# INTRODUCTION TO DATA SCIENCE 01526105

Jakapun Tachaiya (Ph.D.)
Department of Computer Science
Faculty of Science, KMITL

**Intro To Data Science**
**Discord 2/2566**

# Outline

01 – What will be the focus  on this course?

02 – Course Logistics

03 – Example tasks on data analytic

04 – Revise python knowledge

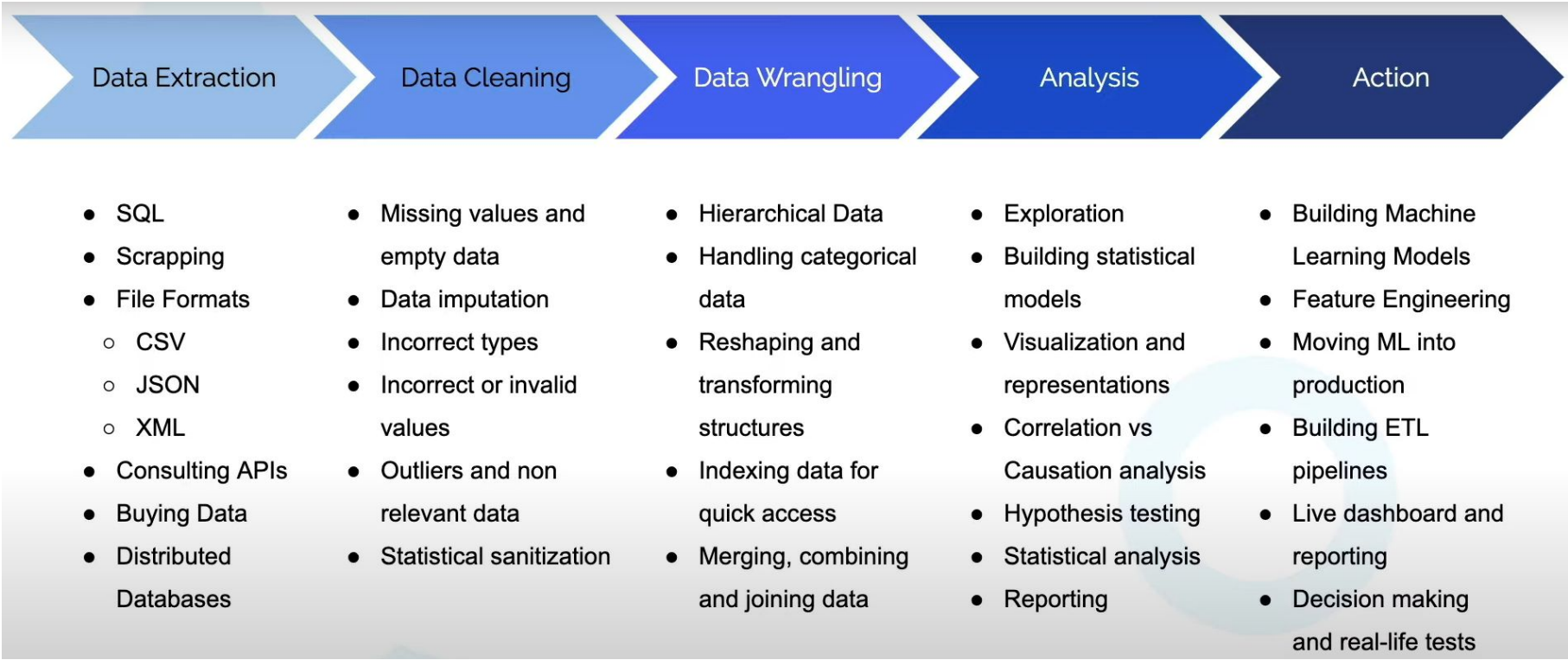
# 01

# What is data science?

# What is on data analytic?

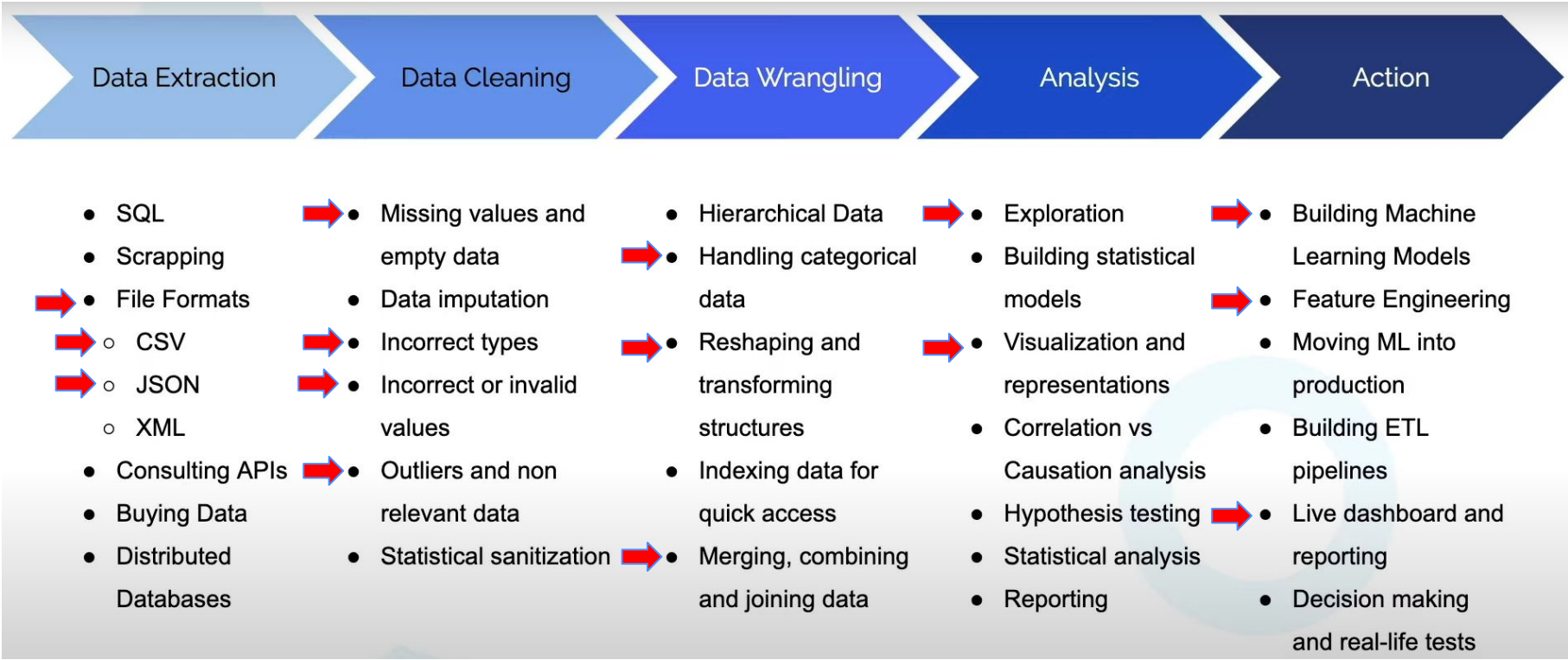
**Data analysis** is a process of **inspecting**, **cleansing**, **transforming**, and **modeling** data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

# Data analysis process (pipeline)




cite: https://www.freecodecamp.org/learn/data-analysis-with-python/data-analysis-with-python-course/introduction-to-data-analysis

# Data analysis process (pipeline)




cite: https://www.freecodecamp.org/learn/data-analysis-with-python/data-analysis-with-python-course/introduction-to-data-analysis

# Data analysis process (pipeline)


python™

NumPy

pandas

matplotlib

+ a b | e a u

Beautiful Soup

Natural Language
Tool Kit (NLTK)
Basic Text Analytics

python

scikit learn

# 02

# Course logistics

# Course logistic

## Course Management

- Slides and Assignment
  - Google Drive
- Google Colab: for coding and assignment.

## Lecture

- **Lecture:** Tue: 13:00 - 17:00 PM (in-class/ Zoom)

## My Info

- **Email:** jakapun.ta@kmitl.ac.th

# Course logistic

## Course Grading (100%)

**Assignments (30%)**

- HW1 - 15%
- HW2 - 15%

**Midterm (35%)**

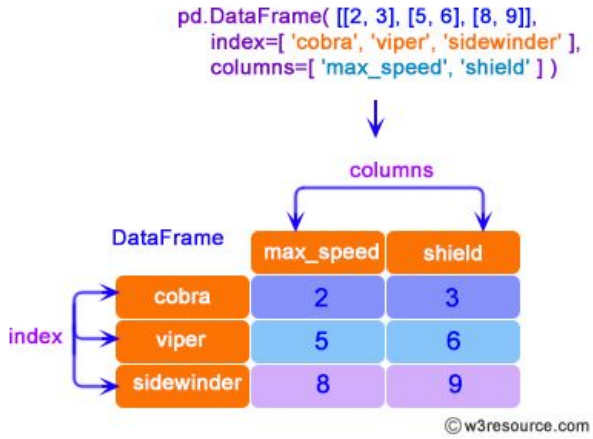
**Final (35%)**

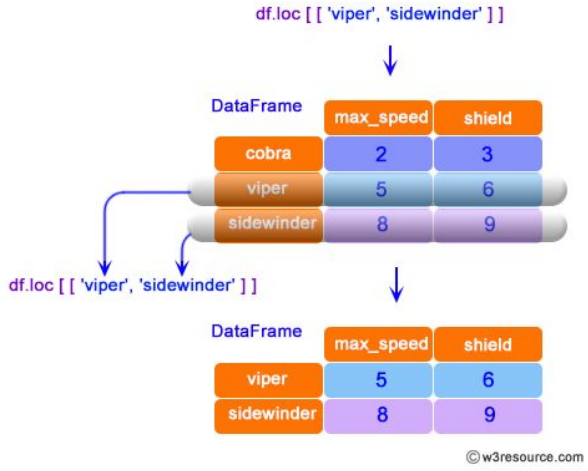
No Textbook require

# Tentative Schedule

| Week | Date | Topic | HW |
|---|---|---|---|
| 1 | 28 Nov | Introduction & outline<br>Revise Python basic | |
| 2 | 5 Dec | **---- No class (public holiday)----** | |
| 3 | 12 Dec | Pandas - Data Manipulation | |
| 4 | 19 Dec | Pandas - Merge Join table | **HW1** |
| 5 | 26 Dec | Data Visualization | |
| 6 | 2 Jan | **---- No class (New Year) -----** | **HW 1 Due** |
| 7 | 9 Jan | Text preprocessing | |
| 8 | 16 Jan | Natural language processing Tasks | |
| 9 | 23 Jan | **---- Midterm Exam ----** | |
| 10 | 30 Jan | Introduction to Machine Learning | |
| 11 | 6 Feb | Regression & evaluation | |
| 12 | 13 Feb | Regression & evaluation | |
| 13 | 20 Feb | Supervised Learning: Classification | |
| 14 | 27 Feb | Supervised Learning: Classification | **HW 2** |
| 15 | 5 Mar | Unsupervised learning: Clustering | |
| 16 | 12 Mar | Text mining | **HW 2 Due** |
| 17 | 19 Mat | **---- Final Exam ----** | |

pandas

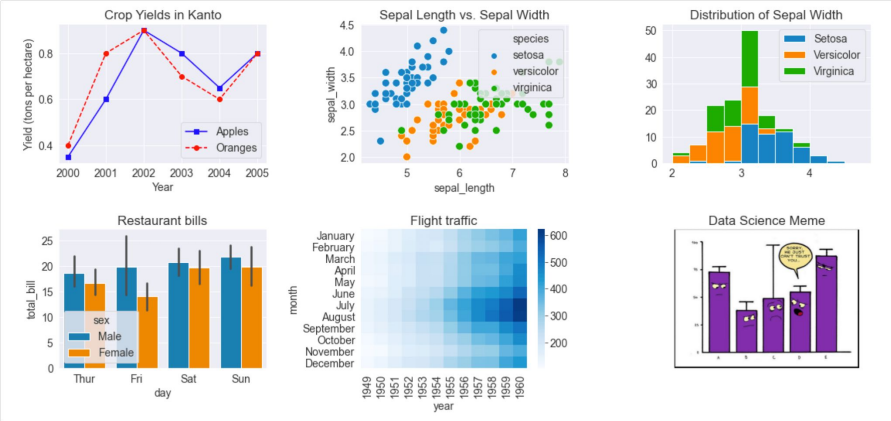

pd.DataFrame( [[2, 3], [5, 6], [8, 9]],
index=[ 'cobra', 'viper', 'sidewinder' ],
columns=[ 'max_speed', 'shield' ] )
columns
DataFrame
max_speed
shield
index
cobra
2
3
viper
5
6
sidewinder
8
9
©w3resource.com


df.loc [ [ 'viper', 'sidewinder' ] ]
DataFrame
max_speed
shield
cobra
2
3
viper
5
6
sidewinder
8
9
df.loc [ [ 'viper', 'sidewinder' ] ]
DataFrame
max_speed
shield
viper
5
6
sidewinder
8
9
©w3resource.com

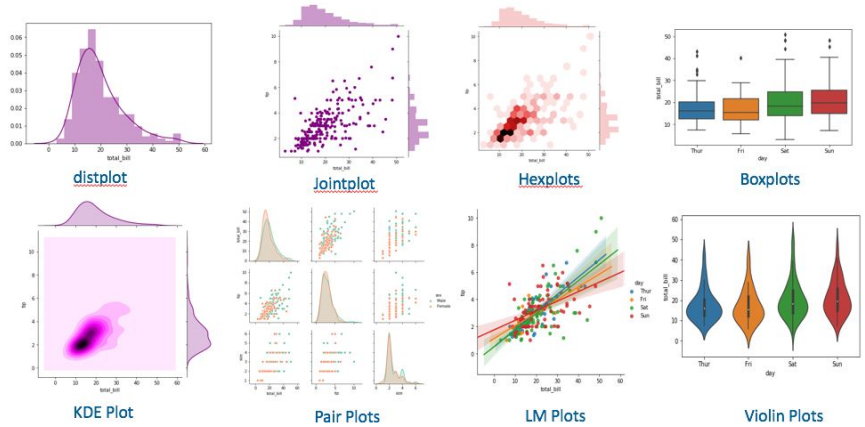
# Data Visualization

seaborn

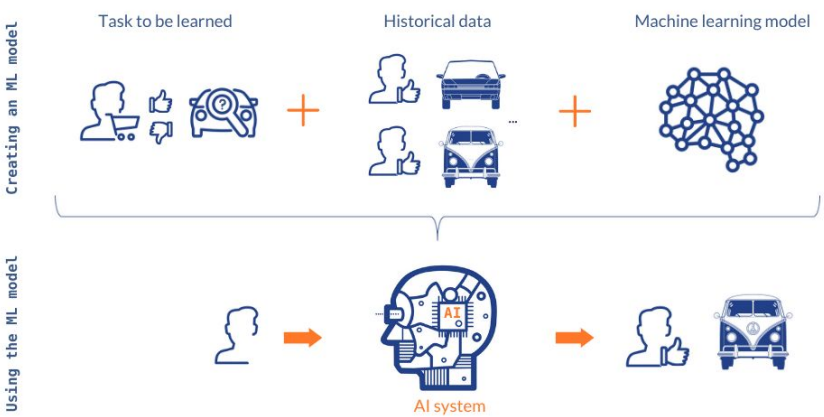
matplotlib

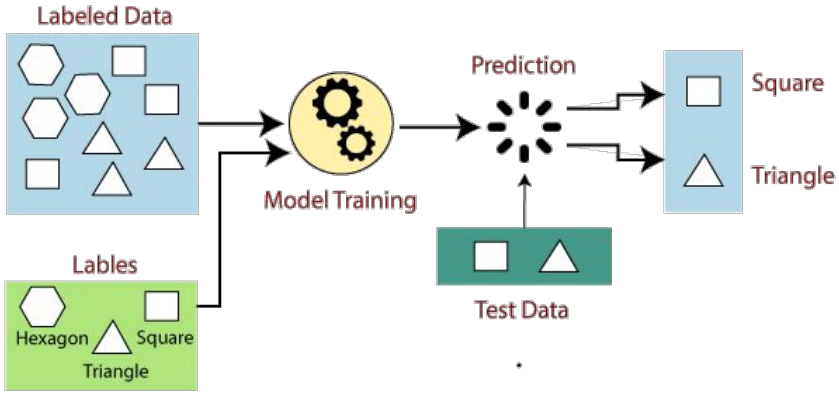



## Seaborn Plots



# Machine learning (AI)



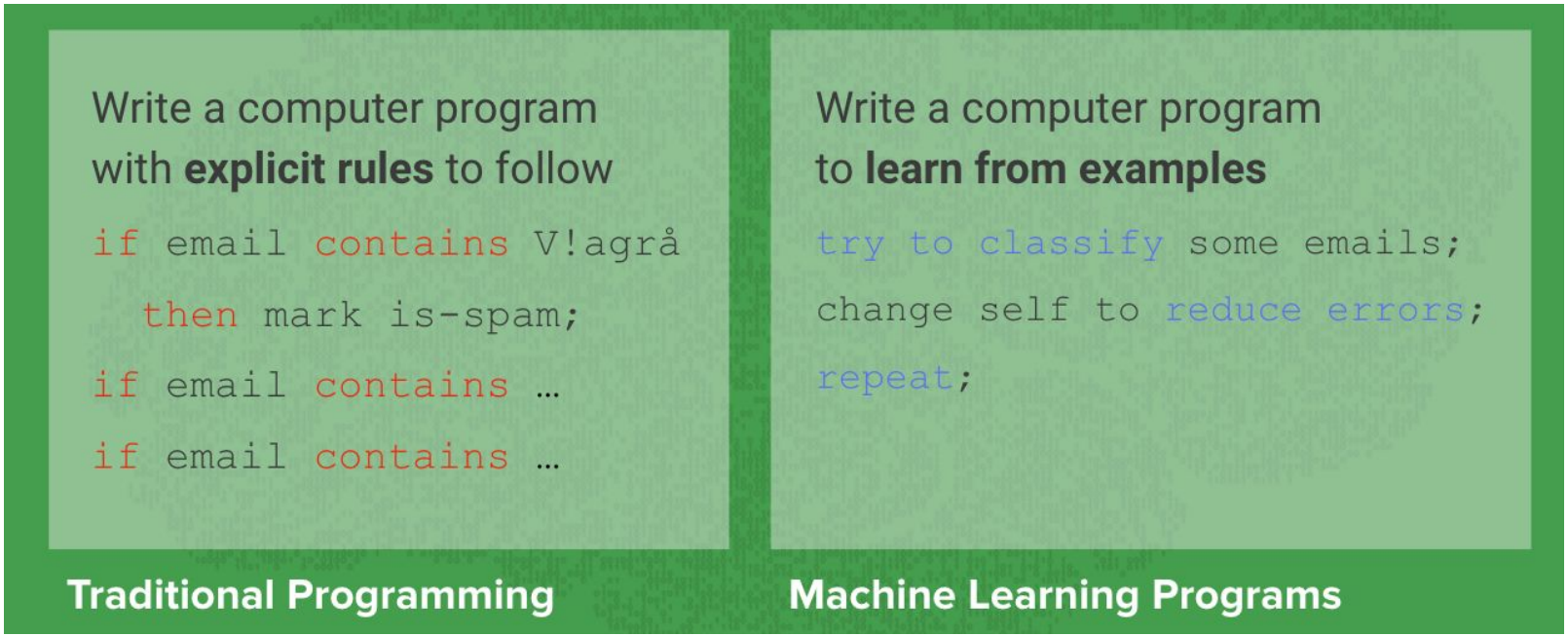



# What is Machine Learning?

"Machine learning … gives computers the ability to learn without being explicitly programmed."

**Arthur Samuel**

# Conditioning VS machine learning



# 03

# Examples on data analytic

# Example – Titanic Survival Analysis

```
df = pd.read_csv('assets/train.csv')
df.head()
```

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

# Example – Titanic Survival Analysis

Use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

- Classical tasks
  - Feature **Importance** and **Correlations**
  - Predict **survival class (0/1)** of passengers based on **features.**

- Dataset **https://www.kaggle.com/c/titanic**
  - 1309 passengers (891 for train & 418 for test)
  - 11 features
  - 10 source variable vs Target variable (survived)

**Data Dictionary**

| Variable | Definition | Key |
|---|---|---|
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex | |
| Age | Age in years | |
| sibsp | # of siblings / spouses aboard the Titanic | |
| parch | # of parents / children aboard the Titanic | |
| ticket | Ticket number | |
| fare | Passenger fare | |
| cabin | Cabin number | |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

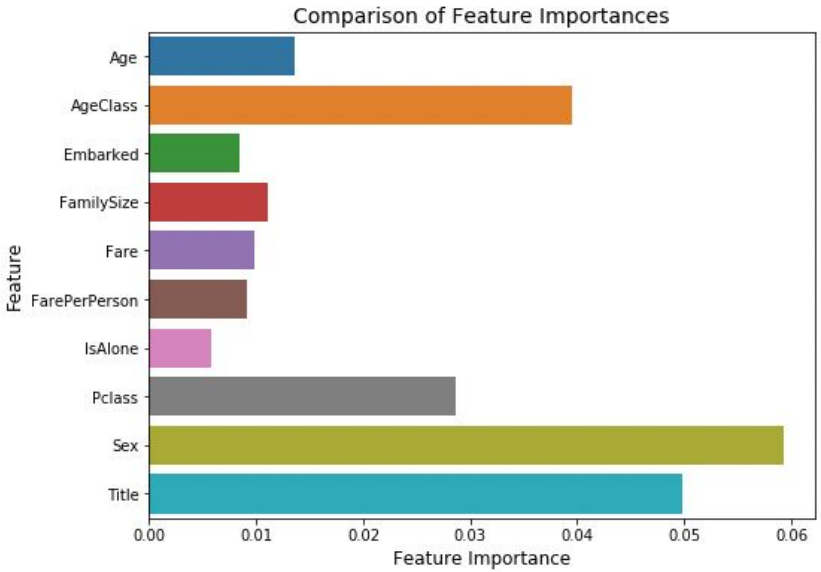
train.csv

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
```

# Example – Titanic Survival Analysis

Feature **Importance** (selection)

- Identify the most significant features from a given dataset.
- **Sex** and **Title** are significant factors in surviving the Titanic Disaster.

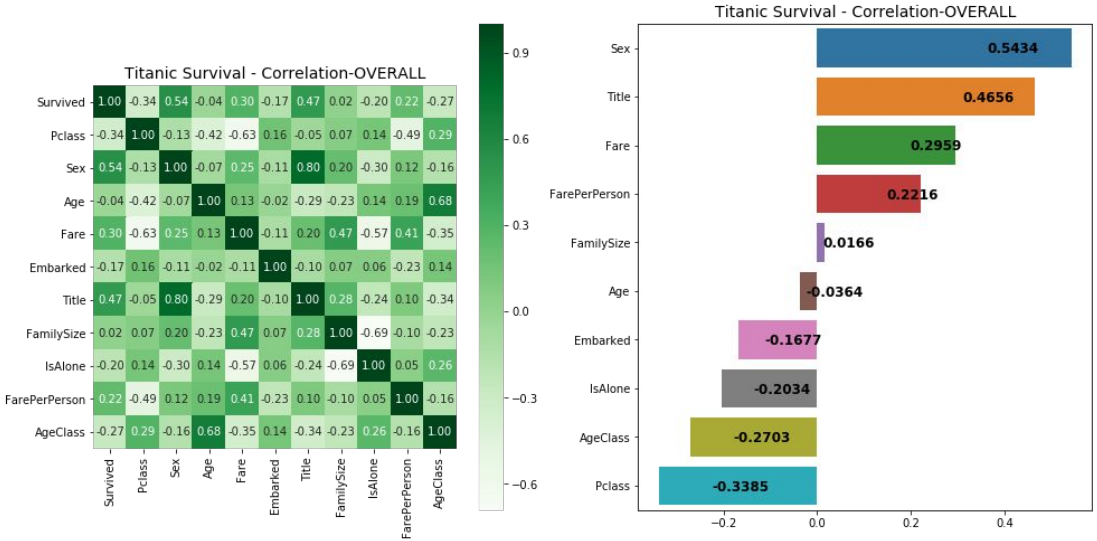



https://www.kaggle.com/code/asimislam/titanic-disaster-fe-eda-corr-ml/notebook

# Example – Titanic Survival Analysis

Feature **Correlations**

- Measure of the linear relationship (y=ax) of 2 or more variables.
  - E.g. Correlation between the height of parents and their offspring,
- Statistical test of association between variables. (-1 to 1 scale)
- **Strong** association is value close to -1 or 1.
- **Weak** association is value close to 0



# Example – Titanic Survival Analysis

Predict **survival class (0/1)** of passengers based on **features.**

- Dataset 1309 passengers
  - (891 for train & 418 for test)
- KNN as a classifier model with the
  - best F1 score on cross validation.
  - 10 – fold cross validation.

```
Train set: (712, 10) (712,)
Test set: (179, 10) (179,)
```

]:

| | Pclass | Sex | Age | Fare | Embarked | Title | FamilySize | IsAlone | FarePerPerson | AgeClass |
|---|---|---|---|---|---|---|---|---|---|---|
| 205 | 1.0 | 1.0 | 0.00 | 0.333333 | 1.0 | 1.0 | 0.1 | 0.0 | 0.010211 | 0.000000 |
| 718 | 1.0 | 0.0 | 0.25 | 0.666667 | 0.5 | 0.0 | 0.0 | 1.0 | 0.030254 | 0.250000 |
| 835 | 0.0 | 1.0 | 0.50 | 1.000000 | 0.0 | 1.0 | 0.2 | 0.0 | 0.054105 | 0.166667 |
| 851 | 1.0 | 0.0 | 1.00 | 0.000000 | 1.0 | 0.0 | 0.0 | 1.0 | 0.015176 | 1.000000 |
| 773 | 1.0 | 0.0 | 0.25 | 0.000000 | 0.0 | 0.0 | 0.0 | 1.0 | 0.014102 | 0.250000 |

| | model | CV-mean | CV-std | AccuracyScore |
|---|---|---|---|---|
| 1 | KNN | 0.823142 | 0.036083 | 0.798883 |

https://www.kaggle.com/code/asimislam/titanic-disaster-fe-eda-corr-ml/notebook

# Example – Text analysis on The Simpsons

The Simpsons meets data analytic.

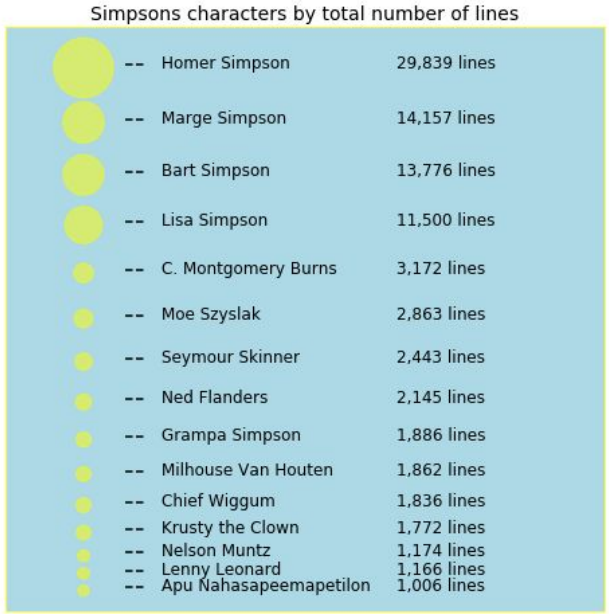
- Extract all lines of characters in popular TV series.

```
id,episode_id,number,raw_text,timestamp_in_ms,speaking_line,character_id,location_id,raw_character_text,raw_location_text,spoken_words,normalized_text,word_count
9549,32,209,"Miss Hoover: No, actually, it was a little of both. Sometimes when a disease is in all the magazines and all the news shows, it's only natural that you think you have it.",848000,true,464,3,Miss Hoover,Springfield Elementary School,"No, actually, it was a little of both. Sometimes when a disease is in all the magazines and all the news shows, it's only natural that you think you have it.",no actually it was a little of both sometimes when a disease is in all the magazines and all the news shows its only natural that you think you have it,31
9550,32,210,Lisa Simpson: (NEAR TEARS) Where's Mr. Bergstrom?,856000,true,9,3,Lisa Simpson,Springfield Elementary School,Where's Mr. Bergstrom?,wheres mr bergstrom,3
9551,32,211,Miss Hoover: I don't know. Although I'd sure like to talk to him. He didn't touch my lesson plan. What did he teach you?,856000,true,464,3,Miss Hoover,Springfield Elementary School,I don't know. Although I'd sure like to talk to him. He didn't touch my lesson plan. What did he teach you?,i dont know although id sure like to talk to him he didnt touch my lesson plan what did he teach you,22
9552,32,212,Lisa Simpson: That life is worth living.,864000,true,9,3,Lisa Simpson,Springfield Elementary School,That life is worth living.,that life is worth living,5
9553,32,213,"Edna Krabappel-Flanders: The polls will be open from now until the end of recess. Now, (SOUR) just in case any of you have decided to put any thought into this, we'll have our final statements. Martin?",864000,true,40,3,Edna Krabappel-Flanders,Springfield Elementary School,"The polls will be open from now until the end of recess. Now, just in case any of you have decided to put any thought into this, we'll have our final statements. Martin?",the polls will be open from now until the end of recess now just in case any of you have decided to put any thought into this well have our final statements martin,33
```

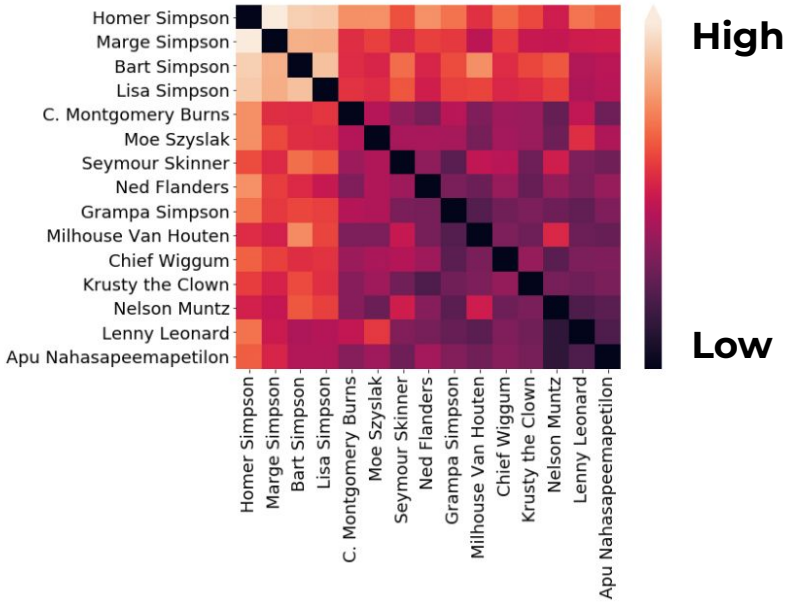
https://towardsdatascience.com/the-simpsons-meets-data-visualization-ef8ef0819d13

# Example – Text analysis on The Simpsons

- Who has most to say?
- 15 characters with the most lines

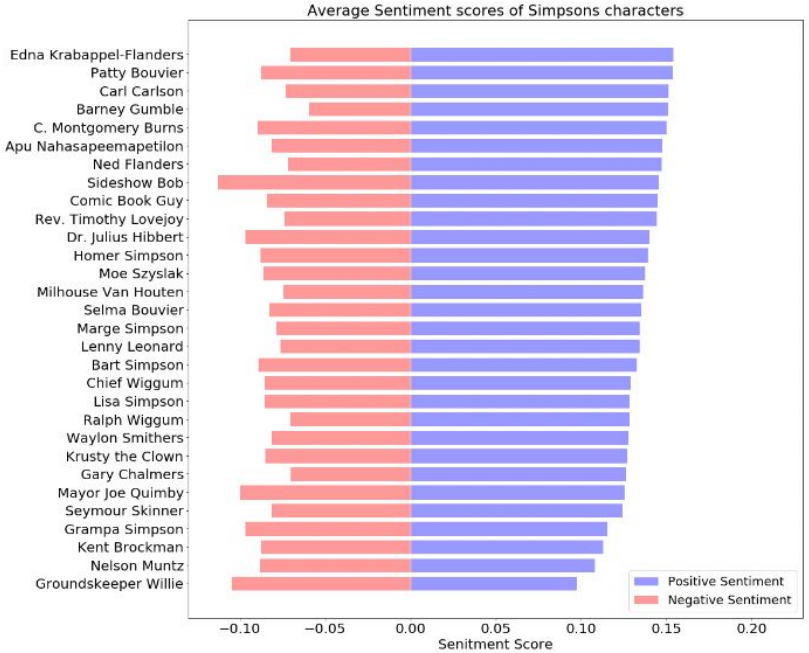



- Who speak to whom?
  a. **many** conversations internally between the Simpson family.
  b. **medium** amount of conversations between Simpson family members and the side characters.
  c. **few** conversations that do not involve the Simpson family.



# Example – Text analysis on The Simpsons

- What are they saying?
  - VADER Sentiment Analysis




- What are the most used words on characters?
  - Word clouds


Word Clouds — Word Spoken by MARGE, BART AND HOMER respectively

https://medium.com/datafloq/which-28-seasons-data-tells-us-about-simpson-characters-e1440f5f8788

# Tools and libraries







Natural Language
Tool Kit (NLTK)
Basic Text Analytics

# Why using open source tools?

## Commercial tools

- 🔼 Easy to use/learn
- 🔻 Closed source
- 🔻 Costly
- 🔻 Limited

## Open source tools (Programming)

- 🔼 Open source (mostly free)
- 🔼 Flexibility/adaptability
- 🔼 Tuning performance (Faster)
- 🔻 Hard to learn (know how to code)

# Why Python for data science and text analysis?

🔼 Simple and easy to learn

🔼 Flexibility/connectivity

🔼 Free and open source

🔼 accessibility to docs, references, and tutorials

🔻 Slower than C

**Python**

```
print("Hello world.")
```

vs.

**Java**

```
public class HelloWorld {
    public static void main (String[]args) {
      System.out.println.("Hello world");
    }
}
```

# 04
# Revise python knowledge

# Ways to use python

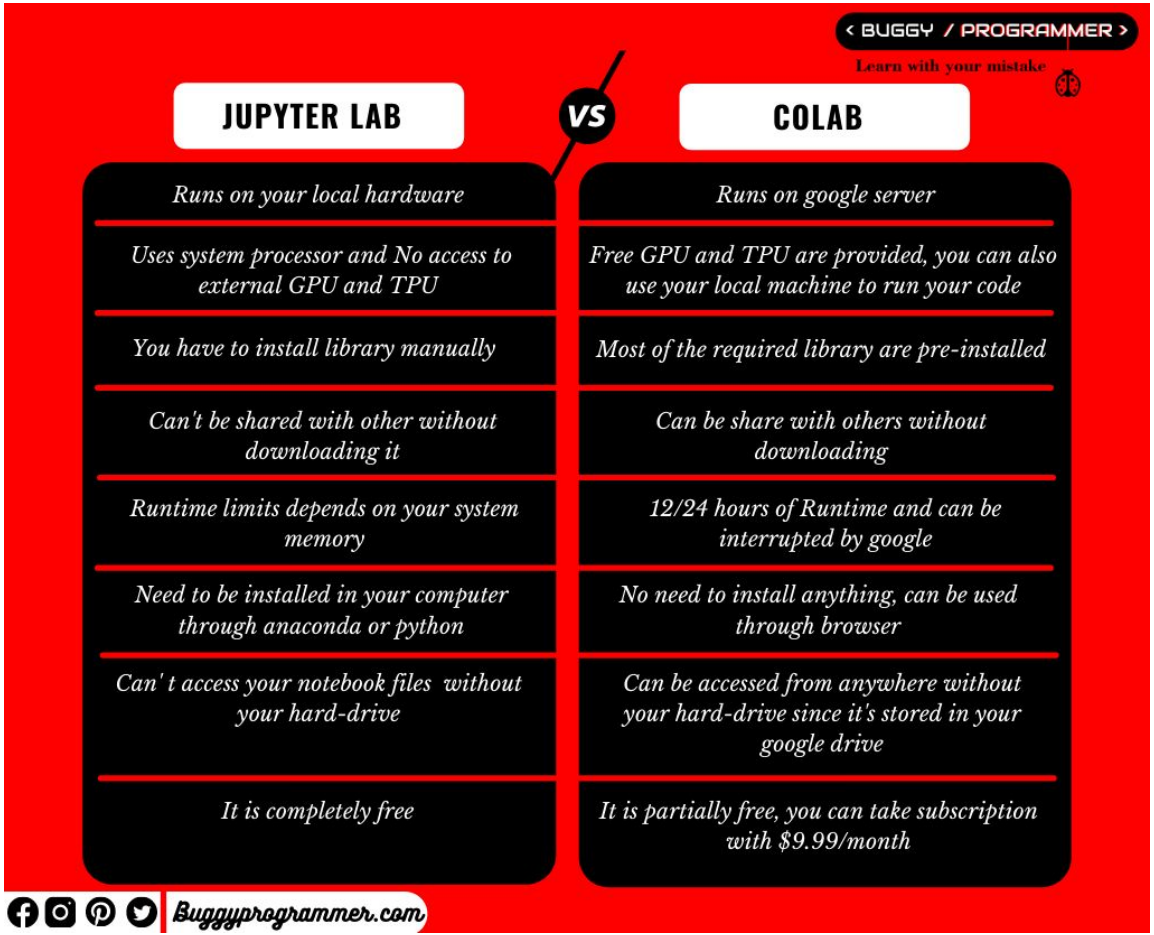
**Python in local machine**

1. Install python in local machine + text editor(Jupyter Notebook), IDE(Pycharm)
2. Install python in local machine via package management (Anaconda)
    a. Limited by machine performance.

**Python in Cloud**

1. Google Colab
    a. Pre-Installed Libraries
    b. Saved on the Cloud
    c. Collaboration
    d. Free GPU and TPU Use

**Python in local machine** **Python in Cloud (Google Colab)**



# How to use python in Google Colab

DEMO