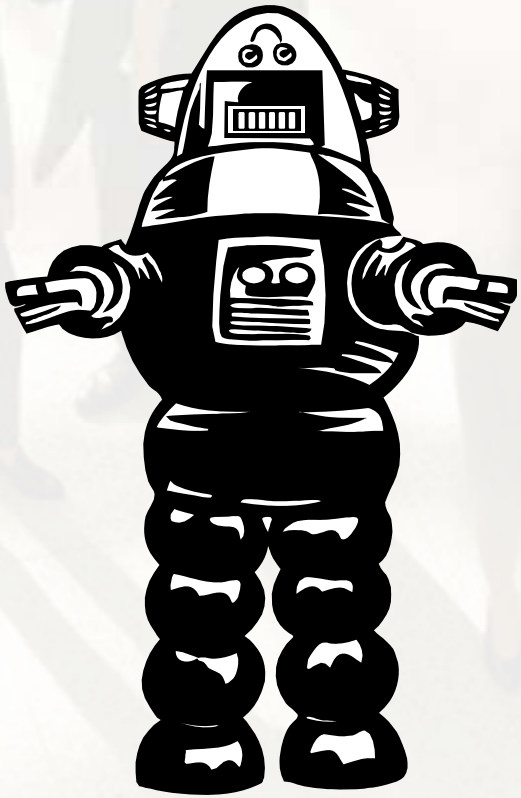
Lection 10

Data Mining and Artificial Intelligence



We are in the 21st century…
So where are the robots?

Data mining is the one really successful application of artificial intelligence technology. It's out there, it's used every day, it makes lots of money and it saves lots of lives.

Definition

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Definition (Cont.)

- Valid: The patterns hold in general.
- Novel: We did not know the pattern beforehand
  - *Married people buy baby-food*
- Useful: We can devise actions from the patterns
  - *More medicine sales after earthquakes*
- Understandable: We can interpret and comprehend the patterns.

- Counter example pattern (Census Bureau Data):
  If (relationship = husband), then (gender = male). 99.6%

Insight and a predictive capability

- Data Mining produces two kinds of results
  - Insight
    - New knowledge, better understanding
  - Predictive capability
    - Spot risks, potentials, anomalies, links

Example: data mining and customer processes

- Insight: Who are my customers and why do they behave the way they do?
- Prediction: Who is a good prospect, for what product, who is at risk,
  what is the next thing to offer?
- Uses: Targeted marketing, mail-shots, call-centres, adaptive web-sites

Example: data mining and fraud detection

- Insight: How can (specific method of) fraud be recognised? What constitute normal, abnormal and suspicious events?
- Prediction: Recognise similarity to previous frauds – how similar?
  Spot abnormal events – how suspicious?
- Used by: Banks, telcos, retail, government…

Why Use Data Mining Today?

"Business Intelligence" as the battlefield

Human analysis skills are inadequate:

- Volume and dimensionality of the data
- High data growth rate
- End-user as the analyst

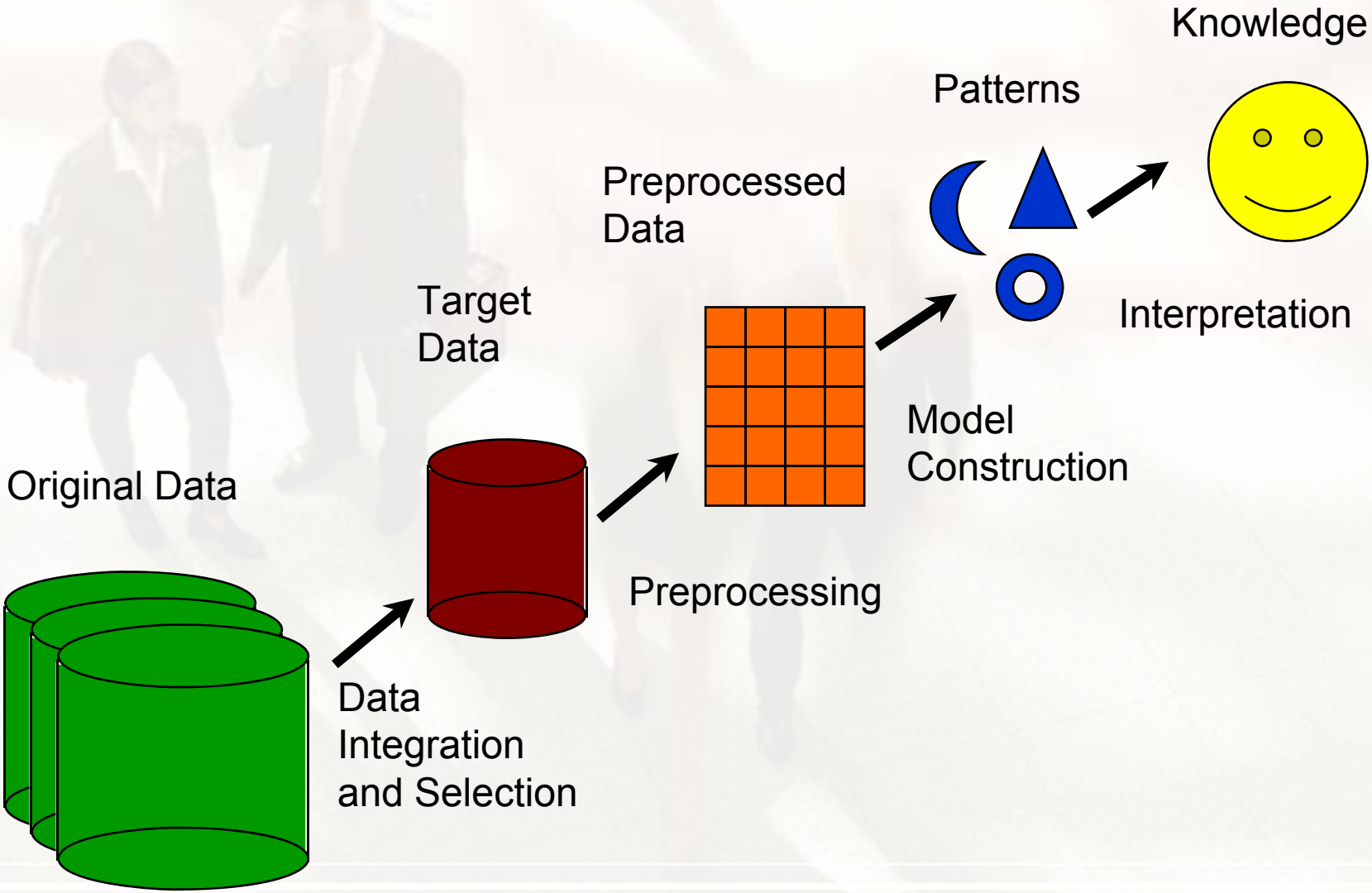
Availability of:

- Data
- Storage
- Computational power
- Off-the-shelf software
- Expertise

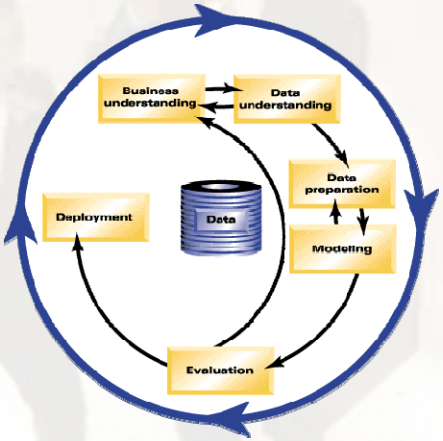
The Evolution of Data Analysis

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	SPSS, Comshare, Arbor, Cognos, Microstrategy,NCR	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	SPSS/Clementine, Lockheed, IBM, SGI, SAS, NCR, Oracle, numerous startups	Prospective, proactive information delivery

Preprocessing and Mining



CRISP-DM Overview



- An industry-standard process model for data mining.
- Not sector-specific
- Non-proprietary

CRISP-DM Phases:

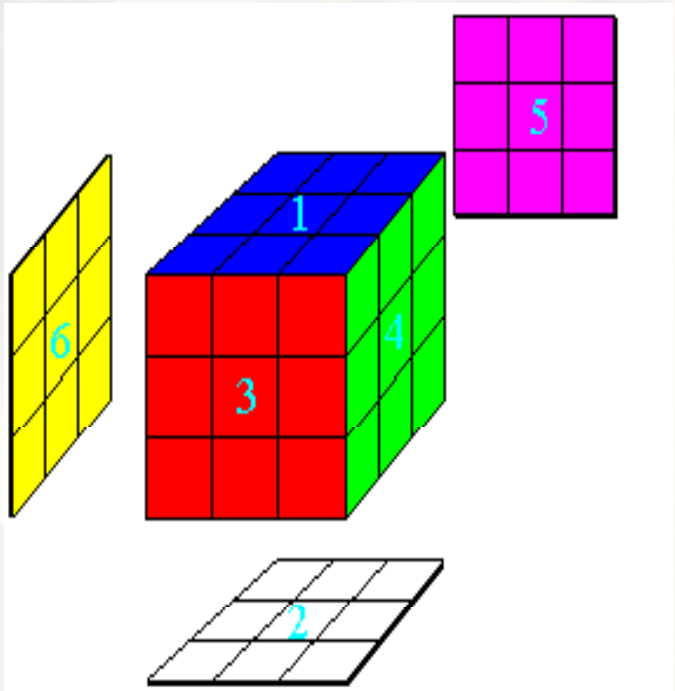
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

➢ Not strictly ordered - respects iterative aspect of data mining

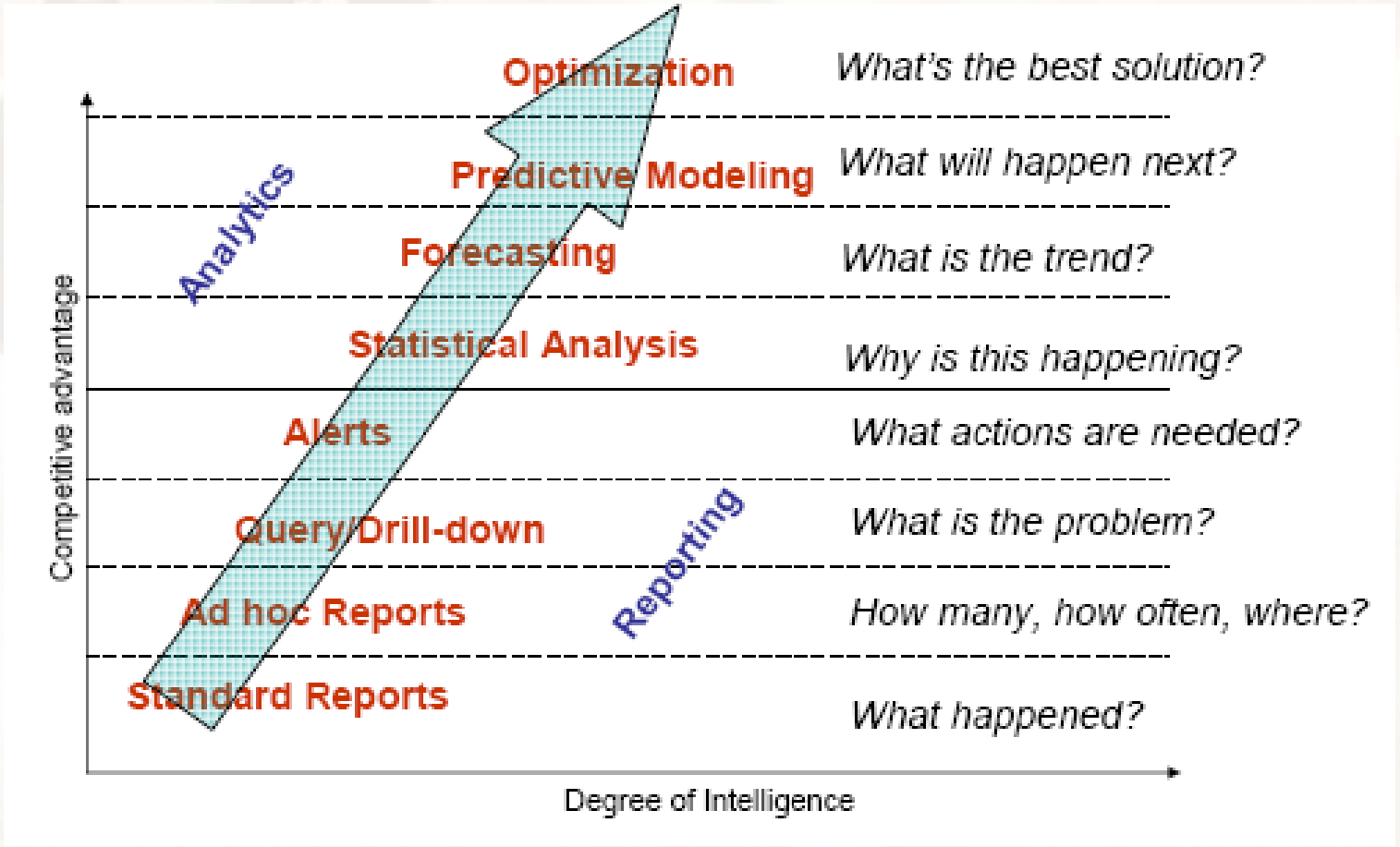
www.crisp-dm.org

Data Mining versus OLAP

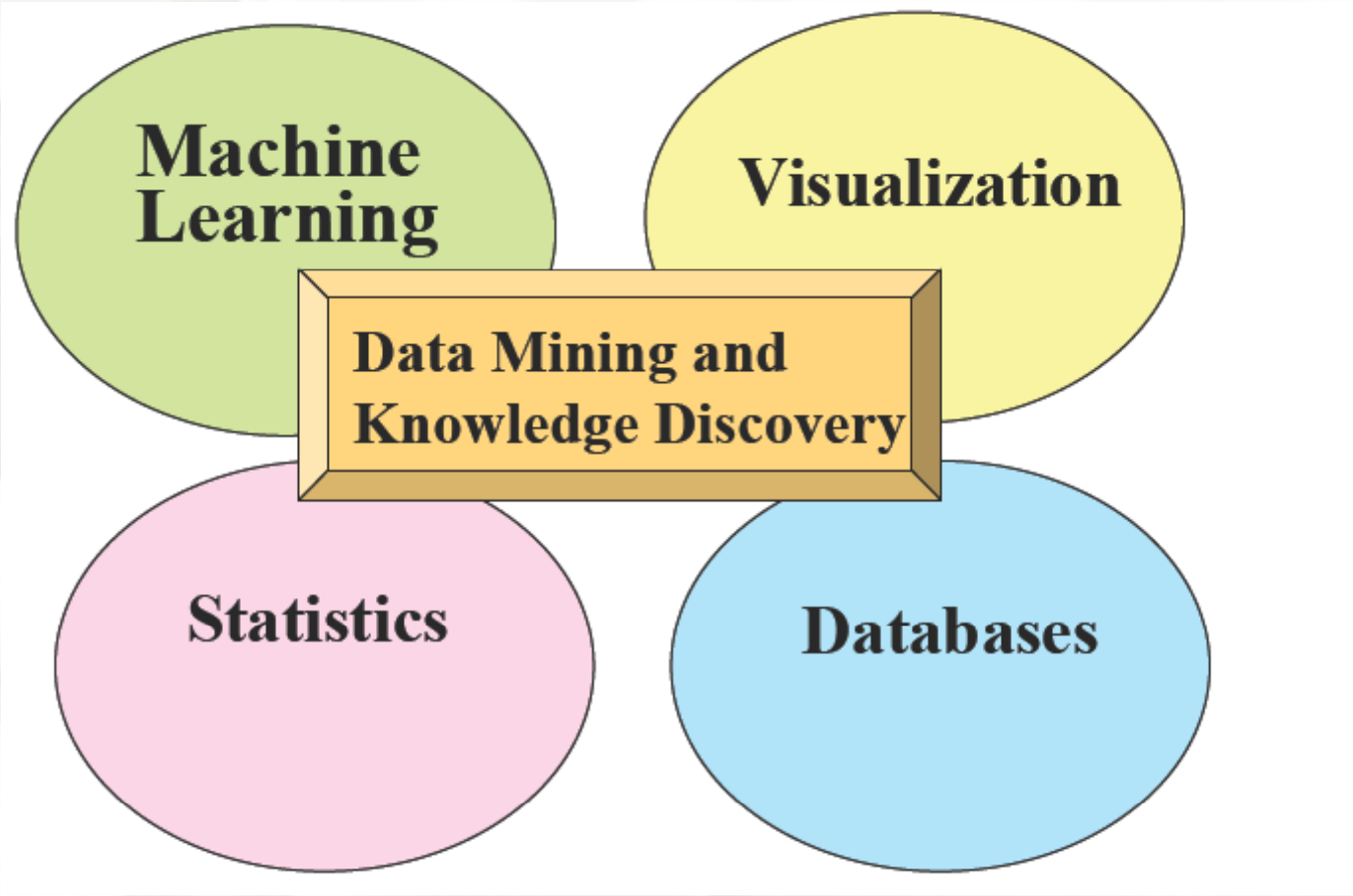
- OLAP - On-line Analytical Processing
  - Provides you with a very good view of what is happening, but can not predict what will happen in the future or why it is happening



Data mining in BI



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

Data Mining Techniques

- Supervised learning
  - Classification and regression
- Unsupervised learning
  - Clustering
- Dependency modeling
- Associations, summarization, causality
- Outlier and deviation detection
- Trend analysis and change detection

Supervised Learning

- F(x): true function (usually not known)
- D: training sample drawn from F(x)

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	0

Supervised Learning

- F(x): true function (usually not known)
- D: training sample (x,F(x))

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	1

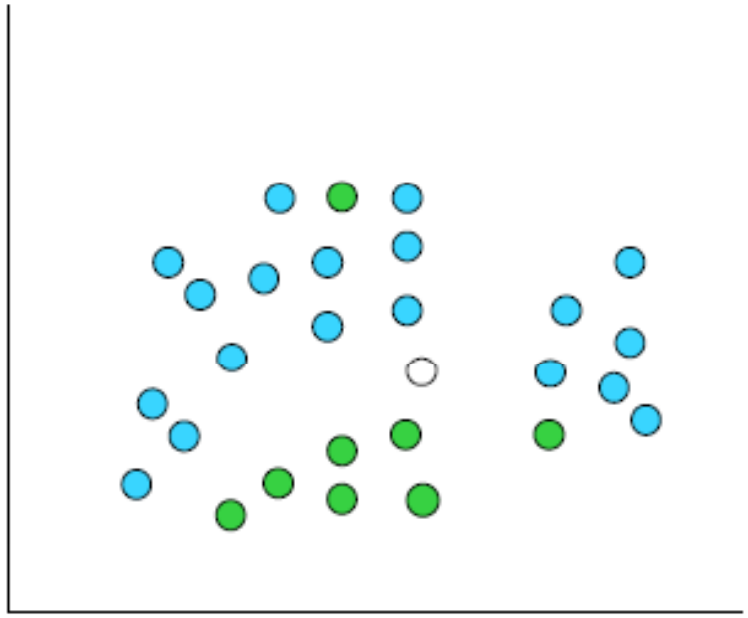
- G(x): model learned from D

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	?

- Goal: $E[(F(x)-G(x))^2]$ is small (near zero) for future samples

Classification

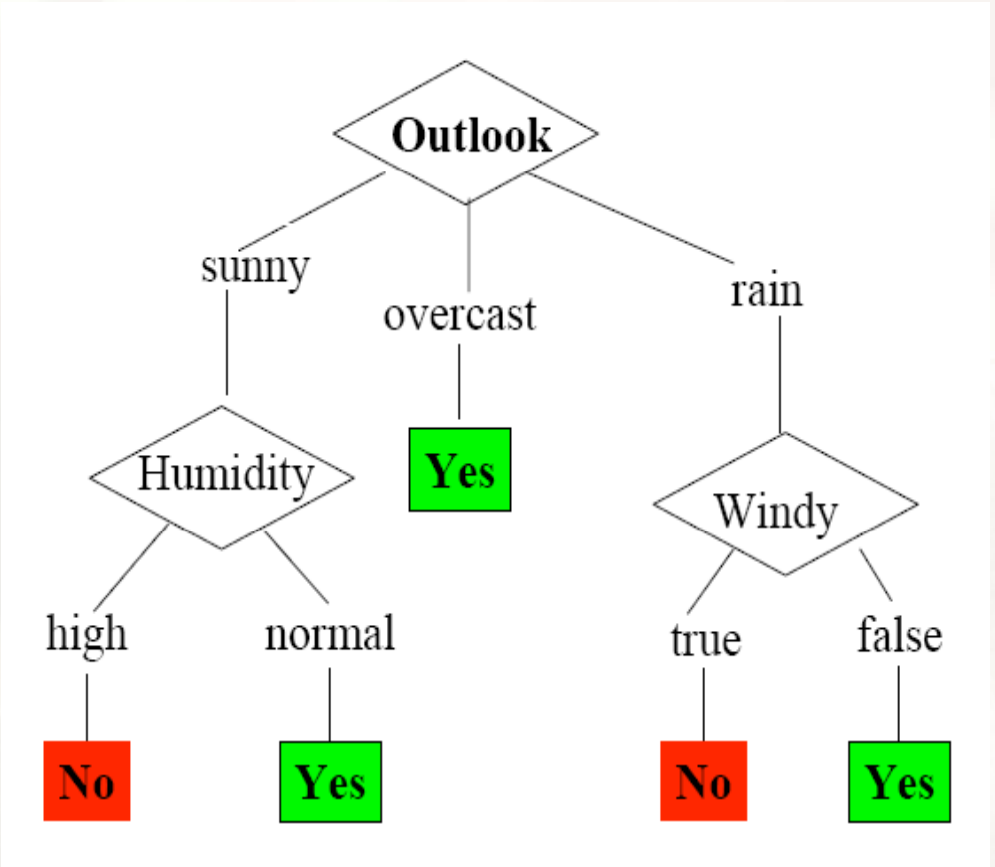
Learn a method for predicting the instance class from pre-labeled (classified) instances



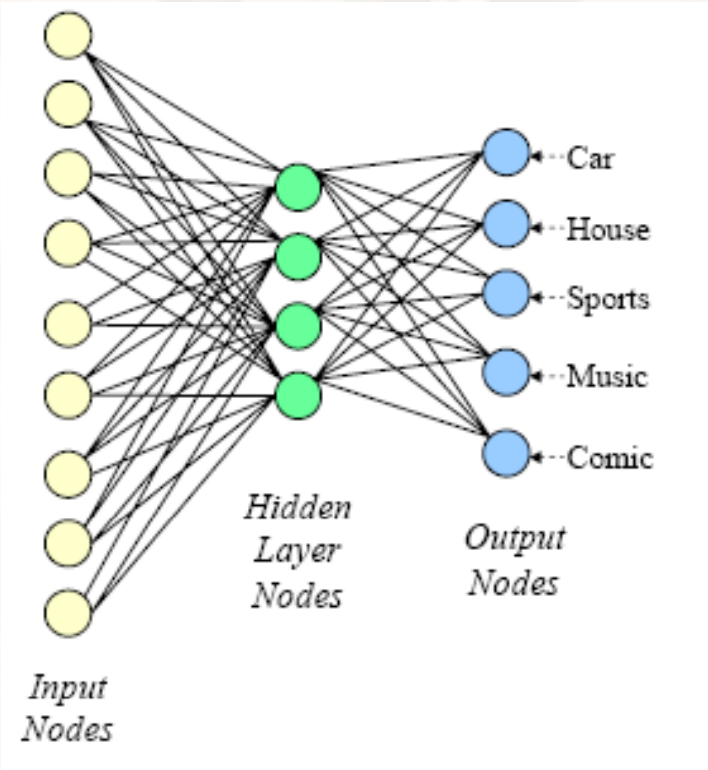
Many approaches:
Statistics,
Decision Trees,
Neural Networks,
...

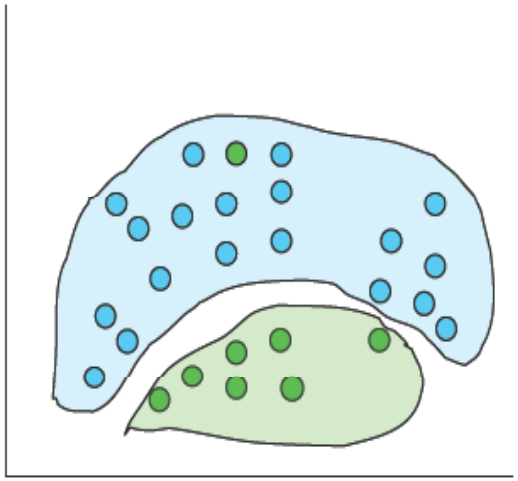
Decision trees

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No



Neural networks





- Can select more complex regions
- Can be more accurate
- Also can overfit the data – find patterns in random noise

Un-Supervised Learning

Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0	
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0

?

Supervised vs. Unsupervised Learning

Supervised

- y=F(x): true function
- D: labeled training set
- D: $\{x_i, F(x_i)\}$
- Learn:
  G(x): model trained to predict labels D
- Goal:
  $E[(F(x)-G(x))^2] \approx 0$
- Well defined criteria:
  Accuracy, RMSE, ...

Unsupervised

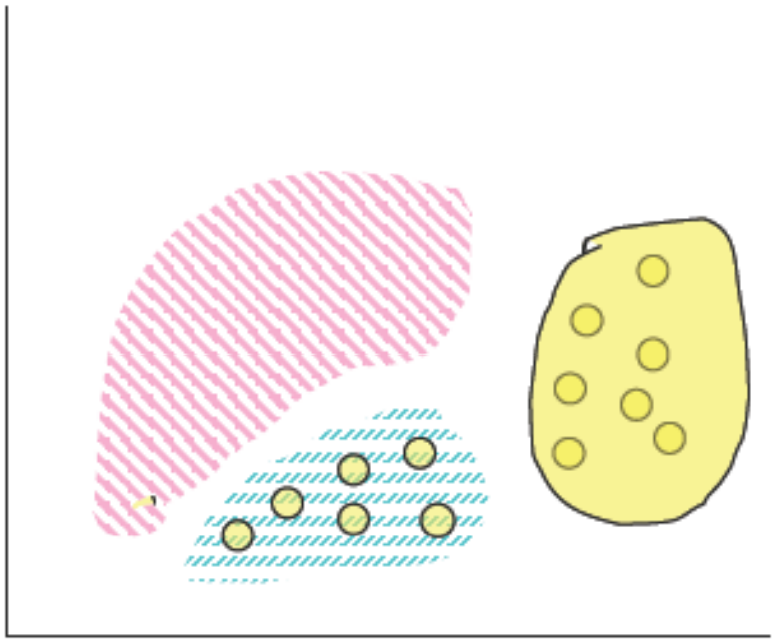
- Generator: true model
- D: unlabeled data sample
- D: $\{x_i\}$
- Learn
  ??????????
- Goal:
  ??????????
- Well defined criteria:
  ??????????

Clustering: Unsupervised Learning

- Given:
  - Data Set D (training set)
  - Similarity/distance metric/information
- Find:
  - Partitioning of data
  - Groups of similar/close items

Clustering

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



Similarity?

- Groups of similar customers
  - Similar demographics
  - Similar buying behavior
  - Similar health
- Similar products
  - Similar cost
  - Similar function
  - Similar store
  - …
- Similarity usually is domain/problem specific

Clustering: Informal Problem Definition

Input:

- A data set of *N* records each given as a *d*-dimensional data feature vector.

Output:

- Determine a natural, useful "partitioning" of the data set into a number of (k) clusters and noise such that we have:
  - High similarity of records within each cluster (intra-cluster similarity)
  - Low similarity of records between clusters (inter-cluster similarity)

Market Basket Analysis

- Consider shopping cart filled with several items
- Market basket analysis tries to answer the following questions:
  - Who makes purchases?
  - What do customers buy together?
  - In what order do customers purchase items?

Market Basket Analysis

Given:

- A database of customer transactions
- Each transaction is a set of items

- Example: Transaction with TID 111 contains items {Pen, Ink, Milk, Juice}

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Market Basket Analysis (Contd.)

- Coocurrences
  - 80% of all customers purchase items X, Y and Z together.
- Association rules
  - 60% of all customers who purchase X and Y also buy Z.
- Sequential patterns
  - 60% of customers who first buy X also purchase Y within three weeks.

Example

Examples:

- {Pen} => {Milk}
  Support: 75%
  Confidence: 75%
- {Ink} => {Pen}
  Support: 100%
  Confidence: 100%

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Market Basket Analysis: Applications

- Sample Applications
  - Direct marketing
  - Fraud detection for medical insurance
  - Floor/shelf planning
  - Web site layout
  - Cross-selling

BI applications: customer profiling

- **Who are most profitable customers and what distinguishes them?**
- **Who are the happiest customers and how are they different?**
- **Which customers are changing categories and what distinguishes them?**
- **How do our customers compare to our competitors' customers?**
- **How do our prospects differ from our customers?**

BI Applications: Direct Marketing and CRM

- Most major direct marketing companies are using modeling and data mining
- Most financial companies are using customer modeling
- Modeling is easier than changing customer behaviour
- Example
  - Verizon Wireless reduced customer attrition rate (churn rate) from 2% to 1.5%, saving many millions of $

BI Applications: e-commerce

- Amazon.com recommendations
  - if you bought (viewed) X, you are likely to buy Y
- Netflix

  - If you liked "Monty Python and the Holy Grail", you get a recommendation for "This is Spinal Tap"

- Comparison shopping
  - Froogle, mySimon, Yahoo Shopping, ...

BI applications: Banking

- Corporate Credit Risk Assessment
  - Should only low risk customers be served?
  - Predicting a risk profile implies predicting default – an outlier
  - Mistakes can be expensive (both + and -)
    - Argentina, B of NE, Kmart, real estate in the 80's
- – Data sources are not always reliable
    - Rating agencies tend to postpone downgrades
    - Low frequency data

BI applications: Banking

- Consumer Lending
  - Traditionally very crude
  - NN models are common today
  - Based on high frequency data
    - Updated with every event
  - Each decision can be used to tune the model
    - Decisions made by managers AND defaults
- Credit Card
  - Very large databases 30*1 MM /month
  - Acquisition, retention, cross-selling, fraud detection, customer service

BI Applications: Security and Fraud Detection

- Credit Card Fraud Detection
  - over 20 Million credit cards protected by Neural networks (Fair, Isaac)
- Securities Fraud Detection
  - NASDAQ KDD system
- Phone fraud detection
  - AT&T, Bell Atlantic, British Telecom/MCI

Data Mining with Privacy

- Exploitation of the customer
  - How is data about me being used?
- Data Mining looks for patterns, not people!
- Technical solutions can limit privacy invasion
  - Replacing sensitive personal data with anon. ID
  - Give randomized outputs
  - Multi-party computation – distributed data

The hype-curve