EXPLORE AI ACADEMY

An introduction to data science

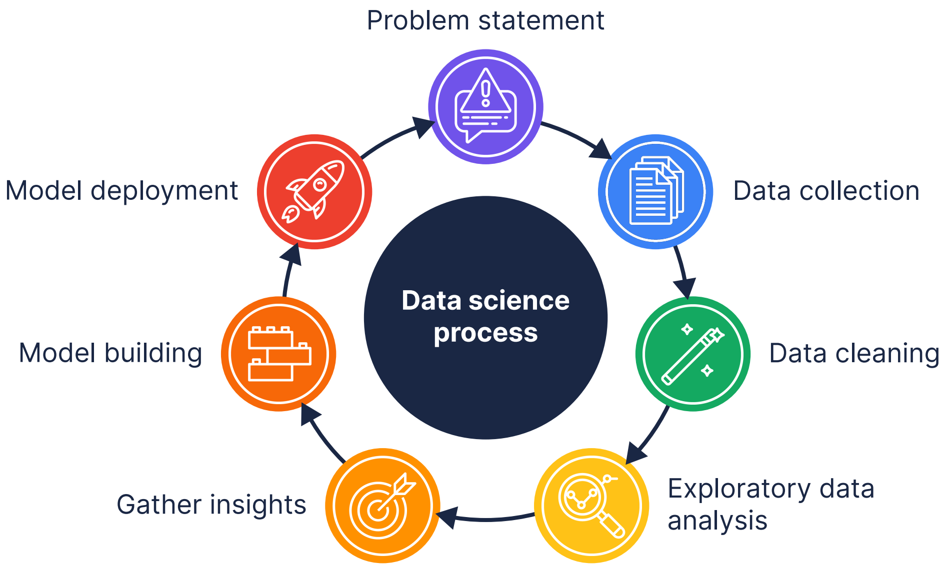
Approaches in data science

Overview

It is important that we are able to make **informed decisions** and **derive appropriate insights** from data.

We therefore need a **structured framework** for working with data and extracting valuable insights from it.

How our data science process is **applied** and **interpreted** depends on several factors, including whether we are doing a **quantitative** or **qualitative** analysis, and whether we need **hindsight, insight, foresight, or context**.



Quantitative and qualitative data analyses

> Quantitative and qualitative data analyses are important because they enable us to gain a **more comprehensive understanding** of complex phenomena and make **data-driven decisions**.

Quantitative data analysis involves numerical measurement and statistical analysis.

It allows us to **measure and analyse numerical data** using **statistical methods**, enabling us to **identify patterns**, trends, and relationships between variables.

It is useful for making predictions, testing hypotheses, and identifying cause-and-effect relationships.

Qualitative data analysis involves exploring patterns and themes in non-numerical data.

It allows us to **explore and interpret non-numerical data**, such as text, images, or videos.

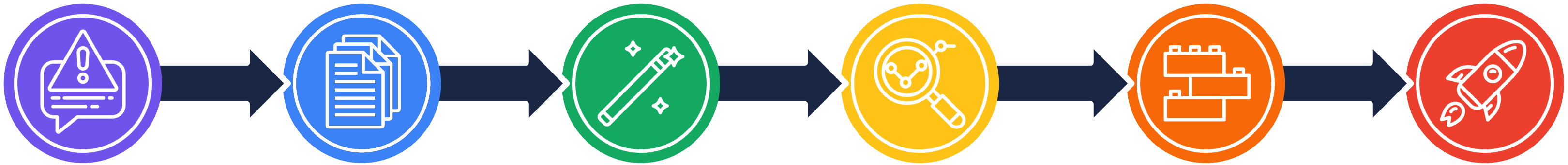
It is useful for understanding the **context of a problem** and people's attitudes, behaviours, etc.

Both types of analysis are important because they provide **different ways of understanding** and **interpreting data**.

An introduction to data science

The data science process

> The data science process is a systematic approach to **transforming a data problem** into a **data-driven solution**.



Problem statement
State the problem or hypothesis

Data collection
Find the right data sources

Data cleaning
Remove, fix, and filter

Exploratory data analysis
Understand the data

Model building
Select features, build, train, and validate

Model deployment
Deploy, test, and communicate

This approach to data science helps us to **discover meaningful** patterns, relationships, and trends and helps us develop **accurate** and **robust** models. **Various forms** of this process are used **across different data disciplines**, including data analytics, science, and engineering, under various names, such as OSEMN and CRISP-DM.

Problem statement

The problem statement helps us **define the scope and objectives** of our analysis and ensures that our insights are **relevant**.

A **problem statement** identifies the gap between the **current (problem) state** and the **desired (outcome) state**. It should be specific, brief, concise, clear, unbiased, and measurable.

A problem statement may also be in the form of a **hypothesis**, which is a **proposed cause and effect** for a particular phenomenon or problem which has not yet been proven correct.

Examples:

Statement: We need to report on estimated water and electricity income from different customer groups.

Hypothesis: The estimated water and electricity income from domestic customers are 30% lower than from other customers.

Question: How much water and electricity income can we expect from commercial customers per month?

Data collection

Data collection includes **identifying and acquiring applicable data sources**, both internally and externally, which can help answer the problem statement.

We can use company data or open-source data, or collect our own data **depending on the nature** of our **problem** and the **analysis** we would like to do.

Examples:

Data acquired from **surveys** such as market research and customer satisfaction surveys.	Queried data from **databases** or **APIs** (Application Programming Interfaces) such as sales data and employee information.	Downloaded data from **open sources** and **cloud repositories** such as general census data.

Data cleaning

Data cleaning, also known as **data wrangling**, involves **transforming raw data into usable formats**.

We can use **several cleaning techniques to ensure** that our data are indeed **accurate** and of the required **quality**. If our data are inaccurate, so will our insights be.

Examples:

Using **spreadsheets** or a **programming language** to remove irrelevant observations, handle missing values, fix structural issues, etc.

Using **regular expressions** for pattern matching and replacing data.

Using **data visualisation tools** such as PowerBI or spreadsheets for identifying outliers and anomalies.

Exploratory data analysis

Exploratory data analysis (EDA) is an approach used to **summarise the main characteristics of a dataset** using aggregations, fundamental statistics, and visualisation techniques.

Before we can gather insights or build a model, we first need to **understand our data**. We can use **non-graphical methods**, such as descriptive statistics and correlation, or **graphical (visualisation) methods** to investigate our data.

Examples:

Descriptive statistics	`Standard dev.`
Aggregations	`Count`
Measures of central tendency	`Mean`
Measures of distribution	`Kurtosis`
Correlation	`Pearson`

Bar Scatter Density Violin

Univariate and multivariate analyses

In EDA, we do either a univariate or multivariate analysis, depending on what we want to investigate.

Univariate analysis is the exploration of individual variables in a dataset, i.e. we only consider one variable at a time.

Non-graphical

We can use **descriptive statistics** such as the standard deviation, central tendency, and measures of distribution.

Graphical

We can use **visualisations** such as **histograms**, **density plots**, and **box plots** to understand the characteristics of a variable.

In a **multivariate analysis** we're more interested in the relationship between the different variables of our dataset.

Non-graphical

We use **correlation** to understand the strength and direction between variables.

Graphical

We can use **visualisations** such as **heatmaps**, **scatter plots**, and **pair plots** to investigate the relationship.

Gather insights

Gathering insights, also known as **data dissemination**, involves **gathering** and **reporting** the **insights** derived from the analysis.

Insights may be gathered in and reported to stakeholders through dashboards and reports that include text and data visualisations.

Examples:

Using **spreadsheets** or a **programming language** to summarise data and construct insights to form a report.

Using **data visualisation tools** such as PowerBI or spreadsheets to visualise and report the insights.

Model building

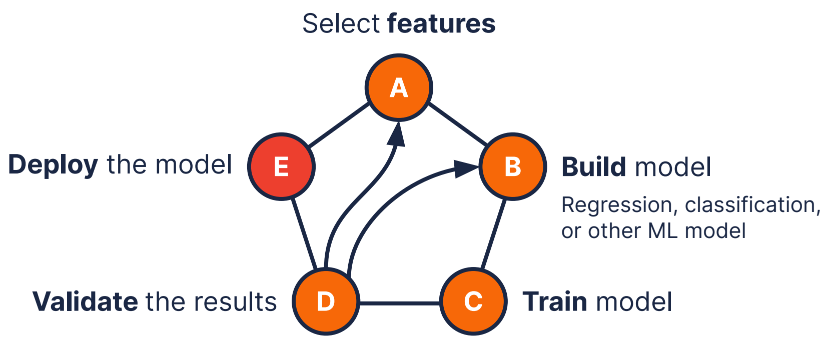
Model building involves **selecting an appropriate algorithm** and **training the model** on the data.

Model building often **involves reiteration** since a model will **rarely give us the results we seek on the first try**. This means that we train and test a model until we've found a suitable model before deploying it into a larger system.

Some common **tools** and **skills required for data collection** include:

Machine learning libraries such as Scikit-learn and TensorFlow for building models in Python.

Deep learning libraries such as Keras and PyTorch for building neural networks in Python.



Model deployment

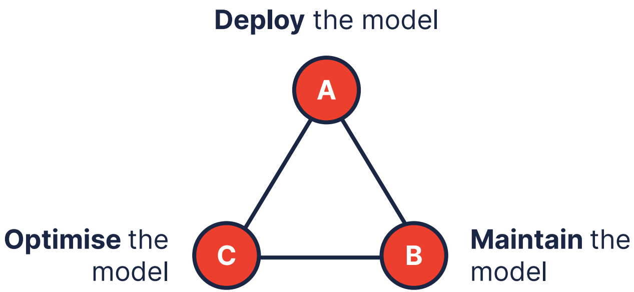
Model deployment involves **integrating the model** into a large system or application.

Deployment bridges the gap between data science and real-world applications. **Effective testing** and **communication** ensure the model is useful, reliable, and understood.

Although we have reached the end of the process, it is crucial to **maintain** and **optimise** the model.

Maintenance: Monitor and maintain the model, archiving insights to facilitate future endeavours.

Optimisation: Regularly retrain the model with new data sources and make adjustments to improve performance.



Type of analytics

> The type of analytics we apply depends on our **goal** and prescribes our **approach** to the data analytics or data science process.

Descriptive

Hindsight

Used to describe **what** has happened in the **past**.

It's a summary of historical data that provides insights into patterns, trends, and relationships within the data

Examples: Dashboards and reports.

Diagnostic

Insight

Used to determine **why** something has happened in the **past**.

Helps organisations understand the factors that contributed to a particular outcome.

Examples: Data mining and drill-down analysis.

Predictive

Foresight

Used to forecast **what** will happen in the **future**.

Uses statistical models and machine learning algorithms to identify patterns and trends in historical data to predict future outcomes.

Examples: Forecasting and risk modelling.

Prescriptive

Context

Used to recommend the best **course of action** for a given situation.

Uses advanced algorithms and optimisation techniques to suggest the most optimal solution based on a variety of factors and constraints.

Examples: Optimisation