

# Report of the Probabilistic Databases Benchmarking Working Group

CK = Christoph Koch

CR = Christopher Re

DO = Dan Olteanu

H-JL = Hans-Joachim Lenz

MvK = Maurice van Keulen

PH = Peter Haas

Jeff Z. Pan

# Create a benchmark now?

- It may be too early for a real benchmark.
  - No convergence of approaches yet.
  - Convergence is also not a priority, we do not want to discourage creativity.
  - We do not understand query and update languages yet but need at least a common core language that the data generators and queries are designed for.

# Create a benchmark now?

- What we really want is credibility for the probabilistic databases field:
  - The existence of many exciting applications is postulated in the various papers, but not many realistic use cases have been proposed.
  - We need such use cases to convince the database community and outsiders.
  - It would be good to have widely known and accepted use cases so that we can refer to them in our papers and use them in our experiments.

# Decisions

- Create a benchmarks/use cases WG (PMark or “Probabilistic Data Processing Council”) that persists beyond Dagstuhl.
- Create a repository for data generators, a wiki, mailing lists, use case definitions, further examples, possibly smaller data sets.
- Create a repository for larger datasets at Twente?

# Use Cases #1

- **TPC-H generator [DO]:** From MayBMS project: each possible world satisfies the TPC-H integrity constraints. Uncertainty is somewhat unrealistic in what is mainly an OLTP database: ok for challenging experiments, but ultimately not a credible use case.
- **TPC-H + modeling/prediction (?) [PH]:** the MCDB TPC-H data generator itself is standard, but there is a model for forecasting, and there are queries.
- **ERP + text extraction [PH]:** automotive, health care.

# Use Cases #2

- **Social networks [CK]:** Get data from Jon Kleinberg and Cornell sociologists; ask which probabilistic queries network scientists want to do.
- **Human resources risk management [CK]:** talk to Myra S. to understand how to make a realistic risk management use case for the HR domain (generalization of the companies-employees-skills management example from CK's talk).
- **Data integration [MvK]:** turn probabilistic information integrator into a data generator. Produce a dataset in the movie rating domain.

# Use Cases #3

- **Option pricing [PH]:** provide use case description, VG function definition for option simulation; Black-Scholes model encoding, etc.: probably late in spring, has to be published first.
- **RFID data [CR]:** contribute Washington dataset.
- **IPUMS US census data [CK]:** the data is cleaned, but we have a generator for introducing uncertainty.

# To-dos

- [PH] talk to IBM people what more may be revealed, specifically in the risk mgmt and healthcare domain.
- [H-JL] Features of uncertain data that need to be varied in a benchmark.
- [everyone] Mattis Neiling (TU Cottbus, PhD thesis) – 3 examples/use cases in object identification.
- Further possible use cases that we now do not have a contributor for:
  - Biology: genomics; gene expression analysis