STATISTICS & ML WITH R

Mapping causal & predictive approaches

2024

M. Chiara Mimmi & Luisa M. Mimmi

DAY 4 – LECTURE OUTLINE

- Mapping causal & predictive approaches

From Correlation/Association to Prediction/Causation → *experimental studies*

- So far, we worked on "**OBSERVATIONAL STUDIES**" (i.e. data where you have *no controlled assignment of the treatment*) measuring variables of interest
  - We may find CORRELATION OR ASSOCIATION, but it DOES NOT IMPLY CAUSATION!
  - WHY? there can be "hidden variables" that affect the relationship between the explanatory variable and the response variable

- Instead, **"EXPERIMENTAL STUDIES"** help us studying CAUSATION in that they are "*designed to provoke a response*"
  - here, researchers ***assign the treatment*** to an ***experimental unit*** (or ***subject***) and observing its ***effect***
  - these studies follow some ***design principles*** to provide robust evidence for causation

A conceptual framework to understand different types of statistical modeling (part 1/2)

1. **ASSOCIATION/CORRELATION** → observational studies
   - aimed at **summarizing or representing the data structure**, *without* an underlying causal theory
   - may help **form hypotheses** for explanatory and predictive modeling
2. **CAUSAL EXPLANATION** → experimental studies
   - aimed at **testing "explanatory connection"** between treatment and outcome variables
   - prevalent in "**causal theory-heavy**" fields (like: economics, psychology, environmental science, etc.)

- **Note**:
  - ✓ The **same modeling approach** (e.g. fitting a regression model) can be used for **different goals**
  - ✓ While they shouldn't be confused, **explanatory power** and **predictive accuracy** are complementary goals: e.g. in bioinformatics (which has little theory and abundance of data), predictive models are pivotal in generating avenues for causal theory.

3. **EMPIRICAL PREDICTION** → algorithmic machine learning and data-mining modeling

A conceptual framework to understand different types of statistical **modeling** (part 2/2)

1. **ASSOCIATION/CORRELATION** → observational studies
2. **CAUSAL EXPLANATION** → experimental studies
3. **EMPIRICAL PREDICTION** → algorithmic machine learning and data-mining modeling
   - aimed at **predicting new or future observations** (without necessarily explaining how)
   - relies on **big data**
   - prevalent in fields like natural language processing, bioinformatics, etc.. In epidemiology, there is more of a mix causal explanation & empirical prediction

- **Notes**:
  - ✓ "Prediction" does not necessarily refer to future events, but rather to *future* datasets that were previously unseen to the algorithm

A framework for CAUSAL ANALYSIS

The conceptual framework for causal analysis

- We need 2 key elements:
  1. HOW DO WE DEFINE THE EFFECT OF AN INTERVENTION?
  2. WHAT ARE THE MOST IMPORTANT FEATURES OF THE DATA TO UNCOVER THAT EFFECT?

- And we must define:
  - Intervention (business, policy) or Treatment (medicine) = decisions and actions that change the behaviors or situation of people/firms/other subjects
  - Subjects = those that may be affected (at lest in principle), in fact are
    - TREATED subjects
    - UNTREATED subjects
  - Outcome = variable(s) that may be affected by the intervention

Defining potential outcomes at the *subject (experimental unit)* level

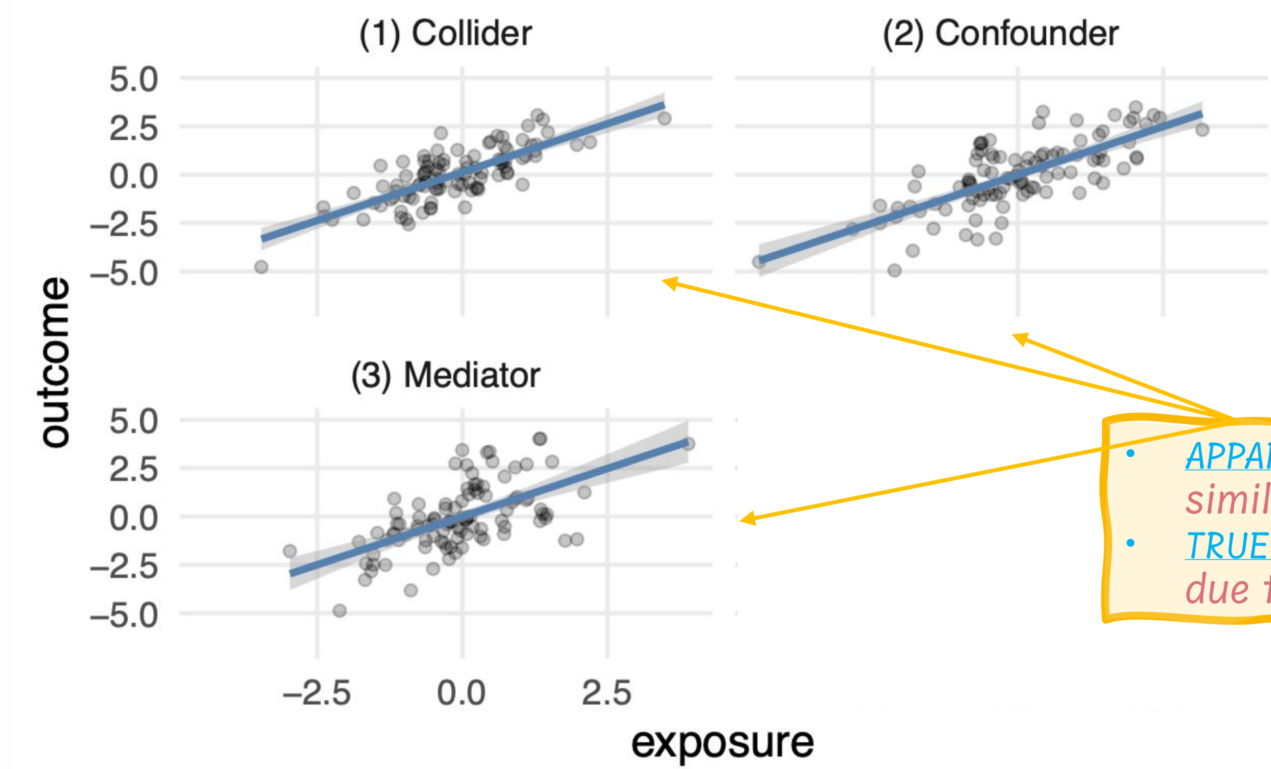
- **ITE = Individual Treatment Effect** (*) = potential outcome for subject if treated – outcome if untreated
    - (*) ITE is never observable!
- **ATE = Average Treatment Effect** = average of ITE differences across subjects
    - (*) The Avg of the differences is – the difference of Averages!
    - Important to have a well-defined group or population
    - ATE can hide different distributions of ITEs (e.g., positives and negatives that cancel each outer out)
- **ATET = Average Treatment effect on the Treated =** average treatment effect across all subjects that end up TREATED

Causal maps

A tool that can be helpful in guiding out subsequent analysis

Typical challenges in estimating causal effects: visual intuition

- Consider 3 distinct datasets: while their statistical summaries and visualizations are very similar, the **true causal effect differs!**
- **Deciding the** correct model requires knowledge of the data-generating mechanism (i.e. the random assignment to exposure/not exposure in experiments)



- *APPARENT relationship similar*
- *TRUE relationship different due to ≠ causal nexus*

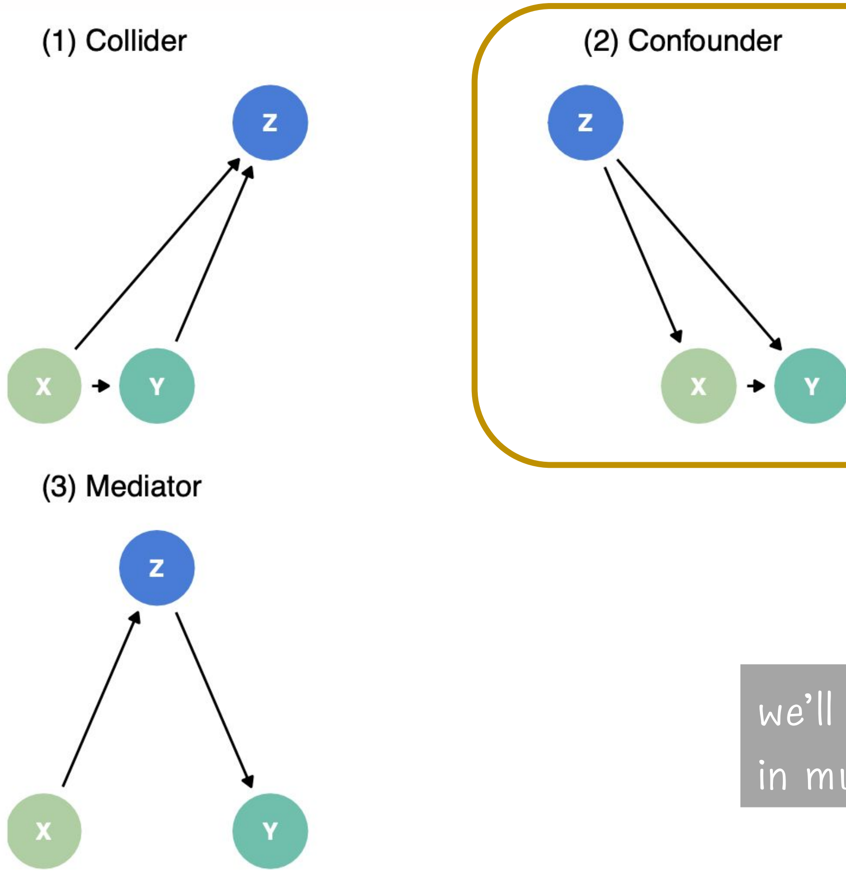
Source: Barrett, M., McGowan, L. D., & Gerke, T. (2024). *Causal Inference in R*. Retrieved from https://www.r-causal.org/

Typical challenges in estimating causal effects: visual intuition

- Directed acyclic graphs (DAGs) can offer visual intuition of the causal nexus at play in the 3 datasets. Failure to adjust models to these situation leads to BIAS
  - X is some continuous exposure of interest, Y a continuous outcome, and Z a known, <u>measured</u> factor

(1) a **"COLLIDER"** <u>is caused by both X and Y</u> (it inadvertedly connects the 2). E.g.:
- X = sodium intake
- Y = systolic blood pressure
- Z = urinary protein excretion

(3) a **"MEDIATOR"** is <u>caused by X</u> and then it <u>causes Y</u>. E.g.:
- X = screen time
- Y = obesity
- Z = physical exercise



(2) a "CONFOUNDER" <u>causes both X and Y.</u> E.g.:
- X = smoking
- Y = lung cancer
- Z = alcohol (consumers also tend to be smokers)

we'll revisit this later in multivariate regression...

Source: Barrett, M., McGowan, L. D., & Gerke, T. (2024). *Causal Inference in R*. Retrieved from https://www.r-causal.org/

Various types of bias

……

Time invariant bias

Shifting emphasis on empirical outcome prediction

Introduction to **Machine Learning (ML)** models

A conceptual framework to understand different types of statistical modeling (part 2/2)

1. **association/correlation** → observational studies
2. **causal explanation** → experimental studies
3. **empirical prediction** → algorithmic machine learning and data-mining modeling
   - aimed at **predicting new or future observations** (without necessarily explaining how)
   - relies on **big data**
   - prevalent in fields like natural language processing, bioinformatics, etc.. In epidemiology, there is more of a mix causal explanation & empirical prediction

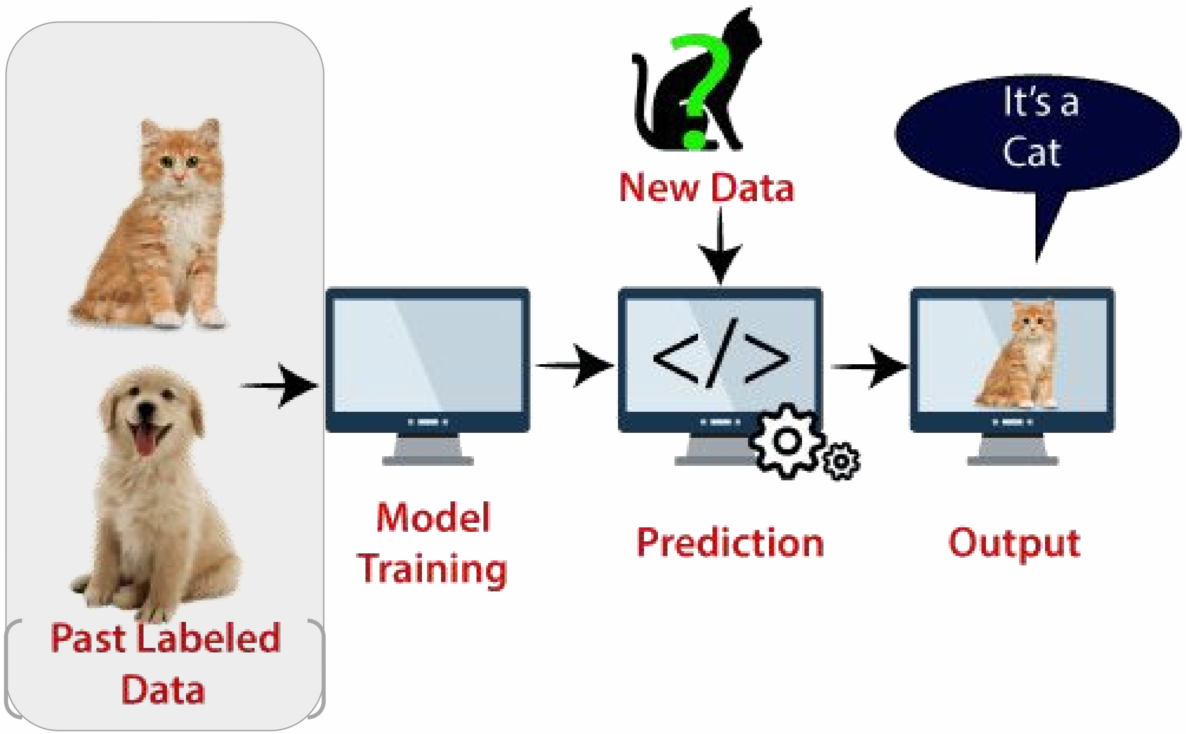
- **NOTES**:
  - ✓ "Prediction" does not necessarily refer to future events, but rather to *future* datasets that were previously unseen to the algorithm

MACHINE LEARNING

Defining Machine Learning (ML)

"At its core, Machine Learning is just a "thing-labeler", taking something and telling you what label It should get."

(Cassie Kozyrkov)



Source: Image from https://entri.app/blog/what-is-svm-algorithm-in-machine-learning/

Defining Machine Learning (ML)

- **Machine Learning** is a broad and highly active research field. (In the life sciences, "*precision medicine*" is an application of machine learning to biomedical data)
- The **general idea** is to predict or discover outcomes from measured predictors, in problems like:
  - *Can we discover new types of cancer from gene expression profiles?*
  - *Can we predict drug response from a series of genotypes?*
  - *How do we classify a set of images/spectrometry outputs, etc.*
  - *Given various clinical parameters, how can we use them to predict heart attacks?*
- The **ML is a data-driven (inductive) approach**, where a machine **learns** the rules/patterns from a set of **training data** and (then) **validates** findings on a set of **testing data**
- In contrast with inferential statistics, **ML *doesn't worry* about assumptions on parameters** (probability distribution, error, correlation, etc.), **nor the causal nexus** between specific predictor(s) and response, **nor the data collection strategy**
- In contrast with standard statistics, **in ML the rules are not necessarily specified**… hence ML = a subfield of AI

Stylized comparison between statistics and machine-learning

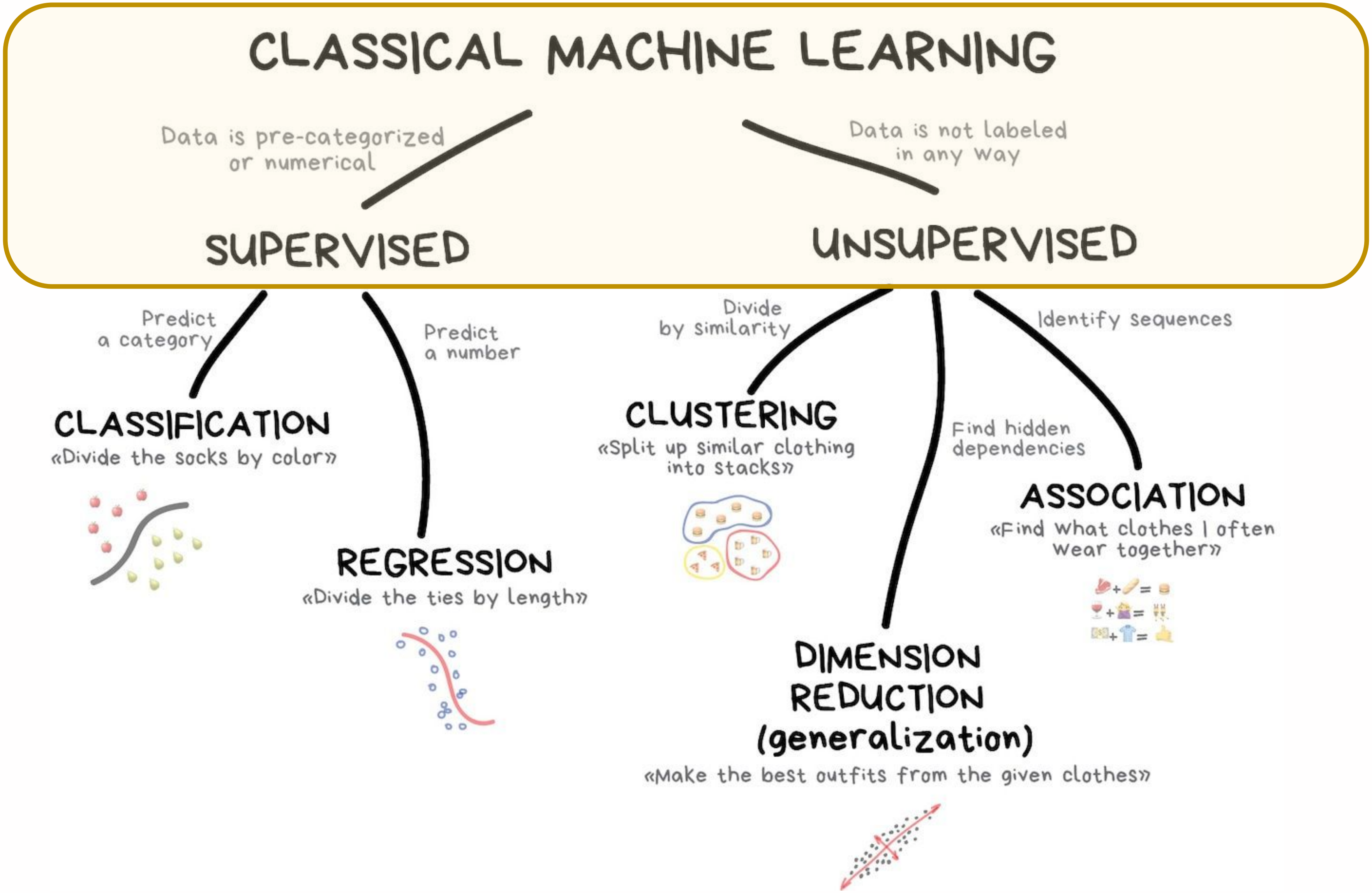
	Standard (causal inference) Statistics	Machine Learning
Typical Goal	Explanation, uncovering causal relationships	Predicting an outcome as accurately as possible
Typical Task	Research based on a theory to identify the causal effect (better: pre-register your hypothesized model).	Try out and tune many different algorithms in order to maximize predictive accuracy in new and unseen test datasets.
Data generating process	Designed ex-ante based on study goal (e.g. randomized control trial, or observational study with statistical control variables)	Useful but not strictly necessary, and often not available
Parameters of interest:	Causal effect size and statistical significance, p-value of treatment X for outcome Y	Model's accuracy (%), precision/recall, sensitivity/specificity, in predicting Y
Dataset	Use **ALL AVAILABLE DATA** to calculate effect of interest (it was designed to be representative of a population).	It is critical to **SPLIT THE DATA** (usually 75% for training and 25% for testing the algorithms) leaving aside a sub-sample to test the model with unseen new data

Source: Adapted from https://forloopsandpiepkicks.wordpress.com/2022/02/10/beginners-guide-to-machine-learning-in-r-with-step-by-step-tutorial/

Supervised or Unsupervised ML algorithms?

….another conceptual framework

A fundamental distinction: supervised and unsupervised ML



Source: Image from https://vas3k.com/blog/machine_learning/index.html

A fundamental distinction: supervised and unsupervised ML

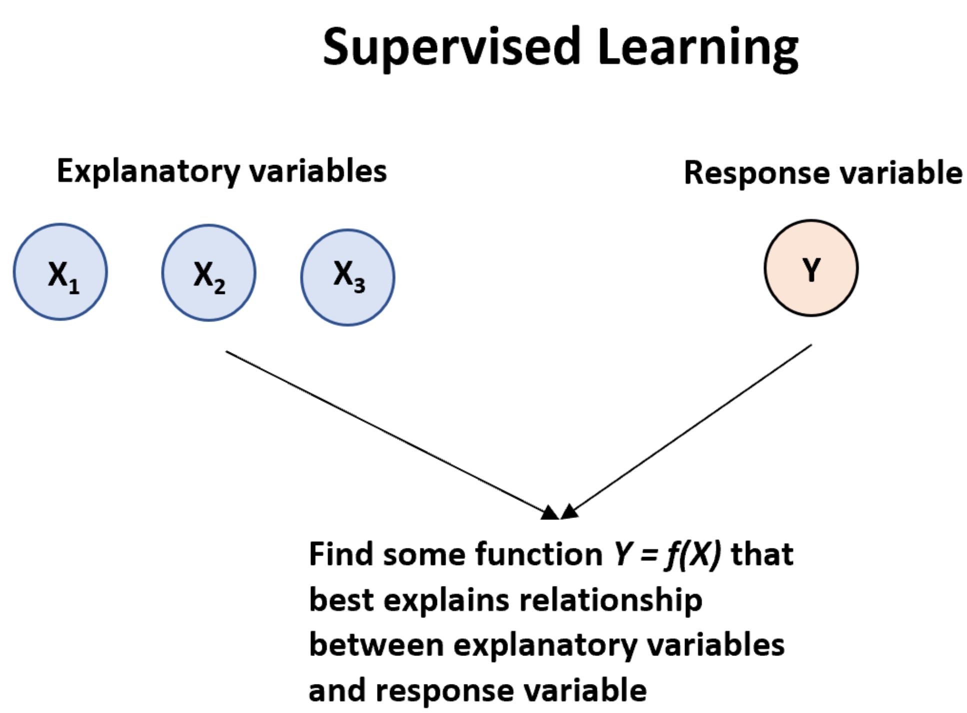
- ML includes many different algorithms that can be used for understanding data. These algorithms can be classified as:

- **Supervised Learning Algorithms**:
  - building a model to estimate or predict an output based on one or more inputs
    - **Regression**: Modeling a relationship, the typical output variable is continuous (e.g. weight, height, time, etc.) or dichotomous.
    - **Classification**: Splits objects based on one of the attributes known beforehand. The the typical output variable is categorical (e.g. male or female, pass or fail, benign or malignant, etc.)
- **Unsupervised Learning Algorithms**:
  - finding structure and relationships among inputs. There is no "supervising" output
    - **Clustering:** Finding "clusters" of observations in a dataset that are similar to each other (*based on unknown features*).
    - **Association:** Finding "rules" that can be used to draw associations. For example, if a patient has a high biomarker X, he will have a low biomarker Y.
    - **Dimension reduction**: Assembling specific features into more high-level ones (e.g. PCA)

Supervised ML algorithms

Supervised Learning Algorithms mechanics

- A supervised learning algorithm can be used when we have **one or more explanatory variables** (X_1, X_2, X_3, …, X_p) and a **response variable** (Y) and we would like to find some function that describes the relationship between the explanatory variables and the response variable:
- $Y = f(X) + \varepsilon$
- where
  - f () represents **systematic information that X provides about Y** and where
  - ε is a random error term independent of X with a mean of zero.



Source: https://www.statology.org/supervised-vs-unsupervised-learning/

Supervised Learning Algorithms purpose

There are two main reasons to use supervised learning algorithms:

1. **Prediction**: We often use a set of explanatory variables to predict the value of some response variable (e.g. using square footage and number of bedrooms to predict home price)
2. **Inference**: We may be interested in understanding the way that a response variable is affected as the value of the explanatory variables change (e.g. how much does home price increase, on average, when the number of bedrooms increases by one?)

- *Depending on whether our goal is inference or prediction (or a mix of both), we may use different methods for estimating the function f. For example, linear models offer easier interpretation but non-linear models that are difficult to interpret may offer more accurate prediction.*

Supervised Learning: commonly used algorithms

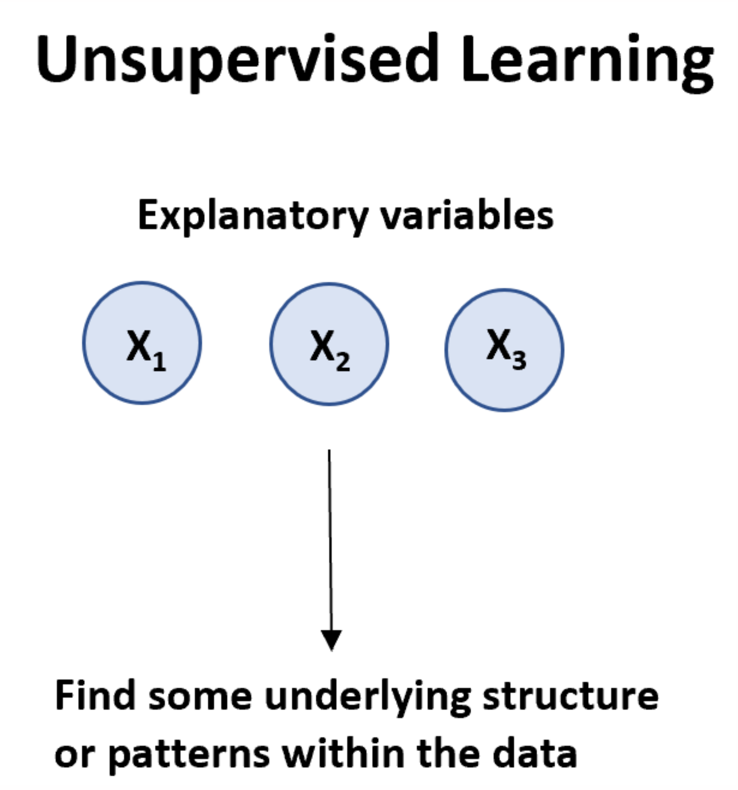
Most commonly used supervised learning algorithms:

- Linear regression
- Logistic regression
- Linear discriminant analysis
- Quadratic discriminant analysis
- Decision trees
- Naive bayes
- Support vector machines
- Neural networks

Unsupervised ML algorithms

Example of PCA

An unsupervised learning algorithm can be used when we have a list of variables (X_1, X_2, X_3, …, X_p) and we would simply like to find underlying structure or patterns within the data.



Source: https://www.statology.org/supervised-vs-unsupervised-learning/

There are two main types of unsupervised learning algorithms:

1. **Clustering:** Using these types of algorithms, we attempt to find “clusters” of observations in a dataset that are similar to each other. This is often used in retail when a company would like to identify clusters of customers who have similar shopping habits so that they can create specific marketing strategies that target certain clusters of customers.
2. **Association:** Using these types of algorithms, we attempt to find “rules” that can be used to draw associations. For example, retailers may develop an association algorithm that says “if a customer buys product X they are highly likely to also buy product Y.”

- Most commonly used unsupervised learning algorithms:
  - Principal component analysis
  - K-means clustering
  - K-medoids clustering
  - Hierarchical clustering
  - Apriori algorithm

Summary: Supervised vs. Unsupervised Learning

- Here are the key differences between supervised and unsupervised learning algorithms:

	Supervised Learning	Unsupervised Learning
Description	Involves building a model to estimate or predict an output based on one or more inputs.	Involves finding structure and relationships from inputs. There is no “supervising” output.
Variables	Explanatory and Response variables	Explanatory variables only
End goal	Develop model to **(1)** predict new values or **(2)** understand existing relationship between explanatory and response variables	Develop model to **(1)** place observations from a dataset into a specific cluster or to **(2)** create rules to identify associations between variables.
Types of algorithms	**(1)** Regression and **(2)** Classification	**(1)** Clustering and **(2)** Association

Source: https://www.statology.org/supervised-vs-unsupervised-learning/