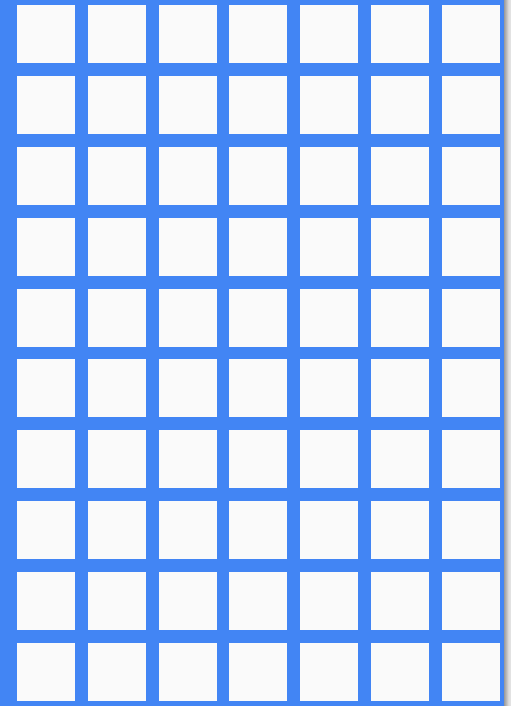
Certification Training Courses for Graduates and Professionals

www.signitivetech.com

Data Science



Data Science Concepts

Business Analytics:

- A scientific process that performs data into insight
- Used for fact-based or data-driven decision making
- Uses tools such as reports & graphs to visualize, analysis & finding optimized solution for the problem

Types of Business Analytics:

- Descriptive Analytics: Includes techniques that explain what has happened in the past
- Predictive Analytics: Includes techniques that use data models created from past data to predict the future or determine the impact of one variable on another
- Prescriptive Analytics: Specifies the best course of action for a business activity in the form of a prescriptive model. The models used in this analytics are called optimization models.

Business Analytics

Descriptive Analytics:

- Companies like MakeMyTrip collect data which in terms of passengers' data to analysis on various factors or properties of passengers like age range, gender, travelling destinations etc to check their business trends, ups-downs, what services are good or bad, customer's feedback etc.

Predictive Analytics:

- Companies plans their goals and targets for the future on the basis of past data to sustain in the competitive age.

Prescriptive Analytics:

- Using both past data and future predictions, companies make their strategies and build a business model including customer satisfaction, promoting best deals and etc to grow the revenues considering optimized solutions like cost-cutting.

Business Analytics

Other types of business analytics:

- ➔ Supply Chain Analytics (Flipkart Logistics, Delhivery)
- ➔ Healthcare Analytics (Hospitals, Life Insurance Companies)
- ➔ Marketing Analytics (Advertising Agencies)
- ➔ Human Resource Analytics (Employee Management)
- ➔ Web Analytics (Facebook, Google)

Data Science

Need of decision making:

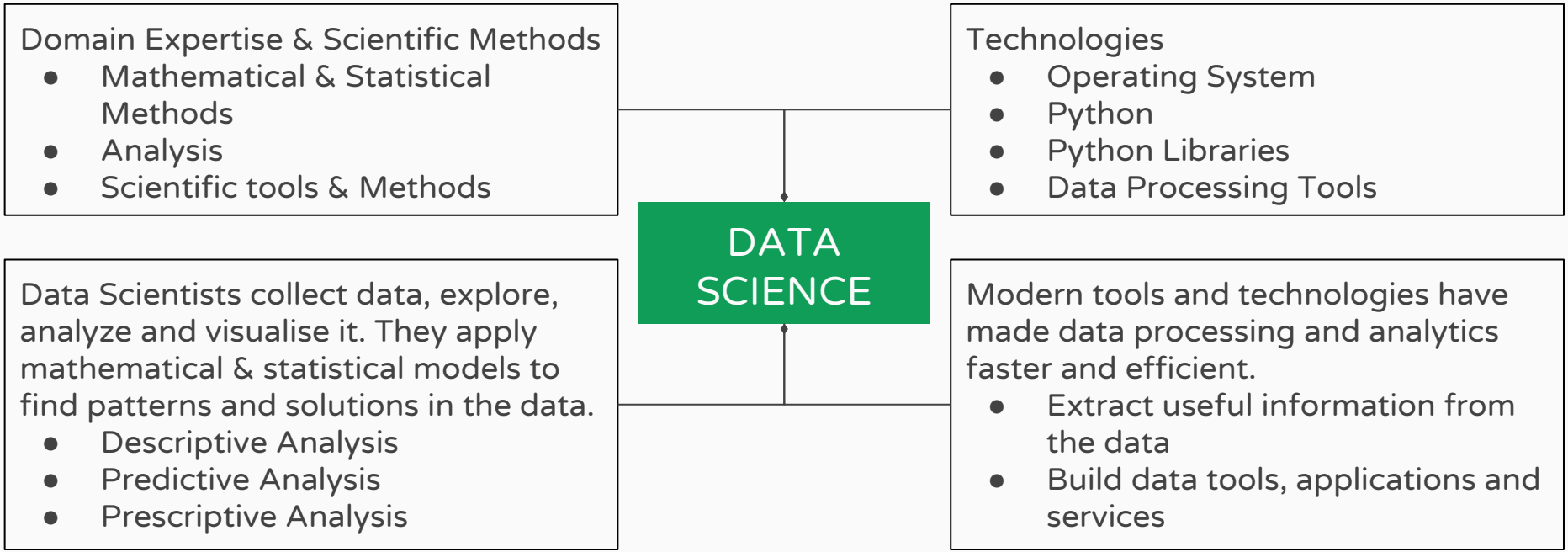
- Low costs
- High revenue
- Reduces expenses
- Growing profit share
- Increase stakeholders

Data Science:

- Includes processes, principles, and methods to understand phenomena through automated data analysis.
- Allows Data-Driven Decision Making (DDD) which determines the productivity of an organisation.

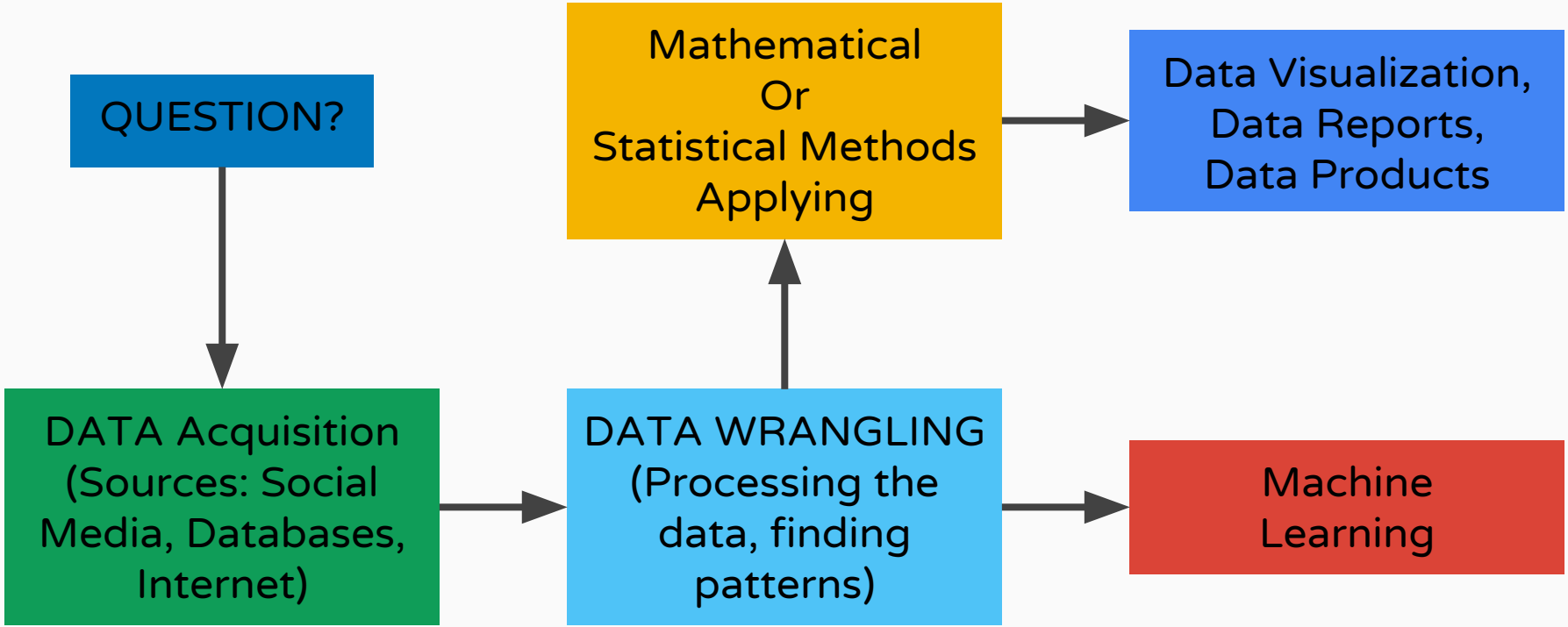
Data Science

Components of Data Science:

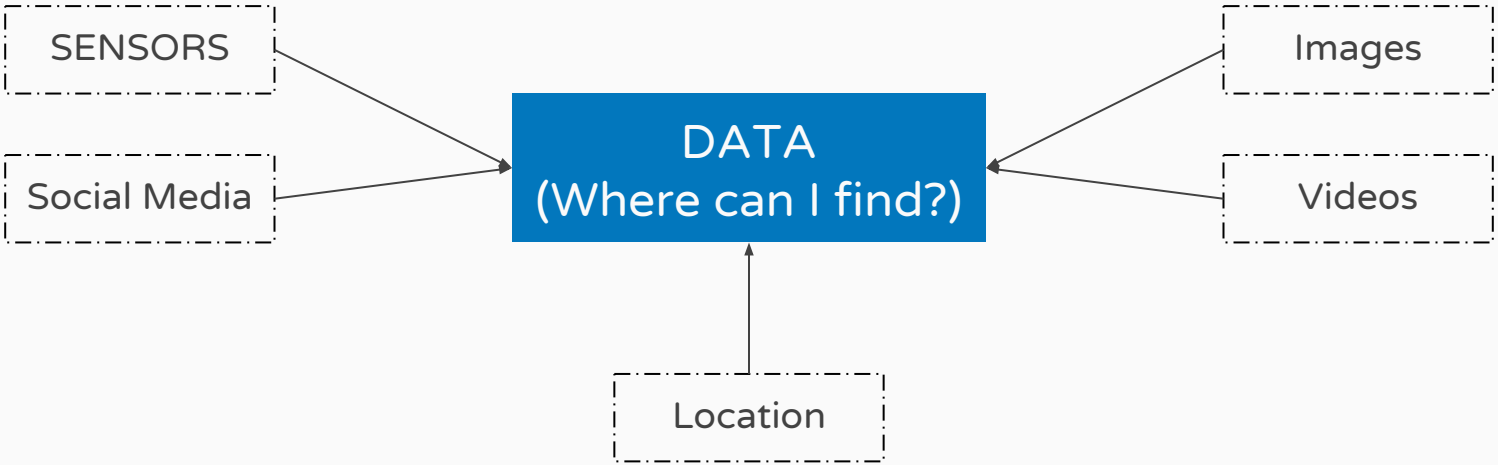


Data Science

Data Science Process:



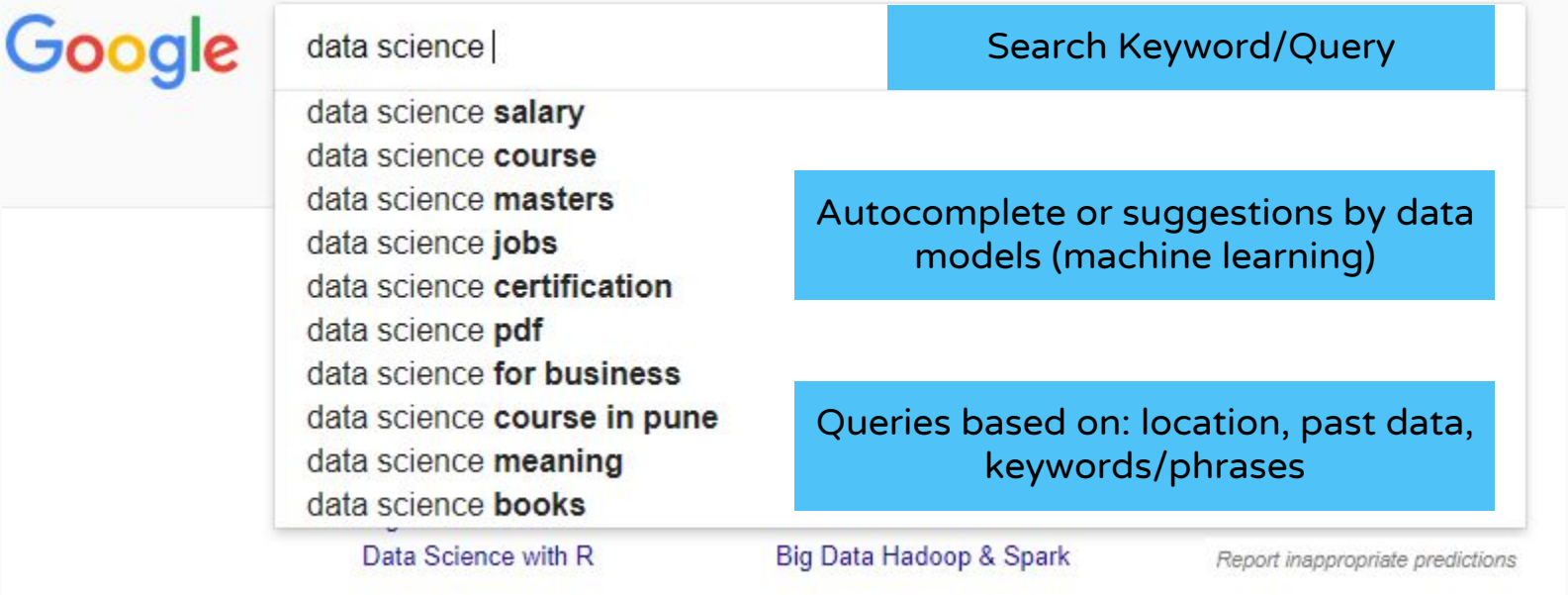
Sources of Data:



3 V's of Data (Big Data):

- Volume: Enormous amount of data generated from various sources.
- Velocity: Large amount of data streaming in at great speeds, which require quick data processing
- Variety: Different formats of data: Structured, Semi-Structured and Unstructured.
  - Structured: Data in the form of tables, database tables, excel sheets
  - Unstructured: Data from social medias including videos, audios, images streaming contents

Using Data Science:



Data Science

Data Analytics and Python:

- ➔ Data Acquisition: PYTHON Language
- ➔ Data Wrangling: pandas, numpy, scipy
- ➔ Data Explore: matplotlib, seaborn
- ➔ Model: scikit-learn (machine learning)
- ➔ Data Visualize: bokeh, seaborn

Data Analytics Flow or Lifecycle:

1. Business Problem
2. Data Acquisition
3. Data Wrangling
4. Exploratory Data Analysis (EDA Techniques)
5. Data Exploration
6. Conclusion or Prediction
7. Communication

Data Analytics Flow or Lifecycle:

★ **Business Problem:**

- The process of analytics begin with questions or business problems.
- Business problems trigger the need to analyze data and find answers.
- This process includes Problems regarding customers, sales, inventory, traffic volume, costs and expenses, revenue generation, profit-loss statements, feedbacks, employees' satisfactions.

Data Analytics Flow or Lifecycle:

★ **Data Acquisition:**

- In this stage we collect data from various sources regarding business questions asked in the step 1 to analyze the data.
- Data can be collected through social media and other sources. Also file handling and file formats conversion are done in this stage.

Data Analytics Flow or Lifecycle:

★ Data Wrangling and Exploration:

Data Wrangling:

- ❖ Data Cleansing: We remove empty or duplicate records, delete unauthorized or illegal data in this stage.
- ❖ Data manipulation: We structure the data in the well readable format, index the data and we do data validation in this stage.

Data Exploration:

- ❖ Data discovery: We discover data in the various formats like graphs, reports, excel sheets, images etc. using different tools or libraries.
- ❖ Data patterns: We find similarities between records or trends in data.

Data Analytics Flow or Lifecycle:

★ Exploratory Data Analysis (EDA):

- Data Cleansing: This is the “Approach” towards best fit the data.
- Data Manipulation: This gives “Focus” on the structure of data.
- Data Discovery: This provides “Assumptions” made on the data.
- Data Patterns: Using “EDA Techniques” it provides numeric and graphical outputs. EDA Techniques can be Quantitative or Graphical.

Data Analytics Flow or Lifecycle:

★ Quantitative EDA Techniques:

Measurement of central tendency:

- Mean: Mean indicates how centralized the data points are.
- Median: Median is the exact middle value.
- Mode: Mode is the most common value in the data i.e frequency.

Measurement of spread:

- Variance and Standard Deviation: It indicates by how much the data point is deviated by Mean.
- Interquartile Range: Distance between 75th and 25th percentile.

Data Analytics Flow or Lifecycle:

★ Graphical EDA Techniques:

Histograms: It shows:

- Mean, median, mode in the data
- Spread of the data
- Presence of multiple modes in the data

Scatter Plot: It represents relationship between two variables. For example if we plot the scatter plot between x and y, then it indicates: are variables x and y related? Does the change in y depends on x and vice versa?

Data Analytics Flow or Lifecycle:

★ **Conclusion and Prediction:**

❖ After doing all stuffs, we reach to a conclusion and/or prediction based on the data analysis after heavily use of statistical and mathematical functions, forecasting, machine learning processes.

Data Analytics Flow or Lifecycle:

★ **Communication:**

- After making conclusion we launch a product, or depict the quarter or annual results of companies, or give more derivatives to the stakeholders.
- All these communications can be happened by launching an advertisement or taking press conference or using social media platforms.

Data Visualization: These techniques are used for effective communication of data.

Benefits of Data Visualization:

- Simplifies quantitative information through visuals
- Shows the relationship between data points and variables
- Identifies patterns
- Establishes trends
- Data Visualization tells a story about the data using different colors, shapes, sizes

Data Science

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

Using Statistics we can:

- Analyze the primary data
- Build a Statistical Model
- Predict the future outcome

➔ Statistical Analysis: It is the scientific process, based on numbers or statistical values, and useful in providing complete insight to the data.

➔ Non Statistical Analysis: It is based on very generic information, and exclusive of numbers, or statistical or quantitative data.

Basics of Statistics:

- ★ Statistical Population: A collection of all probable observations of a specific characteristic of interest.
  - Example: All students learning at specific college.
- ★ Sample: A subset of population.
  - Example: A group of students selected from various branches having similar types of hobbies.
- ★ Variable: An item of interest that can acquire various numerical values.
  - Examples are blood groups, disabilities, pan numbers, aadhaar numbers of students in the sample set.
- ★ Parameter: A population characteristic of interest.
  - Example: The average weight and height of students of sample set

Types of Statistics:

- Descriptive Analytics/Statistics: It means to organize the data and focus on the main characteristics of the data.
- Inferential Analytics/Statistics: It means to use the probability theory to arrive at the conclusion.
  - Random sample is drawn from population which will describe population thoroughly.
  - By using statistical tools, describe the data and make inferences upon the data.

Population and Sample:

- Population: Consists of Various Samples. The samples together represent the population
- A Sample is:
  - The part/piece drawn from the population
  - The subset of population
  - A random selection to represent the characteristics of the population
  - Representative analysis of the entire population

Data Science

Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

Predictive causal analytics:

- If we want a model which can predict the possibilities of a particular event in the future, we need to apply predictive causal analytics.
- Say, if we are providing money on credit, then the probability of customers making future credit payments on time is a matter of concern for us.
- Here, we can build a model which can perform predictive analytics on the payment history of the customer to predict if the future payments will be on time or not.

Prescriptive analytics:

- If we want a model which has the intelligence of taking its own decisions and the ability to modify it with dynamic parameters, we certainly need prescriptive analytics for it.
- This relatively new field is all about providing advice. In other terms, it not only predicts but suggests a range of prescribed actions and associated outcomes.
- The best example for this is Google's self-driving car. The data gathered by vehicles can be used to train self-driving cars. We can run algorithms on this data to bring intelligence to it. This will enable our car to take decisions like when to turn, which path to take, when to slow down or speed up.

Machine learning for making predictions:

- If we have transactional data of a finance company and need to build a model to determine the future trend, then machine learning algorithms are the best bet.
- This falls under the paradigm of supervised learning. It is called supervised because we already have the data based on which we can train our machines.
- For example, a fraud detection model can be trained using a historical record of fraudulent purchases.

Machine learning for pattern discovery:

- If we don't have the parameters based on which we can make predictions, then we need to find out the hidden patterns within the dataset to be able to make meaningful predictions.
- This is nothing but the unsupervised model as we don't have any predefined labels for grouping. The most common algorithm used for pattern discovery is Clustering.
- Let's say we are working in a telephone company and we need to establish a network by putting towers in a region. Then, we can use the clustering technique to find those tower locations which will ensure that all the users receive optimum signal strength.

Lifecycle of Data Science: Lifecycle of Data Science includes following cycles.

1. Discovery
2. Data Preparation
3. Model Planning
4. Model Building
5. Operationalize
6. Communicate Results

Discovery:

- Before we begin the project, it is important to understand the various specifications, requirements, priorities and required budget.
- We must possess the ability to ask the right questions. Here, we assess if we have the required resources present in terms of people, technology, time and data to support the project.
- In this phase, we also need to frame the business problem and formulate initial hypotheses (IH) to test.

Data Preparation:

- In this phase, we require analytical sandbox in which we can perform analytics for the entire duration of the project.
- We need to explore, preprocess and condition data prior to modeling. Further, we will perform **ETLT** (extract, transform, load and transform) to get data into the sandbox.
- We use Python for data cleaning, transformation, and visualization. This will help us to spot the outliers and establish a relationship between the variables.
- Once we have cleaned and prepared the data, it's time to do exploratory analytics on it.

Model Planning:

- Here, we will determine the methods and techniques to draw the relationships between variables. These relationships will set the base for the algorithms which we will implement in the next phase.
- We will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.
- Now that we have got insights into the nature of our data and have decided the algorithms to be used, in the next stage, we will apply the algorithm and build up a model.

Model Building:

- In this phase, we will develop datasets for training and testing purposes.
- We will consider whether our existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing).
- We will analyze various learning techniques like regression, classification, association or clustering to build the model.

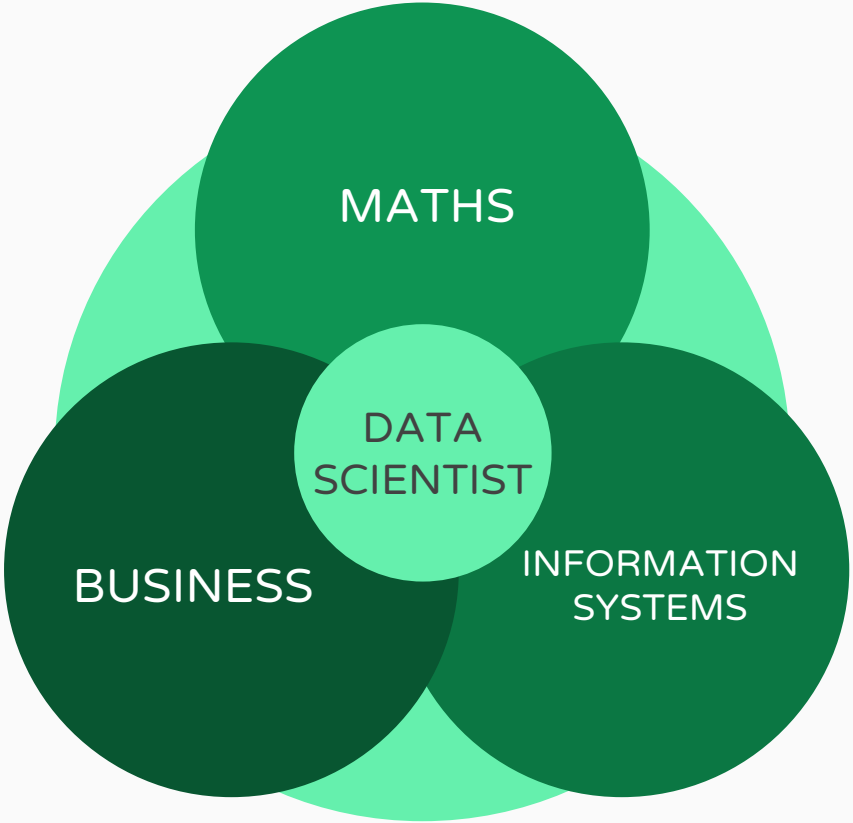
Operationalize:

- In this phase, we deliver final reports, briefings, code and technical documents.
- In addition, sometimes a pilot project is also implemented in a real-time production environment. This will provide us a clear picture of the performance and other related constraints on a small scale before full deployment.

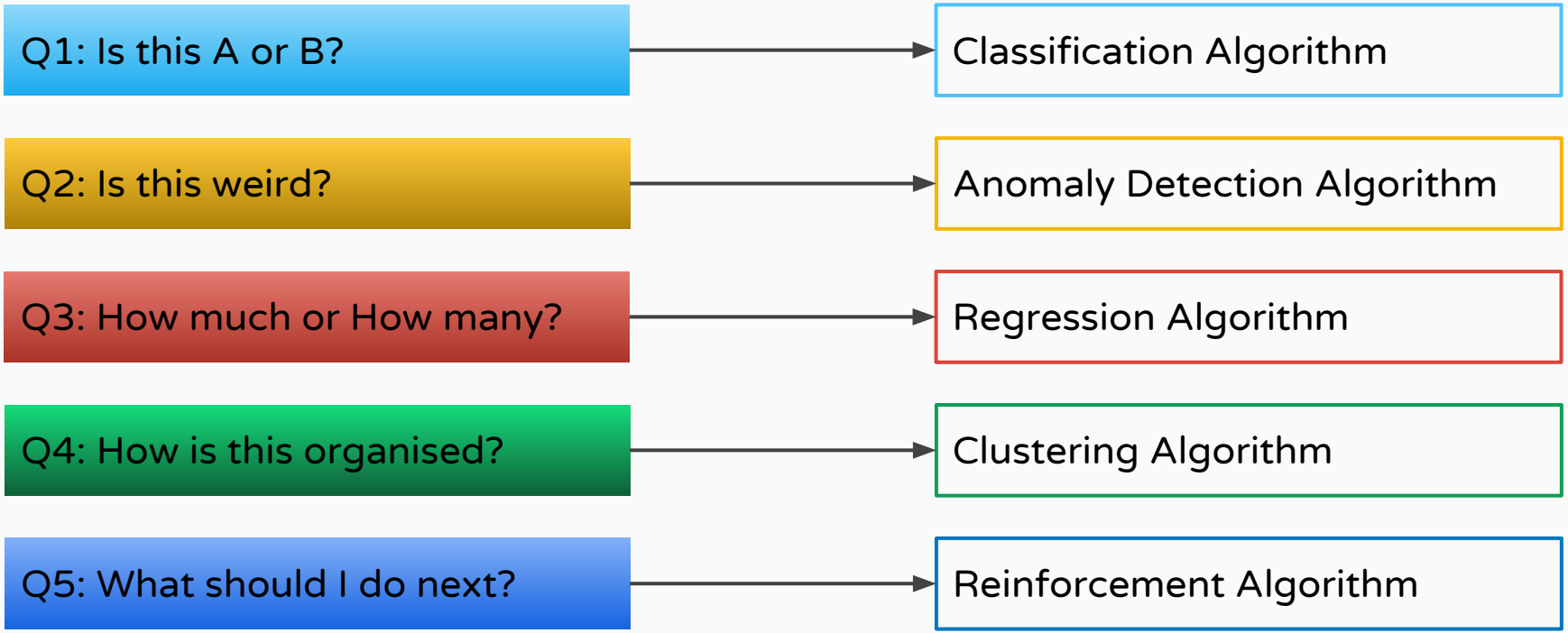
Communicate Results:

- Now it is important to evaluate if we have been able to achieve our goal that we had planned in the first phase.
- So, in the last phase, we identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Step 1.

Who is the Data Scientist?



How to solve a problem in Data Science?



Is this A or B?

- With this question, we are referring to problems which have a categorical answer, as in problems which have a fixed solution, the answer could either be a yes or a no, 1 or 0, interested, maybe or not interested.

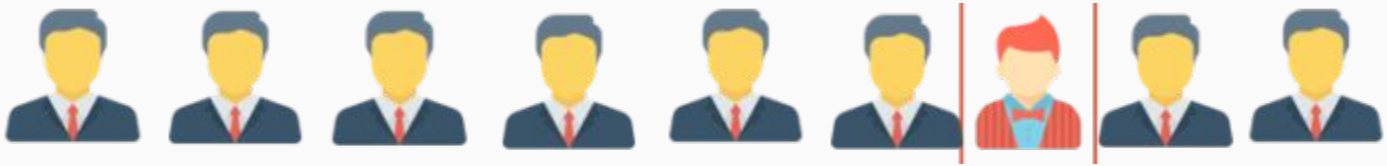
For Example:

- Q. What will you have, Tea or Coffee? Here, we cannot say we would want a coke! Since the question only offers tea or coffee, and hence we may answer one of these only.
- When we have only two type of answers i.e yes or no, 1 or 0, it is called 2 – Class Classification. With more than two options, it is called Multi Class Classification.
- Concluding, whenever we come across questions, the answer to which is categorical, in Data Science we will be solving these problems using Classification Algorithms.

Is this weird?

- Questions like these deal with patterns and can be solved using Anomaly Detection algorithms.

For Example:



- What is weird in the above pattern? The red guy, isn't it?
- Whenever there is a break in pattern, the algorithm flags that particular event for us to review. A real world application of this algorithm has been implemented by Credit Card companies where in, any unusual transaction by a user is flagged for review. Hence implementing security and reducing human effort on surveillance.

How much or How many?

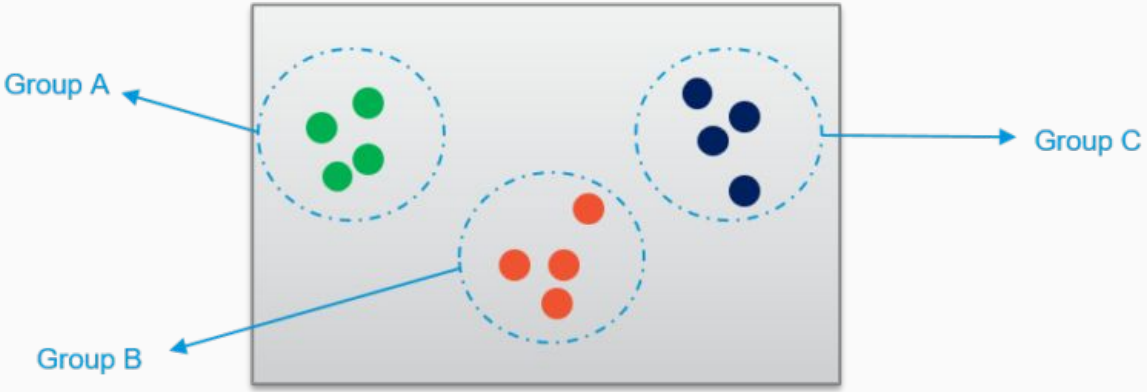
- Regression algorithms are used to predict the Numerical Value.
- So, whenever there is a problem which may ask for figures or numerical values, we solve it using Regression Algorithms.

For Example:

- What will be the temperature for tomorrow?
- Since we expect a numeric value in the response to this problem, we will solve it using Regression Algorithms.

How is this organised?

- Say we have some data, now we don't have any idea, how to make sense out of this data. Hence the question, how is this organised?



- Clustering algorithms group the data in terms of characteristics which are common.
- For example in the above diagram, the dots are organised based on colors. Similarly, be it any data, clustering algorithms try to apprehend what is common between them and hence "clusters" them together.

What should I do next?

- Whenever we encounter a problem, wherein our computer has to make a decision based on the training that we have given it, it involves Reinforcement Algorithms.

For Example:

- Our temperature control system, when it has to decide whether it should lower the temperature of the room, or increase it.

End

Thanks!

Signitive Technologies

13, Gawande Layout, New Sneh Nagar, Behind ICICI Bank, Chhatrapati Square, Nagpur 15

Landmark: Chhatrapati Square Petrol Pump

Contact: 9011033776

www.signitivetech.com

“Keep Learning, Happy Learning”

Best Luck!

Have a Happy Future