Uncertainty Management in Information Systems – Executive Summary by the Organizers

Christoph Koch[1], Birgitta König-Ries[2], Volker Markl[3], and Maurice van Keulen[4]

[1] Cornell University - Ithaca, USA
[2] Universität Jena, Germany
[3] TU Berlin, Germany
[4] University of Twente, The Netherlands

This executive summary provides a brief overview of the topic, the organization, and the outcome of the Dagstuhl Seminar on Uncertainty Management in Information Systems.

1 Topic and Aims of the Seminar

Computer science has long pretended that information systems are perfect mirror images of a perfect world. Database management systems, e.g., work under the assumption that the data stored represent a correct subset of the real world. Of course, this idealized assumption is rarely true. Information systems contain

- wrong information caused, e.g., by data entry errors: This is a common problem for instance in genomic databases
- imprecise or falsely precise information, e.g., a measuring device will provide information with a certain precision only. Typically, information systems store the measured date, but do not store information about the conditions under which this data is true and the precision achieved.
- incomplete information. A certain piece of information may not be available to the information system.
- inconsistent information. Different information systems may contain contradictory information.

In the past information systems have worked around these flaws by extensive consistency checking, plausibility checks, or human discovery and correction. These solutions are bound to fail as systems become ever more distributed, the information more globalized, and the individual systems more autonomous. Hence, we need to find ways for our information systems to directly deal with the uncertainty induced by them.

Nor is imperfection necessarily a bad thing. Inconsistent or unknown information has been addressed by more or less ad-hoc concepts like "NULL" values in SQL. Take inconsistent information. It may reflect information collected under different circumstances or in different contexts, i.e., it may represent different views on the same phenomenon, and the sum total may very well carry more information than any single one.

Dagstuhl Seminar Proceedings 08421
Uncertainty Management in Information Systems
http://drops.dagstuhl.de/opus/volltexte/2009/1940

The challenge, then, is to make system operation resilient to imperfect data. Resilience is not simply a matter of correction but more so of reconciling what appears contradictory information.

Meeting the challenge becomes particularly pressing when we consider the modern development of the computing environment into large-scale, open, mobile, extremely widely distributed systems. Even if everything works correctly, it will no longer be possible to guarantee consistency across such systems. Consider as an example a large-scale peer-to-peer organization. Each peer observes part of the environment only. Even collectively the peers will never observe the environment at the same instance in time. Hence, there is neither individually nor collectively a consistent image of the environment. This reminds one of the uncertainty principle in physics which states that predictions can be made only within certain probabilities. Consequently, what such systems need to incorporate is what is referred to as uncertainty management. We will need mechanisms that allow the individual components to function despite the fact that they have incomplete and maybe incorrect knowledge, and that the system as a whole reaches its goals by limiting the collective uncertainty of collaborating subsystems to acceptable levels of uncertainty.

Uncertainty is an issue that appears in many disciplines. The aim of the seminar was to bring together researchers from all these communities. Some of these communities have a long history of dealing with this problem, for others, it is a new challenge. These are in particular:

- Database Management Systems
- Multi-Agent Systems
- Peer to Peer Systems
- Sensor Networks
- Data Stream Management Systems
- Reputation Systems
- Context-Aware Systems
- Artificial Intelligence
- Information Retrieval
- Self-organizing Systems
- Semantic Web
- Market Economics, Decision Science
- Fuzzy Systems

Each of these communities needs to deal with the issues described above, and many of them do in their own ways. Unfortunately, up to now, there has been little exchange between the communities on their approaches. The outcome of this seminar was a classification of the different types of uncertainty, an overview of the state of the art on dealing with them in the different communities, the applicability of these solutions to other types of systems, and an identification of promising avenues of research.

The seminar was roughly structured along the following three areas:

- Fundamentals, e.g., models for representing uncertain data, impact of uncertainty on (database) operations, consistency models, error correction

- Methods for uncertainty reduction and inconsistency tolerance
- Applications, e.g., obtaining information from sensor networks or stream management systems, structuring unstructured data, object localization, belief revision in agent and reputation systems, personal information management, data integration.

2 Organization of the Seminar

In this section, we will give a brief summary of interesting organizational aspects of the seminar, in particular who was attending and how the working groups were established. Overall, the seminar consisted of a mix of plenary sessions (partly talks, partly discussions), parallel sessions with talks, demos, and working groups.

2.1 Participants

While the original aim of the seminar was to have attend researchers from a very wide range of communities, in the end, the seminar was somewhat "database-heavy". The database community – in particular groups working on probabilistic databases – was very well represented. In fact, to the best of our knowledge, we had representatives from every major effort in this area attending the seminar. However, database dominance turned out to be very beneficial and there was a sufficient number of participants from other areas (AI, fuzzy systems, semantic web, bioinformatics, reputation systems), to foster interdisciplinary discussion and exchange of ideas. Overall, 44 researchers attended the seminar. A majority of those was from Europe, but the US was very well represented, too, and one participant came all the way from South Korea.

2.2 Workgroup topic selection process

Aim The organizers of the Dagstuhl seminar formulated the aim to devote a significant amount of time to parallel discussions in smaller workgroups. The workgroups would have to plenary report to all participants. Moreover, there would have to be a tangible result of the workgroup discussion in the form of an extended abstract in the proceedings. Two rounds of three or four workgroups were considered to be desirable.

The topics for the workgroups, however, would rather be determined by the participants themselves instead of given upfront by the organizers. Therefore, a plenary interactive topic selection session was held, which had the aim of determining two sets of workgroup topics for the two rounds. The topic selection process was divided into three phases:

1. Free inventarisation of candidate topics.
2. Topic clustering to come to 6 to 8 broader but coherent topics.
3. Popularity vote on the topics.
4. Selection of topics, division into two rounds, and distribution of participants over the workgroups

1	Killer applications and challenges for uncertainty
2	Semantic retrieval and uncertainty
3	Contextual information
4	Sensors/networks
5	Quality improvement for symbolic and metric data
6	Aggregation/integration
7	Benchmarks (performance + measures of uncertainty)
8	Where probabilities come from (e.g., inconsistency)
9	Similarity search
10	Present uncertainty to the user
11	Uncertainty in queries and updates
12	Efficient processing and indexing of uncertain data
13	Trust and reputation
14	Standard for update and query languages
15	Uncertainty models and reasoning (classification; simulation; incl. graphical)
16	Lineage, provenance and explanation
17	Connections between uncertainty models
18	Advantages and challenges of uncertainty
19	Uncertainty in multi agent systems
20	Uncertainty in machine learning
21	Uncertainty in query optimization
22	Uncertainty about uncertainty
23	Data mining on uncertain data
24	Uncertainty management and change (partial information fusion, decomposition)

Fig. 1. Initial list of candidate topics

Phase 1. The first phase had a divergent purpose. The participants were invited to freely volunteer possible candidate topics that were of their own interest, but for which they expected that other participants may also like to discuss. The result of this phase was the list of candidate topics of Figure 1.

Phase 2. The second phase had a convergent purpose. The participants were asked to suggest sets of related topics that could be merged into one topic. Discussion naturally arises when less topics of bigger size take shape. The result of this phase was the following list of 7 topics:

25 Agents and trust (13 + 19 + prediction of uncertainty + reasoning)
26 Classification, representation, and modelling (8 + 11 + 15 + 17)
27 Provenance and explanation (6 + 7 + 10 + 16 + 17)
28 Data quality and and and (5 + 6 + 7 + 18 + 22 + 23 + 24)
29 Analysis of uncertain data and systems issues (2 + 7 + 9 + 12 + 20 + 23 + 24)
30 Context and sensor data management (3 + 4)

31 Benchmarking[5] (7 + 14)

Nr	Topic	#1st	#2nd	#3rd
25	Agents and trust	6	3	1
26	Classification, representation, and modelling	12	7	11
27	Provenance and explanation	7	10	6
28	Data quality and and and	4	3	7
29	Analysis of uncertain data and systems issues	9	7	3
30	Context and sensor data management	0	6	6
31	Benchmarking	3	5	4

Fig. 2. Outcome of vote casting phase.

Phase 3. To assess the popularity and relevancy of the topics, a round of voting was performed. Each participant was asked to cast three votes: a first, second and third choice. Figure 2 shows the results of this round.

Phase 4. Because no topic proved too uninteresting, the participants decided to form working groups and discuss all topics. Based on a hand-raising process, it was determined that the following distribution of topics over the scheduled two rounds allowed most people to participate in working groups of their preference.

- *Round 1*: topics 25, 26, and 28.
- *Round 2*: topics 27, 29, 30, and 31.

3 Talks

We had scheduled beforehand a number of plenary talks with the aim to provide participants with a wide overview of topics in uncertainty management. In addition, participants had the opportunity to give shorter talks on their research topics in parallel sessions.

During the seminar, the following plenary talks were given:

- Anish Das Sarma, Stanford University: Trio: A System for Data, Uncertainty, and Lineage
- Amol Deshpande, University of Maryland - College Park: Uncertain Data Management for Sensor Networks

[5] Before the seminar, several participants already contacted each other with the intention to discuss this topic at the seminar, so this topic was not a result of the topic selection process.

- Norbert Fuhr, Universität Duisburg-Essen: Vague Predicates, Probabilistic Rules and 4-Valued Logic for Probabilistic Databases
- Peter Haas , IBM Almaden Center - San José: A Monte Carlo Approach to Managing Uncertain Data
- Ihab Ilyas , University of Waterloo: URank: Ranking Uncertain Data
- Christoph Koch, Cornell University: MayBMS: A System for Managing Large Uncertain and Probabilistic Databases
- Christopher Re , University of Washington: An Overview of the Mystiq Probabilistic Database
- Maurice van Keulen , University of Twente: Probabilistic Data Integration

4 What is special about Uncertainty Management in Information Systems? Advantages, Killer applications, and Challenges

In a plenary session, the seminar participants attempted to collect what is so different about uncertainty management. Where do the challenges lay? Are there advantages to dealing with uncertain data? And: Are there killer applications and what are they?

Advantages
Improve recall without affecting precision
No need to repair/cleanse
Offer DBMS advantages to the "real world" (reliability, scalability, concurrency, integrity)
Less information loss (inflexibility for aggregation)

Challenges
Locate and repair uncertainty
Connect uncertainty with semantics
High dimensionality, massive data sets, many constraints
Representation of uncertainty to the user (UI / HCI)
How much to trust a probabilistic result (confidence)?
Which functionality can be pushed/integrated into the DBMS?

"Killer" applications
Business Intelligence (IR + DBMS)
Surveillance
Healthcare
Integrity management
Fraud analysis
Long-term use of experimental data
Environmental studies
Web querying ("question answering") → reduce information overflow
Data management for experimental data → Biology
Data management for sensor data → Transportation

5 Demos

One highlight of the seminar was a very lively demo session. We asked the participants beforehand to bring their running systems and to give the participants an opportunity to experience hands-on uncertainty management. In this session, the following demos were shown:

- Ela Hunt: Fast Approximate String Searching in Databases
- Dan Olteanu, Christoph Koch: MayBMS in Action
- Amish Das Sarma, Martin Theobald: TrioOne: Layering Uncertainty and Lineage on a Conventional DBMS
- Hans-Joachim Lenz: Management of Imprecise, Incomplete and Uncertain Metric Data
- Maurice van Keulen: IMPrECISE: Good-is-good-enough Data Integration

6 Workgroups

Based on the selection process described above, workgroups were formed. The first set of groups met on Tuesday and Wednesday and then reported back to the plenary, the second set on Thursday. Details on the workgroup results can be found in the individual reports by the workgroups. We include here just a very brief description based on these reports on what the workgroups were about, hoping to encourage the reader to take a closer look at these reports.

Uncertainty and Trust. The aim of the working group was to analyze the relationship between trust and uncertainty in distributed reputation systems. Starting from the identification of sources and types of uncertainty in such systems, the group discussed their relationship to trust. Afterwards, a list of desirable properties of trust representations was compiled and finally open research challenges in the area were identified by the participants.

Lineage/Provenance in Probabilistic Databases. The group discussed the different usages of lineage information in probabilistic databases, different ways to represent lineage depending on the use case, as well as a number of open issues including approximation of lineage, uncertainty in lineage information, the relationship between lineage and privacy, and the implications non-independent input data have on lineage.

Explanation. This group's discussion was structured along three main questions: Why is explanation of results needed in (uncertain) information systems? What should such an explanation contain? How can it be provided, more precisely how should uncertainty be represented?

Probabilistic Databases Benchmark. This group was attended by representatives of groups working on probabilistic databases and discussed first steps towards a common benchmark that shall allow to compare different approaches. A number of concrete steps that will be taken in the near future were agreed upon.

Imprecision, Diversity, and Uncertainty. The main goal of this workgroup was to discuss how to measure uncertainty in the data and the model and how to determine the quality of uncertain data.

Classification, Representation, and Modeling. The discussion in this group was split up into two subgroups: the first subgroup studied how different representation and modeling alternatives currently proposed can fit in a bigger picture of theory and technology interaction, while the second subgroup focused on contrasting current system implementations and the reasons behind such diverse class of available prototypes.

7 Conclusion

The seminar confirmed our belief that uncertainty management is an extremely important area of computer science research that will need contributions from a number of disciplines to be successfully tackled. The seminar identified a number of potential killer applications and many advantages of incorporating uncertainty management in information systems. The seminar provided an excellent basis for the initiation of such interdisciplinary work, but also for the exchange of ideas and the organisation of future collaboration among groups working in the same area, as is evidenced for instance by the probabilistic databases benchmarking initiative. Here, the ”database-heavy” nature of the group turned out to be very beneficial to achieving a concrete outcome of the seminar. For a more detailed description of the results, please refer to the workgroup reports included in the seminar proceedings.

Acknowledgements.

The organizers would like to thank Peter Lockemann and Heinz Schweppe for their valuable input and support during the proposal and seminar preparations. Thanks are due also to the entire Dagstuhl staff for their great support before and during the seminar.

List of Participants

- Lyublena Antova, Cornell University - Ithaca, USA
- Hidir Aras, Universität Bremen, Germany
- Yassen Assenov, MPI für Informatik, Saarbrücken, Germany
- Clemens Beckstein, Universität Jena, Germany
- Sonja Buchegger, Deutsche Telekom Laboratories, Germany
- Christian Böhm, LMU München, Germany
- Anish Das Sarma, Stanford University, USA
- Ander De Keijzer, University of Twente, The Netherlands
- Guy De Tre, Ghent University, Belgium

- Amol Deshpande, University of Maryland, USA
- Peter Dittrich, Universität Jena, Germany
- Norbert Fuhr, Universität Duisburg-Essen, Germany
- Todd J. Green, University of Pennsylvania, USA
- Peter Haas, IBM Almaden Center, USA
- Thomas Hubauer, Siemens, München, Germany
- Ela Hunt, University of Strathclyde - Glasgow, UK
- Anthony Hunter, University College London, UK
- Seung-won Hwang, POSTECH - Pohang, South Korea
- Ihab Ilyas, University of Waterloo, Canada
- Andranik Khachatryan, Universität Karlsruhe, Germany
- Friederike Klan, Universität Jena, Germany
- Christoph Koch, Cornell University - Ithaca, USA
- Rudolf Kruse, Universität Magdeburg, Germany
- Birgitta König-Ries, Universität Jena, Germany
- Hans-Joachim Lenz, FU Berlin, Germany
- Volker Markl, TU Berlin, Germany
- Tom Matthé, Ghent University, Belgium
- Thomas Neumann, MPI für Informatik, Saarbrücken, Germany
- Dan Olteanu, University of Oxford, UK
- Jeff Z. Pan, University of Aberdeen, UK
- Claudia Plant, TU München, Germany
- Christopher Re, University of Washington, USA
- Matthias Renz, LMU München, Germany
- Heinz Schweppe, FU Berlin, Germany
- Thomas Seidl, RWTH Aachen, Germany
- Pritviraj Sen, University of Maryland, USA
- Pierre Senellart, Telecom Paris Tech, France
- Myra Spiliopoulou, Universität Magdeburg, Germany
- Mirco Stern, Universität Karlsruhe, Germany
- Martin Theobald, Stanford University, USA
- Vasilis Vassalos, Athens University of Economics and Business, Greece
- Jef Wijsen, Université de Mons, Belgium
- Ouri Wolfson, University of Illinois - Chicago, USA
- Maurice van Keulen, University of Twente, The Netherlands