

# STATISTICS & ML WITH R

Mapping causal & predictive  
approaches

**2024**

**M. Chiara Mimmi & Luisa M. Mimmi**

# DAY 4 – LECTURE OUTLINE

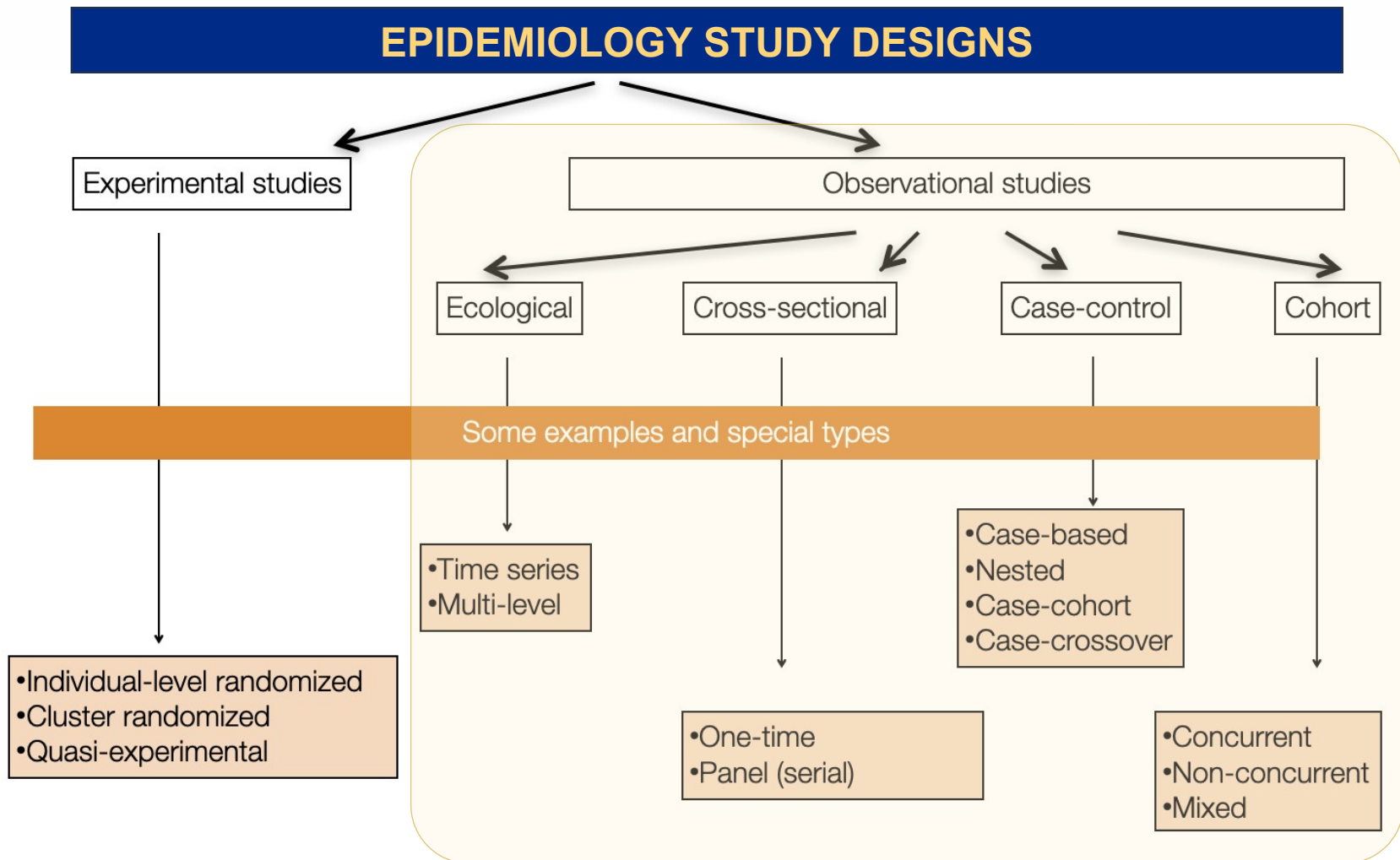
## Mapping causal & predictive approaches

- Illustrating different study designs
- Learning the vocabulary of causal analysis
- Visual understanding of causal pathways, including:
  - Collider variables
  - Confounder variables
  - Mediator variables
- Learn how to address causal pathways in modeling, based on the research question
- Defining causal outcomes and commonly used “estimands” (ATE, ATT, ATU)
- Understand proper statistical methods to estimate ATE, ATT, ATU based on research question and target group
- Introducing Machine Learning (ML)
  - purpose
  - key algorithms' categories

# From *observational* to *experimental* studies

- “**OBSERVATIONAL STUDIES**” on variables of interest and their relationships have *no controlled assignment of the treatment*
  - We may find **CORRELATION / ASSOCIATION**, but it DOES NOT IMPLY CAUSATION! Why?
  - ... **hidden variables** may affect the relationship between the **explanatory variable** and the **response variable**
  - ...*but often used (implicitly or not) to estimate causal effect of an exposure!*
- “**EXPERIMENTAL STUDIES**” seek to uncover **CAUSATION**, so they are *designed to provoke a response*
  - Researchers **assign** the treatment to an **experimental unit** (or **subject**) and observing its **effect**
  - These studies use some *ad hoc* **design principles** and **controlled independent variables**

# Experimental and non-experimental study designs...



Source: <https://bookdown.org/jbrophy115/bookdown-clinepi/design.html>

# Different goals 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 (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

# Different goals 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

Key terminology and visual causal maps

# The conceptual framework for causal analysis (1/3)

- **Fundamental vocabulary:**
  - **Intervention** decisions and actions that change the behaviors or situation of people/firms/other subjects (drug, vaccine, program participation)
    - **TREATMENT** = commonly used in experimental studies when researchers directly “assigns” the **causal variable**
    - **EXPOSURE** = commonly used observational studies when participants “naturally” experience the **causal variable**
  - **Subjects** = those that may be affected (at least in principle), in fact are
    - TREATED subjects
    - UNTREATED subjects
  - **Outcome** = variable(s) that may be affected by the intervention
    - can be caused by exposure either directly or through an intermediate process
  - **Causation** = causal processes that lead to the development of outcomes



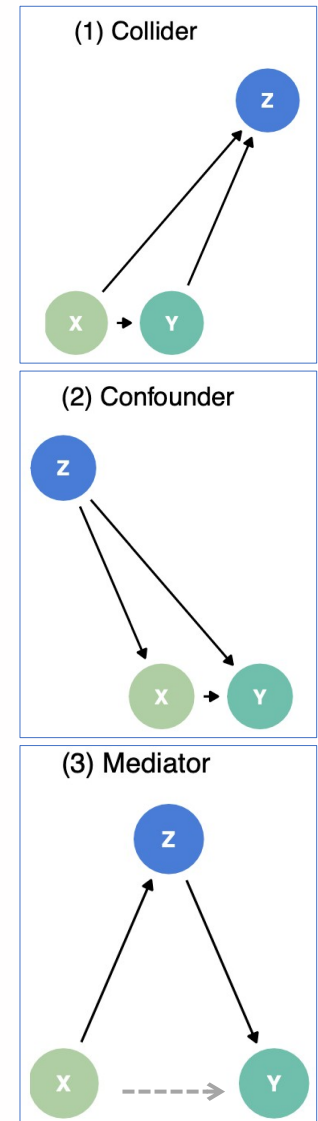
# The conceptual framework for causal analysis (2/3)

- Fundamental vocabulary (“tricky ones” 😊):
  - **Bias** = systematic error that can occur at different stages of the study: *data collection, analysis or interpretation* of the causal relationship exposure-outcome.
    - **Selection bias** occurs when both the exposure and the outcome affect whether an individual is included in the sampled population
    - **Information bias** occurs when there is misclassification or inaccurate measurement (e.g., patients underreporting smoking habits)
    - **prevalence-incidence bias** ...
    - **ecological bias**

# The conceptual framework for causal analysis (3/3)

- Fundamental vocabulary (“tricky ones” 😊):

- Collider** = variable that is **influenced by treatment and outcome** (like a “common effect”)
  - EXAMPLE:** sleepiness (Z), with shift work (X) and apnea (Y)
  - Conditioning on or controlling for a collider in the causal model can create a distortion (“*collider bias*”)
- Confounder** = variable that **affect both treatment and outcome** (“apparent” cause), but it is **not in the causal pathway**
  - EXAMPLE:** smoking (Z), with exercise (X) and lung cancer (Y)
  - Most confounder variables involve some **kind of selection** (e.g., self-selection) that can be addressed stratifying subjects by it
- Mediator** = is a variable that is **in the causal pathway** and “explains” why **treatment affects outcome** (like a “mechanisms”)
  - EXAMPLE:** immune function (Z), with exercise (X) and lung cancer (Y)
  - Conditioning on or controlling for a mediator can be done to assess what *part of the effect* they play



# Estimands, Estimators, Estimates

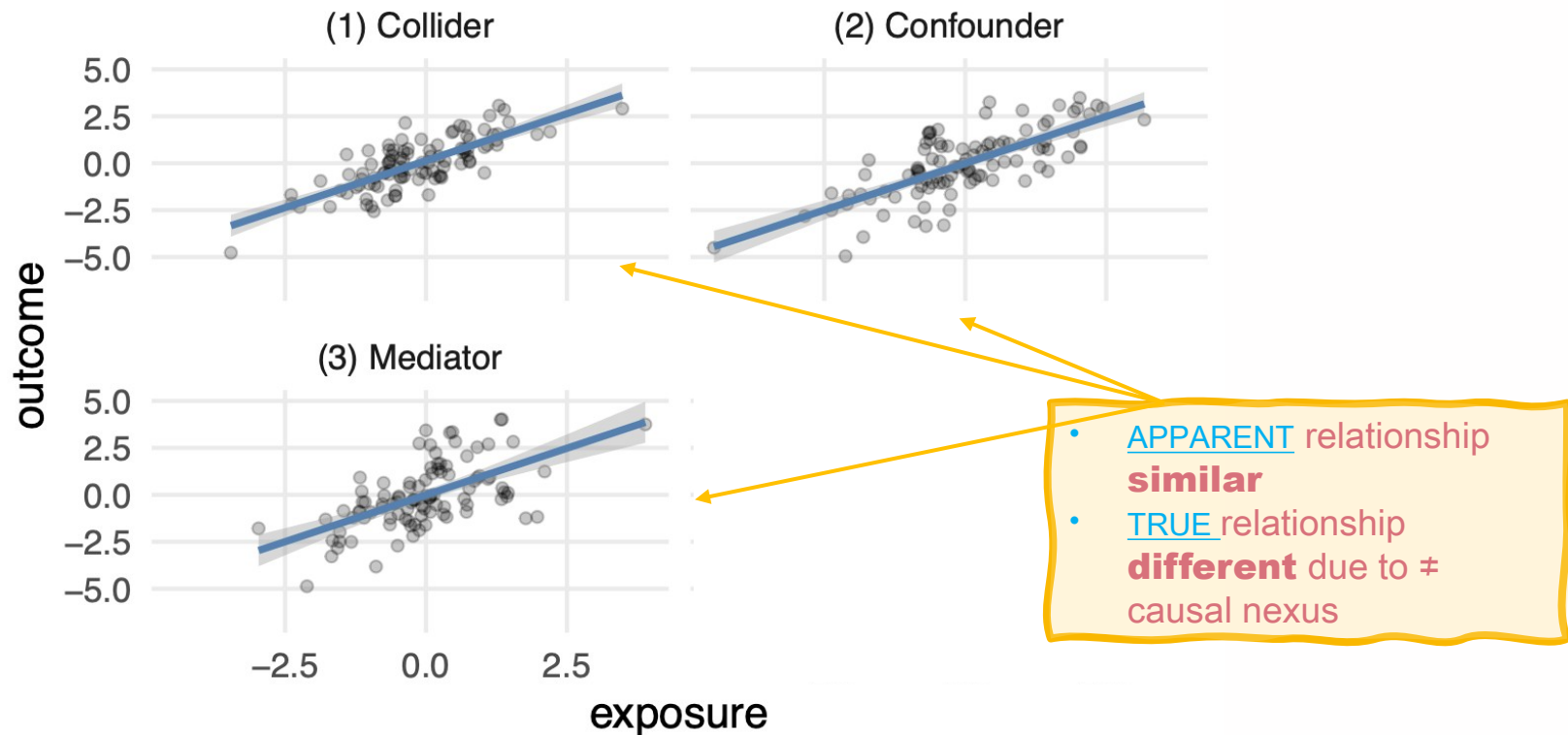
- The **estimand** is the *target of interest*
  - **EXAMPLE:** expected value of the difference in potential outcomes across all individuals
- The **estimator** is the method by which we approximate this estimand using data (“recipe”)
  - **EXAMPLE:** in randomized controlled trial, our estimator could just be the average outcome among those who received the exposure A minus the average outcome among those who receive exposure B
- The **estimate** is the value we get when we plug our data into the estimator
  - **EXAMPLE:** randomized controlled trial, our estimator could just be the average outcome among those who received the exposure A minus the average outcome among those who receive exposure B

# Visualizing causal maps

A helpful tool in guiding statistical modeling

# 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)



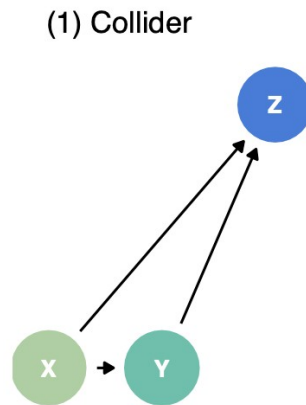
Source: Barrett, M., McGowan, L. D., & Gerke, T. (2024). *Causal Inference in R*. Retrieved from <https://www.r-causal.org/>

# 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, measured factor

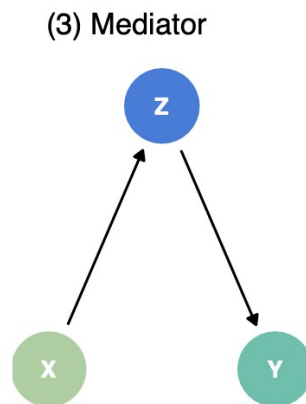
(1) a **“COLLIDER”** is caused by both  $X$  and  $Y$  (it inadvertently connects the 2). E.g.:

- $X$  = sodium intake
- $Y$  = systolic blood pressure
- $Z$  = urinary protein excretion

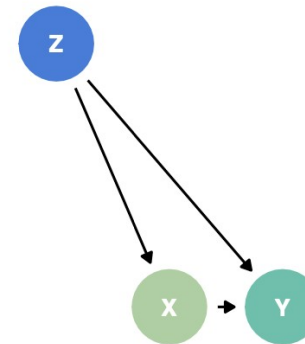


(3) a **“MEDIATOR”** is caused by  $X$  and then it causes  $Y$ . E.g.:

- $X$  = screen time
- $Y$  = obesity
- $Z$  = physical exercise



(2) Confounder



(2) a **“CONFOUNDER”** causes both  $X$  and  $Y$ . 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/>

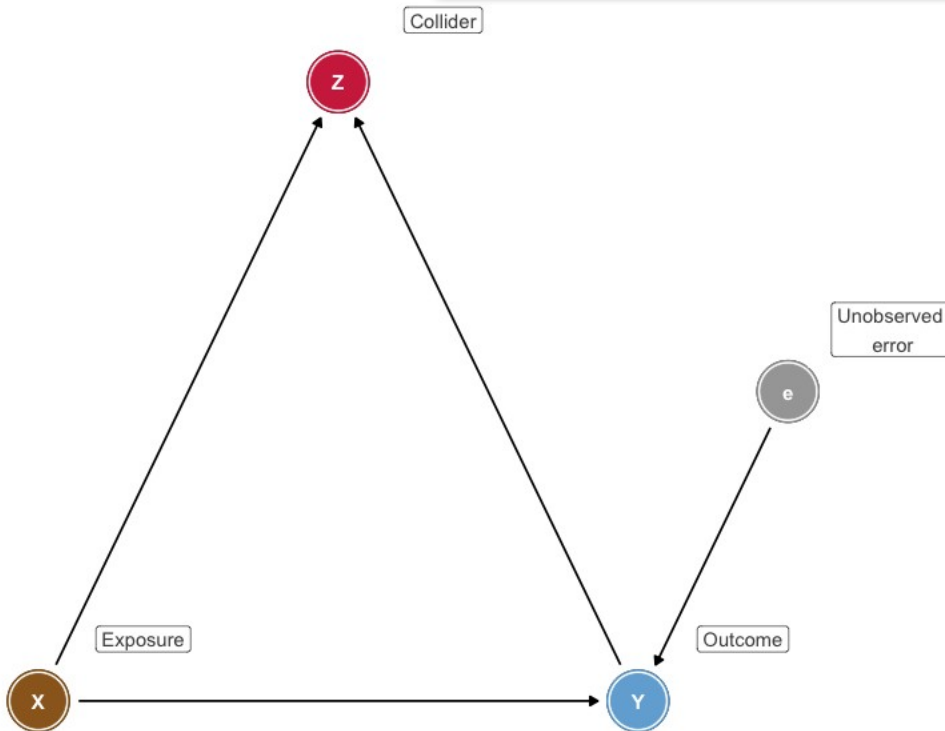
# How to address causal pathways in modeling

This will be re-visited through practical  
examples in Lab 4

# How to deal with collider when modeling?

## RIFARE

Causal map with COLLIDER (Z)



We must NOT control for collider.

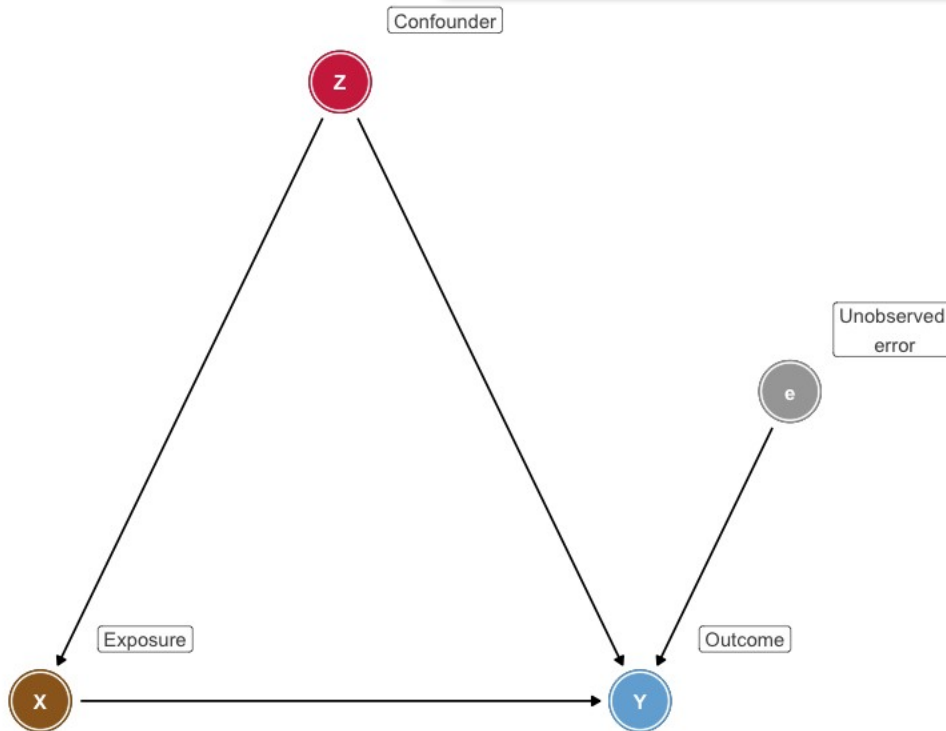
- **colliders CAN HIDE REAL CAUSE EFFECTS**
  - if we adjust/control for a collider we increase bias bc it would distort the true relationship between the exposure and the outcome



# How to deal with **confounder** when modeling?

## RIFARE

Causal map with CONFOUNDER (Z)



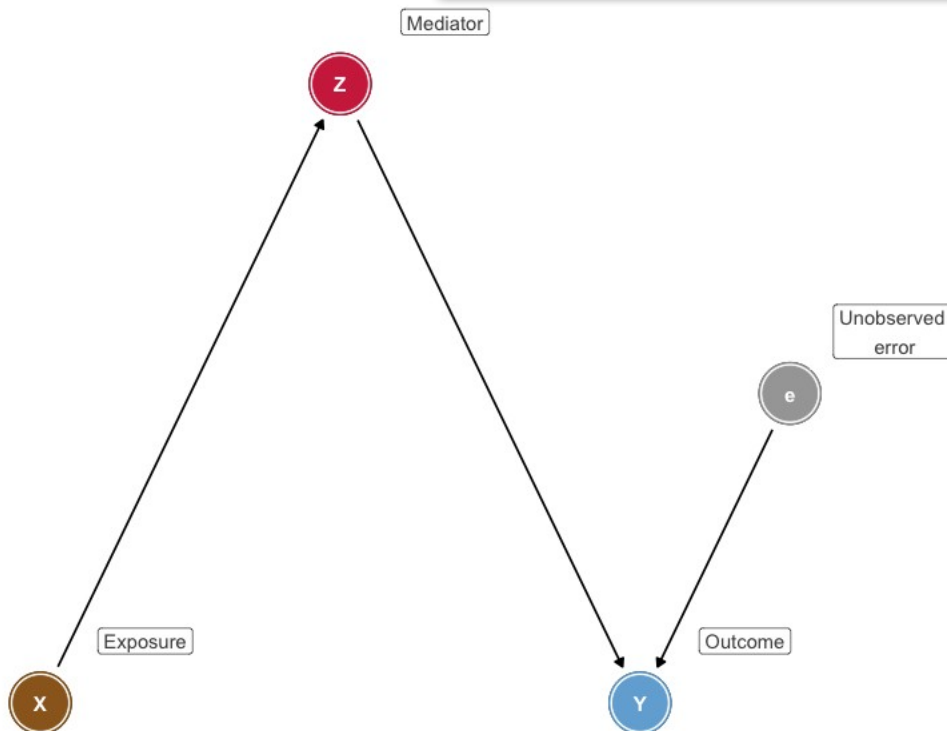
We must control for a confounding variable, so we reduce bias, i.e.:

- include term in regression
- matching
- stratifying...

# How to deal with mediator when modeling?

## RIFARE

Causal map with MEDIATOR (Z)



We could control for the variable named “M” (mediator/mechanism), depending on which effect you want to focus on:

- we should not adjust for mediator (you will not find the total (indirect) effect from x to Y) |
- Conditioning on a mediator closes one of the causal paths between X and Y and distorts the overall relationship between the two.

# Measuring causal outcomes of interest

Commonly used “estimands” (ATE, ATT, ATU)  
and how to select and interpret them  
correctly for making valid inferences

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

- **NOTATION:**

- $y_{0i}$  and  $y_{1i}$  are the potential outcomes in the *absence* and *presence* of treatment
- for patient  $i$  in a study on a new drug on blood pressure,
  - $y_{0i}$  = without new drug
  - $y_{1i}$  = with new drug

- **ITE = Individual Treatment Effect (\*)** = difference, for subject  $i$ , between potential outcome if treated and if untreated

where treatment is

- (\*) ITE is never observable!!
- Hence, we will look at averages...

- **ATE = Average Treatment Effect** = average of ITE differences across subjects

- (\*) The Avg of the differences = the difference of Averages!
- ATE can hide different distributions of ITEs (e.g., positives and negatives that cancel each other out)
- Important to have a well-defined group or population

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

- **ATT (or ATET) = Average Treatment effect on the Treated** = average treatment effect across all subjects that end up TREATED

]

- This refers to the avg of the differences conditionally on the fact that both groups “received” the treatment (“” )
- is essentially the counterfactual for in a '**parallel universe**' where **exactly the same people** who were treated in this universe would not get the treatment
- **ATU = Average Treatment effect on the Untreated** = average treatment effect across all subjects who were NOT TREATED

]

- This time we seek the Avg of the differences (“” ) conditionally on the fact that both groups were “assigned” to the treatment
- is essentially the counterfactual for in a 'parallel universe' where **exactly the same people** who were NOT treated in this universe would get the treatment

# BY THE WAY !

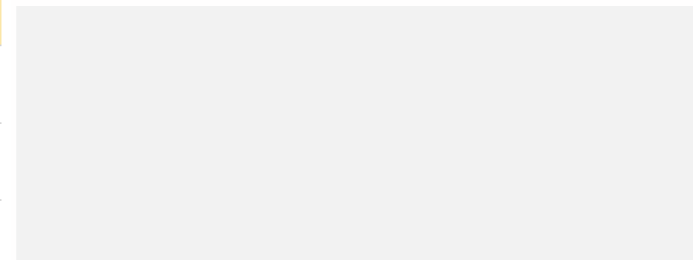
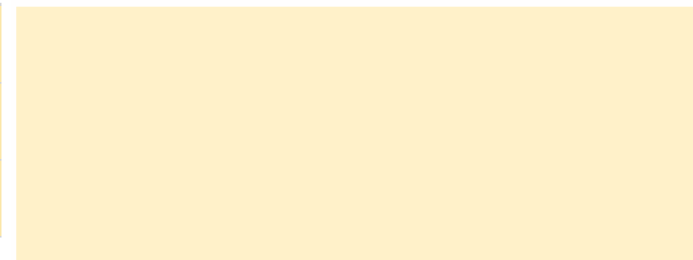
- treatment is a binary random variable
- outcome of interest is
- **ATE = Average Treatment Effect** = average of ITE differences across subjects
- **ATT/ATET = Average Treatment effect on the Treated** = average treatment effect across all subjects that end up TREATED

Ex: Does hospitalization (T) increase health (Y) ?	
(ATE)	Avg health of hospitalized group – avg health of <b>NOT hospitalized group</b>
(ATT)     + ...	Avg health of treated group – [counterfactual] avg health of <b>treated group</b> <b>IF NOT hospitalized</b>
(Selection bias) ... + <i>(hospitalized have worse than non hospitalized)</i>	Difference in [counterfactual] avg health of <b>treated group IF NOT hospitalized</b> - those who were <b>NOT hospitalized</b>

# EXE. potential causal outcomes ( depends on patients characteristics)

	Confounder	Treatment	Unobservable			Realized
	Age	Treated	Potential outcomes		ICE or $\delta_i^*$	Outcome
ID	$Z_i$	$X_i$	$Y_i^1$	$Y_i^0$	$Y_i^1 - Y_i^0$	$Y_i$
1	Old	1	80	60	20	80
2	Old	1	75	70	5	75
3	Old	1	85	80	5	85
4	Old	0	70	60	10	60
5	Young	1	75	70	5	75
6	Young	0	80	80	0	80
7	Young	0	90	100	-10	100
8	Young	0	85	80	5	80

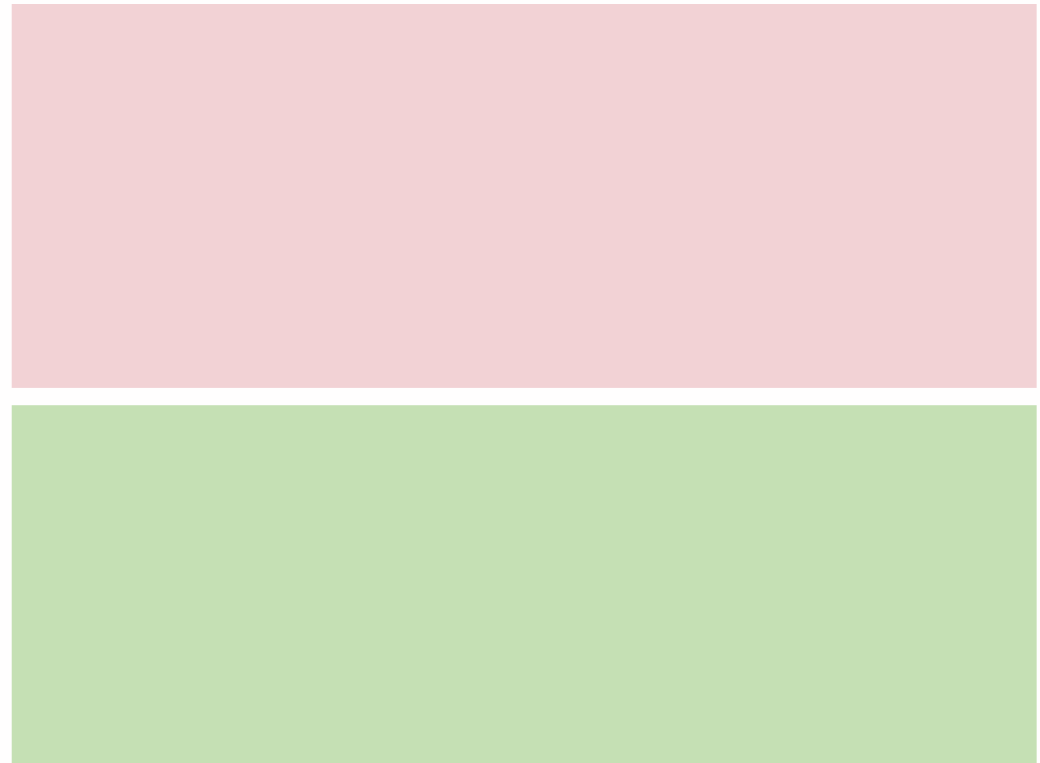
\* ICE = individual causal effect



(ATE decomposition)

# Stratification to deal with **confounder** (i.e. combining the weighted averages for old and young people)

	<i>Confounder</i>	<i>Treatment</i>	<i>Realized</i>
	<b>Age</b>	<b>Treated</b>	<b>Outcome</b>
<b>ID</b>	$Z_i$	$X_i$	$Y_i$
1	Old	1	80
2	Old	1	75
3	Old	1	85
4	Old	0	60
5	Young	1	75
6	Young	0	80
7	Young	0	100
8	Young	0	80



$\tau = 4.1667$

After stratification based on the **confounder** we get a very close **approximation of the ATE**



# Other ways to deal with confounders

- GIVEN THAT IN REAL LIFE WE NEVER HAVE THE ALTERNATIVE POTENTIAL OUTCOMES FOR EACH SUBJECT  $i$ , HOW DO WE DEAL?
  - **STRATIFYING** by Age was easy, but what if there is >1 confounder? What if it is continuous?
- WE'VE GOT TO DO SOMETHING ELSE TO GET TO COMPARABLE GROUPS:
  - **MATCHING** methods = dropping units from the sample or partitioning units into pairs or subclasses (e.g., *Propensity Score Matching*)
  - **WEIGHTING** methods = weighting the units so that the weighted distributions are similar between treatment groups (e.g., *Inverse probability weighting*)
- After *adjustment* the treatment effect is estimated in the resulting sample (incorporating the weights resulting from the matching or weighting)

# Choosing the **estimands** and the proper statistical method to estimate the effect

- In a **randomized trial**, the treated and untreated groups will, on average, have the same distributions of patient characteristics, so the **ATT, ATU, and ATE** will be the same
- **Without randomization**, however, the treatment groups can have quite different distributions of characteristics, **ATT, ATU, and ATE** will differ when these characteristics also relate to the treatment effect
  - So, when using observational data: *for whom should the treatment effect be estimated?*
  - ~~Some methods, such as PSM in its most commonly used form, cannot target the ATE, and so are inappropriate when the ATE is of interest!~~

# Choosing the **estimands** based on the research question

BEFORE analyzing an observational dataset, let's consider **which question** we are asking, and **about which target** population group,

THEN choose a **statistical method** that corresponds to the chosen estimand.

Estimands	Target Population	Example research question and research/policy addressed
ATT	Treated patients	<i>Examining an intervention that would only reach those currently receiving it:</i> <i>- e.g. decision to replace / withhold a treatment for currently treated patients</i>
ATU	Untreated patients (control)	<i>How would untreated patients respond to a new potential treatment/exposure?</i> <i>- e.g. decision to extend a medical practice (drug prescription/vaccine) to a group that would not otherwise receive it</i>
ATE	<i>Full sample / population</i>	<b>Should a specific policy be applied to all eligible patients? How would the outcome be on average?</b> <i>- e.g. regulating a system-wide policy for a previously unregulated practice</i> <i>- useful when treatment decisions are not well informed (ATE does not depend on current treatment assignment)</i> <i>- NOT OK when patients' benefit depend on clinical judgment</i>

## EXE [see LAB 4]: how to exploit “\**matched*” untreated observation to estimate the ATT

- .... ANDREW HEISS ESEMPIO DI PSM (uso il suo che e' troppo bello!!!!)

# Shifting emphasis on empirical outcome prediction

Introduction to Machine Learning (ML)  
models

- <https://statisticalhorizons.com/the-machine-learning-foundations-of-artificial-intelligence/>

# 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

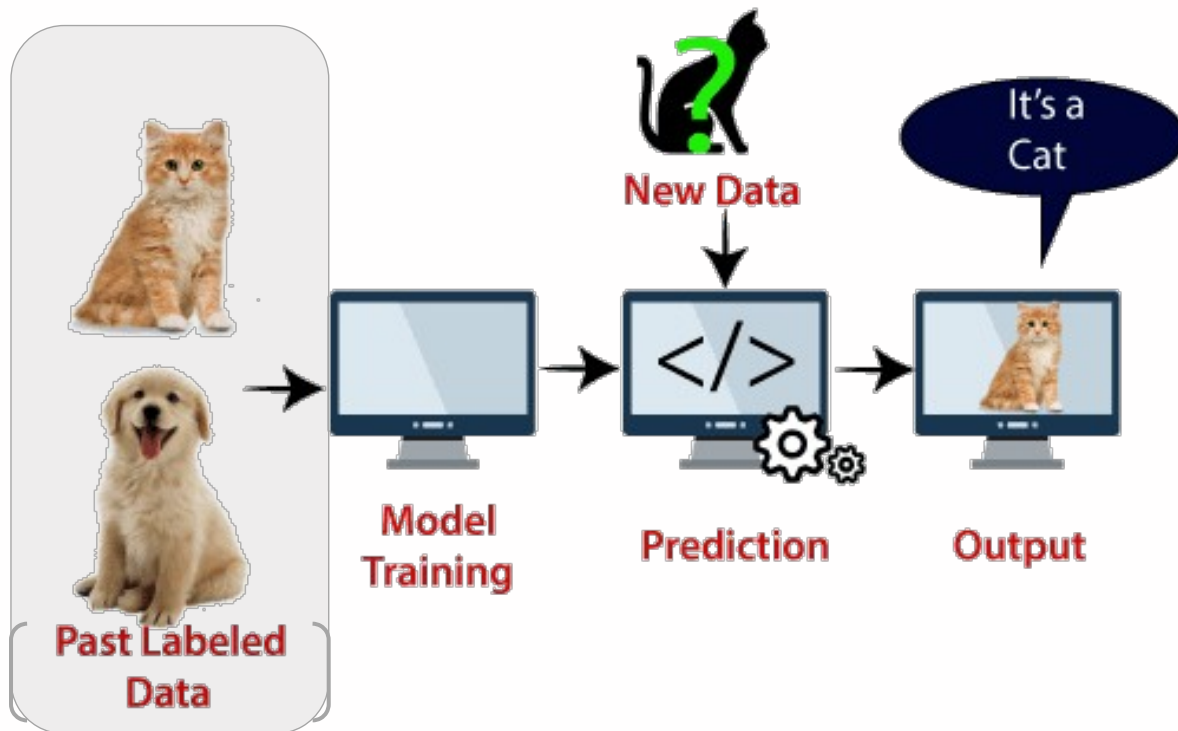


# Defining Machine Learning (ML)



**“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

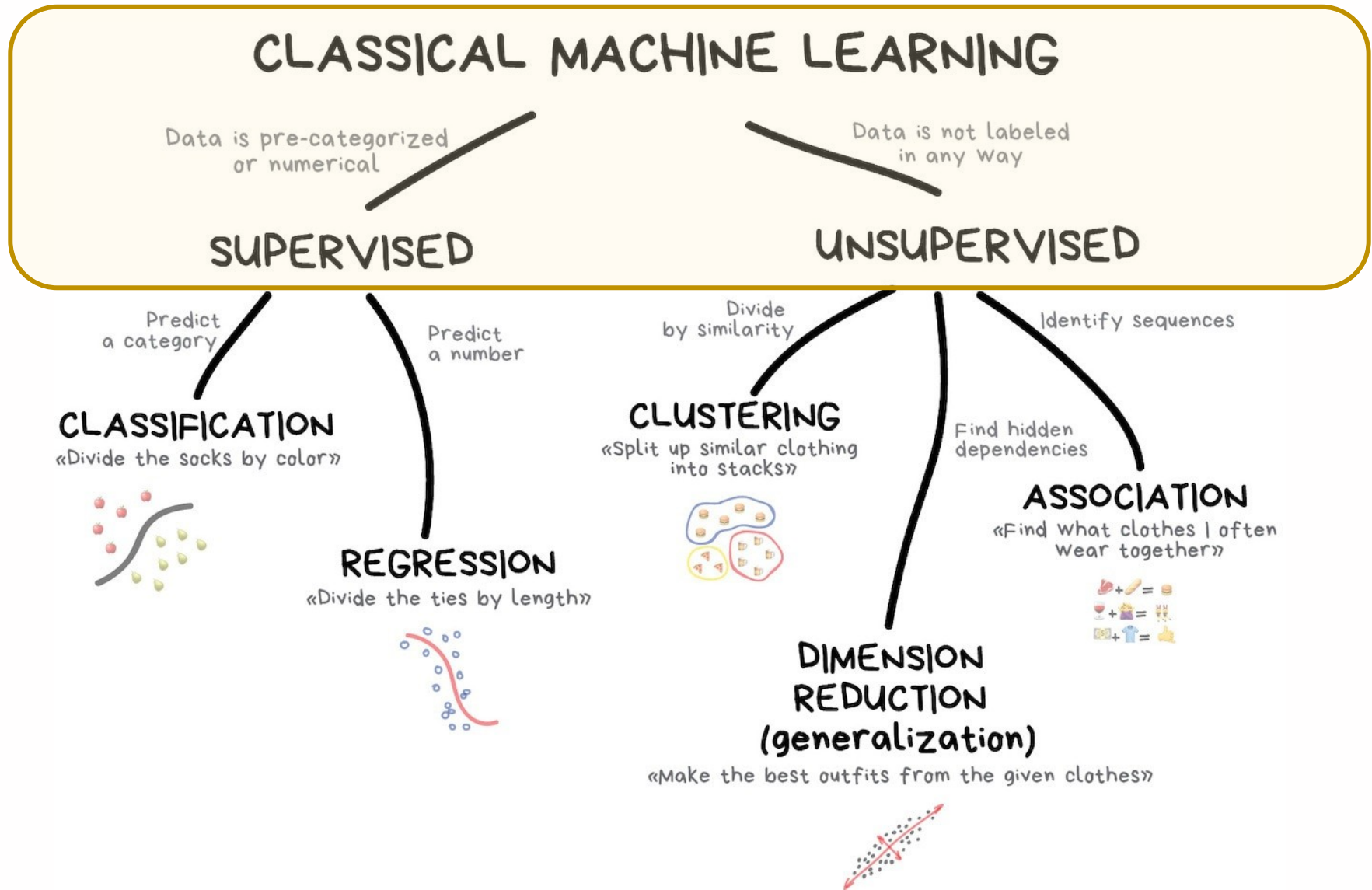
	Standard (causal inference) Statistics	Machine Learning
Typical Goal	Explanation, uncovering causal relationships	Predicting an outcome as accurately as possible
Typical Task	Research based on a theory to identify the <u>causal effect</u> (better: pre-register your hypothesized model).	Try out and tune many different algorithms in order to <u>maximize predictive accuracy</u> 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 <u>treatment X</u> for outcome Y	Model's accuracy (%), precision/recall, sensitivity/specificity, in <u>predicting Y</u>
Dataset	Use <b>ALL AVAILABLE DATA</b> to calculate effect of interest (it was designed to be representative of a population).	It is critical to <b>SPLIT THE DATA</b> (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://forloopsandpiekicks.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



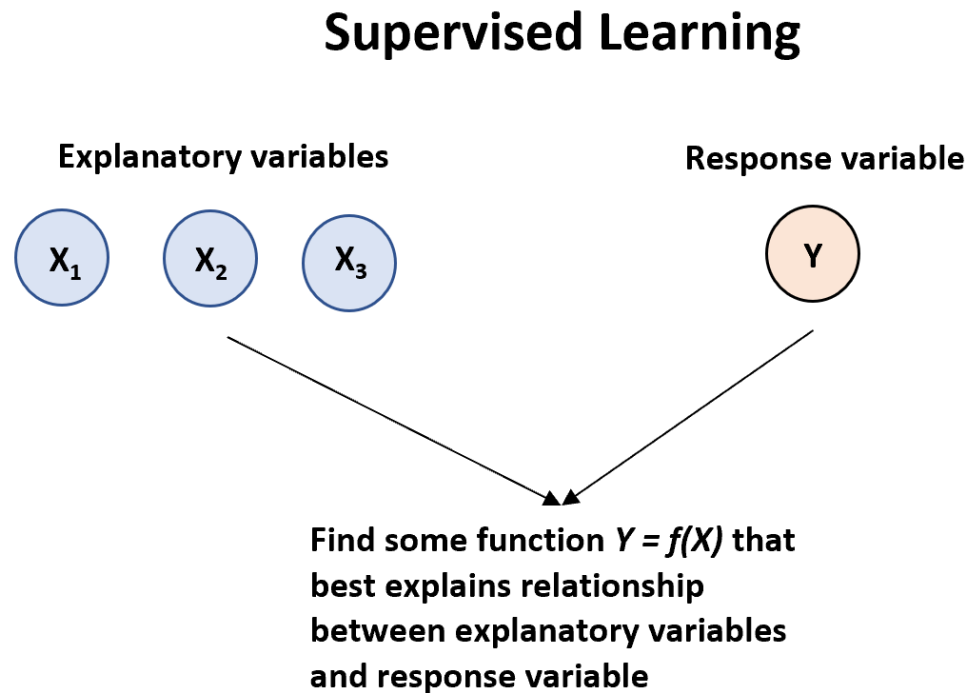
Source: Image from [https://vas3k.com/blog/machine\\_learning/index.html](https://vas3k.com/blog/machine_learning/index.html)

# 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

- A supervised learning algorithm can be used when we have **one or more explanatory variables** ( $X_1, X_2, X_3, \dots, X_p$ ) 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:
- $Y = f(X) + \epsilon$
- where
  - $f()$  represents **systematic information that X provides about Y** and where
  - $\epsilon$  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:

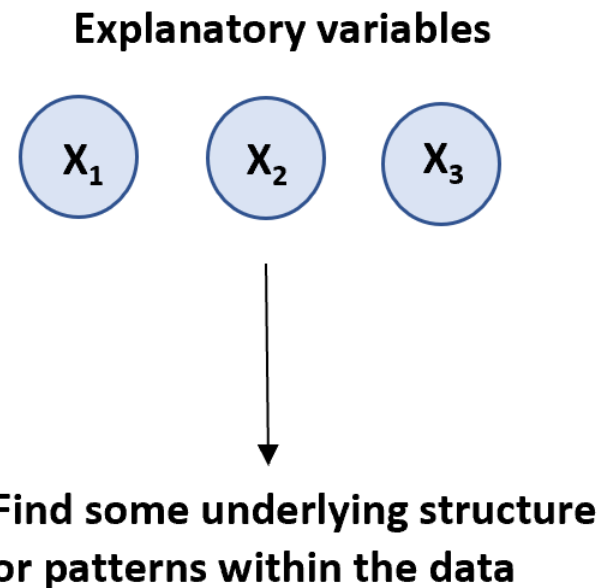
- Linear regression
- Logistic regression
- Linear discriminant analysis
- Quadratic discriminant analysis
- Decision trees
- Naive bayes
- Support vector machines
- Neural networks

# Unsupervised ML algorithms

## Example of PCA

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

## Unsupervised Learning



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
<b>Description</b>	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.
<b>Variables</b>	Explanatory and Response variables	Explanatory variables only
<b>End goal</b>	Develop model to <b>(1)</b> predict new values or <b>(2)</b> understand existing relationship between explanatory and response variables	Develop model to <b>(1)</b> place observations from a dataset into a specific cluster or to <b>(2)</b> create rules to identify associations between variables.
<b>Types of algorithms</b>	<b>(1)</b> Regression and <b>(2)</b> Classification	<b>(1)</b> Clustering and <b>(2)</b> Association

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