



Mr. Dnyandeo S Lavhkare

BE.IT (Govt. College of Engg, Karad.)

ME.CSE (SPPU, Pune.)

Id: dnyaneshlavhkare@gmail.com

ESSENTIALS OF DATA SCIENCE

TEACHING SCHEME	EXAMINATION SCHEME:	CREDITS ALLOTTED:
Theory: 03	End Semester Exam 60-Marks	Credits : 03
Practical: 00 Tutorial: 00	Internal Assessment 40-Marks	
	Total: 100 Marks	Total Credits: 03

EDS Pre-requisites

The Students should have knowledge of,

- Database Concepts.
- Python programming.
- Probability & Statistics.

EDS Objectives

- Introduce R as a programming language.
- Introduce the mathematical foundations required for data science.
- Introduce the first level data science algorithms.
- Introduce a data analytics problem solving framework.
- Introduce a practical real time case study.

ESSENTIALS OF DATA SCIENCE

UNIT – I	Introduction to Data Science	06 - Hrs
UNIT – II	Linear Algebra & Statistical Modeling for Data Science	06 - Hrs
UNIT – III	Optimization for Data Science	06 - Hrs
UNIT – IV	Regression and Classification	06 - Hrs
UNIT – V	Data Analysis and Visualization	06 - Hrs
UNIT – VI	Machine Learning	06 - Hrs

Introduction to Data Science

UNIT
I

Data Science Fundamentals: Data, Data Science Process, Components of Data Science, Data Scientist roles and responsibilities, Introduction to R and R Studio, Variables and Datatypes in R, Data frames, Recasting and Joining of Data frames, Arithmetic, Logical and Matrix Operations in R, **Advanced Programming in R** : Functions, Data Visualization in R Basic Graphics.

06 - Hrs

What is Data?



What is Data?

- Data is a raw fact-----need to process.

OR

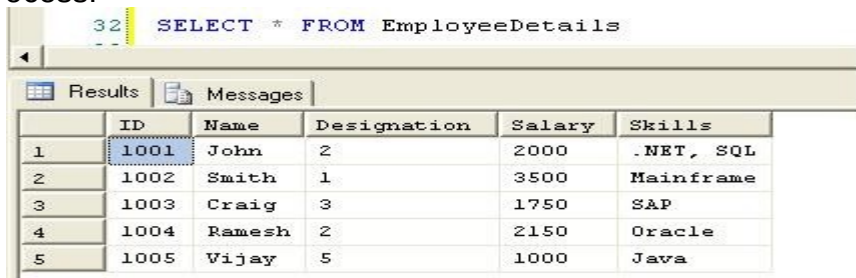
- Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.

Need of Data

- Analysis- Business
- Processing- Unwanted Data Removal
- Accuracy- Business growth

Example: Businesses today are accumulating new data at a rate that exceeds their capacity to extract value from it. The question being faced by every organization is, how data can be used effectively - not just their own data, but all of the data that is available and relevant.

- **Information:** Information is processed, organized and structured data. It provides context for data and enables decision making process.



The screenshot shows a database query interface. At the top, a text box contains the SQL query: `32 SELECT * FROM EmployeeDetails`. Below the text box are two tabs: "Results" and "Messages". The "Results" tab is active, displaying a table with 6 columns: ID, Name, Designation, Salary, and Skills. The table contains 5 rows of data. The first row is highlighted with a blue border.

	ID	Name	Designation	Salary	Skills
1	1001	John	2	2000	.NET, SQL
2	1002	Smith	1	3500	Mainframe
3	1003	Craig	3	1750	SAP
4	1004	Ramesh	2	2150	Oracle
5	1005	Vijay	5	1000	Java

- **Program:** program is a sequence of instructions in a programming language that a computer can execute or interpret.



What is Data Science?



What is Data Science?

What is Data Science?

- Data Science is primarily a combination of Data & Science.
- Data Science is an 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 analytics.

(Ref. Wikipedia)

- Data Science is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process. **(Ref. NIST - The National Institute of Standards and Technology)**

What is Data Science Continue...

- Data science is an amalgamation of different scientific methods, algorithms and systems which enable us to gain insights and derive knowledge from data in various forms.
- Various organizations like Google, Facebook, Uber, Netflix, etc. are already leveraging data science to provide better experiences to their end users.
- **Conclusion:**
The power of Data Science in today's business world we need to acquire skills in various components to master the subject!

Why Data Science?

Why Data Science?

- Data, data everywhere!!!!
- It's a digital world!

In the past few years the amount of data that has been generated in 2.5 quintillion bytes of data. Right at this moment there is more digital data being generated every single second than ever before! No wonder we live in a digital world.

- Data is the oil for today's world. With the right tools, technologies, algorithms, we can use data and convert it into a distinct business advantage
- Data Science can help you to detect fraud using advanced machine learning algorithms.

Few interesting facts about digital data

- There are 40,000 search queries every second (on Google alone), which makes it 3.5 billion searches per day and 1.2 trillion searches per year.
- In **2017**, a staggering 1 trillion photos were taken and billions of them were shared online.
- By **2020**, over 50 billion smart devices will be connected to collect, analyze and share data.
- By **2020**, universe of data will grow from 4.4 zettabytes today to around 44 zettabytes, or 44 trillion gigabytes.
- **Currently, less than 0.5% of all data is analyzed and used! Can you imagine the potential here?**

Case Study: Instant insurance and claims management

- **Lemonade Insurance Company** is an American property and casualty insurance company headquartered in **New York**, offering renters and home insurance policies for homes, apartments, co-ops. Their offerings include a mobile app that tailors insurance deals based on the information that the customers volunteer. This app is embedded with their chat-bot which can further process insurance claims with minimal human interference.



Lemonade Insurance Company

Settled an insurance claim in **3 seconds** in December 2016



What did the company do?

- Spent months learning how claims are handled by humans in old insurance companies
- Used the knowledge to craft an instant claims experience for their claims bot called "AI Jim".



How did AI Jim achieve the world record?

- The bot understands the nature of claims, their severity, and whether the user is in a state of emergency
- It also tries to assess the likelihood of a claim being fraudulent
- It even nudges people to be more honest by incorporating years of behavioral economics research into every little detail in the conversation and the UI
- It settled a claim in just three seconds by running 18 fraud algorithms

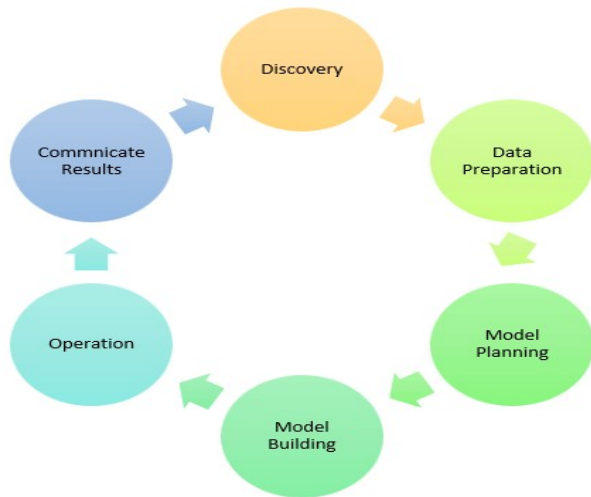
What are the advantages?

- Adoption of **Data Science** lead to the following benefits in business:
 1. Cost reduction
 2. Increase in productivity
 3. Reduction in time taken to solve problems
 4. Process improvements
 5. Competitive advantages



Data Science process: Putting the concepts together

- A data science process helps data scientists use the tools to find unseen patterns, extract data, and convert information to actionable insights that can be meaningful to the company.



1. Discovery

- Discovery step involves acquiring data from all the identified internal & external sources, which helps you answer the business question.
- Logs from web servers
- Data gathered from social media
- Census datasets
- Data streamed from online sources using APIs

2.Data Preparation

- Data can have many inconsistencies like missing values, blank columns, an incorrect data format, which needs to be cleaned. You need to process, explore, and condition data before modelling. The cleaner your data, the better are your predictions.

3. Model Planning

- In this stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

4. Model Building

- In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model, once prepared, is tested against the “testing” dataset.

5. Operationalize

- You deliver the final baselined model with reports, code, and technical documents in this stage. Model is deployed into a real-time production environment after thorough testing.

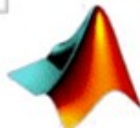
6. Communicate Results

- In this stage, the key findings are communicated to all stakeholders. This helps you decide if the project results are a success or a failure based on the inputs from the model.

Tools for Data Science



SQL



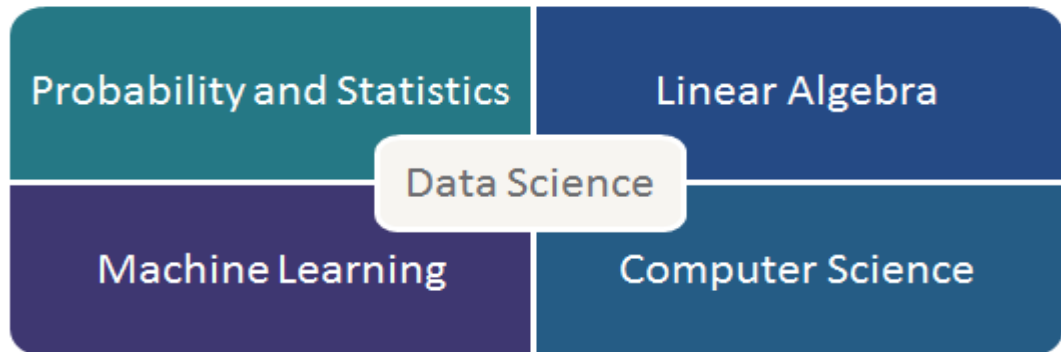
MATLAB

Applications of Data Science

- Some application of Data Science are listed below,
- **Internet Search:** Google search uses Data science technology to search for a specific result within a fraction of a second.
- **Recommendation Systems:** To create a recommendation system. For example, “suggested friends” on Facebook or suggested videos” on YouTube, everything is done with the help of Data Science.
- **Image & Speech Recognition:** Speech recognizes systems like Siri, Google Assistant, and Alexa run on the Data science technique. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.
- **Gaming world:** EA Sports, Sony, Nintendo are using Data science technology. This enhances your gaming experience. Games are now developed using Machine Learning techniques, and they can update themselves when you move to higher levels.

Components of Data Science

- Following are the various components of data science which act like tools to enable a data scientist to draw meaningful insights from data.



- In addition to these we must acquire knowledge about the domain or industry vertical in which we plan to apply Data Science, such as retail, banking & finance, healthcare, e-commerce, life sciences, telecom etc.

Components of Data Science

