Machine Learning, Data Mining, and Knowledge Discovery: An Introduction

AHPCRC Workshop - 8/17/10 - Dr. Martin

Based on slides by Gregory Piatetsky-Shapiro from Kdnuggets

http://www.kdnuggets.com/data_mining_course/

Course Outline

- Machine Learning
  - input, representation, decision trees, other learning algorithms
- Weka
  - machine learning workbench
- Data Mining
  - associations, deviation detection, clustering, visualization
- Case Studies
  - targeted marketing, genomic microarrays
  - Data Mining, Privacy and Security
- Final Project: Microarray Data Mining Competition

Lesson Outline

- **Introduction: Data Flood**
- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- Data Mining Tasks

Trends leading to Data Flood

- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific data: astronomy, biology, etc
  - Web, text, and e-commerce
- Much faster than out ability to analyze it in a useful or meaningful way

Big Data Examples

- 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 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 ~ 26 TB
- Web
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - IBM WebFountain, 160 TB (2003)
  - Internet Archive (www.archive.org),~ 300 TB

From terabytes to exabytes to …

- UC Berkeley 2003 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

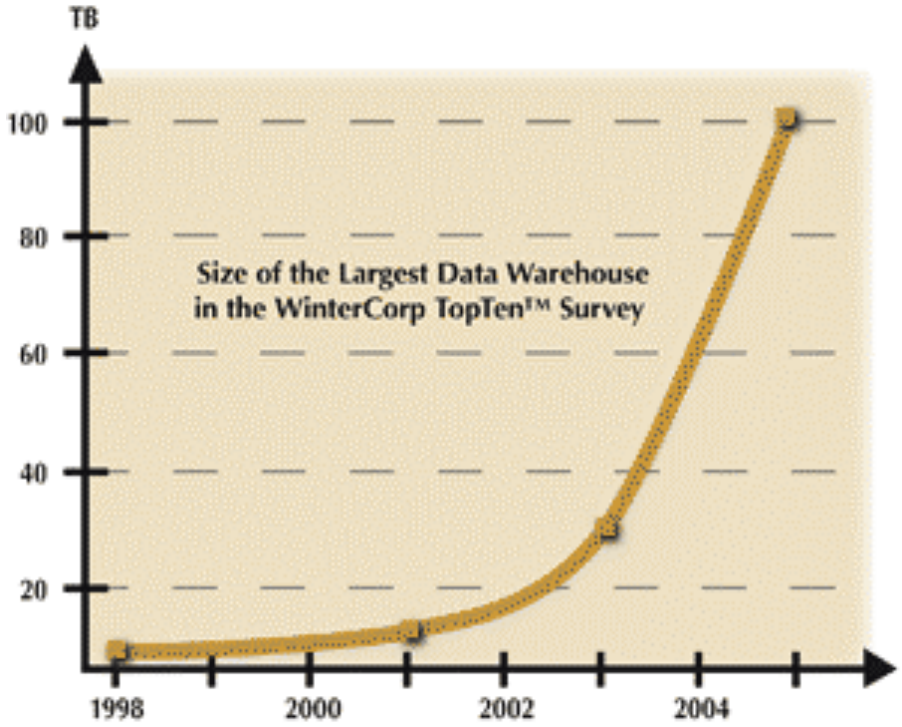
Largest Databases in 2005

Winter Corp. 2005 Commercial Database Survey:

- Max Planck Inst. for Meteorology , 222 TB
- Yahoo ~ 100 TB (Largest Data Warehouse)
- AT&T ~ 94 TB

http://www.wintercorp.com/vldb/2005_topten_survey/toptenwinners_2005.asp

Data Growth



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

Data Growth Rate

- Twice as much information was created in 2002 as in 1999 (~30% growth rate)
- Other growth rate estimates even higher
- Very little data will ever be looked at by a human

Knowledge Discovery is **NEEDED** to make sense and use of data.

Lesson 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, recommender systems, spam filtering …
- 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

- Customer Tasks:
  - attrition prediction
  - targeted marketing:
    - cross-sell, customer acquisition
  - credit-risk
  - fraud detection
- Industries
  - banking, telecom, retail sales, …

Customer Attrition: Case Study

- Situation: Attrition rate at for mobile phone customers is around 25-30% a year!
- With this in mind, what is our task?
  - Assume we have customer information for the past N months.

Customer Attrition: Case Study

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
- Q: Characterize visitors who spend more than $12 on an average order at the site
- Dataset of 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
- Google "kdd cup 2000 gazelle"

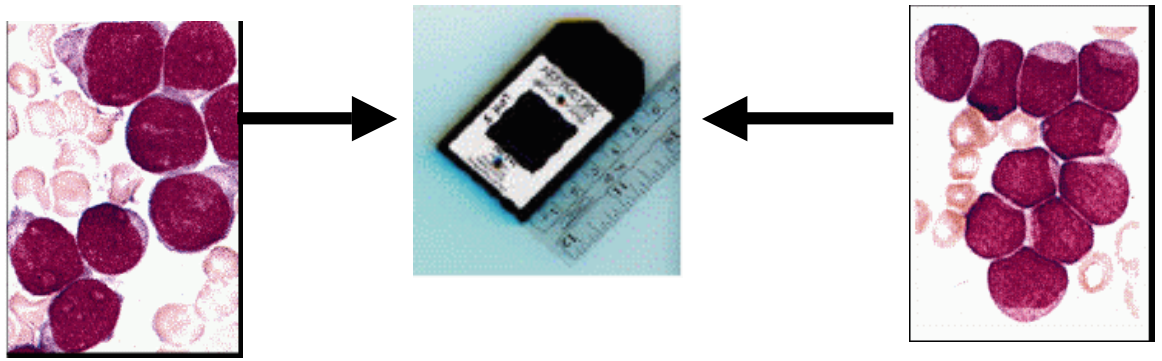
Genomic Microarrays – Case Study

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

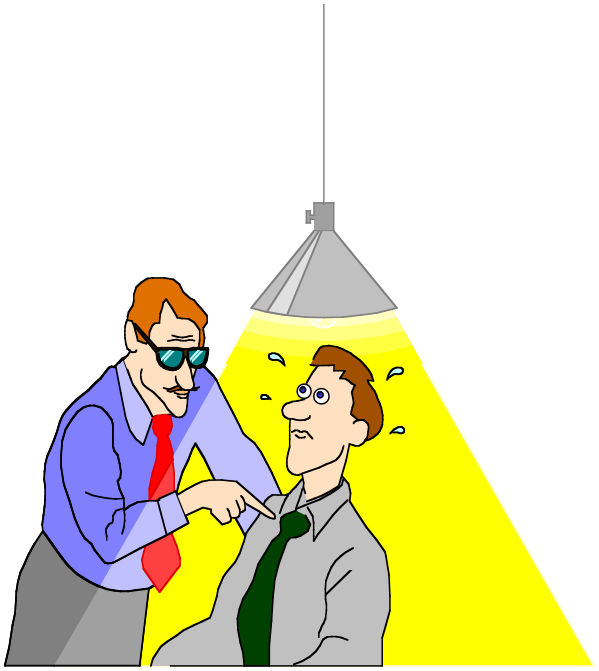
- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model

ALL  AML

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 gov 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 important if personal data is involved

Lesson Outline

- Introduction: Data Flood
- Data Mining Application Examples
- **Data Mining & Knowledge Discovery**
- Data Mining Tasks

Knowledge Discovery Definition

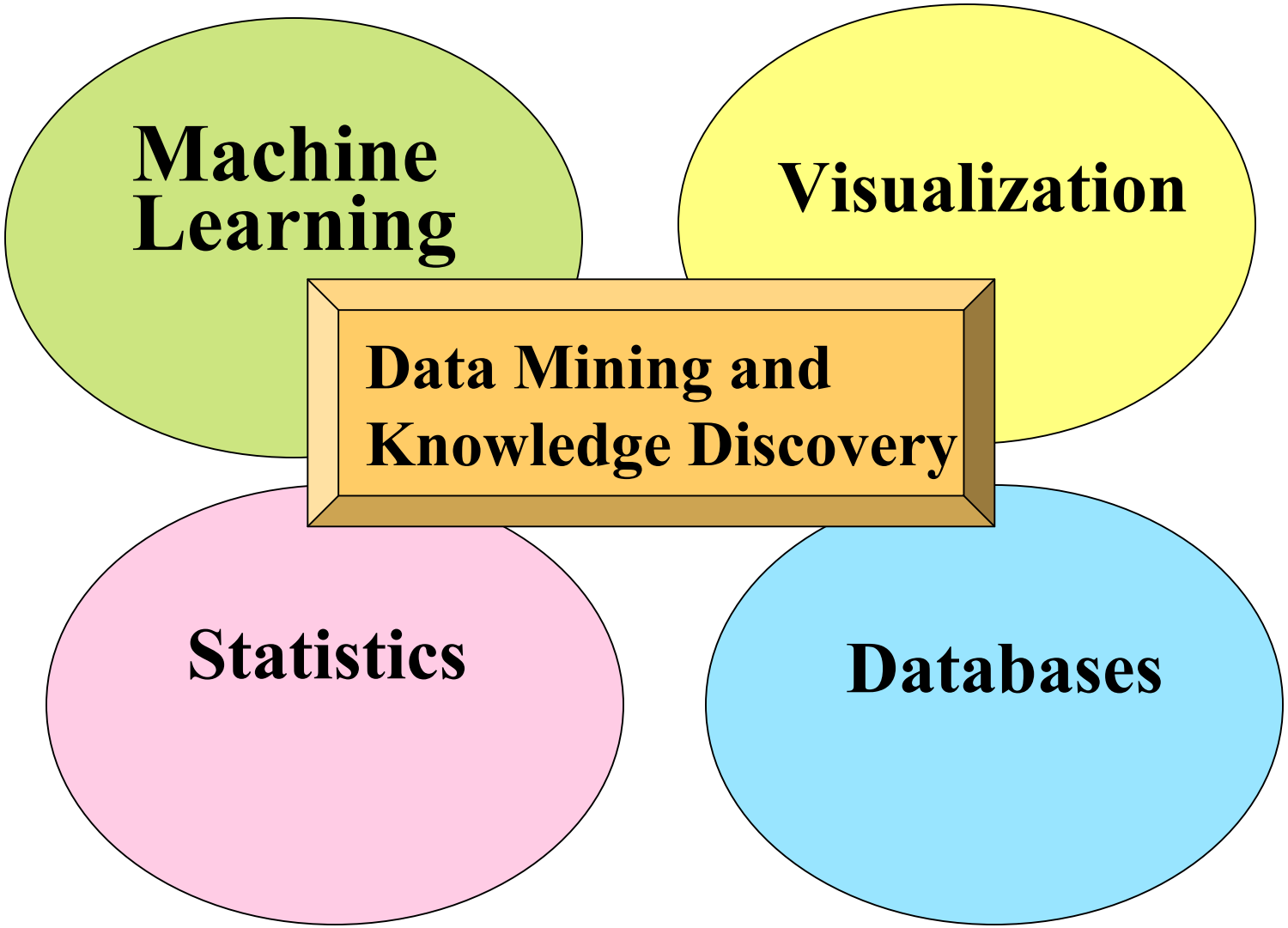
Knowledge Discovery in Data is the

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

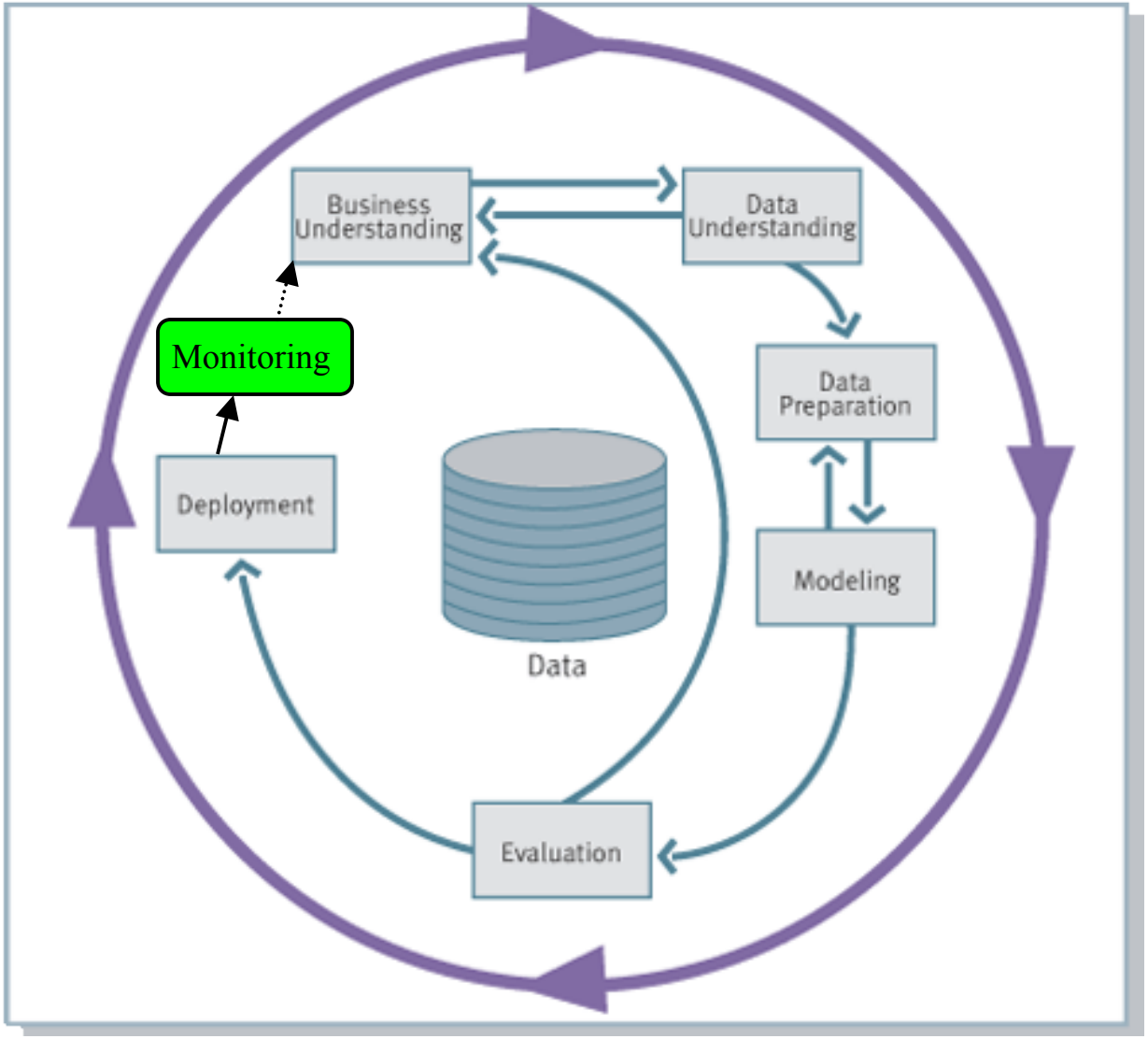
Related Fields



Statistics, Machine Learning and Data Mining

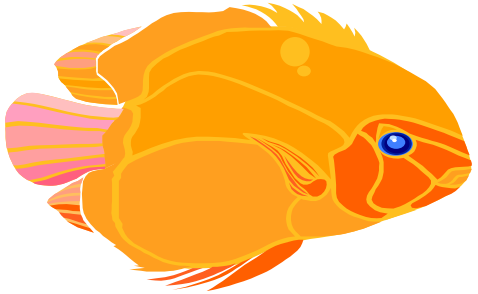
- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates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

Knowledge Discovery Process flow, according to CRISP-DM



see www.crisp-dm.org for more information

Historical Note: Many Names of Data Mining



- Data Fishing, Data Dredging: 1960-
  - used by Statistician (as 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

Lesson Outline

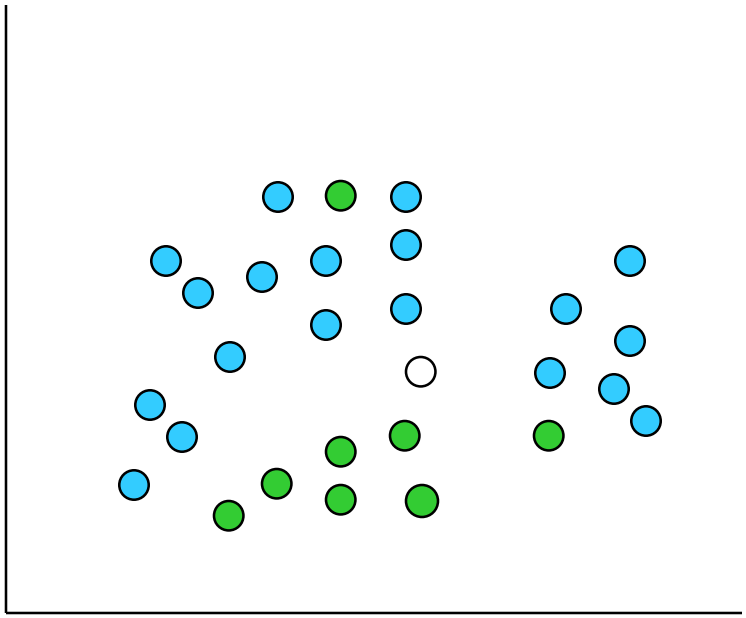
- Introduction: Data Flood
- Data Mining Application Examples
- Data Mining & Knowledge Discovery
- **Data Mining Tasks**

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
- Estimation: predicting a continuous value
- Link Analysis: finding relationships
- …

Data Mining Tasks: Classification

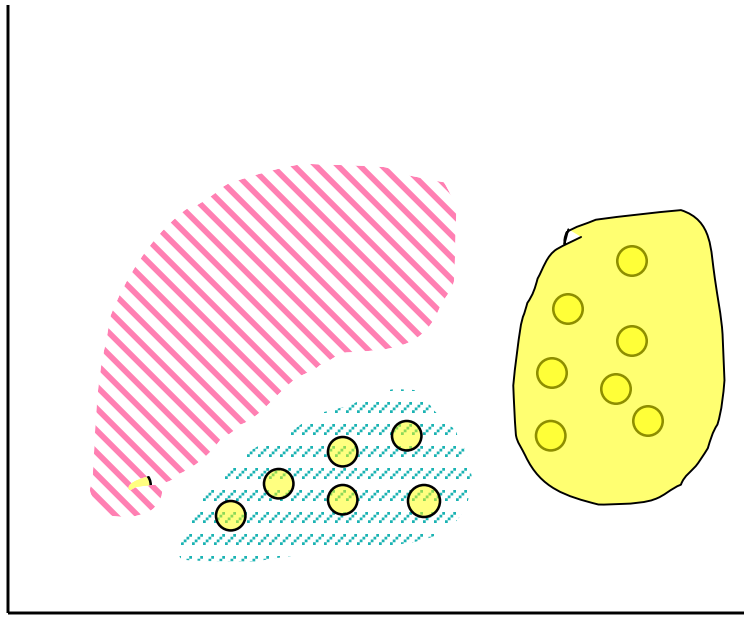
Learn a method for predicting the instance class from pre-labeled (classified) instances



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 Process
- Data Mining Tasks
  - classification, clustering, …

More on Data Mining and Knowledge Discovery

KDnuggets.com

- News, Publications
- Software, Solutions
- Courses, Meetings, Education
- Publications, Websites, Datasets
- Companies, Jobs
- …

Data Mining Jobs in KDnuggets