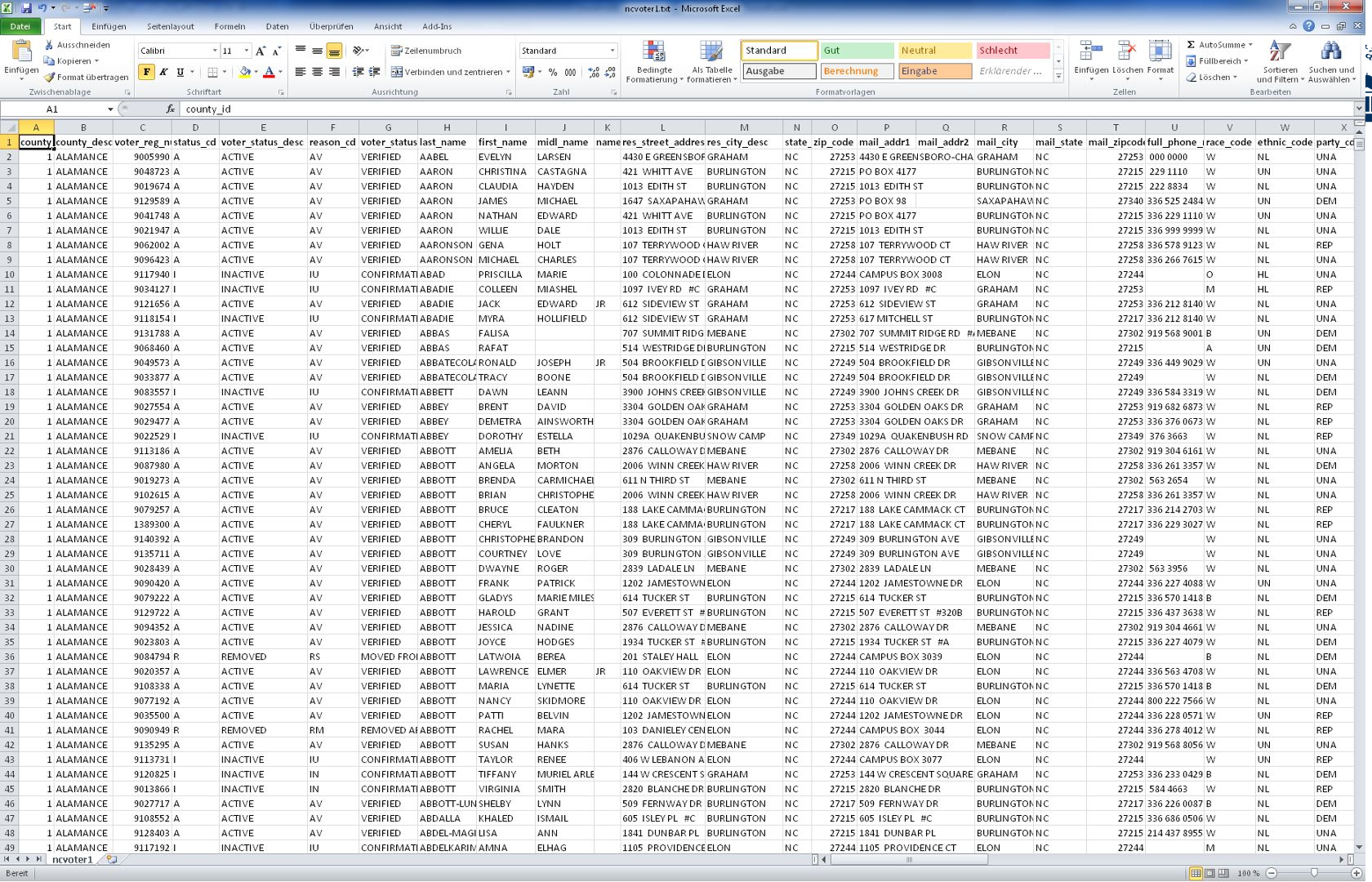
Photo by Maksym Kaharlytskyi on Unsplash

Approximate Data Profiling

Prof. Dr. Felix Naumann
Tobias Bleifuß
Youri Kaminsky

Winter 2022/23


ncvoter1.txt - Microsoft Excel
Datei
Start
Einfügen
Seitenlayout
Formeln
Daten
Überprüfen
Ansicht
Add-Ins
Einfügen
Ausschneiden
Kopieren
Format übertragen
Zwischenablage
Calibri
11
Schriftart
Zeilenumbruch
Verbinden und zentrieren
Ausrichtung
Standard
Zahl
Bedingte Formatierung
Als Tabelle formatieren
Standard
Gut
Neutral
Schlecht
Ausgabe
Berechnung
Eingabe
Erklärender ...
Formatvorlagen
Einfügen
Löschen
Format
Zellen
AutoSumme
Füllbereich
Löschen
Sortieren und Filtern
Suchen und Auswählen
Bearbeiten
A1
county_id
county_id
county_desc
voter_reg_n
status_cd
voter_status_desc
reason_cd
voter_status
last_name
first_name
midl_name
name
res_street_addres
res_city_desc
state_zip_code
mail_addr1
mail_addr2
mail_city
mail_state
mail_zipcod
full_phone_r
race_code
ethnic_code
party_cd
1 ALAMANCE 9005990 A ACTIVE AV VERIFIED AABEL EVELYN LARSEN 4430 E GREENSBOF GRAHAM NC 27253 4430 E GREENSBORO-CHA GRAHAM NC 27253 000 0000 W NL UNA
1 ALAMANCE 9048723 A ACTIVE AV VERIFIED AARON CHRISTINA CASTAGNA 421 WHITT AVE BURLINGTON NC 27215 PO BOX 4177 BURLINGTON NC 27215 229 1110 W UN UNA
1 ALAMANCE 9019674 A ACTIVE AV VERIFIED AARON CLAUDIA HAYDEN 1013 EDITH ST BURLINGTON NC 27215 1013 EDITH ST BURLINGTON NC 27215 222 8834 W NL UNA
1 ALAMANCE 9129589 A ACTIVE AV VERIFIED AARON JAMES MICHAEL 1647 SAXAPAHAW GRAHAM NC 27253 PO BOX 98 SAXAPAHAW NC 27340 336 525 2484 W UN DEM
1 ALAMANCE 9041748 A ACTIVE AV VERIFIED AARON NATHAN EDWARD 421 WHITT AVE BURLINGTON NC 27215 PO BOX 4177 BURLINGTON NC 27215 336 229 1110 W UN UNA
1 ALAMANCE 9021947 A ACTIVE AV VERIFIED AARON WILLIE DALE 1013 EDITH ST BURLINGTON NC 27215 1013 EDITH ST BURLINGTON NC 27215 336 999 9999 W NL UNA
1 ALAMANCE 9062002 A ACTIVE AV VERIFIED AARONSON GENA HOLT 107 TERRYWOOD (HAW RIVER NC 27258 107 TERRYWOOD CT HAW RIVER NC 27258 336 578 9123 W NL REP
1 ALAMANCE 9096423 A ACTIVE AV VERIFIED AARONSON MICHAEL CHARLES 107 TERRYWOOD (HAW RIVER NC 27258 107 TERRYWOOD CT HAW RIVER NC 27258 336 266 7615 W NL UNA
1 ALAMANCE 9117940 I INACTIVE IU CONFIRMATI ABAD PRISCILLA MARIE 100 COLONNADE I ELON NC 27244 CAMPUS BOX 3008 ELON NC 27244 O HL UNA
1 ALAMANCE 9034127 I INACTIVE IU CONFIRMATI ABADIE COLLEEN MIASHEL 1097 IVEY RD #C GRAHAM NC 27253 1097 IVEY RD #C GRAHAM NC 27253 M HL REP
1 ALAMANCE 9121656 A ACTIVE AV VERIFIED ABADIE JACK EDWARD JR 612 SIDEVIEW ST GRAHAM NC 27253 612 SIDEVIEW ST GRAHAM NC 27253 336 212 8140 W NL UNA
1 ALAMANCE 9118154 I INACTIVE IU CONFIRMATI ABADIE MYRA HOLLIFIELD 612 SIDEVIEW ST GRAHAM NC 27253 617 MITCHELL ST BURLINGTON NC 27217 336 212 8140 W NL UNA
1 ALAMANCE 9131788 A ACTIVE AV VERIFIED ABBAS FALISA 707 SUMMIT RIDG MEBANE NC 27302 707 SUMMIT RIDGE RD #/ MEBANE NC 27302 919 568 9001 B UN DEM
1 ALAMANCE 9068460 A ACTIVE AV VERIFIED ABBAS RAFAT 514 WESTRIDGE DI BURLINGTON NC 27215 514 WESTRIDGE DR BURLINGTON NC 27215 A UN DEM
1 ALAMANCE 9049573 A ACTIVE AV VERIFIED ABBATECOLA RONALD JOSEPH JR 504 BROOKFIELD I GIBSONVILLE NC 27249 504 BROOKFIELD DR GIBSONVILLE NC 27249 336 449 9029 W UN UNA
1 ALAMANCE 9033877 A ACTIVE AV VERIFIED ABBATECOLA TRACY BOONE 504 BROOKFIELD I GIBSONVILLE NC 27249 504 BROOKFIELD DR GIBSONVILLE NC 27249 W NL DEM
1 ALAMANCE 9083557 I INACTIVE IU CONFIRMATI ABBETT DAWN LEANN 3900 JOHNS CREEK GIBSONVILLE NC 27249 3900 JOHNS CREEK DR GIBSONVILLE NC 27249 336 584 3319 W NL DEM
1 ALAMANCE 9027554 A ACTIVE AV VERIFIED ABBEY BRENT DAVID 3304 GOLDEN OAK GRAHAM NC 27253 3304 GOLDEN OAKS DR GRAHAM NC 27253 919 682 6873 W NL REP
1 ALAMANCE 9029477 A ACTIVE AV VERIFIED ABBEY DEMETRA AINSWORTH 3304 GOLDEN OAK GRAHAM NC 27253 3304 GOLDEN OAKS DR GRAHAM NC 27253 336 376 0673 W NL REP
1 ALAMANCE 9022529 I INACTIVE IU CONFIRMATI ABBEY DOROTHY ESTELLA 1029A QUAKENBU SNOW CAMP NC 27349 1029A QUAKENBUSH RD SNOW CAMP NC 27349 376 3663 W NL REP
1 ALAMANCE 9113186 A ACTIVE AV VERIFIED ABBOTT AMELIA BETH 2876 CALLOWAY D MEBANE NC 27302 2876 CALLOWAY DR MEBANE NC 27302 919 304 6161 W NL UNA
1 ALAMANCE 9087980 A ACTIVE AV VERIFIED ABBOTT ANGELA MORTON 2006 WINN CREEK HAW RIVER NC 27258 2006 WINN CREEK DR HAW RIVER NC 27258 336 261 3357 W NL DEM
1 ALAMANCE 9019273 A ACTIVE AV VERIFIED ABBOTT BRENDA CARMICHAEI 611 N THIRD ST MEBANE NC 27302 611 N THIRD ST MEBANE NC 27302 563 2654 W NL UNA
1 ALAMANCE 9102615 A ACTIVE AV VERIFIED ABBOTT BRIAN CHRISTOPHE 2006 WINN CREEK HAW RIVER NC 27258 2006 WINN CREEK DR HAW RIVER NC 27258 336 261 3357 W NL UNA
1 ALAMANCE 9079257 A ACTIVE AV VERIFIED ABBOTT BRUCE CLEATON 188 LAKE CAMMA BURLINGTON NC 27217 188 LAKE CAMMACK CT BURLINGTON NC 27217 336 214 2703 W NL REP
1 ALAMANCE 1389300 A ACTIVE AV VERIFIED ABBOTT CHERYL FAULKNER 188 LAKE CAMMA BURLINGTON NC 27217 188 LAKE CAMMACK CT BURLINGTON NC 27217 336 229 3027 W NL REP
1 ALAMANCE 9140392 A ACTIVE AV VERIFIED ABBOTT CHRISTOPHE BRANDON 309 BURLINGTON GIBSONVILLE NC 27249 309 BURLINGTON AVE GIBSONVILLE NC 27249 W NL UNA
1 ALAMANCE 9135711 A ACTIVE AV VERIFIED ABBOTT COURTNEY LOVE 309 BURLINGTON GIBSONVILLE NC 27249 309 BURLINGTON AVE GIBSONVILLE NC 27249 W NL UNA
1 ALAMANCE 9028439 A ACTIVE AV VERIFIED ABBOTT DWAYNE ROGER 2839 LADALE LN MEBANE NC 27302 2839 LADALE LN MEBANE NC 27302 563 3956 W NL UNA
1 ALAMANCE 9090420 A ACTIVE AV VERIFIED ABBOTT FRANK PATRICK 1202 JAMESTOWN ELON NC 27244 1202 JAMESTOWNE DR ELON NC 27244 336 227 4088 W UN UNA
1 ALAMANCE 9079222 A ACTIVE AV VERIFIED ABBOTT GLADYS MARIE MILES 614 TUCKER ST BURLINGTON NC 27215 614 TUCKER ST BURLINGTON NC 27215 336 570 1418 B NL DEM
1 ALAMANCE 9129722 A ACTIVE AV VERIFIED ABBOTT HAROLD GRANT 507 EVERETT ST # BURLINGTON NC 27215 507 EVERETT ST #320B BURLINGTON NC 27215 336 437 3638 W NL REP
1 ALAMANCE 9094352 A ACTIVE AV VERIFIED ABBOTT JESSICA NADINE 2876 CALLOWAY D MEBANE NC 27302 2876 CALLOWAY DR MEBANE NC 27302 919 304 4661 W NL UNA
1 ALAMANCE 9023803 A ACTIVE AV VERIFIED ABBOTT JOYCE HODGES 1934 TUCKER ST # BURLINGTON NC 27215 1934 TUCKER ST #A BURLINGTON NC 27215 336 227 4079 W NL DEM
1 ALAMANCE 9084794 R REMOVED RS MOVED FROI ABBOTT LATWOIA BEREA 201 STALEY HALL ELON NC 27244 CAMPUS BOX 3039 ELON NC 27244 B NL DEM
1 ALAMANCE 9020357 A ACTIVE AV VERIFIED ABBOTT LAWRENCE ELMER JR 110 OAKVIEW DR ELON NC 27244 110 OAKVIEW DR ELON NC 27244 336 563 4708 W NL UNA
1 ALAMANCE 9108338 A ACTIVE AV VERIFIED ABBOTT MARIA LYNETTE 614 TUCKER ST BURLINGTON NC 27215 614 TUCKER ST BURLINGTON NC 27215 336 570 1418 B NL DEM
1 ALAMANCE 9077192 A ACTIVE AV VERIFIED ABBOTT NANCY SKIDMORE 110 OAKVIEW DR ELON NC 27244 110 OAKVIEW DR ELON NC 27244 800 222 7566 W NL UNA
1 ALAMANCE 9035500 A ACTIVE AV VERIFIED ABBOTT PATTI BELVIN 1202 JAMESTOWN ELON NC 27244 1202 JAMESTOWNE DR ELON NC 27244 336 228 0571 W UN REP
1 ALAMANCE 9090949 R REMOVED RM REMOVED AF ABBOTT RACHEL MARA 103 DANIELEY CEN ELON NC 27244 CAMPUS BOX 3044 ELON NC 27244 336 278 4012 W NL REP
1 ALAMANCE 9135295 A ACTIVE AV VERIFIED ABBOTT SUSAN HANKS 2876 CALLOWAY D MEBANE NC 27302 2876 CALLOWAY DR MEBANE NC 27302 919 568 8056 W UN UNA
1 ALAMANCE 9113731 I INACTIVE IU CONFIRMATI ABBOTT TAYLOR RENEE 406 W LEBANON A ELON NC 27244 CAMPUS BOX 3077 ELON NC 27244 W UN REP
1 ALAMANCE 9120825 I INACTIVE IN CONFIRMATI ABBOTT TIFFANY MURIEL ARLE 144 W CRESCENT S GRAHAM NC 27253 144 W CRESCENT SQUARE GRAHAM NC 27253 336 233 0429 B NL DEM
1 ALAMANCE 9013866 I INACTIVE IN CONFIRMATI ABBOTT VIRGINIA SMITH 2820 BLANCHE DR BURLINGTON NC 27215 2820 BLANCHE DR BURLINGTON NC 27215 584 4663 W NL REP
1 ALAMANCE 9027717 A ACTIVE AV VERIFIED ABBOTT-LUN SHELBY LYNN 509 FERNWAY DR BURLINGTON NC 27217 509 FERNWAY DR BURLINGTON NC 27217 336 226 0087 B NL DEM
1 ALAMANCE 9108552 A ACTIVE AV VERIFIED ABDALLA KHALED ISMAIL 605 ISLEY PL #C BURLINGTON NC 27215 605 ISLEY PL #C BURLINGTON NC 27215 336 686 0506 W NL DEM
1 ALAMANCE 9128403 A ACTIVE AV VERIFIED ABDEL-MAGI LISA ANN 1841 DUNBAR PL BURLINGTON NC 27215 1841 DUNBAR PL BURLINGTON NC 27215 214 437 8955 W NL UNA
1 ALAMANCE 9117192 I INACTIVE IU CONFIRMATI ABDELKARIM AMNA ELHAG 1105 PROVIDENCE ELON NC 27244 1105 PROVIDENCE CT ELON NC 27244 M NL UNA
ncvoter1
Bereit
100 %

Introduction: Data Profiling

"Data profiling is the process of examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data" [0]

- Metadata is essential
  - For data integration, query optimization, data cleaning, …
- Metadata is often incomplete, outdated or completely unavailable
- → discover useful metadata, given just the (relational) dataset
- Our Focus: Unique Column Combinations (UCCs), Functional Dependencies (FDs) and Inclusion Dependencies (INDs)

Prof. Dr. Felix Naumann
Tobias Bleifuß
Youri Kaminsky

[0] https://en.wikipedia.org/wiki/Data_profiling

Introduction: Unique Column Combinations (UCCs)

ID	Student ID	First Name	Last Name	Birthday
1	457825	John	Doe	14.02.1966
2	127894	John	Doe	13.11.1990
3	487992	Jim	Smith	17.04.1993
…	…	…	…	…

Unique Column Combination R.X:

For a set of attributes X in Relation R, X is a UCC, if for all t ∈ R there exists no t' ∈ R, so that t[X] = t'[X].

Example UCCs:

- {ID}
- {Student ID}
- {First Name, Last Name, Birthday}

Prof. Dr. Felix Naumann
Tobias Bleifuß
Youri Kaminsky

Introduction: Functional Dependencies (FDs)

ID	Company	Street	Zip-Code	City
…	…	…	12307	Berlin
…	…	…	12307	Berlin
…	…	…	12359	Berlin
			23966	Wismar

ID	First Name	Last Name	Birthday
1	John	Doe	14.02.1966
2	John	Doe	13.11.1990
3	Jim	Smith	17.04.1993

Functional Dependency X → A:

For sets of attributes X and A, the FD X → A holds if:
$t_1[X] = t_2[X] \Rightarrow t_1[A] = t_2[A]$.

Example FDs:

- Zip-Code → City
- First Name, Last Name, Birthday → ID

Prof. Dr. Felix Naumann
Tobias Bleifuß
Youri Kaminsky

Introduction: Inclusion Dependencies (INDs)

Object	Density	Semi-major Axis
Mercury	…	…
Venus	…	…
Earth	…	…
Saturn	…	…

Name	Satellite of	Distance [km]
…	Saturn	…
…	Saturn	…
…	Earth	…

Inclusion Dependency X ⊆ Y:

For sets of attributes $R_1.X=\{X_1,...,X_n\}$ and $R_2.Y=\{Y_1,...,Y_n\}$ the IND $R_1.X \subseteq R_2.Y$ holds if $\Pi_X(R_1) \subseteq \Pi_Y(R_2)$

Example INDs:

- Satellite of ⊆ Object

Prof. Dr. Felix Naumann
Tobias Bleifuß
Youri Kaminsky

Data Profiling Algorithms: Basic Strategies

Automatic discovery algorithms

- Typically discover minimal FDs/UCCs and maximal INDs
- Search space is still exponentially large
- → Employ pruning strategies

Examples for pruning:

- If {A,B} is no UCC, then neither {A} nor {B} can be UCCs
- If AB ↛ C then we also know that A ↛ C and B ↛ C

Data Profiling: Using Sampling

Motivation

- Runtime: even with state of the art algorithms, the discovery of dependencies can be infeasible for very large datasets
- Privacy: for data protection reasons, we may not access all data
- Cost: to collect or purchase all data might be too expensive
- Data quality: our database may be incomplete

Problem Statement

- Discover dependencies on a subset / sample

Research questions

- How close can we get to the true dependencies?
- What properties must our subset / sample have, so that we get reliable results?
- What is the tradeoff between sample size and correctness of results?

Approximate Data Profiling: Related Work

State of the art dependency discovery algorithms on full data sets:

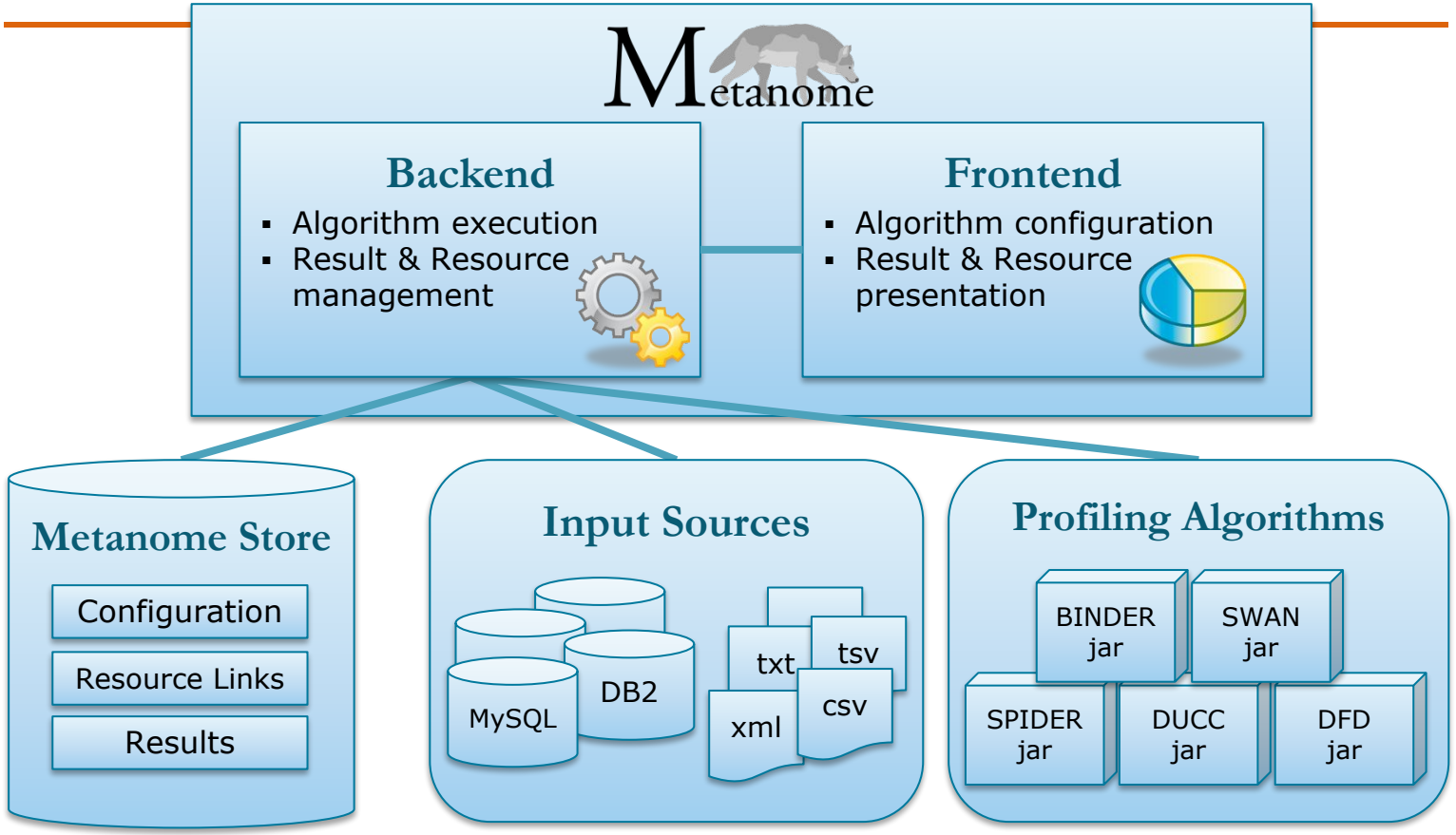
- Functional Dependencies: HyFD [1]
- Inclusion Dependencies: BINDER [2]
- Unique Column Combinations: HyUCC [3]

Sampling/Sketching has already been applied to reduce runtime:

- AIDFD [4]
- FAIDA [5]

[1] Papenbrock, Thorsten, and Felix Naumann. "A hybrid approach to functional dependency discovery." Proceedings of the 2016 International Conference on Management of Data. 2016.
[2] Papenbrock, Thorsten, et al. "Divide & conquer-based inclusion dependency discovery." Proceedings of the VLDB Endowment 8.7 (2015): 774-785.
[3] Papenbrock, Thorsten, and Felix Naumann. "A hybrid approach for efficient unique column combination discovery." Datenbanksysteme für Business, Technologie und Web (BTW 2017) (2017).
[4] Bleifuß, Tobias, et al. "Approximate discovery of functional dependencies for large datasets." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016.
[5] Kruse, Sebastian, et al. "Fast approximate discovery of inclusion dependencies." Datenbanksysteme für Business, Technologie und Web (BTW 2017) (2017).

Metanome


Metanome
Backend
Algorithm execution
Result & Resource management
Frontend
Algorithm configuration
Result & Resource presentation
Metanome Store
Configuration
Resource Links
Results
Input Sources
MySQL
DB2
txt
tsv
xml
csv
Profiling Algorithms
BINDER jar
SWAN jar
SPIDER jar
DUCC jar
DFD jar

Metanome

http://www.metanome.de

https://owncloud.hpi.de/s/WSmgz18Z6yceG7h

Metanome Resources

Repositories:

- Metanome https://github.com/HPI-Information-Systems/metanome
- Algorithms https://github.com/HPI-Information-Systems/metanome-algorithms
- Metanome-CLI https://github.com/sekruse/metanome-cli

Datasets:

- UCC/FD discovery: https://hpi.de/naumann/projects/repeatability/data-profiling/fds.html#c168191
- IND discovery: https://hpi.de/naumann/projects/repeatability/data-profiling/metanome-ind-algorithms.html

Goals and grading

Goals

- Learn about the research area data profiling
- Read papers and understand them
- Craft a novel solution to the problem of sample-based profiling
- Run experiments and evaluate results
- Present results in written and oral form

Grading

- Approach (35%)
- Written report (35%)
- Presentations and discussions in the seminar (30%)

Requirements

You **bring**:

- Java (at least basic skills or willing to learn)

You want to **learn** about:

- Data Profiling
- Algorithmic problems and how to solve them
- Experimental evaluation on large datasets

You do **not** need to bring:

- Prior knowledge about Data Profiling

You should **agree** to:

- Publish the results as open-source

Seminar Roadmap (preliminary)

Week of Oct 24
- Technical introduction
- Introduce Metanome and algorithms repository

Week of Oct 31
- Present related work (1 previously assigned paper)
- Report initial execution of related work (modify input data)
- Potentially: Initial ideas for a sampling-based variant

Week of Nov 7
- Discuss metrics for evaluation
- Discuss initial ideas / general approaches

Week of Nov 14
- Implement evaluation metrics and evaluate related work on sample

Around Christmas
- Milestone: Technical Evaluation
- Present preliminary results

Beginning of February
- Final presentation (chair is invited)

10.02.2023
- Deadline paper-style submission (max 12 pages)
- If preferred, we can postpone the deadline until after the exam periods

Every week we offer to meet for status updates, questions and feedback

Register for the seminar

Structure:

- Work in teams of **two** students

Register:

- Email to tobias.bleifuss@hpi.de until **October 25**
- Use subject: **Approximate Data Profiling seminar registration**
- Optionally, include topics that are interesting for you
- Optionally, include team partner (if you already have one)
- Optionally, include another time slot that fits your schedule better