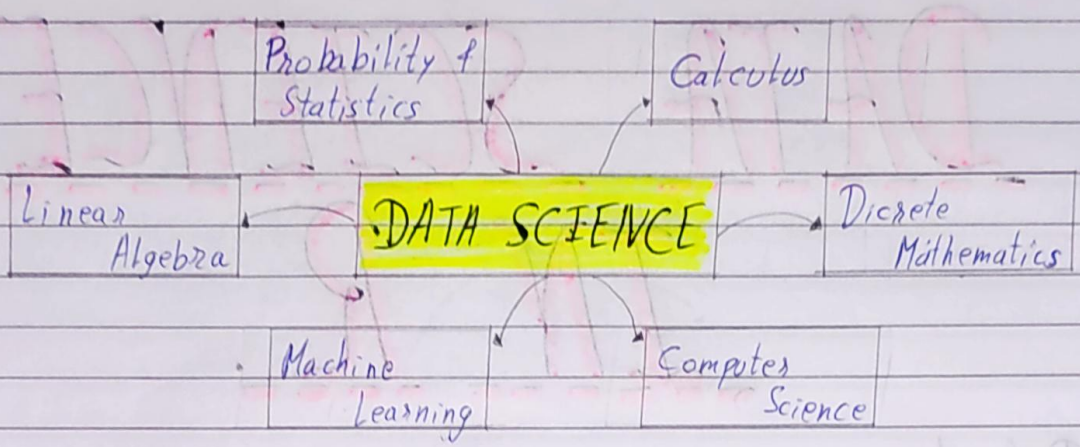
DATA SCIENCE IN R

Definition:

- Data : Collection of facts or information.
- Science : Systematic application of knowledge of the world around us.
- Data Science :- Interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining and analysis (Wikipedia).
  - Combination of various scientific methods, algorithms, systems from ~~be~~ data for better user experience.
  - Methodically aligned data.

Q Why is Data Science in demand?

A. In today's era, there is an abundance of digital data. Currently, less than 0.5% of this data is actually put in use. Data Science can help analyze, manage and use this data, effectively.



Q. Example of a Data Science application.

A. UPS (United Parcel Services) developed a software in 2016 called Orion (On-road Integrated Optimization and Navigation). The software uses algorithms to find the best routes for delivery drivers. The advantage of using Orion - shorter distance, less fuel usage, less engine idle time, more packages delivered.

Other applications can be in sectors such as-

- Retail
- Banking
- Healthcare
- Stock Market
- E-commerce
- Life-science
- Telecommunication
- Etc.

M T W T F S S

Page No.: YOUVA

Date:

Q. What are the advantages of Data Science?

A. The advantages of Data Science are as follows -

- Low cost - Businesses can find the right target to address. - Does not require hefty expenditure overall.
- More productivity - Reaching the larger mass and understanding their needs.
- Less time to solve problems - All the combined data helps derive a better conclusion
- Improved ~~prosse~~ processes - Know what your customer needs
- Competetive advantage - Builds a strong and trusting relation with customers.

o Steps to Align data methodically:

Step 1: Identify your objective

Step 2: Collect suitable information required from the mass.

Step 3: Test the possible hypothesis.

Step 4: ~~Res~~ Find reason for the conclusions derived.

Step 5: Discover patterns.

Step 6: Critical analysis

Step 7: Verify the results.

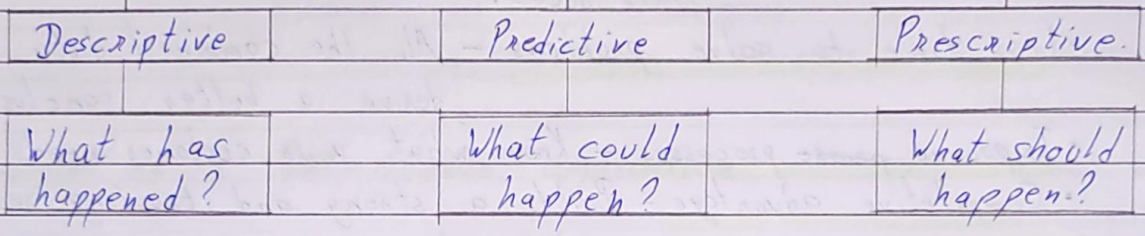
★ Understanding the importance of Data science with -

Market Basket Analysis

We know a general choice of condiment with bread is butter or jam. We can place these items close to each other in super-markets

★ Ask several questions based on the above steps to find a solution.

○ **Types of Analysis.**



○ Factors to consider in Data Science inferences: (conclusion)

- The Problem
- Predefined categories.
- Quantity & Quality of data
- Readiness of data
- Sources of data
- Amount and type of observation
- Attributes of data
- Missing values, if any
- Co-relation of different data.
- Authenticity of result.
- Tools and softwares.

M T W T F S S

Page No.: YOUVA

Date:

PROBABILITY & STATISTICS

Definitions :

- **Probability** : - Likeliness of an outcome to occur.
  - Prediction of a possible event happening.
- **Statistics** : - Organizing, analysing and presenting data to interpret a proper conclusion.

* Probability of an event lies between 0 to 1.

o Relation between probability, statistics and Data Science:

- Quatify the likeliness of an ~~in~~ event occuring in Data Science is predicted by probability.
- Statistics is used to draw a conclusion about a population.
- Statistical analysis requires probability distribution.
- Computation is required for quantitative analysis.
- Computers are used to process complex statistical data.

LINEAR ALGEBRA

Definition :

- **Linear Algebra :** Theory of systems of linear equations, matrices, vector spaces and linear transformations.

◎ Linear Algebra and Data Science :

- Complex scientific problems can be converted into vectors and matrices and solved linearly.
- Statistical computing algorithms can be tedious when it comes to data and linear algebra or iterative methods can be used.
- Linear algebra is a computational engine in data science and ~~preferred~~ preferred over iterative methods.

★ Linear algebra in big data.

- Graphical transformation
- Face morphing
- Object detection / tracking
- Audio & image compression.
- Edge detection
- Blurring
- Signal processing.

○ Iterative Method vs Linear Algebra:

(Frobenius norm in R)

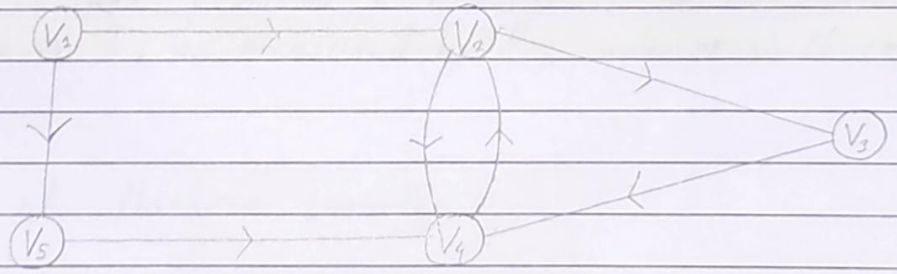
Iterative Method	Linear Algebra

sum ← 0.0
for (i in 1: No_Rows) {
for (j in 1: No_Cols) {
sum1 ← sum1 +
as.numeric (mat[i, j])^2
}
}
	frobenius_norm ←
	norm(as.numeric(mat), type="F")
	
	// norm() is a function for linear algebra in R.

Thus, Linear Algebra is more efficient.

Q. Show implementation of Linear Algebra in real life.

A. Let us check the following graph that shows relation between 5 social media users.



Let us establish the adjacency matrix of the given graph-

×	V_1	V_2	V_3	V_4	V_5
V_1	0	1	0	0	1
V_2	0	0	1	1	0
V_3	0	0	0	1	0
V_4	0	1	0	0	0
V_5	0	0	0	1	0

The above ~~graph~~ matrix can be used in data structure in computer programs for manipulation. Otherwise, it can be used to advertize similar posts to their connections.

Linear Algebra can similarly be used for message transmission or electrical networking.

- More about linear Algebra:

- Linear Algebra transforms large data to matrices for better visual analysis
- This method helps process tons of data points at once rather than each individual part.

MACHINE LEARNING

Definition:

- Machine Learning: "A field of study that gives computers the ability to learn without being explicitly programmed." – Arthur Samuel, 1959.

○ About Machine Learning:

- Concentrates on induction algorithms and on other algorithms that can be said to 'learn'
- Tom M. Mitchell said, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

○ Advantages of Machine Learning:

- large amount of data can be analysed quickly and efficiently.
- Understanding and analysis of subjects out of human expertise; without human intervention.
- Adaptibility to a particular case.

M T W T F S S

Page No.: YOUVA

Date:

Q. Give an example of Machine learning application.

A. IBM's (International Business Machines) or called the Big Blue introduced a program called **Deep Blue**. The chess-playing application beat the chess master, Garry Kasparov in the second match.

- In 1996 → Kasparov (4) - Deep blue (2)
- In 1997 → Kasparov ($2\frac{1}{2}$) - **Deep blue** ($3\frac{1}{2}$)

Other examples of Machine Learning Application -

- Facial recognition
- Handwriting expert.

Q Types of Machine Learning:

MACHINE LEARNING

SUPERVISED LEARNING	UNSUPERVISED LEARNING
• Linear regression	• k-means clustering
• K-nearest neighbours	• Hierarchial clustering
• Decision trees	• Mixture models

SEMI-SUPERVISED LEARNING	REINFORCEMENT LEARNING
• Graph-based methods	• Markov decision process
• Generative models.	• Monte Carlo methods
• Low-density separation	• Temporal-Difference learning

○ Supervised Learning

- Applications with prior data given for analysis.
- Step 1: Machine is given known data to learn
- Step 2: Machine uses what it learned and analyses new data points.
- Step 3: Accuracy of analysis is determined.
- Types: - Classification: Sorted into discrete results.
  - Regression: Continuous and numeric results

○ Unsupervised Learning.

- Prior data not provided; patterns are analysed.
- Step 1: Set some variables or parameters to arrange
- Step 2: Machine groups similar data points using parameter
- Step 3: This grouping is clustered, hence called clustering.

○ Semi-supervised Learning

- Same as supervised; uses labelled and unlabelled data for training.
- Step 1: In case of labelled data points being less than unlabelled; Machine is trained with labelled data points.
- Step 2: Uses above knowledge to label the unlabelled data points.
- Step 3: Combines data of Step 1 & 2 and retrains.
- Step 4: Repeat till converged.

○ Reinforcement Learning.

- Reward-based; finds optimum action where numeric value is highest
- Step 1: No supervision is provided to the machine
- Step 2: The machine determines the best/optimum action.
- Step 3: The analysis is determined on the basis of the most rewarding option through trial and error.

M T W T F S S

Page No.: YOUVA

Date:

COMPUTER SCIENCE

#Definition:

- Computer Science: The study of of principles and use of computers.

○ Importance of Computer Science in Data science:

- Algorithms to be written in programming language for implementation.
- Algorithms use basics of linear algebra
- Statistical computation of data.
- Data management of structured, semi-structured or unstructured data

○ Computer science tools for Data Science.

Programming	Machine-Learning	Data base	Statistics
• R language	• R language	• SQL (My SQL, Oracle, DB2, Maria)	• R language
• Python	• Python (Scikit-Learn)	• No SQL (Cassandra, MongoDB)	• Python (Pandas, Matplotlib)
• Java	• Spark		• MatLab
• Julia	• Weka		• Julia
• Fortran	• Julia		
• C++			

M T W T F S S

Page No.: YOUVA

Date:

DATA SCIENCE: PROCESS & ARCHITECTURE

Intelligence

Data

Information Management

DS Technologies

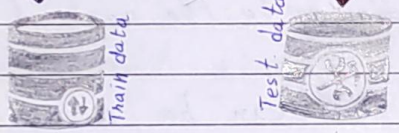
Models

Action

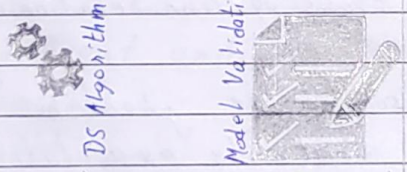
- **Historical data**: Data provided to the machine contains insights from various sources such as apps & sensors or devices to increase efficiency.

Feature Engineering

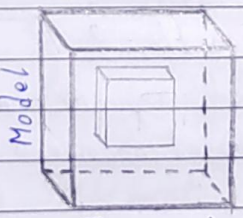
- **Feature Engineering**: To create or select significant features in historical data.



- Use of database, spark, hadoop
- Cleans data
- Divided/parallel computing for huge amount of data



- Following components create a model on the basis of the training data:
- Exploratory data analysis, Statistical Inference, Association analysis, Regression analysis, Time Series analysis & Machine Learning.



- **Model**: Pattern determined/analyzed from the given data.
- New data provided will produce required results

- Use of the model in apps, web, bots, people, devices and other automated systems.

M T W T F S S

Page No.: YOUVA

Date:

DATA SCIENCE: LIFE CYCLE

Define Goal
- What problem needs to be solved?
- What is being done regarding the issue currently?
- What is missing in present solution?

Collect / Manage Data
- What / How much resources / information we have?
- Is it enough and helpful for analysis?
- Is the quality of data up to the mark & clean?

Build Model
- Is there a pattern that can be a solution?
- Use following (any one) modelling techniques — scoring, classification, ranking, clustering, find relation or characterize

Evaluate Model
- Does the model solve the problem?
- Does the model meet the expectation?
- Is the model accurate and/or better than current?

Present results and documents
- How can we solve the problem?
- Does the model cover all requirements by clients?
- Prepare and distribute well-versed documents to team.

Deploy Model
- Now the model must be thoroughly tested.
- It should adapt to unforeseen changes.
- Should be deployed and errors corrected in pilot program.

USE CASES

Definition :

- Use Case : Written description or outline of how a task will be processed / performed.

Q. What are the characteristics of a data science project?

A. Following are must in order to achieve a successful data science project:

- Clear and precise goal.
- Realistic expectations
- Sufficient and unbaised data set
- Right model for the use case
- Correctly represented and deployable model.

o Real - life example / implementation:

- Problem Statement : Country Bank of India needs to reduce their losses due to bad loans.

Let us analyze the problem step - by - step in this business use case.

→ Define the goal → A tool that accurately identifies bad loan applicants, hence reducing the numbers
> Reduce rate of loan charge offs by at least x%.

→ Collect and manage data → > Collect data like duration of loan and employability of applicant;
> Collect data for a certain tenure.
> Visualize, summarize, clean data.
> Do not disregard any data.

→ Build a model → > Here, the method used will be classification of defaulters or non-defaulters.
> Now choose appropriate approach - logistic regression, naive Bayes, k-nearest neighbour, etc.
> Know why and how the model takes decisions.

→ Model evaluation → > Calculate accuracy and precision of the model.
> Compare the predicted and actual values.

→ Present results & documents → > Present efficacy of model to bank officials
> Emphasize study by the model in executive summary
> Show any unique analysis as well.

→ Deploy model → > Experience vs digital analysis maybe in opposition.
> One must look for the more reliable method.

Top 10 use case : Data Science

Use Case	Problem	Solution
Churn Prediction	Losing customers to competition	Repeating customer base
Sentiment analysis/ Opinion mining	How customer feels about products / services	Analyze their likes/dislikes
Online Ads	Rising complexity in ad industry	Study user preference/online trends
Recommendations	Improve return on investment for businesses	Recommender systems based on customer's browsing pattern
Truth and Veracity	Online misinformation affecting business' reputation	Data veracity a.k.a verify accuracy & context of data
News Aggregation	Plethora of news / fake news	Gather news of same topics / Find genuine source of news / Refine by region or preference
Scalability	Handling customers at large	Use chat bots to scale down options
Content Discovery	Right contents for user	Predictive algo(s) for recommedations of content based on history
Intelligent learning	Undesignated / Real time data points / inputs	Real-time analysis through various smart devices
Personalized Medicine	Effects of treatment on patient is un-predictable	Study of individual's genes to suggest treatments and effects

Big Techs providing Data Science solutions:

The 4 leading tech giants providing Data Science solutions are –

IBM

Product: IBM Watson

Clients:
- Moose Jaw
- Staples

AMAZON

Product: Amazon Lex

Clients:
- BuildFax
- Art Finder

GOOGLE

Product: Google Cloud

Clients:
- Johnson & Johnson
- VictoriaPlum

MICROSOFT

Product: Microsoft Azure

Clients:
- Schneider Electric
- Fast Shop

Few other products include –
- Infosys Nia
- Wipro Holmes
- TCS Ignio
- IPSoft
- H2O