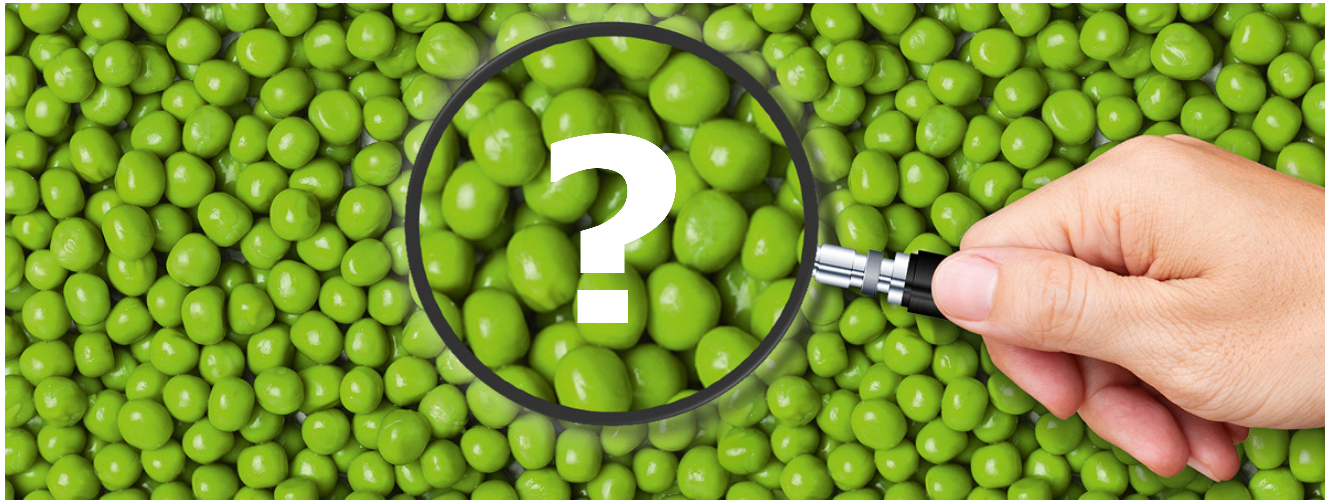
Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

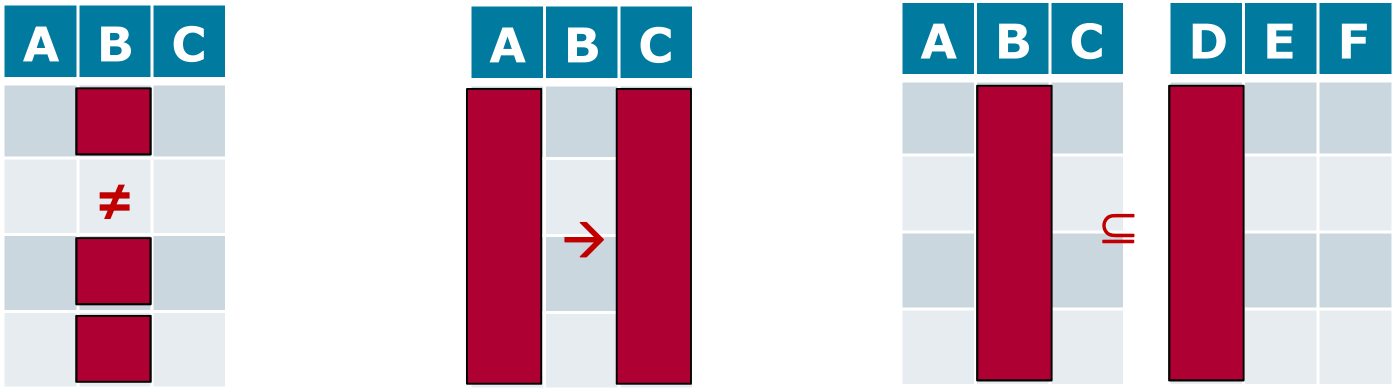
Advanced Data Profiling

Introduction



Data Profiling

- Data Profiling:
  “Process of automatically analyzing a given dataset for metadata”
- Metadata:



intra-column properties

inter-column dependencies

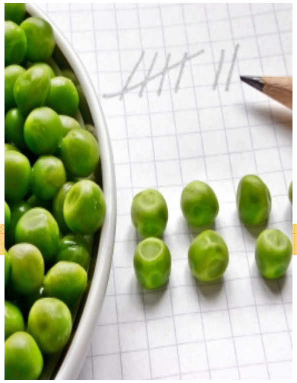
table-wide characteristics

- Motivation:
  Enabling data owners to detect errors, normalize schemata, define additional attribute properties, or integrate other sources

Vision and Goal

- Metanome:
  - Profiling platform developed in the information systems group
  - Incorporates various algorithms for Inclusion Dependencies, Functional Dependencies, Unique Column Combinations, and various other metrics
- Goal of this seminar:
  - Investigate algorithms for *Functional Dependencies*
  - Implement them in Metanome
  - Improve their performance
  - Evaluate and compare the results

Prerequisites



- Needed:
  - Knowledge in programming Java, because Metanome is written in Java
- Nice-to-have:
  - Knowledge in data profiling and in particular functional dependencies

Topics and Subprojects

- **TANE**

  Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen, "TANE: An efficient algorithm for discovering functional and approximate dependencies," *The Computer Journal*, vol. 42, no. 2, pp. 100-111, 1999.

- **fdep**

  P. A. Flach, and I. Savnik, "Database Dependency Discovery: A Machine Learning Approach," *AI Communications*, vol. 12, no. 3, pp. 139-160, 1999.

- **Dep-Miner**

  S. Lopes, J. Petit, and L. Lakhal, "Efficient Discovery of Functional Dependencies and Armstrong Relations," in *Proceedings of the International Conference on Extending Database Technology (EDBT),* 2000.

- **FastFDs**

  C. M. Wyss, C. Giannella, and E. L. Robertson "FastFDs: A Heuristic-Driven, Depth-First Algorithm for Mining Functional Dependencies from Relation Instances," in *DaWaK*, 2001.

- **FUN**

  N. Novelli, and R. Cicchetti, "FUN: An Efficient Algorithm for Mining Functional and Embedded Dependencies," in *Proceedings of the International Conference on Database Theory (ICDT),* 2001.

- **FD_Mine**

  H. Yao, H. J. Hamilton, and C. J. Butz, "FD_Mine: Discovering functional dependencies in a database using equivalences," in *Proceedings of the IEEE International Conference on Data Mining (ICDM),* 2002.

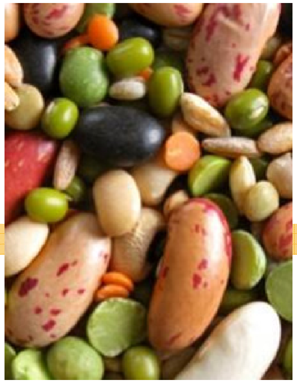
Organisation

- General:
  - 6 participants (selected randomly, if there is a superior number)
  - 3 teams of 2 students
- First half:
  - Study your individual profiling algorithm(s) from given and further literature
  - Implement the algorithm conform to the Metanome-interface
  - Find or generate an own dataset to test your implementation
  - Evaluate your baseline algorithm
  - Give a short mid-term presentation

Organisation

- Second half:
  - Enhance your algorithm:
    - Possible directions:
      - *Conditional FDs*
      - *Heuristical calculation*
      - *Incremental calculation*
      - *Scalability improvement*
    - Enhancements should be switchable!
  - Measure and evaluate your improvements
  - Give an end-term presentation
  - Prepare a paper-style submission of 4 pages per team

Details on Existing Work



Group 1: TANE, FUN, FD-Mine

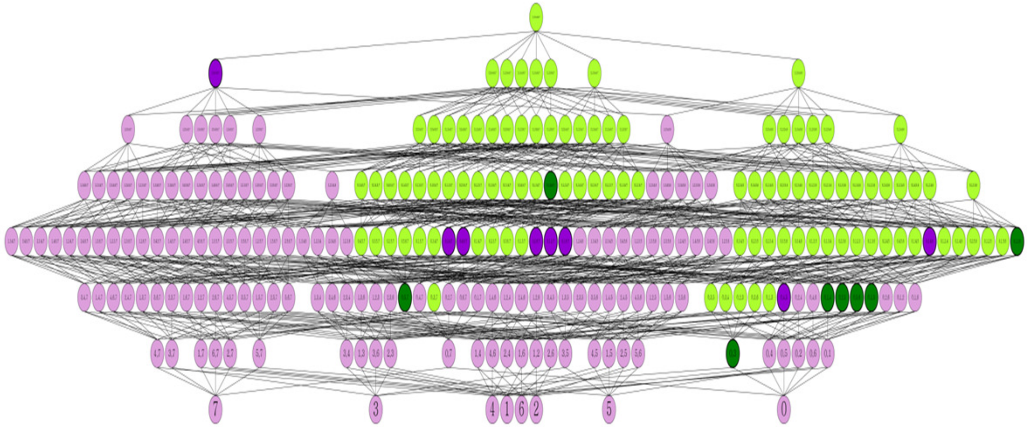
- Candidate generate-and-test approaches
- Pruning based

Group 2: Dep-Miner, FastFDs

- Formal concept analysis approaches

Group 3: fdep

- Minimal cover approach
- Machine learning concepts



Grading

- Active participation in meetings and discussions
- Implementation of the baseline algorithm using the Metanome interface
- Implementation of (at least one) algorithmic enhancement
- Mid-term presentation
- End-term presentation
- Final paper-style submission

Further Procedure

- To bindingly apply for this seminar:
  - Send an email to thorsten.papenbrock@hpi.uni-potsdam.de
  - Deadline: 20.10.2013
  - In case of too many applications, we need to choose randomly