Metanome

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Data Profiling with Metanome

Exercise: FDs

Thorsten Papenbrock
PhD Candidate
Hasso-Plattner-Institute

Advanced Profiling
Three important metadata

Hasso Plattner Institut

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 ⊆ 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 → Rings

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **2**

Exercise 2
Discovery of foreign key candidates

- All teams have passed the exercise:
  - 35 submissions
  - No duplicate algorithm names!
  - No incorrect results (on given datasets)!
  - No import errors in Metanome (apart from execution errors)!

Exercise 2
Short presentations – Part 1

Exercise 2
Our evaluation

- DELL Optiplex 9010
  - CPU: Intel i5 3.2 GHz
  - RAM: 8 GB (2 GB for Metanome JVM)
  - OS: Debian 64-bit
  - JVM: **Java 1.8**

AlexoFredInclusions
curunir
dennis_marius.ind
DJ_IND
dpdc-cnms-ind
DreamteamInd
FrohnOttoInc-preconfigured
GansGans
HorLehInDeBinder
IncDep
InclusionDetector
ind_grundke_wiese
IND_Kirsten_Zwerg
IND_schaeffer_zoellner_hashmap
IND_SPIRO
IndBotheReissaus
INDianaJonesRaiderOfTheLostInclusion
INDIE
INDJRFM
IND-MoritzChristian
IndPerchykSchmidt
Jung
JungRohloffInd
Kankra
klinger_marten_inclusion
LucieKerstinIND
MMINDs
MuchDiscoVeryDisco
MyInd
PCInd
PracticeIndJoerke
RT_BindBuck
SBMMIND
smart-data-cat-IND
Tsun12Ind
YuckInd

WDC-planets dataset

AlexoFredInclusions
curunir
dennis_marius.ind
DJ_IND
dpdc-cnms-ind
DreamteamInd
FrohnOttoInc-preconfigured
GansGans
HorLehInDeBinder
IncDep
InclusionDetector
ind_grundke_wiese
IND_Kirsten_Zwerg
IND_schaeffer_zoellner_hashmap
IND_SPIRO
IndBotheReissaus
INDianaJonesRaiderOfTheLostInclusion
INDIE
INDJRFM
IND-MoritzChristian
IndPerchykSchmidt
Jung
JungRohloffInd
Kankra
klinger_marten_inclusion
LucieKerstinIND
MMINDs
MuchDiscoVeryDisco
MyInd
PCInd
PracticeIndJoerke
RT_BindBuck
SBMMIND
smart-data-cat-IND
Tsun12Ind
YuckInd

Only One InputGenerator

No working parameterization found

WDC-planets dataset

AlexoFredInclusions

dennis_marius.ind
DJ_IND
dpdc-cnms-ind
DreamteamInd
FrohnOttoInc-preconfigured
GansGans
HorLehInDeBinder
IncDep
InclusionDetector
ind_grundke_wiese
IND_Kirsten_Zwerg
IND_schaeffer_zoellner_hashmap
IND_SPIRO
IndBotheReissaus
INDianaJonesRaiderOfTheLostInclusion
INDIE
INDJRFM
IND-MoritzChristian
IndPerchykSchmidt
Jung
JungRohloffInd
Kankra
klinger_marten_inclusion
LucieKerstinIND
MMINDs
MuchDiscoVeryDisco

PCInd
PracticeIndJoerke
RT_BindBuck
SBMMIND
smart-data-cat-IND
Tsun12Ind
YuckInd

TPC-H dataset

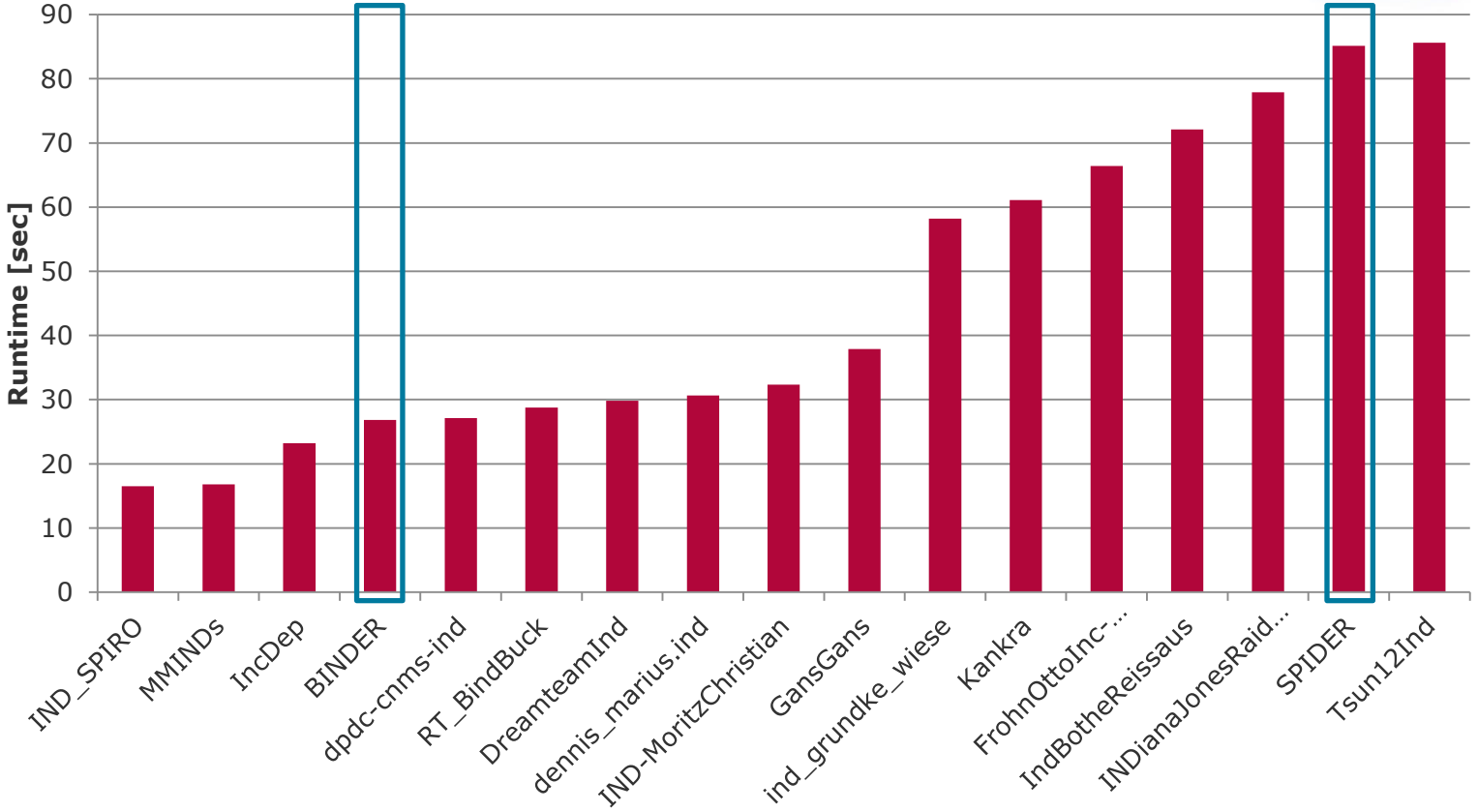
(Customer, Supplier, Nation, Part)

AlexoFredInclusions	**OutOfMemory**
dennis_marius.ind	
DJ_IND	**OutOfMemory**
dpdc-cnms-ind	
DreamteamInd	
FrohnOttoInc-preconfigured	
GansGans	
HorLehInDeBinder	**OutOfMemory**
IncDep	
InclusionDetector	**OutOfMemory**
ind_grundke_wiese	
IND_Kirsten_Zwerg	**>15 min**
IND_schaeffer_zoellner_hashmap	**OutOfMemory**
IND_SPIRO	
IndBotheReissaus	
INDianaJonesRaiderOfTheLostInclusion	
INDIE	**OutOfMemory**
INDJRFM	**Wrong Result**
IND-MoritzChristian	
IndPerchykSchmidt	**OutOfMemory**
Jung	**OutOfMemory**
JungRohloffInd	**OutOfMemory**
Kankra	
klinger_marten_inclusion	**>15 min**
LucieKerstinIND	**OutOfMemory**
MMINDs	
MuchDiscoVeryDisco	**OutOfMemory**
PCInd	**OutOfMemory**
PracticeIndJoerke	**>15 min**
RT_BindBuck	
SBMMIND	**OutOfMemory**
smart-data-cat-IND	**>15 min**
Tsun12Ind	
YuckInd	**>15 min**

TPC-H dataset

(Customer, Supplier, Nation, Part)

Exercise 2
Our evaluation – TPC-H (reduced)



dennis_marius.ind

dpdc-cnms-ind
DreamteamInd
FrohnOttoInc-preconfigured
GansGans

IncDep

ind_grundke_wiese

IND_SPIRO
IndBotheReissaus
INDianaJonesRaiderOfTheLostInclusion

IND-MoritzChristian

Kankra

MMINDs

RT_BindBuck

Tsun12Ind

TPC-H dataset

(complete)

dennis_marius.ind **OutOfMemory**

dpdc-cnms-ind **OutOfMemory**
DreamteamInd
FrohnOttoInc-preconfigured
GansGans

IncDep

ind_grundke_wiese

IND_SPIRO
IndBotheReissaus
INDianaJonesRaiderOfTheLostInclusion **Too many open files**

IND-MoritzChristian

Kankra **OutOfMemory**

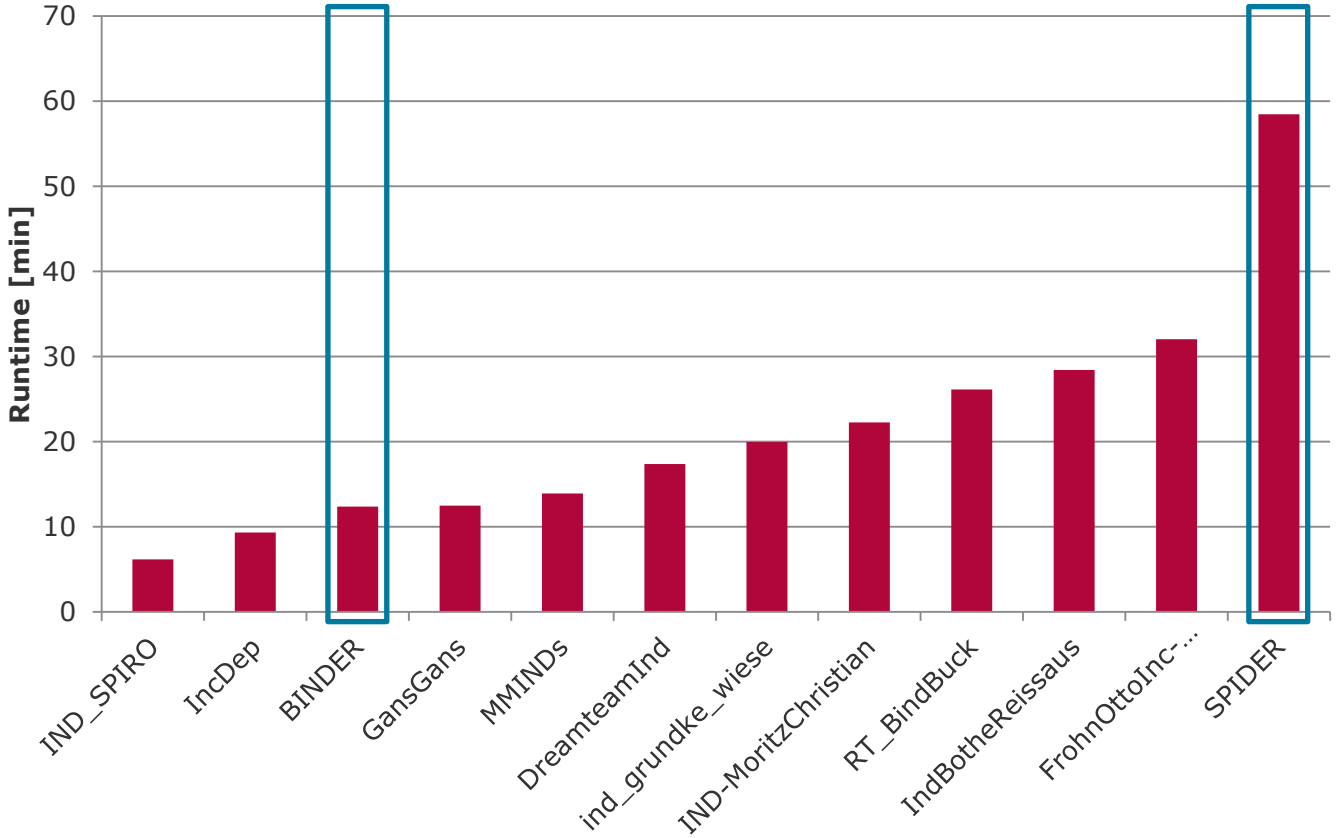
MMINDs

RT_BindBuck

Tsun12Ind **OutOfMemory**

TPC-H dataset
(complete)

Exercise 2
Our evaluation – TPC-H (complete)



DreamteamInd
FrohnOttoInc-preconfigured
GansGans

IncDep

ind_grundke_wiese

IND_SPIRO
IndBotheReissaus

IND-MoritzChristian

MMINDs

RT_BindBuck

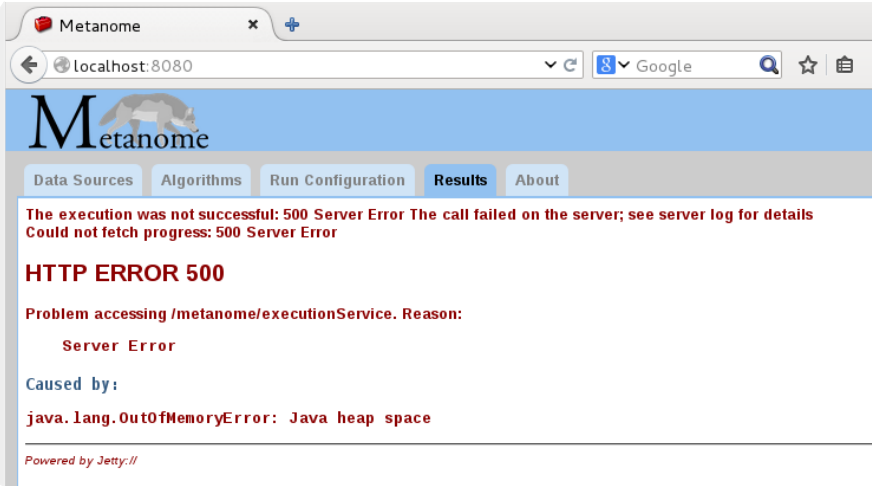
TPC-H and PLISTA dataset

(complete and sample; 1GB RAM)

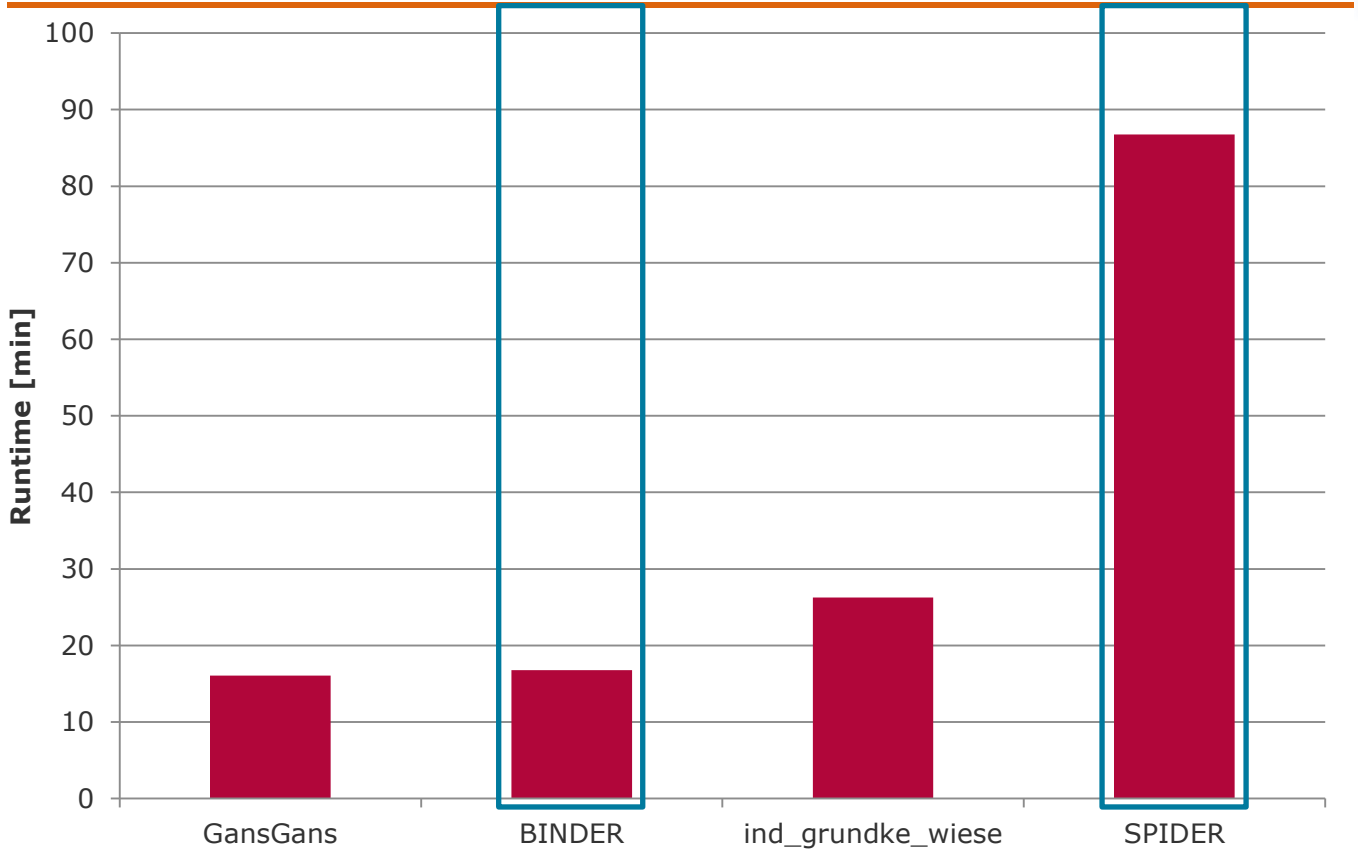
DreamteamInd	**OutOfMemory**
FrohnOttoInc-preconfigured	**>12 h**
GansGans	
IncDep	**OutOfMemory**
ind_grundke_wiese	
IND_SPIRO	**OutOfMemory**
IndBotheReissaus	**OutOfMemory**
IND-MoritzChristian	**NumberFormatException**
MMINDs	**OutOfMemory**
RT_BindBuck	**OutOfMemory**

Also happens to SPIDER and BINDER:
size(candidates) > size(RAM)

TPC-H and PLISTA dataset
(complete and sample; 1GB RAM)



Exercise 2
Our evaluation – TPC-H and PLISTA



Exercise 2
Short presentations – Part 2

Advanced Profiling
Three important metadata

Hasso Plattner Institut

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 ⊆ Name

Functional Dependencies

Normalization criterion

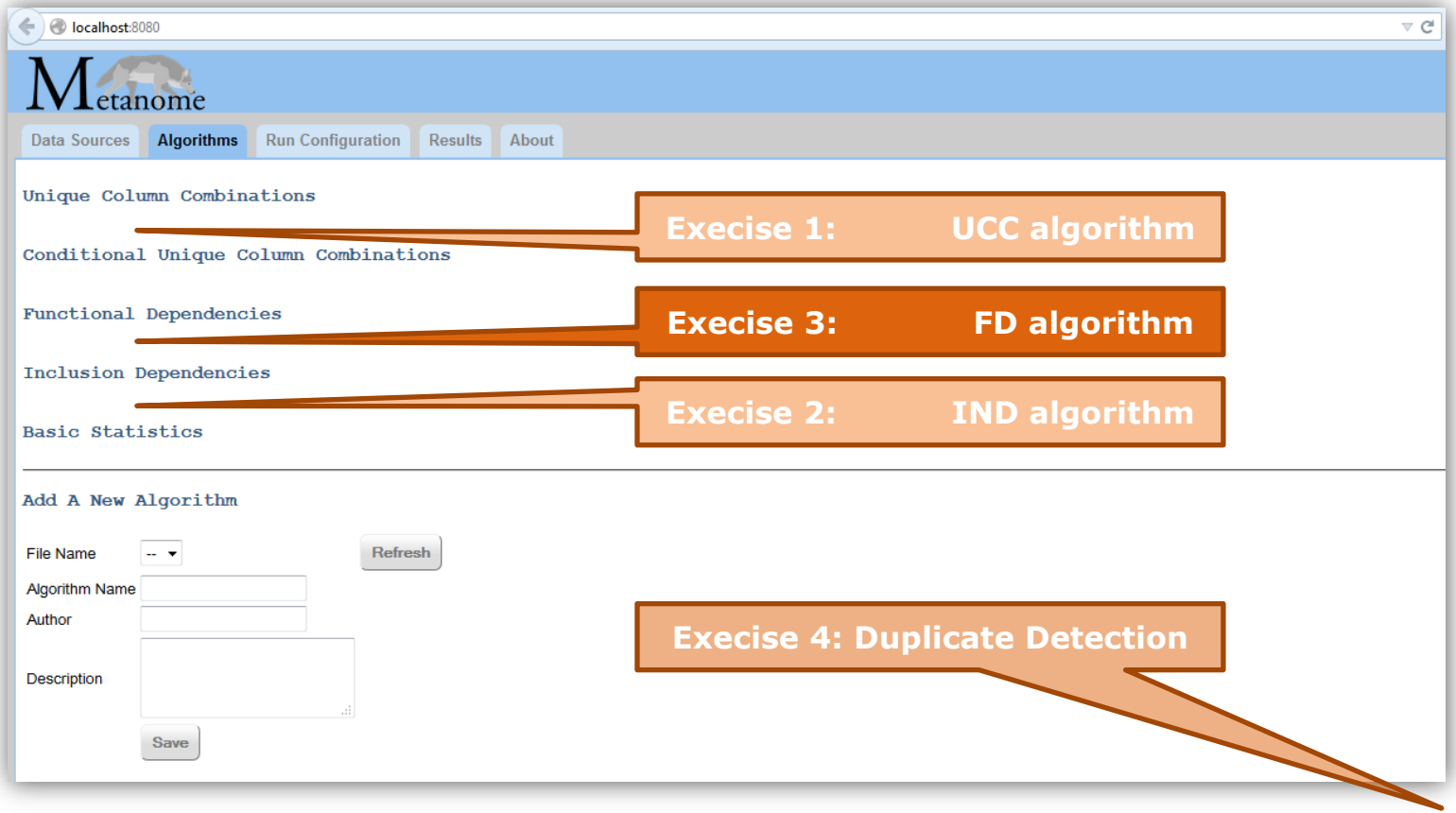
Name	Atmosphere	Rings
Mercury	minimal	no
Venus	CO_2, N_2	no
Earth	N_2, O_2, Ar	no
Mars	CO_2, N_2, Ar	no
Jupiter	H_2, He	yes
Saturn	H_2, He	yes
Uranus	H_2, He	yes
...	...	...

Atmosphere → Rings

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014

Metanome
The exercise: Build your own profiling tool



Exercise 3
Discovery of functional dependencies

Exercise 3
Functional Dependencies

- Deadline: **Monday, 05.01.15**
- 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.

Exercise 3
Discovery of functional dependencies

Task 1: Functional Dependencies - A discovery algorithm

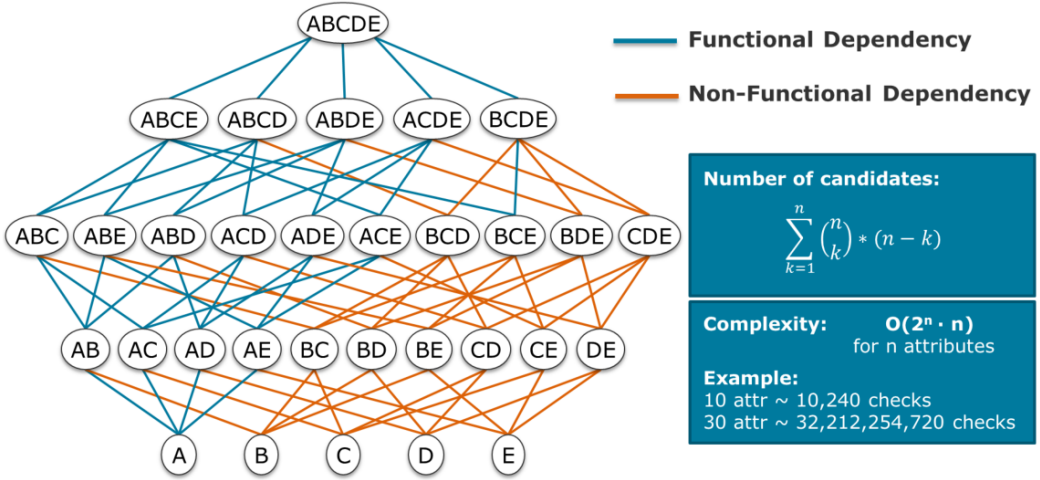
Write an algorithm that discovers *all unary and n-ary* functional dependencies on the given data-sets. The rules for your implementation are as follows:

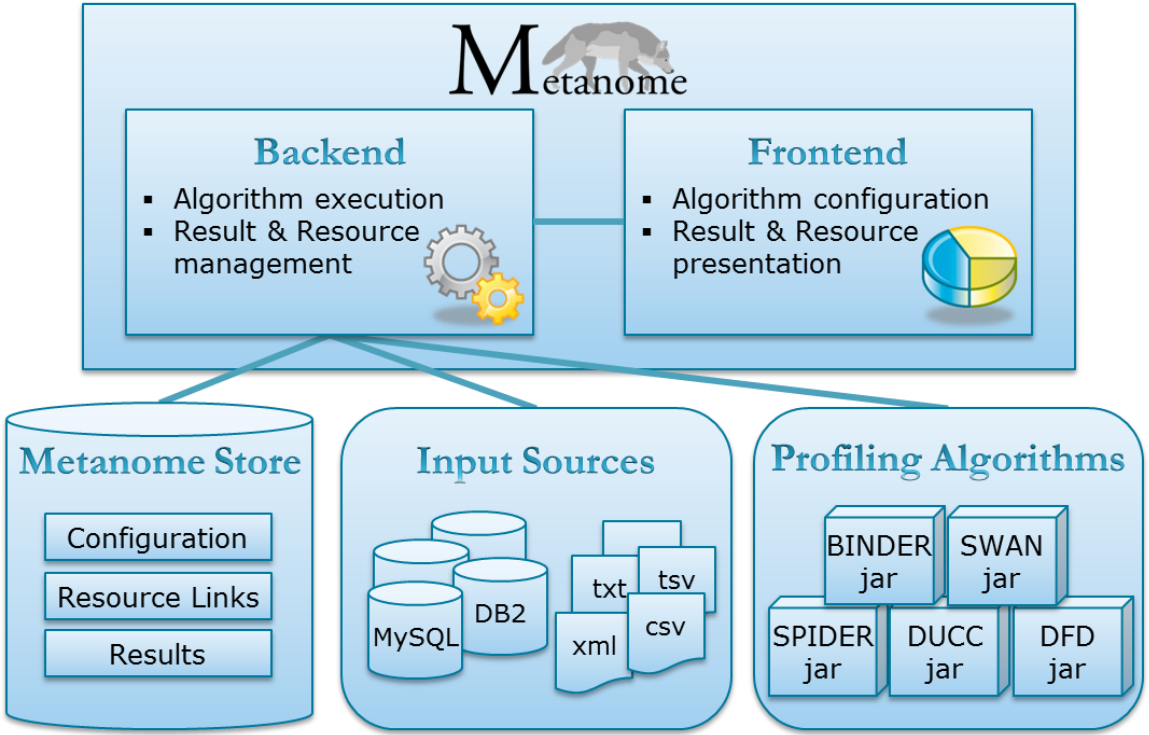
a) The algorithm discovers *exact* results, so no approximate or fuzzy results are allowed.

b) The algorithm is not allowed to use parallelization.

c) The algorithm implements the Metanome interface and is compatible with Metanome.

d) For the NULL semantic, assume NULL = NULL.





HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Data Profiling with Metanome

www.metanome.de

Thorsten Papenbrock
PhD Candidate
Hasso-Plattner-Institute