Data Mining: Presentation

Inês Dutra
ines@dcc.fc.up.pt
Office: 1.31
Office hours:
Mon, Thu: 2-4 pm

Evaluation

- 2 Assignments: total 8 points
- 2 Tests (6 points each):
  - Oct 29th
  - Dec 12th
- Final Exam: 12 points
- Best score between Test and Exam is considered
- Paper reading and discussion

Communication

- In person
- Email: **ines@dcc.fc.up.pt**
  (PLEASE, **DO NOT** SEND EMAIL TO **dutra@fc.up.pt**)
- Always use a subject prefix DM1 in your messages
- Sign your messages, so that I can identify you by more than a number ☺
- Other means:
  - Moodle (warnings, news, and forum)
- Discipline web page:

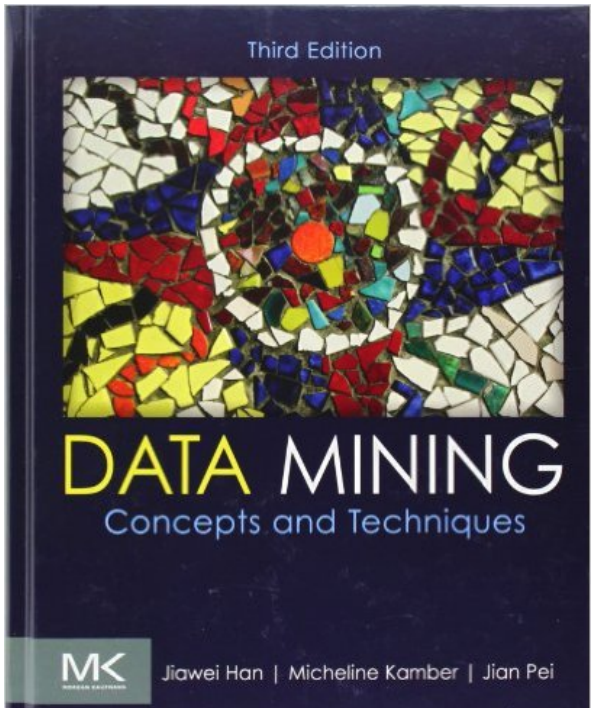
http://www.dcc.fc.up.pt/~ines/aulas/1819/DM1/DM1.html

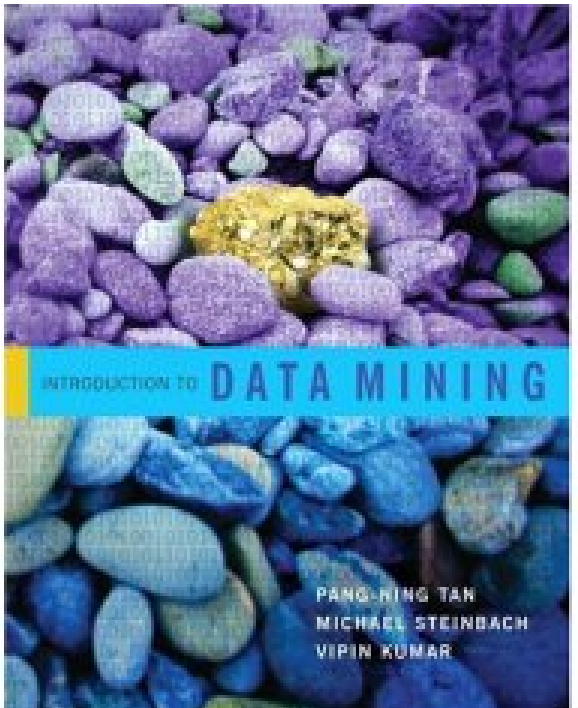
Syllabus

- What is data mining?
- Data versus knowledge
- Kinds of data
- Phases of data mining
- Data Preprocessing
- Descriptive Statistics
- Association rules
- Clustering
- Predictive Models
- Performance Metrics and model validation

Bibliography

- **Data Mining Concepts and Techniques (3rd ed)**
  Jiawei Han, Micheline Kamber and Jian Pei
- **Introduction to Data Mining**
  Pang-Ning Tan, Michael Steinbach and Vipin Kumar





Resources

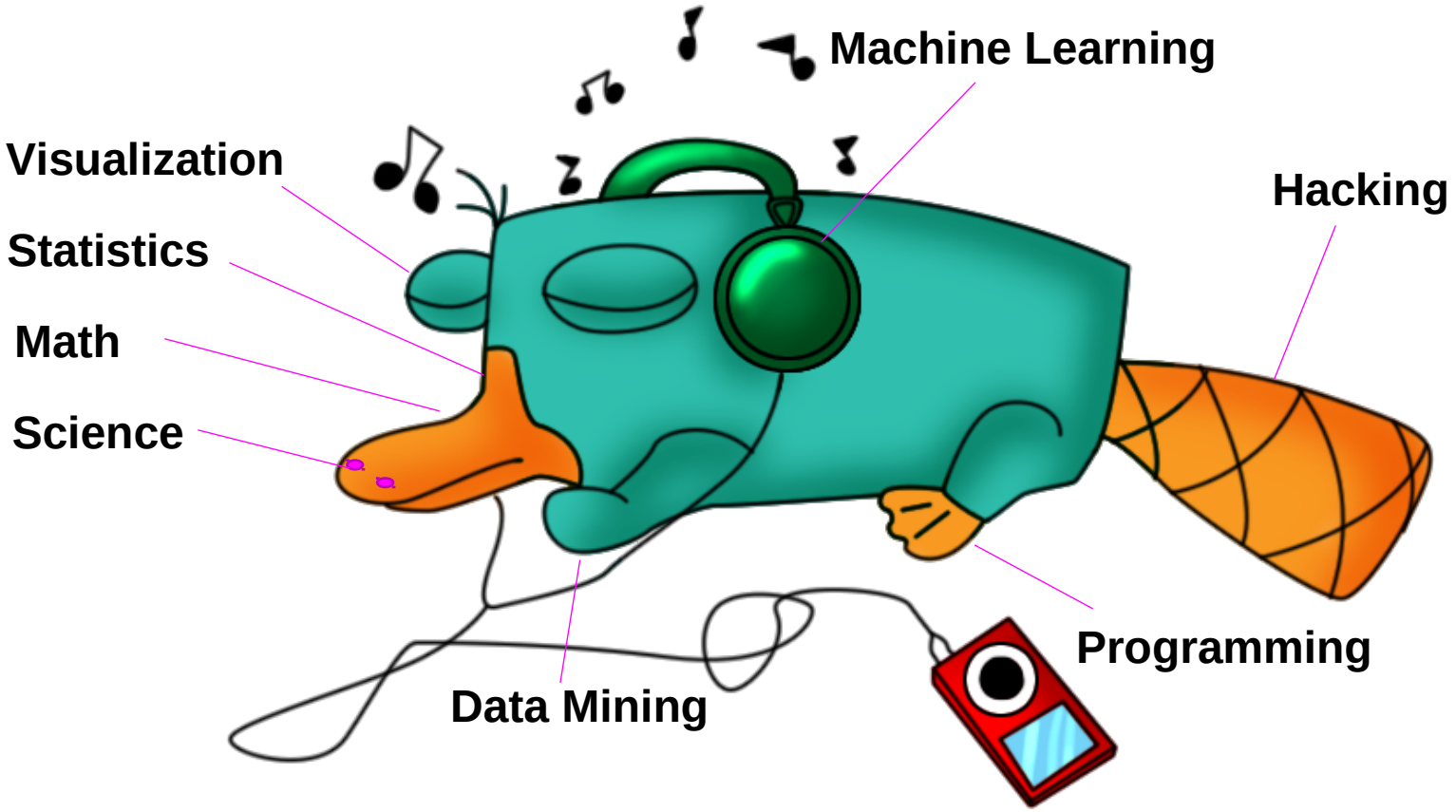
- For programming and libraries
  - R and stats and machine learning packages
  - PyML
- For data visualization and machine learning
  - WEKA
  - KNIME
  - RapidMiner
- For relational learning
  - Aleph and YAP
  - GILPS

Useful links

- KDD nuggets: http://www.kdnuggets.com
- Data Sets at UCI: http://archive.ics.uci.edu/ml/
- http://www.acm.org/sigs/sigkdd/explorations/
- https://www.kaggle.com/

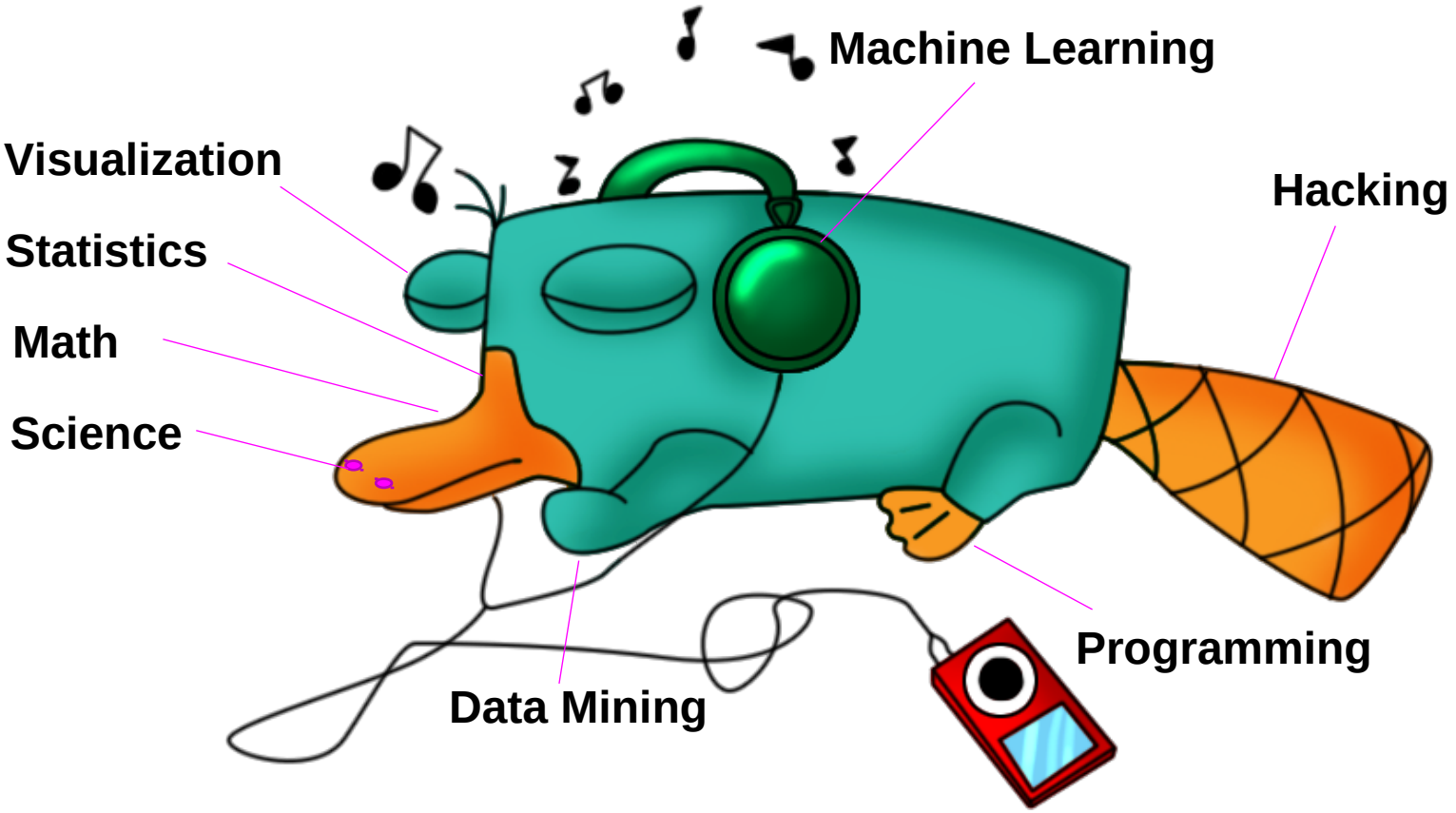
The Homo Platipus ☺

(excellent insight by Carlos Somohano, Founder of DataScience London)



The Homo Platipus ☺

(excellent insight by Carlos Somohano, Founder of DataScience London)



More commonly called: Data Scientist!

Course Requirements

- Motivation
- Willingness to learn
- Lots of patience
  - Interact with other areas
  - Data preprocessing
- Creativity
- Rigor and correctness

Let’s have fun!

Data x knowledge

- Data:
  - refer to single and primitive instances (single objects, people, events, points in time, etc)
  - describe individual properties
  - are often easy to collect or to obtain (e.g., scanner cashiers, internet, etc)

Data x Knowledge

- Knowledge
  - refers to **classes** of instances (sets of...)
  - describes general patterns, structures, laws, principles, etc
  - consists of as few statements as possible
  - is often difficult and time-consuming to find or to obtain

Criteria to assess Knowledge

- correctness (probability, success in tests)
- generality (domain and conditions of validity)
- usefulness (relevance, predictive power)
- comprehensibility (simplicity, clarity, parsimony)
- novelty (previously unknown, unexpected)

Quote

- In the science domain, focus is on:
  - correctness, generality and simplicity
- In economy and industry, focus is on:
  - usefulness, comprehensibility and novelty

"We are drowning in information, but starving for knowledge"

(*John Naisbitt*)