



## (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



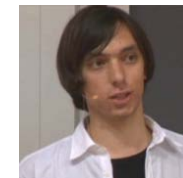
Anja Jentzsch



Dr. Ralf Krestel



Hazar Harmouch

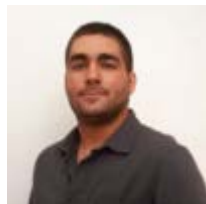


Toni Grütze

Felix Naumann  
Data Profiling  
Canada, 2017



Tobias Bleifuß



John Koumarelas



Michael Loster



Ahmad Samiei



Zhe Zuo



Konstantina Lazaridou



Maximilian Jenders

project **DuDe**  
Data Fusion Duplicate Detection  
project **Stratosphere** Entity Search  
Data Profiling Information Integration  
Data Scrubbing Data as a Service  
project **Metanome** Web Data Information Quality Data Cleansing Web Science  
project **GovWILD** Dependency Detection Linked Open Data RDF Data Mining  
Service-Oriented Systems Entity Recognition Opinion Mining ETL Management  
Text Mining



nvoter.Lbt - Microsoft Excel

Datei Start Einfügen Seitenlayout Formeln Daten Überprüfen Ansicht Add-Ins

Ausschneiden Kopieren Format übertragen Zwischenablage Einfügen

Calibri 11 A A Zeilenumbruch Standard Bedingte Formatierung Als Tabelle formatieren Ausgabe Berechnung Eingabe Erklärender...

AutoSumme Füllbereich Löschen Sortieren und Filtern Suchen und Auswählen

	A1	county_id																							
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
1	county	county_desc	voter_reg_n	status_cd	voter_status_desc	reason_cd	voter_status	last_name	first_name	midl_name	name	res_street_addr	res_city_desc	state	zip_code	mail_addr1	mail_addr2	mail_city	mail_state	mail_zipcode	full_phone	race_code	ethnic_code	party_cd	
2	1	ALAMANCE	9005990	A	ACTIVE	AV	VERIFIED	AABEL	EVELYN	LARSEN	4430 E GREENSBOW	GRAHAM	NC	27253	4430 E GREENSBORO-CHA		GRAHAM	NC	27253	000 0000	W	NL	UNA		
3	1	ALAMANCE	9048723	A	ACTIVE	AV	VERIFIED	AARON	CHRISTINA	CASTAGNA	421 WHITT AVE	BURLINGTON	NC	27215	PO BOX 4177		BURLINGTON	NC	27215	229 1110	W	UN	UNA		
4	1	ALAMANCE	9019674	A	ACTIVE	AV	VERIFIED	AARON	CLAUDIA	HAYDEN	1013 EDITH ST	BURLINGTON	NC	27215	1013 EDITH ST		BURLINGTON	NC	27215	222 8834	W	NL	UNA		
5	1	ALAMANCE	9129589	A	ACTIVE	AV	VERIFIED	AARON	JAMES	MICHAEL	1647 SAXAPAHAW	GRAHAM	NC	27253	PO BOX 98		SAXAPAHAW	NC	27340	336 525 2484	W	UN	DEM		
6	1	ALAMANCE	9041748	A	ACTIVE	AV	VERIFIED	AARON	NATHAN	EDWARD	421 WHITT AVE	BURLINGTON	NC	27215	PO BOX 4177		BURLINGTON	NC	27215	336 229 1110	W	UN	UNA		
7	1	ALAMANCE	9021947	A	ACTIVE	AV	VERIFIED	AARON	WILLIE	DALE	1013 EDITH ST	BURLINGTON	NC	27215	1013 EDITH ST		BURLINGTON	NC	27215	336 999 9999	W	NL	UNA		
8	1	ALAMANCE	9062002	A	ACTIVE	AV	VERIFIED	AARONSON	GENA	HOLT	107 TERRYWOOD	HAW RIVER	NC	27258	107 TERRYWOOD CT		HAW RIVER	NC	27258	336 578 9123	W	NL	REP		
9	1	ALAMANCE	9096423	A	ACTIVE	AV	VERIFIED	AARONSON	MICHAEL	CHARLES	107 TERRYWOOD	HAW RIVER	NC	27258	107 TERRYWOOD CT		HAW RIVER	NC	27258	336 266 7615	W	NL	UNA		
10	1	ALAMANCE	9117940	I	INACTIVE	IU	CONFIRMATI	ABAD	PRISCILLA	MARIE	100 COLONNADE	ELON	NC	27244	CAMPUS BOX 3008		ELON	NC	27244		O	HL	UNA		
11	1	ALAMANCE	9034444	I	INACTIVE	IU	CONFIRMATI	ABADIE	COLLEEN	MIASHEL	1097 IVEY RD	GRAHAM	NC	27253	1097 IVEY RD	#C	GRAHAM	NC	27253		M	HL	REP		
12	1	ALAMANCE	9034444	I	INACTIVE	AV	VERIFIED	ABADIE	JACK	EDWARD	612 SIDEVIEW ST	GRAHAM	NC	27253	612 SIDEVIEW ST		GRAHAM	NC	27253	336 212 8140	W	NL	UNA		
13	1	ALAMANCE	9034444	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	1	ALAMANCE	9034444	A	ACTIVE	AV	VERIFIED	ABBAS	FALISA		707 SUMMIT RIDG	MEBANE	NC	27302	707 SUMMIT RIDGE RD	#	MEBANE	NC	27302	919 568 9001	B	UN	DEM		
15	1	ALAMANCE	9034444	A	ACTIVE	AV	VERIFIED	ABBAS	RAFAT		514 WESTRIDGE	DIBURLINGTON	NC	27215	514 WESTRIDGE DR		BURLINGTON	NC	27215		A	UN	DEM		
16	1	ALAMANCE	9034444	A	ACTIVE	AV	VERIFIED	ABBATECOLA	RONALD	JOSEPH	504 BROOKFIELD	GIBSONVILLE	NC	27249	504 BROOKFIELD DR		GIBSONVILLE	NC	27249	336 449 9029	W	UN	UNA		
17	1	ALAMANCE	9033877	A	ACTIVE	AV	VERIFIED	ABBATECOLA	TRACY	BOONE	504 BROOKFIELD	GIBSONVILLE	NC	27249	504 BROOKFIELD DR		GIBSONVILLE	NC	27249		W	NL	DEM		
18	1	ALAMANCE	9083557	I	INACTIVE	IU	CONFIRMATI	ABBETT	DAWN	LEANN	3900 JOHN'S CREEK	GIBSONVILLE	NC	27249	3900 JOHN'S CREEK DR		GIBSONVILLE	NC	27249	336 584 3319	W	NL	DEM		
19	1	ALAMANCE	9027554	A	ACTIVE	AV	VERIFIED	ABBEY	BRENT	DAVID	3304 GOLDEN OAK	GRAHAM	NC	27253	3304 GOLDEN OAKS DR		GRAHAM	NC	27253	919 682 6873	W	NL	REP		
20	1	ALAMANCE	9029477	A	ACTIVE	AV	VERIFIED	ABBEY	DEMETRA	AINSWORTH	3304 GOLDEN OAK	GRAHAM	NC	27253	3304 GOLDEN OAKS DR		GRAHAM	NC	27253	336 376 0673	W	NL	REP		
21	1	ALAMANCE	9022529	I	INACTIVE	IU	CONFIRMATI	ABBEY	DOROTHY	ESTELLA	1029A QUAKENBUSH	SNOW CAMP	NC	27349	1029A QUAKENBUSH RD		SNOW CAMP	NC	27349	376 3663	W	NL	REP		
22	1	ALAMANCE	9113186	A	ACTIVE	AV	VERIFIED	ABBOTT	AMELIA	BETH	2876 CALLOWAY	DMEBANE	NC	27302	2876 CALLOWAY DR		MEBANE	NC	27302	919 304 6161	W	NL	UNA		
23	1	ALAMANCE	9087980	A	ACTIVE	AV	VERIFIED	ABBOTT	ANGELA	MORTON	2006 WINN CREEK	HAW RIVER	NC	27258	2006 WINN CREEK DR		HAW RIVER	NC	27258	336 261 3357	W	NL	DEM		
24	1	ALAMANCE	9019273	A	ACTIVE	AV	VERIFIED	ABBOTT	BRENDA	CARMICHAEL	611 N THIRD ST	MEBANE	NC	27302	611 N THIRD ST		MEBANE	NC	27302	563 2654	W	NL	UNA		
25	1	ALAMANCE	9102615	A	ACTIVE	AV	VERIFIED	ABBOTT	BRIAN	CHRISTOPHE	2006 WINN CREEK	HAW RIVER	NC	27258	2006 WINN CREEK DR		HAW RIVER	NC	27258	336 261 3357	W	NL	UNA		
26	1	ALAMANCE	9079257	A	ACTIVE	AV	VERIFIED	ABBOTT	BRUCE	CLEATON	188 LAKE CAMMA	BURLINGTON	NC	27217	188 LAKE CAMMACK CT		BURLINGTON	NC	27217	336 214 2703	W	NL	REP		
27	1	ALAMANCE	1389300	A	ACTIVE	AV	VERIFIED	ABBOTT	CHERYL	FAULKNER	188 LAKE CAMMA	BURLINGTON	NC	27217	188 LAKE CAMMACK CT		BURLINGTON	NC	27217	336 229 3027	W	NL	REP		
28	1	ALAMANCE	9140392	A	ACTIVE	AV	VERIFIED	ABBOTT	CHRISTOPHE	BRANDON	309 BURLINGTON	GIBSONVILLE	NC	27249	309 BURLINGTON AVE		GIBSONVILLE	NC	27249		W	NL	UNA		
29	1	ALAMANCE	9135711	A	ACTIVE	AV	VERIFIED	ABBOTT	COURTNEY	LOVE	309 BURLINGTON	GIBSONVILLE	NC	27249	309 BURLINGTON AVE		GIBSONVILLE	NC	27249		W	NL	UNA		
30	1	ALAMANCE	9028439	A	ACTIVE	AV	VERIFIED	ABBOTT	DWAYNE	ROGER	2839 LADALE LN	MEBANE	NC	27302	2839 LADALE LN		MEBANE	NC	27302	563 3956	W	NL	UNA		
31	1	ALAMANCE	9090420	A	ACTIVE	AV	VERIFIED	ABBOTT	FRANK	PATRICK	1202 JAMESTOWN	ELON	NC	27244	1202 JAMESTOWN DR		ELON	NC	27244	336 227 4088	W	UN	UNA		
32	1	ALAMANCE	9079222	A	ACTIVE	AV	VERIFIED	ABBOTT	GLADYS	MARIE MILES	614 TUCKER ST	BURLINGTON	NC	27215	614 TUCKER ST		BURLINGTON	NC	27215	336 570 1418	B	NL	DEM		
33	1	ALAMANCE	9129722	A	ACTIVE	AV	VERIFIED	ABBOTT	HAROLD	GRANT	507 EVERETT ST	#BURLINGTON	NC	27215	507 EVERETT ST	#320B	BURLINGTON	NC	27215	336 437 3638	W	NL	REP		
34	1	ALAMANCE	9094352	A	ACTIVE	AV	VERIFIED	ABBOTT	JESSICA	NADINE	2876 CALLOWAY	DMEBANE	NC	27302	2876 CALLOWAY DR		MEBANE	NC	27302	919 304 4661	W	NL	UNA		
35	1	ALAMANCE	9023803	A	ACTIVE	AV	VERIFIED	ABBOTT	JOYCE	HODGES	1934 TUCKER ST	#BURLINGTON	NC	27215	1934 TUCKER ST	#A	BURLINGTON	NC	27215	336 227 4079	W	NL	DEM		
36	1	ALAMANCE	9084794	R	REMOVED	RS	MOVED FRO	ABBOTT	LATWOIA	BEREA	201 STALEY HALL	ELON	NC	27244	CAMPUS BOX 3039		ELON	NC	27244		B	NL	DEM		
37	1	ALAMANCE	9020357	A	ACTIVE	AV	VERIFIED	ABBOTT	LAWRENCE	ELMER	110 OAKVIEW DR	ELON	NC	27244	110 OAKVIEW DR		ELON	NC	27244	336 563 4708	W	NL	UNA		
38	1	ALAMANCE	9108338	A	ACTIVE	AV	VERIFIED	ABBOTT	MARIA	LYNETTE	614 TUCKER ST	BURLINGTON	NC	27215	614 TUCKER ST		BURLINGTON	NC	27215	336 570 1418	B	NL	DEM		
39	1	ALAMANCE	9077192	A	ACTIVE	AV	VERIFIED	ABBOTT	NANCY	SKIDMORE	110 OAKVIEW DR	ELON	NC	27244	110 OAKVIEW DR		ELON	NC	27244	800 222 7566	W	NL	UNA		
40	1	ALAMANCE	9035500	A	ACTIVE	AV	VERIFIED	ABBOTT	PATTI	BELVIN	1202 JAMESTOWN	ELON	NC	27244	1202 JAMESTOWN DR		ELON	NC	27244	336 228 0571	W	UN	REP		
41	1	ALAMANCE	9090949	R	REMOVED	RM	REMOVED A	ABBOTT	RACHEL	MARA	103 DANIELEY	CENELON	NC	27244	CAMPUS BOX 3044		ELON	NC	27244	336 278 4012	W	NL	REP		
42	1	ALAMANCE	9135295	A	ACTIVE	AV	VERIFIED	ABBOTT	SUSAN	HANKS	2876 CALLOWAY	DMEBANE	NC	27302	2876 CALLOWAY DR		MEBANE	NC	27302	919 568 8056	W	UN	UNA		
43	1	ALAMANCE	9113731	I	INACTIVE	IU	CONFIRMATI	ABBOTT	TAYLOR	RENEE	406 W LEBANON A	ELON	NC	27244	CAMPUS BOX 3077		ELON	NC	27244		W	UN	REP		
44	1	ALAMANCE	9120825	I	INACTIVE	IN	CONFIRMATI	ABBOTT	TIFFANY	MURIEL ARLE	144 W CRESCENT S	GRAHAM	NC	27253	144 W CRESCENT SQUARE		GRAHAM	NC	27253	336 233 0429	B	NL	DEM		
45	1	ALAMANCE	9013866	I	INACTIVE	IN	CONFIRMATI	ABBOTT	VIRGINIA	SMITH	2820 BLANCHE DR	BURLINGTON	NC	27215	2820 BLANCHE DR		BURLINGTON	NC	27215	584 4663	W	NL	REP		
46	1	ALAMANCE	9027717	A	ACTIVE	AV	VERIFIED	ABBOTT-LUN	SHELBY	LYNN	509 FERNWAY DR	BURLINGTON	NC	27217	509 FERNWAY DR		BURLINGTON	NC	27217	336 226 0087	B	NL	DEM		
47	1	ALAMANCE	9108552	A	ACTIVE	AV	VERIFIED	ABDALLA	KHALED	ISMAIL	605 ISLEY PL	#C BURLINGTON	NC	27215	605 ISLEY PL	#C	BURLINGTON	NC	27215	336 686 0506	W	NL	DEM		
48	1	ALAMANCE	9128403	A	ACTIVE	AV	VERIFIED	ABDEL-MAGI	LISA	ANN	1841 DUNBAR PL	BURLINGTON	NC	27215	1841 DUNBAR PL		BURLINGTON	NC	27215	214 437 8955	W	NL	UNA		
49	1	ALAMANCE	9117192	I	INACTIVE	IU	CONFIRMATI	ABDELKARIM	AMNA	ELHAG	1105 PROVIDENCE	ELON	NC	27244	1105 PROVIDENCE CT		ELON	NC	27244		M	NL	UNA		

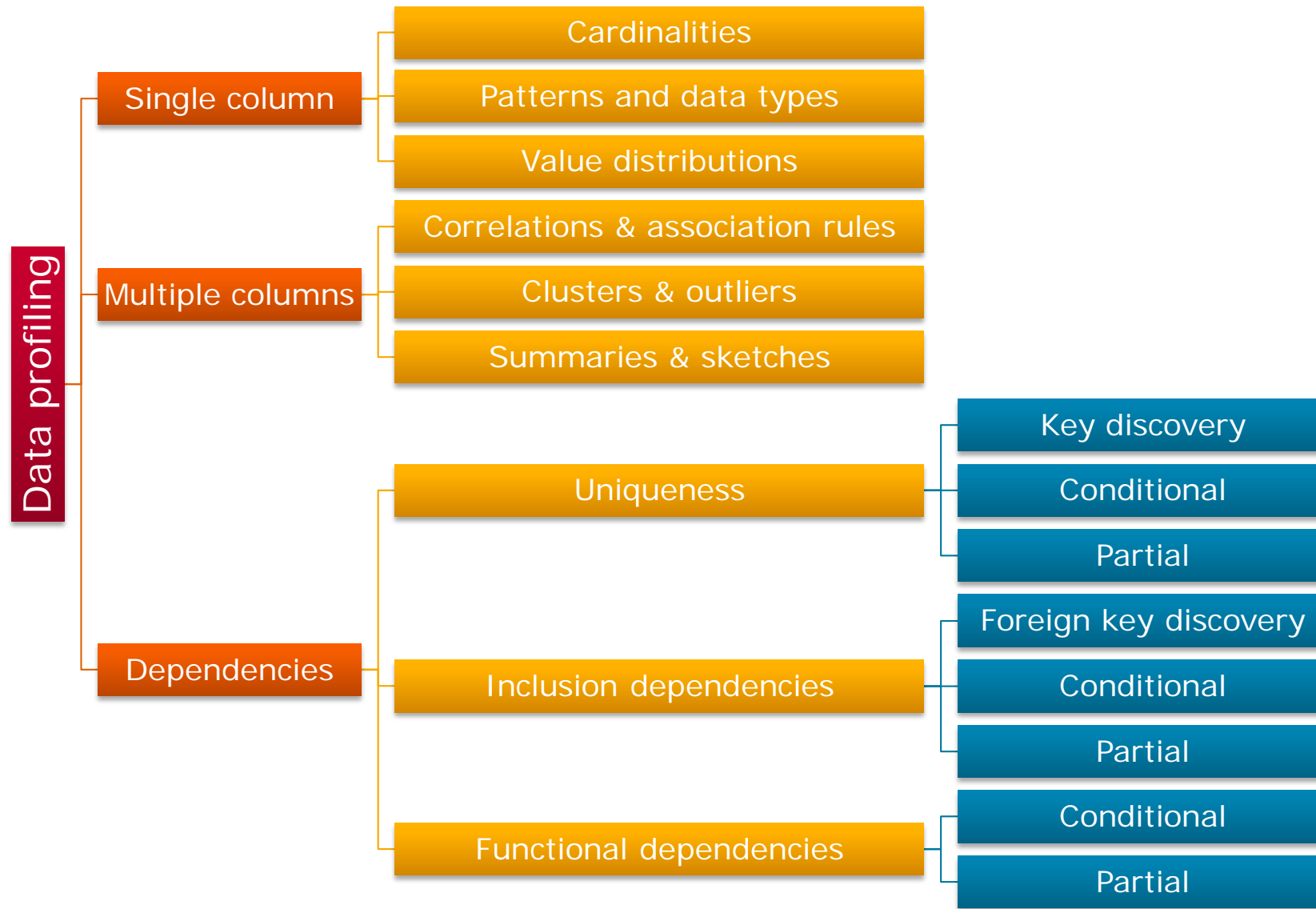
Column labels







# Data Profiling: Classification of Tasks



## Use Cases for Data Profiling

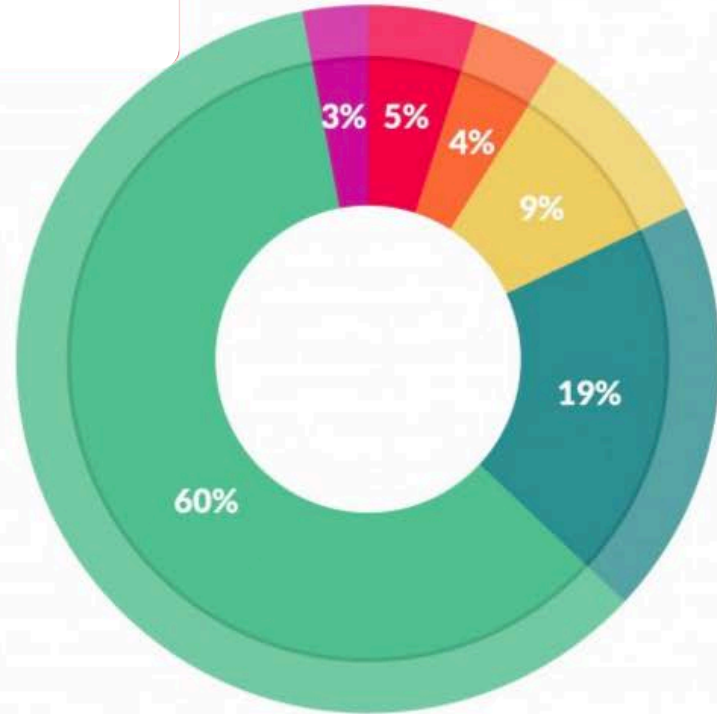
---

- **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



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



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Felix Naumann  
Data Profiling  
Canada, 2017

## 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

IBM Information Server File e.g., IBM Information Analyzer 9.43.86.77

IA\_OVERVIEW\_PROJECT

INVESTIGATE Foreign Key Analysis

Select Data Source to Work With

EMPLOYEE DEPARTMENT

Open Foreign Key Analysis View Details

You can use this pane to view analysis details about a primary key column and the foreign key column that is associated with the primary key column.

Frequency Values Analysis Details

Foreign Key Candidate Pair		
	Base Column	Paired Column
Column	EMPNO	MGRNO
Table	EMPLOYEE	DEPARTMENT
Source	IA	IA
Primary Key	Yes	No
Foreign Key	No	Yes
Data Class	Identifier	Quantity
Data Type	INT32	INT8
Length	0	0
Precision	0	0
Scale	0	0
Cardinality	48	9
Unique	No	No
Constant	No	No
Definition	No	No

Paired to Base:  
Common Data Values: 8 100.0000% Common Domain: Yes

Base to Paired:  
Common Data Values: 8 16.6667% Common Domain: No

Common Domain:

40 8 1

Base Column Paired Column

Close

## 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



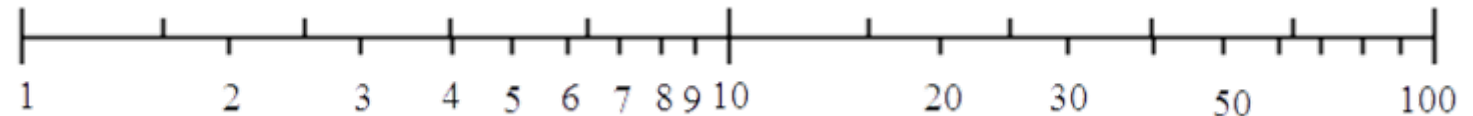
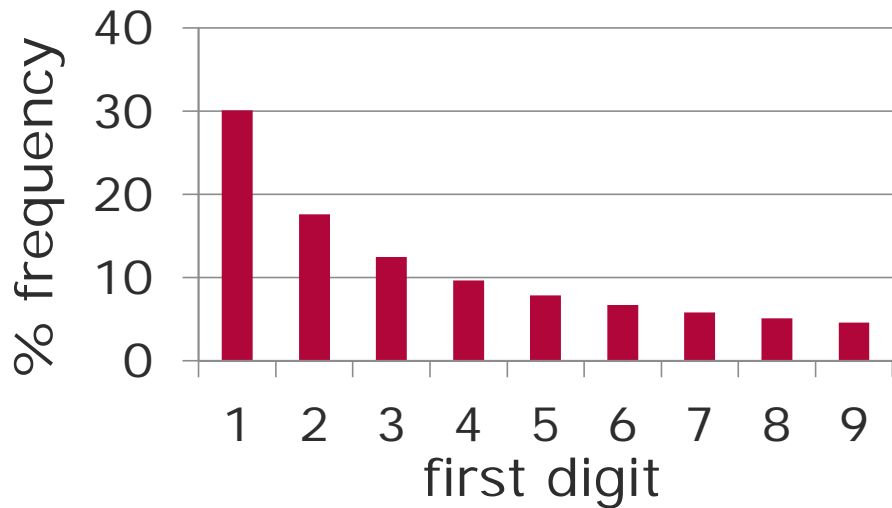
Felix Naumann  
Data Profiling  
Canada, 2017

## Cardinalities, distributions, and patterns

Category	Task	Description
<b>Cardinalities</b>	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
<b>Value distributions</b>	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)
<b>Patterns, data types, and domains</b>	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.

# 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

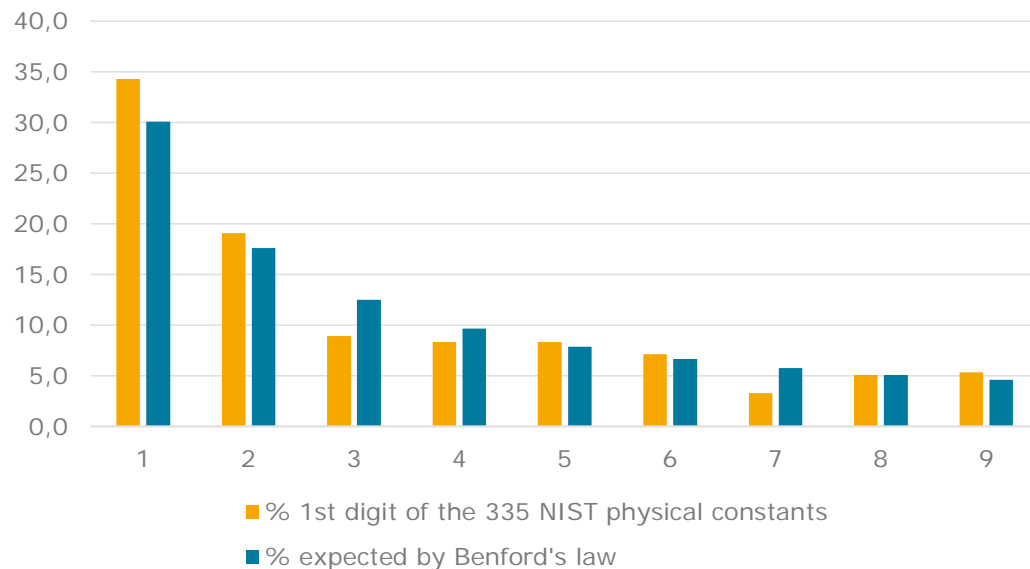


Felix Naumann  
Data Profiling  
Canada, 2017



# 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	
	Count	%
1	26	43.3%
2	7	11.7%
3	9	15.0%
4	6	10.0%
5	4	6.7%
6	1	1.7%
7	2	3.3%
8	5	8.3%
9	0	0.0%

In Benford's law
30.1%
17.6%
12.5%
9.7%
7.9%
6.7%
5.8%
5.1%
4.6%

[http://en.wikipedia.org/wiki/List\\_of\\_tallest\\_buildings\\_and\\_structures\\_in\\_the\\_world#Tallest\\_structure\\_by\\_category](http://en.wikipedia.org/wiki/List_of_tallest_buildings_and_structures_in_the_world#Tallest_structure_by_category)



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

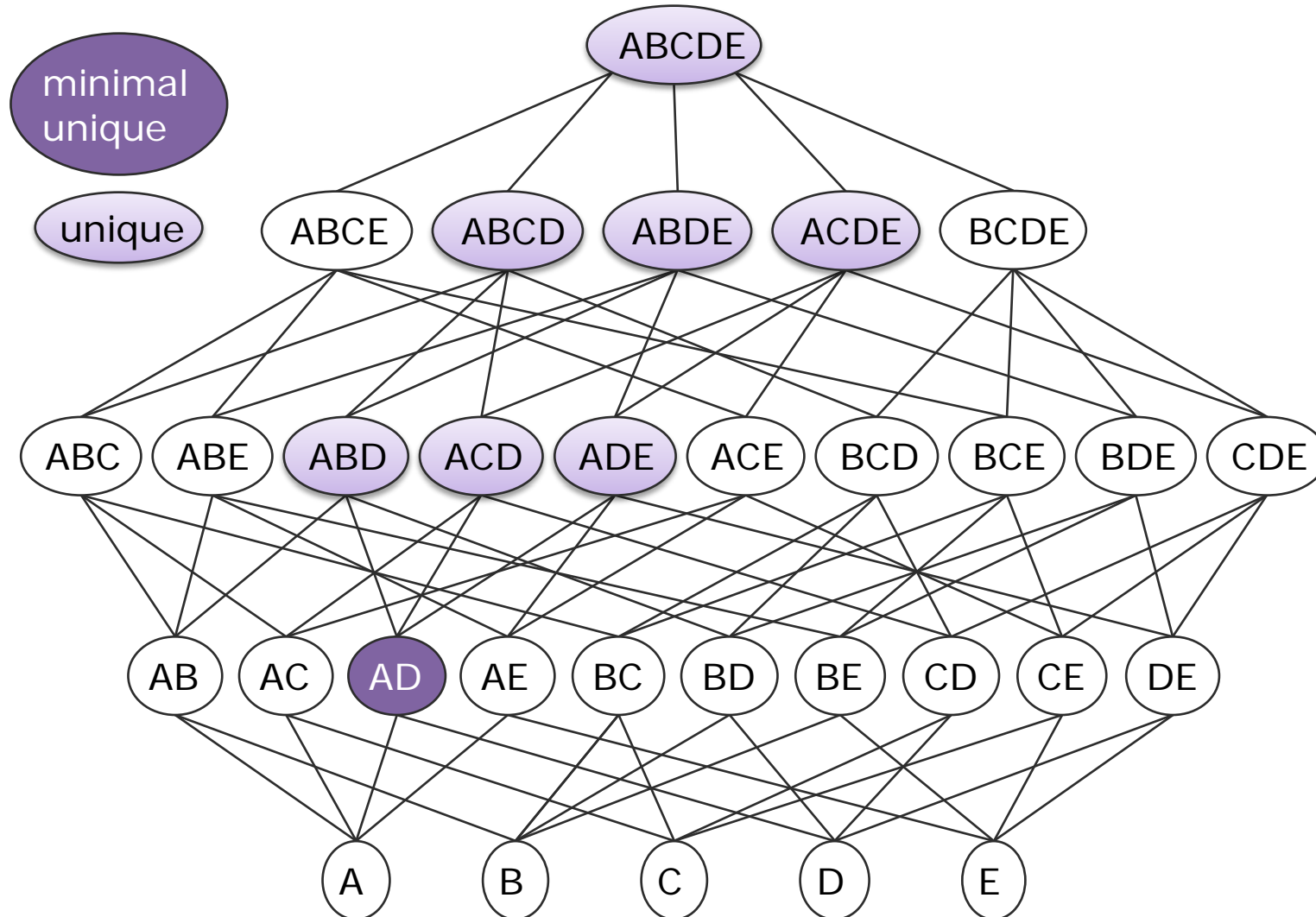
- Uniques: {A, AB, AC, BC, ABC}
- Minimal uniques: {A, BC}
- (Maximal) Non-uniques: {B, C}

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

Felix Naumann  
Data Profiling  
Canada, 2017

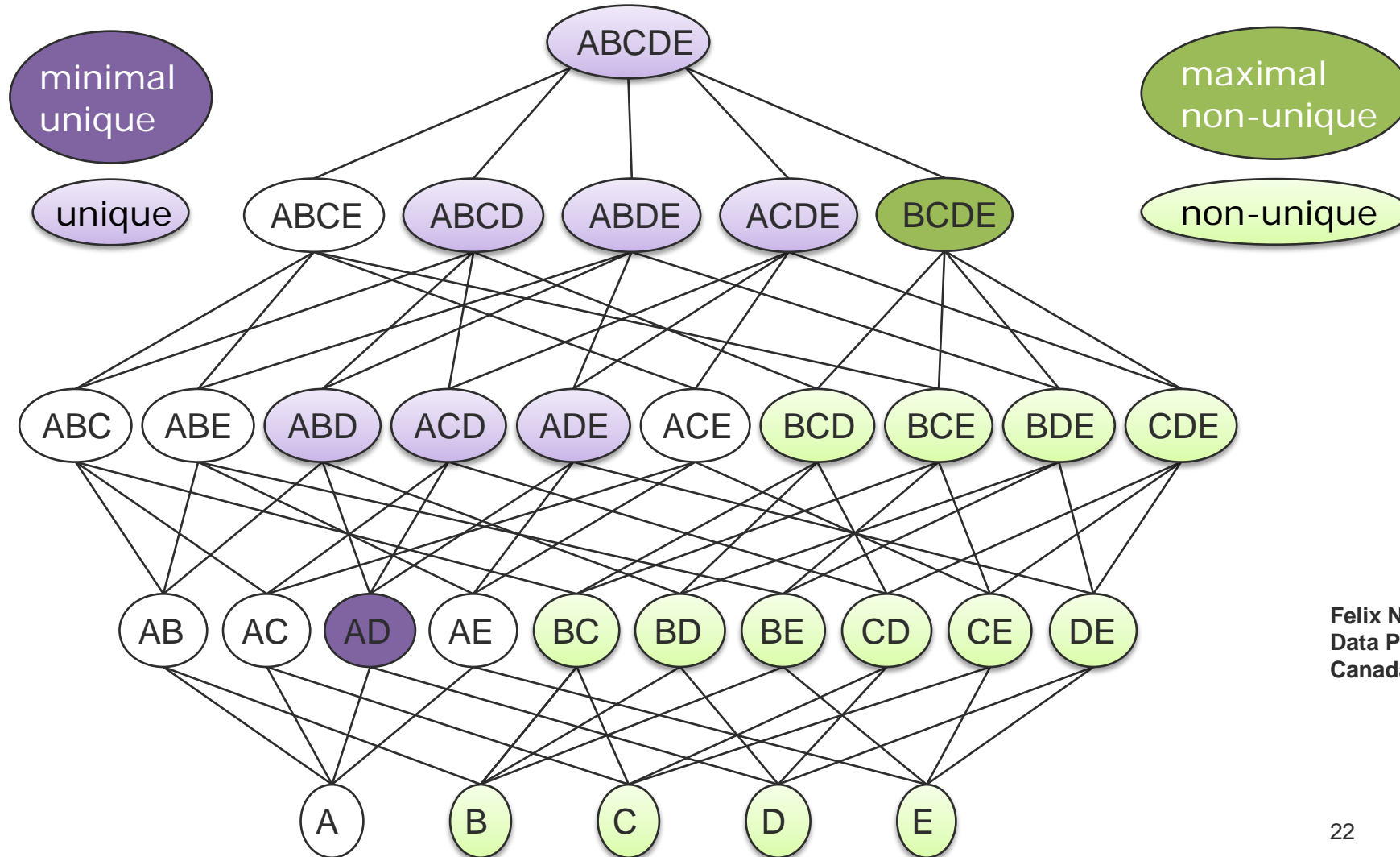


# Pruning effect of a pair



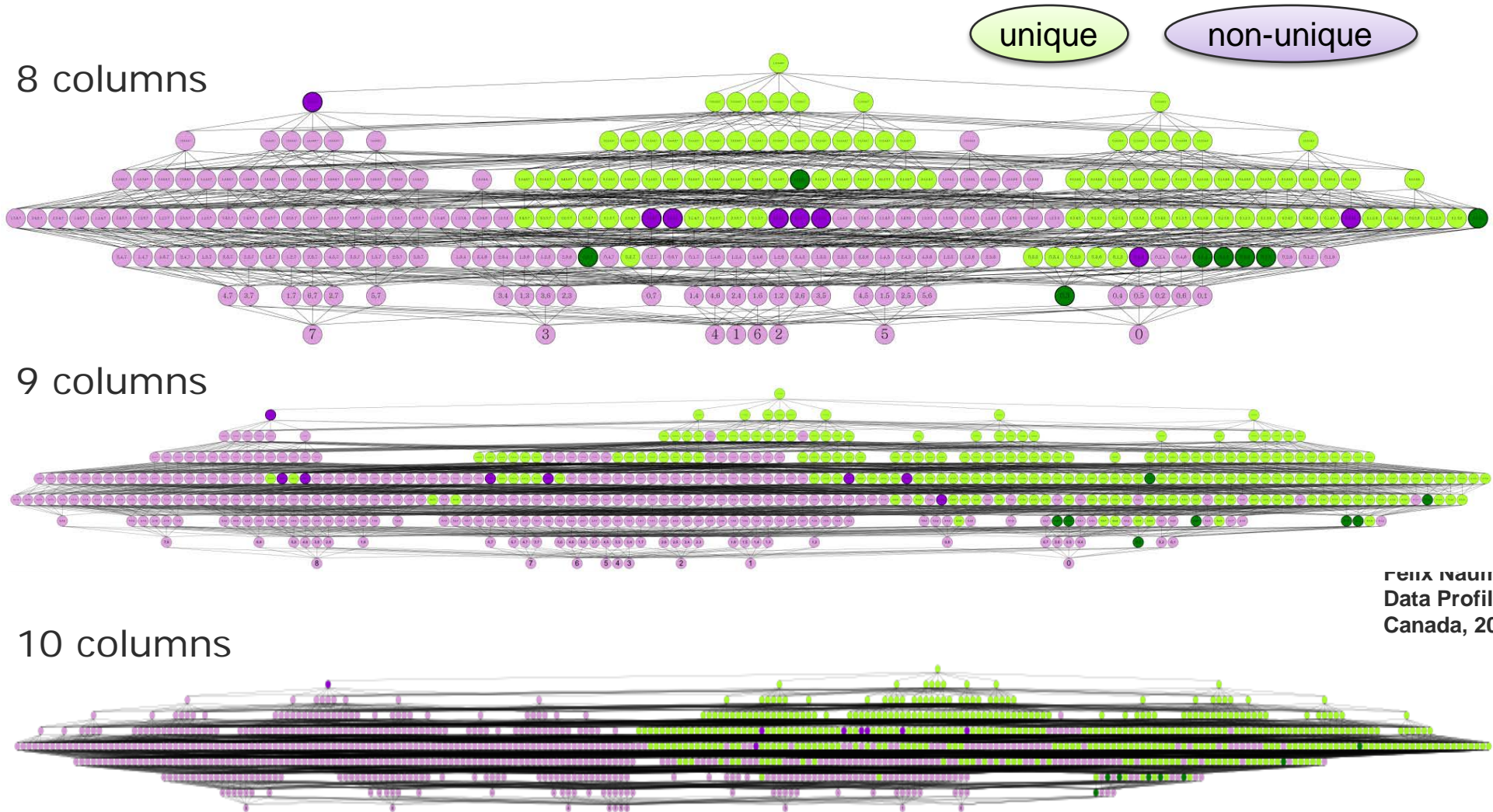
Felix Naumann  
Data Profiling  
Canada, 2017

# Pruning both ways

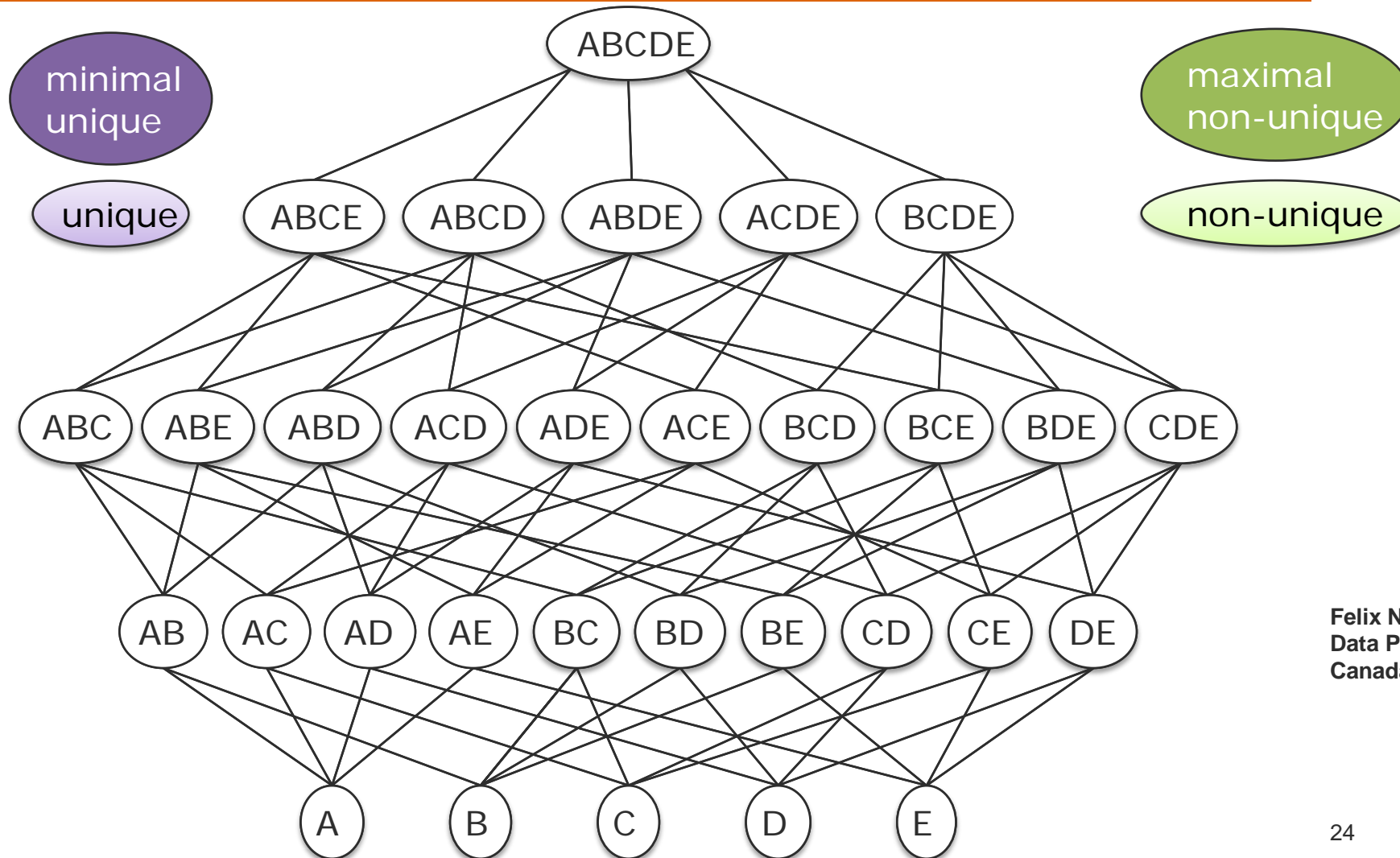


Felix Naumann  
Data Profiling  
Canada, 2017

# TPCH – Uniques and Non-Uniques



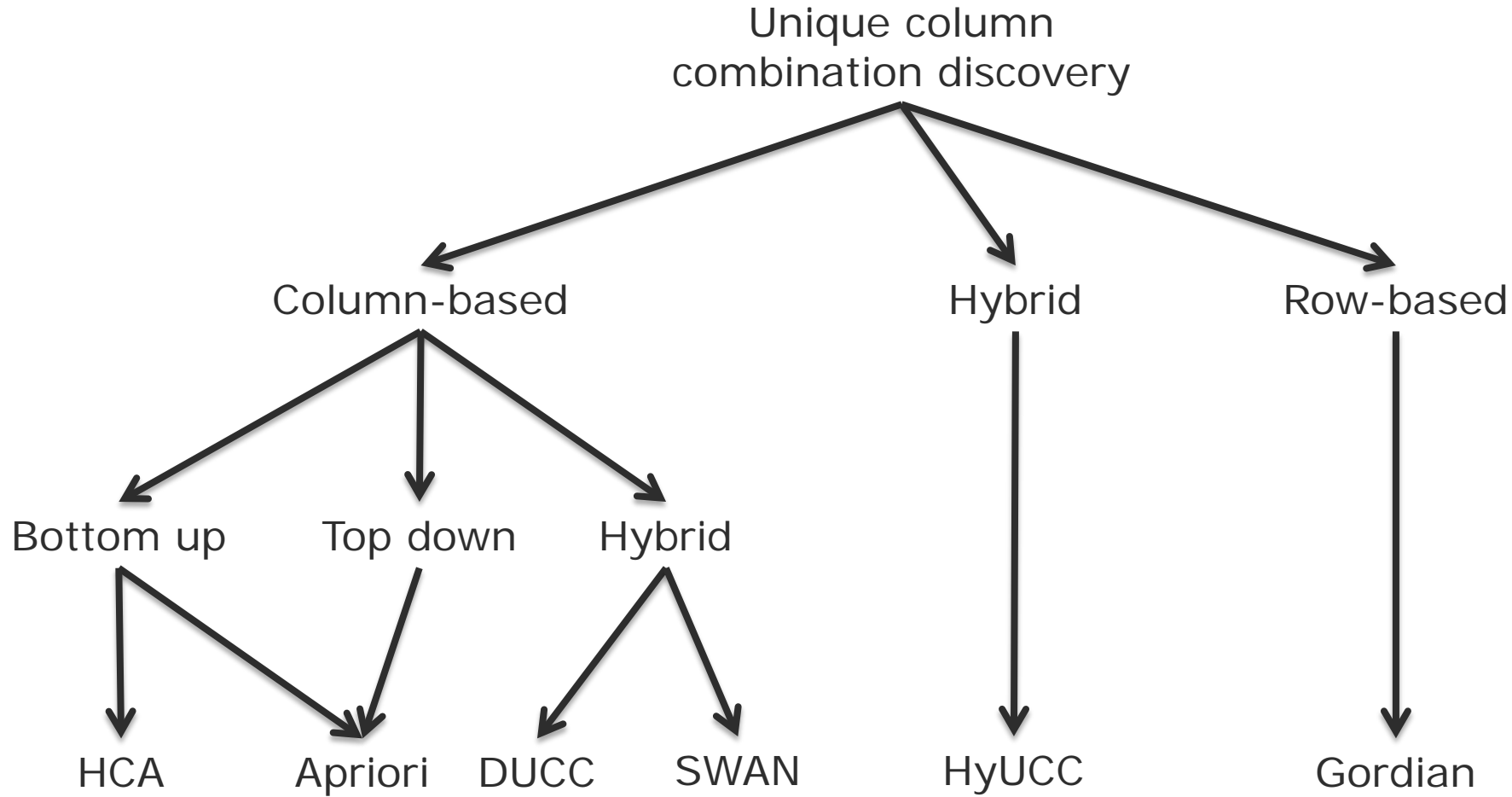
# Apriori visualized



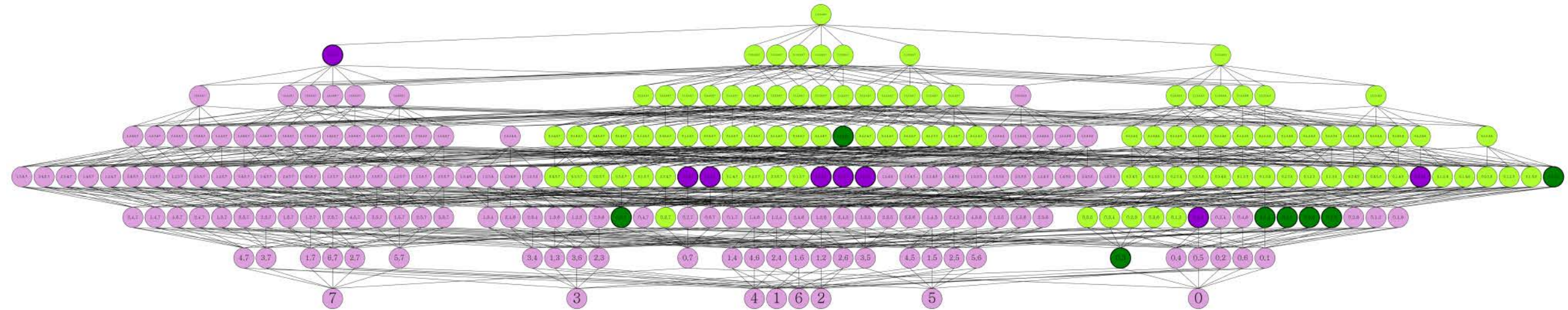
Felix Naumann  
Data Profiling  
Canada, 2017



# Discovery Algorithms



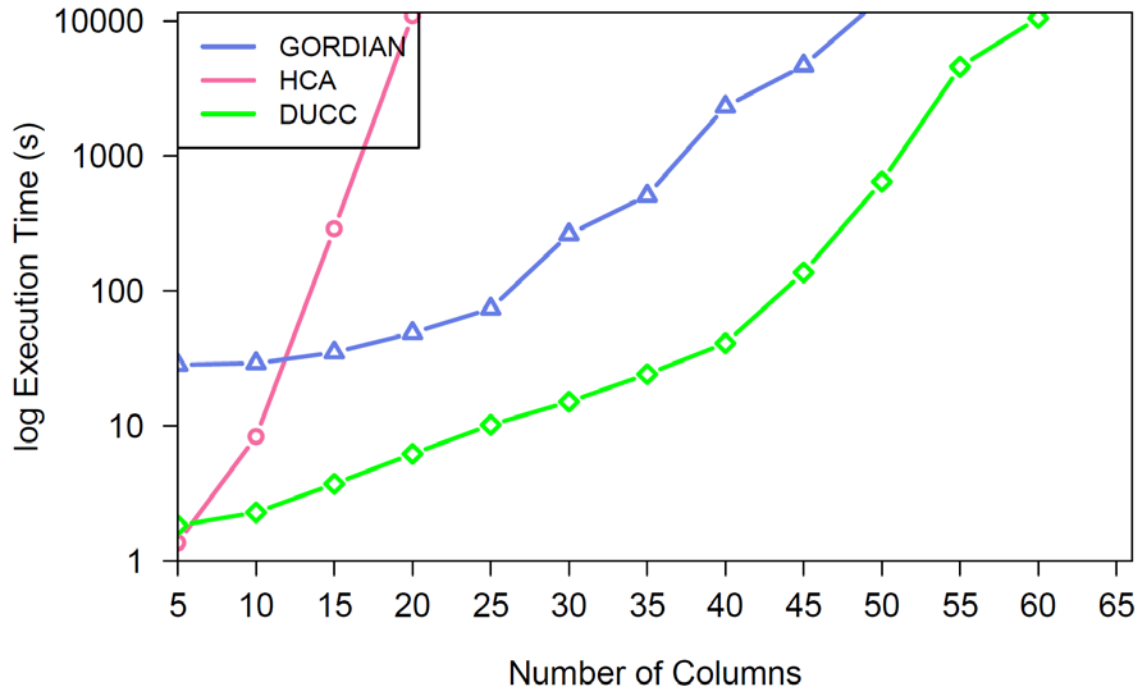
# DUCC – Detecting Unique Column Combinations



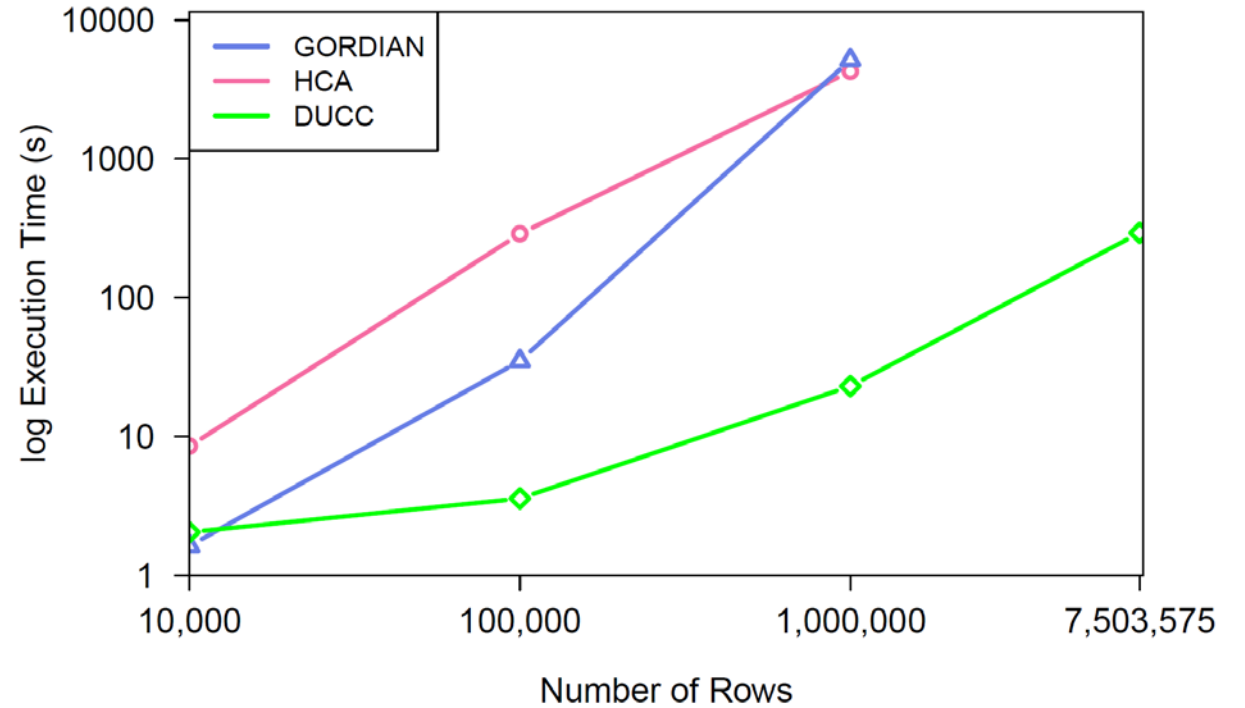
Felix Naumann  
Data Profiling  
Canada, 2017

# Scalability in the number of columns and rows

■ NCVoter data, 100k rows

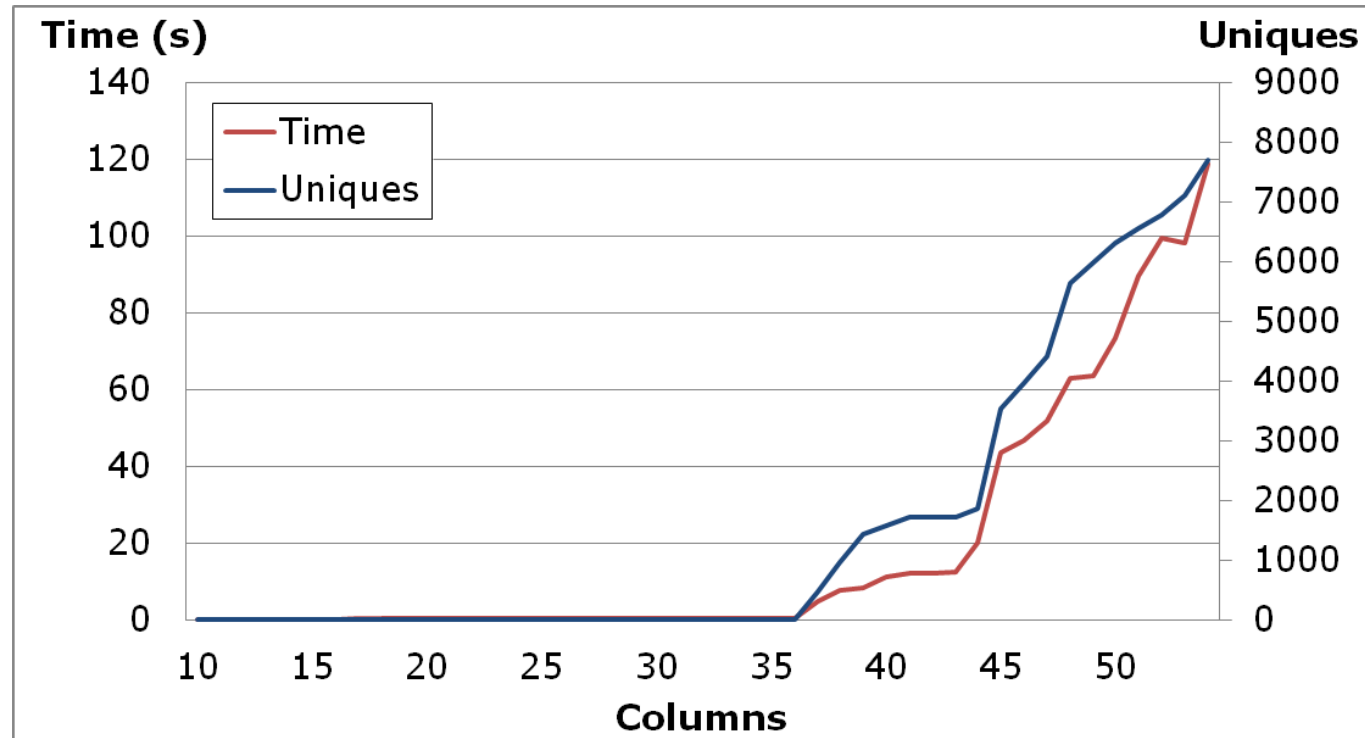


■ NCVoter, 15 columns



# Analysis of DUCC

- Runtime mainly depends on size of solution set



Felix Naumann  
Data Profiling  
Canada, 2017

- 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\_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

# 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

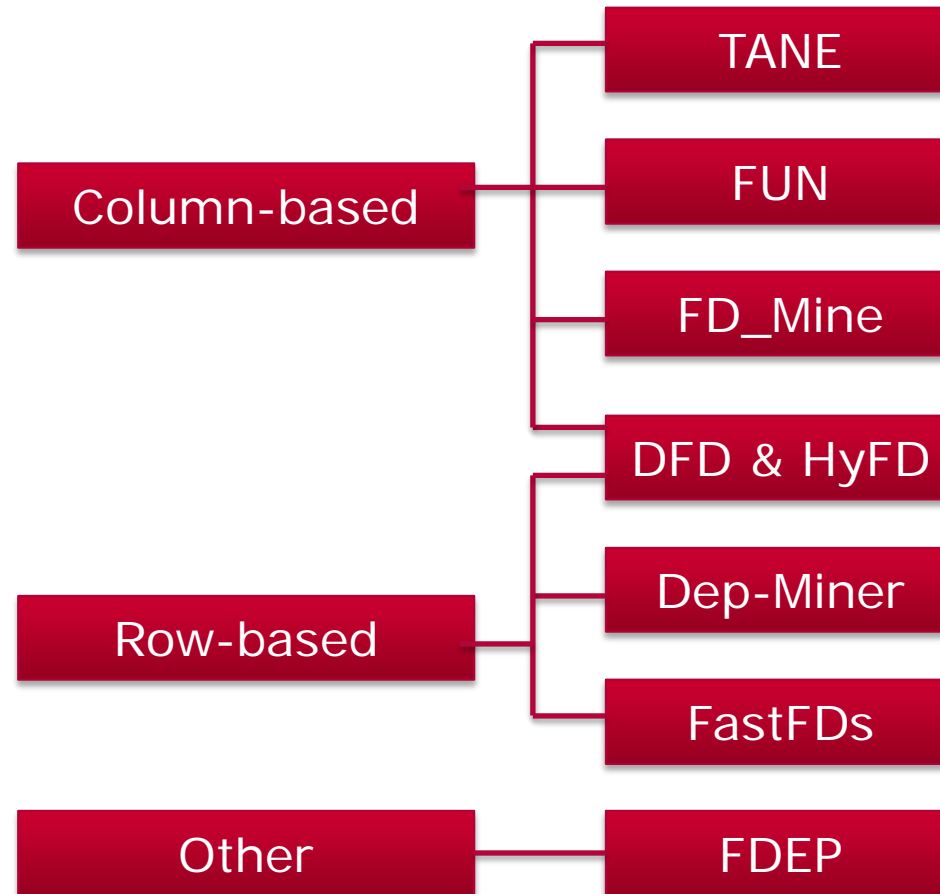
---

- 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

### Naive discovery approach

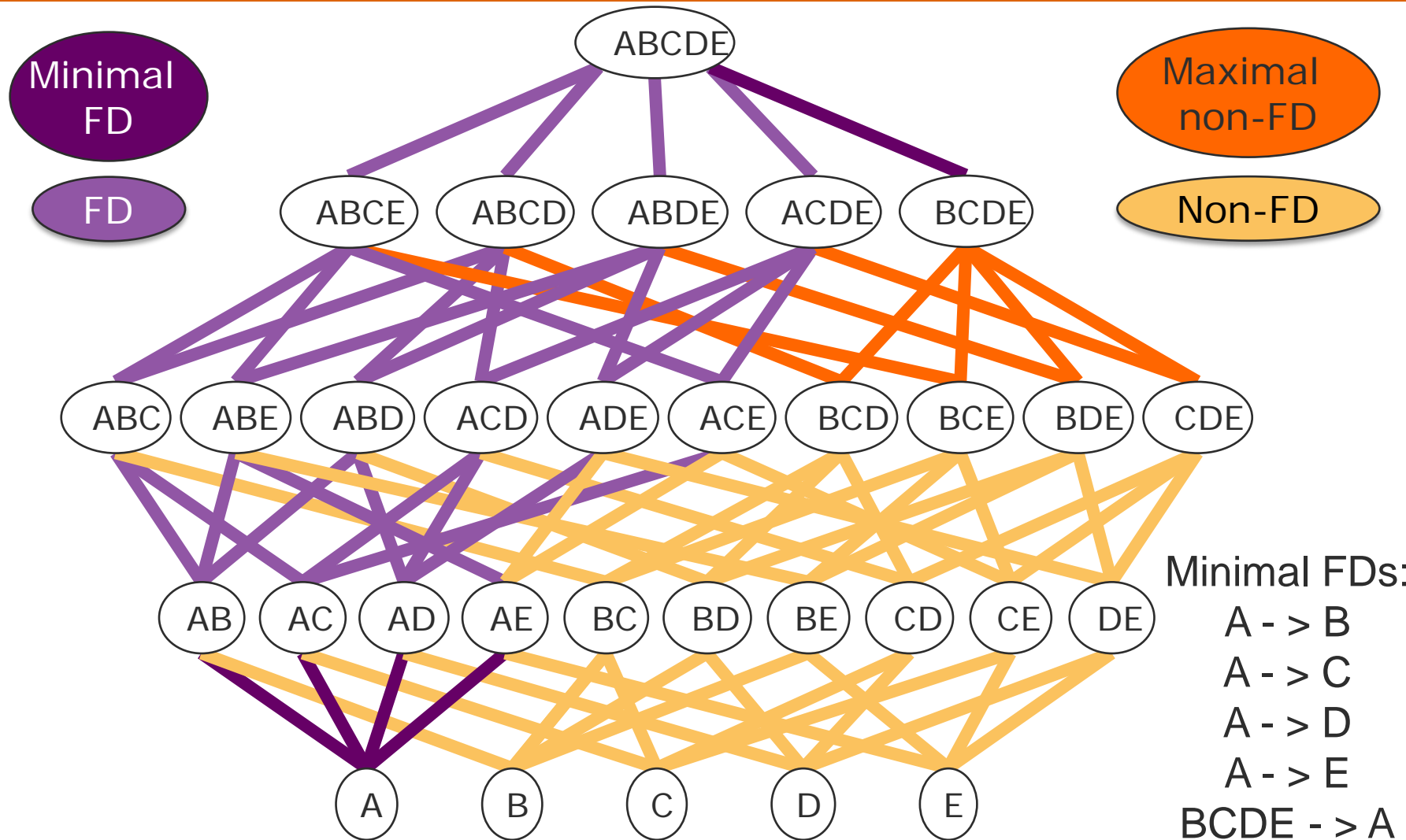
- For each column combination  $X$ 
  - For each pair of tuples  $(t_1, t_2)$ 
    - If  $t_1[X \setminus A] = t_2[X \setminus A]$  and  $t_1[A] \neq t_2[A]$ : Break
- Complexity
  - Exponential in number of attributes times number of rows squared

## Current FD discovery approaches



Felix Naumann  
Data Profiling  
Canada, 2017

Again: Model in lattice – edges represent FDs



Minimal FD

FD

Maximal non-FD

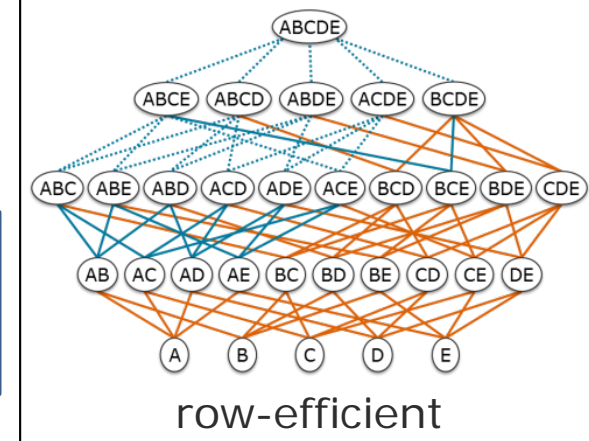
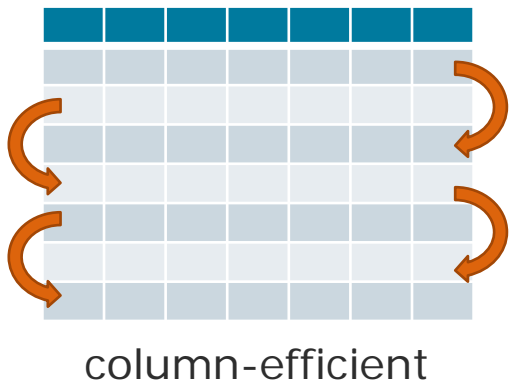
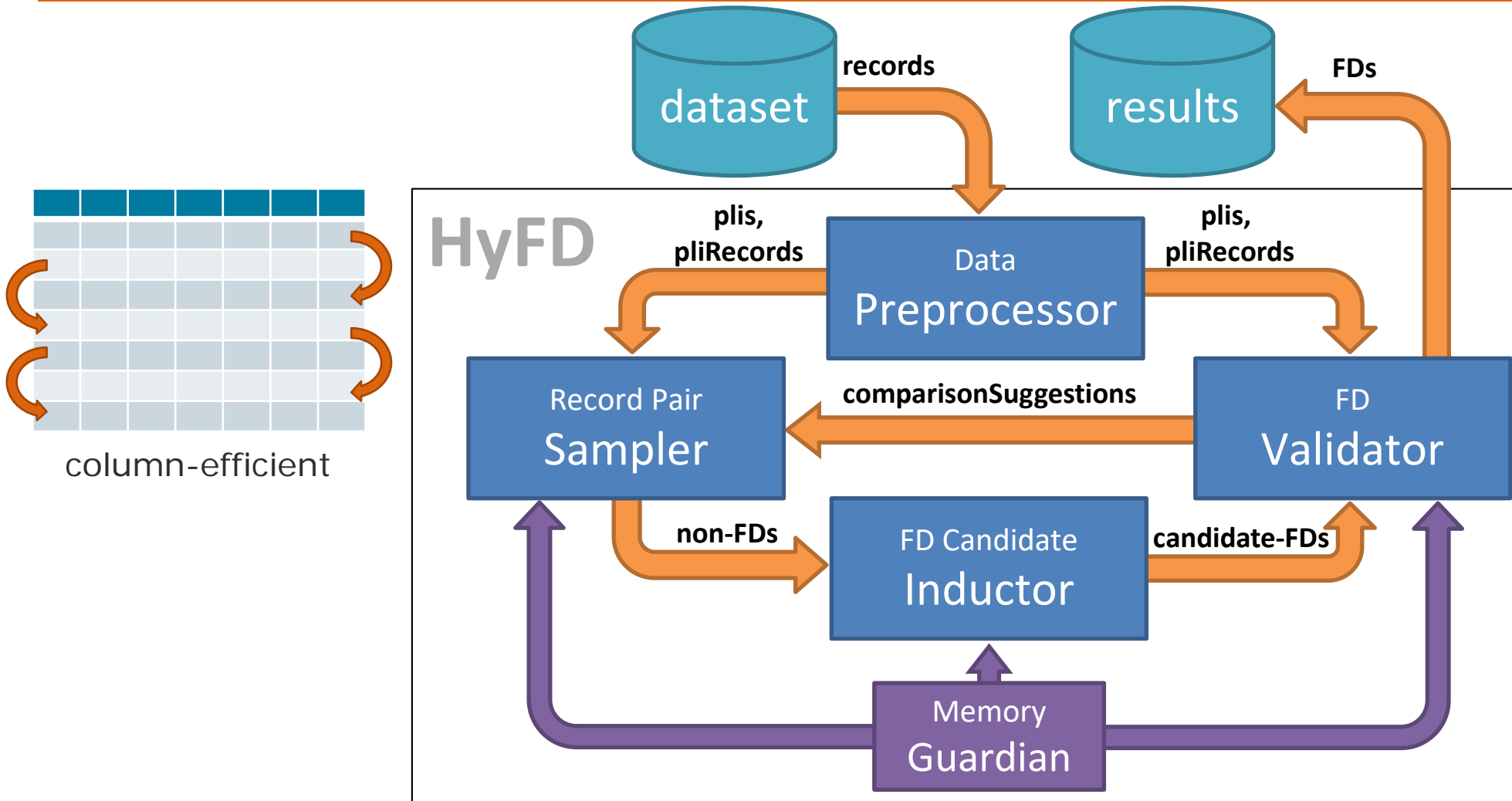
Non-FD

Minimal FDs:

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

Felix Naumann  
Data Profiling  
Canada, 2017

# HyFD: Hybrid FD Discovery



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	<b>0.1</b>	0.2	0.2	0.2	0.2	<b>0.1</b>	<b>0.1</b>
balance-scale	5	625	7	1	1.2	<b>0.1</b>	0.2	0.3	0.3	0.3	0.2	<b>0.1</b>
chess	7	28,056	519	1	2.9	1.1	3.8	1.0	174.6	164.2	125.5	<b>0.2</b>
abalone	9	4,177	187	137	2.1	0.6	1.8	1.1	3.0	2.9	3.8	<b>0.2</b>
nursery	9	12,960	1,024	1	4.1	1.8	7.1	0.9	121.2	118.9	46.8	<b>0.5</b>
breast-cancer	11	699	20	46	2.3	0.6	2.2	0.8	1.1	1.1	0.5	<b>0.2</b>
bridges	13	108	6	142	2.2	0.6	4.2	0.9	0.5	0.6	0.2	<b>0.1</b>
echocardiogram	13	132	6	527	1.6	0.4	69.9	1.2	0.5	0.5	0.2	<b>0.1</b>
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	<b>3.4</b>
ncvoter	19	1,000	151	758	4.3	4.0	ML	5.1	11.4	1.9	1.1	<b>0.4</b>
hepatitis	20	155	8	8,250	12.2	175.9	ML	326.7	5576.5	9.5	0.8	<b>0.6</b>
horse	27	368	25	128,727	457.0	TL	ML	TL	TL	385.8	7.2	<b>7.1</b>
fd-reduced-30	30	250,000	69,581	89,571	<b>41.1</b>	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	<b>21.8</b>
flight	109	1,000	575	982,631	ML	ML	ML	TL	TL	TL	216.5	<b>53.4</b>
uniprot	223	1,000	2,439	>2,437,556	ML	ML	ML	TL	TL	TL	ML	<b>&gt;5254.7</b>

Results larger than 1,000 FDs are only counted

TL: time limit of 4 hours exceeded

ML: memory limit of 100 GB exceeded



## IND discovery $R[X] \subseteq S[Y]$

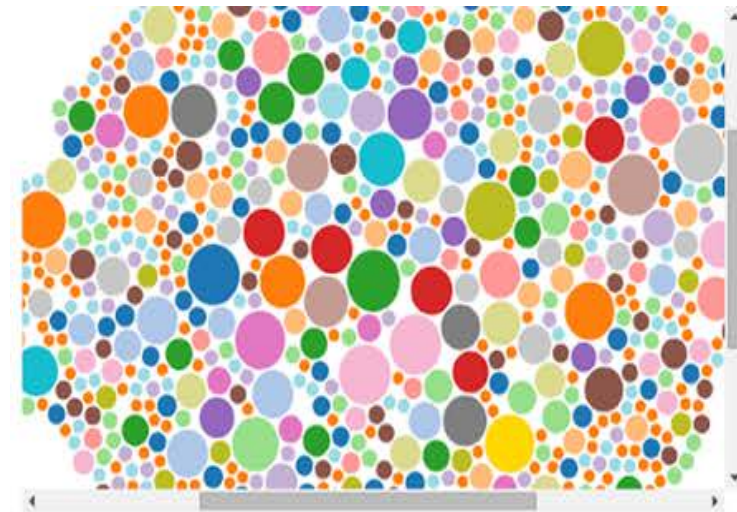
- Unary and n-ary INDs:  $R[A] \subseteq S[B]$  and  $R[ABC] \subseteq S[DEF]$
- Detect unknown foreign keys
- Example: PDB – Protein Data Bank
  - OpenMMS provides relational schema, 175 tables
  - Not a single foreign key constraint!
- Example: Ensembl – genome database
  - Shipped as MySQL dump files: >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_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.'Astrolgy_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







## Many Other Kinds of Dependencies

**dependency**, 157  
 afunctional, 234  
 algebraic, 228–233  
 axiomatization, 166, 171, 172, 186, 193, 202–207, 227, 231  
 capturing semantics, 159–163  
 classification, 218  
 conditional table, 497  
 and data integrity, 162  
 and domain independence, 97  
 dynamic, 234  
 embedded, 192, 217, 233  
 embedded implicational (eid), 233  
 embedded join (ejd), 218, 233  
 embedded multivalued (emvd), 218, 220, 233  
 equality-generating (egd), 217–228  
 extended transitive, 234  
 faithful, 232, 233, 239  
 finiteness, 306  
 full, 217  
 functional (fd), 28, 159, 163–169, 163, 186, 218, 250, 257, 260

general, 234  
 generalized dependency constraints, 234  
 generalized mutual, 234  
 implication  
   in view, 221  
 implication of, 160, 164, 193, 197  
 implicational (id), 233  
 implied, 234  
 inclusion (ind), 161, 192–211, 193, 218, 250  
   acyclic, 207, 208–210, 211, 250  
   key-based, 250, 260  
   typed, 213  
   unary (uind), 210–211  
 inference rule, 166, 172, 193, 227, 231  
   ground, 203  
 join (jd), 161, 169–173, 170, 218  
 key, 157, 163–169, 163, 267  
 logical implication of, 160, 164  
   finite, 197  
   unrestricted, 197  
 multivalued (mvd), 161, 169–173, 170, 186, 218  
 mutual, 233  
 named vs. unnamed perspectives, 159  
 order, 234  
 partition, 234

projected join, 233  
 and query optimization, 163  
 satisfaction, 160  
 satisfaction by tableau, 175  
 satisfaction family, 174  
 and semantic data models, 249–253  
 and schema design, 253–262  
 single-head vs. multi-head, 217  
 sort set, 191, 213, 234  
 subset, 233  
 tagged, 164, 221, 241  
 template, 233, 236  
 transitive, 234  
 trivial, 220  
 tuple-generating (tgd), 217–228  
 typed, 159  
   vs. untyped, 192, 217  
 unirelational, 217  
 and update anomalies, 162  
 and views, 221, 222  
 vs. first-order logic, 159, 234  
 vs. integrity constraint, 157  
 vs. tableaux, 218, 234  
 dependency basis, 172  
 dependency preserving decomposition, 254  
 dependent class, 246  
 dereferencing, 557, 558  
 derivation, 290

[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

RFD abbrev.	RFD name
ACOD	Approximate comparable dependency
ADD	Approximate differential dependency
AFD	Approximate functional dependency
COD	Comparable dependency
CFD	Conditional functional dependency
CFD <sup>P</sup>	CFD with built-in predicates
CFD <sup>C</sup>	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 <sub>K</sub>	OD satisfied within bound $k$
OD <sub>EA</sub>	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 $\sigma$ and $\theta$ approximation

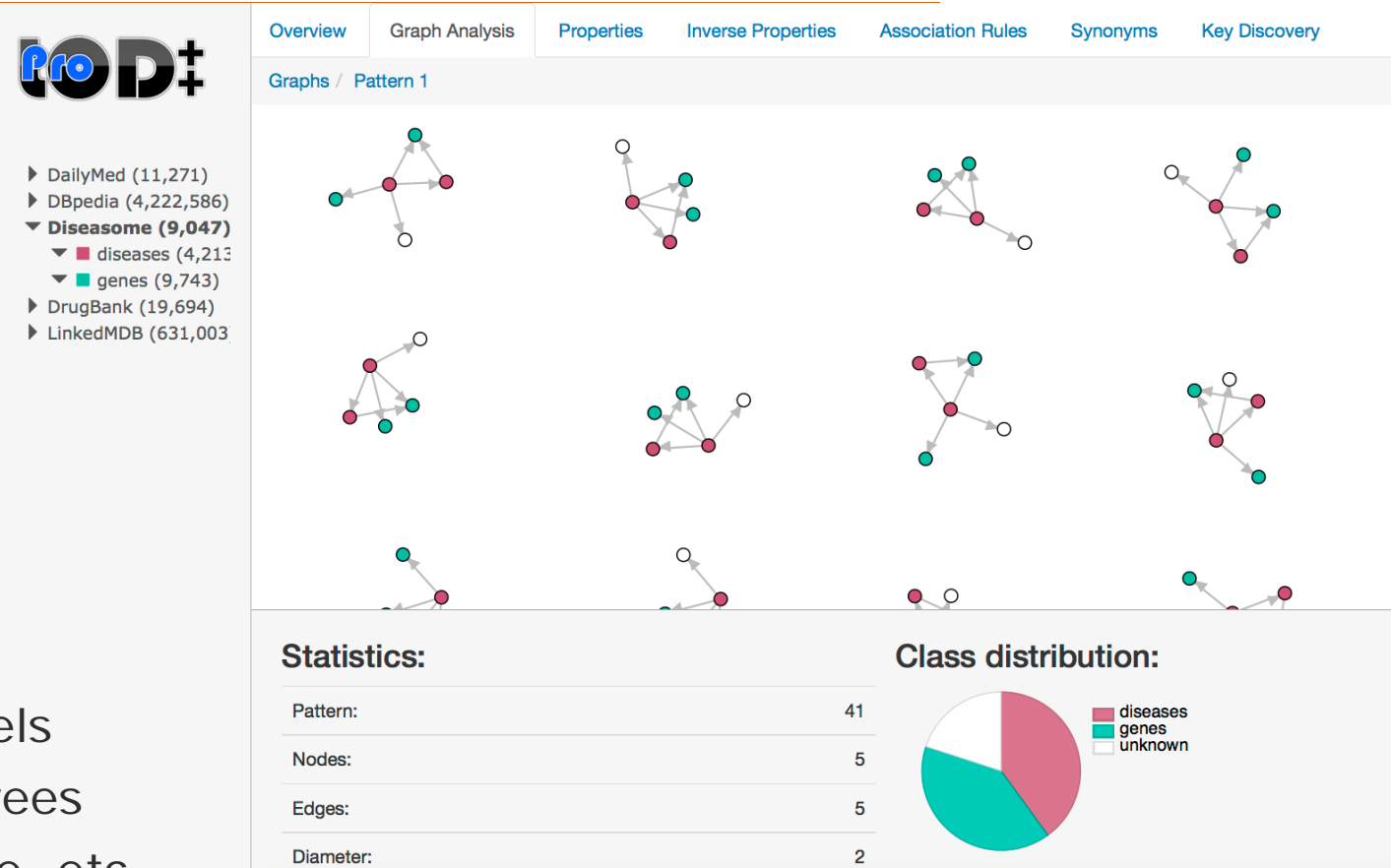
## Conditional Dependencies

---

- 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



**ProD+**

- ▶ DailyMed (11,271)
- ▶ DBpedia (4,222,586)
- ▼ **Diseasome (9,047)**
  - ▼ diseases (4,213)
  - ▼ genes (9,743)
- ▶ DrugBank (19,694)
- ▶ LinkedMDB (631,003)

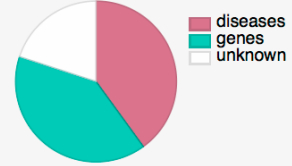
Overview | Graph Analysis | Properties | Inverse Properties | Association Rules | Synonyms | Key Discovery

Graphs / Pattern 1

**Statistics:**

Pattern:	41
Nodes:	5
Edges:	5
Diameter:	2

**Class distribution:**



- diseases
- genes
- unknown



## Outlook: Profiling Challenges

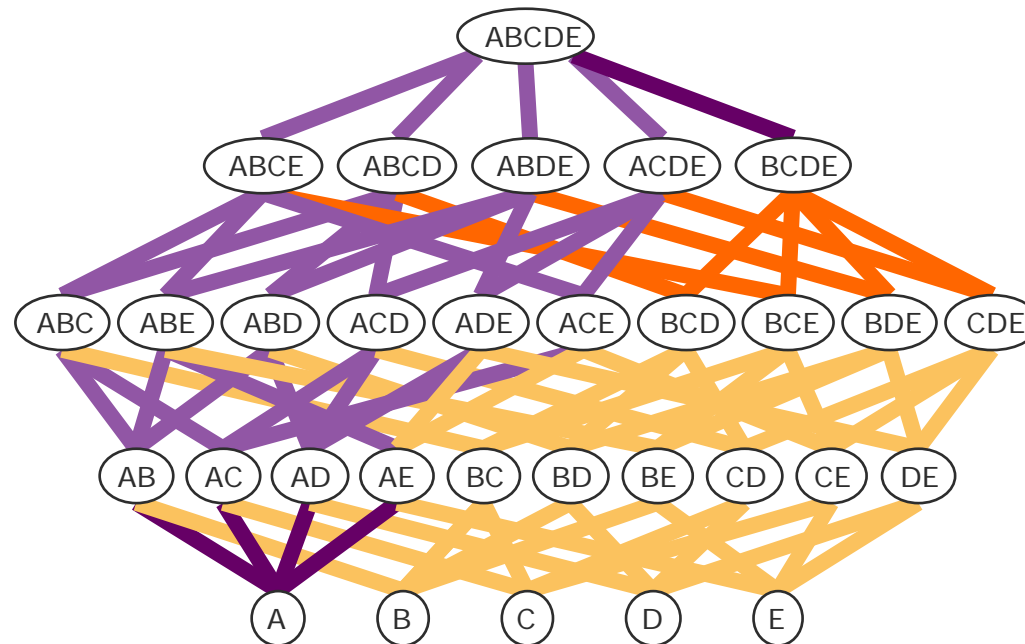
---

- 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

## 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
- RFind: 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

**Felix Naumann**  
**Data Profiling**  
**Canada, 2017**