Introduction & Overview

Data Mining & Machine Learning
(F20DL/F21DL)

Overview

- Course description and Important Information
  - Assessment
- Overview on data mining and machine learning
  - Scope and Goals
- Example applications
- Ethical issues related to data
- Summary/Discussion

VLE (Virtual Learning Environment) - CANVAS

https://canvas.hw.ac.uk

- We ONLY use F21DL as Canvas Course
- If you enrolled in F20DL you should also see
  - F21DL 2021-2022 (Data Mining and Machine Learning) on your list of courses
- You will find all relevant information so please make sure to check it regularly:
  - Course Outline
  - Learning Materials / Reading List
  - Python Tutorials
  - Weekly lab tasks
  - Coursework description and guidelines
  - Supplementary material,…more
  - (Campus specific information will be labelled accordingly)

Assessment

TBC

Data Mining and Machine Learning Portfolio
(Group Work)

Why we want you to work in Groups?

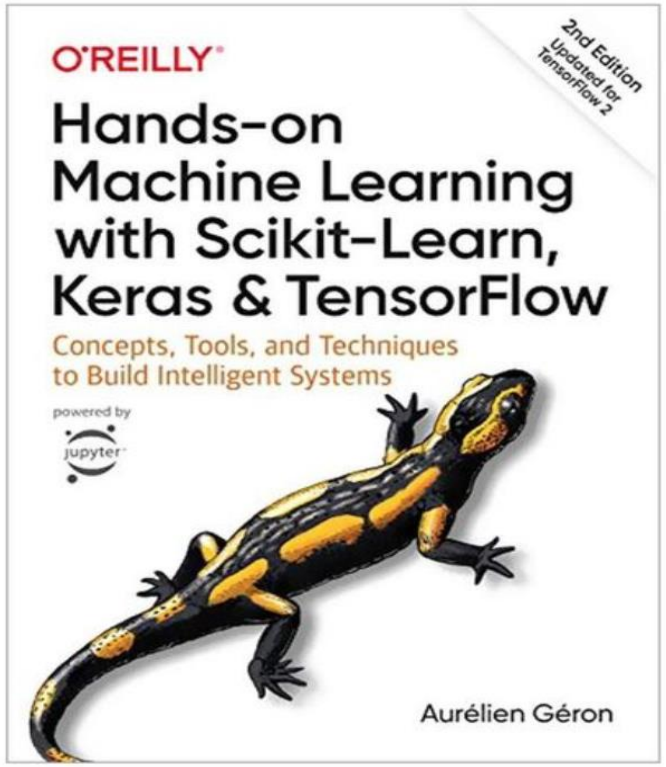
Data Mining and Machine Learning Portfolio (Group Work)

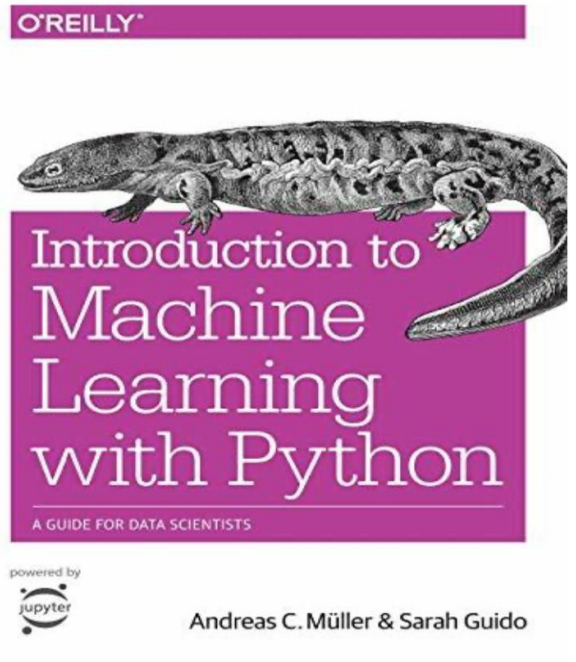
Why we want you to work in Groups?

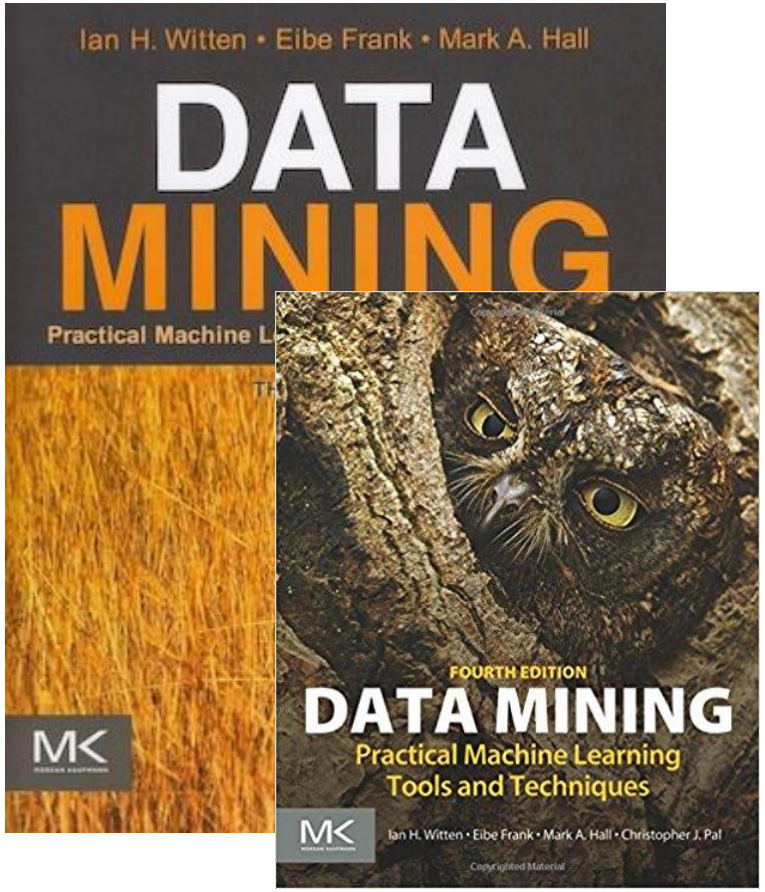
- To encourage discussions for justify findings and explaining conclusions
- To share experiences and explore wider ideas
- To divide tasks and do better research on each topic

Software & Reading

- Python
  - 6 Python tutorial on data processing & scikit-learn ML
- Important links to online resources are provided
  - Recommended books with GitHub repositories







- **Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition** Ian H. Witten, Eibe Frank and Mark A Hall, Elsevier 2011
- New 4th Edition
  - "Deep learning"

Course Structure

Part 1

- Introduction
- Input Preparation
- Knowledge Representations
- Algorithms and Basic Statistics
- Evaluation and Testing

- Week 6: No formal lecture

Part 2 Weeks 5-11

- 4 learning approaches
  - Probability: Bayesian Networks
  - Unsupervised Learning: Clustering
  - Supervised Linear Learning: Decision Trees
  - Supervised Non-Linear Learning: Neural Networks

- Lab tasks & weekly practice exercises & test (check schedule)

How to Succeed on this Course?

How to Succeed on this Course?

- Attend your lectures and follow up on course material (WEEKLY !!!)

  This course is very FAST PACE and things will build up very quickly

- Complete your weekly practice exercises as per schedule !! Only then you will pass your final exam!!

- Make use of your Python material and read the instructions well. Come prepared to the labs & ASK QUESTIONS

- Start working on your CW on the DAY it gets released !! ASK QUESTIONS. Your tutors are happy to help.

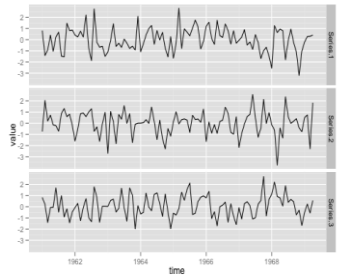
  (your coursework will require experimentation and research beyond classroom time)

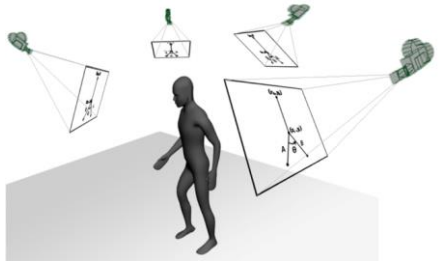
Introduction

Data Mining and Machine Learning

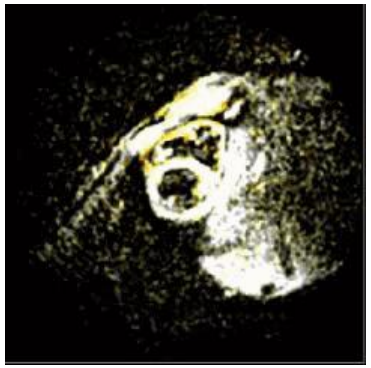
DATA

(Volume-Variety- Veracity-Velocity)





Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



How to Make Sense of Massive Amounts of Data?

IoTX 2015

- 90% of world's data in last 2 years
- 2020, 40% of data from sensors
- 500 million tweets per day
- YouTube: 48 hours of video every minute
- Hospitals: 665 terabytes of patient data, 80% unstructured (CT scans and x-rays)
- Over 100bn emails per day
- Facebook: 100 terabytes of data daily.
- 571 new websites per minute.
- More data created in 2012 than in prior 5,000 years
- Now: More than 5 billion people virtually interacting.

IBM:
Everyday in 2017 2.5 exabytes of data were generated
(1,000,000,000 GB)

300 MB / person to process /day

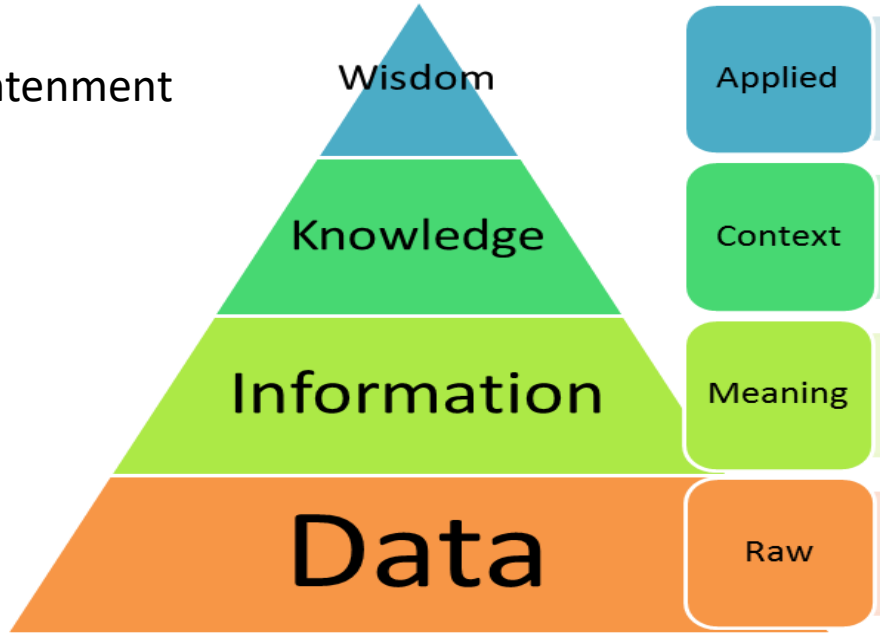
Careem (March 2019)
TB of data processed everyday
structured and unstructured data highly diversified
high velocity
(100 M GPS updates/day – 30K events/sec - 50 M calculations/day)

Towards Data Science June 2019: daily data output is more than 2.5 quintillion bytes.

2020 and beyond: "1.7 Mb of data will be created every second for every person on the planet."

Gaining Insights from Large Volumes of Data

Wisdom ----> Enlightenment



Data Analytics and Machine Learning is the KEY!!!