Data Profiling and Data Cleansing
Introduction

9.4.2013

Felix Naumann

Overview

- Introduction to research group
- Lecture organisation
- (Big) data
  - Data sources
  - Profiling
  - Cleansing
- Overview of semester

Information Systems Team

DFG

IBM

Katrin **Heinrich**

Prof. Felix **Naumann**

Arvid **Heise**

Dustin **Lange**

Johannes **Lorey**

Christoph **Böhm**

project **DuDe**

Duplicate Detection

Data Fusion

Data Profiling

project **Stratosphere**

Entity Search

Information Integration

Data Scrubbing

project **GovWILD**

Data as a Service

Data Cleansing

Information Quality

Web Data

Linked Open Data

RDF Data Mining

Dependency Detection

ETL Management

Service-Oriented Systems

Entity Recognition

Opinion Mining

Anja **Jentzsch**

Ziawasch **Abedjan**

Tobias **Vogel**

HPI Research School

Dr. Gjergji **Kasneci**

Zhe **Zuo**

Maximilian **Jenders**

Toni **Grütze**

Other courses in this semester

Lectures

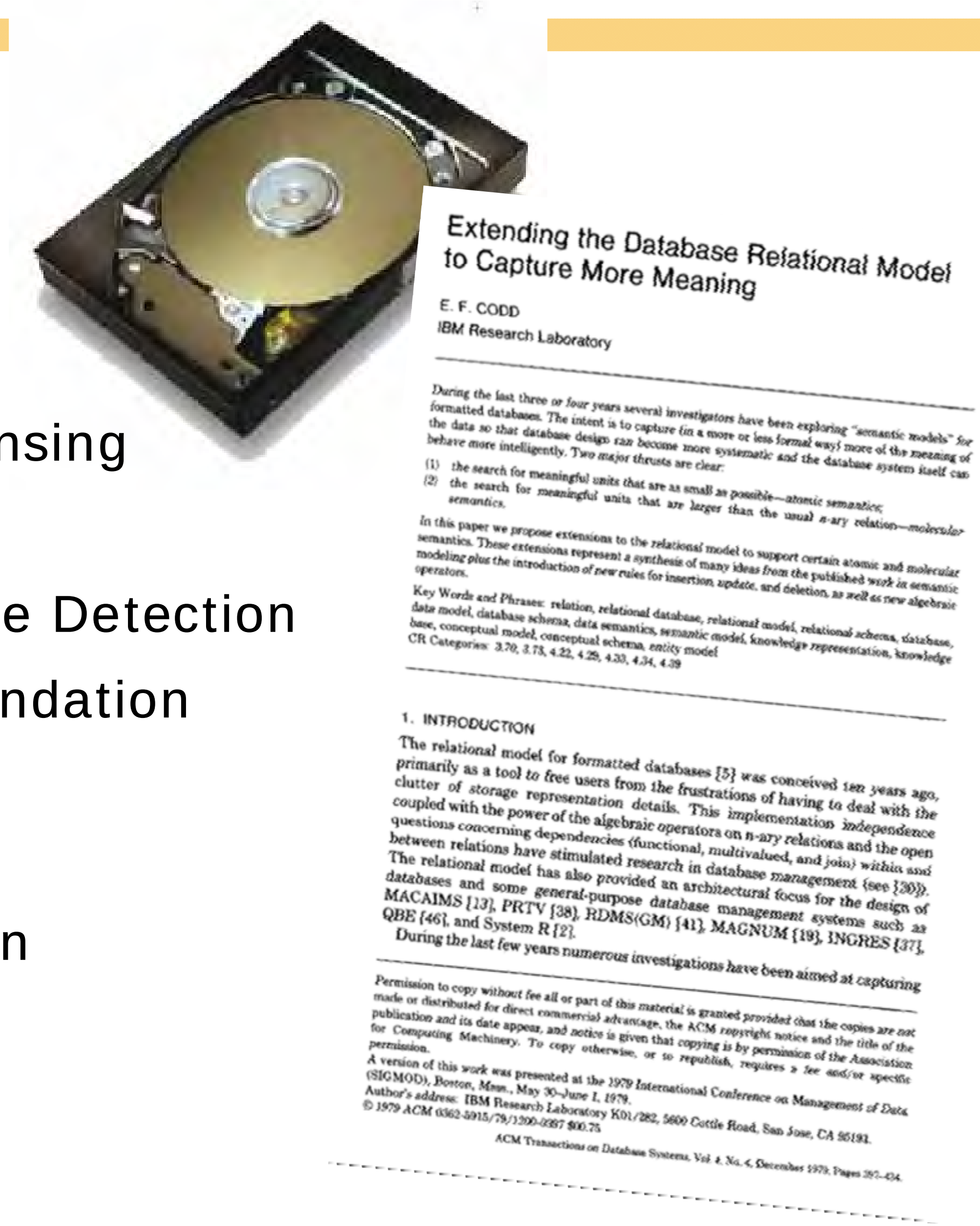
- DBS I (Bachelor)
- Data Profiling and Data Cleansing

Seminars

- Master: Large Scale Duplicate Detection
- Master: Advanced Recommendation Techniques

Bachelorproject

- VIP 2.0: Celebrity Exploration



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.

- Goal: Cross-platform recommendation for posts on the Web
  - Given a post on a website, find relevant (i.e., similar) posts from other websites
  - Analyze post, author, and website features
  - Implement and compare different state-of-the-art recommendation techniques

Sim	P_1	...	P_j	...	P_n
P_1					
P_2					
...					
P_i			?		
...					
P_n					

Calculate $Sim(P_i, P_j)$ (i.e., the similarity between posts P_i and P_j)

Recommend top-k posts

Overview

- Introduction to research group
- Lecture organization
- (Big) data
  - Data sources
  - Profiling
  - Cleansing
- Overview of semester

Dates and exercises

- Lectures
  - Tuesdays 9:15 – 10:45
  - Thursdays 9:15 – 10:45
- Exercises
  - In parallel
- First lecture
  - 9.4.2013
- Last lecture
  - 11.7.2013
- Holidays
  - 9.5. Ascension
- Exam
  - Oral exam, 30 minutes
  - Probably first week after lectures
- Prerequisites
  - To participate
    - Background in databases (e.g. DBS I)
  - 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 15-16
  - Email: naumann@hpi.uni-potsdam.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

Exercise

- Algorithm design and programming exercises
- Data profiling (emphasis on efficiency and scalability)
  - Unique column combinations
  - Inclusion dependencies
  - Functional dependencies
- Data Cleansing (emphasis on quality)
  - Duplicate detection
- Self-motivation wrt good solutions!

Introduction: Audience

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

Overview

- Introduction to research group
- Lecture organization
- (Big) data
  - Data sources
  - Profiling
  - Cleansing
- 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/

Big Data (according to Wikipedia)

- 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.
  - Capture
  - Curation
  - Storage
  - Search
  - Sharing
  - Analysis
  - Visualization

  No transaction management
- Gartner: Big data are **high-volume, high-velocity, and/or high-variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

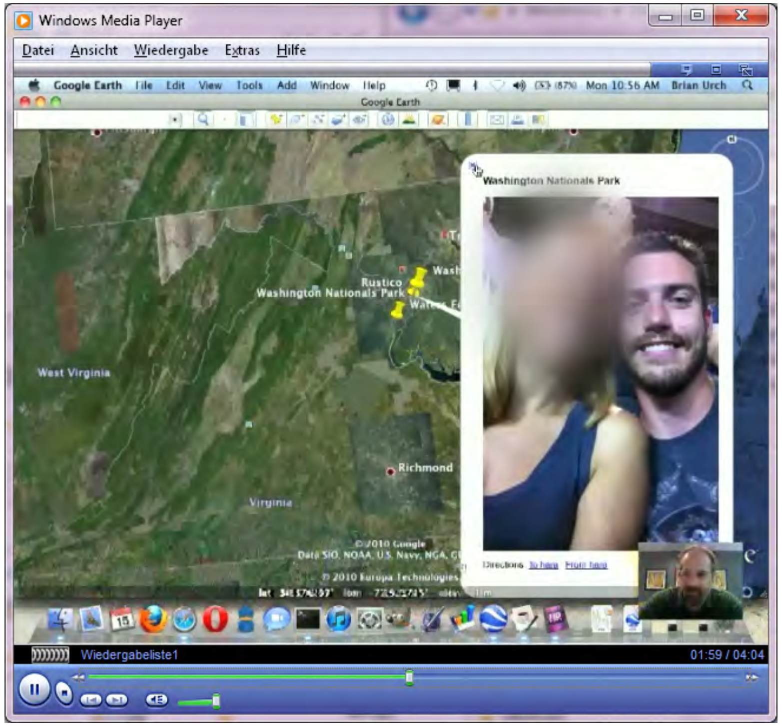
Big and Small

- Big Data can be very small
  - Streaming data from sensor aircrafts
  - 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
  - See video of Raytheon’s RIOT software



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 week.

Examples from Wikipedia – Government

- In 2012, the Obama administration announced the Big Data Research and Development Initiative, which explored how big data could be used to address important problems facing the government. The initiative was composed of 84 different big data programs spread across six departments.
- The United States Federal Government owns six of the ten most powerful supercomputers in the world.
- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.

Examples from Wikipedia – Business

- Amazon.com
  - Millions of back-end operations every day
  - Queries from more than half a million third-party sellers
  - In 2005: the world’s three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- Walmart
  - more than 1 million customer transactions every hour
  - 2.5 petabytes (2560 terabytes) of data
- Facebook handles 50 billion photos from its user base.
- FICO
  - Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide

Big data in science – Chris Anderson

- The End of Theory: The Data Deluge Makes the Scientific Method Obsolete (Chris Anderson, Wired, 2008)
  - http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
- All models are wrong, but some are useful. (George Box)
- All models are wrong, and increasingly you can succeed without them. (Peter Norvig, Google)
- Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.
- Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence).
- Faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.
- Petabytes allow us to say: "Correlation is enough."

Big data by IBM

- Every day, we create 2.5 quintillion bytes of data.
- 90% of all data was created in the last two years.
- Sources
  - Sensors used to gather climate information
  - Posts to social media sites
  - Digital pictures and videos
  - Purchase transaction records
  - Cell phone GPS signals
- Big data is more than simply a matter of size
  - *Opportunity* to find insights in new and emerging types of data and content,
  - Make businesses more *agile*
  - *Answer questions* that were previously considered beyond your reach.

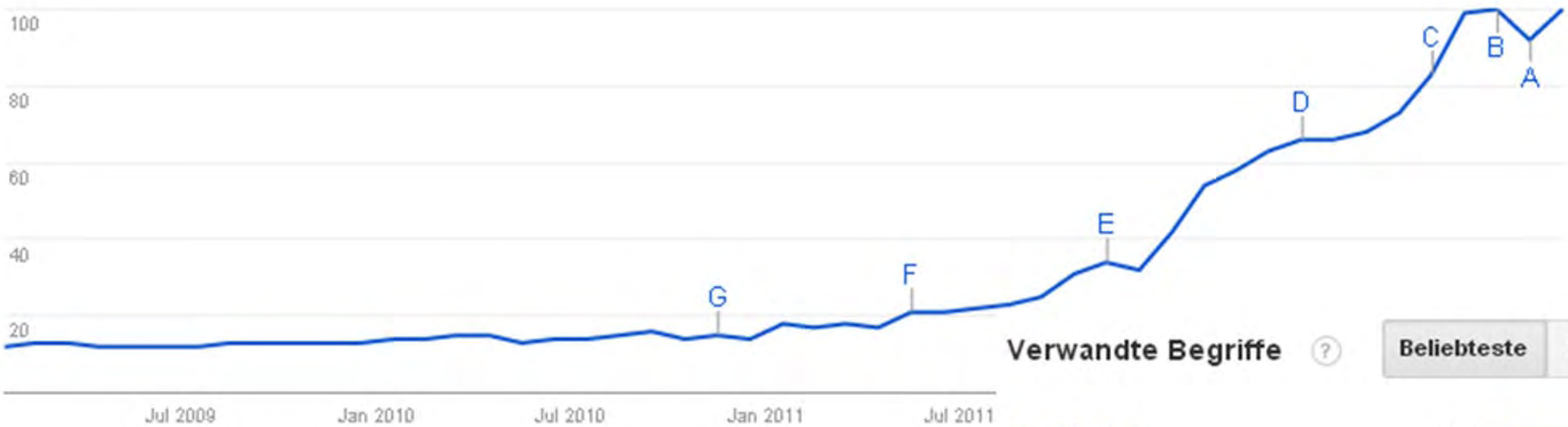
Shallow buzzword…

The four Vs – examples

- Volume
  - Turn 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.

From: http://www-01.ibm.com/software/data/bigdata/

Google trends (Jan 2013)



Verwandte Begriffe | **Beliebteste** | **Zunehme...**

the big data	100
data analytics	50
big data analytics	45
hadoop	40
big data hadoop	35
big data google	25
oracle big data	20
big bang data	20
big data ibm	20
big data cloud	15

Adressing Big Data: Paralleization

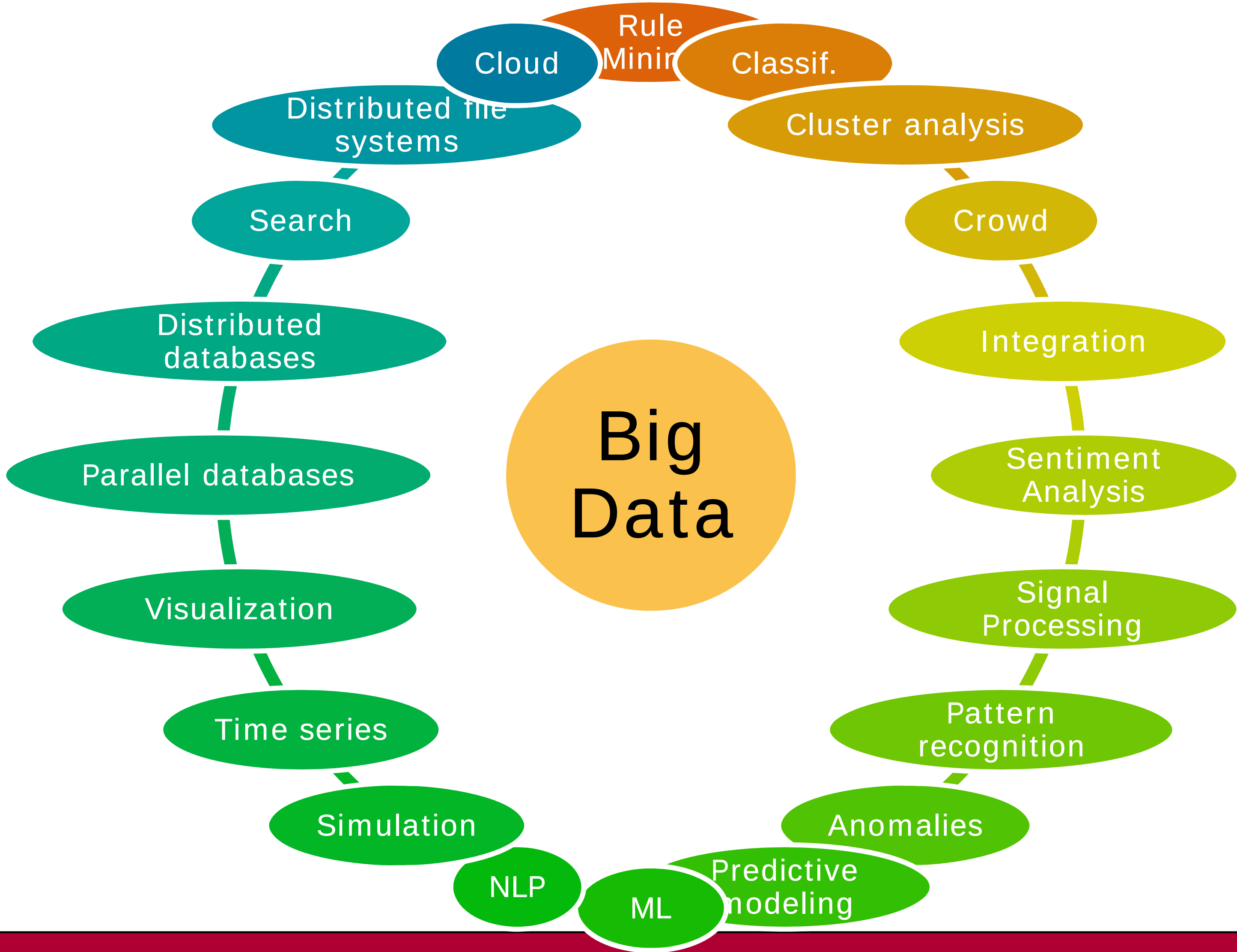
- Long tradition in databases
- Vertical and horizontal partitioning
- Shared nothing
- Each machine runs same single-machine program

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

Overview

- Introduction to research group
- Lecture organisation
- (Big) data
  - Data sources
  - Profiling
  - Cleansing
- Overview of semester

Examples sources

- Science
  - Astronomy
  - Atmospheric science
  - Genomics
  - Biogeochemical and biological research
- Web
  - Web logs
  - Internet text and documents
  - Web indexing
- Business
  - Transactions
- Sensors
  - RFID
  - Sensor networks
  - Military surveillance
- Person data
  - Social networks, social data
  - Call detail records
  - Medical records
- Multimedia
  - Photo and video archives

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

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

Hasso Plattner Institut

Visualization of edits by user „Pearle“

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

Overview

- Introduction to research group
- Lecture organisation
- (Big) data
  - Data sources
  - Profiling
  - Cleansing
- 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.

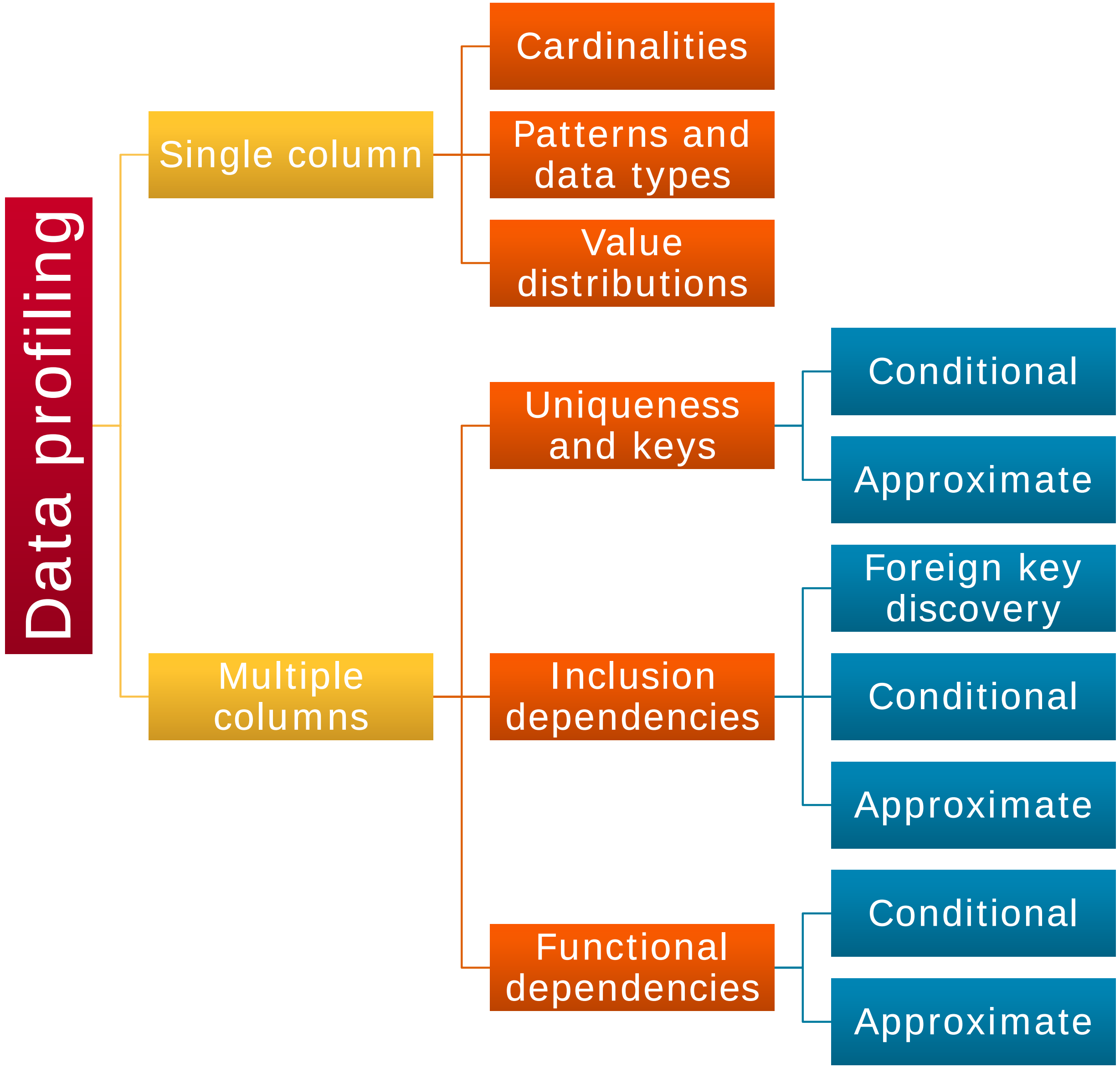
  Wikipedia 03/2013
- Data profiling refers to the activity of creating small but informative summaries of a database.

  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)
- New requirements
  - User-oriented
  - Streaming
  - Etc. – see nest slide set

Overview

- Introduction to research group
- Lecture organization
- (Big) data
  - Data sources
  - Profiling
  - Cleansing
- Overview of semester

Data Cleansing – Definition

- Detect and correct errors in data sets

- Scrubbing
  - Value- and tuple-level operations
  - Outliers
  - Rule-violations
  - Dependency violations

- Cleansing
  - Relation-level operations
  - In particular: duplicate detection and data fusion

Datenqualität: Probleme

Repräsentation

Widersprüche

Ref. Integrität

KUNDE

KNr	Name	Geb.datum	Alter	Geschlecht	Telefon	PLZ
1234	Kuhn, Mark	18.2.80	37	m	999-9999	98693
1234	Anne Will	32.2.70	37	f	768-4511	55555
1235	Mark Kuhn	18.2.80	27	m	567-3211	98693

Eindeutigkeit

Fehlende Werte

Duplikate

ADRESSE

PLZ	Ort
98693	Potsdam
98684	Postdam
98766	BRD

Schreibfehler

Falsche Werte

Data Cleansing Use-cases

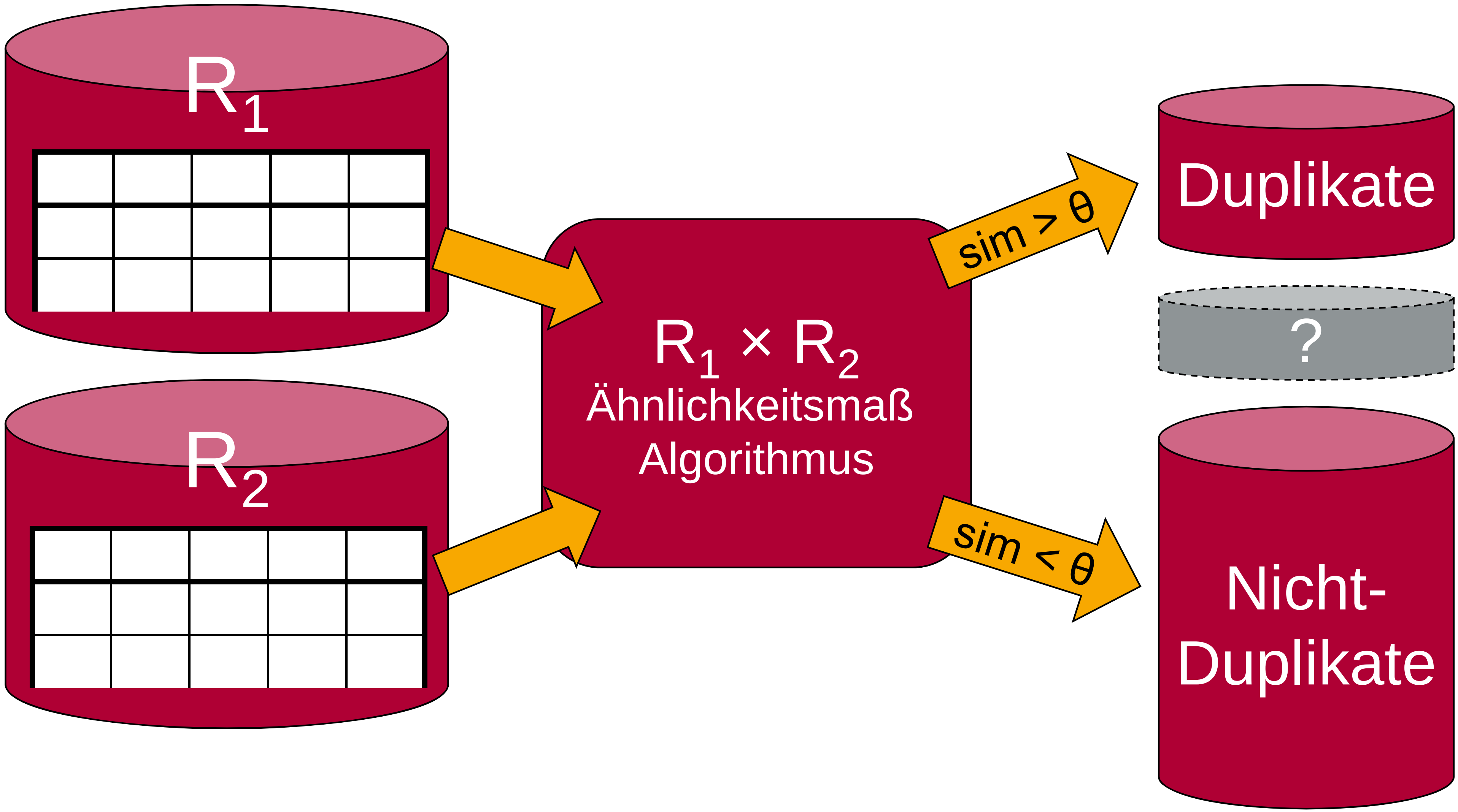
- Master Data Management MDM
- Customer Relationship Management CRM
- Data Warehousing DWH
- Business Intelligence BI

- Examples
  - Inventory levels
  - Banking risks
  - IT overhead
  - Incorrect KPIs
  - Poor publicity

Data Cleansing Challenges

- Defining data quality
  - Data profiling to the rescue
- Semantic complexity
  - Often only an expert can determine the correct value
  - Any techniques are dependent on data set and desired result
    - Much fine-tuning
- Computational complexity
  - Duplicate detection is quadratic
- Evaluation is difficult
  - No gold-standard
  - Anecdotal evidence: „Why don‘t you detect this problem, here?“

Main Challenge: Duplicate Detection



Overview

- Introduction to research group
- Lecture organization
- (Big) data
  - Data sources
  - Profiling
  - Cleansing
- Overview of semester

Schedule

- Part I: Data Profiling
  - Introduction and overview (1)
  - Value patterns and data types (1)
  - Uniqueness detection (1)
  - IND detection (2)
  - FD and CFD detection (1)
  - LOD Profiling (1)
- Industry lectures
  - IBM (1)
  - SAP (1)
- Part II: Data Cleansing
  - Introduction and overview (1)
  - Scrubbing and normalization (1)
  - Data sets and evaluation (1)
  - Similarity measures (1)
  - Similarity indexes (1)
  - Generalized duplicate detection (Stanford) (1)
  - Blocking (2)
  - SNM-based methods (2)
  - Clustering-based methods (Getoor) (1)
  - (Big) Data ethics (1)