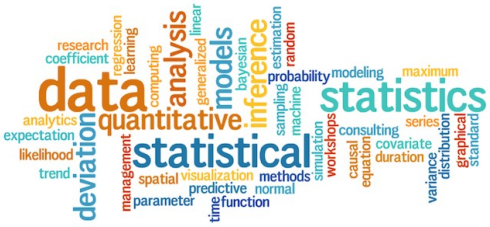
Applied Data Science

Session 1: Introduction

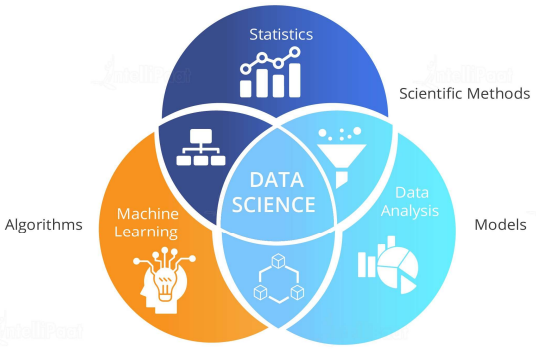
Dr. Soharab Hossain Shaikh



1

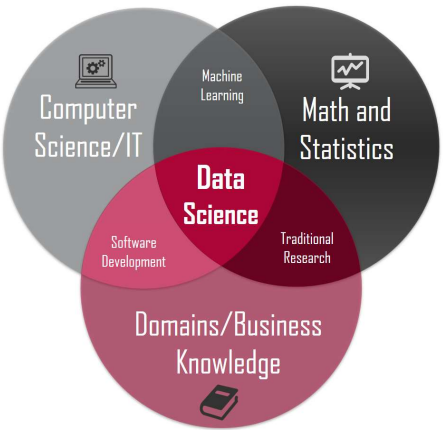
What is Data Science?

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



2

What is Data Science?



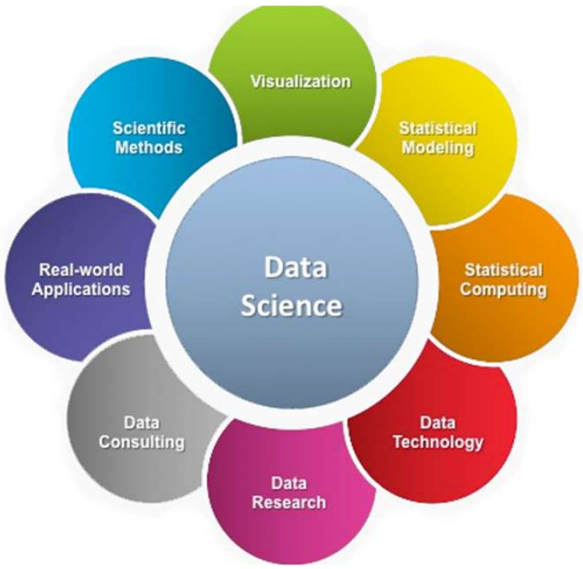
A "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.

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

3

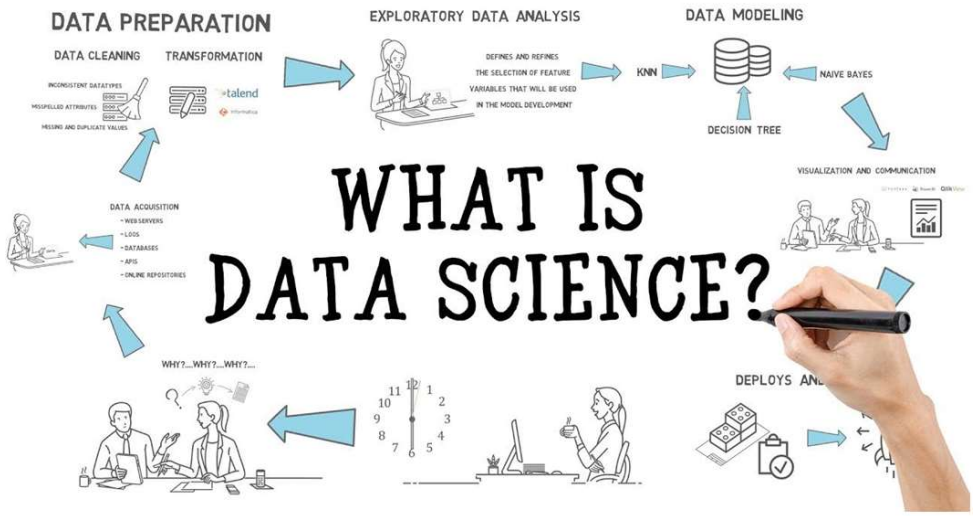
What is Data Science?

- A systematic process to find meaningful patterns from data to get actionable insight.



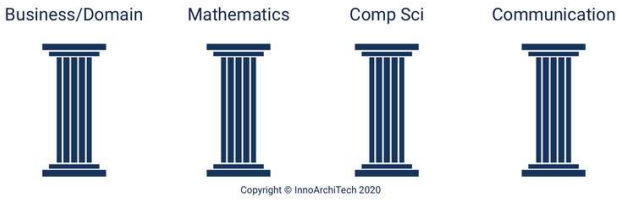
4

What is Data Science?



5

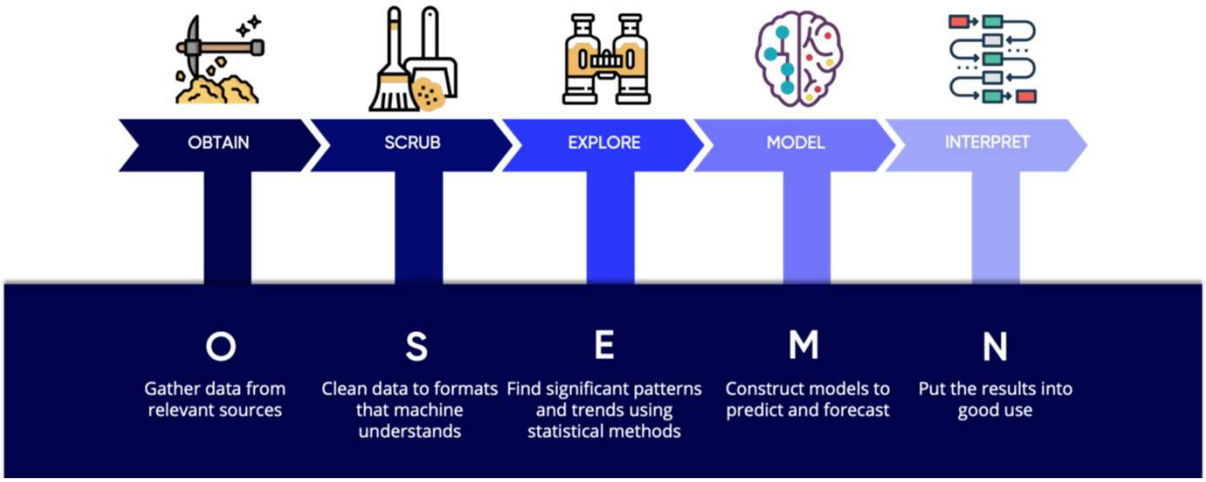
Pillars of Data Science Expertise



- Business/Domain
- Mathematics (includes statistics and probability)
- Computer science (e.g., software/data architecture and engineering)
- Communication (both written and verbal)

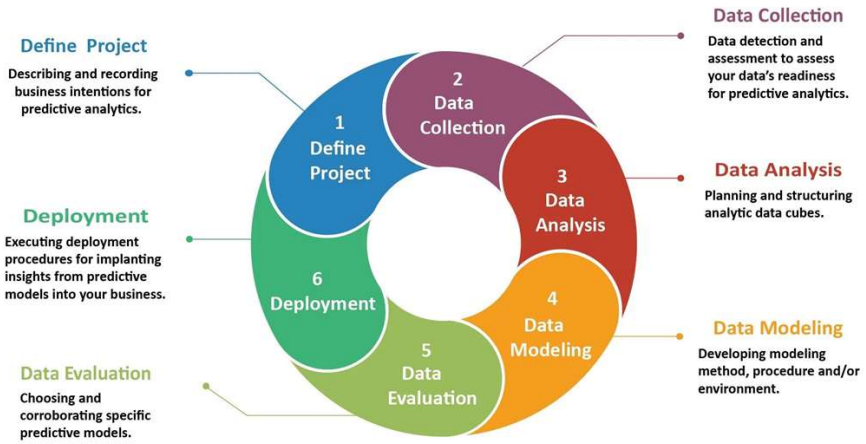
6

Data Science Life Cycle



7

Data Science Life Cycle

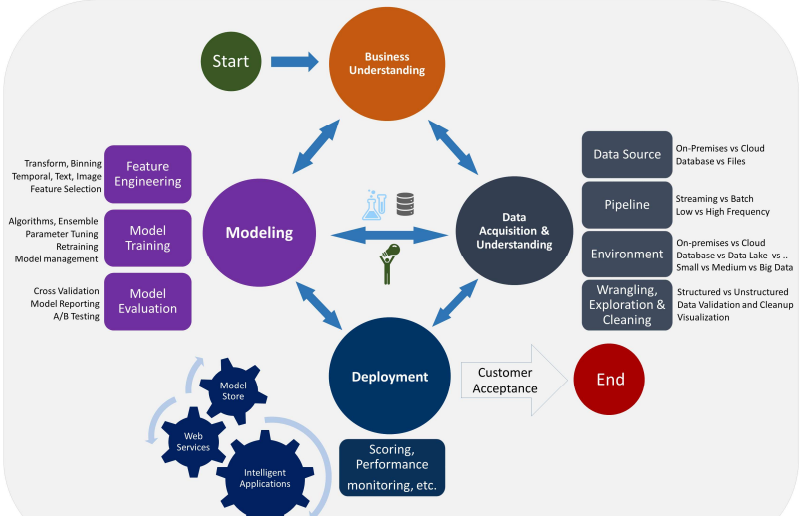


8

Data Science Life Cycle

1. **Define Project Objectives**
   - Specify business problem
   - Acquire subject matter expertise
   - Define unit of analysis and prediction target
   - Prioritize modeling criteria
   - Consider risks and success criteria
   - Decide whether to continue
2. **Acquire & Explore Data**
   - Find appropriate data
   - Merge data into single table
   - Conduct exploratory data analysis
   - Find and remove any target leakage
   - Feature engineering
3. **Model Data**
   - Variable selection
   - Build candidate models
   - Model validation and selection
4. **Interpret & Communicate**
   - Interpret model
   - Communicate model insights
5. **Implement, Document & Maintain**
   - Set up batch or API prediction system
   - Document modeling process for reproducibility
   - Create model monitoring and maintenance plan

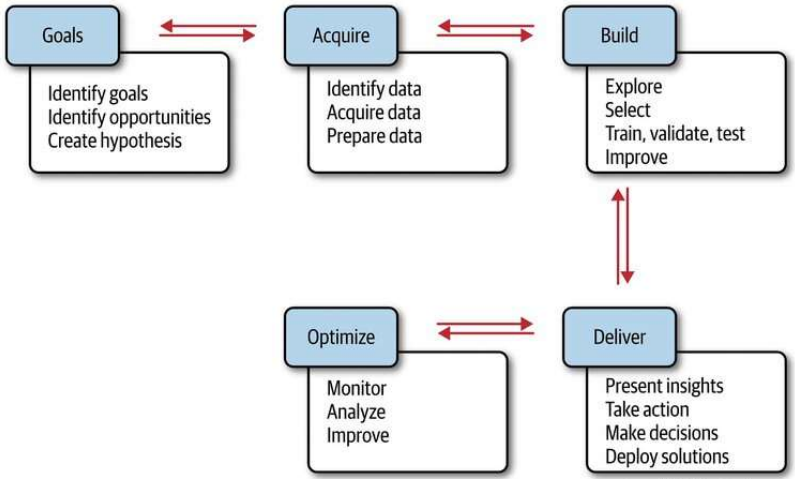
9



Data Science Life Cycle – a closer look

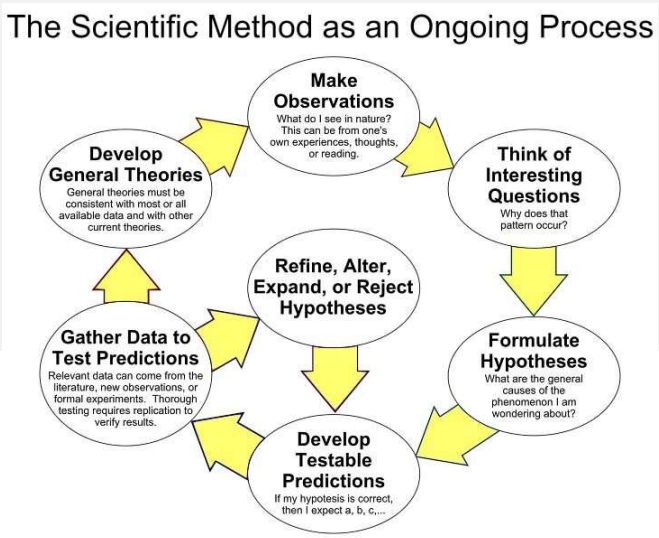
10

The Data Science Process



11

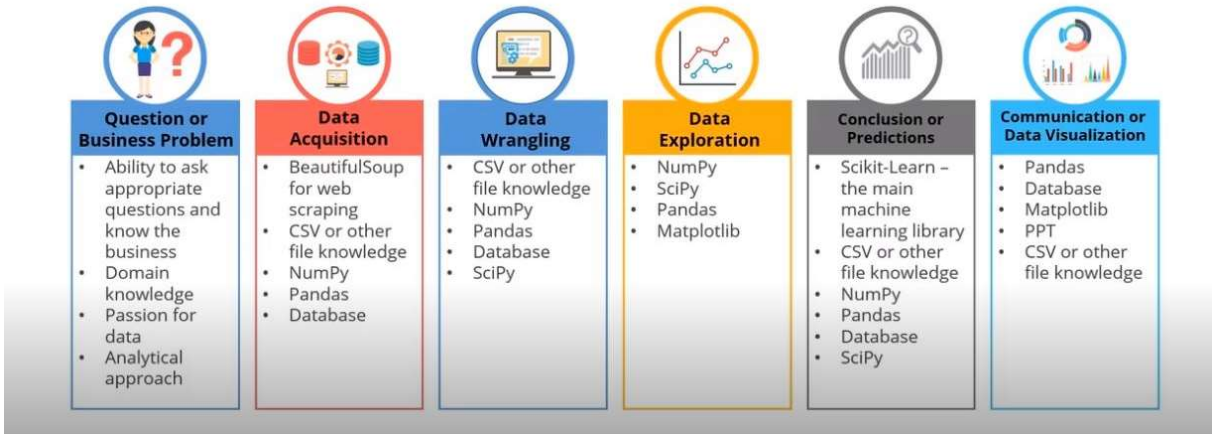
"Science" in Data Science



The term science is usually synonymous with the scientific method, notice that the process outlined in the previous slides is very similar to the process characterized by the expression, scientific method.

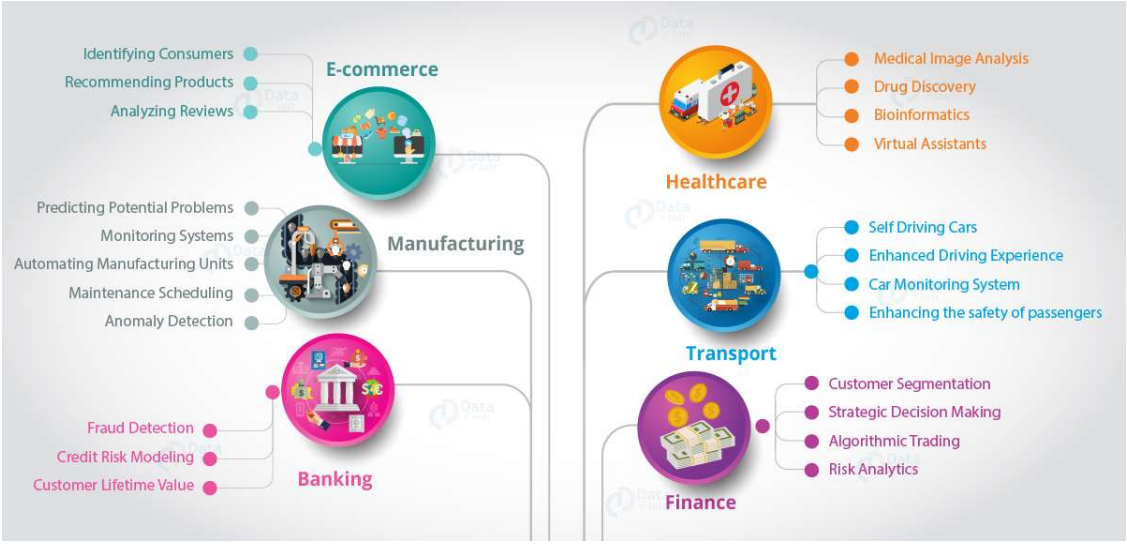
12

Data Analytics Tools



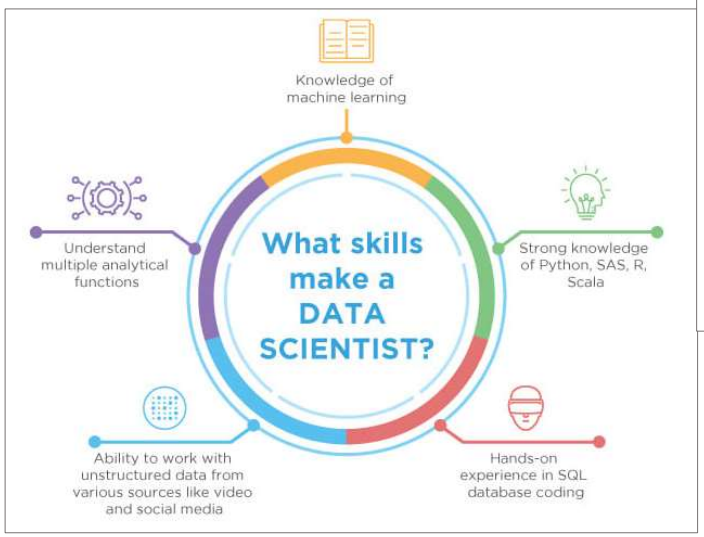
13

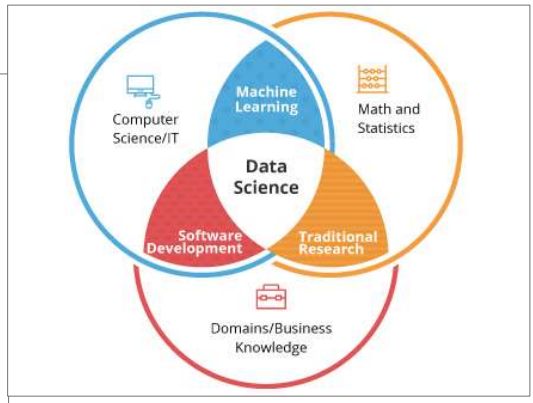
Applications



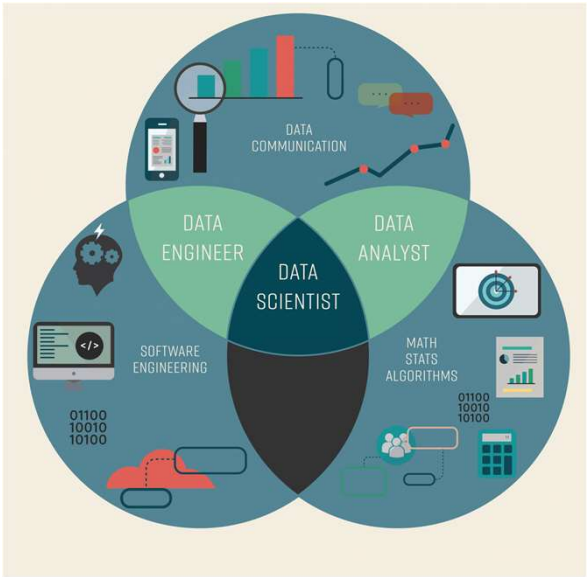
14

Skills to be Developed





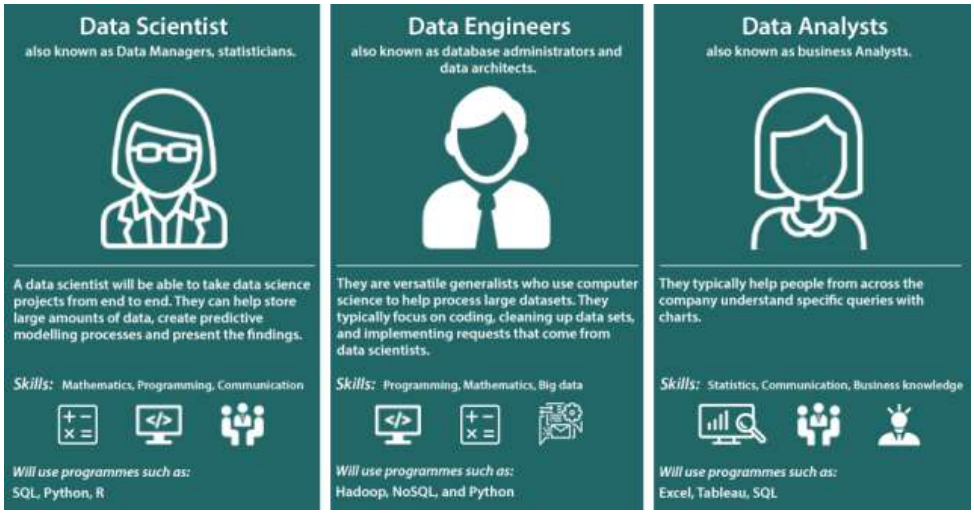
15



Similar Job Roles

16

Similar Job Roles



17

Data Science vs. Data Mining

- Data science is a broad field that includes the processes of capturing of data, analyzing, and deriving insights from it. On the other hand, data mining is mainly about finding useful information in a dataset and utilizing that information to uncover hidden patterns.
- Data Science is a multidisciplinary field that consists of statistics, social sciences, data visualizations, natural language processing, and mining data etc. while the data mining is a subset of the former.
- The role of a data science professional can be considered as a combination of an AI researcher, a deep learning engineer, a machine learning engineer, or a data analyst, to some extent. The person might be able to perform the role of a data engineer as well. On the contrary, a data mining professional doesn't necessarily have to be able to perform all these roles.
- Another notable difference lies in the type of data used by these professionals. Usually, data science deals with every type of data whether structured, semi-structured, or unstructured. On the other hand, data mining mostly deals with structured data.

18

Common Data Science Deliverables

- Prediction (predict a value based on inputs)
- Classification (e.g., spam or not spam)
- Recommendations (e.g., Amazon and Netflix recommendations)
- Pattern detection and grouping (e.g., classification without known classes)
- Anomaly detection (e.g., fraud detection)
- Recognition (image, text, audio, video, facial, …)
- Actionable insights (via dashboards, reports, visualizations, …)
- Automated processes and decision-making (e.g., credit card approval)
- Scoring and ranking (e.g., FICO score)
- Segmentation (e.g., demographic-based marketing)
- Optimization (e.g., risk management)
- Forecasts (e.g., sales and revenue)

19

20

To be continued in the next session…..

21