

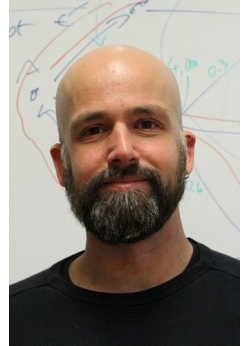
# Data Mining II Organization



**Heiko Paulheim, Nicolas Heist**

# Hello

- Heiko Paulheim
- Professor for Data Science
- Research Interests:
  - Semantic Web and Linked Open Data
  - Data Mining with Linked Open Data
  - Ontology Matching
  - Data Quality and Data Cleaning
- Consultation: Tuesdays, 9-10am
  - Please make an appointment via e-mail to Ms. Lerner
- Heiko will teach the lectures



# Hello

- M.Sc. Nicolas Heist
- Graduate Research Associate
- Research Interests:
  - Semantic Web Technologies
  - Knowledge Graphs and Linked Data
- eMail: [nico@informatik.uni-mannheim.de](mailto:nico@informatik.uni-mannheim.de)
- Nico will teach the exercises and co-supervise the 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 2019
  - with four teams
    - the two best performing teams submit their solutions
  - regular presentations of your approaches
  - paper and final presentation
- Exercise
  - weekly with warm up on DMC tasks from previous years

# Requirements

- Final exam
  - 100 % written exam
  - project is not graded, but mandatory!
- Project work
  - work on DMC tasks
- Presentations
  - up to three intermediate presentations
    - open questions, problems, current results (numbers!)
  - everybody has to present once during those presentations
- Final report
  - 10 pages
  - solutions, results, lessons learned

different to last years!

# 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)
  - 2018: 197 participating teams from 47 countries
- Timeline
  - DMC registration on March 5<sup>th</sup>
  - tasks are published on April 4<sup>th</sup>
  - submissions are due on May 16<sup>th</sup> (internal submission: May 13<sup>th</sup>)
- Further information: <http://www.data-mining-cup.de/en>

# The Data Mining Cup

- 2017: both Uni Mannheim teams among top 10 (out of 202)
- 2018: team from Uni Mannheim scores 2<sup>nd</sup> place (out of 197)
- Prices are awarded at a conference in Berlin in June
  - Top 10 teams are invited to present their solutions





# Schedule

- 19.02.18 Lecture: Preprocessing
- 26.02.18 Lecture: Regression
- 05.03.18 Lecture: Anomaly Detection
- 12.03.18 Lecture: Ensembles
- 19.03.18 Lecture: Time Series
- 26.03.18 Lecture: Neural Networks
- 02.04.18 Lecture: Parameter Tuning
- 09.04.18 DMC intermediate presentation
  - Easter Break
- 29.04.18 DMC intermediate presentation
- 06.05.18 DMC intermediate presentation

DMC task  
published  
on 04.04.

final DMC  
submission  
16.05.



# Deadlines at a Glance

- March 5<sup>th</sup>: DMC registration
- April 4<sup>th</sup>: you know the DMC tasks and your team
- May 13<sup>th</sup>: submission of your DMC solution to Nico and Heiko
- May 16<sup>th</sup>: official submission of your DMC solution
- May 20<sup>th</sup>: submission of your final report



# Lecture Contents

- Data Preprocessing (today!)
- Regression
- Anomaly Detection
- Ensemble Learning
- Time Series Analysis
- Neural Networks and Deep Learning
- Parameter Tuning


# Course Organization

- Lecture Webpage: Slides, Announcements
  - <http://dws.informatik.uni-mannheim.de/en/teaching/courses-for-master-candidates/ie-672-data-mining-2/>
  - hint: look at version tags!
- Additional Material
  - ILIAS eLearning System, <https://ilias.uni-mannheim.de/>

The screenshot shows the 'Data Mining II' course page on the University of Mannheim's Data and Web Science Group website. The page has a teal header with the university logo and group name. A navigation bar includes links for Home, People, News, Focus Areas, Teaching, Projects, Resources, Thesis, Career, and Contact. The main content area is titled 'Data Mining II' and describes the course as a deepening of data mining theory and practice. It lists topics such as Data Preprocessing, Regression and Forecasting, Dimensionality Reduction, Anomaly Detection, Time Series Analysis, Parameter Tuning, Ensemble Methods, and Deep Learning. The page also mentions the 'Data Mining Cup (DMC)' competition and provides details about the lecture and exercises, including dates and locations. A sidebar on the left lists various courses and seminars offered by the group.

# Video Recordings of Last Year's Lecture

- <http://dws.informatik.uni-mannheim.de/en/teaching/lecture-videos/>
  - Accessible from within university network and VPN



Data Mining II  
**Anomaly Detection**

Prof. Dr. Heiko Paulheim  
Data and Web Science Group

UNIVERSITY OF  
MANNHEIM

## Interquartile Range

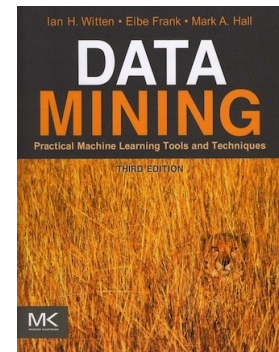
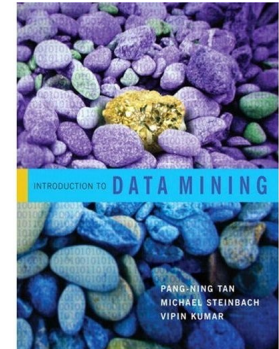
- Divides data in quartiles
- Definitions:
  - $Q1$ :  $x \geq Q1$  holds for 75% of all  $x$
  - $Q3$ :  $x \geq Q3$  holds for 25% of all  $x$
  - $IQR = Q3 - Q1$
- Outlier detection:
  - All values outside  $[\text{median} - 1.5 \cdot IQR ; \text{median} + 1.5 \cdot IQR]$
- Example:
  - $0, 1, 1, 3, 3, 5, 7, 42 \rightarrow \text{median}=3, Q1=1, Q3=7 \rightarrow IQR = 6$
  - Allowed interval:  $[3 - 1.5 \cdot 6 ; 3 + 1.5 \cdot 6] = [-6 ; 12]$
  - Thus, 42 is an outlier

Heiko Paulheim, Robert Meusel

15

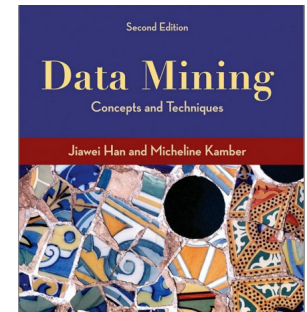
# Literature & Slide Sources

- 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
- Ian H. Witten, Eibe Frank, Mark A. Hall:  
Data Mining: Practical Machine Learning  
Tools and Techniques, 3rd Edition, Morgan Kaufmann.
  - 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/](http://www.kdnuggets.com/data_mining_course/)
- Jiawei Han and Micheline Kamber:  
Data Mining – Concepts and Techniques
  - free e-book access via university library



# Questions?

