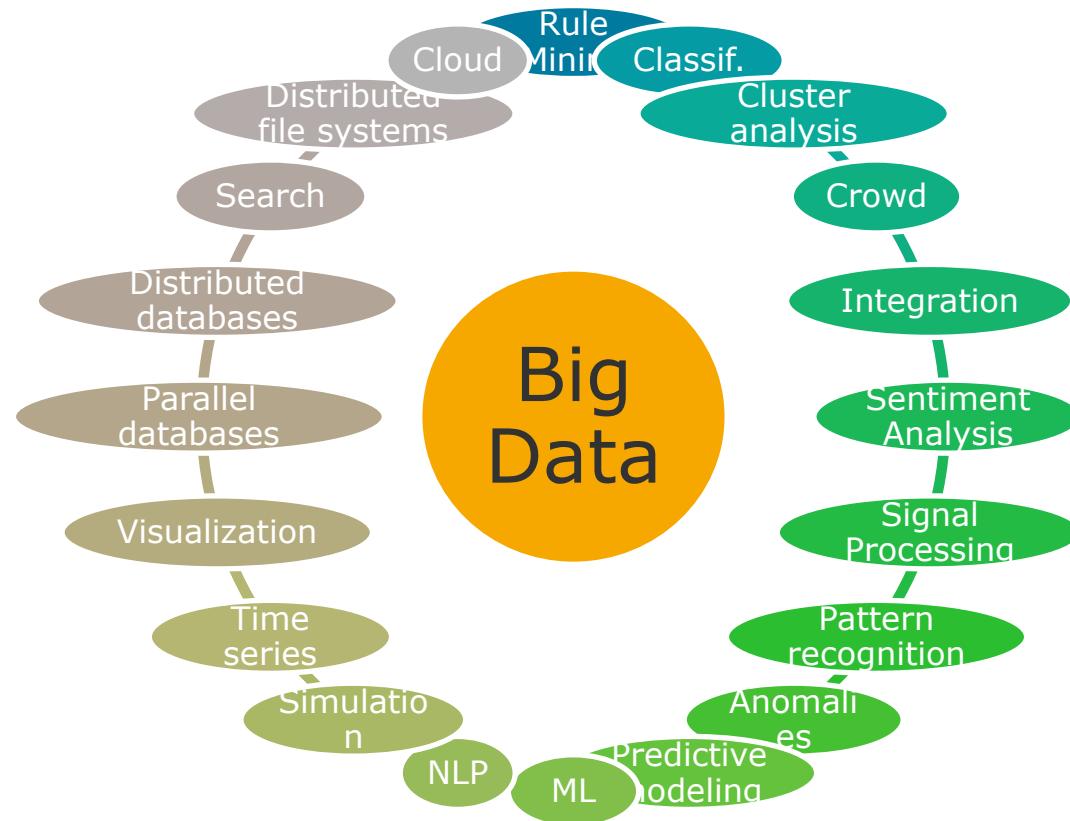


Fondamenti di Data Science e Machine Learning

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

Datei Start Einfügen Seitenlayout Formeln Daten Überprüfen Ansicht Add-Ins

Ausschneiden Kopieren Format übertragen Zellenumbruch Standard Bedingte Formatisierung Ausgabe Gut Neutral Schlecht Einfügen Löschen Format Füllbereich Sortieren Suchen und Filtern Auswählen

Zwischenablage Schriftart Ausrichtung Zahl Formatvorlagen

A1 county_id

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|address|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 GREENSBOO 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 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 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 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 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 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 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 ACTIVE AV CONFIRMATIABAD PRISCILLA MARIE 100 COLONNADE ELON NC 27244 CAMPUS BOX 3008 ELON NC 27244 O HL UNA

11 1 ALAMANCE 9034172 A ACTIVE IU CONFIRMATIABADIE COLLEEN MIASHEL 1097 IVEY RD #C GRAHAM NC 27253 1097 IVEY RD #C GRAHAM NC 27253 M HL REP

12 1 ALAMANCE 9034173 A ACTIVE AV VERIFIED ABADIE JACK EDWARD JR 612 SIDEVIEW ST GRAHAM NC 27253 612 SIDEVIEW ST GRAHAM NC 27253 336 212 8140 W NL UNA

13 1 ALAMANCE 9038377 A ACTIVE IU CONFIRMATIABADIE MYRA HOLLIFIELD 612 SIDEVIEW ST GRAHAM NC 27253 617 MITCHELL ST BURLINGTON NC 27217 336 212 8140 W NL UNA

14 1 ALAMANCE 9038377 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 9038377 A ACTIVE AV VERIFIED Abbas RAFAT 514 WESTRIDGE D BURLINGTON NC 27215 514 WESTRIDGE DR BURLINGTON NC 27215 A UN DEM

16 1 ALAMANCE 9038377 A ACTIVE AV VERIFIED ABATECOLA RONALD JOSEPH JR 504 BROOKFIELD G GIBSONVILLE NC 27249 504 BROOKFIELD DR GIBSONVILLE NC 27249 336 449 9029 W UN UNA

17 1 ALAMANCE 9038377 A ACTIVE AV VERIFIED ABATECOLA TRACY BOONE 504 BROOKFIELD G GIBSONVILLE NC 27249 504 BROOKFIELD DR GIBSONVILLE NC 27249 W NL DEM

18 1 ALAMANCE 9083557 I INACTIVE IU CONFIRMATIABBETT DAWN LEANN 3900 JOHNS CREEK GIBSONVILLE NC 27249 3900 JOHNS 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 902529 I INACTIVE IU CONFIRMATIABBEY DOROTHY ESTELLA 1029A QUAKENBUSH SNOW CAMP NC 27349 1029A QUAKENBUSH RD SNOW CAM NC 27349 376 3663 W NL REP

22 1 ALAMANCE 9113186 A ACTIVE AV VERIFIED ABBOTT AMELIA BETH 2876 CALLOWAY D MEBANE 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 COURTENEY 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 JAMESTOWNE 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 D MEBANE 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 FROI ABBOTT LATWOIA BEREKA 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 JR 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 JAMESTOWNE DR ELON NC 27244 336 228 0571 W UN REP

41 1 ALAMANCE 9090949 R REMOVED RM REMOVED AI ABBOTT RACHEL MARA 103 DANIELEY CEN ELON 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 D MEBANE NC 27302 2876 CALLOWAY DR MEBANE NC 27302 919 568 8056 W NL UNA

43 1 ALAMANCE 9113731 I INACTIVE IU CONFIRMATIABBOTT TAYLOR RENEE 406 W LEBANON A ELON NC 27244 CAMPUS BOX 3077 ELON NC 27244 W NL UN REP

44 1 ALAMANCE 9120825 I INACTIVE IN CONFIRMATIABBOTT 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 CONFIRMATIABBOTT 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-MAGLISA ANN 1841 DUNBAR PL BURLINGTON NC 27215 1841 DUNBAR PL BURLINGTON NC 27215 214 437 8955 W NL UNA

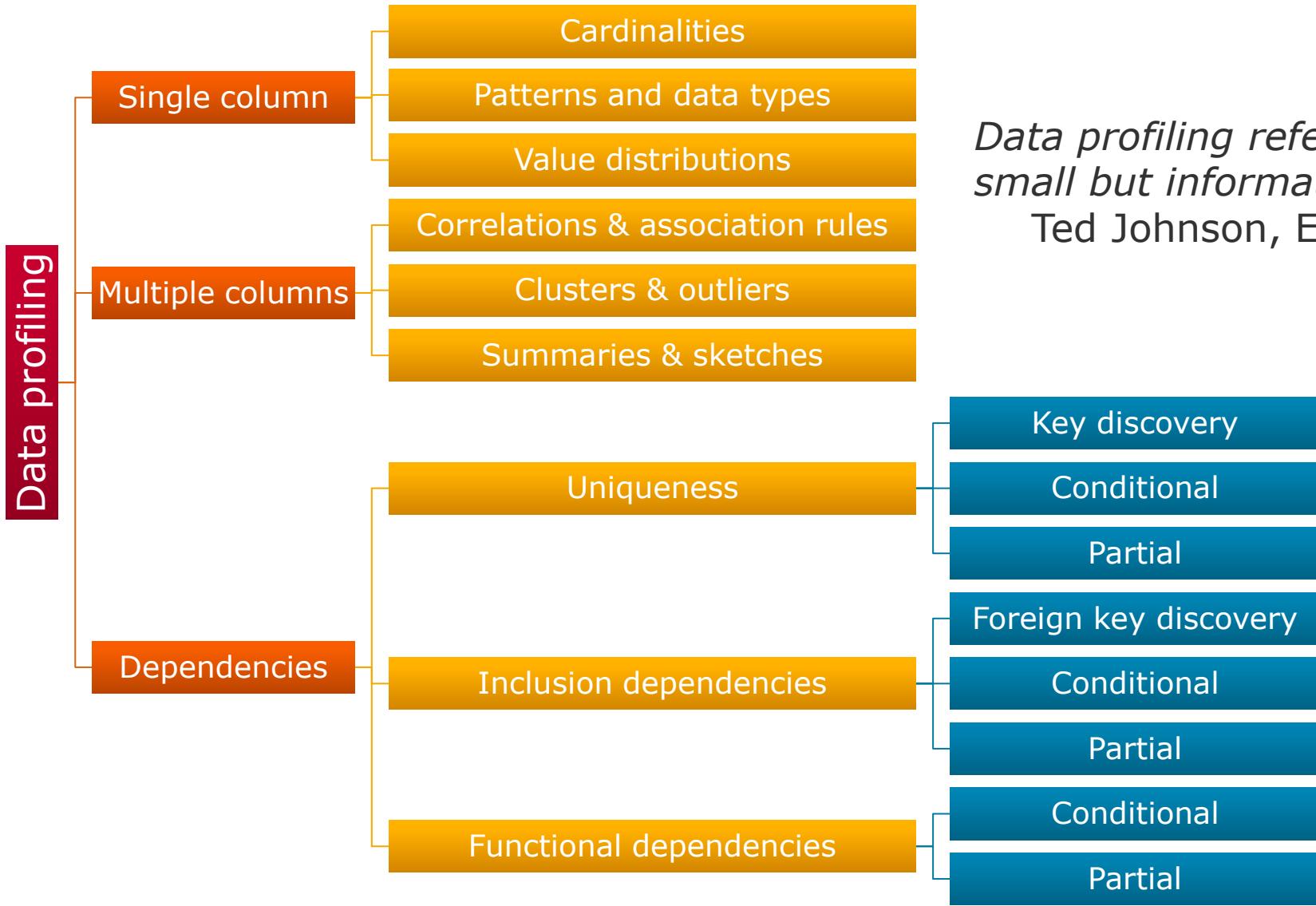
49 1 ALAMANCE 9117192 I INACTIVE IU CONFIRMATIABDELKARIM AMNA ELHAG 1105 PROVIDENCE ELON NC 27244 1105 PROVIDENCE CT ELON NC 27244 M NL UNA

Excel screenshot showing a large dataset in the 'ncvoter1' sheet. A red arrow points from the bottom left towards the top-left cell, labeled 'Number of rows'.

The dataset contains approximately 100,000 rows, starting with row 106185. The first few columns (A, B, C, D, E) are highlighted in yellow. The last few columns (T, U, V, W) are highlighted in light blue. The first column (A) contains numerical IDs (e.g., 106185, 106186), while columns B through W contain various categorical and descriptive data. Row 106185 is highlighted in yellow, and the cell containing '1 ALAMANCE' is also highlighted.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
106185	1 ALAMANCE	9129972 A	ACTIVE	AV	VERIFIED	ZLUCHOWSK	AARON MICHAEL			3551 FORESTDALE BURLINGTON NC	27215 3551 FORESTDALE DR #M BURLINGTON NC	27215 336 270 6878 W	NL UNA									
106186	1 ALAMANCE	9106623 A	ACTIVE	AV	VERIFIED	ZMIJEASKI SEAN				4872 THOM RD MEBANE NC	27302 4872 THOM RD MEBANE NC	27302 336 376 1987 O	UN REP									
106187	1 ALAMANCE	9112148 A	ACTIVE	AV	VERIFIED	ZMIJEWSKI DENNIS AL				4872 THOM RD MEBANE NC	27302 4872 THOM RD MEBANE NC	27302 W	UN UN									
106188	1 ALAMANCE	9094109 I	INACTIVE	IU	CONFIRMATI	ZMIJEWSKI DENNIS				4872 THOM RD MEBANE NC	27302 4872 THOM RD MEBANE NC	27302 336 376 1987 W	UN DEM									
106189	1 ALAMANCE	9128345 A	ACTIVE	AV	VERIFIED	ZMIJEWSKI KEVIN ADAM				4872 THOM RD MEBANE NC	27302 4872 THOM RD MEBANE NC	27302 336 380 5768 W	NL UNA									
106190	1 ALAMANCE	9120294 A	ACTIVE	AV	VERIFIED	ZMIJEWSKI SEAN CHRISTOPHE				4872 THOM RD MEBANE NC	27302 4872 THOM RD MEBANE NC	27302 W	HL UNA									
106191	1 ALAMANCE	9094116 A	ACTIVE	AV	VERIFIED	ZMIJEWSKI IN VIRGINIA LOURDES				4872 THOM RD MEBANE NC	27302 4872 THOM RD MEBANE NC	27302 336 376 1987 U	UN UNA									
106192	1 ALAMANCE	9089250 R	REMOVED	RD	DECEASED	ZOCOLANTIRENIS PIZZOTTI				2502 S NC HWY 119 MEBANE NC	27302 2502 S NC HWY 119 MEBANE NC	27302 W	UN REP									
106193	1 ALAMANCE	9083629 R	REMOVED	RD	DECEASED	ZOCOLANTIRENATO				3141 SHELLY GRAH GRAHAM	27253 3141 SHELLY GRAH GRAHAM DR GRAHAM NC	27253 336 227 7168 W	NL REP									
106194	1 ALAMANCE	9083630 A	ACTIVE	AV	VERIFIED	ZOCOLANTIRITA MARIE				3141 SHELLY GRAH GRAHAM	27253 3141 SHELLY GRAH GRAHAM DR GRAHAM NC	27253 336 227 7168 W	NL REP									
106195	1 ALAMANCE	9100545 I	INACTIVE	IU	CONFIRMATI	ZOLEGMANN ANGELA LYNNE				706 HUFFMAN MII BURLINGTON NC	27215 706 HUFFMAN MILL RD # BURLINGTON NC	27215 336 227 1261 W	NL UNA									
106196	1 ALAMANCE	9137285 A	ACTIVE	AV	VERIFIED	ZOLAYVAR ERIC WATSON				910 COLONIAL DR BURLINGTON NC	27215 910 COLONIAL DR BURLINGTON NC	27215 336 585 0248 O	NL DEM									
106197	1 ALAMANCE	9081869 A	ACTIVE	AV	VERIFIED	ZOLAYVAR RUPERTO BENEDICTO				910 COLONIAL DR BURLINGTON NC	27215 910 COLONIAL DR BURLINGTON NC	27215 336 585 0248 O	NL DEM									
106198	1 ALAMANCE	9109021 A	ACTIVE	AV	VERIFIED	ZOLAYVAR STEPHANIE WATSON				910 COLONIAL DR BURLINGTON NC	27215 910 COLONIAL DR BURLINGTON NC	27215 336 585 0248 W	NL UNA									
106199	1 ALAMANCE	9108096 A	ACTIVE	AV	VERIFIED	ZOLLARS EVELYN NADINE				6830 TOM WOODY SNOW CAMP NC	27349 6830 TOM WOODY RD SNOW CAMF NC	27349 336 376 5754 W	NL UNA									
106200	1 ALAMANCE	9125044 A	ACTIVE	AV	VERIFIED	ZOLLARS MATHEW DAVID				6830 TOM WOODY SNOW CAMP NC	27349 6830 TOM WOODY RD SNOW CAMF NC	27349 W	NL UNA									
106201	1 ALAMANCE	9113912 A	ACTIVE	AV	VERIFIED	ZOLLICOFFEE ANTONIO MARK				108 OAKGROVE D GRAHAM	27253 108 OAKGROVE DR GRAHAM NC	27253 336 260 6673 B	UN DEM									
106202	1 ALAMANCE	9107068 A	ACTIVE	AV	VERIFIED	ZOLLICOFFEE VALERIE				108 OAKGROVE D GRAHAM	27253 108 OAKGROVE DR GRAHAM NC	27253 B	UN DEM									
106203	1 ALAMANCE	9097324 A	ACTIVE	AV	VERIFIED	ZORNES ASHLEY DENICE				5556 N NC HWY 49 MEBANE NC	27302 5556 N NC HWY 49 MEBANE NC	27302 336 578 1157 W	NL UNA									
106204	1 ALAMANCE	9038407 A	ACTIVE	AV	VERIFIED	ZORNES KENNETH ELWOOD				5556 N NC HWY 49 MEBANE NC	27302 5556 N NC HWY 49 MEBANE NC	27302 W	NL UNA									
106205	1 ALAMANCE	9104969 I	INACTIVE	IU	CONFIRMATI	ZORNES MICHELLE LEE				3117 COMMERCIAL BURLINGTON NC	27215 3117 COMMERCIAL PL #L BURLINGTON NC	27215 336 675 0520 W	UN UNA									
106206	1 ALAMANCE	9018738 A	ACTIVE	AV	VERIFIED	ZORNES SHERRIE AVERETTE				5556 N NC HWY 49 MEBANE NC	27302 5556 N NC HWY 49 MEBANE NC	27302 W	NL DEM									
106207	1 ALAMANCE	9027412 I	INACTIVE	IU	CONFIRMATI	ZORNES TERRY LEE				148 N STATE ST HAW RIVER NC	27258 148 N STATE ST HAW RIVER NC	27258 570 1633 W	NL DEM									
106208	1 ALAMANCE	9110367 D	DENIED	DU	VERIFICATIO	ZORNES TINA				801 TROLLINGWOOD HAW RIVER NC	27258 801 TROLLINGWOOD RD HAW RIVER NC	27258 336 578 0646 W	UN UNA									
106209	1 ALAMANCE	9132758 A	ACTIVE	AV	VERIFIED	ZORNES TINA MARIE				801 TROLLINGWOOD HAW RIVER NC	27258 801 TROLLINGWOOD RD HAW RIVER NC	27258 336 420 7630 W	NL UNA									
106210	1 ALAMANCE	9131499 A	ACTIVE	AV	VERIFIED	ZOUFALY EVE				602 E HAGGARD A'ELON	27244 CAMPUS BOX 8911 ELON NC	27244 U	UN UNA									
106211	1 ALAMANCE	9124446 A	ACTIVE	AV	VERIFIED	ZSUPPAN ETELKA HALASZ				1929 HAW VILLAGE GRAHAM	27253 1929 HAW VILLAGE DR GRAHAM NC	27253 W	NL REP									
106212	1 ALAMANCE	9121554 A	ACTIVE	AV	VERIFIED	ZSUPPAN FERENC				1929 HAW VILLAGE GRAHAM	27253 1929 HAW VILLAGE DR GRAHAM NC	27253 W	NL REP									
106213	1 ALAMANCE	9127457 A	ACTIVE	AV	VERIFIED	ZSUPPAN LEVENTE FERENC				1929 HAW VILLAGE GRAHAM	27253 1929 HAW VILLAGE DR GRAHAM NC	27253 336 376 1365 W	NL REP									
106214	1 ALAMANCE	9131401 A	ACTIVE	AV	VERIFIED	ZUBLER LINDSAY BROOKE				3172 CARRIAGE CF HAW RIVER NC	27258 3172 CARRIAGE CREEK CT HAW RIVER NC	27258 U	UN UNA									
106215	1 ALAMANCE	9081728 A	ACTIVE	AV	VERIFIED	ZUBLER TAMI LAJEAN				3172 CARRIAGE CF HAW RIVER NC	27258 3172 CARRIAGE CREEK CT HAW RIVER NC	27258 336 578 8028 W	NL UNA									
106216	1 ALAMANCE	9089569 A	ACTIVE	AV	VERIFIED	ZUBLER TIMOTHY JAMES				3172 CARRIAGE CF HAW RIVER NC	27258 3172 CARRIAGE CREEK CT HAW RIVER NC	27258 W	UN UNA									
106217	1 ALAMANCE	9070674 A	ACTIVE	AV	VERIFIED	ZUBOV ALEX				229 ENGLEMAN A' BURLINGTON NC	27215 229 ENGLEMAN AVE BURLINGTON NC	27215 336 437 9776 W	NL UNA									
106218	1 ALAMANCE	9070288 A	ACTIVE	AV	VERIFIED	ZUBOV LYNN R				229 ENGLEMAN A' BURLINGTON NC	27215 229 ENGLEMAN AVE BURLINGTON NC	27215 336 437 9776 W	NL REP									
106219	1 ALAMANCE	9008787 A	ACTIVE	AV	VERIFIED	ZUMER FRANK EDWARD				801 QUAKER RIDGE MEBANE NC	27302 801 QUAKER RIDGE RD MEBANE NC	27302 919 563 3766 W	UN UNA									
106220	1 ALAMANCE	9008785 A	ACTIVE	AV	VERIFIED	ZUMER LOUISE TURNER				801 QUAKER RIDGE MEBANE NC	27302 801 QUAKER RIDGE RD MEBANE NC	27302 919 563 3766 W	NL DEM									
106221	1 ALAMANCE	9141817 A	ACTIVE	AV	VERIFIED	ZUNG PATRICK BATE				2604 WOODS LN GRAHAM NC	27253 2604 WOODS LN GRAHAM NC	27253 919 357 3896 W	NL DEM									
106222	1 ALAMANCE	9119438 A	ACTIVE	AV	VERIFIED	ZUNIGA JOSE RAMON SAL				714 ROSS ST BURLINGTON NC	27217 714 ROSS ST BURLINGTON NC	27217 336 227 3108 O	HL DEM									
106223	1 ALAMANCE	9108610 A	ACTIVE		VERIFIED	ZUNIGA VANESA ELIZABETH				512 PIEDMONT W BURLINGTON NC	27217 512 PIEDMONT WAY BURLINGTON NC	27217 336 270 0181 W	HL DEM									
106224	1 ALAMANCE	9112637 A	ACTIVE		VERIFIED	ZUNIGA YANET SALAS				3845 MAE DOUGLAS MEBANE NC	27302 3845 MAE DOUGLAS DR MEBANE NC	27302 O	HL DEM									
106225	1 ALAMANCE	9141392 A			VERIFIED	ZUPANCICH MONICA ANITA				2326 N NC HWY 49 BURLINGTON NC	27217 2326 N NC HWY 49 BURLINGTON NC	27217 330 310 0151 W	NL REP									
106226	1 ALAMANCE	9141404 A			VERIFIED	ZUPANCICH RONALD JAMES II				2326 N NC HWY 49 BURLINGTON NC	27217 2326 N NC HWY 49 BURLINGTON NC	27217 757 254 3773 W	NL REP									
106227	1 ALAMANCE	9099261 A	ACTIVE	AV	VERIFIED	ZURFACE ROSSELL EUGENE				2074 TURNER RD MEBANE NC	27302 2074 TURNER RD MEBANE NC	27302 W	UN UNA									
106228	1 ALAMANCE	9099499 A	ACTIVE	AV	VERIFIED	ZWIER ANDREW MICHAEL				1497 LONGEST ACISNOW CAMP NC	27349 1497 LONGEST ACRES RD SNOW CAMF NC	27349 336 376 8830 W	NL REP									
106229	1 ALAMANCE	90977804 R	REMOVED	RL	MOVED FRO	ZWIER MARC				1497 LONGEST ACISNOW CAMP NC	27349 1497 LONGEST ACRES RD SNOW CAMF NC	27349 831 207 9222 W	NL REP									
106230	1 ALAMANCE	9099261 A	ACTIVE	AV	VERIFIED	ZWIER CHRISTY ANN				1497 LONGEST ACISNOW CAMP NC	27349 1497 LONGEST ACRES RD SNOW CAMF NC	27349 W	NL REP									
106231	1 ALAMANCE	9099261 A	ACTIVE	AV	VERIFIED	ZWIER KAREN JEAN				1497 LONGEST ACISNOW CAMP NC	27349 1497 LONGEST ACRES RD SNOW CAMF NC	27349 831 207 9222 W	NL REP									
106232	1 ALAMANCE	9099261 A	ACTIVE	AV	VERIFIED	ZWIER KAREN MARC				1210 WILLOW BRC MEBANE NC	27302 1210 WILLOW BROOK CT MEBANE NC	27302 336 578 8580 W	NL REP									

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

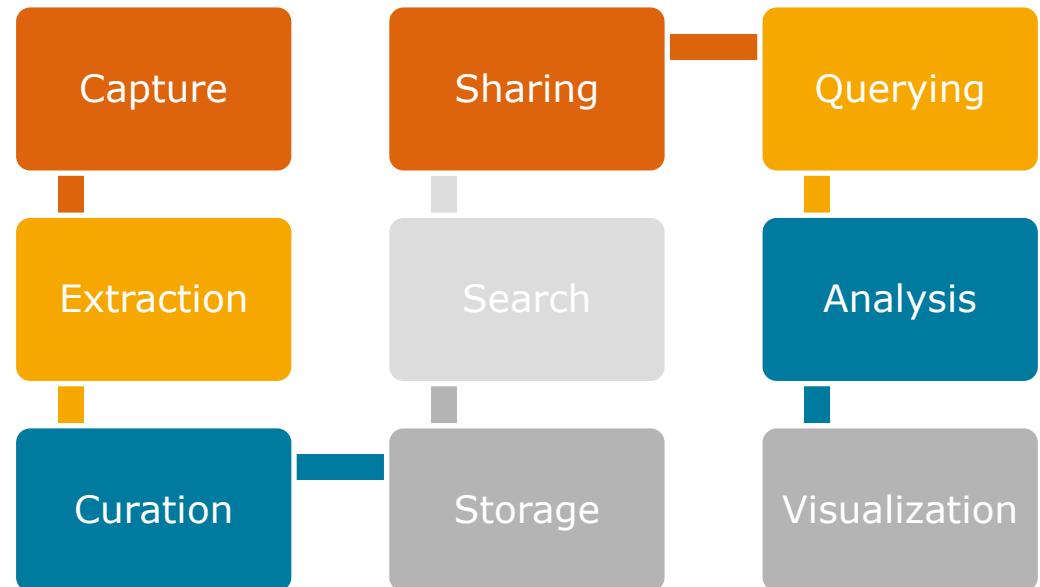
- **Query optimization:** Counts and histograms, functional dependencies, ...
- **Data cleansing:** Patterns, rules, and violations
- **Data integration:** Cross-DB inclusion dependencies
- **Scientific data management:** Inspect new datasets
- **Data analytics and mining:** Profiling as preparation 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...”

Defining Big Data

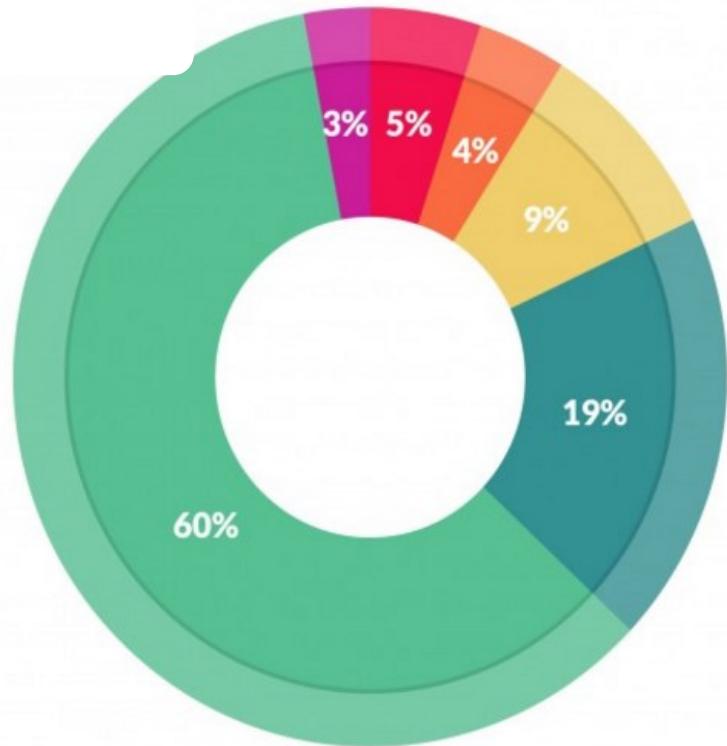
Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.



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



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%*

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

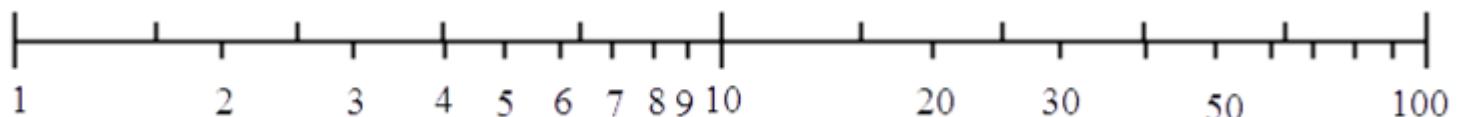
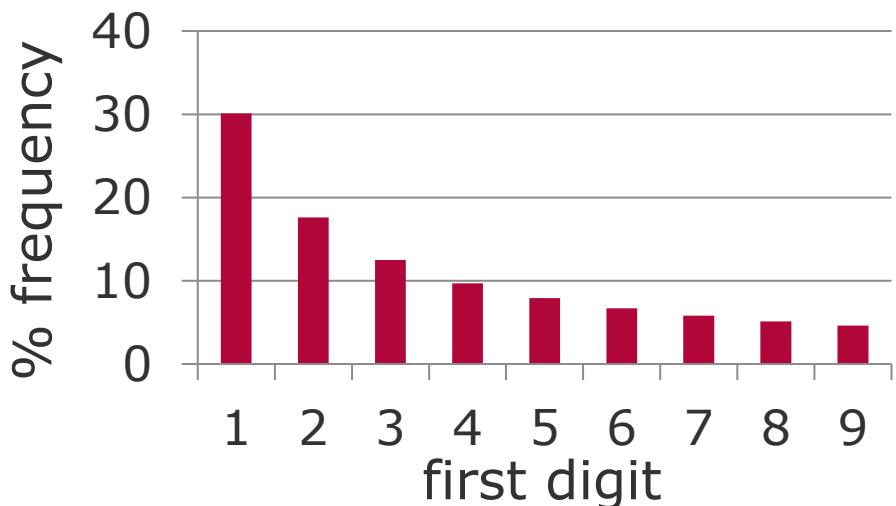


Cardinalities, distributions, and patterns

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.

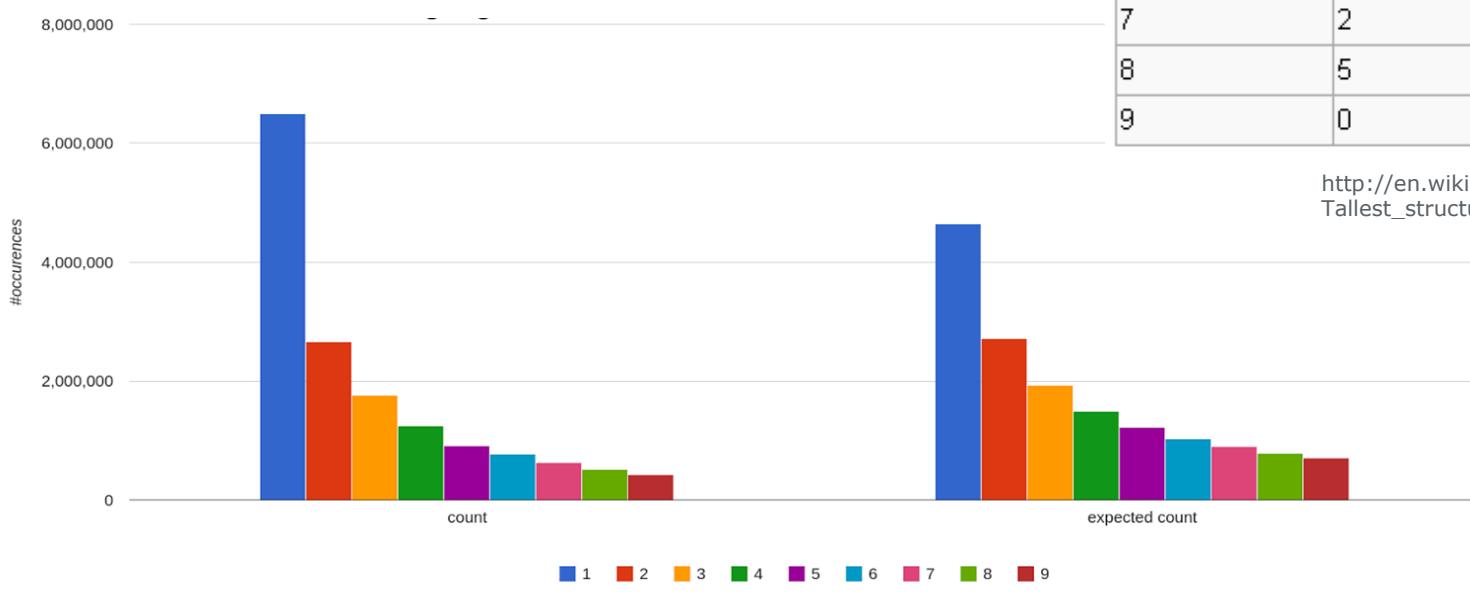
Benford Law Frequency , a.k.a. “first digit law”

- Statement about the distribution of first digits d in (many) *naturally occurring* numbers:
 - $P(d) = \log_{10}(d + 1) - \log_{10}(d) = \log_{10}(1 + \frac{1}{d})$
 - 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	
	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

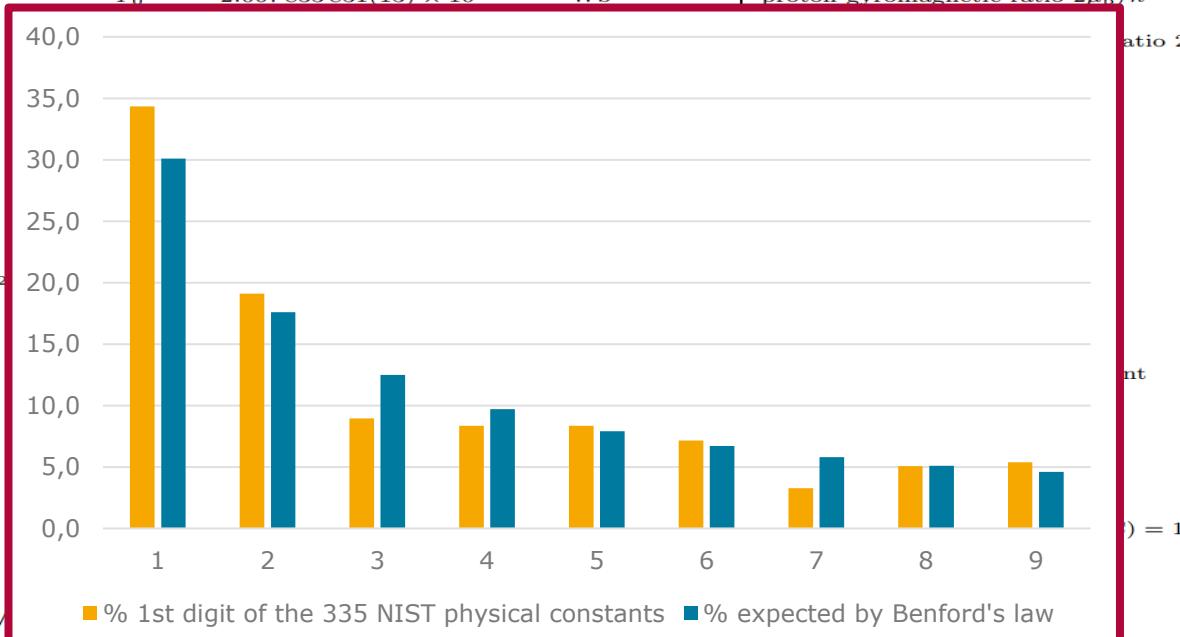
CODATA RECOMMENDED VALUES OF THE FUNDAMENTAL PHYSICAL CONSTANTS: 2014

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	Quantity	Symbol	Numerical value	Unit
speed of light in vacuum	c, c_0	299 792 458 (exact)	m s^{-1}	muon g -factor $-2(1 + a_\mu)$	g_μ	-2.002 331 8418(13)	
magnetic constant	μ_0	$4\pi \times 10^{-7}$ (exact)	N A^{-2}	muon-proton magnetic moment ratio	μ_μ/μ_p	-3.183 345 142(71)	
electric constant $1/\mu_0 c^2$	ϵ_0	$= 12.566 370 614\dots \times 10^{-7}$	N A^{-2}	proton mass in u	m_p	$1.672 621 898(21) \times 10^{-27}$	kg
Newtonian constant of gravitation	G	$6.674 08(31) \times 10^{-11}$	F m^{-1}	energy equivalent in MeV	$m_p c^2$	1.007 276 466 879(91)	u
Planck constant in eV s	h	$6.626 070 040(81) \times 10^{-34}$	J s	proton-electron mass ratio	m_p/m_e	938.272 0813(58)	MeV
$h/2\pi$ in eV s		$4.135 667 662(25) \times 10^{-15}$	eV s	proton magnetic moment to nuclear magneton ratio	μ_p	1836.152 673 89(17)	
$h/2\pi$ in eV s	\hbar	$1.054 571 800(13) \times 10^{-34}$	J s	proton magnetic moment to nuclear magneton ratio	μ_p/μ_N	1.410 606 7873(97) $\times 10^{-26}$	J T^{-1}
elementary charge	e	$6.582 119 514(40) \times 10^{-16}$	eV s	proton magnetic shielding correction $1 - \mu'_p/\mu_p$	σ'_p	2.792 847 3508(85)	
magnetic flux quantum $h/2e$	Φ_0	$1.602 176 6208(98) \times 10^{-19}$	C	(H_2O , sphere, 25 °C)		$25.691(11) \times 10^{-6}$	
Josephson constant $2e/h$		$2.067 833 831(13) \times 10^{-15}$	Wb	proton gyromagnetic ratio $2\mu_p/\hbar$	γ_p	$2.675 221 900(18) \times 10^8$	$\text{s}^{-1} \text{T}^{-1}$
von Klitzing constant $h/e^2 = \mu_0 c/2\alpha$				ratio $2\mu'_p/\hbar$	$\gamma_p/2\pi$	42.577 478 92(29)	MHz T^{-1}
Bohr magneton $e\hbar/2m_e$ in eV T ⁻¹					γ'_p	$2.675 153 171(33) \times 10^8$	$\text{s}^{-1} \text{T}^{-1}$
nuclear magneton $e\hbar/2m_p$ in eV T ⁻¹					$\gamma'_p/2\pi$	42.576 385 07(53)	MHz T^{-1}
fine-structure constant $e^2/4\pi\epsilon_0\hbar c$ inverse fine-structure constant					m_n	1.008 664 915 88(49)	u
Rydberg constant $\alpha^2 m_e c/2h$ energy equivalent in eV					$m_n c^2$	939.565 4133(58)	MeV
Bohr radius $\alpha/4\pi R_\infty = 4\pi\epsilon_0\hbar^2/m_e e^2$					m_n/m_p	1.001 378 418 98(51)	
Hartree energy $e^2/4\pi\epsilon_0 a_0 = 2R_\infty hc = \alpha^2$ in eV					μ_n	$-0.966 236 50(23) \times 10^{-26}$	J T^{-1}
electron mass in u energy equivalent in MeV					μ_n/μ_N	-1.913 042 73(45)	
electron-muon mass ratio					m_d	2.013 553 212 745(40)	u
electron-proton mass ratio					$m_d c^2$	1875.612 928(12)	MeV
electron charge to mass quotient					m_d/m_p	1.999 007 500 87(19)	
Compton wavelength h/m_ec $\lambda_C/2\pi = \alpha a_0 = \alpha^2/4\pi R_\infty$					μ_d	0.433 073 5040(36) $\times 10^{-26}$	J T^{-1}
classical electron radius $\alpha^2 a_0$					μ_d/μ_N	0.857 438 2311(48)	
Thomson cross section $(8\pi/3)r_e^2$					m_h	3.014 932 246 73(12)	u
electron magnetic moment to Bohr magneton ratio					$m_h c^2$	2808.391 586(17)	MeV
to nuclear magneton ratio					μ'_h	$-1.074 553 080(14) \times 10^{-26}$	J T^{-1}
electron magnetic moment anomaly $ \mu_e $ $ \mu_e /(e\hbar/2m_e) - 1$					μ'_h/μ_B	-1.158 671 471(14) $\times 10^{-3}$	
electron g-factor $-2(1 + a_e)$					μ'_h/μ_N	-2.127 497 720(25)	
electron-proton magnetic moment ratio					m_a	4.001 506 179 127(63)	u
muon mass in u energy equivalent in MeV	μ_e/μ_p	-658.210 6866(20)			$m_a c^2$	3727.379 378(23)	MeV
muon-electron mass ratio	m_μ	0.113 428 9257(25)	u		N_A, L	$6.022 140 857(74) \times 10^{23}$	mol^{-1}
muon magnetic moment to Bohr magneton ratio	$m_\mu c^2$	105.658 3745(24)	MeV		m_u	1.660 539 040(20) $\times 10^{-27}$	kg
to nuclear magneton ratio	m_μ/m_e	206.768 2826(46)			$m_u c^2$	931.494 0954(57)	MeV
muon magnetic moment anomaly $ \mu_\mu /(e\hbar/2m_\mu) - 1$	μ_μ	$-4.490 448 26(10) \times 10^{-26}$	J T^{-1}		F	96 485.332 89(59)	C mol^{-1}
muon magnetic moment to Bohr magneton ratio	μ_μ/μ_B	-4.841 970 48(11) $\times 10^{-3}$			R	8.314 4598(48)	$\text{J mol}^{-1} \text{ K}^{-1}$
to nuclear magneton ratio	μ_μ/μ_N	-8.890 597 05(20)			k	$1.380 648 52(79) \times 10^{-23}$	J K^{-1}
muon magnetic moment anomaly $ \mu_\mu /(e\hbar/2m_\mu) - 1$	a_μ	$1.165 920 89(63) \times 10^{-3}$			V_m	$8.617 3303(50) \times 10^{-5}$	eV K^{-1}
						$22.413 962(13) \times 10^{-3}$	$\text{m}^3 \text{ mol}^{-1}$
Energy equivalents							
$(1 \text{ m}^{-1})c$		299 792 458 Hz		$(1 \text{ J}) = 6.241 509 126(38) \times 10^{18} \text{ eV}$		$(1 \text{ eV})/c^2 = 1.073 544 1105(66) \times 10^{-9} \text{ u}$	
$(1 \text{ m}^{-1})hc/k = 1.438 777 36(83) \times 10^{-2} \text{ K}$		$(1 \text{ Hz})h = 4.799 2447(28) \times 10^{-11} \text{ K}$		$(1 \text{ eV}) = 1.602 176 6208(98) \times 10^{-19} \text{ J}$		$(1 \text{ kg}) = 6.022 140 857(74) \times 10^{26} \text{ u}$	
$(1 \text{ m}^{-1})hc = 1.239 841 9739(76) \times 10^{-6} \text{ eV}$		$(1 \text{ Hz})h = 4.135 667 662(25) \times 10^{-15} \text{ eV}$		$(1 \text{ eV})/hc = 8.065 544 005(50) \times 10^5 \text{ m}^{-1}$		$(1 \text{ u}) = 1.660 539 040(20) \times 10^{-27} \text{ kg}$	
$(1 \text{ m}^{-1})h/c = 1.331 025 049 00(61) \times 10^{-15} \text{ u}$		$(1 \text{ K})/h = 69.503 457(40) \text{ m}^{-1}$		$(1 \text{ eV})/h = 2.417 989 262(15) \times 10^{14} \text{ Hz}$		$(1 \text{ u})/c/h = 7.513 006 6166(34) \times 10^{14} \text{ m}^{-1}$	
$(1 \text{ Hz})/c = 3.335 640 951 \dots \times 10^{-9} \text{ m}^{-1}$		$(1 \text{ K})k/h = 2.083 6612(12) \times 10^{10} \text{ Hz}$		$(1 \text{ eV})/k = 1.160 452 21(67) \times 10^4 \text{ K}$		$(1 \text{ u})/c^2 = 931.494 0954(57) \times 10^6 \text{ eV}$	
		$(1 \text{ K})k = 8.617 3303(50) \times 10^{-5} \text{ eV}$					



$$(1 \text{ m}^{-1})c = 299 792 458 \text{ Hz}$$

$$(1 \text{ m}^{-1})hc/k = 1.438 777 36(83) \times 10^{-2} \text{ K}$$

$$(1 \text{ m}^{-1})hc = 1.239 841 9739(76) \times 10^{-6} \text{ eV}$$

$$(1 \text{ m}^{-1})h/c = 1.331 025 049 00(61) \times 10^{-15} \text{ u}$$

$$(1 \text{ Hz})/c = 3.335 640 951 \dots \times 10^{-9} \text{ m}^{-1}$$

$$(1 \text{ Hz})h/k = 4.799 2447(28) \times 10^{-11} \text{ K}$$

$$(1 \text{ Hz})h = 4.135 667 662(25) \times 10^{-15} \text{ eV}$$

$$(1 \text{ K})/h = 69.503 457(40) \text{ m}^{-1}$$

$$(1 \text{ K})k/h = 2.083 6612(12) \times 10^{10} \text{ Hz}$$

$$(1 \text{ K})k = 8.617 3303(50) \times 10^{-5} \text{ eV}$$

$$(1 \text{ J}) = 6.241 509 126(38) \times 10^{18} \text{ eV}$$

$$(1 \text{ eV}) = 1.602 176 6208(98) \times 10^{-19} \text{ J}$$

$$(1 \text{ eV})/hc = 8.065 544 005(50) \times 10^5 \text{ m}^{-1}$$

$$(1 \text{ eV})/h = 2.417 989 262(15) \times 10^{14} \text{ Hz}$$

$$(1 \text{ eV})/k = 1.160 452 21(67) \times 10^4 \text{ K}$$

$$(1 \text{ eV})/c^2 = 1.073 544 1105(66) \times 10^{-9} \text{ u}$$

$$(1 \text{ kg}) = 6.022 140 857(74) \times 10^{26} \text{ u}$$

$$(1 \text{ u}) = 1.660 539 040(20) \times 10^{-27} \text{ kg}$$

$$(1 \text{ u})/c/h = 7.513 006 6166(34) \times 10^{14} \text{ m}^{-1}$$

$$(1 \text{ u})/c^2 = 931.494 0954(57) \times 10^6 \text{ eV}$$

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

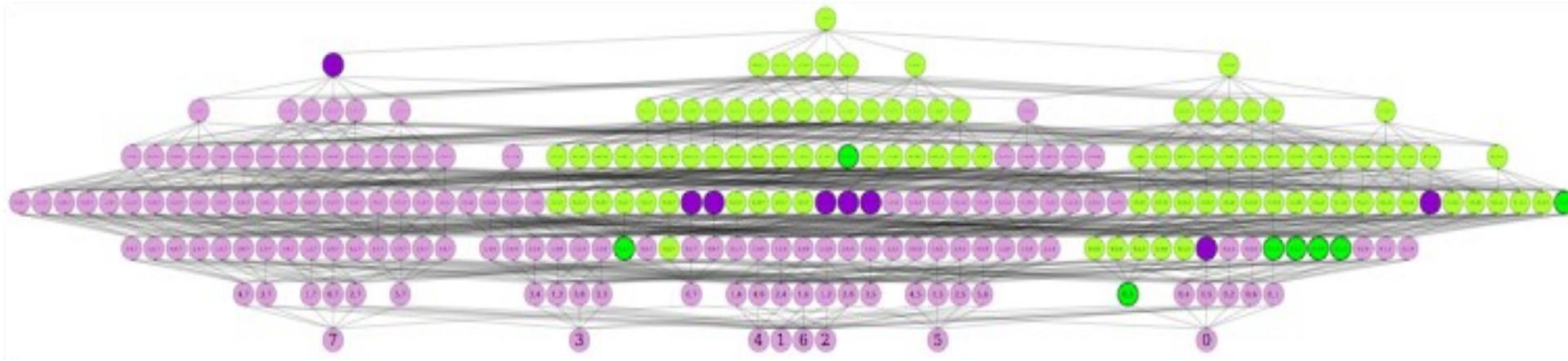
- Uniqueness and keys
 - Unique column: Only unique values
 - Unique column combination: Only unique value combinations
 - Minimality: No column subset is unique
 - Key candidate: No null values
 - Key: Only human expert can decide
 - UCC is prerequisite

- 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

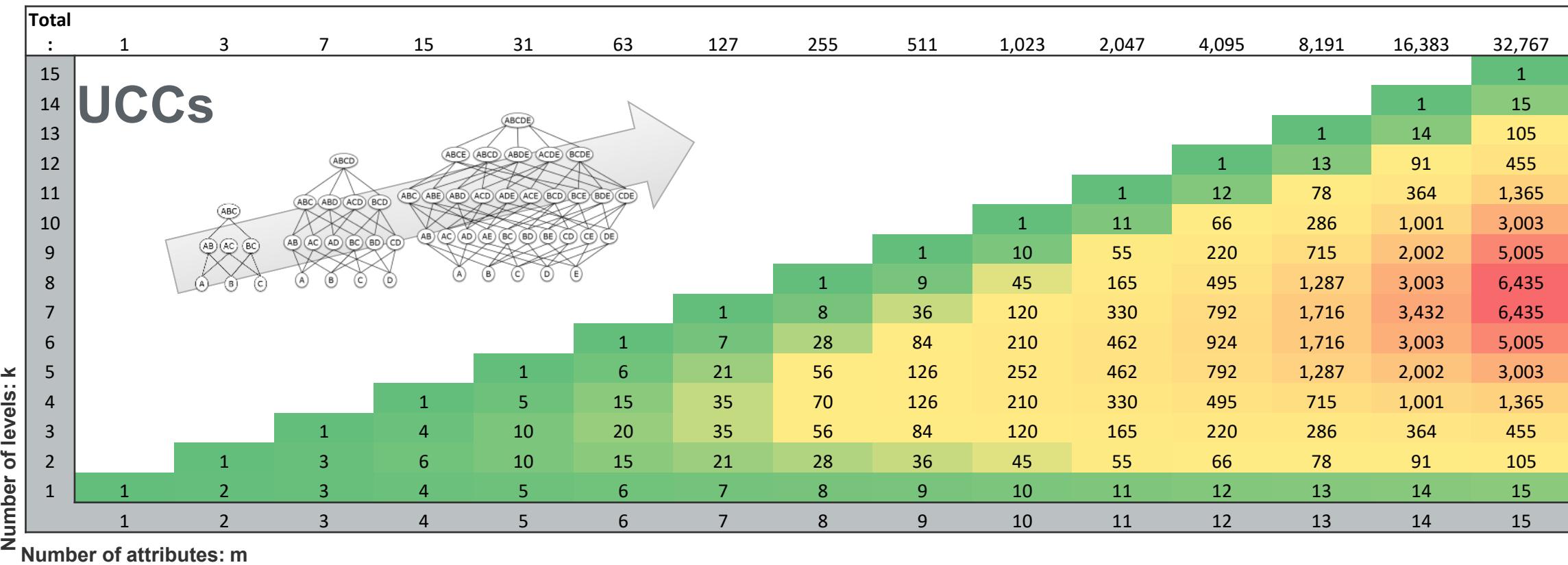
- Useful for
 - Schema design, data integration, indexing, optimization
 - Inverse: non-uniques are duplicates

Large solution space

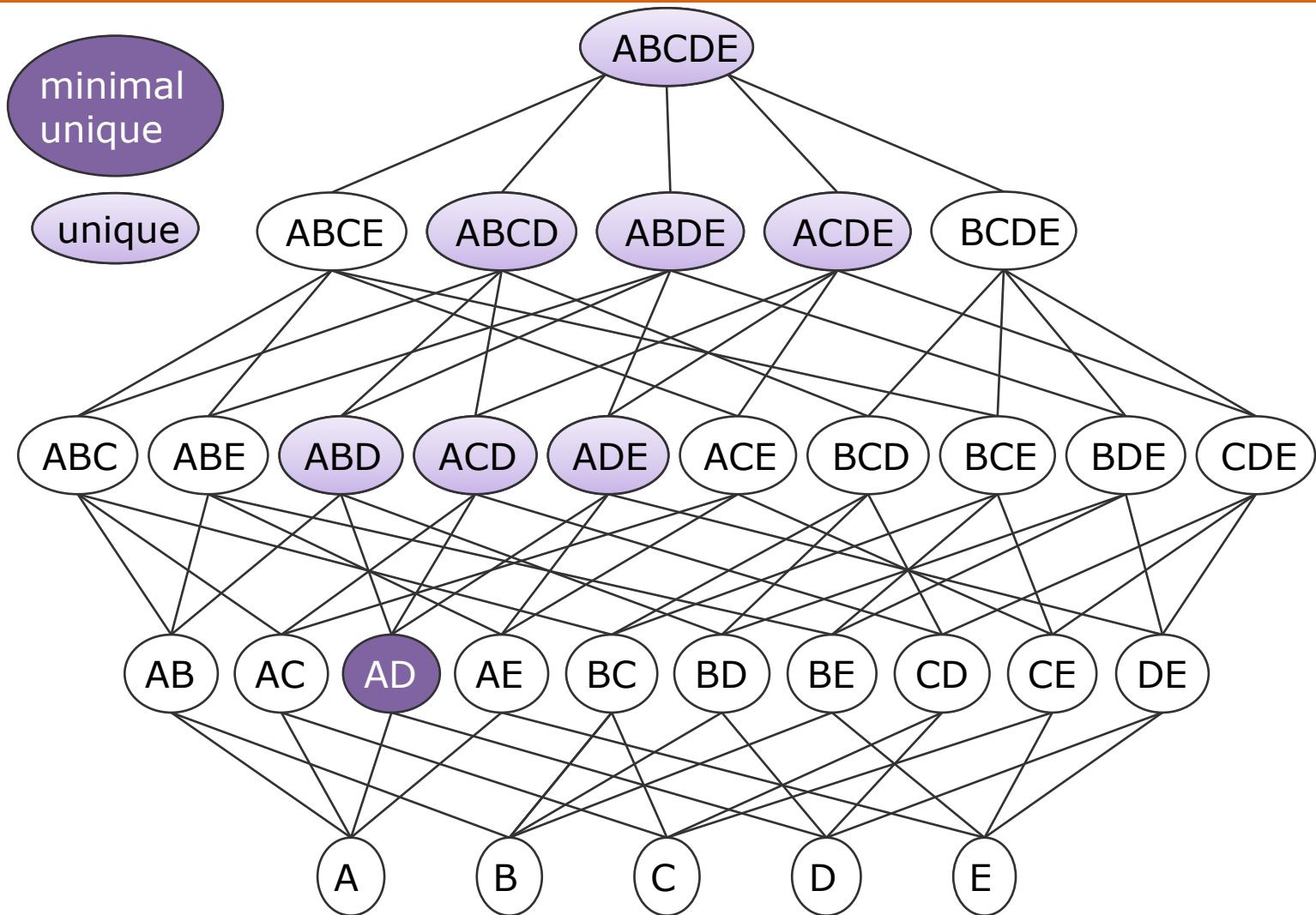


- Size of lattice: $2^n - 1$ (empty set not considered)
- Nodes at level 1: n
- Nodes at level n : 1
- Nodes at level k : $\binom{n}{k} = \frac{n!}{(n-k)!k!}$
- Largest level at $n/2$: $\binom{n}{n/2} = \frac{n!}{\left(\frac{n}{2}\right)^2}$

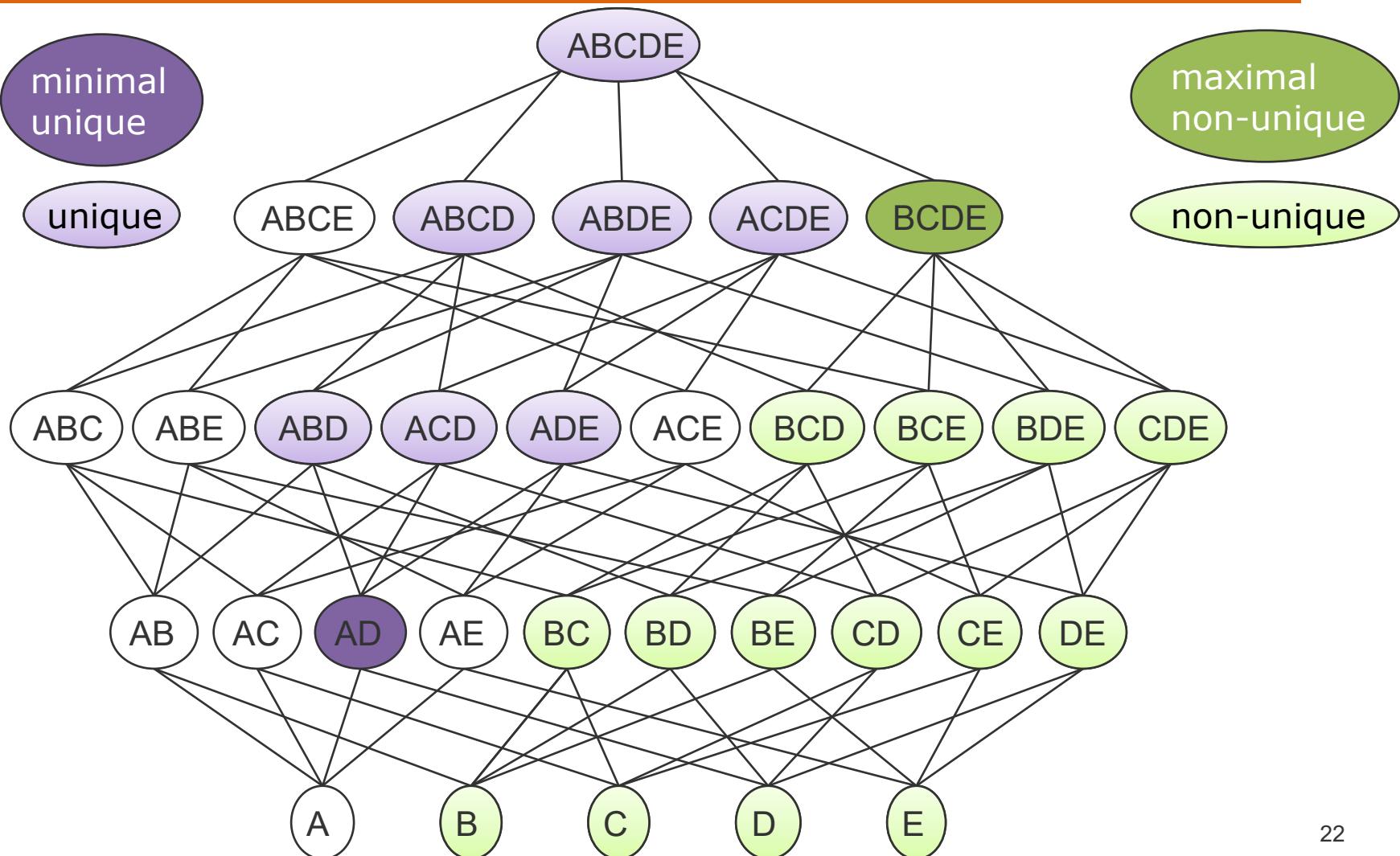
Candidate Set Growth for UCCs



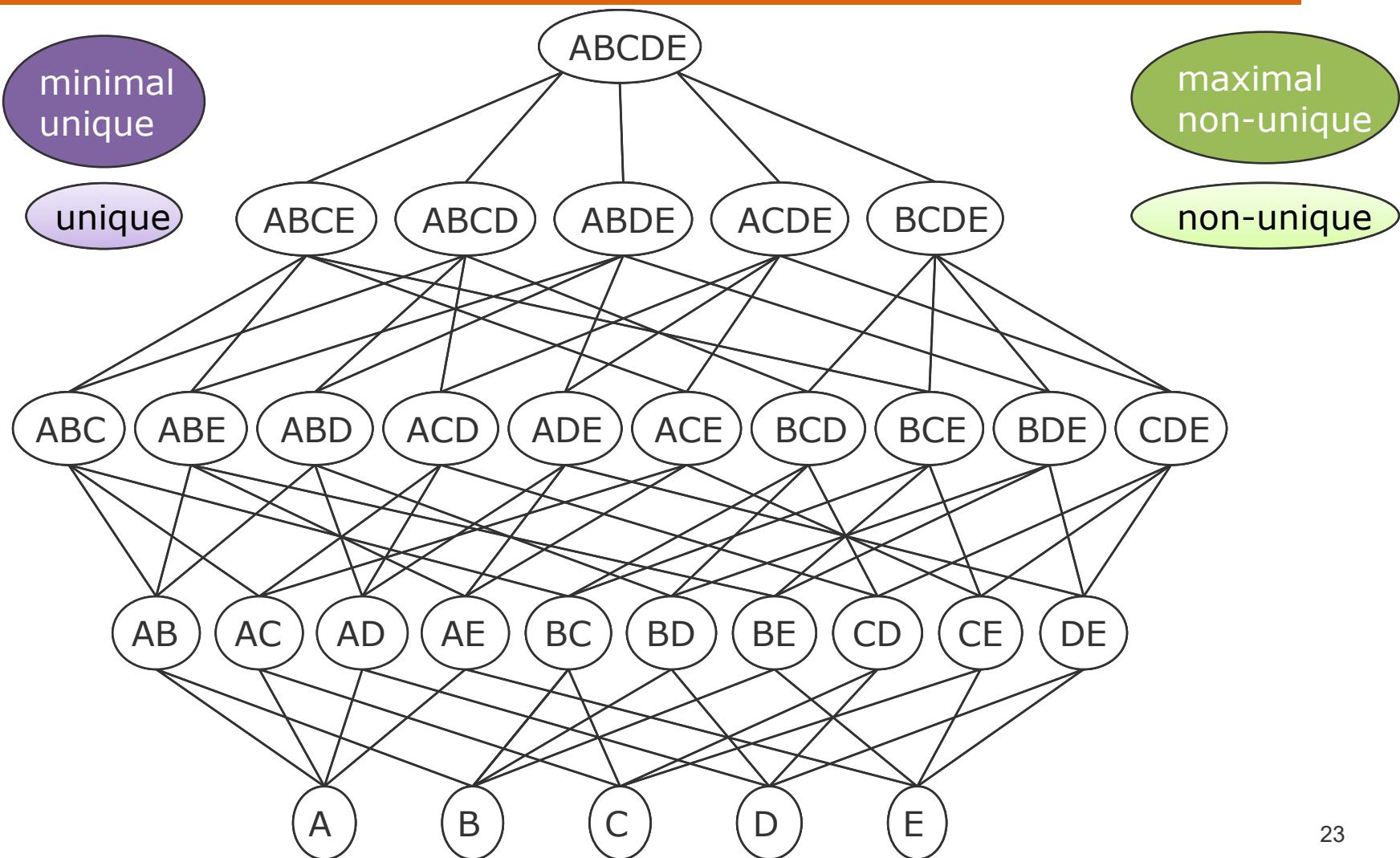
Pruning effect of a pair



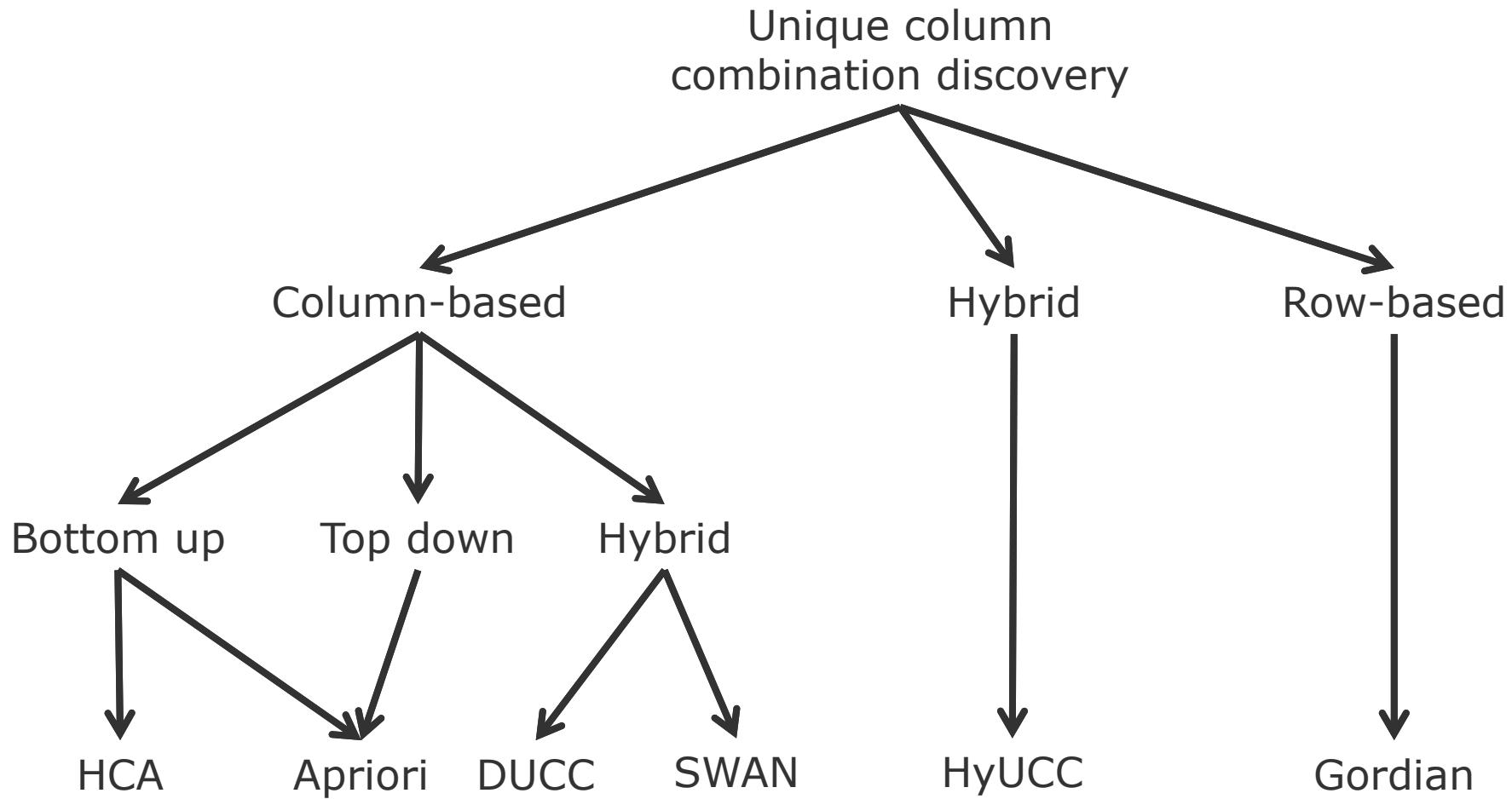
Pruning both ways



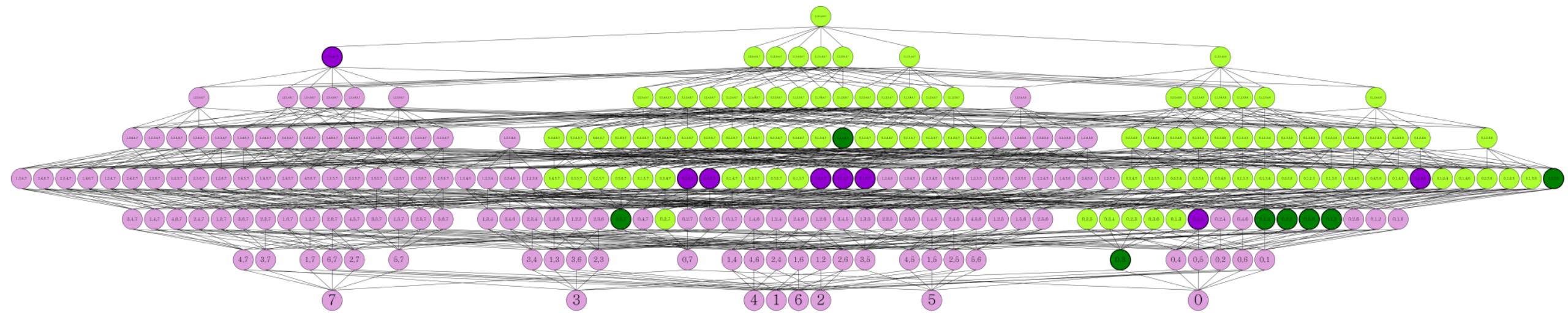
Apriori visualized



Discovery Algorithms

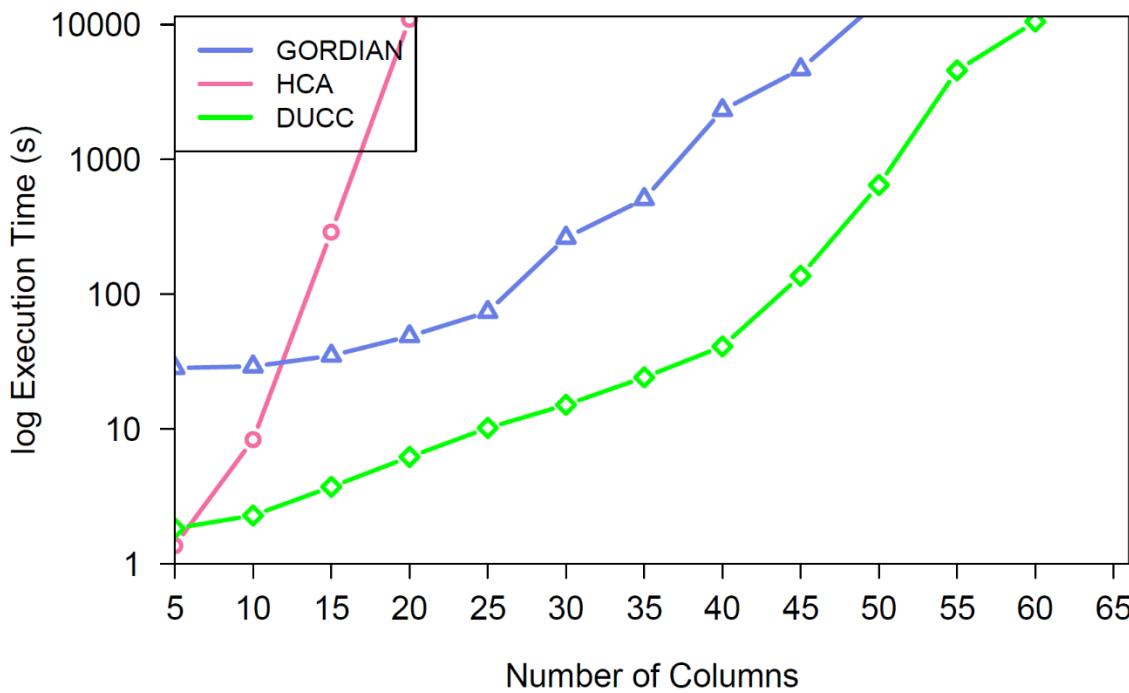


DUCC – Detecting Unique Column Combinations

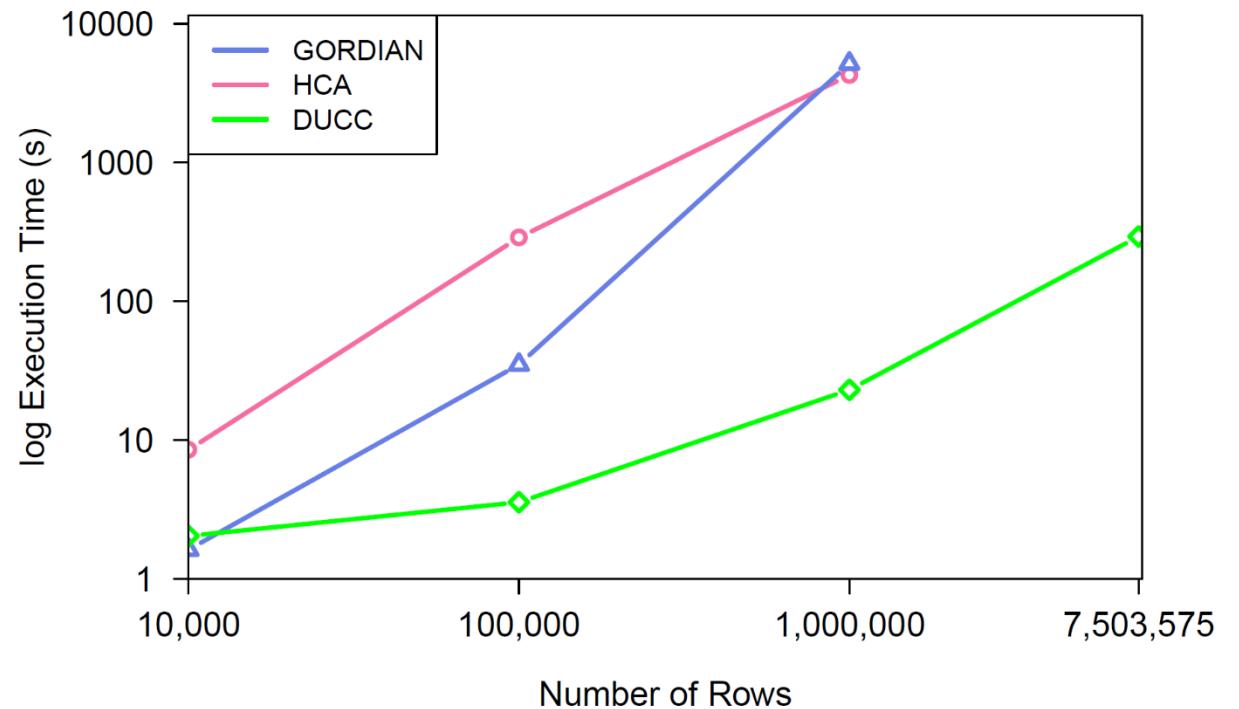


Scalability in the number of columns and rows

■ NCVoter data, 100k rows



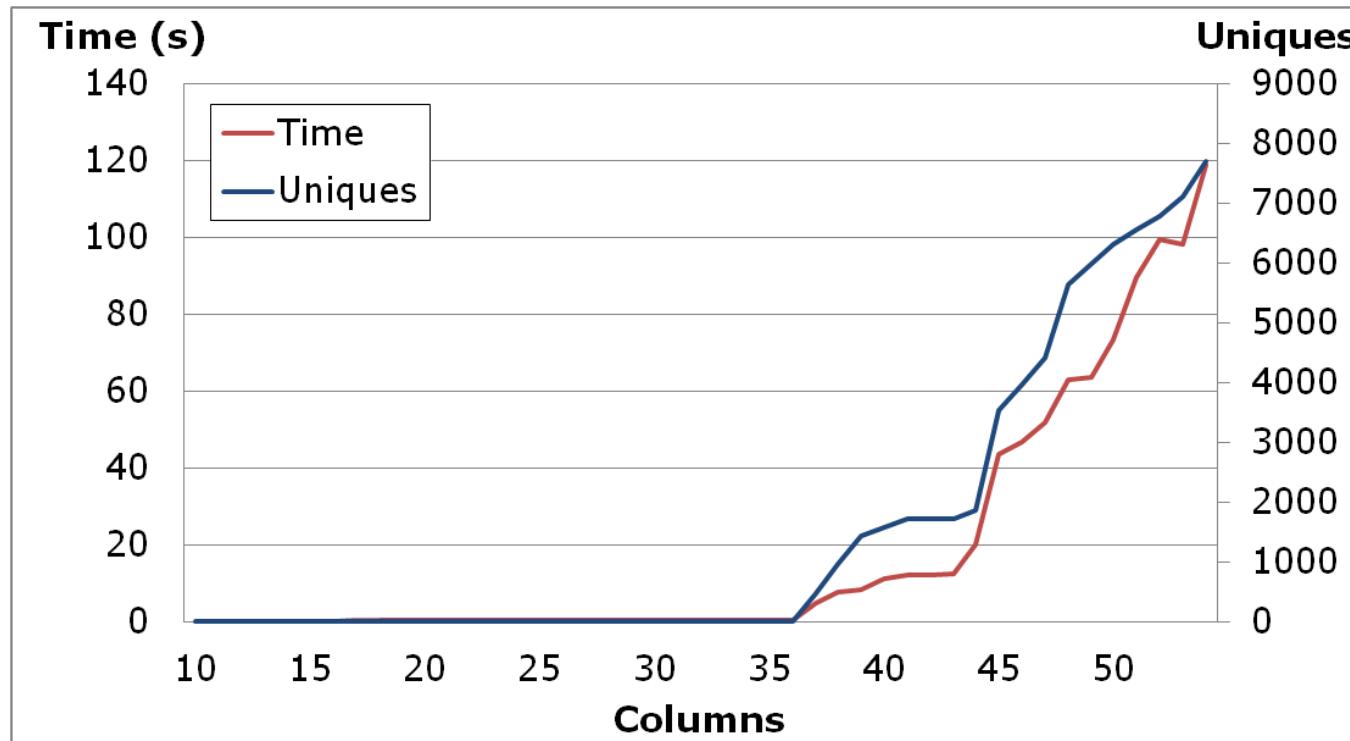
■ NCVoter, 15 columns



■ 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_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_abrv, precinct_desc, municipality_abrv, municipality_desc, ward_abrv, ward_desc, cong_dist_abrv, cong_dist_desc, super_court_abrv, super_court_desc, judic_dist_abrv, judic_dist_desc, nc_senate_abrv, nc_senate_desc, nc_house_abrv, nc_house_desc, county_commiss_abrv, county_commiss_desc, township_abrv, township_desc, school_dist_abrv, school_dist_desc, fire_dist_abrv, fire_dist_desc, water_dist_abrv, water_dist_desc, sewer_dist_abrv, sewer_dist_desc, sanit_dist_abrv, sanit_dist_desc, rescue_dist_abrv, rescue_dist_desc, munic_dist_abrv, munic_dist_desc, dist_1_abrv, dist_1_desc, dist_2_abrv, dist_2_desc, confidential_ind, age, vtd_abrv, 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:

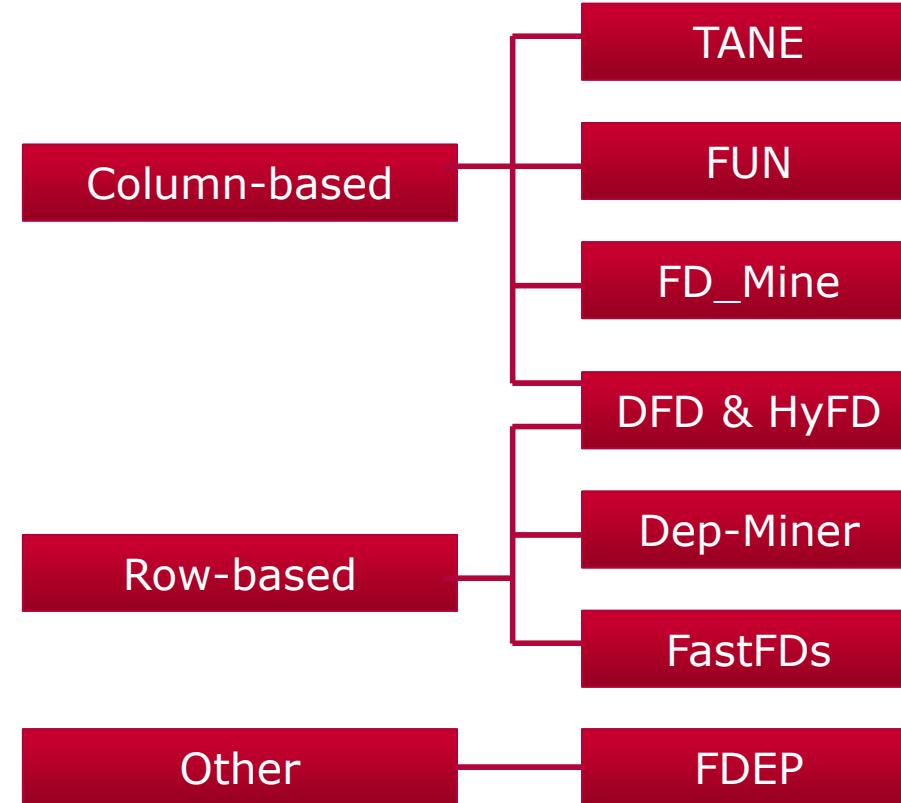
1. Person → Lineage
2. Person → Hair
3. Person → Religion
4. Lineage → Hair
5. Religion, Hair → Lineage
6. ...

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

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

Uses and Algorithms for FDs

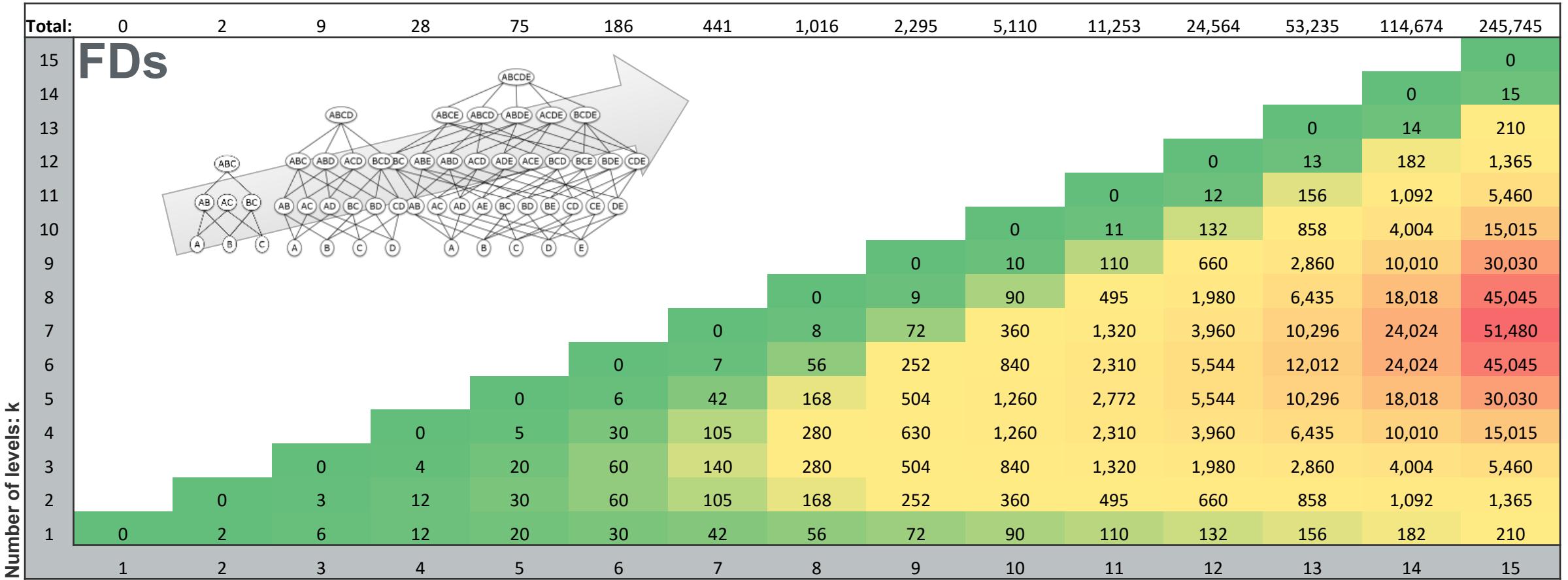
- Schema design
 - Normalization
 - Keys
- Data cleansing
- Query optimization
- Schema design and normalization
- Key discovery



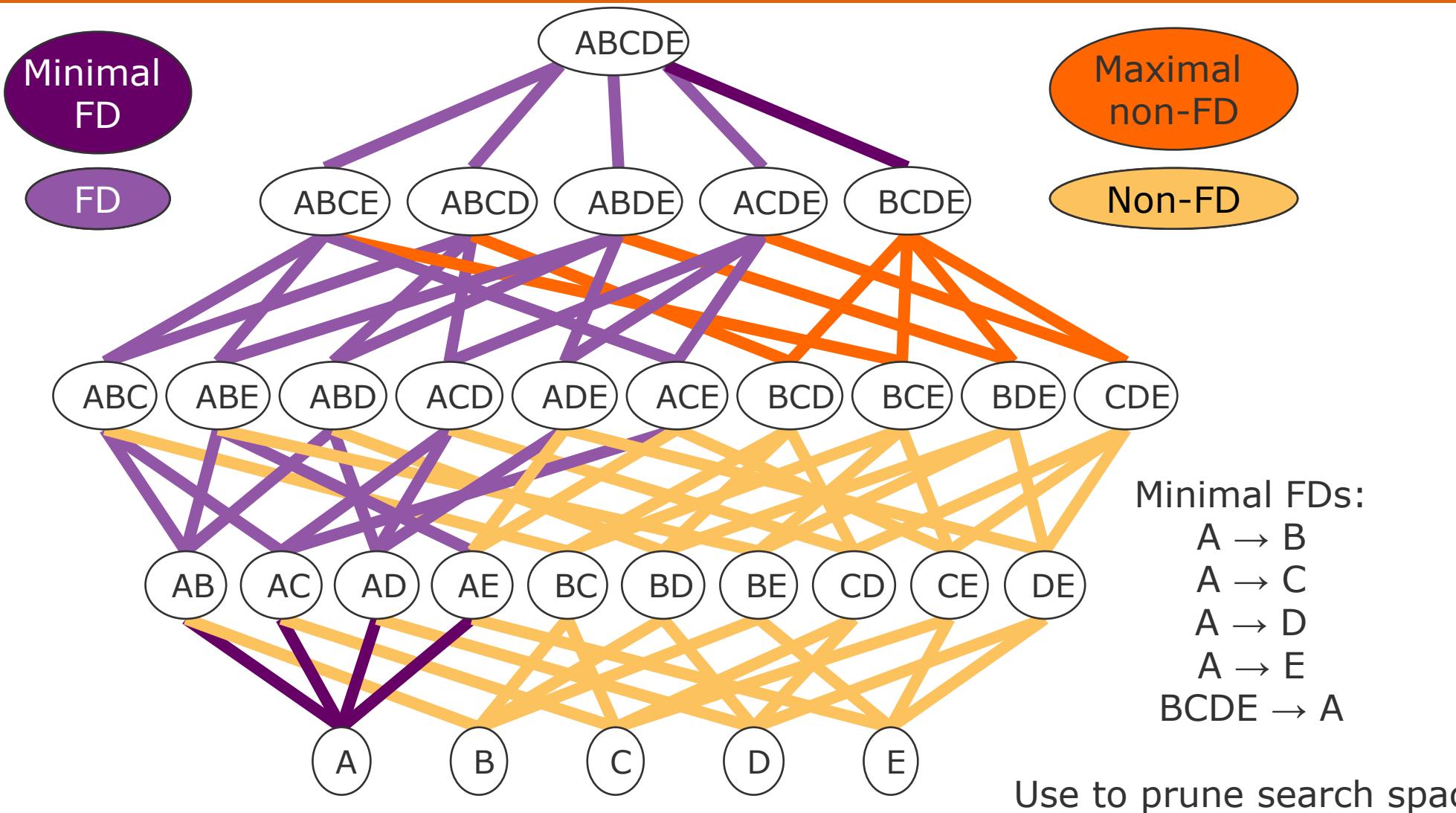
Naïve discovery approach

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

Candidate Set Growth for FDs



Again: Model in lattice – edges represent FDs



Row-based Discovery

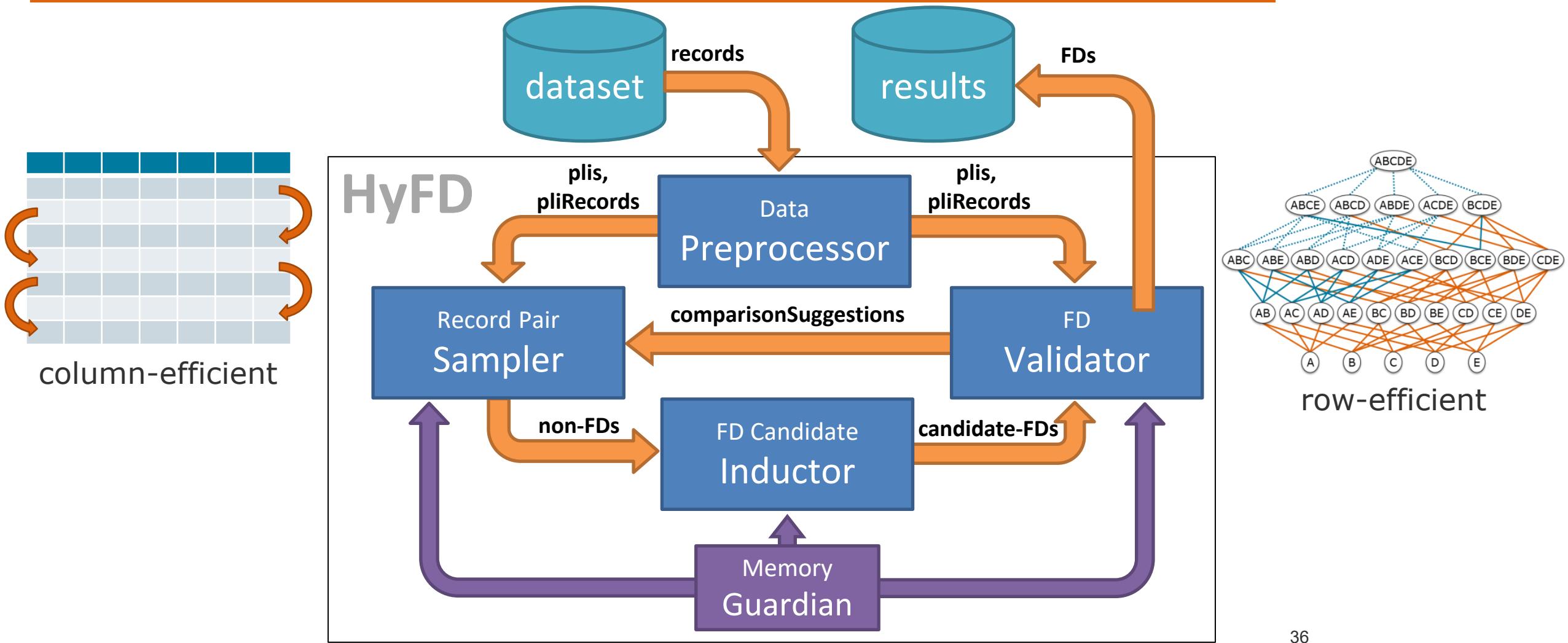
Name	Surname	Postcode	City	Mayor
Thomas	Miller	14482	Potsdam	Jakobs
Sarah	Miller	14482	Potsdam	Jakobs
Peter	Smith	60329	Frankfurt	Feldmann
Jasmine	Cone	01069	Dresden	Orosz
Thomas	Cone	14482	Potsdam	Jakobs
Mike	Moore	60329	Frankfurt	Feldmann

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



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

HyFD: Hybrid FD Discovery



Functional Dependencies: State of the Art

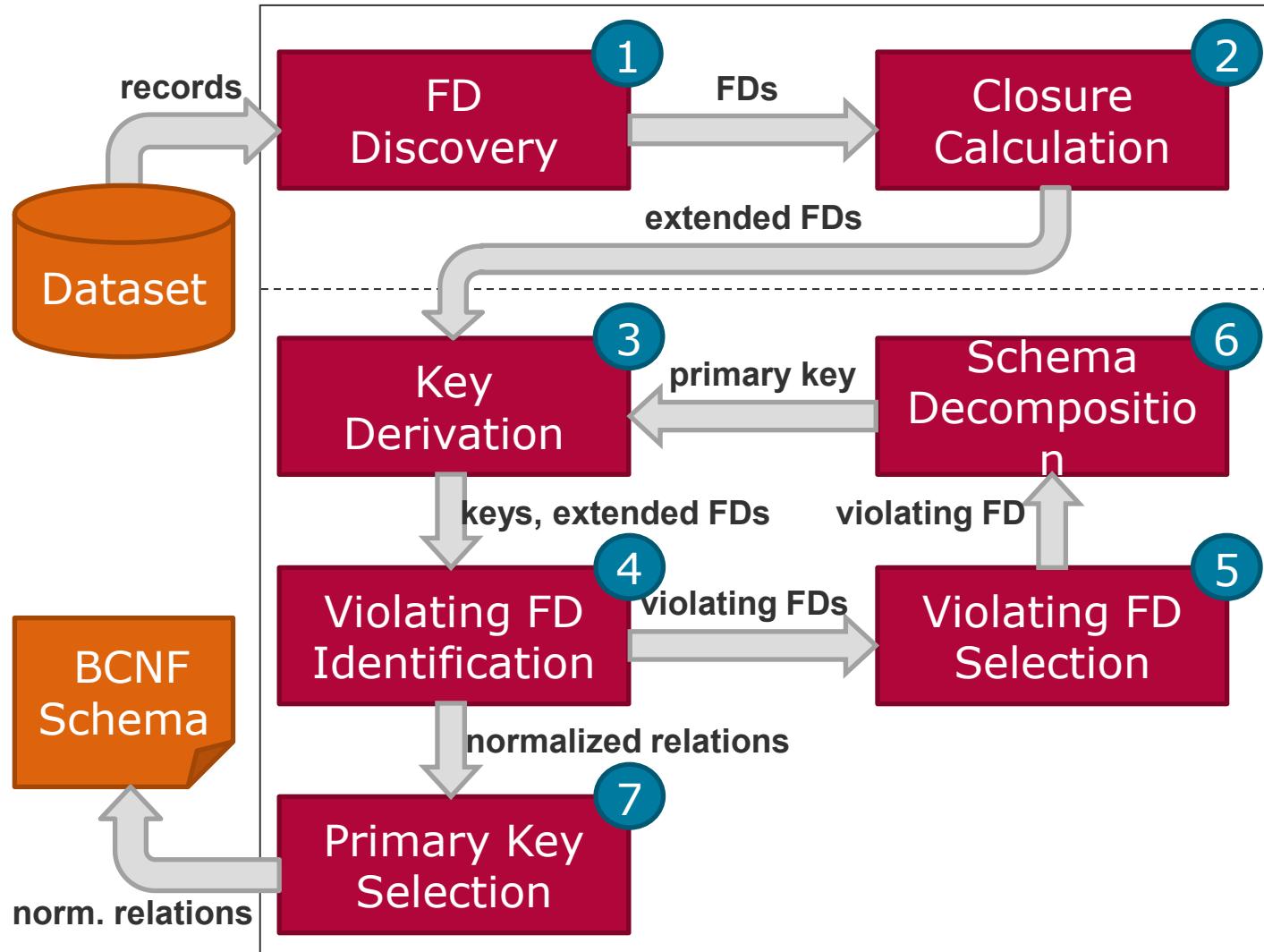
Dataset	Cols [#]	Rows [#]	Size [KB]	FDs [#]	TANE [12]	FUN [18]	FD_MINE [25]	DFD [1]	DEP-MINER [16]	FASTFDs [24]	FDEP [9]	HyFD
iris	5	150	5	4	1.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1
balance-scale	5	625	7	1	1.2	0.1	0.2	0.3	0.3	0.3	0.2	0.1
chess	7	28,056	519	1	2.9	1.1	3.8	1.0	174.6	164.2	125.5	0.2
abalone	9	4,177	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

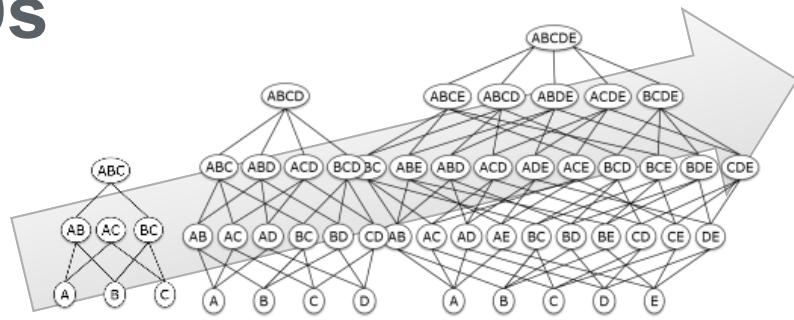
(<u>linenumber</u> , <u>extendedprice</u> , <u>discount</u> , <u>tax</u> , <u>returnflag</u> , <u>shipdate</u> , <u>commitdate</u> , <u>receiptdate</u> , <u>comment</u> , <u>orderkey</u> , <u>partkey</u>)	LINEITEM
→ (<u>linenumber</u> , <u>extendedprice</u> , <u>tax</u> , <u>commitdate</u> , <u>receiptdate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

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

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
Number of levels: k	INDs														
Number of attributes: m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	2	6	12	20	30	42	56	72	90	110	132	156	182	210
1	0	0	0	12	60	180	420	840	1,512	2,520	3,960	5,940	8,580	12,012	16,380
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	120	840	3,360	10,080	25,200	55,440	110,880	205,920
4	0	0	0	0	0	0	0	0	0	1,680	15,120	75,600	277,200	831,600	2,162,160
5	0	0	0	0	0	0	0	0	0	0	30,240	332,640	1,995,840	8,648,640	30,270,240
6	0	0	0	0	0	0	0	0	0	0	0	665,280	8,648,640	60,540,480	302,702,400
7	0	0	0	0	0	0	0	0	0	0	0	0	0	17,297,280	259,459,200
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



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

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

Pruning for INDs

- Subsequences of INDs are INDs

- $R[A_1, \dots, A_n] \subseteq S[B_1, \dots, B_n] \Rightarrow R[A_{i1}, \dots, A_{im}] \subseteq S[B_{i1}, \dots, B_{im}]$

- 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



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.attaccama.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

IBM. Information Server File Edit View Help 9.43.86.77

IA_OVERVIEW_PROJECT INVESTIGATE Column Analysis

Select Data Sources to Work With

EMPLOYEE

View Analysis Summary

View Details

i View the frequency distribution, data classes, properties, domain and completeness information, and formats for the column.

Select View:

- EMPNO
- FIRSTNAME
- MIDINIT
- LASTNAME
- WORKDEPT
- PHONE NO
- HIREDATE
- JOB
- EDLEVEL
- SEX
- BIRTHDATE
- SALARY**
- BONUS
- COMM
- SALUTATION
- EMERGENCY_CONTACT
- BLOOD_TYPE
- HAIR_COLOR

Properties

i Shows inferred and defined structural properties of a column. You can choose new property values to apply to a column.

Data Type

Defined: Inferred: Selected:

DECIMAL	DECIMAL	DECIMAL
---------	---------	---------

Inferred Summary

Inferred Data Type

Data Type	Count	Percent
DECIMAL	46	100

Length

Defined: Inferred: Selected:

9	9	9
---	---	---

Inferred Summary

Minimum: 8
Median: 8
Average: 8.0217
Maximum: 9
Range: 1

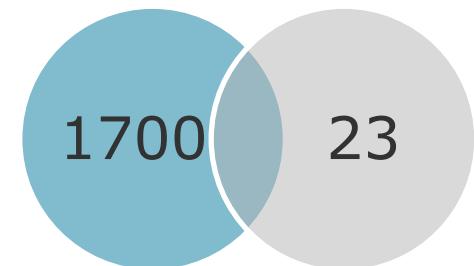
Reviewed

Close Rebuild Inferences Reference Tables Save

A large blue sphere icon is positioned next to the 'DECIMAL' value in the length section.

„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

IBM. Information Server File Edit View Help 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: Common Domain:  
8 100.0000% Yes

Base to Paired:

Common Data Values: Common Domain:  
8 16.6667% No

Common Domain :

Base Column Paired Column

40 8 1

51 Nose

# 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

# 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

| 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 $k$                      |
| ODEA               | 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    |

## Profiling New Types of Data

---

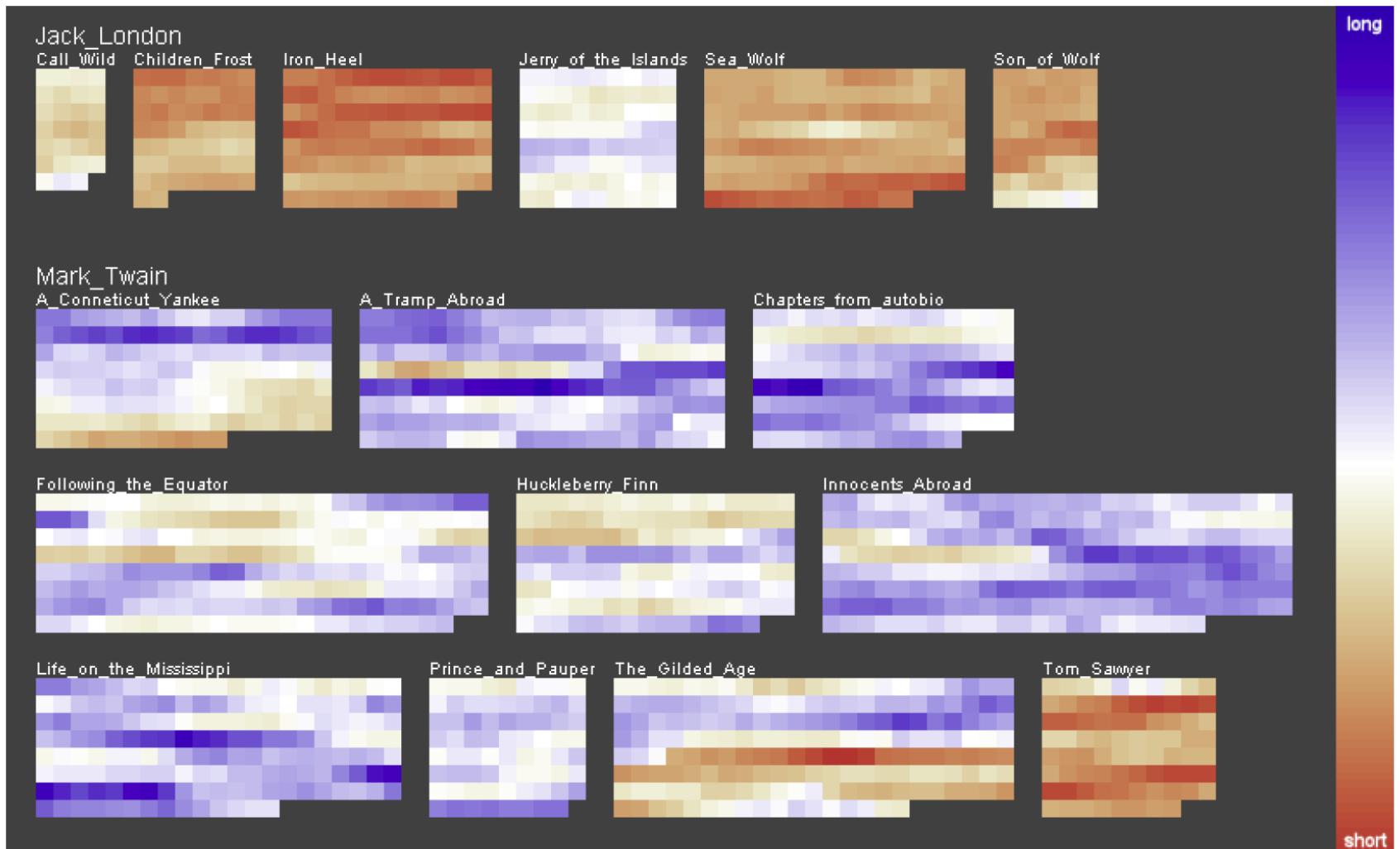
- Traditional data profiling: Single table or multiple tables
- More and more data in other models
  - XML / nested relational / JSON
  - RDF triples
  - Textual data: Blogs, Tweets, News
  - Multimedia data
- Different models offer new dimensions to profile
  - XML: Nestedness, measures at different 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

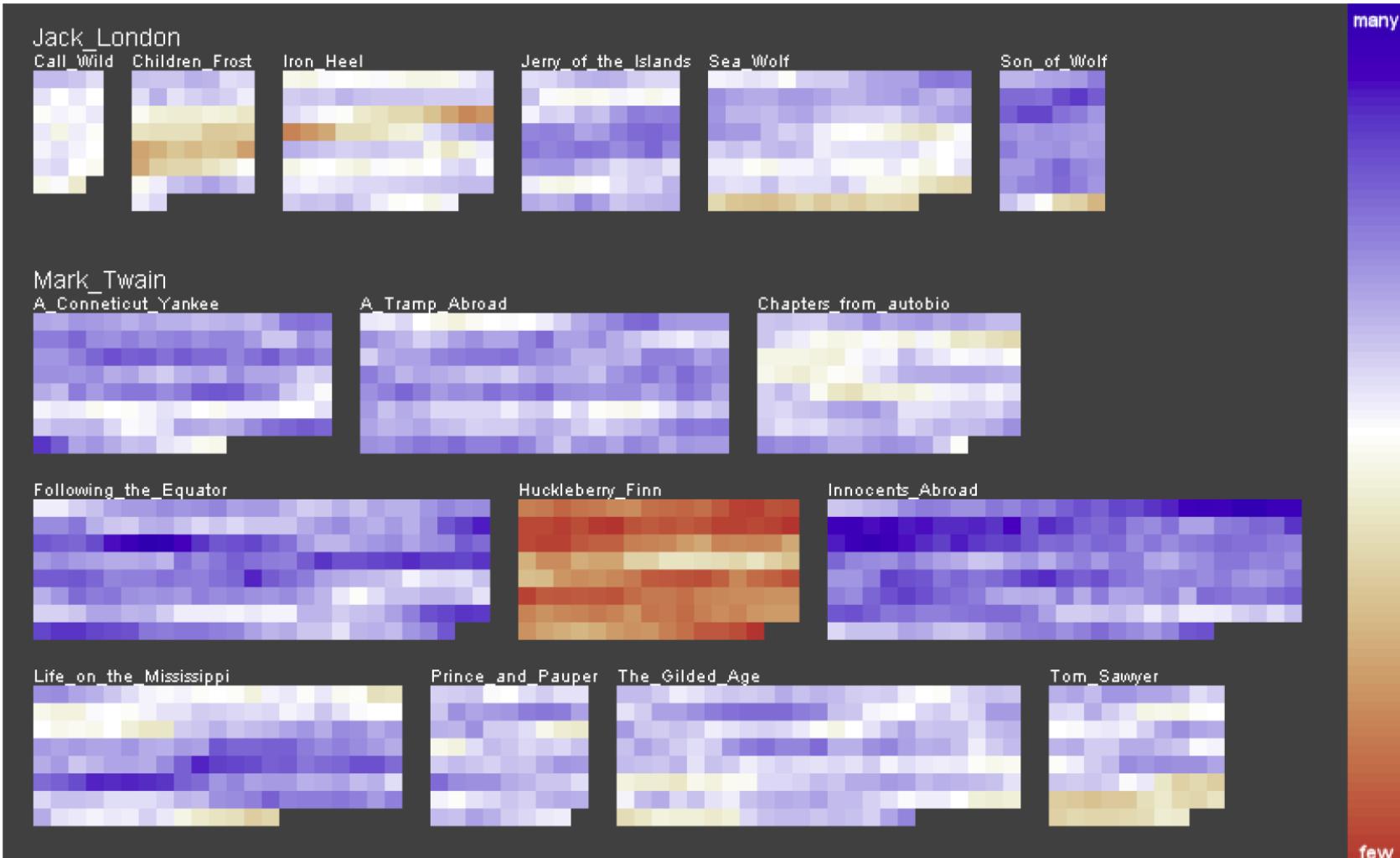
---

- 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)

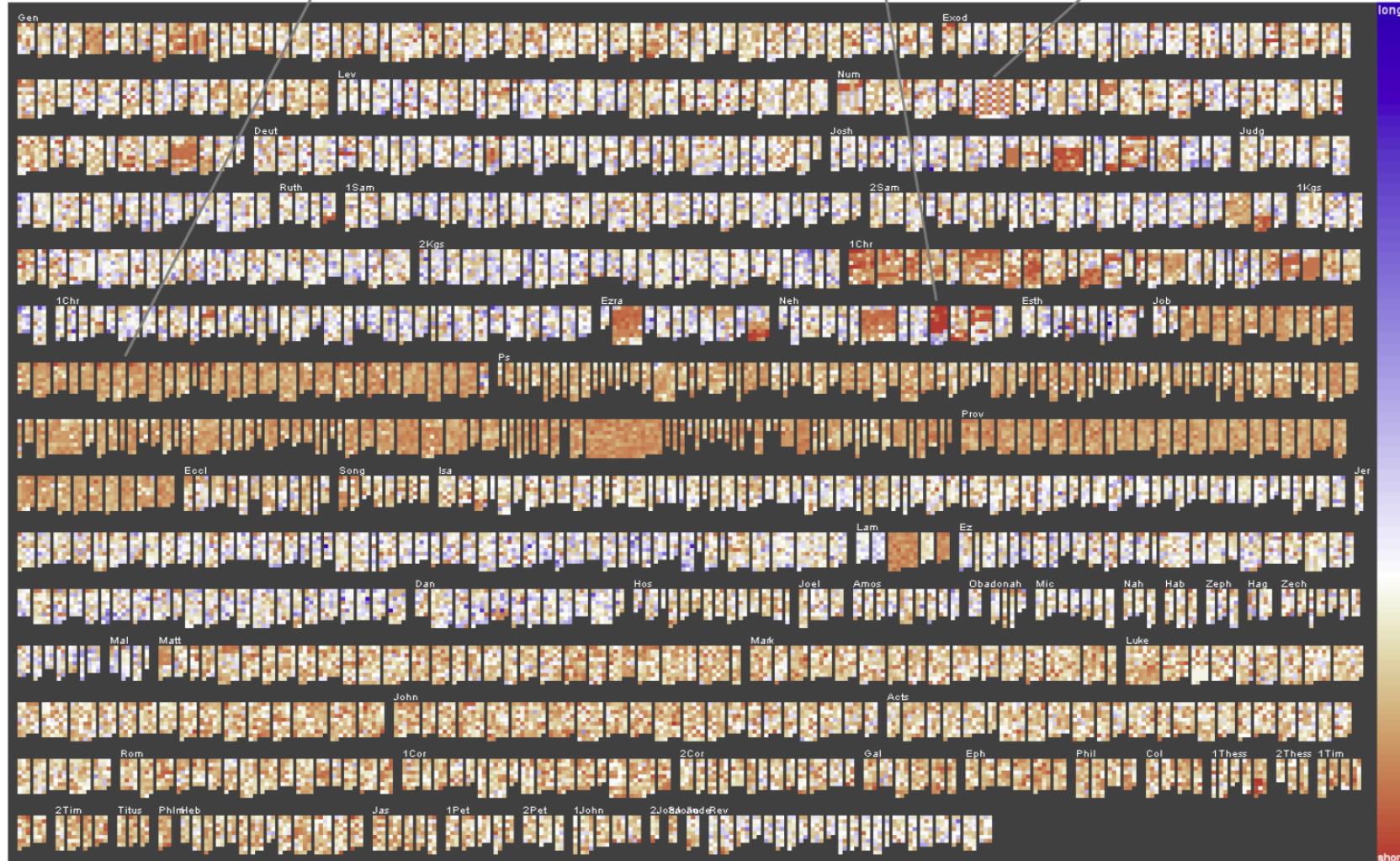
## Average sentence length



# Hapax Legomena



# Verse length



60

## 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

# 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 Frohnhofer, 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 Quiane-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

