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**

- **fdep**

- **Dep-Miner**

- **FastFDs**

- **FUN**

- **FD_Mine**
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
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
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