SORBONNE UNIVERSITY ABU DHABI

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

A GUIDED TOUR OF AI: FROM FOUNDATIONS TO LATEST APPLICATION

HANDS ON SESSION MANIPULATING DATASETS

The World is Data Rich

Recent Buzz Words

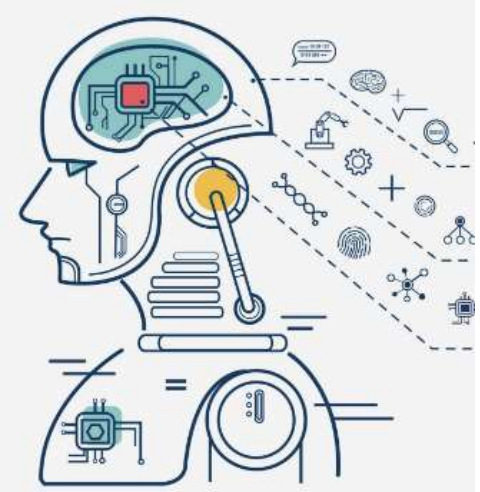
1. Statistics is the study of the collection, analysis, interpretation, presentation and organization of data.

2. Data science is the study of the generalizable extraction of knowledge from data, yet the key word is science.

3. Machine learning is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed.

4. Artificial Intelligence research is defined as the study of intelligent agents: any device that perceives its environment and takes actions that maximize its chance of success at some goal.

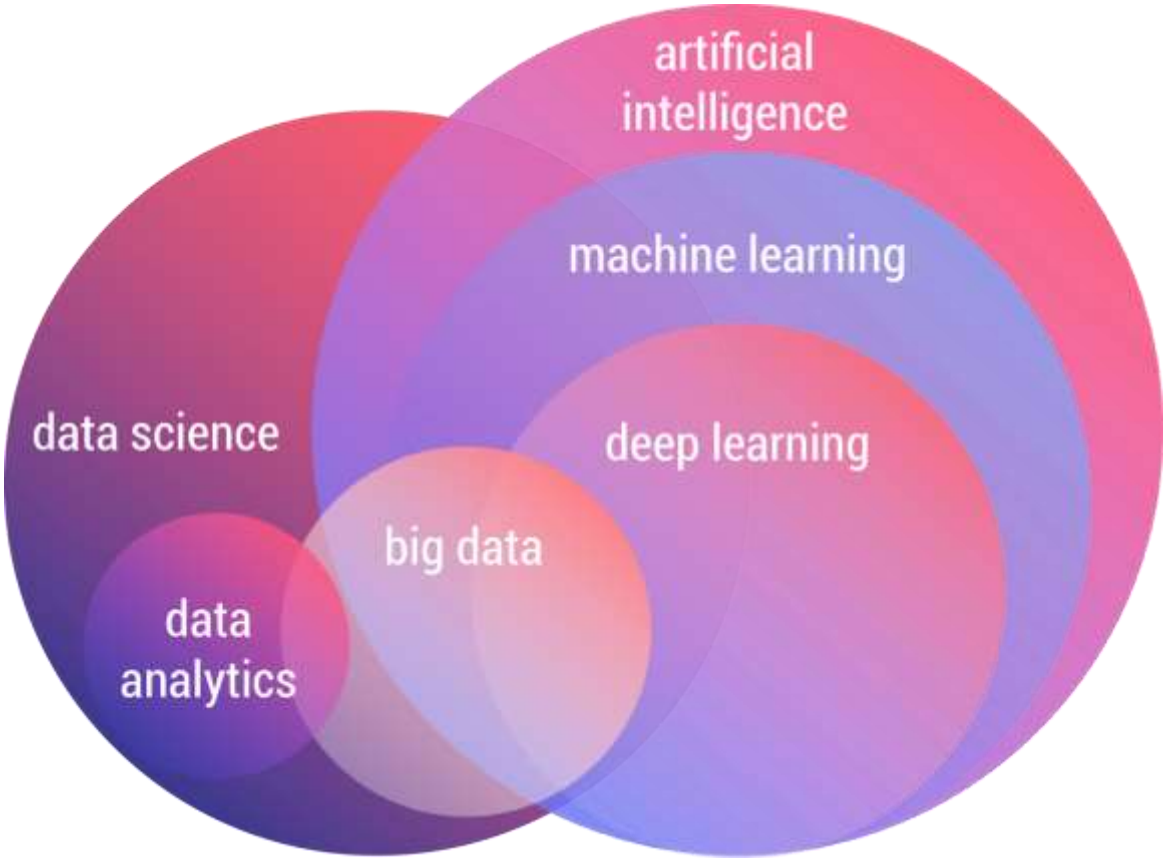
5. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.



Statistics

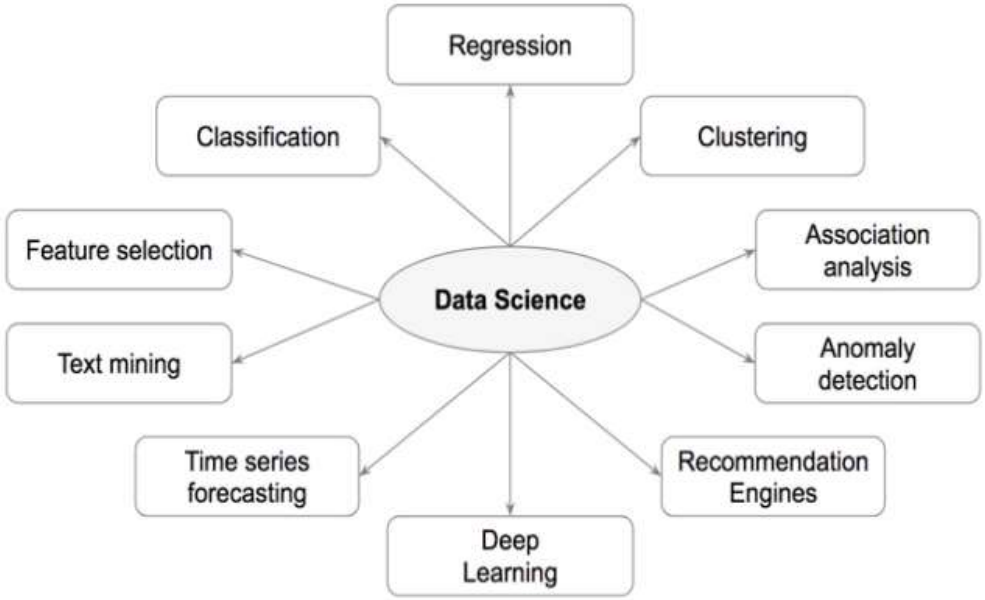
- Data : Large bodies of data with complex data structures are generated from computers, sensors, manufacturing industries, etc.
- Models : Non/Semiparametric models but in complex probability spaces / high-dimensional functional spaces (e.g., deep neural net, reinforcement learning, decision trees, etc.).
- Emphases : Making predictions, causation, algorithmic convergence.
- Data are necessary and at the core of Statistical Learning, Data Science & Machine Learning.
- Statistics : Not only has strong interactions with Probability but also other parts of Data Science (Machine Learning, Artificial Intelligence, etc.).

Data Science



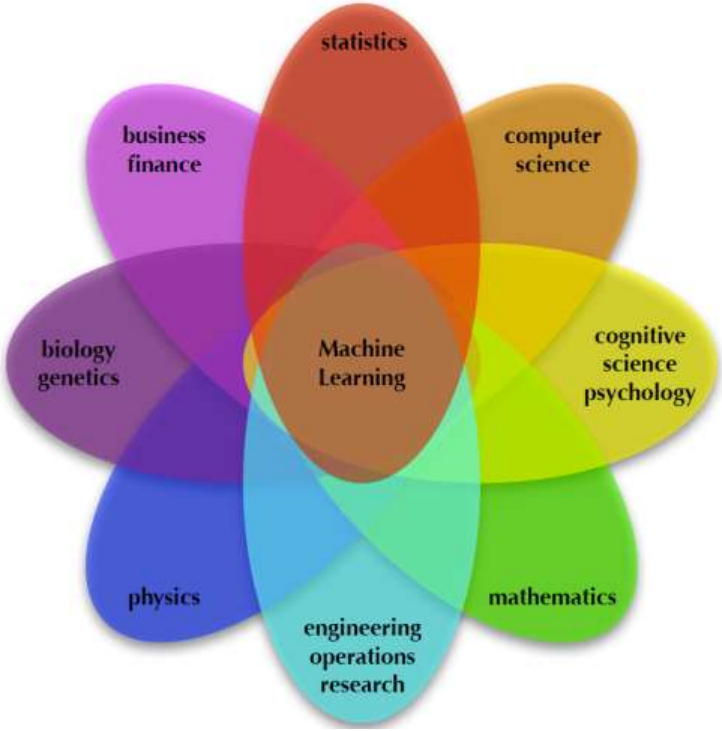
"***Data is the new oil.*** It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."
- Clive Humby, UK Mathematician and Architect of Tesco's Clubcard.

"***When you're fundraising,*** it's AI.
When you're hiring, it's ML.
When you're implementing, it's Linear Regression.
When you're debugging, it's printf()."
- Baron Schwartz, Founder and CEO of VividCortex, 2017.

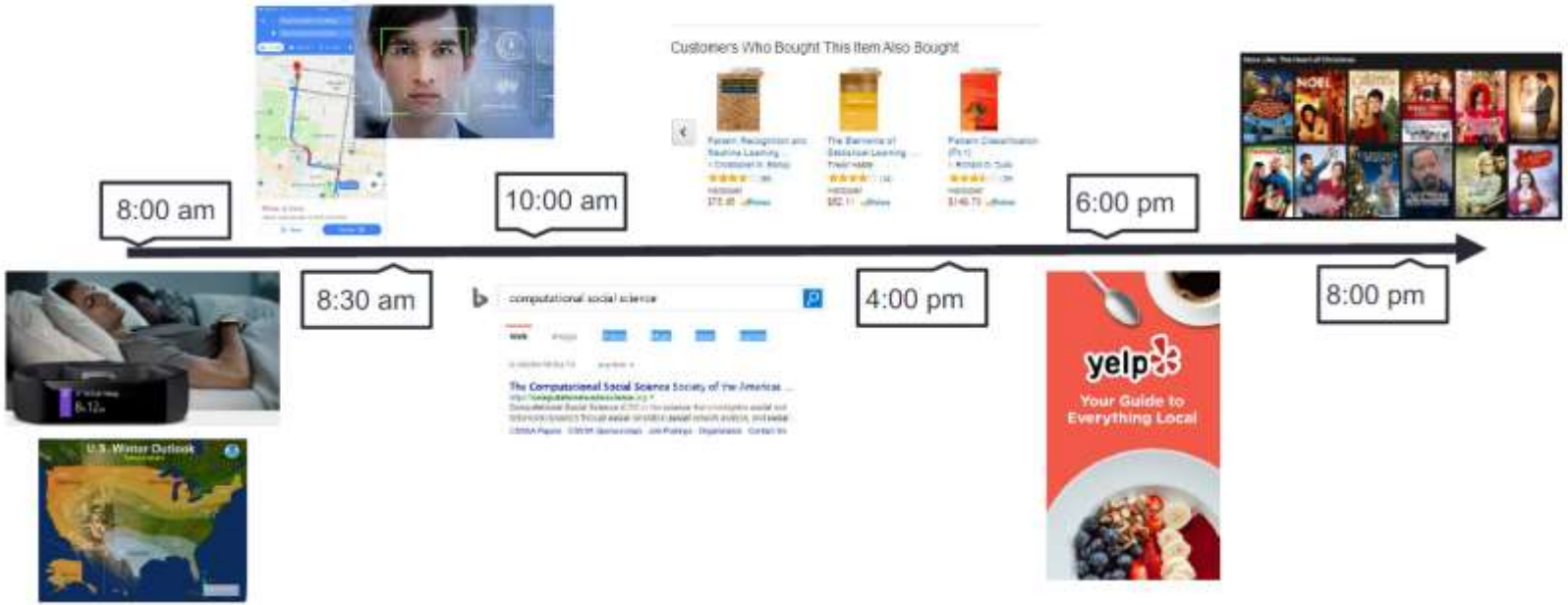


Machine Learning

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.



Machine Learning is impacting our life



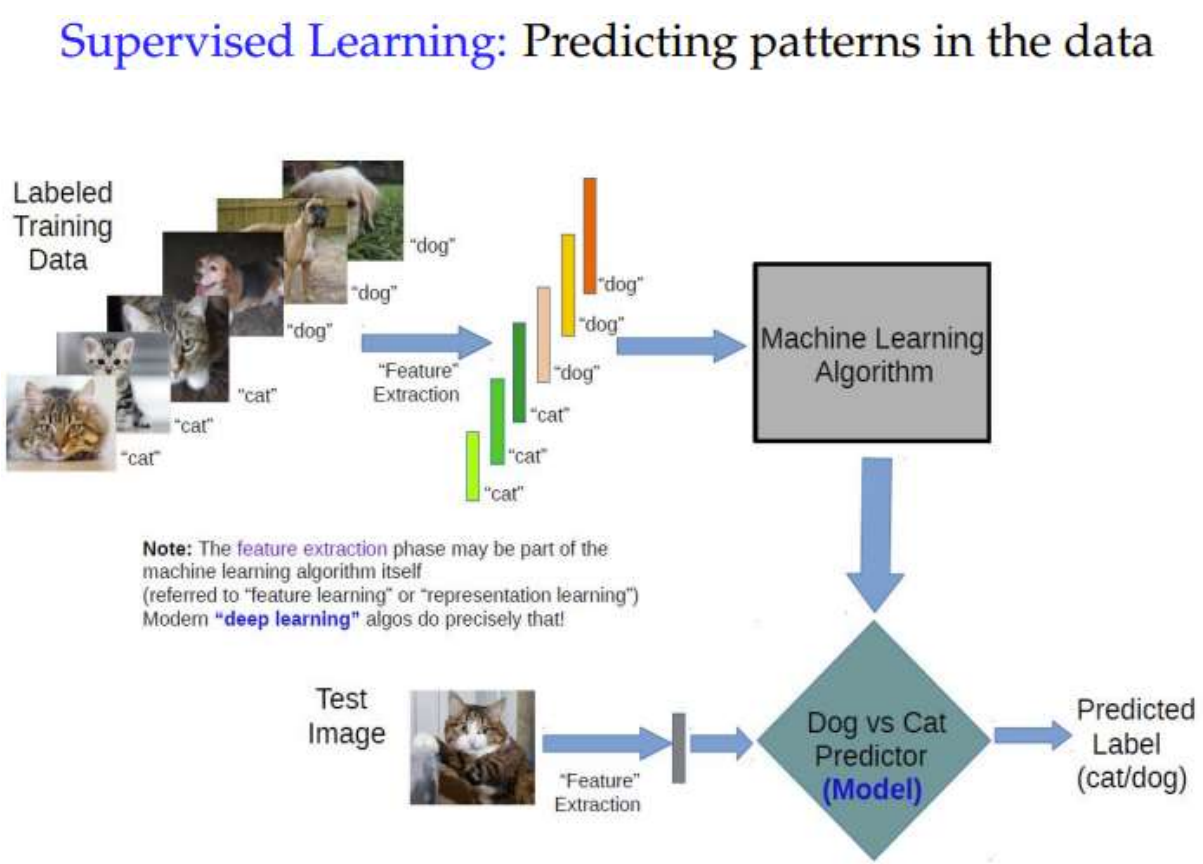
What Machine Learning can do?

Yes

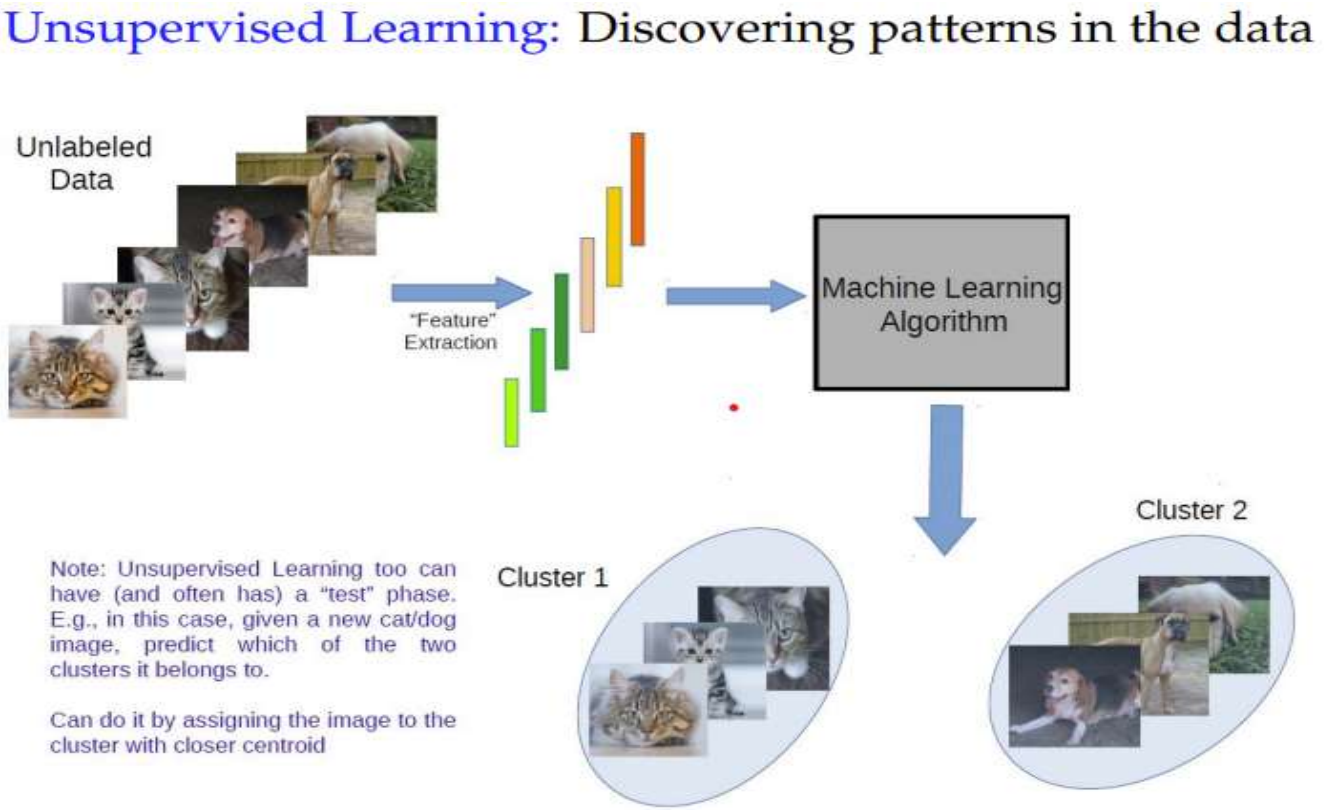
Maybe

No

Supervised Learning Workflow



Unsupervised Learning Workflow



Data Summarization

- To identify the typical characteristics of data (i.e., to have an overall picture).
- To identify which data should be treated as noise or outliers.
- The data summarization techniques can be classified into two broad categories:
  - Measures of **location**
  - Measures of **dispersion**

Measurement of location

- It is also alternatively called as measuring the central tendency.
  - A function of the sample values that summarizes the location information into a single number is known as a measure of location.

- The most popular measures of location are
  - Mean
  - Median
  - Mode
  - Quartile

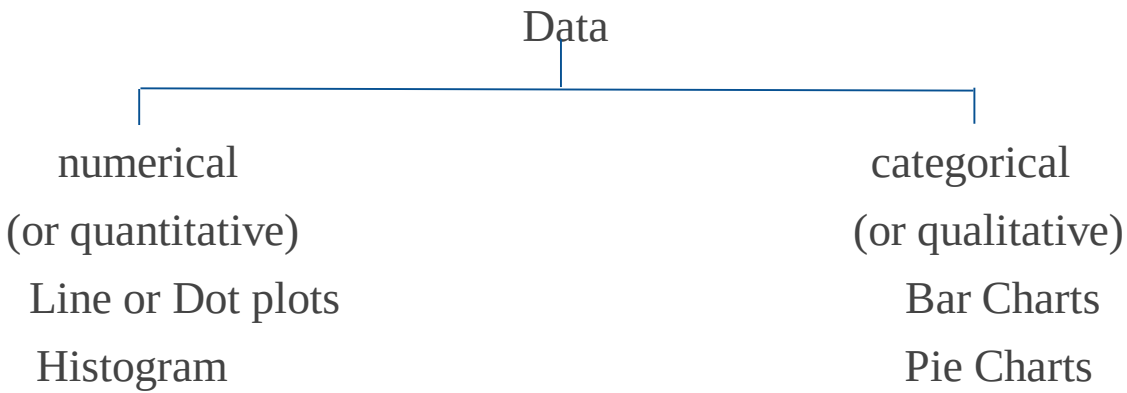
Measures of dispersion

- Location measure are far too insufficient to understand data.
- Another set of commonly used summary statistics for continuous data are those that measure the dispersion.
- A dispersion measures the extent of spread of observations in a sample.

- Some important measure of dispersion are:
  - Range
  - Variance and Standard Deviation
  - Mean Absolute Deviation (MAD)
  - Absolute Average Deviation (AAD)
  - Interquartile Range (IQR)

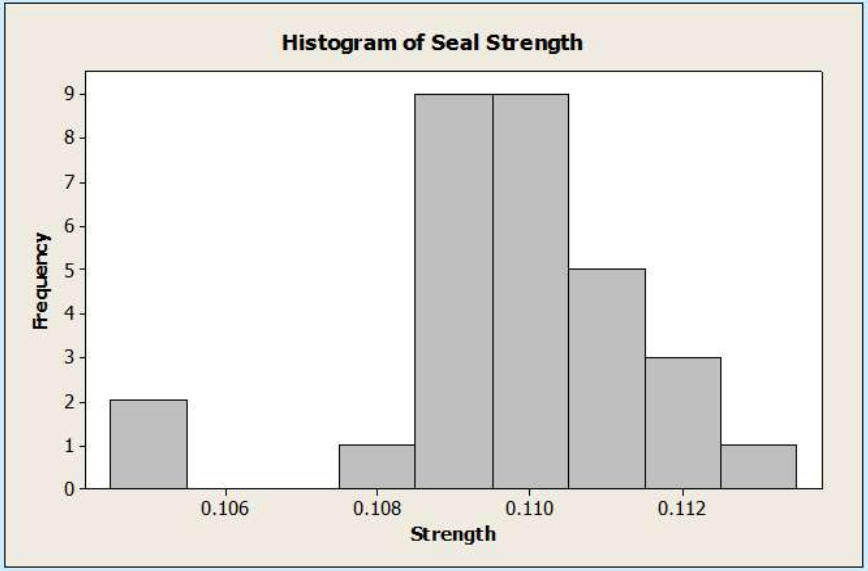
Graphical Representation of Data

- Visualization techniques are ways of creating and manipulating graphical representations of data.
- We use these representations in order to gain better insight and understanding of the problem we are studying - pictures can convey an overall message much better than a list of numbers.



Histogram

Histogram is a basic graphing tool that displays the relative frequency or occurrence of continuous data values showing which values occur most and least frequently.



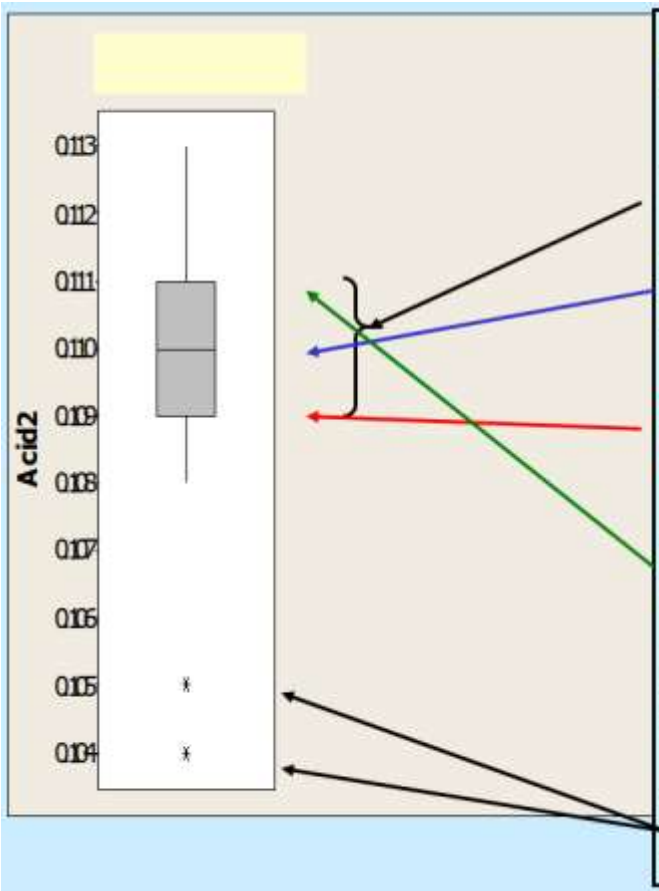
A histogram illustrates the

Shape,
Centering, and
Spread

of data distribution

and indicates whether there are any outliers.

Box Plot



A box and whisker plot provides a 5 point summary of the data.

1)The box represents the middle 50% of the data.

2) The median is the point where 50% of the data is above it and 50% below it.

3) The 1st quartile is where, 25% of the data fall below it.

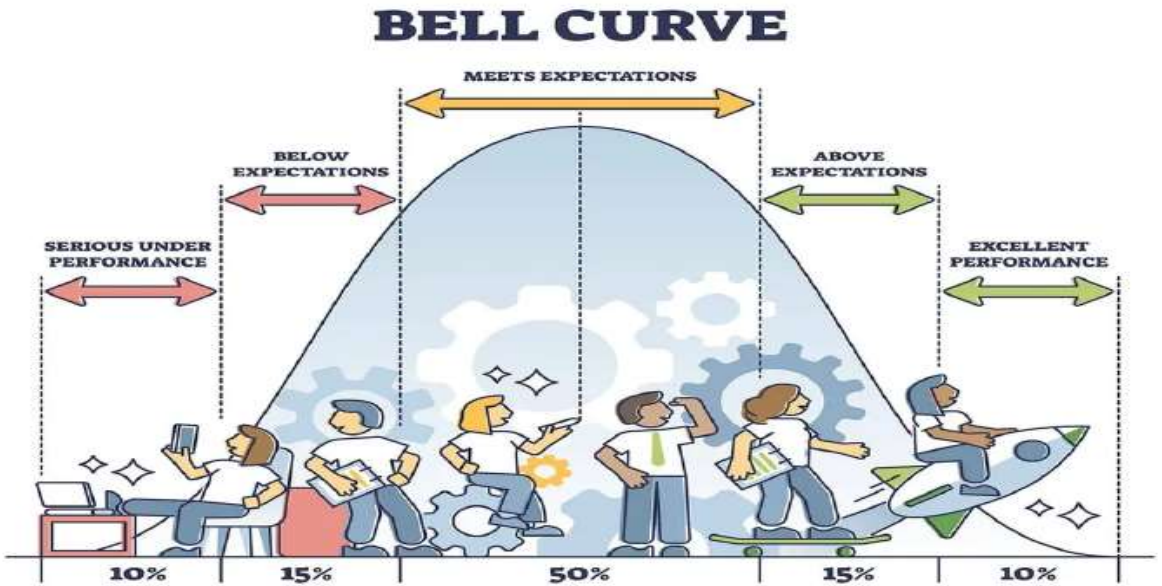
4) The 3rd quartile is where, 75% of the data is below it.

5) The whiskers cannot extend any further than 1.5 times the length of the inner quartiles.

If you have data points outside this, they will show up as outliers.

Normality and Statistics

Normality is a paved road. It is easy to walk but no flowers grow on it. — Vincent Van Gogh.



By Dr. Saul McLeod (2019)

Lets try some exercise in Python

Exercise 1: The monthly credit card expenses of an individual in 1000 rupees is given in the file Credit_Card_Expenses.csv.

a. Read the dataset to Python

b. Compute mean, median minimum, maximum, range, variance, standard deviation, skewness, kurtosis and quantiles of Credit Card Expenses

c. Compute default summary of Credit Card Expenses

d. Draw Histogram of Credit Card Expenses

Descriptive Statistics Using Python

Reading a csv file from local drive

```
from google.colab import files
uploaded = files.upload()
import io
import pandas as pd
data = pd.read_csv(io.BytesIO(uploaded['Credit_Card_Expenses.csv']))
data.head(5)  #shows the first 5 examples of the data
```

To read a particular column or variable of data set to a new variable

Example: Read CC_Expenses to CC

```
cc = mydata.CC_Expenses
cc
```

Descriptive Statistics Using Python

Descriptive Statistics

Statistics	Code
Summary	cc.describe()

Statistics	Value
Count	20
Mean	59.2
Standard Deviation	3.1052
Minimum	53
Q1	57
Median	59
Q3	61
Maximum	65

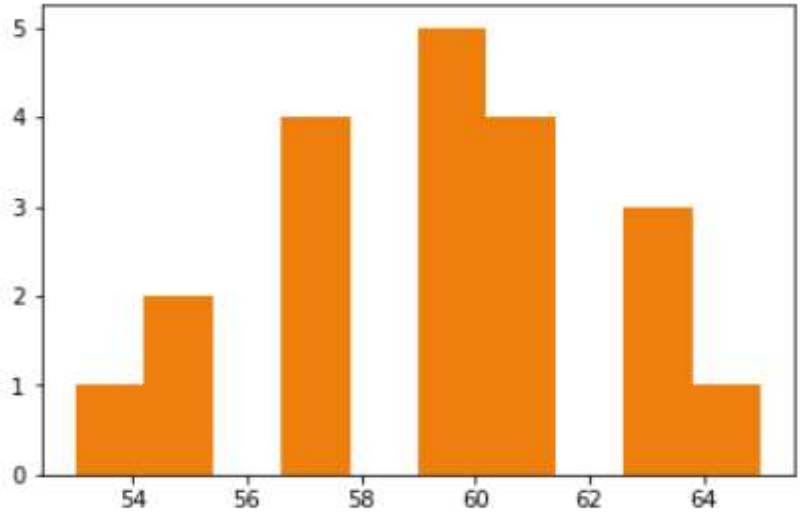
Descriptive Statistics Using Python

Graphs:

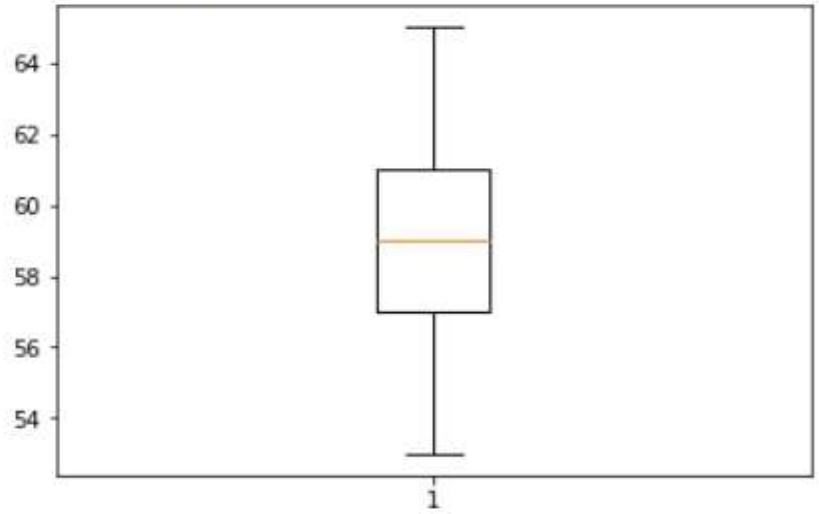
Graph	Code
Histogram	import matplotlib.pyplot as myplot myplot.hist(cc) myplot.show()
Box Plot	myplot.boxplot(cc) myplot.show()

Graphs:

Histogram



Box plot



Correlation Analysis

In statistics, the word correlation is used to denote some form of association between two variables.

- Example: **Weight** is correlated with **height?**

The correlation may be positive, negative or zero.

- **Positive correlation:** If the value of the attribute *A* increases with the increase in the value of the attribute *B* and vice-versa.
- **Negative correlation:** If the value of the attribute *A* decreases with the increase in the value of the attribute *B* and vice-versa.
- **Zero correlation:** When the values of attribute *A* varies at random with *B* and vice-versa.

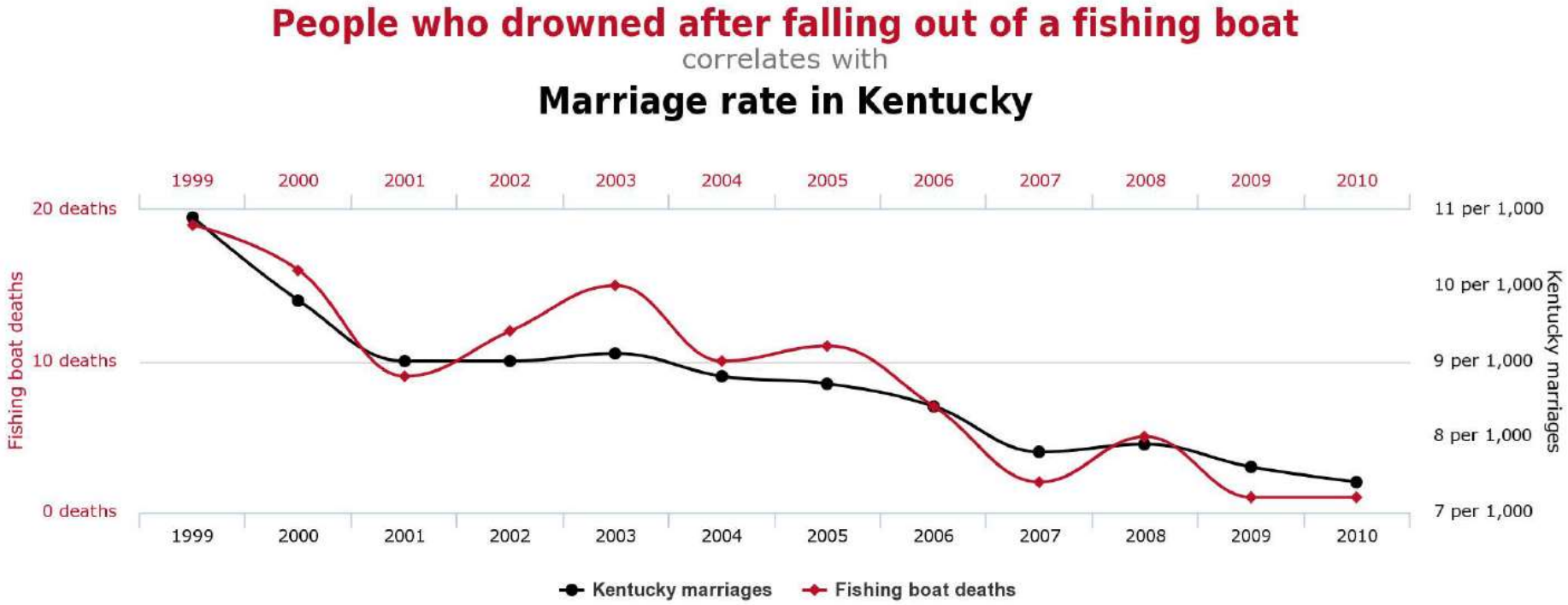
Correlation Coefficient

- **Correlation coefficient is used to measure the degree of association.**
- It is usually denoted by r.
- The value of r lies between +1 and -1.
- Positive values of r indicates positive correlation between two variables, whereas, negative values of r indicate negative correlation.
- $r = +1$ implies perfect positive correlation, and otherwise.
- The value of r nearer to +1 or -1 indicates high degree of correlation between the two variables.
- $r = 0$ implies, there is no correlation

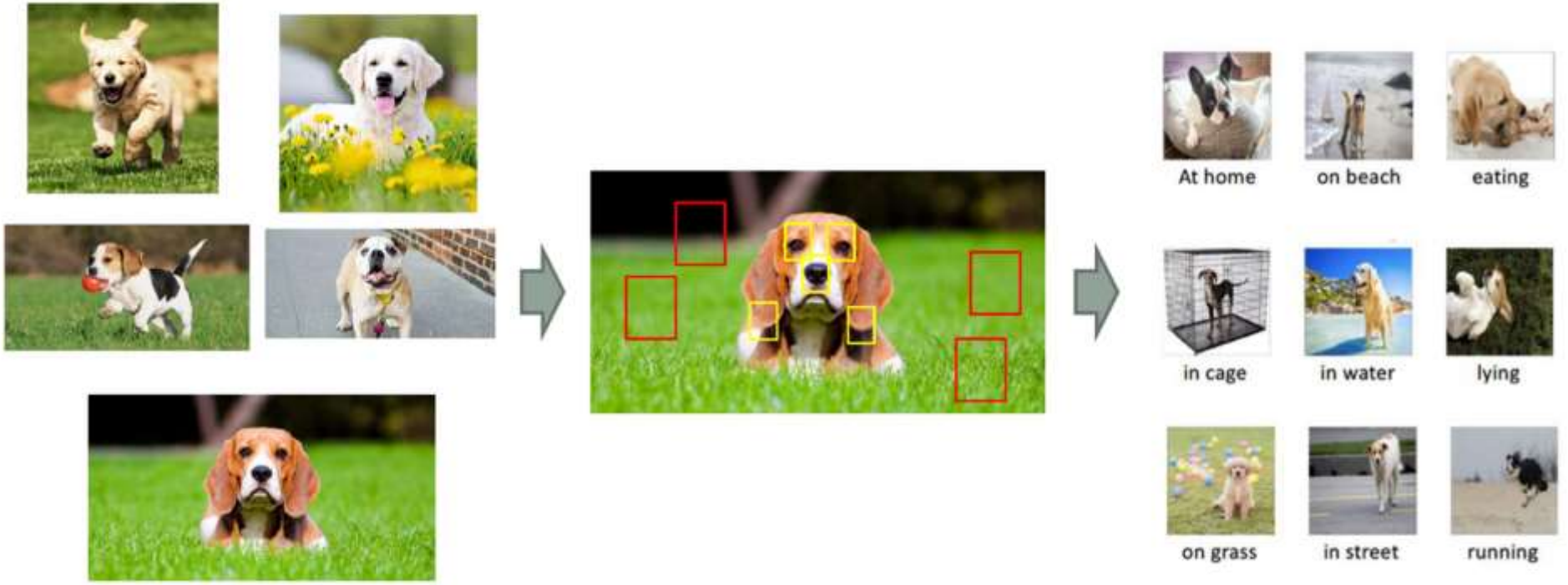
A plausible reason: Correlation

Correlation is the basics of Machine Learning

Correlation does not imply causation



Correlation is unstable

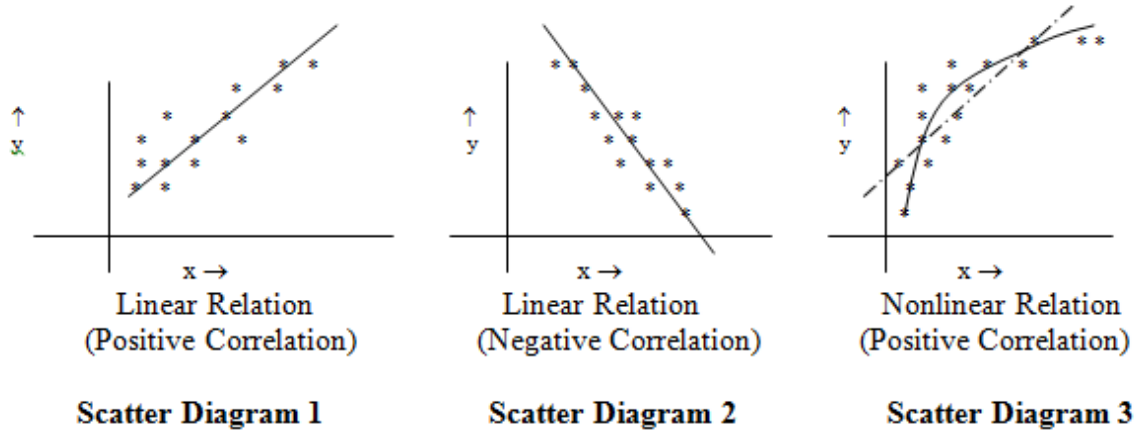


Data

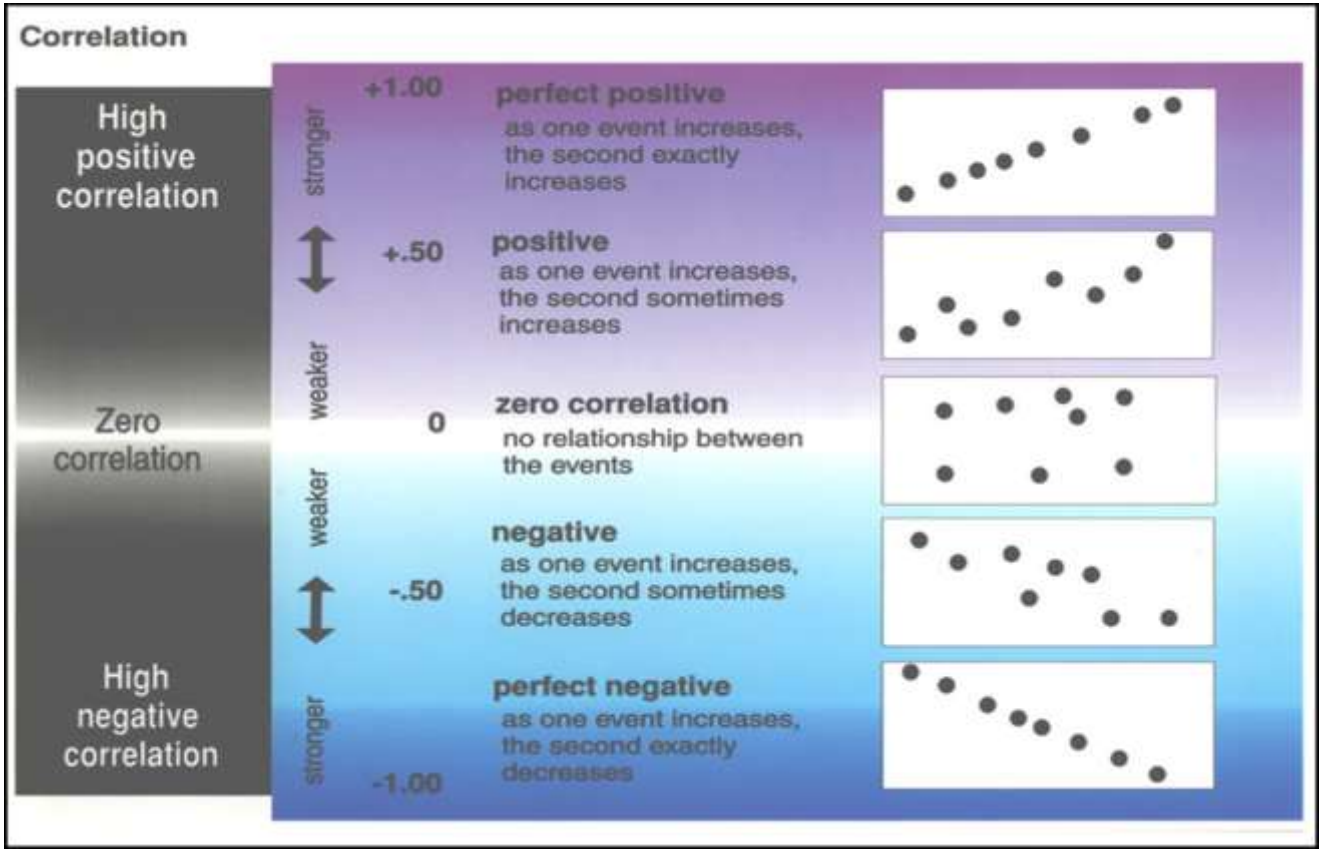
- Let us assume that we have a random sample of size n on two variables namely x and y. The sampled values are $\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$.
- It is necessary to have ordered pairs if we want to find relations between x and y; i.e., each observation consists of two values – one on x and the other on y, i-th observation being (x_i,y_i), i = 1,2, …, n.
- It may sometime be necessary to assume that the sample observations are coming from continuous distribution.

Scatter or dot diagram

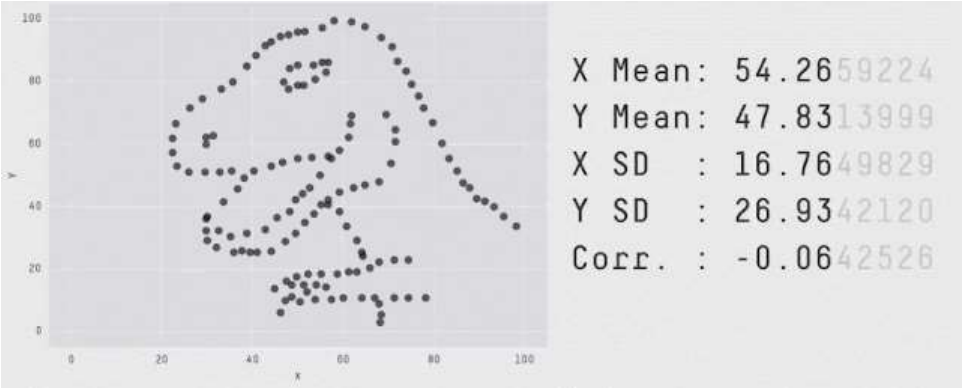
- The first step to get an idea about whether there exists any relation between x and y and if exists, the degree of relation, is to draw **scatter** or **dot diagram**.
- The scatter diagram is nothing but the set of n points of (x,y) pairs shown on a graph paper. The following are some examples of scatter diagrams:



Remarks from Scatter Plots & Correlation Coefficient



What stories can scatter plot tell?

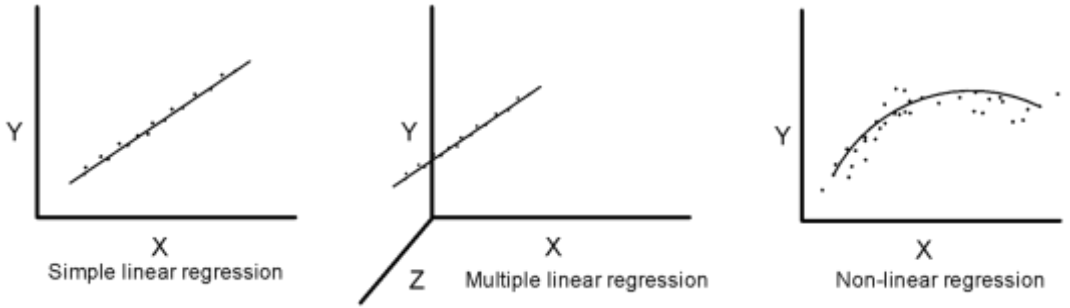


- In 1973, a famous statistician, Francis Anscombe, demonstrated how important it is to visualize the data. The concept got extended later to create Datasaurus Dozen.
- It is a collection of 12 scatterplots with the same means, standard deviations, and correlation coefficient for X and Y (up to 2 decimal places).
- However, the shape of the data is very different from each other. Therefore, the scatterplots tell very different stories about the behavior and interrelationships of X and Y.
- Data available at https://cran.r-project.org/web/packages/datasauRus/vignettes/Datasaurus.html

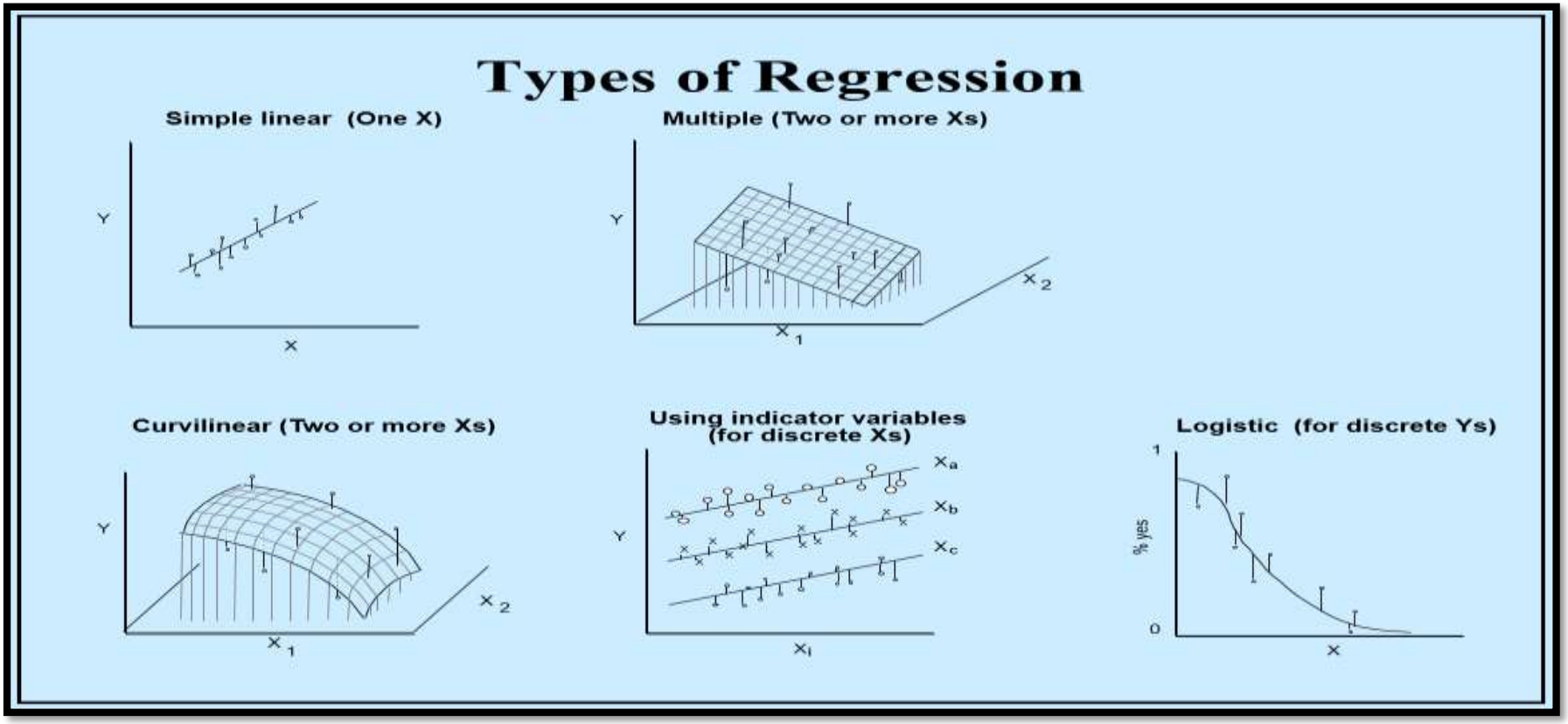
Regression Analysis

- The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

- **Classification of Regression Analysis Models**
  - Linear regression models
    1. Simple linear regression
    2. Multiple linear regression
  - Non-linear regression models



Regression Analysis



Simple Linear Regression

- Regression Analysis is a Statistical tool for investigating the relationship between a dependent variable and one or more independent variables.

- Scatter plots are used to investigate the possible relationship between the variables.

- The simple linear regression model is

$$Y = \alpha + \beta x + \epsilon$$

- For a given x, the corresponding observation Y consists of the value $\alpha + \beta x$ plus an amount ϵ.

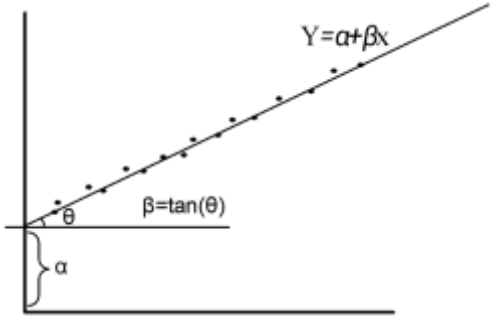
Linear Regression

- The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable.
- That means we want to understand the relationship.
- The line of regression is the line of "best fit" and is obtained by **the principle of least squares**.
- Y - the variables you are predicting
  - i.e. dependent variable
- X - the variables you are using to predict
  - i.e. independent variable
- $\hat{Y}$ - your predictions (also known as Y')

Simple Linear Regression Model

In simple linear regression, we have only two variables:

- Dependent variable (also called Response), usually denoted as Y.
- Independent variable (alternatively called Regressor), usually denoted as x.
- A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$



Note:

- There are infinite number of lines (and hence α_s and β_s)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α and β) quantifying the strength of that relationship.

Multiple Linear Regression

To model output variable y in terms of two or more variables.

General Form:

$$y = a + b_1x_1 + b_2x_2 + - - - + b_kx_k + \varepsilon$$

Two variable case:

$$y = a + b_1x_1 + b_2x_2 + \varepsilon$$

Where

a: intercept (the predicted value of y when all x's are zero)

b_j: slope (the amount change in y for unit change in x_j keeping all other x's constant, j = 1,2,---,k)

Lets try out in Python

Exercise : The effect of temperature and reaction time affects the % yield. The data collected in given in the Mult-Reg_Yield file. Develop a model for % yield in terms of temperature and time?

Step 1: Import packages

```
from google.colab import files
import io
import pandas as mypd
from scipy import stats
import matplotlib.pyplot as myplot
import math as mymath
from pandas.plotting import scatter_matrix
from statsmodels.formula.api import ols
```

Lets try out Multiple Linear Regression in Python

Step 2: Read Data

```
 uploaded = files.upload()
mydata = mypd.read_csv(io.BytesIO(uploaded['Mult_Reg_Yield.csv']))
mydata.head()
time = mydata.Time
temp = mydata.Temperature
output = mydata["Yield"]
```

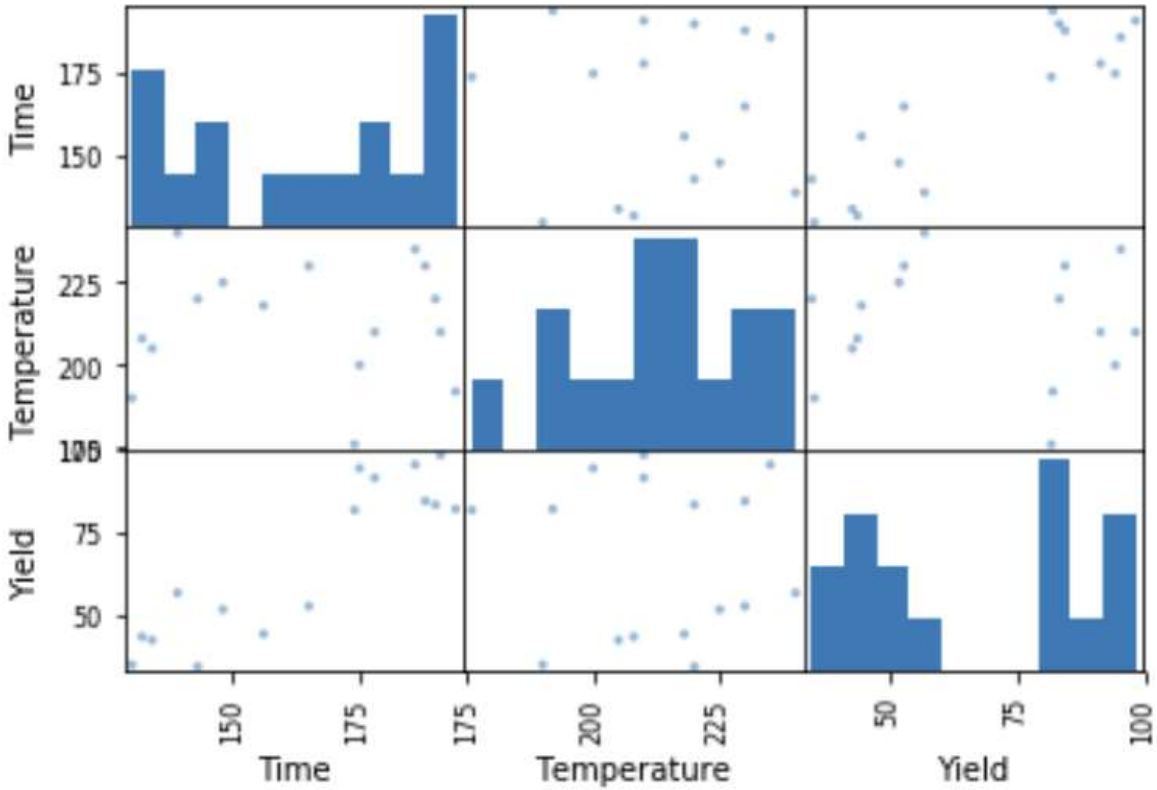
Visualizing the correlation

```
scatter_matrix(mydata)
myplot.show()
```

Correlation between xs & y should be high

Correlation between xs should be low

That’s the Plot



Regression Model

Regression Output

```
mymodel = ols("output ~ time + temp", mydata).fit()
mymodel.summary()
```

Model Yield= 0.9065 x Time - 81.621

Statistics	Value	Criteria
R-squared:	0.806	≥ 0.6
Adj. R-squared:	0.777	≥ 0.6
F-statistic:	27.07	
Prob (F-statistic):	2.32e-05	< 0.05
Log-Likelihood:	-59.703	
AIC:	125.4	
BIC:	127.7	

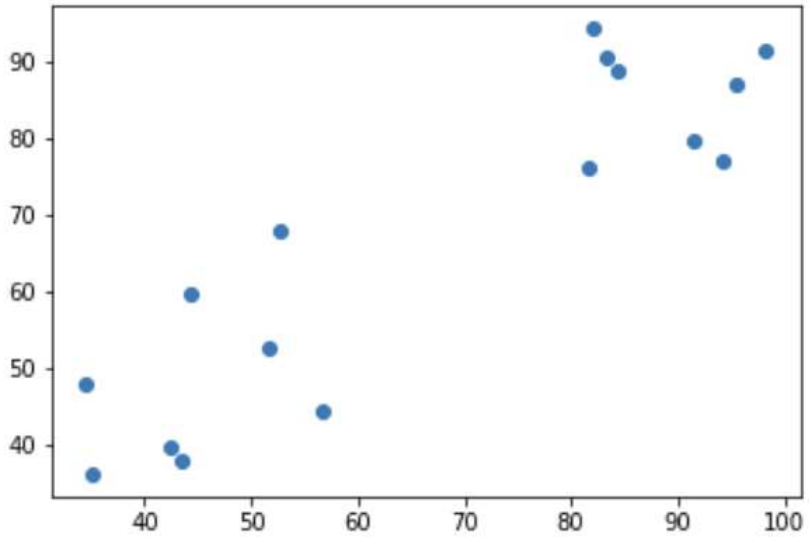
Checking the Residuals

```
pred = mymodel.predict()
res = output - pred
```

SL No	Actual	Predicted	Residuals
1	35	36.22	-1.22
2	81.7	76.10	5.60
3	42.5	39.84	2.66
4	98.3	91.51	6.79
5	52.7	67.94	-15.24
6	82	94.23	-12.23
7	34.5	48.00	-13.50
8	95.4	86.98	8.42
9	56.7	44.38	12.32
10	84.4	88.79	-4.39
11	94.3	77.01	17.29
12	44.3	59.79	-15.49
13	83.3	90.61	-7.31
14	91.4	79.73	11.67
15	43.5	38.03	5.47
16	51.7	52.53	-0.83

Checking the Residuals

```
myplot.scatter(output, pred)
myplot.show()
```



Note: There need to be strong positive correlation between actual and fitted response

Aggregating the Residuals

```
res_sq = res**2
mse = res_sq.mean()
print(mse)
import math as mymath
rmse = mymath.sqrt(mse)
print(rmse)
```

Statistic	Value
MSE	102.005
RMSE	10.099

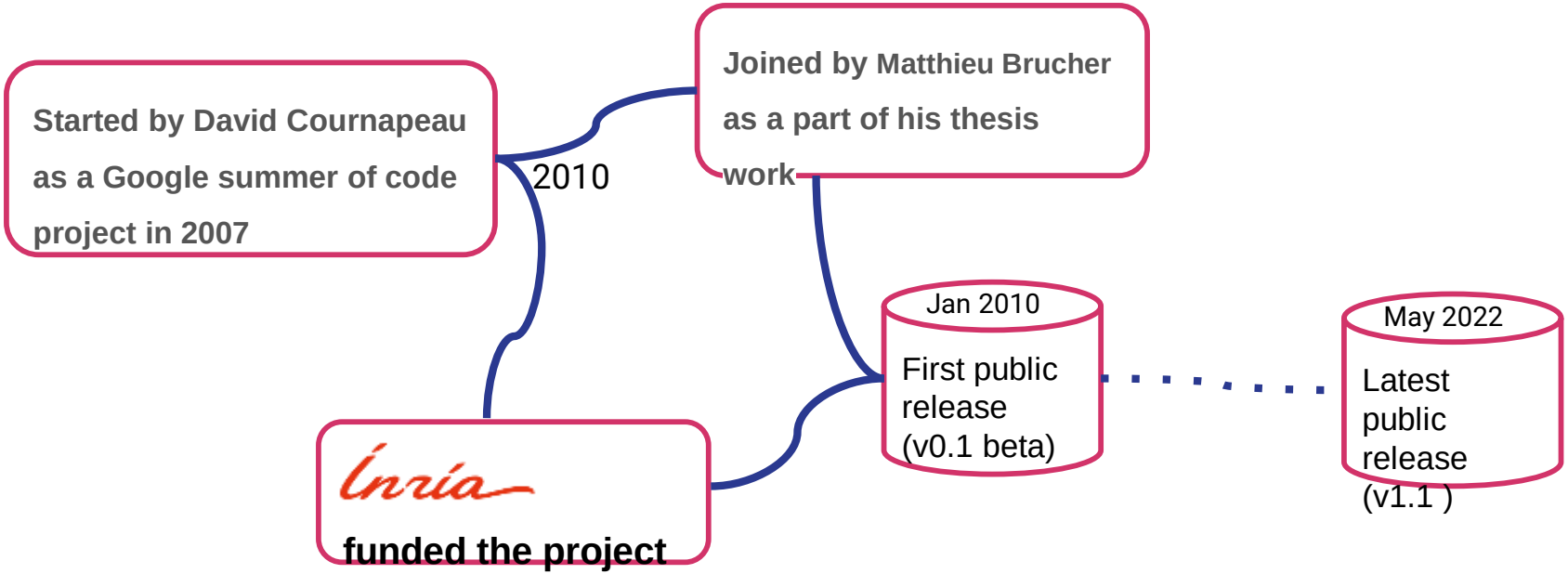
Scikit Learn

Extensions to SciPy (Scientific Python) are called SciKits. Scikit-Learn provides machine learning algorithm.

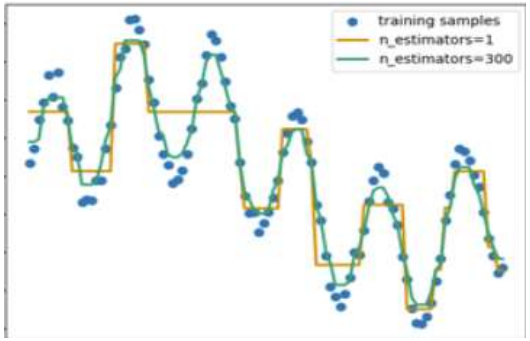
- Algorithms for supervised and unsupervised learning
- Built on SciPy and Numpy
- Standard Python API interface
- Open Source
- Focused package for data modeling

Probably the best general ML framework out there

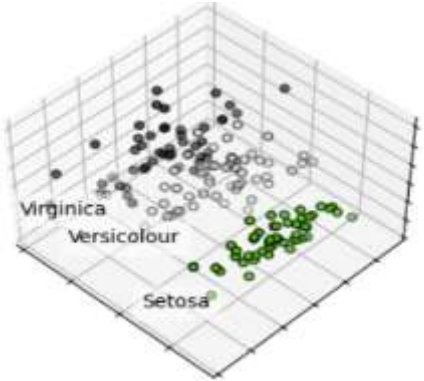
Where did Scikit Learn came from?



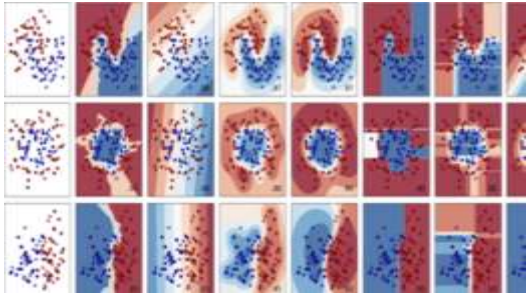
What we can do with Scikit Learn?



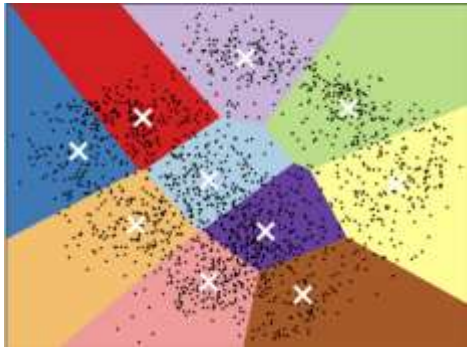
Regression



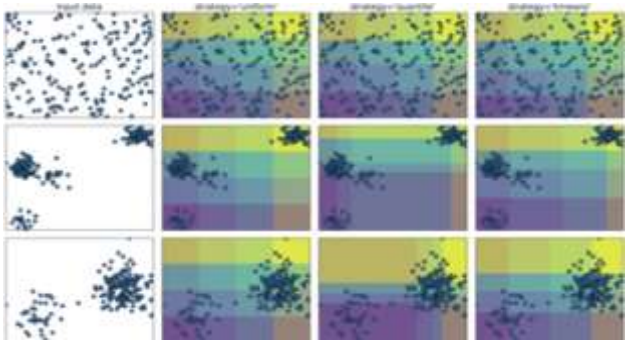
Dimensionality Reduction



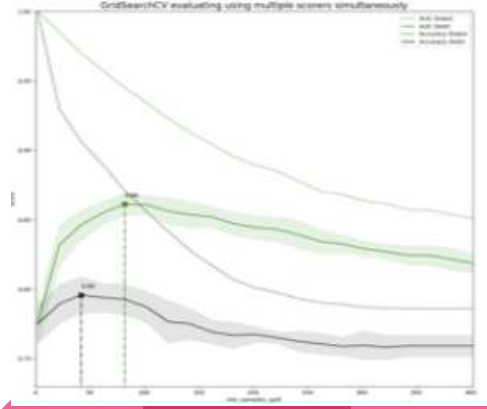
Classification



Clustering



Preprocessing



Model Selection

Let’s check out how to use ML in real world datasets

Thank you !!