

Awarding
Great British
Qualifications



Introduction to Data Science and Big Data


Topic 1: Lecture 1

Data Science and Big Data Fundamentals

Data Science and Big Data Fundamentals Topic 1 - 1.2

The Unit Roadmap

Unit Aim: The aim of this unit is to introduce you to the fundamental concepts and techniques of data science and big data, enabling you to understand the role and significance of data in various domains, apply data analysis methods, and utilize big data tools and technologies to extract meaningful insights and support decision-making.



Data Science and Big Data Fundamentals Topic 1 - 1.3

Unit Syllabus

1. Data Science and Big Data Fundamentals

2. Introduction to Data

3. Understanding Data & Exploration

4. Data Pre-Processing 1

5. Data Pre-Processing 2

6. Data Processing 1

7. Data Processing 2


8. Model Selection and Evaluation

9. Data Visualisation

10. Business Intelligence and Tools

11. Data Science Ethical and Privacy Issues

12. Unit Summary



Data Science and Big Data Fundamentals Topic 1 - 1.4

Unit Delivery


• The teacher-led time for this unit is comprised of lectures and Tutorial/laboratory sessions.

• Lectures are designed to start each topic.

✓ You will be encouraged to be active during lectures by raising questions and taking part in discussions.

• Tutorial/Laboratory sessions are designed to follow the respective topic lecture.

✓ During these sessions, you will be required to work through practical tutorials and various exercises.



V0.0

Visuals Handout – Page 1

Data Science and Big Data Fundamentals Topic 1 - 1.5

Private Study

- You are also expected to undertake private study to consolidate and extend your understanding.
- Exercises are provided in your Student Guide for you to complete during this time.



Data Science and Big Data Fundamentals Topic 1 - 1.6

Assessment

This unit will be assessed by:

- An examination worth 50% of the total mark
- An assignment worth 50% of the total mark



Data Science and Big Data Fundamentals Topic 1 - 1.7

Scope and Coverage

This topic will cover:

- What is Data Science?
- What is Big Data?
- What is Data Analytic?
- Data Science vs Big Data vs Data Analytics
- Components of Data Science
- Data Science Process/lifecycle



Data Science and Big Data Fundamentals Topic 1 - 1.8

Learning Outcomes

By the end of this topic students will be able to:

- Understand the concept of Data Science and Big Data along with the relationship among two.
- Understand the need and applications of Data Science and Big Data.
- Understand the role and responsibility of data scientist, big data expert and data analyst.
- Identify the components and life cycle of Data Science.



Data Science and Big Data Fundamentals Topic 1 - 1.9

Data Science

“Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.”

Mining large amount of structured and unstructured data to identify patterns

Includes a combination of programming, statistical skills, machine learning, and algorithms

Ley, C., Borda, S.P.A. What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences. Int J Data Sci Anal 6, 167–175 (2018)



Data Science and Big Data Fundamentals Topic 1 - 1.10

Data Mining

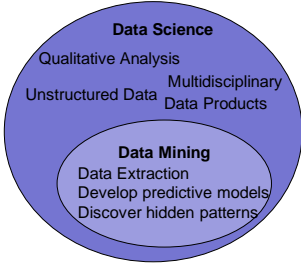
- Data Mining is a process of discovering patterns in large structured data sets involving methods at the intersection of machine learning, statistics, and database systems.
- Data mining is an **inter-disciplinary** subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.



Data Science and Big Data Fundamentals Topic 1 - 1.11

Data Science vs Data Mining

- **Data science** includes the processes of capturing of data, analysing, and deriving insights from it. **Data mining** is finding useful information in a dataset and utilising that information to uncover hidden patterns.
- **Data science** is multidisciplinary and **Data mining** is interdisciplinary and subset of data science.
- **Data science** deals with structured, semi-structured, or unstructured. **Data mining** mostly deals with structured data.



Data Science and Big Data Fundamentals Topic 1 - 1.12

Big Data

- Refers to homogenous volume of data
- Includes capturing data, data storage, data sharing and data querying

What is the difference between Data science and Big data ?




Data Science and Big Data Fundamentals Topic 1 - 1.13

Data Science vs Big Data

- They are not the “same thing”
- Big data = crude oil

Big data is about extracting “crude oil”, transporting it in “mega tankers”, siphoning it through “pipelines”, and storing it in “massive silos”

- Data science is about refining the “crude oil”




Data Science and Big Data Fundamentals Topic 1 - 1.14

Data Analytics

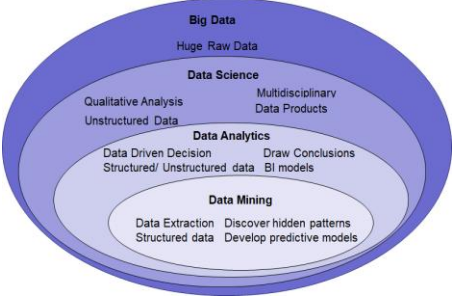
- Perform and process statistical analysis of data
- Discover how data can be used to draw conclusion
- Data analytics and data mining are related fields, but they have different goals and methods.


Data Analytics	Data Mining
Examining data set to draw conclusion	Discover hidden pattern in data set
Make data driven decisions	Make data useable
Use analytical and business intelligence (BI) models	Use mathematical and scientific methods
Any size unstructured, semi structured or structured	Large structured data



Data Science and Big Data Fundamentals Topic 1 - 1.15

Big Picture





Data Science and Big Data Fundamentals Topic 1 - 1.16

Class Activity 1

Can you identify the role of data scientist, big data expert and data analyst?

- Explore and examine data from multiple disconnected sources
- Design and Create data reports
- Architect high scalable distributed systems
- Acquire, process and summarize data
- Develop new learning algorithms


Data Scientist

Data Analyst

Big Data Expert

Data Analyst

Data Scientist



Role of Data Scientist, Big Data Expert and Data Analyst

Data Scientist	Predict the future based on past patterns	Explore data from multiple disconnected sources	Develop new analytical and learning models
Big Data Expert	Analyse system bottlenecks	Architect highly scalable distributed systems	Build large scale data processing systems
Data Analyst	Acquire, process and summarise data	Pack data for insights	Design and create data reports



Target Industries

Data Scientist	Search Engines	Financial Services	E-commerce
Big Data Expert	Communications	Financial Services	Retail
Data Analyst	Healthcare	Travel	IT industry



What Skills Need For These Roles?

Data Scientist	Big Data Expert	Data Analyst
Programming skills like R, Python, SAS, etc.	Programming skills like Scala, Java, etc.	Programming skills like R, Python, SAS etc.
Statistical and mathematical skills	NoSQL databases like MongoDB, CassandraDB	Statistical and mathematical skills
Storytelling and data visualisation	Hadoop, PySpark	Data Wrangling
Hadoop, PySpark, SQL	Excellent grasp of distributed systems	Data Visualisation
Machine Learning		Excellent grasp of distributed systems



Data Science Application Examples

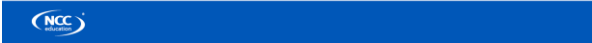
- Fraud detection**
- Investigate fraud patterns in past data
 - Early detection is important
 - Before damage propagates
 - Harder than late detection
 - Precision is important
 - False positive and false negative are both bad
 - Real-time analytics



Data Science and Big Data Fundamentals Topic 1 - 1.21

Data Science Application Examples

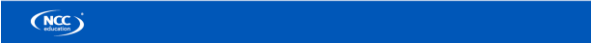
- Recommender systems
- The ability to offer unique personalised service
 - Increase sales, click-through rates, conversions, ...
 - ✓ Netflix recommender system valued at \$1B per year
 - ✓ Amazon recommender system drives a 20-35% lift in sales annually
 - Collaborative filtering at scale



Data Science and Big Data Fundamentals Topic 1 - 1.22

Data Science Application Examples

- Predicting why patients are being readmitted
- Reduce costs
 - Improve population health
 - Find the “why” behind specific populations being readmitted
 - Database of multiple data sources
 - Investigate ties between readmission and socioeconomic data points, patient, history, genetics, ...



Data Science and Big Data Fundamentals Topic 1 - 1.23

Data Science Application Examples

- “Smart cities”
- Refers to using data and ICT to
 - ✓ Better plan communities
 - ✓ Better manage assets
 - ✓ Reduce costs
 - ✓ Deploy open data to better engage with community
 - Some use cases:
 - ✓ Intelligent Transport System
 - ✓ Smart Homes - Energy Monitoring
 - ✓ Digital Government
 - ✓ Smart Farming, etc.



Data Science and Big Data Fundamentals Topic 1 - 1.24

Checkpoint Summary

- Big data refers to any large and complex collection of data.
- Data science is a multidisciplinary field that aims to produce broader insights.
- Data analytics is the process of extracting meaningful information from data to make data driven decisions
- Data Mining is a process of discovering patterns in large structured data sets to make it usable.



NCC

education

Awarding
Great British
Qualifications

Introduction to Data Science and Big Data
Topic 1: Lecture 2
Component of Data Science and Lifecycle

Data Science and Big Data Fundamentals Topic 1 - 1.26

Component of Data Science

Statistics

Programming Language

Learning Techniques

Visualisation

Domain Expertise

NCC

education

Data Science and Big Data Fundamentals Topic 1 - 1.27

Statistics

Statistics is one of the most important components of data science.

Statistics is a way to collect and analyse the numerical data in a large amount and finding meaningful insights from it.

The main advantage of statistics is that information is presented in an easy way.

Some common statistic concepts that every data scientist should know, and you will learn in this module: Probability, Central Tendency, Variability, Relationship Between Variables, Probability Distribution.

NCC

education

Data Science and Big Data Fundamentals Topic 1 - 1.28

Programming Language

Programming languages help data scientists draw insight from large datasets.

Data science programming languages may specialise in performing certain tasks.

Data scientists may learn one versatile language or combine languages for better results.

Some common programming languages that data scientist should know: Python, R, Javascript, SQL, NoSQL, Scala, etc.

NCC

education



Learning Techniques

- Data Scientists must understand Machine Learning for quality predictions and estimations.
- This can help machines to take right decisions and smarter actions in real time with zero human intervention.
- Some machine learning algorithms used in data science: Regression, Decision Tree, Clustering, Principal Component Analysis, Support Vector Machines, Artificial Neural Network, etc.

Visualisation

- Data visualisation is the graphical representation of information and data.
- By using visual elements like charts, graphs and maps, data visualisation tools provide an accessible way to see and understand trends, outliers and patterns in data.



Domain Expertise

- Domain expertise is the knowledge and understanding of a particular field.
 - As data scientists, you may be working in a wide variety of industries, each of which has its own intricacies that can only be learned gradually over time.
 - For simplicity: Look at these groups of words for different industries:
 - ✓ Industry A: loss ratio, combined ratio, conversion rate, price elasticity, price optimisation. **Insurance**
 - ✓ Industry B: Orderbook, arbitrage, short, Sharpe ratio, volume weighted average price, time weighted average price. **Trading**
 - ✓ Industry C: Genomic, clinical/phenotypic, pharmacokinetic, and other molecular data. **Genomics**
- Can you guess what industry each group of words comes from?

CRISP DM Process

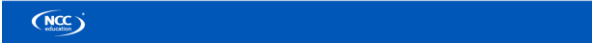
The **C**Ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) is a process model that serves as the base for a data science process.



Data Science and Big Data Fundamentals Topic 1 - 1.33

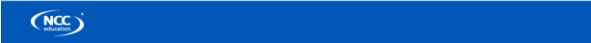
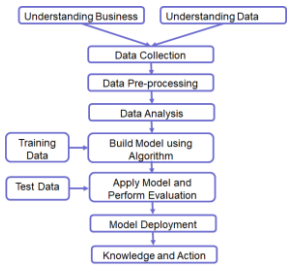
CRISP DM Process

- It has six sequential phases:
- Business Understanding – What does the business need?
 - Data Understanding – What data do we have / need? Is it clean?
 - Data Preparation – How do we organise the data for modeling?
 - Modeling – What modeling techniques should we apply?
 - Evaluation – Which model best meets the business objectives?
 - Deployment – How do stakeholders access the results?



Data Science and Big Data Fundamentals Topic 1 - 1.34

Data Science Lifecycle



Data Science and Big Data Fundamentals Topic 1 - 1.35

Business Understanding

- Any good project starts with a deep understanding of the customer's needs. Data science projects are no exception.
- The Business Understanding phase focuses on understanding the objectives and requirements of the project and includes:
 - ✓ **Determine business objectives:** You should first "thoroughly understand, from a business perspective, what the customer really wants to accomplish."
 - ✓ **Assess situation:** Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
 - ✓ **Determine goals:** In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
 - ✓ **Produce project plan:** Select technologies and tools and define detailed plans for each project phase.



Data Science and Big Data Fundamentals Topic 1 - 1.36

Data Understanding and Collection

- Data understanding phase drives the focus to identify, collect, and analyse the data sets that can help you accomplish the project goals. This phase also has four tasks:
 - ✓ **Collect initial data:** Acquire the necessary data and (if necessary) load it into your analysis tool.
 - ✓ **Describe data:** Examine the data and document its surface properties like data format, number of records, or field identities.
 - ✓ **Explore data:** Dig deeper into the data. Query it, visualise it, and identify relationships among the data.
 - ✓ **Verify data quality:** How clean/dirty is the data? Document any quality issues.



Data Science and Big Data Fundamentals Topic 1 - 1.37

Data Understanding and Collection

- The data captured can be either in structured or unstructured form.
- The methods of collecting the data might come from – logs from websites, social media data, data from online repositories, and even data streamed from online sources via APIs, web scraping or data that could be present in excel or any other source.
- This phase include **data exploration** using Numerical Methods of Exploratory Data Analysis (EDA): Central Tendency, Measurement of Variability and Graphical Methods of EDA: histogram, box plot, scatter plot, stem, and leaf plot.



Data Science and Big Data Fundamentals Topic 1 - 1.38

Data Pre-processing

- A common rule of thumb is that **80%** of the project is data preparation.
- This phase, which is often referred to as "data munging", prepares the final data set(s) for modeling. It has five tasks:
 - ✓ **Select data:** Determine which data sets will be used and document reasons for inclusion/exclusion.
 - ✓ **Clean data:** Often this is the lengthiest task. Without it, you'll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
 - ✓ **Construct data:** Derive new attributes that will be helpful. For example, derive someone's body mass index from height and weight fields.
 - ✓ **Integrate data:** Create new data sets by combining data from multiple sources.
 - ✓ **Format data:** Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.



Data Science and Big Data Fundamentals Topic 1 - 1.39

Data Pre-processing

- ✓ Processing and fine-tuning the raw data, critical for the goodness of the overall project.
- ✓ Data scientists analyse the data collected for biases, patterns, ranges, and distribution of values.
- ✓ This phase includes handling missing values, categorical encoding, feature engineering, dimensionality reduction, outlier analysis, class imbalance.



Data Science and Big Data Fundamentals Topic 1 - 1.40

Data Analysis

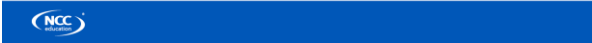
- ✓ The pre-processed data is subjected to various data processing methods using machine learning (ML) and artificial intelligence (AI) algorithms to generate a desirable output.
- ✓ This step may vary slightly from process to process depending on the source of data being processed (data lakes, online databases, connected devices, etc.) and the intended use of the output.
- ✓ Types of Machine Learning: supervised learning, unsupervised learning, reinforcement learning, NLP, deep learning, etc.



Data Science and Big Data Fundamentals Topic 1 - 1.41

Modeling

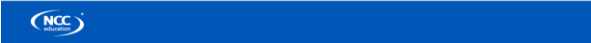
- Build and assess various models based on several different modeling techniques. This phase has four tasks:
 - ✓ **Select modeling techniques:** Determine which ML algorithms to try (e.g., regression, neural net).
 - ✓ **Generate test design:** Pending your modeling approach, you might need to split the data into training, test, and validation sets.
 - ✓ **Build model:** As glamorous as this might sound, this might just be executing a few lines of code like "reg = LinearRegression().fit(X, y)".
 - ✓ **Assess model:** Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.



Data Science and Big Data Fundamentals Topic 1 - 1.42

Evaluation

- Evaluation phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:
 - ✓ **Evaluate results:** Do the models meet the business success criteria? Which one(s) should we approve for the business?
 - ✓ **Review process:** Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarise findings and correct anything if needed.
 - ✓ **Determine next steps:** Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.



Data Science and Big Data Fundamentals Topic 1 - 1.43

Model Deployment

- "Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise."
- A model is not particularly useful unless the customer can access its results.**
- The complexity of this phase varies widely. This final phase has four tasks:
- ✓ **Plan deployment:** Develop and document a plan for deploying the model.
 - ✓ **Plan monitoring and maintenance:** Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.
 - ✓ **Produce final report:** The project team documents a summary of the project which might include a final presentation of data mining results.
 - ✓ **Review project:** Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.



Data Science and Big Data Fundamentals Topic 1 - 1.44

Checkpoint Summary

- The data science lifecycle involves several key steps:
- Problem Definition: Clearly define the objective and goals of project.
 - Data Acquisition: Gather data from sources i.e., databases or APIs.
 - Data Exploration: Explore and understand the data through visualisations and statistical summaries.
 - Data Preparation: Clean, preprocess, and transform the data.
 - Data Analysis: The pre-processed data is subjected to data processing methods using ML and AI algorithms to generate a desirable output.
 - Modeling: Develop and train predictive or analytical models based on the prepared data.
 - Model Evaluation and Deployment: Evaluate model performance, select the best model, and deploy it into a production environment.



Data Science and Big Data Fundamentals Topic 1 - 1.45

Quiz

1. What is the first step in the data science lifecycle?
- a) Data Acquisition
 - b) Problem Definition **Correct Answer**
 - c) Data Preparation
 - d) Exploratory Data Analysis



Data Science and Big Data Fundamentals Topic 1 - 1.46

Quiz

2. Which stage involves cleaning, preprocessing, and transforming the data?
- a) Problem Definition
 - b) Data Acquisition
 - c) Data Preparation **Correct Answer**
 - d) Exploratory Data Analysis



Data Science and Big Data Fundamentals Topic 1 - 1.47

Quiz

3. During which stage do data scientists explore and understand the data through visualisations and statistical summaries?
- a) Data Acquisition
 - b) Exploratory Data Analysis **Correct Answer**
 - c) Model Evaluation
 - d) Model Deployment



Data Science and Big Data Fundamentals Topic 1 - 1.48

Quiz

4. Which stage involves developing and training predictive or analytical models?
- a) Problem Definition
 - b) Data Acquisition
 - c) Modeling **Correct Answer**
 - d) Model Evaluation



Data Science and Big Data Fundamentals: Topic 1 - 1.49

Quiz

5. What is the final step in the data science lifecycle?
- a) Model Evaluation
 - b) Model Deployment Correct Answer
 - c) Data Preparation
 - d) Exploratory Data Analysis



Data Science and Big Data Fundamentals: Topic 1 - 1.50

Discussion Session

What are some common challenges faced when working with data? How can these challenges impact the effectiveness of data-driven solutions in real-world scenarios?



Data Science and Big Data Fundamentals: Topic 1 - 1.51

Discussion Session

Some common challenges when working with data include:

Data quality issues: Inaccurate, incomplete, or inconsistent data can lead to biased or unreliable insights.

Data privacy and security: Safeguarding sensitive data while maintaining its usefulness is crucial to comply with regulations and protect individuals' privacy.

Data volume and complexity: Dealing with large volumes of data and diverse data types can pose challenges in storage, processing, and analysis.

Data integration: Integrating data from multiple sources with different formats and structures requires careful mapping and alignment.

Lack of domain knowledge: Understanding the context and domain-specific nuances is essential to interpret the data correctly and derive meaningful insights.



Data Science and Big Data Fundamentals: Topic 1 - 1.52

Discussion Session

- These challenges can impact the effectiveness of data-driven solutions by:
- Leading to inaccurate or biased conclusions, resulting in flawed decision-making.
 - Increasing the risk of privacy breaches or non-compliance with data regulations.
 - Slowing down analysis and decision-making processes due to data processing and integration complexities.
 - Impeding the ability to uncover hidden patterns or insights due to data quality issues.
 - Limiting the applicability and interpretability of the results without proper domain knowledge.
- Addressing these challenges requires a combination of technical skills, data governance practices, domain expertise, and effective collaboration between data scientists, domain experts, and stakeholders.



Topic Summary

Data Science and Big Data Fundamentals: Topic 1 - 1.53

- Data Science is not same as Big Data.
- Data science is a multidisciplinary field that aims to produce broader insights from the data.
- Data science is a superset of Data Analytics and Data mining.
- Data Science Process includes understanding business and data, data exploration, data pre-processing, data analysis, data modelling, model evaluation and deployment, knowledge & action.



Next Topic 2:
Introduction to Data

Data Science and Big Data Fundamentals: Topic 1 - 1.54

- What is Data?
- How big is Big Data? Some examples
- Sources of Data
- Big Data Challenges
- Data quality and issues
- 5V's of Big Data
- Types of Data: Structured, Unstructured, Semi Structured
- Categories of Data Types: Continuous, Categorical, Text Data, Time Series, Binary.



References

Data Science and Big Data Fundamentals: Topic 1 - 1.55

- Grus, J. (2019). [Data Science from Scratch](#). O'Reilly Media.
- Holmes, D. E. (2013). [Big Data: A Very Short Introduction](#). Oxford University Press.
- Kotu, V., & Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with RapidMiner. Morgan Kaufmann.
- Marz, N., & Warren, J. (2015). Big Data: Principles and best practices of scalable real-time data systems. Manning Publications.
- Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.



Topic 1 – Data Science and Big Data Fundamentals

Any Questions?