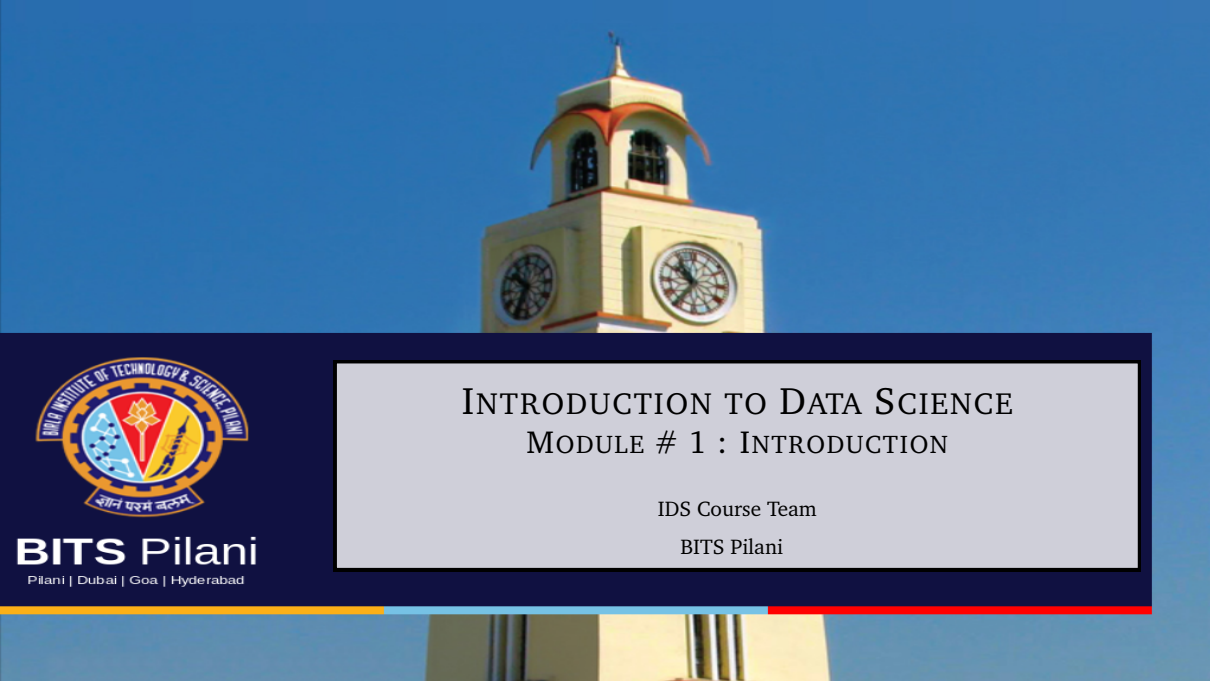


Introduction to Data Science

Module # 1 : Introduction

IDS Course Team

BITS Pilani

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

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

Table of Contents

Course Objectives

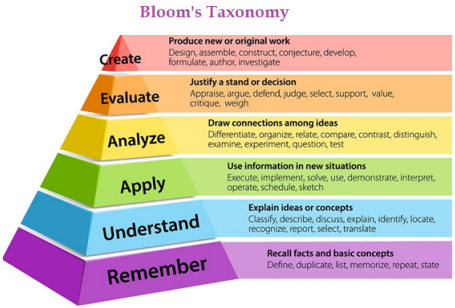
Understand, Apply, Analyze, Evaluate

CO1 The role of Data Science in various scenarios in the real-world of business, industry and government.

CO2 Various roles and stages in a Data Science Project and ethical issues to be considered.

CO3 The processes, tools and technologies for collection and analysis of structured and unstructured data.

CO4 The importance of techniques like data visualization, storytelling with data for the effective presentations of the outcomes with the stakeholders.



Course Structure

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

Text and Reference Books

Text Books

T1	Introducing Data Science by Cielen, Meysman and Ali
T2	Storytelling with Data, A data visualization guide for business professionals, by Cole, Nussbaumer Knaflic; Wiley
T3	Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar

Reference Books

R1	The Art of Data Science by Roger D Peng and Elizabeth Matsui
R2	Ethics and Data Science by DJ Patil, Hilary Mason, Mike Loukides
R3	Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas
R4	KDD, SEMMA and CRISP-DM: A Parallel Overview , Ana Azevedo and M.F. Santos, IADS-DM, 2008

Evaluation Schedule

No	Name	Type	Duration	Weight	Day, Date, Time
EC1	Quiz I	Online	1 hr	5%	19-Dec-21 Release
	Quiz II	Online	1 hr	5%	27-Feb-22 Release
		Best 1 of 2 quizzes will be considered for evaluation.			
	Mini-project	Online	4 weeks	25%	28-Nov-22 Release
EC2	Mid-sem Regular	Online	2.5 hrs	30%	2-Jan-2022 AN
	Mid-sem Makeup	Online	2.5 hrs	30%	22-Jan-2022 AN
EC3	Compre-sem Regular	Online	3 hrs	40%	27-March-2022 AN
	Compre-sem Makeup	Online	3 hrs	40%	9-April-2022 AN

Canvas

Most relevant and up to date info on Canvas

- Handout
- Schedule for Webinar, Quiz, and Assignments.
- Session Slide Deck
- Demo Lab Sheets
- Quiz-I, Quiz-II
- Mini-project

The video recording will be available in Impartus.

Platform / Dataset

- Platform
  - Python / Jupyter Notebook
- Dataset
  - Datasets as we deem appropriate.
- Webinar
  - As per schedule

Table of Contents

Why Data Science?

- ”Data Science is the sexiest job in the 21st century” – IBM.
- Data Science is one of the fastest growing fields in the world.
- According to the U.S. Bureau of Labor Statistics, 11.5 million new jobs will be created by the year 2026.
- Even with COVID-19 situation, and the amount of shortage in talent, there might not be a dip in data science as a career option.

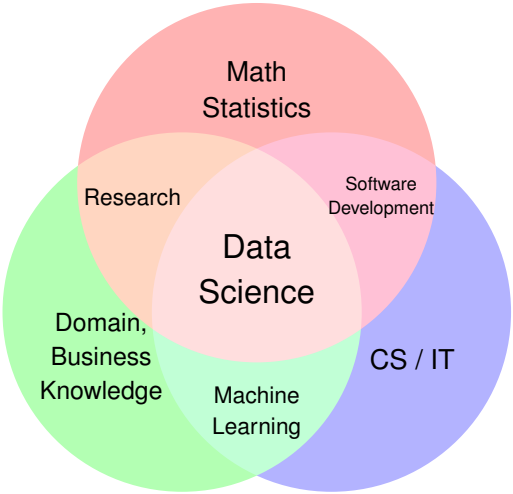
Why Data Science?

- In India, the average salary of a data scientist as of January 2020 is Rs.10L/yr. – Glassdoor, 2020.
- The increase in data science as a career choice in 2020 will also see the rise in its various job roles.
  - Data Engineer
  - Data Administrator
  - Machine Learning Engineer
  - Statistician
  - Data and Analytics Manager

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 helps to translate data into a story. The story telling helps in uncovering insights. The insights help in making decision or strategic choices.
- Data Science is the process of using data to understand different things.
  - Requires a major effort of preparing, cleaning, scrubbing, or standardizing the data.
  - Algorithms are then applied to crunch pre-processed data.
  - This process is iterative and requires analysts' awareness of the best practices.
  - The most important aspect of data science is interpreting the results of the analysis in order to make decisions.

Data Science – Multiple Disciplines



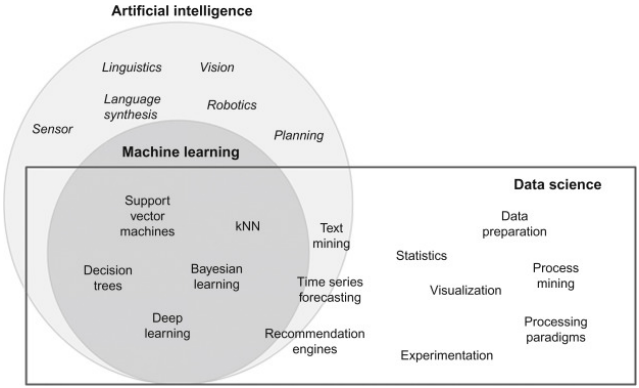
Need of Data Science

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

Data Science, AI and ML

- Artificial Intelligence
  - AI involves making machines capable of mimicking human behavior, particularly cognitive functions like facial recognition, automated driving, sorting mail based on postal code.
- Machine Learning
  - Considered a sub-field of or one of the tools of AI.
  - Involves providing machines with the capability of learning from experience.
  - Experience for machines comes in the form of data.
- Data Science
  - Data science is the application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics to uncover insights from data to enable better decision marking.

Data Science, AI and ML



https://www.sciencedirect.com/topics/physics-and-astronomy/artificial-intelligence

Table of Contents

Use cases of Data Science

DataFlair

Data Science in Facebook

Social Analytics

- Utilizes quantitative research to gain insights about the social interactions of among people.
- Makes use of deep learning, facial recognition, and text analysis.
- In facial recognition, it uses powerful neural networks to classify faces in the photographs.
- In text analysis, it uses “DeepText” to understand people’s interest and aligns photographs with texts.
- It uses deep learning for targeted advertising.
- Using the insights gained from data, it clusters users based on their preferences and provides them with the advertisements that appeal to them.

Data Science in Amazon

Improving E-Commerce Experience

- Personalized recommendation
  - Ppredictive analytics (a personalized recommender system) to increase customer satisfaction.
  - Purchase history of customers, other customer suggestions, and user ratings are analyzed to recommend products.
- Anticipatory shipping model
  - Predict the products that are most likely to be purchased by its users.
  - Analyzes pattern of customer purchases and keeps products in the nearest warehouse which the customers may utilize in the future.

Data Science in Amazon – contd...

Improving E-Commerce Experience

- Price discounts
  - Using parameters such as the user activity, order history, prices offered by the competitors, product availability, etc., Amazon provides discounts on popular items and earns profits on less popular items.
- Fraud Detection
  - Detect fraud sellers and fraudulent purchases.
- Improving Packaging Efficiency
  - Optimize packaging of products in warehouses and increases efficiency of packaging lines through the data collected from the workers.

Data Science in Uber

Improving Rider Experience

- Uber maintains large database of drivers, customers, and several other records.
- Makes extensive use of Big Data and crowdsourcing to derive insights and provide best services to its customers.
- Dynamic pricing
  - Use of big Data and data science to calculate fares based on specific parameters.
  - Uber matches customer profile with the most suitable driver and charges them based on the time it takes to cover the distance rather than the distance itself.
  - The time of travel is calculated using algorithms that make use of data related to traffic density and weather conditions.
  - When the demand is higher (more riders) than supply (less drivers), the price of the ride goes up.

Data Science in Bank of America

Improving Customer Experience

- Erica – a virtual financial assistant (BoA)
  - Erica serves as a customer advisor to over 45 million users around the world.
  - Erica makes use of Speech Recognition to take customer inputs.
- Fraud detection
  - Uses data science and predictive analytics to detect frauds in payments, insurance, credit cards, and customer information.
- Risk modeling
  - Use data science for risk modeling to regulate financial activities.
- Customer segmentation
  - Segment their customers in the high-value and low-value segments.
  - Data scientists makes use of clustering, logistic regression, decision trees to help the banks to understand the Customer Lifetime Value (CLV) and take group them in the appropriate segments.

Data Science in Airbnb

Improving Customer Experience

- Providing better search results
  - Uses big data of customer and host information, homestays and lodge records, and website traffic.
  - Uses data science to provide better search results to its customers and find compatible hosts.
- Detecting bounce rates
  - Use of demographic analytics to analyze bounce rates from their websites.
- Providing ideal lodgings and localities
  - Uses knowledge graphs where the user's preferences are matc hed with the various parameters to provide ideal lodgings and localities.

Data Science in Spotify

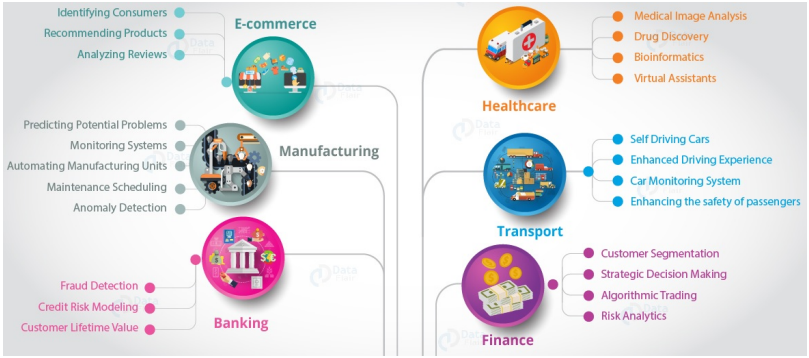
Improving Customer Experience and recommendation

- Providing better music streaming experience
  - Provide personalized music recommendations.
  - Uses over 600 GBs of daily data generated by the users to build its algorithms to boost user experience.
- Improving experience for artists and managers
  - Spotify for Artists application allows the artists and managers to analyze their streams, fan approval and the hits they are generating through Spotify's playlists.

Data Science in Spotify... contd..

- Spotify uses data science to gain insights about which universities had the highest percentage of party playlists and which ones spent the most time on it.
- ”Spotify Insights” publishes information about the ongoing trends in the music.
- Spotify’s Niland, an API based product, uses machine learning to provide better searches and recommendations to its users.
- Spotify analyzes listening habits of its users to predict the Grammy Award Winners.

Applications of Data Science



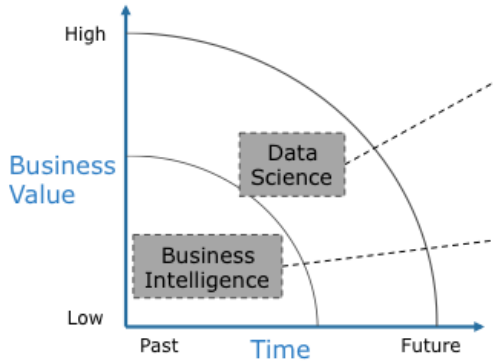
DataFlair

Applications of Data Science

edureka.co

Table of Contents

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 business question	Unlimited
Expertise	Data scientist	Business analyst
Deliverable	Insight or story	Table or report
Applicability	Future, correction for influences	Historic, confounding factors

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 unstructured	Well structured
Inference model	Explicit inference	No inference
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)

Table of Contents

Who is a Data Scientist?

- A data scientist is someone who extracts insights from messy data.
- A data scientist is responsible for guiding a data science project from start to finish.
- Success in a data science project comes not just from an one tool, but from having quantifiable goals, good methodology, cross-discipline interactions, and a repeatable workflow.

Role of a Data Scientist

- 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.
- 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.
- Develop insights that lead to actionable recommendations.
  Learn how to draw insights out of data and communicate them effectively.

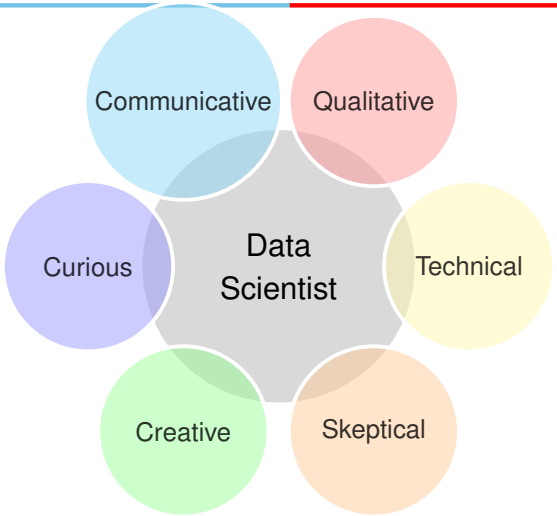
Responsibilities of a Data Scientist

- Data scientists work closely with business stakeholders to understand their goals and determine how data can be used to achieve those goals.
- They design data modeling processes, create algorithms and predictive models to extract the data the business needs, then help analyze the data and share insights with peers.

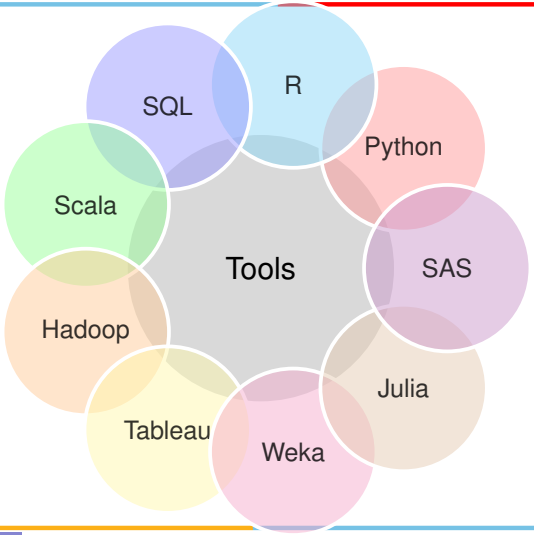
Job Titles of a Data Scientist

- Data scientists – Design data modeling processes to create algorithms and predictive models and perform custom analysis.
- Data analysts – Manipulate large data sets and use them to identify trends and reach meaningful conclusions to inform strategic business decisions.
- Data engineers – Clean, aggregate, and organize data from disparate sources and transfer it to data warehouses.
- Business intelligence specialists – Identify trends in data sets.
- Data architects – Design, create, and manage an organization's data architecture.

Skills required for a Data Scientist



Tools available to a Data Scientist



Algorithms for a Data Scientist

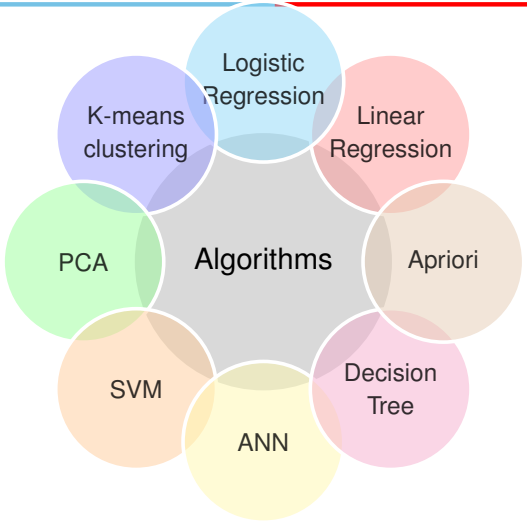
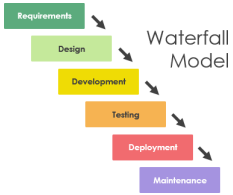


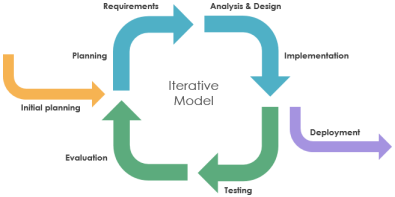
Table of Contents

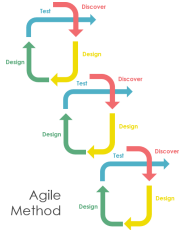
Software Engineering

In general,

- Software engineering is an engineering discipline that is concerned with all aspects of software production.
- Software includes computer programs, all associated documentation, and configuration data that are needed for software to work correctly.
- Waterfall model, Iterative models, Agile models







Data Science Process

Data Science vs. Software Engineering

Software Engineering	Data Science
Software engineering focuses on creating software that serves a specific purpose.	Data science involves analyzing huge amounts of data, with some aspects of programming and development.
Uses a methodology involving various phases beginning from requirements specification through software deployment into production.	Uses a methodology involving various phases beginning from requirements specification through model deployment to better decision making.

Data Science vs. Software Engineering

Data Science	Software Engineering
Involves collecting and analyzing data	Concerned with creating useful applications
Data scientists utilize the ETL (Extract, Tranform, Load) process	Software engineers use the SDLC process
More process-oriented	Uses frameworks like Waterfall, Agile, and Spiral
Data scientists use tools like Amazon S3, MongoDB, Hadoop, and MySQL	Software engineers use tools like Rails, Django, Flask, and Vue.js
Skills include machine learning, statistics, and data visualization	Skills are focused on coding languages

Table of Contents

Data Science Challenges

Data science challenges can be categorized as:

- Data related
- Organization related
- Technology related
- People related
- Skill related

Challenges in Data Science

- Complexity of Data Reality
- Identifying the problem
- Access to right data – Data quantity
- Data Cleansing – Data quality - Data Security
- Granularity, Consistency Availability of Data
- Lack of domain expertise
- Cognitive Bias
- Content and Source Bias

Cognitive Bias

- Cognitive Biases are the distortions of reality because of the lens through which we view the world.
- Each of us sees things differently based on our preconceptions, past experiences, cultural, environmental, and social factors. This doesn't necessarily mean that the way we think or feel about something is truly representative of reality.

References:

- Introducing Data Science by Cielen, Meysman and Ali
- The Art of Data Science by Roger D Peng and Elizabeth Matsui
- https://data-flair.training/blogs/data-science-use-cases/
- https://www.northeastern.edu/graduate/blog/what-does-a-data-scientist-do/
- https://www.visual-paradigm.com/guide/software-development-process/what-is-a-software-process-model/

THANK YOU