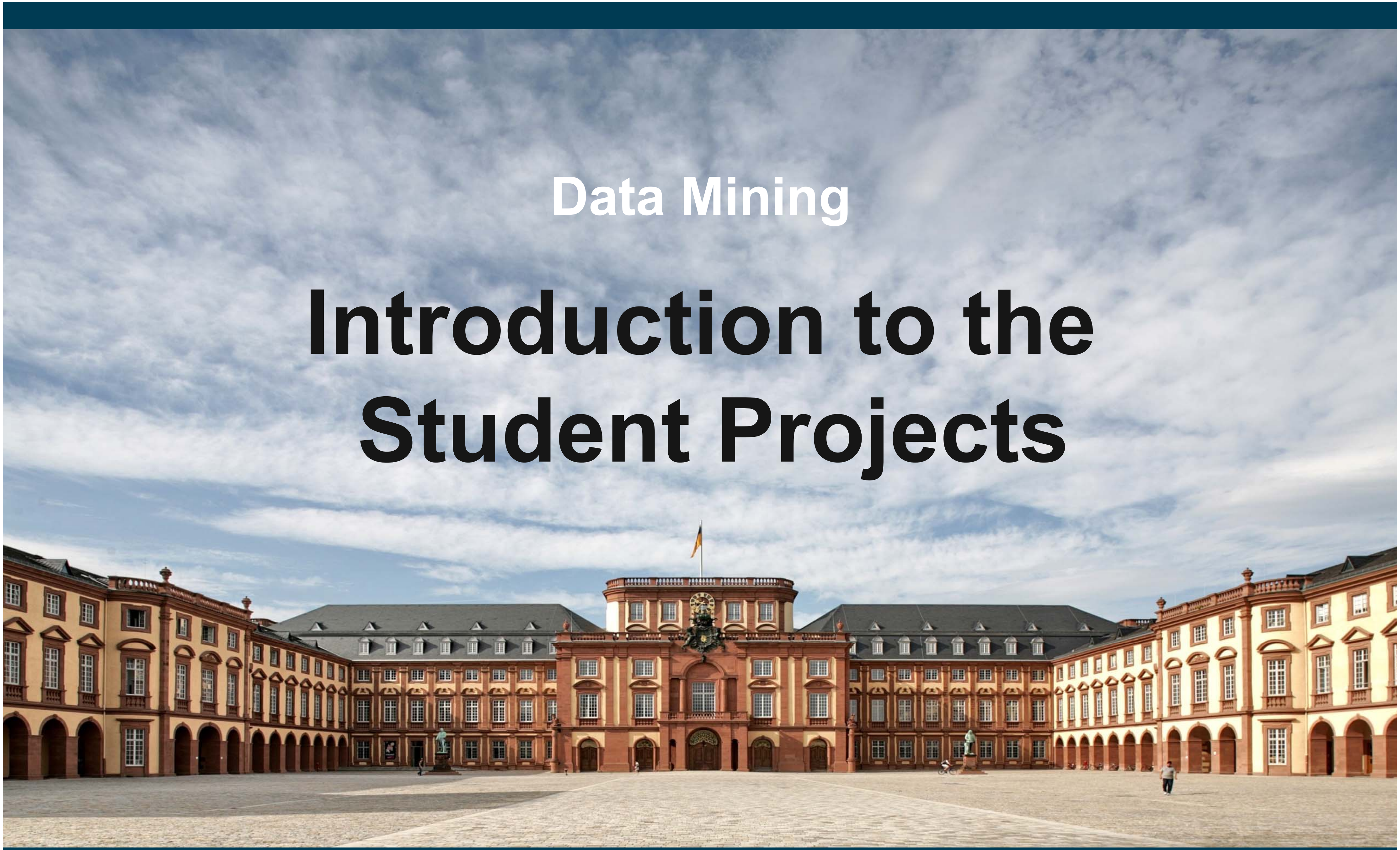


Data Mining

Introduction to the Student Projects

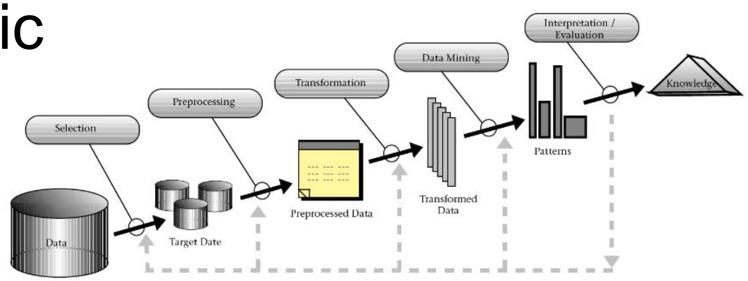
Outline

1. Requirements for Student Projects
2. Requirements for Project Reports
3. Final Exam

Student Projects

– Goals

- Gain practical experience with the complete data mining process
- Get to know additional problem-specific
  - preprocessing methods
  - data mining methods



– Expectation

- Select an interesting data mining problem of <u>your choice</u>
- Solve the problem using
  - the data mining methods that we have learned so far, including
    - proper parameter optimization
    - problem-specific pre-processing and smart feature creation
  - additional data mining methods which might be helpful for solving the problem and build on what we learned in class

Procedure

- Teams of six students
  1. realize a data mining project
  2. write a 10 page summary of the project and the methods employed in the project
  3. present the project results to the other students (10 minutes presentation + 5 minutes discussion)

- Final mark for the course
  - 30 % written summary about the project
  - 10 % project presentation
  - 60 % written exam

Schedule

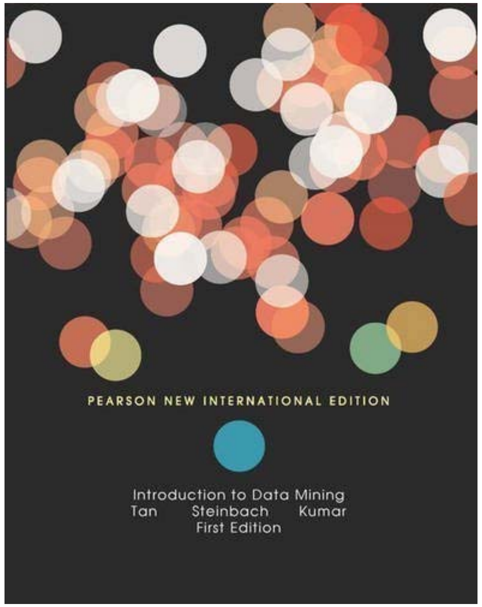
Week	Wednesday	Thursday
03.04.2019	Introduction to Student Projects	Preparation of Project Outline
Sunday, April 7th 2019, 23:59: Submission of Project Outlines		
10.04.2019	Lecture Association Analysis **Feedback Student Projects (13:45-15:15)**	Exercise Association Analysis
06.05.2019	Project Work	Feedback on demand
13.05.2019	Project Work	Feedback on demand
20.05.2019	Project Work	Feedback on demand
Sunday, May 26th 2019, 23:59: Submission of Project Reports		
29.05.2019	Presentation of Project Results (10:15 to 15:15)	
03.06.2018	**Exam**	

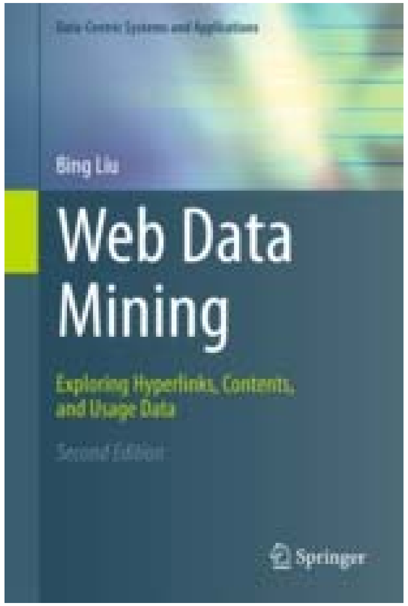
Where to find interesting Data Sets?

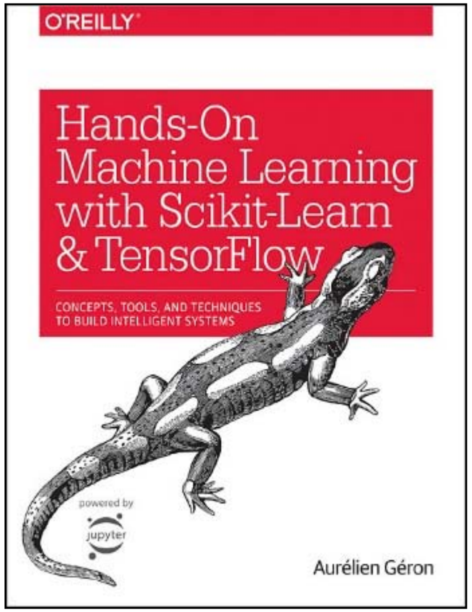
- KDnuggets Dataset List
  - https://www.kdnuggets.com/datasets/index.html
  - References to various data catalogs and datasets
- Data.gov, data.gov.uk, govdata.de
  - Public sector data provided by the government bodies
- Programmable Web
  - Website giving an overview about 13000 public Web APIs
- KDD Cup and Data Mining Cup
  - Data mining competitions providing data sets and solutions
  - http://www.kdd.org/kdd-cup
  - https://www.data-mining-cup.com
- Kaggle
  - Website running commercial and educational data science competitions
  - Offers datasets as well as solutions for older competitions
  - https://www.kaggle.com/
  - Please compare your results to results from the competition’s forum

Where to Find Information about Additional Methods?

1. Pang-Ning Tan, Michael Steinback, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
2. Bing Liu: Web Data Mining, 2nd Edition, Springer.
3. Aurélien Géron: Hands-on Machine Learning with Scikit-Learn. O'Reilly.


PEARSON NEW INTERNATIONAL EDITION
Introduction to Data Mining
Tan Steinbach Kumar
First Edition


Data-Centric Systems and Applications
Bing Liu
Web Data Mining
Exploring Hyperlinks, Contents, and Usage Data
Second Edition
Springer


O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
powered by jupyter
Aurélien Géron

Where to Find Information about Additional Methods?

– Check out the solutions to your problem that other people have tried.
  - for instance by looking at submissions of the KDD Cup or Data Mining Cup as well as Kaggle discussion groups
  - or search for relevant scientific papers using

Google™

Some Project Ideas (not binding)

- Web Log Mining
  - Learn a classifier for the categorizing the visitors of your website.
  - Which features matter? Number of pages visited, time on site, ..
    (Bing Liu Chapter 12.x)
  - Preprocess some web log data outside RapidMiner
  - Learn and evaluate classifier within RapidMiner
- Wikipedia Contributors / Hoax Articles
  - Examine the edit history of Wikipedia contributors
  - Cluster users by different attributes (no of edits, edits/day, topic, ...)
  - Or learn a classifier for the categorizing Wikipedia contributors
- Sentiment Analysis for Discussion Forum / Rating Site / Tweets
  - Are people positive or negative about topic / product? (Bing Liu 11.x)
- Estimate House or Car Prices
  - using different regression methods or transfer learning to localize method

Some Projects realized in previous Semesters

- Mannheim Police Reports
  - Learn classifiers for police reports
  - Identify type of incident, severity of incident, location of incident
- Bundesliga Betting Rules
  - Find rules that help you to predict the outcome of a Bundesliga game
- last.fm Playlist Analysis
  - Cluster last.fm users according to the style of the songs they are listening to
  - Find commons sets of songs for the different clusters
- Analysis of Training Data of a Fitness Center
  - Find different customer groups by clustering exercise data
  - Find frequent combinations of exercises
- Sentiment Analysis of Tweets about Movies
  - Learned classifier from IMDB movie reviews
  - Applied and tested with tweets afterwards
- Classifying a Document‘s Perspective
  - using the example of Israeli – Palestinian Essays

Project Outlines

- maximum 4 pages using Springer Computer Science Proceedings layout
  - Include a project name and your team number on the first page!
- due Sunday, April 7th 2019, 23:59 (in 4 days!)
- send by eMail to Chris, Anna, Oliver
- answer the following questions:

1. What is the problem you are solving?
2. What data will you use?
   - Where will you get it?
   - How will you gather it?
3. How will you solve the problem?
   1. What preprocessing steps will be required?
   2. Which algorithms do you plan to use?
   - Be as specific as you can!
4. How will you measure success? (Evaluation method)
5. What do you expect your results to look like? (Model/Clusters/Patterns)

- Feedback about your project outlines: Wednesday, 10.04.2019, 13:45-15:15

Coaching Sessions

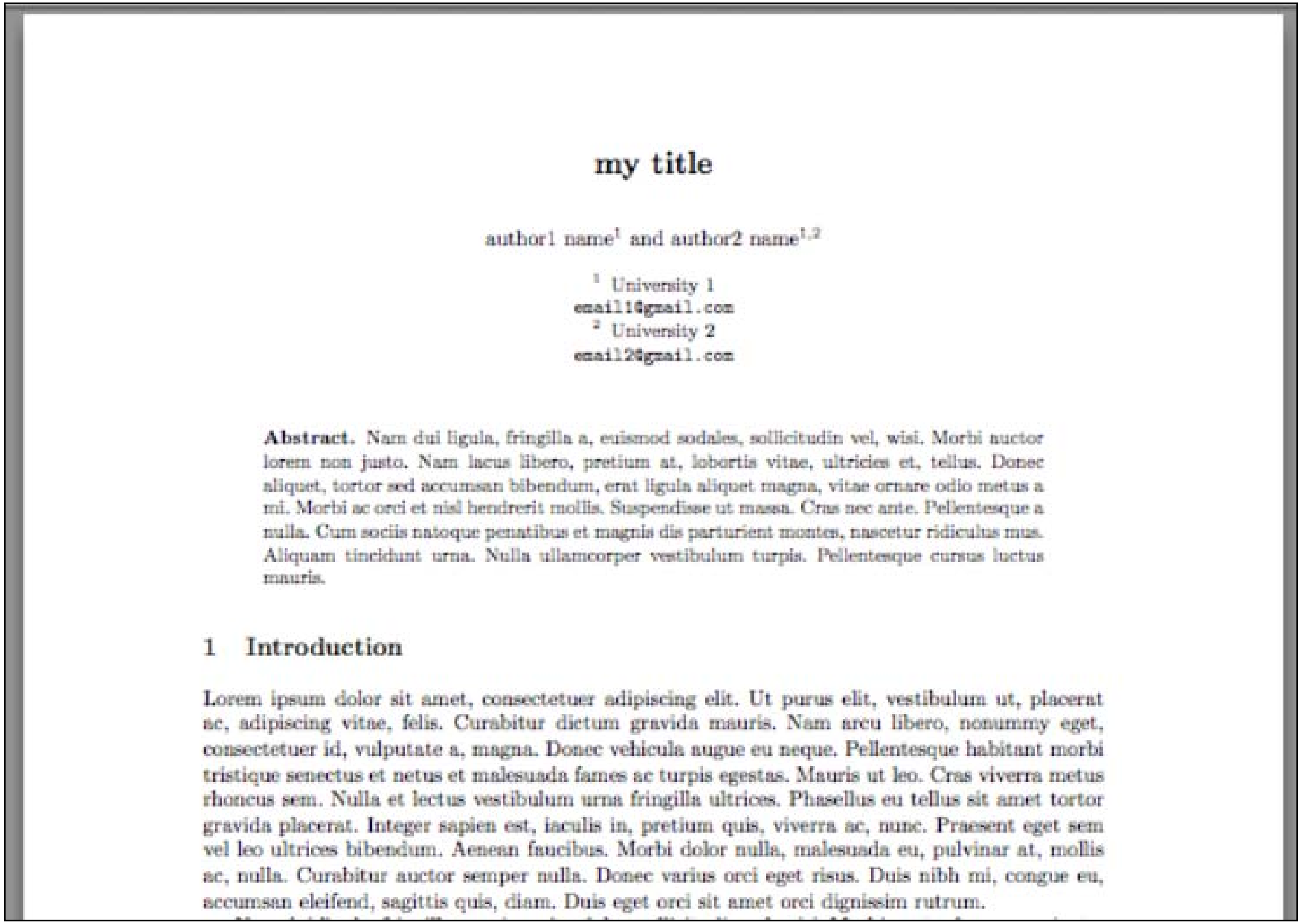
- We will give you tips and answer questions concerning your project.
- Registration via email to Oliver & Anna is mandatory!
  - until Tuesday night!
  - including the questions that you like to discuss
  - including which session you prefer (Thursday B2/B3)
- We will assign you a time slot afterwards and inform you about the slot via email.

- Every team has to attend at least one coaching session!

Project Report

- 10 pages (exactly!) plus references page, no appendix ➔ document length: 11 pages
- Each extra page and each day of late submission downgrades your mark by 0.3!
- due Sunday, May 26th 2019, 23:59
- send by email to Chris, Anna & Oliver
- Outline for project report:
  1. Application area and goals
  2. Structure and size of the data set (minimum 1 page)
  3. Preprocessing and Mining
     - describe different approaches and parameter settings that you tried
     - including evaluation setup and evaluation results
     - including discussion of the results
- Requirements
  1. You must use the latex template of the Springer Computer Science Proceedings
  2. Please cite sources properly and use your references page
  3. Also submit your RapidMiner processes and (a subset) of your data
  4. Include your names and your team number on the first page!

Template: Springer Computer Science Proceedings



my title

author1 name[1] and author2 name[1,2]

[1] University 1
email1@gmail.com
[2] University 2
email2@gmail.com

Abstract. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

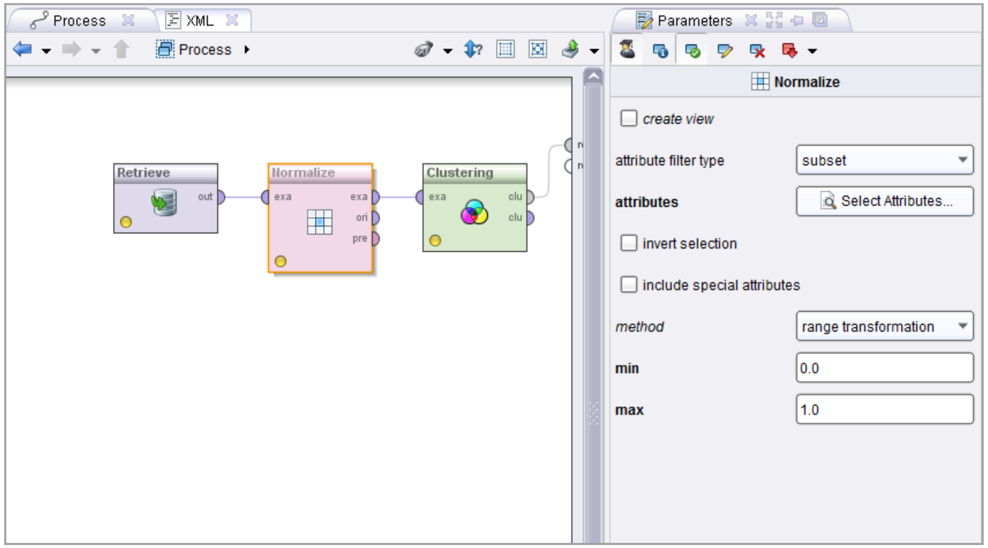
1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines

Severe Errors to Avoid

1. Normalize numeric data before calculating any similarity metrics



Normalize

□ create view

attribute filter type: subset

attributes: Select Attributes...

□ invert selection

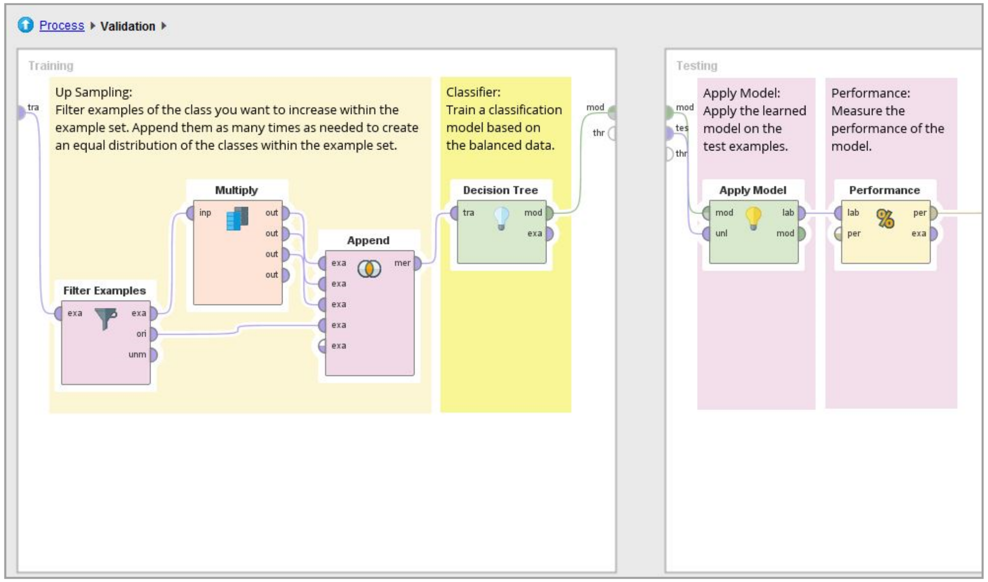
□ include special attributes

method: range transformation

min: 0.0

max: 1.0

2. If your data is unbalanced
   - balance your training data
   - do NOT balance your test data
   - report P/R/F1, not accuracy



Up Sampling:
Filter examples of the class you want to increase within the example set. Append them as many times as needed to create an equal distribution of the classes within the example set.

Classifier:
Train a classification model based on the balanced data.

Apply Model:
Apply the learned model on the test examples.

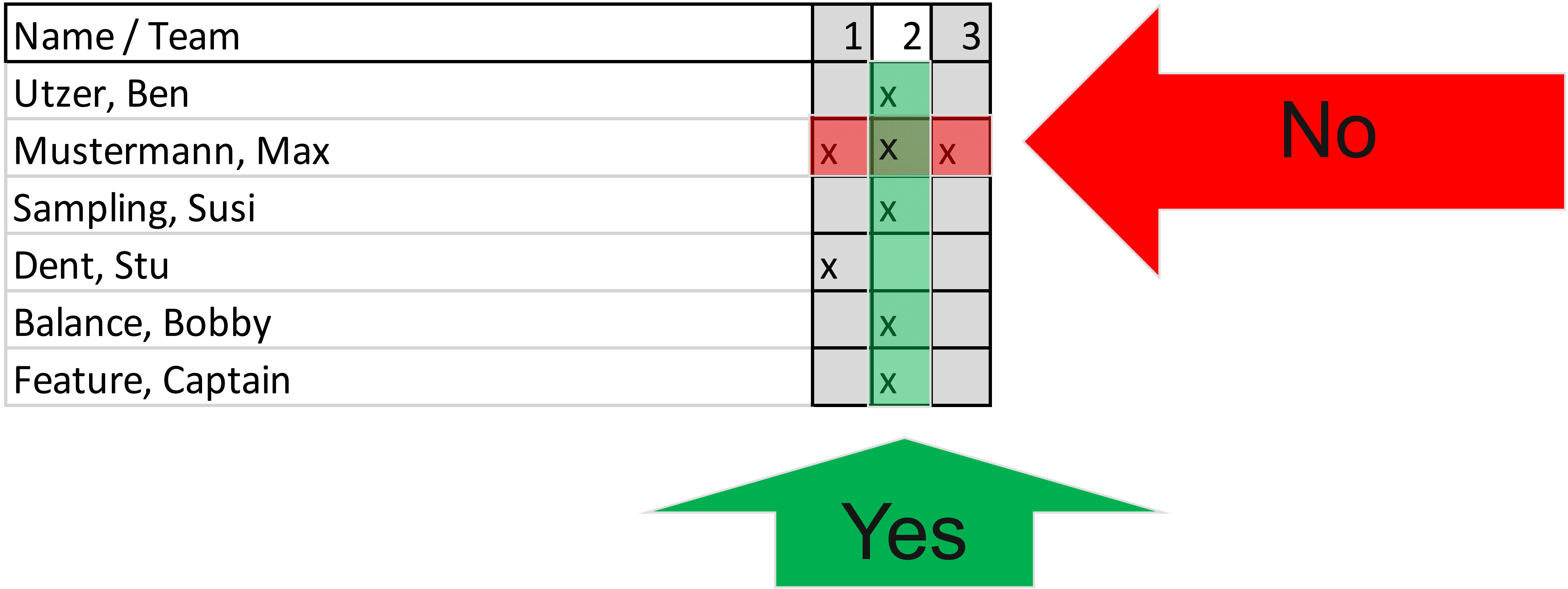
Performance:
Measure the performance of the model.

Final Exam

- Date: Monday, June 3rd
- Duration: 60 minutes
- Structure: 6 open questions that
  - check whether you have understood the content of the lecture
  - require you to describe the ideas behind algorithms and methods
  - might require you to do some simple calculations

Team Assignment

- Find your team now!
- Then enter your team in the student/team matrix!
  - Only enter if you have a team (don’t make random crosses!)
  - There can only be **one cross per row** (you can’t be in two teams!)
  - There should be **six crosses per column** (six students per team!)



Name / Team	1	2	3
Utzer, Ben		x	
Mustermann, Max	x	x	x
Sampling, Susi		x	
Dent, Stu	x		
Balance, Bobby		x	
Feature, Captain		x	