GEORGE MASON UNIVERSITY | Computational and Data Sciences

CSI 695: Scientific Databases

Fall Term 2017

Lecture 1: Overview

Lectures: Prof. Dr. Matthias Renz

Exercises: TBA

About the course

- Class schedule
  - Lectures: Wednesday, 4:30 pm - 7:10 pm, Exploratory Hall 3301.
  - Homework & Exercises: Assignments almost every 2nd week. Discussion in class
- Office hours:
  - Matthias: Mondays, 3:00 pm - 4:00 pm, Research Hall, 2nd floor, room 242 or by appointment (email: mrenz@gmu.edu, please use "CSI 695" in your email subject)
- Exam:
  - Exams will be based on the material discussed in the class plus the exercises.
  - Final Exam: TBA
- Grade
  - Final written exam at the end.

Who am I?

- Diploma in Electrical Engineering

Munich University of Applied Sciences, Germany, 1997-2002.

- Diploma in Computer Science

Department of Computer Science, Ludwig-Maximilians University Munich, Germany, 2002.

- PhD in Computer Science

Department of Computer Science, Ludwig-Maximilians Universität (LMU) Munich, Germany, 2006.

- Habilitation in Computer Science

Department of Computer Science, Ludwig-Maximilians Universität (LMU) Munich, Germany, 2011.

- Acting chair of the Database Systems Group

Department of Computer Science, Ludwig-Maximilians Universität (LMU) Munich, Germany, 2015.

- Associate Professor

Department of Computational & Data Science, George Mason University, 01/15/2016 – present.

Research interests

- My Portfolio:

Applications
- Big Data Analytics
- Industrie4.0

Industry

- eScience
- ***Data-driven Science***

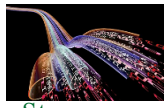
Science

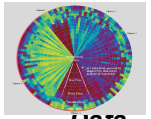
Environments
- Sensor Networks
- Embedded Systems
- Mobile Devices

HW

- Privacy
- Reliability

Meta

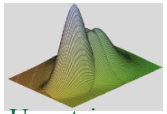

Streams

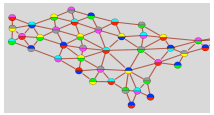

High Dimensional *Data*


Text


Enriched Geo


Timeseries


Uncertain


Graphs


Multimedia


Spatiotemporal

Methods

Managing:
- Data Management
- Similarity Models
- Indexing

Searching:
- Similarity Search
- Query Processing

Pattern Mining
- Clustering
- Outlier Detection
- Frequent Pattern Mining

Conferences (Top-10 Venues)

Application-Oriented:
- SSDBM
- ***DASFAA***
- ***SSTD***
- ***SIGSPATIAL***

Data Engineering and Mgmt.
- ICDE
- CIKM
- EDBT
- ....

DB & Databases Flag Ship Conf.
- ***SIGMOD***
- VLDB
- KDD

....

Outline

- Course Overview
- Why we need Databases?
- What's next

Course Overview

- Course Objectives
  - Acquire knowledge of scientific database and data management techniques:
    - 1) to model scientific databases for science applications,
    - 2) to design, access, and query scientific data,
    - 3) to manage very large sets of scientific data.
  - Become familiar with basic database concepts to model and manage scientific data.
  - Understand unique requirements of Scientific Databases.
  - Understand techniques for non-standard query processing methods to solve scientific problems.

Course Overview

- This is an interdisciplinary course that focuses on the use of data and databases in scientific research, including:
  - Relational Databases (database concepts, modeling, building, searching)
  - Integration of scientific data
  - Scientific Databases vs. Relational Databases
  - General concepts for managing and searching in scientific data
  - Concepts for managing geo-spatial data
  - Concepts for managing time series and sequence data
- Your homework assignments will be exercises that provide learn-by-doing experience. At most one assignment every second week.
- Solutions of exercises will be discussed during the lecture. At most every second week in sync. with assignments.

Course Overview

- Course syllabus @ http://mymason.gmu.edu/ …
- Blackboard website @ http://mymason.gmu.edu/ …
- Log in this week (with your GMU email username & password)
  - to verify your access, and
  - to check for first Homework Assignments (HW#01)
- Log in at least 2x each week throughout the semester to check for new material and class announcements
- Course Material will be added throughout the semester…
- Lecture notes, homework assignments, exam review materials, your grades, and important class announcements through Blackboard.
- The initial Course Material will be added soon.

Course Overview

Date	Lec. #	Topic
		Overview
	1	Course Overview - Introduction to the course
		Introduction to Databases
	2	Introduction to Databases, From Data to Data Management
	3	Modeling scientific data and Ontologies, Entity/Relationship-Model
	4	Relational Algebra, Introduction to SQL (DDL/DML)
		Principles of Scientific Data Management
	5	Introduction to Scientific Data Management, Searching in Scientific Data: Feature Based Similarity Search
	6	Algorithmic Similarity Search Paradigms, Efficient Managing Scientific Data, Range Query Algorithms
	7	Nearest-Neighbor Query Algorithms
	8	Reverse Nearest-Neighbor Query Algorithms, Skyline Query Algorithms, Assessment Methods for Similarity Search Methods
		Managing Spatially Extended Objects
	9	Efficient Representation of Spatial Objects for Spatial Join Queries
	10	Spatial Object Management based on Space Partitioning, Spatial Object Management based on Data Partitioning
	11	Similarity Models for Spatially Extended Objects
		Similarity Search in Time-Series Databases
	12	Introduction to Time-Series Matching Queries, Dynamic Time Warping
		Thanksgiving Break
	13	Dimensionality Reduction Techniques, Review I: Database modeling and Querying
	14	Review II: Similarity search algorithms and Spatial and Temporal Object Modeling

Please note that this schedule is tentative and it will change according to the needs, pace and interest of the class. Dates of midterm and quizzes might change. Students will be notified in time.

Course Overview

notes, retrieve assignment data and, review links to additional materials, and receive special announcements. Assignments will be posted in Blackboard.

Please be aware that innocent remarks can be easily misconstrued. Sarcasm and humor can be easily taken out of context. When communicating, please be positive and diplomatic. I encourage you to learn more about Netiquette.

7. Grades

Partial credit is given on all problems provided work is shown. In grading, the correct answer and explanation counts. If the answer is not correct, evidence that students understood something about the problem is taken into account. The more evidence that is provided, the more partial credit is given.

Final grades at the end of the course will be assigned using **a combination of absolute achievements and relative effort in class's activities.**

7.1 Weighing

At the end of the term all the marks will be totaled as a weighted average according to the following weights:

Relative in class performance: Participation in discussions	**30%**
Homework Assignments and Exercises	**30%**
Final Exam	**40%**

7.2 Attendance

Attendance is not mandatory, however, it helps tremendously in comprehending the course material and with the assignment. Please keep in mind that since no textbook covers the course material, it is very difficult to follow it if you are more than once or twice absent.

7.3 Assignments

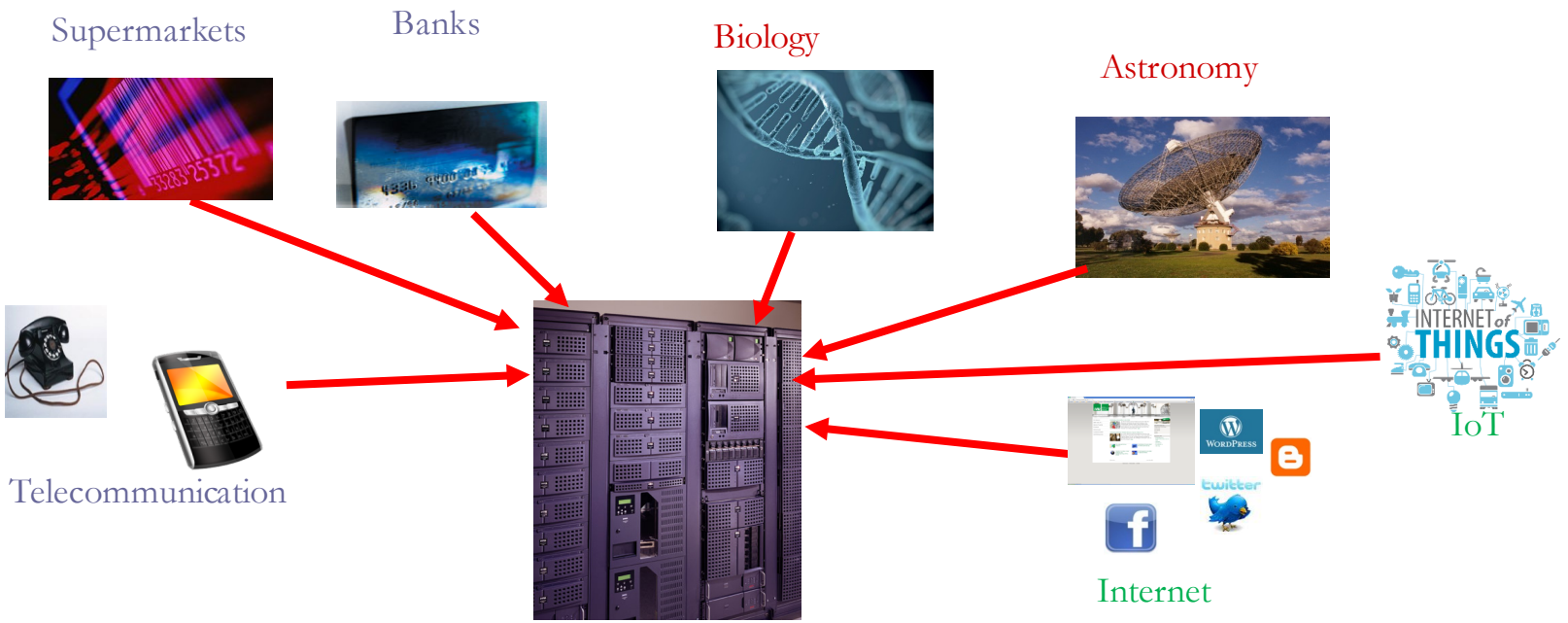
- There will be assignments which will be posted in Blackboard.
- Assignments are usually to be uploaded in Blackboard as PDF file - **DO NOT email assignments directly to the instructor.**
- Assignments are due one week after they are posted in Blackboard if not . Due dates and method of delivery will also be posted in Blackboard.
- **Late assignment submission**: Assignments submitted **after the due date will not be accepted**. Exceptions to this policy may be made given serious circumstances at the discretion of the instructor.

Outline

- Course Overview
- Why we need Databases?
- What's next

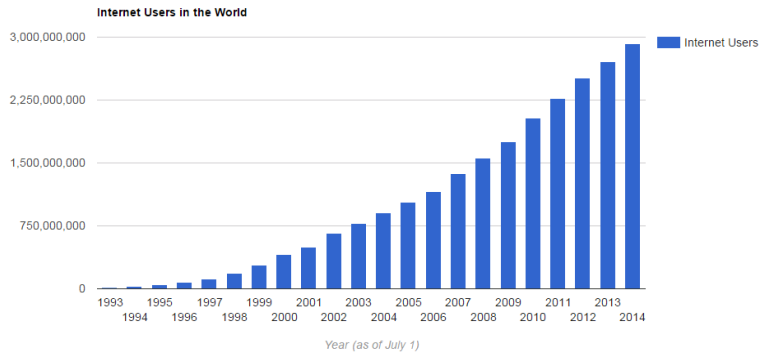
Why we need Databases

- Huge amounts of data are collected nowadays from different application domains
- *"We are drowning in information but starving for knowledge"* John Naibett link
- The amount and the complexity of the collected data does not allow for manual analysis.

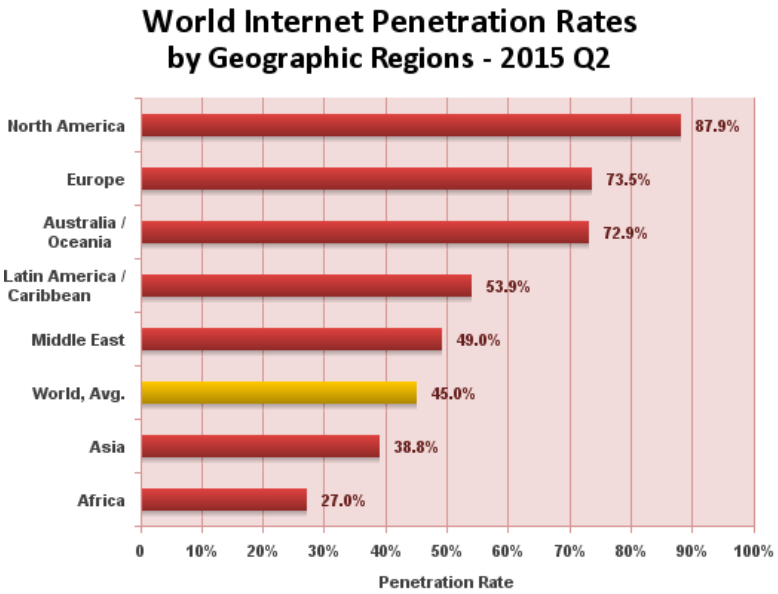


Examples of data sources: The Internet

- Internet users (Source: http://www.internetlivestats.com/internet-users/)



Web 2.0: A world of opinions



Source: Internet World Stats - www.internetworldststs.com/stats.htm
Penetration Rates are based on a world population of 7,260,621,118 and 3,270,490,584 estimated Internet users on June 30, 2015.
Copyright © 2015, Miniwatts Marketing Group

Examples of data sources: Internet of things

- The Internet of Things (IoT) is the network of physical objects or "things" embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data.

Source: https://en.wikipedia.org/wiki/Internet_of_Things

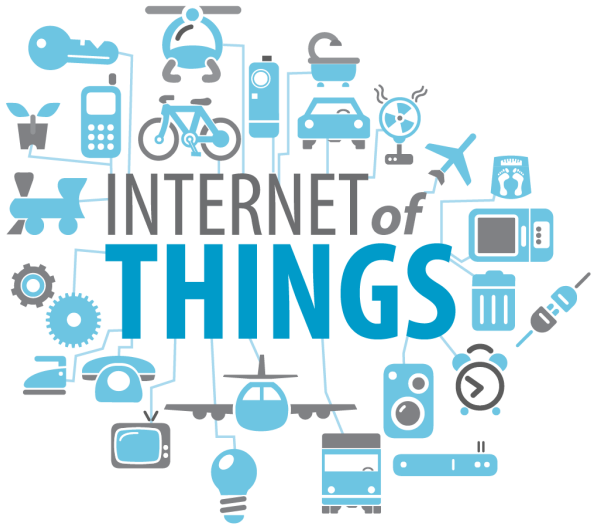


Image source:http://tinyurl.com/prtfqxf

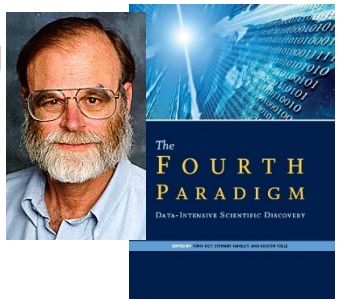
During 2008, the number of things connected to the internet surpassed the number of people on earth… By 2020 there will be 50 billion … vs 7.3 billion people (2015).

These things are everything, smartphones, tablets, refrigerators …. cattle.

Source: http://blogs.cisco.com/diversity/the-internet-of-things-infographic

Examples of data sources: data intensive science

[Comp. Science Pionier Jim Gray]



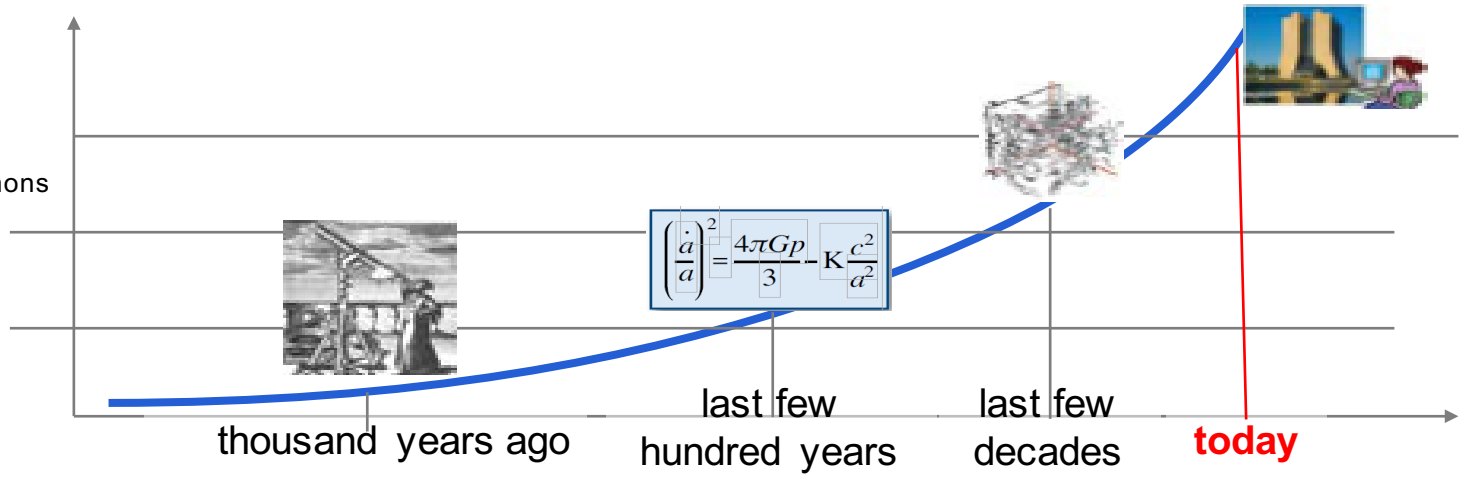
- The Fourth Paradigm:
  Age of data driven exploration
  → Data Science
- Science Paradigms

Data driven exploration
→ **Data Science**

Computer-driven –
Simulation complex phenomenons

Theoretical –
Development of models

Empirical -
Description of natural phenomenons

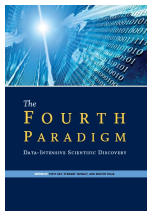


source:http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt

Examples of data sources: data intensive science

"Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets."

"*Modern science increasingly relies on integrated information technologies and computation to collect, process, and analyze complex data.*"



-The Fourth Paradigm – Microsoft

Examples of e-science applications:

- Earth and environment
- Health and wellbeing
  - E.g., The Human Genome Project (HGP)
- Citizen science
- Scholarly communication
- Basic science
  - E.g., CERN

Slide from:http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt

Why Scientific Databases?

- Data enable scientific discovery
  - Data Handling and Archiving (management of massive data resources)
  - Data Discovery (finding data wherever they exist)
  - Data Access (WWW-Database interfaces)
  - Data/Metadata Browsing
  - Data Sharing and Reuse (within project teams; and by other scientists – scientific validation)
  - Data Integration (from multiple sources)
  - Data Fusion (across multiple modalities & domains)
  - Data Mining (Knowledge Discovery in Databases)

From data to knowledge

	Data	Methods	Knowledge
	Call records	Outlier Detection	Detect fraud cases
	Bank transactions	Classification	Customer credibility for loan applications
	Customer transactions from supermarkets/ online stores	Association rules	Which products people tend to buy together?
	Telescope images	Classification	What is the class of a star? E.g., early, intermediate or late formation

What is Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

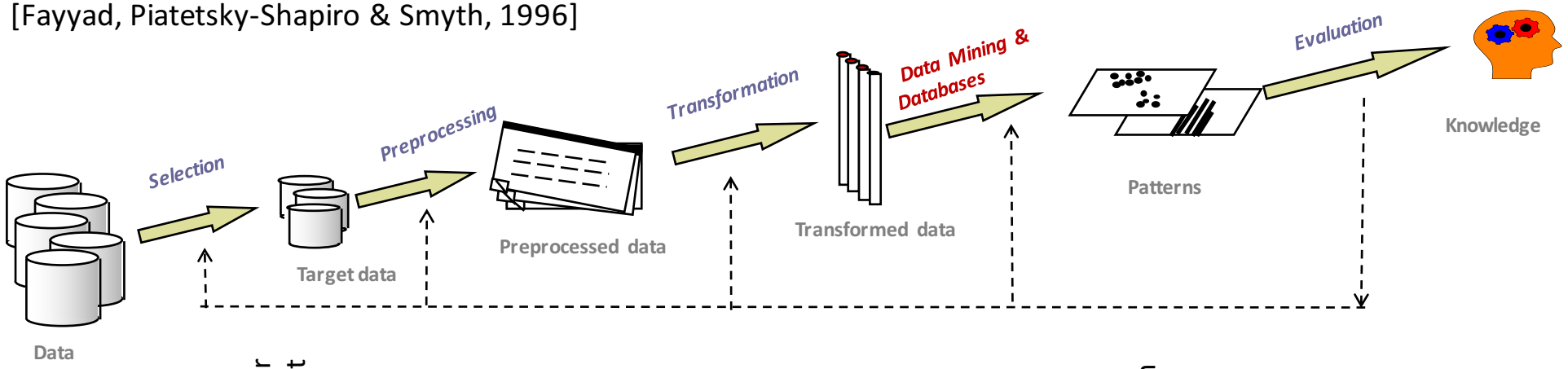
[Fayyad, Piatetsky-Shapiro, and Smyth 1996]

Remarks:

- *valid*: the discovered patterns should also hold for new, previously unseen problem instances.
- *novel*: at least to the system and preferably to the user
- *potentially useful*: they should lead to some benefit to the user or task
- *ultimately understandable*: the end user should be able to interpret the patterns either immediately or after some post-processing

The KDD process and the Databases step

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



Selection:
- Select a relevant dataset or focus on a subset of a dataset
- File / DB/

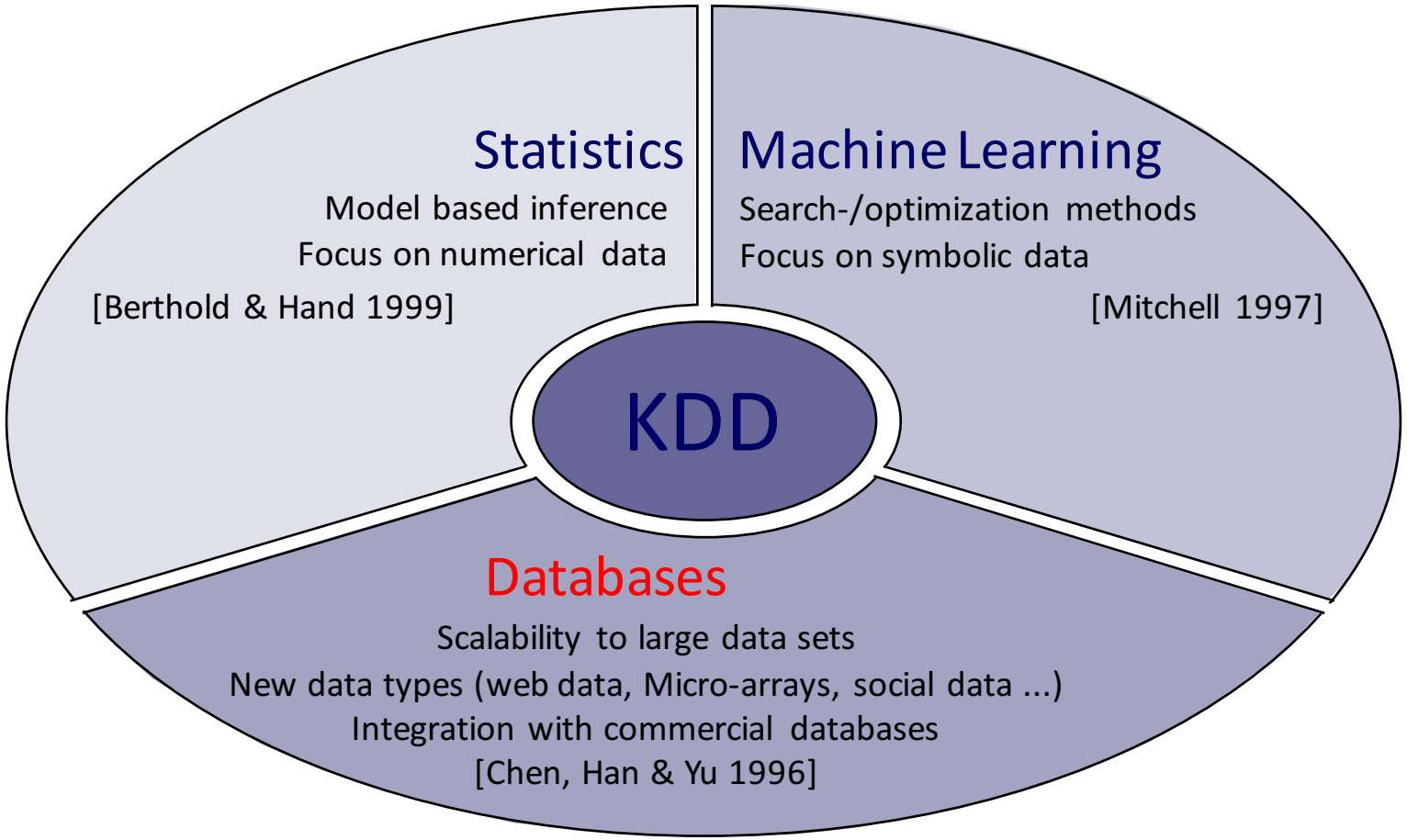
Preprocessing/Cleaning:
- Integration of data from different data sources
- Noise removal
- Missing values

Transformation:
- Select useful features
- Feature transformation/ discretization
- Dimensionality reduction

Data Mining & Databases:
- Search for patterns of interest

Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the Models
- Visualization
- Descriptive Statistics

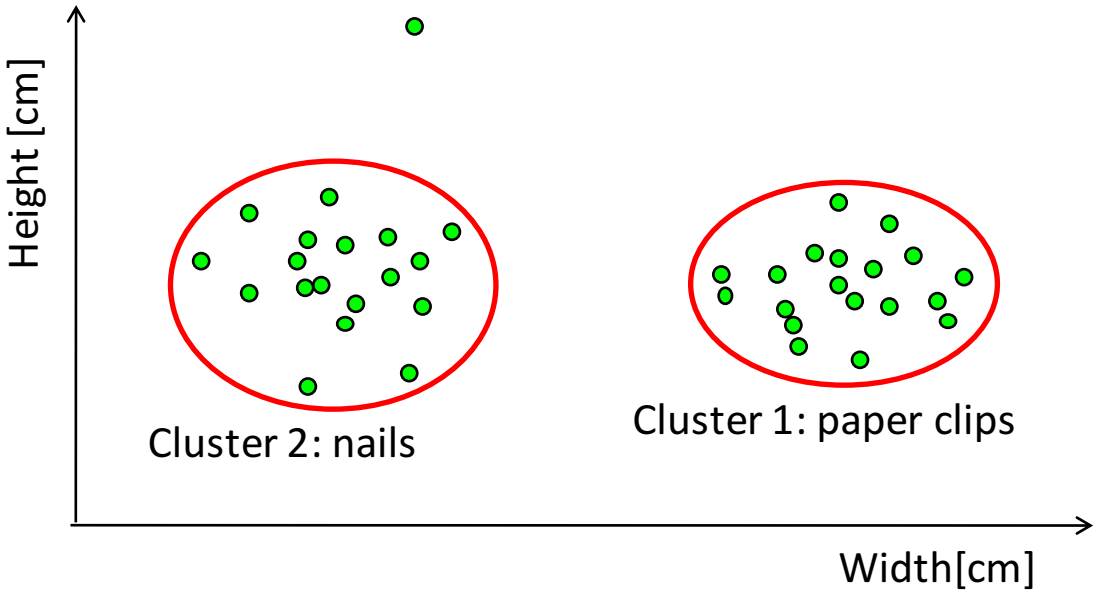
The interdisciplinary nature of KDD



Data Mining (KDD) Examples: Clustering

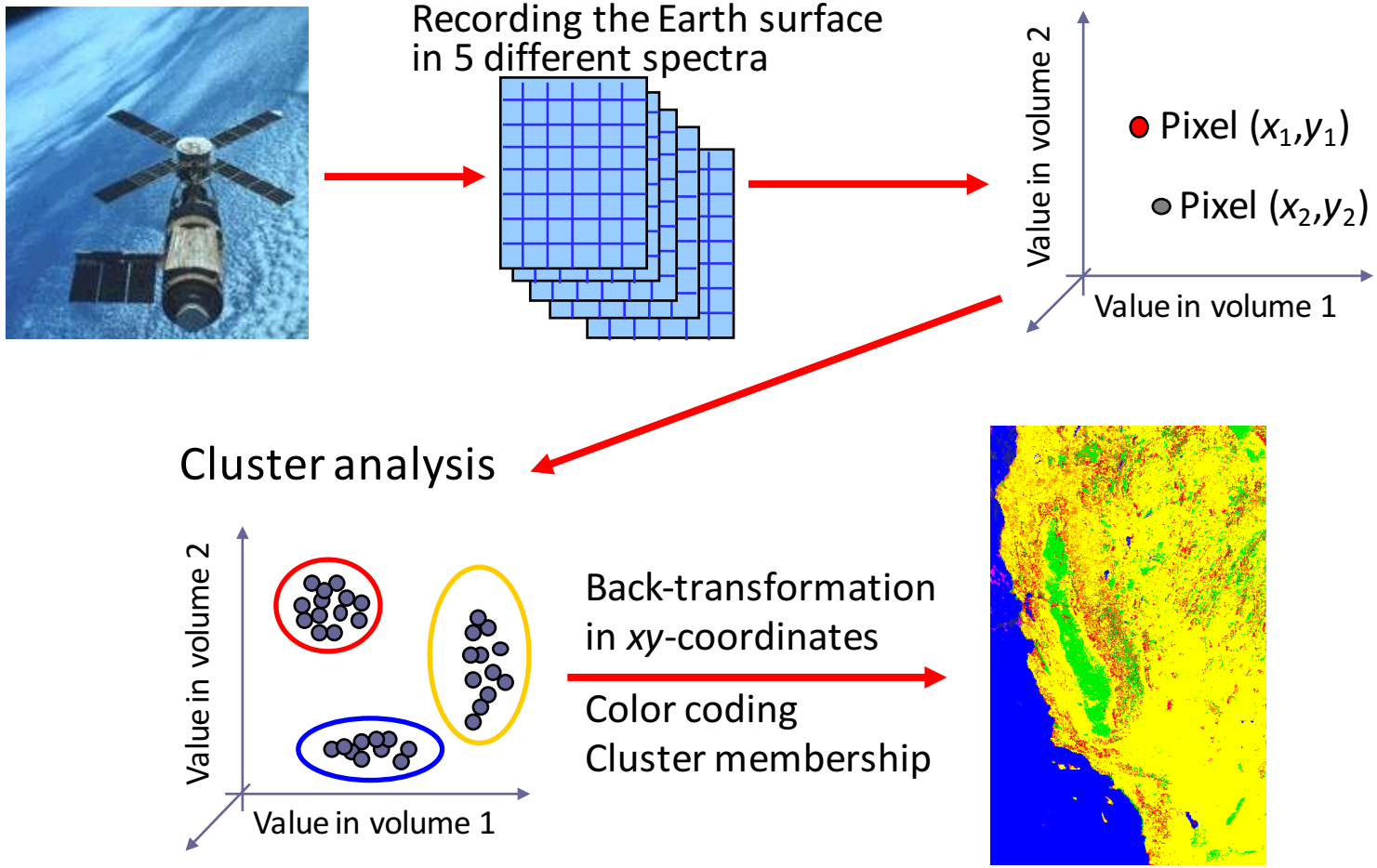
- Clustering can be defined as the decomposition of a set of objects into subsets of similar objects (the so called clusters)
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- The different clusters represent different classes of objects; the number of the classes and their meaning is *not known* in advance.

Clustering: an example



- Each point described in terms of its height and width
- No information on the actual classes (nails, paper clips) is available to the clustering algorithm.

Application: Thematic maps



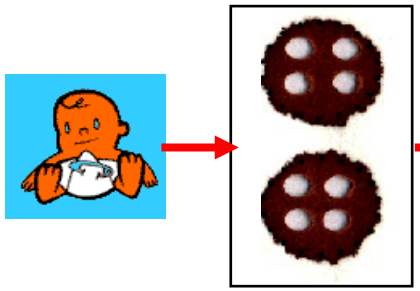
Clustering applications 2/2

Application: Document clustering

- Find groups of documents (topics) that are similar to each other based on the important terms appearing in them.
- Approach:
  - Identify important terms in each document.
  - Form a similarity measure between documents.
  - Cluster based on the similarity measure.
- Gain:
  - Help the end user to navigate in the collection of documents (based on the extracted clusters).
  - Utilize the clusters to relate a new document or search term to clustered documents.
- Check for example, Google News.

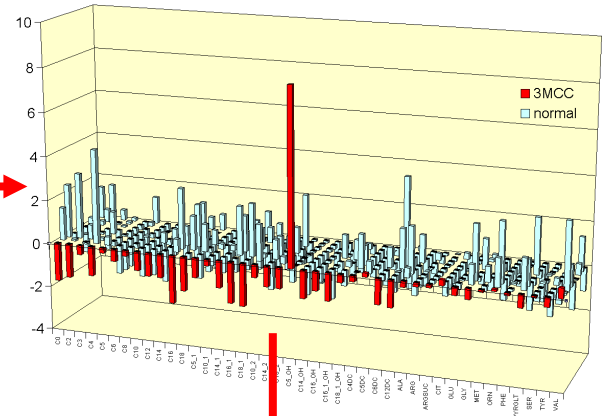
Application: Newborn screening

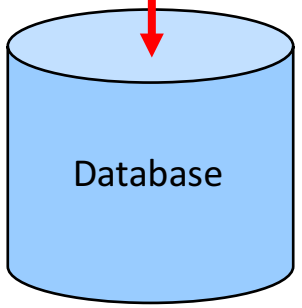
Blood samples of newborns



Mass spectrometry

Metabolite spectrum

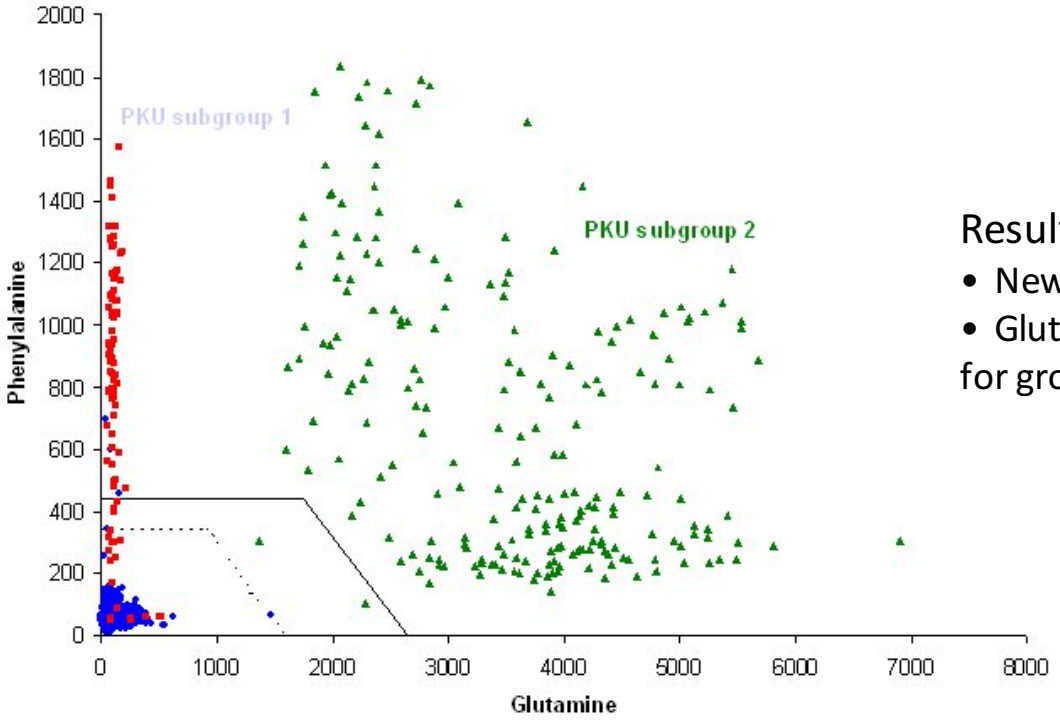




14 analysed amino acids:

alanine
arginine
argininosuccinate
citrulline
glutamate
glycine
methionine
phenylalanine
pyroglutamate
serine
tyrosine
valine
leucine+isoleucine
ornitine

Application: Newborn screening



Result:

- New diagnostic tests
- Glutamine is a new marker for group differentiation

We are now facing a huge problem !

The Tsunami

[Material: Dr. Kirk D. Borne, CDS @ GMU]

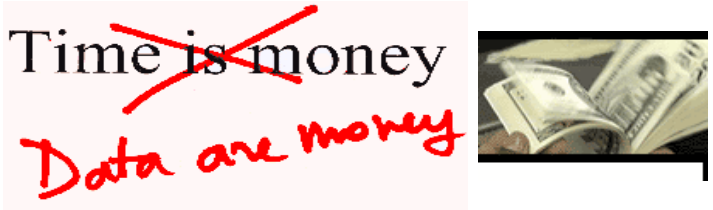
We are now facing a huge problem !

The
Data
Tsunami

[Material: Dr. Kirk D. Borne, CDS @ GMU]

We are now facing a huge problem !

- Huge quantities of data are being generated in all business, government, and research domains:
  - Banking, retail, marketing, telecommunications, health, homeland security, computer networks, other business transactions ...
  - **Scientific data: genomics, astronomy, physics, etc.**
  - Web, text, and e-commerce



[Material: Dr. Kirk D. Borne, CDS @ GMU]

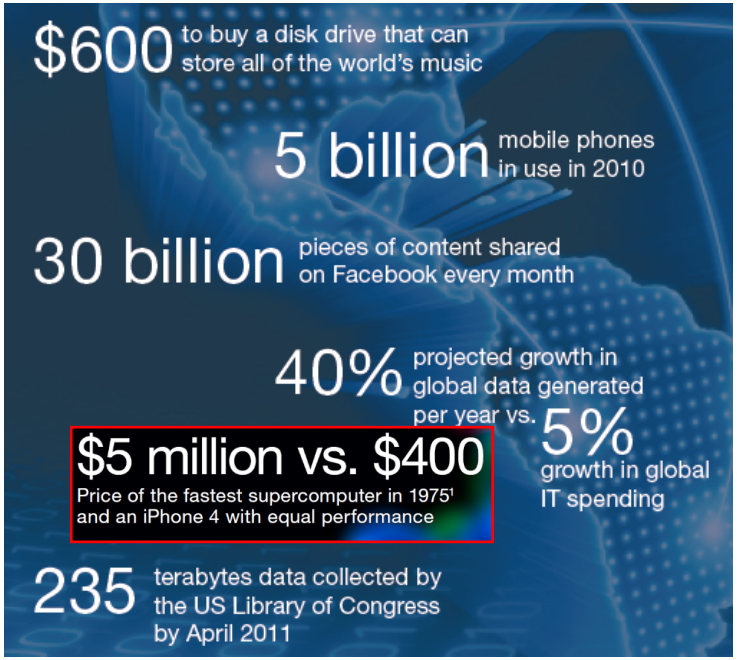
Measuring Data Quantities … to get an idea of data volume

Byte	8 bits	1	one byte = one character (A,B,C...) one bit = 0/1 or Y/N or T/F
Kilobyte	1000 bytes	2^{10}	half a page of text
Megabyte	10^6 bytes	2^{20}	small digital photo, or small book, or 3.5-inch diskette
Gigabyte	10^9 bytes	2^{30}	DVD with broadcast quality movie
Terabyte	10^{12} bytes	2^{40}	50,000 trees made into paper and printed into text
Petabyte	10^{15} bytes	2^{50}	all U.S. academic research libraries
Exabyte	10^{18} bytes	2^{60}	all words ever spoken by human beings throughout all of history

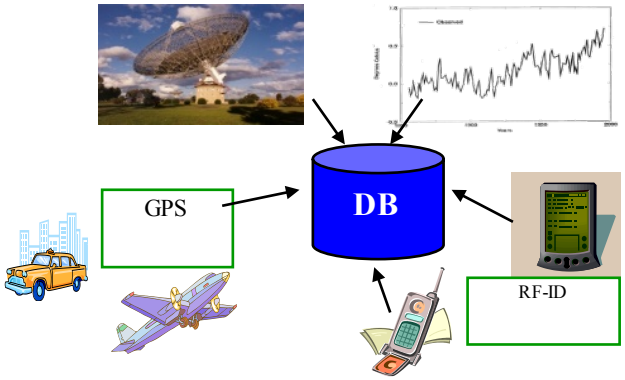
[Material: Dr. Kirk D. Borne, CDS @ GMU]

How much data are there in the world?

- Exponential grows in data



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

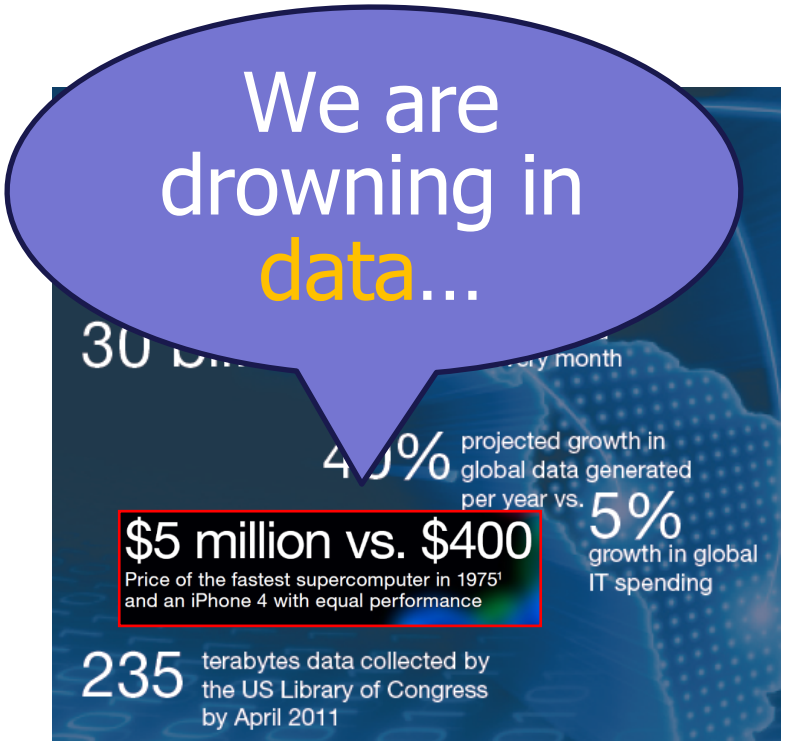


http://www.popsci.com/announcements/article/2011-10/november-2011-data-power

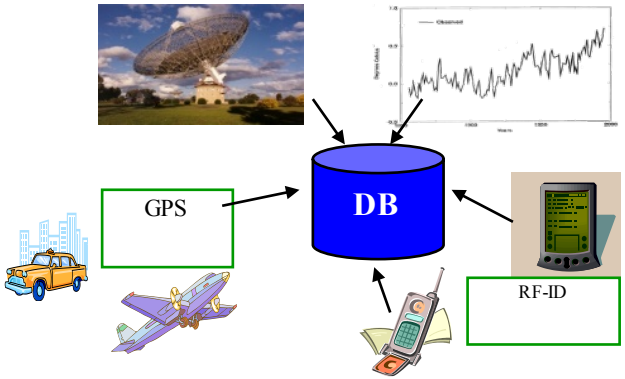
- Data contains value and knowledge

How much data are there in the world?

- Exponential grows in data



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org



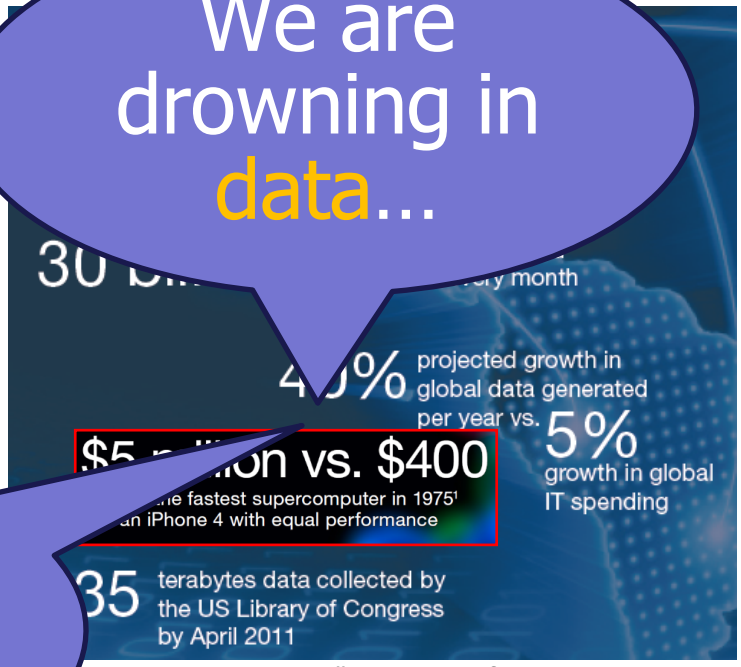
http://www.popsci.com/announcements/article/2011-10/november-2011-data-power

- Data contains value and knowledge

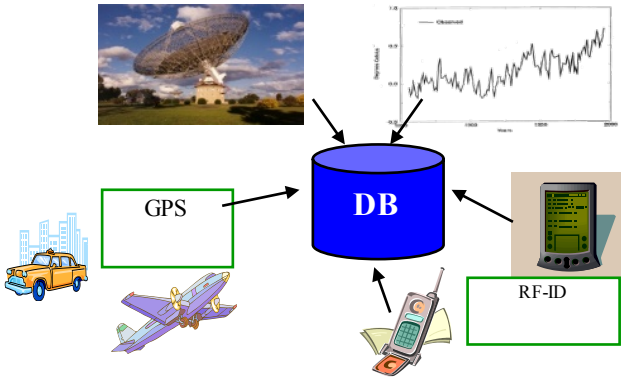
How much data are there in the world?

- Exponential grows in data

We are drowning in data...



kovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, ://www.mmds.org



http://www.popsci.com/announcements/article/2011-10/november-2011-data-power

...but starving for information

- nowledge

How much data are there in the world?

- UC Berkeley 2003 estimate:

  5 exabytes* created in 2002

  - http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

- Updated … 2014 estimate by IDC.com:

2011:	1800 exabytes* (1.8 zettabytes)
2012:	2.8 zettabytes
estimate for 2020:	40 zettabytes

  - http://www.datamation.com/applications/big-data-analytics-overview.html

*** 1 exabyte = 1000 petabytes = 1 million terabytes = 1 billion gigabytes !!**

[Material: Dr. Kirk D. Borne, CDS @ GMU]

How much data are there in the world?

- 2 zettabytes = about 4 trillion CDs of data
- 4 trillion CDs are hard to imagine …

- So, try to visualize just a tiny fraction (**1/7,000,000th**) of that amount …

[Material: Dr. Kirk D. Borne, CDS @ GMU]

How much data are there in the world?

The CD Sea in Kilmington, England
(600,000 CDs)

[Material: Dr. Kirk D. Borne, CDS @ GMU]

Outline

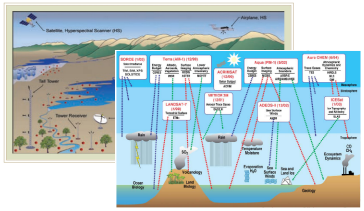
- Course Overview
- Why we need Databases?
- What's next

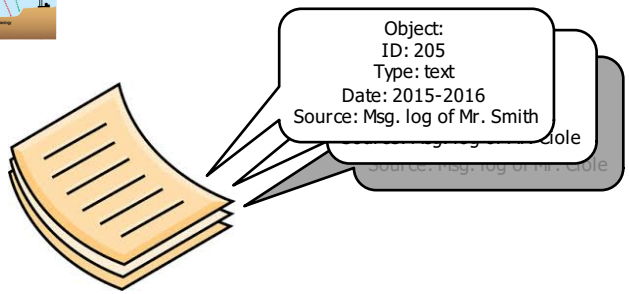
Overview of the lectures (current planning)

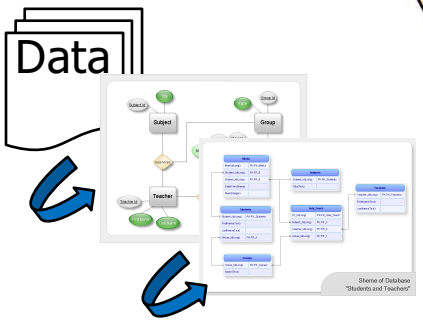
1. Overview
2. Introduction to Scientific Databases
3. Modeling Data and Databases
4. Relational Database Systems
5. Feature spaces and Management of Scientific Data
3. Searching in Scientific Databases
4. Searching in Geo-Spatial Data
5. Searching in Time Sequence Data

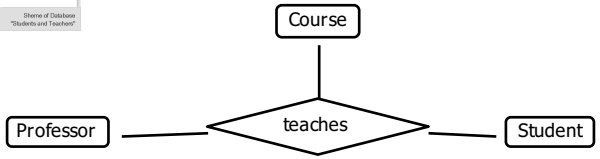
Brief Preview

- Introduction to Databases
  - From Data to Data Management
  - Data, Meta data and Ontologies
  - Data Modeling
    - From data to models to schemes
    - Entity/Relationship Model (E/R-Model)









Brief Preview

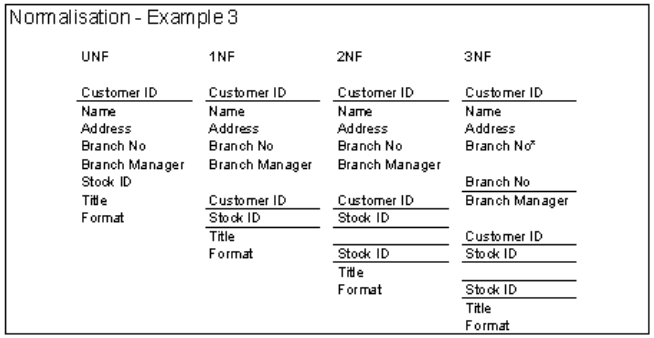
- Relational Databases and Query Languages
  - Relational Algebra

$$\pi_{Student}(\sigma_{PID = "Renz" \wedge PID\ teaches\ SID}(Professor \times Student))$$

  - SQL

```
SELECT Student
FROM Teacher T, Student S
WHERE T.Name = "Renz" and T.Teaches = S.Name
```

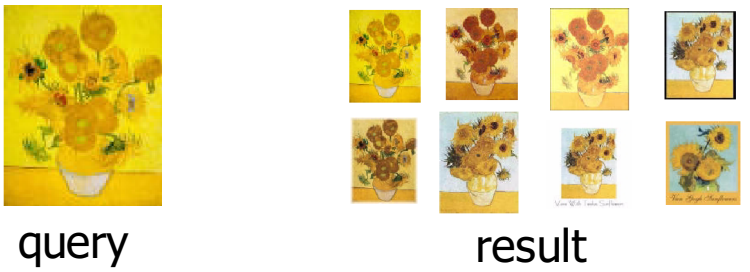
  - Normalisation



Brief Preview

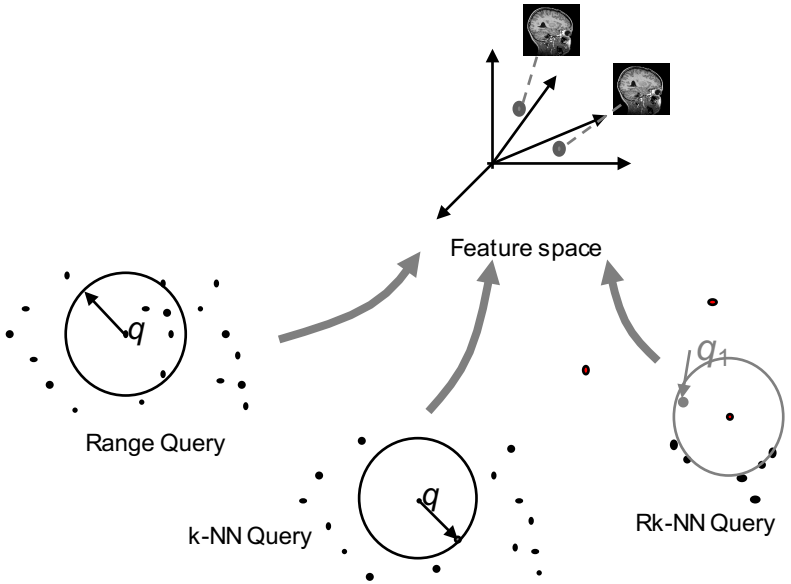
- Scientific Datamanagement

- Searching in Scientific Data

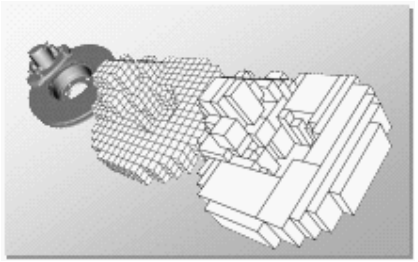


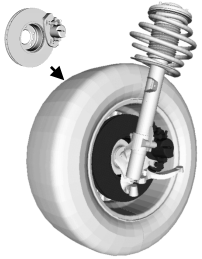
Brief Preview

- Feature-based Similarity Search Methods



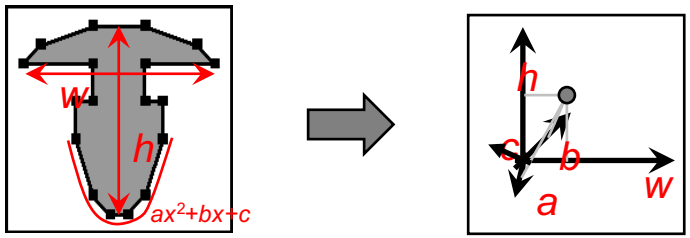
- Spatial Data Management



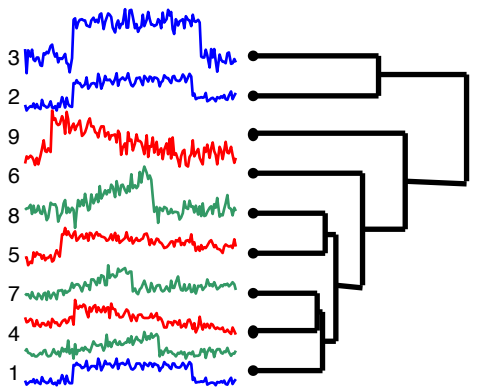


Brief Preview

- Spatial Similarity Models and Similarity Search

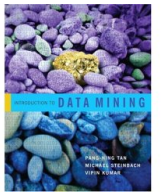


- Time-Series Data Management and Similarity Search

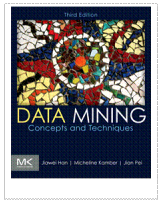


Textbook and recommended readings

- Textbook:
  - Tan P.-N., Steinbach M., Kumar V., *Overview to Databases,* Addison-Wesley, 2006



- Recommended readings
  - Han J., Kamber M., Pei J., *Databases: Concepts and Techniques,* 3rd ed., Morgan Kaufmann, 2011



Online resources

- *Overview to Databases* class by Jennifer Widom, Stanford
  - http://www.db-class.org/course/auth/welcome

- Kdnuggets: Databases and Analytics resources
  - http://www.kdnuggets.com/