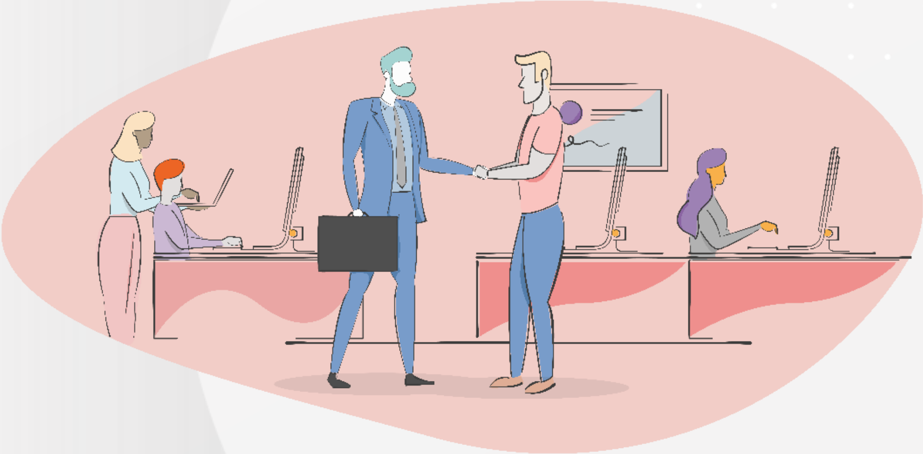
BPP

Fundamentals of the Data-Driven Enterprise



L5 Data Engineer Higher Apprenticeship
Module 1 / 12 (**“Data Fundamentals”**)
Topic 1 / 4

rev. 1 (2024)

Webinar agenda

This webinar will cover the following:

- Building a data-driven culture
- Fundamentals of data
- Standards and engineering best practices

Webinar length: 3 hours

Introductions

Building Careers Through Education

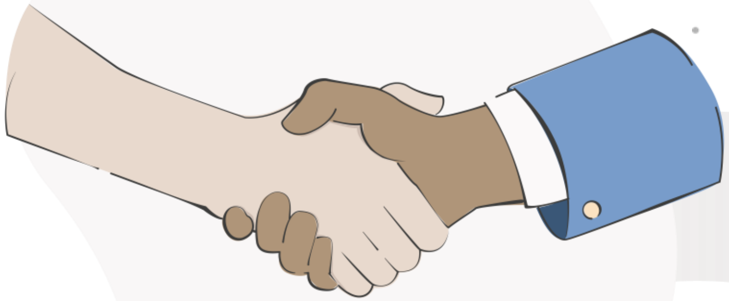
Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

We are very excited to have you starting your journey with Data Engineering!

To give everyone an idea of who you are, try and take a minute to answer the following questions:

- What is your name and what would you like to be called?
- Your personal Goals and Success Metrics for this programme (3-5 sentences)
- Your background (Is it technical? Non-technical? What projects have you collaborated on, if any?)
- Think about Data and Engineering, what makes you excited about these things?
- Anything else you'd like your class and your tutor to know about you?

Your tutor will also introduce themselves!



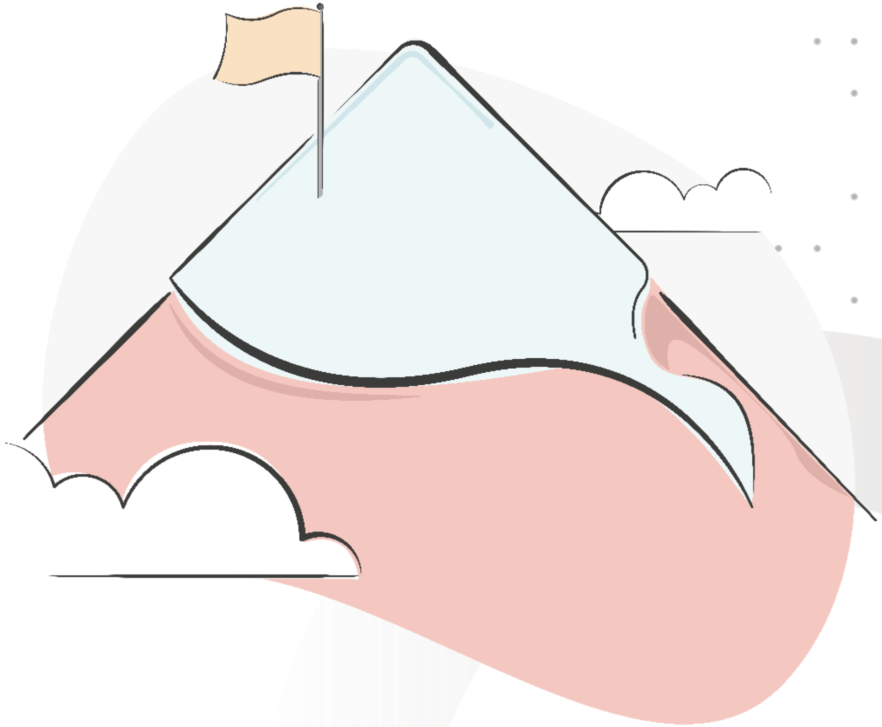
Welcome to the programme!

Learning objectives

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.

By the end of today’s webinar, you will be able to:

- Understand the **value** of data in modern organisations
- Become familiar with different **types** and **sources** of data
- Appreciate the significance of standards, best practices and **regulations**



Sounds like a lot? Don’t worry! We will provide real-world examples for each of the key concepts that you learn about today.

Data-Driven Enterprises

A real-world success story…

- Netflix leverages data analytics for user experience, content recommendations, and streaming optimisation
- Through analysis of user behavior and preferences, Netflix tailors content and predicts successful shows
- This data-driven approach fuels subscriber growth, cementing Netflix's dominance in streaming

***Netflix:** A Highly Successful Data-Driven Organisation*

The Vital Role of Data Engineering

Data engineering in action…

Harnessing the power of big data

Collecting, storing, and processing data at scale

Integrating data sources, and ensuring data quality

Driving innovation and growth

The role of data engineering and data engineers

Understanding Poll

Building Careers Through Education

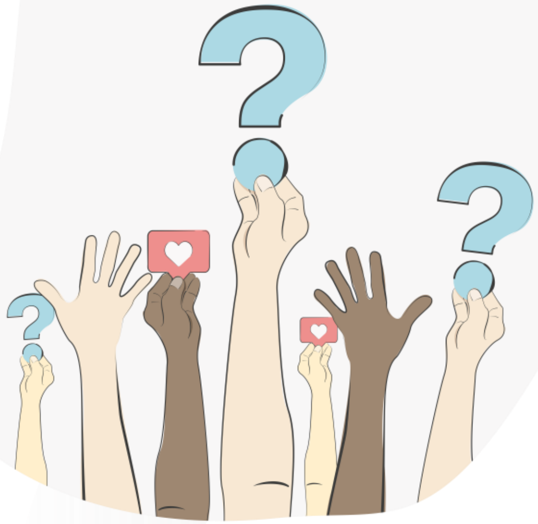
Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

The e-learning for this week’s topic covered a wide range of concepts.

So, are there any concepts you would like a further explanation or support with?

Concepts covered included:

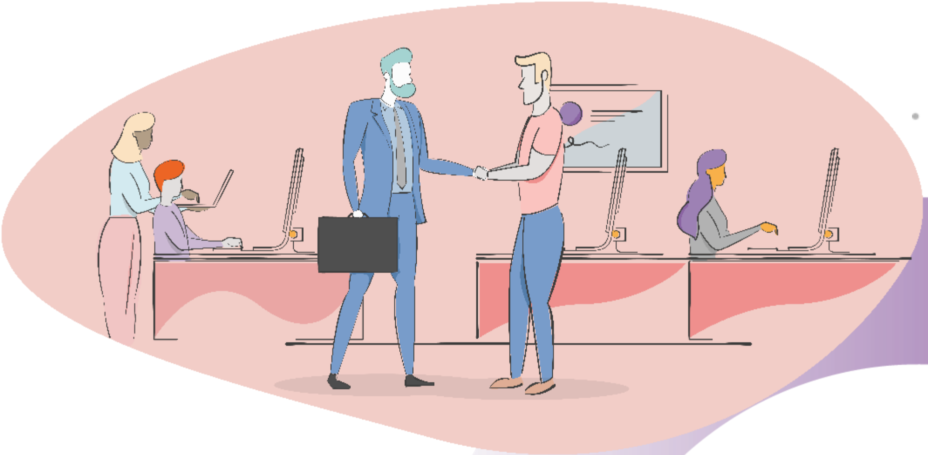
- What is a data-driven culture?
- What are the 5 Vs of Big Data?
- What is the value of Big Data?
- What are some examples of small data?
- What is the value of small data?



Submit your responses to the chat!

BPP

Building a Data-Driven Culture



Module Case Study

An introduction

- Credit Bank Corporation aimed to gain insights into employee performance, satisfaction, and HR processes
- The HR team planned to build a comprehensive analytics dashboard to track key metrics
- Data engineers must integrate diverse data and follow best practices
- The team must determine data types, build a suitable data ecosystem, visualise results, and articulate business value

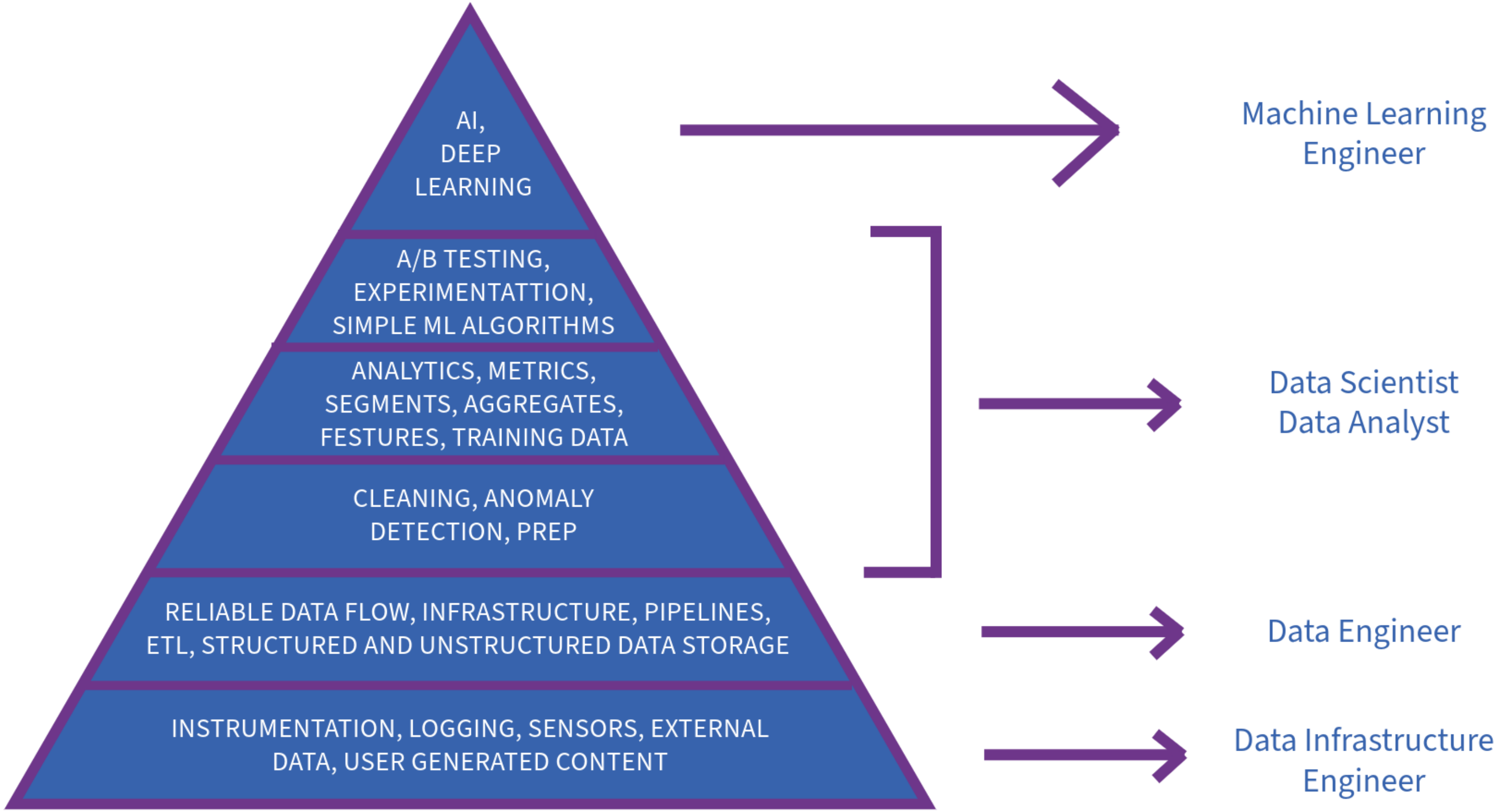
This 'Data Fundamentals' module will provide you with the knowledge and skills to overcome these challenges!

An analytics dashboard

The Data Science Hierarchy of Needs

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.



AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATTION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FESTURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

Machine Learning Engineer

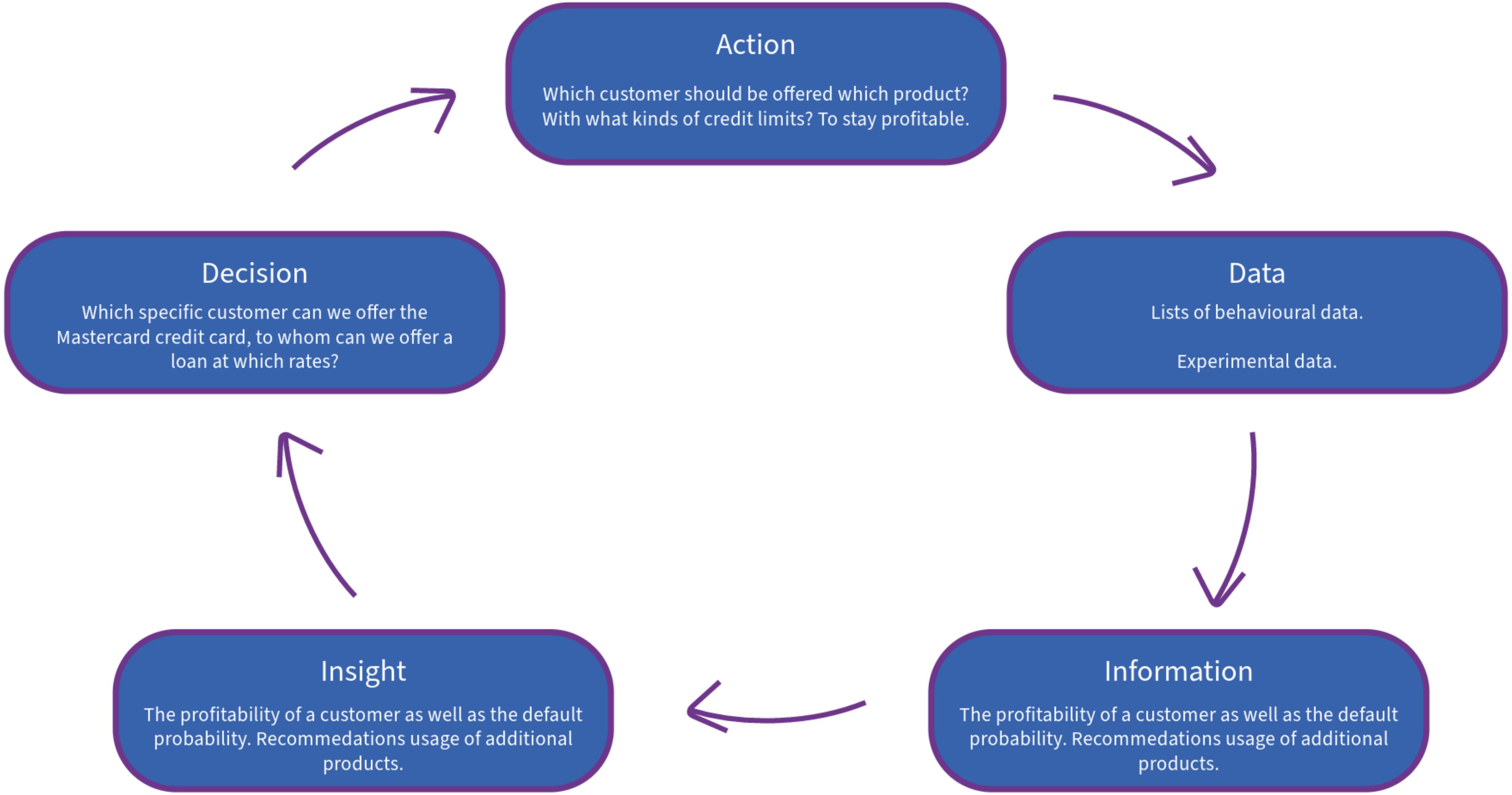
Data Scientist
Data Analyst

Data Engineer

Data Infrastructure Engineer

BPP

Turning Data into Actionable Insight



Action

Which customer should be offered which product? With what kinds of credit limits? To stay profitable.

Data

Lists of behavioural data.

Experimental data.

Information

The profitability of a customer as well as the default probability. Recommedations usage of additional products.

Insight

The profitability of a customer as well as the default probability. Recommedations usage of additional products.

Decision

Which specific customer can we offer the Mastercard credit card, to whom can we offer a loan at which rates?

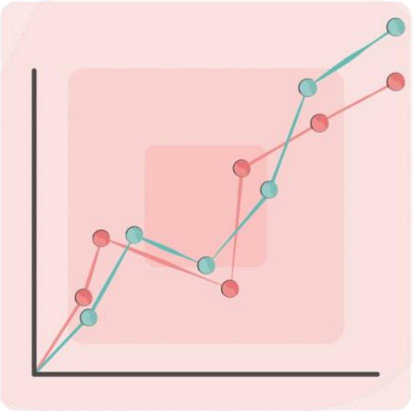
First Steps

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

What strategies should Credit Bank use to generate a data-driven culture?

Build relationships



Choose transparency in algorithms

Celebrate and embrace small wins

Raise data literacy

BPP

Transitioning from Small to Big Data

Why does this happen?

Normally, we'd expect the transition from small to big data to happen because of the following:

- Overwhelming Data
- Performance Bottlenecks
- Missed Opportunities

Big Data

Defining the Extremes

Can you identify small and big data?

Data Examples A
Personal fitness tracker recording daily steps and calories burned
A collection of contacts in a mobile phone’s address book
A simple spreadsheet tracking monthly expenses

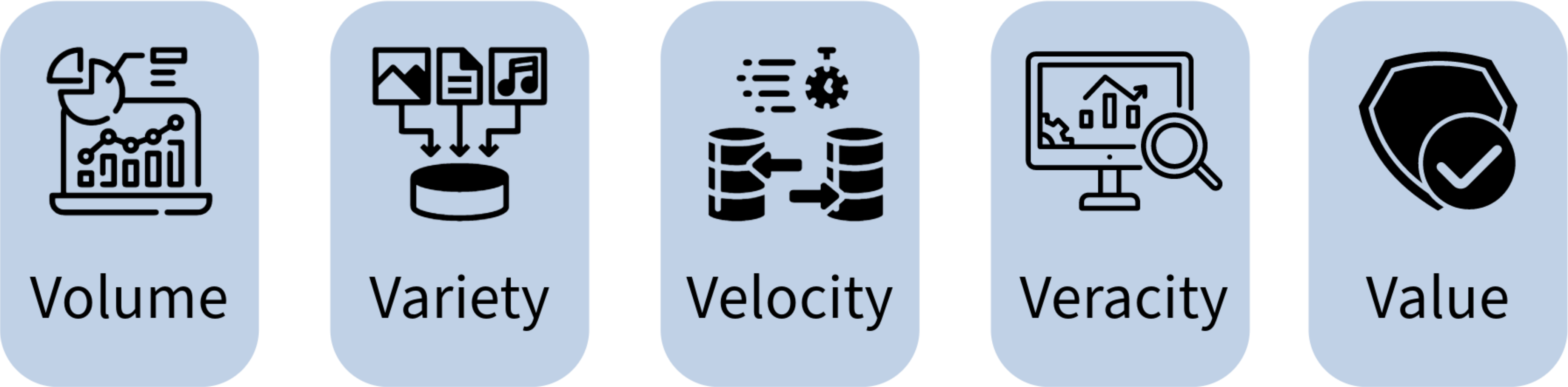
Data Examples B
Social media platforms analysing billions of user interactions daily
Weather monitoring stations collecting vast amounts of climate data worldwide
Online retailers tracking millions of transactions and customer behaviors

***Small Data:** like a crystal-clear lake*

Big Data: like an ocean

The Characteristics of Data

The 5 Vs of data



Volume

Variety

Velocity

Veracity

Value

The 5 Vs of data framework

Small Data vs Big Data

In the context of **volume…**

Big Data	Small Data
Big data refers to the sheer quantity of data generated, stored, and processed	Small data datasets are relatively modest in size, often measured in gigabytes or smaller
Big data datasets are typically massive, ranging from terabytes to petabytes and beyond	They can be managed using conventional storage and processing techniques without overwhelming computational resources
This abundance of data presents challenges in storage, management, and analysis	

Volume

Variety

Velocity

Veracity

Value

Case study: *Credit Bank Corporation's data analytics dashboard*

Small Data vs Big Data

In the context of **variety…**

Big Data	Small Data
Big data encompasses diverse data types and sources, including structured, semi-structured, and unstructured data	Small data consists of homogeneous data with limited variety
Big data datasets often originate from multiple sources like social media, resulting in varied data formats and structures	Small data datasets are often well-structured and uniform, originating from a single source or system

Volume

Variety

Velocity

Veracity

Value

Case study: *Credit Bank Corporation's data analytics dashboard*

Small Data vs Big Data

In the context of **velocity…**

Big Data	Small Data
Big data refers to the speed at which data is generated, collected, and processed	In small data environments, data is generated and processed at a slower pace
Big data streams in rapidly from various sources	Batch processing and periodic analysis are common practices in small data environments
Real-time or near-real-time processing and analysis are necessary to extract actionable insights	

Volume

Variety

Velocity

Veracity

Value

***Case study:** Credit Bank Corporation’s data analytics dashboard*

BPP

Knowledge Check Poll

In the context of Credit Bank Corporation's HR analytics dashboard project…

Which of the following best represents the "volume" characteristic of big data?

A) The diverse data sources used, such as employee performance data, HRIS data, and engagement survey responses

B) The speed at which data is generated and processed from various HR systems and applications

C) The large and growing amount of employee data being collected and processed

D) The accuracy and reliability of the data used in the dashboard

Submit your responses to the chat!

Feedback: C – The large and growing amount of employee data being collected and processed.

Small Data vs Big Data

In the context of **veracity…**

Big Data	Small Data
Big data concerns the reliability, accuracy, and trustworthiness of data	Small data generally exhibits higher data quality and veracity
Big data datasets may include noisy, incomplete, or erroneous data due to their volume and diversity	Small datasets are more manageable and easier to validate and clean
Ensuring data quality and integrity poses a significant challenge in big data environments	This leads to increased reliability, accuracy, and trustworthiness of the data in small data environments

Volume

Variety

Velocity

Veracity

Value

***Case study:** Credit Bank Corporation's data analytics dashboard*

Small Data vs Big Data

In the context of **value…**

Big Data	Small Data
Big data offers valuable insights and business value through advanced analytics	Small data delivers immediate value through straightforward analysis
It uncovers hidden patterns, trends, and correlations	It addresses specific operational questions efficiently
This enables predictive modeling and data-driven decision-making	Small data supports day-to-day decision-making processes effectively
Big data drives innovation and competitive advantage	Small data supports day-to-day decision-making processes effectively

Volume

Variety

Velocity

Veracity

Value

Case study: *Credit Bank Corporation's data analytics dashboard*

Knowledge Check Poll

Which of the following best represents the "veracity" characteristic of big data in the context of Credit Bank Corporation's HR analytics dashboard project?

A) The speed at which employee data is generated from different HR systems

B) The accuracy and trustworthiness of the employee data collected

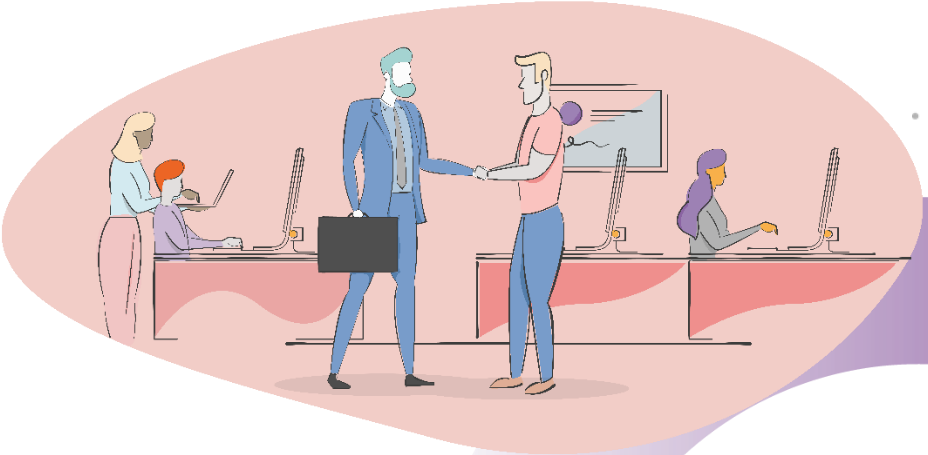
C) The diverse range of data sources used

D) The large size of the employee data being used

Feedback: B - The accuracy and trustworthiness of the employee data collected.

Submit your responses to the chat!

Fundamentals of Data



Section Introduction

Appreciating data sizes

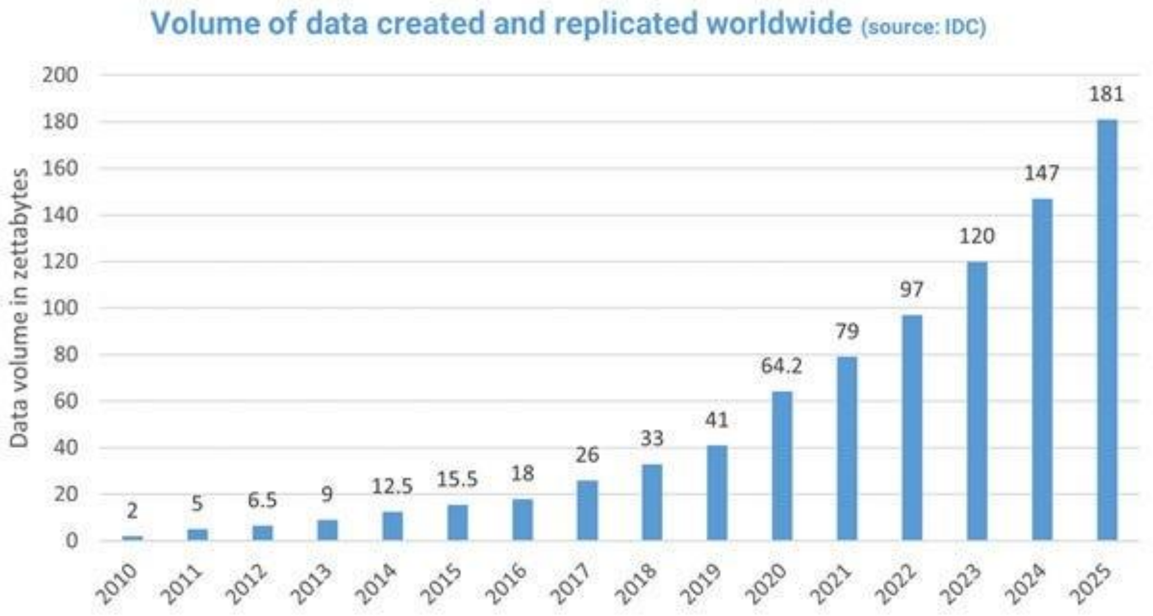
As we live in an increasingly data-driven world, it's important to understand how we quantify the vast amounts of data being generated and stored.

Modern data is growing rapidly, every minute:

500 thousand tweets are sent

500 million instant messages are sent

5 terabytes of data is posted on Facebook



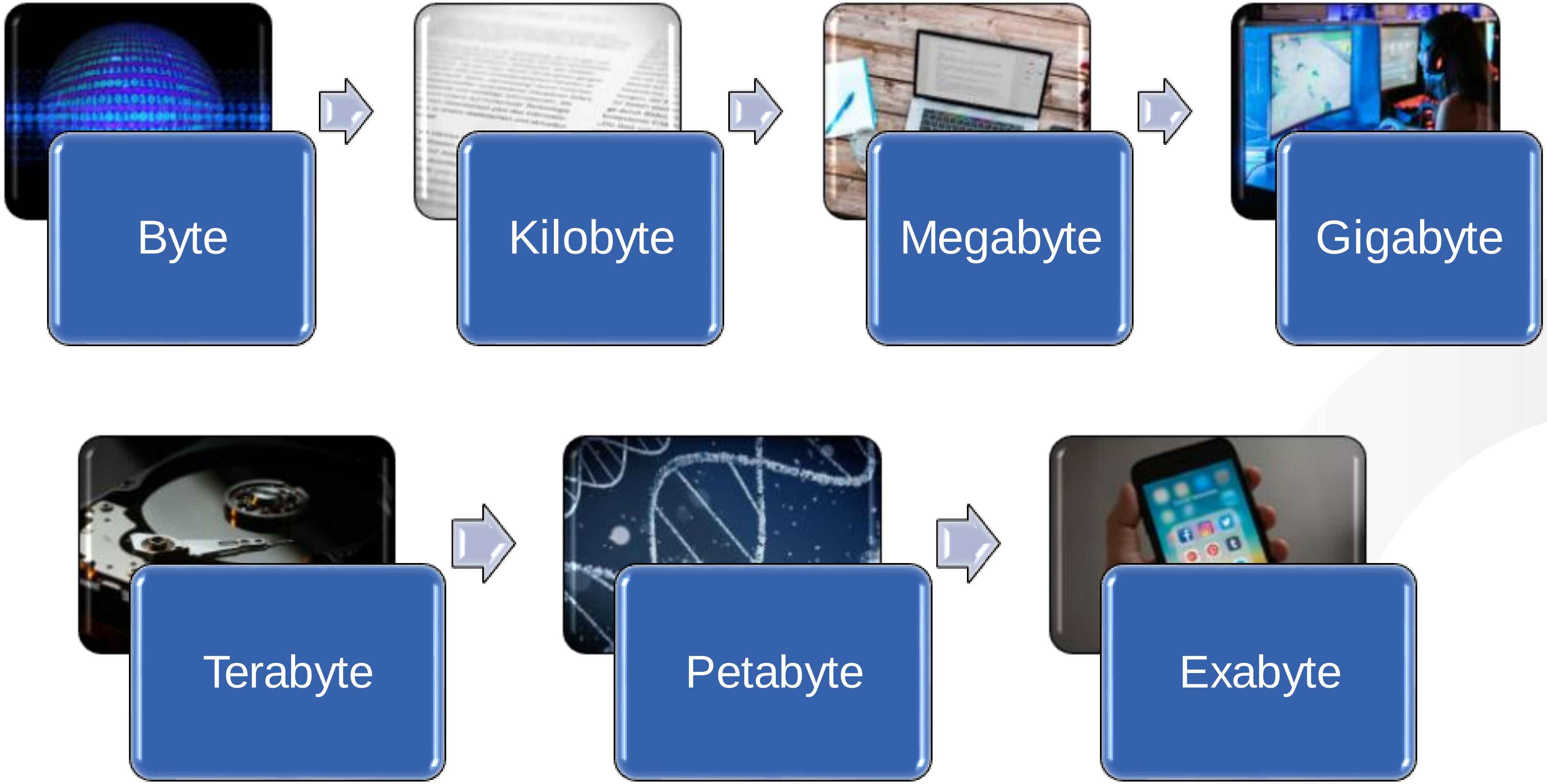
Modern data is growing rapidly!

Image source: *Medium.com*

BPP

Appreciating Data Sizes

From Byte to Exabyte…



Byte

Kilobyte

Megabyte

Gigabyte

Terabyte

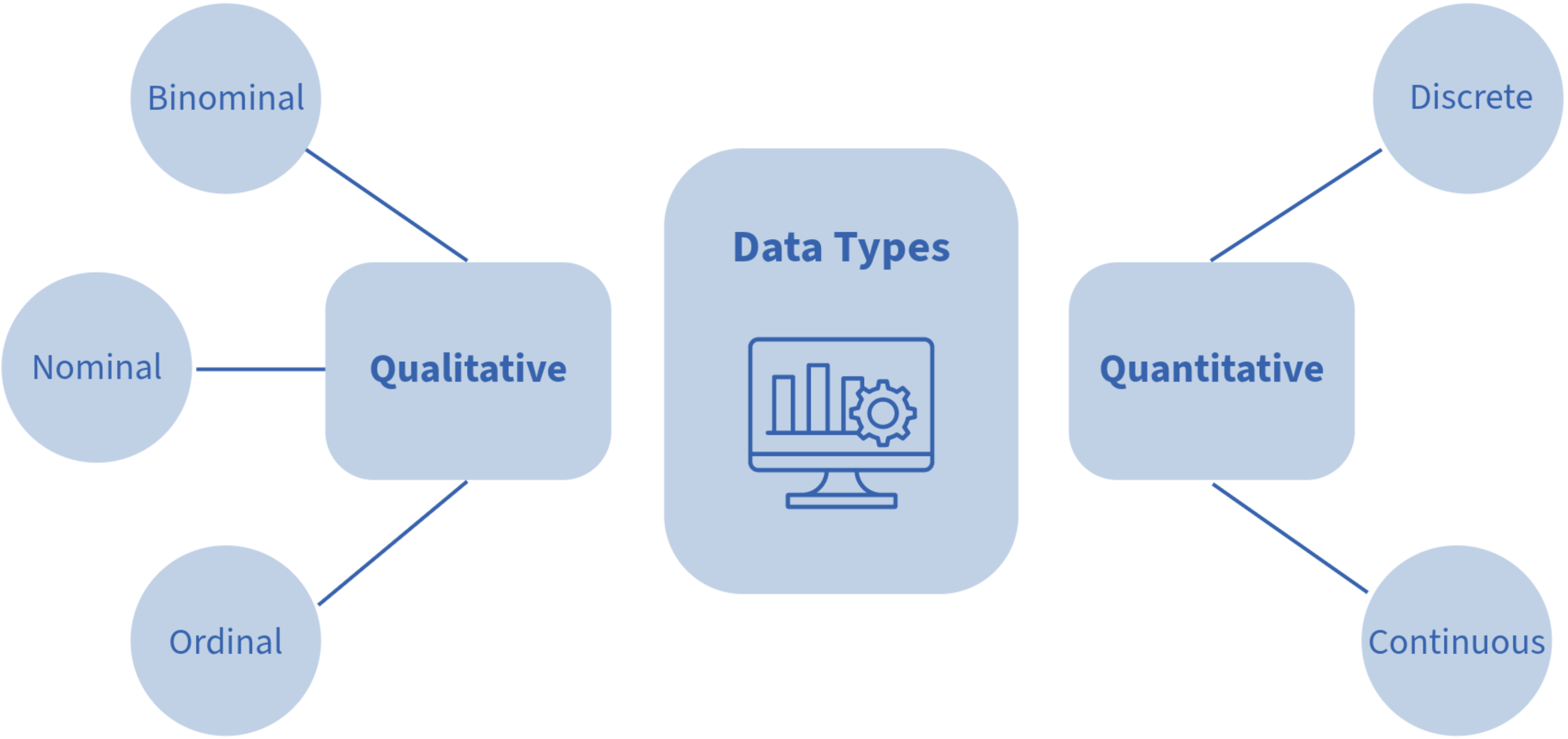
Petabyte

Exabyte

Data sizes and everyday examples from Byte to Exabyte

The Different Types of Data

A hierarchy of two main groups…



The Hierarchy of Qualitative and Quantitative Data Types

Knowledge Check Poll

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

If a dataset containing employee performance reviews from the past 10 years is approximately 50 Gigabytes (GB) in size…

What would be the next larger data size unit that could represent this dataset?

A) Megabyte (MB)

B) Terabyte (TB)

C) Petabyte (PB)

D) Exabyte (EB)

Submit your responses to the chat!

Feedback: B – Terabyte (TB)

BPP

Knowledge Check Poll

Which type of data would be most suitable for tracking employee attendance in Credit Bank Corporation's HR analytics dashboard?

A) Qualitative data (nominal)

B) Qualitative data (ordinal)

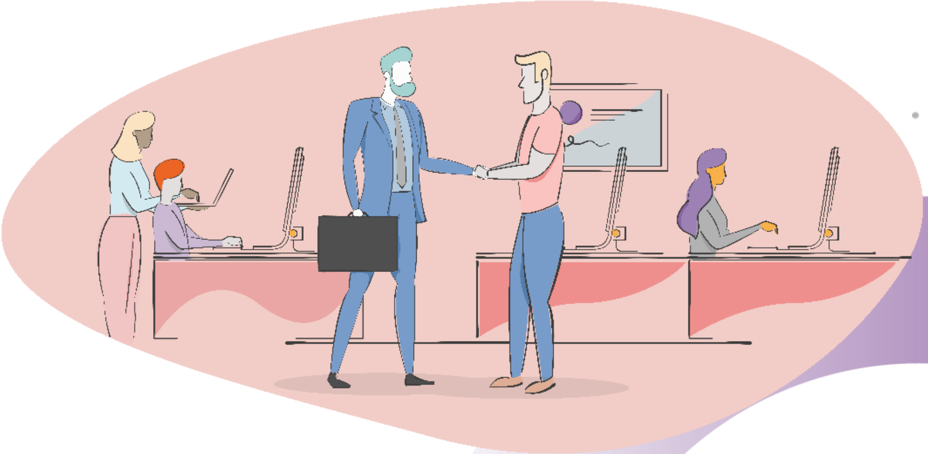
C) Quantitative data (discrete)

D) Quantitative data (continuous)

Feedback: C – Quantitative data (discrete).

Submit your responses to the chat!

Identifying Standards and Engineering Best Practice



Section Introduction

Identifying standards and best practice

Data Formats

API Standards

Cloud Computing Standards

Technological Standards and Engineering Practices

Regulatory Requirements

Data Stewardship Principles

Engineering Best Practices

Data Formats

JSON, CSV and XML…

JSON

Social Media

CSV

Financial Institutions

XML

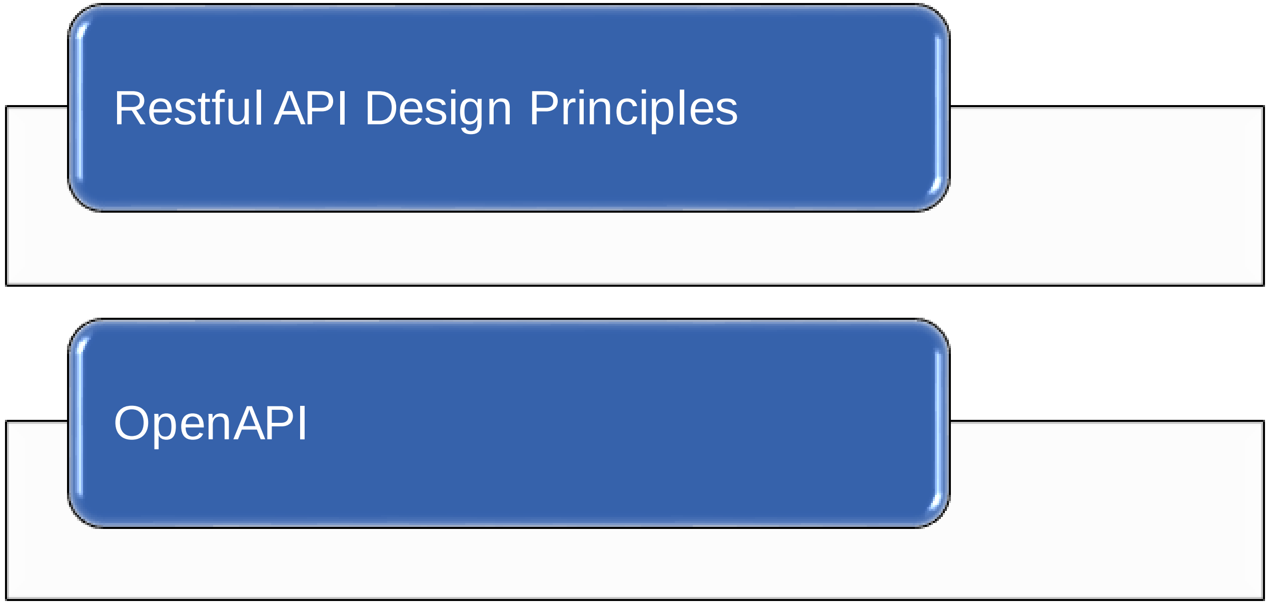
Healthcare

API Standards

And the impact on business…

APIs are the foundational building blocks enabling interconnected digital systems and data exchange.

As businesses increasingly rely on APIs for digital experiences, innovation, and third-party integration, standardisation is critical.



Restful API Design Principles

OpenAPI

The impact of API standards example:

E-commerce

Cloud Computing Standards

And the impact on business…

As cloud adoption accelerates, standards and best practices ensure consistent, reliable, and secure cloud implementations.

We will now explore two prominent cloud computing standards:

- AWS Well-Architected Framework
- OpenAPI

The impact of API standards example:

Healthcare

Regulatory Requirements

Privacy, security and compliance…

General Data Protection Regulations (GDPR)

Information Security Management System (ISO) 27001

Engineering Best Practices

Vital for developing robust, scalable, and dependable systems…

Scalability

Engineering Best Practices

Reliability

Security

Performance Optimisation

Data Documentation

Data Stewardship Principles

Quality, governance and ethics

As data continues to proliferate exponentially, establishing a robust framework for data stewardship becomes imperative.

There are three fundamental principles of responsible and effective data stewardship, as follows:

Data quality

Data governance

Data ethics

Knowledge Check Poll

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Credit Bank Corporation is planning to migrate its HR analytics dashboard to a cloud-based platform for better scalability and accessibility.

Which of the following cloud computing standards would be most relevant for the data engineering team to follow during this migration?

A) OpenAPI Specification

B) GDPR

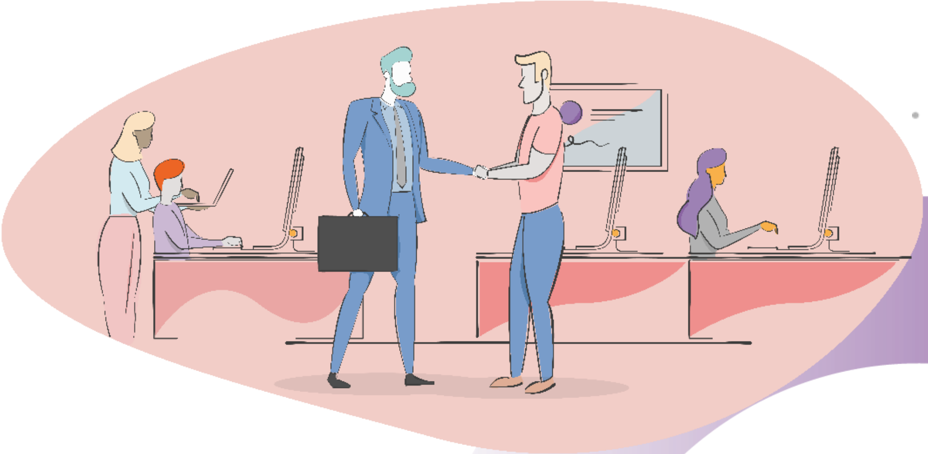
C) AWS Well-Architected Framework

D) HIPAA

Feedback: C - AWS Well-Architected Framework.

Submit your responses to the chat!

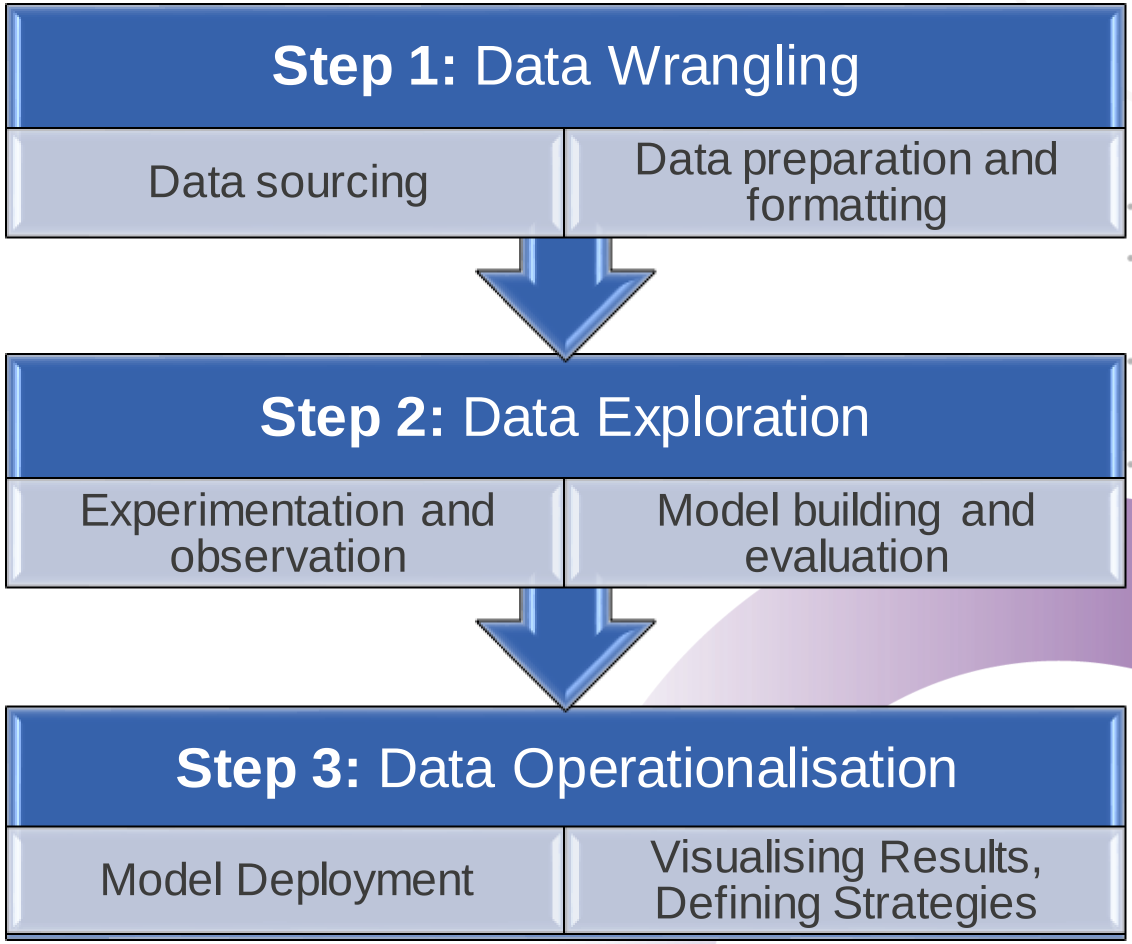
Adding Value and Making Decisions with Data



Section Introduction

The three stages of adding value with data

- The tangible results (deliverables) of adding value with data are Data Products and Services
- Data Engineering is primarily active during Data Wrangling and Data Operationalisation
- Data Science and Data Analytics are mainly concerned with Data Exploration



Step 1: Data Wrangling

Data sourcing

Data preparation and formatting

Step 2: Data Exploration

Experimentation and observation

Model building and evaluation

Step 3: Data Operationalisation

Model Deployment

Visualising Results, Defining Strategies

The three stages of adding value with data

Data Teams

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.

Who do they include?

Software Engineer

Data Teams

Chief Data Officer

Data Scientist

Data Engineer

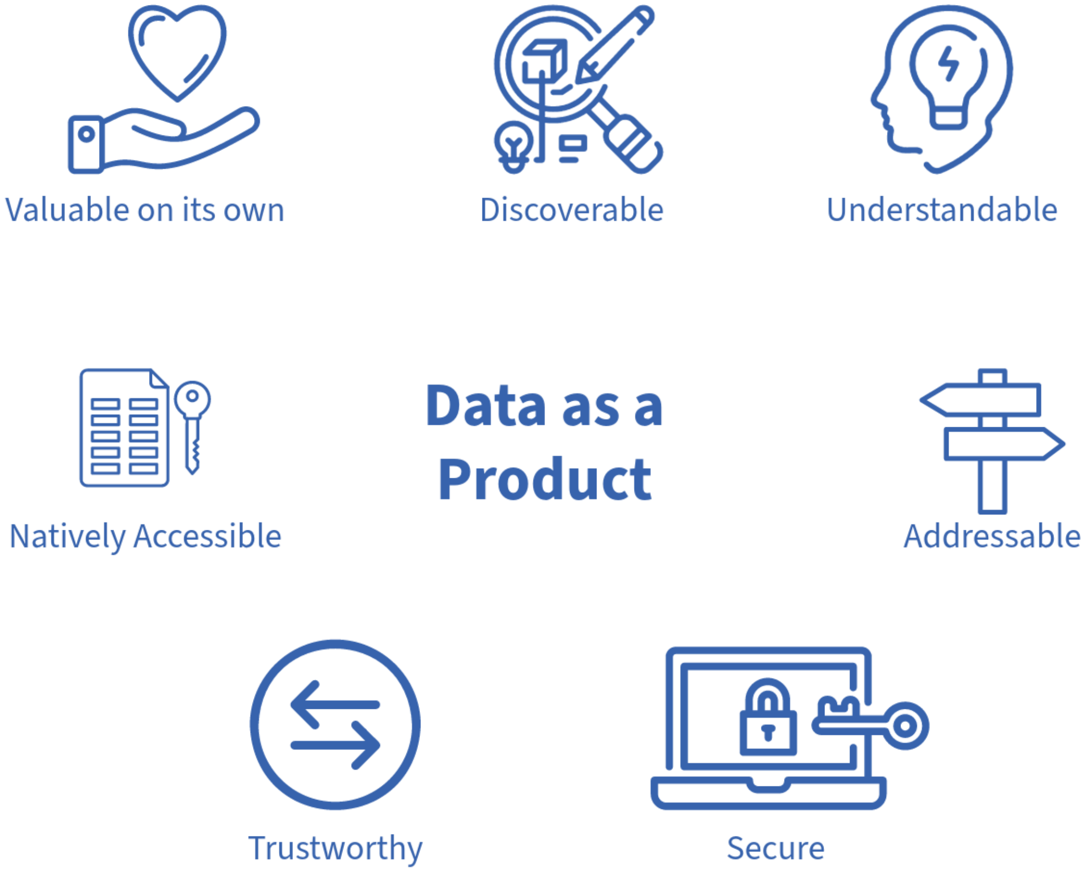
Data Analyst

BPP

Data as a Product

A product in itself and by itself…

- Data can be monetised directly or indirectly
- Data brokers collect and sell aggregated data for targeted advertising
- Companies like Google and Facebook monetise user data through advertising
- Data engineers transform raw data into valuable products
- They ensure data is valuable, accessible, trustworthy, discoverable, and interoperable



Valuable on its own

Discoverable

Understandable

Natively Accessible

Data as a Product

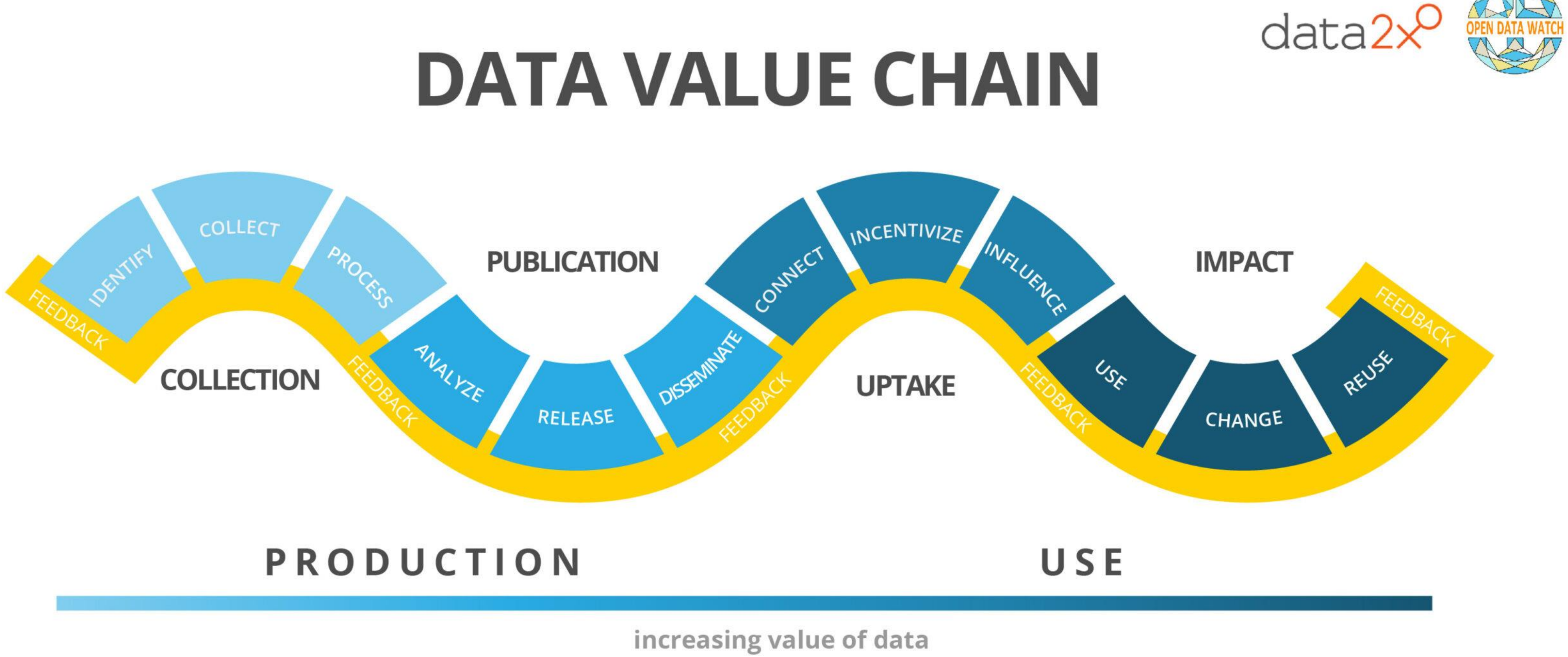
Addressable

Trustworthy

Secure

Data Value Chain

From collection to impact…



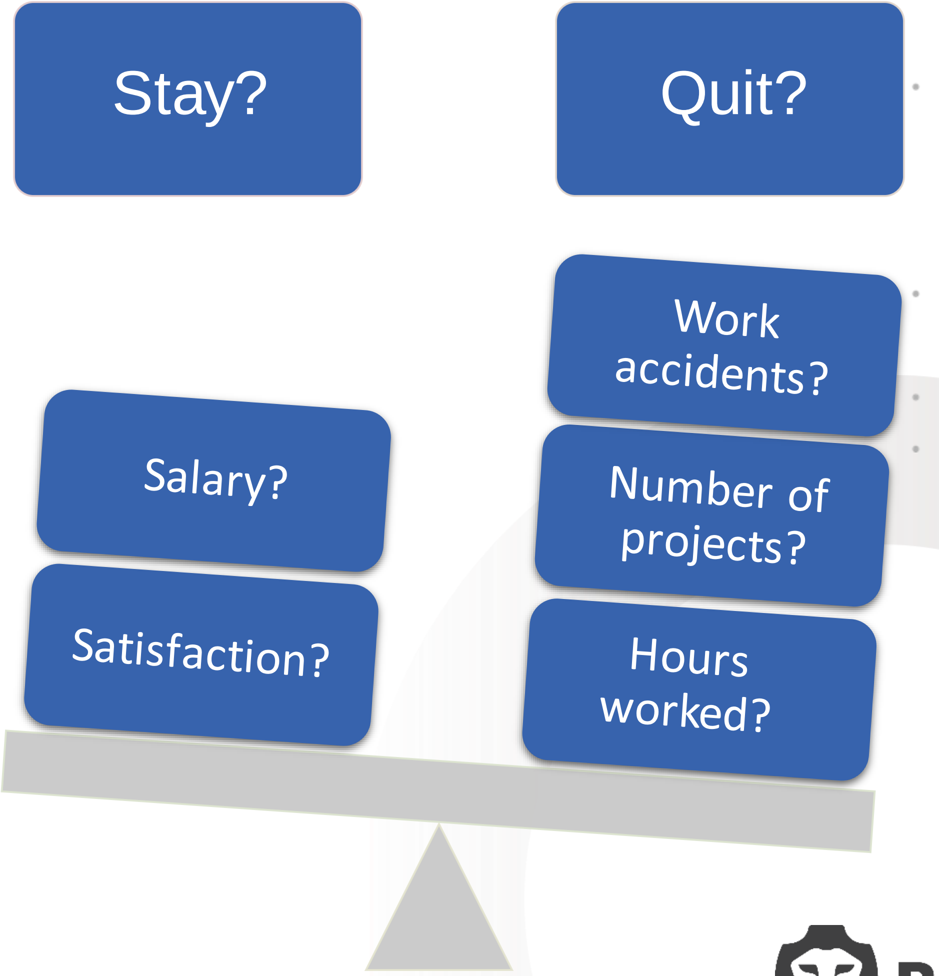
Data as a product

Image source: *Open Data Watch*

Staff Retention Example

Will Peter stay with his employer…?

Datapoint	Value
Name:	Peter
Satisfaction Level:	0.80
Last Evaluation Score:	0.86
Number of Projects:	5
Average Monthly Hours	262
Time with Company (Years)	6
Work Accidents	0
Promotions in Last 5 Years	0
Department	Sales
Salary	£45,000
Did he resign?	?



Stay?

Quit?

Salary?

Satisfaction?

Work accidents?

Number of projects?

Hours worked?

Building Careers Through Education

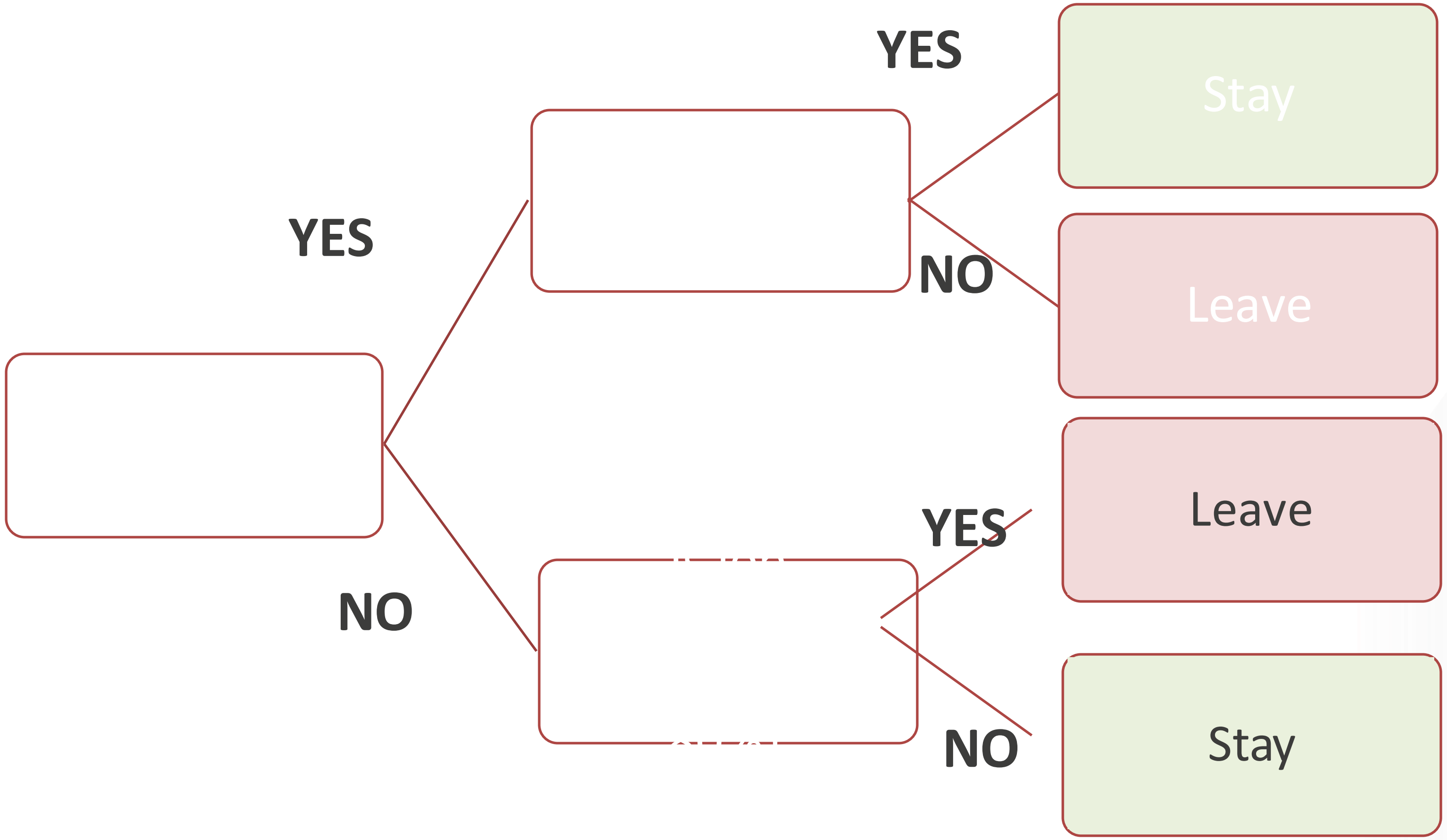
Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

BPP

Decision Trees

An introduction…

Decision trees are machine learning algorithms that represent decision-making processes in a flowchart-like structure, facilitating clear and interpretable understanding of outcomes.

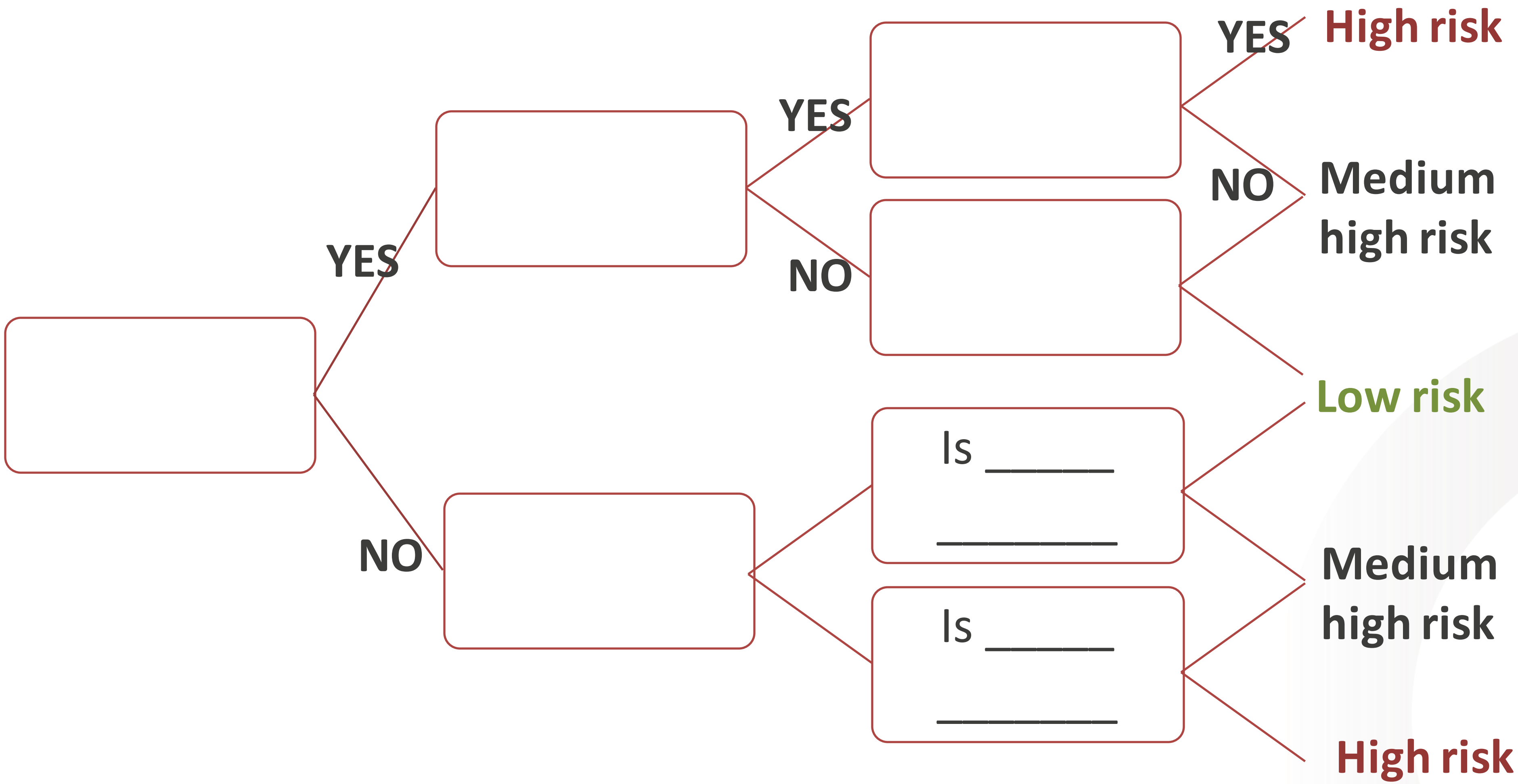


An example decision tree

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

Decision Trees

The deep decision tree model…



An example deep decision tree model

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

Building a Decision Tree

Group practice

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Building an Employee Churn Decision Tree

Objective: To illustrate how actionable insights are generated from data.

Brief: Working in small groups, identify the most useful factors for HR employee churn. Try to build a decision tree that classifies employees into risk groups based on these factors.

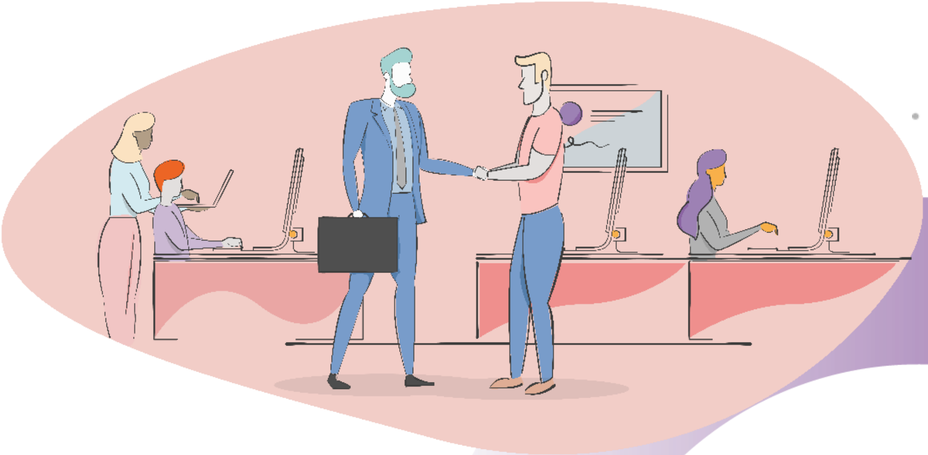
Final output:

- By completing this exercise you should produce a decision tree that can be tested on new (unseen) data to deliver insights
- You should also be able to develop a strategy of next steps based on the decision tree

Group practice

BPP

Practical Lab: Working With Data



Integrating Diverse Data Sources

Group practice

We must now imagine we are part of a team of data engineers within Credit Bank Corporation tasked with integrating employee performance data from various sources into the HR analytics dashboard. The data sources include:

- Employee performance reviews (CSV file)
- Learning Management System (LMS) data (SQL database)
- Employee engagement survey responses (JSON file)

Instructions:

1. Collect data from provided sources
2. Clean and preprocess data for quality and consistency
3. Transform data into suitable format for analysis
4. Integrate transformed data into unified dataset

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Group practice

Brief available here

BPP

Key Learning Summary

The key takeaways from this session are as follows:

- Building a data-driven culture involves strategies like building relationships, choosing transparency in algorithms, celebrating small wins, and raising data literacy

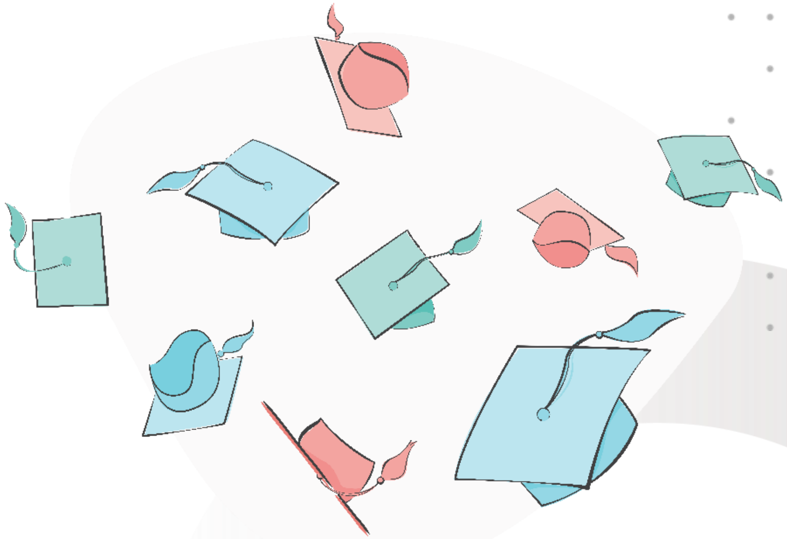
- The 5 Vs of Big Data are Volume, Variety, Velocity, Veracity, and Value

- Data types are categorised into two main groups: Qualitative (nominal and ordinal) and Quantitative (discrete and continuous)

- Appreciating different data size units, from bytes to exabytes, is crucial for quantifying modern data volumes

- API standards like RESTful design principles and OpenAPI Specification enable interconnected systems and data exchange

- Cloud computing standards like the AWS Well-Architected Framework ensure consistent, reliable, and secure cloud implementations



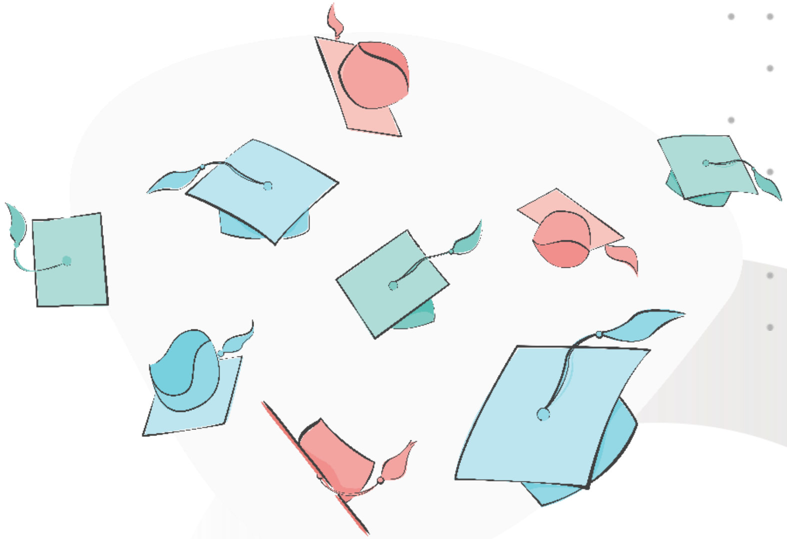
Key Learning Summary (Cont)

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

The key takeaways from this session are as follows:

- Regulatory requirements like GDPR and ISO 27001 govern data privacy, security, and compliance
- Data stewardship principles include data quality, data governance, and data ethics
- Adding value with data involves three stages: data wrangling, data exploration, and data operationalisation
- Data engineering plays a vital role in the data wrangling and data operationalisation stages
- Data can be monetised directly or indirectly, treated as a product itself
- Decision trees are machine learning models that represent decision-making processes in a flowchart structure, facilitating interpretable understanding of outcomes



BPP

Thank you

Do you have any questions, comments, or feedback?