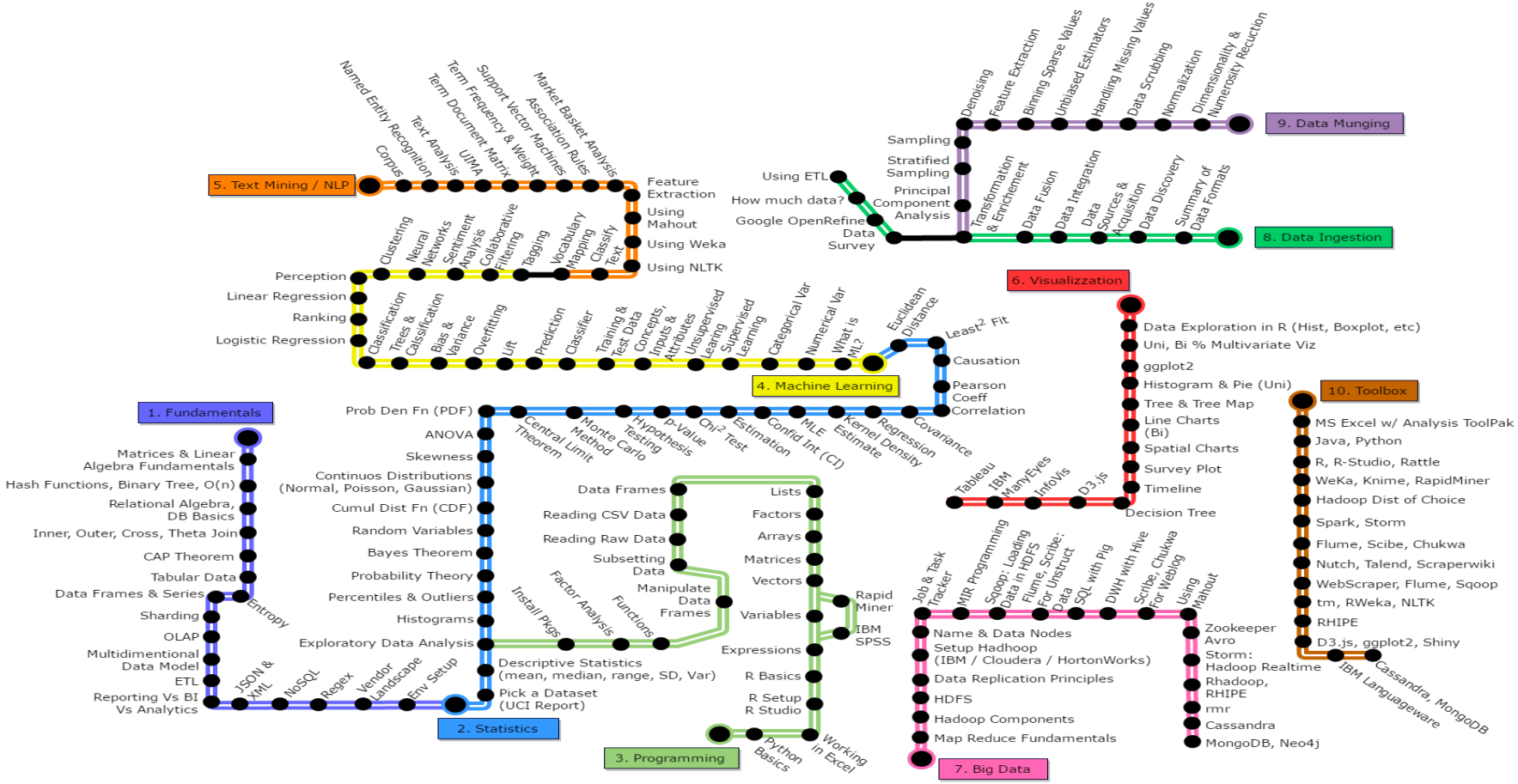
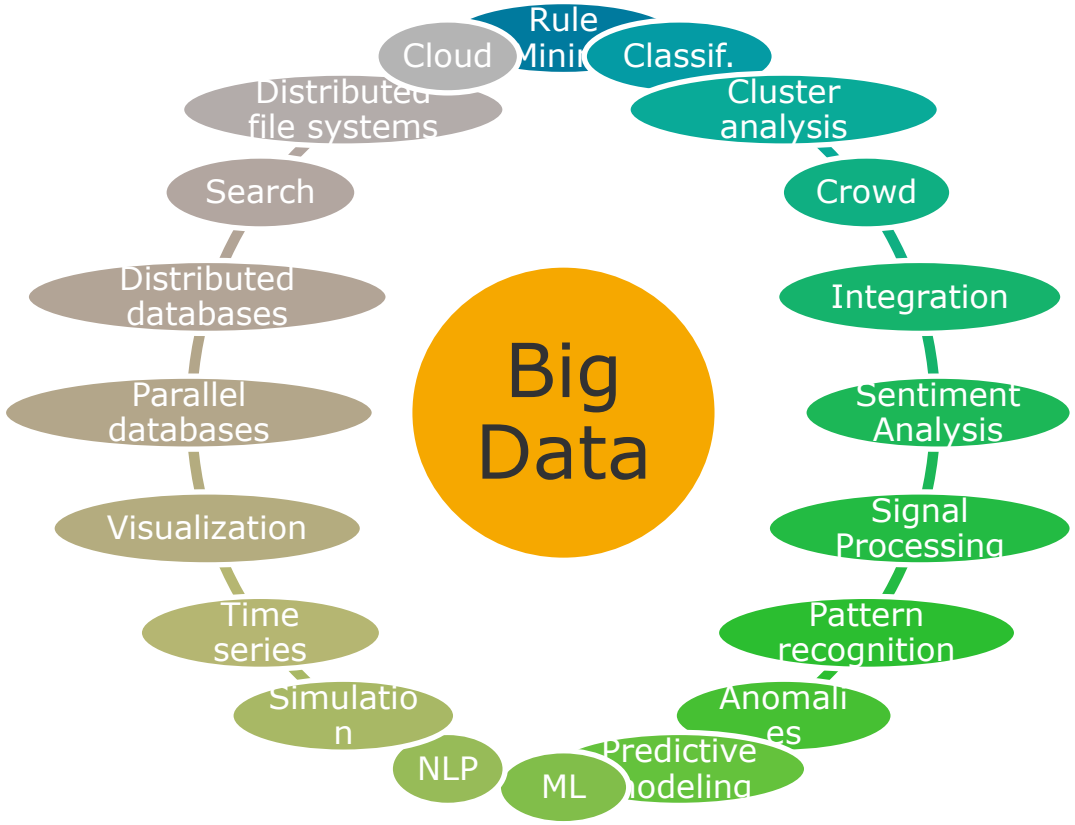


Fondamenti di Data Science e Machine Learning

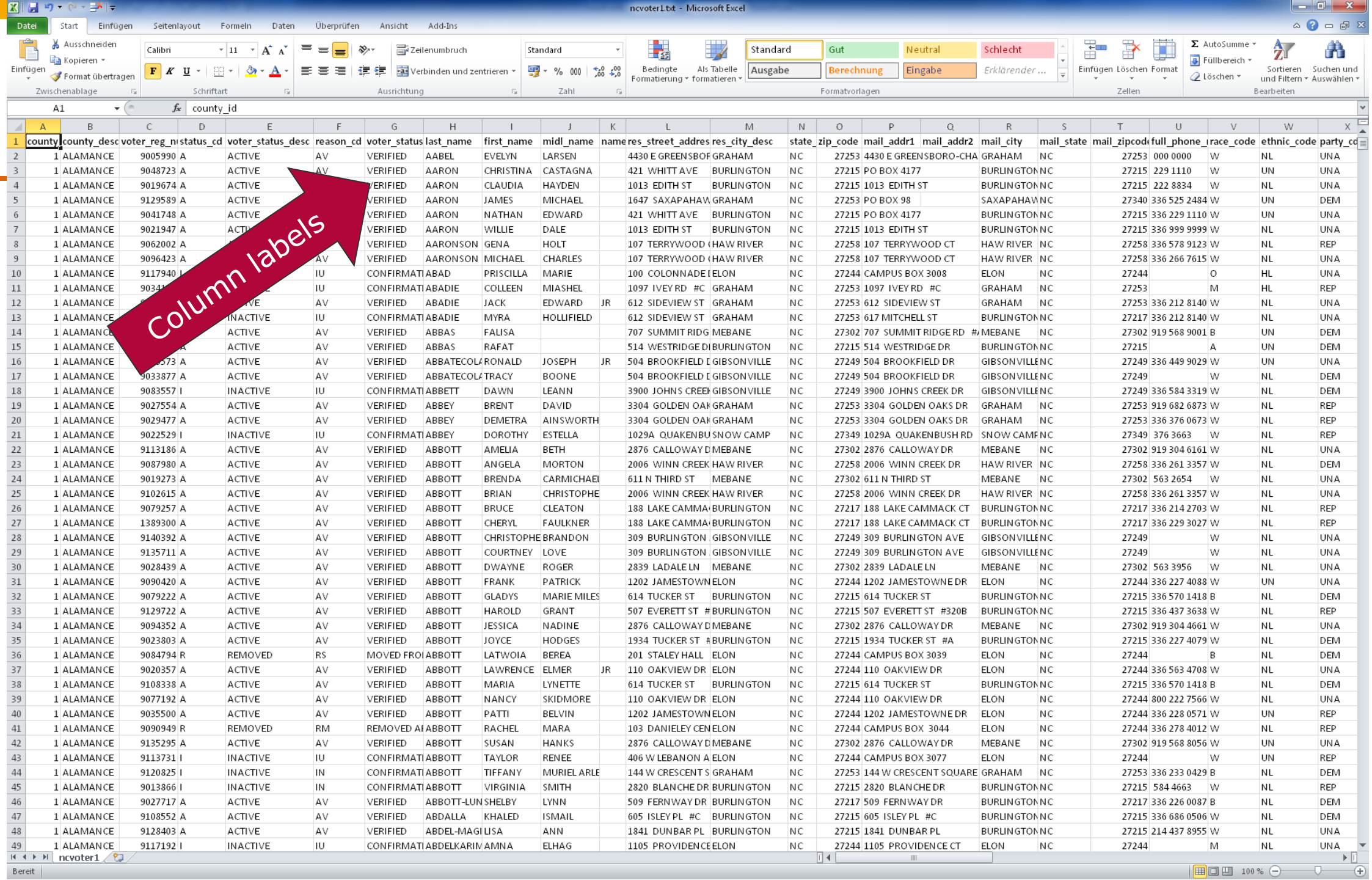
Data Profiling (Seminar Prof. F. Naumann, HPI, Potsdam, Germany)

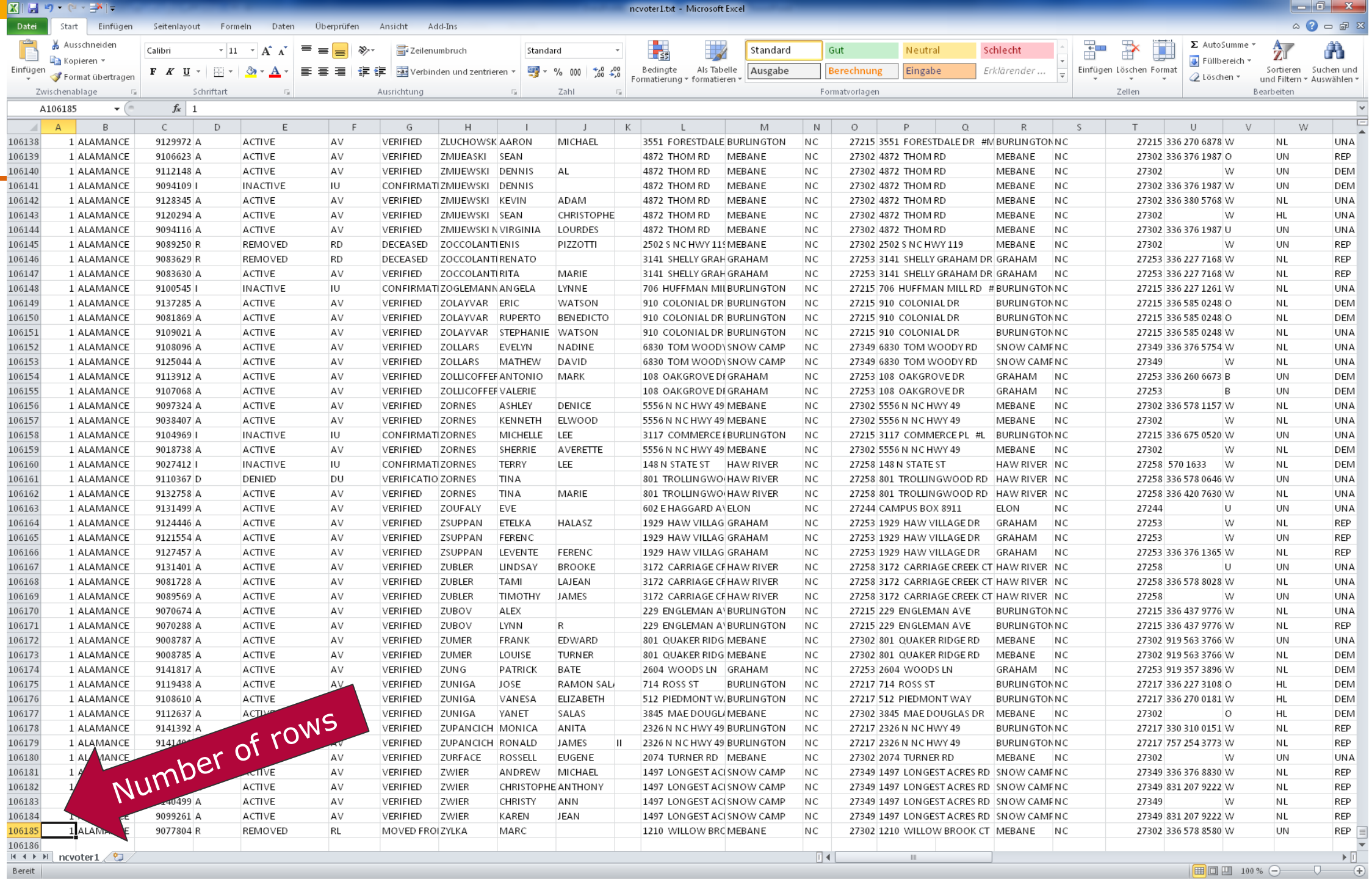
Lecturer: Prof. Giuseppe Polese, aa 2024-25

Technologies to approach big data/data science



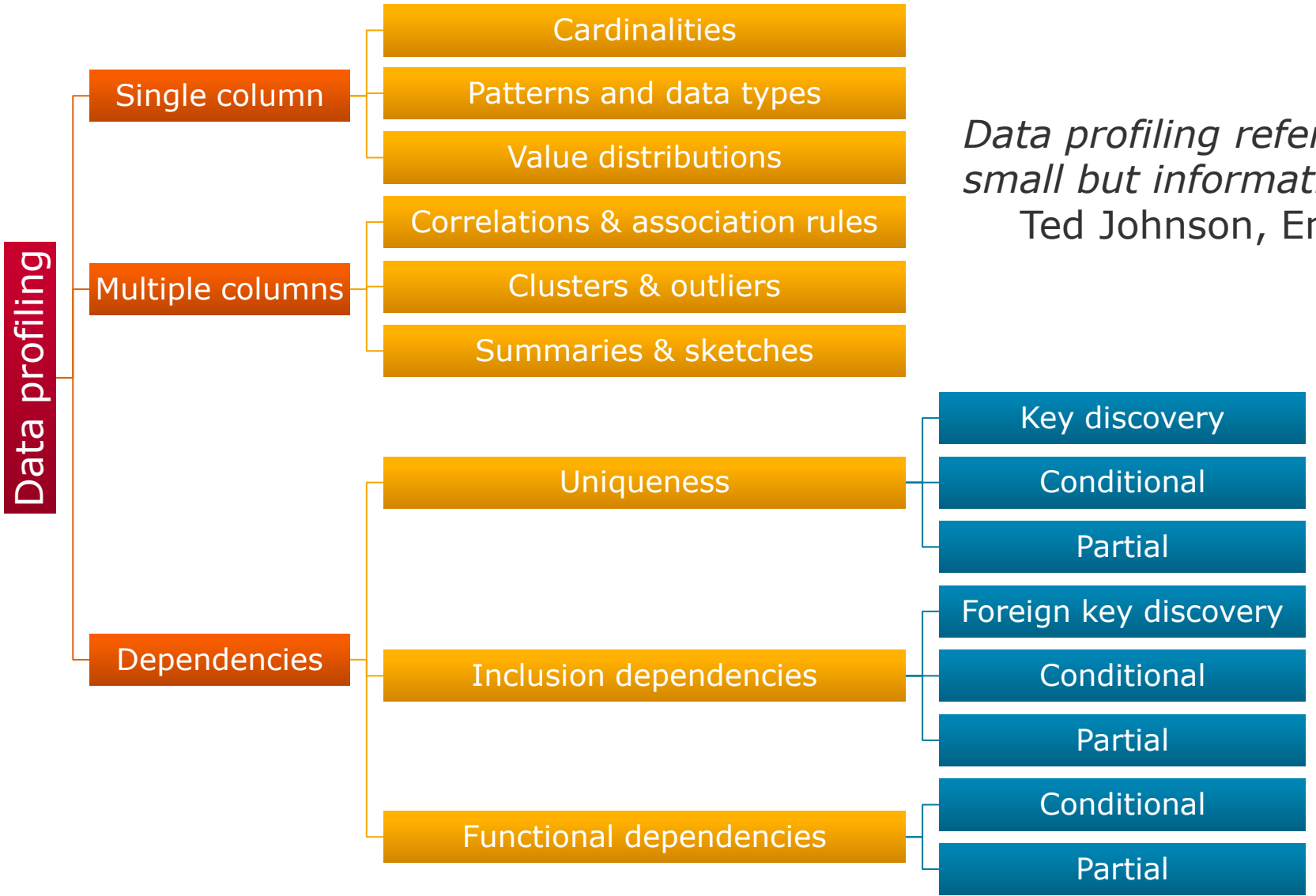
Data profiling and data cleansing are prerequisites for all of these!


ncvoter1.txt - Microsoft Excel
Column labels


ncvoter1.txt - Microsoft Excel
Number of rows


Patterns

Data Profiling: Classification of Tasks



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

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**

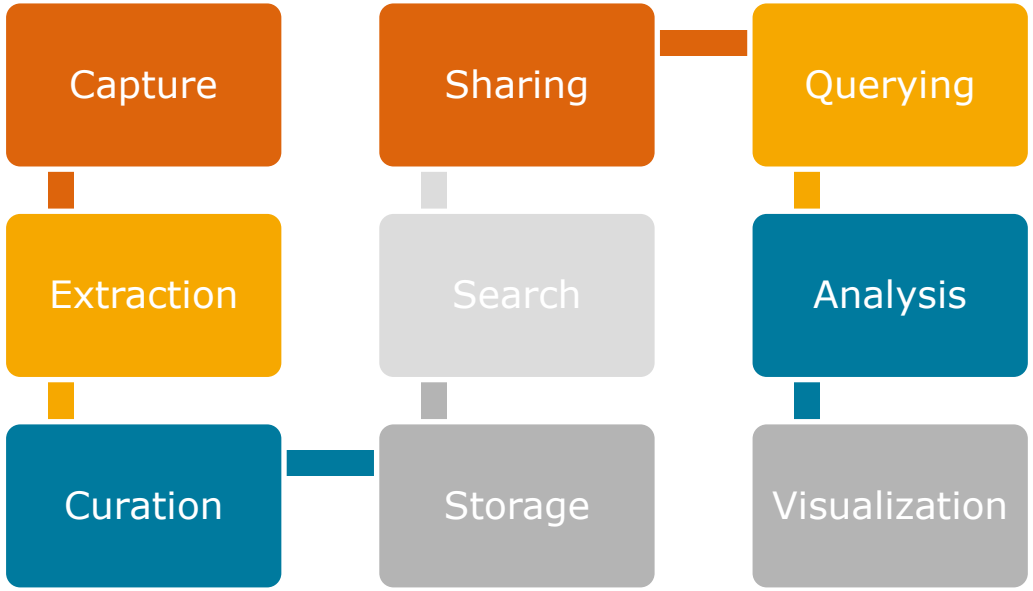
In summary: **Data preparation**

- "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…"

[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.]

Defining Big Data

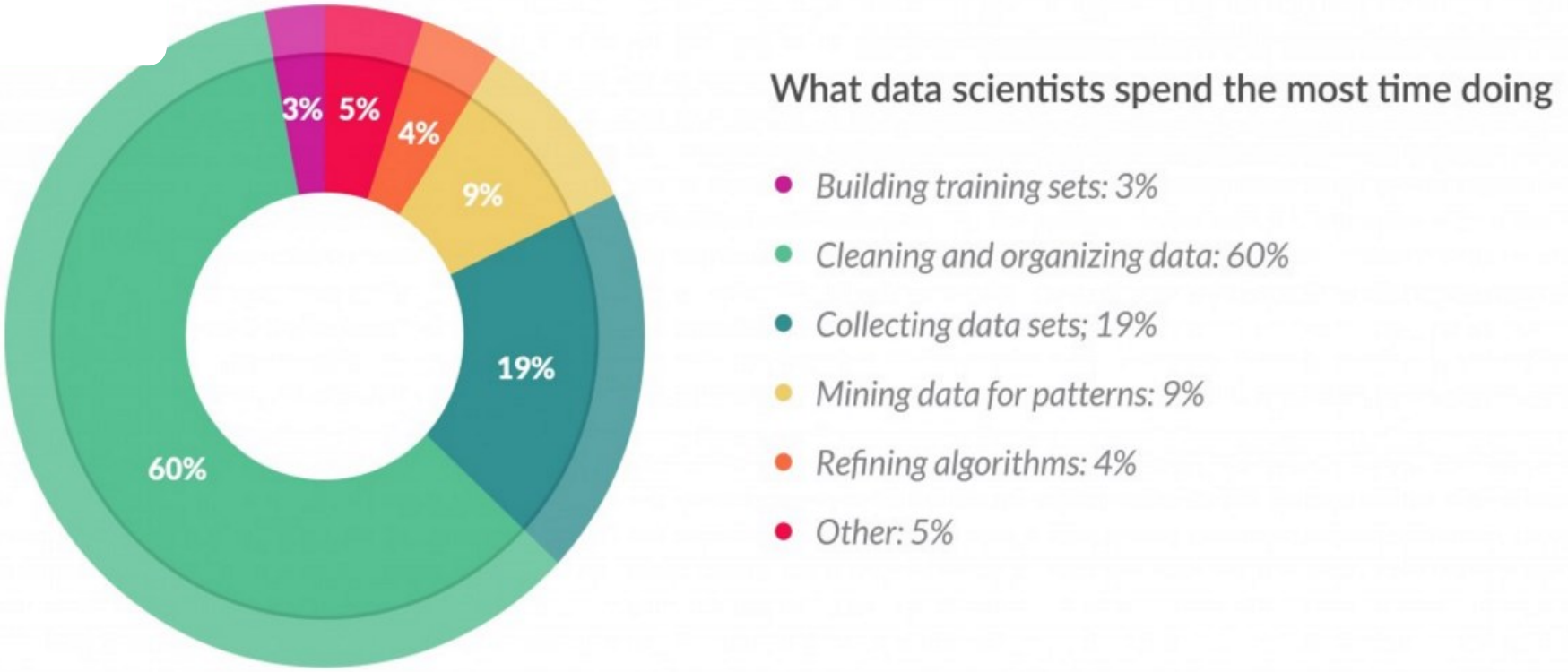
Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.



If data is **too big**, **too fast**, or **too hard** for existing tools to process, it is Big Data.

Data Profiling as Data Preparation

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



"Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task", Gil Press, Forbes, March 23rd, 2016
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

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

Agenda

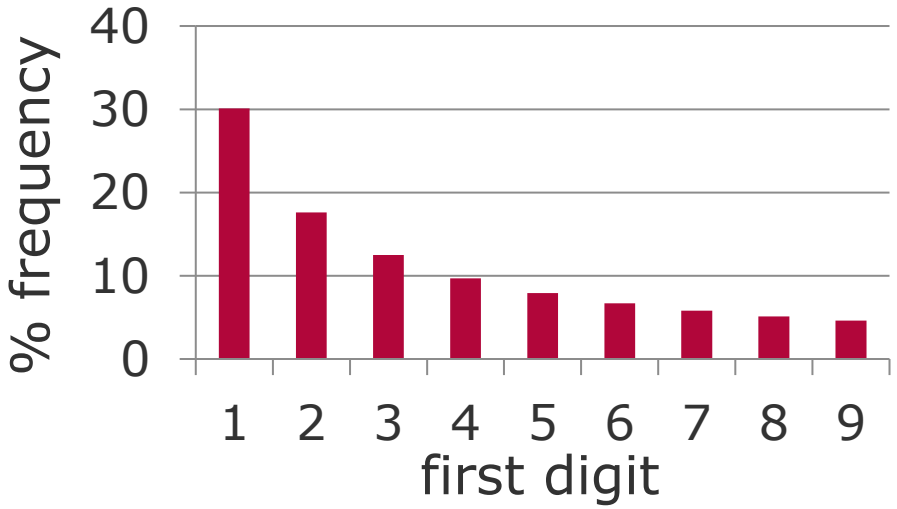
1. **Basic statistics**
2. Uniques and keys
3. Functional dependencies
4. Inclusion dependencies and foreign keys
5. Profiling tools
6. Outlook: Other dependencies and more

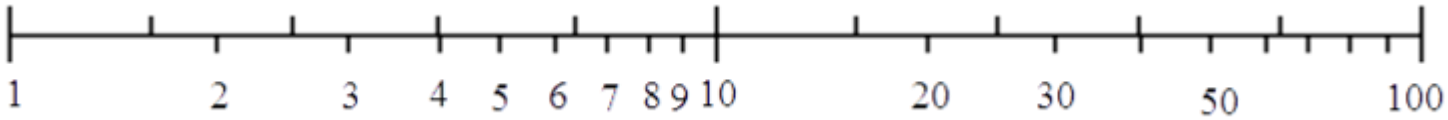
Cardinalities, distributions, and patterns

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.

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





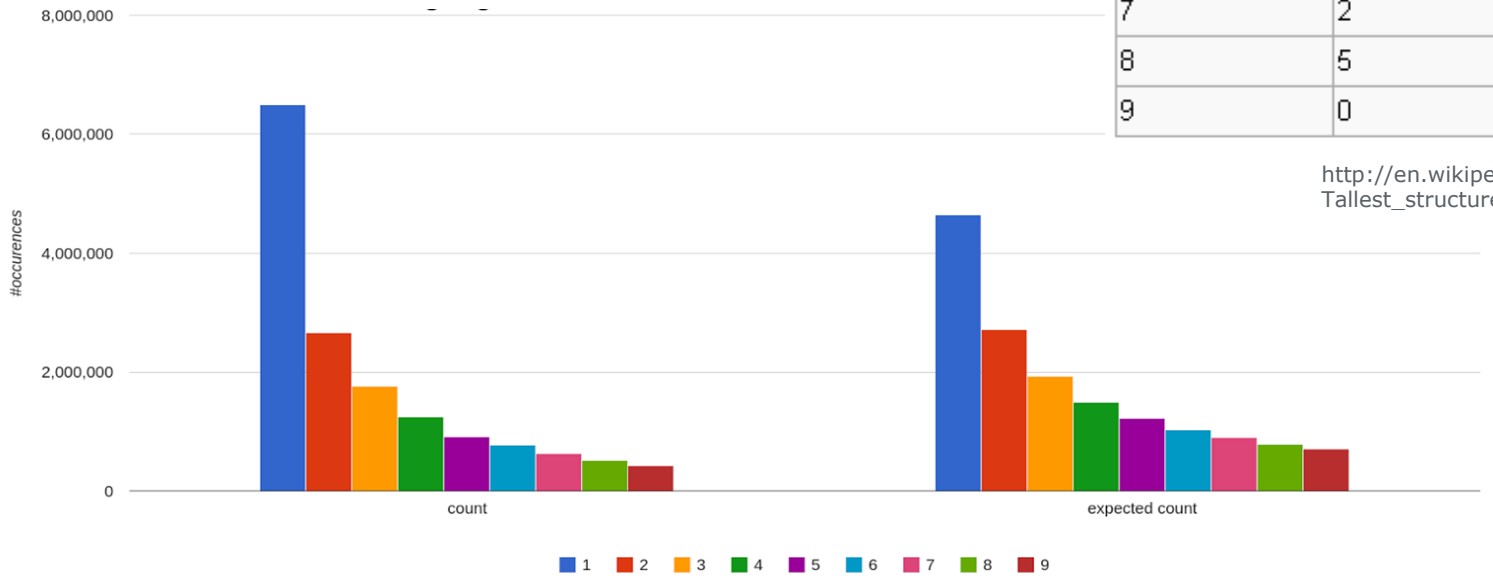
Examples for Benford's Law

- Surface areas of 335 rivers
- Sizes of 3259 US populations
- 1800 molecular weights
- 5000 entries from a mathematical handbook
- Street addresses of the first 342 persons listed in American Men of Science
- 2^n

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



Occurrences of leading digits in WikiTable numbers

CODATA RECOMMENDED VALUES OF THE FUNDAMENTAL PHYSICAL CONSTANTS: 2014

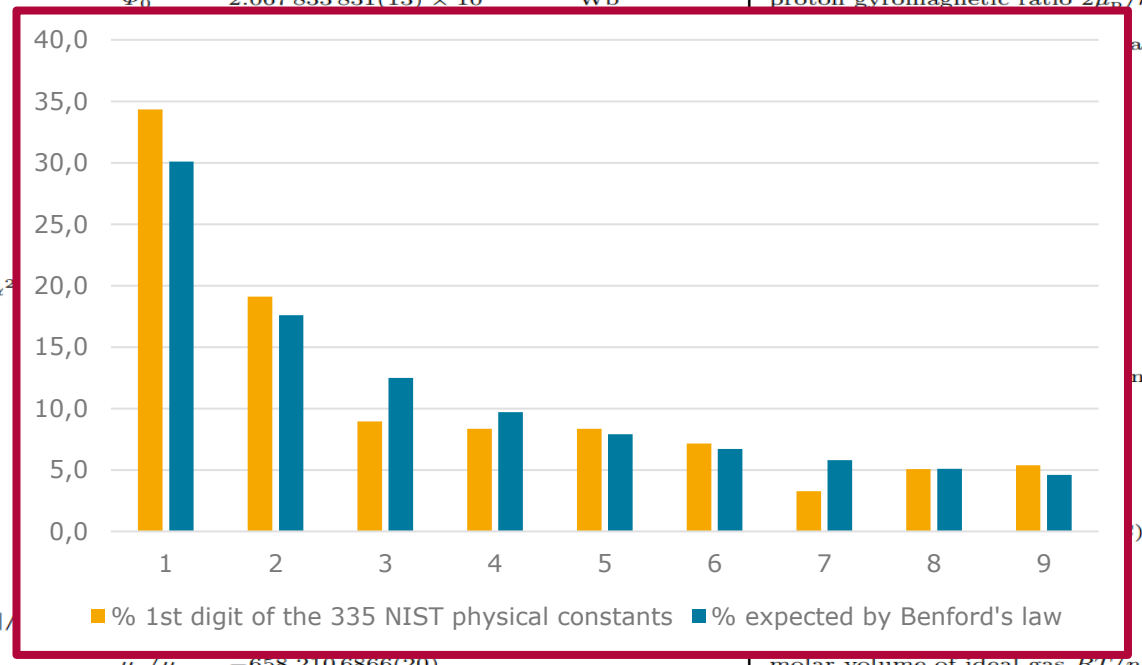
NIST SP 961 (Sept/2015) Values from: P. J. Mohr, D. B. Newell, and B. N. Taylor, arXiv:1507.07956

A more extensive listing of constants is available in the above reference and on the NIST Physics Laboratory Web site physics.nist.gov/constants.

The number in parentheses is the one-standard-deviation uncertainty in the last two digits of the given value.

Quantity	Symbol	Numerical value	Unit
speed of light in vacuum	c, c_0	299 792 458 (exact)	m s^{-1}
magnetic constant	μ_0	$4\pi \times 10^{-7}$ (exact)	N A^{-2}
		$= 12.566\,370\,614\ldots \times 10^{-7}$	N A^{-2}
electric constant $1/\mu_0 c^2$	ϵ_0	$8.854\,187\,817\ldots \times 10^{-12}$	F m^{-1}
Newtonian constant of gravitation	G	$6.674\,08(31) \times 10^{-11}$	m^3 kg^{-1} s^{-2}
Planck constant	h	$6.626\,070\,040(81) \times 10^{-34}$	J s
in eV s		$4.135\,667\,662(25) \times 10^{-15}$	eV s
$h/2\pi$	$\hbar$	$1.054\,571\,800(13) \times 10^{-34}$	J s
in eV s		$6.582\,119\,514(40) \times 10^{-16}$	eV s
elementary charge	e	$1.602\,176\,6208(98) \times 10^{-19}$	C
magnetic flux quantum $h/2e$	Φ_0	$2.067\,833\,831(13) \times 10^{-15}$	Wb
Josephson constant $2e/h$			
von Klitzing constant $h/e^2 = \mu_0 c/2\alpha$			
Bohr magneton $e\hbar/2m_e$			
in eV T^{-1}			
nuclear magneton $e\hbar/2m_p$			
in eV T^{-1}			
fine-structure constant $e^2/4\pi\epsilon_0\hbar c$			
inverse fine-structure constant			
Rydberg constant $\alpha^2 m_e c/2h$			
energy equivalent in eV			
Bohr radius $\alpha/4\pi R_\infty = 4\pi\epsilon_0\hbar^2/m_e e^2$			
Hartree energy $e^2/4\pi\epsilon_0 a_0 = 2R_\infty hc = \alpha^2$			
in eV			
electron mass			
in u			
energy equivalent in MeV			
electron-muon mass ratio			
electron-proton mass ratio			
electron charge to mass quotient			
Compton wavelength $h/m_e c$			
$\lambda_C/2\pi = \alpha a_0 = \alpha^2/4\pi R_\infty$			
classical electron radius $\alpha^2 a_0$			
Thomson cross section $(8\pi/3)r_e^2$			
electron magnetic moment			
to Bohr magneton ratio			
to nuclear magneton ratio			
electron magnetic moment anomaly $\lvert\mu_e\rvert/$			
electron g-factor $-2(1 + a_e)$			
electron-proton magnetic moment ratio	μ_e/μ_p	$-658.210\,6866(20)$	
muon mass in u	m_μ	0.113 428 9257(25)	u
energy equivalent in MeV	$m_\mu c^2$	105.658 3745(24)	MeV
muon-electron mass ratio	m_μ/m_e	206.768 2826(46)	
muon magnetic moment	μ_μ	$-4.490\,448\,26(10) \times 10^{-26}$	J T^{-1}
to Bohr magneton ratio	μ_μ/μ_B	$-4.841\,970\,48(11) \times 10^{-3}$	
to nuclear magneton ratio	μ_μ/μ_N	$-8.890\,597\,05(20)$	
muon magnetic moment anomaly $\lvert\mu_\mu\rvert/(e\hbar/2m_\mu) - 1$	a_μ	$1.165\,920\,89(63) \times 10^{-3}$	

Quantity	Symbol	Numerical value	Unit
muon g-factor $-2(1 + a_\mu)$	g_μ	$-2.002\,331\,8418(13)$	
muon-proton magnetic moment ratio	μ_μ/μ_p	$-3.183\,345\,142(71)$	
proton mass	m_p	$1.672\,621\,898(21) \times 10^{-27}$	kg
in u		$1.007\,276\,466\,879(91)$	u
energy equivalent in MeV	$m_p c^2$	938.272 0813(58)	MeV
proton-electron mass ratio	m_p/m_e	1836.152 673 89(17)	
proton magnetic moment	μ_p	$1.410\,606\,7873(97) \times 10^{-26}$	J T^{-1}
to nuclear magneton ratio	μ_p/μ_N	2.792 847 3508(85)	
proton magnetic shielding correction $1 - \mu'_p/\mu_p$ (H_2O, sphere, 25 °C)	σ'_p	$25.691(11) \times 10^{-6}$	
proton gyromagnetic ratio $2\mu_p/\hbar$	γ_p	$2.675\,221\,900(18) \times 10^8$	s^{-1} T^{-1}
	$\gamma_p/2\pi$	42.577 478 92(29)	MHz T^{-1}
atio $2\mu'_p/\hbar$	γ'_p	$2.675\,153\,171(33) \times 10^8$	s^{-1} T^{-1}
	$\gamma'_p/2\pi$	42.576 385 07(53)	MHz T^{-1}
	m_n	1.008 664 915 88(49)	u
	$m_n c^2$	939.565 4133(58)	MeV
	m_n/m_p	1.001 378 418 98(51)	
	μ_n	$-0.966\,236\,50(23) \times 10^{-26}$	J T^{-1}
	μ_n/μ_N	$-1.913\,042\,73(45)$	
	m_d	2.013 553 212 745(40)	u
	$m_d c^2$	1875.612 928(12)	MeV
	m_d/m_p	1.999 007 500 87(19)	
	μ_d	$0.433\,073\,5040(36) \times 10^{-26}$	J T^{-1}
	μ_d/μ_N	0.857 438 2311(48)	
	m_h	3.014 932 246 73(12)	u
	$m_h c^2$	2808.391 586(17)	MeV
nt	μ'_h	$-1.074\,553\,080(14) \times 10^{-26}$	J T^{-1}
	μ'_h/μ_B	$-1.158\,671\,471(14) \times 10^{-3}$	
	μ'_h/μ_N	$-2.127\,497\,720(25)$	
	m_α	4.001 506 179 127(63)	u
	$m_\alpha c^2$	3727.379 378(23)	MeV
	N_A, L	$6.022\,140\,857(74) \times 10^{23}$	mol^{-1}
) = 1 u	m_u	$1.660\,539\,040(20) \times 10^{-27}$	kg
	$m_u c^2$	931.494 0954(57)	MeV
	F	96 485.332 89(59)	C mol^{-1}
	R	8.314 4598(48)	J mol^{-1} K^{-1}
	k	$1.380\,648\,52(79) \times 10^{-23}$	J K^{-1}
		$8.617\,3303(50) \times 10^{-5}$	eV K^{-1}
molar volume of ideal gas RT/p (T = 273.15 K, p = 101.325 kPa)	V_m	$22.413\,962(13) \times 10^{-3}$	m^3 mol^{-1}
Stefan-Boltzmann constant $\pi^2 k^4/60\hbar^3 c^2$	σ	$5.670\,367(13) \times 10^{-8}$	W m^{-2} K^{-4}
first radiation constant $2\pi hc^2$	c_1	$3.741\,771\,790(46) \times 10^{-16}$	W m^2
second radiation constant hc/k	c_2	$1.438\,777\,36(83) \times 10^{-2}$	m K
Wien displacement law constant $b = \lambda_{max} T = c_2/4.965\,114\,231\ldots$	b	$2.897\,7729(17) \times 10^{-3}$	m K
Cu x unit: $\lambda(\text{Cu K}\alpha_1)/1\,537.400$	xu(Cu Kα_1)	$1.002\,076\,97(28) \times 10^{-13}$	m
Mo x unit: $\lambda(\text{Mo K}\alpha_1)/707.831$	xu(Mo Kα_1)	$1.002\,099\,52(53) \times 10^{-13}$	m



Energy equivalents

$(1\ m^{-1})c = 299\,792\,458$ Hz	$(1\ \text{Hz})h/k = 4.799\,2447(28) \times 10^{-11}$ K	$(1\ \text{J}) = 6.241\,509\,126(38) \times 10^{18}$ eV	$(1\ \text{eV})/c^2 = 1.073\,544\,1105(66) \times 10^{-9}$ u
$(1\ m^{-1})hc/k = 1.438\,777\,36(83) \times 10^{-2}$ K	$(1\ \text{Hz})h = 4.135\,667\,662(25) \times 10^{-15}$ eV	$(1\ \text{eV}) = 1.602\,176\,6208(98) \times 10^{-19}$ J	$(1\ \text{kg}) = 6.022\,140\,857(74) \times 10^{26}$ u
$(1\ m^{-1})hc = 1.239\,841\,9739(76) \times 10^{-6}$ eV	$(1\ \text{K})k/hc = 69.503\,457(40)\ m^{-1}$	$(1\ \text{eV})/hc = 8.065\,544\,005(50) \times 10^5\ m^{-1}$	$(1\ \text{u}) = 1.660\,539\,040(20) \times 10^{-27}$ kg
$(1\ m^{-1})h/c = 1.331\,025\,049\,00(61) \times 10^{-15}$ u	$(1\ \text{K})k/h = 2.083\,6612(12) \times 10^{10}$ Hz	$(1\ \text{eV})/h = 2.417\,989\,262(15) \times 10^{14}$ Hz	$(1\ \text{u})c/h = 7.513\,006\,6166(34) \times 10^{14}\ m^{-1}$
$(1\ \text{Hz})/c = 3.335\,640\,951\ldots \times 10^{-9}\ m^{-1}$	$(1\ \text{K})k = 8.617\,3303(50) \times 10^{-5}$ eV	$(1\ \text{eV})/k = 1.160\,452\,21(67) \times 10^4$ K	$(1\ \text{u})c^2 = 931.494\,0954(57) \times 10^6$ eV

Agenda

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

Uniqueness, keys, and foreign keys

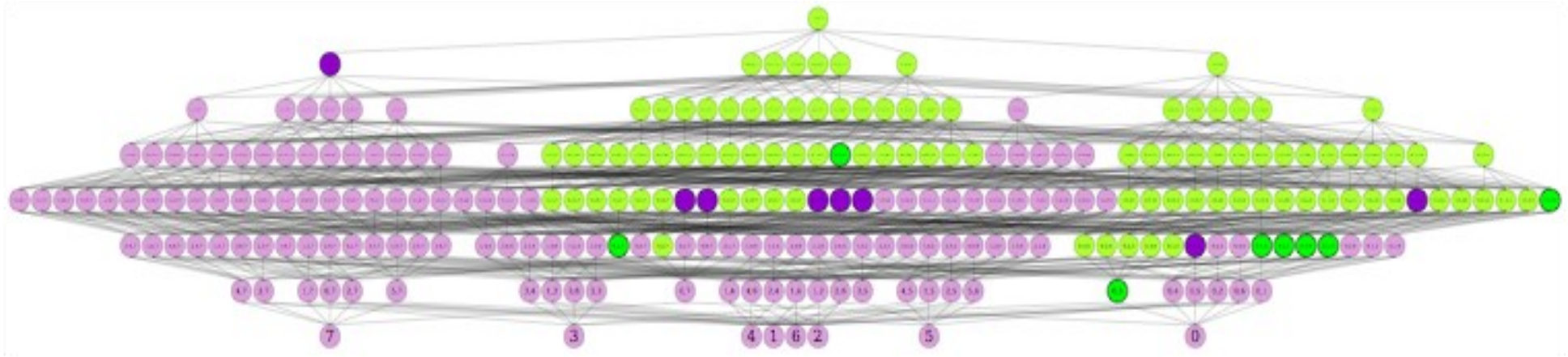
- Uniqueness and keys
  - Unique column: Only unique values
  - Unique column combination: Only unique value combinations
    - Minimality: No column subset is unique
  - Key candidate: No null values
  - Key: Only human expert can decide
    - UCC is prerequisite

- 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

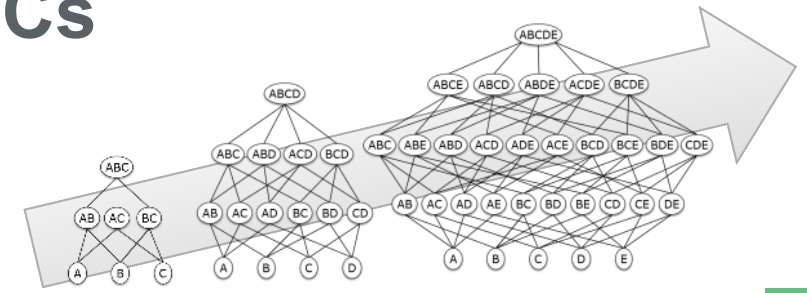
- Useful for
  - Schema design, data integration, indexing, optimization
  - Inverse: non-uniques are duplicates

Large solution space



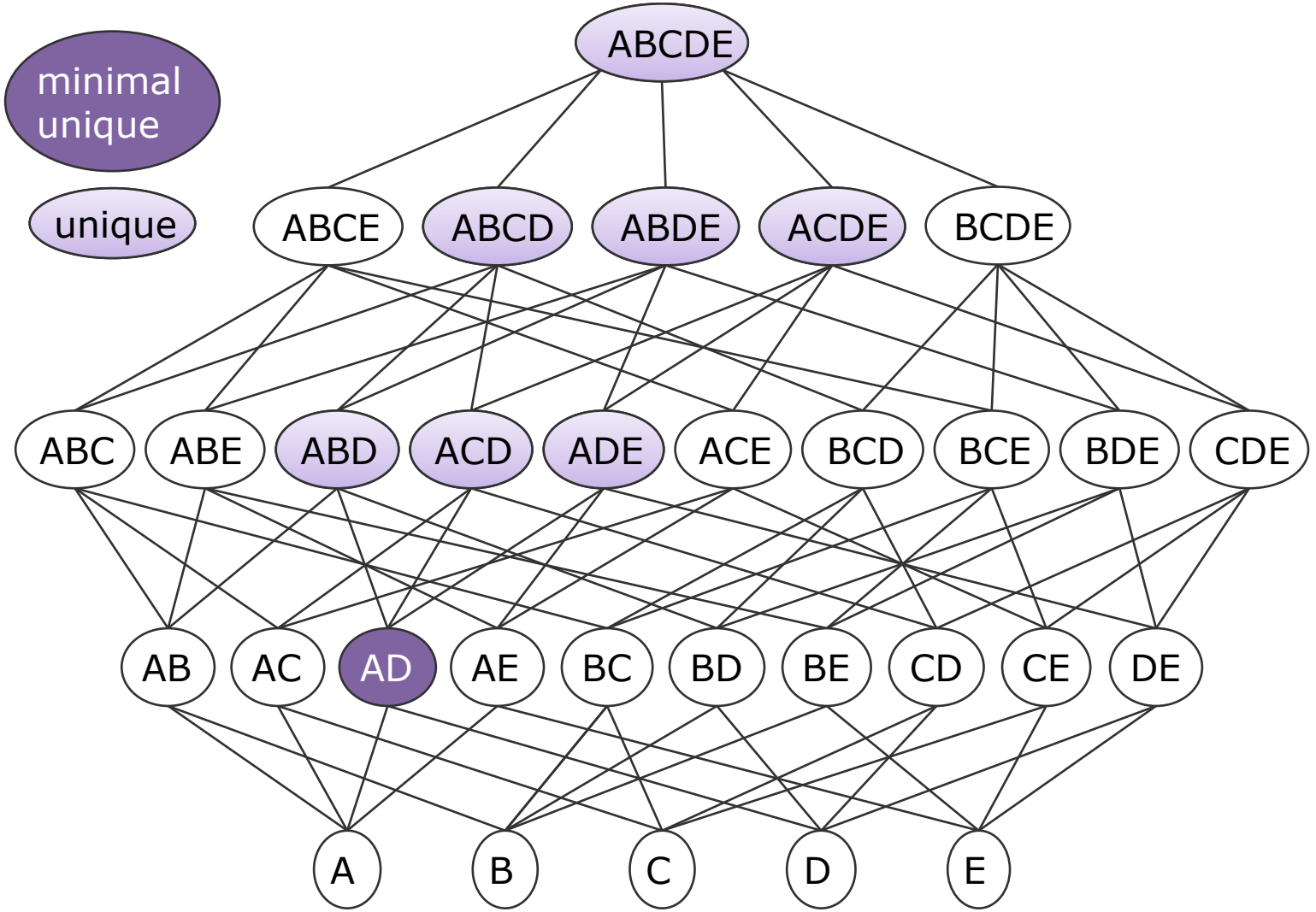
- Size of lattice: 2^n-1 (empty set not considered)
- Nodes at level 1: n
- Nodes at level n: 1
- Nodes at level k: $\binom{n}{k} = \frac{n!}{(n-k)!k!}$
- Largest level at n/2: $\binom{n}{n/2} = \frac{n!}{\left(\frac{n}{2}!\right)^2}$

Candidate Set Growth for UCCs

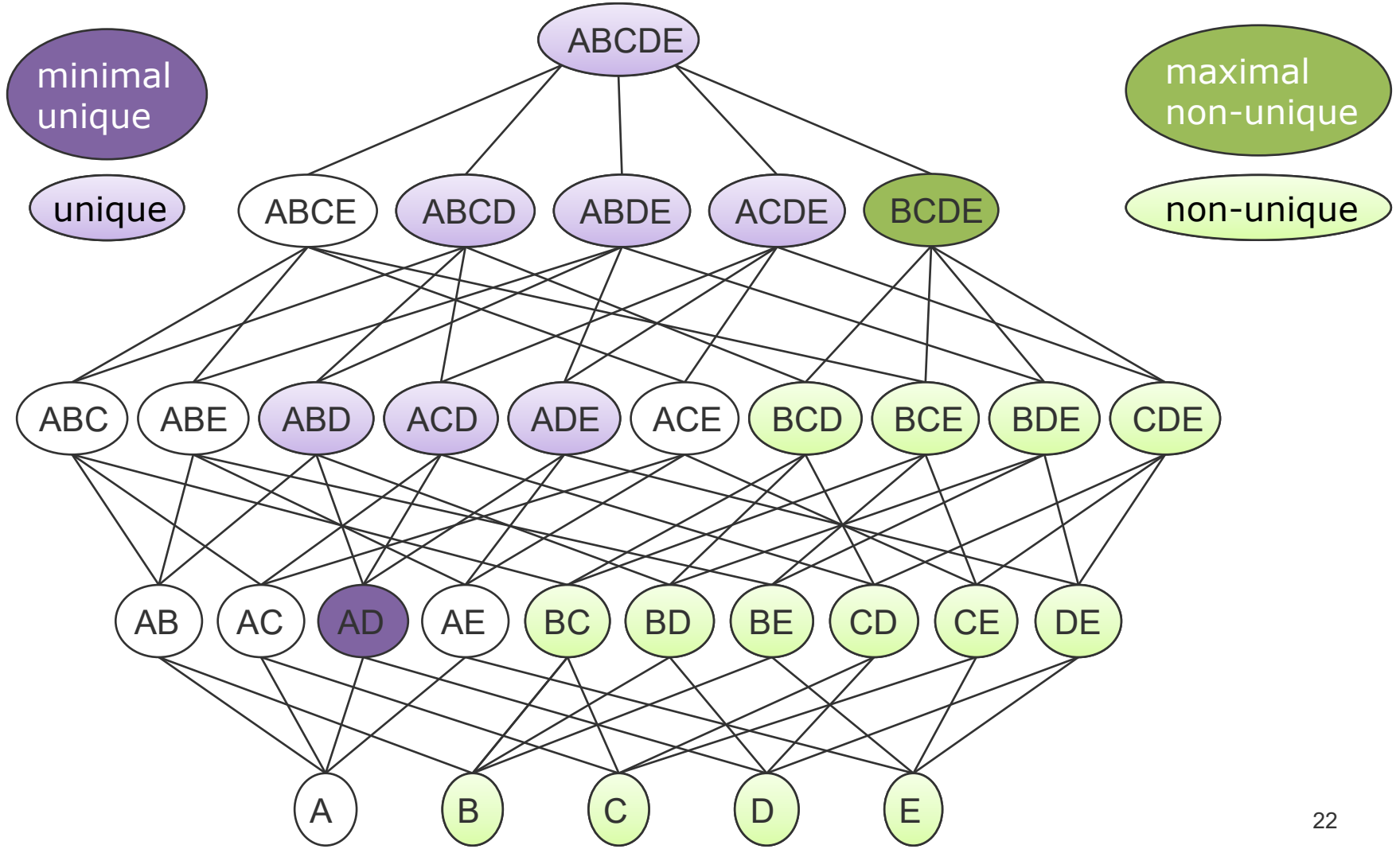


Number of levels: k \ Number of attributes: m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Total :	1	3	7	15	31	63	127	255	511	1,023	2,047	4,095	8,191	16,383	32,767
15															1
14														1	15
13													1	14	105
12												1	13	91	455
11											1	12	78	364	1,365
10										1	11	66	286	1,001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

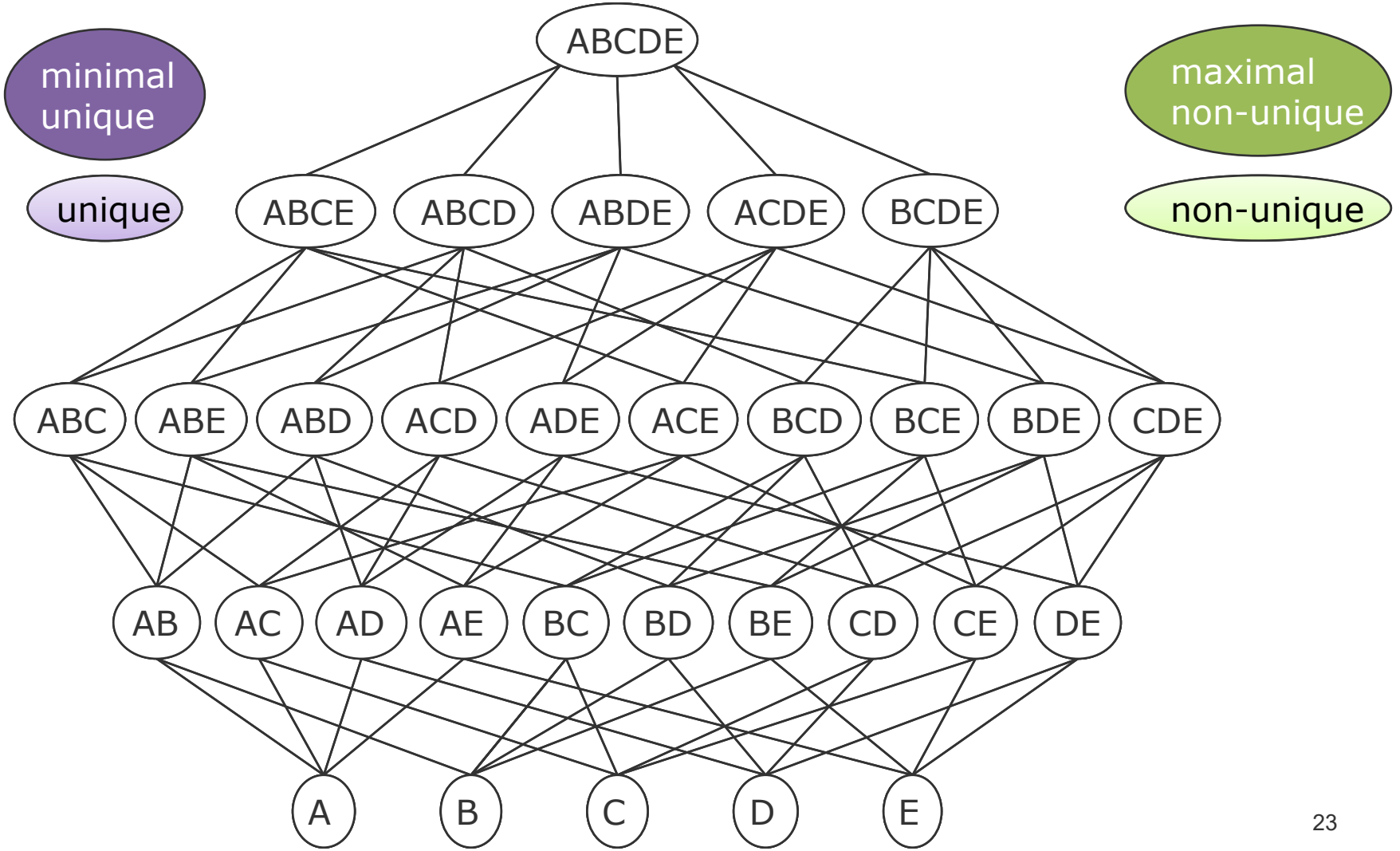
Pruning effect of a pair



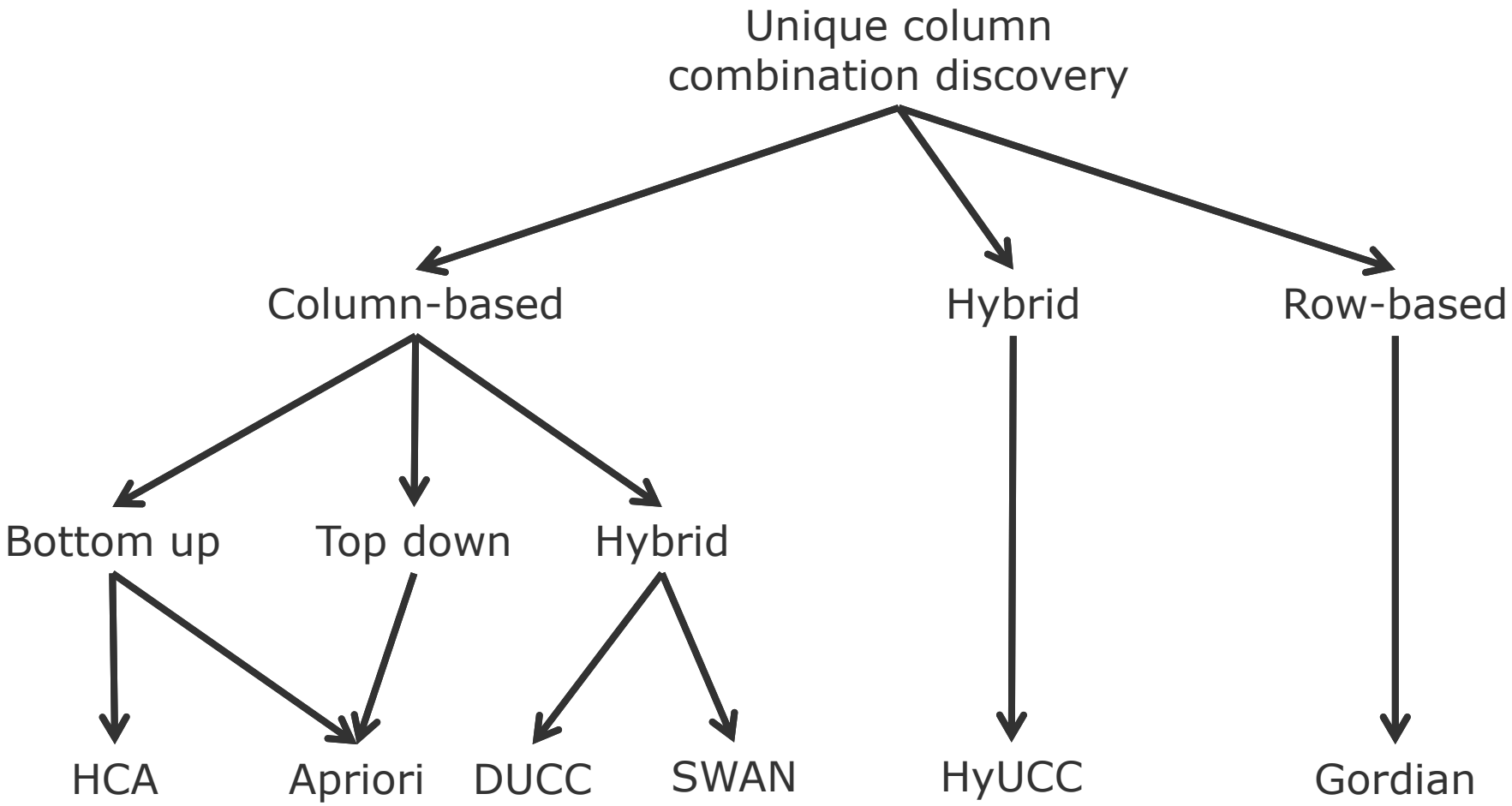
Pruning both ways



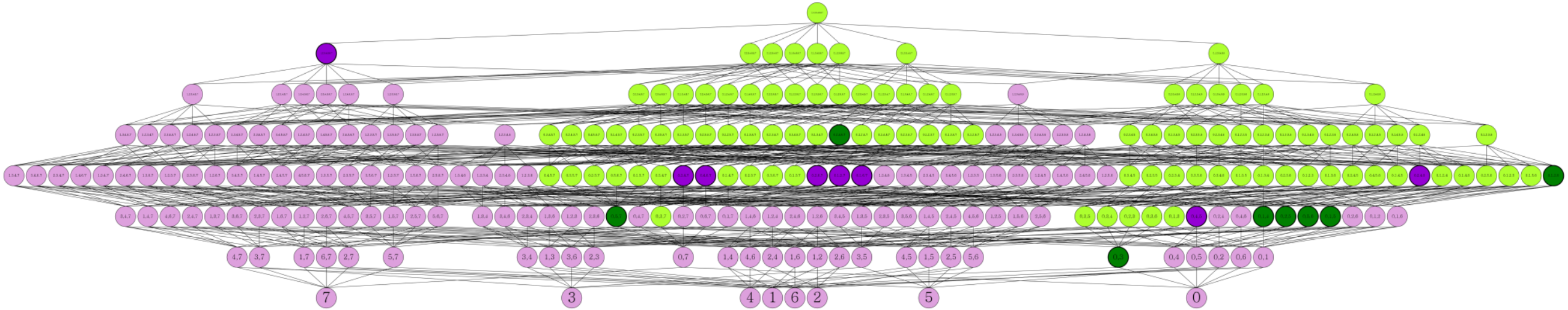
Apriori visualized



Discovery Algorithms

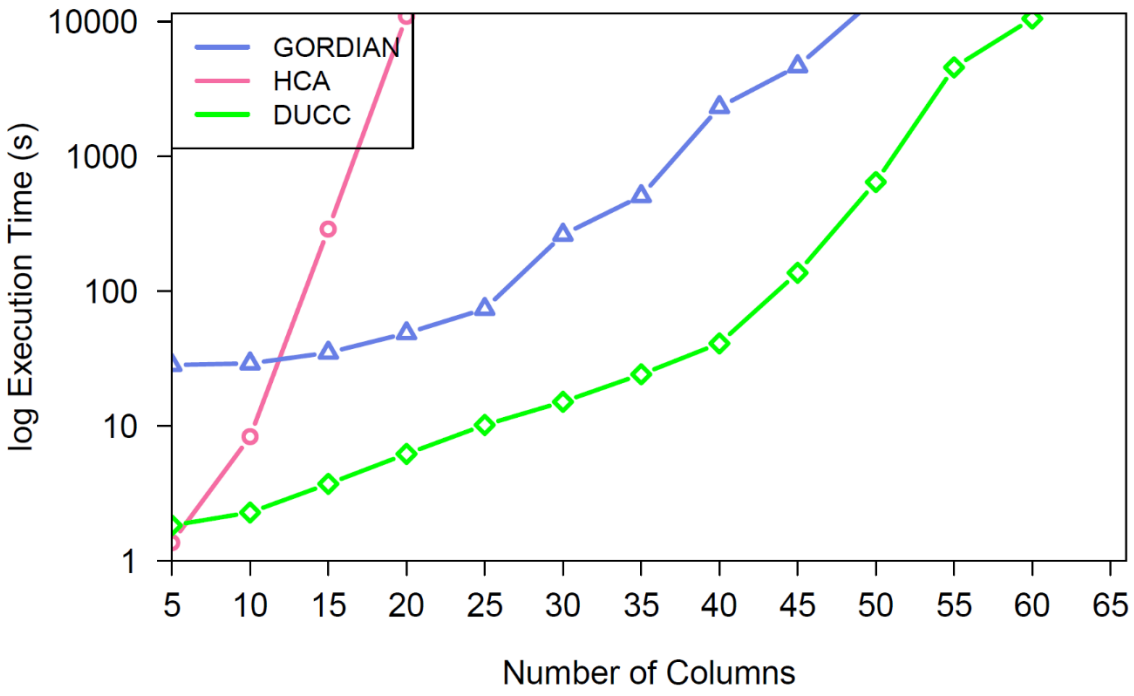


DUCC – Detecting Unique Column Combinations

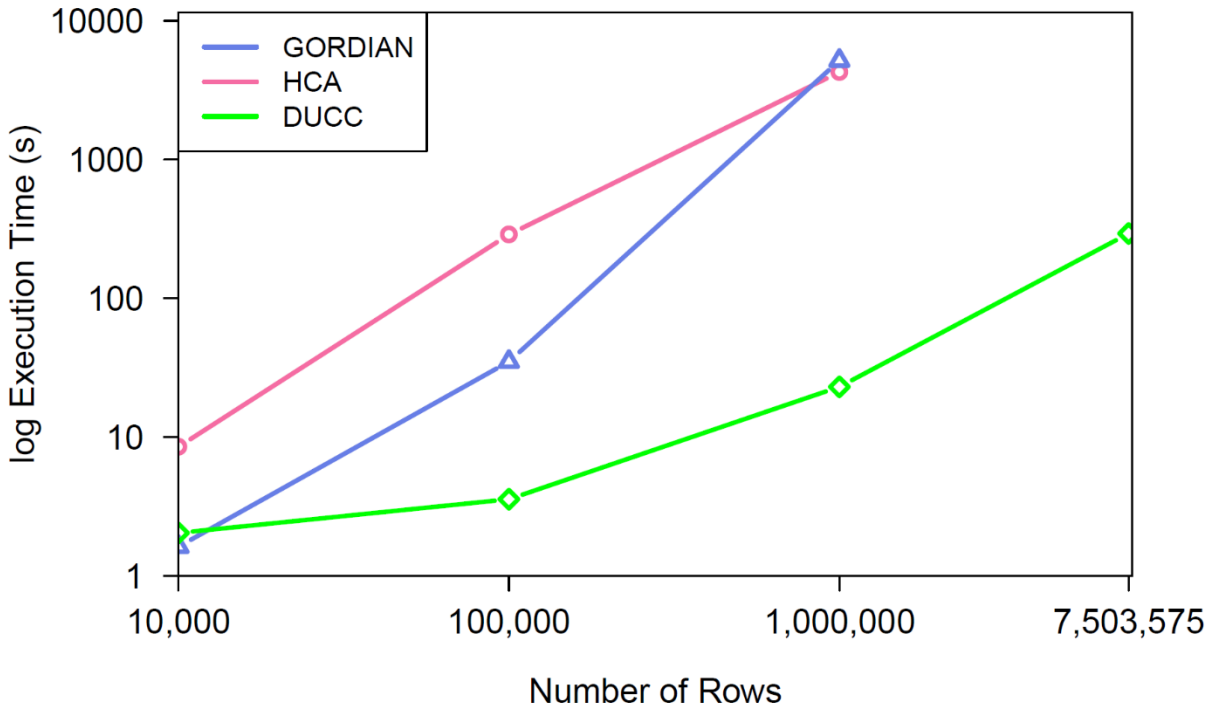


Scalability in the number of columns and rows

- NCVoter data, 100k rows



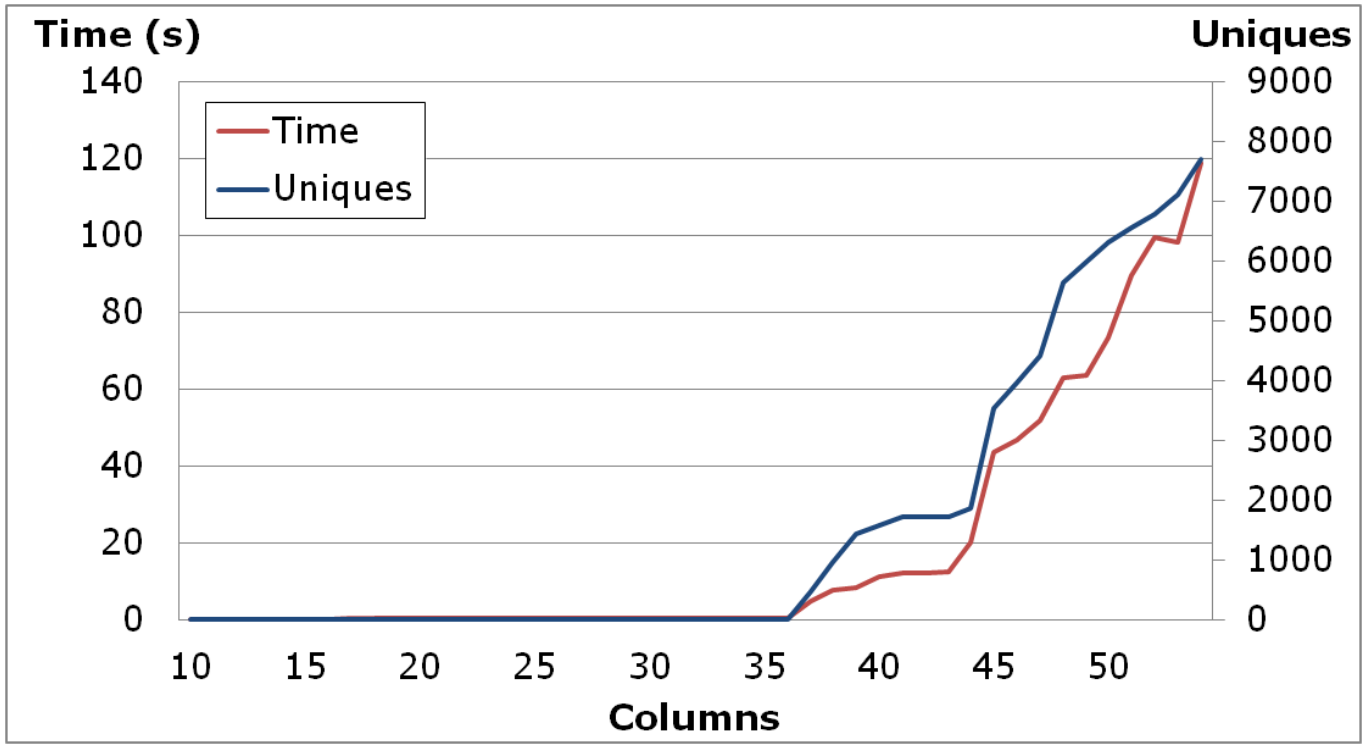
- NCVoter, 15 columns



- New hybrid version shaves off another order of magnitude

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

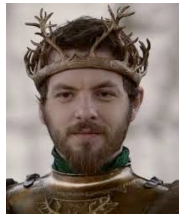
- **A minimal unique**: voter_reg_num, zip_code, race_code
- **A maximal non-unique**: voter_reg_num, status_cd, voter_status_desc, reason_cd, voter_status_reason_desc, absent_ind, name_prefx_cd, name_sufx_cd, half_code, street_dir, street_type_cd, street_sufx_cd, unit_designator, unit_num, state_cd, 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. Profiling tools
6. Outlook: Other dependencies and more

Functional Dependencies

Functional Dependencies

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

Some Functional Dependencies:

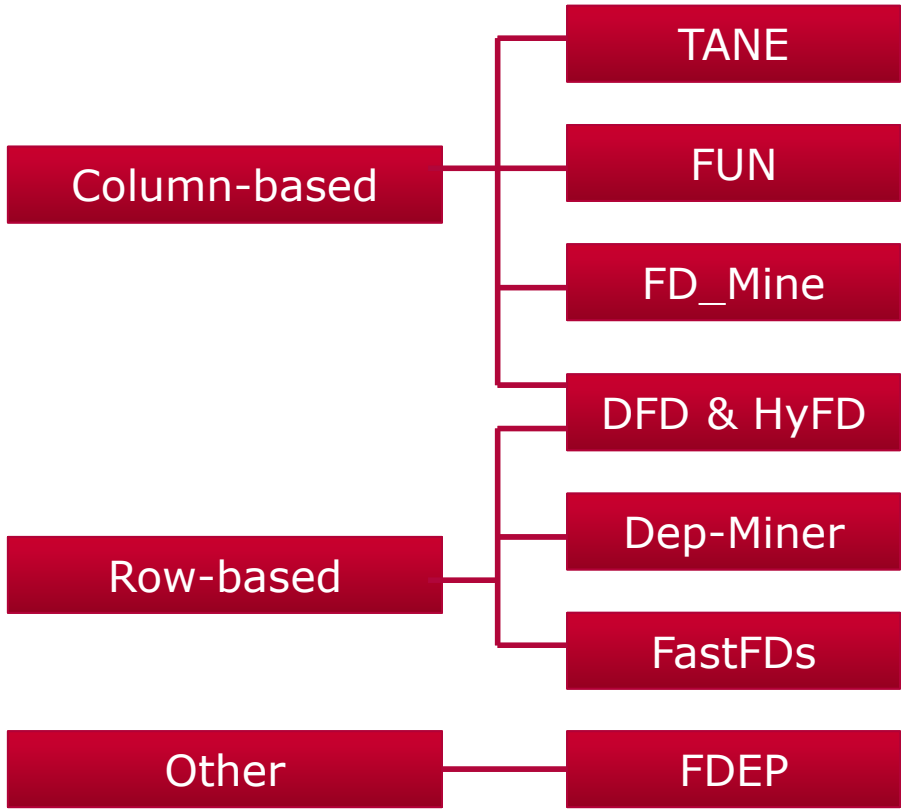
1. Person → Lineage
2. Person → Hair
3. Person → Religion
4. Lineage → Hair
5. Religion, Hair → Lineage
6. …

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

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

Uses and Algorithms for FDs

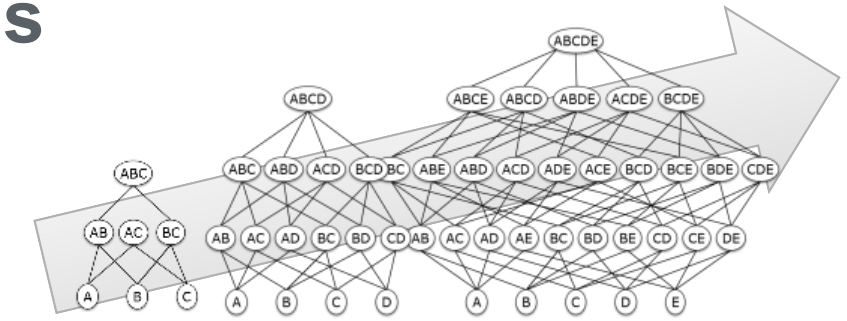
- Schema design
  - Normalization
  - Keys
- Data cleansing
- Query optimization
- Schema design and normalization
- Key discovery



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
- Exponential in number of attributes times number of rows squared

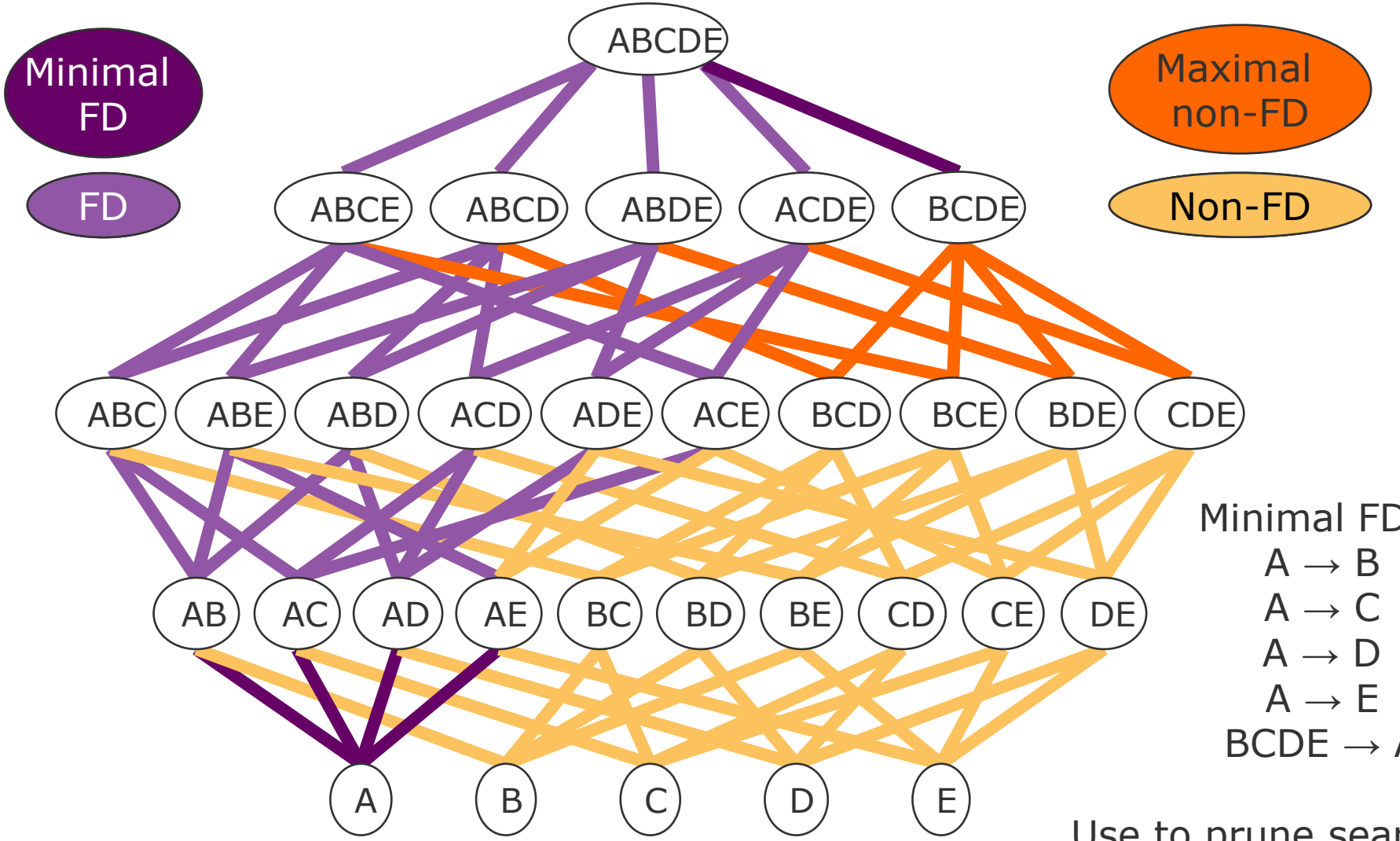
Candidate Set Growth for FDs



Number of levels: k	Total:	0	2	9	28	75	186	441	1,016	2,295	5,110	11,253	24,564	53,235	114,674	245,745
15																0
14															0	15
13														0	14	210
12													0	13	182	1,365
11												0	12	156	1,092	5,460
10											0	11	132	858	4,004	15,015
9										0	10	110	660	2,860	10,010	30,030
8									0	9	90	495	1,980	6,435	18,018	45,045
7								0	8	72	360	1,320	3,960	10,296	24,024	51,480
6							0	7	56	252	840	2,310	5,544	12,012	24,024	45,045
5						0	6	42	168	504	1,260	2,772	5,544	10,296	18,018	30,030
4					0	5	30	105	280	630	1,260	2,310	3,960	6,435	10,010	15,015
3				0	4	20	60	140	280	504	840	1,320	1,980	2,860	4,004	5,460
2			0	3	12	30	60	105	168	252	360	495	660	858	1,092	1,365
1		0	2	6	12	20	30	42	56	72	90	110	132	156	182	210
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Number of attributes: m

Again: Model in lattice – edges represent FDs



Minimal FDs:

$A \rightarrow B$

$A \rightarrow C$

$A \rightarrow D$

$A \rightarrow E$

$BCDE \rightarrow A$

Use to prune search space

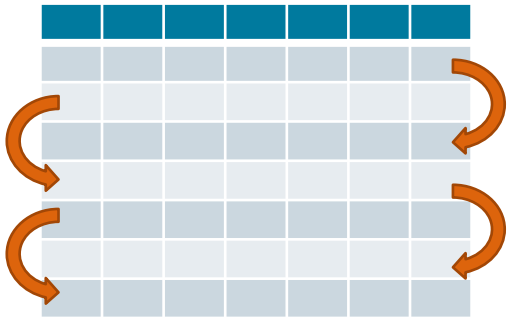
Row-based Discovery

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

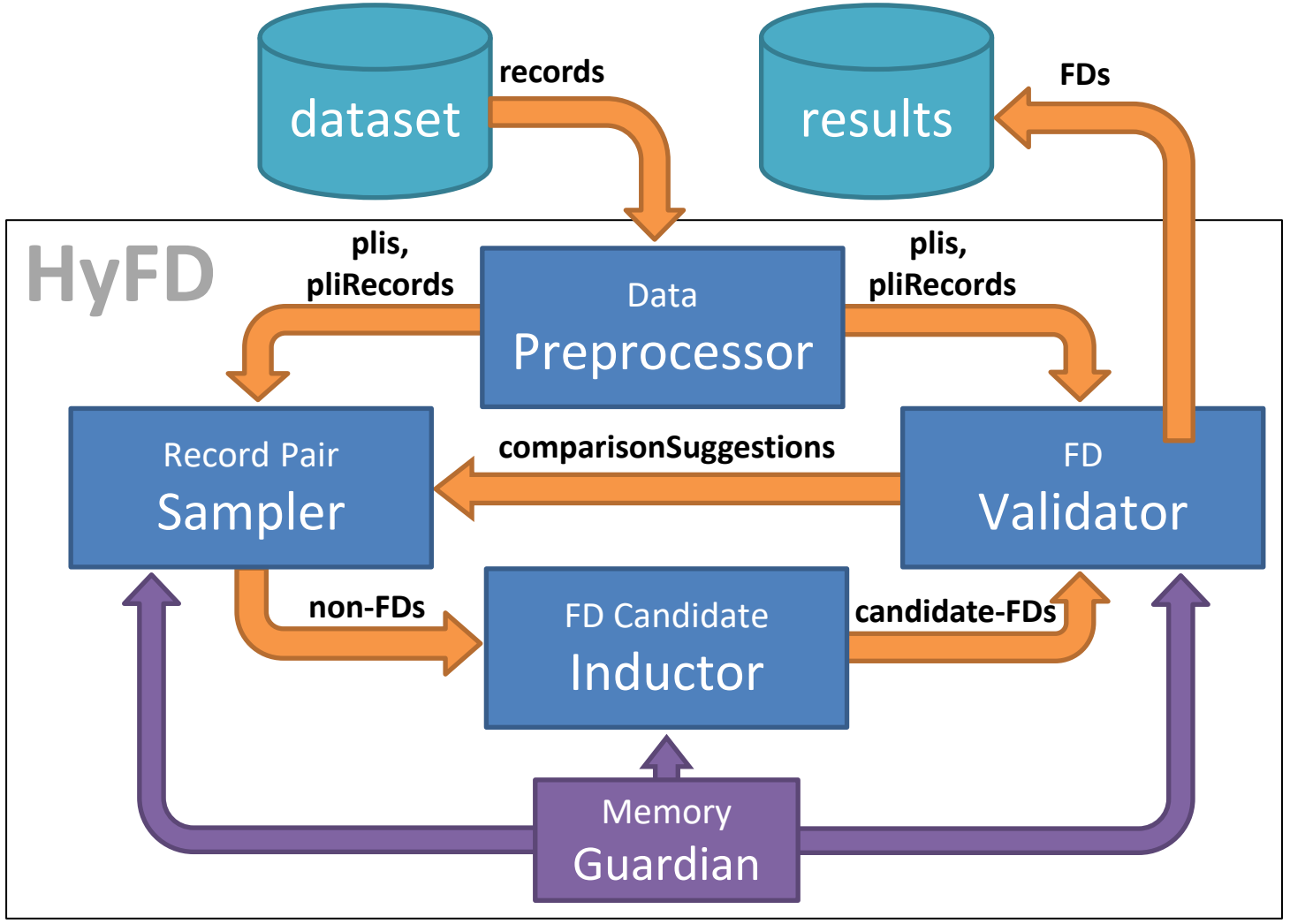
- Surname, Postcode, City, Mayor↛Name
- Name, Postcode, City, Mayor↛Surname
- Surname↛Name, Postcode, City, Mayor

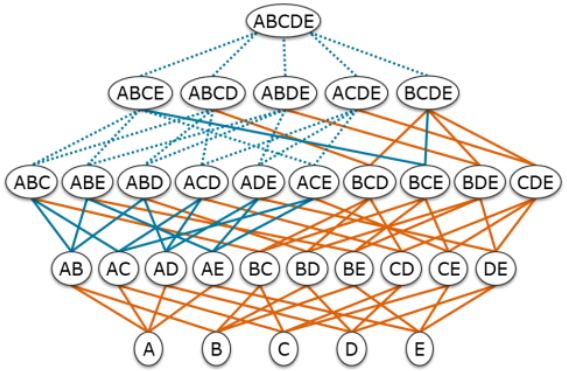
Postcode → City
Postcode → Mayor
Name → Mayor, …

HyFD: Hybrid FD Discovery



column-efficient





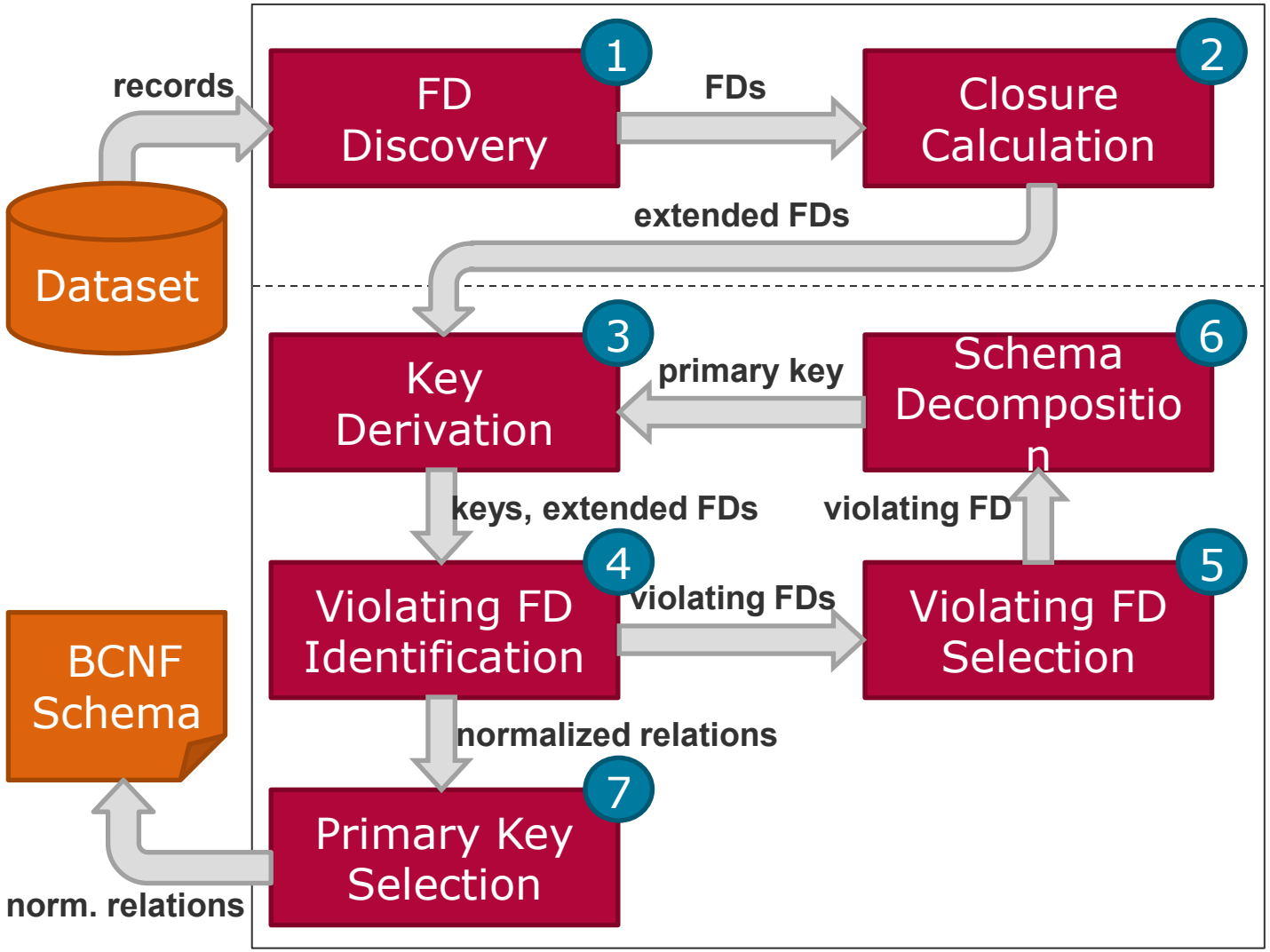
row-efficient

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

Automatic BCNF Normalization



Normalization results: TPC-H

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

Agenda

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

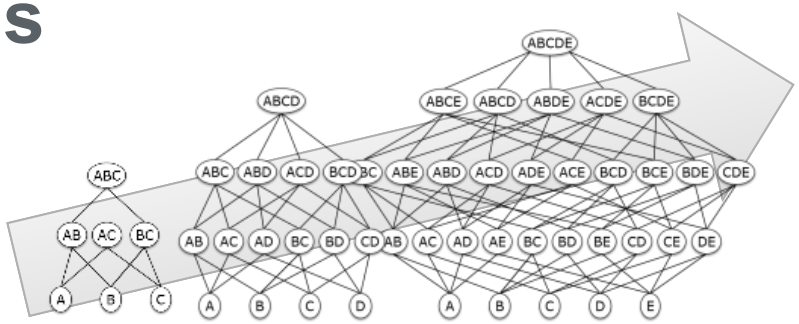
IND discovery $R[X] \subseteq S[Y]$

- 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 with 175 tables
  - Not a single foreign key constraint!
- Example: Ensembl – genome database with >200 tables
  - Not a single foreign key constraint!
- Web tables: No schema, no constraints, but many connections

- Why are FKs missing?
  - Lack of DBMS support for foreign key constraints
  - Fear of performance drop for constraint checking
  - Lack of database knowledge

Candidate Set Growth for INDs



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

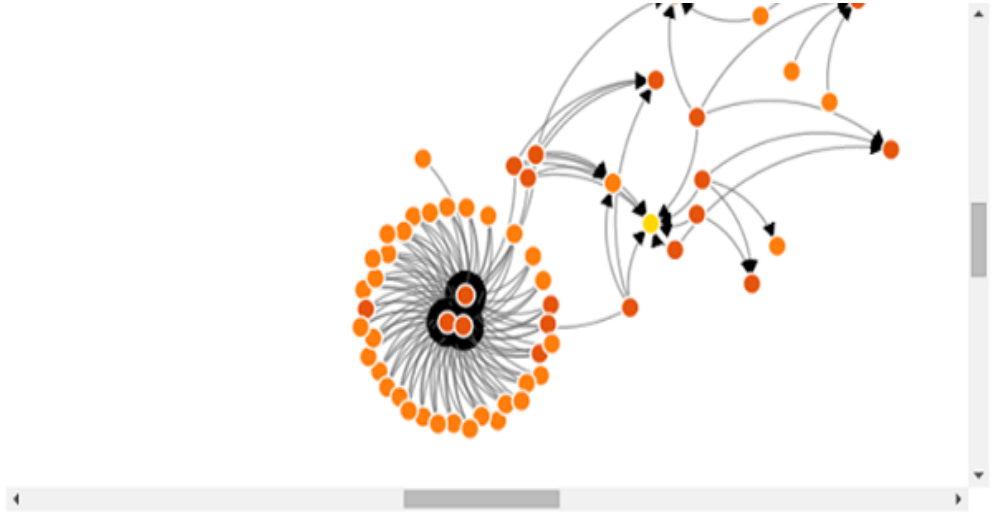
N-ary IND detection:
$O(2^n \cdot n!)$
for n attributes

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

Pruning for INDs

- Subsequences of INDs are INDs
  - $R[A_1, \ldots, A_n] \subseteq S[B_1, \ldots, B_n] \Rightarrow R[A_{i1}, \ldots, A_{im}] \subseteq S[B_{i1}, \ldots, B_{im}]$
- Pruning down
  - Example: $R[AB] \subseteq S[DE] \Rightarrow R[A] \subseteq S[D]$ and $R[B] \subseteq S[E]$
- Pruning up
  - Example: $R[AB] \not\subseteq S[DE] \Rightarrow R[ABC] \not\subseteq S[DEF]$
  - Apriori: Use only satisfied INDs to generate next level candidates
- Pruning laterally
  - $R[AB] \subseteq S[DE] \Rightarrow R[BA] \subseteq S[ED]$
  - Define permutation strategy
    - E.g., lexicographic ordering of attribute labels for LHS

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. **Profiling tools**
6. Outlook: Other dependencies and more

Data profiling tools and algorithms

- **IBM InfoSphere Information Analyzer**
  - http://www.ibm.com/software/data/infosphere/information-analyzer/
- Oracle Enterprise Data Quality
  - http://www.oracle.com/us/products/middleware/data-integration/enterprise-data-quality/overview/index.html
- Talend Data Quality
  - http://www.talend.com/products/data-quality
- Ataccama DQ Analyzer
  - http://www.ataccama.com/en/products/dq-analyzer.html
- **SAP BusinessObjects Data Insight and SAP BusinessObjects Information Steward**
  - http://www.sap.com/germany/solutions/sapbusinessobjects/large/eim/datainsight/index.epx
  - http://www.sap.com/germany/solutions/sapbusinessobjects/large/eim/information-steward/index.epx
- Informatica Data Explorer
  - http://www.informatica.com/us/products/data-quality/data-explorer/
- Microsoft SQL Server Integration Services Data Profiling Task and Viewer
  - http://msdn.microsoft.com/en-us/library/bb895310.aspx
- Trillium Software Data Profiling
  - http://www.trilliumsoftware.com/home/products/data-profiling.aspx
- CloverETL Data Profiler
  - http://www.cloveretl.com/products/profiler
- OpenRefine
  - http://www.openrefine.org
- and many more…

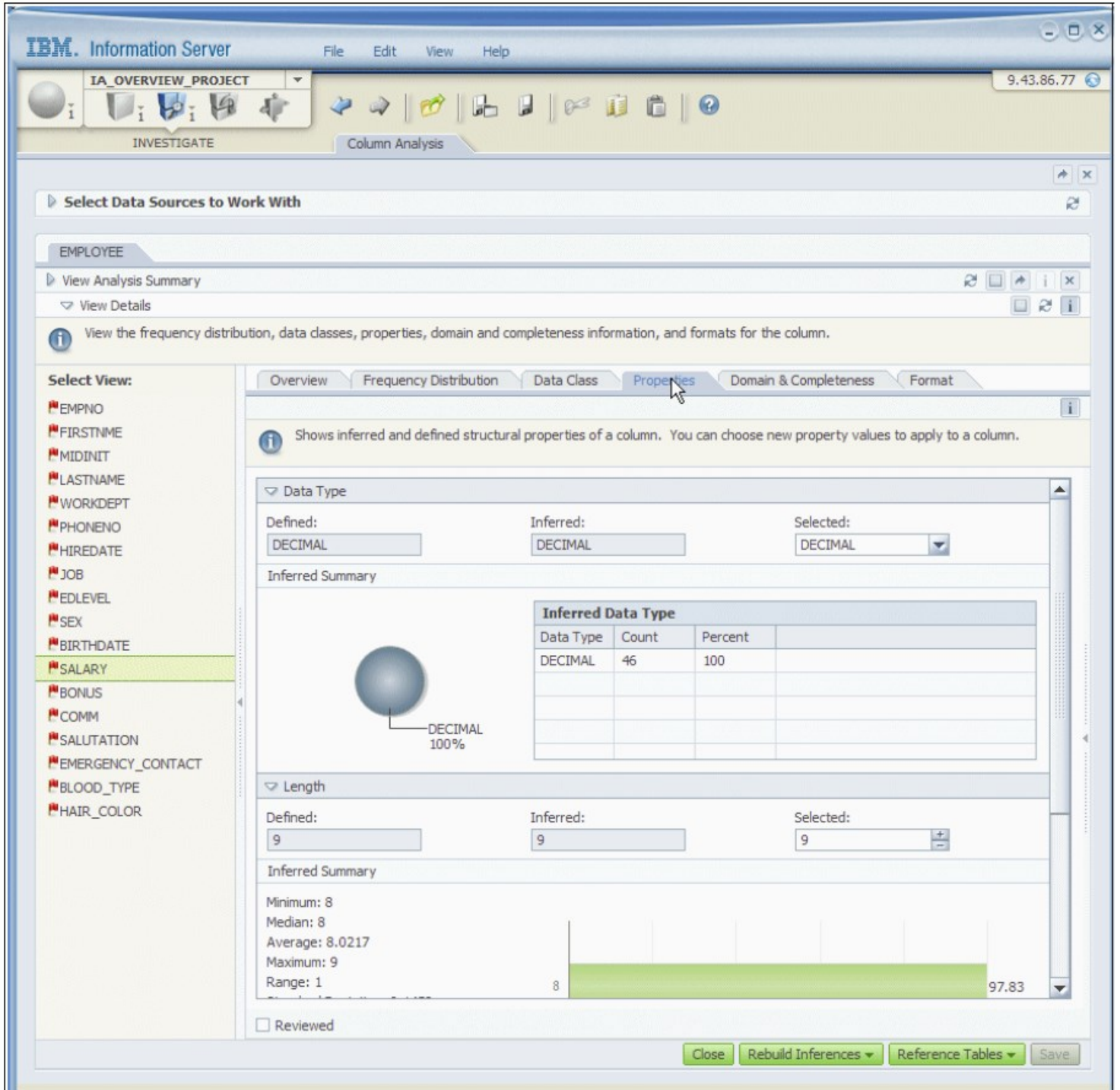
Often packaged with data quality / data cleansing software

Very long feature lists

- Num rows
- Min value length
- Median value length
- Max value length
- Avg value length
- Precision of numeric values
- Scale of numeric values
- Quartiles
- Basic data types
- Num distinct values ("cardinality")
- Percentage null values
- Data class and data type
- Uniqueness and constancy
- Single-column frequency histogram
- Multi-column frequency histogram
- Pattern discovery (Aa9)
- Soundex frequencies
- Benford Law Frequency

- Single column primary key discovery
- Multi-column primary key discovery
- Single column IND discovery
- Inclusion percentage
- Single-column FK discovery
- Multi-column IND discovery
- Multi-column FK discovery
- Value overlap (cross domain analysis)
- Single-column FD discovery
- Multi-column FD discovery
- Text profiling

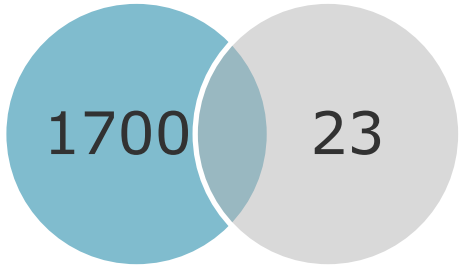
Screenshots for IBM Information Analyzer



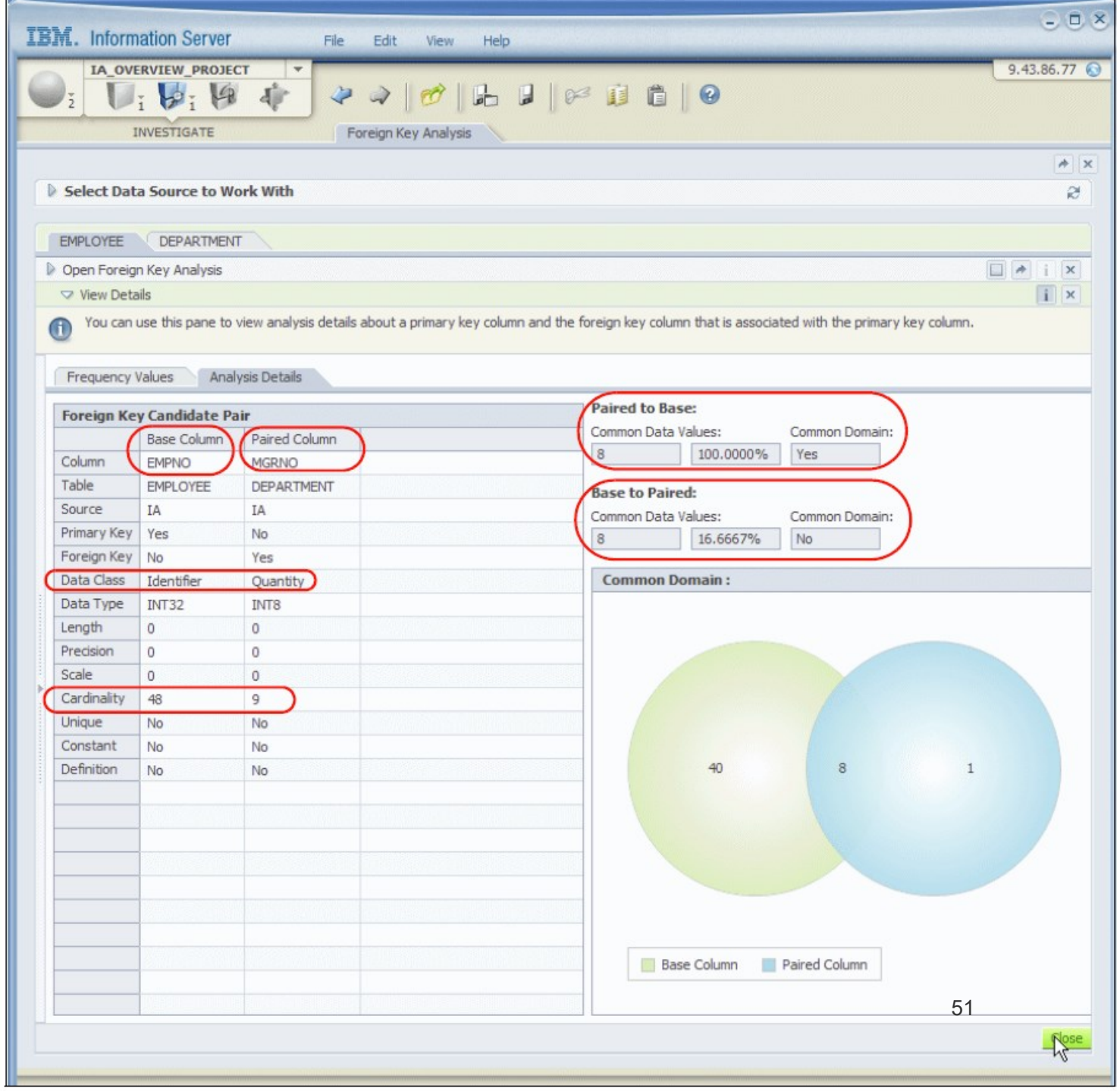
„Checking“ vs. „discovery“

- An anonymous tool:
  - “Cross-table analysis enables you to identify matching or orphaned records between two tables, based on a fully-customizable join condition and optional filter on either table.”

- Corresponds to
  - SELECT COUNT(*)
    FROM A, B
    WHERE A.x = B.y
    AND cond
  - Plus some fancy visualization:



Screenshots for IBM Information Analyzer



Typical Shortcomings of Tools (and research methods)

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

Why are DBMS a poor choice here?

That's the big one

Metanome

Agenda

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

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

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
  - Concise description for which data the partial dependency is valid
- 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

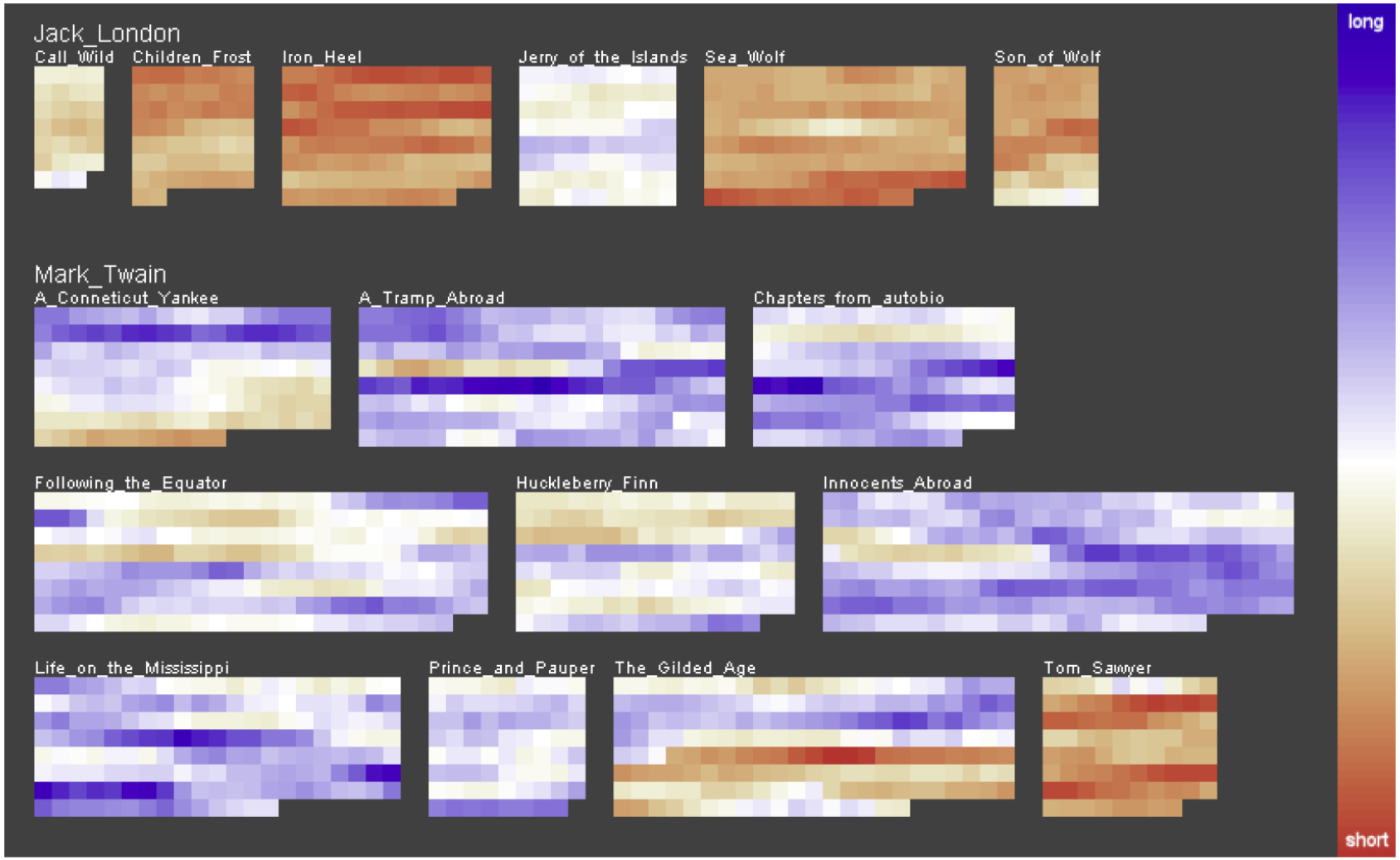
Profiling New Types of Data

- 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
- Different models offer new dimensions to profile
  - XML: Nestedness, measures at different nesting levels
  - **RDF**: Graph structure, in- and outdegrees
  - Multimedia: Color, video-length, volume, etc.
  - **Text**: Sentiment, sentence structure, complexity, and other linguistic measures

Example: Text profiling

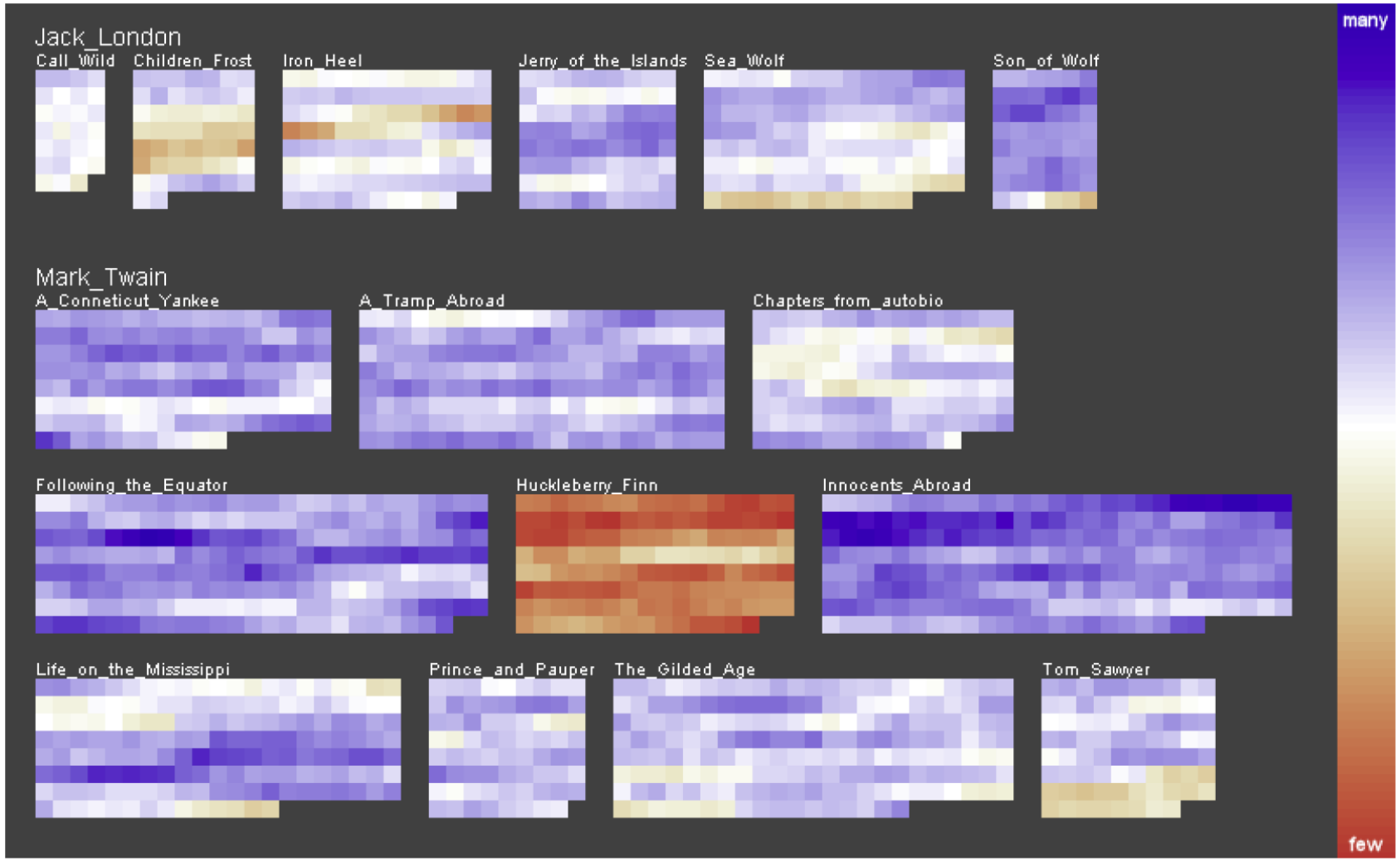
- Statistical measures
  - Syllables per word
  - Sentence length
  - Proportions of parts of speech
- Vocabulary measures
  - Frequencies of specific words
  - Type-token ratio
  - Simpson's index (vocabulary richness)
  - Number of hapax (dis)legomena
    - Token that occurs exactly once (twice) in the corpus
    - Characterize style of an author
- Idea and following figures based on:
  - „*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)

Average sentence length



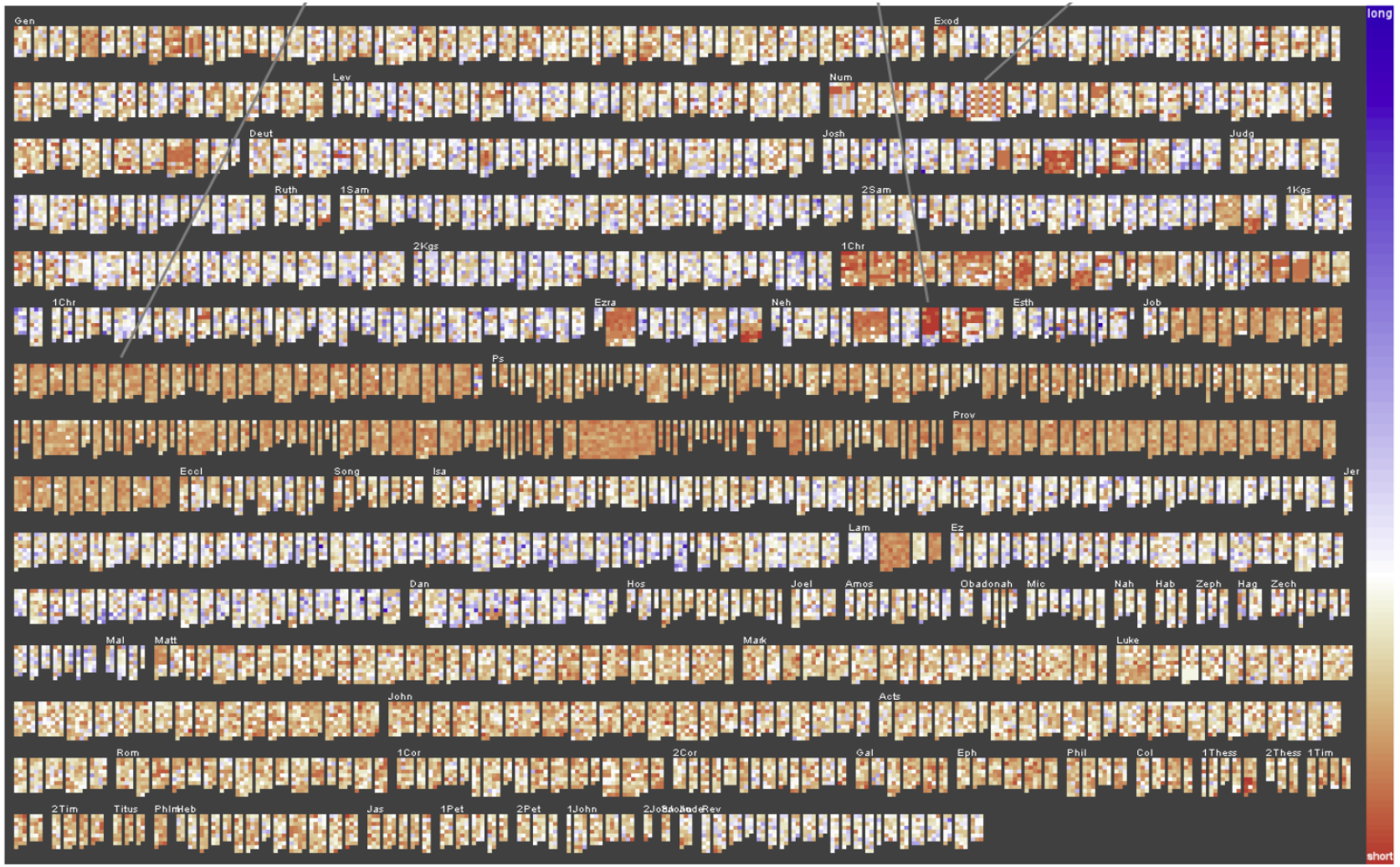
„Literature Fingerprinting: A New Method for Visual Literary Analysis" by Daniel A. Keim and Daniela Oelke

Hapax Legomena



„Literature Fingerprinting: A New Method for Visual Literary Analysis" by Daniel A. Keim and Daniela Oelke

Verse length



„Literature Fingerprinting: A New Method for Visual Literary Analysis" by Daniel A. Keim and Daniela Oelke

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

www.metanome.de

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