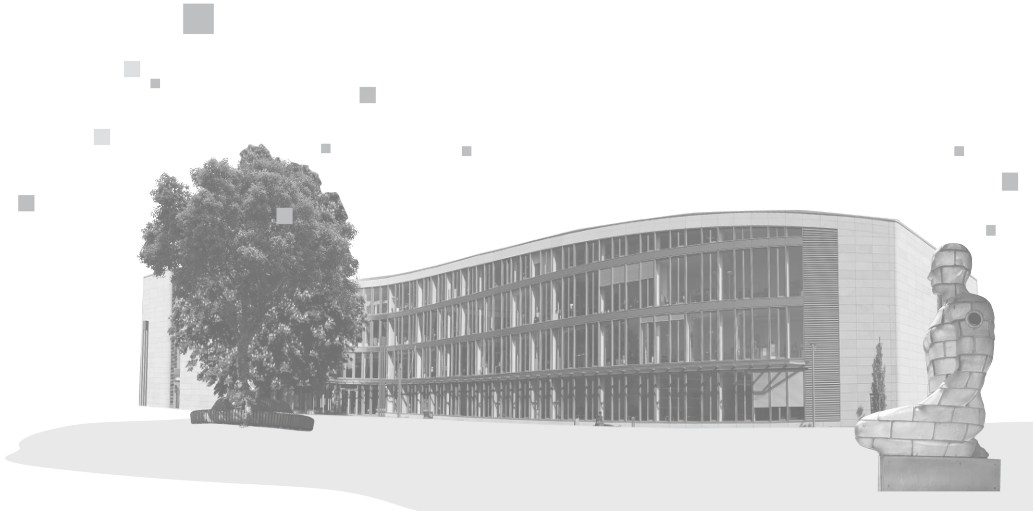
Advanced data profiling:
Classifying genuineness

Youri Kaminsky, Lukas Laskowski, Prof. Dr. Felix Naumann

Hasso Plattner Institut
Digital Engineering · Universität Potsdam



Design IT.
Create Knowledge.

Code of Conduct - Overview

At DEF/HPI, we are committed to providing a high-quality learning as well as research environment and building a community where students and staff can thrive scientifically and personally. Everyone should expect a safe, supportive, and inclusive environment in all our spaces.

Our Code of Conduct helps us meet this goal. Words or actions that are disrespectful, racist, discriminatory, hostile, or harassing are not acceptable.

Examples of these include:

- Offensive comments about others' ethnicity, accent, religion, nationality, gender, sexual orientation, or other personal traits
- Refusing to work with someone based on these personal traits
- Physical or verbal threats and assaults
- Using sexualized or vulgar language or actions
- Disrupting another person's work experience

Code of Conduct - Help and Support

Violations of this code are taken seriously. If you witness or experience any inappropriate behavior, report it to a lecturer or any DEF/HPI contact point. All reports will be handled confidentially and with care.

Please be aware of further contact points and support structures at DEF/HPI, including:

- Equal Opportunities Officers (Charlotte Weiss, Florence Böttger)
- Diversity Manager (Dr. Imke Leicht)
- Ombudsman for good scientific practice (Prof. Tilmann Rabl)
- Student Trusted Advisors (Zero, Ronja, Paul, Florina, Shirin and Anna)
- Psychological counseling hotline (0800 7777015)
- Incident Response System (safecampus.hpi.de)
- as well as the respective offers of the University of Potsdam (Mental Health Counseling Service, Psychosocial Counseling of Studentenwerk, Nightline)
  www.uni-potsdam.de/en/discrimination-free-university/consulting-and-support/overview-of-consueling-and-advising-services

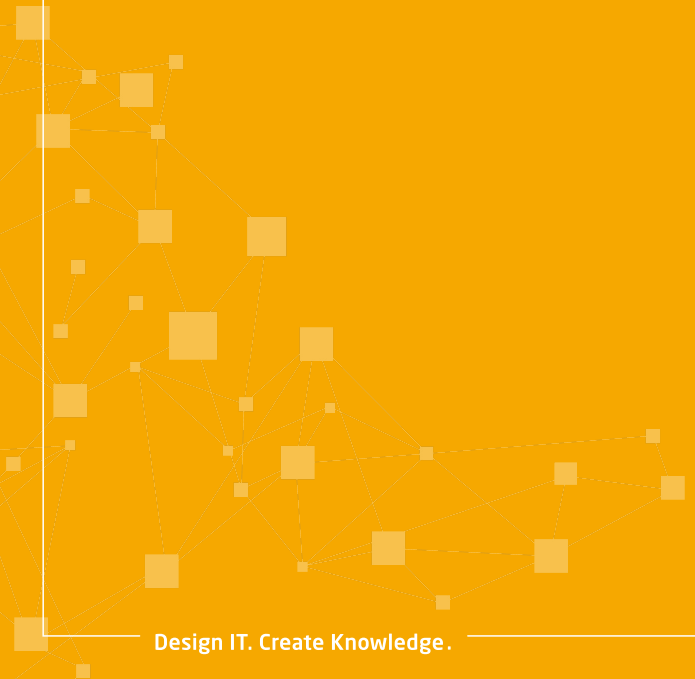
Team Introduction

Youri
4th year PhD student
Data Profiling

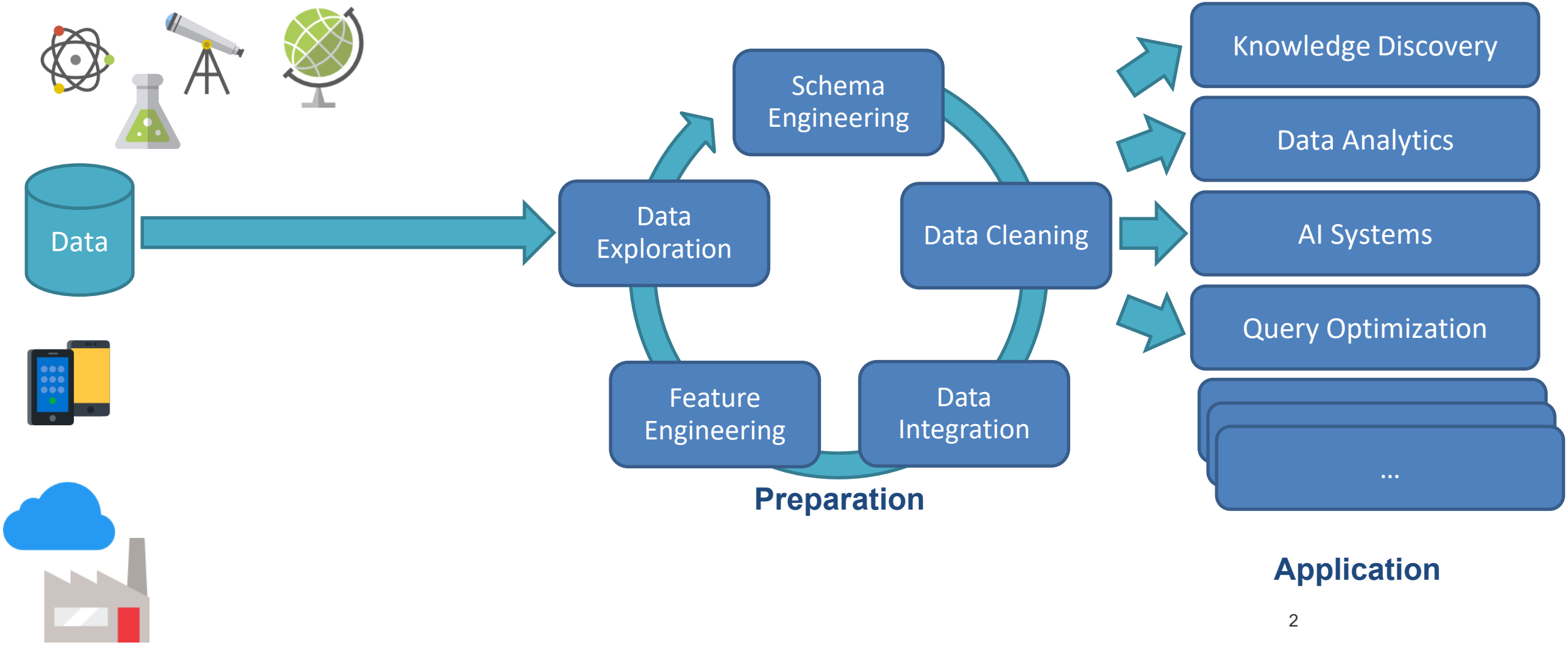
Lukas
3rd year PhD student
Artificial Intelligence

Prof. Felix Naumann
Head of Information
Systems Group

Data Profiling Introduction

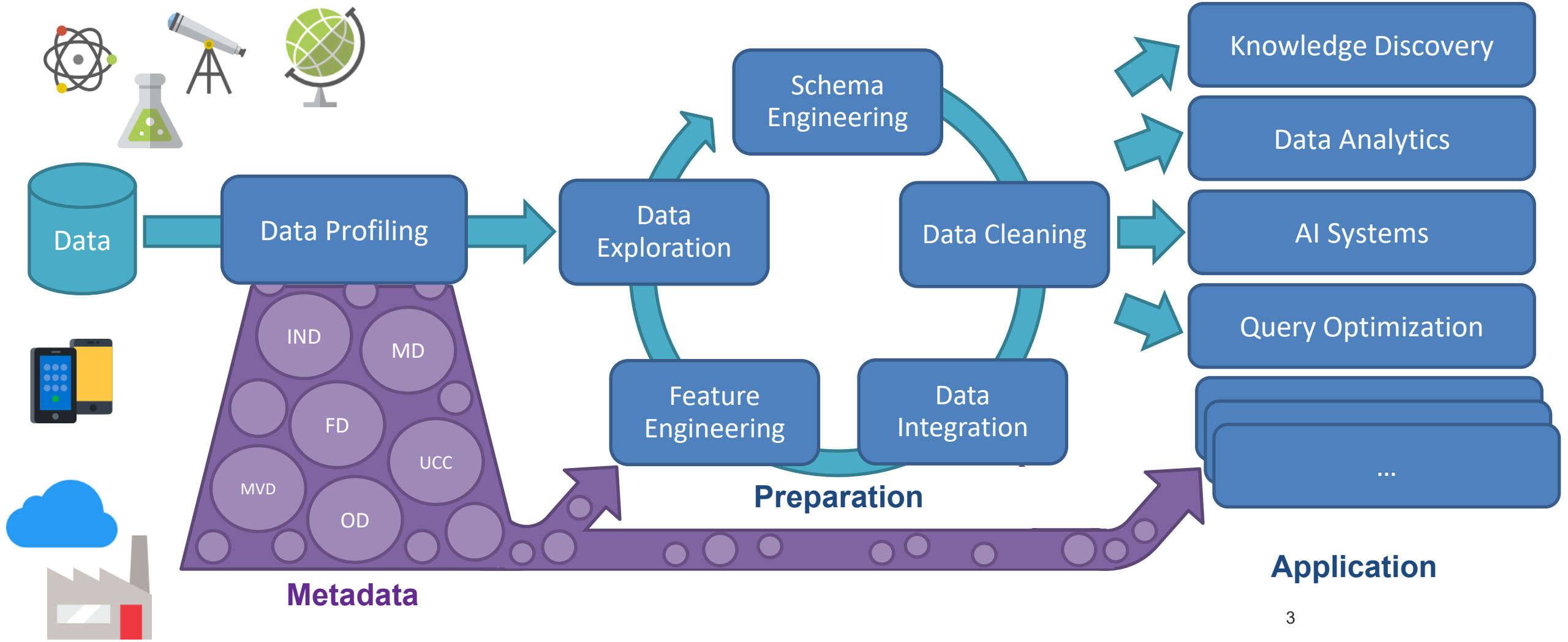


Data Profiling for Data Engineering



Data Profiling for Data Engineering

See here for
Open Research Questions



Use Case: Query Optimization

Open Research Question

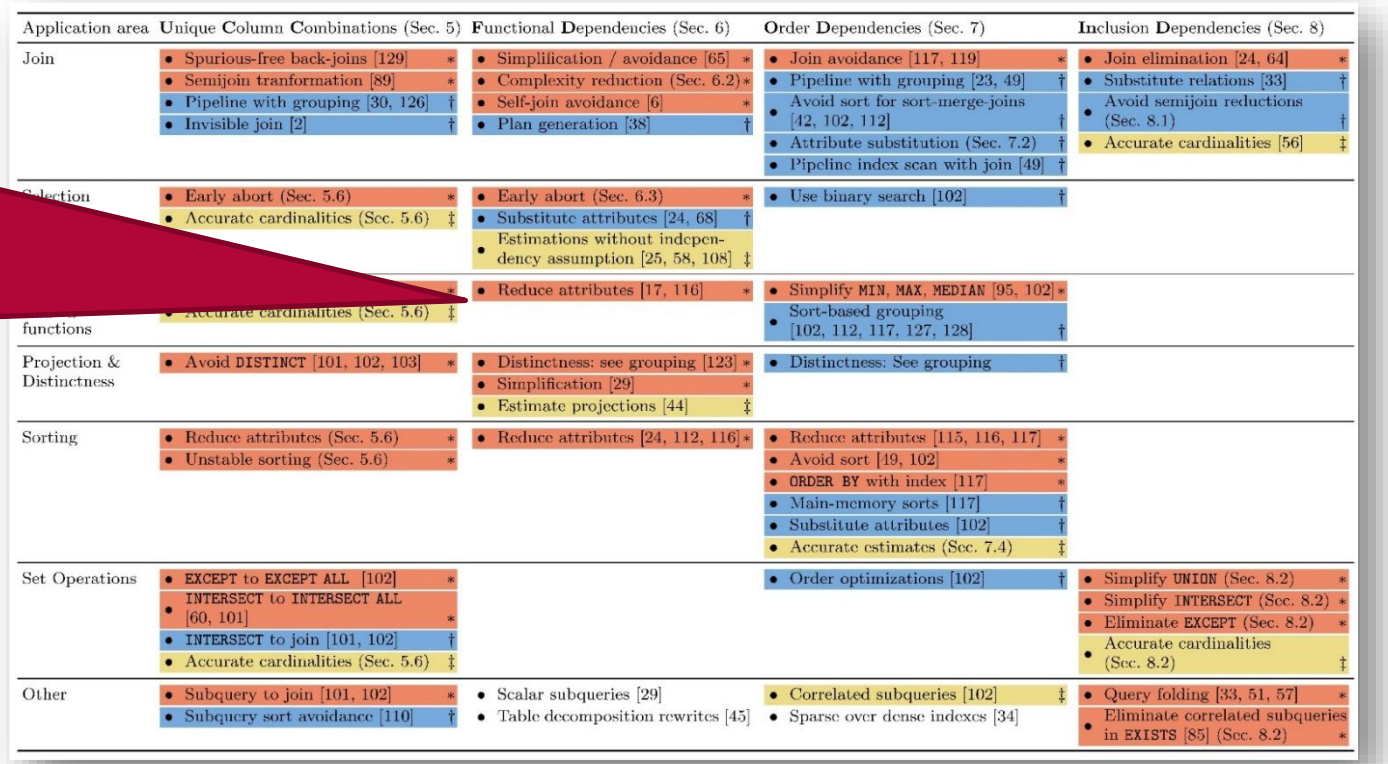
How to introduce dependency-based optimization to DBMS?

- 58 optimization opportunities
  - Using unique column combinations (UCCs), functional dependencies (FDs), order dependencies (ODs), and inclusion dependencies (INDs)

```
SELECT A, B,
AVG(C)
FROM R
GROUP BY A, B

A → B

SELECT A, B,
AVG(C)
FROM R
GROUP BY A
```



Application area	Unique Column Combinations (Sec. 5)	Functional Dependencies (Sec. 6)	Order Dependencies (Sec. 7)	Inclusion Dependencies (Sec. 8)
Join	• Spurious-free back-joins [129] *	• Simplification / avoidance [65] *	• Join avoidance [117, 119] *	• Join elimination [24, 64] *
	• Semijoin tranformation [89] *	• Complexity reduction (Sec. 6.2) *	• Pipeline with grouping [23, 49] †	• Substitute relations [33] †
	• Pipeline with grouping [30, 126] †	• Self-join avoidance [6] *	• Avoid sort for sort-merge-joins [42, 102, 112] †	• Avoid semijoin reductions (Sec. 8.1) †
	• Invisible join [2] †	• Plan generation [38] †	• Attribute substitution (Sec. 7.2) †	• Accurate cardinalities [56] ‡
			• Pipeline index scan with join [49] †	
Selection	• Early abort (Sec. 5.6) *	• Early abort (Sec. 6.3) *	• Use binary search [102] †	
	• Accurate cardinalities (Sec. 5.6) ‡	• Substitute attributes [24, 68] †		
		• Estimations without independency assumption [25, 58, 108] ‡		
functions	*	• Reduce attributes [17, 116] *	• Simplify MIN, MAX, MEDIAN [95, 102] *	
	• Accurate cardinalities (Sec. 5.6) ‡		• Sort-based grouping [102, 112, 117, 127, 128] †	
Projection & Distinctness	• Avoid DISTINCT [101, 102, 103] *	• Distinctness: see grouping [123] *	• Distinctness: See grouping †	
		• Simplification [29] *		
		• Estimate projections [44] ‡		
Sorting	• Reduce attributes (Sec. 5.6) *	• Reduce attributes [24, 112, 116] *	• Reduce attributes [115, 116, 117] *	
	• Unstable sorting (Sec. 5.6) *		• Avoid sort [49, 102] *	
			• ORDER BY with index [117] *	
			• Main-memory sorts [117] †	
			• Substitute attributes [102] †	
			• Accurate estimates (Sec. 7.4) ‡	
Set Operations	• EXCEPT to EXCEPT ALL [102] *		• Order optimizations [102] †	• Simplify UNION (Sec. 8.2) *
	• INTERSECT to INTERSECT ALL [60, 101] *			• Simplify INTERSECT (Sec. 8.2) *
	• INTERSECT to join [101, 102] †			• Eliminate EXCEPT (Sec. 8.2) *
	• Accurate cardinalities (Sec. 5.6) ‡			• Accurate cardinalities (Sec. 8.2) ‡
Other	• Subquery to join [101, 102] *	• Scalar subqueries [29]	• Correlated subqueries [102] ‡	• Query folding [33, 51, 57] *
	• Subquery sort avoidance [110] †	• Table decomposition rewrites [45]	• Sparse over dense indexes [34]	• Eliminate correlated subqueries in EXISTS [85] (Sec. 8.2) *

Use Case: Data Cleansing

1. Discover approximate/relaxed dependencies
2. Verify their genuineness
3. Detect violating records/values
4. Correct the values

Functional dependency violation

Denial constraint violation

Name	ID	LVL	ZIP	ST	SAL
Alice	ID1	5	10001	NM	90k
Bob	ID2	6	87101	NM	80k
Chris	ID3	4	10001	NY	80k

Chu, Xu, Ihab F. Ilyas, and Paolo Papotti. "Holistic data cleaning: Putting violations into context." *ICDE*'13.

Classifying Data Profiling Tasks

> *Data profiling refers to the activity of creating small but informative summaries of a database.*
> Ted Johnson, Encyclopedia of Database Systems

Single-column tasks

- Cardinalities
- Uniqueness
  - Key discovery
- Patterns and data types
- Distributions
- Domain Classification
- …

Multi-column tasks

- Uniqueness (UCCs)
  - Key discovery
- Inclusion dependencies (INDs)
  - Foreign key discovery
- Functional dependencies (FDs)
- Order dependencies (ODs)
- Denial constraints (DCs)
- …

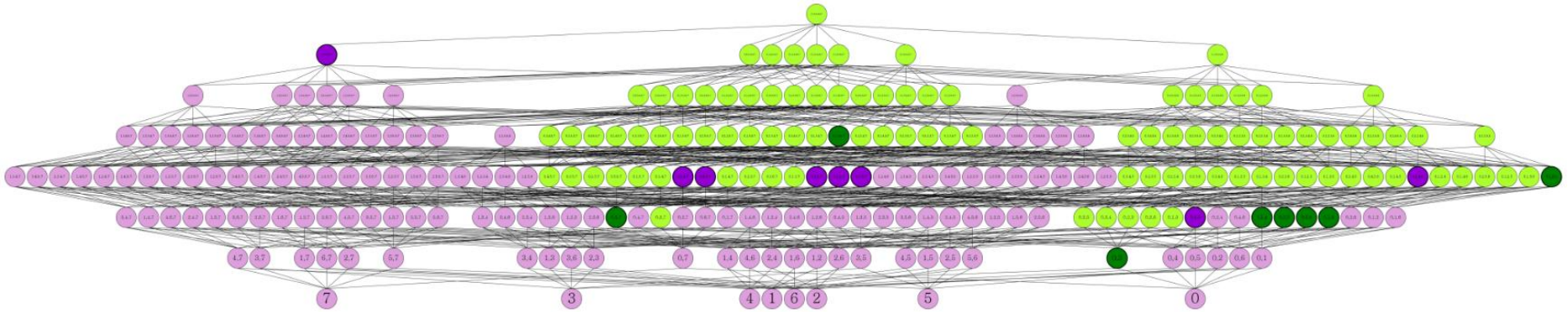
Scalable Profiling

Open Research Question

Can we predict problem and solution size for given data?

- Scalability in number of rows

- Scalability in number of columns
  - "Normal" table with 100 columns:
    $2^{100} - 1 = 1{,}267{,}650{,}600{,}228{,}229{,}401{,}496{,}703{,}205{,}375$
    = 1.3 nonillion column combinations (the power set)



- Large solution space: e.g. exponential number of FDs

Many Results, e.g. for Functional Dependencies

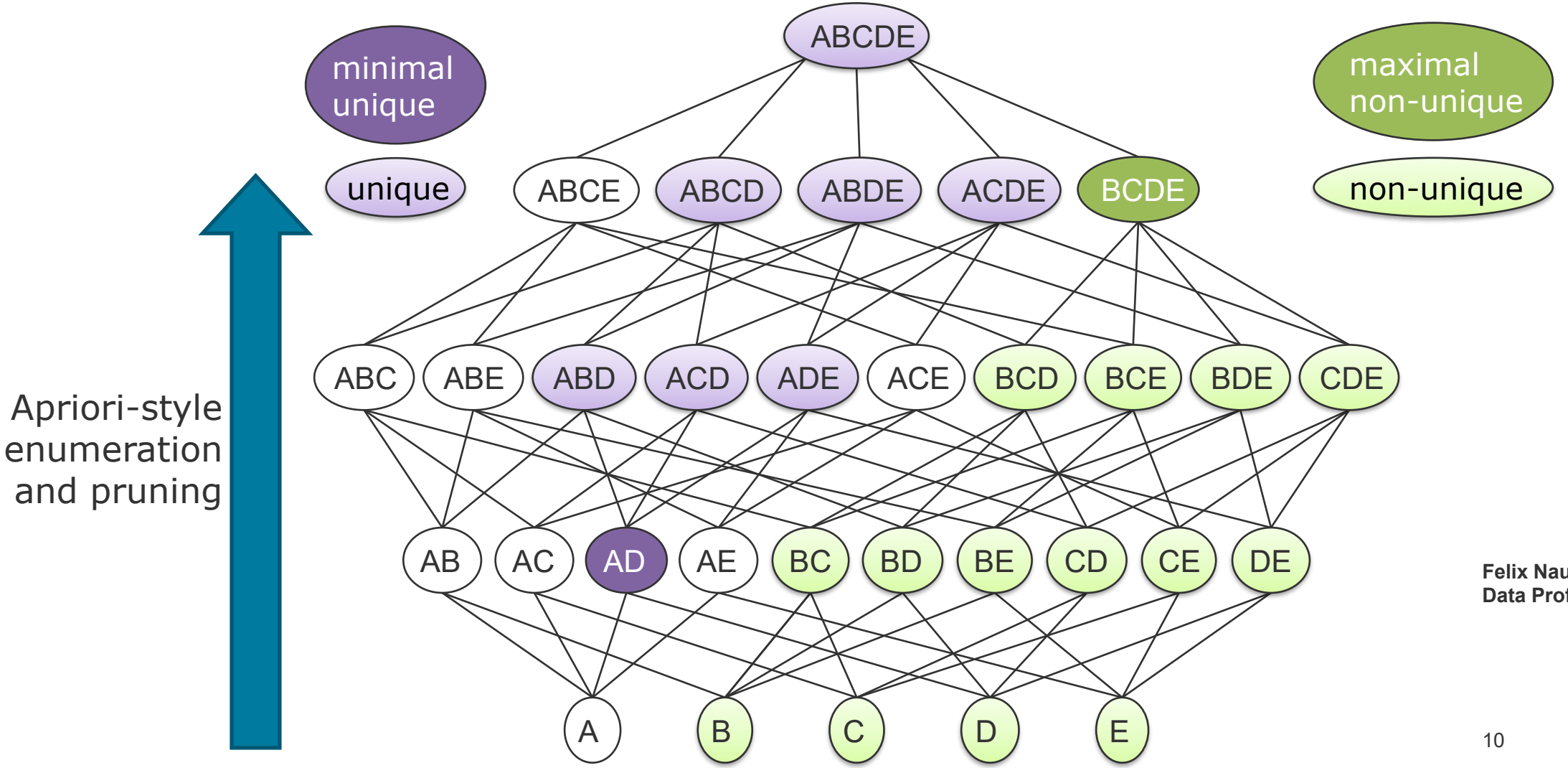
Dataset	Cols [#]	Rows [#]	Size [KB]	FDs [#]	TANE [12]	FUN [18]	FD_MINE [25]	DFD [1]	DEP-MINER [16]	FASTFDS [24]	FDEP [9]	HYFD
iris	5	150	5	4	1.1	**0.1**	0.2	0.2	0.2	0.2	**0.1**	**0.1**
balance-scale	5	625	7	1	1.2	**0.1**	0.2	0.3	0.3	0.3	0.2	**0.1**
chess	7	28,056	519	1	2.9	1.1	3.8	1.0	174.6	164.2	125.5	**0.2**
abalone	9	4,177	187	137	2.1	0.6	1.8	1.1	3.0	2.9	3.8	**0.2**
nursery	9	12,960	1,024	1	4.1	1.8	7.1	0.9	121.2	118.9	46.8	**0.5**
breast-cancer	11	699	20	46	2.3	0.6	2.2	0.8	1.1	1.1	0.5	**0.2**
bridges	13	108	6	142	2.2	0.6	4.2	0.9	0.5	0.6	0.2	**0.1**
echocardiogram	13	132	6	527	1.6	0.4	69.9	1.2	0.5	0.5	0.2	**0.1**
adult	14	48,842	3,528	78	67.4	111.6	531.5	5.9	6039.2	6033.8	860.2	**1.1**
letter	17	20,000	695	61	260.0	529.0	7204.8	6.0	1090.0	1015.5	291.3	**3.4**
ncvoter	19	1,000	151	758	4.3	4.0	ML	5.1	11.4	1.9	1.1	**0.4**
hepatitis	20	155	8	8,250	12.2	175.9	ML	326.7	5576.5	9.5	0.8	**0.6**
horse	27	368	25	128,727	457.0	TL	ML	TL	TL	385.8	7.2	**7.1**
fd-reduced-30	30	250,000	69,581	89,571	41.1	77.7	ML	TL	377.2	382.4	TL	513.0
plista	63	1,000	568	178,152	ML	ML	ML	TL	TL	TL	26.9	**21.8**
flight	109	1,000	575	982,631	ML	ML	ML	TL	TL	TL	216.5	**53.4**
uniprot	223	1,000	2,439	>2,437,556	ML	ML	ML	TL	TL	TL	ML	**>5254.7**

Results larger than 1,000 FDs are only counted **TL**: time limit of 4 hours exceeded **ML**: memory limit of 100 GB exceeded

Uniques and Non-uniques (in North Carolina Voter Data)

- **Unique column combination (UCC)**
  No pair of records has same value combination when projected to those columns
- **A minimal unique**: `voter_reg_num, zip_code, race_code`
- **A maximal non-unique**: `voter_reg_num, status_cd, voter_status_desc, reason_cd, voter_status_reason_desc, absent_ind, name_prefx_cd, name_sufx_cd, half_code, street_dir, street_type_cd, street_sufx_cd, unit_designator, unit_num, state_cd, mail_addr2, mail_addr3, mail_addr4, mail_state, area_cd, phone_num, full_phone_number, drivers_lic, race_code, race_desc, ethnic_code, ethnic_desc, party_cd, party_desc, sex_code, sex, birth_place, precinct_abbrv, precinct_desc, municipality_abbrv, municipality_desc, ward_abbrv, ward_desc, cong_dist_abbrv, cong_dist_desc, super_court_abbrv, super_court_desc, judic_dist_abbrv, judic_dist_desc, nc_senate_abbrv, nc_senate_desc, nc_house_abbrv, nc_house_desc, county_commiss_abbrv, county_commiss_desc, township_abbrv, township_desc, school_dist_abbrv, school_dist_desc, fire_dist_abbrv, fire_dist_desc, water_dist_abbrv, water_dist_desc, sewer_dist_abbrv, sewer_dist_desc, sanit_dist_abbrv, sanit_dist_desc, rescue_dist_abbrv, rescue_dist_desc, munic_dist_abbrv, munic_dist_desc, dist_1_abbrv, dist_1_desc, dist_2_abbrv, dist_2_desc, confidential_ind, age, vtd_abbrv, vtd_desc`

Pruning Subsets



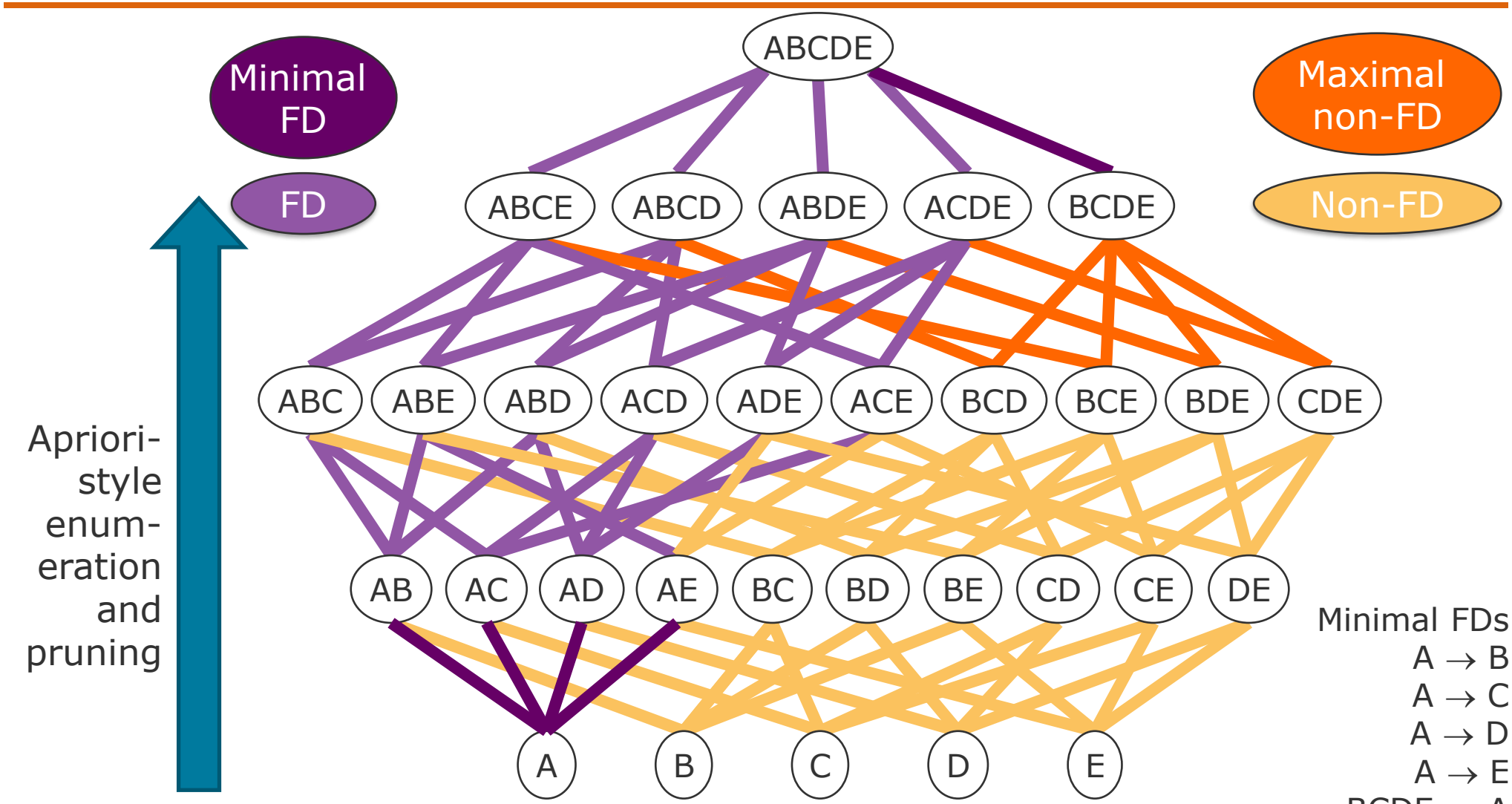
Discovering Functional Dependencies

- „X → A“
  When two tuples have same value in attribute set X,
  they must have same values in attribute A.
  - E.g., ZIP → City

Naïve discovery approach

```
For each column combination X
  For each A∈X
    For each pair of tuples (t1,t2)
      If t1[X\A] = t2[X\A] and t1[A] ≠ t2[A]: Break
```

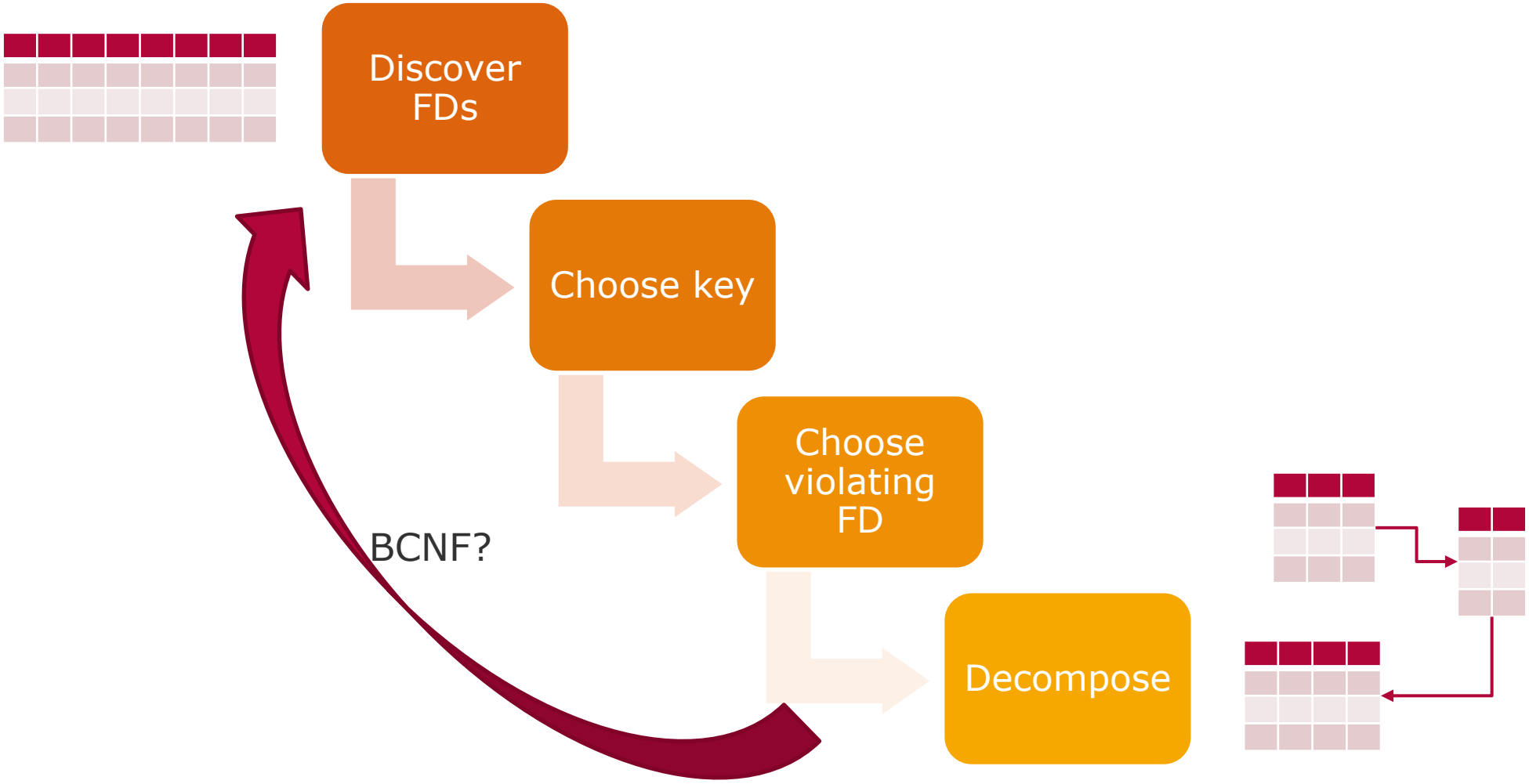
Model in Lattice – Edges Represent FDs



Minimal FDs

$A \rightarrow B$
$A \rightarrow C$
$A \rightarrow D$
$A \rightarrow E$
$BCDE \rightarrow A$

Use case: BCNF Normalization



[EDBT'17]

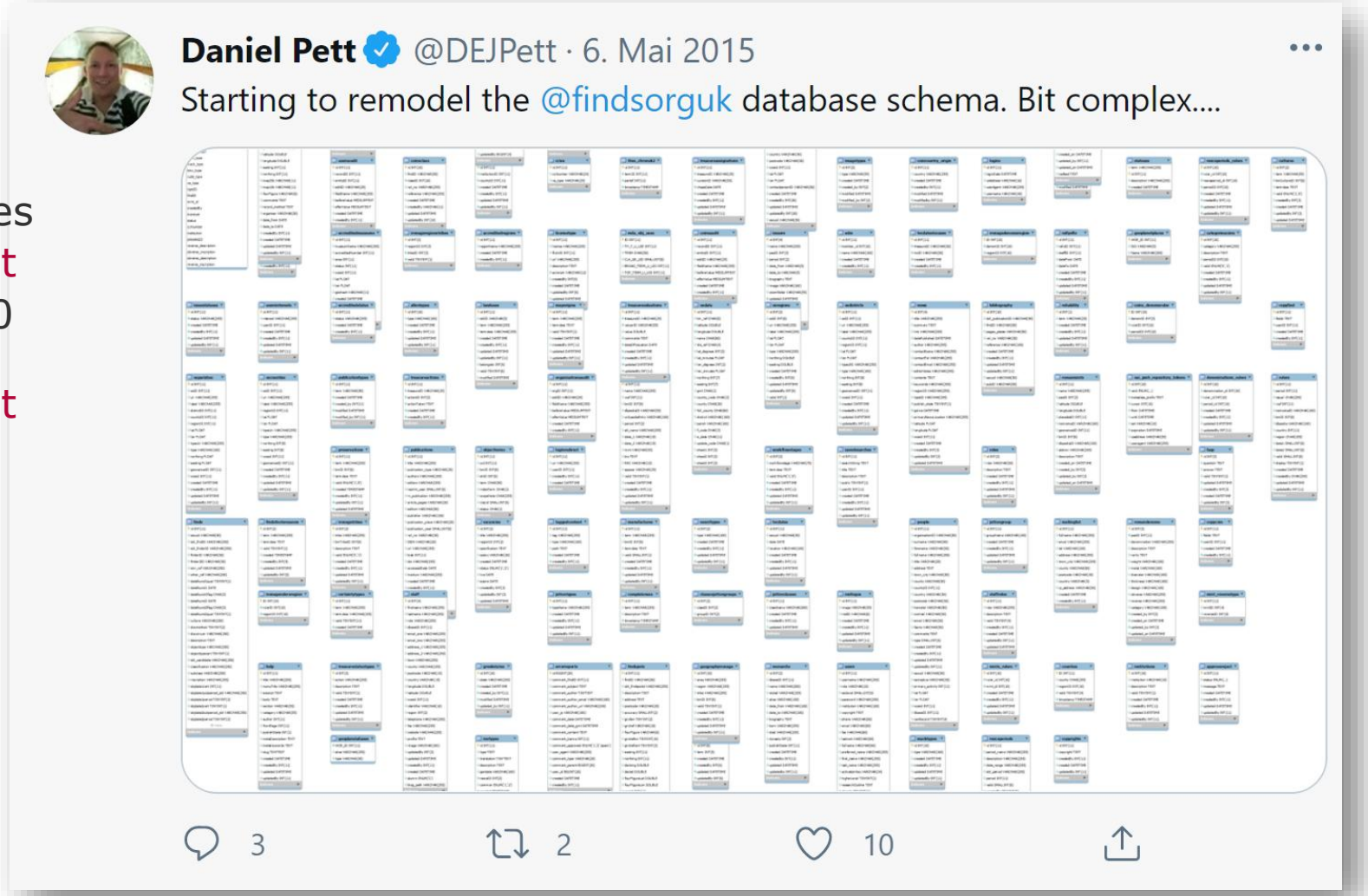
Normalization Results: TPC-H

Inclusion Dependencies for Foreign Key Discovery

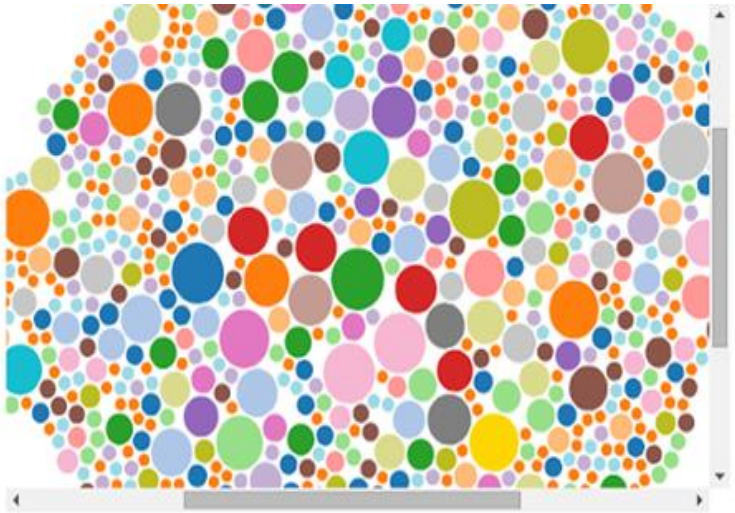
- Unary and n-ary INDs
  $R[A] \subseteq S[B]$ and $R[ABC] \subseteq S[DEF]$

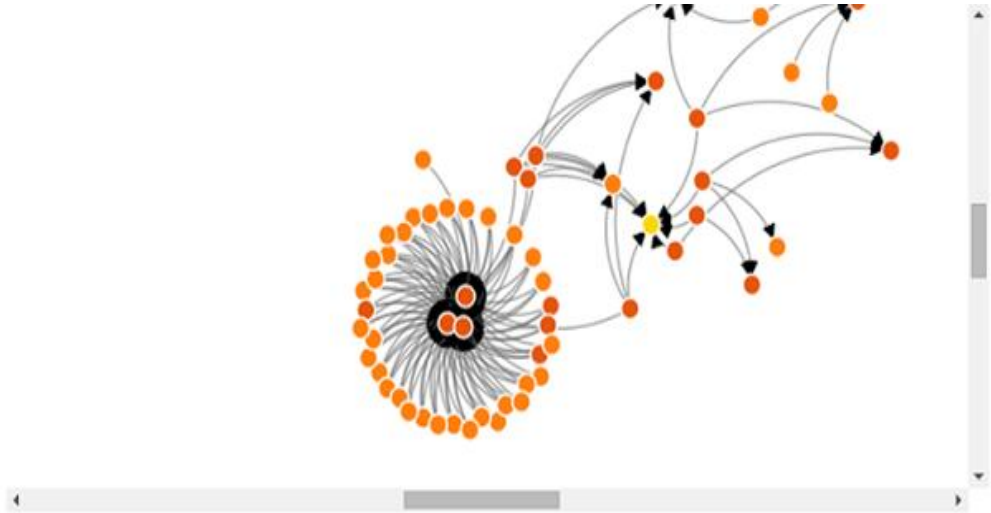
- Use cases
  - PDB – Protein Data Bank: 175 tables
    - Not a single foreign key constraint
  - Ensembl – genome database: >200 tables
    - Not a single foreign key constraint
  - Web tables:
    - No schema, no constraints, but many connections

- Why are FKs missing?
  - Lack of database knowledge
  - Lack of FK-support in DBMS
  - Fear of performance drop
  - Independent origin



Use Case: Web Data Integration
Discovering INDs among millions of web tables





96242-1	Association'.csv
43666-3	43666-3.'BBC_Radio_Stoke'.'Programming'.csv
53064-1	53064-1.'Rotation_period'.'Rotation period of selected objects'.csv
562884-4	562884-4.'Planets_in_astrology'.'Ruling planets of the astrological signs and houses'.csv
175797-1	175797-1.'Sun_sign_astrology'.'Sun signs'.csv
177750-2	177750-2.'BBC_Radio_Manchester'.'Programming'.csv
89462-4	89462-4.'Astrology_and_the_classical_elements'.'Triplicities by season'.csv
213213-1	213213-1.'Dalton_Park'.'Opening times'.csv
470402-	

Celestial Objects	Rotation period	Rotation period
Sun	25.379995 days (equatorial) 35 days (high latitude)	25 d 9 h 7 m 11.6 s 35 d
Mercury	58.6462 days	58 d 15 h 30 m 30 s
Venus	?243.0187 days	?243 d 0 h 26 m
Earth	0.99726968 days	0 d 23 h 56 m 4.100 s
Moon	27.321661 days (synchronous toward Earth)	27 d 7 h 43 m 11.5 s
Mars	1.02595675 days	1 d 0 h 37 m 22.663 s
Ceres	0.37809 days	0 d 9 h 4 m 27.0 s
Jupiter	0.4135344 days (deep interior) 0.41007 days (equatorial) 0.41369942 days (high latitude)	0 d 9 h 55 m 29.37 s 0 d 9 h 50 m 30 s 0 d 9 h 55 m 43.63 s
Saturn	0.44403 days (deep interior) 0.426 days (equatorial) 0.443 days (high	0 d 10 h 39 m 24 s 0 d 10 h 14 m 0 d 10 h 38 m

Zoom (1-5)

Range (logarithmic)

Dataset

allFilters

[TODS'17]

But aren't these just arbitrary observations on one instance?
Genuine Dependencies

Open Research Question

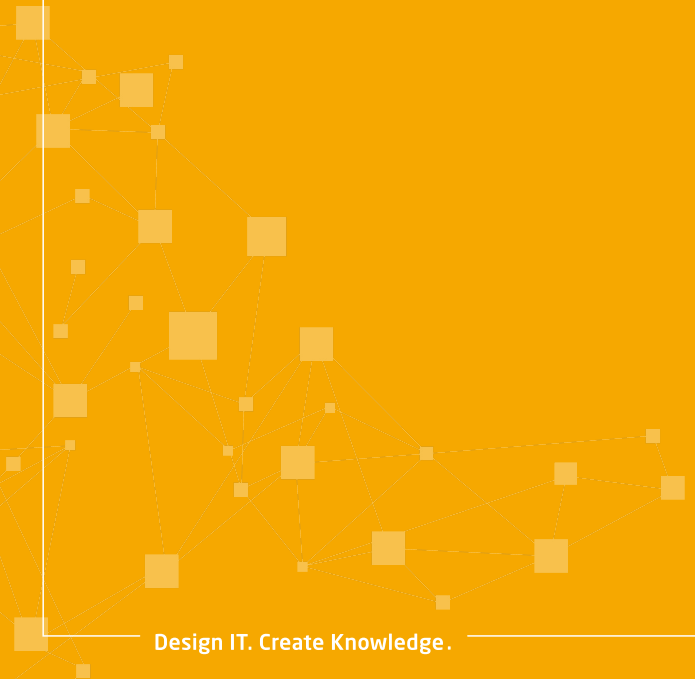
Can we effectively identify genuine dependencies?

- Features for UCCs as keys
  - „ID", „PK", etc. in name
  - Few columns, short data types
  - Early in schema
  - Serves as reference
- Features for INDs as foreign keys
  - „FK", „ID", etc. in name
  - Referenced columns are a UCC or key
  - Random or even distribution of values
  - …
- Temporal data
  - Dependency should have held at every point in time.

…	Sensor_status	Temperature
	1	23.343455
	1	23.454676
	0	24.001135
	1	24.173099
	…	…

Cust_ID	Cust_status	…
0	2	
1	0	
2	1	
3	2	
…	…	

Example in Metanome



Papers

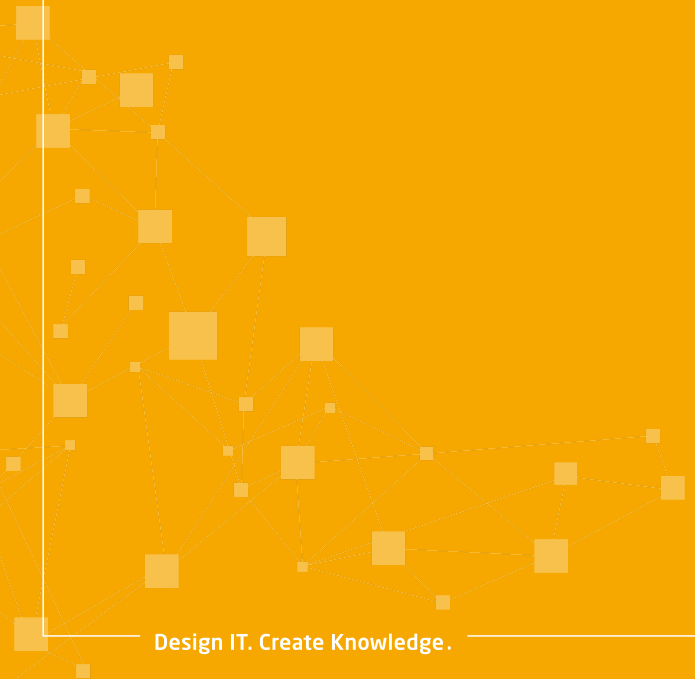
Paper overview

- Unique Column Combinations / Keys
  - Leon Bornemann et al. 2020. **Natural Key Discovery in Wikipedia Tables**.
    In Proceedings of The Web Conference (WWW). https://doi.org/10.1145/3366423.3380039
  - Henning Koehler et al. 2025. **Orthogonal Keys High Precision and Recall for Mining Database Keys From Inconsistent and Incomplete Relations**. In IEEE Transactions on Knowledge & Data Engineering (TKDE). https://doi.org/10.1109/TKDE.2025.3608680
- Inclusion Dependencies / Foreign Keys
  - Lan Jiang et al. 2020. **Holistic primary key and foreign key detection**.
    In the Journal of Intelligent Information Systems. https://doi.org/10.1007/s10844-019-00562-z
  - Alexandra Rostin et al. 2009. **A Machine Learning Approach to Foreign Key Discovery**.
    In International Workshop on the Web and Databases (WebDB).
    https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2009_rostin_a.pdf
  - Meihui Zhang et al. 2010. **On multi-column foreign key discovery**.
    In Proceedings of the VLDB Endowment (PVLDB) https://doi.org/10.14778/1920841.1920944

Paper overview

- Functional Dependencies
  - Jyrki Kivinen et al. 1995. **Approximate Inference of Functional Dependencies from Relations**.
  - In Proceedings of the International Conference on Database Theory (ICDT).
    https://doi.org/10.1016/0304-3975(95)00028-U
  - Panagiotis Mandros et al. 2017. **Discovering Reliable Approximate Functional Dependencies**.
    In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)
    https://doi.org/10.1145/3097983.3098062
  - Thorsten Papenbrock et al. 2017. **Data-driven Schema Normalization**. (Section 7)
    In Proceedings of the International Conference on Extending Database Technology (EDBT)
    https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2017_papenbrock_datadriven.pdf
- Denial Constraints
  - Albert Martin et al. 2025. **How and Why False Denial Constraints are Discovered**.
    In Proceedings of the VLDB Endowment (PVLDB) https://doi.org/10.14778/3748191.3748209

Organizational Introduction



Goals of the seminar

- Learning goals:
  - Get to know the research area of data profiling and machine learning
  - Read and write scientific papers
  - Conducting scientific experiments and presenting their results
- Team activities:
  - Identifying a research question given a topic and relevant papers
  - Crafting and evaluating a novel solution to answer the research question
  - Writing a report
- Deliverables
  - Paper-style report
  - Code, models, and datasets produced
  - Midterm and final presentations
  - Paper presentation

Prior Experience

1. Data Profiling
2. Data Profiling / Data Integration Course @ HPI?
3. Machine Learning
4. Machine Learning Course @ HPI?

Requirements

- Prior knowledge in data profiling
  - preferably having completed the Data Profiling or Data Integration lecture
- Prior experience with machine learning or deep learning
  - preferably completed some related course at HPI
- Good programming skills in a major programming language
  - support for Python, Java, C++
  - focus of AI tools on Python

Grading

In the seminar, each team will develop an approach and write a short report. The final grade consists of the following:

- Paper presentation (pass or fail)
- Quality of approach (35%)
- Written report (25%)
- Midterm presentation (10%)
- Final presentation (30%)

You can withdraw from the seminar without consequences until 27th of October.

Schedule - Overview

- “How to read a paper” session jointly with master's project:

  21st of October, 15:15 - 16:45

- Teaching end: 6th of February
- Midterm presentation: 11th of December
- Final presentation: TBD (presumably 9th to 13th of February)
- Final deadline report: TBD (preferences?)
- Regular meetings in this slot: Tuesdays 15:15 - 16:45
  - Is there a better slot?
  - Individual or group meetings?

Schedule - Next steps

- 14th of October: Introduction
- 21st of October: How to read a paper
- 28th of October: Paper presentation and research idea discussion
  ⇒ withdraw from course until 27th of October
- 4th of November: HPI Teaching Day ⇒ no session
  Submit project proposal
- 11th of November: Finalization of research ideas and kick-off project phase