3050571 Practical Clin Data Sci

Session 1: Course introduction

January 30, 2024

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Self introduction

Computational Molecular Biology Group

Sira Sriswasdi

(สิระ ศรีสวัสดิ์) [sira.sr at chula.ac.th]

Research Affairs, Faculty of Medicine

Postdoctoral Researcher, the University of Tokyo (2013-2017)
Ph.D., Genomics and Computational Biology, University of Pennsylvania (2013)

Ekapol Chuangsuwanich

(เอกพล ช่วงสุวนิช) [ekapolc at cp.eng.chula.ac.th]

Department of Computer Engineering, Faculty of Engineering

Ph.D., Electrical Engineering and Computer Science, MIT (2016)

Juthamas Chaiwanon

(จุฑามาศ ชัยวนนท์) [juthamas.c at chula.ac.th]

Department of Botany, Faculty of Sciences

Ph.D., Biology, Stanford University (2015)

Naruemon Pratanwanich

(นฤมล ประทานวณิช) [naruemon.p at chula.ac.th]

Department of Mathematics and Computer Science, Faculty of Science

Ph.D., Computer Science, University of Cambridge (2017)

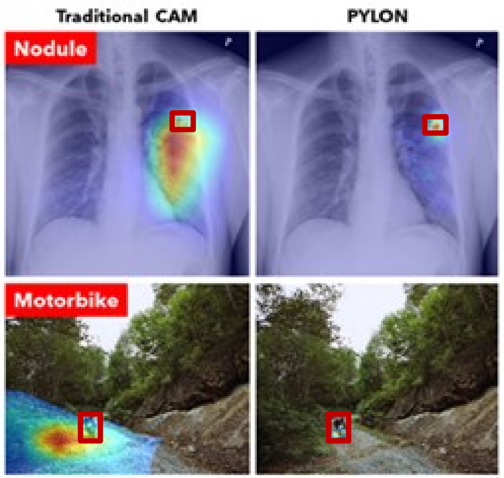
- ❑ Combine basic knowledge in mathematics, computer sciences, biology, and bioinformatics to solve problems
- ❑ 2 Postdoc, 4 MEng, 8 graduate students

George Genchev

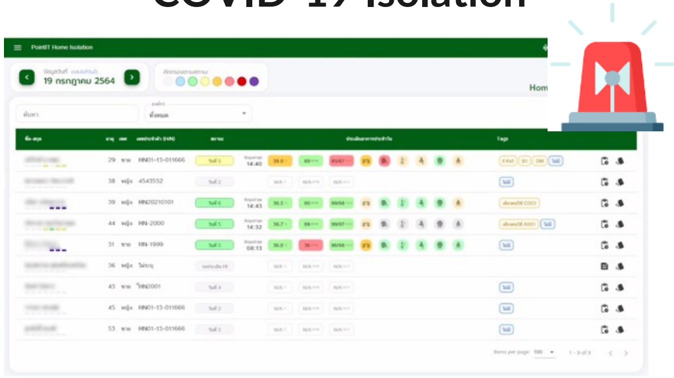
Aijaz Ahmad Malik

Center for Artificial Intelligence in Medicine

Explainable CXR AI

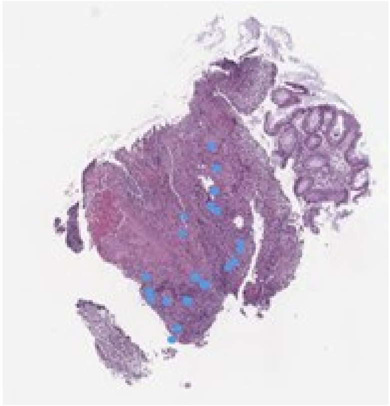


Remote Monitoring for COVID-19 Isolation



With WeSAFE and Burapha Univ.

Digital Pathology



With Institute of Pathology

- Powered by **NVIDIA DGX-A100** and **HPC**
- Provide computing resources and consultation

CHULA ENGINEERING
Foundation toward Innovation
COMPUTER

UNIVERSITY TECHNOLOGY CENTER

โรงพยาบาลจุฬาลงกรณ์ สภากาชาดไทย

NVIDIA

About this course

- 6-week elective for 5th year medical student with interest in data

- **Key topics**
  - Computational thinking
  - Problem solving with computer programming (Python)
  - Data analysis, visualization, and storytelling
  - Machine learning and deep learning

- **Learning styles**
  - Assigned online videos, readings, and Python practices
  - In-class recitation, discussion, and Python workshops
  - Internship with KCMH data team

Objectives

- This course is designed for you:
  - Introduce you to key foundations in data science and machine learning
  - Give you tools to handle the data

- You should understand:
  - The assumption and motivation behind a technique
  - How to use the Python library

- Get you to ask a lot of questions, **both at me and at the data**

Internship with KCMH data team

- **Learning style**
  - Learn about data-driven projects at KCMH
  - Observe how the data team approach the problems
  - Identify where you can contribute

- **Expectations**
  - Pick one project to work on personally or as a team
  - Develop a proposal on how to approach the project
  - Identify available data to evaluate your proposal
    - I will help supervise on the technical aspects

Weekly contents

Week	Topics	Practices
1	Computational thinking	Basic Python programming
2	Data exploration, visualization, and storytelling	Extract knowledge and tell story from data
3	Unsupervised machine learning - Dimensionality reduction - Clustering	Identify patterns and patient subpopulations from various datasets
4	Supervised machine learning - Linear models - Tree models	Predict hospital admission using linear and tree models
5	Introduction to deep learning and AI	Build a small artificial neural network Predict future pneumonia in COVID-19
6	Explainability and AI project design	Full machine learning project pipeline

A typical weekly schedule

Week 1: Introduction and Python programming						
	9-10	10-11	11-12	13-14	14-15	15-16
Monday						
Tuesday				Lecture		
Wednesday				Internship with KCMH data		
Thursday		Internship with KCMH data		Lecture		
Friday				Python workshop		

Grading criteria

- Assignments [60%]
  - Can ask for guidance
  - Can work with each other
  - Can use AI to help, but report how you used it

- Internship [40%]
  - Performance evaluation [20%]
    - Participation
    - Effort
  - Final presentation [20%]

cmb-chula / clini-data-science-3050571 Public

Code | Issues | Pull requests | Discussions | Actions | Projects | Security | Insights

main | 1 Branch | 0 Tags | Go to file | Code

sirasris Update README.md | 1233e11 · 2 days ago | 10 Commits

3050571_assignment.pdf	Assigned study and practice	3 days ago
3050571_syllabus_schedule.pdf	Syllabus and schedule	3 days ago
README.md	Update README.md	2 days ago

README

Welcome to 3050571 Pracical Clinical Data Science

This is the repository for the learning materials from the **3050571: Pracical Clinical Data Science** course taught by our group at the Faculty of Medicine, Chulalongkorn University in Bangkok, Thailand in Spring 2024.

Week 1 - Computational Thinking

Keep these learning points in mind as you study the contents

Key learning points

- What is computational thinking and how do you apply it to solve problem?
- How to systematically approach a problem?

Assigned study

These videos cover more than what I expect you to learn, but they are all beneficial for you in the long term

- Computational Thinking video and reading
- A perspective on programming vs coding first 3 min
- Three (3) things to do when starting out in Data Science
- Optimization problem from MIT 6.0002 Lecture 1 and Lecture 2

Assigned practice

Assignments WILL take time. Get started early.

- Python code editors
- Kaggle Intro to programming and Python lessons

Example of assigned task

There may be only one primary goal, but there are many stories and hypotheses that can be told

Titanic - Machine Learning from Disaster

Overview Data Code Models Discussion Leaderboard Rules

The Challenge

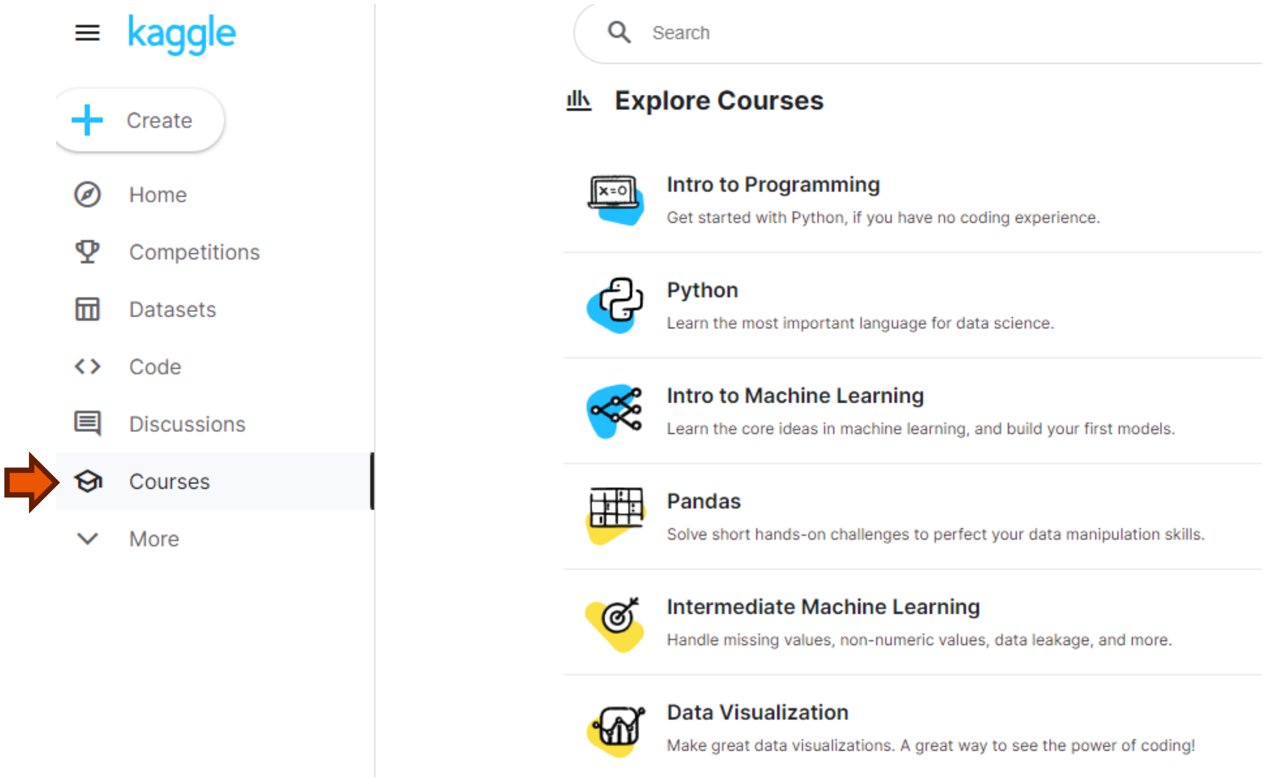
The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

Kaggle


kaggle
Create
Home
Competitions
Datasets
Code
Discussions
Courses
More
Search
Explore Courses
Intro to Programming
Get started with Python, if you have no coding experience.
Python
Learn the most important language for data science.
Intro to Machine Learning
Learn the core ideas in machine learning, and build your first models.
Pandas
Solve short hands-on challenges to perfect your data manipulation skills.
Intermediate Machine Learning
Handle missing values, non-numeric values, data leakage, and more.
Data Visualization
Make great data visualizations. A great way to see the power of coding!

Companion resources

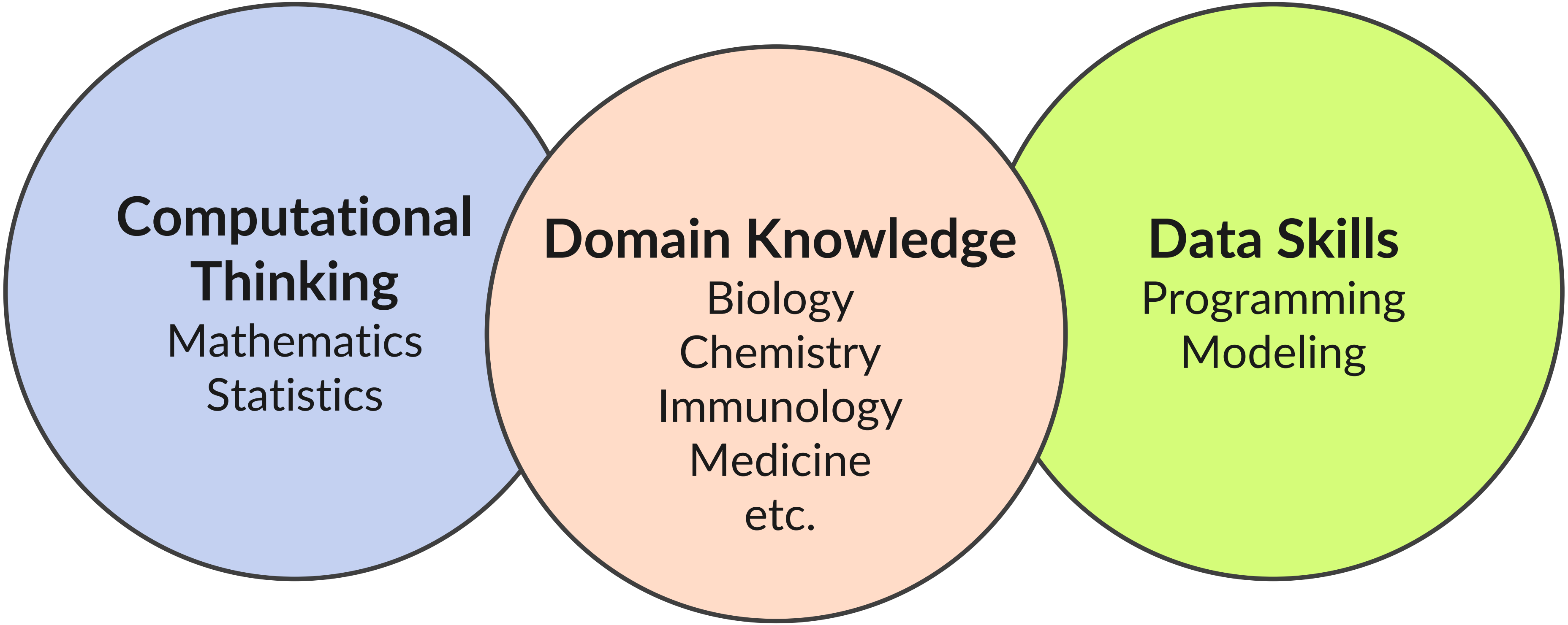
- **MIT 6.0002** – Computational thinking
- **MIT 6.S191** – Deep learning
- **MIT 6.S897** – Machine learning for healthcare

- **StatQuest** YouTube – Explanations of statistical and machine learning concepts

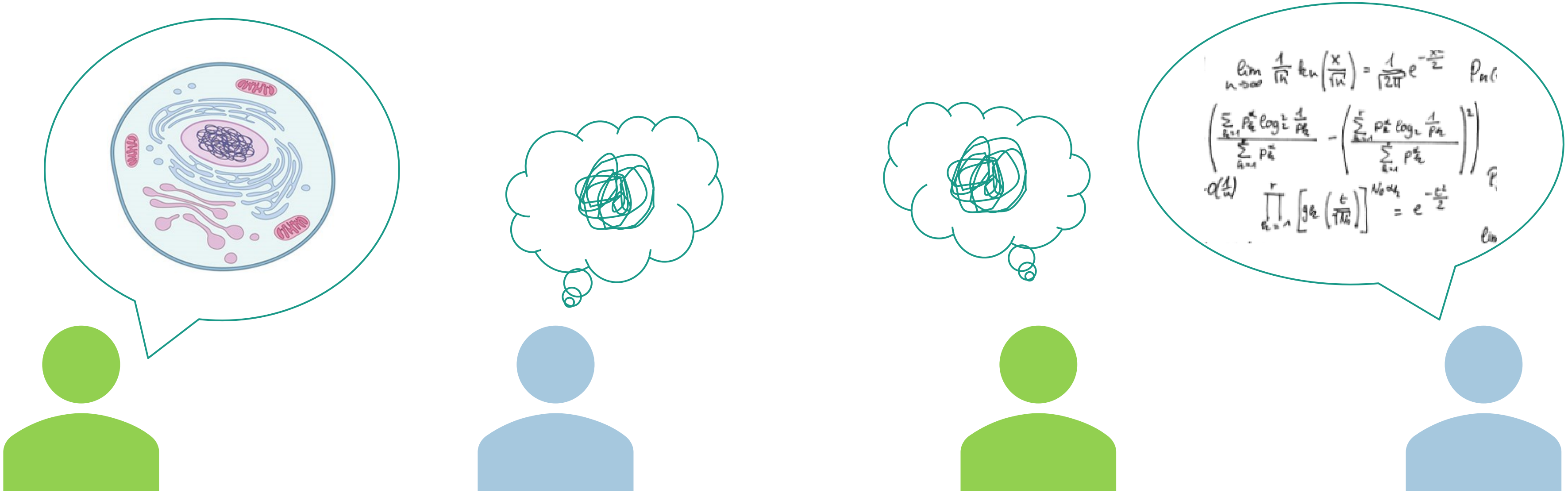
- Machine learning and deep learning courses from **University of Tubingen** and **Stanford University** on YouTube

Computational thinking?

The Trinity of a great data scientist



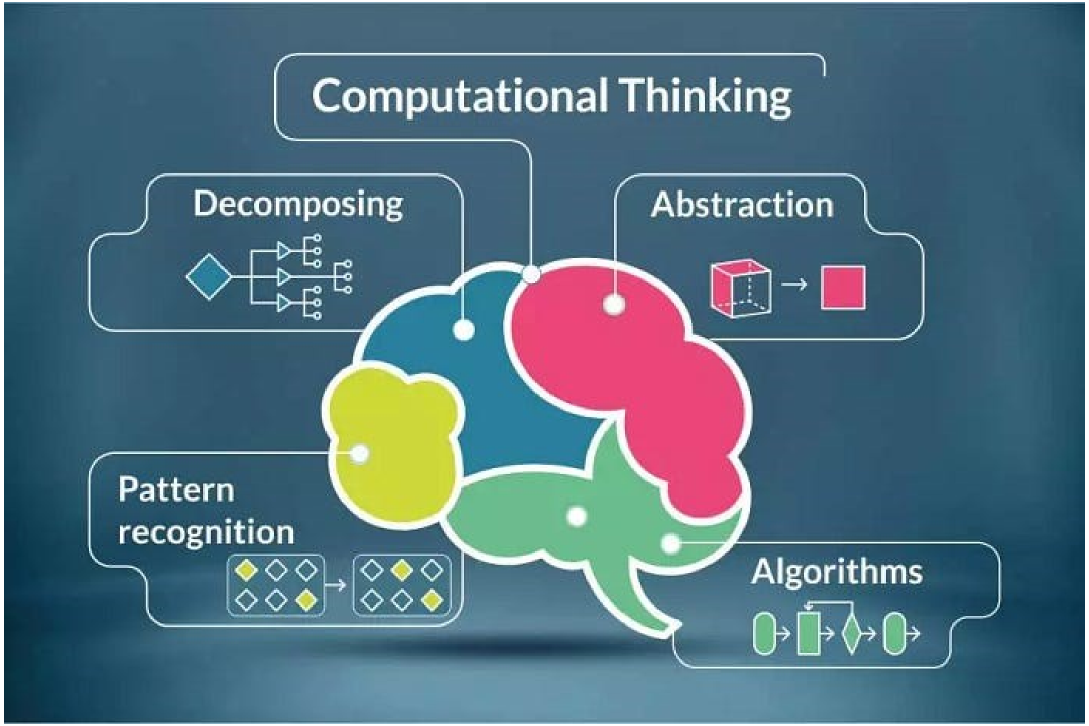
Knowledge enables communication



What is computational thinking?

Breaking down a complex problem into smaller components and relationships

Identifying similarities or patterns in the data to utilize and learn from them



https://www.nextgurukul.in/thenextworld/

Simplifying the variables to focus on the most important factors

Formulating a well-defined step-by-step process to solve the problem

Statistics and hypothesis testing

Same topics but different perspectives

- Revisit the motivation and assumption behind standard techniques
  - How were the p-values calculated?
  - Maximum likelihood principle

- Develop your own tests that fit your data and your hypothesis
  - Permutation test

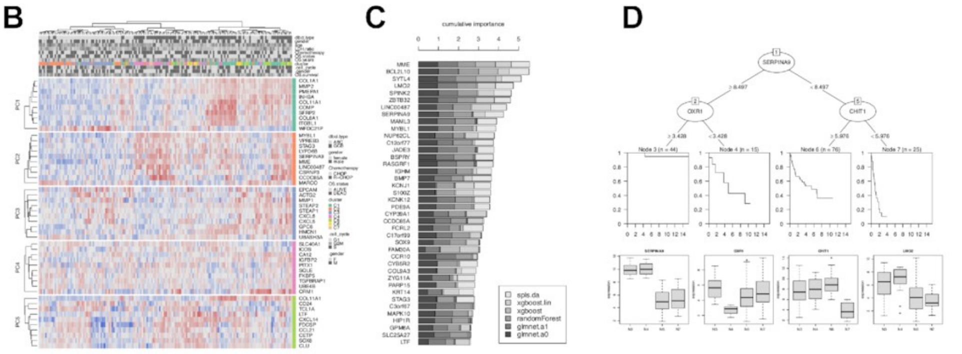
- Integrate statistics with data exploration and visualization on the fly

- Transform statistical knowledge into machine learning knowledge

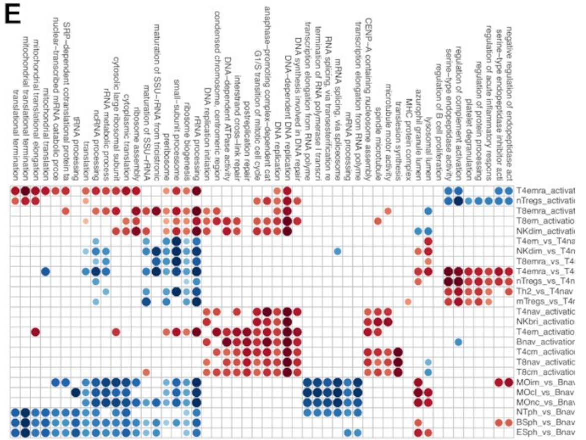
Data exploration, visualization, and storyelling

From raw data to informative graphs

Gene ID	P61_2_C	P62_2_C	P63_2_C	P64_2_C	P68_2_C
ENSG00000000003.14	4.637576	6.183992	5.237635	2.372719	5.665966
ENSG00000000005.5	0	0	0	0	0
ENSG00000000419.12	11.22781	4.813792	2.99782	10.99452	10.7482
ENSG00000000457.13	7.656414	5.082675	7.710682	9.014404	8.488388
ENSG00000000460.16	3.172546	2.245954	5.974815	3.501081	4.162024
ENSG00000000938.12	0	0	0	0.042488	0
ENSG00000000971.15	6.626259	8.19511	5.904925	11.7748	2.050394
ENSG00000001036.13	1.790445	0.76823	3.670635	0.68115	1.894823
ENSG00000001084.11	19.53907	25.08378	11.04872	5.815902	20.23763
ENSG00000001167.14	15.34717	20.00867	17.10001	25.31168	27.41216
ENSG00000001460.17	0.889852	3.090642	0.744581	3.439525	2.417934
ENSG00000001461.16	3.771195	3.12468	1.385353	2.767444	2.973217
ENSG00000001497.16	16.75059	9.662455	15.4965	14.34071	10.62035
ENSG00000001617.11	2.998366	3.712208	3.885852	17.50663	3.019686



Akhmedov, M. et al. NAR Genom and Bioinfor, 2(1):lqz019 (2019)



Everyone struggles with open-endedness

- What to analyze? What to visualize?

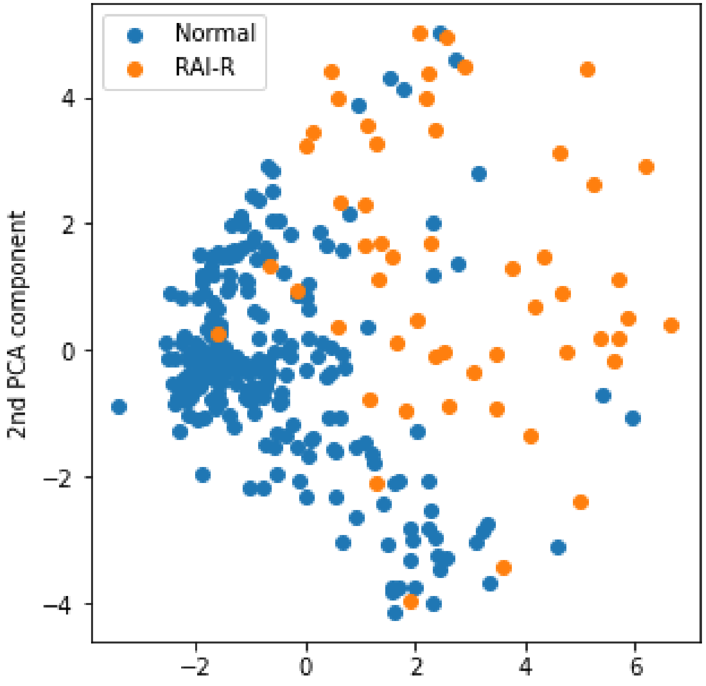
- How to interpret the numbers and graphs?

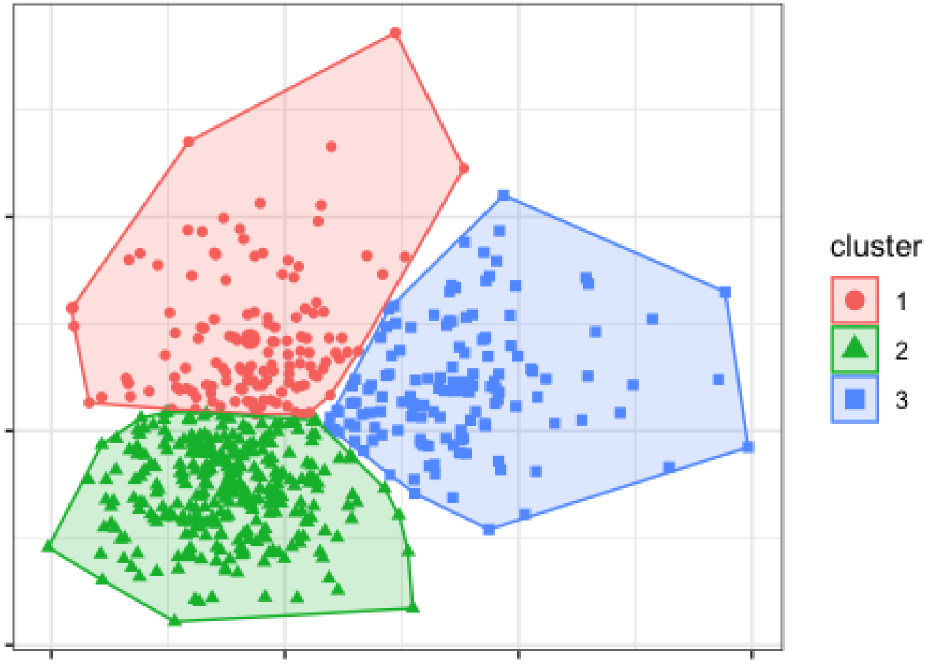
- How to best present to other people?

- How to strengthen your conclusion?
  - Could the association occur by chance?
  - Was there a confounding factor?
  - Is your results specific to the technique used?

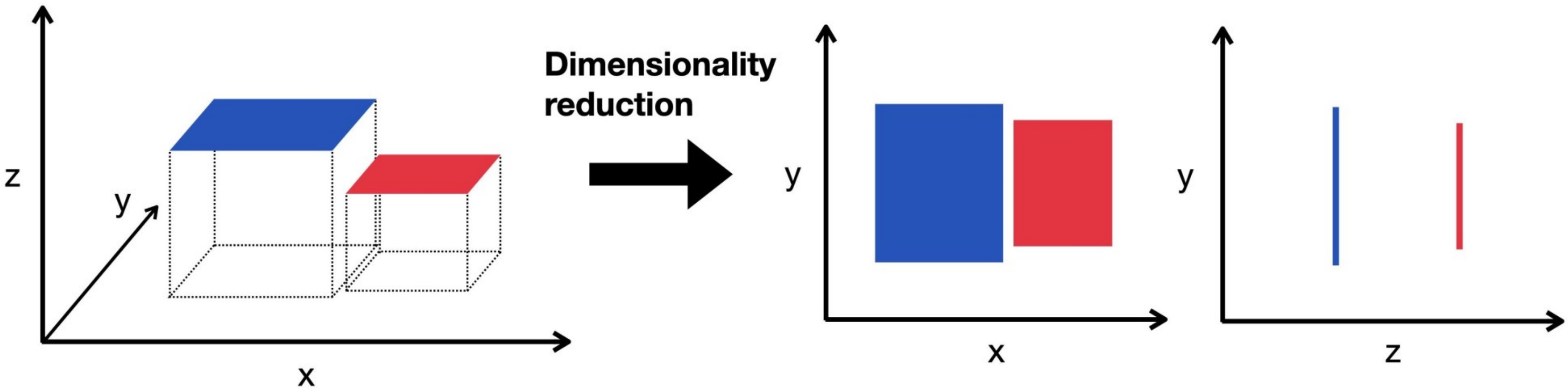
Unsupervised learning

Dimensionality reduction and clustering





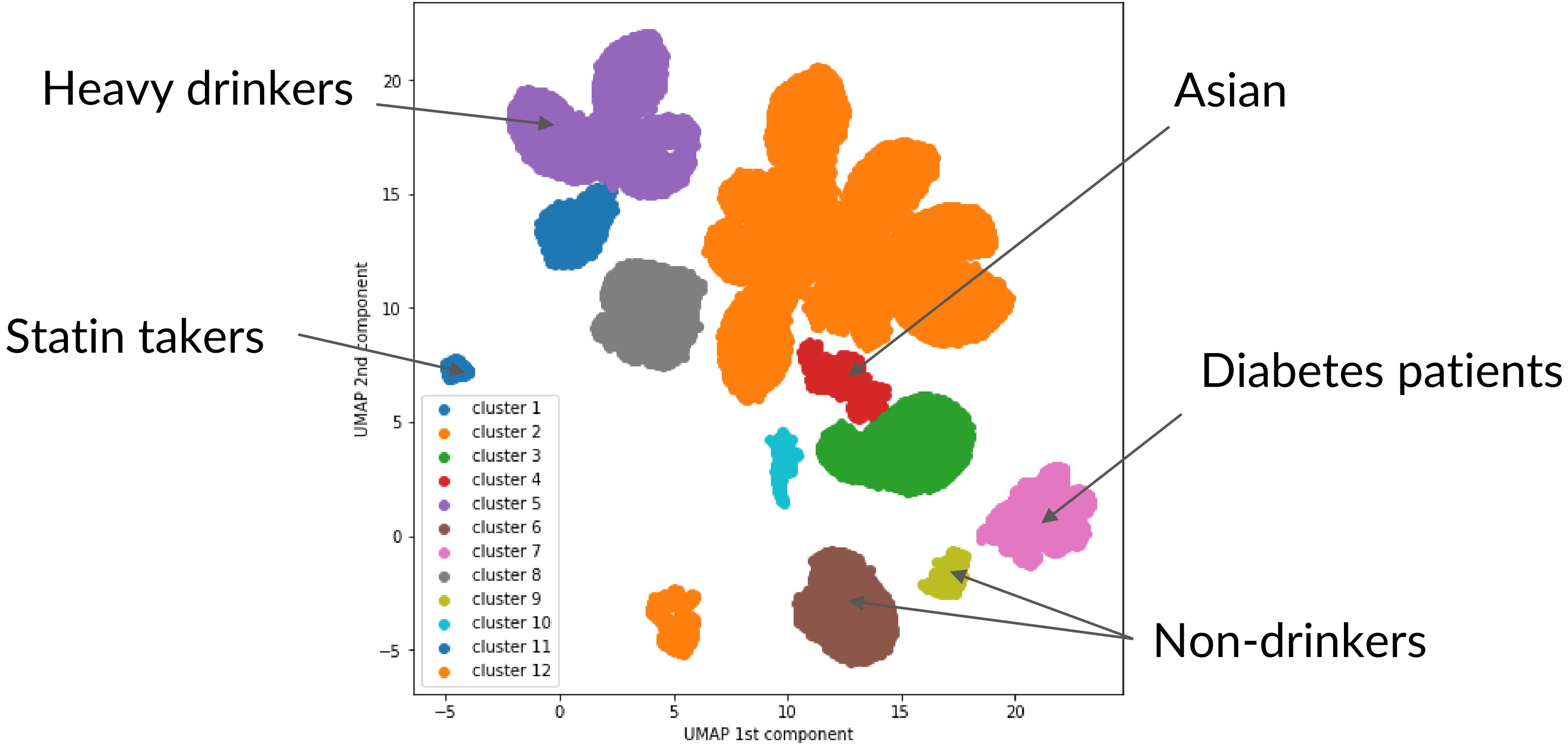
Dimensionality reduction



https://www.sc-best-practices.org/preprocessing_visualization/dimensionality_reduction.html

- Reduce dimension (number of features) while maintaining information
- Patient with similar symptoms also exhibit similar lab tests or have similar demographics or similar medical history

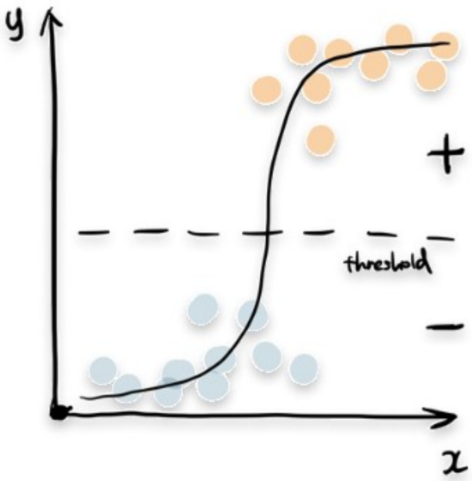
DBSCAN on patient data

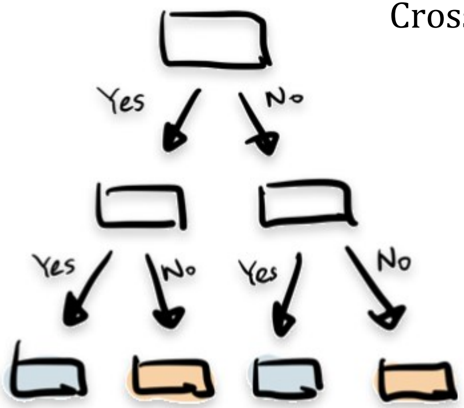


Supervised learning

The cores of supervised learning

Model





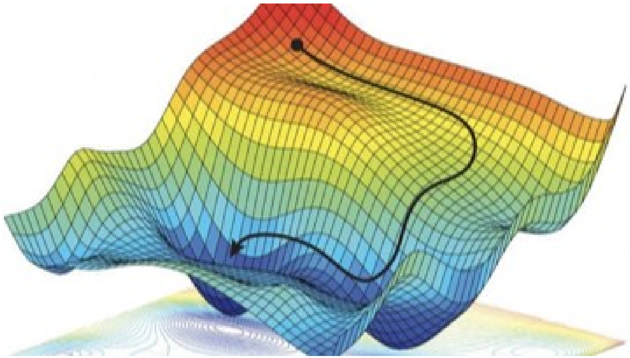
https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

Objective / Loss Function

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

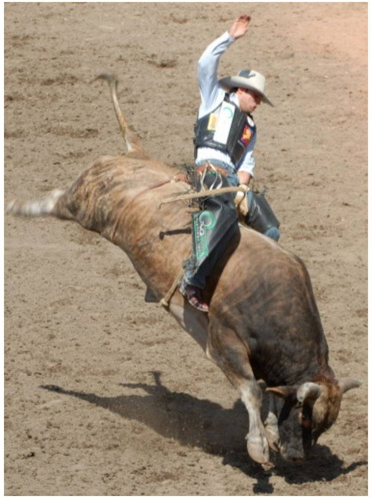
$$\mathrm{Crossentropy} = -\frac{1}{n}\sum_{i=1}^{n} y_i \ln(\hat{y}_i) + (1 - y_i)\ln(1 - \hat{y}_i)$$

Optimization Algorithm

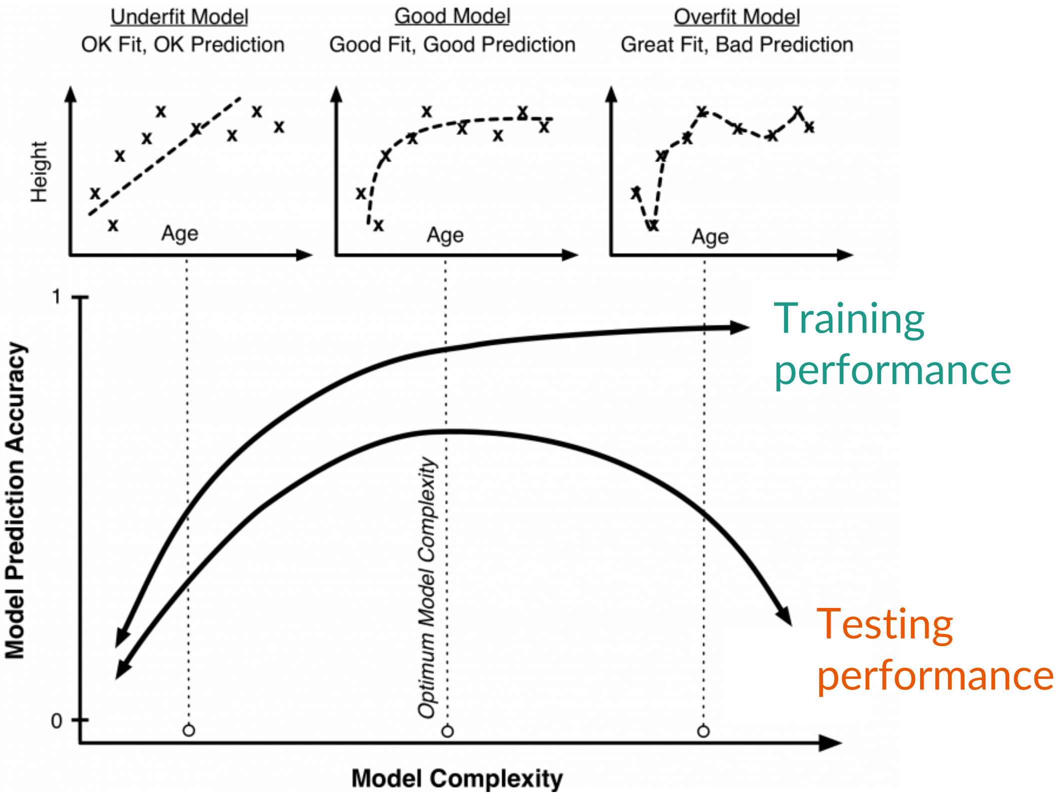


https://medium.com/analytics-vidhya/gradient-descent-b0dc1af33517

Supervised learning is all about control



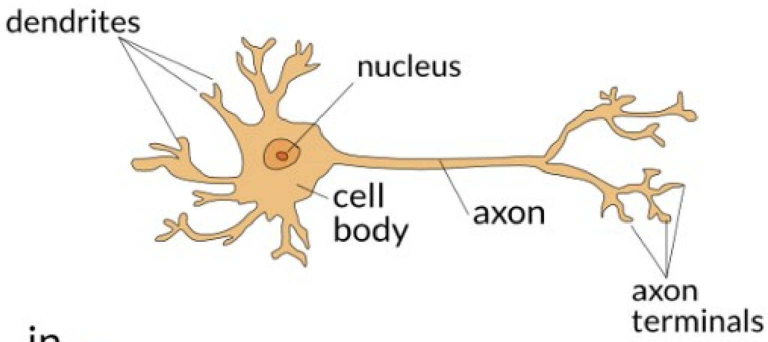
https://en.wikipedia.org/wiki/Bull_riding

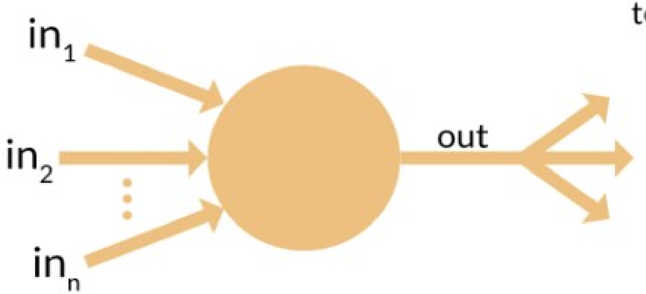


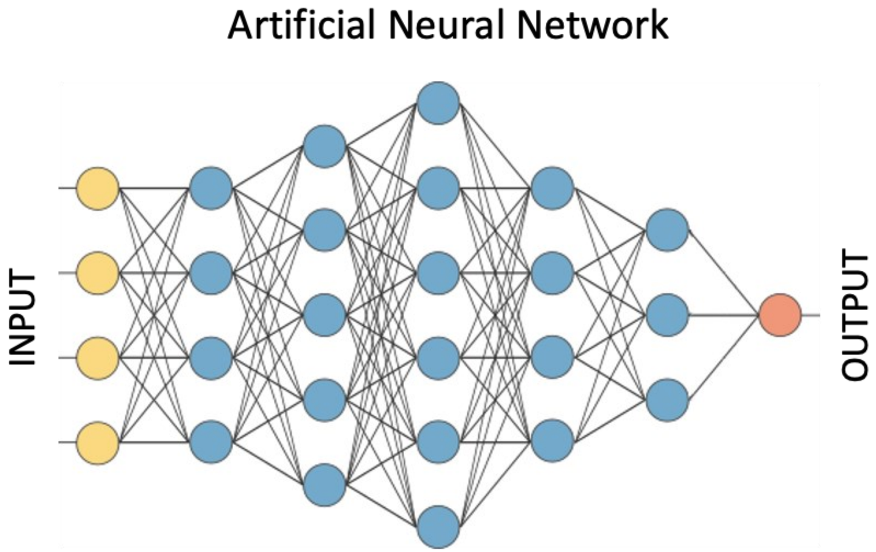
https://realkm.com/2018/04/23/optimization-and-complexity-the-cost-of-complexity-systems-thinking-modelling-series/

Deep learning and AI

Artificial neural network

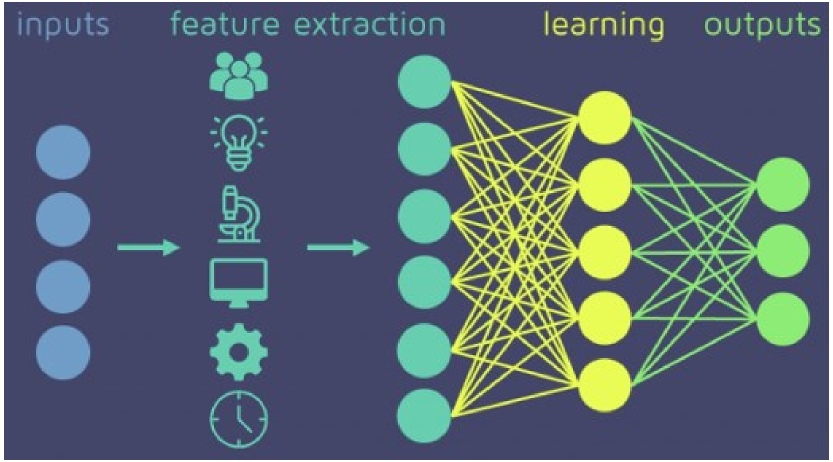


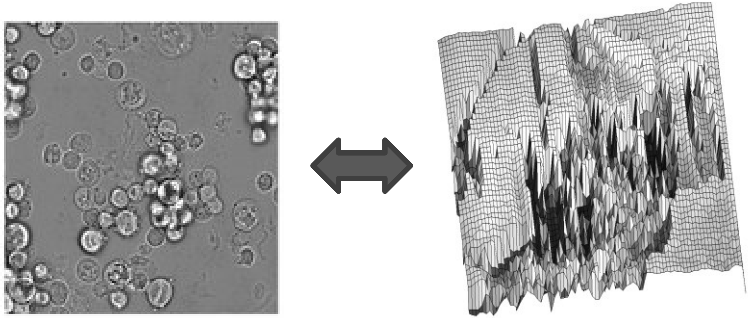


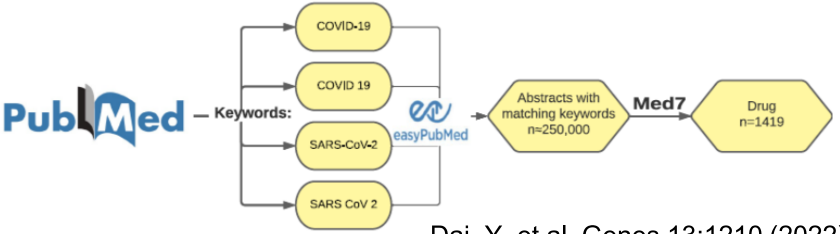


- Network of simple computation nodes: $out = f(w_1 in_1 + w_2 in_2 + \ldots + w_n in_n)$

Limitation of classical (non-deep) learning



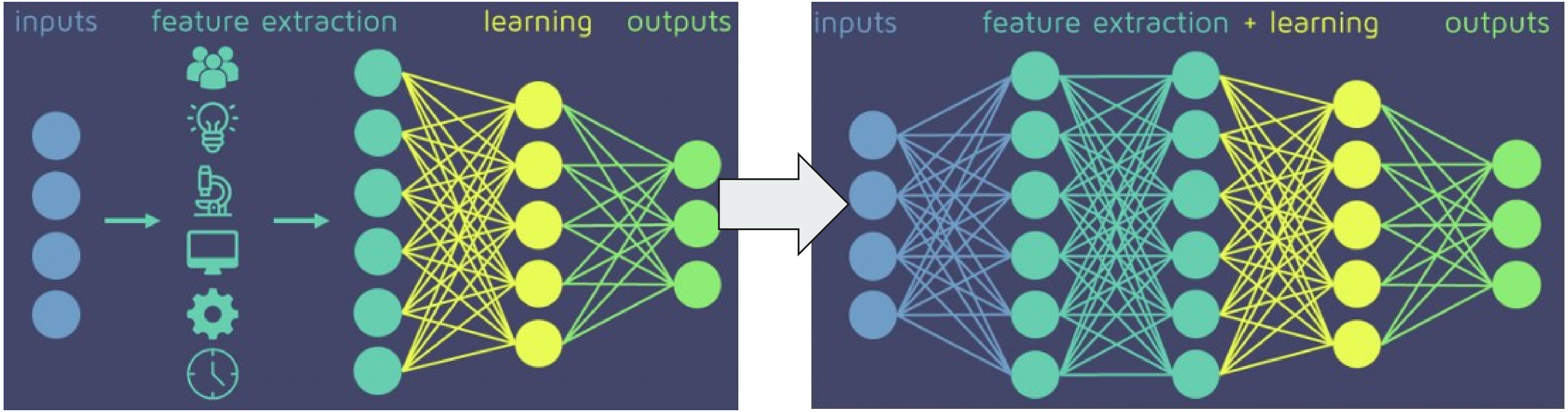




Dai, Y. et al. Genes 13:1210 (2022)

- Classical machine learning requires the input to be formatted and pre-processed by human

End-to-end / representation learning

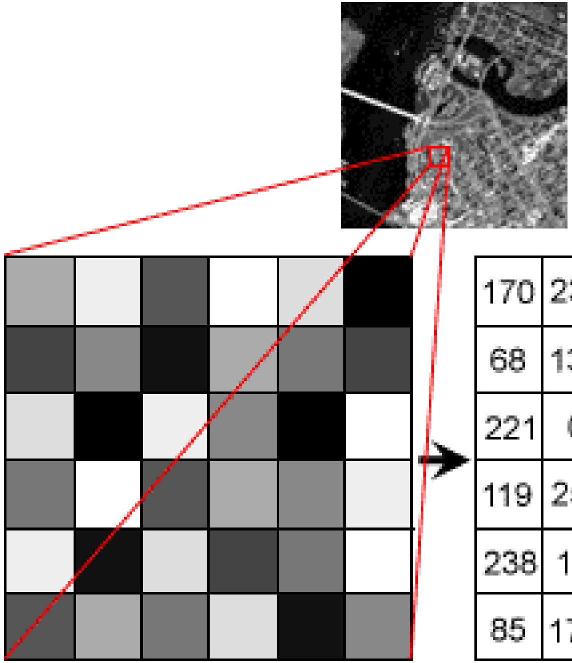


- Deep learning, via artificial neural network models, can learn to extract useful information from raw input directly
- The catch is a lot of data and supervision is needed

Naïve representations

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Image from hackermoon.com

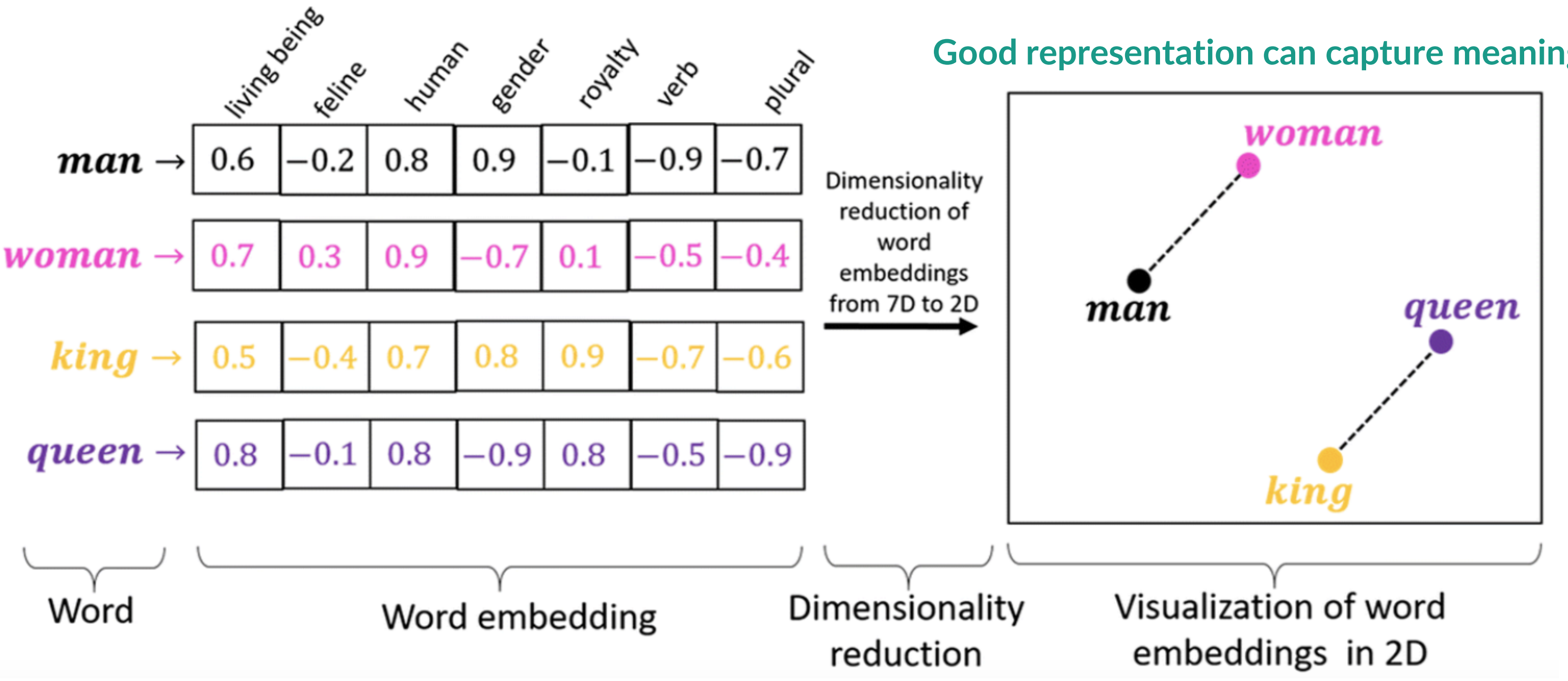


170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

Image from naushardsblog.wordpress.com

Meaningful word embeddings

Good representation can capture meaning!



Explainability and AI project design

AI (silently) makes mistakes and biases

But can you spot them?

Alkaissi, H. et al. Cureus 15:e35179 (2023)

> Late onset Pompe disease (LOPD) is a rare genetic disorder characterized by the deficiency of acid alpha-glucosidase (GAA), an enzyme responsible for the breakdown of glycogen in lysosomes. The accumulation of glycogen in various tissues leads to progressive muscle weakness, primarily affecting the skeletal and respiratory muscles. However, recent studies have also reported liver involvement in LOPD, which is thought to occur as a result of the accumulation of glycogen in liver cells.

- There was no prior publication about liver involvement with LOPD
- However, the authors of this paper have an unpublished manuscript showing a link between liver disease and LOPD
    - *Did ChatGPT just synthesized new knowledge? Or simply hallucinated?*

Huge gap between development and actual use

Healthcare, Law, Regulation, and Policy, Machine Learning

“Flying in the Dark”: Hospital AI Tools Aren’t Well Documented

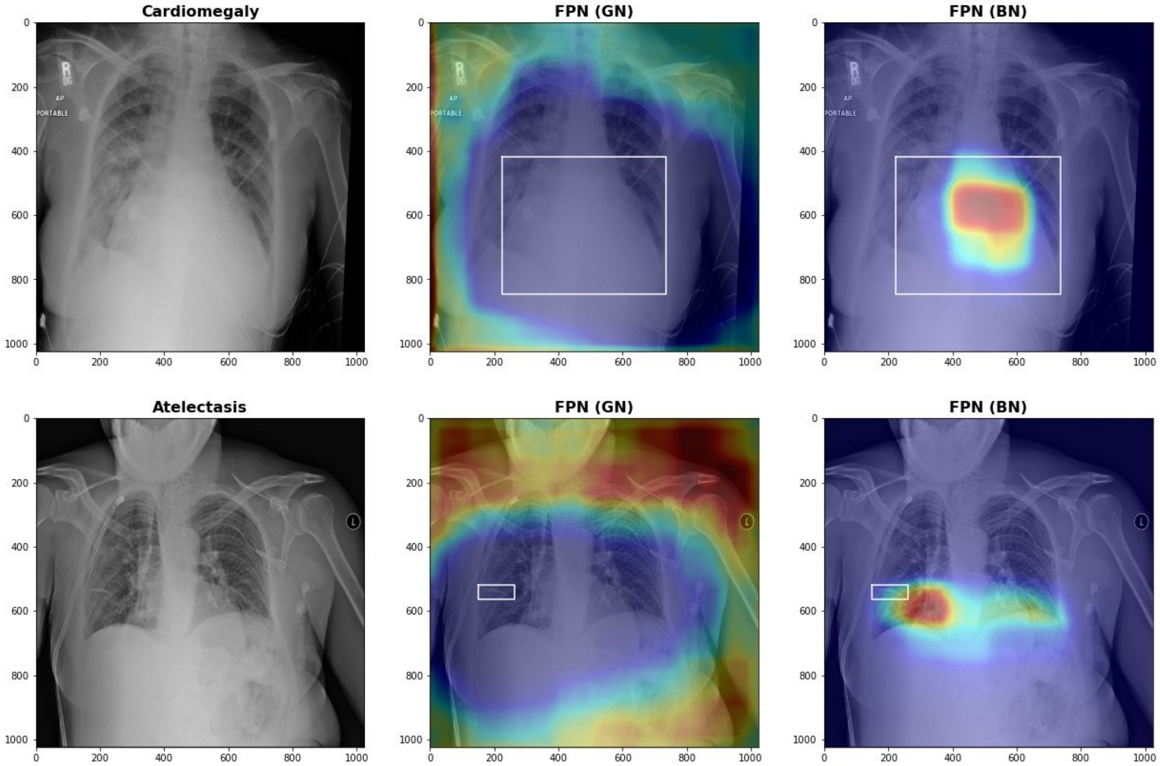
MODEL REPORTING GUIDELINES	EPIC MODEL BRIEFS											
	Deteriorati on Index	Early Detection of Sepsis	Risk of Unplanned Readmission	Risk of Patient No-Show	Pediatric Risk of Hospital Admission or ED Visit	Risk of Hospital Admission or ED Visit	Inpatient Risk of Falls	Projected Block Utilization	Remaining Length of Stay	Risk of Admission of Heart Failure	Risk of Hospital Admission or ED Visit for Asthma	Risk of Hypertension
TRIPOD	63%	63%	61%	48%	42%	61%	47%	36%	55%	48%	44%	51%
CONSORT-AI	63%	43%	63%	60%	33%	67%	53%	47%	47%	49%	42%	51%
SPIRIT-AI	61%	55%	54%	54%	38%	61%	44%	49%	51%	41%	39%	46%
Trust and Value	46%	33%	39%	50%	29%	42%	38%	46%	46%	25%	33%	46%
ML Test Score	27%	15%	33%	24%	9%	33%	15%	6%	18%	12%	9%	15%

Evaluation of sepsis diagnosis AI

Results We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclusion criteria, of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level area under the receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM identified 183 of 2552 patients with sepsis (7%) who did not receive timely administration of antibiotics, highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating a large burden of alert fatigue.

- AUC of 0.63 in practice
- Missed 67% of sepsis

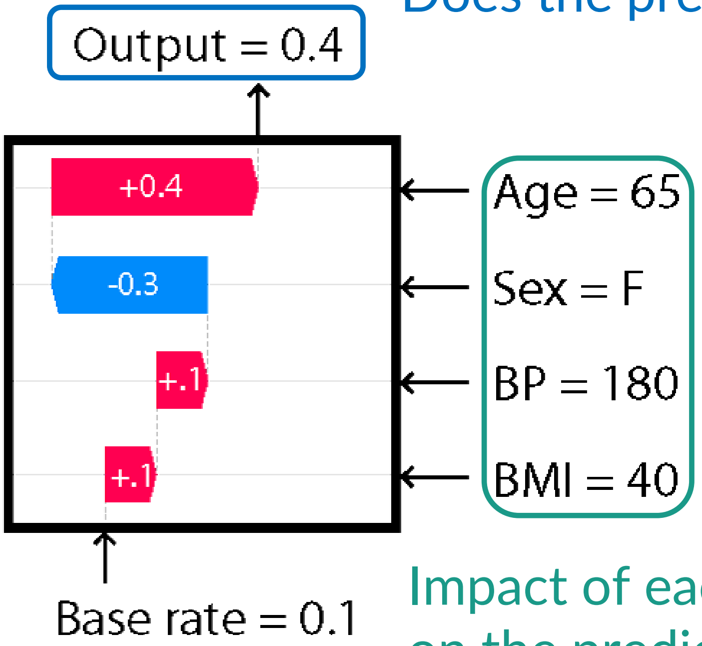
Correct prediction is not enough



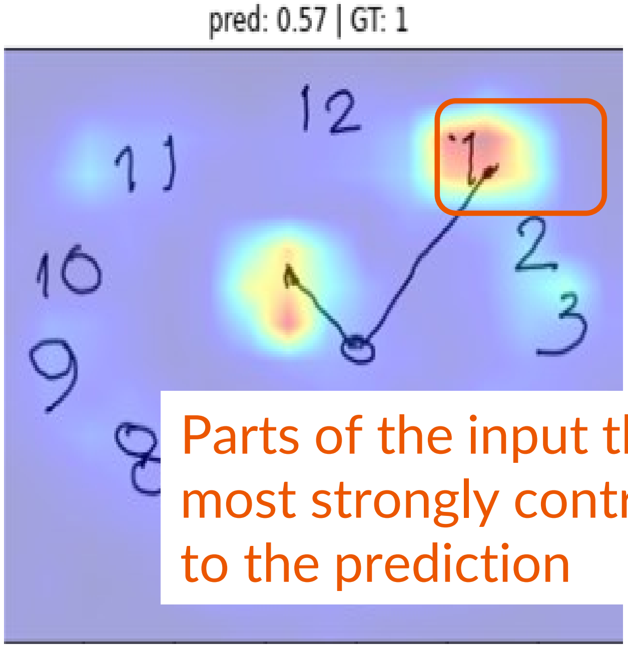
- Two models with the **same classification performance**
- Both images were **correctly classified**
- But the **explanations** complete differ

Explainability

Does the predicted confidence match your expectation?



Impact of each input on the prediction



Parts of the input that most strongly contribute to the prediction

Sources of unexpected behaviors

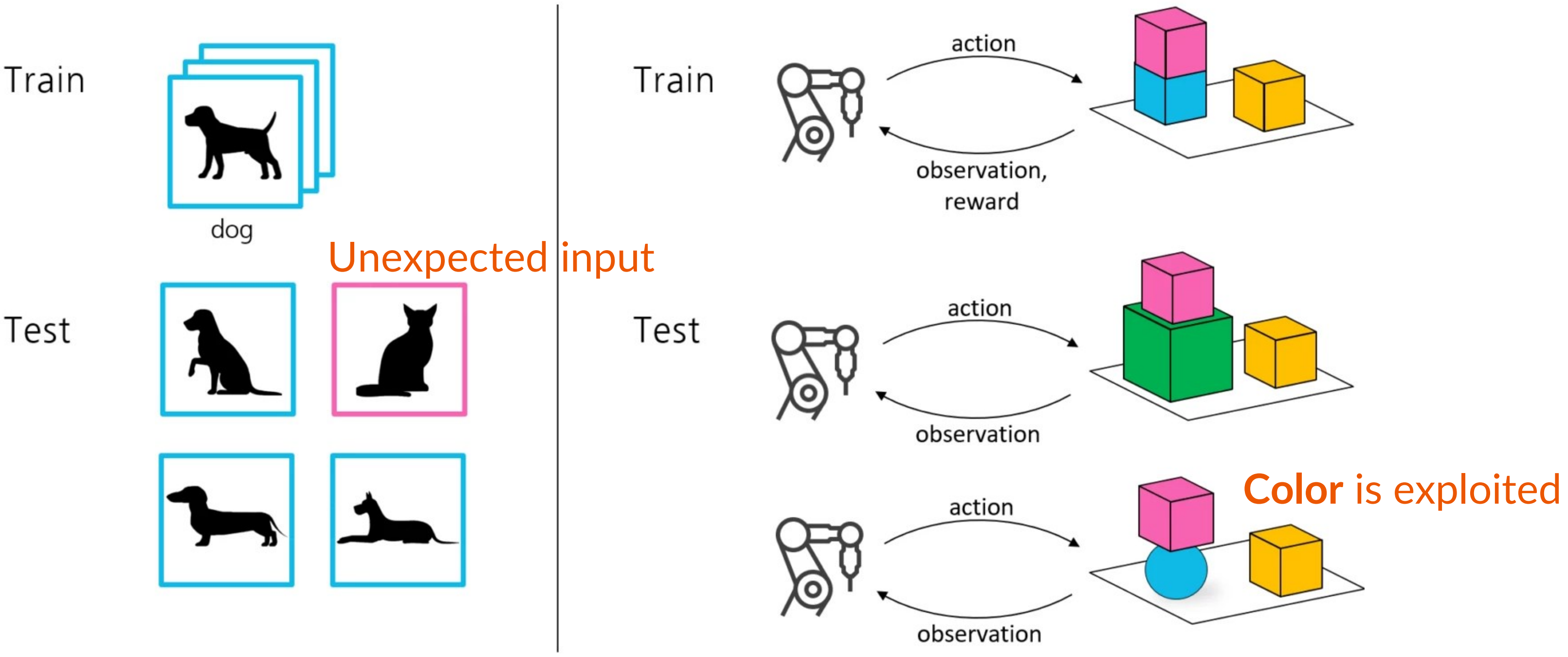


Image from https://safe-intelligence.fraunhofer.de/en/articles/out-of-distribution-detection-for-reinforcement-learning

Summary

- This course gives you the foundation to advance yourself
- Communicate with me and TA
- Make the most out of this course and internship experience
- Have fun!

Any questions?

See you on February 1st