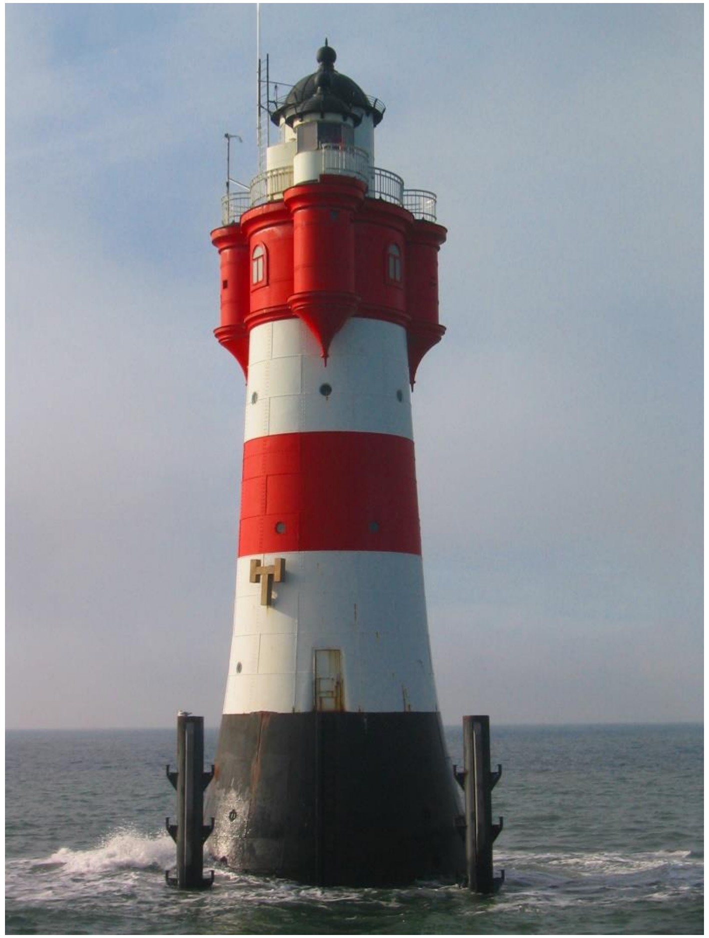
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Big Data and Data Profiling
Introduction

20.4.2017
Felix Naumann

Overview

1. **Introduction to research group**
2. Lecture organisation
3. (Big) data
   - Data sources
   - Profiling
4. Overview of semester



Information Systems Team

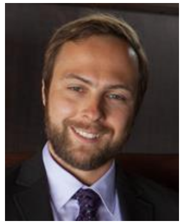

Diana Stephan


Prof. Felix Naumann


Sebastian Kruse


Tim Repke


Thorsten Papenbrock


Dr. Ralf Krestel


Tobias Bleifuß


Hazar Harmouch


John Koumarelas


Toni Grütze


Michael Loster


Ahmad Samiei


Zhe Zuo


Konstantina Lazariduo


Lan Jiang

project DuDe
Duplicate Detection
Data Change
Data Fusion
project Stratosphere
Entity Search
Data Profiling
Information Integration
Web Science
Data Scrubbing
project DataChEx
Data as a Service
Data Cleansing
Information Quality
Text Mining
Web Data
Linked Open Data
RDF Data Mining
Dependency Detection
ETL Management
Service-Oriented Systems
Entity Recognition
Opinion Mining
project Metanome

Other courses in this semester

- Lectures
  - DBS I (Bachelor)
  - Data Profiling
- Seminars
  - Bachelor: Text Mining
  - Master: Recommender Systems
- Bachelorproject
  - Ingestion – Commerzbank
- Masterproject
  - Hate Speech Detection

Extending the Database Relational Model to Capture More Meaning

E. F. CODD
IBM Research Laboratory

During the last three or four years several investigators have been exploring "semantic models" for formatted databases. The intent is to capture (in a more or less formal way) more of the meaning of the data so that database design can become more systematic and the database system itself can behave more intelligently. Two major thrusts are clear:

(1) the search for meaningful units that are as small as possible—*atomic semantics*;
(2) the search for meaningful units that are larger than the usual *n*-ary relation—*molecular semantics*.

In this paper we propose extensions to the relational model to support certain atomic and molecular semantics. These extensions represent a synthesis of many ideas from the published work in semantic modeling plus the introduction of new rules for insertion, update, and deletion, as well as new algebraic operators.

Key Words and Phrases: relation, relational database, relational model, relational schema, database, data model, database schema, data semantics, semantic model, knowledge representation, knowledge base, conceptual model, conceptual schema, entity model
CR Categories: 3.70, 3.73, 4.22, 4.29, 4.33, 4.34, 4.39

1. INTRODUCTION

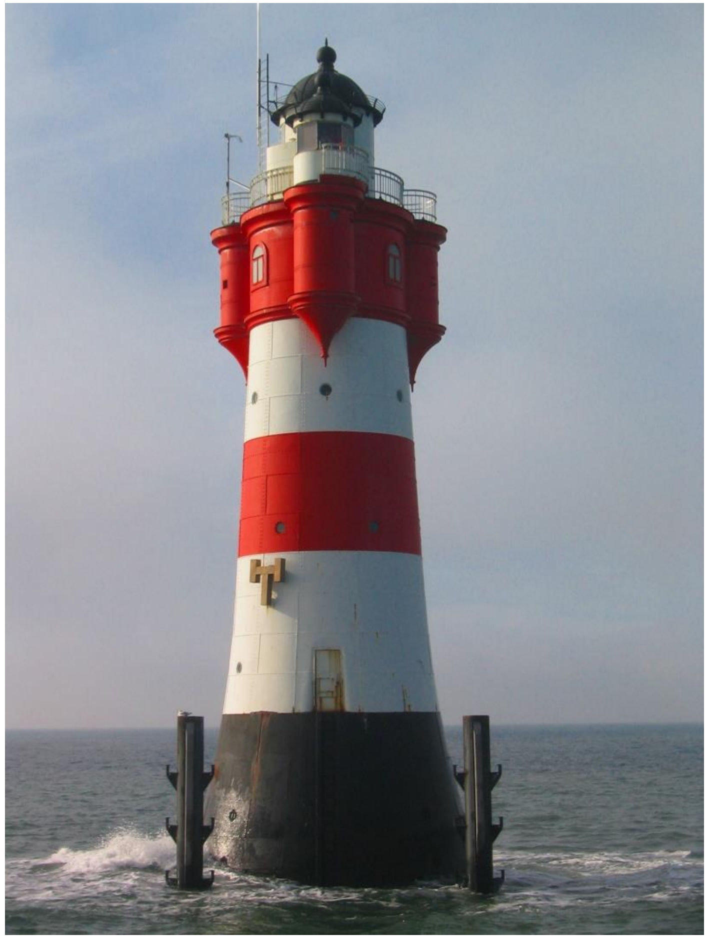
The relational model for formatted databases [5] was conceived ten years ago, primarily as a tool to free users from the frustrations of having to deal with the clutter of storage representation details. This implementation independence coupled with the power of the algebraic operators on *n*-ary relations and the open questions concerning dependencies (functional, multivalued, and join) within and between relations have stimulated research in database management (see [30]). The relational model has also provided an architectural focus for the design of databases and some general-purpose database management systems such as MACAIMS [13], PRTV [38], RDMS(GM) [41], MAGNUM [19], INGRES [37], QBE [46], and System R [2].
During the last few years numerous investigations have been aimed at capturing

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
A version of this work was presented at the 1979 International Conference on Management of Data (SIGMOD), Boston, Mass., May 30–June 1, 1979.
Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95193.
© 1979 ACM 0362-5915/79/1200-0397 $00.75

ACM Transactions on Database Systems, Vol. 4, No. 4, December 1979, Pages 397–434.

Overview

1. Introduction to research group
2. **Lecture organisation**
3. (Big) data
   - Data sources
   - Profiling
4. Overview of semester



Dates and exercises

- Lectures
  - Mondays 15:15 – 16:45
  - Thursdays 13:30 – 15:00
- Exercises
  - In parallel
- First lecture
  - 20.4.2017
- Last lecture
  - 27.7.2017
- See Web for timetable updates!

- Exam
  - Oral or written exam
  - Probably first week after lectures
- Prerequisites
  - To participate
    - Background in databases and their implementation (e.g. DBS I and II)
  - For exam
    - Attend lectures
    - Active participation in exercises
    - “Successfully” complete exercise tasks

Feedback

- Evaluation at end of semester

- Question any time please!
  - During lectures
  - During consultation: Tuesdays 13-15
  - Email: naumann@hpi.de

- Also: Give feedback about
  - improving lectures
  - informational material
  - organization

Literature

- No single textbook
- References to various papers during lecture
- All papers are available either via email from me or (preferred) from
  - Google Scholar: http://scholar.google.com/
  - DBLP: http://www.informatik.uni-trier.de/~ley/db/index.html
  - CiteSeer: http://citeseer.ist.psu.edu/
  - ACM Digital Library: www.acm.org/dl/
  - Homepages of authors

- Profiling relational data: a survey. Ziawasch Abedjan, Lukasz Golab, Felix Naumann, VLDB Journal, vol. 24(4):557-581 2015
  - https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2015/dataprofiling_main.pdf

Exercise

- Algorithm design and programming exercises
  - Data profiling (emphasis on efficiency and scalability)
  - Unique column combinations
  - Inclusion dependencies
  - Functional dependencies

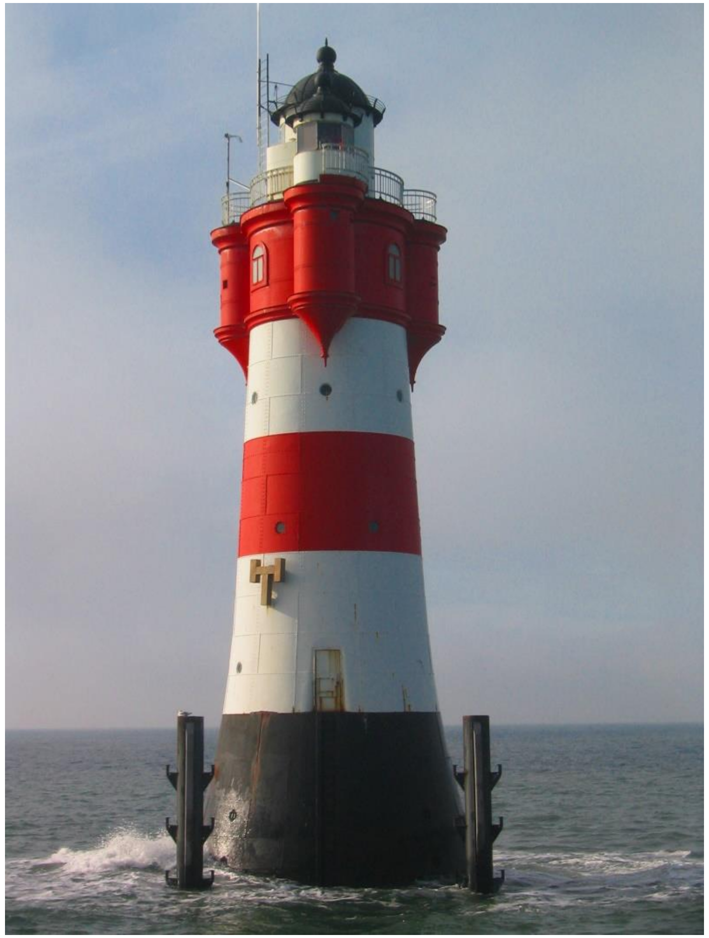
- Self-motivation wrt. good solutions!

Introduction: Audience

- Which semester?
- HPI or UP?
- Erasmus o.ä.?
  - English?
- Database knowledge?
  - Which other related lectures?
- Your motivation?

Overview

1. Introduction to research group
2. Lecture organisation
3. **(Big) data**
   - Data sources
   - Profiling
4. Overview of semester

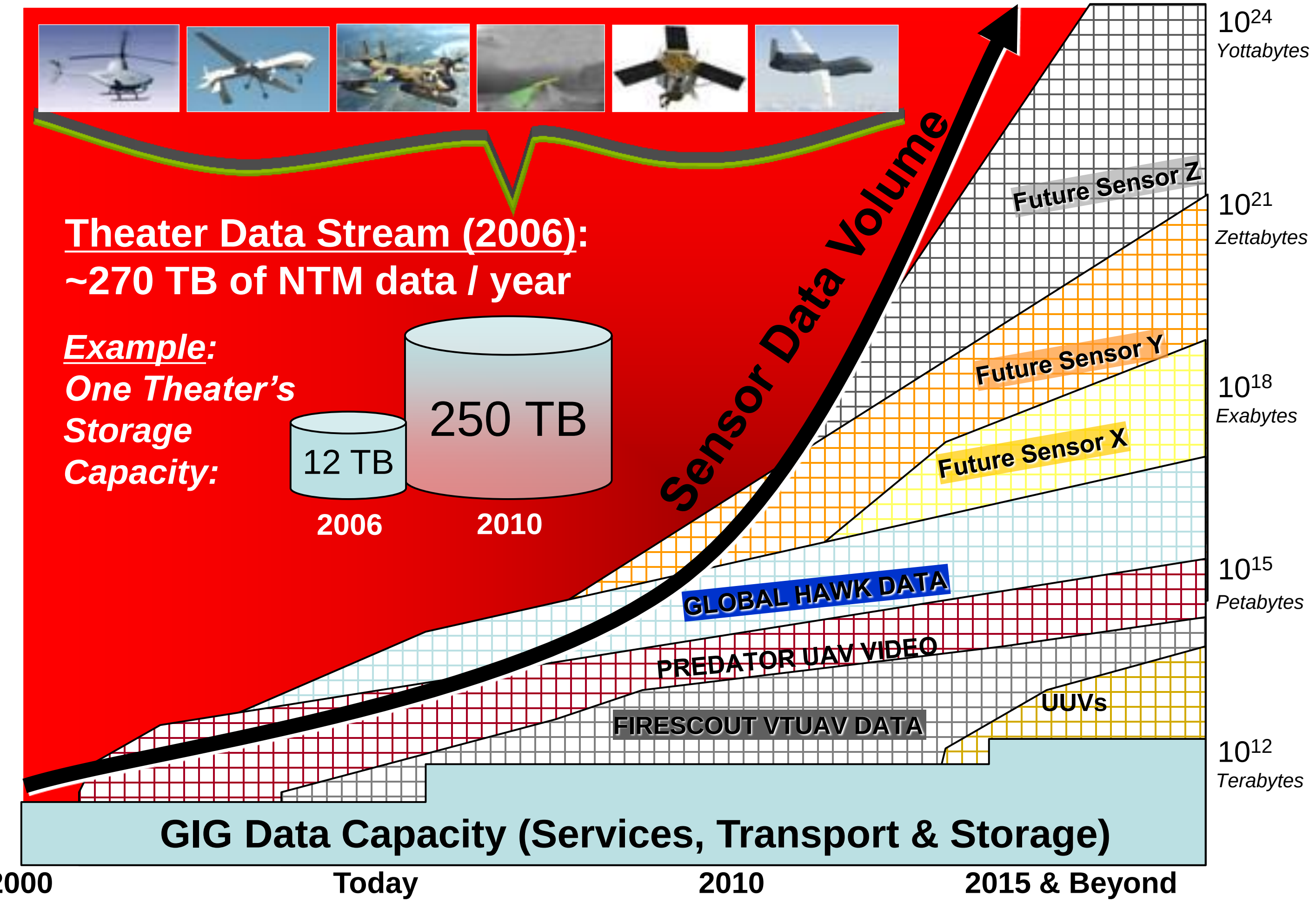


Big Data Motivation

We're now entering what I call the "Industrial Revolution of Data," where the majority of data will be stamped out by machines: software logs, cameras, microphones, RFID readers, wireless sensor networks and so on.
These machines generate data a lot faster than people can, and their production rates will grow exponentially with Moore's Law. Storing this data is cheap, and it can be mined for valuable information.

- Joe Hellerstein
  http://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/

Military Projection of Sensor Data Volume (later refuted)

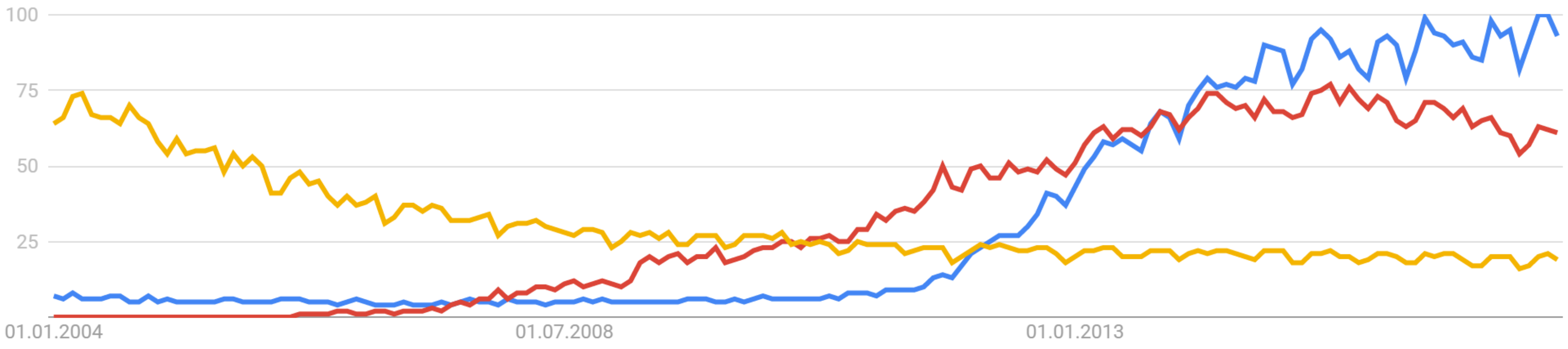


Using 1TB drives, this would require 1 trillion (10^{12}) drives!

Bob Gourley: Thoughts on the future of Information Sharing Technology

HPI Hasso Plattner Institut

Big Data trends



data warehouse

big data

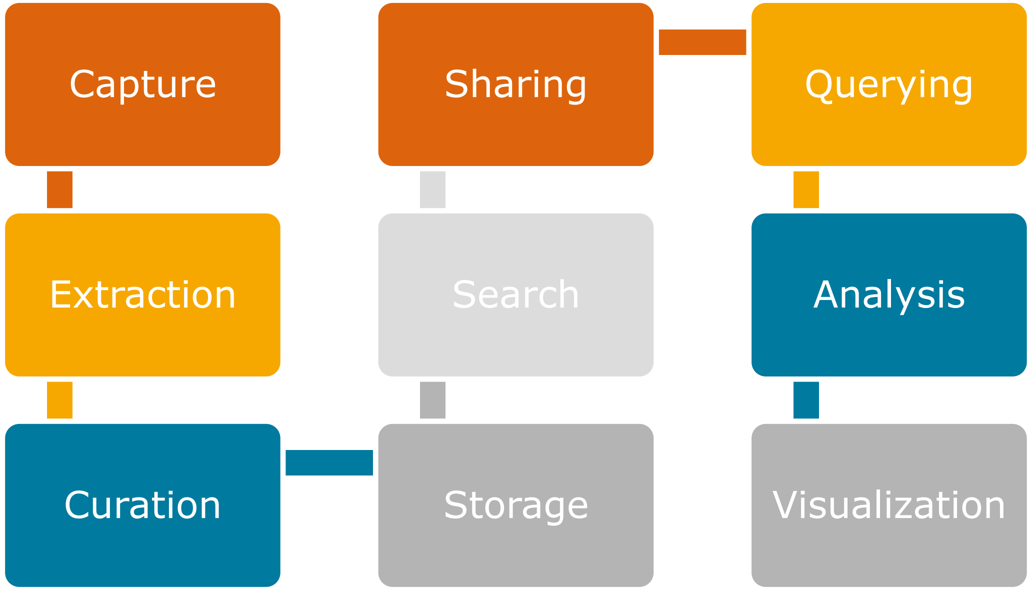
hadoop

Felix Naumann
Data Profiling
Summer 2017

Quelle: Google trends, Apr. 2017

Defining Big Data

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



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

Gartner's 3 (+ 1) V's – Properties of Big Data

- **Volume**
  - 12 terabytes of Tweets: product sentiment analysis
  - 350 billion annual meter readings: predict power consumption
- **Velocity**
  - 5 million daily trade events: identify potential fraud
  - 500 million daily call detail records: predict customer churn faster
- **Variety**
  - 100's of live video feeds from surveillance cameras
  - 80% data growth in images, video and documents to improve customer satisfaction
- **Veracity** (Wahrhaftigkeit)
  - 1 in 3 business leaders don't trust the information they use to make decisions.

http://www.ibm.com/software/data/bigdata/

More V‘s

- **Viscosity**
  - Integration and dataflow friction
- **Venue**
  - Different locations that require different access & extraction methods
- **Vocabulary**
  - Different language and vocabulary
- **Value**
  - Added-value of data to organization and use-case
- **Virality**
  - Speed of dispersal among community
- **Variability**
  - Data, formats, schema, semantics change

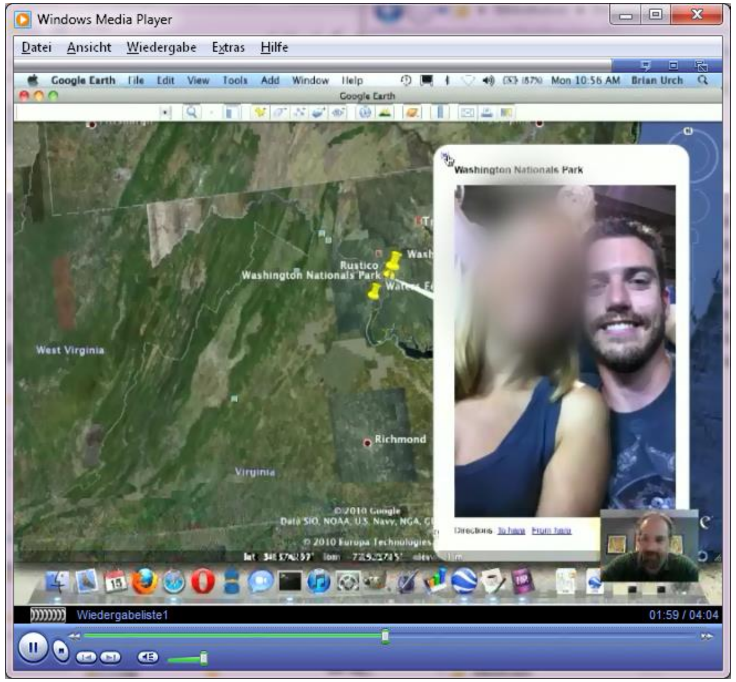
Big and Small

- Big Data can be very small
  - Streaming data from aircraft sensors
  - Hundred thousand sensors on an aircraft is “big data”
  - Each producing an eight byte reading every second
  - Less than 3GB of data in an hour of flying
    - (100,000 sensors x 60 minutes x 60 seconds x 8 bytes).
- Not all large datasets are “big”.
  - Video streams plus metadata
  - Telco calls and internet connections
  - Can be parsed extremely quickly if content is well structured.
  - From http://mike2.openmethodology.org/wiki/Big_Data_Definition
- The task at hand makes data “big”.

„Big data" in business

- Has been used to sell more hardware and software.
- Has become a shallow buzzword.

- But: The actual big data is there, has added-value, and can be used effectively.
  - Data mining
  - Marketing / advertising
  - Collaborative filtering
  - Raytheon's RIOT software
  - NSA, etc.
  - Kreditech, Lenddo, Klout, …
  - …

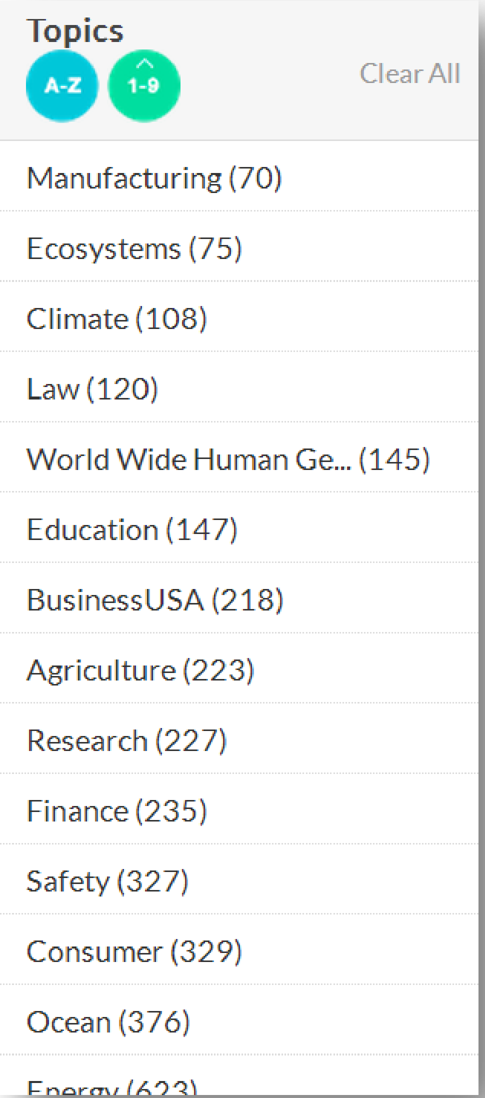


„Big data“ in business

- Amazon.com
  - Millions of back-end operations every day
  - Catalog, searches, clicks, wish lists, shopping carts, third-party sellers, …
- Walmart
  - > 1 million customer transactions per hour
  - 2.5 petabytes (2560 terabytes)
- Facebook
  - 250 PB, 600TB added daily (2013)
  - 1 billion photos on one day (Halloween)
- FICO Credit Card Fraud Detection
  - Protects 2.1 billion active accounts

Big Government Data (USA)

- Big Data Research and Development Initiative
  - Explored how big data addresses important problems facing the government.
  - 84 different big data programs spread across six departments
- Data.gov
  - > 104.000 datasets
- Government owns six of the ten most powerful supercomputers in the world.
- NASA Center for Climate Simulation
  - 32 petabytes of climate observations and simulations



Examples from Wikipedia – Big Science

- Large Hadron Collider
  - 150 million sensors; 40 million deliveries per second
  - 600 million collisions per second
  - Theoretically: 500 exabytes per day (500 quintillion bytes)
  - Filtering: 100 collisions of interest per second
    - Reduction rate of 99.999% of these streams
  - 25 petabytes annual rate before replication (2012)
  - 200 petabytes after replication

Examples from Wikipedia - Science

- Sloan Digital Sky Survey (SDSS)
  - Began collecting astronomical data in 2000
  - Amassed more data in first few weeks than all data collected in the history of astronomy.
  - 200 GB per night
  - Stores 140 terabytes of information
  - Large Synoptic Survey Telescope, successor to SDSS
    - Online in 2016
    - Will acquire that amount of data every five days.
- Human genome
  - Originally took 10 years to process;
  - Now it can be achieved in one day.

Big Data = Science?

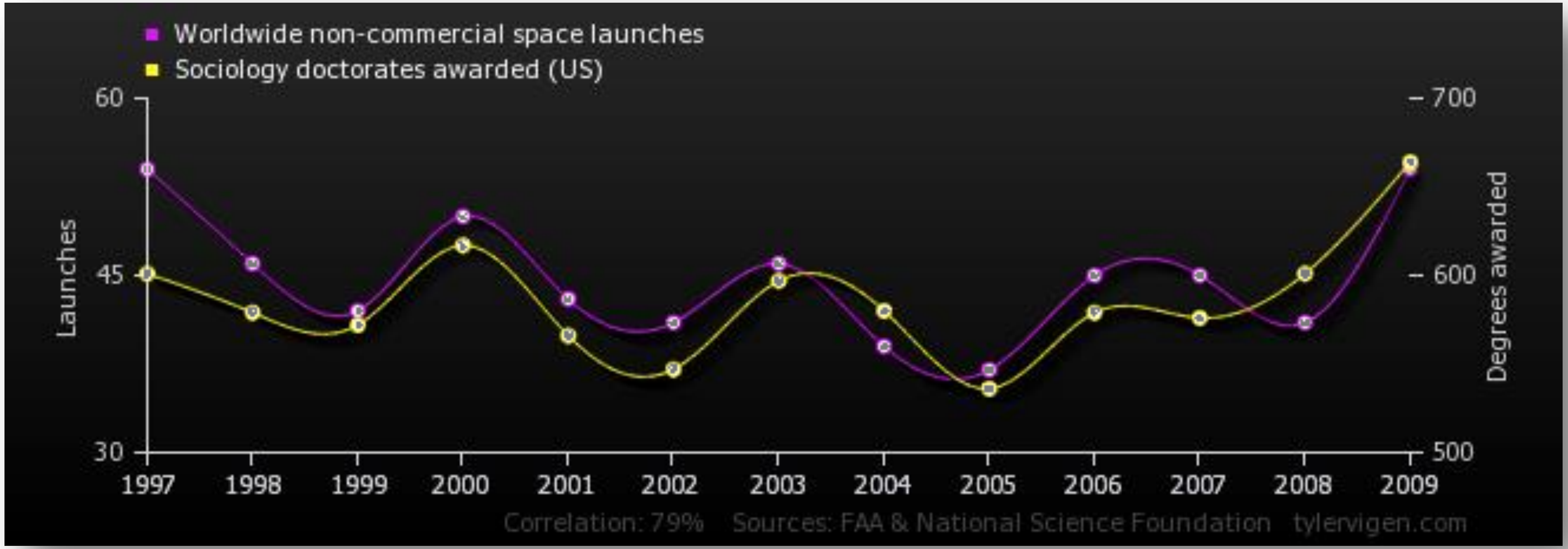
- The End of Theory: The Data Deluge Makes the Scientific Method Obsolete (Chris Anderson, Wired, 2008)
  - All models are wrong, but some are useful. (George Box)
  - All models are wrong, and increasingly you can succeed without them. (Peter Norvig, Google)

- Before Big Data: Correlation is not causation!
- With Big Data: Who cares?
  - Traditional approach to science — hypothesize, model, test — is becoming obsolete.

- Petabytes allow us to say: "**Correlation is enough.**"

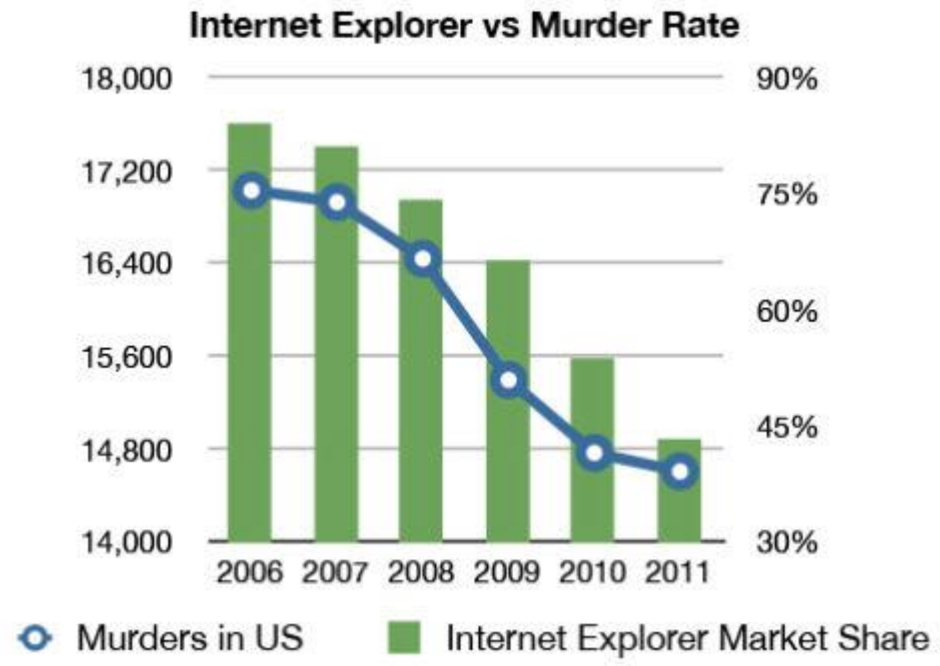
http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Correlation vs. Causation



Quelle: Spurious correlations (www.tylervigen.com)

Correlation vs. Causation



Addressing Big Data: Parallelization

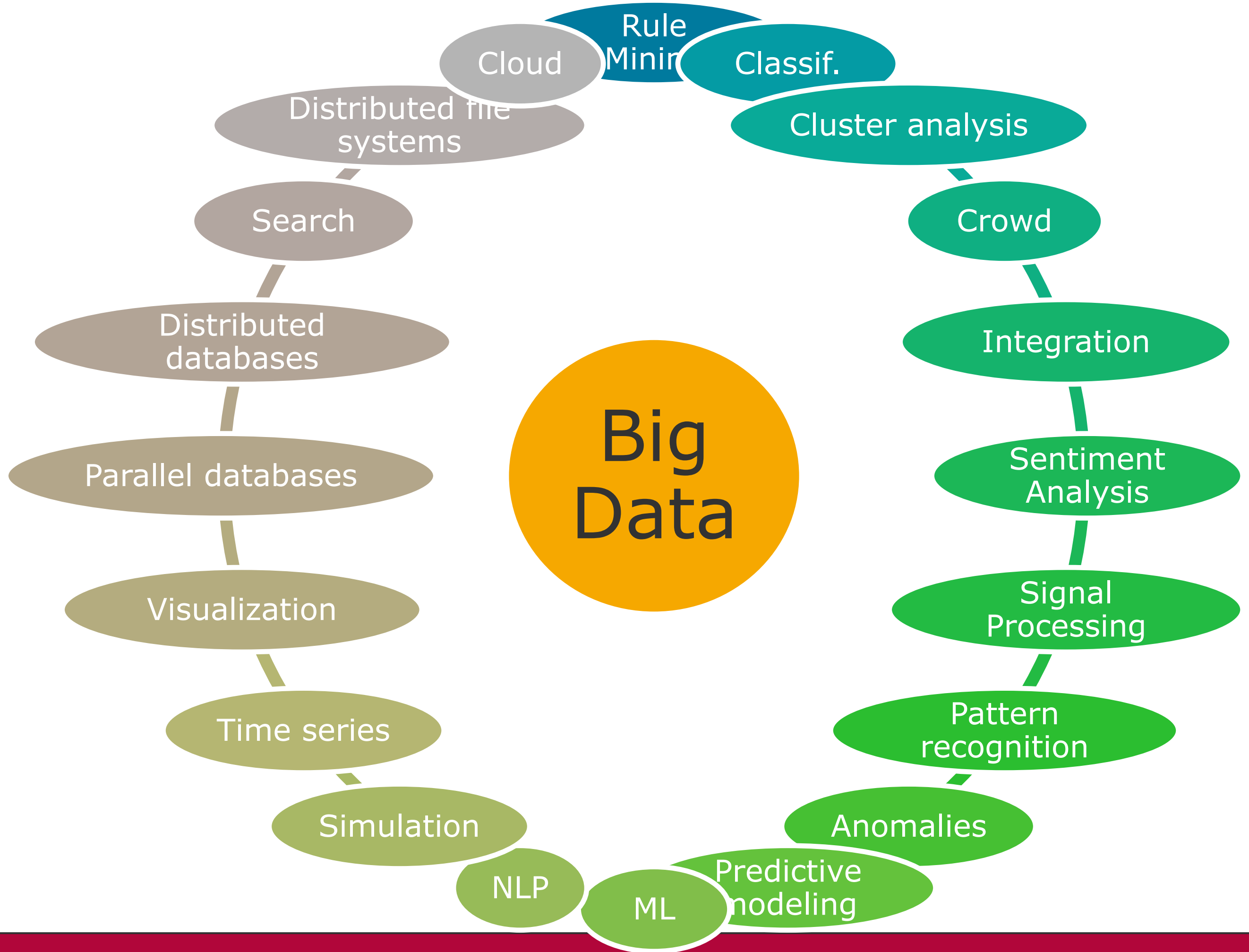
- Long tradition in databases
- Vertical and horizontal partitioning
- Shared nothing
- Each machine runs same single-machine program

- Other trends
  - Map/Reduce / Hadoop
  - Multicore CPUs
  - GPGPUs

Levels of Parallelism on Hardware

- Instruction-level Parallelism
  - Single instructions are automatically processed in parallel
  - Example: Modern CPUs with multiple pipelines and instruction units.
- Data Parallelism
  - Different data can be processed independently
  - Each processor executes the same operations on its share of the input data.
  - Example: Distributing loop iterations over multiple processors
  - Example: GPU processing
- Task Parallelism
  - Different tasks are distributed among the processors/nodes
  - Each processor executes a different thread/process.
  - Example: Threaded programs.

Other technologies to approach big data / data sciences

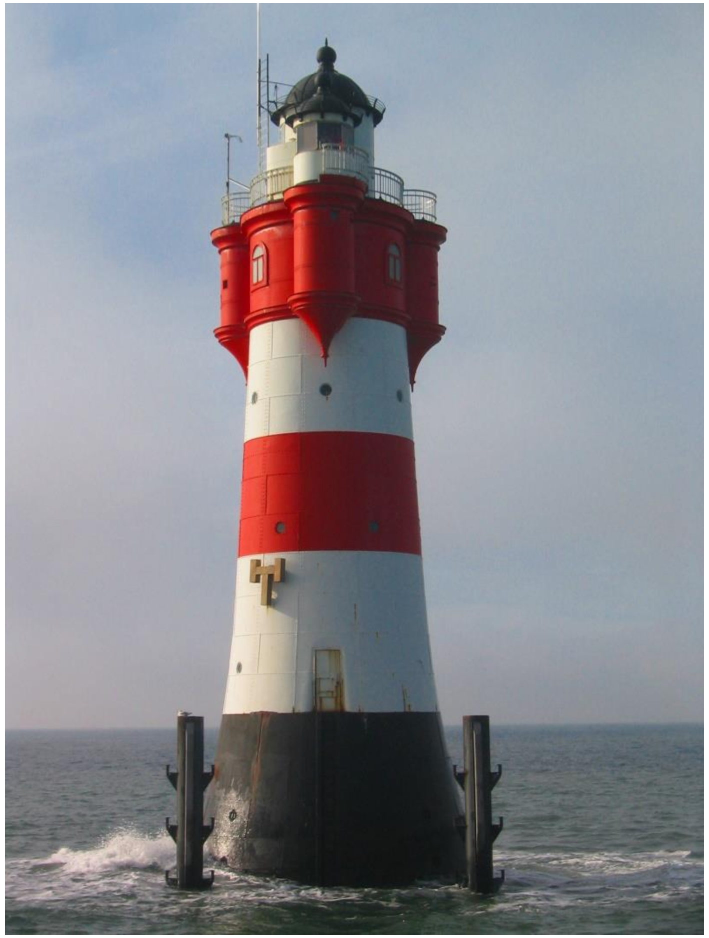


Data profiling and data cleansing are prerequisites for all of these!

- Industry keynote speakers on credit ratings using big data
  - “If the data is out there, we will find it.”
  - “… and that is why I closed my Twitter account.”
  - “… and that is why I had my son close his Twitter account.”

Overview

1. Introduction to research group
2. Lecture organisation
3. (Big) data
   - **Data sources**
   - Profiling
4. Overview of semester



Open vs. closed source

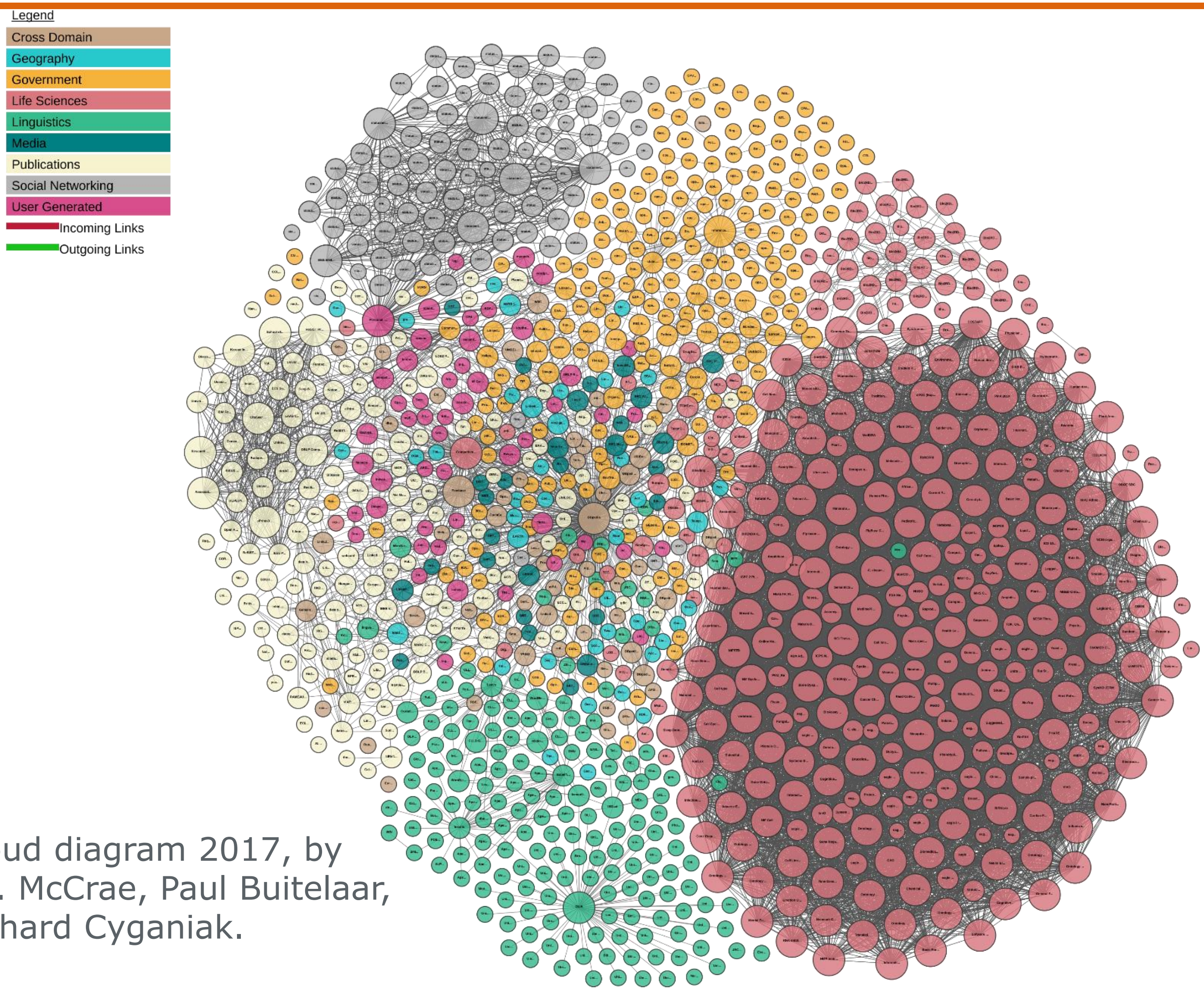
Open

- Linked data
  - http://linkeddata.org/
- Government data
  - data.gov, data.gov.uk
  - Eurostat
- Scientific data
  - Genes, proteins, chemicals
  - Scientific articles
  - Climate
  - Astronomy
- Published data
  - Tweet (limited)
  - Crawls
- Historical data
  - Stock prices

Closed

- Transactional data
  - Music purchases
  - Retail-data
- Social networks
  - Tweets, Facebook data
  - Likes, ratings
- E-Mails
- Web logs
  - Per person
  - Per site
- Sensor data
- Military data

The Linking Open Data cloud diagram



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Wikipedia Infoboxes

```
{{Infobox company
|name             = International Business Machines Corporation
|logo             = <br />[[File:IBM logo.svg|200px]]<br />
|caption          = Logo since 1972, designed by [[Paul Rand]]
|type             = [[Public company|Public]]
|traded_as        = {{New York Stock Exchange|IBM}}<br />[[Dow Jones Industrial Average|Dow Jones Compon
Component]]
|industry         = [[Personal computer hardware|Computer hardware]], [[Software|Computer software]], [[
services]], [[Information technology consulting|IT consulting]]
|products         = [[List of IBM products|See IBM products]]
|founder          = [[Charles Ranlett Flint]]
|foundation       = [[Endicott, New York|Endicott]], New York, U.S.<br />({{Start date|1911|06|16}})
|location_city    = [[Armonk, New York|Armonk]], New York
|location_country = U.S.
|area_served      = Worldwide
|key_people       = [[Ginni Rometty]]<br />({{small|Chairman, President, and CEO}})
|revenue          = {{Increase}} US$ 106.91 [[1000000000 (number)|billion]] <small>(2011)</small><ref na
|url=http://rcpmag.com/articles/2012/01/20/intel-ibm-exceed-earnings-estimates-google-falls-short.aspx|
International Business Machines Corporation |work=United States Securities and Exchange Commission}}</re
|operating_income = {{Increase}} US$ {{0|0}}20.28 billion <small>(2011)</small><ref name=10K/>
|net_income       = {{Increase}} US$ {{0|0}}15.85 billion <small>(2011)</small><ref name=10K/>
|assets           = {{Increase}} US$ 116.43 billion <small>(2011)</small><ref name=10K/>
|equity           = {{Decrease}} US$ {{0|0}}20.13 billion <small>(2011)</small><ref name=10K/>
|num_employees    = 433,362 <small>(2012)</small><ref name="Fortune 500: IBM employees"/>
```

International Business Machines Corporation

Logo since 1972, designed by Paul Rand

Type	Public
Traded as	NYSE: IBM Dow Jones Component S&P 500 Component
Industry	Computer hardware, Computer software, IT services, IT consulting
Founded	Endicott, New York, U.S. (June 16, 1911)
Founder(s)	Charles Ranlett Flint
Headquarters	Armonk, New York, U.S.
Area served	Worldwide
Key people	Ginni Rometty (Chairman, President, and CEO)
Products	See IBM products
Revenue	▲ US$ 106.91 billion (2011)[1]
Operating income	▲ US$ 20.28 billion (2011)[1]
Net income	▲ US$ 15.85 billion (2011)[1]
Total assets	▲ US$ 116.43 billion (2011)[1]
Total equity	▼ US$ 20.13 billion (2011)[1]
Employees	433,362 (2012)[2]
Divisions	Financing, Hardware, Services, Software
Website	IBM.com

Hasso Plattner Institut

Felix Naumann
Data Profiling
Summer 2017

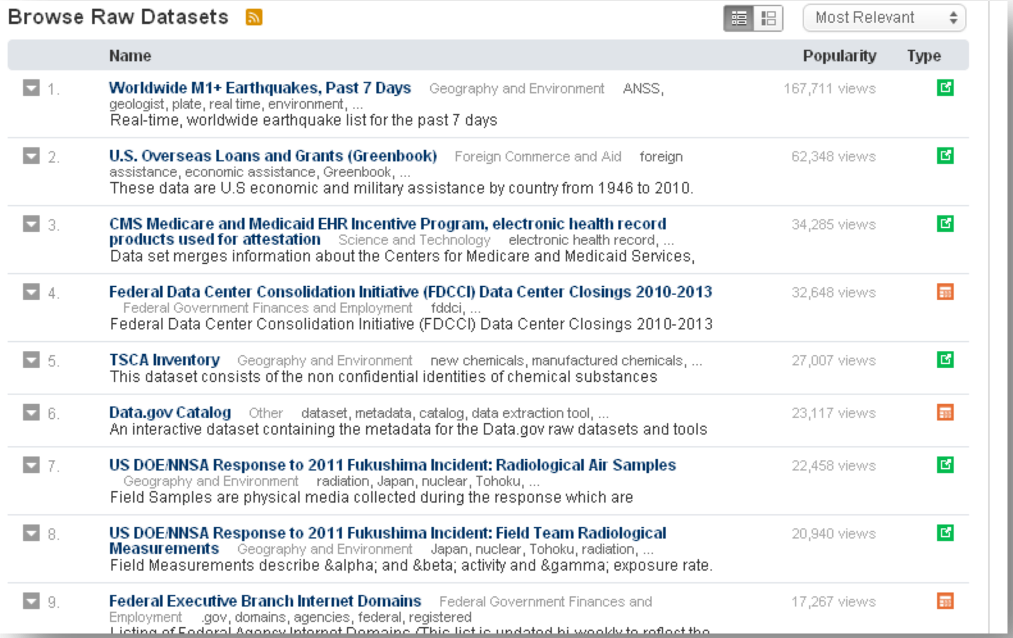
DBpedia statistics

- From 125 languages of Wikipedia
- 3 billion triples
  - 580 million English
- English DBpedia
  - 4.6 million things
  - 1,445,000 persons
  - 735,000 places
  - 411,000 creative works
  - 241,000 organizations
  - 251,000 species
  - …

- http://wiki.dbpedia.org/about/facts-figures

And more sources

- Government data
  - www.data.gov (380k data sets)
  - data.gov.uk (9k)
  - ec.europa.eu/eurostat
- Finance / business data
- Scientific databases
  - www.uniprot.org
  - skyserver.sdss.org
- The Web
  - HTML tables and lists: billions
  - General sources: Dbpedia (3.7m), freebase (23m), microformats…
  - Domain-specific sources: IMDB, Gracenote, isbndb, …



Getting the data

- Download
  - Data volumes make this increasingly infeasible
  - Fedex HDDs
  - Fedex tissue samples instead of sequence data
- Generating big (but synthetic) data
  1. Automatically insert interesting features and properties
  2. Then „magically" detect them
- Sharing data
  - Repeatability of experiments
  - Not possible for commercial organizations

Pathologies of Big Data

- Store basic demographic information about each person
  - **`age`**, **`sex`**, **`income`**, **`ethnicity`**, **`language`**, **`religion`**, **`housing status`**, **`location`**
  - Packed in a 128-bit record
- World population: 6.75 billion rows, 10 columns, 128 bit each
  - About 150 GB
- What is the median age by sex for each country?
  - Algorithmic solution
    - 500$ Desktop: I/O-bound; 15min reading the table
    - 15,000$ Server with RAM: CPUI-bound; <1min
  - Database solution
    - Aborted bulk load to PostgreSQL – disk full (bits vs. integer and DBMS inflation)
  - Small database solution (3 countries, 2% of data)
    - ```
      SELECT country,age,sex,count(*)
      FROM people GROUP BY country,age,sex;
      ```
    - *> 24h, because of poor analysis: Sorting instead of hashing*
    - "PostgreSQL's difficulty here was in **analyzing** [=profiling] the stored data, not in storing it."

- From http://queue.acm.org/detail.cfm?id=1563874
      ```

Big data in Wikipedia

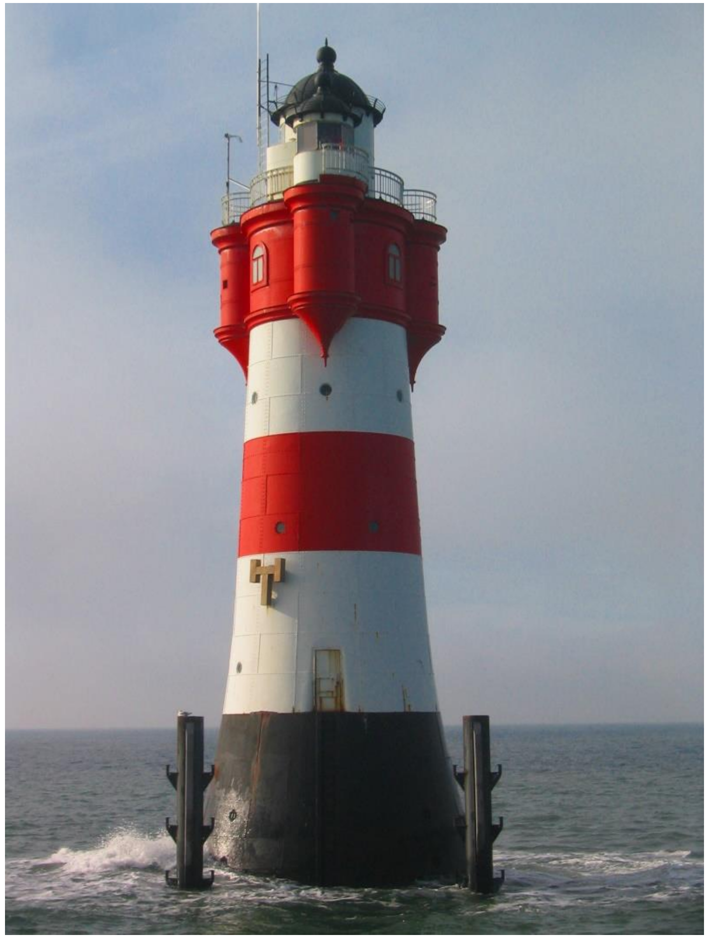
Visualization of edits by user „Pearle“

http://en.wikipedia.org/wiki/File:Viegas-UserActivityonWikipedia.gif

Felix Naumann
Data Profiling
Summer 2017

Overview

1. Introduction to research group
2. Lecture organisation
3. (Big) data
   - Data sources
   - **Profiling**
4. Overview of semester



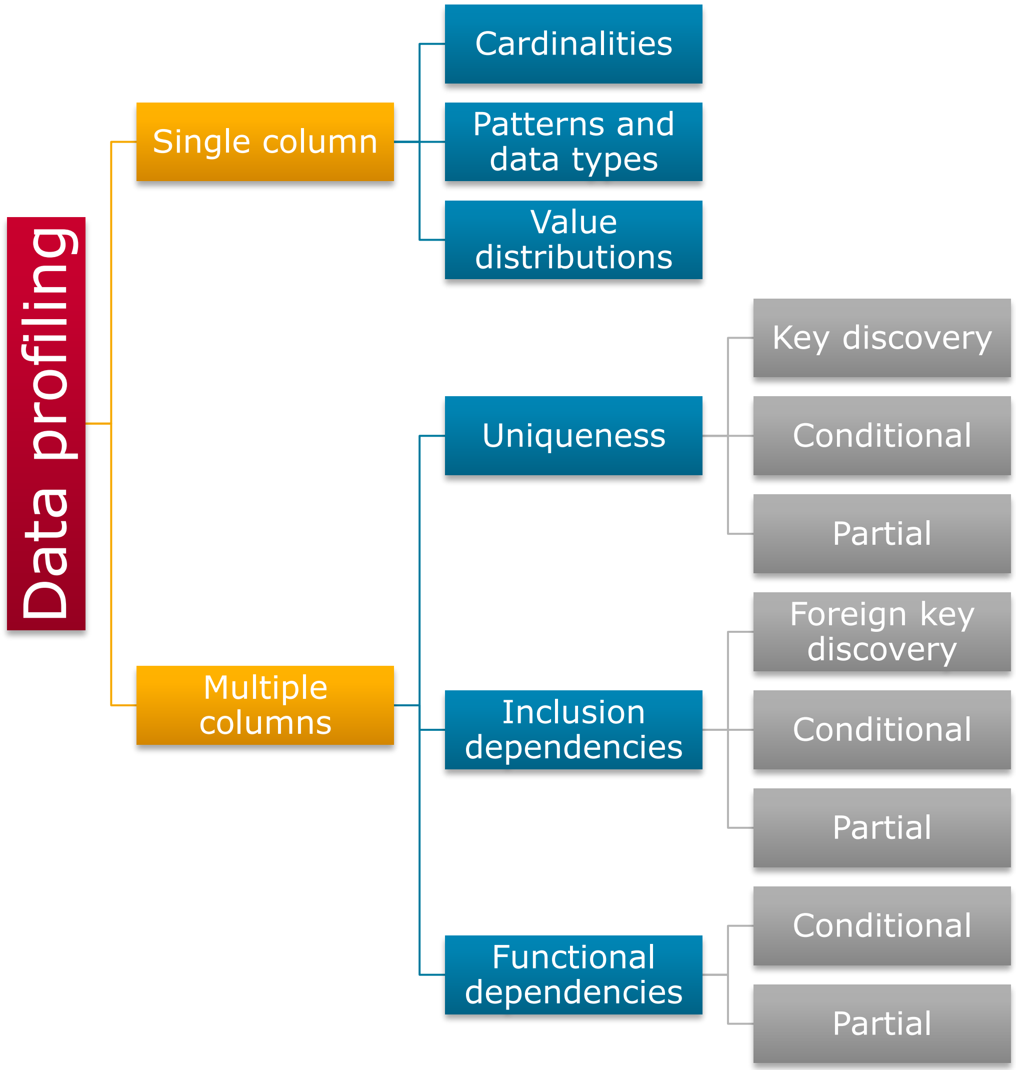
Definition Data Profiling

- 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

- 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, etc.)

- Define as a set of data profiling tasks / results

Classification of Profiling Tasks



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 and mining
  - Profiling as preparation to decide on models and questions
- Database reverse engineering

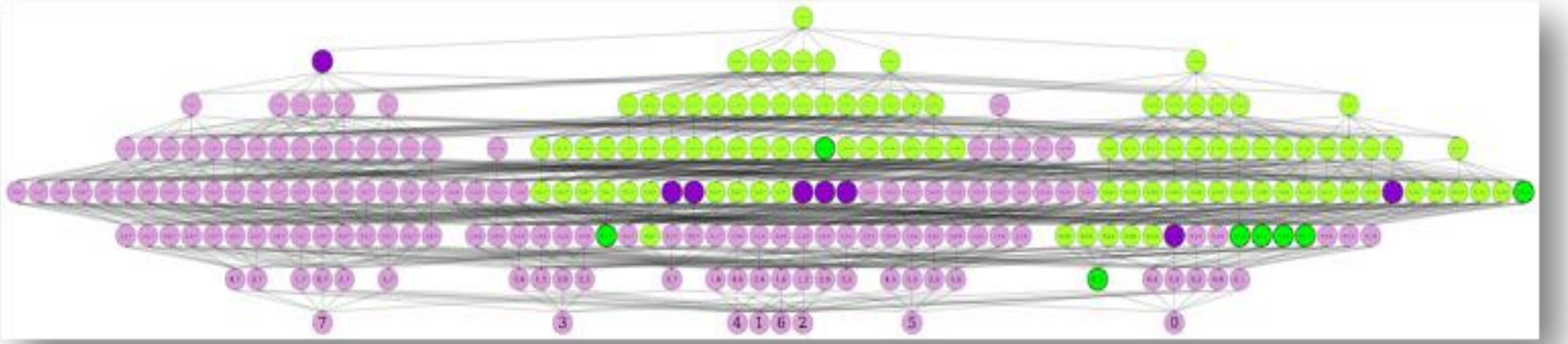
- Data profiling as preparation for any other data management task

Challenges of (Big) Data Profiling

- Computational complexity
  - Number of rows
    - Sorting, hashing
  - Number of columns
    - Number of column combinations
- Large solution space
- I/O-bound due to large data sets and distribution

- New data types (beyond strings and numbers)
- New data models (beyond relational): RDF, XML, etc.
- New requirements
  - User-oriented
  - Streaming
  - Etc. – see next slide set

Large solution space

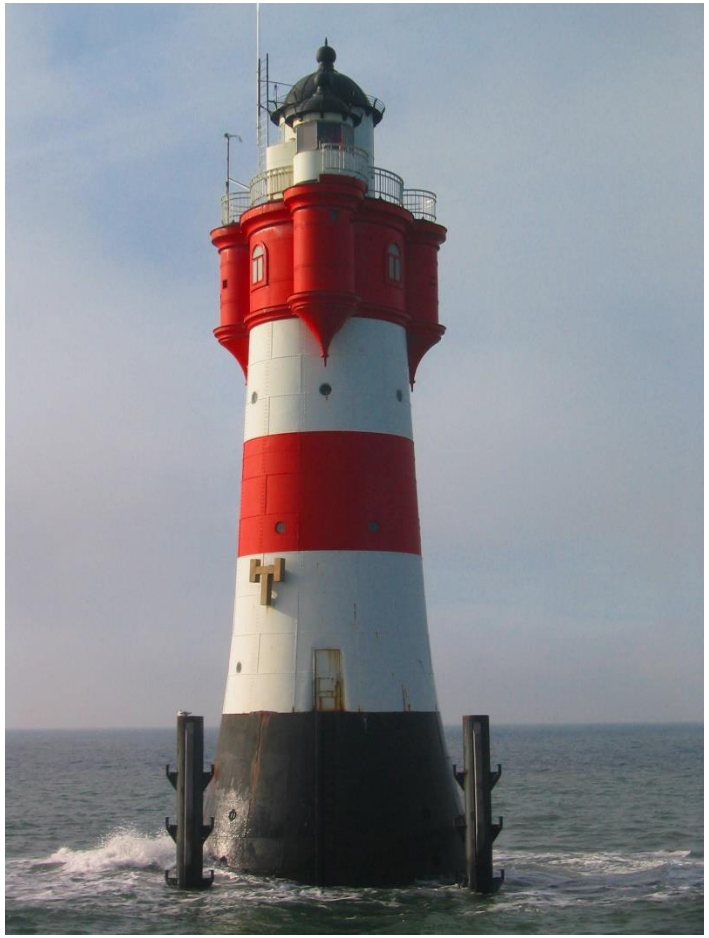


- Size of lattice: 2^n-1 (empty set not considered)
- Nodes at level 1: n
- Nodes at level n: 1
- Nodes at level k: $\binom{n}{k} = \frac{n!}{(n-k)!k!}$
- Largest level at n/2: $\binom{n}{n/2} = \frac{n!}{\left(\frac{n}{2}!\right)^2}$
- $\binom{n}{k} \in \Theta(n^k)$, i.e., bounded from above and below

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1

Overview

1. Introduction to research group
2. Lecture organisation
3. (Big) data
   - Data sources
   - Profiling
4. **Overview of semester**



Schedule

- Big Data (today) and Data Profiling Introduction
- Data Stuctures
  - Lattices, Apriori traversal, complexity, agree-sets/evidence sets, PLIs, Bloom filters
- Unique Column Combinations (UCCs)
  - A-Priori, DUCC, HCA?, Gordian?, Swan?
- Scientific experiments
- Functional Dependencies (FDs)
  - TANE, FD-Mine, FDep, HyFD + approximate TANE
  - FD-measures for ranking (g1-3, support & confidence)
- Inclusion Dependencies (INDs)
  - Spider, Binder, Find2, zigzag, SINDY?, MANY, cINDs?
- Semantics
  - Key & FK detection, normalization, interpretation

Schedule

- Order Dependencies (ODs)
  - ORDER, Szlichta-paper
- Denial Constraints
  - FastDC, Hydra
- Data Synopses
- Column Uniqueness: Approximately counting number of unique values
- Approximation
  - partial, conditional, approximate as concepts, selected approaches
- RDF-Profiling (optional)
- Outlook

- Guest lectures
  - Thomas Bläsius
  - Giuseppe Polese