

## Lecture 1: Data Mining Overview and Process

- What is data mining?
  - Example applications
  - Definitions
  - Multi disciplinary
  - Techniques
  - Major challenges
- The data mining process
- History of data mining
- Data mining resources  
(conferences, journals, Web resources)

## Example application (1) – Telecommunication

- Huge amount of data is collected daily
  - Transactional data (about each phone call)  
(Data on mobile phones, house based phones, Internet, etc.)
  - Other customer data (billing, personal information, etc.)
  - Additional data (network load, faults, etc.)
- Questions
  - Which customer group is highly profitable, which one is not?
  - To which customers should we advertise what kind of special offers?
  - What kind of call rates would increase profit without losing good customers?
  - How do customer profiles change over time?
  - Fraud detection (stolen mobile phones or phone cards)

## Example application (2) – Health

- Different aspects of the health system
  - Personal health records (at GPs, specialists, etc.)
  - Hospital data (e.g. admission data, midwives data, surgery data)
  - Billing information (Medicare, PBS)
- Questions
  - Are doctors following the procedures (e.g. prescription of medication)?
  - Adverse drug reactions (analysis of different data collections to find correlations)
  - Are people committing fraud (e.g. doctor shoppers)
  - Correlations between social and environmental issues and people's health?  
(temporal and spatial analysis of linked data collections)

## Example application (3) – Astronomy

- Terabytes of image and other data from telescopes and satellites  
(large-area sky surveys in optical, infrared, and radio wavelengths)
- Questions
  - Classification of objects (stars, galaxies, pulsars, quasars, etc.)
  - Detect (large scale) structures in the data
  - Find rare, unusual, or even previously unknown types of astronomical objects and phenomena
- MACHO (MAssive Compact Halo Objects) (ANU and US)  
(search for *dark matter*, objects like brown dwarfs or planets in the milky way)

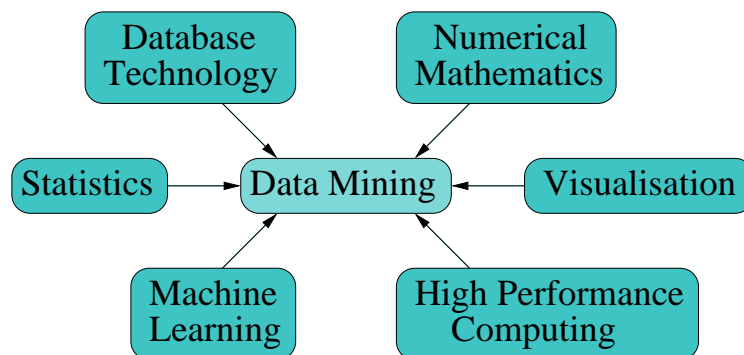
## Definitions of data mining (1)

- *Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.* (Fayyad, Piatetsky-Shapiro and Smyth, 1996)
- *An information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis.* (<http://www.twocrows.com/glossary.htm>)
- Try also: <http://www.google.com>, search term: "define: data mining"

## Definitions of data mining (2)

- Data mining is often also called *Knowledge discovery in databases* (KDD)  
(some say data mining is only one essential step in the KDD process)
- Essential in definitions is:
  - ... non-trivial extraction ...
  - ... previously unknown or novel ...
  - ... potentially useful information ...
  - ... understandable and interesting ...
  - ... large amounts of data ...
  - ... prediction and modelling ...

## Data mining is multi disciplinary



## Data mining methods and techniques (1)

- What they do  
Detect patterns in data: Rules, patterns, classes, associations and functional dependencies, outliers, data distributions
- How they do it  
Search through data and pattern space, non-parametric modelling, filtering, aggregation
- How well they do it  
Errors and biases, over-fitting, confounding effects, speed, scalability

## Data mining methods and techniques (2)

- Cluster analysis (unsupervised learning)  
Group data to form classes, maximise intra-class similarity and minimise similarity between clusters
- Association rules  
Find frequent rules in the data; popular with *market basket analysis*, e.g. "if a customer buys chips he also buys beer (with 50% likelihood)"
- Decision trees  
Build (binary) tree where each node corresponds to a split of attribute values, e.g. "if the weather is sunny play golf else don't play"
- Predictive modelling  
Build mathematical models (functions) of the data in order to predict some unknown or missing values (or future outcomes)

## Data mining methods and techniques (3)

- Neural networks / genetic algorithms
- Outlier detection  
Find unusual, rare events (often regarded as noise, these can be the most interesting objects or events in the data), e.g. fraud detection, network intrusion detection, etc.
- Sequence / time series mining  
Find patterns over time (e.g. episodes, clusters)
- Spatial mining  
Geographical data analysis
- Stream mining  
Where access to the data is limited to once (e.g. network data, telecommunications data, etc.), special algorithms are necessary

## Major challenges in data mining

- Data size
  - Size of data collections grows more than linear, doubling every 18 months (similar to Moore's law of CPU speed)
  - Scalable algorithms are needed
- Data complexity
  - Different types of data (free text, HTML, XML, multimedia)
  - Dimensionality of the data increases (more attributes)
  - The *curse of dimensionality* affects many algorithms (for example find nearest neighbours in high dimensions)
- Privacy and confidentiality

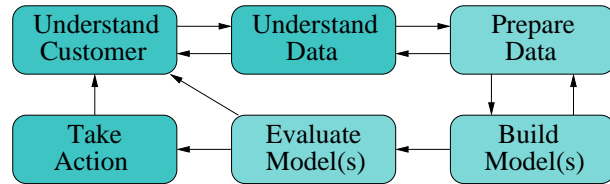
## Ten grand challenges in data mining (U. Fayyad)

- Technical challenges
  1. How does the data grow?
  2. Complexity/understandability trade-off
  3. Interestingness
  4. Scalability
  5. A theory for what we do
- Pragmatic challenges
  6. Where is the data?
  7. Embedding algorithms and solutions within operational systems
  8. Integrating domain knowledge
  9. Managing and maintaining models
  10. Effectiveness measurement

(Source: <http://www.acm.org/sigs/sigkdd/explorations/>, Editorial, volume 5, issue 2, Dec. 2003)

## The data mining / KDD process (1)

- Data mining is an interactive process



- Data mining = "Build Model(s)"
- Typically 90% of time and efforts are spent in the first 3 steps

(Follows: *CRoss Industry Standard Process for Data Mining*, <http://www.crisp-dm.org/>)

## The data mining / KDD process (2)

- An iterative sequence of the following steps
  1. Data cleaning
  2. Data integration
  3. Data selection
  4. Data transformation
  5. Data mining
  6. Pattern evaluation
  7. Knowledge presentation

(Follows: *Data Mining: Concepts and Techniques*, Han/Kamber)

## History of data mining

- The term *data mining* was first mentioned by statisticians several decades ago, but with a different meaning compared to today: *data dredging* (inappropriate (sometimes deliberately so) search for *statistically significant* relationships in large quantities of data; from Wikipedia)
- First workshops on knowledge discovery in databases in early 1990 (part of ACM SIGMOD (management of data) conferences)
- First data mining conferences in mid 1990
- Many more conferences since early 2000
- Data mining is around 15 years old

## Data mining resources (1)

- Conferences
  - ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (since 1995)
  - European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) (since 1997)
  - Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) (since 1997)
  - SIAM (Society for Industrial and Applied Mathematics) International Conference on Data Mining (since 2001)
  - IEEE (International Institute of Electrical Engineers) International Conference on Data Mining (since 2001)
  - AusDM (workshop since 2002, conference since 2004)

## Data mining resources (2)

- Journals
  - <http://www.kluweronline.com/issn/1384-5810/>  
(Kluwer Data Mining and Knowledge Discovery)
  - <http://www.computer.org/tkde/>  
(IEEE Transactions on Knowledge and Data Engineering)
  - <http://www.acm.org/sigs/sigkdd/explorations/issue.php?issue=current>  
(ACM SIGKDD Explorations)
  - <http://www.cs.uvm.edu/~kais/>  
(Springer Knowledge and Information Systems)

## Data mining resources (3)

- (Some) Web resources
  - <http://www.kdnuggets.com/>
  - <http://www.dmg.org/> (Data mining group, PMML)
  - <http://www.acm.org/sigs/sigkdd/>
  - <http://datamining.anu.edu.au/>
  - <http://www.togaware.com/datamining/catalogue.html>
  - <http://www.togaware.com/analytics/>  
Canberra Analytics Group
  - <http://kdd.ics.uci.edu/>  
(UCI Knowledge Discovery in Databases Archive)