SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Data Science

PROF. D. HERREMANS

50.038 Computational data science

About the class

- Prof. Dorien Herremans
- Prof. Soujanya Poria

- Teaching assistants:
  - Raven: kinwai_cheuk@mymail.sutd.edu.sg
  - Jyun: yinjyun_luo@mymail.sutd.edu.sg

- Lecture: Tuesday 8:30 – 11:30am
- Labs: be present and on time!

Assessments

- Lab checkoffs (20%)
- Final exam (40%)
- Final project (40%): solve data science problem
  - Initial presentation – week 8
  - Final presentation – week 13
  - Final report – week 12

Brief overview of classes

- Week 1 (DH):
  - L: Introduction and Big data, Hadoop and MapReduce
  - Lab: MapReduce and Hadoop
- Week 2 (SP)
  - L: Feature vectors, dimension reduction, evaluation
  - Lab: feature handling in Python
- Week 3 (DH)
  - L: Data visualization + Data handling (unix/parsing) / guest speaker
  - Lab: visualization in Python
- Week 4 (SP)
  - Regression algorithms – Time series
  - Lab: time series + regression in Python

Brief overview of classes II

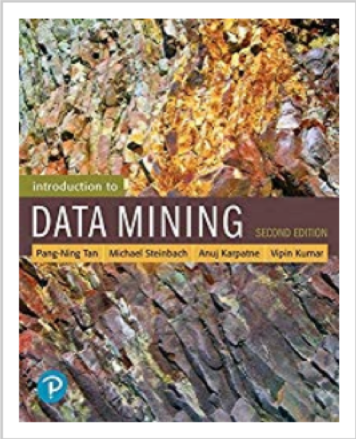
- Week 5 (DH)
  - L: Classification algorithms
  - Lab: Classification in Python
- Week 6 (SP)
  - L: Intro to deep learning
  - Lab: Multilayer perceptron in Python
- Week 8 (DH): project check off. Students present their current work.
- Week 9 (DH)
  - L: Word2vec / NLP
  - Lab: word2vec in Python

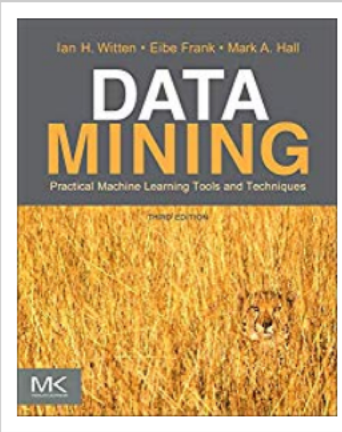
Brief overview of classes III

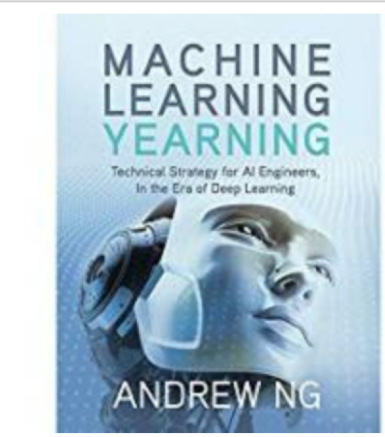
- Week 10 (DH)
  - L: Convolutional neural networks
  - Lab: CNN in Python
- Week 11 (SP)
  - L: Clustering and community detection
  - Lab: practice in Python
- Week 12 (SP)
  - L: temporal sequences / memory models
  - Lab: RNN/LSTM/self-attention
- Week 13 (SP)
  - Final student presentations

Useful references

- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- Tan, P. N., V. Kumar, M. Steinbach (2013). *Introduction to data mining.* Pearson Education India.
- Andrew Ng. 2019. Machine learning yearning.







Intro to data science

What is data science?

- 1960: term used as a substitute for ***computer Science*** (Peter Naur, 1960).
  → datalogy
- 1996: International Federation of Classification Societies (IFCS) ***conference*** in Kobe: “Data Science, classification, and related methods”
- 2012: Harvard Business Review called it:
  "The Sexiest Job of the 21st Century” (Davenport et al., 2012)

Interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured
(Dahr, 2013)

Interdisciplinary buzzwords

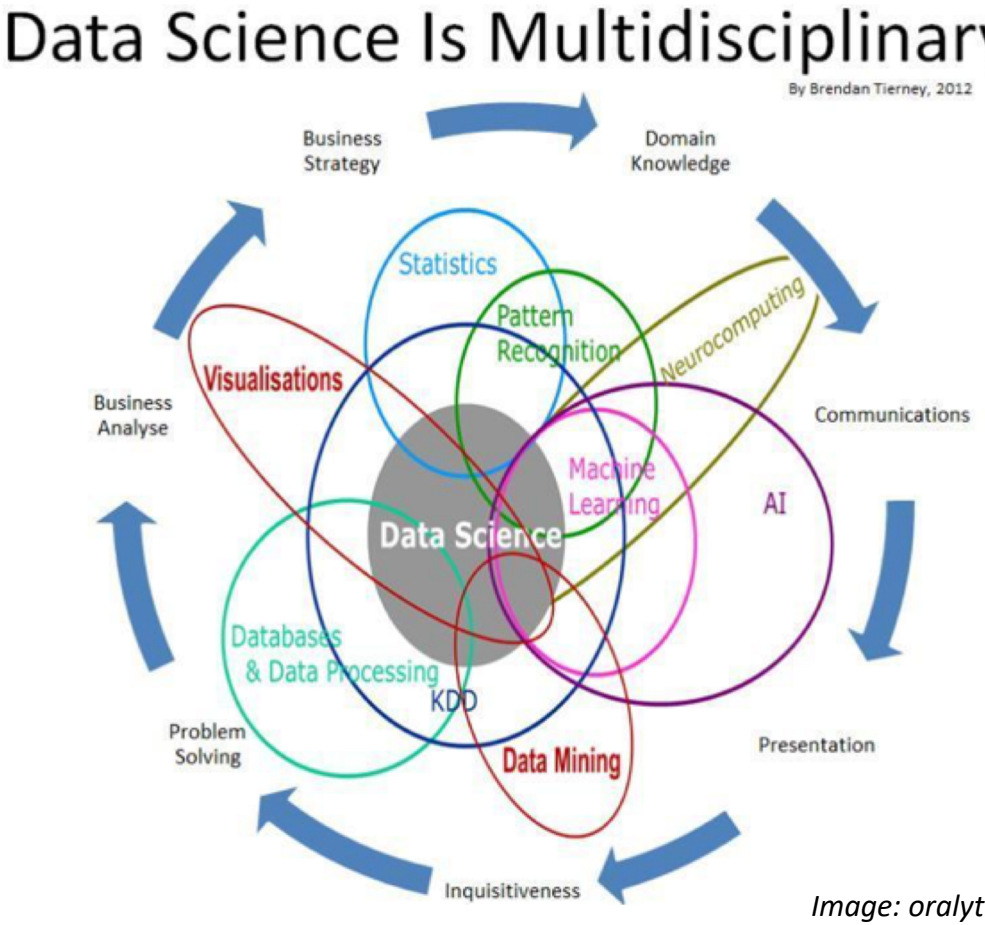


Image: oralytics.com

KDD Process

1. Develop understanding of application, goals
2. Create dataset for study (often from Data Warehouse)
3. Data Cleaning and Preprocessing
4. Data Reduction and projection
5. Choose Data Analysis task
6. Choose Data Analysis algorithms
7. Use algorithms to perform task
8. Interpret and iterate through1-7 if necessary
9. Deploy: integrate into operational systems.

Data mining

SEMMA Methodology (SAS)

- **S**ample from datasets
- **E**xplore datasets, e.g. visualisation
- **M**odify data, e.g. create/transform features
- **M**odel → use algorithms to fit model
- **A**ssess: compare models, test datasets, evaluate reliability/usefulness

Applications of data science

- Fraud prediction
- Document classification (e.g. spam filters)
- Customer churn prediction (e.g. customer leaves)
- Bioinformatics (e.g. Disease prognosis)
- Counter terrorism
- Automatic image captioning
- …

Common data analysis tasks

Task	Supervised Methods	Unsupervised Methods
*Classification	✓	
*Regression	✓	
Causal Modeling	✓	
Similarity Matching	✓	✓
Link Prediction	✓	✓
Data Reduction	✓	✓
*Clustering		✓
Co-occurrence Grouping		✓
Profiling		✓

Data analysis tasks

- “No Free Lunch Theorem”

  => No one algorithm works best for every problem, especially relevant for supervised learning

Hacking skills required

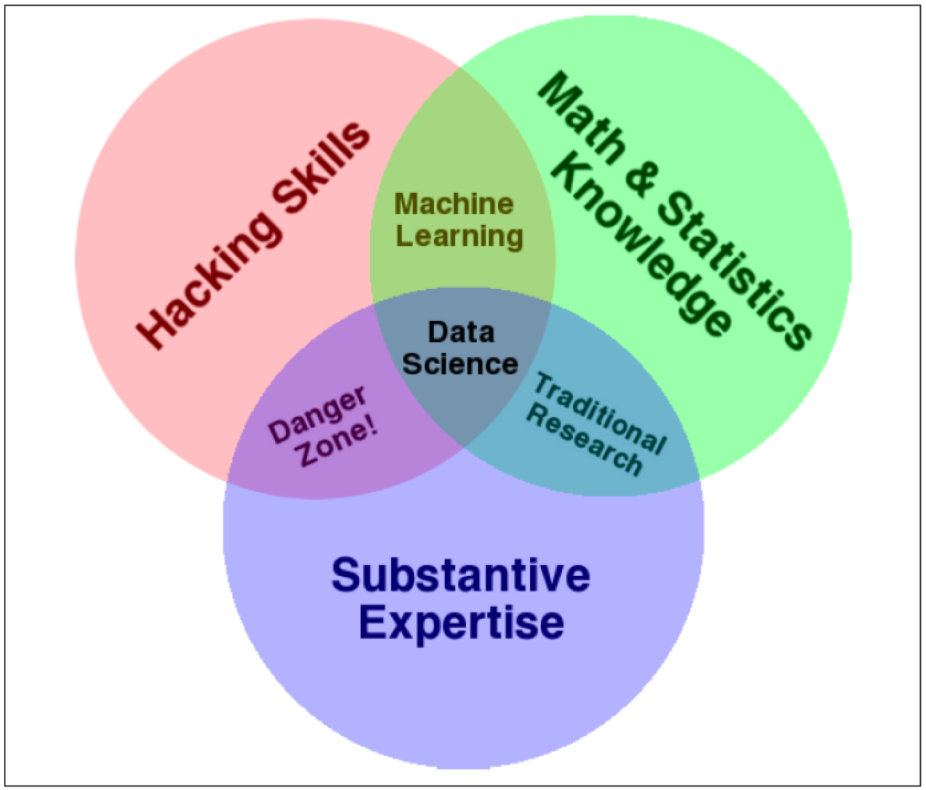
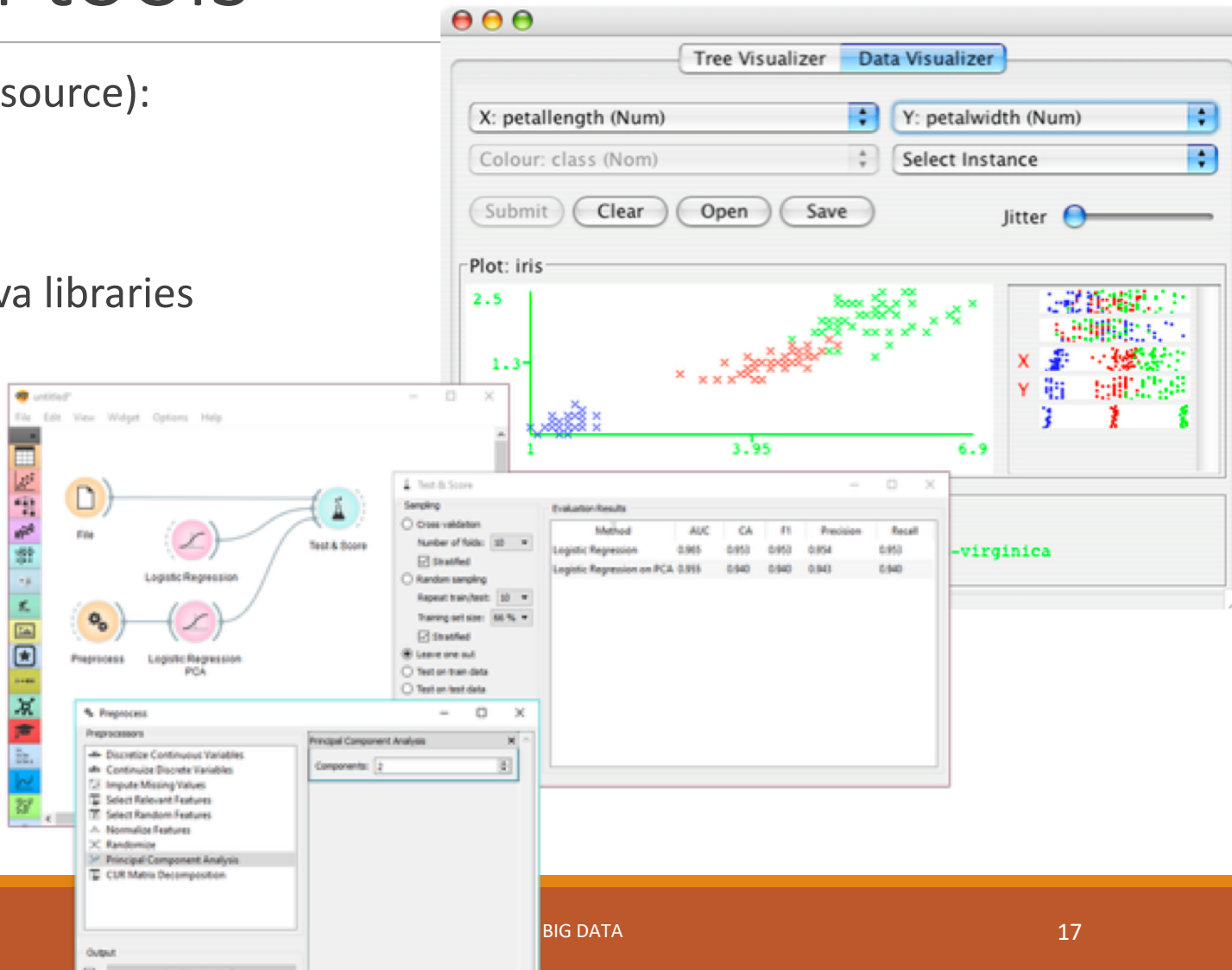


Figure 1-1. Drew Conway's Venn diagram of data science

Useful tools

- Weka (open source):
  - Gui
  - Java library
- R/Python/java libraries
- Rapidminer
- Orange
- Tableau



Upcoming labs

- Access to Unix command line (week 3):
  - Either through (free) Virtualbox software with linux image (Mint, Ubuntu,…)
  - Unix terminal emulator for windows
- All other labs: Python (pref. version 3) -> can be through Google Colab
- This week's lab: Hadoop, see virtual machine install instructions