Data Mining I
Introduction and Course Outline

Heiko Paulheim
Hello

• Prof. Dr. Heiko Paulheim
• Assistant Professor
• Research Interests:
  – Semantic Web and Linked Open Data
  – Data Mining with Linked Open Data
  – Ontology Matching
  – Data Quality and Data Cleaning
  – Outlier Detection
• Room: B6 – C1.08
• Consultation: by appointment
• Heiko will teach the lectures
Hello

• M.Sc. Wi.-Inf. Oliver Lehmberg
• Graduate Research Associate
• Research Interests:
  – Data and Web Mining
  – Network Analysis
  – Web Data Integration
• eMail: oli@informatik.uni-mannheim.de
• Oliver will teach the exercise at 1:45.
  and co-supervise the team projects.
Hello

• Kirli Gashterovski
• Graduate Research Associate
• Research Interests:
  – Data Mining
  – Pattern Extraction
  – NLP
  – Knowledge Extraction
• eMail: k.gashterovski@uni-mannheim.de
• Kiril will teach the exercise blocks at noon and co-supervise the team projects.
Introduction and Course Outline

• Course Outline and Organization
• What is Data Mining?
• Methods and Applications
• The Data Mining Process
Course Organization

• Lecture
  – introduces the principle methods of data mining
  – discusses how to evaluate generated models
  – presents practical examples of data mining applications from the corporate and Web context

• Exercise
  – students experiment with data sets using RapidMiner

• Project Work
  – teams of four students realize a data mining project
  – teams may choose their own data sets and tasks (in addition, we will propose some suitable data sets and tasks)
  – write summary about project, present project results

• Final exam
  – 50 % written exam
  – 50 % project work
# Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>07.09.</td>
<td>Introduction/Course Outline</td>
<td>Introduction to RapidMiner</td>
</tr>
<tr>
<td>14.09.</td>
<td>Lecture: Clustering</td>
<td>Exercise: Clustering</td>
</tr>
<tr>
<td>21.09.</td>
<td>Lecture: Classification 1</td>
<td>Exercise: Classification</td>
</tr>
<tr>
<td>28.09.</td>
<td>Lecture: Classification 2</td>
<td>Exercise: Classification</td>
</tr>
<tr>
<td>29.09.</td>
<td>Lecture: Classification 3</td>
<td>Exercise: Classification</td>
</tr>
<tr>
<td>05.10.</td>
<td>Lecture: Association Analysis</td>
<td>Exercise: Association Analysis</td>
</tr>
<tr>
<td>12.10.</td>
<td>Lecture: Text Mining</td>
<td>Exercise: Text Mining</td>
</tr>
<tr>
<td>19.10.</td>
<td>Intro Student Projects</td>
<td>Project Work</td>
</tr>
<tr>
<td>26.10.</td>
<td>Feedback: Student Projects</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>02.11.</td>
<td>Project work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>09.11.</td>
<td>Project work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>16.11.</td>
<td>Project work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>23.11.</td>
<td>Project work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>30.11.</td>
<td><em>Sunday night: report submission</em></td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>07.12.</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
</tbody>
</table>
Deadlines

• Submission of project work proposal
  – Sunday, October 25th, 23:59

• Submission of final project work report
  – Sunday, November 29th, 23:59

• Project presentations
  – schedule to be announced
  – everyone has to attend
Course Organization

• Lecture Webpage: Slides, Announcements
  – hint: look at version tags!

• Additional Material

• Time and Location
  – Lecture: Monday, 10.15 – 11.45, B6 23-25, A1.01
    • these are parallel groups, you only have to attend one
Course Organization

• Registration
  – if not yet done, please register online at ILIAS
  – ...and decide for one exercise group

• Policy: three strikes out
  – applies if there are >60 registered participants (there is a waiting list)
  – you have to attend at least one of the first three lectures (incl. today)
  – otherwise, we will give your place away

• If you are on the waiting list
  – you may be assigned a place after three weeks
  – waiting list is cleared after this semester (i.e., no priority next semester!)
# Lecture Contents

<table>
<thead>
<tr>
<th>Introduction to Data Mining</th>
<th>What is Data Mining? Methods and Applications The Data Mining Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>K-means Clustering, Hierarchical Clustering, Proximity Measures</td>
</tr>
<tr>
<td>Classification</td>
<td>Nearest Neighbor, Decision Trees, Rule Learning, Model Evaluation, Naïve Bayes, Support Vector Machines</td>
</tr>
<tr>
<td>Association Analysis</td>
<td>Frequent Item Set Generation, Rule Generation, Interestingness Measures</td>
</tr>
<tr>
<td>Text Mining</td>
<td>Preprocessing Text, Feature Creation, Feature Selection, RapidMiner Text Extension</td>
</tr>
<tr>
<td>Introduction to student projects</td>
<td>Overview of provided data sets Overview of proposed tasks</td>
</tr>
</tbody>
</table>
Literature & Slide Sources

• Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
  – 10 copies in university library.
  – we provide scans of important chapters via ILIAS

  – several copies in university library
  – we provide scans of important chapters via ILIAS
Literature & Slide Sources

  – several copies in university library
  – electronic edition available via the library

• Gregory Piatetsky-Shapiro, Gary Parker: KDNuggets Data Mining course:
  http://www.kdnuggets.com/dataMiningCourse/
Additional Material

- Video recordings from FSS 2015
Software

- Well-known data mining software
- Powerful data mining suite
- **We are using Version 6.5 in the exercise!**
- License key will be provided via ILIAS
   • Explains along case studies how to use simple and advanced Rapidminer features.
   • Website with data and processes: http://rapidminerbook.com

   • Free PDF version available online.

   • Introduction to user interface and basic features
   • http://rapidminer.com/learning/getting-started/
Outlook: Data Mining II

• Will be taught every FSS

• Topics
  – Regression
  – Sequential Pattern Mining, Time Series Prediction
  – Anomaly Detection
  – Online Data Analysis
  – Advanced Data Preprocessing

• Practical project
  – The annual Data Mining Cup
  – Worldwide competition of student teams
  – Real-world data mining tasks

• Bonus opportunity
  – Possibility of obtaining RapidMiner Analyst certification
Questions?
IE 560: Decision Support

– Decision-making is an important part of all science-based professions, where specialists apply their knowledge in a given area to make informed decisions.

– In the Lecture, we look at models that help to formulate and algorithmically solve decision making problems, i.e. that find a solution that maximizes the expected benefit of the outcome.

– Topics include:
  • Planning Problems and Algorithms
  • Probabilistic Graphical Models
  • Decision Theory and Decision Networks
  • Game Theory and Mechanism Design
IE 673: Data Mining and Matrices

- Matrices appear everywhere
  - Linear maps, data points, sets, graphs, relational data, ...

- Swiss army knife of data mining: *Matrix decompositions*
  - Analyzing and understanding matrices is central task
  - Denoising, discovery of latent structure, visualization, prediction, clustering, pattern mining, topic modelling, …

\[
\begin{align*}
\text{Customer transactions} & \\
\text{Anna} & \begin{pmatrix} 1 & 1 & 0 \end{pmatrix} & \text{Book 1} & \begin{pmatrix} 5 & 0 & 3 \end{pmatrix} \\
\text{Bob} & \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} & \text{Book 2} & \begin{pmatrix} 0 & 0 & 7 \end{pmatrix} \\
\text{Charlie} & \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} & \text{Book 3} & \begin{pmatrix} 4 & 6 & 5 \end{pmatrix}
\end{align*}
\]

\[
\begin{align*}
\text{Document-term matrix} & \\
\text{Avatar} & \begin{pmatrix} 4 & 2 \end{pmatrix} & \text{Saarbrücken} & \begin{pmatrix} 1 & 11 & 10 \end{pmatrix} \\
\text{Bob} & \begin{pmatrix} 3 & 2 \end{pmatrix} & \text{Helsinki} & \begin{pmatrix} 6.5 & 10.9 & 8.7 \end{pmatrix} \\
\text{Charlie} & \begin{pmatrix} 5 & 3 \end{pmatrix} & \text{Cape Town} & \begin{pmatrix} 15.7 & 7.8 & 8.7 \end{pmatrix}
\end{align*}
\]

\[
\begin{align*}
\text{Incomplete rating matrix} & \\
\text{Avatar} & \begin{pmatrix} 4 & 2 \end{pmatrix} & \text{Saarbrücken} & \begin{pmatrix} 1 & 11 & 10 \end{pmatrix} \\
\text{Bob} & \begin{pmatrix} 3 & 2 \end{pmatrix} & \text{Helsinki} & \begin{pmatrix} 6.5 & 10.9 & 8.7 \end{pmatrix} \\
\text{Charlie} & \begin{pmatrix} 5 & 3 \end{pmatrix} & \text{Cape Town} & \begin{pmatrix} 15.7 & 7.8 & 8.7 \end{pmatrix}
\end{align*}
\]

\[
\begin{align*}
\text{Cities and monthly temperatures} & \\
\text{Avatar} & \begin{pmatrix} 4 & 2 \end{pmatrix} & \text{Saarbrücken} & \begin{pmatrix} 1 & 11 & 10 \end{pmatrix} \\
\text{Bob} & \begin{pmatrix} 3 & 2 \end{pmatrix} & \text{Helsinki} & \begin{pmatrix} 6.5 & 10.9 & 8.7 \end{pmatrix} \\
\text{Charlie} & \begin{pmatrix} 5 & 3 \end{pmatrix} & \text{Cape Town} & \begin{pmatrix} 15.7 & 7.8 & 8.7 \end{pmatrix}
\end{align*}
\]
IE 660: Text Analytics

- Methods to automatically process natural language from a computational / algorithmic perspective

**Key topics**

- Modeling the role and function of words within sentence structure (morphology, syntax/parsing)
  - Understanding meaning in context (computational semantics)
  - Applications
    - Information extraction
    - Machine Translation
    - Topic modeling
Data integration is the process of consolidating data from a set of heterogeneous data sources into a single uniform representation.

Course Topics
1. The Data Integration Process
2. Web Data Formats
3. Schema Mapping and Data Translation
   - How to normalize the structure of data?
4. Identity Resolution
   - How to find records in different data sources describing the same real-world entity?
5. Data Quality and Data Fusion
   - How to combine data from many sources about a single entity?
IE674: Hot Topics in Machine Learning

• New Lecture by Rainer Gemulla and Laura Dietz
  – Basics of machine learning and probability theory
  – Probabilistic graphical models
  – Inference and parameter estimation
  – Neural networks
• We are looking for students who
  – are fluent in Java
  – like data mining
  – want to work on cutting-edge research
  – want to get their hands dirty with web data
• Get in touch with us!
A Bit of History

• **We are drowning in data, but starving for knowledge.**
  
  (John Naisbitt, 1982)

• **Computers have promised us a fountain of wisdom but delivered a flood of data.**

• **It has been estimated that the amount of information in the world doubles every 20 months.**

  (Frawley, Piatetsky-Shapiro, Matheus, 1992)
“We are Drowning in Data...”

More and more data is generated:

- Transaction data from banking, telecommunication, e-commerce
- Scientific data from astronomy, physics, biology
- The public Web, Twitter, ...
- Social network sites
- Application logs
Data, Information, Knowledge, and Wisdom

Gene Bellinger, Durval Castro and Anthony Mills. "Transforming Data to Wisdom."
A Historical Example

- Cholera disease
- From beginning of 19th century
- ~100,000 deaths per year
  - until today!
- For a long time, there was little knowledge
  - on ways of infection
  - on causes of the disease

http://fieldnotes.unicefusa.org/2008/09/newsnet_combating_cholera_1.html
A Historical Example

- August Heinrich Petermann
- 1822-1878
- Geographer and Cartographer
- Geographic maps as a means
  - to understand data
  - to gather knowledge

http://commons.wikimedia.org/wiki/File:August_Heinrich_Petermann.jpg
A Historical Example

• 1848 map of Cholera deaths in London
  – finding: Cholera is more likely in densely populated areas
  – where there is no functioning sewage system
  – conclusion: Cholera is transmitted through contaminated water

A More Recent Example: SARS

- SARS: Severe acute respiratory syndrome
- Outbreak: 2012 in Hong Kong

A More Recent Example: SARS

- Which paths do SARS infection take?
- Max Planck Institute for Dynamics and Self-Organization:
  - SARS infections follow major airline routes

[Image: Map showing airline routes with a color scale indicating the number of infections, with a URL: http://www.mpg.de/483574/pressemitteilung20041014]
A Very Recent Example: the NSA

- Communication data from all over the world
- Searching for suspects and terrorists

http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data
“We are Drowning in Data…”

The following slides are taken from Aidan Hogan's course on “Massive Data Processing”

**Wikipedia**

\[ \approx 5.9 \text{ TB of data} \]

*(Jan. 2010 Dump)*

1 Wiki = 1 Wikipedia
“We are Drowning in Data...”

Human Genome
≈ 4 GB/person
≈ 0.0006 Wiki/person
“We are Drowning in Data…”

US Library of Congress
≈ 235 TB archived
≈ 40 Wiki
"We are Drowning in Data..."

Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year
≈ 12 Wiki/year
“We are Drowning in Data...”

NASA Center for Climate Simulation
≈ 32 PB archived
≈ 5,614 Wiki
“We are Drowning in Data...”

Facebook
≈ 12 TB/day added
≈ 2 Wiki/day
≈ 782 Wiki/year
(as of Mar. 2010)
“We are Drowning in Data...”

Large Hadron Collider
≈15 PB/year
≈2,542 Wikipedias/year
“We are Drowning in Data…”

**Google**
- ≈ 20 PB/day processed
- ≈ 3,389 Wiki/day
- ≈ 7,300,000 Wiki/year
  (Jan. 2010)
“We are Drowning in Data...”

Internet (2016)

\[ \approx 1.3 \text{ ZB/year} \]
\[ \approx 220,338,983 \text{ Wiki/year} \]

(2016 IP traffic; Cisco est.)
...but starving for knowledge!

←Rate at which data are produced

←Rate at which data can be understood
manual interpretation is hardly feasible!
Data Mining: Definitions

- Idea: mountains of data
  - where knowledge is mined
Data Mining: Definitions

• Data Mining is a non-trivial process of identifying
  – valid
  – novel
  – potentially useful
  – ultimately understandable
patterns in data.

(Fayyad et al. 1996)

• Data mining is nothing else than torturing the data until it confesses
  (Fred Menger, year unknown)
• ...and if you torture it enough, you can get it to confess to anything.
Origins of Data Mining

- Draws ideas from machine learning, statistics, and database systems.
- Traditional techniques may be unsuitable due to
  - large amount of data
  - high dimensionality of data
  - heterogeneous, distributed nature of data
Data Mining Application Fields

• Business
  – Customer relationship management, e-commerce, fraud detection, manufacturing, telecom, targeted marketing, health care, …

• Science
  – Data mining helps scientists to analyze data and to formulate hypotheses.
  – Astronomy, physics, bioinformatics, drug discovery, …

• Web and Social Media
  – Advertising, search engine optimization, spam detection, web site optimization, personalization, sentiment analysis, …

• Government
  – Surveillance, crime detection, profiling tax cheaters, …
A Hype Topic: Big Data

• Everybody can analyze large amounts of data at low costs in the cloud

• Technical realization
  – massive parallelization using hundreds or thousands of machines
  – using tools like Hadoop, Hive, Hbase, Mahout, ...

• Open Data and Data Market Places
  – infochimps: >14,000 data sets
  – CKAN DataHub: >5,000 data sets
Data Mining Methods

• Descriptive methods
  – find patterns in data
  – e.g., which products are often bought together?

• Predictive methods
  – predict unknown or future values of a variable
    • given observations (e.g., from the past)
  – e.g., will a person click a banner?
    • given his/her browsing history

• Machine learning terminology:
  – descriptive = unsupervised
  – predictive = supervised
Data Mining Tasks

- Clustering (descriptive)
- Classification (predictive)
- Association Rule Mining (descriptive)
- Text Mining (both descriptive and predictive)

- Covered in Data Mining 2
  - Regression (predictive)
  - Anomaly Detection (descriptive)
  - Sequential Pattern Mining (descriptive)
  - Time Series Prediction (predictive)
Clustering

• Given a set of data points, and a similarity measure among them, find clusters such that
  – Data points in one cluster are similar to one another
  – Data points in separate clusters are different from each other

• Result
  – a descriptive grouping of data points
Clustering: Applications

• Application area: Market segmentation
• Goal: Subdivide a market into distinct subsets of customers
  – where any subset may be conceived as a marketing target to be reached with a distinct marketing mix

• Approach:
  – Collect information about customers
  – Find clusters of similar customers
  – Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters
Clustering: Applications

• Application area: Document Clustering
• Goal: Find groups of documents that are similar to each other based on the important terms appearing in them
• Approach
  – Identify frequently occurring terms in each document
  – Define a similarity measure based on the frequencies of different terms
• Application Example: Grouping of stories in Google News
Classification

• Given a collection of records (training set)
  – each record contains a set of attributes
  – one of the attributes is the class (label) that should be predicted
• Find a model for class attribute as a function of the values of other attributes
• Goal: previously unseen records should be assigned a class as accurately as possible
  – A test set is used to validate the accuracy of the model
  – Training set may be split into training and validation data
### Classification Example

#### Class/Label Attribute

<table>
<thead>
<tr>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Single</td>
<td>75K</td>
<td>?</td>
</tr>
<tr>
<td>Yes</td>
<td>Married</td>
<td>50K</td>
<td>?</td>
</tr>
<tr>
<td>No</td>
<td>Married</td>
<td>150K</td>
<td>?</td>
</tr>
<tr>
<td>Yes</td>
<td>Divorced</td>
<td>90K</td>
<td>?</td>
</tr>
<tr>
<td>No</td>
<td>Single</td>
<td>40K</td>
<td>?</td>
</tr>
<tr>
<td>No</td>
<td>Married</td>
<td>80K</td>
<td>?</td>
</tr>
</tbody>
</table>

#### Training Set

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

#### Unseen Data

- Model
- Learn Classifier
- Training Set
Classification: Applications

• Application area: Direct Marketing
• Goal: Reduce cost of mailing by targeting a set of consumers which are likely to buy a new cell phone
• Approach:
  – Use the data for a similar product introduced before
  – We know which customers decided to buy and which did not
  – Collect various demographic, lifestyle, and company-interaction related information about all such customers
    • Type of business, where they stay, how much they earn, etc.
  – Use this information as input attributes to learn a classifier model
Classification: Applications

- Application area: Fraud Detection
- Goal: Recognize fraudulent cases in credit card transactions
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes
    - When and where does a customer buy? What does he buy?
    - How often he pays on time? etc.
  - Label past transactions as *fraud* or *fair* transactions
    This forms the *class attribute*
  - Learn a model for the class of the transaction
  - Use this model to detect fraud by observing credit card transactions on an account
Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
- produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, M</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Di</td>
</tr>
</tbody>
</table>

Rules Discovered
{Diaper, Milk} → {Beer}
{Milk} → {Coke}
Association Rule Discovery: Applications

- Application area: Marketing and Sales Promotion
- Example rule discovered:
  \{Bagels, Coke\} \rightarrow \{Potato Chips\}
- Insights:
  - promote bagels to boost potato chips sales
  - if selling bagels is discontinued, this will affect potato chips sales
  - coke should be sold together with bagels to boost potato chips sales
Association Rule Discovery: Applications

- Content-based recommendation
  - requirement: much data
  - e.g., Amazon transactions, Spotify logfiles
Association Rule Discovery: Applications

- Real world example:
  - Customer loyalty programs


Association Rule Discovery: Applications

• Real example:
  – Target (American grocery store)
  – Analyzes customer buying behavior
  – Sends personalized advertisement

• Famous case in the USA:
  – Teenage girl gets advertisement for baby products
  – ...and her father is mad

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
Association Rule Discovery: Applications

• Bottom line of the Target teenage girl story:
  – Janet Vertesi, Princeton university
  – Tried to hide her pregnancy from computers

• Measures taken:
  – using Tor for online surfing
  – no social media posts about her pregnancy
  – paying all pregnancy/baby related products in cash
  – a fresh Amazon account delivering to a local locker
    • paying with cash-payed gift cards

• Outcome:
  – massive buying of gift cards in a convenience store was reported to tax authorities

read the full story at http://mashable.com/2014/04/26/big-data-pregnancy/
The Data Mining Process

Source: Fayyad et al. (1996)
The Data Mining Process

• Note that none of those steps actually requires a computer

• Recall Petermann's Cholera maps
  – Data Selection: find data on cholera deaths
  – Data Preprocessing: organize data by geographic area
  – Transformation: draw data on a map
  – Data Mining: look at the map and find patterns
    • possibly step back: add more data (population, water system, ...)
  – Interpretation: Cholera is transmitted via contaminated water

• However, computers make things easier
  – mainly: scalability
  – large data sets
  – large number of possible patterns
Selection and Exploration

• Selection
  – What data is available?
  – What do I know about the provenance of this data?
  – What do I know about the quality of the data?

• Exploration
  – Get an initial understanding of the data
  – Calculate basic summarization statistics
  – Visualize the data
  – Identify data problems such as outliers, missing values, duplicate records
Selection and Exploration

• Visual Data Mining
  – For example as maps
  – Map showing the most popular photo locations in the world, generated from Panoramio logs

http://www.sightsmap.com/
Preprocessing and Transformation

• Transform data into a representation that is suitable for the chosen data mining methods
  – number of dimensions
  – scales of attributes (nominal, ordinal, numeric)
  – amount of data (determines hardware requirements)

• Methods
  – Aggregation, sampling
  – Dimensionality reduction / feature subset selection
  – Attribute transformation / text to term vector
  – Discretization and binarization

• Good data preparation is key to producing valid and reliable models

• Data preparation estimated to take 70-80% of the time and effort of a data mining project!
Data Mining

- Input: Preprocessed Data
- Output: Model / Patterns

1. Apply data mining method
2. Evaluate resulting model / patterns
3. Iterate:
   - Experiment with different parameter settings
   - Experiment with different alternative methods
   - Improve preprocessing and feature generation
   - Combine different methods
Deployment

- Use model in the business context

CRISP-DM Process Model
Questions?