DAGSTUHL
REPORTS

Volume 1, Issue 8, August 2011

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany.
Online available at http://www.dagstuhl.de/dagrep

Publication date
December, 2011

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `http://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.1.8.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e.g. summaries from panel discussions or open problem sessions.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de

www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 11321

Information Management in the Cloud

Edited by

Anastassia Ailamaki[1], Michael J. Carey[2], Donald Kossmann[3], Steve Loughran[4], and Volker Markl[5]

1 **EPFL – Lausanne, CH,** anastasia.ailamaki@epfl.ch
2 **University of California – Irvine, US,** mjcarey@ics.uci.edu
3 **ETH Zürich, CH,** donald.kossmann@inf.ethz.ch
4 **HP Lab – Bristol, GB,** steve.loughran@hp.com
5 **TU Berlin, DE,** volker.markl@tu-berlin.de

Abstract

Cloud computing is emerging as a new paradigm for highly scalable, fault-tolerant, and adaptable computing on large clusters of off-the-shelf computers. Cloud architectures strive to massively parallelize complex processing tasks through a computational model motivated by functional programming. They provide highly available storage and compute capacity through distribution and redundancy. Most importantly, Cloud architectures adapt to changing requirements by dynamically provisioning new (virtualized) compute or storage nodes. Economies of scale enable cloud providers to provide compute and storage powers to a multitude of users. On the infrastructure side, such a model has been pioneered by Amazon with EC2, whereas software as a service on cloud infrastructures with multi-tenancy has been pioneered by Salesforce.com.

The Dagstuhl Seminar 11321 "Information Management in the Cloud" brought together a diverse set of researchers and practitioners with a broad range of expertise. The purpose of this seminar was to consider and to discuss causes, opportunities, and solutions for technologies, and architectures that enable cloud information management. The scope ranged from web-scale log file analysis using cluster computing techniques to dynamic provisioning of resources in data centers, covering topics from the areas of analytical and transactional processing, parallelization of large scale data and compute intensive operations as well as implementation techniques for fault tolerance.

Seminar 07.–12. August, 2011 – www.dagstuhl.de/11321
1998 ACM Subject Classification H.0 [Information Systems] General
Keywords and phrases Cloud Technologies, Information Management, Distributed Systems, Parallel Databases
Digital Object Identifier 10.4230/DagRep.1.8.1

1 Executive Summary

Anastassia Ailamaki
Michael J. Carey
Donald Kossmann
Steve Loughran
Volker Markl

Cloud computing is emerging as a new paradigm for highly scalable, fault-tolerant, and adaptable computing on large clusters of off-the-shelf computers. Cloud architectures strive to massively parallelize complex processing tasks through a computational model motivated by functional programming. They provide highly available storage and compute capacity through distribution and redundancy. Most importantly, Cloud architectures adapt to changing requirements by dynamically provisioning new (virtualized) compute or storage nodes. Economies of scale enable cloud providers to provide compute and storage powers to a multitude of users. On the infrastructure side, such a model has been pioneered by Amazon with EC2, whereas software as a service on cloud infrastructures with multi-tenancy has been pioneered by Salesforce.com.

The Dagstuhl seminar on Information Management in the Cloud brought together a diverse set of researchers and practitioners with a broad range of expertise. The purpose of this seminar was to consider and to discuss causes, opportunities, and solutions for technologies, and architectures that enable cloud information management. The scope ranged from web-scale log file analysis using cluster computing techniques to dynamic provisioning of resources in data centers, covering topics from the areas of analytical and transactional processing, parallelization of large scale data and compute intensive operations as well as implementation techniques for fault tolerance.

The seminar consisted of keynotes, participant presentations, demos and working groups. The first two seminar days consisted of a keynote by Helmut Krcmar on “Business Aspects of Cloud Computing” as well as 33 short presentations on various aspects of cloud computing. On the evening of the second day, the participants formed working groups on economic aspects, programming models, benchmarking. The third day of the seminar consisted of two keynotes, by Dirk Riehle on “Open Source and Cloud Computing" and by Donald Kossmann on “Benchmarking”. After these keynotes, working groups discussed their respective topics. In the evening, an industrial panel with Miron Livny, Steve Loughran, Sergey Melnik, Russell Sears, and Dean Jacobs discussed research challenges in Cloud Computing from an industrial point of view. On the fourth day, a keynote by Miron Livny discussed Cloud Computing from a distributed systems and high-performance computing point of way. After the keynote, a demo session presented the following systems:

- HyPer: A Cloud-scale Main Memory Database System (Team from TUM)
- Asterix and Hyrax (Team from UCI)
- Stratosphere (Team from TU Berlin, HU Berlin and HPI)
- Myriad Parallel Data Generator (Team from TU Berlin)

After these demos, working groups continued during the day and presented their results in the evening. The last day of the seminar, participants continued in working groups and discussed further collaborations with respect to papers and project proposals. During this

day, several abstracts for papers have been prepared, and discussions about several joint research project proposals have started.

The organizers hope that the seminar has helped to organize the research space in cloud computing and identified new research challenges. We look forward towards research collaborations and papers that were bootstrapped during this intensive week.

2 Table of Contents

3 Overview of Talks

This section lists the talks and abstracts of all seminar participants. The titles and abstracts were taken from the seminar's material web site whenever available.

3.1 Facilitating Scientific Analytics in the Cloud

Magdalena Balazinska (University of Washington – Seattle, US)

Sciences are becoming increasingly data rich and data analysis is becoming the bottleneck to discovery. The cloud holds the promise to facilitate large-scale data analysis because it provides easy access to compute resources and data management software with a flexible pay-as-you-go charging mechanism. There are, however, several challenges in leveraging the cloud for scientific analytics. We discuss three challenges in this talk. First, it is extremely challenging to get high-performance from today's data management systems out-of-the box. Second, data management systems can be hard to use even after the cloud takes away the installation and operations tasks. Finally, the interplay between data management and cloud economics raise several interesting new challenges and opportunities. In this talk, we will explain these three challenges and present recent research results from the database group at the University of Washington related to addressing them.

3.2 Web Data Cleaning

Felix Naumann (Hasso Plattner Institut – Potsdam, DE)

The wealth of freely available, structured information on the Web is constantly growing. Driving domains are public data from and about governments and administrations, scientific data, and data about media, such as articles, books and albums. In addition, general-purpose datasets, such as DBpedia and Freebase from the linked open data community, serve as a focal point for many data sets. Thus, it is possible to query or integrate data from multiple sources and create new, integrated data sets with added value. Yet integration is far from simple: It happens at technical level by ingesting data in various formats, at structural level by providing a common ontology and mapping the data source structures to it, and at semantic level by linking multiple records about same real world entities and fusing these representations into a clean and consistent record. The talk highlights the extreme heterogeneity of web data and points to three research directions: (i) Domain-specific Integration Projects, such as govwild.org, (ii) ad-hoc and declarative data cleansing, such as in the Stratosphere project, and (iii) dynamic provisioning of Linked Data in a Data as a Service (DaaS) fashion.

3.3 Cloud Computing Support for Massively Social Gaming

Alexandru Iosup (TU Delft, NL)

Cloud computing is an emerging commercial infrastructure paradigm that promises to eliminate the need for maintaining expensive computing hardware. Through the use of virtualization and resource time-sharing, Infrastructure as a Service (IaaS) clouds address with a single set of physical resources a large user base with different needs. Similarly, Platform as a Service (PaaS) clouds focus on providing platforms that address each the needs of a large and varied community. Thus, clouds promise to enable for their owners the benefits of an economy of scale and, at the same time, reduce the operating costs for many applications. For example, clouds may become for scientists an alternative to clusters, grids, and parallel production environments. In this presentation we focus on three main research questions related to cloud computing:

1. What is the performance of virtualized cloud resources, as perceived by their users? Many production clouds, including some of the largest publicly-accessible commercial clouds such as the Amazon Web Services and the Google App Engine, use virtualized resources to address diverse user requirements with the same set of physical resources. Virtualization can introduce performance penalties, either due of the additional middleware layer or to the interaction of workloads belonging to different virtual machines. Do virtualized resources deliver the same performance regardless of the application? In particular, are applications affected by execution on virtualized resources? We present here our findings from a large-scale performance evaluation study that focuses on four commercial IaaS clouds.
2. What guarantees do we have about the good performability of clouds over long periods of time? A major impediment to cloud adoption at large is their perceived instability, due, in lack of hard evidence, to novelty ("clouds are a technology too immature to be reliable"). Even if a cloud is avilable and works well today, it may well happen that it will not tomorrow. Does performance change over time (for the worse)? Are clouds really available all the time? We present here our findings from a long-term performance evaluation study that focuses on two commercial clouds, one IaaS and one PaaS.
3. Which new applications can make use of clouds? (By new applications we understand applications with a workload different from the applications of the past, including the workloads typical for grid computing.) Commercial clouds are new to the public. What applications that we could not previously afford to run are now enabled by clouds? What applications can function well under the availability and performance profiles of the current production cloud services? We focus in this presentation on Massively Multiplayer Online Games (MMOGs) and Massively Social Games (MSGs), which have recently emerged as a novel Internet-based entertainment application. Hundreds of MMOGs and MSGs already serve over a quarter of a billion paying customers world-wide, with virtual worlds such as World of Warcraft, FarmVille, and Runescape hosting daily several millions of players. These players want fast-paced entertainment delivered through the Internet, which raises important content and resource requirements; when these are not met in full and on time, players are likely to quit. However, the current industry approach in addressing these requirements, of building and maintaining large data centers, has high cost and limited scalability. The high cost makes the market inaccessible for amateur and small game developers. The limited scalability means that even the largest game

developers cannot support this rapidly growing community. We present our early results in understanding if IaaS and PaaS clouds can provide a scalable, dependable, yet low-cost computational technology for MMOGs and MSGs.

The loosely coupled team who has done the work presented here is: Undergraduate Students at TU Delft: Martin Biczak, Arnoud Bakker, Nassos Antoniou, Thomas de Ruiter, etc. Graduate students at TU Delft: Siqi Shen, Nezih Yigitbasi, Ozan Sonmez. Staff at TU Delft: Henk Sips, Dick Epema, Alexandru Iosup. Collaborators Ion Stoica and the Mesos team (UC Berkeley), Vlad Nae, Thomas Fahringer, Radu Prodan (U. Innsbruck), Nicolae Tapus, Mihaela Balint, Ad. Lascateu, Vlad Posea (UPB), Derrick Kondo, Emmanuel Jeannot (INRIA), etc.

3.4 Genome Data Preprocessing with MapReduce

Keijo Heljanko (Helsinki University of Technology, FI)

We describe joint work between Aalto University and CSC done to visualize genomic data using our new preprocessing tool based on the MapReduce programming framework. The work uses the Apache Hadoop system to build a tool for preprocessing sequence alignment map in BAM file format resulting in an opensource tool Hadoop-BAM. We also describe future directions on research in the area.

3.5 HyPer: Hybrid OLTP & OLAP High-Performance Database System

Alfons Kemper (TU München, DE)

The HyPer prototype demonstrates that it is indeed possible to build a main-memory database system that achieves world-record transaction processing throughput and best-of-breed OLAP query response times in one system in parallel on the same database state. The two workloads of online transaction processing (OLTP) and online analytical processing (OLAP) present different challenges for database architectures. Currently, users with high rates of mission-critical transactions have split their data into two separate systems, one database for OLTP and one so-called data warehouse for OLAP. While allowing for decent transaction rates, this separation has many disadvantages including data freshness issues due to the delay caused by only periodically initiating the Extract Transform Load-data staging and excessive resource consumption due to maintaining two separate information systems. We present an efficient hybrid system, called HyPer, that can handle both OLTP and OLAP simultaneously by using hardware-assisted replication mechanisms to maintain consistent snapshots of the transactional data (see the figure on the right). HyPer is a main-memory database system that guarantees the full ACID properties for OLTP transactions and executes OLAP query sessions (multiple queries) on arbitrarily current and consistent snapshots. The utilization of the processor-inherent support for virtual memory management (address translation, caching, copy-on-write) yields both at the same time: unprecedentedly high transaction rates as high

as 100000 per second and very fast OLAP query response times on a single system executing both workloads in parallel. The performance analysis is based on a combined TPC-C and TPC-H benchmark.

We have developed the novel hybrid OLTP & OLAP database system HyPer that is based on snapshotting transactional data via the virtual memory management of the operating system. In this architecture the OLTP process owns the database and periodically (e.g., in the order of seconds or minutes) forks an OLAP process. This OLAP process constitutes a fresh transaction consistent snapshot of the database. Thereby, we exploit operating systems functionality to create virtual memory snapshots for new, cloned processes. In Unix, for example, this is done by creating a child process of the OLTP process via the fork system call.

The forked child process obtains an exact copy of the parent processes address space. This virtual memory snapshot that is created by the fork-operation will be used for executing a session of OLAP queries. These queries can be executed in parallel threads or serially, depending on the system resources or client requirements. In essence, the virtual memory snapshot mechanism constitutes a OS/hardware supported shadow paging mechanism as proposed decades ago for disk-based database systems. However, the original proposal incurred severe costs as it had to be software-controlled and it destroyed the clustering on disk. Neither of these drawbacks occurs in the virtual memory snapshotting as clustering across RAM pages is not an issue. Furthermore, the sharing of pages and the necessary copy-on-update/write is managed by the operating system with effective hardware support of the MMU (memory management unit) via the page table that translates VM addresses to physical pages and traps necessary replication (copy-on-write) actions. Therefore, the page replication is extremely efficiently done in 2μs as we measured in a micro-benchmark.

HyPer's OLTP throughput is better than VoltDB's published TPC-C performance and HyPer's OLAP query response times are superior to MonetDB's query response times. It should be emphasized that HyPer can match (or beat) these two best- of-breed transaction (VoltDB) and query (MonetDB) processing engines at the same time by performing both workloads in parallel on the same database state. HyPer's performance is due to the following design:

- HyPer relies on in-memory data management without the ballast of traditional database systems caused by DBMS-controlled page structures and buffer management. The SQL table definitions are transformed into simple vector-based virtual memory representations – which constitutes a column oriented physical storage scheme.
- The OLAP processing is separated from the mission-critical OLTP transaction processing by fork-ing virtual memory snapshots. Thus, no concurrency control mechanisms are needed – other than the hardware-assisted VM management – to separate the two workload classes.
- Transactions and queries are specified in SQL and are efficiently compiled into efficient LLVM assembly code.
- As in VoltDB, the parallel transactions are separated via lock-free admission control that allows only non-conflicting transactions at the same time.
- HyPer relies on logical logging where, in essence, the invocation parameters of the stored (transaction) procedures are logged via a high-speed network.

3.6 Making Sense at Scale with Algorithms, Machines & People

Tim Kraska (University of California – Berkeley, US)

The creation, analysis, and dissemination of data have become profoundly democratized. Social networks spanning 100s of millions of users enable instantaneous discussion, debate, and information sharing. Streams of tweets, blogs, photos, and videos identify breaking events faster and in more detail than ever before. Deep, on-line datasets enable analysis of previously unreachable information. This sea change is the result of a confluence of Information Technology advances such as: intensively networked systems, cloud computing, social computing, and pervasive devices and communication.

The key challenge is that the massive scale and diversity of this continuous flood of information breaks our existing technologies. State-of-the-art Machine Learning algorithms do not scale to massive data sets. Existing data analytics frameworks cope poorly with incomplete and dirty data and cannot process heterogeneous multi-format information. Current large-scale processing architectures struggle with diversity of programming models and job types and do not support the rapid marshalling and unmarshalling of resources to solve specific problems. All of these limitations lead to a Scalability Dilemma: beyond a point, our current systems tend to perform worse as they are given more data, more processing resources, and involve more people, exactly the opposite of what should happen.

To address these issues, we are starting a new five-year, multi-faculty research effort called the AMPLab, where AMP stands for "Algorithms, Machines, and People". AMPLab envisions a world where massive data, computing, communication and people resources can be continually, flexibly and dynamically be brought to bear on a range of hard problems by huge numbers of people connected to the cloud via mobile and other client devices of increasing power and sophistication. In this talk, I will give an overview of the AMPLab motivation and research agenda and discuss several of our initial projects. One such project, PIQL, is a declarative query language that also provides scale-independence in addition to data-independence by calculating an upper bound on the number of key/value store operations that will be performed for any query. Coupled with a service level objective (SLO) compliance prediction model and PIQL's scalable database architecture, these bounds make it easy for developers to write applications that support an arbitrarily large number of users while still providing acceptable and predictable performance.

3.7 Optimization of PACT Programs

Fabian Hueske (TU Berlin, DE)

The PACT Programming Model is a generalization and extension of the well-known MapReduce Programming Model. Both models have a common ground: they use a key-value pair data model and are based on parallelizable second-order functions which call first-order user functions. While MapReduce offers only two of such second-order functions Map and Reduce, PACT has an extended set of parallelization primitives that also handle multiple inputs. Furthermore, PACT supports so-called Output Contracts which are annotations that reveal

certain characteristics of the black-box user code. Finally, PACT programs are composed as arbitrary acyclic graphs. In contrast, MapReduce jobs have a static structure.

Data processing tasks implemented in PACT are compiled into parallel data flows. During this step, some degrees of freedom enable the compiler to perform physical optimization. These opportunities come from the declarative character of the parallelization primitives and knowledge that is derived from user code annotations. The compiler performs cost-based optimization and aims to reduce network and disk I/O. It chooses shipping (broadcast vs. repartition) and local strategies (sort-merge join vs. hash join) and reuses of existing physical data properties. In this regard the compiler is very similar to the physical optimizer of a traditional PDBMS. However, in contrast to well-defined SQL queries, PACT programs are arbitrary data flows sorely consisting of UDFs. The talk concludes by giving a short overview of upcoming features of the PACT programming model and motivates the need for robust optimization in the context of massively parallel analytics in cloud environments.

3.8 The ASTERIX Project: Cloudy DB Research at UC Irvine

Mike Carey (University of California – Irvine, US)

The ASTERIX project is developing new technologies for ingesting, storing, managing, indexing, querying, analyzing, and subscribing to vast quantities of semi-structured information. The project is combining ideas from three distinct areas - semi-structured data, parallel databases, and data-intensive computing - to create a next-generation, open source software platform that scales by running on large, shared-nothing commodity computing clusters. ASTERIX targets a wide range of semi-structured information, ranging from "data" use cases, where information is well-tagged and highly regular, to "content" use cases, where data is irregular and much of each datum is textual. ASTERIX is taking an open stance on data formats and addressing research issues including highly scalable data storage and indexing, semi-structured query processing on very large clusters, and merging parallel database techniques with today's data-intensive computing techniques to support performant yet declarative solutions to the problem of analyzing semi-structured information. This presentation will provide a whirlwind overview of the project, including its three-layer architecture - the ASTERIX parallel information system (with its ADM data model and AQL query language), the Algebricks query processing layer (which aims to support other implementors of data-intensive computing languages as well), and the Hyracks data-intensive computing platform (an alternative to such platforms as Hadoop and Dryad).

3.9 Algebricks + Hyracks: An efficient Data-Centric Virtual Machine for the Cloud

Vinayak Borkar (University of California – Irvine, US)

In order to harness the power of the cloud for data-intensive tasks, we need a higher level of abstraction that eases the specification of jobs. In this talk we present two systems: Hyracks,

a low-level runtime infrastructure that provides APIs to implement data-parallel operators. In addition, the Hyracks platform includes some commonly useful operators along with a Hadoop compatibility layer to transparently run Hadoop jobs on Hyracks. The second layer, Algebricks, is a higher level of abstraction that provides logical operators which get optimized and compiled down into Hyracks jobs. Algebricks provides a rewriting framework that allows users to implement new rewrite rules.

3.10 Extending Map-Reduce for Efficient Predicate-Based Sampling

Raman Grover (University of California – Irvine, US)

Data analysts today want to grab every bit of data and extract useful information from it. The collected data may scale tera or even petabytes. Sampling has been established as an effective tool in avoiding the subsequent processing cost. A fixed size random sample may not suffice as the sampled data is often required to satisfy additional predicates in order for the collected sample to be useful. We refer this kind of sampling as "Predicate-Based" sampling and is a widely occurring pattern at Facebook. We desire to be able to produce such samples from large scale data in a manner such that the response time is independent of the size of the input dataset. This allows to produce desired samples from increasingly large sizes of input data. Predicate-based sampling can be expressed as a Map-Reduce task. Hadoop as a Map-Reduce implementation provides inefficient execution as it assumes that all input must be processed for a job to produce the required result. Predicate-Based sampling belongs to a class of jobs that can potentially produce the required result by processing partial input. We present an extension of Map-Reduce execution model (as implemented in Hadoop) that allows incremental processing wherein input is added dynamical to a running job in accordance with the need and the load on the cluster. The extended model allows us to produce predicate-based samples from increasingly large quantities of data with response time being independent of the size of the input.

3.11 Challenges for Cloud Benchmarking

Enno Folkerts (SAP AG – Walldorf, DE)

We develop guidelines for designing and running cloud benchmarks. A cloud benchmark is a benchmark, which makes it possible to compare cloud services of a certain domain. We will not define a benchmark. We will state, what cloud benchmarks may have in common and what differentiates cloud benchmarks from traditional benchmarks. We will also check which traditional benchmarking principles are still valid for the cloud and which principles may have to be altered. We will argue, that it is not sufficient to run well established benchmarks in the cloud, but that the cloud calls for a new generation of benchmarks. We will also see, that there may be different challenges for consumers and providers in the domain of cloud benchmarks.

3.12 Trade-Offs in Cloud Application Architecture

Stephan Tai (KIT – Karlsruhe Institute of Technology, DE)

There are diverse objectives in cloud computing – however, not always can all of these objectives be met at the same time. This includes, for example, the traditional question of data consistency versus high availability in distributed data storage. Other potentially conflicting (classes of) objectives include cost efficiency, dependability, performance, or security. We study cloud application architectures from a service-oriented computing perspective and discuss the problem of trade-offs between conflicting objectives. We argue for a novel service engineering model that incorporates trade-offs as first-class abstractions in application architecture design, and call for additional runtime features ("tuning knobs") to flexibly manage trade-offs at runtime.

3.13 Benchmarking Large-Scale Parallel Processing Systems

Alexander Alexandrov (TU Berlin, DE)

This presentation captures an overview of recent work done in the areas of cloud benchmarking and parallel data generation. We first present Myriad – a toolkit for massively parallel generation of synthetic datasets. We show how the parallelization approach implemented by the toolkit relies on horizontal partitioning of the generated data sequences and is alleviated by the use of efficient SeedSkip operations on the underlying PRNG streams. In addition, we also explain how the toolkit fits into our general-purpose benchmark for high-level analytics languages running on top of Hadoop or similar parallelization frameworks.

3.14 To Cloud or Not To. Musings on Cloud Deployment Viability and Cost Models

Radu Sion (Stony Brook University, US)

In this talk we explore the economics of technology outsourcing in general and cloud computing in particular. We identify cost trade-offs and postulate the key principles of outsourcing that define when cloud deployment is appropriate and why.

3.15 Building Large XML Stores in the Amazon Cloud

Jesus Camacho-Rodriguez (INRIA Saclay – Orsay, FR)

It has been by now widely accepted that an increasing part of the world's interesting data is either shared through the Web or directly produced through and for Web platforms using formats like XML (structured documents). At the same time, cloud storage and computing platforms such as Amazon Web Services (AWS) have gained traction and attracted interest for their elastic scalability. In particular, AWS provides a set of basic sub-systems (such as storage for bulk, respectively, small-grained data, queue systems etc.) on top of which one can build more complex applications.

We present our ongoing work on designing an architecture and associated algorithms for efficiently managing large corpora of XML documents based on the AWS components. We consider different indexing strategies to use in order to facilitate the access to a collection of XML documents stored within AWS and efficiently support query processing on these documents. Work is ongoing, in particular on enabling our indexing algorithms to scale through the boundaries of AWS structures, and to experimentally evaluate the trade-offs brought by each strategy.

3.16 Storage for End-user Programming

Dirk Riehle (Universität Erlangen-Nürnberg, DE)

This talk illustrates our vision for end-user programming taking a wiki-style approach. We show some of the challenges that arise for a backing database.

3.17 Adaptive Query Processing in Stratosphere

Johann-Christoph Freytag (HU Berlin, DE)

The talks presents the first results on adaptive query processing in Stratosphere. Our approach is based on the SCORE operator, and extension of Hellerstein's Eddy operator, and on a competition model which is motivated by the work of G. Antoshekov (1992). We show that the two approaches together improve the overall response time when considering join processing.

3.18 Nephele: (Cost) Efficient Parallel Data Flows in the Cloud

Daniel Warneke (TU Berlin, DE)

The world of parallel computing faces data sets which are increasing rapidly in complexity and size. While many different higher-level programming abstractions have recently been introduced to facilitate domain-specific application development on these data sets, the underlying execution engines are still heavily tailored towards cluster-centric long-running batch jobs. This talk highlights the different directions for future research in the field of data-intensive execution engines and sketches our ongoing efforts in the scope of the Stratosphere project.

3.19 Information Extraction in Stratosphere

Astrid Rheinlaender (HU Berlin, DE)

Large scale analytical text processing is important for many real-world scenarios. In drug development, for instance, it is extremely helpful to gather as much information as possible on the drug itself and on other, structurally similar drugs. Such information is contained in various large text collections like patent or scientific publication databases. As a part of the StratoSphere project, we therefore investigate query-based analysis of large quantities of unstructured text. Such a query is parsed, optimized, parallelized, and executed on a cloud infrastructure. Our extraction operators are configurable to embrace different IE strategies, either geared towards high throughput, high precision, or high recall. On the other hand, we also develop optimization strategies such as rewrite rules or cost estimates that allow an efficient execution of IE queries.

3.20 Cloud Computing and Next Generation Sequencing

Ulf Leser (HU Berlin, DE)

The Life Sciences, and in particular the recent advances in DNA sequencing (Next Generation Sequencing, NGS) create an ever growing amount of data. Interestingly, the rate at which data production is increasing is much higher than Moore's law predicts for the increase in computational power - while sequencing throughput doubles roughly every 6-9 months, CPU power is doubling only every 18 months. This poses considerable challenges to the analysis of sequencing data sets. The talk explains the problem, presents the state-of-the-art, and discusses the opportunities Cloud Computing might offer for sequence analysis and well as the problems that have to be tackled.

3.21 Cost-aware data management in the cloud

Verena Kantere (Cyprus University of Technology – Lemesos, CY)

The success of offering data services in the cloud is achieving to perform both cost-efficient and traditionally time-efficient data management. We have proposed a novel economy model for a cloud provider, where users pay on-the-go for the data services they receive and user payments can be used for service provision, infrastructure operation and profit. The economy employs a cost model that takes into account all the available resources in a cloud, such as disk space and I/O operations, CPU time and network bandwidth. In order to ensure the economic viability of the cloud, the cost of offering new services has to be amortized to prospective users that will use them. We have proposed a novel cost amortization model that predicts the extent of amortization in time and number of users. The economy is completed with a dynamic pricing scheme that achieves optimal cloud profit while ensuring user satisfaction with service prices.Â We envision a cloud data service provider with three conceptual layers that should interact closely; namely, the cloud DBMS, the service and the economy layer. There are many open research issues on all the layers. Coarsely, it is necessary to provide techniques for offering and pricing groups and workflows of services that are customizable for various user needs and cloud environments taking into account risk factors.

3.22 Building high performance indexes for key value storage

Russell Sears (Yahoo! Research – Santa Clara, US)

This talk provides an overview of Yahoo!'s distributed key-value store, PNUTS. We are in the process of implementing a new log structured index for PNUTS, and discuss its implementation, and a number of issues that arise when benchmarking of log structured storage systems.

3.23 MuTeDB - A dbms that shows quiet on multi-tenancy

Bernhard Mitschang (Universität Stuttgart, DE)

Software as a Service (SaaS) facilitates acquiring a huge number of small tenants by providing low service fees. To achieve low service fees, it is essential to reduce costs per tenant. For this, consolidating multiple tenants onto a single relational schema instance turned out beneficial because of low overheads per tenant and scalable manageability. We contribute first features of an extended RDBMS to support tenant-aware data management natively. We introduce tenants as first-class database objects and propose the concept of a tenant co text to isolate

a tenant from other tenants. We present a schema inheritance concept that allows sharing a core application schema among tenants while enabling schema extensions per tenant.

3.24 Yes, but does it work?

Steve Loughran (HP Lab – Bristol, GB)

Coverage of the testing issues related to Cloud infrastructures and how applications deployed in such a world don't work the way they should, because they contain assumptions about their environment that are no longer valid.

3.25 ScalOps: Cloud Computing in a High-Level Programming Language

Tyson Condie (Yahoo! Inc., US)

Machine learning systems take part in billions of page views every day at Yahoo!. Examples include recommended reading on Yahoo! News, personalized advertisements and, possibly most well known, the Yahoo! Mail spam filter and the personalized assembly of Yahoo! Frontpage. Building the underlying models includes several distinct phases, currently accomplished using different tools: 1. Feature Extraction / Data preparation: A ETL-style feature extraction and data joining phase that is typically accomplished by Apache Pig. 2. Modeling: Yahoo! uses any number of different machine learning techniques and algorithms to model the data. They all share one key characteristic that makes them unsuitable for DAG-based systems such as Hadoop or Dryad: They perform multiple passes over the data, changing state along the way. It has been demonstrated numerous times in the machine learning community that speedups of at least 10x can be achieved by custom MPI-style implementations when compared to Hadoop MapReduce. 3. Evaluation: This, again, typically performs ETL-style computations and can be accomplished using the large scale data processing tools widely available today.

Scalops is a new machine learning toolkit currently under development. It provides an API and a runtime that can natively express and execute iterations and recursion over Big Data. This in turn allows us to unify all three steps outlined above in a single, concise and easily approachable programming interface in the form of a internal domain specific language hosted by the Scala programming language. We expect the latter to be of great benefit to machine learning practitioners at Yahoo! and beyond. We also envision unifying several now disparate computational paradigms under a single runtime.

3.26 Cloud-based Web data management (it's all about how you view it)

Ioana Manolescu (Université Paris Sud – Orsay, FR)

Joint work of Dario Colazzo, Francois Goasdoue, Jesus Camacho-Rodriguez, Andres Aranda Andujar, and Zoi Kaoudi

The development of the Web led to a strong increase in the volumes of Web-style data being produced, exchanged, analyzed and consumed daily; thus, it is estimated that we now produce every two days the same amount of data that was produced from the beginning of humanity until 2003 . Moreover, most of this continuously produced data does not reside in databases but in Web content such as Web pages, social networking sites, blogs, user videos etc. This wealth of data leads to great interest in efficiently and reliably storing, querying, analyzing and transforming such data. By "Web data", we designate document data, in the style of Web pages, and which we views as XML documents, as well as Semantic Web style data, represented by RDF triples, possibly endowed with RDF Schemas.

In this context, cloud platforms provide a distributed framework, providing at least some lower (file-) level storage of the data. Typical cloud infrastructure provide several levels of storage, one dedicated to very large (unstructured) data objects, and another one built for storing numerous small items, typically structured as sets of attribute-value pairs. Also within the context of cloud computing, frameworks and programming languages have emerged, typically with an emphasis on parallel processing, distributing parallel computations and gathering back results, along the lines of "and typically generalizing or extending" the Map/Reduce paradigm. The starting point of our research agenda in this context is the observation that the complex, heterogeneous or missing structure of Web data raises many challenges for large-scale efficient data management platforms. Indeed, in a large distributed setting, it is not clear how one user' or application's data should be organized on the available storage layers, in order to support efficiently the required query/transactions mix. The complex shape of Web data formats is typically not a good format for storing the data, thus various segmentations or fragmentation strategies are often applied to re-organize the content in smaller, more manageable fragments. In turn, these fragments may be replicated on several sites and adaptively placed across the distributed storage units. When available, schema information as well as information about the workload of each user can also be used to this purpose. The design of such indexes and views should be made with parallelism in mind, so that no single point of contention is introduced when searching for the data structures suited for a given query.

We plan to investigate the design, algorithmic and performance properties of efficient storage structures in the context of cloud-based data management, in particular distributed indexes and distributed materialized views. This work is to be performed in particular within the ICT Labs Europa activity.

4 Overview of Demos

During the demo session three system demos were given. Each demo is shortly described in the remainder of this section.

4.1 Asterix

Vinayak Borkar, Raman Grover

URL http://asterix.ics.uci.edu

The Asterix system is developed at UC Irvine, UC Riverside, and UC San Diego. It is a platform to execute queries on semi-structured data in a massively parallel fashion. The basic system consists of three components, an execution engine called Hyracks, an algebraic optimization layer named Algebrix, and the Asterix query language (AQL). The Hyracks engine is published as open source. The demo showed how data schemas and analytical (OLAP-style) queries are specified by AQL, how they are optimized and executed on Hyracks.

A second demo showed a use case that computes a geo-spatial frequency aggregation of twitter feeds which contain a certain keyword. The result was visualized as heat-map using a web-based map service.

4.2 Stratosphere

Daniel Warneke, Fabian Hueske

URL http://www.stratosphere.eu

Stratosphere is a joint research project by TU Berlin, HU Berlin, and HPI Potsdam. The project researches data management in the cloud and builds a prototype that is publicly available as open source. The Stratosphere system consists of the parallel PACT programming model, an database-inspired optimizer, and a flexible execution engine called Nephele. PACT is a generalization of the MapReduce programming model. Nephele can request computing nodes on demand from Infrastructure-as-a-service (Iaas) providers. The demo showed how an analytical query is defined a PACT program, how it is optimized, and how it is executed on the Amazon EC2 Iaas environment.

A second use case demonstrated a biomedical information extraction pipeline that was defined as a PACT program.

4.3 Parallel Data Generation with Myriad

Alexander Alexandrov

URL http://www.myriad-toolkit.com

Myriad is a development framework for parallel data generators. Myriad relies on an efficient skip-ahead pseudo-random number generator (PRNG) sequence to create virtual pseudo-random sequences for user-defined data types that can be partitioned and randomly accessed at constant computational cost. Myriad-based generators can therefore generate skewed, correlated, and referencing data without any communication between generator instances

running in parallel. This feature makes Myriad a viable framework to define workloads and benchmarks for massively parallel systems such as Hadoop, Asterix, or Stratosphere.

The demo session showed how Myriad can be extended to generate graphically structured data in parallel. The statistical constraints implemented by the data generator make the produced datasets a good fit for testing certain types of analytical queries in a large-scale environment.

5 Break-Out Group Reports

This section lists the abstracts of the break-out sessions. The abstracts and figures were taken from the seminar's material web site.

5.1 Cloud Benchmarking

Alexandru Iosup, Alexander Alexandrov, Enno Folkerts, Donald Kossmann, Seif Haridi, Volker Markl, Tim Kraska, Radu Sion, Anastasia Ailamaki, Dean Jacobs

The goal of this breakout session was to begin work on providing a procedure for rating cloud Infrastructures and Platforms. An important target was to consider ways through which clouds could receive ratings that IT consumers, especially small companies, can use to guide their IT provisioning processes. The need for new benchmarks and benchmarking practices derives from the need of these IT consumers to understand the performance-, the availability-, the reliability-, the scalability-, the elasticity-related, etc. characteristics of clouds. Without a standard benchmarking suite, cloud operators are unable to demonstrate their claims; conversely, potential buyers are not persuaded to buy.

Our group has focused on three main tasks:

1. Defining a framework for the process of benchmarking, which can guide the creation, use, and reporting based on a suite (family) of benchmarks.
2. Understanding the main cloud characteristics that may require new approaches to benchmarking. For example, due to the performance variability exhibited by many clouds, benchmarking metrics have to focus on both expectation and variability. Similarly, elasticity, which encompasses the behavior of the system under varying load, needs possibly new metrics. Other notions discussed were: scalability (including the time needed to reach the desired scale), reliability, availability, robustness (against a "TNT" test), information availability (knowing partially the status of the system), the data management lifecycle (including backups under load and archival), the ability to benchmark data consistency, etc.
3. Asking the questions that can guide the creation of new cloud-related benchmarks. What are good Key Performance Indicators and how to build a Single-Value Rating? Should we test under (external) load? Should we use test drivers located in the cloud? How to benchmark for different provisioning and allocation models/policies? How to benchmark for interactive workloads? (the distance to customer is now part of cloud location) How to build scalable benchmarks and tools for them? etc.
4. Creating a plan for continuation. We have agreed on a plan for continuation.

5.2 Biomedical Analytics in the Cloud

Jim Dowling, Johann-Christoph Freytag, Keijo Heljanko, Ulf Leser, Felix Naumann, Astrid Rheinländer

Recent improvements in both the cost and throughput of sequencing machines has caused a mismatch between the increasing rate at which they can generate data and the ability of our existing tools and computational infrastructure to both store and analyse this data. Currently, organizations are investing significant amounts of resources in sequencing machines before they either have the necessary storage infrastructure or analysis tools that can archive and process the resultant data.

The goal of this breakout-group is to propose both a cloud-computing infrastructure and parallel-programming support that will enable the secure storage and parallel analysis of the coming flood of sequence data. We anticipate that an infrastructure of only 100 machines should cost at most the same as an existing sequencing machine and, with the help of recent cloud computing technologies, it should have minimal administration costs. Such an infrastructure will enable organizations to support the long-term archival of sequence data and reduce the time required to process sequence data by a factor of around one hundred. A sample worflow on top of our parallel infrastructure is depicted on Figure 1.

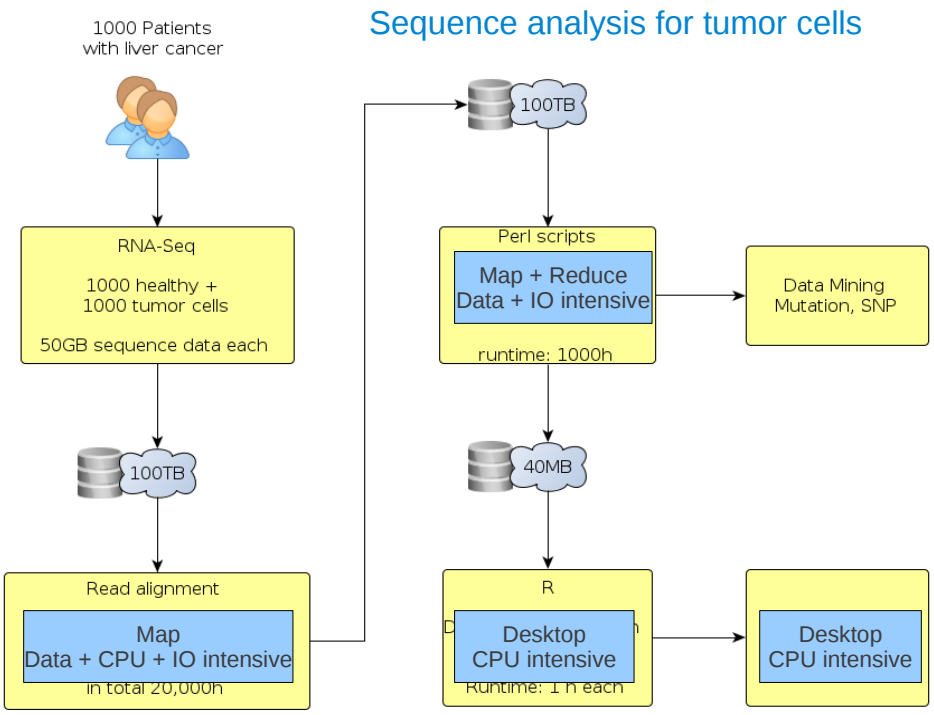


Figure 1 Sequence analysis worflow

5.3 Data and Programming Models

Vinayak Borkar, Jesus Camacho-Rodriguez, Mike Carey, Tyson Condie, Raman Grover, Arvid Heise, Fabian Hueske, Dean Jacobs, Steve Loughran, Ioana Manolescu-Goujot, Sergey Melnik, Bernhard Mitschang, Daniel Warneke

The Data and Programming Model breakout session tried to come up with characterizations of large-scale data applications and platforms. These characterizations should be used to describe and specify the requirements of applications and features of platforms in order to find matches between both. Figure 2 shows how to derive application platform matches based on their characteristics. After a set of characteristics had been derived, they were applied to a couple of example applications and platforms. In addition a 'map' of software stacks of selected parallel data processing platforms was created.

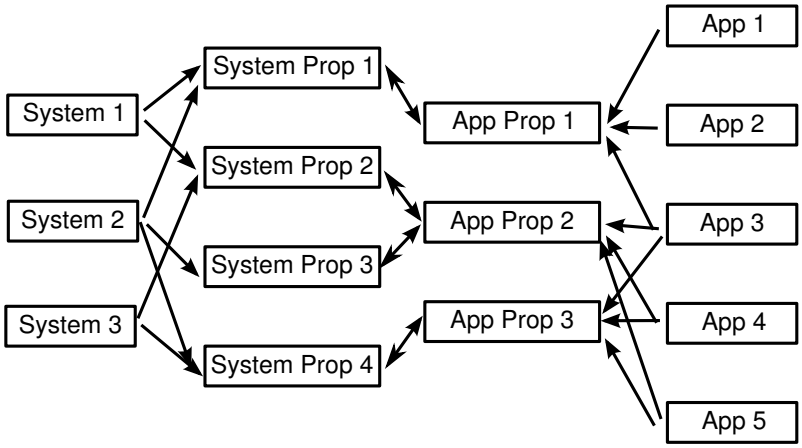


Figure 2 Matching of Application and Platform Characteristics

Abstract

The era of the mainframe and the cluster may seem over, but their concepts are being applied to large-scale datacentres, offering massively-parallel, data-intensive computing and storage services. The challenge in this world is what algorithms can scale up to this environment, tolerate the frequent failures, and support the complex analysis and computational needs of the latest generation of applications.

The goal of this breakout session is to characterise the algorithms and the programming models that have been built to work in this environment. For some popular problems, we show their characteristics, and therefore how their needs match the feature set of these programming models. The characteristics show gaps in the feature set of today's technologies; features that future systems could address.

Application Characteristics

1. Application types: **A**nalyze vs. **T**ransform vs. **E**xtract

Table 1 Application Characteristics

System	1	2	3	4	5	6
WordCount	A	B	M	D	S	
PageRank	A	B	M	D	I	
TPC-H	A	O	S+M	D	S	
K-Means	A	B	M	D	I	
Tile Rendering	T	B	M	C	S	
ETL	T	B	M	D	S	
Recommendation (SGD/LDA)	A	B	M	C	I	
TrendAnalysis (Twitter)	A	B	S	D+C	S	

2. Operation response mode: **O**nline (sync) vs. **B**atch (async)
3. Request types: **S**elective vs. **M**assive
4. Operation step types: **D**ata Intensive vs. **C**ompute Intensive
5. Operation processing mode: **S**ingle flow vs. **I**terative / recursive
6. Data access modes: **G**et vs. **F**ilter vs. **Q**uery

Platform Characteristics

1. Application types: **A**nalyze vs. **T**ransform vs. **E**xtract
2. Data types: **S**tatic **T**yped vs. **D**ynamic Typed vs. **U**ntyped (to be updated in Table 2)
3. Operation response mode: **O**nline (sync) vs. **B**atch (async)
4. Operation semantics: **T**ransparent vs. **O**paque
5. Request types: **S**elective vs. **M**assive
6. Operation step types: **D**ata Intensive vs. **C**ompute Intensive
7. Operation processing mode: **S**ingle flow vs. **I**terative / recursive
8. Data access modes: **G**et vs. **F**ilter vs. **Q**uery (to be updated in table 2)

Characterization of Selected Applications

List of potential example applications:

- Genome Alignment
- Data Cleansing
- Enterprise OLAP
- E-Health Record Management
- Twitter Analysis
- (Ad) Recommendation Systems
- Indexing
- Auditing / Sensor Networks
- (Realtime) Log & Click Analysis
- Tile Rendering
- PageRank
- Social Network Analysis
- Shortest Path

Table 1 shows the characterization of selected applications.

Table 2 System Characteristics

System	1	2	3	4	5	6	7	8
Pig	T	U	B	O$^-$	S+M	D+C	S	
Hive	A	T	B	T	S+M	D	S	
AQL	A+E	T	O+B	T	S+M	D	I	
PACT	A+T	U	B	O$^-$	M	D+C	S	
MR PM	A+T	U	B	O	M	D+C	S	
SQL	A+E	T	O	T	S	D	I$^-$	
Pregel	A	T$^-$	B	O	M	D+C	I	
Nephele	A+T	U	B	O	M	D+C	S	
MPI	–	U	B	O	M	C	I	
Dremel	A	T	O	T	S	D	S	
SimpleDB	E	T	O	T	S+M	D	S	
HBase	E	U	O	T$^-$	S	D	S	

Characterization of Existing Platforms

Table 2 shows the characterization of selected platforms.

Selected Parallel Data Processing Stacks

Figure 3 shows the processing stacks of selected parallel data processing platforms.

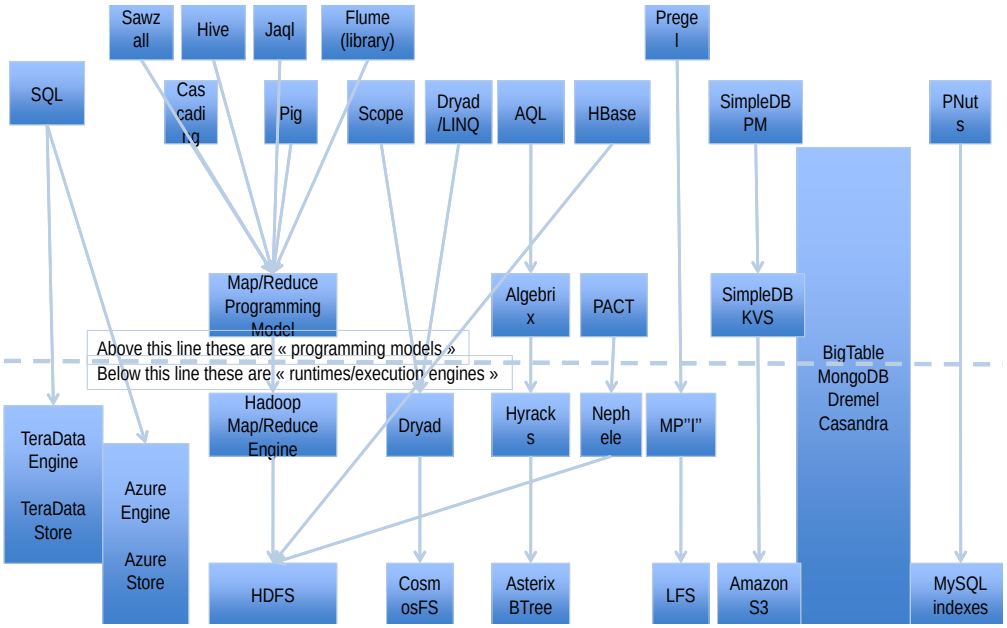


Figure 3 Parallel Data Processing Stacks

5.4 Transactions in the Cloud

Anastisia Ailamaki, Seif Haridi, Alfons Kemper, Tim Kraska, Simon Loesing, Sergey Melnik, Russell Sears

The increasing scale of data management and high-availability requirements have led large Internet services to deploy scalable storage systems that span many data-centers. Such systems relax transactional semantics, such as atomicity and consistency, for scalability. Based on experiences with these systems, we want to explore the fundamental trade-offs these systems face. This includes defining the design requirements that systems of this scale have to fulfill. Some of these requirements are universal, such as manageability and fault-tolerance, while others vary with the application. In particular, these application-dependent differences lead to different data models, consistency properties, data placement and programming models which directly impact the approaches applications can use to transactionally modify the data.

5.5 Cloud Economics

Verena Kantere, Magdalena Balazinska, Athanasios Papaioannou, Helmut Krcmar, Miron Livny

Cloud-computing has recently emerged as a new paradigm for delivering compute infrastructures and software in an "elastic" (i.e. flexible, scalable, and pay-as-you-go) manner. While the service elasticity offers significant advantages, such as reducing the time-to-market and allowing adaptive capacity planning, it also creates important challenges for the platform and software design. For instance, software must effectively take advantage of the possibility to grow and shrink resources as needed. In this work, we study the case of data-management-as-a-service. Today, cloud providers offer data management solutions but they are either feature limited (e.g., Amazon SimpleDB) or lack scalability (e.g., SQL Azure). We identify key design challenges (e.g. system adaptivity to agile resource planning, setup and data migration costs, etc.) and sketch possible architectures.

5.6 Cloud Storage

Anastisia Ailamaki, Russell Sears

Gaps in solid state disk, network and magnetic disk performance are growing exponentially. In the long term, solid state storage performance will outpace networking, while networking will outpace magnetic media. Given these trends, and the need for both magnetic and solid state media, it is unclear whether it will continue to be possible to build general purpose clouds in the future, or how many types of specialized clouds will make sense.

It is also unclear whether the underlying hardware architecture should be homogeneous, so that each machine contains multiple types of storage devices, or heterogeneous, with many classes of machines provisioned for distinct workloads. In a homogeneous system, interference from different types of applications may severely impact long-tail latencies and overall throughput. However, heterogeneous designs statically partition applications into silos, preventing capacity sharing. Also, different classes of applications lead to different bottlenecks; the heterogeneous approach amplifies these bottlenecks.

We intend to benchmark a number of configurations and systems based on both approaches, and to see which of the above problems are most serious. We will use these results to inform the design of new cloud-based storage hardware and software stacks.

Participants

- Alexander Alexandrov
TU Berlin, DE
- Alexandru Iosup
TU Delft, DE
- Alfons Kemper
TU Munich, DE
- Anastassia Ailamaki, EPFL
Lausanne, CH
- Arvid Heise
Hasso Plattner Institute
Potsdam, DE
- Astrid Rheinlaender
HU Berlin, DE
- Athanasios Papaioannou
EPFL Lausanne, CH
- Bernhard Mitschang
Univerity Stuttgart, DE
- Daniel Warneke
TU Berlin, DE
- Dean Jacobs
SAP AG, Walldorf, DE
- Dirk Riehle
Univ. Erlangen-Nuernberg, DE
- Donald Kossmann, ETH
Zurich, CH
- Enno Folkerts
SAP AG, Walldorf, DE
- Fabian Hueske
TU Berlin, DE
- Felix Naumann
Hasso Plattner Institute
Potsdam, DE
- Helmut Krcmar
TU Munich, DE
- Ioana Manolescu-Goujot
Universite Paris Sud-Orsay, FR
- Jesus Camacho-Rodriguez
INRIA Saclay-Orsay, FR
- Jim Dowling
Swedish Institute of Computer
Science, Kista, SE
- Johann-Christoph Freytag
HU Berlin, DE
- Keijo Heljanko
Helsinki Univ. of Technology, FI
- Magdalena Balazinska
University of Washington,
Seattle, US
- Mike Carey,
UC Irvine, US
- Miron Livny
Univ. of Wisconsin-Madison, US
- Radu Sion
Stony Brook University, US
- Raman Grover
UC Irvine, US
- Russell Sears
Yahoo! Research, US
- Seif Haridi
Swedish Institute of Computer
Science, Kista, SE
- Sergey Melnik
Google, US
- Simon Loesing
ETH Zürich, CH
- Stefan Tai
Karlsruhe Institute of
Technology, DE
- Steve Loughran, HP Labs
Bristol, UK
- Tim Kraska
UC Berkeley, US
- Tyson Condie
Yahoo! Inc., US
- Ulf Leser
HU Berlin, DE
- Verena Kantere
Cyprus University of Technology,
Lemesos, CY
- Vinayak Borkar
UC Irvine, US
- Volker Markl, TU Berlin, DE

Report from Dagstuhl Perspectives Workshop 11331

The Future of Research Communication

Edited by

Tim Clark[1], Anita De Waard[2], Ivan Herman[3], and Eduard Hovy[4]

1 **Harvard Medical School & Massachusetts General Hospital, US,** twclark@nmr.mgh.harvard.edu
2 **Elsevier Labs – Jericho, US,** a.dewaard@elsevier.com
3 **W3C/CWI – Amsterdam, NL,** ivan@w3.org
4 **University of Southern California – Marina del Rey, US,** hovy@isi.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Perspectives Workshop 11331 "The Future of Research Communication". The purpose of the workshop was to bring together researchers from these different disciplines, whose core research goal is changing the formats, standards, and means by which we communicate science.

Seminar 15.–18. August, 2011 – www.dagstuhl.de/11331
1998 ACM Subject Classification H.2.8 Database Applications, H.3 Information Storage and Retrieval, H.5 Information and Interfaces and Presentation, I.7.4 Electronic Publishing, J.7 Computers in other Systems, K.4 Computers and Society
Keywords and phrases science publishing, online communities, science policy, new forms of publishing, bioinformatics, digital repositories, semantic publishing, citation analysis, data publication, information access and integration, reporting standards
Digital Object Identifier 10.4230/DagRep.1.8.29
Edited in cooperation with Aliaksandr Birukou

1 Executive Summary

Philip E. Bourne
Tim Clark
Robert Dale
Anita de Waard
Ivan Herman
Eduard Hovy
David Shotton

Research and scholarship lead to the generation of new knowledge. The dissemination of this knowledge has a fundamental impact on the ways in which society develops and progresses, and at the same time it feeds back to improve subsequent research and scholarship. Here, as in so many other areas of human activity, the internet is changing the way things work: it opens up opportunities for new processes that can accelerate the growth of knowledge, including the creation of new means of communicating that knowledge among researchers and within the wider community. Two decades of emergent and increasingly pervasive information technology have demonstrated the potential for far more effective scholarly communication.

The Future of Research Communication, *Dagstuhl Reports*, Vol. 1, Issue 8, pp. 29–52
Editors: Tim Clark, Anita De Waard, Ivan Herman, and Eduard Hovy
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

However, the use of this technology remains limited; research processes and the dissemination of research results have yet to fully assimilate the capabilities of the web and other digital media. Producers and consumers remain wedded to formats developed in the era of print publication, and the reward systems for researchers remain tied to those delivery mechanisms.

Force11 (the Future of Research Communication and e-Scholarship) is a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. Individually and collectively, we aim to bring about a change in scholarly communication through the effective use of information technology. Force11 has grown from a small group of like-minded individuals into an open movement with clearly identified stakeholders associated with emerging technologies, policies, funding mechanisms and business models. While not disputing the expressive power of the written word to communicate complex ideas, our foundational assumption is that scholarly communication by means of semantically-enhanced media-rich digital publishing is likely to have a greater impact than communication in traditional print media or electronic facsimiles of printed works. However, to date, online versions of 'scholarly outputs' have tended to replicate print forms, rather than exploit the additional functionalities afforded by the digital terrain. We believe that digital publishing of enhanced papers will enable more effective scholarly communication, which will also broaden to include, for example, better links to data, the publication of software tools, mathematical models, protocols and workflows, and research communication by means of social media channels.

This document reports on the presentations and working groups that took place during the Force11 workshop on the Future of Research Communication and e-Scholarship held at Schloss Dagstuhl, Germany, in August 2011. More about Force11 can be found at http://www.force11.org. This document is structured as follows. Sections 3-5 report on the presentations of the participants. The presentations discuss, respectively, the past (Section 3), present (Section 4) and future (Section 5) of research communication. Section 6 presents the notes from the working groups. The notes are presented with only minor modifications, to capture the spirit of what was happening "in situ". Section 7 lists the websites and other documents related to the workshop. Section 8 contains the timetable of the workshop. Finally, we list the participants of the workshop.

2 Table of Contents

3 Overview of Talks. The Past

3.1 The Future of Research Communications: The Past

Anita de Waard (Elsevier Labs – Jericho, US)

URL http://slidesha.re/pkspZZ

To see where we need to go in the future, it can be useful to look at the past with a critical eye. For instance, the concept of hypertext: selecting portions of text and linking them to other portions of text, has been around, conceptually, since Vannevar Bush, and practically, since Engelbart's 1968 seminal work. Yet apart from the web, which is a low-hanging fruit realisation of this idea – with only simple links that bring you to another page; not the conceptual networks that were originally conceived – the idea has never really come about, although it is reinvented with startling regularity. Why is this the case? We define four factors that contribute to a technology being accepted:

- Commercial support (e.g. Microsoft Word)
- Community uptake (e.g. LaTeX)
- Ease of use (e.g. the web)
- Academic Credit (e.g. grant proposals)

and discuss how these played a role for the topics discussed at Force11: New Formats, Research Data, Tools and Standards, Business Models, and Attribution and Credit.

3.2 Net-Centric Scholarly Discourse?

Simon Buckingham Shum (The Open University – Milton Keynes, GB) – twitter `@sbskmi`

URL http://slidesha.re/qvoqoU

To make science and scholarship into a more agile sensemaking and problem solving system, better able to respond to the demands of a rapidly shifting environment, we need tools designed for an infrastructure unimaginable in the 17th Century when the first scholarly journals were born. However, the paradigm these established still dominates how we continue to disseminate knowledge. The founding fathers of hypertext, Vannevar Bush (1945) and Doug Engelbart (1963), clearly had the future of scholarly communication in mind when they presented use cases for their pioneering intellectual technologies. In this talk I will trace the core ideas which research since has sought to bring to reality. The essence of the idea is that scholarly communication is the crafting and contesting of networks of ideas, such as claims, concepts, evidence, arguments, and that linear prose is only one way in which to express knowledge. I will give a few examples of how new contributions to the long term reflective conversation of scholarly communication can now be made using the social, semantic web operating across many kinds of device.

3.3 A Brief History of E-Journal Preservation

David S. H. Rosenthal (Stanford University Libraries, US)

URL http://blog.dshr.org/2011/08/brief-history-of-e-journal-preservation.html

Overview of the evolution of e-journal preservation from the initial Mellon Foundation projects to the present. How well did the various business models and technologies work? Where do the costs come from? What are the implications for the future?

4 Overview of Talks. The Present

4.1 Open Citations

David Shotton (University of Oxford, GB)

URL http://bit.ly/vnRNEQ

The Open Citations Corpus (http://opencitations.net/) contains references to 3.4 million biomedical papers, representing 20% of all PubMed Central papers published between 1950 and 2010, and including all the highly cited papers in every biomedical field. The Open Citations web site provides access to the entire corpus with various search and browse options. The entire dataset is downloadable in various formats, including RDF and BibJSON, for reuse. Incoming and outgoing citation networks of selected references can be displayed in different ways and downloaded in various formats. The citation contexts of in-text citation pointers can be used to text mine the cited article and pull back sentences of relevance, to assist the reader in evaluating the quality of the citation and the cited article.

4.2 Scholarly Communication in the Present

Paul Groth (VU University Amsterdam, NL)

URL http://bit.ly/uHWmE9

Current scholarly communication practices can be broadly classified into four main categories: papers, professional meetings, databases, and informal communication. We briefly describe these categories to provide a picture of communication practice in the year 2011.

Papers are the predominate category of scholarly communication and still follow roughly the same form as for the past 200 years. Books and monographs take the role of papers in some disciplines. The Internet has changed the manner in which papers are distributed and managed. Digital libraries and search engines are the primary means to find papers in many disciplines. Social media is playing an increasing role in surfacing particular papers. Interestingly, papers are now often referred to, not by a citation, but using a URL of the paper on the Web. Papers are managed by specific reference management software. Publication of

papers is still largely journal oriented and mediated through peer review and other editorial processes. Open access journals have become more common.

Professional meetings such as conferences, symposiums, and workshops play an important role as they provide forums for scientists to meet and discuss their latests findings and approaches without the lag of publication. This is particular important as research is often international in nature and thus requires face-to-face meetings. Increasingly, conferences leave traces on the Web through the posting of slides and other material as well as live conversations in social media.

Databases have become a primary mechanism for communicating results across scientific disciplines. Many journals in the life sciences, for example, require the deposition of data within on-line databases before a paper can be published.

Informal communication is an important part of the scholarly communication life cycle. The internet and in particular social media (blogs, microblogging, email forums) have become increasingly prevalent. However, the primary means of informal communication is email. Indeed, it can be safely said that email is the main means for scholarly communication today.

Finally, it is important to note that scholarly communication acts as one of a central proxy by which scientific performance is measured. Indeed, the publication of papers in journals is the single proxy often used and is often the basis for career advancement decisions.

While the Internet has changed the way scholarly communication is done. The journal paper still dominates as the primary trackable product of this communication.

4.3 What is holding us back? A short exploration of current impediments to integrated publishing of data and primary research

Fiona Murphy (Wiley-Blackwell, UK)

Others have also highlighted these points — towards promoting discussion. The issues/stakeholders are: Technology/systems, Funding bodies/mandates, Researcher behaviour, Publishers, Other.

Tech/systems: People collect data ad hoc on laptops. Often not collected with the final deposit/site in mind so incurring expense and difficulty, Interfaces may be unhelpful (BADC), Formats issue — danger of outdated media.

Funders: Historically unhelpful. Remote, not communicating or incentivising. Demanding compliance but not following through. In the process of changing gradually.

Researchers: Suspicious of sharing IP/politics (Climategate), Anecdotally data underground/siege mentality, No time, Do not see benefits. There is a missing member of the team. Not trained.

Publishers: Not facilitating — hesitant to invest do not see the benefits either, Used to dealing with libraries rather than end users, Locked into traditional mind-sets — incunabular, Not yet built expertise to required level, Partnerships unknown.

Other: Confusion about where data should sit: who is responsible?

4.4 Making "Beyond the PDF" Current Practice

Philip E. Bourne (UC San Diego, US)

URL http://www.slideshare.net/pebourne/dagstuhl

I report on my perspective as a computational biologists on what I consider major developments in scholarly communication that have happened in the past 7 months since the beyond the PDF workshop. Notable is the announcement of SciVerse from Elsevier which in my opinion has the potential to change the model for how we interact with scholarly content. I also describe my experiences and approach to the established notion of a data journal and how I propose to contribute. Finally, I describe recent experiences with workflows and my perceived impact that they might have on the reproducibility of science.

4.5 A (very) short history of the ADS

Michael J. Kurtz (Harvard-Smithsonian Center for Astrophysics, US)

Main reference Michael J. Kurtz, "The Emerging Scholarly Brain," in Future Professional Communication in Astronomy II, Astrophysics and Space Science Proceedings, 2011, Volume 1, pp. 23–35.
URL http://dx.doi.org/10.1007/978-1-4419-8369-5_3

The Smithsonian/NASA Astrophysics Data System is a sophisticated digital library/ information system; it is used at least daily by nearly every astronomer. It was conceived in 1987, and came on-line in 1992. It is a central engine of astronomy's large and complex information environment, linking together literature and data.

The ADS is in the process of a massive re-engineering. The prototype for the new system can be found at: http://adslabs.org/ui

4.6 How to communicate the data described in publications? The Dryad model

Todd Vision (University of North Carolina – Chapel Hill, US)

Of the tens of millions of research articles that have been published, the underlying data for validation and reuse are available for only small fraction. This compromises the quality and credibility of science. To realise a world in which the publication of research data is customary, it will be necessary to adopt a multifaceted strategy. This includes technological innovations in data repositories, alterations to the landscape of researcher incentives, experimenting with new models of sustainability, and exploring new roles for publishers, learned societies, and funders. Leveraging the close relationship between research data and scholarly publication lessens these challenges, and we are experimenting with such a model in Dryad, a repository for data associated with articles in the basic and applied biosciences.

4.7 More than just data!

Cameron Neylon (Rutherford Appleton Lab. – Didcot, GB)

Much of the discussion of enhanced research communication turns on the availability of digital assets, mostly data, but with an increasing emphasis on software and workflows as well, and the exploitation of these assets to provide a rich media experience, enhanced functionality and discoverability or other benefits of online interactions. Less explored are the issues of how the data was collected, what the relevant physical artefacts are, and how best to capture the information on this in a useful way. As is also the case for effective data and digital process publication this requires systems that help the user to think about publication earlier than is traditional but there are unique challenges to capturing the record of physical processes and in particular the physical world provenance trail that leads to the first relevant digital artefact. This means that effective laboratory recording systems that enhance communication as opposed to just record keeping need to be built and configured in a way that makes those recording processes easy, automatically captures records of physical and digital artefacts via data model that can deliver immediate benefits to the user, but also renders the ultimate aggregation and collation of records into a useful form for communication easy as well. These challenges are not yet sufficiently addressed by the tooling that supports the capture and communication of digital research artefacts and processes.

5 Overview of Talks. The Future

5.1 The Future. Or: What I would Like from Publications of the Future

Eduard H. Hovy (University of Southern California – Marina del Rey/ISI, US)

URL http://bit.ly/t6N2NI

This talks presents the overall vision of the enterprise, which it defines as "To improve the communication of knowledge between scholars using new informatics technology", and lists the general kinds of communicative services that a Publication of the Future (PoF) should provide. These include:

1. Better knowledge access
   - Using terminology standards
   - Automating access
2. Better knowledge communication
   - Reflecting the foundational theory and methodology
   - Contextualising the work in relation to current world
   - Using the best media at hand
3. Better knowledge verification/extension
   - Exposing the reasoning
   - Providing non-text info and tools

The talk illustrates each point with examples, taken from both the sciences and the humanities. It ends with a draft outline of the eventual report.

5.2 Introduction to the Future of Research Communication

Tim Clark (Harvard Medical School & Massachusetts General Hospital, US)

URL http://www.slideshare.net/twclark/dagstuhl-future-sesssion-intro-slides

Research Communication exists in a complex web of technology, information, people and activities. It is currently in a transitional state between print media and Web media. A number of problems are posed for its future development. These include research reproducibility and data provenance, interoperability, dealing with masses of data on previously unknown scales, algorithmic assistance to readers, and in general dealing with the issue of volume of publications, which is intractably large for even highly specialized disciplines.

Technological solutions alone will not be sufficient. The most productive solutions to these and other problems will adopt the "ecosystemic" perspective. They will emphasize the interaction of technology, information, and social formations in mutually beneficial ecosystems, or more correctly, "activity systems", in which value chains are built and sustained for participants from multiple interacting disciplines and communities.

5.3 Networked Knowledge

Stefan Decker (National University of Ireland – Galway, IE)

A new publishing paradigm as a social-technical system. A first approach to the necessary infrastructure for Networked Knowledge – initial ranking, abstractions and access mechanisms.

5.4 The Execution of Dave 2.0

David De Roure (University of Oxford, GB)

URL http://www.myexperiment.org/packs/206.html

What happens when there are millions and millions of executable papers, sitting there and executing away...? "Executable journals" are a step towards this vision – a world of inter-related executable papers, in an altered ecosystem of scholarly publishing with new intermediaries like observatories and a new role for existing intermediaries like libraries and publishers. What will that world be like? It will help us do science-on-demand ("press this button to re-run your thesis"), and equally the papers can process new data autonomously, generating new results which in turn get processed by other papers. You'll receive an email notification when the paper you wrote five years ago is re-run with new inputs from other

people's papers, and so will the people who used yours. Automated execution assists curation and indeed validation and quality checking – and whatever replaces peer review as we know it. Is this crazy or inevitable? The co-evolutionary design of the myExperiment website (http://www.myexperiment.org) for sharing computational workflows gives us as glimpse into this world of executable "Research Objects", which is being further developed under the Wf4Ever project.

5.5 "Towards Horizons 2020" — The Framework Programme for Research and Innovation 2014 to 2020 and Role of scientific data

Mike W. Rogers (European Commission Brussels, BE)

Europe's aim to be the leading knowledge based economy will be supported by the new Framework Programme. The development of the specific programme will be an outcome of intensive public debate and stakeholder participation, based on a number of guiding principles which are rapidly emerging after the first wave of consultations:

- Strong support for bringing research and innovation together in an integrated funding programme.
- Simplification is a key priority for all stakeholders.
- All stages in the innovation chain should be supported, with more attention for close to the market activities (e.g. demonstration, piloting).
- Continuity for the successful elements of current programmes, e.g. European Research Council, Marie Curie, collaborative research.
- EU funding should be tied closely to societal challenges and EU policy objectives (climate change, ageing, energy security, ...).
- More openness and flexibility is needed, less prescriptive calls, better use of bottom-up instruments (also in programme parts guided by clear policy objectives).

The presentation developed the rationales and scope of the various consultations in order to enable participants to better understand the future roadmaps for European Research Models where the connectivity from research to Innovation will be addressed systemically. More specifically, the current consultation on the future of Scientific Data was presented and a number of themes highlighted which the Workshop could develop as a core to its response to the European Commission, both as individuals, as representatives of organisations and as a body of expertise in its own right.

6 Working Groups

The seminar participants formed several working groups that tried to focus on various issues related to the future of research communication and e-scholarship. This section presents the notes of these groups. Note that since the working groups took place in parallel, there was no single terminology: for instance, digital artefact, research object, publication of the future are likely to have the same meaning. The seminar participants agreed to publish a white paper based on these notes (see Force11 white paper in Section 7).

6.1 Data

This group aimed at brainstorming on the main issues related to the creation and publishing of data. Below we include the list of main questions raised during the discussion. More notes can be seen at http://bit.ly/usiQOE.

- How much does the research domain matter when thinking about new publication forms?
- How do we effectively collaborate?
- How do we relate the kind of tool we have and integrate them so that the scientist can play with them?
  - We need to formulate use cases, we could generate a vision on what happens if we put all this together
- What we are already doing in this community to improve scientific publishing and what could we do next?
- How do we get researchers to buy into new ways of publishing?
  - Should we aim to be contagious? People can register and share their things
- Information and data curation
  - How can we maximise the input and first pass of information curation?
  - What is the role of curators to validate NLP results?
- Issues of time
  - How rapid should be the science production loop, from data to publication to science communication? There is a pre- and post- publication aspect, the quicker a publication can be devolved the faster is the impact on people citing that data.
- What would researchers need to know from others?
- What are the bottlenecks of open science?
  - Sharing is a bottleneck: scientists are not available to share they are fine to collaborate and publish but not sharing, because there is no recognition and there are potentially negative effects. There is no policing and penalty.
  - We need funding bodies and journals to penalise those who do not share
  - We can invent mechanism to detect who does not share, i.e if a pub derives form a work it is not collaborative, probably it does not connect to the other research artefacts out there
- We do not have a good value proposition: reservation for self use
- Knowing what I have in my lab
- What do scientists need?
  - Recovery and archive of data, plus access control to data
  - Productivity, I want to be helped into publishing more
  - How do we make the literature more effectively used
  - People do not like paper summarisation, they do not trust the conceptual model presented they think it is limited
  - There is no accreditation for doing annotation, knowledge curation or any king of paper summary
  - What are the incentives we propose for doing this activities?
- Information complexity: I want to be helped to read the right papers in the right (also interdisciplinary papers)
  - Would be good knowing what the most relevant paper are

6.2 Tools and Technologies

This group aimed at organising and predicting the requirements of tools and technologies for Scholarly Communications. This group was made up of: Carole Goble (chair), David De Roure, Anna De Liddo (notes), Phil Bourne, Paolo Ciccarese, David Shotton, Herbert Van de Sompel, Tim Clark, Gully Burns, Udo Hahns. Below we include the list of main discussion items. The list of tools and notes are available at

https://sites.google.com/site/futureofresearchcommunications/force11-tools-framework

and the participants hope to put there a systematic profiling of the tools later.

What are the communication artefact we use in science?

What are the communication functionalities and their integration?

What is the lifecycle of Digital Artefact?

The lifecycle of a Digital Artefact includes the following stages:

1. Registration
2. Certification
3. Archiving
4. Rewarding
5. Enactment of the Digital Artefact: presentations, videos
6. Writing
7. Discoursing
8. Reuse/reproduceable
9. Formal/informal
10. Granularity of publications

How would you alter these tools so that they may become more valuable for the publication lifecycle?

What are the main categories of tools for supporting the entire lifecycle of scientific publication?

Tools that deal with the Digital Artefact and that are used formally and informally to support the lifecycle. There is also another dimension that is the speed of production of Digital Artefact and their development, including the issue of granularity of Digital Artefact.

What are the main tools in place now to support the entire lifecycle of scientific communication?

1. Literature programming
2. Scientific publication
3. Spreadsheets
4. Reference management system
5. Web pages + Web sites
6. Powerpoint
7. Word/LaTeX Google docs
8. Supercomputing
9. Gmail
10. Digital library

11. Poster
12. Analysis workflow + R scripts, codes
13. Amazon for papers/books
14. Catalogues: s/w library, N/F, Yellow pages
15. YouTube
16. Recordings/broadcast/webinars of talks and presentations
17. Dropbox/SlideShare/Flickr/Twitter
18. Terminologies, thesauri, mapping, ontology
19. Search services
20. Analytical tools to survey the landscape, understand the science landscape, i.e mapping and research literature mapping (Compendium, Cohere, knowledge mapping tools)
21. Technologies thesauri
22. Hubs for communication: centres of communities (automated versions of it)
23. What are the more formal tools
24. EasyChair: Conference reviewing tools
25. Journals
26. Grant repositories/applications: generating documents
27. Database schema, data repositories
28. Google+, Facebook, social networks
29. Learned society
30. Conference call (Skype)
31. Publishers
32. Chat
33. Directories of WhoIs/yellow pages

How can these tools be categorised?

- Social Technologies
- Info tech-tool

Off the shelf: What is different in how those tools are used in scholarly communication, compared to other forms of informal communication?

What needs to be added to make of this tools recognised scientific tools: i.e., so that tweets on the last paper you published would be considered by your boss

- Twitter
  - Self-promotion
  - Conference reporting
  - Community intelligence
  - Data observations, cities sensors
  - Reluctant to negative critics
- Dropbox
  - View data in real-time
  - Easy data maintenance
  - Writing

What are the properties necessary to move a tool from formal to informal tool for scientific communication/publishing?

1. Citeability
2. Preservation
3. Highly shareable
4. Known provenance
5. Trustable
6. Accessibility
7. Stability
8. Granularity
9. Cost (or lack of it)
10. Speed
11. IP restrictions
12. Inherent rewardability
13. Annotatabilility
14. Protectability
15. Staking claims
16. Portability
17. Palpability
18. Easiness
19. Capacity
20. Multimediality

What are the categories of tools that are emerging?

1. Communication Instant Discourse
2. Training tools
3. Document composition, editing, authoring
4. Sensemaking
5. Scientific publication/research sharing
6. Preservation/storage
7. Presentation
8. Search tools
9. Digital artefact/ file sharing
10. Terminology services
11. Curation: metadata/indexing/ managing tools
12. Social
13. Certification tools and commenting
14. Execution tools

What are the Media Types of Digital Artefacts?

1. Image
2. Video
3. Text
4. Code
5. To be continued...

Figure 1 A Wordle map from the notes of day 1 represents the most frequent keywords mentioned in the discussion

Back office

1. Citation
2. Identifiers
3. Interoperability and best practices
4. Capability matrix: a map of the skills and tools we have in the group, for understanding, when I need something, what are the interoperability conventions between the tools
5. Machine actionability
6. Economic sustainability and community involvement (SWAT analysis)
7. Problem of VERSIONING at all layers (FLUX) – What is the CONTRACT OF INTEROPERABILITY SERVICES? What is the change protocol/standards of a tool? What is the contractual cleanliness and coherence of the tool? (Within the group)
8. Making use of cutting edge computer science technologies

6.2.1 Actionable recommendations to

6.2.1.1 Funders, policy makers and reviewers of projects

- Reward the funding of tweaks recomposition and interoperability of already existing software. Reward REUSE and REPURPOSE
- Proactively identify missing components and services that proposals should be focusing on. Assess risk. Do not leave all to market!
- Source Material:
  - Software sustainability
  - Putting all the recommendations in an e-infrastructure policy document
- Specific fund archives (AHRC counter example)
- Citation effort
- Best strategies models for assessing the success of scholarly communication
- Identify the obstacles to scholarly communication

6.2.1.2 Scholars

- Put your data in a open repository and cite it and include it in your CV
- Promote tools and propagate practice to scholars
- Get your colleagues to do the same
- Complain and engage in the battle (senior scholars to advocate and promote sharing and defend young scholars that do that by rewarding them for doing that)
- Enlightenment

6.2.1.3 Publishers

- No walled gardens
- Metadata/splash pages should be open including references
- Allow open annotation schemes and name entities access
- Enable citeability of components
- Provide APIs and encourage developers to build applications
- Provide a unified standards
- Exposing content for text mining
- Embrace linked data
- Expose item level download data

Following these recommendations will drive better bigger and access to scholarly contents.

6.2.1.4 Technology developers

- Place your software in the Force11 roadmap and framework at at https://sites.google.com/site/futureofresearchcommunications/force11-tools-framework
- Reuse of existing components and standards
- Collaborate to develop new components that do not exist yet
- Place your software in the value chain of improving research and science communication
- Encourage "enlightened self-interest" in your users

6.2.2 Vision: Making scholarship useful and usable

An interoperable serviced based ecosystem of sustainable core components as the basis for a healthy, innovative and vibrant market of interoperable and usable tools fit for scholarship in the 21st century. These tools and technologies will exploit the full potential of information and communication technologies to serve and not hinder scholarship, thus improving the quality and productivity and dissemination of research. This ecosystem will provide a basis for more rapid and cost-effective innovation of software for scholarship.

6.2.3 Questions raised during the discussion

- How much does the research domain matter when thinking about new publication forms?
- How do we collaborate effectively?
- How we relate the kind of tool we have and integrate them so that scientists can play with them?
- We need to formulate a use cases, we could generate a vision on what happen if we put all these things together
- What we are already doing in this community to improve scientific publishing and what we could do next?
- How do we maximise the input and first pass of information curation?

- What is the role of curators to validate NLP results?
- How do we have researchers buying in new ways of publishing?
- We should aim to be contagious? People can register and share their things
- Issues of time
- How rapid should the science production loop be, from data to publication to science communication?
- There are pre- and post- publication aspects, the quicker a publication can be devolved the faster is the impact on people citing that data.
- What do researchers would need to know from others?
- What are the bottlenecks of open science
- Sharing is a bottleneck: scientists are not available to share, they are fine to collaborate and publish but not share, because there is no recognition and there are potentially negative effects. There is no policy and penalty
- We need funding body and journals that penalise those who do not share
- We can invent mechanism to detect who does not share, i.e if a publication derives from a work which is not collaborative, probably, it does not connect to the other research artefacts out there
- We do not have a good value proposition: reservation for self use, knowing what I have in my lab
- What do scientists need?
- Recovery and archive of data, plus access control to data
- Productivity, I want to be helped into publishing more
- How do we make the literature more effectively used, and how we make people understandable and useful to them.
- People do not like paper summarisation, they do not trust the conceptual model presented they think it is limited
- There are no credits for doing annotation, knowledge curation or any kind of paper summary
- What are the incentive we propose for doing this activities?
- Information complexity I want to be helped to read the right papers in the right (also interdisciplinary thing)
- Would be good knowing what the most relevant paper are.

6.3 Business models for the research communications in the future

This group aimed at brainstorming on possible business models for the research communication, taking into account the changes happening in scholarly publishing nowadays. This group included the following participants: Bradley P. Allen (notes), Aliaksandr Birukou, Philip E. Bourne, Leslie Chan, Olga Chiarcos, Robert Dale, Eve Gray, Paul Groth, Ivan Herman, Eduard H. Hovy, Fiona Murphy, David S. H. Rosenthal (chair), Jarkko Siren. Below we reproduce the summary of the discussion (also found at http://bit.ly/tyaWcL). More notes can be found at http://bit.ly/usiQOE.

Building a sustainable approach to research communications of the future will require the exploration of the space of potential business models. By business model, we mean a conceptual description of how an organisation provides value to customers — and gets paid for doing so. This last consideration of describing how the money flows is key to

understanding and resolving the sustainability and access issues that dog the ecosystem of researchers, institutions, publishers and funding agencies today.

In keeping with the ideas discussed at the Workshop about the future of research communications, the group focused on the notion of the research object as the contained of information being communicated. A research object is composed of one or more of the following types of sub-objects:

- Documents (textual, multimedia, pictures, etc)
- Experimental data
- Methods and procedures
- Relationships among constituents
- Context metadata
- Asset metadata
- Relational metadata
- Provenance

The group used the Business Model Generation [2] methodology to describe the space of potential business models for research communication in the future. Specifically we developed a set of optional choices for elements of the nine components of a Business Model Canvas [1]. These are:

- Value Propositions: What value is being delivering to the customers?
- Customer Segments: Who pays for that value?
- Channels: How is this value delivered to the paying customers?
- Customer Relationships: How is the relationship with the customers managed, and by whom?
- Revenue Streams: In what ways do customers pay us for this value, and optionally, how much?
- Key Resources: Who and/or what is required to build and operate the systems and organisations need to deliver the value?
- Key Activities: What tasks need to performed to deliver the value to the customers?
- Key Partners: Who are the key partners needed for the organisation to be able to deliver value?
- Cost Structure: What costs does the organisation incur to operate and deliver the value?

Table 1 illustrates the sketch of the business model designed during the working group.

Table 1 A sketch of the business model in Business Model Canvas format.

Business Model Component	**Possible Values**
Value Proposition	Seamless Management of Research Objects • Discovery • Preservation • Version control • Exploration & Integration • Metrics • Review, Evaluation, Annotation Seamless Management of Researchers • Reputation • ID

	• Profile • Aggregation, Syndication • Personalisation
Customer Segments	One of: • Creators • Funding agencies • Consumers (Researchers, Public, Industry) • Evaluators (Reviews, Tenure, Regulators) • Advertisers • Sponsorship

Customer Relationships
(who is accountable to the customer)
	One of:
	• Institutional support organisation (universities, research organisations,...)
	• Independent non-profit support organisation
	  • learned societies
	  • foundations
	  • self-organising communities
	• For profit publishers

Channels	One of:
	• Software-as-a-service platform
	• Direct software distributor
	  • shrink-wrap
	  • open source

Revenue Streams	Metered access to objects • "All you can eat" (one time, recurring, ...) • per object • per use Payment for service & support • subscription (one time, recurring, ...) • pay per call Software purchase
Key Activities	All of: • Build and maintain platform • Run platform • Develop and support communities
Key Resources	One or more of: • Platform/software developers • Operations staff • Content experts • Community managers • Marketing, PR • Business development
Key Partners	One or more of: • Learned societies • Funding agencies • Institutions (universities, research institutes, ...)

	• Subject matter experts • General public (crowdsourcing, citizen science, professional/amateur collaboration) • Government
Cost Structure	All of: • Hardware • Software • Network • Power • Staff costs • Market communications
Business Model Component	**Possible Values**

6.4 Assessment and Impact

This group aimed at identifying the critical issues pertaining to the research assessment and impact. The following people took part in this group: Eve Gray, Laura Czerniewicz, Ivan Herman (chair), Herbert van den Sompel (notes), Michael Kurtz, Jarkko Siren, Peter van den Besselaar, Anita de Waard. Below we include the list of main issues discussed. More notes can be seen at https://sites.google.com/site/futureofresearchcommunications/contributions-1/contributions.

6.4.1 Opening statements

- Current assessment mechanism is counter productive to scholarly communication. Need to make policy makers realise and accept that. Only formal citations count. Not other impact.
- What is impact assessment? Assess based on what? Do we need assessment of individuals?
- Impact factor doesn't work for across disciplines. Metrics on people or on artefacts?
- Perspective should be about value and how value relates to business and impact. Measure value! But how?
- Scholarly communication system is skewed by impact assessment as it is.
- The system is counter productive. But measures are essential because of assessing individuals, setting funding policy. Question how to come up with other metrics that can be generated in an open and scaleable way. Question how to get those metrics accepted.

6.4.2 Questions raised

- Do institutions, funding agencies base decisions on impact factor? Not element in decision whether a project gets funded, but does it play role in setting funding policies?
- Do we need to also talk about e.g. service to community as part of assessment? Is that science communication?
- What are metrics to assess research communication system, rather than to assess individuals?

6.4.3 Issues

- Need a multidimensional metrics model to count various things. If possible, the model should apply across disciplines. Simplicity of metric is important.
- What are those new dimensions? Is Altmetrics an answer?
- Why current system is broken?
  - Africa can not publish in ranking journals even if paper is about millions of people dying of some disease
  - The way we conducts science has changed so fundamentally that a metrics mechanism that ignores this change is totally passe
  - Real impact is manifested in different ways now (e.g., we know who the core players are in a scientific community and that is not based on "objective" metrics)
- The stellar researchers are known by their community. All the others not necessarily. Metrics can help.
- Innovations systems thinking. Research => Patent => Commercial. Need to change that thinking.
- Accessibility of metrics (or data from which to derive metrics) across systems is big issue, e.g. download data not consistently available; API to obtain metric only allows limited amount of calls a day.
- Author disambiguation – ORCID?
- Reputation management
- Need to define output types and metrics for output types

6.4.4 Possible dimensions

- How do we measure how research contributes to society (e.g., development goal in Africa)
- Netherlands: "evaluate research in context" effort. Quality of communication between research and community at large determines societal impact.
- local versus global impact
- economical impact
- quality of communication to general public
- measures depend on goals. In many cases citations are good. But, for example, in nursing, readership becomes important.
- need to be able to get at metrics otherwise you have done nothing
- download counts (better to measure social impact). Can be gamed. Can use under right conditions.
- crowd sourcing evaluation (e.g. Faculty of 1000)
- used for teaching (knowledge with of being transferred)
- used in lectures
- general level of reuse
- openness

The problem we see with Impact Factor and other simple metrics are individually taken with grain of salt. But if we would use multiple dimensions we might get a more just system. Decisions makers may choose which dimensions to use.

References

1 Wikipedia: http://en.wikipedia.org/wiki/Business_Model_Canvas, last checked 2011-11-04.
2 Alexander Osterwalder and Yves Pigneur, *Business Model Generation*, John Wiley and Sons, 2010.

7 Relevant links

- A huge collection of relevant links is maintained at http://bit.ly/tFQnkL.
- Future of Research Communication and e-Scholarship (FORCE11) website: http://force11.org/
- Force11 Manifesto *"Improving Future Research Communication and e-Scholarship"*: http://force11.org/node/1688

8 Agenda

time	August 15	August 16	August 17	August 18
8:00 – 9:00		Breakfast	Breakfast	Breakfast
9:00 – 10:30		The Past: Herman (Chair) De Waard, Buckingham Shum, Rosenthal	The Future: Hovy, Clark, Decker, De Roure, Rogers	Presentations Working groups 1, 2, 3
10:30 – 10:45		Break	Break	Break
10:45 – 12:15		The Present: Shotton, Groth, Murphy, Bourne, Kurtz	Working groups II	Presentations Working groups 4,5
12:15 – 13:00		Lunch/Demo's	Lunch/Demo's	Lunch
13:00 – 14:00	Arrive, register, settle in	Network, email, leisure; demo's	Network, email, leisure; demo's	Summary of items from working group, Action items/Calendar for next steps
14:00 – 15:30		The Present: Vision, Neylon; Planning working groups	The Future: Hovy, Clark	
16:00 – 18:00	Welcome/Introductions	Working groups I	Working groups III	Departure
18:00 – 19:30	Dinner	Dinner	Dinner	
19:30 – 20:30	Discuss goals for the week, divide into Working groups	Recap; touch base Working groups, settle questions	Recap day; plan calendar/tasks after Dagstuhl	
20:30	Wine and cheese (and music)	Wine and cheese (and music)	Presentation of working groups – prequel	

Working groups:

number	planned name	final name
1	Research data & code	Data
2	Assessment and impact	Assessment and impact
3	New forms and tools	Tools and technologies
4	Business models	Business models for the research communications in the future
5	Social platform & continuity	*this group was merged with other groups*

Participants

- Bradley P. Allen
Elsevier – Manhattan Beach, US
- Aliaksandr Birukou
CREATE-NET – Povo, Trento, IT
- Judith A. Blake
The Jackson Laboratory – Bar Harbor, US
- Philip E. Bourne
UC San Diego, US
- Simon Buckingham Shum
The Open University – Milton Keynes, GB
- Gully Burns
Univ. of Southern California – Marina del Rey, US
- Leslie Chan
University of Toronto, CA
- Olga Chiarcos
Springer-Verlag – Heidelberg, DE
- Paolo Ciccarese
Harvard University, US
- Timothy W. Clark
Harvard Medical School & Massachusetts General Hospital, US
- Laura Czerniewicz
University of Cape Town, ZA
- Robert Dale
Macquarie University, AU
- Anna De Liddo
The Open University – Milton Keynes, GB
- David De Roure
University of Oxford, GB
- Anita De Waard
Elsevier Labs – Jericho, US
- Stefan Decker
National University of Ireland – Galway, IE
- Alex Garcia Castro
Universität Bremen, DE
- Carole Goble
University of Manchester, GB
- Eve Gray
University of Cape Town, ZA
- Paul Groth
VU University Amsterdam, NL
- Udo Hahn
Universität Jena, DE
- Ivan Herman
W3C/CWI – Amsterdam, NL
- Eduard H. Hovy
Univ. of Southern California – Marina del Rey/ISI, US
- Michael J. Kurtz
Harvard-Smithsonian Center for Astrophysics, US
- Fiona Murphy
Wiley-Blackwell, UK
- Cameron Neylon
Rutherford Appleton Lab. – Didcot, GB
- Steve Pettifer
University of Manchester, GB
- Mike W. Rogers
European Commission Brussels, BE
- David S. H. Rosenthal
Stanford University Libraries, US
- David Shotton
University of Oxford, GB
- Jarkko Siren
European Commission Brussels, BE
- Herbert van de Sompel
Los Alamos National Lab., US
- Peter van den Besselaar
Free Univ. – Amsterdam, NL
- Todd Vision
University of North Carolina – Chapel Hill, US

Report from Dagstuhl Seminar 11332

Security and Rewriting

Edited by

Hubert Comon-Lundh[1], Ralf Küsters[2], and Catherine Meadows[3]

1 ENS – Cachan, FR
2 Universität Trier, DE
3 Naval Research – Washington, US, meadows@itd.nrl.navy.mil

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 11332 "Security and Rewriting".

Seminar 15.–18. August, 2011 – www.dagstuhl.de/11332
1998 ACM Subject Classification F.4.2 Grammars and Other Rewriting Systems
Keywords and phrases Rewriting, Security, Access Control, Protocol Verification
Digital Object Identifier 10.4230/DagRep.1.8.53
Edited in cooperation with Benedikt Schmidt

1 Executive Summary

Hubert Comon-Lundh
Ralf Küsters
Catherine Meadows

Security is a fundamental problem in computer science. Because of the possible catastrophic problems that can arise from poor security, the ability to mathematically prove and formally verify the security of computer systems is vital. Research has been ongoing in this area since the 1970's and has been the subject of many Dagstuhl seminars, including (in the last three years) "Theoretical Foundations of Practical Information Security" (November 2008)[1], and "Formal Protocol Verification Applied" (October 2007)[2].

Research on formal proofs of security has demonstrated that rewriting techniques, including completion, narrowing, unification, play a central role in this area, for example:

- Formally modeling the properties of cryptographic primitives: fundamental properties of the cryptographic primitives are presented as algebraic theories and used as a basis for security analysis.
- Automatically proving security protocols: both the protocol and the attacker's possible actions can be modeled as a rewrite system and unification algorithms play a central role in the security analysis of such systems.
- Formally specifying and verifying security policies: the (possibly infinite) set of allowed transitions may be a represented as a finite rewriting system. The views on a documents or a class of documents may be specified by tree automata.

[1] http://www.dagstuhl.de/08491
[2] http://www.dagstuhl.de/07421

- Modeling and analysis of other security-critical applications: rewrite techniques are used to model and analyze the security of web services, APIs and systems for access control.

The goal of this seminar was (i) to bring together researchers who have a background in rewriting techniques and researchers who have a background in security applications (or both) (ii) to answer, among others, the following questions:

- Are there specific problems in rewriting that stems from security applications and would deserve some further research? For instance, do the algebraic theories of cryptographic primitives enjoy some specific properties? Are there restrictions that are relevant to the applications and that would yield more efficient unification/rewriting algorithms? Which new challenges does the addition of an arbitrary attacker context bring? What are the specific problems on tree automata that are brought by security applications?
- What are the limits/successes/failures of rewriting techniques in security applications?
- What are the emerging research areas at the intersection of security and rewriting?

2 Table of Contents

3 Overview of Talks

3.1 Automated Analysis of Access Control Policies

Alessandro Armando (University of Genova, IT)

Joint work of Armando, Alessandro; Ranise Silvio
Main reference F. Alberti, A. Armando, and S. Ranise, "Efficient Symbolic Automated Analysis of Administrative Role Based Access Control Policies," Proc. of the 6th ACM Symposium on Information, Computer, and Communications Security (ASIACCS), Hong Kong, March 22–24, 2011.
URL http://dx.doi.org/10.1145/1966913.1966935

Automated techniques for the security analysis of Role-Based Access Control (RBAC) access control policies are crucial for their design and maintenance. In this talk, we describe an automated symbolic security analysis technique for Administrative RBAC policies. A class of formulae of first-order logic is used to symbolically encode both the policies and the administrative actions upon them. State-of-the-art automated theorem proving techniques are used (off-the-shelf) to mechanize the security analysis procedure. Besides discussing the assumptions for the effectiveness and termination of the procedure, we demonstrate its efficiency through an extensive empirical evaluation.

3.2 Model Checking of Browser-based Single Sign-On Protocols: an Experience Report

Alessandro Armando (University of Genova, IT)

Joint work of Armando, Alessandro; Carbone, Roberto; Compagna, Luca; Cuellar, Jorge; Giancarlo Pellegrino; Sorniotti, Alessandro
Main reference A. Armando, R. Carbone, L. Compagna, J. Cuellar, G. Pellegrino, and A. Sorniotti, "From Multiple Credentials to Browser-based Single Sign-On: Are We More Secure?" Proceedings of the 26th IFIP TC-11 International Information Security Conference (SEC 2011), pp. 68–79Luzern, Switzerland, June 7–9, 2011.
URL http://dx.doi.org/10.1007/978-3-642-21424-0_6

I will report on my experience in formal modeling and model checking one of the most popular web-based SSO protocols, the SAML 2.0 Web Browser SSO Profile. I will outline the challenges posed to model checkers by this type of security protocols. I will then discuss our findings: the discovery of a serious man-in-the-middle attack on the SAML-based SSO for Google Apps and, more recently, the discovery of an authentication flaw in the prototypical use case described in the SAML standard.

3.3 Real-world Key Exchange versus Symbolic Analysis - Where do we stand?

Cas Cremers (ETH Zürich, CH)

Joint work of Basin, David, Cremers, Cas; Feltz, Michele; Meier, Simon, Schmidt, Benedikt

Real-world applications that require key exchange often use protocols from international standards, such as IEEE P1363, NIST SP800-56, ANSI, or ISO. The design of these protocols is driven by cryptographers; choosing among the proposed protocols also involves engineering considerations, such as efficiency. We study the relation between the desired properties of such protocols, and cryptographic security notions for key exchange, such as the CK and eCK models. We provide symbolic formalizations of the majority of these properties with corresponding automatic tool support. Our symbolic methods have lead to several new results in the cryptographic domain. Although our methods are sufficiently mature to be useful to the designers of key exchange protocols, there are also some types of attack on relevant security properties that are outside of the scope of our symbolic methods.

3.4 Security Analysis in Geometric Logic: To Models via Rewriting

Dan Dougherty (Worcester Polytechnic Institute, US)

Starting from the observations that strand spaces—embodying protocol executions—are models for a certain first-order language and that security goals can be captured by first-order sentences, we present an approach to protocol analysis based on model-finding.

A central role is played by "geometric logic, a logic of finite observations previously studied in the context of denotational semantics.

An important strategic aspect of our approach is the interplay between (i) certain canonical theories incorporating inductive definitions and well-foundedness assumptions and (ii) purely first-order companion theories supporting a model-finding method based on The Chase. The latter is a model-finding method jointly inspired by database theory and rewriting in quasi-equational theories.

3.5 The Margrave policy-analysis tool

Dan Dougherty (Worcester Polytechnic Institute, US)

Joint work of Dougherty, Dan; Fisler, Kathi; Krishnamurthi, Shriram; Nelson, Timothy

Margrave is a policy-analysis tool providing query-based verification and query-based views of policies. It supports "change-impact analysis", allowing a user to compare the effects of multiple policies. It supports reasoning about the combined effects of policies written in different configuration languages, such as a firewall filter and a static router, or a firewall combined with an access-control policy (perhaps on a different component).

In this talk we will focus on the foundations of Margrave: model-finding in order-sorted first-order logic, and describe how Margrave relies on a finite-model theorem, whose proof is based on tree automata.

3.6 Rewrite Specifications of Access Control Policies in Distributed Environments

Maribel Fernandez (King's College – London, GB)

Joint work of Fernandez, Maribel; Bertolissi, Clara

We present a meta-model for access control that takes into account the requirements of distributed environments, where the resources and the access control policies may be distributed across several sites. This distributed meta-model is an extension of the category-based meta-model studied in [1], from which standard centralised access control models such as MAC, DAC, RBAC, Bell-Lapadula can be derived. We use term rewriting to give an operational semantics to the distributed meta-model, and then show how various distributed access control models can be derived as instances.

References

1 Clara Bertolissi and Maribel Fernández. *Category-Based Authorisation Models: Operational Semantics and Expressive Power.* In Proc. of 2nd Int'l Symposium on Engineering Secure Software and Systems (ESSoS), 2010, Pisa, Italy, February 3-4, 2010. Lecture Notes in Computer Science, vol. 5965, pp. 140–156, Springer.

3.7 Logical Protocol Analysis for Authenticated Diffie-Hellman

Joshua D. Guttman (Worcester Polytechnic Institute, US)

Joint work of Guttman, Joshua D.; Dougherty, Daniel J.
Main reference unpublished

Diffie-Hellman protocols for authenticated key agreement construct a shared secret with a peer using a minimum of communication and using limited cryptographic operations. However, their analysis has been challenging in computational models and especially in symbolic models.

In this paper, we develop a framework for protocol analysis that combines algebraic and strand space ideas. We show that it identifies exact assumptions on the behavior of a certifying authority. These assumptions establish the confidentiality and authentication properties for two protocols, the Unified Model and Menezes-Qu-Vanstone (MQV). For MQV, we establish a stronger authentication property than previously claimed, using a stronger (but realistic) assumption on the certifying authority.

Verification within our framework implies that the adversary has no strategy that works uniformly, independent of the choice of the cyclic group in which the protocol operates. Indeed, we provide an equational theory which constitutes an analysis of these uniform

strategies. We provide an abstraction, the notion of indicator, which leads to easy proofs of protocol correctness assertions.

Computational soundness awaits further investigation.

3.8 Formal Specification and Analysis of Security Policies

Helene Kirchner (INRIA, FR)

Joint work of Bourdier, Tony; Cirstea, Horatiu; Jaume, Mathieu
Main reference T. Bourdier, H. Cirstea, M. Jaume, H. Kirchner, "Formal Specification and Validation of Security Policies," 4th Canada-France MITACS Workshop on Foundations & Practice of Security (FPS 2011), Paris (France), May 12–13, 2011. Lecture Notes in Computer Science, vol. 6888.
URL http://hal.inria.fr/inria-00507300/PDF/FormalValidation2010.pdf

A general approach to model a secured system is to consider a transition system whose transitions are guarded by a security policy.

More precisely the evolution of security information in the system is described by transitions triggered by authorization requests and the policy is given by a set of rules describing the way the corresponding decisions are taken.

Policy rules are constrained rewrite rules whose constraints are first-order formulas on finite domains, which provides enhanced expressive power compared to classical security policy specification approaches like the ones using Datalog, for example.

Such specifications have an operational semantics based on transition and rewriting systems and are thus executable.

Non-termination, conflicts or under-specification of policies are easy to detect. Syntactic conditions over the policy rules, satisfied by a large class of policies, can be given for ensuring consistency and completeness.

The presented framework provides ability to

- Specify a security system and an associated security policy: this clear separation is useful for reusability and composition.
- Execute the specification of a secured system, since it can be compiled into term rewriting rules.
- Analyse the specification related properties: experiments can be performed with existing tools, like model-checkers or invariant verifiers.
- Check security requirements, even when they are expressed in a different specification.

More details can be found in [1].

References

1 Bourdier, T., Cirstea, H., Jaume, M., Kirchner, H.: *Formal Specification and Validation of Security Policies,*, 4th Canada-France MITACS Workshop on Foundations & Practice of Security (FPS 2011), Paris (France), May 12-13, 2011. Lecture Notes in Computer Science, vol. 6888.

3.9 Transforming Password Protocols to Compose

Steve Kremer (ENS - Cachan, FR)

Joint work of Chevalier, Céline; Delaune Stéphanie; Kremer, Steve;
Main reference Céline Chevalier, Stéphanie Delaune, and Steve Kremer, "Transforming Password Protocols to Compose," Proceedings of the 31st Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS'11), Mumbai, India, December 2011, Leibniz International Proceedings in Informatics (LIPIcs), Leibniz-Zentrum für Informatik. To appear.

Formal, symbolic techniques are extremely useful for modelling and analyzing security protocols. They improved our understanding of security protocols, allowed to discover flaws, and also provide support for protocol design.

However, such analyses usually consider that the protocol is executed in isolation or assume a bounded number of protocol sessions. Hence, no security guarantee is provided when the protocol is executed in a more complex environment.

In this paper, we study whether password protocols can be safely composed, even when a same password is reused. More precisely, we present a transformation which maps a password protocol that is secure for a single protocol session (a decidable problem) to a protocol that is secure for an unbounded number of sessions. Our result provides an effective strategy to design secure password protocols: (i) design a protocol intended to be secure for one protocol session; (ii) apply our transformation and obtain a protocol which is secure for an unbounded number of sessions. Our technique also applies to compose different password protocols allowing us to obtain both inter-protocol and inter-session composition.

3.10 A procedure for verifying equivalence-based properties of cryptographic protocols

Steve Kremer (ENS – Cachan, FR)

Joint work of Chadha, Stefan; Ciobaca, Stefan; Kremer, Steve

Indistinguishability properties are essential in formal verification of cryptographic protocols. They are needed to model anonymity properties, strong versions of confidentiality and resistance against offline guessing attacks. We present a procedure for verifying observational equivalence for determinate cryptographic protocols when the number of sessions is bounded. For determinate cryptographic protocols, observational equivalence coincides with trace equivalence. The cryptographic protocols are formalized in a fragment of applied pi-calculus without replication and all communication is over public channels.

As in applied pi-calculus, this fragment is parametrized by a first-order sorted term signature and an equationnal theory which allows formalization of algebraic properties of cryptographic primitives. Our procedure is sound and complete for subterm convergent theory which can model several used cryptographic primitives.

The procedure is based on a fully abstract modelling of the traces of a bounded number of sessions of the protocols in first-order Horn clauses on which a dedicated resolution procedure is used to decide both reachability properties and observational equivalence. Currently we were unable to prove termination of the procedure which is conjectured. The procedure has been implemented and tested in the KiSs tool.

3.11 Asymmetric Unification: A New Unification Paradigm for Cryptographic Protocol Analysis

Christopher Lynch (Clarkson University – Potsdam, US)

A new extension of equational unification, called *asymmetric unification*, is introduced.

In asymmetric unification, the equational theory is divided into a set R of rewrite rules and a set E of equations. A substitution σ is an asymmetric unifier of a set of equations P iff for every $s = t \in P$, $s\sigma$ is equivalent to $t\sigma$ modulo $R \cup E$, and furthermore $t\sigma$ is in $E\backslash R$ normal form.

This problem is at least as hard as the unification problem modulo $R \cup E$ and sometimes harder. The problem is motivated from cryptographic protocol analysis using unification techniques for handling equational properties of operators such as XOR.

3.12 My Own Little Hilbert's Program

Sebastian Moedersheim (Technical University of Denmark – Lyngby, DK)

Main reference Sebastian Moedersheim, "Diffie-Hellman without Difficulty," FAST 2011, to appear
URL http://www.imm.dtu.dk/ samo/dh-d.pdf

There have been a number of relative soundness results that make verification of protocols easier, basically showing that certain restrictive models are without loss of generality. The first of these results concerns only the use of certain typing restrictions in protocol analysis, but it turns out that very similar concepts are helpful for compositional reasoning as well. Another application is to reduce the amount of algebraic reasoning in protocols such as those based on Diffie-Hellman: allowing for the use of pattern matching when receiving Diffie-Hellman half-keys even though actually in reality the agent could not check for such patterns. What these results have in common is to exploit good protocol engineering practice for protocol verification: a good protocol suite should be designed such that every message and non-atomic message part has a unique interpretation or type. My own Hilbert's program tries to recognize all protocols as samt und sonders wohlgetypt (completely well-typed); I present some examples where this typing is sound and set out the challenge to find interesting counter-examples.

3.13 On the Complexity of Linear Authorization Logics (Preliminary Results)

Vivek Nigam (LMU München, DE)

Linear Authorization Logics have been used in the Proof-Authentication Framework (PCA) to specify effect-based policies, such as policies involving consumable resources. A key requirement of PCA is the need to construct proof-objects, which requires proof search. We

demonstrate that the propositional multiplicative fragment of linear authorization logics is undecidable.

Therefore, PCA using simple linear policies might already not be feasible. However, we also identify a first-order fragment for which the provability problem is decidable. In particular, we capitalize on capitalizes on the recent work on the decidability of the reachability problem for MSR systems with balanced actions to identify a fragment of linear authorization logics that is PSPACE-complete, namely the fragment of balanced bipolars. This is accomplished by first formalizing a (sound and complete) correspondence between linear authorization logic provability and MSR reachability and then showing that MSR reachability is PSPACE-complete.

3.14 Timed Collaborative Systems

Vivek Nigam (LMU München, DE)

Time is often a key component used in specifying the rules and the requirements of a collaboration. In this talk, we report on our initial steps in extending with explicit time our previous work on models for collaborative systems with confidentiality. In particular, we discuss conditions for PSPACE-completeness of previous compliance problems extended with explicit time. Finally, we identify and discuss in detail a possible application of our model, namely for clinical investigations.

3.15 Unbounded Verification and Falsification of Protocols that use Diffie-Hellman Exponentiation

Benedikt Schmidt (ETH Zürich, CH)

Joint work of Schmidt, Benedikt; Meier, Simon; Cremers, Cas; Basin, David

We present a method for the automatic analysis of protocols specified as multiset rewriting rules. Our approach accounts for algebraic properties of Diffie-Hellman Exponentiation. We support an expressive fragment of two-sorted first-order logic to formalize security properties. Given a protocol and a property such that our method terminates, it either returns a counterexample or proves that all traces of the protocol satisfy the security property. To illustrate the applicability of the method, we sketch the analysis of the NAXOS authenticated key exchange protocol.

3.16 Intruder Deduction in Sequent Calculus

Alwen Tiu (Australian National University - Canberra, AU)

Main reference Alwen Tiu, Rajeev Gore and Jeremy Dawson, "A proof theoretic analysis of intruder theories," Logical Methods in Computer Science, 6(3), 2010.
URL http://dx.doi.org/10.2168/LMCS-6(3:12)2010

An approach to modeling the intruder in analysing security protocols is to formalise the capabilities of the intruder via a natural deduction calculus, or equivalently, via a rewrite system capturing the proof normalisation processes of the natural deduction system. In proof theory, it is well known that natural deduction systems can be equivalently presented in Gentzen's sequent calculus.

Sequent calculus enjoys the so-called subformula property, which in many cases entail bounded proof search. Some preliminary results in using sequent calculus as a framework to structure proof search for intruder deduction problems, under a range of intruder models involving extensions of Dolev-Yao model with AC-convergent theories, are presented. Extensions of these sequent-calculus-based techniques to solve deducibility constraints and symbolic trace equivalence problems are also discussed.

3.17 Automated Validation of Trust and Security in the Internet of Services

Luca Vigano (Università di Verona, IT)

Joint work of AVANTSSAR
URL http://www.avantssar.eu

The AVANTSSAR Project (www.avantssar.eu) has developed an automated platform that provides a rigorous technology for the formal specification and Automated VAlidatioN of Trust and Security of Service-oriented ARchitectures. This technology, which is being tuned on a number of relevant industrial case studies so to allow for the migration into the development process for software solutions for the Internet of Services, aims at speeding up the development of new network and service infrastructures, enhance their security and robustness, and increase the public acceptance of emerging IT systems and applications based on them. I will present some of the main techniques and technologies that are part of the AVANTSSAR Platform and some of the case studies it has been applied on. In particular, to illustrate the platform on the field, I will discuss some of our industrial case studies, including a brief account of our formal analysis of a SAML Web Browser Single Sign-On Protocol. I will also present the first results of the SPaCIoS Project (www.spacios.eu) that has been combining the AVANTSSAR Platform with techniques and tools for penetration and vulnerability testing to allow for the automated validation of services at provision and consumption time.

3.18 An Environmental Paradigm for Defending Security Protocols

Luca Vigano (Università di Verona, IT)

Joint work of Fiazza, Maria-Camilla; Peroli, Michele; Vigano, Luca
Main reference Maria-Camilla Fiazza and Michele Peroli and Luca Vigano, Attack Interference in Non-Collaborative Scenarios for Security Protocol Analysis, Proceedings of SECRYPT 2011, 144–156, SciTePress, 2011

Although computer security typically revolves around threats, attacks and defenses, the subfield of security protocol analysis (SPA) has so far focused almost exclusively on the notion of attack. We wish to show that such focus on attacks depends on few critical assumptions that have been characteristic of the field and have governed its mindset, approach and developed tools. We motivate that indeed there is room in SPA for a fruitful notion of defense and that the conceptual bridge lies in the notion of multiple non-collaborating attackers. To support SPA for defense-identification, we propose a paradigm shift that brings security closer to the conceptual tools of fields that have a rich notion of agent, such as robotics and AI — in contrast to the weak notion of agent that is typical of SPA. These fields, however, lack the required understanding of how to instantiate their tools in a manner that is informative for security analysis. Hence, our main contribution is a novel paradigm for defending security protocols, based on importing into SPA well-established techniques and tools from robotics and AI. At the conceptual and methodological level these techniques form a cohesive picture, which can prompt a parallel development in our understanding of protocols as environments.

Participants

- Myrto Arapinis
University of Birmingham, GB
- Alessandro Armando
University of Genova, IT
- Yannick Chevalier
Université Paul Sabatier – Toulouse, FR
- Hubert Comon-Lundh
ENS – Cachan, FR
- Cas Cremers
ETH Zürich, CH
- Stéphanie Delaune
ENS – Cachan, FR
- Dan Dougherty
Worcester Polytechnic Inst., US
- Santiago Escobar
Universidad Politécnica – Valencia, ES
- Maribel Fernandez
King's College – London, GB
- Cédric Fournet
Microsoft Research UK – Cambridge, GB
- Joshua D. Guttman
Worcester Polytechnic Inst., US
- Hélène Kirchner
INRIA, FR
- Steve Kremer
ENS – Cachan, FR
- Ralf Küsters
Universität Trier, DE
- Christopher Lynch
Clarkson Univ. – Potsdam, US
- Catherine Meadows
Naval Res. – Washington, US
- José Meseguer
Univ. of Illinois – Urbana, US
- Sebastian Mödersheim
Technical University of Denmark – Lyngby, DK
- Paliath Narendran
Univ. of Albany – SUNY, US
- Vivek Nigam
LMU Múnchen, DE
- Michaël Rusinowitch
INRIA Lorraine, FR
- Mark D. Ryan
University of Birmingham, GB
- Ralf Sasse
Univ. of Illinois – Urbana, US
- Benedikt Schmidt
ETH Zürich, CH
- Helmut Seidl
TU München, DE
- Slawomir Staworko
University of Lille III, FR
- Carolyn L. Talcott
SRI – Menlo Park, US
- Sophie Tison
Université de Lille I, FR
- Alwen Tiu
Australian National University – Canberra, AU
- Tomasz Truderung
University of Wroclaw, PL
- Luca Vigano
Università di Verona, IT
- Christoph Weidenbach
MPI für Informatik – Saarbrücken, DE

Report from Dagstuhl Seminar 11341

Learning in the context of very high dimensional data

Edited by

Michael Biehl[1], Barbara Hammer[2], Erzsébet Merényi[3], Alessandro Sperduti[4], and Thomas Villmann[5]

1 **University of Groningen, NL,** `m.biehl@rug.nl`
2 **Universität Bielefeld, DE,** `bhammer@techfak.uni-bielefeld.de`
3 **Rice University, US,** `erzsebet@rice.edu`
4 **University of Padova, IT,** `sperduti@math.unipd.it`
5 **Hochschule Mittweida, DE,** `thomas.villmann@hs-mittweida.de`

With real world data it all stands and falls
if knowledge and insight are truly your goals.
But if you like greatly
delicious sweet pastry
you might just as well resort to Swiss Rolls.

Michael Biehl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 11341 "Learning in the context of very high dimensional data". The aim of the seminar was to bring together researchers who develop, investigate, or apply machine learning methods for very high dimensional data to advance this important field of research. The focus was be on broadly applicable methods and processing pipelines, which offer efficient solutions for high-dimensional data analysis appropriate for a wide range of application scenarios.

Seminar 22.–26. August, 2011 – www.dagstuhl.de/11341
1998 ACM Subject Classification I.2.6 [Artificial Intelligence] Learning
Keywords and phrases Curse of dimensionality, Dimensionality reduction, Regularization Deep learning, Visualization
Digital Object Identifier 10.4230/DagRep.1.1.67

Learning in the context of very high dimensional data, *Dagstuhl Reports*, Vol. 1, Issue 1, pp. 67–95
Editors: M. Biehl, B. Hammer, E. Merényi, A. Sperduti, and T. Villmann

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Michael Biehl
Barbara Hammer
Erzsébet Merényi
Alessandro Sperduti
Thomas Villmann

Goals of the seminar

Rapidly increasing sensor technology, greatly enhanced storage capabilities, and dedicated data formats have lead to a dramatic growth of the size of electronic data available today and, even more so, its dimensionality. Examples include diverse formats such as spectral data, micro- and macroarrays, biotechnological sequence data, or high resolution digital images. Due to its dimensionality and complexity, these data sets can hardly be addressed by classical statistical methods; nor do standard presentation and visualization tools allow an adequate direct inspection by humans. Thus, the need for efficient and reliable automatic processing and analysis tools for very high dimensional data arises in different areas such as bioinformatics, medicine, multi-band image analysis, robotics, astrophysics, geophysics, etc.

The aim of the seminar was to bring together researchers who develop, investigate, or apply machine learning methods for very high dimensional data to advance this important field of research. The focus was put on broadly applicable methods and processing pipelines which offer efficient solutions for high-dimensional data analysis appropriate for a wide range of application scenarios.

Questions tackled in the seminar included the following areas:

1. **Sparse representation and regularization:**
   a. Which general principles (such as information theory, preservation of inherent data structures) offer suitable frameworks in which to achieve a compact representation of high dimensional data? Is it possible to turn these general principles into efficient algorithmic form as mathematical regularization conditions?
   b. Which models are suitable to represent high dimensional data in a dense form (such as prototype based methods, functional data representation, dedicated algebraic structures, decomposition methods)? What are their adaptive parameters and how can they be adapted?
   c. How can the number of free model parameters be restricted by regularization such that the available data provides a sufficient statistics for the resulting model? Is it possible to derive explicit mathematical bounds on the generalization ability?
2. **Dedicated metrics and kernels for high-dimensional data:**
   a. How can the inherent non-Euclidean structure be inferred from the data in the presence of high dimensionality? Which aspects of particular relevance for the application should be emphasized by the corresponding similarity structure and how can this information be estimated with robust statistical tools?
   b. How can this information be embedded into metrics or kernels? Do there exist particularly suited approaches for high dimensional data of specific form such as kernels which make use of sparsity or functional dependencies of the data? How can this be realized algorithmically in an efficient way regarding the high dimensionality?

   c. Is it possible to partially automate the detection of a suitable similarity structure for the analysis tool and accompany this with guarantees such as consistency, or bounds on the generalization ability in the context of high-dimensionality?
3. **Efficient realizations:**
   a. How can robust learning algorithms be designed for the model parameters, ones that can deal with noise and uncertain data, missing values, etc., particularly pronounced in high dimensional data? Are these techniques insensitive with respect to the choice of the metaparameters (such as learning rate, degree of regularization), such that generic methods suitable for non-experts in the field result?
   b. How can the methods be realized efficiently in the context of very high dimensionality? Methods which are linear in the number of dimensions are probably already too slow in this context.
   c. Can the learning algorithms be realized in such a way that adaptation to new data and life-long learning become possible? What are characteristic time scales required for learning certain parameters in dependence of the data dimensionality?
4. **Evaluation of methods:**
   a. What are inherent evaluation criteria for the reliability of the models, also when the number of data points is small compared to the dimensionality? What are stable conformal predictors for model adequacy and accuracy?
   b. How can the results be presented to experts such that humans can judge the reliability and quality of the model? How can, in turn, user feedback be integrated into the models?
   c. Can simplifying models of learning scenarios give insight into the performance of practical algorithms? Which information visualization methods are suitable in this context?

Structure

39 experts from 12 different countries joined the seminar, including a good mixture of established scientists and promising young researchers working in the field. Thereby, the special interests of the researchers ranged from dedicated algorithmic design connected to diverse areas such as dimensionality reduction, data visualization, metric learning, functional data analysis, to various application scenarios including diverse areas such as the biomedical domain, hyperspectral image analysis, and natural language processing.

This setup allowed us to discuss salient issues in a way that integrated perspectives from several points of views and scientific approaches, thereby providing valuable new insights and research contacts for the participants. Correspondingly, a wide range of topics was covered during discussions and presentations in the seminar.

During the week, 29 talks were presented which addressed different aspects of how to deal with high dimensional data and which can be grouped according to the following topics:

- Dimensionality reduction techniques and evaluation measures
- Biomedical applications
- Distances, metric learning, and non-standard data
- Functional data processing
- Probabilistic models for dimensionality reduction
- Feature selection and sparse representation of data

The talks were divided into a variety of tutorials which gave introductory overviews and opinions on important research directions and shorter talks which focused on specific recent (partially yet unpublished) scientific developments.The talks were supplemented by vivid discussions based on the presented topics as well as the traditional social event on Wednesday afternoon in the form of a visit to the beautiful town of Trier.

Results

A variety of open problems and challenges came up during the week. The following topics were identified as central issues in the context of the seminar:

- **Desired properties of dimensionality reduction techniques, possibilities of their formal evaluation:** The topic of dimensionality reduction and data visualization has been addressed in several presentations including several tutorial talks.
  It became apparent that the topic is currently a very rapidly emerging field in machine learning, with a manifold of advanced algorithms being published in the recent literature. Key issues, however, remain a challenge: to use advanced methods in applications, there is the need for widely parameterless techniques, clear interpretability of the results, and comfortable usage e.g. regarding processing speed or uniqueness and robustness of the results.
  For these reasons, advanced methods are often not used in practice.
  It has been discussed that the desired properties and results of dimensionality reduction techniques depend on the given task at hand and cannot universally be formalized. Nevertheless, there is a need for formal evaluation methods of dimensionality reduction to compare techniques, and to guide parameter choice and optimization.
  Promising general evaluation schemes have been proposed in recent years as presented in the seminar, but an extensive evaluation of their suitability is so far lacking.
- **Good scientific benchmarks and evaluation criteria:** In this context, is has been raised that good, accessible benchmarks are rare. Albeit high dimensional electronic data are ubiquitous, these data often require complex preprocessing, they do not allow evaluation of formal methods due to the lack of objective evaluation information, or they are even sometimes subject to restrictions. For that reason, real life data are partially not accessible, and there is the risk that methods are over-adapted according to the available benchmarks which do not necessarily mirror the demands in practical applications. During the seminar, however, it has been raised that quite a few benchmarks have become available in the context of contest data.
- **Where to use complex models as compared to well-established linear techniques:** During the seminar, it became apparent that there is a gap in between advanced techniques proposed in the context of machine learning for high dimensional data and methods which are actually used in application domains such as biomedical data analysis. Often, in practical applications simple linear techniques seem sufficient to reliably detect important and interpretable information in high dimensional data collections.
  A variety of reasons has been discussed in this context: in particular in bioinformatics, the technology to gather data and large data collections are often comparably novel such that information still lies ‘at the surface’ of the data.
  For very high dimensional data, linear techniques sometimes seem the only methods for which sufficient reliability can be guaranteed, more complex nonlinear methods likely focusing on noise in the data due to the lack of appropriate regularizers which suite

the given setting. In this context, it has been raised that the embedding of data into high-dimensions where linear methods often suffice constitutes one of the most prominent approaches to actually solve standard nonlinear problems, popular examples in this context being the support vector machine, the extreme learning machine, or reservoir computing.
Further, linear techniques seem to focus on 'universally important' issues which are relevant independent of the context due to universal statistical properties. For more advanced techniques, domain knowledge is required to set up the models or to interpret the results in a reliable way. This argument has been substantiated in the seminar by several presentations. For example, knowledge about biological networks and metabolic pathways can be integrated into biological data processing and it can greatly enhance the performance.

- **How to deal with complex structures:** It has been raised that an intelligent preprocessing of the data is often more important than the choice of the model. Alternatively, data can be tackled with appropriate problem adapted metrics, followed by rather simple machine learning techniques. In the seminar, quite a number of complex data formats have been presented in applications, such as e.g. data with inherent functional form connected to spectrometry data. Besides the necessity to come up with suitable metrics, this also leads to very interesting theoretical problems. For example, it can be proved that it is not possible to learn in the space of functions at all, unless very strong requirements are fulfilled.
  Nevertheless, impressive applications have been achieved in this context. Thus, it seems worthwhile to investigate which constraints are fulfilled in practical applications such that learnability is guaranteed. On the other side, a variety of powerful metric adaptation schemes exist ranging from fast and efficient convex techniques to nonlinear cost functions. In most cases, however, a simple Mahalanobis distance is used and, often, only basic machine learnings are integrated such as simple k nearest neighbor.
  Further, a unifying theory and guidelines, which technique is best suited in which scenarios, remain unsolved problems.
- **How to model in the correct way for very high dimensions:** Statistics being the universal background for almost all machine learning techniques, it has been raised that statistical models are often rather uniform in their principled design, not taking advantage of the rather flexible way to model dependencies of data and dimensions. While it is common in the context of simple PCA to swap the role of input dimensions and data points in case of very high dimensions, more complex nonlinear models stick to the classical setting as used in the case of comparably low dimensionality. It could be worthwhile to put different principled modeling paradigms into general patterns which allow to reformulate established techniques such that they become suitable for very high dimensional data.

Altogether, the seminar opened quite a few perspectives pointing into important research directions in the context of very high dimensional data.

2 Table of Contents

3 Overview of Talks

3.1 Some biology applications for the analytically minded

Gyan Bhanot (Rutgers University – Piscataway, US)

Joint work of Bhanot, Gyan and many others
Main reference Greenbaum B.; Levine AJ., Bhanot G. and Rabadan R., PLoS Pathogens. 2008 Jun 6;
G. Alexe, et al., Can. Res.67, 10669-76, 2007;
Alexe G, et al, 2008, J. Mol Evol, 67 (5), 465-87.

At the seminar, I talked about several successful projects where simple analytical methods and ideas from bioinformatics were able to reveal novel patterns in biological data. Many of these discoveries were validated in wet lab experiments conducted with biologist and medical colleagues and led to novel understanding. I will describe briefly the projects I discussed.

1. The first project was entitled: "Inferring the past" and involved a simple use of PCA and consensus clustering on mitochondrial sequence data to infer the phylogeny of human migration patterns (Alexe G, Vijaya-Satya R, Seiler M, Platt D, Bhanot T, Hui S, Tanaka M, Levine AJ, Bhanot G. PCA and Clustering Reveal Alternate mtDNA Phylogeny of N and M Clades. 2008, J. Mol Evol, 67 (5), 465–487. PMID: 18855041). We found that the tree emerged as a hierarchical pattern of embedded clusters at each level of PCA. Data bootstrapping and consensus clustering was used to compute the robustness of branches and identify the most informative SNPs.
2. The second project was entitled: "PCA, Clustering reveal clinical subtypes in breast cancer". Here, we used the same methodology to identify clinically relevant classes of breast cancer. The significant discovery here was that HER2+ breast cancers, which are more aggressive, have two subtypes, distinguished by the presence or absence of a lymphocytic infiltration into the tumor which is visible in pathology specimens using a simple staining assay. The clinical utility of this discovery was that the tumors with lymphocytic infiltration were responsive to Herceptin (a drug used in treating HER2+ breast cancers) while those without the infiltrate were less responsive. The paper on this work is :
   G. Alexe, et al. 'High Expression of Lymphocyte Associated Genes in Node Negative HER2+ Breast Cancer correlates with lower Recurrence rates.' Cancer Research, 67, 10669–10676, 2007.
3 The third project was entitled: "Evolution and Mimicry in Influenza and Other RNA Viruses" which was based on:Greenbaum B, Levine AJ, Bhanot G and Rabadan R, PLoS Pathogens. 2008 Jun 6;4(6):e1000079. In this paper, we used a simple permutation method on CpG in the 3rd and 1st coding positions on the viral genome to identify significant di-nucleotide patterns under selection in Flu. We showed that the 1918 Flu virus (H1N1) has evolved to lower CpG content to make it less visible to the immune system. Our research suggested that specific Toll like receptors must exist which can identify ssRNA in cells. This prediction was subsequently validated by experiments. We were also able to characterize the virulence of emerging strains based on the CpG content and flanking sequence context.

3.2 Admire LVQ: The DREAM6 AML prediction challenge

Michael Biehl (University of Groningen, NL)

Joint work of Biehl, Michael; Bunte, Kerstin; Schneider, Petra

We briefly present and discuss our entry to the DREAM6 AML prediction challenge, 2011.

The construction of feature vectors from the flow cytometry counts based on statistical moments is briefly described. Generalized Matrix Relevance Learning is used to obtain a classification scheme and evaluated with respect to ROC performance. The obtained relevance profile provides further insight into the data and the nature of the problem.

This brief presentation is meant as an addendum to the talk of Marc Strickert, who presents an alternative approach to the problem.

3.3 Adaptive Matrices for Color Texture Classification

Kerstin Bunte (University of Groningen, NL)

Joint work of Bunte, Kerstin; Giotis, Ioannis; Petkov, Nicolai; Biehl, Michael
Main reference Kerstin Bunte, Ioannis Giotis, Nicolai Petkov and Michael Biehl, "Adaptive Matrices for Color Texture Classification, Computer Analysis of Images and Patterns," 14th International Conference, CAIP 2011, Seville, Spain, pp. 489–497.
URL http://dx.doi.org/10.1007/978-3-642-23678-5_58

In this paper we introduce an integrative approach towards color texture classification learned by a supervised framework. Our approach is based on the Generalized Learning Vector Quantization (GLVQ), extended by an adaptive distance measure which is defined in the Fourier domain and 2D Gabor filters. We evaluate the proposed technique on a set of color texture images and compare results with those achieved by simple gray value transformation on the color images with a comparable dissimilarity measure and the same filter bank. The features learned by GLVQ improve classification accuracy and they generalize much better for evaluation data previously unknown to the system [1].

References

1 Kerstin Bunte, Ioannis Giotis, Nicolai Petkov, and Michael Biehl. *Adaptive Matrices for Color Texture Classification*. Computer Analysis of Images and Patterns – 14th International Conference, CAIP 2011, Seville, Spain, August 29–31, 2011, Proceedings, Part II

3.4 Challenges of feature representation in Natural Language Processing tasks

Richard Farkas (Universität Stuttgart, DE)

Natural Language Processing (NLP) tasks are usually traced back to classification or ranking problems. The feature representation of instances (which are words, sentences or documents) usually consists of several millions of features. The features set is heterogeneous, the

features can be binary (e.g. occurrence of a particular word) or continuous (e.g. information theoretic measures calculated on unlabeled large document sets); they are usually highly inter-dependent; they are sparse. In spite of these special characteristics, the NLP community employs simple and standard machine learning techniques to handle the feature space. In this talk I focus on the challenges of feature representation in NLP tasks and introduce some promising (task-dependent) approaches to handle the high-dimensionality of these problems.

3.5 Optimization of Parametrized Divergences in Fuzzy c-Means

Tina Geweniger (University of Groningen, NL)

Joint work of Geweniger, Tina; Kästner, Marika; Villmann, Thomas

In many scientific fields like biology, medicine, geology etc. clustering of data plays an important role. Sets of multi-dimensional data samples are grouped to detect or visualize the underlying structure. One family of algorithms are prototype based methods, where each prototype represents one cluster center.

Famous representatives are c-means and neural gas. In my talk I focused on the fuzzy c- means as a variant of the standard c-Means algorithm determining fuzzy memberships to the cluster centers. Usually the Euclidean distance is applied to calculate distances between prototypes and data samples. Yet, if these samples represent functions, i.e. high-dimensional data vectors with spatially correlated components, generalized divergences could be a more appropriate distance measures. Furthermore, incorporation of relevance learning, i.e. weighting of function intervals, leads to improved clustering solutions. We modified the fuzzy c-means such, that the two concepts - divergences and relevance learning - are integrated and presented the theory in combination with some examples. To compare the respective results different cluster evaluation measures for fuzzy clustering were applied and discussed briefly.

3.6 Kernel t-SNE

Andrej Giesbrecht (Universität Bielefeld, DE)

Joint work of Giesbrecht, Andrej; Lueks, Wounter; Hammer, Barbara

The visualization of high-dimensional data is an important challenge in the current data mining research field. A low-dimensional representation of data (e.g. 2D) is an accessible and intuitive way for humans to analyze the information hidden in the complex data. For this reason many algorithms emerged recently to address this problem. However most of these techniques have quadratic or even cubic complexity. One way to overcome this problem is to train the mapping on a manageable subset of the data and then project the remaining data using this mapping. In our contribution we investigate this approach using t-distributed Stochastic Neighbor Embedding [1], which is one of the most well-known methods in the area. We compare the performance of the out-of-sample extension of t-SNE, which is based on the gradient descend, with a simple interpolation technique and SVM regression. All the three

methods are using the result of the trained t-SNE mapping. Also we propose a new method called kernel t-SNE, which is based on t-SNE. Instead of computing the low-dimensional representation of the training data, it learns a kernel mapping, which can be directly applied to the new data. The results show that the interpolation technique and kernel t-SNE achieve good performance, comparable to the out-of-sample extension of t-SNE, being faster than the latter.

References

1 L. van der Maaten and G. Hinton. *Visualizing data using t-sne.* Journal of Machine Learning Research, vol. 9, pp. 2579–2605, November 2008.

3.7 Functional Relevance Learning in Generalized Learning Vector Quantization

Marika Kästner (Hochschule Mittweida, DE)

Joint work of Kästner, Marika; Villmann, Thomas
Main reference M. Kästner, T. Villmann, "Functional Relevance Learning in Learning Vector Quantization for Hyperspectral Data," Proceedings of WHISPERS 2011.

Classification accuracy of functional data frequently depends on only a few dimension windows distinguishing different classes. Hence, classifier systems should not only achieve a good performance but also figure out what is essential for this decision. Learning vector quantization is a robust prototypebased classification method which, together with the relevance learning strategy, assesses the relative contribution of spectral bands for efficient classification. Original relevance learning is based on the scaled Euclidean distance, weighting each band independently. This yields a vectorial relevance profile indicating those dimensions which distinguish the classes best. We propose the use of the functional Sobolev distance instead of the Euclidean, together with a relevance function as profile taking into account the functional properties of data. The relevance function is a superposition of a small set of simple basis functions like Gaussians or Lorentzians. In this way the number of parameters to be optimized in relevance learning is drastically decreased such that an inherent stabilization is obtained while the classification accuracy level is retained. We demonstrate the ability of the functional approach for ground cover classification of an AVIRIS hyperspectral data set (Lunar Crater Volcanic Field). In particular it is emphasize model sparsity in terms of structural sparsity and feature selection.

References

1 M. Kästner, B. Hammer & T. Villmann. Generalized Functional Relevance Learning Vector Quantization. In M.Verleysen (Edt.) *Proc. of European Symposium on Artificial Neural Networks (ESANN'2011)*, pages 93.98, Evere, Belgium, 2011.
2 M. Mendenhall and E. Merényi. Relevance-based feature extraction for hyperspectral images *IEEE Transactions on Neural Networks*, 19(4), 658-672, 2008.
3 B. Hammer & T. Villmann. Generalized Relevance Learning Vector Quantization. *Neural Networks*, 15,1059-1068, 2002.
4 A. S. Sato & K. Yamada. Generalized Learning Vector Quantization *Advanced in neural information processing systems*, 7, 423-429,1995.

3.8 Generative modeling of dependencies between high-dimensional data sets

Arto Klami (Aalto University, FI)

Canonical correlation analysis (CCA) finds maximally correlating linear components of two data sets. Formulated as explicit maximization of correlation, the model severely overfits to small sample sizes and high-dimensional data.

Solving the same problem via a Bayesian formulation helps for small sample sizes, but the associated generative model cannot be reliably estimated for high-dimensional data, severely limiting the applicability of Bayesian CCA. In this talk we show how Bayesian CCA can be used also for high-dimensional data by re-formulating the model as a simpler matrix factorization with group-wise sparsity structure. We present an efficient variational Bayesian algorithm for inference, provide a new kind of multi-set CCA extension, and demonstrate the model in analysis of, e.g., high-dimensional brain imaging data.

3.9 Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants

John A. Lee (University of Louvain, BE)

Joint work of Lee, John A.; Verleysen, Michel;

Dimensionality reduction aims at representing high-dimensional data in low-dimensional spaces, mainly for visualization and exploratory purposes. As an alternative to projections on linear subspaces, nonlinear dimensionality reduction, also known as manifold learning, can provide data representations that preserve structural properties such as pairwise distances or local neighborhoods. Very recently, similarity preservation emerged as a new paradigm for dimensionality reduction, with methods such as stochastic neighbor embedding and its variants. Experimentally, these methods significantly outperform the more classical methods based on distance or transformed distance preservation.

This talk explains both theoretically and experimentally the reasons for these performances. In particular, it details (i) why the phenonomenon of distance concentration is an impediment towards efficient dimensionality reduction and (ii) how SNE and its variants circumvent this difficulty by using similarities that are invariant to shifts with respect to squared distances. The talk also proposes a generalized definition of shift-invariant similarities that extend the applicability of stochastic neighbour embedding to noisy data.

3.10 Scale-independent quality criteria for dimensionality reduction

John A. Lee (University of Louvain, BE)

Joint work of Lee, John A.; Verleysen, Michel;

Dimensionality reduction aims at representing high-dimensional data in low-dimensional spaces, in order to facilitate their visual interpretation. Many techniques exist, ranging from simple linear projections to more complex nonlinear transformations. The large variety of methods emphasizes the need of quality criteria that allow for fair comparisons between them. This talk extends previous work about rank-based quality criteria and proposes to circumvent their scale dependency. Most dimensionality reduction techniques indeed rely on a scale parameter that distinguishes between local and global data properties. Such a scale dependency can be similarly found in usual quality criteria: they assess the embedding quality on a certain scale. Experiments with various dimensionality reduction techniques eventually show the strengths and weaknesses of the proposed scale-independent criteria.

3.11 Multiple Instance Learning

Marco Loog (TU Delft, NL)

Joint work of Loog, Marco; Tax, David M.J.; Sørensen, Lauge; Cheplygina, Veronika; Lee, Wan-Jui; Duin, Robert P.W.;

We present multiple instance learning, or rather multiset classification, as a technique that can play a key role in modeling high dimensional, complicated classification tasks. Where in the standard classification every object is described by a single feature vector, in the multiset classification setting, every object is represented by a collection, a multiset, of feature vectors.

The number of feature vectors may differ from object to object. Example problems where a single feature vector typically does not suffice, but certain object parts can be robustly characterized by feature vectors, are web page classification, document labeling, movie rating, music classification, and gesture recognition. It is classically applied to the problem of molecule classification, but nowadays many of its applications can be found within the fields of computer vision, medical image analysis, and computer-aided diagnosis.

As an illustration, we sketch a complete, yet basic, medical image classification pipeline. Going through various dimensionality reduction steps, our original four million dimensional problem is reduced to a multiset classification problem, which dimensionality is in the order of tens. A condensed overview of different approaches to multiset classification is provided. We advocate the use of fusion-based and dissimilarity-based approach to multiset classification, both of which rely on standard classification methods, avoiding special purpose multiple instance learning routines.

References

1 T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

2 M. Loog and B. van Ginneken. Static posterior probability fusion for signal detection: applications in the detection of interstitial diseases in chest radiographs. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 1, pages 644–647, 2004.

3 O. Maron and A.L. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the 15th International Conference on Machine Learning*, 1998.
4 S. Ray and M. Craven. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 697–704, 2005.
5 L. Sørensen, M. Loog, D. Tax, W.J. Lee, M. de Bruijne, and R. Duin. Dissimilarity-based multiple instance learning. In *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS, pages 129–138. Springer, 2010.
6 D. Tax and R. Duin. Learning curves for the analysis of multiple instance classifiers. In *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS, pages 724–733. Springer, 2008.
7 D.M.J. Tax, M. Loog, R.P.W. Duin, V. Cheplygina, and W.-J. Lee. Bag dissimilarities for multiple instance learning. In *1st SIMBAD Workshop*, LNCS. Springer, 2011.

3.12 On the Problem of Finding the Least Number of Features by L1-Norm Minimisation

Thomas Martinetz (Universität Lübeck, DE)

Joint work of Martinetz, Thomas; Klement, Sascha
Main reference Sascha Klement, Thomas Martinetz, "A new approach to classification with the least number of features," ICMLA 2010, Washington, D.C, USA, 12–14 December, 2010, IEEE Computer Society, pp. 141–146.

We proposed the so-called Support Feature Machine (SFM) as a novel approach to feature selection for classification. It relies on approximating the zero-norm minimising weight vector of a separating hyperplane by optimising for its one-norm. In contrast to the L1-SVM it uses an additional constraint based on the average of data points.

In experiments on artificial datasets we observe that the SFM is highly superior in returning a lower number of features and a larger percentage of truly relevant features. Here, we derive a necessary condition that the zero-norm and 1-norm solution coincide. Based on this condition the superiority can be made plausible.

3.13 How to Evaluate Dimensionality Reduction? - Improving the Co-ranking Matrix

Bassam Mokbel (Universität Bielefeld, DE)

In order to make very high-dimensional data accessible for visual inspection and exploration, dimensionality reduction tools can embed the data points in a low-dimensional space, and thereby produce a visualization, e.g., in the Euclidean plane. The existing methods for dimensionality reduction all have unique characteristics and favor certain data properties, often these characteristics are controlled via rather unintuitive parameters. However, since the general problem is ill-posed, it is unclear which is the best embedding solution for a given visualization task. Recently, this has inspired the development of quality assessment

measures, in order to evaluate visualization results independently from the methods' inherent criteria. Several quality measures can be (re)formulated based on the so-called co-ranking matrix, which subsumes all rank errors, i.e., differences between the ranking of distances from every point to all others, comparing the low-dimensional representation to the original data. Some measures use a parameter K to divide the co-ranking matrix at the K-th row and column into rectangular submatrices, calculating weighted combinations from the sums of each submatrix's elements. The evaluation process typically involves plotting a graph over several (or even all possible) settings of K. Considering simple artificial examples, we argue that this parameter controls two notions at once, that need not necessarily be combined, and that the rectangular shape of submatrices is disadvantageous for an intuitive interpretation of the parameter. We debate that quality measures, as general and flexible evaluation tools, should have parameters with a direct and intuitive interpretation as to which specific error types are tolerated or penalized for a particular visualization task. Therefore, we propose to replace the parameter K with two distinct parameters to control these notions separately, and introduce a differently shaped weighting scheme on the co-ranking matrix. The two new parameters can then directly be interpreted as a threshold up to which rank errors are tolerated, and a threshold up to which the rank-distances are significant for the quality evaluation. Moreover, we propose a color representation of local quality to support the evaluation process for a given mapping, where every point is colored according to its local contribution to the overall quality value.

3.14 DLVQ and its application to Crop Surveillance

Ernest Mwebaze (Makerere University – Kampala, SAF)

Joint work of Mwebaze, Ernest; Biehl, Michael; Quinn, John A.

DLVQ is a variant of LVQ that uses divergences in the distance measure. This is applicable for non-negative usually normalized data for example histograms and spectra data. We use DLVQ with increasingly complicated 'distance' formulations specified by a combination of partial distances related to more than one type of data. We apply this to Crop Surveillance by representing cassava plant leaf images as histograms that are combined in a compound distance and trained using DLVQ. We implement this classifier on $100 Android mobile phones for automated visual based classification. We also propose dimensionality reduction using causal analysis and present some on-going problems and datasets.

3.15 Is the k-NN classifier in high dimensions affected by the curse of dimensionality?

Vladimir Pestov (University of Ottawa, CA)

Main reference Vladimir Pestov, "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?" arXiv:1110.4347v1 [stat.ML]
URL http://arxiv.org/abs/1110.4347

There is an increasing body of evidence suggesting that exact nearest neighbour search in high-dimensional spaces is affected by the curse of dimensionality at a fundamental level. Does it necessarily mean that the same is true for k nearest neighbours based learning algorithms such as the k-NN classifier? We analyse this question at a number of levels and show that the answer is different at every layer that we peel. As our first main result, we show the consistency of a k approximate nearest neighbour classifier. However, the performance of the classifier in very high dimensions is provably unstable.

As our second major result, we point out that the existing model for statistical learning is oblivious of dimension of the domain and so every learning problem admits a universally consistent reduction to the one-dimensional case.

3.16 Bongard problems: learning in unlimited feature space

John Quinn (Makerere University – Kampala, SAF)

In machine learning, some or all aspects of the type of model and features used for any particular problem are usually selected through human judgment. It is interesting to consider the issues that would need to be solved in order to make entirely automated learning possible. In this talk I discuss Bongard problems, the essence of which is to find classification rules in a setting where the representation cannot be fixed a priori. The need to search for the right representation makes some of the issues in fully automated learning explicit.

3.17 Functional data analysis and learnability in arbitrary spaces

Fabrice Rossi (TELECOM-Paristech – Paris, FR)

I will summarize in this talk recent results about learning in infinite dimensional spaces and more generally in arbitrary spaces. The main motivation for studying this theoretical problem is given by functional data analysis: in this context, each observation point is a function and the goal is to learn e.g., to classify those functions. In practice, one might one for instance to detect some particular compound based on a near infrared spectrum of samples under study.

In practice, functional data are obtained as high dimensional vectors through a natural sampling strategy. For instance a function f is given by $(f(t_1), ..., f(t_d))$. This raises two

questions: 1) can we learn to classify exact functions (that is assuming that each function f is completely known) based on a finite learning set? 2) can we learn to classify exact functions based on a finite learning set of sampled functions?

I will first recall that many machine learning algorithms do not need strong assumptions on the input space, at least to be "conceptually" implemented. For instance the K nearest neighbors (KNN) algorithm uses only a dissimilarity on the data and can therefore be applied to any metric space. Linear models can defined simply in Hilbert spaces and then extended to nonlinear mapping using the classical multi-layer perceptron trick (see [1]).

Then I will introduce impossibility examples from [2] and [3]. Those papers show that KNN is not consistent in arbitrary metric spaces and therefore that it cannot be used to learn in this context. [4] shows that the classical kernel estimator is also non consistent in arbitrary metric spaces.

I will give conditions on the space that bring consistency back to KNN. Firstly it is well known since Cover and Hart [5] that the metric space has to be separable (http://en.wikipedia.org/wiki/Separable_space). This is a rather mild condition for functional spaces, but some important spaces such as the one of functions with bounded variations are not separable. Secondly, we need a complex condition that involves both the distribution of the data points and regression function, the so called Besicovitch condition, detailed in [2]. While this condition is automatically true in R^d, it is not in infinite dimensional spaces. Strong assumptions on the regression function (for instance continuity) and/or on the distribution of the data points (for instance a Gaussian distribution with eigenvalues that decrease exponentially quickly) are needed to ensure the consistency of the KNN rule in separable metric spaces of infinite dimension, i.e., in functional spaces considered in functional data analysis.

I will conclude the talk by briefly explaining how one can obtain consistency for functional data analysis with sampled functions. The main idea is to consider regular functions, for instance a Sobolev space such as H^2 (http://en.wikipedia.org/wiki/Sobolev_space). Then natural hypotheses can be used to obtain consistency, for instance a uniform bound on the second derivatives of the data points. Details can be found in [6]. Another solution consists in using projection of the data points on the first functions of a Hilbert basis, as described in [7] and [8].

References

1 Fabrice Rossi and Brieuc Conan-Guez *Functional Multi-Layer Perceptron: a Nonlinear Tool for Functional Data Analysis*. Neural Networks, volume 18, number 1, pages 45-60. January 2005. http://fr.arxiv.org/abs/0709.3642v1

2 F. Cérou and A. Guyader *Nearest Neighbor Classification in Infinite Dimension*. Neural Networks, volume 18, number 1, pages 45-60. January 2005. http://fr.arxiv.org/abs/0709.3642v1

3 Biau, G., Cérou, F. and Guyader, A. *Rates of convergence of the functional k-nearest neighbor estimate*.IEEE Transactions on Information Theory, Vol. 56, pp. 2034–2040. 2010. http://www.lsta.upmc.fr/BIAU/bcg.pdf

4 Abraham, C., Biau, G. and Cadre, B. *On the kernel rule for function classification, Annals of the Institute of Statistical Mathematics*. Vol. 58, pp. 619–633. 2006. http://www.lsta.upmc.fr/BIAU/abc5.p

5 Fabrice Rossi and Nathalie Villa-Vialaneix *Nearest Neighbor Pattern Classification*. T.M. Cover and P.E. Hart. IEEE Transactions on Information Theory, IT-13(1):21–27, January 1967. http://www.stanford.edu/~cover/papers/transIT/0021cove.pdf

6 *Consistency of Functional Learning Methods Based on Derivatives.* Pattern Recognition Letters, volume 32, number 8, pages 1197-1209. June 2011. http://fr.arxiv.org/abs/1105.0204
7 Biau, G., Bunea, F. and Wegkamp, M.H *Functional classification in Hilbert Spaces.* IEEE Transactions on Information Theory, Vol. 51, pp. 2163–2172. 2005. http://www.lsta.upmc.fr/BIAU/bbw.ps
8 Fabrice Rossi and Nathalie Villa-Vialaneix *Support Vector Machine For Functional Data Classification.* Neurocomputing, volume 69, number 7-9, pages 730-742.

3.18 Supervised learning of short and high-dimensional temporal sequences

Frank-Michael Schleif (Universität Bielefeld, DE)

Joint work of Schleif, Frank-Michael; Hammer, Barbara; Gisbrecht, Andrej

Temporal data, with many measurement points and only one or few variables are common in different applications and have been widely studied in the last years. The analysis of short temporal data with many variables is however a quite new field of research and mostly focused on unsupervised approaches like temporal clustering techniques. Here we review first approaches for such data and present a method for the supervised analysis of short and high-dimensional temporal data. The method is evaluated for different artificial data sets.

3.19 Acquisition and processing of high-dimensional data by means of hyperspectral imaging

Udo Seiffert (Fraunhofer IFF Magdeburg, DE)

Hyperspectral imaging is an evolving technology that provides a quantitative assessment of the molecular composition of a wide range of different samples in a non-invasive (optical) manner. The result of the image acquisition process is a stack of images containing the local distribution of reflection spectra of the recorded scenery. Each image pixel becomes a vector along the acquired wavelength range. In contrast to standard colour imaging or multispectral imaging this vector represents a dense and equidistant sampling of a wide wavelength range. This typically leads to a number of characteristic properties of hyperspectral data:

1. High-dimensionality, typically several hundred dimensions;
2. Individual positions of these vectors are rather sampling points of complex patterns than more or less independent features, leading to patterns instead of feature sets;
3. Extensive and proportionate sampling (large number of patterns) due to spatial resolution.

Machine learning offers a powerful framework for pattern recognition and statistical modelling within this context. In order to derive meaningful information from hyperspectral data and comprehensively exploit this imaging technology, novel and adapted data processing approaches and particularly learning paradigms are desired. Hence, hyperspectral image analysis offers both novel perspectives in an increasingly wide range of real-world applications and a challenging playground to develop and test novel methods in machine learning and beyond.

3.20 Functional MRI Analysis

Diego Sona (Fondazione Bruno Kessler – Trento, IT)

Joint work of Sona, Diego; Avesani, Paolo
Main reference Diego Sona, Paolo Avesani, "Feature Rating by Random Subspaces for Functional Brain Mapping," International Conference on Brain Informatics (BI), LNCS Vol. 6334, pp. 112–123, 2010.
URL http://dx.doi.org/10.1007/978-3-642-15314-3_11

Functional magnetic resonance imaging produces datasets presenting high dimensionality in the feature space, and low dimensionality in the sample space.

In the neuroscience community this issue is addressed with a simple and powerful univariate analysis approach, allowing to find areas in the brain maximally activated by specific cognitive or perceptual tasks. We may refer this analysis as "brain mapping". This approach however limits the retrieval of dependencies between the variables. The application of machine learning (ML) may solve this problem, however, two other difficulties arise. On one side ML models usually suffer the curse of dimensionality. On the other side, the experts always require maps indicating the areas of the brain relevant for the investigated cognitive function. In ML terms this corresponds to the set of features allowing a classifier to have good performance. For this reason we need models able to cope with high dimensionality and exhibiting the grouping effect, i.e., similar features must have similar relevance in the trained models.

This would allow to retrieve all relevant features (also the redundant ones) with a posterior analysis of the trained models.

3.21 Correlative matrix mapping connects high-dimensional data sets

Marc Strickert (Universität Siegen, DE)

Many tasks of data analysis concern the linking of high-dimensional data and their labels or, more generally, their multivariate regression targets.

Canonical correlation analysis is a common approach for mapping both vector spaces in a linear fashion into latent spaces where correlation is optimized.

The traditional approach often fails if the number of data points is smaller than the number of data or label dimensions. Furthermore, high-dimensional data, such as spectral measurements, are often redundant and may undergo transformations, while regression targets like metabolite concentrations are often acquired with greater efforts and should be kept constant.

Correlative matrix mapping (CMM) is an alternative formulation based on learning metrics for directed linear transformation from data to targets. Overfitting is reduced by integrating pairwise data relationships into the mapping: the pairwise distances of data transformed by a metric induced by quadratic forms is aimed to be in maximum correlation with the pairwise distances between the regression targets. Second-order learning, l-BFGS, is well-suited to optimize the required mapping parameters. Matrix ranks less than four related to the quadratic form can be used for directly visualizing the transformed discriminative data space. For further reducing overfitting in the CMM model, a very shallow network approach to sub-linear modeling by k-means clustering of the optimized matrix parameters

can be considered. CMM can faithfully address supervised visualizations of classification and regression problems, and it can also act for auto-association, that is, as alternative to principal component mappings. Because of the structural simplicity, optimized model parameters can be interpreted as (pairwise) attribute contributions of input data vectors.

Applications to the identification of molecular descriptors and of relevant document terms are provided in the references, and a MATLAB/GNU-Octave implementation is available at https://mloss.org as package CMM.

References

1 Axel J. Soto and Gustavo E. Vazquez and Marc Strickert and Ignacio Ponzoni. *Target-driven subspace mapping methods and their applicability domain estimation*. Molecular Informatics, 2011

2 Strickert, M.; Soto, A. J. & Vazquez, G. E.; Verleysen, M. (Ed.). *Adaptive matrix distances aiming at optimum regression subspace*.
European Symposium on Artificial Neural Networks (ESANN), D-facto Publications, 2010, 93-98

3 Soto, A. J.; Strickert, M.; Vazquez, G. E. & Milios, E. Butz, C. & Lingras, P. (Eds.). *Subspace Mapping of Noisy Text Documents*. Lecture Notes in Artificial Intelligence, Springer-Verlag Berlin Heidelberg, 2011, LNCS 6657, 377-383

3.22 Verification of Cluster Structure: Escalation of Need and Difficulty for Real, High-Dimensional Data, and Recent Developments

Kadim Taşdemir (EC Joint Research Centre – Ispra, IT) and Erzsébet Merényi (Rice University, US)

Joint work of Taşdemir, Kadim; Merényi, Erzsébet
Main reference Taşdemir, K., Merényi, E, "A Validity Index for Prototype-Based Clustering of Data Sets With Complex Cluster Structures," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 41(4), 1039–1053.

The purpose of this work is evaluation of unsupervised clustering results without reference data, i.e., quantification of how well the clusters returned by an algorithm fit the true data partitions. This is a fundamental challenge of clustering because the data structure and the number of clusters are unknown a priori. Cluster validity indices are commonly used tools for relative cluster validation: for ranking the quality of different clustering results. Real-world applications are increasingly dependent on automatic clustering algorithms for finding intricate structure in high-dimensional, large, complicated data spaces, consequently need reliable validation of the discovered structure. Since labeled reference data for problems with many clusters (let alone unexpected clusters) is rarely available, one has to increasingly turn to cluster validity indices. Many existing indices work well for simple data (clusters are well separated or have parametrical shapes or distributions). Most indices, however, do not work well for data sets with complicated cluster structure (variety of clusters of different shapes, sizes, densities, or overlaps), for one or more of the following reasons: The measures of within-cluster scatter and between-clusters separation - which are combined in various ways in the index formulae - are in most cases based on metric distances, thus directly depend on dimensionality.

They involve large numbers of pair wise distances, which results in poor scaling properties. Many use parametric assumptions, or work with extremes, consequently cannot handle irregular cluster structure.

To alleviate some of these problems, we present Conn_Index (Taşdemir & Merényi, 2009, 2011). It works with a connectivity-based similarity measure, CONN, derived from local density distribution and thus possesses considerable immunity to the curse of dimensionality. It is a prototype- based index, i.e., works only for prototype-based clustering (but for any Vector Quantization prototypes). A positive consequence is that it scales well (the number of pair wise distances that need to be computed scales linearly with the number of data points). Since the scatter and separation measures are density-based Conn_Index handles irregular structure quite well. Given that prototype-based clustering has significant performance advantages for huge data sets, development of prototype-based validity indices is strongly motivated. We demonstrate the superior performance of Conn_Index on simple synthetic data, on some of the UCI benchmark machine learning data sets, as well as on real hyperspectral images with complex cluster structure.

3.23 Topographic Mapping and Dimensionality Reduction of Binary Tensor Data of Arbitrary Rank

Peter Tino (University of Birmingham, GB)

Main reference J. Magut, P. Tino, M. Boden, H. Yan. "Multilinear Decomposition and Topographic Mapping of Binary Tensors," in Artificial Neural Networks (ICANN 2010), pp. 317–326, Lecture Notes in Computer Science, Springer-Verlag, LNCS 6352, 2010.

URL http://www.cs.bham.ac.uk/~pxt/TALKS/bin_tensor_talk.pdf

Current data processing tasks often involve manipulation of multi-dimensional objects - tensors. In many real world applications such as gait recognition, document analysis or graph mining (with graphs represented by adjacency tensors), the tensors can be constrained to binary values only. To the best of our knowledge at present there is no principled systematic framework for topographic maps and dimensionality reduction through decomposition of binary tensors. We propose to achieve this through a generalized multi-linear model for binary tensors.

In the model formulation, to account for binary nature of the data, each tensor element is modeled by a Bernoulli noise distribution. To extract the dominant trends in the data, we constrain the natural parameters of the Bernoulli distributions to lie in a sub-space spanned by a reduced set of basis tensors. Bernoulli distribution is a member of exponential family with helpful analytical properties that allow us to derive an iterative scheme for estimation of the basis tensors and other model parameters via maximum likelihood.

We evaluate and compare the proposed technique with existing real-valued tensor decomposition methods in two scenarios: (1) in a series of controlled experiments involving synthetic data; (2) on a real world biological dataset of DNA sub-sequences from different functional regions, with sequences represented by binary tensors.

3.24 Bayesian Models for Variable Selection that Incorporate Biological Information

Marina Vanucci (Rice University - Houston, US)

Main reference Stingo, F.C., Chen Y.A., Tadesse, M.G. and Vannucci, M., "Incorporating Biological Information into Linear Models: A Bayesian Approach to the Selection of Pathways and Genes," Annals of Applied Statistics, 2011, accepted.

URL http://www.stat.rice.edu/~marina/publications.html

The analysis of the high-dimensional genomic data generated by modern technologies, such as DNA microarrays, poses challenges to standard statistical methods. In this talk I will describe how Bayesian methodologies can be successfully employed in the analysis of such data. I will look at linear models that relate a phenotypic response to gene expression data and employ variable selection methods for the identification of the predictive genes. The vast amount of biological knowledge accumulated over the years has allowed researchers to identify various biochemical interactions and define different families of pathways. I will show how such information can be incorporated into the model for the identification of pathways and pathway elements involved in particular biological processes.

3.25 A brief tutorial on (linear) Distance Metric Learning

Kilian Weinberger (Washington University, US)

One of the fundamental questions of machine learning is how to compare examples. If an algorithm could perfectly determine whether two examples were semantically similar or dissimilar, most subsequent machine learning tasks would become trivial. For example, in classification settings, one would only require one labeled example per class and could then, during test-time, categorize all similar examples with the same class-label. An analogous reduction applies to regression if a continuous estimate of the degree of similarity were available.

It is not surprising that many popular machine learning algorithms, such as Support Vector Machines, Gaussian Processes, kernel regression , k-means or k-nearest neighbors (kNN) fundamentally rely on a representation of the input data for which a reliable, although not perfect, measure of dissimilarity is known. A common choice of dissimilarity measure is an uninformed norm, like the Euclidean distance. Here it is assumed that the features are represented in a Euclidean subspace in which similar inputs are close and dissimilar inputs are far away. Although the Euclidean distance is convenient and intuitive, it ignores the fact that the semantic meaning of "similarity" is inherently task- and data- dependent. Often, domain experts adjust the feature representations by hand - but clearly, this is not a robust approach. It is therefore desirable to learn the metric (or data representation) explicitly for each specific application.

Guided by this motivation, a surge of recent research has focused on learning metrics for the kNN classification (or regression) rule.

Learning a metric instead of (or in addition to) a classifier can have substantially different implications than learning only the classifier directly. For example, in contrast to most

classifiers, metrics can generalize to class categories that were unknown during training time. In this tutorial I review a series of recently published algorithm that learn a Mahalanobis metric explicitly from the data. I highlight general trends and outline future directions of metric learning as a research field.

3.26 Relational Extensions of Learning Vector Quantization

Xibin Zhu (Universität Bielefeld, DE)

Joint work of Zhu, Xibin; Schleif, Frank-Michael; Hammer, Barbara

Prototype-based learning algorithms represent given data in a (sparse) way by means of prototypes, they form decisions based on the similarity of data to prototypes, and training is very intuitive based on Hebbian principles. In addition, prototype-base models have excellent generalization ability, and prototypes offer a compact representation of data which can be beneficial for life-long learning.

Unsupervised prototype-based learning such as k-means, topographic mapping, neural gas, or self-organizing map infer prototypes based on input data only.

Supervised techniques incorporate additionally class labels and try to form class boundaries describing priorly known class labels. One of the most popular techniques in this context is learning vector quantization (LVQ), and extensions thereof which are derived from explicit cost function, generalized LVQ (GLVQ), or statistical models, robust soft LVQ (RSLVQ). These learning algorithms, however, are restricted to Euclidean vectors. Thus they are unsuitable for complex or heterogeneous data sets where input dimensions have different relevance or a high dimensionality yields to accumulated noise which can disrupt the classifications. Although this problem can be partially avoided by appropriate metric learning or by kernel variants, however, if data are inherently non-Euclidean, these techniques can not be applied. In addition, in modern applications, data are often addressed using dedicated non-Euclidean dissimilarities such as dynamic time warping for time series, alignments for symbolic strings, the compression distance to compare sequences based on an information theoretic ground, and similar. These settings do not allow an Euclidean representation of data at all, rather data are given implicitly in terms of pairwise dissimilarities or relations.

In the contribution, we propose extensions of GLVQ and RSLVQ, which can directly deal with relational data sets which are characterized in terms of a symmetric dissimilarity matrix only. The optimization can take place using gradient techniques. We test these techniques on several biomedical benchmark data sets, and the results are comparable to SVM while providing prototype based presentation.

3.27 Agents Learning a Complex Task/ Dispersive PSO

Jort van Mourik (Aston University – Birmingham, GB)

Joint work of van Mourik, Jort; Goldingay, Harry
Main reference H. Goldingay, J. van Mourik, "The effect of load on agent-based algorithms for distributed task allocation," Information Sciences, Elsevier.
URL http://dx.doi.org/10.1016/j.ins.2011.06.011

We present an overview of recent developments concerning various algorithms based on autonomous agents learning a complex optimisation task. We show that relatively simple algorithms based on autonomous agents can be competitive with the best centralised optimisation algorithms for which communication costs may soon become prohibitive for large systems. We propose a hybrid algorithm that combines the best features of both threshold- and market- based algorithms, and by introduction of a simple form of memory we obtain 98.5% of the theoretical efficiency limit of the optimal centralised algorithm, while keeping excellent scalability. As various algorithms are compared, each of which has a set of parameters that need tuning for fair comparison, the need for a good general purpose optimisation method arises. We propose a version of particle swarm optimisation (PSO) that avoids early convergence: Dispersive PSO. We show that this form of PSO generally outperforms existing ones in cases where early convergence (to a local optimum) is an issue.

3.28 Machine Learning for Data Visualization

Laurens van der Maaten (TU Delft, NL)

The talk gives an (incomplete) overview of machine-learning techniques that can be used for the visualization of high-dimensional data. In particular, it focuses on two types of dimensionality reduction techniques: (1) generative models that reduce dimensionality by performing maximum-likelihood learning in a (non)linear Gaussian model and (2) manifold learners that reduce dimensionality by preserving local properties of the data manifold. The last part of the talk focuses on some problems that arise specifically in visualization settings; for instance, with the question of how to visualize non-metric similarities or contingency tables.

4 Preliminary follow up publications resulting from the seminar

4.1 Supervised learning of short and high-dimensional temporal sequences for life science measurements

Frank-Michael Schleif (Universität Bielefeld, DE)

Joint work of F.-M.Schleif, A. Gisbrecht; B. Hammer
Main reference F.-M. Schlief, A. Gisbrecht, B. Hammer, "Supervised learning of short and high-dimensional temporal sequences for life science measurements," Dagstuhl Preprint Archive, arXiv:1110.2416v1 [cs.LG].
URL http://arxiv.org/abs/1110.2416v1

The analysis of physiological processes over time are often given by spectrometric or gene expression profiles over time with only few time points but a large number of measured variables. The analysis of such temporal sequences is challenging and only few methods have been proposed. The information can be encoded time independent, by means of classical expression differences for a single time point or in expression profiles over time. Available methods are limited to unsupervised and semi-supervised settings. The predictive variables can be identified only by means of wrapper or post-processing techniques. This is complicated due to the small number of samples for such studies. Here, we present a supervised learning approach, termed Supervised Topographic Mapping Through Time (SGTM-TT). It learns a supervised mapping of the temporal sequences onto a low dimensional grid. We utilize a hidden markov model (HMM) to account for the time domain and relevance learning to identify the relevant feature dimensions most predictive over time. The learned mapping can be used to visualize the temporal sequences and to predict the class of a new sequence. The relevance learning permits the identification of discriminating masses or gen expressions and prunes dimensions which are unnecessary for the classification task or encode mainly noise. In this way we obtain a very efficient learning system for temporal sequences. The results indicate that using simultaneous supervised learning and metric adaptation significantly improves the prediction accuracy for synthetically and real life data in comparison to the standard techniques. The discriminating features, identified by relevance learning, compare favorably with the results of alternative methods. Our method permits the visualization of the data on a low dimensional grid, highlighting the observed temporal structure.

4.2 PAC learnability versus VC dimension: a footnote to a basic result of statistical learning

Vladimir Pestov (University of Ottawa, CA)

Main reference V. Pestov, "PAC learnability versus VC dimension: a footnote to a basic result of statistical learning," arXiv:1104.2097v1 [cs.LG].
URL http://arxiv.org/abs/1104.2097v1

A fundamental result of statistical learning theory states that a concept class is PAC learnable if and only if it is a uniform Glivenko-Cantelli class if and only if the VC dimension of the class is finite. However, the theorem is only valid under special assumptions of measurability of the class, in which case the PAC learnability even becomes consistent. Otherwise, there is a classical example, constructed under the Continuum Hypothesis by Dudley and Durst and further adapted by Blumer, Ehrenfeucht, Haussler, and Warmuth, of a concept class of VC dimension one which is neither uniform Glivenko-Cantelli nor consistently PAC learnable.

We show that, rather surprisingly, under an additional set-theoretic hypothesis which is much milder than the Continuum Hypothesis (Martin's Axiom), PAC learnability is equivalent to finite VC dimension for every concept class.

Comments: Revised submission to IJCNN 2011.

4.3 How to Evaluate Dimensionality Reduction?

Bassam Mokbel

Joint work of Wouter Lueks, Michael Biehl, Barbara Hammer
Main reference W. Lueks, B. Mokbel, M. Biehl, B. Hammer, "How to Evaluate Dimensionality Reduction? – Improving the Co-ranking Matrix," Dagstuhl Preprint Archive, arXiv:1110.3917v1 [cs.LG].
URL http://arxiv.org/abs/1110.3917v1

The growing number of dimensionality reduction methods available for data visualization has recently inspired the development of quality assessment measures, in order to evaluate the resulting low-dimensional representation independently from a methods' inherent criteria. Several (existing) quality measures can be (re)formulated based on the so-called co-ranking matrix, which subsumes all rank errors (i.e. differences between the ranking of distances from every point to all others, comparing the low-dimensional representation to the original data). The measures are often based on the partitioning of the co-ranking matrix into 4 submatrices, divided at the K-th row and column, calculating a weighted combination of the sums of each submatrix. Hence, the evaluation process typically involves plotting a graph over several (or even all possible) settings of the parameter K. Considering simple artificial examples, we argue that this parameter controls two notions at once, that need not necessarily be combined, and that the rectangular shape of submatrices is disadvantageous for an intuitive interpretation of the parameter. We debate that quality measures, as general and flexible evaluation tools, should have parameters with a direct and intuitive interpretation as to which specific error types are tolerated or penalized. Therefore, we propose to replace K with two parameters to control these notions separately, and introduce a differently shaped weighting on the co-ranking matrix. The two new parameters can then directly be interpreted as a threshold up to which rank errors are tolerated, and a threshold up to which the rank-distances are significant for the evaluation. Moreover, we propose a color representation of local quality to visually support the evaluation process for a given mapping, where every point in the mapping is colored according to its local contribution to the overall quality.

4.4 About Generalization of the Conn-Index for Fuzzy Clustering Validation

Thomas Villmann (Computational Intelligence Group, University of Applied Sciences Mittweida, DE)

Joint work of T. Geweniger, M.ästner, M. Lange, and T. Villmann
Main reference T. Geweniger, M. Kästner, M. Lange, and T. Villmann, "Derivation of a Generalized Conn-Index for Fuzzy Clustering Validation," Machine Learning Reports, Report 07/2011.
URL http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_07_2011

Clustering and cluster validation strongly depends on the underlying model, the used dissimilarity measure, complexity constraints and other limiting options. In context of very high dimensional data and large data sets preprocessing and compression of the data by vector

quantization is one possibility to deal with this complexity. Subsequent clustering of the compressed data should take into account this information as well as cluster validation. One approach in this direction and presented at the seminar is the so-called Conn-Index invented by *Erzsébet Merényi* and *Kadim Taşdemir* [11]. In this approach, for the evaluation of the clusters consisting of vector quantization prototypes, the topological structure information between the prototype vectors acquired during the vector quantization learning is used to assess the quality of the cluster solution. This information is available by the Delaunay-graph with respect to the Voronoi tessellation of the data space according to the prototypes.

We discussed in a participant group (E. Merényi, T. Geweniger, M. Kästner, K. Taşdemir, and T. Villmann), how an extension of this approach could be designed such that fuzzy clustering also would be covered. Fuzzy vector quantization is mainly influenced by the fuzzy c-means algorithm for probabilistic fuzzy assignments [1, 5], its probabilistic counterpart [7, 8], and variants integrating neighborhood cooperativeness for better stability and convergence [3, 2, 4, 9, 10, 12, 13]. In consequence of these discussion we agreed that the topological structure between the prototypes in fuzzy vector quantization is implicitly contained in the fuzzy assignments, and, therefore, could be used to extend the Conn-Index for those vector quantization models.

As a preliminary result we can present a first publication, where we stated these thoughts more precisely formulating the underlying theoretical concepts of the new fuzzy Conn-index, see [6].

References

1 J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.

2 J. C. Bezdek and N. R. Pal. A note on self-organizing semantic maps. *IEEE Transactions on Neural Networks*, 6(5):1029–1036, 1995.

3 J. C. Bezdek and N. R. Pal. Two soft relatives of learning vector quantization. *Neural Networks*, 8(5):729–743, 1995.

4 J. C. Bezdek, E. C. K. Tsao, and N. R. Pal. Fuzzy Kohonen clustering networks. In *Proc. IEEE International Conference on Fuzzy Systems*, pages 1035–1043, Piscataway, NJ, 1992. IEEE Service Center.

5 J. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

6 T. Geweniger, M. Kästner, M. Lange, and T. Villmann. Derivation of a generalized Conn-index for fuzzy clustering validation. *Machine Learning Reports*, 5(MLR-07-2011):1–12, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_07_2011.pdf.

7 R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(4):98–110, 1993.

8 N. Pal, K. Pal, J. Keller, and J. Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 13(4):517–530, 2005.

9 N. R. Pal, J. C. Bezdek, and E. C. K. Tsao. Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks*, 4(4):549–557, 1993.

10 N. R. Pal, J. C. Bezdek, and E. C. K. Tsao. Errata to Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks*, 6(2):521–521, March 1995.

11 K. Taşdemir and E. Merényi. A validity index for prototype-based clustering of data sets with complex cluster structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(4):1039 – –1053, 2011.

12 E. Tsao, J. Bezdek, and N. Pal. Fuzzy Kohonen clustering networks. *Pattern Recognition*, 27(5):757–764, 1994.
13 T. Villmann, T. Geweniger, M. Kästner, and M. Lange. Theory of fuzzy neural gas for unsupervised vector quantization. *Machine Learning Reports*, 5(MLR-06-2011):27–46, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_06_2011.pdf.

Participants

- Gyan Bhanot
Rutgers Univ. – Piscataway, US
- Michael Biehl
University of Groningen, NL
- Kerstin Bunte
University of Groningen, NL
- Gert-Jan de Vries
Philips Research Lab. – Eindhoven, NL
- Klaus Dohmen
Hochschule Mittweida, DE
- Richard Farkas
Universität Stuttgart, DE
- Tina Geweniger
University of Groningen, NL
- Andrej Gisbrecht
Universität Bielefeld, DE
- Sven Haase
Hochschule Mittweida, DE
- Barbara Hammer
Universität Bielefeld, DE
- Marika Kästner
Hochschule Mittweida, DE
- Arto Klami
Aalto University, FI
- Neil D. Lawrence
Sheffield University, GB
- John A. Lee
University of Louvain, BE
- Marco Loog
TU Delft, NL
- Thomas Martinetz
Universität Lübeck, DE
- Erzsébet Merényi
Rice University, US
- Bassam Mokbel
Universität Bielefeld, DE
- Ernest Mwebaze
Makerere Univ. – Kampala, SAF
- Oliver Obst
CSIRO ICT Centre – Marsfield, AU
- Vladimir Pestov
University of Ottawa, CA
- John Quinn
Makerere Univ. – Kampala, SAF
- Fabrice Rossi
Télécom Paris Tech, FR
- Frank-Michael Schleif
Universität Bielefeld, DE
- Petra Schneider
University of Birmingham, GB
- Udo Seiffert
Fraunhofer IFF Magdeburg, DE
- Diego Sona
Fondazione Bruno Kessler – Trento, IT
- Marc Strickert
Universität Siegen, DE
- Kadim Tasdemir
EC Joint Research Centre – Ispra, IT
- David M. J. Tax
TU Delft, NL
- Peter Tino
University of Birmingham, GB
- Laurens van der Maaten
TU Delft, NL
- Jort van Mourik
Aston Univ. – Birmingham, GB
- Marina Vanucci
Rice University – Houston, US
- Michel Verleysen
UC Louvain-la-Neuve, BE
- Thomas Villmann
Hochschule Mittweida, DE
- Kilian Weinberger
Washington University, US
- Xibin Zhu
Universität Bielefeld, DE
- Dietlind Zühlke
Fraunhofer Institut FIT – St. Augustin, DE

Report from Dagstuhl Seminar 11351

Computer Science & Problem Solving: New Foundations

Edited by

Iris van Rooij[1], Yll Haxhimusa[2], Zygmunt Pizlo[3], and Georg Gottlob[4]

1 **Radboud University Nijmegen, NL,** i.vanrooij@donders.ru.nl
2 **Vienna University of Technology, AT,** yll@prip.tuwien.ac.at
3 **Purdue University – West Lafayette, US,** pizlo@psych.purdue.edu
4 **University of Oxford, GB,** Georg.Gottlob@comlab.ox.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 11351 "Computer Science & Problem Solving: New Foundations". This seminar was the first Dagstuhl seminar that brought together a balanced group of computer scientists and psychologists to exchange perspectives on problem solving. In the 1950s the seminal work of Allen Newell and Herbert Simon laid the theoretical foundations for problem solving research as we know it today, but the field had since become disconnected from contemporary computer science. The aim of this seminar was to promote theoretical progress in problem solving research by renewing the connection between psychology and computer science in this area.

Seminar 29. July–02. August, 2011 – www.dagstuhl.de/11351
1998 ACM Subject Classification F.1.1 Models of Computation, F.1.3 Complexity Measures and Classes, F.2.0 [Analysis of Algorithms and problem complexity] General, I.2.4 Knowledge Representation Formalisms and Methods, I.2.8 Problem Solving, Control Methods, and Search.
Keywords and phrases Problem solving, Cognitive psychology, Cognitive systems, Vision Representations, Computational complexity
Digital Object Identifier 10.4230/DagRep.1.8.96

1 Executive Summary

Iris van Rooij
Yll Haxhimusa
Zygmunt Pizlo
Georg Gottlob

This Dagstuhl seminar brought together a group of computer scientists and psychologists to discuss their perspectives on problem solving. The seminar was inspired by two previous Problem Solving workshops in 2005 and 2008 at Purdue University, USA. These workshops, organized primarily by psychologists, laid bare some fundamental theoretical questions in problem solving research. The organizers believed that research on these questions could benefit from more involvement of computer scientists in the area of problem solving. This motivated the organization on this seminar, which aimed to bring together computer scientists and psychologists to help build new formal foundations for problem solving research.

Computer Science & Problem Solving: New Foundations, *Dagstuhl Reports*, Vol. 1, Issue 8, pp. 96–124
Editors: Iris van Rooij, Yll Haxhimusa, Zygmunt Pizlo, and Georg Gottlob
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Of the 36 participants at the seminar about half were computer scientists and the other half were psychologists, though many identified as interdisciplinary researchers (e.g., cognitive scientists). To facilitate cross-disciplinary perspectives, computer science and psychology talks were alternated in the program of the seminar. There were 7 longer featured talks and 15 shorter talks.

On Day 1 of the seminar, Iris van Rooij opened the seminar by explaining its history and motivation. She discussed how computational complexity theory gives a formal framework for quantifying the difficulty of solving different types of search problems (e.g., the Traveling Salesman Problem and Minimum Spanning Tree), but that no analogous formal framework exists yet for quantifying the difficulty of solving so-called 'insight' problems (e.g, the Nine-dot problem), or more generally, quantifying the difficulty of representing a problem in the right way. The question of how one could develop such a formal framework was an overarching theme of the seminar. All participants were invited to think about this question. The topic resurfaced in several talks and workgroup discussions.

The featured talks on Day 1 were by Todd Wareham and Bill Batchelder. Wareham presented novel ideas on how a formal theory of 'insight difficulty' may take shape. His analysis was based on existing ideas in the psychology literature, such as the Representational Change Theory of Knoblich and et al., and his formalisms were inspired by Gentner's Structure-Mapping Theory. Batchelder presented a list of 19 classic examples of insight problems (these problems can be found in Appendix 7.1. These problems served as illustrations of problems that are not 'search problems' in the sense of Newell and Simon, yet for which problem solving researchers should nevertheless like to be able to model and explain the processes involved. Four shorter talks on Day 1 were given by Georg Gottlob, Sarah Carruthers, Sashank Varma and Jakub Szymanik. Gottlob introduced conceptual tools from computational complexity theory, graph theory and probabilistic computation that could inspire new ways of thinking about problem solving. Carruthers presented novel experimental data on how humans solve the graph problem Vertex Cover (see Appendix 7.3 for a definition). Varma presented a methodology for modeling the resource requirements of different brain areas invoked during problem solving. Szymanik presented a generalization of the Muddy Children Problem and explained how its solution can be modeled using logic.

On Day 2 of the seminar there were 3 featured talks. In the first featured talk, Niels Taatgen presented the ACT-R modeling architecture and illustrated how it could model the development of more general problem solving skills as a re-combination of more basic skills. Rina Dechter and Ken Forbus each gave a different AI perspective on problem solving in their featured talks. Dechter presented several sophisticated algorithmic techniques for solving NP-hard problems, such as Constraint Satisfaction and Bayesian Inference. Forbus proposed to consider 'analogy' as a new foundation for problem solving research and illustrated his perspective using the Companion framework. There were 3 short talks on this day. The two short talks by Johan Kwisthout and Marco Ragni (like Wareham's talk on Day 1) touched clearly on the theme of the seminar. Kwisthout proposed a formal framework for capturing the notion of 'relevance' when it comes to finding a suitable problem representation, and Ragni proposed a framework for quantifying the *a priori* difficulty of problem items on an IQ test based the notion of 'representational transformation'. In the third short talk, Jelle van Dijk gave a designers' perspective on problem solving. Van Dijk made the case that much real-world problem solving is probably best studied from an embodied embedded cognitive perspective. The official program for this day was closed with a working group discussion on meanings of common terms used throughout the talks (see Section 4.1).

Day 3 opened with a featured talk by Dedre Gentner. The talk by Gentner complemented

the talk by Forbus on Day 2 as she laid out the experimental evidence for the idea that analogical thinking (comparison and matching) lies at the foundation of human learning and reasoning. The featured talk was followed by two short talks, one by Liane Gabora and one by Daniel Reichman. Gabora presented a perspective on problem solving that is quite unlike the traditional view of problem solving as search through a well-defined space for a well-defined solution. Her perspective is that (creative) problem solving can perhaps best be seen as the recognition and actualization of a solution that before only existed in a state of potentiality. Reichman presented a theoretical computer science perspective on the well-known phenomenon of speed-accuracy tradeoffs in psychology. He proposed that algorithmic techniques from computer science can help predict what shape curves describing speed-accuracy tradeoffs will have in a variety of experimental conditions. In the afternoon of Day 3 there was no official program, and instead participants enjoyed the surroundings of Schloss Dagstuhl and/or went for a hike on one of the hills near the Schloss.

Day 4 started with the featured talk by Yun Chu, who gave an overview of the psychological research on insight problem solving (the talk had originally been scheduled for Day 1, but due to unforeseen circumstances Chu could not arrive at the seminar earlier). The talk by Chu helped build further common ground between the computer scientists and psychologists as it explained in more detail common paradigms and concepts used in psychological research on problem solving. The rest of the day consisted of two workshop sessions aimed at stimulating the formation of new interdisciplinary perspectives and collaborations (for details see Sections 4.2 and 4.3) and several short talks. Ute Schmid presented a framework for what Chomsky called a 'competence level model' of learning to problem solve, based on analytical inductive functional programming. Ulrike Stege gave a survey of typical computer science problems that are or could be used to investigate human problem solving strategies and pointed out some research challenges. Among them is the problem that researchers may think their participants are solving the problem that they posed, but the participants may in fact be solving an altogether different problem which the participants *think* the researchers have posed. Brendan Juba presented a new formal framework for heuristic rules based on PAC semantics. Nysret Musliu presented the concept of a (hyper)tree decomposition, a concept that can be utilized in algorithmic techniques for solving NP-hard problems efficiently. Jered Vroon presented a non-standard formalism in which problem solving is regarded as producing a solution rather than as a search through a search space. Last, Zyg Pizlo presented new algorithmic ideas for modeling human performance on the Traveling Salesman Problem based on the notion of multiresolution-multiscale pyramids. The day closed with a session in which participants brainstormed about novel interdisciplinary collaborations and open problems in the field. Some of these ideas were presented the same day, others were presented in the morning session of Day 5.

The morning of the last day of the seminar, Day 5, was reserved for short presentations of new collaborative ideas that the participants came up with, as well as the presentation of new ideas and open problems (see Sections 4.3 and 5 for details). The seminar closed with a wrap-up session in which participants reflected on the process and outcomes of the seminar (see Section 4.4 for a summary). To conclude, the seminar was successful in several ways: (1) It has resulted in a renewed awareness of how computer science and psychology can complement each other in the study of problem solving; (2) it has created a new impetus for more involvement of computer scientists in contemporary problem solving research; (3) it has created more common ground between computer science and psychologist in the study of problem solving; (4) it has produced several novel ideas on how to conceptualize 'problem solving' and, in particular, 'problem solving by insight;' (5) it has produced several

novel ideas on how to formalize these new conceptualizations; (6) it has produced concrete suggestions for new experimental paradigms for studying problem solving in the lab; (7) it has inspired new cross-disciplinary collaborative research projects; and last but not least (8) it has provided the groundwork on which follow-up Dagstuhl seminars can build in the future. With this seminar, the organizers hope to have contributed to an increased and sustained collaborative research effort between computer science and psychology in the domain of problem solving.

2 Table of Contents

3 Overview of Talks

3.1 Some Issues in Developing a Theory of Human Problem Representation

William H. Batchelder (University of California, US)

First I will describe some of my personal experiences interacting with computer scientists concerning the games of Chess and Go. There once was a strong belief that human knowledge could aid computational approaches to such games as these; however, approaches based on that notion ultimately failed, and instead brute force appears to have won the day. Next I will discuss some of the barriers that I see as standing in the way of developing a satisfactory formal theory of human problem solving. They include: (1) The lack of general experimental paradigms to study problem solving as found, for example, in other areas of cognitive psychology such as human memory, human attention, or visual perception. (2) The lack of a rich behavioral base that accompanies the act of solving a problem. (3) The lack of a formal theory of how problem solvers initially represent and possibly re-represent problems during solution efforts. Finally, in the last half of the talk, I will discuss variations on a set of twelve or so problems drawn from the folklore of brain teasers. I selected these problems because they are not move problems in the sense of Newell and Simon, but instead they require creative problem representations. For each of these problems, once a good representation is achieved, the solution follows pretty easily. I will organize the problems and variations on them around aspects human cognition that must be utilized or overcome to achieve a productive problem representation. These aspects of cognition include the nature of the human senses, imagery, memory, cognitive biases, and reasoning processes. In particular, I will use the problems to draw out issues that may need to be handled in constructing a formal theory of human problem representation. I will file a list of the brain teasers for your possible interest the week before the workshop starts.

3.2 Vertex Cover and Human Problem Solving

Sarah Carruthers (Univ. of Victoria, CA)

In this seminar, we will look at preliminary results from a study of human solutions to Vertex Cover problems. The purpose of this pilot study is to identify: select strategies employed by participants, and features of instances which may impact performance. We will also discuss what measures of performance are of interest, as well as related problems which may be of interest.

3.3 Human Performance on Insight Problem Solving: A Review

Yun Chu (Univ. of Hawaii at Manoa, US)

A review of recent research on insight problem-solving performance. We discuss what insight problems are, the different types of classic and newer insight problems, and how we can classify them. We also explain some of the other aspects that affect insight performance, such as hints, analogs, training, thinking aloud, and individual differences. In addition, we describe some of the main theoretical explanations that have been offered. Finally, we present some measures of insight and relevant neuroscience contributions to the area over the last decade.

3.4 Advanced Reasoning in Graphical models

Rina Dechter (Univ. California – Irvine, US)

Graphical models, including constraint networks, belief networks, Markov random fields and influence diagrams, are knowledge representation schemes that capture independencies in the knowledge base and support efficient, graph-based algorithms for a variety of reasoning tasks, including scheduling, planning, diagnosis and situation assessment, design, and hardware and software verification. Algorithms for processing graphical models are of two primary types: inference-based and search-based. Inference-based algorithms (e.g., variable-elimination, join-tree clustering) are time and space exponentially bounded by the tree-width of the problem's graph. Search-based algorithms can be executed in linear space and often outperform their worst-case predictions. The thrust of advanced schemes is in combining inference and search yielding a spectrum of memory-sensitive algorithms universally applicable across graphical models. The talk will provide an overview of principles of reasoning with graphical models developed in the last decade in constraints and probabilistic reasoning.

3.5 Analogy as a Computational Foundation for Problem-solving and Learning

Kenneth D. Forbus (Northwestern University – Evanston, US)

Classical approaches to problem-solving focus on first-principles reasoning (e.g. logic), using quantified representations and chaining. Analogy and similarity, to the extent they are considered at all, are viewed as rare operations which can safely be ignored, at least to first order. This talk, which describes joint work with Dedre Gentner, argues that the opposite is true: That analogy and similarity should be viewed as primary reasoning operations, with logic and chaining being used in support of them. We start by using a series of examples (including visual problem solving, physics problem solving, counterterrorism, and moral

decision-making) to introduce the primitive operations of analogical processing: Matching, retrieval, and generalization. The importance of qualitative representations, which facilitate analogical reasoning and learning, will be outlined. Finally, we describe some of the new issues that this framework raises for computational models of problem solving, including experience, learning encoding, rerepresentation, and skolem resolution.

3.6 Problem Solving as the Recognition and Actualization of Potentiality

Liane Gabora (University of British Columbia – Vancouver, CA)

Theories of creativity, such as Finke, Ward and Smith's Geneplore model and Simonton's Darwinian model, generally assume that creative problem solving involves searching through memory, selecting well-defined candidate ideas, and then tweaking them in response to problem constraints. I would like to discuss an alternative: that creative individuals wrestle with issues or ideas that are, for them, not well-defined, or in a state of potentiality. Over time, as these ideas are considered from different perspectives, they come to assume a form that is more fully actualized, or well-defined. This suggests that (in accordance with Einstein's assertion that finding the right question is a more significant step than answering it) the challenge is to construct one's mental model of reality in such a careful and precise way that one is led directly to the frontiers of what is known, and thus able to recognize and engage in informed speculation about what lies beyond these frontiers, i.e. what is currently in a state of potentiality for everyone. Problem solving is thus viewed as redefining the unknown in terms of what is known. However, the 'unknown' can take the form of not just gaps in knowledge but experiences that are so traumatic or unusual that we have not fully come to terms with them. This leads to a related question that I would like to explore: can artistic endeavors be understood within a problem-solving framework? I suggest that artistic tasks are those for which the topology of the fitness landscape is determined not by objective, agreed-upon aspects of the world, but by personal experiences and the emotions surrounding them. The problem-solving task for the artist is to translate information patterns underlying the dynamics of, for example, neurotransmitter release that rendered the particular emotional impact of an experience or situation into the constraints of the artistic form; for example, the tragic experience of losing a family member might be translated into the constraints of music. In so doing, one re-frames the experience in terms of what one has experienced before, and thus incorporates it into the fabric of one's understanding of the world, and comes to terms with it.

3.7 The Analogical Mind

Dedre Gentner (Northwestern University – Evanston, US)

Analogical processes are central in human learning and reasoning. Analogical comparison engages a process of structural alignment and mapping that fosters learning and reasoning in

at least three distinct ways: it highlights common relational systems; it promotes inferences; and t calls attention to potentially important differences between situations. It can also lead to re-representing the situations in ways that reveal new facets.

An important outcome of analogical comparison is that the common relational structure becomes more salient and more available for transfer in short, a portable abstraction is formed. Thus, structure-mapping processes bootstrap much of human learning.

The power of analogy is amplified by language learning. Hearing a common label invites comparison between the referents, and this structure-mapping process yields insight into the meaning of the term. The mutual facilitation of analogical processing and relational language contributes to the power and flexibility of human learning.

Finally, although analogy is sometimes thought of as a clever, somewhat effortful process, in fact it is pervasive in human processing, In this talk, I present psychological studies showing the role of analogy processes in human learning and reasoning. I will try to convey not only the power and importance of analogical processes but also their ubiquity in human cognition.

3.8 Living with Computational Complexity

Georg Gottlob (University of Oxford, GB)

Many computational problems that are important in real life are intractable. There are different ways of coping with intractability. This talk will focus, in particular, on methods of recognizing large 'islands of tractability' for NP-hard problem, i.e., large tractable subclasses. We will illustrate the use of graph-theoretic concepts such as tree-width and hyper-tree width in order to obtain large polynomial classes of intractable problems. In addition, we will mention a number of other ways of coping with complexity and illustrate how both computer programs and fruit flies (rather than bees) can solve certain 'complex' problems. The talk will end with some thoughts of more philosophical nature about computational complexity.

3.9 PAC Semantics: A Framework for Heuristic Rules

Brendan Juba (MIT – Cambridge, US)

Valiant's PAC Semantics [1] provides a clean standard that captures the utility of 'rules of thumb' about a given domain that may, for example, be derived from a sample of typical experiences in the domain; we suggest that it may be useful for the analysis of the acquisition and use of heuristic rules. In support of this suggestion, we show that PAC Semantics also features some natural tractable cases for inference. We describe a simple and efficient algorithm that tests the validity of candidate assertions given access to 'partially obscured' samples from the domain, correctly classifying all assertions except for good rules of thumb that cannot be established by a 'simple proof' using typical obscured examples.

References

1 Leslie G. Valiant. Robust logics. *Artificial Intelligence*, 117:231–253, 2000.

3.10 Relevant Representations

Johan Kwisthout (Radboud University Nijmegen, NL)

When computer scientists discuss the computational complexity of, e.g., finding the shortest path from A to B, their starting point typically is a formal description of the problem at hand, e.g., a graph with weights on every edge.

Given such a formal description, either tractability or intractability of the problem is established, by proving that the problem enjoys a polynomial time algorithm, respectively is NP-hard. However, this problem description is in fact an abstraction of the actual problem of being in A and desiring to go to B: it focuses on the relevant aspects of the problem (e.g., distances between landmarks and crossings) and leaves out a lot of irrelevant details.

This abstraction step is often overlooked, but may well contribute to the overall complexity of solving the problem at hand. For example, it appears that 'going from A to B' is easier to abstract: it is fairly clear that the distance between A and the next crossing is relevant, and that the color of the roof of B is typically not. However, when the problem to be solved is 'make X love me', where the current state is (assumed to be) 'X does not love me', it is hard to agree on all the relevant aspects of this problem.

In this talk, I will propose a framework for capturing the notion of relevance when it comes to finding a suitable problem representation, with the ultimate goal of formally separating 'hard to represent' and 'easy to represent' problem instances.

3.11 Algorithms for Computing (Hyper)tree Decompositions

Nysret Musliu (TU Wien, AT)

Constructing decompositions of small width is crucial to solve efficiently problems based on their (hyper)tree decomposition. In recent years, several methods have been proposed to generate good (hyper)tree decompositions. Such methods include exact methods that are used to find optimal decompositions for small problems, and (meta)heuristic algorithms that find (hyper)tree width upper bounds for larger problems. In this talk, we will first give a survey of existing techniques for constructing (hyper)tree decompositions and compare these algorithms on benchmark problems from the literature. Further, we will discuss the following open questions: (1) Can we find more efficient methods to compute upper bounds for (hyper)tree width? (2) Can the existing techniques be easily adapted to generate (hyper)tree decompositions of small width that fulfill other specific conditions?

3.12 Multiresolution-multiscale Pyramids and the Traveling Salesman Problem

Zygmunt Pizlo (Purdue University – West Lafayette, US)

After presenting representative results from experiments on how human subjects produce near-optimal tours, I will describe the main aspects of pyramid models of the human visual system. One of the two main operations in pyramid models of TSP is hierarchical clustering. The second operation is a top-down sequence of approximations of a TSP tour, where centers of clusters are used in lieu of cities. The tour is produced sequentially by moving the model's attention from one city to another. Decisions on finer representations are guided by coarse representations. The model stores in its memory minimal amount of information related to the currently analyzed part of the problem. When additional information is needed, the model 'looks' at the TSP problem again. The errors and memory requirements will be presented and discussed. I will conclude by conjecturing that such a pyramid algorithm is a plausible model of human problem solving, in general.

3.13 In Search of a Cognitive Complexity Measure for Matrix Reasoning Problems

Marco Ragni (Universität Freiburg, DE)

Reasoning difficulty for items in IQ-tests is generally determined empirically: The item difficulty is measured by the number of reasoners who are able to solve the problem. Although this method has proven successful (nearly all IQ-Tests are designed this way) it is desirable to have an inherent formal measure reflecting the reasoning complexity involved. This talk will present some geometrical analogy reasoning problems and based on the types of functions necessary to solve these problems, a difficulty measure is introduced. This is finally compared to the empirical difficulty ranking as determined by Cattell's Culture Fair Test, Evans Analogy problems, and an own experiment.

3.14 Speed-Accuracy Tradeoffs: A computational Perspective

Daniel Reichman (Weizmann Institute – Rehovot, Israel)

Speed-Accuracy questions are central in several subfields of cognitive psychology such as problem solving, decision making and perception. We address several algorithmic techniques (e.g., property testing, stochastic optimization) as well as hardness results in addressing how will speed-accuracy curves look like when handling chalenging problems in psychological contexts.

3.15 Learning Productive Rules from Problem Solving Experience

Ute Schmid (Universität Bamberg, DE)

Joint work of Schmid, Ute; Kitzelmann, Emanuel
Main reference Ute Schmid, Emanuel Kitzelmann, "Inductive Rule Learning on the Knowledge Level," in: Cognitive Systems Research 12 (2011), Nr. 3, pp. 237–248.
URL http://dx.doi.org/10.1016/j.cogsys.2010.12.002

One specific characteristic of human autonomous learning is that humans are able to extract productive rule sets from experience which often is a stream of only positive examples. Following Chomsky, a productive rule set allows a person to produce systematic behavior in situations of arbitrary complexity, for example being able to build towers of sorted blocks for an arbitrary number of blocks. Such productive rules represent the competence of a person – in contrast to a person's performance which is open to unsystematic variations and errors. Furthermore, productive rule sets often are verbalizable, that is, a person can explain a general solution procedure to another person. I propose to use an approach to analytical inductive functional programming to model this type of high-level learning. Analytical inductive programming provides algorithms with clearly defined restriction and preference biases for learning recursive rule sets from small sets of positive examples.

3.16 Human Problem Solving of (hard) Computational Problems-A Computer Scientist's Thoughts and Interests

Ulrike Stege (University of Victoria, CA)

We survey typical computer science problems that are or could be used to investigate human problem solving strategies, such as the Traveling Salesperson problem and the Minimum Spanning Tree problem, as well as other graph problems. We discuss research questions and approaches that are investigated, highlight possible difficulties with current approaches and pose a set of questions that we believe are realistic to investigate in the near future.

3.17 Generalizing Muddy Children Puzzle

Jakub Szymanik (University of Groningen, NL)

Joint work of Gierasimczuk, Nina, Szymanik, Jakub;
Main reference Nina Gierasimczuk, Jakub Szymanik, "Invariance Properties of Quantifiers and Multiagent Information Exchange," Proceedings of the 12th Meeting on Mathematics of Language, Lecture Notes in Artificial Intelligence 6878, M. Kanazawa et al. (Eds.), pp. 72–89, 2011.
URL http://dx.doi.org/10.1007/978-3-642-23211-4_5

We study a generalization of the Muddy Children puzzle by allowing public announcements with arbitrary generalized quantifiers [1, 2]. We propose a new concise logical modeling of the puzzle based on the number triangle representation of quantifiers. Our general aim is to discuss the possibility of epistemic modeling that is cut for specific informational dynamics.

Moreover, we show that the puzzle is solvable for any number of agents if and only if the quantifier in the announcement is positively active (satisfies a form of variety).
Slides can be found at

http://prezi.com/96_wd3mgx_d1/a-generalization-of-the-muddy-children-puzzle/.

References

1 Nina Gierasimczuk and Jakub Szymanik. A Note on a Generalization of the Muddy Children Puzzle. *Proceeding of the 13th Conference on Theoretical Aspects of Rationality and Knowledge, K. Apt (Ed.), ACM Digital Library*, pp. 257– 264, 201. http://dx.doi.org/10.1145/2000378.2000409

2 Nina Gierasimczuk and Jakub Szymanik. Invariance Properties of Quantifiers and Multiagent Information Exchange. *Proceedings of the 12th Meeting on Mathematics of Language, Lecture Notes in Artificial Intelligence 6878, M. Kanazawa et al. (Eds.)*, pp. 72–89, 2011. http://dx.doi.org/10.1007/978-3-642-23211-4_5

3.18 Human Problem Solving: The Search for the Right Toolkit

Niels A. Taatgen (University of Groningen, NL)

Cognitive architectures have two approaches to modeling problem solving. One approach is to consider problem solving as a fundamental property of the architecture, and assume we approach new and unknown problems with set of weak methods that is the same for any individual. The Soar architecture (Newell, 1990) is an example of this approach. Other architectures, like the ACT-R architecture (Anderson, 2007), assume no architectural mechanisms for problem solving it all, but assume that problem solving consists a set of cognitive skills that have to be learned. However, models within such architectures typically encode the problem-solving strategy necessary for the task at hand, and therefore contribute little to a general account of human problem solving.

In my talk I will present a modeling framework that can serve as a starting point for a general theory of how human problem-solving skills develop within an architecture with no architectural problem-solving strategies. The idea is that the model starts with the most basic skills that are possible within a rule-based architecture, which is making single comparisons and simple elementary actions. Guided by declarative knowledge and the process of production compilation, these elementary skills can be combined to more complex skills. I will demonstrate this idea with models of cognitive transfer, in which knowledge learned in one task is used in for another.

3.19 Spatial Problem Solving: The Optimal Deployment of Cortical Resources

Sashank Varma (University of Minnesota, US)

Newell and Simon's analysis of problem solving as search through problem spaces is foundational for artificial intelligence and cognitive psychology. Because problem spaces are typically large, cognitive agents must deploy their limited resources judiciously, through planning and heuristic reasoning. The current research extends the classical conception of problem solving to the level of brain function. The cortex is understood as a set of centers, each possessing a finite supply of computational resources. Problem states, heuristics, and goals are mapped to different centers. In this view, problem solving is the optimal deployment of limited cortical resources across a network of collaborating centers. This is formalized as a linear programming problem that the brain is hypothesized to solve on a moment-by-moment basis. The resulting model provides a good account of the solution times, error rates, and brain activation fluctuations of normal adults and patients with lesions as they solve spatial problems. The implications of this research for artificial intelligence, cognitive psychology, and cognitive neuroscience are discussed.

3.20 Problem Solving as Producing a Solution

Jered Vroon (Radboud University Nijmegen, NL)

Joint work of Vroon, Jered; van Rooij, Iris; Wareham, Todd; Haselager, Pim

A new formalism for describing the structure and (associated) hardware of a system will be introduced. Within this formalism, problem solving is regarded as producing a solution rather than as a search through search space.

This formalism might be more limited than approaches that regard problem solving as a search through search space as it seems to require that a solution-producing structure is already in place. Nonetheless, even within this formalism we can distinguish between systems that require more or less structures and (associated) hardware. In this talk I will discuss these considerations and their relevance to the bigger field of problem solving.

3.21 What Does (and Doesn't) Make Problem Solving by Insight Easy? A Complexity-Theoretic Investigation

H. Todd Wareham (Memorial Univ. of Newfoundland, CA)

Problem solving is a very important and commonly-invoked cognitive ability. There are several recognized types of problem solving, e.g., by analogy, by search, by insight, and each is successful to various degrees in particular situations. Several information-processing

theories have been proposed for these types. However, it is very difficult to use empirical studies to characterize the situations in which these types do and do not work, let alone link such situations to (and hence verify) the mechanisms proposed by these theories.

In this talk, we will describe an approach complementary to empirical studies which uses computational complexity analysis to assess the situations under which the mechanisms proposed by a theory can and cannot operate efficiently. Such assessments, in turn, suggest both predictions that can be verified by experiment as well as viable refinements to the theories. We will illustrate this methodology with an analysis of problem solving by insight as formulated under the Representation Change Theory of Knoblich et al.

3.22 The Way of the Ouroboros: How to Represent Problems by Solving Them

Jelle van Dijk (TU Eindhoven, NL)

This talk is based on observations and design efforts in support of the practice of 'creative problem solving' in groups. I will set the scene discussing the difference between the well-known map and the territory, and I ask whether research on human problem solving should ask how one navigates the former, or rather the latter, or even how one deals with both. Perhaps we have been concerning ourselves too much with the map. I then speculate on three alternative ways by means of which people deal with problems in everyday practice. These 'real-world tactics' may not always be in the central attention of problem solving research. They are: (1) The way of the Oyster: Not solving the problem, but encapsulating it (2) The way of the River: To let things implicitly flow into a solution; and (3) The way of the Ouroboros: Representing the problem by first executing the solution. This strange option I call the Ouroboros, i.e. creating a representation by executing a problem solution, seems to put the cart before the horse and therefore nonsense. I will nonetheless discuss a number of variations of this strategy that I think exist, and actually work, in everyday practical circumstances. In order to ground the idea I will relate it to some theoretical notions from Embodied and Situated Cognition theory, as well as to my empirical observations of creative group sessions (aka brainstorms). My question for the seminar is whether these ideas implies a completely new line of research, or whether it is possible to integrate these ideas into existing models and theories on problem solving (or whether they are really just nonsense).

4 Working Groups

4.1 Key terms and their meanings

During Day 1 it became clear that speakers used terms whose meanings were unclear for some or many participants in the audience. It was decided to keep track of these terms by listing them on the blackboard. All subsequent speakers were asked to define these terms whenever they used them in their talks. Over the course of the seminar, the list grew to include the following terms:

- optimization
- representation
- search
- heuristic
- problem
- insight
- chunk
- cognition
- relevant
- complexity
- embodied cognition
- model
- situated cognition
- knowledge
- distributed cognition
- communication

In a workgroup session on Day 2, small groups of 4 participants (mixed groups of 2 computer scientists and 2 psychologists) were invited to pick one word from the above list and discuss all of its possible meanings. Participants were explicitly instructed not to try to decide on one 'proper' or agreed upon meaning, but rather to generate as many possible different meanings as seemed relevant to the domain of problem solving. Interestingly, of the long list of words there were three words that were a recurrent topic of discussion. These were 'problem', 'representation', and 'model'. It became clear that the meanings of even these three central words differed both between and within computer science and psychology.

The exercise was intended to raise awareness of the different usages of words by different researchers in the area of problem solving. The reasoning of the organizers was as follows: For interdisciplinary collaborations to get off the ground researchers need to be able to speak each other's languages, negotiate meanings, and develop new terminology as the need arises. This exercise helped to foster such an open minded atmosphere.

4.2 Promoting interdisciplinary discussion

On Day 3 small groups of 2 computer scientists and 2 psychologists were formed to discuss the following questions.

1 Why study problem solving in an interdisciplinary manner? Where could one discipline help the other?

2 What do you need to understand each other? What do you need to know to know how to help the other?
3 Do you see insurmountable differences, obstacles, or challenges?
4 What is (human) problem solving? Why study it? What are important research questions?
5 Can we identify different classes of problem solving, and characterize their relative difficulty?

There was general agreement among the participants that an interdisciplinary approach to problem solving would be desirable and possible, and that seminars like this one are useful for building the necessary 'common ground'. Subsequently, the participants added several more questions to the list:

6 When is a problem solved?
7 What type of problems do we want to include in this area of study?
8 What computational methods may be used to investigate these research questions?
9 What experimental paradigm to use?
10 Is there to be one or multiple theories of problem solving?

Question 6 was motivated by the observation that in computer science a problem $f : X \to Y$ is said to be solved when an input $x \in X$ is translated to an output $f(x) \in Y$. Yet, solving the problems presented by the psychologist Batchelder sometimes meant something like 'understand the reason or motive for the behavior or situation in the scenario'. Can the latter notion also be mapped to the input-computation-output paradigm? Question 7 was raised because many different cognitive abilities could count as examples of problem solving, e.g., visual problem solving and common sense reasoning. Should we focus our research on some of these, or consider all of them? Question 8 was raised because a variety of computational methods could be adopted in problem solving research, such as models, architectures and algorithms. Question 9 was raised because fruitful research in psychology often depends on a stable experimental paradigm with interpretable dependent measures (such as accuracy and speed). Last, Question 10 was raised because some participants felt that problem solving theories may be as diverse as the different types of problems out there, whereas other participants were committed to building unified or integrative accounts of problem solving in general.

4.3 New Ideas and Collaborations

On Day 4 of the seminar a working group was scheduled in which participants were invited to think about and try to come up with new cross-disciplinary collaborative projects and/or identify important open problems in the field of problem solving. The ideas generated in this workgroup were presented either the same day or, when ideas needed to be first further developed, on the morning of Day 5. A large group of participants presented ideas inspired by the seminar and/or new collaborations. All participants have furthermore been invited to be submit their work presented at or inspired by the seminar for consideration for publication in *The Journal of Problem Solving* (see Section 6). In the Section 5, we present a selection of the ideas that were presented at the seminar that we judge to be particularly original or important, and of general interest.

4.4 Wrap-up session: Evaluation and outlook

Iris van Rooij chaired the wrap-up session, and considered the following questions.

- Did we get (closer to) a shared notion of 'problem solving' (and 'the study of problem solving')?
- Did we get (closer to) new foundations?
- Did we get (closer to) new ideas for formalizing notions such as 'representational complexity' (e.g., 're-representation', 'insight', 'ill-defined', etc.).

She argued that we could answer all these questions in the affirmative. The questions 'what is problem solving?' and 'what distinguishes problem solving from other cognitive domains?' was also a recurring topic of discussion at the two preceding Purdue workshops on Problem Solving in 2005 and 2008. It appears that these questions plague the domain of problem solving more than other cognitive domains (for reasons unknown, though speculations range from the idea that problem solving is not a unified category of cognitive processes, to the idea that it is but that we have too few good experimental paradigms for studying problem solving in the lab). Even though this seminar has not produced definite answers to these questions, there does seem to be a convergence of ideas on what defines the different types of key mental processes involved in problem solving. This consideration of subprocesses of problem solving has even motivated a new classification of areas of problem solving research that may inform and guide future research and theorizing in the field (see section 5.3). As for the hoped-for progress in the formal foundations of problem solving research, and the notion of 'representational complexity' in particular, novel ideas have also been put forth, for instance in the presentations by Wareham, Kwisthout, and Ragni, and the open problem proposed by Haxhimusa and van Rooij.

In the wrap-up session participants were also asked for their feedback about the seminar and for recommendations for a follow-up seminar. One idea that was raised was that a follow-up seminar could aim to prepare for a Handbook on Problem Solving, and that the first day of the program could include tutorials, e.g., about computer science for psychologists and about psychology for computer scientists. A brainstorm on topics for a follow-up seminar resulted in the following list: Problem solving in the real-world versus the lab; Cognitive architectures and problem solving; Problem solving in the large; Problem solving of dynamic problems; Social problem solving; Human-inspired machine problem solving; Spatial problem solving; Problem solving with bounded resources (Bounded Rationality).

5 New Ideas and Open Problems

In this section we present a selection of open problems and new ideas that were presented at the morning session on Day 5 of the seminar.

5.1 A 'Turing Test' for Problem Solving

Ken Forbus presented an analog of the Turing test for intelligence for the domain of problem solving. He coined this the *Dagstuhl-Batchelder test*, as it was specifically inspired by the computational challenges posed by the 19 problems presented by Bill Batchelder at this seminar. The idea behind this test is that a system could be said to engage in genuine problem solving if it could solve at least these 19 problems. Importantly, the test should be

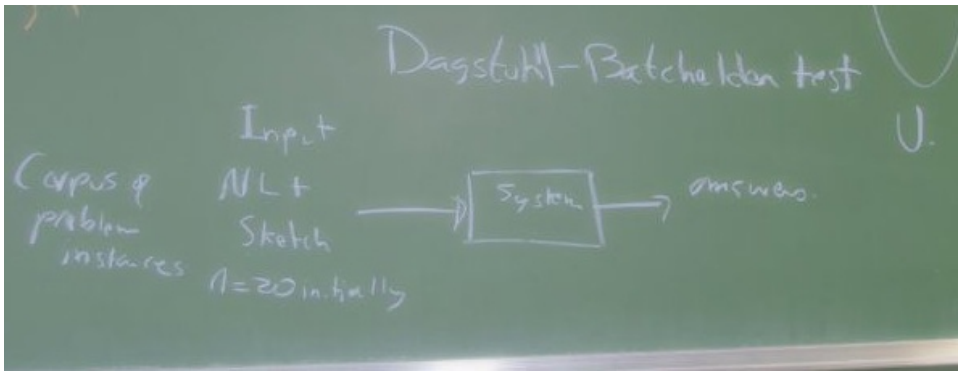


Figure 1 The Dagstuhl-Batchelder test for Problem Solving

fair and representative of how humans can solve the problems. Therefore Forbus imposed the constraint that the inputs to the system should be the raw text and images as presented in Appendix 7.1. In addition the system is allowed to have a knowledge data base, which could for instance consist in sketches of situations etc. Figure 1 illustrates the idea.

5.2 A Complexity Hierarchy of Insight Problems

Haxhimusa and van Rooij posed an open question: Is it possible to formulate a hierarchy of complexity classes for insight problems analogous to the computational complexity classes for search problems? They proposed that such a hierarchy may define classes in terms of the number of changes c to the input representation required to turn an insight problem (conceived as an ill-defined search problem) into a well-defined (potentially trivial) search problem. Here c may be thought of as the number of basic 'insights', 'pieces of information', 'hints' or 're-representation steps' needed to turn an insight problem into a well-defined problem (cf. Todd Wareham's proposal for a similar framework). In the proposed hierarchy $C_0 \subset C_1 \subset C_2 \subset ... \subset C_{n-1} \subset C_n$, the class C_0 would denote the class of well-defined problems, i.e., problems requiring no change in the input in order to become well-defined. Further, each class C_k in the hierarchy consists of those problems that are at most $c = k$ changes away from some problem in C_0. Interesting open questions are the following: Can this idea for a complexity hierarchy for insight problems be worked out to a formal framework? And if so, would it possible to use the framework to characterize, explain and/or predict the difficulty of different classes or types of insight problems for humans?

5.3 A Classification of Problem Solving Research(ers)

Todd Wareham proposed a classification of problems in terms of the nature of the subprocesses that are (assumed to be) invoked during the problem solving process (see Table 1). He observed a scale of problem classes ranging from search only, to restructuring combined with search, to problems that require access to and processing of world knowledge over and above the restructuring and search processes involved. Wareham observed that each class of problem seems to have at least one associated representative researcher, each of which was present at the seminar. This way of conceptualizing different classes of problems seems very

intuitive and may prove useful for the field to understand how different problems and models of problem solving differ and relate to each other.

Table 1 Classification of Problems.

Processes assumed to be involved in problem solving	Representative researcher in psychology	Representative researcher in computer science
search	Pizlo	Stege
restructuring + search	Chu	Wareham
restructuring + search + world knowledge	Batchelder	Forbus

5.4 Verbal Reports Revisited

Frank Jäkel proposed that, lacking a firm theoretical foundation to date, research on problem solving by *insight* may do well to revisit the methodology that lay the foundations for the theory of problem solving by *search* developed by Newell and Simon, viz., verbal reports made by humans about what they are thinking while they are engaged in problem solving. Jäkel pointed out that this methodology has fallen out of favor in psychology because it is a form of introspection and therefore considered unreliable for understanding the nature of cognitive processes. As a consequence, the methodology has been replaced by methodologies using simpler behavioral measures such as speed and accuracy of problem solving. Jäkel makes an important observation however. Even though verbal protocols are based on a form of introspection and may be to some extent unreliable, they are very rich in information that can be useful for hypothesis generation and theory formation. For instance, revisiting the 1972 book by Newell and Simon on problem solving reveals that many of their hypotheses about 'means-end analysis' and 'heuristic search' were a direct consequence of the meticulous analysis of verbal reports of people solving search problems. It is also noteworthy that the methodology for verbal reports has been refined considerably since the heyday of introspection in early psychology without the mainstream of cognitive psychology really taking notice of these developments [1]. In addition, verbalizing and inner speech are an important part of problem solving anyway. Hence, even if verbal reports do not constitute a rigorous *test* of theories of problem solving, they may prove useful for *coming up* with new theories of insight problem solving, which later can be tested using other measures.

References

1 K.A. Ericsson and H.A. Simon. Verbal Reports as Data. *Psychological Review*, 87(3):215–251, 1980.

6 Dissemination of Results

All participants have been invited to submit their research presented at this seminar or inspired by this seminar for consideration for publication in *The Journal of Problem Solving* (JPS). JPS is an open access journal with an interdisciplinary readership. We plan to have two special issues: one in the Spring and the other in the Fall of 2012. Considering the fact that papers in JPS can be accessed by everyone (no subscription is required), the proceedings from this workshop are expected to be read widely and have large impact. Once the special

issues are published, Purdue University Press (the publisher) will print a book with the published papers.

JPS (ISSN 1932-6246) is a multidisciplinary journal that publishes empirical and theoretical papers on mental mechanisms involved in problem solving. The journal welcomes original and rigorous research in all areas of human problem solving, with special interest in solving difficult problems (e.g., problems in which human beings outperform artificial systems). Examples of topics include (but are not limited to) optimization and combinatorial problems, mathematics and physics problems, theorem proving, games and puzzles, knowledge discovery problems, insight problems and problems arising in applied settings. JPS encourages submissions from psychology, computer science, mathematics, operations research and neuroscience. More information on journal web site: http://docs.lib.purdue.edu/jps/

Editor-in-Chief: Zygmunt Pizlo, Department of Psychological Sciences, Purdue University.

7 Appendices

7.1 19 Classic Insight Problems, by Bill Batchelder

Classic insight problems presented by Bill Batchelder are shown in Figures 2, 3, 4. In these figures Batchelder has adapted several famous problems from the folklore of brain teasers. Batchelder selected these problems because they are not move problems in the sense of Newell and Simon, and the main barrier to solution is finding a productive representation. Batchelder acknowledges the original author of some of these problems, even though they are not shown in the figures.

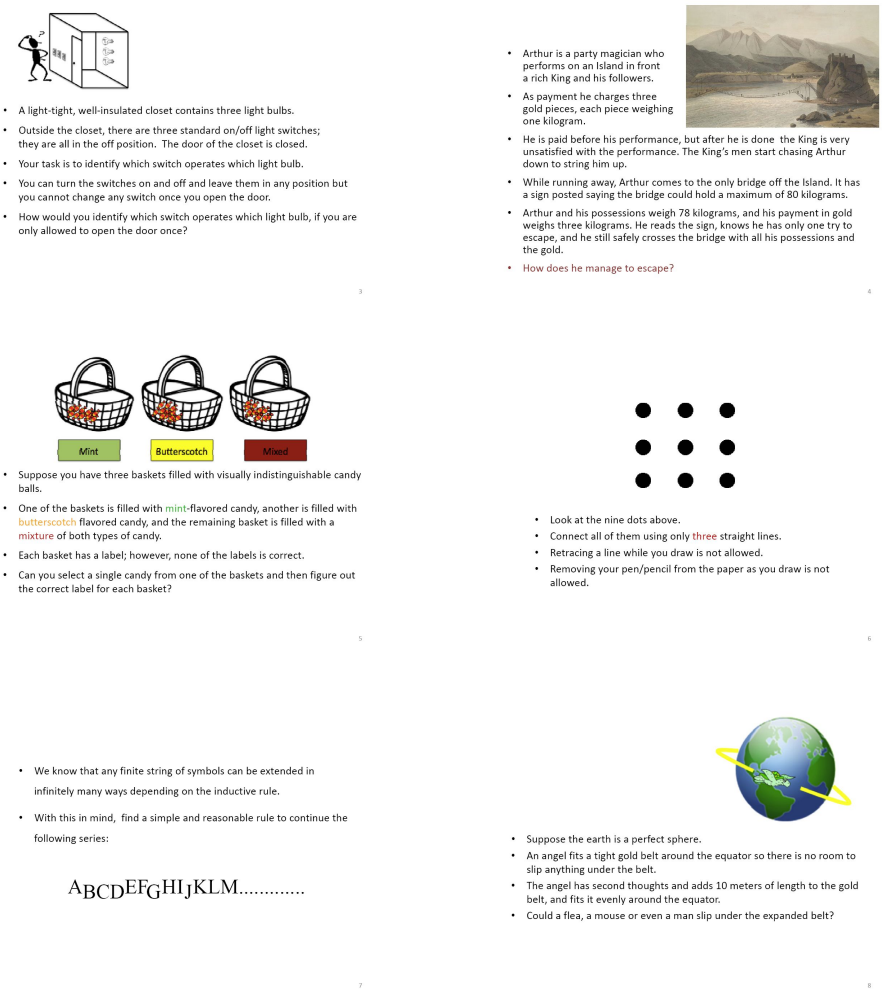
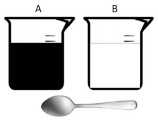


Figure 2 19 insight problems.



- You have two quart-size beakers; beaker A has a pint of coffee and beaker B has a pint of cream in it.
- First, you take a tablespoon of coffee from A and poor it into B. Mix thoroughly.
- Then you take a tablespoon of the mixture in B and poor it into A. Mix thoroughly.
- Which beaker, if any, has less diluted content after the two transfers? The coffee in A or the cream in B? (Forget issues about the chemistry of miscibility)

9

Can you find any interesting sense of the following paradoxical statement?

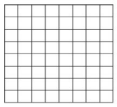
It is impossible to draw a perfect map of England while standing in a London flat, but it might be possible to do it in a New York City pad.

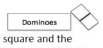
11

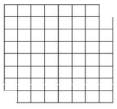
- During a recent census, a man told the census taker that he had 3 children.
- When asked for their ages, he replied, "The product of their ages is 36."
- The census taker said, "I need to know each of their ages."
- The man said, "well The sum of their ages is the same as my house number."
- The census taker looked at the house number and complained, "I still can't tell."
- "Oh, that's right. I forgot to tell you that the oldest one taught the younger ones to play hide-and-seek."
- The census taker promptly wrote down the ages of the three children.
- How old are they?

13

- Imagine that you have an 8" x 8" array of 1" little squares and you also have a large box of 2" x 1" dominoes.
- Of course you can cover each of the 64 squares with the dominoes without any overlaps hanging off the array.
- Now imagine cutting out the upper right hand corner square and the lower left hand corner square. In this new configuration, is it possible to cover the 62 remaining squares with the dominoes with no overlaps or overhangs?

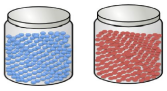






16

- There are two large jars. One jar is filled with a large number of blue beads, and the other is filled with the same number of red beads.
- Five beads from the red-bead jar are scooped out and dumped into the blue-bead jar. Someone then puts a hand in the blue-bead jar, scoops out five beads without knowing what color they are, and dumps them into the red-bead jar.
- Are there the same number of red beads in the red-bead jar as there are blue beads in the blue-bead jar?



10

- A closet has two red hats and three white hats. Three participants and a Game Master know that these 5 hats are the only ones in play.
- Three men sit on chairs and face each other. The first man has two good eyes, the second man has only one working eye, and the third man is totally blind.
- The Game Master places one hat from the closet on each man's head in such a way that no man can see the color of their own hat. Then she offers a deal as follows: Each participant is given the option of guessing the color of their hat. A correct guess brings a $50,000 prize; however, a false guess brings immediate death.
- The first man looks around the room and says, "I am not going to guess". Then the second man looks around the room and says, "I am not going to guess". Finally the third man says, "From what my friends with eyes have said, I can clearly see that my hat is _____".
- He wins the $50,000 and your task is to fill in the blank and explain how the blind man guessed the color of his hat correctly.

12

- A giant cheese cube is made up 27 smaller cheese cubes of various flavors so that it looks like a Rubik's cube.
- A worm first eats through a top corner flavor cube.
- After eating through any given cube, it goes on to eat an adjacent cube (one that shares a wall).
- The middle most cube is the most delicious.
- Is it possible for the worm to eat through all 27 cubes and finish last with the middle most cube?
- Are there other starting cubes that would allow him to finish last with the middle most cube?

15

- George lives at the bottom of a mountain, and there is a single narrow trail from his house to the top of the mountain where there is a campsite.
- At 6AM on Saturday he starts up the trail and without stopping or backtracking, reaches the top, and pitches his tent before 6PM.
- The next morning, on Sunday, he wakes up at 5AM, eats breakfast, and at exactly 6AM starts down the same trail as he had hiked up the previous day.
- He descends without stopping or backtracking and arrives home before 6PM.
- Must there be a time of day on Sunday where he was at exactly the same place on the trail at the same time as he was on the previous day?
- Could there be more than one such place?

17

Figure 3 19 insight problems, cont.

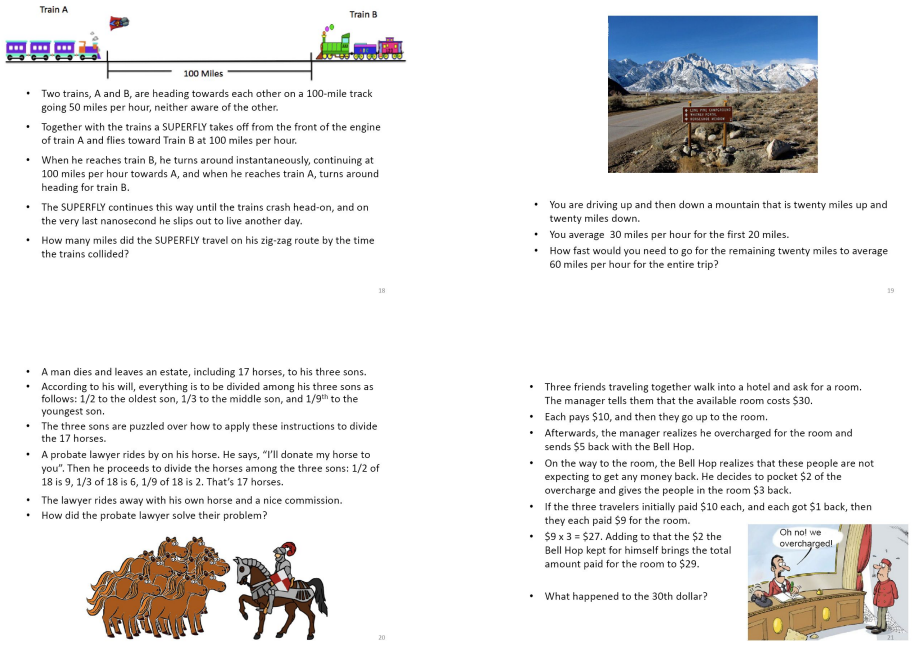


Figure 4 19 insight problems, cont.

7.2 Definition of Basic Terms in Insight Problem Solving, by Yun Chu

Some possible definitions for terms often used in insight problem solving.

▶ **Definition 1** (Problem). A problem occurs when there is an obstacle between a present state and a goal state and it is not immediately obvious how to get around the obstacle.

▶ **Definition 2** (Well-defined problem). It is a clear problem representation with the initial state, goal state, obstacles to the goal state, and the solution path stated.

▶ **Definition 3** (Ill-defined problem). It lacks a clear path to the solution or the operators are not specified.

▶ **Definition 4** (Insight problem). It is one type of ill-defined problem. 'obvious' solutions do not work, usually low solution rates, sudden realization of the solution.

▶ **Definition 5** (Metacognition). It is 'thinking about thinking.'

▶ **Definition 6** (Verbalization). It is talking about what you are doing/thinking while in the problem solving process.

▶ **Definition 7** (Feelings-of-warmth). It is asking the problem solver to provide this rating in answer to the question, 'how close do you feel to the solution?'

Below a list of insight problems is shown.
Some classic insight problems are listed below. For more see Appendix 7.1.

Verbal: Marsha and Marjorie were born on the same day of the same month of the same year to the same mother and father yet they are not twins. How is this possible?

Math: There are 10 bags, each containing 10 gold coins, all of which look identical. In 9 of the bags, each coin is 16 ounces, but in one of the bags, the coins are 17 ounces each. How is it possible (in a single weighing) to determine which bag contains the 17-ounce coins?

Spatial: 9-dot problem. Connect 3 rows of 3 dots each with 4 straight lines without lifting your pencil or retracing any lines.

Some recent insight problems are shown below.

Matchstick arithmetic: Move 1 matchstick to make the following statement true: `IV = III = I`

Compound remote associates: Find the solution word associated with all words of the triad forming 3 compound words: age mile sand

Rebus: What is the common saying (fill the empty part)? iii _____ ooo

Cheap necklace problem: Make a closed necklace with 4 chains of 3 links each. You have 15 cents total. It costs 2 cents to open a link and 3 cents to close it.

8-Ball problem: There are 8 balls in front of you. One of them is slightly heavier than the other 7. Using a balance scale only two times, how can you find the heavy ball?

7.3 Computational Search Problems, by Ulrike Stege

▶ **Definition 1** (Euclidean TSP)**.** Input: A set of points in the Euclidean Plane. Output: A shortest tour connecting all the points.

▶ **Definition 2** (Euclidean MST)**.** Input: A set of points in the Euclidean Plane. Output: A shortest network/graph connecting all the points.

▶ **Definition 3** (Vertex Cover)**.** Input: A(n undirected) graph $G = (V, E)$ Output: A smallest vertex cover V' for G. That is, a subset V' of V where for each edge xy in E, x or y is in V' and V' is as small as possible.

▶ **Definition 4** (Independent Set)**.** Input: A(n undirected) graph $G = (V, E)$ Output: A largest independent set V' for G. That is, a subset V' of V such that for each pair x, y of vertices in V', xy is not an edge in E and V' is maximized.

▶ **Definition 5** (Dominating Set)**.** Input: A(n undirected) graph $G = (V, E)$ Output: A smallest dominating set V' for G. That is, a subset V' of V such that for each vertex $x \in V$, x is in V' or x is adjacent to a vertex y that is in V' and V' is as small as possible.

8 Seminar Program

Monday, 29th of August 2011

Chair: Iris van Rooij

09:00 – 09.45	Introduction of the Seminar. Short presentation of the participants.
09:45 – 10:45	Todd Wareham. *What Does (and Does not) Make Problem Solving by Insight Easy? A Complexity-Theoretic Investigation.*
11:15 – 11:45	Georg Gottlob. *Living with Computational Complexity.*
11:45 – 12:15	Sarah Carruthers. *Vertex Cover and Human Problem Solving.*
14:00 – 14:30	Discussions
14:30 – 15:30	William Batchelder. *Some Issues in Developing a Theory of Human Problem Representation.*
16:00 – 16:30	Sashank Varma. *Spatial Problem Solving: The Optimal Deployment of Cortical Resources.*
16:30 – 16:40	Jakub Szymanik *Generalizing Muddy Children Puzzle.*
16:40 – 18:00	Working group and Discussions.

Tuesday, 30th of August 2011

Chair: Yll Haxhimusa

09:00 – 10.00	Niels Taatgen. *Human problem solving: the search for the right toolkit.*
10:00 – 10:30	Johan Kwisthout. *Relevant Representations.*
11:00 – 12:00	Rina Dechter. *Advanced Reasoning in Graphical models.*
14.00 – 15.00	Ken Forbus. *Analogy as a computational foundation for problem-solving and learning.*
15.00 – 15:30	Jelle van Dijk. *The way of the Ouroboros: How to represent a problem by solving it.*
16:00 – 16:30	Marco Ragni. *In Search of a Cognitive Complexity Measure for Matrix Reasoning Problems.*
16:30 – 18.00	Discussions and Working groups.

Wednesday, 31st of August 2011

Chair: Iris van Rooij

09:00 – 10.00	Dedre Gentner. *The Analogical Mind.*
10:00 – 10:30	Liane Gabora. *Problem Solving as the Recognition and Actualization of Potentiality.*
11:00 – 11:30	Daniel Reichman. *Speed-Accuracy Tradeoffs: A computational Perspective.*
14.00 –	Hiking.

Thursday, 1st of September

Chair: Yll Haxhimusa

09:00 – 09:30	Yun Chu. *Human Performance on Insight Problem Solving: A Review.*
09:30 – 10:00	Ute Schmid. *Learning Productive Rule Sets from Problem Solving Experience.*
10:00 – 10:30	Ulrike Stege. *Using (even more) Foundations from Computer Science to study Aspects of Human Problem Solving.*
11:00 – 12:00	Open Discussion.
14:00 – 14:30	Brendan Juba. *PAC Semantics: a Framework for Heuristic Rules.*
14:30 – 15:00	Nysret Musliu. *Algorithms for Computing (Hyper)tree Decompositions.*
15:00 – 15:15	Jered Vroon. *Problem Solving as Producing a Solution.*
15:15 – 15:30	Zygmunt Pizlo. *Multiresolution-multiscale pyramids and the traveling salesman problem.*
16:00 – 17:00	Preparation: Open Problem Session.
17:00 – 18:00	Open Problem Session.

Friday, 2nd of September 2011

Chair: Iris van Rooij

09:00 – 10.30	Working group, Discussions and Short Talks.
11:00 – 12:00	Wrap-up Session.

Participants

- William H. Batchelder
University of California, US
- Mark Blokpoel
Radboud Univ. Nijmegen, NL
- Sarah Carruthers
University of Victoria, CA
- Yun Chu
La Crescenta, US
- Rina Dechter
Univ. California – Irvine, US
- Kenneth D. Forbus
Northwestern University – Evanston, US
- Liane Gabora
University of British Columbia – Vancouver, CA
- Dedre Gentner
Northwestern University – Evanston, US
- Noah D. Goodman
Stanford University, US
- Georg Gottlob
University of Oxford, GB
- Marcin Grzegorzek
Universität Siegen, DE
- Yll Haxhimusa
TU Wien, AT
- Jens Hedrich
Universität Koblenz-Landau, DE
- Adrian Ion
TU Wien, AT
- Frank Jäkel
Universität Osnabrük, DE
- Brendan Juba
MIT – Cambridge, US
- Markus Knauff
Universität Giessen, DE
- Walter Kropatsch
TU Wien, AT
- Johan Kwisthout
Radboud Univ. Nijmegen, NL
- David Landy
University of Richmond, US
- Zoltan Miklos
EPFL – Lausanne, CH
- Nyrset Musliu
TU Wien, AT
- Zygmunt Pizlo
Purdue University – West Lafayette, US
- Marco Ragni
Universität Freiburg, DE
- Daniel Reichman
Weizmann Inst. – Rehovot, IL
- Ute Schmid
Universität Bamberg, DE
- Ulrike Stege
University of Victoria, CA
- Jakub Szymanik
University of Groningen, NL
- Niels A. Taatgen
University of Groningen, NL
- Susanne Tak
University of Canterbury – Christchurch, NZ
- Jelle van Dijk
TU Eindhoven, NL
- Iris van Rooij
Radboud Univ. Nijmegen, NL
- Sashank Varma
University of Minnesota, US
- Jered Vroon
Radboud Univ. Nijmegen, NL
- H. Todd Wareham
Memorial Univ. of Newfoundland, CA
- Gerhard Woeginger
TU Eindhoven, NL