Data Profiling – A look back and a look forward

EDBT 2021

Felix Naumann

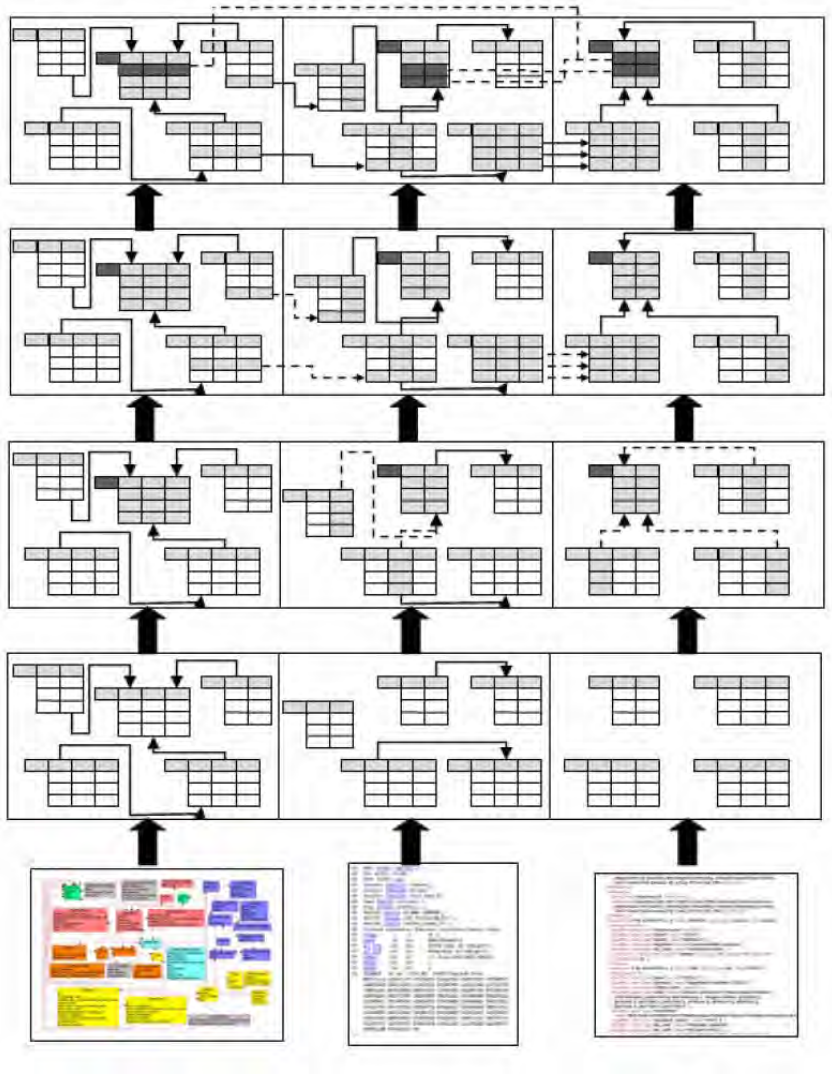
(Almost) Hands-Off Information Integration for the Life Sciences

Entity Resolution

Inclusion dependency discovery across sources

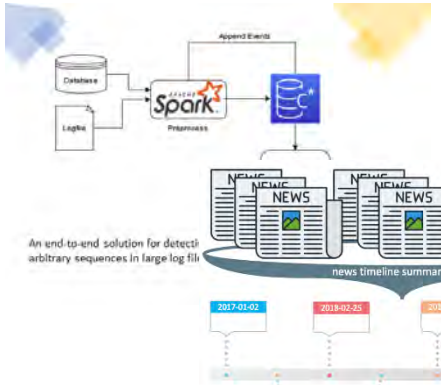
Key and inclusion dependency discovery within sources

Parsing

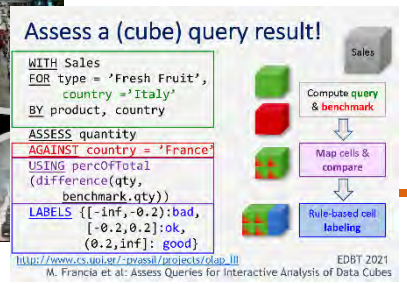


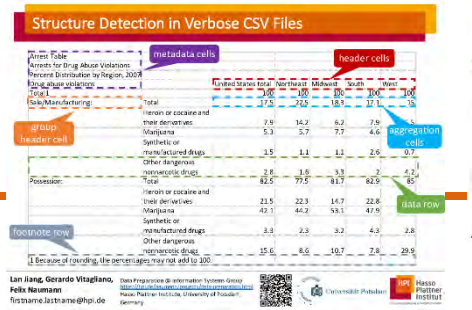
Felix Naumann
Data Profiling
EDBT 2021

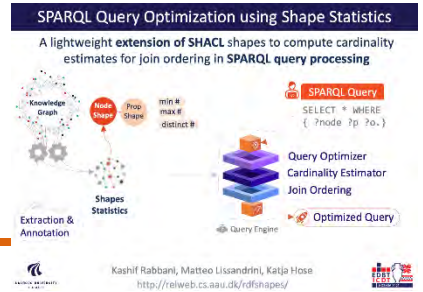
[CIDR'05]

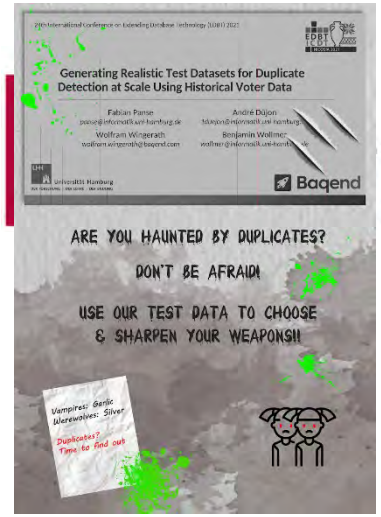

Append Events
Database
Logfile
Spark
Preprocess
NEWS
news timeline summarization
2017-01-02
2018-02-25
2018-03-08
2018-06-03

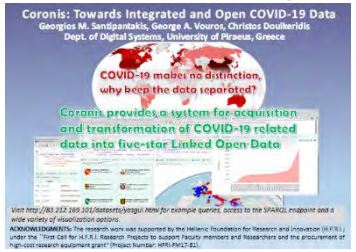

Querying
Top-k Dominant Traffic Flows
on Large Urban Road Networks
Help traffic optimization techniques
by identifying the dominant flows!


Assess a (cube) query result!
WITH Sales
FOR type = 'Fresh Fruit',
country ='Italy'
BY product, country
ASSESS quantity
AGAINST country = 'France'
USING percOfTotal
(difference(qty,
benchmark.qty))
LABELS {[-inf,-0.2):bad,
[-0.2,0.2]:ok,
(0.2,inf]: good}
Sales
Compute query & benchmark
Map cells & compare
Rule-based cell labeling
http://www.cs.uoi.gr/~pvassil/projects/olap_ll
EDBT 2021
M. Francia et al: Assess Queries for Interactive Analysis of Data Cubes

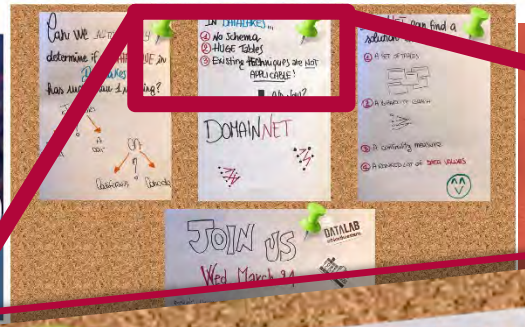

Structure Detection in Verbose CSV Files
metadata cells
header cells
group header cell
aggregation cells
data row
footnote row


SPARQL Query Optimization using Shape Statistics
A lightweight extension of SHACL shapes to compute cardinality estimates for join ordering in SPARQL query processing
SPARQL Query
Extraction & Annotation
Shapes Statistics
Query Optimizer
Cardinality Estimator
Join Ordering
Optimized Query
Query Engine
Kashif Rabbani, Matteo Lissandrini, Katja Hose
http://reiweb.cs.aau.dk/rdfshapes/


Generating Realistic Test Datasets for Duplicate Detection at Scale Using Historical Voter Data
Baqend
ARE YOU HAUNTED BY DUPLICATES?
DON'T BE AFRAID!
USE OUR TEST DATA TO CHOOSE & SHARPEN YOUR WEAPONS!!
Vampires: Garlic
Werewolves: Silver
Duplicates? Time to find out


Coronis: Towards Integrated and Open COVID-19 Data
Georgios M. Santipantakis, George A. Vouros, Christos Doulkeridis
Dept. of Digital Systems, University of Piraeus, Greece
COVID-19 makes no distinction, why keep the data separated?
Coronis provides a system for acquisition and transformation of COVID-19 related data into five-star Linked Open Data


INTERNATIONAL CONFERENCE
EDBT 2021
HOW TO FIND THE TRUTH

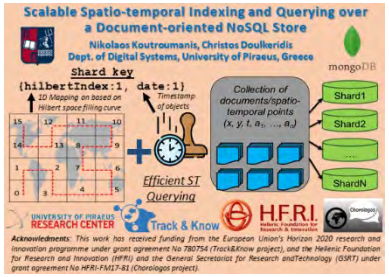

DOMAINNET
JOIN US
Wed March 24

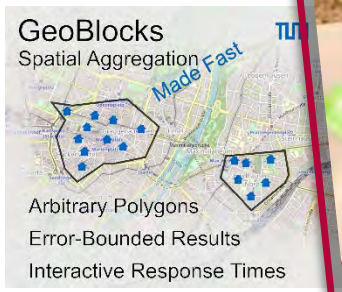

NextiaJD
Based on data profiles & learning models
Scalability powered by Spark!


SotA
Not adapted to generated heterog

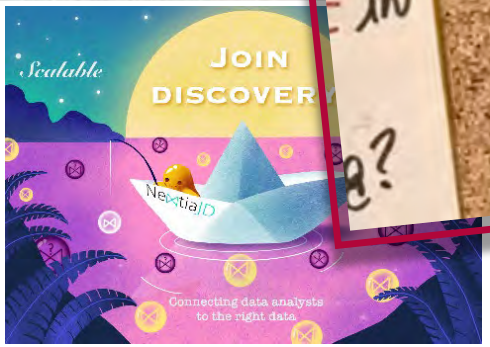
Efficient Exploratory Clustering Analyses with Qualitative Approximations
Manuel Fritz, Dennis Tschechlov, Holger Schwarz
Acceptable Quality
High Quality
High Runtime
Short Runtime
ops of up to 669x
ty is met in most cases

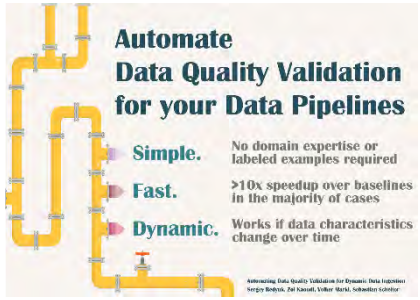

Data scientists and analysts
Shared data
Sharing redundant work
in Microsoft's Cosmos big data platform


Scalable Spatio-temporal Indexing and Querying over a Document-oriented NoSQL Store
Nikolaos Koutroumanis, Christos Doulkeridis
Dept. of Digital Systems, University of Piraeus, Greece
Shard key
{hilbertIndex:1, date:1}
mongoDB
Shard1
Shard2
ShardN
Efficient ST Querying
UNIVERSITY OF PIRAEUS RESEARCH CENTER
Track & Know
HFRI


GeoBlocks
Spatial Aggregation
Made Fast
Arbitrary Polygons
Error-Bounded Results
Interactive Response Times

IN DATALAKES...
1 No Schema
2 HUGE Tables
3 Existing Techniques are NOT APPLICABLE!
AID NOW?

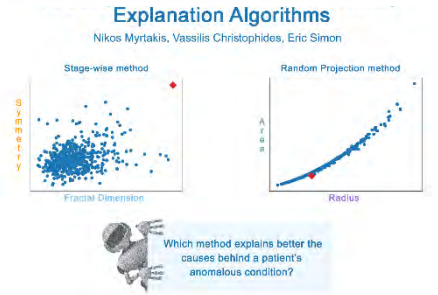

Scalable
JOIN
DISCOVER
NextiaJD
Connecting data analysts to the right data


Automate
Data Quality Validation
for your Data Pipelines
Simple.
No domain expertise or labeled examples required
Fast.
>10x speedup over baselines in the majority of cases
Dynamic.
Works if data characteristics change over time

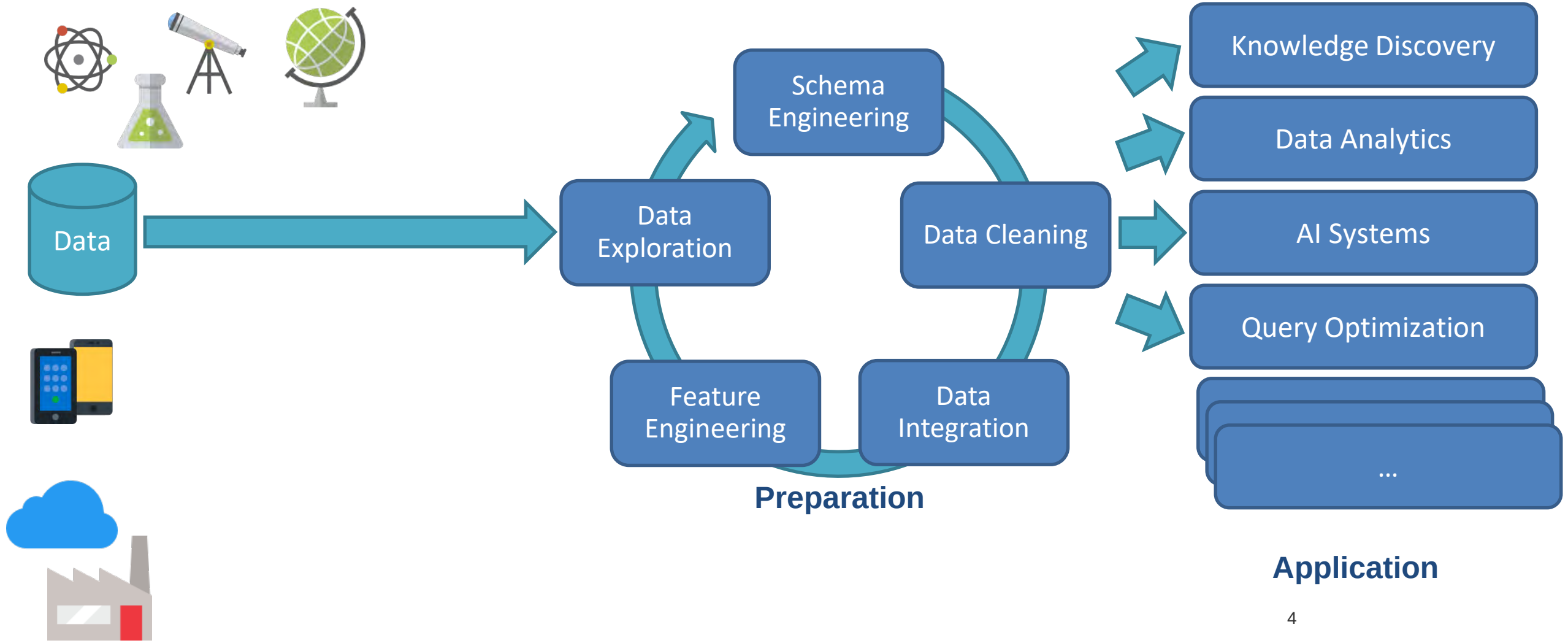
Biased data may lead to unfair classification.
We restore fairness thru data preprocessing.

KNOWING WHEN A RELATIONAL DATABASE NEEDS TUNING: IDENTIFYING PERFORMANCE THROTTLES
TAKING DB TUNERS TO PRODUCTION LANDSCAPE


Need to automatically configure your database?
Have many configuration knobs?
Your workloads are changing?
Our smart agent will take care of it!
OLTP W/O
OLTP R/W
OLAP
Paper #264: Adaptive Multi-Model Reinforcement Learning for Online Database Tuning
Yahiv Gur, Dongsheng Yang, Frederik Stalschus, Berthold Reinwald

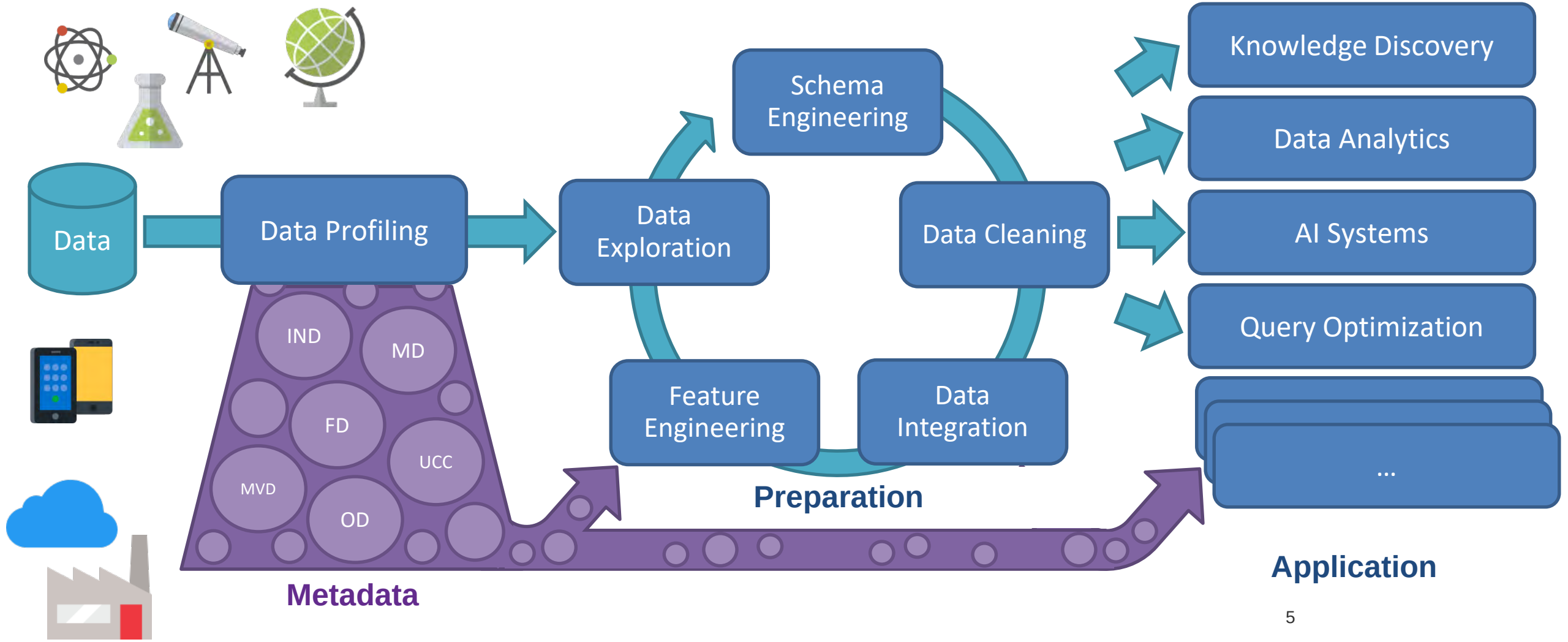

Explanation Algorithms
Nikos Myrtakis, Vassilis Christophides, Eric Simon
Stage-wise method
Random Projection method
Symmetry
Fractal Dimension
Area
Radius
Which method explains better the causes behind a patient's anomalous condition?

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

Felix Naumann
Data Profiling
EDBT 2021

Use Case: Data Cleansing

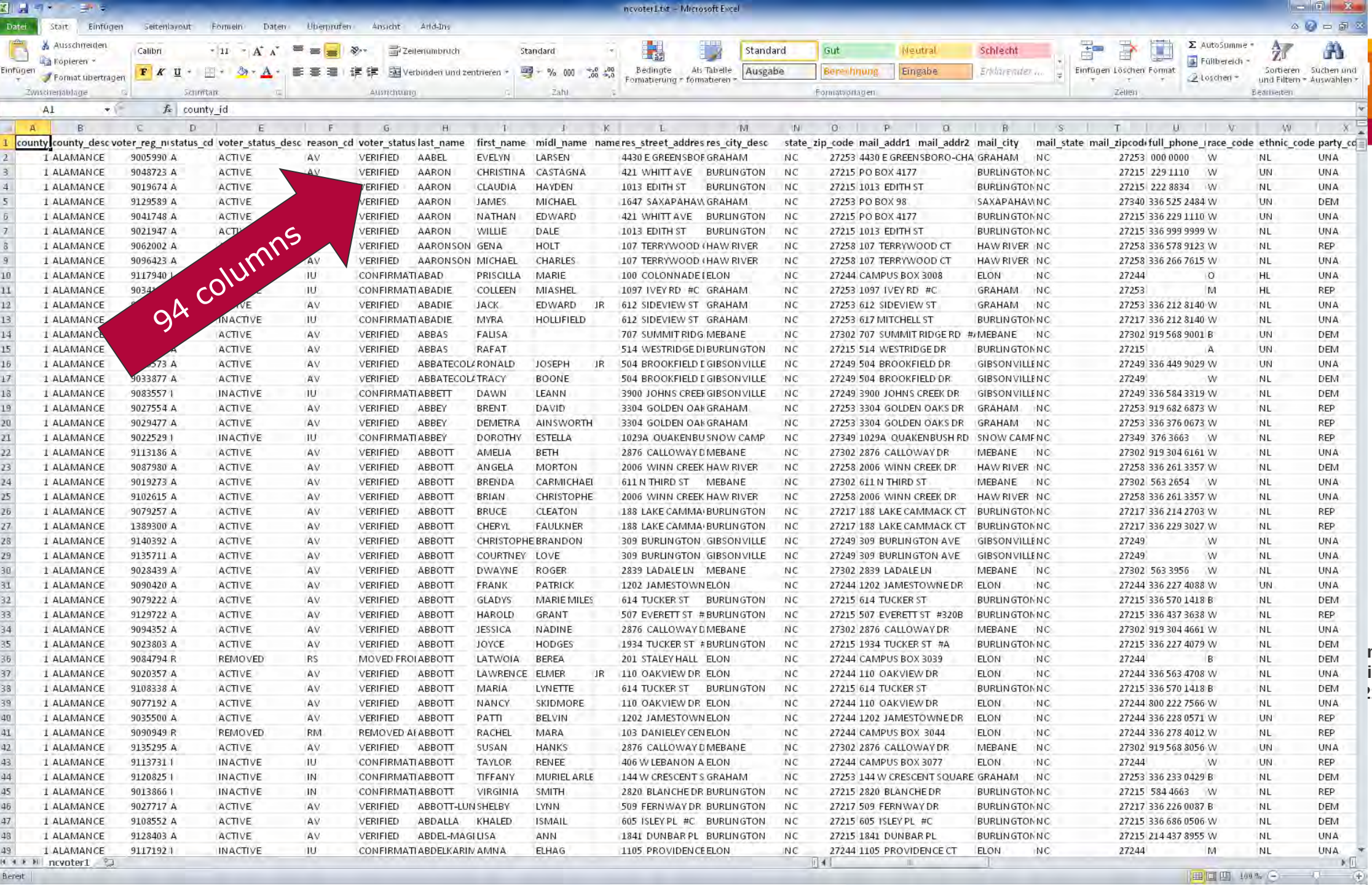
1. Discover approximate/relaxed dependencies
2. Verify their genuineness
3. Detect violating records/values
4. Correct the values

Functional dependency violation

Denial constraint violation

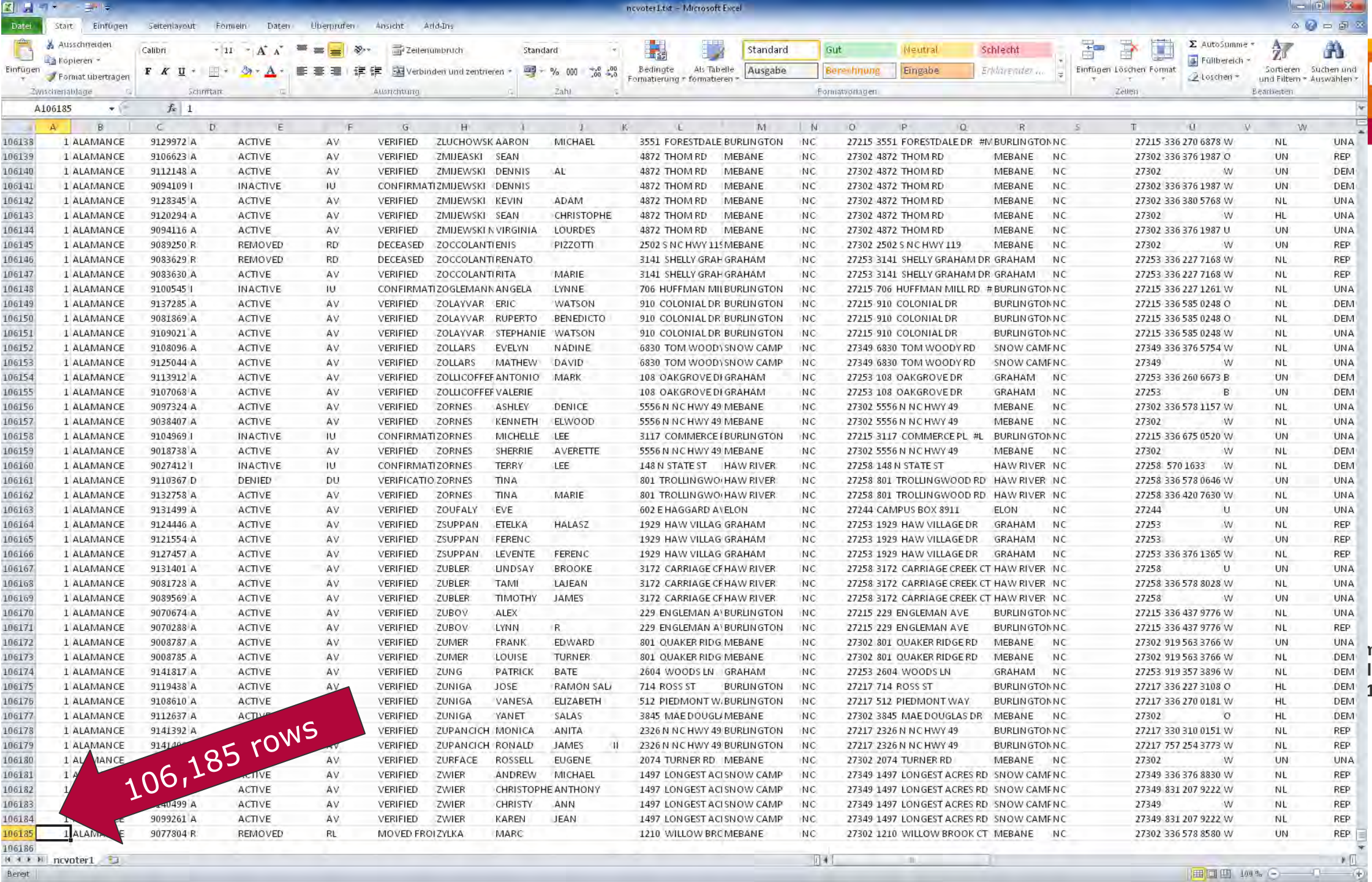
Name	ID	LVL	ZIP	ST	SAL
Alice	ID1	5	10001	NM	90k
Bob	ID2	6	87101	NM	80k
Chris	ID3	4	10001	NY	80k

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



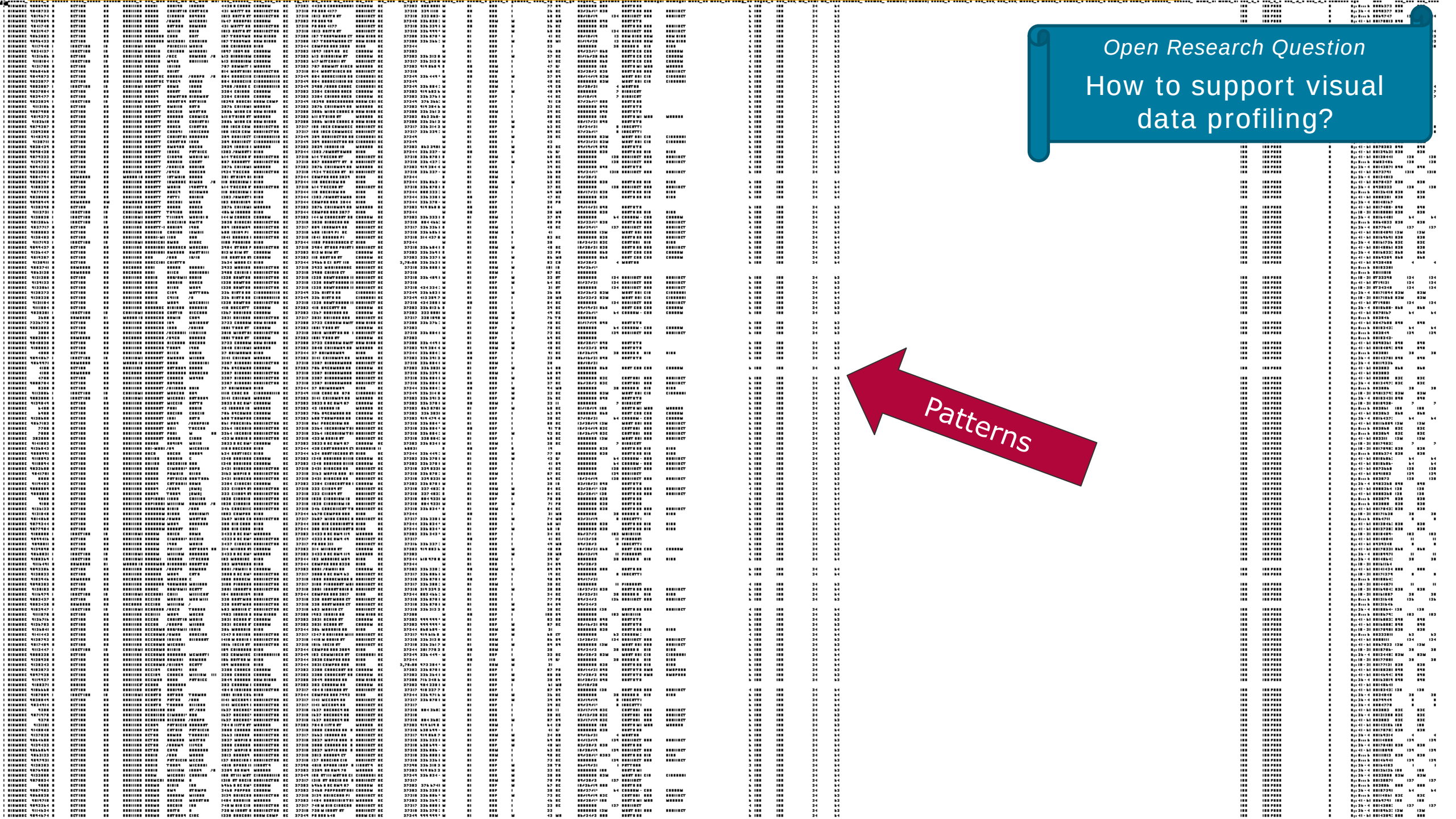
Hasso Plattner Institut

mann
iling
21



Hasso Plattner Institut

nann
ling
1


Open Research Question
How to support visual data profiling?
Patterns

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)
- ...

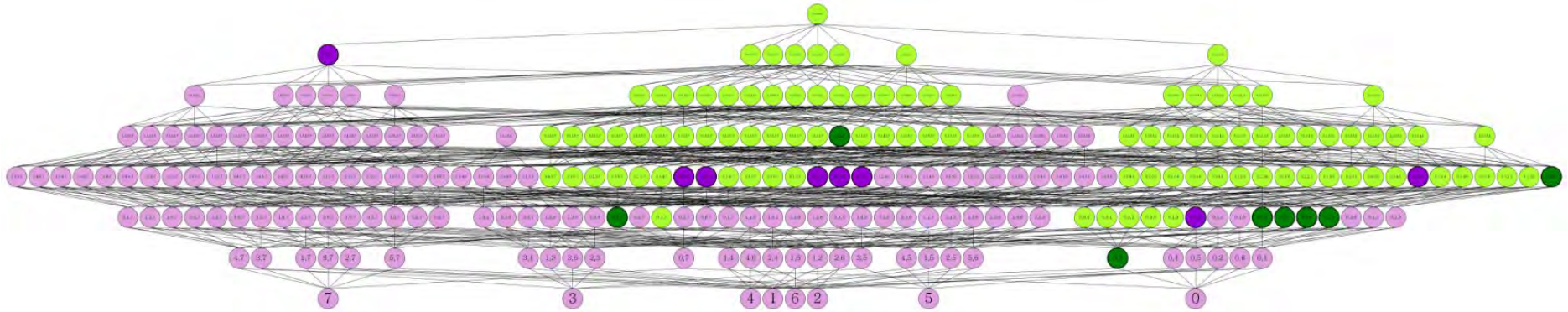
Scalable Profiling

Open Research Question

Can we predict problem and solution size for given data?

- Scalability in number of rows

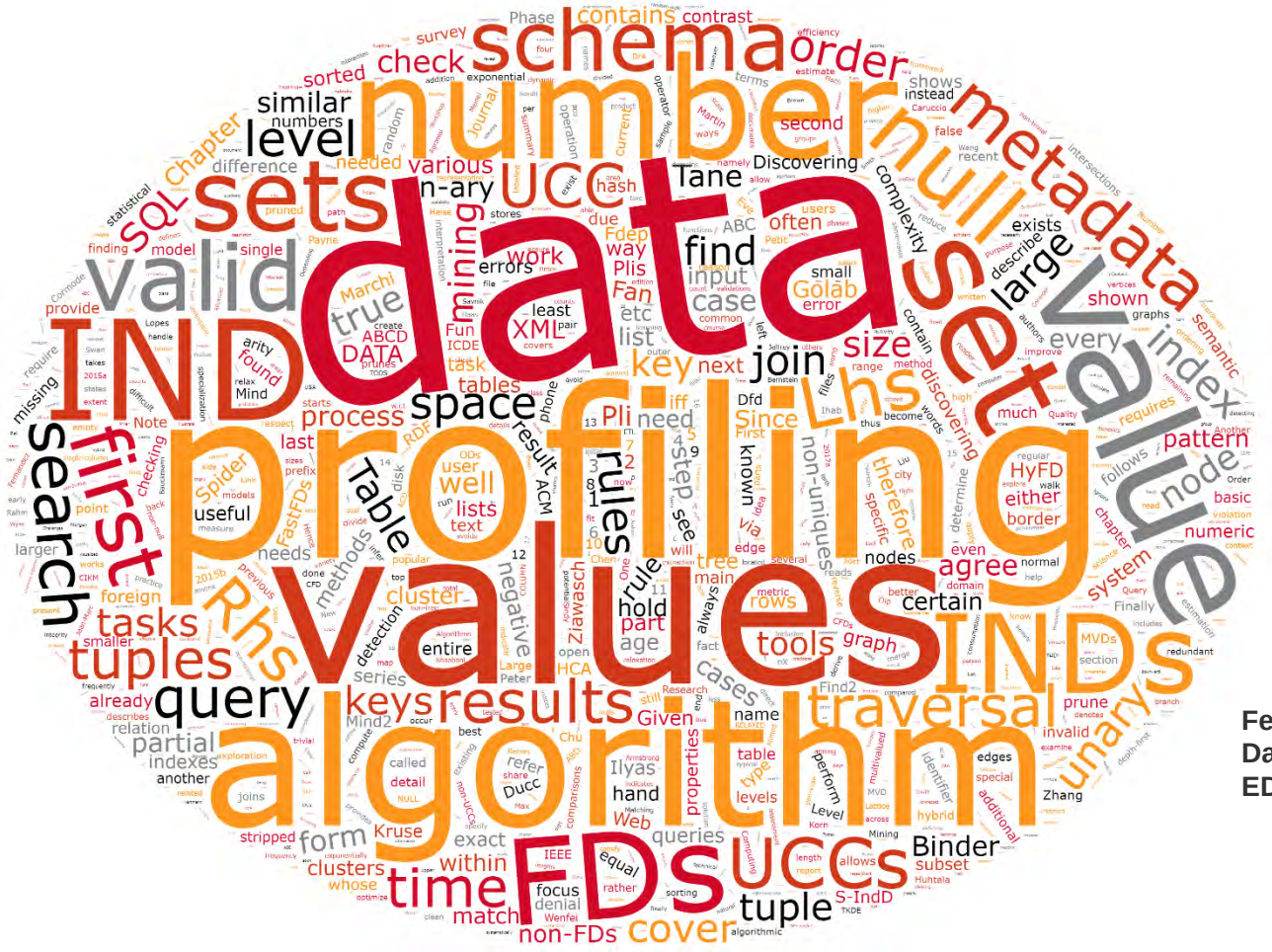
- Scalability in number of columns
  - “Normal” table with 100 columns:
    $2^{100} - 1 = 1{,}267{,}650{,}600{,}228{,}229{,}401{,}496{,}703{,}205{,}375$
    = 1.3 nonillion column combinations (the power set)



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

Agenda

1. **Basic statistics**
2. Uniques and keys
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. New directions



Cardinalities, Distributions, and Patterns

Open Research Question

Can we find useful regular expresssions for data columns?

Patterns and types

- String vs. number
- String vs. number vs. date
- Categorical vs. continuous
- SQL data types
- Domains
- Regular expressions
- Semantic data types

Increa-sing semantic level

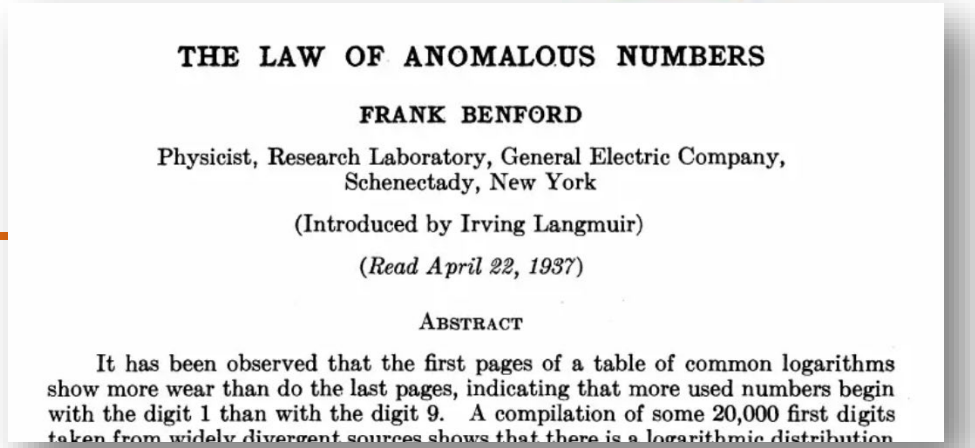
Cardinalities

- num-rows
- value length
- null values
- distinct
- uniqueness

Value distributions

- histograms
- constancy
- quartiles
- soundex
- first digit

Benford Law Frequency (“first digit law”)



THE LAW OF ANOMALOUS NUMBERS

FRANK BENFORD

Physicist, Research Laboratory, General Electric Company, Schenectady, New York

(Introduced by Irving Langmuir)

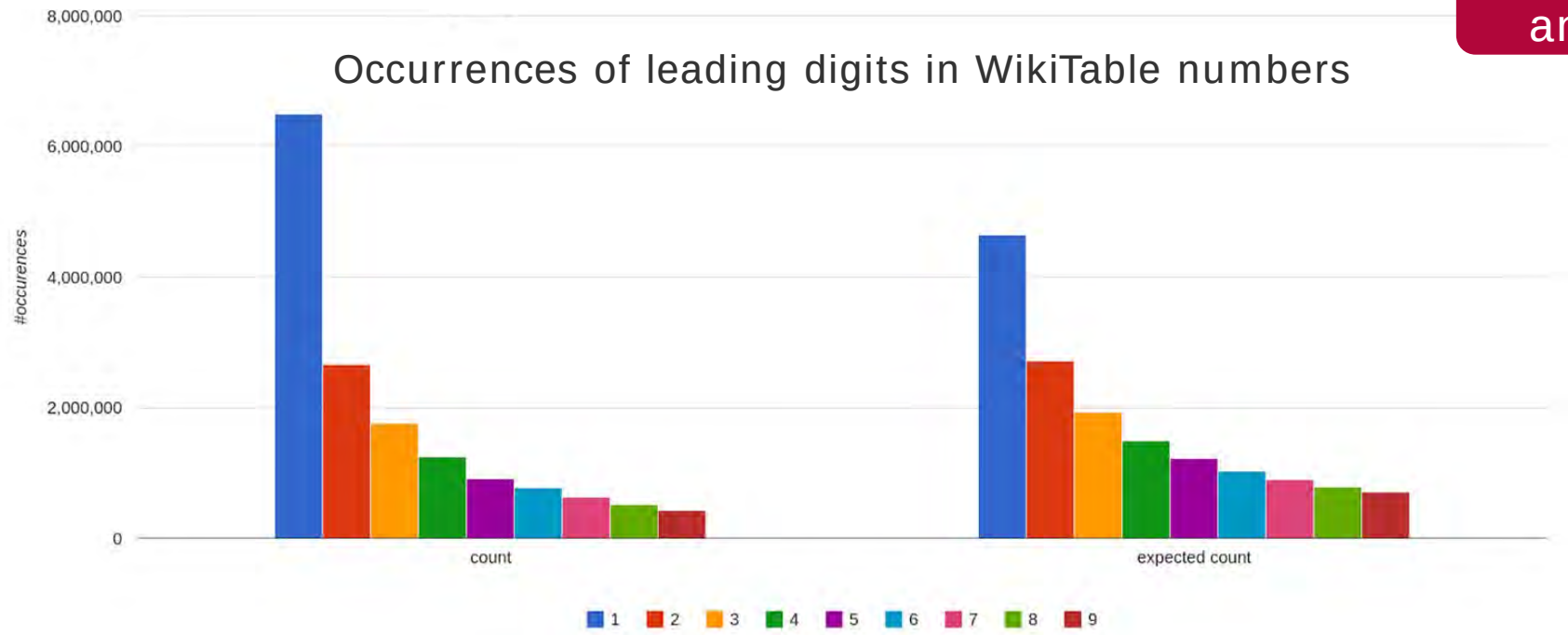
(*Read April 22, 1937*)

ABSTRACT

It has been observed that the first pages of a table of common logarithms show more wear than do the last pages, indicating that more used numbers begin with the digit 1 than with the digit 9. A compilation of some 20,000 first digits taken from widely divergent sources shows that there is a logarithmic distribution

- Distribution of first digits in naturally occurring numbers:
  - $P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}(1 + {}^{1}/_{d})$
  - Holds if log(x) is uniformly distributed
  - Street addresses of the first 342 persons listed in *American Men of Science*
  - 335 NIST physical constants

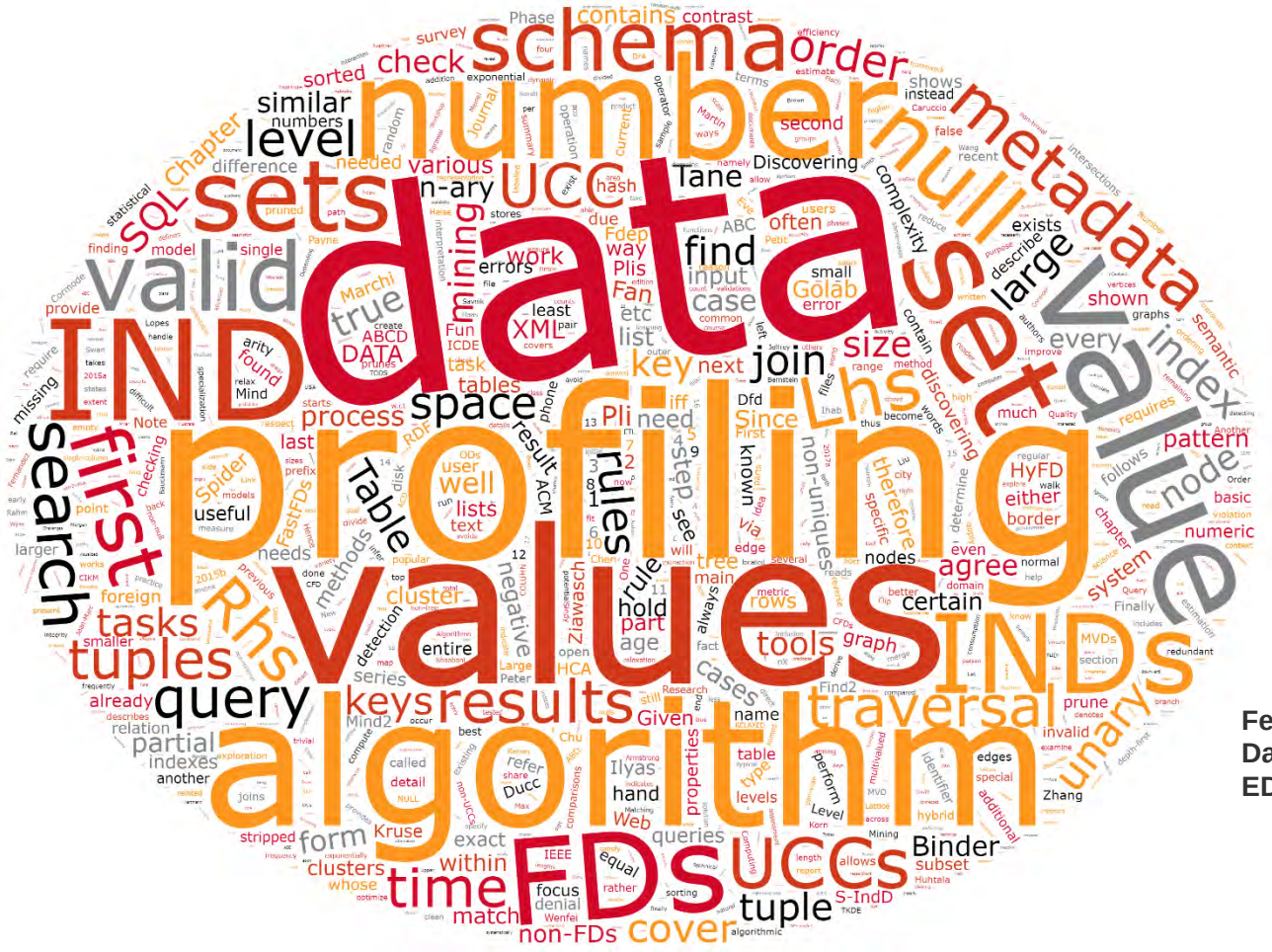
Since 1971 “American Men and Women of Science”



Felix Naumann
Data Profiling
EDBT 2021

Agenda

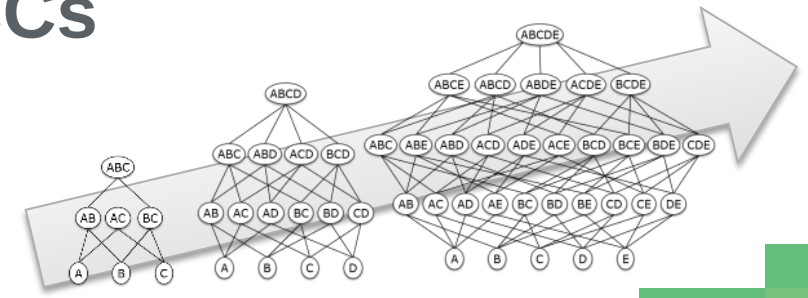
1. Basic statistics
2. **Uniques and keys**
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. New directions



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`

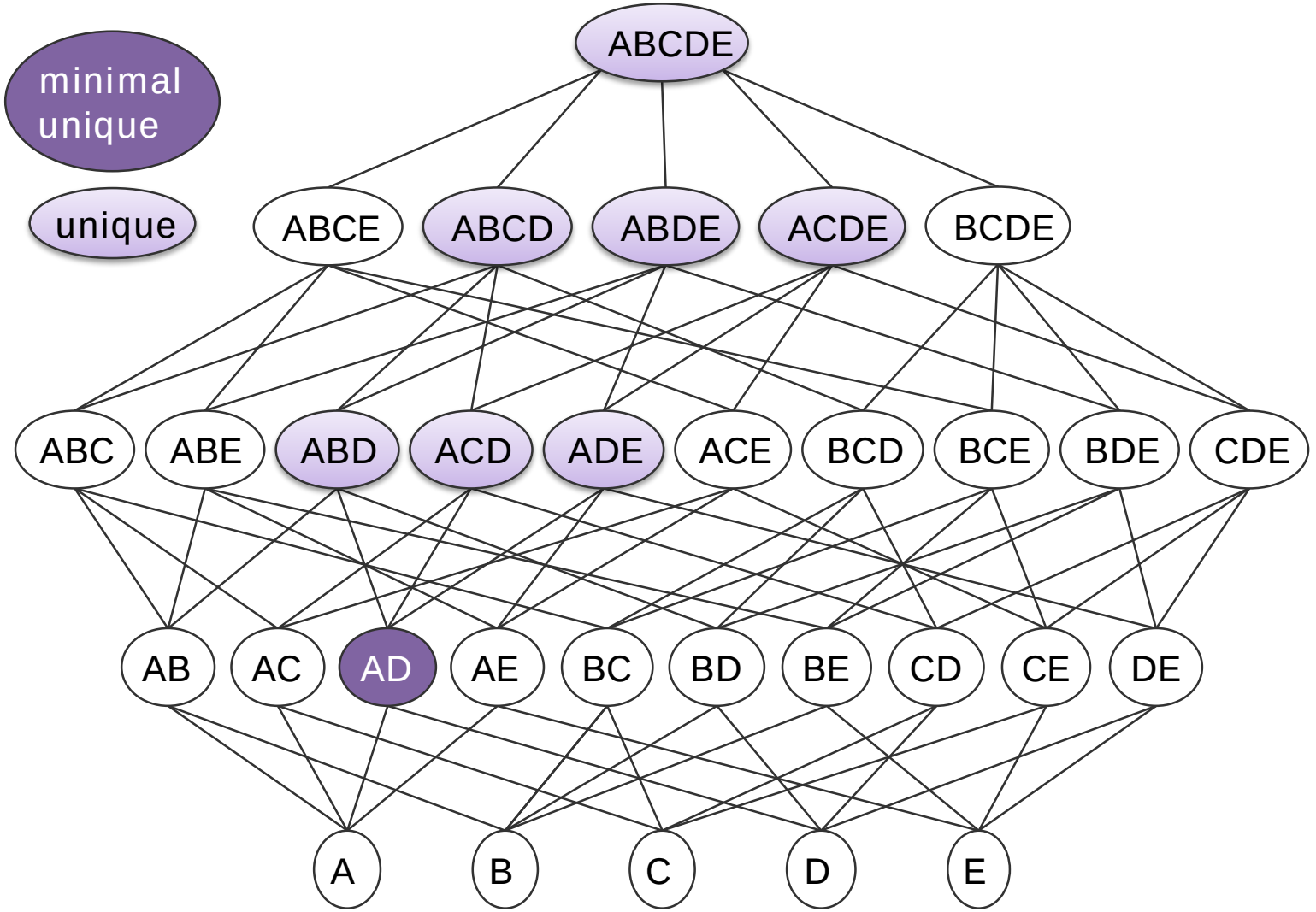
Candidate Set Growth for UCCs



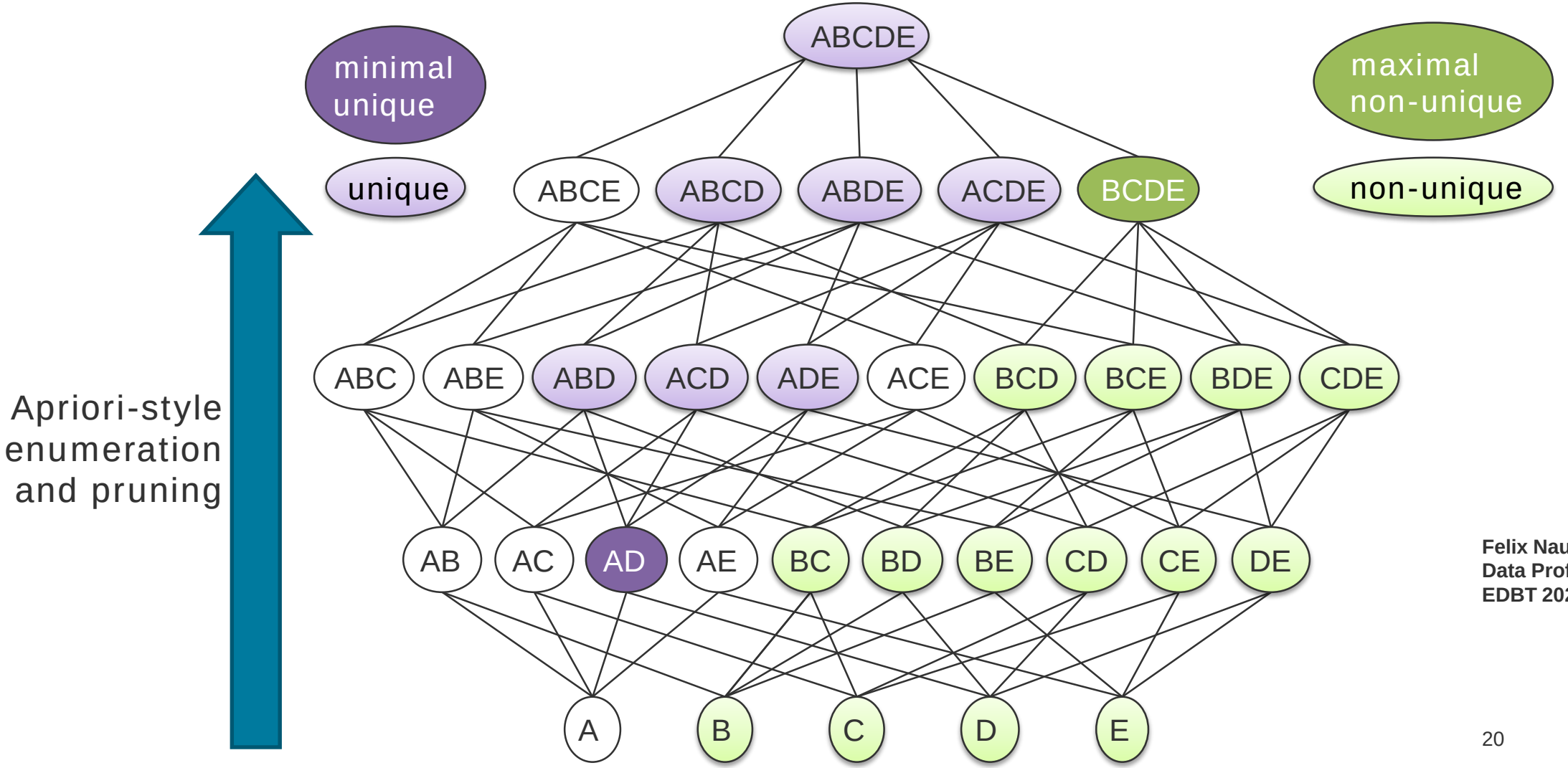
Number of levels: k \ Number of attributes: m	1	3	7	15	31	63	127	255	511	1,023	2,047	4,095	8,191	16,383	32,767
Total:	1	3	7	15	31	63	127	255	511	1,023	2,047	4,095	8,191	16,383	32,767
15															1
14														1	15
13													1	14	105
12												1	13	91	455
11											1	12	78	364	1,365
10										1	11	66	286	1,001	3,003
9									1	10	55	220	715	2,002	5,005
8								1	9	45	165	495	1,287	3,003	6,435
7							1	8	36	120	330	792	1,716	3,432	6,435
6						1	7	28	84	210	462	924	1,716	3,003	5,005
5					1	6	21	56	126	252	462	792	1,287	2,002	3,003
4				1	5	15	35	70	126	210	330	495	715	1,001	1,365
3			1	4	10	20	35	56	84	120	165	220	286	364	455
2		1	3	6	10	15	21	28	36	45	55	66	78	91	105
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Felix Naumann
Data Profiling
EDBT 2021

Pruning Supersets



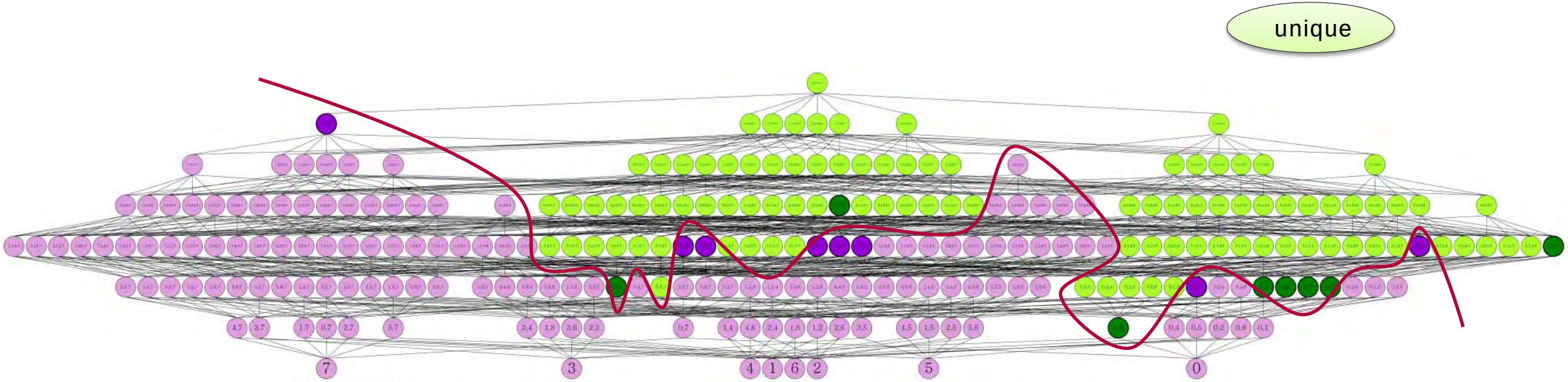
Pruning Subsets



DUCC – Detecting Unique Column Combinations

Open Research Question

What are effective distribution strategies for data profiling?

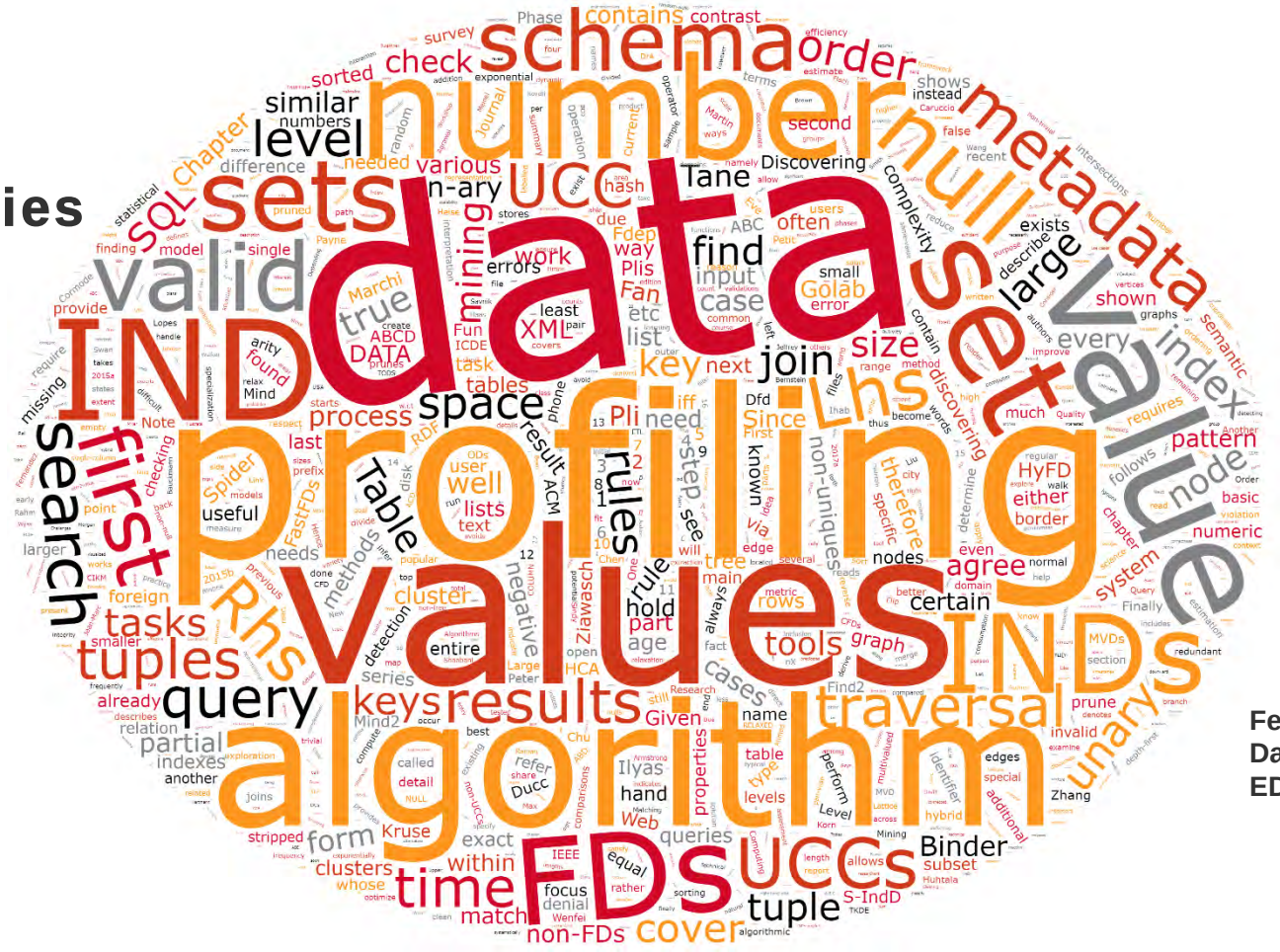


Felix Naumann
Data Profiling
EDBT 2021

[PVLDB'13]

Agenda

1. Basic statistics
2. Uniques and keys
3. **Functional dependencies**
4. Inclusion dependencies and foreign keys
5. New directions

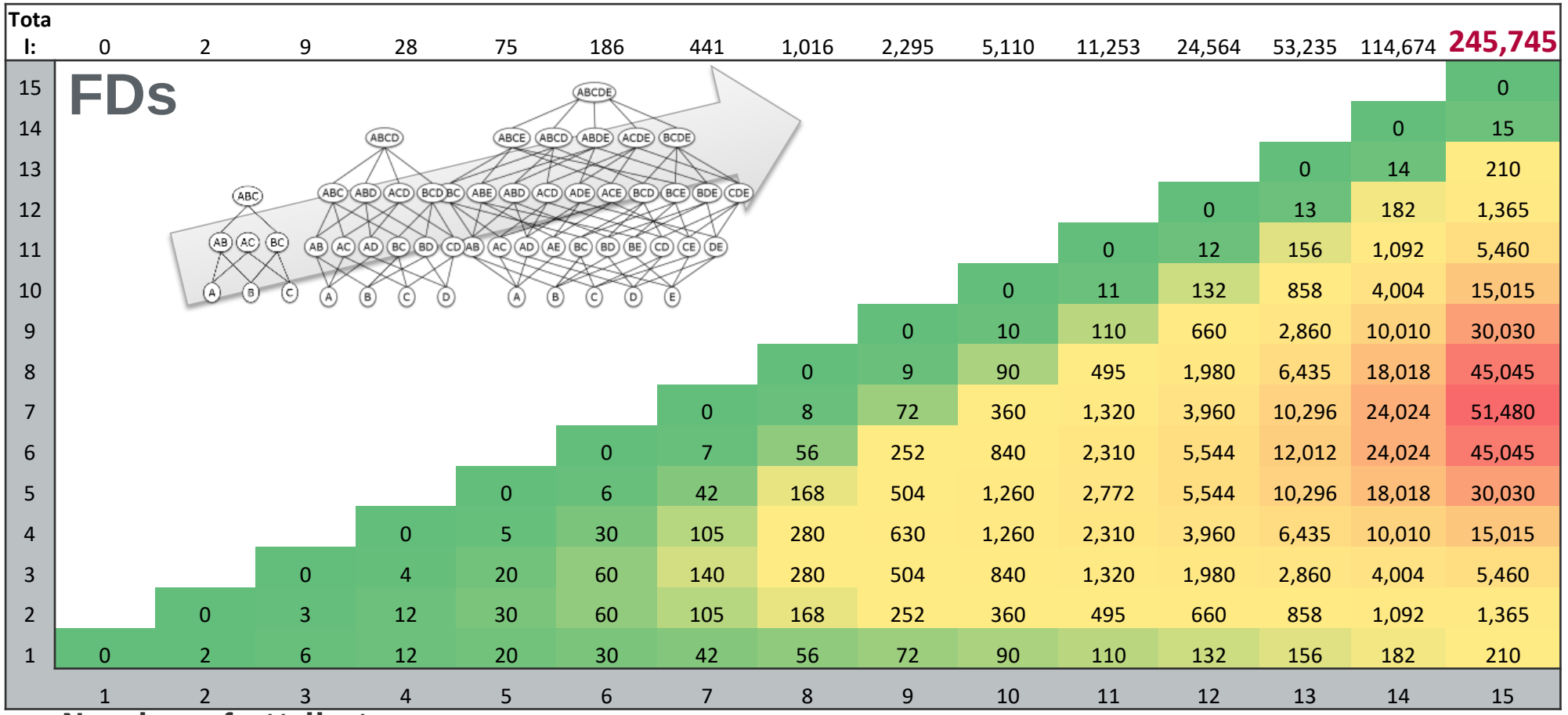


- „X → A“
  When two tuples have same value in attribute set X,
  they must have same values in attribute A.
  - E.g., ZIP → City

Naïve discovery approach

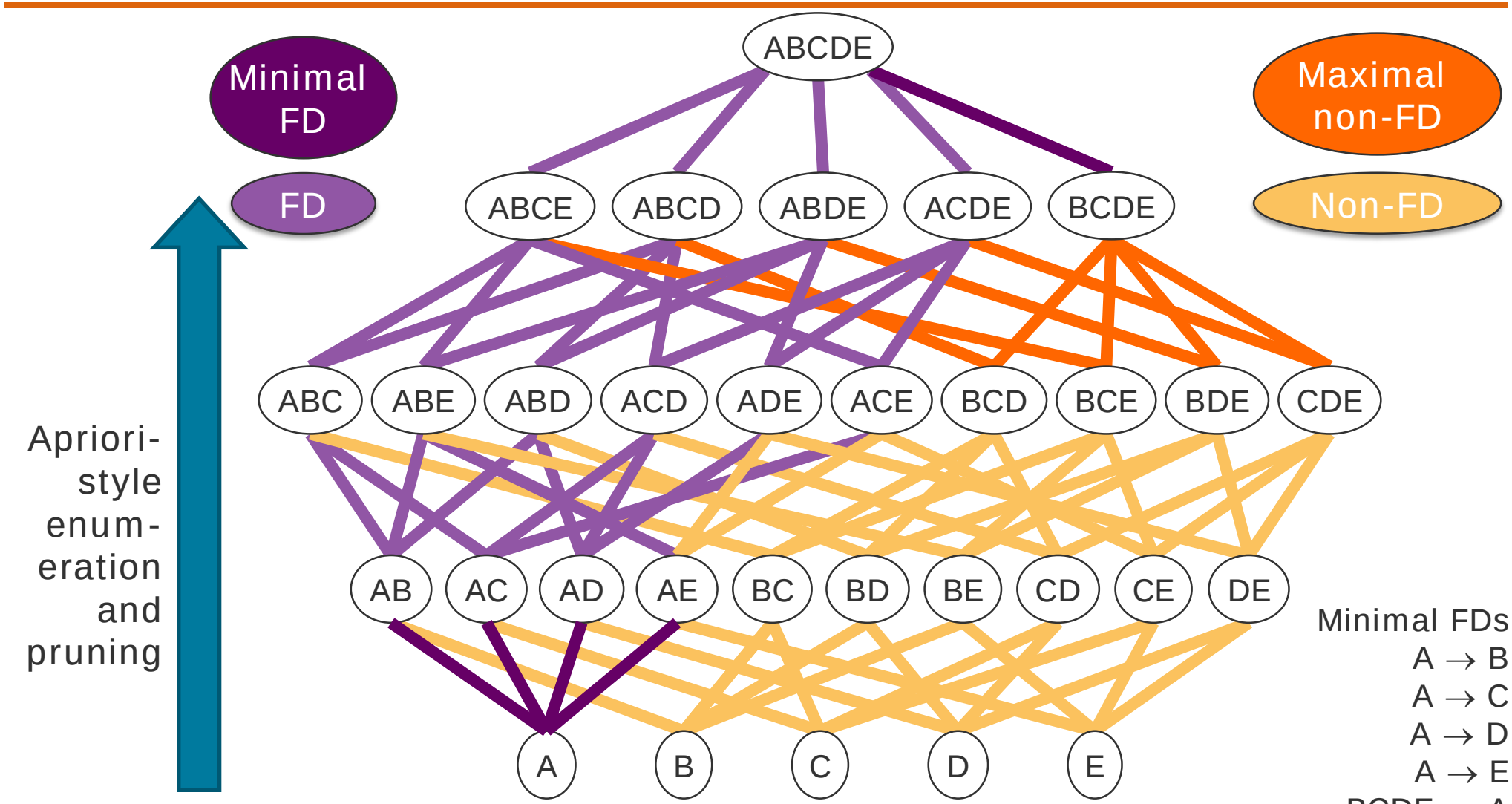
```
For each column combination X
  For each A∈X
    For each pair of tuples (t1,t2)
      If t1[X\A] = t2[X\A] and t1[A] ≠ t2[A]: Break
```

Candidate Set Growth for FDs



Felix Naumann
Data Profiling
EDBT 2021

Model in Lattice – Edges Represent FDs



Minimal FDs

$A \rightarrow B$

$A \rightarrow C$

$A \rightarrow D$

$A \rightarrow E$

$BCDE \rightarrow A$

Row-based Algorithms

Name	Surname	Postcode	City	Mayor
Thomas	Miller	14482	Potsdam	Jakobs
Sarah	Miller	14482	Potsdam	Jakobs
Peter	Smith	60329	Frankfurt	Feldmann
Jasmine	**Cone**	**01069**	**Dresden**	**Orosz**
Thomas	**Cone**	**14482**	**Potsdam**	**Jakobs**
Mike	Moore	60329	Frankfurt	Feldmann

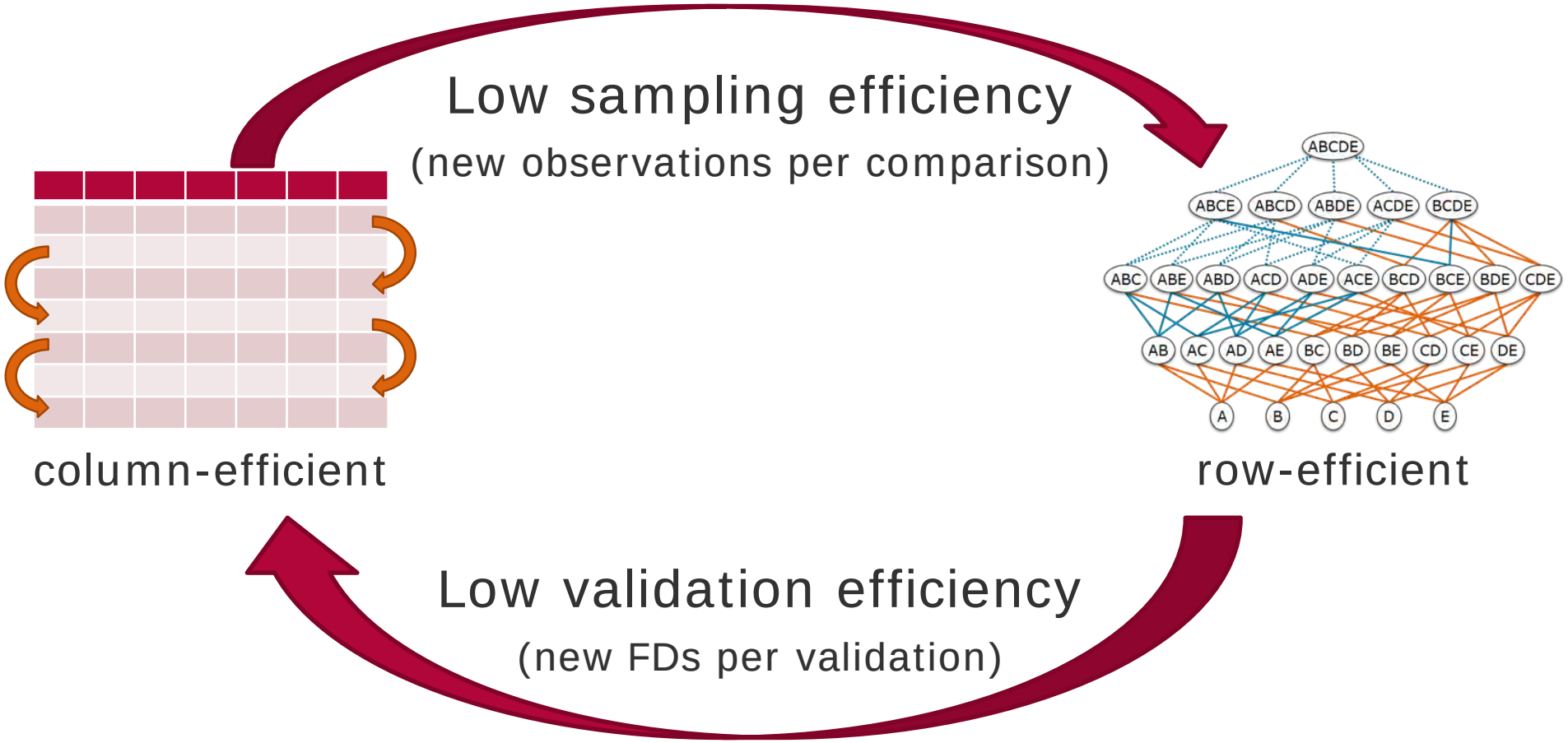
- Surname, Postcode, City, Mayor ↛ Name
- Name, Postcode, City, Mayor ↛ Surname
- Surname ↛ Name, Postcode, City, Mayor

→

Postcode → City
Postcode → Mayor
Name → Surname, ...

Flach, Peter A., and Iztok Savnik. "Database dependency discovery: a machine learning approach." *AI communications* 12.3 (1999): 139-160.

HyFD: Hybrid FD Discovery



Felix Naumann
Data Profiling
EDBT 2021

[SIGMOD'16]

Functional Dependencies: State of the Art

Open Research Question

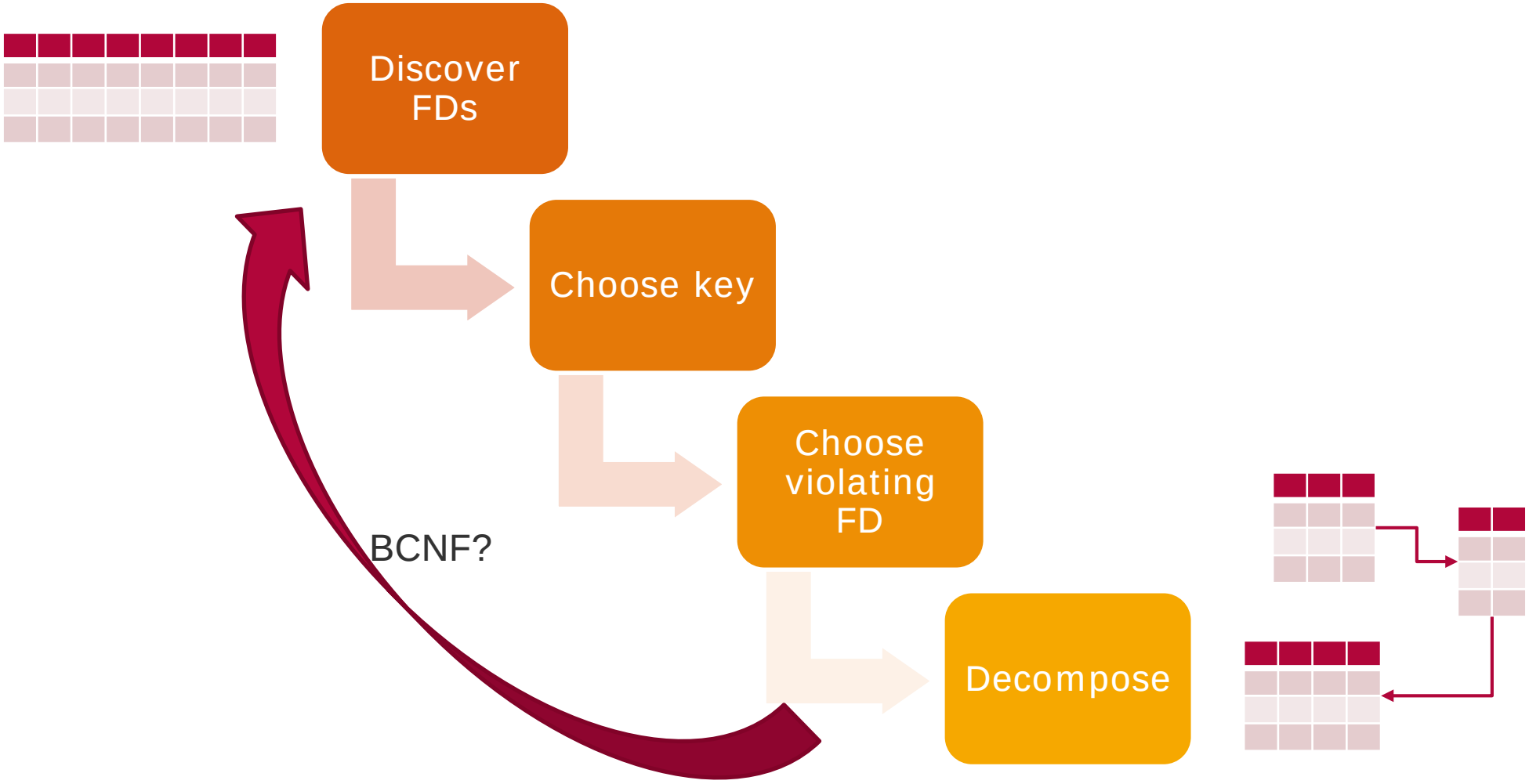
What is a useful and fair benchmark for data profiling?

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

Felix Naumann
Data Profiling
EDBT 2021

Use case: BCNF Normalization

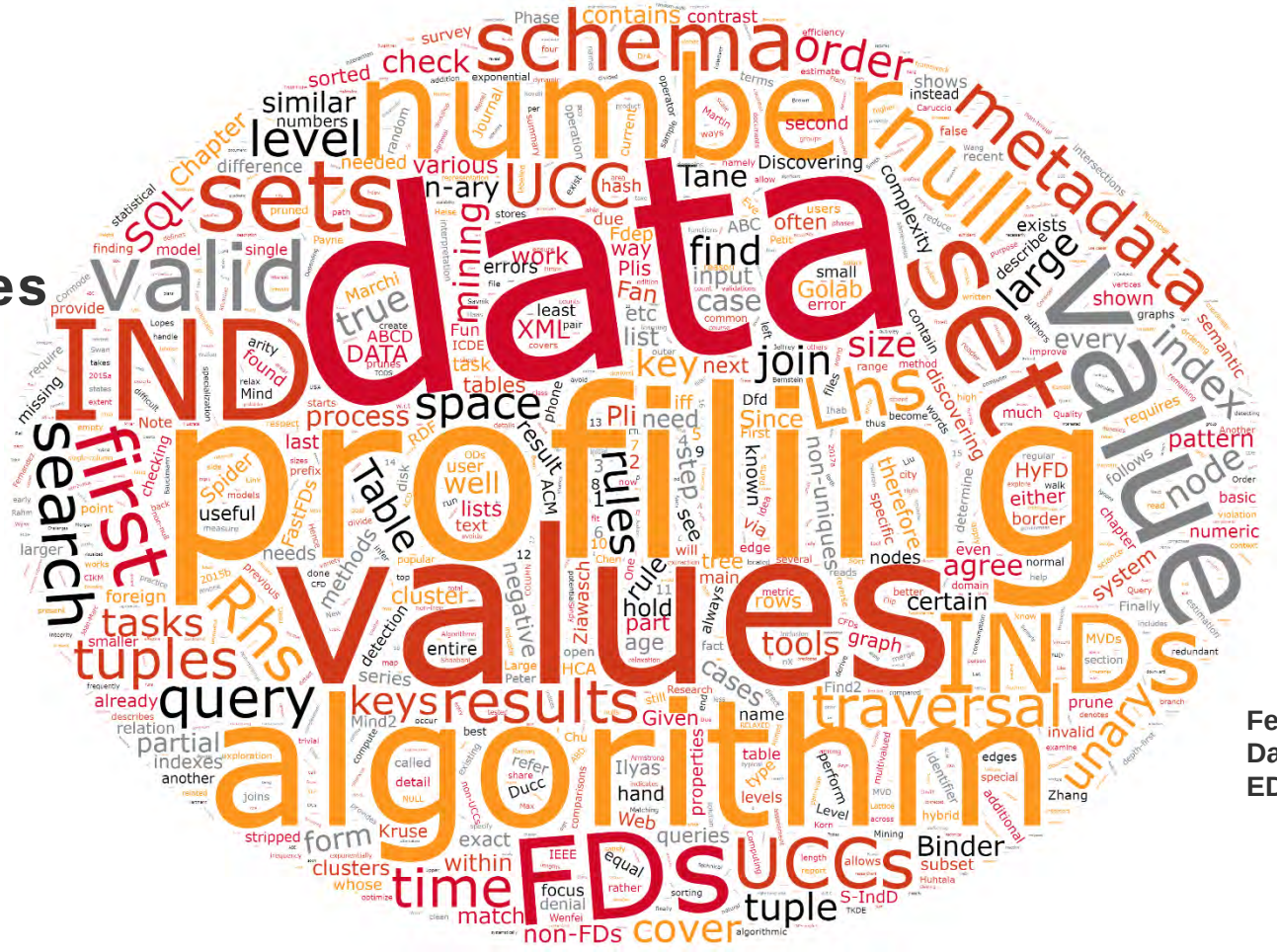


[EDBT'17]

Normalization Results: TPC-H

Agenda

1. Basic statistics
2. Uniques and keys
3. Functional dependencies
4. **Inclusion dependencies and foreign keys**
5. New directions



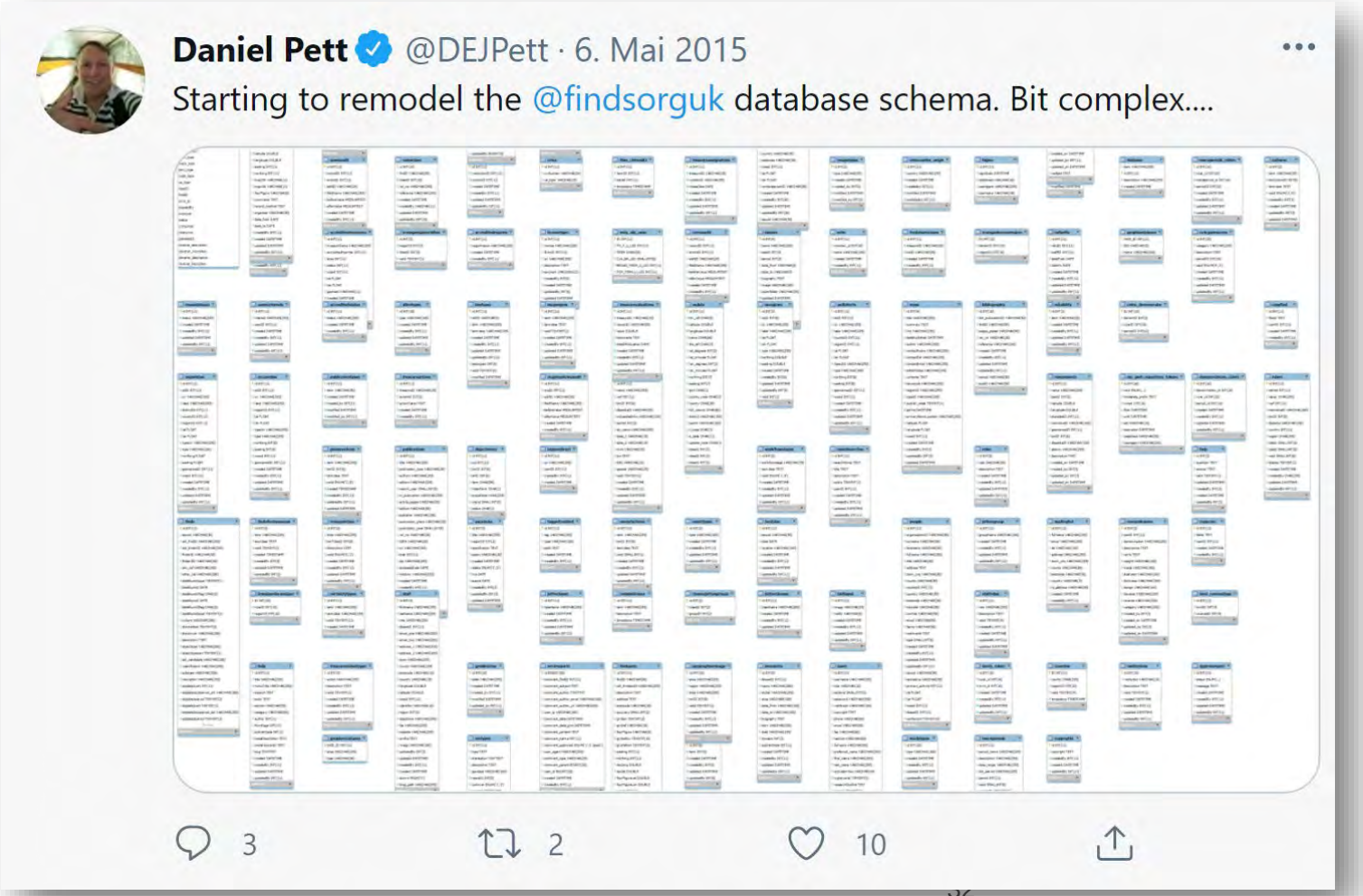
Felix Naumann
Data Profiling
EDBT 2021

Inclusion Dependencies for Foreign Key Discovery

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

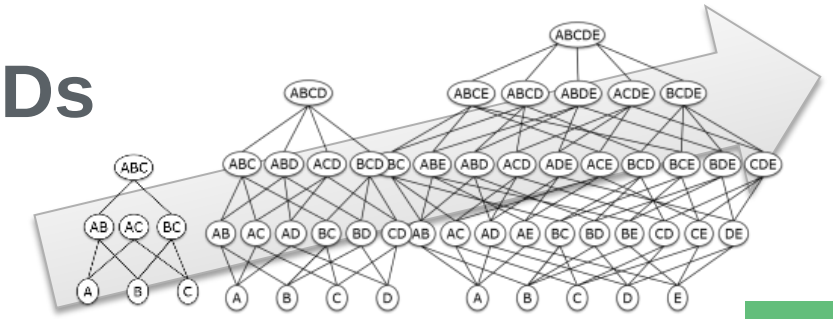
- Use cases
  - PDB – Protein Data Bank: 175 tables
    - Not a single foreign key constraint
  - Ensembl – genome database: >200 tables
    - Not a single foreign key constraint
  - Web tables:
    - No schema, no constraints, but many connections

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



Candidate Set Growth for INDs

k \ m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Total:	0	2	6	24	80	330	1,302	5,936	26,784	133,650	669,350	3,609,672	19,674,096	113,525,594	664,400,310
15															0
14														0	0
13													0	0	0
12												0	0	0	0
11											0	0	0	0	0
10										0	0	0	0	0	0
9									0	0	0	0	0	0	0
8								0	0	0	0	0	0	0	0
7							0	0	0	0	0	0	0	17,297,280	259,459,200
6						0	0	0	0	0	0	665,280	8,648,640	60,540,480	302,702,400
5					0	0	0	0	0	30,240	332,640	1,995,840	8,648,640	30,270,240	90,810,720
4				0	0	0	0	1,680	15,120	75,600	277,200	831,600	2,162,160	5,045,040	10,810,800
3			0	0	0	120	840	3,360	10,080	25,200	55,440	110,880	205,920	360,360	600,600
2		0	0	12	60	180	420	840	1,512	2,520	3,960	5,940	8,580	12,012	16,380
1	0	2	6	12	20	30	42	56	72	90	110	132	156	182	210



Number of levels: k

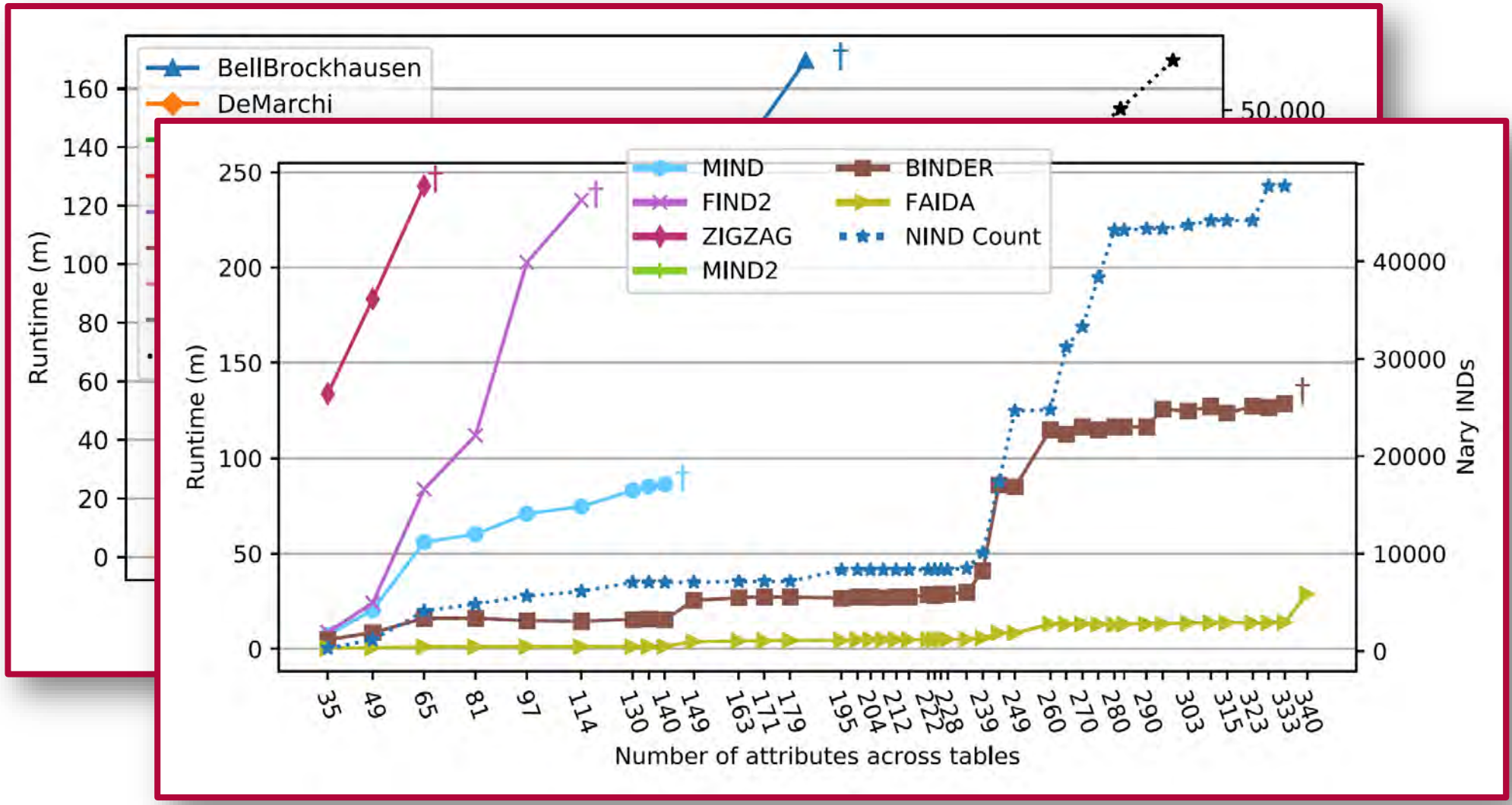
Number of attributes: m

Felix Naumann
Data Profiling
EDBT 2021

IND Detection Algorithms

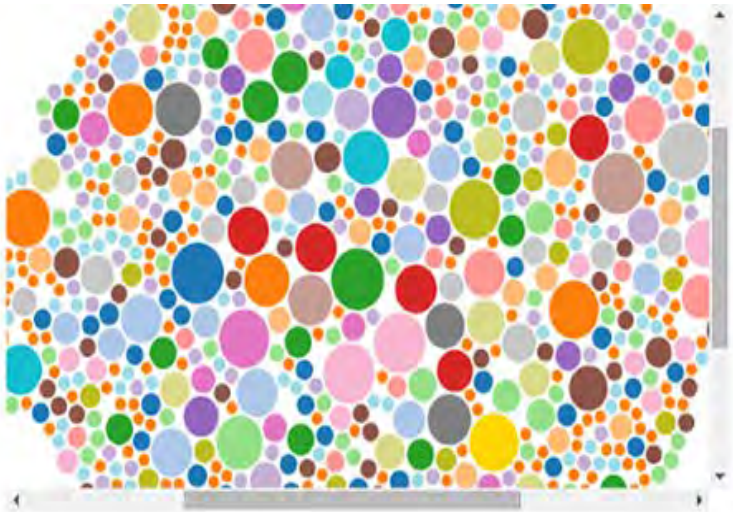
Open Research Question

How to efficiently discover all n-ary INDs?



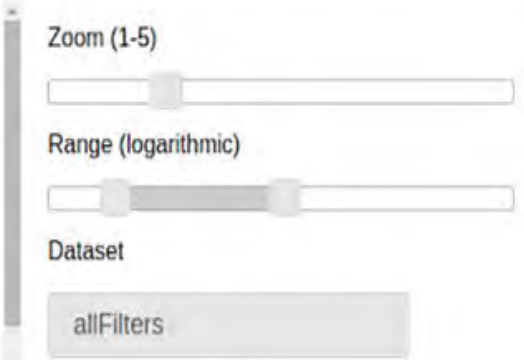
[CIKM'19]

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-	

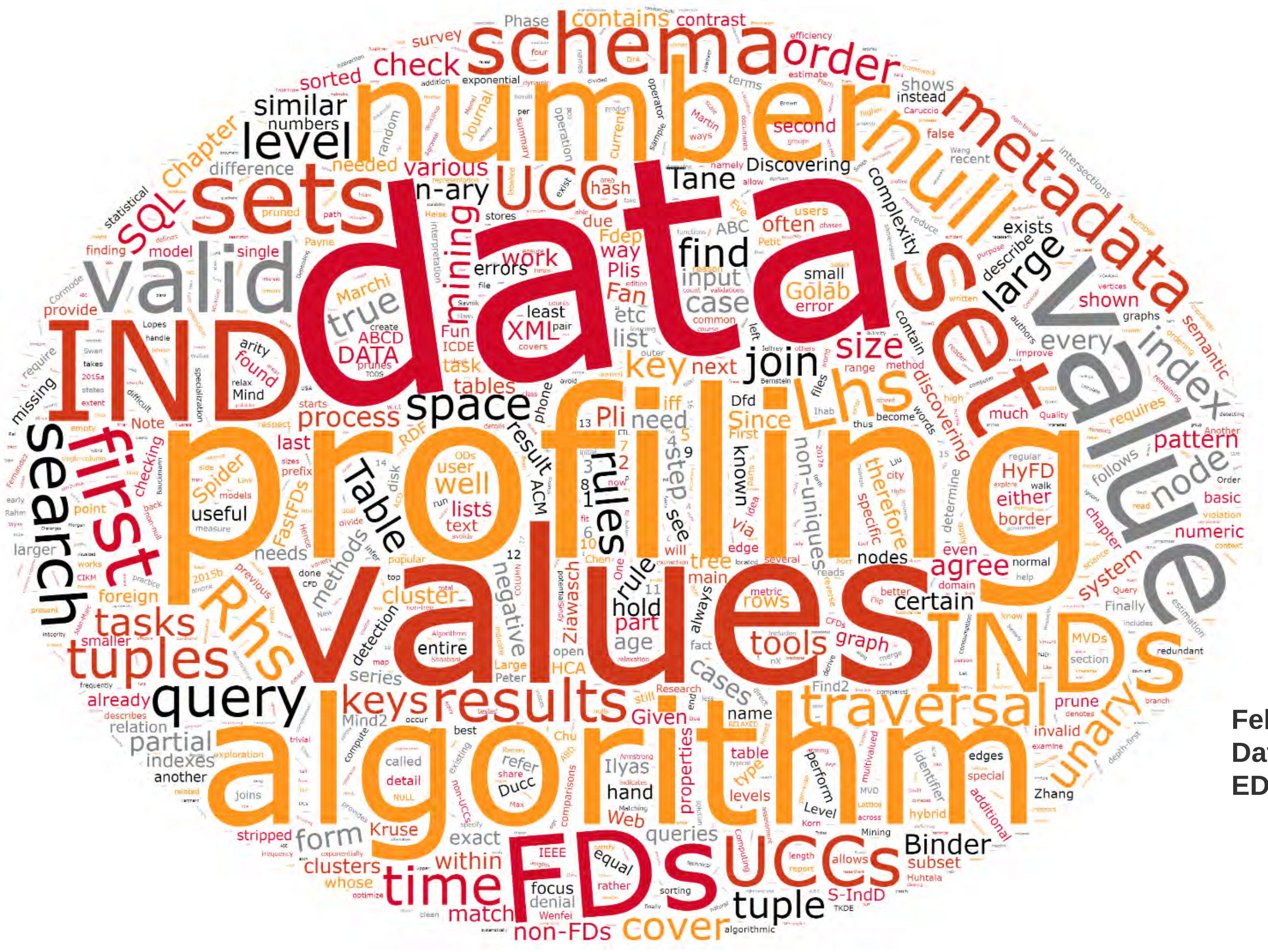
Celestial Objects	Rotation period	Rotation period
Sun	25.379995 days (equatorial) 35 days (high latitude)	25 d 9 h 7 m 11.6 s 35 d
Mercury	58.6462 days	58 d 15 h 30 m 30 s
Venus	?243.0187 days	?243 d 0 h 26 m
Earth	0.99726968 days	0 d 23 h 56 m 4.100 s
Moon	27.321661 days (synchronous toward Earth)	27 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



[TODS'17]

Agenda

1. Basic statistics
2. Uniques and keys
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. **New directions**



Are there not more types of dependencies? Detecting Other Dependencies

Open Research Question

How to combine the discovery of multiple dependency types?

- Multi-valued dependencies (MVDs) and join dependencies
- Denial constraints (DCs)
- Detecting order dependencies (ODs)
  - salary „orders“ rank
  - **SELECT emp_name**
    FROM employees
    ORDER BY rank, salary

emp_name	rank	salary
Smith	1	40k
Johnson	1	40k
Williams	1	45k
Brown	2	60k
Davis	2	60k
Miller	3	70k
Wilson	4	100k

But what if the data changes?
Incremental Dependency Discovery

Open Research Question

Can we efficiently maintain metadata?

- Insertions, deletions and updates
  - Invalidate existing dependencies
    - Check all dependencies
  - Create new (minimal/maximal) dependencies
    - Re-run entire algorithm?

First	Last	City
John	Doe	Berlin
Jane	Smith	Berlin
John	Miller	Paris
John	Smith	Berlin
John	Doe	Paris

But what if the data contains errors?

Relaxed Dependencies

- Relaxed
  - Partial dependencies
  - Conditional dependencies
  - Matching dependencies
- Approximate dependencies
- Dependencies on uncertain data
- Dependencies on incomplete data

[Caruccio, Deufemia, Polese: Relaxed Functional Dependencies - A Survey of Approaches. TKDE '16]

RFD abbrev.	RFD name
ACOD	Approximate comparable dependency
ADD	Approximate differential dependency
AFD	Approximate functional dependency
COD	Comparable dependency
CFD	Conditional functional dependency
CFD^p	CFD with built-in predicates
CFD^c	CFD with cardinality constraints and synonym rules
CMD	Conditional matching dependency
CSD	Conditional sequential dependency
CD	Constrained functional dependency
DD	Differential dependency
eCFD	Extended conditional functional dependency
FFD	Fuzzy functional dependency
MD	Matching dependency
MFD	Metric functional dependency
ND	Neighborhood dependency
NuD	Numerical dependency
OD	Order dependency
OD_K	OD satisfied within bound k
OD_{EA}	OD satisfied almost everywhere
OFD	Ordered functional dependency
PD	Partial determination
POD	Polarized order dependencies
preFD	Preference functional dependency
PAC	Probabilistic approximate constraint
pFD	Probabilistic functional dependency
PuD	Purity dependency
RUD	Roll-up dependency
SD	Sequential dependency
SFD	Similarity functional dependency
soft FD	Soft functional dependency
TD	Trend dependency
TMFD	Type-M functional dependency
XCFD	XML conditional functional dependency
$\sigma\theta$XFD	XML FD with σ and θ approximation

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
  - ...

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

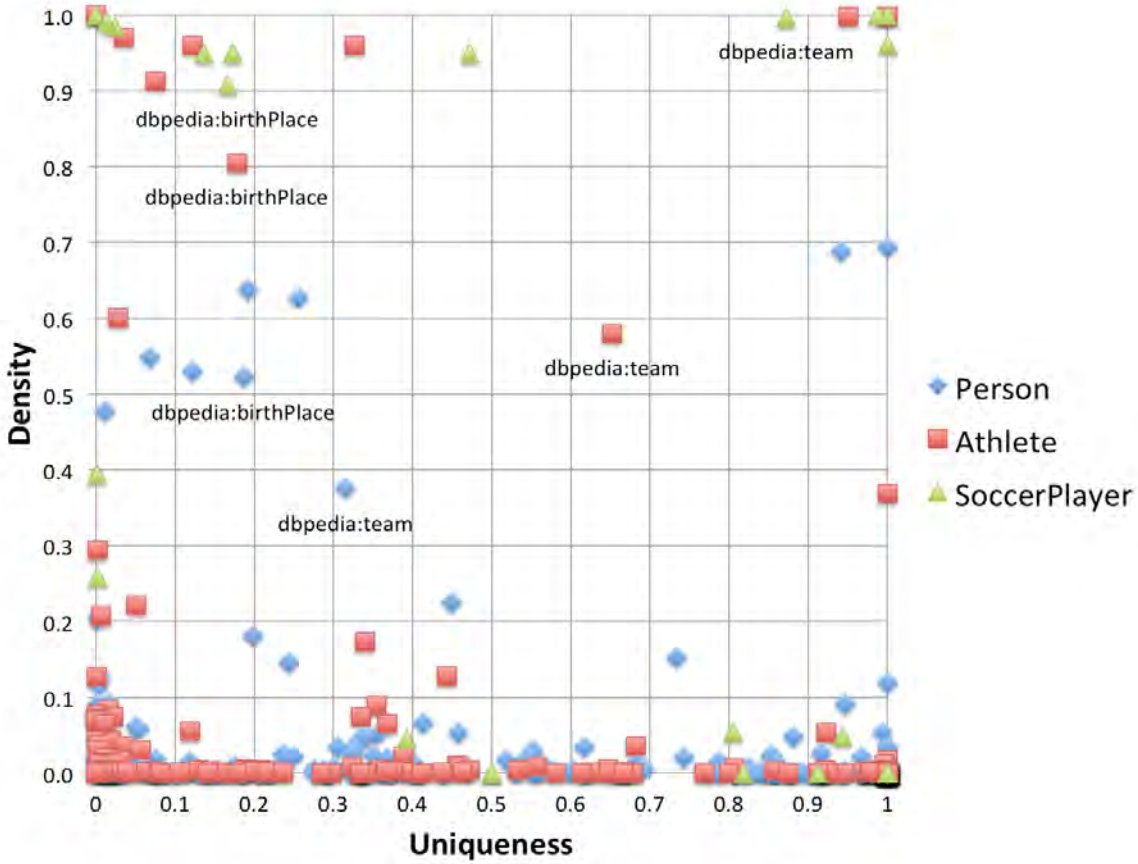
Cust_ID	Cust_status	...
0	2	
1	0	
2	1	
3	2	
...	...	

Profiling for Other Data Models

Open Research Question

Can we define UCCs, INDs, FDs, etc. for trees and graphs?

- Traditional data profiling: Single table or multiple tables
- "Newer" data models
  - XML / nested relational / JSON
  - Graphs
  - RDF triples
- New metadata types to profile
  - XML: Nestedness; measures at each nesting level
  - RDF: Graph structure; node-degrees

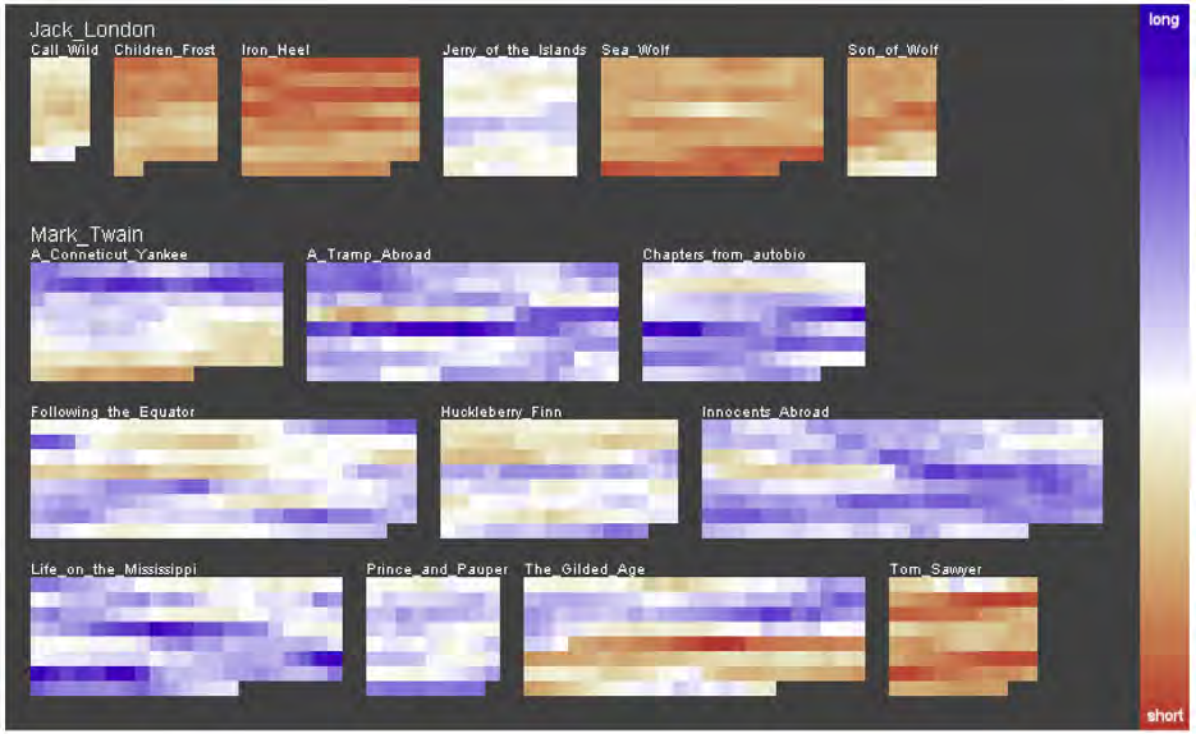


Profiling Multimedia Data

Open Research Question

What even are interesting metadata for multimedia?

- Images and videos
  - Color, video-length, volume, etc.
- Textual data
  - Statistical measures
    - Syllables per word
    - Sentence length
    - Parts of speech
  - Vocabulary measures
    - Frequencies of specific words
    - Simpson's index
  - Content
    - Sentiment



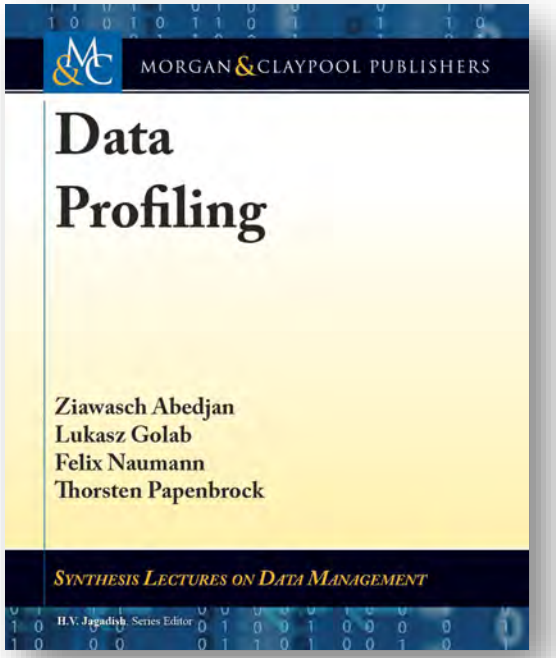
„*Literature Fingerprinting: A New Method for Visual Literary Analysis*" by Daniel A. Keim and Daniela Oelke (IEEE Symposium on Visual Analytics Science and Technology 2007)

Team and Collaborators

Metanome
www.metanome.de

HPI Hasso Plattner Institut

- PhD students at HPI
  - Anja Jentzsch
  - Arvid Heise
  - Hazar Harmouch
  - Ioannis Koumarelas
  - Jan Kossmann
  - Jana Bauckmann
  - Lan Jiang
  - Leon Bornemann
  - Sebastian Kruse
  - Sebastian Schmidl
  - Thorsten Papenbrock
  - Tobias Bleifuß
  - Ziawasch Abedjan
- Data profiling collaborators
  - Edson R.L. Filho
  - Eduardo C. de Almeida
  - Eduardo Pena
  - Giuseppe Polese
  - Hannes Mühleisen
  - Heiko Müller
  - Johann Birnick
  - Jorge-Arnulfo Quiané-Ruiz
  - Laure Berti-Équille
  - Loredana Caruccio
  - Lukasz Golab
  - Martin Schirneck
  - Noël Novelli
  - Paolo Papotti
  - Thomas Bläsius
  - Tobias Friedrich
  - Ulf Leser
  - Vincenzo Deufemia
  - Zoi Kaoudi
  - Saravanan Thirumuruganathan



Plus many great masters students

Summary and Outlook

Current focus on efficiency
- New algorithms
- Approximation

Future focus on efficiency
- Distribution
- Profiling dynamic data
- Query optimization

Current focus on semantics
- Keys and foreign key
- Use cases

Future focus on semantics
- Genuineness
- Data cleaning

Current focus on new problems
- Relaxed dependencies
- New dependency types

Future focus on new problems
- New data models

Open Research Questions

Hasso Plattner Institut

How to support **visual data profiling**?

How to introduce dependency-based **optimization** to DBMS?

Can we **predict** problem and solution size for given data?

Can we find useful **regular expresssions** for data columns?

What are effective **distribution** strategies for data profiling?

Can we efficiently **maintain** metadata?

How to efficiently discover all **n-ary INDs**?

How to **combine the discovery** of multiple dependency types?

What even are interesting metadata for **multimedia**?

Can we effectively identify **genuine dependencies**?

Can we define UCCs, FDs, etc. for **trees and graphs**?

What is a useful and fair **benchmark** for data profiling?