Chapter XI: Data Mining Outro

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2011/12

Chapter XI: Data Mining Outro

1. Data mining's position in sciences
2. Using data mining in practice
   - 2.1. KDD process
3. Data mining applications
   - 3.1. In biology
   - 3.2. In medicine
   - 3.3. In economy
   - 3.4. In the Internet
   - 3.5. In secret services
4. Data mining and privacy

Data mining's position in sciences

- Data mining uses statistical tools and methods to infer from data
  - Is data mining just fancy name for statistics?
- Data mining uses methods to learn unseen
  - Is data mining just boring name for machine learning?

Data mining's position in sciences

- Data mining uses statistical tools and methods to infer from data
  - Is data mining just fancy name for statistics?
- Data mining uses methods to learn unseen
  - Is data mining just boring name for machine learning?

Is data mining a voodoo science?

Data mining vs. the scientific method

- The scientific method:
  - Form a hypothesis
  - Collect the data
  - Test the hypothesis
- The data mining method:
  - Get the data
  - Select a data mining method that makes sense in your data
  - Apply the method to the data

Data mining vs. the scientific method

- The scientific method:
  - Form a hypothesis
  - Collect the data
  - Test the hypothesis
- The data mining method:
  - Get the data
  - Select a data mining method that makes sense in your data
  - Apply the method to the data

Selects a "family" of hypotheses

Data mining vs. the scientific method

- The scientific method:
  - Form a hypothesis
  - Collect the data
  - Test the hypothesis
- The data mining method:
  - Get the data
  - Select a data mining method that makes sense in your data
  - Apply the method to the data

Selects a "family" of hypotheses

Finds the "valid" hypotheses for the data

The voodoo science

The response from several social scientists has been rather unappreciative along the following lines: “Where is your hypothesis? What you’re doing isn’t science! You’re doing DATA MINING !”

http://andrewgelman.com/2007/08/a_rant_on_the_v/

Data mining vs. statistics

- Statistics provides tools to validate the hypotheses
- Data mining generates the hypotheses
- But data mining uses tools from statistics
  - Toolbox of mathematical methods
  - Validation of results
  - Formalization of methods

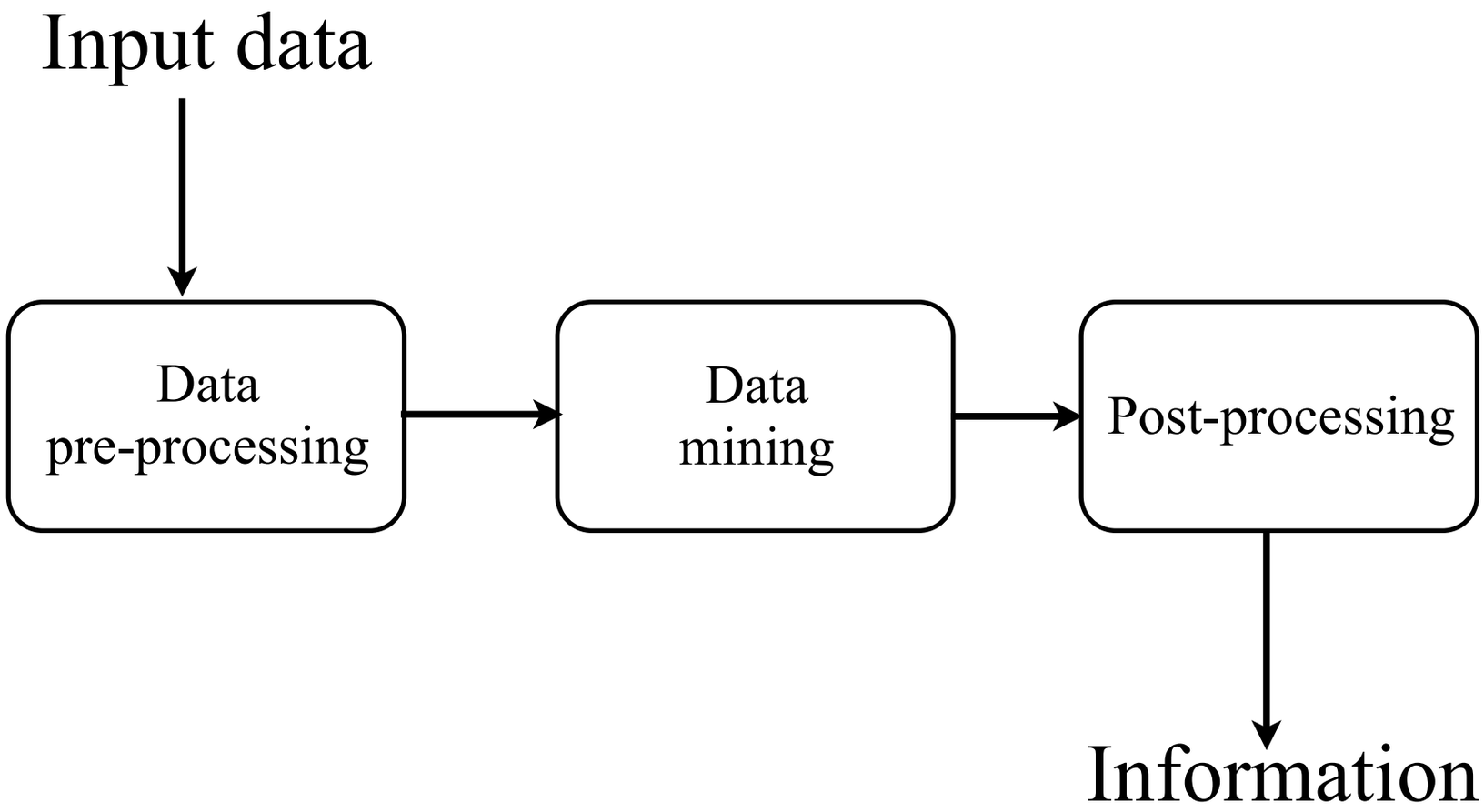
Data mining vs. machine learning

- Data mining uses machine learning methods
  - An application
  - More practical issues
    - Missing values
    - Odd correlations
    - Scalability
    - Domain knowledge
- No clear difference between developing data mining methods vs. developing machine learning methods

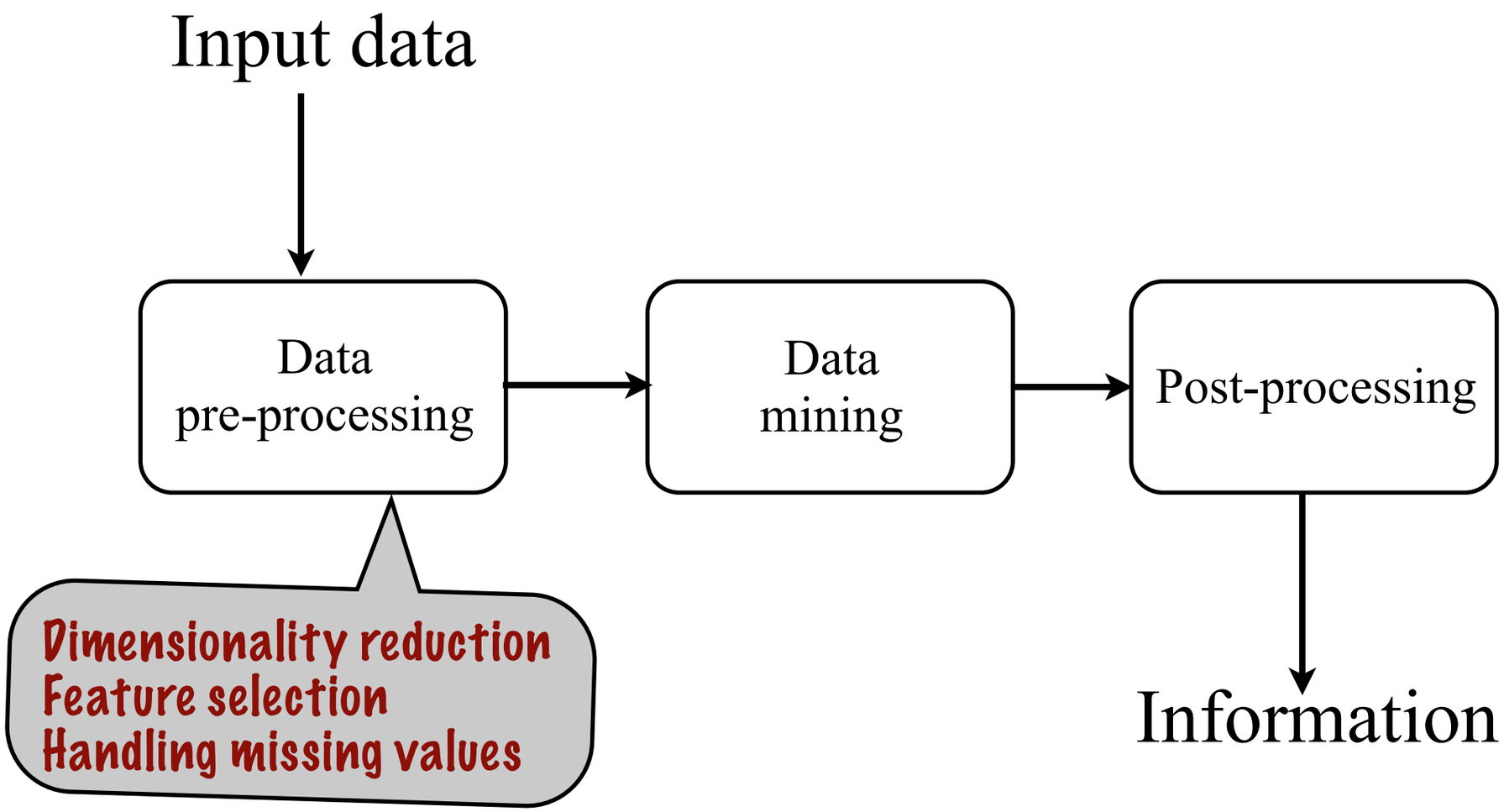
Data mining in practice

- Real world is a messy place
  - Real-world data is even messier
  - Data needs pre-processing
- Applications have (hopefully) domain experts
  - Domain knowledge should be incorporated
  - Domain experts should be able to interpret the results
    - Not too many results
    - Post-processing

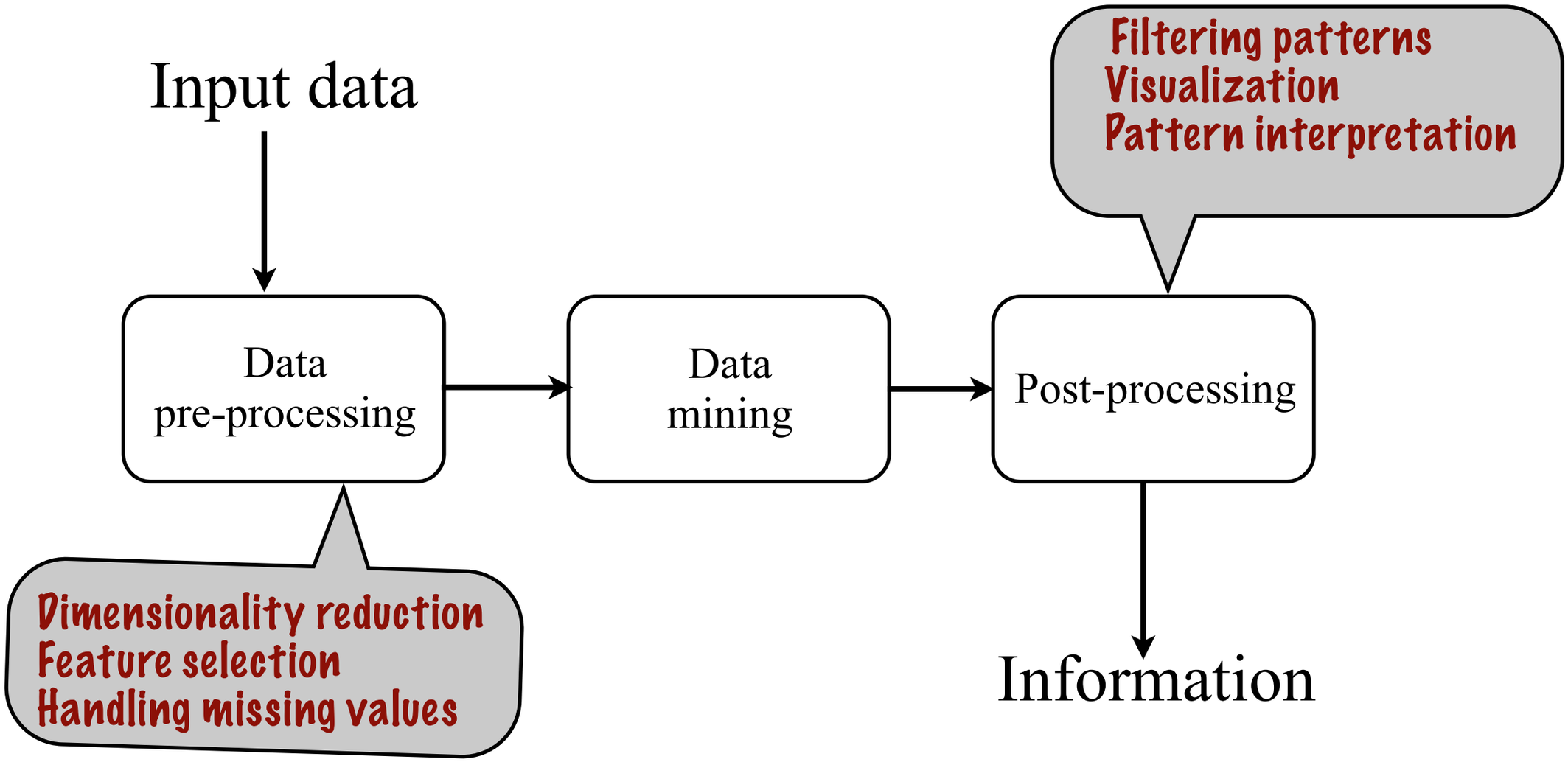
The KDD process



The KDD process



The KDD process



Data pre-processing

- Garbage in, garbage out
- Many issues
  - What to do with missing values
    - Are missing values clearly marked?
  - What's the dimensionality vs. sample size
    - Anyway, which way the observations are?
  - Do some features correlate with each other in an uninteresting way
    - Record ID and class label
  - Is data type suitable for our algorithm
    - Binary, categorical, numerical
  - And many, many more…

Post-processing

- Humans can only interpret so many results
  - Computers are a different thing
- Select top-*k* results
  - What criteria?
- Are the results significant?
  - Statistics
- Are the results meaningful?
  - Domain expert
- Visualization
- Humans are great at finding patterns (even when they don't exist)
  - Computers are a different thing

Leakage

- *Leakage in data mining* refers to the case when prediction algorithm learns from data is should not have access to
  - Problem as the quality is assessed using already-historical test data
  - E.g. INFORMS'10 challenge: predict the value of a stock
    - Exact stock was not revealed
    - But "future" general stock data was available!
      ⇒ 99% AUC (almost perfect prediction!)
  - More subtle one's exist
    - E.g. removing a crucial feature creates a new type of correlation

Data mining applications

- Data mining is quite commonplace
  - Often not called like that
  - Sometimes something other is meant by data mining: ”An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results” —W.S. Brown “Introducing Econometrics”

- Many sciences are turning into *data-driven sciences*
  - How to deal with all the data obtained?

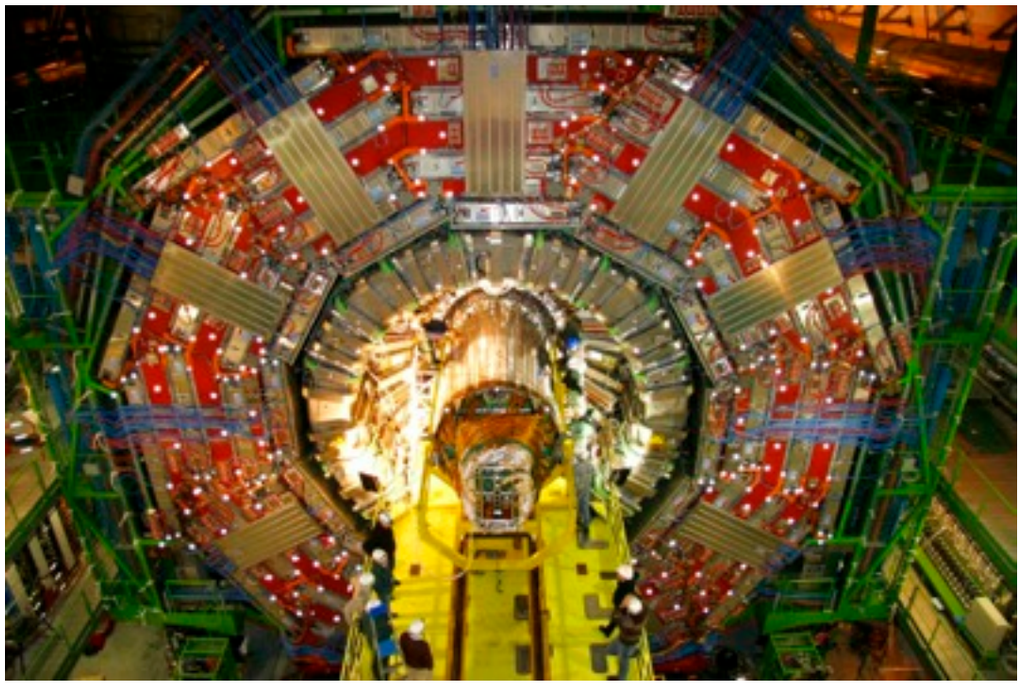


Image: CERN

Data mining and biology

- Genome databanks
  - Identify genes (pattern mining)
  - Identify groups of related genes (graph mining)
- Protein activity
  - Developmental biology (Episode mining)
- Protein-protein interactions
  - Thousands of different proteins
  - Proteins have different roles in different situations in different compartments

Data mining and medicine

- Hospitals
  - Patient diagnoses (decision trees)
  - House, M.D., in a computer (probability estimation)
- Pharmaceutical companies
  - Drug development (much like bioinformatics)
  - Not all drug prototypes can be tested
    - Too many
    - Potentially lethal
    - Learning

Data mining and economy

- Recommender systems
  - Netflix, Amazon, well, anybody
- User segments
  - Clustering
- Machine learning for stock prices
- Part of algorithmic trading
  - Faster than humans
  - Not prone to human errors?
    - But…

Data mining and the Internet

- Social networks (FB, LinkedIn, MySpace, …)
  - Social scientists love!
  - Link prediction
  - Supply recommender systems
- Searching
  - Ad targeting
  - User profiling

Data mining and secret services

- Terrorist profiling/detection

Data mining and privacy

- Very important!
  - Everybody wants to do data mining, nobody wants to be data mined
- Often imposed by laws
  - Medical data, personal information records, …
- *Privacy-preserving data mining*
  - Data provider anonymizes data
  - Data miner does not know (and can't learn) the identities of the data entities
  - Hard to guarantee
    - Leakage!

Preserving privacy in data mining

- Remove sensitive features
  - Can be re-mapped using publicly available data
- *k*-anonymity
  - All released records are similar to at least *k* other records in the released features
    - Homogenous sensitive data can still be learned
- Differential privacy
  - *Differentially private algorithm* will behave (approximately) the same in two data sets that differ only on a tiny subset
    - Presence or absence of single individual don't matter

~~Summary~~ Picture of data mining

Image: Wikipedia