



Data Profiling with Metanome

Exercise: UCCs

Thorsten Papenbrock

PhD Candidate

Hasso-Plattner-Institute

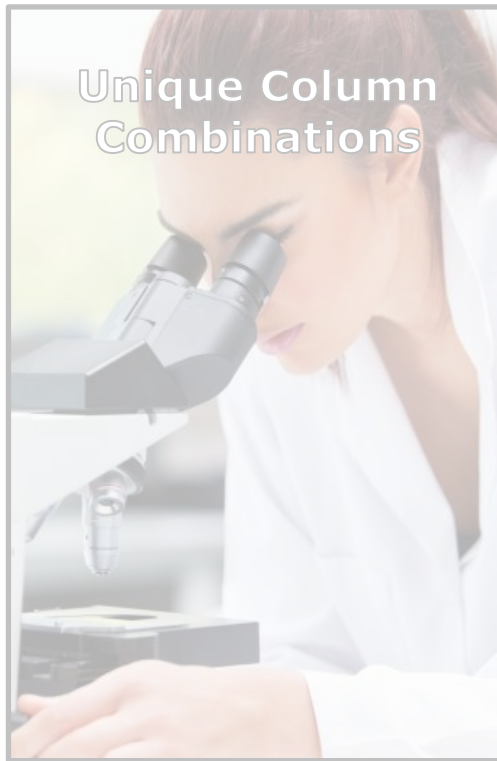
Data Profiling



Metanome



Unique Column Combinations



Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart 2

Data Profiling

The need to know your data



Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 3

Data Profiling

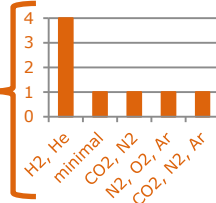
The meaning of "knowing data"

format

range { min = 0.382
max = 11.209

aggregation { sum = 173
avg = 21.625

distribution



Name	Type	Equatorial diameter	Mass	Orbital radius	Orbital period	Rotation period	Confirmed moons	Rings	Atmosphere
Mercury	Terrestrial	0.382	0.06	0.47	0.24	58.64	0	no	minimal
Venus	Terrestrial	0.949	0.82	0.72	0.62	-243.02	0	no	CO ₂ , N ₂
Earth	Terrestrial	1.0	1.0	1.0	1.0	1.00	1	no	N ₂ , O ₂ , Ar
Mars	Terrestrial	0.53	0.1	1.2	1.03	1.03	2	no	CO ₂ , N ₂ , Ar
Jupiter	Giant	11.209	317.8	5.2	11.86	9.93	67	yes	H ₂ , He
Saturn	Giant	9.449	95.2	9.54	29.46	0.43	62	yes	H ₂ , He
Uranus	Giant	4.007	14.6	19.22	84.01	-0.72	27	yes	H ₂ , He
Neptune	Giant	3.883	17.2	30.06	164.8	0.67	14	yes	H ₂ , He

SIMPLE

size

CHAR(32)

CHAR(16)

FLOAT

FLOAT

FLOAT

FLOAT

FLOAT

INTEGER

BOOLEAN

VARCHAR

data types

Chart 4

Data Profiling

The meaning of “knowing data”

inter table dependencies

Planet.Name \subseteq Moon.Planet

relationships

Mass \sim Confirmed Moons

keys

unique

Name	Type	Equatorial diameter	Mass	Orbital radius	Orbital period	Rotation period	Confirmed moons	Rings	Atmosphere
Mercury	Terrestrial	0.382	0.06	0.47	0.24	58.64	0	no	minimal
Venus	Terrestrial	0.949	0.82	0.72	0.62	-243.02	0	no	CO ₂ , N ₂
Earth	Terrestrial	1.2756	1.0	1.0	1.0	1.0	1	no	N ₂ , O ₂ , Ar
Mars	Terrestrial	0.532	0.343	1.52	1.88	1.025	2	no	CO ₂ , N ₂ , Ar
Jupiter	Giant	13.982	318	5.2	11.86	0.995	67	yes	H ₂ , He
Saturn	Giant	9.449	95.2	9.54	29.46	0.43	62	yes	H ₂ , He
Uranus	Giant	4.007	14.6	19.22	84.01	-0.72	27	yes	H ₂ , He
Neptune	Giant	3.883	17.2	30.06	164.8	0.67	14	yes	H ₂ , He

COMPLEX

Equatorial diameter x Mass > 0

rules

Atmosphere \rightarrow Rings

intra table dependencies

Chart 5

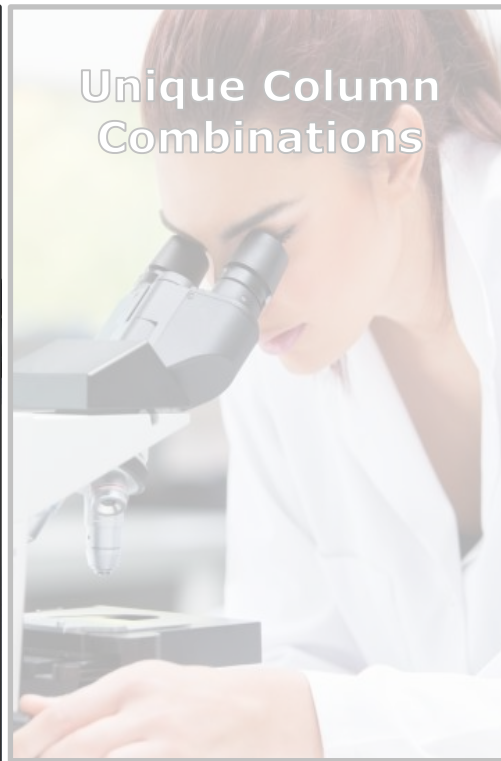
Data Profiling



Metanome



Unique Column Combinations

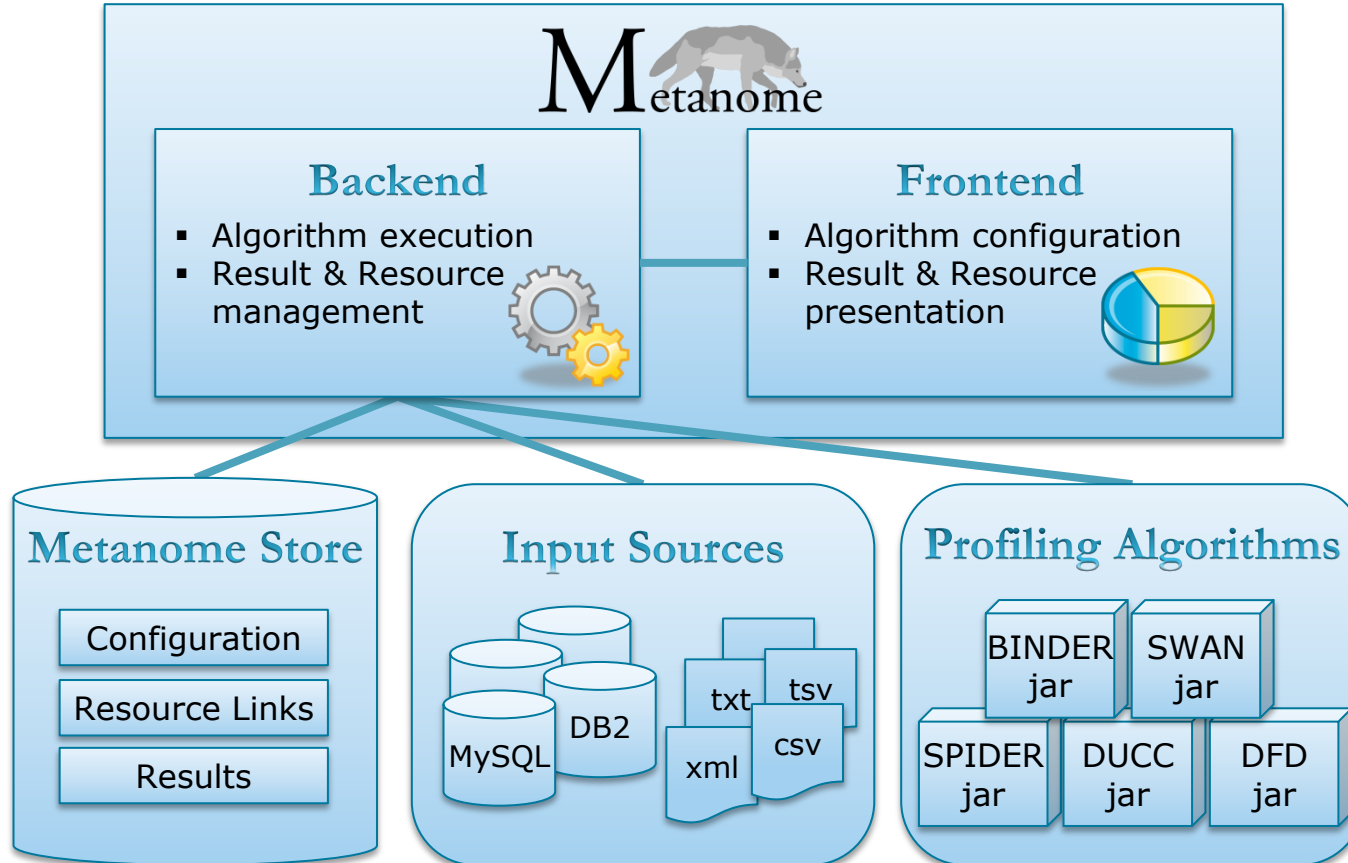


Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart 6

Metanome

An extensible architecture

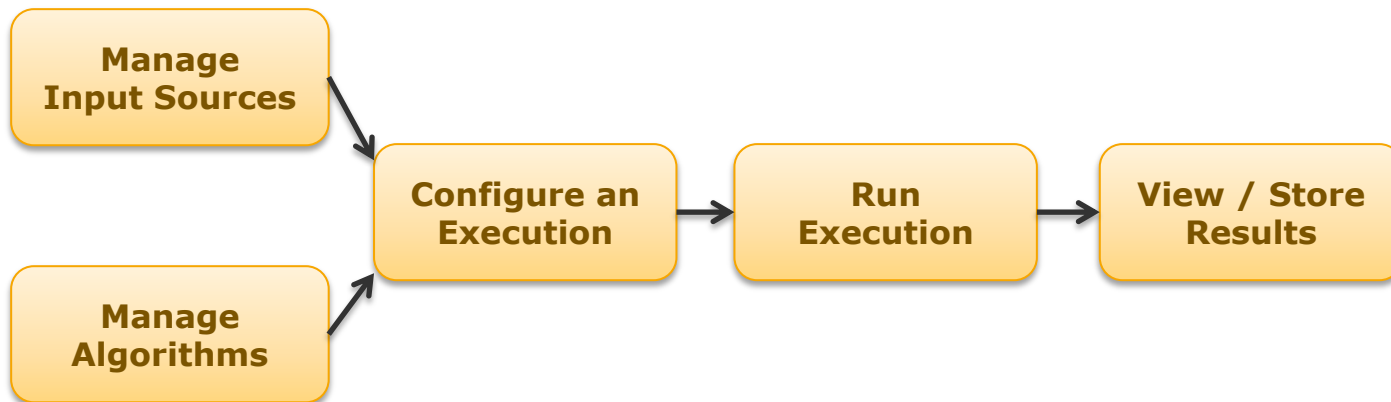


Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 7

Metanome

The typical workflow



Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 8

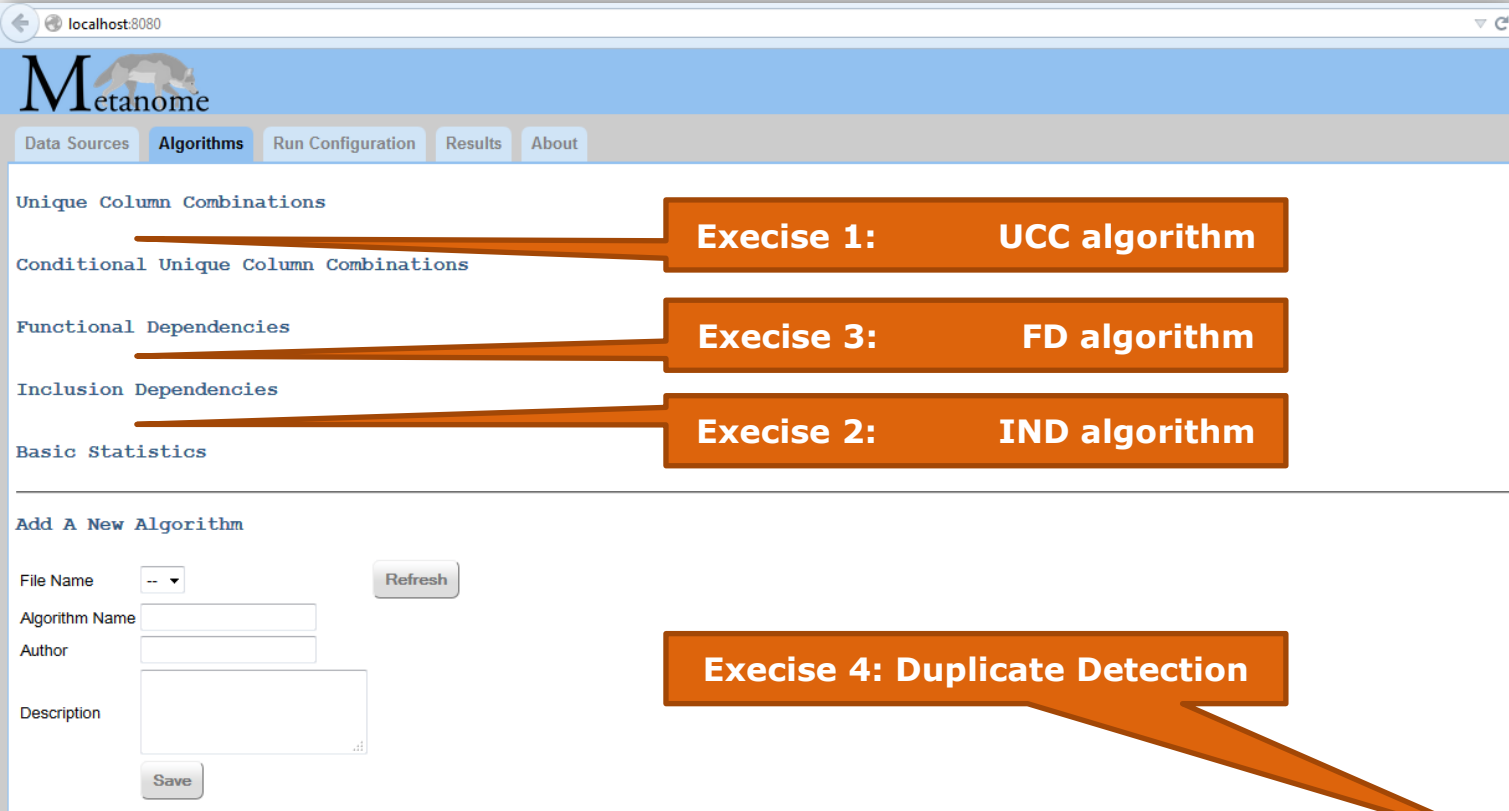
M etanome

Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 9

Metanome

The exercise: Build your own profiling tool



The screenshot shows the Metanome web application interface. The browser address bar displays 'localhost:8080'. The application header includes the Metanome logo and navigation tabs for 'Data Sources', 'Algorithms', 'Run Configuration', 'Results', and 'About'. The main content area lists several algorithm categories: 'Unique Column Combinations', 'Conditional Unique Column Combinations', 'Functional Dependencies', 'Inclusion Dependencies', and 'Basic Statistics'. Three orange callout boxes point to these categories: 'Exercise 1: UCC algorithm' points to 'Unique Column Combinations', 'Exercise 3: FD algorithm' points to 'Functional Dependencies', and 'Exercise 2: IND algorithm' points to 'Inclusion Dependencies'. Below this list is a section titled 'Add A New Algorithm' with input fields for 'File Name', 'Algorithm Name', 'Author', and 'Description', along with 'Refresh' and 'Save' buttons. A fourth orange callout box, 'Exercise 4: Duplicate Detection', points to the 'Add A New Algorithm' section.

Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart 10

Data Profiling



Metanome



Unique Column Combinations



Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart **11**

Advanced Profiling

Three important metadata



Unique Column Combinations

Key candidates

Name	Type	Equatorial diameter	Mass
Mercury	Terrestrial	0.382	0.06
Venus	Terrestrial	0.949	0.82
Earth	Terrestrial	1.000	1.00
Mars	Terrestrial	0.532	0.11
Jupiter	Giant	11.209	317.8
Saturn	Giant	9.449	95.2
Uranus	Giant	4.007	14.6
...

|Name|

Inclusion Dependencies

Foreign key candidates

Sign	Domicile
Aries	Mars
Taurus	Venus
Gemini	Mercury
Cancer	Moon
Leo	Sun
Virgo	Mercury
Libra	Venus
Scorpio	Pluto
Sagittarius	Jupiter
Capricorn	Saturn
Aquarius	Uranus
...	...

Name	Type
Mercury	Terrestrial
Venus	Terrestrial
Earth	Terrestrial
Mars	Terrestrial
Jupiter	Giant
Saturn	Giant
Uranus	Giant
...	...

Domicile \subseteq Name

Functional Dependencies

Normalization criterion

Name	Atmosphere	Rings
Mercury	minimal	no
Venus	CO ₂ , N ₂	no
Earth	N ₂ , O ₂ , Ar	no
Mars	CO ₂ , N ₂ , Ar	no
Jupiter	H ₂ , He	yes
Saturn	H ₂ , He	yes
Uranus	H ₂ , He	yes
...

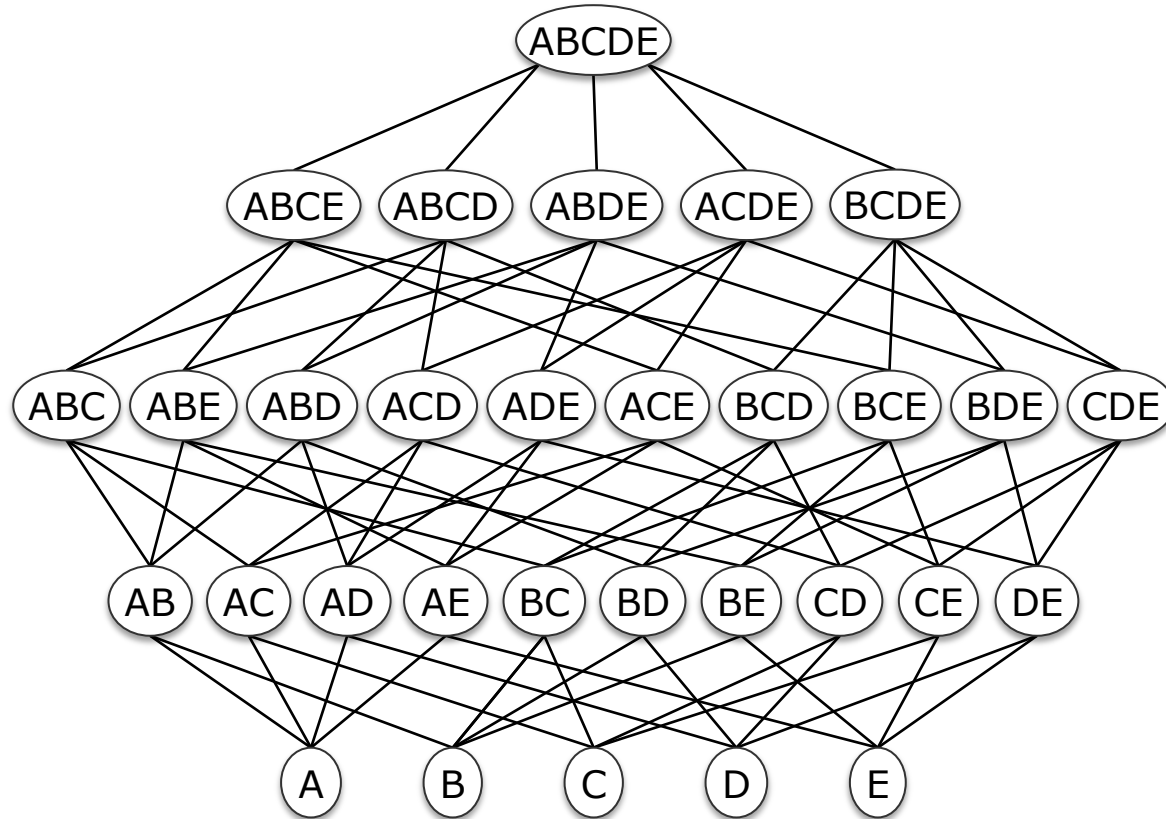
Atmosphere \rightarrow Rings

Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart 12

Advanced Profiling

Discovery of key candidates

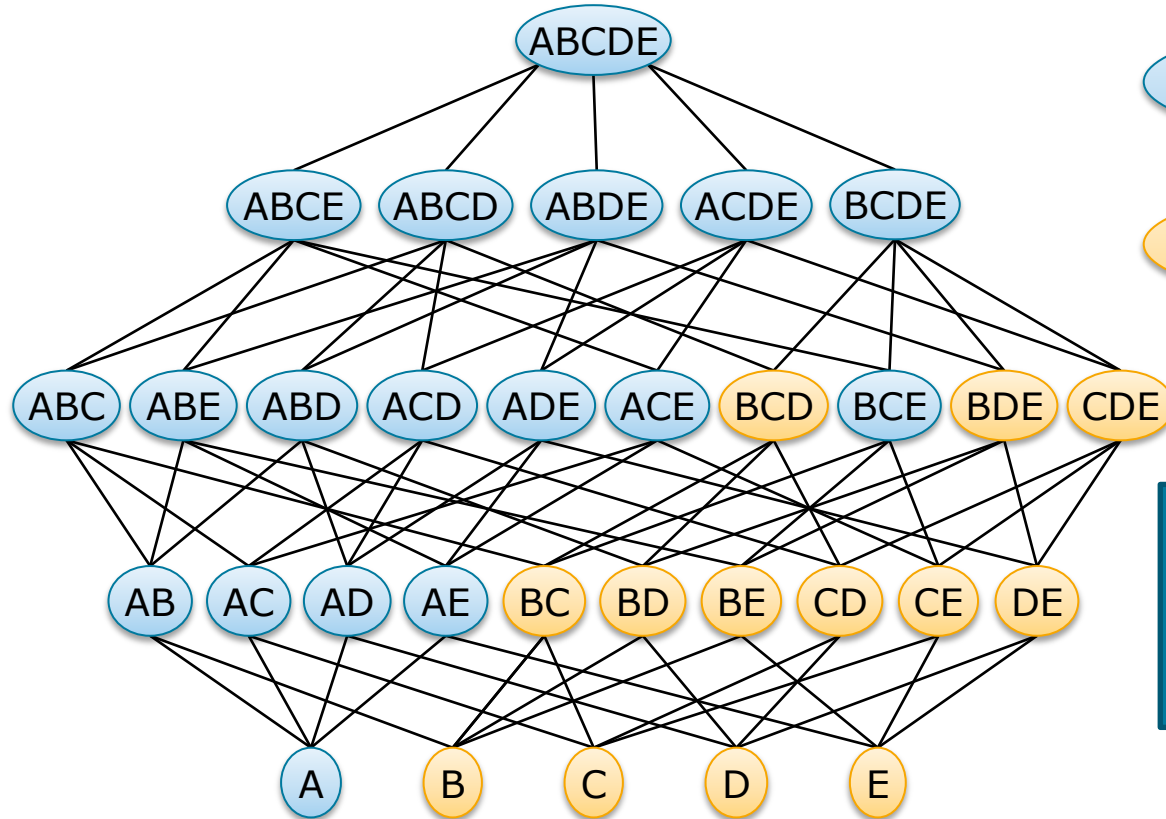


Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 13

Advanced Profiling

Discovery of key candidates



Unique
Column Combination

Non-Unique
Column Combination

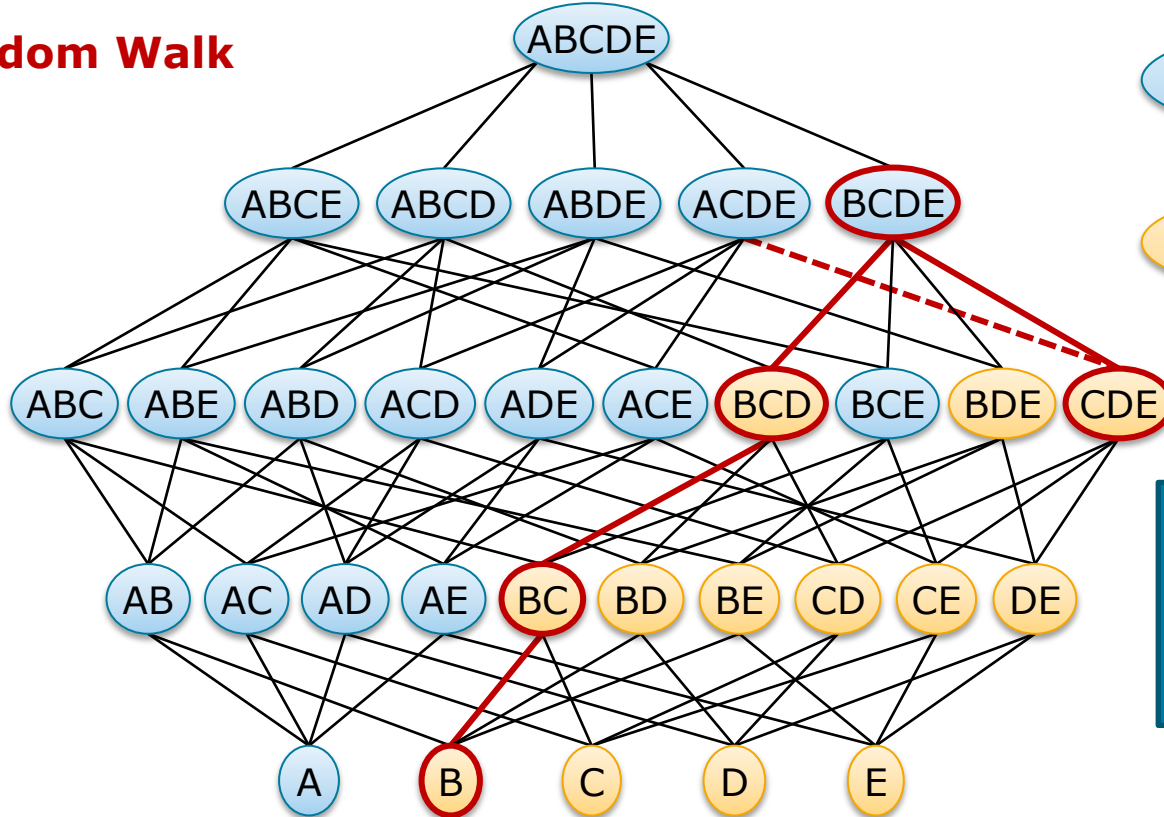
Complexity: $O(2^n - 1)$
for n attributes

Example:
10 attr ~ 1,023 checks
30 attr ~ 1,073,741,823 checks

Advanced Profiling

Discovery of key candidates – DUCC

Random Walk



Unique
Column Combination

Non-Unique
Column Combination

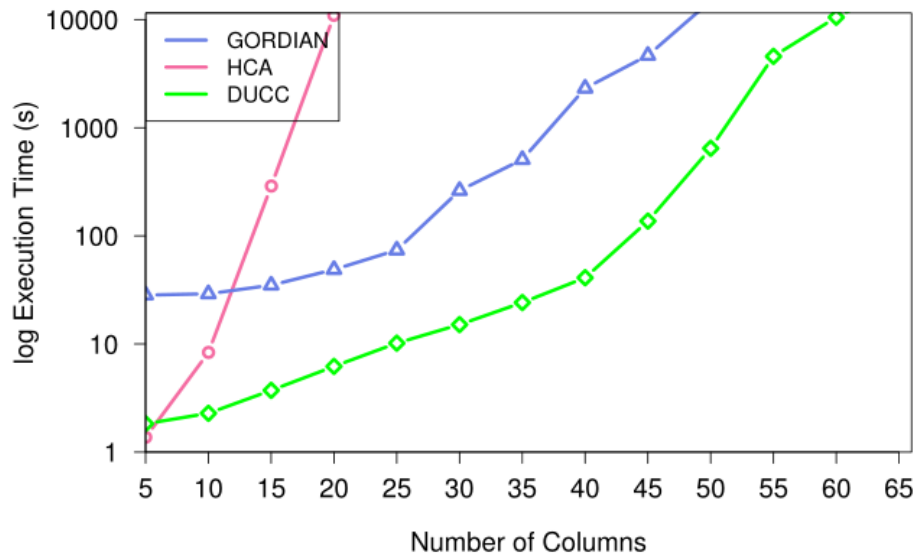
Complexity: $O(2^n - 1)$
for n attributes

Example:
10 attr \sim 1,023 checks
30 attr \sim 1,073,741,823 checks

Advanced Profiling

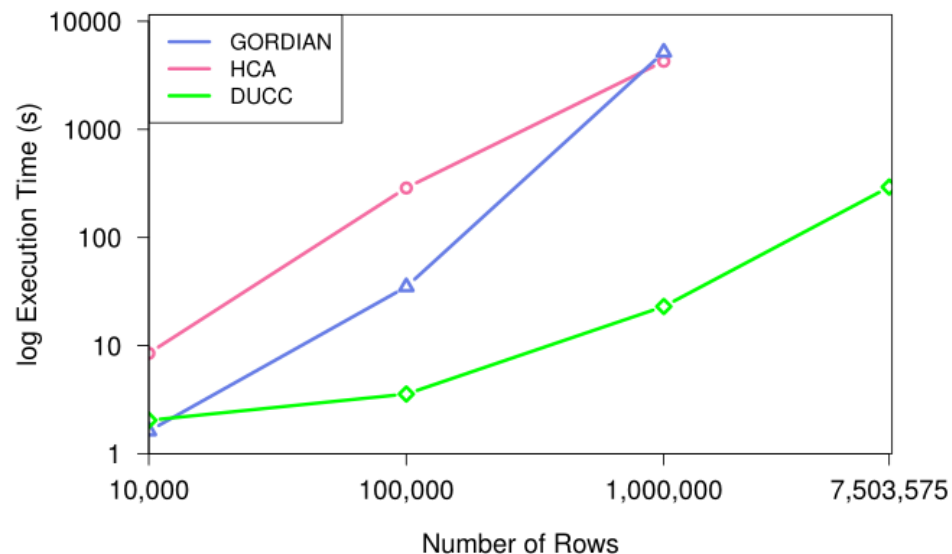
Discovery of key candidates – DUCC

Column Scalability



NCVoter

Row Scalability



NCVoter

Advanced Profiling

Discovery of key candidates – Exercise

Exercise 1

Unique Column Combinations

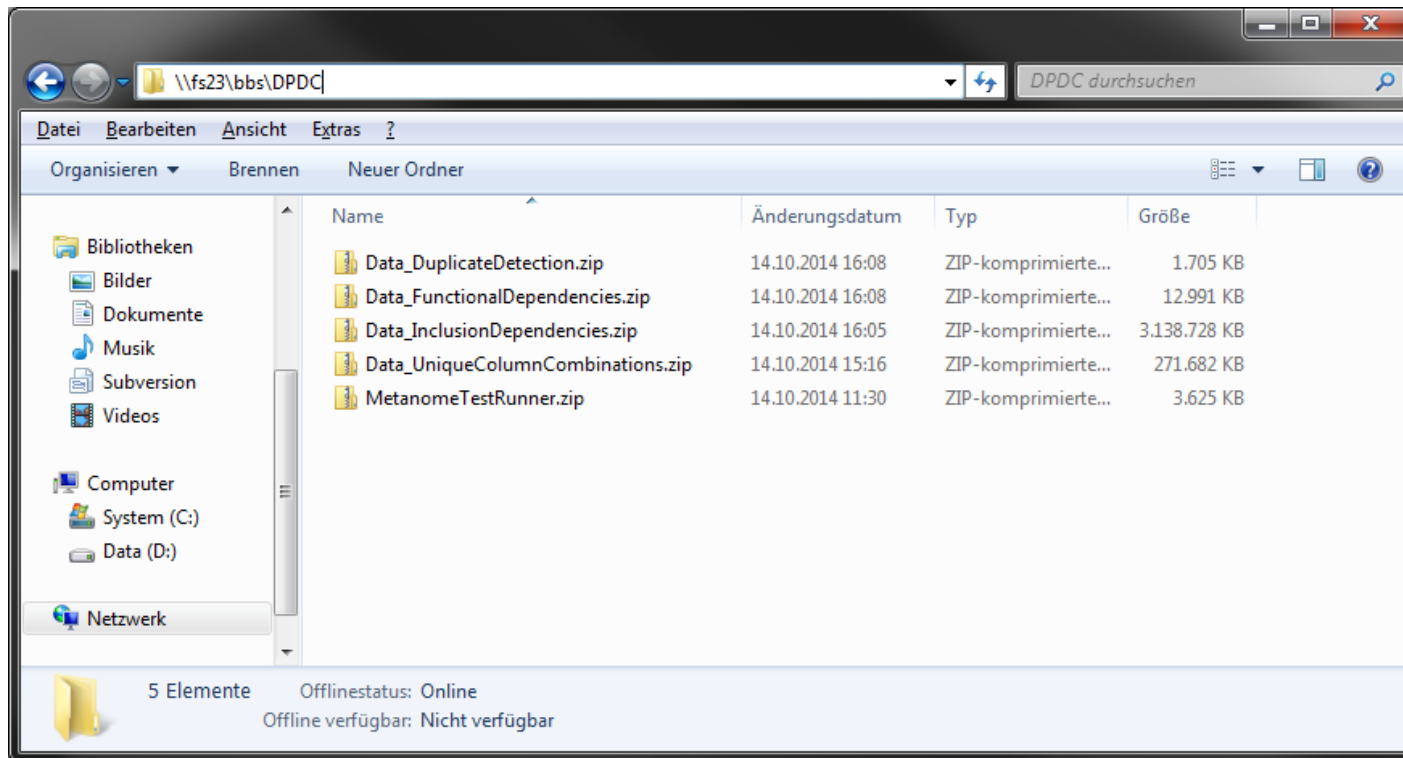
- Deadline: **Monday, 10.11.14**
- The admission to the exam requires *all* exercises to be solved.
- The exercises should be solved in teams of two students.
- The Metanome project is available at GitHub:
<https://github.com/HPI-Information-Systems/Metanome>
- The datasets and supplemental material can be found at network drive S:
\\fs3\bbs\DPDC
- The submission system can be found at:
<https://www.dcl.hpi.uni-potsdam.de/submit/>
- To solve an exercise, please submit a zip file containing the following items:
 - **<algorithm_name>.jar**: An executable Metanome algorithm.
 - **<algorithm_name>.zip**: The algorithm's source code (maven project).
 - **<algorithm_name>.docu.pdf**: Short documentation of the algorithm.
 - **<algorithm_name>.pres.pptx/ppt/pdf**: Two slides presentation of the algorithm.

Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart **17**

Advanced Profiling

Discovery of key candidates – Exercise

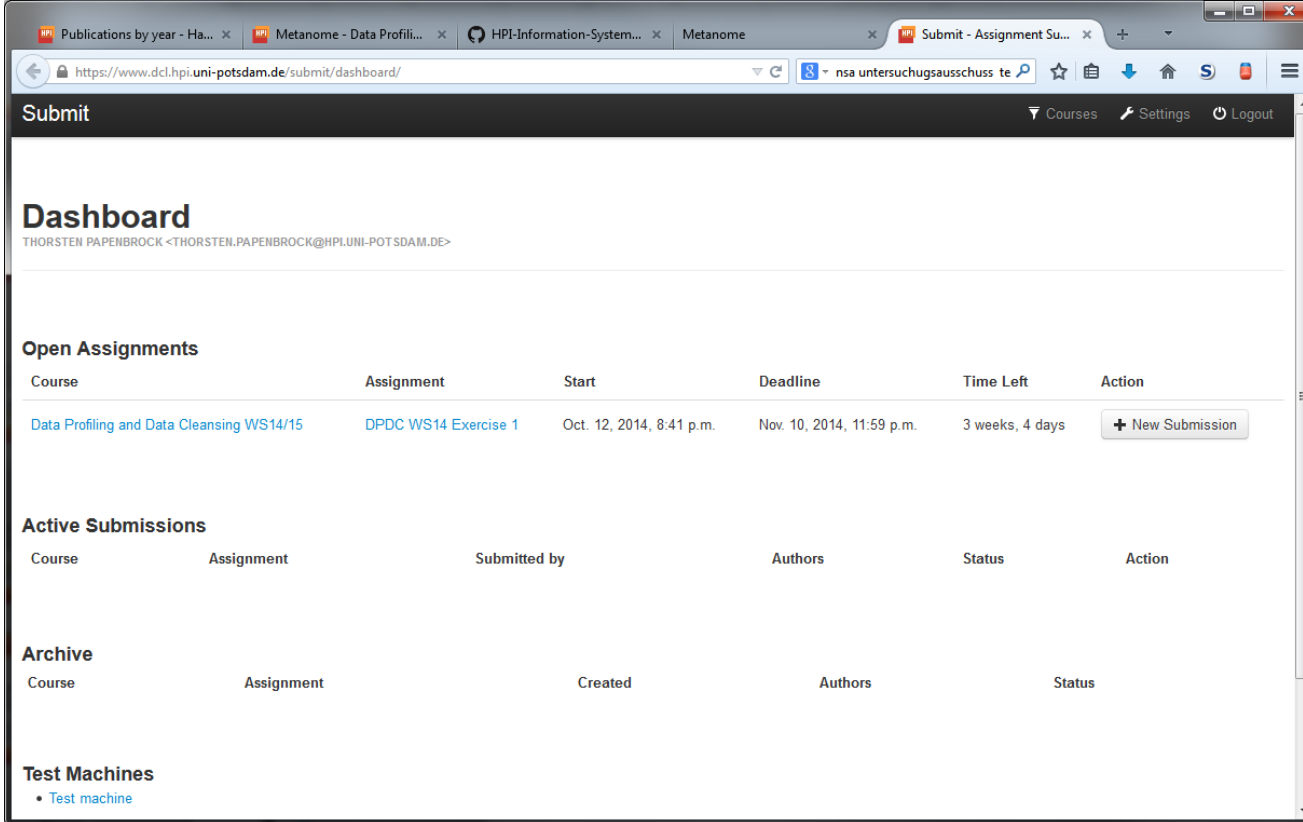


Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 18

Advanced Profiling

Discovery of key candidates – Exercise



Submit

Courses Settings Logout

Dashboard

THORSTEN PAPIENBROCK <THORSTEN.PAPIENBROCK@HPI.UNI-POTSDAM.DE>

Open Assignments

Course	Assignment	Start	Deadline	Time Left	Action
Data Profiling and Data Cleansing WS14/15	DPDC WS14 Exercise 1	Oct. 12, 2014, 8:41 p.m.	Nov. 10, 2014, 11:59 p.m.	3 weeks, 4 days	+ New Submission

Active Submissions

Course	Assignment	Submitted by	Authors	Status	Action
--------	------------	--------------	---------	--------	--------

Archive

Course	Assignment	Created	Authors	Status
--------	------------	---------	---------	--------

Test Machines

- [Test machine](#)

Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 19

Advanced Profiling

Discovery of key candidates – Exercise

Task 1: Unique Column Combinations - A discovery algorithm

Write an algorithm that discovers *all* unique column combinations on the given datasets. The rules for your implementation are as follows:

- The algorithm discovers *exact* results, so no approximate or fuzzy results are allowed.
- The algorithm is not allowed to use parallelization.
- The algorithm implements the Metanome interface and is compatible with Metanome.
- For the NULL semantic, assume $\text{NULL} \neq \text{NULL}$.

You can re-implement an existing algorithm from literature or find your own algorithm. To test and evaluate the algorithm, use the datasets provided on the network share. Your algorithm should at least be able to process the WDC_planets dataset! Before submitting your algorithm, check that it correctly executes within Metanome!

BONUS TASK: If you like to dive deeper, you can try to parallelize your discovery algorithm. What effect does parallelization have on your pruning strategies? You can also try to make the search approximate or conditional. Which changes make your algorithm faster and which make it slower? If you made changes to your main algorithm, please submit them as a separate algorithm, e.g. <algorithm_name>_conditional.jar

Template Algorithm:
<https://github.com/HPI-Information-Systems/Metanome>

Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart 20

Advanced Profiling

How to start

BINDER [papenbrock master]

- src/main/java
 - de.uni_potsdam.hpi.metanome.algorithms.binder
 - BINDERDatabase.java
 - BINDERFile.java
 - de.uni_potsdam.hpi.metanome.algorithms.binder.core
 - BINDER.java
 - de.uni_potsdam.hpi.metanome.algorithms.binder.io
 - FileInputIterator.java
 - InputIterator.java
 - SqlInputIterator.java
 - de.uni_potsdam.hpi.metanome.algorithms.binder.structures
 - Attribute.java
 - AttributeCombination.java
 - IntSingleLinkedList.java
 - Level.java
 - PruningStatistics.java
- src/test/java
- JRE System Library [JavaSE-1.7]
- Maven Dependencies
- src
- target
- pom.xml

MetanomeTestRunner [papenbrock master]

- src/main/java
 - de.uni_potsdam.hpi.metanome_test_runner
 - Main.java
 - MetanomeTestRunner.java
 - de.uni_potsdam.hpi.metanome_test_runner.config
 - Config.java
 - de.uni_potsdam.hpi.metanome_test_runner.mocks
 - MetanomeMock.java
- src/test/java
- JRE System Library [JavaSE-1.7]
- Maven Dependencies
- Referenced Libraries
- dependencies
- src
- target
- pom.xml

Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart 21

Advanced Profiling

Discovery of key candidates – Exercise

Task 2: Documentation

Write a short (max one A4 page) documentation for your algorithm describing the algorithm that you implemented:

- a) Describe the algorithm's basic idea. How does the algorithm cope with the complexity of the given task?
- b) If you used an algorithm from literature, provide a reference to the according publication.
- c) If you came up with an own approach, provide one or two arguments why it is or could be better than related algorithms.
- d) If your algorithm implements an adaption or optimization of existing approaches, describe these briefly.
- e) State if your algorithm (jar) should be published as Metanome algorithm providing you as the authors.
- f) *If you solved a bonus task, please discuss your findings here as well.*

In the same document, answer the following questions:

- a) How many unique column combinations did your algorithm find on the provided datasets?
- b) How long did the discovery take on each dataset and what machine did you use?
- c) Did you discover any limitations of your approach (e.g. runtime or memory consumption) that made computing a certain dataset impossible?
- d) What are the conceptual differences between $NULL \neq NULL$ and $NULL = NULL$ and how

Data Profiling with Metanome

Thorsten Papebrock,
PhD Candidate,
9th of October, 2014
Chart **22**

Advanced Profiling

Discovery of key candidates – Exercise

Task 3: Presentation

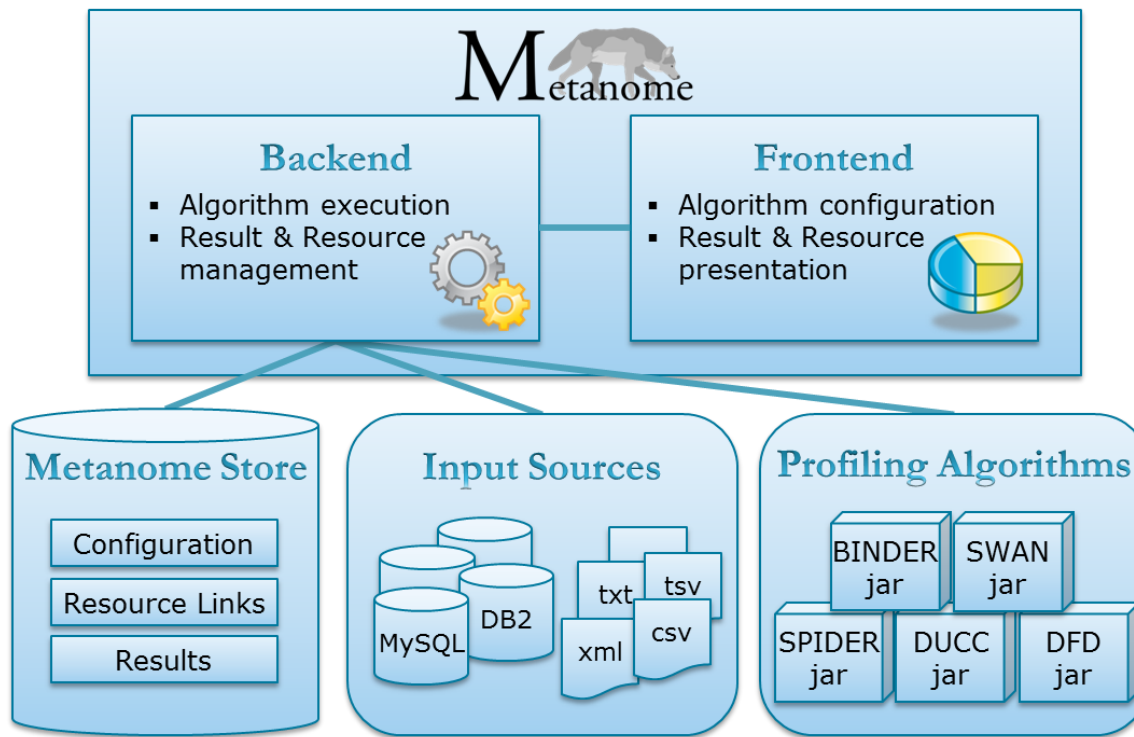
Prepare two slides for a short, 5 min presentation of your algorithm in the lecture. One slide about the algorithm and one slide about its performance. Here are some ideas for these slides:

- a) Explain the idea of your approach and why you think it works well in comparison to others.
- b) How did your algorithm perform on the given datasets?
- c) Did you make any interesting observations?
- d) What lets your algorithm crash?
- e) How hard was it to implement your algorithm?
- f) *This is also a good time to introduce an optimization or adaption of your algorithm.*

Note that each team will present its work once!

Data Profiling with Metanome

Thorsten Papenbrock,
PhD Candidate,
9th of October, 2014
Chart **23**



Data Profiling with Metanome

www.metanome.de

Thorsten Papenbrock

PhD Candidate

Hasso-Plattner-Institute