



# Data Science Workshop

# Workshop agenda:

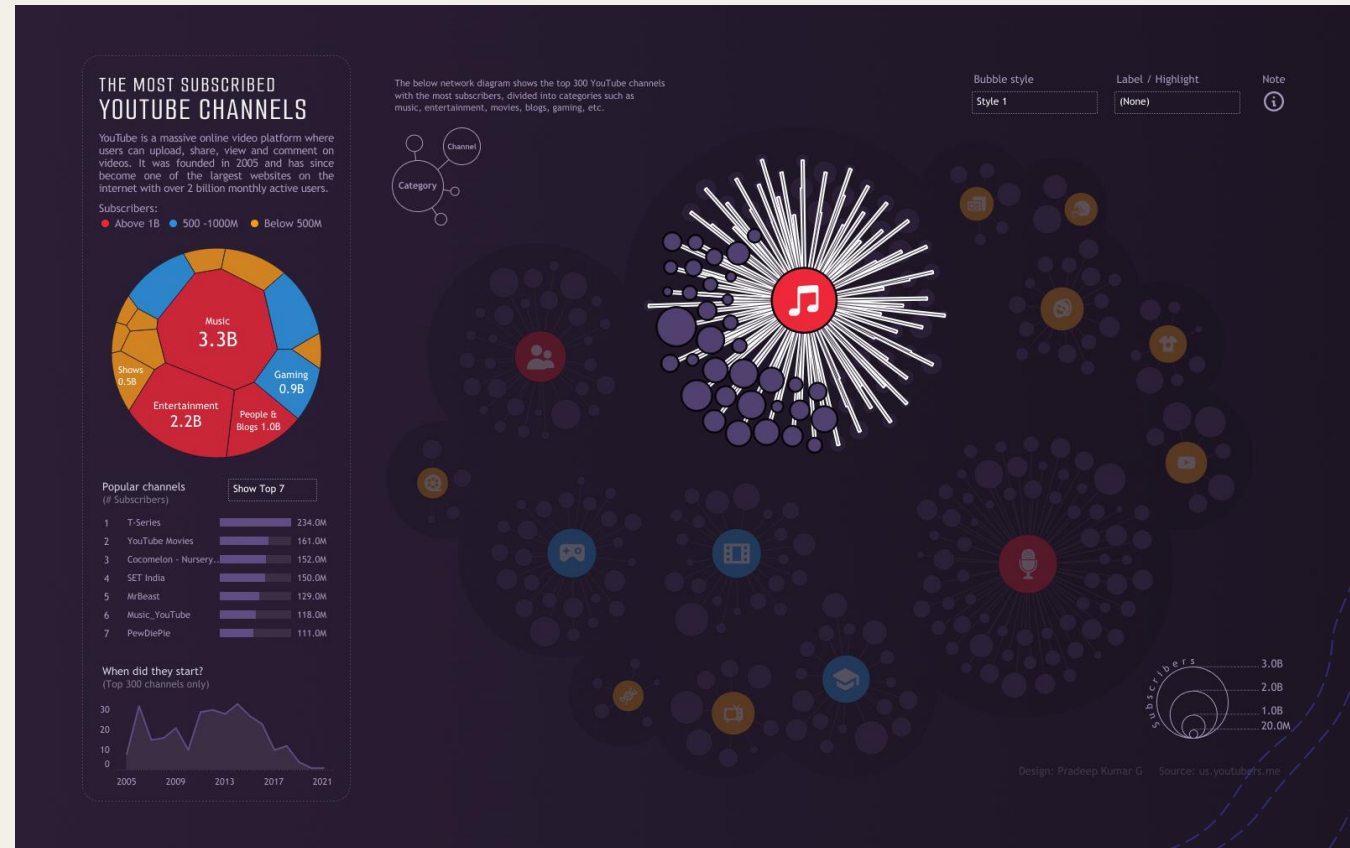
- + 1<sup>st</sup> day: web scrapping.
- + 2<sup>nd</sup> day: data preprocessing.
- + 3<sup>rd</sup> day: Data Analysis.
- + 4<sup>th</sup> day: machine learning & modeling.



# Data Analysis vs Data Science

Data Analyst responsibilities:

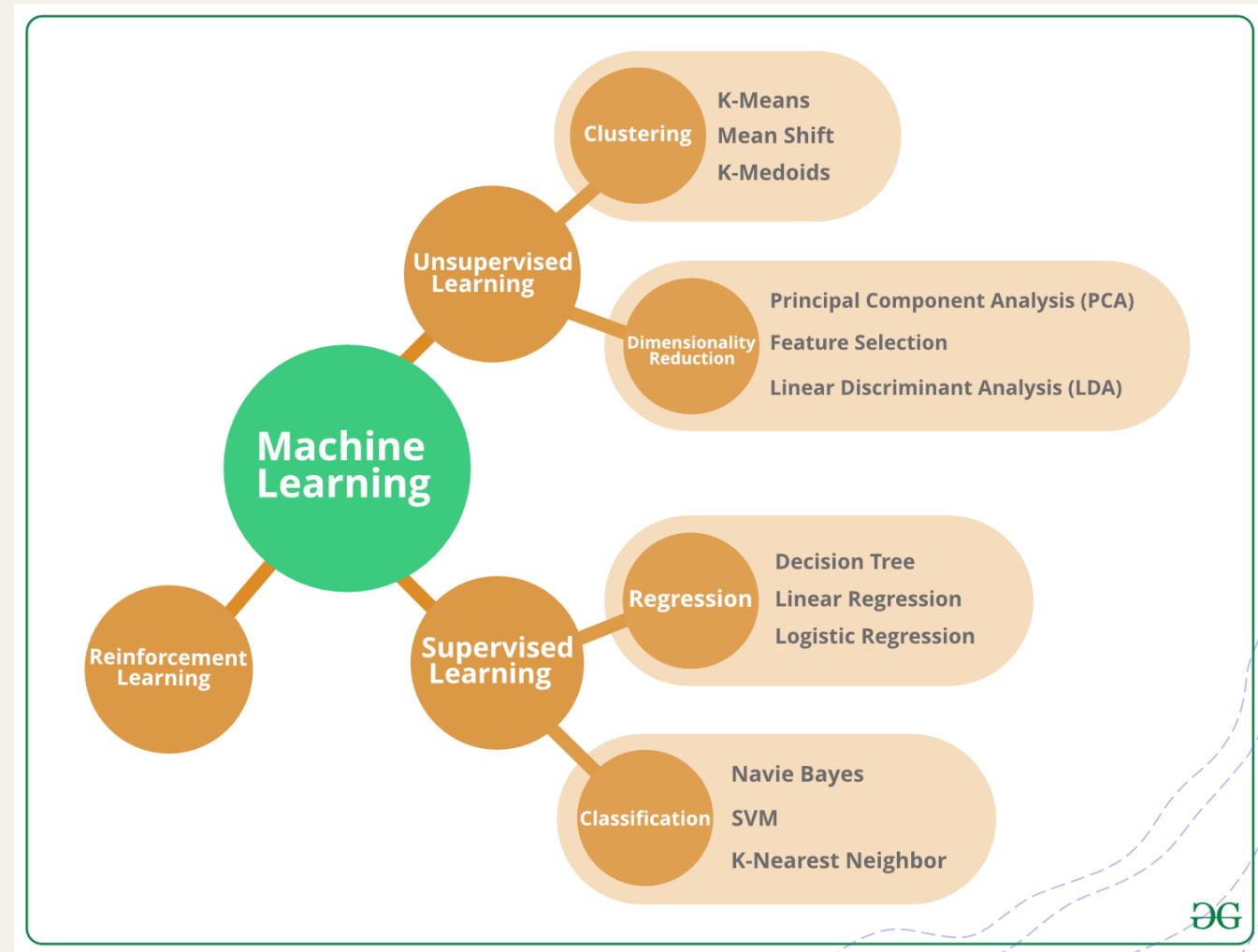
- + Collect, organize and maintain data.
- + Use statistics and other techniques to gain insights.
- + Provide visualizations and dashboards.
- + Help stakeholders make better decisions.



# Data Analysis vs Data Science

## Data Scientist responsibilities

- + Collect, organize, maintain data.
- + Basic analysis and visualizations.
- + Train machine learning models to predict target variables.



# Data Gathering

**A dataset:** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table.

## Types of data:

- Numerical data :
  - Continuous. (price)
  - Discrete. (Age)
- Categorical data:
  - Nominal (color).
  - Ordinal (economic status).

# Data Gathering

## + **Data Sources:**

- Kaggle Datasets. (powered by Google).
- Google Datasets Search engine.
- UCI Machine Learning Repository (University of California).
- Web Scrapping.

# Web Scraping

**Scraping:** Using tools to gather data you can see on a webpage.

A wide range of web scraping techniques and tools exist. These can be as simple as copy/paste and increase in complexity to automation tools, HTML parsing



# Web Scraping

**HTML: Hyper Text Mark-up Language**

```
1  <!DOCTYPE html>
2  <html>
3    <head>
4      <meta charset="UTF-8">
5      <title>Title goes here</title>
6    </head>
7    <body>
8
9    </body>
10 </html>
```



# Web Scraping

## + Web Scraping Ethics:

- Use a Public API when available.
- Check robots.txt file.
- Scrape data at a reasonable rate and throttle/control the number of requests per second. The website owner must not think it is a DDoS attack.
- Make sure website allows scraping and avoid scraping private data.
- Some websites allow scraping:
  - [Wuzzuf](#), [IMDB](#), [rottentomatoes](#), [transfermarkt](#)

# Web Scraping

## + Tools & APIs

- Scraper API. (5000 free API credits per month).
- Selenium.
- BeautifulSoup.
- Scrapy.

# Selenium

## + Requirements:

- Install a Selenium library:

```
Pip install selenium
```

- Install browser driver:
  - [Chrome version 111 web driver.](#)
  -



# Selenium

- + Set-up your environment:
- + I am going to scrape data from Wuzzuf web page, you are free to choose your field of interest.
- + For scraping script visit my notebook.

**Thank You**

