University of Waterloo

Hasso Plattner Institut

Digital Engineering • Universität Potsdam

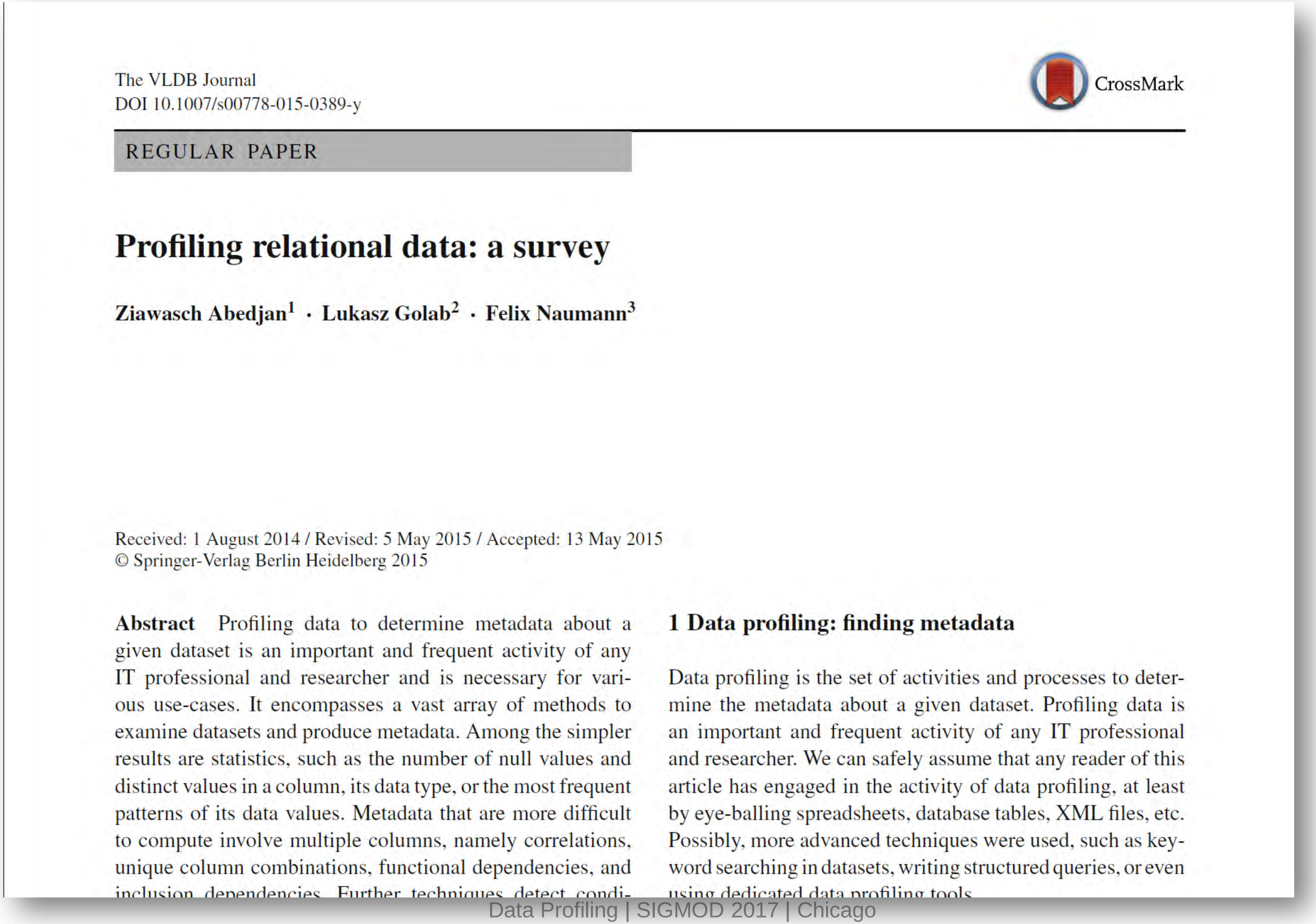
# Data Profiling

## a SIGMOD 2017 Tutorial

Ziawasch Abedjan (TU Berlin)
Lukasz Golab (University of Waterloo)
Felix Naumann (HPI)

> “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.]



The VLDB Journal
DOI 10.1007/s00778-015-0389-y

REGULAR PAPER

# Profiling relational data: a survey

Ziawasch Abedjan[1] · Lukasz Golab[2] · Felix Naumann[3]

Received: 1 August 2014 / Revised: 5 May 2015 / Accepted: 13 May 2015

**Abstract** Profiling data to determine metadata about a given dataset is an important and frequent activity of any IT professional and researcher and is necessary for various use-cases. It encompasses a vast array of methods to examine datasets and produce metadata. Among the simpler results are statistics, such as the number of null values and distinct values in a column, its data type, or the most frequent patterns of its data values. Metadata that are more difficult to compute involve multiple columns, namely correlations, unique column combinations, functional dependencies, and inclusion dependencies. Further techniques detect condi-

## 1 Data profiling: finding metadata

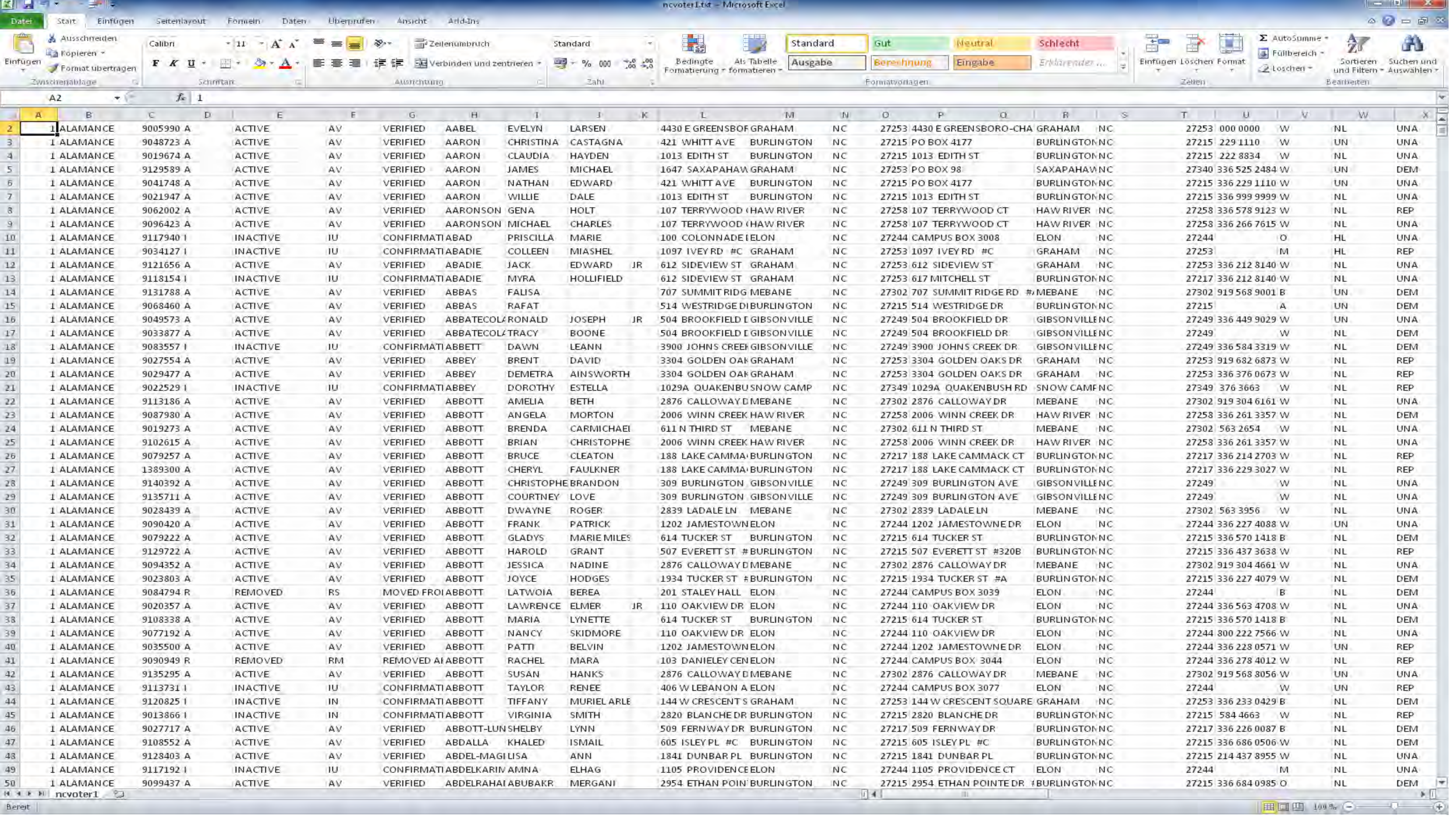
Data profiling is the set of activities and processes to determine the metadata about a given dataset. Profiling data is an important and frequent activity of any IT professional and researcher. We can safely assume that any reader of this article has engaged in the activity of data profiling, at least by eye-balling spreadsheets, database tables, XML files, etc. Possibly, more advanced techniques were used, such as keyword searching in datasets, writing structured queries, or even using dedicated data profiling tools.

# Tutorial Overview

- Motivation
    - Task classification
    - Use cases
- Tools
    - Research and industry
    - Shortcomings
- Single and Multiple Column Analysis
    - Cardinalities and datatypes
    - Co-occurrences and summaries
- Dependencies
    - UCCs, FDs, ODs, INDs
    - and their discovery algorithms
- Outlook
    - Functionality
    - Semantics

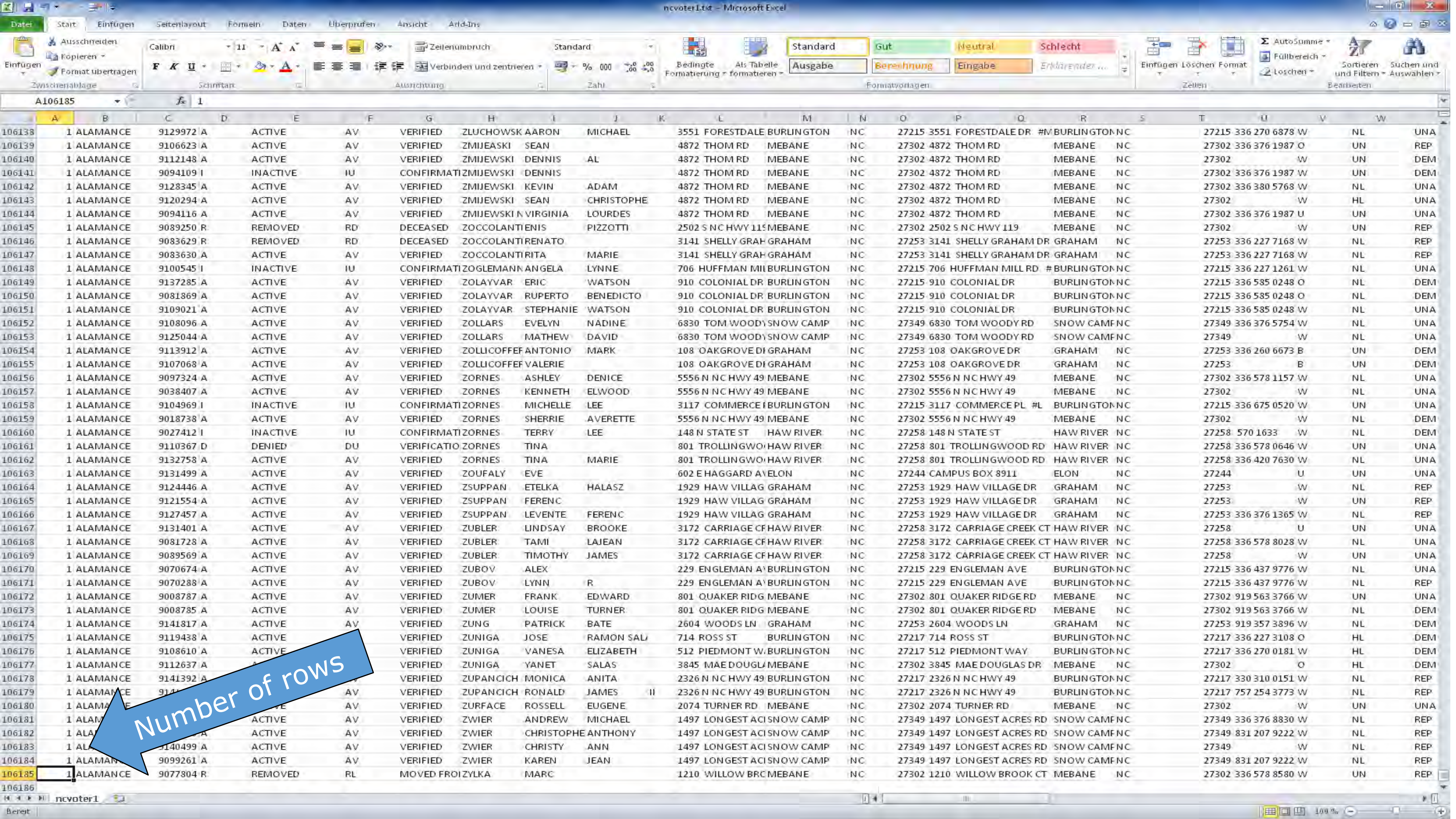
## References

1. Ziawasch Abedjan, Toni Grütze, Anja Jentzsch, and Felix Naumann. Mining and profiling RDF data with ProLOD++. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014. Demo.
2. Ziawasch Abedjan, Johannes Lorey, and Felix Naumann. Reconciling ontologies and the web of data. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 1532–1536, 2012.
3. Ziawasch Abedjan and Felix Naumann. Advancing the discovery of unique column combinations. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 1565–1570, 2011.
4. Ziawasch Abedjan and Felix Naumann. Synonym analysis for predicate expansion. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, pages 140–154, 2013.
5. Ziawasch Abedjan, Jorge-Arnulfo Quiané-Ruiz, and Felix Naumann. Detecting unique column combinations on dynamic data. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1036–1047, 2014.
6. Ziawasch Abedjan, Patrick Schulze, and Felix Naumann. DFD: Efficient functional dependency discovery. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 949–958, 2014.
7. Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwar Dayal, Michael Franklin, Johannes Gehrke, Laura Haas, Alon Halevy, Jiawei Han, H. V. Jagadish, Alexandros Labrinidis, Sam Madden, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Kenneth Ross, Cyrus Shahabi, Dan Suciu, Shiv Vaithyanathan, and Jennifer Widom. Challenges and opportunities with Big Data. Technical report, Computing Community Consortium, `http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf`, 2012.
8. Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 487–499, 1994.
9. Periklis Andritsos, Renée J. Miller, and Panayiotis Tsaparas. Information-theoretic tools for mining database structure from large data sets. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 731–742, 2004.
10. Marcelo Arenas, Jonny Daenen, Frank Neven, Martin Ugarte, Jan Van den Bussche, and Stijn Vansummeren. Discovering XSD keys from XML data. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 61–72, 2013.
11. Morton M. Astrahan, Mario Schkolnick, and Whang Kyu-Young. Approximating the number of unique values of an attribute without sorting. *Information Systems*, 12(1):11 – 15, 1987.
12. Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. LODStats – an extensible framework for high-performance dataset analytics. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 353–362, 2012.
13. Jana Bauckmann, Ziawasch Abedjan, Heiko Müller, Ulf Leser, and Felix Naumann. Discovering conditional inclusion dependencies. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 2094–2098, 2012.
14. Jana Bauckmann, Ulf Leser, Felix Naumann, and Veronique Tietz. Efficiently detecting inclusion dependencies. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1448–1450, 2007.
15. Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551 – 572, 1938.
16. Laure Berti-Equille, Tamraparni Dasu, and Divesh Srivastava. Discovery of complex glitch patterns: A novel approach to quantitative data cleaning. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 733–744, 2011.
17. Geert Jan Bex, Frank Neven, and Stijn Vansummeren. Inferring XML schema definitions from XML data. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 998–1009, 2007.
18. Christoph Böhm, Johannes Lorey, and Felix Naumann. Creating voiD descriptions for web-scale data. *Journal of Web Semantics*, 9(3):339–345, 2011.
19. Loreto Bravo, Wenfei Fan, and Shuai Ma. Extending dependencies with conditions. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 243–254, 2007.
20. Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.
21. Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. *SIGMOD Record*, 26(2):265–276, 1997.
22. Peter Buneman, Susan B. Davidson, Wenfei Fan, Carmem S. Hara, and Wang Chiew Tan. Reasoning about keys for XML. *Information Systems*, 28(8):1037–1063, 2003.
23. Varun Chandola and Vipin Kumar. Summarization – compressing data into an informative representation. *Knowledge and Information Systems*, 12(3):355–378, 2007.
24. Fei Chiang and Renée J. Miller. Discovering data quality rules. *Proceedings of the VLDB Endowment (PVLDB)*, 1:1166–1177, 2008.
25. Roger H.L. Chiang, Chua Eng Huang Cecil, and Ee-Peng Lim. Linear correlation discovery in databases: A data mining approach. *Data and Knowledge Engineering (DKE)*, 53(3):311–337, June 2005.
26. Byron Choi. What are real DTDs like? In *Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB)*, pages 43–48, 2002.
27. Peter Christen. *Data Matching*. Springer Verlag, Berlin – Heidelberg – New York, 2012.
28. Xu Chu, Ihab Ilyas, Paolo Papotti, and Yin Ye. RuleMiner: Data quality rules discovery. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1222–1225, 2014.
29. Xu Chu, Ihab F. Ilyas, and Paolo Papotti. Discovering denial constraints. *Proceedings of the VLDB Endowment (PVLDB)*, 6(13):1498–1509, 2013.
30. Gao Cong, Wenfei Fan, Floris Geerts, Xibei Jia, and Shuai Ma. Improving data quality: Consistency and accuracy. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 315–326, 2007.
31. Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends in Databases*, 4(13):1–294, 2011.
32. Graham Cormode, Lukasz Golab, Korn Flip, Andrew McGregor, Divesh Srivastava, and Xi Zhang. Estimating the confidence of conditional functional dependencies. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 469–482, 2009.
33. Graham Cormode, Flip Korn, S. Muthukrishnan, and Divesh Srivastava. Space- and time-efficient deterministic algorithms for biased quantiles over data streams. In *Proceedings of the Symposium on Principles of Database Systems (PODS)*, pages 263–272, 2006.
34. Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. NADEEF: A commodity data cleaning system. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 541–552, 2013.
35. Amitabha Das, Wee-Keong Ng, and Yew-Kwong Woon. Rapid association rule mining. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 474–481, 2001.
36. Tamraparni Dasu and Theodore Johnson. Hunting of the snark: Finding data glitches using data mining methods. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 89–98, 1999.
37. Tamraparni Dasu, Theodore Johnson, and Amit Marathe. Database exploration using database dynamics. *IEEE Data Engineering Bulletin*, 29(2):43–59, 2006.
38. Tamraparni Dasu, Theodore Johnson, S. Muthukrishnan, and Vladislav Shkapenyuk. Mining database structure; or, how to build a data quality browser. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 240–251, 2002.
39. Tamraparni Dasu and Ji Meng Loh. Statistical distortion: Consequences of data cleaning. *Proceedings of the VLDB Endowment (PVLDB)*, 5(11):1674–1683, 2012.
40. Tamraparni Dasu, Ji Meng Loh, and Divesh Srivastava. Empirical glitch explanations. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 572–581, 2014.
41. Amol Deshpande, Minos Garofalakis, and Rajeev Rastogi. Independence is good: Dependency-based histogram synopses for high-dimensional data. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 199–210, 2001.
42. Thierno Diallo, Noel Novelli, and Jean-Marc Petit. Discovering (frequent) constant conditional functional dependencies. *International Journal of Data Mining, Modelling and Management*, 4(3):205–223, 2012.
43. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, and Xiaowei Xu. Incremental clustering for mining in a data warehousing environment. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 323–333, 1998.
44. Jerome Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer Verlag, Berlin – Heidelberg – New York, 2nd edition, 2013.
45. Wenfei Fan, Floris Geerts, and Xibei Jia. Semandaq: A data quality system based on conditional functional dependencies. *Proceedings of the VLDB Endowment (PVLDB)*, 1(2):1460–1463, 2008.
46. Wenfei Fan, Floris Geerts, Xibei Jia, and Anastasios Kementsietsidis. Conditional functional dependencies for capturing data inconsistencies. *ACM Transactions on Database Systems (TODS)*, 33(2):1–48, 2008.
47. Wenfei Fan, Floris Geerts, Jianzhong Li, and Ming Xiong. Discovering conditional functional dependencies. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 23(4):683–698, 2011.
48. Wenfei Fan, Floris Geerts, Shuai Ma, and Heiko Müller. Detecting inconsistencies in distributed data. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 64–75, 2010.
49. Wenfei Fan, Xibei Jia, Jianzhong Li, and Shuai Ma. Reasoning about record matching rules. *Proceedings of the VLDB Endowment (PVLDB)*, 2(1):407–418, 2009.
50. Wenfei Fan, Jianzhong Li, Nan Tang, and Wenyuan Yu. Incremental detection of inconsistencies in distributed data. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 318–329, 2012.
51. Henning Fernau. Algorithms for learning regular expressions from positive data. *Information and Computation*, 207(4):521–541, 2009.
52. Peter A. Flach and Iztok Savnik. Database dependency discovery: a machine learning approach. *AI Communications*, 12(3):139–160, 1999.
53. Sumit Ganguly. Counting distinct items over update streams. *Theoretical Computer Science*, 378(3):211–222, 2007.
54. Minos Garofalakis, Daniel Keren, and Vasilis Samoladas. Sketch-based geometric monitoring of distributed stream queries. *Proceedings of the VLDB Endowment (PVLDB)*, 6(10), 2013.
55. Chris Giannella and Catherine Wyss. Finding minimal keys in a relation instance, 1999. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.41.7086`.
56. Seymour Ginsburg and Richard Hull. Order dependency in the relational model. *Theoretical Computer Science*, 26:149–195, 1983.
57. Lukasz Golab, Howard Karloff, Flip Korn, Avishek Saha, and Divesh Srivastava. Sequential dependencies. *Proceedings of the VLDB Endowment (PVLDB)*, 2(1):574–585, 2009.
58. Lukasz Golab, Howard Karloff, Flip Korn, and Divesh Srivastava. Data Auditor: Exploring data quality and semantics using pattern tableaux. *Proceedings of the VLDB Endowment (PVLDB)*, 3(1-2):1641–1644, 2010.
59. Lukasz Golab, Howard Karloff, Flip Korn, Divesh Srivastava, and Bei Yu. On generating near-optimal tableaux for conditional functional dependencies. *Proceedings of the VLDB Endowment (PVLDB)*, 1(1):376–390, 2008.
60. Lukasz Golab, Flip Korn, and Divesh Srivastava. Discovering pattern tableaux for data quality analysis: a case study. In *Proceedings of the International Workshop on Quality in Databases (QDB)*, pages 47–53, 2011.
61. Lukasz Golab, Flip Korn, and Divesh Srivastava. Efficient and effective analysis of data quality using pattern tableaux. *IEEE Data Engineering Bulletin*, 34(3):26–33, 2011.
62. Gösta Grahne and Jianfei Zhu. Discovering approximate keys in XML data. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 453–460, 2002.
63. Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data Cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. *Data Mining and Knowledge Discovery*, 1(1):29–53, 1997.
64. Dimitrios Gunopulos, Roni Khardon, Heikki Mannila, and Ram Sewak Sharma. Discovering All Most Specific Sentences. *ACM Transactions on Database Systems (TODS)*, 28:140–174, 2003.
65. Peter J. Haas, Jeffrey F. Naughton, S. Seshadri, and Lynne Stokes. Sampling-based estimation of the number of distinct values of an attribute. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 311–322, 1995.
66. Jean-Luc Hainaut, Jean Henrard, Vincent Englebert, Didier Roland, and Jean-Marc Hick. Database reverse engineering. In *Encyclopedia of Database Systems*, pages 723–728. Springer, Heidelberg, 2009.
67. Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *SIGMOD Record*, 29(2):1–12, 2000.
68. Pat Hanrahan. Analytic database technology for a new kind of user – the data enthusiast (keynote). In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 577–578, 2012.
69. Jan Hegewald, Felix Naumann, and Melanie Weis. XStruct: Efficient schema extraction from multiple and large XML databases. In *Proceedings of the International Workshop on Database Interoperability (InterDB)*, 2006.
70. Arvid Heise, Jorge-Arnulfo Quiané-Ruiz, Ziawasch Abedjan, Anja Jentzsch, and Felix Naumann. Scalable Discovery of Unique Column Combinations. *Proceedings of the VLDB Endowment (PVLDB)*, 7(4):301 – 312, 2013.
71. Joseph M. Hellerstein, Christopher Ré, Florian Schoppmann, Daisy Zhe Wang, Eugene Fratkin, Aleksander Gorajek, Kee Siong Ng, Caleb Welton, Xixuan Feng, Kun Li, and Arun Kumar. The MADlib analytics library or MAD skills, the SQL. *Proceedings of the VLDB Endowment (PVLDB)*, 5(12):1700–1711, 2012.
72. Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64, 2000.
73. David I. Holmes. Authorship attribution. *Computers and the Humanities*, 28:87–106, 1994.
74. Ming Hua and Jian Pei. Cleaning disguised missing data: A heuristic approach. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 950–958, 2007.
75. Ykä Huhtala, Juha Kärkkäinen, Pasi Porkka, and Hannu Toivonen. TANE: An efficient algorithm for discovering functional and approximate dependencies. *Computer Journal*, 42(2):100–111, 1999.
76. Ihab F. Ilyas, Volker Markl, Peter J. Haas, Paul Brown, and Ashraf Aboulnaga. CORDS: Automatic discovery of correlations and soft functional dependencies. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 647–658, 2004.
77. Yannis Ioannidis. The history of histograms (abridged). In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 19–30, 2003.
78. Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3), 1999.
79. Theodore Johnson. *Encyclopedia of Database Systems*, chapter Data Profiling, pages 604–608. Springer, Heidelberg, 2009.
80. Holger Kache, Wook-Shin Han, Volker Markl, Vijayshankar Raman, and Stephan Ewen. POP/FED: Progressive query optimization for federated queries in DB2. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 1175–1178, 2006.
81. Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of Advanced Visual Interfaces (AVI)*, pages 547–554, 2012.
82. Jaewoo Kang and Jeffrey F. Naughton. On schema matching with opaque column names and data values. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 205–216, 2003.
83. Daniel A. Keim and Daniela Oelke. Literature fingerprinting: A new method for visual literary analysis. In *Proceedings of Visual Analytics Science and Technology (VAST)*, pages 115 –122, 2007.
84. Nodira Khoussainova, Magdalena Balazinska, and Dan Suciu. Towards correcting input data errors probabilistically using integrity constraints. In *Proceedings of the ACM International Workshop on Data Engineering for Wireless and Mobile Access (MobiDE)*, pages 43–50, 2006.
85. Jyrki Kivinen and Heikki Mannila. Approximate inference of functional dependencies from relations. In *Proceedings of the International Conference on Database Theory (ICDT)*, pages 129–149, 1995.
86. Henning Koehler, Uwe Leck, Sebastian Link, and Henri Prade. Logical foundations of possibilistic keys. In *Logics in Artificial Intelligence*, volume 8761 of *Lecture Notes in Computer Science*, pages 181–195. Springer International Publishing, 2014.
87. Andreas Koeller and Elke A. Rundensteiner. Heuristic strategies for the discovery of inclusion dependencies and other patterns. *Journal on Data Semantics V*, pages 185–210, 2006.
88. Flip Korn, Barna Saha, Divesh Srivastava, and Shanshan Ying. On repairing structural problems in semi-structured data. *Proceedings of the VLDB Endowment (PVLDB)*, 6(9):601–612, 2013.
89. Nick Koudas, Avishek Saha, Divesh Srivastava, and Suresh Venkatasubramanian. Metric functional dependencies. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1275–1278, 2009.
90. Douglas Laney. 3D data management: Controlling data volume, velocity and variety. Technical report, Gartner, 2001.
91. Jiuyong Li, Jixue Liu, Hannu Toivonen, and Jianming Yong. Effective pruning for the discovery of conditional functional dependencies. *The Computer Journal*, 56(3):378–392, 2013.
92. Yunyao Li, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, and H. V. Jagadish. Regular expression learning for information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21–30, 2008.
93. Bing Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2nd edition, 2010.
94. Jixue Liu, Jiuyong Li, Chengfei Liu, and Yongfeng Chen. Discover dependencies from data – a review. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 24(2):251 – 264, 2012.
95. Stéphane Lopes, Jean-Marc Petit, and Lotfi Lakhal. Efficient discovery of functional dependencies and Armstrong relations. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, pages 350–364, 2000.
96. Stéphane Lopes, Jean-Marc Petit, and Farouk Toumani. Discovering interesting inclusion dependencies: application to logical database tuning. *Information Systems*, 27(1):1–19, 2002.
97. Claudio L. Lucchesi and Sylvia L. Osborn. Candidate keys for relations. *Journal of Computer and System Sciences*, 17(2):270 – 279, 1978.
98. Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm. Generic schema matching with Cupid. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 49–58, 2001.
99. Michael V. Mannino, Paicheng Chu, and Thomas Sager. Statistical profile estimation in database systems. *ACM Computing Surveys*, 20(3):191–221, 1988.
100. Fabien De Marchi, Stéphane Lopes, and Jean-Marc Petit. Efficient algorithms for mining inclusion dependencies. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, pages 464–476, 2002.
101. Fabien De Marchi, Stéphane Lopes, and Jean-Marc Petit. Unary and n-ary inclusion dependency discovery in relational databases. *Journal of Intelligent Information Systems*, 32:53–73, 2009.
102. Fabien De Marchi and Jean-Marc Petit. Zigzag: A new algorithm for mining large inclusion dependencies in databases. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 27–34, 2003.
103. Victor M. Markowitz and Johann A. Makowsky. Identifying extended entity-relationship object structures in relational schemas. *IEEE Transactions on Software Engineering*, 16(8):777–790, 1990.
104. Arkady Maydanchik. *Data Quality Assessment*. Technics Publications, New Jersey, 2007.
105. Laurent Mignet, Denilson Barbosa, and Pierangelo Veltri. The XML web: A first study. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 500–510, 2003.
106. Irena Mlynkova, Kamil Toman, and Jaroslav Pokorný. Statistical analysis of real XML data collections. In *Proceedings of the International Conference on Management of Data (COMAD)*, pages 15–26, 2006.
107. Kristi Morton, Magdalena Balazinska, Dan Grossman, and Jock Mackinlay. Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment (PVLDB)*, 7(6):453–456, 2014.
108. Felix Naumann. Data profiling revisited. *SIGMOD Record*, 42(4):40–49, 2013.
109. Felix Naumann, Ching-Tien Ho, Xuqing Tian, Laura Haas, and Nimrod Megiddo. Attribute classification using feature analysis. In *Proceedings of the International Conference on Data Engineering (ICDE)*, page 271, 2002.
110. Noel Novelli and Rosine Cicchetti. FUN: An efficient algorithm for mining functional and embedded dependencies. In *Proceedings of the International Conference on Database Theory (ICDT)*, pages 189–203, 2001.
111. Nikos Ntarmos, Peter Triantafillou, and Gerhard Weikum. Distributed hash sketches: Scalable, efficient, and accurate cardinality estimation for distributed multisets. *ACM Transactions on Computer Systems (TOCS)*, 27(1):1–53, 2009.
112. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
113. Thorsten Papenbrock, Jens Ehrlich, Jannik Marten, Tommy Neubert, Jan-Peer Rudolph, Martin Schönberg, Jakob Zwiener, and Felix Naumann. Functional dependency discovery: An experimental evaluation of seven algorithms. *Proceedings of the VLDB Endowment (PVLDB)*, 8(10), 2015.
114. Thorsten Papenbrock, Sebastian Kruse, Jorge-Arnulfo Quiané-Ruiz, and Felix Naumann. Divide & conquer-based inclusion dependency discovery. *Proceedings of the VLDB Endowment (PVLDB)*, 8(7), 2015.
115. Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. Using a Hash-Based Method with Transaction Trimming for Mining Association Rules. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 9:813–825, 1997.
116. Jean-Marc Petit, Jacques Kouloumdjian, Jean-François Boulicaut, and Farouk Toumani. Using queries to improve database reverse engineering. In *Proceedings of the International Conference on Conceptual Modeling (ER)*, pages 369–386, 1994.
117. Leo Pipino, Yang Lee, and Richard Wang. Data quality assessment. *Communications of the ACM*, 4:211–218, 2002.
118. Viswanath Poosala, Peter J. Haas, Yannis E. Ioannidis, and Eugene J. Shekita. Improved histograms for selectivity estimation of range predicates. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 294–305, 1996.
119. Viswanath Poosala and Yannis E. Ioannidis. Selectivity estimation without the attribute value independence assumption. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 486–495, 1997.
120. Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.
121. Erhard Rahm and Hong-Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000.
122. Vijayshankar Raman and Joseph M. Hellerstein. Potters Wheel: An interactive data cleaning system. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 381–390, 2001.
123. Alexandra Rostin, Oliver Albrecht, Jana Bauckmann, Felix Naumann, and Ulf Leser. A machine learning approach to foreign key discovery. In *Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB)*, 2009.
124. Arnaud Sahuguet and Fabien Azavant. Building lightweight wrappers for legacy Web data-sources using W4F. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 738–741, 1999.
125. Sunita Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008.
126. Yannis Sismanis, Paul Brown, Peter J. Haas, and Berthold Reinwald. GORDIAN: Efficient and scalable discovery of composite keys. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 691–702, 2006.
127. Kenneth P. Smith, Michael Morse, Peter Mork, Maya Hao Li, Arnon Rosenthal, M. David Allen, and Len Seligman. The role of schema matching in large enterprises. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, 2009.
128. Shaoxu Song and Lei Chen. Differential dependencies: Reasoning and discovery. *ACM Transactions on Database Systems (TODS)*, 36(3):16:1–16:41, 2011.
129. Michael Stonebraker, Daniel Bruckner, Ihab F. Ilyas, George Beskales, Mitch Cherniack, Stan Zdonik, Alexander Pagan, and Shan Xu. Data curation at scale: The Data Tamer system. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, 2013.
130. Ming syan Chen, Jiawei Hun, and Philip S. Yu. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 8:866–883, 1996.
131. Pauray S.M. Tsai, Chih-Chong Lee, and Arbee L.P. Chen. An efficient approach for incremental association rule mining. In *Methodologies for Knowledge Discovery and Data Mining*, volume 1574 of *Lecture Notes in Computer Science*, pages 74–83. Springer Berlin Heidelberg, 1999.
132. Millist W. Vincent, Jixue Liu, and Chengfei Liu. Strong functional dependencies and their application to normal forms in XML. *ACM Transactions on Database Systems (TODS)*, 29(3):445–462, 2004.
133. Tobias Vogel and Felix Naumann. Instance-based "one-to-some" assignment of similarity measures to attributes. In *Proceedings of the International Conference on Cooperative Information Systems (CoopIS)*, pages 412–420, 2011.
134. Shyue-Liang Wang, Wen-Chieh Tsou, Jiann-Horng Lin, and Tzung-Pei Hong. Maintenance of discovered functional dependencies: Incremental deletion. In *Intelligent Systems Design and Applications*, volume 23 of *Advances in Soft Computing*, pages 579–588. Springer Berlin Heidelberg, 2003.
135. Xindong Wu, Chengqi Zhang, and Shichao Zhang. Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, 22(3):381–405, 2004.
136. Catharine Wyss, Chris Giannella, and Edward L. Robertson. FastFDs: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances. In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 101–110, 2001.
137. Rui Xu and Donald C. Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.
138. Mohamed Yakout, Ahmed K. Elmagarmid, Jennifer Neville, and Mourad Ouzzani. GDR: A system for guided data repair. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 1223–1226, 2010.
139. Hong Yao and Howard J. Hamilton. Mining functional dependencies from data. *Data Mining and Knowledge Discovery*, 16(2):197–219, 2008.
140. Cong Yu and H. V. Jagadish. Efficient discovery of XML data redundancies. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 103–114, 2006.
141. Mohammed J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 12(3):372–390, 2000.
142. Meihui Zhang and Kaushik Chakrabarti. InfoGather+: Semantic matching and annotation of numeric and time-varying attributes in web tables. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 145–156, 2013.
143. Meihui Zhang, Marios Hadjieleftheriou, Beng Chin Ooi, Cecilia M. Procopiuc, and Divesh Srivastava. On multi-column foreign key discovery. *Proceedings of the VLDB Endowment (PVLDB)*, 3(1-2):805–814, 2010.
144. Meihui Zhang, Marios Hadjieleftheriou, Beng Chin Ooi, Cecilia M. Procopiuc, and Divesh Srivastava. Automatic discovery of attributes in relational databases. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 109–120, 2011.



A2 | 1

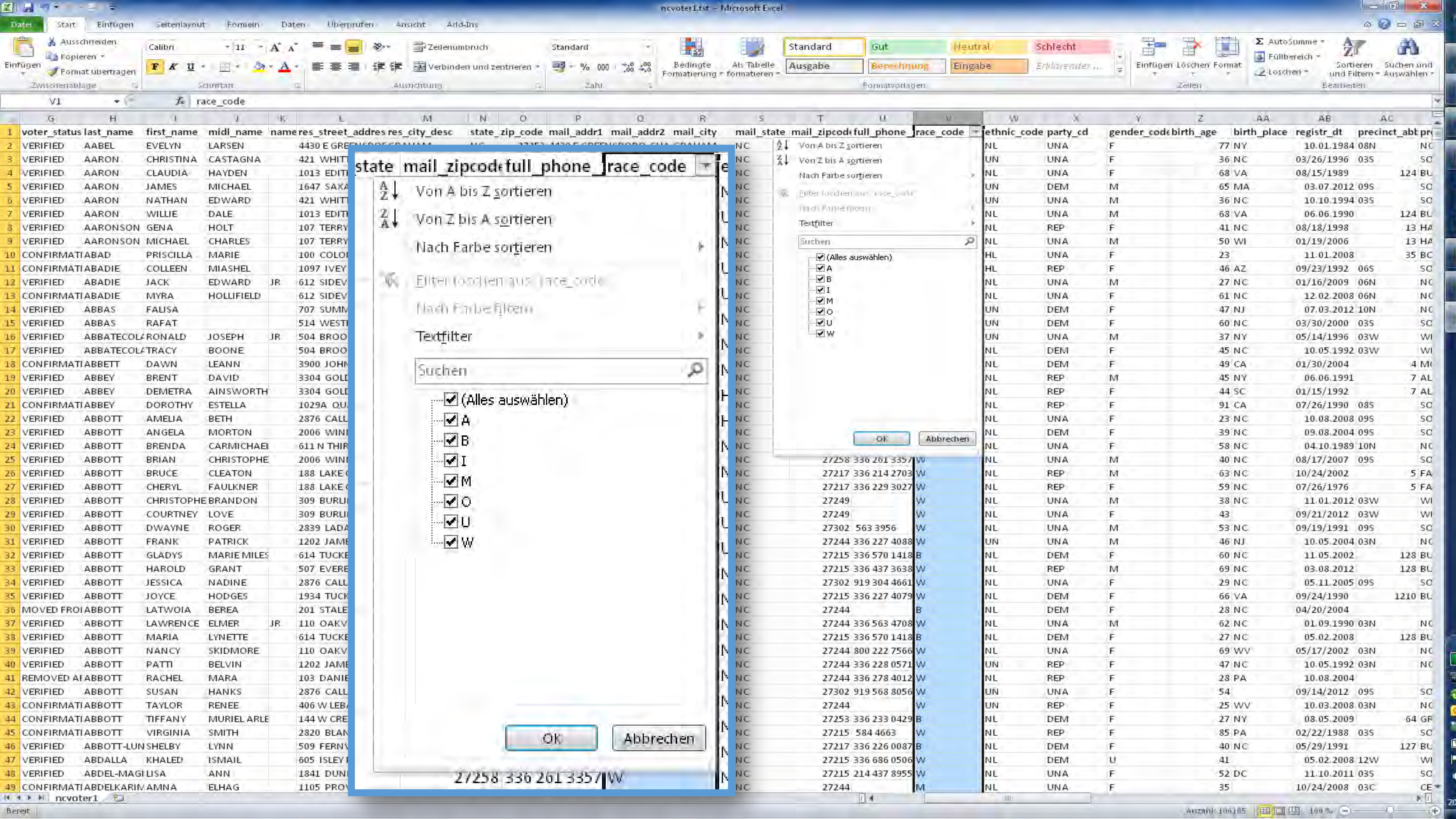
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | ALAMANCE | 9005990 | A | ACTIVE | AV | VERIFIED | AABEL | EVELYN | LARSEN | | 4430 E GREENSBOF | GRAHAM | NC | 27253 | 4430 E GREENSBORO-CHA | | GRAHAM | NC | 27253 | 000 0000 | W | NL | UNA |
| 3 | 1 | ALAMANCE | 9048723 | A | ACTIVE | AV | VERIFIED | AARON | CHRISTINA | CASTAGNA | | 421 WHITT AVE | BURLINGTON | NC | 27215 | PO BOX 4177 | | BURLINGTON | NC | 27215 | 229 1110 | W | UN | UNA |
| 4 | 1 | ALAMANCE | 9019674 | A | ACTIVE | AV | VERIFIED | AARON | CLAUDIA | HAYDEN | | 1013 EDITH ST | BURLINGTON | NC | 27215 | 1013 EDITH ST | | BURLINGTON | NC | 27215 | 222 8834 | W | NL | UNA |
| 5 | 1 | ALAMANCE | 9129589 | A | ACTIVE | AV | VERIFIED | AARON | JAMES | MICHAEL | | 1647 SAXAPAHAW | GRAHAM | NC | 27253 | PO BOX 98 | | SAXAPAHAW | NC | 27340 | 336 525 2484 | W | UN | DEM |
| 6 | 1 | ALAMANCE | 9041748 | A | ACTIVE | AV | VERIFIED | AARON | NATHAN | EDWARD | | 421 WHITT AVE | BURLINGTON | NC | 27215 | PO BOX 4177 | | BURLINGTON | NC | 27215 | 336 229 1110 | W | UN | UNA |
| 7 | 1 | ALAMANCE | 9021947 | A | ACTIVE | AV | VERIFIED | AARON | WILLIE | DALE | | 1013 EDITH ST | BURLINGTON | NC | 27215 | 1013 EDITH ST | | BURLINGTON | NC | 27215 | 336 999 9999 | W | NL | UNA |
| 8 | 1 | ALAMANCE | 9062002 | A | ACTIVE | AV | VERIFIED | AARONSON | GENA | HOLT | | 107 TERRYWOOD ( | HAW RIVER | NC | 27258 | 107 TERRYWOOD CT | | HAW RIVER | NC | 27258 | 336 578 9123 | W | NL | REP |
| 9 | 1 | ALAMANCE | 9096423 | A | ACTIVE | AV | VERIFIED | AARONSON | MICHAEL | CHARLES | | 107 TERRYWOOD ( | HAW RIVER | NC | 27258 | 107 TERRYWOOD CT | | HAW RIVER | NC | 27258 | 336 266 7615 | W | NL | UNA |
| 10 | 1 | ALAMANCE | 9117940 | I | INACTIVE | IU | CONFIRMATI | ABAD | PRISCILLA | MARIE | | 100 COLONNADE I | ELON | NC | 27244 | CAMPUS BOX 3008 | | ELON | NC | 27244 | | O | HL | UNA |
| 11 | 1 | ALAMANCE | 9034127 | I | INACTIVE | IU | CONFIRMATI | ABADIE | COLLEEN | MIASHEL | | 1097 IVEY RD #C | GRAHAM | NC | 27253 | 1097 IVEY RD #C | | GRAHAM | NC | 27253 | | M | HL | REP |
| 12 | 1 | ALAMANCE | 9121656 | 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 | 9118154 | I | INACTIVE | IU | CONFIRMATI | ABADIE | MYRA | HOLLIFIELD | | 612 SIDEVIEW ST | GRAHAM | NC | 27253 | 617 MITCHELL ST | | BURLINGTON | NC | 27217 | 336 212 8140 | W | NL | UNA |
| 14 | 1 | ALAMANCE | 9131788 | 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 | 9068460 | A | ACTIVE | AV | VERIFIED | ABBAS | RAFAT | | | 514 WESTRIDGE DI | BURLINGTON | NC | 27215 | 514 WESTRIDGE DR | | BURLINGTON | NC | 27215 | | A | UN | DEM |
| 16 | 1 | ALAMANCE | 9049573 | A | ACTIVE | AV | VERIFIED | ABBATECOLA | RONALD | JOSEPH | JR | 504 BROOKFIELD D | GIBSONVILLE | NC | 27249 | 504 BROOKFIELD DR | | GIBSONVILLE | NC | 27249 | 336 449 9029 | W | UN | UNA |
| 17 | 1 | ALAMANCE | 9033877 | A | ACTIVE | AV | VERIFIED | ABBATECOLA | TRACY | BOONE | | 504 BROOKFIELD D | GIBSONVILLE | NC | 27249 | 504 BROOKFIELD DR | | GIBSONVILLE | NC | 27249 | | W | NL | DEM |
| 18 | 1 | ALAMANCE | 9083557 | I | INACTIVE | IU | CONFIRMATI | ABBETT | 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 | 9022529 | I | INACTIVE | IU | CONFIRMATI | ABBEY | DOROTHY | ESTELLA | | 1029A QUAKENBU | SNOW CAMP | NC | 27349 | 1029A QUAKENBUSH RD | | SNOW CAMP | NC | 27349 | 376 3663 | W | NL | REP |
| 22 | 1 | ALAMANCE | 9113186 | A | ACTIVE | AV | VERIFIED | ABBOTT | AMELIA | BETH | | 2876 CALLOWAY 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 | CARMICHAEI | | 611 N THIRD ST | MEBANE | NC | 27302 | 611 N THIRD ST | | MEBANE | NC | 27302 | 563 2654 | W | NL | UNA |
| 25 | 1 | ALAMANCE | 9102615 | A | ACTIVE | AV | VERIFIED | ABBOTT | BRIAN | CHRISTOPHE | | 2006 WINN CREEK | HAW RIVER | NC | 27258 | 2006 WINN CREEK DR | | HAW RIVER | NC | 27258 | 336 261 3357 | W | NL | UNA |
| 26 | 1 | ALAMANCE | 9079257 | A | ACTIVE | AV | VERIFIED | ABBOTT | BRUCE | CLEATON | | 188 LAKE CAMMA | BURLINGTON | NC | 27217 | 188 LAKE CAMMACK CT | | BURLINGTON | NC | 27217 | 336 214 2703 | W | NL | REP |
| 27 | 1 | ALAMANCE | 1389300 | A | ACTIVE | AV | VERIFIED | ABBOTT | CHERYL | FAULKNER | | 188 LAKE CAMMA | BURLINGTON | NC | 27217 | 188 LAKE CAMMACK CT | | BURLINGTON | NC | 27217 | 336 229 3027 | W | NL | REP |
| 28 | 1 | ALAMANCE | 9140392 | A | ACTIVE | AV | VERIFIED | ABBOTT | CHRISTOPHE | BRANDON | | 309 BURLINGTON | GIBSONVILLE | NC | 27249 | 309 BURLINGTON AVE | | GIBSONVILLE | NC | 27249 | | W | NL | UNA |
| 29 | 1 | ALAMANCE | 9135711 | A | ACTIVE | AV | VERIFIED | ABBOTT | COURTNEY | LOVE | | 309 BURLINGTON | GIBSONVILLE | NC | 27249 | 309 BURLINGTON AVE | | GIBSONVILLE | NC | 27249 | | W | NL | UNA |
| 30 | 1 | ALAMANCE | 9028439 | A | ACTIVE | AV | VERIFIED | ABBOTT | DWAYNE | ROGER | | 2839 LADALE LN | MEBANE | NC | 27302 | 2839 LADALE LN | | MEBANE | NC | 27302 | 563 3956 | W | NL | UNA |
| 31 | 1 | ALAMANCE | 9090420 | A | ACTIVE | AV | VERIFIED | ABBOTT | FRANK | PATRICK | | 1202 JAMESTOWN | ELON | NC | 27244 | 1202 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 | BEREA | | 201 STALEY HALL | ELON | NC | 27244 | CAMPUS BOX 3039 | | ELON | NC | 27244 | | B | NL | DEM |
| 37 | 1 | ALAMANCE | 9020357 | A | ACTIVE | AV | VERIFIED | ABBOTT | LAWRENCE | ELMER | 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 | UN | UNA |
| 43 | 1 | ALAMANCE | 9113731 | I | INACTIVE | IU | CONFIRMATI | ABBOTT | TAYLOR | RENEE | | 406 W LEBANON A | ELON | NC | 27244 | CAMPUS BOX 3077 | | ELON | NC | 27244 | | W | UN | REP |
| 44 | 1 | ALAMANCE | 9120825 | I | INACTIVE | IN | CONFIRMATI | ABBOTT | TIFFANY | MURIEL ARLE | | 144 W CRESCENT S | GRAHAM | NC | 27253 | 144 W CRESCENT SQUARE | | GRAHAM | NC | 27253 | 336 233 0429 | B | NL | DEM |
| 45 | 1 | ALAMANCE | 9013866 | I | INACTIVE | IN | CONFIRMATI | ABBOTT | VIRGINIA | SMITH | | 2820 BLANCHE DR | BURLINGTON | NC | 27215 | 2820 BLANCHE DR | | BURLINGTON | NC | 27215 | 584 4663 | W | NL | REP |
| 46 | 1 | ALAMANCE | 9027717 | A | ACTIVE | AV | VERIFIED | ABBOTT-LUN | SHELBY | LYNN | | 509 FERNWAY DR | BURLINGTON | NC | 27217 | 509 FERNWAY DR | | BURLINGTON | NC | 27217 | 336 226 0087 | B | NL | DEM |
| 47 | 1 | ALAMANCE | 9108552 | A | ACTIVE | AV | VERIFIED | ABDALLA | KHALED | ISMAIL | | 605 ISLEY PL #C | BURLINGTON | NC | 27215 | 605 ISLEY PL #C | | BURLINGTON | NC | 27215 | 336 686 0506 | W | NL | DEM |
| 48 | 1 | ALAMANCE | 9128403 | A | ACTIVE | AV | VERIFIED | ABDEL-MAGI | LISA | ANN | | 1841 DUNBAR PL | BURLINGTON | NC | 27215 | 1841 DUNBAR PL | | BURLINGTON | NC | 27215 | 214 437 8955 | W | NL | UNA |
| 49 | 1 | ALAMANCE | 9117192 | I | INACTIVE | IU | CONFIRMATI | ABDELKARIM | AMNA | ELHAG | | 1105 PROVIDENCE | ELON | NC | 27244 | 1105 PROVIDENCE CT | | ELON | NC | 27244 | | M | NL | UNA |
| 50 | 1 | ALAMANCE | 9099437 | A | ACTIVE | AV | VERIFIED | ABDELRAHAI | ABUBAKR | MERGANI | | 2954 ETHAN POIN | BURLINGTON | NC | 27215 | 2954 ETHAN POINTE DR # | | BURLINGTON | NC | 27215 | 336 684 0985 | O | NL | DEM |



A106185 fx 1

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 106138 | 1 | 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 |
| 106139 | 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 |
| 106140 | 1 | ALAMANCE | 9112148 | A | ACTIVE | AV | VERIFIED | ZMIJEWSKI | DENNIS | AL | | 4872 THOM RD | MEBANE | NC | 27302 | 4872 THOM RD | | MEBANE | NC | 27302 | | W | UN | DEM |
| 106141 | 1 | 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 |
| 106142 | 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 |
| 106143 | 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 |
| 106144 | 1 | ALAMANCE | 9094116 | A | ACTIVE | AV | VERIFIED | ZMIJEWSKI N | VIRGINIA | LOURDES | | 4872 THOM RD | MEBANE | NC | 27302 | 4872 THOM RD | | MEBANE | NC | 27302 | 336 376 1987 | U | UN | UNA |
| 106145 | 1 | ALAMANCE | 9089250 | R | REMOVED | RD | DECEASED | ZOCCOLANTI | ENIS | PIZZOTTI | | 2502 S NC HWY 11 | MEBANE | NC | 27302 | 2502 S NC HWY 119 | | MEBANE | NC | 27302 | | W | UN | REP |
| 106146 | 1 | ALAMANCE | 9083629 | R | REMOVED | RD | DECEASED | ZOCCOLANTI | RENATO | | | 3141 SHELLY GRAH | GRAHAM | NC | 27253 | 3141 SHELLY GRAHAM DR | | GRAHAM | NC | 27253 | 336 227 7168 | W | NL | REP |
| 106147 | 1 | ALAMANCE | 9083630 | A | ACTIVE | AV | VERIFIED | ZOCCOLANTI | RITA | MARIE | | 3141 SHELLY GRAH | GRAHAM | NC | 27253 | 3141 SHELLY GRAHAM DR | | GRAHAM | NC | 27253 | 336 227 7168 | W | NL | REP |
| 106148 | 1 | ALAMANCE | 9100545 | I | INACTIVE | IU | CONFIRMATI | ZOGLEMANN | ANGELA | LYNNE | | 706 HUFFMAN MIL | BURLINGTON | NC | 27215 | 706 HUFFMAN MILL RD | # | BURLINGTON | NC | 27215 | 336 227 1261 | W | NL | UNA |
| 106149 | 1 | 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 |
| 106150 | 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 |
| 106151 | 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 |
| 106152 | 1 | ALAMANCE | 9108096 | A | ACTIVE | AV | VERIFIED | ZOLLARS | EVELYN | NADINE | | 6830 TOM WOODY | SNOW CAMP | NC | 27349 | 6830 TOM WOODY RD | | SNOW CAMP | NC | 27349 | 336 376 5754 | W | NL | UNA |
| 106153 | 1 | ALAMANCE | 9125044 | A | ACTIVE | AV | VERIFIED | ZOLLARS | MATHEW | DAVID | | 6830 TOM WOODY | SNOW CAMP | NC | 27349 | 6830 TOM WOODY RD | | SNOW CAMP | NC | 27349 | | W | NL | UNA |
| 106154 | 1 | ALAMANCE | 9113912 | A | ACTIVE | AV | VERIFIED | ZOLLICOFFE | ANTONIO | MARK | | 108 OAKGROVE D | GRAHAM | NC | 27253 | 108 OAKGROVE DR | | GRAHAM | NC | 27253 | 336 260 6673 | B | UN | DEM |
| 106155 | 1 | ALAMANCE | 9107068 | A | ACTIVE | AV | VERIFIED | ZOLLICOFFE | VALERIE | | | 108 OAKGROVE D | GRAHAM | NC | 27253 | 108 OAKGROVE DR | | GRAHAM | NC | 27253 | | B | UN | DEM |
| 106156 | 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 |
| 106157 | 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 |
| 106158 | 1 | ALAMANCE | 9104969 | I | INACTIVE | IU | CONFIRMATI | ZORNES | MICHELLE | LEE | | 3117 COMMERCE | BURLINGTON | NC | 27215 | 3117 COMMERCE PL | #L | BURLINGTON | NC | 27215 | 336 675 0520 | W | UN | UNA |
| 106159 | 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 |
| 106160 | 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 |
| 106161 | 1 | ALAMANCE | 9110367 | D | DENIED | DU | VERIFICATIO | ZORNES | TINA | | | 801 TROLLINGWO | HAW RIVER | NC | 27258 | 801 TROLLINGWOOD RD | | HAW RIVER | NC | 27258 | 336 578 0646 | W | UN | UNA |
| 106162 | 1 | ALAMANCE | 9132758 | A | ACTIVE | AV | VERIFIED | ZORNES | TINA | MARIE | | 801 TROLLINGWO | HAW RIVER | NC | 27258 | 801 TROLLINGWOOD RD | | HAW RIVER | NC | 27258 | 336 420 7630 | W | NL | UNA |
| 106163 | 1 | ALAMANCE | 9131499 | A | ACTIVE | AV | VERIFIED | ZOUFALY | EVE | | | 602 E HAGGARD A | ELON | NC | 27244 | CAMPUS BOX 8911 | | ELON | NC | 27244 | | U | UN | UNA |
| 106164 | 1 | ALAMANCE | 9124446 | A | ACTIVE | AV | VERIFIED | ZSUPPAN | ETELKA | HALASZ | | 1929 HAW VILLAG | GRAHAM | NC | 27253 | 1929 HAW VILLAGE DR | | GRAHAM | NC | 27253 | | W | NL | REP |
| 106165 | 1 | ALAMANCE | 9121554 | A | ACTIVE | AV | VERIFIED | ZSUPPAN | FERENC | | | 1929 HAW VILLAG | GRAHAM | NC | 27253 | 1929 HAW VILLAGE DR | | GRAHAM | NC | 27253 | | W | UN | REP |
| 106166 | 1 | ALAMANCE | 9127457 | A | ACTIVE | AV | VERIFIED | ZSUPPAN | LEVENTE | FERENC | | 1929 HAW VILLAG | GRAHAM | NC | 27253 | 1929 HAW VILLAGE DR | | GRAHAM | NC | 27253 | 336 376 1365 | W | NL | REP |
| 106167 | 1 | ALAMANCE | 9131401 | A | ACTIVE | AV | VERIFIED | ZUBLER | LINDSAY | BROOKE | | 3172 CARRIAGE C | HAW RIVER | NC | 27258 | 3172 CARRIAGE CREEK CT | | HAW RIVER | NC | 27258 | | U | UN | UNA |
| 106168 | 1 | ALAMANCE | 9081728 | A | ACTIVE | AV | VERIFIED | ZUBLER | TAMI | LAJEAN | | 3172 CARRIAGE C | HAW RIVER | NC | 27258 | 3172 CARRIAGE CREEK CT | | HAW RIVER | NC | 27258 | 336 578 8028 | W | NL | UNA |
| 106169 | 1 | ALAMANCE | 9089569 | A | ACTIVE | AV | VERIFIED | ZUBLER | TIMOTHY | JAMES | | 3172 CARRIAGE C | HAW RIVER | NC | 27258 | 3172 CARRIAGE CREEK CT | | HAW RIVER | NC | 27258 | | W | UN | UNA |
| 106170 | 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 |
| 106171 | 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 |
| 106172 | 1 | ALAMANCE | 9008787 | A | ACTIVE | AV | VERIFIED | ZUMER | FRANK | EDWARD | | 801 QUAKER RIDG | MEBANE | NC | 27302 | 801 QUAKER RIDGE RD | | MEBANE | NC | 27302 | 919 563 3766 | W | UN | UNA |
| 106173 | 1 | ALAMANCE | 9008785 | A | ACTIVE | AV | VERIFIED | ZUMER | LOUISE | TURNER | | 801 QUAKER RIDG | MEBANE | NC | 27302 | 801 QUAKER RIDGE RD | | MEBANE | NC | 27302 | 919 563 3766 | W | NL | DEM |
| 106174 | 1 | 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 |
| 106175 | 1 | ALAMANCE | 9119438 | A | ACTIVE | | VERIFIED | ZUNIGA | JOSE | RAMON SAL | | 714 ROSS ST | BURLINGTON | NC | 27217 | 714 ROSS ST | | BURLINGTON | NC | 27217 | 336 227 3108 | O | HL | DEM |
| 106176 | 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 |
| 106177 | 1 | ALAMANCE | 9112637 | A | | | VERIFIED | ZUNIGA | YANET | SALAS | | 3845 MAE DOUGL | MEBANE | NC | 27302 | 3845 MAE DOUGLAS DR | | MEBANE | NC | 27302 | | O | HL | DEM |
| 106178 | 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 |
| 106179 | 1 | ALAMANCE | 914 | | | AV | 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 |
| 106180 | 1 | ALAMA | | | | AV | VERIFIED | ZURFACE | ROSSELL | EUGENE | | 2074 TURNER RD | MEBANE | NC | 27302 | 2074 TURNER RD | | MEBANE | NC | 27302 | | W | UN | UNA |
| 106181 | 1 | ALAM | | | ACTIVE | AV | VERIFIED | ZWIER | ANDREW | MICHAEL | | 1497 LONGEST AC | SNOW CAMP | NC | 27349 | 1497 LONGEST ACRES RD | | SNOW CAMP | NC | 27349 | 336 376 8830 | W | NL | REP |
| 106182 | 1 | ALA | | A | ACTIVE | AV | VERIFIED | ZWIER | CHRISTOPHE | ANTHONY | | 1497 LONGEST AC | SNOW CAMP | NC | 27349 | 1497 LONGEST ACRES RD | | SNOW CAMP | NC | 27349 | 831 207 9222 | W | NL | REP |
| 106183 | 1 | AL | 9140499 | A | ACTIVE | AV | VERIFIED | ZWIER | CHRISTY | ANN | | 1497 LONGEST AC | SNOW CAMP | NC | 27349 | 1497 LONGEST ACRES RD | | SNOW CAMP | NC | 27349 | | W | NL | REP |
| 106184 | 1 | ALAMAN | 9099261 | A | ACTIVE | AV | VERIFIED | ZWIER | KAREN | JEAN | | 1497 LONGEST AC | SNOW CAMP | NC | 27349 | 1497 LONGEST ACRES RD | | SNOW CAMP | NC | 27349 | 831 207 9222 | W | NL | REP |
| 106185 | 1 | ALAMANCE | 9077804 | R | REMOVED | RL | MOVED FRO | ZYLKA | MARC | | | 1210 WILLOW BR | MEBANE | NC | 27302 | 1210 WILLOW BROOK CT | | MEBANE | NC | 27302 | 336 578 8580 | W | UN | REP |
| 106186 | | | | | | | | | | | | | | | | | | | | | | | | |

Number of rows



ncvoter1.txt - Microsoft Excel

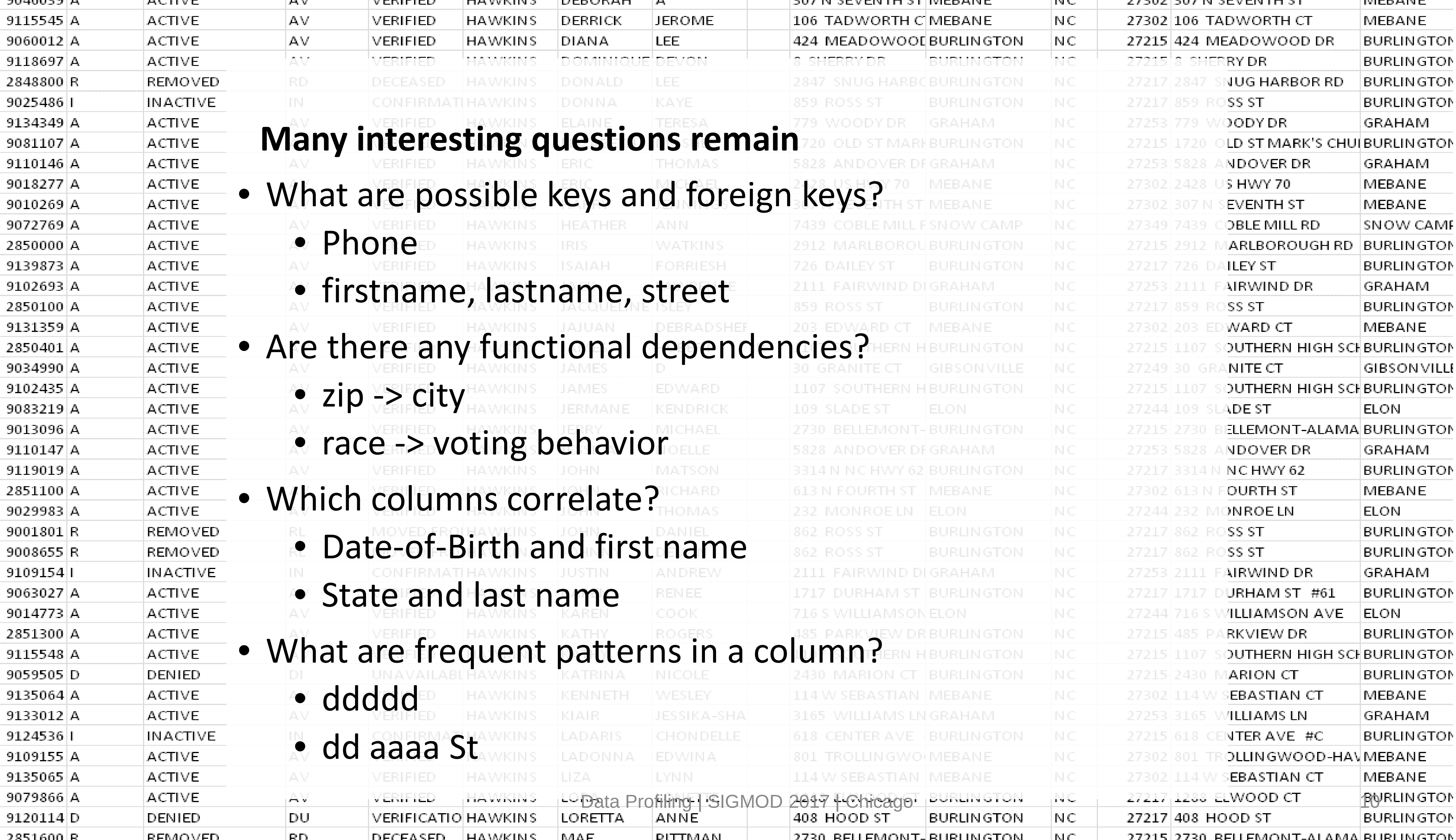
V1 race_code

☑ (Alles auswählen)
☑ A
☑ B
☑ I
☑ M
☑ O
☑ U
☑ W

| | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB | AC | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | voter_status | last_name | first_name | midl_name | name | res_street_addres | res_city_desc | state_ | zip_code | mail_addr1 | mail_addr2 | mail_city | mail_state | mail_zipcode | full_phone_ | race_code | ethnic_code | party_cd | gender_code | birth_age | birth_place | registr_dt | precinct_abb | pr |
| 2 | VERIFIED | AABEL | EVELYN | LARSEN | | 4430 E GRE | | | | | | | NC | | | | NL | UNA | F | 77 | NY | 10.01.1984 | 08N | NC |
| 3 | VERIFIED | AARON | CHRISTINA | CASTAGNA | | 421 WHIT | | | | | | | NC | | | | UN | UNA | F | 36 | NC | 03/26/1996 | 03S | SC |
| 4 | VERIFIED | AARON | CLAUDIA | HAYDEN | | 1013 EDIT | | | | | | | NC | | | | NL | UNA | F | 68 | VA | 08/15/1989 | 124 | BU |
| 5 | VERIFIED | AARON | JAMES | MICHAEL | | 1647 SAXA | | | | | | | NC | | | | UN | DEM | M | 65 | MA | 03.07.2012 | 09S | SC |
| 6 | VERIFIED | AARON | NATHAN | EDWARD | | 421 WHIT | | | | | | | NC | | | | UN | UNA | M | 36 | NC | 10.10.1994 | 03S | SC |
| 7 | VERIFIED | AARON | WILLIE | DALE | | 1013 EDIT | | | | | | | NC | | | | NL | UNA | M | 68 | VA | 06.06.1990 | 124 | BU |
| 8 | VERIFIED | AARONSON | GENA | HOLT | | 107 TERRY | | | | | | | NC | | | | NL | REP | F | 41 | NC | 08/18/1998 | 13 | HA |
| 9 | VERIFIED | AARONSON | MICHAEL | CHARLES | | 107 TERRY | | | | | | | NC | | | | NL | UNA | M | 50 | WI | 01/19/2006 | 13 | HA |
| 10 | CONFIRMATI | ABAD | PRISCILLA | MARIE | | 100 COLO | | | | | | | NC | | | | HL | UNA | F | 23 | | 11.01.2008 | 35 | BC |
| 11 | CONFIRMATI | ABADIE | COLLEEN | MIASHEL | | 1097 IVEY | | | | | | | NC | | | | HL | REP | F | 46 | AZ | 09/23/1992 | 06S | SC |
| 12 | VERIFIED | ABADIE | JACK | EDWARD | JR | 612 SIDEV | | | | | | | NC | | | | NL | UNA | M | 27 | NC | 01/16/2009 | 06N | NC |
| 13 | CONFIRMATI | ABADIE | MYRA | HOLLIFIELD | | 612 SIDEV | | | | | | | NC | | | | NL | UNA | F | 61 | NC | 12.02.2008 | 06N | NC |
| 14 | VERIFIED | ABBAS | FALISA | | | 707 SUMM | | | | | | | NC | | | | UN | DEM | F | 47 | NJ | 07.03.2012 | 10N | NC |
| 15 | VERIFIED | ABBAS | RAFAT | | | 514 WEST | | | | | | | NC | | | | UN | DEM | F | 60 | NC | 03/30/2000 | 03S | SC |
| 16 | VERIFIED | ABBATECOLA | RONALD | JOSEPH | JR | 504 BROO | | | | | | | NC | | | | UN | UNA | M | 37 | NY | 05/14/1996 | 03W | WI |
| 17 | VERIFIED | ABBATECOLA | TRACY | BOONE | | 504 BROO | | | | | | | NC | | | | NL | DEM | F | 45 | NC | 10.05.1992 | 03W | WI |
| 18 | CONFIRMATI | ABBETT | DAWN | LEANN | | 3900 JOHN | | | | | | | NC | | | | NL | DEM | F | 49 | CA | 01/30/2004 | 4 | MC |
| 19 | VERIFIED | ABBEY | BRENT | DAVID | | 3304 GOLD | | | | | | | NC | | | | NL | REP | M | 45 | NY | 06.06.1991 | 7 | AL |
| 20 | VERIFIED | ABBEY | DEMETRA | AINSWORTH | | 3304 GOLD | | | | | | | NC | | | | NL | REP | F | 44 | SC | 01/15/1992 | 7 | AL |
| 21 | CONFIRMATI | ABBEY | DOROTHY | ESTELLA | | 1029A QU | | | | | | | NC | | | | NL | REP | F | 91 | CA | 07/26/1990 | 08S | SC |
| 22 | VERIFIED | ABBOTT | AMELIA | BETH | | 2876 CALL | | | | | | | NC | | | | NL | UNA | F | 23 | NC | 10.08.2008 | 09S | SC |
| 23 | VERIFIED | ABBOTT | ANGELA | MORTON | | 2006 WIN | | | | | | | NC | | | | NL | DEM | F | 39 | NC | 09.08.2004 | 09S | SC |
| 24 | VERIFIED | ABBOTT | BRENDA | CARMICHAEL | | 611 N THIR | | | | | | | NC | | | | NL | UNA | F | 58 | NC | 04.10.1989 | 10N | NC |
| 25 | VERIFIED | ABBOTT | BRIAN | CHRISTOPHE | | 2006 WIN | | | | | | | NC | 27258 | 336 261 3357 | W | NL | UNA | M | 40 | NC | 08/17/2007 | 09S | SC |
| 26 | VERIFIED | ABBOTT | BRUCE | CLEATON | | 188 LAKE | | | | | | | NC | 27217 | 336 214 2703 | W | NL | REP | M | 63 | NC | 10/24/2002 | 5 | FA |
| 27 | VERIFIED | ABBOTT | CHERYL | FAULKNER | | 188 LAKE | | | | | | | NC | 27217 | 336 229 3027 | W | NL | REP | F | 59 | NC | 07/26/1976 | 5 | FA |
| 28 | VERIFIED | ABBOTT | CHRISTOPHE | BRANDON | | 309 BURLI | | | | | | | NC | 27249 | | W | NL | UNA | M | 38 | NC | 11.01.2012 | 03W | WI |
| 29 | VERIFIED | ABBOTT | COURTNEY | LOVE | | 309 BURLI | | | | | | | NC | 27249 | | W | NL | UNA | F | 43 | | 09/21/2012 | 03W | WI |
| 30 | VERIFIED | ABBOTT | DWAYNE | ROGER | | 2839 LADA | | | | | | | NC | 27302 | 563 3956 | W | NL | UNA | M | 53 | NC | 09/19/1991 | 09S | SC |
| 31 | VERIFIED | ABBOTT | FRANK | PATRICK | | 1202 JAME | | | | | | | NC | 27244 | 336 227 4088 | W | UN | UNA | M | 46 | NJ | 10.05.2004 | 03N | NC |
| 32 | VERIFIED | ABBOTT | GLADYS | MARIE MILES | | 614 TUCKE | | | | | | | NC | 27215 | 336 570 1418 | B | NL | DEM | F | 60 | NC | 11.05.2002 | 128 | BU |
| 33 | VERIFIED | ABBOTT | HAROLD | GRANT | | 507 EVERE | | | | | | | NC | 27215 | 336 437 3638 | W | NL | REP | M | 69 | NC | 03.08.2012 | 128 | BU |
| 34 | VERIFIED | ABBOTT | JESSICA | NADINE | | 2876 CALL | | | | | | | NC | 27302 | 919 304 4661 | W | NL | UNA | F | 29 | NC | 05.11.2005 | 09S | SC |
| 35 | VERIFIED | ABBOTT | JOYCE | HODGES | | 1934 TUCK | | | | | | | NC | 27215 | 336 227 4079 | W | NL | DEM | F | 66 | VA | 09/24/1990 | 1210 | BU |
| 36 | MOVED FRO | ABBOTT | LATWOIA | BEREA | | 201 STALE | | | | | | | NC | 27244 | | B | NL | DEM | F | 28 | NC | 04/20/2004 | | |
| 37 | VERIFIED | ABBOTT | LAWRENCE | ELMER | JR | 110 OAKV | | | | | | | NC | 27244 | 336 563 4708 | W | NL | UNA | M | 62 | NC | 01.09.1990 | 03N | NC |
| 38 | VERIFIED | ABBOTT | MARIA | LYNETTE | | 614 TUCKE | | | | | | | NC | 27215 | 336 570 1418 | B | NL | DEM | F | 27 | NC | 05.02.2008 | 128 | BU |
| 39 | VERIFIED | ABBOTT | NANCY | SKIDMORE | | 110 OAKV | | | | | | | NC | 27244 | 800 222 7566 | W | NL | UNA | F | 69 | WV | 05/17/2002 | 03N | NC |
| 40 | VERIFIED | ABBOTT | PATTI | BELVIN | | 1202 JAME | | | | | | | NC | 27244 | 336 228 0571 | W | UN | REP | F | 47 | NC | 10.05.1992 | 03N | NC |
| 41 | REMOVED AF | ABBOTT | RACHEL | MARA | | 103 DANIE | | | | | | | NC | 27244 | 336 278 4012 | W | NL | REP | F | 28 | PA | 10.08.2004 | | |
| 42 | VERIFIED | ABBOTT | SUSAN | HANKS | | 2876 CALL | | | | | | | NC | 27302 | 919 568 8056 | W | UN | UNA | F | 54 | | 09/14/2012 | 09S | SC |
| 43 | CONFIRMATI | ABBOTT | TAYLOR | RENEE | | 406 W LEBA | | | | | | | NC | 27244 | | W | UN | REP | F | 25 | WV | 10.03.2008 | 03N | NC |
| 44 | CONFIRMATI | ABBOTT | TIFFANY | MURIEL ARLE | | 144 W CRE | | | | | | | NC | 27253 | 336 233 0429 | B | NL | DEM | F | 27 | NY | 08.05.2009 | 64 | GR |
| 45 | CONFIRMATI | ABBOTT | VIRGINIA | SMITH | | 2820 BLAN | | | | | | | NC | 27215 | 584 4663 | W | NL | REP | F | 85 | PA | 02/22/1988 | 03S | SC |
| 46 | VERIFIED | ABBOTT-LUN | SHELBY | LYNN | | 509 FERNV | | | | | | | NC | 27217 | 336 226 0087 | B | NL | DEM | F | 40 | NC | 05/29/1991 | 127 | BU |
| 47 | VERIFIED | ABDALLA | KHALED | ISMAIL | | 605 ISLEY | | | | | | | NC | 27215 | 336 686 0506 | W | NL | DEM | U | 41 | | 05.02.2008 | 12W | WI |
| 48 | VERIFIED | ABDEL-MAGI | LISA | ANN | | 1841 DUN | | | | | | | NC | 27215 | 214 437 8955 | W | NL | UNA | F | 52 | DC | 11.10.2011 | 03S | SC |
| 49 | CONFIRMATI | ABDELKARIM | AMNA | ELHAG | | 1105 PROV | | | | | | | NC | 27244 | | M | NL | UNA | F | 35 | | 10/24/2008 | 03C | CE |

ncvoter1

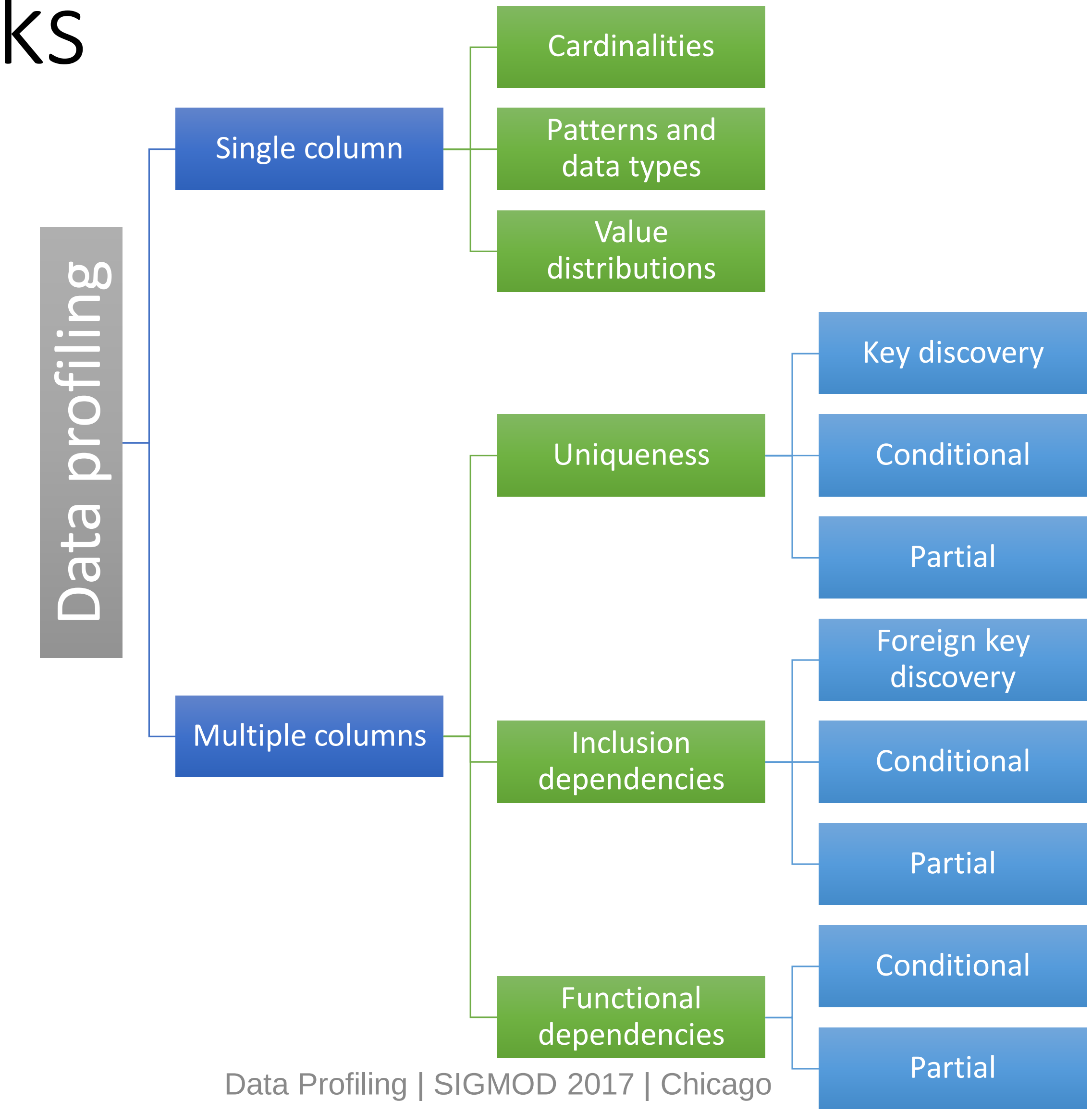
|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ALAMANCE | 9000990 | A | ACTIVE | AV | VERIFIED | [illegible] | [illegible] | [illegible] |  |
| 1 | ALAMANCE | 9040733 | A | ACTIVE | AV | VERIFIED | [illegible] | CHRISTIE | [illegible] |  |
| 1 | ALAMANCE | 9019674 | A | ACTIVE | AV | VERIFIED | [illegible] | [illegible] | [illegible] |  |
| 1 | ALAMANCE | 9139009 | A | ACTIVE | AV | VERIFIED | [illegible] | JAMES | MICHAEL |  |
| 1 | ALAMANCE | 9041740 | A | ACTIVE | AV | VERIFIED | [illegible] | [illegible] | EDWARD |  |
| 1 | ALAMANCE | 9031947 | A | ACTIVE | AV | VERIFIED | [illegible] | WILLIE | [illegible] |  |
| 1 | ALAMANCE | 9063003 | A | ACTIVE | AV | VERIFIED | [illegible] | [illegible] | [illegible] |  |
| 1 | ALAMANCE | 9096433 | A | ACTIVE | AV | VERIFIED | [illegible] | MICHAEL | [illegible] |  |
| 1 | ALAMANCE | 9117940 | I | INACTIVE | IN | CONFIRMATI | [illegible] | PRISCILLA | MARIA |  |
| 1 | ALAMANCE | 9034137 | I | INACTIVE | IN | CONFIRMATI | [illegible] | [illegible] | MICHAEL |  |
| 1 | ALAMANCE | 9131666 | A | ACTIVE | AV | VERIFIED | [illegible] | JACE | EDWARD | JR |
| 1 | ALAMANCE | 9111184 | I | INACTIVE | IN | CONFIRMATI | [illegible] | MARK | [illegible] |  |
| 1 | ALAMANCE | 9131700 | A | ACTIVE | AV | VERIFIED | [illegible] | [illegible] |  |  |
| 1 | ALAMANCE | 9066460 | A | ACTIVE | AV | VERIFIED | [illegible] | [illegible] |  |  |



## Many interesting questions remain

- What are possible keys and foreign keys?
  - Phone
  - firstname, lastname, street
- Are there any functional dependencies?
  - zip -> city
  - race -> voting behavior
- Which columns correlate?
  - Date-of-Birth and first name
  - State and last name
- What are frequent patterns in a column?
  - ddddd
  - dd aaaa St

# Classification of Traditional Profiling Tasks



# Data Profiling vs. Data Mining

- Data profiling gathers technical metadata to support data management
- Data mining and data analytics discovers non-obvious results to support business management

- Data profiling results: information about columns and column sets
- Data mining results: information about rows or row sets
  - clustering, summarization, association rules, …

- Rahm and Do on data cleaning
  - Profiling: Individual attributes
  - Mining: Multiple attributes

[Rahm and Do, Data Cleaning: Problems and Current Approaches, IEEE DE Bulletin, 2000]

# Challenges of (Big) Data Profiling

- Large search space
  - Number of rows AND number of columns (and column combinations)
  - “Small” table with 100 columns:
    $2^{100} – 1$ = 1,267,650,600,228,229,401,496,703,205,375
    = 1.3 nonillion column combinations
- Large solution space: Exponential number of dependencies

- New data types and new data models
- New requirements: User-oriented, interactive, streaming

- Solutions: Scale up, scale out, scale in
- Better: Intelligent enumeration and aggressive pruning

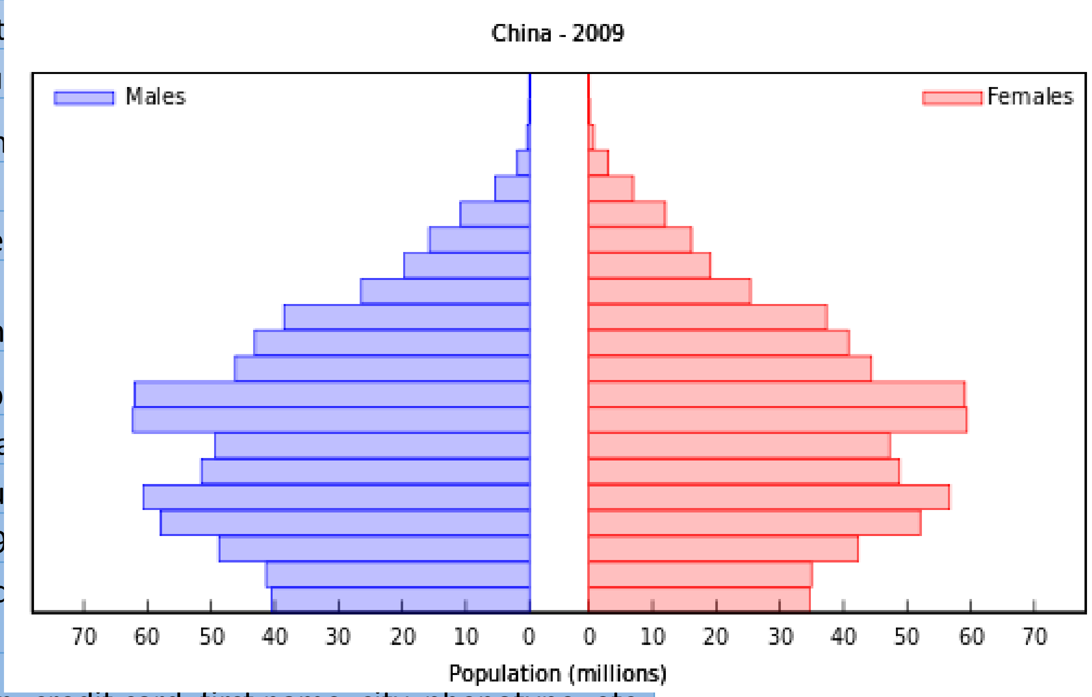
# Use Cases for Profiling

- Query optimization
    - Counts and histograms
- Data cleansing
    - Patterns and violations
- Data integration
    - Cross-DB inclusion dependencies
- Scientific data management
    - Handle new datasets
- Data analytics
    - Profiling as preparation and for initial insights
    - Borderline to data mining
- Database reverse engineering

# Basic Statistics

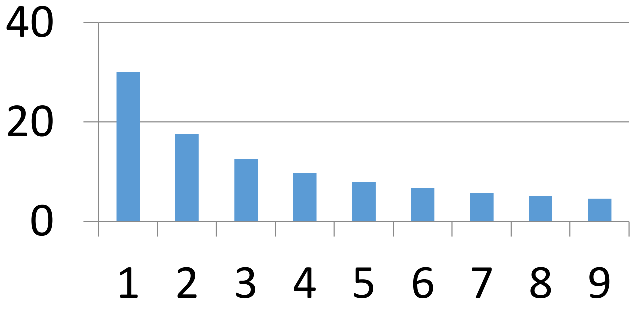
# Cardinalities, Distributions, and Patterns

| 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-widt |
| | constancy | Frequency of most frequent valu |
| | quartiles | Three points that divide the (nun |
| | soundex | Distribution of soundex codes |
| | first digit | Distribution of first digit in nume |
| **Patterns, data types, and domains** | basic type | Generic data type: numeric, alph |
| | data type | Concrete DBMS-specific data typ |
| | decimals | Maximum number of decimal pla |
| | precision | Maximum number of digits in nu |
| | patterns | Histogram of value patterns (Aa9 |
| | data class | Semantic, generic data type: coc<br>identifier, etc. |
| | domain | Classification of semantic domain: credit card, first name, city, phenotype, etc. |

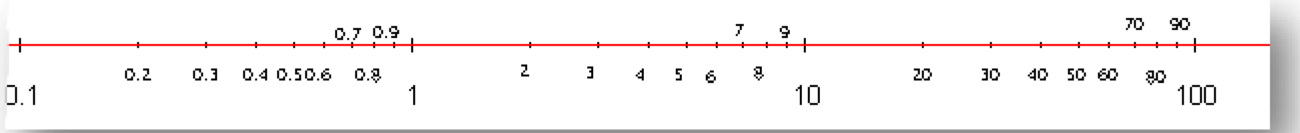


# An Aside: Benford Law Frequency (“first digit law”)

- 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



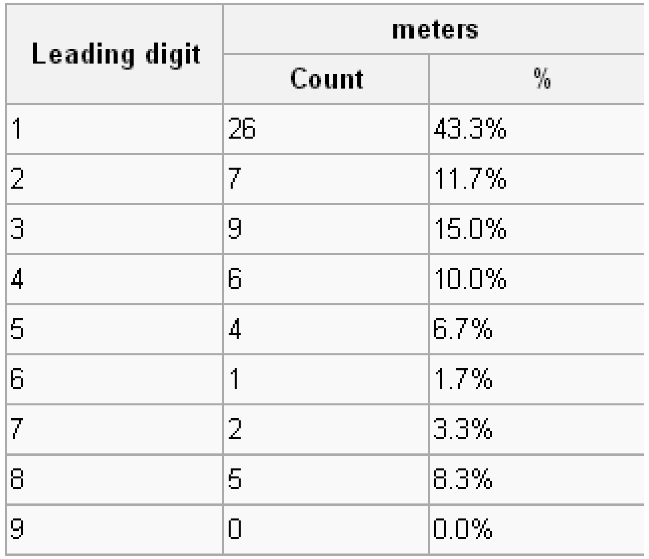
[Benford: The law of anomalous numbers". Proc. Am. Philos. Soc. 78 (4): 551–572, 1938]

# Examples for Benford‘s Law

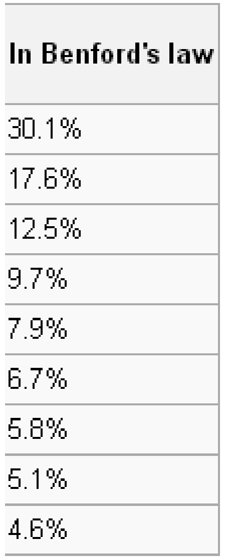
Fraud detection

- Surface areas of 335 rivers
- Sizes of 3259 US populations
- 104 physical constants
- 1800 molecular weights
- 308 numbers contained in an issue of Reader's Digest
- Street addresses of the first 342 persons listed in American Men of Science

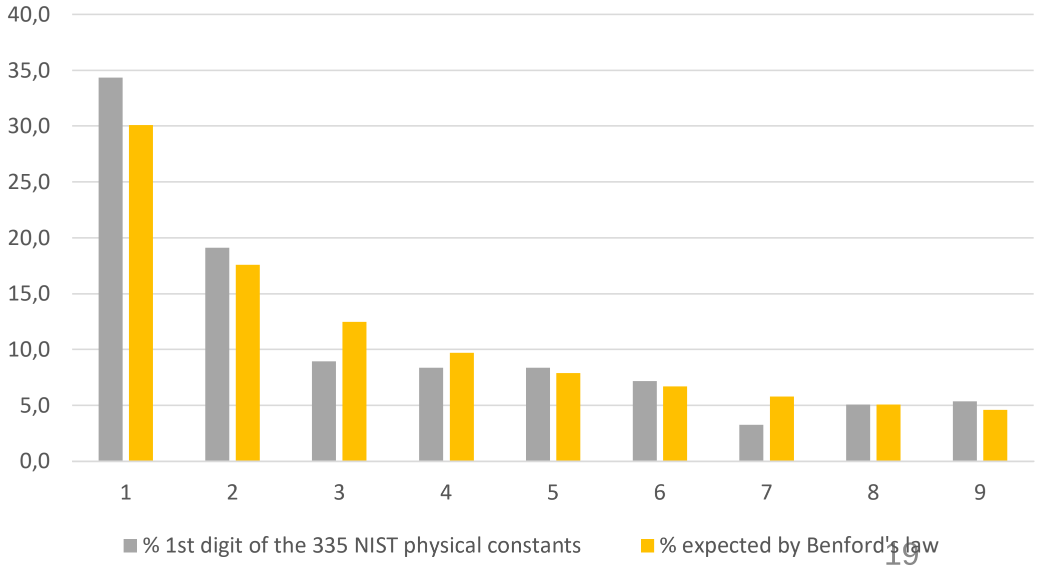
Heights of the 60 tallest structures



| Leading digit | meters | |
|---|---|---|
| | Count | % |
| 1 | 26 | 43.3% |
| 2 | 7 | 11.7% |
| 3 | 9 | 15.0% |
| 4 | 6 | 10.0% |
| 5 | 4 | 6.7% |
| 6 | 1 | 1.7% |
| 7 | 2 | 3.3% |
| 8 | 5 | 8.3% |
| 9 | 0 | 0.0% |

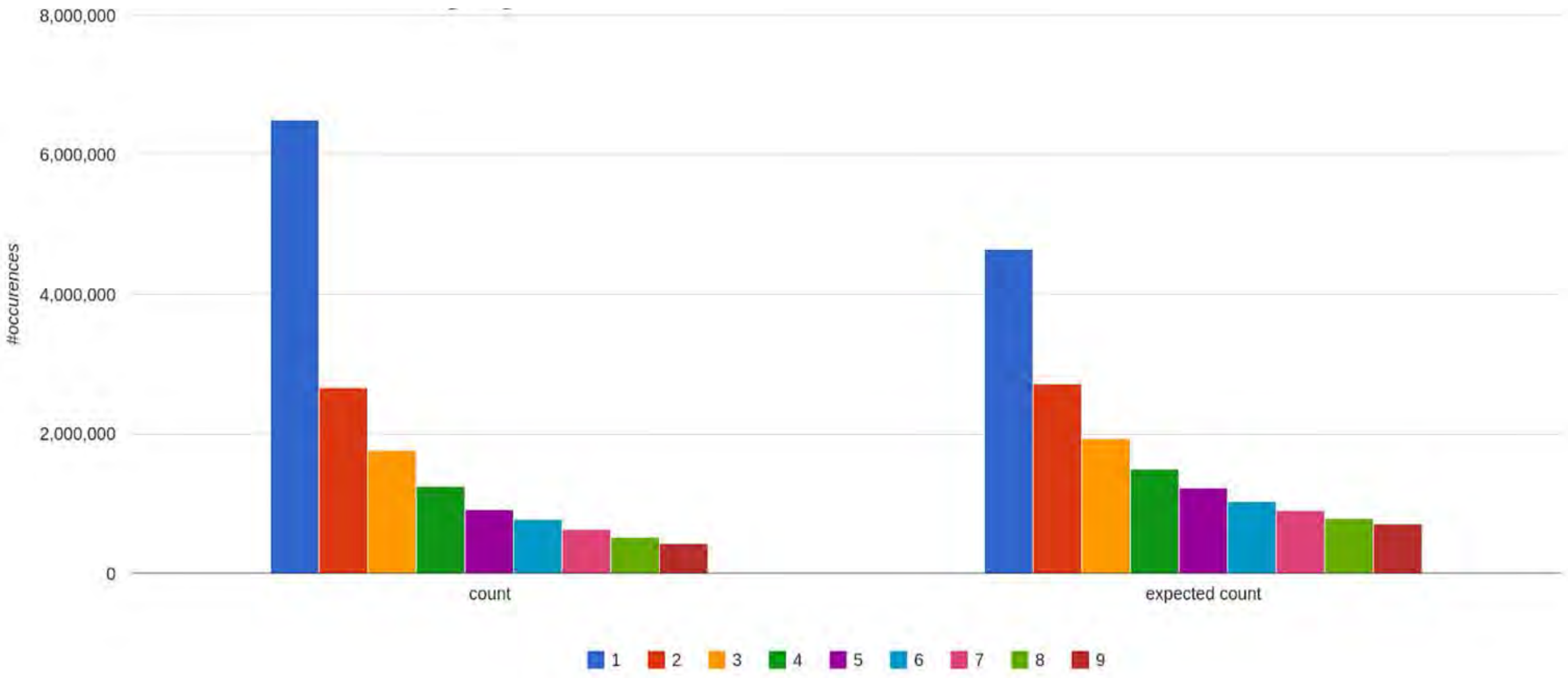


| In Benford's law |
|---|
| 30.1% |
| 17.6% |
| 12.5% |
| 9.7% |
| 7.9% |
| 6.7% |
| 5.8% |
| 5.1% |
| 4.6% |



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

# Occurrences of leading digits in WikiTable numbers



# Unique Column Combinations

# Unique Column Combinations

- Unique column
  - Only unique values
- Unique column combination
  - Only unique value combinations
  - Minimality: No subset is unique
- (Primary) key candidate
  - No null values
  - Uniqueness and non-null in one instance does not imply key: Only human can specify keys (and foreign keys)
- Meaning of NULL values?

# Uses for UCCs

- Learn characteristics of a new data set

- Database management
  - Find a primary key
  - Find unique constraints
- Query optimization
  - Cardinality estimations for joins
- Find duplicates / data quality issues
  - If expected unique column combinations are not unique
  - Or with partial uniques

# Inclusion Dependencies

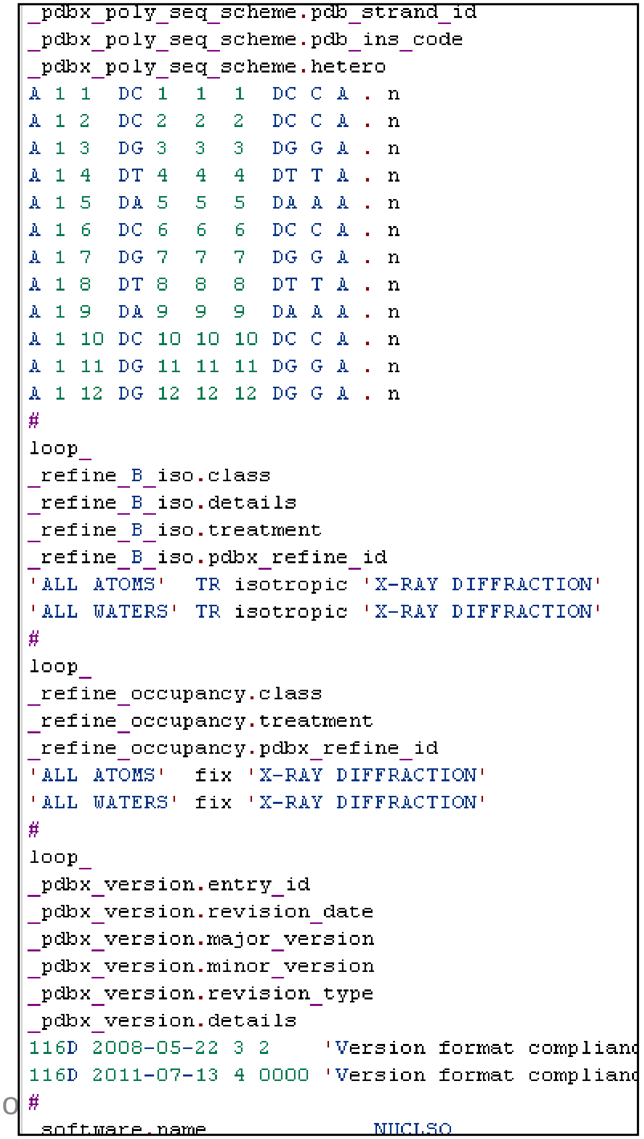
# Inclusion Dependencies

- $A \subseteq B$: All values in A are also present in B
- $A_1,\dots,A_i \subseteq B_1,\dots,B_i$:
  All value combinations in $A_1,\dots,A_i$ are also present in $B_1,\dots,B_i$

- Prerequisite for foreign key
  - Used across relations
  - Use across databases
  - But again: Discovery on a given instance, only user can specify for schema

# Motivation for IND Discovery

- General insight into data
- Detect unknown foreign keys
- Example: PDB – Protein Data Bank
  - OpenMMS provides relational schema
  - 175 tables, 2705 attributes
  - Not a single foreign key constraint!
- Example: Ensembl – genome database
  - Shipped as MySQL dump files
  - More than 200 tables
  - Not a single foreign key constraint!
- Web tables: No schema, no constraints, but many connections



```
_pdbx_poly_seq_scheme.pdb_strand_id
_pdbx_poly_seq_scheme.pdb_ins_code
_pdbx_poly_seq_scheme.hetero
A 1 1  DC 1  1  1  DC C A . n
A 1 2  DC 2  2  2  DC C A . n
A 1 3  DG 3  3  3  DG G A . n
A 1 4  DT 4  4  4  DT T A . n
A 1 5  DA 5  5  5  DA A A . n
A 1 6  DC 6  6  6  DC C A . n
A 1 7  DG 7  7  7  DG G A . n
A 1 8  DT 8  8  8  DT T A . n
A 1 9  DA 9  9  9  DA A A . n
A 1 10 DC 10 10 10 DC C A . n
A 1 11 DG 11 11 11 DG G A . n
A 1 12 DG 12 12 12 DG G A . n
#
loop_
_refine_B_iso.class
_refine_B_iso.details
_refine_B_iso.treatment
_refine_B_iso.pdbx_refine_id
'ALL ATOMS'  TR isotropic 'X-RAY DIFFRACTION'
'ALL WATERS' TR isotropic 'X-RAY DIFFRACTION'
#
loop_
_refine_occupancy.class
_refine_occupancy.treatment
_refine_occupancy.pdbx_refine_id
'ALL ATOMS'  fix 'X-RAY DIFFRACTION'
'ALL WATERS' fix 'X-RAY DIFFRACTION'
#
loop_
_pdbx_version.entry_id
_pdbx_version.revision_date
_pdbx_version.major_version
_pdbx_version.minor_version
_pdbx_version.revision_type
_pdbx_version.details
116D 2008-05-22 3 2    'Version format complianc
116D 2011-07-13 4 0000 'Version format complianc
#
_software.name             NUCLSQ
```

# Functional and other dependencies

# Functional and Other Dependencies

- Functional dependency
  - „X → A“: whenever two records have the same X values, they also have the same A values.
- Multi-valued dependencies
  - Join dependencies
- Order dependencies

  -
    ```
    SELECT emp_name
    FROM employees
    ORDER BY rank, salary
    ```

  -
    ```
    SELECT emp_name
    FROM employees
    ORDER BY rank
    ```

**salary** orders **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 |

# Uses for FDs

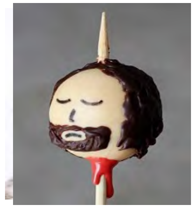
- Schema design
    - Normalization
    - Keys
- Data cleansing
- Schema design and normalization
- Key discovery

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

… and genealogy research!

# 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.“

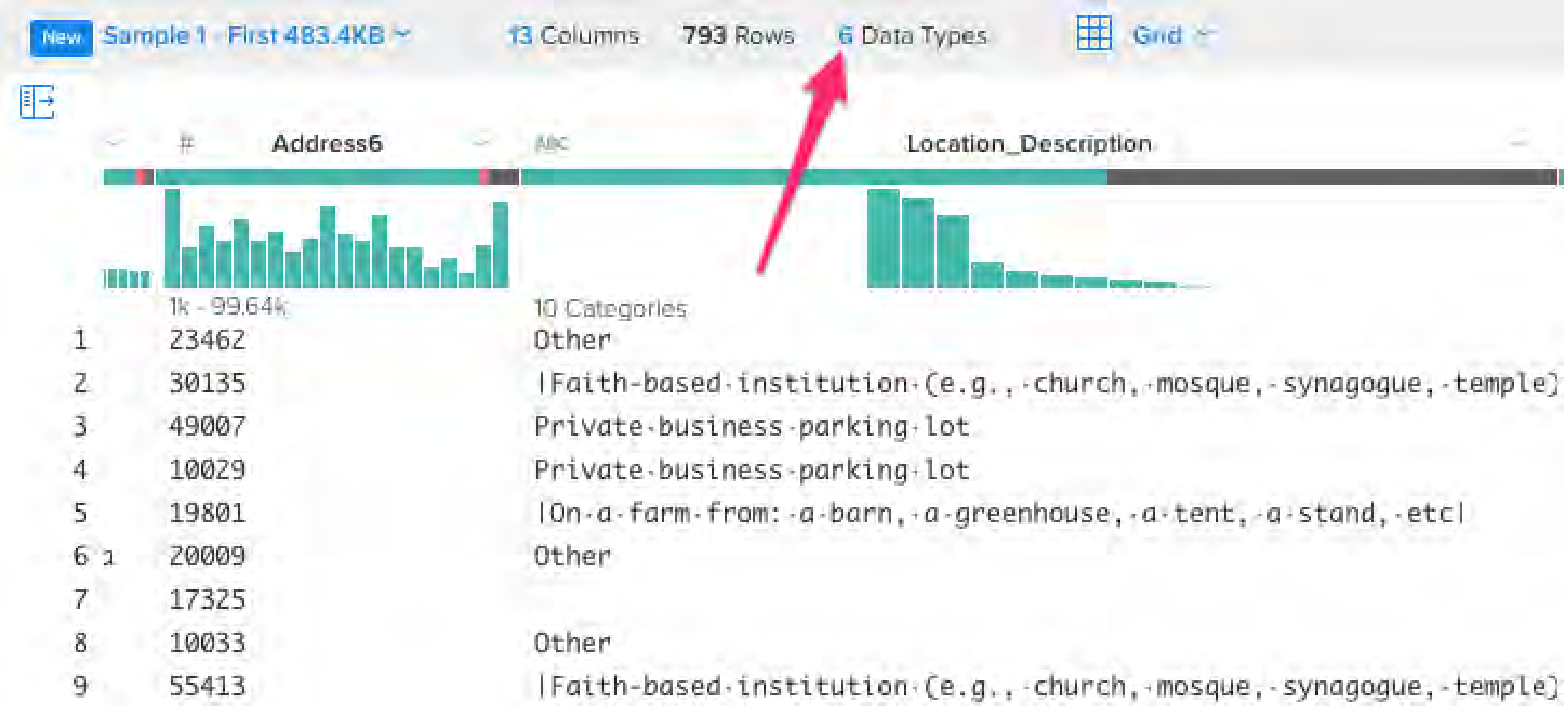
next slide deck

# Tutorial Overview

- Motivation
  - Task classification
  - Use cases
- Tools
  - Research and industry
  - Shortcomings
- Single and Multiple Column Analysis
  - Cardinalities and datatypes
  - Co-occurrences and summaries
- Dependencies
  - UCCs, FDs, ODs, INDs
  - and their discovery algorithms
- Outlook
  - Functionality
  - Semantics

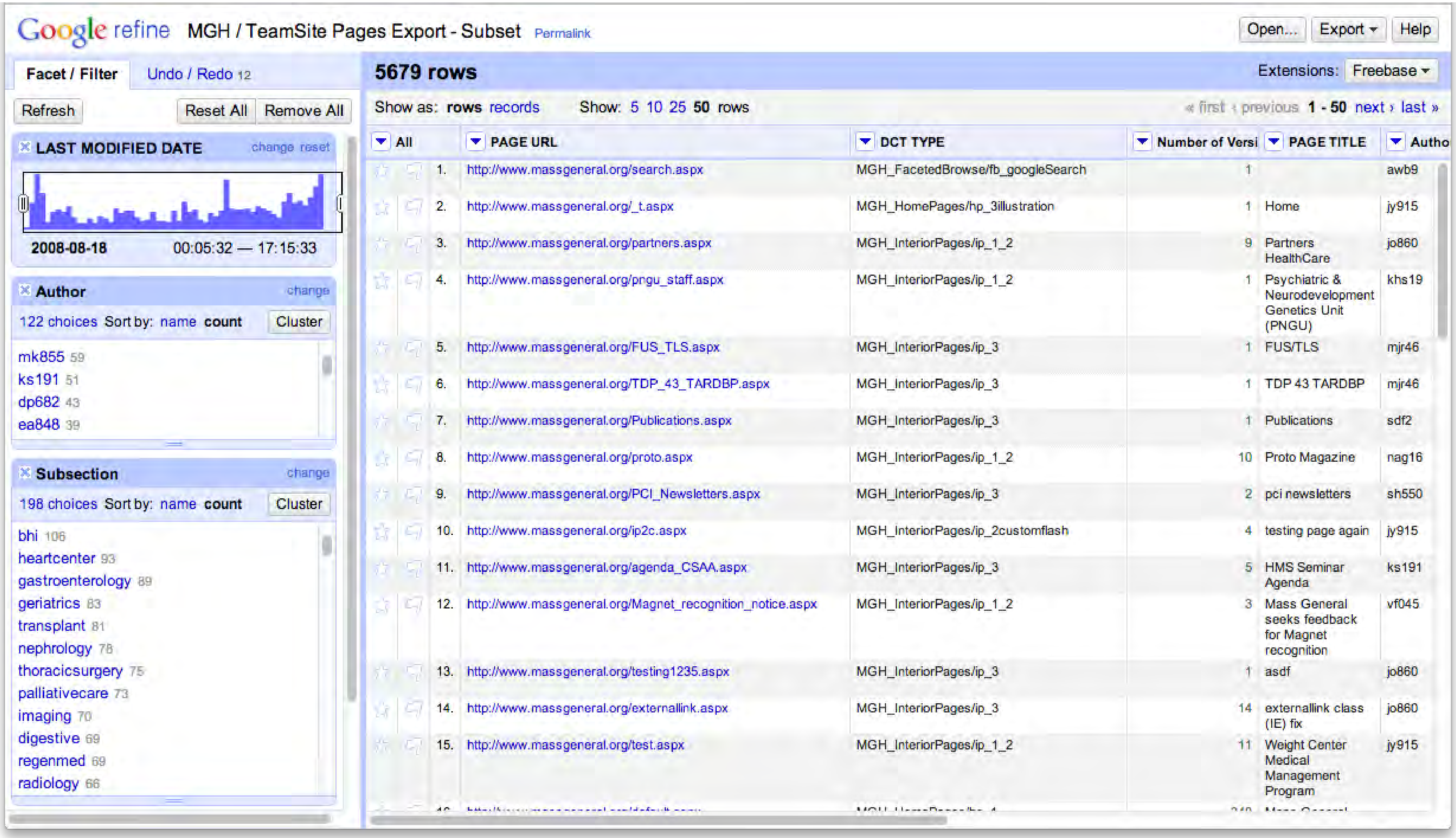
# Tools in Industry

# Trifacta



|  | Address6 | Location_Description |
|---|---|---|
| 1 | 23462 | Other |
| 2 | 30135 | \|Faith-based·institution·(e.g.,·church,·mosque,·synagogue,·temple) |
| 3 | 49007 | Private·business·parking·lot |
| 4 | 10029 | Private·business·parking·lot |
| 5 | 19801 | \|On·a·farm·from:·a·barn,·a·greenhouse,·a·tent,·a·stand,·etc\| |
| 6 ı | 20009 | Other |
| 7 | 17325 |  |
| 8 | 10033 | Other |
| 9 | 55413 | \|Faith-based·institution·(e.g.,·church,·mosque,·synagogue,·temple) |

# Open Refine



Google refine MGH / TeamSite Pages Export - Subset Permalink

Open... Export ▾ Help

Facet / Filter   Undo / Redo 12

Refresh   Reset All   Remove All

LAST MODIFIED DATE   change reset

2008-08-18   00:05:32 — 17:15:33

Author   change

122 choices Sort by: name count   Cluster

| | |
|---|---|
| mk855 | 59 |
| ks191 | 51 |
| dp682 | 43 |
| ea848 | 39 |

Subsection   change

198 choices Sort by: name count   Cluster

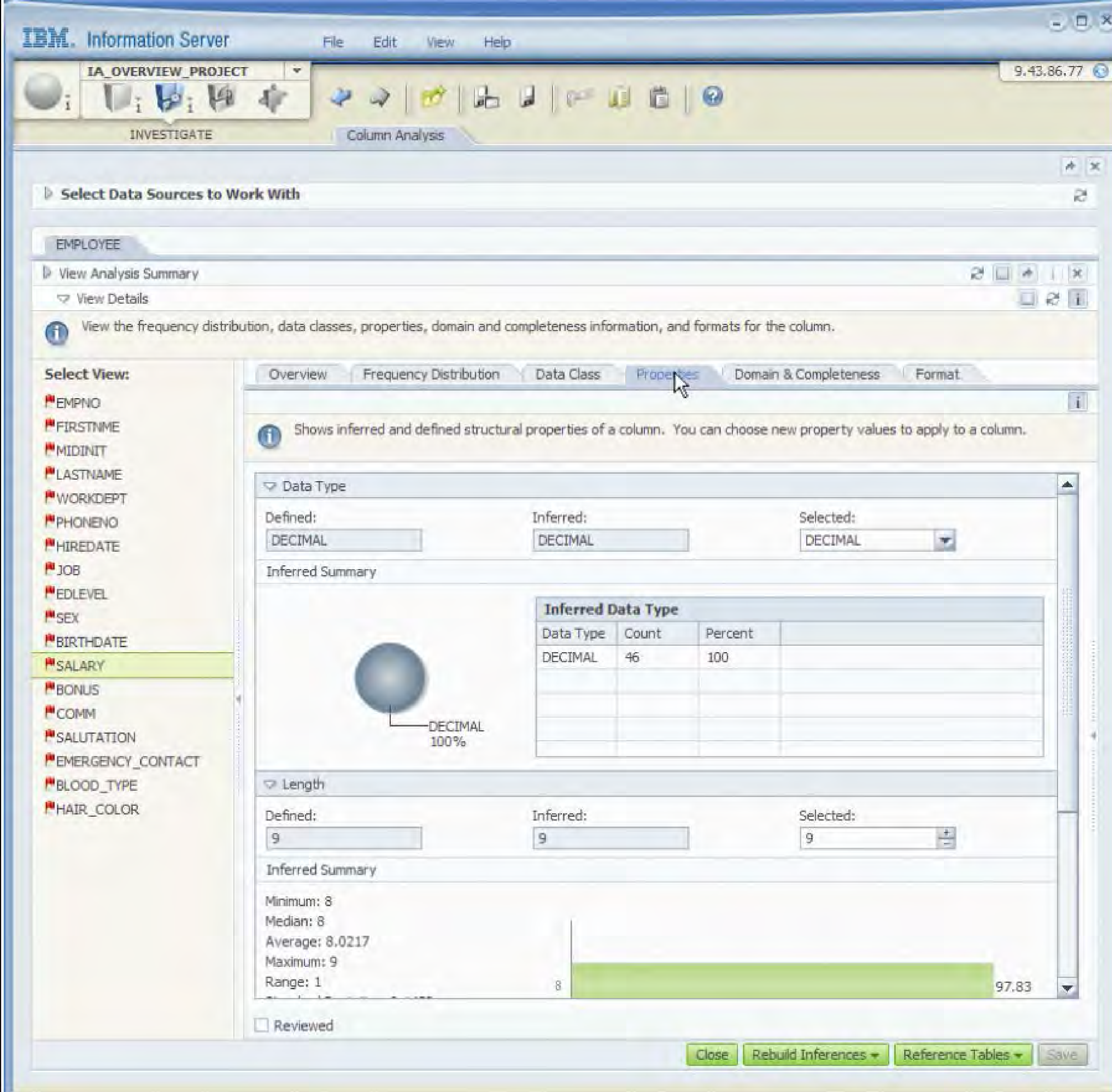
| | |
|---|---|
| bhi | 106 |
| heartcenter | 93 |
| gastroenterology | 89 |
| geriatrics | 83 |
| transplant | 81 |
| nephrology | 78 |
| thoracicsurgery | 75 |
| palliativecare | 73 |
| imaging | 70 |
| digestive | 69 |
| regenmed | 69 |
| radiology | 66 |

5679 rows   Extensions: Freebase ▾

Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 50 next › last »

| All | PAGE URL | DCT TYPE | Number of Versi | PAGE TITLE | Autho |
|---|---|---|---|---|---|
| 1. | http://www.massgeneral.org/search.aspx | MGH_FacetedBrowse/fb_googleSearch | 1 | | awb9 |
| 2. | http://www.massgeneral.org/_t.aspx | MGH_HomePages/hp_3illustration | 1 | Home | jy915 |
| 3. | http://www.massgeneral.org/partners.aspx | MGH_InteriorPages/ip_1_2 | 9 | Partners HealthCare | jo860 |
| 4. | http://www.massgeneral.org/pngu_staff.aspx | MGH_InteriorPages/ip_1_2 | 1 | Psychiatric & Neurodevelopment Genetics Unit (PNGU) | khs19 |
| 5. | http://www.massgeneral.org/FUS_TLS.aspx | MGH_InteriorPages/ip_3 | 1 | FUS/TLS | mjr46 |
| 6. | http://www.massgeneral.org/TDP_43_TARDBP.aspx | MGH_InteriorPages/ip_3 | 1 | TDP 43 TARDBP | mjr46 |
| 7. | http://www.massgeneral.org/Publications.aspx | MGH_InteriorPages/ip_3 | 1 | Publications | sdf2 |
| 8. | http://www.massgeneral.org/proto.aspx | MGH_InteriorPages/ip_1_2 | 10 | Proto Magazine | nag16 |
| 9. | http://www.massgeneral.org/PCI_Newsletters.aspx | MGH_InteriorPages/ip_3 | 2 | pci newsletters | sh550 |
| 10. | http://www.massgeneral.org/ip2c.aspx | MGH_InteriorPages/ip_2customflash | 4 | testing page again | jy915 |
| 11. | http://www.massgeneral.org/agenda_CSAA.aspx | MGH_InteriorPages/ip_3 | 5 | HMS Seminar Agenda | ks191 |
| 12. | http://www.massgeneral.org/Magnet_recognition_notice.aspx | MGH_InteriorPages/ip_1_2 | 3 | Mass General seeks feedback for Magnet recognition | vf045 |
| 13. | http://www.massgeneral.org/testing1235.aspx | MGH_InteriorPages/ip_3 | 1 | asdf | jo860 |
| 14. | http://www.massgeneral.org/externallink.aspx | MGH_InteriorPages/ip_3 | 14 | externallink class (IE) fix | jo860 |
| 15. | http://www.massgeneral.org/test.aspx | MGH_InteriorPages/ip_1_2 | 11 | Weight Center Medical Management Program | jy915 |

# IBM Information Analyzer



Select Data Sources to Work With

EMPLOYEE

View Analysis Summary

View Details

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

**Select View:**

- EMPNO
- FIRSTNME
- MIDINIT
- LASTNAME
- WORKDEPT
- PHONENO
- HIREDATE
- JOB
- EDLEVEL
- SEX
- BIRTHDATE
- SALARY
- BONUS
- COMM
- SALUTATION
- EMERGENCY_CONTACT
- BLOOD_TYPE
- HAIR_COLOR

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

Data Type

Defined: DECIMAL
Inferred: DECIMAL
Selected: DECIMAL

Inferred Summary

DECIMAL 100%

**Inferred Data Type**

| Data Type | Count | Percent |
|---|---|---|
| DECIMAL | 46 | 100 |

Length

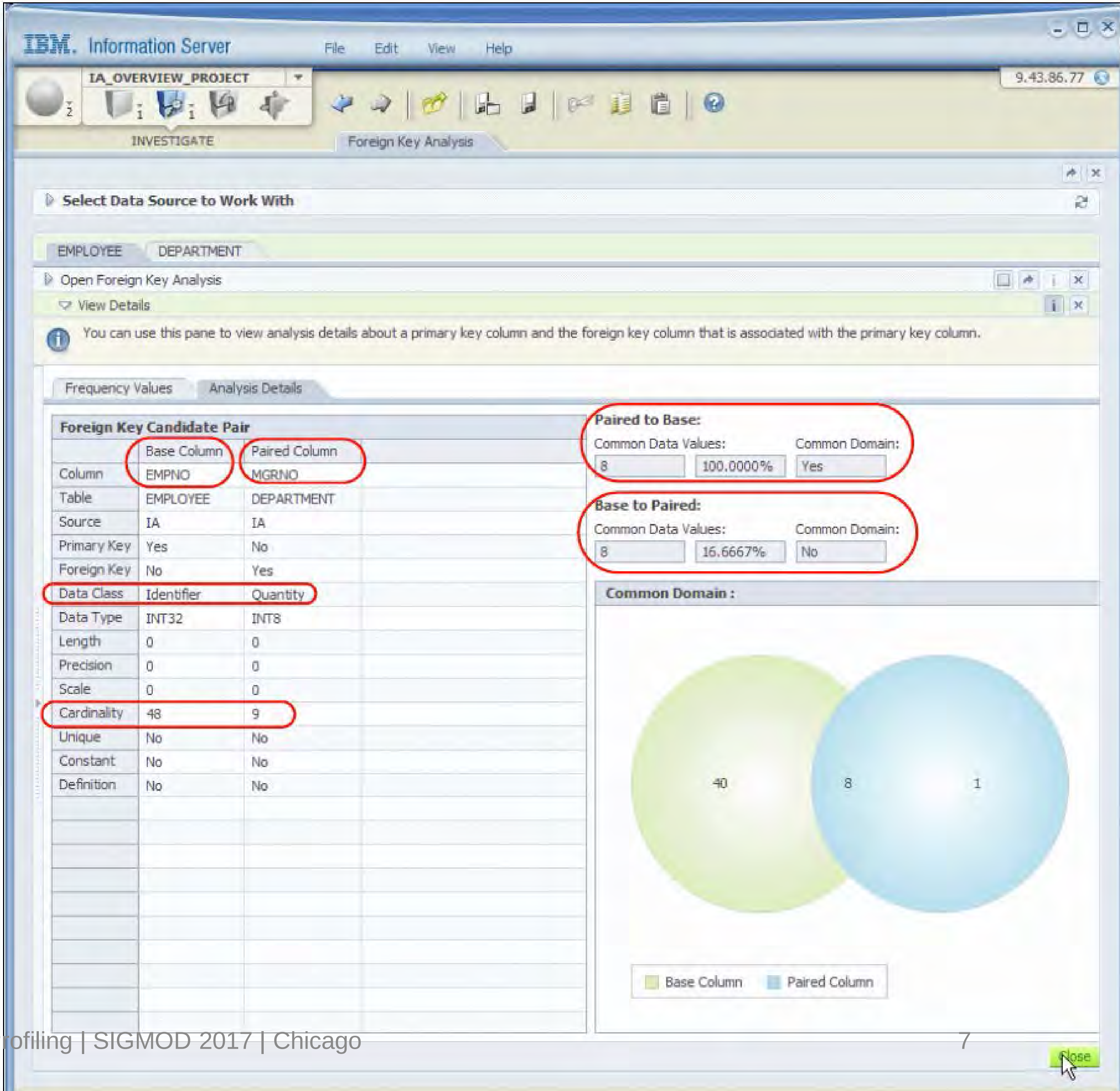
Defined: 9
Inferred: 9
Selected: 9

Inferred Summary

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

□ Reviewed

# IBM Information Analyzer



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.

**Foreign Key Candidate Pair**

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

**Paired to Base:**

Common Data Values: 8 100.0000%  Common Domain: Yes

**Base to Paired:**

Common Data Values: 8 16.6667%  Common Domain: No

**Common Domain :**

🟩 Base Column 🟦 Paired Column

# Uses Cases Covered By Industrial Tools

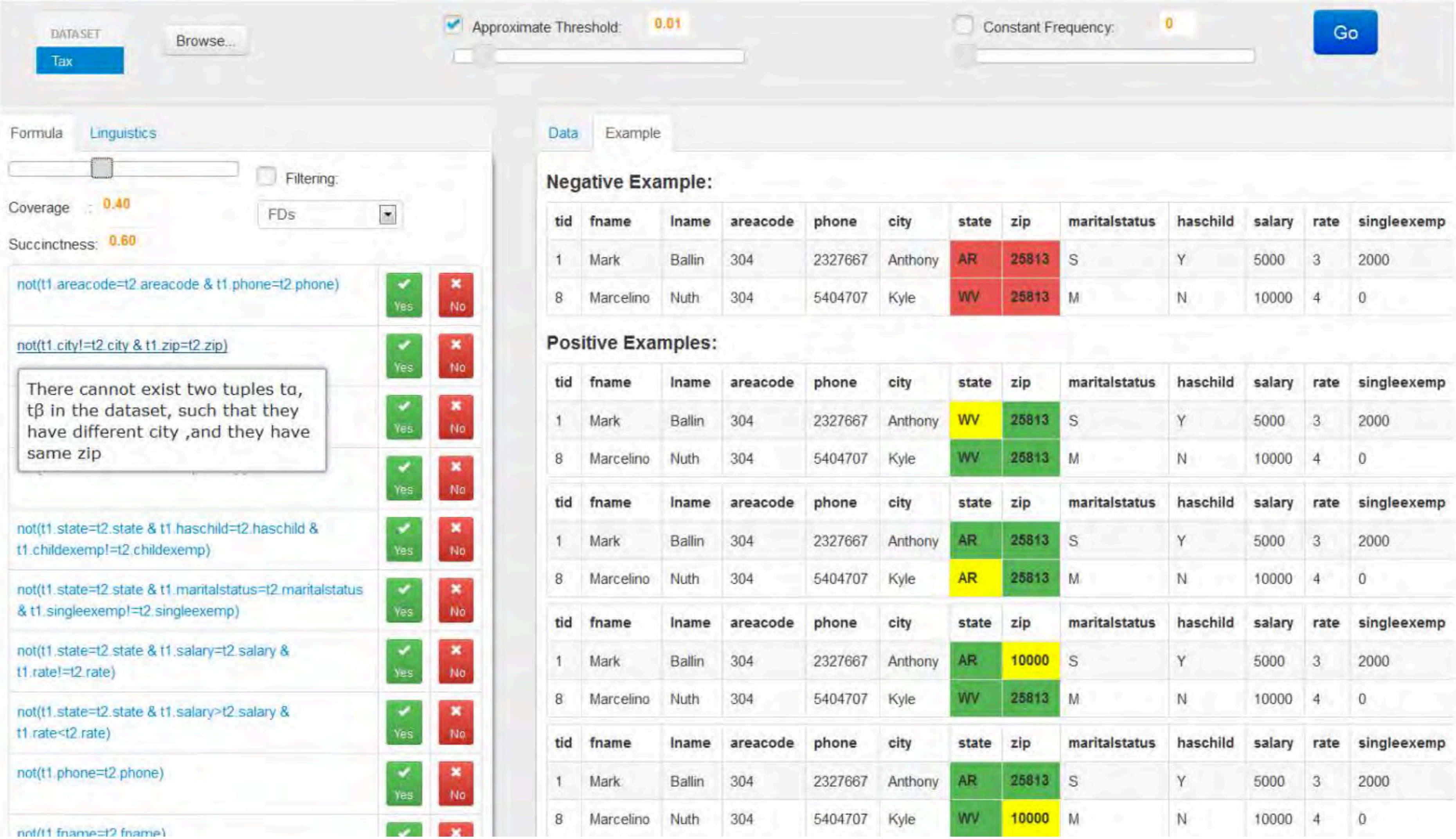
| Tool | Statistics | Patterns | Data types | Uniques | Column dependency | Data dependency |
|---|---|---|---|---|---|---|
| **Attacama**, DQ Analyzer | ✓ | ✓ |  | ✓ |  |  |
| **IBM**, InfoSphere Information Analyzer | ✓ | ✓ |  | ✓ | ✓ |  |
| **Microsoft** SQL Server Data Profiling Task | ✓ | ✓ |  |  | ✓ |  |
| **Oracle** Enterprise Data Quality | ✓ | ✓ |  |  |  |  |
| **Paxata** Adaptive Preparation | ✓ |  |  |  |  |  |
| **SAP** Information Steward | ✓ | ✓ | ✓ |  | ✓ |  |
| **Splunk** Enterprise/Hunk |  | ✓ |  |  |  | ✓ |
| **Talend** Data Profiler | ✓ | ✓ |  |  | ✓ |  |
| **Trifacta** | ✓ | ✓ | ✓ |  |  |  |
| **Tamr** | ✓ |  |  | ✓ |  |  |
| **OpenRefine** | ✓ | ✓ | ✓ |  |  |  |

Restricted data types

Restricted number of columns

# Tools in Research

# RuleMiner



☑ Approximate Threshold: 0.01

□ Constant Frequency: 0

□ Filtering: FDs

Coverage : 0.40

Succinctness: 0.60

- not(t1.areacode=t2.areacode & t1.phone=t2.phone) ✓ Yes ✗ No
- not(t1.city!=t2.city & t1.zip=t2.zip) ✓ Yes ✗ No
  There cannot exist two tuples tα, tβ in the dataset, such that they have different city ,and they have same zip
- ✓ Yes ✗ No
- ✓ Yes ✗ No
- not(t1.state=t2.state & t1.haschild=t2.haschild & t1.childexemp!=t2.childexemp) ✓ Yes ✗ No
- not(t1.state=t2.state & t1.maritalstatus=t2.maritalstatus & t1.singleexemp!=t2.singleexemp) ✓ Yes ✗ No
- not(t1.state=t2.state & t1.salary=t2.salary & t1.rate!=t2.rate) ✓ Yes ✗ No
- not(t1.state=t2.state & t1.salary>t2.salary & t1.rate<t2.rate) ✓ Yes ✗ No
- not(t1.phone=t2.phone) ✓ Yes ✗ No
- not(t1.fname=t2.fname)

**Negative Example:**

| tid | fname | lname | areacode | phone | city | state | zip | maritalstatus | haschild | salary | rate | singleexemp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Mark | Ballin | 304 | 2327667 | Anthony | AR | 25813 | S | Y | 5000 | 3 | 2000 |
| 8 | Marcelino | Nuth | 304 | 5404707 | Kyle | WV | 25813 | M | N | 10000 | 4 | 0 |

**Positive Examples:**

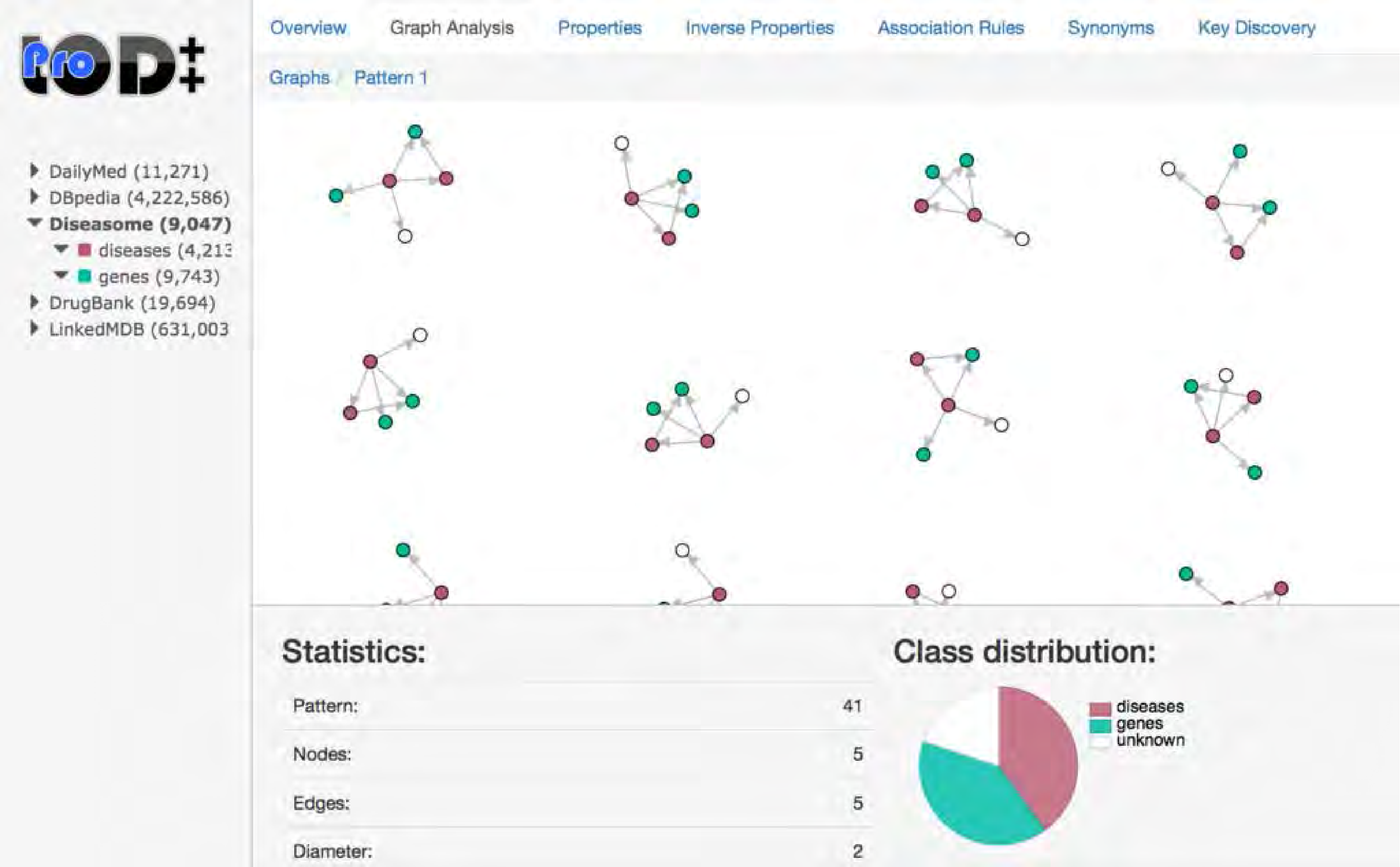
| tid | fname | lname | areacode | phone | city | state | zip | maritalstatus | haschild | salary | rate | singleexemp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Mark | Ballin | 304 | 2327667 | Anthony | WV | 25813 | S | Y | 5000 | 3 | 2000 |
| 8 | Marcelino | Nuth | 304 | 5404707 | Kyle | WV | 25813 | M | N | 10000 | 4 | 0 |

| tid | fname | lname | areacode | phone | city | state | zip | maritalstatus | haschild | salary | rate | singleexemp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Mark | Ballin | 304 | 2327667 | Anthony | AR | 25813 | S | Y | 5000 | 3 | 2000 |
| 8 | Marcelino | Nuth | 304 | 5404707 | Kyle | AR | 25813 | M | N | 10000 | 4 | 0 |

| tid | fname | lname | areacode | phone | city | state | zip | maritalstatus | haschild | salary | rate | singleexemp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Mark | Ballin | 304 | 2327667 | Anthony | AR | 10000 | S | Y | 5000 | 3 | 2000 |
| 8 | Marcelino | Nuth | 304 | 5404707 | Kyle | WV | 25813 | M | N | 10000 | 4 | 0 |

| tid | fname | lname | areacode | phone | city | state | zip | maritalstatus | haschild | salary | rate | singleexemp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Mark | Ballin | 304 | 2327667 | Anthony | AR | 25813 | S | Y | 5000 | 3 | 2000 |
| 8 | Marcelino | Nuth | 304 | 5404707 | Kyle | WV | 10000 | M | N | 10000 | 4 | 0 |

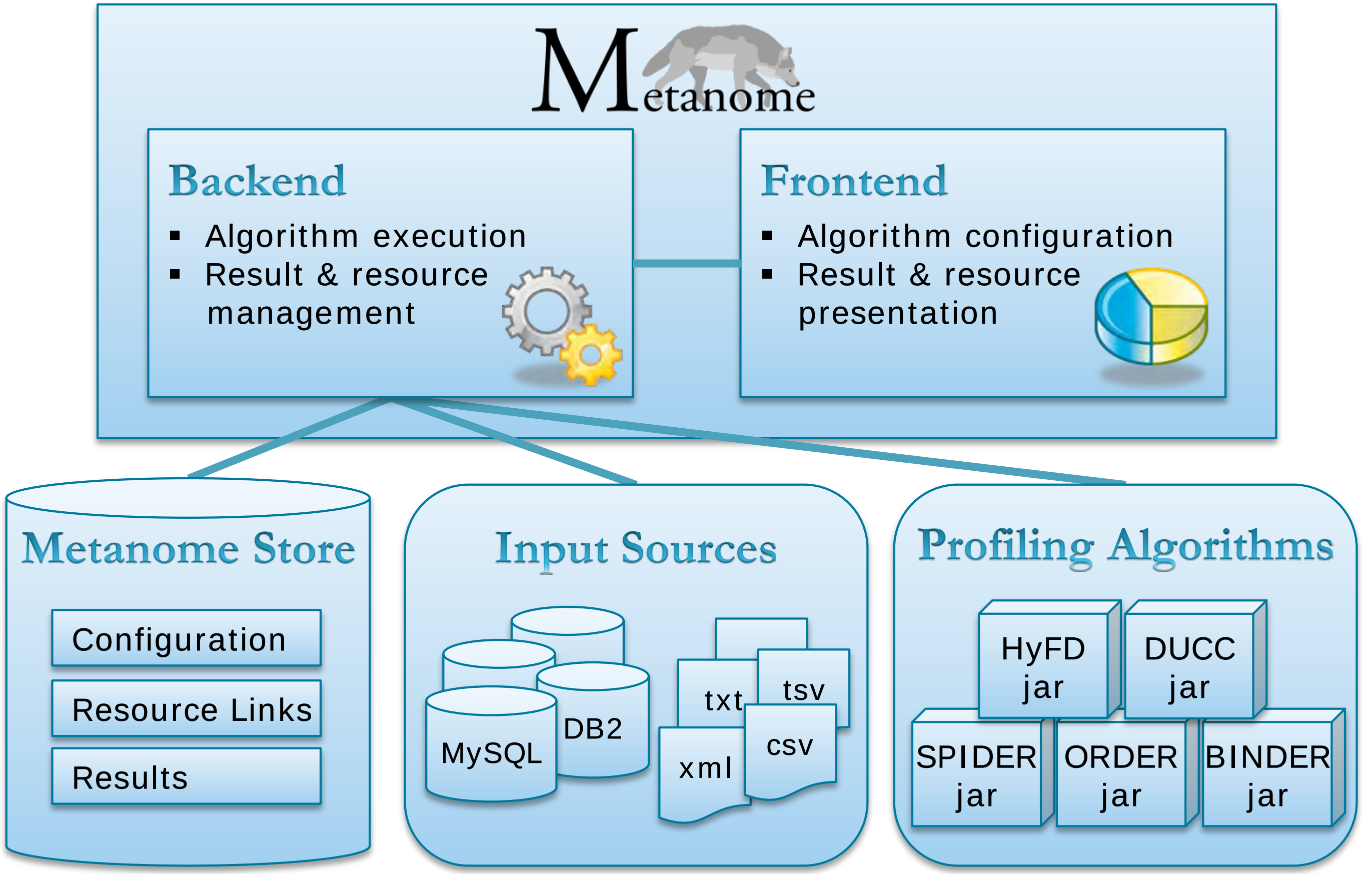
# ProLOD++



Statistics:

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

# Metanome Data Profiling Tool



Metanome

Backend
- Algorithm execution
- Result & resource management

Frontend
- Algorithm configuration
- Result & resource presentation

Metanome Store
- Configuration
- Resource Links
- Results

Input Sources

Profiling Algorithms

Open source framework, tool plus many algorithms

**www.metanome.de**

# Tools in Research

| Tool | Main purpose | Statistics | Patterns | Data types | Uniques | Dependencies | Data Mining |
|---|---|---|---|---|---|---|---|
| **Bellmann** | Data quality browser | ✓ | | | ✓ | | |
| **Potter's Wheel** | ETL tool | ✓ | ✓ | | | | |
| **Data Auditor** | Rule discovery | | | | | | |
| **RuleMiner** | Dependency discovery | | | | | ✓ | |
| **MADLib** | Machine learning | ✓ | | | | ✓ | |
| **Metanome** | Data profiling | ✓ | | | ✓ | ✓ | |
| **ProLOD++** | Profiling and Mining | ✓ | ✓ | | ✓ | ✓ | ✓ |

# Typical Shortcomings

- Usability
- Tools focus on “easy” problems:
  - Statistics
  - Single column or “few” column dependencies
  - „Checking“ vs. „discovery“
- Many industry tools use SQL instead of optimized algorithms
  - Many queries / no early abort
- No tool covers all types of meta-data
- Management of large meta-data results
  - Summarizing meta-data
  - Ranking meta-data based on relevance

next slide deck

# Tutorial Overview

- Motivation
  - Task classification
  - Use cases
- Tools
  - Research and industry
  - Shortcomings
- Single and Multiple Column Analysis
  - Cardinalities and datatypes
  - Co-occurrences and summaries
- Dependencies
  - UCCs, FDs, ODs, INDs
  - and their discovery algorithms
- Outlook
  - Functionality
  - Semantics

# Single Column Analysis

# Cardinalities and distributions

- Number of non-NULL values
- Number of distinct values

Count(*)
count(distinct X)

- MIN and MAX values

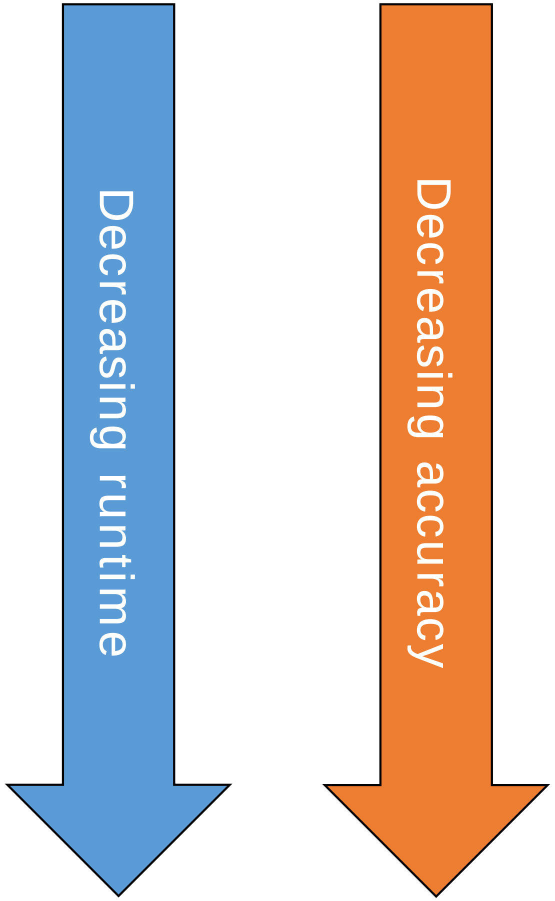
```
For (value in column)
If (value>max)
max=value
```

- Histograms
- Probability distribution for numeric values
- Detect whether data follows some distribution
  - And count the number of outliers

Bottleneck is sorting the data

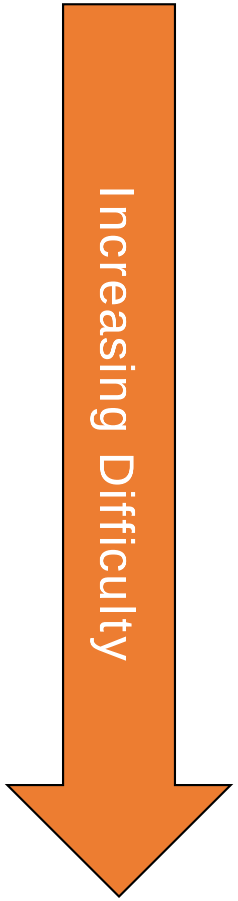
# Count distinct in sublinear time and space?

- Linear Counting
  - [Whang, Vander-Zanden, Taylor: A linear-time probabilistic counting algorithm for database applications. TODS, 1990]
- Stochastic Averaging
  - [Flajolet, Martin: Probabilistic counting algorithms for data base applications. JCSS, 1985]
- Loglog Algorithm
  - [Durand, Flajolet: Loglog counting of large cardinalities. Algorithms-ESA, 2003]
- SuperLogLog Algorithm
  - [Durand, Flajolet: Loglog counting of large cardinalities. Algorithms-ESA, 2003]
- HyperLogLog Algorithm
  - [Flajolet, Fusy, Gandouet, Meunier: Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. DMTCS, 2008]



# Data types and value patterns

- String vs. number
- String vs. number vs. date
- Categorical vs. continuous
  - Days of the week vs. measurements
- SQL data types
  - CHAR, INT, DECIMAL, TIMESTAMP, BIT, CLOB, …
- Domains
  - VARCHAR(12) vs. VARCHAR (13)
- XML data types
  - More fine grained
- Regular expressions  (\d{3})-(\d{3})-(\d{4})-(\d+)
- Semantic domains
  - Address, phone, email, first name
  - Example of ambiguity: phone vs fax



Increasing Difficulty

# Multi Column Analysis

# Pairwise Correlation/Similarity

- Correlations between numeric columns

- Similarity between discrete columns
  - Jaccard similarity of two sets = the size of their intersection divided by the size of their union
  - Careful with strings: phone numbers 123 456 7890 vs. (123) 456-7890
    - May want to use n-grams

# Sketches and Summaries

- Assess column similarity in big (tall and wide) data
  - Want to avoid $N^2$ pairwise comparisons and multiple big table scans
- Techniques:
  - Sampling
  - Hashing:
    - Minhash [Broder: Compression and Complexity of Sequences, 1997]
    - LSH [Gionis, Indyk, Motwani: Similarity search in high Dimensions via hashing, VLDB'99]
  - Sketches [Cormode, Garofalakis, Haas, Jermaine: Synopses for Massive Data:Samples, Histograms, Wavelets, Sketches, FTD'12]

# Column Similarity:
# Jaccard(C1,C2) = intersect(C1,C2)/Union(C1,C2)

- Reduce dimension through Minhash:
  - Find a hash function ***h(·)*** such that:
    - If ***sim*($C_1$,$C_2$)** is high, then with high prob. ***h*($C_1$) = *h*($C_2$)**
    - If ***sim*($C_1$,$C_2$)** is low, then with high prob. ***h*($C_1$) ≠ *h*($C_2$)**
    - Estimate similarity by applying ***k*** different ***$h_i$(·)***
  - Transform table into a Boolean matrix

| Residence (A) | Country (B) | Birthplace (C) |
|---|---|---|
| Chicago | USA | New York |
| New York | Germany | Toronto |
| Berlin | Canada | Chicago |

| Values | A | B | C |
|---|---|---|---|
| Chicago | 1 | 0 | 1 |
| New York | 1 | 0 | 1 |
| Berlin | 1 | 0 | 0 |
| USA | 0 | 1 | 0 |
| Germany | 0 | 1 | 0 |
| Canada | 0 | 1 | 0 |
| Toronto | 0 | 0 | 1 |

# Minhash Example

- Simulate hash through permutation of row numbers
- Pick smallest row number where matrix value equals 1

| **Values** | **A** | **B** | **C** |
|---|---|---|---|
| Chicago | 1 | 0 | 1 |
| New York | 1 | 0 | 1 |
| Berlin | 1 | 0 | 0 |
| USA | 0 | 1 | 0 |
| Germany | 0 | 1 | 0 |
| Canada | 0 | 1 | 0 |
| Toronto | 0 | 0 | 1 |

| h1 | h2 | h3 |
|---|---|---|
| 1 | 7 | 5 |
| 2 | 4 | 6 |
| 3 | 1 | 7 |
| 4 | 5 | 2 |
| 5 | 3 | 3 |
| 6 | 6 | 4 |
| 7 | 2 | 1 |

| **Hash** | **A** | **B** | **C** |
|---|---|---|---|
| h1 | 1 | 4 | 1 |
| h2 | 1 | 3 | 2 |
| h3 | 5 | 2 | 1 |

sim(A,B)= 0

sim(A,C)= 0.33

sim(B,C)= 0

# Single & Multi-Column Analysis

- Cardinalities
- Data types
- Patterns
- Column similarity
- Sketches, summaries
- ….
- Overlap with data mining
- Most techniques:
  - Not very complex but approximations needed for big data/streaming data

next slide deck

# Tutorial Overview

- Motivation
  - Task classification
  - Use cases
- Tools
  - Research and industry
  - Shortcomings
- Single and Multiple Column Analysis
  - Cardinalities and datatypes
  - Co-occurrences and summaries
- Dependencies
  - UCCs, FDs, ODs, INDs
  - and their discovery algorithms
- Outlook
  - Functionality
  - Semantics

# UNIQUE

JUST BECAUSE YOU ARE UNIQUE DOES NOT MEAN YOU ARE USEFUL

# Applications

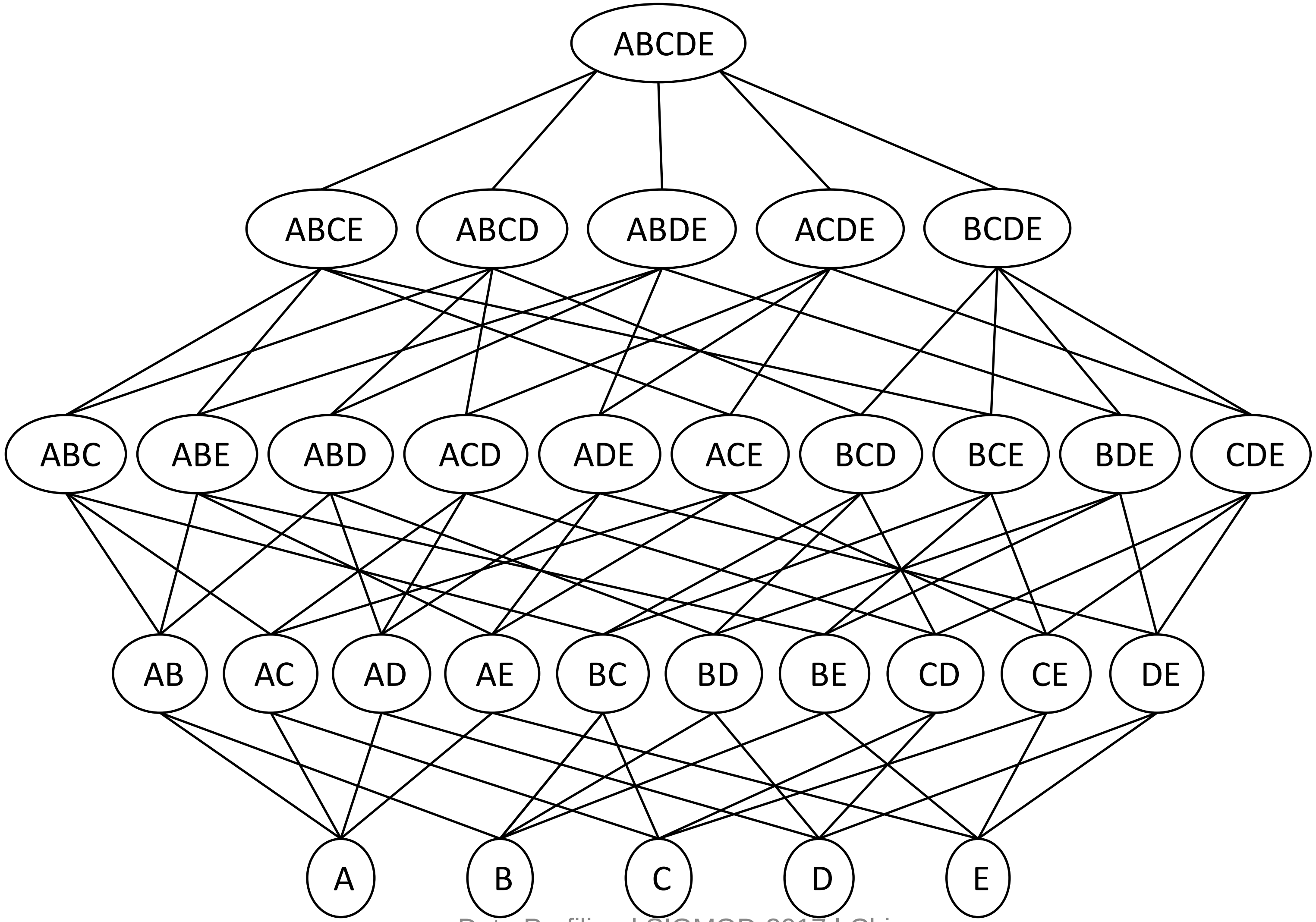
- Learn characteristics of a new data set

- Database management
  - Find candidate keys

- Query optimization
  - Cardinality estimations for joins

- Find duplicates / data quality issues
  - If expected unique column combinations are not unique

# Search Space: Attribute Lattice



$$\binom{5}{5} = 1$$

$$\binom{5}{4} = 5$$

$$\binom{5}{3} = \frac{5 \cdot 4}{2}$$

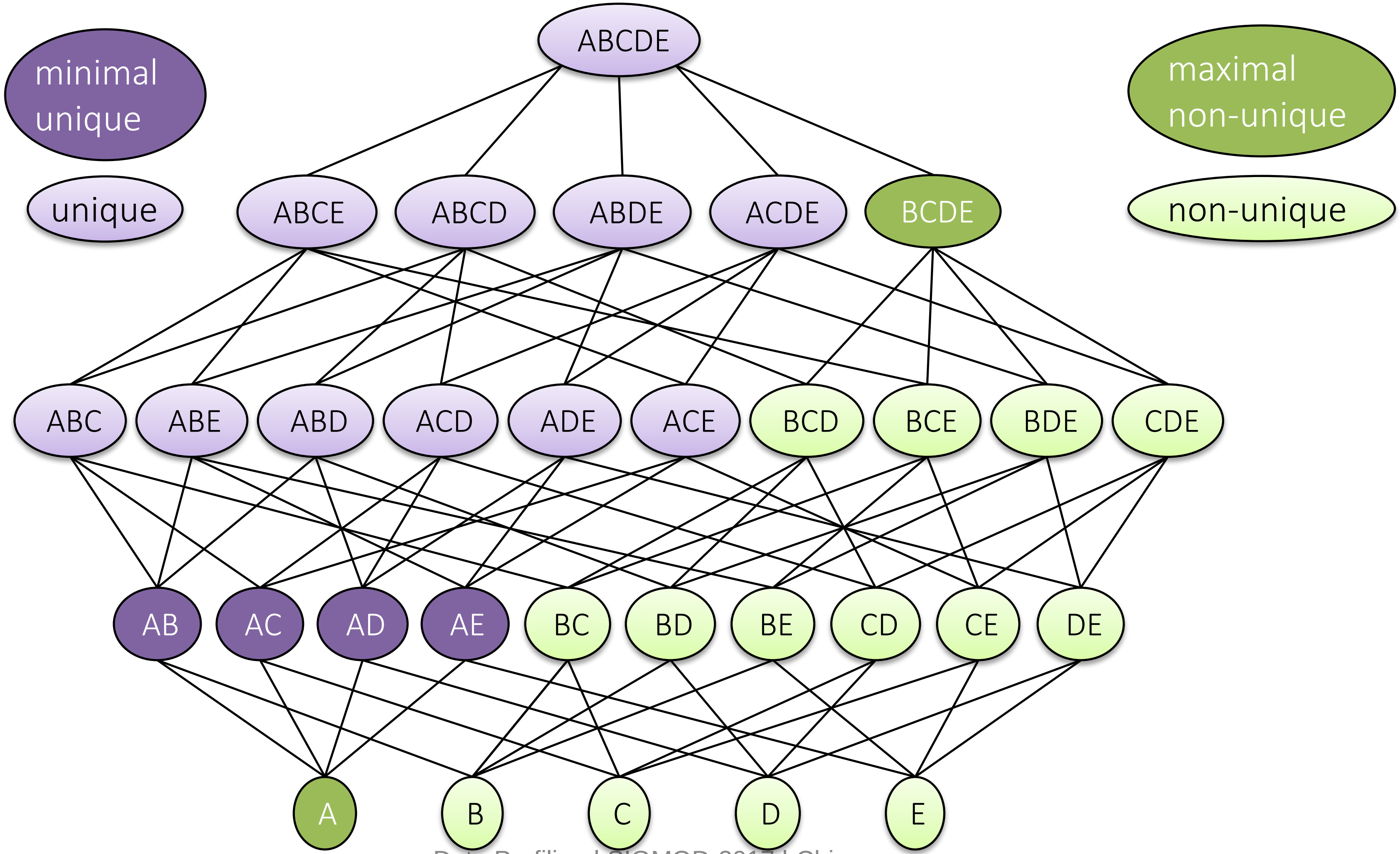
$$\binom{5}{2} = \frac{5 \cdot 4 \cdot 3}{2 \cdot 3}$$

$$\binom{5}{1} = \frac{5 \cdot 4 \cdot 3 \cdot 2}{2 \cdot 3 \cdot 4}$$

# Complexity

- For a lattice over n columns
  - $\binom{n}{k}$ combinations of size k
  - All combinations: $2^n$-1 (let’s ignore the -1 from now on)
  - Largest solution set: $\binom{n}{n/2}$ minimal uniques of size $\frac{n}{2}$
  - Adding a column doubles the search space

# Output

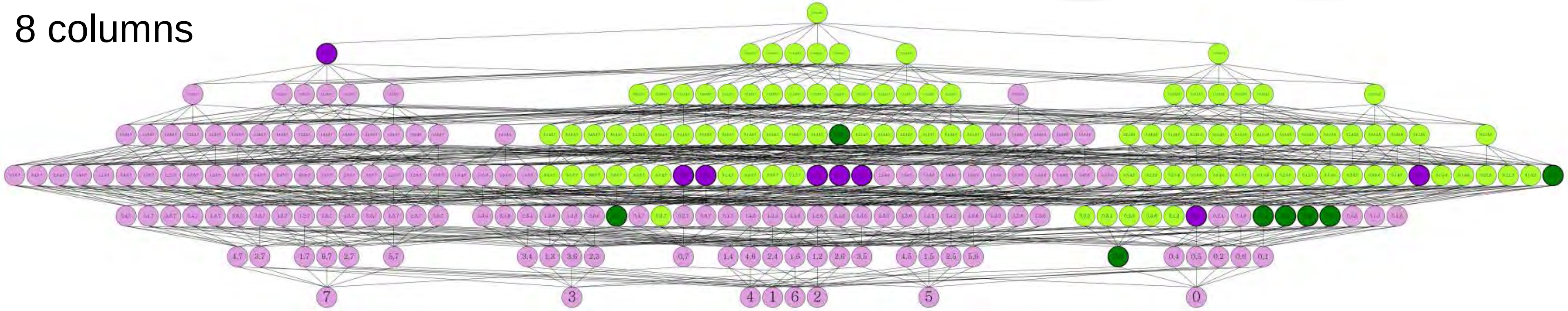

minimal unique
unique
maximal non-unique
non-unique
ABCDE
ABCE
ABCD
ABDE
ACDE
BCDE
ABC
ABE
ABD
ACD
ADE
ACE
BCD
BCE
BDE
CDE
AB
AC
AD
AE
BC
BD
BE
CD
CE
DE
A
B
C
D
E

# TPCH line item

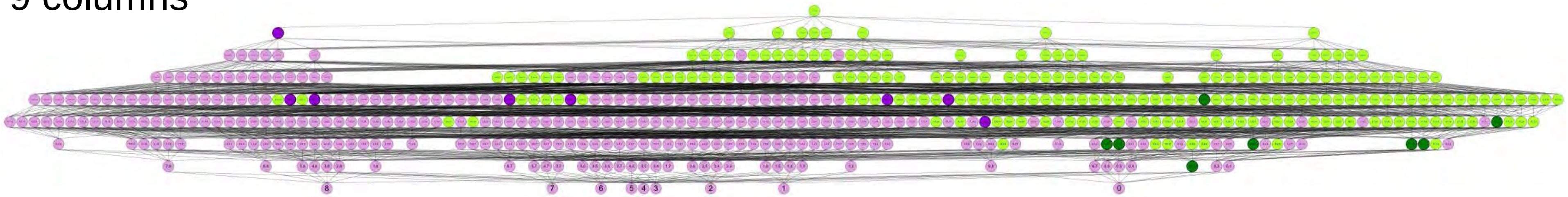
unique

non-unique

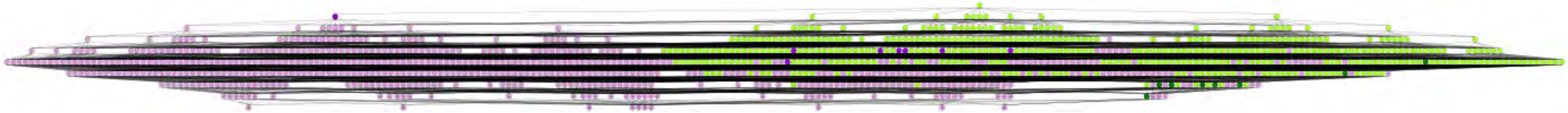
8 columns



9 columns



10 columns

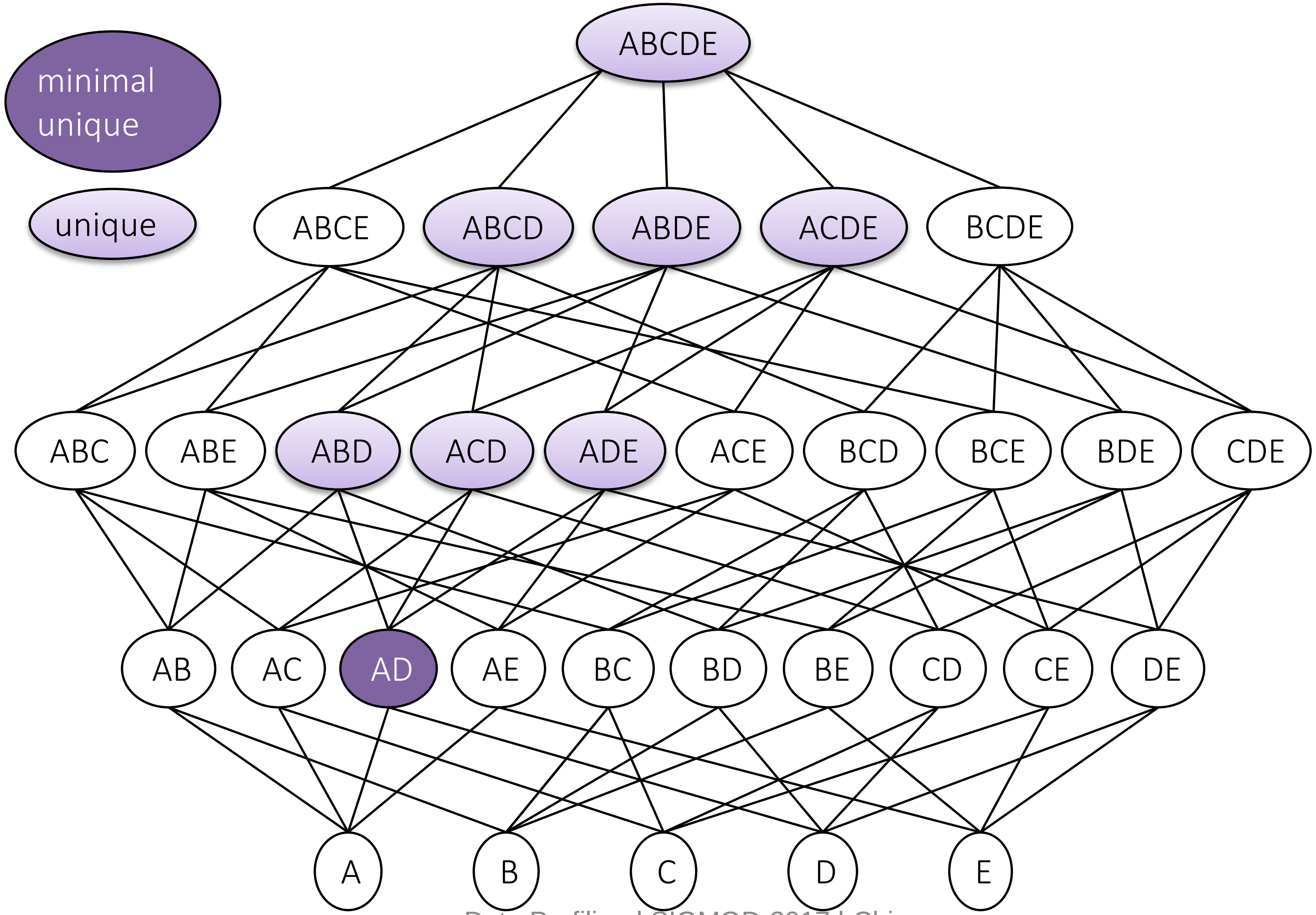


# Pruning

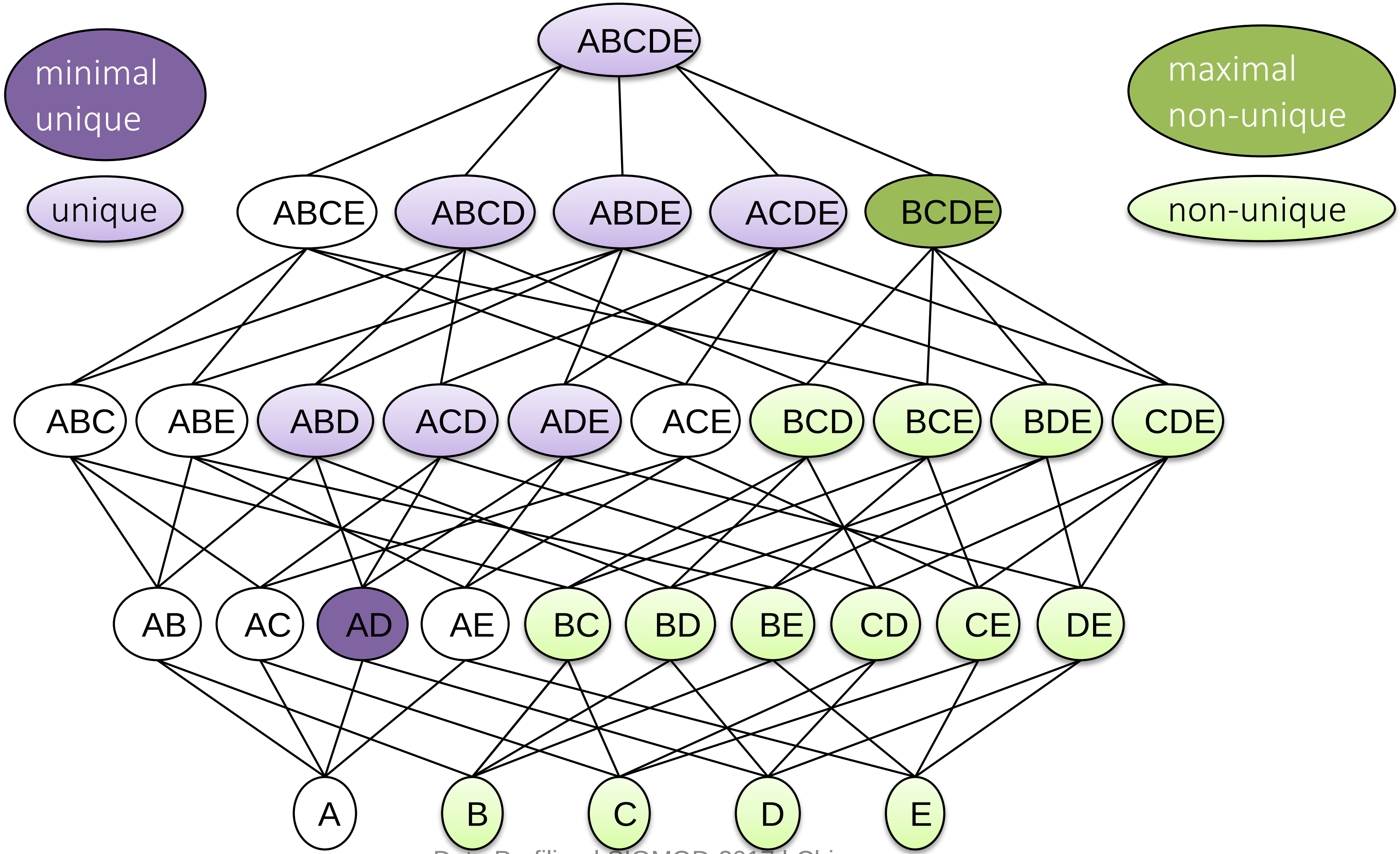
- Pruning:
  - If X is unique, its supersets must be unique
  - If Y is non-unique, its subsets must be non-unique

- Finding a unique column prunes half the lattice
  - Remove column from initial data set and restart

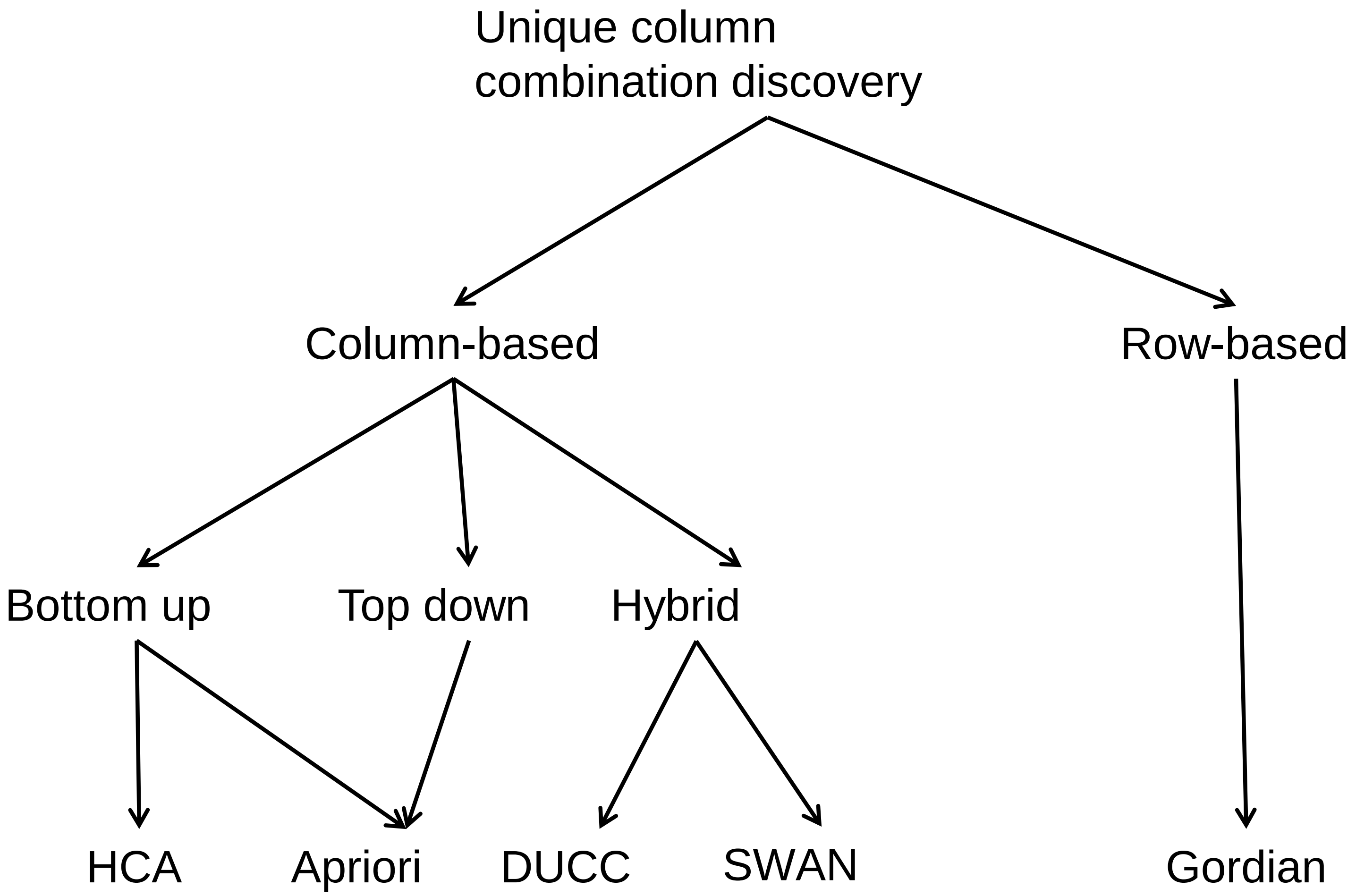
# Pruning effect of attribute pair



# Pruning both ways



# Discovery Algorithms



# Column-based algorithms

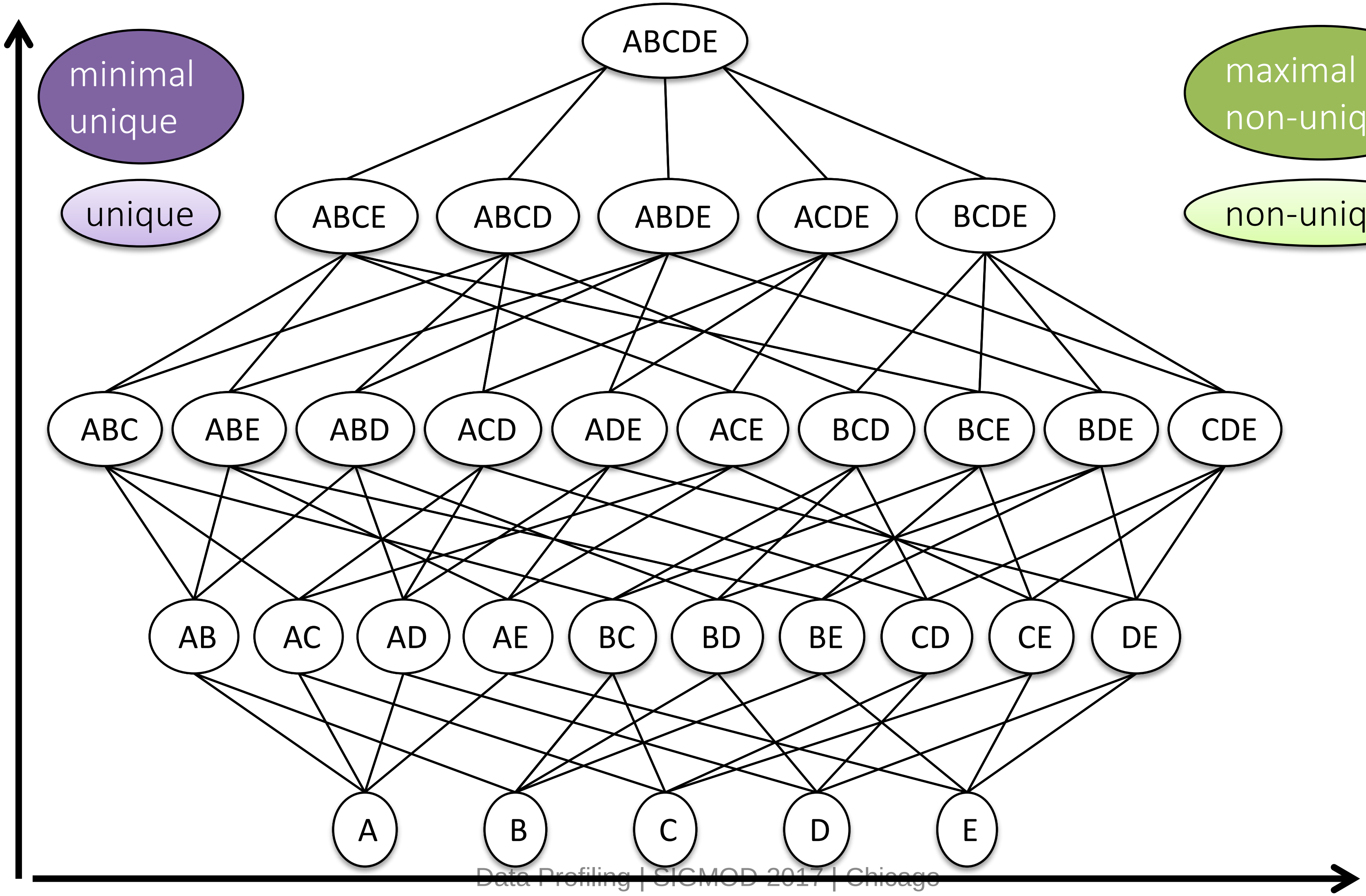
- Traverse through lattice
- Check for uniqueness
  - Can use database backend
    - SELECT COUNT(DISTINCT A, B, C) FROM R
    - Compare with row-count
- Prune lattice accordingly

# Apriori-based

[Giannella, Wyss: Finding minimal keys in a relation instance. (1999)]

- Basic idea:
  - Using the state of combinations of size k
  - We need to visit only unpruned combinations of size k+1
  - Add non-unique columns to combination of size k
- Start with individual columns
- Check pairs of non-unique columns
- Check triples of non-unique pairs …
- Terminate if no new combinations can be enumerated

# Apriori visualized



# Characteristics of Apriori

- Works well for small uniques
  - Bottom-up checks single columns first
- Best case: all columns are unique
  - n checks
- Worst case: no uniques = one duplicate row
  - $2^n$ checks
- Apriori is exponential in n

# Extensions

- Top-down
  - Start from top (all columns)
  - Works well if solution set is high up
- Hybrid[Giannella, Wyss: Finding minimal keys in a relation instance. (1999)]
  - Interleaved bottom-up and top-down
  - Works well if solution set has many small and large combinations
  - Worst case: solution set in the middle
- Statistics-based extensions [Abedjan, Naumann: Advancing the discovery of unique column combinations, *CIKM*'11]
  - Uses histograms for pruning
- Random walk [Heise, Quiané-Ruiz, Abedjan, Jentzsch, Naumann: Scalable Discovery of Unique Column Combinations, PVLDB'14]
  - Pick random superset if current column set is non-unique, random subset otherwise

○ Unique column combination

🟩 Minimal unique column combination

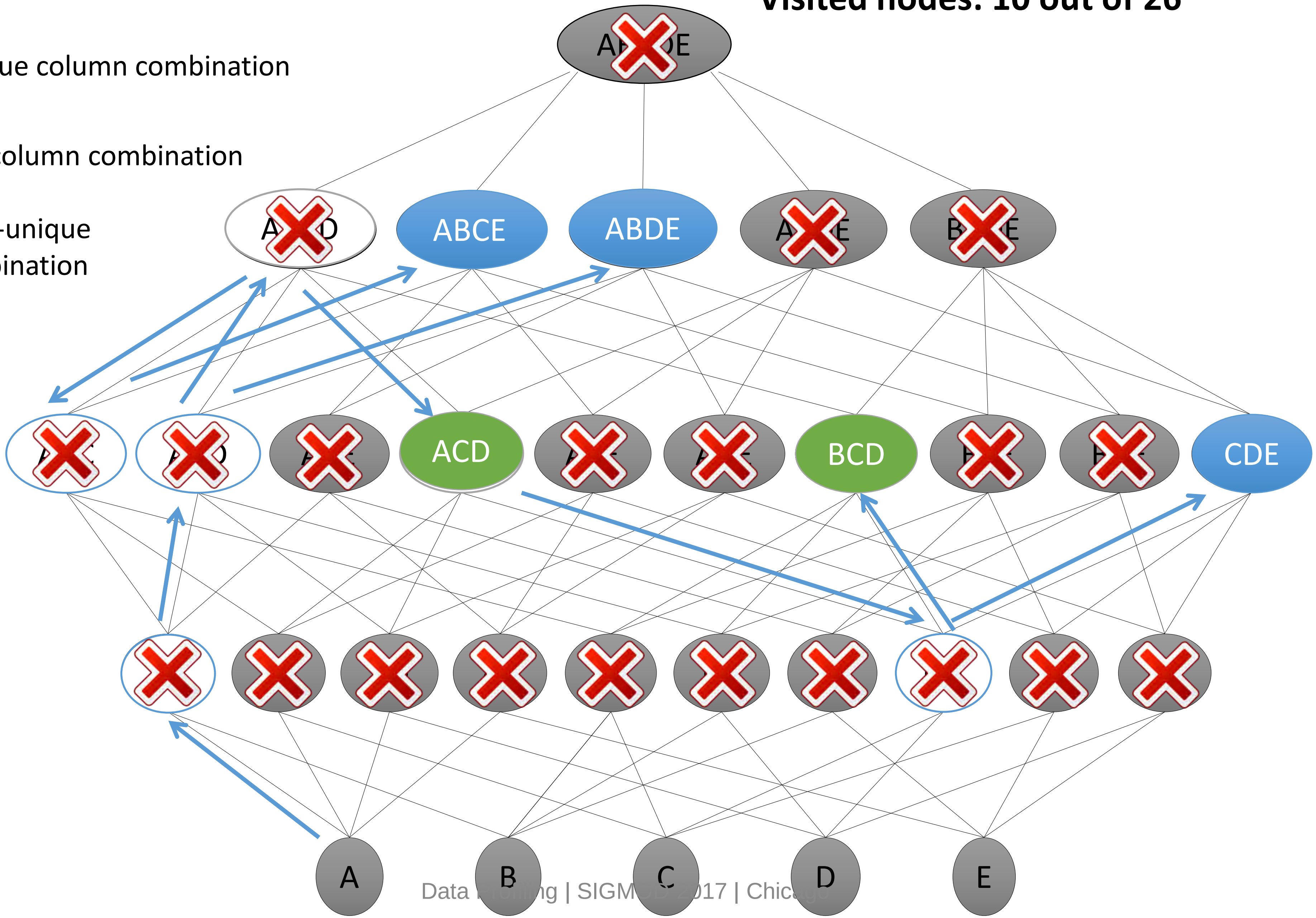
○ Non-unique column combination

🟦 Maximal non-unique column combination

✗ Pruned

## ACD and BCD are minimal uniques

## Visited nodes: 10 out of 26



# Uniques on Dynamic Data: SWAN

[Abedjan, Quanie-Ruiz, Naumann: Detecting Unique Column Combinations on Dynamic Data, ICDE’14]

- **Inserts** may create new duplicate combinations
  - Minimal uniques might become non-unique
  - Maximal non-uniques might lose maximality
- **Deletes** remove duplicate value combinations
  - Non-uniques might become unique
  - Minimal uniques might lose minimality
- SWAN
  - Leverage previously discovered minimal uniques and maximal non-uniques
  - Create appropriate indices

# Functional Dependencies

# Trivial and minimal FDs

- „X → A“ is a statement about a relation R: When two tuples have same value in attribute set X, the must have same values in attribute A.
- Non-trivial: At least one attribute on RHS does not appear on LHS
  - Street, City → Zip, City
- Completely non-trivial: Attributes on LHS and RHS are disjoint.
  - Street, City → Zip
- Minimal FD: RHS does not depend on any subset of LHS
- Typical goal: Given a relation R, find all minimal completely non-trivial functional dependencies.

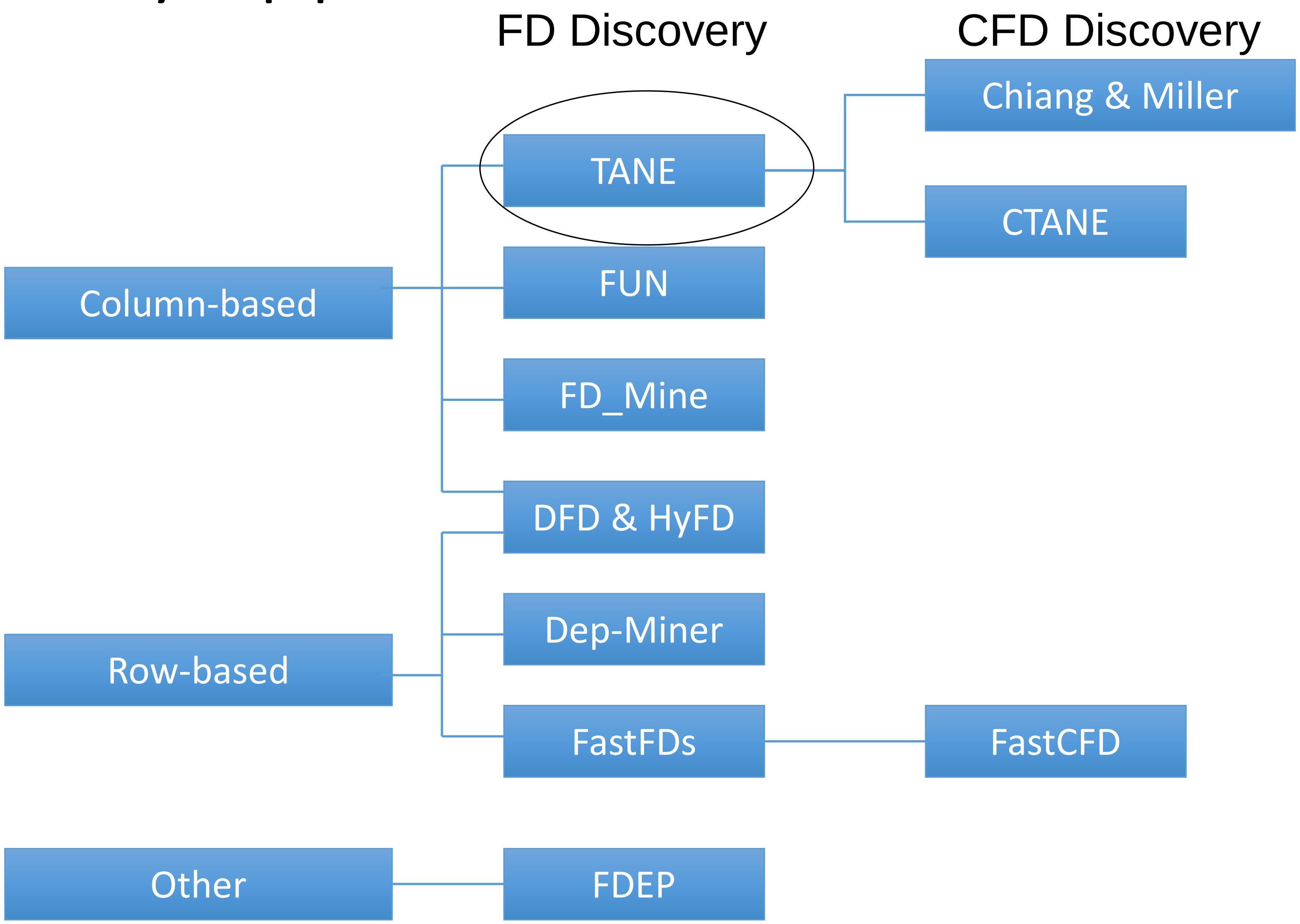
# Naive Discovery Approach

- Task: Given relation R, detect all minimal, non-trivial FDs X → A.

- For each A $\in$ R
  - For each column combination X in R\A
    - If COUNT DISTINCT(X) = COUNT DISTINCT(XA)
      - Return X → A

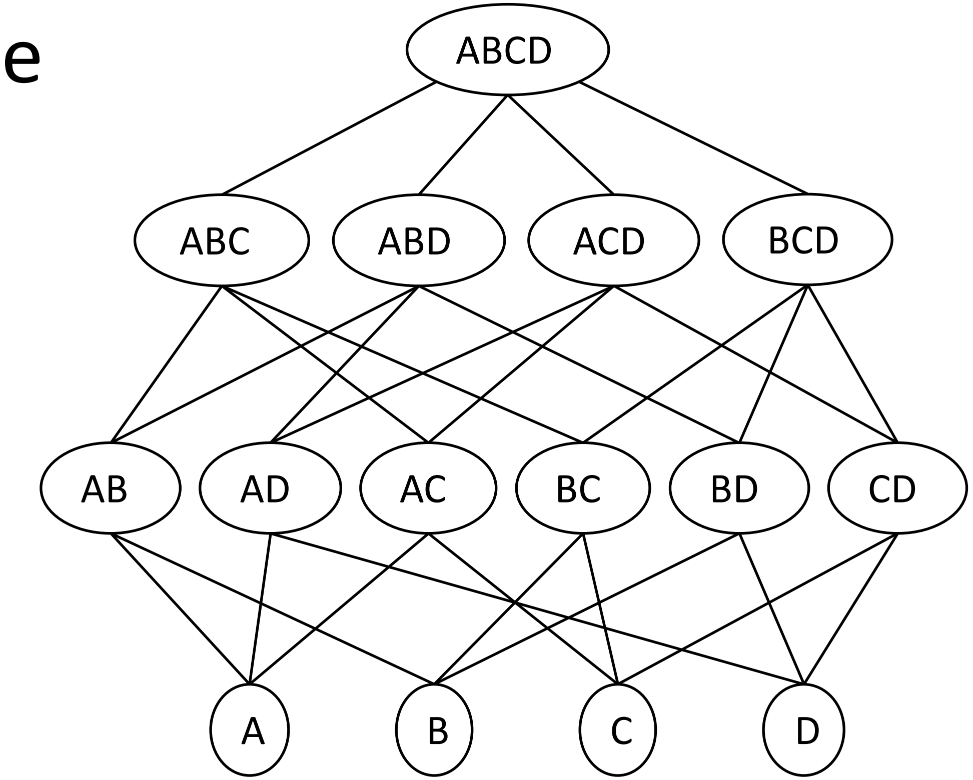
- Complexity
  - For each of the |R| possibilities for RHS
    - check $2^{(|R|-1)}$ combinations for LHS

# FD Discovery approaches



# TANE [Huhtala, Kärkkäinen, Porkka, Toivonen:TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies, Computer Journal’99]

- Bottom up traversal through lattice
  - ⇒ only minimal dependencies
  - Pruning: if B→C, don’t check BD→C
  - Avoids COUNT DISTINCTs
- For a set X, test all X\A → A, A∈X
  - ⇒ only non-trivial dependencies

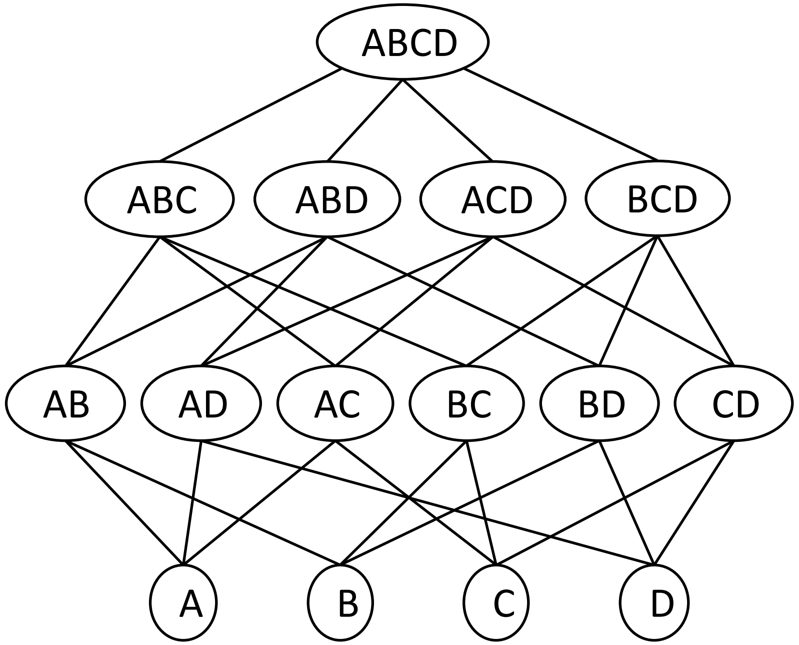


# Candidate Sets

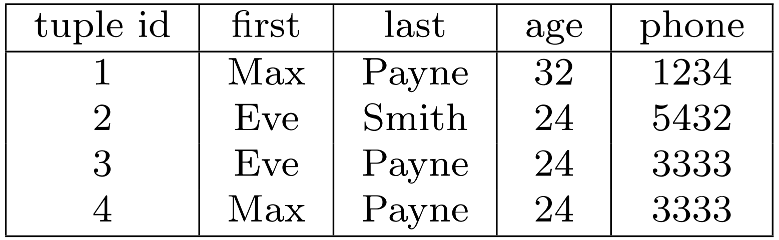
- RHS candidate set C(X)
- Stores only those attributes that might depend on **all** other attributes in X.
  - I.e., those that still need to be checked
  - If $A \in C(X)$ then A does not depend on any proper subset of X.
- $C(X) = R \setminus \{A \in X \mid X \setminus A \rightarrow A \text{ holds}\}$
- Examples: R= {ABCD}, and A → C and CD → B hold
  - C(A) = {ABCD}\{} = C(B) = C(C) = C(D)
  - C(AB) = {ABCD}\{}
  - C(AC) = {ABCD}\{C} = {ABD}
  - C(CD) = {ABCD}\{}
  - C(BCD) = {ABCD}\{B} = {ACD}

# RHS Candidate Pruning

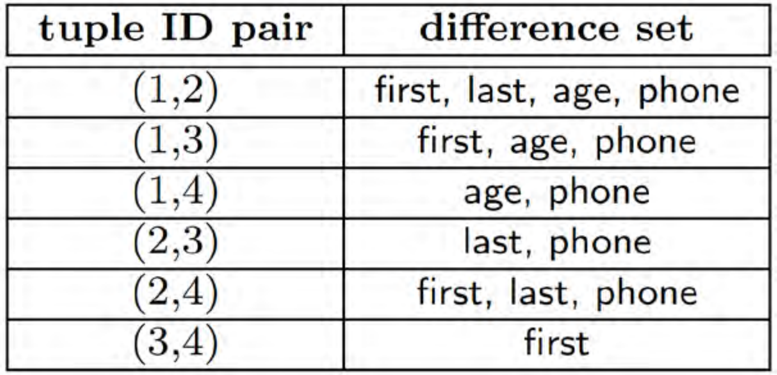
- RHS candidates: $C^+(X) = \{A \in R \mid \forall B \in X: X \setminus \{A,B\} \rightarrow B \text{ does not hold}\}$
  - Special case: A = B corresponds to C(X)
    - Reminder: $C(X) = R \setminus \{A \in X \mid X \setminus A \rightarrow A \text{ holds}\}$
- This definition removes three types of candidates:
  - Minimality
  - Pseudotransitivity
  - Superkey
- Examples: R= {ABCD}, and A → C and CD → B hold
  - C(ABC) = {A}
  - C(BCD) = {ACD}



# Row-Based Algorithms



| tuple id | first | last | age | phone |
|---|---|---|---|---|
| 1 | Max | Payne | 32 | 1234 |
| 2 | Eve | Smith | 24 | 5432 |
| 3 | Eve | Payne | 24 | 3333 |
| 4 | Max | Payne | 24 | 3333 |



| tuple ID pair | difference set |
|---|---|
| (1,2) | first, last, age, phone |
| (1,3) | first, age, phone |
| (1,4) | age, phone |
| (2,3) | last, phone |
| (2,4) | first, last, phone |
| (3,4) | first |

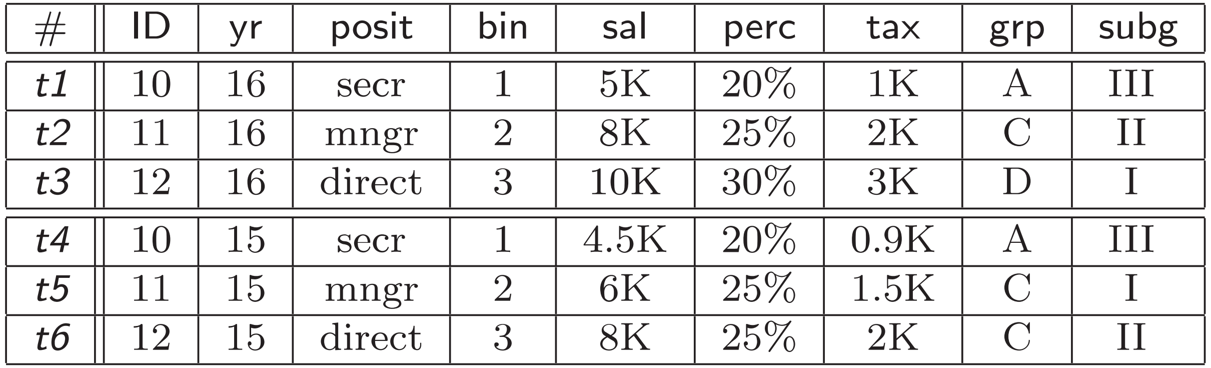
- For each candidate RHS (say, phone)
- Find difference sets including phone, with phone removed
  - {first,last,age}, {fist,age}, {age}, {last}, {first,last}
- So there are pairs of tuples with different phones and different {first,last,age}, different {first,age}, etc.
- Find minimal column subsets that have a non-empty intersection with each difference set
  - {last,age}
- Conclude that {last,age} → phone

# FD Discovery on Dynamic Data [Wang, Tsou, Lin and Hong: Maintenance of discovered functional dependencies: Incremental deletion, ISDA'03]

- Insertions
  - Existing FDs may be violated: Check each one
- Deletions
  - New FD may appear if conflicting tuple deleted: Revisit entire lattice

# Order Dependencies

# Example



| # | ID | yr | posit | bin | sal | perc | tax | grp | subg |
|---|---|---|---|---|---|---|---|---|---|
| *t1* | 10 | 16 | secr | 1 | 5K | 20% | 1K | A | III |
| *t2* | 11 | 16 | mngr | 2 | 8K | 25% | 2K | C | II |
| *t3* | 12 | 16 | direct | 3 | 10K | 30% | 3K | D | I |
| *t4* | 10 | 15 | secr | 1 | 4.5K | 20% | 0.9K | A | III |
| *t5* | 11 | 15 | mngr | 2 | 6K | 25% | 1.5K | C | I |
| *t6* | 12 | 15 | direct | 3 | 8K | 25% | 2K | C | II |

- X→A if sorting on X also sorts on A
- tax→salary
- ODs subsume FDs
  - If X functionally determines Y then X orders XY

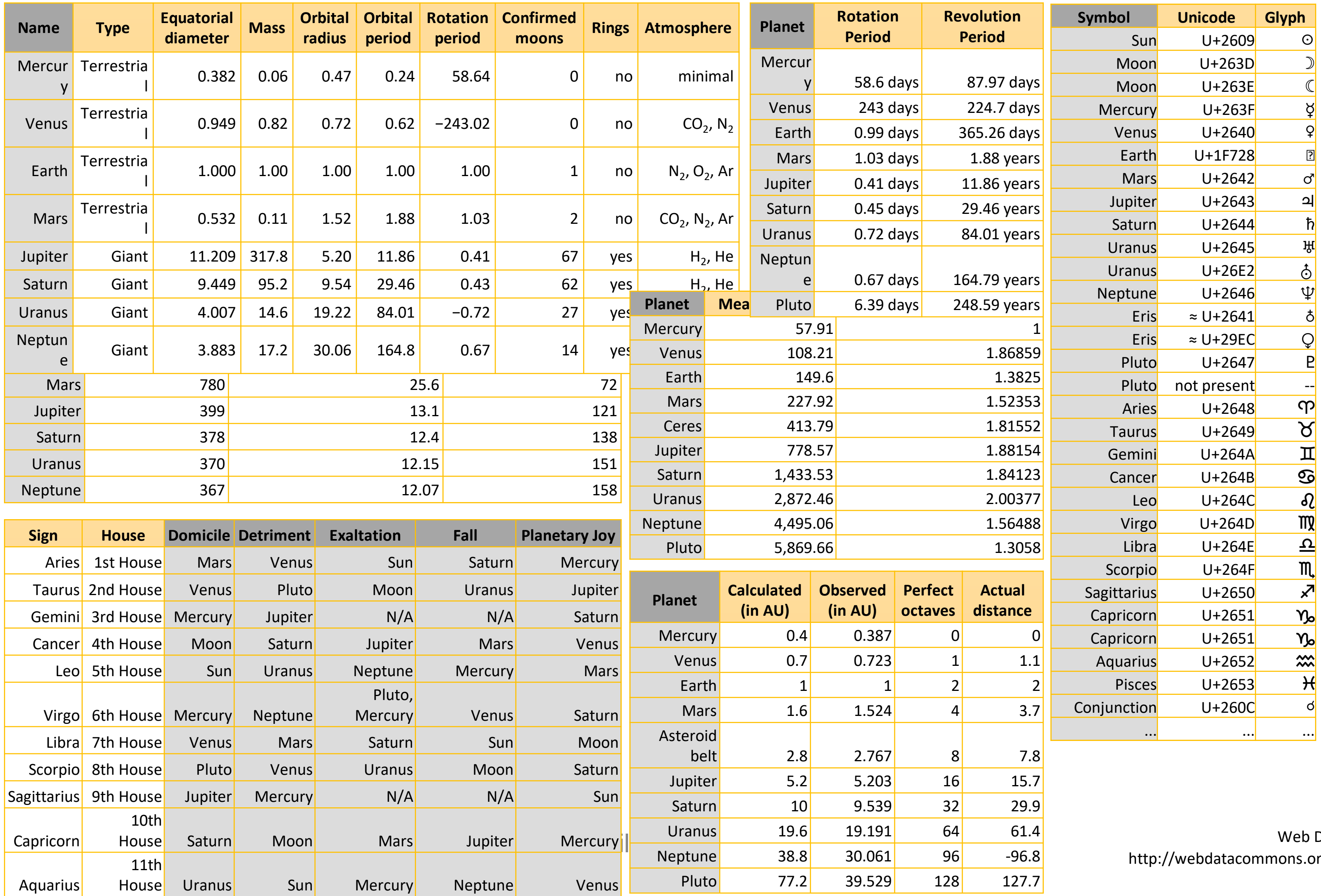
# Discovering Order Dependencies

- List-based lattice approach [Langer, Naumann: Discovering Order Dependencies, VLDBJ’15]
    - Apriori-like, but order matters: XY→A is different from YX→A
- Set-based lattice approach [Szlichta, Godfrey, Golab, Kargar, Srivastava: Effective and Complete Discovery of Order Dependencies via Set-based Axiomatization, PVLDB’17]
    - Rewrite ODs using a set-based canonical form
- Both approaches:
    - New pruning rules based on OD semantics/axioms

# Inclusion Dependencies

CHICAGO
CHICAGO
FRANKIE VALLI & THE 4 SEASONS APR 18
THIS AMERICAN LIFE APR 19

# Linking web tables – an example



| Name | Type | Equatorial diameter | Mass | Orbital radius | Orbital period | Rotation period | Confirmed moons | Rings | Atmosphere |
|---|---|---|---|---|---|---|---|---|---|
| Mercury | Terrestrial | 0.382 | 0.06 | 0.47 | 0.24 | 58.64 | 0 | no | minimal |
| Venus | Terrestrial | 0.949 | 0.82 | 0.72 | 0.62 | −243.02 | 0 | no | $CO_2$, $N_2$ |
| Earth | Terrestrial | 1.000 | 1.00 | 1.00 | 1.00 | 1.00 | 1 | no | $N_2$, $O_2$, Ar |
| Mars | Terrestrial | 0.532 | 0.11 | 1.52 | 1.88 | 1.03 | 2 | no | $CO_2$, $N_2$, Ar |
| Jupiter | Giant | 11.209 | 317.8 | 5.20 | 11.86 | 0.41 | 67 | yes | $H_2$, He |
| Saturn | Giant | 9.449 | 95.2 | 9.54 | 29.46 | 0.43 | 62 | yes | $H_2$, He |
| Uranus | Giant | 4.007 | 14.6 | 19.22 | 84.01 | −0.72 | 27 | yes | |
| Neptune | Giant | 3.883 | 17.2 | 30.06 | 164.8 | 0.67 | 14 | yes | |

| | | | |
|---|---|---|---|
| Mars | 780 | 25.6 | 72 |
| Jupiter | 399 | 13.1 | 121 |
| Saturn | 378 | 12.4 | 138 |
| Uranus | 370 | 12.15 | 151 |
| Neptune | 367 | 12.07 | 158 |

| Planet | Rotation Period | Revolution Period |
|---|---|---|
| Mercury | 58.6 days | 87.97 days |
| Venus | 243 days | 224.7 days |
| Earth | 0.99 days | 365.26 days |
| Mars | 1.03 days | 1.88 years |
| Jupiter | 0.41 days | 11.86 years |
| Saturn | 0.45 days | 29.46 years |
| Uranus | 0.72 days | 84.01 years |
| Neptune | 0.67 days | 164.79 years |
| Pluto | 6.39 days | 248.59 years |

| Planet | Mea | |
|---|---|---|
| Mercury | 57.91 | 1 |
| Venus | 108.21 | 1.86859 |
| Earth | 149.6 | 1.3825 |
| Mars | 227.92 | 1.52353 |
| Ceres | 413.79 | 1.81552 |
| Jupiter | 778.57 | 1.88154 |
| Saturn | 1,433.53 | 1.84123 |
| Uranus | 2,872.46 | 2.00377 |
| Neptune | 4,495.06 | 1.56488 |
| Pluto | 5,869.66 | 1.3058 |

| Symbol | Unicode | Glyph |
|---|---|---|
| Sun | U+2609 | ☉ |
| Moon | U+263D | ☽ |
| Moon | U+263E | ☾ |
| Mercury | U+263F | ☿ |
| Venus | U+2640 | ♀ |
| Earth | U+1F728 | ⍰ |
| Mars | U+2642 | ♂ |
| Jupiter | U+2643 | ♃ |
| Saturn | U+2644 | ♄ |
| Uranus | U+2645 | ♅ |
| Uranus | U+26E2 | ⛢ |
| Neptune | U+2646 | ♆ |
| Eris | ≈ U+2641 | ♁ |
| Eris | ≈ U+29EC | ⧬ |
| Pluto | U+2647 | ♇ |
| Pluto | not present | -- |
| Aries | U+2648 | ♈ |
| Taurus | U+2649 | ♉ |
| Gemini | U+264A | ♊ |
| Cancer | U+264B | ♋ |
| Leo | U+264C | ♌ |
| Virgo | U+264D | ♍ |
| Libra | U+264E | ♎ |
| Scorpio | U+264F | ♏ |
| Sagittarius | U+2650 | ♐ |
| Capricorn | U+2651 | ♑ |
| Capricorn | U+2651 | ♑ |
| Aquarius | U+2652 | ♒ |
| Pisces | U+2653 | ♓ |
| Conjunction | U+260C | ☌ |
| … | … | … |

| Sign | House | Domicile | Detriment | Exaltation | Fall | Planetary Joy |
|---|---|---|---|---|---|---|
| Aries | 1st House | Mars | Venus | Sun | Saturn | Mercury |
| Taurus | 2nd House | Venus | Pluto | Moon | Uranus | Jupiter |
| Gemini | 3rd House | Mercury | Jupiter | N/A | N/A | Saturn |
| Cancer | 4th House | Moon | Saturn | Jupiter | Mars | Venus |
| Leo | 5th House | Sun | Uranus | Neptune | Mercury | Mars |
| Virgo | 6th House | Mercury | Neptune | Pluto, Mercury | Venus | Saturn |
| Libra | 7th House | Venus | Mars | Saturn | Sun | Moon |
| Scorpio | 8th House | Pluto | Venus | Uranus | Moon | Saturn |
| Sagittarius | 9th House | Jupiter | Mercury | N/A | N/A | Sun |
| Capricorn | 10th House | Saturn | Moon | Mars | Jupiter | Mercury |
| Aquarius | 11th House | Uranus | Sun | Mercury | Neptune | Venus |

| Planet | Calculated (in AU) | Observed (in AU) | Perfect octaves | Actual distance |
|---|---|---|---|---|
| Mercury | 0.4 | 0.387 | 0 | 0 |
| Venus | 0.7 | 0.723 | 1 | 1.1 |
| Earth | 1 | 1 | 2 | 2 |
| Mars | 1.6 | 1.524 | 4 | 3.7 |
| Asteroid belt | 2.8 | 2.767 | 8 | 7.8 |
| Jupiter | 5.2 | 5.203 | 16 | 15.7 |
| Saturn | 10 | 9.539 | 32 | 29.9 |
| Uranus | 19.6 | 19.191 | 64 | 61.4 |
| Neptune | 38.8 | 30.061 | 96 | -96.8 |
| Pluto | 77.2 | 39.529 | 128 | 127.7 |

Web Data Commons Project
http://webdatacommons.org/webtables/index.html

# Unary IND detection complexity

| Name | Type | Equatorial diameter | Mass | Orbital radius | Orbital period | Rotation period | Confirmed moons | Rings | Atmosphere |
|---|---|---|---|---|---|---|---|---|---|
| Mercury | Terrestrial | 0.382 | 0.06 | 0.47 | 0.24 | 58.64 | 0 | no | minimal |
| Venus | Terrestrial | 0.949 | 0.82 | 0.72 | 0.62 | −243.02 | 0 | no | $CO_2$, $N_2$ |
| Earth | Terrestrial | 1.000 | 1.00 | 1.00 | 1.00 | 1.00 | 1 | no | $N_2$, $O_2$, Ar |
| Mars | Terrestrial | 0.532 | 0.11 | 1.52 | 1.88 | 1.03 | 2 | no | $CO_2$, $N_2$, Ar |
| Jupiter | Giant | 11.209 | 317.8 | 5.20 | 11.86 | 0.41 | 67 | yes | $H_2$, He |
| Saturn | Giant | 9.449 | 95.2 | 9.54 | 29.46 | 0.43 | 62 | yes | $H_2$, He |
| Uranus | Giant | 4.007 | 14.6 | 19.22 | 84.01 | −0.72 | 27 | yes | $H_2$, He |
| Neptune | Giant | 3.883 | 17.2 | 30.06 | 164.8 | 0.67 | 14 | yes | $H_2$, He |

**Complexity:** $O(n^2-n)$

for n attributes

**Example:**

10 attr ~ 90 checks

1,000 attr ~ 999,000 checks

- Name ⊆ Type ?
- Name ⊆ Equatorial_diameter ?
- Name ⊆ Mass ?
- Name ⊆ Orbital_radius ?
- Name ⊆ Orbital_period ?
- Name ⊆ Rotation_period ?
- Name ⊆ Confirmed_moons ?
- Name ⊆ Rings ?
- Name ⊆ Atmosphere ?

- Type ⊆ Name ?
- Type ⊆ Equatorial_diameter ?
- Type ⊆ Mass ?
- Type ⊆ Orbital_radius ?
- Type ⊆ Orbital_period ?
- Type ⊆ Rotation_period ?
- Type ⊆ Confirmed_moons ?
- Type ⊆ Rings ?
- Type ⊆ Atmosphere ?

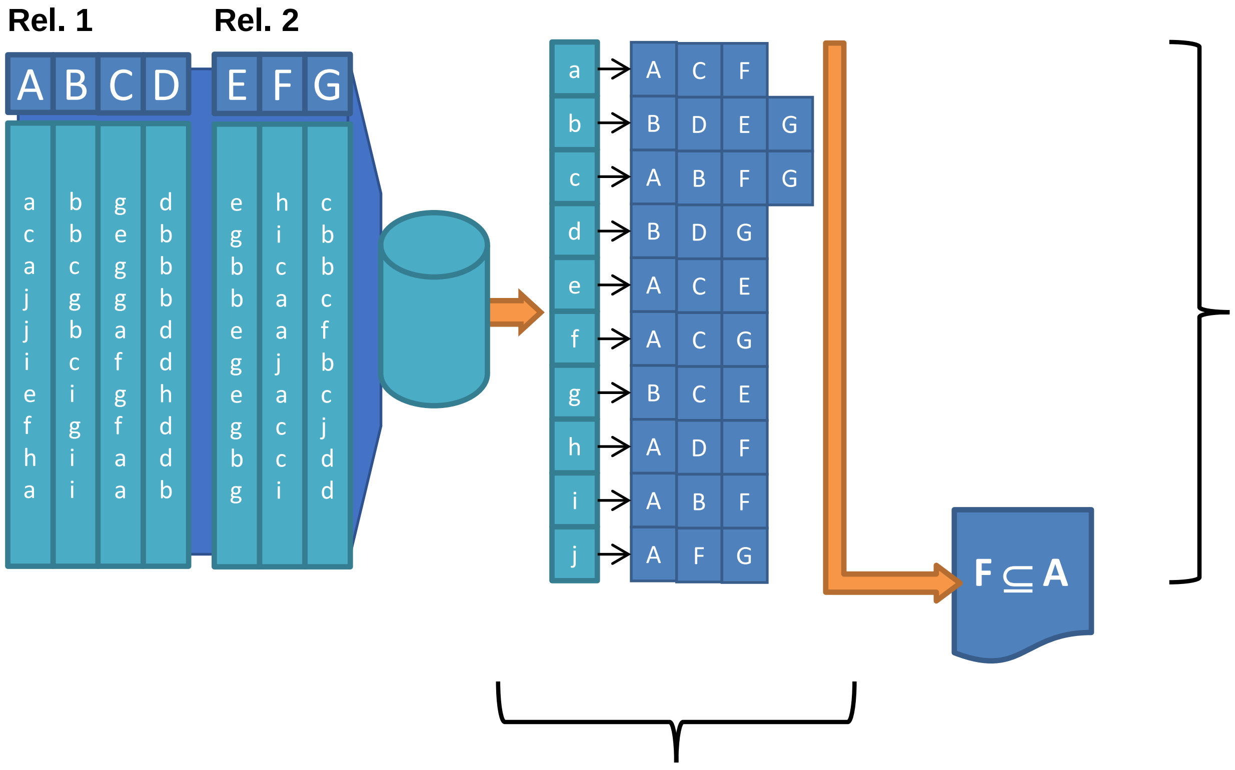
- Mass ⊆ Name ?
- Mass ⊆ Type ?
- Mass ⊆ Equatorial_diameter ?
- …

# MIND

[Marchi, Lopes, Petit: Unary and n-ary inclusion dependency discovery in relational databases,JIIS'09]

🟦 attributes 🟦 values

➡ dataflow ☒ ignored



Rel. 1

| A | B | C | D |
|---|---|---|---|
| a | b | g | d |
| c | b | e | b |
| a | c | g | b |
| j | g | g | b |
| j | b | a | d |
| i | c | f | d |
| e | i | g | h |
| f | g | f | d |
| h | i | a | d |
| a | i | a | b |

Rel. 2

| E | F | G |
|---|---|---|
| e | h | c |
| g | i | b |
| b | c | b |
| b | a | c |
| e | a | f |
| g | j | b |
| e | a | c |
| g | c | j |
| b | c | d |
| g | i | d |

a → A C F

b → B D E G

c → A B F G

d → B D G

e → A C E

f → A C G

g → B C E

h → A D F

i → A B F

j → A F G

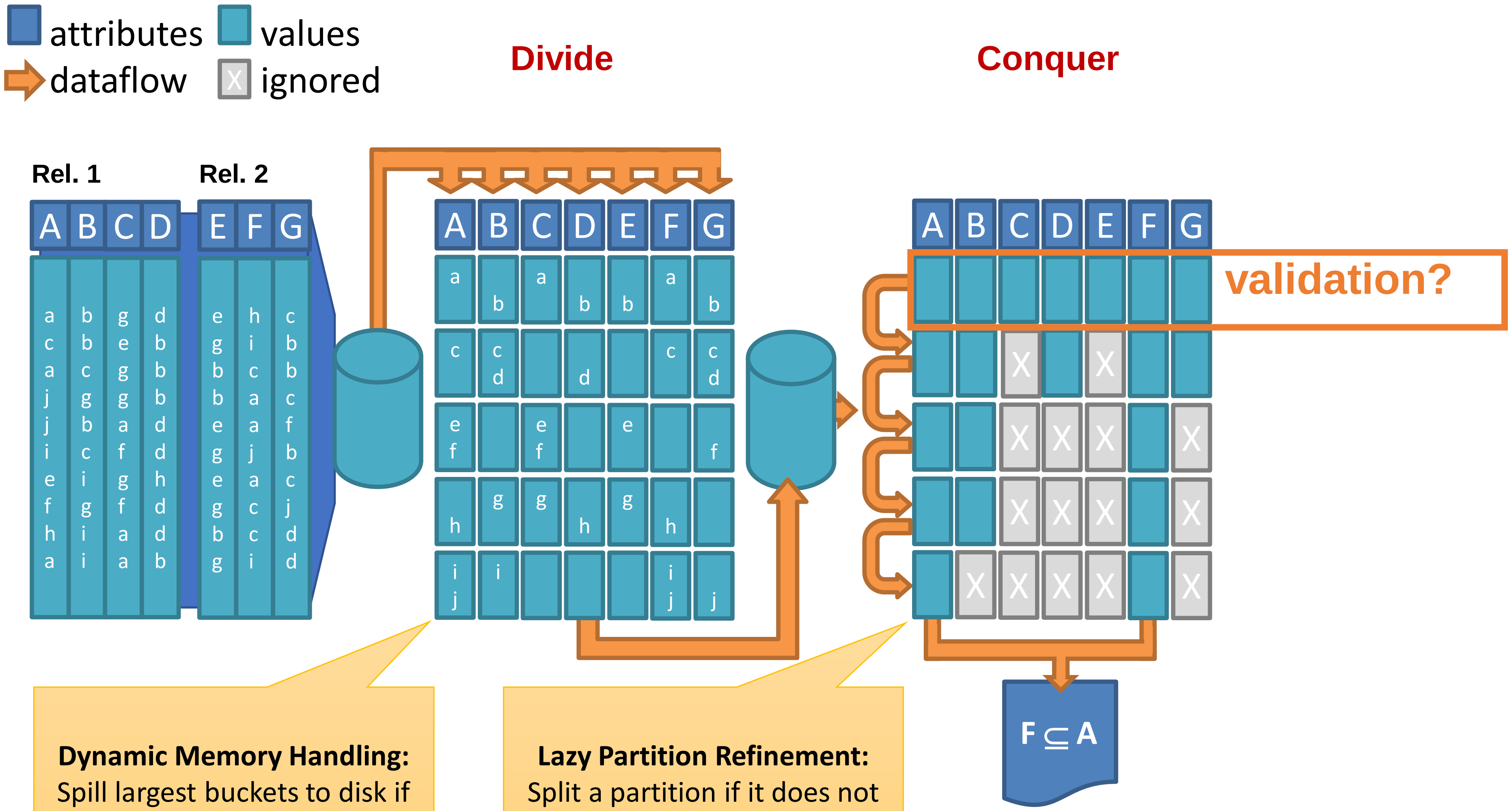
F ⊆ A

All intersections are checked, but not all are necessary!

Needs to fit in main memory!

# BINDER

[Papenbrock, Quiane, Naumann: Divide & Conquer-based Inclusion Dependency Discovery, PVLDB’15]

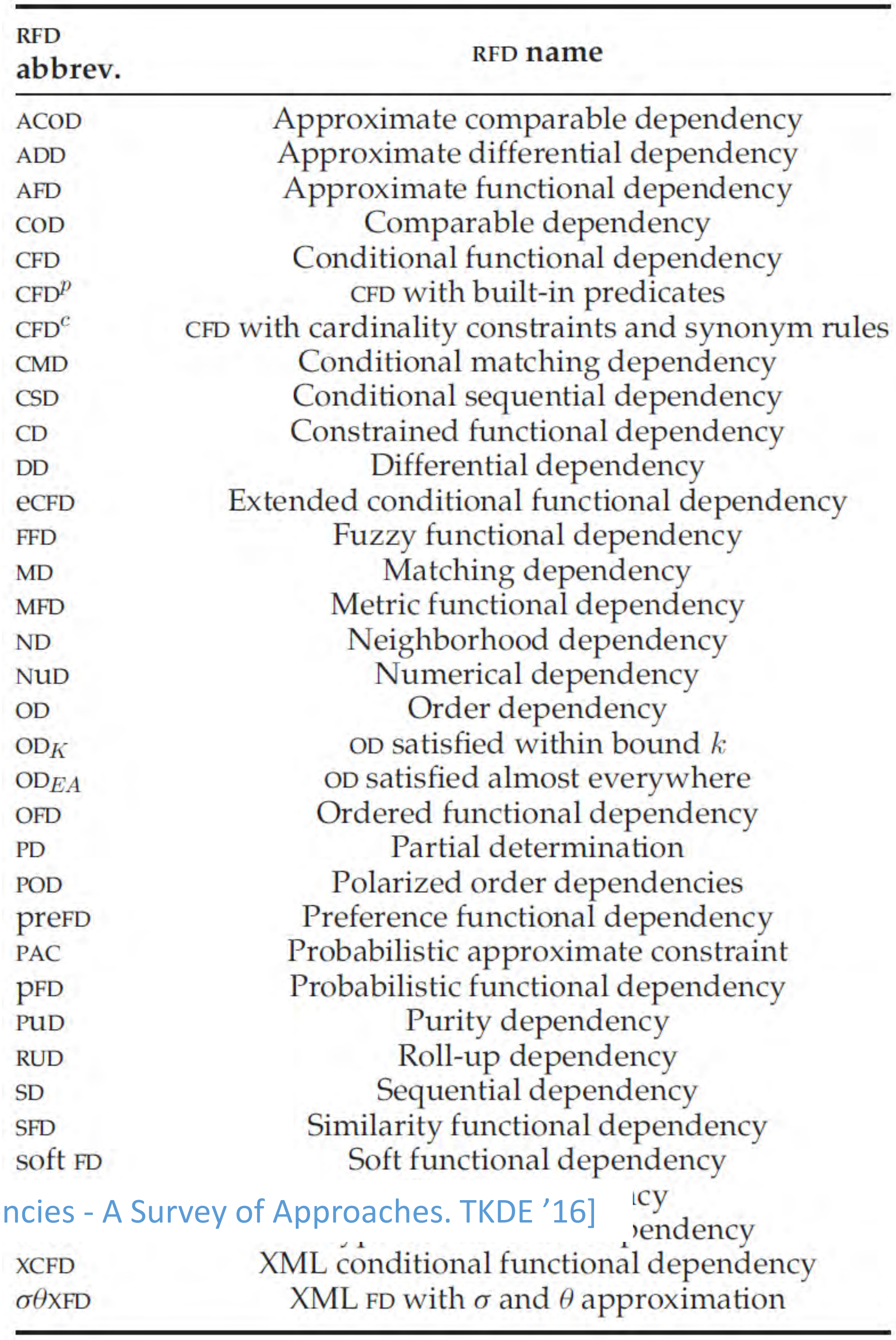


**Dynamic Memory Handling:** Spill largest buckets to disk if memory is exhausted.

**Lazy Partition Refinement:** Split a partition if it does not fit into main memory.

# Extensions

- Dependencies are sensitive to data errors
- Conditional Functional Dependencies
  - X→A but only for X=x1 and X=x5
- Approximate Functional Dependencies
  - How many rows (at a minimum) would have to be removed so the remaining rows satisfy the FD?
- Metric Functional Dependencies
  - X→A holds if tuples that agree on X have As within some distance



| RFD abbrev. | RFD name |
|---|---|
| ACOD | Approximate comparable dependency |
| ADD | Approximate differential dependency |
| AFD | Approximate functional dependency |
| COD | Comparable dependency |
| CFD | Conditional functional dependency |
| $\text{CFD}^p$ | CFD with built-in predicates |
| $\text{CFD}^c$ | 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 |
| $\text{OD}_K$ | OD satisfied within bound $k$ |
| $\text{OD}_{EA}$ | 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 |
|  | ıcy |
|  | endency |
| XCFD | XML conditional functional dependency |
| $\sigma\theta$XFD | XML FD with $\sigma$ and $\theta$ approximation |

[Caruccio, Deufemia, Polese: Relaxed Functional Dependencies - A Survey of Approaches. TKDE ’16]

# Visualization

[1011066.Name] =] [1011057.Name]

[129284.Reference] =] [1223862.null] [586920.Ref.] [1030730.RCDB page] [108435.No.] [1248790.Source] [983315.References] [207338.Home railway (external link)] [975850.Ref] [1375996.Source] [1129539.References] [1168707.References] [744488.Ref] [1169311.Ref] [1068498.Ref] [163214.Reference] [604676.References] [1002900.Ref] [749972.Reference] [951640.References] [939700.Page] [900853.Ref] [788203.Ref] [788409.References] [978758.Ref] [652885.Link] [652377.Ref] [1320358.Reference] [1287392.Ref] [1012269.Report] [1180077.References] [1274408.Ref] [856227.NFL Recap] [1286480.Ref] [1354142.null] [525501.References] [630016.Notes] [762537.Refs] [902406.Report] [1005369.Link] [1255682.Source] [1157534.Source] [1065320.Ref] [956840.Ref] [775466.References] [988811.Ref] [1005838.Link] [1005593.Link] [576411.References] [1134428.Ref] [1170953.Reference(s)] [699144.Note] [268733.References] [931606.Notes] [1284557.Ref.] [1357973.Source] [1238931.Report] [867400.Reference] [794774.Ref] [716064.Refs] [377521.References] [995370.Ref] [1282132.References] [1358158.Ref.] [1120007.Ref] [1342522.Ref] [1319381.null] [889114.Ref] [1004839.Link] [697527.Website] [980509.Ref(s)] [1078901.Ref]

[1390416.Rank] =] [1169921.Rank] [1183098.Rank] [1011765.Rank] [1225076.Rank] [454782.Rank] [1186535.Rank] [1209635.Rank] [1161665.Rank] [708465.Rank] [708648.Rank]

[637307.Date] =] [1311505.Date] [1337020.Date]

[1083420.Event] =] [976659.Event] [976901.Event] [975917.Event] [1060037.Event] [1068182.Event] [1067251.Event] [1067097.Event] [1000067.Event] [972968.Event] [1058267.Event] [988323.Event] [1003312.Event] [1063506.Event] [1027145.Event] [1078507.Event] [1062268.Event]

[302006.Role:] =] [391330.Role:] [703281.Role:] [387497.Role:] [735612.Role:] [151885.Role:] [150598.Role:]

[1083410.Event] =] [983546.Event] [975773.Event] [1071989.Event] [1068219.Event] [1002900.Event] [1074984.Event] [967160.Event] [1052352.Event] [1066949.Event] [1082562.Event] [1151162.Event] [1042660.Event] [1056643.Event] [950860.Event] [958921.Event] [1063309.Event] [973967.Event] [1027145.Event] [1062263.Event]

[73362.State] =] [1185141.State]

[1083402.Event] =] [1083339.Event] [1068498.Event] [1060027.Event] [1002823.Event] [1046135.Event] [1249836.Event] [1000145.Event] [994576.Event] [990543.Event]

[854590.Venue] =] [883202.Venue] [890993.Venue] [1104659.Venue]

[648260.TEAM] =] [1286540.Club] [1308745.Club]

[627822.Division Record] =] [466958.Sets W - L]

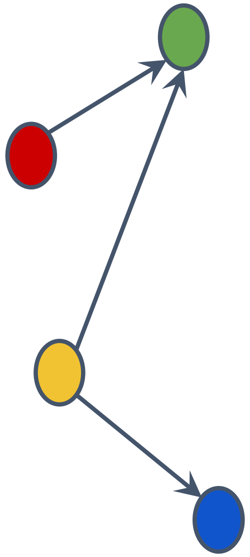
[1236345.Match] =] [1231569.Match]

...

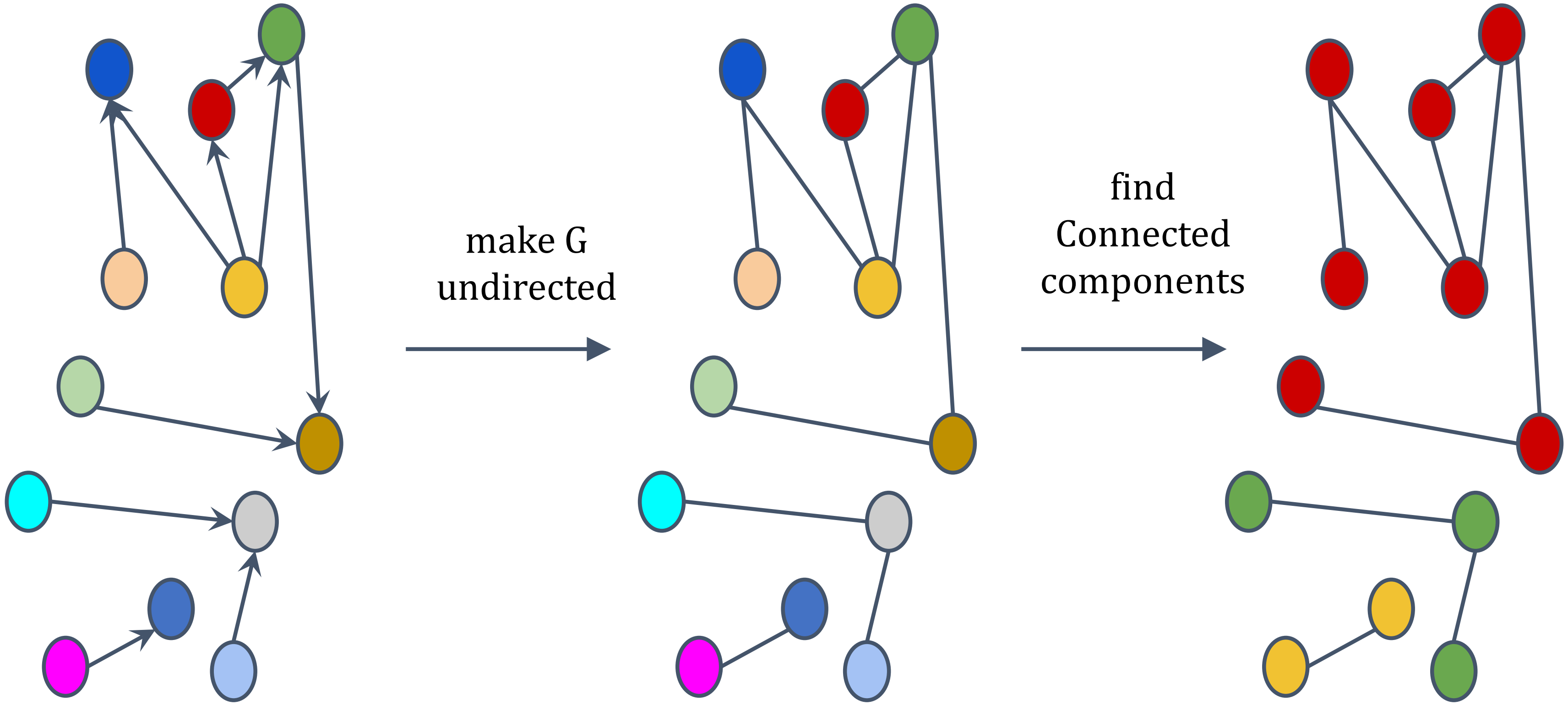
# Visualization

INDs = {
$R_1.A \subseteq R_2.B$,
$R_3.A \subseteq R_1.D$,
$R_3.C \subseteq R_2.A$,
$R_3.B \subseteq R_4.A$
}

G= (
V = {
$R_1$, $R_2$, $R_3$, $R_4$
},
E = {$(R_1, R_2)$, $(R_3, R_1)$,
$(R_3, R_2)$, $(R_3, R_4)$
}
)



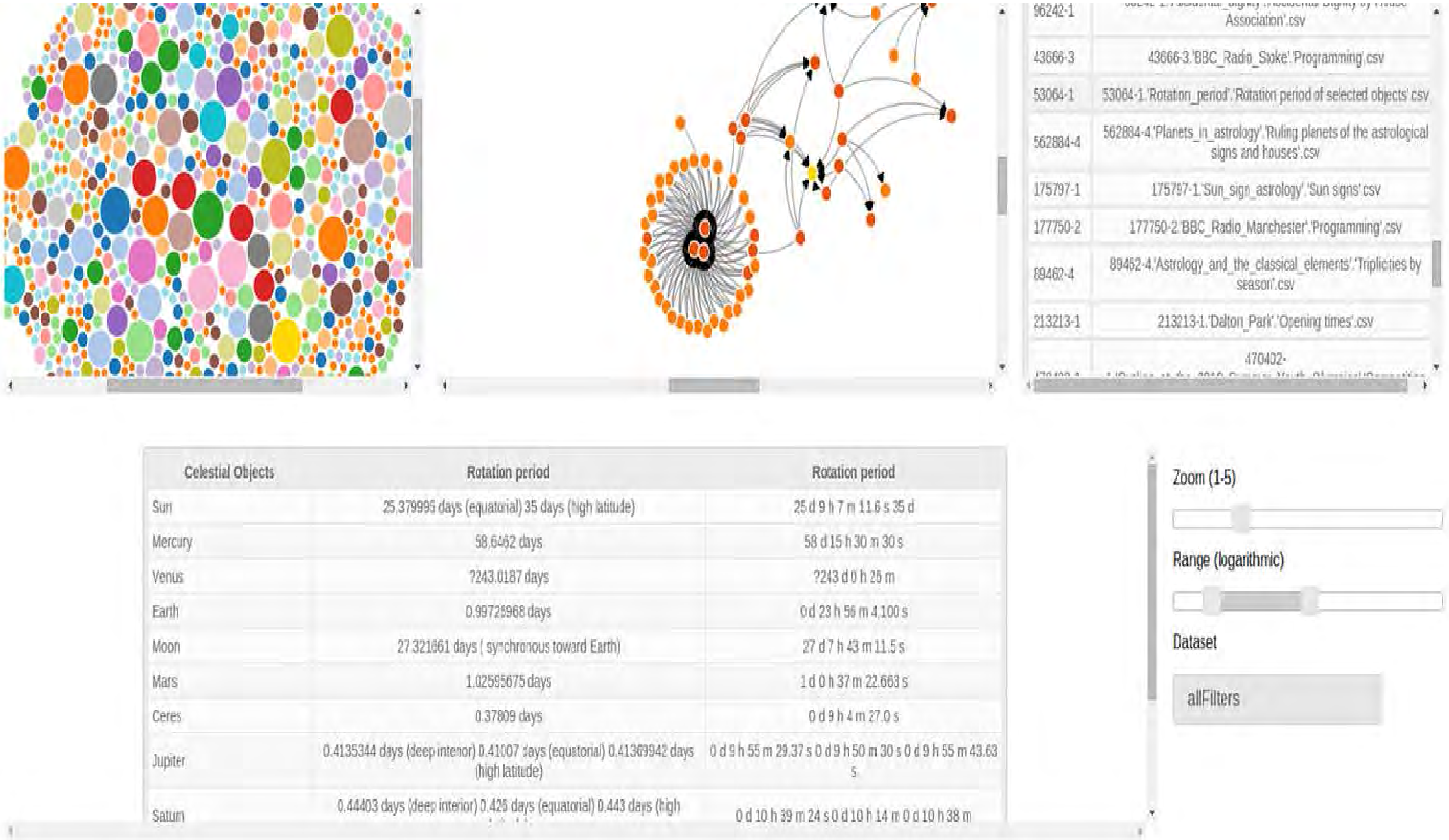
# Visualization



make G undirected

find Connected components

# Interactive Front End



| | |
|---|---|
| 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

# Ranking Dependencies

- Uniques/FDs/ODs
  - rank by size of left hand side (X→A over XYZW→A)
  - rank by position in schema
  - note: apriori-like approaches naturally produce “small” dependencies first
- Inclusion dependencies
  - rank by syntactic similarity: (name ⊆ cust_name)
  - rank by overlap (given A ⊆ B, compute |B/A|)
- Approximate dependencies
  - rank by how many rows satisfy them
- Conditional dependencies
  - rank by support (how many rows they cover)

# More Dependencies

- Denial constraints [Chu, Ilyas, Papotti: Discovering denial constraints, PVLDB’13]
  - First order logic
  - E.g., If two people live in the same province, the one earning a lower salary must pay less tax
- Differential dependencies [Song and Chen: Differential dependencies: Reasoning and discovery, TODS, 2011]
  - X→Y holds when any pair of tuples whose X values are close also have Y values which are close
- Sequential dependencies [Golab, Karloff, Korn, Saha, Srivastava: Sequential dependencies, PVLDB’09]
  - X→[p,q] A holds if sorting by X also sorts by A, and consecutive A values are at least p and at most q apart
  - E.g., Year →[0,1000] Salary means that salaries do not decrease over time and increase by at most 1000/year

next slide deck

# Tutorial Overview

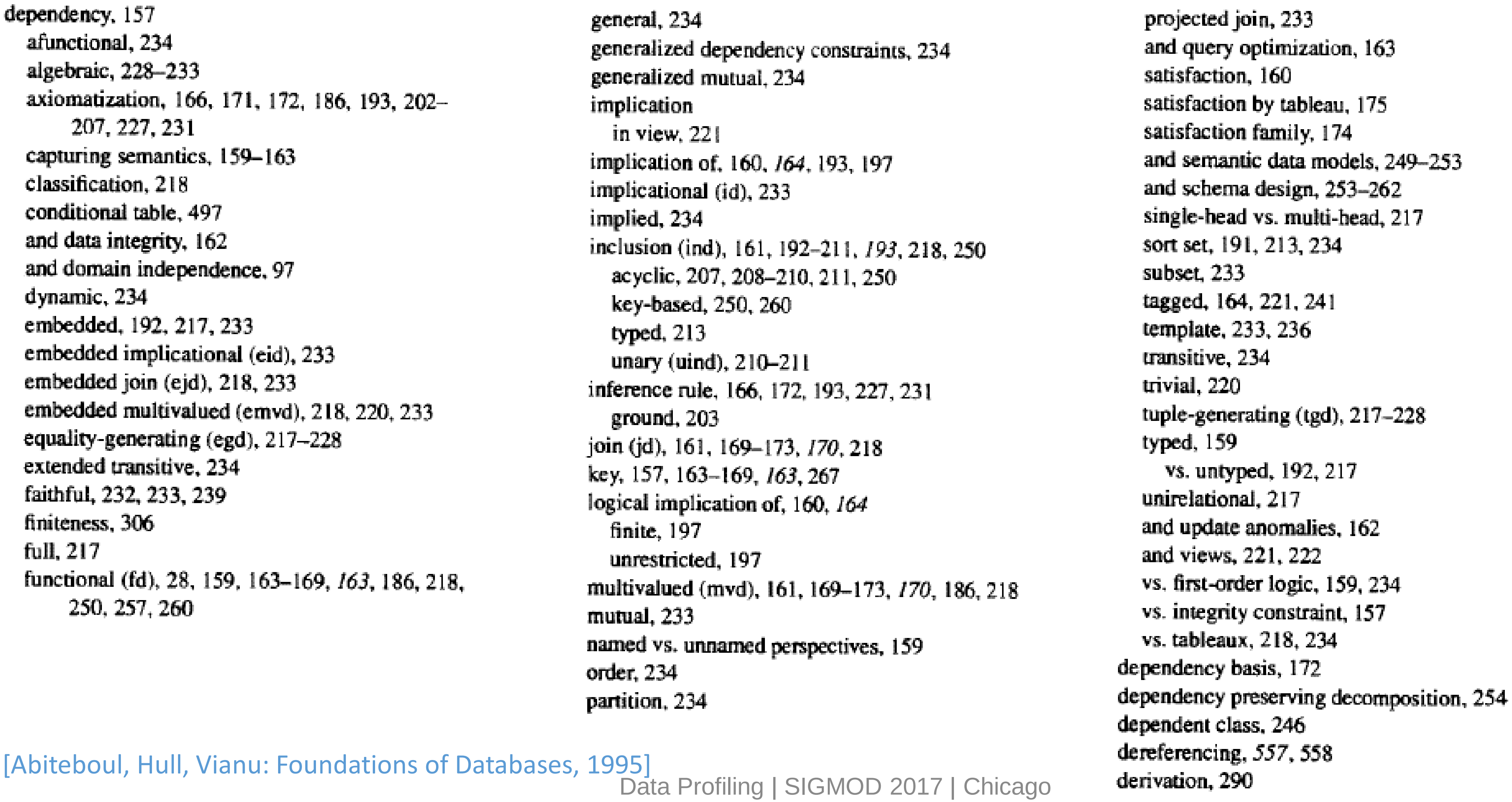
- Motivation
  - Task classification
  - Use cases
- Tools
  - Research and industry
  - Shortcomings
- Single and Multiple Column Analysis
  - Cardinalities and datatypes
  - Co-occurrences and summaries
- Dependencies
  - UCCs, FDs, ODs, INDs
  - and their discovery algorithms
- Outlook
  - Functionality
  - Semantics

# Part Overview

- Functional challenges
- Non-functional challenges
- Semantics of Dependencies

# Extending the Functionality of Data Profiling

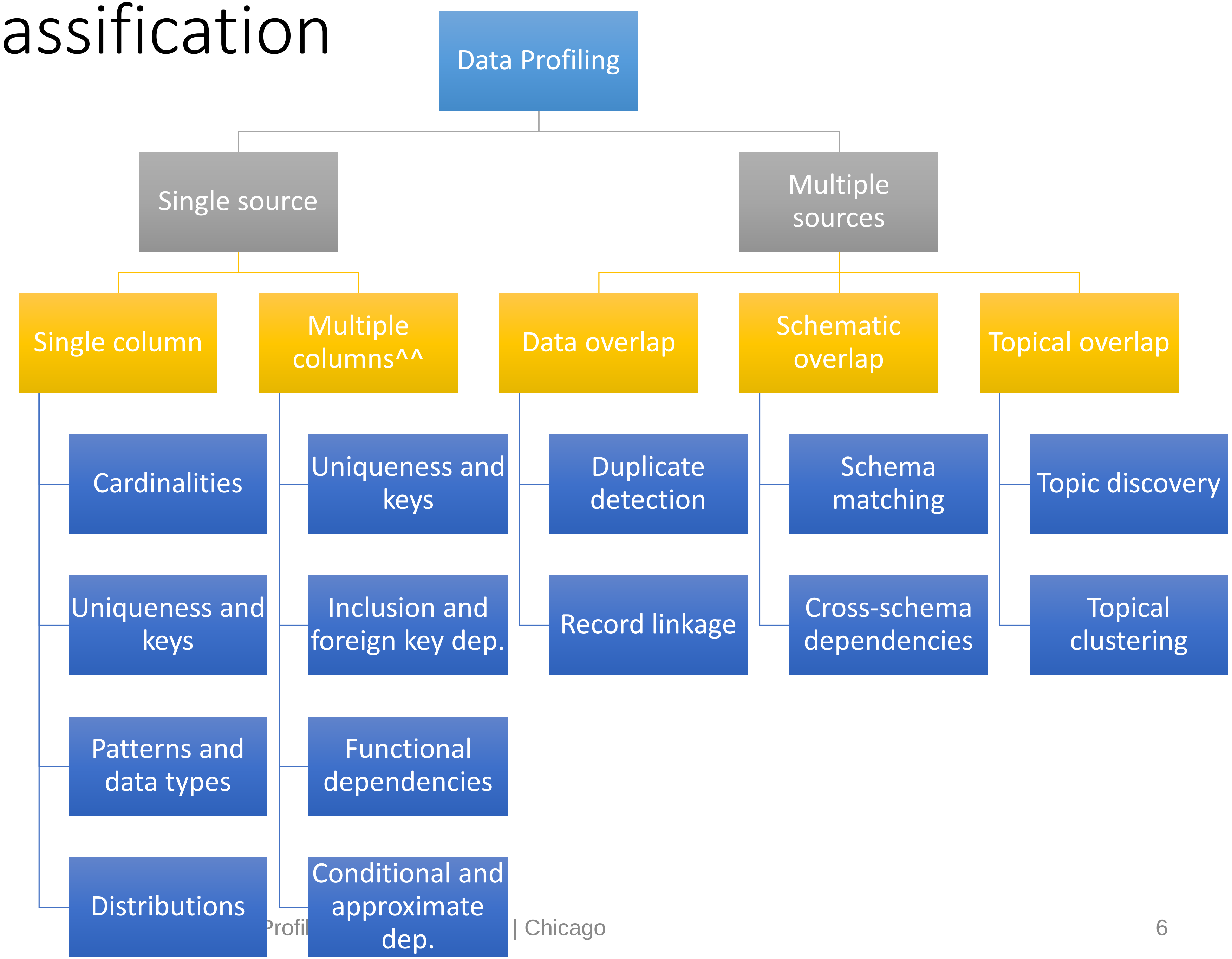
# Many Other Kinds of Dependencies



- dependency, 157
  - afunctional, 234
  - algebraic, 228–233
  - axiomatization, 166, 171, 172, 186, 193, 202–207, 227, 231
  - capturing semantics, 159–163
  - classification, 218
  - conditional table, 497
  - and data integrity, 162
  - and domain independence, 97
  - dynamic, 234
  - embedded, 192, 217, 233
  - embedded implicational (eid), 233
  - embedded join (ejd), 218, 233
  - embedded multivalued (emvd), 218, 220, 233
  - equality-generating (egd), 217–228
  - extended transitive, 234
  - faithful, 232, 233, 239
  - finiteness, 306
  - full, 217
  - functional (fd), 28, 159, 163–169, *163*, 186, 218, 250, 257, 260
  - general, 234
  - generalized dependency constraints, 234
  - generalized mutual, 234
  - implication
    - in view, 221
  - implication of, 160, *164*, 193, 197
  - implicational (id), 233
  - implied, 234
  - inclusion (ind), 161, 192–211, *193*, 218, 250
    - acyclic, 207, 208–210, 211, 250
    - key-based, 250, 260
    - typed, 213
    - unary (uind), 210–211
  - inference rule, 166, 172, 193, 227, 231
    - ground, 203
  - join (jd), 161, 169–173, *170*, 218
  - key, 157, 163–169, *163*, 267
  - logical implication of, 160, *164*
    - finite, 197
    - unrestricted, 197
  - multivalued (mvd), 161, 169–173, *170*, 186, 218
  - mutual, 233
  - named vs. unnamed perspectives, 159
  - order, 234
  - partition, 234
  - projected join, 233
  - and query optimization, 163
  - satisfaction, 160
  - satisfaction by tableau, 175
  - satisfaction family, 174
  - and semantic data models, 249–253
  - and schema design, 253–262
  - single-head vs. multi-head, 217
  - sort set, 191, 213, 234
  - subset, 233
  - tagged, 164, 221, 241
  - template, 233, 236
  - transitive, 234
  - trivial, 220
  - tuple-generating (tgd), 217–228
  - typed, 159
    - vs. untyped, 192, 217
  - unirelational, 217
  - and update anomalies, 162
  - and views, 221, 222
  - vs. first-order logic, 159, 234
  - vs. integrity constraint, 157
  - vs. tableaux, 218, 234
- dependency basis, 172
- dependency preserving decomposition, 254
- dependent class, 246
- dereferencing, *557*, 558
- derivation, 290

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

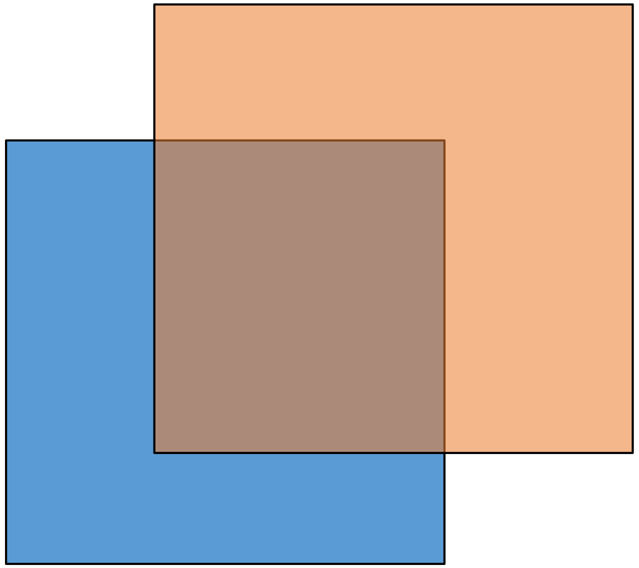
# Extended Classification of Profiling Tasks



- Data Profiling
  - Single source
    - Single column
      - Cardinalities
      - Uniqueness and keys
      - Patterns and data types
      - Distributions
    - Multiple columns^^
      - Uniqueness and keys
      - Inclusion and foreign key dep.
      - Functional dependencies
      - Conditional and approximate dep.
  - Multiple sources
    - Data overlap
      - Duplicate detection
      - Record linkage
    - Schematic overlap
      - Schema matching
      - Cross-schema dependencies
    - Topical overlap
      - Topic discovery
      - Topical clustering

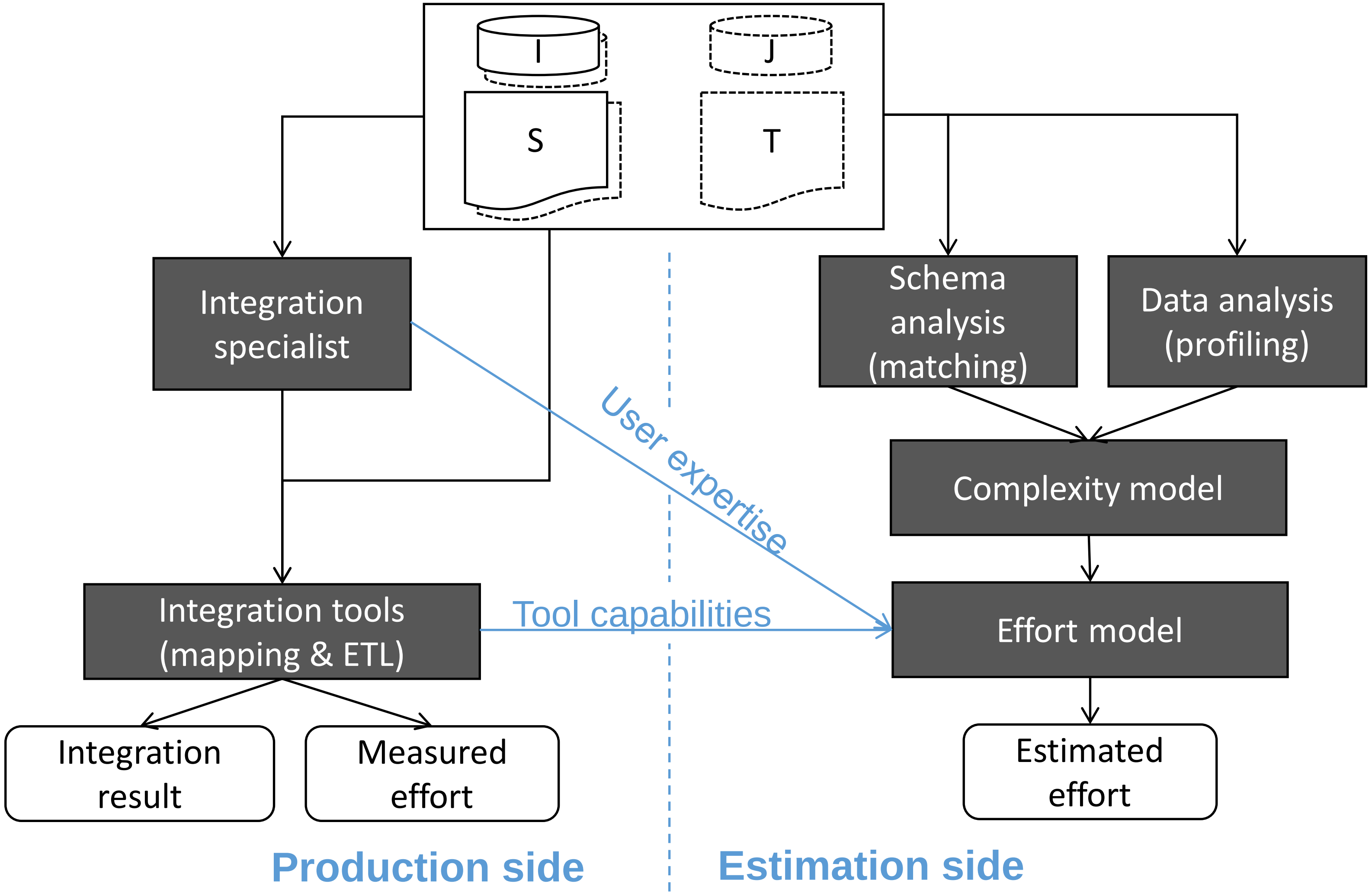
# Profiling for Integration

- Create measures to estimate integration (and cleansing) effort
  - Schema and data overlap
  - Severity of heterogeneity
- Schema matching/mapping
  - What constitutes the “difficulty” of matching/mapping?
- Duplicate detection
  - Estimate data overlap
  - Estimate fusion effort



- Overall: Determine integration complexity and integration effort
  - Intrinsic complexity: Schema and data
  - Extrinsic complexity: Tools and expertise

# Integration Effort Estimation



[Kruse, Papotti, Naumann: Estimating Data Integration and Cleaning Effort. EDBT 2015]

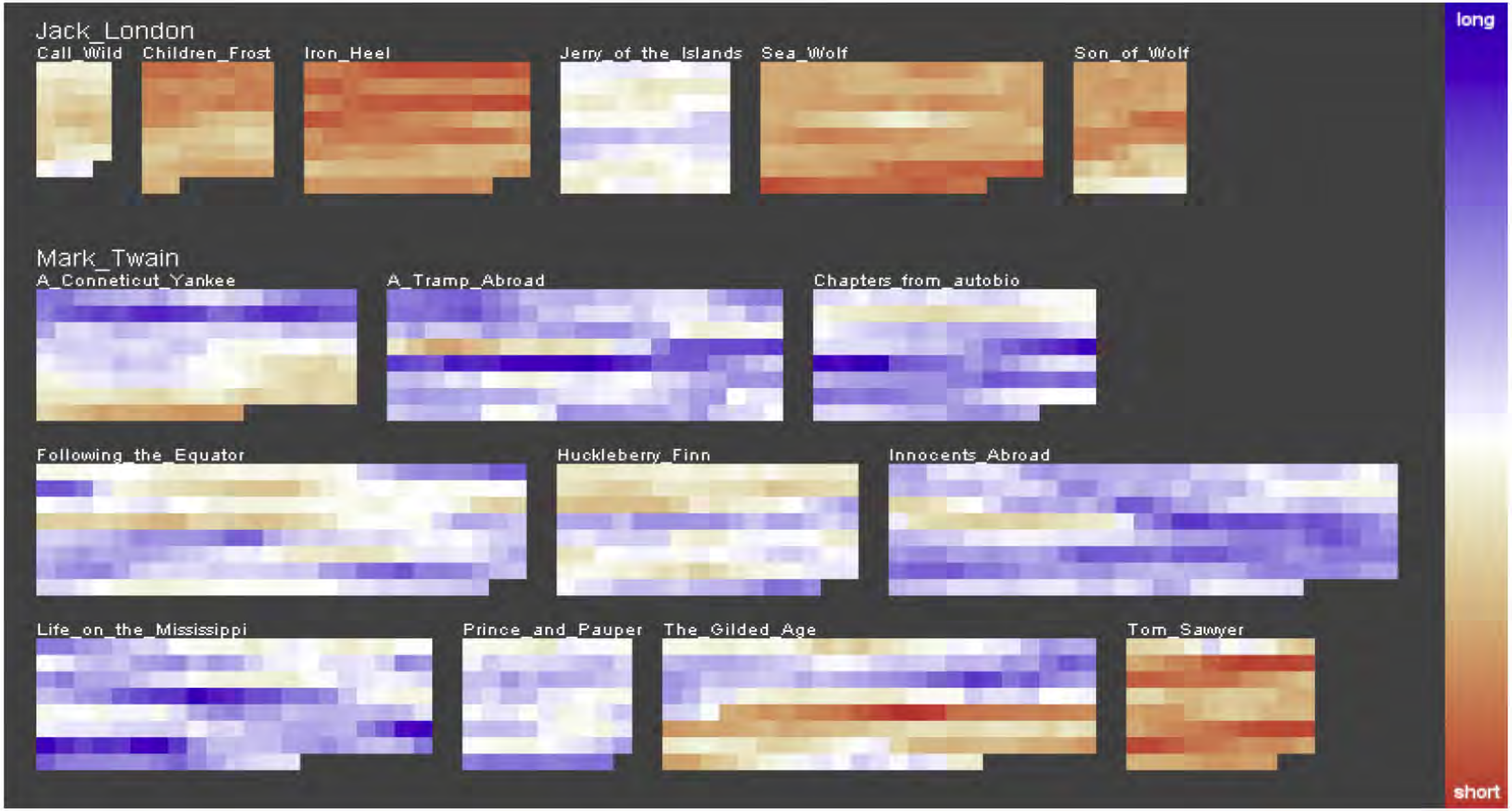
# Profiling new Types of Data

- 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

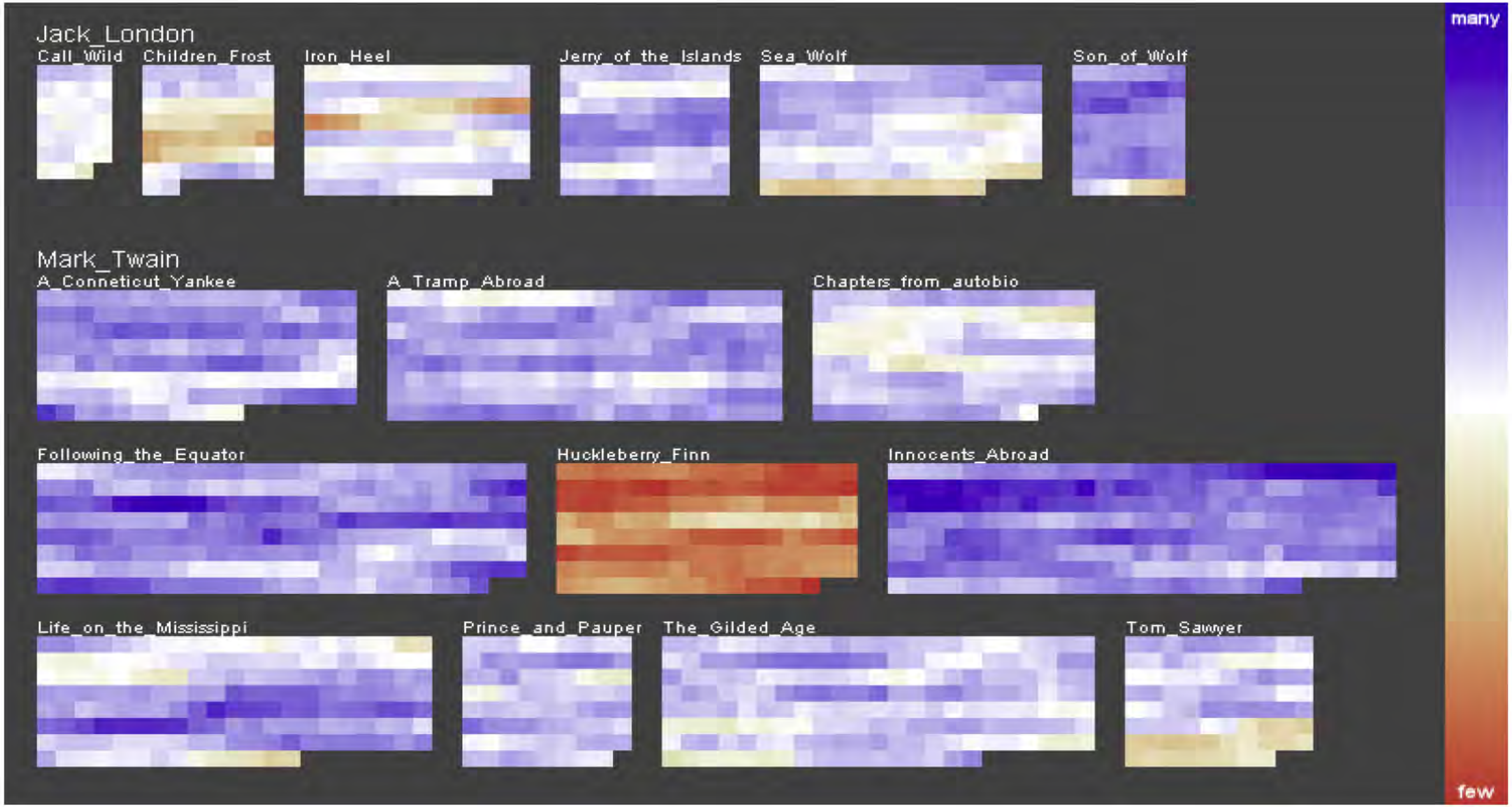
- Statistical measures
  - Syllables per word
  - Sentence length
  - Proportions of parts of speech
- Vocabulary measures
  - Frequencies of specific words
  - Type-token ratio
  - Simpson’s index (vocabulary richness)
  - Number of hapax (dis)legomena
    - Token that occurs exactly once (twice) in the corpus
    - Characterize style of an author

# Average Sentence Length



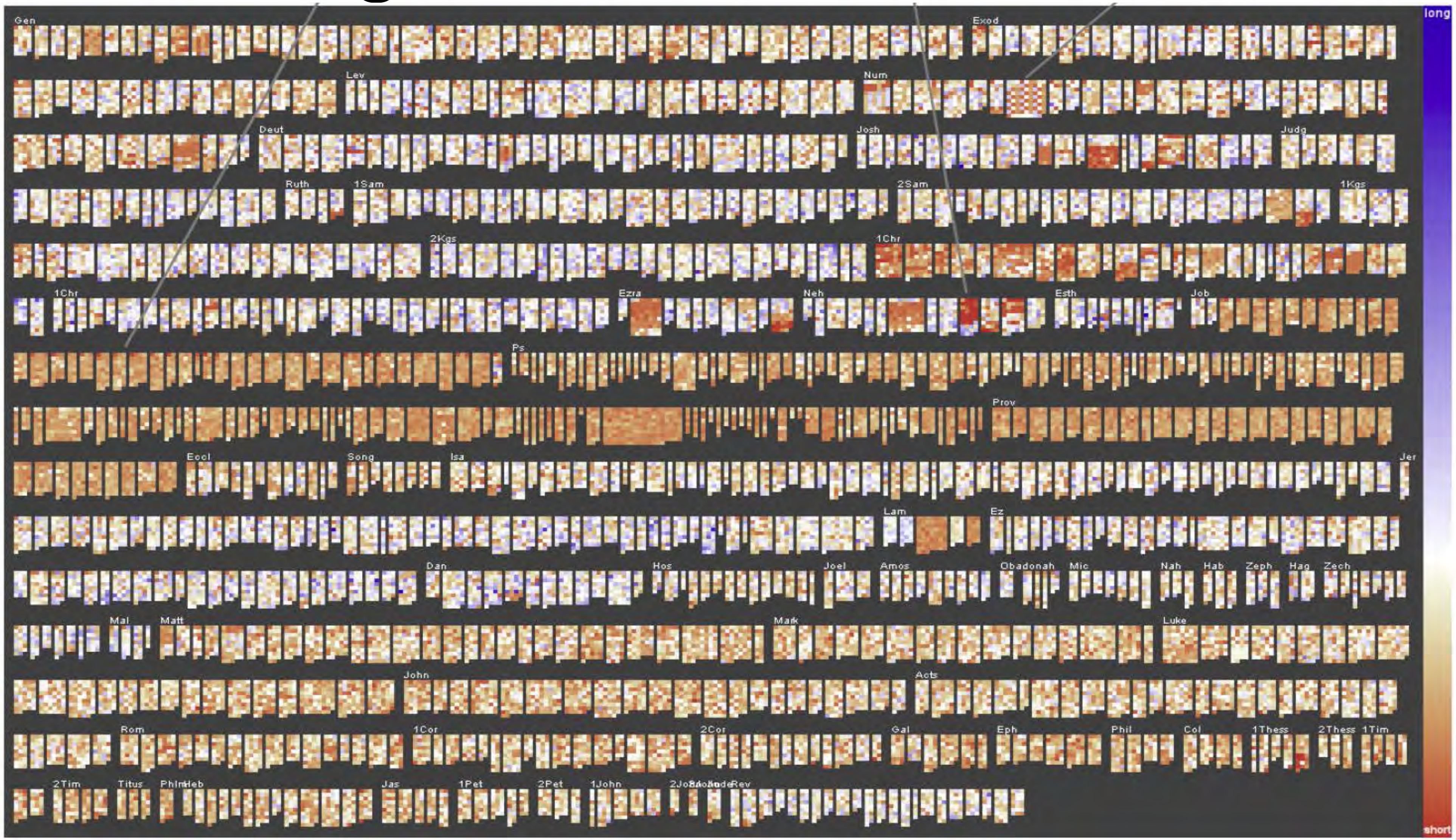
[Keim and Oelke: Literature Fingerprinting: A New Method for Visual Literary Analysis. IEEE VAST 2007]

# Hapax Legomena



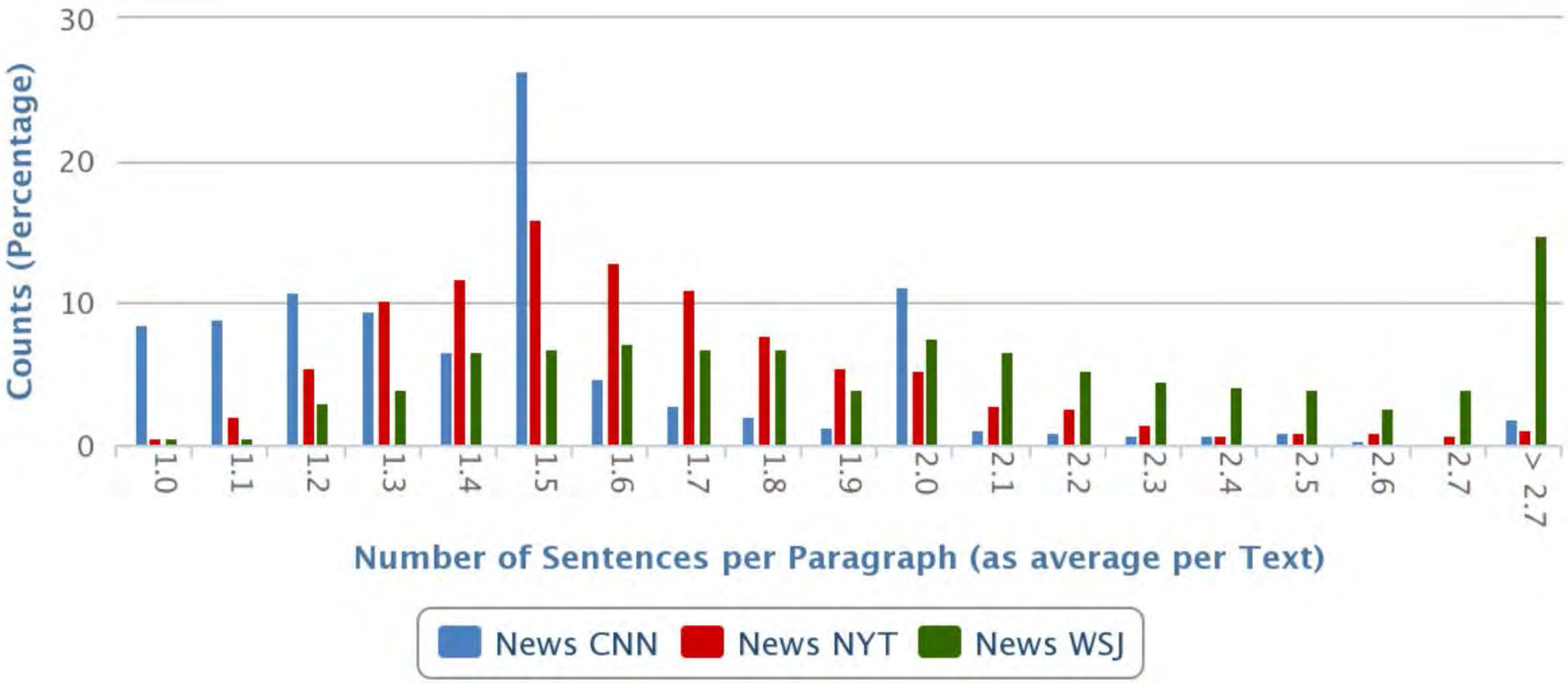
[Keim and Oelke: Literature Fingerprinting: A New Method for Visual Literary Analysis. IEEE VAST 2007]

# Verse Length



[Keim and Oelke: Literature Fingerprinting: A New Method for Visual Literary Analysis. IEEE VAST 2007]

# Example: News Article Statistics



# Improving Non-Functional Properties of Data Profiling

# Holistic Profiling

- Various profiling methods for various profiling tasks
- Commonalities/similarities
    - Search space: All column combinations (or pairs thereof)
    - I/O: Read all data at least once
    - Data structure: Some index or hash table
    - Pruning and candidate generation: based on subset/superset relationships
    - Sortation: Benefit from sorted sets

- Challenge: Develop single method to output all/most profiling results

# Incremental Profiling

- Data is dynamic
  - Insert (batch or tuple-based)
  - Updates
  - Deletes
- Problem: Keep profiling results up-to-date without reprofiling the entire data set
  - Easy examples: SUM, MIN, MAX, COUNT, AVG
  - Difficult examples: MEDIAN, uniqueness, FDs, etc.

# Online Profiling

- Profiling is long procedure
  - Boring for developers
  - Expensive for machines (I/O and CPU)

- Challenge: Display intermediate results
  - … of improving/converging accuracy
  - Allows early abort of profiling run

- Gear algorithms toward that goal
  - Allow intermediate output
  - Enable early output: “progressive” profiling

# Temporal Profiling

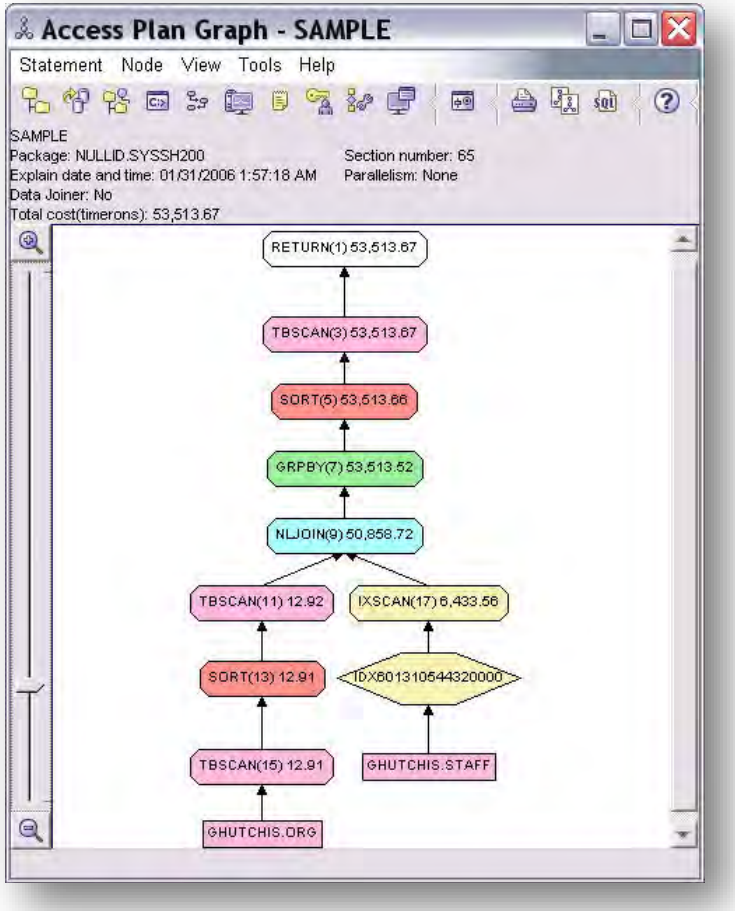
- Observe behavior of dependencies over time
  - Do FDs appear and disappear?
  - Does a partial IND become less partial over time?
  - …

- Metadata monitoring
  - Meta-Metadata

# Profiling Query Results

- Query results are boring: Spruce them up with some metadata
  - Usually only: Row count
  - For each column, give some statistics

- Idea: Piggy-back profiling on query execution
  - Re-use sortations, hash tables, etc.



SAMPLE
Package: NULLID.SYSSH200 Section number: 65
Explain date and time: 01/31/2006 1:57:18 AM Parallelism: None
Data Joiner: No
Total cost(timerons): 53,513.67

# Data Generation and Testing

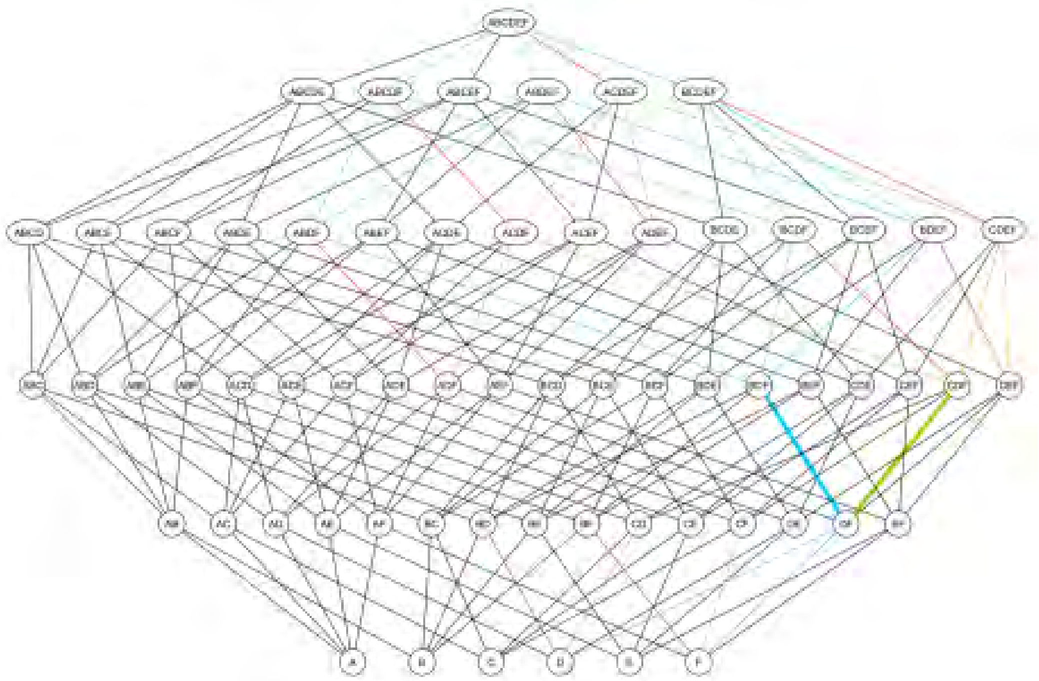
- Generate volumes of data with certain properties
  - Test extreme cases
  - Test scalability

- Problem: Interaction between properties
  - FDs vs. uniqueness
  - Patterns vs. conditional INDs
  - Distributions vs. all others…
- Problem: Create realistic data
  - Distributions, patterns
  - Placement of dependencies (tight or spread out)

Recent work
[Arocena et al. : Messing Up with BART: Error Generation for Evaluating Data-Cleaning Algorithms. PVLDB 9(2), 2015]
[Arocena et al. : The iBench Integration Metadata Generator . PVLDB 9(3), 2015]

# Data Profiling Benchmark

- Define data
  - Data generation
  - Real-world dataset(s)
  - Different scale-factors: Rows and columns
- Define tasks
  - Individual tasks
  - Sets of tasks
- Define measures
  - Speed
  - Speed/cost
  - Minimum hardware requirements
  - Accuracy for approximate approaches



# Summary – much to do

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

# Semantic Interpretation of Profiling Results

# Turning Instance-based Observations to Schema-based Constraints

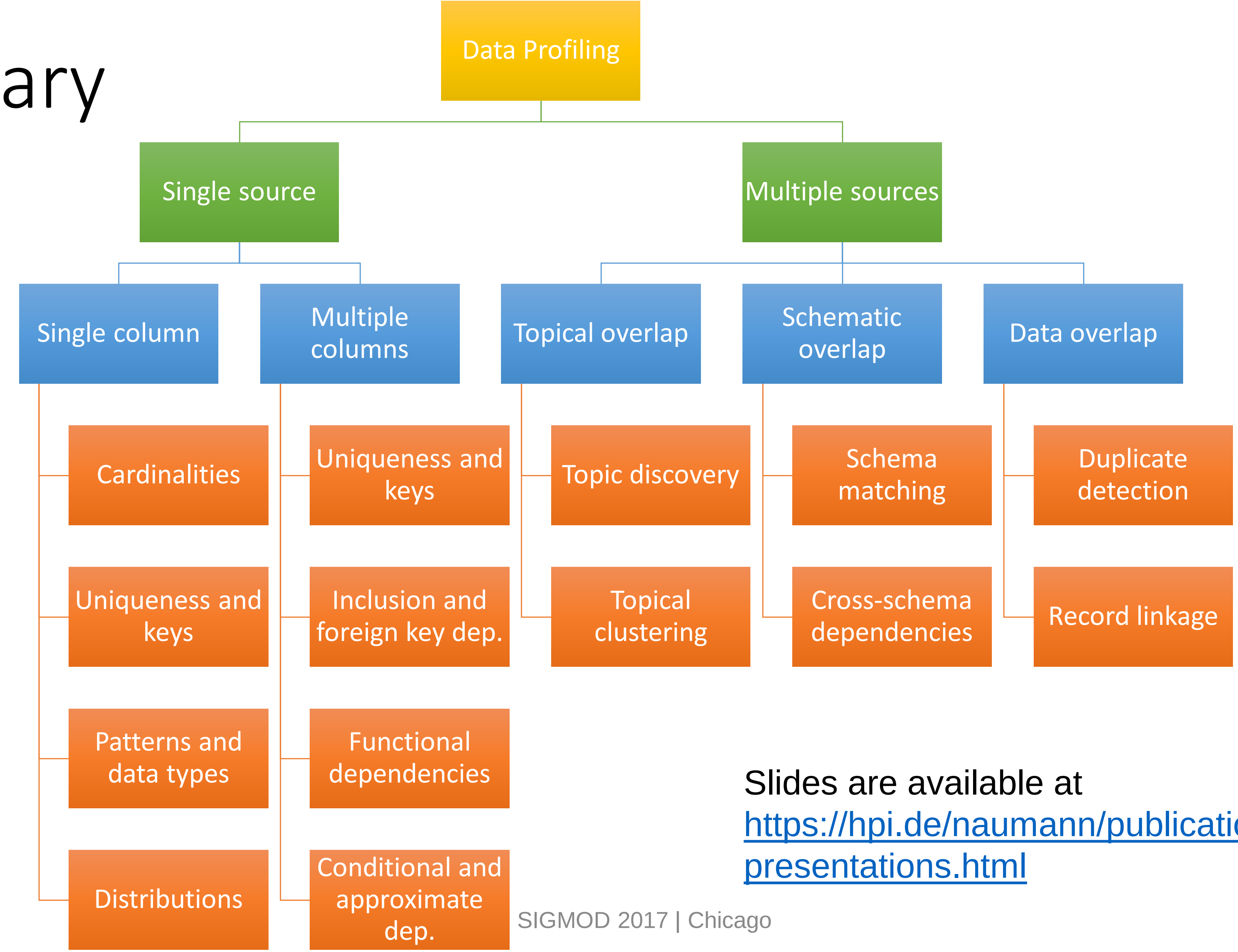
- Hundreds of UCCs – which ones are keys?
- Thousands of FDs – which ones are true?
- Millions of INDs – which ones are foreign keys?

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

# Thanks to co-authors, colleagues and team!

- Carl Ambroselli (Metanome, HPI)
- Jana Bauckmann (INDs, HPI)
- Tanja Bergmann (Metanome, HPI)
- Jens Ehrlich (cUCCs, HPI)
- Claudia Exeler (Metanome, HPI)
- Moritz Finke (Metanome, HPI)
- Toni Gruetze (ProLOD, HPI)
- Hazar Harmouch (Cardinalities, HPI)
- Arvid Heise (UCCs, HPI)
- Anja Jentzsch(ProLOD, HPI)
- Sebastian Kruse (INDs, Metadata Store, HPI)
- Philipp Langer (ODs, HPI)
- Thorsten Papenbrock (INDs, FDs, Metanome, HPI)
- Paolo Papotti (Profiling for Integration, ASU)
- Jorge Quiané-Ruiz (UCCs, QCRI)
- Patrick Schulze (FDs, HPI)
- Fabian Tschirschnitz (INDs, HPI)
- Jakob Zwiener (Metanome, HPI)

# Summary



Slides are available at [https://hpi.de/naumann/publications/selected-presentations.html](https://hpi.de/naumann/publications/selected-presentations.html)