An Introduction to Data Profiling

16.4.2013

Felix Naumann

Overview

- Profiling tasks
- Profiling tools
- Visualization
- Next generation profiling
- Profiling challenges

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.

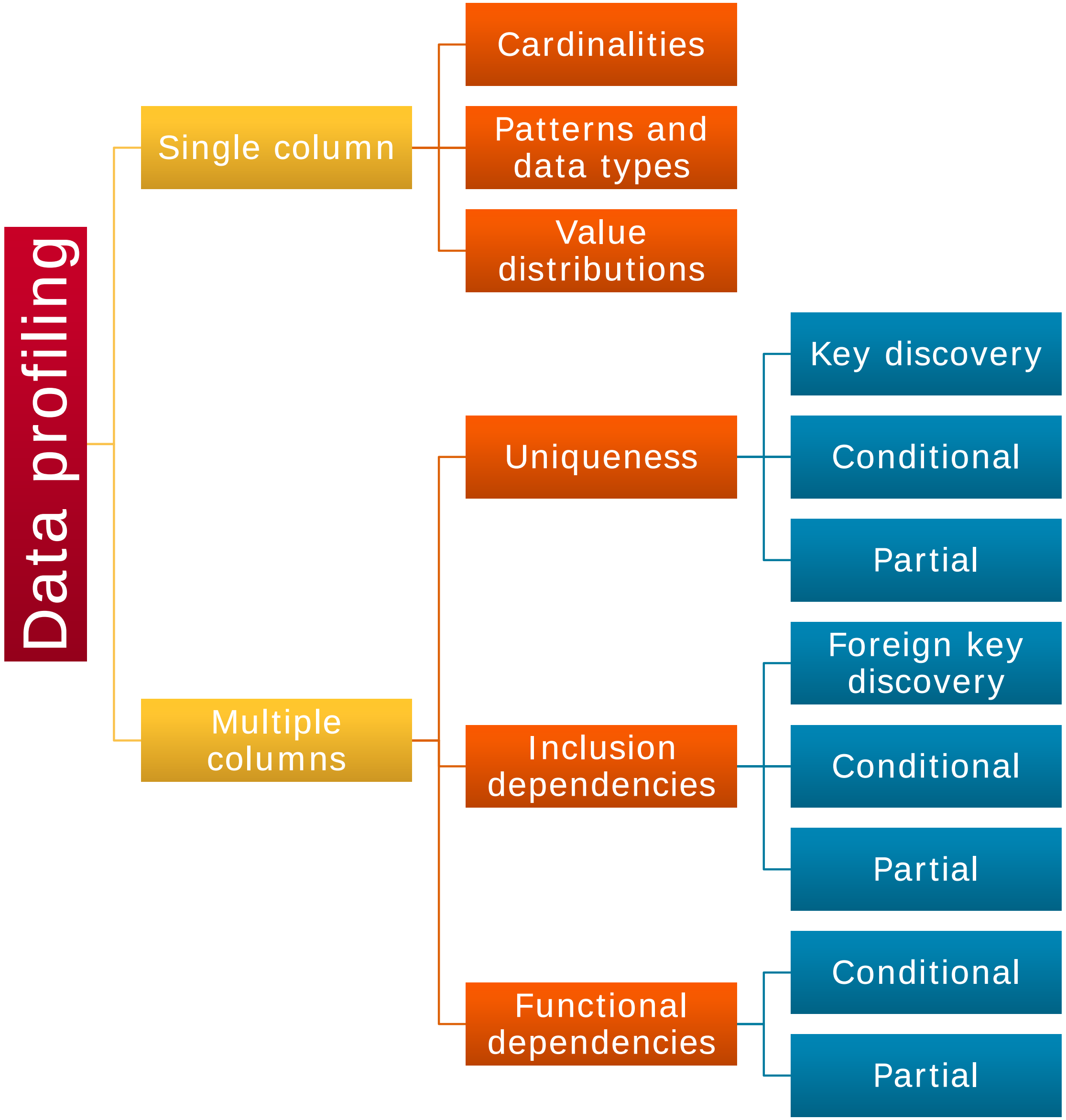
  Wikipedia 03/2013

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

  Ted Johnson, Encyclopedia of Database Systems

- Define as a set of data profiling tasks / results

Classification of Profiling Tasks



Single-column vs. multi-column

- Single column profiling
  - Most basic form of data profiling
  - Assumption: All values are of same type
  - Assumption: All values have some common properties
    - That are to be discovered
  - Often part of the basic statistics gathered by DBMS
  - Complexity: Number of values/rows
- Multicolumn profiling
  - Discover joint properties
  - Discover dependencies
  - Complexity: Number of columns and number of values

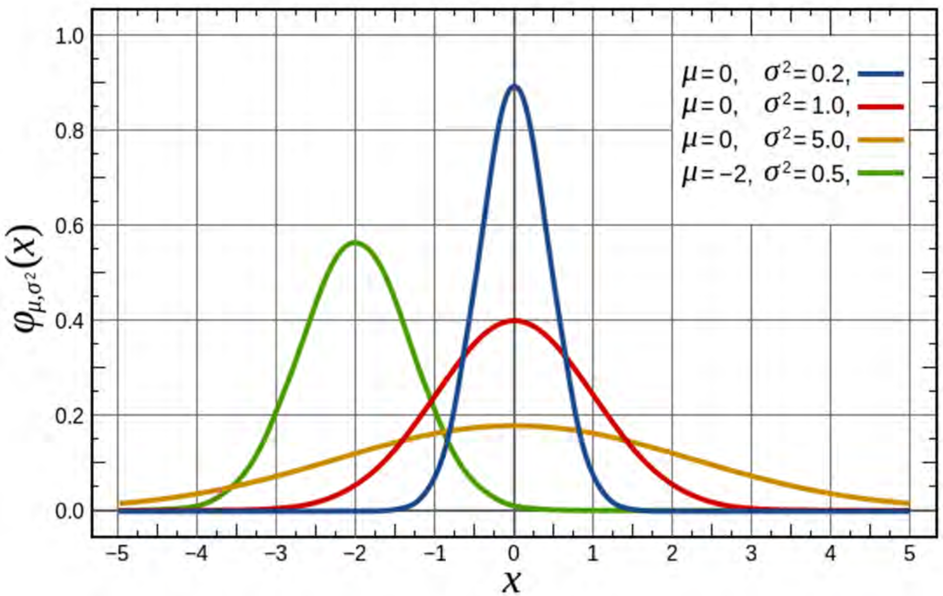
Cardinalities

- Number of values
- Number of distinct values
- Number of NULLs
- MIN and MAX value

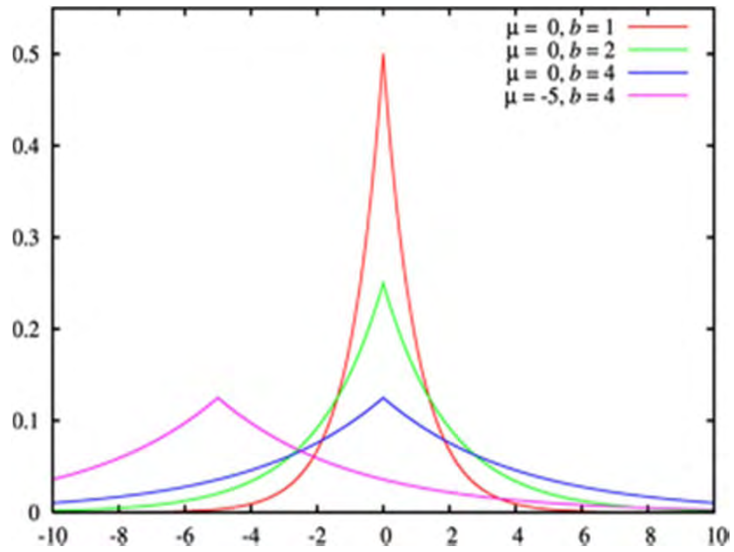
- Useful for
  - Query optimization
  - Categorization of attribute
  - Relevance of attribute

Numeric distributions

- Probability distribution for numeric values
- Detect whether data follows some well-known distribution
  - Determine that distribution function for data values
- If no specific/useful function detectable: histograms



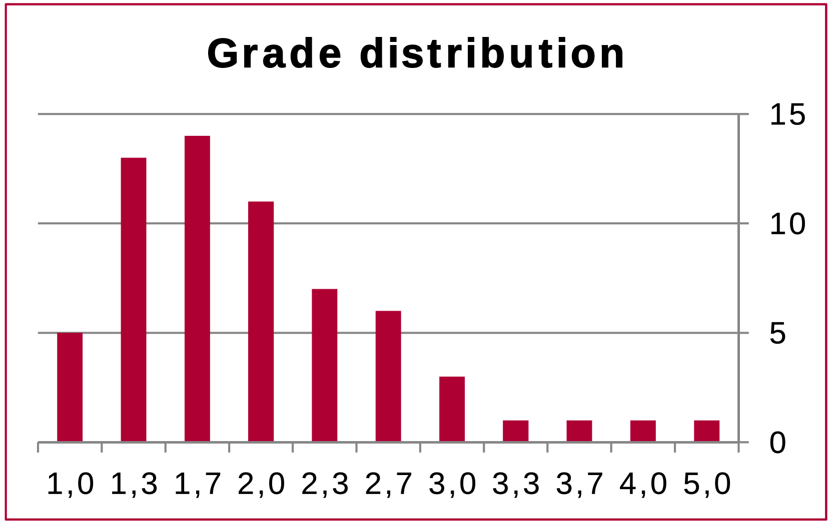
Normal distributions

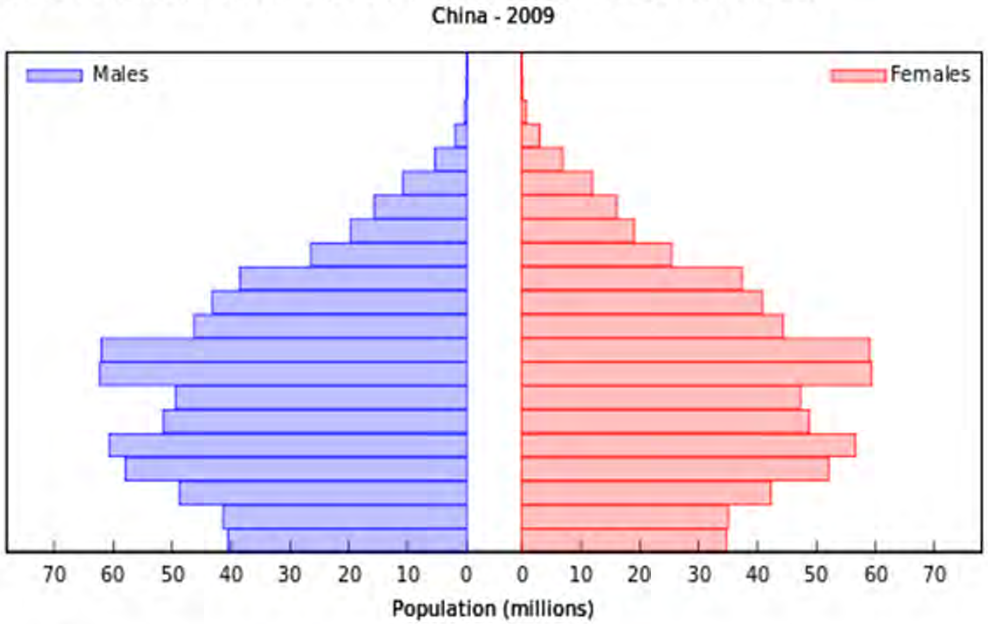


Laplace distributions

Histograms

- Determine (and display) value frequencies for value intervals
  - Or for individual values
  - Estimation of probability distribution for continuous variables





- Useful for
  - Query optimization
  - Outlier detection
  - Visualize distribution

Data types and value patterns

- String vs. number
- String vs. number vs. date
- Categorical vs. continuous
- 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

Uniqueness and keys

- Unique column
  - Only unique values
- Unique column combination
  - Only unique value combinations
  - Minimality: No subset is unique
- 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?
- Useful for
  - Schema design, data integration, indexing, optimization
  - Inverse: non-uniques are duplicates

Inclusion dependencies and foreign keys

- $A \subseteq B$: 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

Functional dependencies

- „X → A“: whenever two records have the same X values, they also have the same A values.

- Useful for
  - Schema design
    - Normalization
    - Keys

Partial 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-source cINDs

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, 2000
  - Profiling: Individual attributes
  - Mining: Multiple attributes

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

- Data profiling as preparation for any other data management task

Overview

- Profiling tasks
- Profiling tools
- Visualization
- Next generation profiling
- Profiling challenges

Data profiling tools and algorithms

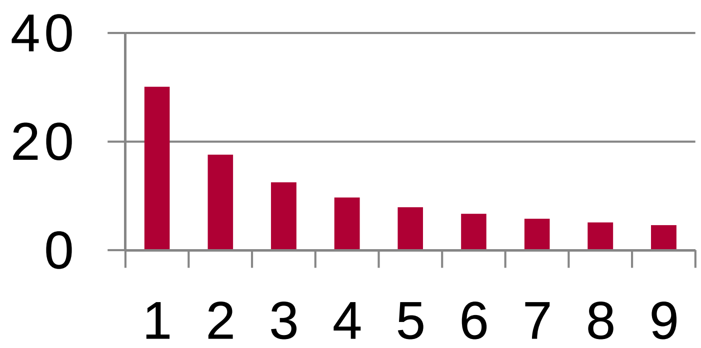
- Often packaged with data quality / data cleansing software
- **IBM InfoSphere Information Analyzer**
  - http://www.ibm.com/software/data/infosphere/information-analyzer/
- Oracle Enterprise Data Quality
  - http://www.oracle.com/us/products/middleware/data-integration/enterprise-data-quality/overview/index.html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

Very long feature lists

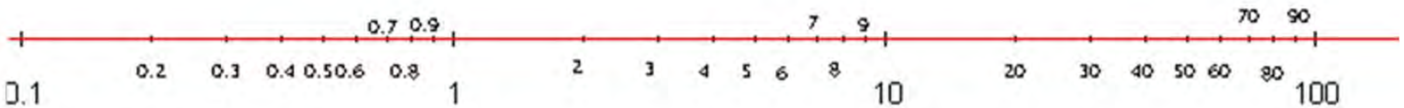
- Num rows
- Min value length
- Median value length
- Max value length
- Avg value length
- Precision of numeric values
- Scale of numeric values
- Quartiles
- Basic data types
- Num distinct values ("cardinality")
- Percentage null values
- Data class and data type
- Uniqueness and constancy
- Single-column frequency histogram
- Multi-column frequency histogram
- Pattern discovery (Aa9)
- Soundex frequencies
- Benford Law Frequency
- Single column primary key discovery
- Multi-column primary key discovery
- Single column IND discovery
- Inclusion percentage
- Single-column FK discovery
- Multi-column IND discovery
- Multi-column FK discovery
- Value overlap (cross domain analysis)
- Single-column FD discovery
- Multi-column FD discovery
- Text profiling

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 distibuted



Picking a random *x* position uniformly on this number line, roughly 30% of the time the first digit of the number will be 1.

Explanation for Benford's Law

- Law is true if the mantissae of the *logarithms* of the numbers are uniformly distributed
  - $1 \leq x < 2 \Rightarrow$ x begins with 1 …
  - $9 \leq x < 10 \Rightarrow$ x begins with 9
  - $\log 1 \leq \log x < \log 2 \Rightarrow$ x begins with 1 …
  - $\log 9 \leq \log x < \log 10 \Rightarrow$ x begins with 9
  - $[\log 1, \log 2] > [\log 9, \log 10]$

- Likely to be (approximately) true if the numbers range over several orders of magnitude.

Explanation for Benford's Law

- Exponential growth
  - Year 1: 1-100
  - Year 2: 101-200
  - Year 3: 201-400
  - Year 4: 401-800
  - Year 5: 801-1600
  - Year 6: 3200
- Function: $2^{(year-1)} \cdot 100$

- First digit
  - 1: 2 years
  - 2: 7 months in year 3
    - $(\log_2(300/100)+1 - 2) * 12$
  - 3: remaining 5 months
  - 4 – 9: few months each
  - 1: again after under 4 months in year 5 for 1 year
    - $(\log_2(10)+1 - 4) * 12$

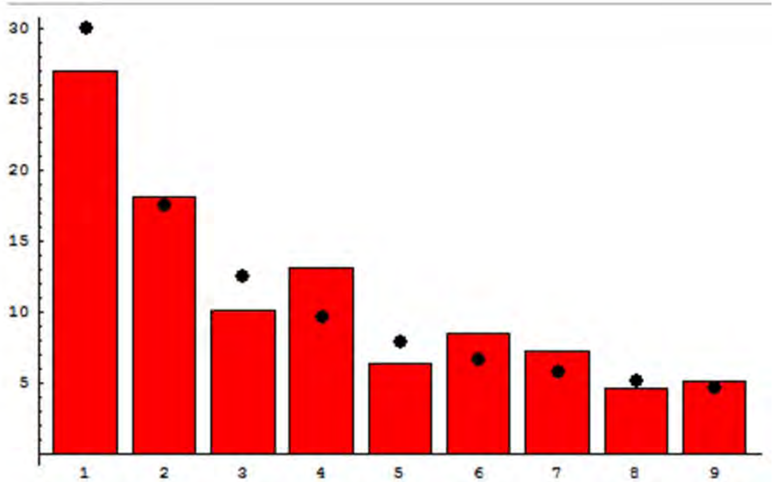
Examples for Benford's Law

Heights of the 60 tallest structures

Leading digit	meters		feet		In Benford's law
	Count	%	Count	%	
1	26	43.3%	18	30.0%	30.1%
2	7	11.7%	8	13.3%	17.6%
3	9	15.0%	8	13.3%	12.5%
4	6	10.0%	6	10.0%	9.7%
5	4	6.7%	10	16.7%	7.9%
6	1	1.7%	5	8.3%	6.7%
7	2	3.3%	2	3.3%	5.8%
8	5	8.3%	1	1.7%	5.1%
9	0	0.0%	2	3.3%	4.6%

- Main application: Fraud detection

- Surface areas of 335 rivers
- Sizes of 3259 US populations
- 104 physical constants
- 1800 molecular weights
- 5000 entries from a mathematical handbook
- 308 numbers contained in an issue of *Reader's Digest*
- Street addresses of the first 342 persons listed in *American Men of Science*

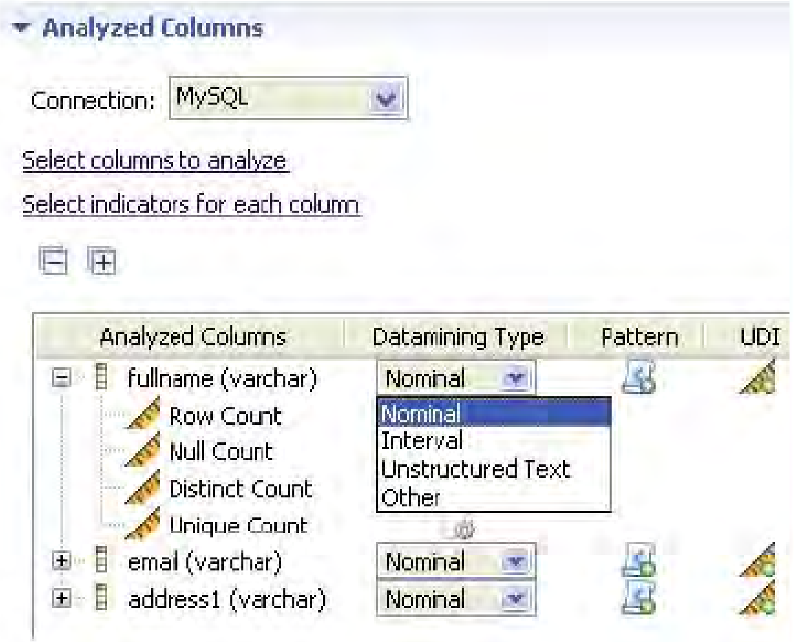


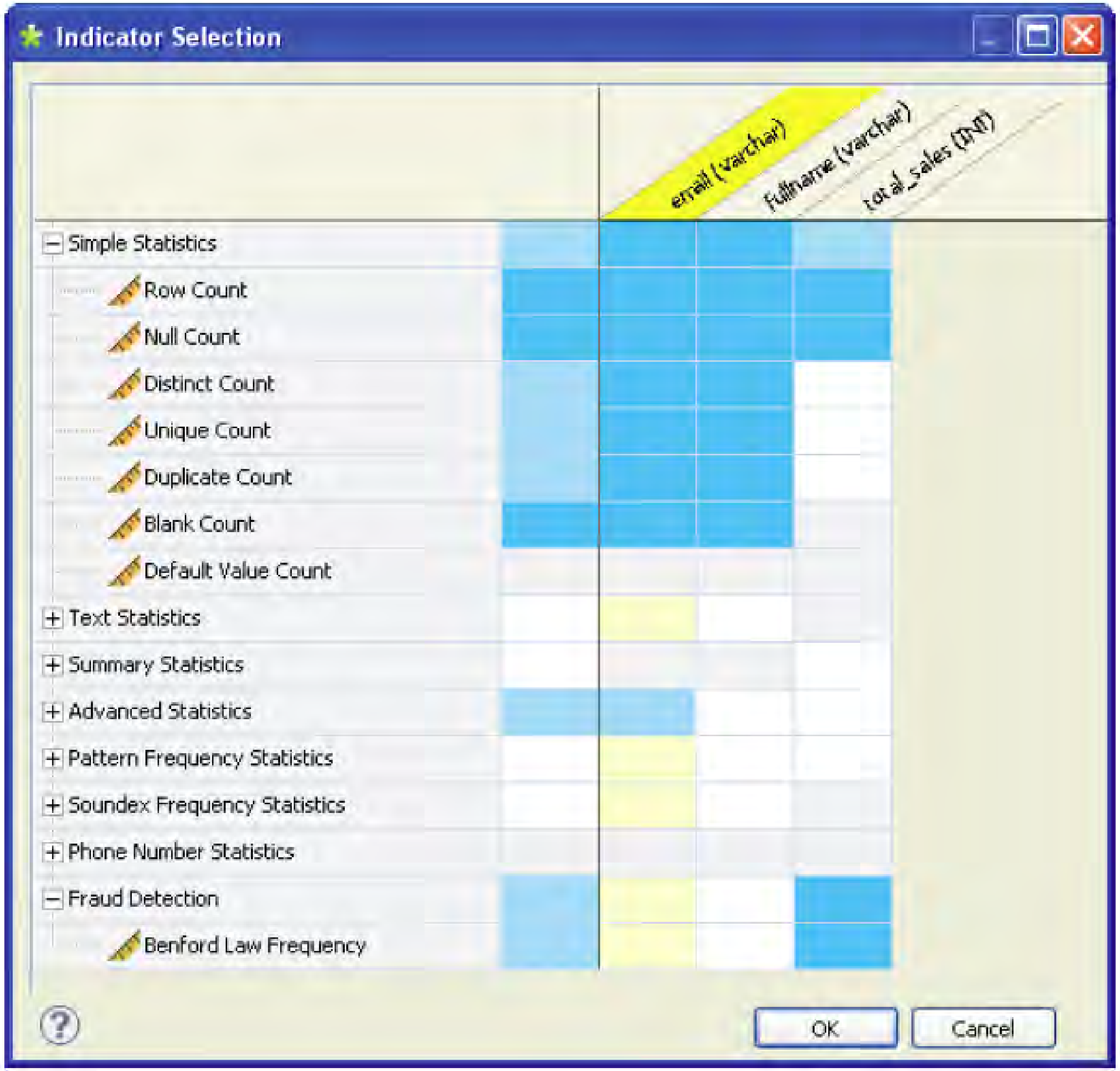
Populations of all 237 countries

Typical Shortcomings of Tools

- Usability
  - Complex to configure
  - Results complex to view and interpret
- Scalability
  - Main-memory based
  - SQL based
- Efficiency
  - Coffee, Lunch, Overnight
- Functionality
  - Restricted to simplest tasks
  - Restricted to individual columns or small column sets
  - „Checking“ vs. „discovery“

Screenshots from Talend

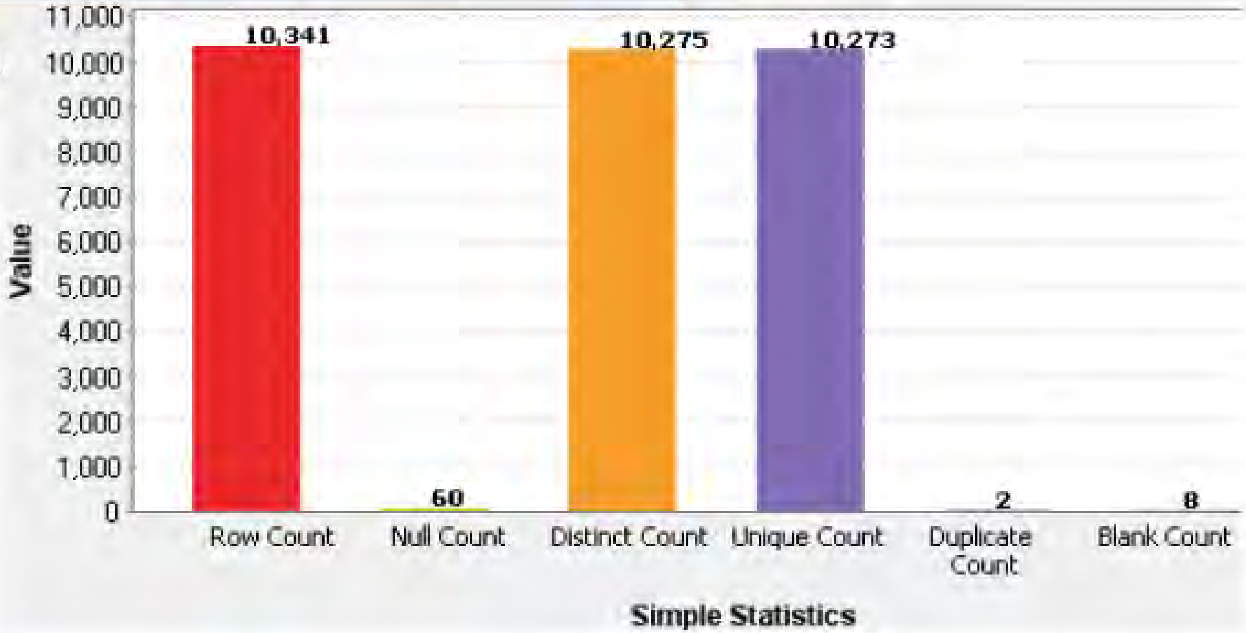




Column:customer.email

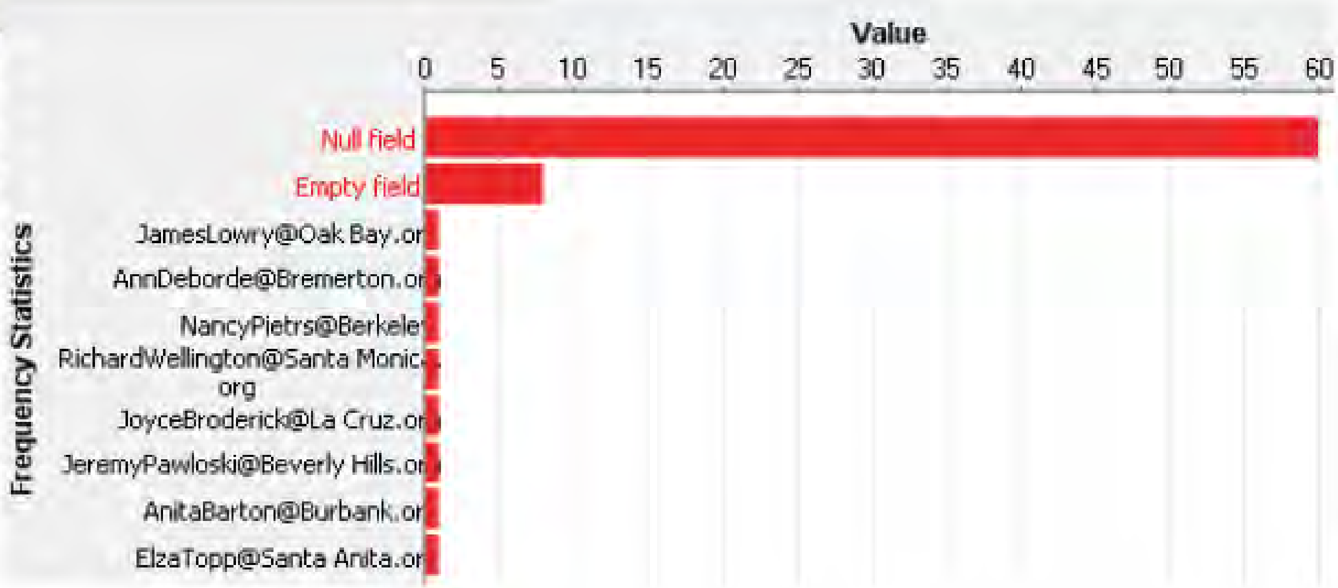
Simple Statistics

Label	Count	%
Row Count	10341.00	100.00%
Null Count	60.00	N/A
Distinct Count	10275.00	N/A
Unique Count	10273.00	N/A
Duplicate Count	2.00	N/A
Blank Count	8.00	N/A



Frequency Statistics

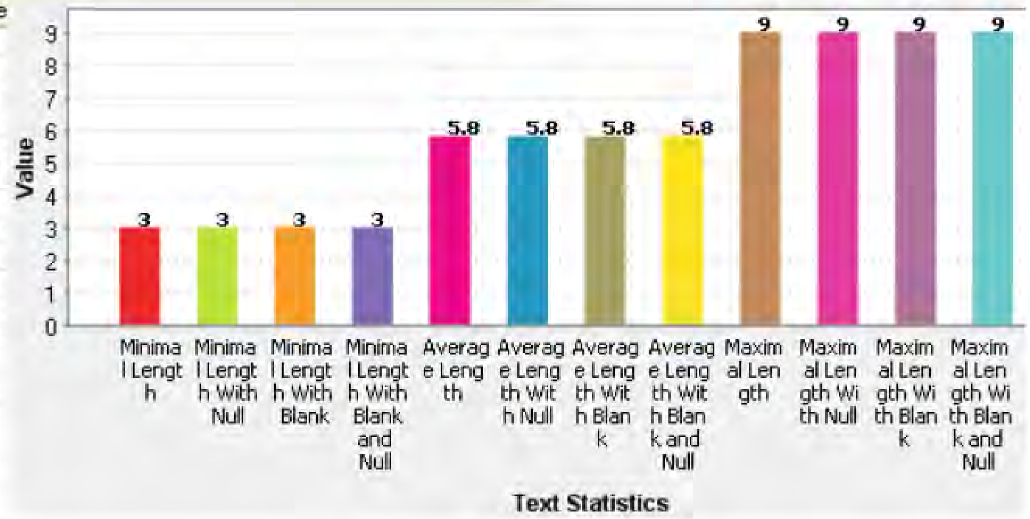
value	count	%
Null field	60.00	N/A
Empty field	8.00	N/A
JamesLowry@Oak Bay.org	1.00	N/A
AnnDeborde@Bremerton.org	1.00	N/A
NancyPietrs@Berkeley	1.00	N/A
RichardWellington@Santa Monica.org	1.00	N/A
JoyceBroderick@La Cruz.org	1.00	N/A
JeremyPawloski@Beverly Hills.org	1.00	N/A
AnitaBarton@Burbank.org	1.00	N/A
ElzaTopp@Santa Anita.org	1.00	N/A



Screenshots from Talend

- Column:metadata.firstname
  - Pattern Matching
  - Simple Statistics
  - Soundex Frequency Table
  - Text Statistics

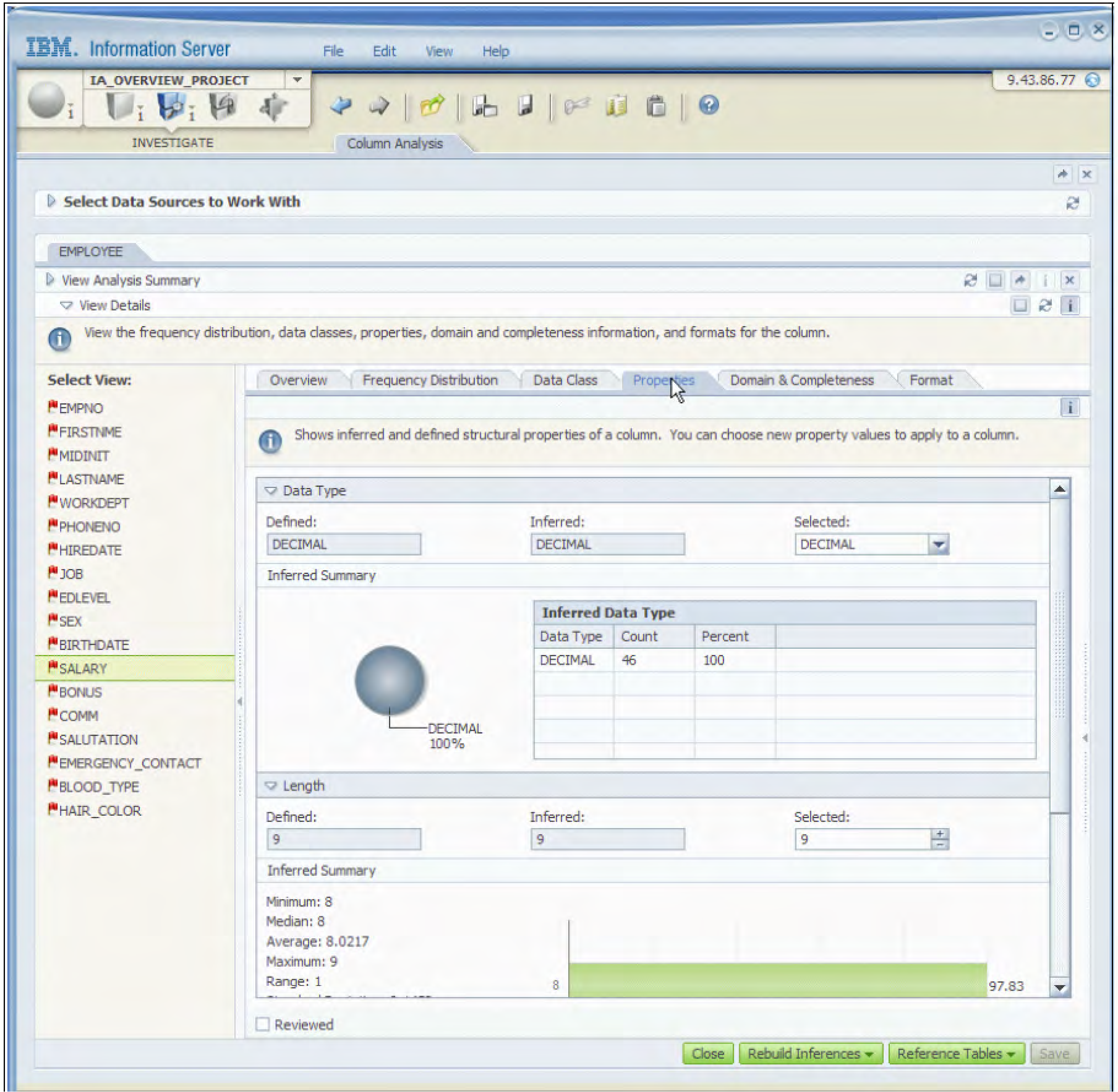
Label	Value
Minimal Length	3.00
Minimal Length With Null	3.00
Minimal Length With Blank	3.00
Minimal Length With Blank and Null	3.00
Average Length	5.80
Average Length With Null	5.80
Average Length With Blank	5.80
Average Length With Blank and Null	5.80
Maximal Length	9.00
Maximal Length With Null	9.00
Maximal Length With Blank	9.00
Maximal Length With Blank and Null	9.00



- Soundex Frequency Table

value	distinct count	count	%
Kylie	4.00	6	N/A
Brent	2.00	3	N/A
Anne	2.00	2	N/A
Aline	2.00	2	N/A
Cole	2.00	2	N/A
Lunea	2.00	2	N/A
Signe	2.00	2	N/A
Colby	2.00	3	N/A
Curran	2.00	2	N/A
Jescie	2.00	2	N/A

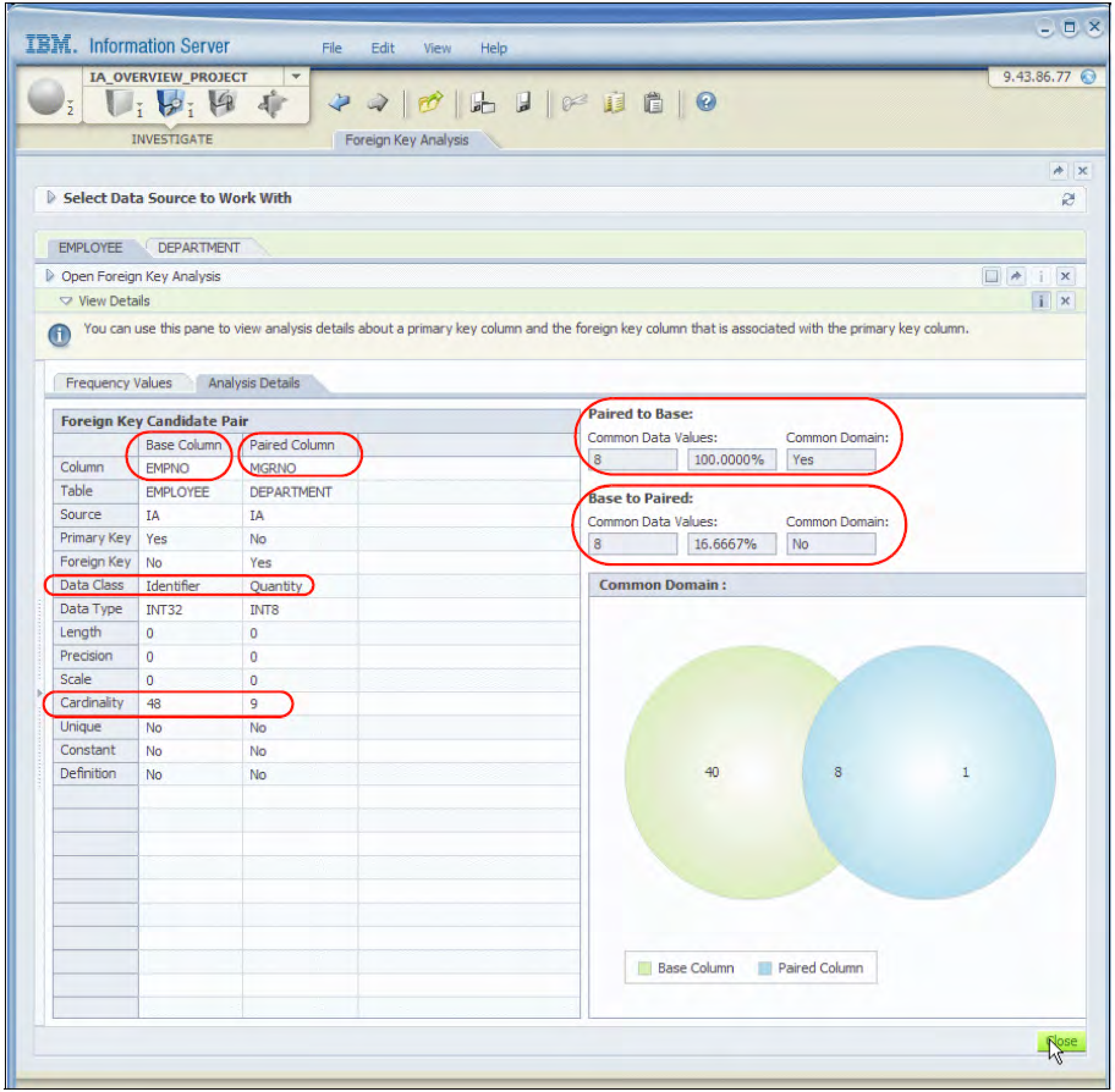
Screenshots for IBM Information Analyzer



F

Screenshots for IBM Information Analyzer

Hasso Plattner Institut



Overview

- Profiling tasks
- Profiling tools
- Visualization
- Next generation profiling
- Profiling challenges

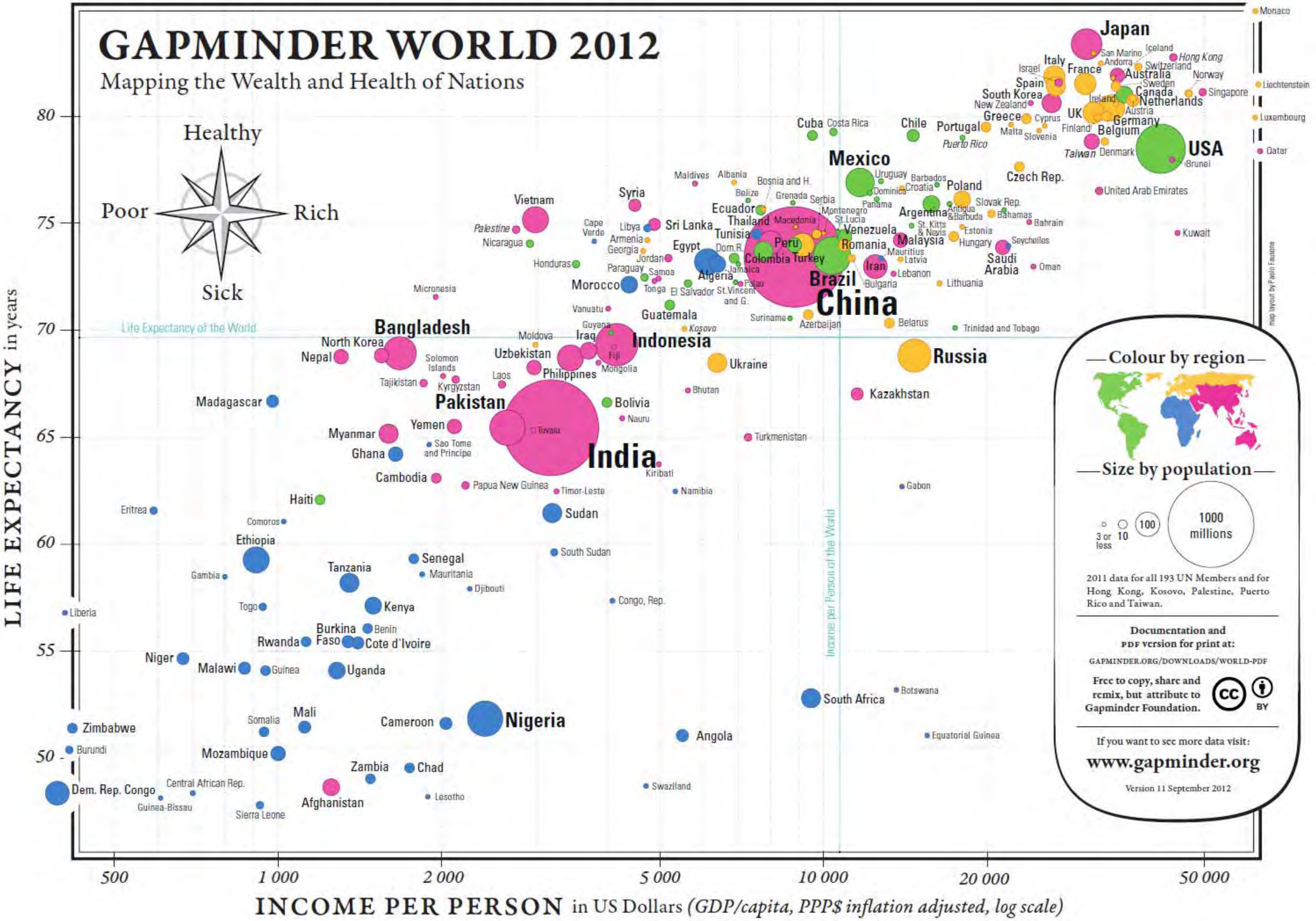
Motivation

- Human in the loop for data profiling and data cleansing.
- Advanced visualization technqiues
  - Beyond bar- and pie-charts
- Interactive visualization
  - Support users in visualizing data, profiling results
  - Support any action taken upon the results
    - Cleansing, sorting, ...
- Further reading: http://vgc.poly.edu/~juliana/courses/cs9223/Lectures/intro-to-visualization.pdf

Massive screens for massive data

- Powerwall, Uni Konstanz
- 5.20 m x 2.15 m; fast neun Millionen Bildpunkte

http://www.informatik.uni-konstanz.de/forschung/powerwall/application-pictures/

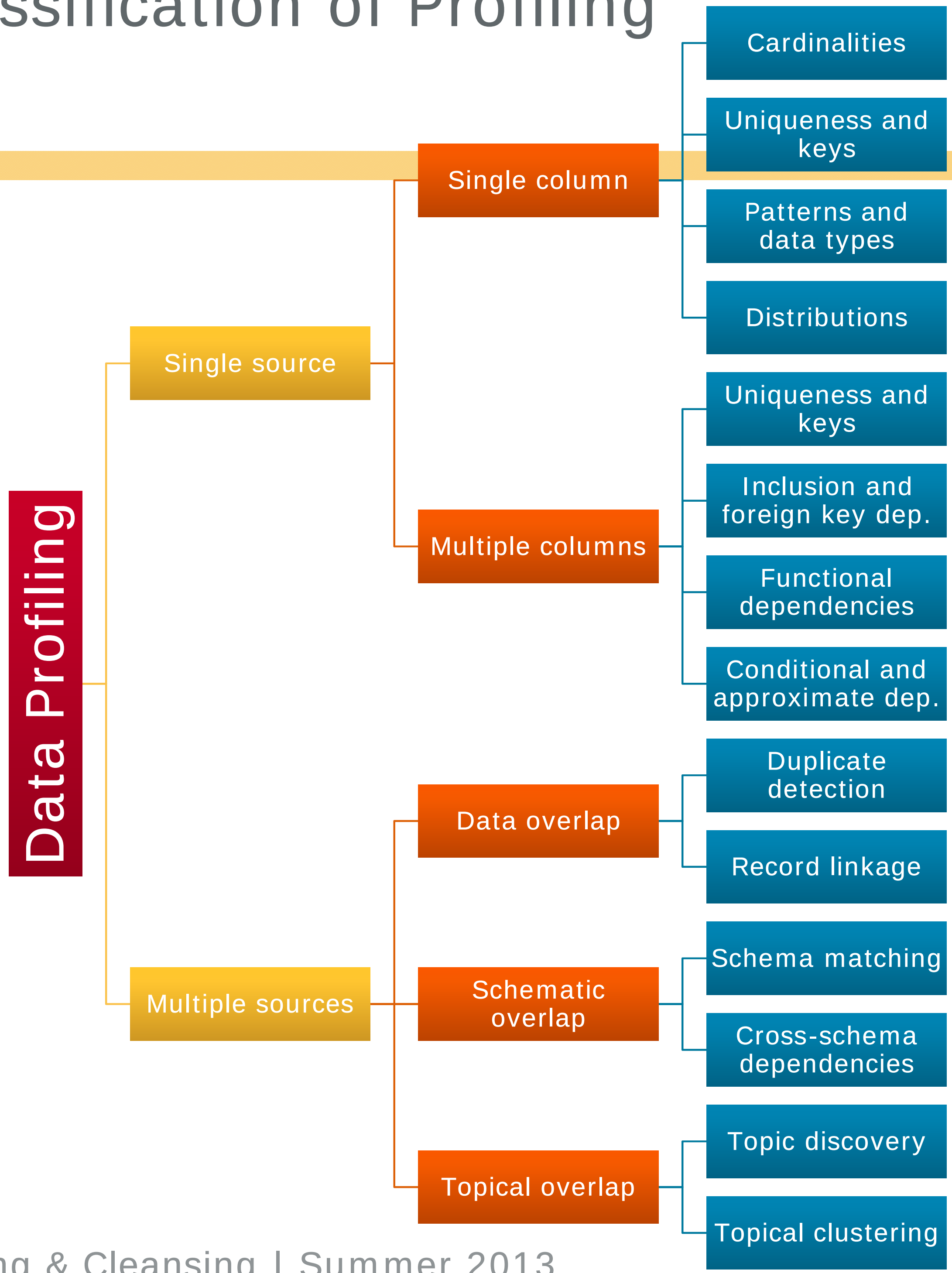


http://www.gapminder.org/GapminderMedia/wp-uploads/Gapminder-World-2012.pdf

Overview

- Profiling tasks
- Profiling tools
- Visualization
- Next generation profiling
- Profiling challenges

Extended Classification of Profiling Tasks



Schema Matching

- Automatically determine cross-schema value correspondences between attributes

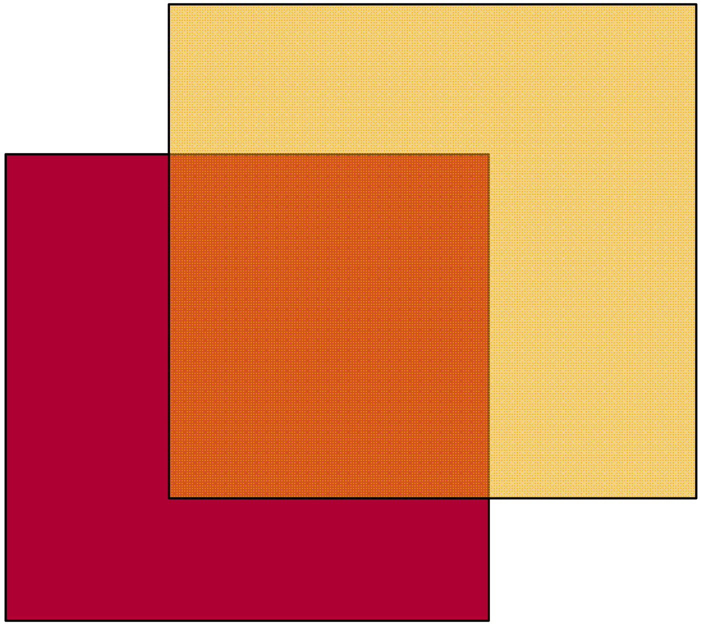
- Traditionally: Input for data transformation and exchange tasks
- For data profiling
  - Determine *closeness* of two schemata
  - Determine „schema fit“
    - Complement or union

- Detect multiple (different) representations of the same real-world entity

- Traditionally: Input for data cleansing and data fusion tasks
- For data profiling
  - Intra-source: Determine duplicity/cleanliness
  - Inter-sources: Determine „data fit“
    - Complement or union

Together: 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

Cross-Schema Dependencies

- Inclusion dependencies across schemata
  - Join paths betwen data sources
- Conditional INDs
  - Typical pattern among crossreferencing sources

Catalog

Unit cost	DBName	ProdID
200 USD	ToyDB	17
50 EUR	ToyDB	18
1000 QAR	FashionDB	18

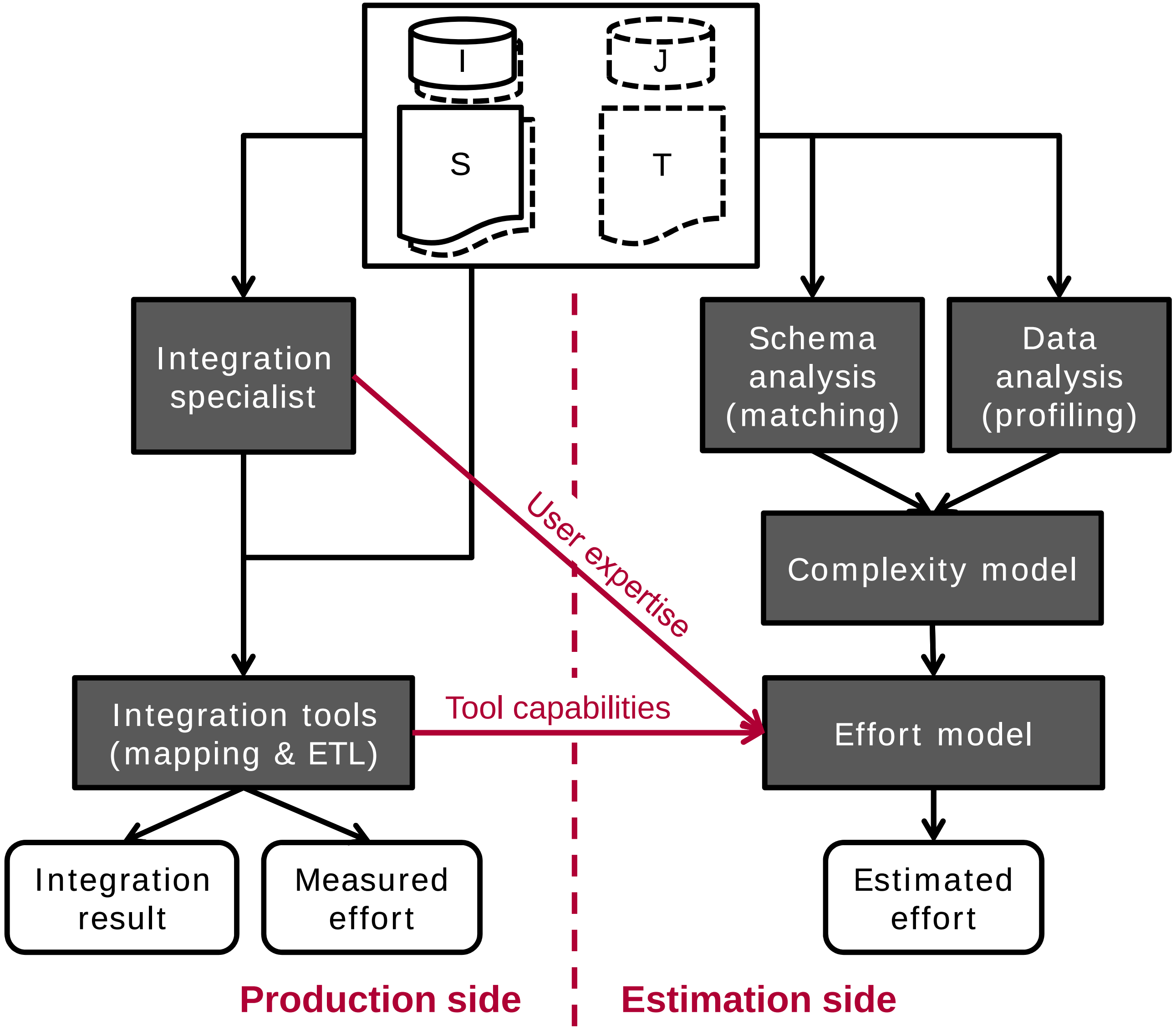
ToyDB

EntityID	further data
17	abcd...
18	efgh...

FashionDB

EntityID	further data
18	abcd...
19	efgh...

Integration effort estimation



Topic discovery and clustering

- What is a data set **about**?
  - Domain(s), topics, entity types
- Single-topic datasets
  - GraceDB, IMDB, Sports databases, etc.
- Multi-topic datasets
  - General purpose knowledge bases: YAGO, Dbpedia
  - Tables crawled from the Web

- What is a topic?
  - Wikipedia categories?
- Topical clustering for
  - Source selection, query processing

Wikipedia Categories

Kategorien: Robotikhersteller | Volkswagen | Unternehmen (Wolfsburg) | Automobilhersteller | Motorenhersteller | Mitglied der Europäischen Bewegung Deutschland

Kategorien: Amtierender Außenminister | Außenminister (Bundesrepublik Deutschland) | Vizekanzler (Deutschland) | Bundestagsabgeordneter | Bundesvorsitzender der FDP | Vorsitzender der FDP-Bundestagsfraktion | Generalsekretär der FDP | Bundesvorsitzender der Jungen Liberalen | Kanzlerkandidat (Deutschland) | Ritter des Ordens wider den tierischen Ernst | Deutscher | Geboren 1961 | Mann

Categories: Companies listed on the New York Stock Exchange | Dow Jones Industrial Average | IBM | Cloud computing providers | Cloud computing vendors | Companies based in Westchester County, New York | Companies established in 1896 | Computer companies of the United States | Computer hardware companies | Computer storage companies | Display technology companies | Electronics companies of the United States | Initialisms | Multinational companies | National Medal of Technology recipients | Point of sale companies | Semiconductor companies | Software companies of the United States | UML Partners

Categories: Ronald Reagan | 1911 births | 2004 deaths | 20th-century presidents of the United States | Actors awarded British knighthoods | Actors from Illinois | American actor-politicians | American anti-communists | American Disciples of Christ | American film actors | American labor leaders | American military personnel of World War II | American politicians of Irish descent | American Presbyterians | American radio personalities | American shooting survivors | American television personalities | Actors from California | California Republicans | Chicago Cubs broadcasters | Cold War leaders | College football announcers | Colorectal cancer survivors | Congressional Gold Medal recipients | Conservatism | Deaths from Alzheimer's disease | Deaths from pneumonia | Delegates to the Republican National Convention | Eureka College | First Motion Picture Unit personnel | Governors of California | History of the United States (1980–1991) | Infectious disease deaths in California | Iran–Contra affair | Knights Grand Cross of the Order of the Bath | Major League Baseball announcers | New Right (United States) | People from Lee County, Illinois | People from Whiteside County, Illinois | People from the Greater Los Angeles Area | People involved in the Soviet war in Afghanistan | Presidential Medal of Freedom recipients | Presidents of the United States | Reagan family | Recipients of Honorary British Knighthoods | Recipients of the Order of the White Eagle (Poland) | Republican Party (United States) presidential nominees | Restoration Movement | Skin cancer survivors | Time magazine Persons of the Year | United States Army Air Forces officers | United States Army officers | United States presidential candidates, 1968 | United States presidential candidates, 1976 | United States presidential candidates, 1980 | United States presidential candidates, 1984

Overview

- Profiling tasks
- Profiling tools
- Visualization
- Next generation profiling
- Profiling challenges

Profiling Challenges

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

Each of these is worth one (or more) master's theses!

Scalable profiling

- Scalable in number of rows
- Scalable in number of columns
  - Number of column combinations is exponential
  - Small table with 100 columns: $2^{100} - 1$ = 1,267,650,600,228,229,401,496,703,205,376 = 1.3 nonillion combinations
    - Impossible to check
    - Impossible even to enumerate
- Solutions
  - Scale up: More RAM, faster CPUs
    - Expensive
  - Scale in: More cores
    - More complex (threads)
  - Scale out: More machines
    - Communication overhead

Holistic Profiling

- Various profiling methods for various profiling tasks
- Commonalities
  - 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

- Idea: 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

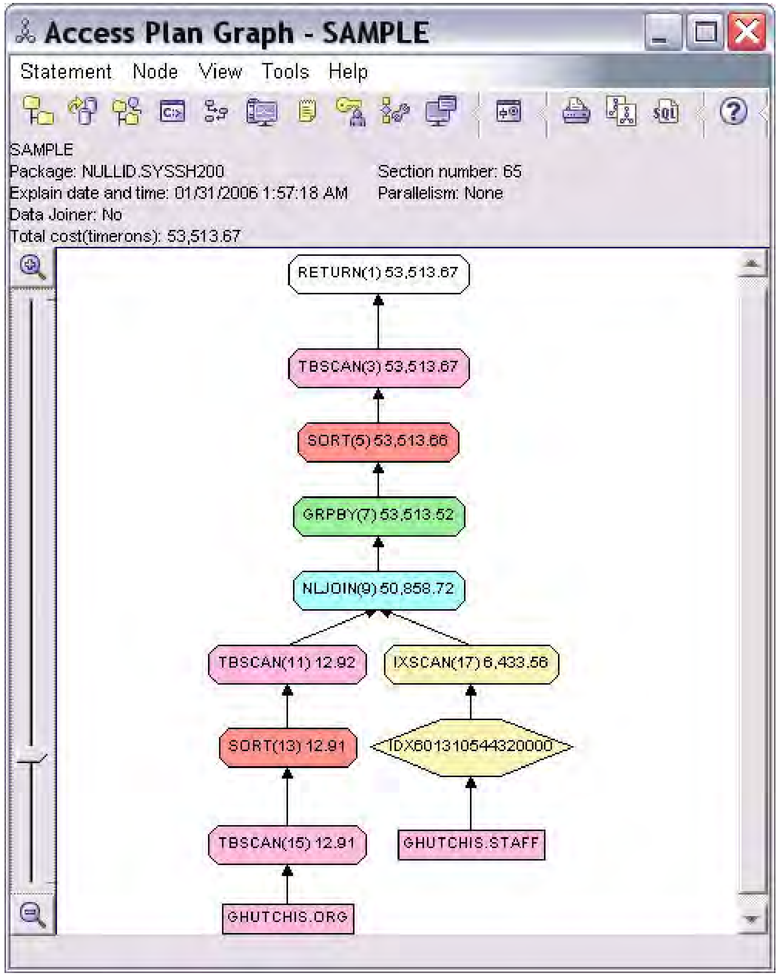
- Profiling is long procedure
  - Boring for developer
  - Expensive for machines (I/O and CPU)

- Problem: Display intermediate results
  - Of improving/converging accuracy
  - Allow early abortion of profiling run

- Gear algorithms toward that goal
  - Allow intermediate output
  - Enable early output (progressive profiling)

Profiling query results

- Query results are boring: Spruce them up with some metadata
  - Usually only: Row count
  - For each column, give some statistics
    - Uniqueness, histogram, AVG, etc.
    - Show FDs



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

Profiling new types of data

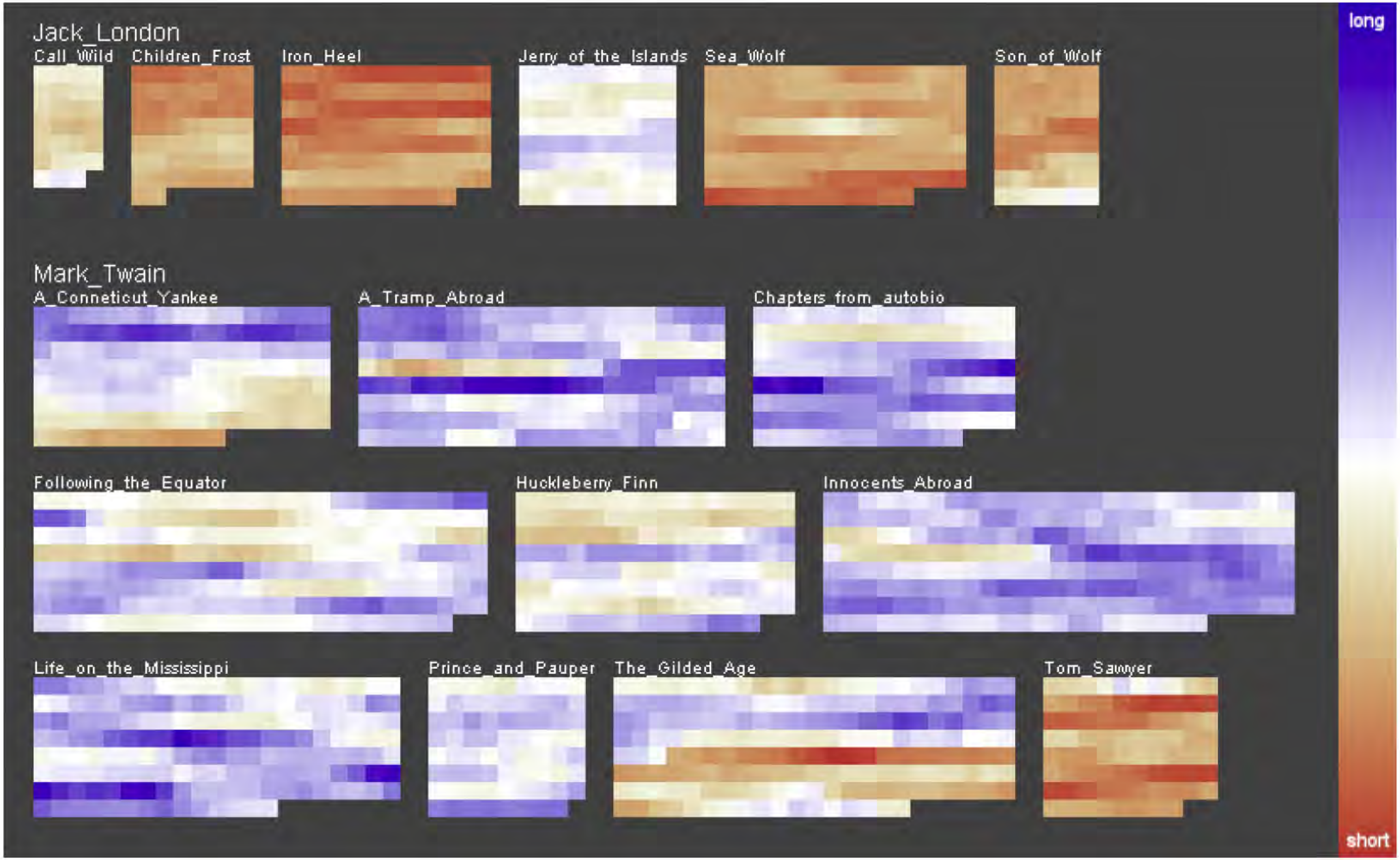
- Traditional data profiling: Single table or multiple tables
- More and more data in other models
  - XML / nested relational / JSON
  - RDF triples
  - Textual data
    - Blogs, Tweets, News
    - In tables (CLOBS): Produkt descriptions, CVs, reviews
  - Multimedia
- All models offer new dimensions to profile
  - Nestedness, measures at different nesting levels
  - Graph structure, in- and outdegrees
  - Sentiment, sentence structure, complexity, and other linguistic measures
  - Color, video-length, volume, etc.

Example: Text profiling

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

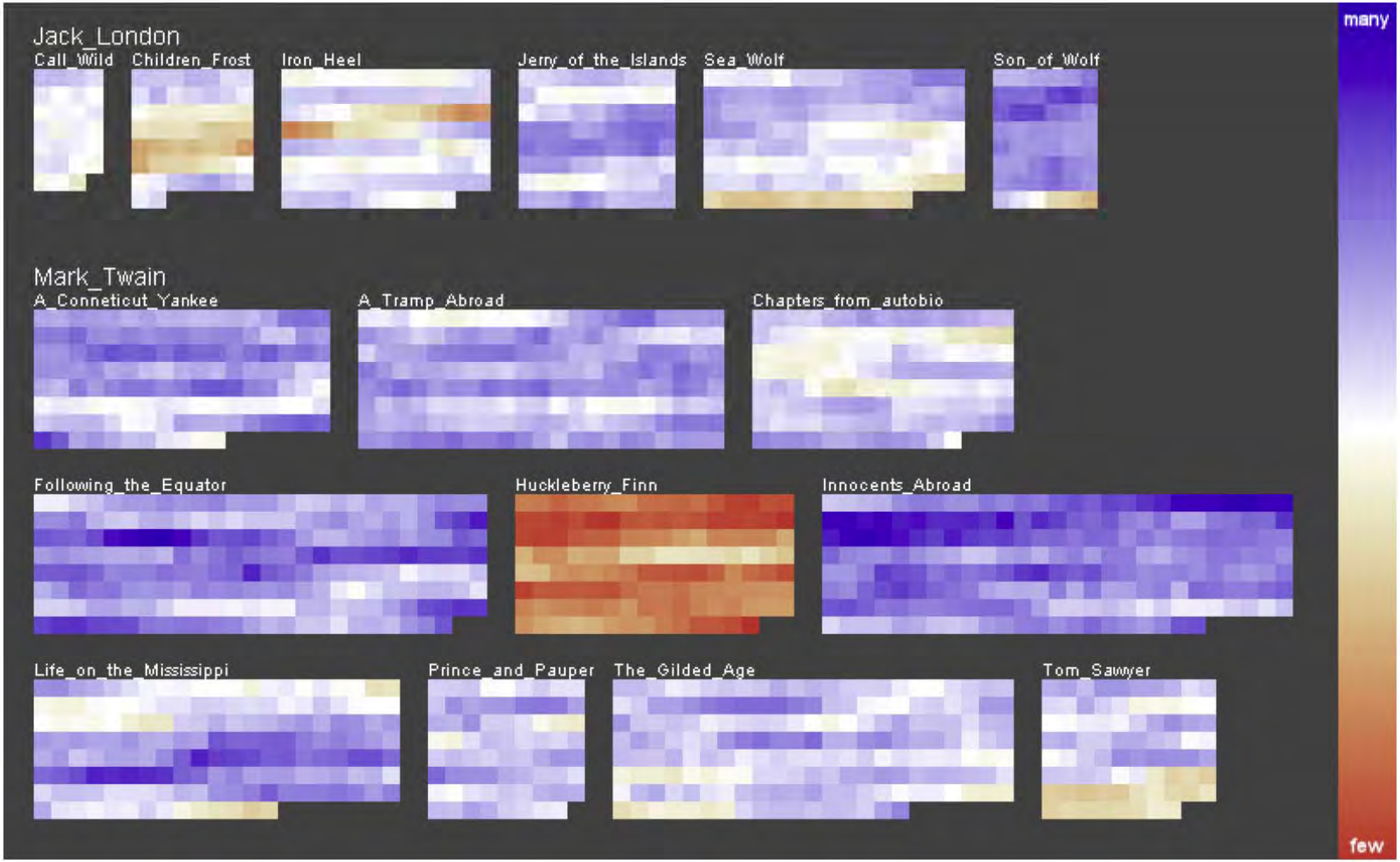
- Idea and following figures based on:
  - „Literature Fingerprinting: A New Method for Visual Literary Analysis" by Daniel A. Keim and Daniela Oelke (IEEE Symposium on Visual Analytics Science and Technology 2007)

Average sentence length



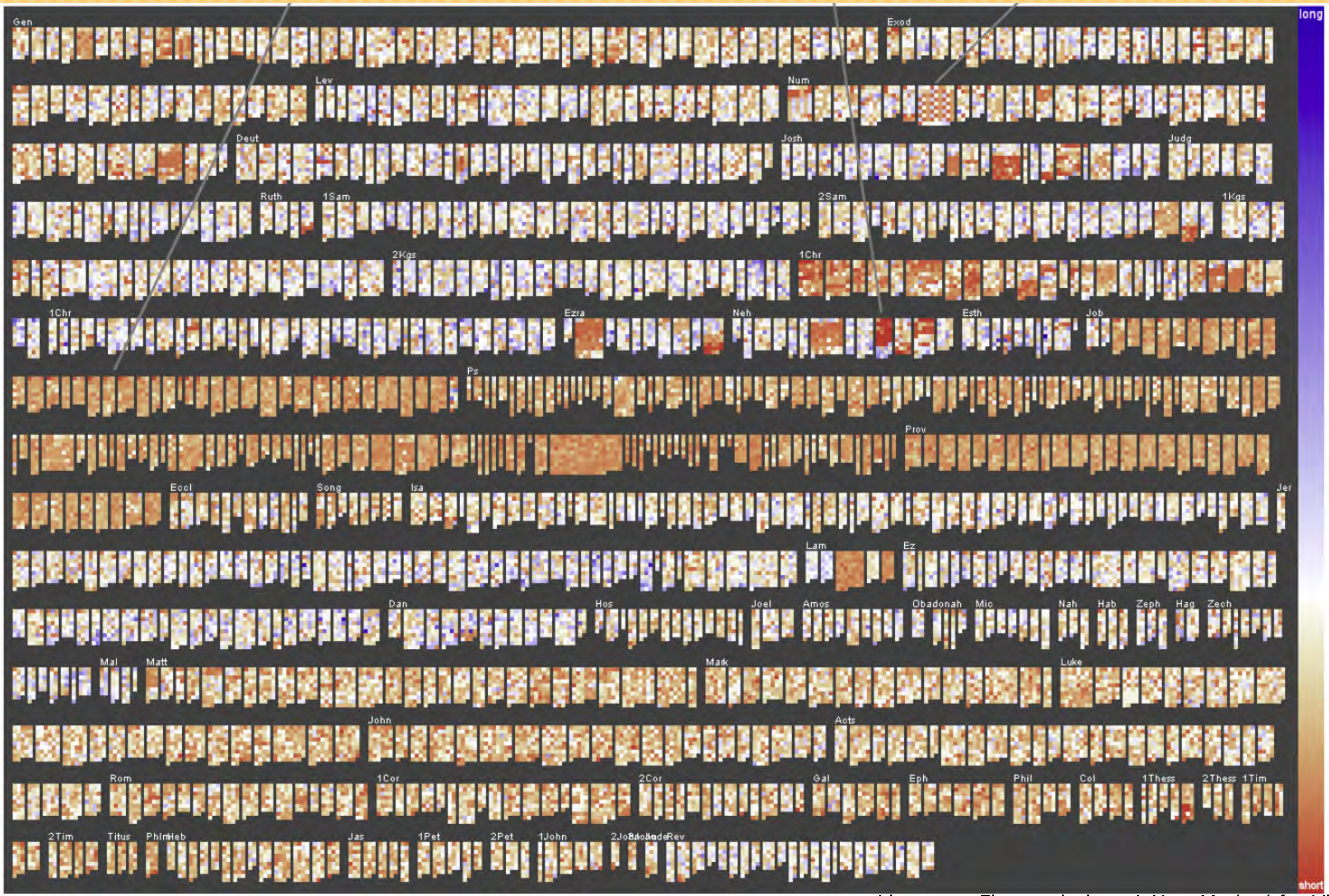
„Literature Fingerprinting: A New Method for Visual Literary Analysis" by Daniel A. Keim and Daniela Oelke

Hapax Legomena



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

Verse length



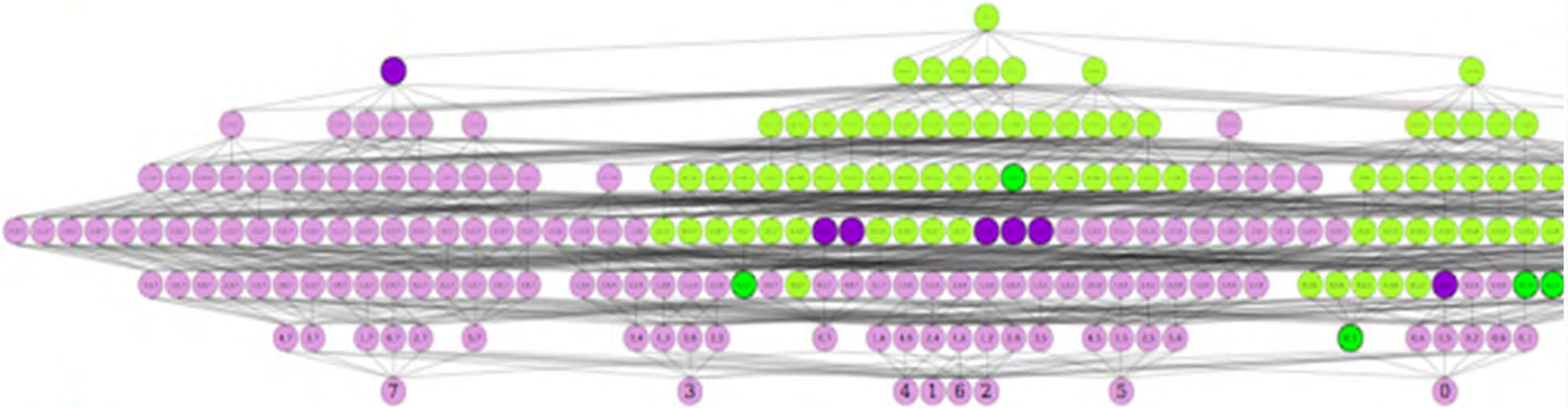
„Literature Fingerprinting: A New Method for Visual Literary Analysis" by Daniel A. Keim and Daniela Oelke

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
  - Distibutions, patterns
  - Example: TPCH (see next slide)
- Problem: Consistently produce same, randomized data

TPCH – Uniques and Non-Uniques

- Using the first 8 columns of the lineitems table



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

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