prevent possible collisions and to provide insights for the motion planner to actively avoid collisions. The time span required for the predictions is evaluated against the reliability of the feedback given to the motion planner of the mobile platform. The mobile systems and the interface to those systems with their models are provided by A4 for data intensive graph model computation in A6.

Finally, the tight relationships of the A4 project with industrial use cases and continuous field validations ensure a sustainable transfer of fundamental research into practical applications, which emphasises the economical benefits of this project together with the whole CRC.

## 3.6 Differentiation from other funded projects

### **Innovations Lab (Innovationslabor)**

**(ten Hompel, Reference number KIT PTKA 02P16Z200) (Funding period: 2016–2019)**

Innovations Lab - hybrid services in logistics: Hybrid service models and new forms of human-machine interaction are explored by developing a research facility with state of the art equipment and interconnecting them. The main focus of Innovations Lab is to build and provide the platforms that are required for interdisciplinary logistics research, whereas this sub project in CRC876 focuses on using specific systems for developing resource constrained models and methods.

### **ConnectedFactories**

**(ten Hompel, Reference number EU 723777) (Funding period: 2016–2019)**

The ConnectedFactories project establishes a structured overview of available and upcoming technological approaches and best practices. The project identifies present and future needs, as well as challenges, of the manufacturing industries. In one hand the main focus is on industrial applications and on the other hand it does not consider constraints which are addressed in CRC876.

### **Safelog**

**(ten Hompel, Reference number EU 688117) (Funding period: 2016–2019)**

Focus of the Safelog project is on safe human-robot interaction in logistic applications for highly flexible warehouses. In particular, it addresses the safety aspects of Automated Guided Vehicles (AGVs) in environments with human interaction.

### **SENSE**

**(ten Hompel, Reference number EU 769967) (Funding period: 2017–2020)**

The SENSE project's strategic objective is to accelerate the path towards the Physical Internet (PI) and enable well-functioning advanced pilot implementations of the PI concept. These implementations will be extended in industry practice by 2030 and contribute to at least 30% reduction in congestion, emissions, and energy consumption.

### **Clusters 2.0**

**(ten Hompel, Reference number EU 723265) (Funding period: 2017–2020)**

In this project, the research is focusing on the development of a network of hyper connected logistics clusters and their influence areas. It mainly focuses on multimodal transportation logistics.

### **LEGOLAS**

**(ten Hompel, Reference number PT ETN IT-1-2-014a) (Funding period: 2017–2020)**

The aim of the LEGOLAS project is to plan assistance systems for modular Industry 4.0 plants in the chemical process industry.

### **InDaSpacePlus**

**(ten Hompel, Reference number BMBF 01IS17031) (Funding period: 2017–2020)**

As a virtual data room, the Industrial Data Space supports the secure exchange and easy linking of data within business ecosystems based on standards and community governance models. Data analysis as the focus of CRC876 is not considered as an aspect of this project.

### **L4MS**

**(ten Hompel, Reference number EU 767642) (Funding period: 2017–2021)**

The ambition of L4MS (Logistics for Manufacturing SMEs) is to reduce the installation cost and time of mobile robots by a factor of 10. Neither machine learning methodology nor constraints from CRC876 are addressed within this project.

**DFG-Forschergruppe 1511**

**(Wietfeld, Reference number Wi 3751/1-1, Wi 3751/2-1) (Funding period: 2014–2018)**

This project addresses innovative wide-area control applications and protection of electrical energy systems, in particular, to avoid large-scale system failures (blackouts). This includes an evaluation and optimisation of the hybrid simulation environment for power grids and communication networks in conjunction with Software-Defined-Networking (SDN).

**BERCOM**

**(Wietfeld, Reference number BMBF 13N13741) (Funding period: 2015–2018)**

BERCOM focuses on hardening and extending LTE for the use in shared critical infrastructure communication networks, e.g., for Smart Grids. It includes a validation of the overall system's increased robustness and performance via a physical demonstrator.

**OPUS**

**(Wietfeld, Reference number EFRE-0800885) (Funding period: 2017–2020)**

OPUS (Optimised Predictive Performance Using Cyber-Physical Systems) intends the development of methods and technologies that enable a predictive maintenance for future generations of permanently installed CPS, such as pumps or engines. Since such installations are widely distributed and frequently located in dead spots, e.g., in basements, a major challenge is to provide methods for reliable wide-area wireless communication in such applications.

**LARUS**

**(Wietfeld, Reference number BMBF 13N14133) (Funding period: 2017–2019)**

LARUS works on the development of an unmanned aerial support system for maritime search and rescue missions. The focus areas are robust long-range communication, aerial base stations, radio-based localisation, and communication-aware mission control.

**IDEAL**

**(Wietfeld, Reference number BMWi 03ET7557A) (Funding period: 2016–2019)**

The aim of the project IDEAL (Impedance Controller and Decentralised Congestion Management for Autonomous Power Flow) is the development of a reactive congestion management system for high and medium voltage networks. However, the included long-range communication is not subject to the resource constraints that are addressed in the A4 project.

**CPS.HUB/NRW**

**(Wietfeld, Reference number EFRE-0400008) (Funding period: 2015–2018)**

CPS.HUB/NRW concentrates the competence and knowledge of all disciplines relevant to the development of cyber-physical systems. The innovation ecosystem provided by CPS.HUB/NRW enables regional actors to benefit from broad CPS-relevant knowledge in order to continuously refine their processes and adapt to new developments.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	2	129,000	2	129,000	2	129,000	2	129,000
Total	—	129,000	—	129,000	—	129,000	—	129,000
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	129,000		129,000		129,000		129,000	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Christian Wietfeld, Prof. Dr., professor	Communication networks	TU Dortmund	4	—	Existing funds
	2	Michael ten Hompel, Prof. Dr., professor	Logistics	TU Dortmund	4	—	Existing funds
	3	Dominik Borst, M.Sc., doctoral researcher	Logistics	TU Dortmund	11.94	—	Existing funds
	4	Karsten Heimann, M.Sc., doctoral researcher	Communication networks	TU Dortmund	9.9	—	Existing funds
	5	Pascal Jörke, M.Sc., doctoral researcher	Communication networks	TU Dortmund	9.9	—	Existing funds
	6	Moritz Roidl, M.Sc., doctoral researcher	Logistics	TU Dortmund	15.92	—	Existing funds
	7	N.N., student assistant	Communication networks	TU Dortmund	10	—	Existing funds
	8	N.N., student assistant	Logistics	TU Dortmund	10	—	Existing funds
Non-research staff	9	Matthias Foese, technical staff	—	TU Dortmund	6	—	Existing funds
	10	Uwe Sondhoff, technical staff	—	TU Dortmund	6	—	Existing funds
<b>Requested staff</b>							
Research staff	11	Robert Falkenberg, M.Sc., doctoral researcher	Communication networks	TU Dortmund	—	Doctoral researcher	—
	12	Aswin Karthik Ramachandran Venkatapathy, M.Sc., doctoral researcher	Logistics	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):****1. Wietfeld, Christian**

Project management. Focus on Communication Networks. Contribution to work packages 2, 4, 5, and 6.

**2. ten Hompel, Michael**

Project management. Focus on heterogeneous logistics platforms. Contribution to work packages 1, 2, 3, 4, and 6.

**3. Borst, Dominik**

Research and development within work packages 2, 3, 4, and 6.

**4. Heimann, Karsten**

Integration of mmWave communication into ultra-dense IoT environments especially in work packages 2, 5, and 6.

**5. Jörke, Pascal**

Research and development of low-power wide-area communication methods for IoT networks, especially within work packages 2, 4, 5, and 6.

**6. Roidl, Moritz**

Research and development of decentralised control of material flow using multi-agent systems within work packages 1, 2, 3, 4, 5, and 6.

**7. N.N.**

Support of implementation within work packages 2, 4, 5, and 6.

**8. N.N.**

Support of implementation within work packages 1, 2, 3, 4, 5, and 6.

**9. Foese, Matthias**

Support with regard to the technical infrastructure.

**10. Sondhoff, Uwe**

For maintenance of technical infrastructure of PhyNetLab and other platforms.

**Job descriptions of staff for the proposed funding period (requested funds):****11. Falkenberg, Robert**

Research and development of novel methods for improvement of the scalability of communication systems in work packages 2, 4, 5, and 6.

**12. Ramachandran Venkatapathy, Aswin Karthik**

Research and development of interoperating network strategies for Industry 4.0 systems in work packages 1, 2, 4, 5, and 6.

**3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from University	8,000	8,000	8,000	8,000
Sum of existing funds	8,000	8,000	8,000	8,000
Sum of requested funds	0	0	0	0

(All figures in euros)

**3.7.5 Requested funding for instrumentation for the new funding period**

This project does not request any funding for major research instrumentation.



### 3.1 General information about Project A6

### 3.1.1 Project title:

Resource-efficient Graph Mining

### 3.1.2 Research area(s):

409-01 (Theoretical Computer Science), 409-06 (Process and Knowledge Management)

### 3.1.3 Principal investigator(s)

Kriege, Nils M., Dr., 09.03.1983, German

LS 11, Fakultät für Informatik, Technische Universität Dortmund  
Otto-Hahn-Straße 14  
44227 Dortmund

Phone: 0231-755-7737  
E-mail: [nils.krieger@tu-dortmund.de](mailto:nils.krieger@tu-dortmund.de)

Mutzel, Petra, Prof. Dr., 13.06.1964, German

LS 11, Fakultät für Informatik, Technische Universität Dortmund  
Otto-Hahn-Straße 14  
44227 Dortmund

Phone: 0231-755-7700  
E-mail: petra.mutzel@tu-dortmund.de

Weichert, Frank, Dr., 07.09.1967, German

LS 7, Fakultät für Informatik, Technische Universität Dortmund  
Otto-Hahn-Straße 16  
44227 Dortmund

Phone: 0231-755-6122  
E-mail: [frank.weichert@tu-dortmund.de](mailto:frank.weichert@tu-dortmund.de)

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

( ) no (x) yes

Dr. Nils Krieger

End date of fixed-term contract: 30.09.2020

Further employment is planned until 30.09.2023.

Dr. Frank Weichert

End date of fixed-term contract: 31.05.2021 (*There is an ongoing application for a permanent position*)

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes    (x) no
2.	clinical trials	( ) yes    (x) no
3.	experiments involving vertebrates.	( ) yes    (x) no
4.	experiments involving recombinant DNA.	( ) yes    (x) no
5.	research involving human embryonic stem cells.	( ) yes    (x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes    (x) no

## 3.2 Summary

The internet of things (IoT) has already started to generate huge amounts of data. Infrastructures, machines, vehicles, and everyday objects such as smartphones or TVs are equipped with intelligent functions that are linked to each other. These objects contain sensors, RFID chips, and cameras that continuously produce data and communicate within these cyber-physical systems (CPSs). A natural representation of a linked data set is provided by a graph, where entities are represented as vertices and their relationships are encoded by edges. Compared to the classical representation of objects as feature vectors, the graph structure additionally allows the representation of the complex relationships between these objects. Project A6 deals with the development of new methods for analysing graphs at a large scale or on a large number of graphs in resource constrained environments.

Our aim is to combine methods from algorithm theory and algorithm engineering with (deep) learning methods from graph mining. The project has started in phase 2 of the Collaborative Research Centre (CRC) 876 with the PIs Kristian Kersting, Petra Mutzel, and Christian Sohler. In 2017, Kristian Kersting moved to a W3 position at the University of Darmstadt. Shortly afterwards Nils Kriege stepped in as project investigator. In phase 2, the project has laid foundations for graph learning. In particular, we have successfully developed new algorithms based on randomised sampling for extracting information out of graphs in a resource-efficient manner.

In the next phase, we would like to bring some of our results into real applications and approach CPSs. In order to strengthen the CPS part, Christian Sohler will leave the project and Frank Weichert will join us. Moreover, we want to focus on *feature learning* techniques for graphs; i.e., our new methods are not based on a predetermined set of features anymore, but learning from the features will be part of the problem. To this end, we want to build on our results in phase 2 on efficient graph kernels and extend them to feature learning. For example, we expand the research of A6 to *geometric deep learning*, which is an emerging field that extends deep learning techniques for Euclidean domains to graph-structured data. In particular, we would like to apply randomised sampling techniques on problems related to graph kernels and also to geometric deep learning. With our attention towards CPSs, the two aspects *dynamic* and *(soft) real-time* are becoming essential. Therefore, we will study learning tasks on dynamic graphs such as sequences and streams of graphs. In order to integrate our new methods into CPSs we need our approaches to obey resource constraints regarding runtime, memory, accuracy, energy, transmission speed and number of labelled data. We will evaluate our methods on specific systems and domains that are relevant in the CRC 876 such as logistic sensor-actuator networks (A4), traffic forecasting (B4), and high-frequent irregularly structured data analysis (C3/C5).

### 3.3 Project progress to date

With the increasing presence of CPSs, linked data and networks increasingly appear, which can be modelled as graphs with or without additional attributes. One example of problems arising in this context is the classification of critical situations in logistic sensor-actuator networks (driverless transport systems (DTSs), unmanned aerial vehicles (UAVs), humans), which is of interest in project A4. Another example is traffic analysis (B4) or concept drift detection in graph streams (A3). Since these networks often are either complex or large or the set of graphs is big, new approaches are needed for the analysis. The area of graph mining is dealing with data mining on graphs and studies learning tasks such as classification, cluster analysis and regression. However, because of the memory assumption or runtime requirements of these graph mining algorithms, only few of them are able to work for problems related to CPSs.

Therefore, during the last three years, we have developed new learning approaches and algorithms for learning tasks on graphs with resource constraints such as runtime or memory consumption. We mainly focused on (multi-label) classification tasks. For this we have combined ideas for randomised algorithms and graph algorithms from the area of algorithm theory and algorithm engineering with learning approaches like graph kernels. We have theoretically analysed as well as realised and experimentally evaluated our new methods, also jointly with projects A1, B2, B4, and C1.

Project A6 started in phase 2 with the PIs Kristian Kersting, Petra Mutzel, and Christian Sohler. In 2017, Kristian Kersting left TU Dortmund University, since he got a W3 professorship at TU Darmstadt. Shortly afterwards, Nils Kriege joined the project A6 as PI. He already finished his dissertation with the CRC 876 in project A6 and is successfully publishing in the area of graph kernels and graph similarity. One of the main goals of the second phase has been to establish connections between graph kernels and randomised sublinear algorithms. This has been the main expertise of Christian Sohler. Since we have successfully established such a connection [A6/179, A6/172], Christian Sohler will resign being PI for project A6; however, we will still cooperate with him via project A2. For phase 3, we want to develop novel learning approaches on graph based data, especially taking into account their deployment on resource constrained real-world applications. To this end, Frank Weichert will join A6 in order to strengthen the CPS part and introduce the concept of geometric deep learning in cooperation with project B2. Project A6 will combine the competences of Frank Weichert with those of Nils Kriege in graph kernels and Petra Mutzel in graph algorithms in order to get new resource-efficient approaches that are both theoretically founded (with guarantees) and practically relevant for CPSs.

#### 3.3.1 Report and current state of research

In the following, a description of the main results achieved in phase 2 is given, followed by a description of the current state of research and the new challenges for the next funding period.

##### Report on phase 2

We have successfully published our research of the last funding period in top-tier (Core-Ranking A\*) computer vision, machine learning and data mining conferences, e.g., AAAI, CVRP, ICDM, IJCAI, KDD, NIPS, and other renowned conferences and journals (e.g., Machine Learning). Several of our manuscripts are currently under review. We have supervised a total of 16 bachelor and master theses closely related to project A6. Nils Kriege (now PI) has finished his PhD as part of the project A6. For a complete list of our publications, see <https://sfb876.tu-dortmund.de/>

[SPP/sfb876-a6.html](#). Due to DFG regulations and space restrictions, we focus the report on our main results relevant for the next funding period and describe their relation to the state of research.

**Graph kernels using neighbourhood aggregation.** A graph kernel is a similarity measure between graphs, which can be represented as a dot product between feature vectors obtained from the graphs [146]. Graph kernels can be used with established learning algorithms such as support vector machines [163] and have proven to be a key technique for solving classification tasks on graphs. We have developed graph kernels using different techniques to associate each vertex with additional information derived from its neighbours (*neighbourhood aggregation*). The most prominent approach to this is the so-called weisfeiler-lehman refinement (WL), which has previously been used to derive graph kernels [165]. The method successively refines the initial vertex labels of a graph as follows. For each vertex the label is replaced by the tuple consisting of its current label and the sorted sequence of labels of its neighbours. This tuple is then mapped to a new label typically represented by an integer. The procedure is iterated for a fixed number of steps or until the number of different labels remains unchanged.

We have proposed *propagation kernels* [A6/173] that associate with each vertex a distribution which is iteratively updated by a propagation scheme [A6/173]. The graph kernel then monitors how information spreads through a set of given graphs. This kernel, just like most other state of the art graph kernels, only takes local graph properties into account, e.g., small substructures. In [A6/175] we developed graph kernels that consider both local and global graph properties using the  $k$ -dimensional generalisation of Weisfeiler-Lehman refinement ( $k$ -WL). The  $k$ -WL defines a labelling function on vertex subsets of cardinality  $k$ . The neighbourhood of a  $k$ -cardinality set  $t$  are the  $k$ -cardinality sets that can be obtained from  $t$  by replacing exactly one vertex. The  $k$ -WL is inherently global, since it labels a set of vertices  $t$  by considering sets whose vertices are not connected to the vertices of  $t$ . Moreover, it does not benefit from the sparsity of the underlying graphs. Our new kernel called the *local  $k$ -WL* considers local neighbourhoods, which are the subsets of the  $k$ -WL neighbourhood containing only the elements having at least one adjacent vertex. Since the new graph kernel may still not scale well on large graphs, we have also devised a stochastic version of the kernel with provable approximation guarantees (using conditional Rademacher averages) for efficient computation. We support our theoretical results with experiments on several graph classification benchmarks, showing that our kernels often outperform the state of the art in terms of classification accuracies.

In collaboration with C1 we have developed a similarity measure based on Weisfeiler-Lehman refinement tailored to the properties of protein complexes. The similarity measure is a convex combination of Jaccard coefficients applied for the individual refinement steps. We have shown empirically that this similarity is in good agreement with edit similarity, a similarity measure derived from graph edit distance, but which can be computed much more efficiently.

**Resource-efficient graph kernels via explicit feature maps.** Support vector machines operate implicitly in the possibly high-dimensional feature space defined by a kernel (*kernel trick*), but are too slow for large-scale data sets. However, linear support vector machines, which use feature vectors instead of kernel functions, can be trained in time linear in the number of training examples [151]. Therefore, we have studied explicit feature maps for graph kernels.

In addition to the advantage of feature vectors mentioned above, known kernels using the kernel trick do not scale to graphs with thousands of vertices, and efficient kernels are typically limited to graphs with discrete labels. To overcome this, we have developed the hash graph kernel framework [A6/176], which turns continuous attributes into discrete labels using randomised hash functions. We have studied the theoretical requirements to guarantee well-defined similarity measures implicitly operating on the annotation. We have employed the Weisfeiler-Lehman subtree kernel [165] and the shortest-path kernel in our framework. An experimental study showed that

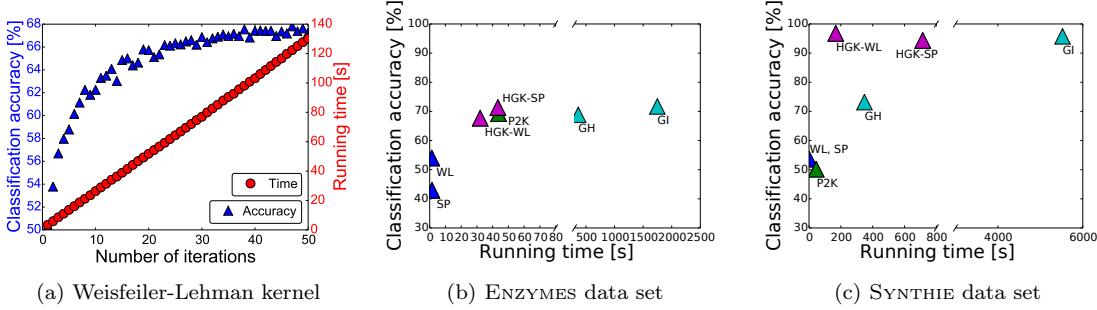


Figure 3.1: Empirical results for the hash graph kernel framework [A6/176]. (a) The influence of the number of hashing iterations on the runtime and classification accuracy using a Weisfeiler-Lehman base kernel. (b), (c) Comparison on two different data sets to other state of the art kernels (GH – GraphHopper, GI – Graph invariant, P2K – Propagation kernel). Our approaches (HGK-SP and HGK-WL) are located in the upper left area, which indicates high accuracy combined with high efficiency.

these achieve state of the art classification accuracies while being orders of magnitude faster than kernels that were specifically designed to handle continuous information, cf. Figure 3.1.

Currently known approximation techniques for explicit feature maps of kernels for continuous vectorial data cannot be used to obtain approximation guarantees in the hash graph kernel framework [A6/176]. Therefore, in a different line of work, we developed explicit feature maps with the goal to lift the known approximation results for kernels on continuous data to kernels for graphs with continuous vertex annotations. We have investigated how general convolution kernels [150] are composed from base kernels and how corresponding feature maps can be constructed. This enabled us to derive approximative, explicit feature maps with convergence guarantees for state of the art kernels supporting real-valued attributes. We have analysed for which kernels and graph properties computation by explicit feature maps is feasible and actually more efficient than functional computation. In extensive experiments, we showed that our approach often achieves a classification accuracy close to the exact methods based on the kernel trick, but requires only a fraction of their runtime. Moreover, we investigated the benefits of employing the kernel trick when the number of features used by a kernel is very large. For random walk and subgraph matching kernels (which we introduced already at *ICML 2012*) we confirmed our theoretical results experimentally by observing a phase transition when comparing runtime with respect to label diversity, walk lengths, and subgraph size, respectively.

The significance of explicit graph feature maps has been confirmed in a recent experimental study [158]. The results show that explicit feature maps are often preferable under strict budget constraints when working with streams of graphs, a setting that is highly relevant for the application in CPSs.

**Graph kernels from optimal assignments.** A leading design paradigm for graph kernels is the convolution kernel, which decomposes graphs into their parts and sums over all pairs of these parts [150]. Assignment kernels, in contrast, are obtained from an optimal bijection between parts and can yield more expressive kernels. Fröhlich et al. [145] first proposed kernels for graphs derived from an optimal assignment between their vertices, where vertex attributes are compared by a base kernel. However, it was shown that the resulting similarity measure is not necessarily a valid kernel [168], which complicates their use in kernel methods. We studied which base kernels lead to valid assignment kernels [A6/178]. To this end, we characterised a specific class of kernels and showed that it is equivalent to the kernels obtained from a hierarchical partition of their domain. When such kernels are used as base kernel the optimal assignment (i) yields a valid

kernel and (ii) can be computed in linear time by histogram intersection given the hierarchy. We have demonstrated the versatility of our results by deriving novel graph kernels based on optimal assignments, which are shown to improve over their convolution-based counterparts w.r.t. classification accuracy. In particular, we proposed the Weisfeiler-Lehman optimal assignment kernel, which performs favourably compared to state of the art graph kernels on a wide range of data sets.

We have presented a summary of some of our advances in kernel-based graph classification including the results from our publications [A6/176, A6/178] at *ECML PKDD 2017* as part of the Nectar Track. For the experimental evaluation of graph kernels, we have established a repository of benchmark data sets publicly available at <https://sfb876.tu-dortmund.de/SPP/sfb876-a6.html>.

**Analysis of the expressivity of graph kernels.** The expressivity of graph kernels is assessed almost exclusively from experimental studies and there is no theoretical justification why one kernel is in general preferable over another. Therefore, we have introduced a theoretical framework for investigating the expressive power of graph kernels, which is inspired by concepts from the area of property testing [A6/179]. We say that a kernel distinguishes a graph property if there is a constant angle between all graphs that have the property and all graphs that differ in more than  $\epsilon \cdot d \cdot n$  edges from every graph that has the property ( $\epsilon$  is considered to be a constant,  $d$  denotes the maximum degree). The idea is that this is a condition we would typically require from a kernel: A graph kernel should clearly separate between graphs from our class of interest and those graphs that are significantly different from that class, and the separation should not be dependent on the size of the graph. In general, we consider a kernel to be more powerful if it distinguishes more properties. We have studied the expressiveness of known established kernels and showed that, surprisingly, many of them do not distinguish basic graph properties such as connectivity. We then developed a new kernel that distinguishes many basic graph properties in bounded degree graphs. We believe that the definition of expressivity can be used to give theoretical indications of the power of graph kernels.

**Graph kernels for text mining.** Jointly with project A1 we have applied graph kernel methods for text mining. Erdmann et al. have brought together methods from machine learning, journalism studies, and statistics to help bridge the segmented data of the international public sphere, using the Transatlantic Trade and Investment Partnership (TTIP) as a case study (*ICML Workshop on #Data4Good 2016*). In a master thesis, we have developed probabilistic theme models for which the themes are weighted graphs. Jointly with journalists we have studied models concerning the temporal attention of themes using epidemiological models. The evaluation has shown that the representation of texts has influence on the text mining results. This work is a nice example of transfer of our results to completely different scientific communities.

**Approximating the spectrum of a graph.** We developed a randomised algorithm to estimate the spectrum (the set of eigenvalues of the adjacency matrix) of an undirected graph [A6/172]. Our algorithm assumes query access to the graph that allows to pick vertices uniformly at random and for a given vertex to pick a neighbour uniformly at random in constant time. The algorithm requires  $2^{O(1/\epsilon)}$  time and returns a compact description of the spectrum with error  $\epsilon n$  with respect to the  $l_1$ -distance of the vectors of eigenvalues sorted by size. We empirically evaluated our algorithm and approximated spectra for different classes of graphs (road networks, co-citation graphs and social networks). Even for graphs with a few million vertices and more than 100 million edges our approach required only a few minutes on standard hardware. Our most interesting finding is that different classes of graphs seem to exhibit different spectra; i.e., the spectra of road networks look different from the spectra of co-citation graphs and the spectra of social networks. We therefore believe that the approximate spectrum can be used as a powerful representation for very large graphs.

**Counting problems and probabilistic inference.** We have investigated whether “Compilation to Graph Databases” could be a practical technique for scaling lifted probabilistic inference and learning methods. Over the past decade, exploiting relations and symmetries within probabilistic models has been proven to be surprisingly effective at solving large-scale data mining problems. One of the key operations inside these lifted approaches is counting – be it for parameter/structure learning or for efficient inference. Since the counts are primarily used in exponents of different functions, computing exact counts is not always necessary, especially when the counts are large. In recent work (*SDM 2016*), we suggest to compiling the logical model (e.g., Markov Logic networks) to a graph represented in resource description framework (RDF) format. This equivalent model allows for both approximate and exact counts to be performed in a fraction of the time required by the original logical model. We have demonstrated that the proposed approach achieves reasonable speed-ups for both inference and learning, without sacrificing accuracy. We proposed *Poisson Sum-Product Networks* for modelling multivariate count data, which allow for fast and tractable inference [B4/A6/174].

The distance distribution of a network counts the number of shortest paths for each length and captures important properties like, e.g., the dynamics of network spreading processes. We have developed a closed-form distribution by applying maximum entropy arguments to derive a general, physically plausible model of path length histograms [A6/171]. Based on the model, we then established the generalised Gamma as a three-parameter distribution for shortest-path distance in strongly connected undirected networks. Extensive experiments confirmed our theoretical results and revealed that the generalised Gamma distribution accounts well for distance distributions of synthesised Erdős-Renyi, Barabasi-Albert, power law, and LogNormal graphs as well as some classes of real-world networks. We studied the attention dynamics of viral videos from the point of view of mathematical epidemiology (*Foundations and Trends in Web Science*). We have introduced a novel probabilistic model of the progression of infective diseases and used it to analyse time series of YouTube view counts and Google searches. Our results on a data set of almost 800 videos show that their attention dynamics are indeed well accounted for by our epidemic model. In particular, we find that the vast majority of videos considered in this study show very high infection rates.

### Current state of research and outlook

We review the related work on feature learning for (dynamic) graphs and give an outlook on the issues that are relevant for our project plan in phase 3.

**Multiple and deep kernel learning.** Kernel methods such as support vector machines are considered traditional “shallow” machine learning techniques compared to deep learning. However, *multiple kernel learning* (MKL) allows learning of linear combinations of multiple base kernels, which may be organised in a hierarchical way according to their level of abstraction [144]. Such methods may therefore be referred to as deep kernels. Recent progress in multiple kernel learning allows to efficiently combine thousands of base kernels with robust  $\ell_p$ -norm mixtures [155]. Motivated by the success of feed-forward neural networks, deep kernel learning for Gaussian processes has been introduced [169]. Deep graph kernels have been proposed in [170], which build on known state of the art kernels, but allow relationships between their features to be respected. This is demonstrated by hand-designed matrices encoding the similarities between features for selected graph kernels.

**Geometric deep learning.** Convolutional Neural Networks (CNNs) are a successful feature learning technique that has led to qualitative breakthroughs on a wide variety of tasks on grid-based structured data. Therefore, the generalisation of convolutional neural networks (CNNs) on non-Euclidean domains, e.g., graphs or manifolds, has recently drawn significant attention [141], as such kinds of data arise in numerous applications. However, basic operations like convolution, which are taken for granted in the Euclidean case, are not even well defined on non-Euclidean

domains. Recent works in this field were brought together under the umbrella term *geometric deep learning* [141] and came up with a set of methods that aim to expand the convolution operator in deep neural networks to handle irregularly structured input data by preserving its main properties: local connectivity, weight sharing, and shift invariance. These methods can loosely be divided into two different subsets: the *spectral* and the *spatial filtering* approaches [141].

Spectral filtering is based on spectral graph theory, where the eigenvectors of the graph's Laplacian matrix are interpreted as Fourier bases. Convolution in the graph domain can thus be represented by multiplication in the Fourier domain. Defferrard et al. [143] proposed a spectral convolution method that avoids the computation of the Laplacian eigenvectors by approximating spectral filters. The number of trainable parameters  $K \in \mathbb{N}$  then corresponds to the range of aggregated local  $K$ -neighbourhoods. Holding  $K = 1$ , Kipf and Welling [153] further simplified this approximation by only considering one-neighbourhoods.

Spatial filtering approaches define convolution directly on groups of spatially close neighbours. Here, we introduced spline-based convolutional neural networks (SplineCNN) [B2/A6/177] with continuous kernel functions based on B-splines. The B-spline formulation allows for a computation time that is independent from the kernel size due to the local support property of the B-spline basis functions. We already validated our convolution operation by addressing the real-world application of classifying cuneiform signs and compared it to the graph edit distance in an extensive experimental evaluation regarding runtime and prediction accuracy [A6/B2/180].

Pooling concepts in the context of graph neural networks were first introduced by Defferrard et al. [143], but have, however, not received much attention since then. A pooling operation requires meaningful neighbourhoods on graphs, where similar vertices are clustered together and preserve local geometric structures. For the *graph coarsening* algorithm, the authors suggest using the greedy *Grclus* algorithm to compute successively coarser versions of a given graph that reduce its size by a factor of two for each level [143]. In the next phase, we plan to develop alternative graph pooling and coarsening techniques, e.g., based on ideas from approaches in graph drawing and graph algorithms.

Variational autoencoders (VAEs) are commonly used in the deep learning community to compress, denoise, or even generate regular structures (from sampled random noise) with the help of transposed convolution layers. generative adversarial networks (GANs) go one step further and attempt to train an image generator by simultaneously training a discriminator challenging it to improve. In contrast to traditional deep generative methods, graph generation is a difficult problem for methods based on gradient optimisation, as graphs are discrete objects after all. Kipf and Welling [154] sidestep these hurdles by letting the decoder output a probabilistic fully connected graph directly from latent space. However, this is only feasible for small graphs due to the loss of sparsity, as both runtime and memory requirements are quadratic in the number of vertices [154]. Johnson [152] and Li et al. [157] learn a variety of neural networks, which are then used to generate graphs by a sequence of decisions. This approach, however, lacks parallelism and cannot be trained in an end-to-end fashion. In the next phase we plan to build on these ideas and study deep *hierarchical* graph synthesis with the help of unpooling layers.

**Dynamics in networks.** Networks (bioinformatics, brain science, social networks, sensor actuator networks) often contain time-dependent information, either explicitly in the form of timestamps or implicitly, e.g., from a process that varies over time. Recently, Michail and Spirakis [159] have provided a survey concerning the theory of dynamic networks. They model a temporal graph as a special case of a labelled graph, where labels give some measure of time by providing the information about the temporal visibility of an edge. They point out that many graph properties and problems become much more difficult on these graphs. For example, important theorems in classical graph theory like Mengers theorem do not hold anymore. However, for the case of Mengers theorem, a natural reformulation was possible. The authors emphasise that the validity

of many other fundamental results of graph theory, like Kuratowski's planarity theorem, need to be checked for validity. Concerning machine learning, Morik [160] provided a discussion of different representations and learning tasks from time phenomena. For graphs, it is always possible to provide feature vectors and then to use any of the approaches for time series or general temporal data points. For example, Borgwardt et al. [140] have suggested a kernel-based approach to the classification of time series of gene expression profiles. Li et al. [156] introduced the *graph kernel tracking problem* in which the adjacency matrices of two time-evolving graphs and a sequence of updates are given. The task is to compute all graph kernels for all time steps. They provide a family of fast algorithms (Cheetah) for solving the problem and analyse its approximation error. Paaßen et al. [161] used graph kernels for predicting the next graph in a dynamically changing series of graphs with applications to intelligent tutoring systems.

recurrent neural networks (RNNs), especially using long short-term memory (LSTM) units or gated recurrent units (GRUs), are capable of exhibiting dynamic temporal behaviour by allowing cyclic connections between neural network units. In addition, many proposed works leverage a combination of CNNs and RNNs to exploit spatial as well as temporal regularities, e.g., for recognising objects in video sequences. On irregularly structured data, Seo et al. [164] proposed a first work of combining graph neural networks with temporal concepts, in which an LSTM or GRU was added to the network. However, this work only deals with time-varying vertex signals, and hence the geometric modification of the graph structure is not taken into account. We want to tackle this issue in the next phase. In addition, *attention mechanisms* have become almost a *de facto* standard in many sequence-based tasks, as they allow for dealing with variable-sized inputs by focusing only on the most relevant parts of the input to make decisions. The term *self-attention* is referred to if attention is used to compute a representation of a single sequence. Due to variable-sized inputs and neighbours, graph neural networks have been recently extended by self-respectively neighbourhood attentional layers and further improve on the current state of the art (e.g., [147, 167]).

**Graph based applications on Cyber-Physical Systems.** The IoT and industry 4.0 are increasingly based on CPSs, a functional cooperation of information technology with mechanical and electronical units. Sensors and actuators form complex *sensor-actuator networks*, which are dependent on each other and are therefore representable by graph based data structures [166]. With the innovation laboratory at TU Dortmund University (project A4), we have the opportunity to transfer our graph based algorithms to various scenarios of sensor-actuator networks. The research hall consists of heterogeneous systems such as DTSs, UAVs, and humans for the use in indoor logistics operations in which these systems interact with the environment and each other through time and space. Here, vertices represent stationary as well as spatio-temporal machines or human beings, and edges encode spatial as well as logical, or respectively logistical, relation. Figure 3.5 on page 144 shows a mapping of a real scenario to a derived graph based representation based on the localisation and communication techniques of project A4. Here the task can be to predict possible collisions in the near future in order to give the involved systems the time to react appropriately. As preparatory work, the student project group 615 was initiated, supervised by the PI Weichert and the Chair of the PI ten Hompel. The project group examines methods for safe and efficient human-machine cooperation in the research hall.

In addition, many optical sensors of CPSs perceive their environment irregularly, e.g., through point clouds. These irregular representations can be naturally transferred to a graph representation (e.g., by a  $K$ -NN graph). For example, the traffic analysis of project B4 is dedicated to the precise prognosis and optimisation of vehicular traffic flow. Here, environment information is gathered irregularly from vehicular sensor data. In contrast to the already existing collaboration of projects A1 and B4 in which spatio-temporal random fields (STRFs) [162] are used to forecast traffic flow, we want to tackle this problem based on spatio-temporal graph kernels and geometric deep learning techniques, where we omit a probabilistic modelling and the graph corresponds to the underlying data. We want our algorithms to be suitable for processing these kinds of data in a resource-efficient and highly accurate way.

**Bibliography**

- [138] V. Arvind, J. Köbler, G. Rattan, and O. Verbitsky. “On the Power of Color Refinement”. In: *Proceedings of the 20th International Symposium on Fundamentals of Computation Theory*. 2015, pp. 339–350 (cit. on p. 166).
- [139] G. Bartel, C. Gutwenger, K. Klein, and P. Mutzel. “An Experimental Evaluation of Multilevel Layout Methods”. In: *Graph Drawing*. Ed. by U. Brandes and S. Cornelsen. Vol. 6502. LNCS. Springer, 2010, pp. 80–91 (cit. on p. 169).
- [140] K. M. Borgwardt, S. V. N. Vishwanathan, and H.-P. Kriegel. “Class Prediction from Time Series Gene Expression Profiles Using Dynamical Systems Kernels”. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2006), pp. 547–58 (cit. on p. 161).
- [141] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric Deep Learning: Going beyond Euclidean data”. In: *IEEE Signal Processing Magazine*. 2017, pp. 18–42 (cit. on pp. 159, 160).
- [142] Y. Cheng, D. Wang, P. Zhou, and T. Zahng. “A Survey of Model Compression and Acceleration for Deep Neural Networks”. In: *ArXiv e-prints* abs/1710.09282 (2017) (cit. on p. 172).
- [143] M. Defferrard, X. Bresson, and P. Vandergheynst. “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2016, pp. 3844–3852 (cit. on pp. 160, 169).
- [144] M. Donini and F. Aiolfi. “Learning Deep Kernels in the Space of Dot Product Polynomials”. In: *Machine Learning* 106.9 (2017), pp. 1245–1269 (cit. on p. 159).
- [145] H. Fröhlich, J. K. Wegner, F. Sieker, and A. Zell. “Optimal Assignment Kernels for Attributed Molecular Graphs”. In: *Proceedings of the 22nd International Conference on Machine Learning (ICML)*. 2005, pp. 225–232 (cit. on p. 157).
- [146] T. Gärtner, T. Horváth, and S. Wrobel. “Graph Kernels”. In: *Encyclopedia of Machine Learning and Data Mining*. 2017, pp. 579–581 (cit. on p. 156).
- [147] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. “Neural Message Passing for Quantum Chemistry”. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 1263–1272 (cit. on p. 161).
- [148] G. Goranci, M. Henzinger, and P. Peng. “Improved Guarantees for Vertex Sparsification in Planar Graphs”. In: *25th Annual European Symposium on Algorithms (ESA)*. 2017, 44:1–44:14 (cit. on p. 169).
- [149] W. Grathwohl, D. Choi, Y. Wu, G. Roeder, and D. Duvenaud. “Backpropagation Through the Void: Optimizing Control Variates for Black-Box Gradient Estimation”. In: *International Conference on Learning Representations (ICLR)*. 2018 (cit. on p. 170).
- [150] D. Haussler. *Convolution Kernels on Discrete Structures*. Tech. rep. ICSC-CRL-99-10. University of California at Santa Cruz, 1999 (cit. on p. 157).
- [151] T. Joachims. “Training Linear SVMs in Linear Time”. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006, pp. 217–226 (cit. on p. 156).
- [152] D. D. Johnson. “Learning Graphical State Transitions”. In: *International Conference on Learning Representations (ICLR)*. 2018 (cit. on pp. 160, 170).
- [153] T. N. Kipf and M. Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *International Conference on Learning Representations (ICLR)*. 2017 (cit. on pp. 160, 172).

- [154] T. N. Kipf and M. Welling. “Variational Graph Auto-Encoders”. In: *NIPS Workshop on Bayesian Deep Learning* (2016) (cit. on pp. 160, 170).
- [155] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. “ $\ell_p$ -Norm Multiple Kernel Learning”. In: *Journal of Machine Learning Research* 12 (2011), pp. 953–997 (cit. on p. 159).
- [156] L. Li, H. Tong, Y. Xiao, and W. Fan. “Cheetah: Fast Graph Kernel Tracking on Dynamic Graphs”. In: *SIAM International Conference on Data Mining (SDM)*. 2015, pp. 280–288 (cit. on p. 161).
- [157] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, and P. Battaglia. “Learning Deep Generative Models of Graphs”. In: *International Conference on Learning Representations (ICLR Workshop)*. 2018 (cit. on pp. 160, 170).
- [158] G. D. S. Martino, N. Navarin, and A. Sperduti. “An Empirical Study on Budget-Aware Online Kernel Algorithms for Streams of Graphs”. In: *Neurocomputing* 216 (2016), pp. 163 –182 (cit. on pp. 157, 168).
- [159] O. Michail and P. G. Spirakis. “Elements of the Theory of Dynamic Networks”. In: *Communications of the ACM* 61.2 (2018), pp. 72–72 (cit. on p. 160).
- [160] K. Morik. “The Representation Race - Preprocessing for Handling Time Phenomena”. In: *Proceedings of the 11th European Conference on Machine Learning (ECML)*. Ed. by R. L. de Mántaras and E. Plaza. Vol. 1810. Lecture Notes in Artificial Intelligence. Berlin, Heidelberg, New York: Springer Verlag Berlin, 2000, pp. 4–19 (cit. on p. 161).
- [161] B. Paafßen, C. Göpfert, and B. Hammer. “Time Series Prediction for Graphs in Kernel and Dissimilarity Spaces”. In: *Neural Processing Letters* (2017), pp. 1–21 (cit. on pp. 161, 171).
- [162] N. Piatkowski, S. Lee, and K. Morik. “Spatio-temporal random fields: compressible representation and distributed estimation”. In: *Machine Learning* 93.1 (2013), pp. 115–139 (cit. on p. 161).
- [163] B. Schölkopf and A. J. Smola. *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. Cambridge MA: MIT Press, 2002 (cit. on p. 156).
- [164] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson. “Structured Sequence Modeling with Graph Convolutional Recurrent Networks”. In: *CoRR* abs/1612 . 07659 (2016) (cit. on p. 161).
- [165] N. Shervashidze, P. Schweitzer, E. van Leeuwen, K. Mehlhorn, and K. Borgwardt. “Weisfeiler–Lehman Graph Kernels”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2539–2561 (cit. on pp. 156, 166).
- [166] H. Song, D. B. Rawat, S. Jeschke, and C. Brecher. *Cyber-Physical Systems: Foundations, Principles and Applications*. Academic Press, 2016 (cit. on p. 161).
- [167] P. Veličković, G. Cucurull, A. Casanova, A. Romero, and Y. B. Liò. “Graph Attention Networks”. In: *International Conference on Learning Representations (ICLR)*. 2018 (cit. on pp. 161, 169, 172).
- [168] S. Vishwanathan, N. Schraudolph, R. Kondor, and K. Borgwardt. “Graph Kernels”. In: *Journal of Machine Learning Research* 11 (2010), pp. 1201–1242 (cit. on pp. 157, 168).
- [169] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. “Deep Kernel Learning”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. 2016, pp. 370–378 (cit. on p. 159).
- [170] P. Yanardag and S. V. N. Vishwanathan. “Deep Graph Kernels”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 1365–1374 (cit. on p. 159).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [A6/171] C. Bauckhage, **K. Kersting**, and **F. Hadji**. “Parameterizing the Distance Distribution of Undirected Networks”. In: *Proceedings of the 31th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by T. Heskes and M. Meila. AUAI, 2015 (cit. on p. 159).
- [A6/172] D. Cohen-Steiner, W. Kong, **C. Sohler**, and G. Valiant. “Approximating the Spectrum of a Graph”. In: *24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2018 (cit. on pp. 155, 158, 167).
- [A6/173] M. Neumann, R. Garnett, C. Bauckhage, and **K. Kersting**. “Propagation Kernels: Efficient Graph Kernels from Propagated Information”. In: *Machine Learning* 102.2 (Feb. 2016), pp. 209–245 (cit. on pp. 156, 166).
- [B4/A6/174] **A. Molina**, S. Natarajan, and **K. Kersting**. “Poisson Sum-Product Networks: A Deep Architecture for Tractable Multivariate Poisson Distributions”. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*. 2017, pp. 2357–2363 (cit. on pp. 159, 247).
- [A6/175] **C. Morris**, **K. Kersting**, and **P. Mutzel**. “Glocalized Weisfeiler-Lehman Graph Kernels: Global-Local Feature Maps of Graphs”. In: *IEEE International Conference on Data Mining (ICDM)*. 2017, pp. 327–336 (cit. on pp. 156, 167, 172).
- [A6/176] **C. Morris**, **N. Kriege**, **K. Kersting**, and **P. Mutzel**. “Faster Kernels for Graphs with Continuous Attributes via Hashing”. In: *IEEE International Conference on Data Mining (ICDM)*. 2016, pp. 1095–1100 (cit. on pp. 48, 156–158, 166, 167).
- [B2/A6/177] **M. Fey**, **J. E. Lenssen**, **F. Weichert**, and **H. Müller**. “SplineCNN: Fast Geometric Deep Learning with Continuous B-Spline Kernels”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on pp. 160, 166–168, 196).
- [A6/178] **N. Kriege**, P. Giscard, and R. C. Wilson. “On Valid Optimal Assignment Kernels and Applications to Graph Classification”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2016, pp. 1623–1631 (cit. on pp. 157, 158, 167, 168).
- [A6/179] **N. Kriege**, **C. Morris**, **A. Rey**, and **C. Sohler**. “A Property Testing Framework for the Theoretical Expressivity of Graph Kernels”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2018 (cit. on pp. 155, 158).
- [A6/B2/180] **N. Kriege**, **M. Fey**, D. Fisseler, **P. Mutzel**, and **F. Weichert**. “Recognizing Cuneiform Signs Using Graph Based Methods”. In: *International Workshop on Cost-Sensitive Learning (COST), SIAM International Conference on Data Mining (SDM)*. Proceedings of Machine Learning Research (PMLR). 2018 (cit. on pp. 48, 160).

### 3.4 Project plan

#### Goals

The overall goal of project A6 is to develop new approaches and algorithms for graph mining for learning tasks such as classification under resource constraints, e.g., memory, training and inference runtime, energy, transmission speed and the number of labelled data. The concrete objective for the next phase is to extend our work to feature learning. We will combine new randomised techniques from algorithmics with modern graph kernel approaches as well as geometric deep learning methods in order to get new resource-efficient approaches that are theoretically founded (e.g., with guarantees) and practically useful for, e.g., CPSs.

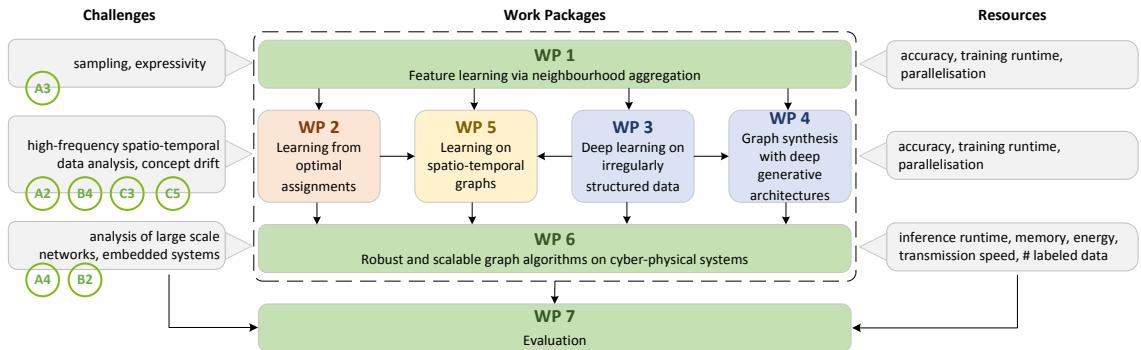


Figure 3.2: Overview of the work packages and their interdependencies. The colours indicate the lead within the work packages in relation to the PIs (■ Kriege, ■ Mutzel, ■ Weichert, ■ all). In addition, we provide the challenges within each work package and their cooperations with other projects (left) as well as their specific resource constraints (right).

#### Work schedule

The work packages, their interdependencies and the main responsibilities of the PIs are shown in figure 3.2. We distinguish two main approaches to achieve our goals: (i) the extension of established graph kernels to feature learning (WP 2) and (ii) the extension of successful feature learning techniques for Euclidean domains to graphs (geometric deep learning) (WP 3). While these two approaches appear to be different at first glance, they are both based on the principle of neighbourhood aggregation (WP 1). We further want to synthesise graphs based on gradient optimisation in WP 4. Learning tasks on spatio-temporal graphs will be considered in WP 5. While resource constraints will already be considered in WPs 1 to 5, the goal of WP 6 is the development of robust and scalable graph algorithms to satisfy the requirements of modern CPSs. We will evaluate our newly developed approaches on state of the art benchmark data and on real applications arising within our CRC 876 (WP 7).

**Work package 1. Feature learning via neighbourhood aggregation** A basic operation used in both, graph kernels and geometric deep learning is neighbourhood aggregation, which associates additional information with each vertex derived from the surrounding graph structure. In this work package we want to study and compare neighbourhood aggregation functions, in theory and practice, in order to derive new generalising techniques. These methods form the basis for the learning algorithms developed in WPs 2 to 5.

**Vertex refinement.** Weisfeiler-Lehman refinement compares vertices according to the degree of agreement between local vertex neighbourhoods [165]. For graph kernels various other approaches have been devised that follow the same spirit, e.g., [A6/173]. These methods can be controlled by a parameter reflecting the extent of the neighbourhood considered to distinguish vertices and can be used to successively refine vertex partitions. The results obtained as part of a bachelor thesis suggest that there is no refinement scheme that is generally preferable over the others, but their success depends on the data set. Therefore, we would ultimately like to *learn* a refinement scheme for a specific task. A unifying view is obtained by considering the set of all vertex partitions partially ordered by the finer-than relation. The techniques then lead to different walks on the associated lattice, and a similarity for two vertices can be determined by the number of steps they reside in the same partition. More generally, we may associate weights with the relations between partitions and measure the similarity by the length of the common walk. We will study the relation between different refinement processes and how they can be generalised. Moreover, we will investigate the effect of different weights and to what extent weights and refinement schemes can be learned in a problem-specific way. The discrete refinement procedures should also be extended to support continuous annotations like Euclidean coordinates. As a first step we will use hashing techniques derived from our previous work [A6/176]. Moreover, hierarchies generated by clustering the data points may be incorporated into the refinement lattice.

**Graph convolution operator.** While traditional Weisfeiler-Lehman refinement provides a fixed aggregation function, graph CNNs introduce additional trainable parameters, which are learned in an end-to-end fashion in order to better adapt to a given data distribution. Preliminary experimental results indicate that these approaches perform better than traditional kernel methods on a number of benchmark tasks. While the power of Weisfeiler-Lehman refinement is well understood [138], this does not apply for graph convolution operations, where trainable parameters and nonlinearities are introduced. Therefore, we want to study the expressive power of graph convolution operations in a systematic way. As a first approach, we want to show that these operations, without end-to-end optimisation, converge to the same vertex partition as the Weisfeiler-Lehman refinement. Subsequently, we want to derive a theoretical framework to investigate why they lead to better empirical results. In addition, we want to integrate trainable parameters to various Weisfeiler-Lehman variants such as the global  $k$ -dimensional WL and our new local  $k$ -WL, and study their effectiveness in a graph neural network setup.

Our spline-based convolution operator [B2/A6/177] weights neighbouring vertex features dependent on the edge attributes from the convolved vertex. Given an adjacency tensor  $\mathbf{A} \in \mathbb{R}^{n \times n \times d}$  with  $d$ -dimensional edge attributes, a feature matrix  $\mathbf{F} \in \mathbb{R}^{n \times m}$  with  $m$ -dimensional vertex features, and  $m$  continuous kernel functions  $\mathbf{g} = (g_1, \dots, g_m)$ , the convolution of SplineCNN over neighbouring features for a vertex  $v_i \in V$  is then defined by

$$(\mathbf{F} \star \mathbf{g})_i = \frac{1}{|\mathcal{N}(v_i)|} \sum_{l=1}^m \sum_{v_j \in \mathcal{N}(v_i)} f_{j,l} \cdot g_l(\mathbf{a}_{i,j,:}), \quad (3.1)$$

where  $\mathcal{N}(v_i)$  denotes the *set of neighbours* of  $v_i$ . For learning on 3D shapes, it is practical to encode the spatial relation between connected vertices in  $\mathbf{a}_{i,j}$ , whereas the edge attributes can be chosen quite freely for arbitrary graphs. We want to study the effectiveness of adding hand-crafted attributes to the edges, which encode local or global structural properties of a given graph. However, the computation of these attributes results in higher runtime and memory costs and increases the number of trainable parameters in each layer. Therefore, we want to evaluate the trade-offs between efficiency, performance, and generality for various hand-crafted edge attributes. Furthermore, we want to expand our operator by resource-efficient concepts, e.g., strided convolution or convolving only over a subset of neighbouring vertices. In addition, the resource-efficient extension of spatial convolution to local  $K$ -neighbourhood aggregation remains an open field of research that we want to deal with.

In addition to [B2/A6/177], we want to investigate the possibilities of developing our own *spectral* convolution operation based on our preliminary work [A6/172] in which we proposed a sublinear time algorithm to approximate the spectrum of a graph (with provable quality guarantees) in constant time. We want to efficiently use this approximation to filter a vertex signal in the spectral domain, given a number of trainable parameters. We expect such a convolution to be very efficient with reasonably low error in contrast to explicitly calculating its spectrum. Since approximations of spectral filters are already well established in the deep learning community, we want to study the advantages as well as limitations of our method.

**Sampled neighbourhood aggregation.** In order to obtain methods that work on large-scale networks, we want to combine the newly developed neighbourhood aggregation approaches with ideas from our randomised sampling techniques [A6/172, A6/175, A6/176] that still guarantee theoretical approximation factors on the quality. We will use these sampling approaches for feature learning on graph kernels as well as on graph CNNs. Here, the challenge is to randomly choose a given set of vertices that are valuable for the learning task. For example, the set can be chosen randomly with respect to certain vertex weights. Once the vertices have been chosen, their neighbourhoods can be aggregated using the methods described above, leading to very efficient approaches. It may even be beneficial to randomly sample the neighbourhoods in addition. However, the quality of generalisation needs still to be investigated in this scenario. Our approach for approximating the spectrum of a graph provides an example of our ideas [A6/172].

**Work package 2. Learning kernels from optimal assignments** Many intuitive similarity measures for graphs are obtained by assigning the vertices of one graph to the vertices of the other graph and then quantifying the resulting structural overlap. These techniques are often referred to as *graph matching* and include, e.g., measures from graph or subgraph isomorphism, maximum common subgraph isomorphism, and the graph edit distance. Although successfully employed for decades in pattern recognition, these methods suffer from three disadvantages: (i) They cannot directly be used with successful contemporary learning algorithms such as support vector machines or neural networks, since the maximum operation leads to an indefinite similarity function. (ii) The underlying graph problems are often NP-hard, such that their solution cannot be computed exactly for large graphs and data sets, in particular in resource constrained settings. (iii) They often depend on various parameters such as label cost functions, which are hand-crafted or determined by extensive grid search.

Using the ideas of graph matching, but overcoming its problems, we consider graph similarity functions of the following form. Let  $G$  and  $H$  be two graphs of the same order with vertex sets  $V$  and  $W$ , then the considered similarity is defined as

$$S_t(V, W) = \max_{B \in \mathfrak{B}(V, W)} W_t(B), \quad \text{where } W_t(B) = \sum_{(v, w) \in B} t(v, w), \quad (3.2)$$

$\mathfrak{B}$  denotes all possible bijections between  $V$  and  $W$  and  $t$  is a similarity defined on vertices. The vertex similarity  $t$  should, for example, consider two vertices as similar when they have a comparable neighbourhood, cf. WP 1, or when they play a similar role in the graph. The vertex similarity  $t$ , the assignment problem, and the classifier building on it are highly independent. We have already shown that, by restricting  $t$ , such that  $t(x, y) \geq \min\{t(x, z), t(z, y)\}$  for all vertices  $x, y, z$ , we can guarantee that  $S_t$  is a valid kernel and can, e.g., be plugged into support vector machines [A6/178]. We will investigate how more general vertex similarity functions can be incorporated into learning algorithms from optimal assignments and to what extent they can be learned end-to-end.

We discuss two starting points to demonstrate the feasibility of the ideas discussed above. First, we will consider a highly restricted variant building on the Weisfeiler-Lehman optimal assignment kernel [A6/178]. Here,  $t$  is realised by Weisfeiler-Lehman refinement leading to a hierarchy with uniform weights. With this approach, optimal assignments yield valid kernels and can be computed

in linear time by histogram intersection. Since the hierarchy fixes the optimal assignment, we can easily introduce weights to the hierarchy controlling the value of the assignment. These can be learned via multiple kernel learning by decomposing the histogram intersection kernel component-wise and finding an optimal linear combination as part of the classification problem. In order to support more general vertex similarity functions, we could follow the ideas of [168] to obtain a kernel that mimics the behaviour of the maximum function while guaranteeing positive semidefiniteness. To this end, we define

$$\tilde{S}_t(V, W; \beta) = \sum_{B \in \mathfrak{B}(X, Y)} e^{\beta W_t(B)}, \quad (3.3)$$

where  $\beta > 0$  is a parameter. Equation (3.3) is positive semidefinite and closely related to Equation (3.2). With increasing parameter  $\beta$  it approaches an exponentiated version of the optimal assignment similarity, since the sum is dominated by the largest addend [168]. The computation of Equation (3.3) requires determining the weight of all assignments instead of finding a single optimal one. We can approximate  $\tilde{S}_t$  by finding the top  $k$  solutions to the assignment problem using combinatorial algorithms.

The most general form allowing graph based representation learning in an end-to-end fashion is likely to be unfeasible for practical applications under resource constraints. Therefore, we will investigate whether the vertex similarity computation can be considered separately without sacrificing accuracy in order to guarantee efficiency. We will develop explicit approximative feature maps for the new assignment kernels. This is possible, for example, when the assignment kernel computation can be reduced to histogram intersection [A6/178]. We will study the trade-off between efficiency and accuracy in the context of graph streams under typical resource constraints in CPSs [158]. The techniques we develop can be used to derive heuristics for classical graph matching problems, e.g., for computing the graph edit distance, and we will systematically compare and evaluate their performance in terms of efficiency and accuracy, cf. WP 7.

**Work package 3. Deep learning on irregularly structured data** This work package deals with the development of deep learning algorithms for the field of irregularly structured data as a generalisation of the concepts known from traditional CNNs, while taking advantage of the sparsity of the input data. As a main objective we want to operate on arbitrary graphs as well as on graphs whose vertices are subject to spatial embedding, such as manifolds, meshes, or which are built up from point clouds. Figure 3.3 provides a comprehensive overview of the geometric deep learning methods discussed in the work packages WP 1, 3, and 4. In particular, the following aspects are of fundamental importance for irregularly structured data and should therefore be addressed in this work package:

- exploration of fast and parallelisable pooling concepts, and
- design of resource-efficient and permutation-invariant self-attention mechanisms.

Our specially developed convolution operator on graphs [B2/A6/177] in our ongoing cooperation with project B2 represents a solid starting point for studying this complex and still challenging research field. In terms of implementation, it needs to be ensured that all proposed methods work in a batch-wise fashion on different sized graphs, are GPU-parallelisable and expose the graph's sparse adjacency representation (cf. WP 6). As deep learning techniques must be very efficient and need to handle huge amounts of data, we want to further investigate the capacity, effectiveness and viability of our methods. As progress is made, we plan to release a GPU-focused, easy-to-use geometric deep learning framework to increase its applicability.

**Pooling layers.** CNNs are typically enriched by pooling layers to work in a hierarchical manner, in which the spatial resolution is progressively reduced. Thus, we want to develop analogous pooling concepts in the graph context, where each level produces a coarser graph which corresponds to the

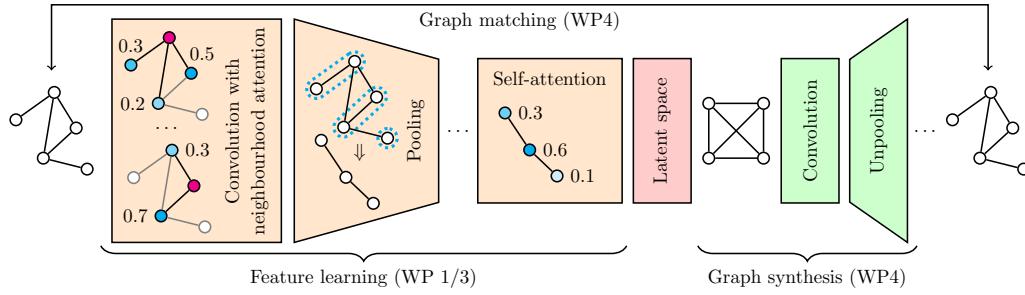


Figure 3.3: Concept graphic of the different graph neural network layers that are addressed in the work packages 1, 3, and 4. Graph convolution, pooling, and attention mechanisms are being used for feature learning (WP 1/3), whereas unpooling layers and graph matching form the fundamentals of graph synthesis (WP 4).

data domain seen at different resolutions. An already proposed pooling operation of graph neural networks only works on precomputed coarsened graphs that need to be permuted, extended by fake vertices, and kept in memory for the whole forward step duration [143]. In order to reduce the costs of preprocessing and allocation of unused memory, we want to examine *dynamic pooling* concepts that enable graph coarsening immediately during runtime. As a further advantage, we gain the ability to augment input graphs in pooling scenarios, something that is not yet possible without a huge amount of additional costs (cf. WP 6), and lay the foundation for hierarchical graph synthesis (cf. WP 4). In addition, we want to investigate alternative graph coarsening algorithms, which are faster and allow for stronger GPU parallelisation than previously used methods, and examine their effectiveness. A good starting point are graph coarsening techniques studied in the area of graph drawing (see, e.g., our work on multi-level layout methods [139]) or graph algorithms (see, e.g., recent research on vertex sparsification [148]).

**Attentional layers.** To handle variable-sized and permutation-invariant input more sensitively we want to design neighbourhood and self-attention mechanisms and explore their effectiveness in combination with graph convolution approaches. In contrast to traditional CNNs, where the inputs are typically cropped to a fixed-sized dimension and obtain a well-defined ordering, this does not apply for inputs such as graphs or manifolds. The design of attention mechanisms on graphs must be, however, invariant to permutations of the graph vertices in order for the neural network to be invariant to graph isomorphism, and therefore poses challenges. Already proposed attention mechanisms, although efficient in theory, underlie strong memory requirements and do not offer major GPU performance benefits in sparse matrix scenarios compared to CPUs, and hence do not scale well to large graphs [167]. Appropriately addressing this constraint is an important step towards resource-efficient applications. In addition, the design of these attentional layers in the graph context will allow us to tackle the extension to recurrent graph convolutional neural networks addressed in WP 5.

**Work package 4. Graph synthesis with deep generative architectures** This work package deals with the synthesis of arbitrary graphs from a fixed number of latent features or from similar examples using deep encoder-decoder or respectively deep generative architectures. The approaches being developed within this work package will allow for learning methods that can, e.g., compress, denoise, generate or modify graphs on the basis of gradient optimisation.

variational autoencoders (VAEs) and GANs define state of the art classes of neural networks used in unsupervised machine learning to solve compression or synthesis tasks on grid-like structured data. Due to their success, the generalisation of these methods to irregularly structured data is a

promising open research issue. Although graph synthesis is a well-studied field, research in tackling this problem using a more general approach like geometric deep learning has only recently begun [152, 154, 157]. The proposed approaches generate graphs either by learning probabilistic fully connected graphs (quadratic memory consumptions) or by learning a variety of neural networks that are used to sequentially derive graph structures based on a sequence of decisions (not end-to-end, missing parallelism).

Considering the results of WPs 1 and 3, we study *deep end-to-end hierarchical graph synthesis* with the help of unpooling layers, that is, the hierarchical generation of graphs from smaller ones. Although unpooling layers are well defined for traditional CNNs on regular grids, challenges arise for irregularly structured data due to their possibly variable number of neighbouring vertices and their dynamic spatial relation to each other.

The major challenge of this work package lies in propagating gradients back through unpooling layers, i.e., discrete graph structures. As a first approach, we want to study the possibilities of approximating gradients of discrete functions using gradient estimators [149]. Another approach might be to force the network into learning correlated graph generation features in the previous convolution operators (i.e., whether to add a vertex or edge to the graph), e.g., by using neighbourhood and self-attention mechanisms, and generate larger graphs based on these features. Note that in this approach, although graph generation is indeed (implicitly) learnable, gradients do not pass through these discrete structures. In addition, it is necessary to approximate graph similarity between the input and its generated version in a differentiable fashion for computing the loss.

Solving the above challenges allows for the design of encoder-decoder and generative architectures for solving graph compression and graph synthesis tasks, respectively. Graph compression is of fundamental interest, where storing of large (temporal) graphs may be intractable. In addition, the latent features can be used not only for graph compression, but also to address the problem of graph classification. Furthermore, graph synthesis finds its use cases in, e.g., the extrapolation of missing data for predictive tasks such as link prediction. Within the Collaborative Research Centre, link prediction is of fundamental interest in project C5, where a task is to find tracks of corresponding vertices from different layers.

**Work package 5. Learning on spatio-temporal graphs** This work package deals with the development of learning methods on temporal and spatio-temporal graphs. *Dynamic graphs* are graphs that change their structure (vertices and/or edges). In the case that the changes depend on discrete time steps, they are called *temporal graphs*. In CPSs, the corresponding graph vertices have locations in space; thus they are called *spatial graphs*. Of interest in our project are so-called *spatio-temporal graphs* that are temporal graphs where the vertices (and edges) have certain locations in space changing over time (see figure 3.5 on page 144 for an example). The challenges involved in this work package are the development of appropriate concepts as well as the integration of the proposed learning methods (cf. WP 1/2/3) to spatio-temporal graphs. We separate these challenges into the following conceptual aspects:

- development of temporal graph kernels, and
- design of deep recurrent graph neural networks.

This further poses challenges for the resource-efficient storage and processing of graph based data. As another hurdle, vertices in graphs may relate to different time scales, and thus require strategies and concepts for standardisation. Moreover, edges in graphs may appear or disappear completely. Within the Collaborative Research Centre, spatio-temporal graphs represent inherent aspects such as the modelling of autonomous CPS-units, IoT networks (cf. A4) or traffic prediction tasks (cf. B4).

**Temporal graph kernels.** Building on our work on developing new graph kernels we can now proceed to new temporal versions of graph kernels for learning tasks on sequences or streams of graphs. Graph kernels have been applied to various learning tasks in social network analysis, neuroscience, and bioinformatics. However, these kinds of networks evolve over time ,and the majority of studies so far do not consider temporal parameters. Our goal is to track the graph kernels on dynamic graphs evolving over time. Thus we will be able to classify graphs with respect to their growth structure. Possible application scenarios include traffic prediction (B4) and the detection of concept drifts in graph streams (A3). An interesting scenario is to recognise (looming) conflict situations in logistic sensor-actuator networks by interpreting them as a stream of changing graphs, where the task is to classify the graph at a certain time step to represent a critical or uncritical situation (A4). In this scenario with spatio-temporal graphs, concept drifts are likely to occur, and we will collaborate with A3 to obtain adaptive learning algorithms. Moreover, we will build on recent advances in time series prediction for graphs using kernel-based methods to explore the potential of tutoring systems [161]. To this end, the problem-solving behaviour of an experienced worker within the logistic sensor-actuator-network should be analysed in order to provide automated hints to trainee employees. Our findings will lead to the development of new graph kernels devised explicitly for graph sequences. We would also like to develop algorithms that dynamically calculate the changes of specific graph kernels (e.g., local  $k$ -WL, graphlet kernel, shortest path kernel) in a sequence of similar graphs/networks. The update time should be guaranteed to be small without sacrificing too much in quality guarantee. For example, we like to advance the approximation algorithm developed in phase 2 of our project in order to track the spectra of a graph sequence. Our methods are based on randomisation, dynamic graph algorithms, and our previous work on graph kernels.

**Deep recurrent graph neural networks.** Deep recurrent neural network architectures, though remarkably capable at modelling sequences, lack an intuitive high-level spatio-temporal structure. Therefore, many proposed works combine traditional CNNs for visual feature extraction followed by RNNs for sequence learning. However, although many problems in computer vision inherently have an underlying high-level structure and can benefit from it, this does not apply for irregularly structured data such as graphs. As such, graph neural networks can not directly be ported to the context of sequence learning, as the removal of vertices and edges needs to be taken into account, and this too holds challenges for future research. Moreover, graph vertices embedded in Euclidean space relate to time-varying positions and thus may modify the underlying irregularly structured data. Therefore, we want to study the combination of graph convolution and recurrent modules (e.g., LSTMs and GRUs) known from traditional RNNs and modify them to fit into the spatio-temporal graph scenario more precisely. As a first approach, we want to derive theoretical evaluation for the possible weaknesses of these modules in the context of graphs compared to regularly structured data.

**Work package 6. Robust and scalable graph algorithms on Cyber-Physical Systems** In this work package, we reinforce the developed algorithms to allow for a deployment on large-scale but resource constrained CPSs and a successful use in real application scenarios. The use of our algorithms on these systems allows (i) single systems to process or classify *irregularly* sampled real-world data, e.g., in autonomous driving scenarios where graphs can be built up from point clouds. In addition, our algorithms can be used for (ii) the orchestration of CPSs linked together (forming a spatio-temporal graph network), e.g., to predict collisions between entities.

We distinguish two different groups of resource constraints in this work package: the availability of labelled data and the hardware and (soft) real-time constraints.

**Availability of labelled data.** A major bottleneck for generalisation is the collection of sizable and reliable labelled real-world data, as resource constraints for labelling include financial costs, time, and data volume as well as the availability and expertise of human beings. Due to the cost-

sensitive task of labelling we want to explore graph based *few-shot learning* (learning from few examples) and *semi-supervised learning* (learning from few labelled and many unlabelled examples) and validate their trade-offs between simplicity, generality, performance, and sample complexity. Most models tackle this problem by learning via low expressive models, i.e., only using a small number of parameters [153, 167]. We want to explore causes and techniques to also allow more expressive models, e.g., SplineCNN, to better generalise in these scenarios. As a first approach, we are studying the challenging field of *graph based data augmentation*. Here, graph data augmentation refers to the regularisation scheme of inflating the data set by label-preserving transformations, e.g., randomly changing the graph connectivity by a small factor. As augmentation typically lowers the quality of a given data set, research questions arise as to how far the aggressiveness of graph data augmentation influences the generalisation of the model.

**Hardware and (soft) real-time constraints.** In this issue our focus will be on hardware as well as (soft) real-time resource constraints relevant to CPSs. In particular, our algorithms will further be adapted to fit constraints such as runtime, memory, inference, accuracy, energy, and transmission speed. Our method for computing our newly suggested local  $k$ -WL kernel can be adapted to run in constant time without sacrificing quality guarantees too much [A6/175]. Due to its structure it can be transferred to parallel distributed systems. However, it still needs to be investigated whether the theoretical assumptions of the model fit to CPSs.

For graph neural networks, we want to develop highly specialised and efficient networks that allow for deployment on mobile hardware (e.g., DTSSs, UAVs). We are interested in complex (large-scale) networks for the orchestration of sensors, actuators and humans in three-dimensional space with additional consideration of time. We examine the transferability of traditional deep learning techniques like pruning, regularisation, and inference optimisation [142] in the graph context, as well as studying graph-specific scalability issues like the GPU-efficient treatment of sparse matrices.

For spatio-temporal graphs, the resource-efficient storage and processing of graph based data need to be investigated. An illustrative example is the autonomous local path planning directly on resource constrained UAVs. In this issue, a focus will be on streaming algorithms and methods for data reduction, e.g., in the form of coresets, together with A2. Jointly with project B2 we want to reduce the number of required input data from the spatio-temporal domain that is required to reliably perform prediction.

**Work package 7. Evaluation** The goal of this work package is the cross-sectional task of evaluating the developed graph algorithms of each work package. A special focus of this work package, however, is the relationship to the requirements of CPSs. Actuator dynamics, multi-sensory, complex interactions among computation units and physical components (i.e., DTSSs, UAVs, humans), computation and (inter-, intra-) communication delays must all be considered carefully to achieve accurate tests. For complex systems with many degrees of freedom, we want to apply a multi-methodical approach. The behaviour  $\phi(\mathbb{S}, \mathbb{A}, p, i)$  of a system  $\mathbb{S}$  can be characterised by the control parameters  $p$ , both for the system  $\mathbb{S}$  and the algorithms  $\mathbb{A}$ , and the input parameters  $i$ . The developed methods of WP 1 to 6 will serve as algorithms  $\mathbb{A}$ . The system  $\mathbb{S}$  can be a model, a simulation or a real system. This allows for various types of experiments (virtual, real, and mixed lab settings) in the form of model in the loop (MIL) and software in the loop (SIL) tests. In principle, we want to investigate whether a behaviour  $\phi(\mathbb{S}, \mathbb{A}, p, i)$  corresponds to a given specification or a specified range  $\tilde{\phi}$ . Requirements include the actuator and sensor behaviour, but especially the adherence to resource limits. If the CPS or the parameter ranges are too complex for a sustainable evaluation, we will at least apply falsification tests ( $\phi \neq \tilde{\phi}$ ).

As virtual experiments, we want to analyse state of the art benchmark libraries and extend our repository of benchmark data sets established in phase 2. We also want to use high level archi-

tures (HLAs) for simulations. For CPSs, the robot operating system (ROS)<sup>1</sup> in combination with the virtual robot experimentation platform (V-REP) is one recommended option. The evaluation will focus on real and mixed lab settings in cooperation with the other projects. Within the Collaborative Research Centre, we want to classify critical situations in large-scale logistic sensor-actuator networks (A4), forecast traffic flow and predict traffic jams (B4), as well as work on various issues related to high-dimensional and high-frequency irregularly structured data from astroparticle physics (C3 and C5). Joint preliminary work by the PIs ten Hompel and Weichert is available for carrying out the experiments, e.g., to use the innovation laboratory at TU Dortmund University (A4). We further want to carry out studies outside of the Collaborative Research Centre, both in joint research activities of the PIs, e.g., the expansion of high-voltage networks, and generally important topics like social networks.

### Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Feature learning via neighbourhood aggregation																	
2. Learning kernels from optimal assignments																	
3. Deep learning on irregularly structured data																	
4. Graph synthesis with deep generative architectures																	
5. Learning on spatio-temporal graphs																	
6. Robust and scalable graph algorithms on Cyber-Physical Systems																	
7. Evaluation																	

### 3.5 Role within the Collaborative Research Centre

The goal of project A6 is the development of resource-efficient algorithms for the analysis of large-scale graphs or a large number of graphs. A6 combines methods from algorithm theory and algorithm engineering with (deep) learning methods from the field of graph mining. Research questions concern the examination of neighbourhood aggregation functions, development of (temporal) graph kernels as well as approaches to generalise deep learning architectures to graph-structured data. The research questions will also be shaped by the leading challenges of CPSs, namely comprehensive resource constraints such as memory requirements, training and inference runtime as well as energy consumption. Project A6 is involved in both the CRC research group “information retrieval” and the CRC research group “data analysis”.

In the past, we have cooperated successfully with projects A1, B2, B4 and C1, which led to several joint publications. There are comprehensive possibilities to expand and deepen the cooperations in the next phase.

<sup>1</sup>ROS: <http://www.ros.org>, accessed: 2018-04-06

## Project A6

We will investigate data reduction techniques, e.g., in the form of coresets, and streaming algorithms for graphs together with project A2. In order to study the relationship between graph features and target changes over time, joint research on concept drifts and limited samples will be carried out with A3. Furthermore, the graph synthesis with deep generative architectures of A6 provides a conceptual link to A3, in which methods for combining real data with additional simulated data are developed. There will be a substantial cooperation with A4. Project A4 examines resource-aware platforms in the context of the innovation laboratory. Here, the group of different entities (DTSSs, UAVs, and humans) is representable by a spatio-temporal graph and hence allows for the analysis of collaborative processes using graph based methods. As such, the lab is used as a central benchmark system for a cooperative evaluation of the developed spatio-temporal graph methods of A6 and the methods for resource-aware heterogeneous systems and indoor localisation of A4.

Since graphs are of far-reaching importance for many projects within the CRC (or their questions can be transferred to graph problems), cooperation takes place not only on the conceptual level (A-area), but also with research questions from the areas B and C. A6 will continue its cooperation with B2 to generalise deep learning techniques to irregularly structured data. In addition, B2 will realise efficient, task-specific deep neural network (DNN) modules for various issues. This is particularly important for the development of methods for spatio-temporal graphs and the use of the methods of A6 on systems with hard resource constraints. Project B4 investigates dynamic transport networks to forecast traffic and predict traffic jams. The learning methods on temporal and spatio-temporal graphs developed in A6 may prove useful for the emerging research questions in B4, where environment information is gathered irregularly from vehicular sensor data. A6 will intensify its cooperation with C1 to analyse dependency graphs between molecular features and their dynamics. Moreover, we will jointly investigate the application of key-value data structures developed for  $k$ -mers for storing and tracking graph features. The methods for graph mining of A6 will be used to examine the high-dimensional and high-frequency irregularly structured data from (astro)particle physics, which are the subjects of projects C3 and C5. The examination of learning issues for the hexagonally arranged data is a central topic in C3, where the methods of A6 can help. For C5, the developed link prediction model in A6 might be of fundamental interest to find tracks of corresponding vertices in different layers.

## 3.6 Differentiation from other funded projects

### GraBaDrug

**(Mutzel, Reference number DFG: MU 1129/10-2) (Funding period: 2014–2019)**

The project “Graph-based methods for the rational drug design” is part of the priority program SPP 1736: “Algorithms for Big Data”. There we concentrate on clustering of sets of molecule structures (unsupervised); as similarity measure, we use the maximum common subgraph paradigm, which is not considered here.

### GRK 1855

**(Mutzel, Reference number DFG: Member of the GRK 1855/1) (Funding period: 2013–2018)**

The graduate school titled *Discrete optimization of technical systems under uncertainty* focuses on optimisation problems under uncertainty with the incorporation of people in the optimisation process. There is no overlap with A6.

### ADJUTANT

**(Weichert, Reference number AiF: ZF4119002DB7) (Funding period: 2018–2020)**

The goal of the project is the development of a robot-based inspection system for the automated non-destructive testing of lightweight components made of fibre plastic composites. Essential issues are the development of multi-criteria optimisation algorithms for path planning based on implicit modelling and a reliable safety concept for collaborative human-robot interaction.

### InÜDosS

**(Weichert, Reference number FOSTA: P 1326/17/2018) (Funding period: 2018–2020)**

The aim of the project is automated condition monitoring of steel structures by Unmanned Aerial Vehicles, to develop methods for automated data analysis and the classification of damaged structures. As a result, the project should enable the creation of a building file in digital form with damage and defect reports as well as instructions for handling.

### CuKa

**(Weichert, Reference number DFG: WE 5036/4-1) (Funding period: 2018–2021)**

The main objective of the research project is to lay the foundations for the provision of a repository/exhibitor spanning both domain-internal (2D photographs, 3D scans) and a cross-domain search function for finding characters and text passages as well as for operationalising the analysis of the cuneiform script.

## 3.7 Project funding

### 3.7.1 Previous funding

The project has been funded within the Collaborative Research Centre since January 2015.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	2	129,000	2	129,000	2	129,000	2	129,000
Total	—	129,000	—	129,000	—	129,000	—	129,000
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	129,000		129,000		129,000		129,000	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Petra Mutzel, Prof. Dr., professor	Algorithm engineering	TU Dortmund	8	—	Existing funds
	2	Nils Kriege, Dr., postdoctoral researcher	Algorithm engineering	TU Dortmund	19.92	—	Existing funds
	3	Frank Weichert, Dr., postdoctoral researcher	Visual computing	TU Dortmund	9	—	Existing funds
	4	Bernd Zey, Dr., postdoctoral researcher	Algorithm engineering	TU Dortmund	4	—	Existing funds
	5	Andre Droschinsky, M.Sc., doctoral researcher	Algorithm engineering	TU Dortmund	15.93	—	Existing funds
	6	N.N., student assistant	Visual computing	TU Dortmund	10	—	Existing funds
	7	N.N., student assistant	Algorithm engineering	TU Dortmund	10	—	Existing funds
Non-research staff	8	Helmut Henning, technical employee	—	TU Dortmund	2	—	Existing funds
	9	Gundel Jankord, secretary	—	TU Dortmund	2	—	Existing funds
<b>Requested staff</b>							
Research staff	10	Matthias Fey, M.Sc., doctoral researcher	Visual computing	TU Dortmund	—	Doctoral researcher	—
	11	Christopher Morris, M.Sc., doctoral researcher	Algorithm engineering	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

**1. Mutzel, Petra**

Project management. Focus on algorithm engineering and graph algorithms. Mainly responsible for WP 5, additional cooperation in WPs 1, 3, 6, 7.

**2. Kriege, Nils**

Project management. Focus on graph comparison and graph kernels. Mainly responsible for WP 2, additional cooperation in WPs 1, 6, 7.

**3. Weichert, Frank**

Project management. Focus on geometric deep learning and algorithms for cyber-physical systems. Mainly responsible for WPs 3, 4, additional cooperation in WPs 1, 5, 6, 7.

**4. Zey, Bernd**

Collaboration in the development of techniques for dynamic graphs within the WP 5.

**5. Droschinsky, Andre**

Collaboration in the development of techniques for graph matching within the WPs 1, 2, and 6.

**6. N.N.**

Support of implementation within the WPs 3, 4, and 6.

**7. N.N.**

Support of implementation within the WPs 1, 2, and 5.

**8. Henning, Helmut**

Support with regard to the technical infrastructure.

**9. Jankord, Gundel**

Document and appointment management and, if necessary, travel planning and accounting.

**Job descriptions of staff for the proposed funding period (requested funds):**

**10. Fey, Matthias**

Development and realisation of approaches to generalise deep learning techniques to graph-structured data. Mainly working on WPs 3, 4, additional cooperation in WPs 1, 5, 6 and 7.

**11. Morris, Christopher**

Investigation of similarity measures from optimal assignments and neighbourhood aggregation functions, development and realisation of feature learning techniques for (temporal) graphs. Mainly working on WP 2, additional cooperation in WPs 1, 3, 5, 6 and 7.

### 3.7.4 Requested funding for direct costs for the new funding period

	2019	2020	2021	2022
TU Dortmund: existing funds from university	10,000	10,000	10,000	10,000
Sum of existing funds	10,000	10,000	10,000	10,000
Sum of requested funds	0	0	0	0

(All figures in euros)

### 3.7.5 Requested funding for instrumentation for the new funding period

This project does not request any funding for major research instrumentation.

### 3.1 General information about Transfer Project TB1

#### 3.1.1 Project title:

Analysis of Spectrometry Data with Restricted Resources

#### 3.1.2 Principal investigator(s)

Baumbach, Jörg Ingo, Prof. Dr., 02.07.1959, German

Hochschule Reutlingen  
Alteburgstraße 150  
72762 Reutlingen

Phone: 07121-271-2043  
E-mail: joerg.baumbach@reutlingen-university.de

Rahnenführer, Jörg, Prof. Dr., 19.05.1971, German

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

Phone: 0231-755-3121  
E-mail: joerg.rahnenfuehrer@tu-dortmund.de

#### 3.1.3 Application partner

##### B & S Analytik GmbH

*Contact:*

Sprave, Gabriele,  
Geschäftsführerin  
Otto-Hahn-Straße 15  
44227 Dortmund  
Phone: 0231-9742-6412  
E-mail: sprave@bs-analytik.de

B & S Analytik GmbH, founded 01/16/2009. Industry: Medical devices, analytical measuring devices. Business of the company: Development, construction, marketing and distribution of technical measurement and analysis instruments, as well as development and distribution of software and databases. Number of employees: 7. Annual turnover: 1,500,000 EUR. No foreign investments.

## 3.2 Project history

### 3.2.1 Report

#### Summary

The aim of the project is the development of resource-aware algorithms for processing and analysis of data generated by multi capillary column – ion mobility spectrometry (MCC-IMS). MCC-IMS is a technology that is cost-effective for the detection of volatile organic metabolites in the exhaled air. The application partner in the transfer project has equipment for and is experienced in measuring MCC-IMS data. In the first phase, methods for preprocessing of MCC-IMS data were developed, in particular for the automatic peak extraction from raw data. In the second phase, the technical aim was the provision of spectrometers that can be used at the bedside, for clinical patients with high risk who must be monitored, and solutions that are integrated in textiles and clothing. The technical preparations for integration in clothing and telemedicine connections include further miniaturisation of all spectrometer components.

The main algorithmic goal of the second phase was the development of overall analysis processes. These processes are combinations of methods for initial data preprocessing and subsequent support of medical decision making. A practicable analysis process must appropriately combine the data in a resource-saving manner. For MCC-IMS data, the most frequent application is classification, often in the situation of two different patient groups with different phenotype, different disease subtype, or different prognosis. Here, both univariate classifiers for the identification of relevant metabolites and multivariate classifiers for minimising prediction errors for the discrimination of the groups are of interest. In the second phase, optimal analysis processes starting with raw measurements and ending with classifier predictions were identified.

The analysis of breath measurements is often influenced by confounders like smoking habits or nutrition of the examined humans, and also by any kind of contamination of the air in the measurement place. In many evaluations, not enough attention is paid to the necessary statistical adjustments for confounders within the analysis processes. We performed a controlled study where the influence of confounders like sex and measurement device on the discrimination ability between two groups was analysed. Normalisation with respect to confounders and alignment of peaks between different devices are crucial for optimising the generalisability of the results.

Furthermore, methods for time series are of interest, where the goal is to detect critical change points in a time series of metabolite concentrations as early as possible. An example is an emerging sepsis whose early detection is crucial for the success of the therapy for an individual. For this task, time series data of the university hospital Homburg from rat models of intubated animals were available. These can be used for modelling the development of a sepsis based on the exhaled air concentrations of certain metabolites. In the second phase, transformed features for a better characterisation of the functional course of the time series of single metabolites were derived and analysed, in order to identify change points earlier. The main improvement was due to standardisation per rat, and for data sets with small sample size also due to stratified cross-validation.

Altogether, the advances in preprocessing of MCC-IMS data and the development of overall analysis processes now allow the implementation of adequate solutions for breath gas analysis for various application scenarios. Such solutions start with the resource-aware automatic preprocessing of raw MCC-IMS data and finish with decision support for disease diagnosis and treatment selection using state of the art statistical classification algorithms.

## Technical advancements of IMS

Considering IMS technology, worldwide more than 100,000 units are in service. Mostly, detection of chemical warfare agents, illegal and legal drugs, and explosives dominate the application fields. Here, false alarm rates of  $\frac{1}{1000}$  are accepted. In the recent years medical applications, especially in the investigations in the field of human exhaled breath, have gained in interest, first scientifically and most recently also commercially. In 2017 the first commercial IMS-based Exhaled Drug Monitor reached classification marked as CE medical – B & S Analytik GmbH, Dortmund, Germany. The Exhaled Drug Monitor EDMON is used for real-time measurement of the volatile propofol concentration in the exhaled air of patients receiving anaesthesia or sedation with propofol. Here, all parts including sampling, detection, and display of the actual value and a trend line were integrated into a single unit to make the EDMON usable online and bedside, directly in the operation theatre and in intensive care units, based on embedded mainboard solutions. In medical studies in total, 7236 exhaled air and 250 blood samples were analysed from 30 patients. The  $R^2$  values between measured and calculated propofol plasma concentrations ranged from 0.80 to 0.99 with a median value of 0.91. The overall median absolute performance error was 6.6% (median performance error: 0%). A correlation analysis between exhaled propofol concentrations and compartment concentrations based on measured plasma concentrations had a median  $R^2$  value of 0.93.

Furthermore, a pillow solution was developed at Reutlingen University to serve as sampling unit directly on the patient. A sampling system was created, requiring no action of the patient. When, e.g., the patient's head is put on the pillow, the system starts a small pump and a sample of some mL of exhaled and surrounding air is taken into the MCC-IMS system, which is integrated or installed in the base of the bed. Recently, the signals of infections were classified using MCC-IMS with exhalations from rats and patients, showing the development of the infection using a time series.

## Optimising the process from raw measurements to classification

The classification of diseases from molecular measurements, here based on metabolite data, typically requires an analysis pipeline. The input is raw noisy measurements, and the output is a prediction. In classification, typically a categorical label is predicted. Here, we focus on MMC-IMS (multi capillary column - ion mobility spectrometry) data that measure volatile organic compounds in the air of exhaled breath; see [TB1/181, TB1/182] for recent reviews on instrumentation and technical parameters. The raw data are measurements where peak regions represent compounds. These peaks have to be identified, quantified, and finally clustered across different experiments. For peak detection, together with project C4, an additive model was used to model possible convolutions of the peaks. After the preprocessing steps, the peak lists are the input for the classification algorithm with the goal to distinguish between (typically two) patient groups; see [TB1/188] for a recent example.

Currently, the single steps of this analysis process, especially for preprocessing, are not automated and thus require subjective manual intervention of human experts. Our goal was to construct a fully automatic pipeline from raw data to classification results, with competitive classification accuracy compared to the semi-manual state of the art approach [TB1/187, TB1/C1/186]. The manual procedure ( $VN^{manual}$ ) is subjective and for data sets with a relevant number of samples also very tedious. Methods for peak detection and for peak clustering were developed in the first phase [TB1/185]. We combined those algorithms and alternatives from the literature with various multivariate classification algorithms into overall analysis pipelines. All feasible combinations of algorithms for the three analysis steps were considered. For the management of multistep workflows we built on the expertise and cooperated with project C1. The combinations yielding the best classification results were considered the best combinations, accepting that this might not coincide with the best results for all single steps when considered separately.

- The first step, peak picking, accomplishes the task of identifying peak regions in the raw measurements and assigning intensity values to them. The selected algorithms covered various procedures, including the algorithms PME and OPME that were developed in the first phase as well as PDSA and SGLTR, which resulted from a project group guided by first phase members [TB1/189]. The software VisualNow also includes an automated algorithm ( $VN^{auto}$ ), based on k-means. Some, but not all of these algorithms, use only a small fraction of the data (5 spectra at a time).
- The second step, peak clustering, solves the task of deciding which peak positions from several raw measurements stem from the same true underlying metabolite. Thereby, the peak positions are aligned and peaks that do not have close peaks in the other measurements are discarded. As a result, the raw measurements are afterwards summarised by a common data table consisting of a previously unknown number of variables (peaks) and their intensities. As algorithms, the established concepts k-means (in  $VN^{auto}$ ), Grid Squares, DBSCAN, and Cluster Editing, adapted to peak clustering, and a new version of EM Clustering developed in the first phase were considered. As semi-automated gold standard, results obtained with the software VisualNow ( $VN^{manual}$ ) were used.  $VN^{manual}$  peak picking and  $VN^{manual}$  peak clustering involve manual decisions about peak positions and peak clusters. An expert examines the raw measurements and checks the respective areas in the other measurements, after a peak has been visually discovered.
- For the third step, i.e., for the construction of the multivariate classifier, competitive state of the art methods like SVMs, K-Nearest-Neighbour and different tree-based algorithms (Classification Tree, Generalised Boosted Models, and Random Forest (RF)) were considered. Hyperparameters were tuned within a nested cross-validation setting.

Not all combinations of algorithms for the three steps were technically feasible. The  $VN^{auto}$  peak picking can be combined with all clustering methods but the other peak picking algorithms cannot be integrated in the software VisualNow and thus also can not be connected with the  $VN^{auto}$  peak clustering step. Consequently, in total, 156 combinations of peak picking, peak clustering, and classification algorithm were analysed.

The analysis processes were evaluated on three different real medical data sets in an unbiased cross-validation setting. They cover different clinically relevant diseases, namely COPD, asbestos and a bacterial infection with pseudomonas aeruginosa. The data sets include 92, 30 and 39 diseased patients and 35, 37, and 30 healthy controls, respectively. In order to achieve unbiased results, we used a 10-fold cross-validation setting with stratified sampling, resulting in training sets with class proportions equal to those in the data set. Since the performance of a cross-validation is dependent on the splits, we repeated it 50 times, such that the variability of the results could be assessed. We used the AUC (area under the receiver operating characteristic curve) as a performance measure to avoid that the results depend on the probability threshold used for the binary classification.

In summary, it turned out that certain combinations yield promising results also across the different data scenarios. The best fully automated analysis process achieved even better classification results than the semi-manual standard approach [TB1/C1/186].

The strongest impact on the outcome can be attributed to the choice of the classification algorithm. The Random Forest clearly outperformed the other algorithms. The best combination, based on the rank sum achieved on the three data sets (in comparison to the other combinations), was SGLTR peak picking, DBSCAN peak clustering, and RF classification. The second and third best combinations also make use of SGLTR and RF but differ with respect to the peak clustering algorithm (EM and CE). The current gold standard, VN manual, is ranked fourth, when combined with the Random Forest classification.

Since the comparison of more than 150 combinations might lead to a random order at least among similarly performing algorithms, the single steps were also evaluated individually, averaging over the other two steps, respectively. For peak picking, VN<sup>manual</sup> achieved the best results (but here it was only averaged over the classification results, since VN<sup>manual</sup> could not be combined with any clustering methods), followed by LM and SGLTR. For the peak clustering task it was not entirely clear which method fits best to the peak picking algorithms, but mostly DBSCAN and EM performed best.

Since the best combination agrees with the results obtained when analysing the impact of the single steps separately, we recommend using it for future classification tasks on MCC-IMS data. It is an online algorithm and therefore allows the peak detection during an ongoing measurement, meeting the resource limitations of a miniaturised MCC-IMS. The SGLTR-DBSCAN algorithm is currently extensively further tested by B & S Analytik.

### **Confounder study**

In many studies concerning the potential of breath gas analysis for the classification of diseases, not much attention is paid to possible biases by confounders. However, the range of potential confounders is wide. Exhaled air can be affected by the metabolites that are present in the room or environment where the measurement is taken. Also, all kinds of technical settings, like the materials of the mouth piece, the temperature within the device, and the device itself as a whole, are likely to be confounders. Especially the nutrition of a person does have a great impact on the exhaled metabolites, which is the reason why several studies use these effects for demonstration purposes. Further examples are characteristics of the individuals themselves, such as sex, weight, age, and smoking habits. These often at the same time have a strong impact on the risk of suffering from a certain disease. In such cases there is a risk that a good disease classification performance is only due to the ability to discriminate between patients with different values for the confounders. Obviously, in this case patients with similar values for the confounders cannot be classified with respect to the disease.

We conducted a controlled study to analyse the role of such confounders for the classification of cohorts based on MCC-IMS data. In total, 49 volunteers participated in the study, with no exclusion criteria except pregnancy (due to laboratory regulations) and allergy to citrus fruits (since the experiment included consuming a glass of orange juice). Measurements were obtained on two different technical devices. Usually, studies are conducted only with one device, in order to minimise the variability of the results, but for broad application, the classification models must generalise to other devices, too. The focus of the study was to assess the influence of confounders such as sex, smoking habits, and the technical device and to propose statistical solutions to adjust for such effects.

Most of the participants in the study were staff of TU Dortmund University or volunteers from the laboratory at B & S Analytik, where the measurements were taken. Each participant filled out a questionnaire about her or his sex, age, smoking habits, food consumption and drinks in the last 12 hours, as well as an agreement to participate in the study. Afterwards, each participant's breath was analysed using MCC-IMS on the two devices. The starting device was chosen using stratified randomisation with respect to the gender. Then, the participant consumed a glass of orange juice and the breath gas was immediately analysed with the second device.

It is well known that some form of calibration for aligning the measurements on different devices is required. Currently, differences between devices can be taken into account by measuring a synthetic component composition on all devices and aligning the raw measurements according to the compared results. The devices are known to underlie stretching factors of both retention time and reduced inverse mobility axis. The stretching factors are currently estimated by visual trial and error. For reproducibility reasons, we replaced this manual procedure with fitting a linear

regression of the peak positions from the six components for both devices, separately for retention time and for inverse reduced mobility. After the peak picking step, performed with SGLTR, the peak positions of one device were aligned to the positions of the other device. The subsequent peak clustering step was then applied on the new positions.

We found no striking effects with respect to sex and smoking habits, but the device had a strong impact on the results. As can be seen in the PCA (principal component analysis) plot in figures 3.1 (left) and 3.2 (left), a large proportion of the variance in the data can be attributed to the device. Whereas the manual peak detection includes an internal alignment for the device, for the automated peak detection a peak position alignment (as described above) is required before peak clustering as an additional step, see figure 3.2 (centre). To reduce the effect of the device, we applied scaling to zero mean and unit variance, separately for both devices. As can be seen for both, the manual and automated peak detection, see figures 3.1 (right) and 3.2 (right), respectively, the scaling causes a shift in the principal components, removing the differences between the devices but maintaining differences between the feature of interest (here juice or no juice).

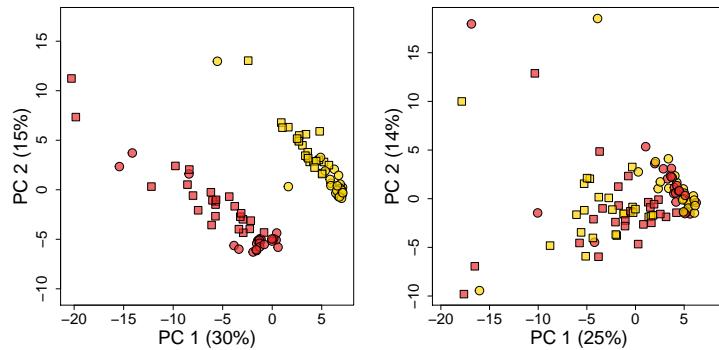


Figure 3.1: PCA for manual peak detection *Left:* without scaling *Right:* with scaling  
*Yellow:* Device A, *Red:* Device B, *Circle:* No juice, *Square:* Juice.

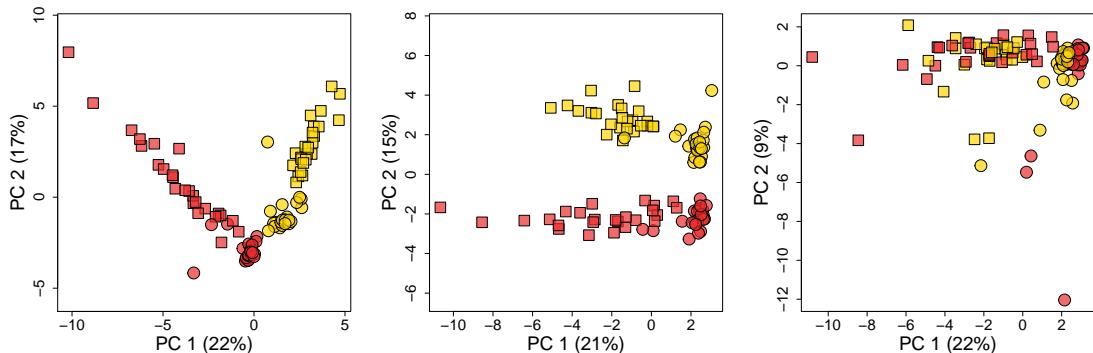


Figure 3.2: PCA for automated peak detection *Left:* without alignment or scaling *Centre:* with alignment but without scaling *Right:* with alignment and scaling  
*Yellow:* Device A, *Red:* Device B, *Circle:* No juice, *Square:* Juice.

### Time series and early detection of diseases

An important application of breath gas analysis is the detection of critical change points in a time series of metabolite concentrations as early as possible. Here, MCC-IMS has a strong advantage

compared to the competing technique gas chromatography, as for MCC-IMS the measurements are available within few minutes on a mobile device. A concrete example is an emerging sepsis, where early detection is crucial for the success of the therapy for a patient.

We performed analyses on measurements from 16 rats obtained in an animal study in Homberg [TB1/183]. For half of the rats, a sepsis was induced, for the other half, only sham surgery was carried out. Then, measurements of 100 metabolites were taken every 20 minutes for 12 hours. As expected, there is a clear negative effect of the induced sepsis on survival time. During the first phase we applied online smoothing filters to these data in order to remove outliers [TB1/190]. Online smoothing means that only values from previous time points are used for smoothing.

A major goal in the second phase, in order to increase the classification accuracy with respect to the two groups, was to better characterise the functional course of the time series and to derive appropriate transformed features for earlier identifying the change point. To derive transformed features, in an online fashion, previous values can be combined in different meaningful ways. As new features we analysed cumulative sums of previous values, first and second differences of the time series (representing slope and curvature of the time series), and sums of those differences.

We performed an extensive study to find out which transformed features improve the classification performance of a Random Forest, evaluated with leave-one-out cross-validation. Further, the impact of standardisation as an initial step was explored. As standardisation, multiple measurements taken slightly before the surgery were averaged per rat and subtracted from the following measurements. In summary, there was no substantial improvement due to the new features, but the standardisation led to higher classification accuracy and thus earlier sepsis detection. Another improvement was obtained by using a stratified version of the cross-validation, especially for the small number of subjects in the present study.

In a simulation study we evaluated the impact both of the feature transformations and of sample size. Typical courses of individual metabolite concentrations over time were modelled with constant and with logarithmic functions, where parameters were set so as to best mimic the real data. In total, we simulated 100 metabolites, a realistic number, from which 6 informative ones exhibited a decay course, whereas the others just represented Gaussian noise. The conclusions from the real data sets could be confirmed on the simulated data. Further, increasing the sample size from 16 to 100 rats resulted in considerably improved classification accuracy, emphasising the need for larger studies than are often carried out in practice.

### 3.2.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [TB1/181] R. Cumeras, E. Figueras, C. E. Davis, **J. I. Baumbach**, and I. Gràcia. “Review on Ion Mobility Spectrometry. Part 1: Current Instrumentation”. In: *Analyst* 140.5 (Mar. 2015), pp. 1376–1390 (cit. on p. 181).
- [TB1/182] R. Cumeras, E. Figueras, C. E. Davis, **J. I. Baumbach**, and I. Gràcia. “Trying to detect gas-phase ions? Understanding Ion Mobility Spectrometry Part 2: Hyphenated Methods and Effects of Experimental Parameters”. In: *Analyst* 140.5 (Mar. 2015), pp. 1391–1410 (cit. on p. 181).
- [TB1/183] T. Fink, A. Wolf, F. Maurer, F. W. Albrecht, N. Heim, B. Wolf, A. C. Hauschild, **B. Bödeker**, **J. I. Baumbach**, et al. “Volatile Organic Compounds during Inflammation and Sepsis in Rats: A Potential Breath Test Using Ion-mobility Spectrometry”. In: *Anesthesiology* 122 (Jan. 2015), pp. 117–126 (cit. on p. 185).

- [TB1/184] M. Gogiashvili, **S. Horsch**, R. Marchan, K. Gianmoena, C. Cadenas, B. Tanner, S. Naumann, D. Ersova, F. Lippek, et al. “Impact of intratumoral heterogeneity of breast cancer tissue on quantitative metabolomics using high-resolution magic angle spinning  $^1\text{H}$  NMR spectroscopy”. In: *NMR in Biomedicine* 31.2 (Dec. 2017).
- [TB1/185] **D. Kopczynski** and **S. Rahmann**. “An online peak extraction algorithm for ion mobility spectrometry data”. In: *Algorithms for Molecular Biology* 10.1 (2015), p. 17 (cit. on p. 181).
- [TB1/C1/186] **S. Horsch**, **D. Kopczynski**, E. Kuthe, **J. I. Baumbach**, **S. Rahmann**, and **J. Rahnenführer**. “A detailed comparison of analysis processes for MCC-IMS data in disease classification—Automated methods can replace manual peak annotations”. In: *PLOS ONE* 12.9 (2017), e0184321 (cit. on pp. 20, 181, 182).
- [TB1/187] **S. Horsch**, **D. Kopczynski**, **J. I. Baumbach**, **J. Rahnenführer**, and **S. Rahmann**. “From raw ion mobility measurements to disease classification: a comparison of analysis processes”. In: *PeerJ PrePrints* 3 (2015) (cit. on p. 181).
- [TB1/188] Y.-i. Yamada, G. Yamada, M. Otsuka, H. Nishikiori, K. Ikeda, Y. Umeda, H. Ohnishi, H. Kuronuma Koji and Chiba, **J. I. Baumbach**, et al. “Volatile Organic Compounds in Exhaled Breath of Idiopathic Pulmonary Fibrosis for Discrimination from Healthy Subjects”. In: *Lung* (2017), pp. 1–8 (cit. on p. 181).

#### b) Other publications

- [TB1/189] **A. Egorov**, A. König, M. Köppen, H. Kühn, I. Kullack, E. Kuthe, S. Mitkovska, R. Niehage, A. Pawelko, et al. *Ressourcenbeschränkte Analyse von Ionenmobilitätsspektren mit dem Raspberry Pi*. Abschlussbericht der Projektgruppe 572 der Fakultät für Informatik. Technischer Bericht 5. TU Dortmund, May 2014 (cit. on p. 182).
- [TB1/190] **S. Horsch**. “Analyse zeitabhängiger IMS-Messungen”. Abschlussarbeit. TU Dortmund, June 2014 (cit. on p. 185).

#### 3.2.3 Documentation of further activities

This project did not carry out any further activities.

### 3.3 Project funding

Funding of this project within the Collaborative Research Centre started in January 2011. The project will be completed by the end of the current funding period.

### 3.3.1 Project staff in the ending funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution, collaboration partner	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Jörg Rahnenführer, Prof. Dr., professor	Statistical methods in genetics and chemometrics	TU Dortmund	4	—	Existing funds
<b>Application partner staff</b>							
Research staff	2	Jörg-Ingo Baumbach, Prof. Dr., professor	Analytics	B & S Analytik GmbH	4	—	—
	3	Rebecca Brehm, M.Sc., doctoral researcher	Analytics	B & S Analytik GmbH	3.3	—	—
	4	Bertram Bödeker, doctoral researcher	Analytics	B & S Analytik GmbH	10	—	—
	5	Laura Klimek, doctoral researcher	Analytics	B & S Analytik GmbH	7.5	—	—
	6	Jessica Kuhlmann, doctoral researcher	Analytics	B & S Analytik GmbH	20	—	—
	7	Andreas Pfannenschmidt, doctoral researcher	Analytics	B & S Analytik GmbH	2.9	—	—
	8	Ann-Kathrin Sippel, doctoral researcher	Analytics	B & S Analytik GmbH	3.1	—	—
	9	Claudia Fauser-Gashi, student assistant	Analytics	B & S Analytik GmbH	2.9	—	—
	10	Erick Franieck, student assistant	Analytics	B & S Analytik GmbH	2.9	—	—
	11	Kathrin Geworski, student assistant	Analytics	B & S Analytik GmbH	2.5	—	—
	12	Yunqing Li, student assistant	Analytics	B & S Analytik GmbH	2.9	—	—
	13	Gianluca Pielsticker, student assistant	Analytics	B & S Analytik GmbH	2.5	—	—

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution, collaboration partner	Project commitment in hours per week	Category	Funding source
<b>Application partner staff</b>							
	14	Jacob Scheuble, student assistant	Analytics	B & S Analytik GmbH	2.5	—	—
	15	Julia Steinbach, student assistant	Analytics	B & S Analytik GmbH	2.5	—	—
	16	Larissa Walter, student assistant	Analytics	B & S Analytik GmbH	2.9	—	—
	17	Denise Wilhelm, student assistant	Analytics	B & S Analytik GmbH	1.6	—	—
	18	Yazhen Yang, student assistant	Analytics	B & S Analytik GmbH	2.5	—	—
Non-research staff	19	Peter Kaiser, technician	—	B & S Analytik GmbH	10	—	—
	20	Carmelo Nieddu, technician	—	B & S Analytik GmbH	10	—	—
<b>Staff funded with approved grant money</b>							
Research staff	21	Salome Horsch, M.Sc., doctoral researcher	Statistical methods in genetics and chemometrics	TU Dortmund	39.83	Doctoral researcher	—

**Job description of staff (supported through existing funds):**

**1. Rahnenführer, Jörg**

Project management. Focus on statistical learning. Cooperation in WP2, WP3, WP4.

**Job descriptions of staff (supported by application partner):**

**2. Baumbach, Jörg-Ingo**

Project management. Focus on metabolomics and clinical diagnostics. Cooperation in WP1 and WP5.

**3. Brehm, Rebecca**

Cooperation in WP1 and WP5.

**4. Bödeker, Bertram**

Cooperation in WP1 and WP5.

**5. Klimek, Laura**

Cooperation in WP1 and WP5.

**6. Kuhlmann, Jessica**

Cooperation in WP1 and WP5.

**7. Pfannenschmidt, Andreas**

Cooperation in WP1 and WP5.

**8. Sippel, Ann-Kathrin**

Cooperation in WP1 and WP5.

**9. Fauser-Gashi, Claudia**

Cooperation in WP1 and WP5.

**10. Franieck, Erick**

Cooperation in WP1 and WP5.

**11. Geworski, Kathrin**

Cooperation in WP1 and WP5.

**12. Li, Yunqing**

Cooperation in WP1 and WP5.

**13. Pielsticker, Gianluca**

Cooperation in WP1 and WP5.

**14. Scheuble, Jacob**

Cooperation in WP1 and WP5.

**15. Steinbach, Julia**

Cooperation in WP1 and WP5.

**16. Walter, Larissa**

Cooperation in WP1 and WP5.

**17. Wilhelm, Denise**

Cooperation in WP1 and WP5.

**18. Yang, Yazhen**

Cooperation in WP1 and WP5.

**19. Kaiser, Peter**

Cooperation in WP1 and WP5.

**20. Nieddu, Carmelo**

Cooperation in WP1 and WP5.

**Job descriptions of staff (funded with approved grant money):**

**21. Horsch, Salome**

Statistical methods development and analysis in WP2, WP3, WP4, Cooperation in WP1 and WP5.



### 3.1 General information about Transfer Project B2

### 3.1.1 Project title:

## Resource-aware real-time analysis of artefact afflicted image sequences for the detection of nano-objects

### 3.1.2 Research area(s):

409-05 (Image and Language Processing), 205-01 (Medical Biometry, Medical Informatics)

### 3.1.3 Principal investigator(s)

Hergenröder, Roland, Dr., 06.08.1961, German

Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.  
Bunsen-Kirchhoff-Straße 11  
44139 Dortmund

Phone: 0231-1392-178  
E-mail: roland.hergenroeder@isas.de

Weichert, Frank, Dr., 07.09.1967, German

LS 7, Fakultät für Informatik, Technische Universität Dortmund  
Otto-Hahn-Straße 16  
44227 Dortmund

Phone: 0231-755-6122  
E-mail: [frank.weichert@tu-dortmund.de](mailto:frank.weichert@tu-dortmund.de)

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

( ) no (x) yes

*Dr. Frank Weichert*

End date of fixed-term contract: 31.05.2021 (There is an ongoing application for a permanent position)

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Application partner

#### ARTES Biotechnology GmbH

*Contact:*

Jenzelewski, Volker,  
Technology director  
Elisabeth-Selbert-Straße 9  
40764 Langenfeld  
Phone: 02173-27587-19  
E-mail: v.jenzelewski@artes-biotechnology.com

The ARTES Biotechnology GmbH is a contract development organisation with focus on biopharmaceutical technology and process development for vaccines and biosimilars. ARTES is specialised in developing biopharmaceutical production processes. The majority of shares of the company are owned by Dr. Michael Piontek (50.05%), founder and managing director, and SUN Pharma Global FZE (45.00%; foreign investment) as strategic investor. ARTES, founded in 2002, has 23 employees and an annual turnover of 2,000,000 EUR (2017).

#### Paul-Ehrlich-Institut

*Contact:*

Stahl, Dorothea, PD. Dr.  
Head of Section Transfusion Medicine  
Paul-Ehrlich-Straße 51-59  
63225 Langen  
Phone: 06103-77-1850  
E-mail: dorothea.stahl@pei.de

The Paul-Ehrlich-Institut (PEI) is the Federal Institute for Vaccines and Biomedicines. PEI reports to the German Federal Ministry of Health and is a senior federal authority in the field of medicinal products providing services in public health. The transfusion medicine section is responsible for the regulation of blood, blood components, and haematopoietic stem cells in Germany and contributes to global issues also by its work in *The world health organization (WHO) Collaborating Centre for Quality Assurance of Blood Products and in vitro Diagnostic Devices* at PEI. Current scientific work of the transfusion medicine section with a staff of 12 employees focuses on profiling quality, safety, and efficacy of blood components and haematopoietic stem cells in dependence of the underlying manufacturing processes and testing methods and their changes over time.

### 3.1.5 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes    (x) no
2.	clinical trials	( ) yes    (x) no
3.	experiments involving vertebrates.	( ) yes    (x) no
4.	experiments involving recombinant DNA.	( ) yes    (x) no
5.	research involving human embryonic stem cells.	( ) yes    (x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes    (x) no

## 3.2 Summary

The long-term goal of the project is to develop and analyse methods for the recognition of specific nano-objects by analysing data-intensive, noisy, and artefact afflicted image sequences from sensor systems, which are subject to several aggravating influences. A challenge is that the processing of the image sequences should be done in (soft) real-time while minimising resource consumption, e.g., of energy and memory. Therefore, multi-objective optimisation is an important topic of our research. Amongst other objectives, we optimise hardware efficiency, time efficiency, and detection quality regarding varying demands. At the centre of attention are the detection and quantification of biological nanoparticles with the plasmon assisted microscopy of nano-sized objects (PAMONO) sensor. Topics of consideration include aspects of technological development and biophysical methods for nano-object detection in vital, real-life scenarios.

A main task in the second funding period has been the development and evaluation of algorithms for a cooperative processing of different temporally and locally independent mobile sensor units. This included methods for resource optimisation on program and platform level. In addition, concepts for simultaneous detection of several types of nano-objects, including different media, have been investigated. The developed adaptive GPU-based algorithms are partly based on different kinds of deep neural networks (DNNs). Notably, the previous analyses have revealed that the PAMONO sensor can also be extremely useful in relevant fields of pharmaceutical quality control and in-process control of biological materials that may contain viruses (blood products) or virus-like particles (vaccines). Therefore, project B2 is to be continued as a transfer project. In the third funding period, the project leaders Jian-Jia Chen (project change) and Heinrich Müller (retirement) will leave the project and Frank Weichert will join the project management.

Based on the experience from the project, we aim to develop methods and technical concepts with respect to robustness and applicability to identify the most feasible methods for fast, user-friendly sensor software adaptation in the field of quality control of medicinal products. This should be achieved in collaboration with the application partners ARTES Biotechnology GmbH (ARTES) and Paul-Ehrlich-Institut (PEI). They are involved in the methodological development and especially responsible for the subsequent commercial aspects of the PAMONO sensor in fields of quality control in the production of vaccines (ARTES) and quality control of blood donations and blood products (PEI). A challenge is to adapt to altering environmental conditions and thus to increasingly diverse characteristics of images. For this purpose, we want to examine methods that increase adaptivity of deep feature learning approaches and develop highly specialised, efficient architectures for DNNs (differentiable programming). At the technical level, we will focus on the development of novel approaches for PAMONO sensor surface functionalisation and of an innovative, adaptive biosensor/actuator unit based on the PAMONO sensor. Therefore, algorithmic solutions for automatic adjustment of the involved actuators and automatic monitoring of the

PAMONO unit sensor behaviour are needed. As a result of the project, we want to achieve a methodical and technical solution that allows for using the PAMONO technology for medicinal product quality control.

### 3.3 Project progress to date

Currently, principal investigators of the second funding period are Jian-Jia Chen, Heinrich Müller, and Roland Hergenröder. In the third funding period of the Collaborative Research Centre (CRC) 876, Roland Hergenröder remains project leader and Frank Weichert will join the project management. In the first and second project phase, he was already an employee in the project and involved in the project coordination. ARTES and the PEI will be participating as application partners for quality control of medicinal products. Since the conceptual focus of the transfer project in the third funding period is no longer in Jian-Jia Chen's research area, he will leave the project. As a cooperation partner, he will remain involved in the project through the projects A1 and A3. Upon retirement, Heinrich Müller will leave the project as project leader.

#### 3.3.1 Report and current state of understanding

The long-term goal of the project is the (soft) real-time recognition of specific nano-object binding signals, especially from biological nano-objects, in image sequences considering limited resources. The analysis is carried out on the basis of data-intensive and very noisy or artefact afflicted image sequences from a sensor system. The essential goal is to optimise the trade-off between hardware efficiency, time efficiency, and detection quality.

One goal of the second funding period was the development of algorithms for a globally controlled image analysis. Additional goals resulted from the focus on distributed systems, for which paradigms at the algorithms and execution level have been developed. Using an execution context-sensitive algorithm and programming paradigm, the resource-aware mapping of algorithms and programs to distributed computing platforms could be made more effective and efficient. In addition, the proposed collaborative execution paradigm has been implemented by employing reactive and proactive offloading concepts. This improves the execution of calculations with real-time requirements on resource-limited, distributed, embedded systems.

Regarding the PAMONO technology, work has been carried out that led to further development and miniaturisation of the PAMONO sensor, which increases the portability. The analytical capability of the sensor could be transferred to other fields of application. Concepts for the simultaneous detection of different types of nano-objects were implemented on sensor and algorithm levels. The work carried out in the second funding period was divided into four sub-areas:

- A. Functional performance enhancement of image and data analysis methods (cf. WP 1, WP 2, WP 3, and WP 4, which have been primarily processed by the Chair of Computer Science 7 (Müller))
- B. Paradigm development on algorithms and execution levels (cf. WP 5, WP 6, and WP 7, which have been primarily processed by the Chair of Computer Science 12 (Chen))
- C. Generalisation and miniaturisation of the PAMONO sensor (cf. WP 8 and WP 9, which have been primarily processed by the ISAS (Hergenröder))
- D. Comprehensive optimisation and validation (cf. WP 10, which has been jointly processed)

In the following, a description of the achievements for each of the sub-areas with references to important published research papers<sup>1</sup> is given, followed by a description of the current state of research and the new challenges for the next funding period.

### A. Functional performance enhancement of image and data analysis methods

This section summarises the work and findings related to image data analysis and global optimisation of data processing (WPs 1, 2, 3, and 4).

**Multi-stage approaches.** For the developed multi-stage approaches [B2/124, B2/212], spatial, temporal, and spatio-temporal features as well as features in frequency domains were analysed. The overall process has been transferred to a generic, adaptive, GPU-based open computing language (OpenCL) pipeline for the detection of nano-objects. The approach is named VirusDetectionCL and is further described in the dissertations of Libuschewski [B2/124] and Siedhoff [B2/212]. Regarding the reduction of data noise, both spatial and temporal noise could be further reduced by a combination of standard image filters and wavelet denoising [B2/209]. Another field of investigation was the analysis of moving particles, e.g., the Influenza A virus (IAV). A multi-target track-prediction linking based on the Kalman filter has been developed for image sequences with high artefact density. The approach allows context-sensitive detection to filter inaccuracies. However, in our studies, we were not able to visualise the movement of IAV particles. One of the possible reasons may be associated with the use of inactivated IAV particles. As an alternative, we studied moving polystyrene particles. We achieved movement of these particles by modifying the gold sensor surface charge and analysed it in different laboratory scenarios.

**Global optimisation of the image analysis.** Our PAMONO signal processing pipeline consists of a large number of image processing methods that can be chosen and configured by parameter sets [B2/212, B2/124]. Therefore, a global multi-objective optimisation approach, termed SYNthesis/OPTimization/analySIS (SynOpSis), has been developed [B2/212] that automatically adapts algorithm choice and parameters to changes in physical sensor parameters. Moreover, with SynOpSis, large amounts of annotated data can be synthesised. SynOpSis has been applied to the analysis of sensor images with the VirusDetectionCL detector (high sensitivity) [B2/124] in combination with a machine learning-based classifier (high precision) [B2/212]. With SynOpSis the detection of particles with a median signal-to-noise-ratio (SNR) below 2.0 has been made possible. This pushed the reliable detection limits from 200 nm particles down to 100 nm. Moreover, further research focused on the development of an unsupervised detector approach, which does not require optimisation and training in case of changes to the sensor setup [B2/207] (best paper award). For the unsupervised detector [B2/207], a recall of 0.76 to 0.98 could be obtained on data with a median SNR of 2.2 down to 1.25. In contrast, the supervised detector achieved a recall of 0.71 to 0.89 and required 28 detector parameters.

**Deep feature learning.** Further research to improve the image processing methodology was carried out through the application of deep neural networks (DNNs). A new real-time capable GPU pipeline has been designed, containing four different DNNs, to process sensor images [B2/210] (best paper award). We combine fully convolutional networks for spatio-temporal proposal generation and time-series analysis, as well as convolutional neural networks (CNNs) for particle classification and size estimation. We were able to automatically detect particles down to the size of 80 nm, improved the  $F_1$ -score from 0.412 to 0.768 and from 0.097 to 0.655, compared to our baseline without DNNs, and achieved results, which were previously achieved on data with an SNR of 1.25, on

---

<sup>1</sup>Publications from phase 2 of the B2 project are listed at <https://sfb876.tu-dortmund.de/SPP/sfb876-b2.html>

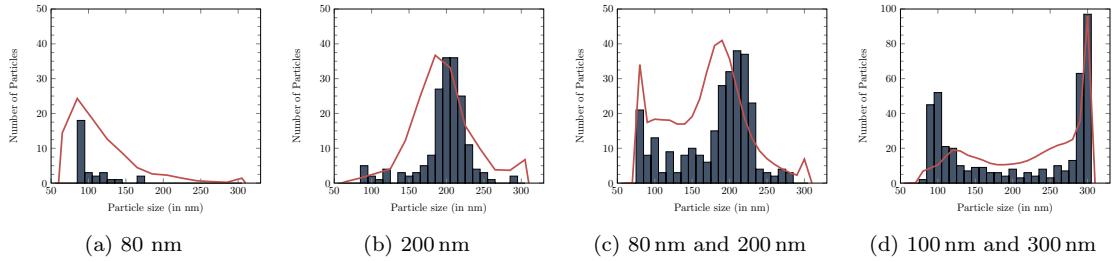
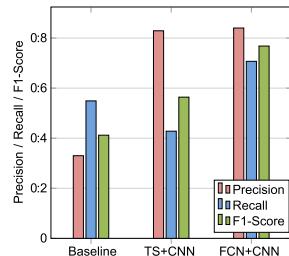


Figure 3.1: Estimated particle size distributions from the output of the neural network compared to reference measurements from an LM10 device (Malvern, UK) shown in red [B2/C1/211].

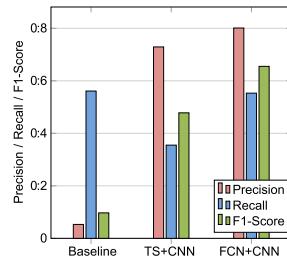
signals with a median SNR of 0.7, cf. figure 3.2. The whole pipeline requires 15.296 ms per frame on an NVIDIA GeForce GTX 1080 Ti. The measurement of individual particle sizes and the compilation of particle size distributions for a given suspension was formulated as an ordered classification problem using classes for particle size intervals and a soft ground truth [B2/208, B2/C1/211]. We were able to estimate size distributions for given suspensions and compare them to reference measurements performed with particle tracking, cf. figure 3.1. This is the first time that particle size distributions were automatically derived using a surface plasmon resonance sensor. Moreover, we developed a training data synthesis procedure that allows training without knowing the exact sizes of particles in image patches [B2/208]. We further used synthetic step functions as training data for the time-series analysis network and compared the results against results obtained with manually labelled real training data, cf. figure 3.2.

**Context-sensitive algorithm and programming paradigms.** The development of context-sensitive algorithms, basically as a cross-sectoral issue, concerned the modelling of degrees of freedom of algorithms and the generic mapping to a concrete use case. For example, the PAMONO sensor can be used in laboratories, at airports, or in hospitals. We developed context-sensitive algorithms for noise reduction, feature extraction, and classification [B2/124]. Moreover, the single-objective GPGPU parameter space exploration (SOG-PSE) approach for platform design with optimisation of resource consumption and the single-objective GPGPU design space exploration (SOG-DSE) approach for optimisation regarding detection performance, processing time, or resource requirements have been developed [B2/124]. As a result, VirusDetectionCL could be optimised towards different scenarios and matching, suitable GPU hardware could be explored. The associated context-sensitive algorithm paradigm is also applied in the created deep resource-aware openCL inference networks (deepRacin) framework [B2/210]. The deep learning library was designed for platform-independent (supports all OpenCL capable GPUs) and parallel execution of data flow graphs. It is especially useful to bring DNNs to mobile GPUs and to make inference computation context-sensitive by providing interfaces for resource-dependent parametrisation. In addition, the focus was on a further generalisation of the examined methods. We developed a method called spline-based convolutional neural networks (SplineCNN) that generalises the convolution operator of CNNs to process irregularly structured and geometric data, e.g., graphs or meshes [B2/A6/177], reaching or exceeding state of the art performance in selected tasks of those fields. Because of the highly relevant scientific questions in connection with graphs that require the development of fundamental methods, the topic will be dealt with further during the next funding period in cooperation with project A6.

**Outlook.** While we were able to improve and extend our qualitative results using deep learning techniques, we identified two topics that demand further research: adaptivity and efficiency. We need robust models for different kinds of setups, methods to adjust existing models to changing



(a) 80 nm data set with a median SNR of 0.716



(b) 80 nm data set with a median SNR of 0.639

Figure 3.2: Detection results evaluated for the baseline method and two different deep neural network methods on two different data sets with low SNR [B2/210].

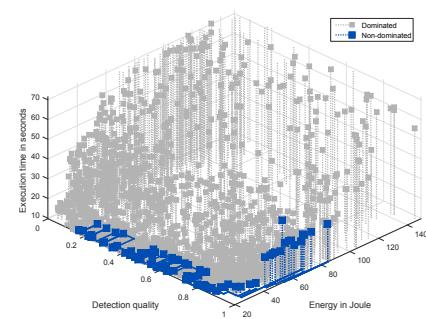


Figure 3.3: Pareto front from an automatic parameter and design space exploration for a mobile virus detection [B2/124].

environments, or a generic training data synthesis procedure that allows an automatic adaptation. Possible solutions for these tasks come from the fields of semi-supervised learning [197], transfer learning [203], and automatic training data synthesis/refinement [199] (combined with our SynOpSis framework). The extension of these concepts for applications with the PAMONO sensor in quality control of medical production processes offers interesting, open research questions.

While very deep CNNs achieved remarkable results for general image processing tasks, we are mainly interested in smaller, more efficient architectures that allow inference computation in soft real-time and on mobile hardware. Those architectures received a lot of attention in recent research [204]. We plan to exploit the flexible nature of the backpropagation algorithm and our background knowledge about the given task to develop more specialised trainable/fixed-function modules in order to further reduce computation complexity. Examples for those highly specialised architectures that make use of task-specific knowledge exist for other tasks [194]. Further interesting research questions concern online parameter optimisation. There is an ongoing cooperation with A3 to combine their model-based optimisation methods with our evolutionary multi-objective optimisation methods [B2/124].

## B. Paradigm development on algorithms and execution levels

This section provides achievements and results regarding paradigm development on algorithms and execution level. The section summarises the results achieved regarding WPs 5, 6, and 7.

**Resource optimisation on program and platform level.** In order to bridge the gap between program and platform layers, methods for a combined hybrid parameter and design space exploration have been developed. Those methods consist of the combined GPU simulator and energy model GPUSimPow, methods for considering offloading, an integration of the context-aware power consumption model (CoPoMo), and the long term evolution (LTE) energy model in cooperation with the A4 project [B2/124]. For the extraction and implementation of pipeline parallelism on resource-restricted embedded systems, the developed methods consist of an OpenMP-inspired infrastructure to efficiently support pipeline parallelism for heterogeneous low-end embedded multiprocessor system-on-chip (MPSoC) systems. Results for OpenMP-based parallelisation executed on an Odroid XU3 show a speed-up of 0.71 up to 3.68 and an energy factor of 1.36 down to 0.28.

**Client/server configurations and collaborative calculation.** Techniques have been derived for analysing schedulability of the tasks that can run in different modes over time to react to the change

in the physical environment. Two types of fixed-priority scheduling in mode change systems have been studied in theory and simulation-based: task-level and mode-level fixed-priority scheduling. An empirical evaluation has been performed and compared to dynamic-priority scheduling on a Raspberry Pi running FreeRTOS. Furthermore, we have developed a set of efficient task partitioning algorithms for suspending tasks [B2/205]. Computation offloading and dynamic voltage and frequency scaling (DVFS) techniques have been adopted to identify the running frequency such that the energy consumption is minimised and the real-time constraints are satisfied. We have applied a middleware approach [B2/206] to make the computation offloading service still reliable in the absence of a fast response from the server and a high-speed connection. The middleware uses an available nearby device either as remote computation unit or as a surrogate to pass the data between the mobile device and the remote server. In addition, an automated computation offloading framework termed distributed openCL (DOCL) has been realised, which is able to dynamically deploy and run existing OpenCL programs on powerful devices in the network [B2/124]. Preliminary evaluations show a speed-up up to 1.9 and energy savings up to 62% for VirusDetectionCL on an Odroid XU4 when compared to local execution. Since the topics are not in the focus of the next funding period, no outlook is provided for this section.

### C. Generalisation and miniaturisation of the PAMONO sensor

This section summarises work that is directly connected to the PAMONO sensor (WPs 8 and 9).

**Generalised nano-object detection.** Concepts to enable the identification of different types of particles, partly on one sensor chip, have been developed, for example, virus-like particles (VLPs) and extracellular vesicles (EV) with diameters of around 120 nm to 140 nm and 100 nm to 115 nm, respectively [B2/C1/211]. The studies, partially carried out in cooperation with project C1, helped to prove the linear dependency between the signal counting rate and the number of nano-vesicles. Model particles were applied for the simultaneous detection on the same PAMONO sensor chip, which showed the sustainability of the newly designed CNN network [B2/210], cf. Section A. Moreover, we studied and improved the capability of the PAMONO sensor. This includes a detection of VLPs in complex biological suspensions (e.g., contaminated solutions) such as serum, the capability and stability of long-time PAMONO measurements, i.e., 20 to 30 minutes, and the sensor surface functionalisation by an application of different self-assembling monolayers. Current specificity of the PAMONO sensor in the measurements of cell-culture-derived VLPs (lacking contaminating content), is around 85% [B2/C1/211]. In ongoing work, we examine how to detect inorganic nano-objects in air samples.

**Miniaturisation and increasing the technical efficiency of the PAMONO sensor.** Reduction of the size of the PAMONO sensor prototype has been achieved by changing the optical scheme from a linear one to a curved one. The total volume of the sensor could be reduced by a factor of four. Moreover, the miniaturised prototype of the PAMONO sensor has been employed to deduce optimal conditions for imaging of nano-objects. Optical engineering principles guided the choice of novel optical elements (CMOS chip, polarising element, etc.) such that the sensitivity of the sensor did not decline due to the minimisation. Further quality control measurements will be performed with non-filtrated complex biological solutions.

**Outlook.** In addition to the fundamental importance, the findings of the current funding period represent ideal preparatory work for quality control of medicinal products in the next funding period, especially to detect different VLPs and EVs. VLPs can boost the immune response without an ability to replicate in cells. Such unique combination of qualities makes VLPs a safer and cheaper platform for vaccine generation than platforms based on inactivated viruses [202]. Microvesicles

(a subgroup of EVs) can carry biologically active compounds (proteins, glycoproteins, etc.), which can influence the immune systems of recipients during blood transfusions. Thus, EVs may serve as a marker for an estimation of blood product immunoreactivity [191].

#### D. Comprehensive optimisation and validation

Regarding the practical use of the PAMONO sensor, concepts for a comprehensive optimisation and validation were examined (WP 10). To solve this task, the multi-objective GPGPU energy-aware design space exploration (MOGEA-DSE) method [B2/124] and the SynOpSis method [B2/212] were combined: SynOpSis has been used to generate training and testing data and to optimise the free parameters of the detection algorithms while MOGEA-DSE has been used to automatically optimise over all relevant objectives of the whole system including the processing device. With evaluation of the methods for various scenarios, a rapid, reliable detection and counting of viruses in PAMONO sensor data have been made possible on high-performance, desktop, laptop, and even hand-held systems [B2/124]. For example, one important result could be achieved for soft real-time processing on the Odroid XU3 device. Figure 3.3 shows the resulting Pareto front and an excerpt of dominated points. For one particularly useful point on the front, 84% energy could be saved compared to a baseline measurement while achieving a speed-up of four and an  $F_1$ -score of 0.995.

**Outlook.** An important prerequisite for using the PAMONO sensor under real conditions is the ability to detect and compensate for sensor errors. However, part of our data consists of very high-dimensional spatio-temporal image sequences that cannot be processed directly with state of the art methods from this field. Assuming that input data and sensor indicators follow a clear distribution for normal behaviour, the goal is to obtain a model for this distribution in order to derive probabilities for data points generated by the running sensor. Then, a low probability value for a new data point is a strong indication of an anomaly. Since the input space is high-dimensional, we need sophisticated methods to model the distribution of standard behaviour. Existing methods for modelling such arbitrary high-dimensional data distributions are generative adversarial networks (GANs) [192] and variational autoencoder [195]. Also, recent work applies GANs for anomaly detection in medical images [198]. Corresponding methods are in focus of the next funding period.

#### Bibliography

- [191] R. Almizraq, J. Seghatchian, and J. Acker. “Extracellular vesicles in transfusion-related immunomodulation and the role of blood component manufacturing”. In: *Transfusion and Apheresis Science* 55 (2016), pp. 281–291 (cit. on p. 199).
- [192] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems* 27. 2014, pp. 2672–2680 (cit. on p. 199).
- [193] P. V. Hornbeck. “Enzyme-Linked Immunosorbent Assays”. In: *Current Protocols in Immunology* 110.2 1 (2015), pp. 1–23 (cit. on p. 205).
- [194] R. Hou, C. Chen, and M. Shah. “Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos”. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017) (cit. on p. 197).
- [195] D. P. Kingma and M. Welling. “Auto-Encoding Variational Bayes”. In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. 2014 (cit. on p. 199).

- [196] S. Racanière, T. Weber, D. Reichert, L. Buesing, A. Guez, D. Jimenez Rezende, A. Puigdomènech Badia, O. Vinyals, N. Heess, et al. “Imagination-Augmented Agents for Deep Reinforcement Learning”. In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 5690–5701 (cit. on p. 206).
- [197] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. “Semi-supervised Learning with Ladder Networks”. In: *Advances in Neural Information Processing Systems 28*. 2015, pp. 3546–3554 (cit. on pp. 197, 204).
- [198] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. “Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery”. In: *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017*. 2017, pp. 146–157 (cit. on pp. 199, 207).
- [199] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. “Learning from Simulated and Unsupervised Images through Adversarial Training”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2242–2251 (cit. on pp. 197, 204).
- [200] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas. “Dueling Network Architectures for Deep reinforcement Learning”. In: *Procs. International Conference on Machine Learning*. Vol. 48. 2016, pp. 1995 –2003 (cit. on p. 206).
- [201] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas. “Sample Efficient Actor-Critic with Experience Replay”. In: *International Conference on Learning Representations (ICLR)* (2017) (cit. on p. 206).
- [202] D. Wetzel, T. Rolf, M. Suckow, A. Kranz, A. Barbian, J.-A. Chan, J. Leitsch, M. Weniger, V. Jenzelewski, et al. “Establishment of a yeast-based VLP platform for antigen presentation”. In: *Microbial Cell Factories* 17.1 (2018) (cit. on p. 198).
- [203] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. “How Transferable Are Features in Deep Neural Networks?” In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*. 2014, pp. 3320–3328 (cit. on pp. 197, 204).
- [204] X. Zhang, X. Zhou, M. Lin, and J. Sun. “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices”. In: *CoRR* abs/1707.01083 (2017) (cit. on p. 197).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [B2/205] Z. Dong, C. Liu, S. Bateni, **K.-H. Chen, J.-J. Chen, G. von der Brüggen, and J. Shi**. “Shared-Resource-Centric Limited Preemptive Scheduling: A Comprehensive Study of Suspension-base Partitioning Approaches”. In: *Proceedings of the 24th IEEE Real-Time and Embedded Technology and Applications Symposium*. 2018 (cit. on pp. 17, 198).
- [B2/206] **A. Toma**, A. Starinow, **J. E. Lenssen**, and **J.-J. Chen**. “Saving Energy for Cloud Applications in Mobile Devices using Nearby Resources”. In: *the 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP 2018)*. Cambridge, UK, Mar. 2018 (cit. on p. 198).
- [B2/207] **D. Siedhoff**, M. Strauch, **V. Shpacovitch**, and D. Merhof. “Unsupervised Data Analysis for Virus Detection with a Surface Plasmon Resonance Sensor”. In: *International Conference on Image Processing Theory, Tools and Applications (IPTA 2017)*. 2017 (cit. on p. 195).

- [B2/208] **J. E. Lenssen, V. Shpacovitch, and F. Weichert.** “Real-Time Virus Size Classification Using Surface Plasmon PAMONO Resonance and Convolutional Neural Networks”. In: *Bildverarbeitung für die Medizin 2017*. 2017, pp. 98–103 (cit. on p. 196).
- [B2/209] **J. E. Lenssen, V. Shpacovitch, D. Siedhoff, P. Libuschewski, R. Hergenröder, and F. Weichert.** “A Review of Nano-Particle Analysis with the PAMONO-Sensor”. In: *Biosensors: Advances and Reviews* (2017), pp. 81–100 (cit. on p. 195).
- [B2/210] **J. E. Lenssen, A. Toma, A. Seibold, V. Shpacovitch, P. Libuschewski, F. Weichert, J.-J. Chen, and R. Hergenröder.** “Real-Time Low SNR Signal Processing for Nanoparticle Analysis with Deep Neural Networks”. In: *International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*. SciTePress, 2018 (cit. on pp. 195–198).
- [B2/A6/177] **M. Fey, J. E. Lenssen, F. Weichert, and H. Müller.** “SplineCNN: Fast Geometric Deep Learning with Continuous B-Spline Kernels”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on pp. 160, 166–168, 196).
- [B2/C1/211] **V. Shpacovitch, I. Sidorenko, J. E. Lenssen, V. Temchura, F. Weichert, H. Müller, K. Überla, A. Zybin, A. Schramm, et al.** “Application of the PAMONO-sensor for Quantification of Microvesicles and Determination of Nanoparticle Size Distribution”. In: *Sensors* 17.2 (2017), pp. 1–14 (cit. on pp. 20, 196, 198, 287).

### b) Other publications

- [B2/212] **D. Siedhoff.** “A Parameter-Optimizing Model-Based Approach to the Analysis of Low-SNR Image Sequences for Biological Virus Detection”. Diss. Dortmund, Germany: TU Dortmund, 2016 (cit. on pp. 195, 199).
- [B2/124] **P. Libuschewski.** “Exploration of Cyber-Physical Systems for GPGPU Computer Vision-Based Detection of Biological Viruses”. Diss. Dortmund, Germany: TU Dortmund, 2017 (cit. on pp. 130, 195–199).

## 3.4 Project plan

### Goals

The long-term goal of the project is to explore and analyse methods for the recognition of specific nano-objects in data-intensive and very noisy or artefact afflicted image sequences from a sensor system that is characterised by several aggravating circumstances. The processing should be performed in compliance with detection quality standards, soft real-time constraints, and low resource consumption (e.g., energy, memory, maintenance demands, and space) of the sensor and the calculation platform. Meeting these requirements requires the identification of trade-offs between the different criteria. The centre of attention is the PAMONO technology, a specific sensor system developed to achieve these goals.

During the third phase of the project, we will develop algorithms, methods, and technical concepts with respect to robustness and applicability to identify configurations for a fast, user-friendly sensor software adaptation. We aim to achieve a methodical and technical development such that the PAMONO technology can be used for pharmaceutical quality control and in-process control of

medical and biological processes covering viruses (PEI) and VLPs (ARTES). Therefore, project B2 is to be continued as a transfer project.

The application partners are involved in methodological development, taking into account the relevant fields of application, but are especially responsible for the possibility of subsequent commercial use. They will utilise the PAMONO technology and its application to draw the attention of commercial manufacturers of diagnostic devices and analytical instruments. In this context, application-centred publications (white papers, application notes, etc.) and product-focused scientific contributions are means to increase public attention. In close cooperative collaboration with the application partners, we will focus on three important application scenarios:

1. **Quality control in the production of vaccines:** The development of vaccines, in particular of novel VLP-based vaccines, is a complex biotechnological process. ARTES actively commercialises the development and production of chimeric VLPs for vaccination purposes. For this purpose, the PAMONO technology could open up completely new possibilities for online and at-line process-analytical quality control. ARTES has development collaborations with several pharmaceutical companies developing human and veterinary vaccines. The PAMONO technology can be further exploited and promoted in the scope of these collaborations. Key challenges for the use of the sensor for quality control include the ability to detect small nano-vesicles, i.e., 60 nm to 70 nm, and to fulfil requirements for self-calibration and automatic (in-)process control. ARTES will have a prototype of the PAMONO technology at its disposal, which will regularly be adapted to the current state of development. This allows ARTES to actively participate in the further development of the PAMONO technology and to use the current prototype for quality control in their production of vaccines. ARTES will conduct experimental evaluations as well as provide corresponding custom-made VLP samples and antibodies.
2. **Quality control of blood donations and blood products:** Considering the importance of blood components for transfusion, a quick and efficient quality control of those components is especially important for diagnosis in early stages of infections and for detection of new virus species. Since current test methods do not fulfil the requirements for those tasks, the PAMONO technology might open a new field in quality control, namely immunological quality control and, ideally, allows verification of immunological efficacy. Here, the PEI is responsible for verifying the suitability of the technology for (clinical) routine screening. It will provide information on relevant viruses and required detection sensitivities. Moreover, PEI will present the results of the transfer project in national, European, and global bodies, e.g., in the WHO, and will establish contacts with manufacturers of blood products and blood donation services. Challenges in data processing are the ability to deal with the small volume of samples that contain only a low number of vesicles as well as enabling high-throughput detection with high specificity and adaptability.
3. **Quality control of vaccines and blood on site or in simple, decentralised facilities:** Besides clinical and laboratory environments, the PAMONO technology could also be used under non-controlled conditions. Examples are an on-site quality control of conventional vaccines, e.g., to detect degradation of vaccines in developing countries, and the examination of blood products, e.g., in crisis regions or in emergency situations. Important prerequisites for further development of the technology are a mobile and easy-to-use system, online function control, and online management during operation under resource limitations. The boundary conditions and important quality measures will be defined by the application partners.

The challenges and possible solutions in the given scenarios are a part of the *One Health* concepts of the WHO and aim to fight antimicrobial resistance as well as the rapid spreading of virus diseases. Besides the medical application, future areas of application of the PAMONO technology are to detect soot particles in the air (e.g., from car exhaust gases) and in the area of environmental control (e.g., waste water) for the identification of excreted pathogenic viruses. In addition to the objectives of the previous funding periods, the application scenarios introduce new objectives in terms of image analysis, system environment, and PAMONO technology. The associated methodological

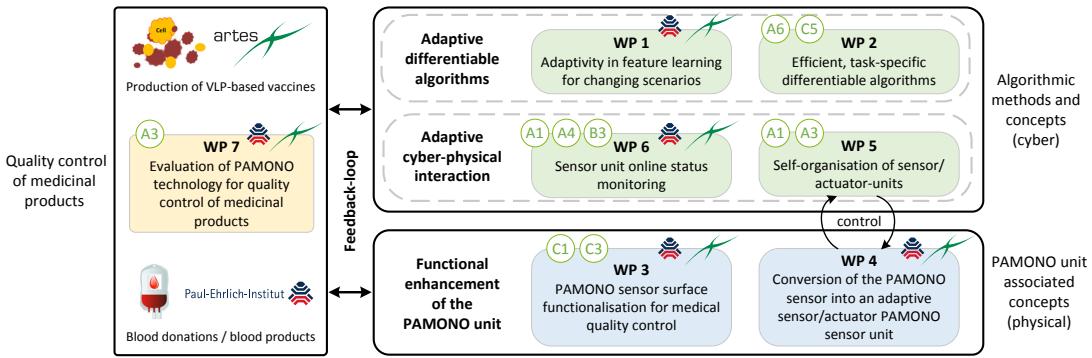


Figure 3.4: Schematic overview of the work packages, their interdependencies, and the cooperations within the CRC. Colours indicate the lead within the work packages in relation to the PIs (■ Hergenröder, ■ Weichert, ■ Hergenröder, Weichert, and application partners).

and technical developments can be classified into the following domains as shown in figure 3.4:

- A. Adaptive differentiable algorithms:** This domain will focus on the development of adaptive and resource-efficient differentiable data analysis algorithms (e.g., specialised CNNs) for processes in quality control of medicinal products. The algorithms should be parametrisable to handle changing environmental conditions (e.g., sample composition, matrix composition, and temperature) and diverse characteristics of images (e.g., noisy and anomalous images) under resource constraints and in (soft) real-time, cf. WPs 1 and 2.
- B. Adaptive cyber-physical interaction:** The application of PAMONO technology for quality control issues during production of VLP-based vaccines and blood products requires maximal possible automation of tilling and tuning processes. Therefore, a challenge is the development of algorithmic solutions for online status monitoring and parameter adaptation of the PAMONO sensor during operation, cf. WPs 5 and 6.
- C. Functional enhancement of the PAMONO unit:** The goal of this domain is the development and technical realisation of novel approaches for PAMONO sensor surface functionalisation as well as the conversion of the PAMONO sensor to a parametrisable sensor/actuator unit, which can be controlled by online adaptation algorithms, cf. WPs 3 and 4.

The success of the individual domains and the overall vision will be evaluated in close cooperation with the application partners, cf. WP 7. The main objectives of the evaluation are to identify suitable parameter settings and to analyse the relevance, impact, and sustainability of the PAMONO technology for quality control in the mentioned fields of application. The feedback provided by the application partners will be used to continuously improve the methods from WPs 1 to 6.

## Work schedule

**Work package 1. Ensuring adaptivity for changing conditions:** The integration of data analysis algorithms into processes for medicinal product quality control introduces the challenge of dealing with different degrees of freedom. Examples are (moderately) altering environmental conditions (e.g., sample composition), variation in particle size (even smaller than the previous limit of 80 nm to 100 nm), and slightly changing tasks like the simultaneous analysis of multiple (at least two) nano-vesicle protein markers on the surface of the same vesicle. Therefore, an issue will be the applicability of the algorithms developed in the previous funding phases as well as to analyse the necessity and ability of those algorithms to adapt to those degrees of freedom. Using feedback from the application partners, adaptive methods and possible optimisations will be explored.

A special focus will be on the development of deep learning algorithms. In current research, DNNs are mostly trained specifically for one setup where they outperform competing methods. We have shown that this is also the case for our image-processing tasks. However, trained neural network models are quite sensitive to concept drift introduced by changing conditions. Therefore, when it comes to the task of bringing deep learning to real-world applications, increasing the adaptivity of trained models is of utter importance. Concerning this topic, a large amount of theoretical research has been performed recently, which will be leveraged to develop practical methods for application in the PAMONO sensor data analysis pipeline.

The first task is the development of novel methods for transferring knowledge from trained models to changing experiment setups and image characteristics. This can either be achieved by mechanisms for efficient online transfer learning, making use of pre-trained models and a few new, labelled training examples, or by one-shot and semi-supervised learning approaches [203, 197]. The possibility of an efficient, adaptive training on mobile sensor hardware will be explored, and a new deep learning framework capable of training on mobile GPU hardware needs to be developed.

The second task is the synthesis of training data for deep feature learning. The SynOpSis framework from the last phase will be applied to create a novel generator for labelled training data. In addition, the applicability of GANs for synthetic training data refinement will be analysed in respect to its impact on model quality [199].

Finally, a study will be conducted (WP 8), evaluating the developed algorithms and methods with respect to robustness and applicability, identifying the most feasible methods for fast sensor software adaptation in the field. This is to be achieved by using feedback from the application partners to gather data regarding real-world application of the PAMONO sensor.

**Work package 2. Efficient, task-specific differentiable algorithms:** The goal of this work package is the development of highly specialised and efficient architectures for DNNs that allow the sensor data analysis pipeline to be efficiently applied on mobile hardware.

In current research, DNNs are seen mostly as black boxes that approximate general image processing functions for arbitrary tasks. We propose an alternative paradigm for designing task-specific CNNs: instead of using very deep and general architectures, we will heavily leverage background knowledge about the given task (e.g., signal model and data characteristics). We identify parts of our deep architectures that learn already known information about the data, which we alternatively can provide to the model in a more direct way. Those redundant trainable model parts will be replaced by newly created fixed-function modules, functions with stronger constraints, or more restrictive in-network routing. We expect those modified architectures (1) to be more resource-saving, (2) to require less training data, and (3) to provide regularisation while preserving the previous levels of accuracy. An additional goal is to reduce the amount of input data from the temporal domain that is required to reliably perform detection for a single frame. Fixed data-aggregation methods, motivated by task-specific knowledge, will be developed and applied before feeding the data into the model to heavily reduce the number of operations.

In order to still provide end-to-end trainable architectures that can be used in conjunction with the backpropagation algorithm, we aim to develop those modules while preserving the differentiability of all modelled functions. Creating those parametrised, end-to-end trainable pipelines that consist of differentiable algorithms is summarised by the term *differentiable programming*.

A part of our research on this topic is the ongoing cooperation with project A6 where we develop and implement efficient, task-specific DNN modules for irregularly structured inputs, generalising the feature learning capabilities of DNNs to new domains. In addition, we will provide differentiable modules for analysis of irregular structured particle accelerator data, which is analysed by project

C5. Depending on our findings with those task-specific, differentiable architectures, we plan to develop a framework for custom neural network modules that is able to categorise, evaluate, and rate those modules with respect to specified criteria. Also, efficient GPU algorithms for the created modules will be developed and integrated into our deepRacin framework.

**Work package 3. PAMONO sensor surface functionalisation for medical quality control:** The objective is to in principle extend the functionalisation of the gold surface within the PAMONO sensor, on which the actual binding of particles occurs, so that the PAMONO sensor can be used for quality control of medicinal products. If synthetic (VLPs) and biological (EVs) targets are analysed, the overall sensor sensitivity and selectivity of the PAMONO sensor can be negatively affected by contaminating biological substances in the samples. Therefore, we want to develop a special buffer composition for the sensor surface to reduce the binding of these undesirable compounds. This *blocking procedure* will be established experimentally by taking theoretical concepts into account [193]. It needs to be carefully adjusted to not significantly affect the sensitivity of the PAMONO sensor. Another important task is to enable reversible binding of captured EVs and VLPs. Reversible binding is a necessary condition to allow a long-term application. It is not yet known how often the gold surface can be reused without a serious loss of surface quality. Moreover, the quality of VLP and EV quantification and size estimation also can be affected by reusing surfaces. This requires further validation. We plan to further adjust the surface functionalisation to match requirements for long-term measurements. Both the changed stability of the detected signals due to the blocking procedure and the dynamic deterioration of the surface during long-term measurements are new challenges for the image analysis in WPs 1 and 2.

Another task is a functionalisation of the surface to allow the simultaneous analysis of multiple (at least two) nano-vesicle protein markers of the same vesicle. For this purpose, we will examine multiple patches with a differing binding capability and derive the individual distributions by combinatorial reasoning. A particular challenge is to design appropriate surface functionalisation spots and to change the flow cell unit to allow measurements of vaccine-carrying VLPs and of EVs from blood products. Purified VLP samples will be provided by ARTES. Further goals include the development of a novel surface functionalisation for multi-marker analysis of EVs derived in tumour cells in cooperation with project C1.

Moreover, it should also be noted that a significant part of a recorded sensor surface is optically disturbed due to the inherent and necessary folded and tilted optical designs. An enlargement of the effective sensor area, i.e., the area of the gold surface that provides signals distinguishable from noise, can help to solve the problem of detecting low-concentrations of nano-vesicles (approx. 1,000 nano-vesicles per  $\mu\text{L}$ ). This is especially crucial for blood products (PEI). Therefore, we will investigate different approaches to enlarge the effective nano-object detection area of the PAMONO sensor without increasing the sensor prototype size, e.g., manipulations of the optical system. It is planned to develop the algorithms in cooperation with the C3 project.

**Work package 4. Conversion of the PAMONO sensor into an adaptive sensor/actuator PAMONO sensor unit:** The objective of this work package is the development of a novel adaptive sensor/actuator unit to provide higher automation of tilling and tuning processes for quality control issues during production. Previously manually set controllers and controller actuating variables of the PAMONO sensor are to be replaced by dynamically controllable actuators. This work package focuses on the technical realisation of those actuators, while WP 5 focuses on the development of the required algorithmic concepts. The automatic adaptation through actuators is a basic requirement to minimise the influence of human factors on the results of PAMONO measurements. Examples of processes that should be automatically controlled by actuators are (1) search of the resonance angle (controlled by the alignment of the incident angle of the laser light to the gold

surface) to minimise the laser light reflectivity, (2) focusing of the scattered plasmon-signal, controlled by the relative position of camera and imaging lens, (3) reduction of background scattering, controlled by the position of the imaging lens relative to an aperture, and (4) flow velocity in the fluidic system. In addition, a vibration sensor and a temperature sensor will be added to detect the status of the PAMONO sensor.

Automatic adjustment of those process variables will be especially effective in combination with novel optical concepts for scattered background reduction, which were recently developed by the ISAS and validated in proof-of-principle experiments. Those concepts allow the detection and quantification of even smaller VLPs around 60 nm to 70 nm (having a refractive index close to that of the buffer), which are often used for vaccine generation (e.g., by ARTES), but need time-consuming, tedious alignment procedures, which can be automated using the developed actuators.

Another challenge is the integration of a computer-operated piezoelectric pump as a substitute for the peristaltic pump in the current fluid system to further miniaturise the sensor prototype. Moreover, the introduction of such a pump is expected to reduce the volume of sample required for the PAMONO measurements to less than 100  $\mu\text{L}$ . This issue is especially important for the control of blood product manufacturing based on the detection of EVs in samples. However, an integration of such a pump in the fluidic system may influence the characteristic of the vesicle binding signal in the recorded images. These changes pose a further challenge for WPs 1 and 7.

**Work package 5. Self-organisation of sensor/actuator-units:** The goal of this work package is to complement WP 4 and develop an algorithmic solution for self-configuration and self-adaptation of biomedical sensor/actuator units. Automatic adjustment of the involved actuators be based on the new sensors (WP 4), optical data (image sequences), and supplementary metadata for optimal image-based recognition of the nano-object bindings. We divide the work package into the tasks of self-configuration and self-adaptation. Self-configuration concerns the automatic calibration of actuators in advance of each measurement series. Reference measurements on a special sensor surface area, the known binding characteristics, and environmental conditions will be used to develop a model that derives optimal actions and settings. The methods for online status monitoring (WP 6) are used to detect functional errors. An open issue is whether the changing functionalisation can be considered in the self-configuration. The binding characteristics of the reference particles are assumed to deviate from those of the real particles. As a result, methods for self-configuration may involve degrees of uncertainty or decisions based on incomplete information.

The second aspect concerns self-adaptation, the automatic adjustment during operation based on changed internal (e.g., probe characteristics), and external (e.g., temperature changes) properties. The goal is to distinguish between *unusual* and *expected* image changes. Substantial preparatory work is available and is a part of the WPs 1 and 7. Present solutions operate with limited autonomy and are mostly restricted to non-visual assessments. Challenges are real-time control, the amount and dimensionality of data to process, and different solutions (adjustment possibilities) to the same problem set (image representation), respectively. Therefore, we also want to examine approaches of contextual learning from demonstrations. The algorithm to be developed represents a cognitive digital twin, which monitors and collects specific parameters, statuses, and contents of interest from the image sensor as well as the supplementary sensors. Furthermore, it makes performance-focused decisions and may be self-improving. We want to study strategies such as model-free reinforcement learning (e.g., duelling network architectures for deep reinforcement learning) [200], or actor-critic algorithms [201]. Moreover, we want to examine hybrid architectures by combining model-free and model-based aspects (e.g., imagination-augmented agents) [196]. Uncertainties and concept drift exist in both the input (sensors data) and the output (actuator settings) of the sensor/actuator unit. Therefore, we want to develop compensatory methods in cooperation with the A3 project. For a fast self-organisation of the sensor/actuator units, dedicated neural network processors are useful. Neural network processors that were previously only available in clusters or in GPUs,

like Google's tensor processing unit (TPU), Intel's Loihi chip, or NVIDIA's Tensor Cores, now appear in mobile devices. This will speed up or even enable new application possibilities on mobile devices. Huawei's neural-network processing unit (NPU) in the Kirin 970 is one example for such a mobile processor that might be incorporated into the (mobile) PAMONO sensor unit. This will be examined in cooperation with the A1 project.

**Work package 6. Sensor unit online status monitoring:** The goal of this work package is to develop an online solution for an automatic monitoring of the PAMONO sensor unit and the sensor/actuator unit status based on image sequences. Automatic self-testing of correct functioning is a basic requirement for using the PAMONO sensor to check blood products or medical production processes (vaccines).

First, the online status monitoring should enable a distinction between normal images (error-free status of the sensor) and anomalous images (faulty status). For this purpose, the normal unit behaviour, consisting of information on the image sensor (image data), the supplementary sensors (e.g., temperature and vibration), and the actuators (e.g., angle of incidence of the laser light) as well as metadata (e.g., type of nano-object and age of the gold surface) will be modelled in order to detect outliers from the standard behaviour (cf. WP 1). Anomaly detection and online monitoring have been and will be subjects of projects A1, A3, and B3. However, the high-dimensional nature of image data provides new challenges and requires more sophisticated and efficient data modelling methods. Therefore, a method will be developed that is capable of describing such high-dimensional data and is able to infer confidence values for normal behaviour, building on existing work of projects A1 and B3. Current research suggests that this method will make use of convolutional GANs, which have shown to be capable of modelling high-dimensional image-based data distributions for anomaly detection [198]. Alternative methods from current literature will be explored in comparison.

Then, based on the model for standard behaviour, the goal is to develop a system for online sensor status monitoring, which provides real-time feedback about the sensor/actuator system state. The possibility to develop state classification methods will be explored where state classification refers to distinguishing between different classes of known/unknown and error/non-error sensor states. Different states may indicate required user or actuator interaction and give an estimate of the expected result quality. It should be noted that we also intend to classify errors in actuator states, which poses a special challenge since error classification results should be directly reused for actuator control in real-time (cf. WP 6). Handling this kind of semantic error classification involves entering new research fields. Moreover, the online monitoring system can be exploited to provide trigger inputs for the robust image and data analysis methods that are developed in WP 1. The findings could also be relevant for further projects within the CRC that are dealing with process chains or complex networks, e.g., projects A4 and B3.

**Work package 7. Evaluation of the PAMONO technology for quality control of medicinal products:** The goals of the work package are evaluation and optimisation of the PAMONO technology with respect to requirements provided by the application partners: (1) characterisation and quality control of purified VLP preparations, (2) process-analytical control of purification step(s) for VLP, and (3) detection of virus particles and exosomes for quality control of blood products.

For evaluation, experiment and working scenarios (in the lab/in the field) and test data for those scenarios need to be created. Experiment scenarios will cover all aspects of the sensor/actuator PAMONO unit: the optical sensor including surface functionalisation, supplementary sensors (e.g., temperature), actuators, algorithmic components for an automatic adjustment of the involved actuators, and automatic data processing. Then, various tests will be implemented and carried out

## Project B2

(e.g., by software/hardware-in-the-loop tests) to analyse the fulfilment of the requirements. These requirements can include response time, resource consumption, reliability, adaptability, and inherent interaction with the system environment. Methods for cross-validation (e.g., Direct ELISA) will be also provided by the application partners. We will develop a holistic validation approach that allows for a flexible way of assessing the system-level aspects by various types of experiments (virtual, real, and mixed settings). Such a real-time hybrid structural testing depends on tests for analysing the component and performance reliability of the internal factors (hardware and algorithms). In a second step, external factors (e.g., environment) are also included.

Another aspect is to evaluate the suitability of the analytical methods for pharmaceutical application, immunological quality control, and quality control of blood products. Currently, there are no reference standards for VLPs and EVs available. There is also no high-throughput method for detecting virus infections at an early stage under resource-limited conditions. Initially, various mixtures of natural biological nano-vesicles (e.g., HIV-VLPs) and silica nanoparticles are prepared. Silica nanoparticles are suited to optically mimic VLPs (range of 60 nm to 120 nm) due to a similar refractive index. Further, biological contaminants (e.g., protein aggregates), as well as different, purified vaccine-carrying VLPs or target vaccine-carrying VLPs, will be provided by ARTES.

Moreover, optimisation will be done for finding the suitable parameter settings with respect to the requirements. During the current funding period, different techniques for global optimisation have been developed, namely, our MOGEA-DSE and SynOpSis methods and the model-based optimisation methods by project A3. Advances in this field will be leveraged and adapted for new systems that need optimisation. An example is the context-aware optimisation of DNN inference algorithms, which are given in the developed deepRacin framework. In a final step, it is planned to estimate critical parameters (e.g., concentration value or degree of contamination) that affect the PAMONO analysis of vaccine-carrying VLP in selected matrices or the analysis of EVs from blood products under stationary and “stress” conditions.

### Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Ensuring adaptivity for changing conditions:																	
2. Efficient, task-specific differentiable algorithms:																	
3. PAMONO sensor surface functionalisation for medical quality control:																	
4. Conversion of the PAMONO sensor into an adaptive sensor/actuator PAMONO sensor unit:																	
5. Self-organisation of sensor/actuator-units:																	
6. Sensor unit online status monitoring:																	
7. Evaluation of the PAMONO technology for quality control of medicinal products:																	

### 3.5 Role within the Collaborative Research Centre

The overall goal of project B2 for the third funding period is the development of algorithmic and technical solutions to enable the PAMONO technology to achieve a sustainable improvement in the medical quality control. In addition to providing solutions for the biophysical challenges, another important goal is to develop deep learning architectures for the analysis of image data under changing environmental conditions, limited resources, and real-time requirements. In accordance with the objectives, the project B2 is involved in both the CRC research group “resource constraints” and the research group “data analysis”. Since we deal with fundamental questions in the area of algorithms, there is extensive cooperation with the A-projects. This is strengthened by Jian-Jia Chen, who was a project leader in B2 in the second funding period and will continue as a project leader of projects A1 and A3 in the third funding period.

A1 examines the relationship between computing architecture and machine learning at a theoretical level. One objective is to investigate resource-aware structures for DNNs. B2 examines this issue for the sensor data analysis pipeline under real-life conditions. We want to develop highly specialised and efficient architectures for DNNs that could be efficiently applied on mobile hardware. With regard to A3, extensive collaborative work has already been carried out in the second funding period on global optimisation, namely, the MOGEA-DSE and SynOpSis methods (B2) and model-based optimisation methods (A3). The cooperation is to be continued with new research questions, such as context-aware optimisation of parameters of DNNs inference algorithms. In addition, methods for recognising and compensating for concept drift characteristics are jointly developed. Further, B2 will integrate the communication models of A4 while the interconnection of PAMONO sensors represents a test case for the resource-aware heterogeneous platforms of A4. A significant part of the research of B2 benefits from an ongoing cooperation with project A6. We will realise efficient, task-specific DNN modules for various issues in relation to irregularly structured inputs such as graphs.

In addition to cooperation with the A-projects, all projects that process sensor data can also benefit from the findings of B2. This provides connections to the B- and C-projects. The collaborations cover the consideration of individual project requirements and the provision of CNNs architectures. One objective of B3 is enhancing the control of industrial processes and the quality of products using machine learning methods based on sensor data. In B2 we want to examine DNN architectures for anomaly detection and online monitoring. Anomaly detection and online monitoring have been and will be subjects of projects A1, A3, and B3, too. Moreover, the methods for sensor data analysis and anomaly detection of B2 can be used to examine the high-dimensional and high-frequent data from (astroparticle) physics and astronomy, which are the subjects of projects C3 and C5. In cooperation with the C3 project, we want to develop the algorithms for enlarging the detection field of the PAMONO sensor and to compensate for the optical distortions. There is also an ongoing cooperation with project C1. We want to develop novel surface functionalisation for multi-marker analysis of EVs derived in tumour cells, a subject of project C1. Thus, the PAMONO technology could function as a monitoring technique for liquid biopsies.

### 3.6 Differentiation from other funded projects

#### **ADJUTANT**

**(Weichert, Reference number AiF: ZF4119002DB7) (Funding period: 2018–2020)**

The goal of the project is the development of a robot-based inspection system for the automated non-destructive testing of lightweight components made of fibre plastic composites. Essential issues are the development of multi-criteria optimisation algorithms for path planning based on implicit modelling and a reliable safety concept for collaborative human-robot interaction.

### InÜDosS

**(Weichert, Reference number FOSTA: P 1326/17/2018) (Funding period: 2018–2020)**

The aim of the project is automated condition monitoring of steel structures by Unmanned Aerial Vehicles to develop methods for automated data analysis and the classification of damaged structures. As a result, the project should enable the creation of a building file in digital form with damage and defect reports as well as instructions for handling.

### CuKa

**(Weichert, Reference number DFG: WE 5036/4-1) (Funding period: 2018–2021)**

The main objective of the research project is to lay the foundations for the provision of a repository/exhibitor spanning both domain-internal (2D-photographs, 3D-scans) and a cross-domain search functions for finding characters and text passages as well as for operationalising the analysis of the cuneiform script.

### AntiThromb

**(Hergenröder, Reference number BMBF: KMU-NetC 03VNE2091F) (Funding period: 2018–2020)**

The task of this project is the development and pre-clinical evaluation of a new coated vascular implant with anti-thrombogetic effect to be used in the case of haemorrhagic stroke and other neurovascular diseases. The new implant will be hem compatible and possibly endothelial positive responding.

### NanoFilter

**(Hergenröder, Reference number DAAD: 57403573) (Funding period: 2018–2018)**

Development of polymeric surfaces with anti-fouling and anti-bacterial capabilities. The first application will be nano-filters for water purification in dry countries such as for instance Jordan.

### MRT-Filter

**(Hergenröder, Reference number DFG: LA 1134/8-1) (Funding period: 2017–2020)**

The task is the development of new spectroscopic imaging techniques for high-resolution NMR allowing measurement of the metabolic distribution pattern and concentration profiles of metabolites in 3D cell culture models such as organoids or spheroids.

## 3.7 Project funding

### 3.7.1 Previous funding

The project has been funded within the Collaborative Research Centre since January 2011.

### 3.7.2 Contribution of the application partner towards the new funding period

	2019	2020	2021	2022
Staff (in hours per week)	10.50	10.50	10.50	10.50
Funding for direct costs	15,400	15,400	15,400	15,400
Instrumentation	0	0	0	0

(Figures on direct costs and instrumentation in euros)

Instrumentation up to 10,000 euros, software and supplies for financial years 2019–2022

Antibodies, MAb 7C12 specific against duck hepatitis envelope protein (VLP surface marker A)	EUR	300
Antibodies, antibody against chimeric VLP surface marker B	EUR	2,500
VLPs, purified, homogeneous duck hepatitis envelope protein VLP (surface marker A)	EUR	1,300
VLPs, purified, chimeric VLP (surface markers A and B)	EUR	5,000
VLPs, purified, homogeneous capsid type VLP (capsid marker C)	EUR	5,000
Biological contaminant samples, samples of protein, and lipoprotein contaminating crude VLP preparations	EUR	1,300

(All figures in euros)

Description of instrumentation	Year of purchase	Cost of purchase
Shimadsu, LC20   HPLC system (SEC, UV detection) for VLP characterisation	2008	21.2
Beckman, DelsaMax Core   light scattering system for nanoparticle characterisation	2014	26.1
BioRad, ChemDoc MP   gel/blot imaging and evaluation system for VLP characterisation	2016	28.7
Tecan, Genios   microtiter plate ELISA reader for VLP characterisation	2003	16.7
GE Health Care, Äkta Purifier   chromatography system for VLP purification	2003	25.2
Beckman, Optima L-90K   Ultracentrifuge for VLP purification/characterisation	2011	25.9

(All figures in thousands of euros)

### 3.7.3 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	1	64,500	1	64,500	1	64,500	1	64,500
Postdoctoral researchers, 100 %	1	69,900	1	69,900	1	69,900	1	69,900
Total	—	134,400	—	134,400	—	134,400	—	134,400
Direct costs	Sum		Sum		Sum		Sum	
Instrumentation up to 10,000 euros, software and supplies	24,000		6,000		6,000		6,000	
Total	24,000		6,000		6,000		6,000	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	158,400		140,400		140,400		140,400	

(All figures in euros)

### 3.7.4 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Roland Hergenröder, Dr., postdoctoral researcher	Biomedical analytics	Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.	9	—	Existing funds
	2	Frank Weichert, Dr., postdoctoral researcher	Visual computing	TU Dortmund	9	—	Existing funds
Non-research staff	3	Maria Becker, non-research staff	—	Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.	10	—	Existing funds
	4	Heike Rapp, secretary	—	TU Dortmund	2	—	Existing funds
<b>Application partner staff</b>							
Research staff	5	Winfried Kammer, Dr., postdoctoral researcher	Quality of blood products	Paul-Ehrlich-Institute	1.5	—	—
	6	Dorothea Stahl, PD Dr., postdoctoral researcher	Quality of blood products	Paul-Ehrlich-Institute	2.5	—	—
	7	Volker Jenzelewski, Dipl.-Biol., doctoral researcher	Biotechnology	ARTES Biotechnology GmbH	2	—	—
	8	Theresa Rolf, M.Sc., doctoral researcher	Biotechnology	ARTES Biotechnology GmbH	2	—	—
	9	N.N., rotating position	Quality of blood products	Paul-Ehrlich-Institute	1	—	—
Non-research staff	10	N.N., non-research staff	—	ARTES Biotechnology GmbH	1.5	—	—

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Requested staff</b>							
Research staff	11	Victoria Shpacovitch, Dr., postdoctoral researcher	Biomedical analytics	Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.	—	Postdoctoral researcher	—
	12	Jan Eric Lenssen, M.Sc., doctoral researcher	Visual computing	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):****1. Hergenröder, Roland**

Project management and conceptual planning at the ISAS. Focus on the functional enhancement of the PAMONO unit. Mainly responsible for WPs 3 and 4, additional cooperation in WPs 1, 5, 6 and 7.

**2. Weichert, Frank**

Project management and coordination of the participants. Focus on adaptive differentiable algorithms and algorithms for adaptive cyber-physical interaction. Mainly responsible for WPs 1, 2, 5 and 6, additional cooperation in WPs 4 and 7.

**3. Becker, Maria**

Setup of experiments in the laboratory.

**4. Rapp, Heike**

Document and appointment management and, if necessary, travel planning and accounting.

**Job descriptions of staff for the proposed funding period (supported by application partner):****5. Kammer, Winfried**

Scientific support on regulatory issues of blood product testing. Cooperation in the WPs 4, 6 and 7.

**6. Stahl, Dorothea**

Acts as principal coordinator for the project at PEI and supervises scientifically. Cooperation in the WPs 1, 3, 4, 6 and 7.

**7. Jenzelewski, Volker**

Acts as principal coordinator for the project at ARTES and supervises scientifical and technical/administrative project personal at ARTES. Cooperation in the WPs 1, 3, 4, 6 and 7.

**8. Rolf, Theresa**

Characterisation of purified VLP samples, cross validation studies, experimental planning to establish sensor unit measuring system at ARTES, sensor evaluation for QC purposes in VLP development and manufacturing. Cooperation in the WPs 3, 4 and 7.

**9. N.N.**

Scientific support on blood product testing, especially for new pathogens. Cooperation in the WPs 1, 3 and 7.

**10. N.N.**

Reproduction studies, measurement of purified VLP samples and treated VLP samples, and data evaluation. Cooperation in the WPs 1, 6 and 7.

**Job descriptions of staff for the proposed funding period (requested funding):****11. Shpacovitch, Victoria**

Head of the laboratory, processing of all biophysical and technical works. Cooperation in the WPs 1, 3, 4, 5 and 7.

**12. Lenssen, Jan Eric**

Development of Architectures for Deep Neural Network. Cooperation in the WPs 1, 2, 5, 6 and 7.

### 3.7.5 Requested funding for direct costs for the new funding period

	2019	2020	2021	2022
Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V.: existing funds from ISAS	6,000	6,000	6,000	6,000
TU Dortmund: existing funds from University	1,500	1,500	1,500	1,500
Application partner	15,400	15,400	15,400	15,400
Sum of existing funds	22,900	22,900	22,900	22,900
Sum of requested funds	24,000	6,000	6,000	6,000

(All figures in euros)

Instrumentation up to 10,000 euros, software and supplies for financial years 2019–2022

Consumables for biological (antibodies, cell culture media, blood) and optical work (gold plates, nanoparticle standards)	EUR	6,000
---	-----	-------

Instrumentation up to 10,000 euros, software and supplies for financial year 2019

Optical equipment (lenses, camera) for two PAMONO sensors	EUR	18,000
---	-----	--------

### 3.7.6 Requested funding for major research instrumentation for the new funding period

This project does not request any funding for major research instrumentation.

### 3.1 General information about Project B3

### 3.1.1 Project title:

Data Mining on Sensor Data of Automated Processes

### 3.1.2 Research area(s):

409-05 (Interactive and Intelligent Systems), 401-06 (Production Automation)

### 3.1.3 Principal investigator(s)

Deuse, Jochen, Prof. Dr.-Ing., 29.12.1967, German

Institut für Produktionssysteme, Professur Arbeits- und Produktionssysteme, Fakultät Maschinenbau, Technische Universität Dortmund

Dortmund  
Leonhard-Euler-Straße 5  
44227 Dortmund

Phone: 0231-755-2652

E-mail: jochen.deuse@tu-dortmund.de

Morik, Katharina, Prof. Dr., 14.10.1954, German

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

Phone: 0231-755-5100  
E-mail: katharina.morik@tu-dortmund.de

Wiederkehr (née Kersting), Petra, Prof. Dr.-Ing., 17.12.1980,  
German

LS 14, Fakultät für Informatik, Technische Universität Dortmund  
Otto-Hahn-Straße 12  
44227 Dortmund

Phone: 0231-755-7208  
E-mail: petra.wiederkehr@tu-dortmund.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

( ) no (x) yes

Do any of the above mentioned persons hold fixed-term positions?

(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes    (x) no
2.	clinical trials	( ) yes    (x) no
3.	experiments involving vertebrates.	( ) yes    (x) no
4.	experiments involving recombinant DNA.	( ) yes    (x) no
5.	research involving human embryonic stem cells.	( ) yes    (x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes    (x) no

## 3.2 Summary

Enhancing the control of industrial processes and the quality of products can be supported by learning from sensor data. Project B3 focuses on the investigation of how decentralised data mining can be used for real-time quality predictions and how it can be integrated into production processes.

The tasks of data analytics in production systems have been ordered according to their difficulty: (1) anomaly detection, (2) diagnostic analytics, (3) predictive analytics, and (4) prescriptive analytics. In the first phase, we succeeded in developing a new decentralised anomaly detection and invented learning from label proportions. In the second phase, we investigated diagnostic and predictive analytics. Diagnostic analytics aims at explaining *not ok* products by process data. Based on time series data from a hot rolling process for steel-bars production, features have been extracted and aggregated that help to distinguish between *ok* and *not ok* products. Modeling the overall preprocessing in the software RapidMiner took the toolbox to its extreme, becoming a high-level programming environment. The advantage is the reproducibility of RapidMiner processes and their understandable documentation. Predictive analytics in flexible production processes has contributed to adequate process control in real-time. For distributed data mining, the training of local models from counts (TLMC) for vertically partitioned data has been developed. The anomaly detection has been enhanced for distributed settings as they are given by the Internet of Things. In the third phase, prescriptive analytics will be the focus of the project; i.e., the learned model changes the process in real-time. An example is not only to recognise a deviation of sensor measurements from the normal, but also to adapt the process parameters accordingly.

The availability of well-prepared industrial data is not always given. Hence, simulations are a means for worry-free experimentation on operating procedures and optimisation strategies that deliver reliable data. The application of active learning is especially promising when combining simulated and measured data. For this purpose, Petra Wiederkehr is included as principal investigator of project B3 for the upcoming period. She is working as professor for “Virtual Machining” at the Department of Computer Science as well as the Faculty of Mechanical Engineering at TU Dortmund University. Her research is focused on developing simulation models for predicting machining processes as well as on analysing and optimising real manufacturing processes. The use of simulations for online process optimisation will be further investigated in the third funding period.

### 3.3 Project progress to date

#### 3.3.1 Report and current state of research

Manufacturing systems within serial production that are organised according to the flow principle often show a high degree of automation. Within those manufacturing systems, failures that are only detected at the end of the value chain may quickly lead to high amounts of reject parts that require laboriously and costly rework or need to be scrapped. To prevent such failures, production processes and the quality of intermediate products have to be monitored continuously. Existing approaches for continuous quality monitoring like statistical process control (SPC) or model predictive control (MPC) require that quality-related product characteristics are known and measurable without timely delay or destruction of the product. In many manufacturing processes, for example hot rolling of steel bars, the characteristics of manufacturing and testing processes do not fulfil these requirements. Therefore, these approaches are not suitable for continuous quality monitoring and control.

A key requirement for continuous quality control is the full coverage of product quality in each stage of the manufacturing process, especially in such stages where quality testing of the product itself is not possible. The linkage of machines in the context of Industry 4.0 through information and communication technology (ICT) to cyber-physical systems with the aim of monitoring, controlling, and optimising complex production systems enables real-time capable approaches for data acquisition, analysis, and process knowledge generation. This makes it possible to collect and analyse sensor data and identify critical process patterns in real-time, as well as to derive control interventions in a timely manner.

For the second funding period seven work packages were defined that focused on the topics of distributed data analysis, prognosis of failure types, active learning for simulation, process control strategies, and prototypical integration into production processes. The following paragraphs summarise the main results of these work packages and relate them to other current approaches.

**Distributed Data Analysis (WP 1)** High-volume sensor data are often transferred to a central node and processed there. However, constraints in the transmission rate and energy consumption may prohibit this procedure, particularly if the required data rate is higher than the maximum transmission rate achievable within the available bandwidth. The scenario of big data from distributed small devices is becoming more and more popular due to increasing interest in the Internet of Things. It requires some analysis and reduction of the data to be performed distributedly on the measuring nodes themselves and limitation of the communication costs of sending local models to the global node. We generalise the specific scenarios of serial production to that of distributed sensors (in some production system), where a global decision needs to be made based on the local measurements. A survey of the Internet of Things investigates such distributed settings in a principled manner [B3/242].

Time series data collected from sensors can be considered as vertical slices of the overall data set. At each point in time, every local sensor contributes the value of its measurement to the overall state of the process. Sensors might be at different production stations, which together measure the production process. This view of vertically partitioned data in production processes has already emerged in the first phase of the project. However, the problem of restricted communication resources was only tackled in the second phase. An overview of distributed variants of the support vector machine (SVM) shows their resource constraints [B3/243]. The new method TLMC exchanges aggregated label values between the neighbouring nodes instead of sensor values [B3/249]. Inside a window over the data stream, the algorithm locally counts the number of each class label and transfers this information to a limited number of neighbours. In the end, the labels are transmitted to the neighbours and then fused locally (e.g., with majority vote) for

prediction. With TLMC, the communication costs could be reduced by an order of magnitude, because only the label proportions (not the sensor values) are transmitted and the algorithm does not require any communication for the selection of hyperparameters. Number of neighbours is the only parameter to be tuned for each node. The trade-off regarding accuracy and communication cost of TLMC was explored using data from a real-time traffic prediction task. Data from about 296 street crossings from Dublin City have been aggregated into intervals of 15 minutes. Locally trained k-nearest-neighbor (kNN) models with a fixed number of six neighbours achieved 84,27% accuracy. With windows of size 50, TLMC achieved 79,62% accuracy and transmitted only a fifth of the data compared to kNN, where all label data was shared among the neighbours. Accuracy similar to that of a globally trained model was achieved with 418 times less data transferred over the network [B3/246]<sup>1</sup>.

**Prognosis of failure types (WP 2)** Detailed and interpretable information of failures within the production chain enables targeted adjustments of the controlling system. Throughout the first funding period, the products were classified binary as *ok* or *not ok* using predefined threshold values. The class of steel blocks in milling process as expressed by a single numerical label, the so called Q-level, which is calculated as a weighted sum of various failure information, is not as informative as a vector of specific failure numbers. Patterns of failure types in time series should be automatically identified and converted into specific failure indicators. We might then detect dependencies between the specific failures.

We received data from quality testing of steel bolts: around 4000 ultra sonic spectra were manually marked with different failure types. We evaluated multiple classification algorithms in combination with various methods of preprocessing. Additionally, we followed the advice of domain experts to remove the back wall echo from the signal. A combination of PCA and wavelets was used to extract features which were then fed into decision trees. This achieved 85% classification accuracy of single failure types. However, a detailed investigation of the trained decision trees revealed that metadata such as the material was favoured throughout the trees. Just by utilising meta information, the decision tree achieved 85% prognosis accuracy, while a classifier based on features extracted from raw data using wavelets obtained only 65%. Using an SVM with RBF (radial basis function) kernel on raw and not preprocessed spectral data achieves 85% accuracy. The accuracy did not improve significantly after adding meta information attributes. Over all components of the failure type vector, the SVM with an RBF kernel performed best, achieving accuracy values between 77% and 85%.

**Active learning for simulation (WP 3)** Simulations are increasingly required to analyse processes with regard to efficiency, safety and the resulting quality of the output with minimal experimental effort. In the second funding period, a Finite Element (FE)-based simulation approach was used to generate additional data for training models, which were based on ML methods. This FE approach can be computationally expensive, especially if it is designed to mirror the process down to the smallest detail [214]. However, building accurate simulation-based ML models requires a large amount of simulation data. Due to the high costs of each FE-simulation run, the collection of such data was expensive [220]. To avoid these, we used Active Learning (AL) to build the most accurate ML model using the lowest amount of simulation data (i.e., minimal number of simulation scenarios).

AL selects a small set of training examples such that a ML model trained on the small set performs (almost) as accurate as a model trained on a large number of examples randomly chosen, while being computationally smaller [234]. Following this idea, AL exploits the user-machine interaction to increase the model accuracy using an optimised training set. Starting from a small and non-optimal training set, traditional AL procedures select unlabelled data points whose inclusion in the

---

<sup>1</sup>An implementation of learning from label proportions by clustering (LLPC) in R is available.

training set improves the performance of the learning model iteratively. Different methods were presented in literature. In Uncertainty Sampling [227], an active learner queries the instances to identify the instance with the highest uncertainty when labelling. Another approach is the Query-By-Committee (QBC) [235]. Thereby, the most informative query is considered to be the instance about which a committee of learners, built following different hypothesis on a current labelled set, disagrees the most. In the second funding period, we used AL to perform informed selection of the minimal number of simulation scenarios whose inclusion in the training set enables to achieve the same performance as a model trained using larger number of scenarios and in best cases to exceed this performance.

In a collaboration with the CRC 837 (Interaction Modeling in Mechanized Tunneling), an AL-based approach was developed using data collected from an FE-based simulation designed specifically for a mechanised tunnelling process at the Institute for Structural Mechanics of the Ruhr University Bochum [214]. The simulation was used to predict tunnelling induced surface settlements during the tunnel construction phase. Each simulation scenario corresponds to a combination of two varying input parameters: the grouting pressure and the support pressure. In the first developed approach, we considered the simulation scenarios themselves as instances and characterised them with the corresponding input parameters as features. First, we trained a model using a small and non-optimal set of scenarios. Second, we iteratively selected the next scenarios according to their maximal Euclidian distance to the already selected training scenarios. We identified the closest training scenario to each unlabelled scenario and selected a user-defined number of the most distant ones. The process was iterated until a defined maximum number of iterations was reached. We were able to reduce the number of training simulation scenarios by one-half while achieving lower error rates in comparison to the model trained with all available simulation scenarios. We have improved this method by adding the performance of the learned model at each iteration and considering only the labelled scenarios of highest model error rates before computing the distances. This resulted in reducing the required scenarios for training by one-half, while the achieved reduction of error rates was up to 13% in comparison to the first approach, applied to the same simulation data.

Another developed approach, which serves as preliminary work for the third funding period, is based on the QBC paradigm [235]. Thereby, a geometric physically-based process simulation system [237] was used to generate data for a milling process. Using this simulation approach, in contrast to FE-based methods, it is possible to predict forces, stabilities, and location errors of the workpiece surface for long-running milling processes in a feasible runtime. We trained a committee of deep learning models using purposely designed training sets based on subsets of features. This committee was utilised to simultaneously select the training scenarios on which the committee disagreed the most and to build a weighted ensemble model. These scenarios were summarised into the weighted ensemble, whose weights were updated with each update of the training set. This approach resulted in reducing the number of training scenarios by 30% while having approximately the same prediction accuracy as an ensemble model that was trained on the overall training set. We could successfully use this approach for stability prediction of milling processes [B3/240]. Furthermore, this model will be used for real-time applications in the third funding period in the context of WP 7. To accomplish this, the model will be refined with measurements of real processes. Since the runtime of simulation scenarios, when utilising the geometric physically-based approach, is not limiting the applicability of developed ML methods and milling processes are costly and time-consuming, AL will be used to identify the contributed improvement of the model when including measurements resulting from processes with a specific parameter value configuration into the training set.

**Production process control strategies (WP 4 – 6)** The integration of quality predictions into intelligent production process control has been developed in the first funding period based on multivariate analyses. On the basis of quality predictions and threshold values, control decisions about the rejection or further processing of a product could be derived. The further the manufacturing process is progressing, the higher is the quality threshold value. The quality predictions have

to fulfil this value due to decreasing process-related degrees of freedom along the progress of the production process. [B3/247] This approach was to be extended towards a comprehensive process control strategy within the second funding period in order to further reduce internal failure costs and to avoid the waste of resources. In order to develop a holistic process control methodology that derives the best control decision across different application scenarios, we investigated multiple strategies.

First, we investigated approaches for process parameter adaptations to reach the desired product quality. The current state of the art includes multiple approaches with focus on intelligent control of automated production processes. However, depending on the complexity of the examined processes these are not generally suitable. A widespread approach for automated control of production processes, especially in the process industry, is the MPC [218]. MPC refers to a class of algorithms that use dynamic models of processes to predict the future behaviour of system output and to determine suitable settings of the manipulated variables on this basis. To successfully apply MPC, an appropriate analytical (usually linear) model is required that realistically reflects the process dynamics [223]. However, in complex, interlinked production processes, finding such a model is usually not possible. Other approaches for process control found in current literature are based on simulation models [229, 222, 230] and the training of causal models using multivariate statistical methods [239, 232]. While approaches based on simulation mostly do not fulfill the real-time requirements, causal approaches are only suitable for processes with a low degree of complexity and if causal relationships between product states are given, process parameter configurations and resulting quality can be clearly stated. Therefore, interlinked processes that contain several process steps are too complex to map causal relationships between product states and process parameter configurations transparently. Accordingly, only observed or simulated process specifications may be taken into account for parameter adaptations.

For this reason, we have developed an approach that does not require any mapping of causal relationships. In [B3/245] a general framework for this approach is proposed interpreting the process chain as a controlled system and endowing the controller with similarity search algorithms. When an intermediate product reaches a control point, the final quality will be predicted based on previously recorded process parameters and available meta information, e.g., on product structure or raw material. If the predicted quality of an intermediate product is not sufficient to fulfil the final quality requirements, a decision on adapting subsequent process parameters has to be made. Therefore, we investigated different applications of quality prediction and similarity search. First we compared the features of the intermediate product with historical data sets. With features we hereby refer to aggregated values from raw process data, the result of the quality prediction, and meta information. Those features can be numerical as well as nominal attributes depending on the type of information. In the first step, we have evaluated different distance and similarity measures for the similarity search. We finally chose the mixed Euclidean distance because of its ability to consider both numerical and nominal attributes and its natural interpretation. Furthermore, various weighting strategies to consider the attributes' importance have been investigated. Based on similarity calculations, the nearest neighbour of the currently processed product is finally chosen as reference product for process adjustments. The decision on whether process parameters are to be adjusted depends on the progress of the production process as well as the nearest neighbour's similarity. In a second approach we selected a sample of historic data sets whose subsequent process parameters are within a certain tolerance compared to the target parameters of the intermediate product. We then apply quality prediction to evaluate if a set of parameters exists such that the final quality will meet the requirements. If such a set of parameters exists, the subsequent process parameters will be adapted accordingly. If multiple parameter sets exist that fulfil the requirements, different criteria like confidence of the model or least adaptation effort (highest similarity) can be taken into account. First applications in simulated scenarios have already shown that the benefit of both approaches decreases as production progresses. While the degree of freedom for adaptations decreases as the production progresses, the influence of the adapted parameters on the final quality also decreases. In further applications of both approaches to different use cases we want to investigate the suitability and benefits depending on application-specific conditions.

In addition to approaches focusing on process parameter adaptations, we investigated organisational control strategies, since physical intervention in production processes is not always admissible and feasible. In this context, new opportunities for quality-based control strategies arise through the increasing customization leading to more individual product requirements and more diverse quality characteristics. It is no longer suitable to derive uniform internal quality standards that the final product has to fulfil in order to be delivered. Instead, measured quality characteristics have to be compared to requirements specified by the customer to determine quality deviations. The integration of knowledge about customer orders and requirements into process control enables new additional decisions: If the predicted quality of an intermediate product does not fulfil the requirements of the assigned customer order, it can possibly be reassigned to a different customer order within the order pool instead of being discarded.

We therefore developed a method that allows reassigning intermediate products to customer orders from the order pool based on predicted quality deviations. Since customers' requirements for a product can sometimes be very diverse, a detailed consideration of all values of quality characteristics can lead to an enormous model complexity, which in turn would lead to long runtimes so that no efficient integration into process control would be possible. In this case, individual customer requirements can be grouped into classes of similar requirements. The classes are represented by the minimal requirements on quality characteristics. The complexity of this grouping task increases with the number of considered quality characteristics, so that these are reduced to few critical features relevant to the customer. For the integration of this method, the production process chain is divided into sections, and a control point is established between each two sections. Whenever a product reaches a control point, the expected quality of the product in its final state is predicted based on previously recorded process parameters. If the predicted quality does not fulfil the requirements of the assigned customer order, a mathematical optimisation model is used to find an optimal reassignment setting. We formulated the mathematical problem as an assignment problem and used the Kuhn-Munkres algorithm to find optimal solutions [B3/248]. Within the development of the optimisation model, different objectives and constraints were investigated and compared. We thereby differentiate between the process-oriented and the customer points of view. From the latter point of view, the key performance indicator (KPI) delivery date deviation has been identified to be a primary indicator for customer satisfaction and has therefore been taken into account for first modelling approaches. From the process-oriented point of view, other KPIs like average order processing time might be of greater interest. The mathematical modelling of the problem, as described in [B3/248], is constructed generically so that it can be applied regardless of the selected objective.

Finally, we investigated how the individual approaches presented above can be aggregated into a holistic process control methodology that derives the best control decision under consideration of all three options (reassignment of intermediate products to customer orders, process parameter adaptation, and rejection of defective parts). The main challenge here is the complexity and the resulting computation time of the optimisation and decision models when all options are taken into account simultaneously. Hence, we are currently evolving a sequential approach that prioritises the three options based on given restrictions and targets. The fulfilment of quality requirements through choosing the best option takes place within the conflicting areas of time and costs. Depending on the main objective of the optimisation, different priorities of options can be obtained. For this reason, we are investigating how the ranking of options and thus the sequence generation can be involved in the approach of adaptive process control. Additionally, the methods developed should be further evaluated and validated by applying them to different (empirical) use cases.

**Prototypical integration (WP 7)** Some of the developed methods are well documented for use by others. In particular, a comprehensive workflow for data cleansing, feature extraction, and feature selection was developed in the software RapidMiner. This allows an efficient preprocessing of sensor data. Even access to ibaPDA, which is a standard data acquisition system in steel industry, has been integrated into RapidMiner. This frequently used process language can readily be used.

In addition, problem-oriented prediction methods such as the learning from label proportions (LLP) and the TLMC algorithms were developed and have already been successfully applied in real scenarios [B3/246]. The results presented in [B3/244] showcase the embedding of quality into interlinked manufacturing processes for higher sustainability.

For integrating quality predictions in machine control the Intelligent Manufacturing Process Control (IMPC) framework has been proposed in the first funding period. Interfaces were created for machine control in order to be able to evaluate data in real-time. This work toward a prototype has continued in the second funding period.

## Bibliography

- [67] A. Bifet and R. Gavaldà. “Learning from time-changing data with adaptive windowing”. In: *In SIAM International Conference on Data Mining*. 2007 (cit. on pp. 111, 228).
- [213] A. Bifet, J. Read, B. Pfahringer, G. Holmes, and I. Zliobaite. “CD-MOA: Change Detection Framework for Massive Online Analysis”. In: *Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013*. Ed. by A. Tucker, F. Höppner, A. Siebes, and S. Swift. Springer, 2013, pp. 92–103 (cit. on p. 228).
- [214] B.-T. Cao, S. Freitag, and G. Meschke. “A Hybrid RNN-GPOD Surrogate Model for Real-time Settlement Predictions in Mechanised Tunnelling”. In: *Advanced Modeling and Simulation in Engineering Sciences* 3.1 (Mar. 2016), p. 5 (cit. on pp. 220, 221).
- [215] D. D’Addona and R. Teti. “Image Data Processing via Neural Networks for Tool Wear Prediction.” In: *Procedia CIRP - Eighth CIRP Conference on Intelligent Computation in Manufacturing Engineering* 12 (2013), pp. 252–257 (cit. on p. 231).
- [216] B. Denkena, M.-A. Dittrich, and F. Uhlich. “Augmenting Milling Process Data for Shape Error Prediction”. In: *Procedia CIRP* 57 (2016), pp. 487–491 (cit. on p. 231).
- [217] D. E. Dimla. “Sensor signals for tool-wear monitoring in metal cutting operations—a review of methods”. In: *International Journal of Machine Tools and Manufacture* 40.8 (2000), pp. 1073–1098 (cit. on p. 232).
- [218] M. Ellis and P. D. Christofides. “Integrating Dynamic Economic Optimization and Model Predictive Control for Optimal Operation of Nonlinear Process Systems”. In: *Control Engineering Practice* 22 (2014), pp. 242 –251 (cit. on p. 222).
- [219] F. Finkeldey, S. Hess, and P. Wiederkehr. “Tool wear-dependent process analysis by means of a statistical online monitoring system”. In: *Production Engineering. Research and Development* 11.6 (2017), pp. 667–686 (cit. on p. 231).
- [220] C. Fischer. “Runtime and Accuracy Issues in Three-dimensional Finite Element Simulation of Machining”. In: *International Journal of Machining and Machinability of Materials* 6.1 (2009), p. 35 (cit. on p. 220).
- [221] D. Freiburg, R. Hense, P. Kersting, and D. Biermann. “Determination of Force Parameters for Milling Simulations by Combining Optimization and Simulation Techniques”. In: *Journal of Manufacturing Science and Engineering* 138.4 (2016) (cit. on p. 231).
- [222] N. Fujita, Y. Kimura, K. Kobayashi, K. Itoh, Y. Amanuma, and Y. Sodani. “Dynamic control of lubrication characteristics in high speed tandem cold rolling”. In: *Journal of Materials Processing Technology* 229 (Mar. 2016), pp. 407–416 (cit. on p. 222).

- [223] J. M. Gálvez, L. E. Zárate, and H. Helman. “A Model-based Predictive Control Scheme for Steal Rolling Mills Using Neural Networks”. In: *Journal of the Brazilian Society of Mechanical Sciences and Engineering* 25 (Mar. 2003), pp. 85–89 (cit. on p. 222).
- [70] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. “A Survey on Concept Drift Adaptation”. In: *ACM Comput. Surv.* 46.4 (Mar. 2014), 44:1–44:37 (cit. on pp. 110, 228).
- [71] J. Gama, P. Medas, G. Castillo, and P. Rodrigues. “Learning with Drift Detection”. In: *Advances in Artificial Intelligence - SBIA 2004*. Ed. by A. Bazzan and S. Labidi. Vol. 3171. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, pp. 286–295 (cit. on pp. 111, 228).
- [224] S. Hess, F. Finkeldey, and P. Wiederkehr. “Elaborated analysis of force model parameters in milling simulations with respect to tool state variations”. In: *Procedia CIRP* 55 (2016), pp. 83–88 (cit. on p. 231).
- [225] R. Klinkenberg and T. Joachims. “Detecting Concept Drift with Support Vector Machines”. In: *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*. Ed. by P. Langley. San Francisco, CA, USA: Morgan Kaufmann, 2000, pp. 487–494 (cit. on p. 228).
- [226] L. I. Kuncheva. “Classifier ensembles for changing environments”. In: *International Workshop on Multiple Classifier Systems*. Springer, 2004, pp. 1–15 (cit. on p. 229).
- [227] D. D. Lewis and 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*. Springer-Verlag New York, Inc., 1994, pp. 1–12 (cit. on p. 221).
- [228] D. C. Montgomery. *Introduction to statistical quality control*. 7. ed. Hoboken, NJ: Wiley, 2013 (cit. on p. 230).
- [229] M. Mosayebi, F. Zarrinkolah, and K. Farmanesh. “Calculation of stiffness parameters and vibration analysis of a cold rolling mill stand”. In: *The International Journal of Advanced Manufacturing Technology* 91.9-12 (Feb. 2017), pp. 4359–4369 (cit. on p. 222).
- [230] M. R. Niroomand, M. R. Forouzan, and M. Salimi. “Theoretical and Experimental Analysis of Chatter in Tandem Cold Rolling Mills Based on Wave Propagation Theory”. In: *ISIJ International* 55.3 (2015), pp. 637–646 (cit. on p. 222).
- [231] T. Pfeifer and R. Schmitt, eds. *Masing Handbuch Qualitätsmanagement*. Carl Hanser Verlag GmbH & Co. KG, May 2014 (cit. on p. 229).
- [232] A. E. Raftery, M. Kárný, and P. Ettler. “Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill”. In: *Technometrics* 52.1 (Feb. 2010), pp. 52–66 (cit. on p. 222).
- [233] J. Schnell and G. Reinhart. “Quality Management for Battery Production: A Quality Gate Concept”. In: *Procedia CIRP* 57 (2016), pp. 568 –573 (cit. on p. 229).
- [85] M. Scholz and R. Klinkenberg. *Boosting Classifiers for Drifting Concepts*. Tech. rep. 6/06. Dortmund, Germany: Collaborative Research Center on the Reduction of Complexity for Multivariate Data Structures (SFB 475), University of Dortmund, Jan. 2006 (cit. on pp. 111, 228).
- [234] B. Settles. *Active Learning Literature Survey*. Tech. rep. 2010 (cit. on p. 220).
- [235] H. S. Seung, M. Opper, and H. Sompolinsky. “Query by Committee”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 287–294 (cit. on p. 221).

- [C1/236] **B. Schowe and K. Morik.** “Fast-Ensembles of Minimum Redundancy Feature Selection”. In: *Ensembles in Machine Learning Applications*. Ed. by O. Okun, G. Valentini, and M. Re. Studies in Computational Intelligence. Springer, 2011, pp. 75–95 (cit. on p. 228).
- [237] P. Wiederkehr and T. Siebrecht. “Virtual Machining: Capabilities and Challenges of Process Simulations in the Aerospace Industry”. In: *Procedia Manufacturing* 6 (2016), pp. 80–87 (cit. on pp. 221, 231).
- [238] T. Wuest, A. Liu, S. C.-Y. Lu, and K.-D. Thoben. “Application of the Stage Gate Model in Production Supporting Quality Management”. In: *Procedia CIRP* 17 (2014), pp. 32 –37 (cit. on p. 229).
- [239] M. A. Younes, M. Shahtout, and M. Damir. “A Parameters Design Approach to Improve Product Quality and Equipment Performance in Hot Rolling”. In: *Journal of Materials Processing Technology* 171.1 (2006), pp. 83 –92 (cit. on p. 222).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [B3/240] **A. Saadallah, F. Finkeldey, K. Morik, and P. Wiederkehr.** “Stability prediction in milling processes using a simulation-based machine learning approach”. In: *51st CIRP conference on Manufacturing Systems*. Elsevier, 2018 (cit. on pp. 20, 221, 306).
- [B3/241] **J. Deuse, J. Schmitt, M. Stolpe, M. Wiegand, and K. Morik.** “Qualitätsprognosen zur Engpassentlastung in der Injektorfertigung unter Einsatz von Data Mining”. In: *Schriftenreihe der Wissenschaftlichen Gesellschaft für Arbeits- und Betriebsorganisation (WGAB) e.V.* (2017) (cit. on pp. 230, 231).
- [B3/242] **M. Stolpe.** “The Internet of Things: Opportunities and Challenges for Distributed Data Analysis”. In: *SIGKDD Explorations* 18.1 (June 2016), pp. 15–34 (cit. on pp. 18, 219).
- [B3/243] **M. Stolpe, K. Bhaduri, and K. Das.** “Distributed Support Vector Machines: An Overview”. In: *Solving Large Scale Learning Tasks: Challenges and Algorithms*. Ed. by S. Michaelis, N. Piatkowski, and M. Stolpe. Vol. 9580. LNAI. Springer International Publishing, 2016, pp. 109–138 (cit. on pp. 18, 219).
- [B3/244] **M. Stolpe, H. Blom, and K. Morik.** “Sustainable Industrial Processes by Embedded Real-Time Quality Prediction”. In: *Computational Sustainability*. Springer, 2016, pp. 201–243 (cit. on pp. 224, 228).
- [B3/245] **M. Wiegand, M. Stolpe, J. Deuse, and K. Morik.** “Prädiktive Prozessüberwachung auf Basis verteilt erfasster Sensordaten”. In: *at-Automatisierungstechnik* 64.7 (2016), pp. 521–533 (cit. on pp. 18, 222, 228, 229).
- [B3/246] **T. Liebig, M. Stolpe, and K. Morik.** “Distributed Traffic Flow Prediction with Label Proportions: From in-Network towards High Performance Computation with MPI”. In: *Proceedings of the 2nd International Workshop on Mining Urban Data (MUD2)*. Vol. 1392. CEUR-WS, 2015, pp. 36–43 (cit. on pp. 18, 21, 220, 224, 261).

## b) Other publications

- [B3/247] **D. Lieber.** "Data Mining in der Gütesicherung der Stabstahlproduktion". Diss. TU Dortmund, 2018 (cit. on p. 222).
- [B3/248] **J. Schmitt**, F. Hahn, and **J. Deuse**. *Mathematical modelling of the quality-based order assignment problem*. Techreport 2. Institute of Production Systems, TU Dortmund University, Mar. 2018 (cit. on p. 223).
- [B3/249] **M. Stolpe**. "Distributed Analysis of Vertically Partitioned Sensor Measurements under Communication Constraints". Diss. Dortmund: TU Dortmund University, 2017 (cit. on p. 219).

## 3.4 Project plan

### Goals

The long-term goal of project B3 is to achieve efficient and zero-defect production through the development of resource-aware procedures for real-time production process control and quality forecasting. The real-time aspect will be the main focus of the third funding period. During the first and second funding periods, trade-offs between runtime, storage consumption, communication costs, energy consumption, data acquisition costs, and forecasting quality were established without taking the real-time constraints into account. However, to cope with the new challenges of producing increasingly individualised products with a short lead-time to market and higher quality, facing smart industries nowadays, optimal and real-time decision-making support methods will be investigated. Given the stages of data analytics ranging from (1) anomaly detection, (2) diagnostic analytics, and (3) predictive analytics to (4) prescriptive analytics, in the third period, prescriptive analytics will be the focus of the project. On the one hand, the learned model will be used to adapt the process in real-time. On the other hand, changing situations will be rapidly be detected and learned models will be modified accordingly.

While bridging gaps in scientific research through new ML methods is the main focus of our work, the verification of applicability and suitability for real-world applications should not be neglected in order to ensure knowledge transfer to industry. Accordingly, the developed ML methods have to be evaluated and validated on real data based on real-life use cases with all related constraints. The requirements of ML applications in manufacturing depend not only on the underlying use case but also on environmental conditions such as the industry sector. Therefore, we intend to utilise multiple use cases. Well-known companies of different industries were successfully acquired as transfer and application partners for project B3 for the next funding period. Our work packages will be applied to address challenges in a wide range of processes in automotive and electronics industries such as the German multinational company BMW. In addition to the acquired industrial use cases, we will investigate issues which arise in manufacturing processes in our local experimental laboratory. This allows data to be generated without a dependency on external application partners. We have also established a collaboration with the CRC 837, given by the MERCUR project "Synthese von maschinellem Lernen und numerischer Simulation zur Echtzeitsteuerung von Tunnelvortriebsprozessen" (2017 - 2019). Our goal is to apply our online diagnostic, predictive and prescriptive analytics methods to the data collected from sensors implanted around a Tunneling Boring Machine (TBM) of a tunnel project "the Wehrhahn-Linie metro" in Düsseldorf in Germany, to ensure safe and sustainable constructions in mechanised tunneling.

## Work schedule

**Work package 1. Real-time aggregation and feature extraction** Aggregation and feature extraction drastically reduces the size of data and accordingly the resulting processing time, with comparable model accuracy [B3/244, B3/245]. Aggregating collected sensor data can be performed offline by gathering statistical indicators using a fixed time window. Such approaches can be extended for online application by using a sliding time window [67]. As we will be dealing with time-evolving data where concepts may shift following the process changes (e.g., modifying the machine input parameters may induce important changes in the process measurements), adaptive windowing is relevant for detecting change [70, 225, 213].

Furthermore, feature extraction is another way to increase the speed of processing sensor data by keeping only the important information while discarding noise and removing correlated measurements. This automatic representation optimisation, however, cannot be performed in real-time. Here, further work is needed. If many extractions are prepared, the online representation change becomes a feature selection task. Based on the algorithm for fast feature selection from the first phase of project C1, [C1/236], we may select the best-suited feature set online. To speed up even the feature selection, we might index feature sets and just select the appropriate set of features.

**Work package 2. Real-time learning and model adaption** Generating accurate and adequate responses in real-time requires building models that are able to capture, filter, analyse and learn from data in real-time. These models can be trained offline but should also be able to learn from changes in distributions of incoming data streams. In a streaming data setting, a predictive model may need to be updated or completely replaced by a new model. We will focus on building models that are able to detect changes, distinguish between noise and drifts, and are adaptive to changes, while remaining robust to noise. Concept drift detection has caught the attention of the machine learning community, and many works have proposed ways to detect and handle such types of changes [71, 225]. Concept drift detection has already been investigated in the framework of streaming data [70, 213]. In addition to concept drift detection, we will design methods for categorising the occurring drifts as global or local by learning rules for constrained regions of the instances space (i.e., set of ranges for the measured attributes). Our research in this direction for the third funding period will cover both regression and classification problems.

Memory management is another aspect of model adaption that will be studied during the next funding period. Learning from evolving data with unknown dynamics requires not only updating the predictive model but also managing indicative data sets and models. In this work package, we investigate how to select the most suitable model and how to monitor its validity in time, with regards to the nature of drifts, the maximum cost of memory, and the reactivity. In conclusion, our aim with this WP is to achieve advanced incremental learning methods that are able to adapt to the evolution of sensor data over time with online efficient memory length choice.

**Work package 3. Online management of many models** The increasing individualisation of products and processes may result in small heterogeneous groups of observations. Learning distinct models is then necessary to represent each group and capture its underlying properties. However, to manage the overall generating process, online management of the resulting models should be established. To this end, the third funding period will focus on the dynamic combination of trained models to respond to changes in data by changing the combination rules. The dynamic combination can be formally established with adaptive ensembles that are built as a weighted combination of distributions characterising the target concepts and enable flexible management of the models from individual model selection to weighted aggregation of the individual models' predictions [70, 85].

Combining models for dynamically changing data can be classified into three main families [226]. The first family is dynamic combination, which is applied by changing the combination rules (e.g., changing weights). The second family relies on the continuous update of the learners such that the learners are either retrained in a batch mode or updated online using new observations (i.e., the combination rules do not change during the process). The last family is based on the structural update, where new learners are added (or existing ones are activated if they have been deactivated before) and inefficient ones are removed. These three categories are not necessarily mutually exclusive. In the third funding period we will investigate several model combination strategies and develop methods for combining two or all three of these strategies, based on the preceding work packages.

The application of adaptive ensemble methods will be made in connection with data drift and online learning. The best possible combination of models will be determined and updated in real time by taking into account the occurrence of drifts (e.g., changing the weighting strategy when a drift is detected). Diversity in adaptive ensemble building will also be explored during the third funding period, along with its impact in covering the distributions characterising the target concepts and in adapting to changes over time. In addition, we will develop clustering strategies for the trained models on various groups of observations. Various clustering criteria will be investigated (e.g., performance of the models on a batch of data, features importance representation). Model clustering will serve as a tool for efficient model selection or aggregation and better representation of the target criteria over time that can help in drawing interpretable results.

**Work package 4. Prediction based quality gate design for process chains** In order to eliminate quality problems directly after occurrence and thus avoid costly rework and waste of resources, quality deviations should be detected as early as possible in the production chain. Effective quality gate design with inline quality testing allows for continuous monitoring of process and product quality. Quality gates mark checkpoints at the beginning or end of critical production stages where the fulfilment of previously agreed performances is measured [231]. They can be seen as decision and control points, offering the possibility for backward and forward feedback loops [238, 233]. In some processes, physical inline tests are not feasible due to technical and organisational restrictions [B3/245]. For this reason, in the first and second funding period methods were developed that allow for inline predictions of final product quality as well as deriving process control decisions. However, these methods did not in particular focus on the design of quality gates, which we will be investigating in the next funding period.

The aim of this work package is to develop a method that allows for efficient quality gate design based on quality predictions for arbitrary production processes. We want to investigate the research questions (1) how to identify and measure relevant product and process parameters for quality prediction and (2) where to position quality gates within the process chain. Traditional approaches for quality gate design in literature identify critical process stages in the first place and derive design decisions for quality gates from this. For example in [238] the stage gate model from the product development domain is adapted and applied to quality gate design. In this approach, quality gates are installed within the process chain depending on criteria like known quality problems, measurability of quality properties, and process layout. In [233] a quality gate system is proposed that is based on the evaluation of product-process correlations using a modified failure mode and effect analysis (FMEA). On the basis of analytical results, relevant measurement steps, appropriate measuring equipment, and the position of quality gates are determined.

However, the approaches found in literature focus on the design of quality gates conducting real measurements of quality properties only and do not include the design of predictive quality gates. Therefore, a method has to be developed that considers the special characteristics of quality predictions as surrogates of real measurements. The effective design of predictive quality gates thereby depends on several criteria. As a starting point, product and process parameters with significant

impact on product quality have to be identified. We want to investigate how the FMEA-approach can be adapted in order to identify relevant input features for training quality prediction models. Besides that, a combination with other tools like value stream analysis or design of experiments will be examined. The measurability of parameters and the derivation of appropriate sensors are also of great importance in this context. For positioning quality gates within the process chain, the reliability of quality predictions as well as the remaining degrees of freedom for control decisions have to be considered as well. Furthermore, in the presence of concept drift and product variety, correlations between process and product parameters and quality properties could change requiring flexible adaptions of quality gate design. For example, additional quality gates or changes of gate positions may be necessary depending on process changes. Thus we want to investigate how a quality gate system has to be designed to cope with time-varying processes. The result of this work package will be a method that allows for predictive quality gate design in complex production processes.

**Work package 5. Quality prediction for bottleneck relief in EOL-testing** Despite advanced quality gate systems, extensive quality testing at the end of the manufacturing process is often essential to ensure the delivery of zero-defect products to the customer. Due to high testing efforts, end-of-line testing (EOL-testing) can easily become the bottleneck of the production system [B3/241]. While the assurance of quality for all products is indispensable, an efficient design of EOL-testing should not decrease the throughput of the production system. The approaches that have been developed so far focus on control decisions within the process chain, ignoring the efforts of EOL-testing.

In this work package we want to investigate how the previously developed methods for quality prediction can be used effectively in order to relieve EOL-testing as the bottleneck of the production line. Therefore, strategies should be developed that allow a reduction of testing efforts. Existing approaches in literature are often based on sample testing of products to reduce testing efforts and costs. A common quality control technique for sample testing in industry is acceptance sampling, which uses statistical sampling to determine whether to accept or reject a production lot. Thereby, the quality of the production lot is evaluated by testing only a sample of products. On the basis of the sample, statements about the overall quality of the entire production lot can be derived with defined statistical certainty [228]. However, complete certainty cannot be guaranteed, so that a residual risk for erroneous conclusions remains. Two types of risks can be distinguished: the risk of rejecting a high-quality lot (pseudo fault) and the risk of accepting a low-quality lot.

In some production processes quality requirements on products are particularly high, and acceptance of a low-quality lot would be extremely critical due to legal provisions for safety-critical products or customer agreements. Sample testing cannot be applied in those situations, so a full-scope testing of all products is indispensable. The approaches developed within the first and second funding period allow for a predictive assessment of quality characteristics based on process parameters for each product. However, due to prediction uncertainty, a complete replacement of physical quality tests by prediction results is not advisable if the risk to accept a low-quality lot should be virtually zero.

Accordingly, testing strategies have to be developed that reduce testing efforts by utilising machine learning and therefore relieve EOL-testing as the bottleneck and at the same time ensure minimal risk of accepting low-quality lots. In order to reduce pseudo faults even in critical areas where the prediction model does not provide reliable results, quality predictions can be complemented with real measurements. To this end we will investigate how machine learning and physical testing on product samples can be combined. Furthermore, we will explore if test plans can be reduced by eliminating single testing parameters or testing points that can be predicted reliably. A key role within the design of testing strategies is the uncertainty of prediction results. Hence, it should be investigated how different performance measures, e.g., sensitivity and specificity, as well as

the prediction's confidence can be considered in the design of time-efficient test plans. Preliminary investigations for this work package can be found in [B3/241]. In light of increasing product variety we will additionally examine how quality predictions can be included in product-individual test plans for a broad range of different products. Therefore, a combination of the results of WP 3 and the testing strategies developed in this work package is necessary.

#### **Work package 6. Combination of machine learning methods and process simulation systems**

Due to the fast evaluation time of models that are based on machine learning methods, an optimisation of industrial manufacturing processes could be achieved in real time [215]. Thereby, these models are capable of predicting upcoming process characteristics and events based on currently measured, previously unseen data. Unfortunately, an enormous amount of data is necessary for learning such models. Using simulation techniques, it is possible to generate an appropriate amount of data with minimal experimental effort. Using geometric physically-based simulation systems [237], milling processes can be analysed regarding, e.g., the resulting cutting forces, process stabilities, and surface qualities in a feasible runtime by utilising simplified geometric and empirical models. The calibration of these models [221] is usually based on simplified milling experiments, so that the identified parameter values have a limited transferability to processes with an increased complexity or different engagement conditions resulting in a deviation between simulated and measured data [224].

The goal of this work package can be divided into two main objectives. The first is to increase the accuracy of simulation results by integrating machine learning methods into a process simulation system. Thereby, several models, e.g., for predicting the process forces, will be replaced by machine learning models. Combining simulation data with measurements is a promising strategy to incorporate additional information into machine learning approaches [216]. For this purpose, a model will be trained using experimental force measurements as target values and simulated data as feature values for engagement situations of different complexities. As these features cannot be measured during the milling experiment, the simulation system is excellently suitable to incorporate additional knowledge into the training of machine learning models. As a consequence, a novel approach for combining these time series has to be developed. In addition, tool wear has a high influence on the cutting forces in milling operations [219]. The geometric consideration of tool wear is not reasonable because of its stochastic characteristic and would deteriorate the simulation runtime. Therefore, an approach for representing the relationship between the simulated tool load, the cutting forces and the flank wear width will be investigated utilising a supervised learning procedure.

The second objective targets the reduction of the experimental effort for training machine learning models. To achieve this, a methodology based on simulated data will be investigated. In addition, the simulation-based model will be refined using a reduced number of experimental measurements to deal with the deviations between simulated and measured data. There are two possible strategies for the refinement of the model. The first strategy consists of a re-training of the initial model to adjust the weights of the model according to the additional data. As a second strategy, another model could be generated, solely using experimental data. The predictions of the two models have to be combined in a weighted manner. This could be realised by designing weights, whose values are chosen inversely to an error metric on experimental test data. Thereby, the approaches regarding active learning that were investigated in WP 3 of the second funding period will be utilised to identify how to design experimental processes, so that the contribution of the experiment to the training set of the refinement can be maximised. For this purpose, a demonstrator workpiece with increasing complexity will be designed, which will be machined and which combines different geometric shapes to generate heterogeneous data with as few experiments as possible. The systematic identification of the number of real experiments necessary to achieve an appropriate prediction accuracy and the eligibility of different geometric shapes for the demonstrator are important aspects for the research, which targets the refinement of the model. In addition, the research question

how the aggregation methods of WP 1 can be incorporated to reduce the data size without losing significant information will be investigated. The result of this work package will be a methodology to generate a model that is based on a combination of experimental measurements, process simulations, and machine learning methods and is capable of predicting characteristics of milling processes with a suitable accuracy.

**Work package 7. Online optimisation of NC-milling processes** Constantly changing process conditions and complex relationships between process characteristics and quality metrics for the machined workpiece are distinguishing challenges when optimising manufacturing processes. Due to the complexity of machining operations, an in-process adaption of the parameter values based on online measurements is required. Different signals can be used to monitor the condition of process characteristics [217]. However, these approaches are only capable of reacting to events that have already occurred, e.g., process instabilities or tool breakage. To prevent these unwanted occurrences and also to identify how to adapt process parameter values without worsening the monitored condition, predictive methods based on a combination of in-process measurements, process simulations, and machine learning models will be developed in the project. These methods will be applied using a 5-axis machining centre to optimise milling processes under real-time constraints. To achieve this, the following research questions have to be investigated.

The prediction of abrupt changes of the process behaviour is complicated, as these changes are difficult to represent as a function of historical data. To include upcoming information as additional features taking these changes into account, the simulation approach has to be modified, so that it is capable of predicting forthcoming geometric engagement information parallel to the regarded process. This is especially challenging in the considered real-time environment where the limited frequency of the possibility to read and adjust tool positions and process parameter values on machining centres requires the realisation of techniques for synchronising measurements, simulation results and model predictions asynchronously, which can manage considerably differing sample frequencies of time series.

Another research question is the determination of how measurement uncertainties can be considered in the simulation model. To this end, the results of an analysis of different process configurations and their influence on the systematic and stochastic uncertainty characteristics will be used to integrate uncertainties into the simulation parameter values.

This work package is significantly interlocked with WP 2 as the model has to be adapted online based on the currently measured data, so that the prediction of process stabilities improve iteratively as the process progresses. As a consequence, the model will be able to cope with alterations of the process behaviour based on algorithms to detect changes and methods to distinguish between noise and drifts. In addition, the findings of this work package are essential for the investigation of the research questions of WP 2 as they serve as the regarded application scenario. Furthermore, a methodology for adapting the process parameter values with a suitable step size is required. In this context, a demonstrator workpiece with increasing complexity will be designed to obtain fundamental knowledge of the process conditions and to comprise complex 5-axis tool movements.

The investigation of the borders of model predictions cannot be conducted using real experiments. Thereby, highly unstable configurations would damage the spindle and drives of the machining centre and endanger the safety of the machine operator. To this end, the geometric physically-based simulation system provides excellent possibilities to retrieve information about these borders, leading to process configurations that lie beyond the experimental feasibility. In this context, we will investigate the research question of determining methods to use these information to analyse the borders of model predictions in a counter factual manner.

## Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Real-time aggregation and feature extraction																	
2. Real-time learning and model adaption																	
3. Online management of many models																	
4. Prediction based quality gate design for process chains																	
5. Quality prediction for bottleneck relief in EOL-testing																	
6. Combination of machine learning methods and process simulation systems																	
7. Online optimisation of NC-milling processes																	

## 3.5 Role within the Collaborative Research Centre

B3 has been the reference project for distributed data analysis. Results have already been used for traffic prognosis (B4). This work is concluded by the PhD thesis of Marco Stolpe. Other projects now commence the investigation of distributed data analysis.

A common interest with A1 is how to handle imbalanced data sets. A collaboration between the two projects has started and will be continued for the third phase, as solving class imbalance problems is crucial for the success of learning tasks. Additional exchanges will be established between the two projects for solving learning tasks under real-time constraints.

While production conditions and therefore the corresponding machine learning tasks change over the production time span, the topic of concept drift has to be taken into account within the design of quality gates involving machine learning applications. Accordingly, we plan to integrate RAMBO-CD developed by A3 into the quality gate design in interlinked manufacturing processes.

C3 uses Monte Carlo-based simulation for example generation in astroparticle physics. To avoid the cost of running these simulations to generate examples carrying no additional information for data analysis and learning tasks, our developed methods in the Active Learning package from the second phase will be exchanged with C3. In fact, Active Learning was used in the context of B3, as a tool to perform an informed sampling that iteratively chooses simulation configurations carrying new knowledge for the learner and guaranteeing the accuracy of the learning task with minimal training effort. As we will also continue using Active Learning for reducing industrial process costs and for refinement of machine learning models, an exchange between the two projects will also be established in the third phase regarding these purposes.

The developed GAN active learning and anomaly detection approaches, which can be utilised to deal with concept drift, will be used and adapted to be applicable to image data by project B2 in the context of WP 1 of the third funding period. Furthermore, B2 is planning to investigate a methodology for self-configuration and self-adaption of sensor-actuator units based on online monitoring. In this context, the findings of WP 2 and WP 7 can be generalised for transferring the developed methods to an application for the sensor-actuator scenario.

## 3.6 Differentiation from other funded projects

### AIM

#### (Deuse, Reference number BMBF 02L14A160) (Funding period: 2016–2019)

Within the project “Work assistance system for the individualisation of work design and method training”, a context-sensitive work assistance system is being developed which is based on smart devices and is intended to represent a new component of company health management.

### InDaS

#### (Deuse, Reference number BMBF 01IS17063A) (Funding period: 2017–2019)

The “Industrial Data Science” project is developing a teaching concept for the qualification of young academics and specialists from industry in the field of machine learning. Within the framework of the teaching concept, the special challenges of manufacturing companies are taken into account, and the participants are thus taught the necessary competencies for solving machine learning problems in industrial practice.

### KoMPI

#### (Deuse, Reference number BMBF 02P15A066) (Funding period: 2017–2019)

Within the scope of the project “Facility-based, digital planning of collaborative assembly systems and integration into variable production scenarios”, a methodology for the planning process of human-robot collaboration systems supporting the entire development process is to be developed. Suitable analogue and digital work aids should enable companies with little experience and limited resources in particular to operate collaborative work systems economically.

### PHASE

#### (Deuse, Reference number ZF4101110LF7) (Funding period: 2017–2019)

As part of the “Personalized hybrid Assembly as a Service” (PHASE) subproject of the international collaborative research project “Manufacturing and Assembly as a Service” (MAaaS), various services for reconfiguring cyber-physical production systems are developed and validated in a hybrid assembly system.

### ROBOTOP

#### (Deuse, Reference number 01MA17009H) (Funding period: 2017–2020)

The development of a modular, internet-based and open robot platform (ROBOTOP) serves to open up the mass market for robots in service and production applications. Through intelligent standardization and reuse of software, hardware and peripheral components, as well as a significant reduction in supply and engineering costs, significant cost reductions can be achieved in the planning and design of industrial robotics solutions.

### Sysmag

#### (Deuse, Reference number AIF 19185 N/1) (Funding period: 2017–2019)

The research project “Development of a planning system for standardized material supply in the complex, varied assembly of large equipment in small quantities” (Sysmag) develops a system for planning and controlling standardized material provision (MBS) in the single-part and small-series production of large-scale equipment.

**VariPro****(Deuse, Reference number IGF-Vorhaben 19683 N) (Funding period: 2017–2019)**

The aim of the research project VariPro, “Variability-based machine allocation planning for customer-order-specific production in SMEs”, is to develop variability-based machine allocation planning for SMEs in customer-order-specific production that will be used in operative production planning. This methodology enables users to make profitable use of value-adding variability so that a considerable increase in efficiency is achieved in MBP and thus in production. The Institute of Production Systems is the executive research unit.

**UA Ruhr-Professur “Virtual Machining”****(Wiederkehr, Reference number MERCATOR Pe-2016-0024) (Funding period: 2017–2022)**

The goal of this project is to establish a new research centre focusing on the simulation and optimisation of machining processes. The establishment of a central contact point for industrial companies is planned, which will enable the transfer into industrial applications. There is no overlap with the Collaborative Research Centre, since no machine learning aspects or resource restrictions are considered.

**Modelling and Simulation of the NC Grinding Process for the Controlled Generation of Work-piece Surfaces Under Consideration of Tool Topography and Wear****(Wiederkehr, Reference number DFG WI 4762/5-1) (Funding period: 2017–2019)**

For the analysis and optimisation of cutting processes with geometrically undefined cutting edges, a simulation system is being developed that will be able to model the removed material for complete grinding processes as well as the influences of different shapes of grains on the process forces and surface topography. There is no overlap with the Collaborative Research Centre, since a different manufacturing process is analysed and no machine learning methods or restrictions regarding resources are investigated.

**Adaption Intelligence of Factories in a Dynamic and Complex Environment (Spokesperson: Prof. Dr. Jakob Rehof)****(Wiederkehr, Reference number DFG GRK 2193) (Funding period: 2016–2020)**

In this project, a simulation system for the analysis of milling processes is qualified for application in the factory planning process. This project supports a collaborative PhD candidate position with Dirk Biermann. There is no overlap with the Collaborative Research Centre as no machine learning aspects are considered and the resources are not restricted.

**Stochastic Modeling of the Interaction of Tool Wear and the Machining Affected Zone in Nickel-Based Superalloys, and Application in Dynamic Stability****(Wiederkehr, Reference number DFG WI 4762/7-1) (Funding period: 2018–2021)**

The objective of this project is the development of wear-evolution models for machining nickel-based super alloys based on contrived tool wear experiments, decoupling the variability from the process force and surface quality models. The project was submitted together with Laine Mears of the Clemson University, South Carolina, USA. Since the integration of machine learning methods is not investigated and there are no resource restrictions, there is no overlap with the Collaborative Research Centre.

**Variety, Veracity, Value: Handling the Multiplicity of Urban Sensors (VaVeL)****(Morik, Reference number Horizon2020-688380) (Funding period: 2016–2020)**

The goal of the VaVeL project is to advance our ability to use urban data in applications that can identify and address citizen needs and improve urban life. The motivation comes from problems in urban transportation. Basic research from A1 has been applied within this European project. This is one way of transferring basic research to real-world applications. At the same time, the data from the city of Dublin and the city of Warsaw have been used by our learning methods. In particular, spatio-temporal random fields have been successfully applied within VaVeL.

**Modellierung von Themen und Strukturen religiöser Online-Kommunikation**  
**(Morik, Reference number MERCUR, PR-2015-0046) (Funding period: 2016–2018)**  
The project addresses two main questions: What are the structures of religious communication in online contexts, and how do religious topics spread across these structures? The study is based on computer-mediated communication (online forums, social media) of neo-conservative Christian and Muslim groups, e.g., Evangelical and Salafi communities. Text data in social media challenge the efficiency of learning. We have developed a more efficient learning of low-rank representations based on convex optimisation. Instead of explicitly learning low-dimensional features, we compute a low-rank representation implicitly by regularising full-dimensional solutions. Lukas Pfahler, the PhD student from this expiring project, has become a member of the CRC 876, who is financed by the university.

**Synthese von maschinellem Lernen und numerischer Simulation zur Echtzeitsteuerung**  
**(Morik, Reference number MERCUR, PR-2016-0039) (Funding period: 2017–2019)**  
As a collaboration with the CRC 837, this MERCUR project investigates the use of machine learning for real-time prediction. The physical relationships obtained from a simulation model and the knowledge gained through process-accompanying data analysis from monitoring and measurement data are merged in order to significantly improve process control in mechanical tunnel construction. The project is related to our work in B3 and meetings between the groups have taken place. The particular work on the tunnel data is on time series abstractions through clustering. The active learning framework from B3 could also be applied to data from tunnel processes.

**Kompetenzzentrum maschinelles Lernen Rhein Ruhr – ML2R**  
**(Morik, Reference number BMBF) (Funding period: 2018–2022)**  
The German Federal government has accepted four competence centres for machine learning which have the double function of achieving scientific excellence and transferring results into practice. Of course, stimulating discussions between members of the CRC876 who work on machine learning and members of ML2R will be possible. This may strengthen Dortmund and attract excellent scientists.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	3	193,500	3	193,500	3	193,500	3	193,500
Total	—	193,500	—	193,500	—	193,500	—	193,500
Direct costs	Sum		Sum		Sum		Sum	
Instrumentation up to 10,000 euros, software and supplies	3,900		3,100		2,200		1,000	
Total	3,900		3,100		2,200		1,000	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	197,400		196,600		195,700		194,500	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Jochen Deuse, Prof. Dr.-Ing., professor	Production systems	TU Dortmund	4	—	Existing funds
	2	Katharina Morik, Prof. Dr., professor	Data mining	TU Dortmund	4	—	Existing funds
	3	Petra Wiederkehr, Prof. Dr.-Ing. Dipl.-Inform., professor	Virtual machining	TU Dortmund	4	—	Existing funds
	4	N.N., doctoral researcher	Data mining	TU Dortmund	19.92	—	Existing funds
	5	N.N., doctoral researcher	Virtual machining	TU Dortmund	19.92	—	Existing funds
	6	N.N., doctoral researcher	Production systems	TU Dortmund	9.96	—	Existing funds
	7	Mario Wiegand, M.Sc., doctoral researcher	Production systems	TU Dortmund	9.96	—	Existing funds
	8	N.N., student assistant	Data mining	TU Dortmund	8	—	Existing funds
	9	N.N., student assistant	Production systems	TU Dortmund	8	—	Existing funds
	10	Florian Priebe, B.Sc., student assistant	Data mining	TU Dortmund	10	—	Existing funds
Non-research staff	11	N.N., secretary	—	TU Dortmund	2	—	Existing funds
<b>Requested staff</b>							
Research staff	12	Felix Finkeldey, M.Sc., doctoral researcher	Virtual machining	TU Dortmund	—	Doctoral researcher	—
	13	Amal Saadallah, M.Sc., doctoral researcher	Data mining	TU Dortmund	—	Doctoral researcher	—
	14	Jacqueline Schmitt, M.Sc., doctoral researcher	Production systems	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):****1. Deuse, Jochen**

Is in charge of the scientific supervision in the field of “Production Systems” and project management. Supports the work packages on real-time learning and model adaption, online management of many models, prediction based quality gate design for process chains and quality prediction for bottleneck relief in EOL-testing with his expertise.

**2. Morik, Katharina**

Project management. Main research focus on online learning and model management. Assistance in WPs: Real-time aggregation and feature extraction, Real-time learning and model adaption, online management of many models and combination of machine learning methods and process simulation systems.

**3. Wiederkehr, Petra**

Is in charge of the scientific supervision in the field of “Virtual Machining” and project management. With her expertise in the field of modelling and optimising machining operations as well as the combination of computer science and mechanical engineering, she contributes significantly to the design of solutions in all areas of activity. She represents the project externally and within the Collaborative Research Centre.

**4. N.N.**

Research work focus on data aggregation with respect to online change detection, adaptive memory management of learning models and model clustering as a tool for efficient online models selection or aggregation.

**5. N.N.**

Works on the development of process simulation approaches in WP Combination of machine learning methods and process simulation systems.

**6. N.N.**

Identification of relevant product and process parameters; positioning of quality gates.

**7. Wiegand, Mario**

Works on quality prediction for bottleneck relief in EOL-testing and supports in application to industrial use cases.

**8. N.N.****9. N.N.**

Supports with literature research and implementation in the field of prediction based quality gate design for process chains and quality prediction for bottleneck relief in EOL-testing.

**10. Priebe, Florian**

Works on learning from sensor data for an industrial use case with assistance in WP: real-time learning and model adaption.

**11. N.N.**

Takes care of travel and appointment planning, supports administrative activities, documentation and public relations.

**Job descriptions of staff for the proposed funding period (requested funds):****12. Finkeldey, Felix**

Investigation and development of novel approaches for the improvement of process simulation systems with regard to machine learning methods and strategies for optimising milling processes in real-time. This includes the conduction and systematic analysis of machining experiments. The main responsibilities comprise the investigation of the research questions in the context of WPs: Combination of machine learning methods and process simulation systems and Online optimisation of NC-milling processes.

**13. Saadallah, Amal**

Research work focus on online selection of data features representations, real-time learning and model adaption with respect to changes and drifts occurrence and online model management of many models for dynamically changing data by exploring different combination rules and adaptive ensemble methods constructions.

**14. Schmitt, Jacqueline**

Research in the field of prediction based quality gate design for process chains and quality prediction for bottleneck relief in EOL-testing. Also supports the work packages on real-time learning and model adaption and online management of many models. Investigates the application of developed methods onto different industrial use cases.

**3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from University	8,000	8,000	8,000	8,000
Sum of existing funds	8,000	8,000	8,000	8,000
Sum of requested funds	3,900	3,100	2,200	1,000

(All figures in euros)

Instrumentation up to 10,000 euros, software and supplies for financial year 2019

Material for milling experiments for all four years of the funding period. It is reasonable to acquire the material for the complete funding period in the first year to ensure that it originates out of the same charge. To investigate tool wear mechanics, the hardened high-speed steel 1.3344 will be used to provoke wear with minimal experimental effort.  For further investigations of adhesive wear effects, the analysis of the process stability and the construction of demonstrator workpieces, aluminum 7075 will be used, which is utilised, e.g., to machine structural components in the aerospace industry.	EUR	2,500
Tools for milling operations to conduct fundamental investigations which are necessary in order to develop simulation models and approaches to combine simulation components with machine learning methods and the training of models.	EUR	1,400

Instrumentation up to 10,000 euros, software and supplies for financial year 2020

Tools for experiments to investigate wear mechanics. For the investigation of a reasonable amount of process parameter values, several tools of the same type are needed.	EUR	3,100
---	-----	-------

Instrumentation up to 10,000 euros, software and supplies for financial year 2021

Tools for the investigations that focus the interactions between process stability and tool wear. While in the first two years one single effect (tool wear of process dynamics) will be considered, in the third year the complexity will be increased by analysing the influence of tool wear on process stabilities and vice versa.	EUR	2,200
--	-----	-------

Instrumentation up to 10,000 euros, software and supplies for financial year 2022

Tools for experiments, which target the transfer and evaluation of the developed approaches to more complex demonstrator components, which have to be manufactured. In particular, the feedback of the model predictions to the process will be evaluated.	EUR	1,000
--	-----	-------

### 3.7.5 Requested funding for instrumentation for the new funding period

This project does not request any funding for major research instrumentation.



### 3.1 General information about Project B4

### 3.1.1 Project title:

Analysis and Communication for Dynamic Traffic Prognosis

### 3.1.2 Research area(s):

407-04 (Traffic and Transport Systems, Intelligent and Automated Traffic), 408-02 (Communications, High-Frequency and Network Technology)

### **3.1.3 Principal investigator(s)**

Liebig, Thomas, Dr., 25.10.1980, German

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

Phone: 0231-755-8257

E-mail: thomas.liebig@tu-dortmund.de

Schreckenberg, Michael, Prof. Dr., 28.09.1956, German

Physik von Transport und Verkehr, Fakultät für Physik Universität  
Duisburg-Essen  
Lotharstraße 1  
47048 Duisburg

Phone: 0203-37-93552

E-mail: michael.schreckenberg@uni-due.de

Wietfeld, Christian, Prof. Dr., 22.01.1966, German

( ) no (x) yes

Do any of the above mentioned persons hold fixed-term positions?

Do any of the above mentioned persons hold fixed term  
(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes    (x) no
2.	clinical trials	( ) yes    (x) no
3.	experiments involving vertebrates.	( ) yes    (x) no
4.	experiments involving recombinant DNA.	( ) yes    (x) no
5.	research involving human embryonic stem cells.	( ) yes    (x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes    (x) no

## 3.2 Summary

While the vision of fully automated road traffic implies the total disappearance of traffic jams, the implications of mixed traffic on the traffic flow are still unclear. Currently, the presented automated vehicles react defensively to human misbehaviour. Therefore, a significant reduction of the traffic flow is anticipated for the hybrid traffic - while the overall traffic congestion is expected to grow through empty runnings, especially for the commercial transport. In the following, the term *automated vehicle* describes a vehicle that drives without a human driver but still needs a fixed start- and endpoint as well as a route definition. This means that the car is able to change lanes to overtake, but it does not take turns on its own accord for a more efficient route.

Based on the results of the second phase, the traffic flow will be predicted and optimised in different environments (e.g., highway, inner city) with increasing grades of automation by cooperation of traffic modeling, data analysis, and communication. The overall aim is to minimise the total dwell time of all vehicles inside the traffic network. For this purpose, the microscopic traffic simulation models will be extended by communicating automated vehicles in order to allow the simulation of different realizations of the behaviour of these vehicles and to elaborate the potentials for further optimisation. The traffic prediction is then performed based on these models.

In so-called closed-loop-scenarios, strategies for traffic control will be developed with deep reinforcement learning methods that also consider human drivers, which only follow the developed model sporadically. The central as well as the local exchange of the required data will follow the privacy-by-design paradigm in order to protect the distribution of sensitive mobility patterns, for example, through application of homeomorphic encryption.

For the provision of Ultra-reliable Low Latency Communication (URLLC) in a Vehicle-to-Everything (V2X) context, available knowledge about the mobility behaviour of vehicles and information about the topology of the surrounding environment will be leveraged for predictive routing, handover, and resource allocation as well as dynamic antenna steering in order to optimise decision processes within 5G networks. The respective methods will furthermore be integrated into the predictive Channel-aware Transmission (pCAT) scheme that was developed in the second phase and evaluated in field tests.

Since Kristian Kersting has followed a call for a W3 professorship of the TU Darmstadt, he will not be available as a principal investigator in the third phase. Nevertheless, since the research topics of this project provide an ideal field of application for his developed methods, further cooperation (e.g., co-supervised thesis) with Kristian Kersting will take place in the future.

In the upcoming phase, Thomas Liebig, who is also working at the Artificial Intelligence Group of the TU Dortmund, will cover the data analysis aspects of this project. He will continue the work

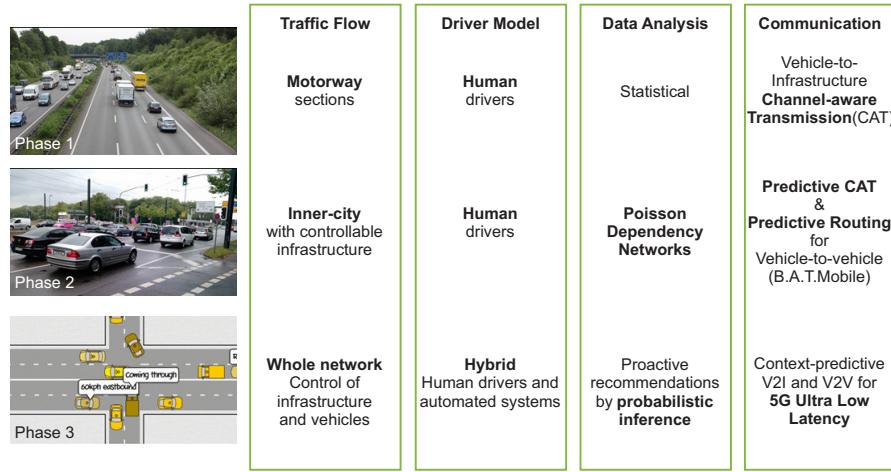


Figure 3.1: Continuity of the research topics over the different phases

of Kristian Kersting with the development of PSPNs and Gaussian models for traffic prognosis and Bandit Learning for routing optimisation. He has gained expertise in analysing traffic data within the EU projects GeoPKDD, MODAP, ESS, INSIGHT, and VaVeL and the BmBF project VASA and has already been integrated into a cooperative publication within project B4 and previous publications with B3 and A1.

### 3.3 Project progress to date

#### 3.3.1 Report and current state of research

In the second phase, three main research priorities were addressed and evaluated in the inner city: (1) the resource-efficient provision of vehicular sensor data using the Long Term Evolution (LTE) standard, (2) lane-accurate localisation, and (3) improved traffic prognosis and optimization. The continuity of the research topics and methods for the project over the different phases of the collaborative research centre is illustrated in figure 3.1.

**Data Acquisition and Processing (WP1):** The first work package of phase 2 dealt with the acquisition and processing of traffic data. On the topic of traffic research, a recent problem is the supplement of missing data from detectors on highways to increase the quality of short-term traffic forecasting. There are different attempts to achieve short-term traffic forecasting (see [282] for a short summary). Many methods work with microscopic traffic flow simulations. Those simulations are based on the real-world topology and real-time traffic data. Especially flow data at on- and off-ramps is important for the reliability.

Missing data of multiple minutes can have a big impact on the quality of the simulations. To increase the quality of the predictions, missing data is compensated by historical data or information from neighbouring detectors, if present. For this purpose, we developed a method [B4/291] to replace missing data based on historical data and the Poisson Dependency Networks (PDN). A test with data samples from the Cologne orbital motorway network in Germany showed that the PDN compared to an exponential smoothing approach needed significant less historical data (1 week compared to 30 weeks that the exponential smoothing approach needed) to supplement missing data. Also, the root-mean-square error and its normalized variant between all predicted traffic flows and observed traffic flows are reduced by about 9%. This improved method will be

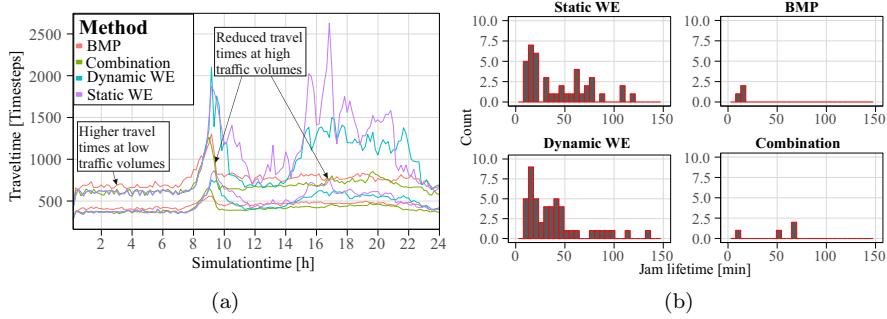


Figure 3.2: Results of a simulated day. a) Average travel time in the network at different times of the day for one day for the different DTA methods. The upper lines marks the value higher than 95% of all travel times and the lower line shows the average travel time over 10 minutes. b) Count of different jam lifetimes for the different DTA methods for one day.

applied on the traffic data that serves as a basis for the simulations in phase 3, to increase the realism of these simulations.

**Optimisation of Travel Times (WP2):** After increasing the accuracy of supplemented missing detector data, in WP 2 we focused our research on the emergence and lifetime of traffic jams. The behaviour of traffic jams can be analysed numerically as in the studies of Nagel [275] or with an analytical approach like that of Gerwinski [262], where the critical density at which jams stabilises was calculated. Recently [252] contributed a probability distribution function (PDF) for vehicle velocities and a velocity-velocity covariance function to identify transitions between free and congested flow. It concludes a probability to find a standing vehicle as a potential order parameter. Another approach is presented in [265] where flow, density, and velocity time series for approximately 2900 days were considered to define three rather stable traffic states by looking into the cross-correlations of these time series. To analyse traffic jams, in [B4/287], we combined the knowledge gained from [275] and [262]. As a result, we simulated traffic jams in a Nagel-Schreckenberg cellular automaton [274] and presented an approach to determine the critical density value at which traffic jams stabilize based on a stability criterion. Further, we were able to show that the suggested power law behaviour for low densities – presented in [262] – originates from the formation of new jams rather than an increased lifetime of existing jams.

Currently we are testing different dynamic traffic assignment (DTA) theories (see [277] for a review) to reduce the number of traffic jams and the average travel time of the system. A great number of theories and models have been presented in the last years that use different approaches. Some of these models are based on node weighting like [261] while others are trying to find an optimal analytic distribution like the "breakdown minimisation principle" (BMP) [267].

In our work in [B4/294], we compared the BMP theory to the simplest method of trip planning that considers the current traffic situation. This trip planning method simply chooses the route with the shortest travel time based on the current traffic situation. For simplicity, we will call it *shortest travel time trip planning* (STTP). The simulation is a cellular automata model [268] and is based on the main roads of the inner city of Duesseldorf. The traffic flow is adapted to the commuter characteristics of the city, with more people commuting into the city for work in the morning and out of the city after work in the afternoon rather than in the opposite direction. We applied the STTP once statically and once dynamically. Static means that every vehicle gets the route with the shortest travel time at the time step it is spawned in the system. Dynamical means that every

time the vehicle drives over a crossing, the route is updated to the route with the shortest travel time at this time step. The results of our simulations are presented for an exemplary simulated day in figure 3.2a and 3.2b. We found out that BMP creates far less traffic jams (most of the time not even one), which greatly reduces the global travel time at times of high traffic flows. The STTP (static and dynamical) routes traffic flow at low traffic volumes better, and existing traffic jams are accounted for because they increase the travel times on the jammed roads. The BMP also needs a preselection of the routes used to prevent high travel times at low traffic volumes and more information than the STTP on the form of breakdown probabilities for each bottleneck. In addition to the study on the BMP and the STTP, we created a new traffic routing method that aims to combine the advantages of the BMP and the static STTP. The combination is presented in figure 3.2a in green. We state that it is reducing the average travel time at times of low traffic volume as well as it reduces the number of breakdowns at high traffic volumes. In phase 3, the results of this work package will be combined with the results of phase 1 as well as WPs 3 and 6 in order to create a simulation that is able to simulate and manipulate inner-city as well as highway traffic.

**Data Analysis and Aggregation (WP3):** In phase 1, we proposed Poisson Dependency Networks (PDN) for traffic imputation at unobserved locations to respect the natural properties of traffic flow and density to be count values and follow a Poisson distribution. In the second phase, the tractability of the prediction model was considered, and Poisson-Sum-Product networks were introduced [B4/A6/174]. These models hold structural similarities to Sum-Product networks [278] and similarly, consist of random variables connected by sum and product nodes. A major extension is the Poisson distributed variables at the leaves, which allows for a tractable inference in a multivariate Poisson distribution. In comparison to PDN, SVM, and LDA, the Poisson-Sum-Product networks performed very well in a traffic imputation scenario [B4/A6/174].

Traffic densities only depend on count values and are often reported as fractions of vehicles per street length by traffic information systems. Thus, a further extension towards Manichean Sum-Product networks (Manichean refers to probabilistic models that combine two different probability distributions in one model) have also been introduced; in future cooperation with Kristian Kersting (TU Darmstadt) we envision its development towards a probabilistic traffic model. In [292], we highlight how reinforcement learning can be used in routing to reduce individual travel times and increase of network performance. In a realistic simulation scenario, we validate the algorithm versus a NASH equilibrium [280] and the standard shortest path routing by the simulation with various information from static sensors or moving vehicles under consideration of different penetration rates. The average individual travel time is decreased while reducing waiting times and increasing vehicle throughput of the whole street network. This validation guides towards our plans in the third phase where different ratios of automated and human-driven vehicles are the focus of our analysis. The work received the best paper award at COSIT 2017. The Spatio-Temporal random field developed in A1 was also used in situation-aware trip planning within the European VaVeL project [A1/C3/293]. The graphical model is used to estimate future sensor readings from past and current observations. Combined with a spatial regression (Gaussian Process Regression) we estimate future traffic flow at arbitrary places of the traffic network. These real-time predictions are exploited by a trip-planning component that finds a route preventing the predicted routes. A system that just produces predictions on the basis of historical observations without consideration of the impact of its suggestions could eventually create novel unexpected congestions. In phase 3, we target this by studying the combination of reinforcement learning and the breakdown minimisation principle.

**Efficient Bidirectional Real-time Communication (WP4):** Since modern cars are equipped with a wide range of different on-board sensor modules, their relevance is expanding from being pure means of transportation to leverage them as mobile sensor nodes that provide highly up-to-date data for crowdsensing-based services [286] in a smart-city context. Apart from using location information for dynamic traffic forecasting and control [257], vehicles can be exploited to acquire

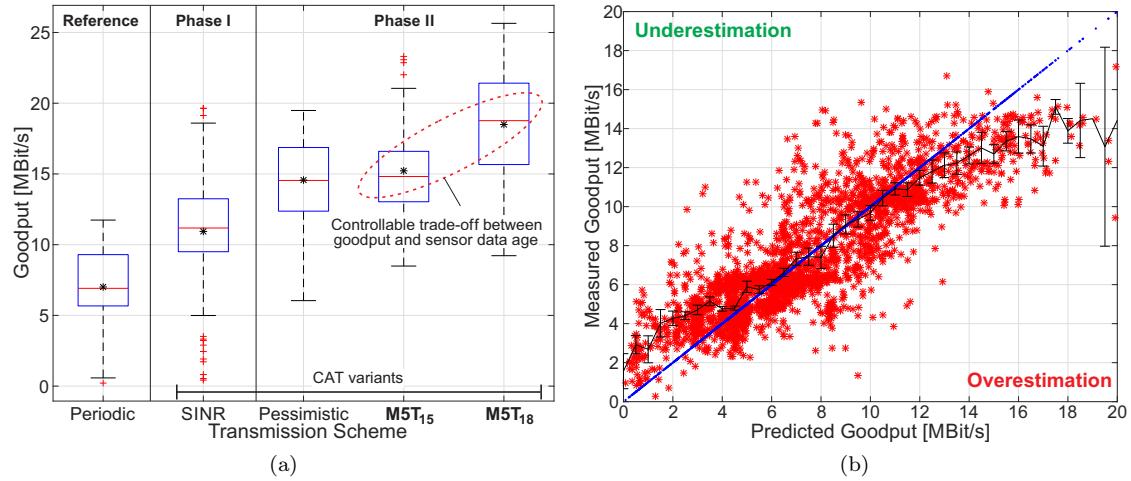


Figure 3.3: Selected results of the second phase regarding the use of machine learning applied to the Channel-Aware Communication (CAT) mechanism (evolved from Phase 1) (c.f. [B4/A4/290])

data that is highly relevant for applications such as road-terrain classification [283] and distributed weather forecast [256].

Within this project, the required data transmission is performed using extended Floating Car Data (xFCD) based on the cellular Long Term Evolution (LTE) standard in public cellular networks. Consequently, resource efficient transmission is highly important in order to minimise interference between this kind of Machine-type Communication (MTC) and Human-to-human (H2H) data traffic [259].

For addressing these challenges, the context-aware communication paradigm [103] has been proposed, taking knowledge about the channel quality and mobility conditions into account to optimise all communication-related decision processes. Machine learning mechanisms are applied [264] to acquire highly accurate predictions of the future state of relevant key performance indicators. Within the first phase of this project, the probabilistic *Channel-aware Transmission (CAT)* scheme was introduced, which schedules data transmissions with respect to the current channel quality based on *Signal-to-noise-plus-interference-ratio (SINR)* measurements in order to increase the resource efficiency by avoiding resource-intense transmissions.

In the second phase and in the context of WP4, the scheme was extended to *predictive CAT (pCAT)* and validated by simulations and field evaluations using real vehicle data obtained from the Controller Area Network (CAN) bus (published in IEEE Transactions on Vehicular Technology). Due to the integration of a-priori information about the SINR behaviour along the predicted route of the vehicle, future connectivity hotspots are proactively identified and exploited which significantly increases the mean throughput (from 42.5% gain with CAT to 61% increase using pCAT). However, in the inner-city scenario, the impact of multipath propagation is much higher than within the highway scenario because of high densities of surrounding buildings and moving obstacles, reducing the significance of the SINR to estimate the channel properties. To overcome this issue, CAT was extended using a context-predictive multi-metric approach that simultaneously takes into account all available passive downlink indicators as well as the size of the data packets by machine learning based data rate prediction (joint work with A4). The transmission decision itself is then performed based on the anticipated data rate instead of a single indicator. In comprehensive field evaluations with a total driven distance of more than 1000 km, we were able to increase the mean data rate by up to 164% (from 6.98 MBit/s to 18.5 MBit/s as shown in figure 3.3a) compared to periodic transmissions [B4/A4/290] (Best Student Paper Award at IEEE VTC Spring 2018). In contrast to other approaches like [271] and [260] that handle resource efficiency in a centralised

manner within the base station, the great advantage of CAT is its decentralised design that does not require communicating additional context information between different network participants.

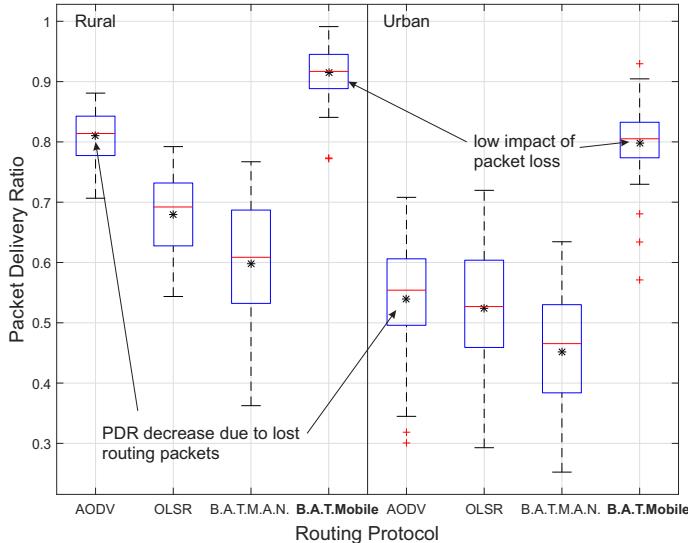


Figure 3.4: Novel mobility-predictive routing protocol B.A.T.Mobile (c) (c.f. [B4/289])

next phase of the Collaborative Research Centre, the principles of CAT and B.A.T.Mobile will be brought together in a 5G context exploiting novel opportunities such as predictive beam steering [279], cellular direct short-range [273] and multi-interface communication [250].

Since the relevance of context-predictive communication goes beyond the considered vehicular scenario, existing methods will be applied from and novel models will be transferred into adjacent research fields. In particular, Unmanned Aerial Vehicle (UAV) networks address similar communication-related challenges in a higher-dimensional mobility context, making those types of vehicles interesting candidates for the augmentation of upcoming Intelligent Traffic Control (ITCS) systems [272].

**Lane-specific Positioning in Inner-city Scenarios (WP5):** Within the B4 project, access to precise and lane-specific location information of individual vehicles is of crucial importance since it is a basic requirement for accurate traffic flow predictions as well as for context-predictive communication. Regular GPS-based positioning methods are not able to fulfill this condition and are severely distorted by multipath effects, especially in interference-intensive environments such as urban canyons. For addressing this issue, methods to optimise the positioning accuracy were developed in the first two phases.

During the last funding period and within WP5, *LOcal interference CompensATion (LOCATE)* was developed to predict all influences, quantify the accruing shift, and compensate the overall error vector. In comparison with stand-alone GPS, LOCATE increases the positioning accuracy by an average of 45%. However, LOCATE includes the very resource-intensive Ray-Tracing technology and is not applicable to tasks that require approximately real-time processing. In order to reduce the required computing time to potentially reach real-time capabilities and to further improve the positioning accuracy, LOCATE was extended to real-time LOCATE (rt-LOCATE). Two further modules, predictive LOCATE (pLOCATE) and differential LOCATE (dLOCATE), were added to the basic LOCATE approach.

As groundwork for the inter-vehicle communication in the next phase, we proposed the novel biology-inspired routing protocol *B.A.T.Mobile* for vehicular mesh networks [B4/289], which leverages cross-layer knowledge about the future trajectory for routing data packets over communication paths with a high temporal robustness. In contrast to other approaches that make use of centralised path determination [276], all routing decisions are performed in a decentralised manner and do not require complete knowledge about the network topology. Therefore, it achieves a significantly higher reliability in challenging environments compared to the established protocols, as shown in figure 3.4c. Since the protocol monitors multiple paths in parallel, it has further been extended to supply decentralized load balancing using multipath communication.

Based on the latest set of Two Line Elements (TLE), pLOCATe estimates future satellite positions and runs Ray-Tracing analyses for future constellations in advance to predict the multipath error, which is stored in relational databases. This prediction replaces the complex and resource-intensive Ray-Tracing analysis with a simple database access during runtime.

Analogous to Differential GNSS, dLOCATe uses two antennas with an exactly known distance to each other. By measuring a position with both antennas in parallel, the real known distance between both antennas is used to correct the probably smaller or higher distance deviation.

After minimising the positioning error with rt-LOCATe, a map-matching approach with highly detailed and up-to-date underlying map material is used to especially match error-affected and unrealistic position measurements to the correct driving lane. Map matching is a well-known technique in navigation devices, and with rt-LOCATe preprocessing, correct lane matching is even more probable, especially in challenging environments like urban canyons.

For an evaluation with scenario-specific considerations, the TU Dortmund university campus is used, with a highly detailed 3D model of the environment and two highly accurate surveyed reference points as measurement locations. A 3D model approach is still a hot research topic and improves the positioning accuracy [251] [263].

Fig. 3.5 classifies LOCATe and rt-LOCATe with map matching according to stand-alone GPS measurements and the theoretical maximum as best-case after compensating all atmospheric and local influences.

Compared to other positioning technologies like Precise Point Positioning [253], the interaction of all components of rt-LOCATe results in a similar positioning accuracy. For the third phase, we consider the satellite-based achieved precision accurate and robust enough for the vehicular scenario and will not work on further improvements within the B4 project. Nevertheless, we will consider the novel Ultra-wideband (UWB) positioning methods developed by the A4 project.

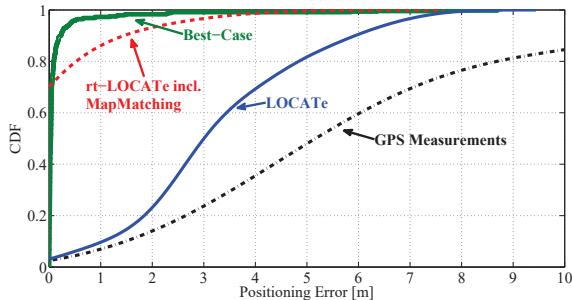


Figure 3.5: Classification of Accuracy Gain using our LOCATe approach (c.f. [B4/288])

#### System Validation by Simulation and Analysis in Inner-city Scenarios (WP6):

In the last work package, we created a complete system based on the results of the previous work packages to test system-relevant dynamical options. In different simulations, presented in [B4/294], we closed lanes or streets of the system at different times of the day to test the impact on the travel time for each routing method. We further created dynamic crossings where the function of lanes is changing depending on the time of the day to adapt to the changes in the traffic volume. For example, in the morning, there are more lanes leading into the city and in the afternoon the lanes are used

to drive out of the city. The most recent simulations show that dynamic lanes have little to no impact on travel times. In most cases, we found no improvements in the travel times because the increased capacity through the dynamic lanes only leads to bigger breakdowns at the following intersections. Only in very few cases, we were able to increase traffic flow through dynamic lanes at only one intersection. To find the streets on which dynamic lanes have the biggest impacts, we looked for bottlenecks that have an increased number of breakdowns compared to the other bottlenecks. There was no bottleneck that had an increased rate of breakdowns for all routing methods, since each method uses streets with a different weighting. A case where dynamic lanes

result in a decrease of the average travel time for the STTP is at an intersection that connects the city ring with the inner city. The simulation continued 100 days once with and once without the dynamic lanes. We were able to decrease the average travel time of the system by up to 8% with the static STTP, 5% with the dynamic STTP. The BMP and the combination of BMP and the STTP showed no improvement for this configuration.

In order to allow the simulation of cooperating automated systems and to bring vehicular mobility and communication networks together in a native way, we proposed the Lightweight ICT-centric mobility simulation (LIMoSim) [B4/123] framework. In contrast to state of the art approaches that couple multiple specialized simulators using Interprocess-Communication (IPC) [281], both components share a common codebase allowing direct information access and control. Consequently, knowledge about the vehicle's mobility behaviour can be leveraged directly to optimise communication processes without causing IPC-related overhead. The framework is therefore an enabler for the simulation of context-predictive models in the hybrid-traffic context of the third phase. In addition to the discussed publications, 27 further research papers (including IEEE Transactions on Vehicular Technology, IEEE GLOBECOM, IEEE VTC, IEEE VNC) were published and are omitted here due to the spatial constraints.

## Bibliography

- [250] K. Abboud, H. A. Omar, and W. Zhuang. "Interworking of DSRC and cellular network technologies for V2X communications: A survey". In: *IEEE Transactions on Vehicular Technology* 65.12 (Dec. 2016), pp. 9457–9470 (cit. on p. 249).
- [251] M. Adjrad and P. D. Groves. "Enhancing least squares GNSS positioning with 3D mapping without accurate prior knowledge". In: *Navigation* 64.1 (2017), pp. 75–91 (cit. on p. 250).
- [252] N. Bain, T. Emig, F.-J. Ulm, and M. Schreckenberg. "Velocity statistics of the Nagel-Schreckenberg model". In: *Phys. Rev. E* 93 (Feb. 2016), p. 022305 (cit. on p. 246).
- [253] P. F. de Bakker and C. C. Tiberius. "Real-time multi-GNSS single-frequency precise point positioning". In: *GPS Solutions* 21.4 (2017), pp. 1791–1803 (cit. on p. 250).
- [254] H. Bast, E. Carlsson, A. Eigenwillig, R. Geisberger, C. Harrelson, V. Raychev, and F. Viger. "Fast routing in very large public transportation networks using transfer patterns". In: *European Symposium on Algorithms*. Springer. 2010, pp. 290–301 (cit. on p. 260).
- [255] P. Buchholz and I. Felko. "PH-graphs for analyzing shortest path problems with correlated traveling times". In: *Computers & Operations Research* 59 (2015), pp. 51–65 (cit. on p. 264).
- [103] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer. "A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques". In: *IEEE Communications Surveys & Tutorials* 19.3 (2017), pp. 1790–1821 (cit. on pp. 130, 131, 143, 248).
- [256] C. T. Calafate, K. Cicenia, O. Alvear, J. C. Cano, and P. Manzoni. "Estimating rainfall intensity by using vehicles as sensors". In: *2017 Wireless Days*. Mar. 2017, pp. 21–26 (cit. on p. 248).
- [257] C. Chen, T. H. Luan, X. Guan, N. Lu, and Y. Liu. "Connected vehicular transportation: Data analytics and traffic-dependent networking". In: *IEEE Vehicular Technology Magazine* 12.3 (Sept. 2017), pp. 42–54 (cit. on p. 247).

- [258] M. P. Deisenroth and J. W. Ng. “Distributed Gaussian processes”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*. JMLR. org. 2015, pp. 1481–1490 (cit. on p. 255).
- [259] S. Djahel, R. Doolan, G. M. Muntean, and J. Murphy. “A communications-oriented perspective on traffic management systems for smart cities: Challenges and innovative approaches”. In: *IEEE Communications Surveys Tutorials* 17.1 (2015), pp. 125–151 (cit. on p. 248).
- [260] Z. Feng, Z. Feng, and T. A. Gulliver. “Biologically inspired two-stage resource management for machine-type communications in cellular networks”. In: *IEEE Transactions on Wireless Communications* 16.9 (Sept. 2017), pp. 5897–5910 (cit. on p. 248).
- [261] R. Geisberger, P. Sanders, D. Schultes, and D. Delling. “Contraction hierarchies: Faster and simpler hierarchical routing in road networks”. In: *Experimental Algorithms*. Ed. by C. C. McGeoch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 319–333 (cit. on p. 246).
- [262] M. Gerwinski and J. Krug. “Analytic approach to the critical density in cellular automata for traffic flow”. In: *Phys. Rev. E* 60 (July 1999), pp. 188–196 (cit. on p. 246).
- [263] Y. Gu and S. Kamijo. “GNSS positioning in deep urban city with 3D map and double reflection”. In: *Navigation Conference (ENC), 2017 European*. IEEE. 2017, pp. 84–90 (cit. on p. 250).
- [264] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, and L. Hanzo. “Machine learning paradigms for next-generation wireless networks”. In: *IEEE Wireless Communications* 24.2 (Apr. 2017), pp. 98–105 (cit. on p. 248).
- [265] J. W. Kantelhardt, M. Fullerton, M. Kämpf, C. Beltran-Ruiz, and F. Busch. “Phases of scaling and cross-correlation behavior in traffic”. In: *Physica A: Statistical Mechanics and its Applications* 392.22 (2013), pp. 5742 –5756 (cit. on p. 246).
- [266] B. S. Kerner. “Failure of classical traffic flow theories: Stochastic highway capacity and automatic driving”. In: *Physica A: Statistical Mechanics and its Applications* 450 (May 2016), pp. 700–747 (cit. on pp. 255, 257).
- [267] B. S. Kerner. “Breakdown minimization principle versus Wardrop’s equilibria for dynamic traffic assignment and control in traffic and transportation networks: A critical mini-review”. In: *Physica A: Statistical Mechanics and its Applications* 466 (2017), pp. 626 –662 (cit. on pp. 246, 260).
- [268] W. Knospe, L. Santen, A. Schadschneider, and M. Schreckenberg. “Towards a realistic microscopic description of highway traffic”. In: *Journal of Physics A: Mathematical and General* 33.48 (2000), p. L477 (cit. on p. 246).
- [269] T. Liebig, S. Peter, M. Grzenda, and K. Junosza-Szaniawski. “Dynamic Transfer Patterns for Fast Multi-modal Route Planning”. In: *Societal Geo-innovation: Selected papers of the 20th AGILE conference on Geographic Information Science*. Ed. by A. Bregt, T. Sarjakoski, R. van Lammeren, and F. Rip. Cham: Springer International Publishing, 2017, pp. 223–236 (cit. on p. 260).
- [270] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang. “Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction”. In: *Sensors* 17.4 (2017), p. 818 (cit. on p. 262).
- [271] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. K. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman. “Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms”. In: *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*. Apr. 2014, pp. 1339–1347 (cit. on p. 248).

- [272] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer. “UAV-enabled intelligent transportation systems for the smart city: Applications and challenges”. In: *IEEE Communications Magazine* 55.3 (Mar. 2017), pp. 22–28 (cit. on p. 249).
- [273] R. Molina-Masegosa and J. Gozalvez. “LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications”. In: *IEEE Vehicular Technology Magazine* 12.4 (Dec. 2017), pp. 30–39 (cit. on p. 249).
- [274] K. Nagel and M. Schreckenberg. “A cellular automaton model for freeway traffic”. In: *Journal de Physique* 1 2 (1992), pp. 2221 –2229 (cit. on p. 246).
- [275] K. Nagel. “Life times of simulated traffic jams”. In: *International Journal of Modern Physics C* 05.03 (1994), pp. 567–580 (cit. on p. 246).
- [276] S. Naimi, A. Busson, V. Vèque, L. B. Slama, and R. Bouallegue. “Anticipation of ETX metric to manage mobility in ad hoc wireless networks”. In: *Proceedings of the 13th International Conference on Ad-hoc, Mobile, and Wireless Networks - Volume 8487*. ADHOC-NOW 2014. New York, NY, USA: Springer-Verlag New York, Inc., 2014, pp. 29–42 (cit. on p. 249).
- [277] S. Peeta and A. K. Ziliaskopoulos. “Foundations of dynamic traffic assignment: The past, the present and the future”. In: *Networks and Spatial Economics* 1.3 (Sept. 2001), pp. 233–265 (cit. on p. 246).
- [278] H. Poon and P. Domingos. “Sum-product networks: A new deep architecture”. In: *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE. 2011, pp. 689–690 (cit. on p. 247).
- [279] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang. “Overview of millimeter wave communications for fifth-generation (5G) wireless networks - With a focus on propagation models”. In: *IEEE Transactions on Antennas and Propagation* 65.12 (Dec. 2017), pp. 6213–6230 (cit. on p. 249).
- [280] T. Roughgarden and É. Tardos. “How bad is selfish routing?” In: *Journal of the ACM (JACM)* 49.2 (2002), pp. 236–259 (cit. on p. 247).
- [281] C. Sommer, R. German, and F. Dressler. “Bidirectionally coupled network and road traffic simulation for improved IVC analysis”. In: *IEEE Transactions on Mobile Computing* 10.1 (Jan. 2011), pp. 3–15 (cit. on p. 251).
- [B3/246] **T. Liebig, M. Stolpe, and K. Morik.** “Distributed Traffic Flow Prediction with Label Proportions: From in-Network towards High Performance Computation with MPI”. In: *Proceedings of the 2nd International Workshop on Mining Urban Data (MUD2)*. Vol. 1392. CEUR-WS, 2015, pp. 36–43 (cit. on pp. 18, 21, 220, 224, 261).
- [282] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias. “Short-term traffic forecasting: Where we are and where we’re going”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 3 –19 (cit. on p. 245).
- [283] S. Wang, S. Kodagoda, L. Shi, and H. Wang. “Road-terrain classification for land vehicles: Employing an acceleration-based approach”. In: *IEEE Vehicular Technology Magazine* 12.3 (Sept. 2017), pp. 34–41 (cit. on p. 248).
- [284] T. Woopen. *UNICARagil*. 2018 (cit. on p. 255).
- [285] L. Ye and T. Yamamoto. “Modeling connected and autonomous vehicles in heterogeneous traffic flow”. In: *Physica A: Statistical Mechanics and its Applications* 490 (2018), pp. 269 –277 (cit. on p. 255).
- [286] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi. “Internet of things for smart cities”. In: *IEEE Internet of Things Journal* 1.1 (Feb. 2014), pp. 22–32 (cit. on p. 247).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [B4/287] H. M. Bette, **L. Habel**, T. Emig, and **M. Schreckenberg**. “Mechanisms of jamming in the Nagel-Schreckenberg model for traffic flow”. In: *Physical Review E* 95.012311 (Jan. 2017) (cit. on p. 246).
- [B4/A6/174] **A. Molina**, S. Natarajan, and **K. Kersting**. “Poisson Sum-Product Networks: A Deep Architecture for Tractable Multivariate Poisson Distributions”. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*. 2017, pp. 2357–2363 (cit. on pp. 159, 247).
- [B4/288] **B. Niehöfer**, **F. Schweikowski**, and **C. Wietfeld**. “LOcal interferenCe compensATion (LOCATE) for GNSS-based Lane-Specific Positioning of Vehicles”. In: *IEEE Vehicular Technology Conference (VTC-Spring)*. IEEE. Nanjing, China, May 2016 (cit. on p. 250).
- [B4/289] **B. Sliwa**, D. Behnke, **C. Ide**, and **C. Wietfeld**. “B.A.T.Mobile: Leveraging mobility control knowledge for efficient routing in mobile robotic networks”. In: *IEEE GLOBECOM 2016 Workshop on Wireless Networking, Control and Positioning of Unmanned Autonomous Vehicles (Wi-UAV)*. Washington D.C., USA: IEEE, Dec. 2016 (cit. on p. 249).
- [B4/A4/290] **B. Sliwa**, **T. Liebig**, **R. Falkenberg**, J. Pillmann, and **C. Wietfeld**. “Efficient machine-type communication using multi-metric context-awareness for cars used as mobile sensors in upcoming 5G networks”. In: *Proceedings of the 87th Vehicular Technology Conference: VTC2018-Spring*. IEEE. 2018 (cit. on pp. 17, 248).
- [B4/123] **B. Sliwa**, J. Pillmann, F. Eckermann, **L. Habel**, **M. Schreckenberg**, and **C. Wietfeld**. “Lightweight joint simulation of vehicular mobility and communication with LIMoSim”. In: *IEEE Vehicular Networking Conference (VNC)*. Torino, Italy, Nov. 2017 (cit. on pp. 141, 251).
- [B4/291] **L. Habel**, **A. Molina**, **T. Zaksek**, **K. Kersting**, and **M. Schreckenberg**. “Traffic simulations with empirical data – How to replace missing traffic flows?” In: *Traffic and Granular Flow '15*. Ed. by V. L. Knoop and W. Daamen. Springer, May 2016, pp. 491–498 (cit. on pp. 245, 318).
- [292] **T. Liebig** and **M. Sotzny**. “On Avoiding Traffic Jams with Dynamic Self-Organizing Trip Planning”. In: *13th International Conference on Spatial Information Theory (COSIT 2017)*. Ed. by E. Clementini, M. Donnelly, M. Yuan, C. Kray, P. Fogliaroni, and A. Ballatore. Vol. 86. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, 17:1–17:12 (cit. on p. 247).
- [A1/C3/293] **T. Liebig**, **N. Piatkowski**, **C. Bockermann**, and **K. Morik**. “Dynamic Route Planning with Real-Time Traffic Predictions”. In: *Information Systems* 64 (2017), pp. 258–265 (cit. on pp. 21, 247).
- [B4/294] **T. Vranken**, **B. Sliwa**, **C. Wietfeld**, and **M. Schreckenberg**. “Performance comparison of dynamic vehicle routing methods for minimizing the global dwell time in upcoming smart cities”. In: *2018 IEEE 88th IEEE Vehicular Technology Conference (VTC-Fall)*. Chicago, USA, Aug. 2018 (cit. on pp. 246, 250).

## 3.4 Project plan

### Goals

The overall goal of project B4 is the precise prognosis and optimisation of the vehicular traffic flow in scenarios with different grades of automation based on data analysis. The required information about the network dynamics is gathered from the vehicles and intelligent road infrastructure and transmitted resource-efficiently as extended Floating Car Data (xFCD) through cellular 5G networks and other communication technologies.

In the first phase, the simulation-based optimisation of the traffic flow was analysed on isolated motorway sections. Based on the obtained insights, the scenario was extended to complex road networks in inner-city scenarios that immanently introduce novel challenges for traffic flow forecast as well as for the resource-efficient communication in the second phase. While the introduction of automated vehicles into traffic creates many interesting opportunities, like those that are analysed in the "UNICARagil" project [284] for example, they also create new challenges that have to be mastered. One of these challenges is the so-called **hybrid vehicular traffic** (or heterogeneous vehicular traffic), the period of coexistence between human drivers and automated systems, which marks a highly challenging step on the transition to completely autonomous traffic due to the heterogeneity in mobility behaviour and the capabilities for mutual coordination of the vehicles. In the third phase, this project will perform fundamental research on this problem.

Most of the current theories and models for hybrid vehicular traffic, like for example [266], do not consider different behaviours of self-driven vehicles from person-driven ones or the reaction of one type of driver to the different (and thus not anticipated) behaviour of the other type. For example, [266] differentiates between the two types of vehicles only with different probability of deceleration, and [285] only considers the different reactions of self-driven vehicles to person-driven ones by an increased safety distance and does not consider different behaviours for person-driven vehicles. To describe and simulate hybrid vehicular traffic, the first step in phase 3 will be to extend the existing models and theories to better consider these points. After the extensions, scenarios that combine motorway sections and inner city systems have to be created to analyse the impact autonomous vehicles have on the traffic flow and the overall dwell time. Based on this research, the existing traffic flow forecast and controlling theories will be changed to adapt to hybrid vehicular traffic with the goal of minimising the overall dwell time. To further optimise traffic flow, different dynamic network infrastructures will be implemented. Since in phase 2, dynamic lanes were analysed and showed little use, and additional lanes are further problematic to build in cities because of the limited space, in phase 3 dynamic traffic lights will be analysed.

In phase 2 novel models for probabilistic modelling of traffic have been proposed. A major challenge of centralized traffic imputation is the tractability of the probabilistic model and its capability to run in parallel on a distributed high-performance cluster. The Sum-Product network is a Graphical Model that provides tractable inference with mixed distributions. In phase 3, this work is continued towards conditional Sum-Product networks that allow factorisation of the joint distribution and combination with cellular automaton models. For large-scale imputation of unobserved traffic values, we analyse distributed Gaussian process regression exploiting the approaches created in the second phase, i.e., Gaussian summary trees, coresets and the distributed Gaussian process approaches from literature (Product of Experts [258]). To avoid isolated applications of the prediction models, limited by the boundaries of the considered regions, we also focus on distributed prediction using vehicle-to-infrastructure communication. With a lightweight prediction model that utilizes learning from label proportions instead of broadcasting every single observation, the communication is reduced and predictions can be made *in situ* within a sensor mesh. Learning optimal control of traffic, e.g., by routing or signalling, based on historical observations is difficult as only the outcome of the actions taken can be monitored and not the result of all the alternative actions. This results in a trade-off between exploiting a strategy and exploring novel strategies.

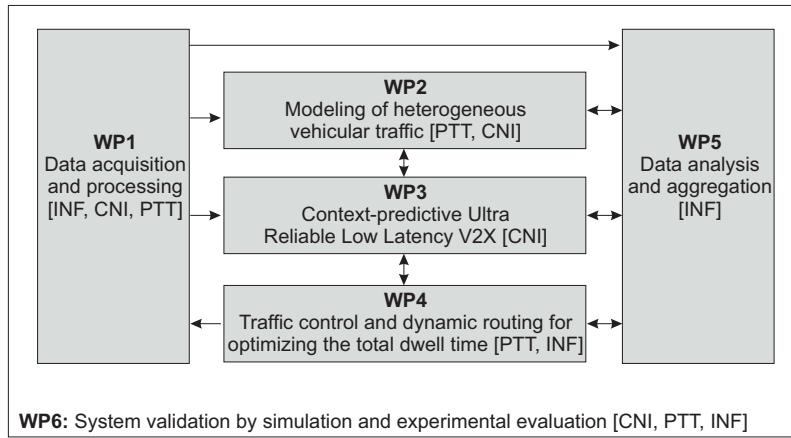


Figure 3.6: Work package planning for the third phase

If only predictions of future situations are considered, without estimation of the reaction of the traffic system to some control action, unexpected hazards could occur. Reinforcement learning, particularly bandit learning, tackles this closed-loop problem. In phase 3, bandit learning will be applied to improve traffic control by situation-aware routing and signalling. A key factor for the realization of these strategies in the hybrid traffic scenario is the availability of reliable and resource-efficient communication mechanisms. On the one hand, highly up-to-date sensor data needs to be transferred to cloud-based traffic management systems while minimising inference with human cell users. On the other hand, real-time traffic flow optimisation requires the propagation of routing adjustments and infrastructure control messages to their respective recipients. In addition, automated vehicles require ultra-reliable low-latency communication for coordinating their mobility behaviour. For addressing these heterogeneous challenges, novel context-predictive mechanisms for resource-efficient 5G communication will be developed in this project that leverage knowledge about the future transmission channel conditions for the anticipatory optimisation of resource allocation, handover and beam steering as well as the choice of communication technology and transmission time with respect to the application requirements.

For validation, the different components will be brought together into an overall simulation system that is capable of integrating live measurement data as well as real hardware using hardware-in-the-loop (HIL) simulation.

### Work schedule

The project plan is illustrated in Fig. 3.6 and illustrates the previously mentioned research goals. The WPs either are assigned to one of the research groups, Communication Networks Institute (CNI), Physics of Transport and Traffic (PTT), or Artificial Intelligence (INF) or are carried out jointly by multiple groups.

**Work package 1. Data acquisition and processing** For addressing the research topics of the hybrid traffic scenario, the B4 project will rely on available static and live data sources as well as on tools for performing own data acquisition campaigns.

- The city of Dublin is equipped with a SCATS system to monitor and control traffic in the city. Within the European VaVeL project, we collaborate with Dublin City Council and are granted real-time and historical access to the traffic loop data in the city of Dublin.

- Every truck over 7.5 tons in Germany is equipped with an on-board unit (OBU) that collects toll-relevant information while the truck is driving on the *Autobahn* or a *Bundesstrasse*. These data will be published by the government and is available for phase 3. This data can then be used to create an origin-destination matrix, which can be used to create realistic simulations.
- The real-time traffic information of around 6000 lanes of the "Bundesstrassenetz" of North Rhine-Westphalia is obtained from the "Mobilitäts Daten Marktplatz". This data is updated once a minute and contains the average velocity as well as the proportion of trucks on the lane.
- Radio-tomographical data from an experimental live deployment of the radio-based Wireless Detection and Warning System (WDWS) on the German highway A9 within an official test field by the German Federal Ministry of Transport and Digital Infrastructure. The system captures time-series data of the attenuation characteristics caused by vehicles passing a setup of multiple radio links. In cooperation with A1, we will analyse the multidimensional time-series data to classify different vehicle types based on their radio fingerprint and derive up-to-date information for traffic forecasting and control. Preliminary work has been presented at IEEE VTC Spring 2017 and IEEE SysCon 2018.
- Cellular network quality indicators in a vehicular context obtained by an application developed for Android and desktop operating systems that has been published as open source software.

**Work package 2. Modeling of heterogeneous vehicular traffic** The behaviour patterns of vehicles in phase 2 only differed between cars, trucks, and motorcycles. Here the focus lay on the analysis, description, and manipulation of the network. This now changes with the introduction of automated vehicles, which create many new challenges in traffic control because of their behaviour pattern that differs from that of non-automated vehicles. Automated vehicles have multiple advantages over human-driven ones. For example, automated vehicles do not dawdle like human drivers, who often need some additional time to start driving because they do not realize that a traffic light has turned green. This reduction or even removal of stochastic elements in traffic makes traffic flow more predictable and reduces the global dwell time. Because of this, it is assumed that in a 100% automated traffic the dwell time of vehicles and the average time of congestion in the system reaches a minimum. The challenges of automated traffic lie in the transitional period from 100% human-driven traffic to 100% automated traffic. Because the coordination between automated vehicles and human-driven ones is different to the interaction between pure human drivers, both types of vehicles have to react defensively in a situation with the other type. This slows them down and reduces the traffic flow in the mixed vehicular traffic. In this work package, a microscopic model for heterogeneous vehicular traffic will be created by identifying and characterizing the differences between automated and human-driven vehicles as well as the differences in their interaction. The microscopic model is particle-based, which means that every vehicle is simulated individually and decides if it wants to speed up, dawdle, and/or do other things. Through simulating every vehicle as a single particle, effects like interactions between two vehicles can better be analysed and understood than in a macroscopic model where only the traffic flow is modelled and the decisions of single vehicles are less important. Previous microscopic traffic models based on a cellular automaton do not include automated vehicles or simplify the differences between them and human-driven vehicles to remove stochastic elements [266].

The first step will be to identify and parameterize the behaviour of automated vehicles in order to simulate them. Methods for reliable communication that are developed in WP3 will be applied for allowing coordination among multiple vehicles. The adapted model is then able to simulate either automated or non-automated vehicles, and the first differences in the traffic flow can be identified. Afterwards, the reaction of automated vehicles to non-automated ones and vice versa

will be characterized in the model in order to achieve the simulation of heterogeneous traffic. Besides the reduced dawdling of automated vehicles, these vehicles are more likely than human-driven ones to follow a travel plan. Thus, besides signalling, individual routing of the vehicles is an important control strategy to reduce dwell times. In preparation for the envisioned smart routing in WP4 (a combination of the minimum breakdown principle with reinforcement learning), in this work package simulated data is exploited to train and validate data mining methods from WP5 for probabilistic traffic imputation at unobserved locations. The results of this work package will be used in WP4 to create routing and traffic control methods with the aim to minimise the global dwell time in heterogeneous vehicular traffic. They will further be used as a basis for the simulations in WP6, which, in turn, will test and evaluate them. Since automated vehicles cannot interact with other vehicles as human drivers can, stable communication between vehicles is an important step for the introduction of automated vehicles. Furthermore, to efficiently apply dynamic network changes and let the automated vehicles react to this, reliable communication between the vehicles and the infrastructure is needed. This means that the behaviour of automated vehicles depends on the vehicle-vehicle and the vehicle-infrastructure communication method, which is investigated in WP3 and integrated into the model.

**Work package 3. Context-predictive Ultra Reliable Low Latency V2X** In this work package, the novel challenges for the communication arising from the mixture traffic scenario are addressed. In parallel to the transition from 4G to 5G networks on the technology side, we move from context-aware to context-predictive communication on the methodological side. While A4 provides cost models for communication technologies and methods for assessing the current network quality, B4 exploits this knowledge for achieving reliable and resource-efficient data transmission in the highly dynamic vehicular scenario.

During the first two phases, the main task for the communication system was the resource-efficient transmission of vehicular sensor data as xfCD to crowdsensing-based services for traffic flow forecasting and optimisation. While those applications define soft deadlines for the age of the sensor data, they do not set strict requirements for the transmission delay, nor does packet loss have a critical impact on the system performance as fleets of vehicles provide massive amounts of data.

In contrast to that, the system requirements change completely for the third phase with the introduction of safety-related vehicle-to-vehicle communication. The overall goal is to provide guaranteed *Ultra-Reliable Low-Latency Vehicle-to-everything* (V2X) communication in a *massive MTC* context, where different kinds of automated systems and communication technologies coexist and compete among the available resources of the radio spectrum.

For addressing these novel challenges, context prediction methods will be integrated as an essential part into the design of all communication-related decision processes. The context is formed by the predicted mobility behaviour of the vehicle as well as the radio channel conditions at the respective locations and will be leveraged with the following approaches.

Knowledge about vehicle's trajectory will be taken into account to predictively avoid ineffective cellular handovers and optimise the resource allocation of the mobile devices by novel real-time scheduling strategies. Analogously, the packet forwarder selection for multi-hop vehicle-to-vehicle communication based on IEEE 802.11p and LTE Device-to-device (D2D) will be improved through consideration of mobility patterns within vehicle fleets.

In the highly-dynamic vehicular scenario, using 5G mmWave pencil beams that make use of extremely directed antennas, brings the dependency of the communication performance to the mobility to a whole new level. While this technology supplies very high data rates using frequencies in the higher GHz-range and system-immanently provides spatial separation, it is highly vulnerable to line of sight obstructions, especially in inner-city scenarios with a large amount of mobile

and static obstacles that cause shadowing and reflection. Therefore, we will exploit the vehicle's anticipated trajectory as well as information about the surrounding environment to **predictively steer the emitted beams** in order to minimise the distortion. In addition, we will derive novel channel models for using mmWave communication in a vehicular context.

Furthermore, we will continue the work on the **predictive choice of the transmission time** for non-time-critical data by leveraging favourable channel conditions for increased resource-efficiency.

It is anticipated that future cars will be equipped with multiple different communication technologies since no single standard is able to fulfil all the given requirements of upcoming vehicle-based applications. In the third phase of this project, the system-immanent heterogeneity will be exploited for choosing and combining network interfaces with respect to the achievable transmission efficiency and the application requirements to enable fast adaption to dynamic channel conditions, e.g., when transiting from motorway to inner-city traffic. Novel multi-interface communication schemes will be developed that are aware of the additional overhead introduced by the technology switch and make use of all the above-mentioned context-predictive approaches.

Based on the results of the second phase, the developed prediction methods have to be brought to the next level in order to guarantee reliable anticipatory communication. In the highly dynamic vehicular environment, the channel conditions are frequently changing due to the vehicle's mobility behaviour. Consequently, accurate estimations of the future spatio-temporal trajectories are required for predicting the radio channel conditions the vehicles will experience in the future. In addition to application-layer knowledge obtained from the vehicle's navigation system as used by the cross-layer mesh routing protocol B.A.T.Mobile in the second phase, traffic flow information will be integrated to optimise the speed-dependent on-lane mobility prediction. Since the individual cars are not able to measure the traffic flow at future locations, the required data is obtained either from cloud-services or through local data exchange with intelligent road infrastructure that captures and analyses traffic indicators.

The channel quality estimation itself will further be improved by integrating knowledge about the fading effects caused by static and mobile obstacles, such as surrounding buildings and other traffic participants, into the channel quality assessment. Furthermore, the data rate prediction scheme will be brought together with the *Client Based Control Channel Analysis for Connectivity Estimation (C3ACE)* of the A4 project to consider estimations of the resources consumed by other active cell users.

Within the B4 project, this work package plays a central role and allows the development of novel models inside the other work packages for optimising the vehicular traffic flow. While established mobility models for human drivers only consider the isolated behaviour mechanisms of individuals, the coordination between automated systems requires means of communication to become an essential part of the system and is therefore an important component of the hybrid mobility models that are developed in WP2. Similarly, the system-wide optimisation of the overall dwell time in WP4 for upcoming Intelligent Traffic Control Systems (ITCS) requires dynamic route adjustments for individual vehicles that need to be communicated.

The Communication Networks Institute runs an extensive 5G-ready laboratory environment that enables the Collaborative Research Centre (CRC) to validate its developed approaches and methods by means of accurate experimental setups. Existing 4G network and channel emulators have recently been extended with the latest cellular IoT (e. g., NB-IoT) and V2X (C-V2X) extensions. In particular, the application of a real-time-capable SDR and PXI-controller setup enables coupling of realistic C-V2X communication links to multi-scale simulation environments, all addressing highly scalable V2X communication. As a major 5G lab component, a fully operational 5G mmWave system operating at a frequency of 28 GHz and equipped with beamforming and pencil-beam antennas facilitates the CRC to validate the impact of beam-tracking performance for different antenna tapers, tracking algorithms, and mobility patterns (preliminary results were published at

WSA 2018). Moreover, an existing mobile 5G lab (modified transporter to mobile laboratory) enables mobile deployment of so-far stationary lab equipment for validation in realistic application area-dependent field trials within WP6.

Within this project, we aim to achieve the highest possible level of transparency and reproducibility by publishing the raw data of our experimental measurements as well as the developed measurement applications in an open source way. Similarly, novel simulation models will be made available inside the respective communities.

#### **Work package 4. Traffic control and dynamic routing of homogeneous and heterogeneous traffic**

The previous works on traffic control and dynamic routing in phases 1 and 2 considered closed systems for highways and inner cities respectively. In this work package, the results from the previous work phases will be combined to achieve and consider a road network that combines city and highway networks. Additional to vehicular traffic, inner-city networks also have pedestrian traffic. Because pedestrian traffic only interacts with vehicular traffic at intersections or other forms of road crossings, the pedestrian traffic can be idealized as a reduction of the road capacity for left or right turning lanes (where the pedestrians have green at the same time as the vehicle). For straight driving traffic, the pedestrian traffic can be idealized as a non-significant impact, because the pedestrians have red when the vehicle traffic has green. Furthermore, this work package aims to implement the model created in WP2 into this network, to consider automated vehicles and heterogeneous traffic. As mentioned in WP2, autonomous vehicles react passively to human driver behaviour; hence a significant reduction in traffic flow is to be expected, which – combined with a possible increase in the traffic volume because of now possible empty trips – would result in an increase of the global dwell time. To reduce these effects, traffic control and dynamic routing are important to prevent congestion. Currently applied methods for traffic control and traffic forecasting use real-time travel data from floating cars and induction loops. This results in a delayed routing for vehicles so that they are routed to avoid a jam only sometime after the jam is created and they are still routed to avoid the area of the jam even after this jam already resolved. The breakdown minimisation principle [267] introduces a critical value for traffic flow, after which traffic can congest. Based on this critical value, the travel time a driver needs to drive through a jam will be approximated. The travel time depends on the traffic flow directed through this jam. In this case the flow is greater than the critical value. Additionally, after the traffic flow decreases below a critical value, a prediction on the lifetime of a jam will be made. Then vehicles can be routed with real-time traffic data as well as these predictions on future traffic. This will further reduce the global dwell time as well as the lifetimes of jams.

Smart routing regimes that incorporate predictions on travel times may reduce dwell time. However, they may also cause novel unexpected jams at various locations throughout the traffic network. Learning optimal control of traffic, e.g., by routing or signalling, based on historical observations is difficult, as only the outcome of the actions taken can be monitored and not the result of all the alternative actions. This results in a trade-off between exploitation and exploration. Reinforcement learning, particularly bandit learning, tackles this closed-loop problem. The work on reinforcement learning for routing towards a reduction of the total dwell time will be continued and combined with the breakdown minimisation principle. By this envisioned combination of the breakdown minimisation principle and bandit learning, situation-aware, dynamic travel time predictions will be created. Incorporation of these predictions into routing will be beneficial; however, trip planning is more complex using dynamic costs, and speed-up heuristics are required to achieve scalable travel plan generation. Our work on efficient trip computations in dynamic settings using Dynamic Transfer Pattern [254, 269] will be continued towards changing transit graphs over time (planned collaboration with A6) and combined with the dynamic predictions from WP5. The required communication for propagating the resulting routing adjustments as well as the crowdsensing-based data acquisition will rely on the V2X-communication results of WP3.

The potential for the results of this work package to lead to a reduction in the global dwell time of vehicles, and thus in reduced traffic volume, will be validated in WP6.

**Work package 5. Data analysis and aggregation** In this work package, the traffic observations are used for imputation and prediction of unobserved values and for training bandit-models that address the exploitation/exploration trade-off in a closed-loop setting, which only reveals information on control actions taken but not on the alternatives. The resource constraints in traffic modelling and prediction hold for the centralized and the decentralized in situ analysis. In the centralized case, a high-performance cluster is applied for traffic prediction on a large scale, the probabilistic prediction models have to be tractable and need to exploit this computer architecture. In phase 2 the initially proposed Poisson Dependency Networks (phase 1) have been improved by Poisson Sum-Product networks, which enjoy a joint probability distribution and tractable inference. However, application to Markov processes could be eased with Conditional sum-product networks that allow for a more flexible factorisation of the distribution. In phase 3 we will study how this conditional sum-product network can be used to combine the rule-based cellular Nagel-Schreckenberg automaton model with data-driven graphical models. Another probabilistic approach for spatial imputation is the use of Gaussian Process Regression (GPR), also known as Kriging. But due to its high computational cost, it does not scale well. In cooperation with A1 we examine the use of distribution strategies (e.g., Gaussian summary trees or coresets) to scale GPR on street network size. In the distributed scenario, which uses vehicle-to-infrastructure communication, ubiquitous resource constrained devices (motes) monitor traffic states (potentially as part of smart city lights) and make real-time traffic predictions. Using low-power wide area networks (LPWAN), observations and predictions can be exchanged among sensor motes, but LPWAN networks as Sigfox bound bandwidth to 140 times 12 bytes per day. This exemplifies the resource constraints in a distributed setting, for which we will rely on the work of A4 that addresses massively scalable communication. In phase 2, we proposed an algorithm that uses learning from label proportions and reduces communication costs by aggregating label proportions in time intervals and communicating these ratios. In phase 3, we plan to extend this to a fully distributed scenario with 1) a real-time algorithm which runs on heterogeneous devices and 2) a local reinforcement learning component that situates within each automated vehicle and uses the predictions to reduce dwell time. In detail, the plans for the third phase comprise:

- **Conditional Sum-Product Networks** Probabilistic modelling of traffic values needs to respect its probability distribution. As traffic flow and density are count values, a Poisson or Cox distribution seems to be appropriate. In the second phase, use of PDNs and sum-product networks as a more tractable method to impute traffic were studied. We continue this work by combining cellular automaton models with these probabilistic approaches by conditional generative models. Conditional Sum-Product Networks will allow for tractable inference and a flexible factorisation of the joint distribution.
- **Bandit Models for routing and signalling** We exploit the use of reinforcement learning for self-organising routing and control (signalling) towards a reduction of dwell times. We combine this with the findings from WP4 on the breakdown minimisation principle.
- **Learning from Label Proportions for distributed prediction** In addition to a centralized setting with a global server, we focus on decentralized predictions that run on resource constrained embedded devices within the infrastructure. In this V2X setting that leverages results of WP3, communication cost for prediction needs to be considered. Instead of broadcasting every observed sensor reading throughout a mesh network, learning from Label Proportions provides a batch-wise communication of Label Proportions [B3/246] and will be developed further towards an online application in collaboration with B3. In WP2 we combine these findings on distributed prediction with simulated traffic data.

- **Secure aggregation and clustering** In a distributed setting, it is important to recognise time slices at which a prediction model can be applied. Horizontal clustering of the traffic values in a region allows identification of time intervals at which the prediction model can be applied. But centralization of the data potentially reveals individual mobility information. We will therefore develop a privacy-preserving aggregation and clustering method, based on homomorphic encryption.
- **Model selection** Various probabilistic models for traffic prediction have been developed in collaboration with the other projects of the CRC (A1, B3, A6), and others exist in the related literature (e.g., usage of Deep Convolutional Neural Networks [270], in A6, we study the use of geometric deep learning models for traffic prediction). However, the distribution of traffic is not stationary but drifts. This could be a slow shift, e.g., turning distribution at an intersection throughout a day or a sudden change caused by lane blockage, hazards, or accidents. In some of these circumstances, different models are beneficial. In collaboration with A3, we will study this and focus on the model selection task under concept drift.

**Work package 6. System validation by simulation and experimental analysis** In this work package, the methods and results of the previous work packages are aggregated into an overall system. For this purpose, we will create a simulation network that combines inner-city and highway networks. In this network, we will then simulate the model created in WP2. It is to be expected that the transition from human-driven vehicles to fully automated ones will take decades and the degree of automation in traffic will increase steadily. For this reason, we will test the results of WP4 and WP5 with different degrees of automation ranging from 0 to 100%. The results of these simulations will then be used as a feedback for the respective work packages, enabling the validation and refinement of the developed models.

In phase 2, locations that are often overloaded could be identified by analysing the average number of breakdowns. Afterwards, dynamic lanes that switched directions based on the time of the day at these locations were tested. It was found that additional lanes shift the location of congestion but do not lead to a significant reduction. To consider the limited space of cities, options that require no additional space like dynamic traffic lights will be looked into in phase 3. The time of the green and red phase at dynamic traffic lights can be changed in real time and adjusted to changes in the traffic flow. Together with a connection of multiple traffic lights, traffic flow can be increased by reducing the average number of times a vehicle has to stop at traffic lights.

In order to be able to adapt the results of the simulations into real traffic scenarios, the simulations will be based on the traffic information gathered in WP1 for increased realism.

While vehicular motion can easily be evaluated in large-scale scenarios by means of simulation, this does not apply for the required novel communication systems (e.g., mmWave), for which simulation models are not widely available yet. Two approaches will be used to overcome this issue. On the one hand, simulation models based on the measured characteristics of real hardware devices that were obtained by experiments in the laboratory and in the field will be created. On the other hand, real-world devices will be directly integrated into the simulation runs using hardware-in-the-loop setups.

## Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Data acquisition and processing																	
2. Modeling of heterogeneous vehicular traffic																	
3. Context-predictive Ultra Reliable Low Latency V2X																	
4. Traffic control and dynamic routing of homogeneous and heterogeneous traffic																	
5. Data analysis and aggregation																	
6. System validation by simulation and experimental analysis																	

## 3.5 Role within the Collaborative Research Centre

Project B4 offers many points of contact to the other projects and benefits greatly from the interdisciplinary character of the CRC.

Together with A1, we will extend the work on resource-efficient sensor data acquisition and use the multidimensional time-series data provided by the Wireless Detection and Warning System to classify vehicles based on their characteristical radio fingerprints. In order to enable future energy-autonomous system deployments, the achievable classification accuracy will be evaluated with respect to the online resource efficiency of the applied machine learning model. In WP5, we continue the cooperation with A1 on scalable Gaussian Process Regression and graphical Spatio-Temporal prediction models (STRF).

With A3 we exploit usage of the MBO method for model selection under concept drift, which naturally occurs in traffic distribution.

For the upcoming third funding phase, we will continue exploiting the strong synergies with A4 which resulted in numerous joint publications in the first two phases. A4 provides the evaluation platform as well as the required resource and cost models for the communication medium that are leveraged within B4 to predictively schedule vehicular data transmissions with respect to the achievable resource efficiency. While providing reliable communication in highly mobile network topologies is one of the focus fields of B4, the applicability of the novel methods developed will be evaluated for Unmanned Aerial Vehicle networks in the logistical environments of A4. Furthermore, the methods for assessing the energy-efficiency of communication mechanisms of A4 will be directly applied for evaluating the novel transmission methods that are developed within B4.

Probabilistic traffic modeling naturally uses graphs to store traffic networks and observations. Together with A6, we study the effects of changing graphs over time for routing problems. This comprises transit schedules that change from day to day and make precomputation of Transfer

Patterns infeasible (compare WP4) and shortest-path computations with dynamically changing costs (derived from the prediction in WP5); the PH-graphs [255] provide a base for time-dependent routing exploiting correlations among the edges. We continue the collaboration with A6 on traffic prediction and study use of geometric deep learning techniques (e.g., Graph Convolutional Neural Networks) for probabilistic traffic prediction.

The communication-efficient distributed prediction with label proportions will be developed further with B3, as it is the basis for our approach towards a distributed situation-aware routing.

The cooperation with project C4 led to a publication at AAAI 2018 on using coresets in sum-product networks. This is an important step towards scalable prediction for traffic scenarios. We intend to continue the cooperation with project C4 and Kristian Kersting (TU Darmstadt) on coresets for deep graphical models from which the traffic prediction mechanisms can benefit.

### 3.6 Differentiation from other funded projects

#### **DFG-Forschergruppe 1511**

**(Wietfeld, Reference number Wi 3751/1-1, Wi 3751/2-1) (Funding period: 2014–2018)**

The project addresses innovative wide-area control applications and protection of electrical energy systems, in particular, to avoid large-scale system failures (blackouts). This includes an evaluation and optimisation of the hybrid simulation environment for power grids and communication networks in conjunction with Software-Defined Networking (SDN).

#### **AutoMat**

**(Wietfeld, Reference number H2020 644657) (Funding period: 2015–2018)**

The core intention of AutoMat is to establish a novel and open ecosystem in the form of a cross-border Vehicle Big Data Marketplace that leverages currently unused information gathered by a large number of vehicles from various brands to reduce the development costs for vehicle data services. In contrast to basic research in project B4, AutoMat is an innovation action that transfers existing methods into market-ready solutions.

#### **BERCOM**

**(Wietfeld, Reference number BMBF 13N13741) (Funding period: 2015–2018)**

BERCOM focuses on hardening and extending LTE for use in shared critical infrastructure communication networks, e.g., for Smart Grids. It includes a validation of the overall system's increased robustness and performance via a physical demonstrator.

#### **OPUS**

**(Wietfeld, Reference number EFRE-0800885) (Funding period: 2017–2020)**

OPUS (Optimised Predictive Performance Using Cyber-Physical Systems) intends the development of methods and technologies that enable predictive maintenance for future generations of permanently installed cyber-physical system (CPS), such as pumps or engines. Since such installations are widely distributed and frequently located at dead spots, e.g., in basements, a major challenge is to provide methods for reliable wide-area wireless communication in such applications.

#### **LARUS**

**(Wietfeld, Reference number BMBF 13N14133) (Funding period: 2017–2019)**

LARUS works on the development of an unmanned aerial support system for maritime search and rescue missions. The focus areas are robust long-range communication, aerial base stations, radio-based localisation, and communication-aware mission control.

**InVerSiV****(Wietfeld, Reference number EFRE-0800422) (Funding period: 2016–2019)**

This project aims to aggregate dynamic local live maps using sensorfusion of sensors within cars and roadside units to enable (semi-)autonomous driving in the challenging areas of megacities. Data from image recognition, radar, and radio-based sensors are transmitted over LTE and LTE V2X to a local edge cloud for the sensorfusion and map broadcasting. In contrast, the B4 project focuses on traffic flow prediction, hybrid vehicular traffic, and efficient communication for the involved data exchange.

**CPS.HUB/NRW****(Wietfeld, Reference number EFRE-0400008) (Funding period: 2015–2018)**

CPS.HUB/NRW concentrates the competence and knowledge of all disciplines relevant to the development of cyber-physical systems. The innovation ecosystem provided by CPS.HUB/NRW enables regional actors with broad CPS-relevant knowledge to participate and continuously refine their processes and adapt to new developments.

**MEC-View****(Schreckenberg, Reference number 19A16010H) (Funding period: 2017–2019)**

This project, funded by the "Bundesministerium für Wirtschaft und Energie" (BMWI), deals with the application of high-accuracy road maps and the necessary infrastructure of sensor systems to update them with the current traffic flow, providing additional information for automated vehicle: e.g., important changes like a slow driving vehicle, that functions as a moving bottleneck, temporarily reducing the local traffic capacity. The focus here lies on the gathering of real time data, while the analysis of real-time dynamic routing or changes in the network would play no role.

**VAVEL****(Liebig, Reference number EU-H2020 Grant Agreement No 688128) (Funding period: 2016–2018)**

The project focuses on proactive control, prediction, and event detection using heterogeneous urban sensor streams. Dynamic travel times are estimated and incorporated into multimodal trip planning using Dynamic Transfer Pattern.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Staff								
Doctoral researchers, 100 %	3	193,500	3	193,500	3	193,500	3	193,500
Total	—	193,500	—	193,500	—	193,500	—	193,500
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
<b>Grand total</b>	<b>193,500</b>		<b>193,500</b>		<b>193,500</b>		<b>193,500</b>	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Michael Schreckenberg, Prof. Dr., professor	Traffic modelling	Duisburg-Essen University	10	—	Existing funds
	2	Christian Wietfeld, Prof. Dr., professor	Communication networks	TU Dortmund	4	—	Existing funds
	3	Thomas Liebig, Dr., junior research group leader	Data mining	TU Dortmund	19	—	Existing funds
	4	Merlin Becker, M.Sc., doctoral researcher	Traffic modelling	Duisburg-Essen University	19.9	—	Existing funds
	5	Karsten Heimann, M.Sc., doctoral researcher	Communication networks	TU Dortmund	19.9	—	Existing funds
	6	N.N., student assistant	Traffic modelling	Duisburg-Essen University	10	—	Existing funds
	7	N.N., student assistant	Communication networks	TU Dortmund	10	—	Existing funds
Non-research staff	8	Matthias Foese, technical staff	—	TU Dortmund	6	—	Existing funds
<b>Requested staff</b>							
Research staff	9	N.N., doctoral researcher	Data mining	TU Dortmund	—	Doctoral researcher	—
	10	Benjamin Sliwa, M.Sc., doctoral researcher	Communication networks	TU Dortmund	—	Doctoral researcher	—
	11	Tim Vranken, M.Sc., doctoral researcher	Traffic modelling	Duisburg-Essen University	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

**1. Schreckenberg, Michael**

Project management. Focus on the development of a hybrid traffic model and its application. Cooperation in the WPs 1, 2, 4, and 6.

**2. Wietfeld, Christian**

Project management. Focus on the further development of context-predictive communication mechanisms. Cooperation in the WPs 1, 2, 3 and 6.

**3. Liebig, Thomas**

Project management. Focus on spatio-temporal prediction models. Cooperation in the WPs 1, 4, 5, and 6.

**4. Becker, Merlin**

Focus on the development of a hybrid traffic model. Cooperation in the WPs 1, 2 and 6.

**5. Heimann, Karsten**

Integration of mmWave communication into vehicular scenarios.

**6. N.N.**

Support of implementation within WPs 1, 2, 4 and 6.

**7. N.N.**

Support of implementation within WPs 1, 2, 3 and 6.

**8. Foese, Matthias**

Support with regard to the technical infrastructure.

**Job descriptions of staff for the proposed funding period (requested funds):**

**9. N.N.**

Research and development of data analysis methods for spatio-temporal traffic prediction in work packages 1, 4, 5, and 6

**10. Sliwa, Benjamin**

Research and development of novel context-predictive communication approaches within the WPs 1, 2, 3 and 6.

**11. Vranken, Tim**

Focus on the development of a hybrid traffic model and its application. Cooperation in the WPs 1, 2, 4, and 6.

### 3.7.4 Requested funding for direct costs for the new funding period

	2019	2020	2021	2022
TU Dortmund: existing funds from University	4,000	4,000	4,000	4,000
Duisburg-Essen University: existing funds from University	4,000	4,000	4,000	4,000
Sum of existing funds	8,000	8,000	8,000	8,000
Sum of requested funds	0	0	0	0

(All figures in euros)

### 3.7.5 Requested funding for instrumentation for the new funding period

This project does not request any funding for major research instrumentation.

### 3.1 General information about Project C1

### 3.1.1 Project title:

## Feature selection in high dimensional data for risk prognosis in oncology

### 3.1.2 Research area(s):

201-07 (Bioinformatics), 205-14 (Oncology)

### 3.1.3 Principal investigator(s)

Rahmann, Sven, Prof. Dr., 01.08.1974, German

Bioinformatik, Informatik XI, Technische Universität Dortmund;  
Genominformatik, Institut für Humangenetik, Universitätsklinikum  
Essen, Universität Duisburg-Essen  
Hufelandstr. 55  
45147 Essen

Phone: 0231-755-7713  
E-mail: Sven.Rahmann@tu-dortmund.de

Schramm, Alexander, Prof. Dr., 21.09.1968, German

Molekulare Onkologie, Innere Klinik/Tumorforschung, Universitätsklinikum Essen, Universität Duisburg-Essen  
Hufelandstr. 55  
45147 Essen

Phone: 0201-723-1630  
E-mail: Alexander.Schramm@uni-due.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material. A copy of the required approval of the responsible ethics committee is included with the proposal.	(x) yes <input type="checkbox"/> no (x) yes <input type="checkbox"/> no
2.	clinical trials	<input type="checkbox"/> yes (x) no
3.	experiments involving vertebrates.	<input type="checkbox"/> yes (x) no
4.	experiments involving recombinant DNA.	(x) yes <input type="checkbox"/> no
5.	research involving human embryonic stem cells.	<input type="checkbox"/> yes (x) no
6.	research concerning the Convention on Biological Diversity.	<input type="checkbox"/> yes (x) no

## 3.2 Summary

Recent advances in molecular biotechnologies have fundamentally changed how cancer patients are diagnosed and treated. The development of targeted therapies has increased patients' life expectancy and quality of life with the majority of cancer types. However, predicting treatment efficacy and selecting the optimal personalised therapy for each patient remains a challenge for clinicians. Mainly, the development of resistance to therapy and intratumoural heterogeneity limit successful long-term remissions and cures. The early prediction of therapy resistance or relapse is thus deemed crucial for further improving therapy outcome. Identification of features termed *biomarkers*, which are derived from patient samples by high-throughput analyses, is an important means to achieve this goal. Project C1 builds and optimises models for clinically relevant decisions in oncology by selecting features from high-dimensional feature spaces, extracted from raw data created on different molecular platforms.

In the past, highly parallel ("next generation") DNA sequencing technology allowed researchers with access to specialised sequencing core facilities to discover tumour-specific mutations. As DNA sequencing capacity continues to increase and costs continue to drop faster than computational capacity and storage can keep up, new algorithmic paradigms are needed for the analysis of very large genomic data sets. In project C1, we investigate new algorithms to extract relevant features for biomarker discovery from whole-genome data sets in the 10–100 terabyte range on commodity hardware by streaming the sequence data and filtering for features of interest using novel string hashing methods.

Recent developments in *nano*pore sequencing are democratising DNA sequencing and genome analysis. The new nanopore sequencers are of size comparable to a USB stick, are inexpensive, and can be used without specialised lab equipment. While nanopore sequencing offers lower throughput and higher error rates than established technologies at the moment, it has the potential to turn DNA sequencing and subsequently genomic analysis into a commodity. In oncology, the vision is that nanopore sequencing, together with non-invasive patient monitoring techniques such as "liquid biopsies" drawn from blood or urine, will allow for detection of small amounts of circulating tumour DNAs, allowing an accurate assessment of patient risk and therapy options. In principle, such an assessment would be possible anywhere at any time, given standard equipment of moderate costs, i.e., the sequencer and a laptop or embedded system.

For this vision to become a reality, several data analysis challenges must be overcome: In addition to the constraints imposed by small sample size  $n$  compared to the high dimensionality  $p$  of the feature space ( $n \ll p$  problem), the cyber-physical systems for nanopore sequencing create novel resource

constraints: The raw data generated by this new technology is a large-volume high-frequency signal of ion currents, which is difficult to translate directly into a DNA sequence. Therefore, to identify tumour “fingerprints” or biomarkers based on tracing tumour-derived nucleic acids, either better methods for DNA base calling from ion currents are needed, or a different representation of the tumour fingerprints has to be considered, such as features in signal space. We will follow both avenues in parallel and in particular consider novel features derived from a discretised compressed ion current signal space.

### 3.3 Project progress to date

#### 3.3.1 Report and current state of research

Previously, feature selection from high-dimensional data mainly focused on gene expression data using microarrays (Phase 1) and on identification of genetic mutations and transcriptional profiles from high-throughput DNA and RNA sequencing (Phase 2), using neuroblastoma tumour samples as applications. Mapping and alignment of sequence data (10 to 50 GB of raw data per sample) proved to be very demanding in terms of computation time and memory requirements. Therefore we identified and established resource-efficient methods for sequence analysis and feature extraction. To model and discover relations between features of different types, e.g., mutations, structural genome changes, gene expression changes, and methylation changes between tumour cells and normal controls or primary and relapse tumours, we developed complexity reduction methods based on probabilistic graphical models. Targeted analyses of paired samples from primary and relapse tumours allowed us to extract patterns of tumour evolution that indicate cancer progression or the escape of tumours from therapeutic intervention. For the model tumour neuroblastoma, this resulted in unprecedented insights into the genomic plasticity of a tumour under attack by therapy [C1/321]. Project leader Sangkyun Lee left Germany early in 2017 when he accepted a professorship in South Korea, which led to changes in parts of WP 6, as discussed below.

**Provision and generation of data (WP1):** We obtained access to 200 exomes from neuroblastoma, a solid tumour of childhood, and corresponding normal tissue as well as to 30 whole genomes sequenced with low coverage. These sequencing efforts were separately financed by the BMBF and the German Cancer Aid. Additionally, we obtained access to a data set comprising RNA sequencing data and mRNA microarray profiling data from 498 patients, which were provided as part of the SEQC consortium [318]. To pave the way for analysing neuroblastoma data in comparison to other cancer types from The Cancer Genome Atlas (TCGA), an R package “EasyTCGA” was developed that facilitates batch downloading of TCGA level-3 data.

With contributions from A. Schramm, colleagues from Cologne and Heidelberg identified a novel mechanism of aggressive tumour growth involving activation of the enzyme telomerase by genetic rearrangements [311].

We also confirmed that tumours can become aggressive by different strategies, and so we decided to focus on the genetic dynamics of neuroblastoma by comparing paired samples at primary diagnosis and at relapse with corresponding normal controls. Additionally, samples from different tumours within the same patient and from different time points after treatment were available for analyses. We used the workflow management system Snakemake (developed in phase 1 of this CRC, see also WP3) to write and at the same time reproducibly document a data analysis process for matched neuroblastoma triples (normal controls, primary tumours, and relapse samples of the same patients). This workflow consisted of quality control steps, resource-efficient read mapping and alignment, and the subsequent feature extraction steps for whole exome sequencing (WES) data sets, gene expression data from RNA sequencing (RNA-seq), and whole genome methylation

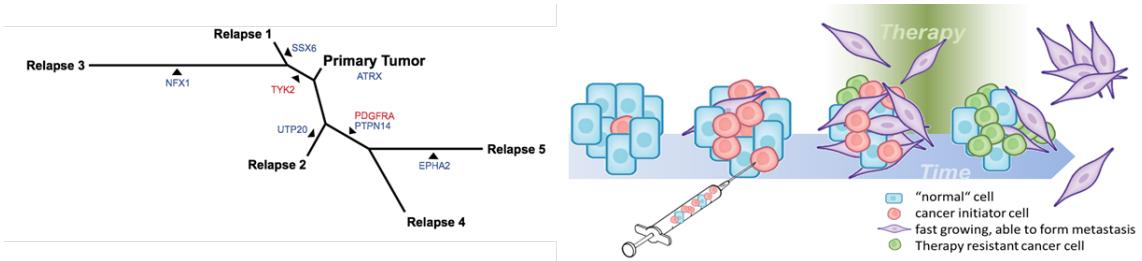


Figure 3.1: Left: Evolutionary tree derived from binary features (presence/absence of informative single nucleotide variants) of a primary neuroblastoma tumour and several relapse samples taken from different tissues and at different times from the same patient [C1/321]. Right: Illustration of evolution of tumour heterogeneity under therapy over time [C1/322].

arrays (Illumina 450k arrays). We thus demonstrated for the first time that neuroblastoma tumours acquire significantly more mutations during the disease course and that intratumoural heterogeneity is a feature contributing to genetic plasticity of the tumour [C1/321]. This in turn is considered as a hallmark of therapy resistance, preventing an even greater success of contemporary treatment protocols in oncology.

**Resource-efficient high-throughput sequence data analysis (WP2):** The basic tasks of *read mapping* and *alignment* of sequenced DNA fragments (“reads”) against the reference genome require significant computational resources. *Read mapping* consists of finding a read’s origin on a reference genome, using error-tolerant pattern matching based on index data structures. *Read alignment* then produces a basepair-by-basepair comparison of the read and the corresponding reference segment to highlight differences. These steps are independent of the type of downstream feature extraction.

At the end of the first phase (after the review for the second phase), we published a GPGPU-based read mapper and aligner called “PEANUT”. The main methodological contribution of this work is a GPU-memory-friendly index data structure for locating  $q$ -grams (DNA strings of given length  $q$ ) in the genome, the “ $q$ -group index”. The  $q$ -group index better supports data parallelism than a classical  $q$ -gram index, and it has been included in NVIDIA’s NVBio library of reusable components designed to accelerate bioinformatics applications using CUDA. Because of the new index, PEANUT is resource-efficient in the sense that its GPU occupancy reaches values up to 1.0 during key steps. The remaining bottleneck was identified as file input/output.

Recently, read mapping methods based on min-hashing of  $k$ -mers have emerged, e.g., minimap2 [306]. Here, each window of a read and the reference genome corresponds to a set of  $k$ -mers, and each such set is represented by (an encoding or a hash value of) its minimising  $k$ -mer with respect to a randomised ordering of  $k$ -mers. Then the probability that the hash values of two  $k$ -mers sets  $A, B$  are equal is their Jaccard similarity coefficient  $|A \cap B|/|A \cup B|$ . Using multiple (randomised) orderings and/or hash functions, near-perfect discrimination between true read origins and random matches can be achieved. With respect to the tolerance against genetic variants, an important design question is whether the variant  $k$ -mers should be added to the index (higher space requirements, larger set union) or not (smaller set intersection if the variant occurs). Adding a variant is therefore beneficial on average if its frequency in the examined population is sufficiently high. We presented a detailed analysis with exact thresholds [C1/319], which received WABI’s Best Paper Award. In addition, a student project group (a practical course over two semesters) developed a min-hashing based read mapper, which was further refined in the master’s thesis of Jens Quedenfeldt.

**Feature extraction from sequence data (WP3):** In a large-scale neuroblastoma study with several external collaborators, in which two of the project leaders participated as first and (equally contributing) last authors, we traced neuroblastoma tumour development from normal cells to primary tumour to relapse after therapy in 16 patients. We have discovered features that characterise this evolution [C1/321] (Figure 3.1).

For tumour risk prediction and treatment decisions, the following molecular features derived from genomic data are of primary interest: (1) single nucleotide mutations (or variants, SNVs), (2) short insertions or deletions of DNA, (3) large structural variants (e.g., chromosomal translocations), (4) copy number changes (gain or loss of genetic material in tumour cells), (5) epigenetic changes such as DNA methylation changes, and (6) differences in gene expression.

We have developed feature extraction workflows and data analysis processes for each type of feature mentioned above. The basis of each of these processes is a workflow management system called Snakemake [305], which was developed by Johannes Köster (a former PhD student in project TB1 in this CRC and now a postdoc in project C1) and Sven Rahmann during the first phase of this CRC and is now widely used worldwide, especially for large-scale DNA sequence analysis workflows. Somatic mutations are discovered and extracted using the APE/EAGLE system developed internally by Christopher Schröder and used in several clinical projects in addition to our study [C1/321]. The overall variant set is naturally encoded by a binary vector, indicating the presence or absence of each mutation in each sample. Thereby, we obtain an  $(n \times p)$ -dimensional binary matrix, where  $n$  is the number of patients and  $p \approx 3.6 \cdot 10^6 \gg n$  is the number of occurring variants. From this set, we have identified a small number of variants as recurring and characteristic for tumour progression.

We initially hypothesised that there might be relevant changes in DNA methylation and hence asked if there were differentially methylated DNA regions between primary and relapse tumours and if they could be explained by corresponding single nucleotide mutations in spatial proximity. For this, it was necessary to classify a continuum of observed methylation levels of a single CpG dinucleotide or a region, from 0.0 (completely unmethylated) to 1.0 (fully methylated) into three categories (named unmethylated, semi-methylated, fully methylated). The thresholds must be set dynamically based on the observed histogram because different methods (e.g., methylation arrays vs. bisulfite sequencing) introduce different biases. We have proposed to use a modified expectation maximisation (EM) algorithm on mixtures of beta distributions, where the maximum likelihood step is replaced by moment-based estimation to avoid the singularities of beta densities at their extreme points and proved fundamental invariance properties of this method [C1/320].

**Identification of sparse feature graphs (WP4):** The identification of dependencies between the expression levels of different genes is an important step towards reducing data complexity and increasing interpretability when analysing tumour progression. Assuming that the continuous feature vectors, representing gene expressions, are samples of a  $p$ -dimensional Gaussian distribution, dependencies of interest are given by the inverse covariance matrix  $\Theta := \Sigma^{-1}$ . This matrix is most likely sparse, since its non-diagonal entries  $\theta_{ij}$  ( $i \neq j$ ) are only nonzero if the corresponding feature variables are conditionally dependent given all other variables.

A state of the art method to estimate a sparse inverse covariance matrix is the graphical LASSO [301], solving the convex optimisation problem

$$\min_{\Theta \in \mathbb{R}^{p \times p}} -\log \det \Theta + \text{tr}(S\Theta) + \lambda \|\Theta\|_1,$$

where  $S$  is the sample covariance matrix and the last  $\ell_1$ -norm term, defined element-wise, induces sparse solutions. The regularisation parameter  $\lambda > 0$  is either estimable by cross validation or by controlling the Family Wise Error Rate (FWER). The first approach results in data- and accuracy-dependent sparsity, and the second approach returns models that are too sparse because of the extreme high feature dimension, so important dependency relations may remain undiscovered.

This is why we have considered an alternative regularisation framework called the Sorted L-One norm Penalisation Estimation (SLOPE) [296]. Here, the regularisation term of the graphical LASSO is substituted by  $J_\lambda(\theta) := \sum_{i=1}^p \lambda_i |\theta|_{(i)}$ , and  $|\theta|_{(i)}$  refers to the  $i$ -th largest absolute value of the vector  $\theta$ . This enables controlling the less restrictive False Discovery Rate (FDR) by a suitable estimation of the parameter vector  $\lambda$ . A successful collaboration with the original authors of SLOPE resulted in the design of an efficient saddle point optimisation algorithm for the Dantzig selector form of SLOPE [C1/324]. In a master's thesis, the application of SLOPE has been examined for the analysis of dependencies among gene expressions.

**Tumour subtype identification by probabilistic graphical models (WP5):** Given a (binary)  $n \times p$  matrix  $D$  of genetic variants (cf. WP3), multiple methods are applicable to extract latent groups of variants, which may represent an independent tumour-shaping process or a tumour subtype.

Boolean Matrix Factorisation (BMF) [309] is a method for this task. BMF uses Boolean algebra for the computation of the matrix product; it aims to find binary matrices  $X \in \{0, 1\}^{p \times k}$  and  $Y \in \{0, 1\}^{n \times k}$  with  $k \ll \min\{n, p\}$  such that

$$D \approx Y \odot X^\top = \left( \bigvee_{l=1}^k Y_{jl} X_{il} \right)_{1 \leq j \leq n, 1 \leq i \leq p},$$

where  $a \vee b$  denotes the logical **or** of truth-values  $a$  and  $b$ . Each outer product  $Y_{\cdot l} \odot X_{\cdot l}^\top$  indicates a group of variants, indicated by  $X_{\cdot l}$ , which occur often together in a subset of the patients, indicated by  $Y_{\cdot l}$ . The usage of Boolean algebra allows for overlap between the outer products, opposed to binary matrix factorisation, where the data is approximated by the product of binary matrices in standard algebra. In contrast to other methods, such as LDA (Latent Dirichlet Allocation) [295], the automatic determination of the BMF rank  $k$  (corresponding to the number of topics in LDA) is better understood, and the minimum description length principle has proven to be effective for model order selection [308].

The drawback of methods for BMF had been the reliance on multiple heuristics, resulting in greedy algorithms that do not scale well and return unstable results. We have been able to overcome these drawbacks by using a novel numerical optimisation technique for the minimisation of a smooth, non-convex relaxation of the typically noncontinuous description length. The novelty of our approach is the application of a regulariser that penalises non-binary values in the  $\ell_1$ -norm and ensures that the factor matrix values are bounded between zero and one. We have derived the proximal mapping of this penalising function and could henceforth employ the Proximal Alternating Linearised Minimisation (PALM) [3], which guarantees the convergence to a critical point of our relaxed objective. We developed the BMF optimisation framework PAL-TILING [A1/C1/32], whose highly parallel implementation for GPUs has been made publicly available (PRIMP: sfb876.tu-dortmund.de/primp). An experimental analysis has shown that PAL-TILING returns robust factorisations that are not sensitive to the initialisation and noise.

We have also explored FDR control in the unsupervised setting of BMF [A1/C1/33]. Assuming that the data is constituted of a Boolean matrix factorisation and Bernoulli-distributed noise, we have proven two bounds on the probability that an outer product covers mostly noise. Each bound exploits a specific property of a factorisation approximating the data, and a theoretical analysis has shown that both bounds are not trivial. In summary, we have established a new method to determine the rank in matrix factorisation under quality guarantees.

**Modelling the sequence of events in tumour development (WP6):** Applying LDA and BMF, as described above, to the available curated data set from [C1/321] was surprisingly challenging. No interpretable variant groups were found when applying these methods to the tumour progression

stages primary tumour (T), relapse (R), and normal control (N) separately; and only patient-specific variant groups were discovered when the methods were applied to the joint data set. Comparing the BMF results for each progression stage suggested a complex structure where parts of the separately obtained (variant groups) occur for every progression stage N/T/R, and additional variants show the differences for a subgroup of patients. However, state of the art methods that assume that the variation groups are either common or discriminative among the classes (i.e., the progression stages) [302, 307, 304] and are not suited to find such a structure. Non-negative matrix factorisation is known to favour near-orthogonal solutions [299].

This motivated a new approach for the derivation of similar and discriminating patterns. For each progression stage  $C \in \{N, T, R\}$ , we attempt to factor the stage-specific observed binary matrix  $D^C$  into

$$D^C \approx Y^C \odot (X + V^C)^T$$

with stage-specific  $Y^C \in \{0, 1\}^{n \times k}$  and  $V^C \in \{0, 1\}^{p \times k}$ , but a common  $X \in \{0, 1\}^{p \times k}$ . The proposed method is called C-SALT [C1/20] and also implemented for GPUs (sfb876.tu-dortmund.de/csalt). We have been able to detect 16 variant groups together with their class-specific variants.

**Biological validation of the relevance of identified features in tumour genomes (WP7):** Previously, regulation of cell cycle control was identified as an Achilles heel of cancer cells, also in neuroblastoma. Our subsequent experiments validated that pharmacological inhibition of the identified genes, including cyclin-dependent kinases (CDKs), represent bona fide strategies for the treatment of neuroblastoma [C1/323]. As these findings have been validated by others, clinical trials have been initiated to evaluate CDK inhibitors as personalised treatment options in neuroblastoma (<https://clinicaltrials.gov/ct2/show/NCT02780128>, accessed 2018-05-28). However, technical limitations in generating gene expression data, not only at our site, but worldwide, have prevented the application of gene expression-based classifiers in clinical practice so far.

Based on the analyses of the mutational dynamics in relapsing tumours and the accompanying results from the graphical modelling of the changes in tumour genomes, we initiated further studies and re-analysed the data from [C1/321]. This revealed novel potential roles for non-coding DNA and the mitochondrial part of the cancer genome. We further validated the biological functions of identified genes in cellular models of neuroblastoma that were available in the lab of A. Schramm. We set up the CRISPR/Cas9 toolbox for genetic editing of tumour cells. CRISPR/Cas9 has been identified as a breakthrough technology for genetic manipulation in the past years, as is now considered as a possible route to precisely edit genomes for the correction of genetic diseases. For our purposes, we established CRISPR/Cas9-mediated overexpression, downregulation or knockout of individual tumour-associated genes. We thereby analysed the role of the potential key genes on proliferation, clonal outgrowth, migratory capacity, invasiveness and drug resistance. A major focus was set on a gene designated as PRKCI, which codes for the protein kinase C iota. This gene was identified to be differentially expressed between relapsing tumours and primary tumours. Using CRISPR/Cas9, tumour cells without or with enhanced levels of PRKCI were created and validated. A complex set of biological experiments with different read-outs provided evidence that PRKCI is involved in significant changes of tumour cells' ability to migrate and to grow without being anchored and connected to other cells. These features are typically seen only in the most aggressive, metastasising cancers, and we could assign a functional role for PRKCI in these processes.

PRKCI upregulation is not restricted to neuroblastoma, but can also be found in lung tumours, in which the majority of small-cell cancers harbour multiple copies of the PRKCI gene, which is also accompanied by a poor overall survival of affected patients [303]. Thus, data from us and others have identified upregulation of PRKCI as a common mechanism in tumour progression associated with fatal outcome of patients independent of the tumour type.

**Current State of Research:** So far, the main resource constraints for this project have been a limited number  $n$  of samples versus an extremely high number  $p$  of potential features (each

	phase 1 exon arrays	phase 2		phase 3	
		WES	mRNA-seq	WGS	nanopore
#probes	200 000	50 000 000	25 000 000	3 000 000 000	3 000 000 000
coverage	N/A	100x	200x	30x	0.33x
read length	N/A	100	100	100	10 000
#reads	N/A	50 000 000	50 000 000	1 000 000 000	100 000 000
raw size [GB]	0.1	5.0	5.0	100.0	500.0
#features	20 000	25 000	25 000	2 000 000	to be determined
error rate	1%	0.1%	0.1%	0.1%	10%

Table 3.1: Typical order-of-magnitude statistics of data sets produced by different technologies used in the different phases of this CRC. Numbers are per sample. (WES: whole exome sequencing; the exome consists of approx. 2% of the genome; WGS: whole genome sequencing; #probes: number of queried locations in the human genome; coverage: number of times each queried location is queried on average; raw size: approximate file size in gigabytes; #features: average number of features extractable from one sample).

observed genetic variant; expression changes of each transcript) and the high volume of the raw sequence data. These challenges remain and become more pronounced as high-coverage whole-genome sequencing (WGS) becomes more commonplace. Table 3.1 gives an overview of data set statistics from different technologies used in project C1 over time. In contrast to whole-exome sequencing, which only interrogates 2% to 3% of the genome at high coverage, with WGS one can discover variants in the entire genome at a reasonable average coverage of 30 reads per position. Such data sets are produced and often processed in core sequencing facilities (PI Sven Rahmann is a bioinformatics expert of the recently established DFG-funded West German Genome Center).

In 2016, bioinformatic sequence analysis experienced a fundamental shift when it became clear that for many questions, precise alignments are not required. For example, to estimate gene expression in an RNA-seq sample, it is sufficient to identify to which gene each read belongs and count the number of reads mapping to each gene. Often but not always, a single  $k$ -mer (string of length  $k$ ) of the read suffices to uniquely assign the originating gene (typically  $k = 23$  for the human genome). Such an approach is implemented, for example, in kallisto [297]. The main advantage of  $k$ -mer methods is their much lower resource usage in terms of runtime, memory, I/O bandwidth, and hence energy in large computing centres. Developing  $k$ -mer based methods for extracting different types of features from genome-wide data sets has become a very active area of research, to which we have already contributed [C1/319], as we will continue to do in the next phase.

On the other hand, clinicians prefer that *diagnosis, risk prediction and therapy decisions* are based on *a few selected biomarkers*. Typically, these are discovered in exploratory genome-wide studies, validated in larger cohorts and then, for diagnosis, checked with *targeted sequencing*. The latest nanopore sequencers (e.g., the MinION from Oxford Nanopore; see figure 3.2, left side) for this task are small cyber-physical systems (slightly larger than a USB stick) that turn DNA sequencing into a commodity that will become ubiquitously available. Physically, a single DNA molecule is threaded through a nanopore. Changes in the resulting ion current are recorded in real-time on a connected laptop as the DNA fragment moves through the pore (figure 3.2 right). Ideally, real-time algorithms would convert the raw signal immediately into the corresponding DNA sequence. While the previous high-volume and high-throughput sequencing methodologies relied on specialised chemistry that required laboratory infrastructure and non-mobile devices and workstations, the developments in nanopore sequencing allow for analyses of single RNA and DNA molecules using, in principle, mobile devices: The MinION system has been used to track Ebola outbreaks in the field [312]. Although mobile sequencers are anticipated to work virtually anywhere in the near future, current challenges include data acquisition, real-time analyses, and conversion of raw measured ion currents at the nanopore into interpretable DNA/RNA sequence data.

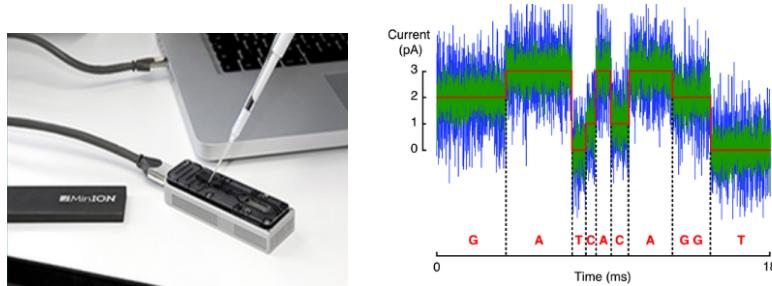


Figure 3.2: Left: MinION device from Oxford Nanopore. Right: Illustration of DNA base calling from an ion current signal: the DNA sequence is inferred from smoothed piecewise constant signal regions. Source: Blog “Omic frontiers: Tales from the Exeter Sequencing Service” by Konrad Paszkiewicz, <https://konradpaszkiewicz.wordpress.com/2014/04/10/nanopore/>, accessed 2018-05-28.

Presently, the interpretation of the raw ion current data is both computationally demanding and an active area of research. state of the art algorithms include Hidden Markov Models and Recurrent Neural Networks to transform the ion currents into DNA sequences. These models must be trained on known DNA sequences, and the characteristics change for each chemistry update. For Oxford Nanopore’s MinION, even the best available methods only achieve error rates in the range of 10% for single reads, which is too high to reliably discover rare single nucleotide variants that may indicate minimal residual disease after therapy. Ryan R. Wick and colleagues maintain a living document with performance comparisons of base-calling methods [317].

## Bibliography

- [295] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022 (cit. on p. 274).
- [296] M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès. “SLOPE – adaptive variable selection via convex optimization”. In: *The annals of applied statistics* 9.3 (2015), p. 1103 (cit. on p. 274).
- [3] J. Bolte, S. Sabach, and M. Teboulle. “Proximal alternating linearized minimization for nonconvex and nonsmooth problems”. In: *Mathematical Programming* 146.1-2 (2014), pp. 459–494 (cit. on pp. 68, 274).
- [297] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. “Near-optimal probabilistic RNA-seq quantification”. In: *Nat. Biotechnol.* 34.5 (May 2016), pp. 525–527 (cit. on p. 276).
- [298] R. Chikhi and G. Rizk. “Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter.” In: *WABI*. Vol. 7534. Lecture Notes in Computer Science. Springer, 2012, pp. 236–248 (cit. on p. 284).
- [299] C. H. Ding, X. He, and H. D. Simon. “On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering.” In: *SDM*. Vol. 5. SIAM. 2005, pp. 606–610 (cit. on p. 275).
- [300] D. Eppstein. “Cuckoo Filter: Simplification and Analysis”. In: *15th Scandinavian Symposium and Workshops on Algorithm Theory, SWAT 2016, June 22-24, 2016, Reykjavik, Iceland*. Ed. by R. Pagh. Vol. 53. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2016, 8:1–8:12 (cit. on p. 282).
- [301] J. Friedman, T. Hastie, and R. Tibshirani. “Sparse inverse covariance estimation with the graphical lasso”. In: *Biostatistics* 9.3 (July 2008), pp. 432–441 (cit. on p. 273).

- [302] S. K. Gupta, D. Phung, B. Adams, and S. Venkatesh. “A matrix factorization framework for jointly analyzing multiple nonnegative data sources”. In: *Data Mining for Service*. Springer, 2014, pp. 151–170 (cit. on p. 275).
- [303] V. Justilien, M. P. Walsh, S. A. Ali, E. A. Thompson, N. R. Murray, and A. P. Fields. “The PRKCI and SOX2 oncogenes are coamplified and cooperate to activate Hedgehog signaling in lung squamous cell carcinoma”. In: *Cancer Cell* 25.2 (Feb. 2014), pp. 139–151 (cit. on p. 275).
- [304] H. Kim, J. Choo, J. Kim, C. K. Reddy, and H. Park. “Simultaneous Discovery of Common and Discriminative Topics via Joint Nonnegative Matrix Factorization”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. Ed. by L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams. ACM, 2015, pp. 567–576 (cit. on p. 275).
- [305] J. Köster and S. Rahmann. “Snakemake: a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19 (2012), pp. 2520–2522 (cit. on p. 273).
- [306] H. Li. “Minimap2: versatile pairwise alignment for nucleotide sequences”. In: *ArXiv e-prints* (Aug. 2017) (cit. on p. 272).
- [307] P. Miettinen. “On Finding Joint Subspace Boolean Matrix Factorizations”. In: *SDM*. 2012, pp. 954–965 (cit. on p. 275).
- [308] P. Miettinen and J. Vreeken. “Model order selection for boolean matrix factorization”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 51–59 (cit. on p. 274).
- [309] P. Miettinen, T. Mielikainen, A. Gionis, G. Das, and H. Mannila. “The discrete basis problem”. In: *Knowledge and Data Engineering, IEEE Transactions on* 20.10 (2008), pp. 1348–1362 (cit. on p. 274).
- [310] I. Müller, P. Sanders, R. Schulze, and W. Zhou. “Retrieval and Perfect Hashing Using Fingerprinting”. In: *Experimental Algorithms - 13th International Symposium, SEA 2014, Copenhagen, Denmark, June 29 - July 1, 2014. Proceedings*. Ed. by J. Gudmundsson and J. Katajainen. Vol. 8504. Lecture Notes in Computer Science. Springer, 2014, pp. 138–149 (cit. on p. 282).
- [311] M. Peifer, F. Hertwig, F. Roels, D. Dreidax, M. Gartlgruber, R. Menon, A. Kramer, J. L. Roncaioli, F. Sand, et al. “Telomerase activation by genomic rearrangements in high-risk neuroblastoma”. In: *Nature* 526.7575 (Oct. 2015), pp. 700–704 (cit. on p. 271).
- [312] J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, et al. “Real-time, portable genome sequencing for Ebola surveillance”. In: *Nature* 530.7589 (Feb. 2016), pp. 228–232 (cit. on p. 276).
- [313] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp. “Detecting DNA cytosine methylation using nanopore sequencing”. In: *Nature Methods* 14.4 (Apr. 2017), pp. 407–410 (cit. on pp. 282, 283).
- [314] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. “Measuring and testing dependence by correlation of distances”. In: *The Annals of Statistics* 35.6 (2007), pp. 2769–2794 (cit. on p. 285).
- [C4/315] T. Treppmann, **K. Ickstadt**, and M. Zucknick. “Integration of multiple genomic data sources in a Bayesian Cox model for variable selection and prediction”. In: *Computational and Mathematical Methods in Medicine* Vol. 2017 (2017), pp. 1–19 (cit. on pp. 285, 287, 316, 321).
- [316] L. Wang, Y. You, and H. Lian. “A simple and efficient algorithm for fused lasso signal approximator with convex loss function”. In: *Computational Statistics* 28.4 (Aug. 2013), pp. 1699–1714 (cit. on p. 283).

- [317] R. R. Wick, L. M. Judd, and K. E. Holt. *Comparison of Oxford Nanopore base-calling tools*. 2018 (cit. on pp. 277, 283).
- [318] H. Zhang, G. Kim, and E. P. Xing. “Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. Ed. by L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams. ACM, 2015, pp. 1425–1434 (cit. on p. 271).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [C1/319] J. Quedenfeld and **S. Rahmann**. “Analysis of Min-Hashing for Variant Tolerant DNA Read Mapping”. In: *17th International Workshop on Algorithms in Bioinformatics, WABI 2017, August 21-23, 2017, Boston, MA, USA*. Ed. by R. Schwartz and K. Reinert. Vol. 88. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017, 21:1–21:13 (cit. on pp. 19, 272, 276).
- [C1/320] C. Schröder and **S. Rahmann**. “A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification”. In: *Algorithms for Molecular Biology* 12 (2017), p. 21 (cit. on p. 273).
- [C1/321] **A. Schramm**, J. Köster, Y. Assenov, K. Althoff, M. Peifer, E. Mahlow, A. Odersky, D. Beisser, C. Ernst, et al. “Mutational dynamics between primary and relapse neuroblastomas”. In: *Nature Genetics* 47.8 (Aug. 2015), pp. 872–877 (cit. on pp. 19, 271–275, 321).
- [C1/322] **M. Schulte**, J. Köster, **S. Rahmann**, and **A. Schramm**. “Cancer evolution, mutations, and clonal selection in relapse neuroblastoma”. In: *Cell Tissue Research* 372.2 (May 2018), pp. 263–268 (cit. on p. 272).
- [C1/323] **M. Schwermer**, **S. Lee**, J. Köster, T. van Maerken, H. Stephan, A. Eggert, **K. Morik**, J. H. Schulte, and **A. Schramm**. “Sensitivity to cdk1-inhibition is modulated by p53 status in preclinical models of embryonal tumors”. In: *Oncotarget* (2015) (cit. on p. 275).
- [C1/20] **S. Hess** and **K. Morik**. “C-SALT: Mining Class-Specific ALTerations in Boolean Matrix Factorization”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2017*. Springer, 2017 (cit. on pp. 80, 275, 287).
- [A1/C1/32] **S. Hess**, **K. Morik**, and **N. Piatkowski**. “The PRIMPING routine—Tiling through proximal alternating linearized minimization”. In: *Data Mining and Knowledge Discovery* 31.4 (July 2017), pp. 1090–1131 (cit. on pp. 68, 71, 80, 274, 287).
- [A1/C1/33] **S. Hess**, **N. Piatkowski**, and **K. Morik**. “The Trustworthy Pal: Controlling the False Discovery Rate in Boolean Matrix Factorization”. In: *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA*. SIAM. 2018, pp. 405–413 (cit. on pp. 71, 80, 274, 287).
- [C1/324] **S. Lee**, D. Brzyski, and M. Bogdan. “Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered l1-Norm”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by A. Gretton and C. C. Robert. JMLR W&CP, 2016, pp. 780–789 (cit. on p. 274).

- [B2/C1/211] **V. Shpacovitch, I. Sidorenko, J. E. Lenssen, V. Temchura, F. Weichert, H. Müller, K. Überla, A. Zybin, A. Schramm, et al.** “Application of the PAMONO-sensor for Quantification of Microvesicles and Determination of Nanoparticle Size Distribution”. In: *Sensors* 17.2 (2017), pp. 1–14 (cit. on pp. 20, 196, 198, 287).

## 3.4 Project plan

### Goals

Given the recent developments described above, our first goal is to enable the *analysis of large WGS data sets above 100 GB per sample on commodity hardware* (see Table 3.1), using  $k$ -mer methods based on ultra-fast hashing (WP 1 and 2). With decreasing sequencing costs and accessible data analysis, we anticipate that WGS and other high-throughput technologies will become the standard choice for *research projects* in oncology, where the goal is to understand mechanisms behind cancer development and evolution, to which we contribute with feature extraction and selection in WPs 6 and 7, respectively.

Our second goal is to overcome the current challenges, especially the high error rates, present in nanopore sequencing and *enable mobile (soft) real-time analysis of high-frequency high-volume nanopore ion current data*. To achieve this goal, we experiment with new suitable algorithms and data structures, using a combination of signal processing methods (building on methods developed previously with project TB1; WP 4) and  $k$ -mer methods applied to discretised signals (WP 5). This entails working in different feature spaces than those previously investigated.

Our third goal is to use nanopore sequence data from non-invasively obtained samples, i.e., *liquid biopsies from blood or urine, to monitor tumour progression, therapy success or failure, and to inform about residual disease after therapy* (WP 3). For this, we characterise molecular tumour fingerprints that are efficiently obtainable and quantifiable by mobile nanopore sequencing and find efficient ways to declare presence or absence of such fingerprints in new samples (WP 6 and 7). Our ongoing efforts in the lab will use advanced genome engineering technologies such as the CRISPR/Cas9 system, which allows for precise introduction of genetic variants either to mimic oncogenic effects or the consequences of targeted therapies to validate the biological function of discovered biomarkers (WP 8). Here, we focus on the evaluation of the different technical possibilities offered by CRISPR/Cas9, including re-activation of genes silenced in tumour cells.

Further goals are to continue the highly fruitful collaborations with projects A1 on machine learning methods for discovering structure in feature matrices and B2 on tumour vesicle detection with the PAMONO sensor, as well as the promising collaborations with projects A6 on feature similarity and network analysis and C4 on applying their sample size reduction techniques to dimensionality reduction. An overview of our goals, methods and work packages is shown in figure 3.3.

### Work schedule

**Work package 1.  $K$ -mer methods for whole genome analysis** An increasing number of paired tumour/normal whole genome (WGS, see Table 3.1) data sets are becoming available, both in-house (currently 56 neuroblastoma data sets, soon also lung cancer data sets) and from public sources such as The Cancer Genome Atlas (TCGA), which contains data on several different tumour entities, including lung cancers, to which we have obtained access permission. Instead of using the established mapping/alignment approach on a typical sample of  $10^{11}$  sequenced basepairs, we propose to pursue the following  $k$ -mer based approach that generates, for each sample, a relatively small set of  $k$ -mers whose count differs from the expected count based on the human reference genome.

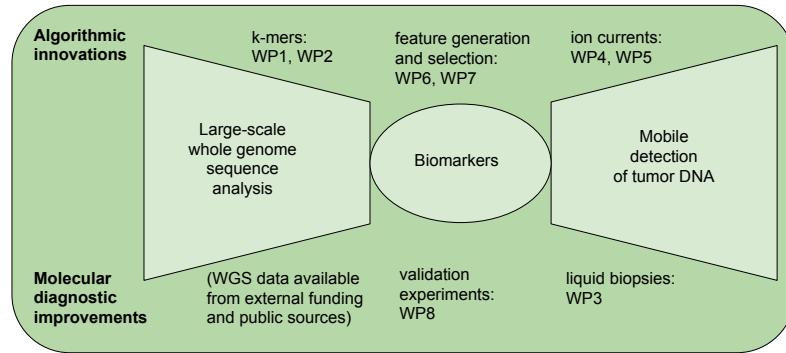


Figure 3.3: Overview of work packages in project C1.

For this, we need a fast and small data structure that supports a key-value store on DNA  $k$ -mers (see WP 2). For each  $k$ -mer of the human reference genome, we store twice its count in a “reference counter” (twice because the normal copy number of a unique  $k$ -mer in the reference genome is 2, referring to the maternal and paternal allele). For each WGS sample, we (separately) create a “sample counter” of  $k$ -mers. Assuming a Poisson distribution for the coverage, we test whether the sample count deviates significantly from the expected count for each  $k$ -mer and only report deviating  $k$ -mers and their estimated copy numbers in the sample. To compare paired tumour/normal samples, the normal sample takes the role of the reference.

Each sample (pair) can be processed independently, in parallel and/or on different machines. Finally, the lists of deviating  $k$ -mers are aggregated into one  $k$ -mer feature table and further filtered, retaining only  $k$ -mers that deviate in a given minimum number of case or control samples. This directly yields features to analyse, continuing our fruitful methodological collaboration with project A1, but a biologist would find lists of  $k$ -mers hard to interpret. Therefore, we investigate whether “classical” features (single nucleotide mutations, short insertions and deletions, copy number alterations of large regions, translocations, etc.) can be recovered from deviating  $k$ -mer statistics. This may not be possible in all cases, but there is evidence that such an approach has a high chance of success. Consider that, for  $k \geq 23$ , over 80% of the  $k$ -mers in the human genome are unique and hence point to a unique location, each of which may serve as an anchor point for overlapping non-unique deviating  $k$ -mers in the sample. For example, an “ideal” isolated heterozygotic single nucleotide mutation would decrease the estimated copy numbers of the  $k$  affected (unique)  $k$ -mers from 2 to 1 and increase the counts of the newly formed  $k$ -mers containing the mutated nucleotide from 0 to 1. In reality, the data is noisy, mutations may not be isolated (i.e., another mutation is less than  $k$  basepairs away), the affected  $k$ -mers may not all be unique, and the mutated  $k$ -mers may exist elsewhere in the genome. Nevertheless, using appropriate Bayesian models and the maximum a posteriori approach, it is reasonable to assume that most of the single nucleotide variants can be discovered in this way. We will use similar statistical counting approaches for other feature types (short insertions or deletions of DNA, copy number variations, translocations, inversions) and investigate the benefits and the limits of the  $k$ -mer based approach for each feature type by comparing it with the classical mapping/alignment approach on selected neuroblastoma and lung cancer WGS data sets.

**Work package 2. Fast and small key-value stores for DNA  $k$ -mers** Each  $k$ -mer based method needs a data structure to retrieve the information associated to a given  $k$ -mer or declare that the  $k$ -mer is not present. Several such data structures have been proposed, and there is a trade-off between memory requirements and lookup speed. The speed is dominated by the number of random access memory lookups (which typically induce cache misses) necessary to retrieve the information for a  $k$ -mer. A cache miss memory lookup may take up to several hundred times more time than

a simple arithmetic operation, such as an addition or multiplication. Small data structures (e.g., FM-index or compressed suffix array) typically need many random-access lookups, while very fast methods (ideally, direct addressing by the base-4 representation of the  $k$ -mer) typically take up too much space to be practical.

We will engineer an optimised  $k$ -mer key-value store, which for most present  $k$ -mers retrieves the associated value with a single cache miss. Preliminary experiments by the group of Sven Rahmann indicate that 3-way bucketed Cuckoo hashing, together with a small overflow table, is a promising approach that allows us to retrieve the information of at least 70% of present  $k$ -mers with a single cache miss, while declaring absence of a  $k$ -mer may require up to six cache misses. Methodologically, we combine and expand on ideas (a) from Cuckoo filters [300], but guarantee that we have no false positives, and (b) from the fingerprint hashing work [310], which proposes using the hash table address as part of the key and explicitly storing only a small fingerprint of the key. Several design options will be evaluated in terms of speed/memory trade-off: two-way or three-way cuckoo hashing, size of buckets, trade-off between overprovisioning in the primary table vs. frequently using the overflow table.

While we optimise the data structure for storing DNA  $k$ -mers, we expect that with modifications, these ideas can be applied in project A6 for storing and counting graph features and tracking them over time. The main design restriction currently is the fixed key size.

**Work package 3. Tumour sequencing and quality control** Based on our discovery that upregulation of PRKCI is a common mechanism in tumour progression associated with fatal outcome, we extend our perspective from neuroblastoma to other tumour types, especially to lung tumours: Patients with advanced or metastatic small-cell lung cancer or non-small-cell lung cancer eligible for systemic cancer therapy at the Lung Cancer Center of the West German Cancer Center will be included. Before initiation of first-line or later-line therapy, blood samples will be obtained and tumour cells will be isolated using techniques established in Alexander Schramm's Molecular Oncology lab. DNA from tumour cells, circulating free DNA (cfDNA), and control blood will be submitted to DNA sequencing analyses using nanopore sequencing technology, which has already been acquired and tested by A. Schramm's lab in early 2018.

We distinguish between *exploratory sequencing*, where the goal is to *discover* features (biomarkers) that distinguish cancer cells from their normal controls, and *diagnostic sequencing*, where the goal is to *detect* a set of predefined features and decide on the presence/absence of tumour cells, or to estimate their concentration for risk prognosis. In particular, our goal is to detect tumour-specific features when they are still rare and hence "hidden" in a background of normal DNA.

For exploratory sequencing, nanopore reads can unveil complementary features to traditional whole genome sequencing because of the much longer reads (tens of thousands in comparison to hundreds of bases), which makes it easier to directly uncover genomic rearrangements and gene fusions. Nanopore sequencing is also able to detect chemical DNA modifications, such as DNA methylation [313]. Mutations, rearrangements and methylation state profiles from sequenced samples will be stored together with clinical and epidemiological data and with the outcomes of therapies, i.e., immune checkpoint inhibitor therapy or prior and subsequent cytotoxic or molecularly targeted therapies. This curated data provides input for feature extraction and selection for the subsequent work packages.

For diagnostic sequencing, we will establish a novel tumour gene panel based on CAPP-Seq data (CAncer Personalised Profiling by deep Sequencing). Present implementations of CAPP-Seq allow for the detection of one molecule of mutant DNA against a background of 10,000 normal DNAs. As lung cancer has both recurrent as well as private mutations (e.g., mutations that occur only in single patients), an optimum of sensitivity and specificity has to be defined in the course of

this project using the MinION technology: Extracellular vesicles (EVs) were shown to contain nucleic acids derived from tumour cells that could serve as a source for tumour-specific DNA markers. In an ongoing collaboration with project B2, EVs will be isolated from plasma samples using the PAMONO sensor in project B2 and provided to us for sequencing and analysis. To increase detection sensitivity, we will exploit a new possibility of nanopore sequencing: ejection of non-informative DNA molecules from the pore and moving on to the next free DNA segments. This requires close-to-real-time analysis of the ion current and could be used to achieve *importance sampling* of DNA molecules. As a consequence, rare events in tumour genomes as well as a lower number of tumour cells against a background of normal cells become detectable. Moreover, this will make it possible to save consumables and lower the overall costs of these assays. Our task is to implement fast rules for early ejection of molecules, which we should be able to achieve using the fast  $k$ -mer indices from WP 2.

**Work package 4. DNA base calling from ion currents** As mentioned in the section *Current state of research*, it is a challenge to convert the ion current signal from a nanopore sequencer into a correct DNA sequence, with state of the art error rates between 10% and 15%. The input to the *base calling problem* consists of a noisy ion current signal sampled at very high frequency, of length approximately proportional to that of the DNA sequence sought, and the task at hand is to infer the DNA basepair sequence (a string over {A,C,G,T}) from the signal, as shown in figure 3.2. Notably, the number of ion current measurements for each basepair may vary. The basic assumption of this work package is that there exists a generative model that, given the DNA  $k$ -mer inside of the nanopore ( $k = 5, 6$  have been used according to [317]), describes the distribution of the ion current signal intensity and its duration. Some of the methods discussed by Wick et al. [317] initially partition the signal into *events*, i.e., time intervals where the signal properties are approximately constant, and compute the signal properties (e.g., mean and variance). A (soft) real-time decision that segments the streaming signal into events has been challenging, and most of the current methods perform an offline analysis of a recorded signal. We propose to use the fused LASSO signal approximator (FLSA) [316] for segmentation, because it can smooth and segment a noisy signal at the same time, retaining sharp edges instead of blurring the signal at change points: Let  $y = (y_i)_{i=1,\dots,n}$  be a noisy signal; let  $x = (x_i)$  be the corresponding unknown original signal, which is assumed to be piecewise constant. Given  $y$ , an estimate for  $x$  can be obtained by solving the convex FLSA problem

$$\text{minimise } f(x) := \sum_{i=1}^n (x_i - y_i)^2 + \lambda \cdot \sum_{i=1}^{n-1} |x_{i+1} - x_i|.$$

An optimal solution will have the property that each  $x_i$  is close to  $y_i$  and at the same time, nonzero differences between consecutive  $x_i$  values are sparse because of the  $\ell_1$  regularisation term on the differences. The parameter  $\lambda \geq 0$  controls the amount of smoothing. In our collaboration with TB1 and in the master's thesis of Elias Kuthe, jointly supervised by PIs Lee and Rahmann, we used FLSA for preprocessing ion mobility spectrometry measurements. To calibrate  $\lambda$ , we examine how the distribution of the duration of piecewise constant segments changes with  $\lambda$  and restrict the parameter to physically plausible ranges (speed of the DNA molecule as it moves through the nanopore).

To model the distribution of the signal intensity during each event, Gaussian distributions with state-dependent parameters have been used, with parameters depending on the DNA  $k$ -mer inside the pore. In addition, chemical modifications of the DNA (in particular, cytosine methylation) have been considered [313]. The transition between DNA states has been modelled by Markov chains, which yields a Hidden Markov model (HMM) for the signal intensity process. We hypothesise that the signal of a single event can be modelled as a Gaussian process whose parameters depend on  $k$ -mers of an *extended* DNA alphabet, taking chemical DNA modifications into account, and  $k \in \{4, 5, 6, 7\}$  from experience. The size of this extended alphabet  $\Sigma$  is currently unknown, but

given large collections of sequenced known DNA, we should be able to estimate it by comparing the model complexity  $|\Sigma|^k$  with the number of observed distinct signal levels in  $x$  and finding the optimal compromise between model fit and complexity using penalised likelihood approaches. Partitioning the segmented signal  $x$  into discrete levels is a new approach that has not been attempted before. Once a level set is given, we hypothesise that the corresponding restricted FLSA problem can be solved more efficiently than a general FLSA problem, and we propose to develop a delayed streaming dynamic programming algorithm to filter the signal while still recording. DNA sequence inference is a two-step process: Using the FLSA segmented signal, infer the state sequence using HMM decoding (Viterbi, forward-backward), and then project the state to the regular DNA alphabet.

Extensions of the basic approach described above may need to be considered: variable-order Markov chains (i.e., dynamic choice of  $k$  depending on context) or mixture models. It may be the case that some DNA states cannot be distinguished by the ion current signal, and only an ambiguous (incompletely resolved) DNA sequence can be obtained. We will consider this possibility and explore to which degree even an incompletely resolved DNA sequence can be mapped to a known reference genome, and whether the presence of single nucleotide variants can be reliably detected.

**Work package 5. Alternative feature spaces** As mentioned in WP 4, the DNA sequence may not be fully resolvable using ion current signals. Here, we literally turn this problem into features: Our idea is to work directly with the discretised ion current level alphabet  $\Gamma$  (which might consist of approx. 500 different symbols), which apparently has not been attempted yet. The result of sequencing a DNA fragment would then be a sequence of symbols from  $\Gamma$ . By sequencing known DNA fragments, the forward transformation  $\Sigma \rightarrow \Gamma$  can be estimated with high reliability. More generally, considering the possibility of chemically modified DNA, the forward transformation would be a mapping  $\Sigma \rightarrow \mathcal{D}(\Gamma)$ , where  $\mathcal{D}(\Gamma)$  denotes the family of probability distributions over  $\Gamma$ . Representing a reference DNA sequence as a sequence  $r$  of values in  $\Gamma$  or  $\mathcal{D}(\Gamma)$ , we can compare a sequenced sample  $s \in \Gamma^*$  to  $r$  using sequence comparison algorithms with a similarity measure on  $\Gamma \times \Gamma$  or  $\Gamma \times \mathcal{D}(\Gamma)$ . The  $k$ -mer methods from WP 1 and 2 can be applied to alphabet  $\Gamma$  to develop fast index-based methods.

The eventual result of this idea would be the discovery of cancer-specific features in  $\Gamma^k$ -space ( $k$ -mers of signal values). As the novelty of this representation will require extensive validation before it can become an accepted part of clinical practice, we will validate proposed features using molecular biology techniques described in WP 8, and, on a case-by-case basis, find the underlying DNA modification.

**Work package 6. Feature generation and extraction** In the past phase, we established reproducible DNA/RNA sequence analysis processes to obtain feature vectors from sequenced samples, including presence/absence of mutations, large-scale structural gains, losses and rearrangements, gene expression and methylation changes. In the upcoming phase, more abstract features will be produced initially: presence or absence of DNA  $k$ -mers (binary) in cases vs. controls, over- or underrepresentation of DNA  $k$ -mers (quantitative) in cases vs. controls, from both whole genome and the smaller-scale nanopore sequencing. Similar features are obtained for the alphabet  $\Gamma$  introduced in WP 5 from nanopore sequencing. The resulting feature matrices from whole-genome data sets are of even higher feature dimension than we previously analysed. In addition,  $k$ -mer based features are hard to interpret for the clinician or biologist and give no direct hints towards the underlying biological processes. To reduce the feature dimension and to provide better interpretable features at the same time, we propose to assemble  $k$ -mers that behave differently between two classes of samples into longer sequences (of DNA or  $\Gamma$ -values). This can be done at the string level, using (at least for DNA) existing genome assemblers, such as Minia [298], that look for

long overlaps between  $k$ -mers, but it can also be complemented by the approaches that discover structure in Boolean matrices that we have developed in collaboration with project A1, to discover co-occurring  $k$ -mers. We will explore combinations of these complementary approaches to obtain reduced feature matrices.

**Work package 7. Feature selection and risk prognosis** From reduced interpretable features (from WP 6), including presence/absence or frequencies of mutations, chromosomal gains, losses and rearrangements, gene expression and methylation changes, (a) dependencies and redundancies need to be identified, and (b) independent prognostically relevant features need to be selected. We point out that *changes in the feature dependency structure* between different tumour subtype samples may be an important prognostic feature.

To discover complex feature dependencies, we will use the *distance correlation* measure [314]. Let  $X$  and  $Y$  be two random variables and  $x = (x_1, \dots, x_n)$ ,  $y = (y_1, \dots, y_n)$  size- $n$  samples of  $X$  and  $Y$ , respectively. The standard Pearson linear correlation  $\rho$  of  $X$  and  $Y$  and the sample correlation  $\rho_n$  of  $x$  and  $y$  are defined, respectively, as

$$\rho(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}, \quad \rho_n(x, y) := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Pearson correlation is restricted to two real-valued random variables and only captures linear dependencies reliably. In contrast, distance correlation  $\delta(X, Y)$  can be defined more generally for random vectors  $X$  ( $p$ -dimensional) and  $Y$  ( $q$ -dimensional) and captures any dependency in the sense that  $\delta(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent [314]. To define it, we need the vector's characteristic functions  $\phi_X(s) := \mathbb{E}[\exp(i\langle s, X \rangle)]$  and  $\phi_Y(t) = \mathbb{E}[\exp(i\langle t, Y \rangle)]$  for  $s \in \mathbb{R}^p$ ,  $t \in \mathbb{R}^q$ , where  $i = \sqrt{-1}$  is the imaginary unit and  $\langle \cdot \rangle$  denotes the inner product. It is known that  $X$  and  $Y$  are independent if and only if  $\phi_{X,Y}(s, t) = \phi_X(s) \cdot \phi_Y(t)$  for all  $s, t \in \mathbb{R}^p \times \mathbb{R}^q$ . Define the *distance covariance* by a weighted integral over the squared deviation from independence,

$$\mathcal{C}(X, Y) := \left( \frac{1}{c_p c_q} \cdot \int_{\mathbb{R}^p \times \mathbb{R}^q} \frac{|\phi_{X,Y}(s, t) - \phi_X(s) \cdot \phi_Y(t)|^2}{\|s\|^{p+1} \|t\|^{q+1}} \, ds \, dt \right)^{1/2},$$

where  $c_p$ ,  $c_q$  are normalisation constants,  $c_p = \pi^{(p+1)/2} / \Gamma((p+1)/2)$ . Finally, the *distance correlation coefficient* is defined as

$$\delta(X, Y) = \mathcal{C}(X, Y) / \sqrt{\mathcal{C}(X, X) \cdot \mathcal{C}(Y, Y)}.$$

In practice, we use the sample version defined by replacing the expectation operator  $\mathbb{E}$  in the characteristic functions  $\phi$  by the sample mean. Surprisingly, the squared sample distance covariance  $\mathcal{C}_n(X, Y)^2$  can be obtained from a simple sum with  $n^2$  terms, even though defined via a complicated integral [314]. Evaluating the distance correlation between genetic events (mutations, losses, gains, translocations) and gene expression changes, we will be able to identify genetic events that are mutually exclusive but have similar effects on the transcriptional program of the cell, pointing at a common underlying mechanism, e.g., the disruption of the same pathway. This should allow us to compile seemingly unrelated genetic events into a single feature with high prognostic relevance. In addition, continuing work on approximate Boolean matrix factorisation in collaboration with A1, we intend to discover complementary structure and further reduce the feature dimension.

With project C4, we will develop regularised regression models to predict the order of magnitude of the event-free survival time, overall survival time, or degree of therapy success for a proposed therapy. While they have used Cox regression models [C4/315], we plan to use LASSO-penalised logistic regression to select informative features with high predictive value for long-term survival. In order to scale this procedure to millions of features, additional techniques from project C4 will

be required, such as random linear sketches of features or coresets of features, turning their sample size reduction methods into dimensionality reduction methods for C1. Naïvely, this amounts to applying the methods to the transposed matrix, but since no class labels would be present after transposition, we need to find ways of encoding labels in the matrix and then use methods that do not rely on labelled data.

**Work package 8. Biological validation of tumour-specific variants** For validation, surplus tumour and lymph node biopsies as well as fresh-frozen, paraffin-embedded samples and fresh-frozen samples deposited in the West German Biobank Essen are available. The lab of A. Schramm has acquired experience of CRISPR/Cas9 technology that can be used to validate the functional relevance of tumour-specific genetic aberrations, and the CRISPR/Cas9 technology is currently being adapted to cultivated lung cancer cells. Feasibility studies involving lung cancer cells and induction of fluorescently labelled proteins along with CRISPR/Cas9 component in these cells have served as proof-of-concept.

Two main questions will be answered in biological validation experiments: First, we will evaluate the contribution of mutations found in patients for the viability and aggressive properties of cancer cells. Second, we will address the issue of intratumoural heterogeneity, which requires molecular barcoding of cells. Here, cells are individually tagged by small and unique DNA sequences, which are introduced into individual cancer cells prior to the experiments. Once tumour cells are equipped with both patient-derived mutations and individual barcodes, they can be treated with different targeted therapies. Surviving cells, representing a therapy-resistant population, can then be recovered, and sequencing of these cells will allow assessment of the contribution of individual mutations to therapy resistance. In addition to the diagnostic value of identifying tumour-specific mutations and early relapses from blood samples, we will thus be able to use *in vitro* models for predicting response to therapy in lung cancer patients.

## Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. K-mer methods for whole genome analysis																	
2. Fast and small key-value stores for DNA <i>k</i> -mers																	
3. Tumour sequencing and quality control																	
4. DNA base calling from ion currents																	
5. Alternative feature spaces																	
6. Feature generation and extraction																	
7. Feature selection and risk prognosis																	
8. Biological validation of tumour-specific variants																	

### 3.5 Role within the Collaborative Research Centre

In the C1 project, we address the challenge of feature extraction from high-volume raw data and feature selection in the resulting extremely high-dimensional feature data that are typically obtained in the life sciences from molecular biology technologies, in particular DNA/RNA sequencing data (gene expression, genetic variation).

In phase 2, we benefitted from methodological cooperations with project A1, as evidenced by several joint publications on Boolean matrix factorisation [A1/C1/32, A1/C1/33], which we adapted to the structure of our genetic variant data [C1/20]. In return, we provided extremely high-dimensional real-world data sets to project A1. We intend to continue this fruitful mutual collaboration concerning the hidden structure of feature matrices. As the dimensionality of the data sets in C1 will increase further, we will need to discuss methods with the PIs of project A1 on how to meaningfully aggregate and project features to lower dimensions to make their methods better applicable.

We contributed our expertise in the management of multistep workflows with large data sets and reproducible data analysis to a collaboration with project TB1 to analyse ion mobility spectrometry (IMS) data sets with different combinations of preprocessing, feature extraction, feature selection and classification algorithms (see the report of TB1). The computational experiments on IMS data motivated us to use the fused LASSO signal approximator on ion current signal data.

We collaborated with project B2, where we provided tumour cell-derived samples, from which tumour-specific vesicles could be detected using the PAMONO sensor [B2/C1/211]. In our upcoming work package 3, we will extend the cooperation with project B2 by aiming to sequence nucleic acids contained in tumour-derived vesicles.

We initiated a collaboration with project C4 to continue research on their method [C4/315] to predict (the order-of-magnitude of) event free survival times after tumour therapy based on molecular features, using Cox regression models, but found that we need to reduce the feature dimension further to successfully apply the method. We want to deepen this collaboration and focus on LASSO-penalised logistic regression for classification. Furthermore, we would like to investigate the possibility of applying the *sample size reduction methods* (e.g., coresets) researched in project C4 to *reducing the dimensionality* of our feature sets.

In additional work, we initiated a cooperation with project A6 to define an efficiently computable similarity measure between (small) labelled graphs, using protein complexes as examples. We plan to intensify this collaboration with A6 in the next phase when we generate (large) dependency graphs between molecular features, where we expect that not only the feature values change with the examined condition (e.g., tumour type, or time since treatment), but also their dependency structure. We intend to use results from A6 on dynamic graphs to both quantify the speed of molecular evolution (from time of therapy) and pinpoint the accumulated differences in the dependency graph structure.

## 3.6 Differentiation from other funded projects

### UA Ruhr Professorship Computational Biology

**(Rahmann, Reference number MERCUR Pe-2013-0012) (Funding period: 2014–2019)**

Sven Rahmann receives funding from Mercator Research Center Ruhr (MERCUR) to establish and maintain a lasting collaboration between the Faculty of Medicine at University of Duisburg-Essen and the Faculty of Computer Science at TU Dortmund University. The funds support compute and storage infrastructure and personnel to transfer new developments in algorithmics to applications in medicine and initiate collaborations between computer scientists and life scientists. The preliminary experiments with fast hashing for compact  $k$ -mer key-value stores were supported by this project, which ends in February 2019.

### OsteoSys

**(Rahmann, Reference number EFRE-0800427) (Funding period: 2016–2019)**

The goal of this consortium, funded by LeitmarktAgentur.NRW and the European Fund for Regional Development (EFRE), is to pave the way for a personalised therapy for osteoporosis, using statistical methods on data obtained in a clinical study. No high-volume or high-dimensional molecular data is produced or analysed. Sequencing occurs for 20–50 selected genes with known effects on bone density. The focus is on clinical data and bone density measurements. There is no overlap with the research in this CRC.

### Role of the tyrosine kinase TrkA and TrkB in checkpoint activation and DSB repair

**(Schramm, Reference number DFG GRK 1739) (Funding period: 2018–2021)**

The project investigates responses of tumour cells to radiation and DNA-damaging agents, focusing on the role of two specific proteins. There is no relation to resource-efficient high-throughput data analysis.

### Modelling primary tumor metabolism in neuroblastoma to identify central nodes for therapeutic intervention

**(Schramm, Reference number BMBF 01ZX1307C) (Funding period: 2017–2019)**

This is a systems medicine project within a consortium aiming to understand the metabolism in tumour cells. There is no methodological overlap to this CRC.

### Interaction of TrkA and MYCN in neuroblastoma

**(Schramm, Reference number Sander-Stiftung 2016.119.1) (Funding period: 2017–2019)**

In cooperation with Barbara Sitek from Ruhr-University Bochum we investigate the changes in protein composition (the proteome) in model systems of neuroblastoma. The project uses completely disjoint technology to project C1 and provides a complementary view on tumours at the protein level instead of the genetic level.

### Chemo-genetic interference with Survivin functions in embryonal tumors

**(Schramm, Reference number Wegener Stiftung Project Nr. 29) (Funding period: 2018–2018)**

This short-term project, initiated by Shirley Knauer from University Duisburg-Essen, evaluates the role of the anti-apoptotic protein Survivin, which is up-regulated in many tumour cells. It is a small-scale molecular biology project.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	1	64,500	1	64,500	1	64,500	1	64,500
Doctoral researchers, 65 %	1	41,900	1	41,900	1	41,900	1	41,900
Non-research staff, 100 %	1	48,000	1	48,000	1	48,000	1	48,000
Total	—	154,400	—	154,400	—	154,400	—	154,400
<b>Direct costs</b>	<b>Sum</b>		<b>Sum</b>		<b>Sum</b>		<b>Sum</b>	
Instrumentation up to 10,000 euros, software and supplies	15,000		15,000		15,000		15,000	
Total	15,000		15,000		15,000		15,000	
<b>Instrumentation</b>	<b>Sum</b>		<b>Sum</b>		<b>Sum</b>		<b>Sum</b>	
Total	0		0		0		0	
<b>Grand total</b>	<b>169,400</b>		<b>169,400</b>		<b>169,400</b>		<b>169,400</b>	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

290

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Sven Rahmann, Prof. Dr., professor	Bioinformatics	University Hospital Essen	10	—	Existing funds
	2	Alexander Schramm, Prof. Dr., professor	Molecular medicine	University Hospital Essen	10	—	Existing funds
	3	Johannes Köster, Dr., postdoctoral researcher	Bioinformatics	University Hospital Essen	9.96	—	Existing funds
	4	Marcel Wiesweg, Dr. med., postdoctoral researcher	Molecular medicine	University Hospital Essen	9.96	—	Existing funds
	5	Till Hartmann, M.Sc., doctoral researcher	Bioinformatics	University Hospital Essen	19.92	—	Existing funds
Non-research staff	6	Annette Parr, medical-technical assistant (mta)	—	University Hospital Essen	19.96	—	Existing funds
	7	Claudia Wolf, secretary	—	University Hospital Essen	5	—	Existing funds
<b>Requested staff</b>							
Research staff	8	N.N., doctoral researcher	Molecular medicine	University Hospital Essen	—	Doctoral researcher	—
	9	N.N., doctoral researcher	Bioinformatics	University Hospital Essen	—	Doctoral researcher	—
Non-research staff	10	Sabine Dreesmann, non-research staff	—	University Hospital Essen	—	Non-research staff	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):****1. Rahmann, Sven**

Supervision of researchers in WPs 1,2,4,5, joint supervision of researchers with A. Schramm in WPs 6 and 7; project management; code reviews; assistance with scientific writing; career advice.

**2. Schramm, Alexander**

Supervision of researchers in WPs 3 and 8; joint supervision of researchers with S. Rahmann in WPs 6 and 7; project management; assistance with experiment planning, laboratory techniques and scientific writing; career advice.

**3. Köster, Johannes**

Proposition of different alternative feature spaces for WP 5; generation of feature matrices from *k*-mers and long nanopore reads, dimensionality reduction by aggregating features in consultation with Marcel Wiesweg and both PIs Rahmann and Schramm in WP 6.

**4. Wiesweg, Marcel**

Dr. Wiesweg is an oncologist with practical experience in bioinformatics and computational learning. In 2017/18, he was on leave from medical duties and developed survival models for lung cancer entities under the supervision of S. Rahmann. He is responsible for the definition of meaningful features from an oncologist's point of view in WP 6, in collaboration with Joannes Köster and both PIs. He will advise the PhD student on the selection of prognostic models in WP 7 together with the PIs, and he will be involved in patient cohort recruitment for validation experiments (WP 8).

**5. Hartmann, Till**

Development of fast compact *k*-mer key-value stores with different memory layouts and their experimental evaluation in WP 2; subsequently development of fused LASSO approaches for segmenting the ion current signal and research of generative models for the ion current signal given the DNA sequence, as well as inference methods for the inverse problem (WP 4).

**6. Parr, Annette**

Sample acquisition, tumor sequencing and clinical documentation in WPs 3 and 8.

**7. Wolf, Claudia**

Administrative assistance with travel booking, purchasing, hirings, documenting work times, vacation, sickness.

**Job descriptions of staff for the proposed funding period (requested funds):****8. N.N.**

Wet lab work, tumor sequencing using the MinION device and quality control, collaboration with project B2 on sequencing of tumor DNA in extracellular vesicles (WP 3); validation experiments of predicted relevant features using larger cohorts and targeted approaches (WP 8); writing publications and a thesis.

**9. N.N.**

Concept, design, implementation and testing of *k*-mer based whole genome analysis (WP 1); construction of discretised signal spaces from ion currents (WP 4); exploration of discretised signal spaces as feature spaces (WP 5); inventing methods feature selection and risk prognosis in extremely high dimensional settings (WP 7) in collaboration with methods from project A1 and using techniques from project C4; writing publications and a PhD thesis.

**10. Dreesmann, Sabine**

Functional evaluation of mutations, e.g. by analyses of survival and death of tumor cells as a function of mutational load, and isolation of extracellular vesicles. The latter additionally enhances cooperation with project B2.

### 3.7.4 Requested funding for direct costs for the new funding period

	2019	2020	2021	2022
University Hospital Essen: existing funds from university	12,500	1,500	1,500	1,500
Sum of existing funds	12,500	1,500	1,500	1,500
Sum of requested funds	15,000	15,000	15,000	15,000

(All figures in euros)

Instrumentation up to 10,000 euros, software and supplies for financial years 2019–2022

The wet lab work requires 15 000 EUR p.a. in addition to university funding, as cell biology and nucleic acid sequencing are complex and comparatively expensive. Breakdown of costs per year: Cell culture media, flasks, serum, required for data acquisition and validation: 1500 EUR. Nucleic acid extraction, required for data acquisition and validation: 1500 EUR. PCR and real-time PCR, required for data acquisition and validation: 2000 EUR. Antibodies /CRISPR/ CAS9 reagents, required for validation: 2000 EUR. Phenotyping assays (Cell Death ELISA, BrdU-Assays FACS-based Apoptosis Assays), required for validation: 2000 EUR.	EUR	15,000
--	-----	--------

### 3.7.5 Requested funding for instrumentation for the new funding period

This project does not request any funding for major research instrumentation.

### 3.1 General information about Project C3

### 3.1.1 Project title:

# Multi-level statistical analysis of high-frequency spatio-temporal process data

### 3.1.2 Research area(s):

409-08 (Massively Parallel and Data-Intensive Systems), 311-01 (Astrophysics and Astronomy)

### 3.1.3 Principal investigator(s)

Morik, Katharina, Prof. Dr., 14.10.1954, German

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

Phone: 0231-755-5100

E-mail: katharina.morik@tu-dortmund.de

Rhode, Wolfgang, Prof. Dr. Dr., 17.10.1961, German

Experimentelle Physik E5, Fakultät Physik, Technische Universität  
Dortmund  
Otto-Hahn-Straße 4  
44227 Dortmund

Phone: 0231-755-3550

E-mail: [wolfgang.rhode@tu-dortmund.de](mailto:wolfgang.rhode@tu-dortmund.de)

Ruhe, Tim, Dr., 08.09.1981, German

Experimentelle Physik E5, Fakultät Physik, Technische Universität  
Dortmund  
Otto-Hahn-Straße 4  
44227 Dortmund

Phone: 0231-755-8500

E-mail: [tim.ruhe@tu-dortmund.de](mailto:tim.ruhe@tu-dortmund.de)

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

*Dr. Tim Ruhe*

End date of fixed-term contract: 30.09.2018

Further employment is planned until 01.10.2024.

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes    (x) no
2.	clinical trials	( ) yes    (x) no
3.	experiments involving vertebrates.	( ) yes    (x) no
4.	experiments involving recombinant DNA.	( ) yes    (x) no
5.	research involving human embryonic stem cells.	( ) yes    (x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes    (x) no

## 3.2 Summary

In astroparticle physics, data analysis is a stepwise procedure, where one major step consists of a preselection of physically relevant signal examples from an overwhelming number of undesired background examples. This step generally requires estimated attributes of the examples like energy and direction. Based on these attributes, which are estimated in a first step, the separation of signal and background examples proceeds in a second step, resulting in a purified set of examples of certain signatures or particles well suited for specific physics analyses. Present telescopes do not require the data to be separated in real time. The next generation of experiments, however, requires a considerable acceleration of the procedures. The analysis is concluded in a third step, consisting in the solution of a specific inverse problem, like the reconstruction of source spectra. Simulated examples, which have to be produced in large numbers and with high precision, provide the basis for all of the aforementioned analysis steps.

The next generation of large-scale telescopes in astronomy and astroparticle physics – CTA [326], SKA [332], and IceCube-Gen2 [331] – is going to acquire data with a new level of precision and at an unprecedented data rate of up to 1 Petabyte per day [330]. This exceptional data rate is more than a factor of 1000 above the present state of the art, necessitating the rejection of certain examples already during data acquisition. The fact that these experiments are not going to consist of single telescopes, but of large-scale telescope arrays, poses outstanding challenges to the analysis of data obtained with these experiments and offers groundbreaking insight into the physical mechanisms that underlie the observed astrophysical sources. Resource constraints arise from the data rate, the inherent limitations in telescope communication, and the isolated telescope sites, like for example the geographic South Pole in case of IceCube.

During the first and second periods of CRC 876, a strong focus was put on the development of methods to solve challenging problems for specific astroparticle detectors (i.e., FACT, MAGIC, IceCube). After thorough tests, initial deployment at one experiment, and validation, these methods were generalised for the use in all operational detectors to which the CRC had access. That way it could be shown that the developed methods also provide appropriate solutions for the planned experiments CTA and IceCube-Gen2. Resource constrained multi-level analysis of high-dimensional and high-frequency data from astroparticle physics and astronomy has become the core capability of C3, and the expertise of the project covers the entire data analysis chain from data acquisition to the reconstruction of spectra.

In the third period, we intend not only to continue providing definite solutions in the construction phase of the experiments. Additionally, we intend to contribute to the solution of problems concerning communication, intelligent interaction, and the common data analysis chains of a multitude of flagship experiments. The extremely accurate application of Convolutional Neural Networks in IceCube will be extended to other use cases, and the new generation of telescope arrays will be

investigated using probabilistic graphical models. The cost-effectiveness of data modeling and simulation will be improved by combining the different approaches of the second funding period.

As an especially challenging but highly rewarding task, we will extend our efforts to include the data processing and analysis chains of the planned SKA radio telescope array, which will be by far the largest and most sensitive radio telescope ever constructed. The SKA will eventually have an effective area of about a square kilometer and facilitate not only unprecedented surveys of the radio sky to complement CTA and IceCube Gen-2, but also searches for low-frequency gravitational waves. It will deliver data at a rate of 3 TB per second already in its first phase, thus offering the greatest data processing challenge in physical sciences worldwide.

In summary, the outstanding success of data analyses will continue in the third phase of the project, providing the latest experiments with highly reliable and highly flexible analysis schemes. To exploit the full physics potential of the telescopes, this already becomes necessary during planning and construction phases. Contributions from the groups in CRC 876 towards solving these challenges have already been strongly welcomed by our national and international partners.

### 3.3 Project progress to date

#### 3.3.1 Report and current state of research

##### Data Aquisition and Real-Time Analysis

Cosmic rays were discovered by the Austrian physicist Viktor Hess in 1905, and although various theoretical models for their acceleration in astrophysical sources exist, experimental evidence for any of the models is missing. Insight into the aforementioned acceleration processes is expected from the simultaneous observation of gamma rays and neutrinos from one and the same source. These so-called multi-wavelength observations are completed by observations of other telescopes, e.g., in the X-ray or optical regime. The detection of rare physical events, e.g., an increased transient flux of gammas or neutrinos from an astrophysical source, thus needs to be detected in real time, in order to trigger so-called follow-up observations. A fast and reliable real-time data analysis is crucial to the success of multi-wavelength campaigns.

In order to meet the requirements of real-time analyses of current and future telescopes in Imaging Cherenkov astronomy, the development of the `streams` framework and the FACT-Tools library was systematically continued within the second funding phase. The resulting set of analysis tools enables the practitioner to realise the entire analysis chain of the FACT telescope in a real-time streaming environment. Due to the runtime efficiency of FACT-tools, the real-time requirements of the FACT telescope can be met on a single desktop computer.

With regard to future telescope arrays like the Cherenkov Telescope Array (CTA), where the data rate is expected to be on the order of 1 TB per day, the `streams` framework was transferred to the popular Big Data platforms Flink, Storm, and Spark [C3/A1/339]. All of these streaming platforms provide horizontal scalability. That is, as more machines are added to the computing cluster, the throughput of the cluster increases almost linearly with the number of additional machines. As running many machines with low individual computing power is more resource-efficient than running a single machine with the entire power, this property helps in reducing resource requirements. The student project group 594 embedded this finding in a comprehensive reference architecture for the analysis of telescope data, ranging from distributed storage to distributed feature extraction and classification, and also providing a web interface to search for specific observations. Building on this experience, a real-time analysis scheme for CTA was developed, which is the only operating prototype to date [C3/341].

As mentioned above, a certain amount of examples needs to be rejected on-site due to the limited communication resources. In a collaborative effort with project A1, FPGAs and embedded hardware, in combination with a suitable implementation of decision trees and Random Forests, were investigated on their on-site applicability. It was found that approximately 12% of all recorded events can be discarded on-site without removing any of the interesting gamma events [A1/C3/31]. This early and efficient rejection of irrelevant examples was further found to reduce the resource requirements in communication and attribute reconstruction. Utilising machine learning on embedded devices for on-site data filtering is an important step towards intelligent telescopes and is expected to greatly enhance the real-time analyses of future telescopes.

### **Signal-Background Separation**

In astroparticle physics, examples utilised in specific analyses (signal examples) are generally superimposed by irrelevant examples (background examples) that need to be rejected in an efficient and reliable manner. The ratio of signal and background examples ranges from  $10^{-2}$  to  $10^{-6}$ , depending on the analysis level and the target particle. An accurate means of recognising and rejecting background examples is thus a key component in every physics analysis. The work of the second funding phase focused on transferring the core concepts (attribute selection, tree-based ensemble learners, unfolding) of an analysis chain developed for muon neutrinos in IceCube [C3/336] to other types of particles and detectors. Transferring these concepts yielded excellent results and provided proof of the robustness and applicability of the utilised methods. Furthermore, the analysis scope of the project was extended to the reconstruction of particle properties with Convolutional Neural Networks [C3/342].

For the case of neutrino astronomy, the general robustness of machine learning-based analysis chains was proven by successfully applying the concepts originally developed for IceCube in its construction phase [C3/336] to different detector geometries [C3/338, C3/337]. The application of an analysis chain that is based on data mining to data taken with the IceCube detector in its 79-string configuration provided the first model-independent measurement of a flux of astrophysical neutrinos [C3/338]. The observed flux was found to be fully compatible with all corresponding measurements and cross-checks by the IceCube collaboration. Applying the analysis chain to analyses of electron neutrinos and cosmic ray muons provided further proof of the generality and applicability of the analysis concept. As simulated examples are required for the training of the classifiers, their performance on experimental data may decrease if the distributions of simulated attributes do not match the experimental ones. It was shown that these mismatches can be reliably detected and removed by training classifiers to distinguish between experimental and simulated examples.

While the analyses discussed above generally utilise examples that have already undergone a certain amount of processing, so-called high-level data, the work of C3 was further extended to the analysis of examples with a lower level of abstraction. These analyses include the search for  $\tau$ -neutrinos, which are expected to contribute to the flux of astrophysical neutrinos but have not been observed so far. The overwhelming number of  $10^{10}$  background examples expected for a single  $\tau$ -neutrino poses a severe challenge to this type of analysis. Despite its challenging character, the application of random forests to charge pulses obtained in IceCube is the most sensitive search for  $\tau$ -neutrinos in IceCube to date [C3/342]. A reliable detection of  $\tau$ -neutrinos can also be achieved by utilising Deep Neural Networks.

With respect to Imaging Cherenkov astronomy, FACT tools, a library based on the **streams** framework and capable of covering the entire analysis chain from preprocessing to the final prediction of examples, has become the standard platform environment for the FACT telescope. During the second funding phase, FACT tools was extended by adding modules for RapidMiner [334], MOA [325], and WEKA [328], which provide a wide range of machine learning algorithms. This exten-

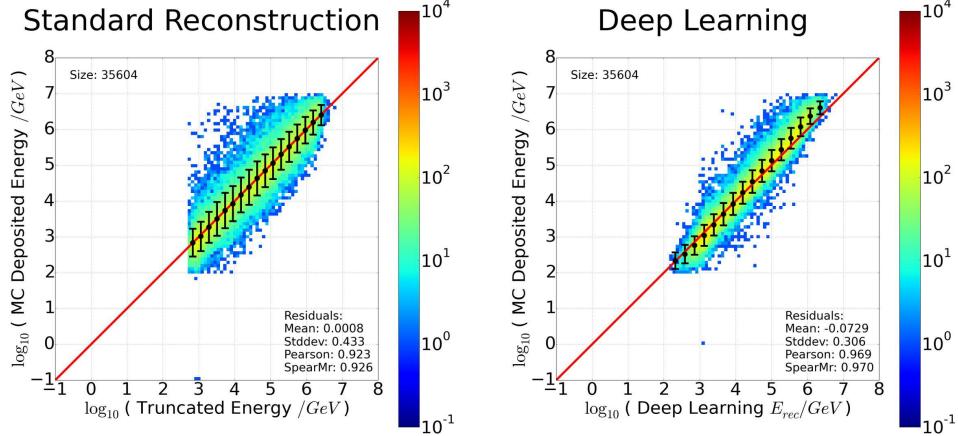


Figure 3.1: Reconstructed energy of atmospheric muons in IceCube obtained with standard reconstruction software (left) and from a Deep Neural Network (right).

sion allows for the estimation of particle types on continuously streamed data, directly obtained from the telescope. The FACT tools environment further allows for real-time processing of single examples directly on arrival. These advances of the platform received recognition with the best paper award in the industrial track of the ECML 2015 [C3/335]. The reliability and performance of the developed analysis concepts were further demonstrated by applying them to data obtained with the FACT telescope. It was shown that the general analysis concept can be applied in analyses focusing on the highly significant detection of gamma ray sources (PhD thesis of Julia Thaele) as well as in analyses concentrating on the high-resolution reconstruction of source spectra (PhD thesis of Fabian Temme).

In addition to the aforementioned investigations, the efforts of the project were extended from two- and multiclass-classification tasks to the reconstruction of particle attributes such as energy and direction. It was shown that the use of Deep Convolutional Neural Networks (CNNs) [327] for the energy reconstruction in IceCube increases the energy resolution of the detector by approximately 20% compared to the best existing energy estimator [C3/342] (see also Fig. 3.1). The reconstruction of particle properties with CNNs was found to require less than 10% of the runtime of existing reconstruction algorithms and is therefore extremely well suited for an on-site application. The aforementioned improvements were established using standard CNN implementations, which do not reflect the specific properties of the IceCube detector. It is expected that accounting for such properties, for example the geometry, will further enhance the performance of the CNNs in the third funding phase. A highly accurate estimation of the particle's energy is crucial for a reliable reconstruction of energy spectra. Therefore, improvements in energy estimation also enhance the accuracy of the reconstructed spectra.

### Spectral Reconstruction

The reconstruction of a distribution  $f(x)$  of a physical quantity  $x$  that cannot be measured directly is a common task in particle and astroparticle physics, as well as in other research areas like medicine. To solve these problems  $f(x)$  has to be obtained using experimentally accessible observables  $y_i$ , which depend on  $x$ . As the relationship between  $x$  and  $y_i$ , e.g., the production of a muon of energy  $E_\mu$  (experimentally accessible via an energy estimate) from a neutrino of energy  $E_\nu$  (experimentally inaccessible), is governed by stochastic processes, this task corresponds to an

inverse problem, generally described by the Fredholm integral equation of the first kind. Solving this integral equation is often referred to as unfolding or deconvolution, and several algorithms for the solution of such inverse problems exist. The algorithms typically applied in particle physics are, however, limited, for example in the number of input variables or in the sense that the information on individual events is lost in the unfolding process.

Within the first funding phase, C3 developed the Dortmund Spectrum Estimation Algorithm (DSEA), which aims at overcoming the limitations of existing algorithms. DSEA essentially translates the inverse problem of unfolding into a classification task, which is solved by a machine learning algorithm. The confidence distributions of individual examples, obtained from the classifier, are then interpreted as a probability density function and aggregated to reconstruct  $f(x)$ .

During the second phase, DSEA was extended such that the inverse problem is solved by using weights on the training examples, which are updated iteratively, depending on the unfolding result. This enables the use of training distributions that differ from the distribution of the test set by eliminating a potential bias towards the training distribution. The iterative use of DSEA thus reduces the assumptions required to obtain an accurate and reliable unfolding result, e.g., by commencing with a uniform distribution. The performance of DSEA was systematically investigated on artificial data as a function of various input parameters (smearing of the data, number of input attributes, number of bins, number of training examples). It was found that DSEA reliably reconstructs the spectrum, even for cases where the amount of smearing is so large that the shape of  $f(x)$  is no longer visible in the  $y_i$ . Too many iterations, however, may lead to a decrease in performance, and a suitable number of iterations must be picked by hand [C3/343].

This shortcoming of DSEA was further investigated, and it was shown that the algorithm smoothly converges to an optimal estimate if the influence of each iteration step is appropriately varied. In this case, the iterative process can be stopped smoothly when the reconstructed spectrum is close to its optimum. Several ways of varying the influence of the individual iterations were systematically investigated on artificial data, and it was found that the algorithm converges quickly and reliably towards the correct  $f(x)$  when the above-mentioned alterations are applied (master's thesis of Mirco Bunse). Our studies on this specific topic are ongoing. In addition to the studies of test problems, the algorithm was successfully applied to simulated examples from various experiments. The application of DSEA to examples from IceCube and FACT yields a highly accurate and reliable reconstruction of the underlying spectra. Furthermore, a close collaboration with project C5 was established in which the algorithm was successfully applied to simulated examples from LHCb.

Binned likelihood fits are a different, but equally effective approach to the solution of inverse problems where the actual binning of the observable space is crucial as it needs to be fine enough to account for the detector effects. A too fine binning, however, may result in sparse sampling of the observable space, which then requires an unnecessarily large number of input events. Detailed studies on the binning showed that the accuracy of the reconstructed spectrum can be greatly enhanced if the generally utilised equidistant binning is exchanged for a binning obtained from the use of decision trees (PhD thesis of Mathis Börner).

## **Data Modelling and Simulation**

Examples in astroparticle physics cannot be annotated by human experts. This is due to the extreme data rate (more than  $10^3$  events per second, after trigger), the requirement to be precisely aware of all selection probabilities, and the difficult decision situation, which requires the use of a large number of examples (up to  $10^6$ ) in the learning processes. The label of an experimentally obtained example is therefore inherently unknown. Therefore, the input for training and validation of learning algorithms has to be simulated. The utilised Monte Carlo simulations generally consist of three steps. In the first one, so-called generators sample particles of all desired or necessary types according to physically predefined distributions in energy and zenith angle. Each sampled

particle is then propagated through the detection medium to the actual detector in the second step by a so-called propagator. The third and final step involves simulating the detector response to a certain signal, initiated by particles of a specific type and energy, which includes a detailed description of the detector geometry and the entire read-out electronics. Whether an example will later provide useful information to a specific analysis task was ignored for a long time. Therefore, an overwhelming quantity of examples that do not provide useful information for the analysis are nevertheless generated and propagated. The expenses for computing in astroparticle physics (in the order of millions of euros per year in Germany alone) are still dominated by the generation of examples that do not help in any of the analysis steps.

One special example, regarding simulations for the FACT telescope, is that a fraction of approximately 80% of the generated events do not even contribute to the emission of Cherenkov light to be detected in the atmosphere. To avoid the propagation of such particles, an interface to the widely used simulation framework *CORSIKA* [329], which is commonly used for the simulation of atmospheric particles, was developed. This interface allows (in this application) to abort the simulation of particles that are unlikely to contribute to the Cherenkov light emission during the propagation process. The decision on the termination is based on the machine learning analysis of precedent cases (heuristics). In this application, the runtime of *CORSIKA* for the FACT telescope was reduced by  $\approx 70\%$  [C3/340] (see also master's thesis of D. Baack). Although *CORSIKA* was not developed in Collaborative Research Centre 876, the discussed interface (CRC 876 development) has been released as a regular part of the code in March 2017. Since *CORSIKA* is used in most experiments in astroparticle physics, this development is of great importance to the entire astroparticle physics community.

An approach to avoid the simulation of examples that carry little to no information on a certain classification problem is based on an iterative generation of examples, which evaluates the performance of a classifier. This approach follows the idea of *Active Class Selection* [333], where the examples are not sampled from a predefined distribution but according to sampling weights, which are updated after every iteration. To do so, the classifier's performance is evaluated as a function of one or more parameters of the simulation, e.g., energy and zenith angle. The weights of examples that improve the classifier's performance only marginally – or not at all – are decreased, whereas the weights of examples that carry useful information are increased accordingly. This idea was adapted during the second funding phase, and it allows a production, that focuses on highly relevant examples. Active Class Selection is therefore appropriate for the resource-saving simulation of relevant example sets.

In addition to an adequate performance within the simulation chain, particle propagation requires a very precise knowledge of the probability of possible interactions between the particle and the surrounding medium, as these interactions govern the information extracted for classification and property estimation. Knowledge of the interactions needs to be obtained from more and more advanced theoretical calculations and to be continuously included in the simulation. The *Propagator with Optimal Precision and Optimal Speed for All Leptons (PROPOSAL)*, was developed during the first funding phase and has been continuously improved. These improvements include the implementation of the signature class of the decaying  $\tau$ -leptons. This is required in searches for  $\tau$ -neutrinos, which have become an essential part of WP2 (signal-background-separation). Furthermore, *PROPOSAL* was extended for the propagation of the class of the so-called exotic particles, which have not been detected experimentally but are postulated in various physical theories, for example in Supersymmetry. Improvements in the close-to-reality simulation of different event classes directly translate to improvements in signal-background separation and spectral reconstruction.

## Bibliography

- [325] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. “MOA: Massive Online Analysis”. In: *Journal of Machine Learning Research* (2010) (cit. on p. 296).
- [326] CTA Consortium. “Introducing the CTA concept”. In: *Astroparticle Physics* 43 (Mar. 2013), pp. 3–18 (cit. on p. 294).
- [327] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016 (cit. on p. 297).
- [328] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. “The WEKA Data Mining Software: An Update”. In: *SIGKDD Explor. Newsl.* 11.1 (Nov. 2009), pp. 10–18 (cit. on p. 296).
- [329] D. Heck and J. Knapp. *EAS Simulation with CORSIKA: A User’s Manual*. Forschungszentrum Karlsruhe. <http://www-ik.fzk.de/corsika>, 2010 (cit. on p. 299).
- [330] C. Hollitt, M. Johnston-Hollitt, S. Dehghan, M. Frean, and T. Bulter-Yeoman. “An Overview of the SKA Science Analysis Pipeline”. In: *ArXiv e-prints* (Jan. 2016) (cit. on pp. 294, 301).
- [331] IceCube-Gen2 Collaboration. “IceCube-Gen2: A Vision for the Future of Neutrino Astronomy in Antarctica”. In: *ArXiv e-prints* (Dec. 2014) (cit. on p. 294).
- [332] H.-R. Klöckner, M. Kramer, H. Falcke, D. Schwarz, A. Eckart, G. Kauffmann, and A. Zensus, eds. *Pathway to the Square Kilometre Array – The German White Paper*. Max Planck Institute for Radio Astronomy, Dec. 2012 (cit. on p. 294).
- [333] R. Lomasky, C. E. Brodley, M. Aernecke, D. Walt, and M. Friedl. “Active Class Selection”. In: *European Conference on Machine Learning (ECML PKDD 2007)*. 2007, pp. 640–647 (cit. on p. 299).
- [334] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. “YALE: Rapid Prototyping for Complex Data Mining Tasks”. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*. ACM. New York, USA: ACM Press, Aug. 2006, pp. 935–940 (cit. on p. 296).
- [B3/240] **A. Saadallah, F. Finkeldey, K. Morik, and P. Wiederkehr.** “Stability prediction in milling processes using a simulation-based machine learning approach”. In: *51st CIRP conference on Manufacturing Systems*. Elsevier, 2018 (cit. on pp. 20, 221, 306).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [C3/335] **C. Bockermann, K. Brügge, J. Buß, A. Egorov, K. Morik, W. Rhode, and T. Ruhe.** “Online Analysis of High-Volume Data Streams in Astroparticle Physics”. In: *European Conference on Machine Learning (ECML PKDD 2015), Industrial Track*. Springer Berlin Heidelberg, 2015 (cit. on pp. 21, 297).
- [C3/336] **F. Clevermann, K. Frantzen, T. Fuchs, J. H. Köhne, N. Milke, K. Morik, D. Pieloth, W. Rhode, T. Ruhe, F. Scheriau, M. Schmitz, J. Ziermann**, and et al. “Development of a general analysis and unfolding scheme and its application to measure the energy spectrum of atmospheric neutrinos with IceCube”. In: *European Physical Journal C* 75 (Mar. 2015), p. 116 (cit. on p. 296).

- [C3/337] **M. Börner, W. Rhode, T. Ruhe, and K. Morik.** “Discovering Neutrinos through Data Analytics”. In: *European Conference on Machine Learning (ECML PKDD 2015)*. Springer Berlin Heidelberg, 2015 (cit. on p. 296).
- [C3/338] **M. Börner, T. Fuchs, M. Meier, T. Menne, D. Pieloth, W. Rhode, T. Ruhe, A. Sandrock, P. Schlunder,** and et al. “Measurement of the  $\nu_\mu$  energy spectrum with IceCube-79”. In: *European Physical Journal C* 77 (Oct. 2017), #692 (cit. on p. 296).
- [A1/C3/31] **S. Buschjäger** and **K. Morik.** “Decision Tree and Random Forest Implementations for Fast Filtering of Sensor Data”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 65-I.1 (Jan. 2018), pp. 209–222 (cit. on pp. 19, 69, 75, 296, 303).

### b) Other publications

- [C3/A1/339] **C. Bockermann.** “Mining Big Data Streams for Multiple Concepts”. Diss. TU Dortmund, 2015 (cit. on p. 295).
- [C3/340] **D. Baack.** *Data Reduction for CORSIKA*. Tech. rep. 2. E5b, Faculty Physic, TU Dortmund, June 2016 (cit. on p. 299).
- [C3/341] **K. Brügge, A. Egorov, C. Bockermann, K. Morik, and W. Rhode.** “Distributed Real-Time Data Stream Analysis for CTA”. In: *Astronomical Data Analysis Software and Systems (ADASS XXVII)*. 2017 (cit. on p. 295).
- [C3/342] **M. Hünnefeld.** “Deep Learning in Physics Exemplified by the Reconstruction of Muon-Neutrino Events in IceCube”. In: *International Cosmic Ray Conference (ICRC 2017)*. 2017 (cit. on pp. 296, 297, 303).
- [C3/343] **T. Ruhe, M. Börner, M. Wornowizki, T. Voigt, W. Rhode, and K. Morik.** “Mining for Spectra – The Dortmund Spectrum Estimation Algorithm”. In: *Astronomical Data Analysis Software and Systems (ADASS XXVI)*. 2016 (cit. on p. 298).

## 3.4 Project plan

### Goals

In the third funding phase, C3 will focus on the transition from state of the art (IceCube, FACT, MAGIC) to future large-scale experiments in astroparticle physics. These future experiments are currently being planned (SKA, IceCube-Gen2) or already under construction (CTA) and will take data with an unprecedented precision at an also unprecedented rate – approximately 1 TB per day for CTA and 1 PB per day for SKA [330]. Although the experiments utilise different detection techniques and intend to answer different scientific questions, the algorithmic data analysis steps are quite similar and can, as shown in the previous funding phase, be adopted between the experiments. The project plan for the third phase should proceed without delay, so that the pressing data analysis questions are answered before the next-generation telescopes are finished.

In contrast to the second phase, two of the three experiments do no longer consist of single telescopes, but of telescope arrays, distributed over a relatively large area. This results in distributed data, which in combination with the large data rate requires a framework for distributed data analysis on modern Big Data architectures. The expertise gained in C3 during the first and second funding phases will thus be consequently utilised to overcome the data analysis challenges posed

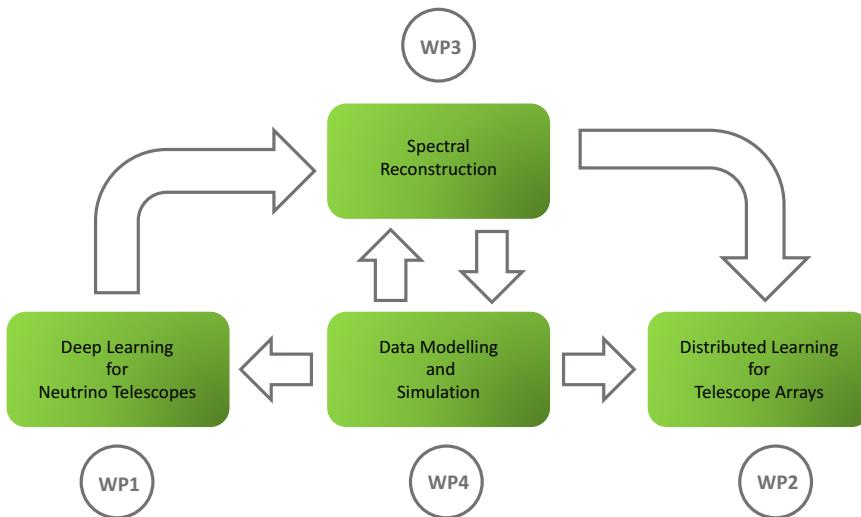


Figure 3.2: Work Packages.

by SKA, CTA, and IceCube-Gen2. A real-time-analysis with **streams**, the application of decision-tree-based analysis chains, the utilisation of Probabilistic Graphical Models and an estimation of physically relevant quantities with Convolutional Neural Networks are the key components of the project. We aim at providing fully optimised, highly flexible, and extendable data analysis concepts for CTA, SKA, and IceCube-Gen2 covering the entire analysis chain from data acquisition to spectral reconstruction.

### Work schedule

The work of C3 is subdivided into four closely connected work packages (see Fig. 3.2), which will be pursued in parallel. **Data Modelling and Simulation** (WP4) provides the common ground for work packages 1 to 3 because it supplies these packages with labelled data. **Spectral Reconstruction** (WP3) is generally the last step of an analysis. Enhanced sensitivities in the reconstruction of source spectra can be utilised to refine the simulation of examples from these sources (WP4) and provide valuable input to observation scheduling of telescope arrays (WP2). **Deep Learning for Neutrino Telescopes** (WP1) utilises Convolutional Neural Networks for the reconstruction of particle properties. Improvements in this work package directly affect and enhance work packages 3 and 4. **Distributed Learning for Telescope Arrays** (WP2) utilises the resemblance of the spatial telescope layout and modern Big Data architectures to enable distributed real-time processing of inherently incomplete examples.

**Work package 1. Deep Learning for Neutrino Telescopes** This work package analyses the properties of astrophysical neutrinos by enhancing Deep Convolutional Neural Networks to be applied to data of a relatively low level of abstraction. The foundation of this work was laid in phases 1 and 2 by the development of a set of well-tuned data analysis methods, which allow for the identification of an astrophysical neutrino flux from pre-analysed (high level of abstraction) data.

Neutrinos from astrophysical sources are now detected as a diffuse flux with a rate of about 20 examples per year. Individual neutrino sources – so-called point sources – have not been conclusively detected to date. Despite their lack of detection, neutrinos from point sources are expected

to provide groundbreaking insight into underlying physics of the sources, e.g., the acceleration of cosmic ray particles. This insight into the acceleration processes is also expected from the multi-wavelength observation of gamma rays and neutrinos from identical sources. In order to achieve such simultaneous observations, IceCube provides information on neutrinos of astrophysical origin to other instruments. For a successful follow-up observation, the neutrino's energy and arrival direction need to be reconstructed as accurately as possible. This task will be accessed using three independent approaches.

The first approach is to use Deep Convolutional Neural Networks (CNNs) to analyse data taken with the IceCube detector in its current state. CNNs have become the state of the art in image and video classification, where local invariances are exploited on an orthogonal grid. Data in astroparticle physics, however, is generally obtained by sensors arranged in a hexagonal coordinate system. Although the same invariances apply, there is no (fast) implementation of a hexagonally shaped kernel to date. Transforming the hexagonal grid to an orthogonal one [C3/342] results in loss of precision and computational efficiency. In order to exploit the full potential of CNNs in astroparticle physics, we will investigate existing convolutional kernels with respect to their applicability in astroparticle physics. In a second step, the gained knowledge will be leveraged in the development of new kernel functions that show the desired properties. Particular attention will be turned towards the applicability of CNNs and their respective kernels in real-time analyses. Improving the energy and directional reconstruction directly affects and improves spectral reconstruction (WP3) as well as the real-time analysis of astrophysical neutrino events (WP2). Furthermore, findings in this area can be readily applied to other applications where data is taken from irregularly spaced sensor arrays, e.g., sensor networks in road traffic or graphs of social networks. The excellent results obtained with CNNs in the second funding phase provide the basis for extending our research in this particular field.

The second approach treated in this work package concerns improvements in the real-time estimation of neutrino properties by implementing fully trained and optimised Deep Convolutional Neural Networks directly on FPGAs. The work of the second funding phase showed that the performance of learning algorithms with respect to the required computing time can be greatly enhanced by implementing the models on FPGAs [A1/C3/31]. We will therefore implement the models obtained from the fully optimised CNNs on FPGAs, which is expected to dramatically reduce the required computing time. The FPGAs itself will be implemented as close to the physical sensors as possible. Due to the fact that IceCube's optical modules are frozen into place in 1.5 kilometers below the ice surface and the IceCube computing facilities cannot be easily exchanged at the South Pole, this is, of course, barely possible for the present embedded system of the IceCube detector. For the next generation of large-scale neutrino telescopes, the hardware development has begun only recently. By developing an FPGA-based reconstruction of neutrino properties, C3 will become an active part of the embedded system development of IceCube-Gen2.

To date, South Pole ice, as a detection medium for the IceCube detector, is described by the scattering and absorption probability for Cherenkov photons, which both affect the propagation of light through the detector. As the South Pole ice was formed naturally, scattering and absorption are functions of the detector depths, which are modelled as discrete properties of 10 m chunks of ice and stored in lookup tables. Although there has been great progress in the modeling of the ice in the past, the coarse modeling in 10 m chunks is rather unsatisfying, especially with respect to the fact that the ice model is one of the dominating systematic uncertainties in all IceCube analyses, independent of the analysis topic. The third area of research within this working package will be the development of a self-calibrating method to model the detector properties, using the properties of South Pole ice as an example. Presently the detector properties are studied as a side branch of the general data analysis. Discrete descriptions of the ice are fed into the simulation and analysis algorithms. Our goal is the development of a method by which the main analysis data stream is used for a continuous self-calibration in space and time. Moreover, despite the fact that the modeling of scattering and absorption has been improved by a better description of the underlying physics, the analysis method – a parameter fit – remains essentially unchanged

since day one. In the third phase, we will obtain a better functional description of the South Pole ice by the systematic application and development of CNNs. Input data will be obtained from using the LED light sources installed on certain DOMs and subsets of the physics analyses well suited for detector calibration. In order to avoid the storage in lookup tables, we intend to obtain a continuous description of the ice properties. Advances in the ice model also enhance the agreement between simulated and experimental data, and therefore directly lead to refinement in work packages 1 to 3, which rely on simulated data.

**Work package 2. Distributed Learning for Telescope Arrays** CTA and the SKA will consist of up to 100 and a few thousand telescopes, respectively. This work package investigates how information obtained from the distributed telescopes in such an array can be analysed with novel machine learning techniques, and how the interaction of this information can be utilised to steer the telescopes in real time. As in the previous funding phase, one of the analysis keystones is the identification of physically relevant (signal) examples from the overwhelming number of irrelevant ones (background).

In addition to the anticipated data rate, further constraints of the analysis arise from the inherently limited communication between telescopes, which also confines the amount of information that can be merged in a given timeframe. Merging the information of *all* telescopes may not be possible at all. Therefore, a machine learning based analysis chain for modern telescope arrays has to be able to make decisions on the basis of inherently incomplete information. The spatial distribution of sensors in telescope arrays, however, is largely resembled by modern Big Data architectures. Thus, the goal of the third phase of the project is to develop a machine learning based analysis scheme that utilises these architectures to analyse data of a subset of a single or a few selected telescopes by means of distributed learning. Within the second funding phase, it was already shown that the `streams` framework can be used on Big Data architectures. The work of the third phase will consequently build on this experience and extend the efforts to meet the requirements of large-scale telescope arrays concerning flexibility, reliability, and scalability.

In order to develop the necessary analysis scheme, we, in a first step, will investigate what and how much information is required to reliably distinguish between signal and background events using incomplete data. In order to solve this task, the telescope arrays will be modelled using Probabilistic Graphical Models. This will enable us to model the signal of a large number of telescopes given the signal of only a few telescopes – or even a single telescope. This approach avoids merging the information of the entire array, and it enables querying data from those telescopes whose information is likely to improve the discrimination between signal and background examples. One particular difficulty in the approach, however, is that fragments of proton showers, when observed by only a few telescopes, may mimic a  $\gamma$ -event. An increased background rate is therefore expected, and studies on the rejection of this type of background are lacking. The overall goal of the third funding phase, to minimise communication and computational cost while maximising the performance of the sensor array, therefore involves efficient rejection fragmented showers with incomplete information.

An equally relevant goal of the third funding phase is the development of a real-time analysis that meets the requirements of CTA. The scope of real-time analyses is the detection of rare astrophysical events like source flares. Within a flare, the gamma ray flux of an astrophysical source may rise by more than an order of magnitude above its typical value within a relatively short timeframe. The total duration of a flare depends on the physical details of the source and typically extends from a few hours to a few days. The detection of flares in real time is of great physical importance, as the physics of the transition processes is poorly understood and may help to reveal the underlying physics of the source as a whole. The scientific intention is thus the observation of a flaring source for as long as possible with as many different instruments as possible, e.g., X-ray or neutrino telescopes. The largest challenge of a flare observation is the detection of

the rising edge of a flare. In the second funding phase, a prototype real-time analysis has been developed which, although it is the fastest real-time analysis to date, is still about a factor of 5 too slow for this purpose. This prototype will be systematically refined, utilising the experience gained in real-time analyses with the FACT telescope.

**Work package 3. Spectral Reconstruction by the Solution of Inverse Problems** The keystone to the extraction of physically meaningful parameters from a measurement is the solution of inherently ill-posed inverse problems. This work package investigates and improves the data mining-based algorithm DSEA, which approaches such problems.

Extracting knowledge from theoretical models for comparison to experimentally obtained spectra of astrophysical sources is a key component of research in astroparticle physics. The successful work on spectral reconstruction will therefore be continued systematically throughout the third funding phase. Building on the excellent results obtained with simulated examples, DSEA will be applied to experimental data obtained from the various experiments. Since the information on individual examples is retained within DSEA, it is expected that the application of DSEA on these data will reveal interesting physical results, for example about the internal structure of astrophysical sources.

For the observation of gamma sources with CTA, one relevant question is how the information from additional telescopes will enhance the accuracy and the resolution of the reconstructed spectrum. The second question is how many examples are required to obtain a certain spectral resolution. Both numbers can directly be converted to observation time required by individual telescopes, which is a valuable resource in Imaging Cherenkov Astronomy. Answering both questions will therefore provide highly relevant input to the planning of source observations with CTA. Since the number of examples expected from neutrino point sources is rather small, the question for IceCube-Gen2 and IceCube is how many examples are required in a stable reconstruction of source spectra from which certain source properties, for example the spectral index, can be reliably extracted.

The work on DSEA will also continue from an algorithmic point of view. DSEA delivers a solution to a discretised version of the Fredholm integral equation, where the discretisation of  $f(x)$ , which corresponds to the assignment of class labels, is somewhat arbitrary. This may, for example, lead to cases where examples located close to the upper boundary of class  $i$  are more similar to examples in bin  $i + 1$  than to examples at the lower boundary of bin  $i$ . Within the third funding phase, these neighbourhood relations will be exploited by assigning additional weights  $w_i$  to the training examples, which reflect the probability of an example to belong class  $i$ . These can, for example, be obtained by extrapolating kernel functions. Furthermore, the class labels in DSEA are ordinal but have to be treated as nominal, as most learning algorithms are incapable of handling ordinal classes. In the third funding phase, we will systematically investigate whether this challenge can be overcome by utilising a ground distance between classes in the classification.

The successful collaboration with C5 on the unfolding of data from LHCb will be continued and intensified. Spectral peaks with widths smaller than the experimental resolution of the particular experiments appear as  $\delta$ -peaks in the histogrammed version of  $f(x)$  and are generally very challenging for unfolding algorithms. The behaviour of DSEA with respect to such  $\delta$ -peaks has not been studied so far but will be investigated by utilising data from the decay of  $B^0$ -mesons obtained with LHCb. It is expected that the reconstructed peak in the spectrum becomes smaller with an increasing number of iterations. Compared to data from astroparticle physics, data from LHCb offers the advantage that the  $\delta$ -peak does not have to be constructed artificially, but appears *naturally* due to the physics governing the decay of  $B^0$ -mesons and due to the resolution of LHCb.

**Work package 4. Data Modeling and Simulation** In astroparticle physics, the most resource-intense computations arise from Monte Carlo simulations, which model the detector response to a signal induced by a primary particle. An accurate simulation of annotated examples provides the basis for all other work packages, but it requires an order of 5 million CPU hours per year within C3 alone. The simulation requirements are expected to increase with the next generation of experiments, due to the increased number of telescopes. This work package therefore focuses on the development of machine learning algorithms that help to reduce the computing resources required for simulation, making simulations for CTA and SKA computationally and economically feasible.

With respect to particle generation, we intend to suppress the calculation of Monte Carlo events carrying no additional information for the later data analysis. Building on the experiences gained with adaptive sampling approaches, we know that particle sampling can be controlled by the analysis task at hand. Proceeding this way avoids the computationally expensive simulation of examples that do not contribute information in the learning process or anywhere else in the analysis chain. As a consequence, processing and storage of irrelevant examples are avoided as well. The Adaptive Sampling Approach will be refined in order to further improve the performance of the learning algorithms. In an intended cooperation with project B3, we will exploit and enhance the concept developed in B3 [B3/240] in the context of astroparticle physics.

The very successful work on the *CORSIKA* extension will be continued in close collaboration with the Karlsruhe Institute of Technology (KIT). The continuation of our work includes the implementation of certain parts of *CORSIKA* on General-purpose Graphics Processing Units (GPGPUs). This is a non-trivial task, due to the complex code structure, which is, however, highly relevant with respect to the simulation requirements of CTA. In a first step, a GPGPU implementation of *CORSIKA* will focus on the production of Cherenkov photons. From our estimates, we expect a factor of three decrease in computing time. Other parts of the code will be investigated towards their suitability for a GPGPU computation and implemented as well for cases where significant increases in performance can be expected.

Within the second phase, the Adaptive Sampling approach and the *CORSIKA* extension were developed in parallel. To exploit the full potential of both concepts, the procedures will be combined and systematically tested towards their performance in the experiment-specific analysis chains. Extrapolating the performance of both concepts, we estimate a 90% decrease in simulation cost.

The work on *PROPOSAL* will continue throughout the third funding phase as well. Within *PROPOSAL*, numerous secondary particles – produced in a single interaction of a primary particle – are propagated through the detection medium. The individual computations carried out for the secondary particles are almost indifferent. This task is predestined for GPGPU computations. The same argument holds for the numerical integration. Within the second phase, the code was restructured, which now allows for a straightforward implementation of parts that are well suited for GPGPU computing. To date *PROPOSAL* is the standard code for particle propagation in IceCube, and we intend to fully implement the GPGPU version of the code into the IceCube simulation chain as well.

## Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Deep Learning for Neutrino Telescopes																	
2. Distributed Learning for Telescope Arrays																	
3. Spectral Reconstruction by the Solution of Inverse Problems																	
4. Data Modeling and Simulation																	

## 3.5 Role within the Collaborative Research Centre

The work of C3 focuses on information acquired at rates of up to 3000 events per second, using different types of sensors (telescopes). This information is digitised, preprocessed, and distributed on-site. Resource constraints directly arise from the remote locations of the experiments, e.g., the South Pole, which limits the bandwidth available for data transmission as well as the overall amount of energy available for data analysis. Indirect constraints affect the computing time and memory available at the various telescope sites.

Due to the large data rates generally acquired in astroparticle physics, C3 is one of the providers of *Big Data* within the Collaborative Research Centre. There are, thus, various prospects and possibilities for cooperation with other projects, with respect to the analysis methods used as well as from a more methodological point of view. For instance, C4 used the *CORSIKA* simulation data from C3 for checking the similarity between simulated and real data.

The successful cooperation of C3 with project A1 has been a key ingredient for the overall success of the project in phase 1 and 2. Analysis methods and algorithms have been adapted to the challenges of large-scale experiments in astroparticle physics. For astroparticle physics this resulted in a set of novel concepts that allow for a robust high-precision analysis of data at various levels of abstraction. From a Computer Science point of view, the development and application of data analysis algorithms in the context of astroparticle physics provide general proof of the applicability of the algorithms in a highly demanding environment.

The extraordinarily strong and productive cooperation with project A1 will therefore be continued in the third funding phase. This cooperation will expand the expertise gained in phase 1 and 2 towards the next generation of telescopes. The cooperative efforts will concentrate on the distributed analysis of incomplete telescope data, required for an effective data analysis for telescope arrays. The expected data rates of CTA and SKA further require a dedicated and cooperative effort of computer scientists and physicists in order to exploit the full scientific potential of the telescope arrays. The overall goal of the collaboration is the development of highly robust, highly accurate data analysis concepts that meet the requirements of large scale telescope arrays but can easily be modified for application in other research areas.

Moreover, the cooperation with A1 in the third funding phase will target the integration of fully optimised learning algorithms on embedded systems, e.g., FPGAs. It is expected that this approach

will reduce the resources required for data analysis. Furthermore, it is anticipated that the use of embedded systems allows for a better integration of machine learning algorithms in the telescope hardware, which is highly relevant for the hardware design of the upcoming telescope arrays. Within the collaboration of A1 and C3, we will carefully validate and evaluate the different options concerning a hardware integration of learning algorithms.

Another close collaboration was established with project C5. Within this cooperation the Dortmund Spectrum Estimation Algorithm (DSEA) was successfully applied to simulated data from the LHCb experiment. For the third funding phase we will continue and intensify the cooperation with C5, by applying DSEA to increasingly challenging sets of data from the LHCb detector. The overall goal of this collaboration is an improvement in the energy resolution of the LHCb experiment.

For the third funding phase, we will further establish a cooperations with projects A6 and B2. The collaboration with A6 will focus on the development of hexagonal kernels for Convolutional Neural Networks, whereas the cooperation with B2 will target the development of a model for the correction of optical artefacts for the PAMONO sensor. With respect to the development of Active Sampling, a collaboration with B3 will be established. We will apply the concept developed in B3 in the context of astroparticle physics and cooperate in order to further enhance the capabilities of this approach with respect to simulations in both fields.

### 3.6 Differentiation from other funded projects

#### **Variety, Veracity, Value: Handling the Multiplicity of Urban Sensors (VaVeL)**

**(Morik, Reference number Horizon2020-688380) (Funding period: 2016–2020)**

The goal of the VaVeL project is to advance our ability to use urban data in applications that can identify and address citizen needs and improve urban life. The motivation of this European project comes from problems in urban transportation, which are successfully approached with spatio-temporal random fields. In contrast, C3 is motivated by astrophysical use cases, which pose different problems to which other methods are applied.

#### **Modellierung von Themen und Strukturen religiöser online-Kommunikation**

**(Morik, Reference number MERCUR, PR-2015-0046) (Funding period: 2016–2018)**

The project addresses two main questions: What are the structures of religious communication in online contexts, and how do religious topics spread across these structures? One major challenge in answering these questions is to find low-rank representations of text data, which is available in social media contexts. The image data primarily analysed in C3 is handled quite differently from such textual data. Lukas Pfahler, the PhD student from this expiring project, has become a member of the CRC 876 who is financed by the university.

#### **Berechnung und experimentelle Analyse der Myon-Wirkungsquerschnitte**

**(Rhode, Reference number DFG, RH 35/9-1) (Funding period: 2017–2020)**

This project is focused on interaction cross-sections for the various interactions of muons in media, which contribute to the overall energy loss of a muon. The project first aims at performing highly accurate theoretical computations of the interaction cross section, by taking into account higher order effects. Second, the scope of the project extends towards a precise measurement of these cross-sections by utilising stopping muons in IceCube.

#### **Verbundprojekt IceCube**

**(Rhode, Reference number BMBF, 05A17PEA) (Funding period: 2017–2020)**

This project supports the construction and maintainance of the IceCube telescope, as well as the development of physical detection methods. Data taken by IceCube are used and distributed as data sets within the CRC 876.

**Kompetenzzentrum maschinelles Lernen Rhein Ruhr – ML2R  
(Morik, Reference number BMBF) (Funding period: 2018–2022)**

The German federal government has accepted four centres for machine learning, which have the double function of achieving scientific excellence and transferring results into practice. Of course, stimulating discussions between members of the CRC876 who work on machine learning and members of ML2R will be possible. It is planned that astrophysical data sets will be used for training in ML2R, because they have no privacy issues.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Staff								
Doctoral researchers, 100 %	3	193,500	3	193,500	3	193,500	3	193,500
Total	—	193,500	—	193,500	—	193,500	—	193,500
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
<b>Grand total</b>	<b>193,500</b>		<b>193,500</b>		<b>193,500</b>		<b>193,500</b>	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Katharina Morik, Prof. Dr., professor	Data mining	TU Dortmund	6	—	Existing funds
	2	Wolfgang Rhode, Prof. Dr., professor	Astroparticle physics	TU Dortmund	8	—	Existing funds
	3	Tim Ruhe, Dr., postdoctoral researcher	Astroparticle physics	TU Dortmund	19.92	—	Existing funds
	4	Dominik Baack, M.Sc., doctoral researcher	Astroparticle physics	TU Dortmund	19.92	—	Existing funds
	5	Sibylle Hess, doctoral researcher	Data mining	TU Dortmund	19.92	—	Existing funds
	6	Lena Linhoff, M.Sc., doctoral researcher	Astroparticle physics	TU Dortmund	19.92	—	Existing funds
	7	Johannes Werthebach, M.Sc., doctoral researcher	Astroparticle physics	TU Dortmund	19.92	—	Existing funds
Non-research staff	8	Matthias Domke, technician	—	TU Dortmund	4	—	Existing funds
	9	Andrea Teichmann, secretary	—	TU Dortmund	4	—	Existing funds
	10	Kai Warda, technician	—	TU Dortmund	4	—	Existing funds
<b>Requested staff</b>							
Research staff	11	Kai Brügge, M.Sc., doctoral researcher	Astroparticle physics	TU Dortmund	—	Doctoral researcher	—
	12	Mirko Bunse, doctoral researcher	Data mining	TU Dortmund	—	Doctoral researcher	—
	13	Alexander Harnisch, doctoral researcher	Astroparticle physics	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

**1. Morik, Katharina**

Project coordination.

**2. Rhode, Wolfgang**

Project coordination.

**3. Ruhe, Tim**

Project coordination. Further development of the Dortmund Spectrum Estimation Algorithm (DSEA). Application of DSEA to data from various experiments (IceCube, FACT, CTA, LHCb). Development of CNN-based real time analyses for IceCube.

**4. Baack, Dominik**

Development of an Active Learning controlled simulation and resource-efficient integration into the simulation chains of existing and future experiments.

**5. Hess, Sibylle**

Development and analysis of machine learning methods for classification of telescope data

**6. Linhoff, Lena**

Development of a real-time-analysis for CTA and combination with point source analyses in MAGIC.

**7. Werthebach, Johannes**

Reconstruction of energy spectra of atmospheric muons using multiple years of data from the IceCube experiment.

**8. Domke, Matthias**

Technical assistance all workpackages.

**9. Teichmann, Andrea**

Secretary

**10. Warda, Kai**

Technical assistance all workpackages.

**Job descriptions of staff for the proposed funding period (requested funds):**

**11. Brügge, Kai**

Development of a machine learning based real-time-analysis for CTA. Development of a general data analysis concept for CTA and SKA.

**12. Bunse, Mirko**

Further development of the Dortmund Spectrum Estimation Algorithm (DSEA). Development of a machine learning based analysis chain for telescope arrays.

**13. Harnisch, Alexander**

Development of a CNN-based real-time-analyis for IceCube and integration at the telescope-site. Ingeration of machine learning algorithms on FPGAs.

**3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from University	7,500	7,500	7,500	7,500
Sum of existing funds	7,500	7,500	7,500	7,500
Sum of requested funds	0	0	0	0

(All figures in euros)

**3.7.5 Requested funding for instrumentation for the new funding period**

This project does not request any funding for major research instrumentation.



### 3.1 General information about Project C4

### 3.1.1 Project title:

Regression approaches for large-scale high-dimensional data

### 3.1.2 Research area(s):

409-08 (Massively Parallel and Data-Intensive Systems), 409-01 (Theoretical Computer Science), 312-01 (Mathematics)

### 3.1.3 Principal investigator(s)

Ickstadt, Katja, Prof. Dr., 18.01.1965, German

Lehrstuhl für Mathematische Statistik und Biometrische Anwendungen,  
Fakultät Statistik, Technische Universität Dortmund  
Vogelpothsweg 87  
44227 Dortmund

Phone: 0231-755-3111

E-mail: katja.ickstadt@tu-dortmund.de

Sohler, Christian, Prof. Dr., 19.02.1973, German

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

Phone: 0231-755-6940

E-mail: christian.sohler@tu-dortmund.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

no       yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes	(x) no
2.	clinical trials	( ) yes	(x) no
3.	experiments involving vertebrates.	( ) yes	(x) no
4.	experiments involving recombinant DNA.	( ) yes	(x) no
5.	research involving human embryonic stem cells.	( ) yes	(x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes	(x) no

## 3.2 Summary

The main objective of project C4 is the development of highly efficient regression approaches. We want to make modern statistical regression methods scalable to very large and high-dimensional data sets and settings where computational resources are scarce.

We focus on algorithmic approaches that can be efficiently implemented in streaming as well as in distributed environments. In particular, we develop methods to aggregate data and to reduce the number of observations using, e.g., random linear projections and sampling, as well as methods to reduce the dimensionality of the underlying, possibly Bayesian, model classes.

Sketching and sampling methods for regression approaches on large-scale data are important areas of research with many interesting open questions. Although basic models are well studied, research on complex and modern statistical methods has just begun. We pursue the study of novel data reduction techniques for, e.g., Bayesian generalised linear models, and aim at the challenging objective of unifying their algorithmic treatment to provide blueprints for broad statistical settings.

Until 9th of August 2019 Christian Sohler will work as a visiting researcher at Google Zürich. His contribution to the project will start after that date.

## 3.3 Project progress to date

In the current funding period, the research in project C4 has focused on extending our streaming algorithms for frequentist and Bayesian regression to more complex modern statistical regression models. A common theme consists in developing data reduction techniques like sketching via random linear projections and coresets retaining the statistical information up to little distortion. Hereby we address resource restrictions like memory access, communication cost, and runtime.

Random linear projections can be maintained dynamically in a data stream, due to their linearity. Also, they have strong aggregation properties which allows to combine individual sketches – stemming from distributed data – to one single sketch for the entirety of the data. Both streaming and distributed computations are fundamental in the analysis of very large data sets.

Coresets are small, possibly weighted data sets that are designed to approximate an input data set with respect to a computational problem. In our project, we study coresets for regression problems, which are often subsets of the input data that are obtained via sampling techniques. The resulting reduced data sets enable better scaling properties and efficient computations via the established (classic) algorithms. In collaboration with project A2 we have published a technical survey on common techniques for obtaining coresets [A2/C4/64].

The highlights of the current funding period include coresets for some generalised linear models as well as graphical models developed in collaboration with projects A2 [A2/C4/66] and B4 [B4/C4/381]. We generalised our Bayesian linear regression models towards hierarchical priors and distributions based on  $\ell_p$ -spaces [C4/383]. We translated the Merge & Reduce principle to maintaining statistical summaries in the streaming model. We imposed structural constraints to handle ion mobility spectrometry (IMS) peak data from project TB1 and the Leibniz Institute for Analytical Sciences (ISAS) [C4/384]. We included multiple genomic data sources in a Bayesian Cox survival model and directly integrated variable selection via a stochastic search algorithm [C4/315]. This methodology is currently employed in an ongoing collaboration with project C1. Tackling former limitations to a small number of variables, we made logic regression approaches viable for genome-wide association studies and other high-dimensional data analysis tasks [C4/386].

### 3.3.1 Report and current state of research

In the following paragraphs we provide a detailed overview of the results achieved in the second funding period of project C4, for every work package.

**Work package 1: Streaming algorithms for generalised linear regression** Generalised linear models (GLMs) extend classical linear regression to more flexible classes of generating distributions. Usually one assumes that the realisations of the dependent variable are generated from a member of the exponential family of distributions, based on the independent observations. Well-known examples of such distributions include the multivariate normal, binomial, Poisson, and gamma distributions. The expectation of the dependent variable  $Y$  is connected to the linear predictor  $X\beta$  via a so-called link function  $h$ ,

$$h(\mathbb{E}(Y)) = X\beta,$$

where  $X$  is the data matrix and  $\beta$  is the unknown parameter vector.

There is extensive work on sampling methods for approximating regression problems including  $\ell_2$ -regression [360] and  $\ell_1$ -regression [355]. These were generalised to  $\ell_p$ -regression for all  $p \in [1, \infty)$  [380]. More recent works study sampling methods for  $M$ -estimators [354] and generalised linear models [45]. We continue this line of research in a joint work with project A2 on logistic regression [A2/C4/66].

**Logistic regression** is an instance of a GLM. The aim of logistic regression is to estimate the parameter  $\beta$  implicitly defining binomial distributions based on the observed data. An exemplary task would be to assess the impact and interactions of variables in predicting the probability of a patient suffering from a certain disease, based on her or his personal, physiological, and diagnostic data. This learning task is based on a fixed set of patient data  $X \in \mathbb{R}^{n \times d}$  and corresponding labels  $Y \in \{-1, +1\}^n$  indicating whether a patient is healthy or not. Folding the labels into the data we write  $Z_i = Y_i X_i$  for all  $1 \leq i \leq n$ .

Our first result in [A2/C4/66] shows the impossibility of compressing the data sublinearly in the input size, which holds in the worst case for any data reduction technique. To go around this limitation, we introduced a novel parameter that can be used to bound the complexity of compressing a data set  $Z$  for logistic regression. This parameter is defined by

$$\mu(Z) = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\|(Z\beta)^+\|_1}{\|(Z\beta)^-\|_1},$$

where  $(Z\beta)^+, (Z\beta)^-$  comprise only the positive and negative entries of  $Z\beta$ , respectively. We call a data set  $\mu$ -complex if it satisfies  $\mu(Z) \leq \mu$ . If the data is  $\mu$ -complex for a small, not necessarily constant,  $\mu$  then there exists a sampling and reweighting scheme based on the sensitivity framework of [365] that produces a  $(1 \pm \varepsilon)$ -coreset of sublinear size  $O(\varepsilon^{-2} \mu \sqrt{n} d^{3/2} \log^{O(1)}(und))$  with high probability. A more involved recursive sampling scheme produces a  $(1 \pm \varepsilon)$ -coreset of size  $O(\varepsilon^{-4} \mu^6 d^3 \log^{O(1)}(und))$ , which is beneficial if the data is well-behaved and the input size is particularly large. These are the first provably sublinear coresets for logistic regression.

The parameter  $\mu(Z)$  has an intuitive statistical interpretation and might be of independent interest as detailed in [A2/C4/66]. It is not uncommon in practice that  $\mu(Z)$  is small, since otherwise logistic regression exhibits methodological weaknesses.

Our experimental evaluation in [A2/C4/66] on real-world benchmark data shows that there is an efficient implementation based on QR-decomposition that is more accurate than uniform random sampling and state of the art heuristic approaches [45] while being competitive in terms of runtime.

**Poisson regression** is another instance of a GLM to deal with count variables [369, 379]. A prominent example within the CRC 876 can be found in project B4, where Poisson models are

used to predict the number of vehicles per minute passing sensors located in the motorway ring around the city of Cologne. The predictions for a single sensor location are made, based on the measurements at all other locations and the parameters learned from a Poisson regression model [B4/291, B4/376]. This can be formalised as a dependency network (DN) [364]. DNs are graphical models comprising a collection of GLMs, where each element of a set of  $d$  variables is regressed on all other variables. DNs have several interesting applications surveyed in a joint publication [B4/C4/381], like collaborative filtering, phylogenetic analysis, genetic analysis, network inference from sequencing data, and traffic as well as topic modelling.

In our work with project B4 [B4/C4/381], we have developed coresets for DNs. Assuming all GLMs in the DN to be ordinary linear regression models, we can subsample and reweight the input points as in [360] to construct a coreset. Maybe surprisingly, we do not need to construct a coreset for each of the  $d$  GLMs separately. We can exploit the common subspace structure of all GLMs to show that it is sufficient to construct one single coreset of size  $O(\varepsilon^{-2}d \log d)$ .

Turning the focus to Poisson GLMs, the situation is different. Again, we can show that in the worst case, any data reduction technique produces either a summary of linear size or fails to approximate the objective function to within a large superconstant factor [B4/C4/381]. Reviewing the statistical modelling for count data, we note that the Poisson lognormal model is a statistical relaxation of the ordinary Poisson model [379]. It introduces a connection to  $\ell_2$ -based linear regression, which we can exploit to show [B4/C4/381] that a reweighted sample of size  $O(\varepsilon^{-2}d \log^2 d)$  gives a good additive approximation to the consistent maximum likelihood estimator of this model.

Our experimental evaluation [B4/C4/381] shows that the sampling scheme outperforms uniform sampling for the normal GLMs. For the Poisson GLMs the result is not as remarkable and the log-likelihood approximation seems worse for large sample sizes at first glance. But as the sample size drops below 20% of the data, our method captures more structure of the data.

**Work package 2: Efficient Bayesian regression approaches** Bayesian regression does not assume a fixed *optimal* solution for a data set as in the frequentist case, but introduces a distribution over the parameter space. The *likelihood* function  $\mathcal{L}(Y|X, \beta)$  models the information that comes from the data. The prior distribution  $p_{\text{pre}}(\beta)$  models problem-specific prior knowledge. Our goal is now to explore and characterise the *posterior* distribution, which, as a consequence of Bayes' Theorem, is a compromise between the observed data situation and the prior knowledge that we assumed for the parameters

$$p_{\text{post}}(\beta|X, Y) \propto \mathcal{L}(Y|X, \beta) \cdot p_{\text{pre}}(\beta).$$

**Our work on random projections for Bayesian regression** [C4/53] extends previous work on frequentist  $\ell_2$ -regression [353] to the Bayesian setting. Certain types of random projections studied in theoretical computer science form a so-called  $\varepsilon$ -subspace embedding for  $\ell_2$ -spaces, which preserves the  $\ell_2$ -norm of all vectors in a linear subspace with little distortion. The idea is to employ an  $\varepsilon$ -subspace embedding  $\Pi$ , which is a linear map capable of compressing the data matrix  $[X, Y] \in \mathbb{R}^{n \times (d+1)}$  into a *sketch*  $[\Pi X, \Pi Y] \in \mathbb{R}^{k \times (d+1)}$  for  $k \in O(\text{poly}(d)/\varepsilon^2)$ , whose dimensions notably do not depend on  $n$ . Our main finding is that the results of a Bayesian analysis on the sketch and on the original data set are similar up to little distortion, depending on the approximation parameter  $\varepsilon$ :

$$\begin{array}{ccc} [X, Y] & \xrightarrow{\Pi} & [\Pi X, \Pi Y] \\ \downarrow & & \downarrow \\ \mathcal{L}(Y|X, \beta) \cdot p_{\text{pre}}(\beta) & \approx_{\varepsilon} & \mathcal{L}(\Pi Y|\Pi X, \beta) \cdot p_{\text{pre}}(\beta). \end{array}$$

We can quantify the approximation via the Wasserstein distance [C4/53]. This choice is especially appealing because it relates the distance of probability measures to properties in the  $\ell_2$ -space over

which they are defined. For normal distributions this comprises that their location parameter as well as their covariances are close to the original.

**Hierarchical regression models** offer an extension of the previous result to a broader class of prior distributions. They present a modern statistical approach that is especially useful when information on different levels is present, e.g., in a meta-analysis, where raw data is available for some studies, but only averages for the others [378]. A hierarchical model is given by

$$p_{\text{post}}(\beta, \theta | X, Y) \propto \mathcal{L}(Y|X, \beta) \cdot p_{\text{pre}}(\beta|\theta) \cdot p_{\text{hyper}}(\theta),$$

where the prior on  $\beta$  on the first level depends on a *hyperparameter*  $\theta$  that is again modelled via a *hyper-prior*  $p_{\text{hyper}}(\theta)$  on the second level of the hierarchy. Such models can be naturally extended to model arbitrary, deep or broad, hierarchies and to model numerous different populations.

**Generalised normal priors** are another modern statistical extension that we study [C4/383]. They result from generalising the inducing norm from  $\ell_2$  to  $\ell_p$  for  $p \in [1, \infty)$ . Their probability density function is given by

$$f(x) = \frac{p}{2\varsigma\Gamma(1/p)} \exp\left(-\frac{|x - \mu|^p}{\varsigma^p}\right),$$

where  $\mu$  is a location parameter and  $\varsigma$  is a scale parameter. The parameter  $p$  determines the shape and heaviness of the tails. Special cases include the normal distribution for  $p = 2$ , the Laplace distribution for  $p = 1$ , and the uniform distribution on  $[\mu - \varsigma, \mu + \varsigma]$  for  $p \rightarrow \infty$ . Generalised normal distributions have been suggested and employed as a robust alternative to model deviations from normality [367] and to model a Bayesian analogue of LASSO regression [372].

**Generalised normal likelihoods** for  $p \in [1, \infty)$  can be approximated in a similar way as in the case of normal distributions via novel subspace embedding techniques for  $\ell_p$  [380]. However, this is technically more challenging. One complication is that the embedding sizes are much larger for  $p > 2$  than for  $p \leq 2$ . The other problem is that the distortion is as large as  $O((d \log d)^{1/p})$  rather than  $(1 \pm \varepsilon)$ . We thus use the random projection only in a preprocessing step [C4/383], to obtain a so-called *well-conditioned basis*, which can be thought of as an  $\ell_p$ -analogue to an orthonormal basis for  $\ell_2$ . From this we can derive sampling probabilities such that by taking  $O(d^{2p+3} \log^2 d \log(1/\varepsilon)\varepsilon^{-2})$  reweighted random samples, we achieve the desired  $(1 \pm \varepsilon)$  distortion.

**Merge & Reduce** is an algorithmic technique from the theory of coresets and streaming algorithms [344]. We have transferred the Merge & Reduce principle to maintain statistical summaries in the streaming model. We can compute the necessary summaries for a regression model by analysing them blockwise and combining the summaries of each block in a structured way.

The efficiency of all our methods mentioned above is assessed comparatively via state of the art Markov Chain Monte Carlo (MCMC) sampling methods, which form the de facto standard in Bayesian regression analysis. For our extensions to hierarchical models and  $p$ -generalised normal distributions as prior or as likelihood [C4/383], subspace embeddings show very similar behaviour to the fundamental  $\ell_2$ -case investigated in [C4/53]: the posterior distributions based on the subspace embeddings approximate the posterior distributions on the full data set well. The location of the posterior distribution is not systematically influenced, and the variation is preserved by the reduction up to little distortion. Similar results were observed for our Merge & Reduce algorithm.

**Work package 3: Non-parametric regression under structural constraints** Shape constraints such as monotonicity or unimodality can be imposed as a form of prior information. Incorporating them in a model leads to a demanding inference task already for small numbers of variables. A univariate regression approach for the shape constraint of unimodality using spline regression was developed in the first funding period. We investigated the applicability of this approach

in clinical dose finding studies together with experts on clinical trials [373] and compared it to classical procedures. An overview of the efficiency of the considered approaches is given for different scenarios typically found in clinical phase II trials.

In addition, we extended the unimodal regression approach for detecting and modelling multiple modes [C4/384, C4/385]. A straightforward way to combine several unimodal regressions when there is no overlap between peaks is piecewise unimodal regression, which was successfully used to model dives of marine animals [368]. For the case of overlapping peaks, that is, when the curves describing the different modes convolve, we investigated different deconvolution models.

We employed an additive model to detect peaks in ion mobility spectrometry (IMS) data from project TB1 and to model possible convolutions of those peaks [C4/384]. The peaks of each IMS spectrum were described by several unimodal functions and were estimated with a backfitting algorithm. The additive model was fitted for several numbers of peaks, and model selection was performed to find a suitable number of peaks. The peaks were combined across the spectra with a heuristical procedure and were used as features for classification of the breath gas samples.

For the deconvolution of loading curves from astroparticle physics data as well as overlapping peaks in IMS spectra, we combined [C4/384, C4/385] the unimodal regression approach with deconvolution models from the literature based on the  $\ell_0$ -penalty [374]. Here, the  $\ell_0$ -penalty facilitates automatic estimation of the number of peaks and their locations. The original model was restricted to estimating one pointwise peak shape, valid for all peaks. The combined model can handle identically as well as diversely shaped peaks that are described by a continuous spline function.

**Work package 4: Incomplete dependent variable** In this work package we assume a setting where the data matrix  $X \in \mathbb{R}^{n \times d}$  is given, but the dependent variable  $Y \in \mathbb{R}^n$  is unknown. We can reveal single values of  $Y_i$  but this is an expensive or restricted resource. The question that we investigate is how to select a smallest possible subset of  $Y$  so as to achieve a  $(1 + \varepsilon)$ -approximation to the optimal  $\ell_2$ -regression error. This selection can be based on  $X$  only, or on information derived from a small preselection that hints to a better subsequent selection.

We compared the performance of existing approaches designed for related topics ranging from experimental design [370] to selection techniques based on statistical leverage scores [362], from the algorithmic theory of coresets for regression problems as well as to methods from the theory of coresets for clustering problems [363].

In an extensive simulation study [C4/382] two approaches provide the best results. One is a combination of  $k$ -means clustering and leverage scores such that in every cluster the observation with the highest leverage score is chosen for the subset. Our second approach is an algorithm inspired by [363], which consists of two steps that are repeated until the desired size of the subset is reached. First, the observation with the highest leverage score is chosen. Then its  $\lceil \frac{n}{k} - 1 \rceil$  nearest neighbours are made ineligible. Both approaches return a subset with  $k$  observations and choose the subset based on a mixture of exploration of the feature space and the observations' importance for the regression model.

**Work package 5: Feature selection** Logic regression is a method designed to find interactions of higher order between variables. The approach seeks for logical connectives of Boolean variables derived from the data  $X \in \mathbb{R}^{n \times d}$ , which can explain the dependent variable  $Y \in \mathbb{R}^n$ . The Boolean formulas are represented as logical trees. The algorithm starts by uniformly selecting a variable and using it as a leaf of a logical tree  $L(X)$ . New trees are formed from elements of the current population by making small local changes and connectives of existing trees. A new tree is accepted or rejected following a *Simulated Annealing* scheme. These steps are repeated for a fixed number of

iterations. Every logical tree thus forms a new variable  $L_i(X) \in \{0, 1\}^n$  by combining a selection of the original variables from the data matrix  $X$ .

Logic regression can be employed for classification tasks (one single tree) as well as for regression analyses. Performing a regression task consists of growing  $T$  logical trees  $L_1, \dots, L_T$  and combining them via a GLM with link function  $h$  to the model,

$$h(E(Y)) = \beta_0 + \sum_{i=1}^T \beta_i L_i(X).$$

We note that logic regression can handle more than only binary data. It can be applied to survival analyses or quantitative dependent variables by choosing the link function appropriately.

One exemplary application is the analysis of genetic single-nucleotide polymorphism (SNP) data, where the aim is to find groups of SNPs interacting in determining the status of a patient with respect to a disease. The current logic regression method does not scale well with high-dimensional data and is limited to a few thousand variables [352]. Modern SNP analyses like genome-wide association studies and exome analyses [C1/321] aim to study a much larger number of variables.

One proposal is to reduce the number of SNPs by deleting redundant SNPs sequentially from every gene according to an association criterion frequently used in genetics [359].

We improved the scalability of the algorithm by sampling a preselection of a subset of the SNPs, which can be considered good if it is a superset of the important SNPs. Our sampling procedure is again derived from subspace-preserving methods. In our work [C4/386], we have studied leverage scores for the preselection both on simulated and benchmark data sets. Theoretical guarantees that exist for  $\ell_2$ -based problems do not apply here, because the logical tree construction and the use of GLMs change the subspace structure substantially. For logic regression, especially the *cross leverage scores* of SNPs with the target variable indicating the disease status proved to be meaningful to detect the important SNPs. Employing the cross leverage scores for an importance sampling procedure as a preprocessing step makes a preselection of influential SNPs, either as main effects or as part of interactions between SNPs. The subsequent logic regression can operate more efficiently and can detect possible higher-order interactions between the SNPs in the selection [C4/386].

In [C4/315] we presented an integrated sampling procedure for variable selection for the Bayesian Cox survival model in the presence of multiple genomic data sources. These have been applied to genetic sequencing and copy number variation data from project C1. Due to very slow convergence rates, we still need to adapt the sampling process for use with these data.

## Bibliography

- [344] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. “Approximating extent measures of points”. In: *Journal of the ACM* 51.4 (2004), pp. 606–635 (cit. on p. 319).
- [345] S. Arora and R. Kannan. “Learning mixtures of separated nonspherical Gaussians”. In: *The Annals of Applied Probability* 15.1A (2005), pp. 69–92 (cit. on p. 329).
- [346] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. “The Hardness of Approximate Optima in Lattices, Codes, and Systems of Linear Equations”. In: *Journal of Computer and System Sciences* 54.2 (1997), pp. 317 –331 (cit. on p. 328).
- [347] M. Badoiu, S. Har-Peled, and P. Indyk. “Approximate clustering via core-sets”. In: *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*. 2002, pp. 250–257 (cit. on p. 329).

- [348] M. Balcan, B. Manthey, H. Röglin, and T. Roughgarden. “Analysis of Algorithms Beyond the Worst Case (Dagstuhl Seminar 14372)”. In: *Dagstuhl Reports* 4.9 (2014), pp. 30–49 (cit. on p. 326).
- [349] P. Berman and M. Karpinski. “Approximating minimum unsatisfiability of linear equations”. In: *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2002, pp. 514–516 (cit. on p. 328).
- [350] T. M. Chan. “Faster core-set constructions and data stream algorithms in fixed dimensions”. In: *Proceedings of the 20th Annual Symposium on Computational Geometry (SoCG)*. 2004, pp. 152–159 (cit. on pp. 326, 328).
- [351] B. Chazelle. “On the convex layers of a planar set”. In: *IEEE Transactions on Information Theory* 31.4 (1985), pp. 509–517 (cit. on p. 326).
- [352] C. C. Chen, H. Schwender, J. Keith, R. Nunkesser, K. Mengersen, and P. Macrossan. “Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.6 (2011), pp. 1580–1591 (cit. on p. 321).
- [353] K. L. Clarkson and D. P. Woodruff. “Numerical linear algebra in the streaming model”. In: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*. 2009, pp. 205–214 (cit. on pp. 318, 326, 327).
- [354] K. L. Clarkson and D. P. Woodruff. “Sketching for  $M$ -Estimators: A Unified Approach to Robust Regression”. In: *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2015, pp. 921–939 (cit. on pp. 317, 327).
- [355] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. “The Fast Cauchy Transform and Faster Robust Linear Regression”. In: *SIAM Journal on Computing* 45.3 (2016), pp. 763–810 (cit. on p. 317).
- [356] R. D. Cook. “Detection of Influential Observation in Linear Regression”. In: *Technometrics* 19.1 (1977), pp. 15–18 (cit. on p. 328).
- [357] V. Damerow and C. Sohler. “Extreme Points Under Random Noise”. In: *Proceedings of the 12th Annual European Symposium on Algorithms (ESA)*. 2004, pp. 264–274 (cit. on p. 327).
- [358] S. Dasgupta. “Learning Mixtures of Gaussians”. In: *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*. 1999, pp. 634–644.
- [359] I. Dinu, S. Mahasirimongkol, Q. Liu, H. Yanai, N. Sharaf Eldin, E. Kreiter, X. Wu, S. Jabbari, K. Tokunaga, et al. “SNP-SNP Interactions Discovered by Logic Regression Explain Crohn’s Disease Genetics”. In: *PLoS ONE* 7.10 (Oct. 2012), pp. 1–6 (cit. on p. 321).
- [360] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. “Relative-Error CUR Matrix Decompositions”. In: *SIAM Journal on Matrix Analysis and Applications* 30.2 (2008), pp. 844–881 (cit. on pp. 317, 318).
- [361] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. “Sampling algorithms for  $\ell_2$  regression and applications”. In: *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2006, pp. 1127–1136 (cit. on pp. 326, 328).
- [362] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. “Fast approximation of matrix coherence and statistical leverage”. In: *Journal of Machine Learning Research* 13 (2012), pp. 3475–3506 (cit. on pp. 320, 328).

- [363] D. Feldman, M. Faulkner, and A. Krause. “Scalable Training of Mixture Models via Coresets”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2011, pp. 2142–2150 (cit. on pp. 320, 329).
- [364] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. M. Kadie. “Dependency Networks for Collaborative Filtering and Data Visualization”. In: *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*. 2000, pp. 264–273 (cit. on pp. 318, 327).
- [45] J. H. Huggins, T. Campbell, and T. Broderick. “Coresets for Scalable Bayesian Logistic Regression”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2016, pp. 4080–4088 (cit. on pp. 88, 317, 327).
- [365] M. Langberg and L. J. Schulman. “Universal  $\epsilon$ -approximators for Integrals”. In: *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2010, pp. 598–607 (cit. on pp. 317, 325–327).
- [366] R. T. S. Laurent and R. D. Cook. “Leverage, local influence and curvature in nonlinear regression”. In: *Biometrika* 80.1 (1993), pp. 99–106 (cit. on p. 328).
- [367] F. Liang, C. Liu, and N. Wang. “A robust sequential Bayesian method for identification of differentially expressed genes”. In: *Statistica Sinica* (2007), pp. 571–597 (cit. on p. 319).
- [368] S. P. Luque and R. Fried. “Recursive filtering for zero offset correction of diving depth time series with GNU R package diveMove”. In: *PLoS One* 6.1 (2011), pp. 1–9 (cit. on p. 320).
- [369] P. McCullagh and J. Nelder. *Generalized linear models*. 2nd ed. Chapman and Hall/CRC, Boca Raton, 1989 (cit. on pp. 317, 326).
- [370] A. J. Miller and N.-K. Nguyen. “A Fedorov exchange algorithm for D-optimal design”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43.4 (1994), pp. 669–677 (cit. on p. 320).
- [371] J. Nelson and H. L. Nguyêñ. “Lower Bounds for Oblivious Subspace Embeddings”. In: *Proceedings of the 41st International Colloquium on Automata, Languages, and Programming (ICALP)*. 2014, pp. 883–894 (cit. on p. 327).
- [372] T. Park and G. Casella. “The Bayesian Lasso”. In: *Journal of the American Statistical Association* 103 (2008), pp. 681–686 (cit. on p. 319).
- [373] J. Rekowski, C. Köllmann, B. Bornkamp, K. Ickstadt, and A. Scherag. “Phase II Dose-Response Trials: A Simulation Study to Compare Analysis Method Performance under Design Considerations”. In: *Journal of Biopharmaceutical Statistics* (2017), pp. 1–17 (cit. on p. 320).
- [374] J. de Rooij and P. Eilers. “Deconvolution of pulse trains with the  $\ell_0$  penalty”. In: *Analytica Chimica Acta* 705.1-2 (2011), pp. 218–226 (cit. on p. 320).
- [375] A. J. Struyf and P. J. Rousseeuw. “Halfspace Depth and Regression Depth Characterize the Empirical Distribution”. In: *Journal of Multivariate Analysis* 69.1 (1999), pp. 135 –153 (cit. on pp. 326, 328).
- [C1/321] **A. Schramm**, J. Köster, Y. Assenov, K. Althoff, M. Peifer, E. Mahlow, A. Odersky, D. Beisser, C. Ernst, et al. “Mutational dynamics between primary and relapse neuroblastomas”. In: *Nature Genetics* 47.8 (Aug. 2015), pp. 872–877 (cit. on pp. 19, 271–275, 321).
- [B4/376] **F. Hadiji, A. Molina**, S. Natarajan, and **K. Kersting**. “Poisson Dependency Networks: Gradient Boosted Models for Multivariate Count Data”. In: *Machine Learning Journal (MLJ)* 100.2 (2015), pp. 477–507 (cit. on p. 318).

- [B4/291] **L. Habel, A. Molina, T. Zaksek, K. Kersting, and M. Schreckenberg.** “Traffic simulations with empirical data – How to replace missing traffic flows?” In: *Traffic and Granular Flow '15*. Ed. by V. L. Knoop and W. Daamen. Springer, May 2016, pp. 491–498 (cit. on pp. 245, 318).
- [377] J. W. Tukey. “Mathematics and the Picturing of Data”. In: *Proceedings of the International Congress of Mathematicians*. Vol. 2. 1974, pp. 523–532 (cit. on pp. 326, 328).
- [378] S. Weber, A. Gelman, D. Lee, M. Betancourt, A. Vehtari, and A. Racine-Poon. “Bayesian aggregation of average data: An application in drug development”. In: *Annals of Applied Statistics* (2018) (cit. on p. 319).
- [379] R. Winkelmann. *Econometric Analysis of Count Data*. 5th ed. Springer, 2008 (cit. on pp. 317, 318).
- [380] D. P. Woodruff and Q. Zhang. “Subspace Embeddings and  $\ell_p$ -Regression Using Exponential Random Variables”. In: *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*. 2013, pp. 546–567 (cit. on pp. 317, 319).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [B4/C4/381] **A. Molina, A. Munteanu, and K. Kersting.** “Core Dependency Networks”. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. 2018 (cit. on pp. 316, 318, 326, 327).
- [A2/C4/64] **A. Munteanu** and **C. Schwiegelshohn**. “Coresets - Methods and History: A Theoreticians Design Pattern for Approximation and Streaming Algorithms”. In: *KI - Künstliche Intelligenz* 32.1 (2018), pp. 37–53 (cit. on pp. 18, 94, 316).
- [C4/53] **L. N. Geppert, K. Ickstadt, A. Munteanu**, J. Quedenfeld, and **C. Sohler**. “Random projections for Bayesian regression”. In: *Statistics and Computing* 27.1 (2017), pp. 79–101 (cit. on pp. 90, 318, 319, 329).
- [C4/315] T. Treppmann, **K. Ickstadt**, and M. Zucknick. “Integration of multiple genomic data sources in a Bayesian Cox model for variable selection and prediction”. In: *Computational and Mathematical Methods in Medicine* Vol. 2017 (2017), pp. 1–19 (cit. on pp. 285, 287, 316, 321).

#### b) Other publications

- [C4/382] N. Lategahn. “Vergleich von Methoden zur Auswahl von Beobachtungen bei Regression mit fehlenden Y-Werten”. Master Thesis. TU Dortmund University, 2016 (cit. on p. 320).
- [C4/383] S. Müller. “Untersuchung von Regression auf eingebetteten Datensätzen unter Verwendung von verschiedenen Abstandsnormen und Penalisierungstermen”. Master Thesis. TU Dortmund University, 2016 (cit. on pp. 316, 319).
- [A2/C4/66] **A. Munteanu, C. Schwiegelshohn, C. Sohler**, and D. P. Woodruff. *On Coresets for Logistic Regression*. Tech. rep. arXiv:1805.08571 [cs.DS], 2018 (cit. on pp. 88, 316, 317, 326).
- [C4/384] **C. Köllmann**. “Unimodal spline regression and its use in various applications with single or multiple modes”. Dissertation. TU Dortmund, 2016 (cit. on pp. 316, 320).

- [C4/385] **C. Köllmann, K. Ickstadt**, and R. Fried. *Beyond unimodal regression: modelling multimodality with piecewise unimodal regression or deconvolution models.* Tech. rep. arXiv:1606.01666 [stat.AP], 2016 (cit. on p. 320).
- [C4/386] A. Wollenberg. “Reduktion hochdimensionaler Datensätze für die logische Regression unter Verwendung von Leverage Scores mit besonderer Berücksichtigung von SNP-Daten”. Master Thesis. TU Dortmund University, 2016 (cit. on pp. 316, 321).

## 3.4 Project plan

### Goals

The main goal of project C4 is the development of highly efficient approaches for modern regression analysis. In particular, we aim at obtaining a common understanding of sketching and sampling methods for regression problems. Our cornerstones to solve these difficult problems are algorithmic techniques, statistical modelling, and geometric relaxations.

Our main algorithmic approaches are sketching via random linear projections and coresets obtained via importance sampling, in particular exploiting the sensitivity framework [365]. These enable us to obtain approximations for complex statistical models in several cases. In other cases we can tailor the statistical models towards the algorithmic treatment while retaining their descriptive and inferential properties. Geometric relaxations provide insights into the structure of the data and interact with the sensitivity measures in a way that makes sublinear data reduction possible for the statistical models. All three cornerstones jointly enable us to develop unified solutions for broad classes of regression models. Our challenging objective is to unify these approaches for the framework of (Bayesian) generalised linear models (GLMs) and beyond.

The work schedule is subdivided into four work packages pursuing individual goals. The highlights include

- to develop a general and unified sampling algorithm dealing with all GLMs and to tackle new approaches for parallelisation and streaming (WP1),
- to develop a novel approximation scheme for Bayesian regression and to close the gap between necessary and sufficient conditions in our existing approaches (WP2),
- to find connections between different importance measures and to classify their strengths in identifying influential observations and describing compressibility aspects of the data (WP3),
- to design an efficient algorithm for regression on normal mixture models leading to an approximation of arbitrary continuous likelihoods (WP4).

### Work schedule

Christian Sohler will work as a visiting researcher at Google Zürich until 9th of August 2019. Therefore, his contribution to work packages 1 and 2 will start in August 2019 when he is back at TU Dortmund. During the first seven months we will mainly focus on statistical importance sampling for generalised linear models and transferring statistical approaches from the second funding period to the Bayesian setting. Work packages 3 and 4 are scheduled to start later following his return. For the first seven months we apply for one additional staff position, for the remainder of the project for two.

In the following we present our four work packages in detail.

**Work package 1. Maximum likelihood estimation in Generalised Linear Models** Recall that a generalised linear model (GLM) is given by

$$h(\mathbb{E}(Y)) = X\beta,$$

where the expectation of the dependent variable is connected to the linear predictor via a so-called link function  $h$ . Our goal is to find a unified approach for approximating the maximum likelihood estimate in GLMs.

Our lower bounds of  $k \in \Omega(n/\log n)$  [B4/C4/381, A2/C4/66] show that there exists no strongly sublinear data reduction with small multiplicative error for Poisson or logistic regression. To overcome these limitations it was necessary to make assumptions on the data or to relax the statistical model placing our work in the setting of *beyond-worst-case analysis* [348]. The hardness result carries over to the general problem of approximating GLMs, but the relaxation techniques do not apply in general and have to be tailored towards the specific model.

We aim at extending to generalised linear models over the exponential family of probability distributions with the aim of obtaining a unified result. The models' loss functions

$$\ell(\beta|X, Y) = \sum_i g(x_i\beta) - Y^T X\beta + C,$$

where  $C$  is a constant independent of  $\beta$ , equal their negative log-likelihood functions  $\ell(\beta|X, Y)$  and have a common form [369] that we want to exploit. The second, linear part can be approximated within the toolbox of randomised linear algebra [353, 361]. The remaining problem lies in the so-called cumulative function  $g$ , which is derived from the link function of the underlying generalised linear model.

Our goal is to design coresets for GLMs to approximate their loss functions within multiplicative  $(1 \pm \varepsilon)$  error for all  $\beta \in \mathbb{R}^d$ . The sensitivity framework of Langberg and Schulman [365] is a general approach for obtaining coresets for any function that is given as a sum of functions  $\sum g_i(\beta) = \sum g(x_i\beta)$ , similar to the loss function. The *sensitivity* of each  $g_i$  corresponds to its worst case importance for approximating the sum over all possible  $\beta$ . It is known that a sufficiently large sample proportional to the distribution given by the sensitivities yields a  $(1 \pm \varepsilon)$ -approximation for all  $\beta$ , thus forming a coreset. The sum of sensitivities is crucial for the number of samples  $k$  that is needed. Obtaining useful, i.e., not too weak bounds on the sensitivities is an important but often non-trivial question. The problem is that for different GLMs we have to deal with functions  $g$  that do not necessarily possess the properties of norms like symmetry, bounded growth, subadditivity, and scale invariance (homogeneity). A natural continuation is to identify common structural properties like monotonicity and convexity.

Our approach is to bound the sensitivities using as little of the functional structure as possible. For example, relying only on monotonicity we can relate the sensitivities to the so-called *convex layers* of a data set [351]. The outmost layer is the set of extreme points lying on the convex hull. The second layer comprises the set of extreme points on the convex hull of the remaining points after removing the first layer, etc. An important observation is that the sensitivity monotonously decreases from the outmost layer to the interior. A related statistical concept that can be used in a similar way to obtain bounds on the sensitivities is the Tukey depth of a point [377]. Moreover, it can hint to statistical interpretations [375] and has been studied independently in computational geometry [350]. The Tukey median of a point set is a point attaining the maximum Tukey depth in the set and provides a meaningful lower bound on the number of samples needed for our approximation.

Now, in the worst case for both measures, all points lie on the convex hull of the input and their sensitivity is thus high. Turning to an average-case perspective, the most important points on the

outer hull make up only a small fraction of the data. Imposing the concept of *smoothed analysis*, it is possible to bound the number of extreme points on the convex hull under a small Gaussian perturbation [357]. Continuing the analysis for the family of convex layers seems plausible, but we will have to deal with technical difficulties in applying the argument recursively, since the removal of the outer hull does not preserve the Gaussian noise of the remaining points.

Another challenge that arises when constructing coresets is that assumptions imposed on the data are not necessarily hereditary with respect to the subset relation. For that reason some aspects of distributed and streaming computation are out of reach. The coreset construction algorithms may not be applicable to arbitrary subsets any more so that we cannot obtain a coreset as the mere union of coresets for the subsets. Our goal is to develop adapted merging techniques for our coresets that allow for communication-efficient distributed computations as well as memory-efficient streaming algorithms with project A2. Another possibility especially suited for the streaming setting is to develop data-oblivious sketching techniques like the ones from [354] for  $M$ -estimators.

**Work package 2. Bayesian Generalised Linear Models** Our goal is to extend the sensitivity framework of [365] mentioned in WP1 to a Bayesian analogon. We aim at introducing the *Bayesian sensitivity* of observation  $i$  given by

$$\varsigma_i = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\ell(\beta|x_i, y_i) + \log(p_{\text{pre}}(\beta))/n}{\ell(\beta|X, Y) + \log(p_{\text{pre}}(\beta))},$$

where  $\ell(\beta|x_i, y_i)$  is the restriction of the negative log-likelihood to observation  $i$ . Imposing a (penalising) prior  $p_{\text{pre}}(\beta)$ , we aim at establishing that points, which attain a high sensitivity at extreme ends of the optimisation domain, could be penalised in such a way that the prior dominates their contribution. Since the prior is present in the objective function, this implies that their contribution tends to uniformity. In such a way we could relax hard subsampling problems to an extent that they become tractable in sublinear space but still retain sufficient problem-specific information to offer a meaningful data analysis.

Our aim is to extend and unify our frequentist results to their Bayesian analogon on the basis of this novel framework for approximating Bayesian GLMs. We intend to introduce a sampling-based approach to construct coresets as a preprocessing step before the computationally demanding analysis. The coresets are supposed to work for the Bayesian models employing almost arbitrary priors. We also intend to extend previous empirical evaluations to the Bayesian setting. Our earlier results on Gaussian dependency networks carry over to Bayesian inference via Gibbs-sampling [364, B4/C4/381]. However, it is unclear and a non-trivial question whether this extension works for Bayesian Poisson regression. There already exists an experimental evaluation of Bayesian logistic regression via MCMC-sampling [45]. It would be interesting to continue the comparison of our methods to theirs in the Bayesian setting and to see whether our method outperforms theirs as in the frequentist case. In addition, the dependency of the coreset size on our novel parameter  $\mu$  might change in the Bayesian setting, which can potentially complicate our data reduction.

Another interesting direction is to study the necessary conditions to obtain a certain error guarantee in Wasserstein distance in terms of the approximation parameter  $\varepsilon$ . For  $\ell_2$  regression, we have shown that we can reduce to a smaller sketch. However, the size of the sketch has to be  $\Theta(\varepsilon^{-2}d)$  in order to retain the guarantees that we used [353, 371]. We thus cannot improve the reduction relying on the existing properties. On the other hand,  $d$  is a lower bound since we need to preserve the rank of the data matrix in order to have any bounded approximation guarantee for GLMs. But preserving the rank only cannot be sufficient to actually bound the error since this does not yield any guarantee on the approximation. This leaves a gap of roughly  $\varepsilon^{-2}$  that we might exploit. The goal is a stronger data reduction with comparable error guarantees. To this end, a possible approach is to relax the requirement to a  $(1 + \varepsilon)$ -approximation only in expectation. If the sets

on which the approximation fails make up only little of the probability mass with respect to the measure given by the prior, then it might be possible to go below  $\Omega(\varepsilon^{-2}d)$ . At the same time, we would be able to bound the Wasserstein distance of the probability measures averaging over three regions. The first region contains large mass where the approximation is good, even better than  $(1 + \varepsilon)$ . The second region contains small mass where the approximation fails, but is still within some reasonable bound. The third region comprises points where the approximation fails completely, but only makes up a null set.

The development of efficient methods for modern regression analyses enables us to work closely together with projects employing such methods. We will continue working on reducing dimensionality and identifying important features or interactions of different types of genetic data. This will be conducted in cooperation with project C1 on special cases of GLMs, in particular logistic regression with  $\ell_2$  and  $\ell_1$  priors and Bayesian Cox regression in a survival context. Our techniques such as random linear sketches and sampling of features, developed in the second funding period for logic regression, will be required to improve the scalability of these procedures.

**Work package 3. Importance measures for observations** From the very beginning of project C4, our research has focused on random projections and sampling techniques to tackle resource restrictions like memory, time, and communication. Random projections still play a role when flexibility in data streaming is mandatory, and help speeding up preconditioning steps, like a QR-decomposition of the data matrix, needed to compute appropriate sampling probabilities. However, our research has shifted to non-uniform sampling techniques as a means to construct coresets. In this light, it is a highly interesting and non-trivial question, how to quantify the importance of single observations and what properties the different approaches exhibit.

The study of importance measures has a rich history in theoretical computer science, computational statistics, and computational geometry. One example are the *statistical leverage scores*, which are importance measures related to several problems in randomised linear algebra [362]. The  $\ell_2$  version has been studied as statistical influence in statistical regression analysis dating back to 1977 [356], and has led more recently to coresets via importance sampling methods for  $\ell_2$  regression in 2006 [361]. Other prominent examples are the Tukey depth of a point, which is the number of points contained in any closed halfspace containing the point, and the Tukey median, which is a point maximising the Tukey depth. These concepts have been studied in robust statistics [377, 375], computational geometry [350], as well as in computational complexity theory [346]. As we have outlined before in WP1, we intend to draw connections between the Tukey depth and the established sensitivity framework in theoretical computer science.

The main focus in this work package is to study the importance measures and their properties, with the goal of creating a landscape of relationships among them. Especially the sensitivity approach seems to be meaningful in many different domains of approximation theory and exploratory statistics. It coincides for example with the statistical leverage scores for norm functions and can be bounded by the convex layer depth or Tukey depth for generalised linear models. A problem might be that actually computing these scores is a hard problem. One possible remedy is the use of approximation algorithms [349]. For these generalised models, which in fact are nonlinear, it would be interesting to draw the connection to smooth and computationally more viable importance measures like Jacobian leverage [366]. Interestingly, these seem to be tractable within the toolbox of randomised linear algebra, particularly using sketching techniques based on random projections.

Eventually, the variable importance measures (VIMs) obtained from our work on logic regression have little theoretical support. Our subsampling approach based on the cross leverage scores as a preprocessing step proved to work well in practice. We therefore plan to compare these scores and

others based on sensitivity to the VIMs and pursue a theoretical corroboration on the scores that show the best correlation in an empirical evaluation.

**Work package 4. Mixtures of normal distributions** Our goal is to design an efficient algorithm for regression on normal mixture models. Mixtures of normal distributions form an important generalisation since they allow to approximate arbitrary continuous distributions given the number of components is large enough. If it is known which component every observation is allocated to, we can apply random projections of [C4/53] component-wise, thus solving the regression problem on the mixture. However, the allocation is usually unknown and a suitable number of mixture components may also be unknown.

We will combine coreset techniques for learning normal mixture distributions with our random projections. If such a combination is not possible, adaptive non-uniform sampling approaches for obtaining coresets [363] are an alternative. Current approaches assume an almost spherical structure or rely on a clear separation of the components [345]. Other approaches restrict the condition number  $\sigma_{\max}/\sigma_{\min}$  of the data matrices or assume  $\varepsilon \leq \sigma_i \leq 1/\varepsilon$  their singular values [363]. It is our aim to achieve an approximation based on as weak assumptions as possible.

When searching for such suitable algorithms, we will also examine their dependency on the number of mixture components. The dependency is of interest, because it influences the difficulty of obtaining a good approximation when the number of components changes. On the other hand this symbolises a trade-off when employing mixtures to approximate arbitrary continuous distributions. A higher number of components allows for a better approximation of the desired continuous distribution, but a higher number of components will presumably increase the target dimension of the embedded data set or the size of the coreset.

Once a suitable coreset is obtained, the standard procedure for learning the normal mixture model is employing the EM-algorithm to obtain a reasonably good result. A probabilistic allocation is given by the EM-algorithm from which the points may be assigned to the components and inserted into the corresponding sketches. We plan to investigate whether it is possible to obtain theoretical guarantees for this procedure. To this end additional information about the mixture components might be required. One approach is to assume an oracle guessing the allocation, which might be removed by exhaustively trying all allocations [347]. Another approach is to assume that a weak approximation of the mixture distribution is incorporated, e.g., via a Bayesian prior learned from previous experiments.

## Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Maximum likelihood estimation in Generalised Linear Models																	
2. Bayesian Generalised Linear Models																	
3. Importance measures for observations																	
4. Mixtures of normal distributions																	

### 3.5 Role within the Collaborative Research Centre

The challenge of solving regression problems on large-scale, high-dimensional data and in resource constrained settings is of high importance since they form the basis for many modern data analysis tasks studied in the CRC 876 and beyond. Our core research directions cover linear regression, generalised linear models, fitting complex physical measurements via splines or Gaussian distributions, and analysis of genetic data. These are central methods for several research goals of the CRC 876 in, e.g., projects TB1, B4, and C1.

We have a fruitful ongoing collaboration with project A2 on the design of streaming and distributed regression algorithms via sketches and coresets. In addition, project A2 studies energy aspects of our sketching methods.

Our findings on Bayesian regression models will help to reduce the dimension in Bayesian optimisation methods. Such methods are studied in project B4 and will be supported theoretically. We intend to continue the cooperation with project B4 and Kristian Kristian Kersting (TU Darmstadt) on coresets for deep graphical models from which the traffic prediction mechanisms can benefit.

We continue to cooperate with project C1 in finding important interactions of variables in high-dimensional genetic data sets via logistic and logic regression as well as on Cox survival analysis. Together with project TB1 we have modelled peak data via shape-constrained regression methods. We share our methodological knowledge on modelling count data and deconvolution approaches for energy spectra with project C3.

### 3.6 Differentiation from other funded projects

#### **Statistical Methods for Damage Processes Under Cyclic Load**

**(Ickstadt, Reference number CRC 823, Project B5) (Funding period: 07/2013–06/2021)**

This project is focused on the modelling and the analysis of processes for questions arising in civil engineering. These works do not overlap with project C4 with regards to content.

#### **Polymorphic Uncertainty Modelling for Stability Quantification of Fluid Saturated Soils and Earth Structures**

**(Ickstadt, Reference number DFG Project IC 5/6-1 in Priority Programme 1886) (Funding period: 01/2017–06/2020)**

This project models uncertainty on different levels for soil conditions arising in civil engineering. There is no overlap with project C4 with regard to content.

#### **ERC Starting Grant**

**(Sohler, Reference number SUBLINEAR 307696) (Funding period: 2012–2018)**

The project deals with the analysis of property testing algorithms for sparse graphs and is not related to this proposal.

#### **LPN-Krypt: Das LPN-Problem in der Kryptografie**

**(Sohler, Reference number Pr-2016-0045) (Funding period: 2017–2019)**

The project studies the learning-parity-with-noise (LPN) problem and its variants from the perspective of (post-quantum) cryptography. It is not related to this proposal.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2011.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	2	91,400	2	129,000	2	129,000	2	129,000
Total	—	91,400	—	129,000	—	129,000	—	129,000
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	91,400		129,000		129,000		129,000	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Katja Ickstadt, Prof. Dr., professor	Statistics	TU Dortmund	5	—	Existing funds
	2	Christian Sohler, Prof. Dr., professor	Algorithmic complexity theory	TU Dortmund	5	—	Existing funds
	3	Julian Riehl, M.Sc., doctoral researcher	Statistics	TU Dortmund	22.28	—	Existing funds
	4	N.N., student assistant	Algorithmic complexity theory	TU Dortmund	8	—	Existing funds
Non-research staff	5	Eva Brune, secretary	—	TU Dortmund	1	—	Existing funds
	6	Alla Stankjawitschene, secretary	—	TU Dortmund	1	—	Existing funds
<b>Requested staff</b>							
Research staff	7	N.N., doctoral researcher	Algorithmic complexity theory	TU Dortmund	—	Doctoral researcher	—
	8	Leo N. Geppert, Dipl.-Stat., doctoral researcher	Statistics	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):****1. Ickstadt, Katja**

Project management. Focus on Bayesian statistics. Cooperation in all WPs.

**2. Sohler, Christian**

Project management. Focus on efficient algorithms. Cooperation in all WPs. Until 9th of August 2019 Christian Sohler will work as a visiting researcher at Google Zürich. His contribution to the project will start after that date.

**3. Riehl, Julian**

Research and development with a focus on (Bayesian) generalised linear models and Cox regression.

**4. N.N.**

Implementation and assistance in evaluation of algorithms.

**5. Brune, Eva**

Secretary.

**6. Stankjawitschene, Alla**

Secretary.

**Job descriptions of staff for the proposed funding period (requested funds):****7. N.N.**

Research and development in all WPs with a focus on efficient algorithms.

**8. Geppert, Leo N.**

Research and development in all WPs with a focus on Bayesian statistics.

**3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from university	7,500	7,500	7,500	7,500
Sum of existing funds	7,500	7,500	7,500	7,500
Sum of requested funds	0	0	0	0

(All figures in euros)

**3.7.5 Requested funding for instrumentation for the new funding period**

This project does not request any funding for major research instrumentation.



### 3.1 General information about Project C5

### 3.1.1 Project title:

# Real-Time Analysis and Storage of High-Volume Data in Particle Physics

### 3.1.2 Research area(s):

409-06 (Information Systems), 309-01 (Particles)

### 3.1.3 Principal investigator(s)

Spaan, Bernhard, Prof. Dr., 25.04.1960, German

Experimentelle Physik 5, Fakultät Physik, Technische Universität Dortmund

Otto-Hahn-Straße 4  
44227 Dortmund

Phone: 0231-755-3662

E-mail: Bernhard.Spaan@tu-dortmund.de

Teubner, Jens, Prof. Dr., 22.03.1976, German

LS 6, Fakultät für Informatik, Technische Universität Dortmund  
Otto-Hahn-Straße 14  
44227 Dortmund

Phone: 0231 755

E-mail: jens.teubner@tu-dortmund.de

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

( ) no (x) yes

Do any of the above mentioned persons hold fixed-term positions?

(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

### 3.1.4 Legal issues

This project includes

1.	research on human subjects or human material.	( ) yes	(x) no
2.	clinical trials	( ) yes	(x) no
3.	experiments involving vertebrates.	( ) yes	(x) no
4.	experiments involving recombinant DNA.	( ) yes	(x) no
5.	research involving human embryonic stem cells.	( ) yes	(x) no
6.	research concerning the Convention on Biological Diversity.	( ) yes	(x) no

## 3.2 Summary

The *Large Hadron Collider (LHC)* at CERN is well known for the massive amounts of physical data it produces. In this project, we collaborate with physicists from the *LHCb* collaboration and address the data processing needs that arise there.

The main objective of the C5 project is to develop new methods to realise the data processing tasks within *LHCb*. We also strive to adapt existing methods to understand whether/how they would apply in the context of *LHCb*. A very important aspect of the project is to evaluate methods in an environment like the one at CERN; specifically, we are only interested in solutions that will carry to a very large scale (such as thousands of compute nodes). The other angle of the project is to connect the application domain with methodology established within the CRC; examples are the use of *compact data summaries* (e.g., to avoid scanning huge amounts of data), *streaming algorithms* (e.g., the “trigger stage” at *LHCb* has to ingest a high-volume data stream and react with real-time characteristics), or the consideration of *low-power platforms*.

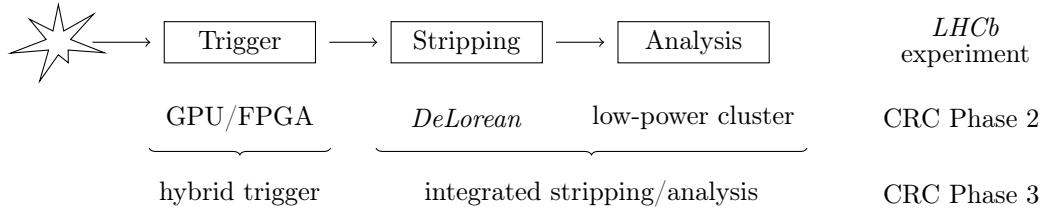
The project entered the CRC at phase 2. During that phase, we developed solutions to individual, isolated problems within *LHCb*. We proposed, e.g., new GPU-based algorithms; a novel filter mechanism (named *DeLorean*); and parallelisation mechanisms for *LHCb* analyses (including a tailor-made low-power hardware platform). Our umbrella goal for phase 3 is to connect those solutions towards a new design for the overall processing pipeline at *LHCb*. To this end, we want to, e.g., integrate our parallel data selection engine *DeLorean* with a Map-Reduce-based analysis framework and extend both parts by GPU acceleration.

## 3.3 Project progress to date

The *LHCb* experiment is one of the four large-scale physical experiments that use the *Large Hadron Collider (LHC)*, the world’s largest particle accelerator. *LHCb* has been designed to investigate the asymmetry between matter and antimatter or, more simply put, to answer the question *Why is there more matter than antimatter in the universe?* As such, *LHCb* aims for nothing less than addressing the central question of our existence. It turns out that the standard model of particle physics is incomplete, as it fails to answer this and other fundamental questions like what is the nature of dark matter. *LHCb* tests the standard model of particle physics with precision in order to find hints for so-called new physics effects.

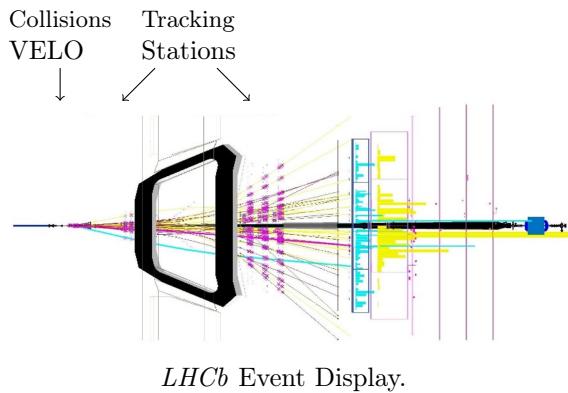
To answer their central questions, particle physicists are typically interested in *rare* decay events, and the key challenge in processing the *LHCb* data sets is to find those rare events within huge data volumes quickly and efficiently. The term “rare,” thereby, has to be taken very literally. The most interesting decay chains happen at most once every  $10^{12} \sim 10^{15}$  proton-proton collisions. In 2015, for example, a decay mode of highest relevance to the search for new physics  $B_s^0 \rightarrow \mu^+ \mu^-$  was observed in an analysis combining the data from the CMS and *LHCb* experiments [391]. It was found to occur approximately three times in 10 billion events.

In the experiment, protons are accelerated to near-light speed, then made to collide. In that situation, new particles form up – only to *decay* a short moment after. A collection of *detectors* is set up to sense the remains of such a *decay chain* and derive information about the particles that had formed. The amounts of data produced by the *LHCb* detectors are massive: 30 million collisions are initiated every second when the experiment is running, and about 100 kilobytes of data are collected for every collision (adding up to a raw data rate of about 3 TB/s). Obviously, it is impossible to store all this data. Thus, the purpose of a trigger is to analyse and filter the data in real time. Resource constraints limit the amount of data stored to disk to an event rate of currently a few kHz, resulting in a data volume of  $\approx 0.6$  GB per second. A first reduction is

Figure 3.1: *LHCb* processing pipeline and summarised contributions.

achieved by a hardware trigger, resulting in a event rate of 1 MHz. These events are processed on a high-performance compute farm, the so-called high-level trigger.

**The LHCb Experiment.** Per collision, hundreds of charged and neutral particles are produced. From their signatures in the detector, their tracks and their physical properties are reconstructed. The tracking trajectories of charged particles pass through various tracking detectors, called Vertex Locator (VELO), TT, inner and outer tracker. They are bent by a magnetic field to allow for a determination of the particle momenta. Two ring imaging Cherenkov detectors (RICH) are used to identify the type of particle. The calorimeters determine the energy deposition of particles. Finally, a muon system identifies muons. In the focus of the measurements are decays of hadrons containing a *b* quark or a *c* quark. The image to the right depicts an event display. In the coming upgrade, the tracking system will be replaced by an upgrade VELO, an Upstream Tracker (UT) for the TT, and a Scintillating Fibre Tracker (SciFi) replacing inner and outer tracker.

*LHCb* Event Display.

**State of the Art.** To cope with the massive data volumes, the physicists of the LHCb collaboration have built up a processing pipeline that can be simplified to three processing steps, illustrated in figure 3.1 (top part). *First*, a *trigger* system *pre-analyses* and *pre-filters* the raw data stream, discarding up to 99.95 % of all data right after acquisition to keep data volumes manageable. *Second*, the so-called *stripping* stage categorises every event into about 200 *stripping lines*. Each stripping line represents the precomputed result of one hard-coded event selection query. *Finally*, individual physicists run their respective *analysis* programs; a typical analysis program reads in one or more stripping lines and refines their predicate(s) to obtain the few collision events the physicist actually looks for.

Beginning in 2019, the experiment will be upgraded substantially [393] with several new subdetectors, among them a new tracking system based on scintillating fibres as sensor [394]. For C5, the upgrades of the data acquisition system and the high-level trigger are of highest relevance. The experiment will operate in a “triggerless” mode, where events from every collision are processed on a high-performance compute farm (that is, the trigger stage is no longer supported by a hardware-based component) [395]. Without dedicated hardware support, the software components will have to cope with the full 30 million events per second. Obviously, it is impossible to store

all this data. Thus, the purpose of the high-level trigger is to analyse and filter the data in real time. Resource constraints limit the amount of data that can be stored to disk to an event rate of  $\approx 20\text{ kHz}$  (corresponding to a data volume of about 2–5 GB per second). A new computing and software concept attempts to overcome this limitation [396].

The quality of the physics potential of the experiment relates directly to the performance of the trigger system. Since only a small amount of events will survive the filtering decisions, it is mandatory to find the proper events with highest precision, and to suppress background as much as possible. The decisions of the high-level trigger farm will be based on various properties of the events. Thus, in a first step, relatively simple and fast algorithms will be employed for a first filtering step. Those events surviving this step will be analysed in more detail, eventually leading to CPU-intensive full reconstruction of all tracks.

There are obvious resource constraints. To make matters worse, in the original planning of the experiment, physicists expected the available compute performance to further improve exponentially (in line with *Moore’s Law*). That trend essentially stopped meanwhile, which further increases the resource pressure in the *LHCb* project.

In conclusion, the three-stage model is hitting serious scalability limitations. The *trigger* stage faces the trade-off between extremely high throughput rates and filter accuracy; to make matters worse, any erroneously discarded data set is irrecoverably lost. *Stripping* avoids sifting through petabytes of data for every analysis; the down side is that *only* precomputed selection criteria can be used as input to an analysis. In spite of the massive compute facilities available at CERN, *analysis* tasks are strongly compute-bound – any improvement in efficiency here will directly translate into better insights on the physics side.

### 3.3.1 Report and current state of research

During the last CRC phase (this project joined the CRC in phase 2), we made contributions to all three stages of the processing pipeline (indicated in the middle of figure 3.1).

**Trigger Stage.** A key challenge in the trigger system of *LHCb* is to sustain the high data rate coming from the experiment. This data rate will significantly increase with coming LHC upgrades and, in fact, physicists would love to increase the data rate further (by increasing the experiment’s *luminosity* configuration) but are currently limited by the ingestion rate of the current trigger, where the simple hardware assisted trigger that limits the rate to 1 MHz becomes increasingly inefficient with higher luminosity. As mentioned above, after 2020, the experiment will run without hardware triggers, which cures this problem but results in an even higher input rate for the software-based triggers.

*Distributed Processing.* The demand to process high-volume data streams in real time resembles the pattern of existing stream-based applications. For instance, the back-end systems at *Twitter* need to deal with a high-volume stream of messages (based on *Twitter’s Storm* technology). Likewise, the *Streams* framework developed in the context of the A1 and C3 projects [A1/C3/398] is prepared to handle massive data streams. During the running phase 2 of the CRC, therefore, we adapted methods from *Storm* and *Streams* to deal with the characteristics of *LHCb* data. The outcome is *ELPACO* [C5/407], a distributed platform to process scientific data streams. A key characteristic of *ELPACO* is its ability to scale linearly with the amount of resources available (i.e., throughput vs. core count) – an important prerequisite to scale to the data volumes expected for *LHCb*.

*Low-Power Hardware.* In addition, *ELPACO* is prepared to run on top of *Eriador*, a novel, ARM-based cluster system that we originally designed to host the *analysis* stage of *LHCb* (more details

later). As we showed in [C5/407], the *ELPACO/Eriador* combination offers a  $2.5 \times$  advantage in energy efficiency, compared to existing Intel-based solutions as they are currently used at CERN.

*Acceleration with FPGAs and GPUs.* In the area of performance enhancement of the trigger by using FPGAs and GPUs, FPGA-based options were initially considered. The initial aim was to use the FPGAs available in the experiment to enable a fast search for clusters of the SciFi tracker across different tracking layers. However, it turned out that the standard event processing demanded increasing capacities of the FPGAs, which left hardly any resources on the available FPGA for the planned project. Therefore, a GPU-based approach was followed.

Since tracking in the trigger requires the most resources, the development of GPU-based tracking software was started. A first version has been developed based on a *hybrid seeding* algorithm. The algorithm searches for tracks using all hits of the 12 detector layers of the SciFi tracker, taking into account the possibilities of noise hits, inefficiencies, and other uncertainties. The tracker consists of three stations, each with four layers with different orientations. The fibres of the two so-called x-layers are oriented perpendicular to the ground, whereas the u- and v-layers are so-called stereo layers with an angle of  $\pm 5^\circ$  with respect to the x-layers. Starting from a hit in a x-layer of the first station, hits in a x-layer of the third station are searched for within a given search window. Based on hits in both layers, the intermediate station (x-layer) is analysed by means of a simple track fit. A track is then made with the association of the hits in the stereo layers. Once a track is found, its associated track hits will not be used in the search for other tracks.

The algorithms starts with a relatively tight search window to find the high momentum tracks first. The search window is subsequently enlarged to identify the tracks with lower momenta, too. Although this approach was successful (Masterthesis J. Surmann), the simple fitting algorithms were identified to become a bottleneck. Thus, alternative approaches to optimise the performance were investigated, taking into account the experience gained with the hybrid seeding algorithm. The approach is now extended to a track-finding algorithm based on so-called *microtracks*, where small tracks elements are identified in a single tracking station, first in the x-layers of the tracking station. The dipole magnet in front of the SciFi tracker bends the charged tracks in one plane, whereas in the perpendicular view most of the tracks can be considered to have straight trajectories, resulting in less demanding computation needs. Until the end of phase 2, the latter approach should be finalised, i.e., optimised for the needs of track finding. In the course of the remaining phase 2 work, the balancing between an increased number of combinations due to numerous track elements and compute needs will be optimised. The parameter space for the optimisation was reduced significantly by these studies, which showed that the combinatorics by an early inclusion of u- and v-layers would be too high for a significant improvement of the performance.

With completion of phase 2, both approaches will have been systematically studied and optimised. We expect that GPU-based tracking for the SciFi tracker will be available and optimised for dual use of CPU and GPU.

*Hybrid Processing on GPUs/CPUs.* Dealing with heterogeneous hardware, such as CPU/GPU combinations, is one of the important unsolved problems in computer science in general. The above-mentioned developments offer efficient solutions to individual problems (such as tracking here). In [C5/406], we also developed an important building block to the general problem of data processing on heterogeneous hardware. To emphasise the generality of our strategy, we devised methods for truly hybrid processing in the context of a full-fledged SQL database engine.

Other than existing work, our novel methods *query chopping* and *data-driven operator placement* seamlessly also handle *concurrency* (e.g., when multiple users compete for the same GPU) and *resource contention* (e.g., when workloads demand more memory than the device can support) and provide *robust runtime characteristics* even under such adverse conditions. The results in [C5/406] will become particularly valuable to *integrate* the tailor-made *LHCb* algorithms that we have developed so far.

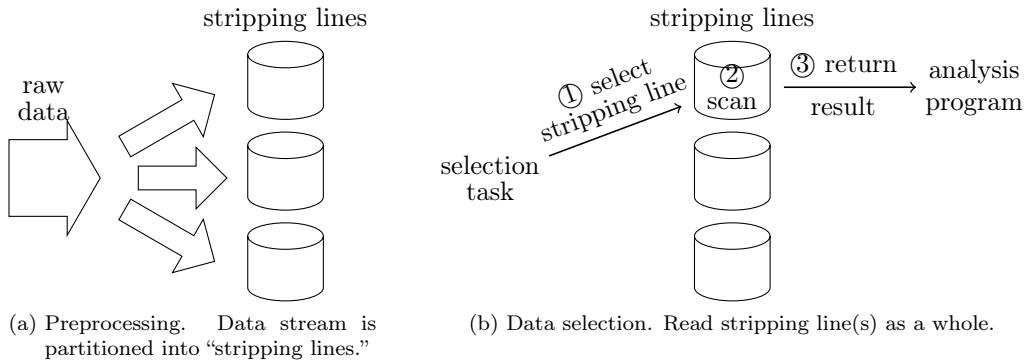


Figure 3.2: *LHCb* “stripping” concept. The input stream is partitioned into a few hundred “stripping” lines after data acquisition (events may also be stored in multiple stripping lines). The only access mechanism allowed is to scan a full stripping line.

**Stripping Stage.** Most physical analyses on the *LHCb* data set – from a data processing perspective – boil down to selecting few relevant collision events out of a vast pile of data. Thereby, (partial) selection criteria can be as simple as “return all events that produced a muon particle with an energy of at least ...” but may also be as complex as graph conditions on the third particle tracks (the latter can be inferred through compute-intensive track-finding algorithms). In the course of the project, it became clear that the resulting query complexity and diversity essentially rule out access structures like (multidimensional) indexes, leaving *scans* as the only viable search mechanism.

Analysis performance is, therefore, heavily influenced by the *volume* of the data that is being scanned. To reduce this volume, the existing platform at CERN uses a mechanism that physicists refer to as *stripping*: a *preprocessing stage* segregates all events into so-called *stripping lines* on the storage cluster; each stripping line corresponds to predefined search criteria.<sup>1</sup> Figure 3.2 illustrates the stripping concept: the high-volume stream coming from the left is partitioned and stored into multiple files (“stripping lines”). Stripping lines, thereby, overlap whenever an event satisfies the predicates stated for more than one stripping line.

To date, a few hundred stripping lines are registered in the *LHCb* system, which was found to be a compromise between selectivity and the cost of preprocessing. In fact, stripping lines need to have a selectivity below 0.5 %, so the cost of materialisation stays manageable. The stripping concept is both a blessing and a curse. While it reduces the scan cost for common analysis (types), stripping is limited to classes of analyses that have been pre-declared to the stripping process. In addition, individual events may be found in more than one stripping line, which increases storage demands and adds an undesired level of complexity in bookkeeping for analyses that use data across stripping lines. Thus, an improved solution would be highly desirable for particle physicists.

With *DeLorean*, we developed a mechanism to *replace* the stripping concept with a novel (and highly successful) strategy. *DeLorean* consists of two parts:

1. A *raw data part* contains the full *LHCb* data set. This part is kept in ROOT, which is the native data format that most analysis programs expect when reading their input.
2. *DeLorean’s relational part* is a compressed *synopsis* of the full data set, stored as a relational *column store*. The synopsis includes only those fields of the data set that are queried frequently (using simple, “sargable” predicates) and with high selectivity.

<sup>1</sup>A stripping line compares best to a *materialised view* in a relational database engine.

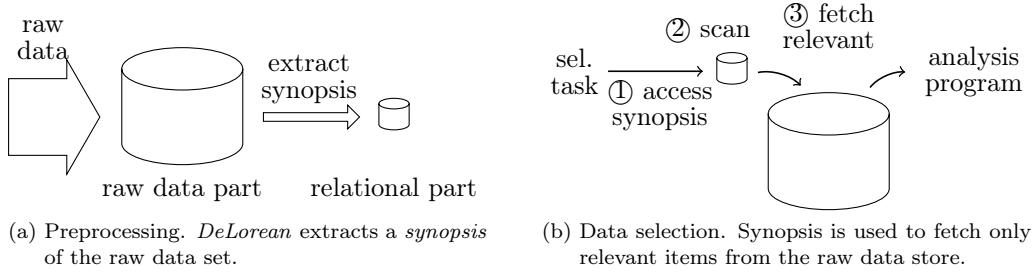


Figure 3.3: *DeLorean* concept. At data acquisition time, *DeLorean* extracts a compact *synopsis* of the raw data. At analysis time, *DeLorean* permits ad hoc selection criteria, fetches only a relevant data subset from the raw data part.

When an analysis program requests data from *DeLorean*, the system uses the synopsis to identify *candidate matches*. Only candidate matches are then fetched in their entirety from the raw data part. For predicates with high selectivity – which is the typical situation – this indirection can result in significant savings in the overall scan volume.

Figure 3.3 illustrates the *DeLorean* concept. The benefit of the mechanism is twofold: focusing on only relevant data items can significantly reduce the accessed data volume, hence improve performance; at the same time, *DeLorean* is not restricted to pre-configured selection criteria but supports *ad hoc analyses*. Another way of looking at the latter characteristic is that *DeLorean* allows for *stripping-less processing* for the first time.

As we detailed in [C5/405], *DeLorean* outperforms the “DaVinci” system at CERN by several factors. What is more, *DeLorean* scales well with the available processing resources and is currently only bound by the available I/O bandwidth capacity.

There is still substantial potential to improve *DeLorean*. The decision, which information to store within *DeLorean*’s synopsis and which aspects of the query to evaluate there, has to be done manually in the current version of *DeLorean*. Clearly, automating this process could further boost productivity in the field. At a different end, the I/O limitation of *DeLorean* is an indicator that performance could further be improved by better *balancing* compute and scan tasks. This is an important driver to our plan of deeper *integrating* the “stripping” and “analysis” stages within phase 3 of the CRC. Similar in spirit, with *Eriador*, we devised a hardware platform to solve *LHCb* analysis task in a highly energy-efficient way; during phase 3, we plan to mirror this success and develop a distributed *intelligent storage system* to improve performance and energy efficiency of the *LHCb* stripping stage.

**Analysis Stage.** The actual data analysis is initiated by individual physicists. To this end, physicists prepare analysis programs that read data from one or more stripping lines, then perform compute-intensive operations on the data. A key ingredient of the data analysis is the selection of the relevant decay channel. Typically, we apply machine learning methods in a later stage of the analysis to optimise the efficiency and background rejection. Although several methods have been investigated, most of the analyses make use of Boosted Decision Trees.

The high compute demand of typical analyses makes the programming of analysis code tedious and error prone. Within phase 2 of the CRC, we therefore experimented with parallelisation/distribution strategies based on the *Map-Reduce* paradigm and showed that they can achieve attractive analysis speeds, combined with an easier development process.

To gauge the potential of an improved, *Map-Reduce*-based computing solution, we prepared three *reference analyses*:

- $B^0 \rightarrow J/\psi K_S^0$  which shows a relatively large yield and serves as a benchmark for the time-dependent CP asymmetry measurements. In addition, the relevant parameter  $\sin 2\beta$  has been measured with precision at the so-called B-factories with the experiments Babar and Belle. Thus, the analysis is a key measurement to demonstrate *LHCb*'s capability to perform this type of measurement relying on B-flavour tagging in a potential hostile hadronic environment. The resulting value for  $\sin 2\beta = 0.731 \pm 0.035 \pm 0.02$  [C5/402] compares well with the world average value. It demonstrated that *LHCb* will soon achieve the best individual measurement of this value. Further improvements resulted from the inclusion of  $\psi(2S)$  modes and  $J/\psi$  decays into electrons ( $J/\psi \rightarrow e^+e^-$ ) instead of muons [C5/404]. A key ingredient for all measurements of time-dependent CP asymmetries is the so-called flavour tagging, which aims at the determination of the production state of a neutral  $B$ -meson ( $B$  or anti- $B$ ) as its quality directly relates to the precision of the measurement. A new framework and new methods have been successfully developed to improve the performance and to cope with the increasing number of tracks per events due to the increased centre-of-mass energy of the LHC in Run II and with the increased luminosity for the upgraded experiment.
- $B^0 \rightarrow D^{(*)} + D^-$ : These particular types of decays are of particular relevance for the precise determination of the  $B_s$ -mixing phase  $\phi_s$ , which carries a large sensitivity to new physics effects. The measurement of time-dependent CP asymmetries in these decay channels measures the parameter  $\sin 2\beta$  plus a small deviation. Together with the measurement of  $\sin 2\beta$  in the decay mode  $B^0 \rightarrow J/\psi K_S^0$ , a precise determination of the deviation follows, which in turn relates to an unknown deviation from the  $\phi_s$ , measured in the decay mode  $B_s^0 \rightarrow D_s^+ D_s^-$ . In the decay  $B^0 \rightarrow D^{(*)} + D^-$  evidence for CP violation has been found with a significance of  $4\sigma$ . The corresponding deviation in  $\sin 2\beta$  was constrained to  $\Delta\phi = -0.16^{+0-19}_{-0.21}$  rad [C5/403, C5/408]

In order to increase the precision of the measurement, preparations are under way to incorporate new data into the analysis. In addition, another decay mode is being analysed:  $B^0 \rightarrow D^{*+} D^-$ . This work is far advanced. A cut-based selection has been developed to allow for better tests with a *Map-Reduce*-based selection. A dedicated fitting framework has been developed to extract the CP observables from data.

- Search for the rare  $B$ -decay  $B_s^0 \rightarrow K_S^0 K_S^0$  in *LHCb* data. This decay mode, which is a rare CP eigenstate, is of highest interest for the *LHCb* physics program due to its sensitivity to new physics effects. Experimental difficulties arise from the subsequent kaon decays  $K_S^0 \rightarrow \pi^+ \pi^-$ . Due to the comparably large lifetime of the  $K_S^0$ , a sizeable fraction of the pions won't have hits in the vertex locator. Kaons from these so-called downstream tracks are not fully considered by the (software) trigger; i.e., events with both kaons formed from downstream tracks have not been considered in the trigger. As preparatory work, the trigger has been modified to include these events, too. We are now looking forward to new data, especially from the upgraded detector.

For the latter analysis, a *Map-Reduce* framework has been developed, and the resulting selection has been tested and compared with the results based on the standard ROOT-based analysis. The results are very promising. Without any optimisation, similar performance figures as by the standard analysis have been obtained, indicating that considerable improvements can be achieved in subsequent optimisation of the *Map-Reduce*-based selection [392]. Thus, the data analysis will be ready for inclusion in *DeLorean* in the near future.

**Analysis on Low-Power Hardware.** The analysis stage at *LHCb* is massively compute- and therefore also energy-intensive. Thus, we investigated the use of *low-power ARM processors* for use in

particle physics. As a first step, together with students we developed a low-power compute cluster based on ARM processors, nicknamed *Eriador* [C5/401]. With this cluster, several studies have been performed:

- Much of the CPU load in *LHCb* analysis programs is spent on *track reconstruction*. To improve runtime and energy efficiency of that task, we developed a novel *hybrid seeding* algorithm for execution on low-power hardware [C5/409]. Our results demonstrate that the use of massively parallel low-power hardware can improve the energy efficiency of the task by a factor of  $\approx 5$  when compared to state of the art Intel processors. By leveraging parallelism, this efficiency can directly be translated into improved runtime performance (while staying within a given power budget).
- As mentioned above, we also used *Eriador* in the context of *ELPACO*. We analysed a typical particle decay chain:  $D_s^+ \rightarrow \phi\pi^+$  with  $\phi \rightarrow K^+K^-$ . The performance was compared to computations on the CRC cluster (running Intel Xeon). Although the Xeon platforms turned out to be slightly faster, the energy consumption for the ARM cluster was greatly reduced; i.e., the number of events per Joule was significantly larger.
- Monte Carlo simulations: In cooperation with C3, CORSIKA, a Monte Carlo code to simulate extensive air showers for cosmic rays was adopted for use with *Eriador*. Very promising results were obtained with an “out-of-the-box” code, indicating again the superior energy efficiency of the ARM processors. In the future, the cooperation with C3 will be extended to use a Monte Carlo code adapted for the needs of the Cherenkov telescopes MAGIC and FACT, as well as the neutrino telescope IceCube.

**Results Outside *LHCb*.** In order to improve the extraction of physics observables, a collaboration with C3 has been launched. In particular, the experimental resolution needs to be modelled with highest accuracy in fits to the data. For example, the determination of the number of events contributing to a certain reaction is typically relying on the analysis of the invariant mass distribution. Since the resolution cannot be described by a single Gaussian distribution, complex functions are used to model the mass distribution. Imperfections in the modelling of the mass distribution will therefore introduce systematic uncertainties in the subsequent determination of the CP violation parameters. The unfolding algorithms of C3 bear the potential to eliminate the need to understand the mass resolution by taking into account other parameters relating to the mass resolution. Based on simulated *LHCb* data with known truth, the mass spectrum for reconstructed  $K^*$  has been successfully unfolded.

The *DeLorean* framework demonstrates how valuable it can be to replace classical, application-specific access structures with a compact and flexible relational data structure. A key reason is that the latter will much better embrace advances in hardware technology, which favour stream-based, parallelisable processing modes.

Discussions with CRC partners from the C1 project inspired us to employ the same mechanism also within the context of *genome data*. In [C5/400], we propose a method to encode genome data sets using a relational data model. Our method is carefully tuned to leverage the strengths of modern *in-memory database engines* (*CoGaDB*, for that matter), such as *column orientation* or *lightweight compression*. The outcome is a system that is performance-competitive with sophisticated, tailor-made systems that have been engineered over years. Yet our system offers the ease and flexibility of an SQL-based system, which may significantly enhance the productivity of genome analyses.

## Bibliography

- [387] A. Alexandrov, A. Kunft, A. Katsifodimos, F. Schüler, L. Thamsen, O. Kao, T. Herb, and V. Markl. "Implicit Parallelism through Deep Language Embedding". In: *Proc. of the 2015 ACM SIGMOD Int'l Conference on Management of Data*. Melbourne, Victoria, Australia, May 2015, pp. 47–61 (cit. on p. 348).
- [388] S. Breß, F. Beier, H. Rauhe, K. Sattler, E. Schallehn, and G. Saake. "Efficient Co-Processor Utilization in Database Query Processing". In: *Information Systems* 38.8 (2013), pp. 1084–1096 (cit. on p. 349).
- [389] Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst. "HaLoop: Efficient Iterative Data Processing on Large Clusters". In: *Proc. of the VLDB Endowment* 3.1 (2010), pp. 285–296 (cit. on p. 348).
- [390] C. Faerber. "Acceleration of Cherenkov angle reconstruction with the new Intel Xeon/FPGA compute platform for the particle identification in the LHCb Upgrade". In: *J. Phys.: Conf. Ser.* 898 (), p. 032044 (cit. on p. 351).
- [391] V. Khachatryan, I. Bediaga, B. Spaan, and andere (CMS and LHCb Collaborations). "Observation of the rare  $B_s^0 \rightarrow \mu^+ \mu^-$  decay from the combined analysis of CMS and LHCb data". In: *Nature* 522 (2015), pp. 68–72 (cit. on p. 336).
- [392] A. Kilic. "Untersuchungen zu neuen Strukturen der LHCb-Daten auf Basis von Hadoop". Abschlussarbeit. Dortmund, Germany: TU Dortmund University, Dec. 2017 (cit. on p. 342).
- [393] LHCb Collaboration. *Framework TDR for the LHCb upgrade*. 2012 (cit. on p. 337).
- [394] LHCb Collaboration. *LHCb Tracker Upgrade Technical Design Report*. 2014 (cit. on p. 337).
- [395] LHCb Collaboration. *LHCb Trigger and Online Upgrade Technical Design Report*. 2014 (cit. on p. 337).
- [396] LHCb Collaboration. *LHCb Upgrade Software and Computing Technical Design Report*. 2018 (cit. on p. 338).
- [397] P. Marwedel. *Embedded system design: Embedded systems foundations of cyber-physical systems*. Springer Science & Business Media, 2010 (cit. on p. 349).
- [A1/C3/398] **C. Bockermann** and **H. Blom**. *The Streams Framework*. Tech. rep. 5. TU Dortmund, 2012 (cit. on p. 338).
- [399] R. Weber, H. Schek, and S. Blott. "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces". In: *Proc. of the 24th Int'l Conference on Very Large Data Bases (VLDB)*. New York City, NY, USA, Aug. 1998, pp. 194–205 (cit. on p. 346).

### 3.3.2 Project-related publications by participating researchers

#### a) Peer-reviewed publications and books

- [C5/400] S. Dorok, **S. Breß, J. Teubner**, H. Läpple, G. Saake, and V. Markl. "Efficiently Storing and Analyzing Genome Data in Database Systems". In: *Datenbank-Spektrum* (June 2017) (cit. on pp. 16, 343).
- [C5/401] M. D. Götz, R. Kühn, O. Zietek, R. Bernhard, M. Bulinski, D. Duman, B. Freisen, U. Jentsch, T. Klöppner, D. Popovic, and L. Xu. "Energy Efficiency of a Low Power Hardware Cluster for High Performance Computing". In: *INFORMATIK 2017*. Ed. by M. Eibl and M. Gaedke. Gesellschaft für Informatik, Bonn, Sept. 2017, pp. 2537–2548 (cit. on pp. 16, 31, 343).

- [C5/402] LHCb Collaboration, R. Aaij, B. Adeva, M. Adinolfi, A. Affolder, **B. Spaan**, and et al. “Measurement of C P Violation in  $B^0 \rightarrow J/\psi K_S^0$  Decays”. In: *Physical Review Letters* 115.3 (July 2015), p. 031601 (cit. on p. 342).
- [C5/403] LHCb Collaboration, R. Aaij, **F. Meier, M. Schellenberg, B. Spaan**, and et al. “Measurement of CP violation in  $B^0 \rightarrow D^+ D^-$  decays”. In: *Physical Review Letters* 117.26 (2016), p. 261801 (cit. on p. 342).
- [C5/404] LHCb Collaboration, R. Aaij, **M. Schellenberg, B. Spaan, H. Stevens**, and et al. “Measurement of CP violation in  $B^0 \rightarrow J/\psi K_S^0$  and  $B^0 \rightarrow \psi(2S) K_S^0$  decays”. In: *Journal of High Energy Physics* 11 (2017), p. 170 (cit. on p. 342).
- [C5/405] **M. Kußmann**, M. Berens, **U. Eitschberger**, A. Kilic, **T. Lindemann, F. Meier, R. Niet, M. Schellenberg, H. Stevens, J. Wishahi, B. Spaan**, and **J. Teubner**. “DeLorean: A Storage Layer to Analyze Physical Data at Scale”. In: *Datenbanksysteme für Business, Technologie und Web (BTW 2017)*. Ed. by B. M. et al. (Hrsg.) Vol. P-265. LNI. GI, Mar. 2017, pp. 413–422 (cit. on p. 341).
- [C5/406] **S. Breß, H. Funke**, and **J. Teubner**. “Robust Query Processing in Co-Processor-Accelerated Databases”. In: *Proceedings of the 2016 ACM SIGMOD Conference on Management of Data*. San Francisco, CA, USA: ACM, June 2016 (cit. on pp. 339, 349).
- [C5/407] **T. Lindemann**, J. Kauke, and **J. Teubner**. “Efficient Stream Processing of Scientific Data”. In: *Proc. of the Joint HardDB & Active ’18 Workshop*. Paris, France, Apr. 2018 (cit. on pp. 338, 339).

### b) Other publications

- [C5/408] **F. Meier**. “Measurement of  $\sin 2\beta$  using charmonium and open charm decays at LHCb”. Diss. TU Dortmund University, 2016 (cit. on p. 342).
- [C5/409] **T. Lindemann**. *Efficient Track Reconstruction on Modern Hardware*. Tech. rep. 1. DBIS Group, Chair 6, Department of Computer Science, Mar. 2018 (cit. on p. 343).

## 3.4 Project plan

### Goals

Modern particle physics is a prime example of data processing on a massive scale. During the passing CRC phase, we made important contributions to the individual stages of the *LHCb* processing pipeline. Our overarching goal for phase 3 is to better connect our results across the pipeline, thereby also making them more generic and applicable to a broader application field.

From an application perspective, the *trigger stage* will face a significantly increased data volume with the coming upgrade of the experiment, demanding maximum algorithm efficiency in the trigger system. The *stripping stage* in its current form is a limitation with regard to both storage space and user flexibility. Finally, the complexity of *analysis tasks* is increasing and requires methods to handle this work in a resource-efficient manner. By connecting our results from phase 2, we expect to address these demands and enable the physicists to gain more insights into their data through better processing efficiency.

On the methodological side, isolated solutions to many of the sub-problems within the application domain exist by now. However, it is not clear whether they are actually ready to run at large

scale, and they are not connected to one another. By example of the real-world *LHCb* problem, we expect to gain insight into the challenges that arise when establishing such connections and develop solutions that we will design to be applicable also beyond our primary application field.

Our results will be of very high relevance to the *LHCb* experiment, but also to other scientific use cases that need to sift through extremely large data volumes. We will present the output of each work package to the *LHCb* collaboration, so they can apply our results to the actual experiments. Conversely, we will benefit from feedback by the *LHCb* community as well as from experience when they apply our work to the full data sets.

## Work schedule

Our work for the coming phase can be divided into two main *research tracks*: in *Track 1* (WPs 1-4), we want to *integrate* the *DeLorean* system with the (Hadoop-based) analysis framework; in *Track 2* (WPs 5-8), we want to marry the execution platform with *hardware and resource awareness*.

**Track 1: Integration of *DeLorean* and (Hadoop-based) Analysis Framework.** We developed *DeLorean* to replace the *stripping* component of the *LHCb* processing pipeline. From a technical perspective, this addresses the most data-intensive stage of the pipeline. Reducing and distributing but also avoiding I/O, therefore, was our principal guide when designing *DeLorean*. We did so by leveraging ideas, e.g., from VA-files [399] and lifting them to the level of massive, cloud-style scale-out systems. Our analysis back-end, including the *Eriador* hardware substrate that we developed in the course of the project, is built on massive scale-out technology, too (Apache Hadoop, for that matter). The focus here, however, is to distribute the compute demand of complex analysis tasks over many processors.

*Integrating* the I/O-bound *DeLorean* and the compute-bound analysis parts will yield important results both from a conceptual as well as from a practical perspective. Most importantly, *balancing* workloads is a good way to maximise the efficiency of a system – but new methods are needed to achieve that balance. On the practical side, an integrated system will yield a playground to better understand the involved trade-offs, while at the same time being an efficient execution platform for the *LHCb* experiment itself.

The track requires a very tight interaction between experts from the computer science and physics domains, in order to connect the application needs for massive scale with the technical solutions that exist in the computer science field. Track 1 will be spearheaded by Jens Teubner from the DBIS group, with his experience in the design of scalable database architecture as well as in language design and (query) compiler construction. Bernhard Spaan (particle physics group) not only will contribute the necessary physics background, but is also highly familiar with existing and previous computing architectures used in large-scale physical experiments.

**Track 2: Hardware and Resource Awareness.** In an era where energy, heat dissipation, and bandwidth constraints limit the achievable compute capacity, the move toward *heterogeneous hardware designs* is seen as the only promising route to achieve sustainable improvements in application performance. As a step towards the large-scale use of novel execution platforms, we already developed highly efficient GPU-based solutions for several sub-problems in the processing of *LHCb* data. During phase 3 of the CRC, we want to intensify these efforts and work towards designs that allow for truly hybrid processing. We envision a setting where different system components (CPUs, GPUs, FPGAs, etc.) each work at their maximum utilisation to jointly achieve optimal application performance.

Bernhard Spaan from the particle physics group has a track record of designing CPU-, GPU- or FPGA-based algorithms for individual physics problems. Jens Teubner (DBIS group), by contrast, has extensive experience in the design of *hybrid* CPU/GPU/FPGA processing, especially in the processing of large data volumes. Bernhard Spaan will coordinate the efforts in this project track. All work packages will jointly be addressed by both groups.

**Work package 1. Technical Foundation** *DeLorean*, based on Apache Drill, and the Hadoop Map-Reduce framework both run on top of the distributed HDFS file system, and both distribute their jobs in a massively parallel fashion. Both also have some notion of *data locality*, which they use to avoid communication across the cluster system.

The seeming similarities use different mechanisms underneath, however. Running *DeLorean* and Map-Reduce cooperatively requires understanding and coordination of the distribution and parallelisation mechanisms of the two. One important aspect here will be the locality aspect, so as to avoid moving data between nodes to the extent possible. A more technical – but equally important – aspect are the data interfaces between the two processing back-ends. *DeLorean* draws parts of its advantages from a *pipelined processing model*, whereas the communication mechanism between Map-Reduce jobs is based on files in HDFS.

In this work package, therefore, we want to establish the technical foundations to build a hybrid analysis engine that integrates both the “stripping” and “analysis” aspects illustrated in Figure 3.1.

On the methodical side, the key challenges of this work package are going to be questions around the *placement* of data and tasks as well as the *communication* between different components. A complete system will have to take placement and communication patterns into consideration when planning and scheduling an analysis task. Here, we lay the foundations for such a reasoning by establishing modes of placement and communication.

On the technical side, the work package will include the development of protocols, data formats, and interfaces between components (and integrate them into the complex world of *DeLorean* and Hadoop).

To get started, we plan to base this work package on a concrete example. Specifically, we want to first realise the reference analysis – which successfully guided our work in CRC phase 2 already – “manually” in the integrated system. As the C5 project continues, manual decisions within this implementation will be replaced step-by-step by optimisation mechanisms within the system.

With the upgraded experiment, the interplay of trigger and analysis becomes more and more important, as selections for many analyses will be integrated into the trigger decisions. With our reference analyses, we can therefore establish close links to the work packages dealing with the improvement of the trigger decision. The determination of the parameter  $\sin 2\beta$  will continue to serve as a guidance for the next step to generalise the concept for a broader use in particle physics, i.e., *LHCb*. Obviously, more than a single analysis is needed to demonstrate the performance of the system. As a first step, we will merge data from all reference analyses and subsequently add other data and corresponding analyses. In addition, we will further develop the analyses by making use of the unfolding algorithms developed in C3 to improve the extraction of parameters or observables from the data.

**Work package 2. Task Representation** (Automatic) reasoning over the characteristics of an analysis program (e.g., over its communication pattern or its bandwidth and data needs) requires that program to be available in a suitable *representation*. The state of the art in the physics

domain, hand-written C++ programs, clearly is not suited to reason over such characteristics. (Relational) database engines, at the other end, have long since used plan representations (e.g., inferred from relational algebra); but those representations typically do not have the expressiveness to cover physical analyses.

To fill the gap, in this work package we want to design a suitable representation to describe analysis tasks. A text-based version of that representation, a *domain-specific language (DSL)*, will be an invaluable tool to further study our integrated *DeLorean*/Map-Reduce platform. The representation we will include means to express which part of an overall task should be executed where in the system (and which data has to be transferred where). Ideally, the representation will come with *equivalence rules*, which could be used to rewrite analysis tasks. Such functionality will be the basis for building an optimisation component later on.

*Data flow-oriented* task representations are known to be a good basis for such scenarios. An example is “Pig Latin”, an early proposal to combine ease of use, data flow characteristics, and the possibility to compile for parallel execution. More recently, *Emma* [387] and *HaLoop* [389] were proposed with similar characteristics. We intend to use these proposals as a starting point for our work. However, the existing solutions are all too abstract to satisfy the physicists’ needs. Based on a library of abstract, database-style operators, they leave a gap to the formula-based notation that physicists are used to work with.

As mentioned above, WP 1 will use a *reference analysis* and hand-code it for execution on the integrated *DeLorean*/Map-Reduce platform. The data of the LHC Run 2 will be analysed. In addition, the analysis of data from the upgraded detector will be developed. In WP 2, we will use this analysis as a guide to design the task representation (and likely the very first task expressed in and supported by our task representation). Further realistic tasks from *LHCb* will then help to gauge the usefulness of our representation for actual analyses.

**Work package 3. Cost Analysis and Cost Modeling** The availability of a task representation language opens up the possibility to study different cost factors quantitatively. In modern, large-scale systems, for instance, it is not immediately clear how communication and compute costs relate; or whether it is desirable to trade compute overhead for an increased data transfer volume (or vice versa) if the analysis task permits such trades.

Based on the representation language from WP 2, we want to study the various cost factors that (might) arise. To this end, we want to first prepare artificial tasks (and their DSL representation) that stress one particular cost factor and quantify their effect via measurements on our evaluation cluster. Following up, we want to carry these results further to realistic analysis tasks (for which we prepared instances in WP 2, too). Our plan is to manually rewrite, possibly optimise, these tasks via our representation language and evaluate the effects on different cost metrics (e.g., latency, response time, system utilisation).

Cost analyses of that type fall well into the growing relevance of “systems” work, especially in the database field. Characterisations of systems – covering software as well as hardware aspects – are increasingly recognised by the community and seen as an important building block for future systems.

Cost analysis will seamlessly transition into the establishment of a *cost model* for the *DeLorean*/Map-Reduce combination. Such a cost model will provide a basis to rewrite analysis tasks in a systematic way. One way of looking at this work package, therefore is the construction of a simple “optimiser” component for our system; however, we think that building an optimiser that would actually deserve that name would go beyond what is feasible within the scope of the project.

Both cost analysis and cost modeling will build on results from project A2. That project also considers “communication” as a cost factor for algorithm analysis (albeit at a much higher level of abstraction). Likewise, the platform models developed in project A4 include cost models for communication; resource models for energy and memory are developed in project B2. Both projects consider platforms of a much smaller scale (i.e., embedded devices). Yet, their results will be highly beneficial for us. Specifically, we expect that communication will play an important role also in our environment, a massively distributed platform integrating *DeLorean* and Map-Reduce.

**Work package 4. Heterogeneous Hardware** Track 2 of our work schedule focuses on the use of heterogeneous hardware platforms in the context of physical analyses. While most of that work will happen outside the *DeLorean*/Map-Reduce platform, the work package *Heterogeneous Hardware* will be concerned with integrating the developments of Track 2 into the common platform. In particular, we want to extend the *DeLorean*/Map-Reduce platform by the possibility to leverage *graphics processors* available in the individual nodes.

Such an integration will touch many aspects of the platform. For instance, the DSL for task representation likely will need extensions to deal with the heterogeneous hardware (including means to express and modify the assignment of tasks to the different execution units). From earlier work [388], we know that cost models may vastly differ between CPUs and GPUs – an aspect that will require a reconsideration of our cost model on the one hand, but offer new opportunities for optimisation on the other.

Heterogeneous hardware platforms are seen as a key building block to satisfy the growing demand for data analysis capacity [397]. With this work package, we expect to contribute to this important field.

**Work package 5. Hybrid GPU/CPU Algorithms for SciFi Tracking** During the running phase of the CRC, we developed a GPU-based tracking-optimised algorithm for the SciFi Tracker of the *LHCb* experiment, which was a major achievement in the current funding period, i.e., phase 2. For the remainder of phase 2, we will tie this algorithm better into the available high-level trigger infrastructure and work towards a trigger component that can off-load parts of its work to an available GPU.

Offloading work to a GPU (or to other processing devices) this way makes it possible to benefit from hybrid processing capabilities. The approach is limited, however, to predefined algorithms or workloads, where compute demands and context are reasonably well defined. In *LHCb*, these conditions are met only in very few situations (such as dedicated trigger machines that always only run trigger tasks). The typical situation, by contrast, is varying workloads in combination with a large and diverse set of analysis tasks, turning the proper allocation of CPU and GPU resources into a real challenge.

In the work package, we will focus on the challenge of where to off-load which algorithm to a GPU (i.e., on the *placement* problem). Such a decision will require information about the expected cost of running a specific algorithm on a specific device as well as information about the current execution context, in particular about the current load situation in the system. On top of that, a fallback mechanism will be needed to cope with the situation when task offloading fails (due to the limited amount of device memory, a situation that arises quite frequently as soon as multiple analysis tasks compete for a single GPU).

In particular with respect to the latter issue, but also with respect to the consideration of execution context, we will significantly benefit from our earlier results [C5/406] (where we studied a much

simpler variant of the problem in a database setting) as well as from our intensified cooperation with project A2 (where communication costs to, from, and within the GPU were studied).

For the prediction of cost, we want not only to investigate the classical path of modeling each algorithm's behaviour, but also to evaluate machine learning methods as an alternative. We expect to benefit from a collaboration with project A1 with respect to the latter approach.

To limit the scope of the work package, we will first consider only the SciFi tracking component within the trigger stage of the *LHCb* processing pipeline. This not only significantly reduces the complexity of the problem in terms of code size and development effort; it also enables us to directly apply our previously developed GPU implementations of the tracking algorithms.

The complete solution will also require mechanisms to generate GPU code for individual problems and to ship code and data to/from the device. We plan to either reuse our previously written tracking code with regard to these aspects or draw from existing work.

A thorough evaluation of the previous steps is mandatory before initiating the next steps. A successful evaluation is a prerequisite to convince the *LHCb* collaboration to consider our results as an alternative to implement the trigger system of *LHCb*. Thus, we aim to involve other *LHCb* members in the final evaluation step.

**Work package 6. Generalised Hybrid GPU/CPU Algorithms** Following up on WP 5, we want to lift the limitation of using hybrid processing capabilities only for SciFi tracking and generalise the approach to the full tracking, including VELO, UT, and SciFi Tracker of the *LHCb* experiment in the trigger. Likewise, the placement algorithms will be adapted for use in the full tracking. Several issues need to be addressed: for example, stand-alone tracking in a subdetector (i.e., VELO) or matching of tracks across different subdetectors. Thus, some parts of the algorithms can be invoked in parallel to the tracking algorithms for the SciFi tracker. Therefore, we want to investigate placement options to minimise the waiting times before executing the final track finding and fitting stage that needs information from all tracking subdetectors.

In order to further speed up the trigger, other detector components will be taken into consideration. Clearly, the optimal use of GPU/CPU will depend strongly on the possibility to process single events in a highly parallel fashion. The most promising detector component seems to be the RICH detectors, where numerous hits need to be associated to individual tracks. In the RICH detectors, Cherenkov photons are emitted with a velocity-dependent angle along the trajectory of a charged particle track. The photons are then focused onto a read-out plane to form rings. Due to the large number of tracks, these rings consisting of a few ( $\leq 10$ ) detected photons overlap significantly. Thus, the task is to identify the ring associated with an individual track and to determine the Cherenkov angle from which the mass of the particle can be inferred once the momentum is known. As the mass is the key to identify a particle, the corresponding significant information gain results in improved trigger precision.

The analysis of RICH detector output represents a particular class of tasks within the *LHCb* framework. Its result will improve the *precision* of the trigger decisions, but decisions can also be made without RICH information. Therefore, in the work package we want to develop methods to either *omit* RICH analysis altogether when the system is tight on compute resources or *abort* RICH analysis when requirements for reaction time cannot be met otherwise. Most likely, RICH analysis will always be placed on the GPU device, so the process can never slow down any tasks on the CPU.

This work is closely linked to the tracking-related work, as the reconstruction algorithm depends on the track parameters determined from the full tracking algorithm.

**Work package 7. Hardware/Software Co-design** Another highly promising approach to optimise the trigger performance is to make use of the possibility to further invoke hard/software co-design based on the new type of CPUs that include GPU or FPGA features. The placement algorithm will be augmented to include this new hardware to dynamically optimise the performance of this new compute platform. The challenge here is that additional types of processing units significantly enlarge the search space to decide the placement of individual analysis tasks: There are  $n^k$  possibilities to place  $k$  analysis tasks on  $n$  different kinds of processing units. Therefore, in the work package we want to develop methods and heuristics to build a placement that can embrace different types of processing units.

The increasing demands on data throughput and compute resources, combined with the fact that there are little improvements in the speed of the single CPU, resulted in several developments of further acceleration based on large-scale parallelisation. Within phase 2 of the CRC and the previous work packages of C5, the additional use of powerful GPGPUs is in the focus of the research.

Companies such as Intel and AMD have developed other approaches of closely integrating this type of parallelisation capabilities in the CPU, resulting in the Intel's Xeon Phi processor and AMD's APUs. With Intel's acquisition of Altera, an FPGA-based acceleration platform has been developed. In those approaches, limitations introduced by the use of PCIe have been overcome with new concepts such as UltraPath Interconnect (UPI).

As part of the development of optimised placement concepts, we will include these new hardware possibilities, beginning with the possibility to include GPU-based accelerators. In addition, we will investigate the possibility to include FPGA accelerators. However, the scope of the work package does not allow for the development of FPGA-based algorithms for all detector components. Within *LHCb*, however, FPGA implementations for several relevant components already exist or will be developed in the context of other *LHCb* activities. For example, an algorithm for the reconstruction of Cherenkov angles for the RICH detectors has been developed and tested with the Intel Xeon/FPGA compute platform, with promising performance characteristics [390]. The data transfer to the FPGA will be a limiting factor, which will be an important challenge within the work package. In addition, the complexity of the application needs to be taken into account, as even relatively simple algorithms consume a sizeable fraction of the FPGA's capacity.

The results of this work package will directly benefit the *LHCb* experiment. However, we aim for heuristics that will also be applicable outside our immediate application. In particular, we plan to work on this work package in close collaboration with other CRC projects that develop algorithms for novel classes of processing units (e.g., FPGA algorithms are developed within project A1).

**Work package 8. Trigger Decisions based on Raw Data** The key challenge in the trigger stage of the *LHCb* experiment is to decide, within a tight time interval, whether a collision's data set should be kept or discarded. To this end, state of the art trigger implementation depend on a *full reconstruction* of the collision event. That is, the trigger software computes the precise track of each particle, together with particle properties such as momentum and mass. Reconstruction is highly compute-intensive. The existing code uses heuristics and learning-based methods to accelerate local decisions (such as particle identification); but the majority of the procedure is guided by the compute-intensive reconstruction.

The larger vision of our work would be a trigger solution that makes decisions *directly* from the raw data set and discards (some) events before reconstructing them expensively. Decisions based on raw data could save a substantial amount of compute resources, which would be a giant leap forward to make more resources available and gain better insights within the analysis stage. Work package 8 is a *side-track* work package where we want to follow the idea. It will run parallel to

## Project C5

the entire project, because the concept will interact with many of the issues that we will explore in the other work packages.

In cooperation with relevant CRC projects, in particular with A1, B2, and A6, we want to evaluate the possibility to deploy machine learning or other classification algorithms on raw data, i.e., data as it is directly recorded by the experiment. Specifically, A1 can provide the necessary background in foundational learning algorithms; B2 has experience in the application of such algorithms to sensor readings and spatial data; with A6, we would like to evaluate whether the graph characteristics of particle decay chains can be leveraged for decisions based on raw data.

Obviously, this work package has a highly explorative character and can be considered as a high risk/high gain project, making use of the excellent collaborative framework of the CRC. Demonstrating the feasibility of this strategy would be a tremendous success.

Some very early steps toward decisions based on raw data have already been made in the second CRC phase. Specifically, the analysis of the *LHCb* data showed that using just raw data with no additional information such as the effective positions of the individual hits adds an unnecessary level of complexity to the problem. Thus, we are currently working on the proper data preprocessing to have an optimised association of hit to the proper location. In principle, 3D information should be appropriate. However, the number of hits in the tracking station within one event is large (order of 10,000). It would be highly beneficial to make use of the fact that the dipole magnet of the experiment bends the charged tracks only in one plane.

In the cooperation with B2, we want to restrict the trigger decision to data from the VELO only, making use of the relatively long lifetime of Hadrons containing b-quarks, which travel on average about 1 cm before they decay. B2 will provide differentiable modules for analysis of irregular structured data. Since most of the particles are produced at the point of the proton-proton collisions, the long lifetime leads to so-called secondary vertices with a comparably smaller number of tracks emerging from them. Since there is only a very small magnetic field in the VELO regions, the tracks can be considered as straight tracks. The upgraded VELO detector will consist of individual sensor planes (perpendicular to the beam axes) made of pixels, thus providing precise 3D information for each hit in the detector. The sensitive region starts at a distance of about 5 mm from the beam axes. Typical tracks within the acceptance of the *LHCb* detector will have 5 to 10 or more hits in the VELO, which is sufficient to allow the detection of straight tracks.

With A6, we want to investigate the possibility to find tracks within the entire tracking system by means of graphs, as tracks can be considered as graphs, in particular as a link prediction task.

The project will make use of simulated data, where true tracks are known. Full events from the upgraded detector will contain on average 5 or more proton-proton collisions. Thus, we have full control of the degree of complexity of the event. Starting with single tracks with and without added noise, more and more tracks will be added subsequently to the analysis.

## Schedule

Work package	Quarter	2019				2020				2021				2022			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. Technical Foundation																	
2. Task Representation																	
3. Cost Analysis and Cost Modeling																	
4. Heterogeneous Hardware																	
5. Hybrid GPU/CPU Algorithms for SciFi Tracking																	
6. Generalised Hybrid GPU/CPU Algorithms																	
7. Hardware/Software Co-design																	
8. Trigger Decisions based on Raw Data																	

## 3.5 Role within the Collaborative Research Centre

The role of this project within the CRC may be summarised to two aspects: the project is a *data provider* with large amounts of real-world data; and we try to *benefit from and leverage techniques* developed in other projects.

With PI Jens Teubner, there is a natural collaborative link to project A2. Specifically, we want to leverage the CPU/GPU hybrid processing mechanisms developed there. Likewise, we plan to intensify our collaboration with B2 with respect to GPU acceleration.

C3 is concerned with particles from astrophysics. Although the particle types they look at are different, many techniques can be adapted to both experiments. We have an extensive exchange with C3 and, e.g., already benefit from their unfolding algorithms. In the other direction, we already demonstrated how our results in the context of the *Eriador* cluster can significantly improve the energy efficiency of the CORSIKA tool chain, which is heavily used by particle physicists.

We started to cooperate with A1 and A6 with respect to making trigger decisions based on learning techniques. Thereby, A6 looks particularly promising, because decay chains can also be viewed as graphs; graph analysis techniques from A6, therefore, look promising for our scenario. Similar in spirit, project B2 has great experience in the analysis of data with spatial information. We plan to study the incorporation of their results into the analysis of SciFi and VELO tracking data, which essentially consists of particles detected at given coordinates.

Conversely, we will provide our data sets to those projects as high-volume real-world data sets.

### 3.6 Differentiation from other funded projects

#### **LHCb: Quark-Flavor-Physik am LHC: Flavorsignaturen in Theorie und Experiment - LHCb: Run 2 und Upgrade**

**(Spaan, Reference number 05 H15PE15CL1) (Funding period: 2015–2018)**

This BMBF-funded collaborative project is the basis for participating in the LHCb experiment. Collaborating institutes from Germany are: RWTH Aachen, University of Heidelberg, University of Rostock, TU Dortmund, and the MPI for Nuclear physics Heidelberg. The BMBF has signed corresponding memoranda of understanding with regard to collaboration and CERN, which ensures the long term participation of German groups in LHCb. Bernhard Spaan's LHCb activities at Dortmund have been supported by the BMBF since 2004. A grant application for the next three years has been submitted recently. BMBF funding ensures that the participating universities can fulfil their obligations to the experiment. This applies in particular to the operation and maintenance of the experiment. In addition, the BMBF has also supported the German contribution to the construction of the experiment as well of the upcoming upgrade.

The BMBF funding is thus the basis of getting access to the data of the experiment in the long term. Thus, detector operations, upgrade and physics analysis are in the focus of the BMBF project, whereas in C5 computer science problems are tackled, based on the challenges from the large data rates and data volumes delivered by the LHCb detector.

#### **DFG Priority Program 2037 · MxKernel**

**(Teubner, Reference number TE 111/2-1) (Funding period: 2017–2020)**

Jens Teubner is a co-initiator of the DFG-funded priority program “Scalable Data Management for Future Hardware” and PI in the project “MxKernel: A Bare-Metal Runtime System for Database Operations on Heterogeneous Many-Core Hardware.” The project targets the co-design of database and operating system software and is not directly related to this proposal.

## 3.7 Project funding

### 3.7.1 Previous funding

Funding of this project within the Collaborative Research Centre started in January 2015.

### 3.7.2 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Doctoral researchers, 100 %	2	129,000	2	129,000	2	129,000	2	129,000
Total	—	129,000	—	129,000	—	129,000	—	129,000
Direct costs	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Instrumentation	Sum		Sum		Sum		Sum	
Total	0		0		0		0	
Grand total	129,000		129,000		129,000		129,000	

(All figures in euros)

### 3.7.3 Requested funding for staff for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Category	Funding source
<b>Existing staff</b>							
Research staff	1	Bernhard Spaan, Prof. Dr., professor	Particle physics	TU Dortmund	6	—	Existing funds
	2	Jens Teubner, Prof. Dr., professor	Information systems	TU Dortmund	6	—	Existing funds
	3	Ulrich Eitschberger, Dr., postdoctoral researcher	Particle physics	TU Dortmund	8	—	Existing funds
	4	N.N., doctoral researcher	Particle physics	TU Dortmund	8	—	Existing funds
	5	Kevin Heinicke, M.Sc., doctoral researcher	Particle physics	TU Dortmund	10	—	Existing funds
	6	Thomas Lindemann, M.Sc., doctoral researcher	Information systems	TU Dortmund	29.88	—	Existing funds
Non-research staff	7	Matthias Domke, technician	—	TU Dortmund	5	—	Existing funds
	8	Alla Stankjawitschene, secretary	—	TU Dortmund	1	—	Existing funds
	9	Britta Stickel, secretary	—	TU Dortmund	1	—	Existing funds
	10	Kai Warda, technician/engineer	—	TU Dortmund	5	—	Existing funds
<b>Requested staff</b>							
Research staff	11	N.N., doctoral researcher	Information systems	TU Dortmund	—	Doctoral researcher	—
	12	N.N., doctoral researcher	Particle physics	TU Dortmund	—	Doctoral researcher	—

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

- 1. Spaan, Bernhard**  
Project coordination, focus on to LHCb activities
- 2. Teubner, Jens**  
Project coordination, focus on to database aspects
- 3. Eitschberger, Ulrich**  
Reference analysis coordination, HLT data preparation
- 4. N.N.**  
Reference analysis, HLT, GPU
- 5. Heinicke, Kevin**  
Reference analysis, Hadoop/Map-Reduce/DeLorean adaption to LHCb
- 6. Lindemann, Thomas**  
hybrid algorithms, hardware/software co-design
- 7. Domke, Matthias**  
Focus on technical support
- 8. Stankjawitschene, Alla**  
administrative support
- 9. Stickel, Britta**  
administrative support
- 10. Warda, Kai**  
Support for GPU/FPGA related work

**Job descriptions of staff for the proposed funding period (requested funds):**

- 11. N.N.**  
Integration of *DeLorean* with other elements of the processing pipeline.
- 12. N.N.**  
Focus on HLT work, GPU, Hard/Software Co-Design

**3.7.4 Requested funding for direct costs for the new funding period**

	2019	2020	2021	2022
TU Dortmund: existing funds from university	6,500	6,500	6,500	6,500
Sum of existing funds	6,500	6,500	6,500	6,500
Sum of requested funds	0	0	0	0

(All figures in euros)

**3.7.5 Requested funding for instrumentation for the new funding period**

This project does not request any funding for major research instrumentation.



### 3.1 General information about Project MGK

### 3.1.1 Project title: Integrated Research Training Group

### 3.1.2 Project leader(s)

Rhode, Wolfgang, Prof. Dr. Dr., 17.10.1961, German  
Experimentelle Physik E5, Fakultät Physik, Technische Universität  
Dortmund  
Otto-Hahn-Straße 4  
44227 Dortmund  
Phone: 0231-755-3550  
E-mail: [wolfgang.rhode@tu-dortmund.de](mailto:wolfgang.rhode@tu-dortmund.de)

Is the employment of the principal investigator(s) at the institution(s) indicated contractually fixed for the duration of the proposed funding period?

Do any of the above mentioned persons hold fixed-term positions?

(x) no ( ) yes

Funding for principal investigator(s) at the institution(s) indicated is covered by core support (state funds or similar):

( ) no (x) yes

## 3.2 Summary

The Integrated research training group (RTG) is a designated training programme for doctoral students and postdocs in order for them to achieve an outstanding international qualification profile. In continuous cooperation with each other and with the Principal Investigators (PIs) of various projects, the young scientists combine their competences regarding data analysis and embedded systems, as well as expert knowledge on the different practical problems studied within the Collaborative Research Centre (CRC). This way the members of the RTG are enabled to increase their visibility in international research. In addition to attending conferences, they are introduced to actively designing workshops and the membership in programme committees.

In its essence, the concept of the RTG for the third funding period is a continuation of the concept, which proved its worth during the first and second funding periods of the CRC. The RTG is open to excellent young scientists, generally selected through international calls for proposals. Applicants are expected to (already) have strong focus on the research areas relevant to the CRC. In Dortmund students are able to acquire the relevant knowledge by attending selected lectures and, if applicable, by writing their master's theses as members of the CRC. The training in the RTG is based on this previous knowledge, which is complemented in a structured form. This way students and postdocs acquire a combined knowlegde in data analysis and embedded systems.

The training in the RTG is based on a canon of colloquial lectures held by CRC members and tailored to the central questions of the CRC. This way, doctoral students are guided towards high international research standards in a quick, effective, and comprehensive manner. The highest priority is given to enabling the young scientists to conduct original and independent research.

The doctoral candidates therefore present their own work in *Topical Seminars*, attached to the individual research groups of the CRC (*Data Analysis, Resource Restriction, and Information Gathering*). These presentations generally include dedicated discussions with project leaders and invited high-level international researchers. Since the CRC intends to already recruit excellent students as doctoral candidates at an early stage of their academic training, students who decided to conduct their bachelor's or master's theses as part of a project within the CRC, are invited to participate in the lectures and *Topical Seminars*. A limited number of doctoral candidates will be given the opportunity to familiarise themselves with their doctoral subject in the presence of outgoing project staff by means of scholarships.

The training within the RTG is concluded by classes on *Soft- and Technical Skills*. Technical Skill classes mainly focus on the handling of relevant software tools, such as RapidMiner, R, ROOT, and others. *Soft Skill* classes focus, for example, on enhancing the the students' communication skills. Furthermore, PhD students are given the opportunity to preferentially participate in training programmes available for the academic staff at TU Dortmund. All developed capabilities are brought together at the annual CRC workshops, where students are strongly encouraged to demonstrate their skills in oral and written form. The training of the young scientists is rounded off by a bi-annual summer school, to which outstanding international scientists are invited.

### 3.3 Tabular report

The means of the CRC 876 were used to support doctoral students and scholarship holders as shown in the Table 3.1.

	PhD students	scholarship holders
Number	86	3
Funding period, minimum	3 months	6 months
Funding period, average	29.84 months	6 months
Funding period, maximum	48 months	6 months
Amount of funding, minimum	25 % E13	1000 Euro/month plus 103 Euro material costs
Amount of funding, average	89.49 % E13	1067 Euro/month plus 103 Euro material costs
Amount of grant, maximum	100 % E13	1100 Euro/month plus 103 Euro material costs

Table 3.1: Tabular report on the funds budgeted for the promotion of doctoral candidates over the entire funding period. Short funding periods or small percentage shares of positions are created by switching from CRC funding to basic funding or by pro-rata funding. The maximum funding time is always the planned funding period.

**List of the completed dissertations** (2015 – 2018, including theses from the basic equipment):

- Helena Kotthaus, Methods for Efficient Resource Utilization in Statistical Machine Learning Algorithms (2018)
- Mathis Börner, Bestimmung des Energiespektrums von atmosphärischen Myonneutrinos mit 3 Jahren Daten des IceCube-Detektors (2018)
- Alexander Munteanu, Large Scale Statistical Data Analysis (2018)
- Ulrich P. Eitschberger, Flavour-tagged Measurement of CP Observables in  $B_s^0 \rightarrow D_s^\mp K^\pm$  Decays with the LHCb Experiment (2018)
- Olaf Neugebauer, Efficient Implementation of Resource-Constrained Cyber-Physical Systems Using Multi-Core Parallelism (2018)
- Dominik Kopczynski, Resource-constrained analysis of ion mobility spectrometry data (2017)
- Olivera Holzkamp, Memory Aware mapping strategies for heterogeneous MPSoC systems (2017)
- Wen-Hung Huang, Scheduling algorithms and timing analysis for hard real-time systems (2017)
- Matthias Meier, Co-Konfiguration von Hardware- und Systemsoftware-Produktlinien (2017)
- Christoph Borchert, Aspect-Oriented Technology for dependable operating systems (2017)
- Pascal Libuschewski, Exploration of cyber-physical systems for GPGPU computer vision-based detection of biological viruses (2017)
- Marco Stolpe, Distributed analysis of vertically partitioned sensor measurements under communication constraints (2017)
- Sabrina Einecke, The Data Mining Guide to the Galaxy-Active Galactic Nuclei in a Multi-Wavelength Context (2017)

## Project MGK

- Eugen Rempel, Statistische Analyse von hochdimensionalen toxikologischen Expressionsdaten (2016)
- Christoph Ide, Resource-Efficient LTE Machine-Type Communication in Vehicular Environments (2016)
- Dominic Siedhoff, A parameter-optimizing model-based approach to the analysis of low-SNR image sequences for biological virus detection (2016)
- Olga Erohin, Knowledge acquisition through data analysis for prospective time determination (2016)
- Brian Niehöfer, Modellbasierte Interferenzkompensation für die satellitengestützte Ortung in urbanen Szenarien (2016)
- Tomasz Fuchs, Charmante Myonen im Eis - Messung des hochenergetischen atmosphärischen Myon-Energiespektrums mit IceCube in der Detektorkonfiguration IC86-I (2016)
- Fabian Temme, On the Hunt for Photons-Analysis of Crab Nebula Data obtained by the first G-APD Cherenkov Telescope (2016)
- Kai Schenetten, Röntgen-Photoelektronenspektroskopie mit Silizium-Photomultipliern (2016)
- Christian Bockermann, Mining big data streams for multiple concepts (2016)
- Frank Meier, Measurement of  $\sin 2\beta$  using charmonium and open charm decays at LHCb (2016)
- Claudia Köllmann, Unimodal spline regression and its use in various applications with single or multiple modes (2016)
- Michel Lang, Automatische Modellselektion in der Überlebenszeitanalyse (2015)
- Nils Kriege, Comparing Graphs: Algorithms & Applications (2015)
- Markus Putzke, Selbstorganisierende Minimierung der Interferenz von Femtozellen in heterogenen Netzen durch zufällige Frequenzsprungverfahren (2014)
- Johannes Köster, Parallelization, Scalability, and Reproducibility in Next Generation Sequencing Analysis (2015)
- Katharina Frantzen, Von der Monte-Carlo-Produktion zur Datenanalyse - Eine Analyse der 2012 genommenen Daten des Aktiven Galaktischen Kerns Mrk 421 (2015)
- Ann-Kristin Overkemping, Messages from a Black Hole - A long-term analysis of the Active Galactic Nucleus Markarian 421 in the light of gamma-rays measured by MAGIC-I (2015)
- Sebastian Breß, Efficient Query Processing in Co-Processor-accelerated Databases (2015)
- Benedikt Konrad, A methodology for family-based balancing of varient flow lines (2014)

### List of the scholarship holders (2015-2018, supervised as part of the doctoral preparation):

- Wei Liu, supervised by Jian-Jia Chen in project B2. Work included design of an OpenCL-based remote offloading framework, resource management schemes on accelerators and energy aware management of embedded systems.

- Junjie Shi, supervised by Jian-Jia Chen in project A3. He supported research on real-time 5G scheduling strategies and efficient utilisation of these.
- Amal Saadallah, supervised by Katharina Morik in project B3. She worked on active learning for automated process sensor data and the potential combination with finite element simulations.

### 3.4 Qualification programme

We assume that the doctoral candidates in the integrated RTG have a qualified degree in one of the disciplines participating in the CRC.

**Overview:** The goal of the broad training available to PhD students within the CRC is to qualify them as competent and internationally recognised scientists. After obtaining their doctorate, these students take professional positions in industry or accept research positions. Such a research position can either be abroad or by setting up a junior research group within the system of publicly funded research and teaching in Germany. In both cases, the young scientists are expected to conduct original and independent research during their doctoral training, which also has to be appropriately communicated. Furthermore, PhD students are expected to develop and demonstrate their teaching abilities.

The individual projects of the CRC are highly interdisciplinary and have core requirements in computer science, statistics, physics, mechanical engineering, medicine, electrical engineering and biology. Every doctoral candidate therefore has to identify the areas in which he or she must acquire content in order to cope with the upcoming research tasks. The identification of the content is of course supported by the PIs. Since these contents may differ from candidate to candidate, we have – with the most emphatic support of the doctoral students – deliberately no obligatory<sup>1</sup> canon of lectures to be heard. Instead, the doctoral candidates commit themselves in a doctoral agreement to acquire the relevant skills and knowledge by attending the appropriate classes and lectures on their own responsibility. Awarding *Credit Points* for signatures on attendance lists during the attendance of events was deliberately waived from the outset in favour of the doctoral students' personal responsibility. Within the first and second funding phases, the doctoral students have proven their ability to independently acquire subject-specific and interdisciplinary knowledge from the canon of lectures offered to them. Furthermore, a lively discussion atmosphere has established itself in RTG-events. Both facts are strong indications that waiving *Credit Points* was a correct decision.

#### Programme components for the qualification phase preceding the PhD phase:

The last year of their master's programme is generally the main qualification phase for future doctoral candidates of the CRC, but other qualifications, e.g., the preparation of the bachelor's thesis or research work carried out as part of the RISE scholarship program, are also taken into account. Within this phase, the students are introduced to the methodology of scientific research under the guidance of doctoral students and other scientists involved in the project. For the students this includes familiarising themselves with the relevant techniques (for example, a specific software) and the opportunity to participate in CRC events. If possible, suitable candidates are further involved in CRC activities by employing them as student assistants of the specific projects.

**The study programme** is designed such that the doctoral students are given the greatest possible freedom for their research. The minimum study load for the teaching period is 4 hours per week. This time is divided into two two-hour events:

---

<sup>1</sup>A compulsory attendance requirement for lectures may not be pronounced for legal reasons at the TU Dortmund.

- The *Topical Seminars* differ from classic colloquia, as they generally include a large proportion of discussion with the invited international researchers and lecturers of the RTG. Students further participate in the *Topical Seminars* by taking responsibilities in the organisation of the events and by suggesting and inviting high-profile guests. In most cases the stay of a guest is also organised by the inviting student, which naturally offers the possibility for further discussions.
- In addition the doctoral students participate in two-hour *Project Seminars*, in which they present their ongoing research and in the form of lectures and discuss current scientific challenges.

Block courses are mainly offered during the lecture-free period and are generally dedicated to training *Soft* and *Technical Skills*. Doctoral students are given priority when attending training programmes open to the entire staff of the university.<sup>2</sup> Competences acquired in *Soft Skill* seminars are readily applied by the doctoral candidates, in their participation in the training and teaching of students in the qualification phase.

The following courses were organised for the doctoral candidates and tailored based on their requirements. All courses have been financed from University funds:

- RapidMiner basic training
- RapidMiner Radoop: Distributed data analysis with RapidMiner and Hadoop
- Scientific writing in english for conference and journal publications
- Presentation of scientific research
- Open Access Publications: Impact, licensing and publishing models
- Plagiarism and Citations: Types of plagiarism, recognition of plagiarism and correct citations
- Arrogance principle: Training for women in male-centric environments

To supplement the professional knowledge of the doctoral students regarding the core questions of the CRC 876, we recommend special seminars and lectures. The listed courses show the programme offered on a yearly basis in the second phase of the CRC. It will be, if possible, continued and amended in the third CRC phase.

- Real-Time Systems (Chen, Computer Science)
- Computational Omics (Rahmann, Computer Science)
- Project group 594 (1 year) Big Data (Bockermann, Blom, Morik Computer Science)
- Bachelor project Data mining and data analytics (Piatkowski, Liebig, Buschjäger, Computer Science)
- Project group Deep News Dive (Kersting Computer Science)
- Software ubiquitärer Systeme (Spinczyk, Computer Science)
- Data Processing on Modern Hardware (Teubner, Computer Science)
- Advanced Topics in Algorithms (Sohler, Computer Science)

<sup>2</sup>[http://www.zhb.tu-dortmund.de/wb/de/home/TU\\_interne\\_Weiterbildung/](http://www.zhb.tu-dortmund.de/wb/de/home/TU_interne_Weiterbildung/)

- Randomized Algorithms (Sohler, Computer Science)
- Graph Algorithms (Mutzel, Computer Science)
- Introduction to Nuclear- and Particle-Physics (Spaan, Physics)
- Astroparticle Physics (Rhode, Physics)
- Statistical Methods of Data Analysis (Rhode, Physics)
- Modellierung und Dimensionierung von Kommunikationssystemen (Wietfeld, Electrical Engineering)
- Mobilfunknetze (Wietfeld, Electrical Engineering)
- Bayes-Statistics (Ickstadt, Statistics)
- Statistics in Genetics (Ickstadt, Rahnenführer, Statistics)
- Statistical Visualisation (Geppert, Statistics)
- Embedded Systems (Marwedel, Chen, Computer Science)
- Cyber-Physical System Fundamentals (Marwedel, Chen, Computer Science)
- Graph Data Mining (Kriege, Computer Science)
- Algorithm Engineering for Graph Data Mining (Kriege, Computer Science)
- Data Mining for Structured Data (Kriege, Computer Science)

Scientific results are presented by the doctoral students in written and oral form at annual workshops, also organised by the students. The papers are published as technical reports by the CRC.<sup>3</sup> Suitable results are also presented at international conferences. Doctoral students are actively involved in the organisation of the bi-annual CRC summer school, which hosts high-level international lecturers and is supported by the executive board of the CRC.

**Internationalisation:** The main focus of the education in the RTG is on the training of skills for independent research at an international level. We strive to create an international atmosphere by means of international calls for doctoral and scientific positions. Lectures and discussions in the *Topical Seminars* and *Project Seminars* are generally in English. Great importance is given to the participation of doctoral candidates in international conferences.

**Active Participation:** The active participation of doctoral candidates is a key concept of the CRC. Their communication among themselves and their guidance given to students during bachelor's and master's theses are characterised by cooperation and peer-to-peer teaching from the very beginning. By inviting guest researchers and organising courses, they learn how to contact researchers and work with them, and how to organise workshops in practice. In these tasks the PhD students are first and foremost supported by the PI envisaged for their training in formal matters, but also by the entire CRC, in particular by the executive board. Doctoral students learn how research works in practice by getting actively involved, but at the same time being secured by the support of the CRC.

PhD students are highly encouraged to follow their own research ideas, in order to learn how to work independently and about the motivation that arises from following ones own ideas. The qualification profile of our students is also in high demand outside the CRC, as they are renowned

---

<sup>3</sup><https://sfb876.tu-dortmund.de/FORSCHUNG/techreports.html>

for their scientific work and that they do not experience any difficulties in finding jobs in or outside academia. In fact, attractive job offers have already been submitted to doctoral students before finishing the PhD.

**Rules of good scientific work:** Doctoral students are officially instructed by the TU Dortmund University<sup>4</sup> and are personally informed in the projects about the rules of good scientific work and about project-specific aspects of ethics and privacy, respectively.

**Measures to ensure the scientific and career-related qualification:**

- The TU Dortmund University is particularly committed to the advancement of women via a mentoring programme focusing on engineering sciences: *mentoring*<sup>3</sup> <sup>5</sup>, which is a joint programme of the three universities TU Dortmund University, Ruhr-University Bochum and the University Duisburg-Essen. Established in 2005, this programme is supported by the Ministry of Culture and Science in NRW and the Association of German Engineers. Doctoral students and mentors (professors, postdoctoral researchers, and industry executives) are brought together for an informal exchange of experience, which is decoupled from the supervision of the doctoral thesis, but rather focused on career planning, development, and counseling in dealing with scientific institutions, e.g., conferences and journals. In total more than ten PhD students from CRC 876 participated in the program.
- At TU Dortmund University, the process towards an internationally oriented course of study is consistently pursued, also for German students. This includes the Studium-International-Certificate, which takes advantage of the diverse cultural potential of TU Dortmund by promoting an exchange between German and foreign students, rewarding individual commitment in an international environment. As part of this certificate programme, doctoral students of the CRC are given the opportunity to profile themselves for their future careers, by acquiring key competences, e.g., in communication and cultural relations. In order to obtain the certificate, all participants must participate in an inter-cultural competence training programme and additionally meet two of the three following conditions: (1) a research stay abroad of at least one month, (2) completion of a language course at the Language Center of TU Dortmund or (3) commitment to the International Campus of the TU Dortmund University. The certificate is awarded by the International Office of TU Dortmund.<sup>6</sup>
- In the third and final phase of the project, suitable doctoral candidates are strongly encouraged to hold retreats at the Leibniz Center for Computer Science, Schloss Dagstuhl, on the joint writing of books on the topic of the CRC.

**Distribution of the courses at the universities of the UA Ruhr:** The UA Ruhr is a strategic alliance of the universities Duisburg-Essen, Bochum, and TU Dortmund University. Two of these locations (University Duisburg-Essen, TU Dortmund University) are involved in the CRC. Students are already able to attend suitable classes and lectures at any of the three universities already as part of their training. The geographic distance between the universities is small and therefore negligible.

**Guest researchers in the Integrated Research Training Group:** Guest researchers usually introduce themselves and their research shortly after their arrival with a lecture at the *Topical Seminar*. They then take part in the events of the project seminars and *Topical Seminars*. This way, communication and discussion between them and the doctoral students extends beyond the narrow circle of their own research. Contacts going beyond these events are of course entirely in the spirit of the CRC and strongly encouraged. The same applies to guest researchers who are invited for lectures at the summer schools.

---

<sup>4</sup>[https://www.tu-dortmund.de/uni/de/Forschung/Gute\\_wissenschaftliche\\_Praxis/](https://www.tu-dortmund.de/uni/de/Forschung/Gute_wissenschaftliche_Praxis/)

<sup>5</sup><http://www.mentoring-hoch3.de>

<sup>6</sup>[http://www.aaa.uni-dortmund.de/cms/de/Internationale\\_Studierende/PROFIN/](http://www.aaa.uni-dortmund.de/cms/de/Internationale_Studierende/PROFIN/)

**Summer Schools:** To date four international summer schools on *Resource-aware Machine Learning* (2012, 2014, 2015, and 2017) were organised by the CRC 876. In 2015 we took the opportunity to organise the school together with the ECML PKDD 2015 (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases) in Porto, Portugal. All courses of the summer school were taught by internationally renowned scientists, among them several young researchers from the CRC. In addition to classical lectures, the summer schools contained hands-on exercises and data analysis challenges. The next international summer schools on *Resource-aware Machine Learning* are planned for 2019 and 2021.

- The programme of the first summer school on *Resource-aware Machine Learning*, Dortmund 2012 is listed below:<sup>7</sup>
  - Introduction to Machine Learning, Michèle Sébag (LRI Univ. Paris-Sud, France)
  - 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, Denmark)
  - From Data Taking to Rapid Mining, Wolfgang Rhode
  - Use of the Power of your GPU: Massively Parallel Programming with CUDA, Nico Piatkowski
- Second summer school on *Resource-aware Machine Learning*, Dortmund 2014:<sup>8</sup>
  - Introduction to Machine Learning, Céline Robardet (INSA, Univ. Lyon, France)
  - 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

---

<sup>7</sup><http://sfb876.tu-dortmund.de/SummerSchool2012/index.html>

<sup>8</sup><http://sfb876.tu-dortmund.de/SummerSchool2014/index.html>

- Third summer school on *Resource-aware Machine Learning*, together with the ECML PKDD
  - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2015, Porto, Portugal<sup>9</sup>
    - Big Data infrastructures, Sebastian Schelter (TU Berlin, Germany)
    - Data stream models: Sketches, and Probability tools, Ricard Gavaldà (Universitat Politècnica de Catalunya, Spain)
    - Understanding Human Mobility with Big Data, Fosca Gianotti, Salvatore Rinzivillo (both Istituto di Scienza e Tecnologie dell'Informazione, Italy)
    - The streams framework, Christian Bockermann
    - Cyber-physical systems: opportunities, problems and (some) solutions, Peter Marwedel
    - Programming on modern hardware, Jens Teubner
    - Execution, energy, and communication models for resource-aware cyber-physical systems, Jian-Jia Chen, Peter Marwedel
    - Preprocessing of large-scale sensor data with RapidMiner, Marco Stolpe
    - Approximating structural properties of graphs by random walks, Christian Sohler
    - Probabilistic Graphical Models, Nico Piatkowski
    - Dense-subgraph discovery, Aristides Gionis (Aalto University, Finland)
    - Reading the Web, Estevam Hruschka (Carnegie Mellon University, USA)
    - Netflix: Recommender Systems, Carlos Gomez-Uribe (Massachusetts Institute of Technology, USA)
    - Big Data Stream Mining, Albert Bifet (LTCI, Télécom ParisTech, France), André Carvalho (Universidade Federal de Pernambuco, Brazil), João Gama (University of Porto, Portugal)
    - Multilabel Classification, Jesse Read (École Polytechnique, France)
    - Advanced topics in predicting structured outputs, Michelangelo Ceci (Università degli Studi di Bari, Italy), Sašo Džeroski (Jozef Stefan Institute, Slovenia)
- Fourth summer school on *Resource-aware Machine Learning*, Dortmund 2017:<sup>10</sup>
  - A Tale of a Quest for Business Intelligence from Social Data, Rakesh Agrawal (Purdue University, USA)
  - Programming Energy-Harvesting-Based Sensor Nodes, Olaf Spinczyk
  - Hands-On: Ultra low power learning, Mojtaba Masoudinejad
  - Field-programmable technology and machine learning, Wayne Luk (Imperial College London, UK)
  - Probabilistic Graphical Models, Nico Piatkowski

<sup>9</sup><http://www.ecmlpkdd2015.org>

<sup>10</sup><http://sfb876.tu-dortmund.de/SummerSchool2017/index.html>

- Tractable Probabilistic Graphical Models, Kristian Kersting
- Generalised LDPC architectures for high-throughput FPGA realization, Rob Maunder (University of Southampton, UK)
- Probabilistic Graphical Models, Nico Piatkowski
- Mobility Models, Thomas Liebig
- Deep Learning, Sebastian Buschjäger
- Graph Streaming, Chris Schwiegelshohn
- Deep Learning on FPGAs, Sebastian Buschjäger

### 3.5 Management and supervision

**Recruitment of Doctoral Students** In order to recruit excellent and appropriately qualified doctoral candidates, two paths are being followed. International candidates are contacted via a permanent call for applications and, wherever possible, via personal contacts of the project leaders. All open positions are permanently advertised on the CRC website. Interested and suitable candidates may test and show their abilities for several months as part of a scholarship in the CRC.

The tendering procedure is carried out by the PIs of the project to which the position is assigned. In addition to a qualifying degree, the candidates are expected to demonstrate thorough knowledge (through lectures and/or seminars) in at least one of the three focal areas of the CRC (*modern database systems, computer architecture, and data analysis*).

After application, suitable candidates are selected and invited to an interview, which is generally accompanied by a presentation given in front of the project members. The candidate to whom the position is offered is selected by the PIs, but taking into account the votes of the other team members. Then, a proposal is submitted to the head of the Research Training Group, who is responsible for checking compliance with the qualification criteria and reporting to the Executive Board and the representatives of the doctoral candidates.

Doctoral students working on topics of the projects as part of the *basic equipment* are admitted to the Research Training Group if they meet the corresponding qualification criteria. In this case an application for admission is also submitted to the head of the Research Training Group, who checks compliance with the qualification criteria and reports to the CRC Executive Board and the representatives of the doctoral candidates.

The rights and duties of each doctoral candidate are defined by the doctoral agreement (see figure 3.1), which is signed by the respective candidate and two supervisors. The doctoral agreement entitles the candidate to participate in the structured training programme of the RTG.

In order to enable doctoral candidates – especially those from abroad – to optimally familiarise themselves with their doctoral subject matter, the CRC was able to successfully offer short-term scholarships for a total of 48 months per funding period. This short-term scholarship may not exceed the duration of six months for a single doctoral candidate, but can be awarded for a shorter period, if appropriate. The duration of this short-term scholarship is not counted towards the three-year doctoral period. Candidates can be proposed to the CRC board by the project leaders, and the Executive Board endeavours to make the funds available for projects of appropriate content and timing.

 <b>technische universität dortmund</b> Integriertes Graduiertenkolleg des Sonderforschungsbereich 876	<b>PROMOTIONSVEREINBARUNG<sup>1</sup></b> zwischen der/dem Promovierenden <b>Frau / Herrn</b> _____ der /dem ersten Betreuer/in <b>Frau / Herrn</b> _____ der / dem zweiten Betreuer/in <b>Frau / Herrn</b> _____ und dem integrierten Graduiertenkolleg des SFB 876 vertreten durch [dessen Leiter] <b>Frau / Herrn</b> _____ <b>S 1 Thema der Dissertation</b> Der/die Promovierende erstellt beginnend am _____ eine Dissertation zum Thema: _____ Das Promotionsvorhaben zum Teilprojekt _____ wurde im Exposé vom _____ beschrieben und von ihm Betreuer/in/en angenommen.	<b>S 2 Zeit- und Arbeitsplan</b> 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.	<b>S 3 Aufgaben und Pflichten der Betreuenden</b> (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 Selbstständigkeit des/des Promovierenden und geeignete Maßnahmen zur Erhöhung der internationalen Sichtbarkeit der/des Promovende/n. (2) Der Betreuer/ die Betreuerin verpflichtet sich zur Betreuung - unabhängig von der Dauer der Finanzierung - bis zum Abschluss der Promotion.	<b>S 4 Aufgaben und Pflichten der/des Promovierenden</b> (1) Der/die Promovierende verpflichtet sich zu einer regelmäßigen Berichterstattung über inhaltliche Teilergebnisse der Dissertation sowie die Einhaltung des Zeit- und Arbeitsplans. 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 Wahrung der besuchten Lehrveranstaltungen und der besuchten Workshops und Konferenzen beinhaltet ist zu den Workshops einzureichen. Diesem Bericht sind die Veröffentlichungen und technischen Noten des/des Promovende/n beizulegen.
---	--	---	---	--

Figure 3.1: Draft of the doctoral agreement to be individualized.

<sup>1</sup> Diese „Muster-Promotionsvereinbarung“ orientiert sich an den Empfehlungen der DFG (DFG-Vordruck 1.90 – 7/08).

Although all PhD positions are advertised internationally, we have found that the recruitment of international candidates is quite challenging, due to the specific profile of the CRC. In international comparison, the level of education in Dortmund seems to be so high that only some applications from foreign universities were qualitatively suitable. On the other hand, highly qualified applicants often had comparative offers with considerably better financial resources. It was therefore not entirely trivial to find qualified international candidates willing to place their professional interest above pecuniary aspects.

It has been apparent since the first funding period that the research topics within the CRC are rather interesting to students at TU Dortmund University. This is likely due to the fact that members of the CRC were given the opportunity to inspire students in lectures and seminars and also to personally offer possible research topics to students. We have observed that some students are systematically preparing themselves for a PhD position within the CRC from the bachelor thesis on.

#### **Assignment of the doctoral candidates to their supervisors:**

Projects within the CRC are lead by at least two PIs with different scientific expertise, and PhD candidates apply for a specific project, which typically suggests the choice of a supervisor. As a rule, two scientists that are particularly well acquainted with the special problems of the doctoral thesis and chosen by the candidate as supervisors, sign the doctoral agreement.

#### **Organisation of the supervision**

Every PhD student is supervised by at least two PIs at all times. Inspired by discussions within the different working groups, we have decided that additional supervision of individual aspects of the dissertation, by a third supervisor may be provided if needed. Such additional supervision goes beyond a simple discussion with internationally renowned experts, which is accessible to the students at any time. This double supervision, which was found to be particularly well suited for the interdisciplinary character of training in the RTG, was also found to promote the scientific independence of the candidates.

The descriptions of the dissertation projects and work packages of the individual projects show that the doctoral candidates must also maintain close professional contact with each other in order to master the scientific challenges posed by their research topic. Teamwork is therefore an essential ingredient of every dissertation in the CRC, which is further intensified by the generally very pleasant climate in discussion between doctoral students and their supervisors. The quality of the results achieved by this interdisciplinary support is excellent.

In order to gain experience in supervising others, doctoral candidates are encouraged to take on the co-supervision of a student during his or her bachelor's or master's thesis.

#### **Promoting the scientific independence of doctoral students**

The supervision structure gives the doctoral students a largely independent position. Scientifically based on their interdisciplinary doctoral projects, they discuss professionally with the CRC members on the various platforms created by the CRC. They have the freedom to invite suitable guests and potential scientific partners for the seminar events via the doctoral speakers. By participating as much as possible in the organisation of the conference-like workshops and summer schools, they can and should assume administrative responsibility. According to financial means, the CRC travel funds are primarily used to present the doctoral students' research results at national and international workshops and conferences. Papers submitted there will, of course, be discussed in the CRC beforehand. Working stays in external groups can also be financed by the CRC. The aim of this funding is to establish a network of contacts between doctoral candidates and to increase their international visibility. The possibility of assuming responsibility for supervising bachelor's and master's students has proven to be a further incentive for independent research.

### **Doctoral Agreement**

A doctoral agreement (according to DFG form 1.90) is made between the supervisors and the doctoral candidates (see figure 3.1)

### **Progress Controls**

The progress controls provided in the supervision agreement take place at the workshops of the CRC, where the doctoral candidates present the status of their work in lectures and in written form. The four-page reports submitted in writing on this occasion are published online.<sup>11</sup> The reports are discussed at the workshops and in the projects, so that the doctoral candidates will receive feedback and criticism from other sources as well as from their supervisors. A certificate is issued for a successful doctorate in the integrated Research Training Group of the CRC 876.

### **Organisational structure of the RTG and possible scopes of design for doctoral students**

The Research Training Group is jointly supported by the CRC's project managers and represented externally by the CRC's spokesperson. The head of the RTG is in charge of day-to-day business and is a member of the CRC's Executive Board, where he or she is particularly involved in representing the interests and the training of the doctoral students. All CRC project managers, whose doctoral candidates have been admitted as doctoral candidates to the RTG irrespective of the source of payment, but with appropriate professional orientation and qualification, are supervisors of the RTG. The doctoral candidates choose a council of three speakers, whereby at least one speaker must be a woman. These speakers can make suggestions for improvements at any time.

They particularly participate in the organisation of the CRC's guest program. The doctoral students are supported by the managing director of the CRC in developing their activities (invitations, the organisation of events). General decisions are made at the meeting of all CRC project managers. Decisions on individual cases are made by the CRC Executive Board. The Research Training Group is always represented by its spokesperson and, as a rule, by the doctoral candidates' spokespersons on the board of the CRC.

The doctoral candidates' spokespersons are not involved in decisions on the orientation of research within the CRC, which could also affect their supervisors. They are also not directly involved in personnel decisions. Meetings of the Executive Board take place monthly during the lecture period and optionally during the lecture-free period. Joint meetings of all supervisors and doctoral students take place at the workshops.

## **3.6 Environment of the Integrated Research Training Group**

**Positioning of the Integrated Research Training Group in relation to other programmes for the promotion of young researchers and structured doctoral programmes.** The career development of the young researchers is embedded in a row of measures and programmes:

- The TU Dortmund, the University of Duisburg-Essen, and the Ruhr-University Bochum merged to form the Ruhr University Alliance (UA Ruhr) in March 2007. The universities bundle their strengths in research and teaching without giving up their independence.
- The universities participating in the CRC 876 attach particular importance to the promotion of young researchers. The TU Dortmund University and the University of Duisburg-Essen

<sup>11</sup><http://sfb876.tu-dortmund.de/FORSCHUNG/techreports.html>

together with the Ruhr-University Bochum support the Science Career Net Ruhr (SCNR). The network offers the *mentoring*<sup>3</sup> mentoring programme for female doctoral students. An annual autumn academy is devoted to questions about doctoral studies and further training. In August 2009, the CoachingPLUS programme was launched, a seminar programme specially tailored to special cultures for advanced young scientists. In addition, SCNR is also offering the Career Forum to develop all aspects of the scientific future of the young scientists since 2010.<sup>12</sup>

- The *Research Academy Ruhr* is one of the largest and most powerful platforms in Germany to support young researchers and to prepare them for careers inside and outside academia. The *Research Academy Ruhr* is sponsored and supported by the three UA Ruhr universities of Bochum, Dortmund, and Duisburg-Essen and is instrumental in the development of UA Ruhr as an academic location. To strengthen the local young researchers from their late master's studies to the early professorships, UA Ruhr created the Research Academy Ruhr. Its synergies offer a locally expanded and networked qualification program, which increases the options and perspective of 10,000 young researchers through specialised qualification and career guidance offers. The *Research Academy Ruhr* combines the know-how of more than ten years of innovative support for young researchers in the Ruhr area. This includes the Ruhr-University Research School as a distinguished and first interdisciplinary graduate school in Germany, the above-mentioned Science Career Net Ruhr as an association for innovative support of all academic career levels, and the Global Young Faculty as an interdisciplinary network of aspiring researchers in the Ruhr area as well as the numerous disciplinary PhD programmes within UA Ruhr. Because the Research Academy Ruhr is a collective platform for supporting young researchers, it creates positive effects in terms of qualification, synergy, and attractiveness of the research location. Furthermore, it fosters the discussion about the status and the profile of PhD students and PhD standards as well as processes of quality assurance.<sup>13</sup>
- As part of the RAR, the Graduate Center (GC) has been created recently at the TU Dortmund University. This center combines all existing, ongoing and future efforts for early qualification and identification of PhD candidates. Courses and consulting services are open to other UA Ruhr PhD candidates and the other university's offers to those in Dortmund alike.
- The continuing education programmes offered by the Center for Higher Education at the TU Dortmund University (Zentrum für Hochschulbildung) are open to young researchers, including seminars on the design of teaching, project management, and life planning.<sup>14</sup>
- The basic equipment of the Research Training Group is planned for subject-specific courses that are not offered by the TU Dortmund University itself, e.g., the RapidMiner Hadoop course.
- Since spring 2010, the doctoral students' forum (ProFor), a coordination office for all doctoral candidates, has been available at the UDE.<sup>15</sup>
- At the TU Dortmund University there are currently the structured graduate programmes listed below, including a DFG Graduate programme and three NRW Forschungsschulen. In addition, there are networks organised within the faculty or across faculties (Faculties 12 - 16). The environment of the RTG of the CRC 876 consists of the following programmes for the promotion of young researchers:

- DFG – Research Training Group Discrete Optimization of Technical Systems under

---

<sup>12</sup><http://www.scn-ruhr.de/uamr.html>

<sup>13</sup><http://www.uamr.org/research-academy-ruhr/>

<sup>14</sup><http://www.zhb.tu-dortmund.de>

<sup>15</sup><http://www.uni-due.de/profor/>

Uncertainty (GRK 1855, 2013 - 2018), Faculty of Computer Science, Speaker: Univ.-Prof. Dr. Peter Buchholz, Chair of Practical Computer Science. In the past, the RTG 1855 offered compact courses on various optimisation methods and the man-machine interface. The doctoral students of both Research Training Groups were encouraged to take part in the compact courses of the other RTG.

- DFG – Research Training Group Adaptation intelligence of factories in a dynamic and complex environment (RTG 2193, since 2016), Faculty of Computer Science, Speaker: Univ. -Prof. Dr. Dr. Jakob Rehof, Chair of Software-Engineering. For the doctoral students the same conditions as for RTG 1855 apply.
- DFG – Coherent manipulation of interacting spin excitations in tailored semiconductors, (TRR 160, since 2015), Physics Faculty, Speaker: Univ. -Prof. Dr. Manfred Bayer, Chair of Experimental Physics II. Concerning research and education, there is no relevant overlap with CRC 876.
- DFG – TU Dortmund University is further participating in the Research Training Group Phenomena of high dimensions in Stochastics - Fluctuations and discontinuity (RTG 2131, since 2015) Faculty of Mathematics, RU Bochum, Speaker: Univ. -Prof. Dr. Peter Eichelsbacher, Chair of Stochastics.
- NRW-Forschungsschule: Energy Efficient Production and Logistics, Ruhr-University Bochum, TU Dortmund University: Site spokesman Prof. Dr. Claus Weihs, Faculty of Statistics. The topic of energy efficiency also concerns the CRC. The Graduate School is an initiative of the Engineering Unit Ruhr, the cooperation platform of the mechanical engineering faculties of the Ruhr-Universität Bochum and the TU Dortmund University, with the participation of Prof. Jochen Deuse (TU Dortmund University, spokesman) and is therefore also linked to the CRC 876 in terms of personnel. Doctoral students of the Research School and PhD students of our Research Training Group have access to the subject-specific courses of the other institution.
- NRW-Forschungsschule: Ruhr Graduate School in Economics RGS Econ Ruhr-University Bochum, University Duisburg-Essen, RWI Rheinisch-Westfälisches-Institut für Wirtschaftsforschung, TU Dortmund University: Spokesman of the University of Applied Sciences Prof. Dr. Wolfgang Leininger, Faculty of Economics and Social Sciences
- Graduate School North Rhine-Westphalia: Graduate Cluster for Industrial Biotechnology TU Dortmund University, University of Bielefeld, Heinrich-Heine-University Düsseldorf, Spokesman: Prof. Dr. -Ing. Gerhard Schembecker, Faculty of Bio- and Chemical Engineering
- Graduate School NRW: didactic development research on diagnosis-guided teaching and learning processes (FUNKEN), Spokesman: Univ. -Prof. Dr. Stephan Hufmann, Dortmund Competence Center for Teacher Education and Teaching/Learning Research (DoKoLL)
- Private industry: Graduate School of Logistics TU Dortmund University, University of Duisburg-Essen, University of Paderborn, WWU Münster, Spokesman: Axel Kuhn, Faculty of Mechanical Engineering, University of Duisburg-Essen, Germany
- Max Planck Society, International Max Planck Research School in Chemical Biology, Max Planck Institute of Molecular Physiology, TU Dortmund University (Faculty of Chemistry and Chemical Biology), Ruhr-University Bochum, Speaker: Dr. Martin Engelhard, Max Planck Institute of Molecular Physiology
- Progress Report on Energy Efficiency in the Neighbourhood – clever supply. conversion.

activation. From July 2014, the Faculties of Spatial Planning, Electrical Engineering and Information Technology as well as the Faculty of Economics and Social Sciences at the Technical University of Dortmund will be participating in the Research Training Group, which is investigating not only technical, but also structural-spatial, economic, legal, and social issues relating to energy efficiency in the quarter.

- In 2009, the universities of the UA Ruhr opened their courses for students from the other universities and launched the e-learning project Ruhr Campus Online.<sup>16</sup>

In particular, the specialist events offered by the Faculties of Computer Science, Statistics, Physics and Logistics in this context are of potential interest to doctoral students from projects of the CRC 876. All these events can of course be attended by doctoral students. Due to the different professional backgrounds of the doctoral students of CRC 876, participation can only be optional and recommended.

**Service of the universities:** The participating universities have committed themselves to providing material and financial support both for the material basic equipment and of crediting the teaching to be provided for the RTG. Of course, rooms and basic equipment are available for the doctoral students. If, after the maximum funding has expired, a doctoral procedure is not yet completed, the participating faculties will ensure the final financing of the project.

**Promotion of doctorates:** The measures to promote families with children have been reported in section 1.4.2 of the proposal. The measures also apply in particular to doctoral candidates.

**Relief of the university teachers involved in the RTG:** The teaching carried out in the Research Training Group, which also includes supervision seminars, is credited against the teaching load. The head of the RTG is relieved by the Faculty of Physics from 2 hours teaching per week.

**Non-university research institutions:** In the field of embedded systems research, collaborations are conducted with the ICD (Informatik Centrum Dortmund e.V.)<sup>17</sup> and IMEC, Leuven, Belgium<sup>18</sup>.

### 3.7 Interim assessment and outlook

In the middle of the second CRC period, the Integrated Research Training Group was evaluated by the doctoral candidates. The key questions of this evaluation were further followed by the spokespersons of the doctoral students.

In these evaluations the interdisciplinary cooperation between the doctoral students was perceived as particularly positive because it allows them to discuss their own work professionally, to look beyond the boundaries of their own research and to become coauthor of joint publications of various projects. In particular, the interdisciplinary orientation of the CRC was seen positive for the unification of expertise in the fields of data analysis and embedded systems.

The *Technical Skill* seminars on R, RapidMiner, and L<sup>A</sup>T<sub>E</sub>Xas well as the technical courses at the summer school and teaching of *Soft Skills* such as the presentation courses or the course on the *The Arrogance Principle* for female members of the CRC were established. The summer schools as a whole were very well received. The doctoral students found the presentations very interesting and also praised the interdisciplinary and international orientation and the exchange with external students.

---

<sup>16</sup><http://www.uamr.de/aims/index.htm>

<sup>17</sup><http://www.icd.de>

<sup>18</sup><http://www.imec.be>

The opportunity to assume responsibility and the opportunities for co-determination of doctoral students were gladly accepted. This includes the organisation of the workshops as well as the possibility to hold their own courses at the workshops and the summer school. Here again, the fact that the events are open to members of all working groups and their students (bachelor/master candidates) and postdocs was found to be very pleasant.

In particular, the *Topical Seminars* were judged positively because of the possibility to invite international guest scientists, thus establishing personal contacts with researchers of international format and opening the possibility to cooperate with these guest scientists. This opportunity was used by more students than in the first phase.

Finally, the doctoral students support the model of deciding on their own responsibility to participate in courses in the RTG within the framework of the doctoral agreement and consider it to be very good.

As foreseen for this RTG period, the doctoral candidates' spokespersons organised meetings of the doctoral candidates at regular intervals.

**Lightning Talks:** For the purpose of supporting exchange between PhD students within the RTG, weekly "blind date lightning talk" meetings were implemented. These meetings aimed to provide a platform for a first contact between PhD students at different stages of their thesis. In these meetings, a small group of fellow PhD students shared and discussed their research topics as well as the latest results in one-minute lightning talks. The subgroup of five students was exchanged each week and selected randomly among the members of the RTG in order to reflect the diversity of research fields within the CRC. One of the graduate speakers was always present to lead the discussion. Furthermore, feedback has been collected about the RTG and the work in the CRC from the PhD students point of view, and suggestions for additional support of the students, e.g., requests for courses, have been gathered. Concerning the latter, successful and efficient publishing, time management, presentation techniques, strategies in research, and career development are still the top-requested topics of training.

**PhD Students Group:** The second goal of the lightning talks was to provide a basis and starting point for fixed collaboration groups of PhD students that provide the space for a more intensive exchange. Aside from that exchange, these groups put focus on dedicated topics that support the work of the students, e.g., time management, situation analysis, or publication strategies. As a result of the lightning talks, a group of PhD students formed that requested to continue with regular meetings and support each other more in their doctoral studies. One challenge for the groups that form was that they must drive themselves. Thus, the participants were asked to list things they need help with and things they can offer to enrich the group. In order to be efficient those meetings have a time limit of 60 to 90 minutes. Possible contents are reports (e.g., short presentations on specific topics) and a review of own progress (e.g., paper, dissertation, software).

Both lightning talk meetings and the PhD students group were perceived as rewarding and worth repeating by the majority of the participants. This concept allows the doctoral candidates' spokespersons to get in contact with the students of the RTG and discuss their needs and opportunities. Thus, there will be an additional iteration of this concept in the third phase.

**CRC workshop as part of the fundamental studies form:** Furthermore, the doctoral students actively reinforce the promotion activities for the CRC in Dortmund for students in the qualification phase, also to increase their share of supervised bachelor's and master's theses. The doctoral candidates' spokespersons therefore organised the 2015 CRC workshop and made it open to students of TU Dortmund. By participating in the workshop, the students collected credit points as part of the fundamental studies form. The lectures were given by mainly the doctoral students and gave the students a good overview of the projects in CRC 876.

Furthermore, discussions with doctoral students show that the core concept with its focal points in interdisciplinary cooperation, scientific visibility, and the individual responsibility of the doctoral students was well received. From the point of view of the supervisors, the RTG was also extremely successful in the training of a technically heterogeneous, but successful research group of doctoral students.

The structured doctoral programme offered in the CRC 876 complies with the requirements of the framework regulations for doctoral studies of the TU Dortmund University. The aspect of internationality in lectures and programme committees, which became relevant only slowly in the start-up phase of the CRC with very young doctoral students, was consistently brought to the fore in the second funding period and shall continue for the third.

## 3.8 Project funding

### 3.8.1 Previous funding

The project has been funded within the Collaborative Research Centre since January 2011.

### 3.8.2 Existing funds for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Funding source
<b>Existing staff</b>						
Research staff	1	Wolfgang Rhode, Prof. Dr. Dr., professor	Astroparticle physics	TU Dortmund	4	Existing funds

**Job descriptions of staff for the proposed funding period (supported through existing funds):**

**1. Rhode, Wolfgang**

MGK administration.

	2019	2020	2021	2022
Existing funds for direct costs	0	0	0	0
Sum of existing funds for direct costs	0	0	0	0
Sum of requested funds for direct costs	51,400	51,400	51,400	51,400

(All figures in euros)

### 3.8.3 Requested funding

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Total	—	0	—	0	—	0	—	0
Direct costs	Sum		Sum		Sum		Sum	
Visiting researchers		27,000		27,000		27,000		27,000
Total		27,000		27,000		27,000		27,000
Fellowships	Sum		Sum		Sum		Sum	
Doctoral researchers		14,400		14,400		14,400		14,400
Total		14,400		14,400		14,400		14,400
Global funds	Sum		Sum		Sum		Sum	
Project-specific workshops		10,000		10,000		10,000		10,000
Total		10,000		10,000		10,000		10,000
<b>Grand total</b>		<b>51,400</b>		<b>51,400</b>		<b>51,400</b>		<b>51,400</b>

(All figures in euros)

### 3.8.4 Requested funding for staff

This project does not request any funding for staff.

### 3.8.5 Requested funding for direct costs

Visiting researchers for financial years 2019–2022

The topical seminar is organised once per week during the lecture period. It provides an important networking measure for graduate students. We expect 30 guests per year, resulting in one guest per week. Given a mix of domestic, European and transatlantic journeys of our guests, we request an average of 900 EUR per person.	EUR	27,000
--	-----	--------

### 3.8.6 Requested funding for fellowships

Fellowships for financial years 2019–2022

Funds are requested for doctoral researchers, while special consideration is given to foreign doctoral researchers. The goal is to simplify the access to the CRC research areas. When available, doctoral researcher position will be offered after successful acquisition of the respective qualification. This position will be financed through either existing or requested funds. Funds are requested for two six-month fellowships per year. Fellows will receive 1.200 EUR per month.	EUR	14,400
---	-----	--------

### 3.8.7 Requested global funds

Project-specific workshops for financial years 2019–2022

A workshop with external participation, yet organised by the CRC, is planned for the first and third year of the funding period. An international summer school is planned for the second and fourth year. These are excellent opportunities for all members to increase the publicity of the CRC. It additionally allows doctoral students to experience and participate in the organisation of a conference. The funds proofed sufficient on the basis of summer schools and workshops during the first and second funding phase.	EUR	10,000
---	-----	--------

### 3.1 General information about Project Z

#### 3.1.1 Project title: Central Tasks of the Collaborative Research Centre

#### 3.1.2 Project leader

Morik, Katharina, Prof. Dr., 14.10.1954, German  
 LS 8, Fakultät für Informatik, Technische Universität Dortmund  
 Otto-Hahn-Straße 12  
 44227 Dortmund  
 Phone: 0231-755-5101  
 E-mail: katharina.morik@tu-dortmund.de

### 3.2 Documentation of the activities of the Collaborative Research Centre

#### Overview of invited Guests

The following lists display guests invited by the Collaborative Research Centre (CRC), grouped by their type of visit. Guests were contextually supervised by project members of the inviting project, supported by members of the management project. The guest's gender is shortened as M/F.

**Summerschool Guests:** The biennially held summerschools are assisted by the invitation of external lecturers. The lectures are contextually chosen according to the topics covered by the CRC. For each lecturer, the lecture topic, and the date of their first lecture is listed. Additionally, the lecturers attending the summerschool 2014 are listed as it was held after the review for phase two of the CRC. The summerschool in 2015 was organised in collocation with the ECMLPKDD 2015 and lecturers external to the CRC were therefore not funded by the CRC.

**Céline Robardet:** F, INSA-Lyon, France, 29.09.2014: Mining patterns in attributed dynamic graphs

**Christian Bauckhage:** M, University of Bonn, Germany, 29.09.2014: From k-means clustering to DESICOM: Matrix Factorization for Data Analysis

**John Duchi:** M, Stanford University, USA, 30.09.2014: Privacy Aware Learning

**Rich Caruana:** M, Microsoft Research, USA, 01.10.2014: Model Compression

**Wayne Luk:** M, Imperial College London, England, 25.09.2017: Field-programmable technology and machine learning

**Rob Maunder:** M, University of Southampton, England, 25.09.2017: Generalised LDPC architectures for high-throughput FPGA realisation

**Rakesh Agrawal:** M, Data Insights Laboratories, San Jose, USA, 26.09.2017: A Tale of a Quest for Business Intelligence from Social Data

**Chris Schwiegelsohn:** M, La Sapienza University, Rome, Italy, 26.09.2017: Graph Streaming

**Guests of the Integrated Graduate School Topical Seminar:** An integral part of the integrated research training group is invitation of guests by doctoral researchers. The following list shows, chronologically ordered, the guests, who were invited within the seminar. For every guest, the date of their presentation is shown. Usually, each guest stayed 2 days, on which the host and other interested project members had the opportunity to discuss the topics of their presentation.

**David Atienza:** M, EPFL, Switzerland, 19.02.2015: Thermal-Aware Design of 2D/3D Multi-Processor System-on-Chip Architectures

**Hannes Mühlleisen:** M, CWI Amsterdam, Netherlands, 09.04.2015: Opening the SQL Kingdom to the R-ebels

**Santiago Pagani:** M, Karlsruhe Institute of Technology, Germany, 28.05.2015: Thermal-Aware Power Budgeting and Transient Peak Computation for Dark Silicon Chip

**Volker Markl:** M, TU Berlin University, Germany, 25.06.2015: Apache Flink and the Berlin Big Data Center

**Giorgio Patrini:** M, Australian National University, Australia, 27.08.2015: Learning with Label Proportions (LLP)

**Ingo Müller:** M, SAP Software Solutions, Germany, 29.10.2015: Cache-Efficient Aggregation: Hashing Is Sorting

**Simona Constantinescu:** F, ETH Zurich, in Switzerland, 05.11.2015: Waiting Time Models for Mutual Exclusivity and Order Constraints in Cancer

**Karsten Borgwardt:** M, ETH Zurich, in Switzerland, 12.11.2015: Significant Pattern Mining

**Thorsten Joachims:** M, Cornell University, USA, 19.11.2015: From Average Treatment Effects to Batch Learning from Bandit Feedback

**Kerstin Eder:** F, University of Bristol, 26.11.2015: Whole Systems Energy Transparency (or: More Power to Software Developers!)

**Giovanni de Micheli:** M, EPF Lausanne, Italy, 14.01.2016: Nano-Tera.ch: Electronic Technology for Health Management

**He Sun:** M, University of Bristol, England, 17.03.2016: Graphs, Ellipsoids, and Balls-into-Bins: A linear-time algorithm for constructing linear-sized spectral sparsification

**Jilles Vreeken:** M, Saarland University, 14.04.2016: Discovering Compositions

**Xue Liu:** M, McGill University, Canada, 19.04.2016: When Bits meet Joules: A view from data center operations' perspective

**Marius Kloft:** M, HU Berlin, Germany, 21.04.2016: Kernel-based Machine Learning from Multiple Information Sources

**Sasa Misailovic:** M, ETH Zürich, Switzerland, 28.04.2016: Analysis and Optimization of Approximate Programs

**Pieter Mestdagh:** M, University of Ghent, Belgium, 19.05.2016: Reprocessing and analysis of high-throughput data to identify novel therapeutic non-coding targets in cancer

**Mark Huber:** M, Claremont Meckenna College, USA, 23.06.2016: Perfect simulation algorithms give a method for sampling exactly from high dimensional distributions.

**Cong Liu:** M, University of Texas at Dallas, USA, 30.06.2016: Predictable Real-Time Computing in GPU-enabled Systems

**Mathias Niepert:** M, NEC Labs Europe Heidelberg, Germany, 21.07.2016: Deep Learning for Big Graph Data

**Wolfgang Rosenstiel:** M, University of Tuebingen, Germany, 28.07.2016: Applications of Machine Learning: From Brain-Machine-Interfaces to Autonomous Driving

**Luca Benini:** M, University of Bologna, Italy, 08.09.2016: Toward zero-power sensing: a transient computing approach

**Paolo Cumani:** M, IFAE Barcelona, Spain, 29.09.2016: The Cherenkov Telescope Array (CTA)

**Bart Goethals:** M, University of Antwerp, Belgium, 20.10.2016: k-Morik: Mining Patterns to Classify Cartified Images of Katharina

**Arno Siebes:** M, University of Utrecht, Netherlands, 20.10.2016: Sharing Data with Guaranteed Privacy

**Jana Giceva:** F, ETH Zürich Systems Group, Switzerland, 27.10.2016: Customized OS support for data processing on modern hardware

**Dirk Koch:** M, Technologies Group University of Manchester, England, 10.11.2016: Runtime Reconfigurable Computing - from Embedded to HPC

**Jan Reineke:** M, Saarland University, Germany, 17.11.2016: On the Smoothness of Paging Algorithms

**Yung-Hsiang Lu:** M, Purdue University, USA, 15.12.2016: Opportunities and Challenges in Global Network Cameras

**Rohit Babbar:** M, Max Planck Institute Tuebingen, Germany, 26.01.2017: Scalable Algorithms for Extreme Multi-class and Multi-label Classification

**Eirini Ntoutsi:** F, Leibniz University Hannover, Germany, 09.02.2017: Learning over high dimensional data streams

**Daniela Huppenkothen:** F, New York University, USA, 04.05.2017: How to Time a Black Hole: Time series Analysis for the Multi-Wavelength Future

**David Woodruff:** M, IBM Almaden Research Center, USA, 08.05.2017: Sketching as a Tool for Geometric Problems

**Luis Moreira-Matias:** M, NEC Laboratories Europe Heidelberg, Germany, 11.05.2017: Real-Time Mobility Data Mining

**Volker Tresp:** M, LMU Munich/Siemens, Germany, 18.05.2017: Learning with Knowledge Graphs

**Pascal Schweitzer:** M, RWTH Aachen University, Germany, 24.05.2017: Algorithmic Symmetry Detection and Exploitation

**Martin Atzmueller:** M, Tilburg University, Netherlands, 01.06.2017: Complex Network Mining on Digital and Physical Information Artefacts

**Jakob Rehof:** M, TU Dortmund University, Germany, 08.06.2017: Probabilistic Program Induction = program synthesis + learning?

**Bashir Al-Hashimi:** M, University of Southampton, England, 22.06.2017: Runtime management for many core embedded systems: the PRiME approach

**Eyal Rozenberg:** M, CWI Amsterdam, Netherlands, 29.06.2017: GPU and coprocessor use in analytic query processing - Why we have only just begun

**Sarmila Dhandapani:** F, Bannari Amman Institute of Technology, India, 29.09.2017: Ultra-Low-Power Wireless Communication Using IR-UWB

**Rameshwari Ramasamy:** F, Bannari Amman Institute of Technology, India, 29.09.2017: Ultra-Low-Power Wireless Communication Using IR-UWB

**Giovanni Beltrame:** M, Polytechnique Montreal, Canada, 27.10.2017: How to program 1000 Robots?

**Thomas Kipf:** M, University of Amsterdam, Netherlands, 30.11.2017: End-to-end learning on graphs with graph convolutional networks

**Samarjit Chakraborty:** M, University of Munich, Germany, 25.01.2018: Resource-Aware Cyber-Physical Systems Design

**Lisa McShane:** F, National Institutes of Health, USA, 22.03.2018: Analysis of high-dimensional Data: Opportunities and challenges

**Tomasz Burzykowski:** M, Hasselt University, Belgium, 22.03.2018: A bird's eye view on processing and statistical analysis of 'omics' data

**Willi Sauerbrei:** M, University of Freiburg, Germany, 22.03.2018: Short introduction of the STREngthening Analytical Thinking for Observational Studies (STRATOS) initiative; Guidance for the selection of variables and functional form for continuous variables - Why and for whom?

**Riccardo de Bin:** M, University of Oslo, Norway, 22.03.2018: Strategies to derive combined prediction models using both clinical predictors and high-throughput molecular data

**Catherine McGeoch:** F, D-Wave Systems, Canada, 10.04.2018: (D-Wave Systems): Performance Evaluation for Annealing Based Quantum Computers

**Marco Zimmerling:** M, TU Dresden, Germany, 17.05.2018: From Best-effort Monitoring to Feed-back Control: How Synchronous Transmissions Enable the Future Internet of Things

**Silvio Lattanzi:** M, Google Zürich, Switzerland, 24.05.2018: Consistent k-Clustering

**Tina Eliassi-Rad:** F, Northeastern University Boston, USA, 05.06.2018: Just Machine Learning

#### **Research visits:**

**Happy Mittal:** M, Indian Institute of Technology, Delhi, India, October - December 2016: Symmetries in statistical relational learning.

Happy Mittal's research focused on probabilistic graphical models and statistical relational learning. His publications at the NIPS 2015 (Lifted Inference Rules With Constraints) and 2014 (New Rules for Domain Independent Lifted MAP Inference) showed a promising relation to the work carried out in the projects A6 and B4.

Under supervision by Kristian Kersting, he focused on the analysis of the visual genome. The visual genome is a knowledge base, which combines 4.2 million regions in 100,000 pictures with 2.1 object instances with 1.8 million attributes. This may allow combining structured data (pictures) with natural language to approach physical symbol grounding. Research focus have been relational models, fractional automorphisms, graph kernels and lifted probabilistic

inference. For projects A6 and B4 this made further data accessible: Large graphs with complex attributes of everyday processes and bag-of-objects representations similar to traffic count data.

**Mostafa Jafari Nodoushan (pending Visa application):** M, Sharif University of Technology, Tehran, Iran, October 2018 - March 2019: Calculus of variations for energy efficient DVFS scheduling.

Embedded systems usually have severe limitations on power consumption. Some reasons are: 1) Many embedded systems are battery-based and because of nonlinear battery behaviour, the battery lifetime depends on how the system pulls power from the battery, 2) Embedded systems are usually fan-less with limited cooling equipment, therefore temperature management is a prominent issue in these systems which can be done via controlling the power consumption, 3) Many embedded systems are used in safety-critical applications where high reliability is crucial. Power consumption has considerable impacts on the reliability of digital systems and both transient (soft errors) and permanent faults (aging) depend on the way the system dissipates power.

A branch of mathematics is calculus of variations (CoV), where system parameters are considered as continuous functions of time. It is planned to use the CoV to address the problem of system-level power management in embedded systems. Using CoV, we consider supply voltage, supply current, operational frequency, and signal activity as continuous functions of time, and we try to provide system level techniques to shape these functions (determine the right curve) to achieve objectives such as improved battery lifetime, improved reliability, and reduced temperature.

### Funds for gender equality and family-friendliness

**Gender survey:** Attracting women is still an important goal, especially for a high-profile fundamental research facility as a Collaborative Research Centre. At the start of the second funding phase in 2015 the CRC conducted a survey with women on all levels, from graduate to principal investigator. Purpose of the survey was to find answers on the motivation behind pursuing a PhD as well as identifying necessary prerequisites to create an environment for gender equality.

**Mentoring<sup>3</sup>:** In order to promote the careers of female scientists, the UA Ruhr University Alliance provides the programme *mentoring<sup>3</sup>*. The programme consists of the components of individual mentoring using personal mentors, seminars about interdisciplinary competences, networking, and workshops.

During the first funding phase of the CRC, a dedicated group of researchers was organised around their specific needs. Due to the tremendous success and positive feedback, for the second phase this concept is repeated with a new group.

The CRC-funds for gender equality have been used to support this group with a dedicated central coordinator and mentor as well as guest speakers and trainers for career advancing topics.

**Child care:** Initiated by the Dortmund Collaborative Research Centre 823, a child care facility for children up to age 3 of parents working in DFG-funded projects was launched in 2011. Since then, a continuous demand for care for CRC 876 researcher's children proved the importance of this extended child care unit.

Personal costs are financed from the CRC equality and family-friendliness funds, remaining non-personnel costs (Material, rent, insurances, around 70% of overall costs) are financed from existing TU Dortmund University funds.

**Arrogance training:** The arrogance training is a course specialised on the interaction of highly qualified women with men, especially in environments where an imbalance in numbers towards men prevails. The course trains interactions with colleagues and supervisors with practical exercises and sessions with sparring partners.

Participation in this course was already offered during the first phase in the CRC and due to its applicability was repeated in 2017.

### **Central activities**

Central coordinated activities, beside the regular topical seminars as documented in project MGK, of the CRC are the periodic workshops of the graduates as well as the summer schools:

**Workshop 26.02.2015:** Review of first funding phase results, lessons learned during preparation for the second phase and outlook on upcoming research.

**Summer School 02.-05.09.2015:** Summer school in collocation with ECMLPKDD 2015 and joint organisation with João Gama and Alípio Jorge.

**Workshop 28.-30.09.2015:** Christian Sohler and his group have organised the 2nd European Meeting on Algorithmic Challenges of Big Data (ACBD 2015) in Dortmund. The goal of the workshop has been to consolidate the European research community in this area and discuss recent scientific advances. Close to 20 leading European researchers gave talks on topics related to the analysis of massive data sets. Topics included external memory algorithms, local and distributed algorithms, online algorithms, clustering algorithms, sketching, and many more.

**Workshop 24.-26.02.2016:** Workshop of the integrated graduate school for reports of the graduates in their respective projects. Additionally, this workshop was opened for students of all disciplines as part of the course program *Studium Fundamentale*. Courses of this type are mandatory for students of computer science, bio- and chemical engineering, electrical engineering and information technology, city and regional planning, and journalism. Goal was to use the interdisciplinary orientation of the CRC's research to broaden the view of students as well as CRC members.

**Workshop 22.10.2016:** Flash talks to second phase highlights as well as plans and outlook for third phase proposal.

**Workshop 05.07.2017:** Presentation of sketches for research in the third phase.

**Summer School 25.09.-28.09.2017:** Fourth Summer School on Resource-Aware Machine Learning: From deep learning, probabilistic graphical models to ultra low power learning. A one and a half-day hands-on session let the participants explore real-life sensor data and deploy trained models on extremely limited devices.

**Workshop 16.10.2017:** Second presentation on project progress and plans for third phase proposal.

## Lump-Sum funds

The lump-sum funds are used for project startup-support, public relations and publications.

**Brochure about the CRC:** Initially created during the first funding phase, the brochure about the CRC's overall research theme as well as each single project, got updated to reflect the changes in topics and projects. The design for the brochure already considered this, as the project's description were separated from the immutable theme of the CRC as a whole. Each project flyer was redesigned for the second phase. Additionally, because of the positive reception of the brochure and to increase international general audiences, an English version was designed from the ground up. With the title *Big Data - Small Devices* the brochure presents the CRC's range from data analysis to cyber-physical systems.

**German Innovation New York:** Beside the individual activities of principal investigators and researchers to present the CRC to the public, a dedicated event was organised at the German Center for Research and Innovation New York on March 7th 2016. Presentations covered the overall CRC research focus and enabled discussion about current trends with Claudia Perlich (Chief Data Scientist Dstillery) and Tina Eliassi-Rad (Northeastern University Boston).

**Open Access Publications:** Most publications are still traditionally published in conference proceedings and subscription journals. Recent trends show an increasing amount of high-rated open access proceedings and journals, where the publication costs are covered by the authors. In cases where a project could not use surplus funds, e.g., costs of PhD students below DFG funding, the lump-sum funds have been used to cover these costs. For the future an ongoing increase of open access publications is expected.

**Starting funds:** The overall structure of the collaborative research is planned to remain mainly constant for the proposed third phase, therefore a low need arised for starting funds to accelerate projects. Notable exception was a joint publication in project B3. Here the proposed new principal investigator, Petra Wiederkehr, and Felix Finkeldey from her group, published an article in 2018 at the CIRP conference on Manufacturing Systems together with Katharina Morik and Amal Saadallah. Travel for presentation at the conference has been supported by the lump-sum funds.

### 3.3 Project funding

#### 3.3.1 Existing funds for the new funding period

	No.	Name, academic degree, position	Field of research	Department of university or non-university institution	Project commitment in hours per week	Funding source
<b>Existing staff</b>						
Research staff	1	Katharina Morik, Prof. Dr., professor	Data mining	TU Dortmund	4	Existing funds
Non-research staff	2	Andreas Greve, non-research staff	—	TU Dortmund	10	Existing funds

#### Job descriptions of staff for the proposed funding period (existing funds):

##### 1. Morik, Katharina

Spokesperson of the CRC 876, who is responsible for the overall organisation and external representation of the Collaborative Research Centre. She initiates international and national contacts and is responsible for CRC-internal cooperation.

##### 2. Greve, Andreas

Support of the technical infrastructure for data exchange and the installation of application programs. Maintenance of web presentation and the CRC cluster.

	2019	2020	2021	2022
TU Dortmund: existing funds from University	3,000	3,000	3,000	3,000
Sum of existing funds	3,000	3,000	3,000	3,000
Sum of requested funds	230,700	230,700	230,700	230,700

(All figures in euros)

### 3.3.2 Requested funding for the new funding period

Funding for	2019		2020		2021		2022	
Staff	QTY	Sum	QTY	Sum	QTY	Sum	QTY	Sum
Coordination, 100 %	1	69,900	1	69,900	1	69,900	1	69,900
Coordination, 50 %	1	24,000	1	24,000	1	24,000	1	24,000
Student assistants	16.0	168,000	16.0	168,000	16.0	168,000	16.0	168,000
Substitute positions	0.5	35,000	0.5	35,000	0.5	35,000	0.5	35,000
Total	—	296,900	—	296,900	—	296,900	—	296,900
Direct costs	Sum		Sum		Sum		Sum	
Travels	87,500		87,500		87,500		87,500	
Visiting researchers	6,000		6,000		6,000		6,000	
Total	93,500		93,500		93,500		93,500	
Global funds	Sum		Sum		Sum		Sum	
Coordination	7,200		7,200		7,200		7,200	
Gender equality measures	30,000		30,000		30,000		30,000	
Lump-sum funds	100,000		100,000		100,000		100,000	
Total	137,200		137,200		137,200		137,200	
<b>Grand total</b>	<b>527,600</b>		<b>527,600</b>		<b>527,600</b>		<b>527,600</b>	

(All figures in euros)

### 3.3.3 Requested funding for staff

#### Student assistants for financial years 2019–2022

Student assistants also support the practical work in the projects. For this purpose, each project and principal investigator (with the exception of industrial partners) should receive funding for a student assistant with 7 hours per week corresponding to half a position. In addition, projects Z and MGK each receive funds for administrative tasks amounting to 24 hours per week, corresponding to 1.5 positions. During the last two funding phases student assistants did not only prove extremely successful in supporting the work in each project, but also provided a major opportunity to integrate qualified students early into the CRC.	EUR	168,000
--	-----	---------

#### Coordination for financial years 2019–2022

The financial administration (material and personnel resources, booking of invoices, preparation of requests for funds and proof of use) and the general administration (organisation of colloquia, workshops and summer schools regarding the booking of rooms, invoicing guests, correspondence by e-mail and directly on the internet, predominantly in English) is handled by this position.	EUR	24,000
--	-----	--------

## Project Z

### Coordination for financial years 2019–2022

Managing Director, who is responsible for communication between the projects and the integrated Graduate School as well as the coordination of the scientific activities. He is controlling submission of reports, courses and seminars in the Graduate School and monitors the internet presence. In particular, he supervises the information technology equipment in the computation cluster, for data storage, bibliographies, software repositories and internal collaboration tools. He organises the working groups with the references to projects and the Graduate School and prepares scientific courses as well as summer schools. He is responsible for the controlling of finances and creates the reports with colleagues from the administration. Stefan Michaelis joined the CRC shortly after the start of the first phase and has a deep knowledge of the structures and collaborations inside the CRC.	EUR	69,900
---	-----	--------

### Substitute positions for financial years 2019–2022

A principal investigator can apply to the board of management for a research semester, whereby the approval of the faculty, a concrete work plan and a list of possible substitute representatives must be submitted. The planned work at another location should bring the CRC a considerable additional amount of exposure. The project leader can be represented by an external colleague during his or her research term, so that the work in the CRC is enriched. It can also be taken up by postdocs who can prove their qualification for a professorship in this way. In this way, we promote young scientists and broaden their perspectives after their doctorate. A substitute for 1 semester per year is paid according to E14.	EUR	35,000
---	-----	--------

### 3.3.4 Requested funding for direct costs

#### Travels for financial years 2019–2022

Participation in conferences abroad is of great importance for all members of the CRC. For example, the following meetings are important for subjecting one's own work to international criticism and for exchanging ideas with colleagues: ICDM, KDD, ICML, ECML PKDD, DATE, VLSI Design, HIPEAC, HICSS, WADS, VISAPP, WABI, COMPSTAT. In addition, there are workshops that are of particular interest to doctoral students. If we only consider the persons financed by the supplementary equipment, the quoted sum results from a mixture of European and American or Asian trips including the conference fees. The overall amount is based on the experience of expenditures during the second phase of the CRC.	EUR	87,500
--	-----	--------

#### Visiting researchers for financial years 2019–2022

The possibility for colleagues to spend their sabbatical at the CRC, in which they receive no salary from their university, is to be made possible in addition to the Topical Seminar. For a longer stay of an American colleague during his research semester we assume 4.000,-EUR per month.	EUR	6,000
--	-----	-------

### 3.3.5 Requested global funds

Coordination for financial years 2019–2022

For office supplies and printing costs we calculate 200,-EUR per project and principle investigator annually.	EUR	7,200
---	-----	-------

Gender equality measures for financial years 2019–2022

To complement the university's efforts, additional childcare measures are organised and financed for workshops or similar events of the CRC, for business trips of CRC scientists, and for evening events of the CRC. Special training courses on scientific careers for women are intended to support female scientists. A third aspect concerns security during business trips, for which increased travel costs may also become necessary due to the use of taxis or booking of the conference hotel.	EUR	30,000
--	-----	--------

Lump-sum funds for financial years 2019–2022

The lump-sum funds are earmarked for unforeseeable changes when the proposal is submitted. These can be new professor positions or the integration of young scientists, but also devices that have become important for projects as a result of technical progress. Additionally, in the previous phases the funds supported open access publications and the public relation activities like the CRC brochure targeted to the general public.	EUR	100,000
--	-----	---------



## **4 Bylaws of the Collaborative Research Centre**

### **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 Vollertrages, 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 Vorstandentscheid 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 Declaration on working space for the Collaborative Research Centre

Is the existing office and/or lab space sufficient to accommodate the Collaborative Research Centre at the time of submitting the proposal? Yes

Will there be sufficient office and/or lab space to accommodate the Collaborative Research Centre including any planned extensions in the financial years...

- |           |     |
|-----------|-----|
| ... 2019? | Yes |
| ... 2020? | Yes |
| ... 2021? | Yes |
| ... 2022? | Yes |

Dortmund, 12/7/18



Albrecht Ehlers  
(Chancellor)



## 6 Declaration on lists of publications

We hereby declare that the lists of publications included in this proposal and in the attached research profiles of the principal investigators/project leaders were compiled in accordance with the rules of the DFG concerning publication lists.

Dortmund, 12.07.2018

  
\_\_\_\_\_  
Katharina Morik

Prof. Dr. Katharina Morik  
(Spokesperson of Collaborative Research Centre)

Dortmund, 12.07.2018

  
\_\_\_\_\_  
Ursula Gather

Prof. Dr. Dr. h.c. Ursula Gather  
(Rector of applicant university)