Course Title:

Data Science (数据科学)

(Semester: Fall 2021/2022)

Dr. Oluwarotimi W. SAMUEL

Research Center for Neural Engineering
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences

Contact: (Email: samuel@siat.ac.cn & timitex92@gmail.com)

Phone: +86-15814491870

(2021.09.23)

Data Science Process

❑Outline for today's lecture

- ✓Fundamental steps in Data Science
- ✓The relevance of each step
- ✓Responsibilities of a Data Scientist

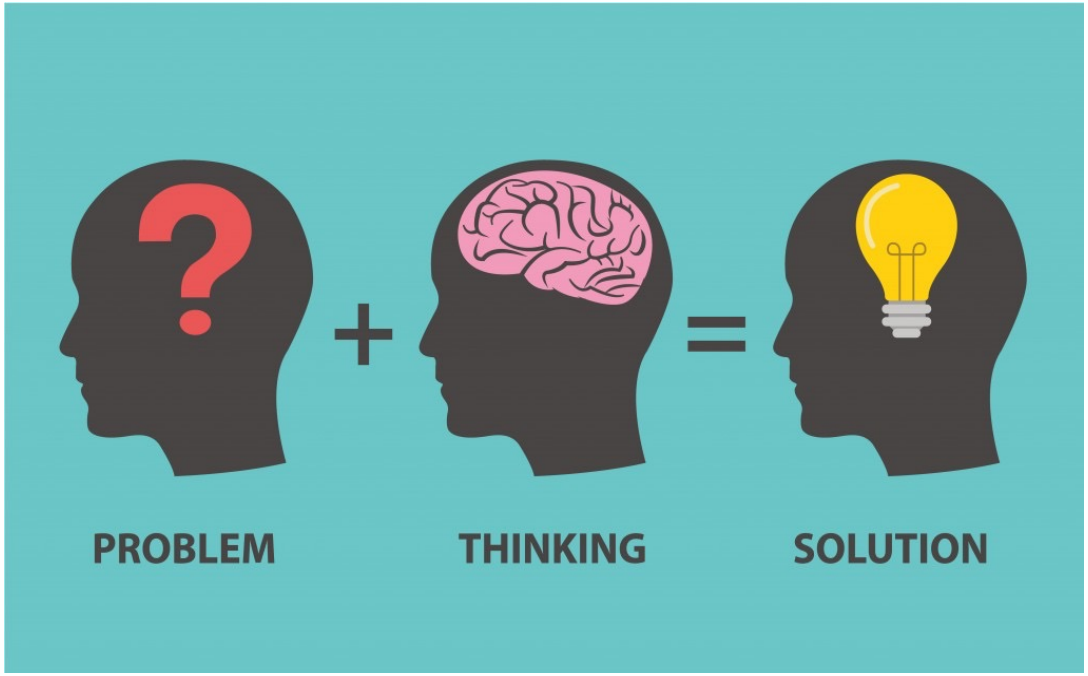
Data Science Process

- **Objective:** This lecture is expected to discuss Data Science process and necessary steps for providing adequate solutions to real world problems.

- **Expectation:** At the end of this lecture, students are expected to understand the important steps that Data Scientists employ in providing Data Science driven solutions.

Data Science Process

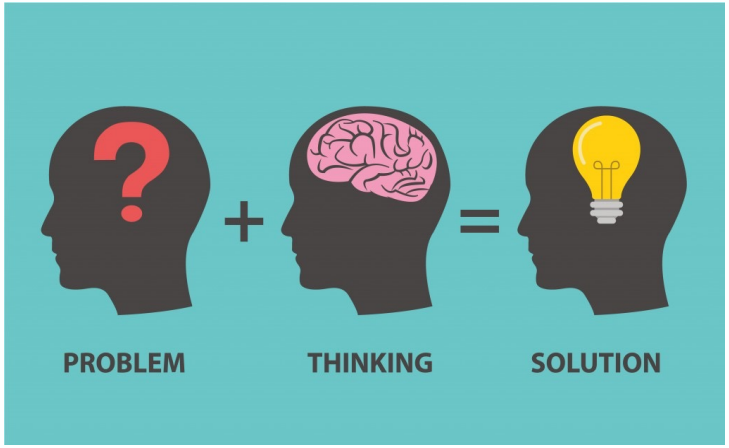
- How does a Data scientist go from analyzing a problem to developing a solution?



Data Science Process

- This leads us to DS process-

"*A sequence of organized activities performed by a Data scientist*" as follows:



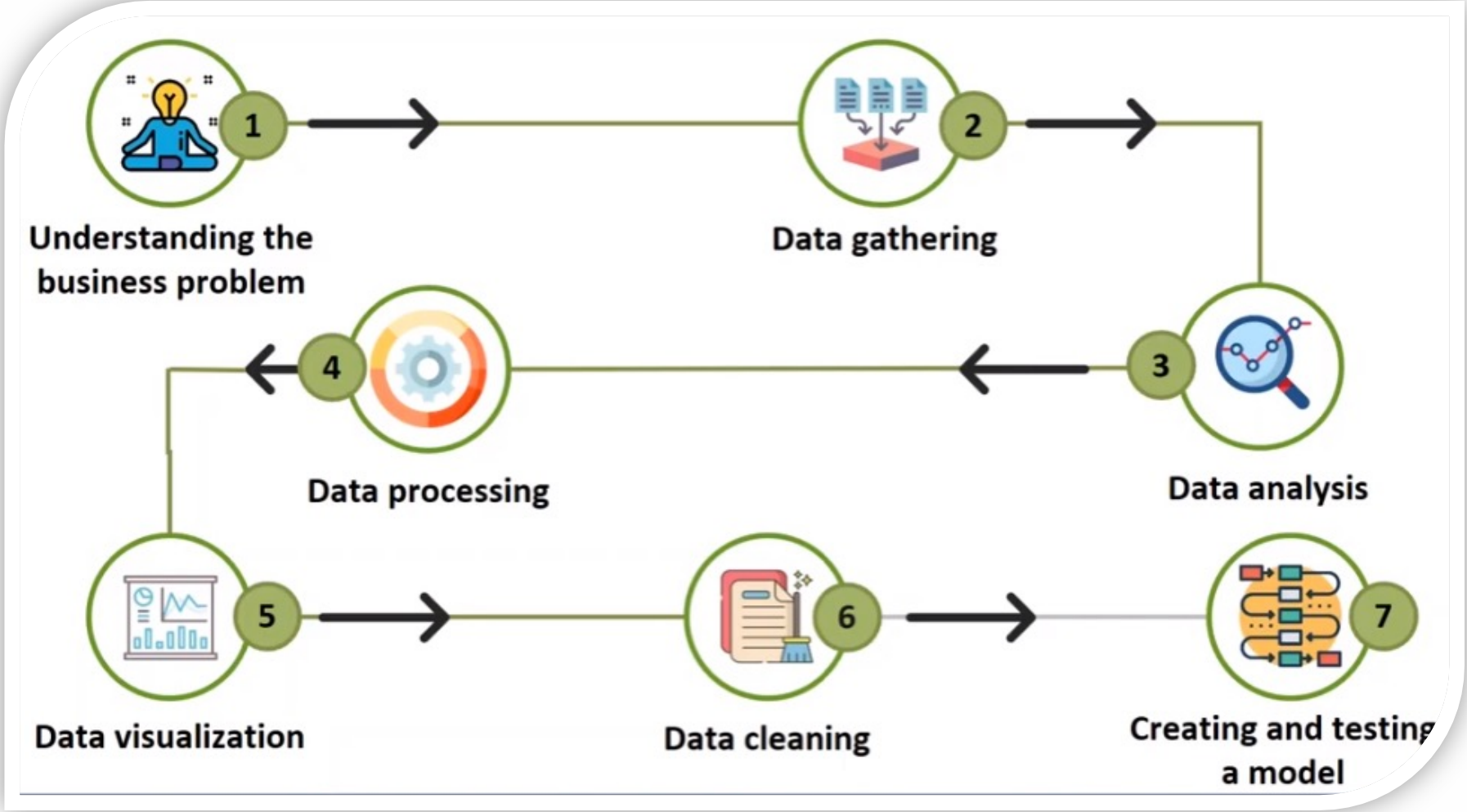
Data Science Process

❑Steps for building a DS driven solution:

- ✓Understanding the business problem
- ✓Data gathering
- ✓Data processing
- ✓Data analysis
- ✓Data visualization
- ✓Data cleaning
- ✓Creating and testing a model

Data Science Process

Pictorial Representation



Data Science Process

❑ Understanding the Business Problem:

- ✓ The first step in a DS solution is *understanding the problem*
- ✓ Proper understanding of the problem helps a Data scientist to develop adequate solution
- ✓ Let us look at some pointers on how to understand the problem

Data Science Process

❑ Understanding the Business Problem:

1. First, ask the WHY questions about the problem at hand
2. Understand the end product required
3. Determine the data sources for this problem
4. Gather all information and context required to solve the problem

NOTE: Once the problem is well understood, appropriate data and information are gathered.

Data Science Process

❑ Data Gathering:

- ✓ Data gathering is the process of retrieving data from various sources to be used in our DS process
- ✓ For example, data can be gathered from multiple sources including:
  - CVS files
  - the Internet through Web API
  - SQL and non-SQL databases

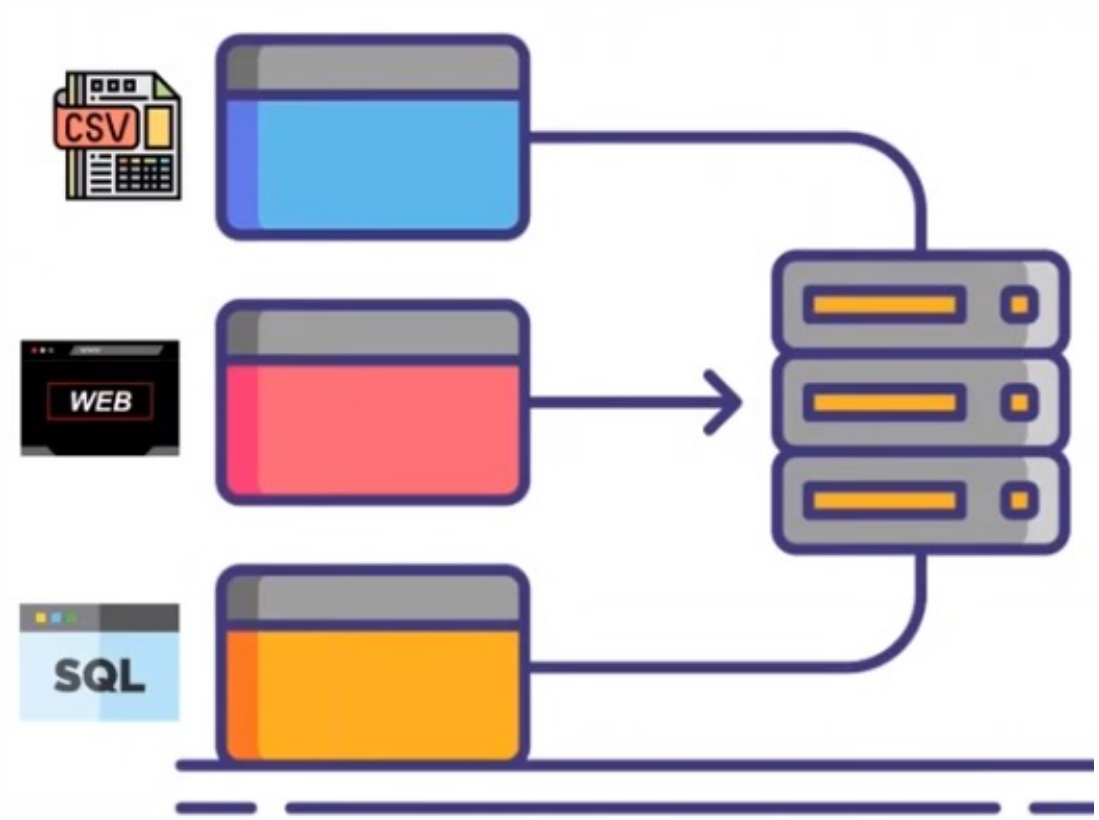
Data Science Process

❑ Multiple Data Sources:

CVS files

Web APIs

SQL DBs



Data Science Process

❑ Data Processing:

✓ Data processing is the process of converting the gathered data into:

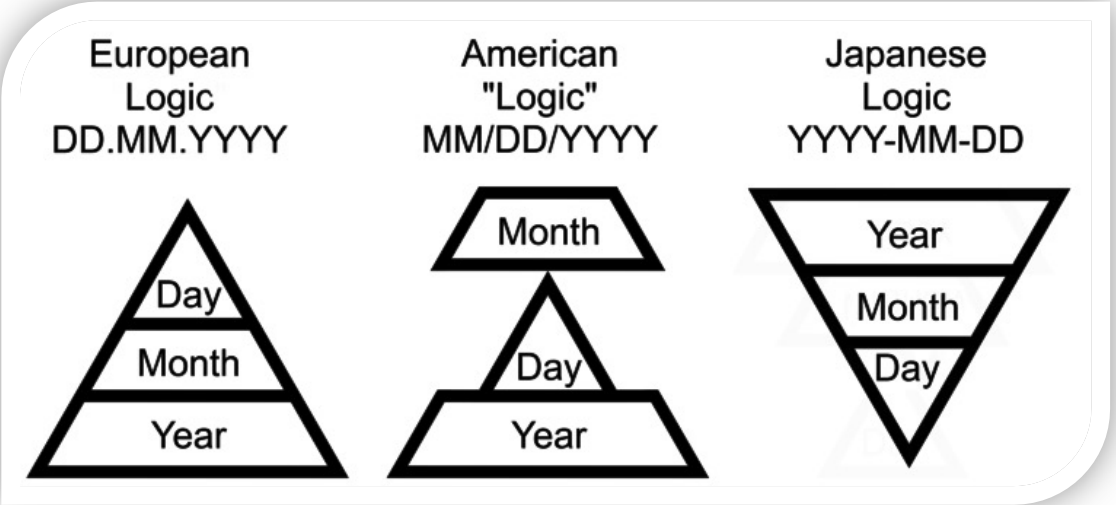
- easily readable formats, or
- easily process-able formats

Data Science Process

❑ Data Processing:

✓ For example, when the datasets gathered from multiple sources, contain dates of different formats:

- *DS course registration date*
- *Data of birth*
- *Date of purchase*



Data Science Process

❑ Data Analysis:

- ✓ Data analysis is the process of analyzing data sets to summarize their main characteristics.

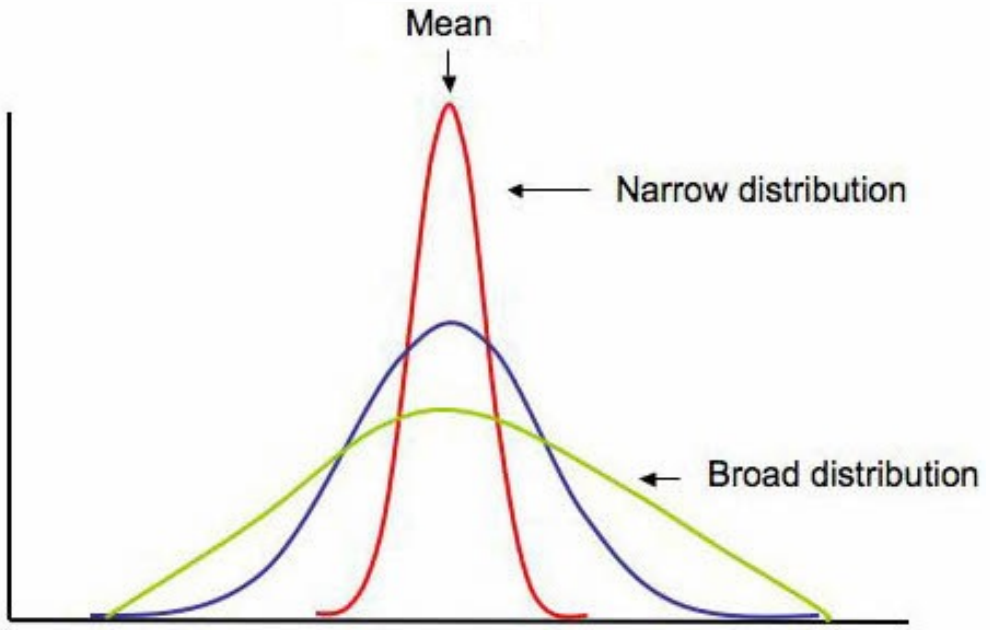
- ✓ Such characteristics allow us to understand certain phenomenon about the data sets.

Data Science Process

❑ Data Analysis:

The characteristics includes:

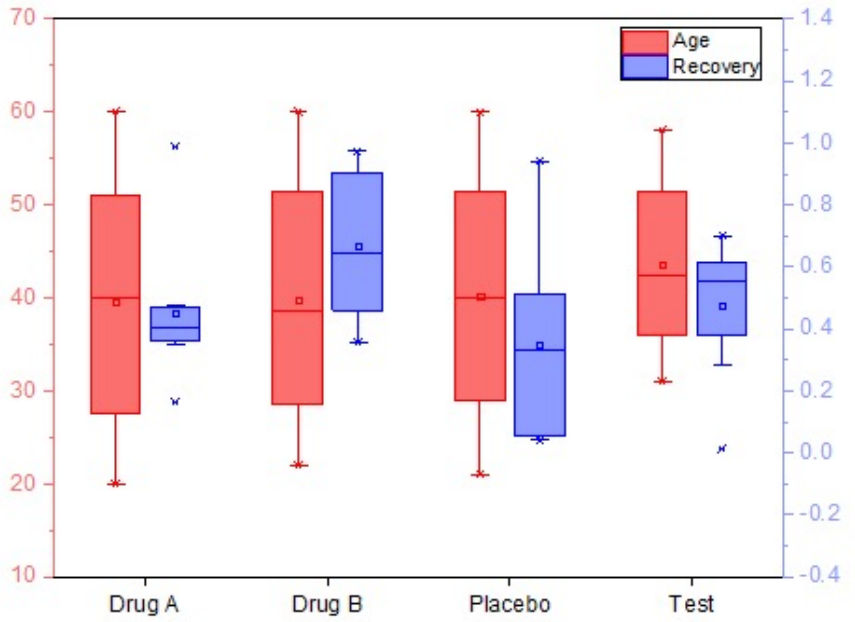
- distribution of the datasets
- mean of the value



❑ Data Analysis:

The characteristics includes:

- most occurring & least occurring values
- minimum and maximum values
- standard deviation and variances of the data



Data Science Process

❑ Data Visualization:

- ✓ Data visualization is the graphical/pictorial representation of the information and data that we have.
- ✓ This allows us to observe trends/patterns in our data set.
- ✓ Taking a look at a large set of numbers in tables may not be that helpful in aiding understanding of inherent trends.

Data Science Process

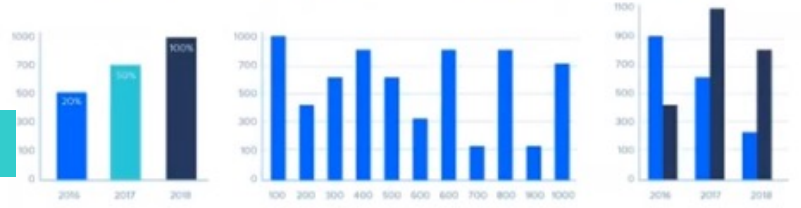
❑ Data Visualization:

✓ Hence, the need for data visualization using graphs and charts, plots.

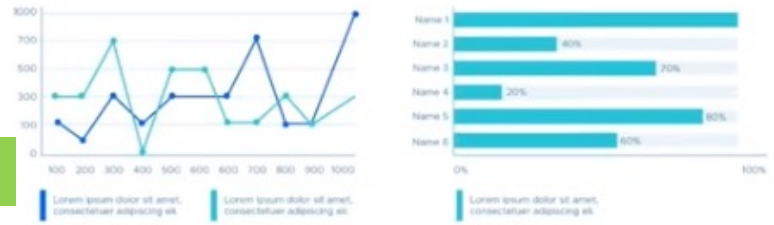
	S.No	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	1	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.460000	0	1	4	4
1	2	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.020000	0	1	4	4
2	3	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.610000	1	1	4	1
3	4	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.440000	1	0	3	1
4	5	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.020000	0	0	3	2
5	6	Valiant	18.1	6	225.0	105	2.76	3.460	17.674828	1	0	3	1
6	7	Duster 360	14.3	8	360.0	245	3.21	3.570	15.840000	0	0	3	4
7	8	Merc 240D	24.4	4	146.7	62	3.69	3.190	20.000000	1	0	4	2
8	9	Merc 230	22.8	4	140.8	95	3.92	3.150	22.900000	1	0	4	2
9	10	Merc 280	19.2	6	167.6	123	3.92	3.440	18.300000	1	0	4	4
10	11	Merc 280C	17.8	6	167.6	123	3.92	3.440	18.900000	1	0	4	4
11	12	Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.400000	0	0	3	3
12	13	Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.600000	0	0	3	3
13	14	Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.000000	0	0	3	3
14	15	Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.980000	0	0	3	4
15	16	Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.820000	0	0	3	4
16	17	Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.420000	0	0	3	4
17	18	Fiat 128	32.4	4	78.7	66	4.08	2.200	17.674828	1	1	4	1
18	19	Honda Civic	30.4	4	75.7	52	4.93	1.615	18.520000	1	1	4	2
19	20	Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.900000	1	1	4	1
20	21	Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.010000	1	0	3	1
21	22	Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.870000	0	0	3	2
22	23	AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.300000	0	0	3	2
23	24	Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.410000	0	0	3	4
24	25	Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.050000	0	0	3	2

Pie chart

Bar plots



Line plots



✓ Facilitates our understanding of such hidden information

Data Science Process

❑ Data Visualization:

	S.No	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	1	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.460000	0	1	4	4
1	2	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.020000	0	1	4	4
2	3	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.610000	1	1	4	1
3	4	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.440000	1	0	3	1
4	5	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.020000	0	0	3	2
5	6	Valiant	18.1	6	225.0	105	2.76	3.460	17.674828	1	0	3	1
6	7	Duster 360	14.3	8	360.0	245	3.21	3.570	15.840000	0	0	3	4
7	8	Merc 240D	24.4	4	146.7	62	3.69	3.190	20.000000	1	0	4	2
8	9	Merc 230	22.8	4	140.8	95	3.92	3.150	22.900000	1	0	4	2
9	10	Merc 280	19.2	6	167.6	123	3.92	3.440	18.300000	1	0	4	4
10	11	Merc 280C	17.8	6	167.6	123	3.92	3.440	18.900000	1	0	4	4
11	12	Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.400000	0	0	3	3
12	13	Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.600000	0	0	3	3
13	14	Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.000000	0	0	3	3
14	15	Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.980000	0	0	3	4
15	16	Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.820000	0	0	3	4
16	17	Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.420000	0	0	3	4
17	18	Fiat 128	32.4	4	78.7	66	4.08	2.200	17.674828	1	1	4	1
18	19	Honda Civic	30.4	4	75.7	52	4.93	1.615	18.520000	1	1	4	2
19	20	Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.900000	1	1	4	1
20	21	Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.010000	1	0	3	1
21	22	Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.870000	0	0	3	2
22	23	AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.300000	0	0	3	2
23	24	Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.410000	0	0	3	4
24	25	Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.050000	0	0	3	2

Data Science Process

❑ Why should we Visualization:

1. To view changes over time easily via a visual aid rather that plain data
2. To discover correlations among two or more variables seamlessly
3. To simplify complex information into user-friendly formats
4. To tell a better story with a bunch of pictures over time

Data Science Process

❑ Data Cleaning/Cleansing:

Data cleaning is the process of removing unwanted or inaccurate records from a table or a dataset.

NOTE: Analysis made on clean data yield more accuracy.

Data Science Process

❑ Data Cleaning/Cleansing:

✓ Take a closer look at the data sets to understand how to clean it particularly when such data contain:

- incomplete records or missing values records would make it difficult to analyze

Data Science Process

❑ Data Cleaning/Cleansing:

- missing values could be replaced with zeroes, ones, or with the mean of the entire dataset

- data that contain null values maybe dropped/deleted

✓ Different kind of ***Normalization Technique*** maybe applied during the data cleaning process.

❑ Creating a Model:

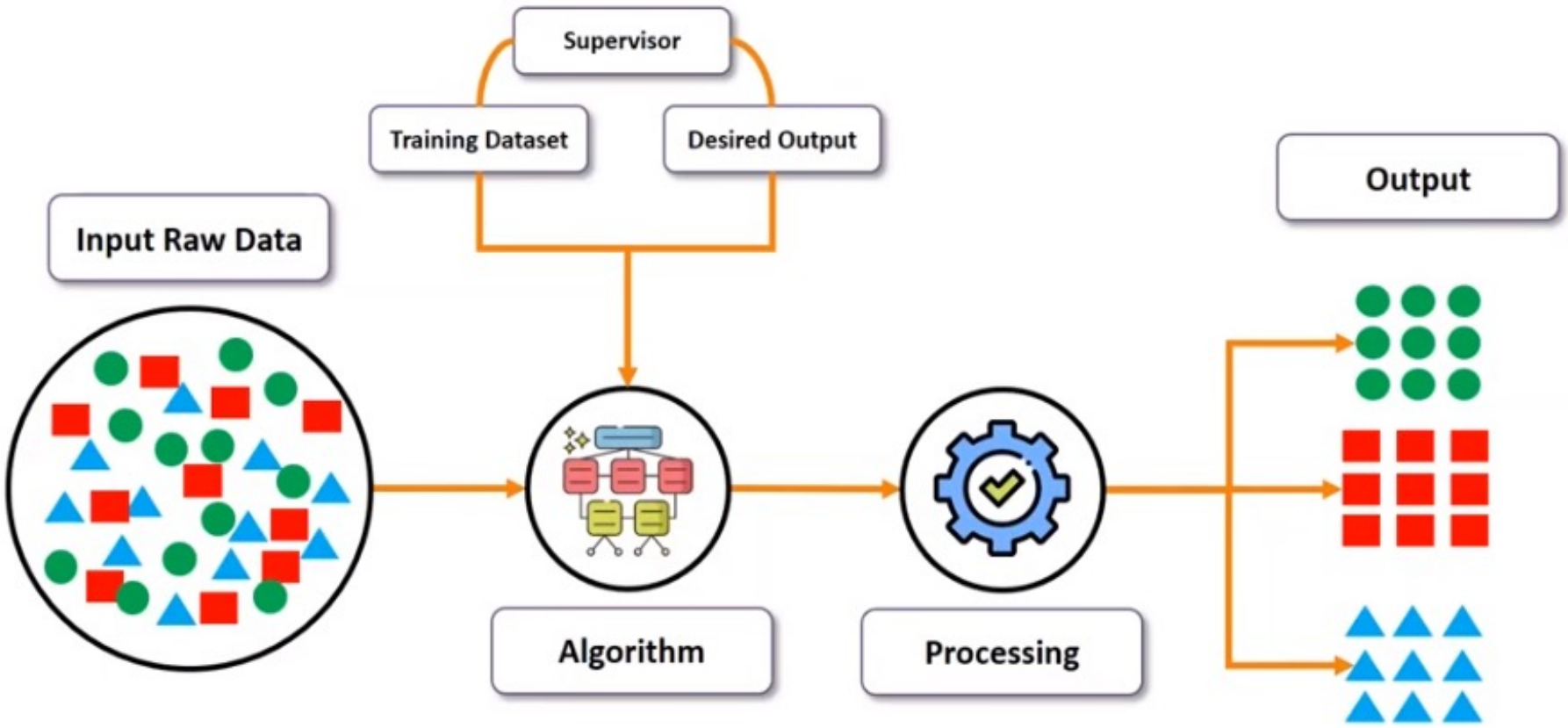
- ✓ A model is a mathematical construct that finds patterns when we feed raw data into it.
- ✓ Models are developed for specific use cases. In order words, a model built for a specific problem mayn't provide adequate solution for other problems.
- ✓ The model enables us to classify the input or make prediction about future trends of things etc.

Data Science Process

❑ Creating a Model:

✓Model building process:

- input, algorithm, processing, output



Data Science Process

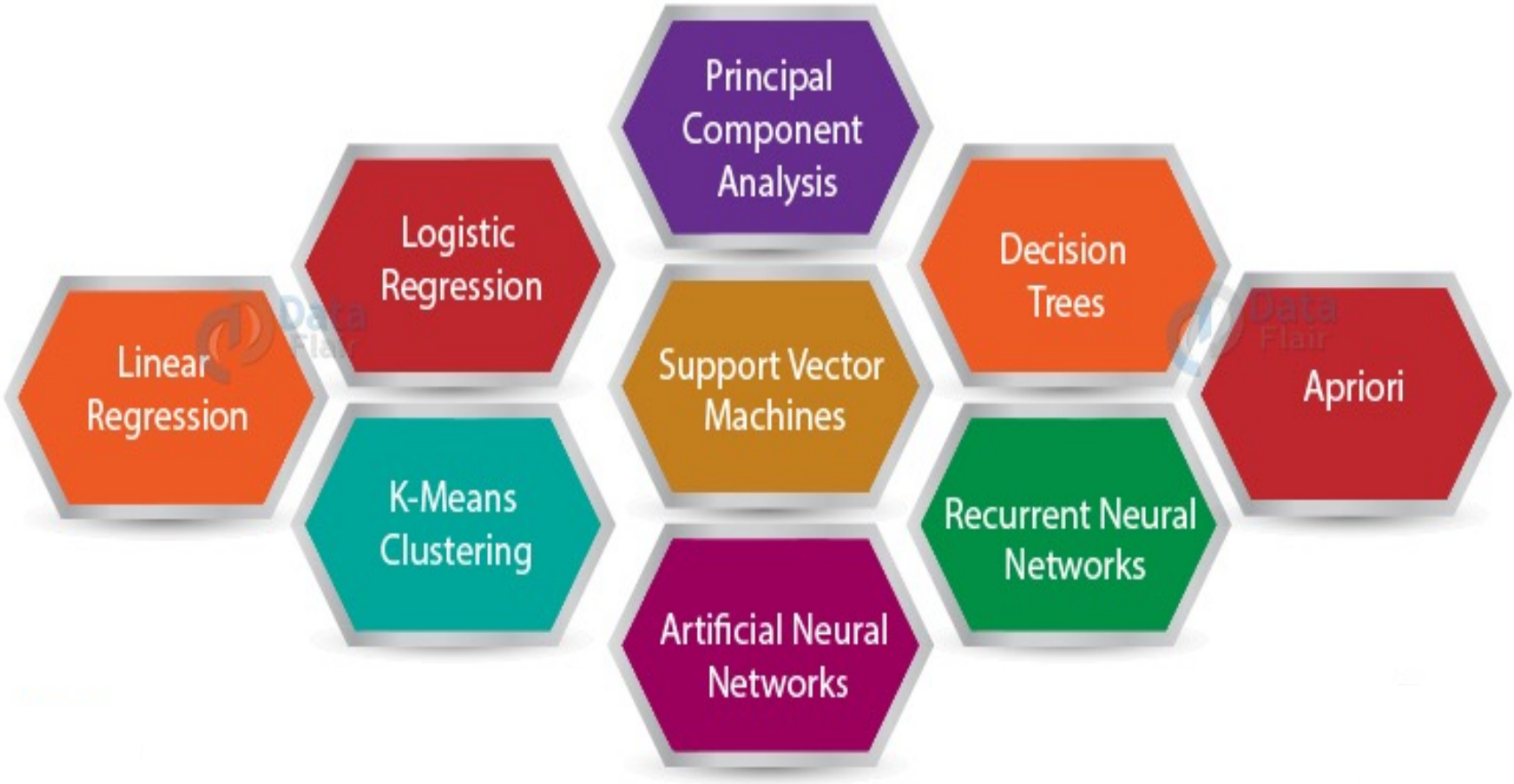
- Creating a Model:
  - ✓ We have to choose an appropriate algorithm and start training it with a subset of the dataset that we have.

  - ✓ We have to segregate between training and testing data, with majority of the dataset is used for training the model.

Data Science Process

- Creating a Model:

  Choosing an appropriate model for our solution:



Data Science Process

❑ Testing the Model:

- ✓ Testing a model involves evaluating its performance and accuracy in the presence of new datasets/test sets and comparing its accuracy with the existing model.
- ✓ Once we create multiple models driven by different algorithms, we have to test their performance.
- ✓ This will enable us obtain the best model that would be deployed for practical applications.

Data Science Process

❑ Testing the Model:

✓ This can be done by checking how much (%) of the testing data that was correctly predicted.

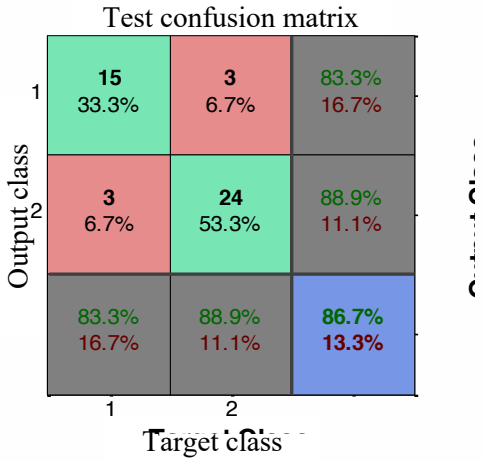
Data Science Process

❑ Testing the Model:

✓ Several key terminologies come into mind when we are trying to test a model as follows:

Confusion Matrix

A confusion matrix is a table layout that allows us to visualize the performance of an algorithm



Accuracy Score

Accuracy score is equal to the percentage of rows in the testing data that are correctly classified

		Predicted/Classified	
		Negative	Positive
Actual	Negative	998	0
	Positive	1	1

❑ Data Scientist Responsibilities:

✓ All the methods of working with asynchronous code handle errors differently.

1. Analytics
2. Strategy
3. Collaboration

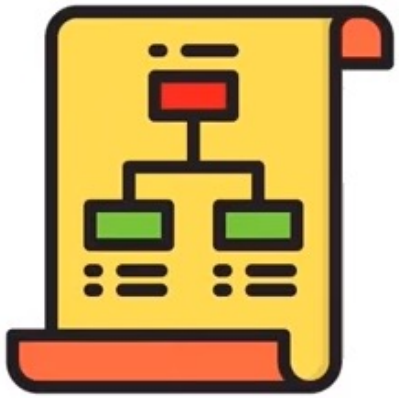
To be able to analyze data and extract useful information out of it

❑ Data Scientist Responsibilities:

✓ All the methods of working with asynchronous code handle errors differently.

1. Analytics
2. Strategy
3. Collaboration

To be able to come up with ways to make the best use of newly obtained information



❑ Data Scientist Responsibilities:

✓ All the methods of working with asynchronous code handle errors differently.

1. Analytics
2. Strategy
3. Collaboration

To be able to collaborate with other people and solve complicated problems

❖ Summary for today's lecture

- ✓ We have learned about the fundamental procedures that a Data Scientist needs to follow to be able to provide adequate solution to a specific problem.

- ✓ Also, we were able to highlight some of the key responsibilities of a Data Scientist.

Questions and Comments!