Hello

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

• Dipl.-Wi.-Inf. Robert Meusel
• Graduate Research Associate
• Research Interests:
  – Data and Web Mining
  – Social Network Analysis
  – Linked Data Technologies
• Robert will teach the exercise blocks and co-supervise the team projects
Hello

• Dipl.-Wi.-Inf. Max Schmachtenberg
• Graduate Research Associate
• Research Interests:
  – Data Profiling
  – Data Cataloging
  – Linked Data Technologies
• Max will teach the exercise blocks 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 two/three 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>Tuesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>02.09.</td>
<td>Introduction/Course Outline</td>
<td>Introduction to RapidMiner</td>
</tr>
<tr>
<td>09.09.</td>
<td>Lecture: Clustering</td>
<td>Exercise: Clustering</td>
</tr>
<tr>
<td>16.09.</td>
<td>Lecture: Classification 1</td>
<td>Lecture: Classification 2</td>
</tr>
<tr>
<td>23.09.</td>
<td>Exercise: Classification</td>
<td>Lecture: Classification 3</td>
</tr>
<tr>
<td>30.09.</td>
<td>Exercise: Classification (ctd.)</td>
<td><em>public holiday</em></td>
</tr>
<tr>
<td>07.10.</td>
<td>Exercise: Evaluation</td>
<td>Association Analysis</td>
</tr>
<tr>
<td>14.10.</td>
<td>Exercise: Association Analysis</td>
<td>Text Mining</td>
</tr>
<tr>
<td>21.10.</td>
<td>Exercise: Text Mining</td>
<td>Intro Student Projects</td>
</tr>
<tr>
<td>28.10.</td>
<td>Project work</td>
<td>Feedback: Student Projects</td>
</tr>
<tr>
<td>04.11.</td>
<td>Project work</td>
<td>Project Work</td>
</tr>
<tr>
<td>11.11.</td>
<td>Project work</td>
<td>Project work</td>
</tr>
<tr>
<td>18.11.</td>
<td>Project work</td>
<td>Project work</td>
</tr>
<tr>
<td>25.11.</td>
<td>Project work</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>02.12.</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
</tbody>
</table>
Deadlines

• Submission of project work proposal
  – Monday, November 4th, 23:59

• Submission of final project work report
  – Monday, November 25th, 23:59
# 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>
Course Organization

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

• Additional Material

• Time and Location
  – Tuesday, 15.30 – 17.00
    Room A101 in B6
  – Thursday, 10.15 – 11.45
    Room A101 in B6

• Bring a laptop each time!
Course Organization

• Waiting list
  – this year, there are >30 people on the waiting list
  – if you decide not to attend, please leave ILIAS group
  – so others can have your place

• Policy
  – you have to decide until Tuesday next week
  – if you do not leave until then, you will have to stay
    (moral obligation towards those who did not get *your* place)

• If you are on the waiting list
  – you may get assigned a place until next week
  – waiting list is cleared after this semester
    (i.e., no priority next semester!)
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/data_mining_course/
• Powerful open-source data mining suite
• Download: http://rapid-i.com/
Outlook: Data Mining II

• New lecture, starting in 2014
• Will be taught every FSS
• Topics
  – Regression
  – Sequential Pattern Mining, Time Series Prediction
  – Online Data Analysis
  – Advanced Data Preprocessing
• Practical project
  – The annual Data Mining Cup
  – Worldwide competition of student teams
  – Real-world data mining tasks
Questions?
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
"We are Drowning in Data..."

• Example: the ATLAS system at CERN
• Purpose: search for the Higgs Boson
• Data: 320 MB/s (i.e., 27 TB per day!)
“We are Drowning in Data…”

• So all the data is there
• Management and storage is tricky, but only part of the story
• So how is the Higgs Boson discovered?

Image: CERN
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

http://www.mpg.de/483574/pressemitteilung20041014
A Very Recent Example

• 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
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: 5.100 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
  – 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)
- Regression (predictive)
- Association Rule Mining (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.

• Similarity Measures
  – Euclidean distance if attributes are continuous.
  – Other problem-specific measures.

• Goals
  – Intracluster distances are minimized
  – Intercluster distances are maximized
Clustering: Application 1

- 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 different attributes of customers based on their geographical and lifestyle related information
  - Find clusters of similar customers
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters
Clustering: Application 2

• 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

<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>

Refund | Marital Status | Taxable Income | Cheat
--------|----------------|----------------|-------
No      | Single         | 75K            | ?     |
Yes     | Married        | 50K            | ?     |
No      | Married        | 150K           | ?     |
Yes     | Divorced       | 90K            | ?     |
No      | Single         | 40K            | ?     |
No      | Married        | 80K            | ?     |

Unseen Data

Training Set

Learn Classifier

Model
Classification: Application 1

• 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 decided otherwise. This \{buy, don’t buy\} decision forms the class attribute
  – 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: Application 2

- 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
Classification: Application 3

• Application area: Customer Relationship Management
• Goal: To predict whether a customer is likely to be lost to a competitor
• Approach:
  – Use detailed record of transactions with each of the past and present customers, to find attributes
    • How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
  – Label the customers as loyal or disloyal
  – Find a model for loyalty
Regression

• Predict a value of a given *continuous* valued variable based on the values of other variables

• Greatly studied in statistics and neural network field

• Examples:
  – Predicting sales amounts of new product based on advertising expenditure
  – Predicting wind velocities as a function of temperature, humidity, air pressure, etc

• Difference to classification:
  – The label attribute is *continuous*
  – Classification is used for *nominal* labels (e.g., yes/no)
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: Application 1

- Application area: Marketing and Sales Promotion
- Let the rule discovered be
  \{\text{Bagels, \ldots}\} \rightarrow \{\text{Potato Chips}\}
- Potato Chips as consequent: Can be used to determine what should be done to boost its sales
- Bagels in the antecedent: Can be used to see which products would be affected if the store discontinues selling bagels
- Bagels in antecedent and Potato chips in consequent: Can be used to see what products should be sold with Bagels to promote sale of

![Frequently Bought Together](amazon.com)

Price For All Three: $87.41

Add all three to Cart  Add all three to Wish List

Show availability and shipping details
Association Rule Discovery: Application 1

• Content-based recommendation
  – requirement: much data
  – e.g., Amazon transactions, Spotify logfiles
Association Rule Discovery: Application 1

- Real world example:
  - Customer loyalty programs

![Bar chart showing the number of distributed bonus cards with multiple partner companies in Germany in millions of pieces](http://de.statista.com/statistik/daten/studie/36618/umfrage/anzahl-herausgegebener-bonuskarten-mehrere-partnerunternehmen/)

Association Rule Discovery: Application 1

• 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: Application 2

- Application area: Supermarket shelf management
- Goal: Identify items that are bought together by sufficiently many customers
- Approach: Process the point-of-sale data collected with bar code scanners to find dependencies among items
- A classic rule:
  - If a customer buys diaper and milk, then he is very likely to buy beer
  - So, don’t be surprised if you find six-packs stacked next to diapers!
Association Rule Discovery: Application 3

- Application area: Inventory Management
- Consumer appliance repair company
  - visits customers for repairs
  - goal: have tools and parts on board
  - to minimize number of visits
- Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns
Sequential Pattern Discovery: Definition

- Given a sequence of events (or sets of objects), find typical patterns:
  - (A,B) (C) (D,E)
  - (A) (B,C) (D)
  - (C) (A,D) (B)
  - Typical pattern: C → D
Sequential Pattern Mining: Application 1

- Web usage mining (navigation analysis)
- Input
  - Server logs
- Patterns
  - typical sequences of pages
- Usage
  - restructuring web sites
Sequential Pattern Mining: Application 2

• Recurring customers
  – Typical book store example:
    • (Twilight) (New Moon) → (Eclipse)

• Recommendation in online stores

• Allows more fine grained suggestions than frequent pattern mining

• Example:
  – mobile phone → charger vs. charger → mobile phone
    • are indistinguishable by frequent pattern mining
  – customers will select a charger after a mobile phone
    • but not the other way around!
    • however, Amazon does not respect sequences...
Time Series Prediction: Definition

- Given a sequence of events
  - predict the next event(s)

<table>
<thead>
<tr>
<th>Day</th>
<th>Weather</th>
<th>Temperature</th>
<th>Wind Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>Sunny</td>
<td>28°C</td>
<td>13 km/h</td>
</tr>
<tr>
<td>Tuesday</td>
<td>Cloudy</td>
<td>25°C</td>
<td>18 km/h</td>
</tr>
<tr>
<td>Wednesday</td>
<td>Cloudy</td>
<td>26°C</td>
<td>21 km/h</td>
</tr>
<tr>
<td>Thursday</td>
<td>Rain</td>
<td>19°C</td>
<td>35 km/h</td>
</tr>
<tr>
<td>Friday</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Saturday</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Sunday</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Time Series Prediction: Simple Approach

http://xkcd.com/605/
Time Series Prediction: Application

- Stock market prediction
- Computer trading

The Data Mining Process

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

• Note that none of the 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?