

COMP8400: Algorithms and Techniques for Data Mining

The data mining process

Lecture outline

- Three example data mining applications
- The data mining / KDD process
- Data mining and business intelligence
- Definitions of data mining
- Major challenges in data mining
- Short history of data mining
- Data mining resources
- What to do next

Example application 1: Telecommunication

- Huge amounts of data are collected on a daily basis
 - Transactional data (about each phone call)
(data on mobile phones, house based phones, Internet, etc.)
 - Customer data (billing, personal information, etc.)
 - Additional data (network load, faults, etc.)
- Possible 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)
 - Network load predictions

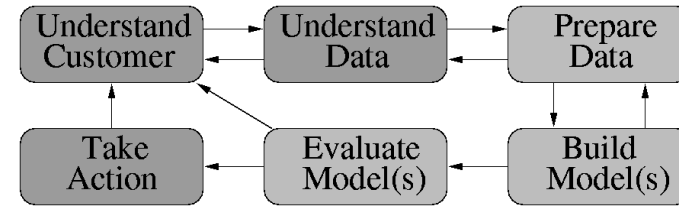
Example application 2: Health

- Different aspects of the health system
 - Personal health records (at general practitioners and specialists)
 - Hospital data (e.g. admission data, midwives data, surgery data, etc.)
 - Nursing homes and death data (admissions, causes, medications, etc.)
 - Billing information (Medicare, Pharmaceutical Benefit Scheme)
 - Private health insurance and ambulance/emergency data
- Possible 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)?
 - Are there correlations between social and environmental issues and people's health? (temporal and spatial analysis of linked data collections)

Example application 3: Astronomy

- Terabytes of images and other data from telescopes and satellites
 - Large-area sky surveys in optical, infrared, and radio wavelengths
 - Time-series data
- Possible 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

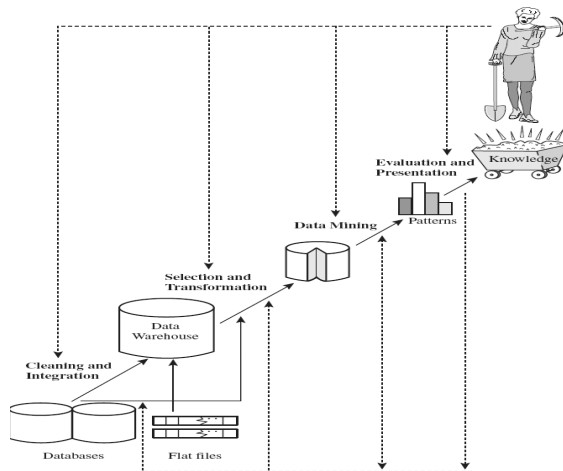
The data mining / KDD process



- Data mining is an interactive process
- Data mining = *Build Model(s)*
- Typically up to 90% of time and effort are spent in the first three steps!

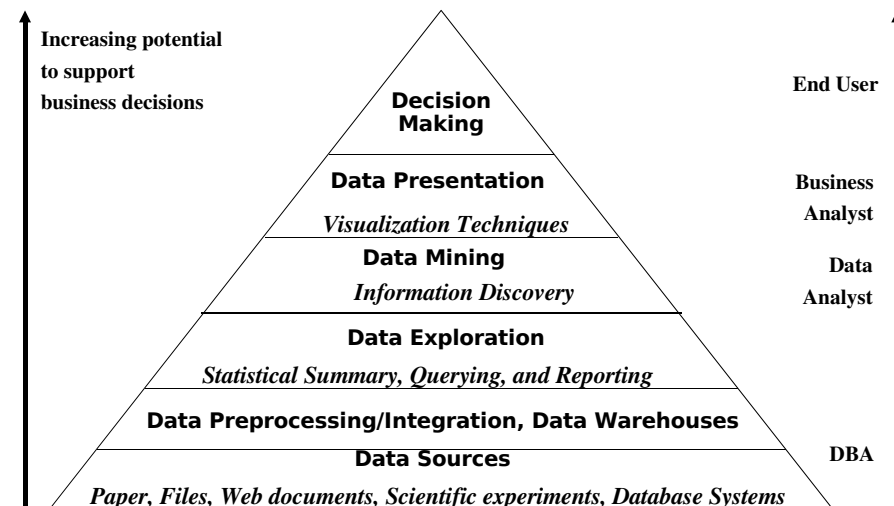
(Follows: *CRoss Industry Standard Process for Data Mining*, <http://www.crisp-dm.org/>)

The data mining / KDD process (2)



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Data mining and business intelligence



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

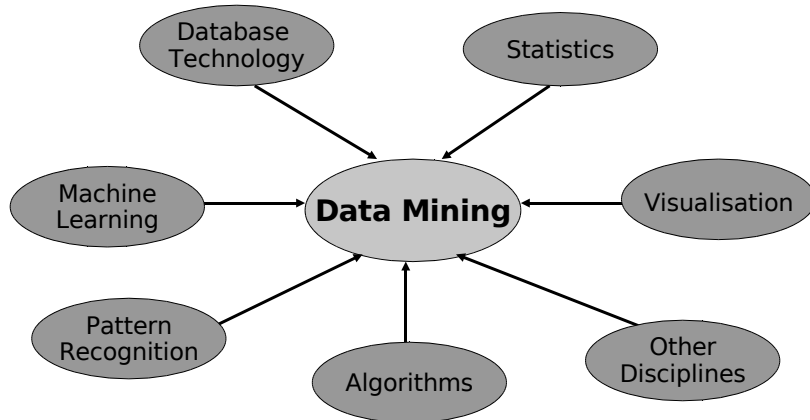
Definitions of data mining

- *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)

- 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 often also called Knowledge Discovery in Databases (KDD)
 - Some say data mining is only one essential step in the KDD process

Data mining is multi-disciplinary



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

Major challenges in data mining

- **Data size**
 - Size of data collections grows more than linear, doubling around every 18 months (similar to Moore's law of CPU speed)
 - Scalable algorithms are needed
- **Data complexity**
 - Different types of data (database tables, 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**
 - Data mining can reveal details about people which is not available otherwise
 - Linking and matching data is especially critical / controversial

Ten grand challenges in data mining (U. Fayyad)

- Technical challenges

- How does the data grow?
- Scalability (of algorithms)
- Complexity/understandability trade-off
- Interestingness
- A theory for what we do

- Pragmatic challenges

- Where is the data?
- Embedding algorithms and solutions within operational systems
- Integrating domain knowledge
- Managing and maintaining models
- Effectiveness measurement

(Source: <http://www.acm.org/sigs/sigkdd/explorations/>, Editorial, vol 5, no 2, Dec. 2003)

- Another, more recent paper is on COMP8400 Web site

Short 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 late 1980s and early 1990s (part of IJCAI (Artificial Intelligence) and ACM SIGMOD (Management of Data) conferences)
- First data mining conferences in mid 1990
- Many more conferences since early 2000
- So data mining is now in it's teen years (around 19 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 (Institute of Electrical and Electronics Engineers) International Conference on Data Mining (ICDM) (since 2001)
- Australasian Data Mining Conference (AusDM) (workshop since 2002, conference since 2004)

Data mining resources (2)

- Journals

- Springer Data Mining and Knowledge Discovery
<http://www.springerlink.com/content/1573-756X>
- Springer Knowledge and Information Systems
<http://www.springerlink.com/content/0219-3116>
- IEEE Transactions on Knowledge and Data Engineering
<http://www.computer.org/tkde/>
- ACM SIGKDD Explorations
<http://www.acm.org/sigs/sigkdd/explorations>
- ACM Transactions on Knowledge Discovery from Data
<http://tkdd.cs.uiuc.edu/>

Data mining resources (3)

- Web resources

- <http://www.kdnuggets.com/> (News, software, jobs, courses, conferences, data repositories, polls, and more)
- <http://www.dmg.org> (Data mining group, PMML)
- <http://www.acm.org/sigs/sigkdd/> (ACM Special Interest group on KDD)
- <http://datamining.anu.edu.au/>
- <http://www.togaware.com/> (Graham Williams, ATO)
- <http://www.iapa.org.au> (Institute of Analytics Professionals of Australia)
- <http://www.togaware.com/analytics/> (Canberra Analytics Group)
- <http://kdd.ics.uci.edu/> (UCI Knowledge Discovery in Databases Archive)

What now... things to do

- Make sure you are enrolled properly!
- Fill in tutorial/lab list
- Check COMP8400 Web site, especially course *Schedule*, *Assessment*, discussion *Forum* and *Links*
- If you plan to use your laptop for labs, download and install *Rattle* and possibly *Weka* / *KNIME*
- Get and read paper *Competing on Analytics* from COMP8400 lectures Web site Web site
- Read chapter 1 in text book