Machine learning, artificial intelligence, data mining, …

1. LECTURE

assoc. prof. Branko Kavšek

Lecture outline

- Introduction: data flood
- Data mining application examples
- Data mining & knowledge discovery
- Data mining tasks

Trends leading to data flood

- More data is generated:
  - In business:
    bank, telecom, other business transactions, …
  - In science:
    astronomy, biology, chemistry, …
  - On the web:
    social networks, e-commerce, …

Big data examples (18 years ago)

- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
  - storage and analysis is a big problem;

- AT&T handles **billions** of calls **per day**
  - so much data, it cannot be all stored – analysis has to be done "on the fly", on streaming data;

Largest databases in 2003

- Commercial databases (Winter Corp. 2003 survey):
  - France Telecom has largest decision-support DB = ~30TB;
  - AT&T has database = ~26 TB;

- Web:
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - IBM WebFountain, 160 TB
  - Internet Archive (www.archive.org), ~300 TB;

From terabytes to exabytes to …

- UC Berkeley - estimate: **5 exabytes**
  (5 million terabytes) of new data was created in 2002.

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

- **US** produces **~40% of new stored data worldwide**.

- 2006 estimate: **161 exabytes** (IDC study)

  www.usatoday.com/tech/news/2007-03-05-data_N.htm

- 2010 projection: **988 exabytes**.

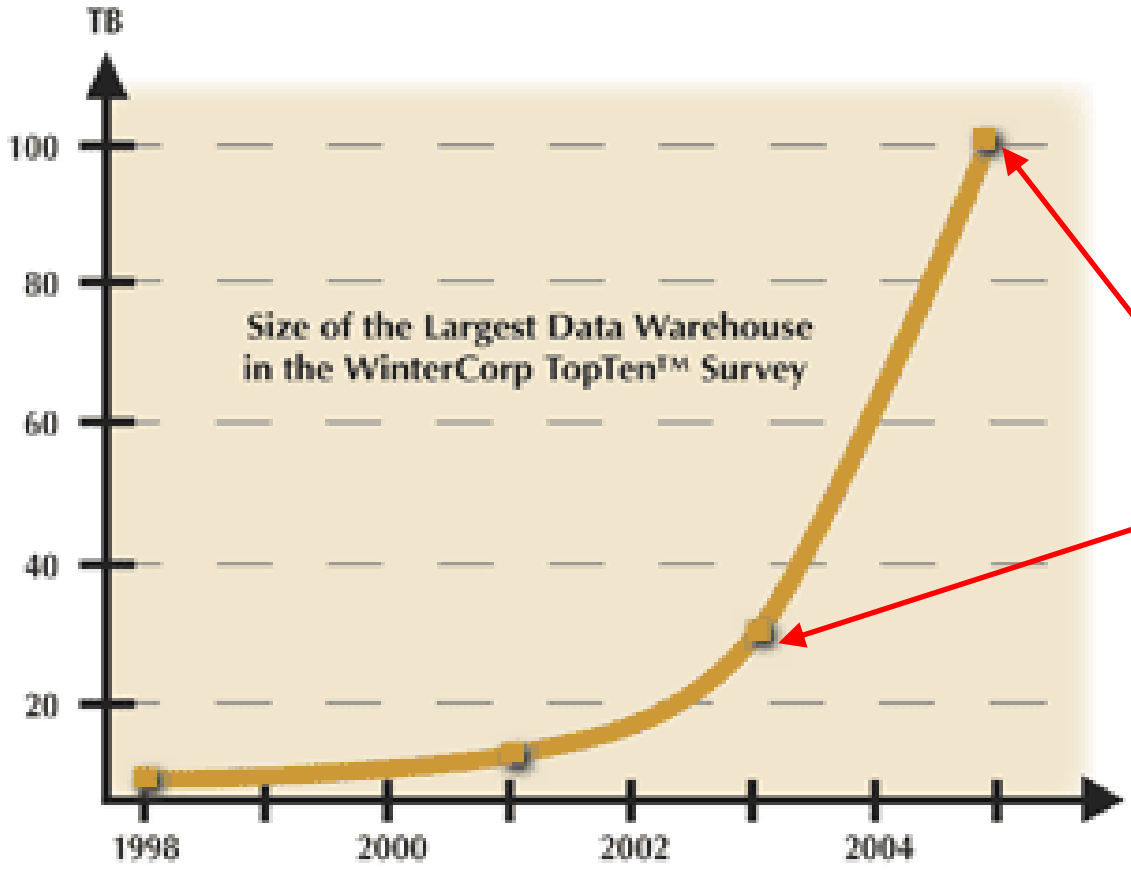
Largest databases in 2005

Winter Corp. 2005 commercial DB survey:

1. Max Planck Inst. for Meteorology: **222 TB**
2. Yahoo: **~100 TB** (largest data warehouse)
3. AT&T: **~94 TB**

http://dssresources.com/news/1010.php

Data growth



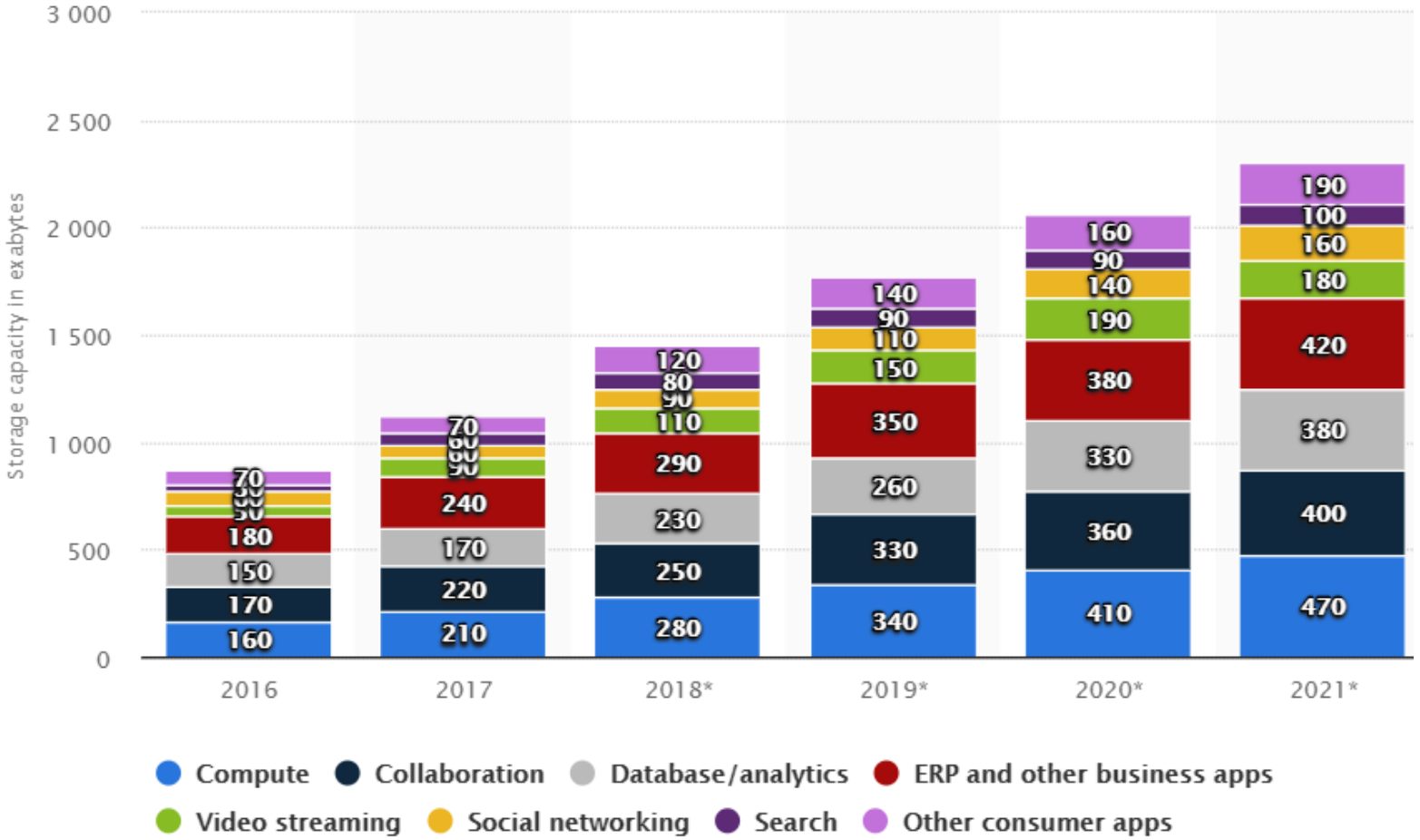
In 2 years,
the size of the
largest database
TRIPLED !!!

The present situation

Example:
end of June, 2017 – CERN‘s data center stores more than **200 petabytes** of data (200 million gigabytes)

https://en.wikipedia.org/wiki/Zettabyte_Era

Data center storage capacity worldwide from 2016 to 2021, by segment



Data growth rate

- In the year 2002 – **2 times more** data has been produced than in the year 1999;
- In the year 2005 – **3 times more** data has been produced than in the year 2003;
- Very little data will ever be looked at by a human;

Knowledge Discovery and Data Mining
are **NEEDED** to make sense and use of data !!!

Lecture outline

- Introduction: data flood
- Data mining application examples
- Data mining & knowledge discovery
- Data mining tasks

Machine learning/data mining application areas

Science:

- astronomy, bioinformatics, drug discovery. …

Business:

- CRM (Customer Relationship management), fraud detection, e-commerce, manufacturing, sports/entertainment, telecom, targeted marketing, health care, …

Web:

- search engines, advertising, web and text mining, …

Government:

- surveillance, crime detection, profiling tax cheaters, …

Application areas

What do you think are some of the most important and widespread business applications of **data mining**?

Data mining for customer modeling

- Attrition prediction,
- targeted marketing (cross-sell, customer acquisition),
- credit-risk,
- fraud detection,
- banking,
- telecom,
- retail sales, …

Customer attrition: case study

- Situation:
  Attrition rate for mobile phone customers is around 25-30% a year (US data)!

- With this in mind, what is the DM task?
  - Assumption: we have customer information for the past N months.

Customer attrition: case study (2)

Task:

- Predict who is likely to attrite next month.
- Estimate customer value and what is the cost-effective offer to be made to this customer.

Customer attrition: results

- Verizon Wireless built a customer data warehouse;
- Identified potential attriters;
- Developed multiple, regional models;
- Targeted customers with high propensity to accept the offer;
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)

(Reported in 2003)

Assessing credit risk: case study

Situation: Person applies for a loan.

Task: Should a bank approve the loan?

Note:
People who have the best credit don't need the loans, and people with worst credit are not likely to repay
➔ bank's best customers are "in the middle".

Credit risk: results

- Banks develop credit models using variety of machine learning methods,
- mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan,
- widely deployed in many countries.

e-commerce

A person buys a book (product) at Amazon.com

What is the task?

Successful e-commerce: case study

Task:
Recommend other books (products) this person is likely to buy.

Amazon does clustering based on books bought:

- Customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**".

Recommendation program is quite successful.

Unsuccessful e-commerce: case study (KDD-Cup 2000)

Data:
clickstream and purchase data from Gazelle.com, legwear and legcare e-tailer.

Question:
Characterize visitors who spend
more than $12 on an average order at the site.

Dataset = 3,465 purchases, 1,831 customers,

Very interesting analysis by Cup participants

- thousands of hours – $X,000,000 (Millions) of consulting,

Total sales: -$Y,000,

Obituary: Gazelle.com out of business, Aug 2000.

Genomic microarrays: case study

Given microarray data for a number of samples (patients), can we:

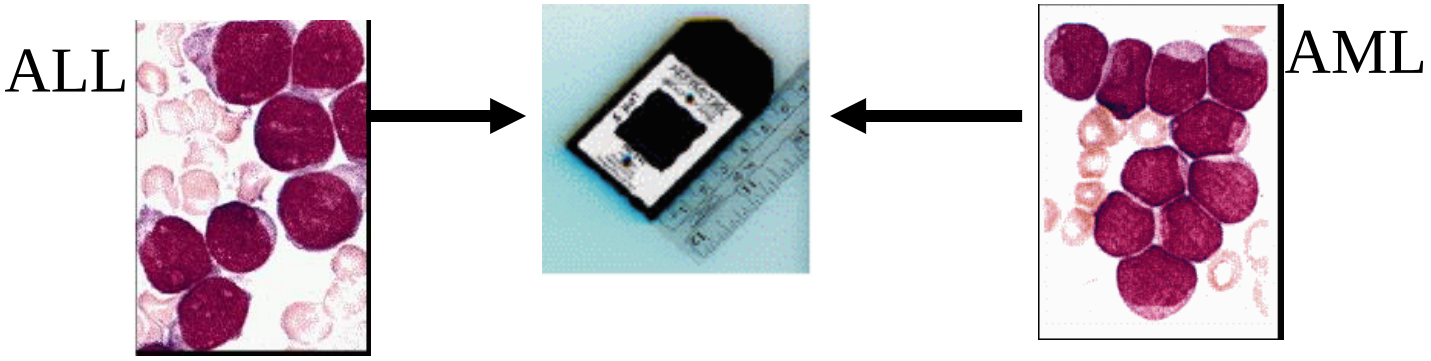
- accurately diagnose the disease?
- predict outcome for given treatment?
- recommend best treatment?

Example: ALL/AML data

38 training cases, 34 test cases, ~7,000 genes

2 classes: Acute Lymphoblastic Leukemia (ALL) vs
Acute Myeloid Leukemia (AML)

Use train data to build diagnostic model

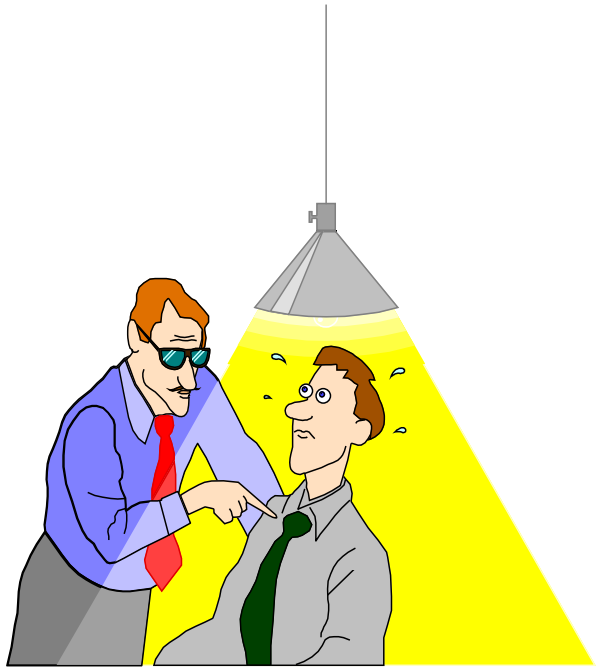


Results on test data:

33/34 correct, 1 error (may be mislabeled)

Security and fraud detection: case study

- Credit card fraud detection
- Detection of money laundering
  - FAIS (US Treasury)
- Securities fraud
  - NASDAQ KDD system
- Phone fraud
  - AT&T, Bell Atlantic, British Telecom/MCI
- Bio-terrorism detection at Salt Lake Olympics 2002



Data mining and privacy

- In 2006, NSA (National Security Agency) was reported to be mining years of call info, to identify terrorism networks;
- Social network analysis has a potential to find networks;
- Invasion of privacy – do you mind if your call information is in a government database?
- What if NSA program finds one real suspect for 1,000 false leads? 1,000,000 false leads?

Problems suitable for data mining

- require knowledge-based decisions
- have a changing environment
- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!

Privacy considerations are important if personal data is involved !!!

Lecture outline

- Introduction: data flood
- Data mining application examples
- Data mining & knowledge discovery
- Data mining tasks

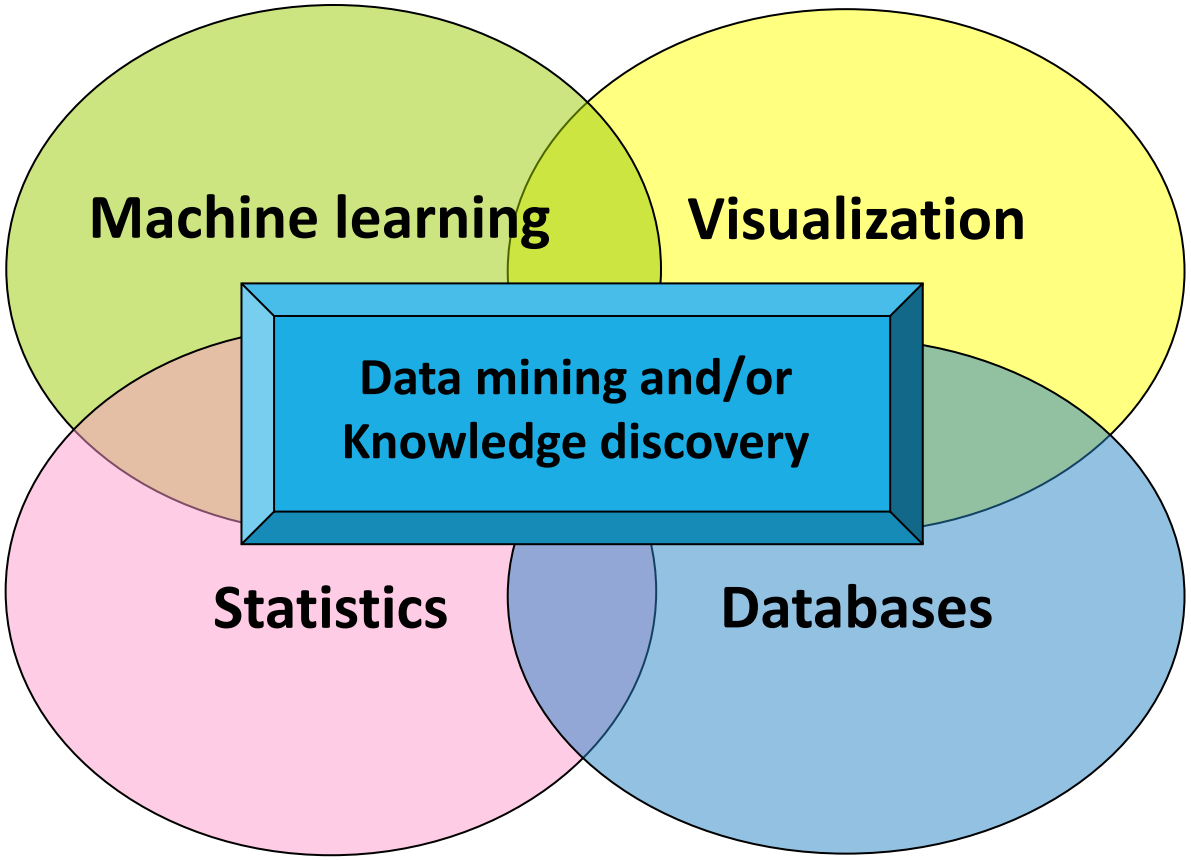
Definition of "knowledge discovery"

Knowledge discovery in data is the *non-trivial* process of identifying:

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

From:
Advances in Knowledge Discovery and Data Mining, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996.

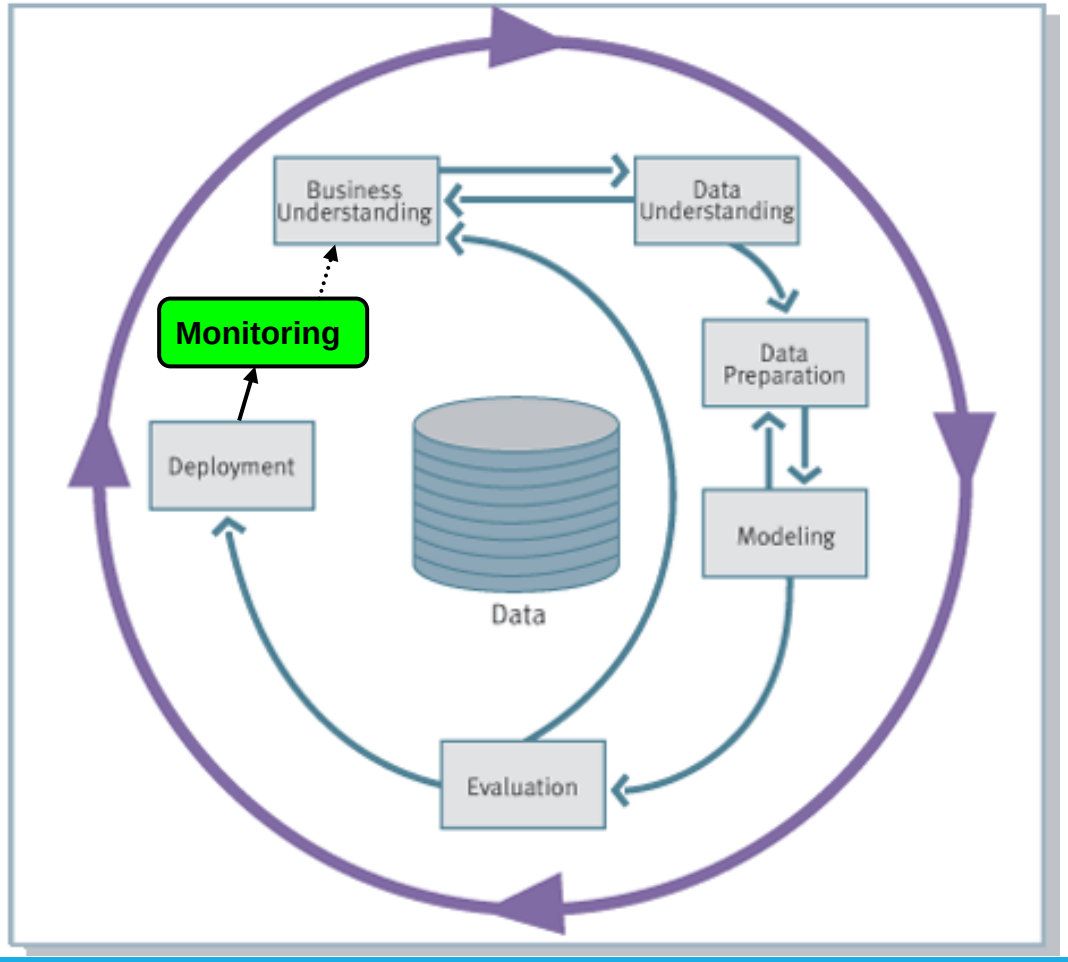
Related fields



Statistics, machine learning and data mining

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning:
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data mining and/or Knowledge discovery in data:
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data "cleaning", learning, integration and visualization of results
- **Distinctions are "fuzzy"**.

Knowledge discovery process flow – according to CRISP-DM



See also:
https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
for more detailed information

Historical note: the many names of data mining

Data Fishing, Data Dredging: 1960 –

- used by statisticians (considered as a bad name);

Data Mining: 1990 –

- used DB, business;
- in 2003 – bad image because of TIA;

Knowledge Discovery in Databases: 1989 –

- used by AI, machine learning community;

also:
Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction ...

Currently:
Data Mining and **Knowledge Discovery**
are used interchangeably (as synonyms).

Lesson outline

- Introduction: data flood
- Data mining application examples
- Data mining & knowledge discovery
- Data mining tasks

Major data mining tasks

Classification: predicting an item class

Clustering: finding clusters in data

Associations: e.g. A & B & C occur frequently

Visualization: to facilitate human discovery

Summarization: describing a group

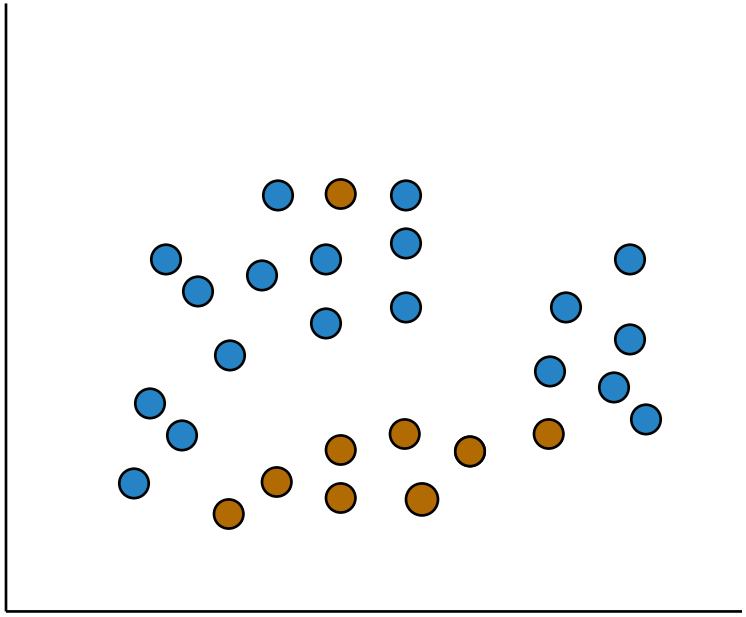
Deviation detection: finding changes

Regression/estimation: predicting a continuous value

Link analysis: finding relationships

Data mining tasks: prediction (classification)

Learn a method for predicting the instance class from pre-labeled (classified) instances

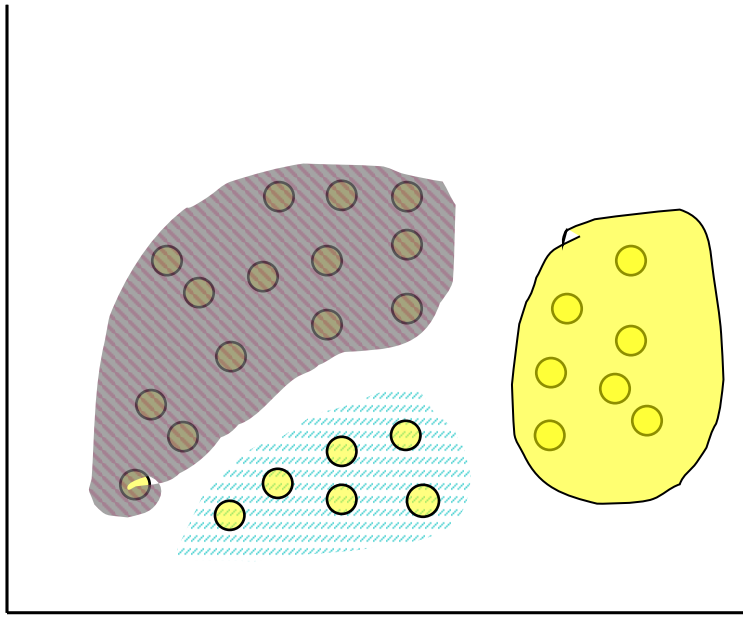


Many approaches:

statistics,
decision trees,
neural networks, ...

Data mining tasks: clustering

Find "natural" grouping of instances given un-labeled data



Summary

- Technology trends lead to data flood
  - data mining is needed to make sense of data;
- Data Mining has many applications, successful and not;
- Knowledge discovery is a process;
- Data mining tasks
  - classification, clustering, …