Data Mining II
Organization

Heiko Paulheim, Robert Meusel
Hello

• 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
• 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
Course Organization

- Lecture
  - addresses advanced data mining topics
  - builds on Data Mining I lecture contents!

- Project Work
  - we will take part in the Data Mining Cup 2015
  - with two teams
  - regular presentations of your approaches
  - paper and final presentation

- Exercise
  - weekly with warm up on DMC tasks from previous years
Requirements

• Final exam
  – 50 % written exam
  – 50 % project work
• Project work
  – work on DMC tasks
• Presentations
  – three intermediate presentations
    • open questions, problems, current results (numbers!)
  – one final presentation
  – everybody has to present once during those four presentations
• Final report
  – 10 pages
  – solutions, results, lessons learned
The Data Mining Cup

• An annual competition
  – for students
  – run since 2002
  – participation from all over the world
  – max. two teams per institution (i.e., university)
  – 2014: 125 participating teams

• Timeline
  – tasks are published on April 2\textsuperscript{nd}
  – submissions are due on May 14\textsuperscript{th} (internal submission: May 9\textsuperscript{th})

• Further information: http://www.data-mining-cup.de/en
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Tuesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.02.15</td>
<td>Lecture: Preprocessing</td>
<td>Exercise: Preprocessing</td>
</tr>
<tr>
<td>16.02.15</td>
<td>Lecture: Regression</td>
<td>Exercise: Regression</td>
</tr>
<tr>
<td>23.02.15</td>
<td>Lecture: Anomaly Detection</td>
<td>Exercise: Anomaly Detection</td>
</tr>
<tr>
<td>02.03.15</td>
<td>Lecture: Ensembles</td>
<td>Exercise: Ensembles</td>
</tr>
<tr>
<td>09.03.15</td>
<td>Lecture: Time Series</td>
<td>Exercise: Time Series</td>
</tr>
<tr>
<td>23.03.15</td>
<td>Lecture: Online Learning</td>
<td>Exercise: Online Learning</td>
</tr>
<tr>
<td>30.03.15</td>
<td></td>
<td><strong>Easter Break</strong></td>
</tr>
<tr>
<td>06.04.15</td>
<td></td>
<td><strong>Easter Break</strong></td>
</tr>
<tr>
<td>13.04.15</td>
<td>Lecture: Parameter Tuning</td>
<td>Exercise: Parameter Tuning</td>
</tr>
<tr>
<td>20.04.15</td>
<td>Task discussion, team building</td>
<td>Work on DMC tasks</td>
</tr>
<tr>
<td>27.04.15</td>
<td>Work on DMC tasks</td>
<td></td>
</tr>
<tr>
<td>04.05.15</td>
<td>Intermediate result presentation</td>
<td>Work on DMC tasks</td>
</tr>
<tr>
<td>11.05.15</td>
<td>Work on DMC tasks</td>
<td></td>
</tr>
<tr>
<td>18.05.15</td>
<td>Work on Final Submission and Presentation</td>
<td></td>
</tr>
<tr>
<td>25.05.15</td>
<td>-</td>
<td><strong>Final presentation</strong></td>
</tr>
</tbody>
</table>
Deadlines at a Glance

• April 7\textsuperscript{th}: you know the DMC tasks and your team
  – Note: this is during Easter break!
• May 14\textsuperscript{th}: submission of your DMC solution to Robert and Heiko
• May 19\textsuperscript{th}: official submission of your DMC solution
• May 25\textsuperscript{th}: submission of your final report
• May 26\textsuperscript{th}: final presentations
RapidMiner Analyst Certification

• Offered for the first time this semester
• Online exam run by RapidMiner
  – voluntary part of this lecture
  – does not replace the DM2 exam
  – likely date: May 25th, during lecture slot
  – free of charge
Lecture Contents

• Data Preprocessing (today!)
• Regression
• Anomaly Detection
• Ensemble Learning
• Time Series Analysis
• Online Learning
• Parameter Tuning
Course Organization

- **Lecture Webpage: Slides, Announcements**
  - hint: look at version tags!

- **Additional Material**
  - ILIAS eLearning System, [https://ilias.uni-mannheim.de/](https://ilias.uni-mannheim.de/)

- **Time and Location**
  - Monday, 10.15 – 11.45
    Room A104 in B6
  - Tuesday, 13.45 – 15.15
    Room A104 in B6

- **Bring a laptop each time!**
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

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

• Jiawei Han and Micheline Kamber: Data Mining – Concepts and Techniques – free e-book access via university library
Literature & Slide Sources

• Albert Bifet: Adaptive Stream Mining
  – we will provide scans of important chapters in ILIAS

• Joao Gama: Knowledge Discovery from Data Streams
  – we will provide scans of important chapters in ILIAS
Software

- Powerful open-source data mining suite
- Download: http://rapid-i.com/
- We use the most recent version of RapidMiner Studio
  - You will obtain a license key
Questions?