Research Directions for Principles of Data Management (Abridged)

Serge Abiteboul Marcelo Arenas Pablo Barceló Meghyn Bienvenu
Diego Calvanese Claire David Richard Hull Eyke Hüllermeier
Benny Kimelfeld Leonid Libkin Wim Martens Tova Milo
Filip Murlak Frank Neven Magdalena Ortiz Thomas Schwentick
Julia Stoyanovich Jianwen Su Dan Suciu
Victor Vianu Ke Yi

In April 2016, a community of researchers working in the area of Principles of Data Management (PDM) joined in a workshop at the Dagstuhl Castle in Germany. The workshop was 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 the 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 article presents a summary of the report created by the workshop [4]. That 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. The report organizes the identified research challenges for PDM around seven core themes, namely *Managing Data 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 exclusive.

The Dagstuhl report is intended for a diverse audience, ranging from funding agencies, to universities and industrial research labs, to researchers and scientists who are exploring the many issues that arise in modern data management. The report 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 themes in data management stand out where principled, mathematical thinking can bring new approaches and much-needed clarity. The first relates to the overall 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. As documented in numerous sources, so-called "data wrangling" can form 50% to 80% of the labor costs in an analytics investigation. As discussed in the Dagstuhl report, the PDM research areas of *Managing Data at Scale*, *Knowledge-enriched Data*, *Multi-model Data*, *Uncertain Information*, and *Data Management and Machine Learning* are all relevant to supporting Big Data Analytics. The second broad theme of data management where principled thinking can help 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 PDM research areas of *Multi-model Data*, *Knowledge-enriched Data*, *Uncertain Information*, and *Process and Data* are all relevant to this theme. These are providing approaches that make it easier to understand and process the myriad kinds of data and updates involved, and to enable higher degrees of confidence in transactional software that is used to process the data. The third broad theme, which is just beginning to emerge, is the development of new principles and approaches in support of ethical data management. Emerging research suggests that the use of mathematical principles in research on *Ethics and Data Management* can lead to new approaches to ensure data privacy for individuals, a broader perspective on notions of "fair" data dissemination and analysis, and compliance with government and societal regulations at the corporate level.

The findings of the Dagstuhl 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 article includes overviews of the seven PDM research areas mentioned above, and a concluding section with comments about the road ahead for PDM research. The interested reader is referred to the full Dagstuhl Report [4] for more detail and references.

1. MANAGING DATA 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 data processing at scale. Still, however, we face important practical challenges such as the following:

New Paradigms for Multi-Way Join Processing.
A celebrated result by Atserias, Grohe, and Marx [17] has sparked a flurry of research efforts in reexamining 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. Several worst-case algorithms have been designed in different computation models [70, 51, 20, 6], all of which have abandoned the binary tree paradigm, while adopting a more *holistic* approach. In particular, *leapfrog join* [87] has been implemented inside a full-fledged database system. We believe that these newly developed algorithms 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 efforts for designing and implementing new query optimizers and cost estimation under the new paradigm.

Approximate Query Processing.

The area of *online aggregation* [49] studies new algorithms that allow to return approximate results (with statistical guarantees) for analytical queries at early stages of the processing, so that the user can terminate it as soon as the accuracy is acceptable. Recent studies have shown encouraging results [48, 61], but there is still much room for improvement: (1) The existing algorithms have only used simple random sampling or sample random walks to sample from 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 properties of the data. These summarization techniques can also be useful in approximate query processing. (3) Integrating these techniques into data processing engines is still a significant 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) [80, 45, 58]. 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 datasets. 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 main measure of complexity for algorithms designed for these models. Recent studies (e.g., [20]) have established tight bounds on the communication cost for computing 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 important to study the communication complexity of classical machine learning tasks under these models. This is developed in more detail in Section 5, which summarizes research opportunites at the crossroads of data management and machine learning.

We think that the following techniques will be useful in handling these challenges: statistics, sampling and approximation theory, communication complexity, information theory, and convex optimization.

2. MULTI-MODEL DATA: 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? Could we create methodologies allowing engineers to design a new data model?

Understanding Data.
How does one make sense of the data? 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? How can we help users formulate queries in a more uniform way?

Processing Data.
How does one evaluate queries efficiently?

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. 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. [23, 26]). The long-standing challenge of entity resolution is exacerbated in the context of finding correspondences between data sets structured according to different models [85].

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. 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. [30, 15]).

Query Languages.
Go beyond bespoke query languages for the specific data models [13] 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 [56, 46] or via static optimization [21, 33].

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 5). The first, second, and third raise knowledge representation issues (see Section 4). 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.

3. UNCERTAIN INFORMATION

Incomplete, uncertain, and inconsistent information is ubiquitous in data management applications. This was recognized already in the 1970s [32], 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 after 40+ years of relational technology, when evaluating SQL queries over incomplete databases one gets results that make people say *"you can never trust the answers you get from [an incomplete] database"* [34]. In fact we know that SQL can produce every type of error imaginable when nulls are present [63].

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 [53] to new applications such as data integration and exchange [59, 14], consistent query answering [22], ontology-based data access [29], 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 – and sometimes even undecidable, see [62]. 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 [81], 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.

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.

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 [78], or ontologies [44], along the lines of approaches described in Section 4, 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 [55]; 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.

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.

Reasoning.

There is much work on this subject; see Section 4 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?

4. 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. Ontology-based data management has been proposed recently as a general paradigm to address this challenge.

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 specific area where this is starting to play an important role is e-commerce, where standard ontologies are already available [50].

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. 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.

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.

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.

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 [27, 18, 25, 16, 69].

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.

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 [28, 16].

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.

5. DATA MANAGEMENT AND MACHINE LEARNING

We believe that Data Management (DM) and Machine Learning (ML) 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 semantics, maintenance and scale. This complementarity in assets is accompanied by a difference in the core mechanisms: the PDM community has largely adopted methodologies driven by logic, while the ML community centralized around probability 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 required methodologies and formal foundations span a variety of related fields such as logic, formal languages, computational complexity, statistical analysis, and distributed computing.

Category DM for ML

Key challenges in this area 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. This is a critical and time-consuming task [57], and a central theme of modern ML methodologies. 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 features optimally suited for the ML task at hand? Is a given language suitable for a certain ML task? Important criteria for the goodness of a feature language include the risks of *under/overfitting* the training data, as well as the computational complexity of evaluation. The PDM community has already studied problems of a similar nature [47].

The promise of deep (neural network) learning brings 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* [74])? Can database queries (of different languages) complement neural networks by means of expressiveness and/or efficiency?

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 and inference in an efficient manner [72]. Research along this direction has amplified in recent years and includes, for example, the use of hashing [88], Bloom filters [31], tree-based data structures [38] in learning algorithms. Related to this is work on distributed machine learning, where data storage and computation is accomplished in a network of distributed units [7], and the support of machine learning by data stream management systems [67].

Complexity Analysis.

The PDM community has established a strong machinery for fine-grained analysis of querying complexity; see, e.g., [10]. 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 [54]. Results along this direction, connecting DM querying complexity and ML training complexity, have been recently shown [75].

Category ML for DM

Data management systems support 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.

Incorporating ML features is a natural evolution for PDM. Database systems with such features have already been developed [77, 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 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, i.e., optimization that does not modify the final result [76, 89]. As mentioned in Section 1, in some scenarios it makes sense to apply *lossy* optimization that guarantees only an approximation of the 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 [9].

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 [77]. 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 VC-dimension, an important theoretical concept in generalization theory, to database queries, Riondato et al. [73] provide accurate bounds for their selectivity estimates that hold with high probability. This direction can leverage the past decade of research on probabilistic databases [82], which can be combined with theoretical frameworks of machine learning, such as PAC learning [86].

6. 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. Over the past half century, computer science research has studied foundational issues of process and of data mainly as separated phenomena, but research into basic questions about systems that combine process and data has been growing over the past decade. Two key areas where data and process have been studied together are scientific workflows and data-aware BPM [52].

In the 1990's and 00's, foundational research in scientific workflow 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 [36].

Foundational work on data-aware BPM began in the mid-00's [24, 41], enabled in part by IBM's "Business Artifacts" model for business process [71], that combines data and process in a holistic manner. Deutch and Milo [39] 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 has provided the conceptual basis for the recent OMG Case Management Model and Notation (CMMN) standard [65]. Rich work on verification for data-aware processes has emerged [28, 40], and the artifact-based perspective is enabling an approach to managing the interaction of business processes and legacy data systems [83].

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.

While many business processes have been automated using techniques from the BPM field, there are many other processes that are still manual – often because high levels of variation make it cost prohibitive to automate using current techniques.

Evolution and Migration of Business Processes.

Managing change of business processes remains largely manual, highly expensive, time consuming, and risk-prone.

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. The recent industrial interest in shared ledger technologies, e.g., Blockchain, highlights the importance of this area.

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, but is today based on a broad variety of largely *ad hoc* and manual techniques [37].

Workflow and Business Process Usability.
Enabling people to understand and work effectively to manage large numbers of process definitions and process instances 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 [28, 40], verification currently relies on faithful abstractions reducing the problem to classical finite-state model checking. Further work is needed to develop more powerful abstractions, address new application areas, enable incremental verification techniques, and enable modular styles of verification that support "plug and play" approaches.

Tools for Design and Synthesis.
Although compilers and relational database design have both benefited from solid mathematical foundations (context free grammars and dependency theory, respectively), there is still no robust framework that supports principled design of business processes in the larger context of data, resources, and workforce.

Models and Semantics for Views, Interaction, and Interoperation.
A robust theory of views for data-aware business processes has the potential to enable substantial advances in the simplification of process comparison, process composition, process interoperation, process out-sourcing, and process privacy (e.g., see [3]).

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, including new approaches for instrumenting processes to simplify data discovery [64], 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, and also extensions along the lines just mentioned. 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.

7. HUMAN-RELATED DATA & 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 [68] and legal scholars [19]. 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 [79] and was the subject of a recent Dagstuhl seminar [5]. 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 [43]. 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 [35, 84], 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 [60]. Diversity [8, 42] 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 [66].

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.

8. 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 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.

9. REFERENCES

[1] D. Abadi et al. The Beckman report on database research. *Commun. ACM*, 59(2):92–99, 2016.

[2] S. Abiteboul, B. André, and D. Kaplan. Managing your digital life. *Commun. ACM*, 58(5):32–35, 2015.

[3] S. Abiteboul, P. Bourhis, and V. Vianu. Comparing workflow specification languages: A matter of views. *ACM Trans. Database Syst.*, 37(2):10, 2012.

[4] S. Abiteboul et al. Research directions for Principles of Data Management (Dagstuhl perspectives workshop 16151). `https://arxiv.org/abs/1701.09007`.

[5] S. Abiteboul et al., editor. *Data, Responsibly*, volume 16291 of *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl – LZI, 2016, forthcoming.

[6] F. N. Afrati and J. D. Ullman. Optimizing multiway joins in a map-reduce environment. *IEEE Trans. Knowl. Data Eng.*, 23(9):1282–1298, 2011.

[7] A. Agarwal et al. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15:1111–1133, 2014.

[8] R. Agrawal et al. Diversifying search results. In *WSDM*, pages 5–14. ACM, 2009.

[9] M. Akdere et al. The case for predictive database systems: Opportunities and challenges. In *Conference on Innovative Data Systems Research (CIDR)*, pages 167–174. www.cidrdb.org, 2011.

[10] A. Amarilli, P. Bourhis, and P. Senellart. Provenance circuits for trees and treelike instances. In *ICALP*, pages 56–68. Springer, 2015.

[11] J. Angwin et al. Machine bias. ProPublica, May 2016.

[12] M. Aref et al. Design and implementation of the LogicBlox system. In *SIGMOD*, pages 1371–1382. ACM, 2015.

[13] M. Arenas, G. Gottlob, and A. Pieris. Expressive languages for querying the semantic web. In *PODS*, pages 14–26. ACM, 2014.

[14] M. Arenas et al. *Foundations of Data Exchange*. Cambridge University Press, 2014.

[15] M. Arenas et al. Faceted search over RDF-based knowledge graphs. *J. Web Sem.*, 37:55–74, 2016.

[16] A. Artale et al. A cookbook for temporal conceptual data modelling with description logics. *ACM Trans. on Computational Logic*, 15(3):25:1–25:50, 2014.

[17] A. Atserias, M. Grohe, and D. Marx. Size bounds and query plans for relational joins. *SIAM J. Comput.*, 42(4):1737–1767, 2013.

[18] J.-F. Baget et al. On rules with existential variables: Walking the decidability line. *Artificial Intelligence*, 175(9–10):1620–1654, 2011.

[19] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

[20] P. Beame, P. Koutris, and D. Suciu. Communication steps for parallel query processing. In *PODS*, pages 273–284. ACM, 2013.

[21] M. Benedikt, W. Fan, and F. Geerts. XPath satisfiability in the presence of DTDs. *J. ACM*, 55(2), 2008.

[22] L. Bertossi. *Database Repairing and Consistent Query Answering*. Morgan&Claypool Publishers, 2011.

[23] G. J. Bex et al. Inference of concise regular expressions and DTDs. *ACM Trans. Database Syst.*, 35(2), 2010.

[24] K. Bhattacharya et al. Towards formal analysis of artifact-centric business process models. In *BPM*, pages 288–304. Springer, 2007.

[25] M. Bienvenu et al. Ontology-based data access: A study through Disjunctive Datalog, CSP, and MMSNP. *ACM Trans. Database Syst.*, 39(4):33:1–33:44, 2014.

[26] M. J. Cafarella, D. Suciu, and O. Etzioni. Navigating extracted data with schema discovery. In *WebDB*, 2007.

[27] D. Calvanese, G. De Giacomo, and M. Lenzerini. Conjunctive query containment and answering under description logics constraints. *ACM Trans. on Computational Logic*, 9(3):22.1–22.31, 2008.

[28] D. Calvanese, G. De Giacomo, and M. Montali. Foundations of data-aware process analysis: a database theory perspective. In *PODS*, pages 1–12. ACM, 2013.

[29] D. Calvanese et al. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. Autom. Reasoning*, 39(3):385–429, 2007.

[30] S. Cebiric, F. Goasdoué, and I. Manolescu. Query-oriented summarization of RDF graphs. *Proc. VLDB Endowment*, 8(12):2012–2015, 2015.

[31] M. Cissé, N. Usunier, T. Artieres, and P. Gallinari. Robust Bloom filters for large multilabel classification tasks. In *NIPS*, 2013.

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

[33] W. Czerwinski et al. The (almost) complete guide to tree pattern containment. In *PODS*, pages 117–130. ACM, 2015.

[34] C. J. Date. *Database in Depth – Relational Theory for Practitioners*. O'Reilly, 2005.

[35] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *PoPETs*, 2015(1):92–112, 2015.

[36] S. B. Davidson and J. Freire. Provenance and scientific workflows: Challenges and opportunities. In *SIGMOD*, pages 1345–1350. ACM, 2008.

[37] U. Dayal et al. Data integration flows for business intelligence. In *EDBT*, pages 1–11. ACM, 2009.

[38] K. Dembczynski, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic

classifier chains. In *ICML*, pages 279–286. Omnipress, 2010.
[39] D. Deutch and T. Milo. A quest for beauty and wealth (or, business processes for database researchers). In *PODS*, pages 1–12. ACM, 2011.
[40] A. Deutsch, R. Hull, and V. Vianu. Automatic verification of database-centric systems. *SIGMOD Record*, 43(3):5–17, 2014.
[41] A. Deutsch et al. Automatic verification of data-centric business processes. In *ICDT*. ACM, 2009.
[42] M. Drosou and E. Pitoura. DisC diversity: result diversification based on dissimilarity and coverage. *Proc. VLDB Endowment*, 6(1):13–24, 2012.
[43] C. Dwork et al. Fairness through awareness. In *ITCS*, pages 214–226. ACM, 2012.
[44] T. Eiter, T. Lukasiewicz, and L. Predoiu. Generalized consistent query answering under existential rules. In *KR*, pages 359–368. AAAI Press, 2016.
[45] J. Feldman et al. On distributing symmetric streaming computations. In *SODA*, pages 710–719. SIAM, 2008.
[46] G. Gottlob, C. Koch, and R. Pichler. Efficient algorithms for processing XPath queries. *ACM Trans. Database Syst.*, 30(2):444–491, 2005.
[47] G. Gottlob and P. Senellart. Schema mapping discovery from data instances. *J. ACM*, 57(2), 2010.
[48] P. J. Haas and J. M. Hellerstein. Ripple joins for online aggregation. In *SIGMOD*, pages 287–298. ACM, 1999.
[49] J. M. Hellerstein, P. J. Haas, and H. J. Wang. Online aggregation. In *SIGMOD*, pages 171–182. ACM, 1997.
[50] M. Hepp. The web of data for e-commerce: Schema.org and GoodRelations for researchers and practitioners. In *ICWE*, pages 723–727. Springer, 2015.
[51] X. Hu and K. Yi. Towards a worst-case i/o-optimal algorithm for acyclic joins. In *PODS*. ACM, 2016.
[52] R. Hull and J. Su. NSF Workshop on Data-Centric Workflows, May, 2009. `http://dcw2009.cs.ucsb.edu/report.pdf`.
[53] T. Imielinski and W. Lipski. Incomplete information in relational databases. *J. ACM*, 31(4):761–791, 1984.
[54] K. Jasinska et al. Extreme F-measure maximization using sparse probability estimates. In *ICML*. JMLR.org, 2016.
[55] A. K. Jha and D. Suciu. Probabilistic databases with MarkoViews. *Proc. VLDB Endowment*, 5(11):1160–1171, 2012.
[56] M. Kaminski and E. V. Kostylev. Beyond well-designed SPARQL. In *ICDT*, pages 5:1–5:18. Schloss Dagstuhl – LZI, 2016.
[57] S. Kandel et al. Enterprise data analysis and visualization: An interview study. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2917–2926, 2012.
[58] P. Koutris, P. Beame, and D. Suciu. Worst-case optimal algorithms for parallel query processing. In *ICDT*, pages 8:1–8:18. Schloss Dagstuhl – LZI, 2016.
[59] M. Lenzerini. Data integration: a theoretical perspective. In *PODS*, pages 233–246. ACM, 2002.
[60] J. Lerman. Big data and its exclusions. *Stanford Law Review Online*, 66, 2013.
[61] F. Li, B. Wu, K. Yi, and Z. Zhao. Wander join: Online aggregation via random walks. In *International Conference on Management of Data (SIGMOD)*, pages 615–629. ACM, 2016.
[62] L. Libkin. Incomplete information: what went wrong and how to fix it. In *PODS*, pages 1–13. ACM, 2014.
[63] L. Libkin. SQL's three-valued logic and certain answers. *ACM Trans. Database Syst.*, 41(1):1, 2016.
[64] R. Liu et al. Business artifact-centric modeling for real-time performance monitoring. In *BPM*, pages 265–280, 2011.
[65] M. Marin, R. Hull, and R. Vaculín. Data-centric BPM and the emerging Case Management standard: A short survey. In *BPM Workshops*, pages 24–30, 2012.
[66] V. Z. Moffitt et al. Collaborative access control in WebdamLog. In *SIGMOD*, pages 197–211. ACM, 2015.
[67] G. D. F. Morales and A. Bifet. SAMOA: Scalable advanced massive online analysis. *Journal of Machine Learning Research*, 16:149–153, 2015.
[68] C. Muñoz, M. Smith, and D. Patil. Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President, The White House*, May 2016.
[69] N. Ngo, M. Ortiz, and M. Simkus. Closed predicates in description logics: Results on combined complexity. In *KR*, pages 237–246. AAAI Press, 2016.
[70] H. Q. Ngo et al. Worst-case optimal join algorithms: [extended abstract]. In *PODS*, pages 37–48. ACM, 2012.
[71] A. Nigam and N. Caswell. Business Artifacts: An Approach to Operational Specification. *IBM Systems Journal*, 42(3), 2003.
[72] Y. Prabhu and M. Varma. FastXML: a fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD*, pages 263–272. ACM, 2014.
[73] M. Riondato et al. The vc-dimension of SQL queries and selectivity estimation through sampling. In *ECML/PKDD*, pages 661–676. Springer, 2011.
[74] R. Salakhutdinov and G. E. Hinton. Semantic hashing. *Int. Journal of Approximate Reasoning*, 50(7):969–978, 2009.
[75] M. Schleich, D. Olteanu, and R. Ciucanu. Learning linear regression models over factorized joins. In *SIGMOD*, pages 3–18. ACM, 2016.
[76] P. G. Selinger et al. Access path selection in a relational database management system. In *SIGMOD*, pages 23–34. ACM, 1979.
[77] J. Shin et al. Incremental knowledge base construction using DeepDive. *Proc. VLDB Endowment*, 8(11):1310–1321, 2015.
[78] S. Staworko, J. Chomicki, and J. Marcinkowski. Prioritized repairing and consistent query answering in relational databases. *Ann. Math. Artif. Intell.*, 64(2-3):209–246, 2012.
[79] J. Stoyanovich, S. Abiteboul, and G. Miklau. Data responsibly: Fairness, neutrality and transparency in data analysis. In *EDBT*, pages 718–719. OpenProceedings.org, 2016.
[80] D. Suciu and V. Tannen. A query language for NC. In *PODS*, pages 167–178. ACM, 1994.
[81] D. Suciu et al. *Probabilistic Databases*. Morgan&Claypool Publishers, 2011.
[82] D. Suciu et al. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.
[83] Y. Sun, J. Su, and J. Yang. Universal artifacts. *ACM Trans. on Management Information Systems*, 7(1), 2016.
[84] L. Sweeney. Discrimination in online ad delivery. *Commun. ACM*, 56(5):44–54, 2013.
[85] B. ten Cate, V. Dalmau, and P. G. Kolaitis. Learning schema mappings. *ACM Trans. Database Syst.*, 38(4):28, 2013.
[86] L. Valiant. A theory of the learnable. *Commun. ACM*, 17(11):1134–1142, 1984.
[87] T. L. Veldhuizen. Triejoin: A simple, worst-case optimal join algorithm. In *ICDT*, pages 96–106. OpenProceedings.org, 2014.
[88] K. Weinberger et al. Feature hashing for large scale multitask learning. In *ICML*, pages 1113–1120. ACM, 2009.
[89] M. Yannakakis. Algorithms for acyclic database schemes. In *VLDB*, pages 82–94. IEEE, 1981.