Manifesto from Dagstuhl Perspectives Workshop 16151

Research Directions for Principles of Data Management

Serge Abiteboul[1], Marcelo Arenas[2], Pablo Barceló[3], Meghyn Bienvenu[4],
Diego Calvanese[5], Claire David[6], Richard Hull[7], Eyke Hüllermeier[8],
Benny Kimelfeld[9], Leonid Libkin[10], Wim Martens[11], Tova Milo[12],
Filip Murlak[13], Frank Neven[14], Magdalena Ortiz[15], Thomas Schwentick[16],
Julia Stoyanovich[17], Jianwen Su[18], Dan Suciu[19], Victor Vianu[20], and
Ke Yi[21]

1   ENS – Cachan, FR
2   Pontificia Universidad Catolica de Chile, CL, `marenas@ing.puc.cl`
3   DCC, University of Chile – Santiago de Chile, CL
4   University of Montpellier, FR
5   Free Univ. of Bozen-Bolzano, IT
6   University Paris-Est – Marne-la-Vallée, FR
7   IBM TJ Watson Research Center – Yorktown Heights, US, `hull@us.ibm.com`
8   Universität Paderborn, DE
9   Technion – Haifa, IL
10 University of Edinburgh, GB
11 Universität Bayreuth, DE, `wim.martens@uni-bayreuth.de`
12 Tel Aviv University, IL, `milo@cs.tau.ac.il`
13 University of Warsaw, PL
14 Hasselt Univ. – Diepenbeek, BE
15 TU Wien, AT
16 TU Dortmund, DE, `thomas.schwentick@udo.edu`
17 Drexel University — Philadelphia, US
18 University of California – Santa Barbara, US
19 University of Washington – Seattle, US
20 University of California – San Diego, US
21 HKUST – Kowloon, HK

Abstract

The area of Principles of Data Management (PDM) has made crucial contributions to the development of formal frameworks for understanding and managing data and knowledge. This work has involved a rich cross-fertilization between PDM and other disciplines in mathematics and computer science, including logic, complexity theory, and knowledge representation. We anticipate on-going expansion of PDM research as the technology and applications involving data management continue to grow and evolve. In particular, the lifecycle of Big Data Analytics raises a wealth of challenge areas that PDM can help with.

In this report we identify some of the most important research directions where the PDM community has the potential to make significant contributions. This is done from three perspectives: potential practical relevance, results already obtained, and research questions that appear surmountable in the short and medium term.

Executive Summary

In April 2016, a community of researchers working in the area of Principles of Data Management (PDM) joined in the Dagstuhl Castle in Germany for a workshop organized jointly by the Executive Committee of the ACM Symposium on Principles of Database Systems (PODS) and the Council of the International Conference on Database Theory (ICDT). The mission of this workshop was to identify and explore some of the most important research directions that have high relevance to society and to Computer Science today, and where the PDM community has the potential to make significant contributions. This report describes the family of research directions that the workshop focused on from three perspectives: potential practical relevance, results already obtained, and research questions that appear surmountable in the short and medium term. This report organizes the identified research challenges for PDM around seven core themes, namely *Query Processing at Scale*, *Multi-model Data*, *Uncertain Information*, *Knowledge-enriched Data*, *Data Management and Machine Learning*, *Process and Data*, and *Ethics and Data Management*. Since new challenges in PDM arise all the time, we note that this list of themes is not intended to be exhaustive.

This report is intended for a diverse audience. It is intended for government and industry funding agencies, because it includes an articulation of important areas where the PDM community is already contributing to the key data management challenges in our era, and has the potential to contribute much more. It is intended for universities and colleges world-wide, because it articulates the importance of continued research and education in the foundational elements of data management, and it highlights growth areas for Computer Science and Management of Information Science research. It is intended for researchers and students, because it identifies emerging, exciting research challenges in the PDM area, all of which have very timely practical relevance. It is also intended for policy makers, sociologists, and philosophers, because it re-iterates the importance of considering ethics in many aspects of data creation, access, and usage, and suggests how research can help to find new ways for maximizing the benefits of massive data while nevertheless safeguarding the privacy and integrity of citizens and societies.

Contents

1   Introduction

In April 2016, a community of researchers working in the area of Principles of Data Management (PDM) joined in the Dagstuhl Castle in Germany for a workshop organized jointly by the Executive Committee of the ACM Symposium on Principles of Database Systems (PODS) and the Council of the International Conference on Database Theory (ICDT). The mission of this workshop was to identify and explore some of the most important research directions that have high relevance to society and to Computer Science today, and where the PDM community has the potential to make significant contributions. This report describes the family of research directions that the workshop focused on from three perspectives: potential practical relevance, results already obtained, and research questions that appear surmountable in the short and medium term. This report organizes the identified research challenges for PDM around seven core themes, namely *Query Processing at Scale*, *Multi-model Data*, *Uncertain Information*, *Knowledge-enriched Data*, *Data Management and Machine Learning*, *Process and Data*, and *Ethics and Data Management*. Since new challenges in PDM arise all the time, we note that this list of themes is not intended to be exhaustive.

This report is intended for a diverse audience. It is intended for government and industry funding agencies, because it includes an articulation of important areas where the PDM community is already contributing to the key data management challenges in our era, and has the potential to contribute much more. It is intended for universities and colleges world-wide, because it articulates the importance of continued research and education in the foundational elements of data management, and it highlights growth areas for Computer Science and Management of Information Science research. It is intended for researchers and students, because it identifies emerging, exciting research challenges in the PDM area, all of which have very timely practical relevance. It is also intended for policy makers, sociologists, and philosophers, because it re-iterates the importance of considering ethics in many aspects of data creation, access, and usage, and suggests how research can help to find new ways for maximizing the benefits of massive data while nevertheless safeguarding the privacy and integrity of citizens and societies.

The field of PDM is broad. It has ranged from the development of formal frameworks for understanding and managing data and knowledge (including data models, query languages, ontologies, and transaction models) to data structures and algorithms (including query optimizations, data exchange mechanisms, and privacy-preserving manipulations). Data management is at the heart of most IT applications today, and will be a driving force in personal life, social life, industry, and research for the foreseeable future. We anticipate on-going expansion of PDM research as the technology and applications involving data management continue to grow and evolve.

PDM played a foundational role in the relational database model, with the robust connection between algebraic and calculus-based query languages, the connection between integrity constraints and database design, key insights for the field of query optimization, and the fundamentals of consistent concurrent transactions. This early work included rich cross-fertilization between PDM and other disciplines in mathematics and computer science, including logic, complexity theory, and knowledge representation. Since the 1990s we have seen an overwhelming increase in both the production of data and the ability to store and access such data. This has led to a phenomenal metamorphosis in the ways that we manage and use data. During this time, we have gone (1) from stand-alone disk-based databases to data that is spread across and linked by the Web, (2) from rigidly structured towards loosely structured data, and (3) from relational data to many different data models (hierarchical,

graph-structured, data points, NoSQL, text data, image data, etc.). Research on PDM has developed during that time, too, following, accompanying and influencing this process. It has intensified research on extensions of the relational model (data exchange, incomplete data, probabilistic data, ...), on other data models (hierachical, semi-structured, graph, text, ...), and on a variety of further data management areas, including knowledge representation and the semantic web, data privacy and security, and data-aware (business) processes. Along the way, the PDM community expanded its cross-fertilization with related areas, to include automata theory, web services, parallel computation, document processing, data structures, scientific workflow, business process management, data-centered dynamic systems, data mining, machine learning, information extraction, etc.

Looking forward, three broad areas of data management stand out where principled, mathematical thinking can bring new approaches and much-needed clarity. The first relates to the full lifecycle of so-called "Big Data Analytics", that is, the application of statistical and machine learning techniques to make sense out of, and derive value from, massive volumes of data. The second stems from new forms of data creation and processing, especially as it arises in applications such as web-based commerce, social media applications, and data-aware workflow and business process management. The third, which is just beginning to emerge, is the development of new principles and approaches in support of ethical data management. We briefly illustrate some of the primary ways that these three areas can be supported by the seven PDM research themes that are explored in this report.

The overall lifecycle of Big Data Analytics raises a wealth of challenge areas that PDM can help with. As documented in numerous sources, so-called "data wrangling" can form 50% to 80% of the labor costs in an analytics investigation. The challenges of data wrangling can be described in terms of the "4 V's" – Volume, Velocity, Variety, and Veracity – all of which have been addressed, and will continue to be addressed, using principled approaches. As we will discuss later, PDM is making new contributions towards managing the Volume and Velocity. As an example, *Query Processing at Scale* (Section 2) talks about recent advances in efficient n-way join processing in highly parallelized systems, which outperform conventional approaches based on a series of binary joins [18, 37]. This section also introduces different paradigms for approximate query processing, sometimes in an *online* or *streaming* setting, in which the user can terminate as long as it is satisfies with the quality of the answer. PDM is contributing towards managing the Variety: *Knowledge-enriched Data* (Section 5) provides tools for managing and efficient reasoning with industrial-sized ontologies [33], and *Multi-model Data* (Section 3) provides approaches for efficient access to diverse styles of data, from tabular to tree to graph to unstructured. Veracity is an especially important challenge when performing analytics over large volumes of data, given the inevitability of inconsistent and incomplete data. The PDM field of *Uncertain Information* (Section 4) has provided a formal explanation of how to answer queries in the face of uncertainty some four decades ago [79], but its computational complexity has made mainstream adoption elusive – a challenge that the PDM community should redouble its efforts to resolve. Provocative new opportunities are raised in the area of *Data Management and Machine Learning* (Section 6), because of the unconventional ways in which feature engineering and machine learning algorithms access and manipulate large data sets. We are also seeing novel approaches to incorporate Machine Learning techniques into database management systems, e.g., to enable more efficient extraction and management of information coming from text [12].

The new forms of data creation and processing that have emerged have led to new forms of data updates, transactions, and data management in general. Web-based commerce has revolutionized how business works with supply chain, financial, manufacturing, and other

kinds of data, and also how businesses engage with their customers, both consumers and other businesses. Social applications have revolutionalized our personal and social lives, and are now impacting the workplace in similar ways. Transactions are increasingly distributed, customized, personalized, offered with more immediacy, and informed by rich sets of data and advanced analytics. These trends are being compounded as the Internet of Things becomes increasingly real and leveraged to increase personal convenience and business efficiencies. A broad challenge is to make it easy to understand all of this data, and the ways that the data are being processed; approaches to this challenge are offered in both *Multi-model Data* (Section 3) and *Knowledge-enriched Data* (Section 5). Many forms of data from the Web, including from social media, from crowd-sourced query answering, and unstructured data in general create *Uncertain Information* (Section 4). Web-based communication has also enabled a revolution in electronically supported processes, ranging from conventional business processes that are now becoming partially automated, to consumer-facing e-commerce systems, to increasingly streamlined commercial and supply chain applications. Approaches have emerged for understanding and managing *Process and Data* (Section 7) in a holistic manner, enabling a new family of automated verification techniques [35]; these will become increasingly important as process automation accelerates.

While ethical use of data has always been a concern, the new generation of data- and information-centric applications, including Big Data Analytics, social applications, and also the increasing use of data in commerce (both business-to-consumer and business-to-business) has made ethical considerations more important and more challenging. At present there are huge volumes of data being collected about individuals, and being interpreted in many different ways by increasing numbers of diverse organizations with widely varying agendas. Emerging research suggests that the use of mathematical principles in research on *Ethics and Data Management* (Section 8) can lead to new approaches to ensure data privacy for individuals, and compliance with government and societal regulations at the corporate level. As just one example, mechanisms are emerging to ensure accurate and "fair" representation of the underlying data when analytic techniques are applied [50].

The findings of this report differ from, and complement, the findings of the 2016 Beckman Report [1] in two main aspects. Both reports stress the importance of "Big Data" as the single largest driving force in data management usage and research in the current era. The current report focuses primarily on research challenges where a mathematically based perspective has had and will continue to have substantial impact. This includes for example new algorithms for large-scale parallelized query processing and Machine Learning, and models and languages for heterogeneous and uncertain information. The current report also considers additional areas where research into the principles of data management can make growing contributions in the coming years, including for example approaches for combining data structured according to different models, process taken together with data, and ethics in data management.

The remainder of this report includes the seven technical sections mentioned above, and a concluding section with comments about the road ahead for PDM research.

2   Query Processing at Scale

Volume is still the most prominent feature of *Big Data*. The PDM community, as well as the general theoretical computer science community, has made significant contributions to efficient query processing at scale (concerning both Volume and Velocity). This is evident

from the tremendous success of parallel algorithms, external memory algorithms, streaming algorithms, etc., with their applications in large-scale database systems. Sometimes, the contributions of theoretical foundations might not be immediate, e.g., it took more than a decade for the *MapReduce* system to popularize Valiant's theoretical *bulk synchronous parallel (BSP)* model [109] in the systems community. But this exactly means that one should never underestimate the value of theory.

Next we review two of the most important practical challenges we face today concerning query processing at scale:

Developing New Paradigms for Multi-way Join Processing. A celebrated result by Atserias, Grohe, and Marx [18] has sparked a flurry of research efforts in re-examining how multi-way joins should be computed. In all current relational database systems, a multi-way join is processed in a pairwise framework using a binary tree (plan), which is chosen by the query optimizer. However, the recent theoretical studies have discovered that for many queries and data instances, even the best binary plan is suboptimal by a large polynomial factor. Meanwhile, worst-case optimal algorithms have been designed in the RAM model [86], the external memory model [65], and BSP models [23, 5]. These new algorithms have all abandoned the binary tree paradigm, while adopting a more *holistic* approach to achieve optimality. Encouragingly, there have been empirical studies [37] that demonstrate the practicality of these new algorithms. In particular, *leapfrog join* [111], a worst-case optimal algorithm, has been implemented inside a full-fledged database system. Therefore, we believe that the newly developed algorithms in the theory community have a potential to change how multi-way join processing is currently done in database systems. Of course, this can only be achieved with significant engineering efforts, especially in designing and implementing new query optimizers and cost estimation under the new paradigm.

Approximate query processing. Most analytical queries on *Big Data* return aggregated answers that do not have to be 100% accurate. The line of work on *online aggregation* [63] studies new algorithms that allow the query processor to return approximate results (with statistical guarantees) at early stages of the processing so that the user can terminate it as soon as the accuracy is acceptable. This both improves interactiveness and reduces unnecessary resource consumption. Recent studies have shown some encouraging results [62, 76], but there is still a lot of room for improvement: (1) The existing algorithms have only used simple random sampling or sample random walks to sample from the full query results. More sophisticated techniques based on Markov Chain Monte Carlo might be more effective. (2) The streaming algorithms community has developed many techniques to summarize large data sets into compact data structures while preserving important properties of the data. These data summarization techniques can be useful in approximate query processing as well. (3) Actually integrating these techniques into modern data processing engines is still a significant practical challenge.

These practical challenges raise the following theoretical challenges:

The Relationship Among Various Big Data Computation Models. The theoretical computer science community has developed many beautiful models of computation aimed at handling data sets that are too large for the traditional random access machine (RAM) model, the most prominent ones including parallel RAM (PRAM), external memory (EM) model, streaming model, the BSP model and its recent refinements to model modern distributed architectures. Several studies seem to suggest that there are deep connections between seemingly unrelated Big Data computation models for streaming computation, parallel processing, and external memory, especially for the class of problems interesting to the PDM community

(e.g., relational algebra) [54, 72]. Investigating this relationship would reveal the inherent nature of these problems with respect to scalable computation, and would also allow us to leverage the rich set of ideas and tools that the theory community has developed over the decades.

The Communication Complexity of Parallel Query Processing.   New large-scale data analytics systems use massive parallelism to support complex queries on large data sets. These systems use clusters of servers and proceed in multiple communication rounds. In these systems, the communication cost is usually the bottleneck, and therefore has become the primary measure of complexity for algorithms designed for these models. Recent studies (e.g., [23]) have established tight upper and lower bounds on the communication cost for computing some join queries, but many questions remain open: (1) The existing bounds are tight only for one-round algorithms. However, new large-scale systems like Spark have greatly improved the efficiency of multi-round iterative computation, thus the one-round limit seems unnecessary. The communication complexity of multi-round computation remains largely open. (2) The existing work has only focused on a small set of queries (full conjunctive queries), while many other types of queries remain unaddressed. Broadly, there is great interest in large-scale machine learning using these systems, thus it is both interesting and important to study the communication complexity of classical machine learning tasks under these models. This is developed in more detail in Section 6, which summarizes research opportunites at the crossroads of data management and machine learning. Large-scale parallel query processing raises many other (practical and foundational) research questions. As an example, recent frameworks for parallel query optimization need to be extended to the multi-round case [10].

We envision that the following theory techniques will be useful in addressing the challenges above (that are not considered as "classical" PDM or database theory): Statistics, sampling theory, approximation theory, communication complexity, information theory, convex optimization.

3   Multi-model Data: Towards an Open Ecosystem of Data Models

Over the past 20 years, the landscape of available data has dramatically changed. While the huge amount of available data is perceived as a clear asset, exploiting this data meets the challenges of the "4 V's" mentioned in the Introduction.

One particular aspect of the *variety* of data is the existence and coexistence of different models for semi-structured and unstructured data, in addition to the widely used relational data model. Examples include tree-structured data (XML, JSON), graph data (RDF, property graphs, networks), tabular data (CSV), temporal and spatial data, text, and multimedia. We can expect that in the near future, new data models will arise in order to cover particular needs. Importantly, data models include not only a data structuring paradigm, but also approaches for queries, updates, integrity constraints, views, integration, and transformation, among others.

Following the success of the relational data model, originating from the close interaction between theory and practice, the PDM community has been working for many years towards understanding each one of the aforementioned models formally. Classical DB topics – schema and query languages, query evaluation and optimization, incremental processing of evolving data, dealing with inconsistency and incompleteness, data integration and exchange, etc. – have been revisited. This line of work has been successful from both the theoretical and

practical points of view. As these questions are not yet fully answered for the existing data models and will be asked again whenever new models arise, it will continue to offer practically relevant theoretical challenges. But what we view as a new grand challenge is the coexistence and interconnection of all these models, complicated further by the need to be prepared to embrace new models at any time.

The coexistence of different data models resembles the fundamental problem of data heterogeneity within the relational model, which arises when semantically related data is organized under different schemas. This problem has been tackled by data integration and data exchange, but since these classical solutions have been proposed, the nature of available data has changed dramatically, making the questions open again. This is particularly evident in the Web scenario, where not only the data comes in huge amounts, in different formats, is distributed, and changes constantly, but also it comes with very little information about its structure and almost no control of the sources. Thus, while the existence and coexistence of various data models is not new, the recent changes in the nature of available data raise a strong need for a new principled approach for dealing with different data models: an approach flexible enough to allow keeping the data in their original format (and be open for new formats), while still providing a convenient unique interface to handle data from different sources. It faces the following four specific practical challenges.

Modelling data. How does one turn raw data into a database? This used to amount to designing the right structure within the relational model. Nowadays, one has to first choose the right data models and design interactions between them. Could we go even further and create methodologies allowing engineers to design a new data model?

Understanding data. How does one make sense of the data? Previously, one could consult the structural information provided with the data. But presently data hardly ever comes with sufficient structural information, and one has to discover its structure. Could we help the user and systems to understand the data without first discovering its structure in full?

Accessing data. How does one extract information? For years this meant writing an SQL query. Currently the plethora of query languages is perplexing and each emerging data model brings new ones. How can we help users formulate queries in a more uniform way?

Processing data. How does one evaluate queries efficiently? Decades of effort brought refined methods to speed up processing of relational data; achieving similar efficiency for other data models, even the most mature ones such as XML, is still a challenge. But it is time to start thinking about processing data combining multiple models (possibly distributed and incomplete).

These practical challenges raise concrete theoretical problems, some of which go beyond the traditional scope of PDM. Within PDM, the key theoretical challenges are the following.

Schema languages. Design flexible and robust multi-model schema languages. Schema languages for XML and RDF data are standardized, efforts are being made to create standards for JSON [90], general graph data [100], and tabular data [82, 16]. Multi-model schema languages should offer a uniform treatment of different models, the ability to describe mappings between models (implementing different views on the same data, in the spirit of data integration), and the flexibility to seamlessly incorporate new models as they emerge.

Schema extraction. Provide efficient algorithms to extract schemas from the data, or at least discover partial structural information (cf. [27, 31]). The long-standing challenge of entity resolution is exacerbated in the context of finding correspondences between data sets structured according to different models [107].

Visualization of data and metadata.   Develop user-friendly paradigms for presenting the metadata information and statistical properties of the data in a way that helps in formulating queries. In an ideal solution, users would be presented relevant information about data and metadata as they type the query. This requires understanding and defining what the relevant information in a given context is, and representing it in a way allowing efficient updates as the context changes (cf. [36, 15]).

Query languages.   Go beyond bespoke query languages for the specific data models [14] and design a query language suitable for multi-model data, either incorporating the specialized query languages as sub-languages or offering a uniform approach to querying, possibly at the cost of reduced expressive power or higher complexity.

Evaluation and Optimization.   Provide efficient algorithms for computing meaningful answers to a query, based on structural information about data, both inter-model and intra-model; this can be tackled either directly [70, 58] or via static optimization [24, 40]. In the context of distributed or incomplete information, even formalizing the notion of a meaningful answer is a challenge [78], as discussed in more detail in Section 4.

All these problems require strong tools from PDM and theoretical computer science in general (complexity, logic, automata, etc.). But solving them will also involve knowledge and techniques from neighboring communities. For example, the second, third and fifth challenges naturally involve data mining and machine learning aspects (see Section 6). The first, second, and third raise knowledge representation issues (see Section 5). The first and fourth will require expertise in programming languages. The fifth is at the interface between PDM and algorithms, but also between PDM and systems. The third raises human-computer interaction issues.

4   Uncertain Information

Incomplete, uncertain, and inconsistent information is ubiquitous in data management applications. This was recognized already in the 1970s [39], and since then the significance of the issues related to incompleteness and 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, both theoretically and practically. 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"* [41]. In fact we know that SQL can produce every type of error imaginable when nulls are present [77].

On the theory side, we appear to have a good understanding of what is needed in order to produce correct results: computing *certain answers* to queries. These are answers that are true in all complete databases that are compatible with the given incomplete database. This idea, that dates back to the late 1970s as well, has become *the* way of providing query answers in all applications, from classical databases with incomplete information [67] to new applications such as data integration and exchange [74, 13], consistent query answering [26], ontology-based data access [33], and others. The reason these ideas have found limited application in mainstream database systems is their complexity. Typically, answering queries over incomplete databases with certainty can be done efficiently for conjunctive queries or some closely related classes, but beyond the complexity quickly grows to intractable

(sometimes even undecidable). Since this cannot be tolerated by real life systems, they resort to ad hoc solutions, which go for efficiency and sacrifice correctness; thus bizarre and unexpected behavior occurs.

While even basic problems related to incompleteness in relational databases remain unsolved, we now constantly deal with more varied types of incomplete and inconsistent data. A prominent example is that of probabilistic databases [103], where the confidence in a query answer is the total weight of the worlds that support the answer. Just like certain answers, computing exact answer probabilities is usually intractable, and yet it has been the focus of theoretical research.

The key challenge in addressing the problem of handling incomplete and uncertain data is to provide theoretical solutions that are *usable in practice*. Instead of proving more impossibility results, the field should urgently address what can actually be done efficiently.

Making theoretical results applicable in practice is the biggest practical challenge for incomplete and uncertain data. To move away from the focus on intractability and to produce results of practical relevance, the PDM community needs to address several challenges.

RDBMS technology in the presence of incomplete data. It must be capable of finding query answers one can trust, and do so efficiently. But how do we find good quality query answers with correctness guarantees when we have theoretical intractability? For this we need new approximation schemes, quite different from those that have traditionally been used in the database field. Such schemes should provide guarantees that answers can be trusted, and should also be implementable using existing RDBMS technology.

To make these scheme truly efficient, we need to address the issue of the performance of commercial RDBMS technology in the presence of incomplete data. Even query optimization in this case is hardly a solved problem; in fact commercial optimizers often do not perform well in the presence of nulls.

Models of uncertainty. What is provided by current practical solutions is rather limited. Looking at relational databases, we know that they try to model everything with primitive null values, but this is clearly insufficient. We need to understand types of uncertainty that need to be modeled and introduce appropriate representation mechanisms.

This, of course, will lead to a host of new challenges. How do we store/represent richer kinds of uncertain information, that go well beyond nulls in RDBMSs? Applications such as integration, exchange, ontology-based data access and others often need more (at the very least, marked nulls), and one can imagine many other possibilities (e.g., intervals for numerical values). This is closely related to the modelling data task described in Section 3.

Benchmarks for uncertain data. 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, making it hard to test efficiency of solutions in practice.

Handling inconsistent data. How do we make handling inconsistency (in particular, consistent query answering) work in practice? How do we use it in data cleaning? Again, there are many strong theoretical results here, but they concentrate primarily on tractability boundaries and various complexity dichotomies for subclasses of conjunctive queries, rather than practicality of query answering techniques. There are promising works on enriching theoretical *repairs* with user preferences [101], or ontologies [51], along the lines of approaches described in Section 5, but much more foundational work needs to be done before they can get to the level of practical tools.

Handling probabilistic data.   The common models of probabilistic databases are arguably simpler and more restricted than the models studied by the Statistics and Machine Learning communities. Yet common complex models can be simulated by probabilistic databases if one can support expressive query languages [69]; hence, model complexity can be exchanged for query complexity. Therefore, it is of great importance to develop techniques for approximate query answering, on expressive query languages, over large volumes of data, with practical execution costs. While the focus of the PDM community has been on deterministic and exact solutions [103], we believe that more attention should be paid to statistical techniques with approximation guarantees such as the sampling approach typically used by the (Bayesian) Machine Learning and Statistics communities. In Section 6 we further discuss the computational challenges of Machine Learning in the context of databases.

The theoretical challenges can be split into three groups.

Modeling.   We need to provide a solid theoretic basis for the practical modeling challenge above; this means understanding different types of uncertainty and their representations. As with any type of information stored in databases, there are lots of questions for the PDM community to work on, related to data structures, indexing techniques, and so on.

There are other challenges related to modeling data. For instance, when can we say that some data is true? This issue is particularly relevant in crowdsourcing applications [95, 61]: having data that looks complete does not yet mean it is true, as is often assumed.

Yet another important issue addresses modeling query answers. 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.

The Programming Languages and Machine Learning communities have been investigating *probabilistic programming* [56] as a paradigm for allowing developers to easily program Machine Learning solutions. The Database community has been leading the development of paradigms for easy programming over large data volumes. As discussed in detail later in Section 6, we believe that modern needs require the enhancement of the database technology with machine learning capabilities. In particular, an important challenge is to combine the two key capabilities (machine learning and data) via query languages for building statistical models, as already began by initial efforts [21, 32].

Reasoning.   There is much work on this subject; see Section 5 concerning the need to develop next-generation reasoning tools for data management tasks. When it comes to using such tools with incomplete and uncertain data, the key challenges are: How do we do inference with incomplete data? How do we integrate different types of uncertainty? How do we learn queries on uncertain data? 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.

Algorithms.   To overcome high complexity, we often need to resort to approximate algorithms, but approximation techniques are different from the standard ones used in databases, as they do not just speed up evaluation but rather ensure correctness. The need for such approximations leads to a host of theoretical challenges. How do we devise such algorithms? How do we express correctness in relational data and beyond? How do we measure the quality of query answers? How do we take user preferences into account?

While all the above are important research topics that need to be addressed, there are several that can be viewed as a priority, not least because there is an immediate connection between theory and practice. In particular, we need to pay close attention to the following

issues: (1) understand what it means for answers to be right or wrong, and how to adjust the standard relational technology to ensure that wrong answers are never returned to the user; (2) provide, and justify, benchmarks for working with incomplete/uncertain data; (3) devise approximation algorithms for classes of queries known to be intractable; and (4) make an effort to achieve practicality of consistent query answering, and to apply it in data cleaning scenarios.

It is worth remarking that questions about uncertain data are often considered in the context of data cleaning, under the assumption that uncertainty is caused by dirty data. The focus of data cleaning is then on eliminating uncertainty, much less on querying data that we are not completely sure about. The latter however cannot be dismissed because data cleaning techniques do not always allow us to deal with uncertain data. Indeed, the fact that data is unclean is only sometimes – but by no means always – the cause of uncertainty, and the field of uncertain data covers many scenarios that data cleaning is not handling. These include the treatment of nulls in databases (which are not always due to dirty data), and probabilistic data, where uncertainty is due to the nature of data rather than it being dirty. The closest subject to data cleaning we cover here is consistent query answering, but even then the focus is different, as one tries to see what can be meaningfully extracted from data if it cannot be fully cleaned.

5 Knowledge-enriched Data Management

Over the past two decades we have witnessed a gradual shift from a world where most data used by companies and organizations was regularly structured, neatly organized in relational databases, and treated as complete, to a world where data is heterogenous and distributed, and can no longer be treated as complete. Moreover, not only do we have massive amounts of data; we also have very large amounts of rich knowledge about the application domain of the data, in the form of taxonomies or full-fledged ontologies, and rules about how the data should be interpreted, among other things. Techniques and tools for managing such complex information have been studied extensively in Knowledge Representation, a subarea of Artificial Intelligence. In particular logic-based formalisms, such as description logics and different rule-based languages, have been proposed and associated reasoning mechanisms have been developed. However, work in this area did not put a strong emphasis on the traditional challenges of data management, namely huge volumes of data, and the need to specify and perform complex operations on the data efficiently, including both queries and updates.

Both practical and theoretical challenges arise when rich domain-specific knowledge is combined with large amounts of data and the traditional data management requirements, and the techniques and approaches coming from the PDM community will provide important tools to address them. We discuss first the practical challenges.

Providing end users with flexible and integrated access to data. A key requirement in dealing with complex, distributed, and heterogeneous data is to give end users the ability to directly manage such data. This is a challenge since end users might have deep expertise about a specific domain of interest, but in general are not data management experts. As a result, they are not familiar with traditional database techniques and technologies, such as the ability to formulate complex queries or update operations, possibly accessing multiple data sources over which the data might be distributed, and to understand performance implications. Ontology-based data management has been proposed recently as a general paradigm to

address this challenge. It is based on the assumption that a domain ontology capturing complex knowledge can be used for data management by linking it to data sources using declarative mappings [91]. Then, all information needs and data management requirements by end users are formulated in terms of such ontology, instead of the data sources, and are automatically translated into operations (queries and updates) over the data sources. Open challenges are related to the need of dealing with distribution of data, of handling heterogeneity at both the intensional and extensional levels, of performing updates to the data sources via the ontology and the mappings, and in general of achieving good performance even in the presence of large ontologies, complex mappings, and huge amounts of data [33, 57, 59].

Ensuring interoperability at the level of systems exchanging data. Enriching data with knowledge is not only relevant for providing end-user access, but also enables direct inter-operation between systems, based on the exchange of data and knowledge at the system level. A requirement is the definition of and agreement on standardized ontologies covering all necessary aspects of specific domains of interest, including multiple modalities such as time and space. A specific area where this is starting to play an important role is e-commerce, where standard ontologies are already available [64].

Personalized and context-aware data access and management. Information is increasingly individualized and only fragments of the available data and knowledge might be relevant in specific situations or for specific users. It is widely acknowledged that it is necessary to provide mechanisms on the one hand for characterizing contexts (as a function of time, location, involved users, etc.), and on the other hand for defining which fragments of data and/or knowledge should be made available to users, and how such data needs to be pre-processed/filtered/modified, depending on the actual context and the knowledge available in that context. The problem is further complicated by the fact that both data and knowledge, and also contextual information, might be highly dynamic, changing while a system evolves. Heterogeneity needs to be dealt with, both with respect to the modeling formalism and with respect to the modeling structures chosen to capture a specific real-world phenomenon.

Bringing knowledge to data analytics and data extraction. Increasing amounts of data are being collected to perform complex analysis and predictions. Currently, such operations are mostly based on data in "raw" form, but there is a huge potential for increasing their effectiveness by enriching and complementing such data with domain knowledge, and leveraging this knowledge during the data analytics and extraction process. Challenges include choosing the proper formalisms for expressing knowledge about both raw and aggregated/derived data, developing knowledge-aware algorithms for data extraction and analytics, in particular for overcoming low data quality, and dealing with exceptions and outliers.

Making the management user friendly. Systems combining large amounts of data with complex knowledge are themselves very complex, and thus difficult to design and maintain. Appropriate tools that support all phases of the life-cycle of such systems need to be designed and developed, based on novel user interfaces for the various components. Such tools should themselves rely on the domain knowledge and the sophisticated inference services over such knowledge to improve user interaction, in particular for domain experts as opposed to IT or data management experts. Supported tasks should include design and maintenance of ontologies and mappings (including debugging support), query formulation, explanation of inference, and data and knowledge exploration [55, 73, 48, 15].

To provide adequate solutions to the above practical challenges, several key theoretical challenges need to be addressed, requiring a blend of formal techniques and tools traditionally studied in data management, with those typically adopted in knowledge representation in AI.

Development of reasoning-tuned DB systems.   Such systems will require new/improved database engines optimized for reasoning over large amounts of data and knowledge, able to compute both crisp and approximate answers, and to perform distributed reasoning and query evaluation. To tune such systems towards acceptable performance, new cost models need to be defined, and new optimizations based on such cost models need to be developed.

Choosing/designing the right languages.   The languages and formalisms adopted in the various components of knowledge-enriched data management systems have to support different types of knowledge and data, e.g., mixing open and closed world assumption, and allowing for representing temporal, spatial, and other modalities of information [34, 19, 29, 17, 87]. It is well understood that the requirements in terms of expressive power for such languages would lead to formalisms that make the various inference tasks either undecidable or highly intractable. Therefore, the choice or design of the right languages have to be pragmatically guided by user and application needs.

New measures of complexity.   To appropriately assess the performance of such systems and be able to distinguish easy cases that seem to work well in practice from difficult ones, alternative complexity measures are required that go beyond the traditional worst-case complexity. These might include suitable forms of average case or parameterized complexity, complexity taking into account data distribution (on the Web), and forms of smoothed analysis.

Next-generation reasoning services.   The kinds of reasoning services that become necessary in the context of knowledge-enriched data management applications go well beyond traditional reasoning studied in knowledge representation, which typically consists of consistency checking, classification, and retrieval of class instances. The forms of reasoning that are required include processing of complex forms of queries in the presence of knowledge, explanation (which can be considered as a generalization of provenance), abductive reasoning, hypothetical reasoning, inconsistency-tolerant reasoning, and defeasible reasoning to deal with exceptions. Forms of reasoning with uncertain data, such as probabilistic or fuzzy data and knowledge will be of particular relevance, as well as meta-level reasoning. Further, it will be necessary to develop novel forms of reasoning that are able to take into account non-functional requirements, notably various measures for the quality of data (completeness, reliability, consistency), and techniques for improving data quality. While such forms of reasoning have already begun to be explored individually (see, e.g., [52, 30], much work remains to bring them together, to incorporate them into data-management systems, and to achieve the necessary level of performance.

Incorporating temporal and dynamic aspects.   A key challenge is represented by the fact that data and knowledge is not static, and changes over time, e.g., due to updates on the data while taking into account knowledge, forms of streaming data, and more in general data manipulated by processes. Dealing with dynamicity and providing forms of inference (e.g., formal verification) in the presence of both data and knowledge is extremely challenging and will require the development of novel techniques and tools [35, 17].

In summary, incorporating domain-specific knowledge to data management is both a great opportunity and a major challenge. It opens up huge possibilities for making data-centric systems more intelligent, flexible, and reliable, but entails computational and technical

challenges that need to be overcome. We believe that much can be achieved in the coming years. Indeed, the increasing interaction of the PDM and the Knowledge Representation communities has been very fruitful, particularly by attempting to understand the similarities and differences between the formalisms and techniques used in both areas, and obtaining new results building on mutual insights. Further bridging this gap by the close collaboration of both areas appears as the most promising way of fulfilling the promises of Knowledge-enriched Data Management.

6   Data Management and Machine Learning

We believe that research that combines Data Management (DM) and Machine Learning (ML) is especially important, because these fields can mutually benefit from each other. Nowadays, systems that emerge from the ML community are strong in their capabilities of statistical reasoning, and systems that emerge from the DM community are strong in their support for data semantics, maintenance and scale. This complementarity in assets is accompanied by a difference in the core mechanisms: the PDM community has largely adopted logic-based methodologies, while the ML community centralized around probability theory and statistics. Yet, modern applications require systems that are strong in *both* aspects, providing a thorough and sophisticated management of data while incorporating its inherent statistical nature. We envision a plethora of research opportunities in the intersection of PDM and ML. We outline several directions, which we classify into two categories: *DM for ML* and *ML for DM*.

The category *DM for ML* includes directions that are aimed at the enhancement of ML capabilities by exploiting properties of the data. Key challenges are as follows.

Feature Generation and Engineering. Feature engineering refers to the challenge of designing and extracting signals to provide to the general-purpose ML algorithm at hand, in order to properly perform the desired operation (e.g., classification or regression). This is a critical and time-consuming task [71], and a central theme of modern ML methodologies, such as kernel-based ML, where complex features are produced implicitly via kernel functions [97], and deep learning, where low-level features are combined into higher-level features in a hierarchical manner [25]. Unlike usual ML algorithms that view features as numerical values, the database has access to, and understanding of, the *queries* that transform raw data into these features. Thus, PDM can contribute to feature engineering in various ways, especially on a semantic level, and provide solutions to problems such as the following: How to develop effective languages for query-based feature creation? How to use such languages for designing a set of complementary, non-redundant features optimally suited for the ML task at hand? Is a given language suitable for a certain class of ML tasks? Important criteria for the goodness of a feature language include the risks of *underfitting* and *overfitting* the training data, as well as the computational complexity of evaluation (on both training and test data). The PDM community has already studied problems of a similar nature [60].

The premise of deep (neural network) learning is that the model has sufficient expressive power to work with only *raw, low-level features*, and to realize the process of high-level feature generation in an automated, data-driven manner [25]. This brings a substantial hope for reducing the effort in manual feature engineering. Is there a general way of solving ML tasks by applying deep learning directly to the database (as has already been done, for example, with *semantic hashing* [94])? Can database queries (of different languages) complement neural networks by means of expressiveness and/or efficiency? And if so, where lies the boundary between the level of feature engineering and the complexity of the network?

Large-Scale Machine Learning. Machine learning is nowadays applied to massive data sets of considerable size, including potentially unbounded streams of data. Under such conditions, an effective data management and the use of appropriate data structures that offer the learning algorithm fast access to the data are major prerequisites for realizing model induction (at training time) and inference (at prediction time) in a time-efficient and space-efficient manner [92]. Research along this direction has amplified in recent years and includes, for example, the use of hashing [112], Bloom filters [38], and tree-based data structures [45] in learning algorithms. As another example, lossless compression of large datasets, as featured by *factorized databases* [89], have been shown to dramatically reduce the execution cost of machine-learning tasks. Also related is work on distributed machine learning, where data storage and computation is accomplished in a network of distributed units [6], and the support of machine learning by data stream management systems [84].

Complexity Analysis. The PDM community has established a strong machinery for fine-grained analysis of querying complexity; see, e.g., [9]. Complexity analysis of such granularity is highly desirable for the ML community, especially for analyzing learning algorithms that involve various parameters like I/O dimension, and number of training examples [68]. Results along this direction, connecting DM querying complexity and ML training complexity, have been recently shown [96].

The motivation for the directions in the second category, *ML for DM*, is that of strengthening core data-management capabilities with ML. Traditionally, data management systems have supported a core set of querying operators (e.g., relational algebra, grouping and aggregate functions, recursion) that are considered as the common requirement of applications. We believe that this core set should be revisited, and specifically that it should be extended with common ML operators.

As a prominent example, motivated by the proliferation of available and valuable textual resources, various formalisms have been proposed for incorporating text extraction in a relational model [53, 98]. However, unlike structured data, textual resources are associated with a high level of uncertainty due to the uncontrolled nature of the content and the imprecise nature of natural language processing. Therefore, ML techniques are required to distill reliable information from text.

We believe that incorporating ML is a natural evolution for PDM. Database systems that incorporate statistics and ML have already been developed [99, 12]. Query languages have traditionally been designed with emphasis on being *declarative*: a query states how the answer should logically relate to the database, not how it is to be computed algorithmically. Incorporating ML introduces a higher level of declarativity, where one states how the end result should behave (via examples), but not necessarily which query is deployed for the task. In that spirit, we propose the following directions for relevant PDM research.

Unified Models. An important role of the PDM community is in establishing common formalisms and semantics for the database community. It is therefore an important opportunity to establish the "relational algebra" of data management systems with built-in ML/statistics operators.

Lossy Optimization. From the early days, the focus of the PDM community has been on *lossless* optimization, that is, optimization that leaves the end result intact. As mentioned in Section 2, in some scenarios it makes sense to apply *lossy* optimization that guarantees only an approximation of the true answer. Incorporating ML into the query model gives further opportunities for lossy optimization, as training paradigms are typically associated with built-in quality (or "risk") functions. Hence, we may consider reducing the execution cost if

it entails a bounded impact on the quality of the end result [8]. For example, Riondato et al. [93] develop a method for random sampling of a database for estimating the selectivity of a query. Given a class of queries, the execution of any query in that class on the sample provides an accurate estimate for the selectivity of the query on the original large database.

Confidence Estimation.   Once statistical and ML components are incorporated in a data management system, it becomes crucial to properly estimate the *confidence* in query answers [99], as such a confidence offers a principled way of controlling the balance between precision and recall. It is then an important direction to establish probabilistic models that capture the combined process and allow to estimate probabilities of end results. For example, by applying the notion of the Vapnik-Chervonenkis dimension, an important theoretical concept in generalization theory, to database queries, Riondato et al. [93] provide accurate bounds for their selectivity estimates that hold with high probability; moreover, they show the error probability to hold simultaneously for the selectivity estimates of all queries in the query class. In general, this direction can leverage the past decade of research on probabilistic databases [104] which can be combined with theoretical frameworks of machine learning, such as PAC (Probably Approximately Correct) learning [110].

Altogether, we have a plethora of research problems, on improving machine learning with data management techniques (DM for ML), and on strengthening data management technologies with capabilities of machine learning (ML for DM). The required methodologies and formal foundations span a variety of related fields such as logic, formal languages, computational complexity, statistical analysis, and distributed computing. We phrased the directions as theoretically oriented; but obviously, each of them is coming with the practical challenge of devising effective solutions over real systems, and on real-life datasets and benchmarks.

7   Process and Data

Many forms of data evolve over time, and most processes access and modify data sets. Industry works with massive volumes of evolving data, primarily in the form of transactional systems and Business Process Management (BPM) systems. Research into basic questions about systems that combine process and data has been growing over the past decade, including the development of several formal models, frameworks for comparing their expressive power, approaches to support verification of behavioral properties, and query languages for process schemas and instances.

Over the past half century, computer science research has studied foundational issues of process and of data mainly as separated phenomena.

In recent years, data and process have been studied together in two significant areas: scientific workflows and data-aware BPM [66]. Scientific workflows focus on enabling repeatability and reliability of processing flows involving large sets of scientific data. In the 1990's and the first decade of the 2000s, foundational research in this area helped to establish the basic frameworks for supporting these workflows, to enable the systematic recording and use of provenance information, and to support systems for exploration that involve multiple runs of a workflow with varying configurations [43]. The work on scientific workflows can also play a role in enabling process support for big data analytics, especially as industry begins to create analytics flows that can be repeated, with relatively minor variation, across multiple applications and clients.

Foundational work on data-aware BPM was launched in the mid-00's [28, 47], enabled in part by IBM's "Business Artifacts" model for business process [88], that combines data and process in a holistic manner. Deutch and Milo [46] provide a survey and comparison of several of the most important early models and results on process and data. One variant of the business artifact model, which is formally defined around logic rather than Petri-nets, has provided the conceptual basis for the recent OMG Case Management Model and Notation standard [81]. Importantly, the artifact-based perspective has formed the basis for a vibrant body of work centered around verification of systems that support processes involving large-scale data [35, 47]. The artifact-based perspective is also beginning to enable a more unified management of the interaction of business processes and legacy data systems [105]. Importantly, there is strong overlap between the artifact-based approach and core building blocks of the "shared ledger" approach to supporting business (and individual) interactions around the exchange of goods and services, as embodied initially by the Blockchain paradigm of Bitcoin [108].

Foundational work in the area of process and data has the potential for continued and expanded impact in the following six practical challenge areas.

Automating manual processes. Most business processes still rely on substantial manual effort. In the case of "back-office" processing, Enterprise Resource Planning systems such as SAP automatically perform the bulk of the work, e.g., for applications in finance and human resource management. But there are still surprisingly many "ancillary processes" that are performed manually, e.g., to process new bank accounts or newly hired employees. In contrast, business processes that involve substantial human judgement, such as complex sales activities or the transition of IT services from one provider to another, are handled today in largely *ad hoc* and manual ways, with spreadsheets as the workflow management tool of choice.

Evolution and migration of Business Processes. Managing change of business processes remains largely manual, highly expensive, time consuming, and risk-prone. This includes deployment of new business process platforms, evolution of business processes, and integration of business processes after mergers.

Business Process compliance and correctness. Compliance with government regulations and corporate policies is a rapidly growing challenge, e.g., as governments attempt to enforce policies around financial stability and data privacy. Ensuring compliance is largely manual today, and involves understanding how regulations can impact or define portions of business processes, and then verifying that process executions will comply.

Business Process interaction and interoperation. Managing business processes that flow across enterprise boundaries has become increasingly important with globalization of business and the splintering of business activities across numerous companies. While routine services such as banking money transfer are largely automated, most interactions between businesses are less standardized and require substantial manual effort to set up, maintain, and troubleshoot. The recent industrial interest in shared ledger technologies highlights the importance of this area and provides new motivation for developing foundational results for data-aware processes.

Business Process discovery and understanding. The field of Business Intelligence, which provides techniques for mining and analyzing information about business operations, is essential to business success. Today this field is based on a broad variety of largely *ad hoc* and manual techniques [44], with associated costs and potential for error. One important direction

on understanding processes focuses on viewing process schemas and process instances as data, and enabling declarative query languages against them [20]. More broadly, techniques from Multi-model Data Management (Section 3), Data Management and Machine Learning (Section 6), and Uncertain Data (Section 4) are all relevant here because of (respectively) the heterogeneity of data about and produced by processes, the importance of anticipating undesirable outcomes and mitigating, and the fact that the information stored about processes is often incomplete.

Workflow and Business Process usability.   The operations of medium- and large-sized enterprises are highly complex, a situation enabled in part by the power of computers to manage huge volumes of data, transactions, and processing all at tremendously high speeds. This raises questions relating to Managing Data at Scale (Section 2). Furthermore, enabling humans to understand and work effectively to manage large numbers of processes remains elusive, especially when considering the interactions between process, data (both newly created and legacy), resources, the workforce, and business partners.

The above practical BPM challenges raise key research challenges that need to be addressed using approaches that include mathematical and algorithmic frameworks and tools.

Verification and Static Analysis.   Because of the infinite state space inherent in data-aware processes [35, 47], verification currently relies on faithful abstractions reducing the problem to classical finite-state model checking. However, the work to date can only handle restricted classes of applications, and research is needed to develop more powerful abstractions enabling a variety of static analysis tasks for realistic data-aware processes. Incremental verification techniques are needed, as well as techniques that enable modular styles of verification that support "plug and play" approaches. This research will be relevant to the first four practical challenges.

Tools for Design and Synthesis.   Formal languages (e.g., context-free) had a profound impact on compiler theory and programming languages. Dependency theory and normal forms had a profound impact on relational database design. But there is still no robust framework that supports principled design of business processes in the larger context of data, resources, and workforce. Primitive operators for creating and modifying data-aware process schemas will be an important starting point; the ultimate goal is partial or full synthesis of process from requirements, goals, and/or regulations. This research will be relevant to the first, second, fourth, and sixth practical challenges.

Models and semantics for views, interaction, and interoperation.   The robust understanding of database views has enabled advances in simplification of data access, data sharing, exchange, integration, and privacy, as well as query optimization. A robust theory of views for data-aware business processes has similar potential. For example, it could support a next generation of data-aware service composition techniques that includes practical verification capabilities. Frameworks that enable comparison of process models (e.g., [3]) can provide an important starting point for this research. This research will be relevant to all of the practical challenges.

Analytics for Business Processes.   The new, more holistic perspective of data-aware processes can help to provide a new foundation for the field of business intelligence. This can include new approaches for instrumenting processes to simplify data discovery [80], and new styles of modularity and hierarchy in both the processes and the analytics on them.

Research in process and data will require on-going extensions of the traditional approaches, on both the database and process-centric sides. New approaches may include models for the creation and maintenance of interoperations between (enterprise-run) services; semi-structured and unstructured forms of data-aware business process (cf. noSQL); new abstractions to enable verification over infinite-state systems; and new ways to apply machine learning. More broadly, a new foundational model for modern BPM may emerge, which builds on the artifact and shared-ledger approaches but facilitates a multi-perspective understanding, analogous to the way relational algebra and calculus provide two perspectives on data querying.

One cautionary note is that research in the area of process and data today is hampered by a lack of large sets of examples, e.g., sets of process schemas that include explicit specifications concerning data, and process histories that include how data sets were used and affected. More broadly, increased collaboration between PDM researchers, applied BPM researchers, and businesses would enable more rapid progress towards resolving the concrete problems in BPM faced by industry today.

8   Human-Related Data and Ethics

More and more "human-related" data is massively generated, in particular on the Web and in phone apps. Massive data analysis, using data parallelism and machine learning techniques, is applied to this data to generate more data. We, individually and collectively, are losing control over this data. We do not know the answers to questions as important as: Is my medical data really available so that I get proper treatment? Is it properly protected? Can a private company like Google or Facebook influence the outcome of national elections? Should I trust the statistics I find on the Web about the crime rate in my neighborhood?

Although we keep eagerly consuming and enjoying more new Web services and phone apps, we have growing concerns about criminal behavior on the Web, including racist, terrorist, and pedophile sites; identity theft; cyber-bullying; and cyber crime. We are also feeling growing resentment against intrusive government practices such as massive e-surveillance even in democratic countries, and against aggressive company behaviors such as invasive marketing, unexpected personalization, and cryptic or discriminatory business decisions.

Societal impact of big data technologies is receiving significant attention in the popular press [11], and is under active investigation by policy makers [85] and legal scholars [22]. It is broadly recognized that this technology has the potential to improve people's lives, accelerate scientific discovery and innovation, and bring about positive societal change. It is also clear that the same technology can in effect limit business faithfulness to legal and ethical norms. And while many of the issues are political and economical, technology solutions must play an important role in enabling our society to reap ever-greater benefits from big data, while keeping it safe from the risks.

We believe that the main inspiration for the data management field in the 21st century comes from the management of human-related data, with an emphasis on solutions that satisfy ethical requirements.

In the remainder of this section, we will present several facets of ethical data management.

Responsible Data Analysis.   Human-related data analysis needs to be "responsible" – to be guided by humanistic considerations and not simply by performance or by the quest for profit. The notion of responsible data analysis is considered generally in [102] and was the subject of a recent Dagstuhl seminar [4]. We now outline several important aspects of the problem, especially those where we see opportunities for involvement by PDM.

Fairness. Responsible data analysis requires that both the raw data and the computation be "fair", i.e. not biased [50]. There is currently no consensus as to which classes of fairness measures, and which specific formulations, are appropriate for various data analysis tasks. Work is needed to formalize the measures and understand the relationships between them.

Transparency and accountability. Responsible data analysis practices must be transparent [42, 106], allowing a variety of stakeholders, such as end-users, commercial competitors, policy makers, and the public, to scrutinize the data collection and analysis processes, and to interpret the outcomes. Interesting research challenges that can be tackled by PDM include using provenance to shed light on data collection and analysis practices, supporting semantic interrogation of data analysis methods and pipelines, and providing explanations in various contexts, including knowledge-based systems and deep learning.

Diversity. Big data technology poses significant risks to those it overlooks [75]. Diversity [7, 49] requires that not all attention be devoted to a limited set of objects, actors or needs. The PDM community can contribute, for instance, to understanding the connections between diversity and fairness, and to develop methods to manage trade-offs between diversity and conventional measures of accuracy.

Verifying Data Responsibility.   A grand challenge for the community is to develop verification technology to enable a new era of responsible data. One can envision research towards developing tools to help users understand data analysis results (e.g., on the Web), and to verify them. One can also envision tools that help analysts, who are typically not computer scientists nor experts in statistics, to realize responsible data analysis "by design".

Data Quality and Access Control on the Web.   The evaluation of data quality on the Web is an issue of paramount importance when our lives are increasingly guided and determined by data found on the Web. We would like to know whether we can trust particular data we found. Research is needed towards supporting access control on the Web. It may build for instance on cryptography, blockchain technology, or distributed access control [83].

Personal Information Management Systems.   A Personal Information Management System is a (cloud) system that manages all the information of a person. By returning part of the data control to the person, these systems tend to better protect privacy, re-balance the relationship between a person and the major internet companies in favor of the person, and in general facilitate the protection of ethical values [2].

Ethical data management raises new issues for computer science in general and for data management in particular. Because the data of interest is typically human-related, the research also includes aspects from other sciences, notably, cognitive science, psychology, neuroscience, linguistics, sociology, and political sciences. The ethics component also leads to philosophical considerations. In this setting, researchers have a chance for major societal impact, and so they need to interact with policy makers and regulators, as well as with the media and user organizations.

9   Looking Forward

As illustrated in the preceding sections, the principled, mathematically-based approach to the study of data management problems is providing conceptual foundations, deep insights, and much-needed clarity. This report describes a representative, but by no means exhaustive, family of areas where research on the Principles of Data Management (PDM) can help to

shape our overall approach to working with data as it arises across an increasingly broad array of application areas.

The Dagstuhl workshop highlighted two important trends that have been accelerating in the PDM community over the past several years. The first is the increasing embrace of neighboring disciplines, including especially Machine Learning, Statistics, Probability, and Verification, both to help resolve new challenges, and to bring new perspectives to them. The second is the increased focus on obtaining positive results, that enable the use of mathematically-based insights in practical settings. We expect and encourage these trends to continue in the coming years.

The PDM community should also continue reinforcing a mutually beneficial relationship with the Data Management Systems community. Our joint conferences (SIGMOD/PODS and EDBT/ICDT) put us in a unique situation in Computer Science where foundational and systems researchers can get in touch and present their best work to each other. PDM researchers should redouble its efforts to actively search for important problems that need a principled approach. Likewise, the organisers of the respective conferences should continue to develop a forum that stimulates the interaction between foundational and systems research.

The need for precise and robust approaches for increasingly varied forms of data management continues to intensify, given the fundamental and transformational role of data in our modern society, and given the continued expansion of technical, conceptual, and ethical data management challenges. There is an associated and on-going expansion in the family of approaches and techniques that will be relevant to PDM research. The centrality of data management across numerous application areas is an opportunity both for PDM researchers to embrace techniques and perspectives from adjoining research areas, and for researchers from other areas to incorporate techniques and perspectives from PDM. Indeed, we hope that this report can substantially strengthen cross-disciplinary research between the PDM and neighboring theoretical communities and, moreover, the applied and systems research communities across the many application areas that rely on data in one form or another.

References

1 Daniel Abadi, Rakesh Agrawal, Anastasia Ailamaki, Magdalena Balazinska, Philip A. Bernstein, Michael J. Carey, Surajit Chaudhuri, Jeffrey Dean, AnHai Doan, Michael J. Franklin, Johannes Gehrke, Laura M. Haas, Alon Y. Halevy, Joseph M. Hellerstein, Yannis E. Ioannidis, H. V. Jagadish, Donald Kossmann, Samuel Madden, Sharad Mehrotra, Tova Milo, Jeffrey F. Naughton, Raghu Ramakrishnan, Volker Markl, Christopher Olston, Beng Chin Ooi, Christopher Ré, Dan Suciu, Michael Stonebraker, Todd Walter, and Jennifer Widom. The Beckman report on database research. *Commun. ACM*, 59(2):92–99, 2016. `doi:10.1145/2845915`.

2 Serge Abiteboul, Benjamin André, and Daniel Kaplan. Managing your digital life. *Commun. ACM*, 58(5):32–35, 2015.

3 Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Comparing workflow specification languages: A matter of views. *ACM Trans. Database Syst.*, 37(2):10, 2012.

4 Serge Abiteboul, Gerome Miklau, Julia Stoyanovich, and Gerhard Weikum. Data, responsibly (dagstuhl seminar 16291). *Dagstuhl Reports*, 6(7):42–71, 2016. `doi:10.4230/DagRep.6.7.42`.

5 Foto N. Afrati and Jeffrey D. Ullman. Optimizing multiway joins in a map-reduce environment. *IEEE Trans. Knowl. Data Eng.*, 23(9):1282–1298, 2011.

6 Alekh Agarwal, Olivier Chapelle, Miroslav Dudik, and John Langford. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15:1111–1133, 2014.

7   Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *International Conference on Web Search and Web Data Mining (WSDM)*, pages 5–14. ACM, 2009.

8   Mert Akdere, Ugur Cetintemel, Matteo Riondato, Eli Upfal, and Stanley B. Zdonik. The case for predictive database systems: Opportunities and challenges. In *Conference on Innovative Data Systems Research (CIDR)*, pages 167–174. www.cidrdb.org, 2011.

9   Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Provenance circuits for trees and treelike instances. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 9135 of *LNCS*, pages 56–68. Springer, 2015.

10 Tom J. Ameloot, Gaetano Geck, Bas Ketsman, Frank Neven, and Thomas Schwentick. Parallel-correctness and transferability for conjunctive queries. In *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS 2015*, pages 47–58, 2015. `doi: 10.1145/2745754.2745759`.

11 Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, May 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

12 Molham Aref, Balder ten Cate, Todd J. Green, Benny Kimelfeld, Dan Olteanu, Emir Pasalic, Todd L. Veldhuizen, and Geoffrey Washburn. Design and implementation of the LogicBlox system. In *International Conference on Management of Data (SIGMOD)*, pages 1371–1382. ACM, 2015.

13 Marcelo Arenas, Pablo Barceló, Leonid Libkin, and Filip Murlak. *Foundations of Data Exchange*. Cambridge University Press, 2014.

14 Marcelo Arenas, Georg Gottlob, and Andreas Pieris. Expressive languages for querying the semantic web. In *Symposium on Principles of Database Systems (PODS)*, pages 14–26. ACM, 2014.

15 Marcelo Arenas, Bernardo Cuenca Grau, Evgeny Kharlamov, Sarunas Marciuska, and Dmitriy Zheleznyakov. Faceted search over RDF-based knowledge graphs. *J. Web Sem.*, 37:55–74, 2016.

16 Marcelo Arenas, Francisco Maturana, Cristian Riveros, and Domagoj Vrgoc. A framework for annotating CSV-like data. *Proceedings of the VLDB Endowment*, 9(11), 2016.

17 Alessandro Artale, Roman Kontchakov, Vladislav Ryzhikov, and Michael Zakharyaschev. A cookbook for temporal conceptual data modelling with description logics. *ACM Trans. on Computational Logic*, 15(3):25:1–25:50, 2014. `doi:10.1145/2629565`.

18 Albert Atserias, Martin Grohe, and Dániel Marx. Size bounds and query plans for relational joins. *SIAM J. Comput.*, 42(4):1737–1767, 2013.

19 Jean-François Baget, Michel Leclère, Marie-Laure Mugnier, and Eric Salvat. On rules with existential variables: Walking the decidability line. *Artificial Intelligence*, 175(9–10):1620–1654, 2011.

20 Eran Balan, Tova Milo, and Tal Sterenzy. BP-Ex: a uniform query engine for business process execution traces. In *International Conference on Extending Database Technology (EDBT)*, pages 713–716. ACM, 2010.

21 Vince Bárány, Balder ten Cate, Benny Kimelfeld, Dan Olteanu, and Zografoula Vagena. Declarative probabilistic programming with datalog. In *International Conference on Database Theory (ICDT)*, volume 48 of *LIPIcs*, pages 7:1–7:19. Schloss Dagstuhl–LZI, 2016.

22 Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

23 Paul Beame, Paraschos Koutris, and Dan Suciu. Communication steps for parallel query processing. In *Symposium on Principles of Database Systems (PODS)*, pages 273–284. ACM, 2013.

24 Michael Benedikt, Wenfei Fan, and Floris Geerts. XPath satisfiability in the presence of DTDs. *J. ACM*, 55(2), 2008.

25 Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

26 Leopoldo Bertossi. *Database Repairing and Consistent Query Answering*. Morgan&Claypool Publishers, 2011.

27 Geert Jan Bex, Frank Neven, Thomas Schwentick, and Stijn Vansummeren. Inference of concise regular expressions and DTDs. *ACM Trans. Database Syst.*, 35(2), 2010.

28 K. Bhattacharya, C.E. Gerede, R. Hull, R. Liu, and J. Su. Towards formal analysis of artifact-centric business process models. In *International Conference on Business Process Management (BPM)*, volume 4714 of *LNCS*, pages 288–304. Springer, 2007.

29 Meghyn Bienvenu, Balder ten Cate, Carsten Lutz, and Frank Wolter. Ontology-based data access: A study through Disjunctive Datalog, CSP, and MMSNP. *ACM Trans. Database Syst.*, 39(4):33:1–33:44, 2014. `doi:10.1145/2661643`.

30 Stefan Borgwardt, Felix Distel, and Rafael Peñaloza. The limits of decidability in fuzzy description logics with general concept inclusions. *Artificial Intelligence*, 218:23–55, 2015. `doi:10.1016/j.artint.2014.09.001`.

31 Michael J. Cafarella, Dan Suciu, and Oren Etzioni. Navigating extracted data with schema discovery. In *International Workshop on the Web and Databases (WebDB)*, 2007.

32 Zhuhua Cai, Zografoula Vagena, Luis Leopoldo Perez, Subramanian Arumugam, Peter J. Haas, and Christopher M. Jermaine. Simulation of database-valued markov chains using simsql. In *International Conference on Management of Data (SIGMOD)*, pages 637–648. ACM, 2013.

33 Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. Autom. Reasoning*, 39(3):385–429, 2007.

34 Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. Conjunctive query containment and answering under description logics constraints. *ACM Trans. on Computational Logic*, 9(3):22.1–22.31, 2008.

35 Diego Calvanese, Giuseppe De Giacomo, and Marco Montali. Foundations of data-aware process analysis: a database theory perspective. In *Symposium on Principles of Database Systems (PODS)*, pages 1–12. ACM, 2013. `doi:10.1145/2463664.2467796`.

36 Sejla Cebiric, François Goasdoué, and Ioana Manolescu. Query-oriented summarization of RDF graphs. *Proceedings of the VLDB Endowment*, 8(12):2012–2015, 2015. URL: http://www.vldb.org/pvldb/vol8/p2012-cebiric.pdf.

37 Shumo Chu, Magdalena Balazinska, and Dan Suciu. From theory to practice: Efficient join query evaluation in a parallel database system. In *International Conference on Management of Data (SIGMOD)*, pages 63–78. ACM, 2015.

38 Moustapha Cissé, Nicolas Usunier, Thierry Artieres, and Patrick Gallinari. Robust Bloom filters for large multilabel classification tasks. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

39 E. F. Codd. Understanding relations (installment #7). *FDT - Bulletin of ACM SIGMOD*, 7(3):23–28, 1975.

40 Wojciech Czerwinski, Wim Martens, Pawel Parys, and Marcin Przybylko. The (almost) complete guide to tree pattern containment. In *Symposium on Principles of Database Systems (PODS)*, pages 117–130. ACM, 2015.

41 Chris J. Date. *Database in Depth – Relational Theory for Practitioners*. O'Reilly, 2005.

42   Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *PoPETs*, 2015(1):92–112, 2015. URL: http://www.degruyter.com/view/j/popets.2015.1.issue-1/popets-2015-0007/popets-2015-0007.xml.

43   Susan B. Davidson and Juliana Freire. Provenance and scientific workflows: Challenges and opportunities. In *International Conference on Management of Data (SIGMOD)*, pages 1345–1350. ACM, 2008.

44   Umeshwar Dayal, Malú Castellanos, Alkis Simitsis, and Kevin Wilkinson. Data integration flows for business intelligence. In *International Conference on Extending Database Technology (EDBT)*, pages 1–11. ACM, 2009.

45   K. Dembczynski, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *International Conference on Machine Learning (ICML)*, pages 279–286. Omnipress, 2010.

46   Daniel Deutch and Tova Milo. A quest for beauty and wealth (or, business processes for database researchers). In *Symposium on Principles of Database Systems (PODS)*, pages 1–12. ACM, 2011.

47   Alin Deutsch, Richard Hull, and Victor Vianu. Automatic verification of database-centric systems. *SIGMOD Record*, 43(3):5–17, 2014. `doi:10.1145/2694428.2694430`.

48   Zlatan Dragisic, Patrick Lambrix, and Eva Blomqvist. Integrating ontology debugging and matching into the eXtreme design methodology. In *Workshop on Ontology and Semantic Web Patterns (WOP)*, volume 1461 of *CEUR Workshop Proceedings*, 2015. URL: http://ceur-ws.org/Vol-1461/WOP2015_paper_1.pdf.

49   Marina Drosou and Evaggelia Pitoura. DisC diversity: result diversification based on dissimilarity and coverage. *Proceedings of the VLDB Endowment*, 6(1):13–24, 2012. URL: http://www.vldb.org/pvldb/vol6/p13-drosou.pdf.

50   Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226. ACM, 2012.

51   Thomas Eiter, Thomas Lukasiewicz, and Livia Predoiu. Generalized consistent query answering under existential rules. In *International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 359–368. AAAI Press, 2016.

52   Corinna Elsenbroich, Oliver Kutz, and Ulrike Sattler. A case for abductive reasoning over ontologies. In *International Workshop on OWL (OWLED)*, volume 216 of *CEUR Workshop Proceedings*, 2006. URL: http://ceur-ws.org/Vol-216/submission_25.pdf.

53   Ronald Fagin, Benny Kimelfeld, Frederick Reiss, and Stijn Vansummeren. Document spanners: A formal approach to information extraction. *J. ACM*, 62(2):12, 2015.

54   Jon Feldman, S. Muthukrishnan, Anastasios Sidiropoulos, Clifford Stein, and Zoya Svitkina. On distributing symmetric streaming computations. In *Symposium on Discrete Algorithms (SODA)*, pages 710–719. SIAM, 2008.

55   Enrico Franconi, Paolo Guagliardo, Marco Trevisan, and Sergio Tessaris. Quelo: an ontology-driven query interface. In *Workshop on Description Logics (DL)*, volume 745 of *CEUR Workshop Proceedings*, 2011. URL: http://ceur-ws.org/Vol-745/paper_58.pdf.

56   Noah D. Goodman. The principles and practice of probabilistic programming. In *Symposium on Principles of Programming Languages (POPL)*, pages 399–402. ACM, 2013.

57   Georg Gottlob, Stanislav Kikot, Roman Kontchakov, Vladimir V. Podolskii, Thomas Schwentick, and Michael Zakharyaschev. The price of query rewriting in ontology-based data access. *Artificial Intelligence*, 213:42–59, 2014. `doi:10.1016/j.artint.2014.04.004`.

58   Georg Gottlob, Christoph Koch, and Reinhard Pichler. Efficient algorithms for processing XPath queries. *ACM Trans. Database Syst.*, 30(2):444–491, 2005.

59 Georg Gottlob, Giorgio Orsi, and Andreas Pieris. Query rewriting and optimization for ontological databases. *ACM Trans. Database Syst.*, 39(3):25:1–25:46, 2014. `doi:10.1145/2638546`.

60 Georg Gottlob and Pierre Senellart. Schema mapping discovery from data instances. *J. ACM*, 57(2), 2010. `doi:10.1145/1667053.1667055`.

61 Benoît Groz, Tova Milo, and Sudeepa Roy. On the complexity of evaluating order queries with the crowd. *IEEE Data Eng. Bull.*, 38(3):44–58, 2015. URL: http://sites.computer.org/debull/A15sept/p44.pdf.

62 Peter J. Haas and Joseph M. Hellerstein. Ripple joins for online aggregation. In *International Conference on Management of Data (SIGMOD)*, pages 287–298. ACM, 1999.

63 Joseph M. Hellerstein, Peter J. Haas, and Helen J. Wang. Online aggregation. In *International Conference on Management of Data (SIGMOD)*, pages 171–182. ACM, 1997.

64 Martin Hepp. The web of data for e-commerce: Schema.org and GoodRelations for researchers and practitioners. In *International Conference on Web Engineering (ICWE)*, volume 9114 of *LNCS*, pages 723–727. Springer, 2015. `doi:10.1007/978-3-319-19890-3_66`.

65 Xiao Hu and Ke Yi. Towards a worst-case i/o-optimal algorithm for acyclic joins. In *Symposium on Principles of Database Systems (PODS)*. ACM, 2016.

66 R. Hull and J. Su. NSF Workshop on Data-Centric Workflows, May, 2009. URL: http://dcw2009.cs.ucsb.edu/report.pdf.

67 Tomasz Imielinski and Witold Lipski. Incomplete information in relational databases. *J. ACM*, 31(4):761–791, 1984.

68 Kalina Jasinska, Krzysztof Dembczynski, , Robert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *International Conference on Machine Learning (ICML)*. JMLR.org, 2016.

69 Abhay Kumar Jha and Dan Suciu. Probabilistic databases with MarkoViews. *Proceedings of the VLDB Endowment*, 5(11):1160–1171, 2012.

70 Mark Kaminski and Egor V. Kostylev. Beyond well-designed SPARQL. In *International Conference on Database Theory (ICDT)*, volume 48 of *LIPIcs*, pages 5:1–5:18. Schloss Dagstuhl – LZI, 2016.

71 Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An interview study. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2917–2926, 2012.

72 Paraschos Koutris, Paul Beame, and Dan Suciu. Worst-case optimal algorithms for parallel query processing. In *International Conference on Database Theory (ICDT)*, volume 48 of *LIPIcs*, pages 8:1–8:18. Schloss Dagstuhl – LZI, 2016.

73 Domenico Lembo, José Mora, Riccardo Rosati, Domenico Fabio Savo, and Evgenij Thorstensen. Mapping analysis in ontology-based data access: Algorithms and complexity. In *International Semantic Web Conference (ISWC)*, volume 9366 of *LNCS*, pages 217–234. Springer, 2015. `doi:10.1007/978-3-319-25007-6_13`.

74 Maurizio Lenzerini. Data integration: a theoretical perspective. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 233–246. ACM, 2002.

75 Jonas Lerman. Big data and its exclusions. *Stanford Law Review Online*, 66, 2013.

76 Feifei Li, Bin Wu, Ke Yi, and Zhuoyue Zhao. Wander join: Online aggregation via random walks. In *International Conference on Management of Data (SIGMOD)*, pages 615–629. ACM, 2016.

77 L. Libkin. SQL's three-valued logic and certain answers. *ACM Trans. Database Syst.*, 41(1):1, 2016.

78 Leonid Libkin. Certain answers as objects and knowledge. *Artificial Intelligence*, 232:1–19, 2016.

79 W. Lipski. On semantic issues connected with incomplete information databases. *ACM Trans. Database Syst.*, 4(3):262–296, 1979.

80 Rong Liu, Roman Vaculín, Zhe Shan, Anil Nigam, and Frederick Y. Wu. Business artifact-centric modeling for real-time performance monitoring. In *International Conference on Business Process Management (BPM)*, pages 265–280, 2011.

81 Mike Marin, Richard Hull, and Roman Vaculín. Data-centric BPM and the emerging Case Management standard: A short survey. In *Business Process Management Workshops*, pages 24–30, 2012.

82 Wim Martens, Frank Neven, and Stijn Vansummeren. SCULPT: A schema language for tabular data on the web. In *International Conference on World Wide Web (WWW)*, pages 702–720. ACM, 2015.

83 Vera Zaychik Moffitt, Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. Collaborative access control in WebdamLog. In *International Conference on Management of Data (SIGMOD)*, pages 197–211. ACM, 2015.

84 G. De Francisci Morales and A. Bifet. SAMOA: Scalable advanced massive online analysis. *Journal of Machine Learning Research*, 16:149–153, 2015.

85 Cecilia Muñoz, Megan Smith, and DJ Patil. Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President, The White House*, May 2016. URL: https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_ data_discrimination.pdf.

86 Hung Q. Ngo, Ely Porat, Christopher Ré, and Atri Rudra. Worst-case optimal join algorithms: [extended abstract]. In *Symposium on Principles of Database Systems (PODS)*, pages 37–48. ACM, 2012.

87 Nhung Ngo, Magdalena Ortiz, and Mantas Simkus. Closed predicates in description logics: Results on combined complexity. In *International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, pages 237–246. AAAI Press, 2016. URL: http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12906.

88 A. Nigam and N.S. Caswell. Business Artifacts: An Approach to Operational Specification. *IBM Systems Journal*, 42(3), 2003.

89 Dan Olteanu and Jakub Závodný. Size bounds for factorised representations of query results. *ACM Trans. Database Syst.*, 40(1):2, 2015. `doi:10.1145/2656335`.

90 Felipe Pezoa, Juan L. Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoc. Foundations of JSON schema. In *International Conference on World Wide Web (WWW)*, pages 263–273. ACM, 2016.

91 Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. Linking data to ontologies. *J. on Data Semantics*, X:133–173, 2008. `doi:10.1007/978-3-540-77688-8_5`.

92 Yashoteja Prabhu and Manik Varma. FastXML: a fast, accurate and stable tree-classifier for extreme multi-label learning. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 263–272. ACM, 2014.

93 Matteo Riondato, Mert Akdere, Ugur Cetintemel, Stanley B. Zdonik, and Eli Upfal. The vc-dimension of SQL queries and selectivity estimation through sampling. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, volume 6912 of *LNCS*, pages 661–676. Springer, 2011.

94 Ruslan Salakhutdinov and Geoffrey E. Hinton. Semantic hashing. *Int. Journal of Approximate Reasoning*, 50(7):969–978, 2009.

95 Akash Das Sarma, Aditya G. Parameswaran, and Jennifer Widom. Towards globally optimal crowdsourcing quality management: The uniform worker setting. In *International Conference on Management of Data (SIGMOD)*, pages 47–62, 2016. `doi:10.1145/2882903. 2882953`.

96 Maximilian Schleich, Dan Olteanu, and Radu Ciucanu. Learning linear regression models over factorized joins. In *International Conference on Management of Data (SIGMOD)*, pages 3–18. ACM, 2016. `doi:10.1145/2882903.2882939`.

97 B. Schölkopf and AJ. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

98 Warren Shen, AnHai Doan, Jeffrey F. Naughton, and Raghu Ramakrishnan. Declarative information extraction using datalog with embedded extraction predicates. In *International Conference on Very Large Data Bases (VLDB)*, pages 1033–1044. ACM, 2007.

99 Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. Incremental knowledge base construction using deepdive. *Proceedings of the VLDB Endowment*, 8(11):1310–1321, 2015. URL: http://www.vldb.org/pvldb/vol8/p1310-shin.pdf.

100 Slawek Staworko, Iovka Boneva, José Emilio Labra Gayo, Samuel Hym, Eric G. Prud'hommeaux, and Harold R. Solbrig. Complexity and expressiveness of shex for RDF. In *International Conference on Database Theory (ICDT)*, volume 31 of *LIPIcs*, pages 195–211. Schloss Dagstuhl – LZI, 2015.

101 Slawek Staworko, Jan Chomicki, and Jerzy Marcinkowski. Prioritized repairing and consistent query answering in relational databases. *Ann. Math. Artif. Intell.*, 64(2-3):209–246, 2012.

102 Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. Data responsibly: Fairness, neutrality and transparency in data analysis. In *International Conference on Extending Database Technology (EDBT)*, pages 718–719. OpenProceedings.org, 2016.

103 D. Suciu, D. Olteanu, C. Re, and C. Koch. *Probabilistic Databases*. Morgan&Claypool Publishers, 2011.

104 Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

105 Y. Sun, J. Su, and J. Yang. Universal artifacts. *ACM Trans. on Management Information Systems*, 7(1), 2016.

106 Latanya Sweeney. Discrimination in online ad delivery. *Commun. ACM*, 56(5):44–54, 2013. `doi:10.1145/2447976.2447990`.

107 Balder ten Cate, Víctor Dalmau, and Phokion G. Kolaitis. Learning schema mappings. *ACM Trans. Database Syst.*, 38(4):28, 2013.

108 Florian Tschorsch and Björn Scheuermann. Bitcoin and beyond: A technical survey on decentralized digital currencies. Cryptology ePrint Archive, Report 2015/464, 2015.

109 Leslie G. Valiant. A bridging model for parallel computation. *Commun. ACM*, 33(8):103–111, 1990.

110 LG. Valiant. A theory of the learnable. *Commun. ACM*, 17(11):1134–1142, 1984.

111 Todd L. Veldhuizen. Triejoin: A simple, worst-case optimal join algorithm. In *International Conference on Database Theory (ICDT)*, pages 96–106. OpenProceedings.org, 2014.

112 K.Q. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *International Conference on Machine Learning (ICML)*, pages 1113–1120. ACM, 2009.