Report from Dagstuhl Perspectives Workshop 16151

Foundations of Data Management

Edited by

Marcelo Arenas[1], Richard Hull[2], Wim Martens[3], Tova Milo[4], and Thomas Schwentick[5]

1 **Pontificia Universidad Catolica de Chile, CL,** marenas@ing.puc.cl
2 **IBM TJ Watson Research Center – Yorktown Heights, US,** hull@us.ibm.com
3 **Universität Bayreuth, DE,** wim.martens@uni-bayreuth.de
4 **Tel Aviv University, IL,** milo@cs.tau.ac.il
5 **TU Dortmund, DE,** thomas.schwentick@udo.edu

Abstract

In this Perspectives Workshop we have explored the degree to which principled foundations are crucial to the long-term success and effectiveness of the new generation of data management paradigms and applications, and investigated what forms of research need to be pursued to develop and advance these foundations.

The workshop brought together specialists from the existing database theory community, and from adjoining areas, particularly from various subdisciplines within the Big Data community, to understand the challenge areas that might be resolved through principled foundations and mathematical theory.

Perspectives Workshop April 10–15, 2016 – http://www.dagstuhl.de/16151
1998 ACM Subject Classification H.2 Database Management
Keywords and phrases Foundations of data management, Principles of databases
Digital Object Identifier 10.4230/DagRep.6.4.39
Edited in cooperation with Pablo Barceló

1 Executive Summary

Marcelo Arenas
Richard Hull
Wim Martens
Tova Milo
Thomas Schwentick

The focus of Foundations of Data Management (traditionally termed Database Theory) is to provide the many facets of data management with solid and robust mathematical foundations. The field has a long and successful history and has already grown far beyond its traditional scope since the advent of the Web.

The recent push towards Big Data, including structured, unstructured and multi-media data, is transforming and expanding the field at an unusually rapid pace. However, for understanding numerous aspects of Big Data, a robust research exploration into the principled foundations is still lacking. This transformation will call upon the Database Theory community to substantially expand its body of tools, techniques, and focal questions and to much more fully embrace several other disciplines, most notably statistics and probability

theory, natural language processing, data analytics, emerging hardware and software supports for computation, and data privacy and security.

Big Data is not the only force that is driving expansion and transformation for the Foundations of Data Management. With the increasing digitization of diverse industries, including "smarter cities", education, healthcare, agriculture and others, many diverse kinds of data usage at large scales are becoming crucial. The push towards data-centric business processes, which are especially important for knowledge-worker driven processes, raise fundamentally new questions at the intersection of data and process. And increasing adoption of semantic web and other ontology-based approaches for managing and using meta-data push the boundaries of traditional Knowledge Representation.

The purpose of this Dagstuhl Perspectives Workshop was to explore the degree to which principled foundations are crucial to the long-term success and effectiveness of the new generation of data management paradigms and applications, and to understand what forms of research need to be pursued to develop and advance these foundations.

For this workshop we brought together specialists from the existing database theory community, and from adjoining areas, such as Machine Learning, Database Systems, Knowledge Representation, and Business Process Management, to understand the challenge areas that might be resolved through principled foundations and mathematical theory.

More specifically, during this workshop we worked on:

- Identifying areas, topics and research challenges for Foundations of Data Management in the forthcoming years, in particular, areas that have not been considered as Database Theory before but will be relevant in the future and of which we expect to have papers at PODS and ICDT, the main conferences in the field.
- Outlining the techniques that will be most fruitful as starting points for addressing the new foundational challenges in Data Management.
- Characterising the major challenge areas in Big Data where a principled, mathematically-based approach can provide important contributions.
- Finding research goals in neighbouring areas that may generate synergies with our own.

The workshop consisted of eight invited tutorials on selected topics: (1) Managing Data at Scale, (2) Uncertainty and Statistics in Foundations of Data Management, (3) Human in the Loop in Data Management, (4) Machine Learning and Data Management, (5) Data-Centric Business Processes and Workflows, (6) Ethical Issues in Data Management, (7) Knowledge Representation, Ontologies, and Semantic Web, and (8) Classical DB Questions on New Kind of Data. The abstracts of these talks can be found below in the document.

There were also seven working groups on theory-related topics, which identified the most relevant research challenges for Foundations of Data Management in the forthcoming years, outlined the mathematical techniques required to tackle such problems, and singled out specific topics for insertion in a curriculum for the area. The topics of these working groups were: (1) Imprecise Data, (2) Unstructured and Semi-structured Data, (3) Process and Data, (4) Data Management at Scale, (5) Data Management and Machine Learning, (6) Knowledge-Enriched Data Management, and (7) Theory and Society. There was also a working group on curriculum related issues, that collected and enriched the information provided by the working groups about the design of a curriculum on Foundations of Data Management. Each one of these groups worked for two consecutive hours in different days. Workshop participants had to participate in at least two working groups, although most of the people participated in four of them. Summaries of the discussions held in each one of these working groups can be found below in the document.

During the first day of the workshop, there were also five working groups that analysed several community-related aspects. In particular: (1) Attraction of women and young mem-

bers, (2) cross-fertilization with neighbouring areas, (3) relationship to industry, (4) impact of our research, and (5) the publishing process. The discussion within some of these working groups gave rise to the creation of specific tasks to be accomplished by the community in the following years. These tasks will be coordinated by the councils of PODS and ICDT, the two main conferences in the field.

This Dagstuhl Report will be accompanied by a Dagstuhl Manifesto, in which the outcome of the different working groups will be explained in more detail and several strategies for the development of our field will be proposed.

2 Table of Contents

3 Overview of Talks

3.1 Issues in Ethical Data Management

Serge Abiteboul (ENS – Cachan, FR)

In the past, database research was driven by data model and performance. In the future, personal/social data and ethics will. What should be done? Change how we deal with personal data? Change the web? We will consider various issues: Privacy, data analysis, data quality evaluation, data dissemination, and data memory.

3.2 (Non-)Classical DB Questions on New Kinds of Data

Marcelo Arenas (Pontificia Universidad Catolica de Chile, CL) and Pablo Barceló (University of Chile – Santiago de Chile, CL)

Different data models have been proposed in the last 20 years for representing semistructured or unstructured data. These include, e.g., graph databases, XML, RDF, JSON, and CSV. We explain that over such models we have (a) some classical DB questions being particularly important (e.g., schema design, query languages, and updates), (b) other classical DB questions gaining renewed interest (e.g., uncertainty, distribution, workloads, etc), and (c) some new questions appearing in full force (e.g., trust, access, variety, schema extraction).

Via examples we show that these problems are not only of theoretical interest, but, more importantly, that theory can have a strong positive impact in their practice. We start by presenting the work done in the standardization of a declarative query language for graph databases, which has concentrated the efforts of academics and practitioners in the last year. In this case theory has helped identifying a reasonable tradeoff between expressiveness and computational cost for the language. We then concentrate on the case of RDF and its query language SPARQL. We briefly review several practical problems that remain open regarding SPARQL evaluation in a distributed and open environment (the Web), and sketch the kind of theoretical understanding that is needed for their solution.

3.3 Knowledge Representation, Ontologies, and Semantic Web

Georg Gottlob (University of Oxford, GB) and Carsten Lutz (Universität Bremen, DE)

The tutorial gave an overview of the recent developments in data access with ontologies. The first part focussed on the case where ontologies are formulated in a description logic (DL), surveying in particular the different families of DLs and their relations, as well as current research topics. The second part concentrated on the case where ontologies are formulated

in existential rule languages, which generalize Horn DLs and other relevant formalisms. It surveyed relevant language families and their relations, as well as recent results.

3.4 Machine Learning and Data Management

Eyke Hüllermeier (Universität Paderborn, DE)

This tutorial-style presentation starts with a brief introduction to machine learning, specifically tailored for an audience with a background in database theory. The main part of the talk is then devoted to an overview of large-scale machine learning, that is, the application of machine learning methods to massive amounts of data, where scalability and efficient data management clearly become an issue. Finally, the talk also addresses the question of what machine learning can contribute to database management.

3.5 Uncertainty and Statistics in Foundations of Data Management

Benny Kimelfeld (Technion – Haifa, IL) and Dan Suciu (University of Washington – Seattle, US)

In this tutorial we outline the state of affairs on the research of managing probabilistic and statistical data, while focusing on the angle of database theory. We begin with an overview of past research, where we cover three main aspects: modeling probabilistic databases, complexity of probabilistic inference, and relevant applications. In particular, we describe established connections between query evaluation over tuple-independent databases, weighted model counting, and inference over Markov Logic Networks. We also discuss some key techniques such as lifted inference, tree decomposition, and knowledge compilation. Finally, we propose several directions for future research, where we make the case for a tighter incorporation of the sources of uncertainty (beyond the uncertainty itself) into the formal models and associated algorithms.

3.6 Human-in-the-loop in Data Management

Tova Milo (Tel Aviv University, IL)

Modern data analysis combines general knowledge stored in databases with individual knowledge obtained from the crowd, capturing people habits and preferences. To account for such mixed knowledge, along with user interaction and optimisation issues, data management platforms must employ a complex process of reasoning, automatic crowd-task generation and result analysis. This tutorial introduces the notion of crowd mining and describes a generic architecture for crowd mining applications. This architecture allows to examine and

compare the components of existing crowdsourcing systems and point out extensions required by crowd mining. It also highlights new research challenges and potential reuse of existing techniques/components. We exemplify this for the OASSIS project, a system developed in Tel Aviv University, and for other prominent crowdsourcing frameworks.

3.7 Data Management at Scale

Dan Suciu (University of Washington – Seattle, US) and Ke Yi (HKUST – Kowloon, HK)

The tutorial starts with a brief discussion of technology trends relevant to scalable data management research. Next, it presents four theoretical results on the complexity of multi-join query evaluation: the AGM bound bound and worst case sequental algorithms, the communication cost for parallel alagorithm, the I/O cost in the external memory model, and recent sampling-based approximation algorithms. It ends with a brief discussion on potential topics to be included in a future course or book on Foundations of Data Management.

3.8 Data-Centric Business Processes and Workflows

Victor Vianu (University of California – San Diego, US) and Jianwen Su (University of California – Santa Barbara, US)

A business process (BP) is an assembly of tasks to accomplish a business goal. A workflow is a software system managing business process executions. Traditional BP/workflow models do not include data, which makes it difficult to accurately specify processes in which data plays a central role, and raises many problems for their analysis, inter-operation, and evolution. This tutorial provides an introduction to the practical issues in data-centric BP/workflow management, and surveys existing theoretical results, addressing the following topics: (1) Formal analysis, verification, and synthesis of workflows, (2) workflow management system design, and (3) other workflow management issues including process mining and interoperation. We argue that the marriage of process and data is a fertile ground for new database theory questions, bringing into play a mix of techniques from finite-model theory, automata theory, computer-aided verification, and model checking. The new setting requires extending classical approaches to views, data exchange, data integration, expressiveness measures, and incomplete information.

4 Working groups

4.1 Theory and Society

Serge Abiteboul (ENS – Cachan, FR)

(*Moderator:* Sege Abiteboul. *Participants:* Sudeepa Roy, Juan Reutter, Ronald Fagin, Jianwen Su, Thomas Schwentick, Stijn Vansummeren, Thomas Eiter, Iovka Boneva).

Society asks questions such as: (a) Who owns my medical data? (b) Is the information I receive of good quality? (c) Can Google influence the US president election? Theory asks by: (a) Access control (b) Tools for fair data analysis (c) Tools for defining and checking bias.

Regarding the first one, access control, it should consider questions such as "who can access sensitive data?" and "how is sensitive data being used?" In this context we think there is a need for developing:

- Models for access control over distributed data.
- Languages and user friendly tools for (i) specifying how the data can be used, (ii) understanding which data has been used for what, and (iii) control whether the data has been used in agreement with the access policy.

This might allow nice users to be guided in order to understand which queries are forbidden. It might also help detecting malicious users a posteriori.

Regarding the second one, fair data analysis, it might be of help for the data analyst (e.g., do I make good quality data analysis?) and the non-expert user (e.g., is this a relevant recommendation?). We think that in this context there is a need for developing:

- User friendly tools that check whether data analysis was fair.
- Helper tools for data analysts that check fairness criteria and guide in correcting the fairness errors.

Regarding the third one, bias and responsibility, this is of great importance not only for society but also for researchers (e.g., what can we do as researchers for designing less biased web services with high responsibility?). We think that in this context there is a need for developing:

- Political solutions (laws, control).
- Definitions of what it means having bias.
- Ranking algorithms that take into account the quality of pages.
- Technical tools for checking for the existence of bias, e.g., verification, testing, etc.

We also believe that several of these topics could be included in the curriculum of undergrad students, e.g., laws related to data, ethics (in general, for computer science), privacy issues and possible solutions, etc.

4.2 Knowledge-enriched Data Management

Diego Calvanese (Free University of Bozen-Bolzano, IT)

(*Moderator:* Diego Calvanese. *Participants:* Andreas Pieris, Carsten Lutz, Claire David, Filip Murlak, Georg Gottlob, Magdalena Ortiz, Marcelo Arenas, Meghyn Bienvenu, Paolo Guagliardo, Reinhard Pichler, Thomas Eiter, Jianwen Su, Ron Fagin, Sudeepa Roy).

The working group identified four important practical challenges:

1. Develop personalized and context-aware data access and management tools. Data in this case is highly heterogenous, multi-model and multi-modal. Here we deal with "small data" at the individual level, tuned to different view points.
2. Providing end users flexible and integrated access to large amounts of complex, distributed, heterogenous data (under different representations and different models). End users should be assumed to be domain experts, not data management experts.
3. Ensure interoperability at the level of systems exchanging data.
4. Bring knowledge to data analytics and data extraction.

It also identified seven relevant theoretical challenges related to the latter:

1. Development of reasoning-tuned DB systems, including new optimizations, new cost models, new/improved database engines optimized for reasoning, approximate answers, distributed evaluation, etc.
2. Choosing/designing the right languages for supporting these tasks. Here we need pragmatic choices motivated by user needs, but also supporting different types of knowledge and data (e.g., mixing CWA+OWA, temporal, spatial, etc.)
3. We need new measures of complexity for understanding easy/difficult cases, that explain better what works in practice. It would be interesting to explore alternative complexity measures, such as parameterized and average complexity, measuring complexity on the Web, smoothed analysis, etc.
4. Building user-friendly interfaces (beyond Protege). In particular, we need tools support geared towards end users (i.e., domain experts, lay people, and not just IT/knowledge engineers), but also support for query formulation, tools for exploration, etc.
5. Developing next-generation reasoning services. Here we need to explore notions related to explanation, abduction, hypothetical reasoning, defeasible reasoning, etc.
6. Reasoning with imperfect data, e.g., reasoning under contradictions and/or uncertainty, reasoning about quality of data, and support for improving data quality.
7. In depth study of temporal and dynamic aspects, such as managing changing data and knowledge, streaming data, reasoning about change, updating data in the presence of knowledge, etc.

4.3 Semistructured and Unstructured Data

Claire David (University Paris-Est – Marne-la-Vallée, FR)

(*Moderator:* Claire David. *Participants:* Iovka Boneva, Wim Martens, Juan Reutter Magdalena Ortiz, Domagoj Vrgoc, Filip Murlak, Frank Neven, Pablo Barcelo, Marcelo Arenas, Serge Abiteboul, Torsten Grust, Thomas Schwentick).

The huge amount of available data is perceived as a clear asset. However, exploiting this data meets several obstacles, as for instance the well known 3V: volume, velocity, variety of the data. One particular aspect of the variety of data is the co/existence of different formats for semi-structured and unstructured data, in addition to the widely used relational data format. A non exhaustive list is tree-structured data (e.g. XML, json), graph data (RDF an others), tabular data (e.g., CSV), temporal and spatial data, text, multimedia. We can expect that in the near future, new data formats will arise in order to cover particular needs.

The existence and coexistence of various formats is not new, but we believe that recent changes in the nature of available data raise a strong need for a new principled approach for dealing with different data models.

The database community has been working actively for many years now towards understanding each of these data formats formally by abstracting them, by developing good notions of schema, query languages, mappings and by studying the complexity of the related problems. These questions are still relevant and constitute important challenges.

Data heterogeneity is also an old problem in the database community, and until now has been tackled by data integration and data exchange solutions. A possible way for integrating data of multiple formats is to import all the data into a relational database, and use an RDMS afterwards for querying and managing the data; this solution is currently widely applied. The working group agreed that such a solution does have advantages, but does not fit all use cases; for instance, some data sources may contain dynamic data, and integration requires a lot of effort.

Additionally, the nature of available data has changed since the classical data integration and data exchange solutions have been proposed. The proliferation of many different data formats, the increase in the amount of available data, and the limited control users have on the data they are using, all challenge these solutions.

Therefore, we need a principled approach for dealing with different data models. For usability reasons, the new approach to be proposed must satisfy the following constraints:

A. It should be possible to keep the data in its original format, while still accessing data from different sources from a unique interface.
B. The approach should allow for adding new data models in the future.

The working group identified a number of problems that need to be solved in order to achieve this goal:

1. Understanding the data:
   - How to abstract and represent information about the structure of such data stored in various format? (Is there a good notion of multimodel schema? Can we use ontology?)
   - As open or third party data comes sometimes with no meta information available, we need tools for extracting such schema or at least partial structural information from data.
   - Develop entity resolution tools.

- Provide mapping between data in various formats.
- Provide tools for extracting statistical properties of data, and tools for data analytic.

2. Accessing the data: There is a need for:
   - Specialized query languages for the different data models.
   - A general query language that can combine various specialized query languages.
   - Provide efficient algorithm for planning and evaluating a query using structural information of data.
   - Methods and tools for data summarization and data visualization.
   - Provide user friendly paradigms for both presenting the data, information about it structure and formulating queries.
3. Orthogonal aspects:
   - Data in some applications can be highly distributed. How to handle distributed data processing, indexing, costs model?
   - Models for representing trust for data/information, privacy and data quality.
   - Do not forget the user who need usable tools to be able to understand and access the data.

4.4 Data Management and Machine Learning

Eyke Hüllermeier (Universität Paderborn, DE)

(*Moderator:* Eyke Hüllermeier. *Participants:* Rick Hull, Floris Geerts, Benny Kimelfeld, Pablo Barcelo, Jens Teubner, Jan Van den Bussche, Stijn Vansummeren).

Machine Learning (ML) often focuses on large-scale learning, but several other directions exist that might be of interest to data management. These include: (a) Logic: Learning of formulas, concepts, ontologies, and so on (work in database theory and description logics), (b) Automata: Learning automata (which might have applications for learning XML schemas, for instance), (c) Workflows: Process mining, process discovery, and others, and (d) Query languages: Extensions of languages used in DBMS by ML functionalities (e.g., query-by-example), ILP, declarative data mining, etc.

We believe that there are several ML problems in which a data management perspective could be oh help. These include:

1. Applying data management techniques in the optimization of ML algorithms, e.g., implementation of operators for data access, sequential vs parallel sampling, etc.
2. A finer complexity analysis of ML algorithms, in particular, their complexity in terms of different parameters such as the number of features, labels, label density, etc.
3. Studying I/O complexity of ML algorithms in external memory.
4. Building models for distributed ML based on frameworks such as MapReduce.

We also believe that data management could benefit from ML. For example:

1. Building query languages for DBMS that incorporate ML functionalities. These can be of importance for optimization, cleaning, analysis, etc.
2. Use of deep networks for semantic indexing/hashing.
3. Learning hash functions.
4. Predictive modeling, performance prediction.
5. Learning approximate dependencies, concepts, ontologies, etc.

4.5 Imprecise data

Leonid Libkin (University of Edinburgh, GB)

(*Moderator:* Leonid Libkin. *Participants:* Meghyn Beinvenu, Angela Bonifati, Diego Calvanese, Paolo Guagliardo, Benny Kimelfeld, Phokion Kolaitis, Maurizio Lenzerini, Carsten Lutz, Tova Milo, Dan Olteanu, Sudeepa Roy, Dan Suciu, Jan Van den Bussche, Victor Vianu).

Incomplete, uncertain, and inconsistent information is ubiquitous in data management applications. This was recognized already in the 1970s, and since then the significance of the issues related to incompleteness/uncertainty has been steadily growing: it is a fact of life that data we need to handle on an everyday basis is rarely complete.

However, while the data management field developed techniques specifically for handling incomplete data, their current state leaves much to be desired. Even evaluating SQL queries over incomplete databases – a problem one would expect to be solved after 40+ years of relational technology – one gets results that make people say *"you can never trust the answers you get from [an incomplete] database"* (Date). And while even such basic problems remain unsolved, we now constantly deal with more varied types of incomplete and inconsistent data. There is thus an urgent need to address these problems from a theoretical point of view, keeping in mind that theoretical solutions must be usable in practice (indeed, this is the field where perhaps too much theoretical work focused on proving various impossibility results – intractability, undecidability – rather that addressing what can actually be done efficiently).

The main challenges are split into three groups.

Modeling uncertainty: This includes several themes, for example: (a) What are types of uncertainty we need to model/understand? (b) How do we store/represent uncertain information? What standard RDBMSs give us, for example, is very limited. (c) When can we say that some data is true? This issue is particularly relevant in crowdsourcing applications: having data that looks complete does not yet mean it is true. (d) How do we rank uncertain query answers? There is a tendency to divide everything into certain and non-certain answers, but this is often too coarse.

Reasoning with uncertainty There is much work on this subject but we want to address points close to data management: (a) How do we do inferences with incomplete data? (b) How do we integrate different types of uncertainty? (c) How do we learn queries on uncertain data? (d) What do query answers actually tell us if we run queries on data that is uncertain (that is, how results can be generalized from a concrete incomplete data set)?

Making it practical This is the most challenging direction. For far too long, theoretical literature identified small classes where queries behave well over incomplete data (often various classes of conjunctive queries) and then concentrated on proving intractability results outside those classes. We need to move on, and look at questions like those below: (a) How do we find good quality query answers even when we have theoretical intractability? For instance, how do we find answers with some correctness guarantees, and do so efficiently? (b) How do we make commercial RDBMS technology work well and efficiently in the presence of incomplete data? Even query optimization in this case is hardly a solved problem. (c) How do we make handling inconsistency (in particular,

consistent query answering) work in practice? How do we use it in data cleaning? (d) How do we achieve practically feasible query evaluation on probabilistic data?

Regarding the question of what problems do we need to solve soon. We believe that while all the above are important research topics that need to be addressed, it appears that there are several that can be viewed as a priority, not least because there is an immediate connection between theory and practice. This gives us a chance to make a good theory to solve practically relevant problems, and to make this theoretical work visible.

1. Can we fix the standard relational technology so that at least textbook writers would stop saying (justifiably!) that *"you're getting wrong answers to some of your queries"*. We need to understand what it means to be right/wrong, and how to adjust the technology to ensure that wrong answers do not appear.
2. What should we use as benchmarks when working with incomplete/uncertain data? Quite amazingly, this has not been addressed; in fact standard benchmarks tend to just ignore incomplete data.
3. How do we devise approximation algorithms for classes of queries known to be intractable? The field is too dependent on producing results for conjunctive queries and leaving the rest for proving high-complexity results, but non-conjunctive queries do exist and need to be handled.
4. Is there any hope of making consistent query answering practical, and relevant (perhaps in data cleaning). Again too much emphasis was on proving dichotomies, even within conjunctive queries, and not making it work.

Regarding neighboring communities, we believe that our closest neighbors are database systems people (in terms of conferences, SIGMOD, VLDB). There are also interesting overlaps with things happening in uncertainty in AI (conferences such as UAI or SUM) and general AI conferences (KR, AAAI, IJCAI).

4.6 Curriculum for Foundations of Data Management

Frank Neven (Hasselt University – Diepenbeek, BE)

(*Moderator:* Frank Neven. *Participants:* Phokion Kolaitis, Thomas Schwentick, Leonid Libkin, Torsten Grust, Wim Martens, Pablo Barcelo, Reinhard Pichler, Marcelo Arenas, Stijn Vansummeren, Juan Reutter).

The working group on a curriculum for Foundations of Data Management (FDM) focused on the following questions. While the main goal of the group was to identify a curriculum for an FDM course on graduate level, the group also explored the relationship with other courses. A driving principle followed by the group was that we should not only prepare the next generation of PhD students in database theory (educate), but also expose the beauty and variety of our field to attract future PhD students (attraction) The group considered the following questions:

1. What are relevant concepts from FDM that could be featured as examples in non-database courses? The group gave the following examples of concepts: (a) trees can be related to XML, (b) graphs to the semantics Web, (c) regular expressions to querying of graphs

and to content models in XML Schema, (d) mathematical structures in a logic course as examples of databases, and (e) laws related to data in an ethics in CS course.

2. What are the relevant concepts from FDM that can be featured in a first DB course (undergraduate level)? The following examples were given:
   - Data models: Relational, semi-structured, graphs.
   - Relational calculus as an abstraction of a query language.
   - Recursion and Datalog (as there is not much syntax involved here, you can quickly explain it and even use a system like LogicBlox to bring the concept in practice).
   - Well-designedness, constraints, normalization / BCNF.
   - The concept that data often doesn't fit into main memory (memory hierarchy / external memory).
   - Concurrency (an example here is a proof of the fact that 2-phase locking works).
   - The concept of incomplete information (e.g., the pitfalls of NULLS in SQL).
3. What are the relevant topics for a graduate course in FDM? As a whole Dagstuhl seminar could be devoted to just this question, the following should just be seen as a start of discussion and would need further input from the community. Possible core topics that at the moment are not treated in depth in the AHV book:
   - Refined ways for studying complexity of queries.
   - Acyclic queries / Structural decomposition / tree decomposition (parametrized complexity).
   - AGM bound; join algorithms; leapfrog.
   - Data exchange and integration / relationship to KR.
   - Data provenance.
   - Graph data, RDF / Tree data, XML, JSON (tree automata).
   - Approximate query answering, top-k.
   - Stream-based processing.
   - External memory.
   - Incomplete information / Probabilistic databases.
   - Possible emerging topics: Map-reduce, parallel computation models.
4. How should we proceed to write a book on FDM? It is safe to say that there was no general agreement on how to proceed. One possibility is to organize the book TATA-style where an editorial board is established that decides the topics, invites writers and ensures global consistency between the material. The effort could start from the AHV book by adapting and adding chapters.

4.7 Process and Data

Victor Vianu (University of California – San Diego, US)

(*Moderator:* Victor Vianu. *Participants:* Thomas Schwentick, Thomas Eiter, Daniel Deutch, Magdalena Ortiz, Diego Calvanese, Jianwen Su, Richard Hull, Wim Martens, Serge Abiteboul).

Traditionally, workflow models have been process centric. Such a workflow specifies the valid sequencing of tasks by some finite-state transition diagram. This is inadequate in applications where data plays a central role, and has led to new workflow models in

which data is treated as a first-class citizen. The marriage of process and data represents a qualitative leap analogous to that from propositional logic to first-order logic.

Participants to the working group identified practical and technical challenges raised by data-centric workflows. The following were identified as important practical challenges:

1. Evolution and migration of business processes.
2. Automating manual processes (including workflow-on-the-fly).
3. Business Process compliance, correctness, non-functional properties, etc.
4. Business Process interaction and interoperation.
5. Business Process discovery and understanding (including analytics).
6. Workflow and business process usability (including human collaboration).

Tackling the above challenges raises a set of technical problems enumerated below (next to each are listed the relevant practical challenges):

A. Verification and Static Analysis (relevant to 1,2,3,4).
- Approximate verification via abstraction.
- Incremental verification under workflow or property modifications.
- Enabling/verifying "plug and play" of process building blocks.
- Checking compliance (financial, government, security).

B. Tools for Design and Synthesis (relevant to 1,2,4,6).
- Operators for creating/modifying process schemas.
- Full or partial synthesis from requirements and/or goals.
- Inferring process dynamically from "digital exhaust".

C. Models and semantics for views, interaction, and interoperation (relevant to 1,2,3,4,5,6).
- Providing customized views of business process schemas and instances.
- Consistent usage of shared data and resources by multiple processes.
- Views and Composition (e.g., orchestration, choreography, etc).
- Inferring/guaranteeing properties of process compositions.

D. Analytics with Process and Data (relevant to 1,2,3,4,5,6).
- "Business Intelligence" – analytics over executions of business processes.
- Process mining, process discovery.
- Automatic identification of best practices.
- Providing guidance and explanation at run time.

The participants also agreed that the lack of real data presents an obstacle to research in the area and that developing closer connections to industry would be beneficial. It is also desirable to explore further areas of applicability of this research, such as social networks or the blockchain approach used in financial transactions.

4.8 Managing data at scale

Ke Yi (HKUST – Kowloon, HK)

(*Moderator:* Ke Yi. *Participants:* Floris Geerts, Reinhard Pichler, Jan van den Bussche, Frank Neven, Filip Murlak).

In this working group we started by identifying different computational models that have been proposed for managing data at scale and some others that might be important to consider:

- **PRAM:** It has had some beautiful theory, but made little practical impact, reasons including 1) too far from reality, 2) hard to program. Nevertheless, there are many algorithmic ideas that are still very useful in other parallel models.
- **External memory:** Considered as a success, used as the standard model for DBMS algorithms. A related model is the cache-oblivious model which tries to model multiple levels of memory hierarchy. It is an elegant model, with some implementation work. But did not gain enough traction due to complication of the algorithms.
- **Streaming model:** Another successful model, with both nice theory and practical systems. More importantly, it has nurtured data sketching (data summarization in general), which is not only useful for data streams, but as a general approach in making big data small.
- **BSP, MapReduce, MPC, Pregel:** The original BSP model did not get enough attention due to complication: it had too many parameters and tried to model various costs. A computational model has to be both simple and realistic in order to be useful. The MPC model simplified the model and seems to hit the right simplicity-reality tradeoff. There is much active research, and more is expected to come.
- **Asynchronous models:** BSP, MPC etc are all synchronous models, but asynchronous models have not been studied fully in the database theory community (better studied in the distributed computing community).
- **Computational models for new hardware:** There are many practical researches in using GPU for speeding up database operations, but little theoretical work due to various architectures. CUDA might be a good model for more theoretical work. There are also many new chip designs that may be worth of looking at.
- **Relationship between the models:** Are there any general reductions between these models? Essentially, all these models are about locality and parallelism. If there is no general reduction, maybe there can be one for a class of algorithms or a class of problems (e.g., CQ, relational algebra).

We also identified several problems, where scalability is a major concern, that it would be relevant for the DB theory community to tackle:

- **Standard DB problems:** Of course, we should continue to work on standard DB problems (relational algebra, in particular joins, semi-structured data, XML, etc). Meanwhile, we should pay attention to the theory-practice gap.
- **Graph problems:** There is a lot of work on page rank, shortest path, mining, etc. Still many more needs to be solved at scale.
- **Machine learning at scale:** Machine learning is becoming more data-intensive. Linear algebra is a basic toolbox that needs to be scalable. New paradigms that are related to this community include distributed machine learning and learning on streams.
- **Transactions:** More theoretical work needs to address transactions. Issues like strong consistency and eventual consistency. Interaction with the distributed computing community might be desirable.
- **Scientific computing:** In particular large-scale simulation, which is becoming a dominating approach for scientific research.

We also identified some key techniques that these problems might require for their solution:

- **Beyond-worst-case analysis:** Worst-case analysis is still the first step in understanding the complexity of a problem, but may not be really relevant to an algorithm's practical performance. Many other angels should be explored, including, e.g., instance optimality, parameterized complexity, and smoothed complexity.

- **Algorithms specifically tailored for big data:** Just saying that a problem is in PTIME may not be enough for big data. We'd like to have algorithms that run in linear-time, near-linear time, or even sub-linear time. Parallel complexity is another important measure.
- **Approximation algorithms:** As data gets big, exact answers are not often required, and approximation algorithms can offer great efficiency speedup in this case. Sampling, MCMC, stratified sampling (BlinkDB), sampling over joins (by random walks) and many statistical techniques can be applied. Sublinear-time algorithms have been extensively studied in the algorithms community, but not fully exploited yet in DB.
- **Convex optimization:** In particular primal-dual methods, which have been recently applied for analyzing the maximum join size (AGM). It would be interesting to see if this can leave to more applications for DB questions.
- **Information theory:** We think it can be useful in proving lower bounds. It has been extensively used in the TCS community (communication complexity, etc), with a few examples in DB (communication lower bound for MPC model). We expect more applications to come.

Participants

- Serge Abiteboul
ENS – Cachan, FR
- Marcelo Arenas
Pontificia Universidad Catolica de Chile, CL
- Pablo Barcelo
University of Chile – Santiago de Chile, CL
- Meghyn Bienvenu
University of Montpellier, FR
- Iovka Boneva
Université de Lille I, FR
- Angela Bonifati
University Claude Bernard – Lyon, FR
- Diego Calvanese
Free Univ. of Bozen-Bolzano, IT
- Claire David
University Paris-Est – Marne-la-Vallée, FR
- Daniel Deutch
Tel Aviv University, IL
- Thomas Eiter
TU Wien, AT
- Ronald Fagin
IBM Almaden Center – San Jose, US
- Floris Geerts
University of Antwerp, BE
- Georg Gottlob
University of Oxford, GB
- Torsten Grust
Universität Tübingen, DE
- Paolo Guagliardo
University of Edinburgh, GB
- Eyke Hüllermeier
Universität Paderborn, DE
- Richard Hull
IBM TJ Watson Res. Center – Yorktown Heights, US
- Benny Kimelfeld
Technion – Haifa, IL
- Phokion G. Kolaitis
University of California – Santa Cruz, US
- Maurizio Lenzerini
Sapienza University of Rome, IT
- Leonid Libkin
University of Edinburgh, GB
- Carsten Lutz
Universität Bremen, DE
- Wim Martens
Universität Bayreuth, DE
- Tova Milo
Tel Aviv University, IL
- Filip Murlak
University of Warsaw, PL
- Frank Neven
Hasselt Univ. – Diepenbeek, BE
- Dan Olteanu
University of Oxford, GB
- Magdalena Ortiz
TU Wien, AT
- Reinhard Pichler
TU Wien, AT
- Andreas Pieris
TU Wien, AT
- Juan L. Reutter
Pontificia Universidad Catolica de Chile, CL
- Sudeepa Roy
Duke University – Durham, US
- Thomas Schwentick
TU Dortmund, DE
- Jianwen Su
University of California – Santa Barbara, US
- Dan Suciu
University of Washington – Seattle, US
- Jens Teubner
TU Dortmund, DE
- Jan Van den Bussche
Hasselt University, BE
- Stijn Vansummeren
Université Libre de Bruxelles, BE
- Victor Vianu
University of California – San Diego, US
- Domagoj Vrgoc
Pontificia Universidad Catolica de Chile, CL
- Ke Yi
HKUST – Kowloon, HK