

# Data Mining II Organization



Heiko Paulheim, Nicolas Heist

# Hello

- Heiko Paulheim
- Professor for Data Science
- Research Interests:
  - Semantic Web and Knowledge Graphs
  - Data Mining and Machine Learning with Knowledge Graphs
  - Ontology Matching
  - Data Quality and Data Cleaning
- Consultation: Tuesdays, 9-10am
  - Please make an appointment via e-mail to Ms. Lermer
- 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 2021
  - with eight 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

# 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)
  - 2020: 162 participating teams from 35 countries
- Timeline
  - DMC registration already running (!)
  - tasks are published on April 13th
  - submissions are due on June 29<sup>th</sup> (internal submission: June 18<sup>th</sup>)
- Further information: <http://www.data-mining-cup.de/>

# 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)
- 2019: team from Uni Mannheim scores 10<sup>th</sup> place (out of 149)
- 2020: team from Uni Mannheim scores 8<sup>th</sup> place (out of 162)
- Prices are awarded at a virtual conference in July





# Schedule

- 9.3. Introduction & Data Preprocessing
- 16.3.Ensembles
- 23.3.Time Series
- *Easter Break*
- 13.4.Neural Networks & Deep Learning
- 20.4.DMC Session 1
- 27.4.Hyperparameter Tuning
- 4.5. DMC Session 2
- 11.5.Anomaly Detection
- 18.5.DMC Session 3
- 25.5.Model Verification
- 1.6. DMC Session 4
- 8.6. DMC Session 5
- 15.6.DMC Session 6

DMC task  
published  
13.4.

final DMC  
submission  
29.6.



# Deadlines at a Glance

- next Monday: DMC team registration
- April 13<sup>th</sup>: you know the DMC tasks and your team
- June 18<sup>th</sup>: submission of your DMC solution and report
- June 29<sup>th</sup>: official submission of your DMC solution



# Lecture Contents

- Data Preprocessing (today!)
- Ensemble Learning
- Time Series Analysis
- Neural Networks and Deep Learning
- Parameter Tuning
- Anomaly Detection
- Model Evaluation, Verification, and Comparison


# 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 website for the Data Mining II course. The header includes the University of Mannheim logo and the group name 'DATA AND WEB SCIENCE GROUP'. The navigation menu lists '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 building on fundamentals. 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 course structure includes a lecture, two alternative exercise dates, and a final exam consisting of a written exam and project work. Slides and exercises will be posted on the website.

# Video Recordings of an earlier 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

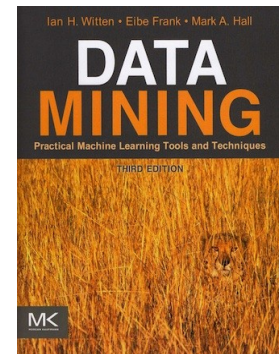
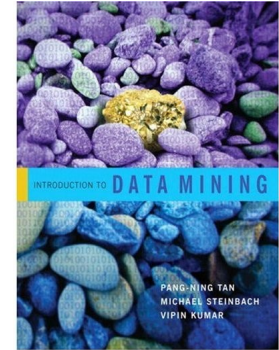
### Interquartile Range

- 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

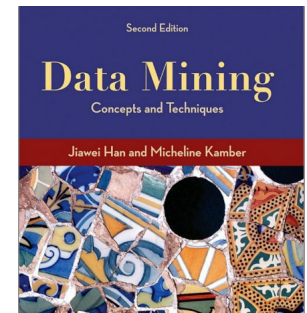
# Literature & Slide Sources

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar:  
Introduction to Data Mining,  
Pearson / Addison Wesley.
  - 10 copies in university library
  
- 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



# 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?

