(Big) Data Profiling

January, 2017

Felix Naumann

The Hasso Plattner Institute

Felix Naumann
Data Profiling
Canada, 2017

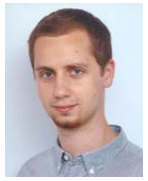
Information Systems Team

http://www.hpi.de/naumann/home.html

Thorsten **Papenbrock**

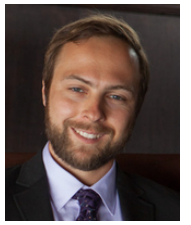
Diana **Stephan**

Prof. Felix **Naumann**



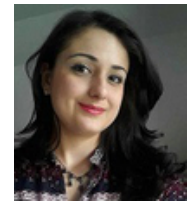
Sebastian **Kruse**

Anja **Jentzsch**



Dr. Ralf **Krestel**

Tobias **Bleifuß**



Hazar **Harmouch**

John **Koumarelas**

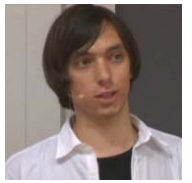
Michael **Loster**

Ahmad **Samiei**

Zhe **Zuo**

Konstantina **Lazariduo**

Maximilian **Jenders**



Toni **Grütze**

project **DuDe**

Duplicate Detection

Data Fusion

Data Profiling

project **Stratosphere**

Entity Search

project **Metanome**

Information Integration

Data Scrubbing

Data as a Service

project **GovWILD**

Data Cleansing

Web Science

Information Quality

Web Data

RDF Data Mining

Dependency Detection

Linked Open Data

ETL Management

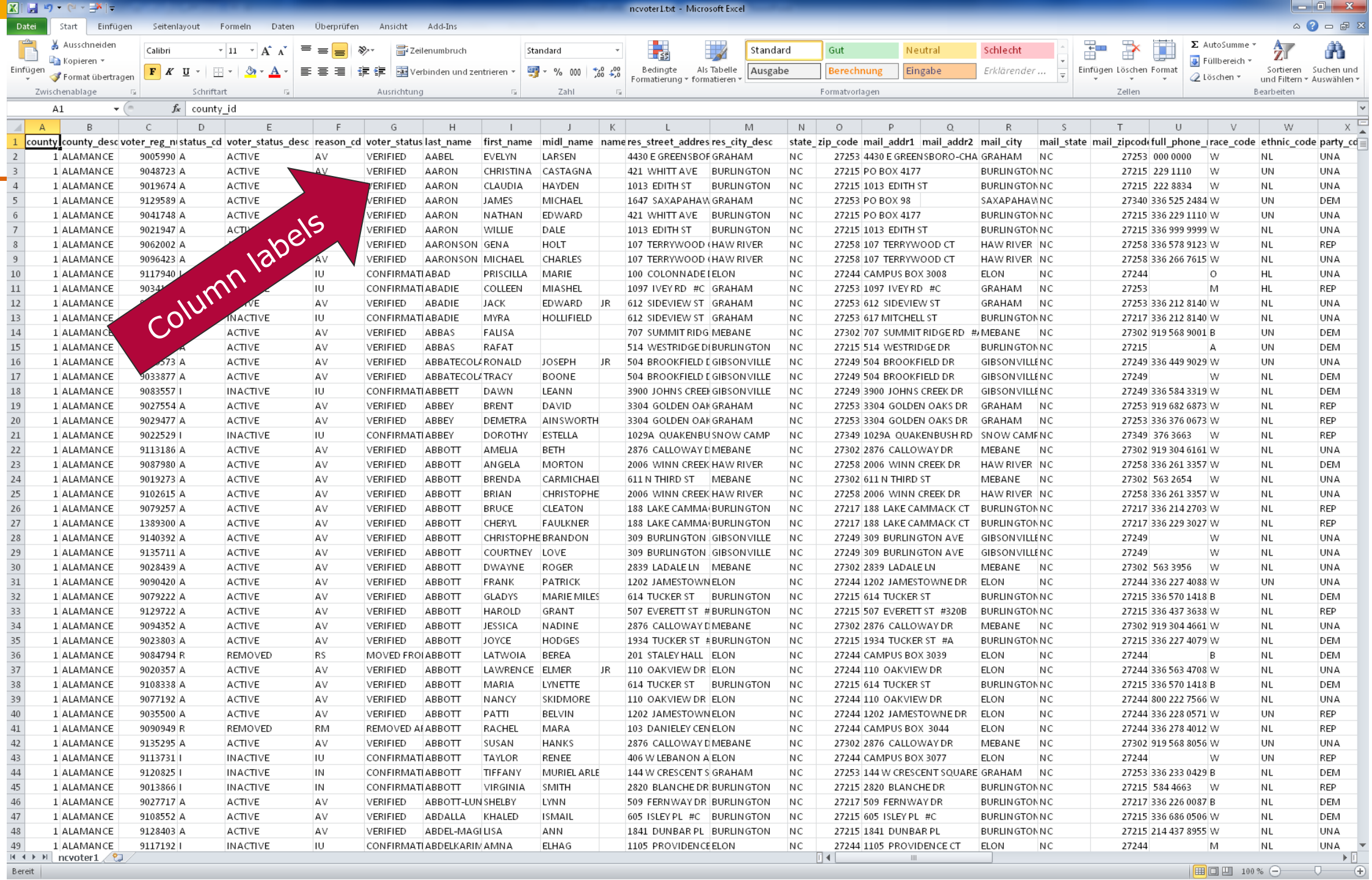
Service-Oriented Systems

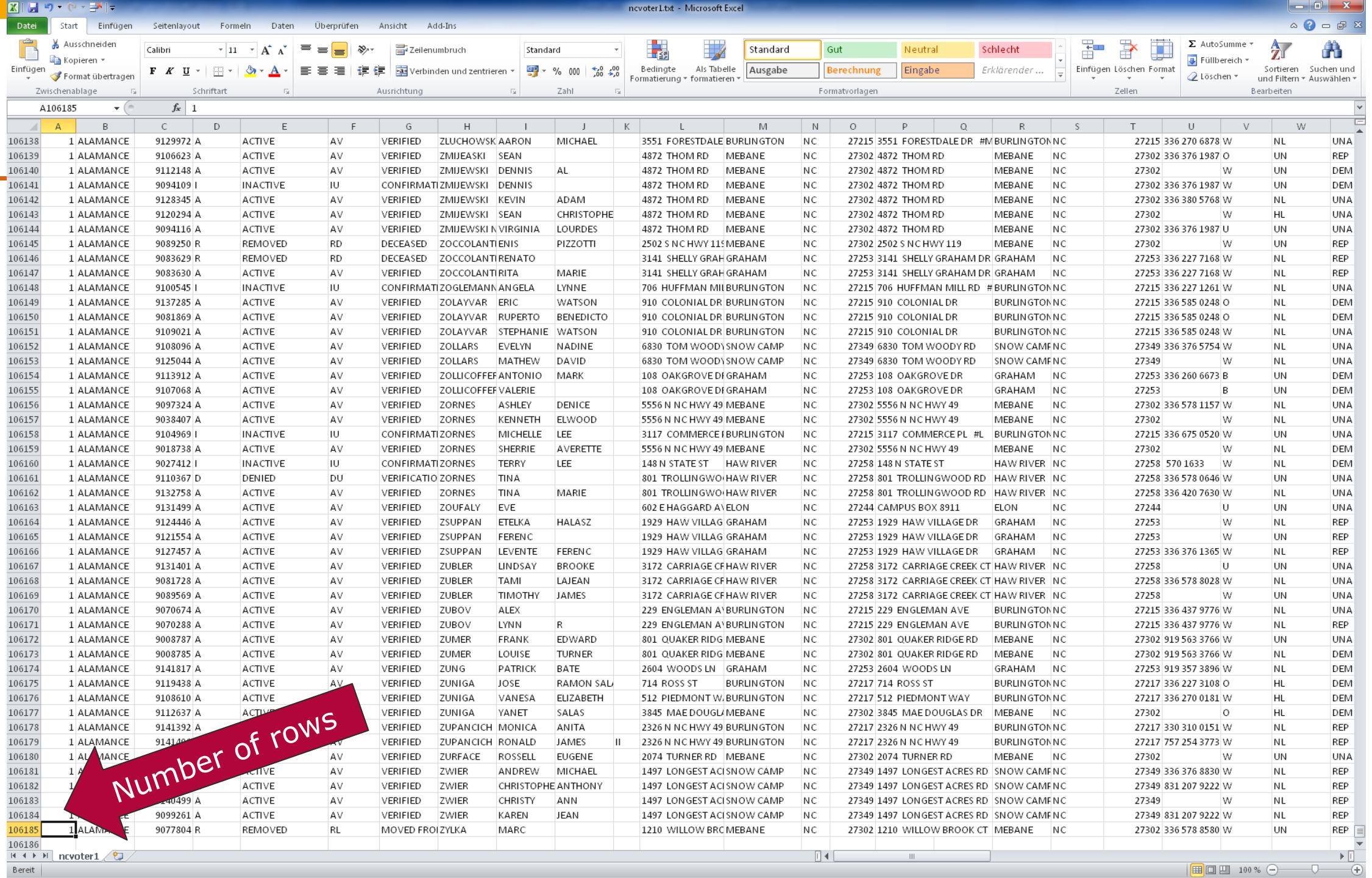
Entity Recognition

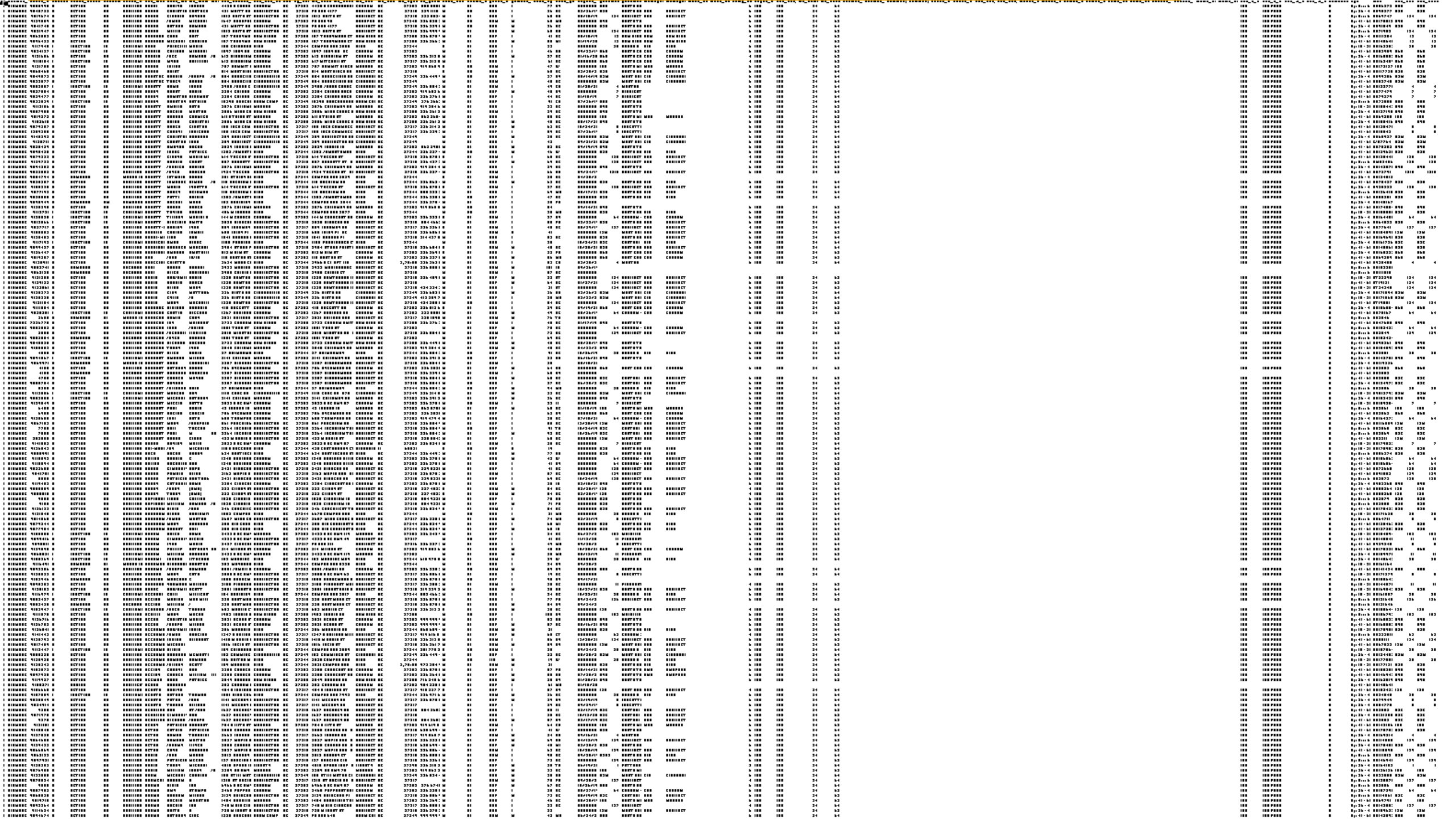
Opinion Mining

Text Mining

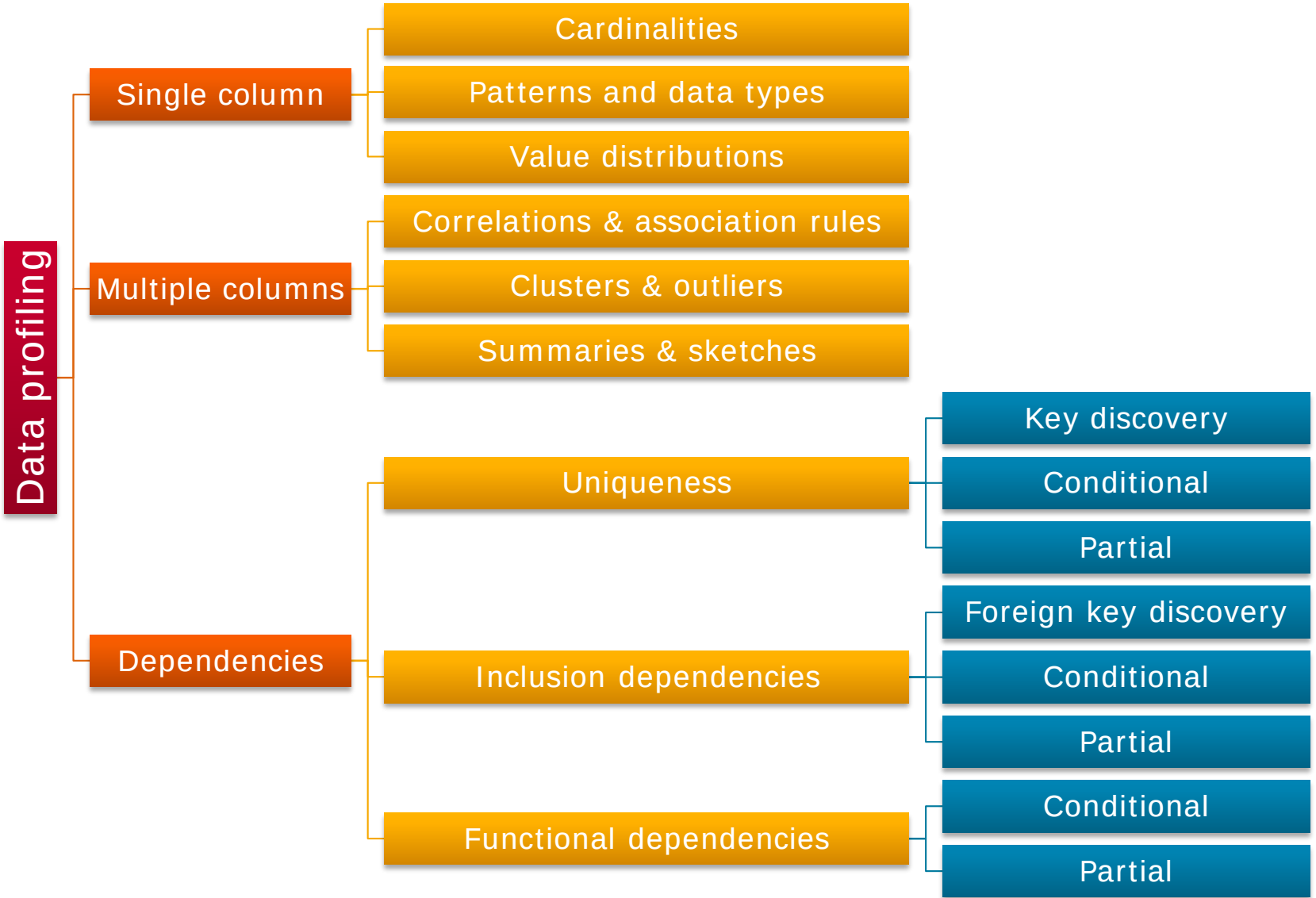
Felix Naumann
Data Profiling
Canada, 2017


ncvoter1.txt - Microsoft Excel
Column labels
county_id
ncvoter1


ncvoter1.txt - Microsoft Excel
Number of rows



Data Profiling: Classification of Tasks



Felix Naumann
Data Profiling
Canada, 2017

Use Cases for Data Profiling

- **Query optimization**: Counts and histograms, functional dependencies, …
- **Data cleansing**: Patterns, rules, and violations
- **Data integration**: Cross-DB inclusion dependencies
- **Scientific data management**: Inspect new datasets
- **Data analytics and mining**: Profiling as preparation to decide on models and questions
- **Database reverse engineering**

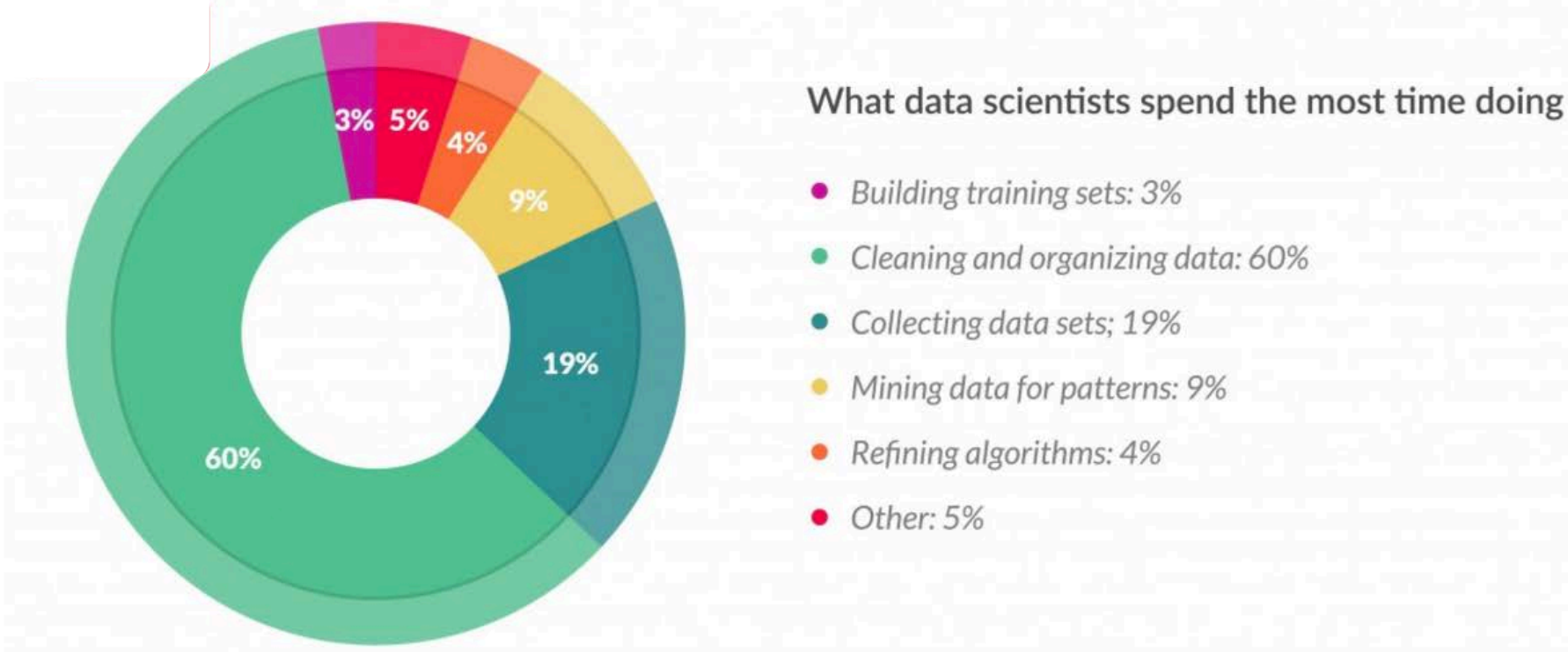
- "If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain…"

Felix Naumann
Data Profiling
Canada, 2017

[D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom. Challenges and opportunities with Big Data. Technical report, Computing Community Consortium, http://cra.org/ccc/docs/ init/bigdatawhitepaper.pdf, 2012.]

Data profiling as data preparation for any data management task – CrowdFlower's Data Science Report 2016

***Data preparation** accounts for about 80% of the work of data scientists*

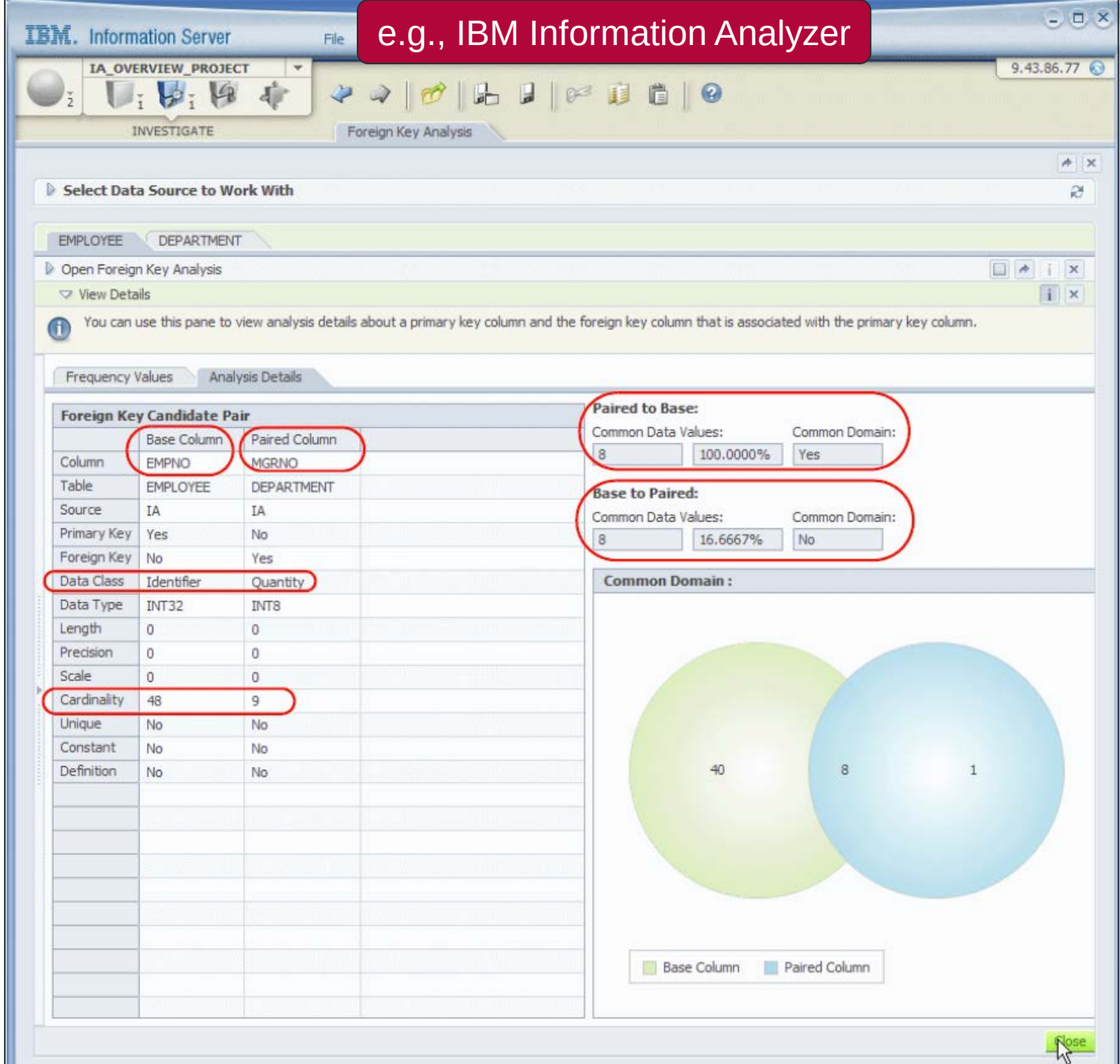


Felix Naumann
Data Profiling
Canada, 2017

"Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task", Gil Press, Forbes, March 23rd, 2016

Shortcomings of commercial and research tools

- Usability
  - Complex to configure
  - Results complex to view and interpret
- Scalability
  - Main-memory based
  - SQL based
- Efficiency
  - Coffee, Lunch, Overnight
- Functionality
  - Restricted to simplest tasks
  - Restricted to individual columns or small column sets
  - „Checking“ vs. „discovery“
- Interpretation of profiling results



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
  - Impossible to check or even enumerate

- Possible solutions
  - Scale up: More memory, faster CPUs
  - Scale in: More cores
  - Scale out: More machines
  - Scale smart: Intelligent enumeration and aggressive pruning

Felix Naumann
Data Profiling
Canada, 2017

Agenda

1. **Basic statistics**
2. Uniques and keys
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. Outlook: Other dependencies and more

Felix Naumann
Data Profiling
Canada, 2017

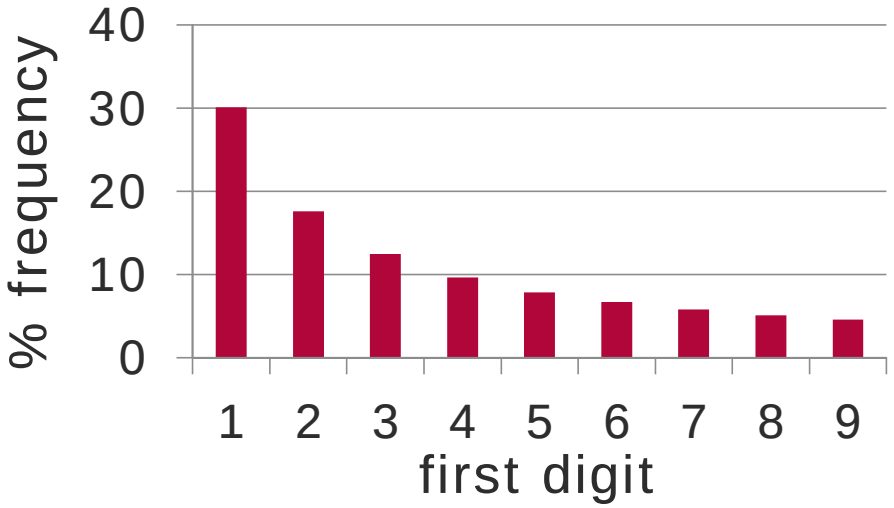
Cardinalities, distributions, and patterns

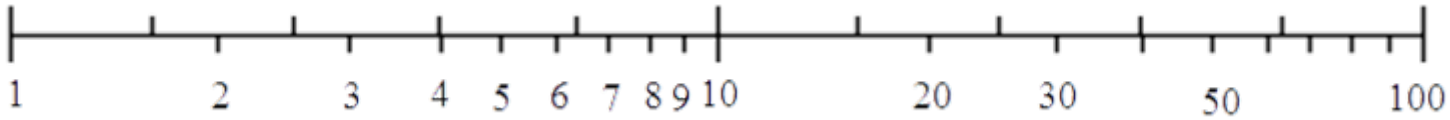
Category	Task	Description
Cardinalities	num-rows	Number of rows
	value length	Measurements of value lengths (min, max, median, and average)
	null values	Number or percentage of null values
	distinct	Number of distinct values; aka “cardinality”
	uniqueness	Number of distinct values divided by number of rows
Value distributions	histogram	Frequency histograms (equi-width, equi-depth, etc.)
	constancy	Frequency of most frequent value divided by number of rows
	quartiles	Three points that divide the (numeric) values into four equal groups
	soundex	Distribution of soundex codes
	first digit	Distribution of first digit in numeric values (Benford's law)
Patterns, data types, and domains	basic type	Generic data type: numeric, alphabetic, date, time
	data type	Concrete DBMS-specific data type: varchar, timestamp, etc.
	decimals	Maximum number of decimal places in numeric values
	precision	Maximum number of digits in numeric values
	patterns	Histogram of value patterns (Aa9...)
	data class	Semantic, generic data type: code, indicator, text, date/time, quantity, identifier, etc.
	domain	Classification of semantic domain: credit card, first name, city, phenotype, etc.

Felix Naumann
Data Profiling
Canada, 2017

Benford Law Frequency , a.k.a. “first digit law”

- Statement about the distribution of first digits *d* in (many) *naturally occurring* numbers:
  - $P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}(1 + {}^{1}/_{d})$
  - Holds if log(x) is uniformly distributed

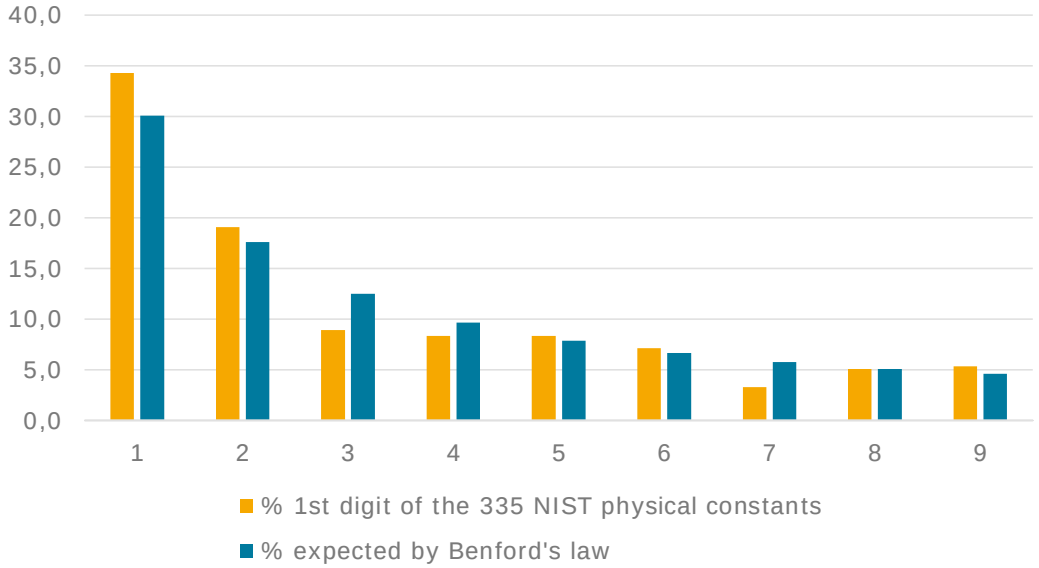




Felix Naumann
Data Profiling
Canada, 2017

Examples for Benford‘s Law

- Surface areas of 335 rivers
- Sizes of 3259 US populations
- 1800 molecular weights
- 5000 entries from a mathematical handbook
- 308 numbers contained in an issue of Reader's Digest
- Street addresses of the first 342 persons listed in American Men of Science



Heights of the 60 tallest structures

Leading digit	meters		In Benford's law
	Count	%	
1	26	43.3%	30.1%
2	7	11.7%	17.6%
3	9	15.0%	12.5%
4	6	10.0%	9.7%
5	4	6.7%	7.9%
6	1	1.7%	6.7%
7	2	3.3%	5.8%
8	5	8.3%	5.1%
9	0	0.0%	4.6%

http://en.wikipedia.org/wiki/List_of_tallest_buildings_and_structures_in_the_world#Tallest_structure_by_category

Felix Naumann
Data Profiling
Canada, 2017

Fraud detection

Agenda

1. Basic statistics
2. **Uniques and keys**
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. Outlook: Other dependencies and more

Felix Naumann
Data Profiling
Canada, 2017

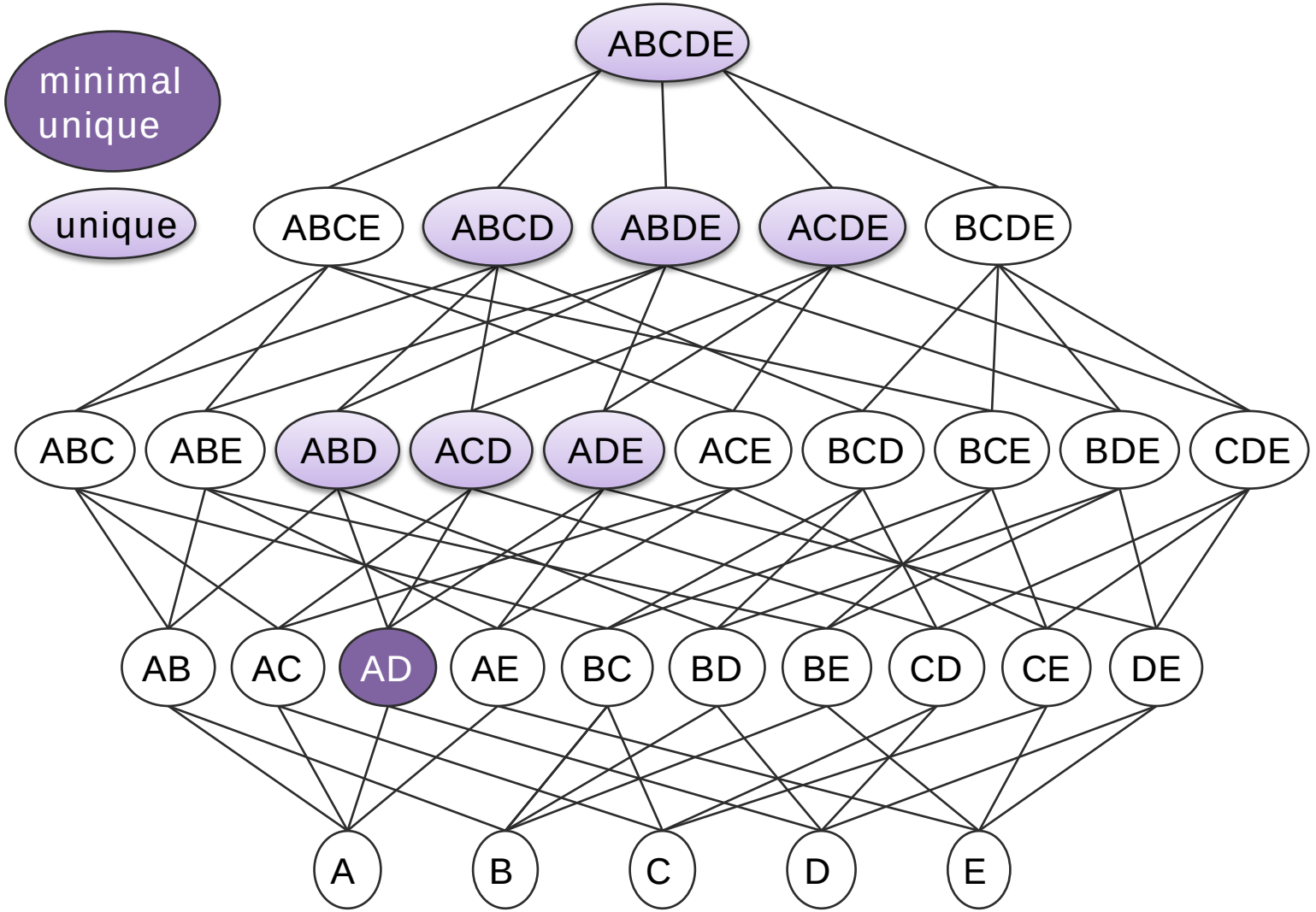
Uniqueness, keys, and foreign keys

- Uniqueness and keys
  - Unique column: Only unique values
  - Unique column combination: Only unique value combinations
    - Minimality: No column subset is unique
  - Key candidate: No null values
  - Key: Only human expert can decide
    - UCC is prerequisite

- Uniques: {A, AB, AC, BC, ABC}
- Minimal uniques: {A, BC}
- (Maximal) Non-uniques: {B, C}

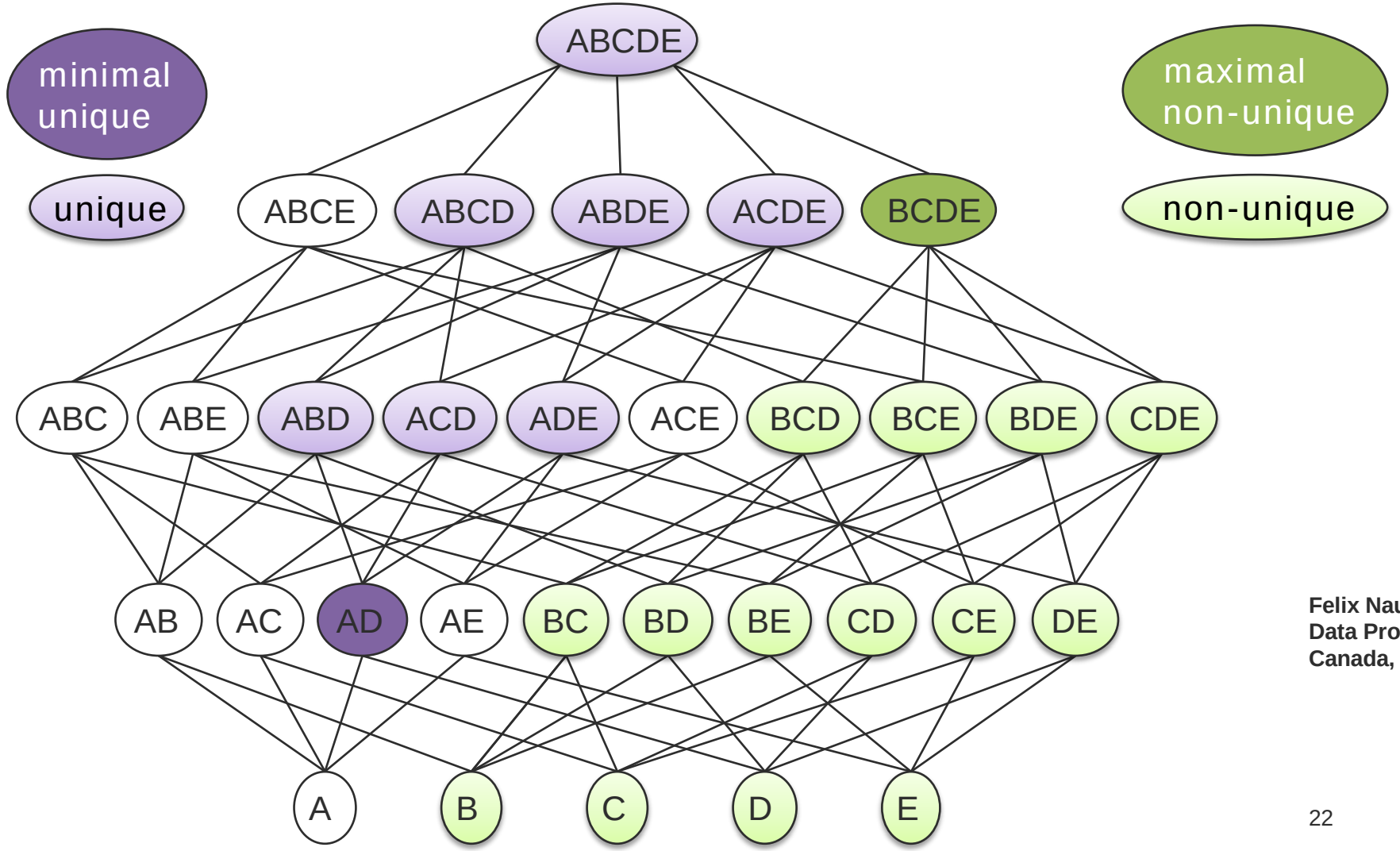
A	B	C
a	1	x
b	2	x
c	2	y

Pruning effect of a pair



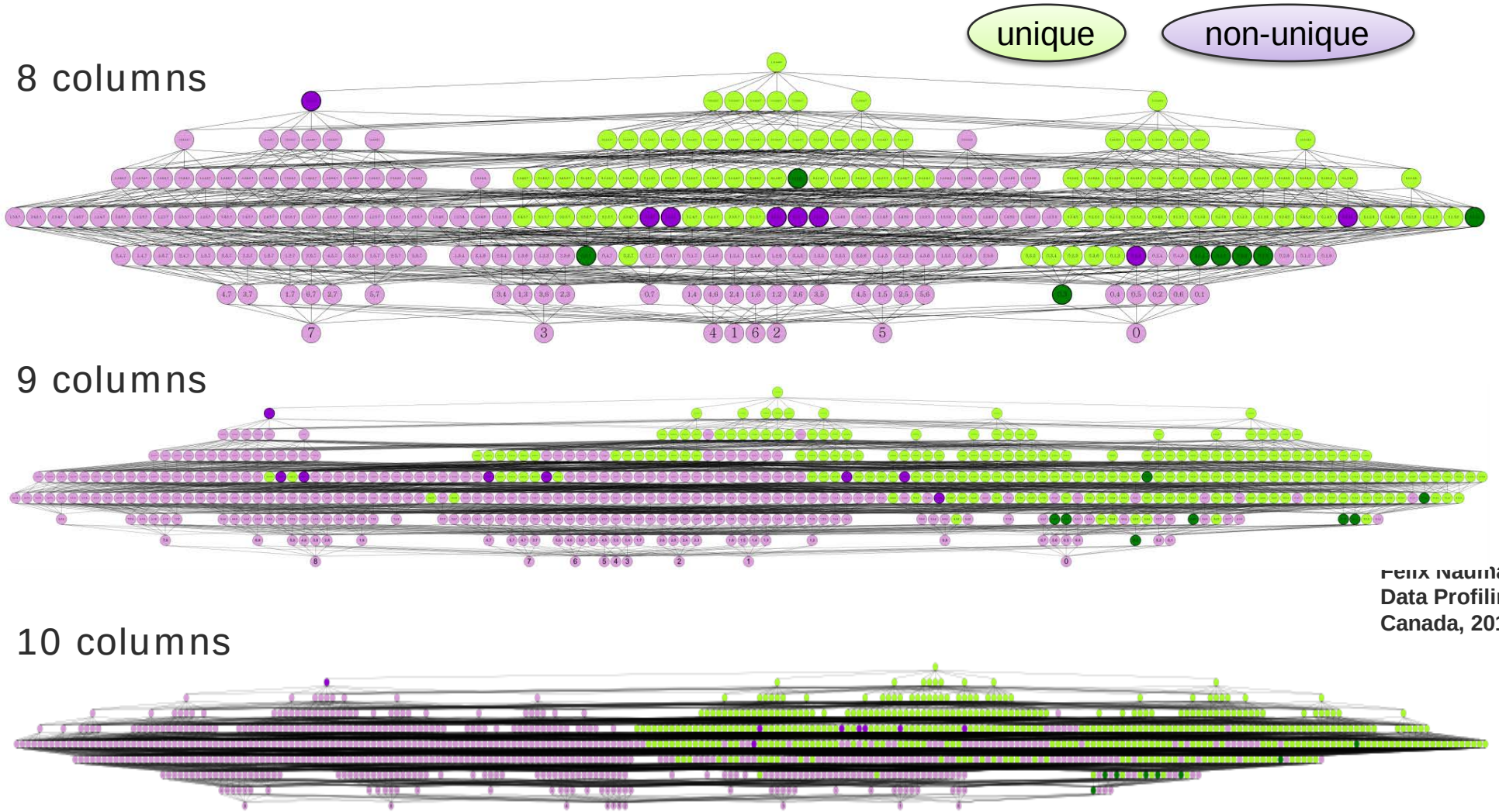
Felix Naumann
Data Profiling
Canada, 2017

Pruning both ways



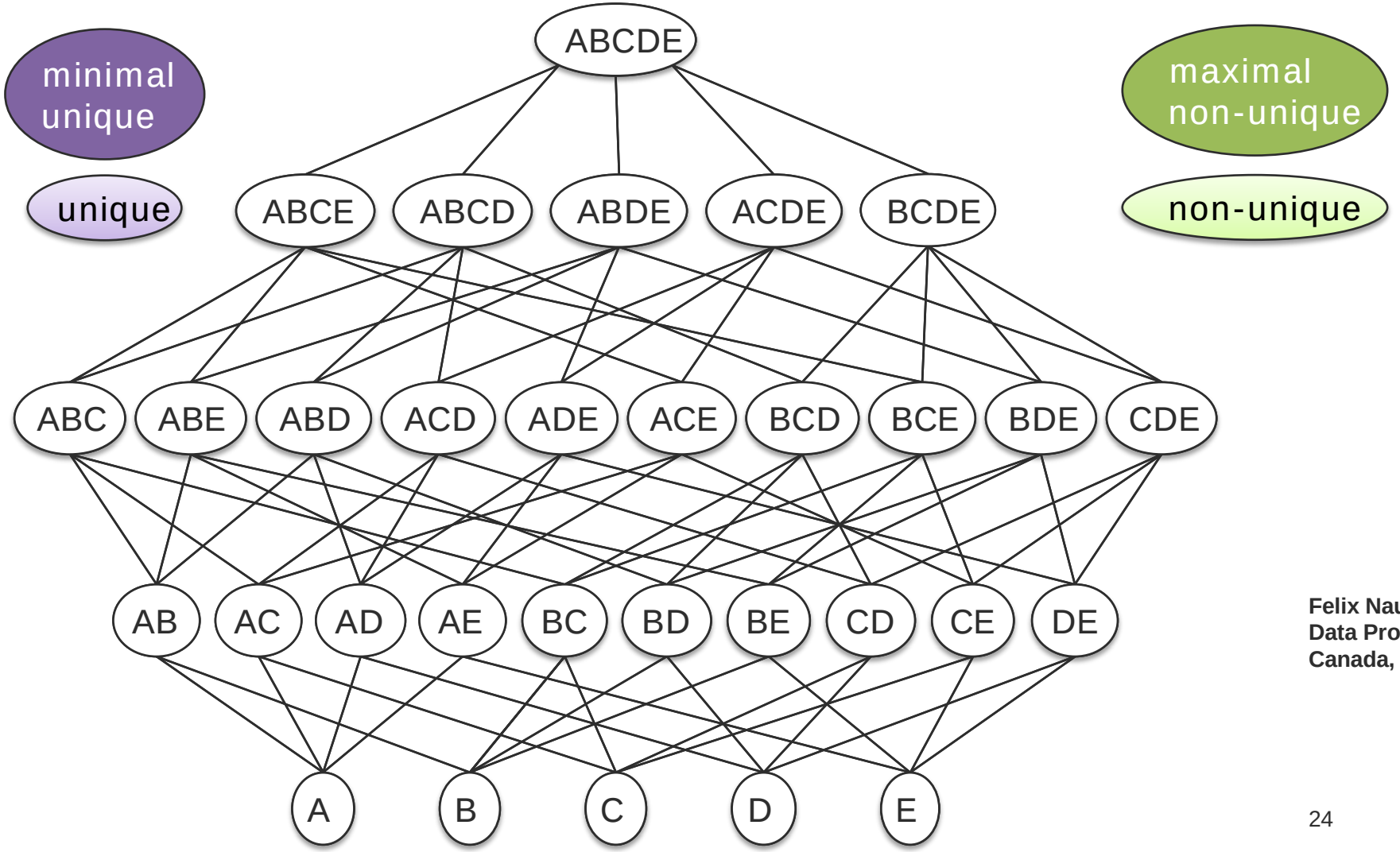
Felix Naumann
Data Profiling
Canada, 2017

TPCH – Uniques and Non-Uniques



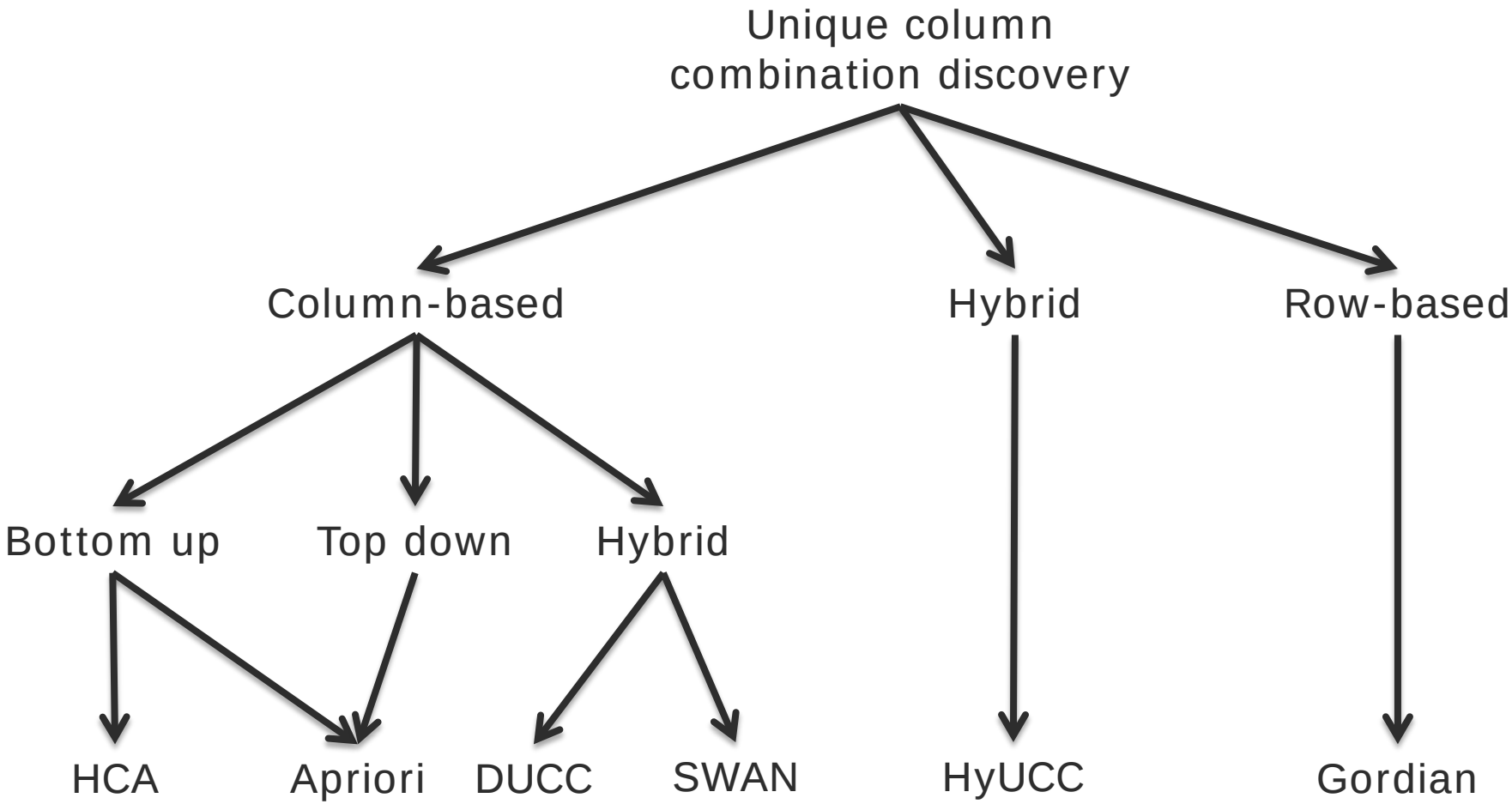
Felix Naumann
Data Profiling
Canada, 2017

Apriori visualized



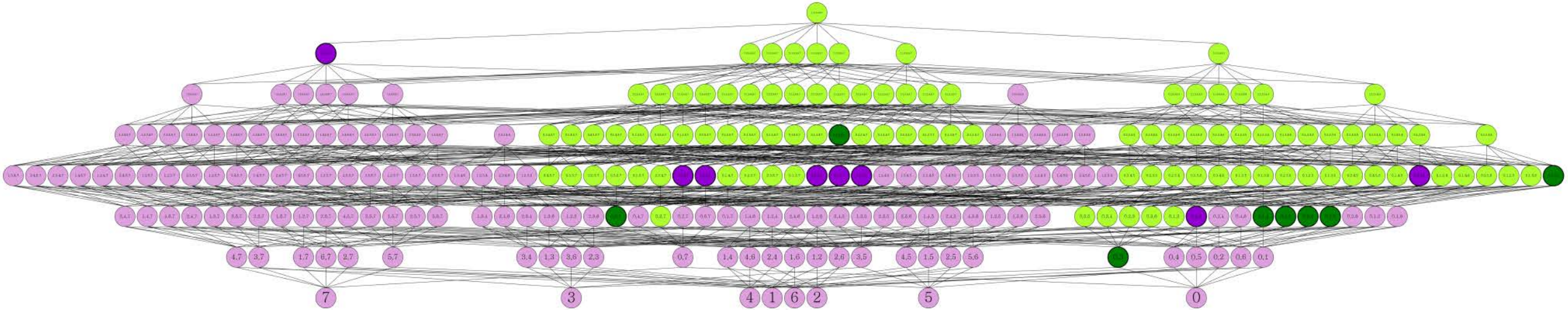
Felix Naumann
Data Profiling
Canada, 2017

Discovery Algorithms



Felix Naumann
Data Profiling
Canada, 2017

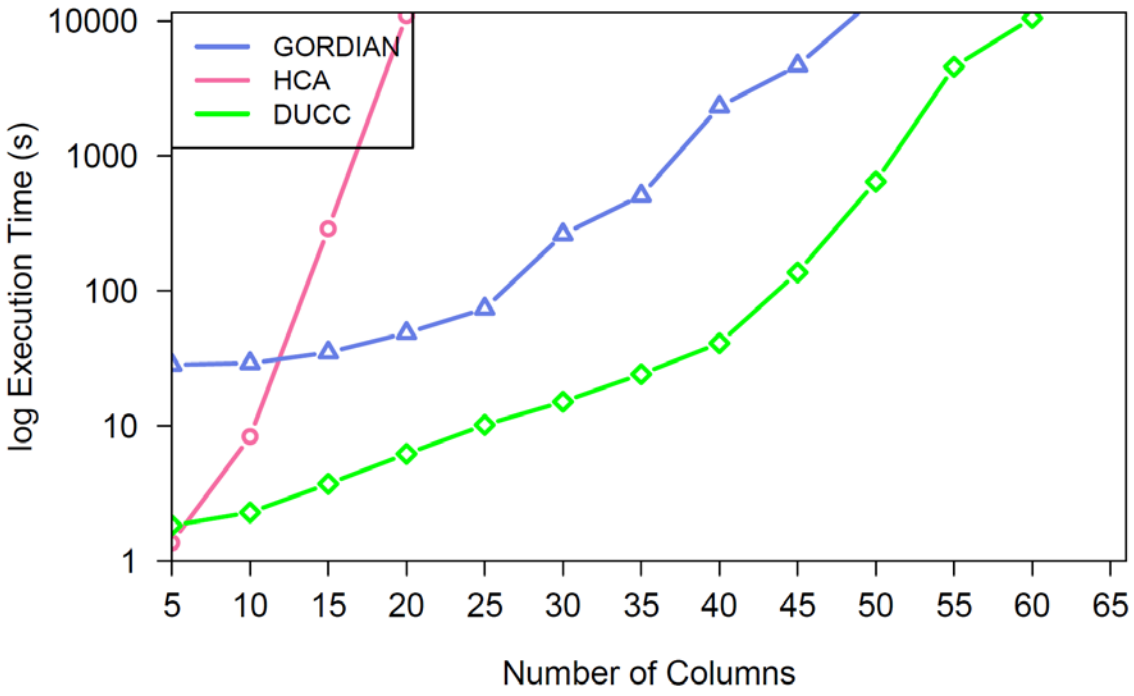
DUCC – Detecting Unique Column Combinations



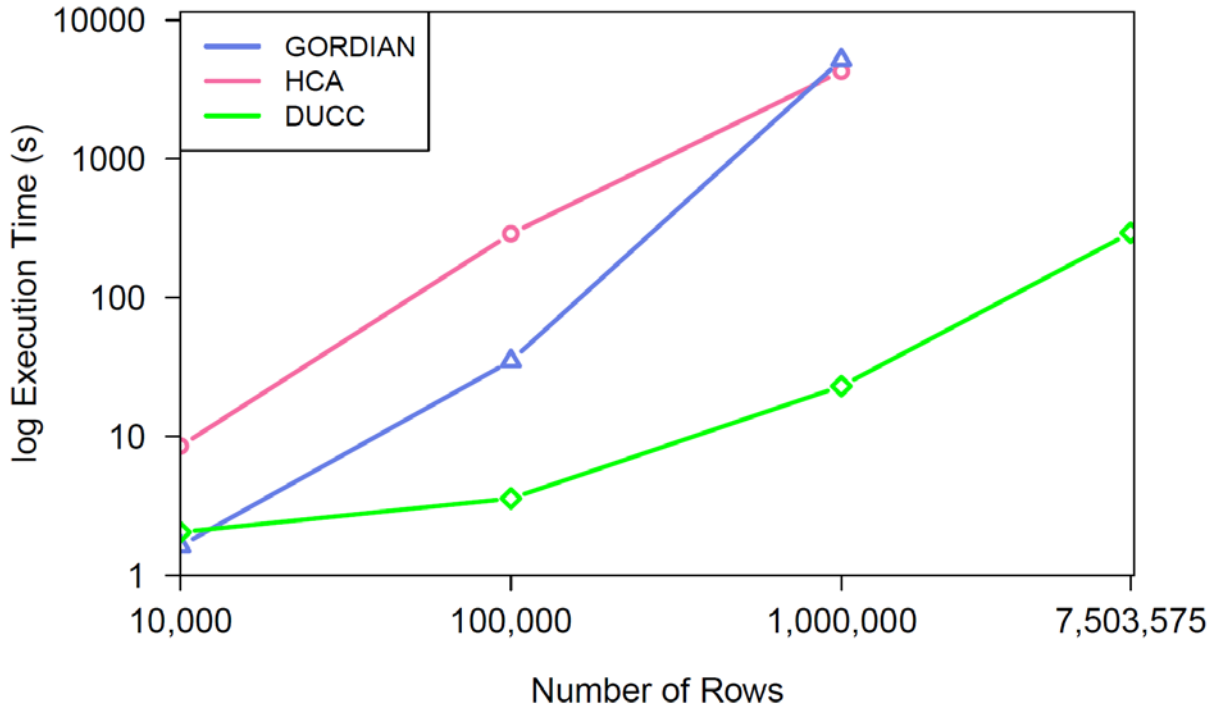
Felix Naumann
Data Profiling
Canada, 2017

Scalability in the number of columns and rows

- NCVoter data, 100k rows

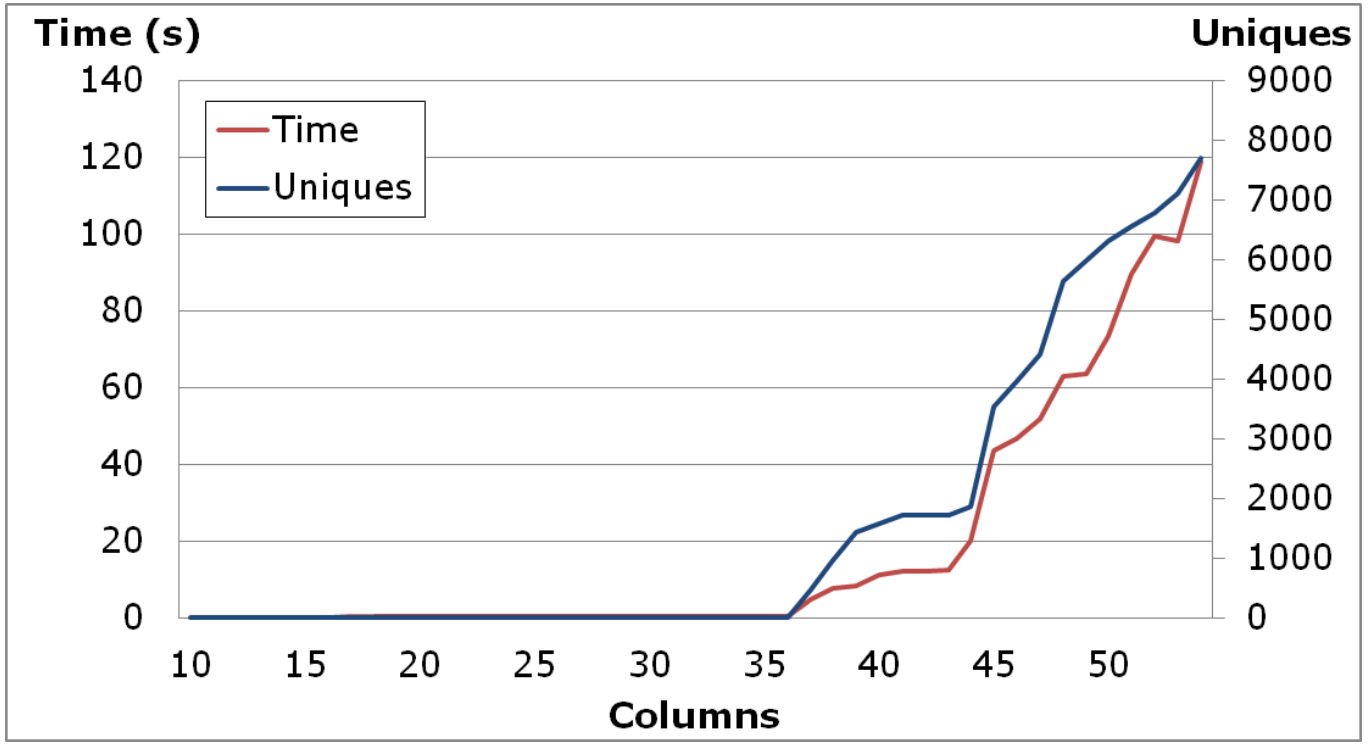


- NCVoter, 15 columns



Analysis of DUCC

- Runtime mainly depends on size of solution set



- Worst case: solution set in the middle of lattice: $\binom{n}{n/2}$ uniques

Uniques and non-uniques in NC-voter data

- **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

Agenda

1. Basic statistics
2. Uniques and keys
3. **Functional dependencies**
4. Inclusion dependencies and foreign keys
5. Outlook: Other dependencies and more

Felix Naumann
Data Profiling
Canada, 2017

Felix Naumann
Data Profiling
Canada, 2017

Functional Dependencies

Person	Lineage	Hair	Religion
			New gods
			New Gods
			Old gods
			New gods
			Old gods

Some Functional Dependencies:

1. Person → Lineage
2. Person → Hair
3. Person → Religion
4. Lineage → Hair
5. Religion, Hair → Lineage
6. …

Ned Stark: „#4 looks like a reasonable quality constraint“

Ned Stark: „I believe Joffrey violates my database constraint.“

Felix Naumann
Data Profiling
Canada, 2017

Uses for FDs

- Schema design
  - Normalization
  - Keys
- Data cleansing
- Schema design and normalization
- Key discovery

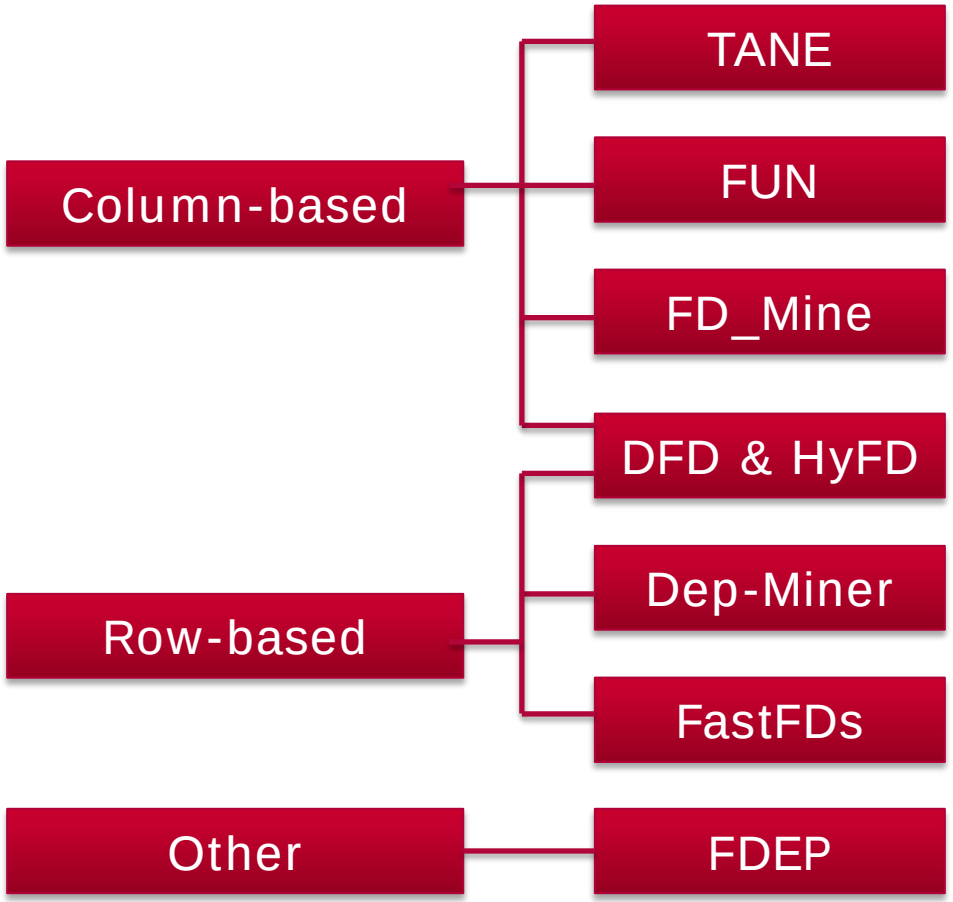
- Data cleansing (especially partial/conditional FDs)
- Anomaly detection
  - Data integrity constraints
  - Data curation rules
- Query optimization: Independence of column attributes
- Index selection

Naïve discovery approach

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

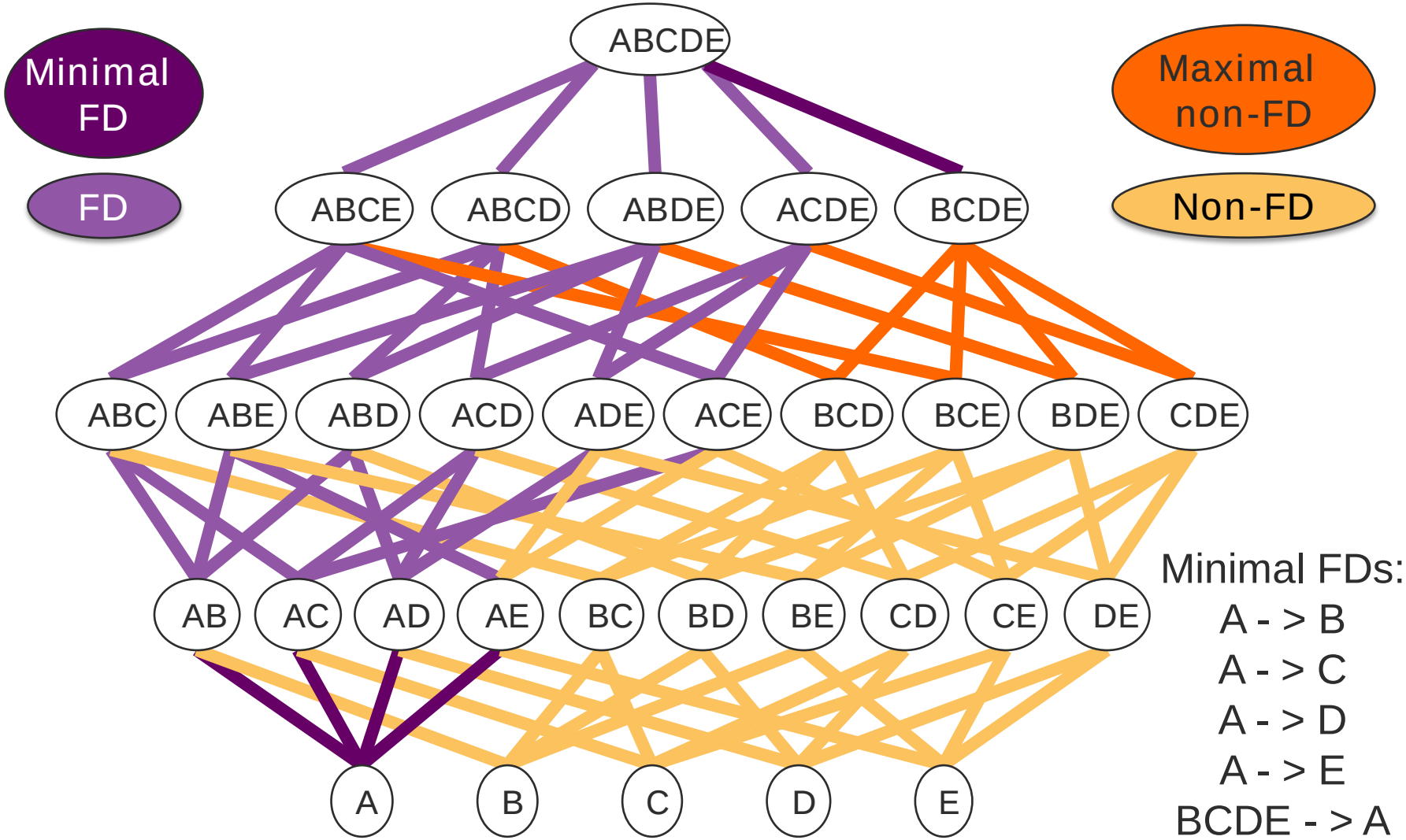
- Complexity
  - Exponential in number of attributes times number of rows squared

Current FD discovery approaches



Felix Naumann
Data Profiling
Canada, 2017

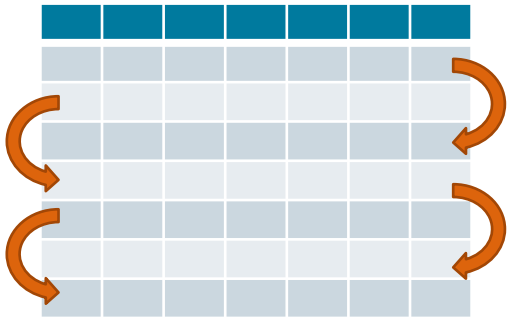
Again: Model in lattice – edges represent FDs



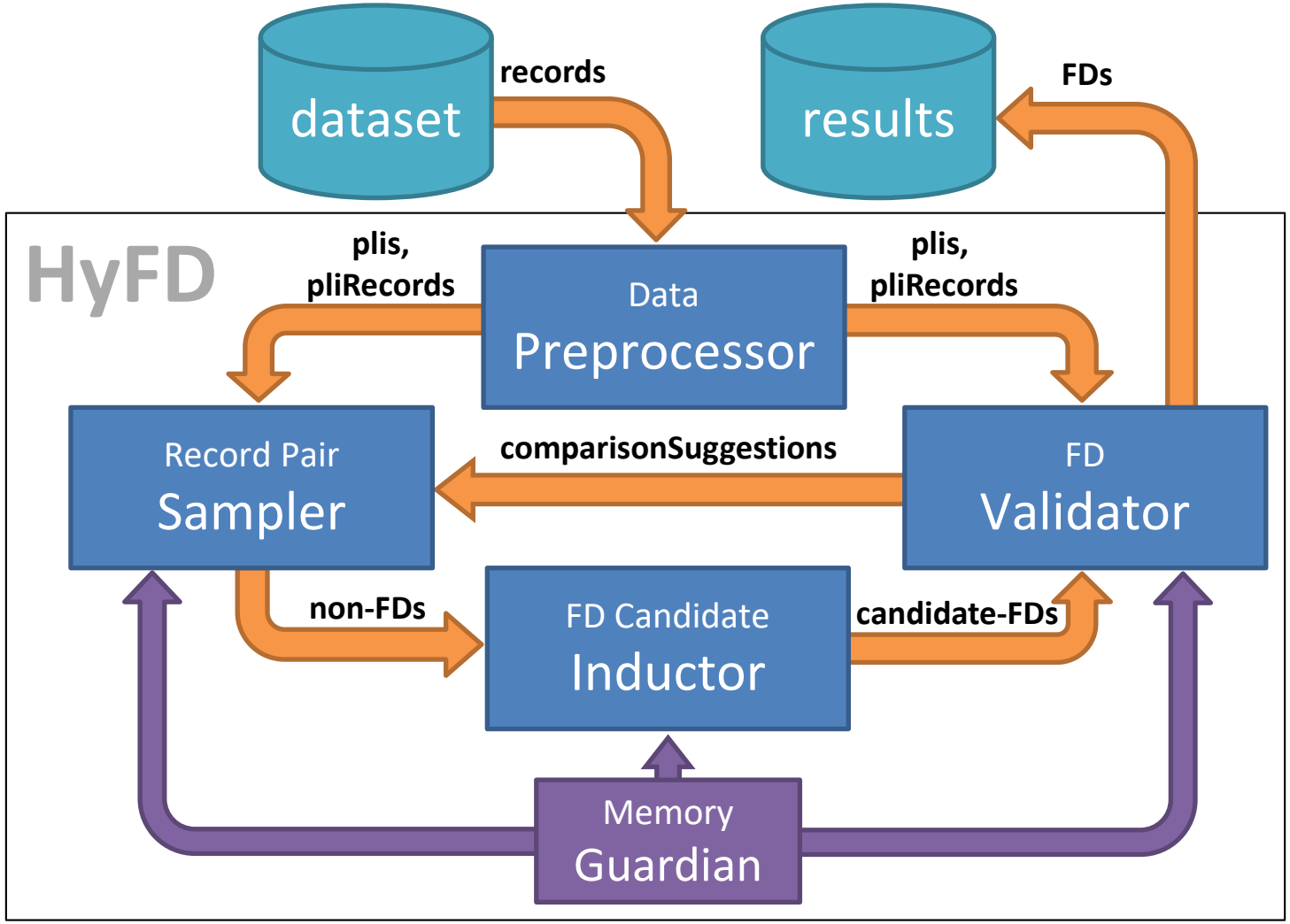
Minimal FDs:

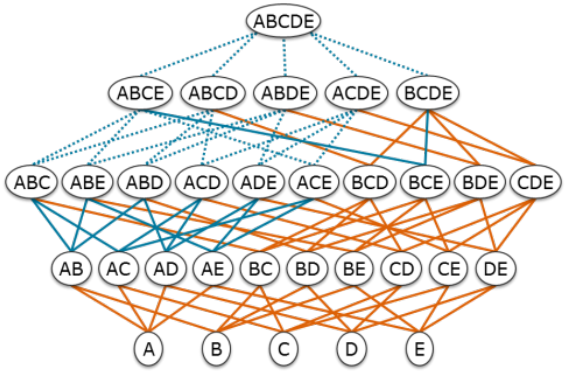
A - > B
A - > C
A - > D
A - > E
BCDE - > A

HyFD: Hybrid FD Discovery



column-efficient





row-efficient

Felix Naumann
Data Profiling
Canada, 2017

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

Agenda

1. Basic statistics
2. Uniques and keys
3. Functional dependencies
4. **Inclusion dependencies and foreign keys**
5. Outlook: Other dependencies and more

Felix Naumann
Data Profiling
Canada, 2017

IND discovery R[X] ⊆ S[Y]

- Unary and n-ary INDs: R[A] ⊆ S[B] and R[ABC] ⊆ S[DEF]

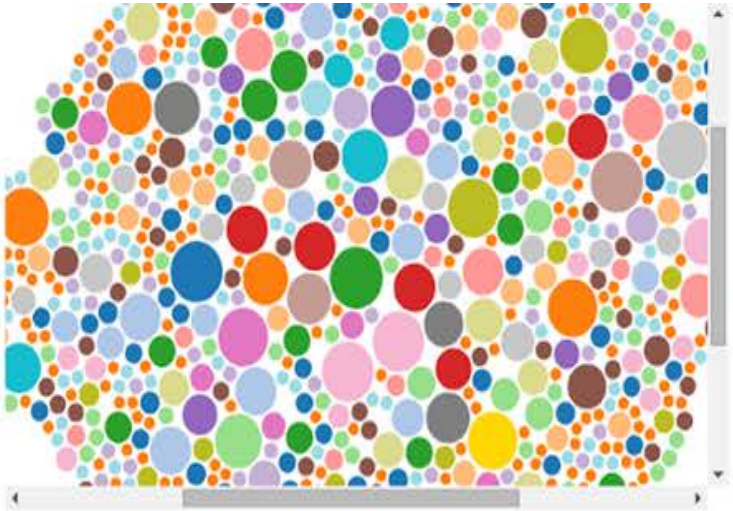
- Detect unknown foreign keys
- Example: PDB – Protein Data Bank
  - OpenMMS provides relational schema, 175 tables
  - Not a single foreign key constraint!
- Example: Ensembl – genome database
  - Shipped as MySQL dump files: >200 tables
  - Not a single foreign key constraint!
- Web tables: No schema, no constraints, but many connections
- Why are FKs missing?
  - Lack of support for foreign key constraints in DBMS
  - Fear of performance drop for constraint checking
  - Lack of database knowledge

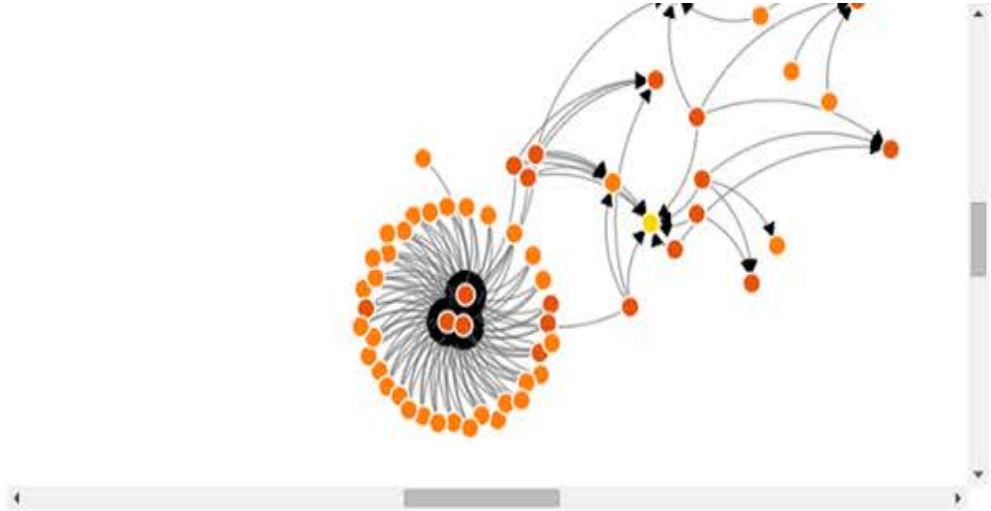
Unary IND detection:
$O(n^2)$
for n attributes

N-ary IND detection:
$O(2^n \cdot n!)$
for n attributes

Felix Naumann
Data Profiling
Canada, 2017

MANY: 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

Agenda

1. Basic statistics
2. Uniques and keys
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. **Outlook: Other dependencies and more**

Many Other Kinds of Dependencies

[Abiteboul, Hull, Vianu: Foundations of Databases, 1995]

Other dependencies

- Detecting multi-valued dependencies (MVDs) and join dependencies
- Detecting denial constraints (DCs)

- Detecting order dependencies (ODs)
  - **SELECT emp_name**
    FROM employees
    ORDER BY rank, salary
  - **SELECT emp_name**
    FROM employees
    ORDER BY rank

Remove rank

Replace with salary (if index only on 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

salary „orders“ rank

Felix Naumann
Data Profiling
Canada, 2017

Partial dependencies

- Aka. “approximate dependencies”
- Do not perfectly hold
  - For all but 10 of the tuples
  - Only for 80% of the tuples
  - Only for 1% of the tuples

- Also: Approximate dependencies
- Conditional dependencies
- Matching dependencies
- Metric dependencies

[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

Conditional Dependencies

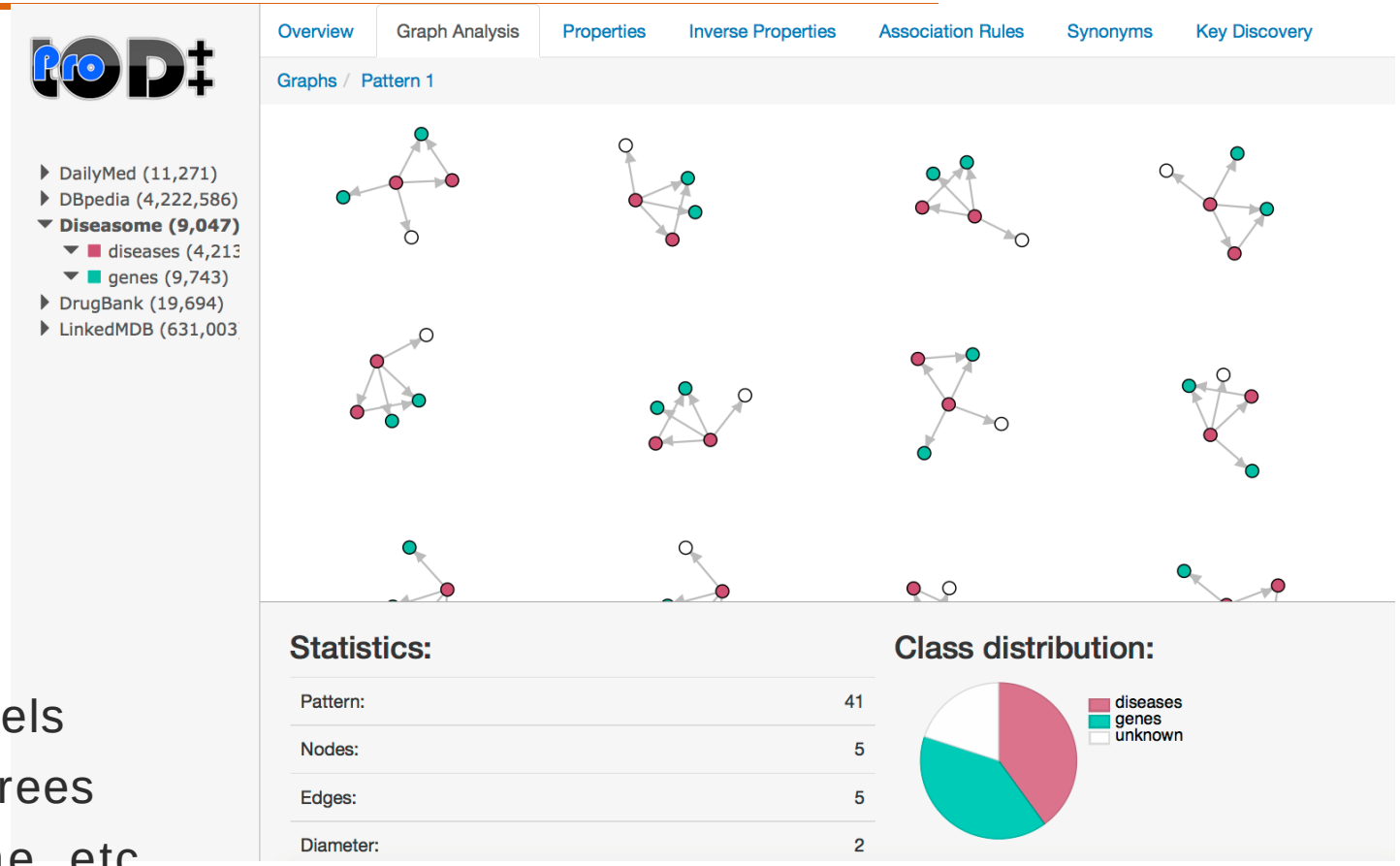
- Given a partial IND or FD: For which part does it hold?
- Expressed as a condition over the attributes of the relation

- Problems:
  - Infinite possibilities of conditions
  - Interestingness:
    - Many distinct values: less interesting
    - Few distinct values: surprising condition – high coverage

- Useful for Integration
  - Cross-database cINDs

Outlook: Profiling new types of data

- Traditional data profiling:
  Single table or multiple tables
- More and more data in other models
  - XML / nested relational / JSON
  - RDF triples
  - Textual data: Blogs, Tweets, News
  - Multimedia data



- New dimensions to profile
  - XML: Measures at different nesting levels
  - RDF: Graph structure, in- and out-degrees
  - Multimedia: Color, video-length, volume, etc.
  - Text: Sentiment, sentence structure, complexity, and other linguistic measures

Outlook: Profiling Challenges

- Efficient profiling
- Scalable profiling
- Holistic profiling
- Incremental profiling
- Online profiling
- Temporal profiling
- Profiling query results
- Profiling new types of data
- Data generation and testing
- Data profiling benchmark

- Hundreds of UCCs – which ones are keys?
- Thousands of FDs – which ones are true?
- Millions of INDs – which ones are foreign keys?

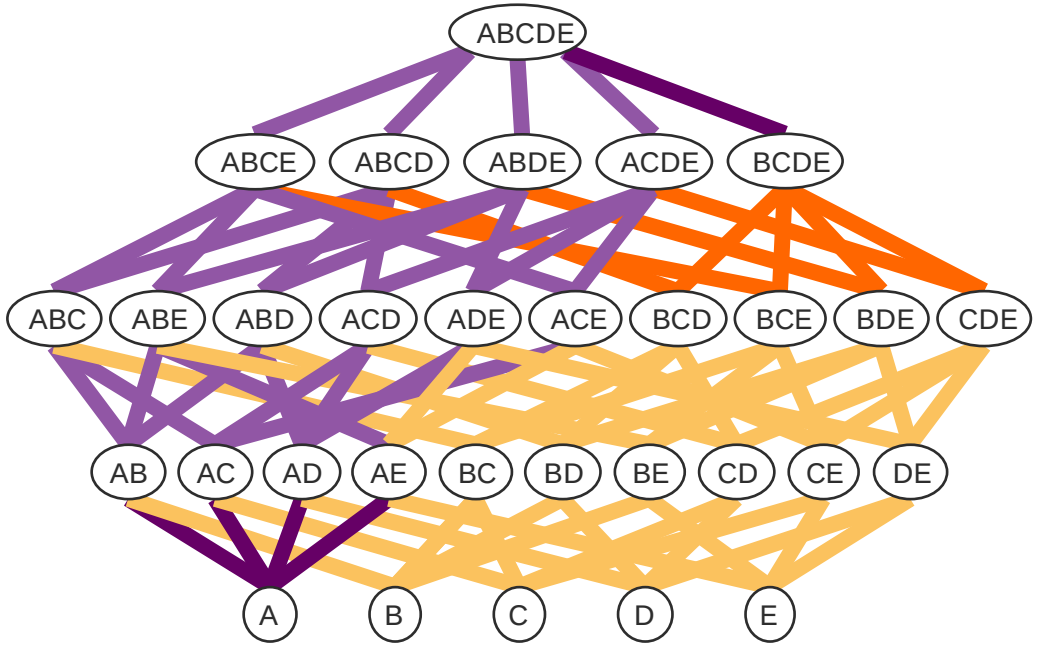
- User-driven interpretation:
  - Rank and visualize metadata
- Machine-driven interpretation
  - Machine learning

Felix Naumann
Data Profiling
Canada, 2017

www.metanome.de

Summary

1. Basic statistics
2. Uniques and keys
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. Outlook: Other dependencies and more



Felix Naumann
Data Profiling
Canada, 2017

References – work at HPI

- A Hybrid Approach for Efficient Unique Column Combination Discovery: Thorsten Papenbrock, Felix Naumann, BTW 2017
- Fast Approximate Discovery of Inclusion Dependencies, Sebastian Kruse, Thorsten Papenbrock, Christian Dullweber, Moritz Finke, Manuel Hegner, Martin Zabel, Christian Zöllner, Felix Naumann, BTW 2017
- Data-driven Schema Normalization, Thorsten Papenbrock, Felix Naumann, EDBT 2017
- Data Anamnesis: Admitting Raw Data into an Organization, Sebastian Kruse, Thorsten Papenbrock, Hazar Harmouch, Felix Naumann, IEEE Data Engineering Bulletin, 2016
- A Hybrid Approach to Functional Dependency Discovery, Thorsten Papenbrock, Felix Naumann, SIGMOD 2016
- Efficient Order Dependency Discovery, Philipp Langer and Felix Naumann, VLDB Journal 2016
- Holistic Data Profiling: Simultaneous Discovery of Various Metadata, Jens Ehrlich, Mandy Roick, Lukas Schulze, Jakob Zwiener, Thorsten Papenbrock, and Felix Naumann, EDBT 2016
- RDFind: Scalable Conditional Inclusion Dependency Discovery in RDF Datasets, Sebastian Kruse, Anja Jentzsch, Thorsten Papenbrock, Zoi Kaoudi, Jorge-Arnulfo Quiane-Ruiz, Felix Naumann, SIGMOD 2016
- Data Profiling (tutorial), Ziawasch Abedjan, Lukasz Golab and Felix Naumann, ICDE 2016
- Approximate Discovery of Functional Dependencies for Large Datasets, Tobias Bleifuß, Susanne Bülow, Johannes Frohnhofen, Julian Risch, Georg Wiese, Sebastian Kruse, Thorsten Papenbrock, Felix Naumann, CIKM 2016
- Divide & Conquer-based Inclusion Dependency Discovery, Thorsten Papenbrock, Sebastian Kruse, Jorge-Arnulfo Quiane-Ruiz, Felix Naumann, PVLDB 2015
- Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms, Thorsten Papenbrock, Jens Ehrlich, Jannik Marten, Tommy Neubert, Jan-Peer Rudolph, Martin Schönberg, Jakob Zwiener, Felix Naumann, PVLDB 2015
- Profiling relational data: a survey, Ziawasch Abedjan, Lukasz Golab, Felix Naumann, VLDB Journal 2015
- Scaling Out the Discovery of Inclusion Dependencies, Sebastian Kruse, Thorsten Papenbrock, Felix Naumann, BTW 2015
- Data Profiling with Metanome (demo), Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, Felix Naumann, PVLDB 2015
- DFD: Efficient Discovery of Functional Dependencies, Ziawasch Abedjan, Patrick Schulze, Felix Naumann, CIKM 2014
- Detecting Unique Column Combinations on Dynamic Data, Ziawasch Abedjan, Jorge-Arnulfo Quanie-Ruiz, Felix Naumann, ICDE 2014
- Profiling and Mining RDF Data with ProLOD++, Ziawasch Abedjan, Toni Gruetze, Anja Jentzsch, Felix Naumann, ICDE Demo 2014
- LODOP - Multi-Query Optimization for Linked Data Profiling Queries., Benedikt Forchhammer, Anja Jentzsch, Felix Naumann, PROFILES 2014
- Scalable Discovery of Unique Column Combinations, Arvid Heise, Jorge-Arnulfo Quiane-Ruiz, Ziawasch Abedjan, Anja Jentzsch, Felix Naumann, PVLDB 2013
- Data Profiling Revisited, Felix Naumann, SIGMOD Record 2013
- Discovering Conditional Inclusion Dependencies. Jana Bauckmann, Ziawasch Abedjan, Heiko Müller, Ulf Leser, Felix Naumann, CIKM 2012
- Advancing the Discovery of Unique Column Combinations, Ziawasch Abedjan, Felix Naumann, CIKM 2011
- A Machine Learning Approach to Foreign Key Discovery, Alexandra Rostin, Oliver Albrecht, Jana Bauckmann, Felix Naumann, Ulf Leser, WebDB 2009
- Efficiently Detecting Inclusion Dependencies, Jana Bauckmann, Ulf Leser, Felix Naumann, Veronique Tietz, ICDE 2007
- Efficiently Computing Inclusion Dependencies for Schema Discovery, Jana Bauckmann, Ulf Leser, Felix Naumann, ICDE 2006