HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

582364 Data mining, 4 cu
Lecture 1: Introduction

Spring 2010

Lecturer: Juho Rousu

Teaching assistant: Taru Itäpelto

Course topics

- Discovery of frequent patterns from data
  - Association rules
  - Sequential patterns
- Data mining process
  - Data preprocessing
    - Discretization, missing value treatment
  - Validation of data mining results
    - Statistical testing, randomization
- Special topics (to be fixed)
  - Graph mining, text mining, web mining

Course organization

- Lectures:
  - Mon 12-14, Tuesday 10-12 (alternating with group work): 16.3., 30.3, 20.4.
  - Juho Rousu (A239B, juho.rousu (ät) cs.helsinki.fi)
- Group work sessions:
  - Tuesdays 10-12 (alternating with lectures): 23.3, 13.4, 27.4
- Exercises & group work debrief:
  - Tuesdays 12-14
  - Taru Itäpelto-Hu (B333, itapelto (ät) cs.helsinki.fi)
- Paper summary deadlines: 22.3, 12.4, 26.4
- Course exam: 4.5. at 9-12, lecture hall B123
- Easter break 1-7.4.: no sessions on 5-6.4

Completing the course & grading

- Exercises, completed as homework, reviewed in the exercise session: 15% of the grade
- Group work, completed during group work session, presented in debrief session: 15% of the grade
- Paper summarizing, completed as homework: 15% of the grade
- Course exam: 55% of the grade

- Grading: 50% of total gives 1/5, 80% of total gives 5/5

- Alternatively: next separate exam 4.6 at 16-20 A111

Group work & exercises

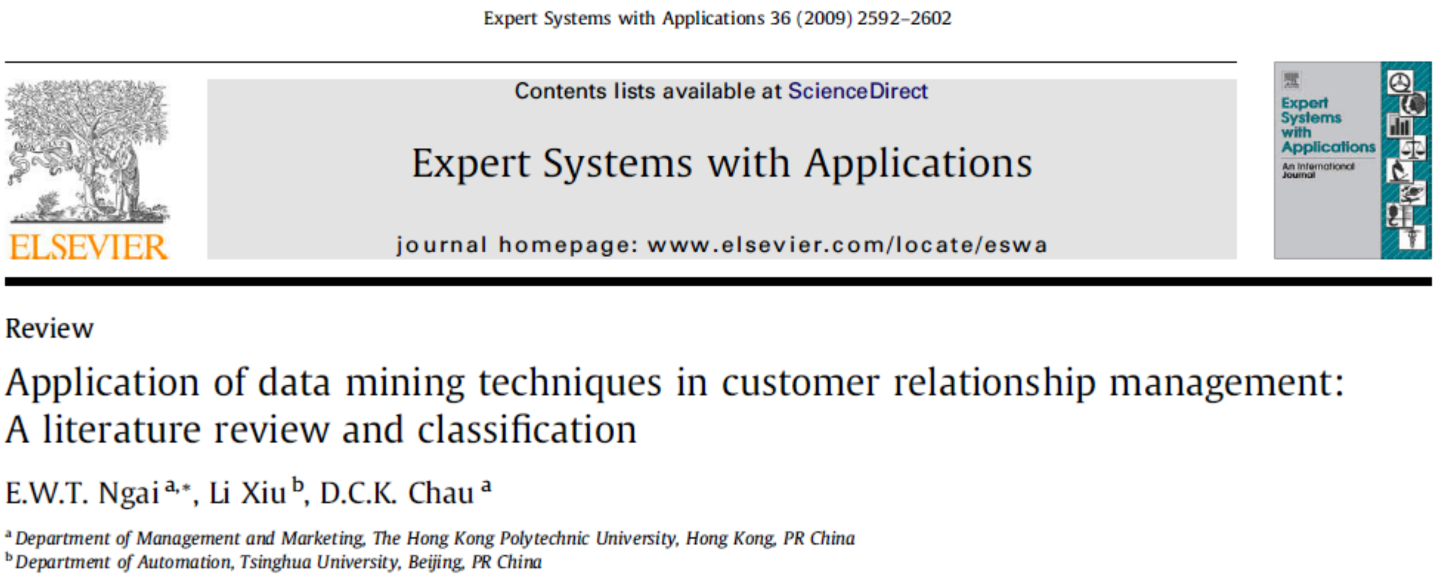
- Three group work sessions: 23.3., 13.4, 27.4 in B222
  - Group work session 10-12
  - Group work debrief 12-14
- Organization into groups and assignments for the groups at the start of each session
- Groups' work presented in the debrief session immediately after the group session
- Type of assignments may vary (Technical questions, brainstorming, figuring out an algorithm, …)
- Two 'normal' exercise sessions: 29.3 and 20.4 in B222
- Exercises given out the previous week, completed at home, reviewed in the exercise session

Writing summaries of scientific papers

- Three scientific papers will be given to read and summarize during the course
- Deadlines: #1: 22.3, #2: 13.4, #3: 26.4
- Summary will looks as the following:
  - Length 2-4 pages
  - Gathers the main contents of a scientific article and rephrases it in your own words
  - Format of a scientific paper with title, author information (you), an abstract, section titles and references.
  - Summary is to be returned as a PDF file, via email to Taru (itapelto (ät) cs.helsinki.fi) by the given deadline.
  - Each paper will be graded on a scale of 1-5. Late submissions will be automatically graded down.

Paper #1

- Available from the course web page
- Access restricted to cs.helsinki.fi and hiit.fi domains, so download it when at the university
- Deadline: Monday 22.3 by 23:59
- Return as pdf file to Taru (itapelto (ät) cs.helsinki.fi)



Expert Systems with Applications 36 (2009) 2592–2602

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Application of data mining techniques in customer relationship management: A literature review and classification

E.W.T. Ngai[a,*], Li Xiu[b], D.C.K. Chau[a]

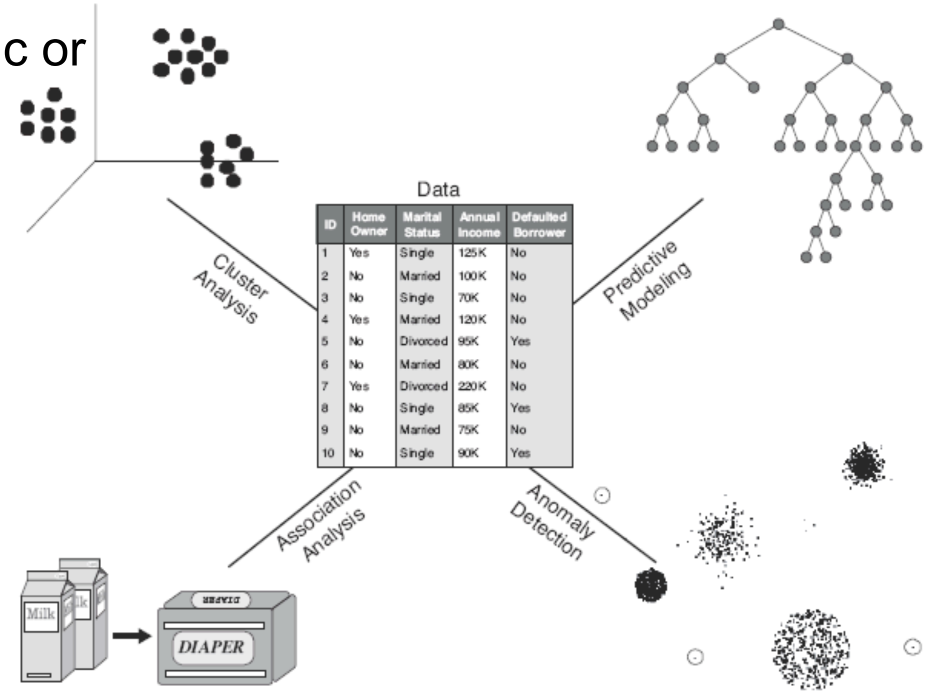
[a] Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong, PR China
[b] Department of Automation, Tsinghua University, Beijing, PR China

582635 Data mining project, 2 cr

- Separate course immediately after this course
- 10.5.-21.5
- Data mining techniques are applied in practice. Students can complete the course in two ways:
  - Either by implementing a data mining algorithm given in the assignment and by analyzing a given data with it,
  - or,by mining given data with a (wider) selection of methods, e.g. using ready-made software.
- In both cases, a research report is written describing the work and a seminar presentation is given.

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



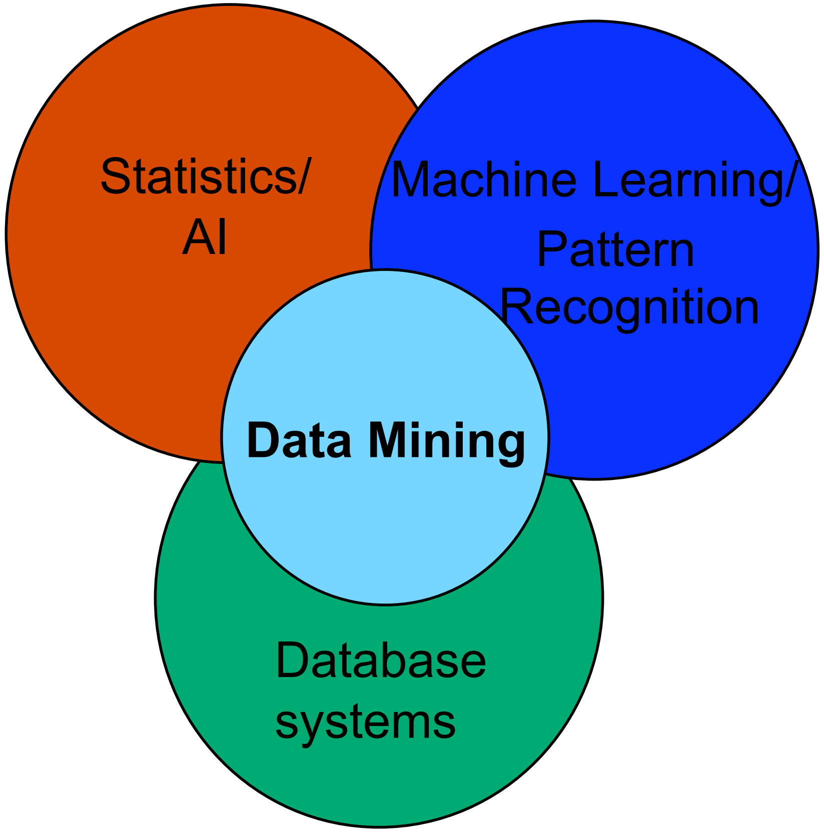
Data

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 1.3. Four of the core data mining tasks.

Neighbor disciplines of Data Mining

- Draws ideas and methods from
  - machine learning/AI,
  - pattern recognition,
  - statistics
  - database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data
- “Algorithms meet statistics”



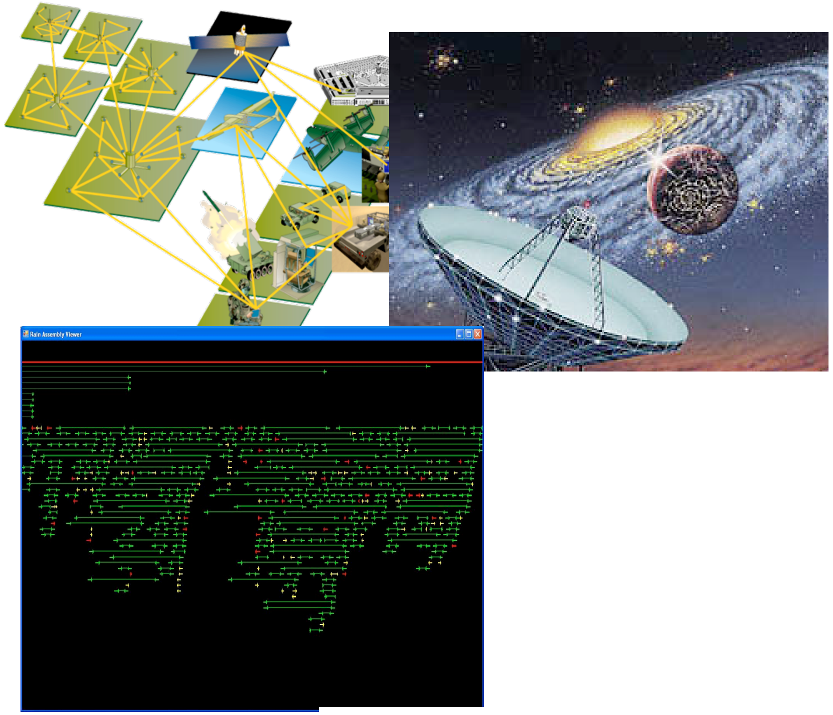
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Loyalty cards
    - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - 3G mobile phones with GPS

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - DNA sequencing robots churning out new genomes
  - Web, social media
- Traditional techniques infeasible for raw data
  - Data too large/heterogeneuos
  - Patterns too complex/numerous
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



Data Mining: Predictive tasks

- Goal: Use some variables to predict *unknown* or *future* values of pre-defined target variables.
- Major types of predictive tasks
  - Classification: predict the value discrete target variable with typically few values, often binary variable
  - Regression: predict the value of a continuous target variable
  - Anomaly/novelty detection: predict a deviation from normal/current state of affairs
- Related concept: *supervised* machine learning
- More information: courses *Introduction to machine learning, Probabilistic models*

Classification application: credit card fraud detection

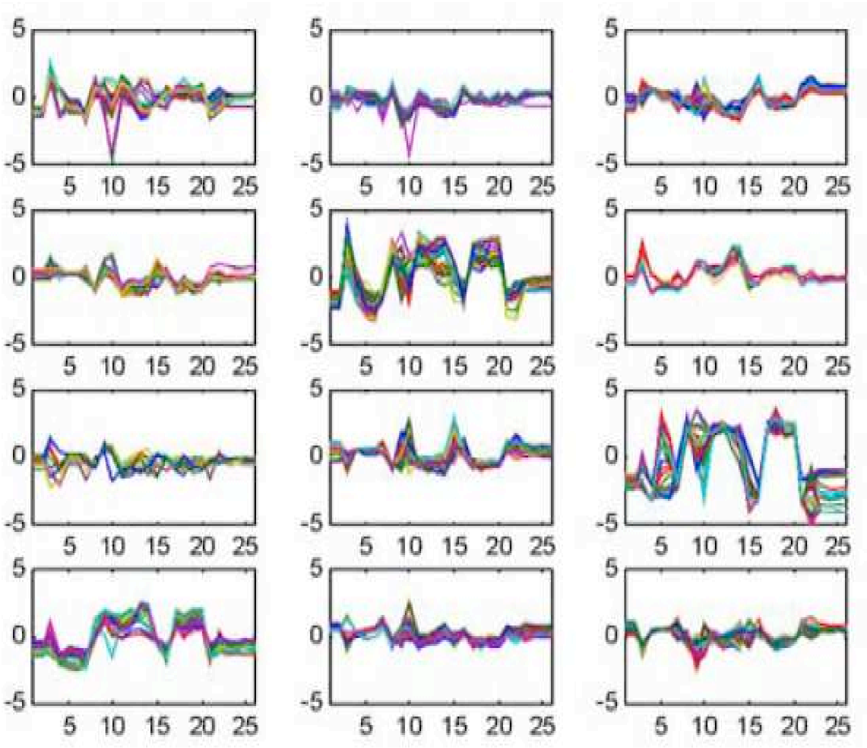
- Goal: detect fraudulent use of credit card (card and/or card details stolen and misused)
- Approach:
  - Database of credit card purchase transactions (date, place, store details, card holder details, price of purchased item, transaction history of the credit card, …)
  - Transactions labeled as “normal” vs. “fraudulent”
- Learn a classifier that predicts the from the transaction data the correct label
  - Many methods: Decision trees, Support vector machine, Nearest neighbor classifier, Naïve Bayes

Regression application: house pricing

- Goal: real estate agent wants to predict the selling price of a house in order to set an appropriate asking price
- Approach:
  - Database of transactions of previously sold property (location, type of house/apartment, details of the property, asking price, time on market,…)
  - Target variable: sale price of the property
- Learn a regression model that predicts the from the transaction data the sale price
- Many methods: Linear regression, Support vector regression, Nearest neighbor regression, Regression trees

Anomaly detection application: Industrial Process Monitoring

- Goal: a tool that checks if the process is running within normal specifications
- Data from problem situations not available
  - e.g. nuclear reactor meltdown
- Collect measurement data from the process and devise prototype profile(s) of normal operation
- Large deviation from the prototype causes an alarm

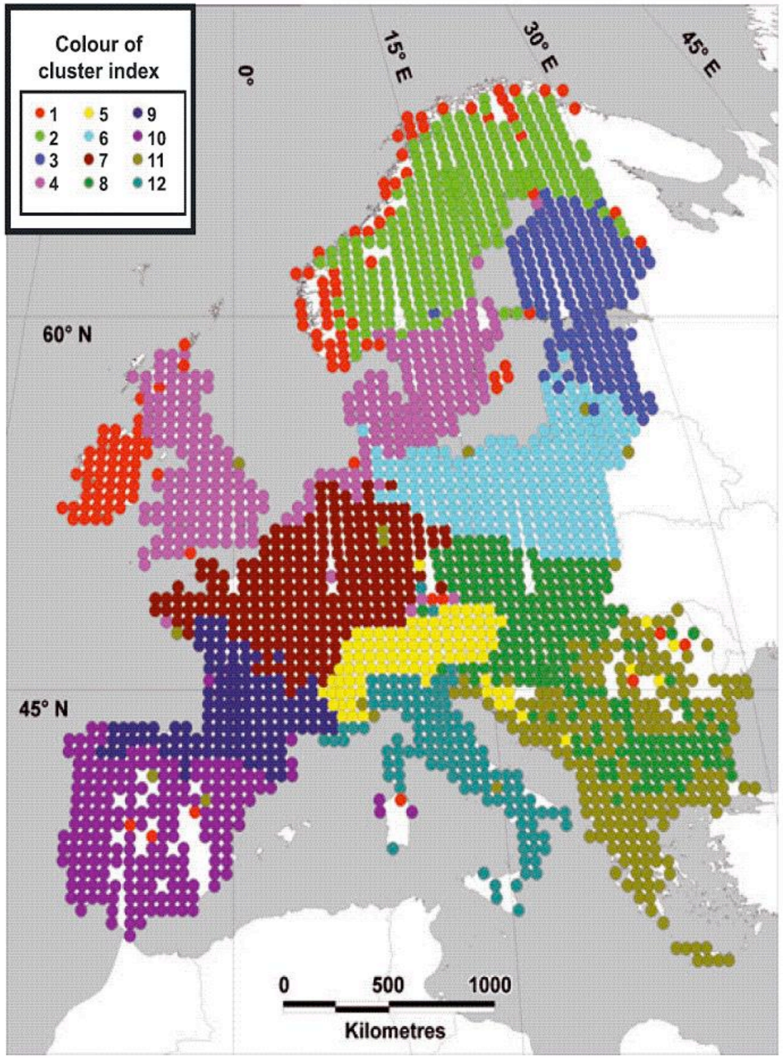


Data Mining: Descriptive tasks

- Goal: Find human-interpretable patterns that describe the data.
- “What has the data have to tell?”
  - Typically no specific target variable
- Major types of descriptive tasks
  - Clustering: divide the data into internally coherent groups
  - Frequent pattern discovery: find combinations of variables’ values that occur more frequently than expected by chance
- Related concepts: unsupervised machine learning, explorative data analysis
- *Note: the division in to descriptive and predictive tasks is not sharp; human experts often want to understand why certain prediction is arrived at – hence we end up the task of describing the predictive model.*

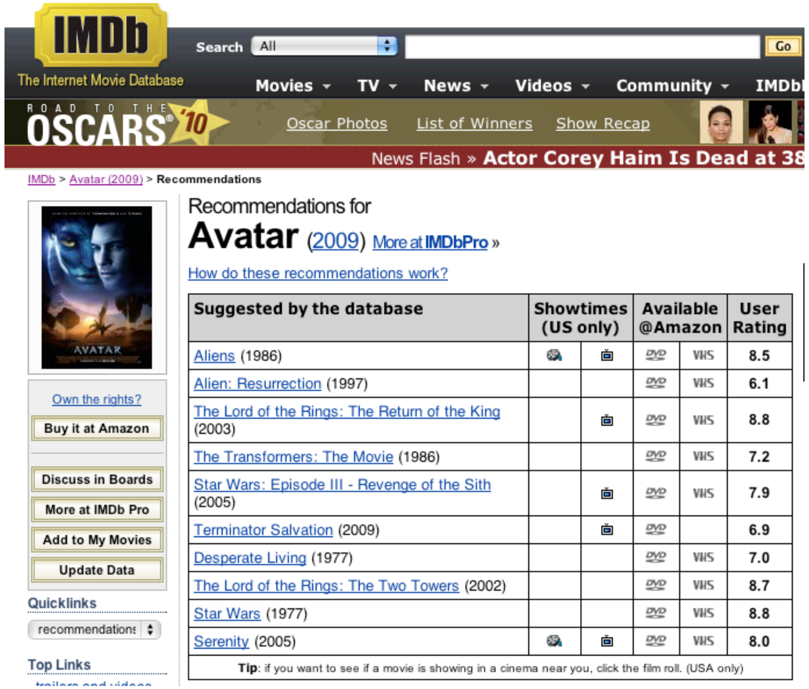
Clustering Application: Ecological data analysis

- Clustering can reveal new groupings in data
- In the picture, clustering of the European map is based on occurrence of species
- Each cell is a 50x50 km area where the occurrence of 124 species has been recorded
- Clusters = similar occurrence profiles
- Many methods available: k-means clustering, hierarchical clustering, …



Frequent Pattern Discovery Application: Recommendation systems

- Many recommendation web sites are based on collecting data on frequently co-occurring items
- If you liked the film “Aliens” you probably like “Avatar”
- People that buy book X, frequently buy book Y



News Flash » Actor Corey Haim Is Dead at 38

IMDb > Avatar (2009) > Recommendations

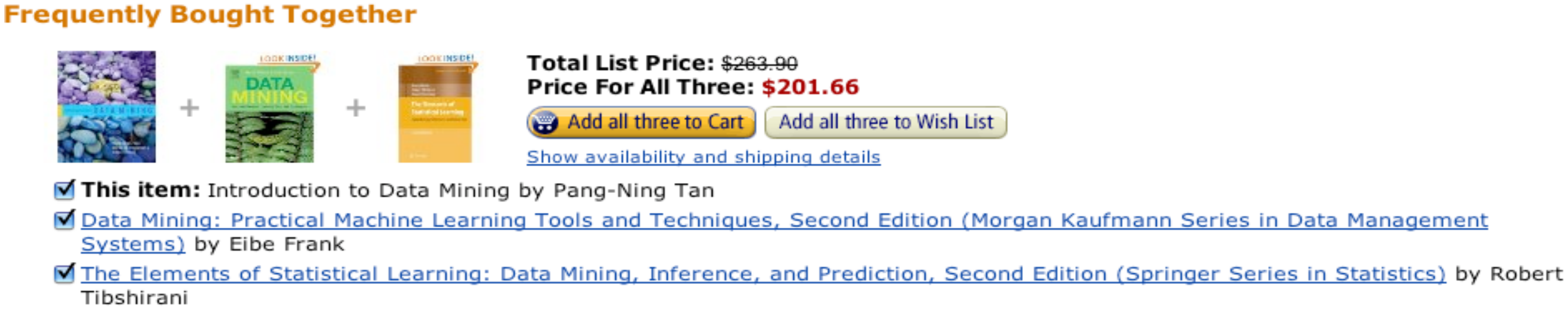
Recommendations for

Avatar (2009) More at IMDbPro »

How do these recommendations work?

Suggested by the database	Showtimes (US only)		Available @Amazon		User Rating
Aliens (1986)	🎞	📺	DVD	VHS	8.5
Alien: Resurrection (1997)			DVD	VHS	6.1
The Lord of the Rings: The Return of the King (2003)		📺	DVD	VHS	8.8
The Transformers: The Movie (1986)			DVD	VHS	7.2
Star Wars: Episode III - Revenge of the Sith (2005)		📺	DVD	VHS	7.9
Terminator Salvation (2009)		📺	DVD		6.9
Desperate Living (1977)			DVD	VHS	7.0
The Lord of the Rings: The Two Towers (2002)			DVD	VHS	8.7
Star Wars (1977)			DVD	VHS	8.8
Serenity (2005)	🎞	📺	DVD	VHS	8.0

Tip: if you want to see if a movie is showing in a cinema near you, click the film roll. (USA only)



Frequently Bought Together

Total List Price: ~~$263.90~~
Price For All Three: $201.66

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

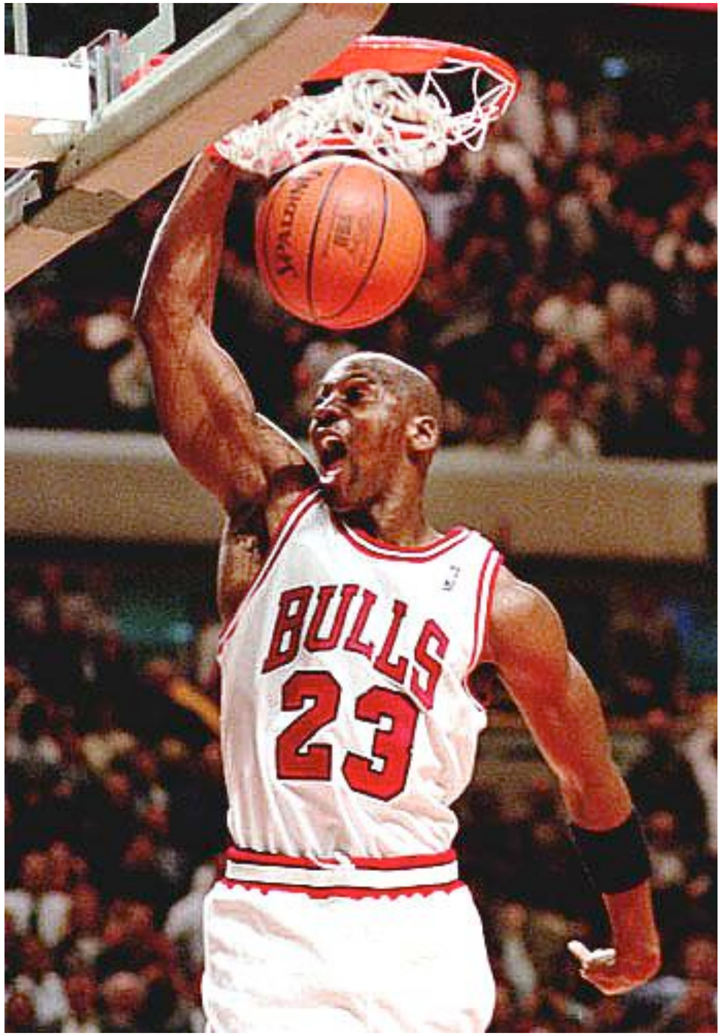
☑ **This item:** Introduction to Data Mining by Pang-Ning Tan

☑ Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems) by Eibe Frank

☑ The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Robert Tibshirani

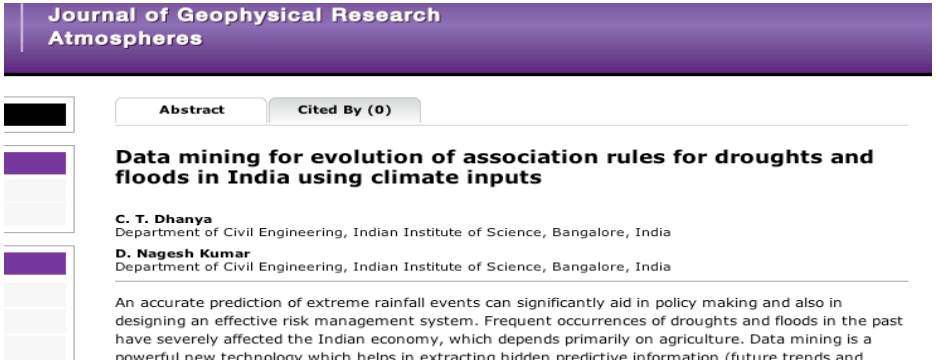
Frequent pattern discovery: Basketball scout

- In NBA, rigorous statistics are kept of events of play and the actions of all players
- Advanced Scout –system discovers interesting patterns
  - e.g.”When player X is on the field the shooting accuracy of player Y drops from 75% to 30%
- Bhandari I., Colet, E., Parker, J., Pines Z., Pratap R., Ramanujam K. (1997): Advanced Scout: datamining and knowledge discovery in NBA data. Data Mining and Knowledge Discovery, 1 (1), 121--125}



Frequent Pattern Discovery Application: Scientific discovery

- Explorative analysis of scientific data can reveal unexpected associations
- Especially useful in “weak theory” domains where human experts do not yet know all the relevant variables



Journal of Geophysical Research
Atmospheres

Data mining for evolution of association rules for droughts and floods in India using climate inputs

C. T. Dhanya
Department of Civil Engineering, Indian Institute of Science, Bangalore, India

D. Nagesh Kumar
Department of Civil Engineering, Indian Institute of Science, Bangalore, India

An accurate prediction of extreme rainfall events can significantly aid in policy making and also in designing an effective risk management system. Frequent occurrences of droughts and floods in the past have severely affected the Indian economy, which depends primarily on agriculture. Data mining is a powerful new technology which helps in extracting hidden predictive information (future trends and

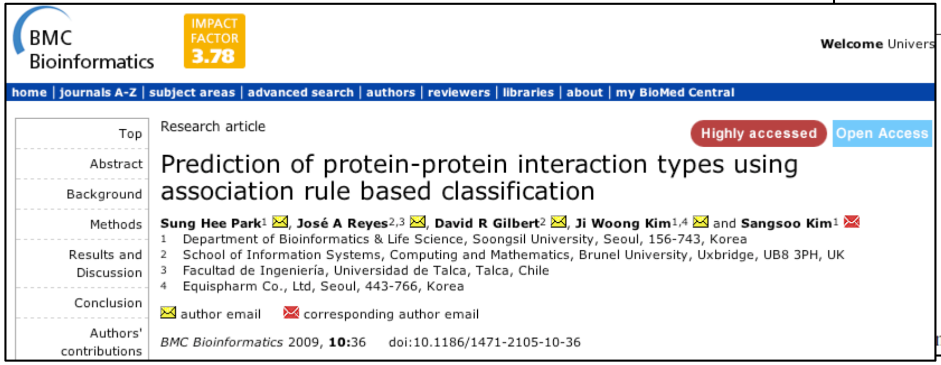
Behavior Research Methods
2007, 39 (2), 259-266

Articles

An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents

Dion H. Goh and Rebecca P. Ang
Nanyang Technological University, Singapore

Association rule mining (ARM) is a technique used to discover relationships among a large set of variables n a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely



Research article

Prediction of protein-protein interaction types using association rule based classification

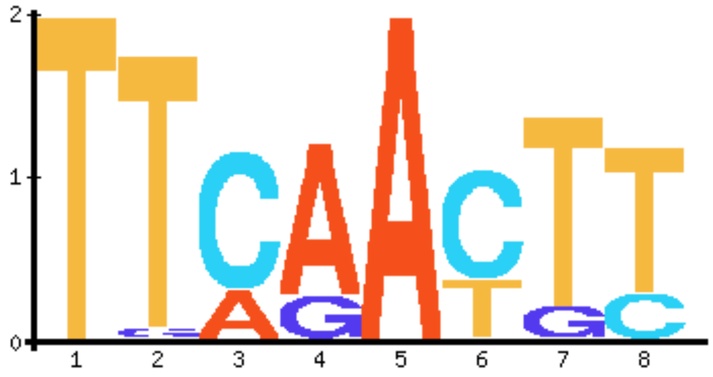
Sung Hee Park[1], José A Reyes[2,3], David R Gilbert[2], Ji Woong Kim[1,4] and Sangsoo Kim[1]
1 Department of Bioinformatics & Life Science, Soongsil University, Seoul, 156-743, Korea
2 School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, UB8 3PH, UK
3 Facultad de Ingeniería, Universidad de Talca, Talca, Chile
4 Equispharm Co., Ltd, Seoul, 443-766, Korea

✉ author email ✉ corresponding author email

BMC Bioinformatics 2009, **10**:36 doi:10.1186/1471-2105-10-36

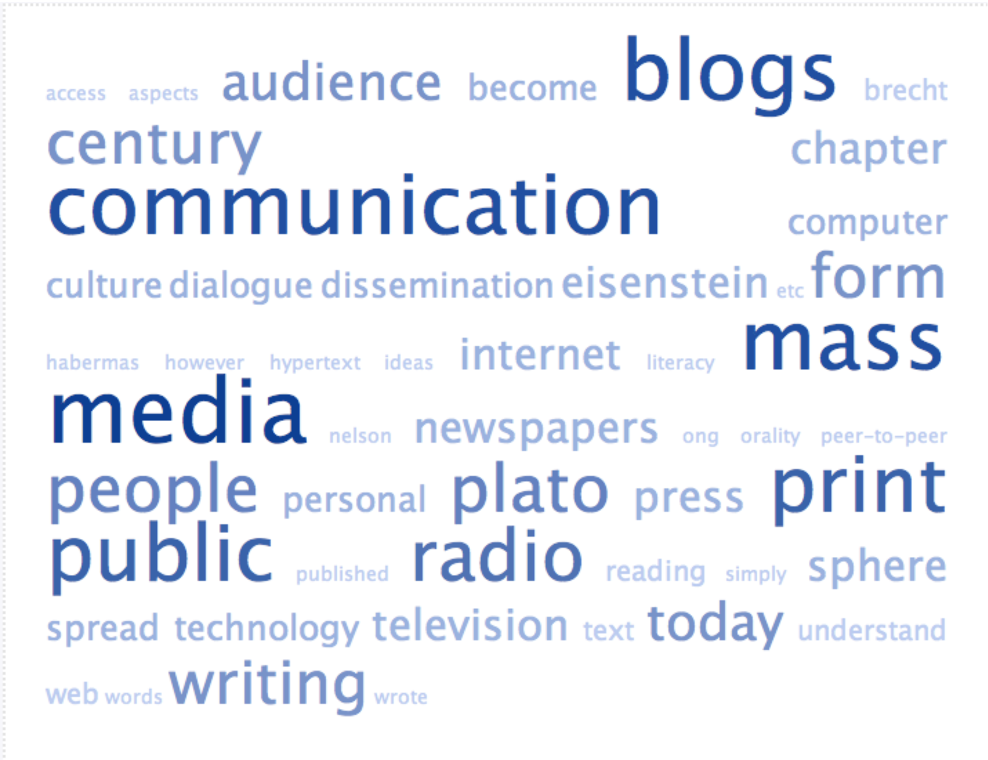
Frequent sequential patterns: Motif discovery in biosequences

- In DNA data, binding sites of regulatory proteins are characterized by distinct subsequences
- Biologically important motifs should occur more frequently than a random subsequence
- The patterns are typically not completely fixed but allow variability in certain locations
- Sequential pattern: order of items (letters) important



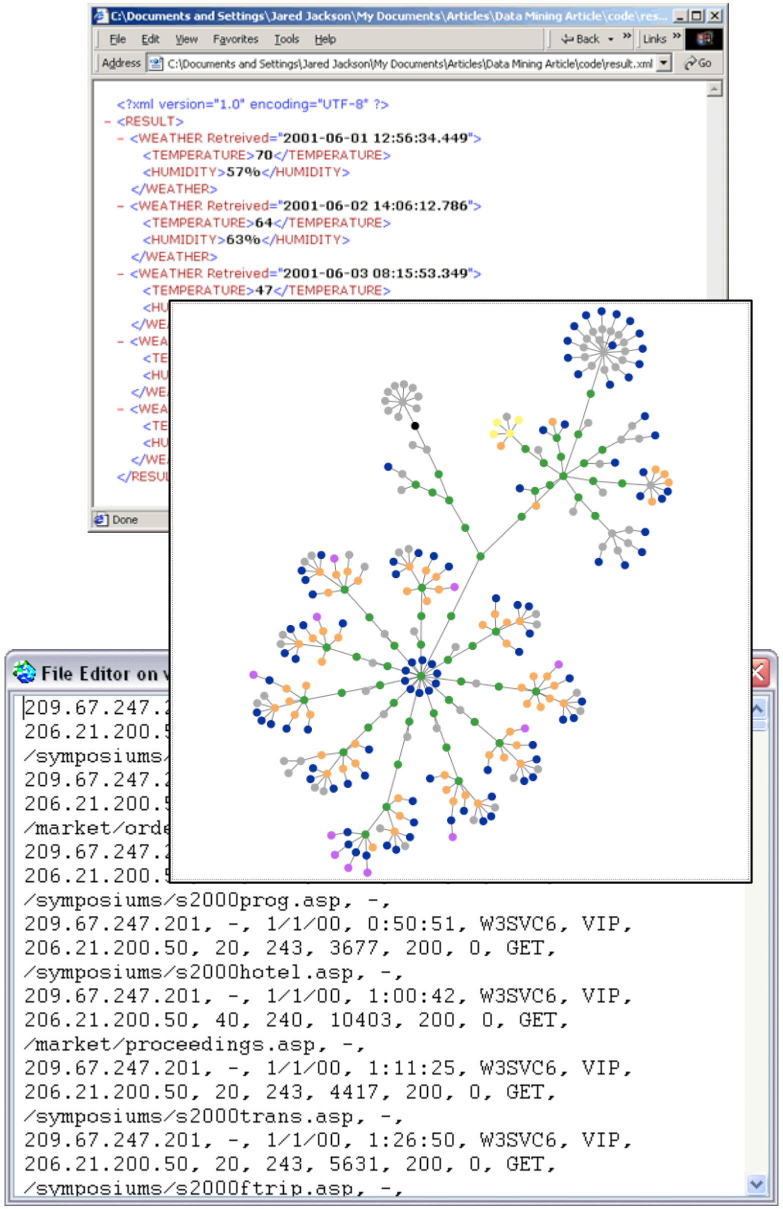
Text mining

- Data mining & Information retrieval for text
- Topics:
  - Text categorization, & clustering
  - Named entity extraction
  - Entity relation modeling
  - Sentiment analysis
- Many applications:
  - Biomedical data mining
  - Media monitoring
  - Marketing


access aspects audience become blogs brecht
century chapter
communication computer
culture dialogue dissemination eisenstein etc form
habermas however hypertext ideas internet literacy mass
media nelson newspapers ong orality peer-to-peer
people personal plato press print
public published radio reading simply sphere
spread technology television text today understand
web words writing wrote

Web mining

- Web provides many data mining challenges
- Topics:
  - Web Content Mining: mining the documents in the web
    - related field of text mining
  - Web Structure Mining: mining the link structure of the web
    - e.g. Google PageRank
  - Web Usage Mining: mining the log files of web for user modeling and web site optimization


C:\Documents and Settings\Jared Jackson\My Documents\Articles\Data Mining Article\code\res...
File
Edit
View
Favorites
Tools
Help
Back
Links
Address
C:\Documents and Settings\Jared Jackson\My Documents\Articles\Data Mining Article\code\result.xml
Go
Done
File Editor on v

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <RESULT>
  - <WEATHER Retreived="2001-06-01 12:56:34.449">
      <TEMPERATURE>70</TEMPERATURE>
      <HUMIDITY>57%</HUMIDITY>
    </WEATHER>
  - <WEATHER Retreived="2001-06-02 14:06:12.786">
      <TEMPERATURE>64</TEMPERATURE>
      <HUMIDITY>63%</HUMIDITY>
    </WEATHER>
  - <WEATHER Retreived="2001-06-03 08:15:53.349">
      <TEMPERATURE>47</TEMPERATURE>
      <HU
    </WE
  - <WEA
      <TE
      <HU
    </WE
  - <WEA
      <TE
      <HU
    </WE
  </RESUL
```

```
209.67.247.2
206.21.200.5
/symposiums/
209.67.247.2
206.21.200.5
/market/orde
209.67.247.2
206.21.200.5
/symposiums/s2000prog.asp, -,
209.67.247.201, -, 1/1/00, 0:50:51, W3SVC6, VIP,
206.21.200.50, 20, 243, 3677, 200, 0, GET,
/symposiums/s2000hotel.asp, -,
209.67.247.201, -, 1/1/00, 1:00:42, W3SVC6, VIP,
206.21.200.50, 40, 240, 10403, 200, 0, GET,
/market/proceedings.asp, -,
209.67.247.201, -, 1/1/00, 1:11:25, W3SVC6, VIP,
206.21.200.50, 20, 243, 4417, 200, 0, GET,
/symposiums/s2000trans.asp, -,
209.67.247.201, -, 1/1/00, 1:26:50, W3SVC6, VIP,
206.21.200.50, 20, 243, 5631, 200, 0, GET,
/symposiums/s2000ftrip.asp, -,
```

Frequent sequential patterns: web log mining

- Find rules that predict strong sequential dependencies among different events.
- Episode rules
  - "When webpage A is accessed, webpage B is accessed within 10 seconds, 70% of times"

Website1,	*Website2*	*Start*	*End*	*Conf (%)*
Citeseer.com,	**Rgtu.net**	**1:00**	**1:20**	**70**
Newsworld.com,	**Citeseer.com**	**2:00**	**2:10**	**75**
Citeseer.com,	**Rgtu.net**	**2:00**	**2:15**	**70**

Challenges of Data Mining

- Scalability
  - Analysis of terabytes of data requires efficient algorithms – parallelization of computation may be needed
  - All data may not fit to the main memory of the computer – need efficient index structures for secondary memory
- (Curse of) Dimensionality
  - Many datasets have thousands, even millions of attributes
  - Statistical problems: easy to find spurious patterns created by random effects
  - Computational problems: many methods scale badly when number of attributes increase

Challenges of Data Mining

- Complex and Heterogeneous Data
  - Web pages with hyperlink structure
  - Images
  - Streaming data: video, speech
  - Sequence data: DNA sequence data, Natural language data
  - Spatio-temporal data: e.g. earth surface temperature over time
- Data Ownership and Distribution
  - Data may not be stored in one geographical location or owned by a single institution
  - Security and Privacy Preservation

Challenges of Data Mining

- Non-traditional analysis
  - Traditional data analysis mostly deals with data arising from well-controlled data gathering process
    - Scientific experiments testing a hypothesis
    - Questionnaries to random population samples
  - Data mining many times analyses data collected for a different original purpose
    - Data may not contain the most interesting attributes
    - Data may be a biased sample of the population

Data mining process

- Data mining process consists of several interdependent steps
  - Preprocessing to make the data suitable for analysis
  - Data mining to find the patterns/build models
  - Postprocessing to make the results suitable for human analysis
- In reality: iterative process with feedback loops and human interaction

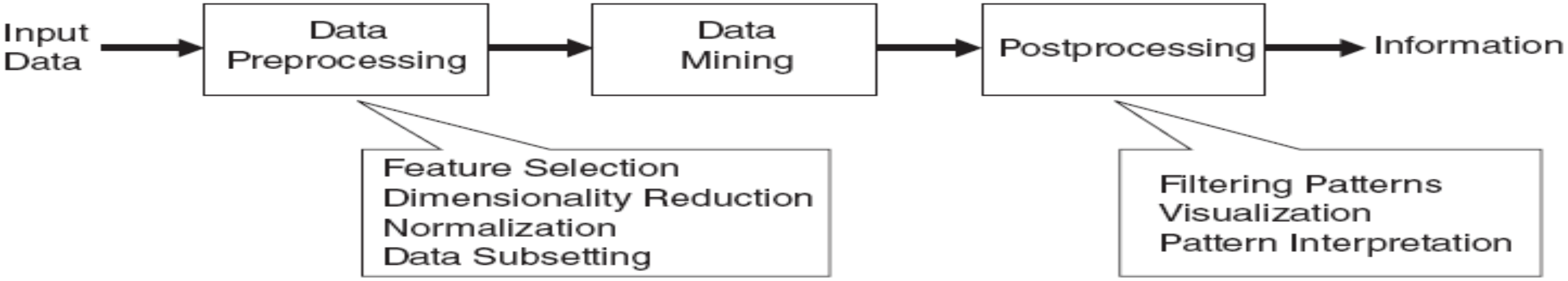


Figure 1.1. The process of knowledge discovery in databases (KDD).

Types of data

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

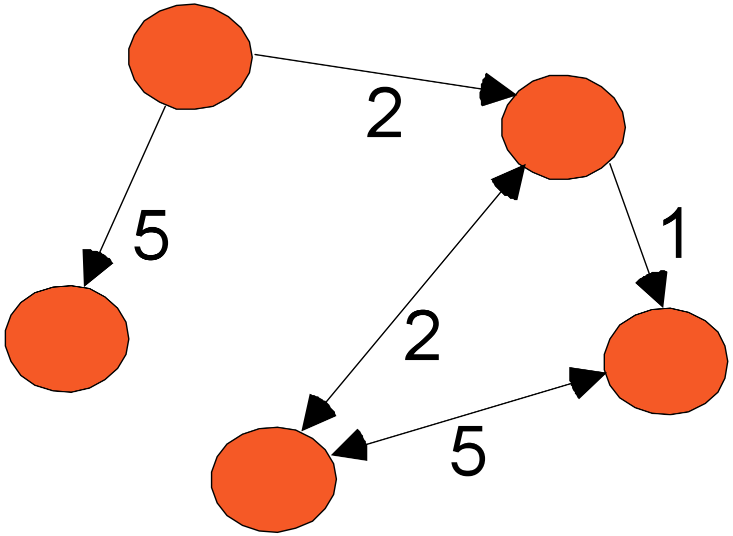
Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Graph Data

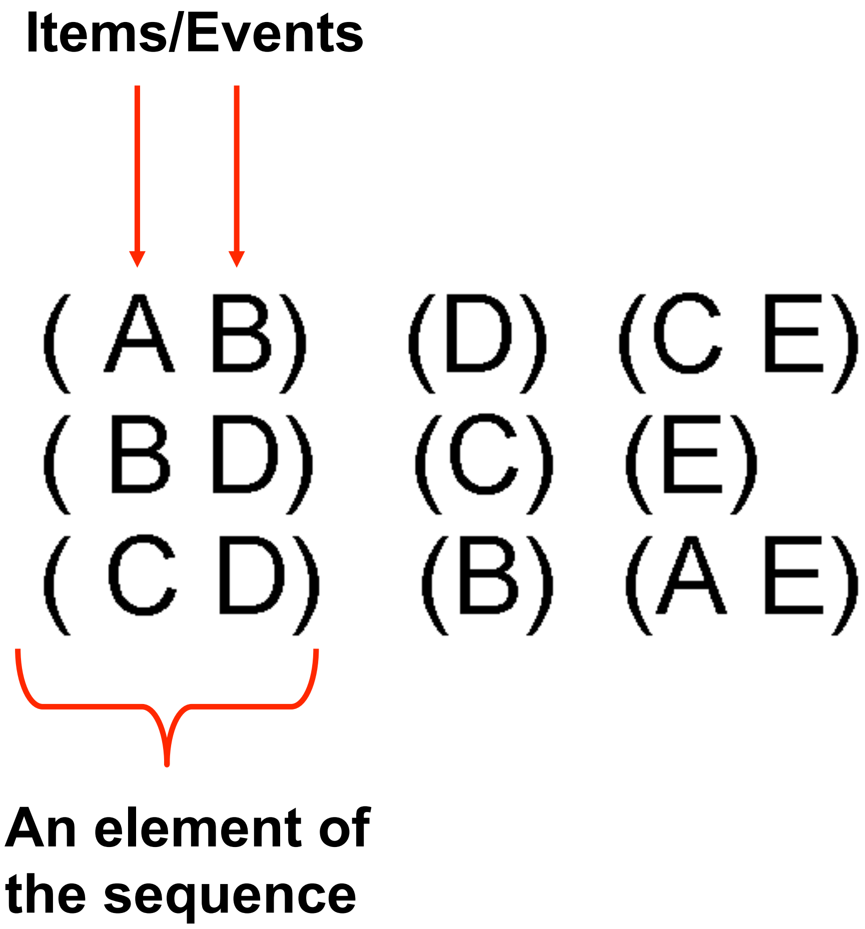
- Examples: Generic graph and HTML Links



```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Ordered Data

- Sequences of transactions



Items/Events

(A B) (D) (C E)

(B D) (C) (E)

(C D) (B) (A E)

An element of the sequence

Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean