DataSci 207– Applied Machine Learning

Cornelia Paulik, PhD

School of Information

UC Berkeley

Introduction and Framing

About me

Current:

- Assistant Professor of Practice in the School of Information at UC Berkley
- Visiting Researcher, Global Policy Lab at Stanford

Something fun about me:

- I love hiking (strenuous trails are my fav)
- Infant boy + two dogs keep our life very busy

About you

- Undergraduate major
- Current job (if any)
- Something fun about you

Course websites

- The Async material is in bCourses: https://bcourses.berkeley.edu
- The Live Session material is on my website: https://corneliailin.github.io/datasci_w207_summer2024/

Live sessions organization

- Each session is 90 minutes
- 30 minutes Q&A related to the topic of the week
- 60 minutes Code demonstration / Breakout room exercises

Today’s learning objectives

- General concepts of Machine Learning (ML)
- Roadmap for building ML systems (1. Introduction_and_framing.ipynb)

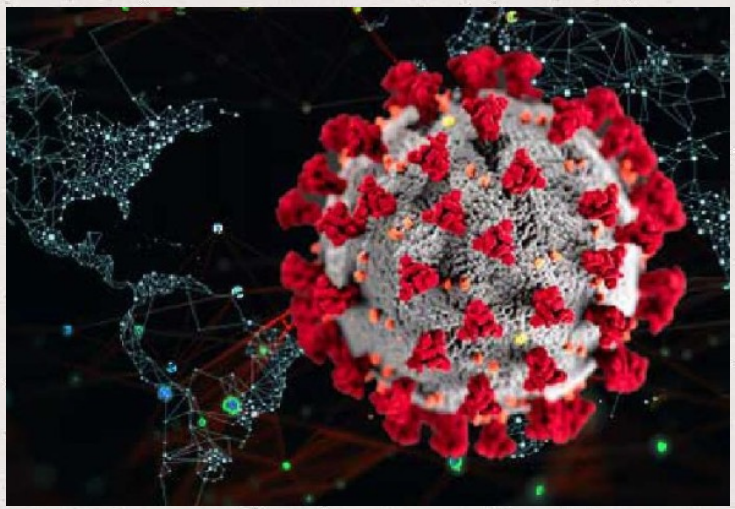
General concepts of ML

Q1: Can you make **predictions** about the **future** using **ML**? Provide an example.

General concepts of ML

Q1: Can you make **predictions** about the **future** using **ML**? Provide an example.

Yes!



→ Example: predict number of COVID-19 cases and deaths

General concepts of ML

Q2: Name and explain the 3 types of ML supervision

General concepts of ML

Q2: Name and explain the 3 types of ML supervision

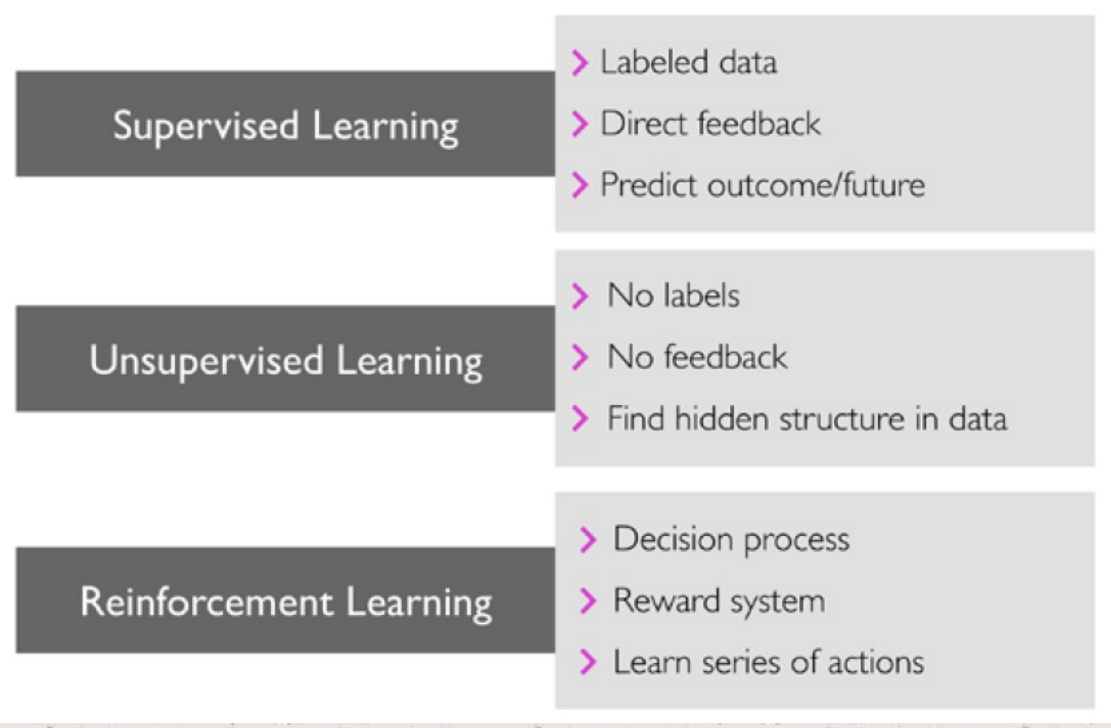


Image source: S. Raschka and V. Mirjalili, Python Machine Learning

General concepts of ML

Q3: Name and describe the **2 types of supervised ML models**

General concepts of ML

Q3: Name and describe the **2 types of supervised ML models**

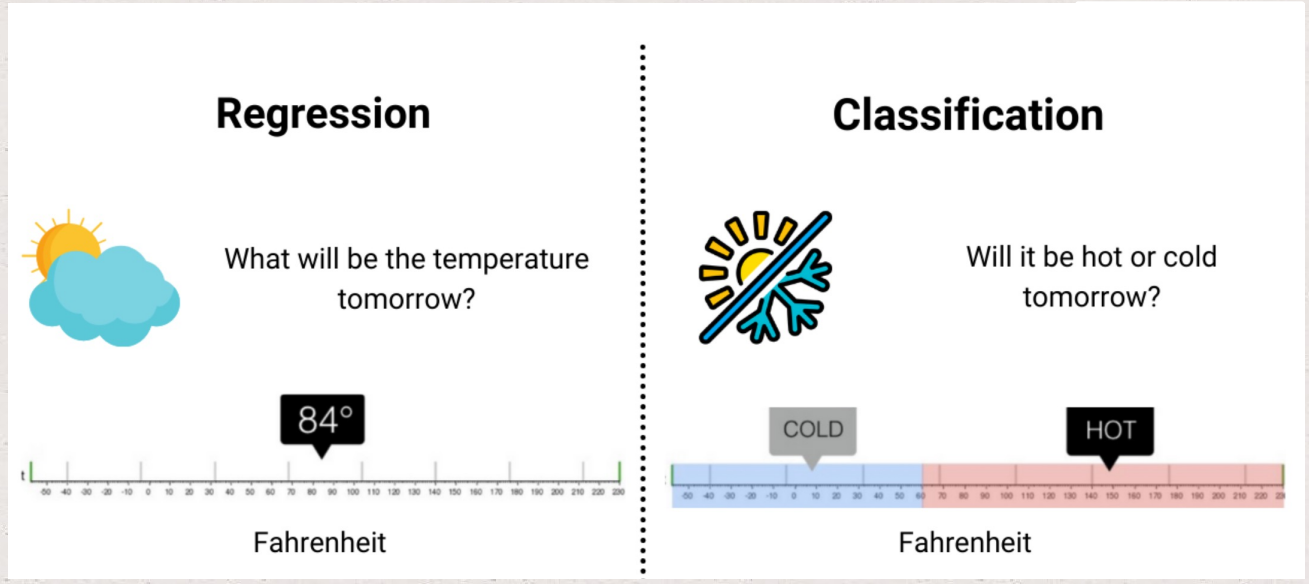


Image source: https://www.enjoyalgorithms.com/

General concepts of ML

Q4: True or False? Clustering is a technique used for structuring information and deriving meaningful relationships from data.

General concepts of ML

Q4: True or False? Clustering is a technique used for structuring information and deriving meaningful relationships from data.

True, it's an unsupervised learning technique used to find subgroups in data.

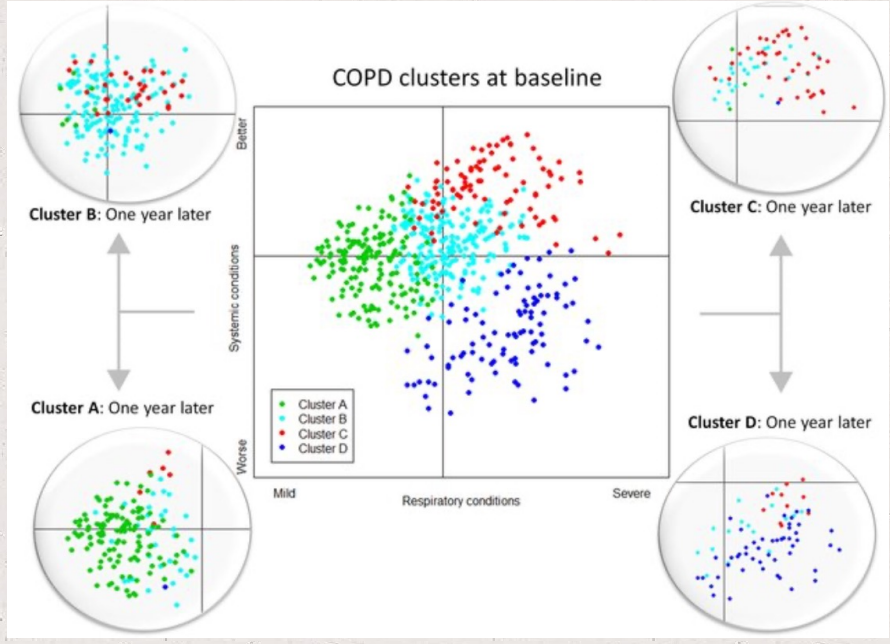


Image source: https://www.researchgate.net/publication/307969853_Chronic_Obstructive_Pulmonary_Disease_Subtypes_Transitions_over_Time/figures

→ Example: discover patient groups based on their diagnosis history in order to develop distinct treatment plans.

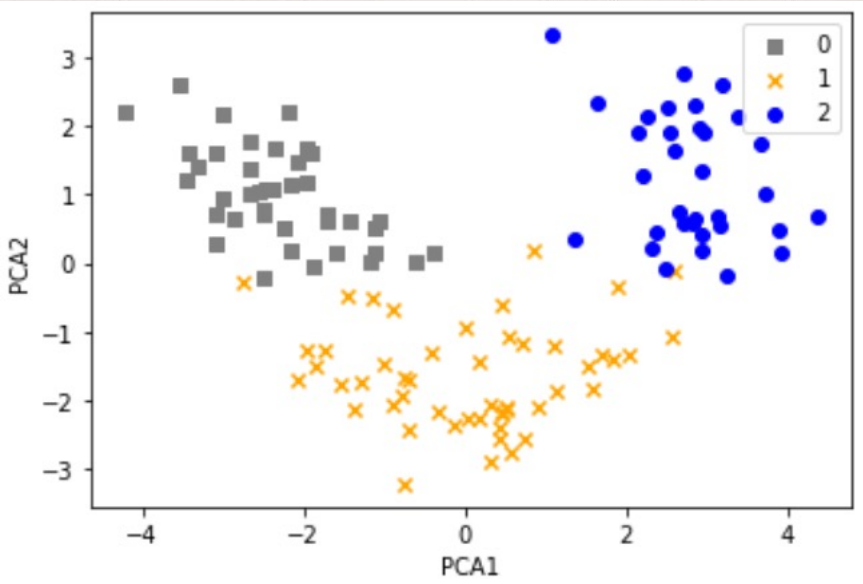
General concepts of ML

Q5: What is Principal Component Analysis (PCA)?

General concepts of ML

Q5: What is Principal Component Analysis (PCA)?

PCA is an unsupervised learning technique, useful when working with data of high dimensionality, for processing and/or visualizing data.



→ Example: Common to combine PCA and clustering!

General concepts of ML

Q6: What is a typical **workflow** for using **ML** in predictive modeling?

General concepts of ML

Q6: What is a typical **workflow** for using **ML** in predictive modeling?

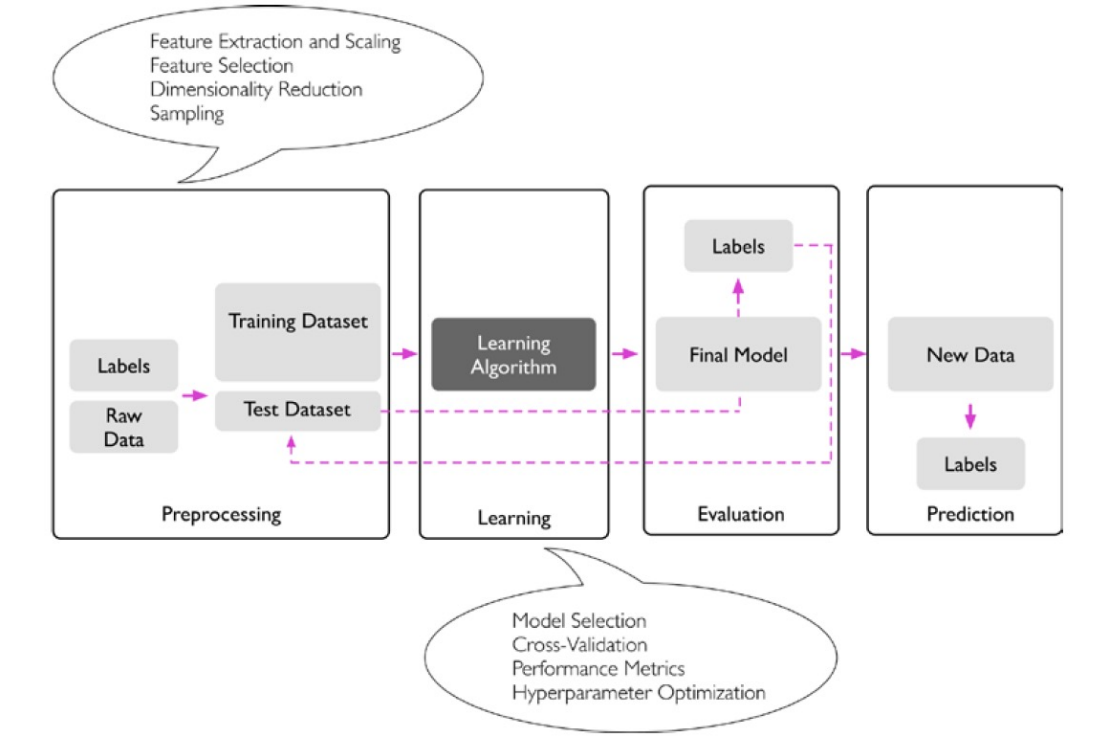


Image source: S. Raschka and V. Mirjalili, Python Machine Learning

→ Example: Introduction_and_framing.ipynb (CI)

General concepts of ML

Q7: Why do we need a **train-test** split? How do you **evaluate prediction**?

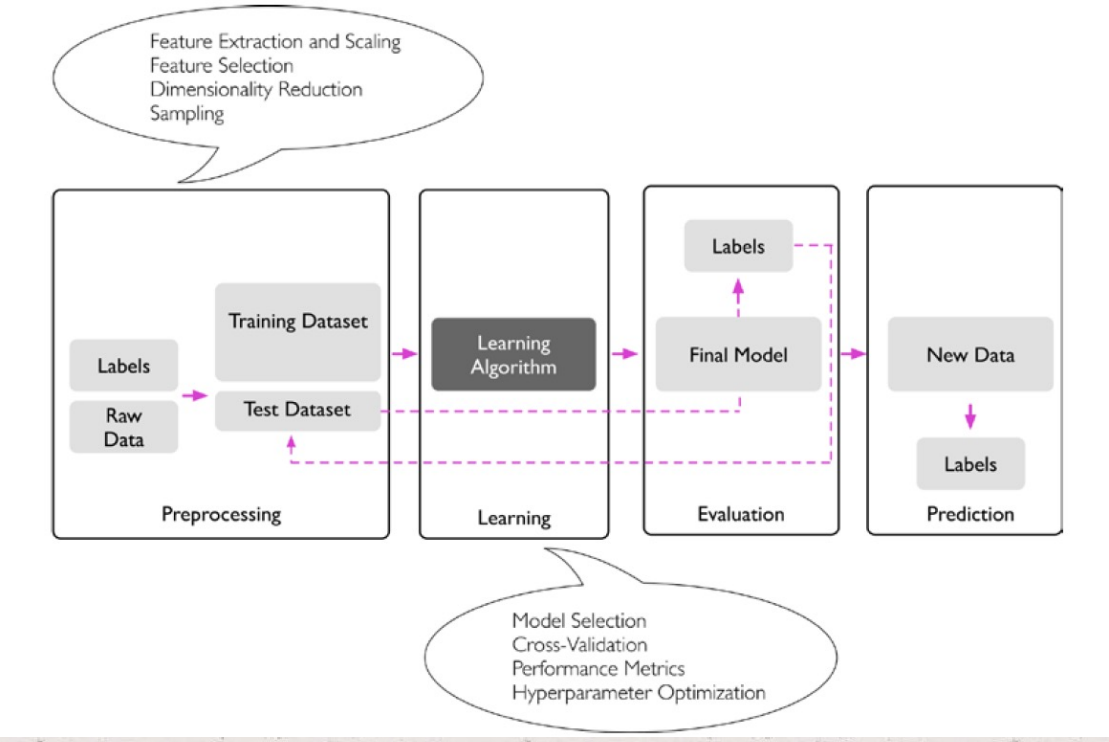


Image source: S. Raschka and V. Mirjalili, Python Machine Learning

What is meant by **generalization**?