Data Mining I

Organization and Course Outline
Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Data and Web Mining
  - Web Data Integration
  - Linked Data Technologies
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30

- I will hold the lecture introducing the principle methods of data mining.
Hallo

- **Dr. Volha Bryl**
- Postdoctoral Researcher
- Research Interests
  - Data Mining
  - Data Quality and Data Cleaning
  - Linked Data Technologies
  - Natural Language Processing
- Room: B6, 26, C 1.02
- eMail: volha@informatik.uni-mannheim.de

  - Volha will teach the exercise group at 12:00 and co-supervise the student projects.
Hallo

- Dipl.-Wi.-Inf. Oliver Lehmberg
- Graduate Research Associate
- Research Interests:
  - Data and Web Mining
  - Network Analysis
  - Web Data Integration
- Room: B6, 26, C 1.04
- eMail: oli@informatik.uni-mannheim.de
- Oliver will teach the exercise group at 10:15 and co-supervise the student projects.
Today’s Lecture

1. Course Organization and Outline

2. Introduction to Data Mining
   1. What is Data Mining?
   2. Methods and Applications
   3. 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.

- **Two alternative Exercise Groups**
  - students experiment with data sets using RapidMiner
  - you need to register for one of them

- **Project Work**
  - teams of three to four students realize a data mining project
  - teams may choose their own data sets and tasks (in addition, I will propose some suitable data sets and tasks)
  - teams write a summary about their project and present the project results

- **Grading**
  - 50 % written exam, 50 % project work
Course Organization

- Course Webpage
  - provides up-to-date information, lecture slides, and exercise material

- Solutions to the Exercises
  - ILIAS eLearning System, https://ilias.uni-mannheim.de/

- Time and Location
  - Lecture: Wednesday, 10.15 - 11.45
    Room A5, B144
  - Exercise: Thursday, 10.15 - 11.45
    Room B6, A104 (Oliver Lehmberg)
  - Alternative exercise:
    Thursday, 12.00 - 13.30,
    Room B6, A101 (Volha Bryl)
Waiting List

- This year, we again have people on the waiting list
  - if you decide not to attend, please leave ILIAS group this week
  - so others can have your place

- Policy: Two strikes out
  - you have to attend this lecture or the lecture next Wednesday (18.02.2015)
  - If you are not attending, you will be deleted from the participants list.

- If you are on the waiting list
  - you may get assigned a place next week
  - The waiting list is cleared after this semester (i.e., no priority for next semester!)
Lecture Recording

- This semester, all lectures are recorded on video
  - you can use the videos for preparing the exam
  - we will also give students in upcoming semesters access to the videos

- You are required to sign a form saying that
  - it is OK that your questions are recorded
  - it is OK that you appear by accident on the video
  - if you do not like to sign the form, please sit behind the camera

- Lots of thanks to the Stabsstelle Studium und Lehre for doing the recording.
## Lectures Contents

| 1. Introduction to Data Mining | What is Data Mining?  
|                              | Methods and Applications  
|                              | The Data Mining Process  
| 2. Clustering                | K-means Clustering, Density-based Clustering,  
|                              | Hierarchical Clustering, Proximity Measures  
| 3. Classification            | Nearest Neighbor, Decision Trees  
|                              | Model Evaluation, Rule Learning, Naïve Bayes,  
|                              | Support Vector Machines  
| 4. Association Analysis      | Frequent Item Set Generation, Rule Generation  
|                              | Interestingness Measures  
| 5. Text Mining               | Preprocessing Text, Feature Generation, Feature  
|                              | Selection, RapidMiner Text Extension  
| 6. Introduction to student projects | Requirements and Organization  
|                              | Overview of proposed data sets and tasks  

<table>
<thead>
<tr>
<th>Date</th>
<th>Wednesday: Lecture</th>
<th>Thursday: Exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.02.2015</td>
<td>Introduction to Data Mining</td>
<td>Introduction to RapidMiner</td>
</tr>
<tr>
<td>18.02.2015</td>
<td>Lecture Clustering</td>
<td>Exercise Clustering</td>
</tr>
<tr>
<td>25.02.2015</td>
<td>Lecture Classification 1</td>
<td>Exercise Classification</td>
</tr>
<tr>
<td>04.03.2015</td>
<td>Lecture Classification 2</td>
<td>Exercise Classification</td>
</tr>
<tr>
<td>11.03.2015</td>
<td>Lecture Classification 3</td>
<td>Exercise Classification</td>
</tr>
<tr>
<td>18.03.2015</td>
<td>Lecture Association Analysis</td>
<td>Exercise Association Analysis</td>
</tr>
<tr>
<td>25.03.2015</td>
<td>Lecture Text Mining</td>
<td>Exercise Text Mining</td>
</tr>
<tr>
<td></td>
<td><strong>Easter break</strong></td>
<td></td>
</tr>
<tr>
<td>15.04.2015</td>
<td>Introduction to Student Projects</td>
<td>Preparation of Project Outline</td>
</tr>
<tr>
<td>22.04.2015</td>
<td>Feedback Student Projects</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>29.04.2015</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>06.05.2015</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>13.05.2015</td>
<td>Project Work</td>
<td>- Public Holiday -</td>
</tr>
<tr>
<td>20.05.2015</td>
<td>Project Work</td>
<td>Submission of project summaries</td>
</tr>
<tr>
<td>27.05.2015</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
<tr>
<td>05.06.2015</td>
<td>Written Exam</td>
<td></td>
</tr>
</tbody>
</table>
Deadlines

- Submission of project proposal
  - Sunday, April 19th, 23:59

- Submission of final project report
  - Friday, May 22nd, 23:59

- Project presentations
  - week of May 25th
  - everyone has to attend the presentations
Literature – Data Mining Methods

   - main reference book for the course!
   - 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

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

4. Website: **KDNuggets**
   - Overview of tools, online courses, events
   - http://www.kdnuggets.com/
Software

- Powerful data mining suite
- We are using Version 6.2 in the exercise
Literature – Rapidminer

   - 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.

3. **Rapidminer – User Manual**
   - Introduction to user interface and basic features
Questions?