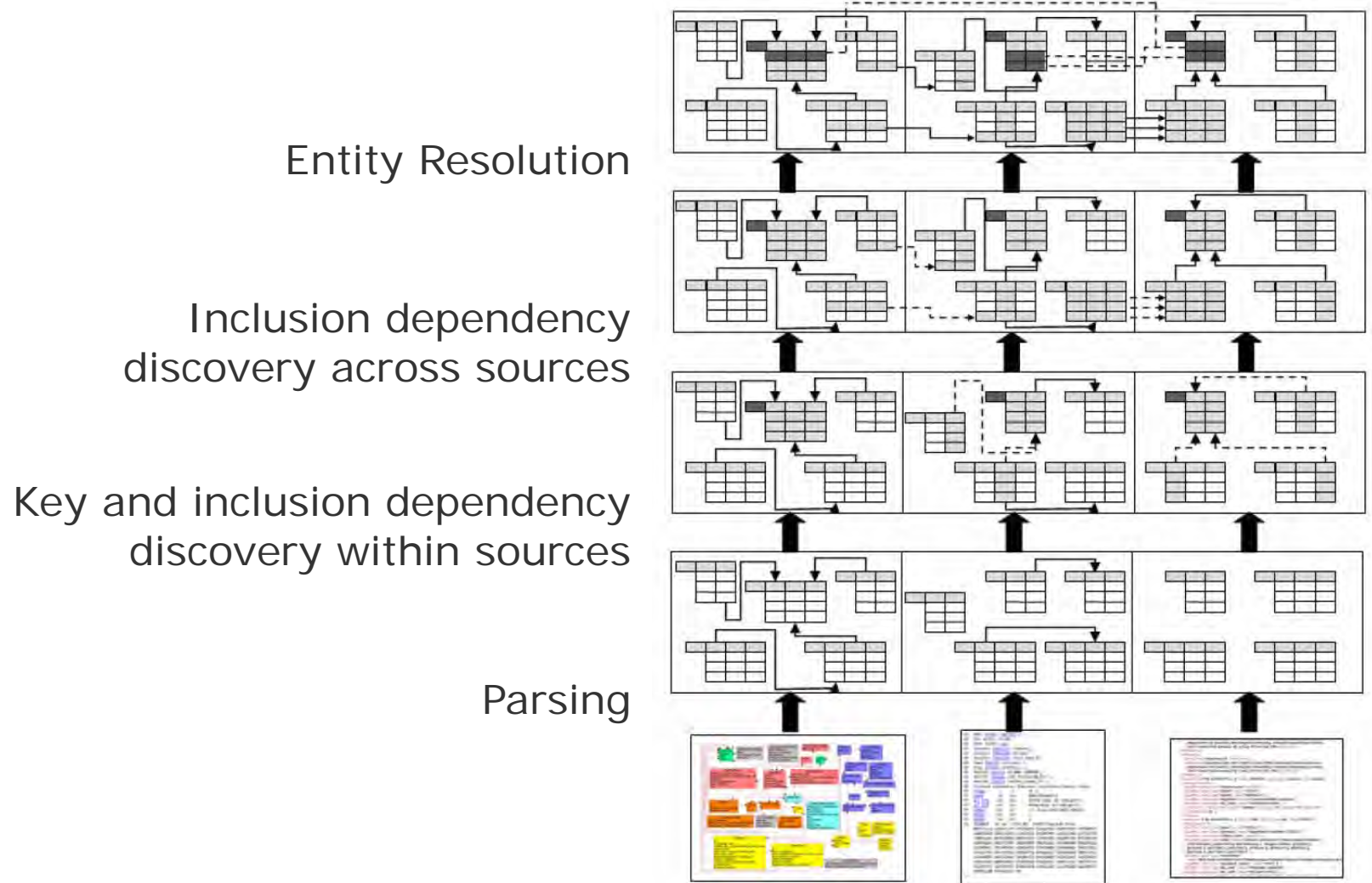


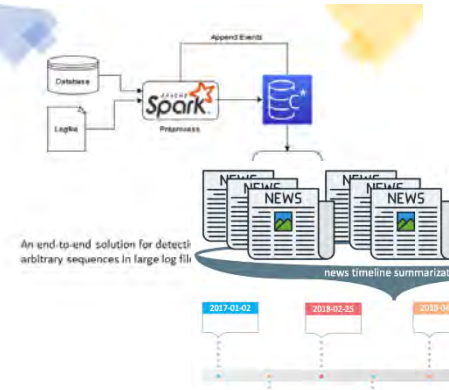


Data Profiling – A look back and a look forward

EDBT 2021
Felix Naumann



Felix Naumann
Data Profiling
EDBT 2021



Querying Top-k Dominant Traffic Flows on Large Urban Road Networks

Help traffic optimization techniques by identifying the dominant flows!

Stavros Sotiropoulos, Nikos Christopoulos, George Vassilakis, University of Piraeus, Greece

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]}

EBDT 2021
M. Francia et al.: Assess Queries for Interactive Analysis of Data Cubes

Structure Detection in Verbose CSV Files

Header Cell	Metadata Cells	Header Cells	Group Header Cell	Footer Row
Product	Product Name	Product Category	Product Sub-Category	Product Sales
Region	Region Name	Region Code	Region Sales	Region Total
Year	Year	Year	Year	Year

SPARQL Query Optimization using Shape Statistics

A lightweight extension of SHACL shapes to compute cardinality estimates for join ordering in SPARQL query processing.

Knowledge Graph, Shapes Statistics, Extraction & Annotation, Query Engine

SPARQL Query: SELECT * WHERE { ?node ?p ?o. }

Query Optimizer, Cardinality Estimator, Join Ordering, Optimized Query

Kashif Rabbani, Matteo Lindorini, Katja Hose
http://relweb.cs.aau.dk/rdfs/shape/

Generating Realistic Test Datasets for Duplicate Detection at Scale Using Historical Voter Data

ARE YOU HAUNTED BY DUPLICATES? DON'T BE AFRAID!
USE OUR TEST DATA TO CHOOSE & SHARPEN YOUR WEAPONS!

Qeios International Conference on Learning Database Technology (2021) 2021

Coronis: Towards Integrated and Open COVID-19 Data

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.

Georgios M. Santopoulou, George A. Vouros, Christos Doukidis, Dept. of Digital Systems, University of Piraeus, Greece

INTERNATIONAL CONFERENCE EBDT 2021

October 12-13, 2021, Athens, Greece

TD-AC: Efficient Data Partitioning based Truth

NexiaJD

Based on data profiles & learning models
Scalability powered by Spark

Efficient Exploratory Clustering Analyses with Qualitative Approximations

Acceptable Quality vs High Quality vs High Runtime

Manuel Fritz, Dennis Tschelchov, Holger Schwarz

Sharing redundant work in Microsoft's Cosmos big data platform

Data scientists and analysts

Scalable Spatio-temporal Indexing and Querying over a Document-oriented NoSQL Store

Nikolaos Koutroumanis, Christos Doukidis, Dept. of Digital Systems, University of Piraeus, Greece

Efficient ST Querying

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!

Ad Now?

Dependencies

Dependencies, Dependencies, Dependencies, Dependencies, Dependencies, Dependencies, Dependencies, Dependencies

Paper #264: Adaptive Multi-Model Reinforcement Learning for Online Database Tuning

Yaniv Gur, Dongsheng Yang, Frederik Stalschus, Berthold Reinwald

Automate Data Quality Validation for your Data Pipelines

Simple, Fast, Dynamic

No domain expertise or labeled examples required
>10x speedup over baselines in the majority of cases
Works if data characteristics change over time

Scalable JOIN DISCOVER

Connecting data analysts to the right data.

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

Explanation Algorithms

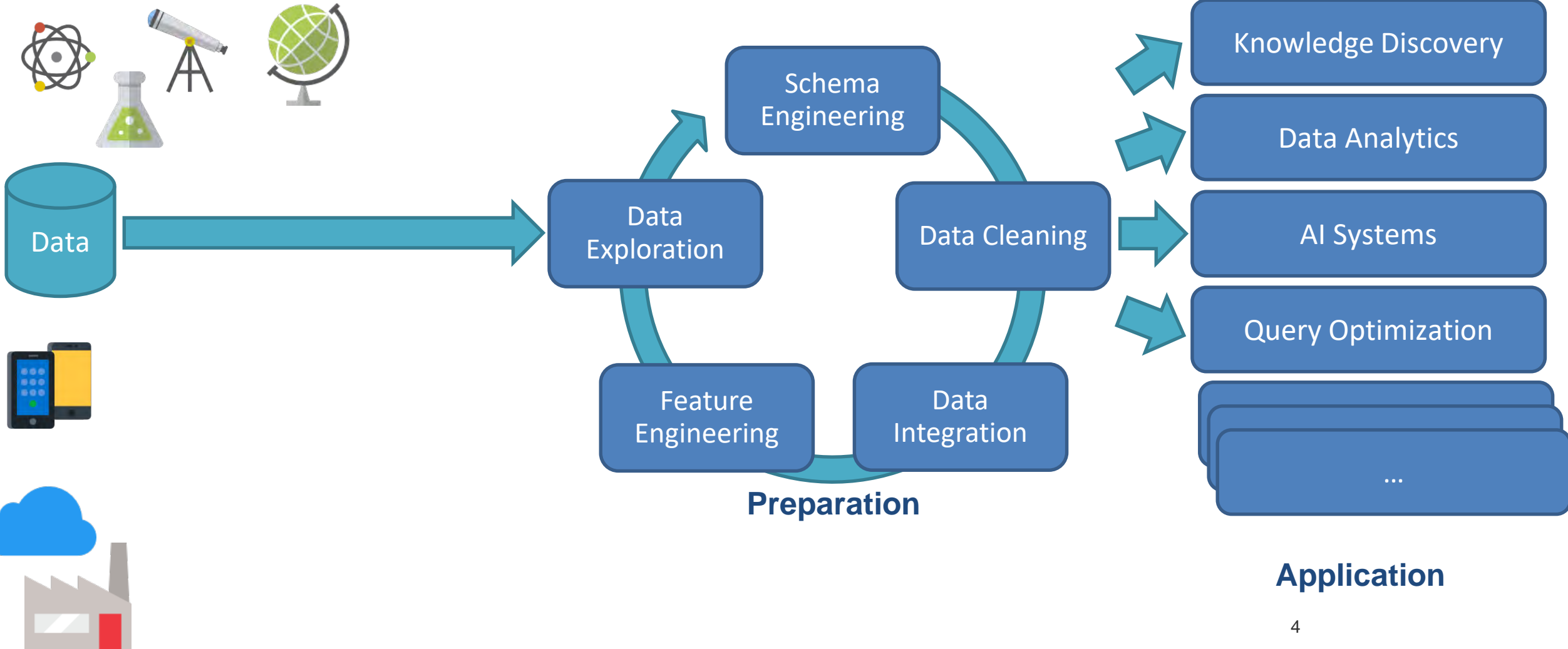
Nikos Myrtales, Vassilis Christophides, Eric Simon

Stage-wise method vs Random Projection method

Fractal Dimension vs Radius

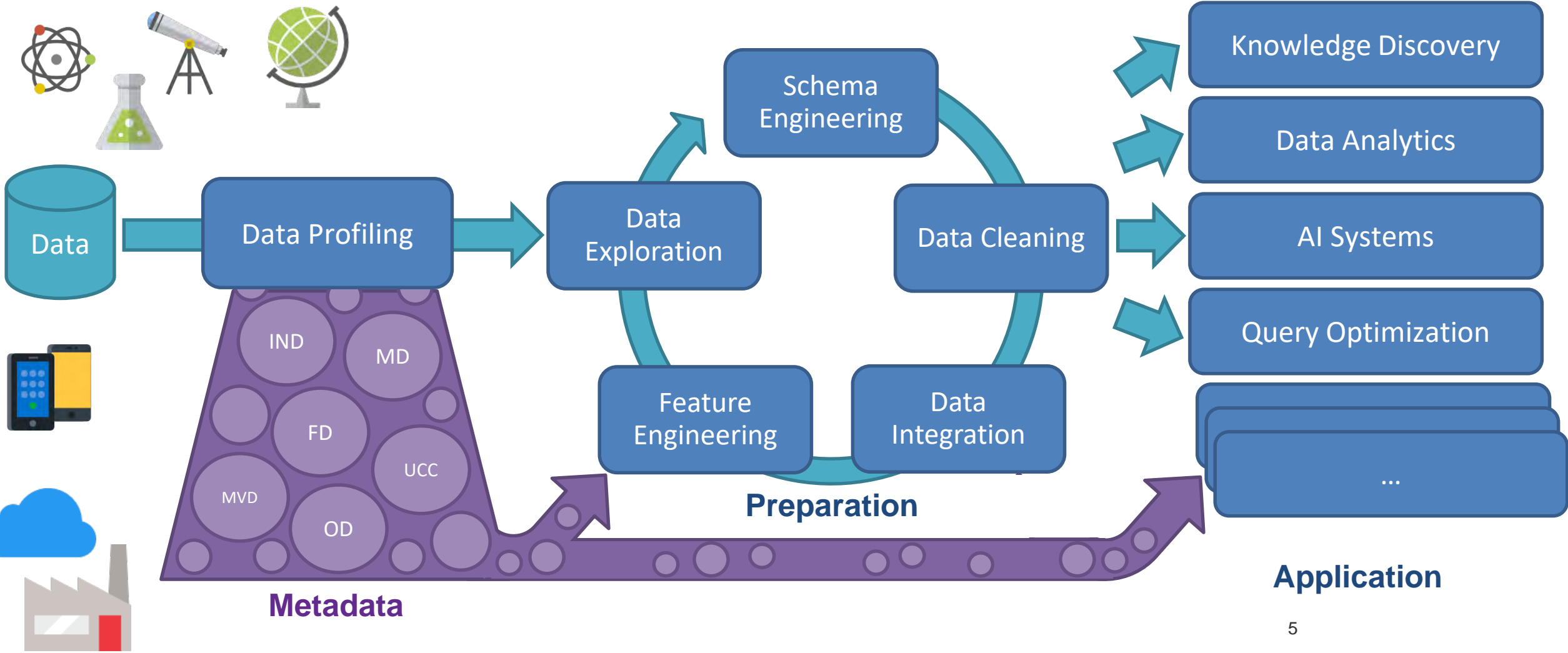
Which method explains better the causes behind a patient's anomalous condition?

Data Profiling for Data Engineering



Data Profiling for Data Engineering

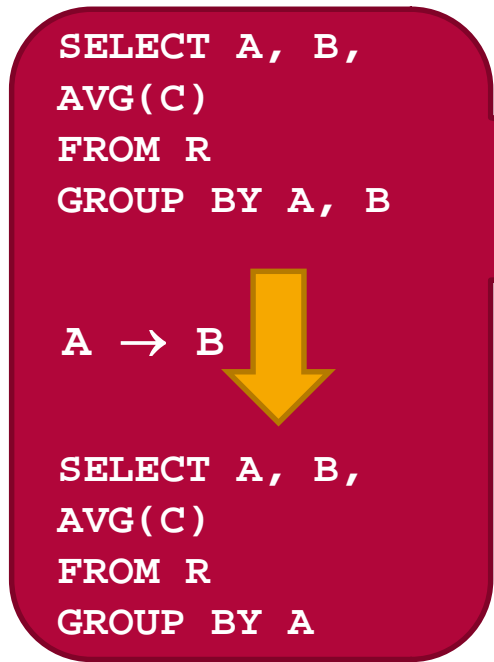
See here for
Open Research Questions



Use Case: Query Optimization

- 58 optimization opportunities

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



Application area	Unique Column Combinations (Sec. 5)	Functional Dependencies (Sec. 6)	Order Dependencies (Sec. 7)	Inclusion Dependencies (Sec. 8)
Join	<ul style="list-style-type: none"> Spurious-free back-joins [129] * Semijoin transformation [89] * Pipeline with grouping [30, 126] † Invisible join [2] † 	<ul style="list-style-type: none"> Simplification / avoidance [65] * Complexity reduction (Sec. 6.2) * Self-join avoidance [6] * Plan generation [38] † 	<ul style="list-style-type: none"> 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] † 	<ul style="list-style-type: none"> Join elimination [24, 64] * Substitute relations [33] † Avoid semijoin reductions (Sec. 8.1) † Accurate cardinalities [56] †
Selection	<ul style="list-style-type: none"> Early abort (Sec. 5.6) * Accurate cardinalities (Sec. 5.6) † 	<ul style="list-style-type: none"> Early abort (Sec. 6.3) * Substitute attributes [24, 68] † Estimations without independence assumption [25, 58, 108] † Reduce attributes [17, 116] * 	<ul style="list-style-type: none"> Use binary search [102] † 	
functions	<ul style="list-style-type: none"> Accurate cardinalities (Sec. 5.6) † 		<ul style="list-style-type: none"> Simplify MIN, MAX, MEDIAN [95, 102] * Sort-based grouping [102, 112, 117, 127, 128] † 	
Projection & Distinctness	<ul style="list-style-type: none"> Avoid DISTINCT [101, 102, 103] * 	<ul style="list-style-type: none"> Distinctness: see grouping [123] * Simplification [29] * Estimate projections [44] † 	<ul style="list-style-type: none"> Distinctness: See grouping † 	
Sorting	<ul style="list-style-type: none"> Reduce attributes (Sec. 5.6) * Unstable sorting (Sec. 5.6) * 	<ul style="list-style-type: none"> Reduce attributes [24, 112, 116] * 	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> EXCEPT to EXCEPT ALL [102] * INTERSECT to INTERSECT ALL [60, 101] * INTERSECT to join [101, 102] † Accurate cardinalities (Sec. 5.6) † 		<ul style="list-style-type: none"> Order optimizations [102] † 	<ul style="list-style-type: none"> Simplify UNION (Sec. 8.2) * Simplify INTERSECT (Sec. 8.2) * Eliminate EXCEPT (Sec. 8.2) * Accurate cardinalities (Sec. 8.2) †
Other	<ul style="list-style-type: none"> Subquery to join [101, 102] * Subquery sort avoidance [110] † 	<ul style="list-style-type: none"> Scalar subqueries [29] Table decomposition rewrites [45] 	<ul style="list-style-type: none"> Correlated subqueries [102] † Sparse over dense indexes [34] 	<ul style="list-style-type: none"> 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

Felix Naumann
Data Profiling
EDBT 2021

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

county	county_desc	voter_reg_ni	status_cd	voter_status_desc	reason_cd	voter_status	last_name	first_name	midl_name	name	res_street	address	res_city_desc	state	zip_code	mail_addr1	mail_addr2	mail_city	mail_state	mail_zipcod	full_phone	race_code	ethnic_code	party_cd
1	ALAMANCE	9005990	A	ACTIVE	AV	VERIFIED	AABEL	EVELYN	LARSEN		4430 E GREENSBOW	GRAHAM	NC	27253	4430 E GREENSBOW-CHA		GRAHAM	NC	27253	000 0000	W	NL	UNA	
2	ALAMANCE	9048723	A	ACTIVE	AV	VERIFIED	AARON	CHRISTINA	CASTAGNA		421 WHITT AVE	BURLINGTON	NC	27215	PO BOX 4177		BURLINGTON	NC	27215	229 1110	W	NL	UNA	
3	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	
4	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	
5	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	
6	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	
7	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	
8	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	
9	ALAMANCE	9117940	I	INACTIVE	IU	CONFIRMATI	ABADIE	PRISCILLA	MARIE		100 COLONNADE	ELON	NC	27244	CAMPUS BOX 3008		ELON	NC	27244		O	HL	UNA	
10	ALAMANCE	9034111	I	INACTIVE	IU	CONFIRMATI	ABADIE	COLLEEN	MIASHEL		1097 IVEY RD	#C GRAHAM	NC	27253	1097 IVEY RD		GRAHAM	NC	27253		M	HL	REP	
11	ALAMANCE	9034111	I	INACTIVE	IU	CONFIRMATI	ABADIE	JACK	EDWARD	JR	612 SIDEVIEW ST	GRAHAM	NC	27253	612 SIDEVIEW ST		GRAHAM	NC	27253	336 212 8140	W	NL	UNA	
12	ALAMANCE	9034111	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	
13	ALAMANCE	9034111	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	
14	ALAMANCE	9034111	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	
15	ALAMANCE	9034111	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	
16	ALAMANCE	9034111	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	
17	ALAMANCE	9034111	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	
18	ALAMANCE	9034111	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	
19	ALAMANCE	9034111	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	
20	ALAMANCE	9034111	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	
21	ALAMANCE	9034111	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	
22	ALAMANCE	9034111	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	
23	ALAMANCE	9034111	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	
24	ALAMANCE	9034111	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	
25	ALAMANCE	9034111	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	
26	ALAMANCE	9034111	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	
27	ALAMANCE	9034111	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	
28	ALAMANCE	9034111	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	
29	ALAMANCE	9034111	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	
30	ALAMANCE	9034111	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	
31	ALAMANCE	9034111	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	
32	ALAMANCE	9034111	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	
33	ALAMANCE	9034111	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	
34	ALAMANCE	9034111	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	
35	ALAMANCE	9034111	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	
36	ALAMANCE	9034111	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	
37	ALAMANCE	9034111	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	
38	ALAMANCE	9034111	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	
39	ALAMANCE	9034111	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	
40	ALAMANCE	9034111	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	
41	ALAMANCE	9034111	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	
42	ALAMANCE	9034111	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	
43	ALAMANCE	9034111	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	
44	ALAMANCE	9034111	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	
45	ALAMANCE	9034111	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	
46	ALAMANCE	9034111	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	
47	ALAMANCE	9034111	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	
48	ALAMANCE	9034111	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	
49	ALAMANCE	9034111	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	
50	ALAMANCE	9034111	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	

94 columns



manniling 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W						
106138	1	ALAMAN	9129972	A	ACTIVE	AV	VERIFIED	ZLUCHOWSK	AARON	MICHAEL	3551	FORESTDALE	BURLINGTON	NC	27215	3551	FORESTDALE	DR	#V	BURLINGTON	NC	27215	336	270	6878	W	NL	UNA	
106139	1	ALAMAN	9106623	A	ACTIVE	AV	VERIFIED	ZMIJEWSKI	SEAN		4872	THOM RD	MEBANE	NC	27302	4872	THOM RD			MEBANE	NC	27302	336	376	1987	O	UN	REP	
106140	1	ALAMAN	9112148	A	ACTIVE	AV	VERIFIED	ZMIJEWSKI	DENNIS	AL	4872	THOM RD	MEBANE	NC	27302	4872	THOM RD			MEBANE	NC	27302				W	UN	DEM	
106141	1	ALAMAN	9094109	I	INACTIVE	IU	CONFIRMATI	ZMIJEWSKI	DENNIS		4872	THOM RD	MEBANE	NC	27302	4872	THOM RD			MEBANE	NC	27302	336	376	1987	W	UN	DEM	
106142	1	ALAMAN	9128345	A	ACTIVE	AV	VERIFIED	ZMIJEWSKI	KEVIN	ADAM	4872	THOM RD	MEBANE	NC	27302	4872	THOM RD			MEBANE	NC	27302	336	380	5768	W	NL	UNA	
106143	1	ALAMAN	9120294	A	ACTIVE	AV	VERIFIED	ZMIJEWSKI	SEAN	CHRISTOPHE	4872	THOM RD	MEBANE	NC	27302	4872	THOM RD			MEBANE	NC	27302				W	HL	UNA	
106144	1	ALAMAN	9094116	A	ACTIVE	AV	VERIFIED	ZMIJEWSKI	N VIRGINIA	LOURDES	4872	THOM RD	MEBANE	NC	27302	4872	THOM RD			MEBANE	NC	27302	336	376	1987	U	UN	UNA	
106145	1	ALAMAN	9089250	R	REMOVED	RD	DECEASED	ZOCCOLANTI	TENIS	PIZZOTTI	2502	S NC HWY 119	MEBANE	NC	27302	2502	S NC HWY 119			MEBANE	NC	27302				W	UN	REP	
106146	1	ALAMAN	9083629	R	REMOVED	RD	DECEASED	ZOCCOLANTI	RENATO		3141	SHELLY GRAH	GRAHAM	NC	27253	3141	SHELLY GRAHAM DR			GRAHAM	NC	27253	336	227	7168	W	NL	REP	
106147	1	ALAMAN	9083630	A	ACTIVE	AV	VERIFIED	ZOCCOLANTI	RITA	MARIE	3141	SHELLY GRAH	GRAHAM	NC	27253	3141	SHELLY GRAHAM DR			GRAHAM	NC	27253	336	227	7168	W	NL	REP	
106148	1	ALAMAN	9100545	I	INACTIVE	IU	CONFIRMATI	ZOGLEMANN	ANGELA	LYNNE	706	HUFFMAN MIL	BURLINGTON	NC	27215	706	HUFFMAN MILL RD			#	BURLINGTON	NC	27215	336	227	1261	W	NL	UNA
106149	1	ALAMAN	9137285	A	ACTIVE	AV	VERIFIED	ZOLAYVAR	ERIC	WATSON	910	COLONIAL DR	BURLINGTON	NC	27215	910	COLONIAL DR				BURLINGTON	NC	27215	336	585	0248	O	NL	DEM
106150	1	ALAMAN	9081869	A	ACTIVE	AV	VERIFIED	ZOLAYVAR	RUPERTO	BENEDICTO	910	COLONIAL DR	BURLINGTON	NC	27215	910	COLONIAL DR				BURLINGTON	NC	27215	336	585	0248	O	NL	DEM
106151	1	ALAMAN	9109021	A	ACTIVE	AV	VERIFIED	ZOLAYVAR	STEPHANIE	WATSON	910	COLONIAL DR	BURLINGTON	NC	27215	910	COLONIAL DR				BURLINGTON	NC	27215	336	585	0248	W	NL	UNA
106152	1	ALAMAN	9108096	A	ACTIVE	AV	VERIFIED	ZOLLARS	EVELYN	NADINE	6830	TOM WOODY	SNOW CAMP	NC	27349	6830	TOM WOODY RD				SNOW CAMF	NC	27349	336	376	5754	W	NL	UNA
106153	1	ALAMAN	9125044	A	ACTIVE	AV	VERIFIED	ZOLLARS	MATHEW	DAVID	6830	TOM WOODY	SNOW CAMP	NC	27349	6830	TOM WOODY RD				SNOW CAMF	NC	27349				W	NL	UNA
106154	1	ALAMAN	9113912	A	ACTIVE	AV	VERIFIED	ZOLLICOFFEF	ANTONIO	MARK	108	OAKGROVE	DI GRAHAM	NC	27253	108	OAKGROVE DR				GRAHAM	NC	27253	336	260	6673	B	UN	DEM
106155	1	ALAMAN	9107068	A	ACTIVE	AV	VERIFIED	ZOLLICOFFEF	VALERIE		108	OAKGROVE	DI GRAHAM	NC	27253	108	OAKGROVE DR				GRAHAM	NC	27253				B	UN	DEM
106156	1	ALAMAN	9097324	A	ACTIVE	AV	VERIFIED	ZORNES	ASHLEY	DENICE	5556	N NC HWY 49	MEBANE	NC	27302	5556	N NC HWY 49				MEBANE	NC	27302	336	578	1157	W	NL	UNA
106157	1	ALAMAN	9038407	A	ACTIVE	AV	VERIFIED	ZORNES	KENNETH	ELWOOD	5556	N NC HWY 49	MEBANE	NC	27302	5556	N NC HWY 49				MEBANE	NC	27302				W	NL	UNA
106158	1	ALAMAN	9104969	I	INACTIVE	IU	CONFIRMATI	ZORNES	MICHELLE	LEE	3117	COMMERCE I	BURLINGTON	NC	27215	3117	COMMERCE PL			#L	BURLINGTON	NC	27215	336	675	0520	W	UN	UNA
106159	1	ALAMAN	9018738	A	ACTIVE	AV	VERIFIED	ZORNES	SHERRIE	AVERETTE	5556	N NC HWY 49	MEBANE	NC	27302	5556	N NC HWY 49				MEBANE	NC	27302				W	NL	DEM
106160	1	ALAMAN	9027412	I	INACTIVE	IU	CONFIRMATI	ZORNES	TERRY	LEE	148	N STATE ST	HAW RIVER	NC	27258	148	N STATE ST				HAW RIVER	NC	27258	570	1633		W	NL	DEM
106161	1	ALAMAN	9110367	D	DENIED	DU	VERIFICATIO	ZORNES	TINA		801	TROLLINGWO	HAW RIVER	NC	27258	801	TROLLINGWOOD RD				HAW RIVER	NC	27258	336	578	0646	W	UN	UNA
106162	1	ALAMAN	9132758	A	ACTIVE	AV	VERIFIED	ZORNES	TINA	MARIE	801	TROLLINGWO	HAW RIVER	NC	27258	801	TROLLINGWOOD RD				HAW RIVER	NC	27258	336	420	7630	W	NL	UNA
106163	1	ALAMAN	9131499	A	ACTIVE	AV	VERIFIED	ZOUFALY	EVE		602	E HAGGARD	A ELON	NC	27244	CAMPUS BOX	8911			ELON	NC	27244				U	UN	UNA	
106164	1	ALAMAN	9124446	A	ACTIVE	AV	VERIFIED	ZSUPPAN	ETELKA	HALASZ	1929	HAW VILLAG	GRAHAM	NC	27253	1929	HAW VILLAGE DR				GRAHAM	NC	27253				W	NL	REP
106165	1	ALAMAN	9121554	A	ACTIVE	AV	VERIFIED	ZSUPPAN	FERENC		1929	HAW VILLAG	GRAHAM	NC	27253	1929	HAW VILLAGE DR				GRAHAM	NC	27253				W	UN	REP
106166	1	ALAMAN	9127457	A	ACTIVE	AV	VERIFIED	ZSUPPAN	LEVENTE	FERENC	1929	HAW VILLAG	GRAHAM	NC	27253	1929	HAW VILLAGE DR				GRAHAM	NC	27253	336	376	1365	W	NL	REP
106167	1	ALAMAN	9131401	A	ACTIVE	AV	VERIFIED	ZUBLER	LINDSAY	BROOKE	3172	CARRIAGE CF	HAW RIVER	NC	27258	3172	CARRIAGE CREEK CT				HAW RIVER	NC	27258				U	UN	UNA
106168	1	ALAMAN	9081728	A	ACTIVE	AV	VERIFIED	ZUBLER	TAMI	LAJEAN	3172	CARRIAGE CF	HAW RIVER	NC	27258	3172	CARRIAGE CREEK CT				HAW RIVER	NC	27258	336	578	8028	W	NL	UNA
106169	1	ALAMAN	9089569	A	ACTIVE	AV	VERIFIED	ZUBLER	TIMOTHY	JAMES	3172	CARRIAGE CF	HAW RIVER	NC	27258	3172	CARRIAGE CREEK CT				HAW RIVER	NC	27258				W	UN	UNA
106170	1	ALAMAN	9070674	A	ACTIVE	AV	VERIFIED	ZUBOV	ALEX		229	ENGLEMAN A	BURLINGTON	NC	27215	229	ENGLEMAN AVE				BURLINGTON	NC	27215	336	437	9776	W	NL	UNA
106171	1	ALAMAN	9070288	A	ACTIVE	AV	VERIFIED	ZUBOV	LYNN	R	229	ENGLEMAN A	BURLINGTON	NC	27215	229	ENGLEMAN AVE				BURLINGTON	NC	27215	336	437	9776	W	NL	REP
106172	1	ALAMAN	9008787	A	ACTIVE	AV	VERIFIED	ZUMER	FRANK	EDWARD	801	QUAKER RIDG	MEBANE	NC	27302	801	QUAKER RIDGE RD				MEBANE	NC	27302	919	563	3766	W	UN	UNA
106173	1	ALAMAN	9008785	A	ACTIVE	AV	VERIFIED	ZUMER	LOUISE	TURNER	801	QUAKER RIDG	MEBANE	NC	27302	801	QUAKER RIDGE RD				MEBANE	NC	27302	919	563	3766	W	NL	DEM
106174	1	ALAMAN	9141817	A	ACTIVE	AV	VERIFIED	ZUNG	PATRICK	BATE	2604	WOODS LN	GRAHAM	NC	27253	2604	WOODS LN				GRAHAM	NC	27253	919	357	3896	W	NL	DEM
106175	1	ALAMAN	9119438	A	ACTIVE	AV	VERIFIED	ZUNIGA	JOSE	RAMON SALI	714	ROSS ST	BURLINGTON	NC	27217	714	ROSS ST				BURLINGTON	NC	27217	336	227	3108	O	HL	DEM
106176	1	ALAMAN	9108610	A	ACTIVE	AV	VERIFIED	ZUNIGA	VANESA	ELIZABETH	512	PIEDMONT W	BURLINGTON	NC	27217	512	PIEDMONT WAY				BURLINGTON	NC	27217	336	270	0181	W	HL	DEM
106177	1	ALAMAN	9112637	A	ACTIVE	AV	VERIFIED	ZUNIGA	YANET	SALAS	3845	MAE DOUGL	MEBANE	NC	27302	3845	MAE DOUGLAS DR				MEBANE	NC	27302				O	HL	DEM
106178	1	ALAMAN	9141392	A	ACTIVE	AV	VERIFIED	ZUPANCICH	MONICA	ANITA	2326	N NC HWY 49	BURLINGTON	NC	27217	2326	N NC HWY 49				BURLINGTON	NC	27217	330	310	0151	W	NL	REP
106179	1	ALAMAN	9141392	A	ACTIVE	AV	VERIFIED	ZUPANCICH	RONALD	JAMES	2326	N NC HWY 49	BURLINGTON	NC	27217	2326	N NC HWY 49				BURLINGTON	NC	27217	757	254	3773	W	NL	REP
106180	1	ALAMAN	9141392	A	ACTIVE	AV	VERIFIED	ZURFACE	ROSSELL	EUGENE	2074	TURNER RD	MEBANE	NC	27302	2074	TURNER RD				MEBANE	NC	27302				W	UN	UNA
106181	1	ALAMAN	9141392	A	ACTIVE	AV	VERIFIED	ZWIER	ANDREW	MICHAEL	1497	LONGEST AC	SNOW CAMP	NC	27349	1497	LONGEST ACRES RD				SNOW CAMF	NC	27349	336	376	8830	W	NL	REP
106182	1	ALAMAN	9141392	A	ACTIVE	AV	VERIFIED	ZWIER	CHRISTOPHE	ANTHONY	1497	LONGEST AC	SNOW CAMP	NC	27349	1497	LONGEST ACRES RD				SNOW CAMF	NC	27349	831	207	9222	W	NL	REP
106183	1	ALAMAN	9141392	A	ACTIVE	AV	VERIFIED	ZWIER	CHRISTY	ANN	1497	LONGEST AC	SNOW CAMP	NC	27349	1497	LONGEST ACRES RD				SNOW CAMF	NC	27349				W	NL	REP
106184	1	ALAMAN	9099261	A	ACTIVE	AV	VERIFIED	ZWIER	KAREN	JEAN	1497	LONGEST AC	SNOW CAMP	NC	27349	1497	LONGEST ACRES RD				SNOW CAMF	NC	27349	831	207	9222	W	NL	REP
106185	1	ALAMAN	9077804	R	REMOVED	RL	MOVED	FROI ZYLKA	MARC		1210	WILLOW BRC	MEBANE	NC	27302	1210	WILLOW BROOK CT				MEBANE	NC	27302	336	578	8580	W	UN	REP
106186																													



106,185 rows

nann
ling
1

Table with columns: ID, Name, Address, City, State, Zip, Phone, Email, etc. The table contains a large volume of data rows, many of which are partially obscured by a large red arrow pointing to the right with the text 'patterns' written on it.

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

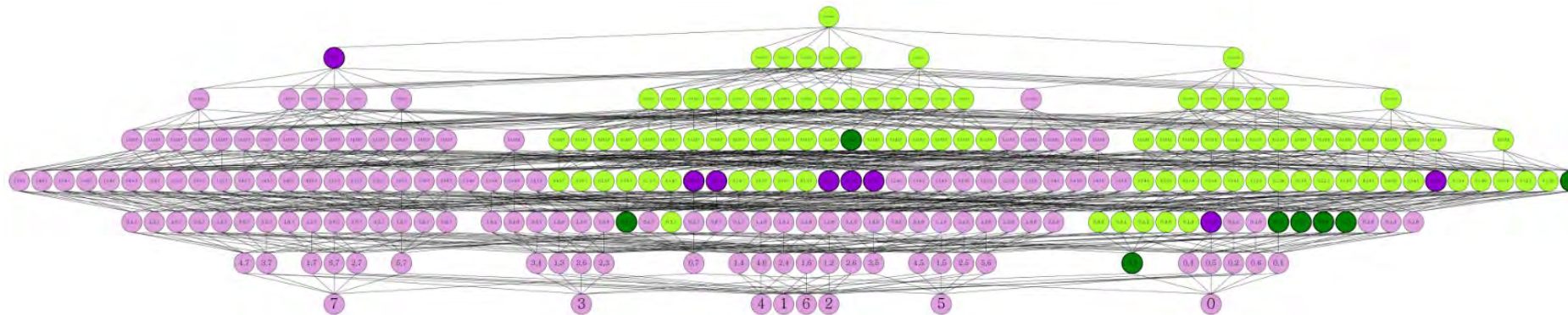
- Cardinalities
- Uniqueness
 - Key discovery
- Patterns and data types
- Distributions
- Domain Classification
- ...

Multi-column tasks

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

Scalable Profiling

- 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)



Felix Naumann
Data Profiling
EDBT 2021

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

Patterns and types

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



Cardinalities

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

Value distributions

- histograms
- constancy
- quartiles
- soundex
- first digit

Benford Law Frequency ("first digit law")

■ 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

THE LAW OF ANOMALOUS NUMBERS

FRANK BENFORD

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

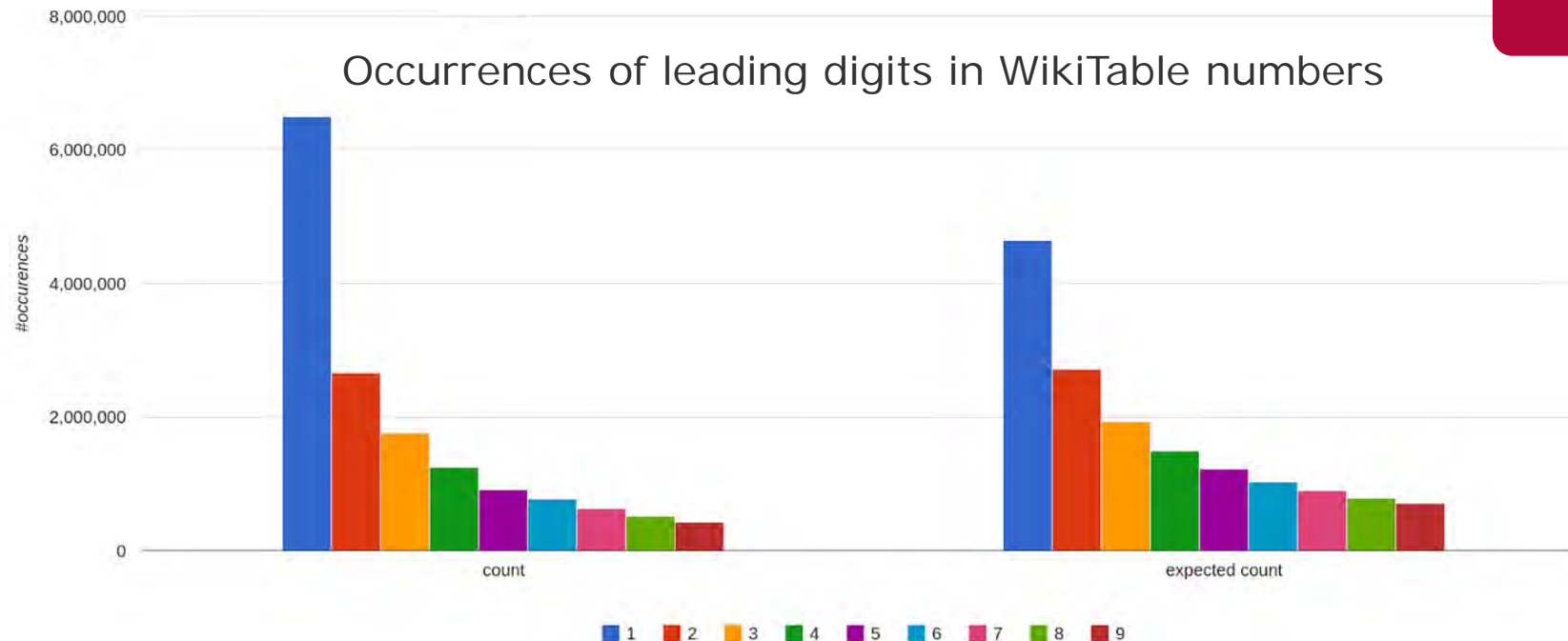
(Introduced by Irving Langmuir)

(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

Since 1971 "American Men and Women of Science"



Felix Naumann
Data Profiling
EDBT 2021

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

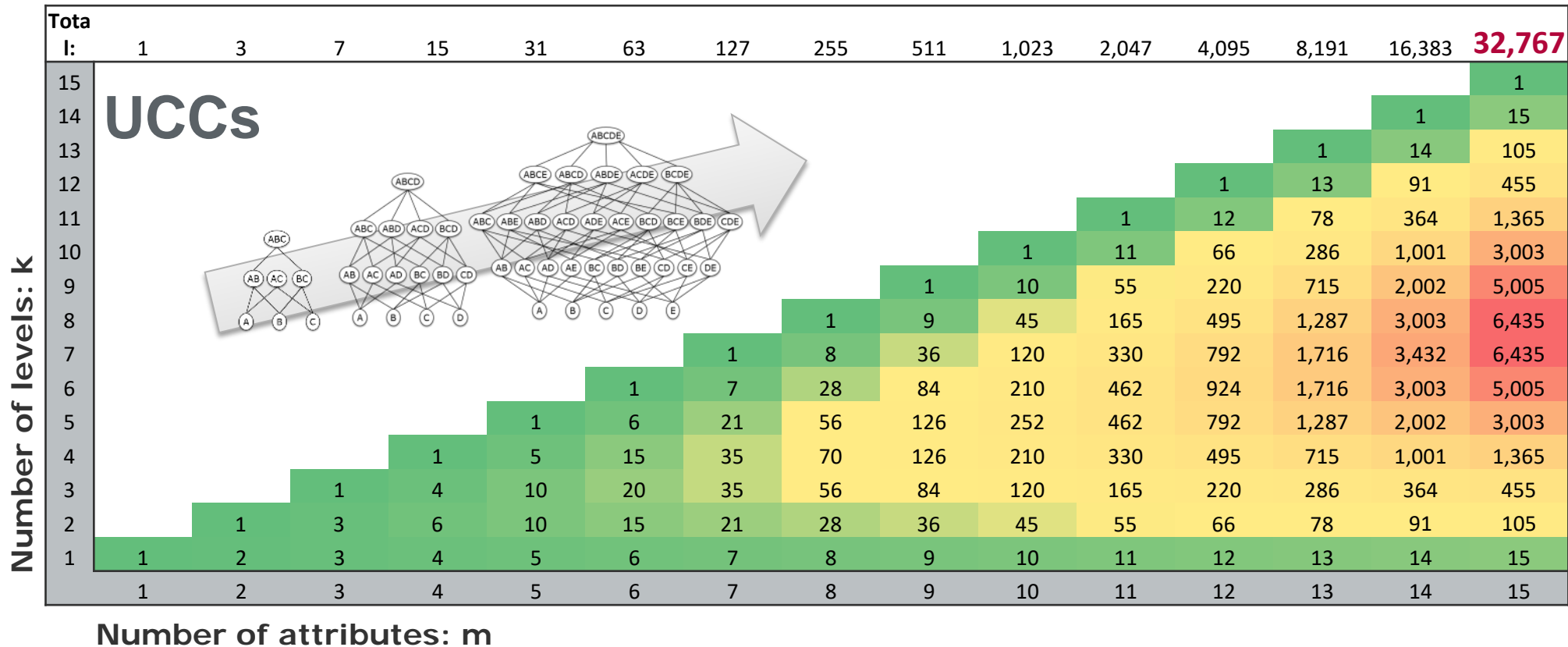
- **Unique column combination (UCC)**

No pair of records has same value combination when projected to those columns

- **A minimal unique:** `voter_reg_num, zip_code, race_code`

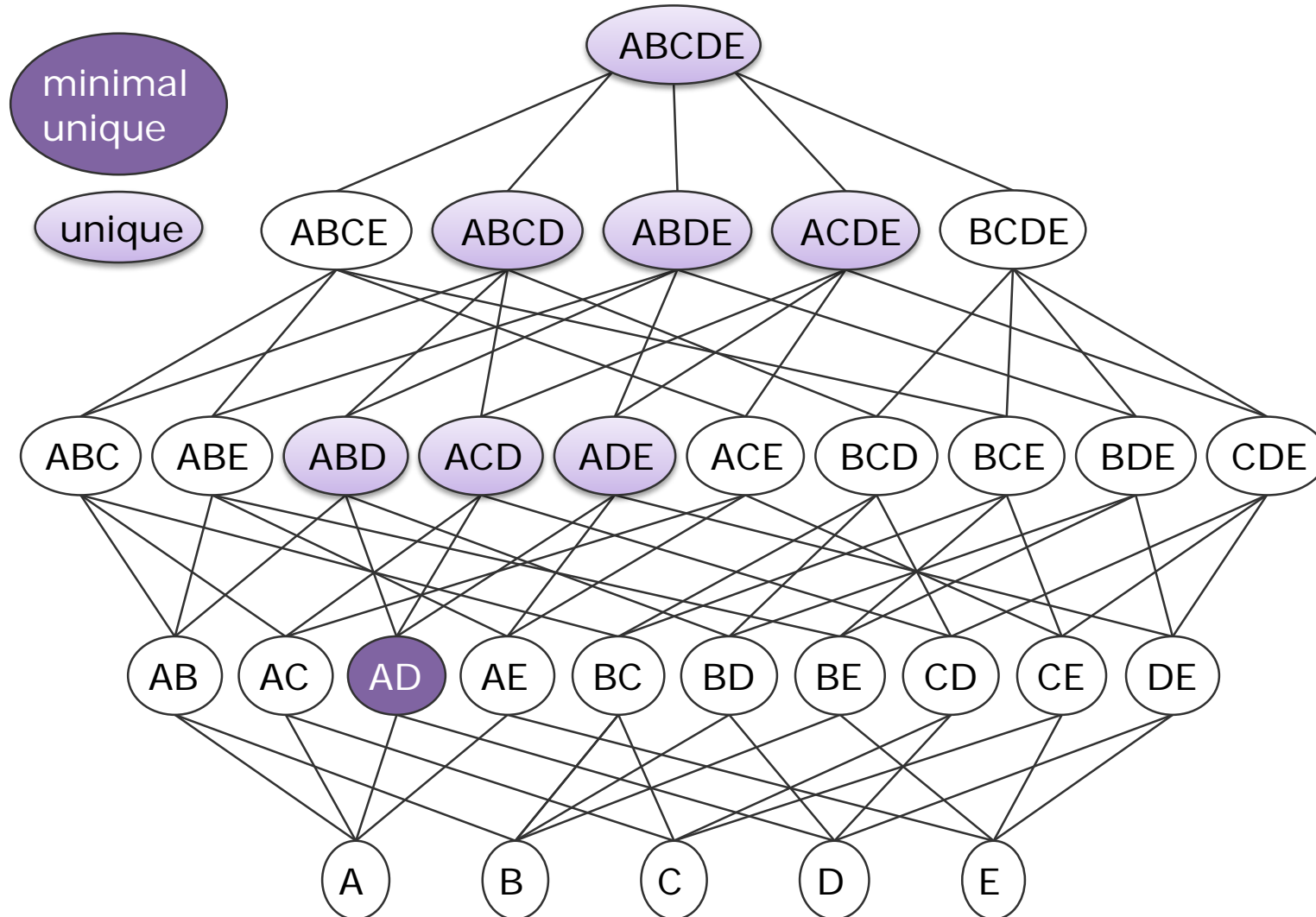
- **A maximal non-unique:** `voter_reg_num, status_cd, voter_status_desc, reason_cd, voter_status_reason_desc, absent_ind, name_prefix_cd, name_sufx_cd, half_code, street_dir, street_type_cd, street_sufx_cd, unit_designator, unit_num, state_cd, mail_addr2, mail_addr3, mail_addr4, mail_state, area_cd, 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_abbrev, precinct_desc, municipality_abbrev, municipality_desc, ward_abbrev, ward_desc, cong_dist_abbrev, cong_dist_desc, super_court_abbrev, super_court_desc, judic_dist_abbrev, judic_dist_desc, nc_senate_abbrev, nc_senate_desc, nc_house_abbrev, nc_house_desc, county_commiss_abbrev, county_commiss_desc, township_abbrev, township_desc, school_dist_abbrev, school_dist_desc, fire_dist_abbrev, fire_dist_desc, water_dist_abbrev, water_dist_desc, sewer_dist_abbrev, sewer_dist_desc, sanit_dist_abbrev, sanit_dist_desc, rescue_dist_abbrev, rescue_dist_desc, munic_dist_abbrev, munic_dist_desc, dist_1_abbrev, dist_1_desc, dist_2_abbrev, dist_2_desc, confidential_ind, age, vtd_abbrev, vtd_desc`

Candidate Set Growth for UCCs



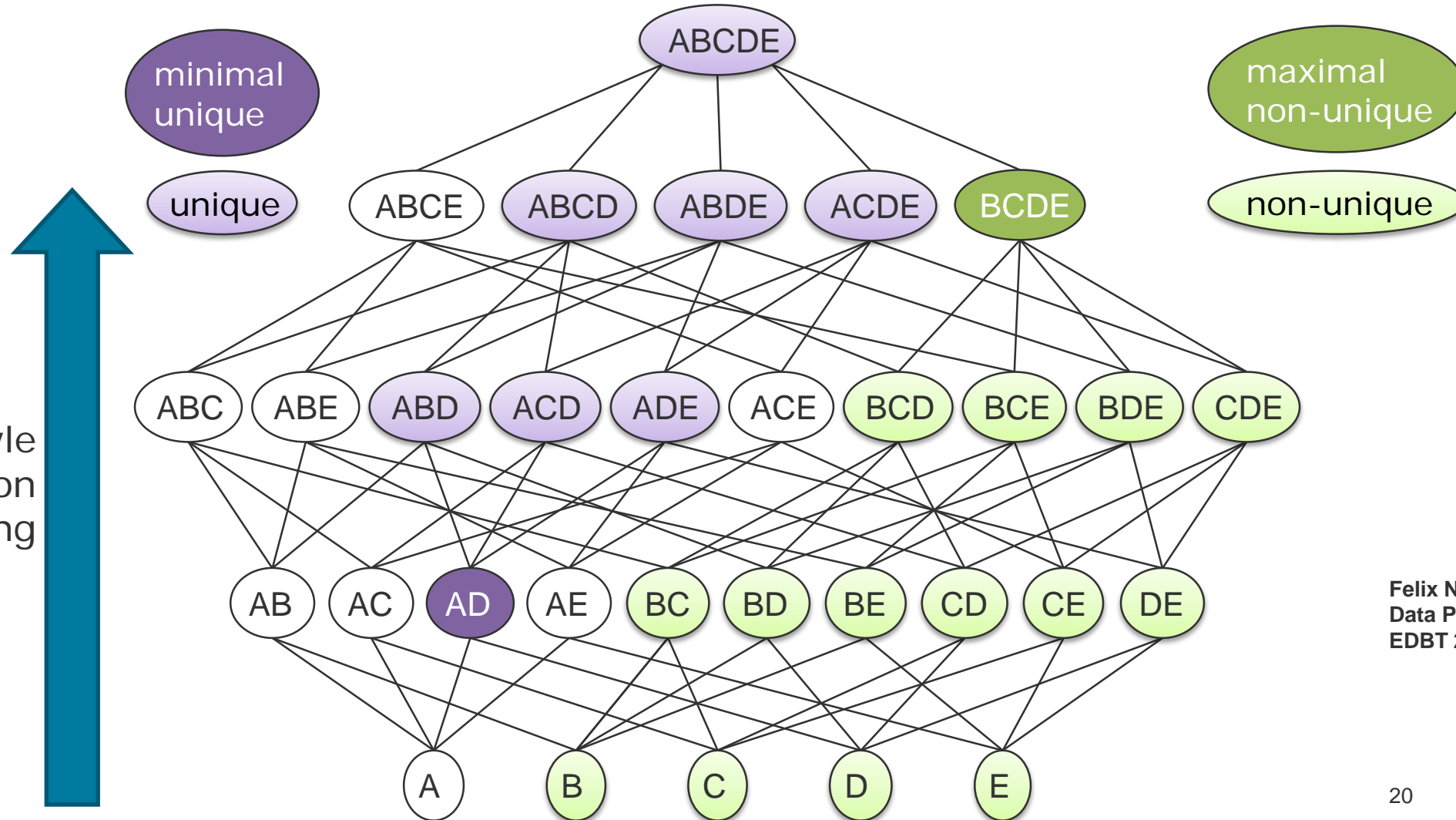
Felix Naumann
Data Profiling
EDBT 2021

Pruning Supersets



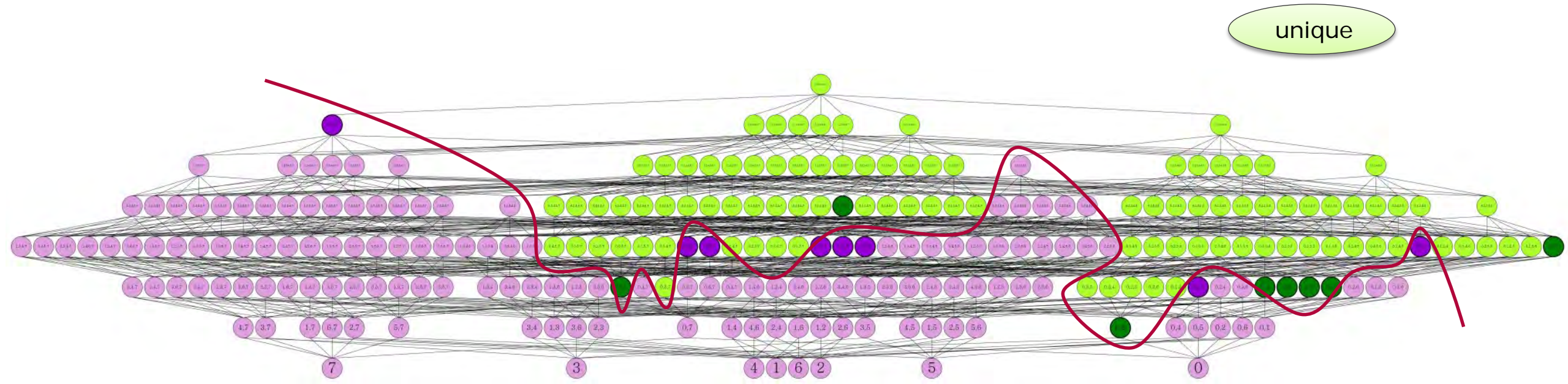
Pruning Subsets

Apriori-style enumeration and pruning



Open Research Question
What are effective distribution strategies for data profiling?

DUCC – Detecting Unique Column Combinations



unique

non-unique

Felix Naumann
Data Profiling
EDBT 2021

Discovering Functional Dependencies

- „ $X \rightarrow A$ “
When two tuples have same value in attribute set X , they must have same values in attribute A .
 - E.g., $\text{ZIP} \rightarrow \text{City}$

Naïve discovery approach

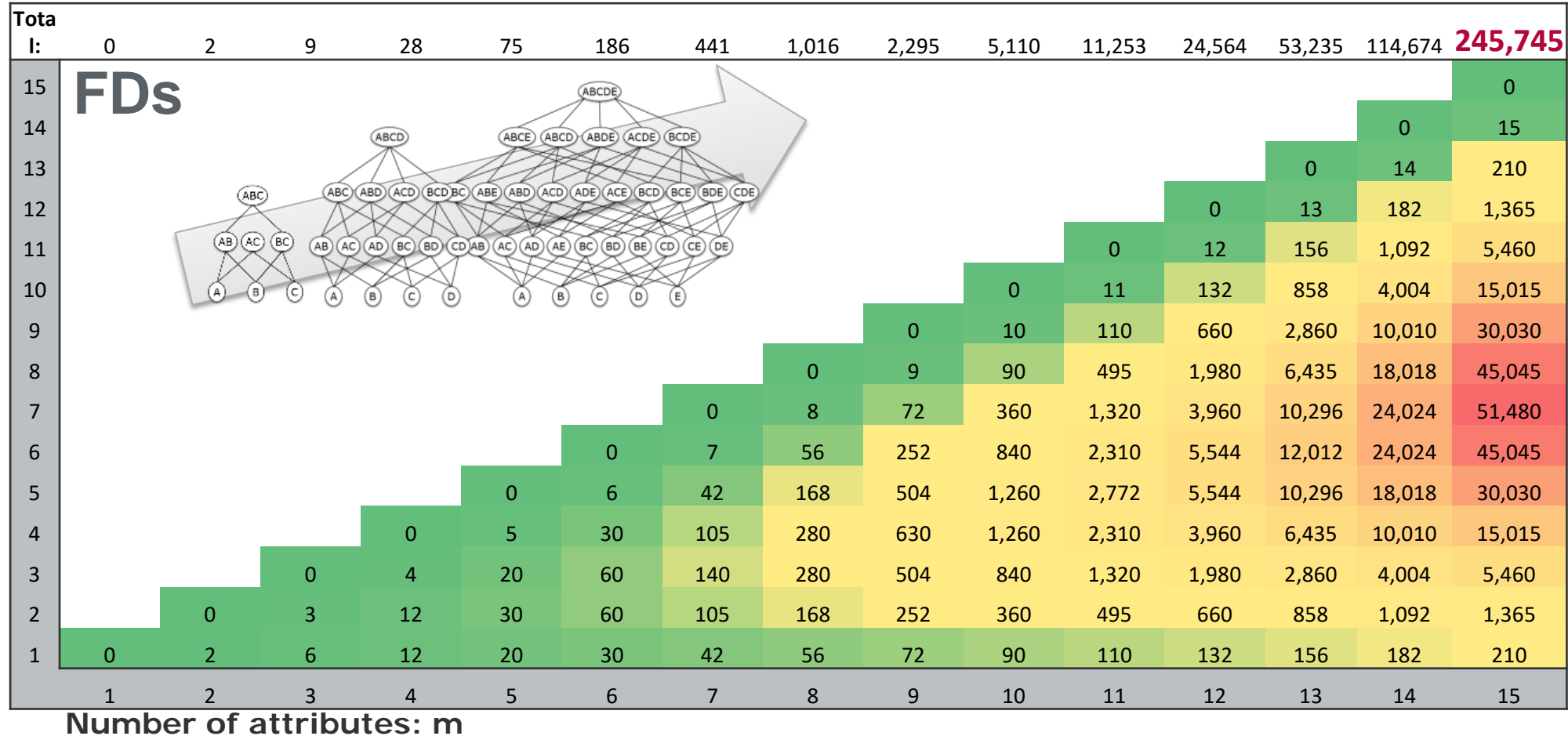
For each column combination X

For each $A \in X$

For each pair of tuples (t_1, t_2)

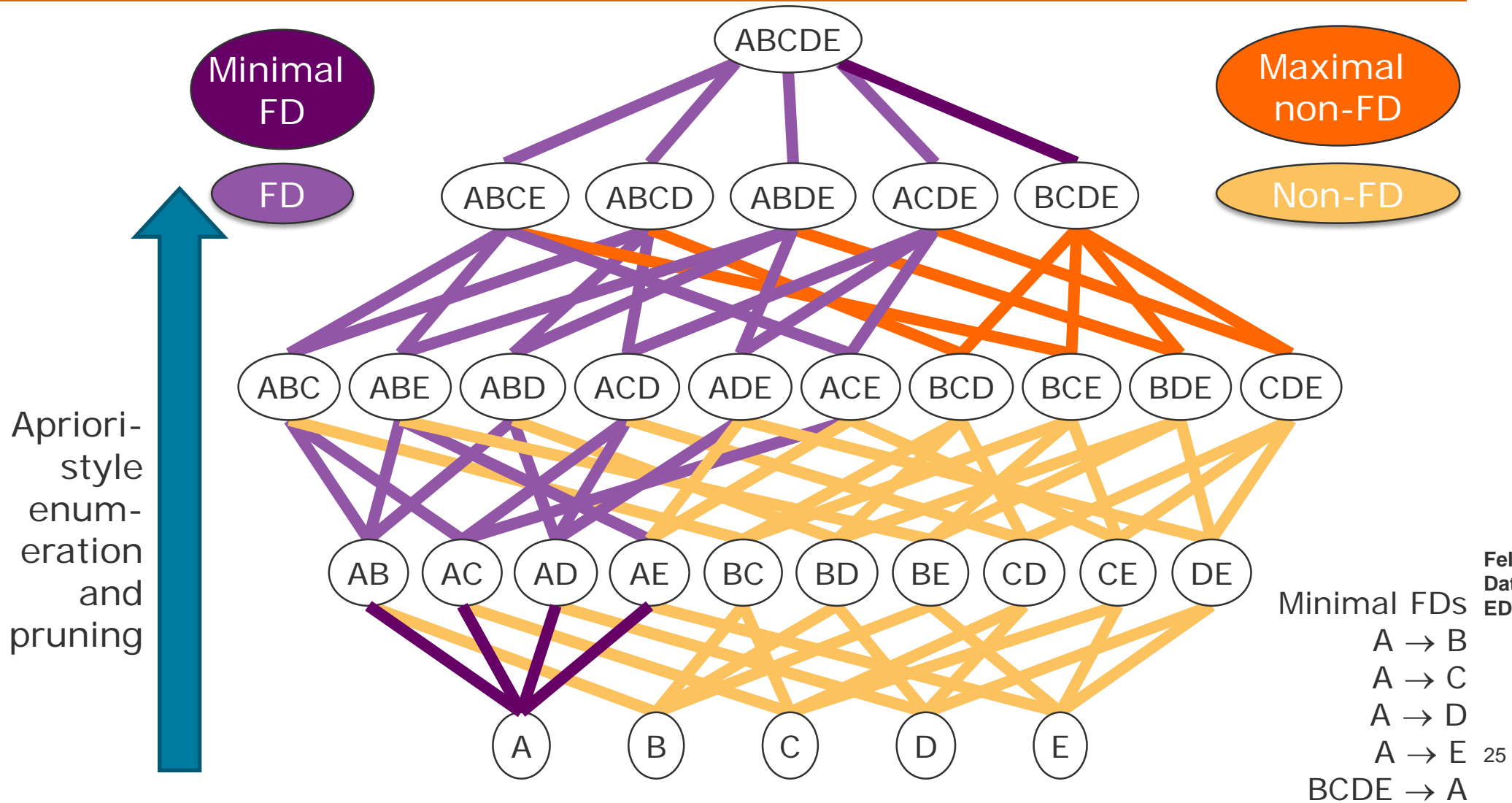
If $t_1[X \setminus A] = t_2[X \setminus A]$ and $t_1[A] \neq t_2[A]$: Break

Candidate Set Growth for FDs



Felix Naumann
Data Profiling
EDBT 2021

Model in Lattice – Edges Represent FDs



Row-based Algorithms

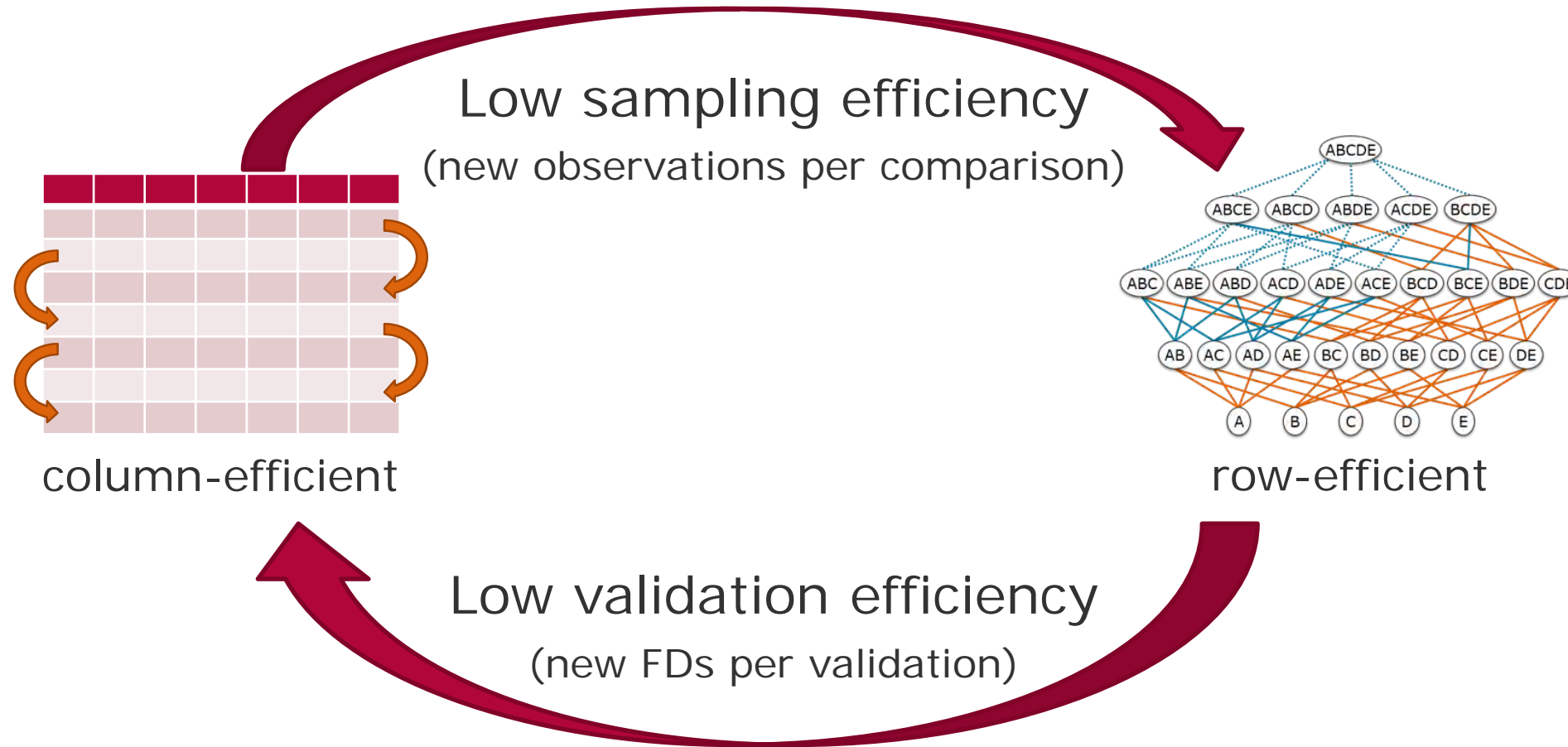
<u>Name</u>	<u>Surname</u>	Postcode	City	Mayor
Thomas	Miller	14482	Potsdam	Jakobs
Sarah	Miller	14482	Potsdam	Jakobs
Peter	Smith	60329	Frankfurt	Feldmann
Jasmine	Cone	01069	Dresden	Orosz
Thomas	Cone	14482	Potsdam	Jakobs
Mike	Moore	60329	Frankfurt	Feldmann

- Surname, Postcode, City, Mayor $\not\rightarrow$ Name
- Name, Postcode, City, Mayor $\not\rightarrow$ Surname
- Surname $\not\rightarrow$ Name, Postcode, City, Mayor



Postcode \rightarrow City
 Postcode \rightarrow Mayor
 Name \rightarrow Surname, ...

HyFD: Hybrid FD Discovery



Felix Naumann
Data Profiling
EDBT 2021

Functional Dependencies: State of the Art

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

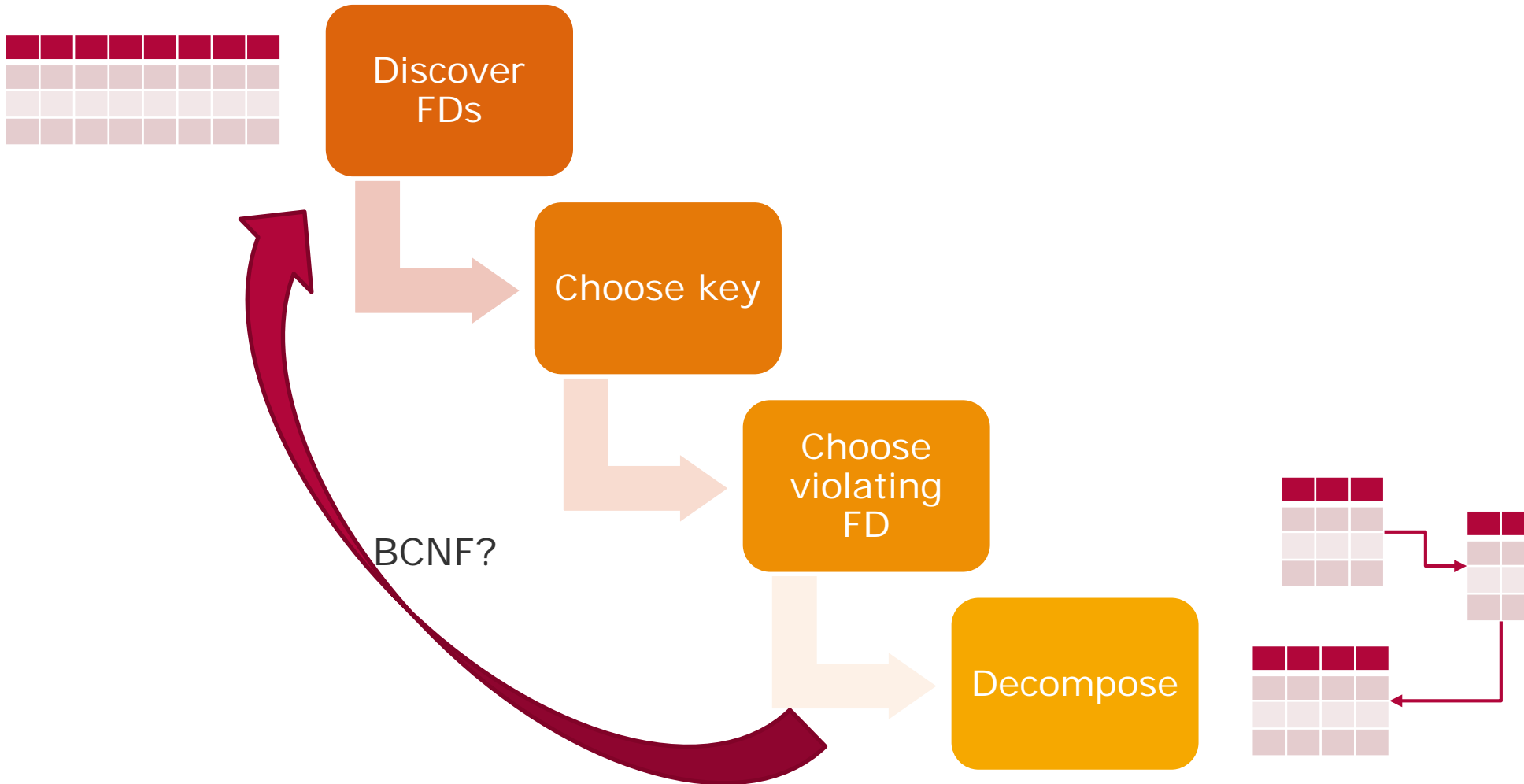
Results larger than 1,000 FDs are only counted

TL: time limit of 4 hours exceeded

ML: memory limit of 100 GB exceeded

Felix Naumann
Data Profiling
EDBT 2021

Use case: BCNF Normalization



Felix Naumann
Data Profiling
EDBT 2021

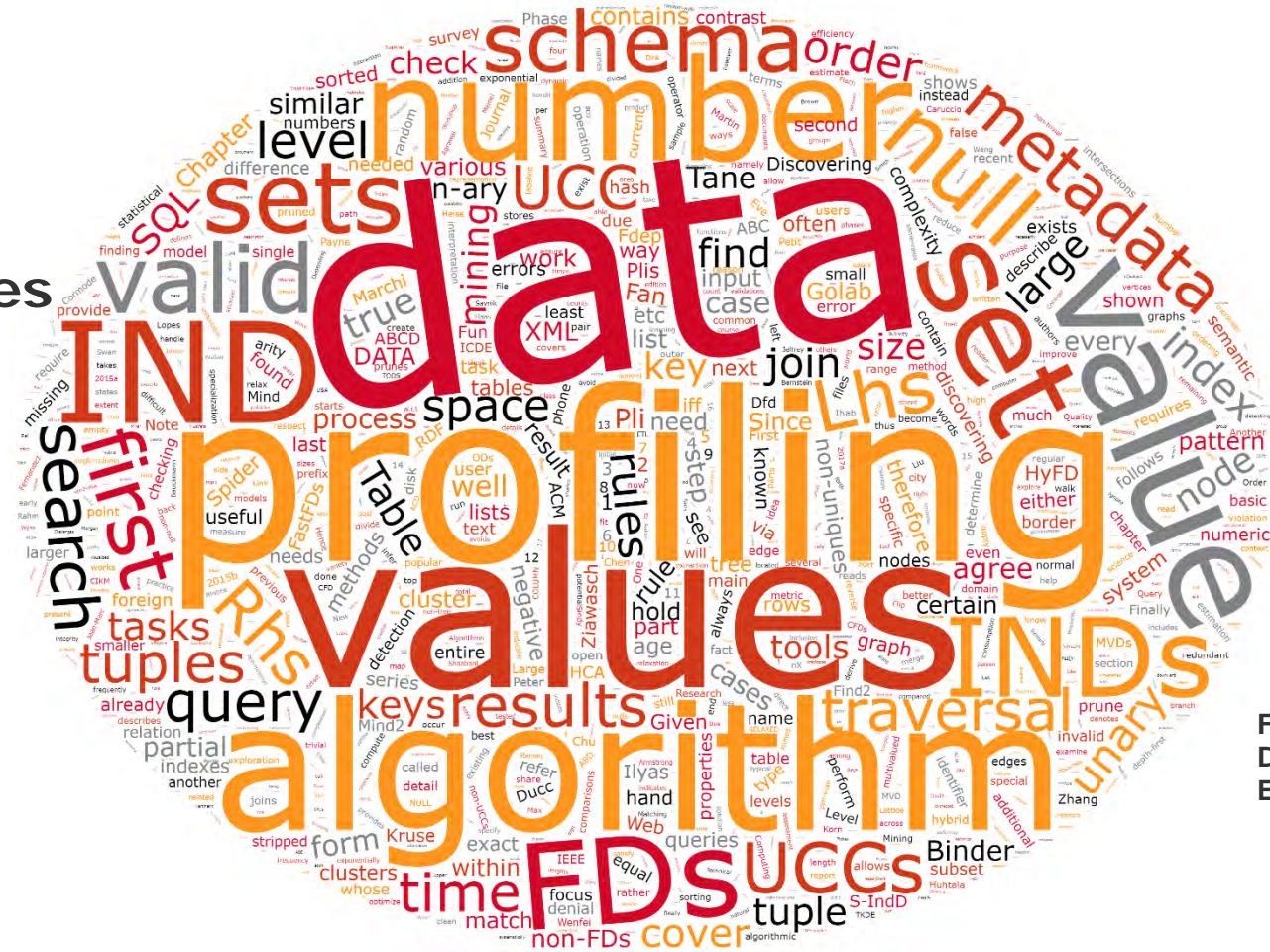
Normalization Results: TPC-H

(<u>linenumber</u> , extendedprice, discount, tax, returnflag, shipdate, commitdate, receiptdate, comment, <u>orderkey</u> , partkey)	LINEITEM
→ (<u>linenumber</u> , <u>extendedprice</u> , <u>tax</u> , <u>commitdate</u> , <u>receiptdate</u> , shipinstruct)	
→ (<u>extendedprice</u> , <u>discount</u> , shipmode, <u>orderkey</u>)	
→ (quantity, <u>extendedprice</u> , <u>partkey</u>)	
→ (linestatus, <u>shipdate</u>)	
→ (<u>tax</u> , <u>returnflag</u> , <u>orderkey</u> , <u>partkey</u> , suppkey)	
↳ (<u>availqty</u> , <u>supplycost</u> , comment, <u>partkey</u> , <u>suppkey</u>)	PARTSUPP
↳ (<u>partkey</u> , name, brand, type, size, container, retailprice, comment)	PART
↳ (mfgr, <u>brand</u>)	
↳ (<u>suppkey</u> , name, address, phone, acctbal, comment, nationkey)	SUPPLIER
↳ (<u>nationkey</u> , name, comment, regionkey)	NATION
↳ (shippriority, <u>regionkey</u> , name, comment)	REGION
→ (<u>orderkey</u> , totalprice, orderdate, orderpriority, clerk, comment, custkey)	ORDERS
↳ (orderstatus, <u>totalprice</u> , <u>orderdate</u>)	
↳ (<u>custkey</u> , name, address, phone, acctbal, mktsegment, comment)	CUSTOMER

Felix Naumann
Data Profiling
EDBT 2021

Agenda

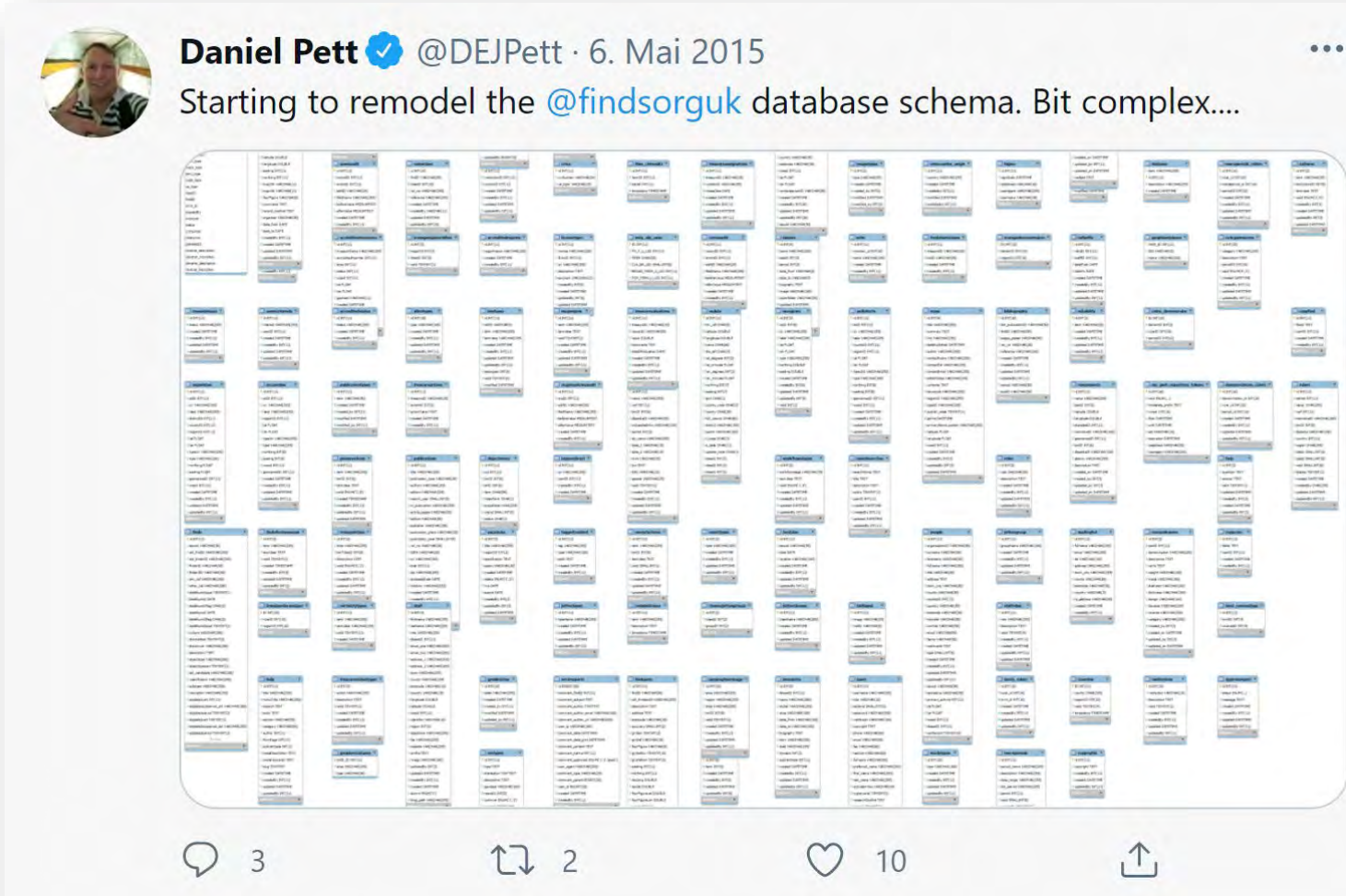
1. Basic statistics
2. Uniques and keys
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. New directions



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

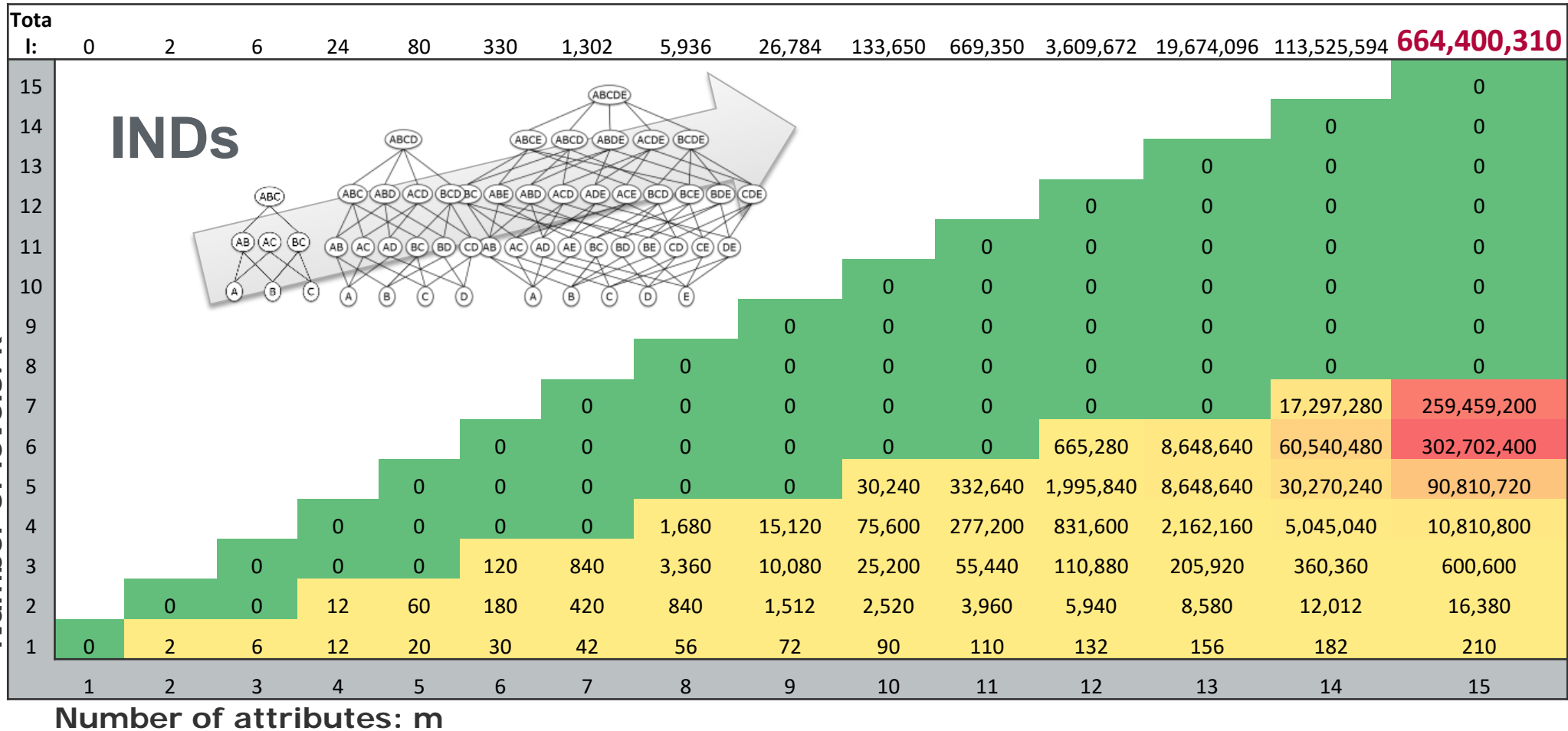


Daniel Pett @DEJPett · 6. Mai 2015

Starting to remodel the [@findsorguk](#) database schema. Bit complex...

3 2 10

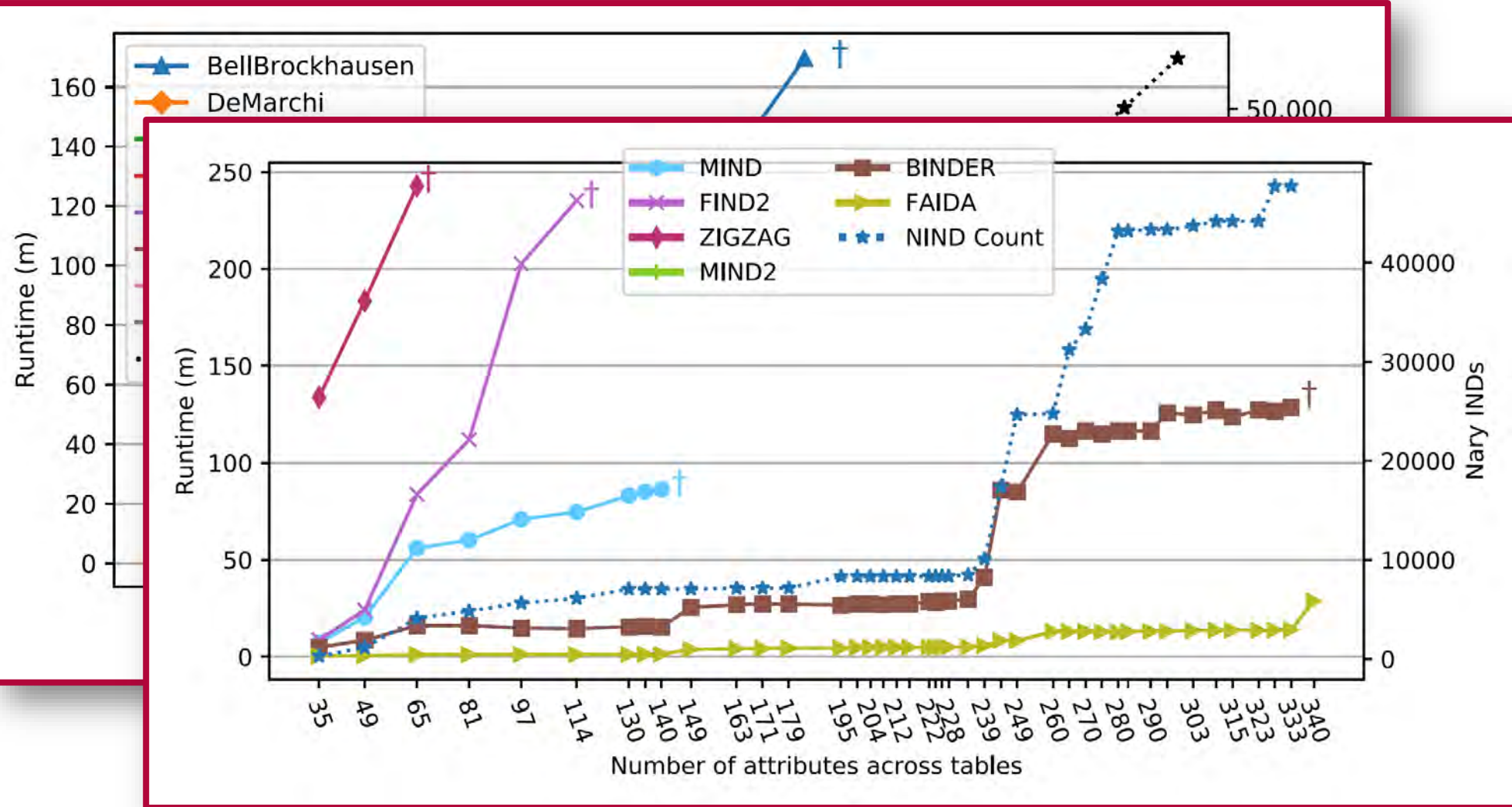
Candidate Set Growth for INDs



Felix Naumann
Data Profiling
EDBT 2021

Open Research Question
 How to efficiently discover all n-ary INDs?

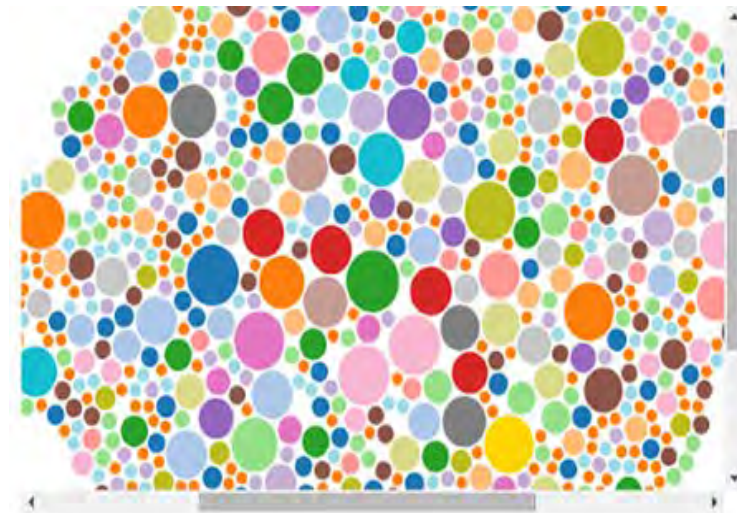
IND Detection Algorithms



Felix Naumann
 Data Profiling
 EDBT 2021

Use Case: Web Data Integration

Discovering INDs among millions of web tables



96242-1	96242-1.'Rotational_period_of_selected_objects'.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_astrolgy'. 'Ruling planets of the astrological signs and houses'.csv
175797-1	175797-1.'Sun_sign_astrolgy'. '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-	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 latitude)	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

Are there not more types of dependencies? Detecting Other Dependencies

- 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

- 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

■ Features for UCCs as keys

- „ID“, „PK“, etc. in name
- Few columns, short data types
- Early in schema
- Serves as reference

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

■ Features for INDs as foreign keys

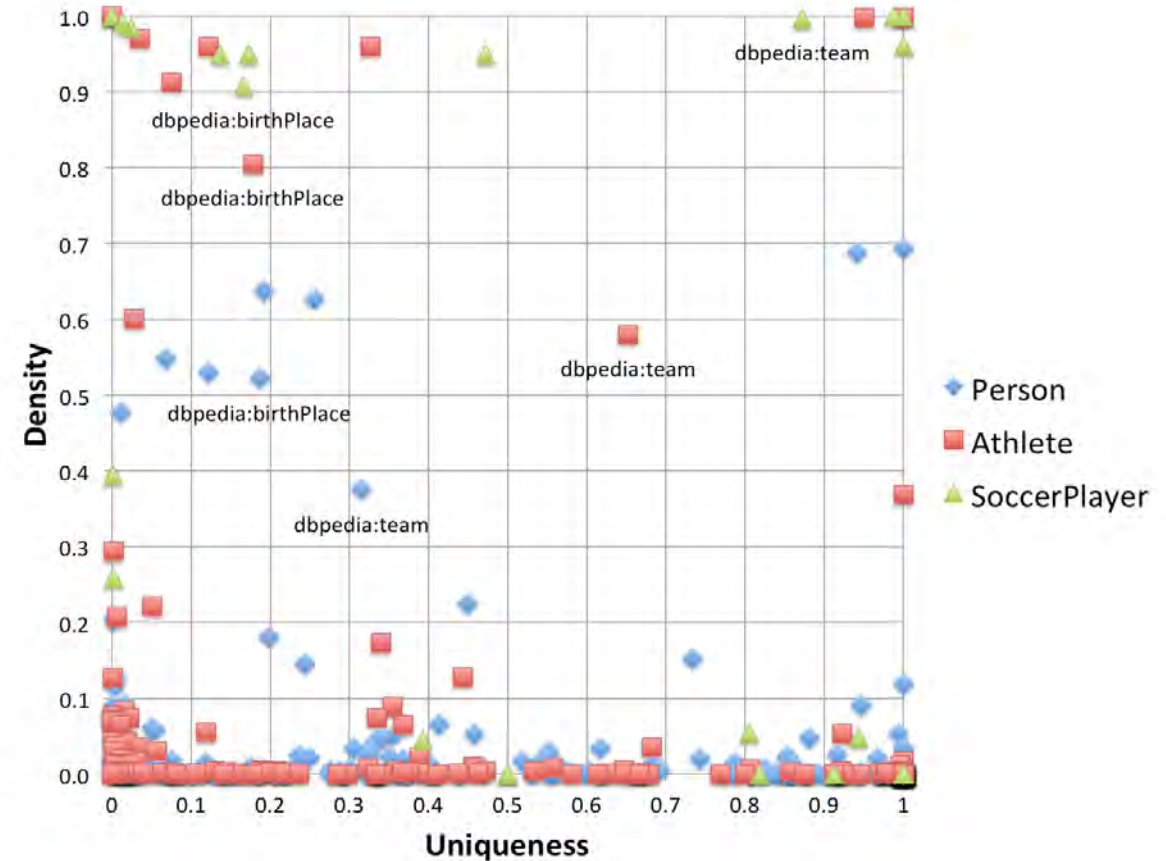
- „FK“, „ID“, etc. in name
- Referenced columns are a UCC or key
- Random or even distribution of values
- ...

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

Felix Naumann
Data Profiling
EDBT 2021

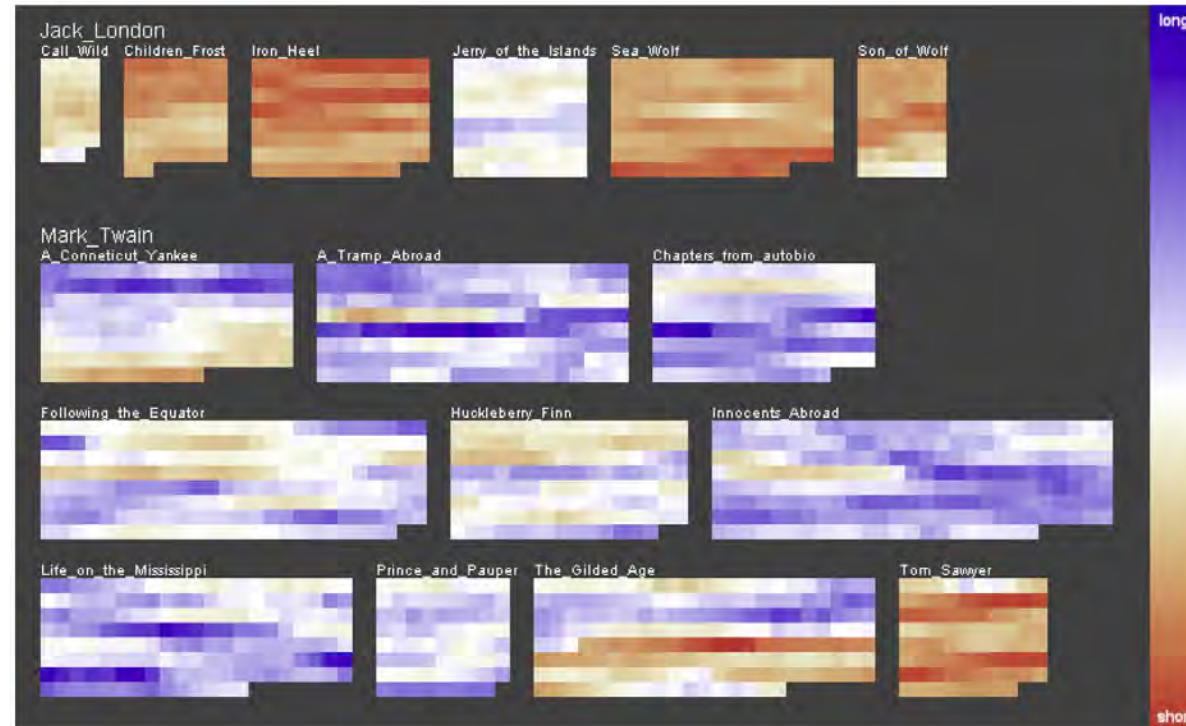
Profiling for Other Data Models

- 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

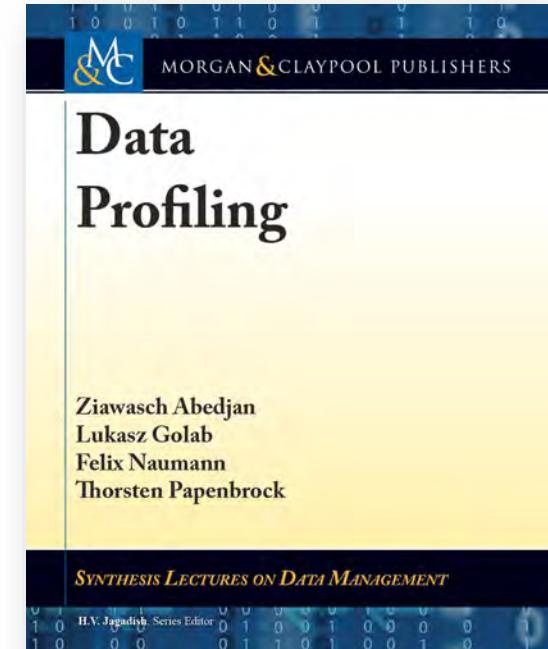
- Images and videos
 - Color, video-length, volume, etc.
- Textual data
 - Statistical measures
 - Syllables per word
 - Sentence length
 - Parts of speech
 - Vocabulary measures
 - Frequencies of specific words
 - Simpson's index
 - Content
 - Sentiment



„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)

Felix Naumann
Data Profiling
EDBT 2021

- 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

Current focus on **semantics**

- Keys and foreign key
- Use cases

Current focus on **new problems**

- Relaxed dependencies
- New dependency types

Future focus on **efficiency**

- Distribution
- Profiling dynamic data
- Query optimization

Future focus on **semantics**

- Genuineness
- Data cleaning

Future focus on **new problems**

- New data models

Open Research Questions

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 expressions** 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?