CHAIR FOR BIOINFORMATICS
AND INFORMATION MINING

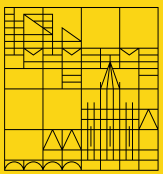
Advanced Data Challenge

Data Mining for Fun and Profit

Christian Borgelt und Christoph Doell

Chair for Bioinformatics and Information Mining
Department of Computer and Information Science
Konstanz University, Germany
First.Last@uni.kn

Universität
Konstanz



Get your hands dirty

- Data Science is as much a craft as it is a science.
- Learning by doing.
- Each Data Science project is unique with different problems and hurdles.
- Real world data sets.
- Money! (well...)

https://www.flickr.com/photos/reallynuts/4385681932 License: CC 2.0 Attribution

Organization/Setup

- Some data set (some suggestions to follow).
- Some data analysis environment, for example,
  KNIME, Weka, MatLab, R, Python + ScikitLearn etc.
- Teams of 2 persons (exactly 2! not 1, not 3, but 2!)
- Weekly meetings to exchange experiences,
  discuss problems, get help... ("Data Analysts Anonymous")
- Mid-term presentation (10%)
  Date: in 6 weeks ($\sim$ 29.05.2018/03.06.2018) Not fixed!
- Final presentation (40%)
  Date: end of semester ($\sim$ 18.07.2018) Not fixed!
- 6-8 page experimental report (50%)
  Date: 31.08.2018

Registration / Presentations

Registration

- ZeUS `https://bit.ly/2HDvAyV`
- StudIS
- ILIAS `https://bit.ly/2Hpmngb`

Presentations

- Mid-term presentation $\sim$ 29.05.2018/03.06.2018
  - 15 mins + 5 mins Discussion
- Final presentation (end of semester)
  - 25 mins + 5 mins Discussion

Experimental Report Submission–the Rough Guide

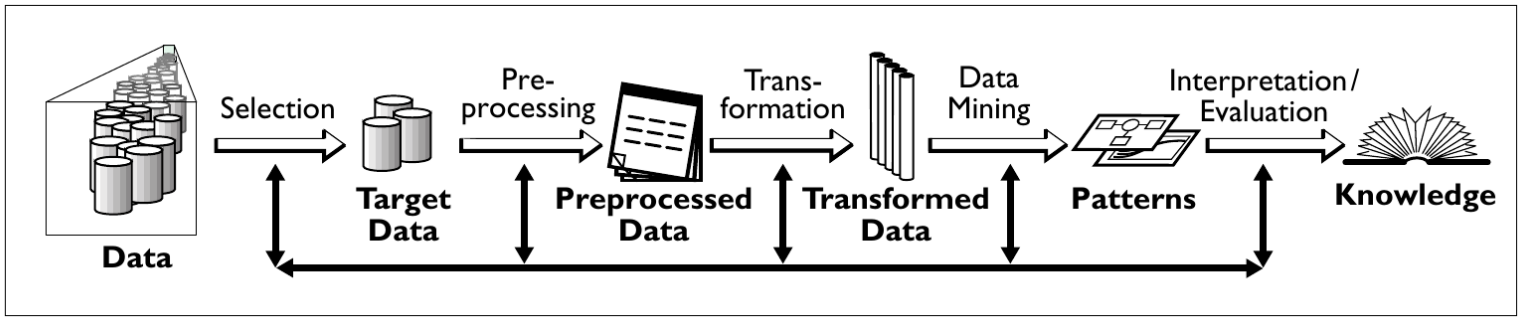
Goal: Write an experimental report so that the experiment can be reproduced.

Structure:

1. Abstract
2. Data description
3. Project description
   - Related Work
   - Models that performed
   - Models that didn't perform
4. Results
5. Conclusion

How to start?

Follow the steps of the KDD Process.
Each of them needs to be included in the final presentation.



Fayyad, Piatetsky-Shapiro, Smyth 1996:
"The KDD Process for Extracting Useful Knowledge from Volumes of Data"
Communications of the ACM

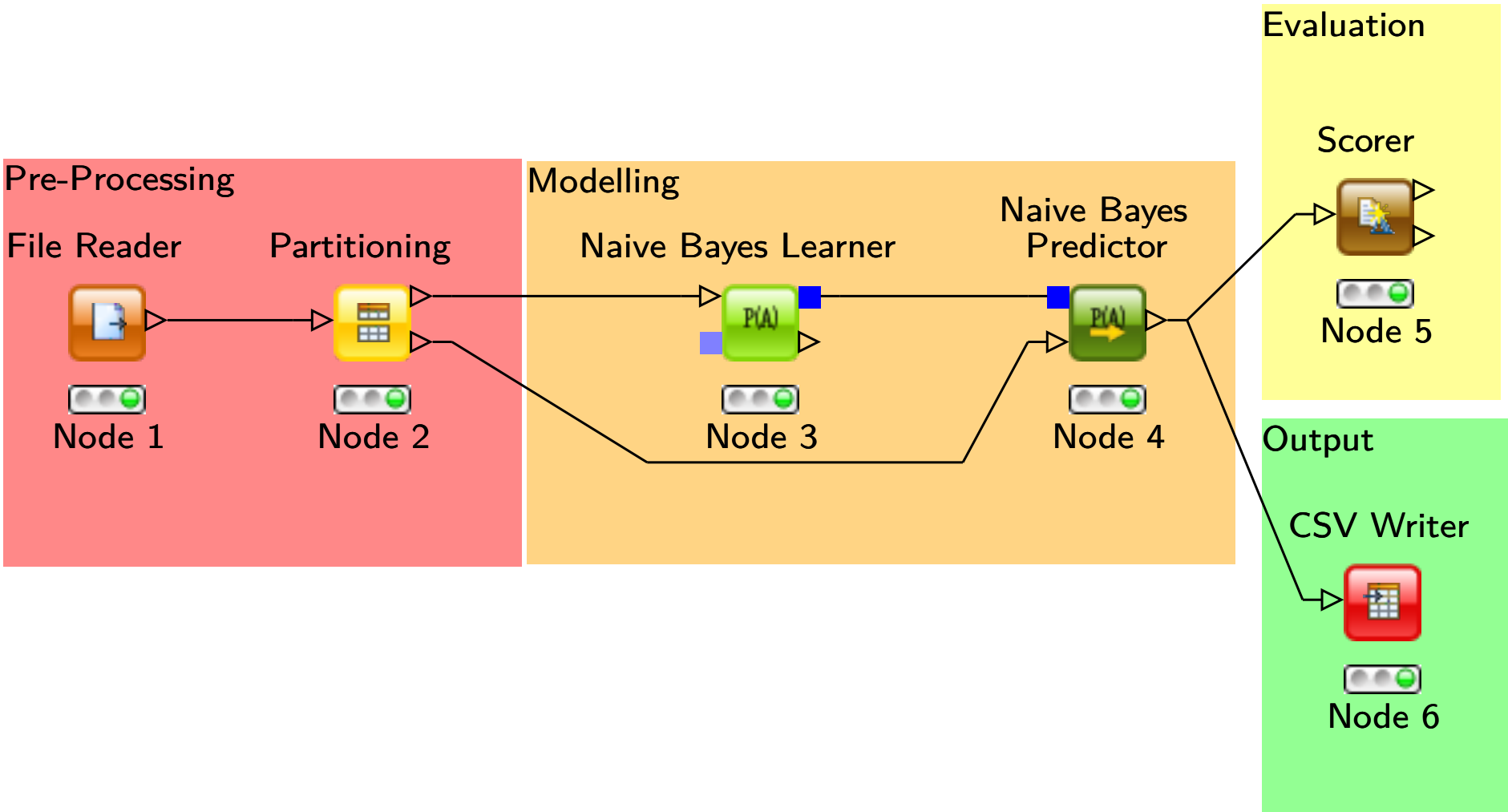
KNIME: A Data Analysis Platform

`http://www.knime.org`

- "leading open platform for data-driven innovation"
- "It's yours: ...customiz[able]...with a broad range of free or commercial applications. Or create and share your own."
- "Shortcut your learning curve:
  KNIME helps build on what you've already developed and learned, without having to start from scratch."
- "More thinking, less tinkering:
  ...spend more quality time with your data."
- "Low cost, zero risk: ...minimal investment."

`http://www.knime.org/knime`

KNIME: How does it work?

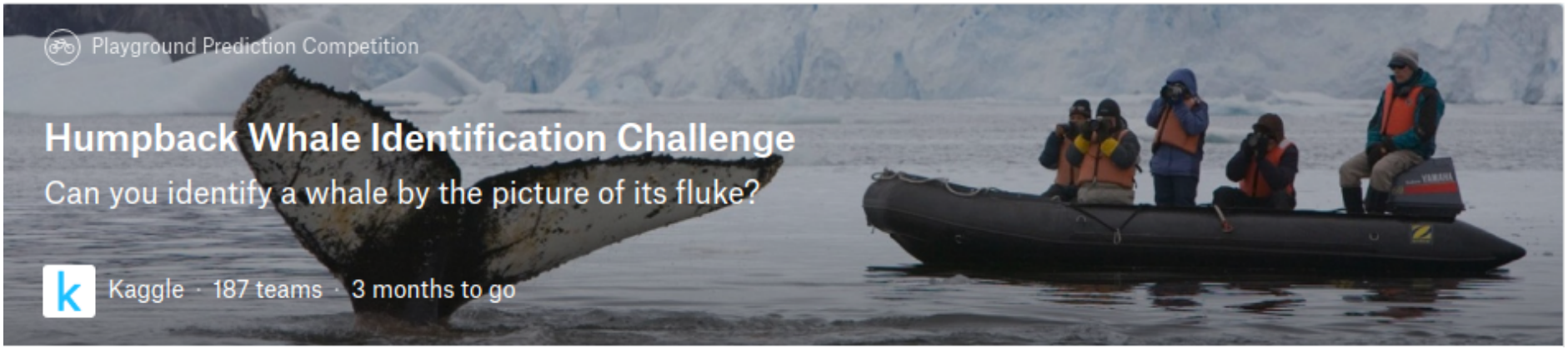


Predict Future Sales

1. You are provided with daily historical sales data.
2. The task is to forecast the total amount of products sold in every shop for the test set.
3. Creating a robust model that can handle such situations is part of the challenge.

Details: `https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data`

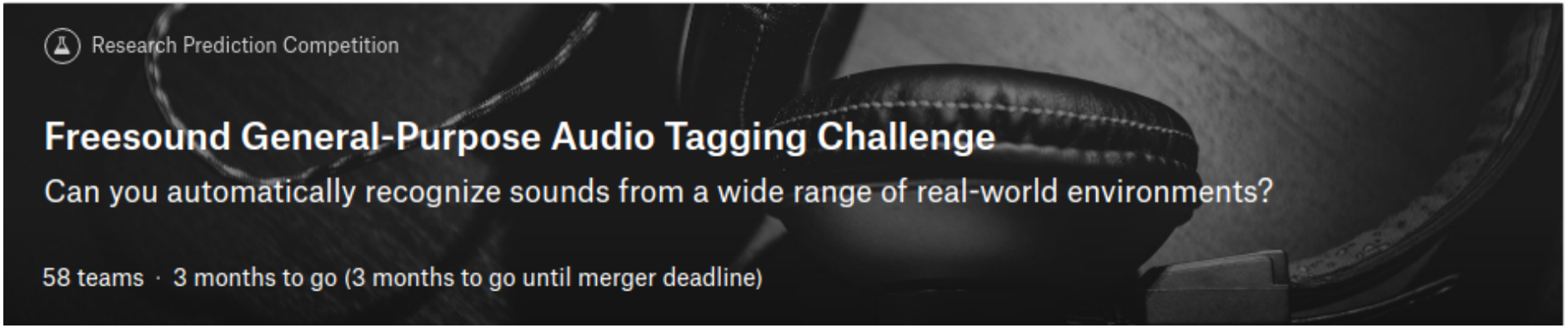
Humpback Whale Identification



1. To aid whale conservation efforts, scientists use photo surveillance systems to monitor ocean activity.
2. They use the shape of whales' tails and unique markings to identify the species of whale.
3. You are challenged to build an algorithm to identify whale species in images.

Details: `https://www.kaggle.com/c/whale-categorization-playground`

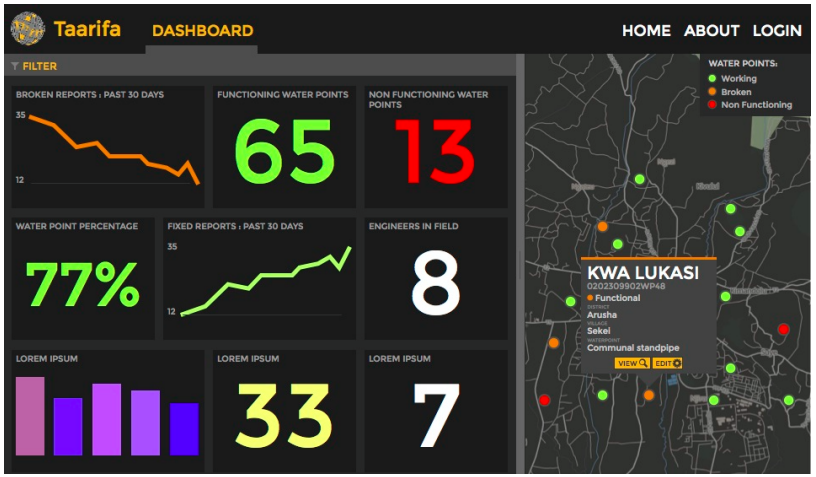
Freesound Audio Tagging



1. Some sounds are distinct and instantly recognizable, other sounds are not clear and are difficult to pinpoint.
2. Currently, a lot of manual effort is required for tasks like annotating sound collections.
3. You are challenged to build a general-purpose automatic audio tagging system.

Details: `https://www.kaggle.com/c/freesound-audio-tagging`

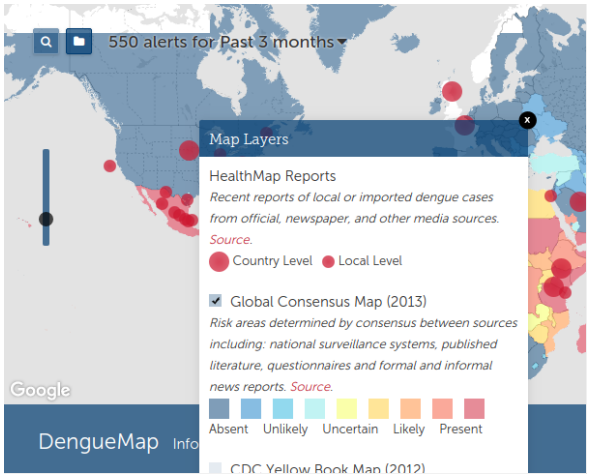
Pump it Up: Data Mining the Water Table



1. Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which don't work at all?
2. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Details: `https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/`

DengAI: Predicting Disease Spread



1. Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world.
2. The transmission dynamics of dengue are related to climate variables such as temperature and precipitation.
3. Can you predict local epidemics of dengue fever? (for two cities, San Juan and Iquitos)

Details: `https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/`

United Nations Millennium Development Goals

1. In 2000, the member states of the United Nations agreed to a set of goals to measure the progress of global development.
2. The UN measures progress towards these goals using indicators such as percent of the population making over one dollar per day.
3. Your task is to predict the change in these indicators one year and five years into the future.

Details: `https://www.drivendata.org/competitions/1/united-nations-millennium-development-goals/`

Other Competitions/Challenges

- Data Analysis Competitions for Biomag 2018

  `http://www.biomag2018.org/2018/01/02/data-analysis-competition-1/`

  `http://www.biomag2018.org/2018/02/13/data-analysis-competition-2/`

- CAMDA Contest Challenges

  `http://camda2018.bioinf.jku.at/doku.php/contest_dataset`

- Teradata University Network Data Challenges

  `http://www.teradatauniversitynetwork.com/Community/Student-Competitions/2018/Data-Challenge/Datasets/`

- Find your own!

  If you have found some data analysis/data science competition on the internet, contect us to see whether it would be suitable for this course.