DAGSTUHL
MANIFESTOS

Volume 7, Issue 1, January – December 2018

ISSN 2193-2433

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

Publication date
January, 2019

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

Aims, Scope
The manifestos from Dagstuhl Perspectives Workshops are published in the *Dagstuhl Manifestos* journal. Each manifesto aims for describing the state-of-the-art in a field along with its shortcomings, strengths. Based on this, position statements, perspectives for the future are illustrated. A manifesto typically has a less technical character; instead it provides guidelines, roadmaps for a sustainable organisation of future progress.

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

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.

Perspectives Workshop April 10–15, 2016 – http://www.dagstuhl.de/16151
2012 ACM Subject Classification Theory of computation → Database theory
Keywords and phrases database theory, principles of data management, query languages, efficient query processing, query optimization, heterogeneous data, uncertainty, knowledge-enriched data management, machine learning, workflows, human-related data, ethics
Digital Object Identifier 10.4230/DagMan.7.1.1

Engineering Academic Software, *Dagstuhl Manifestos*, Vol. 7, Issue 1, pp. 1–29
Authors: S. Abiteboul et al.

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

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 interoperation 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 preprocessed/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.

Manifesto from Dagstuhl Perspectives Workshop 16472

QoE Vadis?

Edited by

Markus Fiedler[1], Sebastian Möller[2], Peter Reichl[3], and Min Xie[4]

1 **Blekinge Institute of Technology, SE,** markus.fiedler@bth.se
2 **TU Berlin, DE,** sebastian.moeller@tu-berlin.de
3 **Universität Wien, AT,** peter.reichl@univie.ac.at
4 **Telenor Research - Trondheim, NO,** min.xie@telenor.com

Abstract

The goal of the Dagstuhl Perspectives Workshop 16472 has been to discuss and outline the strategic evolution of Quality of Experience as a key topic for future Internet research. The resulting manifesto, which is presented here, reviews the state of the art in the Quality of Experience (QoE) domain, along with a SWOT analysis. Based on those, it discusses how the QoE research area might develop in the future, and how QoE research will lead to innovative and improved products and services. It closes by providing a set of recommendations for the scientific community and industry, as well as for future funding of QoE-related activities.

Perspectives Workshop November 20–25, 2016 – http://www.dagstuhl.de/16472
2012 ACM Subject Classification Quality assurance, user models, user studies, heuristic evaluations, user centered design
Keywords and phrases multimedia, network and application management, network quality monitoring and measurement, quality of experience, socio-economic and business aspects, user experience
Digital Object Identifier 10.4230/DagMan.7.1.30

Executive Summary

This Dagstuhl Manifesto is devoted to future trails that Quality of Experience (QoE) research is expected to take, and lines of activities that deserve to be supported by different stakeholders. Indeed, preceding Dagstuhl Seminars (09192, 12181 and 15022) have had strong impacts on the community forming and joint view onto the QoE domain, including its placement in relation to other areas. An overview is given in the Manifesto's state of the art section, together with a review of evaluation methods and a SWOT analysis. We then turn our focus on the question how the QoE research area might develop in the future, with focus on new applications and services, new methodologies, practical systems and relationships to adjacent research areas. Furthermore, innovative aspects and means to yield innovative and improved products and services based on QoE research are discussed, related to short, medium and long terms. In particular, a marriage between the adjacent areas of QoE and User Experience (UX) is proposed. Besides providing the "Fundamental Law of Quality of Experience", the recommendations for stakeholders in the QoE/UX domain address academic communities, industry partners and public funding agencies, respectively.

QoE Vadis?, *Dagstuhl Manifestos*, Vol. 7, Issue 1, pp. 30–51
Editors: Markus Fiedler, Sebastian Möller, Peter Reichl, and Min Xie

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

Table of Contents

1 Introduction

During the recent decade, the transition from the technology-oriented notion of Quality of Service (QoS) to the user-centric concept of Quality of Experience (QoE) has become an important paradigm change in communication networking research. Simultaneously, the field of QoE as such has significantly developed and matured. This is amongst others reflected in the series of three Dagstuhl Seminars 09192 "From Quality of Service to Quality of Experience" (2009), 12181 "Quality of Experience: From User Perception to Instrumental Metrics" (2012) and 15022 "Quality of Experience: From Assessment to Application" (2015).

The QoE-related Dagstuhl Seminars had a significant impact on the understanding, definition and application of the QoE notion and concepts in the QoE community, for instance with respect to redefining fundamental concepts of quality. That work was performed in close collaboration with the COST Action IC1003 Qualinet [1] that has been concentrating on QoE in multimedia systems and services, and is still actively convening experts from all over the world to regular meetings and exchanges. In particular, this collaboration has led to the widely regarded Qualinet White Paper on "Definitions of QoE and related concepts" [4] and to the launch of a new journal entitled "Quality and User Experience" [2], fostering the scientific exchange within and between QoE and User Experience (UX) communities.

Realising the urgent need of jointly and critically reflecting the future perspectives and directions of QoE research, the QoE-related Dagstuhl Seminars were complemented by the Dagstuhl Perspectives Workshop 16472 "QoE Vadis?", whose output is this Dagstuhl Manifesto. Its remainder is structured as follows: Section 2 provides a state-of-the-art and SWOT analysis of the current research landscape for QoE. Section 3 contains projections of how the area of QoE might develop in the future, and Section 4 postulates how it will lead to innovative and improved products and services. Finally, Section 5 provides a set of recommendations for the scientific community and industry as well as for future funding of QoE-related activities.

2 State of the Art

2.1 Background

In the last years grounding work on the definition on QoE has been performed. Before that time the psycho-acoustic community was referring to quality as the result of a perception and judgement process [9]. In parallel, the networking community was focused on the concept of Quality of Service (QoS), mainly related to low-level network metrics which are indicative for network and/or service performance. For example, the Telecommunication Standardisation Sector of the International Telecommunication Union (ITU-T) defined QoS as follows:

> "Quality of Service is the totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service." [8]

However, the practical use and implementation of the QoS concept left unexplained how the needs of the user are taken into account when characterising the service in terms of QoS parameters. In fact, QoS parameters only describe technical performance of the system or service under consideration, and leave out user perception and judgement. As a consequence, the concept of QoE was developed as user-centric counterpart of QoS.

Members of the COST Action IC 1003 "European Network on Quality of Experience in Multimedia Systems and Services" (Qualinet) [1], as well as attendees of the Dagstuhl

Workshop 09192 "From Quality of Service to Quality of Experience" set out to define QoE, and the discussion between these groups led to the now accepted definition of the resulting Qualinet White Paper:

> "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state." [4]

Based on this definition, a more holistic version that emphasises the process of experiencing has been published in [13]:

> "Quality of Experience (QoE) is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfilment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state." [13]

The Qualinet White Paper further elaborates on influence factors (IFs) contained in the definition as follows:

> "Influence Factor: Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user." [4]

This includes the following three types of influence factors:

- "Human IF is any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socio-economic background, the physical and mental constitution, or the user's emotional state." [4]
- "System IFs refer to properties and characteristics that determine the technically produced quality of an application or service [10]. They are related to media capture, coding, transmission, storage, rendering, and reproduction/display, as well as to the communication of information itself from content production to user." [4]
- "Context IFs are factors that embrace any situational property to describe the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics [11, 10]." [4]

Delight and annoyance used in the above definitions are two emotional states that may help to characterise a user's current state. However, other dimensions that refer to a user's state (such as for instance arousal and dominance) may also be under consideration. The underlying assumption is that system performance (quantified in terms of QoS) may influence QoE, which in turn has a reciprocal interrelationship with the user's state: the current state of the user may influence the user's QoE judgement (for instance a user in good mood may be less critical towards quality impairments), and a positive or negative user experience may also lead to a change in the user's state (for example a user might get very annoyed due to performance issues). It is further assumed that the user state will have a relation to their behaviour. Behaviour may either refer to the behaviour when actually using a service (e.g. the click path when browsing a web page, on a micro level), or refer to the intention to use the service or the actual use of a service at all, on a macro level. Use of a service on the macro level can be assessed in terms of behavioural economics, and may result in business.

Figure 1 Causal relationship between QoS, QoE, user state, user behaviour, and business.

The behaviour of users when being confronted with systems or services is also the target of User Experience (UX) research, which addresses both functional and non-functional (or hedonic) aspects of experience, and how they relate to system design. Whereas there is a strong relationship between QoE and UX, and strong advances can be expected by combining both principles (see discussion below), we refrain from discussing the state-of-the-art of UX research, and rather refer to the Dagstuhl White Paper on User Experience [14], and to [16].

2.2 Problem Areas and Purpose

The concept of QoE, as defined above, has been mostly applied to multimedia systems and services in which there was a clear assignment of producers (e.g. a TV station), network operators (e.g. telcos), and users. For such services, the main purpose of QoE was to manage scarcity of resources. E.g., a network operator could decide which channel settings to apply in a given situation, thus potentially optimising the QoE for groups of users. This rather "channel-centric" point of view was recently extended to broader and more interactive services, such as web browsing, video conferencing, or online gaming. Such services are far more complex to deal with, as user behaviour and actions impact perceived quality to a significant degree, while user context may be of highest importance.

The causal relationship between QoS (technical), QoE (experienced), a user's internal state (e.g. emotional), user behaviour (observable and trackable), and business, is illustrated in Figure 1. In fact, the relationships between those items may be rather complex, in particular when the roles of the parties are less clearly defined, and when not all factors can be fully controlled, e.g. in Over-The-Top (OTT) services. Nevertheless, it is frequently helpful to address this relationship from one of two complementary perspectives, namely the producer or the consumer perspective:

1. The commercial goal of the producer is to make profit, and QoE should serve that purpose. In practice, understanding of QoE can be used as a part of a general service development process. In the simplest approach, the producer applies a bottom-up method in which system and service characteristics (QoS) are measured and adjusted. The measurements can be done by the producer or some external entity. They ask from the QoE experts how quality of experience has been affected. That understanding is then used to design better services. In a more complex approach, the producer tries to model the whole business process from QoS through QoE and customer behaviour to the profitability of producer business. Then, the chain depicted above can be used to optimise the business by changing QoS (e.g., picking a codec for a particular application) or application features, or to avoid negative business impact (e.g., churn).
2. The consumer perspective concentrates on the question: How well does a service with certain quality characteristics fulfil the needs of the consumers, and how can the service be adapted according to these needs? QoE is an integral part of that issue, but does not provide the whole picture. The needs include happiness, usefulness, and overall well-being of the consumers, amongst others. As also other aspects such as context, emotional state

and expectations affect the consumer's satisfaction, all in all we are only able to control a small portion of the factors influencing a customer's QoE.

2.3 Methods, Models and Tools

In order to optimise services for QoE, experienced quality needs to be quantified, and – to a certain extent – made projectable (which means especially the identification and quantification of those influence factors which are under our control). For this purpose, a number of methods, models and tools have been developed in the past. The following paragraphs provide an overview of the approaches which were followed.

Evaluation methods for QoE can typically be divided into methods involving actual users, and instrumented measurement approaches. Most QoE studies involving users take place in controlled lab settings, which are characterised by high internal validity and a high level of control, and typically manipulating one or more independent variables. However, other studies have been emerging which aim at increasing the ecological validity in QoE studies, to reach a higher number of users and in some cases to gain a better understanding of relevant influence factors. These include approaches for data capturing (of implicit and/or explicit, self-reported user feedback, data from the application and network conditions, etc.) on a mobile device "in the wild", studies in a lab environment designed to resemble the natural context of use to a higher degree, and analysis through crowdsourcing. The focus of QoE evaluation can moreover have different degrees of granularity, in terms of the considered temporal dimensions (e.g., longitudinal vs. instantaneous, cross-sectional time-span). Currently, most QoE studies focus on a short time span (using short stimuli and evaluations at one moment in time), but the interest in the long-term development of QoE is increasing and requires other methods, outside of the lab. Especially for regularly-used services the consideration of a single usage episode is not enough, and methods capturing a number of subsequent usage episodes need to be used [6].

2.3.1 Evaluation methods involving users

Typically, QoE evaluation studies involving human observers (sometimes called "subjective evaluations") are based on a series of recommendations from ITU regarding the assessment of quality in different application domains (for traditional services), containing information about how the experiment should be conducted, which scales should be used, what the test environment should look like, etc. The ITU-T P series and the ITU-R BS and BT series of Recommendations provide details in this respect. Participants in studies are exposed to certain stimuli (e.g., 10 second video excerpts) or interact with a system under certain test conditions, and are then asked to rate (mostly quantitatively) the experienced quality. Mostly, quantitative feedback is collected from test users, yet there have been some studies adopting a more qualitative approach. Through statistical analysis of the collected data, the impact of the controlled independent variable(s) can then be quantified.

2.3.2 Instrumental evaluation methods

Instrumental evaluation methods do not involve explicit user feedback, but provide data which is expected to be linked to experience, and which potentially allows to estimate QoE. The data can stem from different angles:

- *Measurements taken from the user*
  - Behavioral measurements: Such measurements can include, e.g., the number of clicks, the viewing behaviour, user actions (e.g., muting video, refreshing a page), errors in executing a certain task, collecting gaze information (e.g., through eye-tracking), etc.
  - Physiological measurements: Increasingly, the usefulness of physiological measures and tools for QoE studies in general, and in particular to investigate how it relates to emotion as one aspect of a user's state, has been investigated. These include, e.g., Galvanic Skin Response (GSR), heart rate variability, Electroencephalogram (EEG), Near-InfraRed Spectography (NIRS), Electromyography (EMG), and functional Magnetic Resonance Imaging (fMRI).
- *Measurements taken from the system*
  - Signals: Signals, such as video, speech, audio, but also other environmental signals captured by the system (e.g. in the case of Internet of Things applications) provide a rather comprehensive and continuous description of what information is transported to the user. Access to these signals may, however, sometimes be difficult, and the user reception of these signals (such as viewing or hearing characteristics) need to be taken into account when estimating their impact on QoE.
  - Parameters: Parameters, such as codec, throughput, bandwidth, buffer size, delay, etc., are performance metrics (thus, QoS) which may be related to QoE.
- *Measurements taken from the context*
  - These can include measurements that provide information about the context in which the experience takes place, e.g., location, temperature, static vs. nomadic use, etc., using different types of sensors (often used in combination with the collection of explicit user feedback through self-reports, e.g., in the Experience Sampling Method).

2.3.3 Prediction models

Prediction models *estimate* QoE or certain aspects of it, mostly on the basis of measurements taken from the system. Only few such models are known which use measurements taken from the user or from the context as input information.

- Signal-based models: Models where the signal represents the model input can be distinguished according to the availability (or not) of a clean, non-degraded reference signal
  - Full Reference: Full reference to compare the signal to is available.
  - Reduced Reference: Reduced reference (i.e. a simplified version of the non-degraded reference) is available.
  - No Reference: Only the degraded signal is available, no reference to compare to.
- Parametric models: Parametric models aim to predict QoE for a certain scenario, based on input parameters related to the system or the signal. Depending on whether the parameter values are measured during system operation, or estimated from planning values of a new system, these models can be classified into
  - Monitoring models
  - Planning models

Planning and optimising QoE is the task of **QoE management**. Up to date, this has commonly been addressed from complementary perspectives [15]. On the one hand, QoE-driven application management addresses monitoring, control, and adaptation on the user and application host/cloud level, by optimising the quality of OTT services [7]. On the other hand, QoE-driven network and system management mechanisms concern vendors, providers and operators, with the aims to obtain insight into impairments perceived by users and their

relationships to QoS [5] and to identify root causes of potential QoE problems. QoE control and optimisation mechanisms deployed in the network focus on optimised network resource allocation and efficiency, admission control, QoE-driven routing, etc. Those mechanisms are especially critical for wireless and mobile networks, characterised by variable resource availability and inherent resource limitations [12].

However, there is an ongoing need for research and development efforts in the QoE management domain in order to yield approaches that overarch applications, services, systems and networks. For this, there are promising integrated and cross-layer approaches combining both application and network management mechanisms [3], in particular in the context of new networking paradigms such as Network Function Virtualisation (NFV) and Software Defined Networking (SDN). With services being delivered via a chain of different providers, there is a clear need to address the potential of QoE management mechanisms in the context of new business models. Specifically, collaborative models between the network and application service providers may improve QoE with a positive impact on the user state. This calls for the definition of specific interfaces where QoE-related data is exchanged between the stakeholders, cf. Section 4.2.

2.4 Applications

The application areas for QoE may be classified into consumption and interactive (real-time and non real-time, e.g. email) services. Regarding the former, visual consumption services, such as video streaming, television, and image transmission dominate the field, followed by audio and data transmission such as file transfers. Recently, the incorporation of QoE concepts in multi-sensory, augmented and virtual reality consumption services has been observed. On the other hand, interactive services such as speech (in particular telephony), web browsing and other web applications, online gaming, cloud services, and video conferencing have been in the focus of QoE research. Given the expected major impact of Augmented Reality (AR) and Virtual Reality (VR) applications in the near future, preliminary works on the evaluation of the quality has also been conducted in this area. Furthermore, there are new applications of QoE in emerging contexts such as those where Internet of Things (IoT) applications are deployed. It has to be noted that the level of maturity of QoE research, methods, models and tools for these emerging services is far lower than for the "classical" video, speech and audio services.

2.5 SWOT Analysis

A SWOT analysis has been carried out by the attendees of the Dagstuhl Perspectives Workshop 16472 "QoE Vadis?". The following statements represent unfiltered viewpoints and judgements of the participants.

Strengths

QoE concepts have matured. In particular, QoE definitions have evolved to a stable, well-accepted status. Influence factors and QoE as well as quantification of quality improvements are well-understood. Practically usable methods and tools for a set of applications have been developed, with practical impact. As a consequence, QoS-driven network and service management are gradually being replaced by QoE-driven management techniques, providing

Figure 2 Evolution from QoS to QoE to QoL.

telcos (amongst others) with better methodologies. There is an increased focus on bringing technological innovation closer to the end-user/customer, for instance through more user-centric design processes. The community has evolved towards a multi-disciplinary group, which is reflected by the methods used. Also, there is a clear economic relevance.

Weaknesses

We perceive a set of lacks, for instance of a theoretical framework to guide research and design, especially in new application areas; of large and open databases to be used for QoE analysis; of longitudinal QoE studies and models; and of measures to assess the user state and the implications of QoE for user behaviour. Furthermore, strongly interdisciplinary aspects not sufficiently covered so far, such as interaction design; user emotions; cognition; needs; preferences; and behaviour. Studies suffer from low degrees of generalisability, for instance between lab studies and studies in the wild; between similar services; due to application-specific models; and due to fast changes in the services and their settings.

Opportunities

More interdisciplinary work will enable more accurate models and help to get a better understanding of the influence factors. Knowledge can be transferred to enable QoE-prediction in new application areas (within and beyond multimedia). The business potential of QoE can be enhanced. Consumers can be provided with better consumer information on communication services. An approach for "Quality of Experience by design" can be developed. Likewise, user happiness and well-being can be increased ("happiness by design"). Also, the "tyranny of the *Mean* Opinion Score (MOS)" may end, by modelling and exploiting individuality and variations among users instead of staying with MOS-typical averages and aggregations.

Threats

User privacy might be affected. QoE may be considered solved, or not relevant for new application areas. Implementing QoE might not be cost-effective. There are signs for an identity crisis of QoE: A clear target of QoE is still missing; it is difficult for experts and non-experts alike to capture QoE concepts; and the position and visibility of QoE as compared to adjacent areas (e.g., UX, Customer Experience) may be considered weak.

3 How the QoE Research Area Might Develop in the Future

The state-of-the-art analysis exhibited an evolution trend from QoS to QoE, and lately to QoL (Quality of Life) as sketched in Figure 2, indicating an increase of QoE involvement into the society and into people's daily life. New concepts, methodologies and principles are expected to bring QoE research towards this direction.

Following the SWOT analysis, several research areas are proposed for QoE to

- cover a wider range of applications and services (breadth);
- build more accurate models and develop new methodologies to gain better understanding of users and predict QoE (depth);
- generalize the results to more practical scenarios and provide more feasible solutions to stakeholders (practicality);
- establish a closer partnership with adjacent research areas to broaden the studied perspectives and enhance research efficiency (efficiency and visibility).

3.1 Expansion to New Applications and Services

Traditional QoE is focused on multimedia services. With the development of new enabling technologies (IoT and immersive technologies, such as augmented reality, virtual reality, 3D presentation and capturing), new services emerge with new formats and requirements. QoE research needs to move beyond traditional multimedia services and extend to new emerging services and applications such as E-Health; work experience; learning and education; and immersive services and communications, in new scenarios like smart city and smart home. New models and methodologies are required to describe the QoE of these new services and to capture the key quality issues (or influence factors).

The challenges are two-fold. First, the quality features of these new services are either unclear or have only partly been investigated. As a consequence, it is hard, if not impossible, to predict and characterise the potential quality dimensions in order to model them. Second, new services are developed and launched at a speed (in a scale of weeks/months) much faster than QoE research (in a scale of years) can be performed. Conventional QoE research approaches take a long time to finalise and standardise QoE assessment methods for a specific service (e.g., identify quality dimensions, run lab tests, standardise subjective test methods, develop prediction models, etc.), which is no longer suitable in the new era. In order to reduce the risk of significantly lagging behind the service development, QoE has to come up with new approaches to speed up the process, i.e., building functional/feasible QoE models for new applications quickly.

3.2 Development of New Methodologies

In parallel to the development of QoE models for new services, there is also a need to develop new methodologies to investigate the aspects that are critical for QoE research but were not (fully or precisely) tackled by previous QoE work.

One of the biggest challenges in QoE is the interrelation between human emotions, cognition, attitude and behaviour, and the role of QoE in that context. Its study requires a multi-disciplinary approach, involving expertise from user experience (UX), social science, different sub-strands (e.g. experimental, social, etc.) of psychology, physiology etc. In order to develop techniques that can formalise, model, measure and analyse human behaviour, several factors need to be addressed. The first question is how to describe the user behaviour at both micro- and macro-level. Second, since many services tend to be interactive, then the question arises how to describe the interactive activities between users, between users and machines, and how to evaluate the impact of such interactions. Third, user behaviour is a continuous and complex process. Current work mainly assesses QoE in a short term

(e.g., from ten seconds to a few minutes). Practical service usage spans over a longer time period, which complicates the user behaviour and thus requires new models to capture the user affective state and the dynamic behavioural variations. Fourth, it is still a somewhat open issue how to practically measure the user behaviour, e.g., what data should be collected from the system and from the users, respectively, and how the user data should be collected, e.g., implicitly or explicitly.

User behaviour is one key factor influencing QoE. There are many other factors contributing to QoE which are required to be investigated, some of which have already been mentioned in the state of the art section earlier in this document. Representatives are mobility for mobile broadband services and non-multimedia-type factors for new IoT use cases. Considering that the overall QoE is a compound effect of these factors, research has to be done to understand both the impact of individual influence factors and their combined impact. The resulted high complexity requests more advanced approaches of data analytics. Finally, QoE assessment should also become more capable of evaluating the impact of adaptive approaches (e.g. dynamic adjustment of the performance in real time), which are more commonly used by service providers. Current evaluation methods are not suited for this purpose, and therefore need to be adapted and extended, such that the impact of adaptive operations on the overall QoE (either during a session or over a number of sessions) can be investigated.

3.3 Generalisation to Practical Systems

In order to make QoE visible and valuable to industry, QoE needs to provide more insightful and practical solutions to practitioners besides theoretical and experimental results. As a response to the request from service providers and operators, a repository of objective models and/or a toolbox could be designed and then used by them to predict QoE for different services, quantitatively and qualitatively. However, since many mature QoE models were developed in the lab environment or in a small-scale scenario for specific services, several issues are raised up in regard to how to generalise these models to practical systems with a much larger scale while maintaining similar performance and usability.

First, a large-scale QoE framework cannot support the same complexity and resource consumption as subjective lab tests can. More objective (or hybrid) assessment models are needed to embed QoE functionalities into a system. A possible approach could be to i) define a set of key influence factors or measurable metrics that are sufficiently powerful to study the QoE of the services; ii) give a quantified indicator of QoE; iii) derive a prediction model to calculate the QoE indicator from the defined measurable metrics.

Second, data collections and QoE predictions have to be automated in order to enable large-scale QoE measurement and monitoring. This may require association with other technologies like machine learning and IoT that will allow for automated and intelligent monitoring, prediction, and improvement of QoE.

Third, current QoE models and results are limited by their application range. It is hard to transfer a QoE model from one to another, different service. A more generalised framework is demanded to facilitate QoE prediction in next generation networks with diverse services.

Finally, the innovation cycle of QoE model creation for novel services and application domains clearly needs to speed up, without of course lowering the quality of the models themselves. Several examples from the history of QoE, including, e.g., the evolution of sound quality models; more than one decade of research on the E-Model; or the tedious struggle

towards modelling QoE for IPTV demonstrate the need for significant further effort to be put into innovation cycles with both sufficient speed and quality to render results that are useful to the practitioners in the field.

3.4 Relationship to Adjacent Research Areas

Facing the challenges of speeding up the QoE model development process and automating QoE monitoring and prediction in practical large-scale systems, QoE needs to partner with adjacent areas (e.g., UX and machine learning) to seek more effective methods and models.

UX and QoE have many commonalities, and are complementary to each other (UX is more into qualitative assessment whereas QoE is focused on quantitative evaluation). It is natural to build a bridge between QoE and UX so that transferable knowledge, tools and results can be exchanged and reused in both areas. By identifying the areas with common interest (e.g., VR/AR), QoE may adapt well-developed UX methodologies and tools to assess quality dimensions of new services, and modify/apply UX methodology and results to the engineering/algorithmic perspective of QoE.

Analytical tools are necessary for successful QoE assessment. As a significant use case to improve QoL, the special features and demand of QoE should be brought to the machine learning/AI/big data community. The high complexity, the multi-dimensional and multi-sensory features, the inclusion of user behaviour in the generated data and the demand for explicit interpretation of analytical outcomes may require the development of new advanced machine learning algorithms.

In addition, a physiological point of view is useful to describe how expectations and experiences are formed. Business and economic perspectives will help to reveal the relationships between QoE, satisfaction and service provisioning, e.g., willingness to pay, charging and pricing, resource allocation, operation planning and optimisation (which solution is more cost-efficient, fewer customers with high quality services vs. more customers with low quality services?), and the impact of net neutrality. Specifically, means have to be found to incentivise different stakeholders to cooperate in the effort of improving QoE.

As an example, a concept of "QoE by Design" or "QoE in Design" is proposed that basically covers all the above aspects. The idea behind "QoE in Design" is to integrate QoE into the service design process from the beginning, instead of waiting until the service is launched. During the design phase of new services, QoE dimensions will be identified, including the finer-grained user behaviour changes. Functions will be added to instrument systematic measurement of the identified QoE dimensions in a large-scale context. During the proof of concept phase, beta users will be included in the process of defining service characteristics and field tests will run with representative panels and reliable prototypes. After the services are launched, the system will continue monitoring quality dimension measures and user behaviour, which will feed back to refine and modify the service design.

4 How QoE Research Will Lead to Innovative and Improved Products and Services

As outlined in the state-of-the art, the consideration of QoE leads to several benefits for the stakeholders. From a technical point of view, QoE-driven products and services allow to

minimise annoyance and to solve technical problems that hinder good user experience and QoE, e.g. by utilising QoE monitoring, while user experience monitoring barely exists.

However, QoE research has focused mainly on the **QoE ego-system** rather than on the **QoE eco-system**. This means that QoE has been mainly addressed within a single session on a short-time scale for a single user of one concrete application. Thereby, different facets have been addressed by the research community like subjective user studies to identify QoE influence factors for particular applications, QoE models to quantify and capture the effects of those influence factors, and QoE monitoring approaches to provide means for QoE management for improved QoE.

In this section, the question is addressed how QoE research will lead to innovative and improved products and services. To this end, the entire QoE eco-system and the stakeholders along the service delivery chain to the end user need to be considered. In comparison to the traditional QoE ego-system thinking, the QoE eco-system faces manifold research challenges. It is required to extend current QoE research by the different perspectives of the QoE eco-system, and to incorporate user experience. The following items are the market needs where QoE may have an impact.

- The service / system providers (operator, media content producer, vendors, software developers, communities of users) need methodologies and tools to manage the quality of provided services in order to be more competitive.
- Current and future products and services should focus on customer experience, reflecting the business value of QoE.
- People's quality of life needs to be central in the services and products design, addressing the societal value of QoE.

4.1 Analysis of Technical Infrastructure and Artefacts in Requirements Analysis

To come up with innovative and improved products and services, the workflow in the design process of the service and products needs to be revised in such a way, that QoE is included in the process, and put into a relationship with technological aspects.

As an integral part of the requirements analysis of products and services, ethnographical observations have to be carried out to understand the workflow of a specific domain in context, and to infer recommendations whether and how a technology can be used to improve the workflow, and thus the happiness of stakeholders. Typically, stakeholders' behaviours and activities are the focus of observations and analysis. If existing infrastructure (e.g., internet connectivity) and technological tools (e.g., desktop) are considered as 'background' (i.e. not serving as data collection tools or objects of evaluation), they are not analysed at the same granularity level or as systematically as stakeholders' behaviours/activities. However, such background artefacts can have significant effect on stakeholders. QoE can provide a model how to systematise or parameterise these potential factors to bridge the gap.

4.2 Innovative Aspects Through QoE Research

QoE research introduces a facet of innovative aspects. The transition from the QoE ego-system to the QoE eco-system incorporates all stakeholders and their needs. Thereby, QoE is supposed to **remove technical barriers** and allow for a **better communication** between

stakeholders. **QoE models** enable a-priori testing of new applications, especially interactive applications, in different contexts, and thus provide a better holistic point of view on user delight or annoyance.

The introduction of QoE-enabled Application Programming Interfaces (APIs) and semantics through a **semantic layer** will allow relevant stakeholders – providers, operators and customers – to have **transparent access** to agreed-upon QoE-relevant data. Such a semantic layer with open APIs allows for new or improved services and products in the market, such as applications/services and their management. Thus, the semantic layer is a key enabler for **increased competition on fairer grounds** amongst different providers. Mutually agreed-on QoE data may serve as key differentiator and bring the customer in a stronger position, being able to choose between different competing providers. For instance, **QoE-driven recommendation functionalities** can be implemented on that layer, e.g. to offer the user contents across platforms, while **real-time QoE feedback** allows for dynamic (re-)configuration of applications, services and underlying resources in order to yield a sustainable balance between QoE provisioning and related spendings. Obviously, QoE-enabled APIs have the potential to foster the creation of the QoE eco-system, and to act as key enabler for QoE improvements and innovation.

By taking into account QoE, the provider demonstrates that it **cares about the user**. This has also the effect of making the users keener on making their data available. On the one hand, the use of user data to quality-related goals is restricted to limit privacy concerns. Thereby, data may also be shared at an aggregation level at the upper layers. On the other hand, the users may be provided with information related to QoE which may e.g. bring insights when facing QoE problems or enable the user to overcome QoE issues when using a service, e.g. switching off background applications. As a result, regulators will be pushed to change the **privacy regulations** on the usage of user data, thereby balancing the need for open data against privacy requirements.

4.3 Means and Approaches Fostering QoE-driven Innovation

These aspects can be introduced along with the following items on different time scales.

(a) **On the short term**, within the next five years, a variety of means and approaches will foster QoE-driven innovation and improvement of services.

One key element is **machine learning and data analytics**. This approach can be used to predict QoE on the basis of system and user related data (e.g., user behaviour and status). Thereby, user comments and feedback from external fora can be exploited to assess the perceived quality and user behaviour. **Sentiment analysis** may then be a promising approach for obtaining an enriched data set for QoE assessment. But QoE also represents a useful input to the use of machine learning and data analytics in (i) the assessment of the user experience and user behaviour, and (ii) in the management of the network.

In general, **better QoE models** are another key enabler for innovation and improvement. Key aspects are an extension of QoE models that match different user profiles and implement personalisation. Furthermore, QoE models need to address the different perspectives of the QoE eco-system, e.g. by incorporating user behaviour as part of the model, or by identifying and including relevant internal and external context factors including physical, cultural, social, or economic context. As an example, QoE models used in WebRTC need to be improved, impacting a large number of WebRTC-based applications.

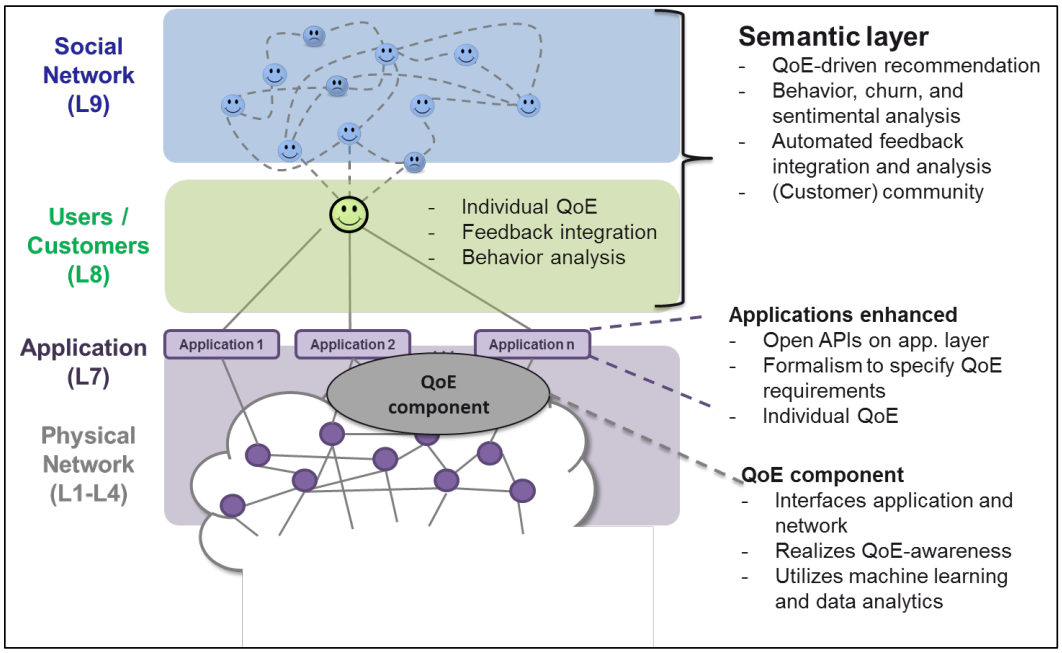


Figure 3 Semantic layer on top of the network layers.

Those QoE models need to result in **direct and operationable methodologies and tools** that improve existing or upcoming products (e.g., concrete adaptive streaming improvements; coding). Another innovation example is the compensation for poor behaviour (in terms of QoE or user experience), leading to an **overall better experience with the service in the longer term**, after facing temporary disruptions (e.g., vouchers; explanations; discounts). QoE can drive the design and implementation of applications and services, for instance to avoid unexpected or aberrant behaviours when the network behaves badly (e.g., by providing tools and mechanisms to allow for graceful degradation of the user experience). The tools and techniques need to be **transferred to practitioners**. Beyond academic dissemination, rather self-contained, **vulgarisation/popularisation** efforts are required to reach all stakeholders and practitioners.

(b) On the medium term (i.e. within 3–7 years), various innovation enablers and technological solutions are foreseen which partly rely on the short term means. Innovation is fostered through improved competitiveness by improving QoE/UX in new or existing services.

Such a major innovation driver is a **semantic layer** which interfaces the different stakeholders and allows exchanging information, which is illustrated in Figure 3. The key elements are **open APIs** in the application layer and a **formalism to specify QoE requirements**, to create QoE-aware services and applications. For example, APIs for telcos will allow services to specify requirements, which can innovate service assurance for OTTs. QoE could be a critical component in these approaches. Getting closer to the user and the user experience can be realized via sophisticated **feedback integration** to collect and analyse user experience data. As part of this semantic layer, measurement approaches and tools are provided incorporating knowledge of the QoE key influence factors and QoE models.

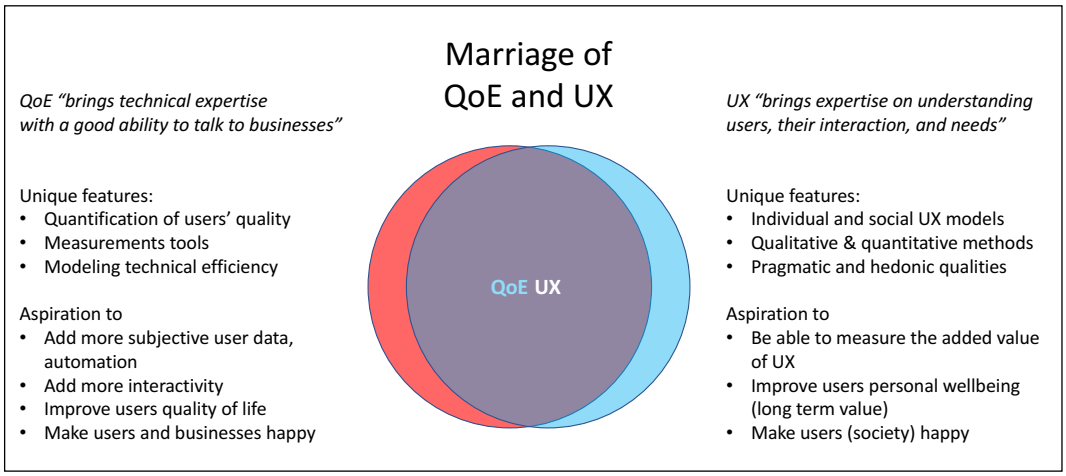


Figure 4 Marriage of QoE and UX.

From the perspective of a provider, QoE is an enabler to **customer relations**. One way is to gather information about the users' personal experiences and to leverage existing models for "average users". **Personalised QoE prediction** may be done in a general enough way, and packaged for use by concrete services. Another way is to **analyse customers' feedback** and mapping it to QoE disruptions (e.g., to check historic issues). Such a **QoE tool for customer relations** supports the improvement and innovation of services and products. This also includes customer communities. The building of customer communities may be promising. For example, if one can provide QoE estimates in real time, that information can be provided to the user, and their feedback can be gathered. It is a research topic to investigate which feedback would be useful to collect or which compensation types would be appropriate for the situation.
With telcos transitioning to **QoE-driven policies** for e.g., network design, base station deployment, etc., tools are required for realizing those policies. Thereby, QoE could complement these activities, focusing on the technology and performance requirements of e.g., proposed designs. QoE research provides a bridge to industry to foster innovation, e.g. in the MPEG-5 standard for multi-sensory services.

(c) **On the long term**, a **"next generation of QoE/UX-aware" designers and engineers** are to be formed, who will be able to use the tools and techniques of QoE research to better develop new products. This requires to educate students accordingly. **QoE by design** or integrating **QoE in the design** should be considered as fundamental part of the workflow in the design process of the service and products. This also means to merge the UX and QoE communities' expertise, objectives and vision, the **"marriage of QoE and UX"**, which would help to improve on existing unique features, follow aspirations, and link addressed stakeholders, as illustrated in Figure 4.
The marriage of QoE and UX may lead in the long-term to the next generation of QoE/UX-aware designers and engineers who are able to fulfil new requirements:

- **Teacher** of future generations: requires to establish the educational environment to train them about QoE/UX.
- **Developer** of new tools to enable innovation: requires to follow an integrated approach of QoE/UX in the development process.
- **Manager** to convey the ideas to businesses: requires to communicate the added value of QoE/UX to businesses and customers.

Finally, the combination of QoE and UX will foster and improve services and products from different domains: Multimedia/Entertainment/Gaming, IoT/Wearable Interfaces, Multisensory Interaction. This may allow to integrate current UX efforts into QoE research towards user acceptance, trust, safety, emotions, user wow, engagement, fun, flow, immersion, and presence.

5 Recommendations for Stakeholders in the QoE/UX Domain

Quality of Experience (QoE) and User Experience (UX) are increasingly gaining importance from several viewpoints corresponding to the different stakeholders in the ecosystem. In this final section, we provide recommendations towards enabling the development of the domain.

More specifically, we consider three stakeholder categories in detail, i.e. scientists working in the field, industry, and public funding agencies. For all them, the "Fundamental Law of Quality of Experience" applies, which, thriving on notorious historical examples, could be formulated as follows:

R0: It's the end user, stupid!

Putting the end user into the centre of the innovation cycle is indispensable for the sustainable success of the future service-oriented industry as a whole, as s/he is the one with complete information about service experience, and who eventually has to pay for it. Hence, we strongly recommend that any stakeholder focus strongly on the end user, his/her expectations and real needs.

5.1 Academic communities

R1: Promote interdisciplinary research.

It has become abundantly clear that much closer collaboration needs to take place between the involved scientific communities, i.e. QoE, UX, and behavioural economics. We recommend the organisation of workshops and symposia involving all these communities, for example in a setting such as Dagstuhl. This, along with joint research efforts, will lead to the sharing of knowledge, methodologies and tools that is needed to further the development of the research agenda and impact. As a result, a solid theoretical and practical foundation for both QoE and UX communities will be achieved. At the same time, joint publication venues for all relevant topics related to QoE and UX shall be provided, along the lines of, e.g., the recently founded "Quality and User Experience" journal [2].

R2: Provide access to open data and tools.

Despite the associated difficulties, we emphasise the importance of gathering QoE-related data from operational services and applications, which will enable us to, e.g., better understand key influence factors, develop more accurate models for QoE, and effective QoE management mechanisms. In addition, open source tools for supporting the creation, sharing and evaluation of data should be developed and maintained by the scientific community.

R3: Drive investigation beyond the comfort zone.

While the current state of the art already provides a comprehensive toolbox for QoE research, it is considered extremely important to emphasize topics and methods outside of the established framework. For instance, future research should address a deeper understanding of the time-scales involved in QoE and UX modelling, as well as the use of bleeding-edge analytical, statistical and modelling methods (including big data, deep learning, and other machine learning techniques). While this might offer an opportunity to speed up the often time-intensive process of creating appropriate QoE models, especially for new application fields, it will be pivotal to also increase the quality of the models themselves, which provides an equally challenging task.

5.2 Industry partners

R4: Turn QoE from reactive to proactive research.

With most current services, QoE is at best an afterthought, often resulting in user frustration and churn. Hence, inspired by the concept of "Security by Design" which has become prevalent in modern services, as it helps dealing with a large number of security problems, we propose fostering a "QoE by Design" approach to service development, whereby QoE informs the service or system design choices, so as to facilitate a positive user experience. Thus, QoE needs to become an integral part of system and service design, which in turn requires resources, dissemination and exploitation efforts, and expertise from other domains, such as UX. Hence, our main recommendation for maximising the impact of QoE and UX in the business domain, is the adoption of the "QoE by Design" approach described above. This will enable the development of innovative QoE-aware offerings.

R5: Implement mechanisms for direct quality feedback.

We strongly recommend investing efforts in raising awareness of the importance of QoE for end users, and get them involved by promoting constructive feedback to the service providers, instead of simply churning. To this end, quality feedback gathering mechanisms could be easily integrated into all sorts of applications, enabling users to directly and easily submit QoE-relevant feedback to the service provider(s). In analogy to the wide-spread "Help" or "Like" facilities, there could be a "Quality feedback" mechanism that provides an intuitive means for users to give feedback about their quality of experience in a timely and unobtrusive manner.

R6: Join forces within industry.

Exchanging QoE-related information between business stakeholders (e.g., telcos, over-the-top providers, infrastructure providers, content providers) towards the implementation of QoE monitoring and management solutions helps optimising services and thus creates a win-win situation for all sides. To this end, we encourage the creation of an openly accessible repository for vulgarisation of domain knowledge and dissemination of tools and methods, especially to foster the adoption of the "QoE by Design" approach. Further, we suggest the integration of customer experience management work-flows. Moreover, the business implications of QoE and UX need to be further studied, and their value communicated

in a clear way to industry players, especially concerning sustainable business models and opportunities. Legal considerations, in particular related to network neutrality, need to be considered in this context as well.

5.3 Public funding agencies

R7: Support QoE research as scientific approach to a substantial and unsolved problem.

For early multimedia services, Quality of Service (QoS) provided a coarse approximation of user-perceived service quality. This has become unsatisfactory especially since the explosive development of new services and applications, each with very different needs. On the other hand, based on the ubiquity of fast Internet access, these services play an ever more important role in the daily life of users and our society. It is therefore essential to ensure that the quality experienced by the users is up to their expectations, both to avoid user frustration, and also its negative impact of business. Hence, QoE has to go far beyond merely being "QoS 2.0", which requires significant on-going interdisciplinary efforts, where – for instance – the "QoE by Design" approach introduced above will provide a significant step forward.

R8: Understand QoE as key paradigm for the future digital society.

New technologies such as virtual and augmented reality, ubiquitous computing and the Internet of Things have a strong potential to improve services in key areas of our society, like e-health, ambient assisted living, smart cities, etc., thus improving the quality of life to citizens. However, if these new applications fail to meet the quality requirements and expectations of their users, their impact may be severely limited, and worse yet, it may have negative and even fatal consequences (in critical areas such as telemedicine, or self-driving vehicles). Hence, we suggest supporting industrial or research endeavours that lead to openly accessible means for implementing the proposed QoE by design approach. At the same time, research into privacy and trust related issues involved in the collection of data for QoE purposes will ensure that the rights of the users are upheld. Finally, from an inclusiveness point of view, QoE technologies will help ensuring that all user groups, including marginalised ones, receive adequate service quality.

R9: Create a cross-disciplinary and cross-institutional research community.

We recommend the promotion of educative actions supporting the formation of new professionals and early-stage researchers in the joint QoE and UX fields, so as to address the needs discussed previously. These efforts should actively involve a broad range of different disciplines, ranging from communication technology to humanities and arts, and should be based on the cooperation of different faculties and/or academic centers.

R10: Support market diversity and sustainability.

QoE and UX are expected to be key aspects for the adoption and sustainability of innovative technologies and services, which will increase user engagement and satisfaction, as well as user acquisition and retention, which in turn will improve the profitability of businesses. Furthermore, easy access to QoE technologies will enable smaller industry actors to differentiate their offerings and be able to compete with larger incumbents. The integration of QoE and

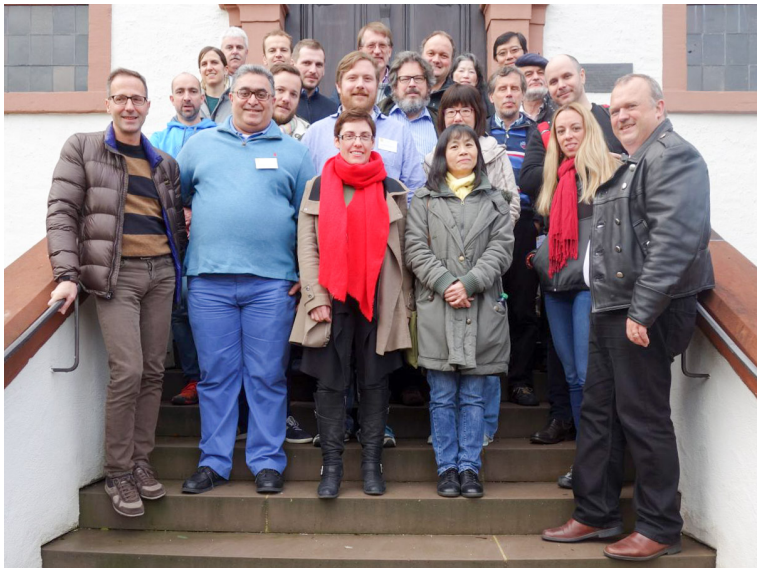
UX will help address the business viability (as per the above), technical feasibility (exploiting QoE enablers) and desirability (considering UX) of new services, and allow for their success.

5.4 Conclusions

Hence, summarising briefly, we believe that further developing QoE has the clear potential to provide a key contribution for the evolution of the future digital society. It will require joining forces both in research and industry through broad interdisciplinarity, enforcing the links between adjacent research areas and communities like QoE and UX, increasing accessibility of data through open data approaches, and integrating innovative methodologies like, for instance, machine learning. Together with the envisaged "Quality by Design" approach and the proposed emphasis on appropriate feedback mechanisms, the "turn to the user" will offer highly promising opportunities for the future networking and service market, which by now has also been acknowledged by the EU in the context of the upcoming "Next Generation Internet" activity.

6 Participants

- Jan-Niklas Antons
TU Berlin, DE
- Luigi Atzori
University of Cagliari, IT
- Katrien De Moor
NTNU – Trondheim, NO
- Touradj Ebrahimi
EPFL – Lausanne, CH
- Sebastian Egger-Lampl
AIT Austrian Institute of Technology – Wien, AT
- Markus Fiedler
Blekinge Institute of Technology – Karlskrona, SE
- Jörgen Gustafsson
Ericsson Research – Luleå, SE
- Tobias Hoßfeld
Universität Duisburg-Essen, DE
- Lucjan Janowski
AGH Univ. of Science & Technology – Krakow, PL
- Kalevi Kilkki
Aalto University, FI
- Udo Krieger
Universität Bamberg, DE
- Effie Lai-Chong Law
University of Leicester, GB
- Sebastian Möller
TU Berlin, DE
- Marianna Obrist
University of Sussex – Brighton, GB
- Peter Reichl
Universität Wien, AT
- Virpi Hannele Roto
Aalto University, FI
- Henning Schulzrinne
Columbia University – New York, US
- Lea Skorin-Kapov
University of Zagreb, HR
- Jan Van Looy
Ghent University, BE
- Martín Varela
VTT Technical Research Centre of Finland – Oulu, FI
- Katarzyna Wac
University of Geneva, CH
- Felix Wu
University of California – Davis, US
- Min Xie
Telenor Research – Trondheim, NO
- Hans-Jürgen Zepernick
Blekinge Institute of Technology – Karlskrona, SE



References

1 European Network on Quality of Experience in Multimedia Systems and Services (COST IC 1003 Qualinet). URL: http://www.qualinet.eu/.

2 Springer Journal Quality and User Experience (QUEX). URL: https://www.editorialmanager.com/quex/mainpage.html.

3 Arslan Ahmad, Alessandro Floris, and Luigi Atzori. QoE-centric service delivery: A collaborative approach among OTTs and ISPs. *Computer Networks*, 110:168–179, 2016. doi:10.1016/j.comnet.2016.09.022.

4 Patrick Le Callet, Sebastian Möller, and Andrew Perkis, editors. *Qualinet White Paper on Definitions of Quality of Experience, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*. White paper, March 2013. Version 1.2 Novi Sad. URL: http://www.qualinet.eu/index.php?option=com_content&view=article&id=45&Itemid=52.

5 Markus Fiedler, Tobias Hoßfeld, and Phuoc Tran-Gia. A generic quantitative relationship between Quality of Experience and Quality of Service. *IEEE Network*, 24(2):36–41, 2010. doi:10.1109/MNET.2010.5430142.

6 Dennis Guse. *Multi-episodic Perceived Quality of Telecommunication Services*. Doctoral dissertation, Technische Universität Berlin, 2016. doi:10.14279/depositonce-5499.

7 Tobias Hoßfeld, Raimund Schatz, Martín Varela, and Christian Timmerer. Challenges of QoE management for cloud applications. *IEEE Communications Magazine*, 50(4):28–36, 2012. doi:10.1109/MCOM.2012.6178831.

8 ITU-T Rec. E.800. Definitions of terms related to Quality of Service, September 2008. URL: https://www.itu.int/rec/T-REC-E.800-200809-I/en.

9 Ute Jekosch. *Voice and Speech Quality Perception – Assessment and Evaluation*. Signals and Communication Technology. Springer, Berlin, Heidelberg, 2005. doi:10.1007/3-540-28860-0.

10 Satu Jumisko-Pyykkö. *User-Centered Quality of Experience and Its Evaluation Methods for Mobile Television*. Doctoral dissertation, Tampere University of Technology, 2011.

11 Satu Jumisko-Pyykkö, Dominik Strohmeier, Timo Utriainen, and Kristina Kunze. Descriptive Quality of Experience for mobile 3D video. In Ebba Þóra Hvannberg, Marta Kristín Lárusdóttir, Ann Blandford, and Jan Gulliksen, editors, *Proceedings of the 6th Nordic Conference on Human-Computer Interaction 2010, Reykjavik, Iceland, October 16-20, 2010*, pages 266–275. ACM, 2010. doi:10.1145/1868914.1868947.

12 Maria G. Martini, Chang Wen Chen, Zhibo Chen, Tasos Dagiuklas, Lingfen Sun, and Xiaoqing Zhu. Guest Editorial QoE-Aware Wireless Multimedia Systems. *IEEE Journal on Selected Areas in Communications*, 30(7):1153–1156, 2012. doi:10.1109/JSAC.2012.120801.

13 Alexander Raake and Sebastian Egger. *Quality and Quality of Experience*, pages 11–33. Springer International Publishing, Cham, 2014. doi:10.1007/978-3-319-02681-7_2.

14 Virpi Roto, Effie Law, Arnold Vermeeren, and Jettie Hoonhout, editors. *User Experience White Paper: Bringing Clarity to the Concept of User Experience*. White paper, 2011. Result of the Dagstuhl Seminar 10373. URL: http://www.allaboutux.org/uxwhitepaper.

15 Raimund Schatz, Markus Fiedler, and Lea Skorin-Kapov. *QoE-Based Network and Application Management*, pages 411–426. Springer International Publishing, Cham, 2014. doi:10.1007/978-3-319-02681-7_28.

16 Ina Wechsung and Katrien De Moor. *Quality of Experience Versus User Experience*, pages 35–54. Springer International Publishing, Cham, 2014. doi:10.1007/978-3-319-02681-7_3.

Manifesto from Dagstuhl Perspectives Workshop 16152

Tensor Computing for Internet of Things

Edited by

Evrim Acar[1], Animashree Anandkumar[2], Lenore Mullin[3], Sebnem Rusitschka[4], and Volker Tresp[5]

1 **University of Copenhagen, DK**
evrim@life.ku.dk
2 **University of California – Irvine, US**
a.anandkumar@uci.edu
3 **University of Albany – SUNY, US**
lmullin@albany.edu
4 **Siemens AG – München, DE**
sebnem.rusitschka@siemens.com
5 **Siemens AG – München, DE**
volker.tresp@siemens.com

Abstract

"The fundamental laws necessary for the mathematical treatment of large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved." – Dirac 1929

The digital world of Internet of Things (IoT) will provide a high-resolution depiction of our physical world through measurements and other data - even high-definition "video," if you consider streaming data frames coming from a myriad of sensors embedded in everything we use. This depiction will have captured our interactions with the physical world and the interactions of digitally enhanced machines and devices. Tensors, as generalizations of vectors and matrices, provide a natural and scalable framework for handling data with such inherent structures and complex dependencies. Scalable tensor methods have attracted considerable amount of attention, with successes in a series of learning tasks, such as learning latent variable models, relational learning, spatio-temporal forecasting as well as training and compression of deep neural networks.

In a Dagstuhl Perspectives Workshop on Tensor Computing for IoT, we validated the fundamental suitability of tensor methods for handling the massive amounts of data coming from connected cyber-physical systems (CPS). The multidisciplinary discourse among academics, industrial researchers and practitioners in the IoT/CPS domain and in the field of machine learning and tensor methods, exposed open issues that need to be addressed to reap value from the technological opportunity. This Manifesto summarizes the immediate action fields for advancement: IoT Tensor Data Benchmarks, Tensor Tools for IoT, and the evolution of a Knowledge Hub. The activities will also be channeled to create best practices and a common tensor language across the disciplines.

In a not so distant future, basic infrastructures for living will be mainly data-driven, automated by digitally enhanced devices and machines. The tools and frameworks used to engineer such systems will ensure production-ready machine learning code which utilizes tensor-based, hence better interpretable, models and runs on distributed, decentralized, and embedded computing resources in a robust and reliable way. We conclude the manifesto with a strategy how to move towards this vision with concrete steps in the identified action fields.

Perspectives Workshop April 10–13, 2016 – http://www.dagstuhl.de/16152

Tensor Computing for Internet of Things, *Dagstuhl Manifestos*, Vol. 7, Issue 1, pp. 52–68
Editors: Evrim Acar, Animashree Anandkumar, Lenore Mullin, Sebnem Rusitschka, and Volker Tresp

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

2012 ACM Subject Classification Computing methodologies → Machine learning, Computer systems organization → Embedded and cyber-physical systems, Hardware → Emerging tools and methodologies

Keywords and phrases Distributed Systems, Real-time and embedded systems, Signal processing systems, Learning, Multiagent systems

Digital Object Identifier 10.4230/DagMan.7.1.52

Executive Summary

Cyber-physical systems (CPS), or the more consumerized Internet of Things (IoT) is a new wave of embedding affordable computing and communication into our previously mechanized world to enable for example adaptive energy efficient buildings fueled by renewable energy sources and connected to smart power grids, factory automation that yields flexible manufacturing and zero down-time connected to adaptive global supply chains, multi-modal on-demand public transportation facilitated by car-sharing and even self-driving cars in the near future.

After years of industrial research, we can pinpoint with confidence that all of the above scenarios of IoT have the following common requirements emerging from a set of common characteristics, i.e., they all require the extraction of actionable information for near real-time automation from multidimensional, spatio-temporal data. This data is only partially stochastic, as much as humans are involved as the users and operators. But mostly the data comes from a human-engineered, but mechanically, increasingly digitally, automated network such as electricity networks, supply chains/networks, transportation networks – commonly referred to as flow networks. The digitalization of these flow networks is what we refer to as CPS or IoT. Such digitalization includes ever more precise sensors, cheaper embedded computing, ubiquitous connectivity, combined with massive amounts of historical data and easy-to-spawn compute clusters in global data centers.

In April 2016, Dagstuhl hosted a Perspectives Workshop on Tensor Computing for the Internet of Things by bringing together academic researchers from the tensor community, distributed computing and machine learning as well as industrial researchers and practitioners from the IoT/CPS domain. The goal of the workshop was to explore the tensor representations and tensor computing as the basis for the machine learning solutions needed to turn massive amounts of IoT/CPS data into useful and actionable information. Tensors, as generalizations of vectors and matrices, provide a natural representation for data with many axes of variation, e.g., multidimensional, spatio-temporal data. The workshop validated the suitability of tensor-based computation for handling data coming from IoT/CPS and concluded with a vision that tensors would be a crucial part of a bigger computational machinery supporting the domain experts of IoT/CPS in the near future and supporting the machines and devices in IoT/CPS in the long term. This manifesto discusses the immediate action areas, i.e., IoT Tensor Benchmark Data & Infrastructure, Tensor Tools for IoT, and Tensor Learn – a knowledge hub, to move towards this vision, and concludes with strategic steps to be taken within the three action areas.

The manifesto is intended for government and industry funding agencies as well as academic and industrial researchers. The manifesto will draw the attention of funding agencies to the open issues needed to be addressed for utilizing the massive amounts of IoT/CPS data from a data science perspective by pointing to tensor computing as a crucial tool. The manifesto will also address to academic and industrial researchers by emphasizing

the open research directions in tensor computing as well as in its use for production-level development and deployment in IoT/CPS.

Table of Contents

1 Introduction

In April 2016, Dagstuhl hosted a Perspectives Workshop on Tensor Computing for the Internet of Things [2]. The prior year, industrial researchers had formulated the challenges of gaining insights from multi-dimensional sensory data coming from large-scale connected energy, transportation networks or manufacturing systems. The sheer amount of streaming multi-aspect data was prompting us to look for the most suitable techniques from the machine learning community: multi-way data analysis.

The workshop focused on the Internet of Things (IoT), i.e. devices, which have the capability to sense, communicate, and even control or interact with their environments. These devices are increasingly becoming parts of complex, dynamic, and distributed systems of electricity or mobility networks, hence our daily lives. Various sensors enable these devices to capture multiple aspects of their surroundings in real-time. For example, phasor measurement units capture transient dynamics and evolving disturbances in the power system in high-resolution, in a synchronized manner, and in real-time. Another example is traffic networks, where a car today can deliver about 250 GB of data per hour from connected electronics such as weather sensors within the car, parking cameras and radars. Experts estimate that the IoT will consist of almost 50 billion objects by 2020 [36], which will trigger the Era of Exascale computing necessitating the management of heat and energy of computing in concert with more and more complex processor/network/memory hierarchies of sensors and embedded computers in distributed systems.

Crucial for the extraction of relevant information is the format in which the raw data from such systems is represented. Crucial for the practicability of information extraction in IoT is which and how operations are used guaranteeing various attributes of resource use and management. Tensors can be viewed as both multidimensional data structures and as multilinear operators. The goal of the workshop was to explore tensor representations and computing as the basis for machine learning solutions for the IoT. Tensors are algebraic objects which describe linear and multilinear relationships, and can be represented as multidimensional arrays. They often provide a natural and compact representation for multidimensional data. In the recent years, tensor and machine learning communities - mainly active in the data-rich domains such as neuroscience, social network analysis, chemometrics, knowledge graphs etc. - have provided a solid research infrastructure, reaching from the efficient routines for tensor calculus to methods of multi-way data analysis, i.e., tensor decompositions, to methods for consistent and efficient estimation of parameters of the probabilistic models.

Some tensor-based models have the intriguing characteristic that if there is a good match between the model and the underlying structure in the data, the models are much better interpretable than alternative techniques. Their interpretability is an essential feature for the machine learning techniques to gain acceptance in the rather engineering heavy fields of automation and control of cyber-physical systems (CPS). Many of these CPS show intrinsically multilinear behavior, which is appropriately modeled by tensor methods and tools for controller design can use these models. The calibration of sensors delivering data and the higher resolution of measured data will have an additional impact on the interpretability of models.

Various presentations on tensor methods by established researchers at the workshop from different application domains assured us that tensor methods are reaching a maturity tipping point. However, knowledge of usage characteristics of tensor models is scattered. Discussions of the currently independent perspectives on the usage of tensor methods showed a potential for convergence, which we want to leverage through the action areas we are describing in

this Dagstuhl Manifesto. During our discussions based on the presentations of the IoT industrial researchers, it quickly became clear that we would need benchmark challenges for cyber-physical systems and benchmark data in order to be able to replicate the successes in machine learning for object recognition and natural language understanding.

The tensor computing community will equally benefit from the new types of data, requirements, and multi-aspect characteristics of IoT, which can lead to techniques that increase success rates of previous applications of tensor methods, as was the case with the challenges of social network data analysis leading to better tensor models/algorithms that can analyze data sets with missing entries, now used in many other fields in addition to social network analysis. Additionally, as opposed to standardized machine learning techniques, tensor computing currently lacks a common language and the homogeneity to flexibly exchange models. Hence, a hub platform bringing data and domain knowledge of cyber-physical systems together with various practitioners of tensor computing would enhance increasing coherence of terms, best practices in data acquisition and structuring methods as well as model benchmarking, cataloging, and exchange of methods.

In the following the Manifesto describes the three action fields of Benchmarks, Tools, and Knowledge Hub, when put together, will make tensors a crucial part of a bigger computational machinery. This machinery will enable first domain experts of IoT/CPS and at a future time also the machines and devices in IoT/CPS to create efficient and sustainable infrastructures for life. We conclude the Manifesto with this vision and a strategy how to move into the right direction now.

2 IoT Tensor Benchmark Data & Infrastructure

Availability of benchmark data has been one of the reasons behind the recent advances in machine learning, e.g. large collections of high-resolution imagery for image recognition in computer vision tasks - or large corpus of written and spoken text for applications that need natural language processing. Although the special - multi-relational - structure of data is at the heart of tensor decompositions, there are no dedicated benchmark tensor data sets. Benchmark data typically is chosen to shed light on an algorithm's critical performance aspects and compare it to other algorithms. Well-known problems with this approach are the problem-specificity and that the computational performance and scalability remain still untested for larger real-world problem data sets.

IoT may indeed bring with it the much needed tensor data in a benchmarkable environment for tensor computing. Until now most effective data sets are known to be from chemometrics, telecommunications networks, neuroscience and social networks. Chemometrics data mainly represent a "closed" environment, e.g. the make-up of a fluid consisting of multiple components with different spectra. Application of tensor decompositions allows for interpretable factorization and analysis results in such closed environments. Cyber-physical systems are made up of such closed environments, which connected to each other build wider networks/systems. Examples are IoT data sets on home energy usage, in which the multi-aspect measurements of power parameters at the home breaker box capture the varying characteristic spectra of all the electrical appliances contributing to a home's energy usage, a "closed" environment. Hence, the application of tensor decompositions to the problem of so called non-intrusive load monitoring should yield similar interpretable analytic results as in chemometrics applications. Furthermore a local power grid network consists of multiple such closed environments, and links to other local grids to make up the bigger power distribution and transmission system.

Similar connectivist view of other CPS domains - such as in manufacturing with factories and supply chains, or in mobility with connected vehicles and multimodal transportation systems, etc. - and the promising nature of tensor methods motivates researchers and potential data providers to organize so called "IoT Tensor Data Challenges," which will

- accommodate larger data sets on real-world problems of IoT/CPS,
- curate for high-accuracy and high-resolution sensor data,
- from a "closed" environment such that factorization yields interpretable results, as well as
- have the potential to capture larger networks in the data

through the inherent connectivity of the IoT/CPS data challenge. The issues with current benchmark data collections should be addressed by standardizing the "IoT Tensor Data Benchmark Infrastructure" with following research & development aspects:

- The infrastructure should enable users to filter problems based on technical similarity, e.g. spatio-temporal problems, multi-class predictions etc. The infrastructure should also enable to browse across others' implementation of algorithms and compare effectively.
- In addition to prediction accuracy, key performance indicators of benchmarking for IoT/CPS applications are interpretability, computational resource consumption, robustness in stream processing and potentially in highly distributed settings.
- The interface to the infrastructure should also enable users to access metadata and analyze metadata to understand how the algorithms perform, e.g. computation cost per training, per prediction, etc.
- Additionally, the interface should enable users to easily understand how different tensor models and algorithms perform in different scenarios.

The organization of IoT Tensor Data Challenges will require coordinated efforts of this community and their extended network. Whereas especially data from industrial partners will be handled with care, and confidentially if required, as to lower the barriers to providing data for the challenge. The design and development of the IoT Tensor Data Benchmark Infrastructure requires an open and iterative approach, which will be improved with every data challenge.

3 Tensor Tools for IoT

Data in many disciplines contains more than two axes of variation, e.g., spatial, temporal and spectral dimensions of multi-channel electroencephalography (EEG) signals represented in both time and frequency domains [3, 27], and can be represented as a multi-way array, also referred to as a higher-order tensor. Exploiting the low-rank structure and capturing the underlying patterns in such higher-order data sets are crucial in some domains in order to extract information from complex data sets. Therefore, tensor factorizations, i.e., extensions of matrix factorizations to multi-way data, have proved useful in a variety of applications, in particular, in chemometrics, neuroscience, signal processing and data mining [4, 23, 34, 32].

In this section we discuss and identify open research questions in two parts: (a) regarding models and algorithms and (b) regarding development for and deployment of these models and algorithms in IoT/CPS.

Models and algortihms. Tensor factorizations have become a popular data mining tool in the last decade. Inter-disciplinary conferences of the tensor community such as TRICAP (Three-way Methods in Chemistry and Psychology) and TDA (Tensor Decompositions and Applications) as well as workshops sponsored by AIM (American Institute of Mathematics)

and NSF (National Science Foundation) have played a key role in promoting and advancing the field by bringing experts from different fields together to identify and solve issues in tensor computing. Significant efforts have been invested in developing tensor factorization models, building algorithms and finding the right tensor models for applications of interest. Among the variety of tensor factorization approaches, the CANDECOMP/PARAFAC (CP) model [17, 11] has proved useful in applications, where the goal is to capture the underlying factors uniquely and use them for interpretation. As a result of its uniqueness properties leading to easily interpretable models, CP has been successfully used in neuroscience, chemometrics, social network analysis and signal processing applications. The CP model has strong assumptions about the underlying structure of the multi-way data, i.e., each slice of the tensor should have the same factors but in different proportions. If there is a good match between the data and the CP model, it is possible to summarize the data in a compact, unique and meaningful way. If the data does not follow a CP model, more flexible tensor factorization models such as a Tucker model [37] can be used for exploratory data analysis. Also, in particular, when the goal is data compression, Tucker-based approaches have proved to be effective. In addition to CP and Tucker models, there are many tensor models (see surveys/books on tensor factorizations [4, 23, 34, 16, 32]), which may be preferred depending on the goal of the application and the underlying structure of the data sets of interest.

While the analysis of data emerging from IoT/CPS applications will benefit from the expertise of the tensor community, new types of data and requirements of the applications will also call for further developments in tensor computing. The CPS/IoT systems are real-time, distributed, networked, and show dynamic behavior. The data coming from the sensors embedded into these systems is streaming, noisy, both high-frequency and high-volume, both sparse and dense. We have identified the following open problems in tensor computing as the challenges to primarily focus on in order to make tensor computations effective tools in IoT/CPS applications:

- Developing efficient streaming tensor models that can analyze real-time data,
- Building algorithms scalable to high-volume data (for both sparse and dense),
- Developing efficient distributed models and algorithms,
- Automating the building blocks of tensor modeling, e.g., model order selection, model selection, to decrease expert inputs in the analysis of IoT/CPS data,
- Uncertainty quantification of model parameters for tensor factorizations,
- Introducing new visualization methods in order to increase the interpretability of tensor factorizations,
- Developing data fusion models and their streaming versions that can jointly analyze coupled heterogeneous data sets, i.e., data sets in the form of matrices and higher-order tensors,
- Building tensor factorization models that can incorporate prior knowledge such as the connectivity structure (topology) of IoT systems,
- Forming a common tensor computing language to facilitate the exchange of expertise.

Development and Deployment. During the workshop we also had the opportunity to exchange on trends in tensor tools and emerging frameworks, which focus on development and deployment support for production-level code. Tensor tools have come a long way since the first version of Tensor Toolbox for Matlab over a decade ago [6].

Whilst new tools for Matlab have emerged with more focus on modularity, documentation and getting users from other research fields up to speed on using tensors [38], more specialized implementations such as Tensor Trains [31] or a distributed version of Tucker computations [22] are increasingly being shared on github as open source. Open source does speed up

research immensely since code and papers are instantly accessible to investigate, learn from, and build upon. We believe that this trend will also assist in disseminating and in creating the common tensor computing language across disciplines. Especially, when data scientists will start adopting and porting some of these tensor-based models and algorithms for use in their favorite programming language and numerical computation libraries.

Matlab is popular with mathematicians and scientists. However, data scientists and machine learning researchers rarely use Matlab. Instead for the longest time Theano [9], a numerical computation library in Python, has been the most popular open source framework. Theano's focus has been deep learning and efficient computations utilizing GPUs. Since November 2015, Google open sourced their numerical computation library called Tensorflow [1], which since then gained considerably in popularity. Tensorflow has Python bindings, whilst the core is written in C++. Tensorflow aims to enable the creation of maintainable production-ready code, which runs on distributed machines, hence highly targeted towards industrial data scientists and applications which deal with massive amounts of data that no single analytics machine can handle effectively. In this latter category of industry-focused tools another framework worth mentioning exists: Deeplearning4J [21]. Deeplearning4J is also a distributed deep learning framework suitable for major companies and large government organizations, which to date still heavily rely on Java or a JVM-based system. Both Tensorflow and Deeplearning4J are designed for use with distributed data management and processing systems such as open source Hadoop and Spark [35] or in the case of Tensorflow also naturally with Google's proprietary cluster scheduling system called Borg [15].

Researchers in the intersection of tensor computing and machine learning have been implementing and open sourcing tensor methods in Python [28] for use in Python projects, or in Scala [18] for use with Spark, or in Julia [5], a language designed to address the needs of high-performance numerical analysis and computational science while also being effective for general-purpose programming, just to name a few. This is a typical sign of the search for a dominant design in this newly converging field of machine learning and tensor computing. A potential research & development direction is to create an abstraction layer. A well-designed API would allow to build tensor-based learning models by clipping together high-level building blocks of tensor decompositions and similar methods. The abstraction layer would be placed on top of numerical computation libraries like Tensorflow or Deeplearning4J etc. TensorLab has such a layer built-in but currently it is only on top of MATLAB.

At this point it is hard to predict, which languages and frameworks will prevail after more experience has been gained in the intersection of machine learning and tensor computing. Yet, the domain of IoT/CPS additionally demands the code deployment to be lightweight and the programming language to be robust and efficient for embedded processors. Java is inherently cross-platform, there is an embedded variant and OSGi suitable for some IoT application classes. However, in CPS domains where the insights gained from tensor decompositions shall translate into controller actions and other near real-time optimization, performance will be the crucial factor. Whilst C++ as a systems programming language seems to be the natural choice, it must be noted that C++ is difficult to optimize and maintain.

The skill set that can break down tensor-based machine learning models and algorithms - even if only for inference - into reliable, high-performant, embedded code is very rare. This realization is a definitive call for developing of frameworks that support the developers. Tensorflow is the only framework, at the time of this writing, which is used in production and supports direct deployment of trained models in embedded and mobile devices [39]. During research for the compilation of the Manifesto, we also found a new machine learning framework called Leaf [26] written in Rust, which is an up and coming safe and parallel

systems programming language that is easy to write and deploy. Interestingly, the initial performance benchmarks affirm our discussions that Tensorflow may be too memory-intensive for embedded environments. Ironically, Leaf's development concluded in May 2016 due to the rapidly increasing popularity of Tensorflow.

One very important realization which is just beginning to surface in the research community is that all of these frameworks depend on the same low-level libraries such as BLAS for efficiently performing linear algebraic routines. BLAS is a library from the 70s, which has added so-called levels over the years for vector operations (Level 1) for matrix-vector operations (Level 2) and for matrix-matrix operations (Level3). BLAS Level 1 operations are computed in linear time, Level 2 in quadratic and Level 3 operations are computed in cubic time. Tensor operations have traditionally been implemented in terms of BLAS operations, e.g. Matrix Multiplication, incurring both a performance and a storage overhead because tensors must be flattened to use matrix-matrix operations and this procedure is repeated multiple times depending on the model/algorithm, the dimension of the data as well as layout of caches and processors of the hardware. This typical memory blowup problem might have been a niche problem until now, but the more data is being processed and the faster analytics result are being expected, the more critical it will become [8] [25]. A promising abstraction we came across during the research after our Workshop is BLIS [40]. The BLIS framework is not a single library or static API, but rather a nearly-complete template for instantiating high-performance BLAS-like libraries.

At the hardware level most of the frameworks again depend on the same abstractions for translating the algebraic routines onto machine instruction sets through libraries such as CUDA and OpenCL. Whilst CUDA is a software layer that gives direct access to the GPU's virtual instruction set and parallel computational elements for NVIDIA hardware, OpenCL aims to deliver comparable abstraction across heterogeneous platforms consisting of central processing units (CPUs), graphics processing units (GPUs), digital signal processors (DSPs), field-programmable gate arrays (FPGAs) and other processors or hardware accelerators. In the application domain of IoT/CPS we have heterogeneous architectures across hierarchies of processors, memory, and network. In our discussions surrounding the workshop, we even questioned traditional processor architectures with hardware managed cache hierarchy, a design principal also from the 70s.

Indeed we are starting to see more innovation even at the processor level, because the cost of moving data across hardware-managed memory layers starts to dwarf the useful computation with that data. This difference was not significant in the early days of computing, and was remedied by scaling techniques via increasing processor clock frequencies and now increasing the number of cores integrated on a single chip. However, the difference in energy used for moving data to the computation versus the energy used for the computation itself becomes very costly when we have machine learning from massive amounts of data. The cost increase is exponential when tensor operations on multidimensional data are necessary. Google, accompanying the open sourcing of their Tensorflow framework for machine learning, unveiled the Tensor Processing Unit (TPU) [10], a custom application-specific integrated circuits (ASIC) built specifically for machine learning. TPUs "only" utilize a clever trick for optimizing performance per watt by allowing the chip to be more tolerant of reduced computational precision, which means it requires fewer transistors per operation. Others redesign processors from the ground up such as the NEO chip from REX Computing [12]. The design of NEO relies on a range of hardware simplifications which are focused on exposing low level functionality. Once a feature exists in software, the reshaping of the tensor could be fused with internal layout of data and packing operations, requiring no explicit reshaping operations or additional workspace and memory.

In summary, we believe three R&D directions will crystallize in the following years in the intersection of mass data-driven machine learning, tensor computing, and IoT/CPS:

- High-level building blocks of tensor decompositions to be used on top of lower level numerical computation libraries
- Basic multilinear algebraic libraries with optimized tensor operations for the currently heterogeneous processor architectures
- New processor architectures redesigned to fundamentally improve balance between extreme efficiency and reconfigurability

As a Tensor Computing for IoT community we will closely follow and co-develop in these R&D directions to also feed in the IoT/CPS requirements for reliability, safety and robustness in highly distributed systems.

4 Tensor Learn - Knowledge Hub

In a recent publication [33], co-authored by two of our participants, the authors state that "After two decades of research on tensor decompositions and applications, the senior co-authors still couldn't point their new graduate students to a single 'point of entry' to begin research in this area." There is this need to provide a comprehensive and deep overview to young researchers and practitioners that will enable them to start developing related algorithms and applying them also to IoT/CPS.

At the same time, another one of our participants has been recently recognized for the two decades of dedication to transforming the process and food industry through actionable insights gained by applying and refining tensor decomposition techniques on multi-way chemometrics data collected in manufacturing facilities. There is this reward for both researchers, industrial practitioners, as well as the society and organizations supporting them - especially "in a time when there is a flood of data, but not the resources to draw out valuable and socially beneficial information from it" [14].

In the few months since the Dagstuhl workshop, one of the organizers joined Amazon's Machine Learning team as principal research scientist, one organizer was called upon as an advisor for the development of a new embeddable chip to disrupt exascale computing, and yet another started her company to enable clean electricity usage and exchange at zero-marginal cost through data-driven automation. These industrial activities signal not only the renaissance but also to some extent the viability of tensor methods for dealing with massive amounts of data coming from an increasingly digitalizing world.

By reviving the tensor decomposition application fields through varied challenges and high-quality data from IoT/CPS, and by creating a focal point of knowledge consolidation and dissemination, we believe that we can considerably shorten the time for breakthrough research in socially beneficial fields such as energy, mobility, cities, and manufacturing to name a few. Through digitalization these areas will be main sources of massive amounts of data coming from high-precision sensors at higher speeds given the advances in communication and computing infrastructures. Many of the established businesses in these areas, especially small and medium enterprises, which do not have the resources for R&D but face the same data deluge, will highly benefit from educational and open source resources available through this international knowledge hub.

As an initial step towards creating the knowledge hub "Tensor Learn," two of our participants organized a workshop co-located with NIPS [24]. The workshop aimed to draw the attention to this recent renaissance of tensor methods in machine learning, availability of

new tensor numerical computation frameworks, and point towards open research questions. In order to make the extension from workshop to Knowledge Hub, we aim to:

- host IoT Tensor Data Challenges and
- call for multidisciplinary discourse on the tensor applications for machine learning
- whilst establishing the common tensor computing language to facilitate such discourse.

5 Vision & Strategy

In the short-term, tensors will be a crucial part of a bigger computational machinery supporting the domain experts of IoT/CPS due to the ability of tensor frameworks to capture and represent the multi-aspect information within raw data sets and streams. Further along the line, also connected machines and devices will be supported by the same machinery to carry on tasks in dynamic, (near) real-time environments along-side domain experts. Data scientists, data engineers and system engineers are already building pieces of this computational machinery.

We as a community will have reached a first milestone when we eventually can qualify the most heard phrase: "It depends on the data", e.g. through recipes and best practices. For example, in IoT/CPS data is always analyzed over time and space. In IoT, prediction (trending) is very important to detect anomalies that deviate from the prediction; e.g. anomalies in massive streams of IP traffic data coming from interconnected routers, or coming from interconnected machines in factories, or in the future from sensorized streets accommodating self-driving cars. In CPS, additionally the control aspect comes into play: Once connected machines and devices recognize objects and can classify those, then they can learn through reinforcement within safe parameters how to interact with their multi-dimensional environment.

Scalable tensor methods have attracted considerable amount of attention, with successes in a series of learning tasks, such as learning latent variable models, relational learning, spatio-temporal forecasting as well as training [19] and compression [20] of deep neural networks. As a community we want to pave the way towards successful application of these methods in IoT/CPS. Our milestones on this way are to:

- Showcase suitability of tensor methods on real-world data coming from IoT/CPS that have the inherent structures and complex dependencies that result from the networked nature of IoT/CPS.
- Identify the new research problems that dynamic, (near) real-time, and/or safety-critical systems of energy, mobility, factories expose – especially w.r.t. deployment and performant, robust computing.
- Motivate our multidisciplinary network to take on these research problems by contributing to tensor tools and frameworks for production-level development and deployment in IoT/CPS.

In the following we depict the strategic and tactical steps within the three action areas:

Develop IoT Tensor Data Challenge & Infrastructure to be plugged into Tensor Learn knowledge hub by

- Communicating the potential and curating data from
  - open data initiatives of cities and regulated governmental bodies through our extended network [7]
  - crowd-sourced open infrastructure data like opengridmap [29] (power system), openstreetmap [30] (mobility) and

  - open environmental sensing from open data APIs of hardware providers developers, e.g. Enphase solar inverter cloud API, or data on public blockchains, e.g. solar power generation data logged into Electricchain [13]
- Forming partnerships with hardware/sensor providers/users who will benefit from tensor decomposition for improving interpretability of sensor data analytics and for compressed sensing and at the same time can explore how machine learning systems improve with the availability of high-accuracy and high-resolution data.
- Applying for an international research grant that allows us to work together on curating the data and to create and host a benchmarking infrastructure, to extract and share best practices discovered through the challenges.

Open Source contributions to available tensor tools alongside tutorials, recipes and best practices of applications of these tensor methods listed along side the completed data challenges/benchmarks on the Tensor Learn knowledge hub. The established communities of available frameworks that we are extending can become sponsors and partners of the researchers and practitioners of Tensor Computing for IoT.

Promote and position Tensor Learn as a knowledge hub that started as a workshop co-located with NIPS in order to advance the multidisciplinary discourse between tensor computing and its applications in machine learning. In the same manner we will co-locate further workshops with renown IoT/CPS conferences of IEEE, ACM, and the International Federation of Automation and Control (IFAC). Along the way gathering significant curated data challenges and benchmarks for typical tasks in multi-aspect IoT/CPS that can be automated through machine learning in a reliable and interpretable way by utilizing tensor methods.

6 Participants

- Evrim Acar (University of Copenhagen, DK)
- Kareem Aggour (General Electric – Niskayuna, US)
- Animashree Anandkumar (University of California – Irvine, US)
- Rasmus Bro (University of Copenhagen, DK)
- Ali Taylan Cemgil (Bogaziçi University – Istanbul, TR)
- Edward Curry (National University of Ireland – Galway, IE)
- Lieven De Lathauwer (KU Leuven, BE)
- Hans Hagen (TU Kaiserslautern, DE)
- Souleiman Hasan (National University of Ireland – Galway, IE)
- Denis Krompaß (Siemens AG – München, DE)
- Gerwald Lichtenberg (HAW – Hamburg, DE)
- Benoit Meister (Reservoir Labs, Inc. – New York, US)
- Lenore Mullin (University of Albany – SUNY, US)
- Morten Mørup (Technical University of Denmark – Lyngby, DK)
- Axel-Cyrille Ngonga-Ngomo (Universität Leipzig, DE)
- Ivan Oseledets (Skoltech – Scolkovo, RU)
- Renato Pajarola (Universität Zürich, CH)
- Vagelis Papalexakis (Carnegie Mellon University, US)
- Christine Preisach (SAP SE – Walldorf, DE)
- Achim Rettinger (KIT – Karlsruher Institut für Technologie, DE)
- Sebnem Rusitschka (Siemens AG – München, DE)
- Volker Tresp (Siemens AG – München, DE)
- Bülent Yener (Rensselaer Polytechnic Institute – Troy, US)

References

1 Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016. URL: http://arxiv.org/abs/1603.04467, `arXiv:1603.04467`.

2 Evrim Acar, Animashree Anandkumar, Lenore Mullin, Sebnem Rusitschka, and Volker Tresp. Tensor computing for internet of things (dagstuhl perspectives workshop 16152). *Dagstuhl Reports*, 6(4):57–79, 2016. `doi:10.4230/DagRep.6.4.57`.

3 Evrim Acar, Canan Aykut-Bingöl, Haluk Bingol, Rasmus Bro, and Bülent Yener. Multiway analysis of epilepsy tensors. In *Proceedings 15th International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB), Vienna, Austria, July 21-25, 2007*, pages 10–18, 2007. doi:10.1093/bioinformatics/btm210.

4 Evrim Acar and Bülent Yener. Unsupervised multiway data analysis: A literature survey. *IEEE Trans. Knowl. Data Eng.*, 21(1):6–20, 2009. doi:10.1109/TKDE.2008.112.

5 Yun-Jhong Wu Alex Williams. A julia implementation of tensor decomposition algorithms, 2016. (accessed January 13, 2017). URL: https://github.com/JuliaTensors.

6 Brett W Bader and Tamara G Kolda. A preliminary report on the development of matlab tensor classes for fast algorithm prototyping. Technical report, Technical Report SAND2004–3487, Sandia National Laboratories, Livermore, CA, 2004.

7 Laure Le Bars, Edward Curry, Thomas Hahn, and Milan Petkovi? Big data value association, 2015. (accessed January 13, 2017). URL: http://www.bdva.eu/.

8 Muthu Manikandan Baskaran, Benoît Meister, Nicolas Vasilache, and Richard Lethin. Efficient and scalable computations with sparse tensors. In *IEEE Conference on High Performance Extreme Computing, HPEC 2012, Waltham, MA, USA, September 10-12, 2012*, pages 1–6. IEEE, 2012. doi:10.1109/HPEC.2012.6408676.

9 Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. Theano: new features and speed improvements. *CoRR*, abs/1211.5590, 2012. URL: http://arxiv.org/abs/1211.5590, arXiv:1211.5590.

10 Google Cloud Platform Blog. Google supercharges machine learning tasks with tpu custom chip, 2016. (accessed January 13, 2017). URL: https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html.

11 J. Douglas Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young" decomposition. *Psychometrika*, 35:283–319, 1970.

12 REX Computing. The rex neo architecture, 2015. (accessed January 13, 2017). URL: http://www.rexcomputing.com/.

13 eletricchain. Electricchain, 2016. (accessed January 13, 2017). URL: http://www.electricchain.org/.

14 University of Copenhagen Faculty of Science. Food professor rasmus bro receives the first nils foss excellence prize, 2016. (accessed January 13, 2017). URL: http://www.science.ku.dk/english/press/news/2016/food-professor-rasmus-bro-receives-the-first-nils-foss-excellence-prize/.

15 Google. Tensorflow: An open source software library for numerical computation using data flow graphs, 2015. (accessed August 16, 2016). URL: https://www.tensorflow.org/.

16 Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitt.*, 36(1):53–78, 2013.

17 Richard A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory" multi-modal factor analysis. *UCLA working papers in phonetics*, 16:1–84, 1970.

18 Furong Huang. A spectral (third order tensor decomposition) learning method for learning lda topic model on spark, 2016. (accessed January 13, 2017). URL: https://github.com/FurongHuang/SpectralLDA-TensorSpark.

19 Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *CoRR abs/1506.08473*, 2015.

20 Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *CoRR*, abs/1511.06530, 2015. URL: http://arxiv.org/abs/1511.06530, arXiv:1511.06530.

21 Alicia Klinvex, Grey Ballard, Tamara Kolda, Woody Austin, and Hemanth Kolla. Deeplearning4j: Open-source distributed deep learning for the jvm, 2014. (accessed January 13, 2017). URL: https://deeplearning4j.org/index.html.

22 Alicia Klinvex, Grey Ballard, Tamara Kolda, Woody Austin, and Hemanth Kolla. Tuckermpi on gitlab: Open-source software for distributed tucker computations on large-scale dense data, 2017. (accessed January 13, 2017). URL: https://gitlab.com/tensors/TuckerMPI.

23 Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. doi:10.1137/07070111X.

24 Anima Anand Kumar, Maximilan Nickel, Rong Ge, Yan Liu, and Rose Yu. Tensor learn workshop@nips - learning with tensors: Why and how?, December 10, 2016. (accessed January 13, 2017). URL: http://tensor-learn.org/.

25 Devin Matthews. Blas for tensors: What, why, and how?, 2015. (accessed January 13, 2017). URL: https://www.cs.utexas.edu/users/flame/BLISRetreat2015/slides/Devin_BLISRetreat_2015.pdf.

26 Michael Hirn Maximilian Goisser. Open machine intelligence framework for hackers, 2014. (accessed January 13, 2017). URL: https://github.com/autumnai.

27 Fumikazu Miwakeichi, Eduardo Martinez-Montes, Pedro A. Valdés-Sosa, Nobuaki Nishiyama, Hiroaki Mizuhara, and Yoko Yamaguchi. Decomposing EEG data into space-time-frequency components using parallel factor analysis. *NeuroImage*, 22:1035–1045, 2004.

28 Maximilian Nickel. scikit-tensor: Python library for multilinear algebra and tensor factorizations, 2013. (accessed August 16, 2016). URL: https://github.com/mnick/scikit-tensor.

29 Open grid map, 2014. (accessed January 13, 2017). URL: http://opengridmap.com/.

30 Open street map, 2004. (accessed January 13, 2017). URL: https://www.openstreetmap.org.

31 Ivan Oseledets. Tt-toolbox (tt=tensor train) version 2.2.2, 2016. (accessed January 13, 2017). URL: https://github.com/oseledets/TT-Toolbox.

32 Evangelos E. Papalexakis, Christos Faloutsos, and Nicholas D. Sidiropoulos. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM TIST*, 8(2):16:1–16:44, 2017. doi:10.1145/2915921.

33 Nicholas D. Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *CoRR*, abs/1607.01668, 2016. URL: http://arxiv.org/abs/1607.01668, arXiv:1607.01668.

34 Age K. Smilde, Rasmus Bro, and Paul Geladi. *Multi-way Analysis with Applications in the Chemical Sciences*. Wiley, 2004.

35 Apache Spark. Apache spark: A fast and general engine for large-scale data processing, 2012. (accessed August 16, 2016). URL: http://spark.apache.org/.

36 Statista. Internet of things (iot): number of connected devices worldwide from 2012 to 2020 (in billions), 2015. (accessed August 16, 2016). URL: http://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/.

37 Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

38 N Vervliet, O Debals, L Sorber, M Van Barel, and L De Lathauwer. Tensorlab 3.0. *available online*, 2016. URL: http://www.tensorlab.net.

39 Pete Warden. How to quantize neural networks with tensorflow, 2016. (accessed January 13, 2017). URL: https://petewarden.com/2016/05/03/how-to-quantize-neural-networks-with-tensorflow/.

40 Field G. Van Zee and Robert A. van de Geijn. BLIS: A framework for rapidly instantiating BLAS functionality. *ACM Trans. Math. Softw.*, 41(3):14:1–14:33, 2015. doi:10.1145/2764454.

Manifesto from Dagstuhl Perspectives Workshop 15362

Present and Future of Formal Argumentation

Edited by

Dov M. Gabbay[1], Massimiliano Giacomin[2], Beishui Liao[3], and Leendert van der Torre[4]

1 **King's College London, GB and University of Luxembourg, LU**
dov.gabbay@kcl.ac.uk
2 **University of Brescia, IT**
massimiliano.giacomin@unibs.it
3 **Zhejiang University, CN and University of Luxembourg, LU**
baiseliao@zju.edu.cn
4 **University of Luxembourg, LU**
leon.vandertorre@uni.lu

Abstract

Formal Argumentation is emerging as a key reasoning paradigm building bridges among knowledge representation and reasoning in artificial intelligence, informal argumentation in philosophy and linguistics, legal and ethical argumentation, mathematical and logical reasoning, and graph-theoretic reasoning. It aims to capture diverse kinds of reasoning and dialogue activities in the presence of uncertainty and conflicting information in a formal and intuitive way, with potential applications ranging from argumentation mining, via LegalTech and machine ethics, to therapy in clinical psychology. The turning point for the modern stage of formal argumentation theory, much similar to the introduction of possible worlds semantics for the theory of modality, is the framework and language of Dung's abstract argumentation theory introduced in 1995. This means that nothing could remain the same as before 1995—it should be a focal point of reference for any study of argumentation, even if it is critical about it. Now, in modal logic, the introduction of the possible worlds semantics has led to a complete paradigm shift, both in tools and new subjects of studies. This is still not fully true for what is going on in argumentation theory. The Dagstuhl workshop led to the first volume of a handbook series in formal argumentation, reflecting the new stage of the development of argumentation theory.

Perspectives Workshop August 30 to September 4, 2015 – https://www.dagstuhl.de/15362
2012 ACM Subject Classification Computing methodologies → Knowledge representation and reasoning
Keywords and phrases Artificial Intelligence, Knowledge Representation and Reasoning, Multi-Agent Systems, Argumentation, Non-monotonic Logic
Digital Object Identifier 10.4230/DagMan.7.1.69

Present and Future of Formal Argumentation, *Dagstuhl Manifestos*, Vol. 7, Issue 1, pp. 69–95
Editors: Dov M. Gabbay, Massimiliano Giacomin, Beishui Liao, and Leendert van der Torre

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

Executive Summary

Dov M. Gabbay
Massimiliano Giacomin
Beishui Liao
Leendert van der Torre

Diverse kinds of reasoning and dialogue activities can be captured by argumentation models in a formal and still quite intuitive way, thus enabling the integration of different specific techniques and the development of applications humans can trust. Formal argumentation lays on the solid basis of extensively studied theoretical models at different levels of abstraction, efficient implementations of these models, as well as a variety of experimental studies in several application fields.

In order to be able to convert the opportunities of the present into actual results in the future, the formal argumentation research community is reflecting on the current assets and weaknesses of the field and is identifying suitable strategies to leverage the former and to tackle the latter. As an example, the definition of standard modeling languages and of reference sets of benchmark problems are still in their infancy, reference texts for newcomers are missing, the study of methodological guidelines for the use of theoretical models in actual applications is a largely open research issue.

From August 30 to September 4, 2015, twenty-two world leading experts in formal argumentation from 10 countries was gathered to develop an analysis of the current state of the research in this field and to draw accordingly some strategic lines to ensure its successful development in the future.

The program included first individual presentations on introductory overviews, logical problems and requirements for formal argumentation, specific formalisms and methodologies, relationship between various approaches and applications. Collective discussions on general issues then arose from individual presentations, mainly focusing on four topics, i.e. basic concepts and foundations, specific formalisms for argumentation, algorithms, and connections both inside the argumentation field and with outside research topics. In the end, discussion groups were aimed at identifying the most important open problems in argumentation. Many of them concerned foundational issues of the theory, e.g, how to formally represent various kinds of arguments and how to identify sets of postulates on the reasoning activity over arguments in specific contexts. However, the relationship between argumentation and other research fields (e.g. natural language processing, machine learning, human computer interaction, social choice) was seen to be of major importance, especially to develop more mature applications.

This document summarizes the discussions and results of the Dagstuhl Perspectives Workshop. We first present the many faces of formal argumentation, highlighting the role of formal argumentation in various disciplines. Then, we introduce the

- state-of-the-art of theories and algorithms of formal argumentation formulated in details in a Handbook of Formal Argumentation, including Dung's abstract argumentation and its extensions, structured argumentation systems (ASPIC$^+$, DeLP, ABA and deductive argumentation), as well as a view on applications with special emphasis on the issue of mining arguments from natural language sources.
- Argumentation mining. Thereafter, we introduce the important roles that formal argumentation has played in the field of artificial intelligence. Finally, we discuss

- challenges and future developments. We identify challenging problems from some important perspectives, including theoretical foundations, and connections between formal argumentation and other areas. Moreover, we provide some methodological considerations for future development.

Table of Contents

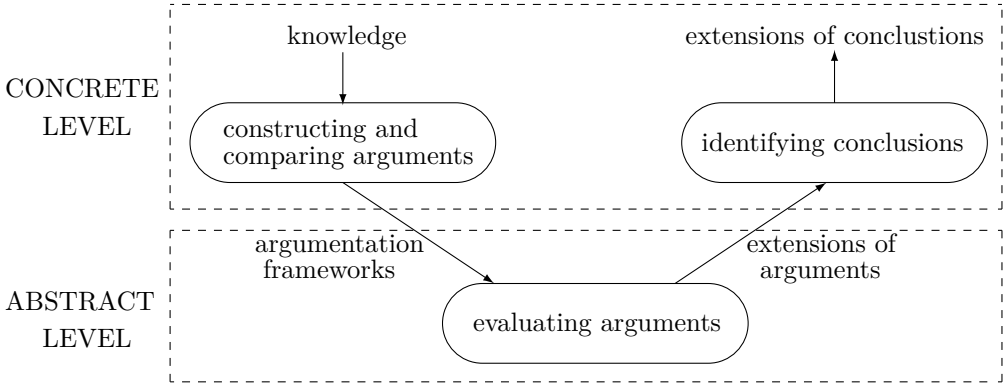


Figure 1 The architecture of a Dung style abstract argumentation system.

1 Introduction to formal argumentation

The Dagstuhl Perspectives Workshop 15362 "Present and Future of Formal Argumentation" was held between August 30 to September 4, 2015, with 22 participants from 10 countries. The goal of this Dagstuhl Perspectives Workshop was to gather the world leading experts in formal argumentation in order to develop an analysis of the current state of the research in this field and to draw accordingly some strategic lines to ensure its successful development in the future. The attendees summarized the state-of-the-art, identified a set of challenging problems, and pointed out possible research directions, ranging from clarifying foundational issues of the theories developed in the literature to integrating argumentation with other research fields, especially in an application-oriented perspective. Following the workshop, the participants have contributed to the first volume of the handbook series of formal argumentation that appeared in 2018, and they are currently involved in the preparation of the second volume.

Formal argumentation is concerned with formalisms for capturing the reasoning in the context of disagreement. We briefly introduce some basic notions of formal argumentation. In general, the study of argumentation is concerned with how assertions are proposed, discussed, and resolved in the context of disagreement [4]. The disagreement or inconsistency may arise during the process of reasoning of an individual agent, or a set of agents interacting each other. In different cases, the nature of inconsistency may vary. In the process of epistemic reasoning and belief revision, the inconsistency of information is mainly due to the uncertainty and incompleteness of information. In the case of practical reasoning such as decision-making or planning, an agent may have several motivations like desires and obligations. Due to the limitation of resources, the agent cannot fulfil all of them, and the conflicts among different motivations arise. In means-end reasoning, there exist different options, which can be mutually exclusive. In the case of inter-agent communication, such as negotiation and discussion, the interests, objectives, preferences or standpoints of different participants might be inconsistent.

Traditional and informal argumentation is concerned with the evaluation of individual arguments. In contrast, Dung introduced his theory of abstract argumentation in which the evaluation of the status of arguments does not depend on the internal structure of the arguments at all, but only on the relation among the arguments with other arguments, and the

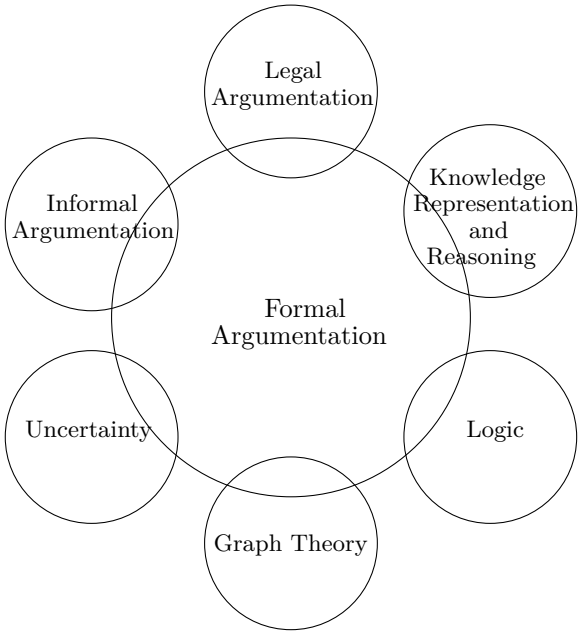


Figure 2 Bridges from formal argumentationn.

status of these related arguments. Consider his two-level architecture illustrated in Figure 1, taken from Liao [17]. In this architecture, the working process of an argumentation system is composed of three steps. First, on the basis of an underlying knowledge base (or from natural text), a set of arguments are constructed and the attacks between them are identified. They form a so-called argumentation framework, which is an abstract representation of arguments and their relationships. Second, given an argumentation framework, the status of arguments is evaluated in terms of a number of criteria, producing sets of extensions of arguments. Each extension may be understood as a set of arguments that are acceptable together. Third, for each extension of arguments the associated set of conclusions is identified, and the justification status for each conclusion is determined on the basis of these sets.

2 Interdisciplinary aspects of formal argumentation

Formal argumentation and formal logic play a central role in the foundations of various disciplines, and they are therefore often used as the methodology for interdisciplinary research projects. Before going into the details of the modern stage of formal argumentation, we highlight the role of formal argumentation in various disciplines, as illustrated in Figure 2. Note that in this figure, we only indicate the overlap between formal argumentation and other disciplines. Rather than giving a comprehensive review, we just provide some examples to show the possibilities and usefulness of bridging formal argumentation with various disciplines.

2.1 Informal argumentation in philosophy and linguistics

Maybe most obviously, formal argumentation can be considered as a candidate for the foundations or theory underlying informal argumentation in philosophy and linguistics. In 1965, Toulmin's much cited book "the uses of argument" led to a criticism on the use of classical logic for reasoning, and the rise of so-called informal logic [25]. Most of the criticism of Toulmin and colleagues has been addressed by non-monotonic logic and more recently, formal argumentation.

Whereas in informal argumentation the evaluation of single argument plays a central role, in formal argumentation the evaluation of argumentation frameworks is the focal point of discussion. Consequently, relations among arguments play a central role in formal argumentation, such as the notion of attack in Dung's theory. Modern formal argumentation offers a kind of interactive argumentation, where the evaluation of individual arguments is enriched with a theory where the evaluation of arguments depends on the evaluation of other arguments. The principle-based approach studies diversity by distinct acceptance semantics, and principles of these semantics [2].

A main challenge is to bridge informal and formal argumentation, in other words to build informal argumentation on top of the new foundations of formal argumentation. This is far from straightforward. From a methodological perspective, the insights of abstract argumentation are a guide, but Dung's theory should not be used as a straight jacket. Researchers in argumentation are free to generalise and adapt it as needed.

Maybe the most promising application in computational argumentation is argumentation mining [20], which is typically build of argumentation schemes developed in informal argumentation. The challenge is to use the foundations of formal argumentation also in this application.

2.2 Legal and ethical argumentation

Legal practice is build on legal argumentation, both in the two branches of roman and case law. It is combined with other kinds of reasoning such as normative and case-based reasoning. This is most explicit in the court room. Formal argumentation has developed in the artificial intelligence community around the ICAIL conference and the legal expert systems studied in JURIX.

Obviously, formal argumentation can be used to reason about legal rules and norms, to decide conflicts or to deal with uncertainty [8]. It is also well suited to deal with one of the main challenges in legal reasoning, called legal interpretation [18, 4]. Legal informatics and LegalTech receive a lot of attention recently, for example to automate regulatory compliance checking. Furthermore, ethical considerations play a role in law, so it may not be a surprise that formal argumentation can play a role in formal ethics as well, including machine ethics.

2.3 Knowledge representation and reasoning in artificial intelligence

Non-monotonic logic and logic programming were adopted in the early eighties as the main methodology in knowledge representation and reasoning, one of the main subareas of artificial intelligence. In the nineties their role was taken over by answer set programming and formal argumentation. Formal argumentation successfully established itself with a large number

of papers in the main journal in the area, called Artificial Intelligence journal, a dedicated journal called Argument & Computation, and a biannual conference called International Conference on Computational Models of Argument, or COMMA.

The modern stage of formal argumentation identifies in the diversity of argumentation and reasoning approaches a common core: Dung's theory of abstract argumentation. This paradigm shift in formal argumentation shows, roughly, how many forms of reasoning can be characterised at an abstract level as an instance of graph reasoning. As a consequence, formal argumentation has been used since the mid nineties as a general framework to classify reasoning methods, besides non-monotonic logic and logic programming also, for example, instances of game theory and social choice.

Algorithms and game-based decision procedures have been developed, together with a formal analysis based on a principle - approach, and complexity analysis. Moreover, various theories of structured argumentation extend Dung's theory with rules and priorities, and a search for a common theory of structured argumentation is currently the main challenge in the area of formal argumentation. There are many open questions in the foundations of formal argumentation, and we are convinced that insights from other formal areas can be used to further develop the theory.

2.4 Reasoning in mathematical logic and graph-theoretic reasoning

Dung's theory of abstract argumentation deals with binary attack relations, and his argumentation framework is a directed graph. As a consequence, abstract argumentation has a close relation to graph theory. However, the relation with graph-theoretic reasoning is relatively unexplored. Likewise the connection of argumentation with logic and liar paradox is yet to be studied in depth. We discuss them in the open problems sections.

2.5 Probabilistic and fuzzy reasoning

Formal argumentation, including Dung's model and its various extensions, can be viewed as a kind of qualitative approach. In recent years, enriching argumentation with uncertainty and fuzziness has attracted attention, since these two aspects are hardly absent in typical knowledge sources. For instance, when considering arguments in natural language texts, uncertainty and fuzziness pervade them both explicitly and implicitly [3]. The explicit presence of uncertainty and fuzziness is exemplified by statements like "I believe that tomorrow will probably be a bit colder than today", where the qualifier "probably" indicates (in a fuzzy way) that the subject's belief is accompanied by a certain degree of uncertainty, while the term "a bit colder" provides a fuzzy specification of tomorrow's expected temperature.

Given that uncertainty and fuzziness and argumentation live side by side, or even permeate each other, in daily discourse, one might expect that this close relationship has a formal counterpart in the models adopted in formal argumentation research, thus supporting the activities of identification and representation of arguments featuring uncertainty and fuzziness starting from natural language expressions.

Since uncertainty and vagueness can be interpreted in different ways by different measures such as probabilities, possibilities, and fuzziness, various kinds of uncertain and fuzzy argumentation have been proposed. Among them, probability-based approaches, including their concepts, formalisms and computational aspects, have been extensively studied.

Given an argumentation framework $F = (A, R)$, and a probability function p that assigns probabilities to arguments or sets of arguments, a basic question is how to interpret probability. There are mainly two approaches in existing literature. One is called constellations approach (external view). The external view is to think of $p(x)$ as the probability of the predicate "$x \in A$". That is, the probability that the argument x is present in A. It imposes probability externally expressing uncertainty on what the network graph is. Another is called epistemic approach (internal view). The internal probability is where the above numbers signify the value of the argument, such as its truth, its reliability, its probability of being effective, etc. [12]. In the epistemic approach, the topology of the graph is fixed but probabilistic assessments on the acceptance of arguments are evaluated with respect to the relations of the arguments in the graph. The core idea of the epistemic approach is that the more likely it is to believe in an argument, the less likely it is to believe in an argument attacking it [15]. The epistemic approach is useful for modeling the belief that an opponent might have in the arguments that could be presented, which is useful for example when deciding on the best arguments to present in order to persuade that opponent.

3 Foundations of formal argumentation

The first volume of the handbook is concerned with the foundations of formal argumentation. Dung's framework and language constitute a turning point for the modern stage of the formal argumentation theory. This means that nothing could remain the same as before Dung—it should be a focal point of reference for any study of argumentation, even if it is critical about it. The handbook reflects the new stage of the development of the argumentation theory. The main content of the first volume of the handbook is as follows.

3.1 Overview

The first three chapters give a general overview of formal argumentation from different perspectives. In Chapter 1, Frans H. van Eemeren and Bart Verheij position formal argumentation in the scope of the larger research of (informal) argumentation. They point out that argumentation has been studied since Antiquity, and modern argumentation theory took inspiration from these classical roots, with Toulmin's 'The Uses of Argument' [26] and Perelman and Olbrechts-Tyteca's 'The New Rhetoric' [23] as representants of a neo-classical development. In the 1970s, a significant rise of the study of argumentation started, often in opposition to the logical formalisms of those days that lacked the tools to be of much relevance for the study of argumentation as it appears in the wild. In this period, argumentation theory, rhetoric, dialectics, informal logic, and critical thinking became the subject of productive academic study. Since the 1990s, innovations in artificial intelligence supported a formal and computational turn in argumentation theory, with ever stronger interaction with non-formal and non-computational scholars. In this chapter, the authors sketch argumentation and argumentation theory as it goes back to classical times, following the developments before and during the currently ongoing formal and computational turn.

In Chapter 2, Henry Prakken gives a historical overview of formal argumentation in terms of a distinction between argumentation-based inference and argumentation-based dialogue. Systems for argumentation-based inference are about which conclusions can be drawn from a given body of possibly incomplete, inconsistent of uncertain information. They ultimately

define a nonmonotonic notion of logical consequence, in terms of the intermediate notions of argument construction, argument attack and argument evaluation, where arguments are seen as constellations of premises, conclusions and inferences. Systems for argumentation-based dialogue model argumentation as a kind of verbal interaction aimed at resolving conflicts of opinion. They define argumentation protocols (the rules of the argumentation game) and address matters of strategy (how to play the game well). In this chapter, the author reviews the main formal and computational models for both aspects of argumentation, sketches their main historical influences, and discusses some main applications areas.

In Chapter 3, Thomas F. Gordon suggests applying software engineering requirements analysis methods to the development and evaluation of formal models of argumentation. Their aim and purpose is to help assure that formal argumentation models the full scope of argumentation as it is understood and studied in the humanities and social sciences, so as to provide a foundation for software tools supporting real argumentation tasks, in a wide variety of application domains.

3.2 Abstract argumentation

In this part, Pietro Baroni, Martin Caminada, Massimiliano Giacomin first present an overview on the state of the art of Dung's abstract argumentation frameworks and their semantics, covering both some of the most influential literature proposals and some general issues concerning semantics definition and evaluation. As to the former point the chapter reviews Dung's original notions of complete, grounded, preferred, and stable semantics, as well as a variety of notions subsequently proposed in the literature namely, naive, semi-stable, ideal, eager, stage, CF2, and stage2 semantics, considering both the extension-based and the labelling-based approaches with respect to their definitions [1]. As to the latter point the chapter analyzes the notions of argument justification and skepticism comparison and discusses semantics agreement.

Then, Gerhard Brewka, Stefan Ellmauthaler, Hannes Strass, Johannes P. Wallner, and Stefan Woltran describe abstract dialectical frameworks, or ADFs for short. ADFs are generalizations of the widely used Dung argumentation frameworks. Whereas the latter focus on a single relation among abstract arguments, namely attack, ADFs allow arbitrary relationships among arguments to be expressed. For instance, arguments may support each other, or a group of arguments may jointly attack another one while each single member of the group is not strong enough to do so. This additional expressiveness is achieved by handling acceptance conditions for each argument explicitly. The semantics of ADFs are inspired by approximation fixpoint theory (AFT), a general algebraic theory for approximation based semantics developed by Denecker, Marek and Truszcynski. After briefly introducing AFT and discussing its role in argumentation, the authors formally introduce ADFs and their semantics. In particular, they show how the most important Dung semantics can be generalized to ADFs. Furthermore, they illustrate the use of ADFs as semantical tool in various modelling scenarios, demonstrating how typical representations in argumentation can be equipped with precise semantics via translations to ADFs. They also present grappa, a related approach where the semantics of arbitrary labelled argument graphs can be directly defined in an ADF-like manner, circumventing the need for explicit translations. Finally, they address various computational aspects of ADFs, like complexity, expressiveness and realizability, and present several implemented systems.

3.3 Structured argumentation

There are four structured argumentation formalisms introduce in the handbook [5]. First, Sanjay Modgil and Henry Prakken reviews abstract rule-based approaches to argumentation, in particular the ASPIC+ framework [19]. In ASPIC+ and its predecessors, going back to the seminal work of John Pollock, arguments can be formed by combining strict and defeasible inference rules and conflicts between arguments can be resolved in terms of a preference relation on arguments. This results in abstract argumentation frameworks (a set of arguments with a binary relation of defeat), so that arguments can be evaluated with the theory of abstract argumentation. First the basic ASPIC+ framework is reviewed, possible ways to instantiate it are discussed and how these instantiations can satisfy closure and consistency properties. Then the relation between ASPIC+ and other work in formal argumentation and nonmonotonic logic is discussed, including a review of how other approaches can be reconstructed as instantiations of ASPIC+. Further developments and variants of the basic ASPIC+ framework are also reviewed, including developments with alternative or generalised notions of attack and defeat and variants with further constraints on arguments. Finally, implementations and applications of ASPIC+ are briefly reviewed and some open problems and avenues for further research are discussed.

Second, Kristijonas Cyras, Xiuyi Fan, Claudia Schulz, and Francesca Toni introduce disputes, explanations, and preferences in Assumption-Based Argumentation (ABA), a form of structured argumentation with roots in non-monotonic reasoning [9]. As in other forms of structured argumentation, notions of argument and attack are not primitive in ABA, but are instead defined in terms of other notions. In the case of ABA these other notions are those of rules in a deductive system, assumptions, and contraries. ABA is equipped with a range of computational tools, based on dispute trees and amounting to dispute derivations, and benefiting from equivalent views of the semantics of argumentation in ABA, in terms of sets of arguments and, equivalently, sets of assumptions. These computational tools can also provide the foundation for multi-agent argumentative dialogues and explanation of reasoning outputs, in various settings and senses. ABA is a flexible modelling formalism, despite its simplicity, allowing to support, in particular, various forms of non-monotonic reasoning, and reasoning with some forms of preferences and defeasible rules without requiring any additional machinery. ABA can also be naturally extended to accommodate further reasoning with preferences.

Third, Alejandro J. García and Guillermo R. Simari introduce argumentation based on logic programming. Among of the programming paradigms based on formal logic, Logic Programming has been a successful effort to create a declarative model of expressing computational processes producing significant theoretical and practical results; as such, the area has contributed computationally attractive systems with remarkable success in many applications. By blending concepts from the areas of Logic Programming and Argumentation, Defeasible Logic Programming (DeLP) proposes a computational reasoning system with an argumentation engine at its core capable of obtaining answers from a knowledge base which is represented with a language that uses logic programming constructs extended with defeasible rules [13]. The careful integration of foundational intuitions and concepts from both areas has formulated a framework that inherits from the logic programming field its expressivity and computational efficiency and receives from argumentation theory a human-like reasoning model facilitating its use in applications. In this chapter, after succinctly recalling the basic elements of logic programming the authors formally introduce the DeLP language and the warranting process that obtains the answers for queries. Then, they present DeLP-Servers, which give possibly distributed client agents running on remote hosts the ability to consult different reasoning services , as well as some extensions and applications of DeLP.

Fourth, Philippe Besnard and Anthony Hunter present a review of argumentation based on deductive arguments [6]. A deductive argument is a pair where the first item is a set of premises, the second item is a claim, and the premises entail the claim. This can be formalized by assuming a logical language for the premises and the claim, and logical entailment (or consequence relation) for showing that the claim follows from the premises. Examples of logics that can be used include classical logic, modal logic, description logic, temporal logic, and conditional logic.

3.4 Argumentation and dialogue

In this part, Martin Caminada first discusses argumentation semantics as formal discussion. He interprets a number of main-stream argumentation semantics by means of structured discussion. The idea is that an argument is justified according to a particular argumentation semantics if and only if it is possible to win a discussion of a particular type. Hence, different argumentation semantics correspond to different types of discussion. He provides an overview of what these discussions look like, and their formal correspondence to argumentation semantics.

Then, Fabrizio Macagno, Douglas Walton, Chris Reed discuss argumentation schemes. The purpose of this chapter is threefold: 1) to describe the schemes, showing how they evolved and how they have been classified in the traditional and the modern theories; 2) to propose a method for classifying them based on ancient and modern developments; and 3) to outline and show how schemes can be used to describe and analyze or produce real arguments. To this purpose, they build on the traditional distinctions for building a dichotomic classifications of schemes, and they advance a modular approach to argument analysis, in which different argumentation schemes are combined together in order to represent each step of reasoning on which a complex argument relies. Finally, they show how schemes are applied to formal systems, focusing on their applications to Artificial Intelligence, AI & Law, argument mining, and formal ontologies.

Finally, Katarzyna Budzynska and Serena Villata introduce approaches for processing natural language argumentation. Although natural language argumentation has attracted the attention of philosophers and rhetoricians since Greek antiquity, it is only very recently that the methods and techniques of computational linguistics and machine learning have become sufficiently mature to tackle this extremely challenging topic. Argument mining, the new and rapidly growing area of natural language processing and computational models of argument, aims at automatic recognition of argument structures in large resources of natural language texts. The goal of this chapter is to familiarise the reader focused on formal aspects of argumentation with this approach, and to show how argument structures, e.g. those studied in abstract argumentation frameworks, can be extracted, providing a bridge between mathematical models and natural language. To this end, they describe the typical argument mining pipeline and related tasks, and present in more detail a specific example of work in this area.

3.5 Computational aspects of formal argumentation

This part is about the computation aspects of formal argumentation. Wolfgang Dvorak and Paul E. Dunne first give an overview of the core computational problems arising in formal argumentation together with a complexity analysis highlighting different sources of

computational complexity. More specifically, they consider three of the previously discussed formalisms, that are Dung's abstract argumentation frameworks, assumption-based argumentation, and abstract dialectical frameworks, each of which allows to highlight different sources of computational complexity in formal argumentation. As most of these problems turn out to be of high complexity they also consider properties of instances, like being in a specific graph class, that reduce the complexity and thus allow for more efficient algorithms. Finally, they show how to apply techniques from parametrized complexity that allow for a more fine-grained complexity classification.

Then, Federico Cerutti, Sarah A. Gaggl, Matthias Thimm and Johannes P. Wallner introduce foundations of implementations if formal argumentation. They survey the current state of the art of general techniques, as well as specific software systems for solving tasks in abstract argumentation frameworks, structured argumentation frameworks, and approaches for visualizing and analysing argumentation. Furthermore, they discuss challenges and promising techniques such as parallel processing and approximation approaches. In addition, they address the issue of evaluating software systems empirically with links to the International Competition on Computational Models of Argumentation.

3.6 Principle-based analysis of formal argumentation

Choice problem: If there are many semantics, then how to choose one semantics from this set of alternatives in a particular application?

Search problem: How to guide the search for new and hopefully better argumentation semantics?

Whereas examining the behaviour of semantics on examples can certainly be insightful, a need for more systematic study and comparison of semantics has arisen. The principles used in a search problem are typically desirable, and desirable properties are sometimes called postulates. For the mathematical development of an principle-based theory, it obviously does not matter whether principles are desirable or not.

The formal analysis in the final five chapters of the first volume of the handbook is based on a principle-based evaluation of argumentation semantics, including dynamic principles and locality and modularity in abstract argumentation. At the structured level, rationality postulates and critical examples are presented. Meanwhile, the respective roles of logic and non-monotonic reasoning in argumentation are explored. Martin Caminada shows how to apply argumentation theory for non-monotonic reasoning using a kind of principles for structured argumentation called rationality postulates. The idea is that arguments are constructed using strict and defeasible inference rules, and that it is then examined how these arguments attack (or defeat) each other. Leendert van der Torre and Srdjan Vesic discuss the principle-based approach to abstract argumentation semantics, Ringo Baumann discusses existence and uniqueness, expressibility, and replaceability, and Pietro Baroni, Massimiliano Giacomin, and Beishui Liao discuss locality and modularity in abstract argumentation. The closing chapter of Alexander Bachman explores the respective roles of logic and nonmonotonic reasoning in argumentation. The notion of collective argumentation is introduced as a logical basis of argumentation frameworks, and provide it with a natural (four-valued) logical semantics. Bochman shows not only that argumentation and logic are important for non-monotonic reasoning, but also the other way round, namely that the main non-monotonic formalisms and argumentation systems constitute actually primary instantiations of Dung's abstract argumentation in appropriately extended logical languages.

4 Open problems and future development

Formal argumentation has developed as a branch of knowledge representation and reasoning within artificial intelligence. As a scientific community and research area, it can be positioned in between informal argumentation and mathematical logic, and it is inspired by applications in legal reasoning, linguistics, computer science, philosophy, and more. As may be expected from a research area in between informal argumentation and mathematical logic, there is a widespread use of different methodologies that are applied in formal argumentation. Moreover, the methodology may differ also on the application for which the formal argumentation models are developed. We consider open questions for the relation between informal and formal argumentation, then we consider questions related to the further development of formal argumentation itself, and finally we consider open questions concerning the relation between formal argumentation and mathematical logic, as well as other formal theories.

4.1 The bridge between informal and formal argumentation

Informal and natural language argumentation has attracted the attention of philosophers and rhetoricians since Greek antiquity. Informal argumentation studies evolving argumentation schemes, classifying them, and using them to describe and analyze or produce real arguments. Informal analysis highlights the role of critical questions and aims to reveal fallacies. Moreover, argument mining is an emerging area of natural language processing and computational models of argument, aiming at automatic recognition of argument structures in large resources of natural language texts. It is only very recently that the methods and techniques of computational linguistics and machine learning have become sufficiently mature to tackle this extremely challenging topic.

Compared to informal and natural language argumentation, the formalisms developed in formal argumentation are highly stylized, abstracting away many aspects characterizing argumentation in daily life. The few remaining concepts are then analyzed with formal rigor, also from a computational point of view. Moreover, an aim of formal argumentation is to develop formal models of argumentation which are useful as a foundation for developing software tools for supporting various argumentation tasks in practical applications. Tom Gordon emphasizes in his chapter that our aim should be to avoid developing a separate technical understanding of argument and argumentation with only a weak connection to how these concepts are understood in the humanities and related fields, both by scholars and practitioners.

The first fundamental distinction in formal argumentation, as highlighted in the historical overview of Prakken, is between argumentation as inference and argumentation as dialogue. Most research reported in the area is of the first kind, though a number of main-stream argumentation semantics can be interpreted by means of structured discussion, in the sense that an argument is justified according to a particular argumentation semantics iff it is possible to win a discussion of a particular type. Hence, different argumentation semantics correspond to different types of discussion.

The formal theory of argumentation as inference has highlighted the attack among arguments as its central concept. This reflects that argumentation is a process where different opinions may conflict, and these conflicts may be explicated and resolved. Consequently, many systematic introductions to argumentation start with Dung's theory of abstract argumentation frameworks, which takes the notions of argument and attack as primitive,

i.e., nothing is assumed about about the structure of arguments or the nature of attack. However, as discussed by Prakken in his historical overview chapter, there had been quite some formal work on argumentation-based inference before Dung's landmark 1995 paper, and all this early work specified the structure of arguments and the nature of attack. According to Prakken, the seminal paper in this respect was Pollock's 1987 article, and many ideas developed in this early body of work are still important today.

The focus in early work on structured argumentation agrees with the usual approaches in informal argumentation, which do not have arguments as the primitive notion but concepts like claims, reasons and grounds. For example, Walton defines the term 'argument' as 'the giving of reasons to support or criticize a claim that is questionable, or open to doubt'.

Nevertheless, the notion of meaning in Dung's theory is radically different from many traditional theories. There are multiple semantics under consideration, and each semantics may present various alternatives. As we explain later, when we consider the relation between formal argumentation and mathematical logic, it means that the mainstream theories of formal argumentation discussed in this area are closer to para-consistent logic developed in philosophical logic, and non-monotonic logic developed in artificial intelligence.

Many relations between the various formalisms of structured argumentation have been discussed, but there is no consensus on a common core going beyond Dung's abstract theory, and there is no consensus on which system should be used in practice for which application. For example, a very expressive approach like ASPIC+ may be useful for a principle based analysis of structured argumentation, but a more restricted approach like ABA or DeLP may be more suited for implementation, or to prove certain formal properties.

The formal analysis discussed in the area is of two kinds. First, algorithms together with complexity results are presented for the defined formal systems. Second, a principle based approach is developed to analyze the formal systems. the principle-based or axiomatic approach is a methodology to choose an argumentation semantics for a particular application, and to guide the search for new argumentation semantics, The study of representation and (im)possibility results for abstract argumentation must be extended for a principle-based approach for extended argumentation such as bipolar frameworks, preference-based frameworks, abstract dialectical frame-works, weighted frameworks, and input/output frameworks.

Coming from informal and natural language argumentation, the theory of formal argumentation presents two challenges. The first challenge is whether the developed theories of formal argumentation can be used as a foundational theory of informal and natural language argumentation. For example, how can argument schemes be used to define arguments in the formal approaches, or how can the formal approaches support the natural language processing techniques and machine learning algorithms?

The second challenge is how the here developed theories of formal argumentation can be adapted or extended such that they cover a wider range of phenomena in informal argumentation. Ideally, these adaptations and extensions should still follow the mathematical elegance and simplicity of the presented theories. Moreover, these innovations should not affect the formal and computational properties of the theories, or at least they should not make large concessions.

4.2 Challenges for formal argumentation

In the past two decades, theories and algorithms of formal argumentation have been extensively developed. However, there are still some fundamental problems to be explored, including the problems related to time and dynamics, rationality postulates, models and semantics

of argumentation, preferences between arguments, and efficient algorithms. Some of the problems in this direction are identified as follows.

- **Validity or status of arguments with respect to time/dynamics**. Since argumentation is intrinsically dynamic, the status of arguments may change upon the changing of underlying knowledge. Since little attention has been devoted to explicitly consider the presence of time and its impact in an argumentation-based context, it would be interesting to further study the model of formal argumentation with respect to time/dynamics.
- **Qualitative postulates that should be satisfied in specific contexts**. Several rationality postulates have been proposed both for abstract and structured argumentation. However, further research should be devoted to study the sets of postulates that should be satisfied in specific application contexts. This may also require the identification of novel postulates.
- **Development of Dung's theory**. Dung's abstract argumentation plays an important role in the community of formal argumentation, and there are already a number of extensions of this theory. Further extensions might include the following.
  - Identifying an elegant formalism encompassing Dung's model and capturing also different ways of evaluating arguments, e.g. balancing considerations.
  - Developing an alternative approach to model the cases where we are interested in only one argument, and focus mainly on explanation and justification.
  - Achieving a clarification on the "semantics of a semantics", to make clear when to adopt a specific semantics instead of another. In this respect, a focus on specific argumentation contexts would be required.
- **Preference relation and defeat relation.** When considering the preference relation over underlying knowledge [14], an important question is how to lift the the preference relation of the underlying knowledge to that of arguments. Meanwhile, preference order between arguments is dynamic and may depend on the labelling of arguments, thus a recursive process may be needed.
- **Efficient algorithms.** Since many natural questions regarding argument acceptability are computationally intractable, developing efficient algorithms for formal argumentation is important. According to the results of a recent competition on Computational Models of Argumentation[1], reduction-based systems (either SAT-based or ASP-based) are more efficient than non reduction-based. However, this may be due to the fact that research focusing on efficient algorithms is not sufficiently mature. Thus it would be interesting to study and develop algorithms for abstract argumentation not based on SAT problem, e.g., fixed-parameter tractable algorithms.
- **Negation of arguments.** It is not clear what is the negation of an argument. One possible way is to define operators, like negation of trust as distrust, negation of attack as support, negation of argument, etc.

4.3 Connection with other theories

Besides, formal argumentation can also be related to other formal theories like computational social choice theory, belief revision, neural networks, and Bayesian networks, etc.

- **Formal argumentation and logics.** The interplay between argumentation and logic has a long history. However, the relation between formal argumentation and various kinds

[1] ICCMA 2015: see http://argumentationcompetition.org/2015/

of logics are not clear. For instance, how to use argumentation to represent preference-based nonmonotonic reasoning, how to use argumentation to represent deontic reasoning, etc.

- **Formal argumentation and mathmatics.** Given a direct graph, mathematicians and researchers in the community of argumentation may have different views. For instance, while the former pay their attention to the number of nodes, number of arrows, topological properties, connectivity, etc., and define the notion of a kernel of the graph, the latter concern more on arguments and look for complete extensions of the graph. It is worth to further study the mutual benefits of these two areas.
- **Formal argumentation and computational social choice.** There are some interesting research questions, e.g., to explore the relation between voting and the semantics of argumentation, or between the kind of democracy and the semantics of argumentation, etc.
- **Formal argumentation and belief revision.** Formal argumentation and belief revision are complementary. The former concerns how an agent changes her beliefs when new information arrives, while the latter deals with the the justification of new beliefs or the strategies to changes the beliefs of other agents. The connections between these two fields are promising and beneficial.
- **Bridge betweens uncertainty, fuzziness and argumentation.** As a combination of qualitative approach and quantitative approach, it is very promising to develop theories and applications by combining argumentation with uncertainty theory, including probability theory, possibility theory and fuzzy theory, etc.
- **Formal argumentation and other networks.** Abstract argumentation framework is a directed graph. It is natural to connect argumentation framework to other networks, such as neural networks, Bayesian Networks, etc.

We give some examples in the remainder of this section.

4.3.1 Connection with graph theory

Since the following sections discuss the connection with mathematical directed graph theory the style of writing needs to be more formal.

Abstract argumentation deals with binary relations R on a set S. The system (S, R) has sometimes the following properties when used by the argumentation community.

(*1) S is finite.

(*2) R is allowed to be reflexive and allowed to be symmetrical.

Also a lot of the mathematics studied in formal argumentation has to do with dealing with cycles arising because of these properties. In graph theory in comparison there is the notion of directed graphs (digraphs). The requirement is that R is irreflexive. There is also the notion of weak ordering where R is also required to be not symmetric $xRy \rightarrow \neg(yRx)$ [24]. Typically, there is no requirement that S be finite.

The abstract argumentation communities and the graph theory mathematicians ask slightly different questions about (S, R). They also use different words/names for sometimes he same concept.

- If $x, y \in S$ and $\neg xRy \wedge \neq yRx$ in argumentation we say $\{x, y\}$ are conflict free. In graph theory we say they are independent.
- In argumentation they consider complete extensions, $E \subseteq S$. These re maximal subsets of conflict free points and researchers look at their existence. Among them are stable extensions. In graph theory such stable extension sets are called kernels and the mathematics of their existence is studied.

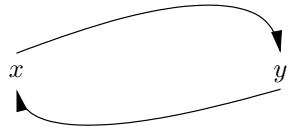


Figure 3

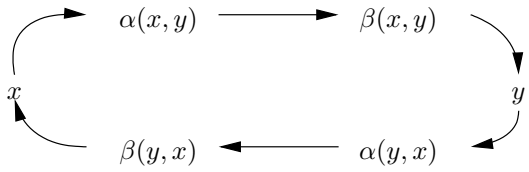


Figure 4

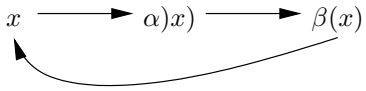


Figure 5

Stable extensions or kernels are subsets $E \subseteq S$ satisfying the following:

1. $\forall x, y \in E(\neg xRy \wedge \neg yRx)$
2. $(\forall z \in S - E)(\exists y \in E)(yRz)$.

In graph theory one studies also perfect kernels, namely kernels E such that also $S - E$ is a kernel. See papers of Walicki and Sjurdyrkolbotn.

Both communities realise that odd cycles in (S, R) cause problems and try to mathematically deal with them. The argumentation people are more algorithmic while the graph theory approach is more set-theoretical. Also in argumentation they deal with numerical graphs as well (papers by Gabbay-Rodrigues and others) while the graph community have less research about numerical annotation in the abstract math (there are many network communities such as flow networks, neural networks, etc., theses are very numerical but they do not stress conflict freeness).

It is important to note that results and concepts in the argumentation community make the requirement of irreflexivity ($\neg xRx$) and a-symmetry $xRy \rightarrow \neg yRx$ mathematically unimportant. In other words, any (S, R) can be rewritten as (S^*, R^*), with $S \subseteq S^*$ and $R \subseteq R^*$ such that any kernel $E \subseteq S$ can be uniquely obtained and extended to a unique kernel $E^* \subseteq S^*$ by $E = E^* \cap S$. The idea is as follows, explained by example. Let $x \twoheadrightarrow y$ means xRy. Consider Figure 3. We may have $x = y$.

Figure 4 considers some new points. Let $\alpha(x, y)$ and $\beta(x, y)$ be $\alpha(y, x), \beta(y, x)$.

If $x = y$ we take only Figure 5.

Let (S^*, R^*) be extended as above for all pairs $\{x, y\}$ with $x \neq y$ and $xRy \wedge yRx$ and any z with zRz and $\alpha(z), \beta(z)$. So for example Figure 6 becomes Figure 7.

(S, R) of Figure 6 has one kernel/stable extension $E = \{x\}$. Figure 7 of (S^*, R^*) has the kernel E^*.

$$E^* = \{x, \beta(x, y), \alpha(y, x), \alpha(y)\}.$$

The above illustrates how mathematically formal argumentation can connected with related mathematical directed graph theory. The formal similarities between the areas and the natural research instinct of the mathematicians involved will push cooperation between some of the individuals in each group.

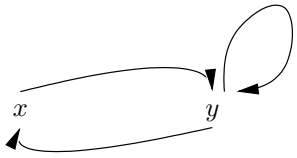


Figure 6 Showing (S, R)

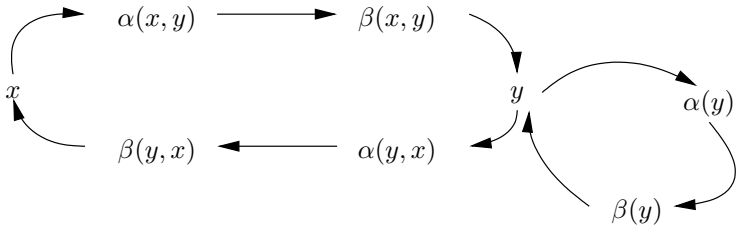


Figure 7 showing (S^*, R^*)

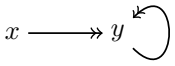


Figure 8

4.3.2 Connection with logic and liar paradox, Gaifman 1988

The basic meaning of xRy in argumentation is that if we accept x as "in" then we must reject y as "out". This fits nicely with the liar paradox basic understanding that xRy means that x is a statement that y is false. In fact, Gaifman's 1988 paper already introduced mathematically the graphs of argumentation networks of Dung's 1995 paper. Gaifman's evaluation, however, is different because of his different intended interpretation.

Consider Figure 8. According to Dung, the interpretation of $x \to y$ can be taken as ecological. x kills y. Thus since x is not attacked, x is "in" or is "alive". Since x is alive and attacks y we have that y is "out" or "dead". According to Gaifman and the liar paradox interpretation, x says that y is "false". Since y says "I am lying" it cannot have a crisp value and so x cannot have a value. So we get $x = y =$ no value $=$ undecided $=$ gap. (gap is the Gaifman terminology for the argumentation case of undecided.)

What about Figure 9? In argumentation $y =$ und and therefore $x =$ und. According to the liar interpretation approach, y says I am lying and furthermore so is x. In comparison, x does not say anything about anyone else lying. We need to agree that if x says nothing about other statements then we let x be true.[2] This corresponds to Clause (C1) of the Caminada labelling. So following this agreement we get that $x = \top$. Let us now consider y. We have that that y is making two statements, namely y is saying"y is false and x is false". Thus since x is $\top$ we get that the second statement of y is false. Thus we have:

$y = [(y \text{ is false }) \wedge x \text{ is false}] = \text{false}.$

Note that Figure 9 is obtained from Figure 8 by reversing its arrows, and then Gaifman evaluation for Figure 9 gives the same result as Dung evaluation for Figure 8.

[2] Gaifman uses a classical model to evaluate x, so we can say in agreement with Gaifman that our classical model gives all such atoms value $\top$.

Figure 9

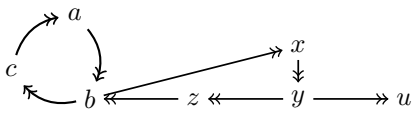


Figure 10

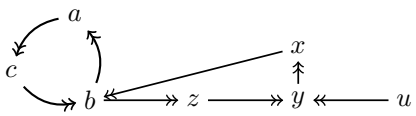


Figure 11

Let us do a more complex example. Consider the network of Figure 10.
If we calculate a Dung extension for this figure, we get that the only extension

x = in, y = out, u = in, z = in, b = out, c = in, a = out.

Let us calculate now the Gaifman extension for this same Figure 10:

$u = \top$ since u says nothing about anyone else
$y = \bot$ since y says u is lying and $u = \top$
$= \top$ since x says y is lying
$b = \bot$ since $x = \top$
$z = \top$ since $b = \bot$
$a = \top$ since $b = \bot$
$c = \bot$ since $a = \top$.

Now let us invert the arrows in Figure 10 and get Figure 11 and then compute according to Dung. u = in (not attacked), y = out, x = in, b = out, z = in, a = in, c = out.

4.3.3 Connection with Saveliev 2017

Some formal "argumentation" work has already been done by mathematicians like Saveliev. This is a good sign for the future. Saveliev considered (S, T, U) with the following properties:

$T(x)$ means $x = \top$
xUy means x says that y is false

He required the following axioms

A1: $Tx \wedge Uy \rightarrow \neg Ty$
A2: $\neg Tx \wedge \exists y(xUy) \rightarrow \exists y(xUy \wedge Ty)$

Let us rewrite $xRy = \text{def}.yUx$. We get

A1: $Tx \wedge yRx \rightarrow \neg Ty$
A2: $\exists y(yRx) \rightarrow [\neg Tx \rightarrow \exists y(yRx \wedge T(y)]$

If we assume that (S, R) is such that every x is attacked (i.e. no start points $\forall x \exists y(yRx)$), then we get

A1: $Ty \wedge yRx \rightarrow \neg Ty$
A2: $\neg Tx \rightarrow \exists y(yRx \wedge Ty)$.

Therefore we get

A3: Tx iff $\exists y(Ty \wedge yRx)$.

This is exactly the Caminada condition which mean that we are dealing with Dung networks where $\forall x \exists y(yRx)$ holds.

It is not a problem to make this condition true. For any x which is not attacked, add a new point $\gamma(x)$ and expand (S, R) to (S^*, R^*) with

$$S^* = S \cup \{\gamma(x) | x \text{ not attacked in } R\}$$
$$R^* = R \cup \{(\gamma(x), x), (x, \gamma(x)) | x \text{ not attacked in } R\}$$

The extensions E of (S, R) are obtained uniquely from those extensions of (S^*, R^*) where all $\gamma(x)$ are in. Let $\Gamma = \{\gamma(x)\}$. So $E \supseteq \Gamma$.

Saveliev proves theorems about his axioms, i.e. on models (S^*, R^*). These can be translated to theorems on (S, R), and vice versa.

5 Applications of formal argumentation

Applications of the theories and algorithms of formal argumentation have been proposed in several domains. This is mainly due to the ability of the theory to handle uncertain and possibly contradictory information, its capability of capturing diverse and heterogeneous reasoning mechanisms, as well as the fact that the basic concepts of the theory are akin to human intuition. In particular, a natural application domain is legal reasoning [4], since legal knowledge is inherently argumentative, while other obvious application domains are medical reasoning and e-democracy (see e.g. [16, 7]).

Whatever domain is considered, a core issue is the identification and/or acquisition of arguments. While these can be manually introduced by the user, a recent thread of research is devoted to automatically identify argumentative structures from textual sources, including arguments components (e.g. premises and conclusions), argumentation schemes related to arguments, and the relationships holding between arguments (such as subargument relations, attacks and support). This is a complex issue which requires (at least) the integration of Computational Linguistics research with the study of computational models of argumentation.

Regarding specifically the argumentation formalisms for argument mining, since structured argumentation systems are meant to capture more closely the actual construction of arguments in argumentation processes, they are more suitable candidate formalisms for this purpose. For instance, in the assumption-based model, argument analysis amounts to the identification, within a given text, of the argument claim, of its supporting assumptions and of the rules used for the claim deduction. Similar "component identification guidelines" could be drawn

for other structured systems. Two difficulties must be acknowledged, however, concerning the use of these formalisms for actual argument mining. First, some of these formalisms are still rather abstract, since, for the sake of generality, they leave unspecified some important aspects (e.g. the actual language adopted) hence they are not applicable without making some further specific choices at the implementation level. Second, and more important, they have typically been conceived to capture argumentation in already formalized settings (e.g. argument-based reasoning on possibly inconsistent knowledge bases) rather than at a natural language level. Indeed, due to the enthymematic nature of most natural arguments, some of the argument components encompassed by the above mentioned models, like the assumptions or the rules used, are left implicit in natural language expressions of arguments. Hence, one might argue that such structured formalisms are a suitable target for a "second level" analysis (and completion) of the natural arguments identified in a text, but, in a sense, can be too demanding as a first target formalism for the argument mining process itself. As a matter of fact, an analysis of the references in the papers presented at the First and Second International Workshop on Argumentation Mining shows that the structured argumentation formalisms are practically absent from current research on argumentation mining, while much more attention has been reserved to the use of semi-formal/diagrammatical schemes.

Semi-formal schemes, often lending themselves to a diagrammatical representation, provide models of argument structure and/or of inter-argument relationships which are typically focused on a few elements, regarded as crucial for the analysis and comprehension of some key aspects of the argumentation process. As such, these schemes do not provide a complete account nor a formal backing of the argumentation process as a whole, to be covered by other models, but rather can be regarded as shedding light on some central points, beneficial for the development of more complete and more formal models. Examples are the Toulmin model [25], subsequently developed by Freeman [10, 11], Wigmore diagrams [29] and Walton's argumentation schemes [27, 28]. A discussion of the uses of this kind of models for argumentation mining is provided by [22], while an analysis of the papers presented at the First and Second International Workshop on Argumentation Mining and of some earlier influential work [21] shows in particular a prevalence in the use of Walton's argumentation schemes.

Argumentation may be a natural way for human reasoning and communication. However, the gap between existing theories and algorithms of formal argumentation and real applications is still surprisingly big. Some research problems are as follows.

- **Natural language interfaces to arguments.** In order to facilitate the applications of theories and algorithms of formal argumentation to daily life reasoning and communication, it is vital to develop human-friendly interfaces.
- **Formal argumentation account of fallacies.** Fallacies are the most efficient way of human reasoning and persuasion in daily life. Formal models of fallacies are still missing.
- **Analyzing and modelling argumentation schemes** As a semi-formal model, argumentation schemes can play an important role to connect arguments in natural language and formal argumentation in AI. However, hot to exploit argumentation schemes in formal argumentation is a problem not completely studied yet.
- **Argumentation mining.** Argumentation mining is a promising direction to apply theories and algorithms of formal argumentation. However, according to the state of the art, there are a lot of challenging problems in this direction, e.g., the identification and formalization of arguments and their components, the identification of various relations between arguments, the measurement and formalization of uncertainties of natural arguments, etc.

- **Software engineering methods for argumentation.** In order to connect the notions developed in argumentation theory with practical domains, the development of software engineering methods (e.g. for requirement analysis) would be useful to drive research in argumentation.

In addition, to study how theories of formal argumentation can be applied to practice, one may consider a more systematic research direction, called "argumentation analysis", which is coined after the word "decision analysis" . The nutshell of decision analysis is the application of decision science to real-world problems through the use of systems analysis and operations research. It describes how people should logically make decisions in simple situations or complex situations. In the setting of formal argumentation, we need study procedures, methods, and tools for identifying, clearly representing, and formally assessing important aspects of argumentation from simple situations to very complex situations:

Reasoning problems Consider a simple reasoning problem, e.g. which people can together go to a party based e.g. on their (possibly temporal) constraints and individual preferences. Argumentation analysis may be considered as a prescriptive approach, especially concerned with dealing with uncertainties qualitatively and/or quantitatively. Prescriptive argumentation researches how optimal arguments could be accepted.

Agent interaction Consider a more complex problem of the argument of sex offenders in therapy. Every human being has reasoning distortion, but sex offenders have unusual and exceptional cognitive distortion. To model this distortion, we need to know how people actually make arguments, regardless of argument quality. Meanwhile, people also use logical fallacies, which are the most effective arguments in daily life. The prescriptive approach is found to be in fact rarely used in the reasoning of individuals. The hiatus between prescriptive argumentation and descriptive approaches is greater in high-stakes argumentation and negotiation, made under time pressure.

Institutions Consider the complex decisions of religious or legal systems over time. The institution builds up information and argumentation evolves into information processing. To model this kind of argument, we need to go beyond the above-mentioned prescriptive and descriptive approaches. An example in this direction is Talmudic logic. The Jewish Talmud is a body of arguments and discussions about all aspects of the human agents social, legal, ethical and religious life. It is a practical and coherent body of laws developed logically to address human behavior.

6 Conclusions

Some researchers seem to believe that the theory of formal argumentation may be more or less finished, and we can focus on computational aspects and the use of formal argumentation, since there is nothing important left to add to it. In our view, this is far from the truth. On the contrary, there is a large number of important open problems in the foundations in this field of research.

An important recommendation coming from this Dagstuhl workshop is that we need to evaluate the current argumentation formalisms. In particular, more investigation on how to apply the formalisms studied in formal computational argumentation to "real" reasoning contexts, including e.g. legal and ethical reasoning. An important issue to connect formal argumentation to real argumentation is to mine arguments from natural language text, which requires building novel applications for argument mining.

Moreover, from these argumentation applications a more systematic research direction may emerge, called "argumentation analysis." In the setting of formal argumentation, we need study procedures, methods, and tools for identifying, clearly representing, and formally assessing important aspects of argumentation from simple situations to very complex situations.

Together, the applications and argumentation analysis may inform the further development of the formal machinery involved in argumentation theory. From a theoretical perspective, we call for more unified theoretical results rather than fragmentation into specific isolated studies. Moreover, the identification of some important applications that should be developed may contribute to this evolution of the community. In short, the research on this topic is active, vibrant, and rich, but the area is vast and diverse and needs to be connected together.

Acknowledgements. We thank the anonymous reviewers for valuable comments. We have received funding from the European Union's H2020 research and innovation programme under the Marie Curie grant agreement No. 690974 for the project MIREL: MIning and REasoning with Legal texts.

7 Participants

- Pietro Baroni
University of Brescia, IT
- Ringo Baumann
Universität Leipzig, DE
- Stefano Bistarelli
University of Perugia, IT
- Alexander Bochman
Holon Institute of Technology, IL
- Gerhard Brewka
Universität Leipzig, DE
- Katarzyna Budzynska
Polish Academy of Sciences – Warsaw, PL
- Martin Caminada
University of Aberdeen, GB
- Federico Cerutti
University of Aberdeen, GB
- Wolfgang Dvorak
Universität Wien, AT
- Dov M. Gabbay
King's College London, GB
- Massimiliano Giacomin
University of Brescia, IT
- Tom Gordon
Fraunhofer FOKUS – Berlin, DE
- Beishui Liao
Zhejiang University, CN
- Henry Prakken
Utrecht University, NL
- Chris Reed
University of Dundee, GB
- Odinaldo Rodrigues
King's College – London, GB
- Guillermo R. Simari
National University of the South – Bahía Blanca, AR
- Matthias Thimm
Universität Koblenz-Landau, DE
- Leendert van der Torre
University of Luxembourg, LU
- Bart Verheij
University of Groningen, NL
- Emil Weydert
University of Luxembourg, LU
- Stefan Woltran
TU Wien, AT

References

1 Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *Knowledge Engineering Review*, 26(4):365–410, 2011. doi:10.1017/S0269888911000166.

2 Pietro Baroni and Massimiliano Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10-15):675–700, 2007. doi:10.1016/j.artint.2007.04.004.

3 Pietro Baroni, Massimiliano Giacomin, and Beishui Liao. Uncertainty and fuzziness from natural language to argumentation models. In *Proceedings of the European Conference on Argumentation (ECA) - Thematic Panel*, Lisbon, Portugal, 2015.

4 Trevor J. M. Bench-Capon, Henry Prakken, and Giovanni Sartor. Argumentation in legal reasoning. In *Argumentation in Artificial Intelligence*, pages 363–382. Springer, 2009. doi:10.1007/978-0-387-98197-0_18.

5 Philippe Besnard, Alejandro Javier García, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Ricardo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4, 2014. doi:10.1080/19462166.2013.869764.

6 Philippe Besnard and Anthony Hunter. Constructing argument graphs with deductive arguments: a tutorial. *Argument & Computation*, 5(1):5–30, 2014. doi:10.1080/19462166.2013.869765.

7 Dan Cartwright and Katie Atkinson. Using computational argumentation to support e-participation. *IEEE Intelligent Systems*, 24(5):42–52, 2009. doi:10.1109/MIS.2009.104.

8 Célia da Costa Pereira, Beishui Liao, Alessandra Malerba, Antonino Rotolo, Andrea G. B. Tettamanzi, Leon van der Torre, and Serena Villata. Handling norms in multi-agent system by means of formal argumentation. *IFCoLog Journal of Logic and its Applications*, 4(9), 2017.

9 Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. Assumption-based argumentation. In *Argumentation in Artificial Intelligence*, pages 199–218. Springer, 2009. doi:10.1007/978-0-387-98197-0_10.

10 James B. Freeman. *Dialectics and the macrostructure of arguments*. De Gruyter Mouton, 1991.

11 James B. Freeman. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer, 2011. doi:10.1007/978-94-007-0357-5.

12 Dov M. Gabbay and Odinaldo Rodrigues. Probabilistic argumentation: An equational approach. *Logica Universalis*, 9(3):345–382, 2015. doi:10.1007/s11787-015-0120-1.

13 Alejandro Javier García and Guillermo Ricardo Simari. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming (TPLP)*, 4(1-2):95–138, 2004. doi:10.1017/S1471068403001674.

14 John F. Horty. Defaults with priorities. *Journal of Philosophical Logic*, 36(4):367–413, 2007. doi:10.1007/s10992-006-9040-0.

15 Anthony Hunter and Matthias Thimm. On partial information and contradictions in probabilistic abstract argumentation. In Chitta Baral, James P. Delgrande, and Frank Wolter, editors, *Proceedings of the Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning KR 2016*, pages 53–62, Cape Town, South Africa, 2016. AAAI Press. URL: http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12780.

16 Anthony Hunter and Matthew Williams. Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, 56(3):173–190, 2012. doi:10.1016/j.artmed.2012.09.004.

17 Beishui Liao. *Efficient Computation of Argumentation Semantics*. Intelligent systems series. Academic Press, 2014. URL: http://store.elsevier.com/product.jsp?isbn=9780124104068.

18 L. Thorne McCarty and N. S. Sridharan. The representation of an evolving system of legal concepts: II. prototypes and deformations. In Patrick J. Hayes, editor, *Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI 1981*, pages 246–253, Vancouver, BC, Canada, 1981. William Kaufmann. URL: http://ijcai.org/Proceedings/81-1/Papers/050.pdf.

19 Sanjay Modgil and Henry Prakken. The $ASPIC^+$ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014. doi:10.1080/19462166.2013.869766.

20 Marie-Francine Moens. Argumentation mining: Where are we now, where do we want to be and how do we get there? In Prasenjit Majumder, Mandar Mitra, Madhulika Agrawal, and Parth Mehta, editors, *Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation, FIRE 2013*, pages 2:1–2:6, New Delhi, India, 2013. ACM. doi:10.1145/2701336.2701635.

21 Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, 2011. doi:10.1007/s10506-010-9104-x.

22 Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31, 2013. doi:10.4018/jcini.2013010101.

23 Chaim Perelman and Lucie Olbrechts-Tyteca. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame, 1969.

24 M. Richardson. On weakly ordered systems. *Bulletin of the American Mathematical Society*, 52(2):113–116, 02 1946. URL: https://projecteuclid.org:443/euclid.bams/1183507698.

25 Stephen E. Toulmin. *The Uses of Argument*. Cambridge: Cambridge University Press, 1958.

26 Stephen E. Toulmin. *The Uses of Argument, Updated Edition*. Cambridge: Cambridge University Press, 2003.

27 Douglas Walton. *Argumentation schemes for presumptive reasoning*. Lawrence Erlbaum Assoc Inc, 1996.

28 Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008. URL: http://www.cambridge.org/us/academic/subjects/philosophy/logic/argumentation-schemes.

29 John Henry Wigmore. *The Principles of Judicial Proof: As Given by Logic, Psychology, and General Experience, and Illustrated in Judicial Trials*. Little, Brown & Co, 1931.

Manifesto from Dagstuhl Perspectives Workshop 17442

From Evaluating to Forecasting Performance: How to Turn Information Retrieval, Natural Language Processing and Recommender Systems into Predictive Sciences

Edited by

Nicola Ferro[1], Norbert Fuhr[2], Gregory Grefenstette[3], Joseph A. Konstan[4], Pablo Castells[5], Elizabeth M. Daly[6], Thierry Declerck[7], Michael D. Ekstrand[8], Werner Geyer[9], Julio Gonzalo[10], Tsvi Kuflik[11], Krister Lindén[12], Bernardo Magnini[13], Jian-Yun Nie[14], Raffaele Perego[15], Bracha Shapira[16], Ian Soboroff[17], Nava Tintarev[18], Karin Verspoor[19], Martijn C. Willemsen[20], and Justin Zobel[21]

1 **University of Padova, Italy** ferro@dei.unipd.it
2 **University of Duisburg-Essen, Germany** norbert.fuhr@uni-due.de
3 **Institute for Human Machine Cognition, USA** ggrefenstette@ihmc.us
4 **University of Minnesota, Minneapolis, USA** konstan@umn.edu
5 **Autonomous University of Madrid, Spain** pablo.castells@uam.es
6 **IBM Research, Ireland** elizabeth.daly@ie.ibm.com
7 **DFKI GmbH, Saarbrücken, Germany** declerk@dfki.de
8 **Boise State University, USA** michaelekstrand@boisestate.edu
9 **IBM Research, Cambridge, USA** werner.geyer@us.ibm.com
10 **UNED, Spain** julio@lsi.uned.es
11 **The University of Haifa, Israel** tsvikak@is.haifa.ac.il
12 **University of Helsinki, Finland** krister.linden@helsinki.fi
13 **FBK, Trento, Italy** magnini@fbk.eu
14 **University of Montreal, Canada** nie@iro.umontreal.ca
15 **ISTI-CNR, Pisa, Italy** raffaele.perego@isti.cnr.it
16 **Ben-Gurion University of the Negev, Israel** bshapira@bgu.ac.il
17 **National Institute of Standards and Technology, USA** ian.soboroff@nist.gov
18 **Delft University of Technology, The Netherlands** n.tintarev@tudelft.nl
19 **The University of Melbourne, Australia** karin.verspoor@unimelb.edu.au
20 **Eindhoven University of Technology, The Netherlands** M.C.Willemsen@tue.nl
21 **The University of Melbourne, Australia** jzobel@unimelb.edu.au

Abstract

We describe the state-of-the-art in performance modeling and prediction for Information Retrieval (IR), Natural Language Processing (NLP) and Recommender Systems (RecSys) along with its shortcomings and strengths. We present a framework for further research, identifying five major problem areas: understanding measures, performance analysis, making underlying assumptions explicit, identifying application features determining performance, and the development of prediction models describing the relationship between assumptions, features and resulting performance.

Perspectives Workshop October 30 to November 03, 2017 – www.dagstuhl.de/17442
2012 ACM Subject Classification Information systems → Information retrieval, Information systems → Recommender systems, Computing methodologies → Natural language processing
Keywords and phrases Information Systems, Formal models, Evaluation, Simulation, User Interaction
Digital Object Identifier 10.4230/DagMan.7.1.96

From Evaluating to Forecasting Performance: How to Turn Information Retrieval, Natural Language Processing and Recommender Systems into Predictive Sciences, *Dagstuhl Manifestos*, Vol. 7, Issue 1, pp. 96–139
Editors: Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan

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

Executive Summary

This workshop brought together experts from information retrieval (IR), recommender systems (RecSys), and natural language processing (NLP). Common to these three neighboring fields is the challenge of modeling and predicting algorithm performance under varying application conditions, measured in terms of result quality. A particular challenge is that these methods create or affect a human experience, and so performance ultimately depends on human judgment of the quality of experience and performance. Progress in performance modeling and prediction would allow us to better design such systems to achieve desired performance under given operational conditions.

In this manifesto, we first consider the state of prediction in the three research disciplines, and then describe a general framework for addressing the prediction problem.

Research in IR puts a strong focus on evaluation, with many past and ongoing evaluation campaigns. However, most evaluations utilize offline experiments with single queries only, while most IR applications are interactive, with multiple queries in a session. Moreover, context (e.g., time, location, access device, task) is rarely considered. Finally, the large variance of search topic difficulty make performance prediction especially hard.

NLP has always engaged in both intrinsic evaluation of the steps in the language processing pipeline (e.g., language identification, tokenization, morphological analysis, part-of-speech tagging, parsing, entity extraction, classification, etc.) and in extrinsic, application-oriented evaluation (such as information retrieval, machine translation, and so on). The different goals of different applications mean that there is no one best NLP processing system, and also call into doubt the usefulness of intrinsic evaluations alone, since the improvement of one pipeline step might have little influence on broader application performance. Added to this, the performance of an NLP system in a new language or domain can be hard to predict, as it may depend on the existence of language resources to implement these pipelines.

RecSys generate predictions and/or recommendations for a particular user from a set of candidate items, often for a particular context. Like the other two areas, the field has a legacy of metrics, user experimentation research, benchmarks, and datasets. At a general level, current RecSys research aims at distilling the current large body of empirical knowledge into more systematic foundational theories and at learning from cumulative research. More specific issues include topics like auto-tuning of systems, exploration vs. exploitation, coping with context-dependent performance, and algorithm vs. system performance.

For a general framework for performance prediction, we identified 5 problem areas:

1. **Measures**: We need a better understanding of the assumptions and user perceptions underlying different metrics, as a basis for judging about the differences between methods. Especially, the current practice of concentrating on global measures should be replaced by using sets of more specialized metrics, each emphasizing certain perspectives or properties. Furthermore, the relationships between system-oriented and user-/task-oriented evaluation measures should be determined, in order to obtain improved prediction of user satisfaction and attainment of end-user goals.
2. **Performance analysis**: Instead of regarding only overall performance figures, we should develop rigorous and systematic evaluation protocols focused on explaining performance differences. Failure and error analysis should aim at identifying general problems, avoiding idiosyncratic behavior associated with characteristics of systems or data under evaluation.
3. **Assumptions**: The assumptions underlying our algorithms, evaluation methods, datasets, tasks, and measures should be identified and explicitly formulated. Furthermore, we need strategies for determining how much we are departing from these assumptions in new cases and how much this impacts on system performance.

4. **Application features**: The gap between test collections and real-world applications should be reduced. Most importantly, we need to determine the features of datasets, systems, contexts, tasks that affect the performance of a system.
5. **Performance Models**: We need to develop models of performance which describe how application features and assumptions affect the system performance in terms of the chosen measure, in order to leverage them for prediction of performance.

These five problem areas call for a research and funding agenda where basic research efforts should address the first three items above by laying new foundations for the IR, NLP, and RecSys fields and adopting a multidisciplinary approach to bridge among algorithmics, data management, statistics, data analysis, human-computer interaction, and psychology. Once these foundations are laid, subsequent research efforts should leverage them and exploit, for example, machine learning and artificial intelligence techniques to address the last two items in the above list.

Overall, the above research agenda outlines a set of "high risk, high gain" research topics and promises to deliver a major paradigm shift for the IR, NLP, and RecSys fields, by embracing a new radical vision of what should be at the foundations of those fields and targeting a technological breakthrough able to change the way in which academia and industry invent, design and develop such kind of systems.

Table of Contents

1 Introduction

Predictability is a fundamental attribute of daily life: we expect familiar things to behave in familiar ways. In science, predictability has taken on more specific meanings; our understanding of a system, model, or method is validated by our ability to predict performance or outcomes, often in a quantified form. A particular challenge for the systems regarded here is that, ultimately, they create or affect a human experience.

Questions we might like to answer in this context include the following:

- How reliable will a system perform over different tasks?
- What test materials (and at what scale) are required to establish performance to standards that imply predictability?
- Will the current performance of a system be robust to changes in its data or use, and what parameters or limits would indicate whether there is a risk to performance?
- Can performance uncertainty be quantified?
- How can we plan a move from a laboratory prototype to a system in operation?
- To what extent do performance metrics match user perceptions and experiences?
- What resources or configuration might be required to adapt a system to a new context or a new application?
- What resources might be required to maintain a system or confirm that it is continuing to perform?

In this paper, we first discuss the state of performance prediction in the areas of Information Retrieval (IR), Recommender Systems (RecSys), and Natural Language Processing (NLP). Then we present a general framework for addressing the prediction problem, and point out the corresponding research challenges.

2 Information Retrieval

2.1 Motivations for Prediction in IR

An IR system is successful if it provides the information that a user needs to complete a task, supports them in learning, or helps the user accomplish a goal. That is, the purpose of an IR system is to have impact on a cognitive state, and thus the value or correctness of an outcome is inherently subjective. A related challenge is that, typically, a single system is often relied on by a user for a wide range of unrelated activities, and that similar interactions from different users may be the consequence of different intents. This is in part a consequence of the fact that tasks can be underspecified, or ill-formed; or may be fluid, shifting during the course of an interaction; or may be progressive.

Validation via users, and inconsistencies in that validation, are therefore an inherent component of prediction. These validations are inherently more complex than specific questions such as comprehensibility of a text or ease of use of an app. [49] describes prediction as a challenge for evolving IR to an engineering science, but the problem in IR is even more complex, referring to human judgment rather than to measurement of certain technical properties.

Several types of prediction may be relevant in IR. One case is that we have a system and a collection and we would like to know what happens when we move to a new collection, keeping the same kind of task. In another case, we have a system, a collection, and a kind of

task, and we move to a new kind of task. A further case is when collections are fluid, and the task must be supported over changing data.

Current approaches to evaluation mean that predictability can be poor, in particular:

- Assumptions or simplifications made for experimental purposes may be of unknown or unquantified validity; they may be implicit. Collection scale (in particular, numbers of queries) may be unrealistically small or fail to capture ordinary variability.
- Test collections tend to be specific, and to have assumed use-cases; they are rarely as heterogeneous as ordinary search. The processes by which they are constructed may rely on hidden assumptions or properties.
- Test environments rarely explore cases such as poorly specified queries, or the different uses of repeated queries (re-finding versus showing new material versus query exploration, for example). Characteristics such as "the space of queries from which the test cases have been sampled" may be undefined.
- Researchers typically rely on point estimates for the performance measures, instead of giving confidence intervals. Thus, we are not even able to make a prediction about the results for another sample from the same population. A related confound is that highly correlated measures (for example, Mean Average Precision (MAP) vs normalized Discounted Cumulated Gain (nDCG)) are reported as if they were independent; while, on the other hand, measures which reflect different quality aspects (such as precision and recall) are averaged (usually with a harmonic mean), thus obscuring their explanatory power.
- Current analysis tools are focused on sensitivity (differences between systems) rather than reliability (consistency over queries).
- Summary statistics are used to demonstrate differences, but the differences remain unexplained. Averages are reported without analysis of changes in individual queries.

Perhaps the most significant issue is the gap between offline and online evaluation. Correlations between system performance, user behavior, and user satisfaction are not well understood, and offline predictions of changes in user satisfaction continue to be poor because the mapping from metrics to user perceptions and experiences is not well understood.

2.2 Successes in Prediction in IR

The IR field has always had a strong evaluation focus. Because we are always trying to measure what we do, and furthermore working on analyzing the measures and the methodologies, we have a lot of experience in thinking about what we would like to predict. Also, IR is fundamentally about supporting people working to complete some kind of task. For example, modeling IR as a ranking problem already makes an assumption on how to present results and how users will access the output of the system. Even when evaluation is abstracted away from the actual user, we realize this measurement gap must be bridged.

Shared evaluation campaigns (TREC[1], CLEF[2], NTCIR[3], FIRE[4]) have always played a central role in IR research. They have produced huge improvements in the state-of-the-art and helped solidify a shared systematic methodology, achieving not only scholarly

[1] http://trec.nist.gov/
[2] http://www.clef-initiative.eu/
[3] http://research.nii.ac.jp/ntcir/
[4] http://fire.irsi.res.in/

impact [9, 103–105] but also economic impact [94]. The model has been adopted by other areas, and the IR field has successfully expanded into broader Information Access problems. Scalability has always been a major concern in the field that has been pushed by evaluation campaigns, and it is not as much a critical problem in prediction for IR systems as it is in other areas of Information Systems.

As a result of a strong evaluation focus, we have built a lot of datasets, and these datasets have closely related characteristics: common data types, common tasks, common experimental setups, common measures. This has let us appreciate the difficulty of predicting effectiveness on unseen data, tasks, or applications. There is extensive research on test collection building and evaluation methodologies, e.g. on robustness of the pooling methodology [117], the sensitivity and reliability of our measures [16,98], the impact of inter-assessor agreement [109], how many topics to use [97], just to name a few, although it is not easy to extract general lessons from it.

These test collections have allowed us to study what types of queries can be predicted to work well [27] and to discover other characteristics of queries (such a temporal distribution of the topic [66]) that can also be used to predict precision on some queries. Query performance prediction [19, 61] is thus concerned with predicting how difficult a query will be rather than the performance of a system for a given query but it can be a useful starting point for more advanced types of prediction.

Modeling score distribution, i.e. determining how relevant and not relevant documents are distributed, can be considered an another potential enabler for prediction, as also suggested by recent work which explicitly links it to query performance prediction [28].

On a more theoretical level, Axiomatics (the formal definition of constraints in a space of solutions for a problem) have been successfully used to predict the performance of IR models [34], to understand the properties and scales of evaluation measures [36–39] and to reduce the search space of available quality metrics [6–8].

Reproducibility is becoming a primary concern in many areas of science [48] and, in particular, in computer science as also witnessed by the recent ACM policy on result and artifact review and badging[5] [42]. Increasing attention is being paid to reproducibility also in IR [40, 118] where discussion is ongoing: use of private data in evaluation [18]; evaluation as a service [59]; reproducible baselines [75] and open runs [110]; considering it as part of the review process of major conferences and in dedicated tracks, such as the new ECIR Reproducibility Track; and, the inception of reproducibility tasks in the major evaluation campaigns[6] [43]. All these aspects contribute a better understanding and interpretation of experimental results and clarify implicit and explicit assumptions made during IR system development, which are key enablers for prediction.

2.3 Priorities for IR Experimentation

The considerations sketched out above, analyzed against existing successes, suggest four broad priorities that should be reflected in experimental methodologies: uncertainty, offline versus online, failure analysis, and reproducibility. Other aspects include use features (of topics, documents, and context), the roles of measures, domain adaptation, and more application-specific issues such as individual queries versus sessions. We consider each of these in turn below.

[5] https://www.acm.org/publications/policies/artifact-review-badging
[6] http://www.centre-eval.org/

2.3.1 Priorities

Uncertainty

Our measures typically produce a point estimate, without confidence intervals or effect sizes. Statistical significance is not predictive, and does not quantify uncertainty, although researchers use them that way. We need measures that indicate bounds as well as averages, where we can indicate confidence bounds in the performance of a system on unseen data.

Offline versus online

Offline metrics at best weakly predict online effectiveness and user satisfaction. We need to understand how online effectiveness can be predicted more reliably, and what factors are responsible for the inconsistency. A particular factor is the single-query nature of most offline evaluation, while online experiences are iterative or progressive, that is, involve a session. Thus, the result of the complete session is what matters for users.

Failure Analysis

Failure analysis typically focuses on individual tasks where performance is extremely bad. The RIA workshop [15] followed this approach, but did not arrive at general conclusions for improving the systems considered. Instead, it was acknowledged that there are topics of varying difficulty, and thus various approaches for estimating query difficulty have been proposed (see e.g. [91]). However, the core problem is still unsolved: which methods would be suitable for improving the results of 'difficult' queries?

Reproducibility and replicability

The ACM policy on Artifact Review and Badging[7] distinguishes between replicability (different team, same experimental setup) and reproducibility (different team, different experimental setup). Reproducibility is a key ingredient for prediction since not only it enables the systematic replication and understanding of experimental results – a key aspect to ensure robustness of prediction – but also studies how robust experimental results can be ported to new contexts and generalized. However, we still lack commonly agreed methodologies to ensure the replicability, reproducibility and generalizability of experimental results, as well as protocols and measures to verify and quantify the reproducibility of experimental results.

2.3.2 Other open issues

Measures and resources

It is clear that measures vary with regard to predictability. We need to develop good practice recommendations for selecting and using evaluation metrics: which metrics are suitable for a given task, scenario, or dataset? How should we interpret inconsistent quality signals? How should we deal with multiple, complementary quality signals (e.g. Precision and Recall)?

System comparisons are somewhat stable on typical small sets of topics, but concerns about the sampling population mean that to increase our understanding we need vastly

[7] https://www.acm.org/publications/policies/artifact-review-badging

larger topic sets; and arguably these should be characterized by kind of task and kind of interaction.

Finally, we also need a better understanding of what IR evaluation measures are and what their properties are. Indeed, we need to go beyond what we empirically [16, 17, 96] and theoretically [6, 7, 36, 38, 100] know today about IR evaluation measures and turn this into knowledge about evaluation measures affect prediction.

Contexts

What factors affect domain adaptation? Is it reasonable and effective to consider a domain as comprised of a number of tasks, where each has its own success criteria that, in turn, is reflected into a measure? What would constitute an actionable description (by means of features, tasks, collections, systems, and measures) of what is required to move from one domain to another? A specific example is the difference between language-dependent and language-independent factors, both at system level and at domain level, since they may require different kinds of prediction techniques.

3 Natural Language Processing

Current research in NLP emphasizes methods that are knowledge-free and lack explanatory power, but are demonstrably effective in terms of task performance. This has the consequence that small changes in the application scenario for an NLP system has an unpredictable impact on performance. We need to make the process of developing NLP systems more efficient.

3.1 Motivations for Prediction in NLP

We regard predictability of Natural Language Processing (NLP) system performance as the capacity to take advantage of known experiences (methodologies, techniques, data) to minimize the effort to develop new high performing systems. A key issue that impacts our ability to predict the performance of an NLP system is *portability*. Under this perspective we need to consider the following portability aspects:

- cross-language portability
- cross-corpus portability
- cross-domain portability
- cross-task portability

As an anecdotal example that motivates the interest in predictability, we can look at a project [52] for automatic classification of radiological reports in a hospital department. It was developed as a supervised system, which required a significant annotation effort by domain experts. The same technology was then proposed to the same department of another hospital, which asked for an estimate of the annotation effort, i.e. time of domain experts needed for adapting the system to a different classification schema. At this point it became clear that there was a lack of predictive methodologies and tools. At the end of the day, the new hospital was not convinced to invest in the technology due to the unclear investment that would be required.

Another anecdotal example with a more positive outcome is the Software Newsroom [65] which is a set of tools and applied methods for automated identification of potential news from textual data for an automated news search system. The purpose of the tool set is to analyze data collected from the Internet and to identify information having a high probability of containing new information. The identified information is summarized in order to help understanding of the semantic content of the data, and to assist the news editing process. The application had been developed for English and initially did not transfer well into Finnish. The problem was attributed to the fact that data was collected from Internet discussions and that the language was probably substandard. Attempts to fix this did not yield performance improvements. Later it became clear that words with certain syntactic and semantic properties are effective when building topic models for English, at which point it could be demonstrated that words with similar properties in Finnish are useful as well. Correctly extracting such words required knowledge about the special characteristics of the Finnish language.

A challenging aspect of typical NLP components, e.g. part of speech tagging, named entity recognition, parsing, semantic role labeling, is that they require a significant amount of human supervision, in the form of annotated data, to train reliable models. This an issue clearly impacts the portability of both individual components and more complex systems that depend on pipelines of such components. Several efforts are being made in the NLP field to reduce and to predict the amount of such supervision, moving towards less supervised algorithms. We mention a few of these research directions:

- the use of unannotated data and distributional representations of words, i.e. embeddings, as features for machine learning algorithms;
- distance learning approaches, exploiting the role of available resources, e.g. taxonomies, dictionaries, background knowledge to infer training examples;
- active learning techniques, in order to select instances to be manually annotated to optimize the performance of a system;
- projections of annotations across aligned corpora, from one source language (typically English) to a target language, to reduce the effort to develop training data.

Although the above research streams are producing significant advancements in terms of portability, we feel the need for fundamental research where predictability of NLP systems is addressed in the broader context of cross-language linguistic phenomena, characteristics of corpora, domain coverage and particular properties of the task.

3.2 Successes in Prediction in NLP

One traditional technique for predicting performance is to perform post-hoc data degradation. In TREC-4 (1995), the 'Confusion Task' compared performance of query retrieval using corrected OCR text against text with 10% and 20% recognition errors [67]. In this way, given an evaluation of the recognition rate of an OCRed collection, one could predict the performance degradation compared with a corrected collection. Similarly, TREC-9 analyzed the effect of spelling errors on retrieval performance, and the absence of word translations in cross-language information [79]. More recently, this method of post-hoc corpus degradation was used to show that at least 8 million words of text is needed to achieve published results in word embedding tasks, such as similarity and analogy [56].

This degradation technique provides a negative prediction of relative performance to a known system and known input testbed, but does not allow us to predict how well a given technique will work on a new language, or a new corpus, or a new domain.

Retrospective analysis can show that different domains have different measurable characteristics that correlate with some system performance. For example, biological texts have a greater entropy that correlates with a degraded performance of named entity recognition compared with performance on edited newspaper text [85]. Word sense disambiguation has been shown to degrade across a number of factors that can be calculated before experimentation [114].

The Software Newsroom [65] is an example of adapting a news discovery system from English to Finnish where there was a need to know linguistics and language technology to understand which parts were similar and which parts needed to be adapted. Similarly, evaluation of NLP tools applied across domains demonstrates that adaptation is required to port tools, e.g. from general English to a more specialized context such as biomedicine [107].

Work in adapting NLP technology to new languages, particularly to low-resource languages, begins with the problem of complex requirements for building NLP systems, including both annotated data sets and tools for analysis of linguistic data at various levels, such as the lexical, syntactic or semantic level. Recent research in *transfer* or *projection* learning has shown that it is possible to leverage data in one language to develop tools for the analysis of other, even quite linguistically distinct, languages [31]. However, this research has also demonstrated the need for resources to facilitate transfer, ranging from complementary resources such as parallel corpora and bilingual dictionaries to broad overarching frameworks such as the Universal POS Tagset [88] and the Universal Dependency representation [84]. In short, NLP system development requires either task-specific annotated data sets, a strategy for inferring annotations over data that can be leveraged, or a framework that facilitates model transfer through shared representation.

Current attempts at learning morphological reinflection for various languages have met some initial success using Deep Learning where it has been shown that approx. 10000 training cases can render a performance around 95% correct results for many languages [25]. Some of the systems benefited from additional unannotated data to boost performance.

In a keynote talk at GSCL 2017 (http://gscl2017.dfki.de/), Holger Schwenk (Facebook, Paris) presented recent advances in deep learning in the field of NLP, showing how Machine Translation could be understood as a cross-lingual document search. As deep learning is performing feature extraction and classification in an automatic fashion, this technology can be deployed in various NLP tasks, for example machine translation. Word embeddings, neural language models and sentence embeddings are leading to an application of multilingual joint sentence embeddings, supporting high quality translation. The further development of such approaches could lead to a better integration of NLP systems, using the generated vector spaces for cross-language, cross-corpus, cross-domain and cross-task information sharing.

3.3 Priority Next Steps in NLP Research

We believe that, in order to improve predictability of NLP systems, research in the next years should focus on innovative, fine-grained, shared, methodologies for error analysis. We advocate evaluation measures as well as techniques able to provide both quantitative and qualitative data that explain the behavior of the system under specific experimental conditions. We expect to move from ad-hoc and mainly manual error analysis to shared and automatic tests through which we determine reliable predictability indicators.

Particularly, it is expected that new error analysis methods can provide empirical evidence of system failures based on the whole complexity of the context in which the system operates,

including linguistic cross-language phenomena, the properties of the data (corpora, resources, knowledge), the domain characteristics, the specific task the system is supposed to addresses, and the role of the human users that interact with the system. Test suites, as discussed in Section 5.5.2, could support error analysis structured by cases.

The kind of expected error analysis has to be enough fine-grained to give precise insight about the causes for a system/tool not to deliver the expected results, including:

- Are the corpus/data sets or other (domain) resources appropriate?
- Has the right/adapted algorithm been selected?
- Is the selected/developed gold standard the relevant one?
- Are we using the right metrics/measures?
- Are we using the right amount of (linguistic) knowledge?
- Are we using the right type of representation of the data and the features, also in combination with data/tools that are not specific to NLP?
- Check the validity of the assumptions of what the system should deliver and at what level of quality, taking into consideration IT-performance, but not focusing only on the measures.

3.3.1 Desiderata

In the last few years, the increasing availability of large amounts of language data for a subset of natural languages as well as the availability of more powerful hardware and algorithmic solutions, has supported the re-emergence of machine learning methods that in certain applications, for example Neural Machine Translation (NMT), has relevantly improved performance in terms of objective measures. In the light of those developments, we need to re-think the way we develop and deploy NLP systems, taking into account not only linguistic knowledge but also technological parameters. Conversely, excitement about the performance of neural network approaches should not close our eyes on specific linguistic features and language properties. We need to embark on a new theory of the field of natural language processing.

To sum up, the NLP field is missing a comprehensive diagnostic theory for NLP systems. A consequence of using powerful diagnostic tools will be a substantial rethink of the way we develop and make NLP systems more adaptable to new languages, data, tasks, applications and scenarios, including when this involves other types of technologies. A long-term opportunity in this direction is that of NLP systems able to auto-adapt themselves to a changing environment, predicting adaptations on the base of diagnostic tools.

4 Recommender Systems

4.1 Motivations for Prediction in RecSys

Introduction and History. Even in the earliest days of recommender systems, predicting performance was seen as critical. The earliest recommender systems companies hired "sales engineers" whose job was to evaluate the potential gain prospective customers would have from deploying a recommender in their applications. A typical example was reported in talks by John Riedl. The recommender systems company Net Perceptions sent a team of sales engineers to work with a large catalog retailer. To make the sale, they had to import the retailer's database into their system and run side-by-side experiments with phone operators making suggestions from the legacy or new system. It was a multi-week, multi-person effort

that fortunately led to a successful sale and deployment. Not all such efforts were successful, highlighting the desirability of predictive models of performance that can more efficiently support deployment and tuning decisions. This section reviews examples of cases where such prediction is needed.

Case 1: ROI Improvement for Mobile News. A company develops start screens for mobile devices providing news items that are pushed to their users once they turn on their devices. They have a few million users and agreements with news agencies in various countries. Typically they provide a list of 8 items when the user turns on the device. Their business model is based on user clicks, so they are interested to improve the Click-Through Rate (CTR) and user engagement. They want to examine whether personalization of the provided list of items would improve these measures and if the investment in the development and implementation of personalization would be returned (ROI). Currently they provide the list based on the nationality of the user, recency of items and some notions of popularity. Several algorithms were considered: variations of content-based and collaborative algorithm, and diversity to expose more items and enhance the ranking of popular items. Performed off-line analysis yielded interesting results that were not always consistent for different parts of the data. A/B tests are very costly and can be done very selectively, since they don't have an experimental infrastructure, thus deploying algorithms and testing different variations places huge effort on production. The challenge is to predict which algorithm or the combination of algorithms would provide the expected ROI. Is it possible to predict for the company if they should invest in personalization. Can that be inferred from the offline test results, from the dataset, task, algorithm features, or from success and failure stories.

Case 2. To Personalize at All? An online education company has a large and expanding library of courses, and currently has no personalized mechanisms for recommending courses to their learners. Their system is based on three forms of discovery: (a) search for courses that match relevant keywords, (b) lists of most-popular courses, both overall and within broad top-level categories (e.g., "most popular computer science courses"), and (c) marketer-selected lists of courses to promote in themes (e.g., the April theme was "new beginnings") with a set of promoted introductory courses in different categories. The company is interested in determining whether there would be significant benefits to adding a personalized recommender system to their site.

Given the characteristics of the education company's dataset (number of learners and courses, distributions of courses-taken and learners-enrolled, etc.), can we model and predict the performance of a recommender system for this application? Today, we cannot. Our choices are to offer "advice from experience" or to offer instead to go through the data engineering effort to implement the recommender in order to justify its feasibility. Neither is a particularly satisfying alternative for a company (or expert) hoping to make an informed decision to invest without incurring substantial cost.

Case 3. How Much Value Do We Have from Certain Data? Data collection is both expensive and potentially interfering with the privacy of clients. With increasing regulation on which data and how data is stored, such as General Data Protection Regulation (GDPR) in the EU, this may also be difficult in practice. In some cases, however adding the right amount and kind of data can improve the quality of predictions. At the moment, we do not have a formal way of assessing how much, and what kind, of data will translate to a specific return. While there is some heuristic consensus on which dimensions may be relevant, and that these depend on dimensions of the domain, client, and tasks, this knowledge is not systematically structured or cataloged.

4.2 Successes in Prediction in RecSys

This section outlines key areas of success with regard to prediction in recommender systems, outlining the state-of-the-art and gaps to motivate the priority areas in Section 4.3.

This section discusses the following topics.

- Noise and inconsistency
- Data Sets
- Metrics and Evaluation Protocols
- Toolkits
- Subjective evaluation
- Meta-learning

4.2.1 Noise and inconsistency

Accuracy of prediction is limited by the by noise (e.g., so called *shilling malicious ratings* [74]), anchoring effects due to original ratings [2], as well as by the inconsistency of rating by end-users [4].

Progress has been made in terms of detecting noise, and in the development of de-noising techniques both in terms of algorithms [5] and interface design [1]. There is an understanding that prediction accuracy may be restricted by the upper bounds of such factors. However, the nature of noise and its role in relation to prediction accuracy is not completely understood. For example, it is still not clear to which extent changes in rating behavior are due to inconsistency, versus how much reflects a genuine change in opinion.

4.2.2 Data Sets

Recommender research been advanced by many public data sets containing user consumption data, suitable for training collaborative filters and evaluating recommender algorithms of various forms. These have enabled direct comparison of algorithms in somewhat standardized environments. These include:

- EachMovie [78], movie ratings from a movie recommender system operated by the Digital Equipment Corporation (DEC).
- MovieLens [60], a series of movie rating data sets released from the MovieLens movie recommender system operated by GroupLens Research at the University of Minnesota. Recent versions include the Tag Genome [108], a dense matrix of inferred relevance scores for movie-tag pairs.
- Jester [54], a set of user-provided ratings of jokes.
- NetFlix [13] (no longer available), user-provided ratings of movies in the NetFlix DVD-by-mail system; this data set was the basis for the NetFlix Prize, which awarded $1M for a 10% improvement in prediction accuracy over NetFlix's internal recommendation algorithm.
- BookCrossing [116], book ratings harvested from an online book community.
- Yahoo! Research publishes a number of data sets, including movie ratings and music ratings.
- CiteULike provided access to user bibliographies of research papers.
- Amazon rating data collected by [62].
- Yelp regularly provides data sets of business ratings [115].
- The Million Song Dataset is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. [14].

In addition, the RecSys Challenge has regularly made new data sets on news, jobs, music, and other domains available to the research community, and there are suitable data sets through a number of Kaggle competitions and other sources, such as the NewsReel labs within CLEF [70, 76].

The "challenge" format around many data sets has provided a boost of energy in recommender systems research, with teams competing to provide the best performance around a single data set.

4.2.3 Toolkits

The recommender systems research community has a long history of publicly available or open-source software for supporting research and development. Early work on item-based collaborative filtering was supported by SUGGEST [30, 69]. Throughout the last decade, a number of open-source packages have been developed. Currently-maintained packages that are used in recommender research include LibRec [58], RankSys [99], LensKit [32], and rrecsys [119, 120]; Rival [95] provides cross-toolkit evaluation capabilities. These toolkits provide varying capabilities: some focus on algorithms or evaluation, while others provide both; some support primarily offline operation and batch evaluation while others have direct support for live use in online systems. There have also been a number of toolkits in the past that are no longer being maintained, such as MyMediaLite [51], and others that have pivoted away from a focus on recommendation such as Apache Mahout [10], in addition to algorithm-specific packages such as SVDFeature [22] and non-recommender-specific software such as XGBoost, Torch, and TensorFlow.

4.2.4 Subjective Evaluation

The goal of a recommender system is to provide personalized support for users in finding relevant content or items. If we want to predict or model whether that goal is actually achieved, researchers have realized that we should move beyond accuracy metrics to see if algorithmic improvements actually change the experience users have with the system [80]. This has led to several conceptual models that also provide subjective measures and scales of users' quality perceptions and evaluations of recommender systems [89, 113]. Building on this earlier work and other work on technology acceptance and attitude models, [73] argue that their user-centric framework [72] provide an "EP type" theory than can [E]xplain and [P]redict user behavior given the specific conditions of the recommender system under investigation. In other words, the framework goes beyond user studies that only qualitatively inspect user satisfaction or large scale A/B tests that only quantitatively look at the impact of a system change on user behavior. It aims at determining the factors why particular experimental conditions (i.e. a change in objective diversity of the algorithm) can change user experience (i.e. choice satisfaction) and user interaction behavior (increased engagement) by looking at the intermediate concept of subjective system perceptions (e.g., subjective perceptions of diversity and accuracy). For example a study [112] shows that user perceptions of accuracy and diversity mediated the effect of the diversification of recommender output on the experienced choice difficulty and user satisfaction showing that only if these subjective perceptions are changing, we can predict user experience to change.

4.2.5 Meta-Learning

Recommender system algorithms can be complex. Usually, they are configured specifically for specific data, users, and tasks and are optimized for specific desired measures. The construction and tuning of RecSys algorithms is typically done manually by human experts through try-and-err testings. It is desired to automatically explore the vast space of possible algorithms with the vision of enabling the prediction of which algorithms will perform well for a given dataset, a set of users, task, and performance metric using meta-learning techniques. A combination of features extracted from a given dataset, task, users, along with algorithm configuration, and the discretized result of a specified performance metric would make a labeled meta-training learning instance. One major challenge would be to learn the set of features of the dataset, users, tasks, that can affect the results. Another major challenge would be to collect enough instances for a large variety of datasets, algorithms, tasks and measures that would enable valuable learning. For this challenge the following possible sources can be considered: 1) a corpus of datasets and the corresponding learning results that will be provided collected by the community as a joint effort that should be promoted 2) data and information extracted from machine learning competitions (e.g. Kaggle). It may be possible to extract relevant information also from academic paper results. Nevertheless, to address the challenge of learning in the vast search space of possible algorithm and their specific configurations, ML techniques should be designed to allow a system to learn and capture insights and experiences in order to guide the selection of algorithms. Previous research showed that meta-learning can be successfully used for selecting the best model for decomposing large learning tasks such as [93], selecting the best setting for multi-label classification tasks and even recently for selecting the best collaborative filtering model for recommender systems However, to have the supervised meta-learning successful, a joint community effort to collect learning instances could be beneficial.

4.3 Priority Next Steps in RecSys Research

This section outlines the identified priority next steps. These are organized into three broad categories:

- Developing the Foundations for Rigor
- Learning from Cumulative Research
- Specific Challenges

4.3.1 Developing the Foundations for Rigor

There are several prerequisites that the field needs to achieve in order to place the remainder of the research we propose on a rigorous foundation.

4.3.1.1 Taxonomizing Cases

Today recommender systems are used in many domains and for different purposes. For example, in addition to recommending content for consumption, researchers have also started using those systems to incentivize user to create content e.g. [53]. Given this broad spectrum of use cases and applications, performance evaluation protocols are often tailored or anchored in the context of the recommender system use case. In order to learn from these cases, we need a rigorous and consistent way of describing them.

There are two levels of description needed in order to facilitate rigorous learning from cases. The first is agreement on the things necessary to describe a *case*, which we take to mean a set of research findings in a particular recommendation situation. The case consists of at least the following properties:

- The domain
- The system or experiment goals
- The target users
- The user task(s) within the domain
- The user's characteristics considered for recommendation
- The data
- The algorithms
- The experimental design and evaluation protocol (online or offline)
- The metrics and statistical analysis

The second level is the means of describing each of these properties. Significant research, detailed in later sections, will be needed in order to make this possible. For example, we need to know what properties of a user task and characteristics need to be captured in order to facilitate generalization and learning. Different tasks can have very different requirements, changing even the direction of certain optimization criteria; while recommending songs in a music recommender that a user has listened to before will usually have a positive impact on user satisfaction, recommending old news in a news recommender system will have a negative impact on satisfaction. We believe it is critical for the future success of recommender systems to develop a taxonomy of tasks that will lend itself to create task models connecting user goals of recommender systems with their objective, subjective evaluation metrics and output metrics. Herlocker et. al. [63] developed an initial description of end user goals and tasks of recommender systems (e.g. annotation in context, find good items, recommend sequence, ...). We suggest evaluating this research as a starting point for the development of a more formal taxonomy and evaluate literature since then to also develop different dimensions of a taxonomy of tasks (e.g. domain, time-sensitivity, content creation, or cost to consume).

Developing this shared understanding will enable the field to move forward and develop predictive knowledge from the current and future body of research findings.

4.3.1.2 Standardizing Evaluations

To ensure the replicability and comparability of results, the field needs to establish standards for measures and evaluation protocols. There are too many different ways to calculate what is labeled as the same measure. Even well-understood metrics such as Root Mean Square Error (RMSE) have important differences in their application: how unscorable items are handled and whether scores are averaged per-user or globally. The community needs to establish standard definitions and protocols for both metrics and for best practices in managing the broader evaluation (e.g., how to split data for cross-fold validation to evaluate performance on common tasks). Further research is needed to establish some of the standards, for example how best to mitigate popularity bias [12] and misclassified decoys [26, 33] and the role of time in splitting data sets. As metrics are standardized, the community should also provide standard test cases for use in acceptance tests to validate new implementations. A standard resource for recomputing or labelling historical results could also be used to assess new individual implementations.

There will likely always be the need to occasionally deviate from standard calculations to answer particular research questions. A standardization effort, however, can lay out the

option space and describe defaults and best practices for standard tasks; authors should provide clear justification when they deviate from the defaults the community agrees upon.

The standardization also needs to lay the decision points in experimental designs, such as what happens when an algorithm cannot produce recommendations for a user: does it get 'forgiveness' and have that user ignored in its assessment, does it get penalized as if it recommended nothing useful, or does it use some fallback method to ensure all users receive recommendations? There is not necessarily a best answer to these questions. Community guidelines should lay them out so that authors are aware that they need to make, describe, and justify a decision about each of them instead of relying on accidental properties of implementation details, and guidance on sensible defaults for common recommendation scenarios would help future researchers.

As the community decides upon standards, toolkit implementers should implement them and ideally make them the default operation, with appropriate thought given to backwards compatibility for their users.

There are two additional immediate first steps to promote rigor even before standardization is achieved. **Paper authors** should report sufficiently complete details on their evaluation protocols, algorithm configuration, and tuning to enable readers to reproduce them with exact replication of original decisions. **Toolkit implementers** should document the expected evaluation results of well-tuned versions of their algorithms on common data sets to provide a reference point for authors and reviewers to assess claimed baseline performance.

4.3.2 Learning from cumulative research

To enable us to learn from previous and ongoing research, we are assuming that there is a definition of a case. Such a case may include a description of the case, a description of the dataset (underlying assumptions, algorithm parameters, outputs, etc.) as well as a link to the resulting paper).

This repository should be both collaborative and machine readable. Users may add and remove content, with moderation of who can edit the repository, and how information is removed, to ensure completeness and consistency. A not trivial aspect in building this repository is the analysis of the existing body of research in order to transform it to standard case descriptions; this can be done either manually, requiring quite a lot of effort, or automatically, even if this is going to be a challenge. However, once this repository has been bootstrapped, adding new cases will require just a low-cost (for humans) and standardized procedure to be followed.

The repository will open up several new interesting possibilities:

- Meta-analysis is a well-understood technique in domains such as medical studies, where statistical confidence accrues through aggregating the evidence from numerous research studies. Medical research has the benefits of both more controlled studies and a long tradition of publishing results in a manner that explicitly supports such meta-analysis. Recommender systems will need to evolve its research to support these techniques, including such steps as:
  - developing standard templates for reporting results from recommender system evaluations and experiments and a disciplinary culture of reporting these with research results
  - developing characteristic parameters for datasets, users, and tasks
  - developing and testing predictive models over the diverse characteristics of these
- Creating a mechanisms for comparing commonalities and variabilities among cases, in order to support researchers in making decisions. This will help us identify recurring successful

cases, and failure cases, and the characteristics of both kinds of cases. Moreover, analysis of previous cases will also allow us to identify aspects of cases that are not sufficiently covered by current research. With time, this mechanism may become increasingly automated.

- As we aim at being able to predict the level of success of a recommender system before using it and even before developing it, failure analysis becomes of major importance as a tool to fix our prediction models. Given the actual results vs. the expected we should be able to reason about the causes of the mismatch or failure. Indeed, failure analysis/correction can be done either by analyzing the model that was discussed or by adding it as a case for deep-learning based meta-analysis where cases of systems that include data characteristics, task characteristics, algorithmic characteristics objective and subjective measures are analyzed to identify "successful" or "to be" systems.

Finally, to keep the repository a primary tool for research, we need to envision mechanisms to encourage contributions to it such as dedicated workshops and tutorials, both physical and online. These workshops will focus on cases analysis in the areas identified as gaps or failures. These workshops will also enable us to share what can be learned from the repository, as well as continue to help grow the case base.

Properties of data sets and algorithms

Different recommender algorithms attempt to exploit different properties of the data, e.g. user-based collaborative filtering leverages the assumption that people who are similar will like similar things. Content-based filtering uses textual features to represent items with the hopes of finding related items. However, the performance of these individual algorithms will depend heavily on the data set and the distribution of the properties being exploited and how they relate to each other. For example sparse datasets mean neighborhood algorithms perform poorly unless some latent factorization model is employed. Similarly, if content filtering is using a textual representation of items or people but if those base feature vectors don't accurately represent or reflect the items or user preferences then no tuning of the similarity measure or ranking will create significant additional value for the user.

Connecting this back to a more formal task taxonomy, we need to create a clear list of measures and distributions that will help guide the identifying the right family of algorithms based on the properties of their dataset and the class of task to be evaluated. By linking the performance of individual classes of algorithms, simply looking at the distribution and measures of the data can help predict performance and identify possible weaknesses before requiring a complete end-to-end evaluation.

Predictability limitations in data sets

Differing assumptions, measures and testing strategies lead to wildly varying performance across datasets. One issue which can be addressed is exploring how much information can realistically be exploited by any algorithm based on the properties of the data. This issue has started to be addressed in the complex networks community, for example Song et. al. explored the **limits of predictability** in human mobility to test the assumption on whether prediction is truly possible [101]. More recently Marting *et al* [77] have asked the same question for social systems in general highlighting that "*the central question of this paper — namely to what extent errors in prediction represent inadequate models and/or data versus intrinsic uncertainty in the underlying generative process — remains unanswered*". If a clear process was in place to test the limits of predictability in the underlying data, then circumstances where experiments demonstrate unrealistically high prediction rate can be

identified as having some flawed experiment design properties. For example if a recommender algorithm can accurately predict a users movie choice 99% of the time but the underlying data shows almost random behavior, then something is awry.

4.3.3 Specific Challenges

This section identifies a set of specific challenges that are central to advancing the goal of predicting system performance or to achieving the above priorities related to rigor and learning.

Auto-tuning. Adding to the challenge of recommender system selection and performance prediction is the difficulty of tuning the underlying algorithms. Nearest neighbor algorithms include a variety of parameters related to neighborhood size, weighting based of extent of commonality, and in the case of model-based approaches the size and truncation of the model. Similarly, latent factor models have extensive parameters in both training and use.

A consequence of the challenge of tuning is that researchers often fail to compare like with like. As parameters can depend on data distributions, it is increasingly important to identify standard ways to tune algorithms–and particularly baseline algorithms for comparison. Fortunately, tuning can be framed as a combination of parameter space exploration and understanding the response curves and sensitivity to parameters. With a systematic exploration of algorithms, we should also explore empirically-tested auto-tuning to ensure both fair comparisons and more efficient exploration.

Exploration vs. Exploitation. Two key challenges in recommender systems are discovering changes or inaccuracies in models of user preference and the cold-start item problem of learning to recommend new items. Both of these can be addressed interactively by presenting users with some set of "exploratory" recommendations – recommendations based not on their current tastes but on the system's need for information. For example, a music player may identify a target set of users to whom a new song should be played to identify the right audience for that song. Or a news recommender may periodically recommend a randomly selected news article to validate or update the user's preference model.

The trade-off between exploration and exploitation–both algorithmically and as a matter of user experience–needs further study. In particular, we may need to create metrics of "realistic user experience" that incorporate system-wide exploration as well as targeted exploitation of profiles.

Temporality and Dependency. Little work has been done in the recommender system community to address changes of user preferences over time, for example, Moore et al. [82] modeled temporal changes of preferences in music recommendation. We need to more systematically explore how we can detect patterns of change and exploit those in our performance prediction models in multiple domains across our use cases. Closely related to the challenge of temporal changes of user preferences is the ability to understand how recommendations based on user preferences influence and change preferences, i.e. we need to take this dependency into account in our performance models. For example, low level measures such as diversity or exploratory recommendations will have an impact on preferences.

Underlying theoretical assumptions in recommender algorithms. Recommender algorithms (and evaluation protocols) are built on statistical and mathematical models that incorporate underlying theoretical assumptions about the distributions and patterns of data. These range from the high-level assumptions of all collaborative filtering algorithms (stable

or predictable preferences, past agreement predicts future agreement) to more complex distributional assumptions (e.g., exponential popularity distributions to support neighbor-finding) to issues of temporal stability (e.g., whether offline evaluation has to be ordered vs, random). These theoretical limitations are often at most informally expressed, and they are rarely explicitly checked or analyzed. Rather, experiments are put in place, and empirical evidence is drawn from them confirming or refuting the effectiveness of recommendation approaches at the end of the algorithmic development pipeline.

An explicit and precise identification and a deeper understanding of the essential assumptions would help assess and document the scope of algorithmic performance evidence and predictions. In practice, recommender algorithms have often "worked well enough" when assumptions are violated, but such boundaries should be tested and understood.

We find it a worthy endeavor to research what the precise (core, simplifying or otherwise) assumptions in the algorithms really are; finding means for checking them in particular cases (data, tasks, users, etc.); and understanding the impact in the algorithm effectiveness to the extent that the assumptions are not met, or not fully. This should help enable and guide principled algorithmic development, diagnosis, deployment and innovation, beyond just assumption-blind trial and error.

Likewise, we should understand whether, to what extent or in what direction the biases may distort the experimental measurements. Further implicit assumptions are made on the purpose for which a recommender system is to be deployed when evaluation metrics are developed and chosen. Understanding and analyzing the consistency between metrics and the ultimate goals the system is conceived for are key to make sure the right thing is being measured.

Algorithm vs. System Performance. One of the major issues in evaluating the performance of a RecSys in a realistic setting, by real users, is the users' inability to distinguish between the system as a whole and the recommendation algorithm itself. Indeed, many of the most successful advances in studying algorithm differences have come from individual research systems where the same system interface could be used with different algorithms.

A challenge today, however, is general lack of access to such systems for the typical researcher. While industry has access to large user bases, companies rarely will allow external researchers to experiment with those users.

We therefore propose a community-wide effort to build and maintain a high-quality, usable recommender system specifically to support the research community. This system would have the ability to integrate different algorithms, and would include instrumentation to allocate users to experimental conditions, record user interaction, log system performance, and administer user experience surveys where needed. Most important, it would report metrics and export data according the to the community-agreed standard.

Such an effort could be launched *ab initio* or could involve creating a consortium to enhance, open, and maintain pre-existing recommender systems for the research community.

5 Cross-Discipline Themes

In order to predict performance, a number of research issues has to be addressed central to all domains (IR/NLP/RecSys), which are sketched in figure 1. First we have to choose the performance criteria and define corresponding *measures*. When performing experiments with different test collections and observing the system's output and the measured performance, we will carry out a *performance analysis*. For that, we will look at violations of the *assumptions*

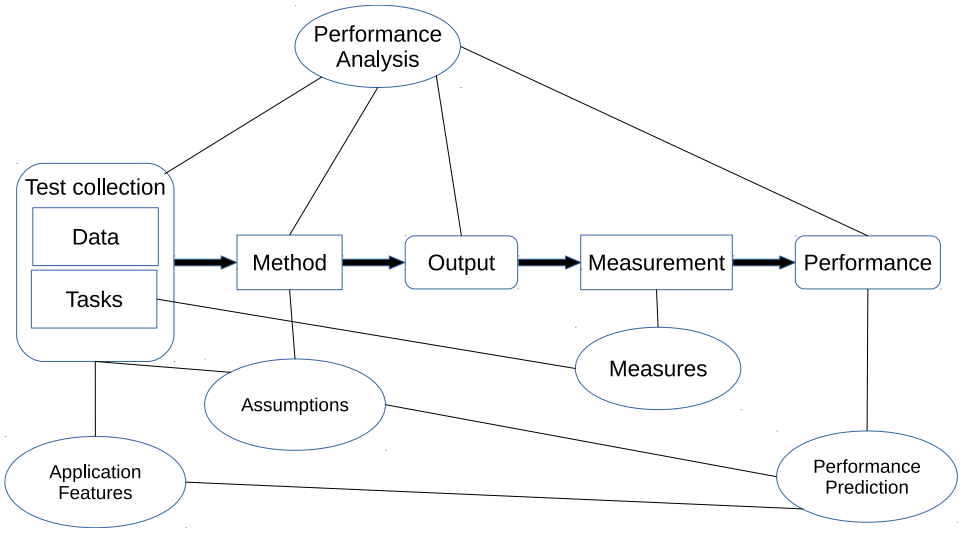


Figure 1 An overarching performance prediction framework.

underlying the method applied. Also, characteristics of the data and tasks will have an important effect on the outcome. Finally, we aim at developing a *performance prediction model* that takes these factors into account.

Different fields have traditionally focused on evaluating specific aspects in this framework, but we believe that understanding the relations between these tasks is essential in achieving adequate performance prediction. Moreover, we have mentioned several times the importance of reproducibility for improving our experimental evaluation practices. It should be understood that reproducibility is just another side of the coin when it comes to performance prediction. Indeed, the possibility of replicating the same results in the same experimental condition, the capability of reproducing them in different conditions, and the ability to generalize them to new tasks and scenarios are just another way of formulating the performance prediction problem.

Once these tasks are well understood we can begin to try and predict performance in an unseen situation if enough of the above still hold. Expecting to be able to test and tune all aspects of this pipeline is a limiting factor for exploring new ideas and solutions. It is our hope that by abstracting stages within the framework, recurrent patterns will emerge to support prediction for unseen cases (combinations of the above aspects).

In the following, we discuss each of these aspects in detail. Besides describing open research issues, we will also point to out some cases where weak current scholarly practices impair our understanding of the matter.

5.1 Measures

5.1.1 What and Why

In this section, we focus on two aspects, namely the definition of the low-level metrics, and the link between low-level and system-level performance evaluation, meant as the connection between more objective and engineering-like measures with more subjective ones, ultimately representing the user satisfaction and experience with a system.

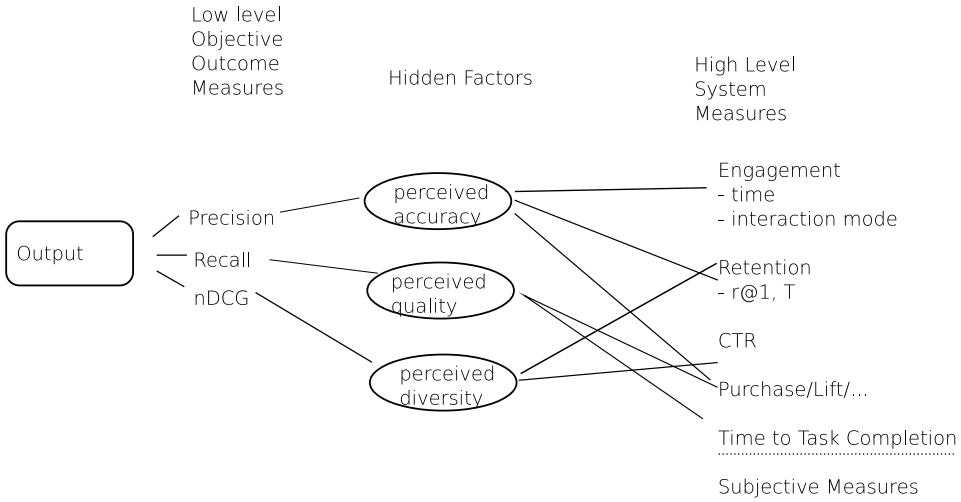


Figure 2 Low- and high-level performance measures.

Metrics definition

The definition of a metric relies on several alternatives and decisions, which happen before the actual measurement takes place, also to avoid any post-hoc bias.

We first have to choose the criteria that reflect the goals of our evaluation, for instance relevance, diversity, or novelty. However, as said before, the performance of a system is not only a matter of goals but also of the "utility" delivered to users. Therefore, we need to identify/choose a prototypical user behavior; for example, when ranking is involved, a stopping point, after which no more recommended items or retrieved documents are considered, introduces a clear separation between seen and unseen items, where only the former influence the measurement outcome. Instead of a deterministic behavior, we might also assume a stochastic model for this aspect as done, for example, by [21, 35, 81]. We also have to define the user's preferences concerning the items seen, like e.g. the total number of useful items, or the ratio between useful and useless items.

Finally, we have to choose an aggregation method like arithmetic or geometric mean, where the former focuses on absolute differences, and the latter on relative changes, paying attention that the aggregation method is admissible when considering the scale properties – ordinal, interval, ratio scales – of a measure [39, 50]

Overall, current research often neglects the fact that each metric represents a specific user standpoint, and often the standpoint may be context-dependent. Thus, an evaluation focusing on a single low-level metric will either ignore many user standpoints, or represent an intransparent mixture of different standpoints.

From low-level to system-level performance

Figure 2 shows a closer look at the measurement aspect, where we distinguish between low-level evaluation measures and expected high-level system outcomes. We see the range of inputs and performance measurements that reflect the performance of a diversity of systems, including IR, NLP, and Recommender Systems. In this section we focus on the link between low-level and system-level performance evaluation, and in turn on the challenge of not just building predictive models but also incorporating and building a deeper understanding of the factors that lead to system-level results.

This deeper understanding is essential to crafting complete systems. For example, to have more effective automated summarization, we need to understand what low-level properties of these summarizations lead users to perceive fluent text. Similarly, this enables us to understand how diversity of a search or recommender result affects the user's confidence in having found the correct result, or help them learn about the scope of possible results.

The model behind this understanding combines statistical analysis with existing theory to posit and evaluate hidden aspects (e.g., perceptions of coverage or fluency) and to measure the relationships among the low-level measurements, the hidden aspects, and the system-level performance measurements. A good model will highlight the most significant links, identify causation when possible, and provide mechanism to both predict the impact of system changes and reverse direction to identify target system changes to achieve desired system-level results.

5.1.2 How

5.1.2.1 Low level performance measures

Most of the research in algorithmic evaluation has focused on low level performance measures such as precision and recall. Our model tries to link low level measures to system level measures, potentially explained by the hidden aspects will tell us what low level measures can best identify system level successes (or failures). Some low level measures might directly relate to system level performances, for example, we might find that a measure of precision directly influences click through rates, but in many cases the relations between low and system level measures might be opaque and can only be understood by unraveling the underlying hidden aspects.

5.1.2.2 Hidden aspects

Hidden aspects are aspects that cannot directly be measured objectively, but that can be measured subjectively from users via surveys or observations. Hidden aspects allow us to understand relations between low level performance measures and system level measures. For example, a particular low-level measure (e.g. novelty) might relate to several hidden aspects in directionally different ways (e.g. improving the perceived diversity but reducing the perceived accuracy of a recommendation or search result). These hidden aspects in turn might either positively or negatively influence system-level performance measures. What hidden aspects to account for and how to measure these is a question for which theoretical understanding of the problem is crucial.

5.1.2.3 System-level performance measures

On the system side, many measures can reflect system performance. We must consider behavioral measures, which can be short-term effects such as click through rates, measures of what items users access or read and when they consume the items, as well as long-term effects reflecting the system's success such as long-term retention or users unsubscribing from a service or reduced or increased usage of the service. Subjective measures such as satisfaction that cover the user experience are also important to measure and predict system-level performance and can be both short-term (are you satisfied with the choice you just made) as well as long-term when measuring overall user satisfaction using infrequent higher-level surveys (e.g. are you happy with our service in the last three months). Sometimes subjective measures can even outperform behavioral measures in predicting, for example user segments of a website [55]. The challenge will be to understand which direct short-term measures best

predict long-term satisfaction and system success. Optimizing the system for any one single short-term measure may in fact harm long-term performance. For example maximizing the number of clicks from a user in a news reading service may actually not reflect they are reading more, it may mean they are trying to find something of interest to read and are failing, so a combination of number of clicks and pause time on the page in combination are better predictions of user satisfaction. Subjective measures such as surveys maybe used to provide training data for machine learning models to capture the complex relationship between system level measures and predicting good user retention and satisfaction.

5.1.2.4 Understanding the relations between the measures via path modeling

Our model explicitly models the hidden aspects as intermediate concepts relating the low-level aspects to the high-level measures. Statistical methods such as path modeling and structural equation modeling allow us to model structural relations between the variables in one single model. This approach in essence just regresses system-level measures on hidden aspects and low-level aspects, and shows whether effects of low level on system-level measures are direct, or indirect and thus mediated by the hidden aspects. As the model fits all relations at once, their relative contributions can be better understood and estimated then one could do by just correlating all measures without a clearly-defined underlying structure. Moreover, the path modeling will allow us to test which underlying structure provides the best explanation and find missing relations. For example, if hidden aspects relate to system level measures but no low level measures directly affect these particular hidden aspects, this will indicate that either better low level measures might be needed or that other external factors might influence our system level measures that are not under our control. Such external factors can be partially controlled for by testing the model in a controlled experiment in which we only manipulate a particular aspect of our algorithms and keep the rest of the system the same, such that we can tease out the one aspect we are interested in. Structural equation modeling moreover allows for hidden aspects to be latent constructs that can be measured through several questionnaire items, rather than by single indicators. This is important when measuring psychological constructs such as perceived diversity or satisfaction, because single items lack 'content validity'. Each user might interpret an item differently and by measuring concepts with several slightly differently phrased items a better measurement of the underlying latent construct can be achieved.

5.1.3 Next Steps

Here we suggest some key next steps that should be undertaken and supported in order to cover the full range from low-level and system-oriented measures to high-level and user-oriented ones:

- Creating a dictionary of higher-level and hidden-aspect measures, including validated and reusable measures that can support comparative research and accumulation of results across studies. We note that certain sub-areas have a longer tradition of higher-level evaluation metrics, and that other sub-areas will need to be engaged to understand the key success measures for their systems.
- Building a library of case studies–examples with constructed models. These cases should be collected in a standard format to promote further analysis and meta-analysis.
- Encourage the study of complex cases–including cases that span more than one technology. To build our understanding of how user perceptions affect performance of complex systems, we need to study a wide range of increasingly complex cases.

- Perform both manual and automated analyses to seek patterns in the case library. By exploring common subgraph patterns, we can develop evidence-based theory to govern system design. The most interesting patterns will likely be ones where the subgraphs depend on specific system attributes (e.g., certain relationships may exist in systems where users have a particular goal in mind, but not in ones where users are simply exploring).
- Ensure the availability of long-standing user cohorts, who can assess over time the systems and whose outcomes can be traced to validate predictions. We will require different user cohorts for different domains/contexts, so that it becomes possible to develop a matrix arrangements of systems across cohorts which can be leveraged for cross-predictability either column-wise, i.e. changing domain/cohort, or row-wise, i.e. changing systems within different rounds of the same cohort. An open question is how to map user cohorts to real users? How good is their external validity?

5.2 Performance Analysis

Reporting of averages and average improvements is often unhelpful, and is uninformative in terms of explaining what system elements contributed to success, what data and queries the method is applicable to, and for which data and queries the method fails. That is, instead of focusing on statistically significant differences in the average from control to treatment, we need to move to understanding the changes in specific tasks and task types, and to understanding the contributions of individual system components.

In this context, researchers also should no longer ignore the problems of multiple testing and sequential testing: When performing multiple significance tests on the same data set, they must adjust the significance level accordingly, using e.g. Bonferroni's method[8] [50]. Even more problematic is the sequential testing case, where the same data is used by other researchers, who have learned from previous results on the same test collection, and then perform significance tests for their new method(s), not considering the large number of tests carried out before. As shown in [20], this usually leads to totally random results. As a consequence, statistically meaningful results cannot be expected from heavily re-used test collections. A similar statement might also hold for multiple qualitative analyses on the same data set. Thus, re-use of a test collection is problematic, which leads to the need for more (and more diverse) collections.

Another important viewing angle is the consideration of measures representing different user standpoints: instead of focusing on universal performance, more emphasis should be put on performance differences wrt. different metrics. E.g., in retrieval, many users will look at the top ranking documents only (e.g. in Web search), while others are aiming at locating all potentially relevant documents (e.g. patent search). Thus, instead of looking at overall performance only, it is more interesting to identify methods that support specific user standpoints.

Classic failure analysis inspects individual tasks where performance is significantly altered, but other data interrogation methods, such as systematic addition of noise, can illustrate the robustness and vulnerabilities of the system that is under investigation.

[8] https://en.wikipedia.org/wiki/Multiple_comparisons_problem

5.3 Documenting and Understanding Assumptions

5.3.1 Role of assumptions

Any method or model is based on certain assumptions, some of which are explicit; usually, there is an even larger number of implicit assumptions. The performance of a method in an application depends mainly on the extent to which these assumptions are true in this setting. Thus, we have to solve three problems:

- Identify the underlying assumptions (make implicit ones explicit).
- Devise methods for determining if or to what extent these assumptions are fulfilled in an application.
- Develop a model that tells us how the violation of an assumption affects performance.

Only when we have answers to these three questions, we are able to make reasonable predictions.

5.3.2 Assumption categories

Assumptions come into play at many points in the design and evaluation of recommender, IR, and NLP systems. At each point, there are at least two broad categories of assumptions: *fundamental* assumptions and *convenience* assumptions. These two categories are transversal to different kinds of assumptions which we can distinguish according to the role they play in data, algorithms/techniques, evaluation protocols and metrics, and their implications on system performance and the validity of research findings. Overall, they determine a taxonomy which we can use to systematically check and make them explicit.

Convenience assumptions are simplifications (or approximations) intended to make problems tractable, reduce their complexity and/or enable evolving some starting point theoretical expression (e.g. a probability) into a computable form (counting things and doing math upon numbers). Examples include the mutual feature independence assumption in Naïve Bayes (of which pairwise word independence in text IR can be seen as a particular case), whereby joint probabilities are decomposed into products of simpler distributions; user, time and context independence as a means to eliminate variables from IR and recommendation problems; or document relevance independence assumption, which enables the definition of simple and easy to compute metrics such as precision. Convenience assumptions may be violated, and yet the algorithm or the metric may still work reasonably. On the other hand, performance differences between collections may be traced back to the violation of certain assumptions.

Convenience assumptions typically represent an opportunity to define new research problems consisting of the elimination of a particular simplifying assumption and dealing with the corresponding complexity. An example is personalized IR, which takes the user variable back into the problem and copes with it; or IR diversity, which removes the document relevance independence assumption; or time-aware or context-aware IR, which do the same with time and context.

By fundamental assumptions we mean hypotheses that algorithms or metrics themselves build upon – they are intrinsic to the underlying model. For instance, content-based recommendation assumes item features can partially explain user choices; IR language model algorithms assume language similarity is related to relevance; most text IR models assume term frequency matters; proximity search algorithms assume word order matters too; metrics like precision assume users want to get relevant documents; average search length (rank of first relevant document) assumes users need just one relevant document or item; recommendation diversity metrics may assume people enjoy variety; novelty metrics assume users wish to be surprised; an experimental protocol may assume each and every user has a non-empty set of

training (or test) observations. When fundamental assumptions fail to be met, the algorithm or the metric may no longer be effective or valid. Content-based recommendation is as good as random if user choices are unrelated to item features; a novelty metric is irrelevant if users are just willing to stick to familiar experiences; lack of data for a single user may result in an undefined evaluation outcome.

Becoming aware of and understanding fundamental assumptions enables a better and more consistent use of the tool (algorithm, metric, protocol) that builds upon them, and may prevent unintentional misuse. It can also help detect spurious confounders (biases that cause the hypothesis to hold for misleading reasons) and experimental flaws that can easily go unnoticed (e.g. a recommendation algorithm's accuracy skyrockets simply because we forgive it refusing to deliver recommendations to certain users; depending on the characteristics of these users, this may result in discriminatory quality of service).

5.3.3 Understanding Violations to Assumptions

A critical aspect in explaining and predicting performance is to understand whether and to what extent the assumptions our methods are based upon have been complied with or violated.

This understanding should happen at both theoretical and experimental level. At theoretical level, among the various assumptions, we should be able to differentiate those that are crucial for a method and whose violations seriously hamper its application from those that are somehow desirable. At experimental level, we should have techniques for assessing each assumption and understand whether and how much it has been violated.

We need to develop commonly agreed *scales* to quantify how much an assumption has been violated. However, given the wide range and diversity in the type of assumptions we have, we should aim at developing assumption checking methods and scales that hold, at best, for families of related assumptions rather than hoping for a single general solution where one-fits-all.

Then, we need to research on the relationship between the severity of departures from assumptions, quantified in the above mentioned scales, and the observed and predicted performances. The final goal is to understand how much resilient are our methods to such violations and how much this impacts on explanation, first, and prediction, after.

Violations of algorithm or technique assumptions are perhaps the easiest to assess: run the algorithm on a data set that violates its assumption(s) and measure its performance and behavior. Violations of evaluation and data assumptions are more challenging, as they undermine the tools by which we measure the behavior of the system in the first place. To assess the impact of these assumptions, we need techniques that allow us to peek behind the curtain and understand the behavior of these components of the experimental process under a range of possible truths, in order to relate their output to our confidence about the relationship of the data and evaluation to the underlying truth and intended task.

An area we can take inspiration from is statistics and the notion of *robustness* in statistical testing, meant as "insensitivity to small deviations from the assumptions" [64]. Robustness is developed both a theoretical level, e.g. by studying it under a null and an alternative hypothesis [68], and defining indicators such as, for example, the breakdown point, i.e. the proportion of incorrect observations an estimator can handle before giving an incorrect result.

Furthermore, simulation and resampling are particularly promising tools for quantifying the importance of assumptions to components of the information processing and evaluation pipeline. Measuring results on different data sets is useful, but only provides a few data points regarding the behavior of a method or evaluation technique, and does not change the

relevant variables in a controlled fashion; further, the data set's relationship to underlying ground truth cannot be controlled and may not be known. Simulation and resampling allows a range of possibilities – some within assumptions, some outside – to be tested, and the relationship of data to truth to be controlled, allowing us to precisely characterize the system response to targeted violations of its assumptions. These experiments can take multiple forms, including wholly-synthetic data, resampling of traditional data sets, and sampling of specialized data sets such as ratings collected on complete or uniformly sampled sets of items. As one example, [106] employed simulation to study the robustness of information retrieval evaluations to violations of statistical assumptions about the underlying data sets and their topic distributions.

5.3.4 Increasing awareness in our communities

We note that across our communities there is a large variance on how assumptions are managed and on the perception itself of their importance. A general recommendation is pushing in any possible way our communities to a greater awareness of the need for making assumptions explicit and clear. Inserting in all scientific works a clear statement permitting a precise identification and a deeper understanding of the essential assumptions made and their scope of validity should become a universal practice. To this regard we recall the effort currently conducted in the IR community toward reproducibility of results [40, 41]: after a consciousness campaign last several years, we now have a reproducibility track in one of the main IR conferences (ECIR), and reproducibility tasks have been just launched in the major evaluation campaigns[9].

The awareness on such an important aspect impacting the validity and reproducibility of results can be disseminated and increased in several ways. A first recommendation is adding an explicit reference to the clarity and completeness of assumptions made in the call for paper and the paper review forms of all conferences. This can have the double effect of educating the reviewers to reserve a particular attention to assumption clarity and, on the other hand, to increase author's awareness on them. Papers claiming results involving assumptions that are not explicitly voiced or understood should not be deemed as solid since no strong conclusion can be drawn from them. As a second step, after a systematization of assumptions and a greater understanding have been reached, the emerging best practices can give origin to commonly accepted requirements to be integrated in the call for papers of specific tracks.

It is significantly harder to test the importance of assumptions in user-facing aspects of the system, such as the presentation of results or the task model, as it is prohibitively expensive to simulate arbitrarily many versions of a system and put them before users. System utility can be remarkably robust to violations of core assumptions – for example, e-commerce vendors obtain great value from collaborative filtering techniques that assume items are functionally interchangeable even when they clearly are not – but rigorous empirical data on this robustness is difficult to obtain. However, measuring hidden factors (cf. Figure 2) might help explain why particular versions of a system perform better, directly testing underlying assumptions.

[9] http://www.centre-eval.org/

5.4 Application features

One common feature of Natural Language Processing, Information Retrieval and Recommender Systems is the wide space of data and task characteristics that have to be accounted for when designing a system. Adapting existing systems to a new domain, a new data set, or a new task, and then predicting their performance in this new setting is particularly challenging in our research fields because there is always some degree of mismatch between testing and development conditions (either in laboratory or real-world settings) used to create the existing systems and new application area.

As a result, measuring only effect sizes and statistical significance is of little help for predicting out-of-the-lab performance. Even moving between two test collections which apparently share the same features often results in different experimental outcomes. In order to have predictive power, evaluation methodologies need a much higher emphasis on explanatory analysis: why, where and how systems fail is more relevant than effect sizes on average measures.

In this section we begin by reviewing a few measurable characteristics that make prediction possible but challenging in our research fields, and we then move to advocating explanatory analysis.

5.4.1 Task & Data features

How will an existing method, algorithm or system perform under conditions different from the ones in which it was tested? There are some easily identifiable features related to the data or the task that, if changed, may affect predictability.

With respect to the task, some relevant characteristics are the **language** involved in the task. Will the task be performed using monolingual or multilingual data. Will the output be in a different language from the input (cross-linguistic)? Or is the task language independent? Are there the necessary language resources for the task? Does the task involve some dialect for which these resources have to be adapted? Is the data based on speech, on written text?

Another characteristic of the task is it **dynamicity**: are we dealing with a static collection, or a stream of data? Is the task a one-off, ad-hoc task, or a long standing task, such as filtering a news stream with a static query? Is the task offline, or online, performed with an active user? Does the task change over time as the user performs it?

Task **context** also plays an important role in many situations: Current Web search engines consider already user history, location, time and end device when computing the search result. The same might be true for other types of tasks.

We can characterize the data as **curated**, for example scientific papers, or edited news stories, or as naturalistic, for example, stemming from social media, or transcribed speech. In the latter case, one can sometimes measure the expected error rate, such as the frequency of spelling errors, or transcription errors. Many language processing tools were developed for curated language, without such errors.

Another dimension of data is its **connectedness** or **structure**. Can each data item be considered as a separate item, or are there links between items? For example, web pages link to other web pages. A collection of movies can contain a series of implicitly linked sequels. Users in a social network have both explicit and implicit connections to other users. Each data item can have internal structure (metadata such as timestamps, hand-assigned classification codes, numerical data; or internal structure, such as abstract, body, supplemental material).

With respect to (textual) data, some measurable features are: readability and comprehensibility; domain; users' expertise; how source and target data correlates; verifiability

of answers; dependence on assumptions to construct ground truth; richness of features; external validations; existence of corner cases and stress factors; parameterizations that impact performance; quality of domain resources (ontologies, dictionaries, taggers, etc.)

These characteristics, however, are not likely to be sufficiently predictive: even when they are the same in the new application as in laboratory conditions, often components of the systems perform differently. One of the main shortcomings of our experimental methodology is the lack of adequate explanatory methods.

5.4.2 Bias and Scaling

Test collections are often not a representative sample of a larger population. Instead, they have been compiled under certain restrictions (e.g. in IR test collections, rather specific or too general topics are not considered). We need to understand the limitations and bias of our sampling methodology across topics, documents, and systems. Can we determine when differences are due to bias, or when we are sampling from separate distributions?

Another problem to be investigated is the effect of scale: methods doing well on small test collections might not work on collections orders of magnitude larger, and vice versa.

5.5 Modeling Performance

Trying to explain and model the performance of systems over different datasets and tasks is a preliminary yet indispensable step towards envisioning how to predict the performance of such systems. However, this is often difficult to do due the lack of appropriate analysis techniques and the need for careful experimental designs and protocols, which may be complex and demanding to carry out.

There is therefore a need for further research providing us with the methods for analyzing and decomposing the performance into those of the affecting factors, such as system components, datasets, tasks, and more. These explanatory models will then constitute the basis for developing predictive models.

Performance prediction can take different forms. We commonly wish to make an *ordinal* prediction, of which of two systems will be superior for a kind of task over a class of collections. For a single system, we might aim at an *interval* prediction, giving us a confidence interval for a certain metric; the most simple case would be a prediction for another sample from the same population. While these two approaches target at average performance, we may alternatively wish to estimate risk or *uncertainty*, that is, predict a likelihood of failure.

5.5.1 Performance factor analysis in IR

In the case of IR, over the years, there have been examples of attempts to decompose performance into constituting factors, based on the use of General Linear Mixed Models (GLMM) and ANalysis Of VAriance (ANOVA).

[11, 102] have shown how to break down the performance of an IR system into a Topic and a System effect, finding that the former has a much bigger impact than the latter.

By using a specific experimental design, [45, 46] also broke down the System effect into those of its components – namely stop lists, stemmers, and IR models. They further demonstrated that we are not actually evaluating these components alone, even when we change only one of them and keep all the rest fixed. Rather, we are evaluating whole pipelines

where these components are inserted and with which they may have positive (or negative) interaction, boosting (or depressing) their estimated impact.

The difficulty in estimating the Topic*System interaction effects is the lack of replicates for each (topic, system) pair in a standard experimental setting. Therefore, [92] used simulation based on distributions of relevant and not relevant documents to demonstrate the importance of the Topic*System interaction effect. Very recently, [111] exploited random partitions of the document corpus to obtain more replicates of each (topic, system) pair, obtaining an estimation of the Topic*System interaction effect which allowed for improved precision in determining the System effect.

Finally, [44] conducted preliminary studies on the effect of Sub-Corpora and the System*Sub-Corpus, showing their impact and how they can be exploited to improve the estimation of the System effect.

All these GLMMs are not connected yet, meaning that they tackle the problem separately from different viewpoints but there is not yet a single model integrating all these facets. So a first required step toward performance prediction is to unify all these explanatory models into a single one. Then, the next step is to turn these models into predictive ones, e.g. by using some of the features discussed in Section 5.4 to learn how to predict the factors described in these models.

5.5.2 Controlled experimentation in NLP

NLP components are often combined into more complex NLP systems (e.g. part of speech detection, entity recognition etc. being used as part of a summarization task). The need to understand both the individual performance of components of pipeline NLP systems, as well as interactions between them, has resulted in evaluation methods involving controlled experimentation with systems in terms of these component parts.

Systematic component evaluation. Evaluating adaptivity by "decomposing" and evaluating it in a "piece-wise" manner can also be adapted from evaluations of interactive adaptive systems [87]. This can be done in a number of ways such as component substitution, ablation, and oracle input data.

Component Substitution: One strategy for doing this is to perform experiments that involve substitution of alternative components for a single component of the pipeline, to measure the impact on that component on the overall system performance.

Ablation: A related approach is the use of "ablation" (also called *lesion*) studies, in which sets of features, or combinations of feature sets, are systematically removed, in order to determine the most effective representation for a given task [24, 83]. This is commonly used in evaluation of machine learning-based methods which make use of substantial feature engineering.

Oracle input data: Pipelining components introduces a ceiling for each component that limits the performance of the overall system. To focus evaluation on a specific component in isolation, the performance of a target component can be measured by assuming that perfect input data is derived from earlier stages of processing. In user studies, this is sometimes also called a "wizard-of-oz" approach, where the component being evaluated is facing an end-user. Most commonly in these systems, the oracle is a human operating as system.

As an example, the performance of a relation extraction system that depends on a named entity recognition system as a precursor step will be limited by the performance of that earlier step. In the BioNLP-Shared Task relation extraction evaluations, gold standard entity

annotations are provided as input for the relation extraction systems [90] to control for this problem. While this represents idealized conditions for the overall system, it allows isolation of the relation extraction algorithms from performance effects resulting directly from the entity recognition step.

Test suites: Test suites have long been used by the NLP community to structure the evaluation of the functionality of specific tools [86], and also to structure efficient development of NLP systems [47]. In this approach, specific test cases are created based on controlled variation of pre-determined phenomena.

Recently, this approach has seen some revival, on the basis that articulation of specific cases identified in linguistic data can be used to guide finer-grained evaluation of systems that process that data, and that evaluation of purely natural data is dominated by high-frequency, possibly "simple" cases [23, 29, 57].

It can be argued that producing meaningful test cases is itself a challenging, resource-intensive activity [71], and also that the corner cases are not possible to define in advance. Nevertheless, this approach may provide a useful strategy to consider for deeper characterization of system performance and performance predictability, by characterizing the types of data that are expected to be seen by a system, and their varying distributions in natural data sets.

6 Conclusion

Performance prediction in the areas of IR, NLP and RecSys is a research problem that has been ignored for many years. In this manifesto, we have presented a framework for starting research in this area. Some problems might require substantial resources before they can be addressed. For instance, the analysis of performance-determining application features requires a large number of testbeds. Most of the problems, however, require primarily a more analytic approach. Instead of focusing only on performance improvement/system tuning, researchers should aim at improving our understanding of why, how and when the investigated methods work.

This manifesto should not only be regarded as a useful account of an important research challenge. We hope that it will also produce valuable fall-outs, such as bringing these issues in the research agenda of the involved communities (as it recently happened in the case of IR [3]), helping funding agencies in envisioning appropriate funding instruments for addressing these challenges, and spurring researchers on to overcome today's limitations.

7 Participants

- Pablo Castells
Autonomous University of Madrid, Spain
- Elizabeth M. Daly
IBM Research – Dublin, Ireland
- Thierry Declerck
DFKI GmbH – Saarbrücken, Germany
- Michael D. Ekstrand
Boise State University, USA
- Nicola Ferro
University of Padova, Italy
- Norbert Fuhr
Universität Duisburg-Essen, Germany
- Werner Geyer
IBM TJ Watson Research Center – Cambridge, USA
- Julio Gonzalo
UNED – Madrid, Spain
- Gregory Grefenstette
IHMC – Paris, France
- Joseph A. Konstan
University of Minnesota – Minneapolis, USA
- Tsvi Kuflik
Haifa University, Israel
- Krister Lindén
University of Helsinki, Finland
- Bernardo Magnini
Bruno Kessler Foundation – Trento, Italy
- Jian-Yun Nie
University of Montreal, Canada
- Raffaele Perego
CNR – Pisa, Italy
- Bracha Shapira
Ben Gurion University – Beer Sheva, Israel
- Ian Soboroff
NIST – Gaithersburg, USA
- Nava Tintarev
TU Delft, The Netherlands
- Karin Verspoor
The University of Melbourne, Australia
- Martijn Willemsen
TU Eindhoven, The Netherlands
- Justin Zobel
The University of Melbourne, Australia

Acknowledgements

We thank Schloss Dagstuhl for hosting us.

References

1 Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. De-biasing user preference ratings in recommender systems. In *RecSys 2014 Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2014)*, pages 2–9, Foster City, CA, USA, 2014. URL: http://ceur-ws.org/Vol-1253/paper1.pdf.

2 Gediminas Adomavicius, Jesse C Bockstedt, Shawn P Curley, and Jingjing Zhang. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research*, 24(4):956–975, 2013. doi:10.1287/isre.2013.0497.

3 James Allan, Jaime Arguello, Leif Azzopardi, Peter Bailey, Tim Baldwin, Krisztian Balog, Hannah Bast, Nick Belkin, Klaus Berberich, Bodo Billerbeck, Jamie Callan, Rob Capra, Mark Carman, Ben Carterette, Charles L. A. Clarke, Kevyn Collins-Thompson, Nick Craswell, W. Bruce Croft, J. Shane Culpepper, Jeff Dalton, Gianluca Demartini, Fernado Diaz, Laura Dietz, Susan Dumais, Carsten Eickhoff, Nicola Ferro, Norbert Fuhr, Shlomo Geva, Claudia Hauff, David Hawking, Hideo Joho, Gareth Jones, Jaap Kamps, Noriko Kando, Diane Kelly, Jaewon Kim, Julia Kiseleva, Yiqun Liu, Xiaolu Lu, Stefano Mizzaro, Alistair Moffat, Jian-Yun Nie, Alexandra Olteanu, Iadh Ounis, Filip Radlinski, Maarten de Rijke, Mark Sanderson, Falk Scholer, Laurianne Sitbon, Mark Smucker, Ian Soboroff, Damiano Spina, Torsten Suel, James Thom, Paul Thomas, Andrew Trotman, Ellen Voorhees, Arjen P. de Vries, Emine Yilmaz, and Guido Zuccon. Research Frontiers in Information Retrieval – Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52(1):34–90, June 2018. doi:10.1145/3274784.3274788.

4 Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. I like it... I like it not: Evaluating user ratings noise in recommender systems. In *17th International Conference on User Modeling, Adaptation, and Personalization, , UMAP 2009*, volume 5535 of *Lecture Notes in Computer Science*, pages 247–258. Springer, 2009. doi:10.1007/978-3-642-02247-0_24.

5 Xavier Amatriain, Josep M. Pujol, Nava Tintarev, and Nuria Oliver. Rate it again: increasing recommendation accuracy by user re-rating. In *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009*, pages 173–180. ACM, 2009. doi:10.1145/1639714.1639744.

6 Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486, 2009. doi:10.1007/s10791-008-9066-8.

7 Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. A general evaluation measure for document organization tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 643–652. ACM, 2013. doi:10.1145/2484028.2484081.

8 Enrique Amigó, Damiano Spina, and Jorge Carrillo de Albornoz. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *Proc. 41th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*, pages 625–634. ACM Press, New York, USA, 2018. doi:10.1145/3209978.3210024.

9 Marco Angelini, Nicola Ferro, Birger Larsen, Henning Müller, Giuseppe Santucci, Gianmaria Silvello, and Theodora Tsikrika. Measuring and Analyzing the Scholarly Impact of Experimental Evaluation Initiatives. In Maristella Agosti, Tiziana Catarci, and Floriana Esposito, editors, *Proc. 10th Italian Research Conference on Digital Libraries (IRCDL 2014)*, volume 38, pages 133–137. Procedia Computer Science, Vol. 38, 2014. doi:10.1016/j.procs.2014.10.022.

10 Apache Software Foundation. Apache Mahout 0.12.2, June 2016. URL: https://mahout.apache.org/.

11 David Banks, Paul Over, and Nien-Fan Zhang. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval*, 1(1-2):7–34, May 1999. doi:10.1023/A:1009984519381.

12 Alejandro Bellogín, Pablo Castells, and Iván Cantador. Statistical Biases in Information Retrieval Metrics for Recommender Systems. *Information Retrieval*, 20(6):606–634, December 2017. URL: http://ir.ii.uam.es/pubs/irj2017.pdf.

13 James Bennett and Stan Lanning. The Netflix Prize. In *Proc. of KDD Work on Large-Scale Rec. Sys.*, 2007. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6998.

14 Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011. URL: http://ismir2011.ismir.net/papers/OS6-1.pdf.

15 Chris Buckley and Donna Harman. Reliable information access final workshop report. *ARDA Northeast Regional Research Center Technical Report*, 3, 2004.

16 Chris Buckley and Ellen M. Voorhees. Evaluating Evaluation Measure Stability. In Emmanuel J. Yannakoudakis, Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, editors, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 33–40. ACM Press, New York, USA, 2000. doi:10.1145/345508.345543.

17 Chris Buckley and Ellen M. Voorhees. Retrieval Evaluation with Incomplete Information. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 25–32. ACM Press, New York, USA, 2004. doi:10.1145/1008992.1009000.

18 Jamie Callan and Alistair Moffat. Panel on Use of Proprietary Data. *SIGIR Forum*, 46(2):10–18, December 2012. doi:10.1145/2422256.2422258.

19 David Carmel and Elad Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool Publishers, USA, 2010. doi:10.2200/S00235ED1V01Y201004ICR015.

20 Ben Carterette. The Best Published Result is Random: Sequential Testing and Its Effect on Reported Effectiveness. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 747–750, New York, NY, USA, 2015. ACM. doi:10.1145/2766462.2767812.

21 Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected Reciprocal Rank for Graded Relevance. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy J. Lin, editors, *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, pages 621–630. ACM Press, New York, USA, 2009. doi:10.1145/1645953.1646033.

22 Tianqi Chen, Weinan Zhang, Qiuxia Lu, Kailong Chen, Zhao Zheng, and Yong Yu. SVDFeature: A toolkit for feature-based collaborative filtering. *Journal of Machine Learning Research: JMLR*, 13(1):3619–3622, December 2012. URL: http://dl.acm.org/citation.cfm?id=2503308.2503357.

23 K. Bretonnel Cohen, Christophe Roeder, William A. Baumgartner Jr., Lawrence E. Hunter, and Karin Verspoor. Test Suite Design for Biomedical Ontology Concept Recognition Systems. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2010/summaries/31.html.

24 Paul R. Cohen and Adele E. Howe. How evaluation guides AI research: The message still counts more than the medium. *AI magazine*, 9(4):35, 1988. URL: http://www.aaai.org/ojs/index.php/aimagazine/article/view/952.

25 Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. *CoRR*, abs/1706.09031, 2017. arXiv:1706.09031.

26 Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, page 39–46, New York, NY, USA, 2010. ACM. doi:10.1145/1864708.1864721.

27 Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM, 2002. doi:10.1145/564376.564429.

28 Ronan Cummins. Document Score Distribution Models for Query Performance Inference and Prediction. *ACM Transactions on Information System (TOIS)*, 32(1):2:1–2:28, January 2014. doi:10.1145/2559170.

29 Dina Demner-Fushman. Adapting Naturally Occurring Test Suites for Evaluation of Clinical Question Answering. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 21–22, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/W/W08/W08-0505.

30 Mukund Deshpande and George Karypis. Item-based top-N Recommendation Algorithms. *ACM Transactions on Information and System Security*, 22(1):143–177, January 2004. doi:10.1145/963770.963776.

31 Long Duong. *Natural language processing for resource-poor languages*. PhD thesis, The University of Melbourne, 2017. URL: http://hdl.handle.net/11343/192938.

32 Michael D. Ekstrand. *Towards Recommender Engineering: Tools and Experiments in Recommender Differences*. PhD thesis, University of Minnesota, Minneapolis, MN, July 2014. URL: http://hdl.handle.net/11299/165307.

33 Michael D. Ekstrand and Vaibhav Mahant. Sturgeon and the Cool Kids: Problems with Top-N Recommender Evaluation. In *Proceedings of the 30th Florida Artificial Intelligence Research Society Conference*. AAAI Press, May 2017. URL: https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/viewPaper/15534.

34 Hui Fang, Tao Tao, and Chengxiang Zhai. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.*, 29(2):7:1–7:42, April 2011. doi:10.1145/1961209.1961210.

35 Marco Ferrante, Nicola Ferro, and Maria Maistro. Injecting User Models and Time into Precision via Markov Chains. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, pages 597–606. ACM Press, New York, USA, 2014. doi:10.1145/2600428.2609637.

36 Marco Ferrante, Nicola Ferro, and Maria Maistro. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In James Allan, W. Bruce Croft, Arjen P. de Vries, and Chengxiang Zhai, editors, *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, pages 21–30. ACM Press, New York, USA, 2015. doi:10.1145/2808194.2809452.

37 Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. An Interval-Like Scale Property for IR Evaluation Measures. In Nicola Ferro and Ian Soboroff, editors, *Proc. 8th International*

Workshop on Evaluating Information Access (EVIA 2017), pages 10–15. CEUR Workshop Proceedings (CEUR-WS.org), 2017. URL: http://ceur-ws.org/Vol-2008/paper_11.pdf.

38 Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. Are IR Evaluation Measures on an Interval Scale? In Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz, editors, *Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017)*, pages 67–74. ACM Press, New York, USA, 2017. doi:10.1145/3121050.3121058.

39 Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. A General Theory of IR Evaluation Measures. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2018. doi:10.1109/TKDE.2018.2840708.

40 Nicola Ferro. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)*, 8(2):8:1–8:4, February 2017. doi:10.1145/3020206.

41 Nicola Ferro, Norbert Fuhr, Kalervo Järvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum*, 50(1):68–82, June 2016. doi:10.1145/2964797.2964808.

42 Nicola Ferro and Diane Kelly. SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum*, 52(1):4–10, June 2018. doi:10.1145/3274784.3274786.

43 Nicola Ferro, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian-Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Nineth International Conference of the CLEF Association (CLEF 2018)*, volume 11018 of *Lecture Notes in Computer Science*, pages 239–246. Springer, 2018. doi:10.1007/978-3-319-98932-7_23.

44 Nicola Ferro and Mark Sanderson. Sub-corpora Impact on System Effectiveness. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 901–904. ACM Press, New York, USA, 2017. doi:10.1145/3077136.3080674.

45 Nicola Ferro and Gianmaria Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*, pages 25–34. ACM Press, New York, USA, 2016. doi:10.1145/2911451.2911530.

46 Nicola Ferro and Gianmaria Silvello. Toward an Anatomy of IR System Component Performances. *Journal of the American Society for Information Science and Technology (JASIST)*, 69(2):187–200, February 2018. doi:10.1002/asi.23910.

47 Dan Flickinger. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28, 2000. doi:10.1017/S1351324900002370.

48 Juliana Freire, Norbert Fuhr, and Andreas Rauber. Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science. *Dagstuhl Reports*, 6(1):108–159, 2016. doi:10.4230/DagRep.6.1.108.

49 Norbert Fuhr. Salton Award Lecture Information Retrieval As Engineering Science. *SIGIR Forum*, 46(2):19–28, December 2012. doi:10.1145/2422256.2422259.

50 Norbert Fuhr. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum*, 51(3):32–41, December 2017. doi:10.1145/3190580.3190586.

51 Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. MyMediaLite: A free recommender system library. In *Proceedings of the Fifth ACM Conference*

on Recommender Systems, RecSys '11, page 305–308, New York, NY, USA, 2011. ACM. doi:10.1145/2043932.2043989.

52 Alfonso Emilio Gerevini, Alberto Lavelli, Alessandro Maffi, Roberto Maroldi, Anne-Lyse Minard, Ivan Serina, and Guido Squassina. Automatic Classification of Radiological Reports for Clinical Care. In *Artificial Intelligence in Medicine - 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings*, pages 149–159, 2017. doi:10.1007/978-3-319-59758-4_16.

53 Werner Geyer, Casey Dugan, David R. Millen, Michael Muller, and Jill Freyne. Recommending Topics for Self-descriptions in Online User Profiles. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 59–66, New York, NY, USA, 2008. ACM. doi:10.1145/1454008.1454019.

54 Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information retrieval*, 4(2):133–151, July 2001. doi:10.1023/A:1011419012209.

55 Mark P. Graus, Martijn C. Willemsen, and Kevin Swelsen. Understanding Real-Life Website Adaptations by Investigating the Relations Between User Behavior and User Experience. In Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless, editors, *User Modeling, Adaptation and Personalization*, number 9146 in Lecture Notes in Computer Science, pages 350–356. Springer International Publishing, June 2015. doi:10.1007/978-3-319-20267-9_30.

56 Gregory Grefenstette. Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research. In *Ninth Workshop on Building and Using Comparable Corpora*, page 26, Portoro, Slovenia, May 2016. ELDA. URL: https://comparable.limsi.fr/bucc2016/pdf/BUCC05.pdf.

57 Tudor Groza and Karin Verspoor. Automated Generation of Test Suites for Error Analysis of Concept Recognition Systems. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 23–31, Melbourne, Australia, 2014. URL: https://aclanthology.info/papers/U14-1004/u14-1004.

58 Guibing Guo, Jie Zhang, Zhu Sun, and Neil Yorke-Smith. LibRec: A java library for recommender systems. In *UMAP Workshops*, volume 1388 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015. URL: http://ceur-ws.org/Vol-1388/demo_paper1.pdf.

59 Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy J. Lin, Simon Mercer, and Martin Potthast. Evaluation-as-a-Service: Overview and Outlook. *CoRR*, abs/1512.07454, December 2015. arXiv:1512.07454.

60 F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, December 2015. doi:10.1145/2827872.

61 Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. A Survey of Pre-Retrieval Query Performance Predictors. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *Proc. 17th International Conference on Information and Knowledge Management (CIKM 2008)*, pages 1419–1420. ACM Press, New York, USA, 2008. doi:10.1145/1458082.1458311.

62 Ruining He and Julian McAuley. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 507–517. International World Wide Web Conferences Steering Committee, April 2016. doi:10.1145/2872427.2883037.

63 Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004. doi:10.1145/963770.963772.

64 Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, USA, 2nd edition, 2009. URL: https://www.wiley.com/en-us/Robust+Statistics%2C+2nd+Edition-p-9781118210338.

65 Juhani Huovelin, Oskar Gross, Otto Solin, Krister Linden, Sami Petri Tapio Maisala, Tero Oittinen, Hannu Toivonen, Jyrki Niemi, and Miikka Silfverberg. Software newsroom–an approach to automation of news search and editing. *Journal of Print Media Technology research*, 2(3):141–156, 2013. URL: http://hdl.handle.net/10138/42754.

66 Rosie Jones and Fernando Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems (TOIS)*, 25(3):14, 2007. doi:10.1145/1247715.1247720.

67 Paul B. Kantor and Ellen M. Voorhees. Report on the TREC-5 Confusion Track. In *Proceedings of The Fifth Text REtrieval Conference, TREC 1996*, volume Special Publication 500-238. National Institute of Standards and Technology (NIST), 1996. URL: http://trec.nist.gov/pubs/trec5/papers/confusion_track.ps.gz.

68 Takeaki Kariya and Bimal K. Sinha. *Robustness of Statistical Tests*. Academic Press, USA, 1989. doi:10.1016/C2013-0-10934-8.

69 George Karypis. SUGGEST recommendation engine, November 2000. URL: http://glaros.dtc.umn.edu/gkhome/suggest/overview.

70 Benjamin Kille, Andreas Lommatzsch, Gebrekirstos G. Gebremeskel, Frank Hopfgartner, Martha Larson, Jonas Seiler, Davide Malagoli, András Serény, Torben Brodt, and Arjen P. de Vries. Overview of NewsREEL'16: Multi-dimensional Evaluation of Real-Time Stream-Recommendation Algorithms. In Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016)*, volume 9822 of *Lecture Notes in Computer Science*, pages 311–331. Springer, 2016. doi:10.1007/978-3-319-44564-9_27.

71 Margaret King. Evaluating Natural Language Processing Systems. *Commun. ACM*, 39(1):73–79, January 1996. doi:10.1145/234173.234208.

72 Bart Knijnenburg, Martijn Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, October 2012. doi:10.1007/s11257-011-9118-4.

73 Bart P. Knijnenburg and Martijn C. Willemsen. *Evaluating Recommender Systems with User Experiments*, pages 309–352. Springer US, Boston, MA, 2015. doi:10.1007/978-1-4899-7637-6_9.

74 Shyong K. Lam and John Riedl. Shilling recommender systems for fun and profit. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *Proceedings of the 13th international conference on World Wide Web*, pages 393–402. ACM, 2004. doi:10.1145/988672.988726.

75 Jimmy J. Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig MacDonald, and Sebastiano Vigna. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*, volume 9626 of *Lecture Notes in Computer Science*, pages 408–420. Springer, 2016. doi:10.1007/978-3-319-30671-1_30.

76 Andreas Lommatzsch, Benjamin Kille, Frank Hopfgartner, Martha Larson, Torben Brodt, Jonas Seiler, and Özlem Özgöbek. CLEF 2017 NewsREEL Overview: A Stream-Based

Recommender Task for Evaluation and Education. In Gareth J. F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eighth International Conference of the CLEF Association (CLEF 2017)*, volume 10456 of *Lecture Notes in Computer Science*, pages 239–254. Springer, 2017. doi:10.1007/978-3-319-65813-1_23.

77 Travis Martin, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web*, pages 683–694. ACM, 2016. doi:10.1145/2872427.2883001.

78 Paul McJones. Eachmovie collaborative filtering data set. *DEC Systems Research Center*, 249:57, 1997.

79 Paul McNamee and James Mayfield. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 159–166. ACM, 2002. doi:10.1145/564376.564406.

80 Sean M. McNee, John Riedl, and Joseph A. Konstan. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1097–1101, New York, NY, USA, 2006. ACM. doi:10.1145/1125451.1125659.

81 Alistair Moffat and Justin Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2:1–2:27, 2008. doi:10.1145/1416950.1416952.

82 Joshua L. Moore, Shuo Chen, Douglas Turnbull, and Thorsten Joachims. Taste Over Time: The Temporal Dynamics of User Preferences. In Alceu de Souza Britto Jr., Fabien Gouyon, and Simon Dixon, editors, *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013*, pages 401–406, 2013. URL: http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/220_Paper.pdf.

83 Allen Newell. A tutorial on speech understanding systems. *Speech recognition*, pages 3–54, 1975.

84 Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May 2016. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/348.html.

85 Chikashi Nobata, Nigel Collier, and Jun'ichi Tsujii. Comparison between tagged corpora for the named entity task. In *Proceedings of the workshop on Comparing corpora*, pages 20–27. Association for Computational Linguistics, 2000. doi:10.3115/1117729.1117733.

86 Stephan Oepen, Klaus Netter, and Judith Klein. TSNLP - test suites for natural language processing. In John Nerbonne, editor, *Linguistic Databases*, chapter 2, pages 13–36. CSLI Publications, 1998.

87 Alexandros Paramythis, Stephan Weibelzahl, and Judith Masthoff. Layered evaluation of interactive adaptive systems: framework and formative methods. *User Modeling and User-Adapted Interaction*, 20(5):383–453, 2010. doi:10.1007/s11257-010-9082-4.

88 Slav Petrov, Dipanjan Das, and Ryan McDonald. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*

(LREC-2012). European Language Resources Association (ELRA), 2012. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.

89 Pearl Pu, Li Chen, and Rong Hu. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 157–164, New York, NY, USA, 2011. ACM. doi:10.1145/2043932.2043962.

90 Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. *BMC Bioinformatics*, 13(11):S2, June 2012. doi:10.1186/1471-2105-13-S11-S2.

91 Fiana Raiber and Oren Kurland. Query-performance prediction: setting the expectations straight. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 13–22. ACM, 2014. doi:10.1145/2600428.2609581.

92 Stephen E. Robertson and Evangelos Kanoulas. On Per-topic Variance in IR Evaluation. In William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 891–900. ACM Press, New York, USA, 2012. doi:10.1145/2348283.2348402.

93 Lior Rokach. Decomposition methodology for classification tasks: a meta decomposer framework. *Pattern Analysis and Applications*, 9(2):257–271, October 2006. doi:10.1007/s10044-006-0041-y.

94 Brent R. Rowe, Dallas W. Wood, Albert N. Link, and Diglio A. Simoni. *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*. RTI Project Number 0211875, RTI International, USA, July 2010. URL: http://trec.nist.gov/pubs/2010.economic.impact.pdf.

95 Alan Said and Alejandro Bellogin. Comparative Recommender System Evaluation: Benchmarking Recommendation Frameworks. In *Proceedings of the Eighth ACM Conference on Recommender Systems (RecSys '14)*, RecSys '14, page 129–136, New York, NY, USA, October 2014. ACM Press. doi:10.1145/2645710.2645746.

96 Tetsuya Sakai. Evaluating Evaluation Metrics based on the Bootstrap. In Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin, editors, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 525–532. ACM Press, New York, USA, 2006. doi:10.1145/1148170.1148261.

97 Tetsuya Sakai. Topic set size design. *Information Retrieval*, 19(3):256–283, June 2016. doi:10.1007/s10791-015-9273-z.

98 Mark Sanderson and Justin Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In Ricardo A. Baeza-Yates, Nivio Ziviani, Gary Marchionini, Alistair Moffat, and John Tait, editors, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 162–169. ACM Press, New York, USA, 2005. doi:10.1145/1076034.1076064.

99 Saúl Vargas Sandoval. *Novelty and diversity evaluation and enhancement in recommender systems*. PhD thesis, Universidad Autonama Demadrid, Madrid, Spain, 2015. URL: http://saulvargas.es/phd-thesis.pdf.

100 Fabrizio Sebastiani. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. In James Allan, W. Bruce Croft, Arjen P. de Vries, and Chengxiang Zhai, editors, *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, pages 11–20. ACM Press, New York, USA, 2015. doi:10.1145/2808194.2809449.

101 Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010. doi:10.1126/science.1177170.

102 Jean Tague-Sutcliffe and James Blustein. A Statistical Analysis of the TREC-3 Data. In Donna K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 1994.

103 Clare Thornley, Andrea C. Johnson, Alan F. Smeaton, and Hyowon Lee. The Scholarly Impact of TRECVid (2003–2009). *Journal of the American Society for Information Science and Technology (JASIST)*, 62(4):613–627, April 2011. doi:10.1002/asi.21494.

104 Theodora Tsikrika, Alba Garcia Seco de Herrera, and Henning Müller. Assessing the Scholarly Impact of ImageCLEF. In Pamela Forner, Julio Gonzalo, Jaana Kekäläinen, Mounia Lalmas, and Maarten de Rijke, editors, *Multilingual and Multimodal Information Access Evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011)*, volume 6941 of *Lecture Notes in Computer Science*, pages 95–106. Springer, 2011. doi:10.1007/978-3-642-23708-9_12.

105 Theodora Tsikrika, Birger Larsen, Henning Müller, Stefan Endrullis, and Erhard Rahm. The Scholarly Impact of CLEF (2000–2009). In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF 2013)*, volume 8138 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2013. doi:10.1007/978-3-642-40802-1_1.

106 Julián Urbano. Test collection reliability: a study of bias and robustness to statistical assumptions via stochastic simulation. *Information Retrieval Journal*, 19(3):313–350, December 2015. doi:10.1007/s10791-015-9274-y.

107 Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L Johnson, Christophe Roeder, Jinho D Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC bioinformatics*, 13(1):1, 2012.

108 Jesse Vig, Shilad Sen, and John Riedl. Computing the Tag Genome. Technical report, GroupLens Research, University of Minnesota, 2010. URL: http://www.grouplens.org/node/447.

109 Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, September 2000. doi:10.1016/S0306-4573(00)00010-8.

110 Ellen M. Voorhees, Shahzad Rajput, and Ian Soboroff. Promoting Repeatability Through Open Runs. In Emine Yilmaz and Charles L. A. Clarke, editors, *Proc. 7th International Workshop on Evaluating Information Access (EVIA 2016)*, pages 17–20. National Institute of Informatics, Tokyo, Japan, 2016. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/evia/04-EVIA2016-VoorheesE.pdf.

111 Ellen M. Voorhees, Daniel Samarov, and Ian Soboroff. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)*, 36(2):12:1–12:21, September 2017. doi:10.1145/3086701.

112 Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, 26(4):347–389, October 2016. doi:10.1007/s11257-016-9178-6.

113 Bo Xiao and Izak Benbasat. E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly*, 31(1):137–209, 2007. URL: http://www.jstor.org/stable/25148784.

114 David Yarowsky and Radu Florian. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310, 2002. doi:10.1017/S135132490200298X.

115 Yelp. Yelp Dataset, September 2017. Accessed: 2017-11-2. URL: https://www.yelp.com/dataset/challenge.

116 Cai-Nicolas Ziegler, Sean McNee, Joseph A Konstan, and Georg Lausen. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pages 22–32, Chiba, Japan, 2005. ACM. doi:10.1145/1060745.1060754.

117 Justin Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 307–314. ACM Press, New York, USA, 1998. doi:10.1145/290941.291014.

118 Justin Zobel, William Webber, Mark Sanderson, and Alistair Moffat. Principles for Robust Evaluation Infrastructure. In Maristella Agosti, Nicola Ferro, and Costantino Thanos, editors, *Proc. Workshop on Data infrastructurEs for Supporting Information Retrieval Evaluation (DESIRE 2011)*, pages 3–6. ACM Press, New York, USA, 2011. doi:10.1145/2064227.2064247.

119 Ludovik Çoba and Markus Zanker. rrecsys: An R-package for Prototyping Recommendation Algorithms. In *RecSys Posters*, volume 1688 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL: http://ceur-ws.org/Vol-1688/paper-12.pdf.

120 Ludovik Çoba and Markus Zanker. Replication and Reproduction in Recommender Systems Research - Evidence from a Case-Study with the rrecsys Library. In *Advances in Artificial Intelligence: From Theory to Practice*, Lecture Notes in Computer Science, pages 305–314. Springer, Cham, June 2017. doi:10.1007/978-3-319-60042-0_36.