DATA, MSML, BIOI 602 Principles of Data Science

Pandas and Databases(SQL)

Lecture 03

2024

[Slides from Fardina Fathmiul Alam]

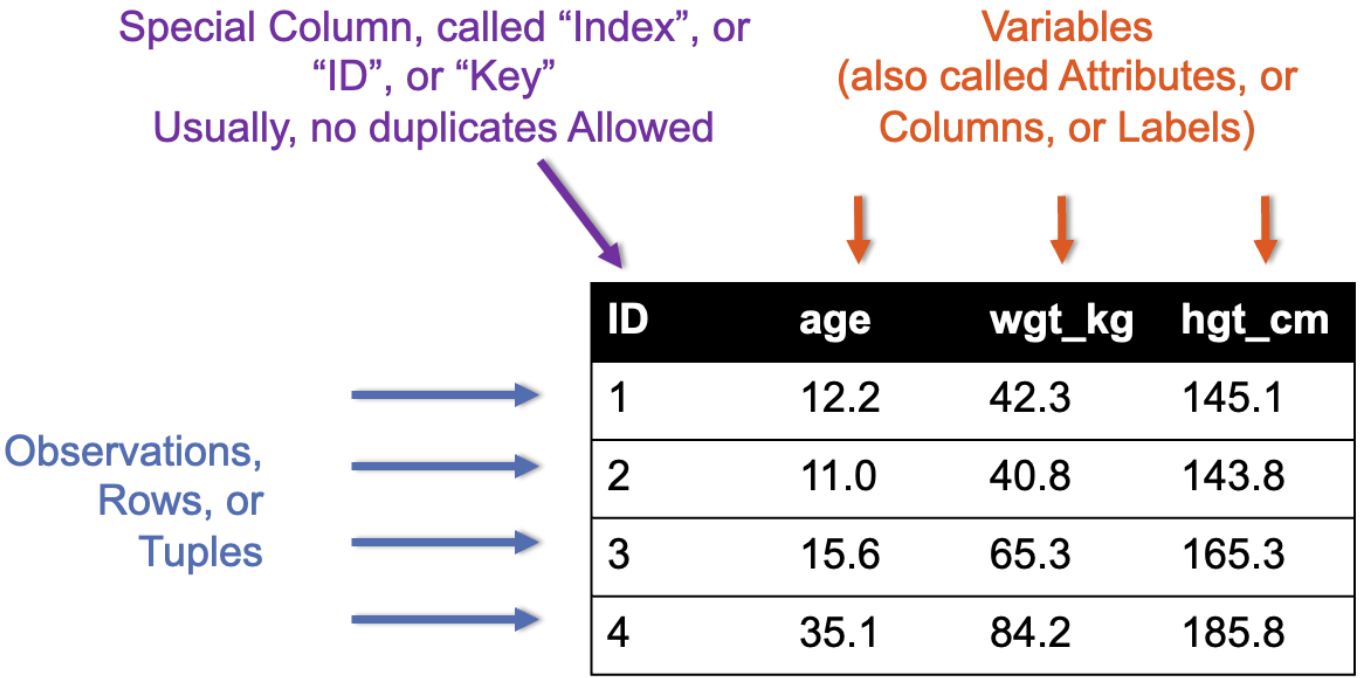
Importance

- **Python**: It's a user-friendly language for data analysis.
- **Git**: Helps manage code and data changes when working with a team.
- **Pandas**: Simplifies data cleaning and manipulation.
- **Databases**: Needed for storing and retrieving data efficiently.

These skills are essential for effective data analysis, collaboration, and handling data in real-world projects.

Basic Concept Review: Table / Tabular Data

In tabular data, information is organized into rows and columns.



ID	age	wgt_kg	hgt_cm
1	12.2	42.3	145.1
2	11.0	40.8	143.8
3	15.6	65.3	165.3
4	35.1	84.2	185.8

Tabular data is crucial for both Pandas and SQL

Provides a structured and organized way to represent, manipulate and visualize data.

- Both pandas and SQL are commonly used for working with tabular data, making it essential for tasks such as data analysis, manipulation, and visualization.

Next

Pandas

Class Colab: https://colab.research.google.com/drive/1DIJXtoeaHndPiykVUJg0CUqj-RZxtj2T?usp=sharing

Dataset: https://drive.google.com/file/d/1au5qJnRR57pOeAt4aefzxwsmG02JyaR_/view?usp=sharing

Pandas: Python Library for Manipulating Tabular Data

"Pandas": Short for "Python Data Analysis Library," used for data analysis.

- Written by Wes McKinney: Started in 2008 to get a high-performance, flexible tool to perform quantitative analysis on financial data

Pandas: Python Library for Manipulating Tabular Data

"Pandas": Short for "Python Data Analysis Library," used for data analysis.

- Open source, free to use
- Analyze data locally that can fit in memory
- Almost entirely functional
- Integrates with most major frameworks in data pipelines (e.g scikit-learn, numpy etc.).
- Compatible with various data formats: CSV, Excel, XML, JSON, and relational databases.

Pandas First Steps: install and import

- Pandas is an easy package to install. Open up your terminal program (shell or cmd) and install it using either of the following commands:

```
$ conda install pandas
        OR
$ pip  install  pandas
```

- For **`jupyter notebook`** users, you can run this cell:
  - The `!` at the beginning runs cells as if they were in a terminal.

```
!pip install pandas
```

- For Google Colab: already pre-installed; just import it
- To import pandas we usually import it with a shorter name since it's used so much:

```
import pandas as pd
```


Installation: https://pandas.pydata.org/pandas-docs/stable/getting_started/install.html

Core components of Pandas:
Series & DataFrames

Pandas: Data Table Representation

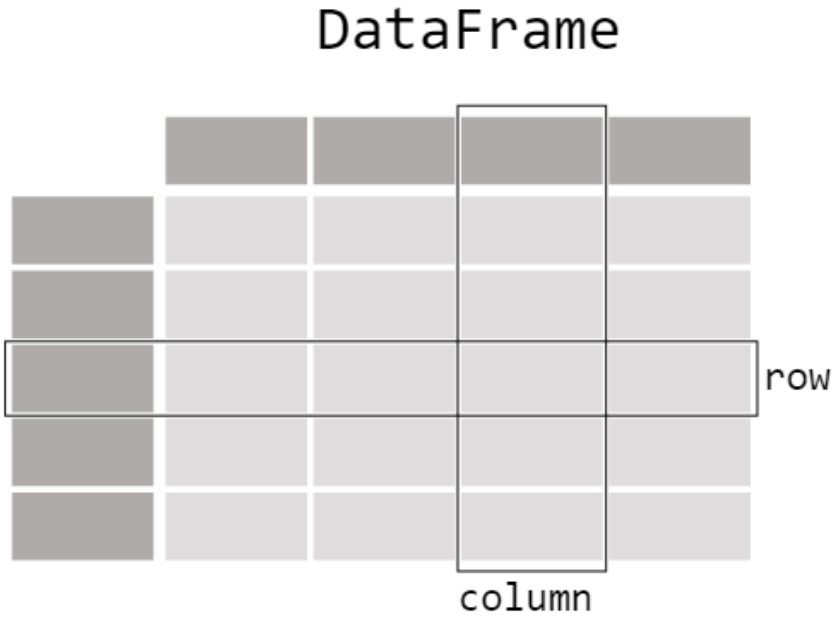
Data in Pandas is stored in two fundamental data types:

- Series is the data type that stores a single column of data.
- DataFrame is the data type that stores an entire dataset.

Notes: They both behave very similar to numpy arrays, so you can use them as input to anything in numpy, scipy, or scikit-learn. The main difference from numpy arrays is indexing.

Pandas: Data Table Representation cont.

In pandas, the most basic object is the DataFrame, that is a collection of Series objects. Each column in a DataFrame is essentially a Series.



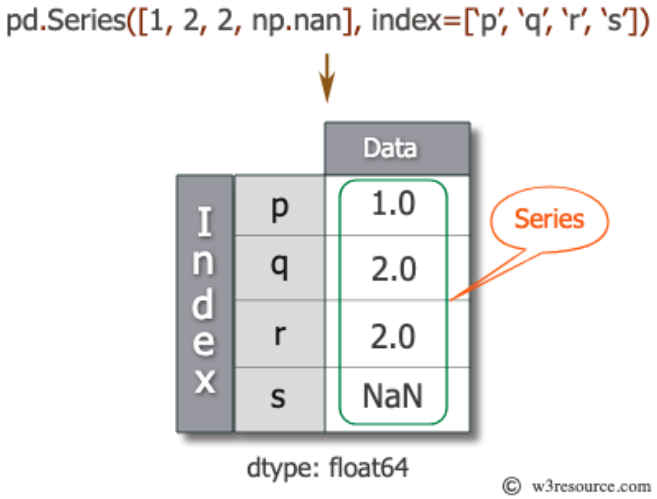
(Series)

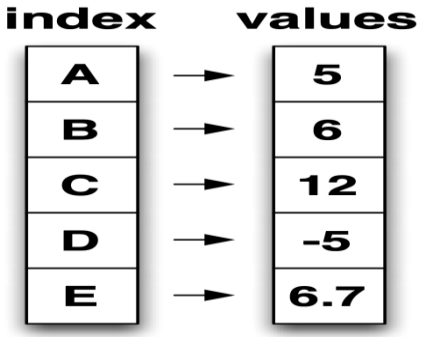
The primary two components of pandas are the Series and DataFrame.

- Series is essentially a column, and
- DataFrame is a multi-dimensional table made up of a collection of Series.

Series

- An one dimensional array of data with an index
- Subclass of numpy.ndarray
- **Data:** any type
- **Index:** most commonly integers, start with 0 (default).
    - Index labels need not be ordered
    - We almost NEVER access a series by an index
- Duplicates possible but result in reduced functionality, may produce inaccurate result.





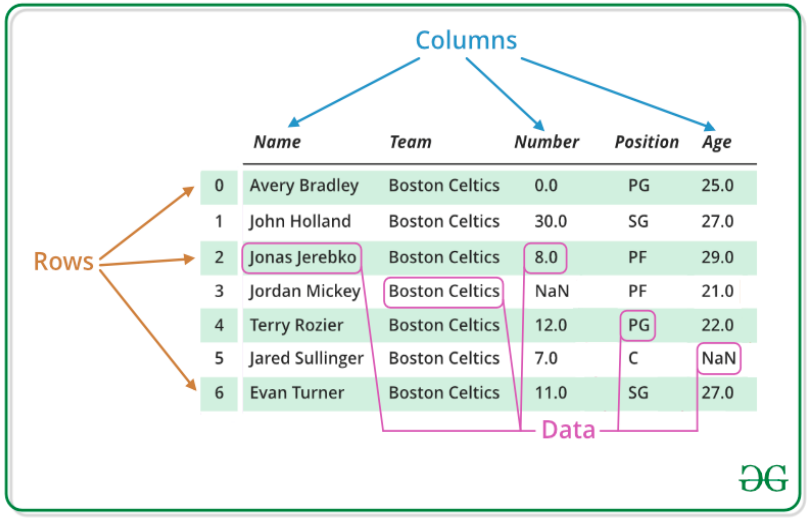
Series Functions

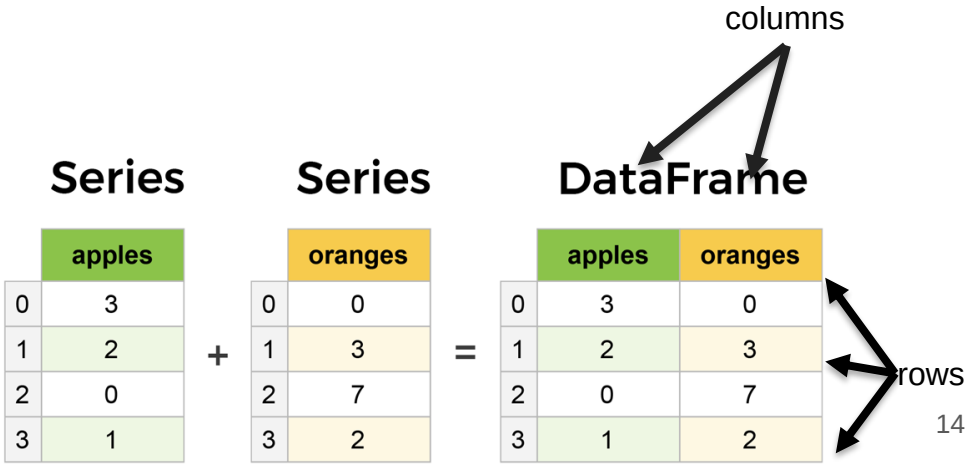
- Pandas can do anything with a series you can do with an array
  - Sum, count, show unique, etc
  - If you have a basic thing you want to do with a series, look it up!
- apply()
  - Apply takes a function as an argument and returns a series with that function applied
- You can also convert a series into a series of booleans by applying a logical condition (which will be useful later)

Dataframes

A collection of series organized by columns

- Each column can have a different type (integers, floats, strings, etc.)
- Mutable size: insert and delete columns
- Can Perform Arithmetic operations on rows and columns
- We do not access rows via index (less common).
- We DO access columns via their names



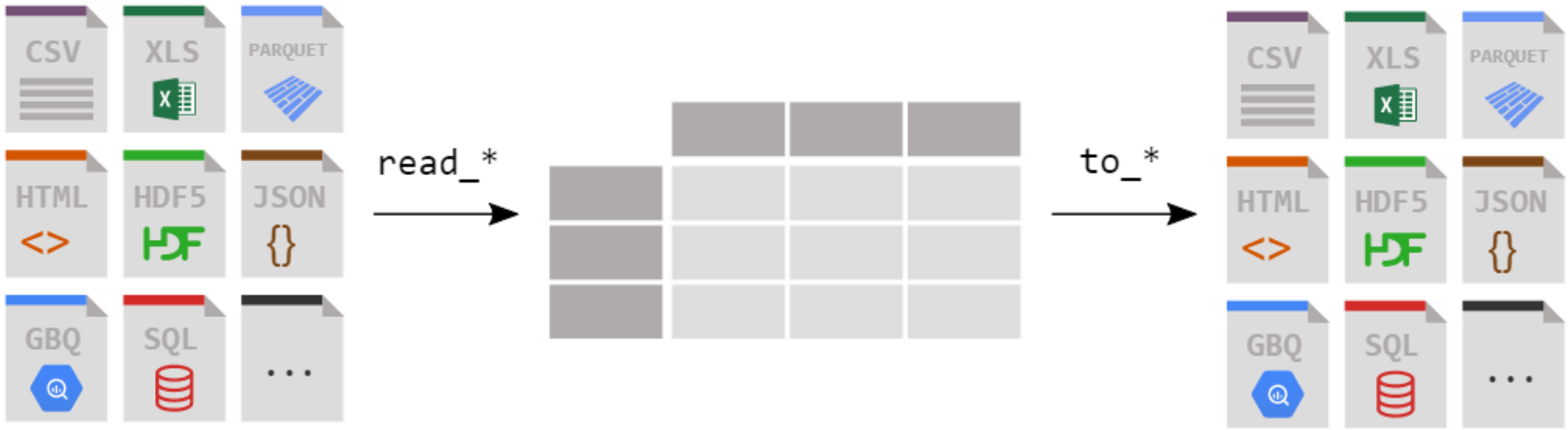


Dataframe Functions

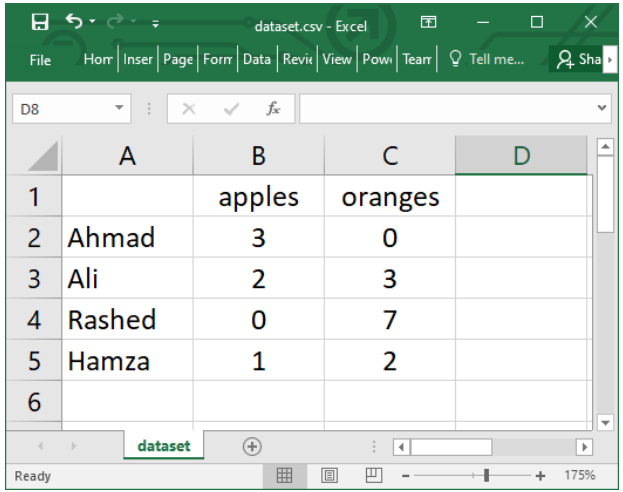
Important functions:

- How to access a column
- How to filter a column
- Apply (LATER TOPIC)
- GroupBy

Loading a DataFrame from files



Load a CSV file into a Pandas DataFrame:



	A	B	C	D
1		apples	oranges	
2	Ahmad	3	0	
3	Ali	2	3	
4	Rashed	0	7	
5	Hamza	1	2	
6				

Reading data from a CSV file

File Edit Format Run Options Window Help

```
1 import pandas as pd
2
3 df = pd.read_csv('dataset.csv')
4 print(df)
5
```

- With CSV files, all you need is a single line to load in the data:

```
df = pd.read_csv('dataset.csv')
```

	Unnamed: 0	apples	oranges
0	Ahmad	3	0
1	Ali	2	3
2	Rashed	0	7
3	Hamza	1	2

Viewing data

Use DataFrame.head() and DataFrame.tail() to view the top (first) and bottom (last) few rows of the frame respectively.

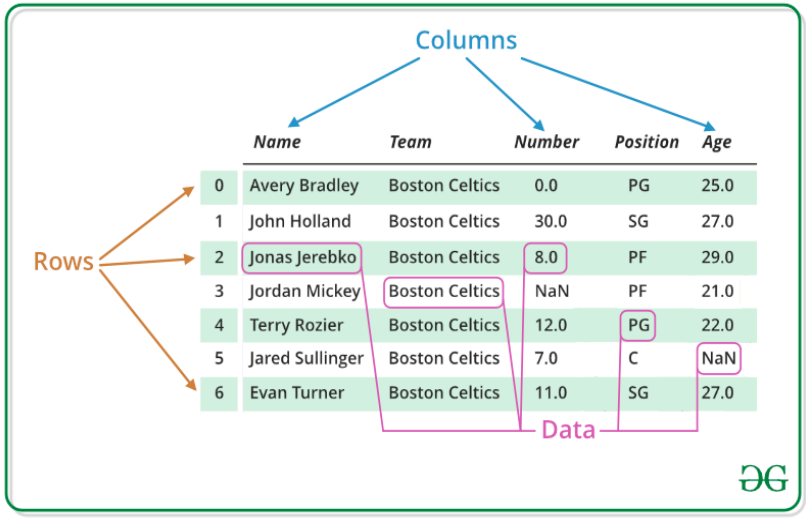
- Default value is 5, unless specified otherwise.

How to access a column

dataframe['column_name']: Access a column by its name.

E.g. : df['age']

Where df=your dataframe



Arithmetic Operations

Can perform arithmetic operations (e.g., addition, subtraction, multiplication, division) on entire columns or between columns.

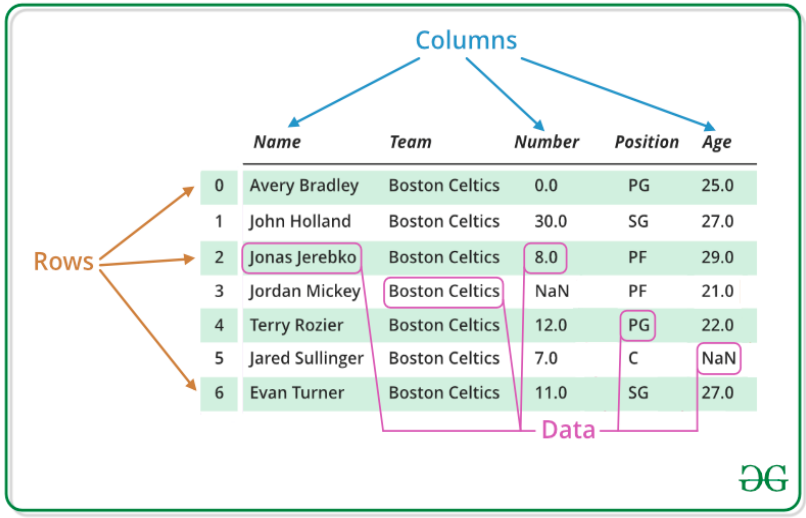
E.g.:

```
df['new_column'] = df['column1'] + df['column2']
df['new_column'] = df['column1'] - df['column2']
```

Applying Functions Directly to the dataframe

You can apply different functions directly to a DataFrame column:

Syntax: **dataframe[column_name].function()**

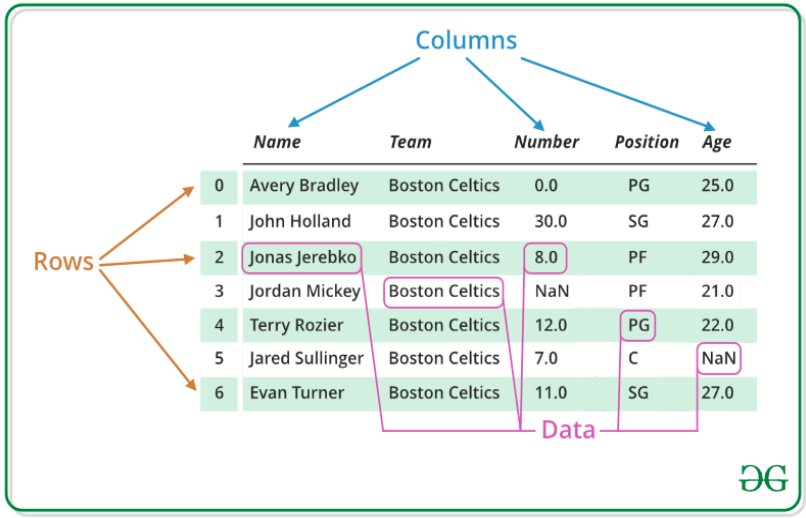


Applying Functions Directly to the dataframe

Aggregation operations are used to compute summary statistics or single values from multiple values in a dataset. E.g:

E.g.: df[“age”].sum()

Try some other aggregation operation!



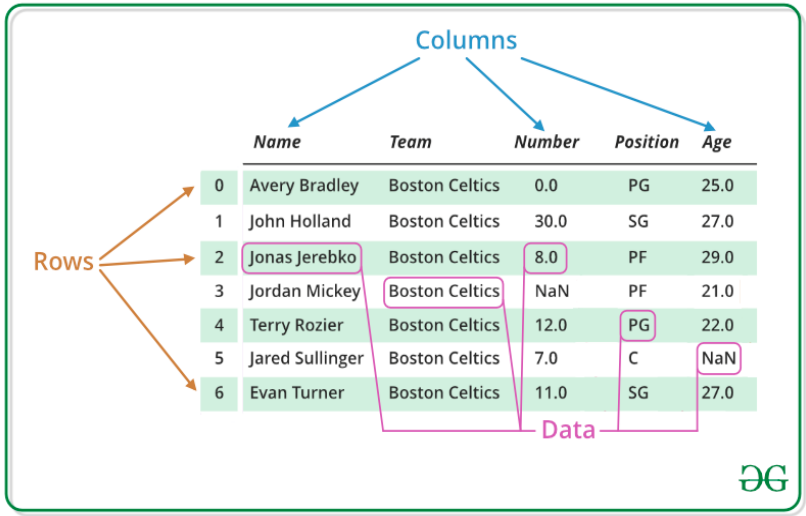
How to Filter a Column based on Condition (s)

To filter a column in a pandas DataFrame, we can use boolean indexing.

E.g.: Filter only the 'Name' column for values equal to 'Jonas Jerebko'

dataframe[‘column name’][boolean condition]

df[‘Name’][df[‘Name’]=='Jonas Jerebko']



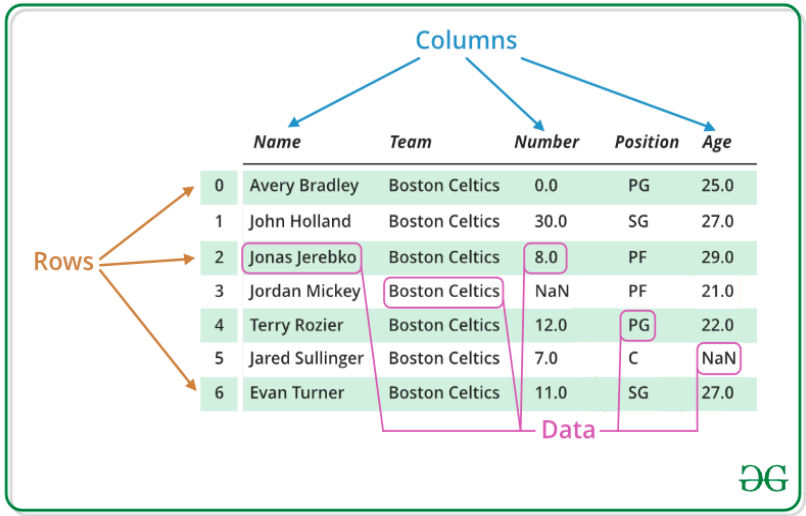
How to Filter Rows based on Condition (s)

E.g.: Filter rows where age is greater than 25.

dataframe [boolean condition]

df [df['Age']>29]

Q: What about multiple condition?



More Advanced: Filtering Data & Apply Statistical Functions

We can filter rows in a DataFrame based on a condition and then apply statistical functions to a specific column:

```
df[df['condition']][column_name].statistics_function()
```

- **Filters the DataFrame** based on the condition specified within the brackets.
- It **selects only the rows** that satisfy the condition.

- Selects the **specific column** from the filtered DataFrame obtained in left, resulting in a Series containing only the values from that column.

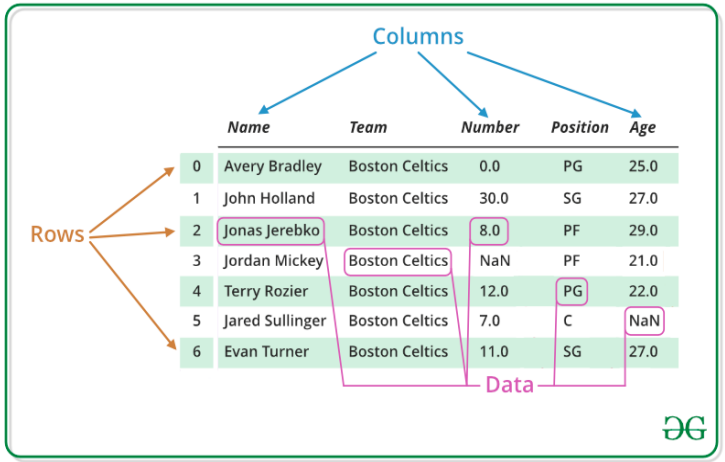
- **Applies** a statistical function to the Series obtained in the left to compute a summary statistic.

More Advanced: Filtering Data & Apply Statistical Functions

We can filter rows in a DataFrame based on a condition and then apply statistical functions to a specific column:

df[df['condition']][column_name].statistics_function()

Try: Calculate the mean of the 'Number' column where the 'Age' is greater than 25



Grouping Data (Groupby)

You can group data by a condition and calculate statistics within each group.

- This process is commonly referred to as "split-apply-combine."

```
df.groupby('condition')[column_name].statistics_function()
```

Splitting

- Groups the DataFrame based on the values in the 'condition' column.
- **Splits the DataFrame** into groups based on unique (identical) values in the specified column.

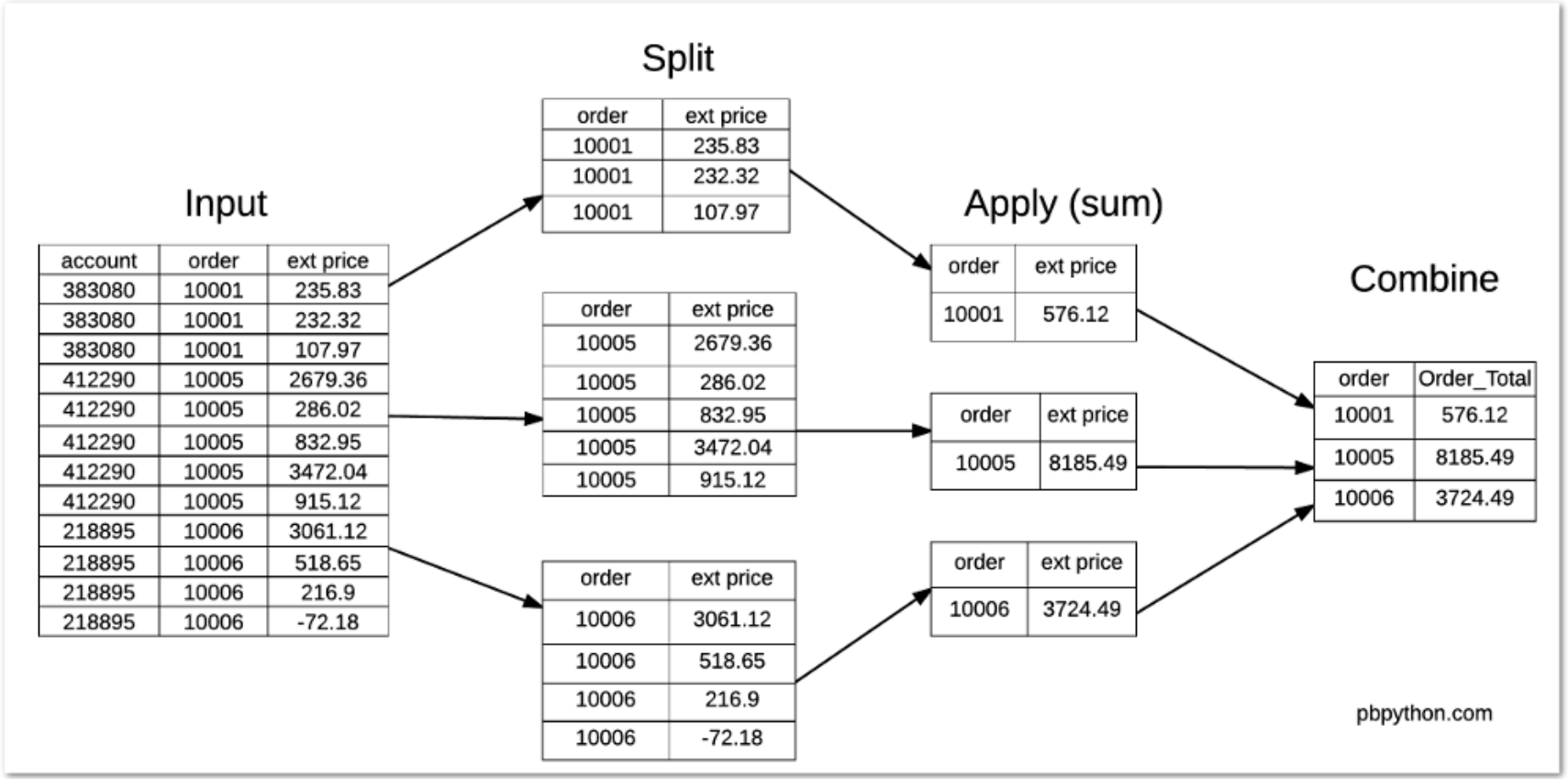
Apply

- Selects a specific column from each group, resulting in a Series containing the values from that column within each group..

Combine

- **Applies** a statistical function to each group of values in the selected column.

Grouping Data (Groupby)



Write data to a CSV (Comma Separated Values) file

For a Series s: s.to_csv("filename.csv")

For a DataFrame df: df.to_csv("filename.csv")

Pandas: Advantages

- A powerful python library for data tasks like analysis and cleaning.
- Simplifies data representation for better understanding.
- Cleans messy datasets for readability and relevance.
- Increases productivity by minimizing writing.
- Offers extensive features for easy data analysis.

References

- pandas documentation
  - https://pandas.pydata.org/pandas-docs/stable/index.html
- pandas: Input/output
  - https://pandas.pydata.org/pandas-docs/stable/reference/io.html
- pandas: DataFrame
  - https://pandas.pydata.org/pandas-docs/stable/reference/frame.html
- pandas: Series
  - https://pandas.pydata.org/pandas-docs/stable/reference/series.html
- pandas: Plotting
  - https://pandas.pydata.org/pandas-docs/stable/reference/plotting.html

Next

Databases & SQL

Class Colab:
https://colab.research.google.com/drive/1XuwWC_SN7RCKs7f8ZU6gxNzFg8ca2Cc7?usp=sharing

Databases

Many tasks in data science involve extracting and manipulating data from a database.Most leading database systems are relational, organizing data in rows and columns of tables.

- Databases are used when:
  - Dealing with large amount of dataset
    - Our data won’t fit in memory
  - We want to access particular parts of our data extremely quickly

Databases

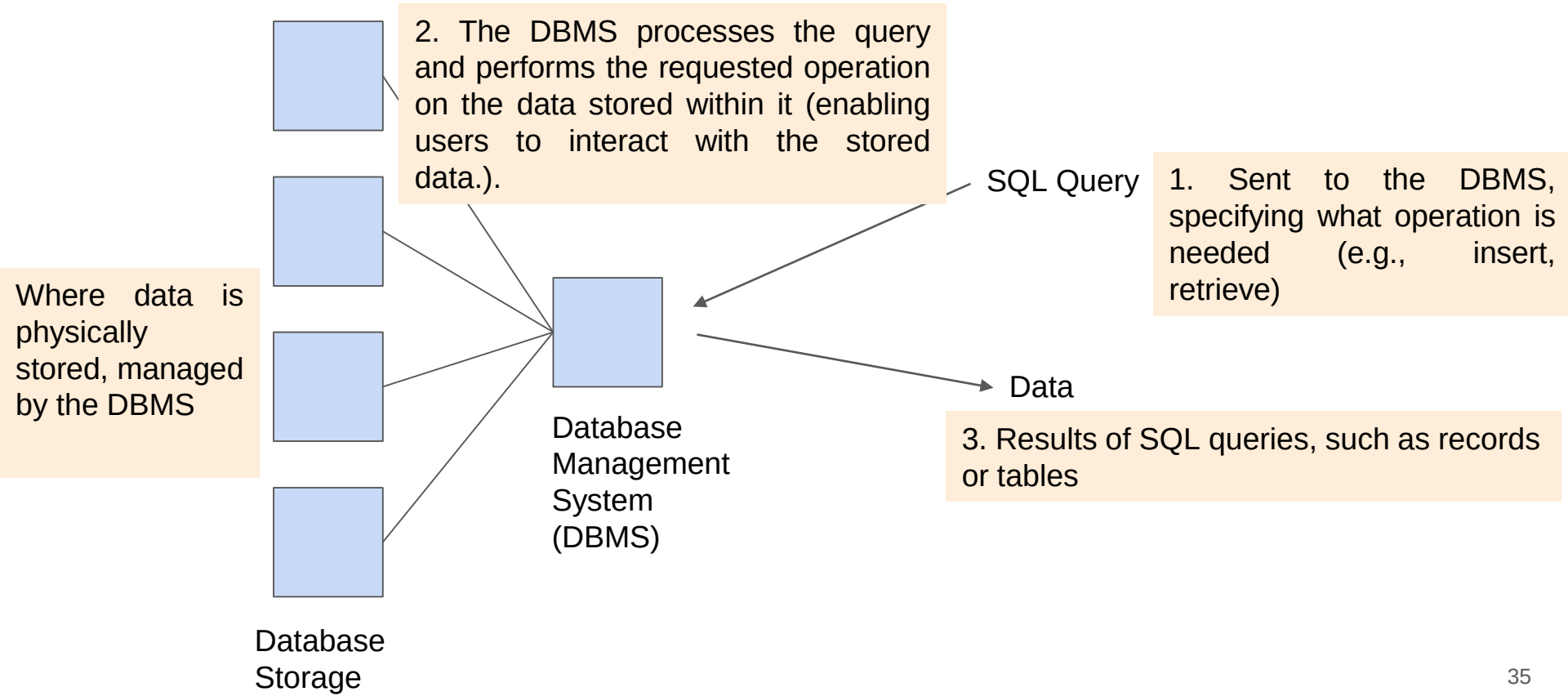
For example, imagine the table on the right contains ten million customers, and we would like to grab every customer in Berlin quickly.

- Allows efficient storage and management of the data on disk.
- Allows retrieval of relevant data without the need to load the entire dataset into memory.

Number of Records: 91

CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden
6	Blauer See Delikatessen	Hanna Moos	Forsterstr. 57	Mannheim	68306	Germany
7	Blondel père et fils	Frédérique Citeaux	24, place Kléber	Strasbourg	67000	France
8	Bólido Comidas preparadas	Martín Sommer	C/ Araquil, 67	Madrid	28023	Spain
9	Bon app'	Laurence Lebihans	12, rue des Bouchers	Marseille	13008	France
10	Bottom-Dollar Marketse	Elizabeth Lincoln	23 Tsawassen Blvd.	Tsawassen	T2F 8M4	Canada
11	B's Beverages	Victoria Ashworth	Fauntleroy Circus	London	EC2 5NT	UK
12	Cactus Comidas para llevar	Patricio Simpson	Cerrito 333	Buenos Aires	1010	Argentina
13	Centro comercial Moctezuma	Francisco Chang	Sierras de Granada 9993	México D.F.	05022	Mexico
14	Chop-suey Chinese	Yang Wang	Hauptstr. 29	Bern	3012	Switzerland
15	Comércio Mineiro	Pedro Afonso	Av. dos Lusíadas, 23	São Paulo	05432-043	Brazil

Database Setup (Very Simplified)



SQL

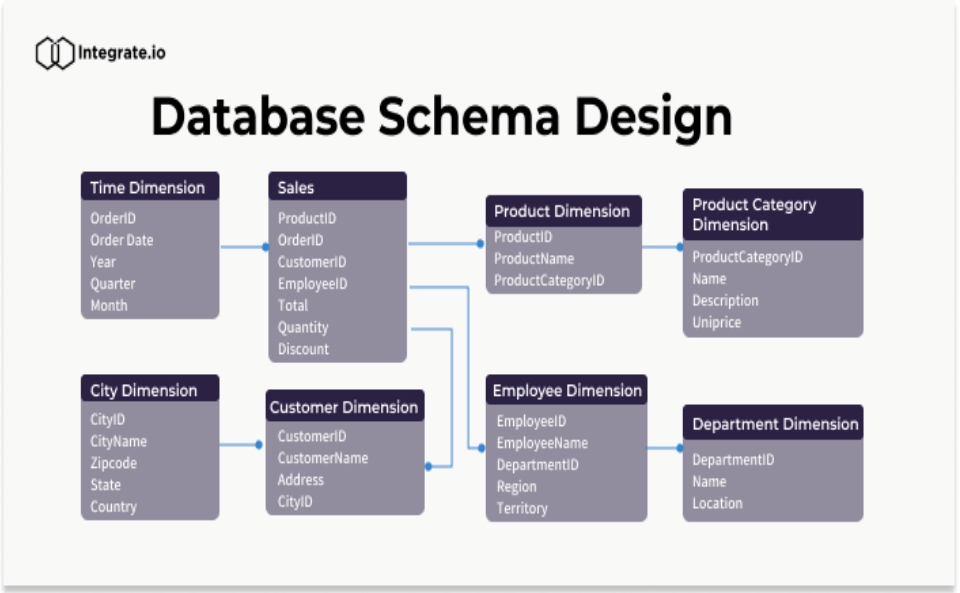
SQL is a computer language for manipulating and retrieving data. It is the standard language for relational databases and is also supported by many non-relational databases. SQL is pronounced either "S-Q-L " or "seekwəl ".

- SQL is how we ask our database questions
- Stands for **Structured Query Language**
- It’s not a programming language
- For example, if we want to ask for all customers from Berlin:

```
SELECT * FROM customers WHERE City = ‘Berlin’
```

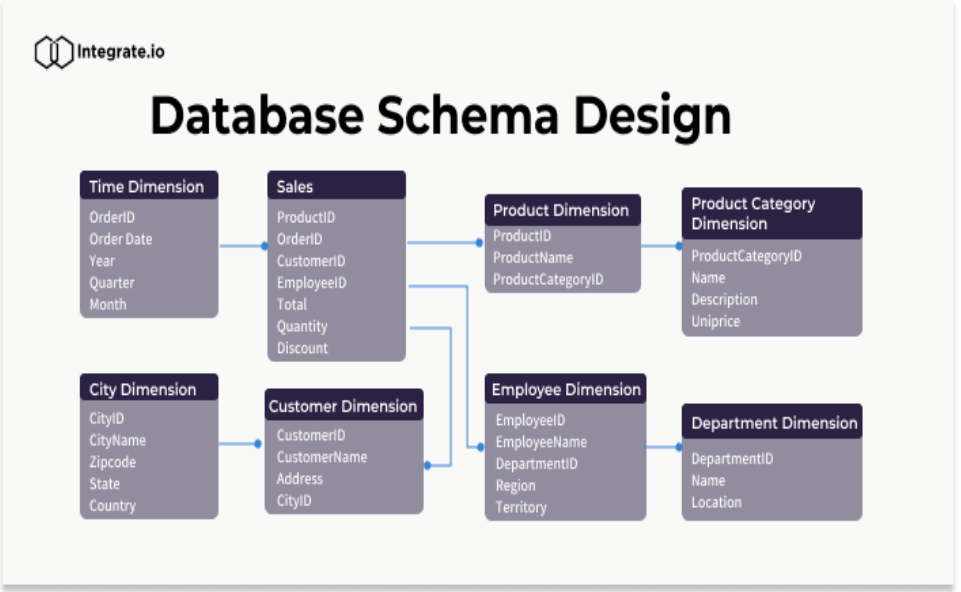
Important Vocab

- Databases store data in **tables**, which are similar to DataFrames in Pandas.
  - Organized into rows and columns
  - Has an **index,** enable quicker data retrieval and faster querying
- A **database schema defines the structure of the database:**
  - including tables, their columns, data types, relationships, and constraints
  - Outline how the data is organized

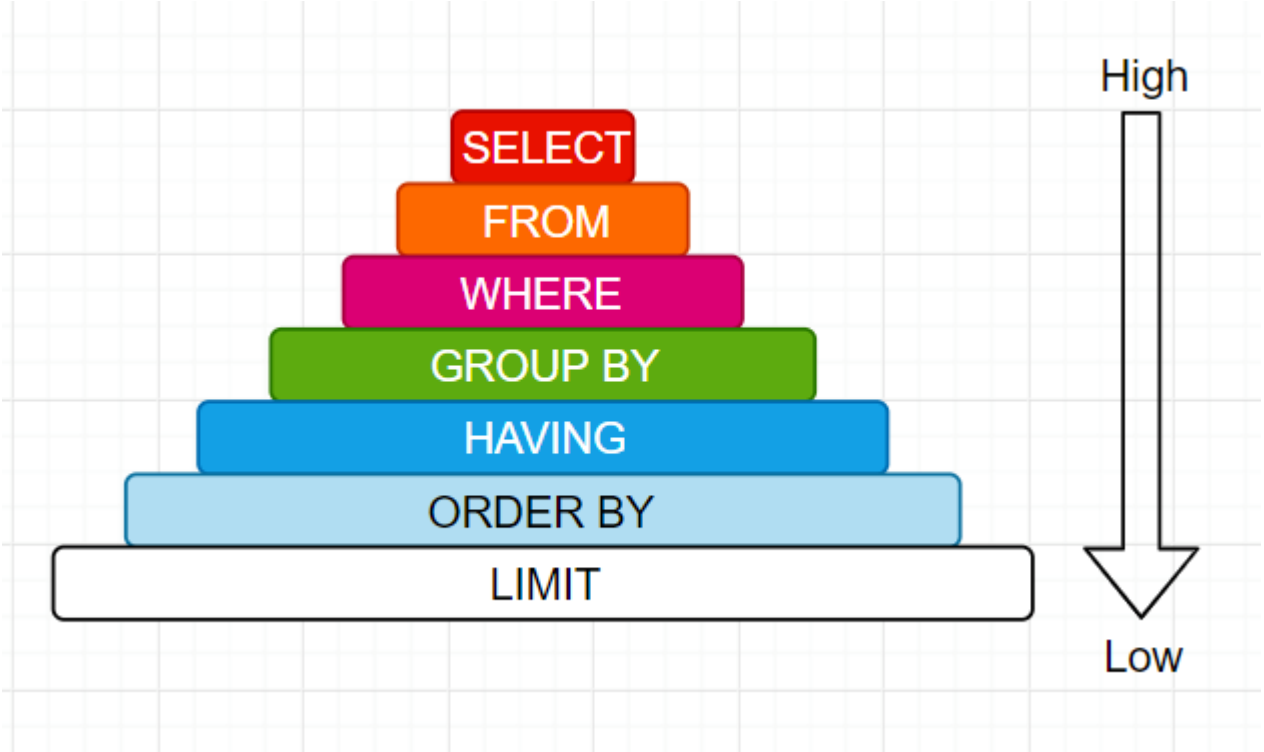


Important Vocab

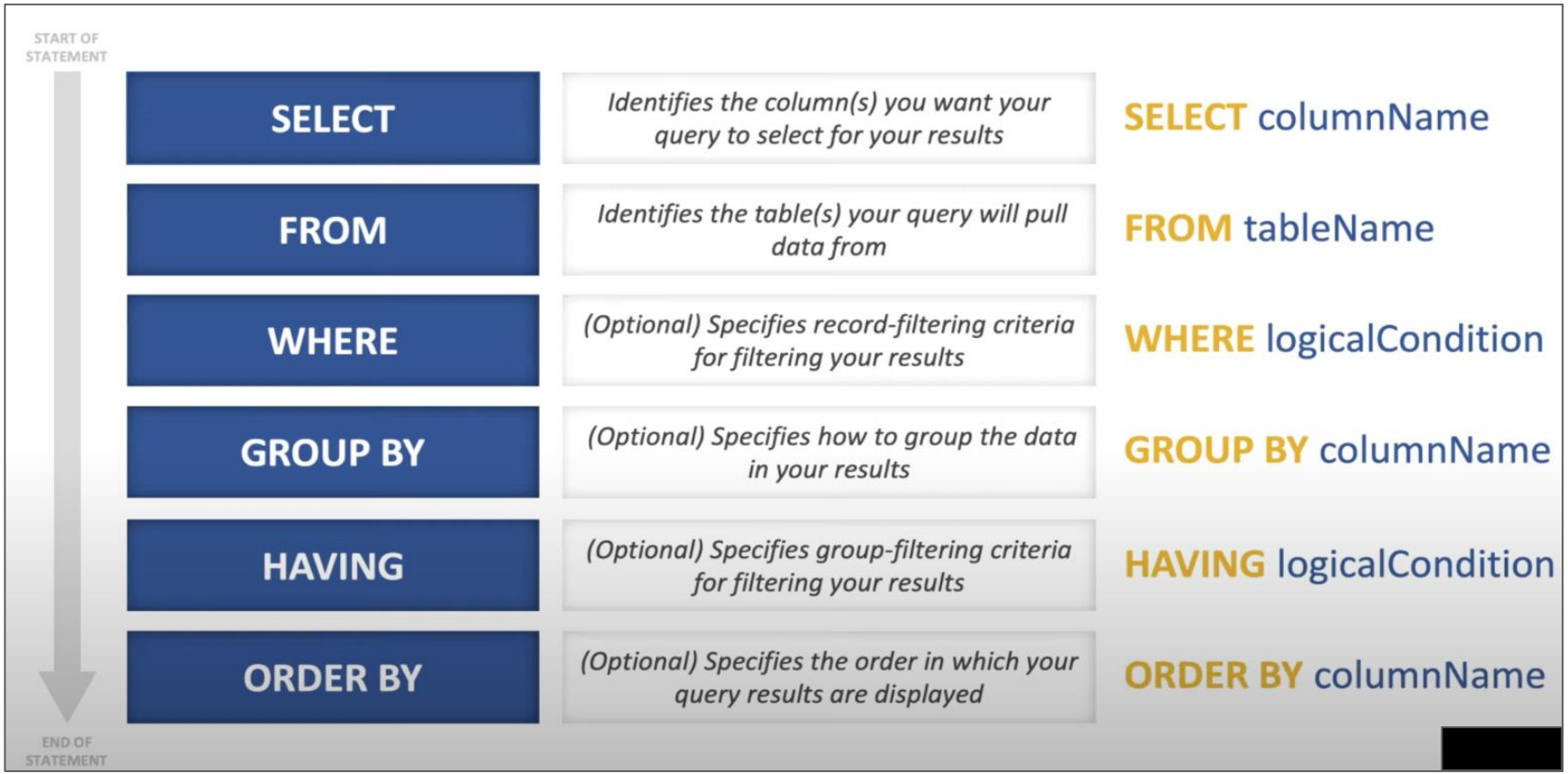
- Two tables may often have the same field despite having different information
  - E.g, a 'customer' table and an 'order' table may both contain a 'customer_id' field to establish a relationship between them.



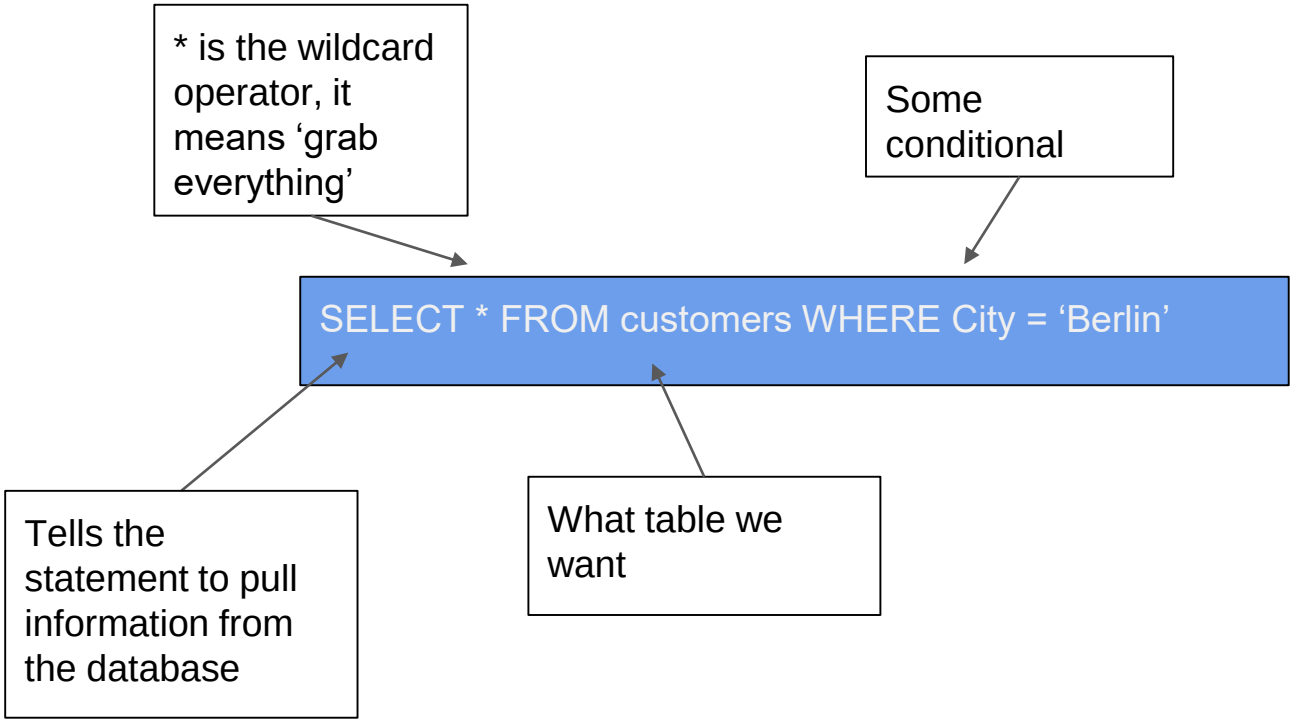
The “Big 6” Elements of a SQL Select Statement



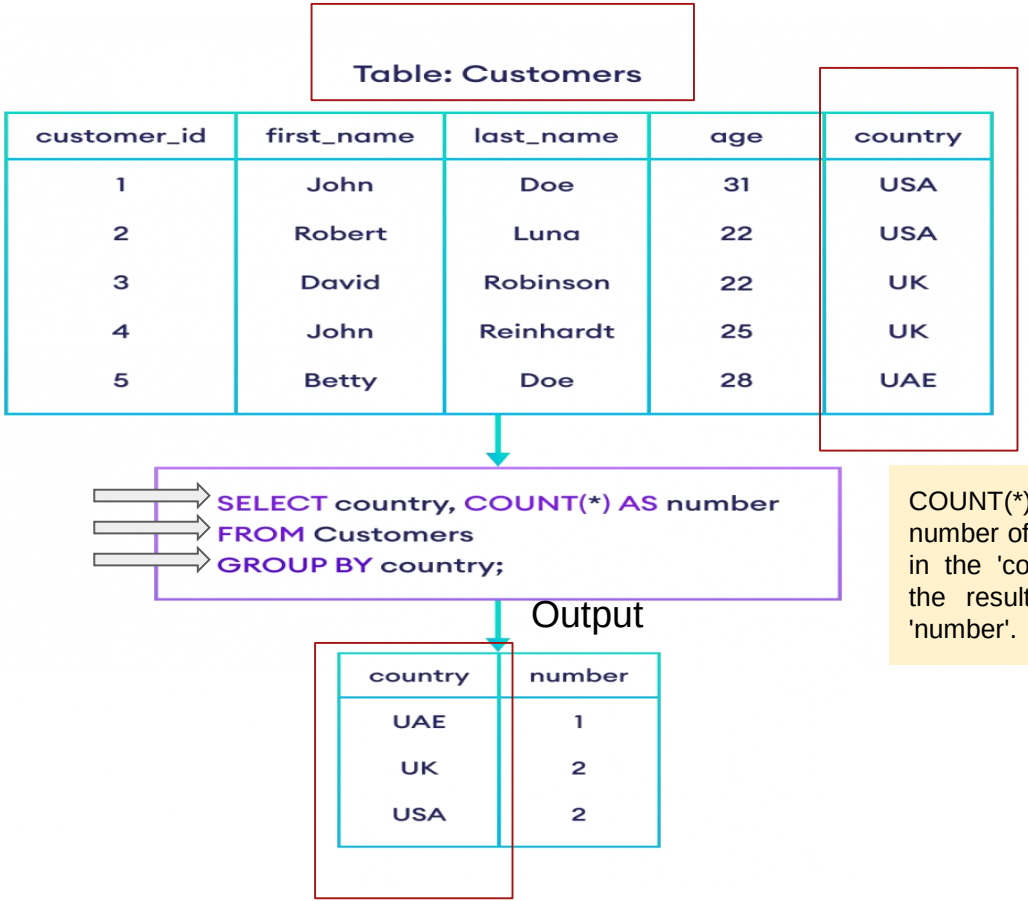
The “Big 6” Elements of a SQL Select Statement



How a query is structured (Where condition)

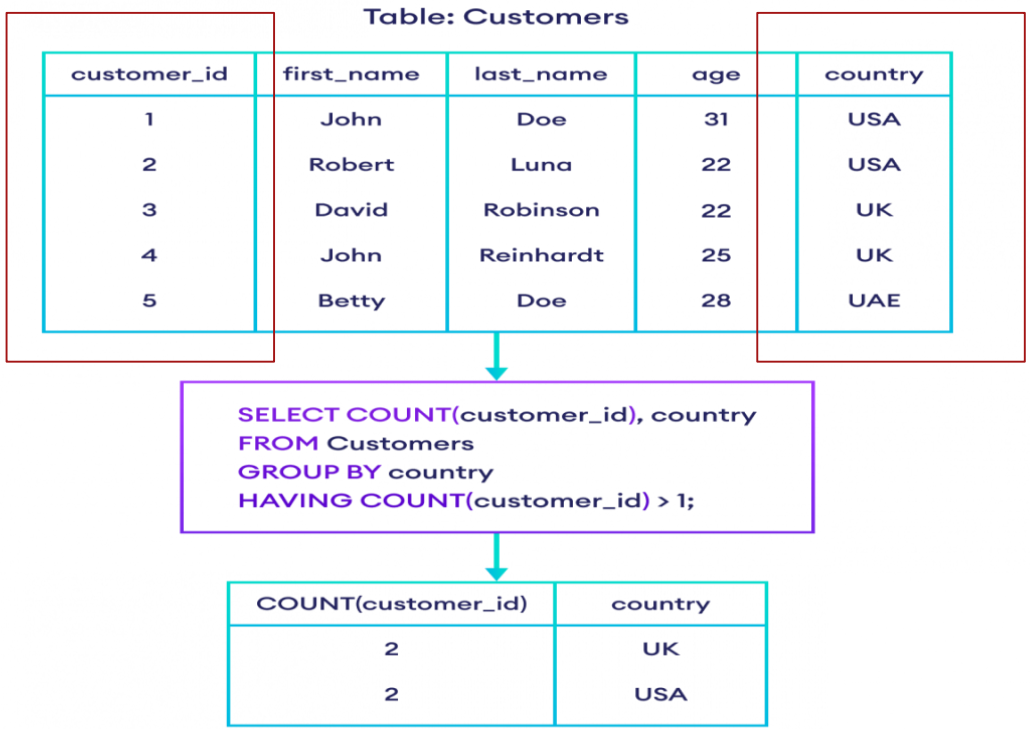


How a Query is Structured (Group By)



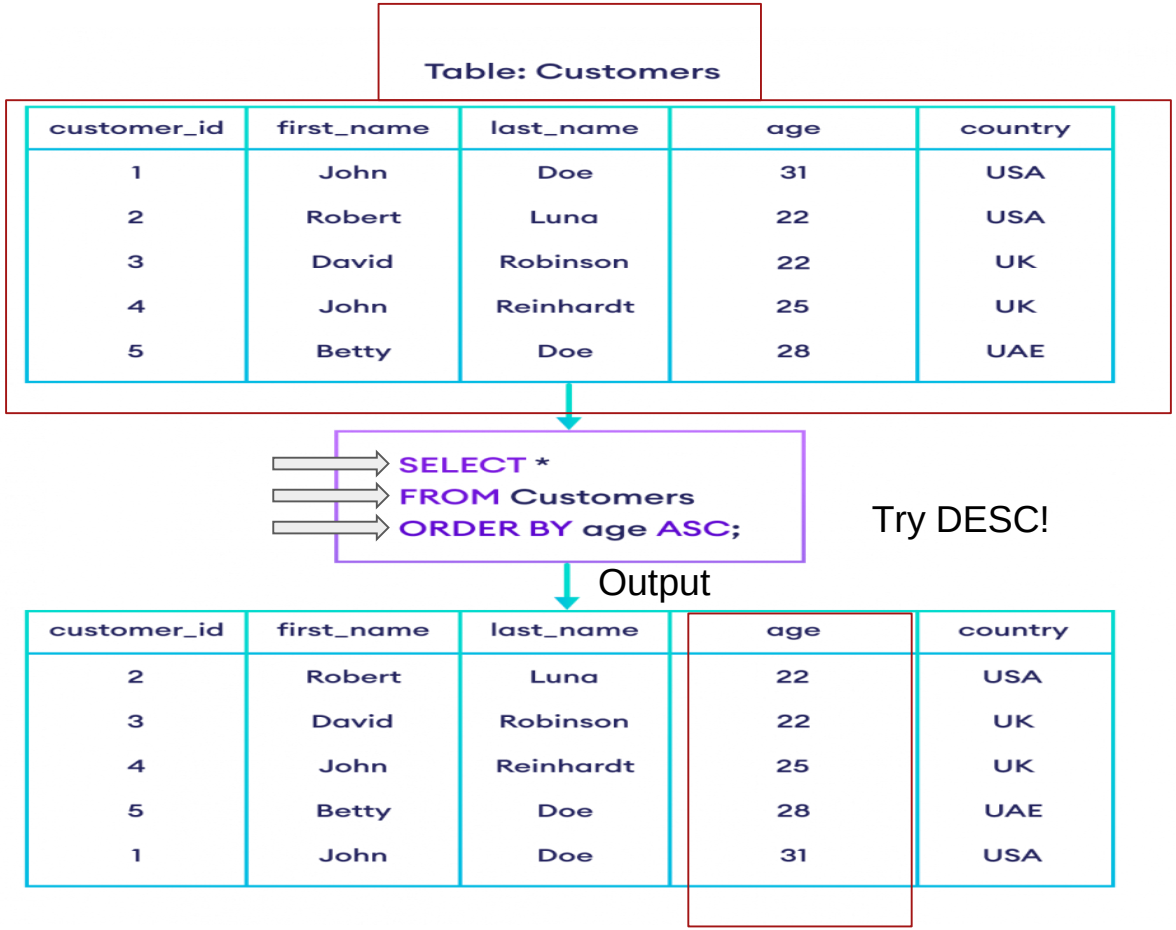
COUNT(*) AS number: counts the number of rows for each unique value in the 'country' column and assigns the result to a new column alias 'number'.

How a query is structured (Group by with Having)



Note: The HAVING clause was introduced because the WHERE clause does not support aggregate functions. Also, GROUP BY must be used before the HAVING clause.

How a query is structured (Order by)



How a query is structured (Order by DESC (Descending Order)

Table: Customers

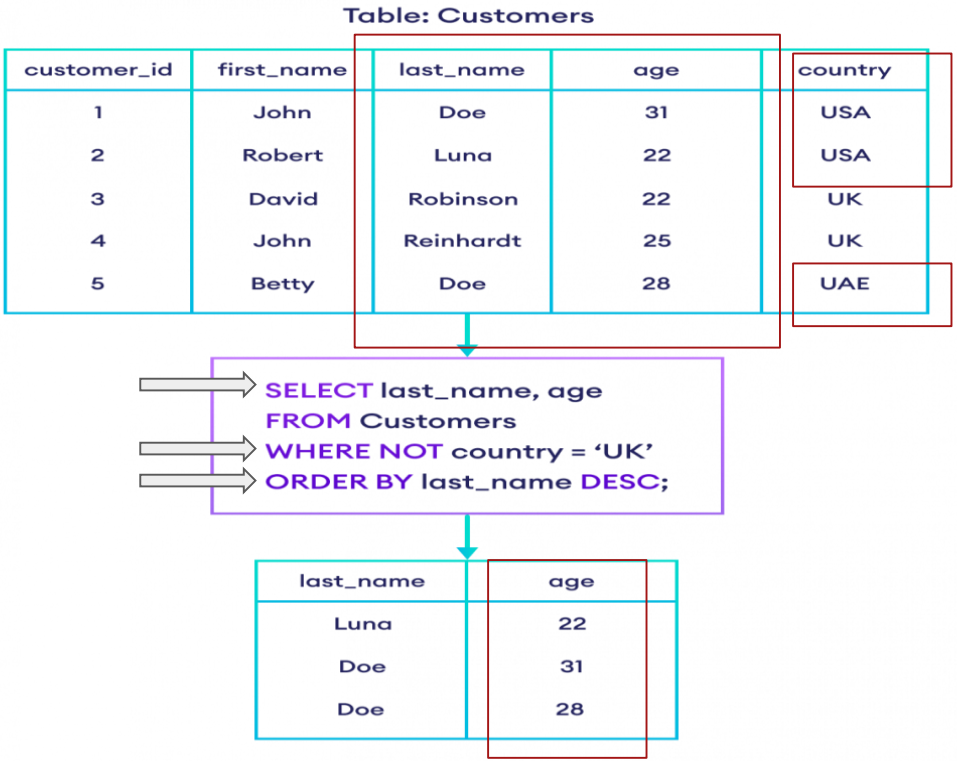
customer_id	first_name	last_name	age	country
1	John	Doe	31	USA
2	Robert	Luna	22	USA
3	David	Robinson	22	UK
4	John	Reinhardt	25	UK
5	Betty	Doe	28	UAE

```
SELECT *
FROM Customers
ORDER BY age DESC;
```

Output

customer_id	first_name	last_name	age	country
1	John	Doe	31	USA
5	Betty	Doe	28	UAE
4	John	Reinhardt	25	UK
2	Robert	Luna	22	USA
3	David	Robinson	22	UK

How a query is structured (Order by with WHERE)



How a query is structured for aggregate fuction (e.g. SUM)

Table: Orders

order_id	item	amount	customer_id
1	Keyboard	400	4
2	Mouse	300	4
3	Monitor	12000	3
4	Keyboard	400	1
5	Mousepad	250	2

```
SELECT SUM(amount) AS total_sales
FROM Orders;
```

total_sales
13350

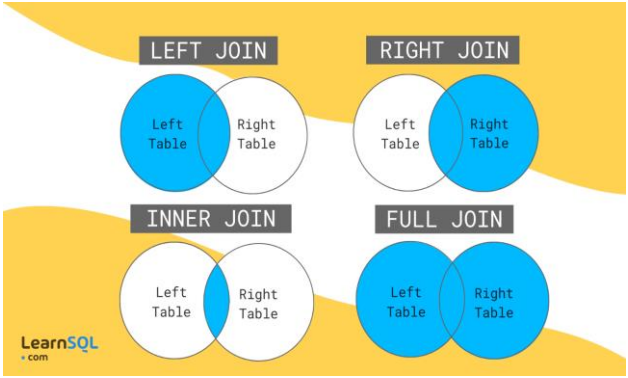
column new name “total_sales”

```
SELECT SUM(column_name)
FROM table;
```

Joins

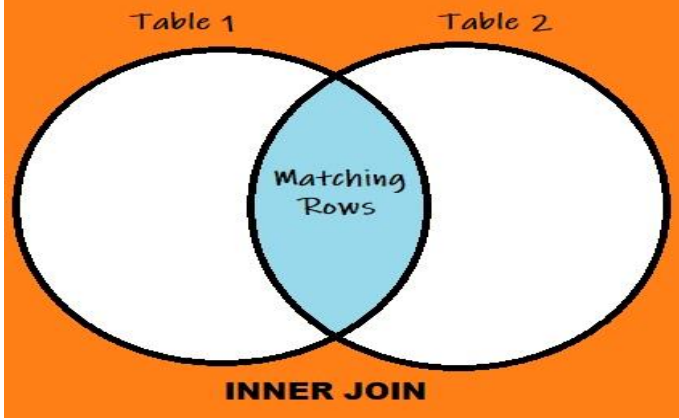
In relational databases, data from multiple tables is combined to generate insights. This is achieved through multi-table queries written with join statements.

- A join is a SELECT statement that combines data (rows) from two tables, known as the left table and right table, into a single result.
- The tables are combined by comparing values in the specified columns from both tables, usually using the = operator, to match corresponding values.
  - **Important: the columns being compared must have compatible data types for the comparison to be valid**



We will focus on: INNER JOIN, which selects only matching left and right table rows.

How a query is structured (INNER JOIN)



Returns only the rows with matching values in both tables.

- INNER join **compares** each row of the first table with each row of the second table, to **find all pairs of rows** that satisfy the join-predicate.
- When the join-predicate is satisfied, the column values from both tables are combined into a new table.

Example: join-predicate :
EmpDetails.EmpID = MaritalStatus.EmpID

EmpDetails

ID	Name	Salary
1	John	40000
2	Alex	25000
3	Simon	43000

MaritalStatus

ID	Name	Status
1	John	Married
3	Simon	Married
4	Stella	Unmarried

ID	Name	Salary	Status
1	John	40000	Married
3	Simon	43000	Married

How a query is structured (INNER JOIN)

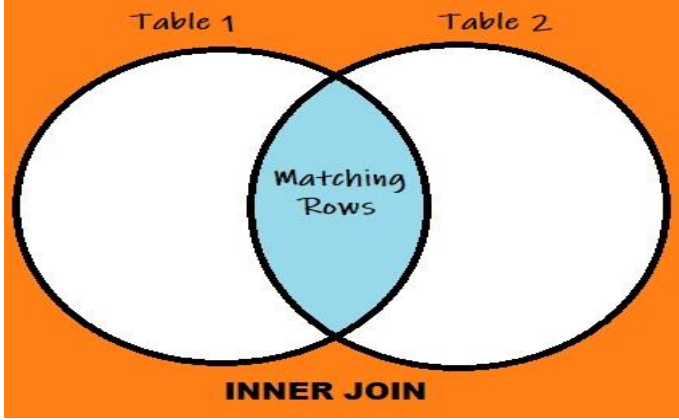


Table: Customers

customer_id	first_name
1	John
2	Robert
3	David
4	John
5	Betty

Table: Orders

order_id	amount	customer
1	200	10
2	500	3
3	300	6
4	800	5
5	150	8

customer_id	first_name	amount
3	David	500
5	Betty	800

```
SELECT columns_from_both_tables
FROM table1
INNER JOIN table2
ON table1.column1 = table2.column2
```

How a query is structured (INNER JOIN)

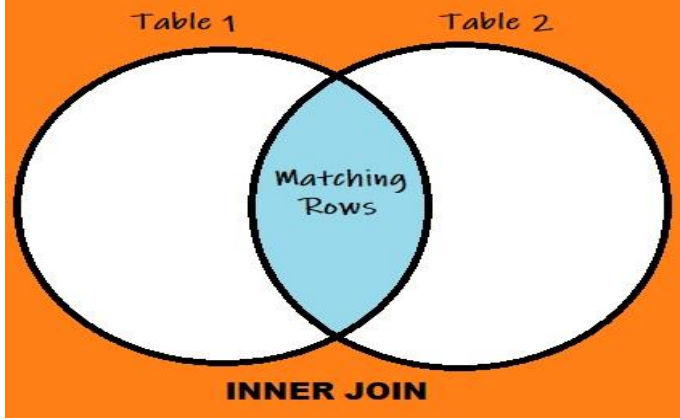


Table: Customers

customer_id	first_name
1	John
2	Robert
3	David
4	John
5	Betty

Table: Orders

order_id	amount	customer
1	200	10
2	500	3
3	300	6
4	800	5
5	150	8

customer_id	first_name	amount
3	David	500
5	Betty	800

```
Ex:
SELECT Customers.customer_id,
Customers.first_name, Orders.amount
FROM Customers
INNER JOIN Orders
ON Customers.customer_id =
Orders.customer;
```

- *You can also use * to select all columns*
- *Use “tablename.columnname” for selecting specific columns*

Summary

- **Python:** A versatile programming language used for data analysis, machine learning, and automation in the data science lifecycle.
- **Git:** A version control system that tracks changes in code and data, ensuring collaboration, reproducibility, and version management.
- **Pandas:** A Python library for data manipulation, cleaning, exploration, and preprocessing.
- **SQL and Databases:** SQL is used to manage, query, and manipulate structured data stored in relational databases, essential for data storage, retrieval, and organization.

Together, these tools and technologies enable data scientists to collect, clean, analyze, and manage data effectively throughout the data science lifecycle.

Additional Reading Slides

Sqllite03

The sqlite3 module is a powerful part of the Python standard library; it lets us work with a fully featured on-disk SQL database without installing any additional software.

- an open-source software, making it accessible and adaptable to various platforms and use case
- Making a connection between sqlite3 database and Python Program
  sqliteConnection = sqlite3.connect(("database.sqlite")
  crsr = conn.cursor() → Facilitate data retrieval, manipulation, and management

Documentation: https://www.sqlite.org/docs.html

DATA, MSML, BIOI 602 Principles of Data Science

Experimental Design

Lecture: 03

[Slides from Fardina Fathmiul Alam]

Today We will talk about

- What is Experimental Design
- Different Steps, associated variables etc. for experimental design
- Briefly: What is Hypothesis (LATER TOPIC)
- Methods for collecting data
- Many more….

Experimental design

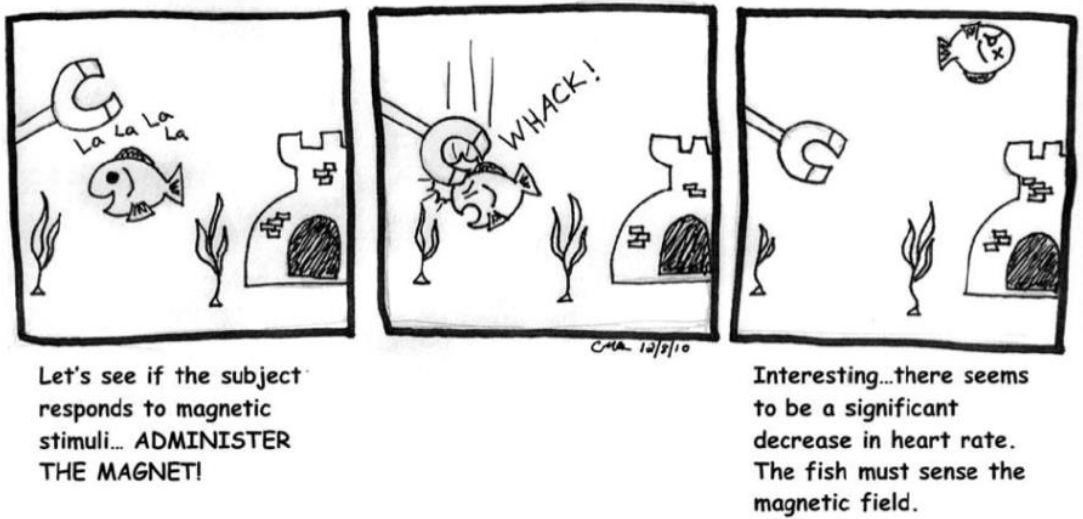
The science and subfield of statistics about how to collect data effectively…

R. A. Fisher (1890-1962)
Founding father of Modern Statistics
Geneticist

"There's a flaw in your experimental design.
All the mice are scorpios."

Experimental Design

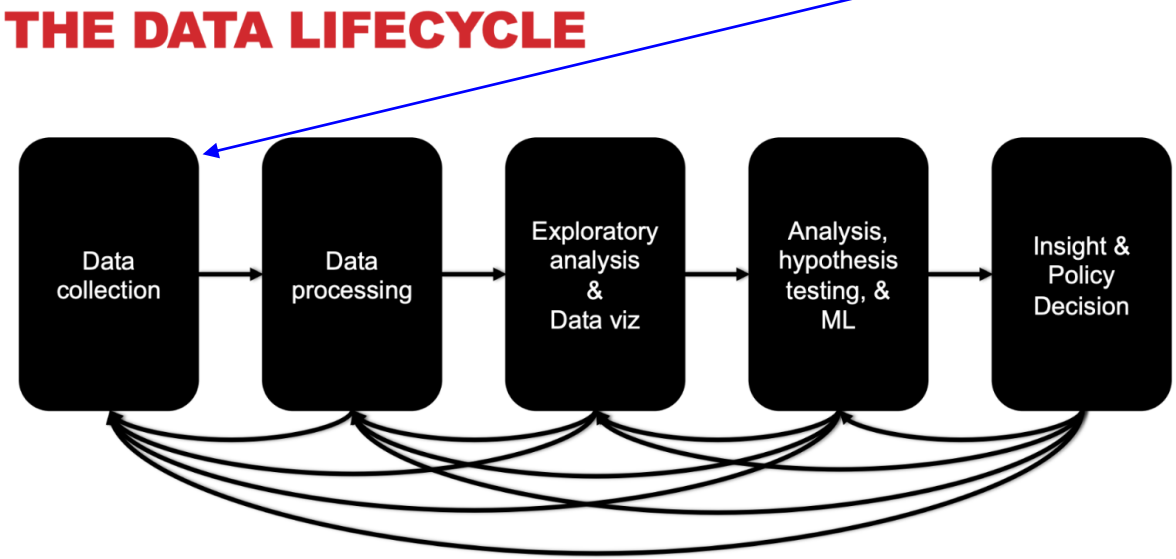
Experimental design is the process of planning, carrying out, and analyzing experiments to test a hypothesis.



From: http://www.hawaii.edu/fishlab/NearsideFrame.htm

Experimental Design in Data Science?

Experimental design is a crucial aspect of data science that focuses on planning and conducting experiments to gather **meaningful data**.

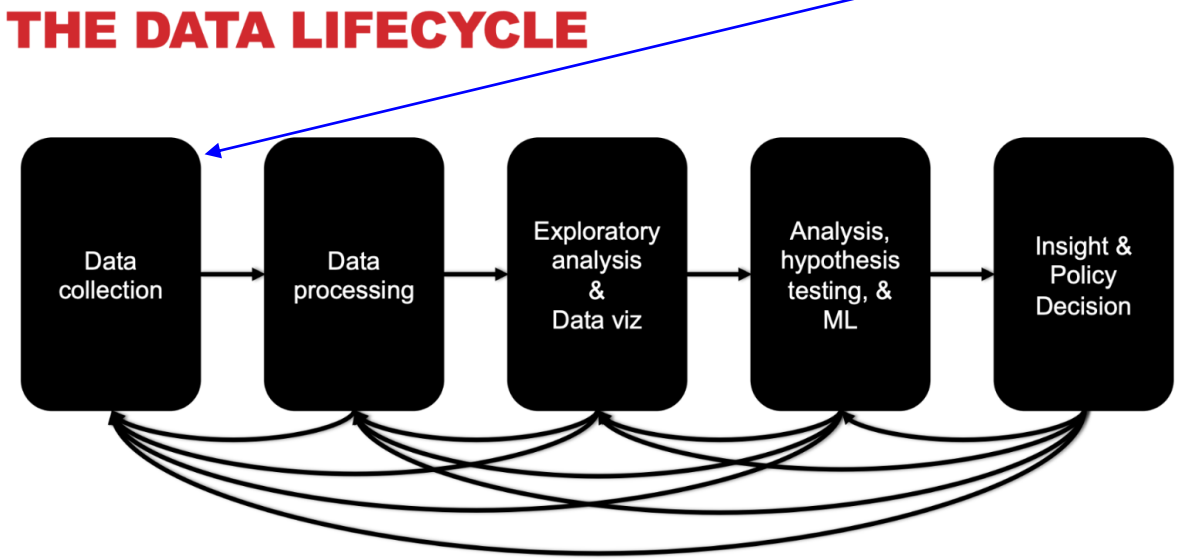


Data science fundamentally involves making decisions **based on data**.

Experiments in Data Science require proper planning and design to best answer its questions.

Experimental Design in Data Science?

Experimental design is a crucial aspect of data science that focuses on planning and conducting experiments to gather **meaningful data**.



- Involves making decisions about **how** to **collect**, **manipulate**, and **analyze data** to answer specific research questions or test hypotheses.
- **Goal:** Ensure collected data is reliable, unbiased, and leads to valid conclusions.

Maximizing the amount of data that can be gathered from an experiment while minimizing the time, costs, and mistakes that are involved is the goal of Data Science and experimental design.

Experimental Design Steps in a data science project

- ❏ Define the Problem or Research Question aims to address.
- ❏ Identify the variable(s) including potential confounding variables and the population/sample you will study. .
- ❏ Formulate Hypothesis
- ❏ Develop a detailed plan for data collection, ensuring representativeness when using a sample.
- ❏ Select Data Collection Methods
- ❏ Then Collect Data

Next DS Life Cycle Step…….

Define the Problem or Research Question aims to address

What will the weather be like for the next 10 days?

Which courses are likely to have the highest demand next semester?

How many visitors will the website named "X" receive over the next week?

Which students are at risk of dropping out of an online course?

Which loan applicants are likely to default in the next year?

Upcoming movie "Avatar 4" will be a box office hit or miss?

Will flights be delayed over the next 10 days?

Which subscribers are likely to cancel their Netflix subscription soon?

Asking the right questions before solving a DS problem is a great start! And be specific!

Need proper planning and good **experimental design**.

Remember that

For most data science applications, ultimately, comes down to predicting the future.

We answer the question, “Given some set of options, which option **maximizes** my **optimization criteria**.”

Objective or goal function → we want to *achieve*

Identify the option/ choice/ decision that maximizes or optimizes the desired outcome based on these predefined criteria.

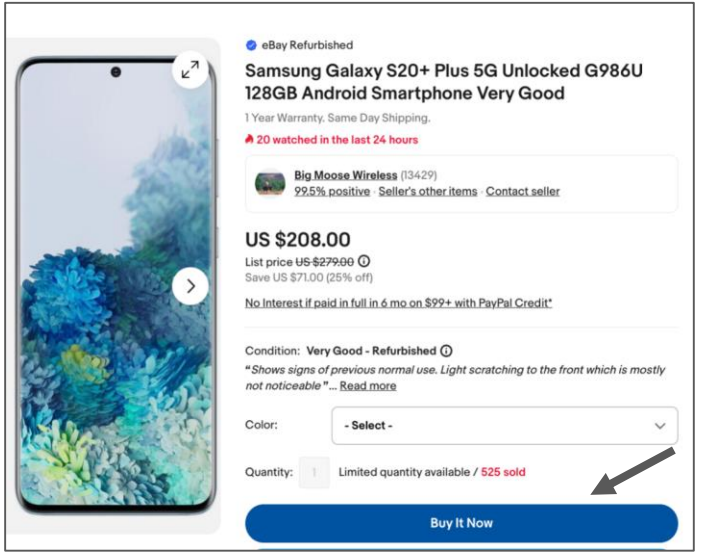
Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Example: Online Retail

Let's say you're a data scientist working for an online retailer, and you want to test whether changing the color of the "Buy Now" button on your website affects the click-through rate (CTR)

- **Click-through rate (CTR) → the percentage of users who click on the ad after seeing it**

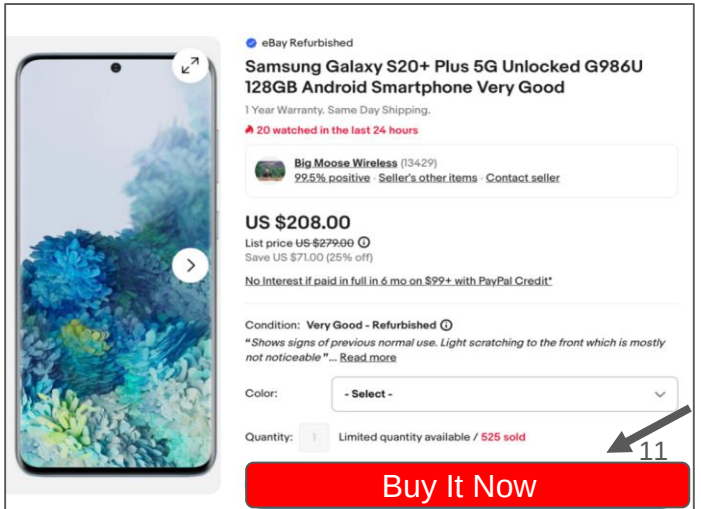


What is your problem definition?

Find which version of the button "Buy Now" (Option A (default) or Option B (red)) is more likely to maximize the CTR

What is you Optimization Criteria? What we want to maximize?

CTR → we want to select the ad options with button "buy now" that leads to **Higher CTR**



Ques: How can we set up an experiment to collect data in this case?

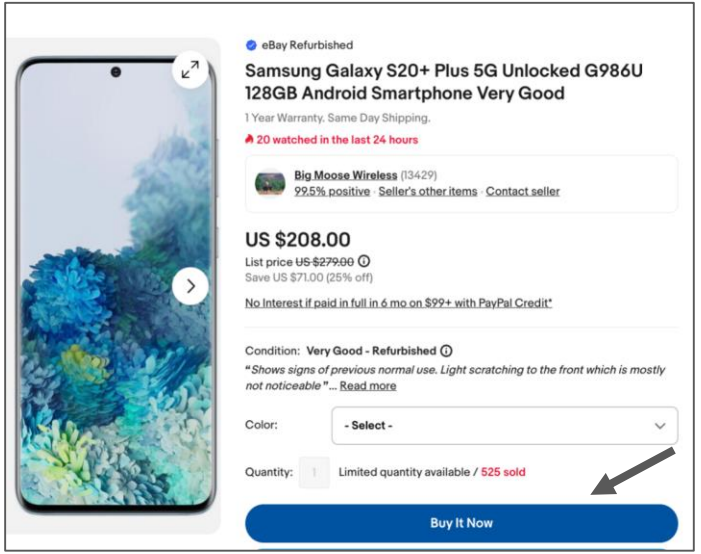
Ques: How can we set up an experiment to collect data in this case?

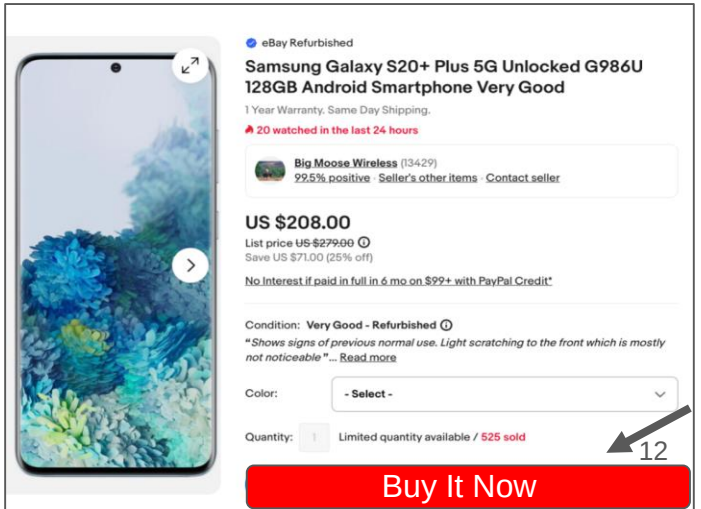
Data Size / Sample ? → No. of website visitors

- ➔ Views the original website with **existing button color**.
- ➔ Experiences no changes; baseline for comparison.

- ➔ Sees the same website but with a **different color for the "Buy Now"** button.
- ➔ Experiences the change you want to test.

- **Collect data** on the click-through rates for both groups over a specific period.
- **Compare** the click-through rates (CTR) between the groups **after** the experiment
- **Check** if there's a significant difference [LATER TOPIC IN THIS COURSE], we can infer that the **change in button color influenced the click-through rate**.





Ques: How can we set up an experiment to collect data in this case?

Data Size / Sample ? → No. of website visitors

Views the original website with **existing button color**. This group experiences no changes.

Sees the same website but with a **different color for the "Buy Now".** This is the group that experiences the change you want to test.

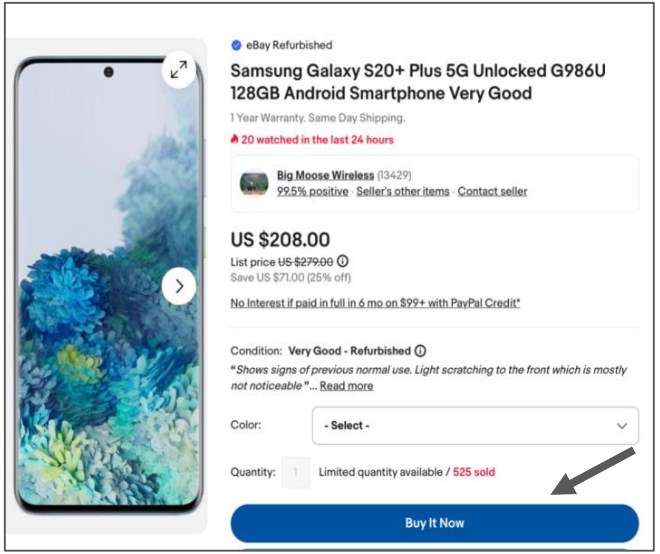
CONTROL GROUP

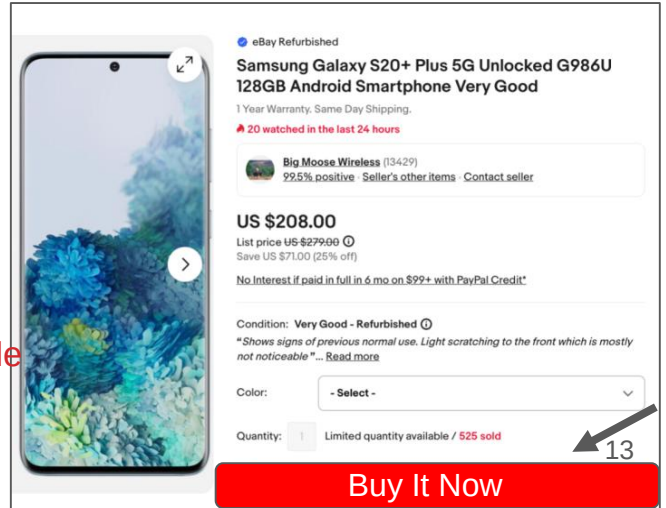
TREATMENT GROUP

Ques: What are the variables here?

- Click-through rate (what we measure) ← Dependent Variable
- Color of the "Buy Now" (what we manipulate) ← Independent Variable

Draw more reliable conclusions about the impact of the independent (manipulated) variable.





Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Identify the variable (or variables) of interest and the population of the study.

Once the problem is defined, identify the variable(s) of interest that are relevant to your research question.

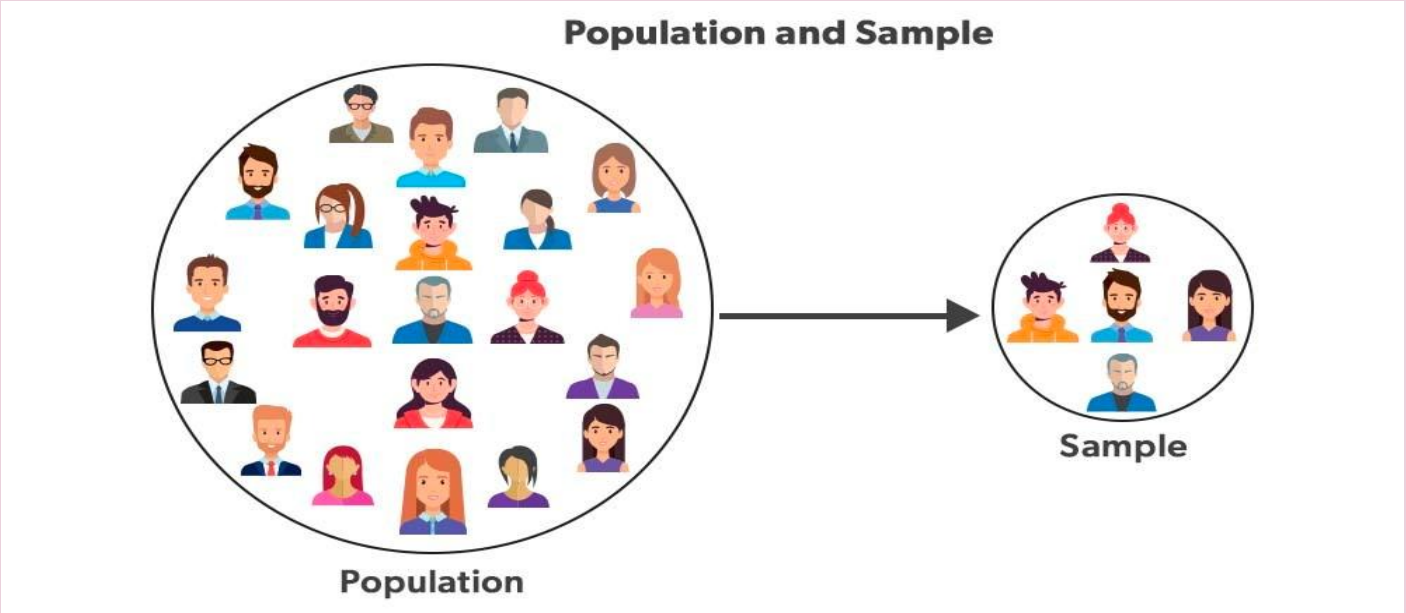
Also, specify the **population or sampl**e that your study will focus on.

Terminology: Variable

A variable is any characteristic that is recorded for subjects in a study.

Example: “gender,” “major,” “age,” and “GPA” might be variables for a study about college students.

Terminology: Population and Sample



Population: The entire group that we want to learn about.

Sample: A smaller group from the population that we actually study to draw conclusions about the entire population.

Identify the variable (or variables) of interest and the population of the study.

2 types:

- **Independent variable (IV):** are the factors/cause that may influence the outcome.
  - We can manipulate or change is to see how it affects the dependent variable.
- **Dependent variable (DV):** outcome of interest (what we measure)
  - expected to change as a result of changes in the independent variable

Important to know: While independent variables produce dependent variables, it is also true that **multiple independent** variables **may influence one another**.

- Why it's a problem? can complicate the interpretation of results (challenging to determine the individual impact of each ind. Variables to dep. variable).

➢ Good experimental design minimizes the presence of correlated variables.

Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Come up with a Hypothesis

The hypothesis of your experiment is the statement you want to test.

“if hypothesis X is right, then Y should be true…..”

- **A testable explanation/statement that addresses the question.**
- Represents an educated guess as to the relationship between variables and outcome of the experiment
- Experiments or Observations are conducted to see whether the predictions are correct. If confirmed, the hypothesis is supported. If not, it may be time for a new hypothesis.

Example

Hypothesis:

"If the amount of **average study time** (___ variable?) is increased, then **average exam scores** (___ variable?) will also increase.”

Recap:

- **Independent variable (IV):** are the factors/cause that may influence the outcome.
- **Dependent variable (DV):** outcome of interest (what we measure)

More Examples

Hypothesis:

As **books read** increases, **average literacy** also increases

If the **exercise duration** is extended, then the **average calories burned** will also increase

If the **temperature** rises, then **average ice cream sales** will also increase.

Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Next:

Select a Data Collection Method and Collect Data to Test your Hypothesis

Before collecting data to test your hypothesis, you first have to consider the problems that can cause errors in your result, one of them being a **confounder**.

Analyze if any other variables (beyond the ones you intended to manipulate or measure) could have impacted / influenced your results (Note that, if is true, we may need to redesign the experimen for ensuring more accurate and reliable conclusions).

What is a confounder?

Extraneous variables that **may affect** the relationships between dependent and independent variables.

- **May not the focus** of the study but **can impact the results** by influencing the dependent variable.
- An additional factor that is not controlled for but could potentially influence the outcomes.

A confounder can lead to incorrect conclusions about the true effect of the independent variable on the dependent variable.

Back to Example

factor that is manipulated or changed; expected to have an impact on the dependent variable

Hypothesis:

"If the amount of **study time** (independent variable) is increased, then **exam scores** (dependent variable) will also increase."

the outcome or response that is measured in the study; expected to change based on variations in the independent variable.

A potential confounding variable? **Prior Knowledge**

Back to More Examples: Find Potential Confounder

As **books read** increases, **literacy** also increases → Age, Socioeconomic status

If the **exercise duration** is extended, then the **calories burned** will also increase → Metabolic Rate

More Example: Experimental Design Flow : Polling

1. **<u>Problem Formulation:</u>** Imagine you're a data scientist tasked with predicting the outcome of a political election using a dataset of voter preferences. Your goal is to **design an experiment** that **accurately represents the entire population's voting behavior.**

   Questions: How do we know which candidate is ahead!

   - Can we create the IDEAL POLLING?

Eliminate confounding variables as much as possible → Sample Bias, geographic representation, Population Proportion Bias, Demographic Mismatch and many more to make the data as accurate as feasible.

Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Deals with Confounder

Considering and addressing confounders in experimental design is crucial to ensure accurate results, preventing external factors from influencing outcomes and ensuring the correct design of experiments.

There are three ways you can deal with confounders:

- Control
- Randomization
- Replication

Control

Manage or eliminate the impact of confounding variables

- involves the use of **control groups** and **treatment groups**

Example 01: control the effect of “prior knowledge”

"If the amount of **study time** is increased, then **exam scores** will also increase.”

We can measure the prior knowledge of each individual; to see the effects/impact of prior knowledge on exam score.

Experimental Design:

- Control group → Participants with no prior knowledge.
- Treatment group →Participants with a range of prior knowledge levels.

Comparison:

Results are compared to determine whether the **no prior knowledge group** differs in exam score from the **group of having varying levels of prior knowledg**e.

Example 02: control the effect of “age”

"As **books read** increases, **literacy** also increases.”

We can measure the age of each individual; to see the effects of age on literacy.

Experimental Design:

- Control group → participants of fixed ages.
- Treatment group →participants of the range of ages.

Comparison:

Results are compared to determine whether the **fixed age group** differs in literacy from the **group of various ages**.

Another general example: a classic experimental design in drug testing

We want to **test the effect of a drug** on patients.

Experimental Design:

- Control group → won't receive the drugs (used a baseline;is not exposed to the independent variable)
- Treatment group → receive the drugs

and the effects are compared.

Randomization

Helps make groups fair by deciding things randomly.

Randomly assign participants to

- The control group (**won't receive the drugs**) or
- The treatment group (**receive the drugs**)

Why? To minimize systematic confounding, reduce the risk of bias in either group being enriched for confounders, and help distribute confounding variables equally

❑ This minimizes the likelihood of systematic confounding;

- Reduces the risk of bias in either group being enriched for confounders.
- Each participant is assigned to one of the two groups, not both simultaneously.

Replication

Doing your experiment more than once to be more certain.

- Repeat the experiment multiple times to assess the consistency and reliability of results, helping to identify and account for potential confounding factors.

If replication is done(with a new set of data) and it **produces the same conclusions**, this shows that the experiment is strong and has a good design and suggests that **confounders are less likely to be influencing** the outcomes.

Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

Methods for Collecting Data

Below are some common methods/way to collect data if pre-existing datasets are not available.

- Observational studies
- Surveys.
- Experiments
- Simulations

Methods for Collecting Data

A. Observational studies → Observe and record data (variables) without intervening or manipulating variables (Observe; don't change anything intentionally).

E.g. Observing animal behavior in a natural habitat without any external influence.

B. Surveys → Collect information through structured questionnaires or interviews.

E.g. Conducting a survey to gather opinions on a political issue.

C. Experiments → We actively change something to see what happens.

D. Simulations → Create artificial scenarios to model real-world situations for data collection.

E.g. Using a computer simulation to study traffic patterns in a city.

Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

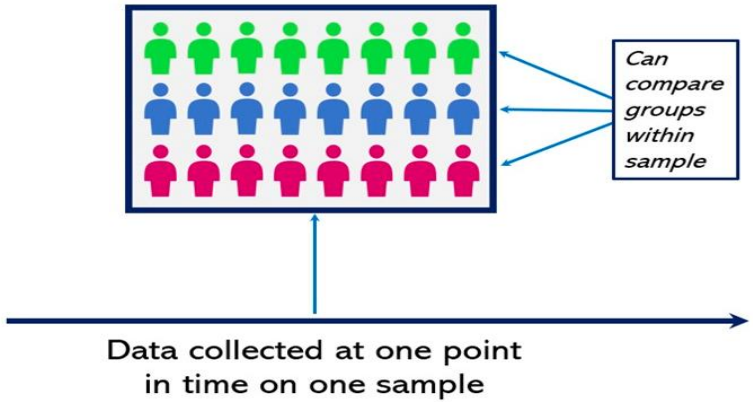
A. Observational Studies

- Cross sectional studies
- Retrospective (case control) studies
- Prospective (longitudinal or cohort) studies

A. Observational Studies

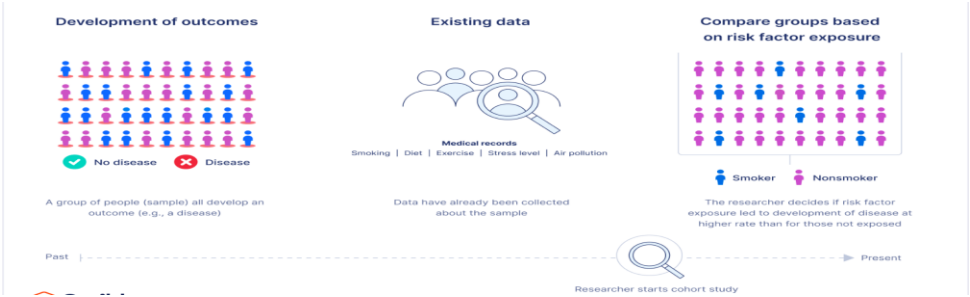
- **Cross sectional studies:** Collects data from many different individuals at one specific single point of time

  - Taking a picture of a group **right now** to see what they're like.
  - Surveying individuals of various age groups **on a single day**



A. Observational Studies

- **Retrospective (case control) studies:** Looking backwards at studies of events in the past to examine the relationship between the exposure and the outcome. E.g. Investigating past events to find out what might have caused something.
    - **Interviewing severe dengue patients** to ask about their experiences and medical history, studying disease spread dynamics.
    - **Studying past smoking history in lung cancer patients** (cases) and non-cancer individuals (controls) to identify associations and understand their impact on health outcomes.



A. Observational Studies

- **Prospective (longitudinal or cohort) studies:** Researchers follow and observe a group(s) of people (called a cohort) closely - ***over a period of time***, collecting data on their exposure to a factor of interest (observe changes; identify potential causes).

Observing a cohort <u>over time to see</u> how things change → tracking their smoking habits and health outcomes to determine the association between smoking and the development of respiratory diseases.

  - but sometimes people stop participating, which can make things a bit tricky to understand.
  - These tend to have high dropout rates

B. Surveys (A specific type of observational study)

A survey is used to investigate characteristics of a population. It is frequently used when the subjects are people, and questions are asked of them.

- When designing a survey, you must be very careful of wording (and sometimes ordering) the questions so that the results are **not biased**.

Topics

- An example of Experimental Design (ED)
- Identifying Variables & Population/Sample of the Study
  - Independent variable
  - Dependent variables
- Hypothesis
- A potential problem in ED: Confounder Variable
- How to Deal with Confounder
  - Control
  - Randomization
  - Replication
- Methods for Collecting Data
  - Observational studies
    - Cross sectional studies
    - Retrospective (case control) studies
    - Prospective (longitudinal or cohort) studies
  - Surveys.
  - Experiments
    - Placebo Effect
    - A common method to minimize bias in Experimental Design
  - Simulations

C. Experiment

In an experiment, a researcher assigns **a treatment** and observes the response.

- Observe effects on subjects after the application of some treatment
  - Might want to compare a treatment versus a control or multiple treatments -

receives intervention/actual treatment

receives no intervention or receive placebo (fake treatment)

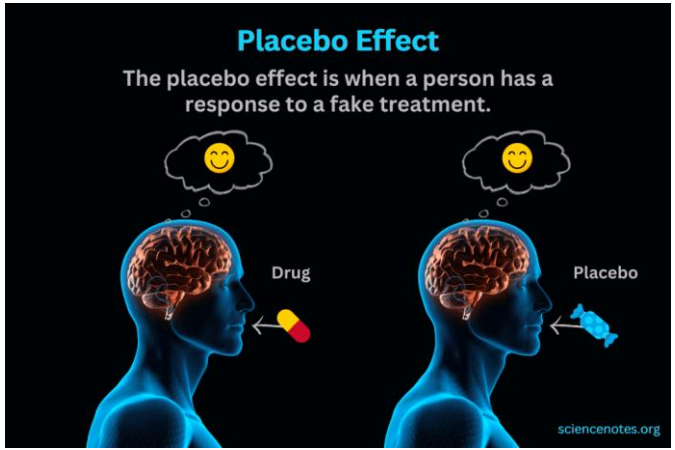
- The more variables you can control for, the better your experiment

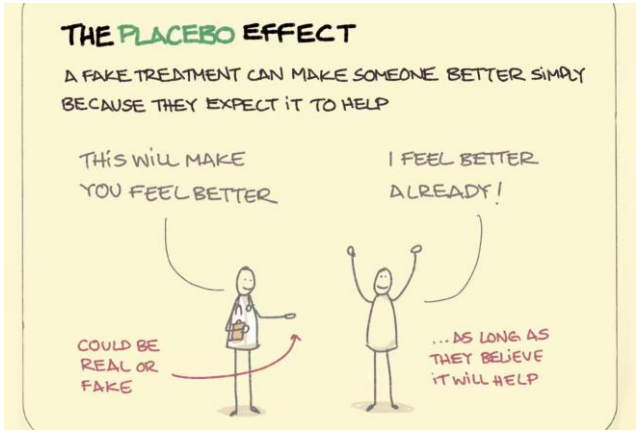
Placebo effect

Placebo Effect: Belief in treatment leads to feeling better even without actual treatment.

- A person believes psychologically that a certain treatment is positively affecting them, even though no treatment was given at all.

E.g: The placebo effect is when a person's physical or mental health appears to improve after taking a placebo or 'dummy' treatment.





A common method to minimize bias in Experimental Design

Blinding: (Subjects would be blinded) when the people involved in an experiment **don't know who's getting the real treatment and who's not.** Prevent Biases. It is a technique used to make the subjects “blind” to which treatment (or placebo) they are being given

- **Single-blinding** is when either the participants or the researchers don't know
- **Double-blinding** is when both don't know.

The use of placebos contributes to the blinding of experiments

D. Simulation

A simulation uses a mathematical, physical, or computer model to replicate the conditions of a process or situation.

- This is frequently used when the actual situation is too expensive, dangerous, or impractical to replicate in real life

The Fundamental Rule of Data Collection

Your data must representative of the population you want to study.

Keep in mind that

It is almost impossible to be certain that your experiment has completely removed all forms of bias. It is necessary to consider possible sources of bias and highlight them in your analysis. Ideally, future experiments would improve upon your method by iteratively eliminating those sources of bias.

Quick Class Task

Identify which method for collecting data (observational study, an experiment, a simulation, or a survey) is **best** in each of the following situations and **explain your answer.**

1. The effect of a severe earthquake would have on the Salt Lake Valley.
2. Whether or not a certain coupon attached to the outside of a catalog makes recipients more likely to order products from a mail-order company.
3. Whether or not smoking has an effect on coronary heart disease.
4. Determining the average household income of homes in Salt Lake City.

Additional Reading Slides

Example: Online Ad Campaign

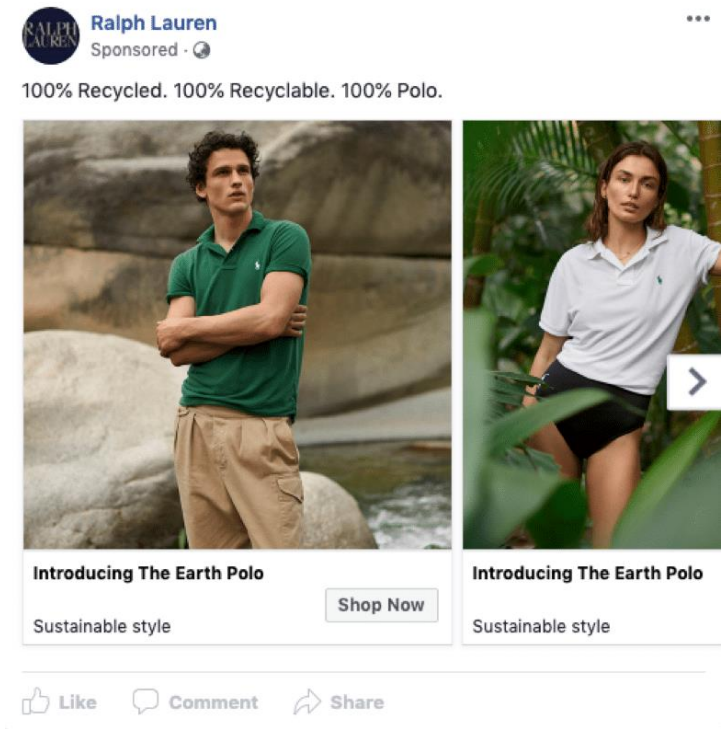
Imagine you're working for an e-commerce company that is planning an online ad campaign to promote a new product. Your goal is to design an effective campaign that maximizes the click-through rate (CTR) → the percentage of users who click on the ad after seeing it

What is your problem definition?

Find which version of the ad (Option A or Option B) is more likely to maximize the CTR

What is you Optimization Criteria? What we want to maximize?

CTR → we want to select the ad options that leads to **Higher CTR**



Try: What is the Optimization Criteria for each following:

- Amazon
- Facebook
- A non-profit treating malaria
- A defense contractor

Using data to make smarter decisions for the future
But it’s hard!

Hypothesis

An educated guess as to the relationship between variables and outcome of the experiment

The hypothesis of your experiment is the statement you want to test. A null hypothesis is a default statement that you make that assumes there is no relationship between the compared variables of your test.

“A hypothesis may be simply defined as a guess. A scientific hypothesis is an educated guess.” – Isaac Asimov, Book of Science and Nature Quotations, 1988

Components of Experimental Design / Experiment Terminology

Dependent variable (DV): Variable the experimenter measures. This is the outcome (i.e., the result) of a study.

- It is measured to assess the effect of the independent variable. It is the outcome or response.

Independent Variable (IV): The variable the experimenter manipulates (i.e., changes) is assumed to have a direct effect on the dependent variable.

- It is manipulated by the researcher to observe its effect. It is the cause in a cause-and-effect relationship.

Components of Experimental Design / Experiment Terminology

A confounding variable: is a variable that is not the focus of the study but can impact the results by influencing the dependent variable.

- Variable(s) that have affected the results (DV), apart from the IV. A confounding variable could be an extraneous variable that has not been controlled.

Components of Experimental Design / Experiment Terminology

A confounding variable: A variable that is connected to both the dependent and independent variables but is not a component of the hypothesis being tested is referred to as a confounder.

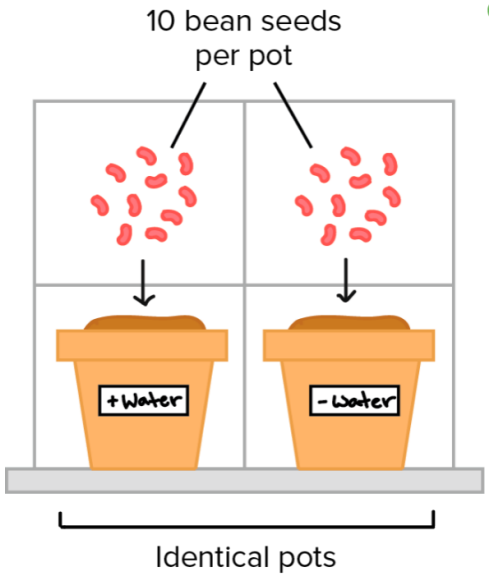
- Confounding factors have the potential to skew an experiment's findings and provide false conclusions.
- Confounding variables must be understood in experimental design and taken into account when creating the experiment.

Control Group: A group in an experiment that does not receive the treatment or intervention, used as a baseline for comparison.

Treatment / Experimental Group: The group in an experiment that receives the treatment or intervention being studied.

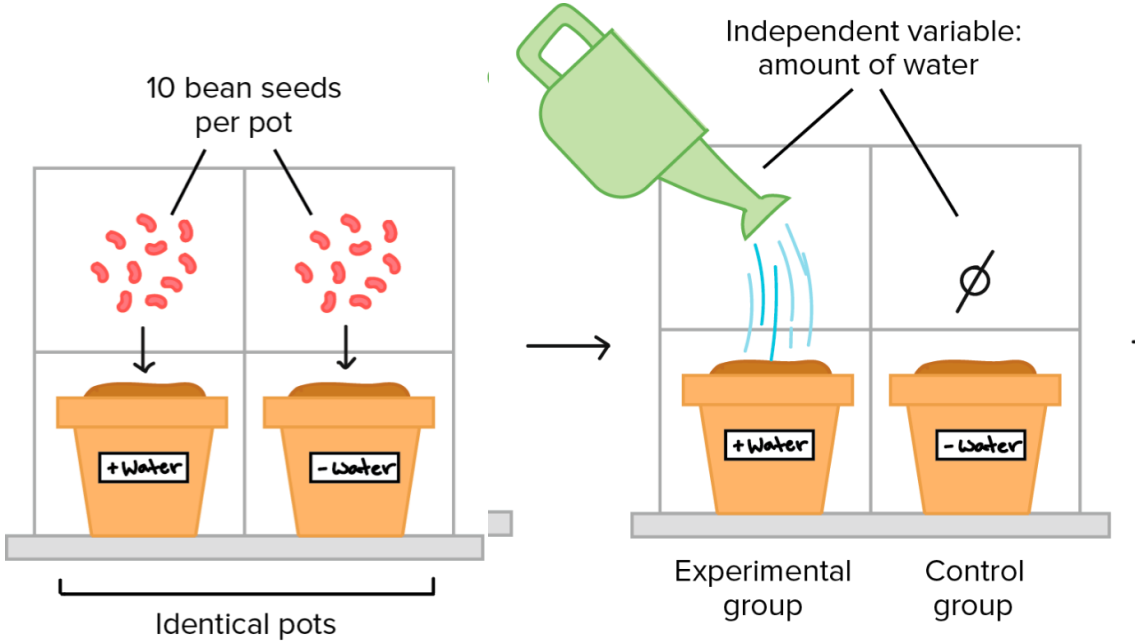
Example: Control Experiment

Let’s hypothesis: “ Water is needed to bean sprout to grow!”



Example: Control Experiment

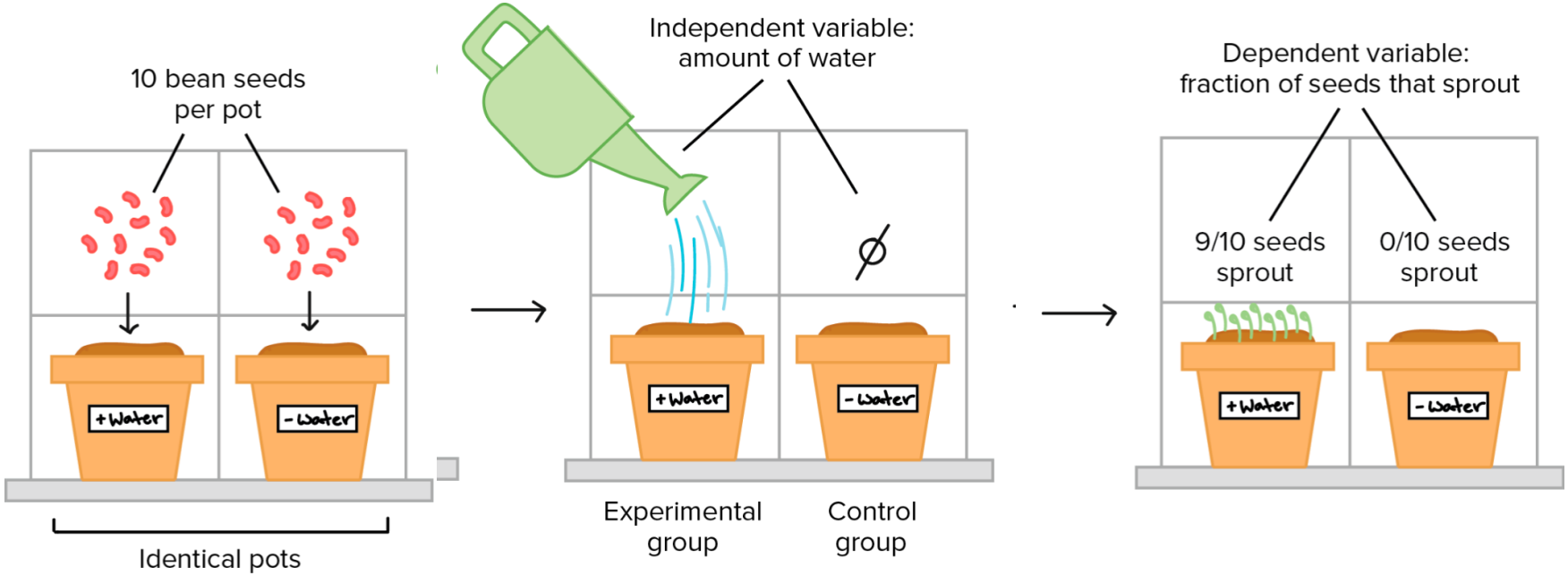
Let's hypothesis: " Water is needed to bean sprout to grow!"



(Treatment Group)

Example: Control Experiment

Let's hypothesis: " **Water** is needed to **bean sprout** to grow!"



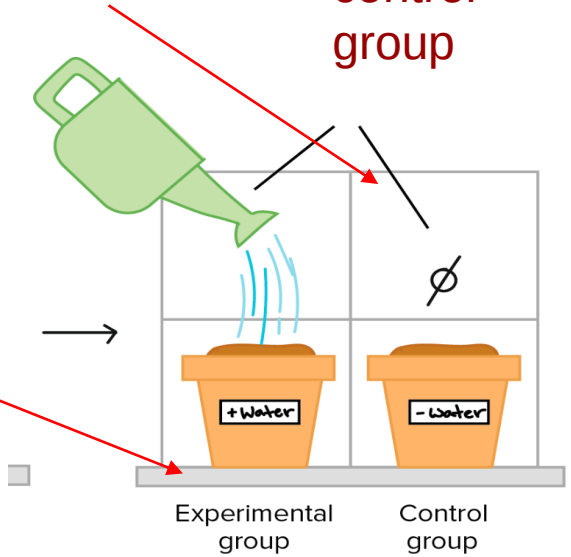
Control and Treatment (experimental) groups

Two identical groups except:

1. one receives a treatment (water) while the other (control) does not.

control group

Treatment (experimental) group

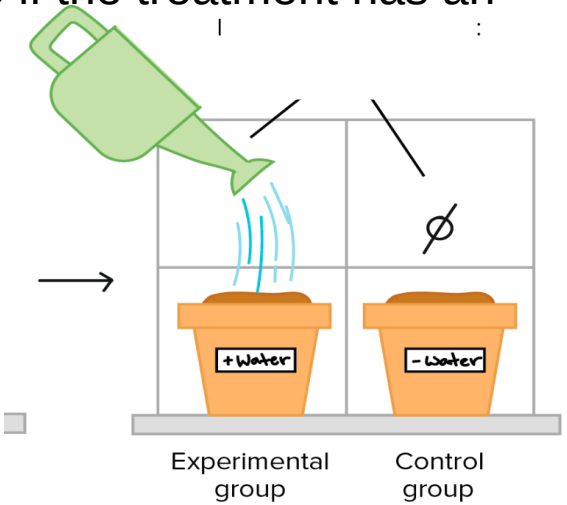


(Treatment Group)

Control and Treatment (experimental) groups

Two identical groups except:

1. one receives a treatment (water) while the other (control) does not.
    - Control group provides a baseline that lets us see if the treatment has an effect.



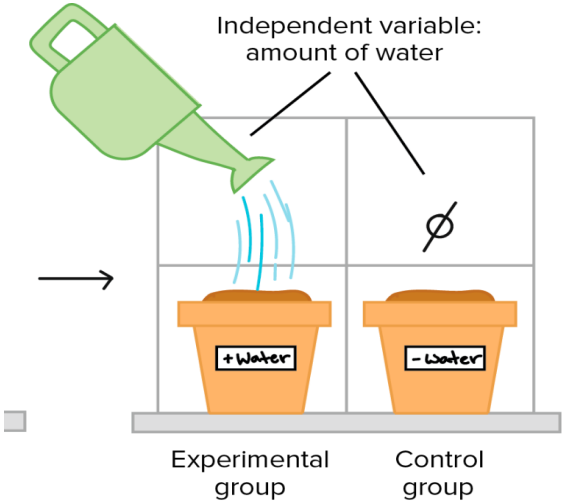
(Treatment Group)

Independent and dependent variables

The factor that is different between the control and experimental groups is known as the **independent variable**.

- What is in this case?

- Why independent ? does not depend on what happens in the experiment. Instead, it is something that the experimenter applies or chooses him/herself.



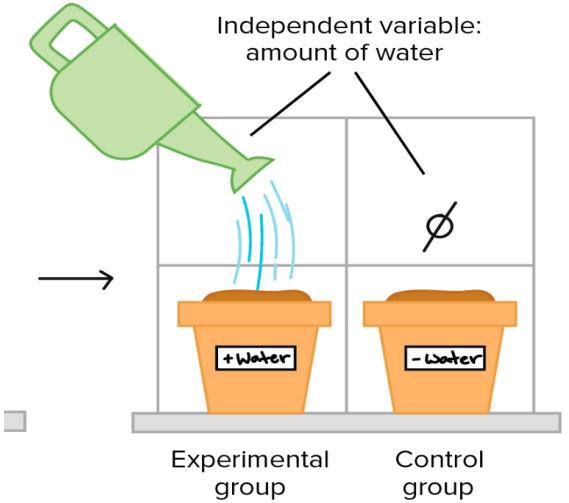
(Treatment Group)

Independent and **dependent** variables

In contrast, **the dependent variable** in an experiment is the response that's measured to see if the treatment had an effect.

- What is in this case?

- Imp. Key points: Remember, the dependent variable depends on the independent variable , and not vice versa.



(Treatment Group)

Randomization: Random assignment of participants to different experimental groups (or to independent variable conditions) to minimize biases and ensure equal representation.

- All participants should have an equal chance of taking part in each condition.

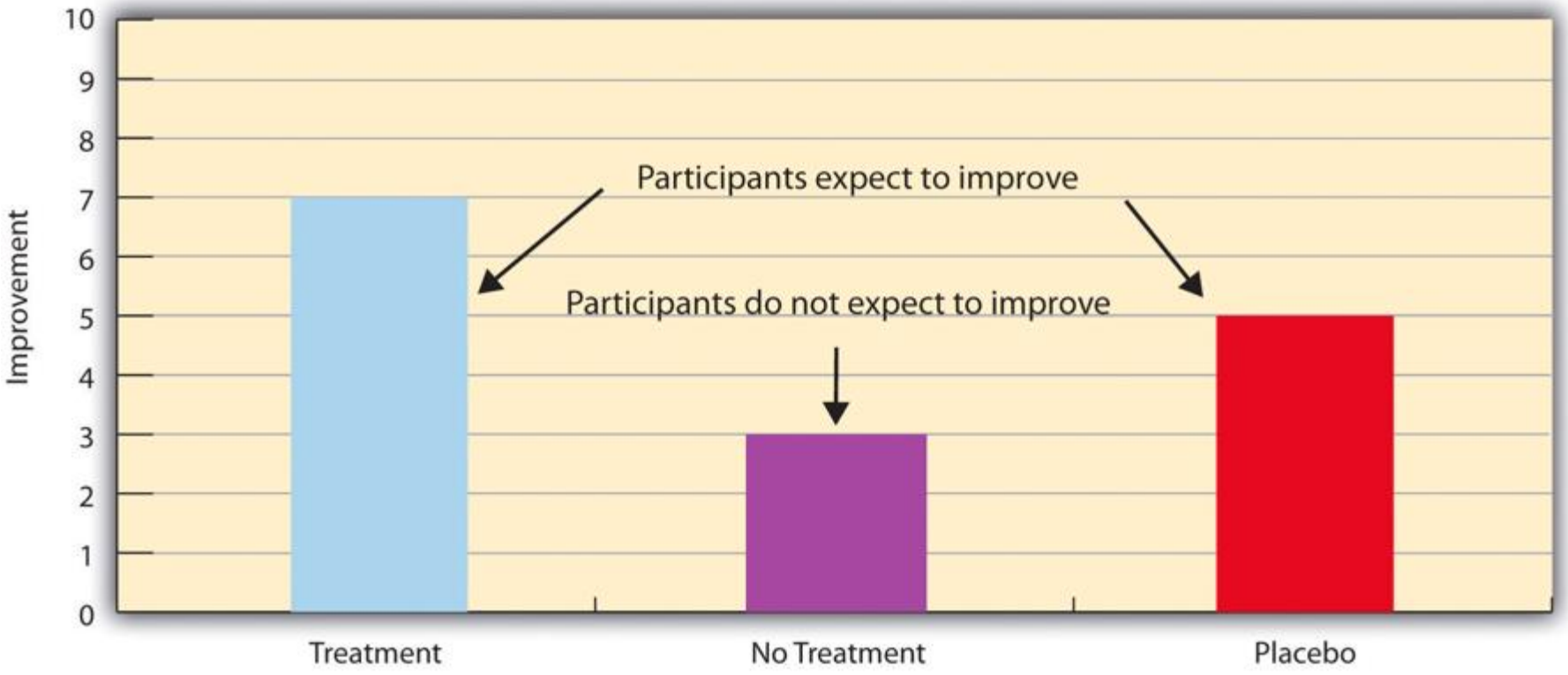
Replication: Repeating the experiment with different samples to enhance reliability.

The Powerful Placebo

Many people are not surprised that placebos can have a positive effect on disorders that seem fundamentally psychological, including depression, anxiety, and insomnia. However, placebos can also have a positive effect on disorders that most people think of as fundamentally physiological. These include asthma, ulcers, and warts (Shapiro & Shapiro, 1999)[2]. There is even evidence that placebo surgery—also called "sham surgery"—can be as effective as actual surgery.

Medical researcher J. Bruce Moseley and his colleagues conducted a study on the effectiveness of two arthroscopic surgery procedures for osteoarthritis of the knee (Moseley et al., 2002)[3].The control participants in this study were prepped for surgery, received a tranquilizer, and even received three small incisions in their knees. But they did not receive the actual arthroscopic surgical procedure. The surprising result was that all participants improved in terms of both knee pain and function, and the sham surgery group improved just as much as the treatment groups. According to the researchers, "This study provides strong evidence that arthroscopic lavage with or without débridement [the surgical procedures used] is not better than and appears to be equivalent to a placebo procedure in improving knee pain and self-reported function" (p. 85).

Placebo Affect



Recap: Definitions and Terminology

• **A confounding variable** occurs when an experimenter cannot tell the difference between the effects of different factors on a variable.

• **The placebo effect occurs** when a subject (or “experimental unit”) reacts favorably to a placebo when no medicated treatment has been given.

• **Blinding** is a technique used to make the subjects “blind” to which treatment (or placebo) they are being given.

• **A double-blind experiment** is one in which neither the experimenter nor the subjects know which treatment is being given.

• **Randomization** is a process of randomly assigning subjects to treatment groups. There are several different techniques for randomization:

- A completely randomized design assigns subjects to different treatment groups through random assignment.
- A randomized block design is sometimes used to make sure that subjects with certain characteristics are assigned to each treatment. For example, when testing a certain medication, you might first want to split subjects in groups according to either gender or age (or both), then randomly assign each of these groups to the different treatments.

• **Sample size** is the number of participants in the experiment. The larger the sample, the more representative of the population the results will be, but the costs of the experiment will also be higher.

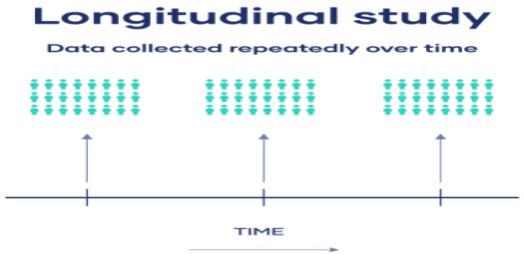
• **Replication** is the ability to reproduce the experiment (and results) under similar conditions

Observational Studies

- **Cross sectional– is a type of observational study that analyzes data from a population, or a representative subset, at a specific point in time**
  - Look at data a a single point in time
  - Ex: Like taking a picture of a group right now to see what they're like.
  - A snapshot of the population at a particular moment rather than a study that tracks changes over time.

Observational Studies

- .**Retrospective (case control) studies**--Looking at studies of events in the past
  - Ex: Investigating past events to find out what might have caused something.
    - Researchers gather information about events that have already happened (*smoking habits*) to better understand their possible impact on health outcomes (*lung cancer*).
- **Prospective (longitudinal or cohort) studies**--Researchers follow and observe groups closely
  - These tend to have high dropout rates
  - Ex: Watching a group over time to see how things change, but sometimes people stop participating, which can make things a bit tricky to understand.



Surveys (A specific type of observational study)

- What type of study are these?
- phone, mail, email, web-based, in person
- some additional issues
  - **Wording of questions can introduce bias (deliberate or unintentional) –** “Do you agree…?" - "that pepperoni is the best pizza topping?"
  - **Ordering of questions (planting ideas)** - Ask about favorite vegetables before meats.
  - **Convenience samples/Self-selected samples**- Surveying solely on a vegetarian forum.
  - **Desire of respondents to please**- “choose pepperoni over mushrooms to please respondents”
  - **Confidentiality concerns may influence responses** - Asking name, address, pizza choices: privacy worries.
  - **Non-response bias**- “sending survey only to 30 out of the 100 people”

Probability, Distributions, and Summary Stats

DATA, MSML, BIOI 602 Principles of Data Science

[Slides from Fardina Fathmiul Alam]

Topics we will cover:

1. Probability Theory:
    a. Basic concepts (events, sample space, probability axioms)
    b. Conditional probability
    c. Bayes' theorem
    d. Probability distributions (discrete and continuous)
    e. Common distributions (e.g., normal, binomial, Poisson)
2. Descriptive Statistics:
    a. Measures of central tendency (mean, median, mode)
    b. Measures of dispersion (variance, standard deviation, range)
    c. Percentiles and quartiles
    d. Skewness and kurtosis

Inferential Statistics: Hypothesis testing (Later Topic)

Part 01 Probability Theory

What is Probability?

- People talk loosely about probability all the time:

  “What are the chances the Orioles will win this weekend?”

  “What’s the chance of rain tomorrow?”

- For scientific purposes, we need to be more specific in terms of defining and using probabilities

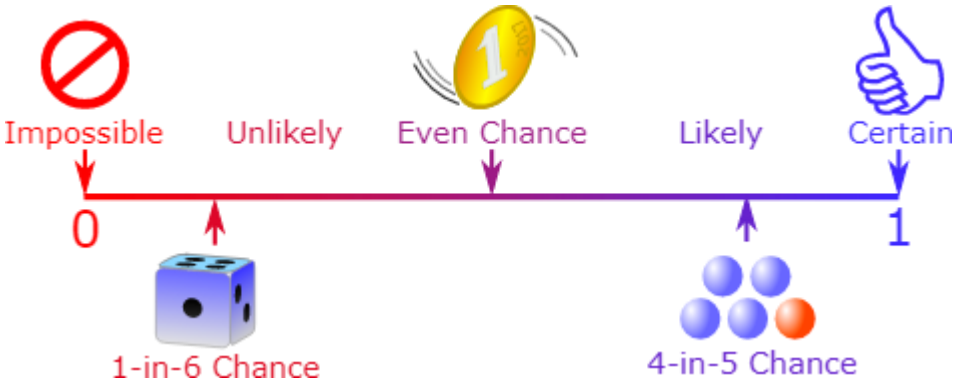
What is Probability?

Probability is simply how likely something is to happen.

Remember, the analysis of events governed by probability is called statistics.

What is Probability? (most commonly used concepts in statistics)

Classic Definition: Probability is a measure between **zero and one** for the **likelihood** that something or some event might occur.



A Classic Example of Probability

- What is the chances of rolling a one (Number 1) on the dice?

Answer is a 1/6

Ques: How do we know that ?

Basic Probability Formula

To find the probability of an event happening we use the formula:

$$P(A) = \frac{\text{Number of times A occurs}}{\text{Total number of possible outcomes}}$$

Here, **P(A)** = The probability of event A

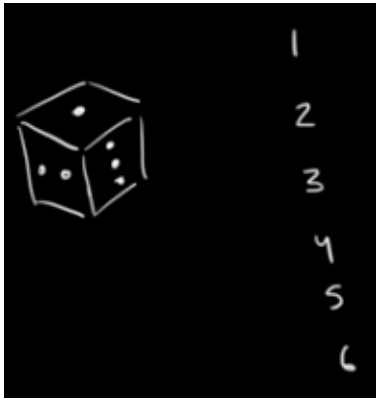
Probabilities range from **0 (impossibility-the event will not occur)) to 1 (certain- the event will definitely occur)** [degrees of likelihood]the and can also be written as a percentage.

Basic Probability Formula

To find the probability of an event happening we use the formula:

$$P(A) = \frac{\text{Number of times A occurs}}{\text{Total number of possible outcomes}}$$

Example: The probability of getting number 1 is: $P(1) = 1/6$



There are six different outcomes. (**sample space**)

Try Yourself: What is probability of getting an even number when a die is rolled?

Some More Examples

- Stock Market ➔ The chance of stock market's rising above some point, or falling below some point is ‘x’%
- Weather Forecast ➔ The chance for rain is 45% tonight.
  - The likelihood of rainfall in terms of probability is 0.45 that the event, rainfall, ,might occur

So Essentially probability is a measure between equate to 1.

Probability allows us to measure and express uncertainty.

In data science, we often deal with incomplete or noisy data, and probability provides a framework to assess the likelihood of different outcomes. By assigning probabilities to events, we can make informed decisions and evaluate the associated risks.

Random Variable and Probability Distribution: Making Sense of Uncertainty

A random variable is a variable in statistics and probability theory that represents the outcomes of a random event.

- Assigns a number to each possible outcome of an uncertain event.
- Functions as a map of outcomes in a probability space.

E.g: Imagine rolling two dice. The outcome (the combination of numbers we get) is uncertain. So, we use a random variable, X to represent the sum of the numbers on the two dice. X could take on values from 2 to 12, depending on the sum of the two dice.

Random Variable and Probability Distribution: Making Sense of Uncertainty

Probability Distributions: Describe the likelihood of different outcomes occurring.

- Tells us how likely each outcome of the random variable is.
- Act like a map showing the chances of each possible value of our random variable.

Example: Probability Distribution

Probability Distributions: Describe the likelihood of different outcomes occurring.

E.g: Imagine rolling two dice. There are 36 possible outcomes (since each die has 6 sides, giving us 6×6=36 combinations). Each outcome has an equal chance of happening if the dice are fair. So, the **probability distribution for our random variable X** would look like this: →

P(X=2)= 1/36
P(X=3)= 2/36 (rolling a 1 and a 2 or rolling a 2 and a 1)
P(X=4)= 3/36
P(X=5)= 4/36
P(X=6)= 5/36
P(X=7)= 6/36
P(X=8)= 5/36
P(X=9)= 4/36
P(X=10)= 3/36
P(X=11)= 2/36
P(X=12)= 1/36

Random Variable Types:

- Discrete Random Variable: Represents outcomes that can only take on a countable number of distinct values.
  - E.g. Rolling a six-sided die. The possible outcomes (1, 2, 3, 4, 5, 6) are finite and countable.

- Continuous Random Variable: Represents outcomes that can take on any value within a certain range.
  - E.g. Measuring the height of people. Height can take on any value within a range (e.g., from 0 to infinity). It can be 160.5 cm, 178.23 cm, or any value in between.

More Probability Formulas

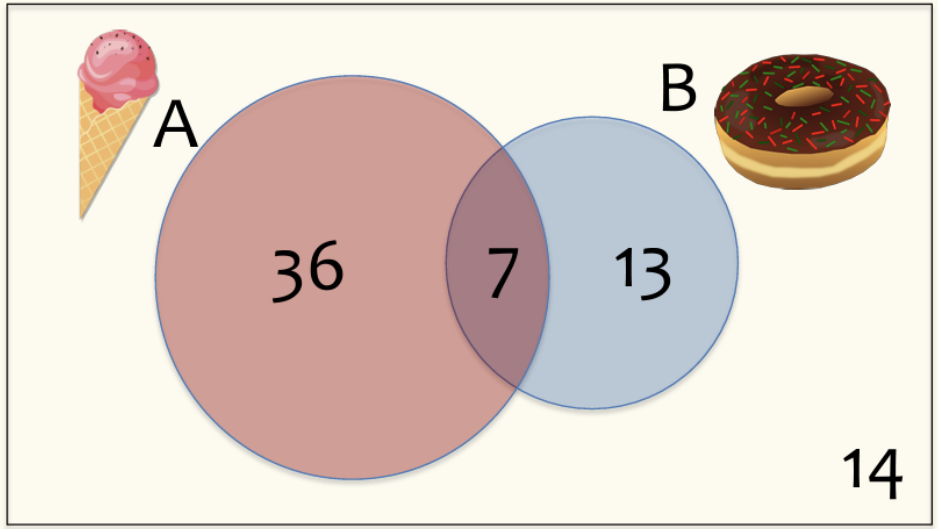
Sometimes you want to

figure out the chance of something happening when we already know else has happened/occured?

"Conditional Probability"

(Understanding Likelihood Given Information)

Conditional Probability (Idea)



What's the probability that someone likes ice cream **given** they like donuts?

Conditional Probability cont.

Example:

What is the probability of passing the class given you didn't sleep the night before?

Conditional Probability cont.

The probability of event A happening given that event B has occurred.

We write it: P(A | B) ➔ read as the probability of “A given B.”

For the example on the previous slide,
let

- A =Passing the class,
- B =Not sleeping the night before the final

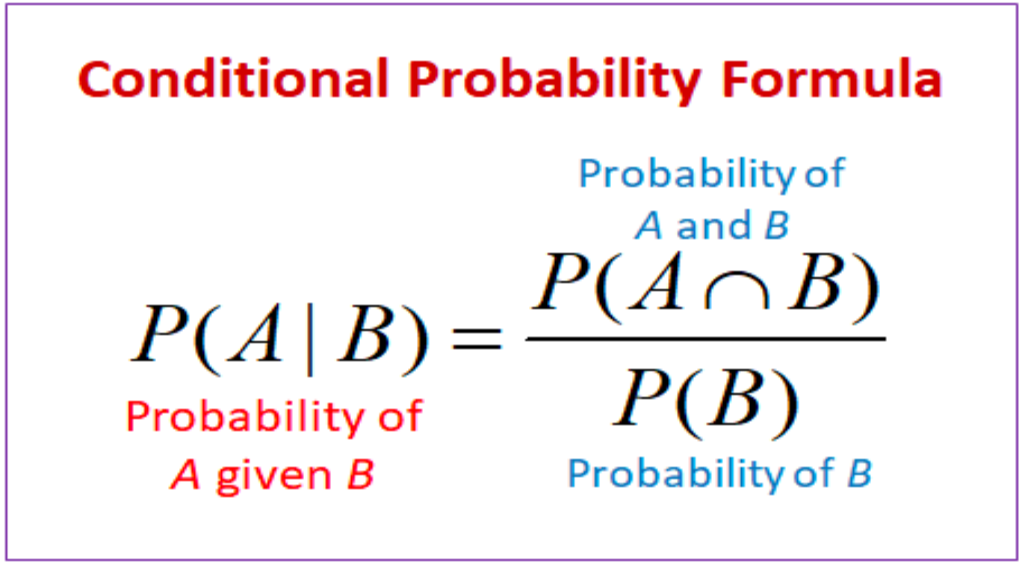
We can expressed as “P(Passing the class | Not sleeping the night before the final)”

Definition: Conditional Probability

Definition. The **conditional probability** of event A **given** an event B happened (assuming $P(B) \neq 0$) is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Formula: Conditional Probability



We take the chance of both things happening together, then divide it by the chance of the thing we know

This formula tells us the probability of event A happening, given that event B has happened.

Formula: Conditional Probability

Imagine two overlapping circles representing two events, **A** and **B**:

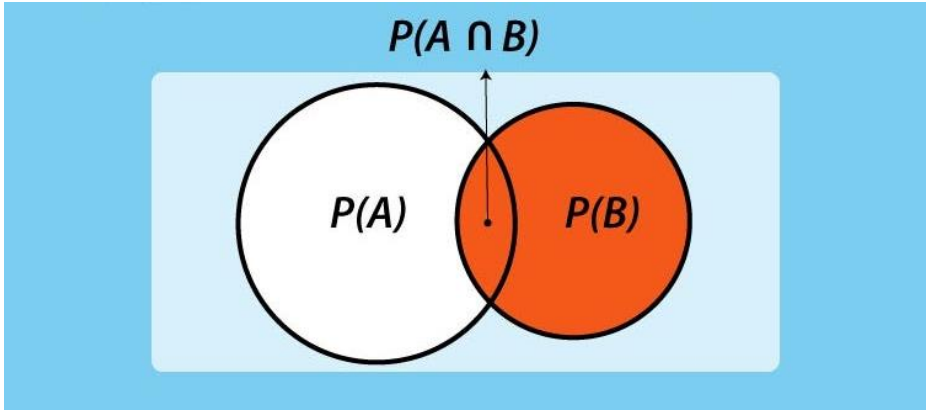
- **Circle A**: All outcomes where event **A** happens (**P(A)** - the size of Circle A).
- **Circle B**: All outcomes where event **B** happens (**P(B)** - the size of Circle B).
- **Overlap between Circles A and B:** Probability of both events happening together (P(A and B)).

Conditional Probability Formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B = Probability of A and B / Probability of B

By dividing by P(B), we focus our attention on the subset of cases where event B is true.

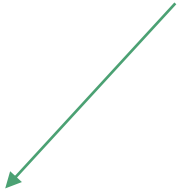


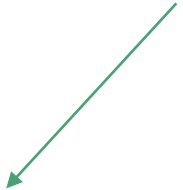
Venn diagram for Conditional Probability, P(A|B)

More Example:

Imagine you have a bag of colored marbles - 5 red and 5 blue. If you know that 3 out of the 5 red marbles are also **shiny**, you might wonder:

"What's the chance of picking a shiny marble from the bag if I know it's red?"





A represents the event of picking a shiny marble

B represents the event the event of picking a red marble.

More Example:

Imagine you have a bag of colored marbles - 5 red and 5 blue. If you know that 3 out of the 5 red marbles are also **shiny**, you might wonder:

"What's the chance of picking a shiny marble from the bag if I know it's red?"

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

Probability of B

P (A ∩ B) = What is the probability of picking a shiny red marble ?

P (A ∩ B) = 3/10

More Example:

Imagine you have a bag of colored marbles - 5 red and 5 blue. If you know that 3 out of the 5 red marbles are also **shiny**, you might wonder:

"What's the chance of picking a shiny marble from the bag if I know it's red?"

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

Probability of B

P (B) = What is the probability of picking a red marble ?

$$P(B) = 5/10$$

More Example:

Imagine you have a bag of colored marbles - 5 red and 5 blue. If you know that 3 out of the 5 red marbles are also **shiny**, you might wonder:

"What's the chance of picking a shiny marble from the bag if I know it's red?"

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

Probability of B

$P(A \cap B) = 3/10$

$P(B) = 5/10$

$P(A|B)?$

More Example:

Imagine you have a bag of colored marbles - 5 red and 5 blue. If you know that 3 out of the 5 red marbles are also **shiny**, you might wonder:

"What's the chance of picking a shiny marble from the bag if I know it's red?"

Conditional Probability Formula

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of A and B

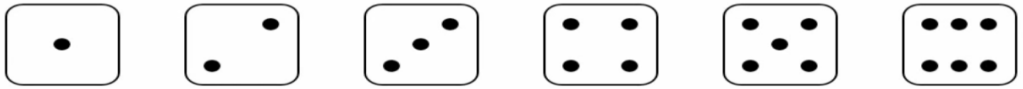
Probability of B

$P(A \cap B) = 3/10$

$P(B) = 5/10$

$P(A|B) = (3/10) / (5/10) = 3/5 = 0.6$

Exercise: Conditional Probability



$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

$$P(B \mid A) = \text{What is the Probability of } \left(\begin{array}{c} \text{rolling a dice and it's} \\ \text{value is less than 4} \end{array} \;\middle|\; \begin{array}{c} \text{knowing that the value is} \\ \text{an odd number} \end{array} \right)$$

Exercise: Conditional Probability

What is the Probability of

rolling a dice and it's value is 1

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

knowing that the value is an odd number

rolling a dice and it's value is 1