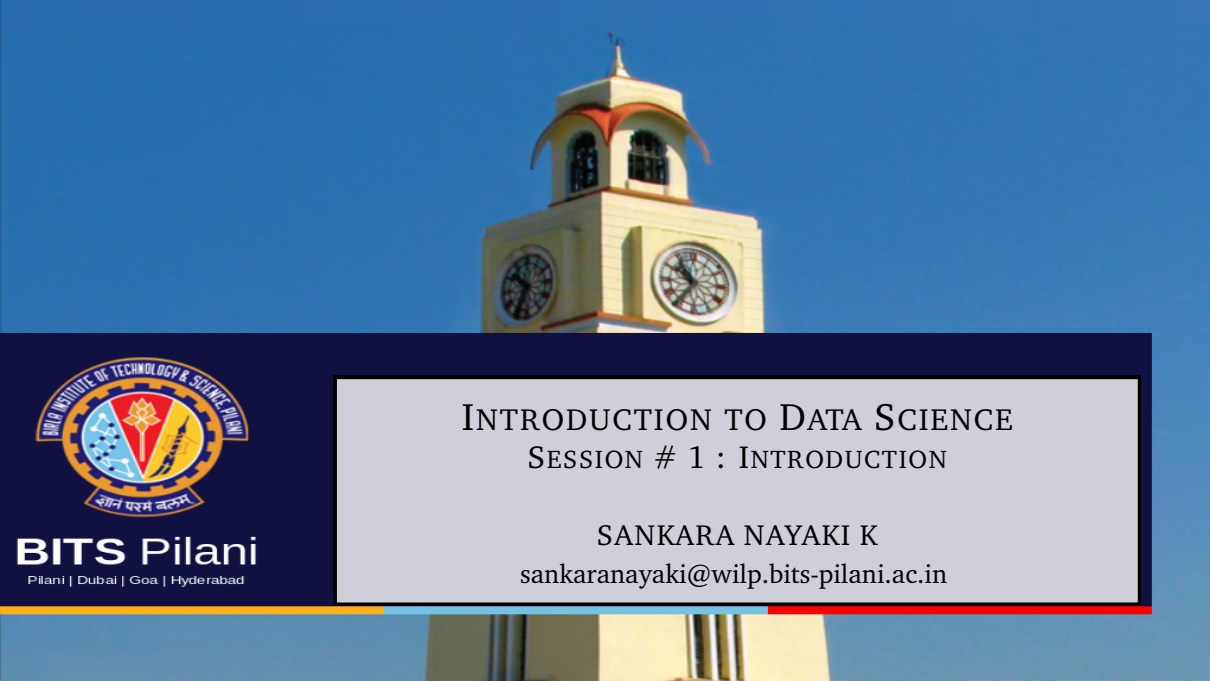


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Session # 1 : Introduction

SANKARA NAYAKI K
sankaranayaki@wilp.bits-pilani.ac.in

The instructor is gratefully acknowledging
the authors who made their course
materials freely available online.

References:

- Introducing Data Science by Cielen, Meysman and Ali
- Storytelling with Data by Cole Nussbaumer Knaflic; Wiley
- Introduction to Data Mining by Tan, Steinbach and Vipin Kumar
- The Art of Data Science by Roger D Peng and Elizabeth Matsui
- Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas

Table of Contents

1. Course Handout
2. Evaluation Components
3. Data Science

Course Handout

M1 Introduction to Data Science
M2 Data Analytics
M3 Data Science Process
M4 Data Science Teams
M5 Data and Data Models
M6 Data wrangling and Feature Engineering
M7 Data visualization
M8 Storytelling with Data
M9 Ethics for Data Science

Lab Sessions

The Lab capsules has to be practised. Details will be posted soon in Impartus.

L1 17-Oct-20

L2 16-Jan-21

Platforms / Tools

1. Python

Dataset

1. Iris dataset

Any other dataset can be used.

Table of Contents

1. Course Handout
2. Evaluation Components
3. Data Science

Evaluation Components

Component	Weightage	Deadline
Quiz-I (Pre-Mid)	5 %	TBA
Quiz-II (Post-Mid)	5 %	TBA
Assignment	20 %	TBA
Mid-Semester Exam	30 %	Per Schedule
Comprehensive Exam	40 %	Per Schedule

Further announcements will be posted in Canvas.

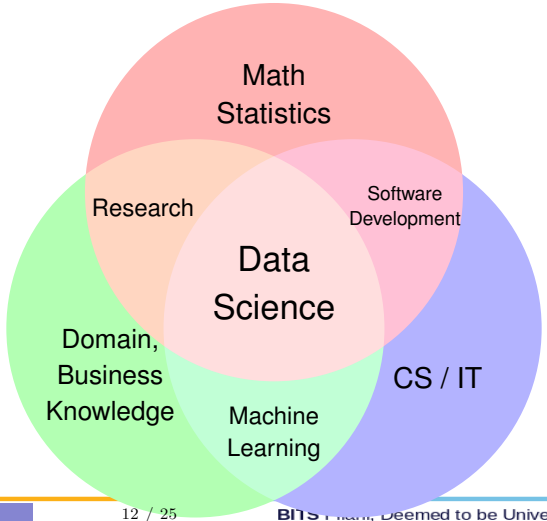
Table of Contents

1. Course Handout
2. Evaluation Components
3. Data Science

Data Science

- Data Science is a study of data.
- Data Science is an art of uncovering insights and trends that are hiding behind the data.
- Data Science is the process of using data to understand different things.
- Data Science helps to translate data into a story. The story telling helps in uncovering insights. The insights help in matking decision or strategic choices.

Data Science – Multiple Disciplines



Need of Data Science

- Data deluge, tons of data.
- Powerful algorithms.
- Open software and tools.
- Computational speed, accuracy and cost.
- Data storage in terms of capacity and cost.

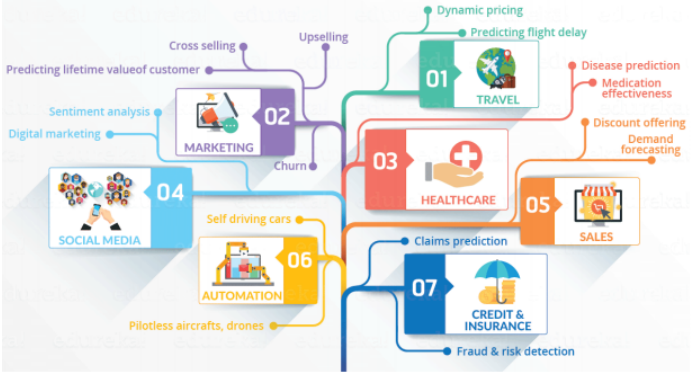
Use cases of Data Science

DataFlair

Applications of Data Science

DataFlair

Applications of Data Science

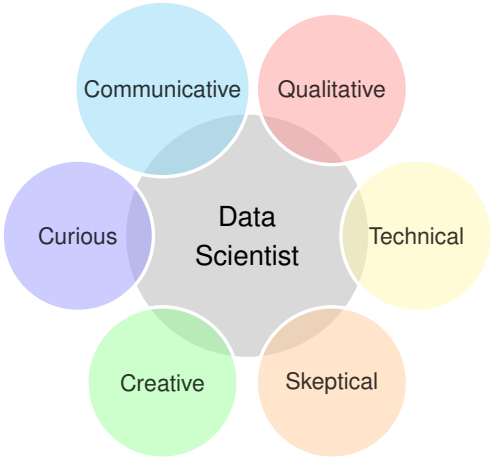


edureka.co

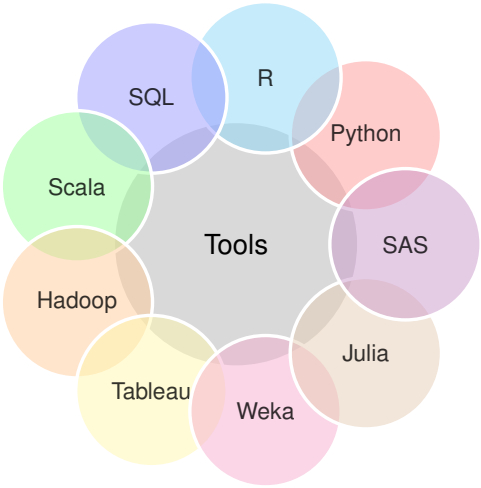
Role of a Data Scientist

1. Reframe business challenges as analytics challenges.
   This is a skill to diagnose the problem, consider the core of a given problem, and determine which kinds of candidate analysis analytical method can be applied to solve it.
2. Design, implement and deploy statistical models and data mining technique on data.
   This activity is mainly the role of data scientist, applying complex or advanced analytical methods to a variety of business problem using data.
3. Develop insights that lead to actionable recommendations.
   Learn how to draw insights out of data and communicate them effectively.

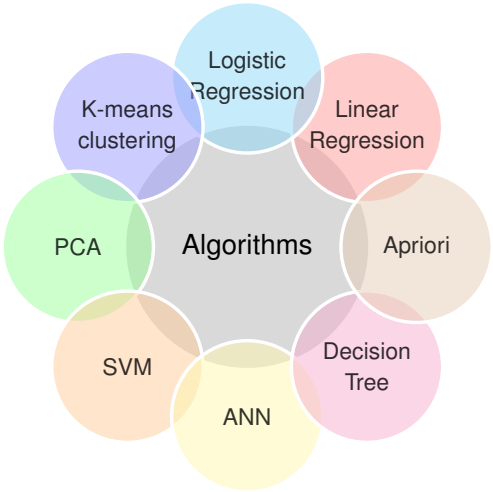
Skills required for a Data Scientist



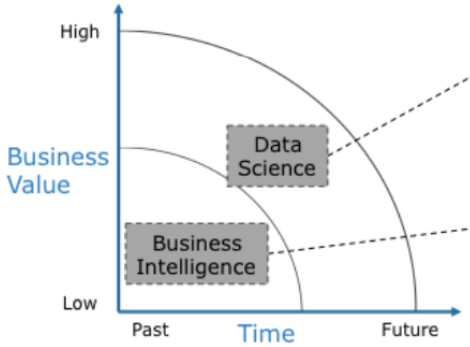
Tools available to a Data Scientist



Algorithms for a Data Scientist



Data Science vs. Business Intelligence



Data Science

- Predictive analysis
  Prescriptive analysis
- Why…? What will…?
  What should I do…?

Business Intelligence

- Descriptive analysis
  Standard reporting
- What happened?

Data Science vs. Business Intelligence

	Data Science	Business Intelligence
Perspective	Looking forward	Looking backward
Analysis	Predictive Explorative	Descriptive Comparative
Data	Same data, New analysis Listens to data Distributed	New Data, Same analysis Speaks for data Warehoused
Scope	Specific to buisness question	Unlimited
Expertise	Data scientist	Business analyst
Deliverable	Insight or story	Table

Data Scientist vs. Business Analyst

Area	BI Analyst	Data Scientist
Focus	Reports, KPIs, trends	Patterns, correlations, models
Process	Static, comparative	Exploratory, experimentation, visual
Data sources	Pre-planned, added slowly	On the fly, as-needed
Transform	Up front, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on load	Schema on query
Analysis	Retrospective, Descriptive	Predictive, Prescriptive

EMC²

Data Science vs. Statistics

	Data Science	Statistics
Type of problem	Semi structured or un-structured	Well structured
Analysis Objective	Need not be well formed	Well formed objective
Type of Analysis	Explorative	Confirmative
Data collection	Data collection is not linked to the objective	Data collected based on the objective
Size of dataset	Large Heterogeneous	Small Homogeneous
Paradigm	Theory and heuristic (deductive & inductive)	Theory based (deductive)

Software Engineering for Data Science

1. For data scientists, software is the generalization of a specific aspect of a data analysis. Software allows for the systematizing and the standardizing of a procedure, so that different people can use it and understand what it's doing, at any given time. Software is the encompasses all required tools into a specific module or procedure that can be repeatedly applied in a variety of settings
2. Software will have an interface, or a set of inputs and a set of outputs that are well understood.
   3 Levels of S/W : Code, Write Function S/W Package or API
3. Application of Skill Sets differ in
   Methodologies, Objectives, Approaches Tools.

Data Science Challenges

1. Complexity of Data Reality
2. Identifying the problem
3. Access to right data – Data quantity
4. Data Cleansing – Data quality - Data Security
5. Granularity, Consistency Availability of Data
6. Lack of domain expertise
7. Cognitive Bias Content Bias

THANK YOU