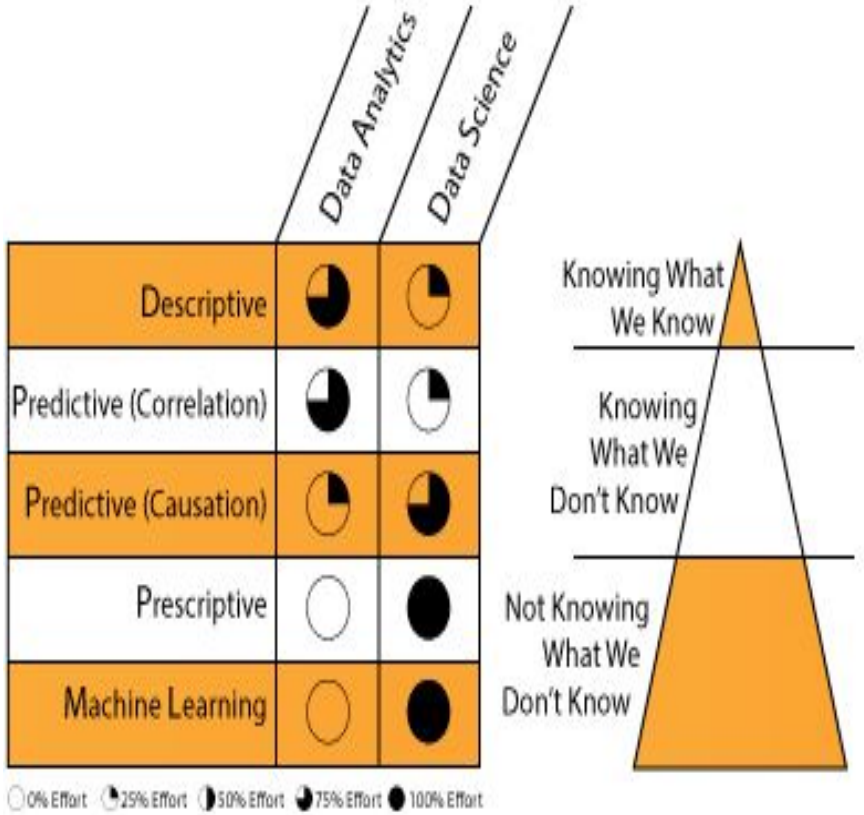
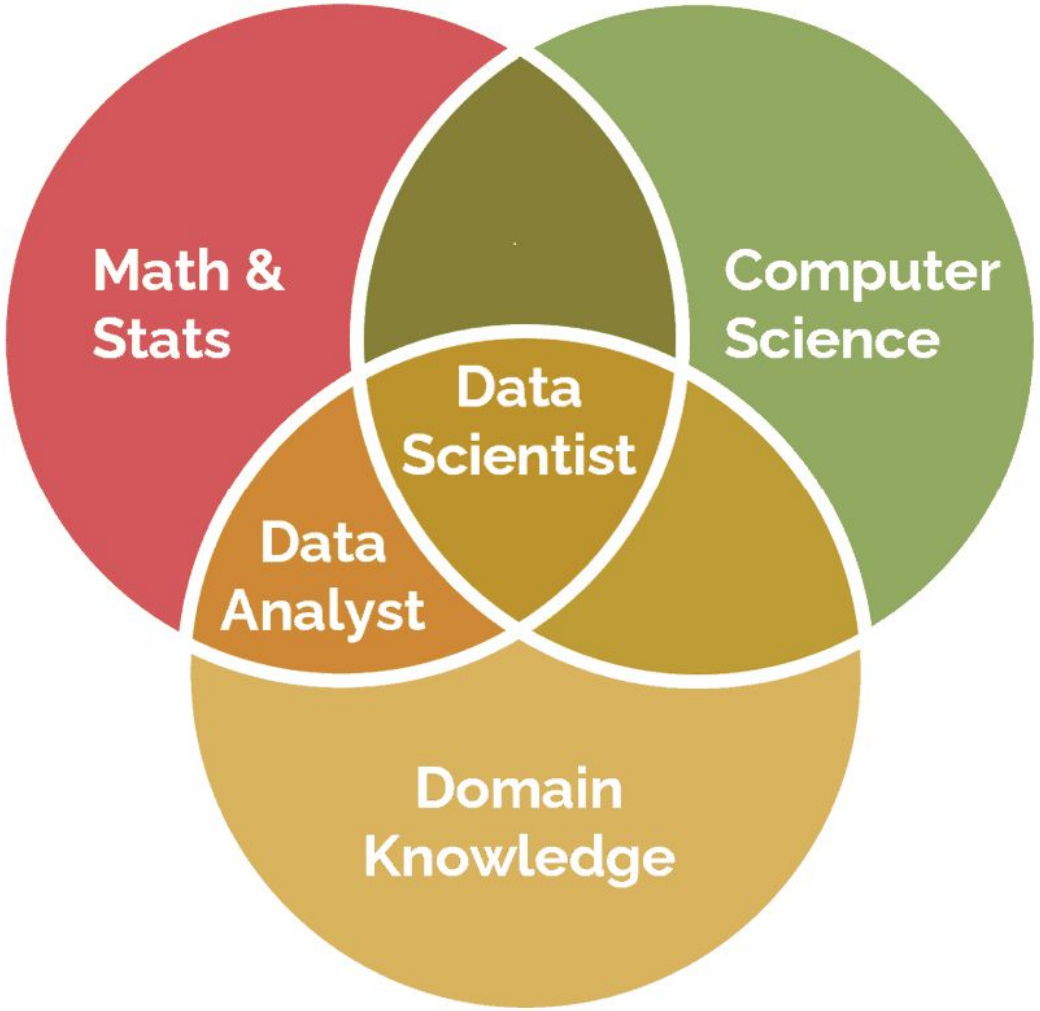
Data Science Using R Programming

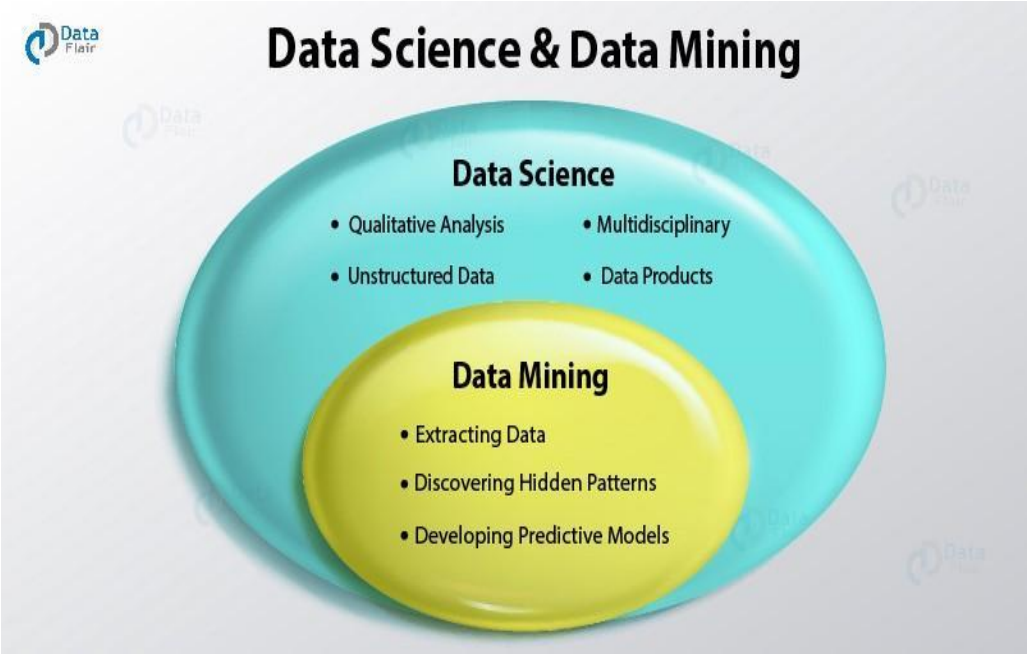
UNIT I

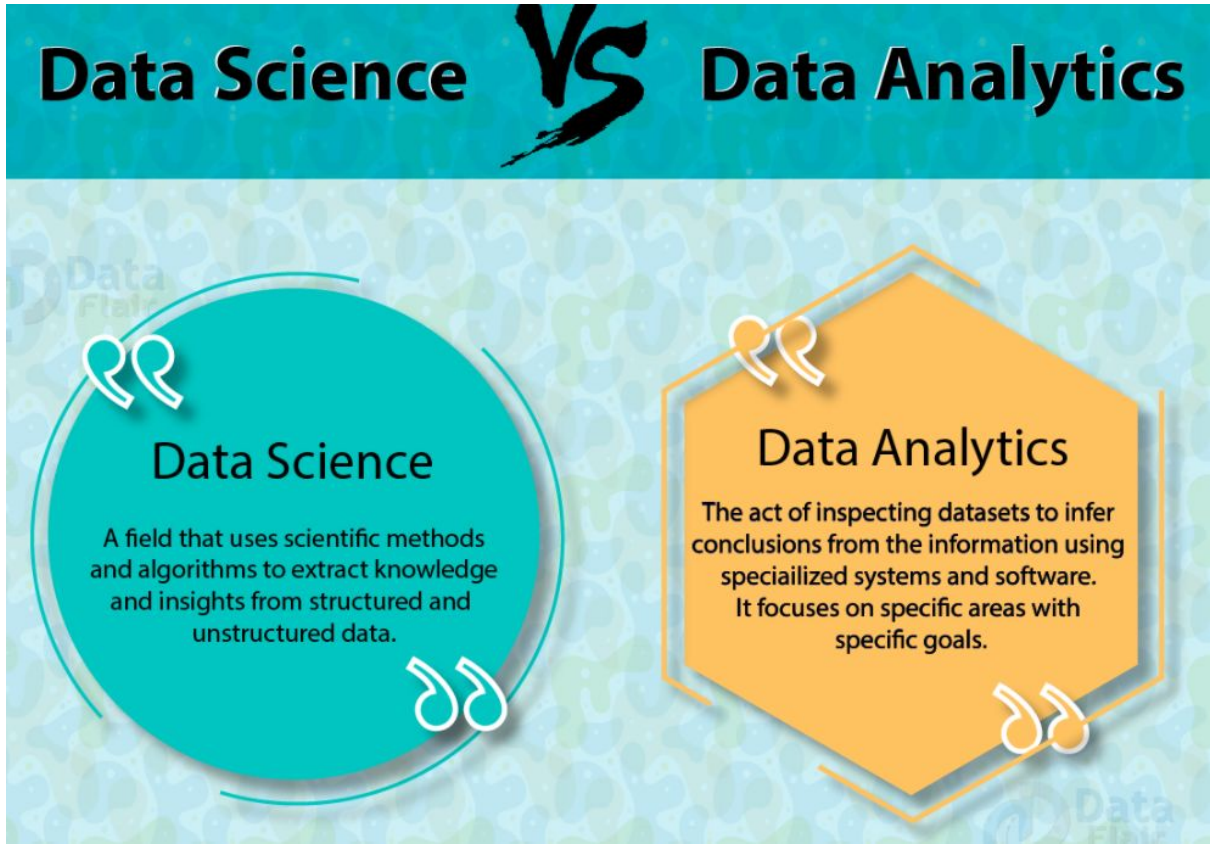
What is Data Science?

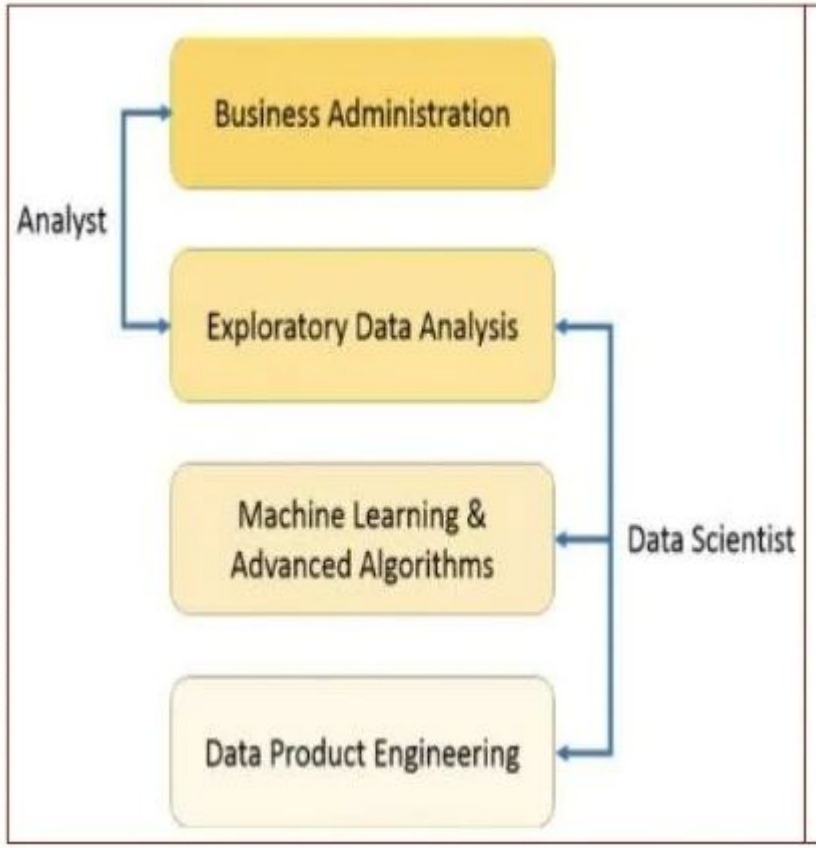


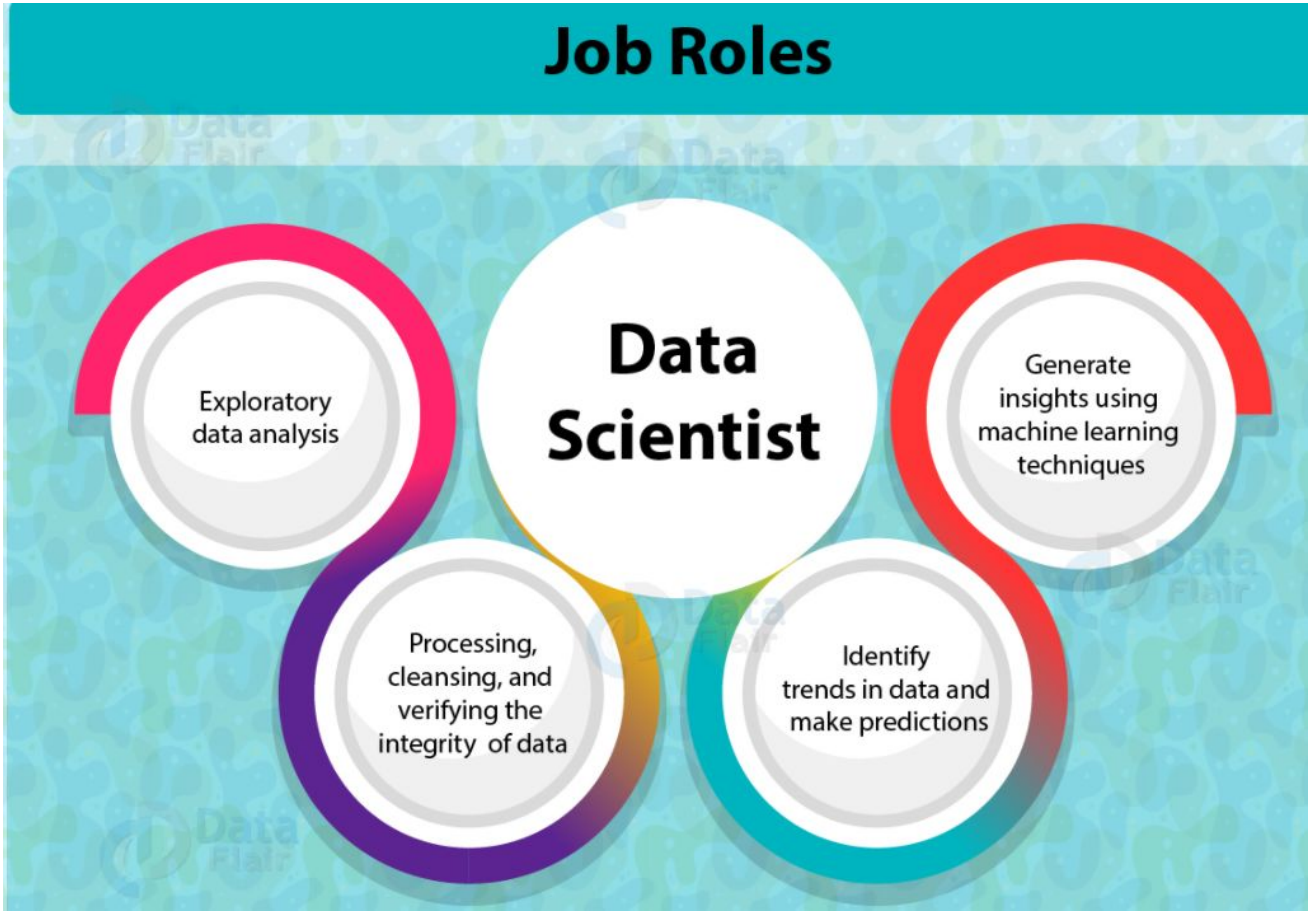


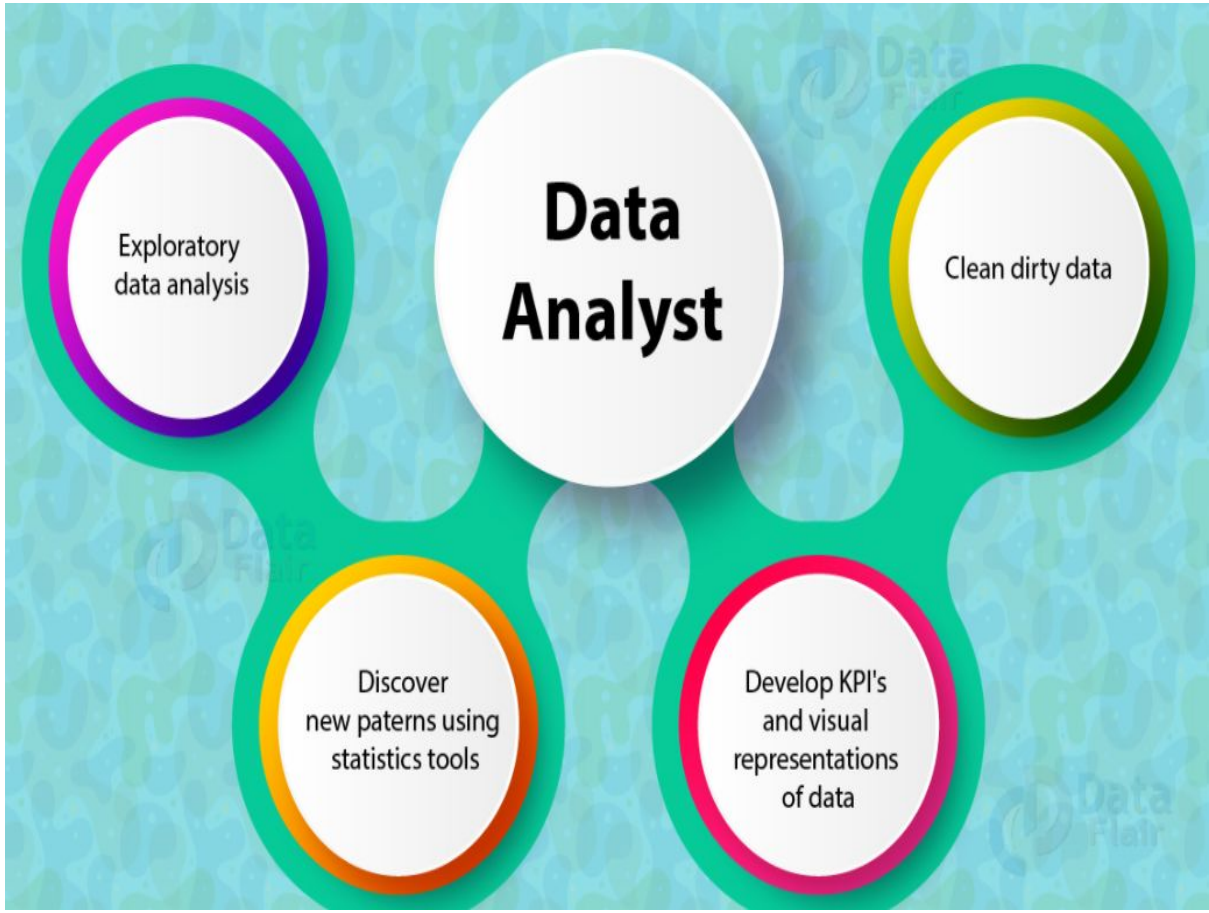
- A

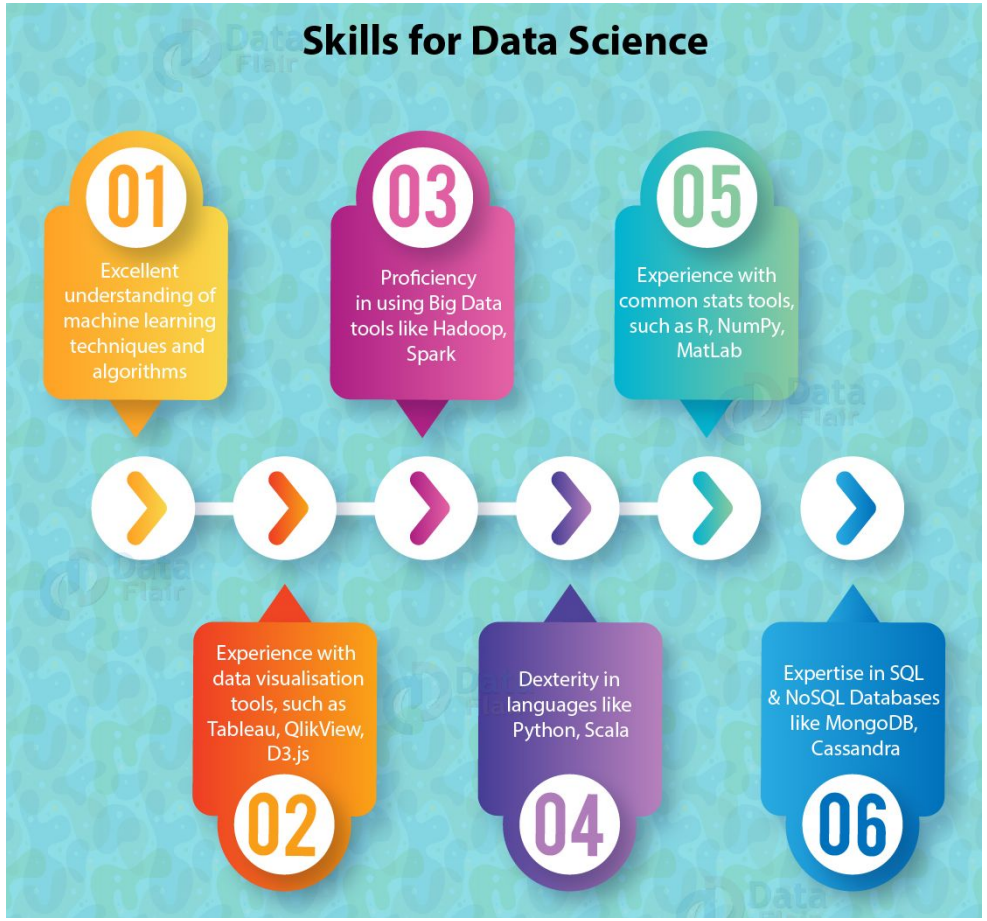

Data Flair
Data Science & Data Mining
Data Science
• Qualitative Analysis
• Multidisciplinary
• Unstructured Data
• Data Products
Data Mining
• Extracting Data
• Discovering Hidden Patterns
• Developing Predictive Models

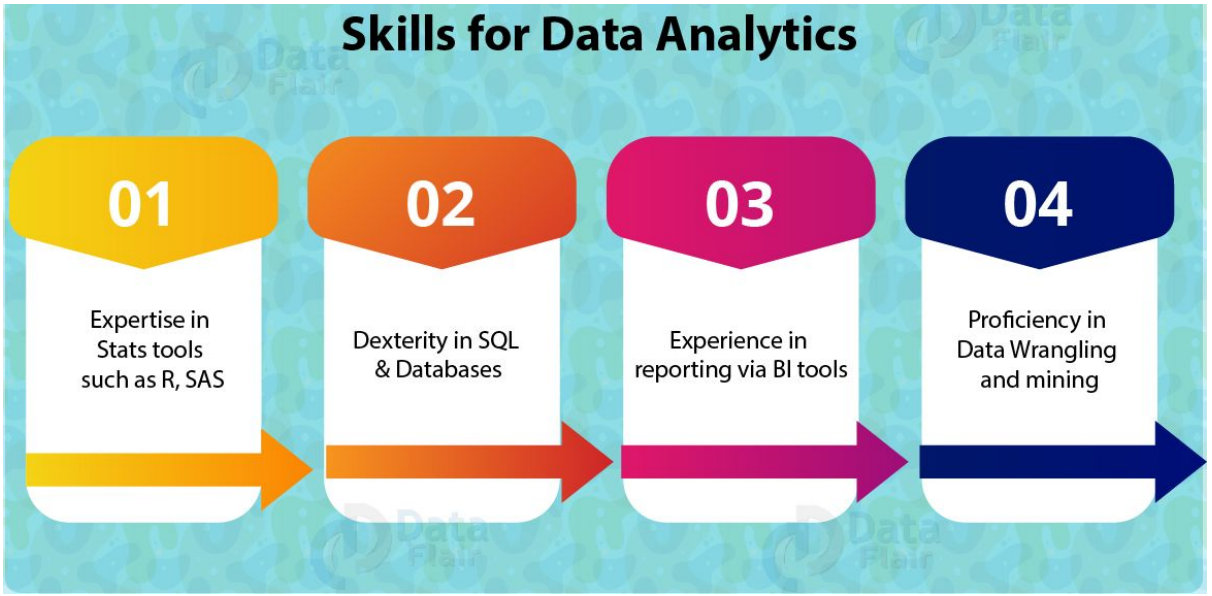

Data Science
VS
Data Analytics
Data Science
A field that uses scientific methods and algorithms to extract knowledge and insights from structured and unstructured data.
Data Analytics
The act of inspecting datasets to infer conclusions from the information using speciailized systems and software. It focuses on specific areas with specific goals.


Analyst
Business Administration
Exploratory Data Analysis
Machine Learning & Advanced Algorithms
Data Product Engineering
Data Scientist


Job Roles
Data Scientist
Exploratory data analysis
Generate insights using machine learning techniques
Processing, cleansing, and verifying the integrity of data
Identify trends in data and make predictions


Data Analyst
Exploratory data analysis
Clean dirty data
Discover new paterns using statistics tools
Develop KPI's and visual representations of data


Skills for Data Science
01
Excellent understanding of machine learning techniques and algorithms
02
Experience with data visualisation tools, such as Tableau, QlikView, D3.js
03
Proficiency in using Big Data tools like Hadoop, Spark
04
Dexterity in languages like Python, Scala
05
Experience with common stats tools, such as R, NumPy, MatLab
06
Expertise in SQL & NoSQL Databases like MongoDB, Cassandra


Skills for Data Analytics
01
Expertise in Stats tools such as R, SAS
02
Dexterity in SQL & Databases
03
Experience in reporting via BI tools
04
Proficiency in Data Wrangling and mining

- Data Science is a field that makes use of scientific methods and algorithms

in order to extract knowledge and discover insights from data (structured on un

- Data Analytics is the process of using specialized systems and software to

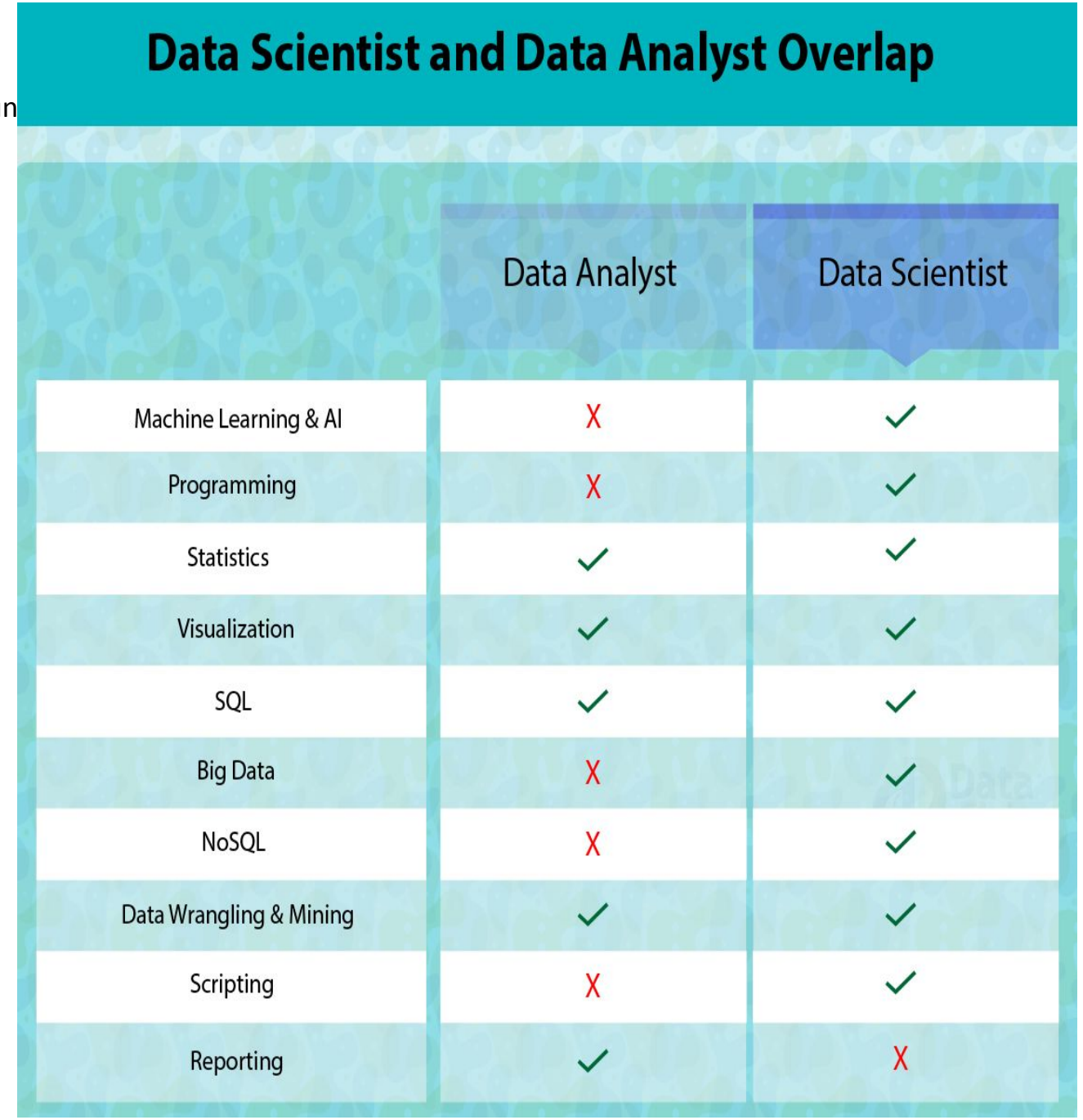
inspect information in datasets in order to derive conclusions.

Data Analyst:

- Perform exploratory data analysis
- Discover new patterns using statistics tools
- Develop KPI's and visual representations of data
- Clean dirty data

Data Scientist:

- Perform exploratory data analysis
- A process, cleanse, and verify the integrity of data
- Identify trends in data and make predictions
- Generate insights using Machine Learning techniques



WHAT IS DATA SCIENCE?

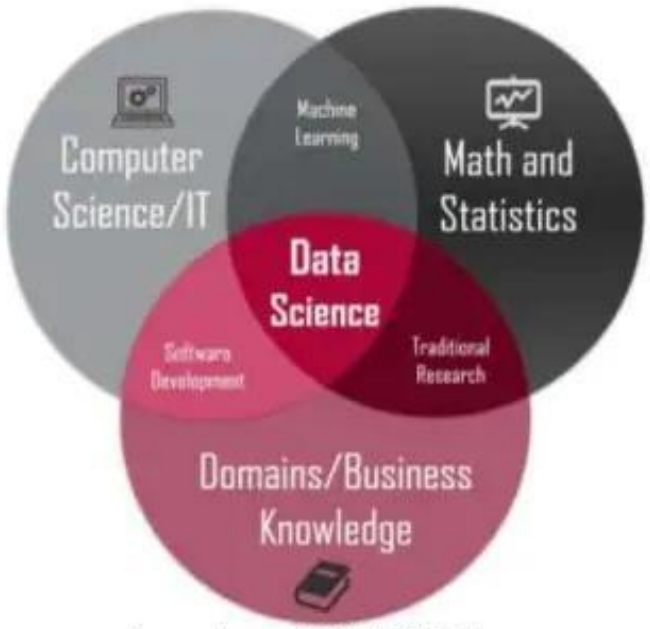
- a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

Source: https://bit.ly/30dekJB

WHAT IS DATA SCIENCE?

- a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.

- employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science.



Source: https://bit.ly/2YTRQ3w

Why learn Linear Algebra?

- You were able to identify the flower because the human brain has gone through million years of evolution. We do not understand what goes in the background to be able to tell whether the colour in the picture is red or black. We have somehow trained our brains to automatically perform this task.

How do you make a computer do the same task ?

How does a machine stores this image?

how can an image such as this with multiple attributes like colour be stored in a computer?

By storing the pixel intensities in a **Matrix**

So any operation which you want to perform on this image would likely use Linear Algebra and matrices at the back end.

- Weights of neural network are stored in the form of a **Matrix**

Cont..

- Representation of problems in Linear Algebra
- 2-variable
- 3-variable
- Multi variable., thousands of variables?

Linear Algebra for data science

- Linear Algebra is a branch of mathematics that is extremely useful in data science and machine learning.
- Most machine learning models can be expressed in matrix form.
- A dataset itself is often represented as a matrix.
- Linear algebra is used in data preprocessing, data transformation, and model evaluation.

Topics you need to be familiar with:

1.Vectors

2.Matrices

- Transpose of a matrix
- Inverse of a matrix
- Determinant of a matrix
- Trace of a matrix
- Dot product
- Eigenvalues
- Eigenvectors

Linear algebra

- Linear algebra is about vectors and matrices and in machine learning we are always working with vectors and matrices (arrays) of data.

- It provides useful shortcuts for describing data as well as operations on data that we need to perform in machine learning methods.

- Linear algebra is the mathematics of data and the notation allows you to describe operations on data precisely with specific operators. You need to be able to read and write this notation.

- Linear Algebra Arithmetic-- need to know how to add, subtract, and multiply scalars, vectors, and matrices
- Linear Algebra for Statistics--The results of some collaborations between the two fields are also staple machine learning methods, such as the PCA for short, used for data reduction.
- Matrix Factorization--some of these may be recognized as "machine learning" methods, such as SVD for short, for data reduction
- Linear Least Squares--Least squares is most known for its role in the solution to linear regression models, but also plays a wider role in a range of machine learning algorithms.

Examples of Linear Algebra in Machine Learning

1. Dataset and Data Files

- 5.1,3.5,1.4,0.2,Iris-setosa
- 4.9,3.0,1.4,0.2,Iris-setosa
- 4.7,3.2,1.3,0.2,Iris-setosa

2. Images and Photographs

3. One Hot Encoding

- Encode categorical variables to make It easier to work with and learn.

  Eg: red, green, blue

  1, 0, 0

  0, 1, 0

  0, 0, 1

4. Linear Regression

- An old method from statistics for describing the relationships between variables. Often used in machine learning for predicting numerical values.
- The most common way of solving linear regression is using matrix factorization methods such as an LU decomposition or SVD.

5. Regularization

- A technique that is often used to encourage a model to minimize the size of coefficients while it is being fit on data is called regularization.
- Common implementations include the $L2$ and $L1$ forms of regularization. Both of these forms of regularization are a measure of the magnitude or length of the coefficients as a vector and are methods lifted directly from linear algebra called

Cont..

6. Principal Component Analysis(PCA)

- Often a dataset has many columns, thousands or more. Modeling data with many features is challenging, and models built from data that include irrelevant features are often less skillful. It is hard to know which features of the data are relevant and which are not. Reducing the number of columns of a dataset Is called dimensionality reduction, and most popular is method is PCA. It uses matrix factorization and eigen decomposition from linear algebra. SVD may also be used in some implementations.

7. Singular-Value Decomposition

- Applications such as feature selection, visualization, noise reduction use SVD.

8. Latent Semantic Analysis

- In natural language processing(NLP), it is common to represent documents as large matrices of word occurrences. Columns of the matrix may be the words and rows may be sentences, paragraphs, pages or documents of text with cells in the matrix marked as the count or frequency of the number of times the word occurred. This is a sparse matrix representation of the text. Matrix factorization methods such as SVD can be applied to this sparse matrix which has the effect of distilling the representation down to its most relevant essence.

9. Recommender Systems

- Predictive modeling problems that involve the recommendation of products are called recommender systems. A simple example is in the calculation of the similarity between sparse customer behavior vectors using distance measures such as Euclidean distance or dot products.
- Matrix factorization methods like the SVD are used widely in recommender systems to distill item and user data to their essence for querying, searching and comparison.

10. Deep Learning

- Deep learning uses artificial neural networks with newer methods and faster hardware on very large datasets. Linear algebra is

Linear equations

- The set of linear equations

  $x1 + x2 = 1,$

  $x1 = -1,$

  $x1 - x2 = 0$

  is written as $Ax = b$ with $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$

Coefficients of linear combinations:

Let $a1, \ldots, an$ denote the columns of A. The system of linear equations $Ax = b$ can be expressed as: $x1a1 + \cdots + xnan = b,$

i.e., b is a linear combination of $a1, \ldots, an$ with coefficients $x1, \ldots, xn$.

So solving $Ax = b$ is the same as finding coefficients that express b as a linear combination of the vectors $a1, \ldots, an$.

Distance

- Distance on Numeric Data: Minkowski Distance

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)

Special Cases of Minkowski Distance

- *h* = 1: Manhattan (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

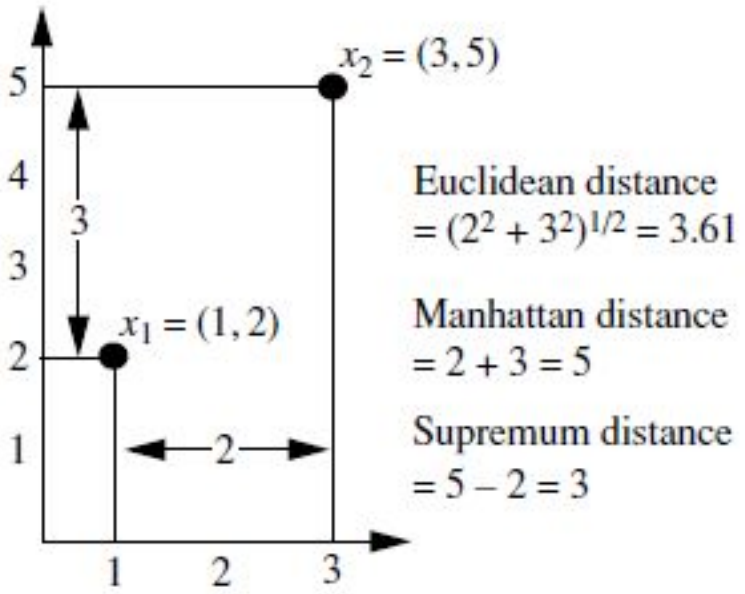
- *h* = 2: (L_2 norm) Euclidean distance

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

- $h \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i,j)=\lim_{h\rightarrow\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}=\max_{f}^{p}|x_{if}-x_{jf}|$$

Ex:



Euclidean, Manhattan, and supremum distances between two objects.

Hyper planes

- Geometrically, a hyperplane is a geometric entity whose dimension is one less than that of its ambient space.
- an ambient space is the space surrounding a mathematical object along with the object itself.

- In 3D space , hyperplane is a geometric entity that is 1 dimension less. So it's 2 dimensions and a 2-dimensional entity in a 3D space would be a plane.
- In 2D space, 1 dimension less would be a single-dimensional geometric entity, which would be a line and so on.

- Hyperplane is usually described by an equation as follows:

$$X^T n + b = 0$$

- If we expand this out for n variables we will get something like this:

$$X_1 n_1 + X_2 n_2 + X_3 n_3 + \ldots\ldots\ldots + X_n n_n + b = 0$$

- In 2D space, we will get following equation which is nothing but an equation of a line.

$$X1n1 + X2n2 + b = 0$$

- Consider a 2D space with $n = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \; and \; b = 4$

- The value of X will be $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

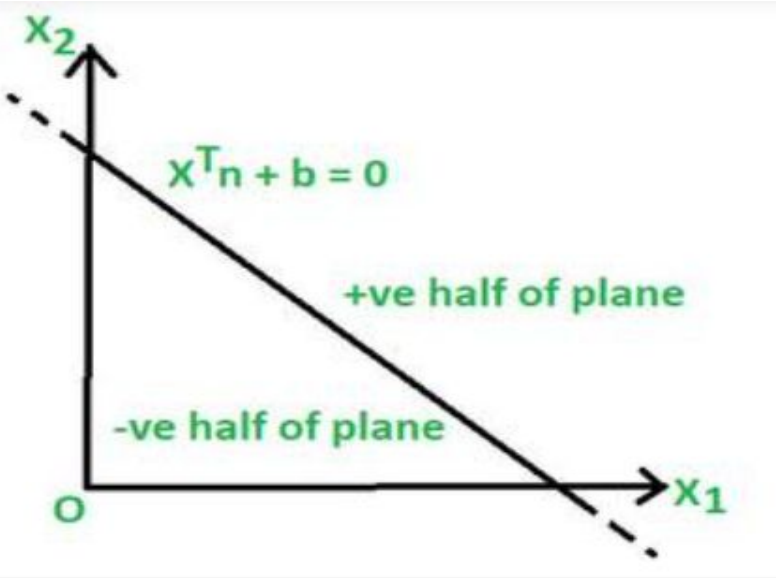
- According to the equation of hyperplane, it can be solved as:

$$X^T n + b = 0$$
$$\begin{bmatrix} x_1 x_2 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} + 4 = 0$$
$$x_1 + 3x_2 + 4 = 0$$

- So from the solution it can be seen that the hyperplane is the equation of a line.

Half-space:

- Consider this 2-dimensional picture:



A hyperplane divides a higher dimensional space into two half-spaces.

A half-plane is one of the halves after the plane has been split in two.

- Solving the previous example, we arrived at the equation of hyper-plane :

  $x1 + 3x2 + 4 = 0$.

There are three possible cases:

- Case 1:(-1,-1)

  $x1 + 3x2 + 4 = 0$ □ point is on the line

- Case 2:(1,-1)

  $x1 + 3x2 + 4 > 0$ □ point is on Positive half-space

- Case 3:(1,-2)

  $x1 + 3x2 + 4 < 0$ □ point is on Negative half-space

Eigen values, Eigen vectors

-