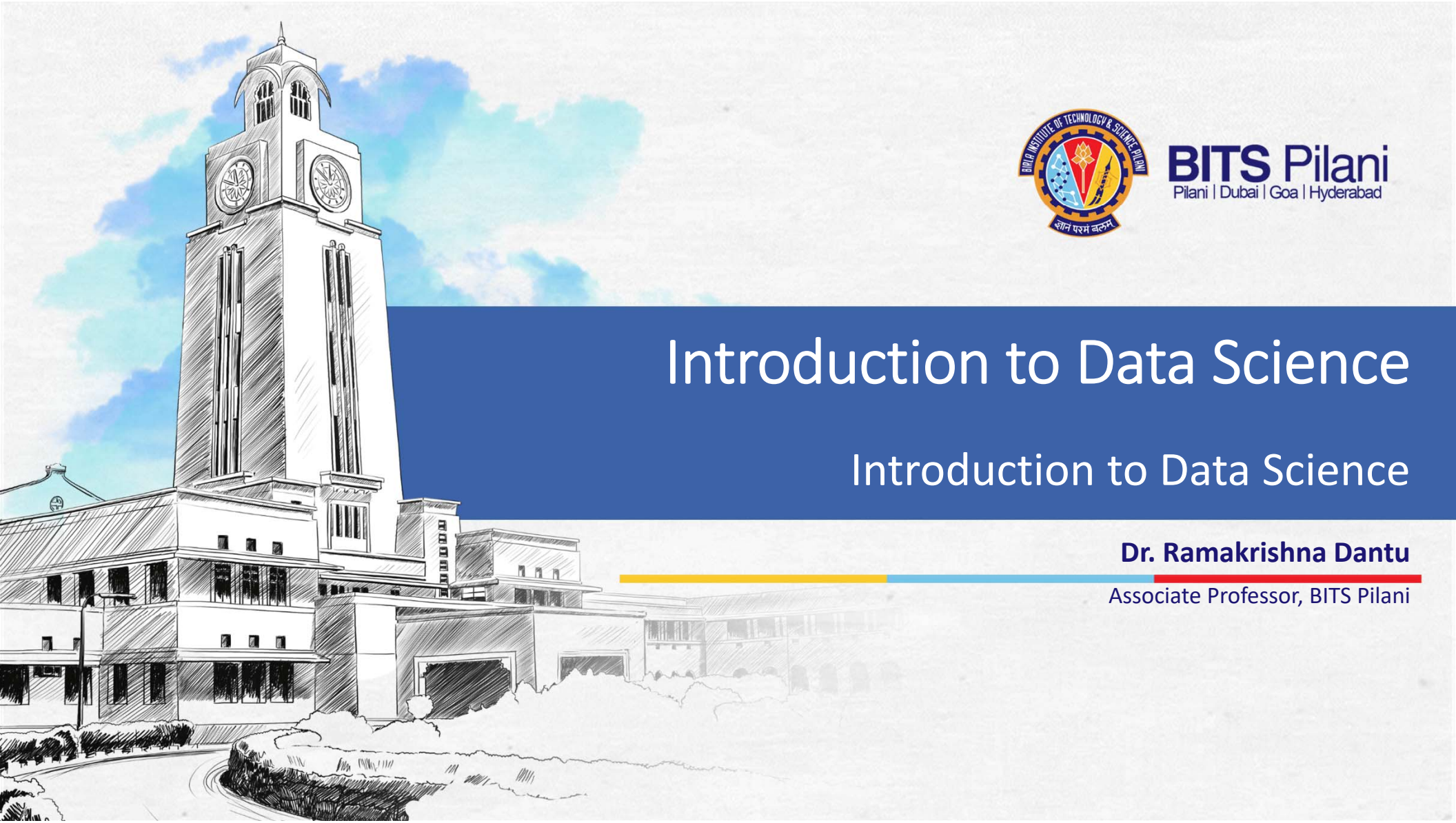


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Introduction to Data Science

Dr. Ramakrishna Dantu
Associate Professor, BITS Pilani

innovate achieve lead

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Introduction to Data Science

innovate achieve lead

Module Topics

- Fundamentals of Data Science
  - Why Data Science
  - Defining Data Science
  - Data Science Process
- Real World applications
- Data Science vs BI
- Data Science vs Statistics
- Roles and responsibilities of a Data Scientist
- Software Engineering for Data Science
- Data Scientists Toolbox
- Data Science Challenges

Why Data Science?

Why Data Science?

- Let's look into this question from the following perspectives:
  - Data Science as a field
  - Various data science job roles
  - Market revenue
  - Skills

Source: https://analyticsindiamag.com/why-a-career-in-data-science-should-be-your-prime-focus-in-2020/

Why Data Science?

- Data Science as a field
  - "Data Science is the sexiest job in the 21st century"
    - --IBM
  - Data Science is one of the fastest growing fields in the world
  - In 2019, 2.9 million data science job openings were required globally
  - 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

Source: https://analyticsindiamag.com/why-a-career-in-data-science-should-be-your-prime-focus-in-2020/

Why Data Science?

- The Hottest Job Roles and Trends In Data Science 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
  - In India, the average salary of a data scientist as of January 2020 is ₹10L/yr, which is pretty attractive (Glassdoor, 2020)

Source: https://analyticsindiamag.com/why-a-career-in-data-science-should-be-your-prime-focus-in-2020/

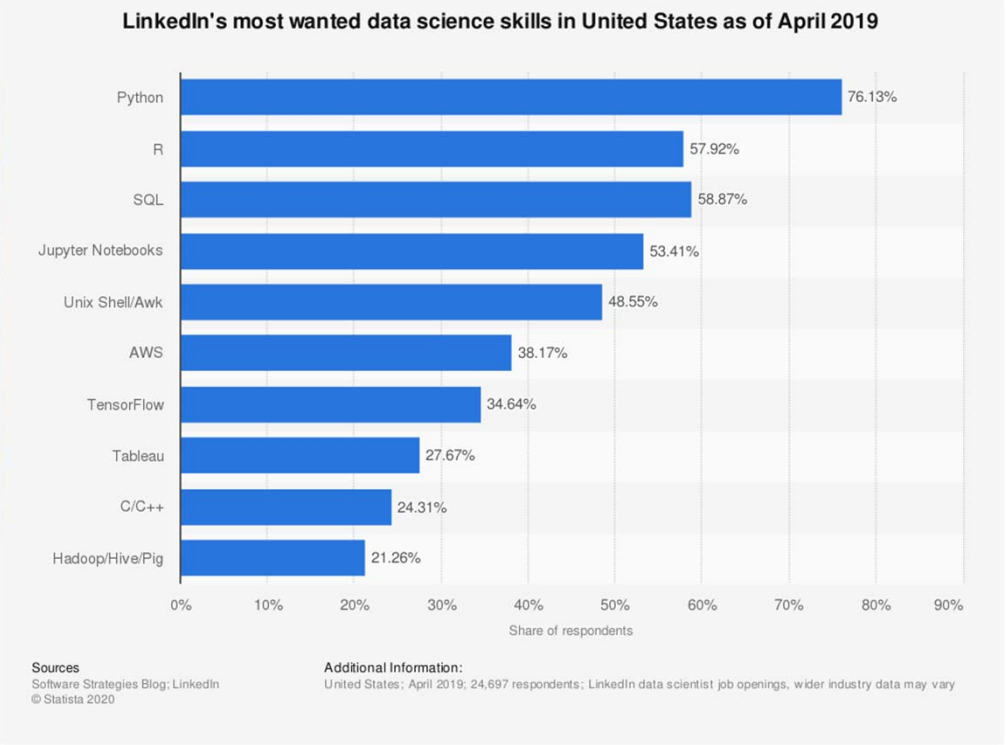
Why Data Science?

- Market Revenue
  - In recent years, with the amount of shift in analytics and data science, the market revenue has increased
  - Analytics, data science, and big data industry in India generated about
    - $2.71 billion annually in 2018, and the revenue grew to
    - $3.03 billion in 2019
  - This 2019 figure, is expected to double by 2025 nearly.

Source: https://analyticsindiamag.com/why-a-career-in-data-science-should-be-your-prime-focus-in-2020/

Most wanted Data Science Skills in 2019

- Survey Details
  - LinkedIn data scientist job openings
  - Survey conducted in the US in April 2019
  - 24,697 respondents
  - 76.13 percent of data scientist job openings on LinkedIn required a knowledge of the programming language Python
  - Wider industry data may vary



Source: https://softwarestrategiesblog.com/2019/06/13/how-to-get-your-data-scientist-career-started/

Data Science Defined

Data Science Defined

- There is no consistent definition to what constitutes data science
- Data Science is a study of data.
- Data Science is an art of uncovering useful patterns, connections, 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 in turn help in making decisions or strategic choices
- Data Science involves extracting meaningful insights from any data
  - 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
  - Automation of tasks allows us focus on the most important aspect of data science:
    - Interpreting the results of the analysis in order to make decisions

Data Science Defined

- Data science is an inter-disciplinary practice that draws from
  - Data engineering, statistics, data mining, machine learning, and predictive analytics
- Similar to operations research, data science focuses on…
  - Implementing data-driven models and managing their outcomes
- To uncover nuggets of wisdom and actionable insights, Data science combines the
  - Data-driven approach of statistical data analysis
  - Computational power of the computer
  - Programming acumen of a computer scientist
  - Domain-specific business intelligence

Source: Data Science using Python and R – Chantal Larose & Daniel Larose

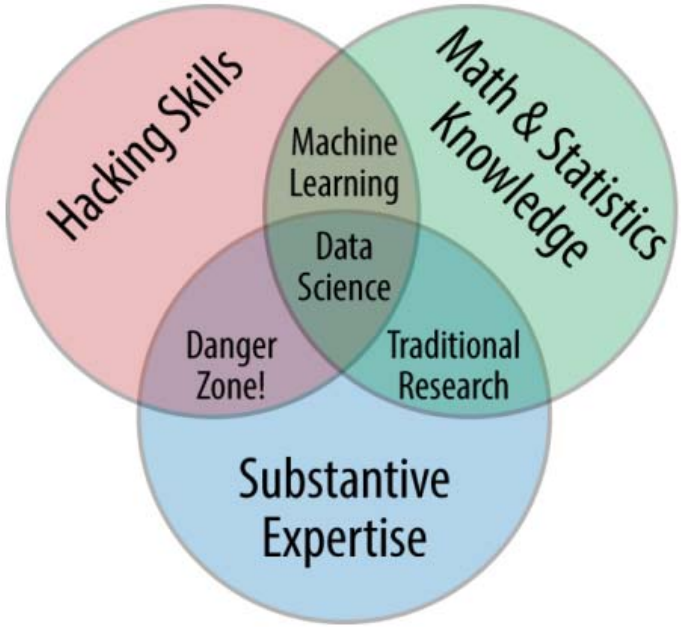
Data Science Defined – Drew Conway's view

- Drew Conway is the CEO and founder of Alluvium
- He is a leading expert in the application of computational methods to social and behavioral problems at large-scale
- Drew has been writing and speaking about the role of data — and the discipline of data science — in industry, government, and academia for several years.
- Drew has advised and consulted companies across many industries:
  - Ranging from fledgling start-ups to Fortune 100 companies
  - Academic institutions, and
  - Government agencies at all levels
- Drew started his career in counter-terrorism as a computational social scientist in the U.S. intelligence community

Source: http://drewconway.com

Data Science Defined - Drew Conway's view

- Data science comprises three distinct and overlapping skills:
  - *Statistics* – for modeling and summarizing datasets
  - *Computer science* – for designing and using algorithms to effectively store, process, and visualize the data
  - *Domain expertise* – necessary both to formulate the right questions and to put their answers in context

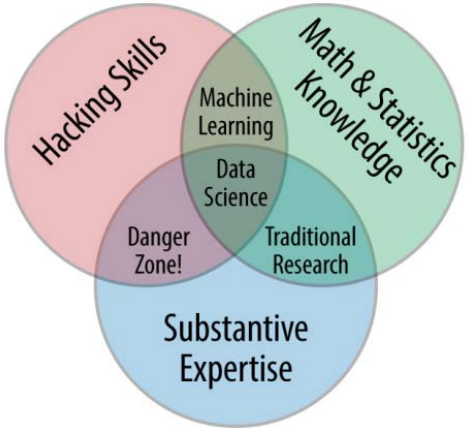


Drew Conway's Data Science Venn Diagram

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science Defined - Drew Conway's view

- Danger Zone
  - Here, people "know enough to be dangerous," and is the most problematic area of the diagram
  - People are perfectly capable of extracting and structuring data in the field they know quite a bit about
  - People even know enough R to run a linear regression and report the coefficients
    - But they lack any understanding of what those coefficients mean
  - It is from this part of the diagram that the phrase "lies, damned lies, and statistics" emanates
    - Because either through ignorance or malice this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created
  - As such, the danger zone is sparsely populated, however, it does not take many to produce a lot of damage.

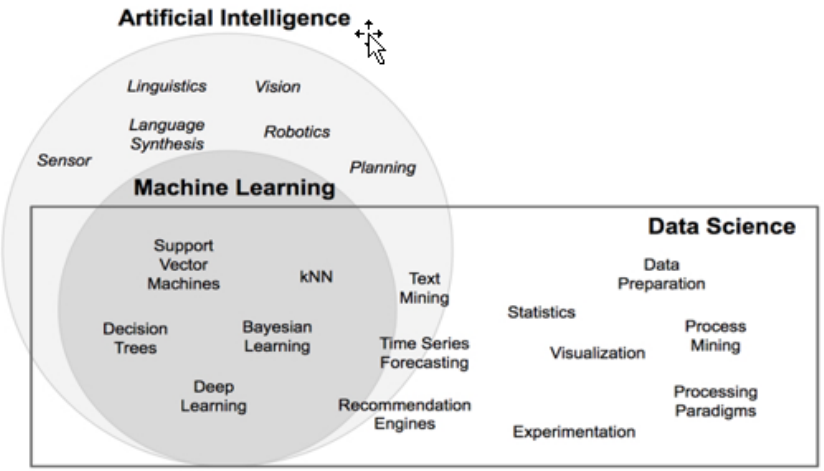


Drew Conway's Data Science Venn Diagram

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Artificial Intelligence, Machine Learning, and Data Science

- Artificial Intelligence
  - AI involves making machines capable of mimicking human behavior, particularly cognitive functions:
    - For e.g., 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 that is used to teach machines is called training data
- Data Science
  - Data science is the application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics



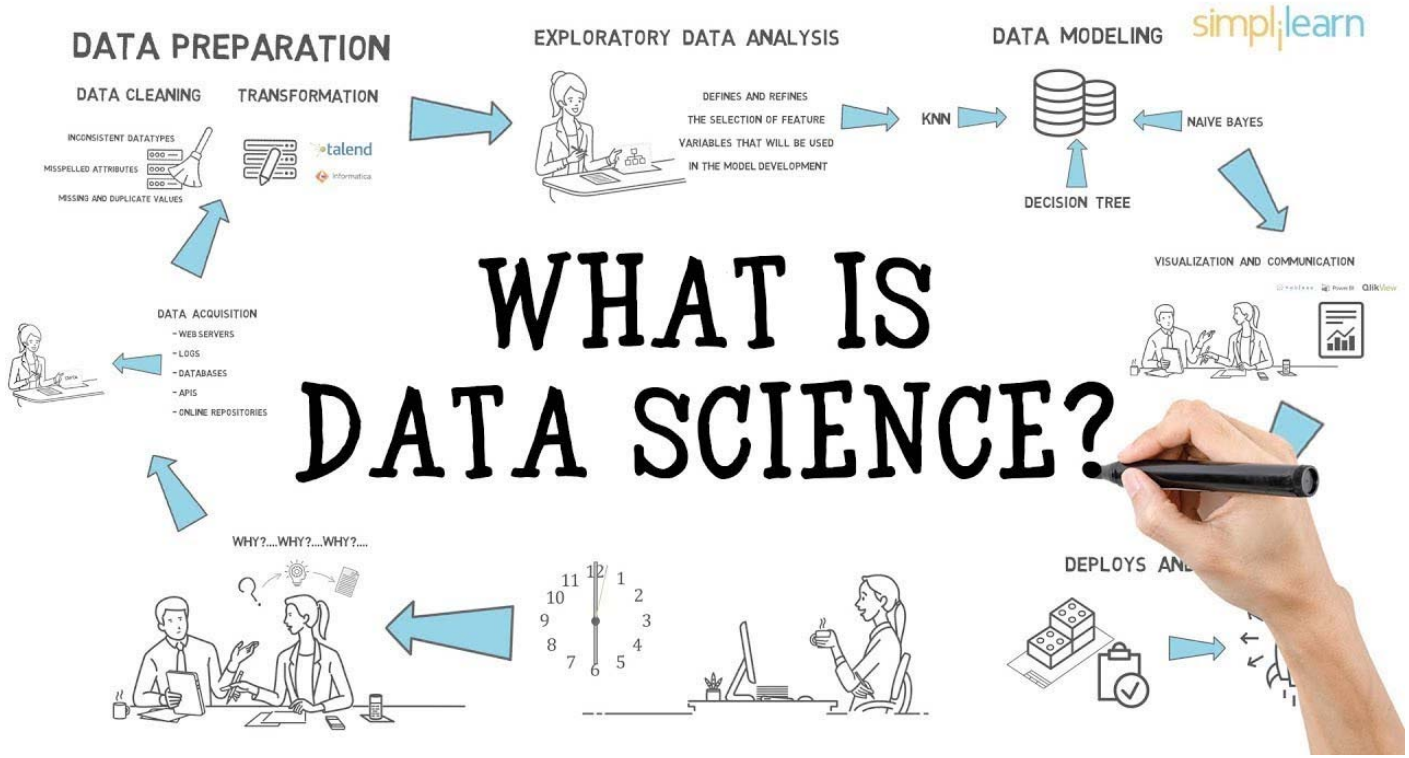
Source: Data Science, 2nd Edition by Bala Deshpande; Vijay Kotu

Data Science Process

Fundamentals of Data Science

innovate achieve lead

Various Stages of Data Science Process

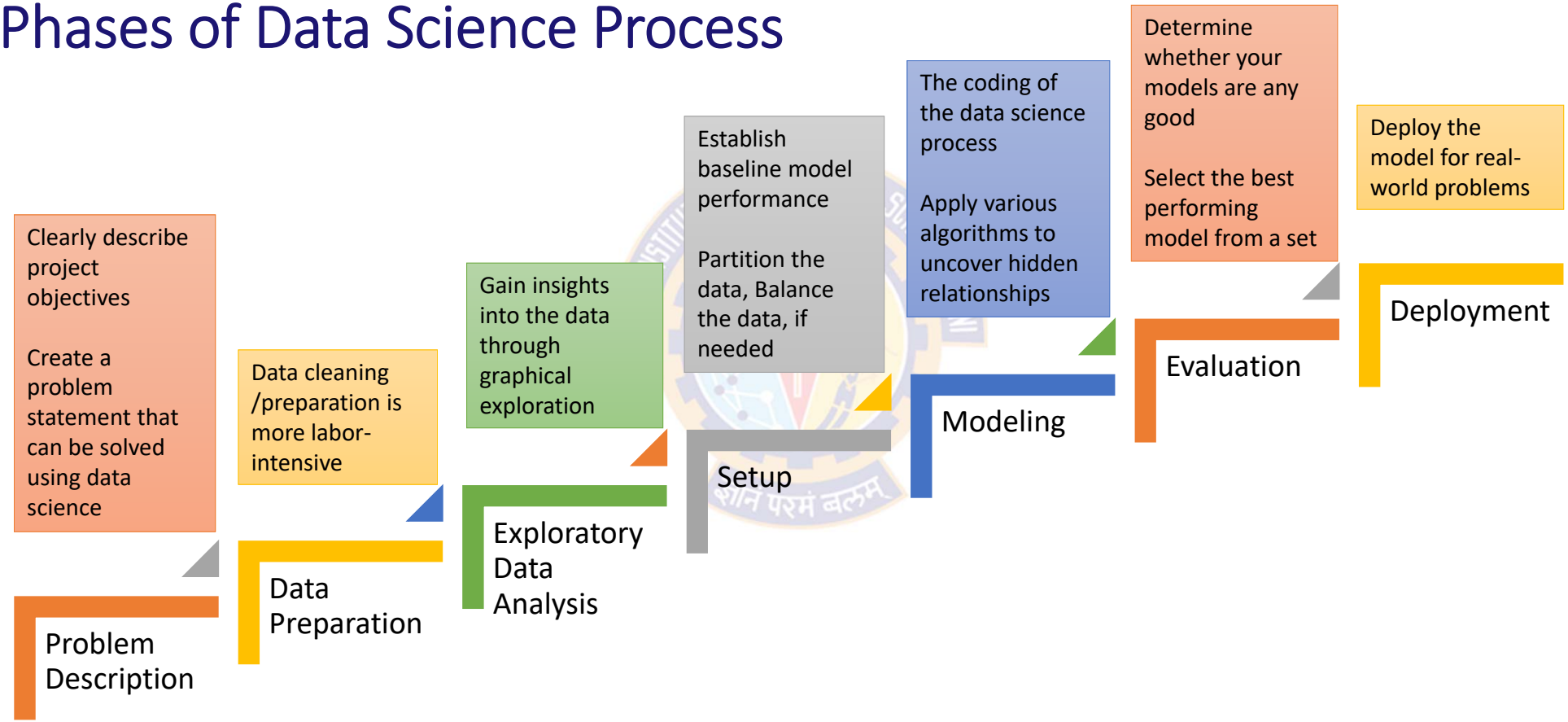


Source: Simplilearn - https://www.youtube.com/watch?v=X3paOmcrTjQ

Fundamentals of Data Science

innovate achieve lead

Phases of Data Science Process



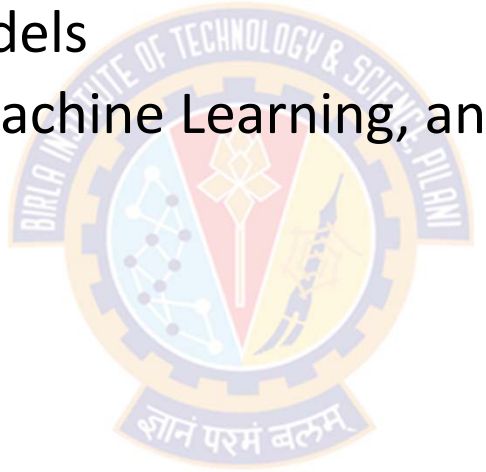
Source: Data Science Using Python & R by Chantal Larose and Daniel Larose

What is Data Science?

innovate achieve lead

Key features and motivations of Data Science

- Extracting Meaningful Patterns
- Building Representative Models
- Combination of Statistics, Machine Learning, and Computing
- Learning Algorithms
- Associated Fields
  - Descriptive Statistics
  - Exploratory Visualization
  - Dimensional Slicing
  - Hypothesis Testing
  - Data Engineering
  - Business Intelligence



Source: Data Science, 2nd Edition, by Bala Deshpande; Vijay Kotu

Fundamentals of Data Science

Data Science Tasks and Examples

Tasks	Description	Algorithms	Examples
Classification	• Predict if a data point belongs to one of the predefined classes. • The prediction will be based on learning from a known dataset	• Decision trees • neural networks • Bayesian models • induction rules • k-nearest neighbors	• Assigning voters into known buckets by political parties, e.g., soccer moms • Bucketing new customers into one of the known customer groups
Regression	• Predict the numeric target label of a data point. • The prediction will be based on learning from a known database	• Linear regression • logistic regression	• Predicting the unemployment rate for the next year • Estimating insurance premium
Anomaly Detection	• Predict if a data point is an outlier compared to other data points in the dataset	• Distance-based • density-based local outlier factor (LOF)	• Detecting fraudulent credit card transactions and network intrusion
Association analysis	• Identify relationships within an item set based on transaction data	• FP-growth algorithm • a priori algorithm	• Finding cross-selling opportunities for a retailer based on transaction purchase history

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors.

Source: Data Science, 2nd Edition, by Bala Deshpande; Vijay Kotu

Data Science Tasks and Examples

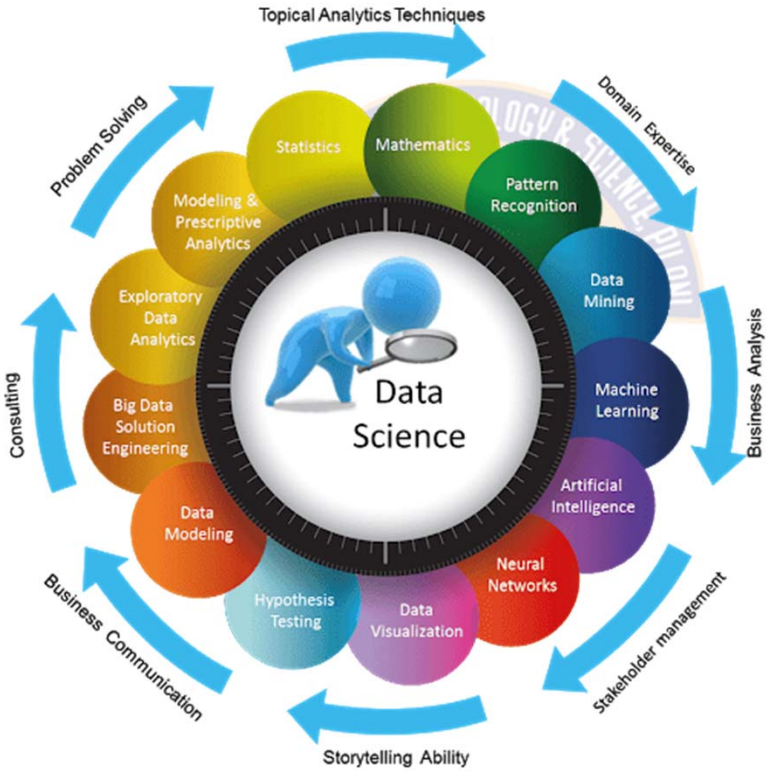
Tasks	Description	Algorithms	Examples
Time series forecasting	• Predict the value of the target variable for a future timeframe based on historical values	• Exponential smoothing • Autoregressive integrated moving average (ARIMA) • Regression	• Sales forecasting • Production forecasting • Virtually any growth phenomenon that needs to be extrapolated
Clustering	• Identify natural clusters within the dataset based on inherit properties within the dataset	• k-Means • Density-based clustering (e.g., DBSCAN)	• Finding customer segments in a company based on transaction, web, and customer call data
Recommendation engines	• Predict the preference of an item for a user	• Collaborative filtering • Content-based filtering • Hybrid recommenders	• Finding the top recommended movies for a user

Source: Data Science, 2nd Edition, by Bala Deshpande; Vijay Kotu

Fundamentals of Data Science

innovate achieve lead

Disciplines involved in Data Science



Source: http://www.verival.com/DataScience

Real World
Data Science Applications

Real-World Applications

innovate achieve lead

Data Science Use Cases

Source: https://data-flair.training/blogs/data-science-use-cases/

Real-World Applications

innovate achieve lead

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 its own understanding engine called “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

Source: https://data-flair.training/blogs/data-science-use-cases/

Real-World Applications

innovate achieve lead

Amazon

- Improving E-Commerce Experience
  - Personalized recommendation
    - Amazon heavily relies on predictive analytics (a personalized recommender system) to increase customer satisfaction
    - Amazon analyzes the purchase history of customers, other customer suggestions, and user ratings to recommend products
  - Anticipatory shipping model
    - Uses big data for predicting 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
  - 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
    - Amazon has its own novel ways and algorithms to detect fraud sellers and fraudulent purchases
  - Improving Packaging Efficiency
    - Amazon optimizes packaging of products in warehouses and increases efficiency of packaging lines through the data collected from the workers

Source: https://data-flair.training/blogs/data-science-use-cases/

Real-World Applications

innovate achieve lead

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
    - The concept is rooted in Big Data and data science to calculate fares based on the parameters
    - Uber matches customer profile with the most suitable driver and charges you 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
    - When the demand for Uber rides is less, then Uber charges lower rate

Source: https://data-flair.training/blogs/data-science-use-cases/

Real-World Applications

innovate achieve lead

Bank of America and other Banks

- Improving Customer Experience
  - Erica – a virtual financial assistant (BoA)
    - Considered as the world's finest innovation in finance domain, Erica serves as a customer advisor to over 45 million users around the world
    - Erica makes use of Speech Recognition to take customer inputs, which is a technological advancement in the field of Data Science.
  - Fraud detection
    - Uses data science and predictive analytics to detect frauds in payments, insurance, credit cards, and customer information
    - Banks employ data scientists to use their quantitative knowledge where they apply algorithms like association, clustering, forecasting, and classification.
  - Risk modeling
    - Banks use data science for risk modeling to regulate financial activities
  - Customer segmentation
    - Using various data-mining techniques, banks are able to 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

Source: https://data-flair.training/blogs/data-science-use-cases/

Real-World Applications

innovate achieve lead

Airbnb

- Providing better search results
  - Airbnb uses a massive big data of customer and host information, homestays and lodge records, as well as website traffic
  - Uses data science to provide better search results to its customers and find compatible hosts
- Detecting bounce rates
  - Airbnb makes use of demographic analytics to analyze bounce rates from their websites
  - In 2014, Airbnb found that users in certain countries would click the neighborhood link, browse the page and photos and not make any booking
  - To mitigate this issue, Airbnb released a different version in those countries and replaced neighborhood links with the top travel destinations
    - This resulted in a 10% improvement in the lift rate for those users
- Providing ideal lodgings and localities
  - Airbnb uses knowledge graphs where the user's preferences are matched with the various parameters to provide ideal lodgings and localities

Source: https://data-flair.training/blogs/data-science-use-cases/

Real-World Applications

innovate achieve lead

Spotify

- Providing better music streaming experience
  - Spotify uses Data Science and leverages big data for providing personalized music recommendations
  - It has over 100 million users and uses a massive amount of big data
  - 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
- Others
  - 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
    - In the year 2013, Spotify made 4 correct predictions out of 6

Source: https://data-flair.training/blogs/data-science-use-cases/

Real-World Applications

innovate achieve lead

Real World Examples

- Anomaly detection
  - Fraud, disease, crime, etc.
- Automation and decision-making
  - Background checks, credit worthiness, etc.
- Classifications
  - In an email server, this could mean classifying emails as “important” or “junk”
- Forecasting
  - Sales, revenue and customer retention
- Pattern detection
  - Weather patterns, financial market patterns, etc.
- Recognition
  - Facial, voice, text, etc.
- Recommendations
  - Based on learned preferences, recommendation engines can refer you to movies, restaurants and books you may like

Real-World Applications

innovate achieve lead

Real World Examples

- Sales and Marketing
  - Google AdSense collects data from internet users so relevant commercial messages can be matched to the person browsing the internet
  - MaxPoint (http://maxpoint.com/us) is another example of real-time personalized advertising
- Customer Relationship Management
  - Commercial companies in almost every industry use data science to gain insights into their customers, processes, staff, completion, and products
  - Many companies use data science to offer customers a better user experience, as well as to cross-sell, up-sell, and personalize their offerings

Source: Introducing Data Science - Cielen, Meysman and Ali

Real World Applications

- Human Resources
  - Human resource professionals use people analytics and text mining to:
    - Screen candidates, monitor the mood of employees, and study informal networks among coworkers
  - People analytics is the central theme in the book
    - *Moneyball: The Art of Winning an Unfair Game*
    - The traditional scouting process for American baseball was random, and replacing it with correlated signals changed everything
    - Relying on statistics allowed them to hire the right players and pit them against the opponents provided the biggest advantage

Real World Applications

- Finance
  - Financial institutions use data science to:
    - Predict stock markets, determine the risk of lending money, and learn how to attract new clients for their services
  - As of 2016, at least 50% of trades worldwide are performed automatically by machines based on algorithms developed by *quants* with the help of big data and data science techniques
    - Data scientists who work on trading algorithms are often called as quants

Source: Introducing Data Science - Cielen, Meysman and Ali

Real World Applications

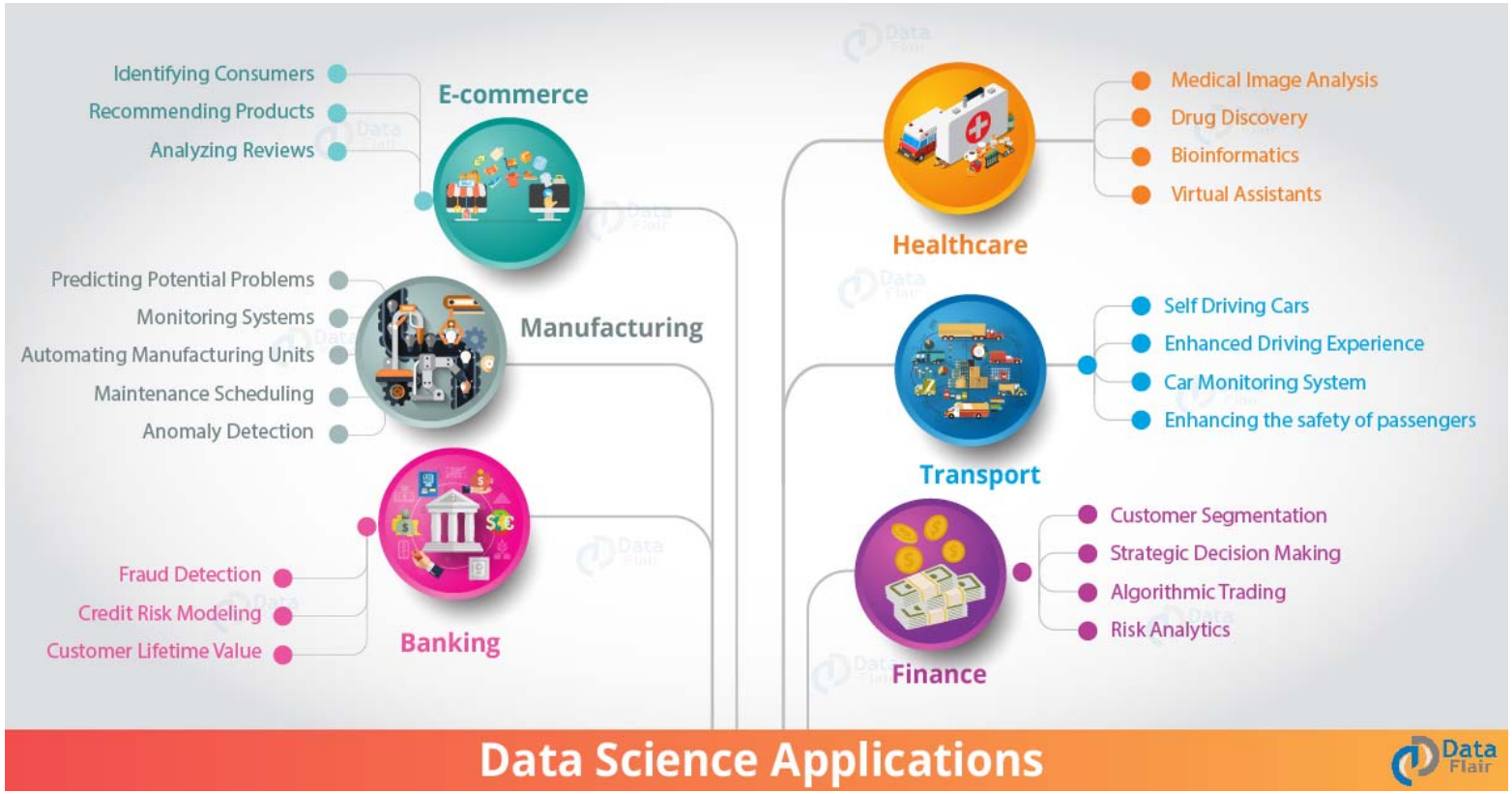
- Non-Governmental Agencies
  - Nongovernmental organizations (NGOs) are also bigtime into data science
  - They use it to raise money and defend their causes
  - The World Wildlife Fund (WWF), for instance, employs data scientists to increase the effectiveness of their fundraising efforts
  - Many data scientists devote part of their time to helping NGOs, because NGOs often lack the resources to collect data and employ data scientists
  - DataKind is one such data scientist group that devotes its time to the benefit of mankind

Source: Introducing Data Science - Cielen, Meysman and Ali

Real World Applications

- Education
  - Universities use data science in their research but also to enhance the study experience of their students
  - The rise of massive open online courses (MOOC) produces a lot of data, which allows universities to study how this type of learning can complement traditional classes
  - MOOCs are an invaluable asset if you want to become a data scientist and big data professional:
    - Coursera, Udacity, and edX
  - The big data and data science landscape changes quickly, and MOOCs allow you to stay up to date by following courses from top universities

Real-World Applications



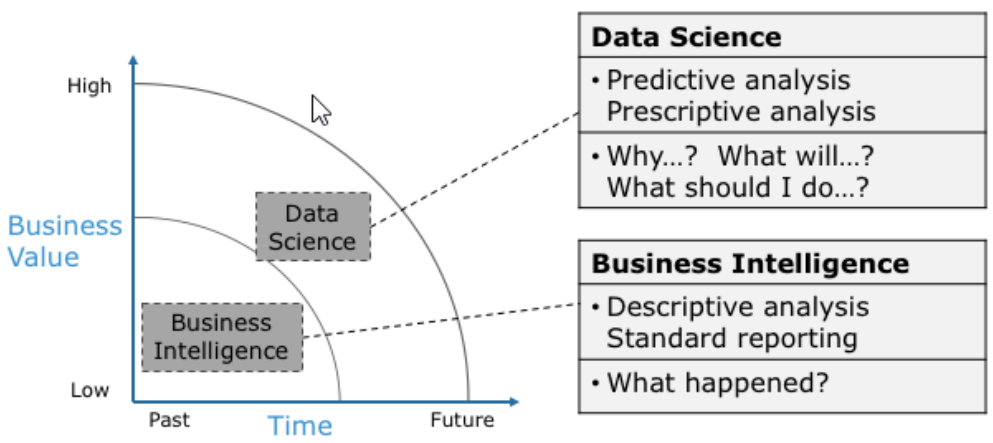
Source: https://data-flair.training/blogs/data-science-applications/

Data Science Vs. Business Intelligence

Data Science Vs. Business Intelligence

innovate achieve lead

Comparing Data Science and BI



- Business Intelligence
  - What happened last quarter?
  - How many units sold?
  - Where is the problem?
  - In which situations?
- Data Science
  - Why is this happening?
  - What if…
  - What would be an optimal scenario for our business?
  - What will happen next?
  - What if these trends continue?

Source: https://www.datasciencecentral.com/profiles/blogs/updated-difference-between-business-intelligence-and-data-science

Data Science Vs. Business Intelligence

Comparing Data Science and BI

	Business Intelligence	Data Science
Perspective	Past and Present	Forward Looking
Involves Statistics	Yes	Yes
Analysis	Descriptive Comparative	Explorative Diagnostic Predictive
Data	Usually structured (SQL) Data Warehouse New Data, Same analysis	Less Structured (Logs, Cloud data, Text, noSQL, SQL) Distributed Data Same Data, New analysis
Scope	General	Specific to business questions
Tools	Statistics, Visualization	Statistics, Machine Learning, Graph Analysis, NLP
Expertise	Business Analyst	Data Scientist
Deliverable	Tabular, Charts	Visualization, Insight, Story

Data Science Vs. Business Intelligence

innovate achieve lead

BI Analyst and Data Scientist Characteristics

	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
Transformation	Up front, Carefully planned	In-database, On-demand, Enrichment
Data Quality	Single version of truth	"Good enough", probabilities
Analytics	Retrospective, Descriptive	Predictive, Prescriptive

Table shows the different attitudinal approaches for each

Source: https://www.datasciencecentral.com/profiles/blogs/updated-difference-between-business-intelligence-and-data-science

Data Science Vs. Business Intelligence

Comparing Data Science and Statistics

	Data Science	Statistics
Type of problem	Semi-structured or Unstructured	Well Structured
Analysis objective	Need not be well defined	Well formed objective
Type of analysis	Explorative	Confirmative
Data collection	Data collection is not linked to the objective	Data collection based on the objective
Size of the dataset	Large Heterogeneous	Small Homogenous
Paradigm	Theory & Heuristic (Deductive & Inductive)	Theory based (Deductive)

Source: https://www.datasciencecentral.com

Roles and Responsibilities of a Data Scientist

R & R of a Data Scientist

innovate achieve lead

Who is a Data Scientist?

- One source writes this…
  - In fact, some data scientists are — for all practical purposes — statisticians, while others are pretty much indistinguishable from software engineers
  - Some are machine-learning experts, while others couldn't machine-learn their way out of kindergarten
  - Some are PhDs with impressive publication records, while others have never read an academic paper (shame on them, though)
  - No matter how you define data science, you'll find practitioners for whom the definition is totally, absolutely wrong
- Nonetheless, 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

Source: Data Science from Scratch, by Joel Grus

R & R of a Data Scientist

innovate achieve lead

Who is a Data Scientist?

- Skills Required for a Data Scientist
  - Math
  - Statistics
  - Computer science
  - Machine learning
  - Domain expertise
  - Data visualization
  - Communication and presentation skills

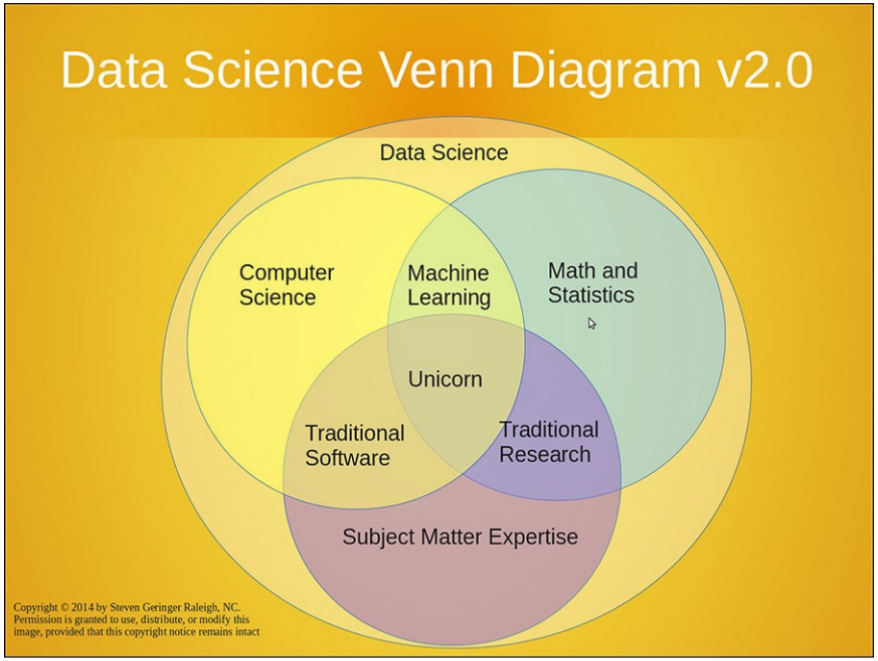


Figure 1.2 Making of a data scientist in a Venn diagram

Image Source: Getting Started with Data Science – Murtaza Haider
Source: Doing Data Science – O'Reilly – Cathey O'Neil & Rachel Schutt

R & R of a Data Scientist

innovate achieve lead

Data Scientist Responsibilities

- 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
- Typically, the process of gathering and analyzing data involves the following path:
  - Ask the right questions to begin the discovery process.
  - Acquire data.
  - Process and clean the data.
  - Integrate and store data.
  - Initial data investigation and exploratory data analysis.
  - Choose one or more potential models and algorithms
  - Apply data science methods and techniques, such as machine learning, statistical modeling, and artificial intelligence.
  - Measure and improve results.
  - Present final results to stakeholders.
  - Make adjustments based on feedback.
  - Repeat the process to solve a new problem

Source: https://www.northeastern.edu/graduate/blog/what-does-a-data-scientist-do/

R & R of a Data Scientist

innovate achieve lead

Data Scientist Job Titles

- The most common careers in data science include the following roles:
  - 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.

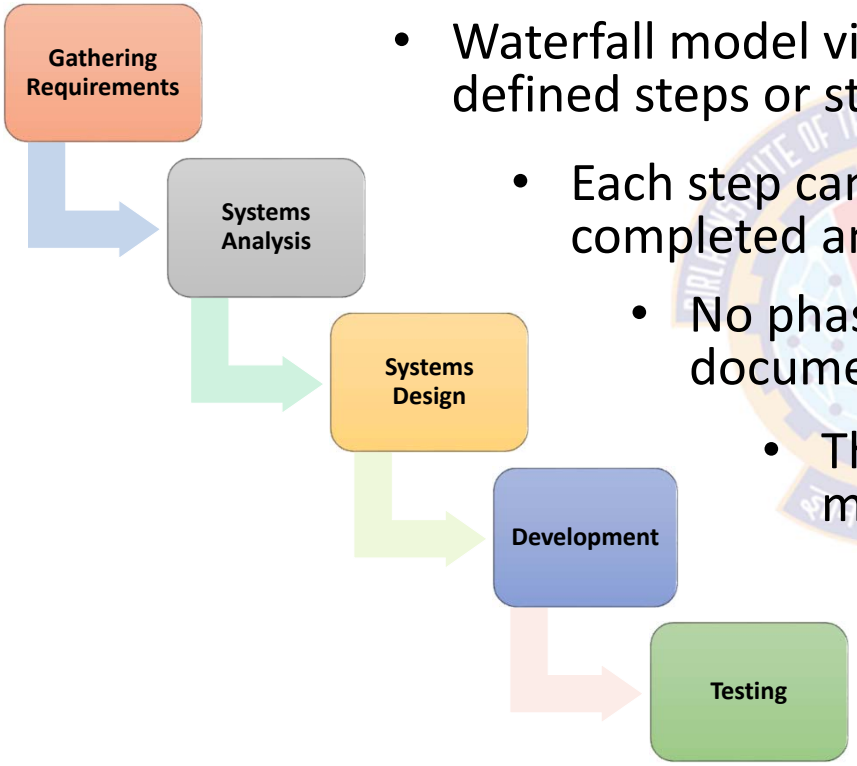
Source: https://www.northeastern.edu/graduate/blog/what-does-a-data-scientist-do/

Software Engineering for Data Science

Software Engineering

- Software engineering is an engineering discipline that is concerned with all aspects of software production
  - From early stages of system specification through to maintaining the system after it has gone into production
- Software includes computer programs, all associated documentation, and configuration data that are needed for software to work correctly
- Software engineers are concerned with developing software products. That is,
  - Generic products that are stand-alone systems
  - Products that are customized for a particular customer

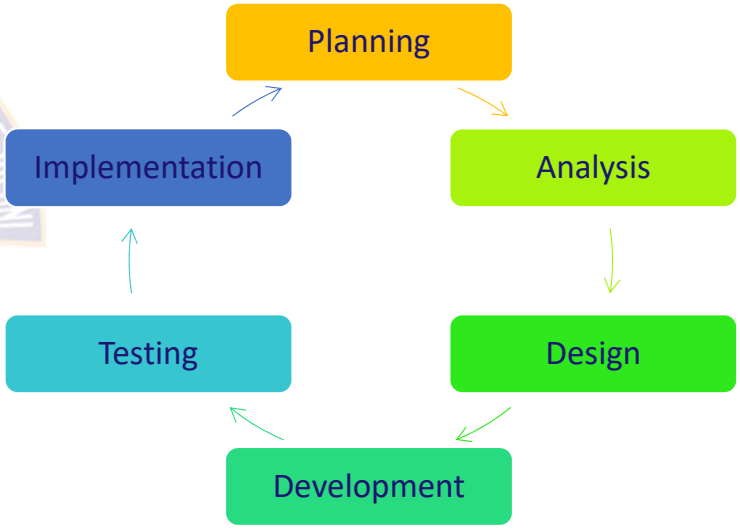
Software Development Process



- Waterfall model views development activities as strictly defined steps or states
  - Each step cannot begin until the previous step is completed and documented
    - No phase is considered complete until its documentation is approved
      - The model is tightly tied to project management phases

Source: Object-Oriented Systems Analysis and Design by Noushin Ashrafi and Hessam Ashrafi

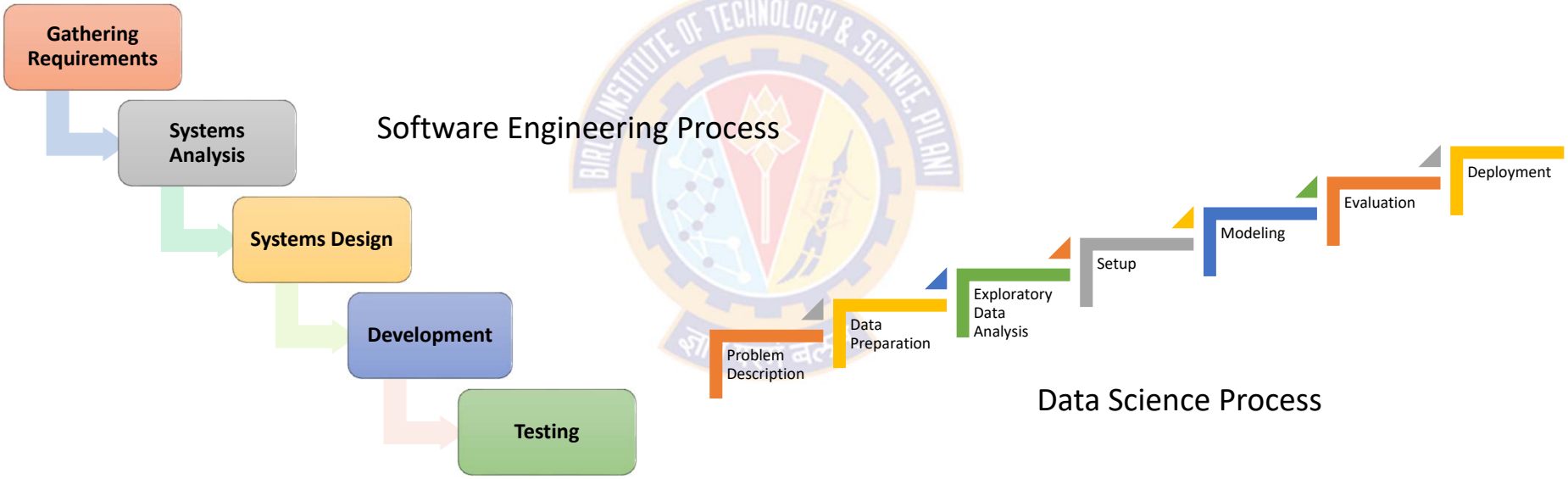
Software Development Process



- The whole project is developed in multiple iterations
- Each iteration implements a selected set of requirements or functionality
- Iterative process groups requirements into layers or cycles
- Within each iteration, the development goes through a complete life cycle: analysis, design, coding, and testing

Source: Object-Oriented Systems Analysis and Design by Noushin Ashrafi and Hessam Ashrafi

Software Engineering Process Vs. 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
Consider a software such as Chrome or FireFox browser that is developed using software engineering principles	You use the browser to search for let's say “Data Science Bootcamp”. The results displayed are probably based on algorithm developed using Data Science
If you wish to find a restaurant, you would use your phone assistant (developed by a software engineer) to find one. Then, an algorithm (developed by a data scientist) searches and finds restaurants close to your location and displays the results in a map application (developed by a software engineer) and tells how exactly you can go	

Source: https://careerkarma.com/blog/data-science-vs-software-engineering/

Software Engineering for Data Science

Data Science

- Involves collecting and analyzing data
- Data scientists utilize the ETL process
- Is more process-oriented
- Data scientists use tools like Amazon S3, MongoDB, Hadoop, and MySQL
- Skills include machine learning, statistics, and data visualization

Software Engineering

- Is concerned with creating useful applications
- Software engineers use the SDLC process
- Uses frameworks like Waterfall, Agile, and Spiral
- Software engineers use tools like Rails, Django, Flask, and Vue.js
- Skills are focused on coding languages

Source: https://careerkarma.com/blog/data-science-vs-software-engineering/

Data Scientist's Toolbox

Data Scientist's Toolbox

innovate achieve lead

Tools Used by Data Scientists for Data Analysis

- Java, R, Python, Clojure, Haskell, Scala…
- Hadoop, HDFS & MapReduce, Spark, Strom…
- HBase, Pig, Hive, Shark, Impala, Cascalog…
- ETL, Webscrapers, Flume, Sqoop, Hume…
- SQL, RDMS, DW, OLAP…
- Knime, Weka, RapidMiner, Scipy, NumPy, scikit-learn, pandas…
- js, Gephi, ggplot2, Tableau, Flare, Shiny…
- SPSS, Matlab, SAS…
- NoSQL, Mongo DB, Couchbase, Cassandra…
- And Yes! … MS-Excel : the most used, most underrated DS tool.

Source: http://makemeanalyst.com/data-scientist-toolkits/

Data Science Challenges

Data Science Challenges

Broad Categorization of Data Science Challenges

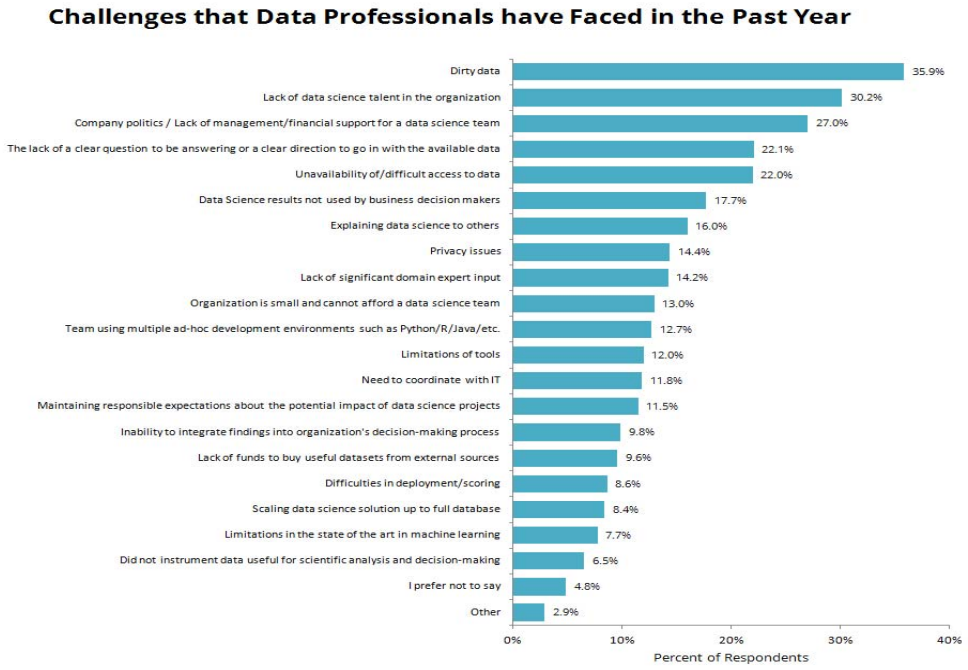
- By reading related material from various sources, data science challenges can be categorized as:
  - Data related
  - Organization related
  - Technology related
  - People related
  - Skill related
  - Can you think of anything else?

Data Science Challenges

innovate achieve lead

Challenges Faced by Data Professionals

- Survey of over 16,000 data professionals
- The question was…
- "At work, which barriers or challenges have you faced this past year? (Select all that apply)."
- The top 10 challenges were:
  - Dirty data (36%)
  - Lack of data science talent (30%)
  - Company politics (27%)
  - Lack of clear question (22%)
  - Data inaccessible (22%)
  - Results not used by decision makers (18%)
  - Explaining data science to others (16%)
  - Privacy issues (14%)
  - Lack of domain expertise (14%)
  - Organization small and cannot afford data science team (13%)



Data are from the Kaggle 2017 The State of Data Science and Machine Learning study. You can learn more about the study and download the data here: https://www.kaggle.com/surveys/2017. Respondents were asked, "At work, which barriers or challenges have you faced this past year? (Select all that apply)." A total of 10153 respondents were asked this questions.

BUSINESS BROADWAY

Source: https://businessoverbroadway.com/2018/03/18/top-10-challenges-to-practicing-data-science-at-work/

Data Science Challenges

innovate achieve lead

Challenges Faced by Data Professionals

- A principal component analysis of the 20 challenges resulted in five categories
  - Insights not Used in Decision Making
    - These challenges include company politics, an inability to integrate study findings into decision-making processes and lack of management support.
  - Data Privacy, Veracity, Unavailability
    - These challenges revolved around the data itself, including how “dirty” it is, its availability as well as privacy issues.
  - Limitations of tools to scale / deploy
    - Challenges in this category are related to the tools that are used to extract insights, deploy models as well as scaling solutions up to the full database.
  - Lack of Funds
    - Challenges around lack of funding impact what the organization can purchase with respect to external data sources, data science talent and, perhaps, domain expertise.
  - Wrong Questions Asked
    - Challenges are about the difficulty in maintaining expectations about the impact of data science projects and not having a clear question to answer or a clear direction to go in with the available

Source: https://businessoverbroadway.com/2018/03/18/top-10-challenges-to-practicing-data-science-at-work/

Data Science Challenges

- Insights not Used in Decision Making
  - Company politics
  - Inability to integrate study findings into decision-making processes
  - Lack of management support.
- Data Privacy, Veracity, Unavailability
  - “Dirty” data
  - Lack of availability
  - Privacy issues
- Limitations of tools to scale / deploy
  - Challenges in this category are related to the tools that are used to extract insights, deploy models as well as scaling solutions up to the full database.
- Lack of Funds
  - Challenges around lack of funding impact what the organization can purchase with respect to external data sources, data science talent and, perhaps, domain expertise.
- Wrong Questions Asked
  - Challenges are about the difficulty in maintaining expectations about the impact of data science projects and not having a clear question to answer or a clear direction to go in with the available

Principal Component Analysis of Challenges Experienced by Data Professionals

	Components				
Challenges	**Insights not Used in Decision Making**	**Data Privacy, Veracity, Unavailability**	**Limitations of tools to scale / deploy**	**Lack of Funds**	**Wrong Questions Asked**
Company politics / Lack of management/financial support for a data science team	**.63**	.18	.03	.26	-.04
Data Science results not used by business decision makers	**.67**	.04	.06	.07	.17
Inability to integrate findings into organization's decision-making process	**.67**	.01	.06	-.06	.22
Need to coordinate with IT	**.49**	.36	.21	-.12	-.27
Dirty data	.16	**.62**	.16	.10	.25
Privacy issues	.13	**.60**	.11	.00	-.07
Unavailability of/difficult access to data	.07	**.65**	.13	.07	.18
Difficulties in deployment/scoring	.15	.11	**.56**	-.13	.10
Limitations in the state of the art in machine learning	-.04	-.07	**.56**	.33	.02
Limitations of tools	.15	.27	**.40**	.21	-.35
Scaling data science solution up to full database	.08	.04	**.61**	.04	.12
Team using multiple ad-hoc development environments such as Python/R/Java/etc.	.13	.27	**.40**	.01	.12
Lack of funds to buy useful datasets from external sources	.02	.18	.14	**.56**	-.05
Lack of significant domain expert input	.08	.06	.22	**.41**	.35
Organization is small and cannot afford a data science team	.13	.00	-.09	**.69**	.04
Maintaining responsible expectations about the potential impact of data science projects	.22	.16	.24	-.05	**.57**
The lack of a clear question to be answering or a clear direction to go in with the available data	.26	.28	.08	.12	**.57**
Did not instrument data useful for scientific analysis and decision-making	.32	.06	.24	.09	.15
Explaining data science to others	.39	.19	.15	.09	.31
I prefer not to say	.00	-.38	.13	-.21	-.11
Lack of data science talent in the organization	**.47**	.14	.09	**.45**	.02

Data are from the Kaggle 2017 The State of Data Science and Machine Learning study. You can learn more about the study and download the data here: https://www.kaggle.com/surveys/2017. Respondents were asked, "At work, which barriers or challenges have you faced this past year? (Select all that apply)." A total of 10153 respondents were asked this questions. Data were coded as 1 if challenge was selected, 0 if challenge was not selected. Table represents principal component matrix (after varimax rotation). Elements of .40 or greater appear in bold.

BUSINESS BROADWAY

Source: https://businessoverbroadway.com/2018/03/18/top-10-challenges-to-practicing-data-science-at-work/

Data Science Challenges

Data Science Challenges

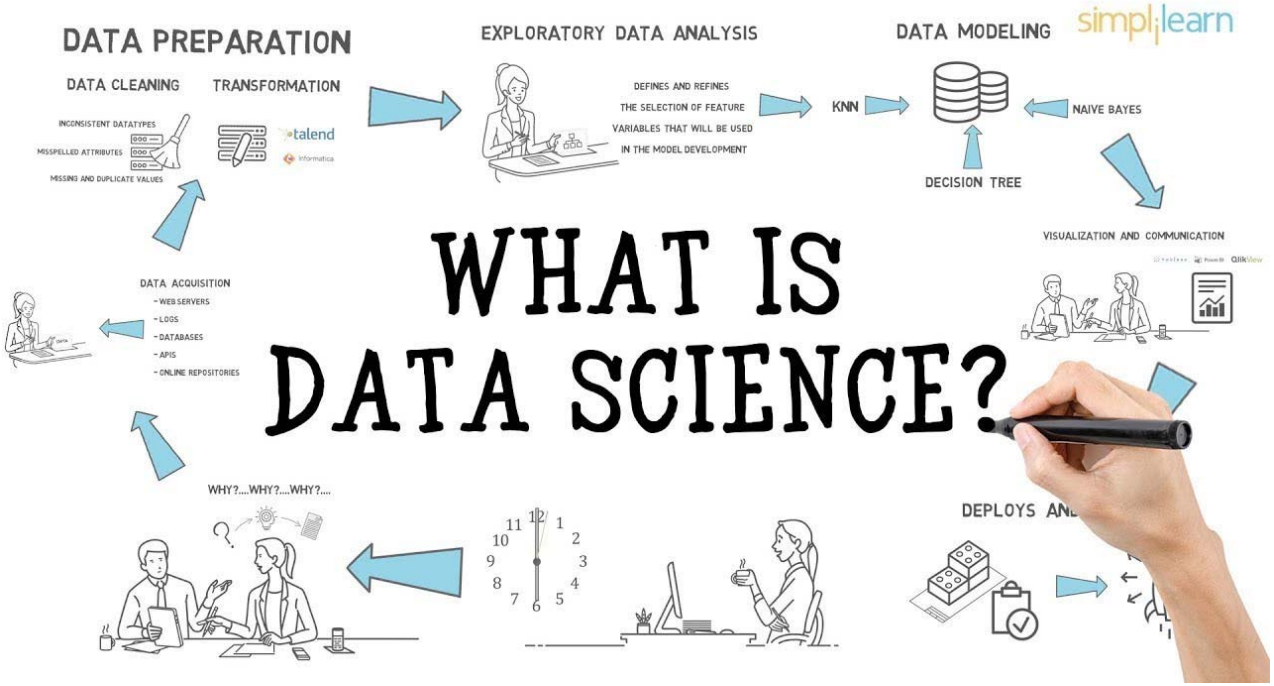
- Unreasonable Management Expectations
- Misunderstanding Data's Significance
- Being Blamed for Bad News
- Having to Convince Management
- Lack of Professionals
- Problem Identification
- Accessing the Right Data
- Cleansing of the Data

Source: https://www.dataquest.io/blog/data-science-problems-fix/

innovate achieve lead

Data Science Challenges

- Complexity of reality
- Cognitive bias
- Data Quality
  - Accuracy
  - Completeness
  - Uniqueness
  - Timeliness
  - Consistency
- Content and Source Bias
- Granularity of Data
  - In terms of time and space
- Availability & Access of Data
  - Data Divide



Source: Coursera
Source: https://dimensionless.in/challenges-faced-by-data-scientist-and-how-to-overcome-them/

Data Science Challenges

Cognitive Bias

- The way we see the world (events, facts, people, etc.,) is based on our own belief system and experiences and may not be reasonable or accurate
- Cognitive biases are ways of thinking about and perceiving the world that may not necessarily reflect reality
- We may think we experience the world around us with perfect objectivity, but this is rarely the case
- 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
- Simply put, cognitive biases are the distortions of reality because of the lens through which we view the world

Source: https://www.wordstream.com/blog/ws/2014/05/22/landing-pages-cognitive-biases

Data Science Challenges

Types of Cognitive Biases

- Selection (or sample) Bias
- Seasonal Bias
- Linearity Bias
- Confirmation Bias
- Recall Bias
- Survivor Bias
- Observer Bias
- Reinforcement Bias
- Availability Bias

Source: https://www.wordstream.com/blog/ws/2014/05/22/landing-pages-cognitive-biases

Data Science Challenges

innovate achieve lead

Cognitive Bias

- MONEYBALL - breaking biases - FREQUENCY BASED PROBABILITY / STATISTICS - MATHEMATICS in the MOVIES
  - https://www.youtube.com/watch?v=KWPhV6PUr9o&ab_channel=Movieclips
- Types of cognitive bias
  - https://www.youtube.com/watch?v=wEwGBIr_RIw&ab_channel=PracticalPsychology
- Graphs can be misleading
  - https://www.youtube.com/watch?v=E91bGT9BjYk&ab_channel=TED-Ed
- Simpson's paradox
  - https://www.youtube.com/watch?v=sxYrzzy3cq8&ab_channel=TED-Ed

Source: https://www.wordstream.com/blog/ws/2014/05/22/landing-pages-cognitive-biases


innovate
achieve
lead

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
ज्ञानं परमं बलम्

Thank You!