2025

Introduction & Course Overview

CPSC 4300/6300: Applied Data Science

Dr. Aaron J. Masino
Associate Professor, School of Computing

College of ENGINEERING, COMPUTING AND APPLIED SCIENCES

Instructors

Course Instructor

Dr. Aaron Masino

Email: amasino@clemson.edu

Teaching Assistant

Luyi Li

Email: luyil@g.clemson.edu

Office Hours: Tuesday’s 2:00-4:00, 102 Barre Hall

A little about me

- Former roles include:
  - VP of Clinical AI at a healthcare startup
  - Assistant Professor of Biomedical Informatics at UPenn
  - Director of Data Science at Children's Hospital of Philadelphia Research Institute
  - Senior Scientist in adaptive optics industry roles
  - Air Force officer – worked on the Delta space launch program
- More about my Clemson lab here

Now it is your turn

- Please complete the “Initial Student Survey” on Canvas in the Quizzes section.

What is this course about?

The material of the course will integrate the five key facets of an investigation using data:

1. data collection; data wrangling, cleaning, and sampling to get a suitable data set
2. data management; accessing data quickly and reliably
3. exploratory data analysis; generating hypotheses and building intuition
4. prediction or statistical learning
5. communication; summarizing results through visualization, stories, and interpretable summaries.

Prerequisites

- **4000-Level Credits**: CPSC 2120, STAT 2300 or STAT 3090
- **6000-Level Credits**: At least one college-level introductory class in statistics

Prerequisites - Computing

- Understand and work with different types of data structures
- Ability to quickly learn a new programming language (Python)
- Ability to use Git and GitHub for version control and software development.

Syllabus Review

What is Data Science?

- “Data Science is an interdisciplinary field about ***processes and systems*** to ***extract knowledge or insights from data*** in various forms, either structured or unstructured […]” (Wikipedia)
- “At it’s core, data science involves using ***automated methods*** to analyze ***massive amounts of data*** and to extract knowledge from them.” (NYU Data Science)
- A data scientist is someone “[…] who is better at statistics than any software engineer and better at software engineering than any statistician.” (John Wills, Cloudera)

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

What is the scientific/business goal?

What would you do if you had **all** of the data?

What do you want to predict or estimate?

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

How were the data sampled?

Which data are relevant?

Are there privacy issues?

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

Plot the data.

Are there anomalies or egregious issues?

Are there patterns?

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

Build a model.

Fit the model.

Validate the model.

The Data Science Process

- Ask an interesting question
- Get the Data
- Explore the Data
- Model the Data
- Communicate/Visualize the Results

What did we learn?

Do the results make sense?

Can we effectively tell a story?

Prerequisites - Statistics

- Understanding the difference between quantitative and qualitative features
- Understand the use of linear regression techniques to analyze the association between two qualitative variables
- Understand the concept of sampling distribution and how it applies to statistical inference
- Calculate and interpret confidence intervals
- Understand the logic of hypothesis testing

Questions?

Machine learning overview

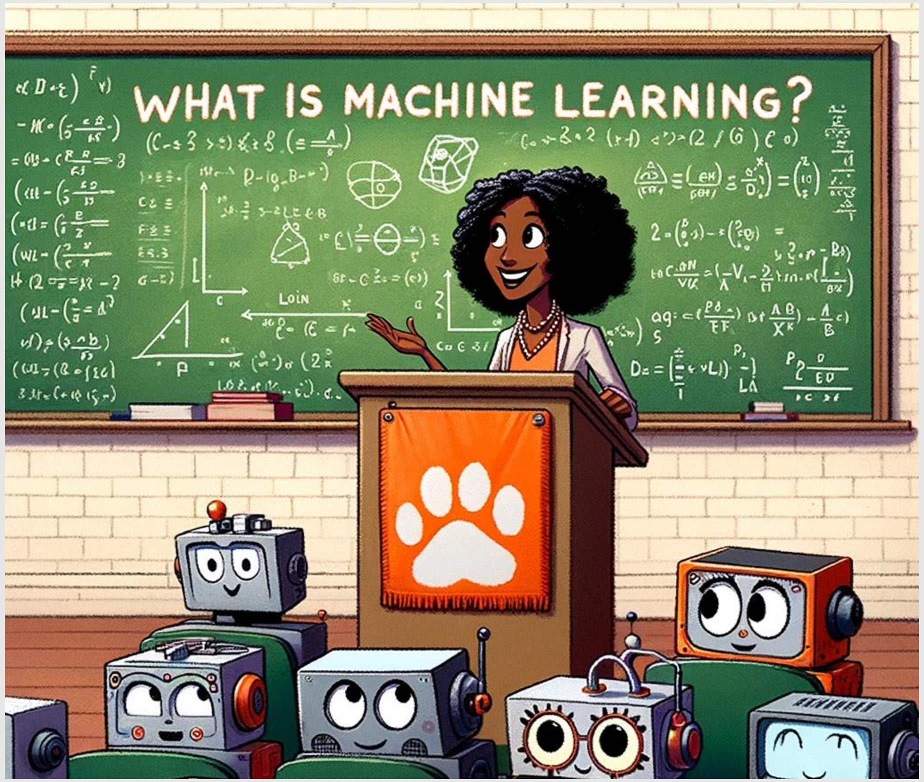
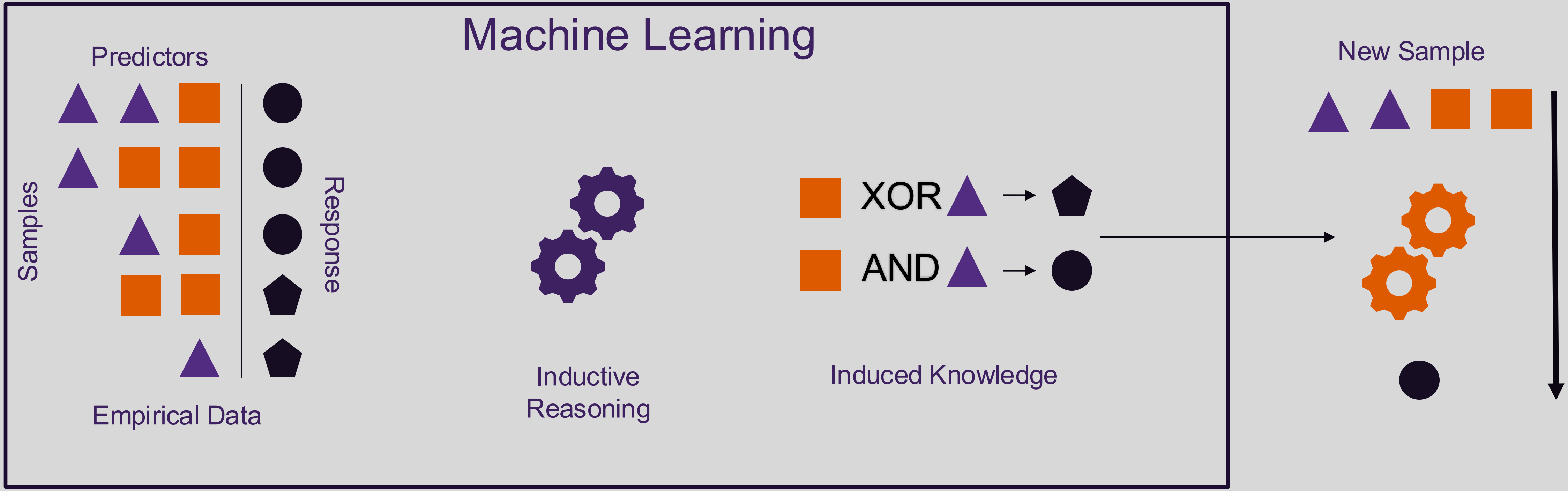


Image generated by openai.com DALL-E

Definitions of machine learning

Any computing method that implements inductive reasoning over empirical data to create knowledge, however represented, that can be used to perform one or more tasks.



Definitions of machine learning

Following *ISL, we can define machine learning mathematically as any method that estimates an **unknown** function f from empirical observations of variables $X = (X_1, X_2, \cdots, X_p)$ and, optionally, Y, where

$$Y = f(X) + \varepsilon$$

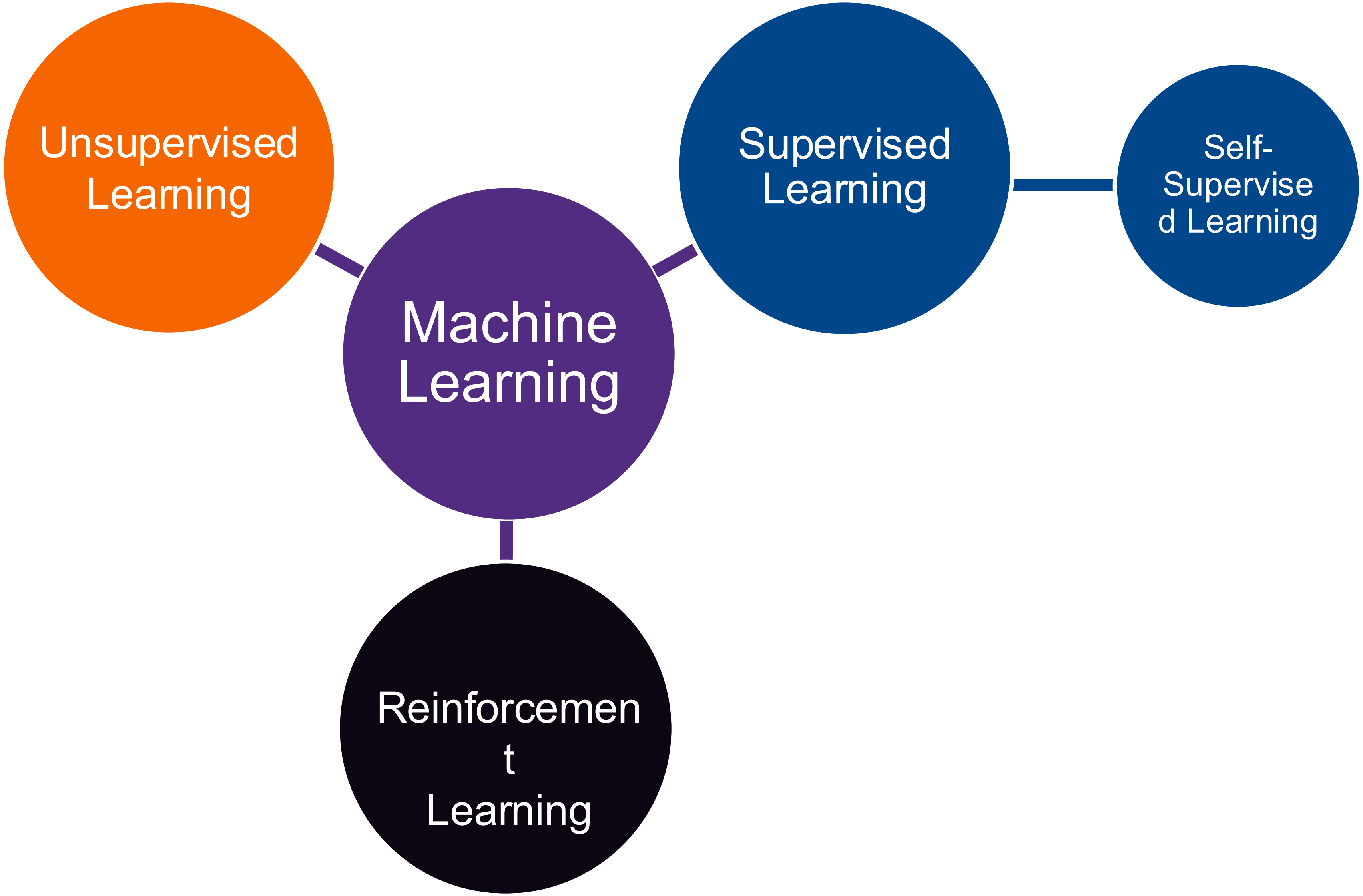
*Jame et al. **I**ntroduction to **S**tatistical **L**earning with Applications in Python. Springer 2023

Motivation

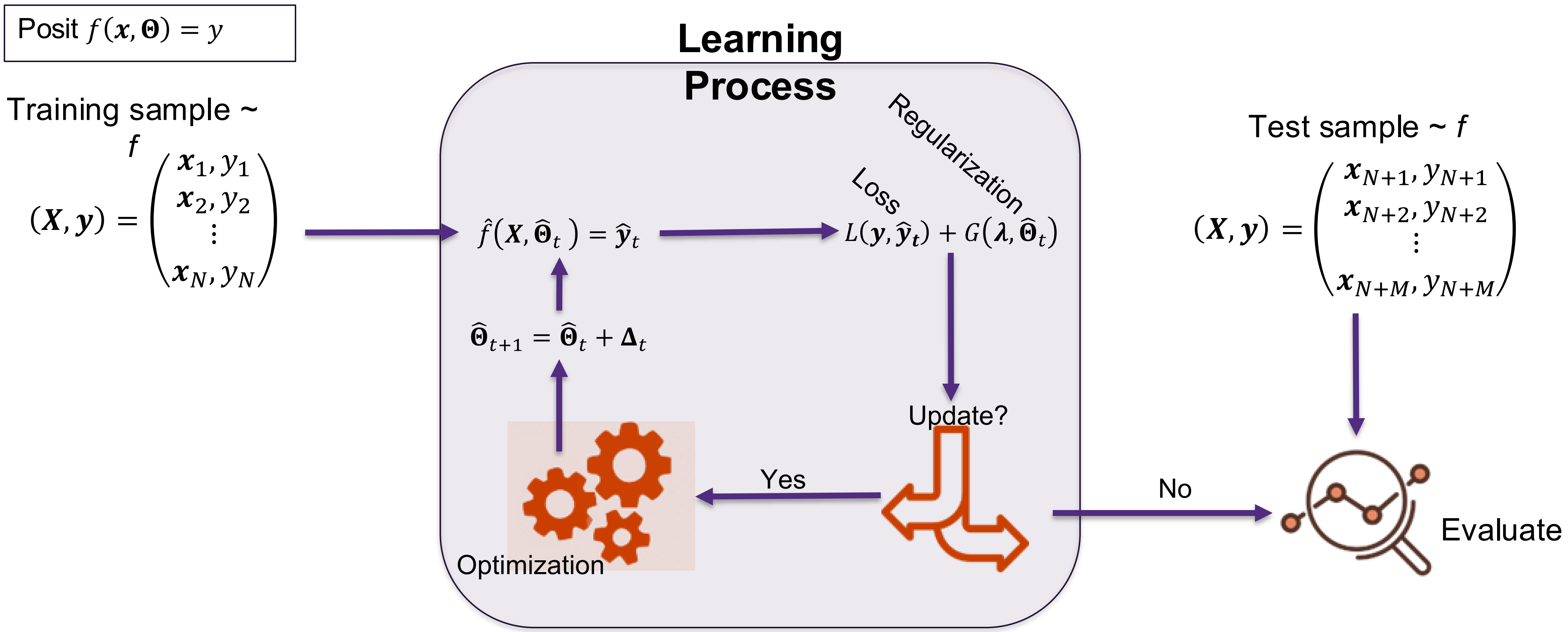
Why do we want to estimate f ?

- It may be easy to obtain X but not Y and we don't know f
    - Referred to as *prediction* in ISL. Somewhat confusing as prediction usually implies a future event.
    - I will refer to this as estimation of Y unless it is an estimate of a future event, in which case I will refer to this as forecasting
- We may want to *infer* information about the relationship between X and Y or about structure in X

Learning Paradigms



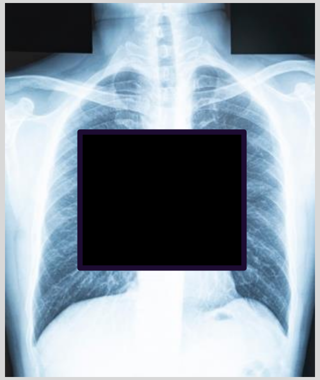
Supervised Learning



Supervised

$\xrightarrow{f}$ Healthy

Self Supervised



$\xrightarrow{f}$

Self Supervised Learning

Supervised

$\boxed{\text{Posit } f(\boldsymbol{x}, \boldsymbol{\Theta}) = y}$

Empirical sample ~ *f*

$$(\boldsymbol{X}, \boldsymbol{y}) = \begin{pmatrix} \boldsymbol{x}_1, y_1 \\ \boldsymbol{x}_2, y_2 \\ \vdots \\ \boldsymbol{x}_N, y_N \end{pmatrix}$$

Self Supervised

$\boxed{\text{Posit } f(\boldsymbol{x}/\boldsymbol{x}_i, \boldsymbol{\Theta}) = \boldsymbol{x}}$

Empirical sample ~ *f*

$$(\boldsymbol{X}/\boldsymbol{xi}, \boldsymbol{X}) = \begin{pmatrix} \boldsymbol{x}_1/x_{1,i}, \boldsymbol{x}_1 \\ \boldsymbol{x}_2/x_{2,i}, \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N/x_{N,i}, \boldsymbol{x}_N \end{pmatrix}$$

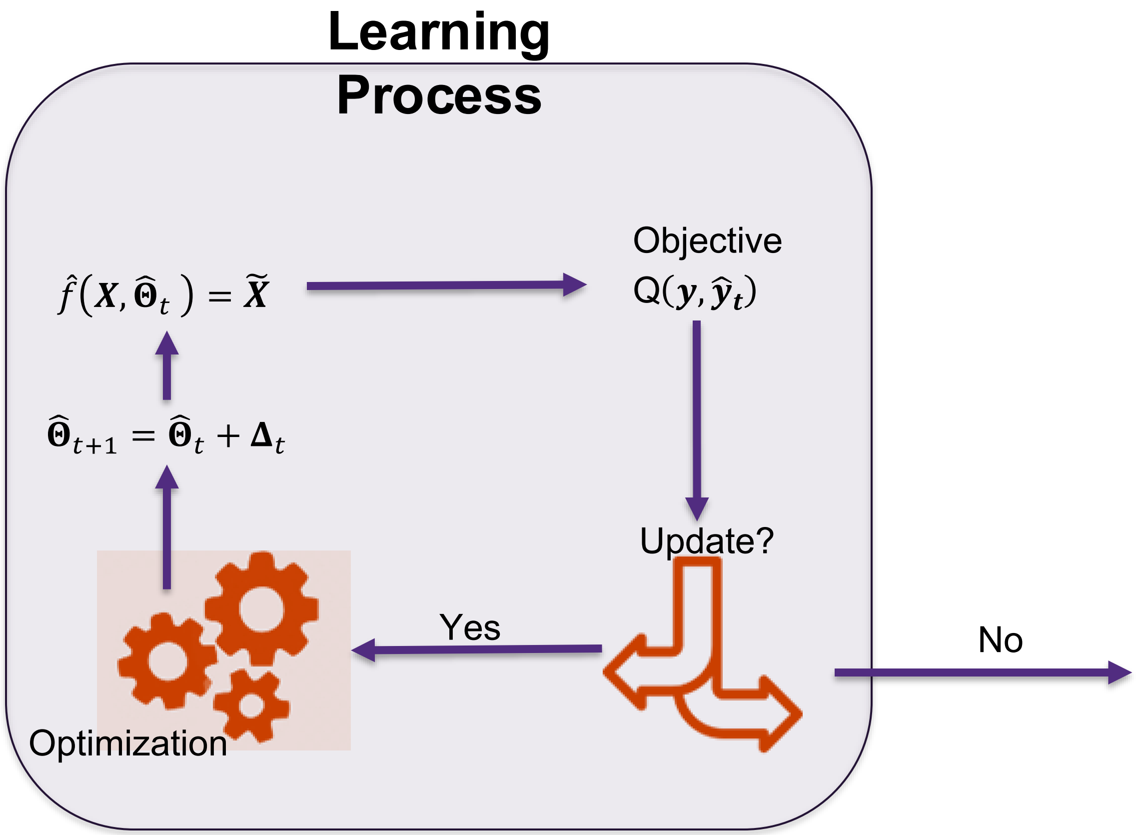
Supervised: The output, y, is distinct from the input and usually human generated

Self supervised: The output, $\boldsymbol{x}$, is a reproduction of the original input given on part of the input or a compressed form of the input

Unsupervised Learning

Posit $f(\boldsymbol{X}, \boldsymbol{\Theta}) = \widetilde{\boldsymbol{X}}$ subject to constraints $\boldsymbol{C}$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix}$$



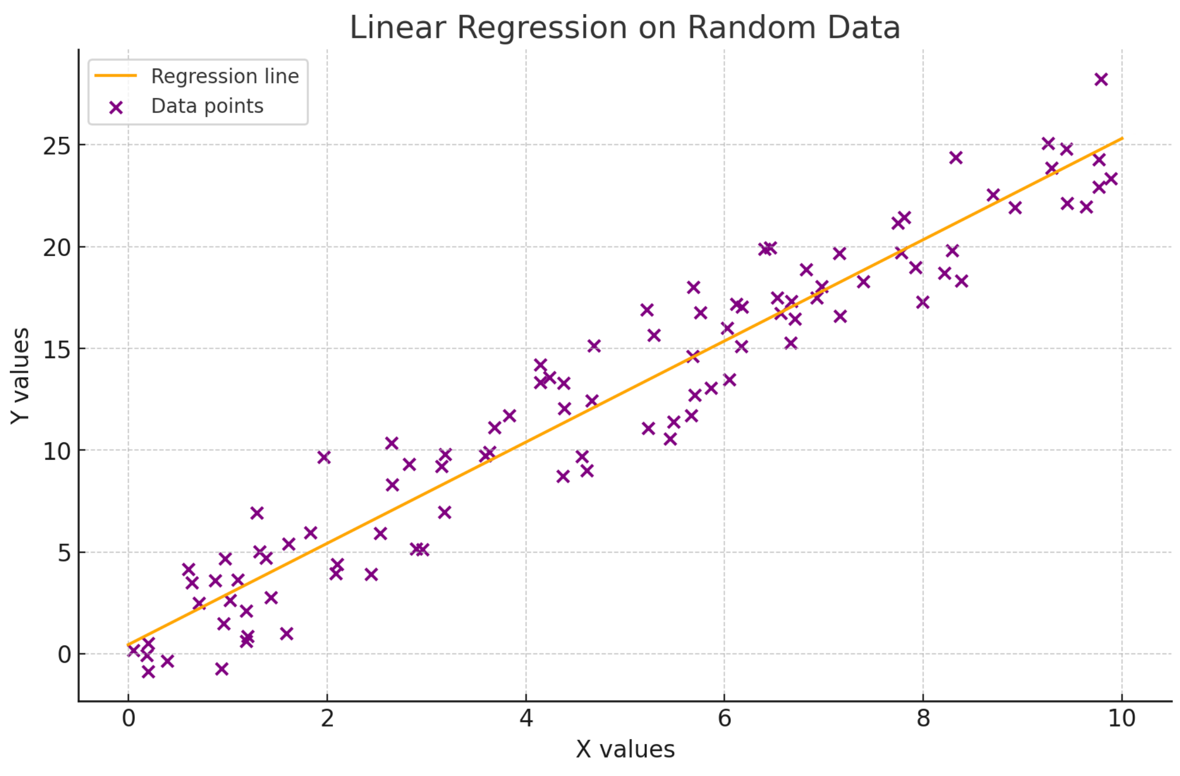
$$\widetilde{\boldsymbol{X}} = \left[f \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix} \right]$$

Some of the task domains addressable with machine learning

- Regression
- Classification
- Clustering
- Dimensionality reduction / representation learning
- Sample Generation

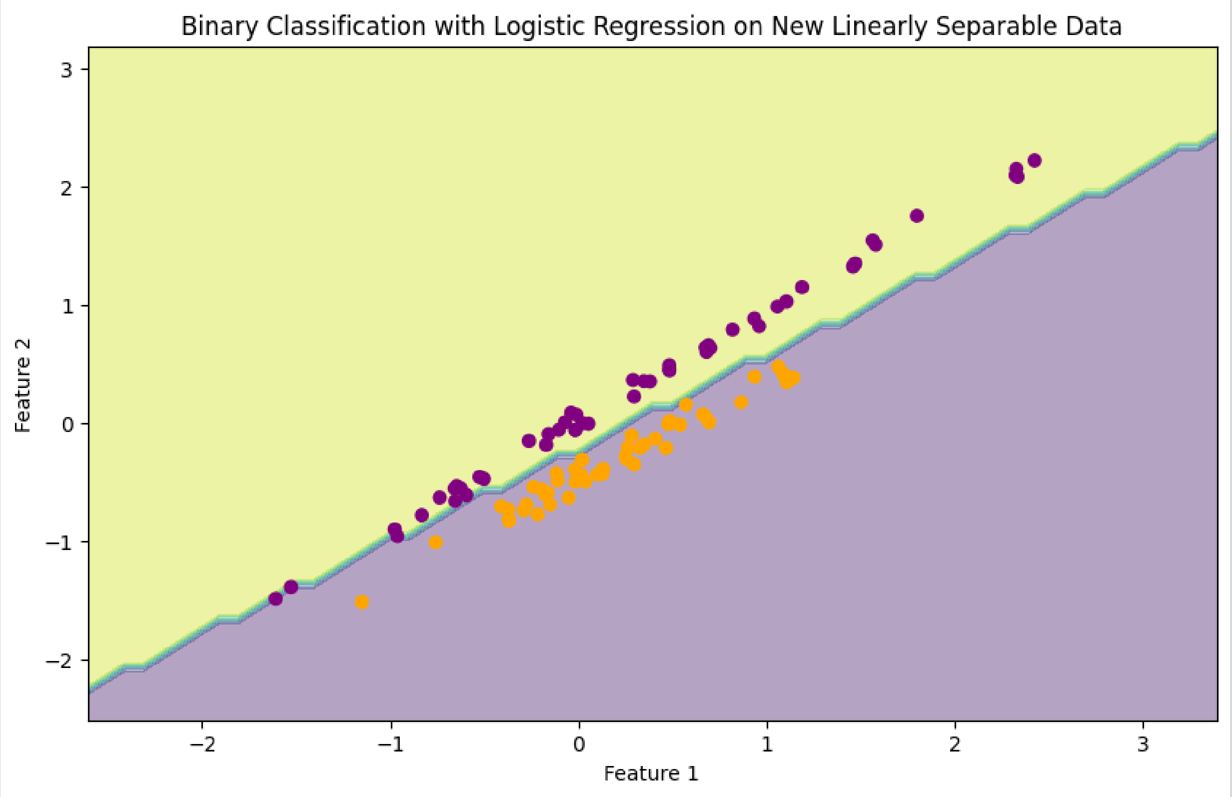
Regression

- Construct a function, f, to estimate the quantitative value, Y, from input X
- Some methods covered in this course
  - Linear regression
  - Tree-based regression
  - LSTMs



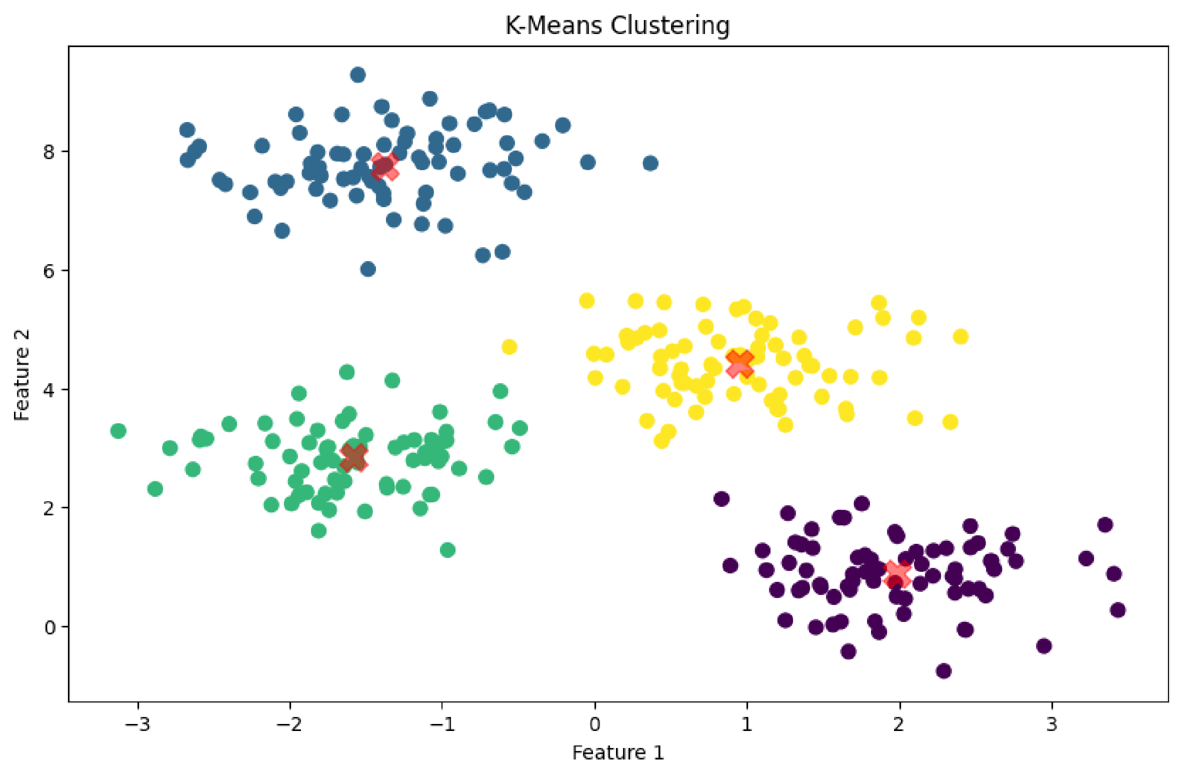
Classification

- Construct a function, f, to estimate the qualitative (class membership) value, Y, from input X
- Some methods covered in this course
    - Logistic regression
    - Tree-based methods
    - CNNs
    - Transformers



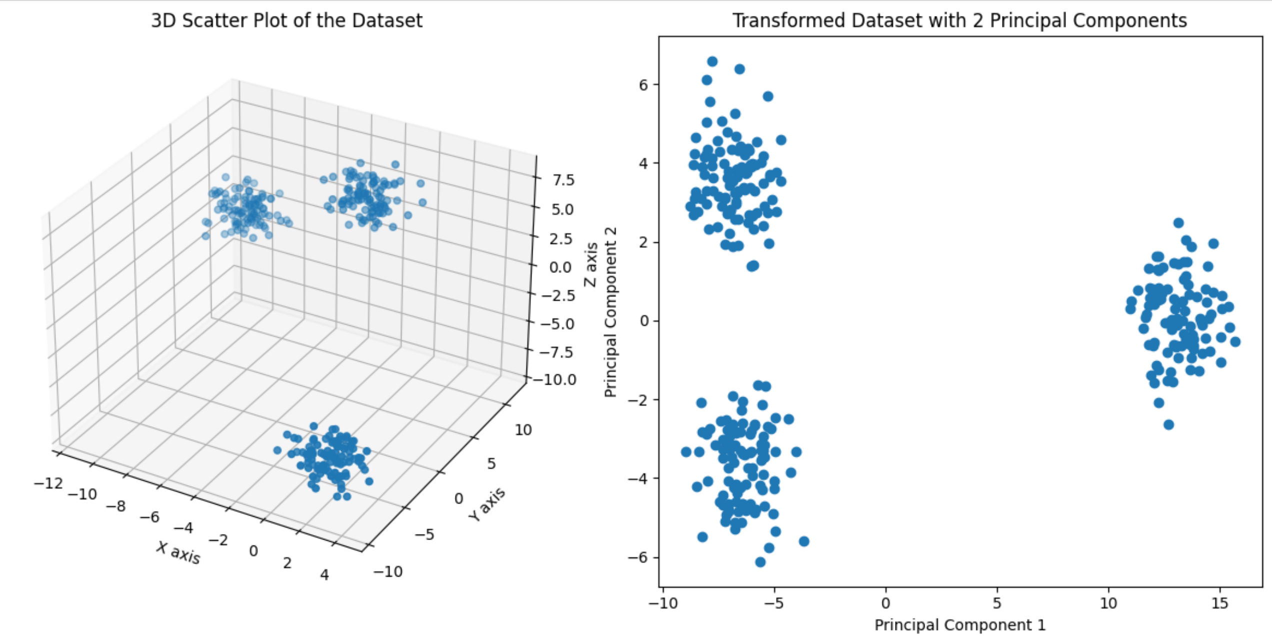
Clustering

- Construct a function, f, to identify structure (e.g., clusters) among samples with input X
- Methods covered in this course
  - K-means
  - Hierarchical



Dimensionality reduction

- Construct a function, f, to transform input X to a lower dimensional representations
- Methods covered in this course
  - Principal components analysis



Generation

- Construct a function, f, that samples values, X, from the input space with approximately the same conditional distributions as the empirical data
- Methods covered in this course
    - Generative language models

You

Are you a generative model?

ChatGPT

Yes, I am a generative model. Specifically, I'm based on GPT-4 (Generative Pre-trained Transformer 4), which is a type of large language model developed by OpenAI. As a generative model, I have several key characteristics:

1. **Training on Diverse Data:** I have been trained on a vast array of text data from books, websites, and other textual sources. This training allows me to understand and generate human-like text across a wide range of topics and styles.
2. **Language Understanding and Generation:** My primary function is to understand natural language input and generate coherent, contextually relevant, and often informative or creative textual output. This includes answering questions, providing explanations, engaging in conversation, and even creating content like stories or essays.

Why so many methods?

- ***No Free Lunch Theorem** (NFLT): under the constraint of no prior knowledge, ALL optimization methods perform the same when averaged across all possible objectives
- Implication: *ALL supervised (including self) machine learning algorithms perform the same* when averaged across all problems
  - A priori knowledge of the problem can guide our selection
  - Suggests the need to try multiple algorithms on a given problem

*Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." IEEE transactions on evolutionary computation 1.1 (1997): 67-82.