EPI4932 Clinical Data Science

Maike Imkamp
Leonard Wee
Sander van Kuijk

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

WHAT IS CLINICAL DATA SCIENCE?

Fundamentals of Clinical Data Science

‘*Features*’

Predictors

‘*High-dimensional feature space*’

…a lot of predictors

‘*Confusion matrix*’

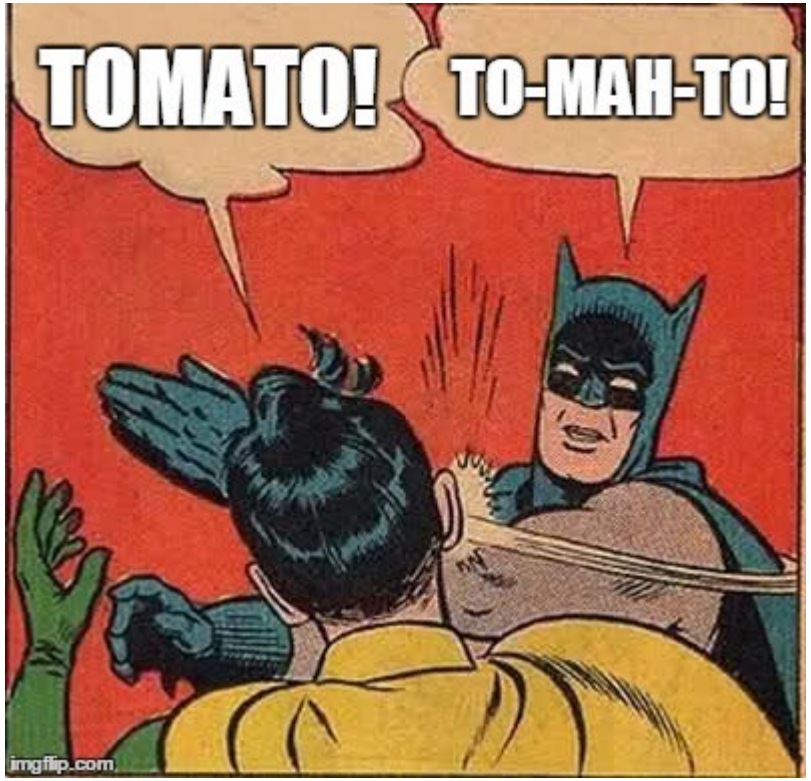
Two-by-two table

		Test Result		
		Positive	Negative	Total
True Condition	Positive	True Positive (A)	False Negative (C)	A + C
	Negative	False Positive (B)	True Negative (D)	B + D
	Total	A + B	C + D	A + B + C + D

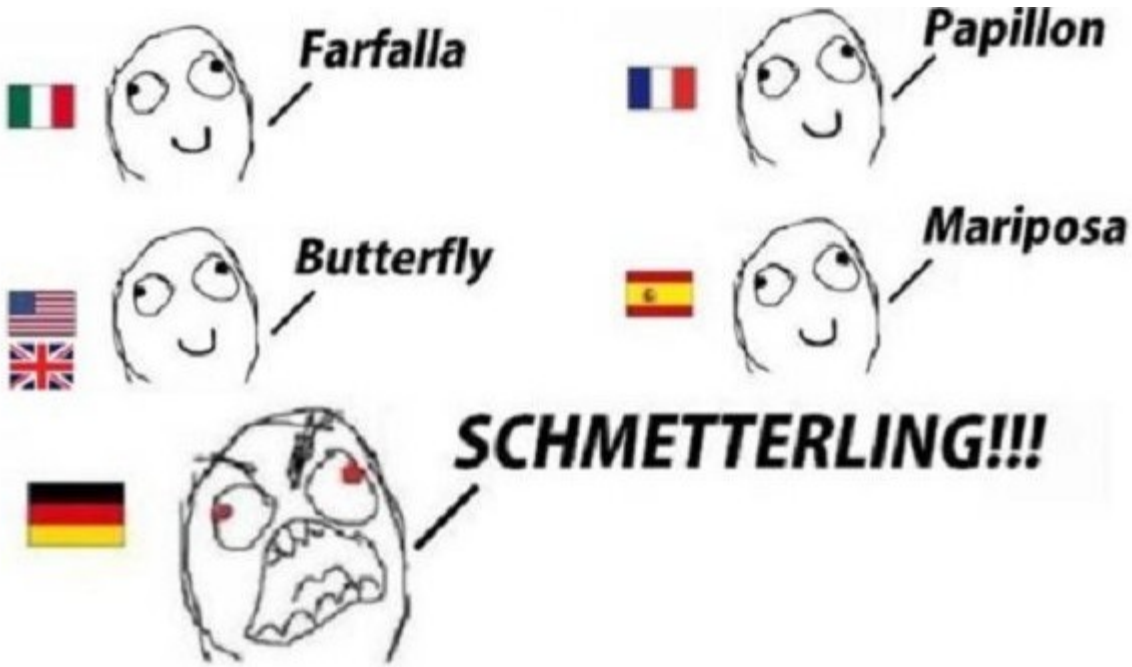
Link to picture

‘*Training and test datasets*’

Development and validation data


TOMATO!
TO-MAH-TO!
imgflip.com

HOWEVER, IT'S NOT *JUST* LANGUAGE DIFFERENCES



HOWEVER, IT'S NOT *JUST* LANGUAGE DIFFERENCES

- Epidemiology: theory-driven
  - Explaining
    - Causal inference
    - Selection bias, information bias, confounding
- Data science: data-driven
  - Learning from data
    - Unsupervised learning
    - Supervised learning
    - Reinforcement learning

Example where we do meet: prediction

- Epidemiology
  - Regression techniques
- Data Science
  - Machine learning/ artifical intelligence

SO WHERE DID WE ALL COME FROM?

- 19th century: biostatistics developed for measuring human traits and quantifying morbidity and mortality
- 19th century: epidemiology evolved from medicine in response to infectious disease crises; diverged to a chronic disease perspective
  - methods were developed to mitigate the effects of bias and confounding (developed by statisticians)
- 20th century: data science emerged due to the need for computer scientist skills in analyses

Goldstein et al. On the Convergence of Epidemiology, Biostatistics, and Data Science. Harv Data Sci Rev. 2020; 2(2)

The association between the number of beds in an intensive care unit and infection risk:

"*… the epidemiologist suggests assembling a retrospective cohort from the electronic medical records (EMR) for a one-year period and the data scientist is able to interface with the EMR, retrieve a patient list, and abstract all of the variables necessary for analysis. The biostatistician conducts a rigorous analysis, including assessing completeness of the data and identifying potential biases in the analysis.*"

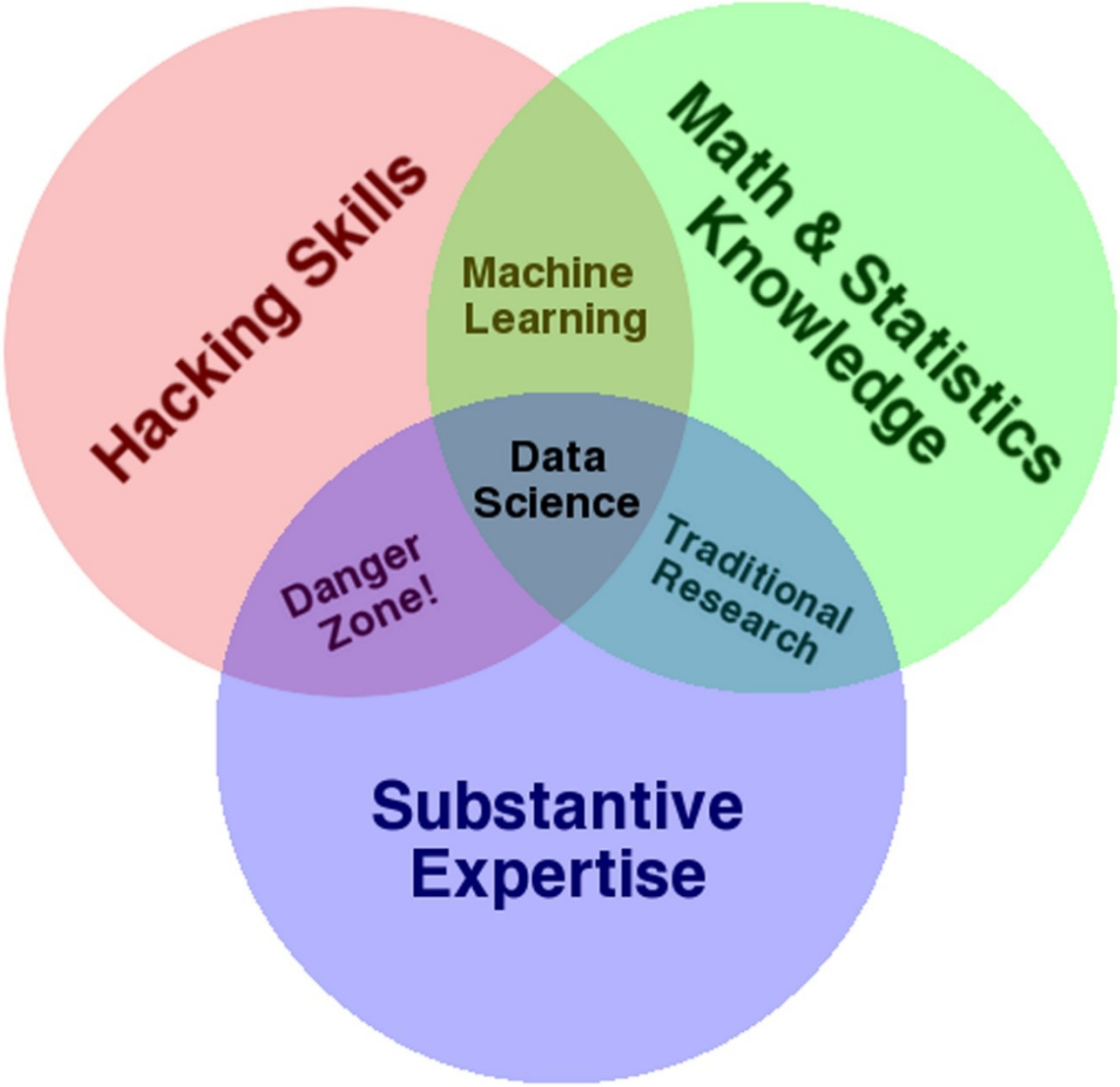
Goldstein et al. On the Convergence of Epidemiology, Biostatistics, and Data Science. Harv Data Sci Rev. 2020 ; 2(2)

The association between the number of beds in an intensive care unit and infection risk

"… the epidemiologist suggests assembling a retrospective cohort from the electronic medical records (EMR) for a one-year period and the data scientist is able to interfaces with the EMR, retrieve patient list, and abstract all of the variables necessary for analysis. The biostatistician conducts a rigorous analysis, including assessing completeness of the data and identifying potential biases in the analysis."

NARROW VIEW!

Goldstein et al. On the Convergence of Epidemiology, Biostatistics, and Data Science. Harv Data Sci Rev. 2020 ; 2(2)



The data science Venn diagram. (Conway, 2013).

BRIDGING THE GAP

- We're all specialists in the analysis of quantitative clinical data
  - Research design
  - Statistical methods
  - Substantive expertise

- The focus may differ!
  - Data scientists often have strong math background
  - Epidemiologists often have strong study design/ bias background

WHAT CAN *WE* LEARN?

- (Clinical) Data Science:
  - Database design, data linkage skills
  - Unsupervised learning techniques are used much more often
  - Lightyears ahead in reproducibility of research

WHAT CAN *WE* LEARN?

- (Clinical) Data Science:
  - Database design, data linkage skills
  - **Unsupervised learning techniques are used much more often**
  - **Lightyears ahead in reproducibility of research**

- **Two main themes:**

1. Reproducible research
2. Unsupervised learning
3. Supervised learning (Elective by Laure)

WE'LL BE YOUR TEACHERS

"An article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result."

- James Buckheit and David Donoho

THE IMPORTANCE OF REPRODUCIBILITY

- Reproducibility enhances replicability
  - Code sharing between research groups!
- Pragmatism: being able to reproduce all steps in virtually no time, easily work in teams
- Journals may require it
- Funding bodies may require it
- FAIR principles (**F**indability, **A**ccessibility, **I**nteroperability, and **R**euse)

REPRODUCIBILITY AT TWO LEVELS

1. Manuscript contains all necessary steps to reproduce study/ experiment
   - E.g. use reporting guidelines
   - Make sure to describe decisions made
   - Will always be a brief summary!

Epi SOP

2. "the full software environment, code and data"

Not yet epi SOP

- Data scientists are software developers

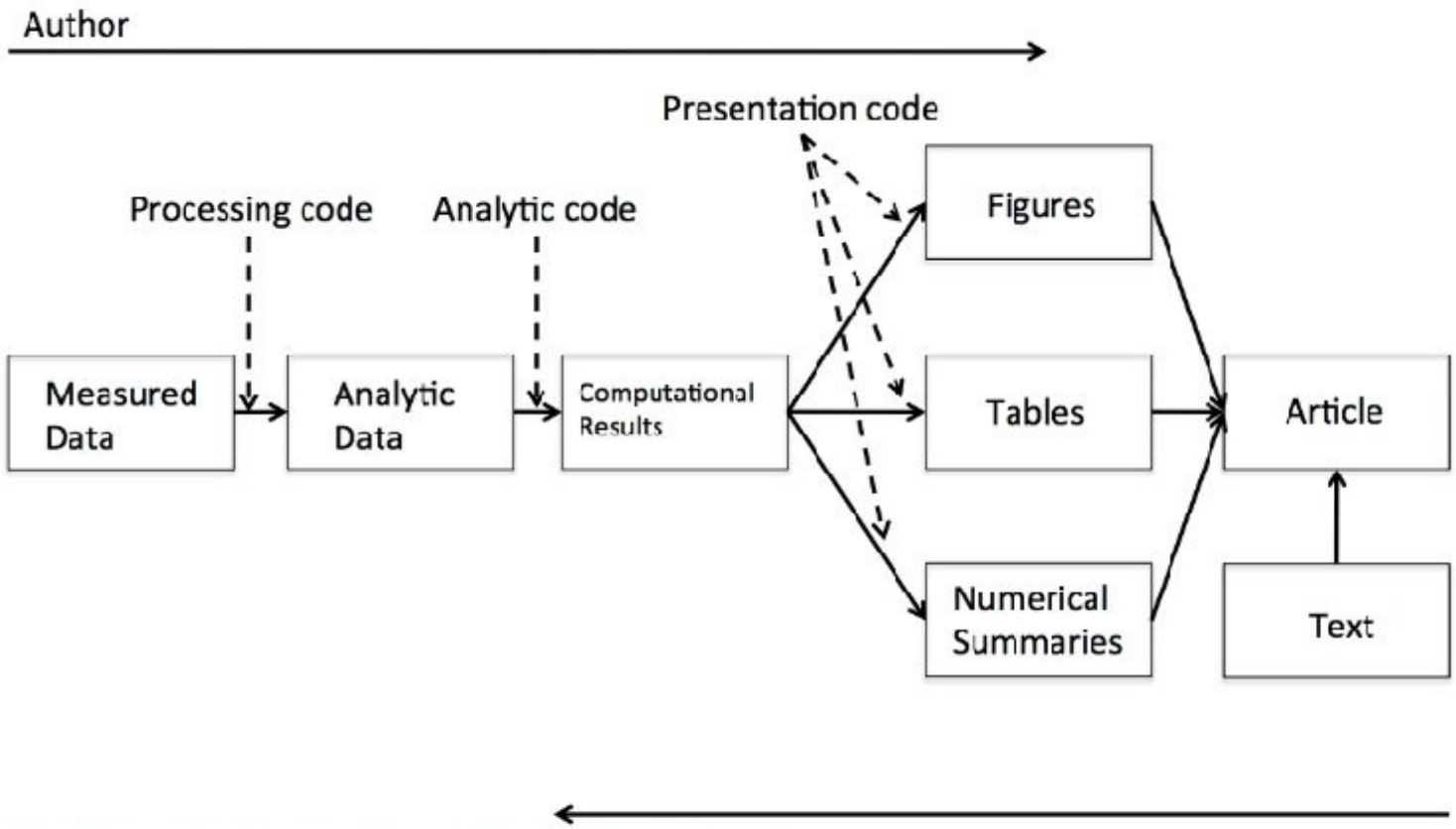
- Are epidemiologists software developers?

GOAL:

You should be able to reproduce all findings of your project.

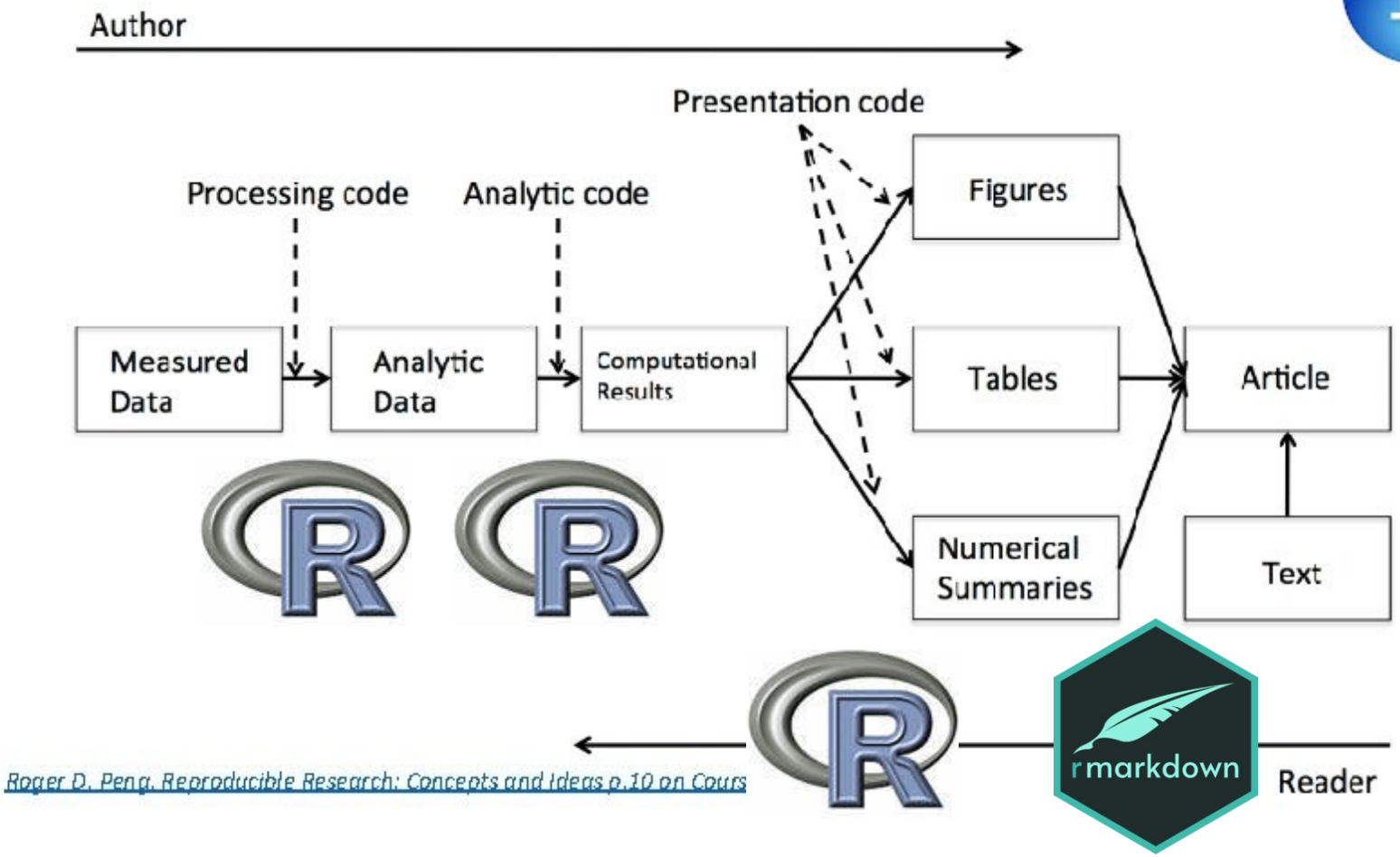
An independent researcher who is provided with data and code should be able to reproduce all findings of your project.

Research Pipeline



Roger D. Peng. Reproducible Research: Concepts and Ideas p.10 on Coursera

Research Pipeline



Roger D. Peng. Reproducible Research: Concepts and Ideas p.10 on Cours

TEACHING ACTIVITIES: REPRODUCIBILITY

- Git/ GitHub practical
- R markdown practical

Take a break!

CODING BEST PRACTICES

1. Write programs for people, not computers
2. Let the computer do the work
3. Make incremental changes
4. Don’t repeat yourself
5. Plan for mistakes
6. Optimize software only after it works correctly
7. Document design and purpose, not mechanics
8. Collaborate

Wilson et al. Best Practices for Scientific Computing. PLOS Biology, 2014.

PROGRAMS FOR PEOPLE, NOT COMPUTERS

- Examples: make code style and formatting consistant
  - E.g.: use Google's R Style Guide for formatting
  - Explain the code u wrote

The future of epi studies?

OXFORD

Briefings in Bioinformatics, 2022, 1–15
https://doi.org/10.1093/bib/bbab571
Problem Solving Protocol

MiRNA–disease association prediction based on meta-paths

Liang Yu, Yujia Zheng and Lin Gao
Corresponding author: Y. Liang, School of Computer Science and Technology, Xidian University, Xi'an 710071, P.R. China. Tel.: +86-13759921156; Fax: 0086-029-88202427. E-mail: lyu@xidian.edu.cn

"*Code and data are available at https://github.com/LiangYu-Xidian/MDPBMP.*"

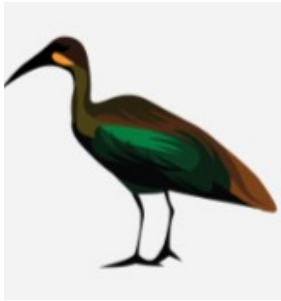
Acknowledgements

"*Thanks to all those who maintain excellent databases and to all experimentalists who enabled this work by making their data publicly available.*"

UNSUPERVISED LEARNING

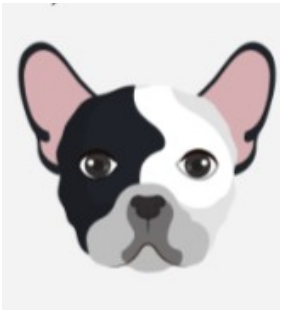
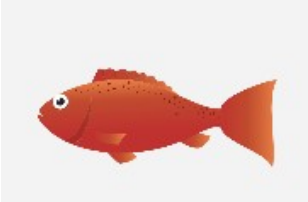
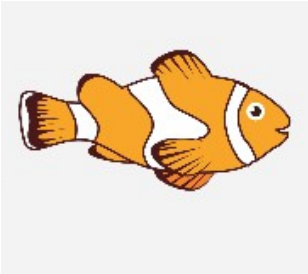
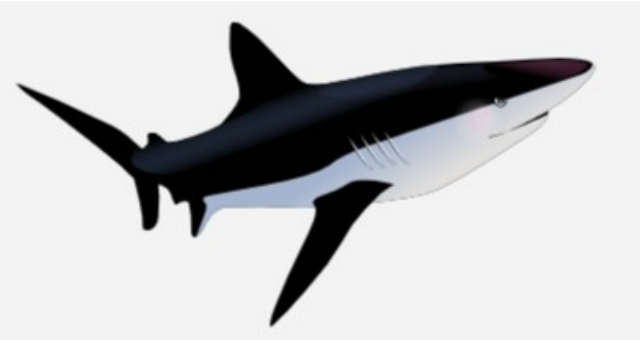
UNSUPERVISED LEARNING







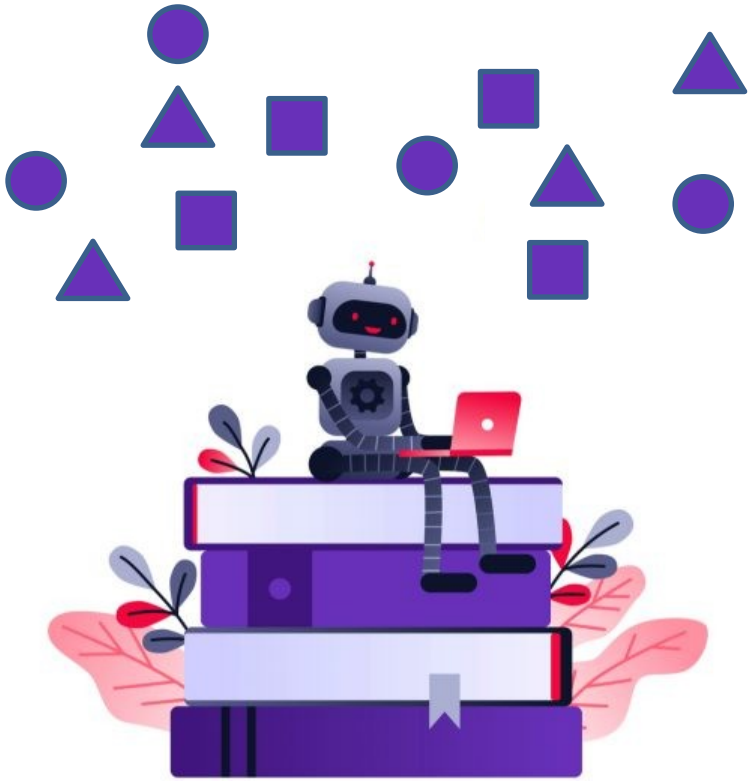


WHAT WOULD E.T. DO?

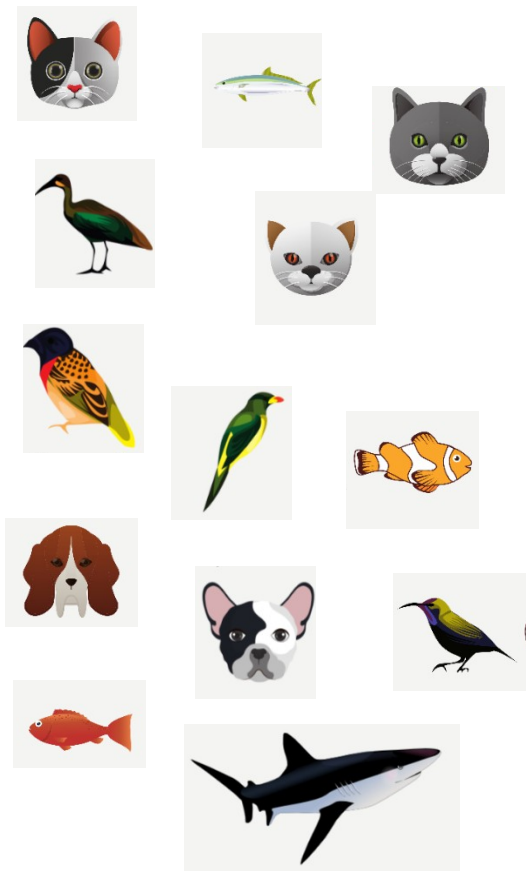
UNSUPERVISED LEARNING

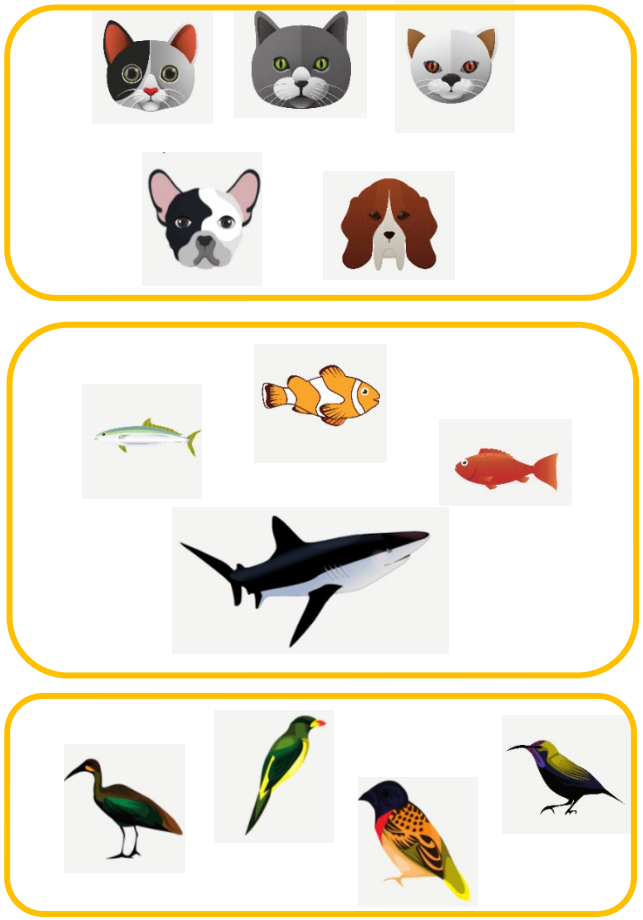
UNSUPERVISED LEARNING



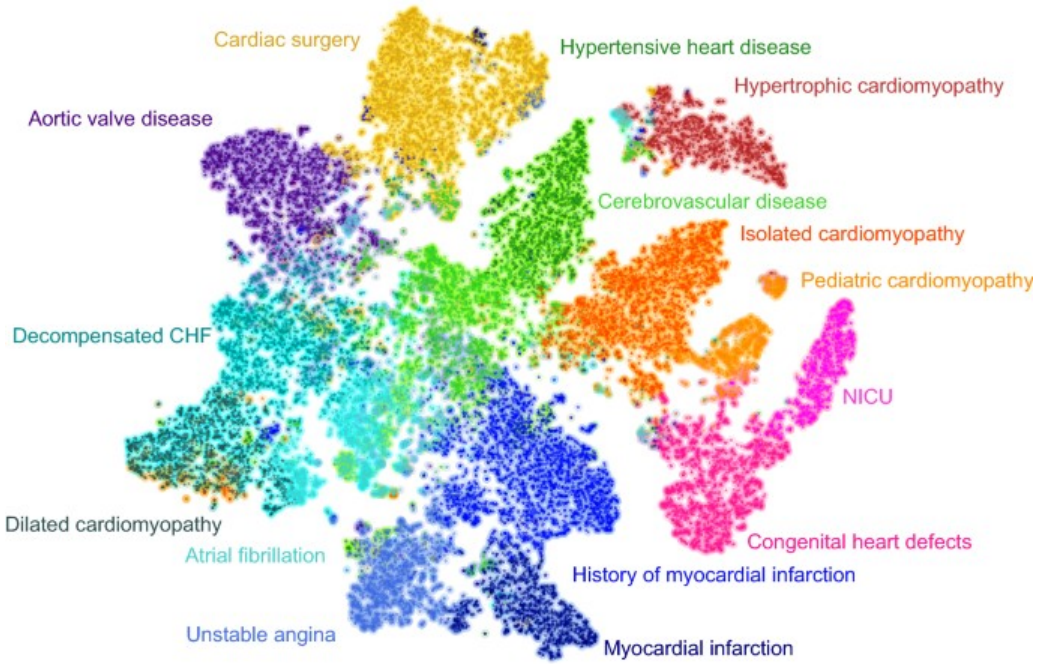
UNSUPERVISED LEARNING

UNSUPERVISED LEARNING





UNSUPERVISED LEARNING



TEACHING ACTIVITIES UNSUPERVISED LEARNING

- Data exploration lecture
- Data visualisation practical
- Clustering lecture and practical

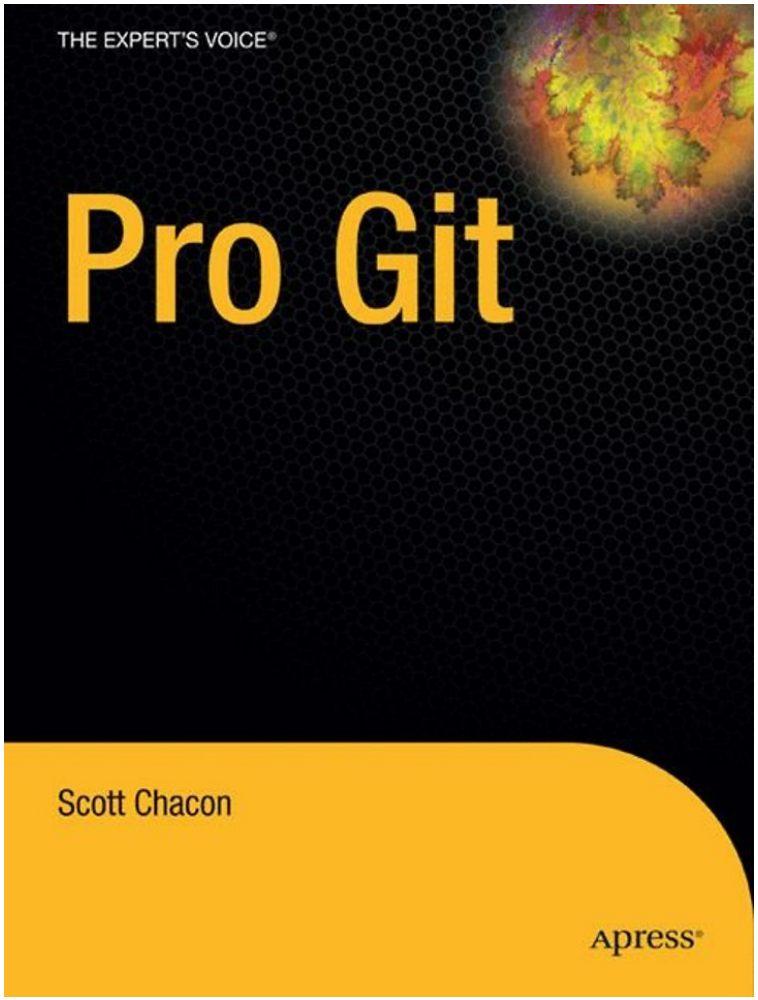
COURSE ASSIGNMENT

- Bottom-up research project!
  - Start with the data
  - End with the research question
- Abstract (one by each research team)
  - Max. 3 tables/ figures, multiple on a grid is allowed!
- Assessment
  - Abstract (pdf/ Word/ HTML)
  - Collaboration! "*the full software environment, code and data, that produced the result*"

RESEARCH TEAMS

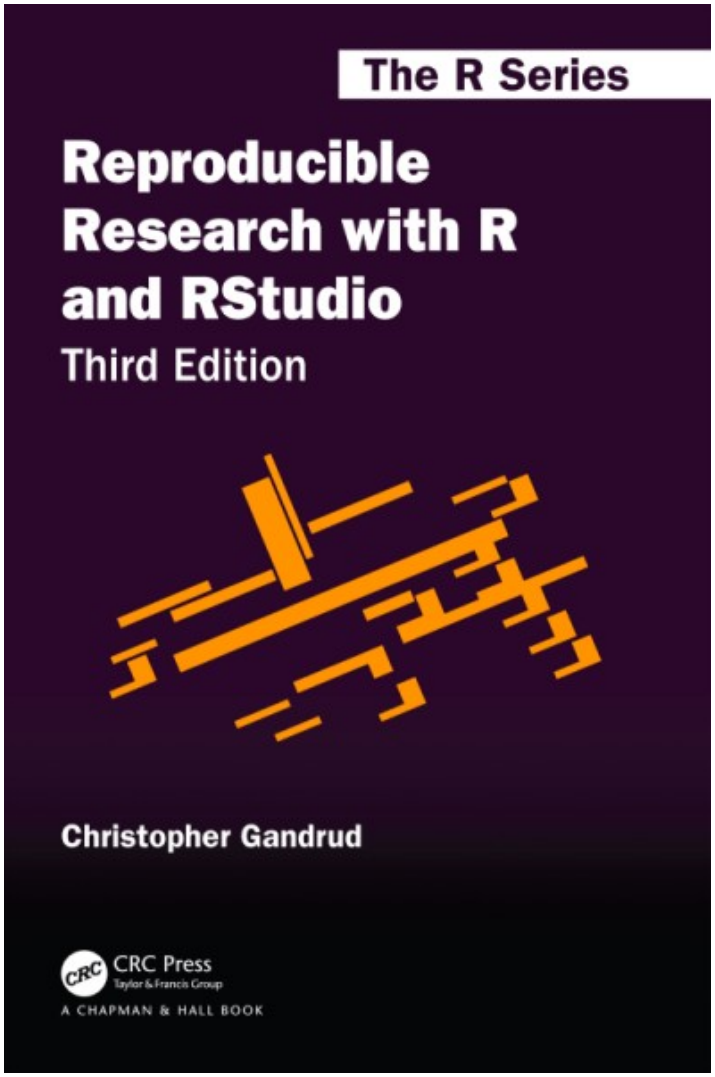
1. Casper, Dick, Christina, Simon
2. Hala, Lara, Jasper, Anke
3. Christine, Maria, Philipp, Paddy

Suggested literature (1)



Free online edition available here

Suggested literature (2)



Free online previous edition available here

Suggested literature (3)

- Interesting papers:

[1] Goldstein et al. On the Convergence of Epidemiology, Biostatistics, and Data Science. doi: 10.1162/99608f92.9f0215e6.

[2] Wilson et al. Best Practices for Scientific Computing. doi: 10.1371/journal.pbio.1001745.

[3] Bi et al. What is Machine Learning? A Primer for the Epidemiologist. doi: 10.1093/aje/kwz189.