Data Profiling

Ziawasch Abedjan
(TU Berlin)

My Background

2005 | 2013 | 2014 | 2016 | now

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

BS, MS, PhD

Internship

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
عضو في مؤسسة قطر
Member of Qatar Foundation

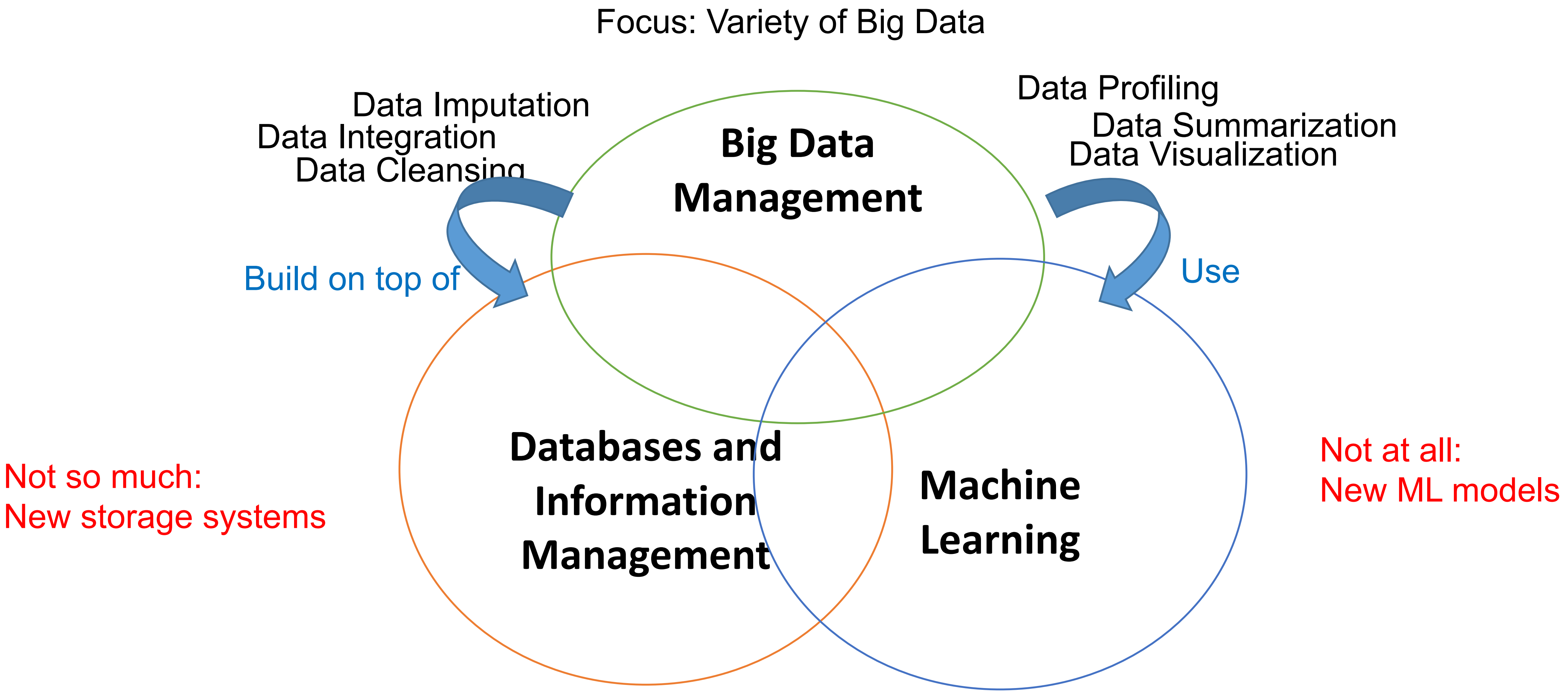
PostDoc

berlin

Juniorprofessor

BBDC
BERLIN BIG DATA CENTER

Big Data Management Group (BigDaMa)



BigDaMa People

Ziawasch Abedjan
Head of the BigDaMa

Mohammad Mahdavi
PhD Student

Data Cleaning

Larysa Visengeriyeva
PhD Student

Data Cleaning

Umar Maqsud
PhD Student, DFKI

Data Streams
Data Cleaning

Maximilian Dohlus
PhD Student, PTB

Smart Data Extraction

Emergence of Data Driven Applications

Big data: The next frontier for innovation, competition, and productivity

McKinsey&Company

Harvard Business Review

DATA

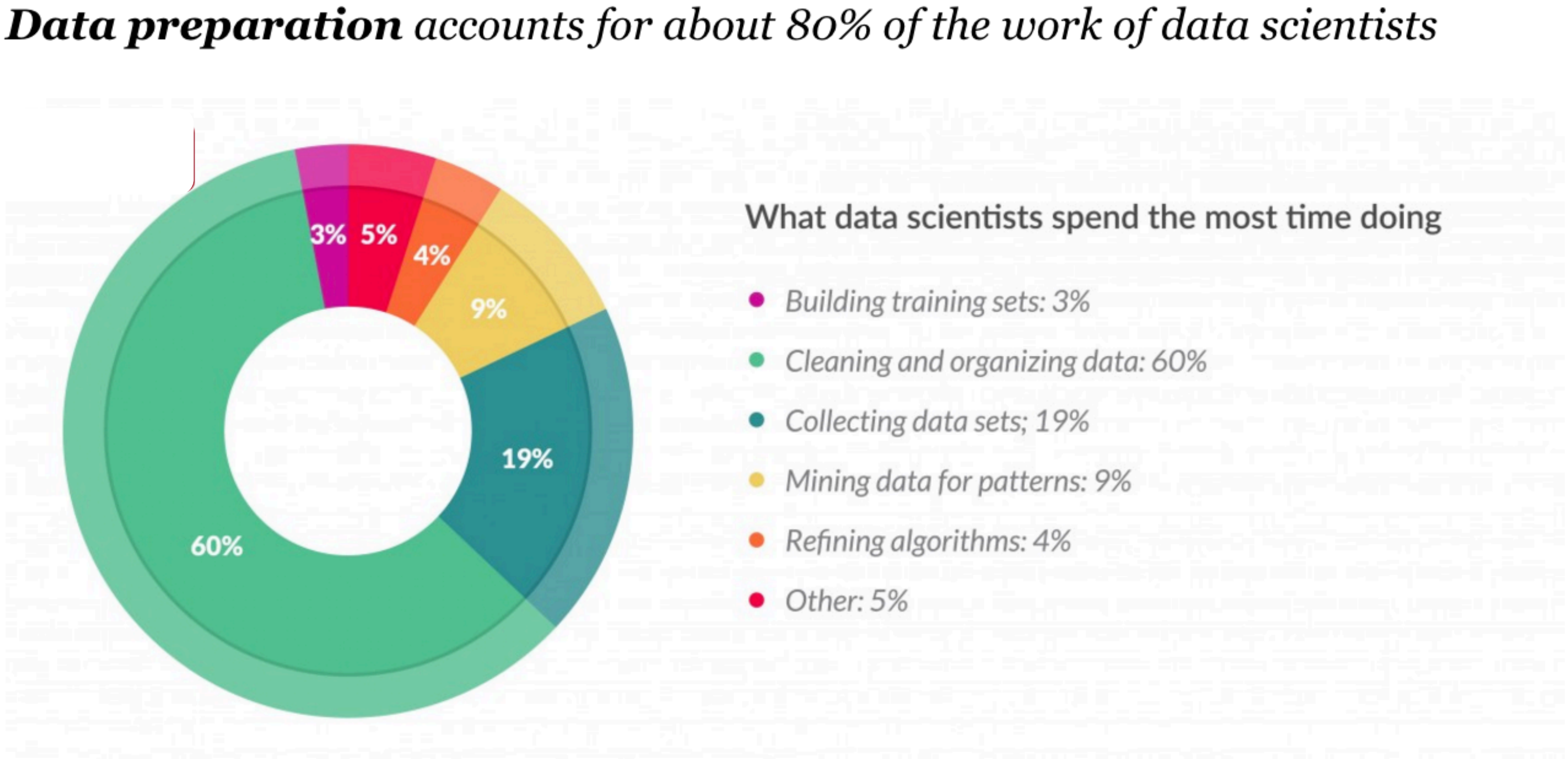
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

But, what do data scientists actually do?

CrowdFlower’s Data Science Report 2016



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

“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

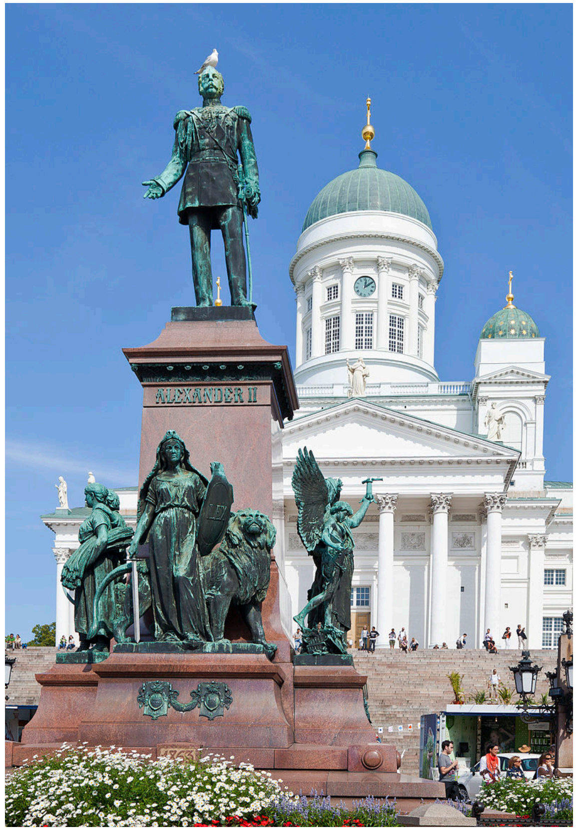
Abstract Profiling data to determine metadata about a given dataset is an important and frequent activity of any IT professional and researcher and is necessary for various use-cases. It encompasses a vast array of methods to examine datasets and produce metadata. Among the simpler results are statistics, 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, INDs, FDs
  - and their discover 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. *Foundations 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 Han, 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.

ncvoter1.txt - Microsoft Excel

Datei | Start | Einfügen | Seitenlayout | Formeln | Daten | Überprüfen | Ansicht | Add-Ins

A2 | f_x | 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 CREEI	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	CARMICHAEL		611 N THIRD ST	MEBANE	NC	27302	611 N THIRD ST		MEBANE	NC	27302	563 2654	W	NL	UNA
25	1	ALAMANCE	9102615	A	ACTIVE	AV	VERIFIED	ABBOTT	BRIAN	CHRISTOPHE		2006 WINN CREEK	HAW RIVER	NC	27258	2006 WINN CREEK DR		HAW RIVER	NC	27258	336 261 3357	W	NL	UNA
26	1	ALAMANCE	9079257	A	ACTIVE	AV	VERIFIED	ABBOTT	BRUCE	CLEATON		188 LAKE CAMMA	BURLINGTON	NC	27217	188 LAKE CAMMACK CT		BURLINGTON	NC	27217	336 214 2703	W	NL	REP
27	1	ALAMANCE	1389300	A	ACTIVE	AV	VERIFIED	ABBOTT	CHERYL	FAULKNER		188 LAKE CAMMA	BURLINGTON	NC	27217	188 LAKE CAMMACK CT		BURLINGTON	NC	27217	336 229 3027	W	NL	REP
28	1	ALAMANCE	9140392	A	ACTIVE	AV	VERIFIED	ABBOTT	CHRISTOPHE	BRANDON		309 BURLINGTON	GIBSONVILLE	NC	27249	309 BURLINGTON AVE		GIBSONVILLE	NC	27249		W	NL	UNA
29	1	ALAMANCE	9135711	A	ACTIVE	AV	VERIFIED	ABBOTT	COURTNEY	LOVE		309 BURLINGTON	GIBSONVILLE	NC	27249	309 BURLINGTON AVE		GIBSONVILLE	NC	27249		W	NL	UNA
30	1	ALAMANCE	9028439	A	ACTIVE	AV	VERIFIED	ABBOTT	DWAYNE	ROGER		2839 LADALE LN	MEBANE	NC	27302	2839 LADALE LN		MEBANE	NC	27302	563 3956	W	NL	UNA
31	1	ALAMANCE	9090420	A	ACTIVE	AV	VERIFIED	ABBOTT	FRANK	PATRICK		1202 JAMESTOWN	ELON	NC	27244	1202 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

ncvoter1

Bereit | 100 %

ncvoter1.txt - Microsoft Excel

Datei | Start | Einfügen | Seitenlayout | Formeln | Daten | Überprüfen | Ansicht | Add-Ins

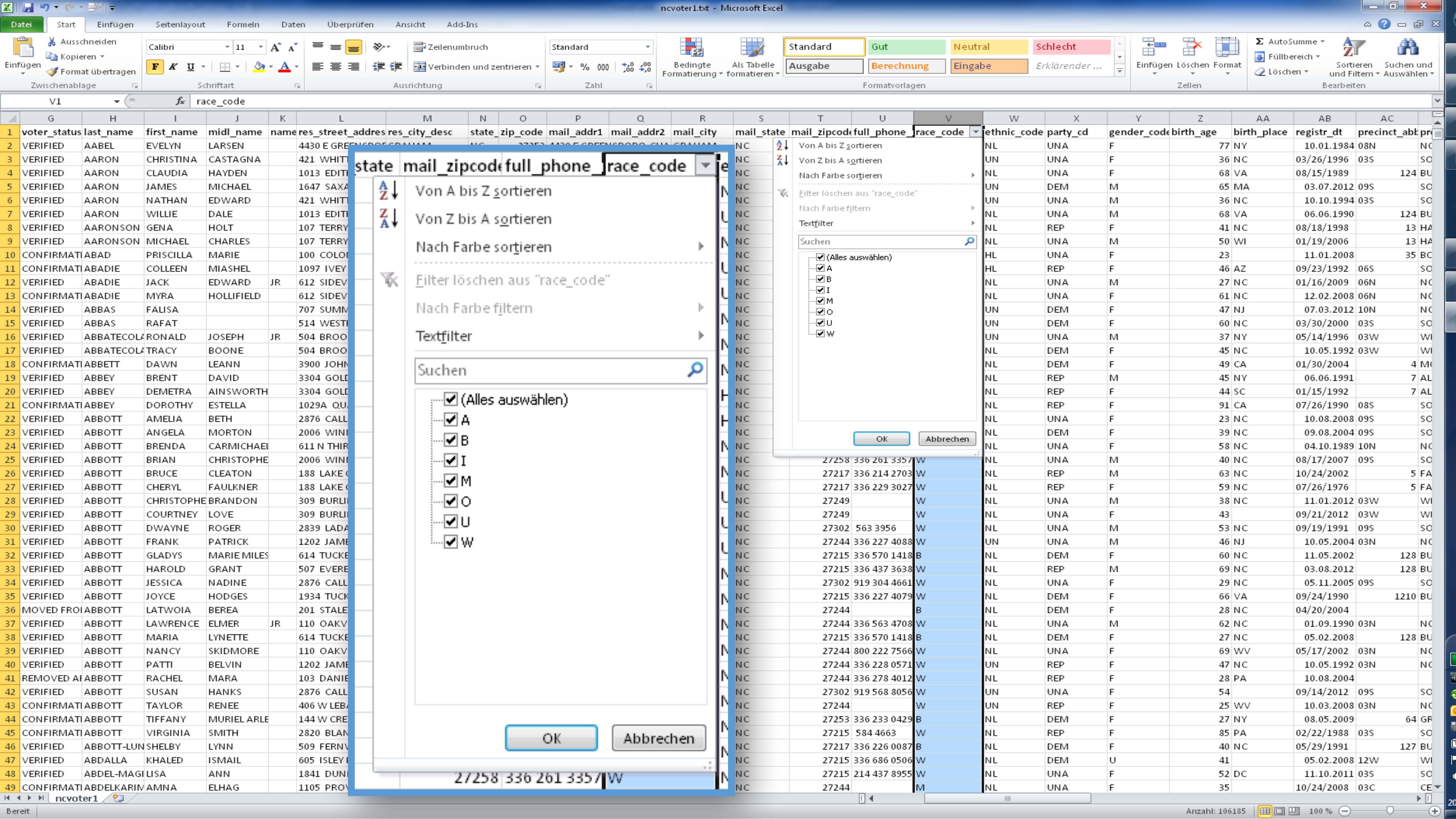
A106185 | f_x | 1

	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	ZOLLICOFFEF	ANTONIO	MARK		108 OAKGROVE DI	GRAHAM	NC	27253	108 OAKGROVE DR		GRAHAM	NC	27253	336 260 6673	B	UN	DEM
106155	1	ALAMANCE	9107068	A	ACTIVE	AV	VERIFIED	ZOLLICOFFEF	VALERIE			108 OAKGROVE DI	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 I	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 CF	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 CF	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 CF	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	[illegible]	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	[illegible]	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	[illegible]	[illegible]	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	[illegible]	[illegible]	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	ALAMA[illegible]CE	914[illegible]	[illegible]	[illegible]	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[illegible]	[illegible]	[illegible]	[illegible]	AV	VERIFIED	ZURFACE	ROSSELL	EUGENE		2074 TURNER RD	MEBANE	NC	27302	2074 TURNER RD		MEBANE	NC	27302		W	UN	UNA
106181	1	ALAM[illegible]	[illegible]	[illegible]	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[illegible]	[illegible]	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[illegible]	[illegible]140499	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[illegible]	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 BRC	MEBANE	NC	27302	1210 WILLOW BROOK CT		MEBANE	NC	27302	336 578 8580	W	UN	REP
106186																								

Number of rows

ncvoter1

Bereit | 100 %


ncvoter1.txt - Microsoft Excel
Datei
Start
Einfügen
Seitenlayout
Formeln
Daten
Überprüfen
Ansicht
Add-Ins
V1
race_code
state
mail_zipcode
full_phone
race_code
Von A bis Z sortieren
Von Z bis A sortieren
Nach Farbe sortieren
Filter löschen aus "race_code"
Nach Farbe filtern
Textfilter
Suchen
(Alles auswählen)
A
B
I
M
O
U
W
OK
Abbrechen
ncvoter1
Bereit
Anzahl: 106185
100 %

[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	I	INACTIVE	[illegible]
[illegible]	[illegible]	I	INACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	I	INACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	A	ACTIVE	[illegible]
[illegible]	[illegible]	[illegible]	[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

Definition Data Profiling

- Data profiling is the process of examining the data available in an existing data source [...] and collecting statistics and information about that data.

  [Wikipedia 04/2016]

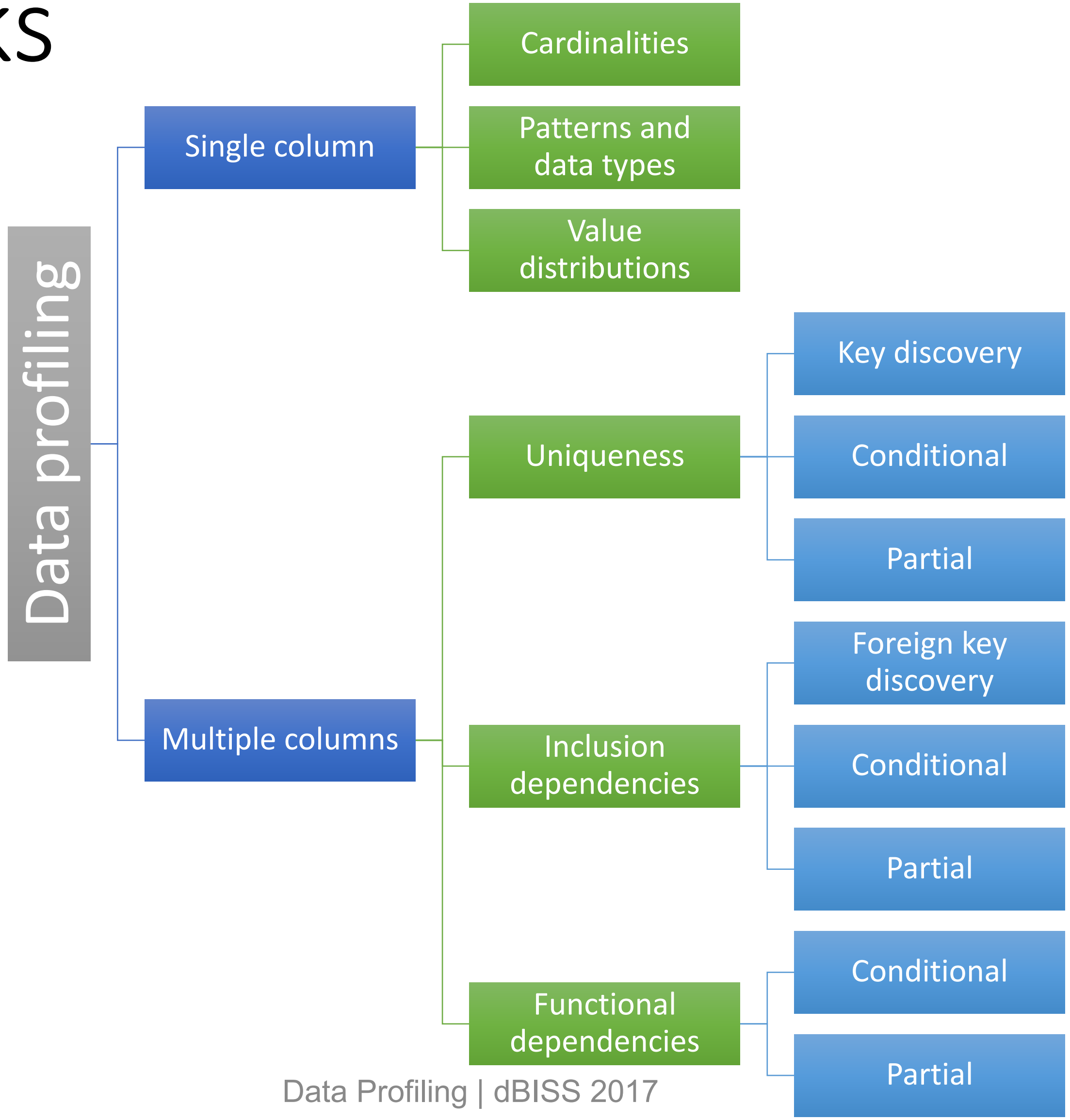
- Data profiling refers to the activity of creating small but informative summaries of a database.

  [Ted Johnson, Data Profiling, Encyclopedia of Database Systems, 2009]

- Data profiling is the set of activities and processes to determine the metadata about a given dataset.

- A fixed set of data profiling tasks / results

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

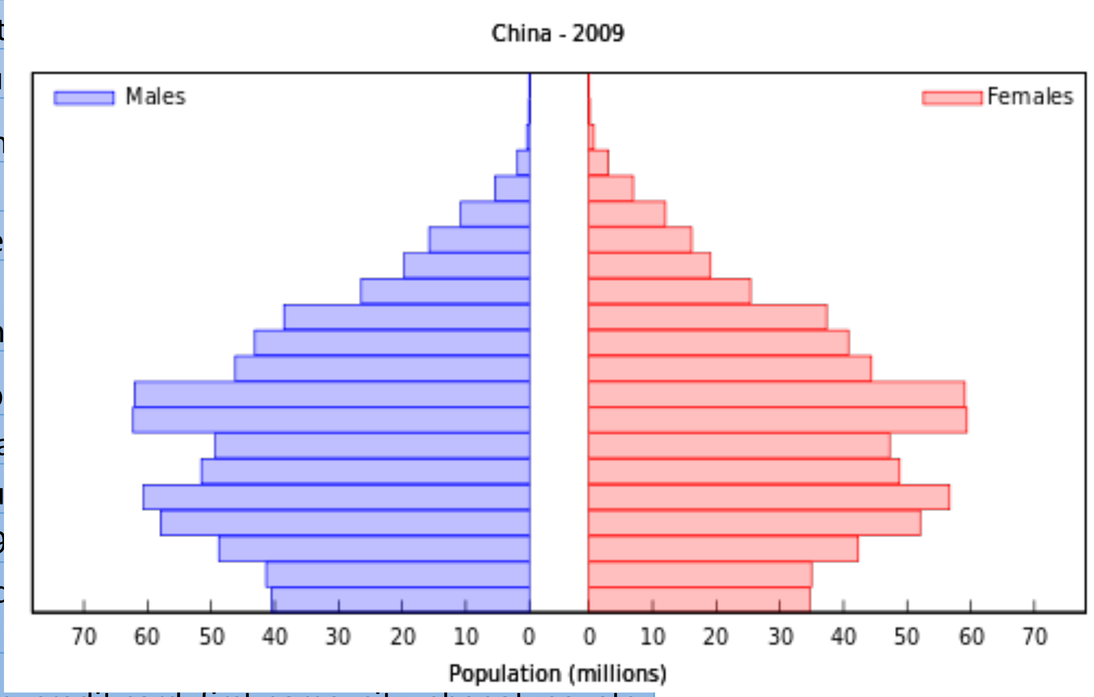
Use Cases for Profiling

- 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

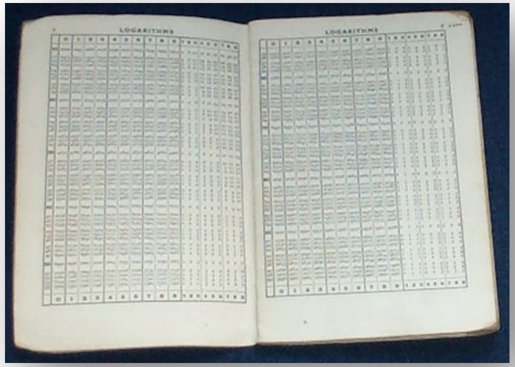
Cardinalities, Distributions, and Patterns

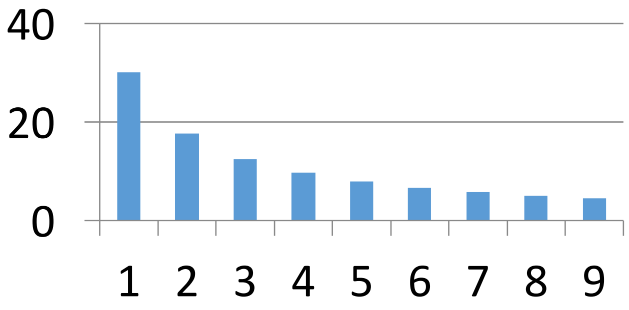
Category	Task	Description
Cardinalities	num-rows	Number of rows
	value length	Measurements of value lengths (min, max, median, and average)
	null values	Number or percentage of null values
	distinct	Number of distinct values; aka “cardinality”
	uniqueness	Number of distinct values divided by number of 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 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

- 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%

Fraud detection

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

Uses for Basic Statistics

- Traditional uses
  - Query optimization
  - Outlier/error detection
  - Visualize distribution

- Semantic uses
  - Categorization of attributes: Data types
  - Relevance of attributes: Completeness and quality
  - Semantics of attributes: Matching and cleansing

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

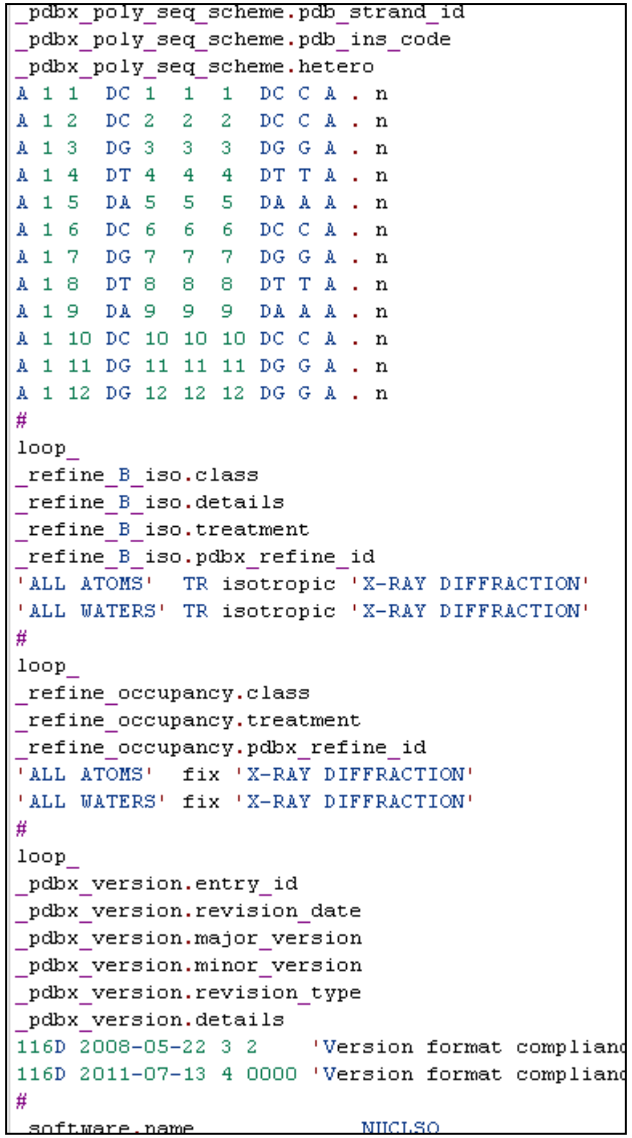
Inclusion Dependencies

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

- 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)

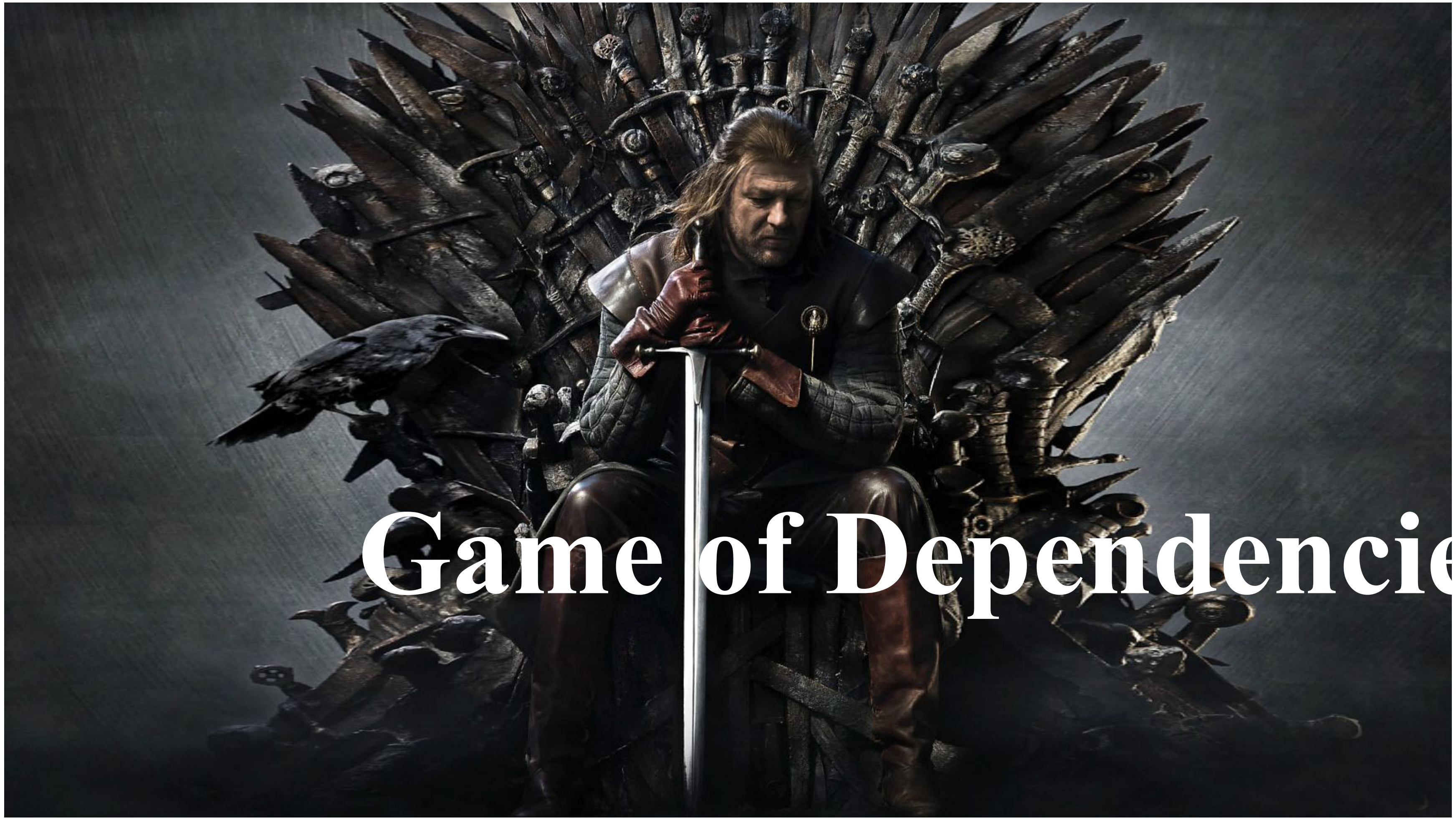
emp_name	rank	salary
Smith	1	40k
Johnson	1	40k
Williams	1	45k
Brown	2	60k
Davis	2	60k
Miller	3	70k
Wilson	4	100k

Uses for FDs

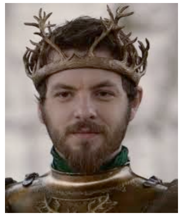
- Schema design
  - Normalization
  - Keys
- Data cleansing
- Schema design and normalization
- Key discovery
- Data cleansing (especially partial/conditional FDs)
- Anomaly detection
  - Data integrity constraints
  - Data curation rules
- Query optimization: Independence of column attributes
- Index 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, H r → Lineage
6. …

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

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


Properties of
Dependencies

Partial Dependencies

- Aka. “approximate dependencies”
- INDs and FDs that do not perfectly hold
  - For all but 10 of the tuples
  - Only for 80% of the tuples
  - Only for 1% of the tuples
- Also for patterns, types, uniques, and other constraints
- Useful for: Data cleansing

Conditional Dependencies

- Given a partial IND or FD: For **which** part do the hold?
- Expressed as a condition over the attributes of the relation
- Problems:
  - Infinite possibilities of conditions
  - Interestingness:
    - Many distinct values: less interesting
    - Few distinct values: surprising condition – high coverage

- Useful for Integration
  - Cross-database cINDs

Other (Relaxed) Dependencies

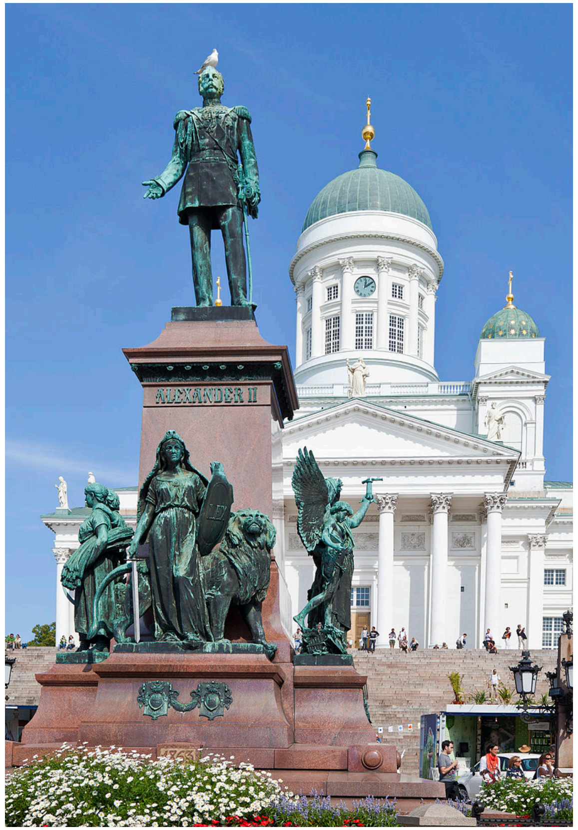
- Partial dependencies
- Approximate dependencies
- Conditional dependencies
- Matching dependencies
- Metric dependencies

RFD abbrev.	RFD name
ACOD	Approximate comparable dependency
ADD	Approximate differential dependency
AFD	Approximate functional dependency
COD	Comparable dependency
CFD	Conditional functional dependency
CFD^p	CFD with built-in predicates
CFD^c	CFD with cardinality constraints and synonym rules
CMD	Conditional matching dependency
CSD	Conditional sequential dependency
CD	Constrained functional dependency
DD	Differential dependency
eCFD	Extended conditional functional dependency
FFD	Fuzzy functional dependency
MD	Matching dependency
MFD	Metric functional dependency
ND	Neighborhood dependency
NuD	Numerical dependency
OD	Order dependency
OD_K	OD satisfied within bound k
OD_{EA}	OD satisfied almost everywhere
OFD	Ordered functional dependency
PD	Partial determination
POD	Polarized order dependencies
preFD	Preference functional dependency
PAC	Probabilistic approximate constraint
pFD	Probabilistic functional dependency
PuD	Purity dependency
RUD	Roll-up dependency
SD	Sequential dependency
SFD	Similarity functional dependency
soft FD	Soft functional dependency
[illegible]	[illegible]cy
TMFD	Type-M functional dependency
XCFD	XML conditional functional dependency
$\sigma\theta$XFD	XML FD with σ and θ approximation

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

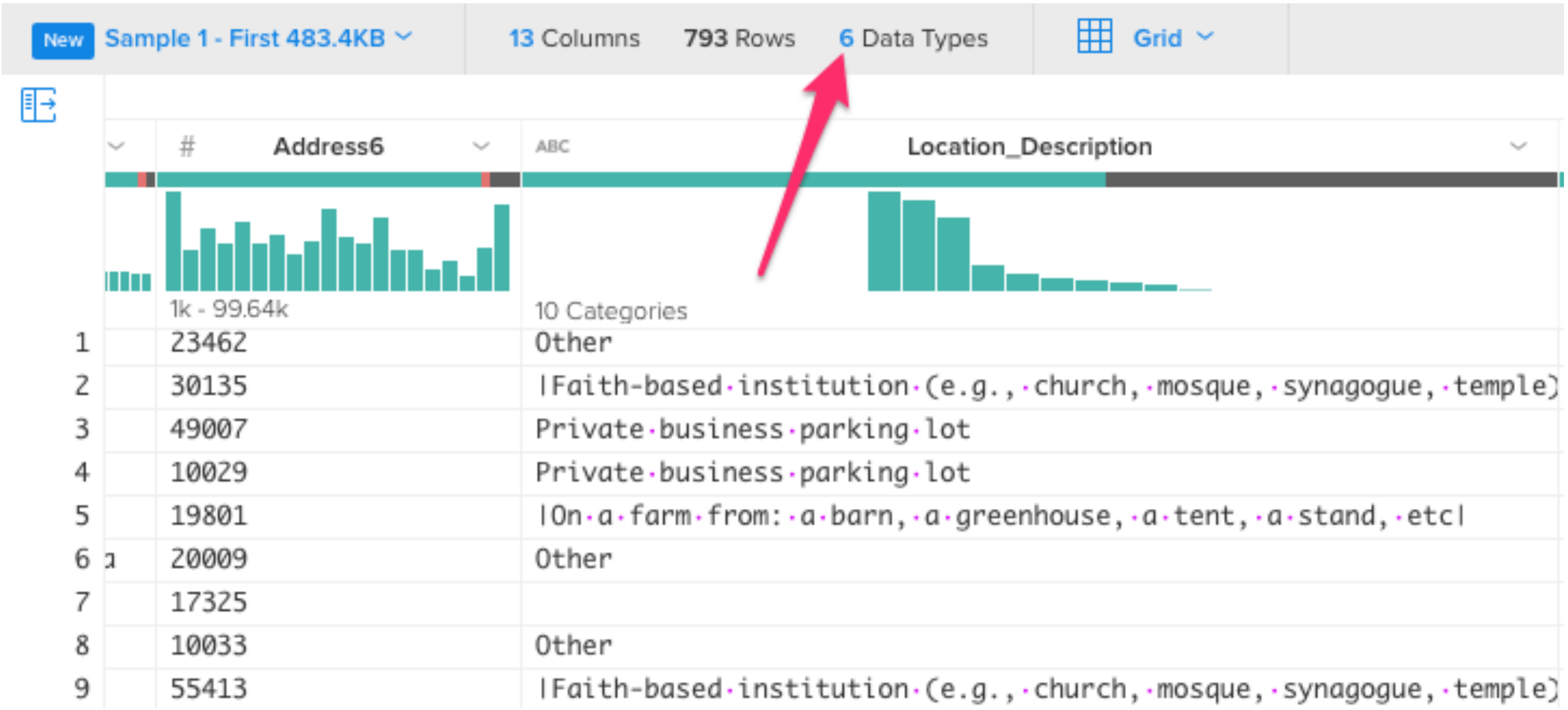
Tutorial Overview

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

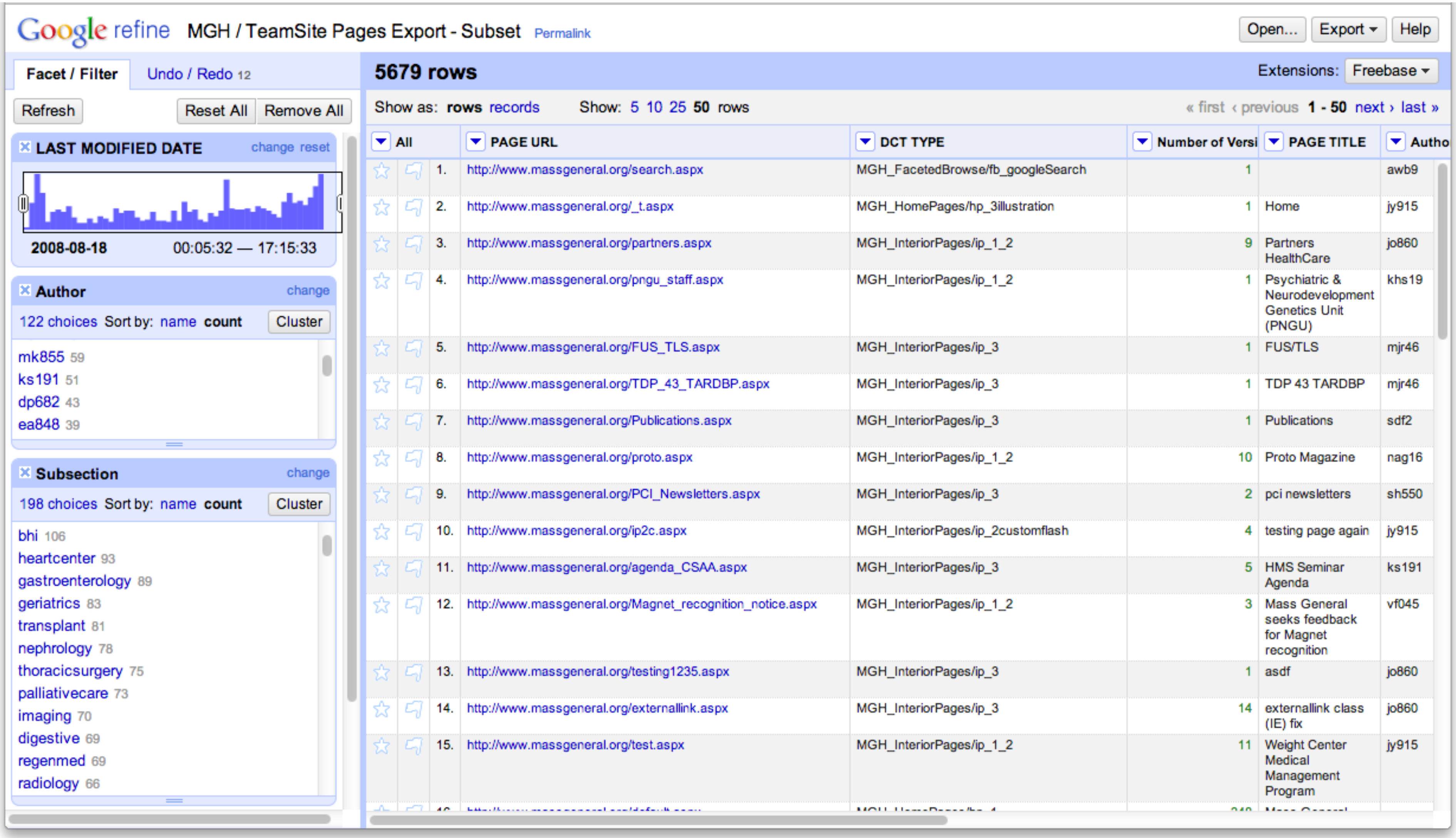


Tools in Industry

Trifacta


New
Sample 1 - First 483.4KB
13 Columns
793 Rows
6 Data Types
Grid
Address6
ABC
Location_Description
1k - 99.64k
10 Categories
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 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

Uses Cases Covered By Industrial

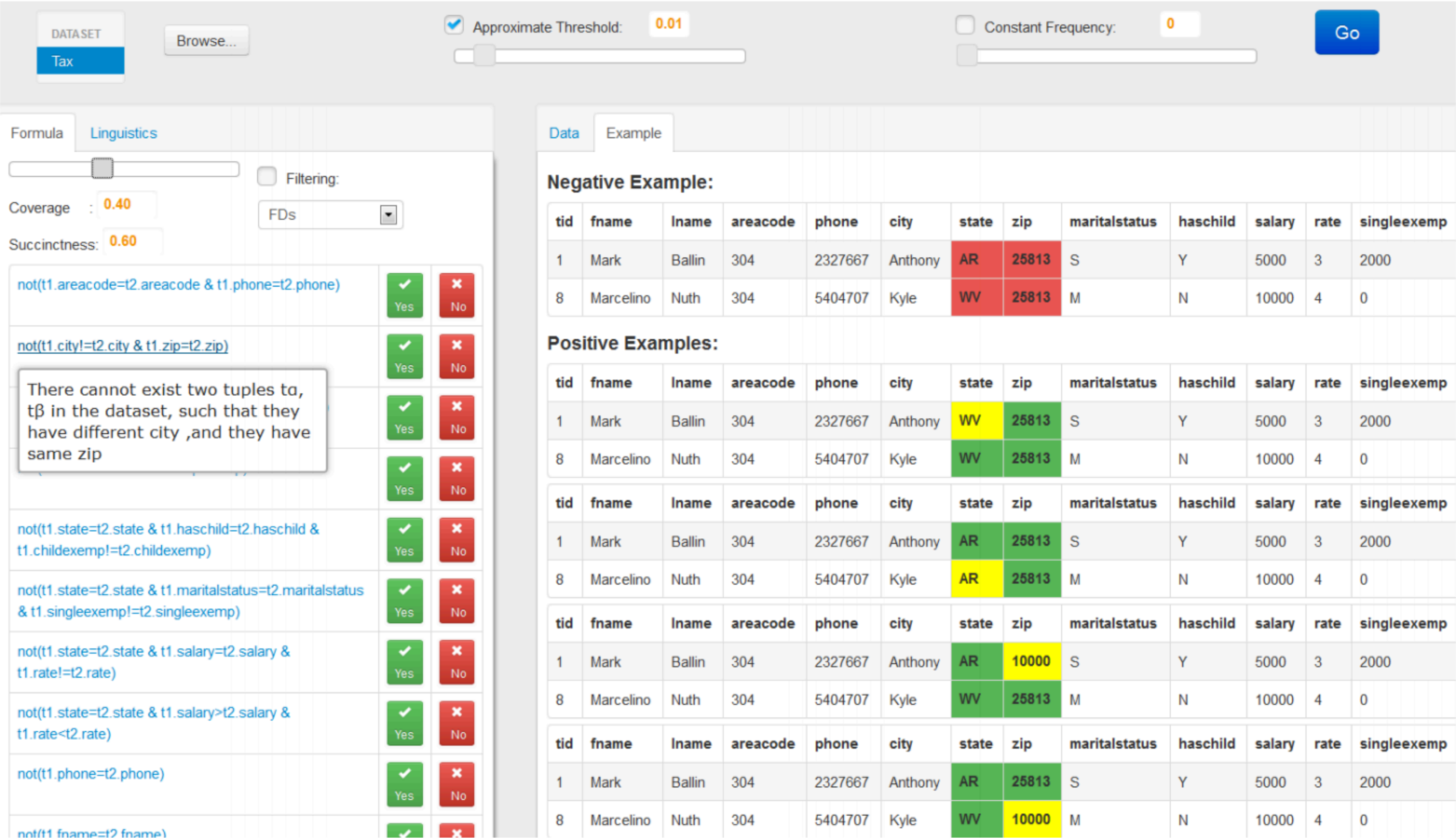
Restricted data types

Restricted number of columns

Tool	Statistics	Patterns	Data types	Uniques	[illegible]	[illegible]
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	✔	✔	✔			

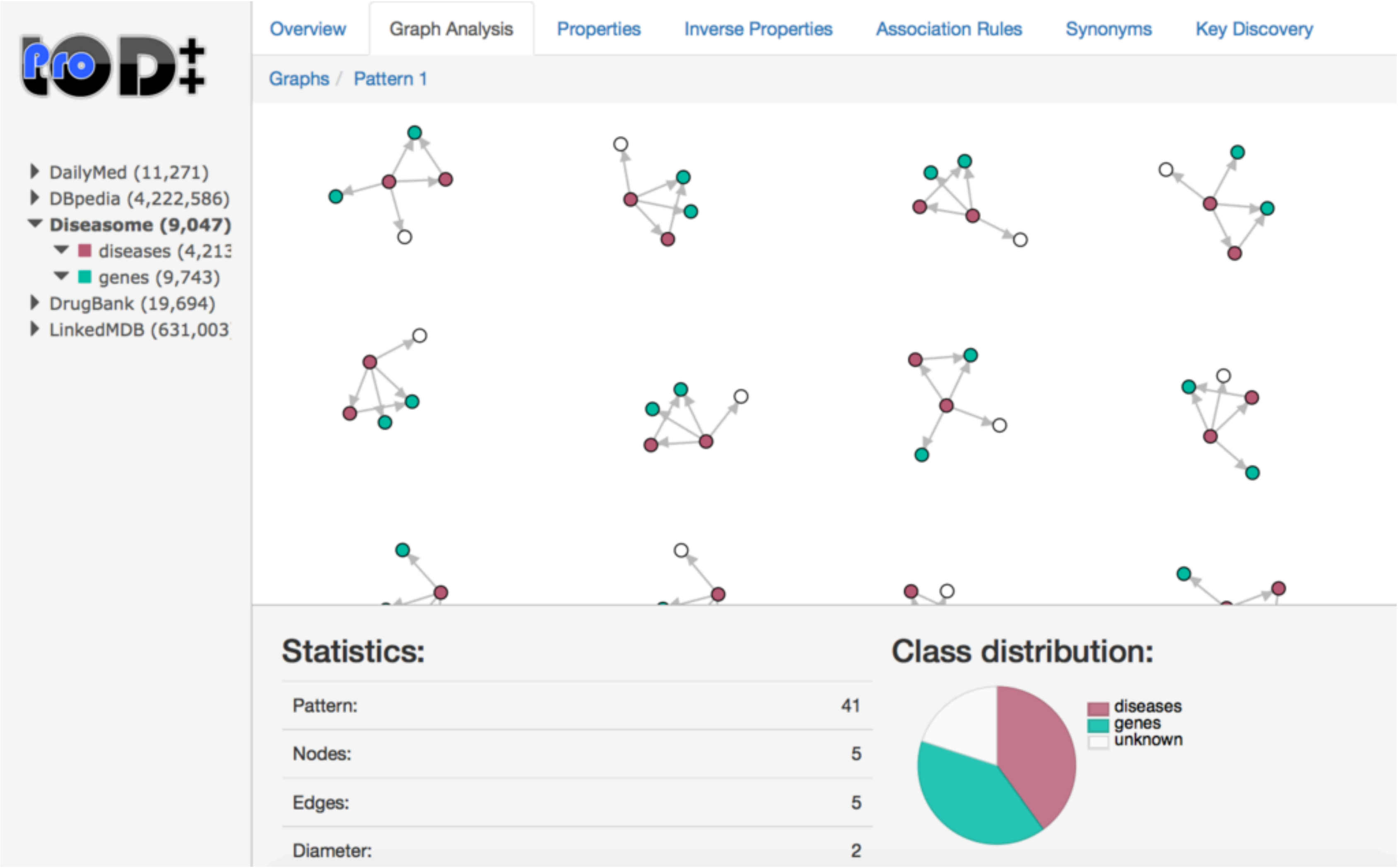

Tools in Research

RuleMiner



DATASET
Tax
Browse...
Approximate Threshold: 0.01
Constant Frequency: 0
Go
Formula
Linguistics
Coverage : 0.40
Succinctness: 0.60
Filtering:
FDs
not(t1.areacode=t2.areacode & t1.phone=t2.phone)
not(t1.city!=t2.city & t1.zip=t2.zip)
There cannot exist two tuples tα, tβ in the dataset, such that they have different city ,and they have same zip
not(t1.state=t2.state & t1.haschild=t2.haschild & t1.childexemp!=t2.childexemp)
not(t1.state=t2.state & t1.maritalstatus=t2.maritalstatus & t1.singleexemp!=t2.singleexemp)
not(t1.state=t2.state & t1.salary=t2.salary & t1.rate!=t2.rate)
not(t1.state=t2.state & t1.salary>t2.salary & t1.rate<t2.rate)
not(t1.phone=t2.phone)
not(t1.fname=t2.fname)
Yes
No
Data
Example
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++



Overview
Graph Analysis
Properties
Inverse Properties
Association Rules
Synonyms
Key Discovery
Graphs / Pattern 1
DailyMed (11,271)
DBpedia (4,222,586)
Diseasome (9,047)
diseases (4,213
genes (9,743)
DrugBank (19,694)
LinkedMDB (631,003
Statistics:
Pattern: 41
Nodes: 5
Edges: 5
Diameter: 2
Class distribution:
diseases
genes
unknown

Tools in Research

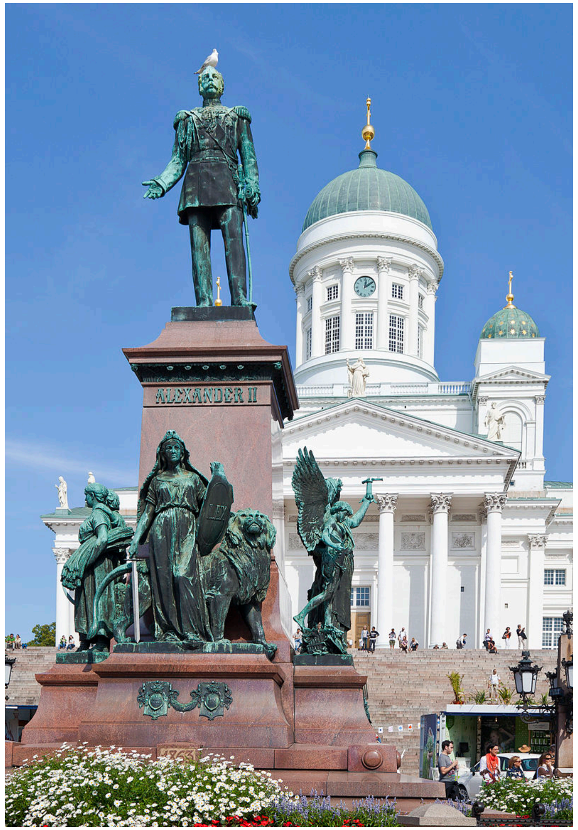
Tool	Main purpose	Statistics	Patterns	Data types	Uniques	Dependencies	Data Mining
Bellmann	Data quality browser	✔			✔		
Potter's Wheel	ETL tool	✔	✔				
Data Auditor	Rule discovery						
RuleMiner	Dependency discovery					✔	
MADLib	Machine learning	✔				✔	
Metanome	Data profiling	✔			✔		
ProLOD++	Profiling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

Shortcomings

- No “real” profiling tool
- Tools focus on “easy” problems:
  - Statistics
  - Single column or “few” column dependencies
  - Many industry tools use SQL instead of optimized algorithms
- No tool covers all types of meta-data
- Management of large meta-data results
  - Summarizing meta-data
  - Ranking meta-data based on relevance

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, INDs, FDs
  - and their discover algorithms
- Outlook
  - Functionality
  - Semantics



Single Column Analysis

Cardinalities and distributions

- Number of values
- Number of distinct values
- Number of NULLs

Count(*)
count(distinct X),
count (X) where X=null

- MIN and MAX value

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

- Histograms
- Probability distribution for numeric values
- Detect whether data follows some well-known distribution

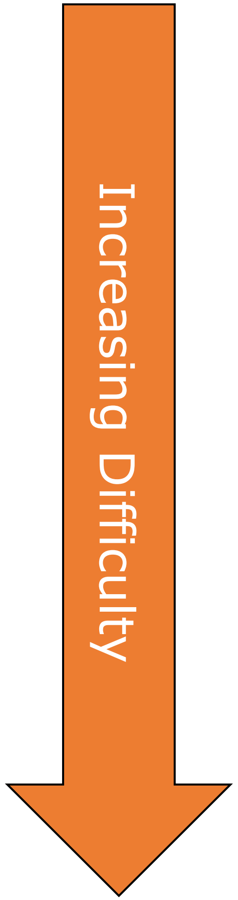
Bottleneck is sorting the data

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
  - Adress, phone, email, first name



Multi Column Analysis

Frequencies, Rules, Correlations

- Frequencies:
  - Which values co-occur with each other?

- Rules:
  - Which values depend on a specific value?

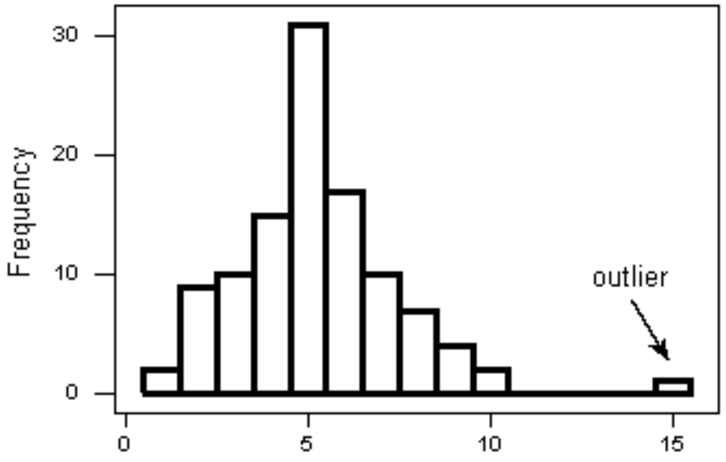
- Correlations:
  - Which values correlate?
  - Which values substitute each other?

Core step: Frequent Itemset Mining

- Origin: Transactional Analysis
  - Which products have been bought together?
- Main step:
  - Find frequencies for all item combinations
- Optimization:
  - Find frequencies for all relevant item combinations, i.e., item combinations with minimum support
- Algorithms:
  - Apriori [Aggrawal, Srikant: fast Algorithms for Mining Association rules, VLDB'94]
  - FP-Growth [Han, Pei, Yin: Mining frequent patterns without candidate generation, SIGMOD'00]
  - ..
  - Survey: [Hipp, Guentzer, Nakhaeizadeh: Algorithms for Association Mining – A General Survey and Comparison, KDD'00]

Outlier detection

- Low-frequent values
- Structural outliers
  - Wrong value representations, e.g.:
    - CA instead of California
- Numerical outliers
  - E.g., according to Gaussian distribution
- Outlier combinations
  - Co-occurrence analysis



- Survey: [Hodge, Austin: A survey of outlier detection methodologies, AI'04]

Sketches and Summaries

- Use cases:
  - Assess column similarity
  - Dimension reduction
  - Data stream samples
- 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)

- N^2 pairwise comparisons
- 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. $h(C_1) \neq h(C_2)$
    - Estimate similarity by applying ***k*** different $h_i(\cdot)$
  - Transform table into a Boolean matrix

Residence (A)	Country (B)	Birthplace (C)
Helsinki	Finland	Oslo
Oslo	Germany	Copenhagen
Berlin	Denmark	Helsinki

Values	A	B	C
Helsinki	1	0	1
Oslo	1	0	1
Berlin	1	0	0
Finland	0	1	0
Germany	0	1	0
Denmark	0	1	0
Copenhagen	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
1	Helsinki	1	0	1
2	Oslo	1	0	1
3	Berlin	1	0	0
4	Finland	0	1	0
5	Germany	0	1	0
6	Denmark	0	1	0
7	Copenhagen	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	4	7
h3	2	5	7

sim(A,B)= 0

sim(A,C)= 0.33

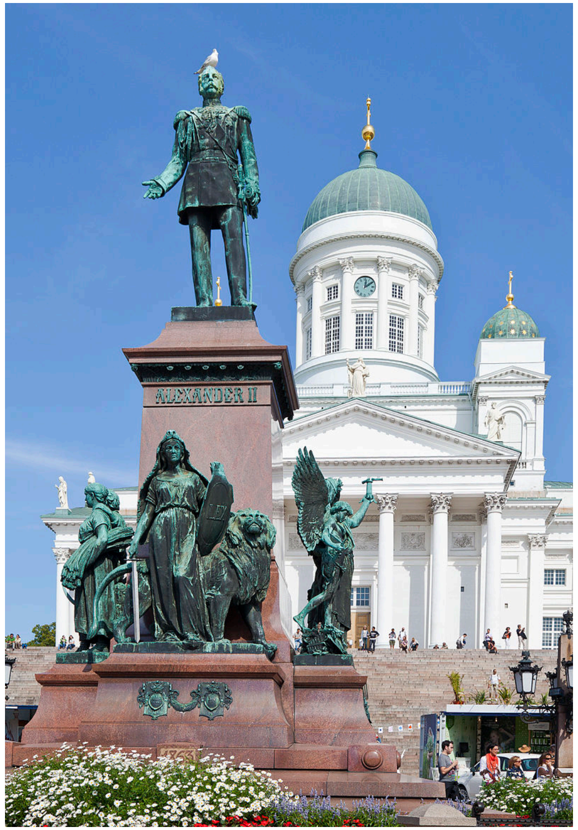
sim(B,C)= 0

Single & Multi-Column Analysis

- Cardinalities
- Data types
- Patterns
- Co-occurrences
- Sketches, summaries
- ….
- Strong overlap with data mining
- Most of them:
  - Not very complex but approximations needed on big data

Tutorial Overview

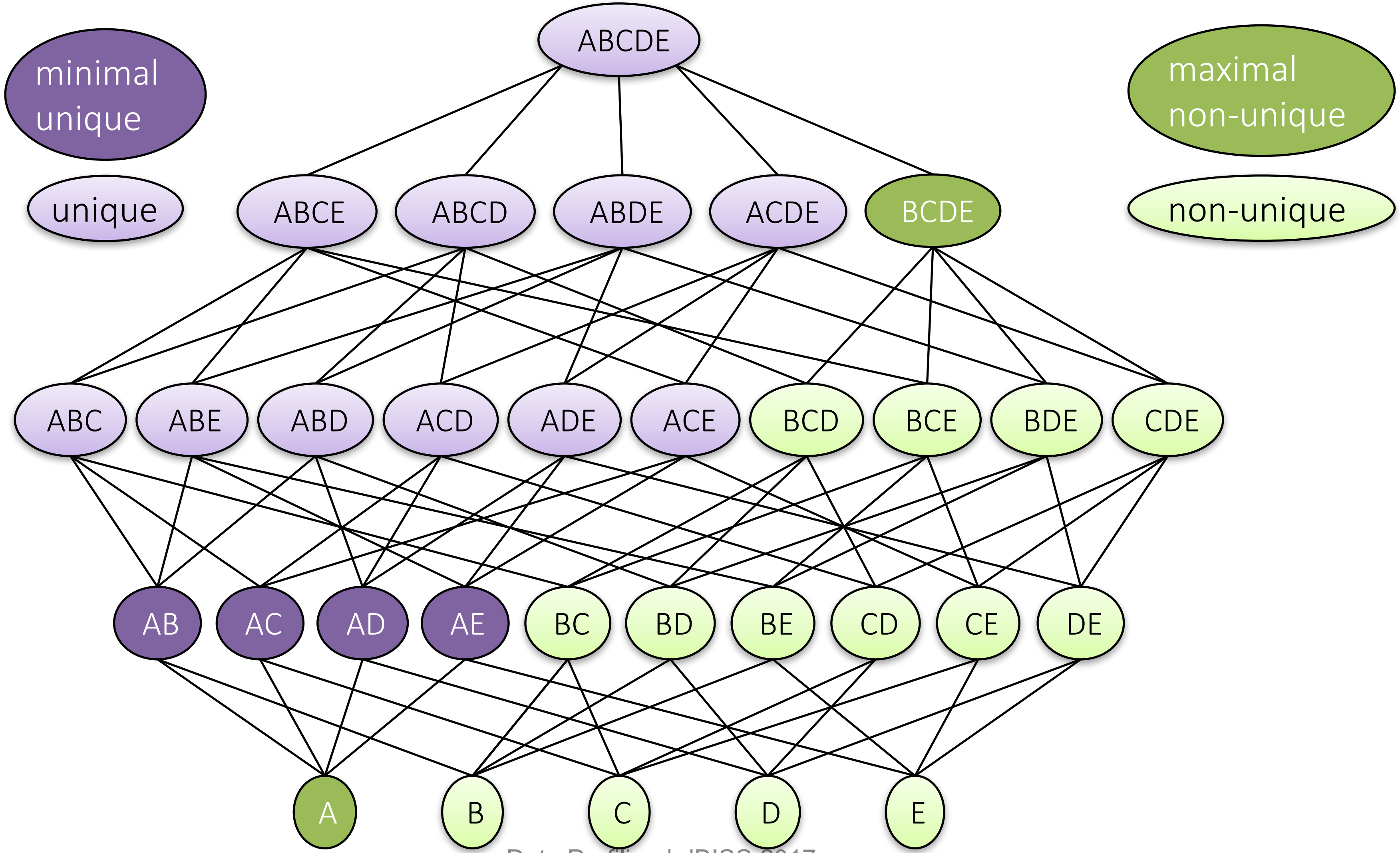
- Motivation
  - Task classification
  - Use cases
- Tools
  - Research and industry
  - Shortcomings
- Single and Multiple Column Analysis
  - Cardinalities and datatypes
  - Co-occurrences and summaries
- **Dependencies**
  - **UCCs, INDs, FDs**
  - **and their discover algorithms**
- Outlook
  - Functionality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



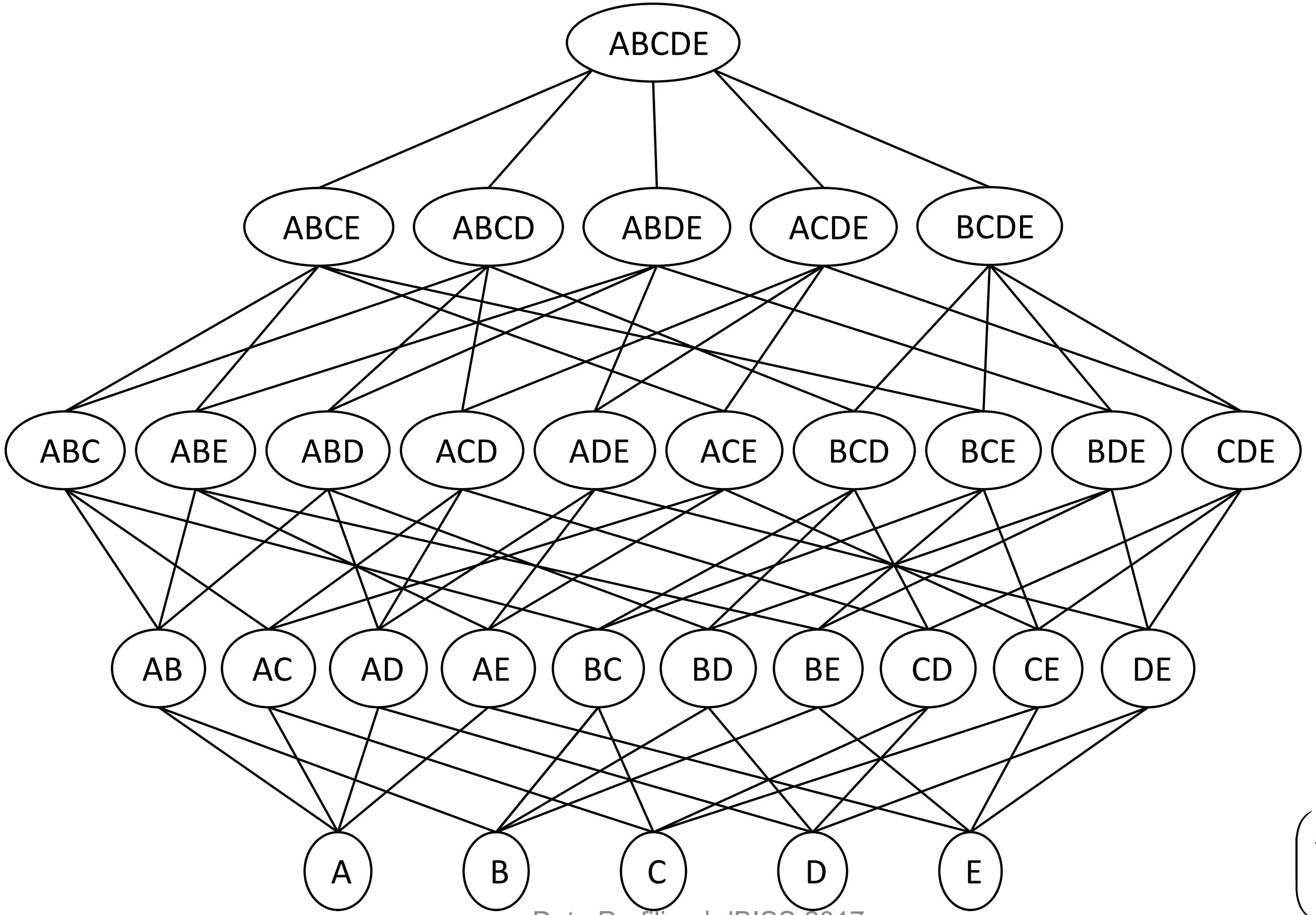
UNIQUE

JUST BECAUSE YOU ARE UNIQUE DOES NOT MEAN YOU ARE USEFUL

Result of algorithm



Challenge: Exponential search space



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

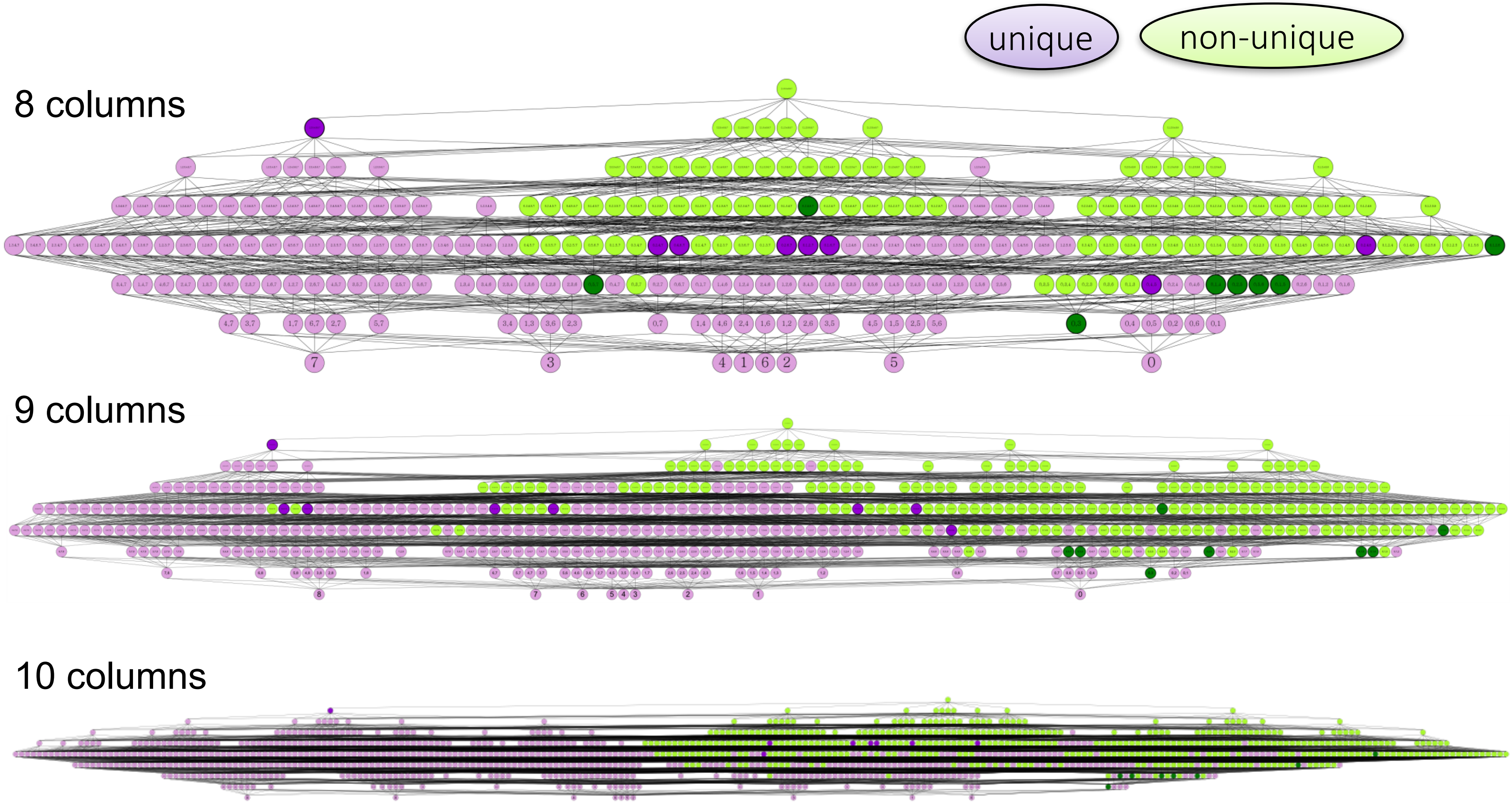
$$\binom{5}{4} = 5$$

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

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

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

TPCH line item



Computational feasibility

- For a lattice over n columns
  - $\binom{n}{k}$ combinations of size k
  - All combinations: 2^n-1 (let's ignore -1 for the remaining slides)
  - Largest solution set: $\binom{n}{n/2}$ minimal uniques are of size $\frac{n}{2}$

$$\binom{n}{k} \in \Theta(n^k) \Rightarrow \binom{n}{n/2} \in \Theta(n^n)$$

    - Verifying minimality, requires to check also all combinations of size $\frac{n}{2} - 1$

  - Adding a column doubles search space (and vice versa)

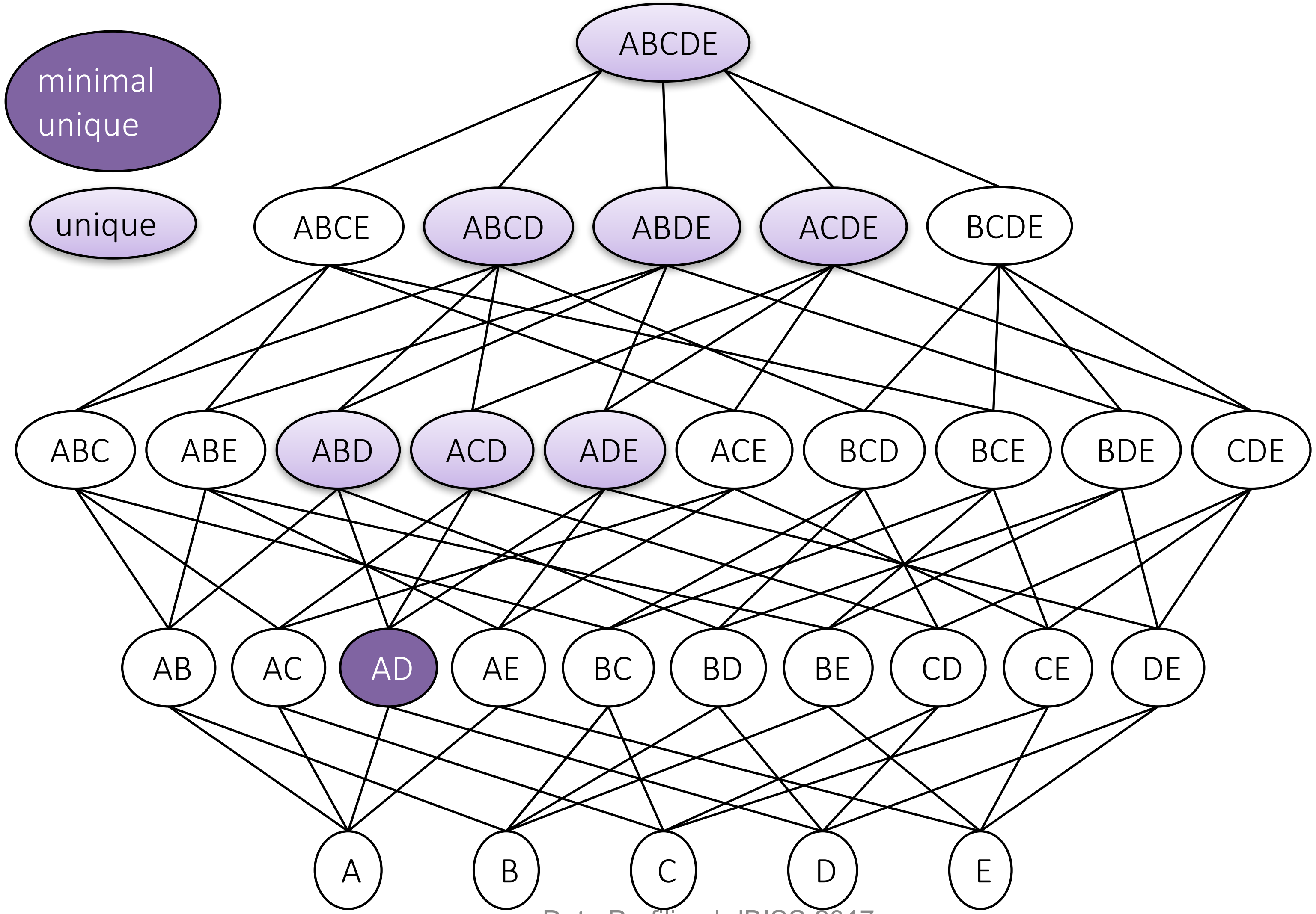
Pruning with uniques #2

- Pruning: inferring the type of a combination without actual verification

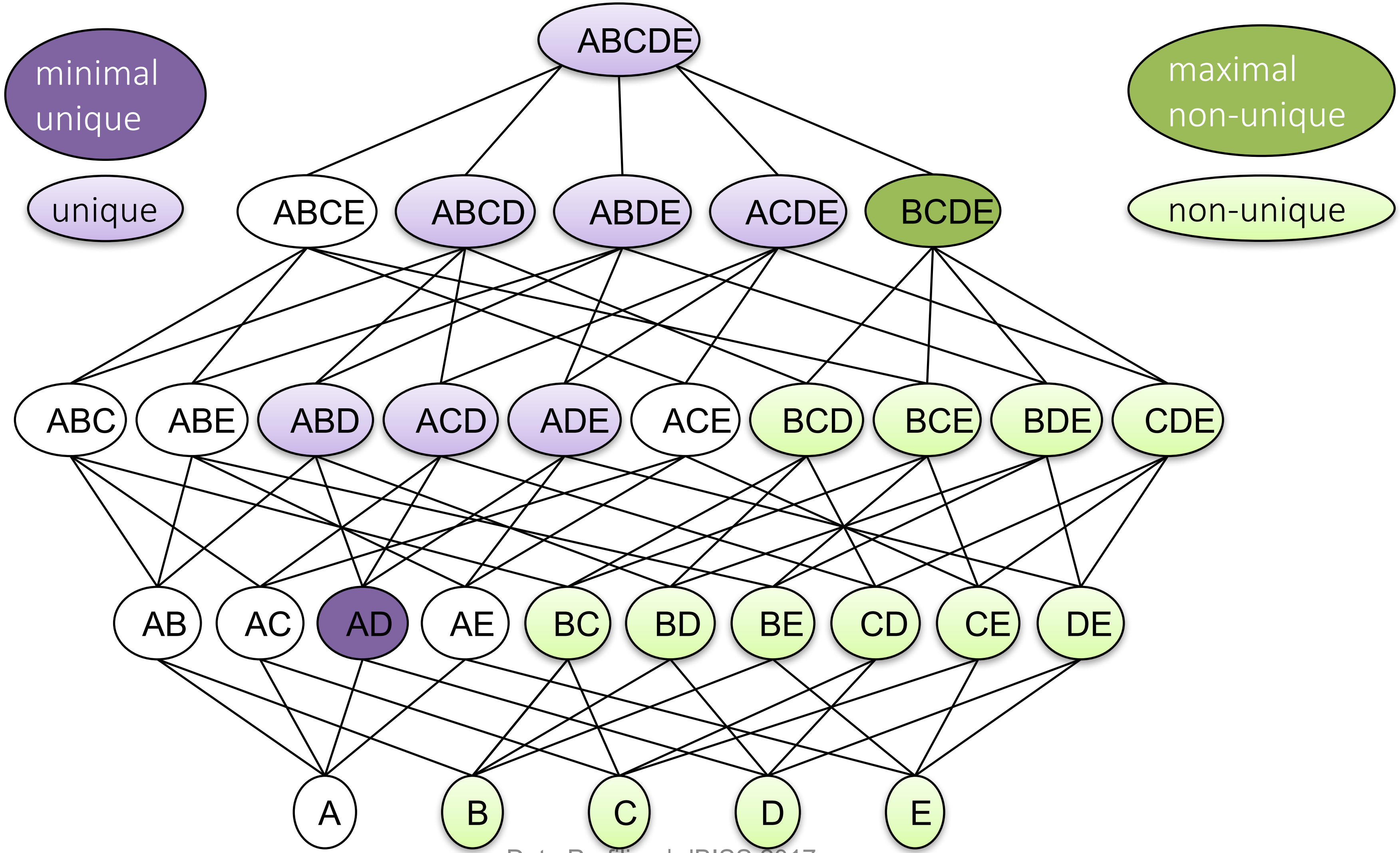
- If A is unique, supersets must be unique
- Finding a unique column prunes half of the lattice
  - Remove column from initial data set and restart
- Finding a unique column pair removes a quarter of the lattice
  - In general, the lattice over the combination is removed

- The pruning power of a combination is reduced by prior findings
  - AB prunes a quarter
  - BC additionally prunes only one eighth
  - ABC was already pruned by AB and constitutes already one eighth

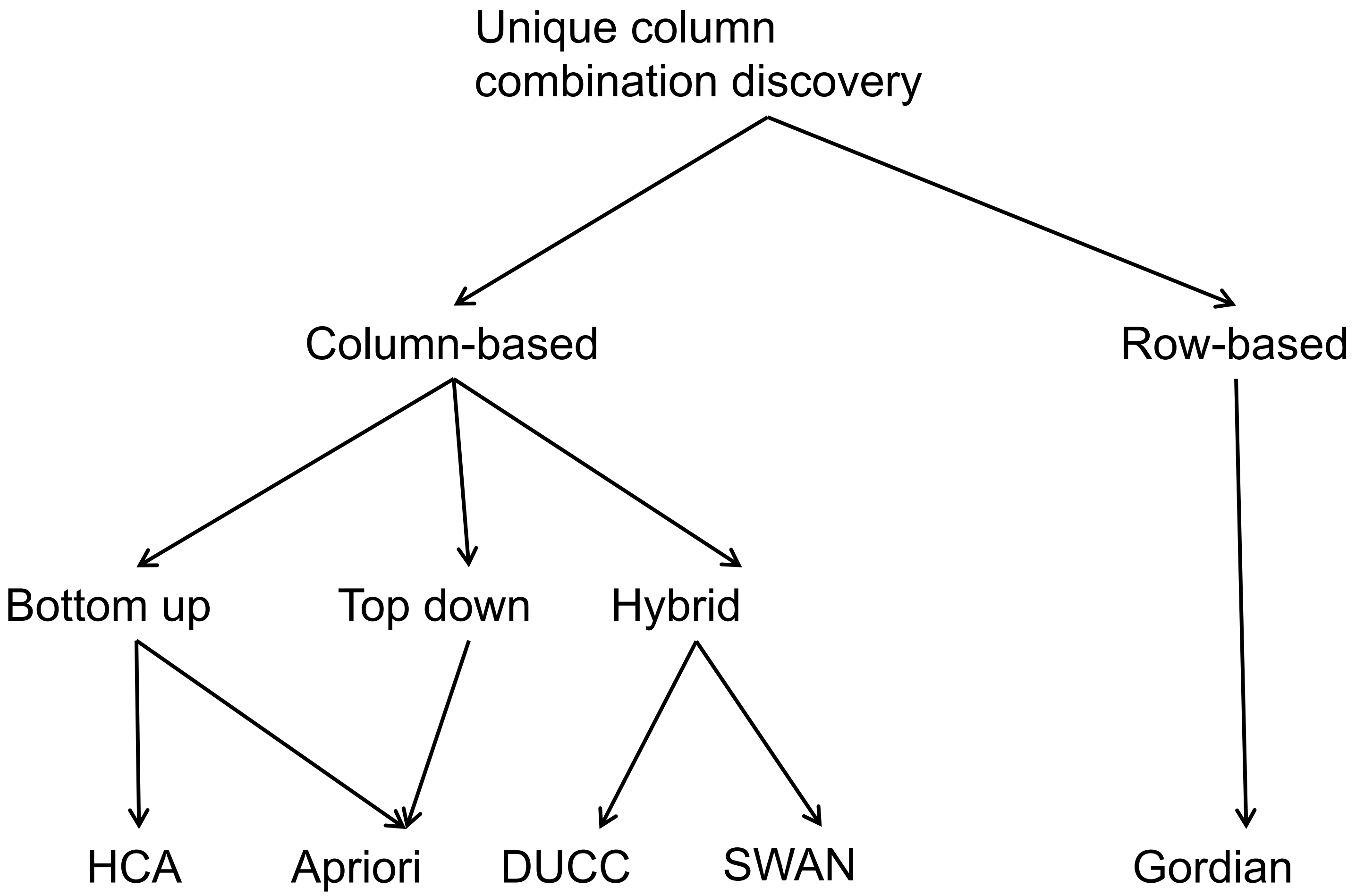
Pruning effect of a pair



Pruning both ways



Discovery Algorithms



Column-based algorithms

- Traverse through lattice
- Check for uniqueness
  - Different approaches possible
  - Use database backend and distinctness query
    - SELECT COUNT(DISTINCT A, B, C) FROM R
    - Compare with row-count
  - Position list indexes (explained later)
  - For now, check is blackbox

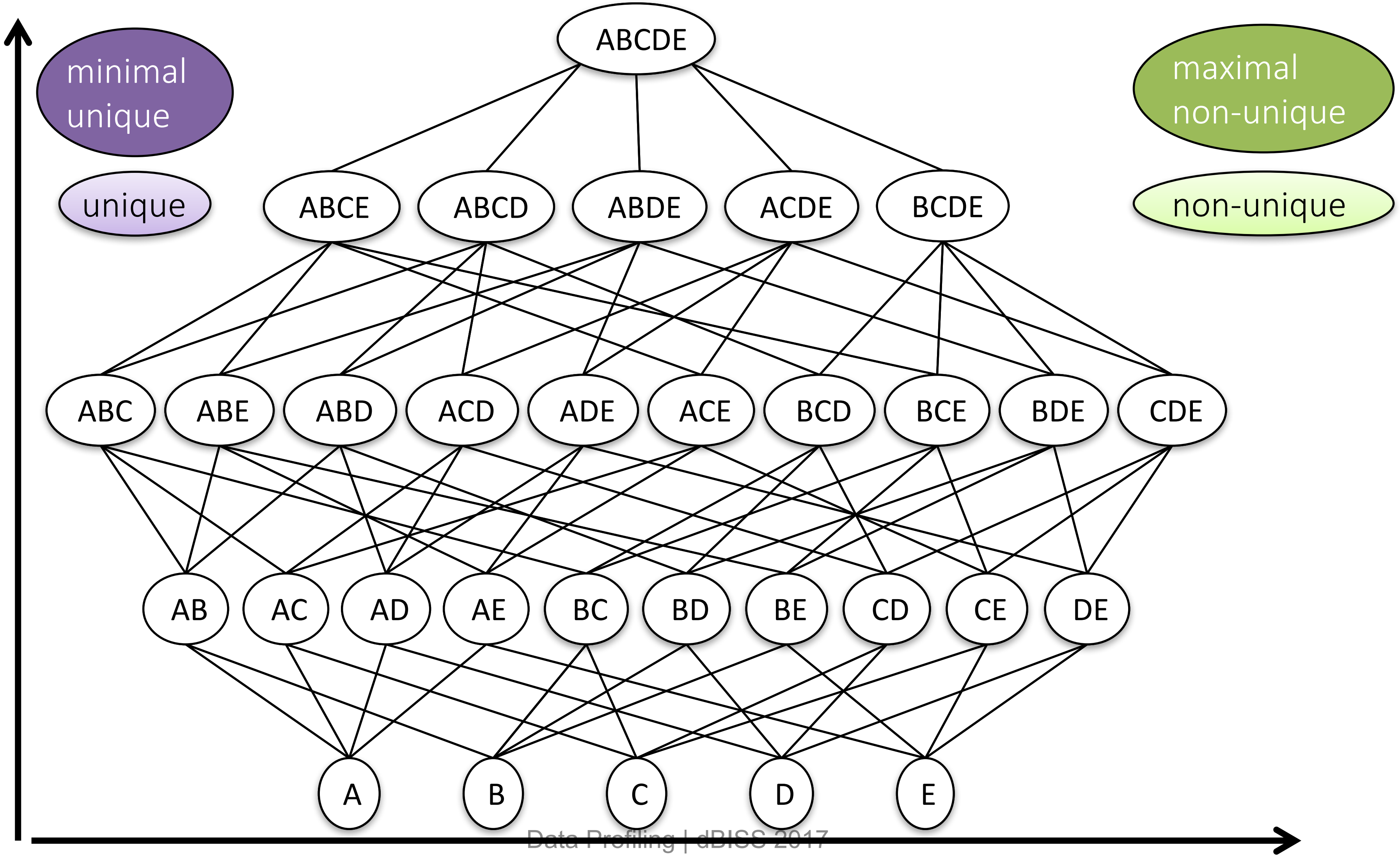
- Prune lattice accordingly

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 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 and go down
  - Performs better if solution set is high up
  - Candidate pruning becomes more tricky
- Hybrid [Giannella, Wyss: Finding minimal keys in a relation instance. (1999)]
  - Combine bottom-up and top-down
  - Interleave checks
  - 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]
  - More sophisticated candidate generation
  - Uses histograms for pruning
  - Finds and uses functional dependencies on-the-fly

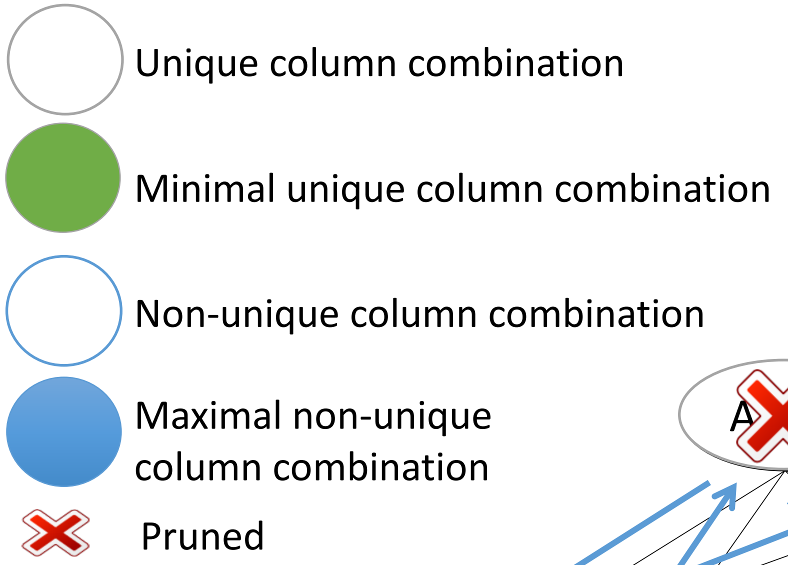
DUCC

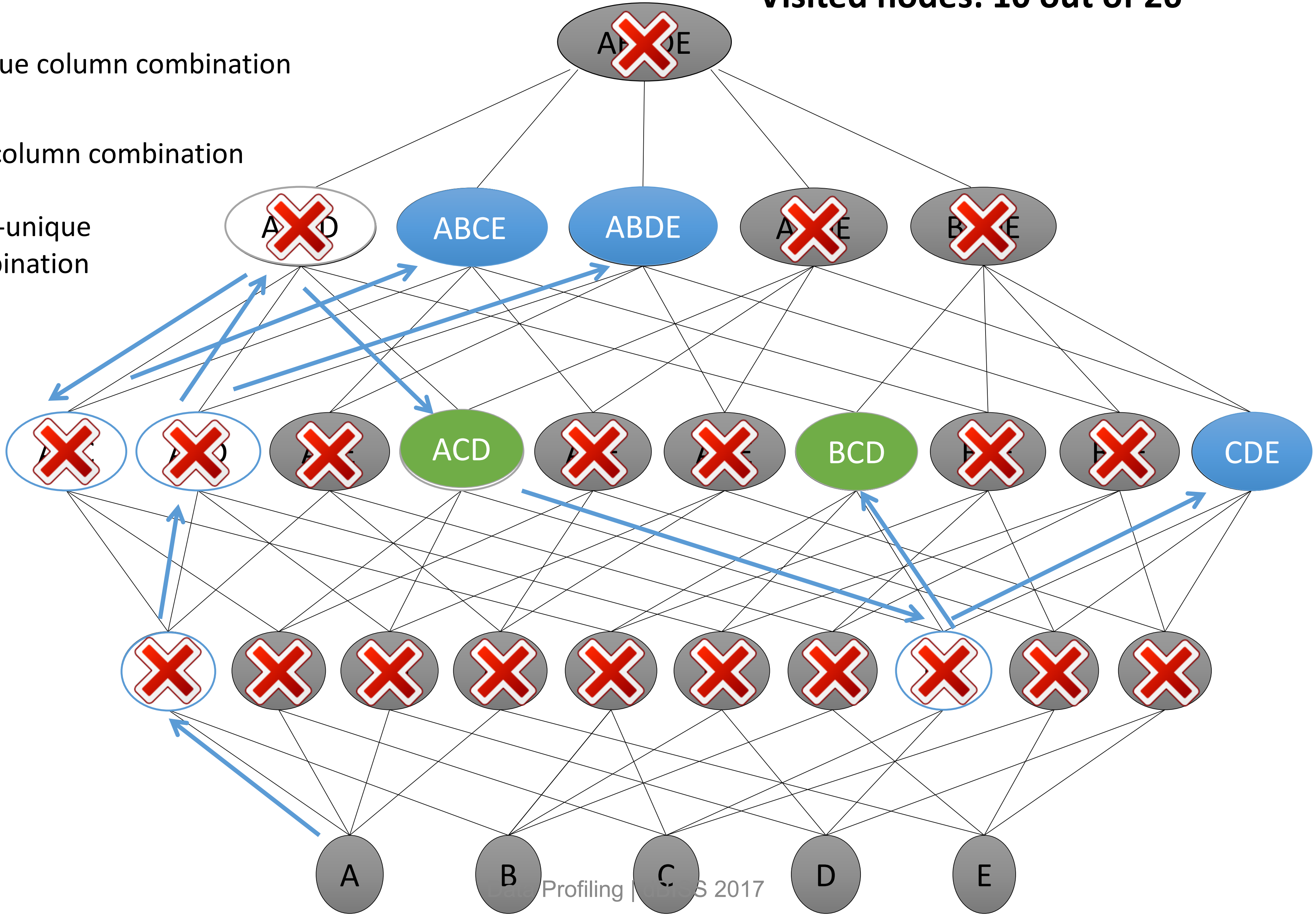
[Heise, Quiané-Ruiz, Abedjan, Jentzsch, Naumann: Scalable Discovery of Unique Column Combinations, PVLDB'14]

- Scalability is major design goal of DUCC
  - Random walk well suited for parallelization
    - Few coordination overhead
  - Threads/worker share findings through event bus
    - Uniques/non-uniques
    - Holes in graph
  - Lock-free to avoid bottlenecks
    - Only memory barrier in local event bus
- Basic idea: random walk through lattice
  - Pick random superset if current combination is non-unique
  - Pick random subset otherwise

ACD and BCD are minimal uniques

Visited nodes: 10 out of 26





Position List Index

- Aka “partitions”
- Incorporates row-based pruning
- Intuition: number of duplicates decrease when going up the lattice
  - Many unnecessary rows are checked again and again

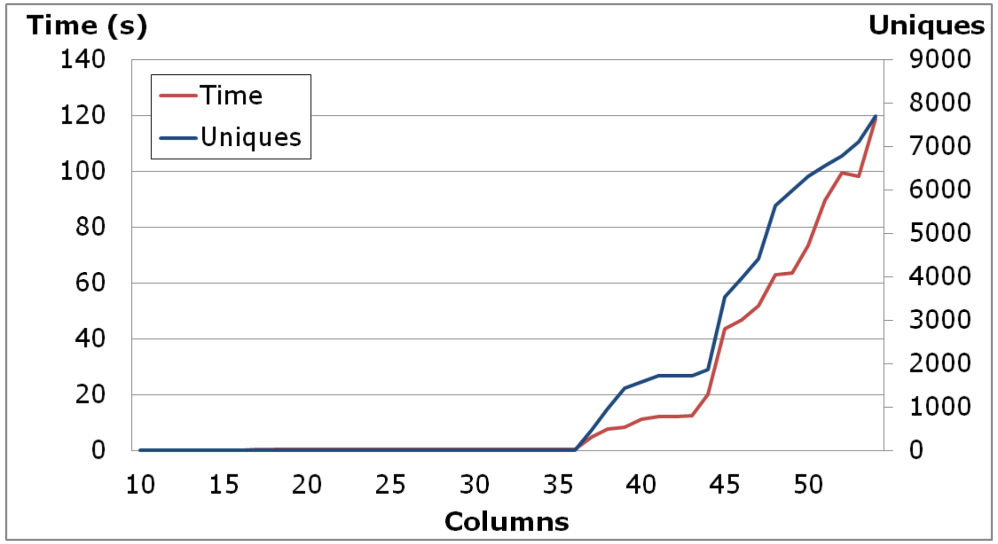
- Keep track of duplicates with inverted index
  - X: a->$\{r_1, r_2, r_3\}$, b->$\{r_4, r_5\}$
  - Y: 1->$\{r_1, r_3\}$, 2->$\{r_2, r_5\}$, 3->$\{r_4\}$

- Stripped partitions:
  - Remove clusters of size 1:
  - X: $\{\{r_1, r_2, r_3\}, \{r_4, r_5\}\}$
  - Y: $\{\{r_1, r_3\}, \{r_2, r_5\}\}$

X	Y
a	1
a	2
a	1
b	3
b	2

Analysis of DUCC

- Runtime mainly depends on size of solution set



- Worst case: Solution set is in the middle: $\binom{n}{n/2}$
- Aggressive pruning may lead to loss of minimal uniques!
  - Gordian's final step can be used to plug these holes

Gordian

[Sismanis, Brown, Haas, Reinwald: GORDIAN: efficient and scalable discovery of composite keys, *VLDB'06*]

- Row-based algorithm
- Builds prefix tree while reading data
  - Discover maximal non-uniques on prefix tree
- Compute minimal uniques from maximal non-uniques
  - Complementation

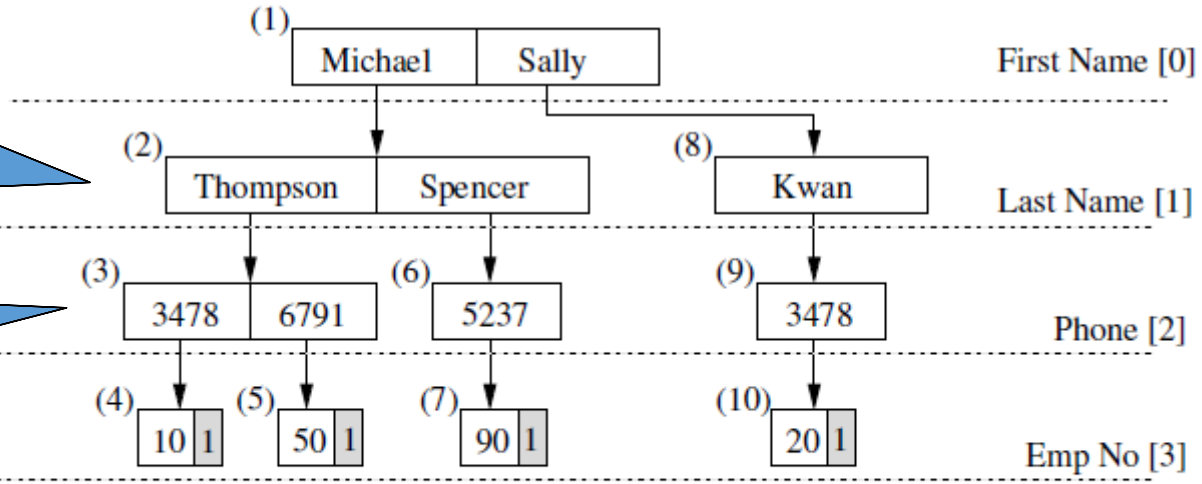
Prefix tree

FirstName	*LastName*	*Phone*	*EmpNo*	*COUNT*
Michael	Thompson	3478	10	1
Sally	Kwan	3478	20	1
Michael	Spencer	5237	90	1
Michael	Thompson	6791	50	1

One tree per attribute order

3 nodes, thus max. non-unique

4 nodes, thus unique



Analysis Gordian

- According to paper, polynomial in the number of tuples for data with a Zipfian distribution of values
  - Can abort scan as soon as duplicate has been found

- Worst case
  - Exponential in the number of columns
  - All data needs to be stored in memory

- Computing minimal uniques from maximal non-uniques
  - O(non-uniques^3 x columns)
  - Can be sped up with presorted list

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 get unique
  - Minimal uniques might lose minimality
- SWAN
  - Leverage the knowledge of 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.

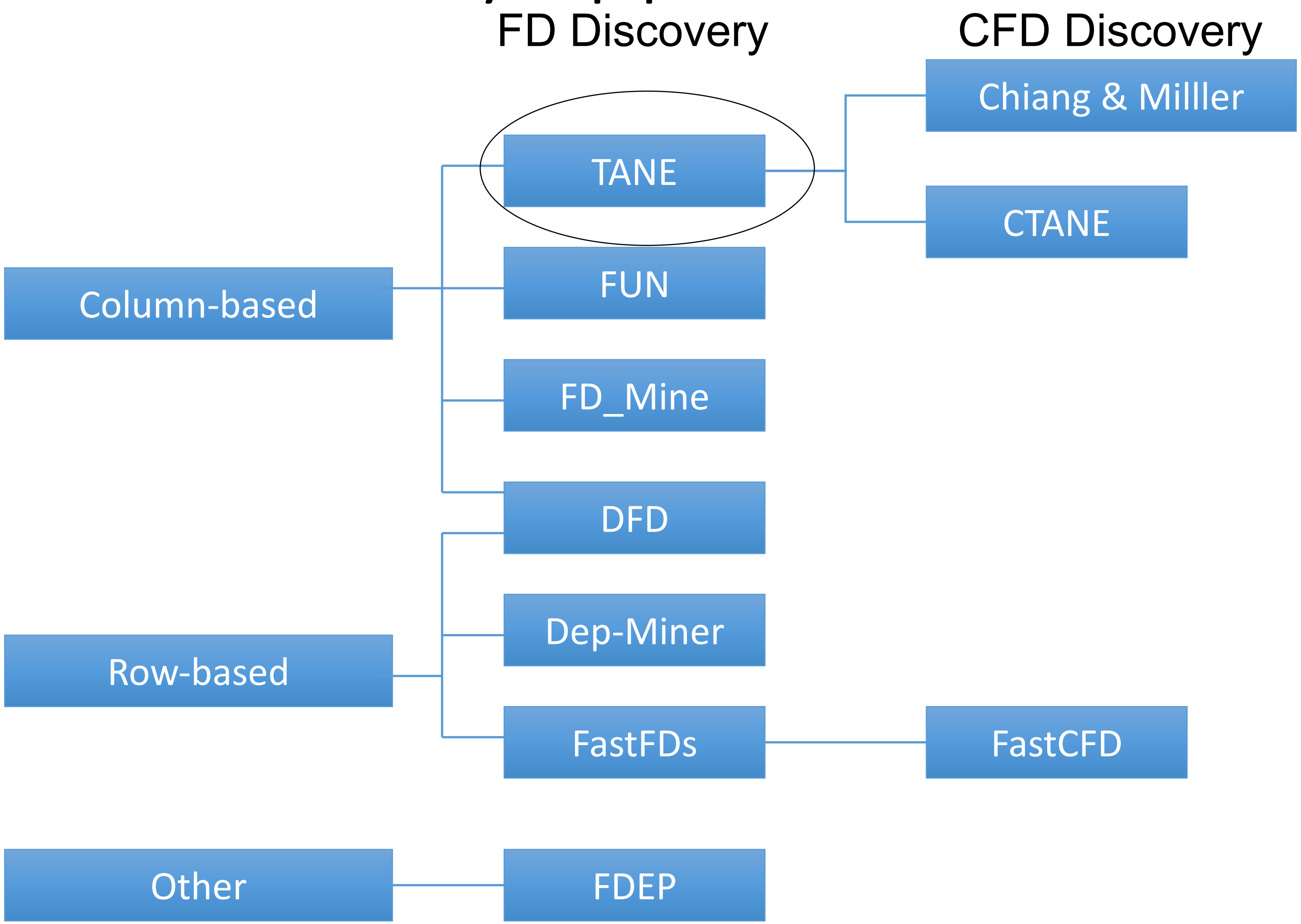
Naive Discovery Approach

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

- For each A ∈ R
  - For each column combination X\A
    - For each pair of tuples (t_1,t_2)
      - If t_1[X] = t_2[X] and t_1[A] ≠ t_2[A]: Break
    - Return X → A

- Complexity
  - For each of the |R| possibilities for RHS
    - check $2^{(|R|-1)}$ combinations for LHS
    - And scan each record pair ($n^2/2$) for each check

Current FD Discovery approaches



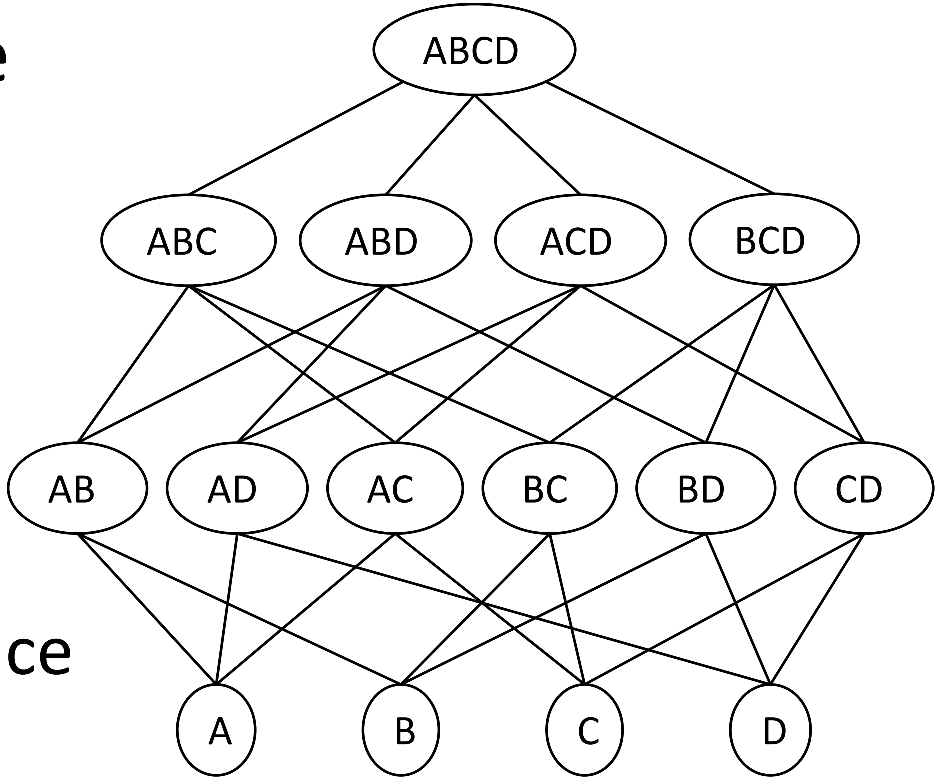
Tane – General Idea

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

- Two key ideas
  1. Reduce column combinations through pruning
     - Reasoning over FDs
  2. Reduce tuple sets through partitioning
     - Partition data according to attribute values
     - Level-wise increase of size of attribute set
       - Consider sets of tuples whose values agree on that set

TANE: Discovery strategy

- Bottom up traversal through lattice
  - $\Rightarrow$ only minimal dependencies
  - Pruning
  - Re-use results from previous level
- For a set X, test all $X \setminus A \rightarrow A$, $A \in X$
  - $\Rightarrow$ only non-trivial dependencies
  - Interpretation: Test each edge in lattice
  - Test on efficient data structure

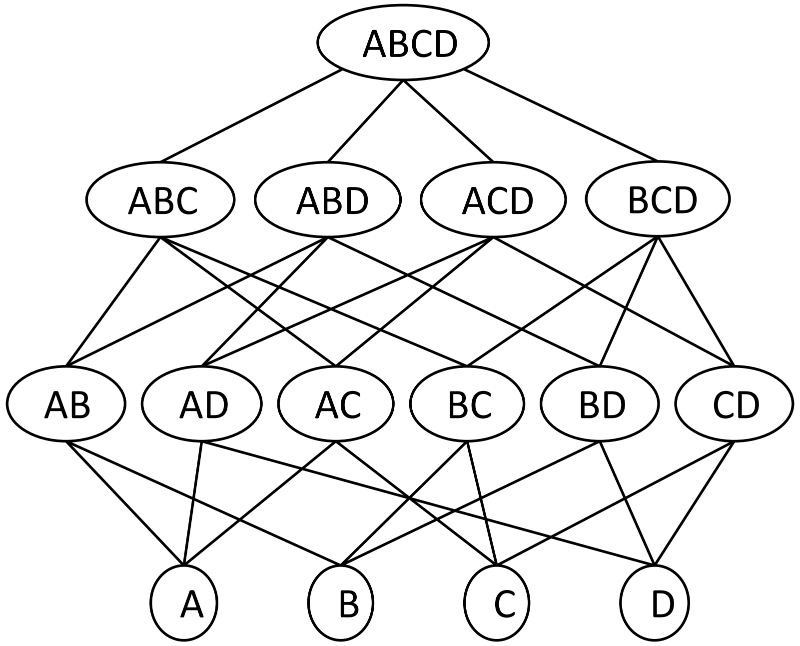


Candidate Sets

- 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} = {ABD}
  - 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) = {D}
  - C(BCD) = {A}

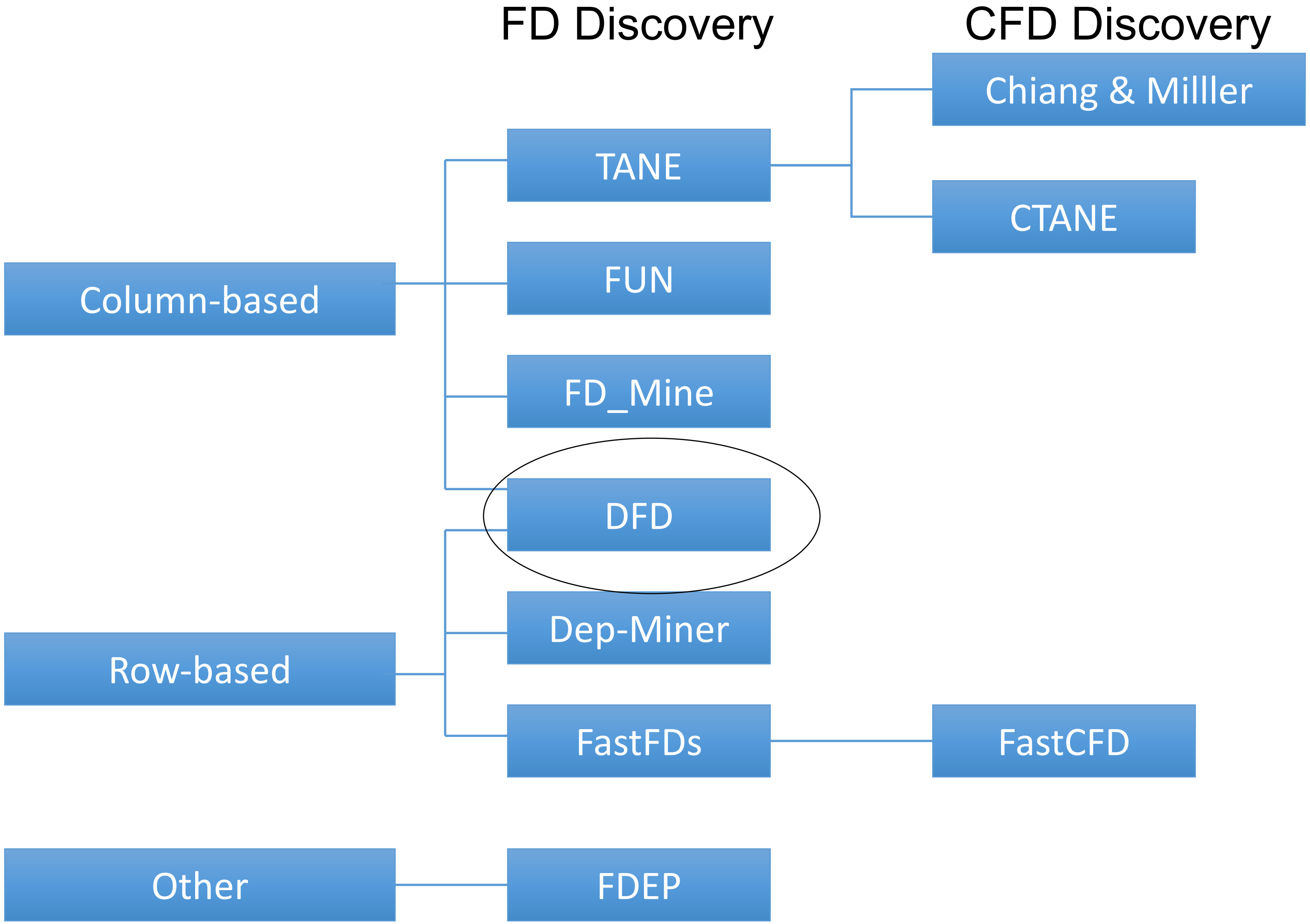


Partial FDs with TANE

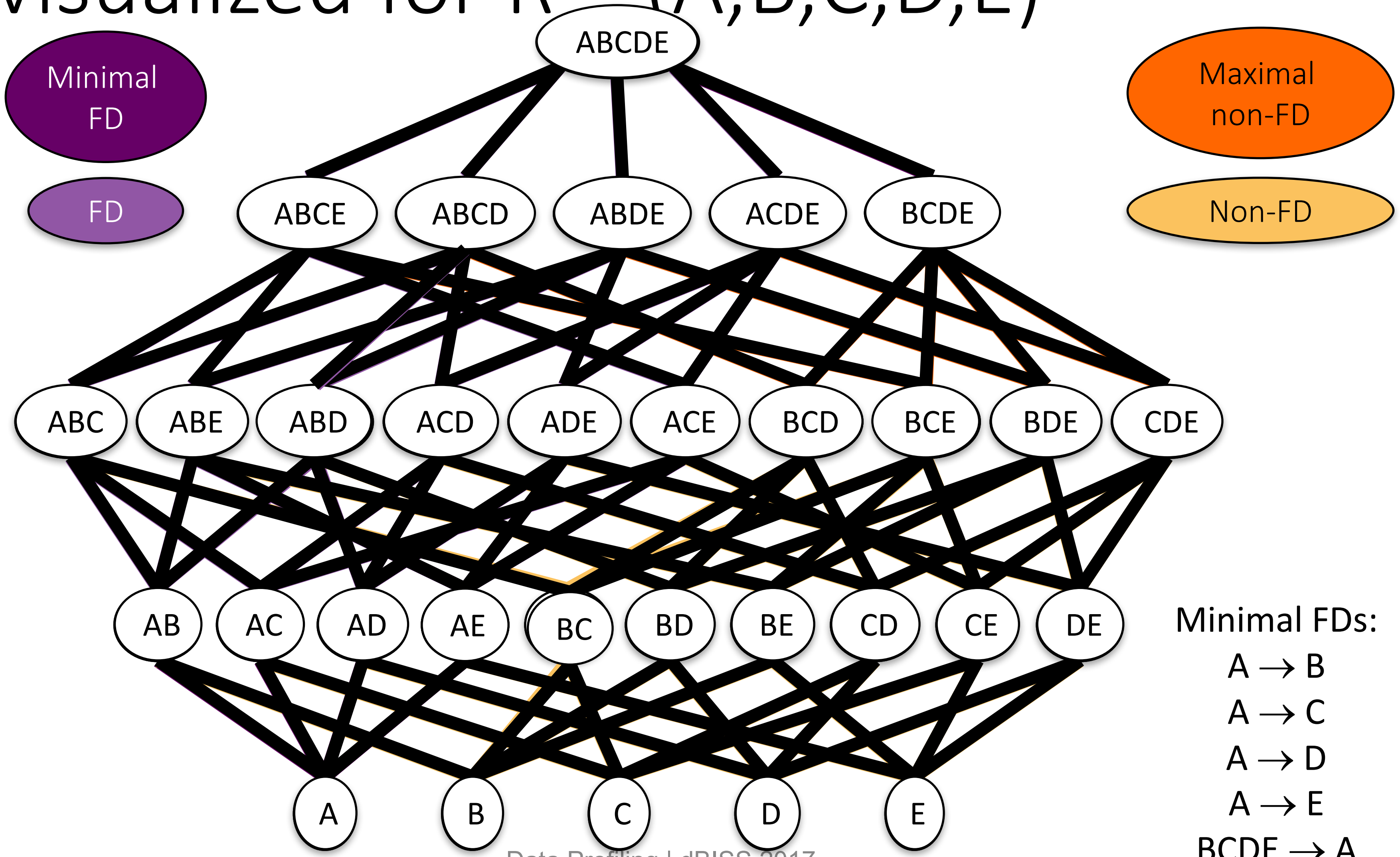
- Definition based on minimum number of tuples to be removed from R for X→A to hold in R.
- Discovery problem:
  - Given relation R and threshold ε, find all minimal non-trivial FDs X→A such that $e(X \rightarrow A) \leq \varepsilon$.
  - Called “approximate” FDs in paper

1. Define error: Fraction of tuples causing FD violation
   - Error $e(X \rightarrow A) = \min\{|S| \mid S \subseteq R,\ R \setminus S \models X \rightarrow A\} / |R|$
2. Specify error threshold $\boldsymbol{\varepsilon}$
3. Modify dependency checking algorithm
   - Efficient algorithm to compute error
   - Bounds to avoid error calculation

Current FD Discovery approaches



DFD Explanation: Tane visualized for R = (A,B,C,D,E)



Minimal FDs:

$A \rightarrow B$

$A \rightarrow C$

$A \rightarrow D$

$A \rightarrow E$

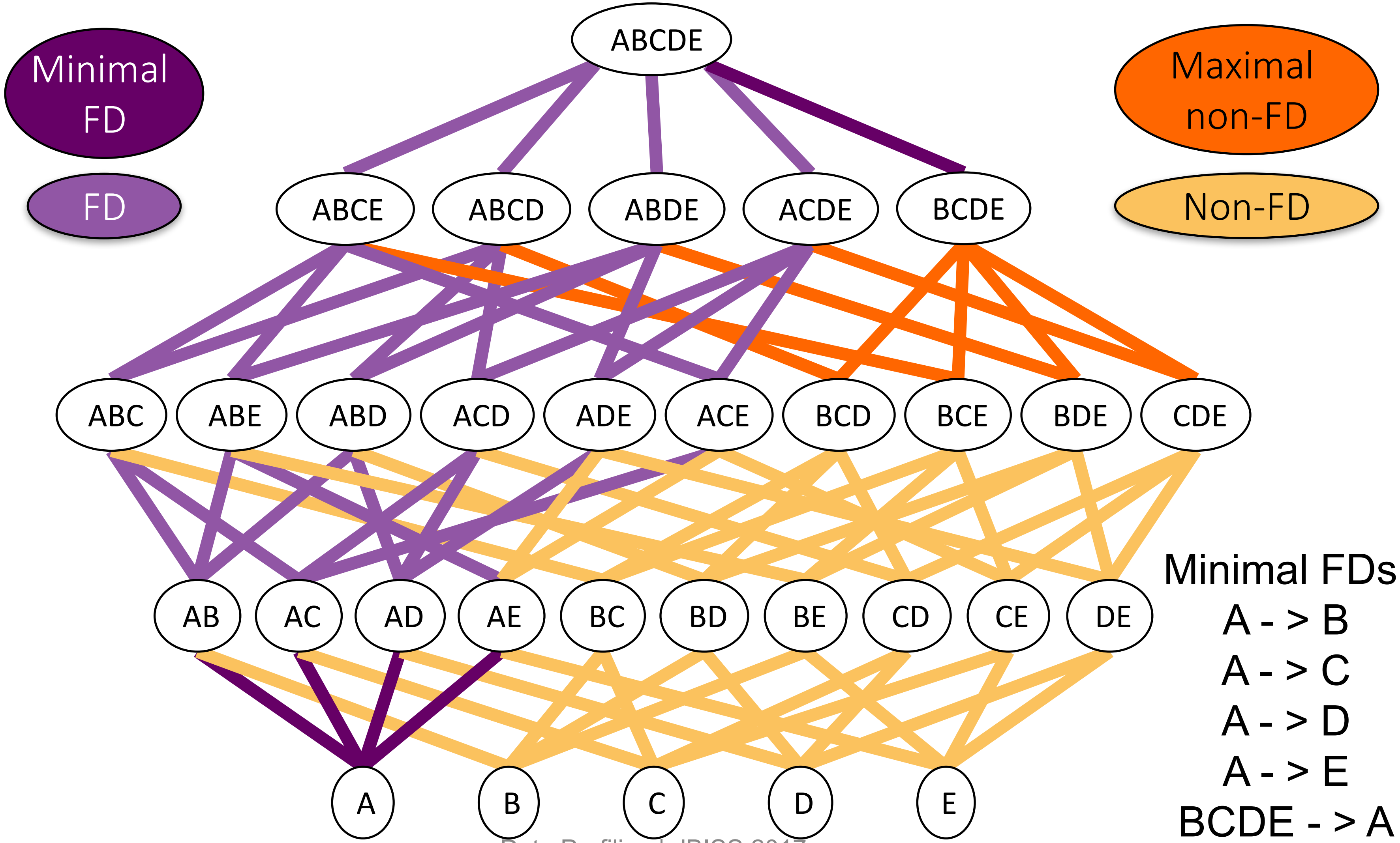
$BCDE \rightarrow A$

DFD: Depth-first approach for functional dependency discovery

[Abedjan,Schulze,Naumann: DFD:Efficient Functional Dependency Discovery, CIKM'14]

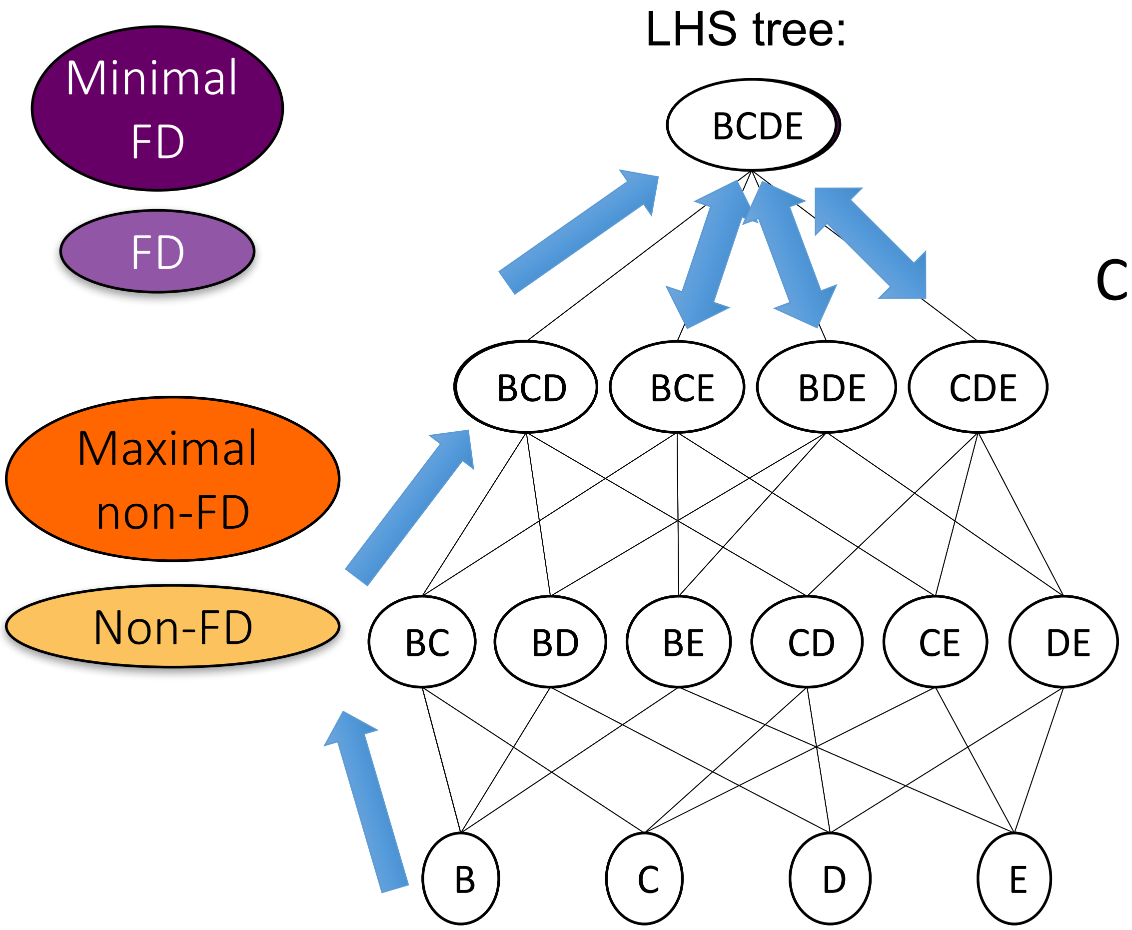
- Traverse depth-first and prune upwards and downwards
- Applied for key/unique discovery: DUCC
  - Key discovery is a subproblem of FD discovery
  - Adapt the concept of minimality in keys to LHS of FDs:
    - An FD $X \rightarrow C$ is minimal if $\forall X' \subset X : X' \xrightarrow{NOT} C$
    - A non-dependency $X \xrightarrow{NOT} C$ is maximal if $\forall X' \supset X : X' \rightarrow C$

Decompose Relation for each RHS



Minimal FDs:
A - > B
A - > C
A - > D
A - > E
BCDE - > A

Decomposition for RHS=A



Checked combinations

Tane: 15

DFD: 7

Minimal FDs:

BCDE $\rightarrow$ A

Traversal Holes

- Aggressive traversal and pruning
  - As for DUCC: Some nodes might never be reached.
- GORDIAN [VLDB’06]:
  - Complement the set of **maximal non-keys**
- **=** set of **minimal keys**
- Key observation from DUCC: the **difference** of one set and the complement of its counterpart delivers the **unvisited nodes**!
- Hole discovery works for FDs too:
  - Consider **minimal FD LHS** and **maximal non-FD LHS**

Execution time - uniprot

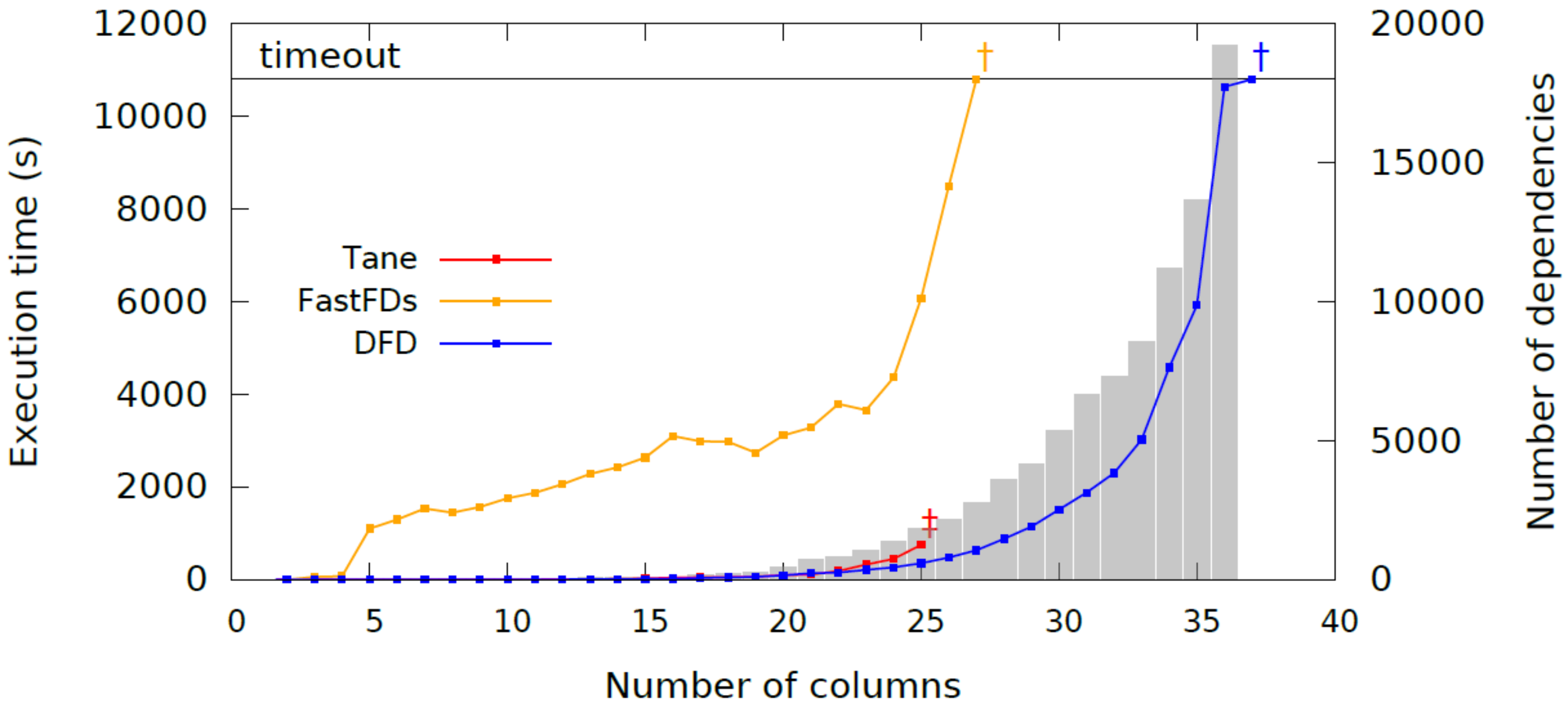


Figure 4.2: Execution time for Tane, FastFDs, and DFD on the first 100,000 rows of the uniprot dataset. († - Time Limit ‡ - Memory Limit)

Functional Dependency Evaluation

dataSet	Columns	Rows	FDs	Tane	FUN	FD_Mine	Dep-Miner	FastFDs	FDep	DFD
iris	5	150	4	0.6s	**0.1s**	**0.1s**	**0.1s**	**0.1s**	**0.1s**	**0.1s**
balance-scale	5	625	1	0.9s	0.4s	0.3s	**0.2s**	0.5s	0.3s	**0.2s**
chess	7	28,056	1	2.0s	1.0s	3.0s	200.8s	200.1s	202.5s	**0.9s**
abalone	9	4,177	137	1.0s	**0.3s**	1.0s	2.9s	3.0s	4.1s	0.9s
nursery	9	12,960	1	3.1s	1.5s	6.0s	132.0s	131.9s	56.6s	**1.1s**
breast-cancer	11	699	46	1.4s	**0.4s**	1.5s	0.9s	1.0s	**0.4s**	0.9s
bridges	13	108	142	1.3s	0.5s	2.9s	**0.2s**	**0.2s**	**0.2s**	0.9s
echocardiogram	13	132	538	0.8s	**0.1s**	69.9s	**0.1s**	**0.1s**	**0.1s**	1.6s
adult	14	48,842	78	81.2s	150.2s	485.3s	5982s	5946s	760.7s	**6.8s**
letter	17	20,000	61	326s	553.9s	ML	865.4s	853.9s	292.3s	**9.1s**
hepatitis	20	155	8,250	10.9s	321.6s	TL	5363.1s	9.3s	**0.5s**	317.8s
horse	27	368	128,726	5451.s	TL	TL	TL	386.8s	**15.7s**	TL
fd-reduced-30	30	250,000	89,571	**41.1s**	78.4s	TL	391.9s	391.3s	TL	TL
flight	109	1,000	982,631	ML	TL	ML	TL	TL	**213.5s**	TL
plista	125	1,000	178,152	ML	TL	TL	TL	TL	**26.4s**	TL

IND Discovery

1. **DeMarchi‘s Algorithm**
2. **Spider**
3. **BINDER & MIND**
   - High performance IND detection
   - Work by Thorsten Papenbrock

BINDER – divide & conquer based IND detection

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	ye	
Neptune	Giant	3.883	17.2	30.06	164.8	0.67	14	ye	

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

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

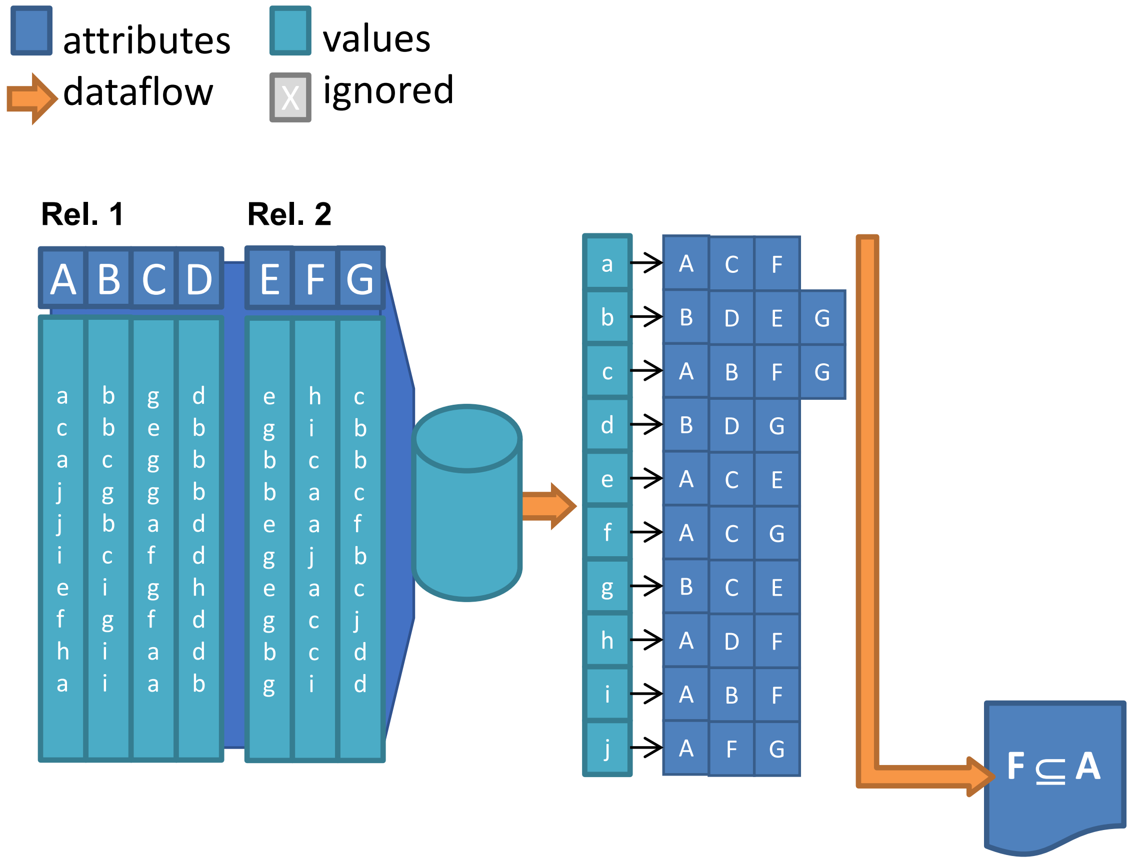
- Name ⊆ Type ?
- Name ⊆ Equatorial_diameter ?
- Name ⊆ Mass ?
- Name ⊆ Orbital_radius ?
- Name ⊆ Orbital_period ?
- Name ⊆ Rotation_period ?
- Name ⊆ Confirmed_moons ?
- Name ⊆ Rings ?
- Name ⊆ Atmosphere ?

- 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]



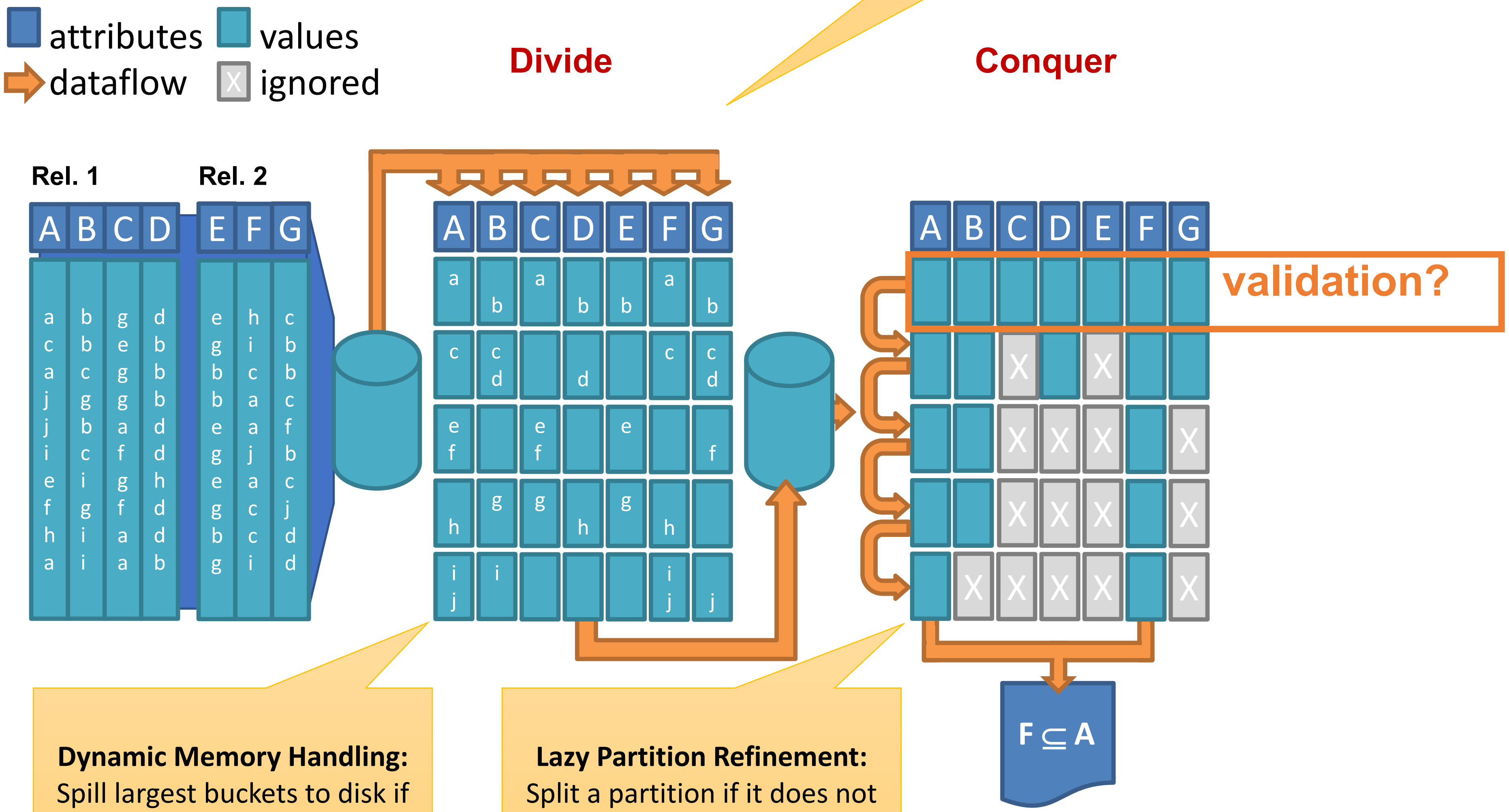
All intersections are executed, but not all are necessary!

Needs to fit into main memory!

BINDER algorithm – workflow

[Papenbrock, Quiane, Naumann: Divide & Conquer-based Inclusion Dependency Discovery, PVL

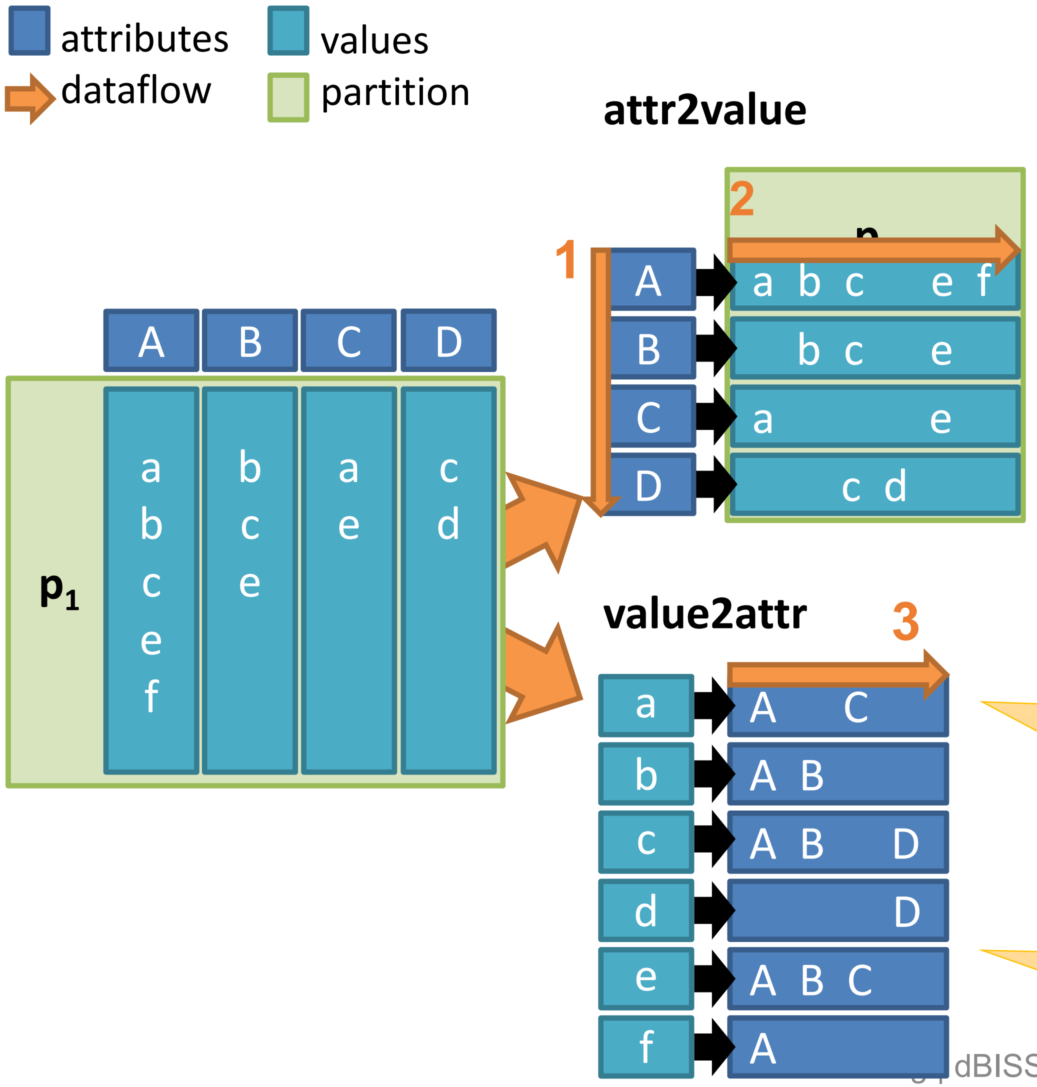
No sortation needed, just hashing



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.

BINDER algorithm – validation

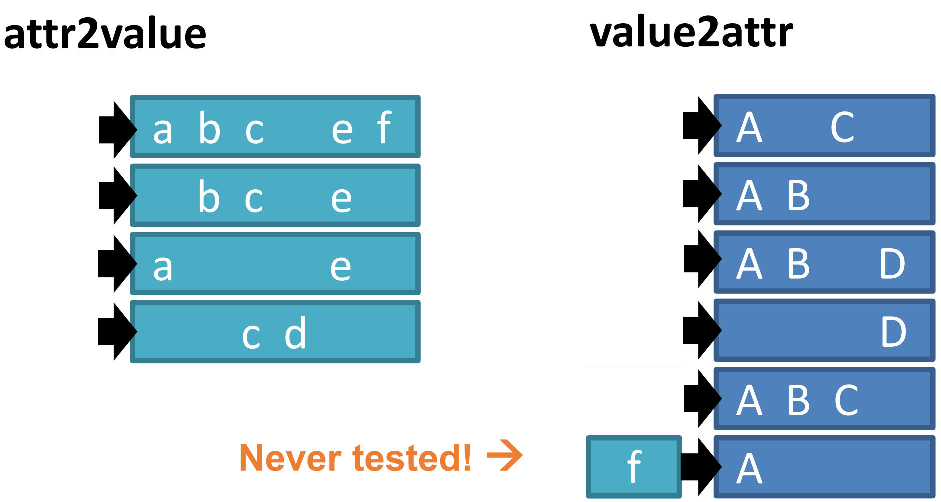


1. Iterate attributes
2. Iterate values
3. If value2attr entry exists
   - Intersect candidates with this list
   - Remove value2attr entry
   - If attribute removed from all candidates
     - Remove entry from attr2value

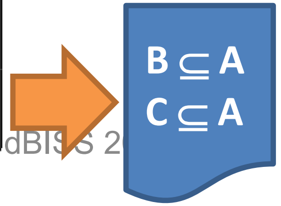
Both indexes fit into main memory due to the partitioning!

see DeMarchi's algorithm

BINDER algorithm – validation example



	A	B	C	D
look up	**B,C,D**	**A,C,D**	**A,B,D**	**A,B,C**

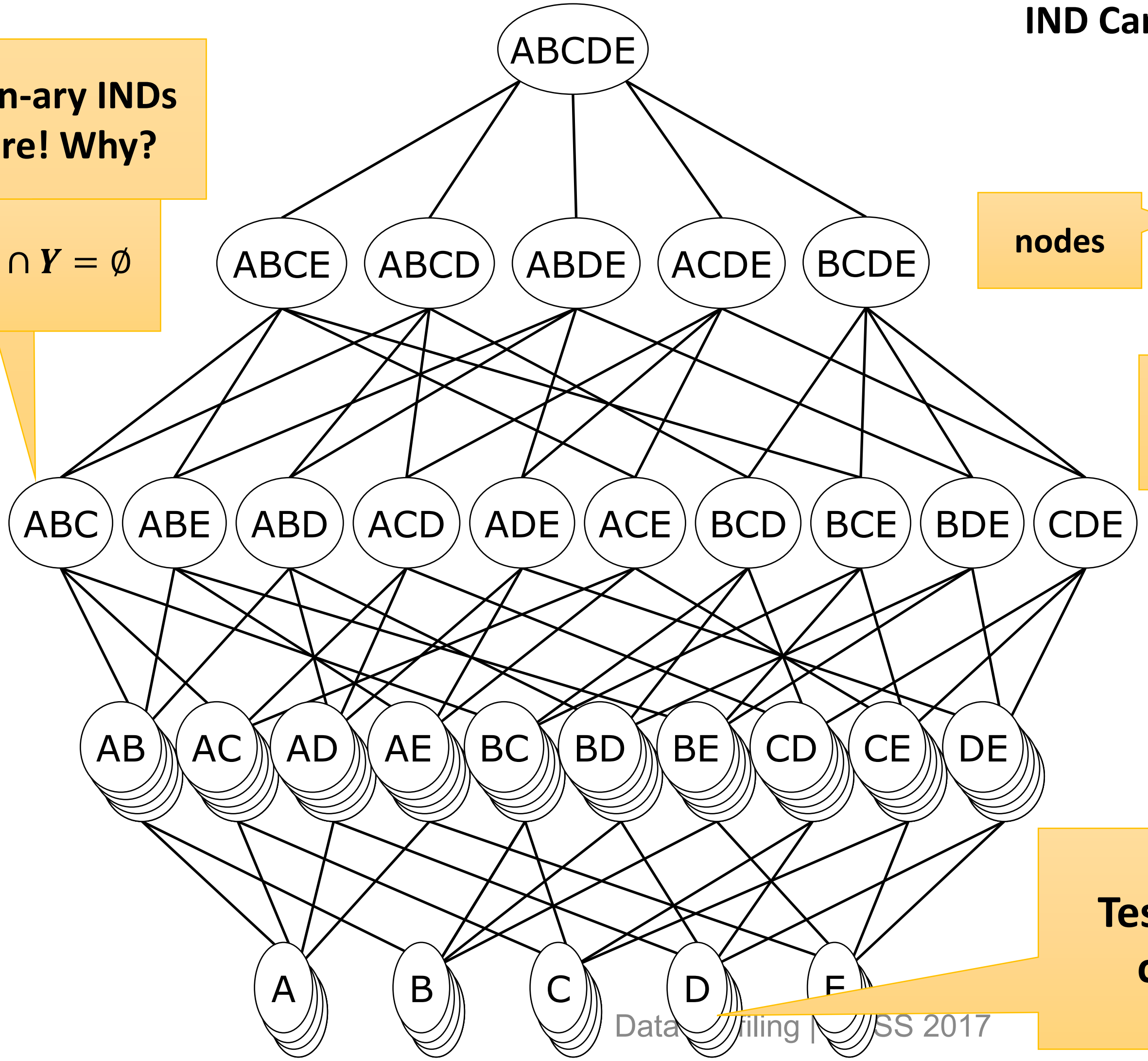


1. Iterate attributes
2. Iterate values
3. If value2attr entry exists
   - Intersect candidates with this list
   - Remove value2attr entry
   - If attribute removed from all candidates
     - Remove entry from attr2value

N-ary IND detection complexity

No n-ary INDs here! Why?

$X \cap Y = \emptyset$



IND Candidates in level k:

$$\overset{X}{\binom{n}{k}} * \overset{Y}{\binom{n-k}{k}} * k!$$

nodes

All permutations

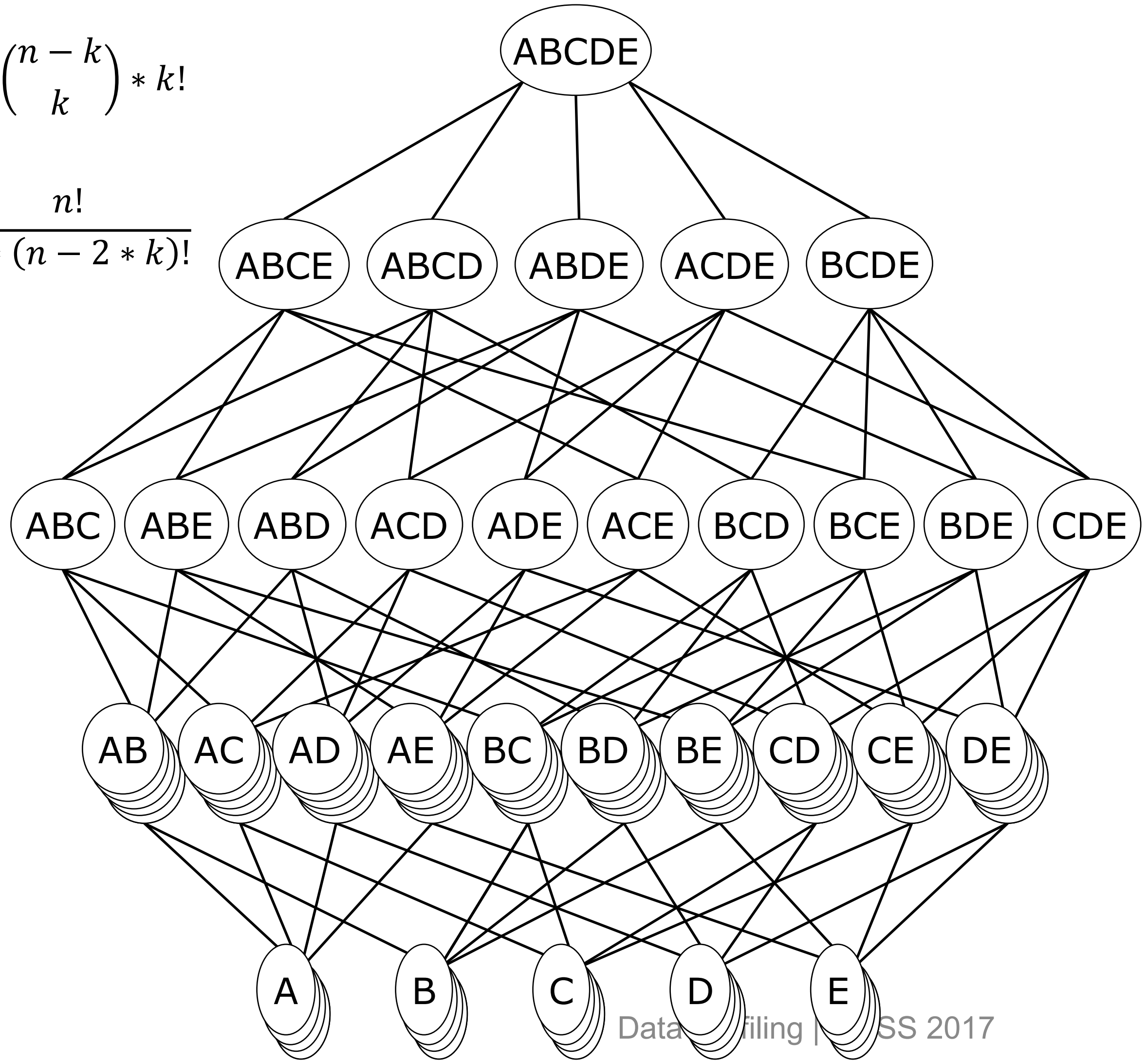
Other, non-overlapping nodes

Test combination with all other combinations of same size!

N-ary IND detection complexity

$$\binom{n}{k} * \binom{n-k}{k} * k!$$

$$= \frac{n!}{k! * (n-2*k)!}$$



$$\binom{5}{5} * \binom{5-5}{5} * 5! \sim 0$$

$$\binom{5}{4} * \binom{5-4}{4} * 4! \sim 0$$

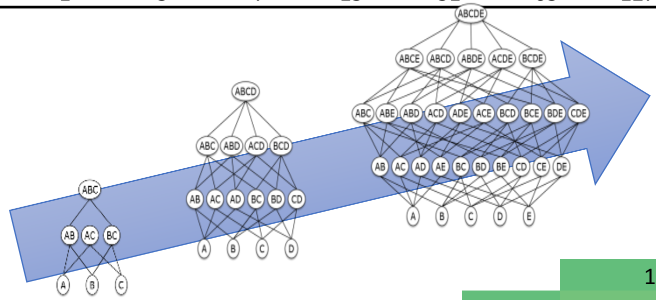
$$\binom{5}{3} * \binom{5-3}{3} * 3! \sim 0$$

$$\binom{5}{2} * \binom{5-2}{2} * 2! = 60$$

$$\binom{5}{1} * \binom{5-1}{1} * 1! = 20 = n^2 - n$$

N-ary IND detection complexity

Unique Column Combinations

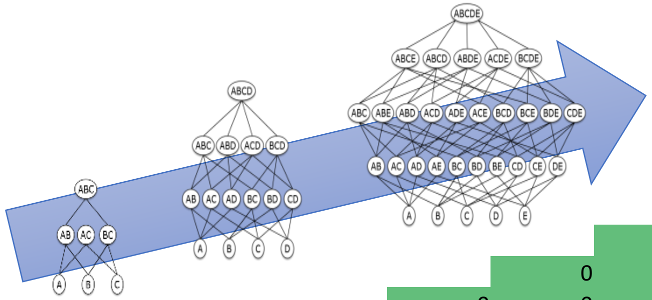


Number of levels: k

total	1	3	7	15	31	63	127	255	511	1023	2047	4095	8191	16383	32767
15															1
14														1	15
13													1	14	105
12												1	13	91	455
11											1	12	78	364	1365
10										1	11	66	286	1001	3003
9									1	10	55	220	715	2002	5005
8								1	9	45	165	495	1287	3003	6435
7							1	8	36	120	330	792	1716	3432	6435
6						1	7	28	84	210	462	924	1716	3003	5005
5					1	6	21	56	126	252	462	792	1287	2002	3003
4				1	5	15	35	70	126	210	330	495	715	1001	1365
3			1	4	10	20	35	56	84	120	165	220	286	364	455
2		1	3	6	10	15	21	28	36	45	55	66	78	91	105
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Number of attributes: n

Inclusion Dependencies



Number of levels: k

total	0	2	6	24	80	330	1302	5936	26784	133650	669350	3609672	19674096	113525594	664400310
15															0
14														0	0
13													0	0	0
12												0	0	0	0
11											0	0	0	0	0
10										0	0	0	0	0	0
9									0	0	0	0	0	0	0
8								0	0	0	0	0	0	0	0
7							0	0	0	0	0	0	0	17297280	259459200
6						0	0	0	0	0	0	665280	8648640	60540480	302702400
5					0	0	0	0	0	30240	332640	1995840	8648640	30270240	90810720
4				0	0	0	0	1680	15120	75600	277200	831600	2162160	5045040	10810800
3			0	0	0	120	840	3360	10080	25200	55440	110880	205920	360360	600600
2		0	0	12	60	180	420	840	1512	2520	3960	5940	8580	12012	16380
1	0	2	6	12	20	30	42	56	72	90	110	132	156	182	210
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Number of attributes: n

MIND & BINDER – candidate generation

- **Apriori algorithm:**
  - Bottom-up lattice traversal strategy
  - Input: all valid attribute combinations of size n
  - Output: all candidate attribute combinations of size n+1

- **Adaption for n-ary IND detection:**
  - Let R_i be the i-th relation in the relational schemata R. For each valid IND $R_j[X] \subseteq R_k[Y]$ with $|X|=|Y|=n$ generate all IND candidates $R_j[XA] \subseteq R_k[YB]$ so that:

1. $R_j[X] \subseteq \subseteq R_k[Y]$ and $R_j[A] \subseteq \subseteq R_k[B]$ (both are valid INDs)
2. $\forall X_i \in X: X_i < A$ (INDs are permutable; do not generate them twice)
3. $A \notin X,\ B \notin Y$ (do not generate trivial candidates)

Intrinsic limitations of IND algorithms

- Observations: all IND algorithms follow a common pattern

Algorithm	Phase 1 Data Reorganization	Phase 2 Comparison
De Marchi	Create Inverted Index	Intersect Attribute Groups
SPIDER	Sort Columns	Value-based Iteration
BINDER	Partition Columns	In-Memory Partition Comparison

- e.g., IND $A \subseteq B$
  - to prove, need to read A completely
  - to disprove, need to read B completely
- Data reorganization is the most expensive phase
  - I/O-heavy workload, but other phase brings considerable I/O as well

Visualisation

[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]

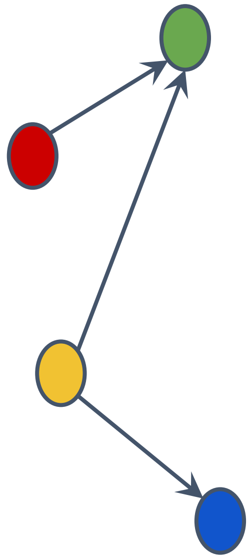
...

Visualisation

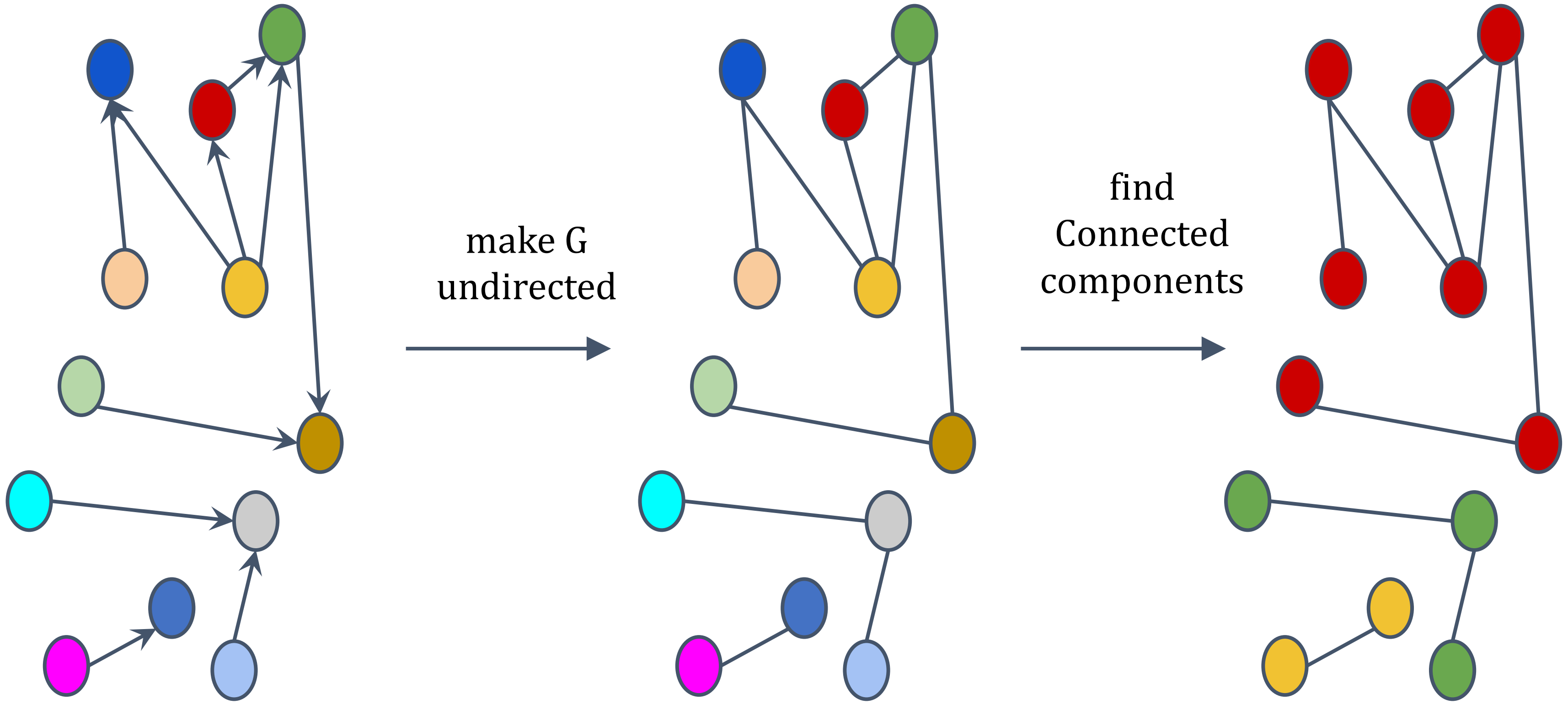
INDs = {
$R_1.A \subseteq R_2.B$,
$R_3.A \subseteq R_1.D$,
$R_3.C \subseteq R_2.A$,
$R_3.B \subseteq R_4.A$
}

G= (
V = {
R_1, R_2, R_3, R_4
},

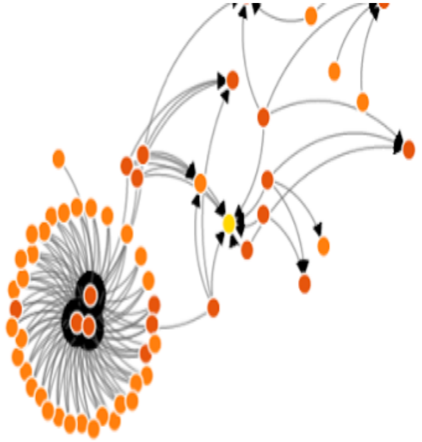
E = { (R_1, R_2), (R_3, R_1),
(R_3, R_2), (R_3, R_4)
}
)



Visualisation



Interactive Application



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

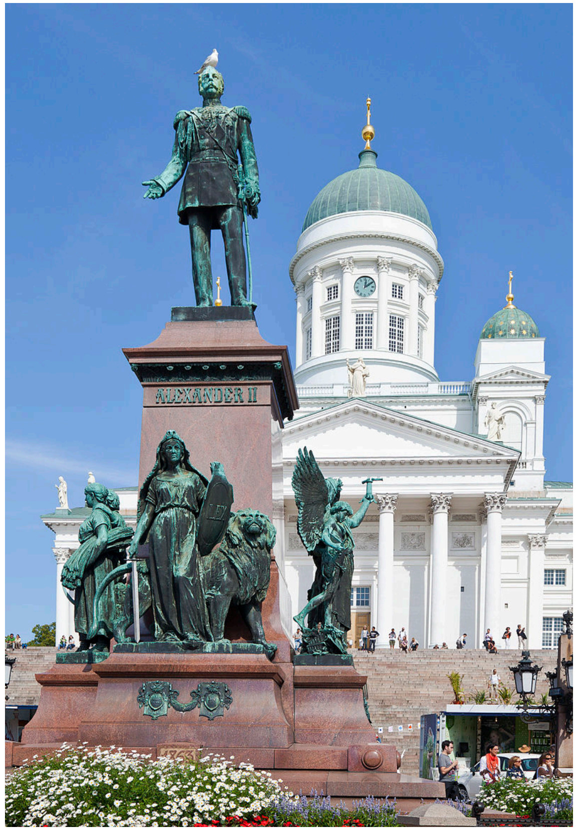
allFilters

More Dependencies

- Conditional …
  - Uniques
  - FDs
  - INDs
- Approximate ..
  - ..
- Order dependencies [Langer, Naumann: Discovering Order Dependencies, VLDBJ'15]
- Matching dependencies [Fan et al.:Reasoning about record matching rules, VLDB'09]

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, INDs, FDs
  - and their discover algorithms
- **Outlook**
  - **Functionality**
  - **Semantics**

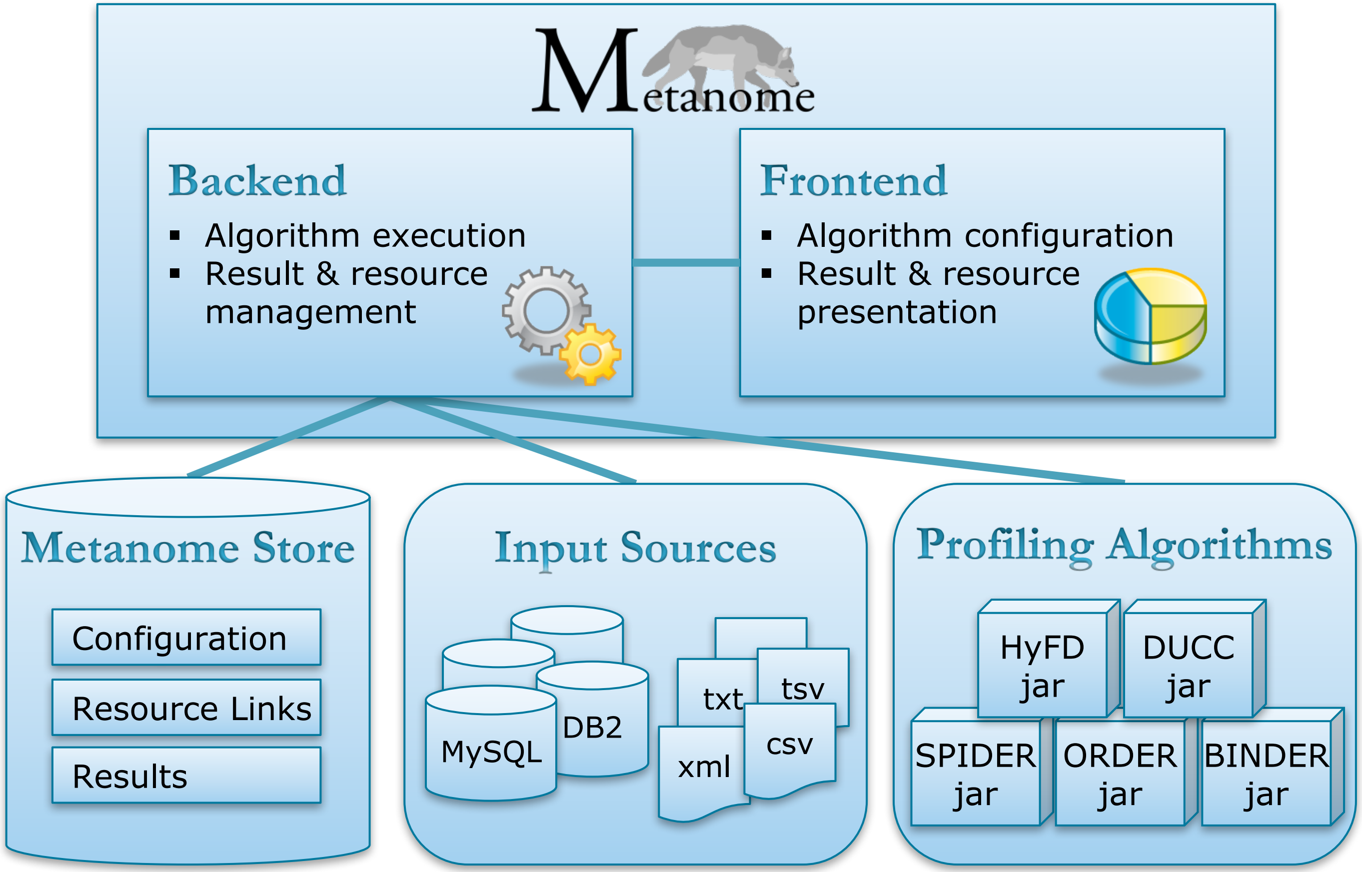


Part Overview

- The Metanome
  Data Profiling Framework
- Functional challenges
- Non-functional challenges
- Semantics of Dependencies

The Metanome Data Profiling Framework

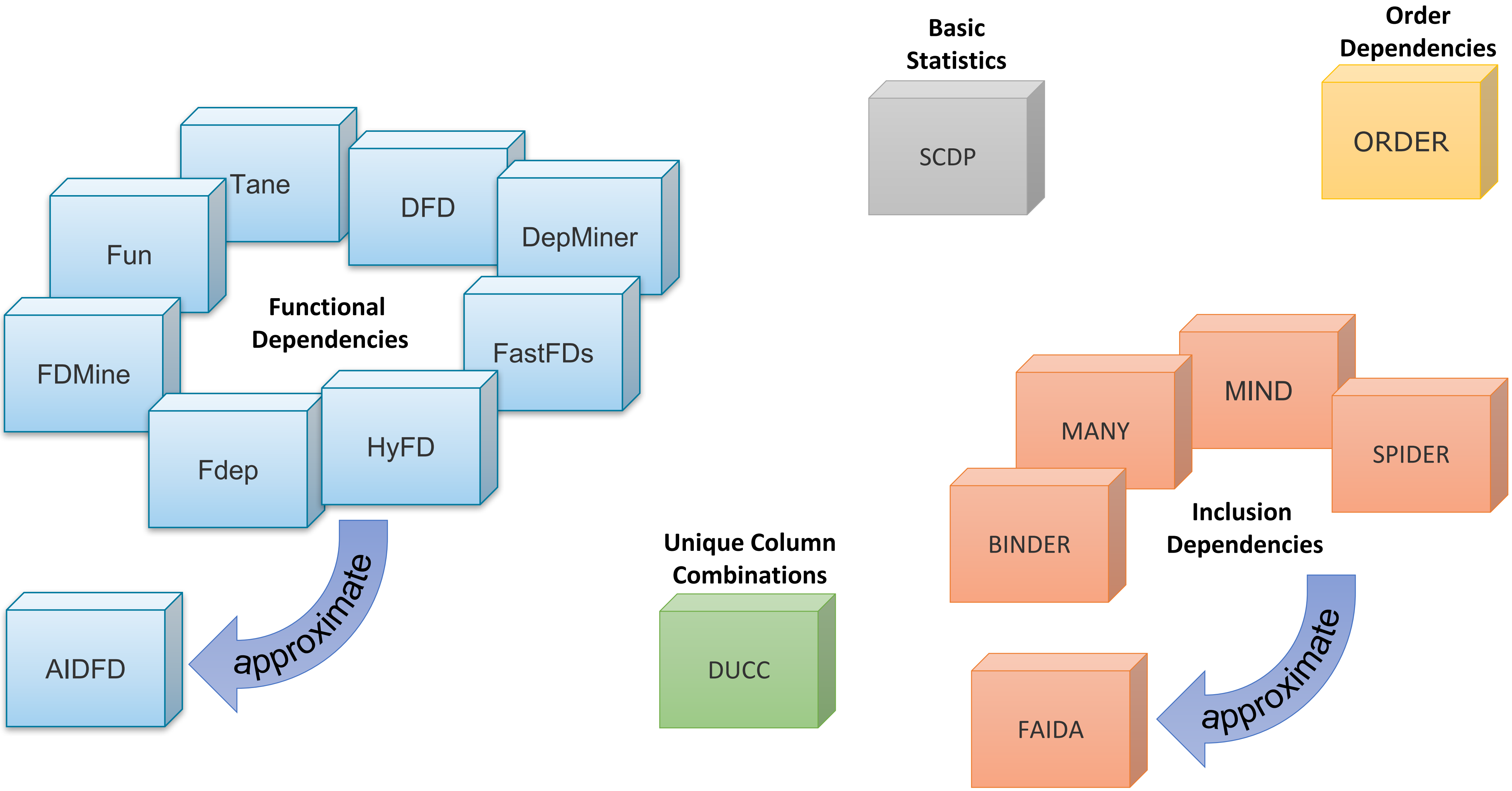
Metanome Data Profiling Tool



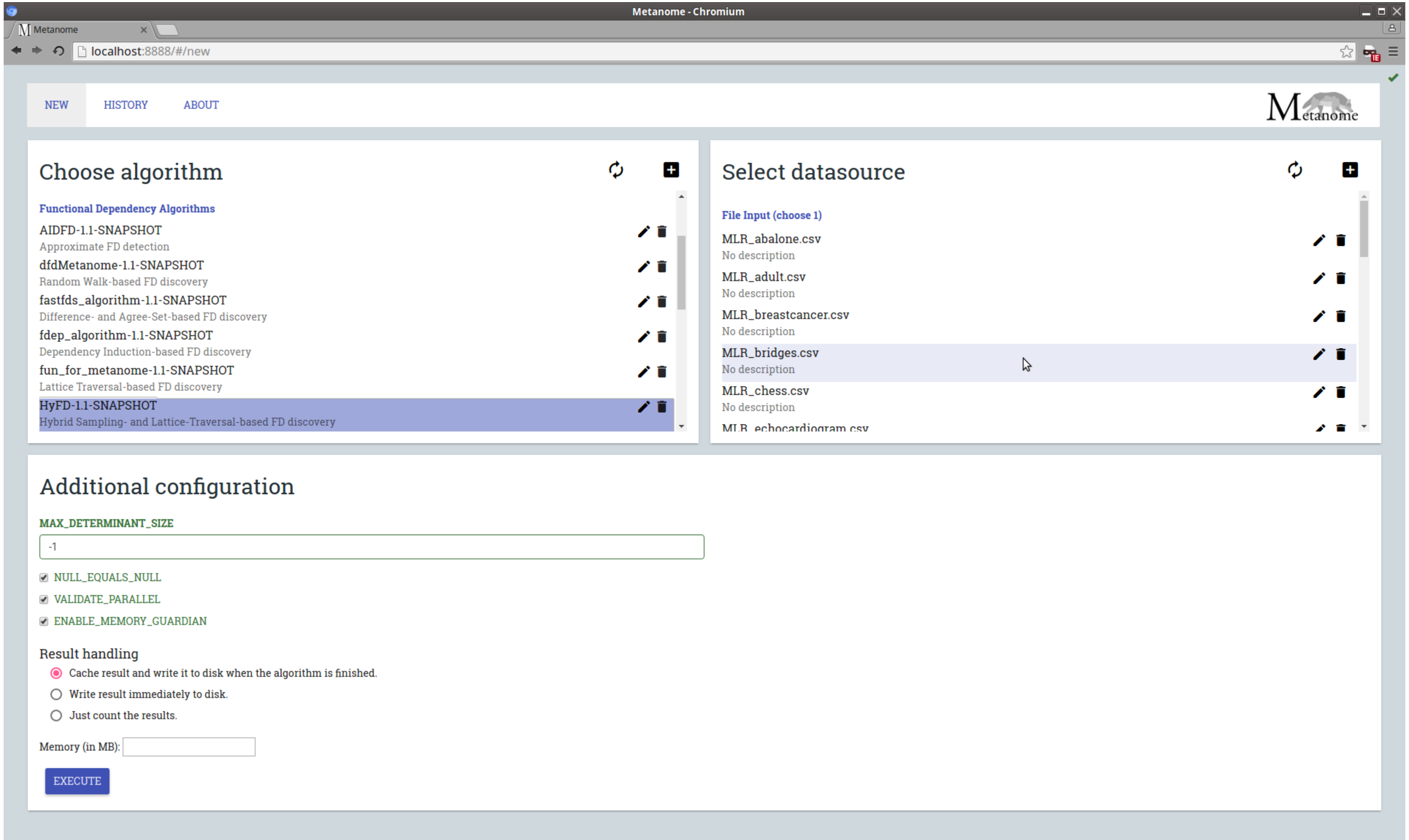
Open source framework, tool plus many algorithms

www.metanome.de

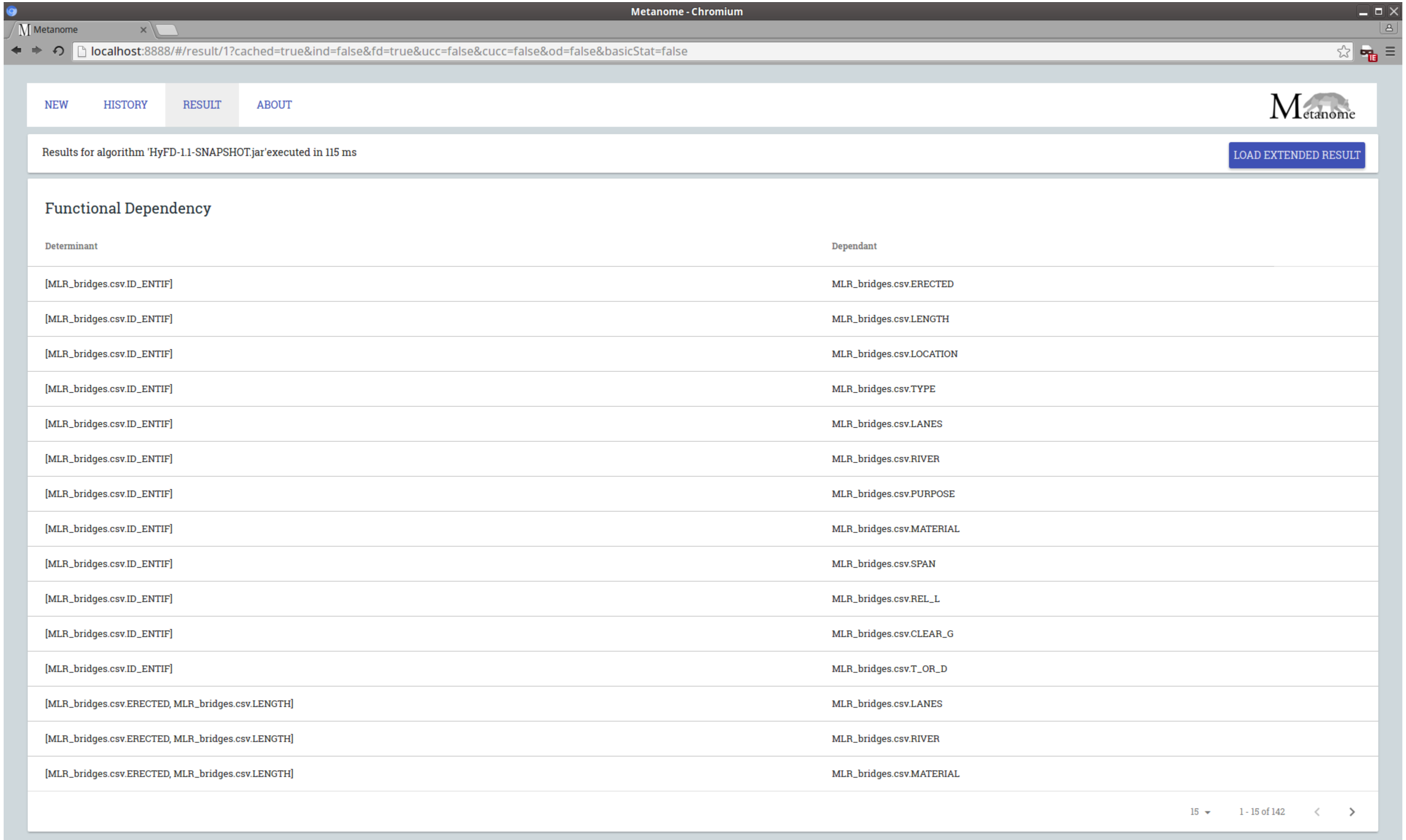
Profiling Algorithms



Metanome User Experience

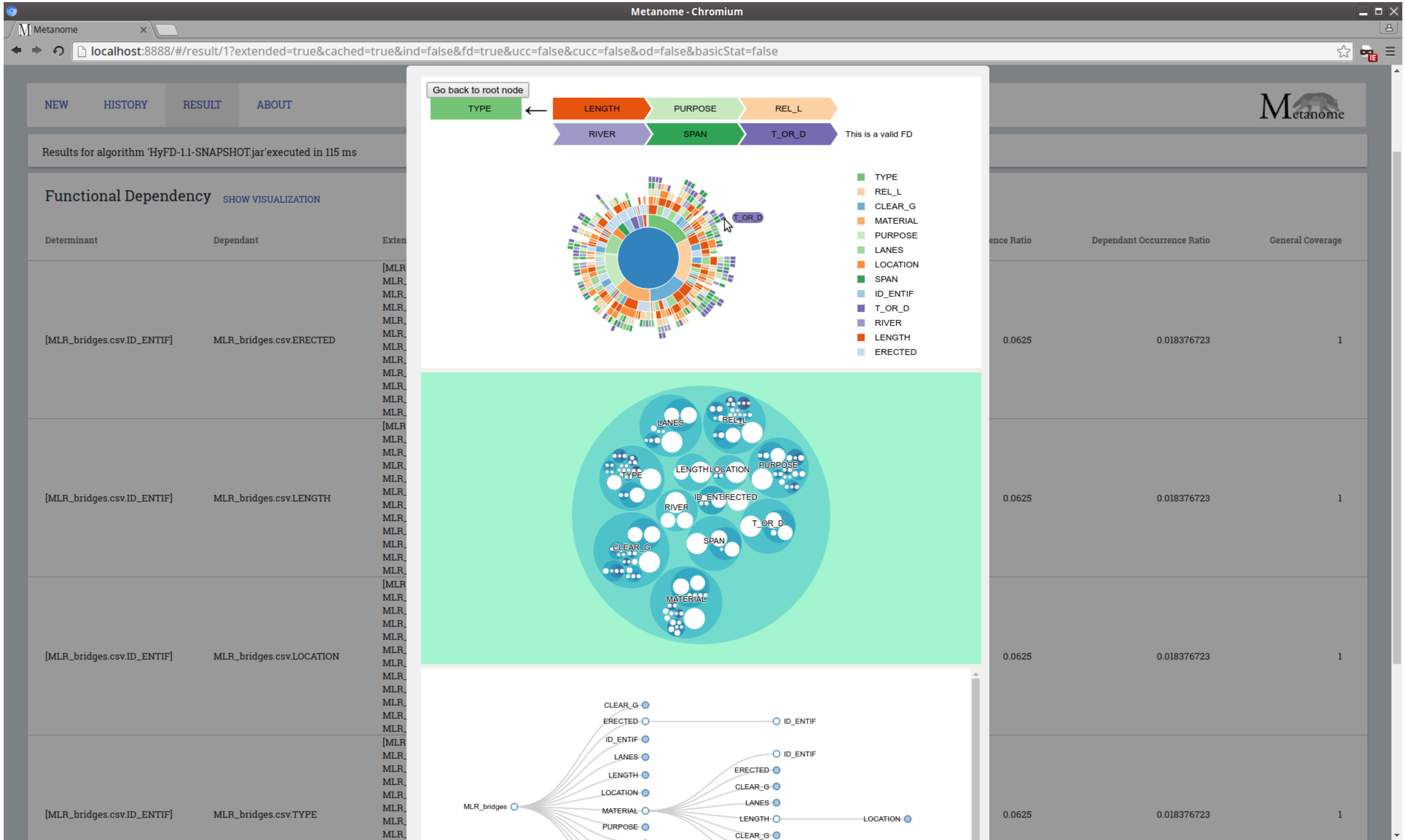


Metanome User Experience



Determinant	Dependant
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.ERECTED
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.LENGTH
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.LOCATION
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.TYPE
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.LANES
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.RIVER
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.PURPOSE
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.MATERIAL
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.SPAN
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.REL_L
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.CLEAR_G
[MLR_bridges.csv.ID_ENTIF]	MLR_bridges.csv.T_OR_D
[MLR_bridges.csv.ERECTED, MLR_bridges.csv.LENGTH]	MLR_bridges.csv.LANES
[MLR_bridges.csv.ERECTED, MLR_bridges.csv.LENGTH]	MLR_bridges.csv.RIVER
[MLR_bridges.csv.ERECTED, MLR_bridges.csv.LENGTH]	MLR_bridges.csv.MATERIAL

Metanome User Experience

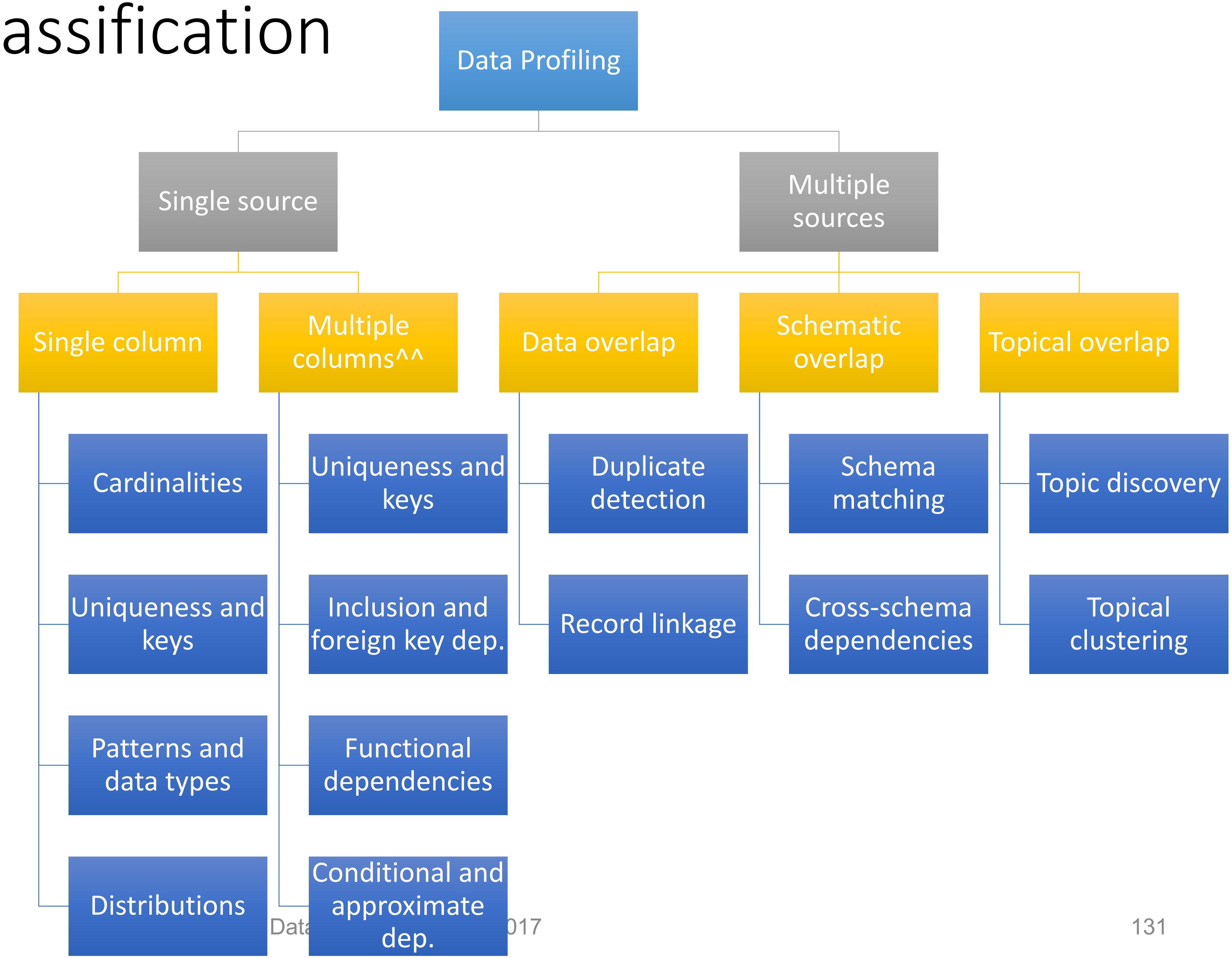


Extending the Functionality of Data Profiling

Many Other Kinds of Dependencies

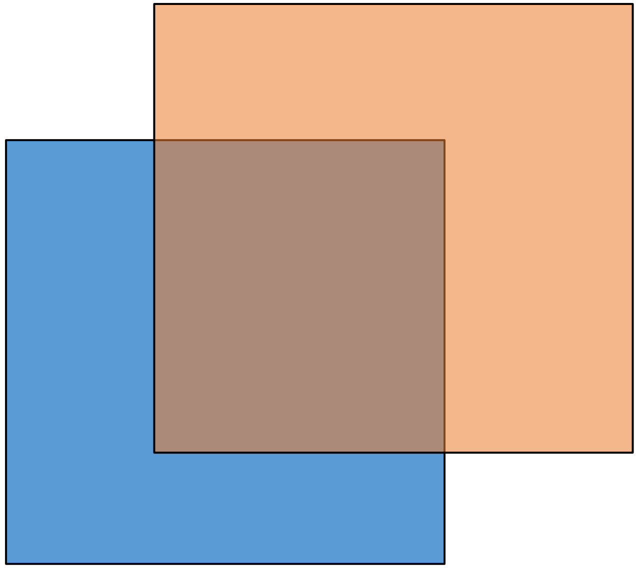
[Abiteboul, Hull, Vianu: Foundations of Databases, 1995]

Extended Classification of Profiling Tasks



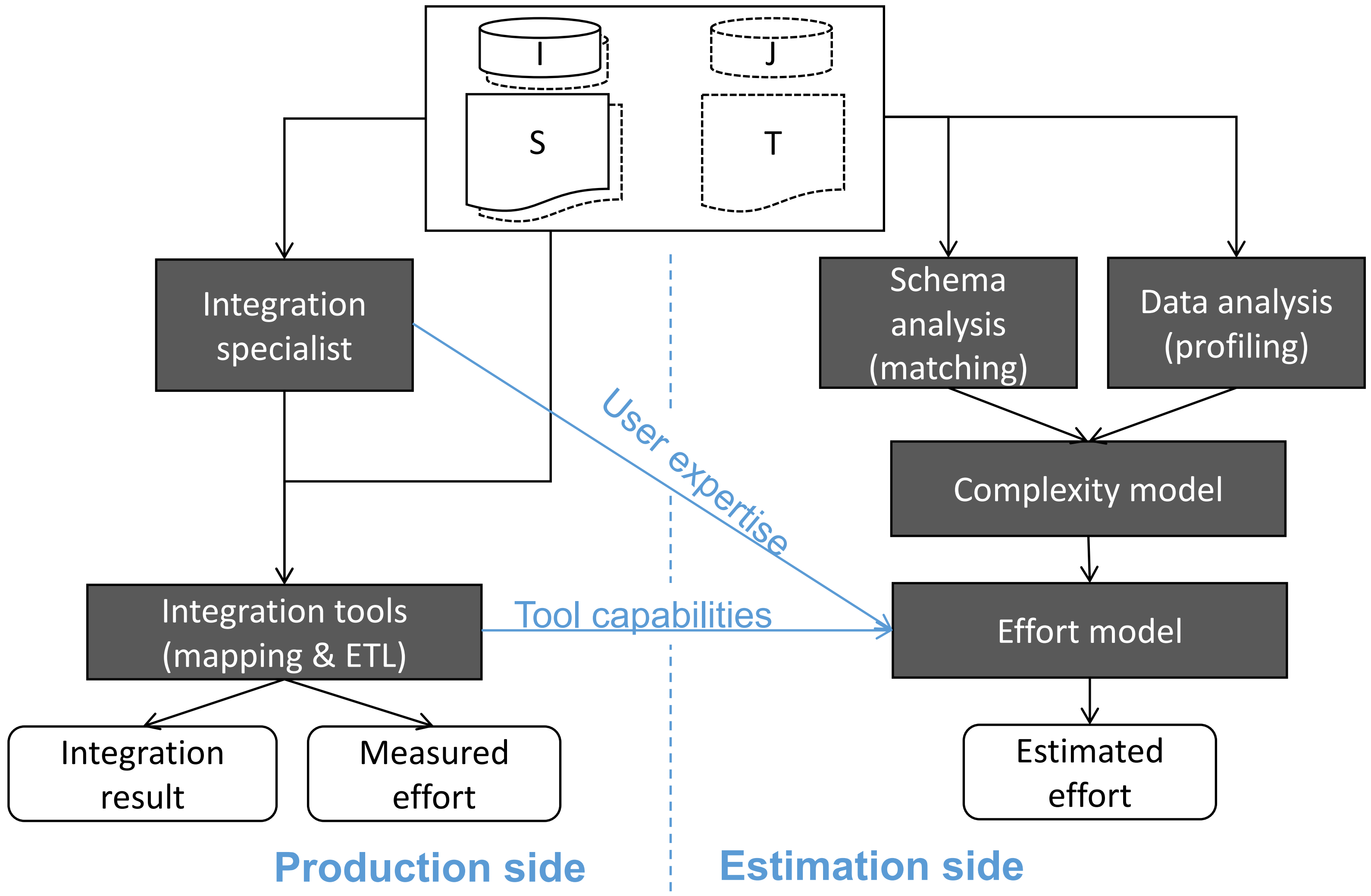
Profiling for Integration

- 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]

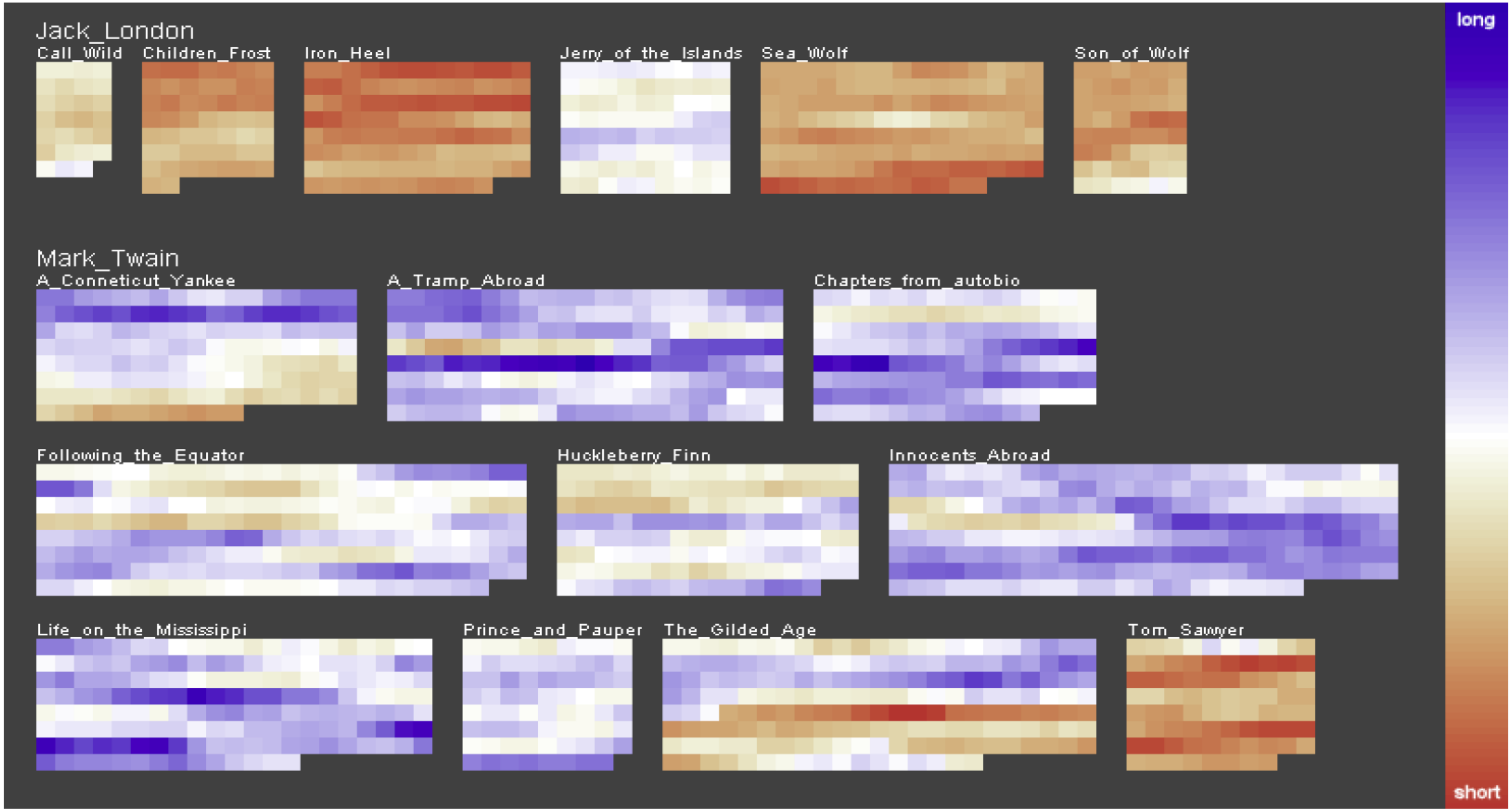
Profiling new Types of Data

- 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

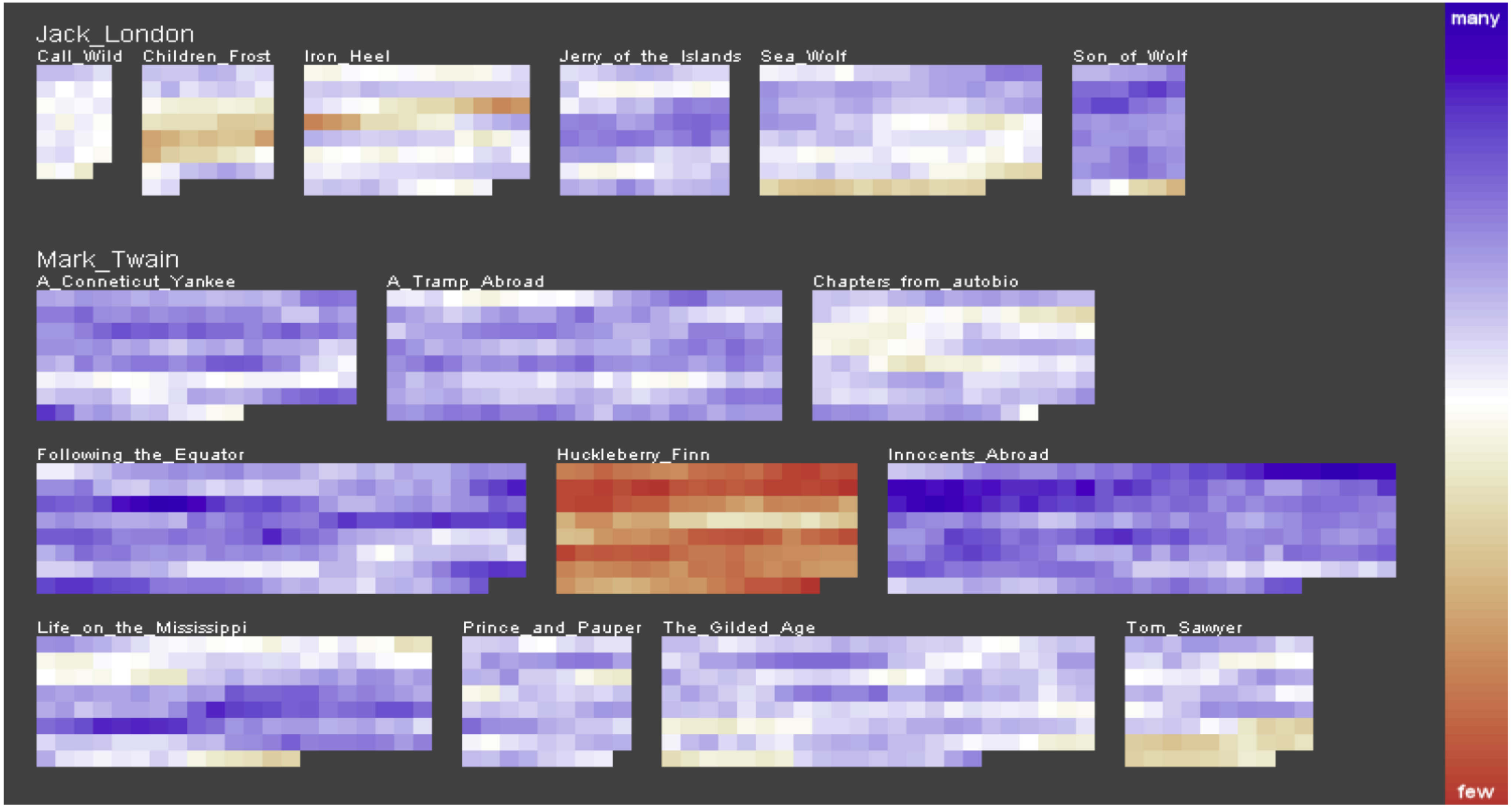
- Statistical measures
  - Syllables per word
  - Sentence length
  - Proportions of parts of speech
- Vocabulary measures
  - Frequencies of specific words
  - Type-token ratio
  - Simpson’s index (vocabulary richness)
  - Number of hapax (dis)legomena
    - Token that occurs exactly once (twice) in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

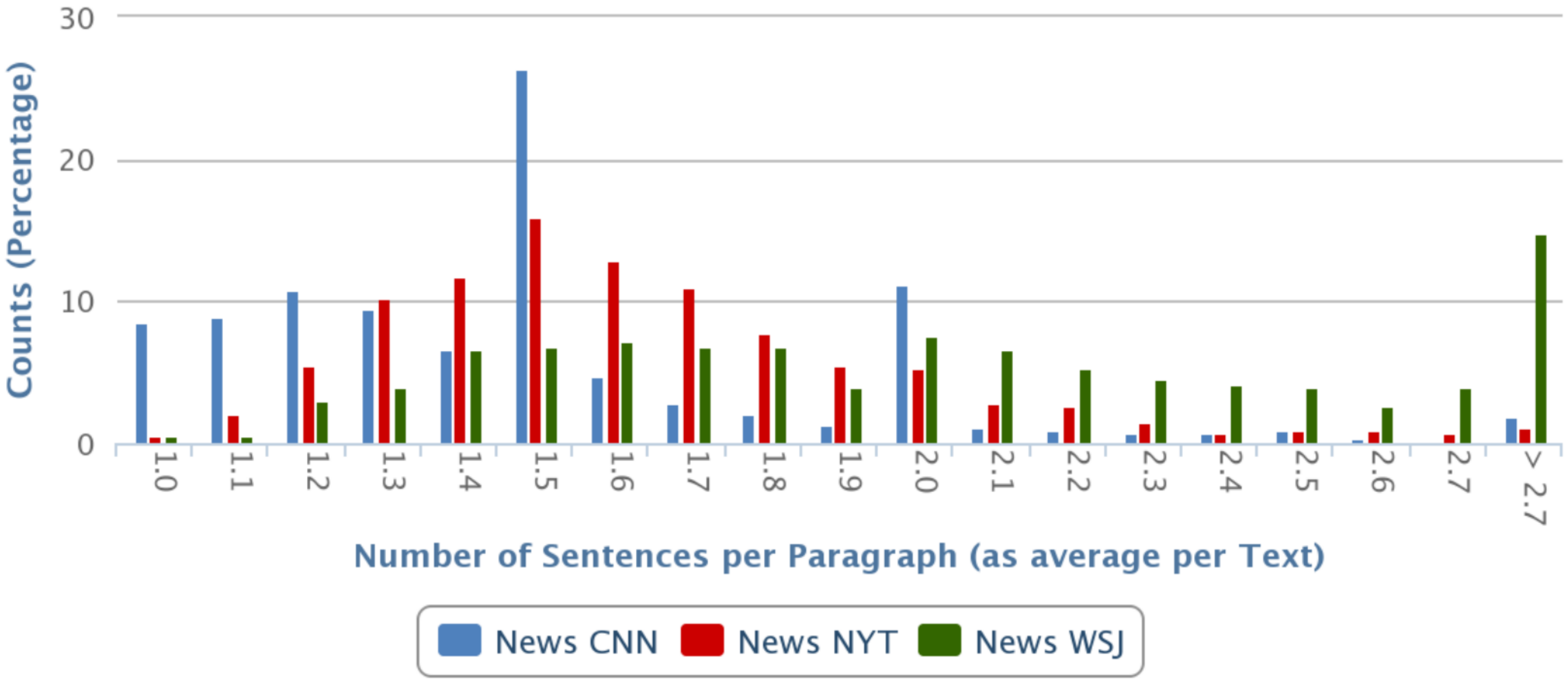


[Keim and Oelke: Literature Fingerprinting: A New Method for Visual Literary Analysis. IEEE VAST 2007]

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

Profiling Challenges

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

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

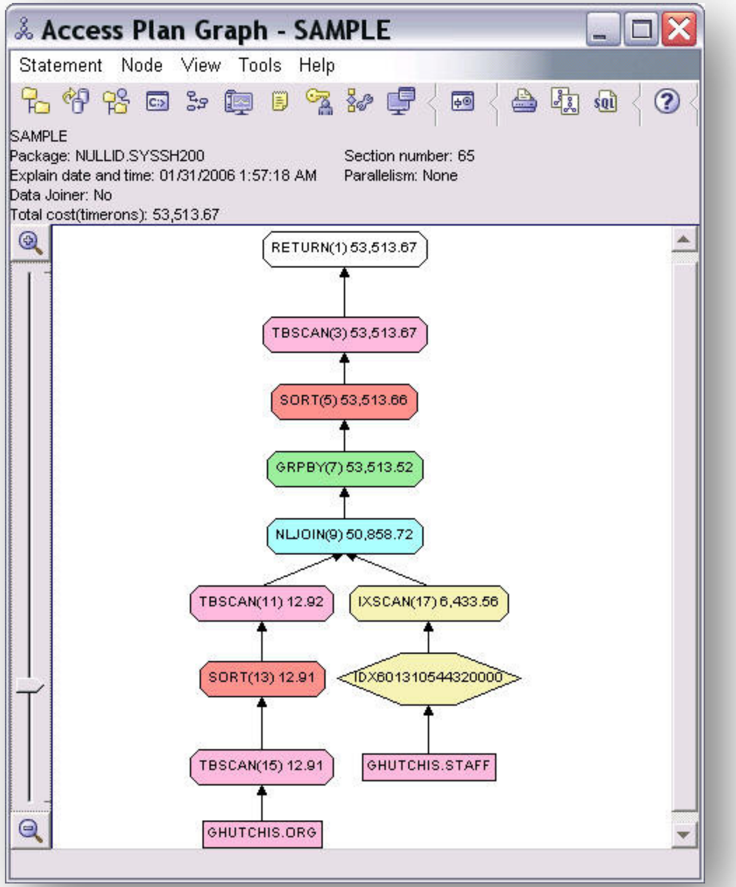
Temporal Profiling

- 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.



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)
  - Example: TPCH (next slide)

Recent work

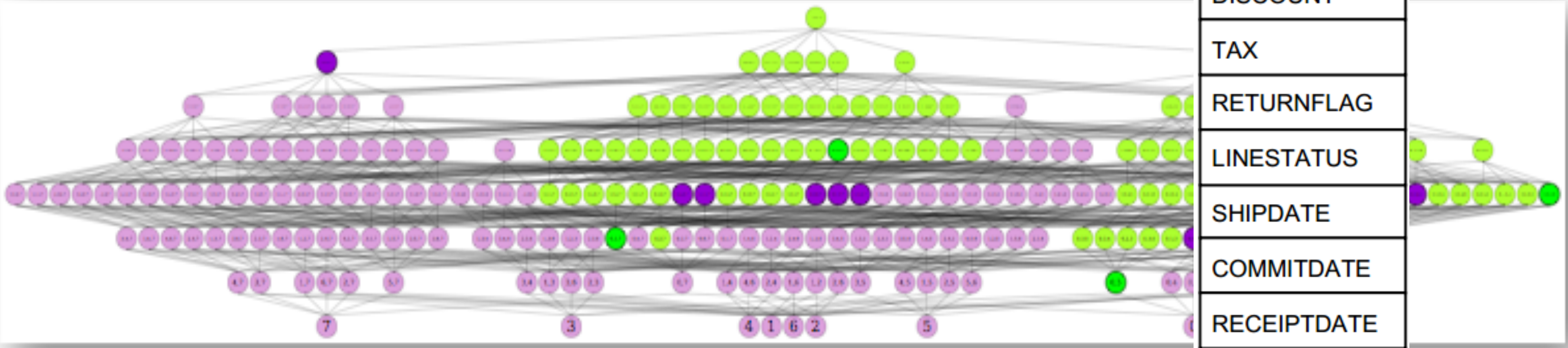
[Arocena et al. : Messing Up with BART: Error Generation for Evaluating Data-Cleaning Algorithms. PVLDB 9(2), 2015]
[Arocena et al. : The iBench Integration Metadata Generator . PVLDB 9(3), 2015]

TPCH – Uniques and Non-Uniques

- Using the first 8 columns of the lineitems table
- Using a scale-factor of 0.1

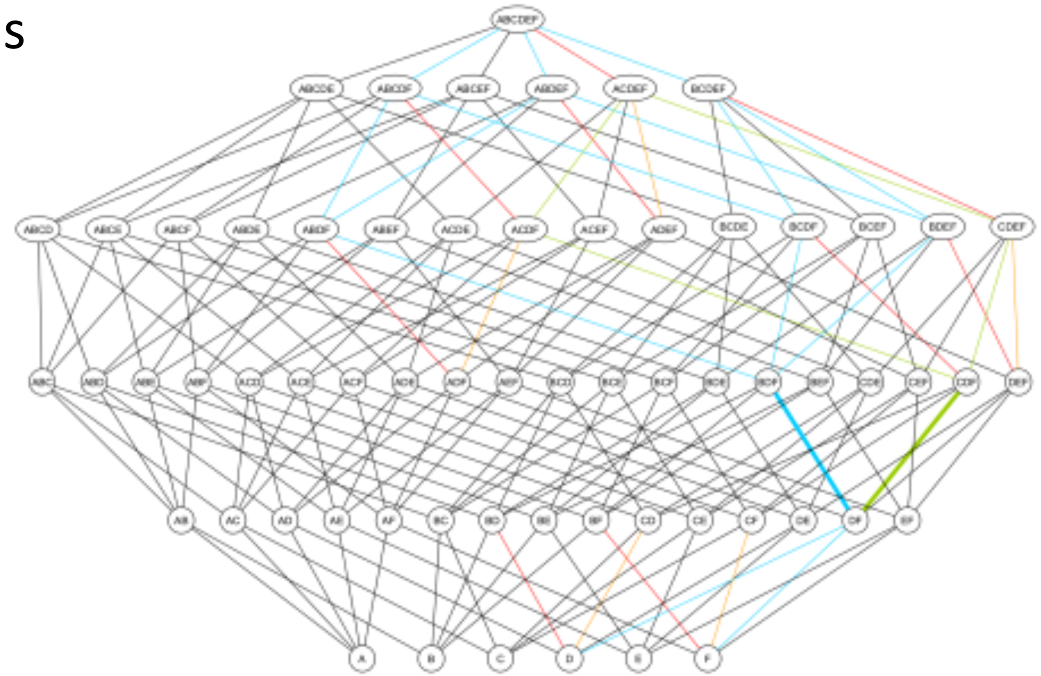
LINEITEM (L_)
SF*6,000,000

ORDERKEY
PARTKEY
SUPPKEY
LINENUMBER
QUANTITY
EXTENDEDPRICE
DISCOUNT
TAX
RETURNFLAG
LINESTATUS
SHIPDATE
COMMITDATE
RECEIPTDATE
SHIPINSTRUCT
SHIPMODE
COMMENT



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



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

Summary

- 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
    - Topical overlap
      - Topic discovery
      - Topical clustering
    - Schematic overlap
      - Schema matching
      - Cross-schema dependencies
    - Data overlap
      - Duplicate detection
      - Record linkage

Contact us for slides – online soon

- abedjan@csail.mit.edu
- felix.naumann@hpi.de