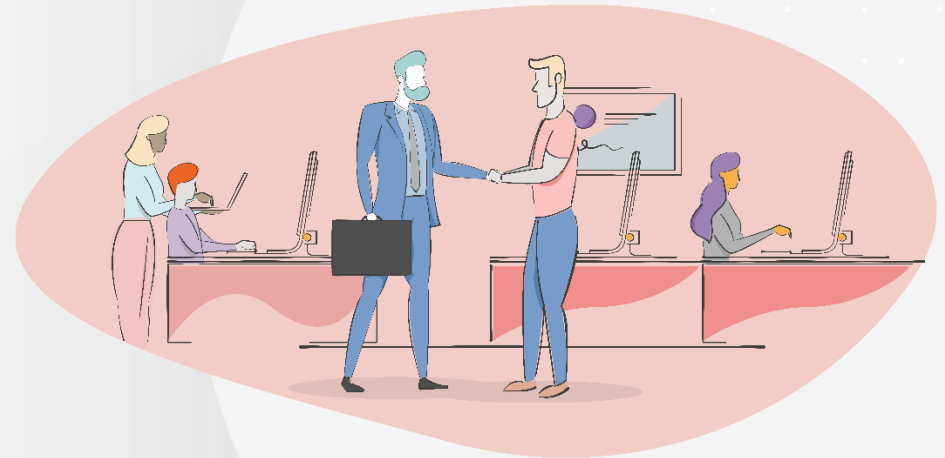


# Fundamentals of the Data-Driven Enterprise



**L5 Data Engineer Higher Apprenticeship**

**Module 1 / 12 (“Data Fundamentals”)**

**Topic 1 / 4**

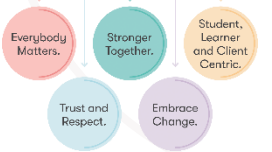
# Webinar agenda

This webinar will cover the following:

- Settling into your new programme
- Building a data-driven culture
- Fundamentals of data
- Standards and engineering best practices

**Webinar length:** 3 hours

Building Careers  
Through Education



# Introductions

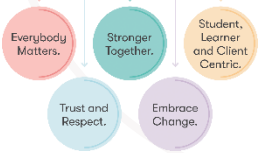
We are very excited to have you starting your journey with Data Engineering!

You have already introduced yourselves during the Induction, but let's continue getting to know one another!

- Intros and catch-up (2 minutes)
- What exciting data projects or products have you worked on (or would like to work on)? (3 minutes)
- Your two Goals and Expectations from this programme (2 minutes)
- How are you managing your time for e-learning? (3 minutes)

**Share your experiences!**

Building Careers  
Through Education



**Welcome to the programme!**



# Time management

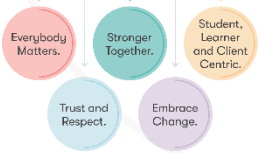
Your tutor will now re-cap the main points that you discussed re: time management.

Do you want to add anything else at this point, or ask any more questions about the time expectations for this course?

Think about:

- are you a morning learner or an evening learner?
- do you need quiet focus or more stimulation?
- how do you communicate your time and space needs at work?
- do you want to track extra learning if you're aiming for a distinction?

Building Careers  
Through Education



# Hub, Programme Handbook and Tracking

Your tutor will remind you how to access the Hub, and will show you your forum, and your Skills Scan (I).

There will be a post on the forum with your submission deadlines for your programme's formatives and portfolio pieces.

Your tutor will also remind you how to access the Programme Handbook via the Hub, and you will explore some sections of it together.

Pay attention to how you can track extra learning activities such as reading, shadowing, internal research, and stretch/extend activities.

Building Careers  
Through Education



# KSBs

You will now review the KSBs for this programme.

These are in your Programme Handbook.

After you come back to the main room, discuss which KSBs did you find the most surprising, or perhaps even unclear at this point?

Remember you are at the very beginning of your journey, and things will start making more sense very soon!

Building Careers  
Through Education



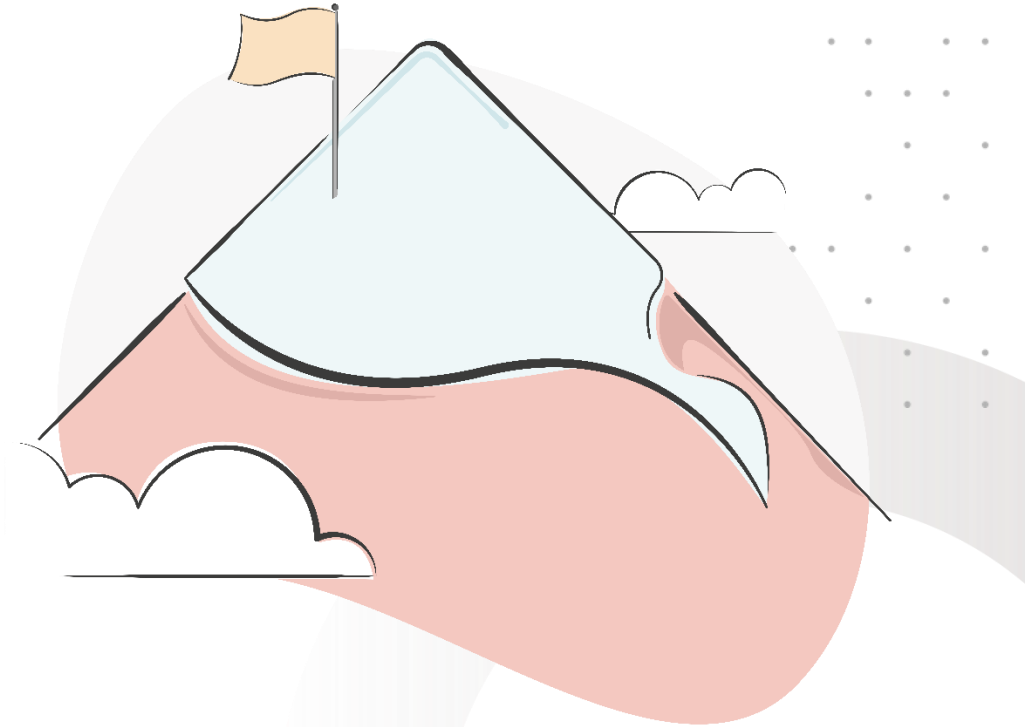
# Learning objectives

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.

Building Careers  
Through Education



# 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*

Building Careers Through Education



# Understanding Poll

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?

Your tutor will now **whiteboard** with you the elements of a data-driven culture, such as:

Clean, usable data

Governance and quality

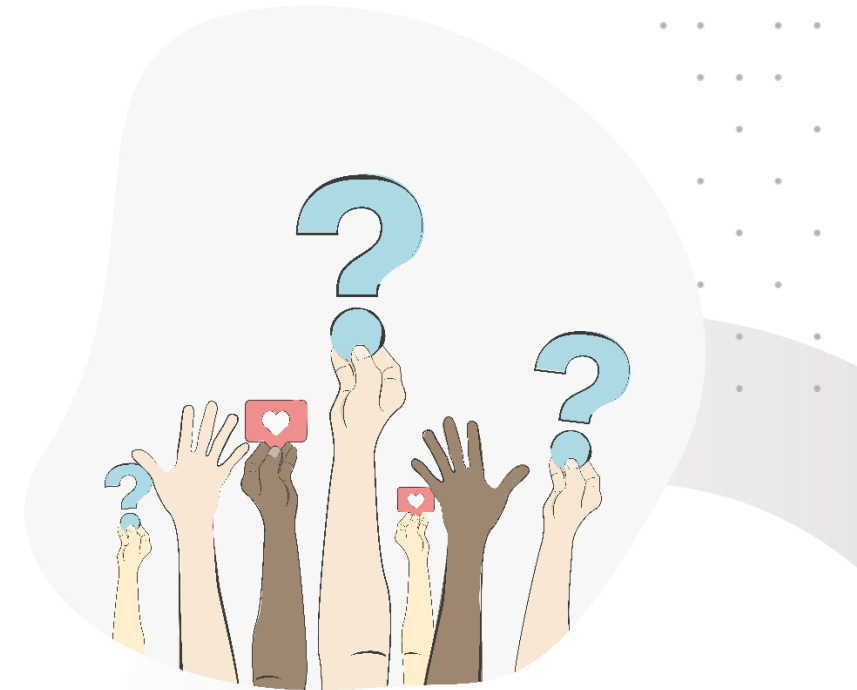
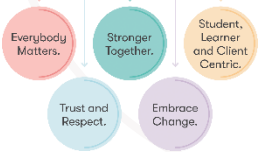
Data stewardship and literacy

Collaboration and stakeholder buy-in

Data Strategy and Executive support

What else can you remember / come up with?

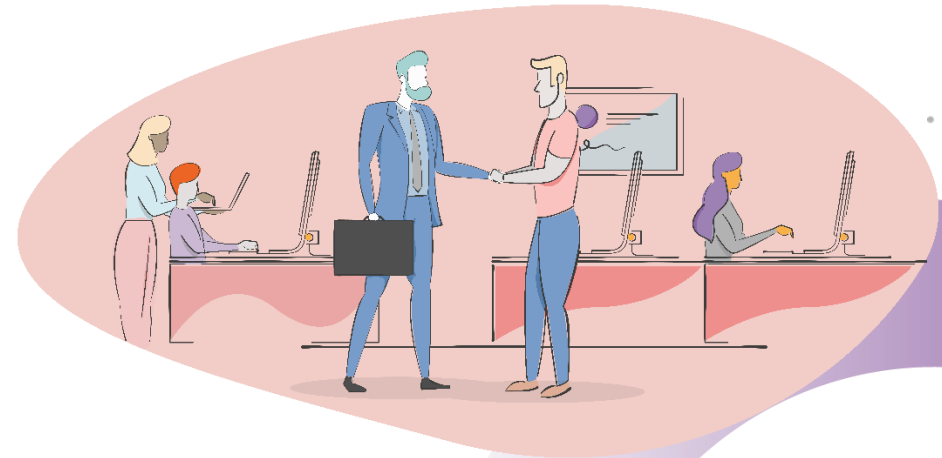
Building Careers  
Through Education



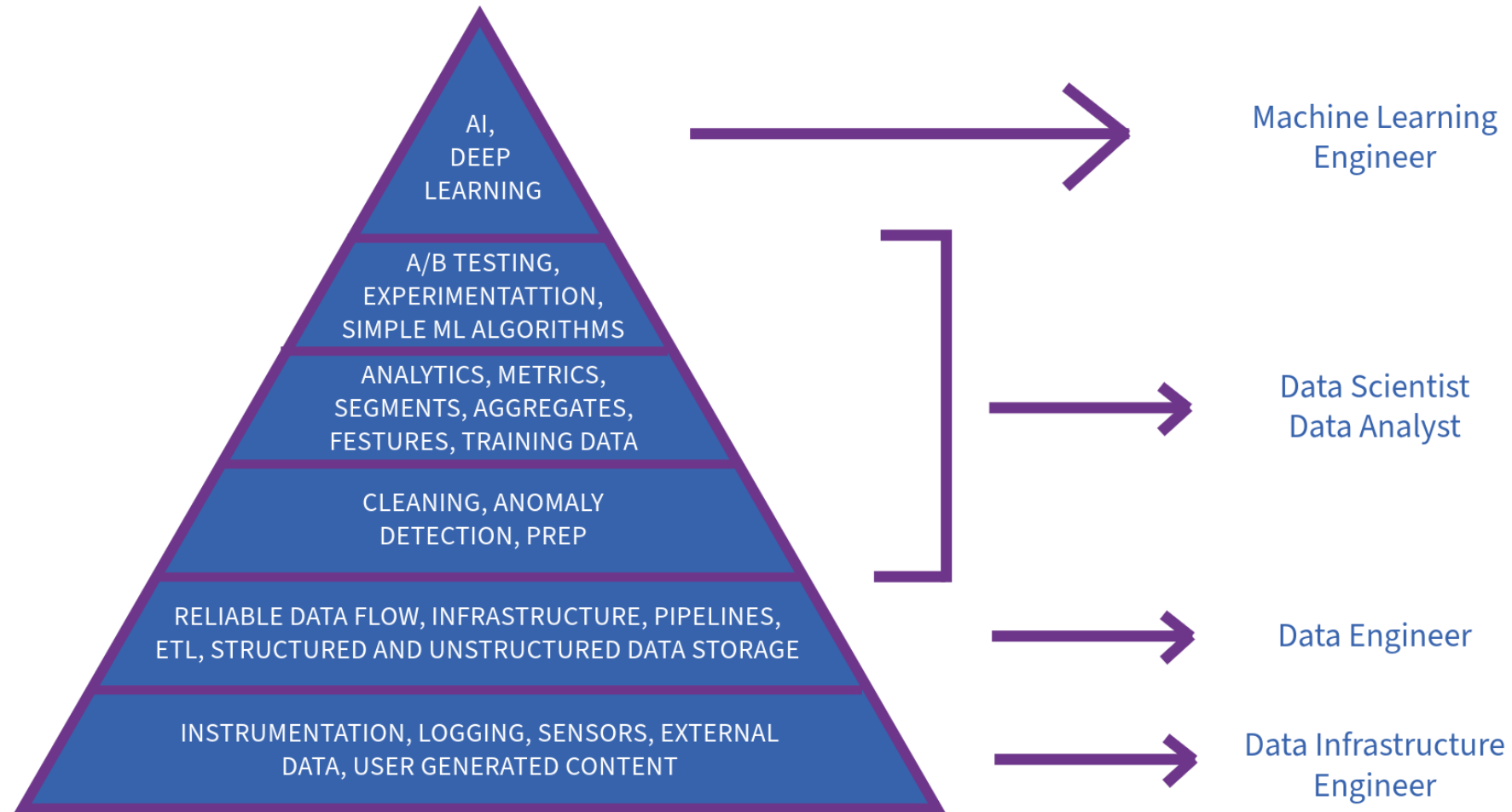
**Submit your responses to the chat!**



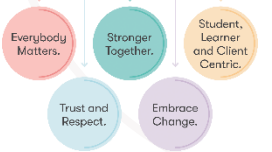
# Building a Data-Driven Culture



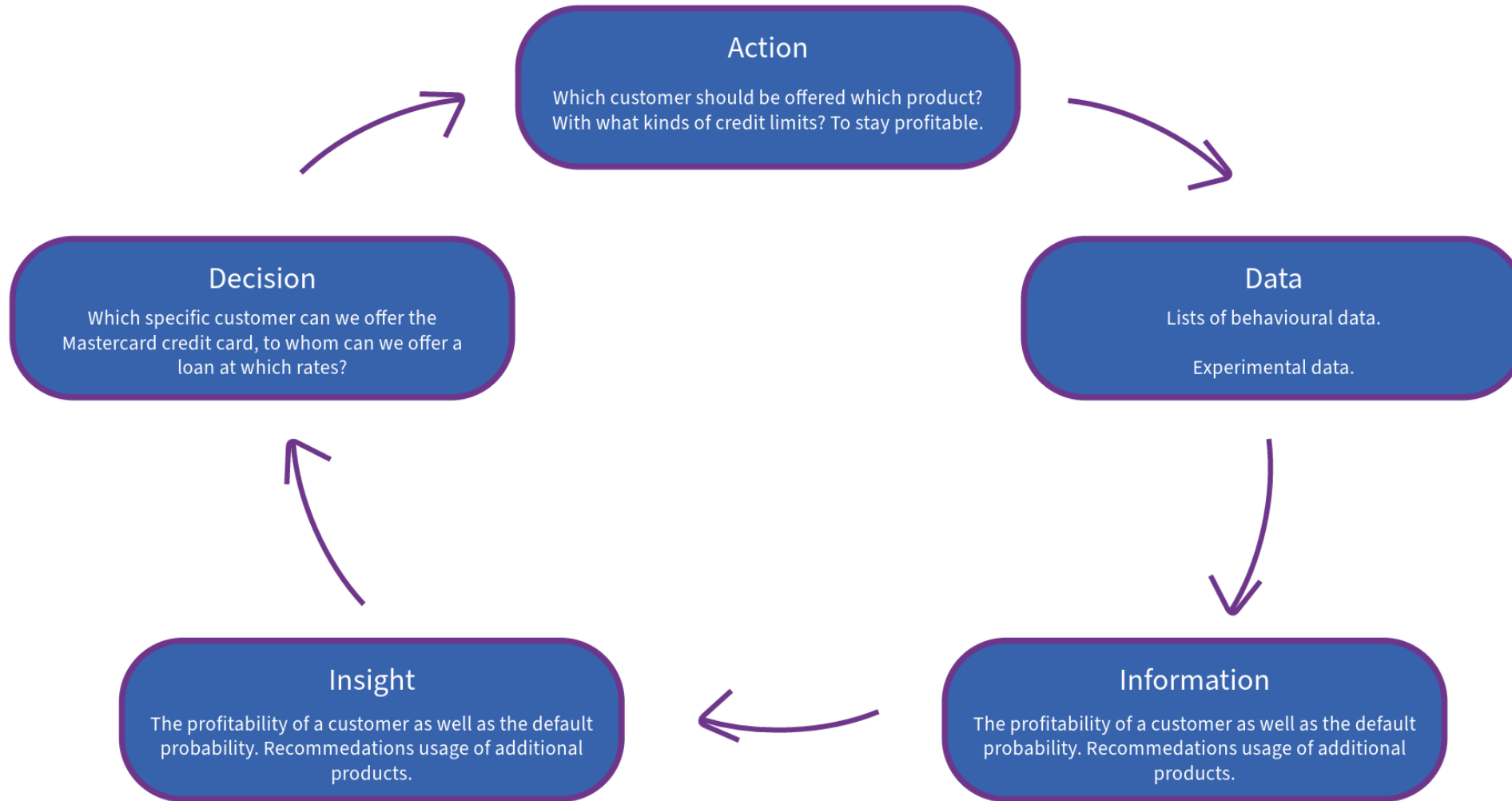
# The Data Science Hierarchy of Needs



Building Careers Through Education



# Turning Data into Actionable Insight

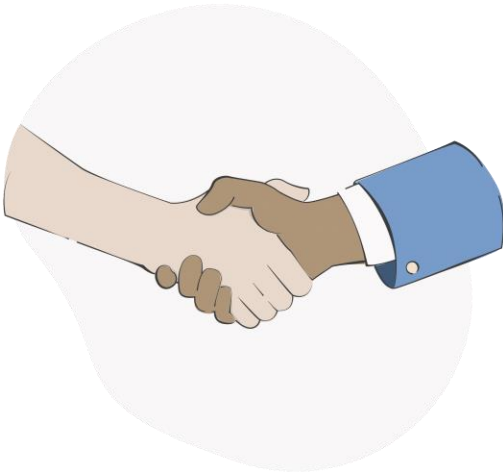


Building Careers  
Through Education

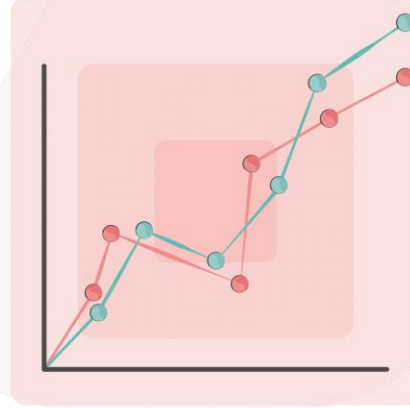


# First Steps

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



*Build relationships*



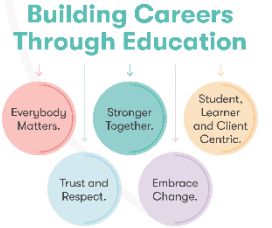
*Choose transparency in algorithms*



*Celebrate and embrace small wins*



*Raise data literacy*



# Learning journal - reminder

Your tutor will remind you about the importance of your learning journal and point you to the section in the Handbook that discusses it.

You will be also shown where the post-webinar activity is on the Hub, which will help you set up your first learning journal on GitHub.

Building Careers  
Through Education



# 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



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

## 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



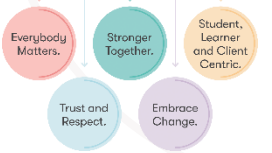
**Big Data:** like an ocean



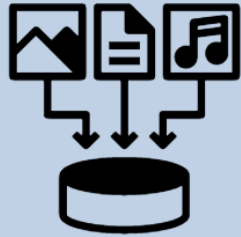
# The Characteristics of Data

## The 5 Vs of data

Building Careers  
Through Education



Volume



Variety



Velocity



Veracity



Value

*The 5 Vs of data framework*


Raise data literacy



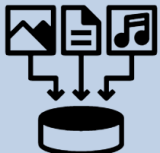
# Small Data vs Big Data

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

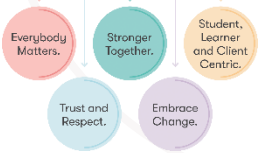
# Knowledge Check Poll

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.

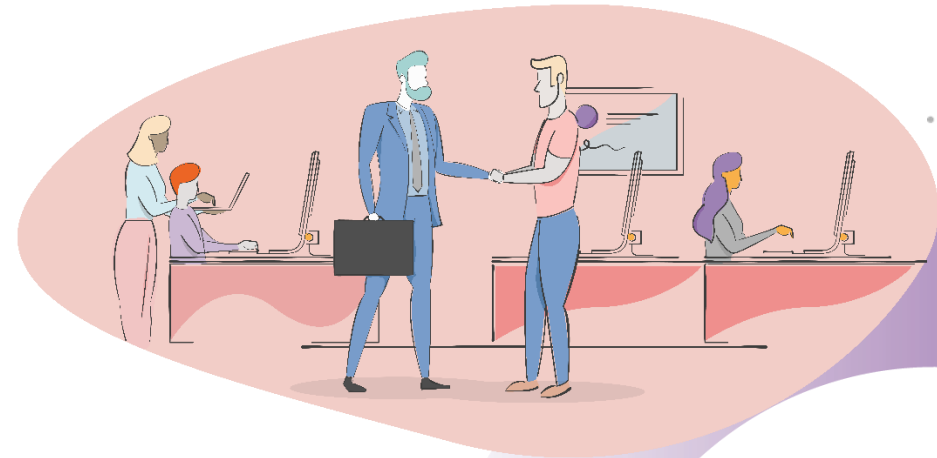
Building Careers  
Through Education



**Submit your responses to  
the chat!**



# Fundamentals of Data



# Data Deluge

## 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

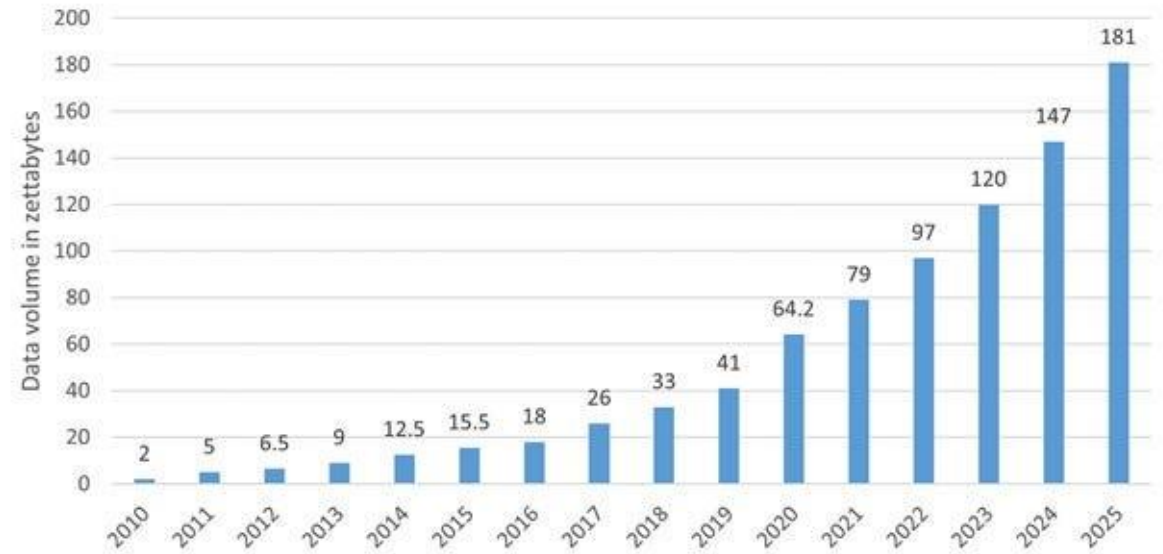


500 million instant messages are sent



5 terabytes of data is posted on Facebook

Volume of data created and replicated worldwide (source: IDC)



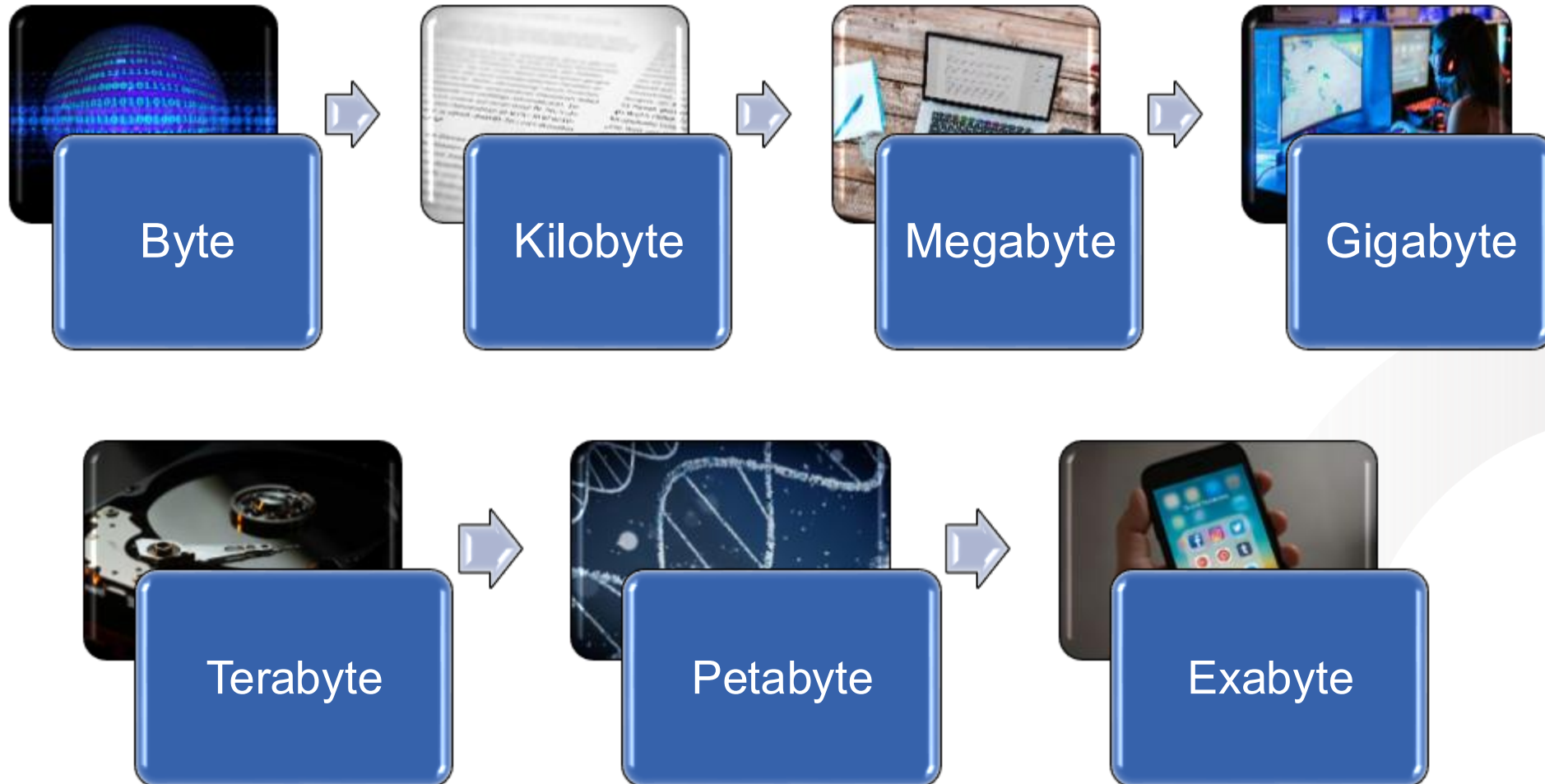
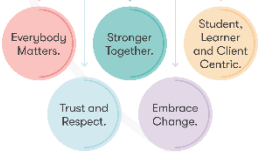
*Modern data is growing rapidly!*

*Image source: Medium.com*

# Appreciating Data Sizes

From Byte to Exabyte...

Building Careers  
Through Education

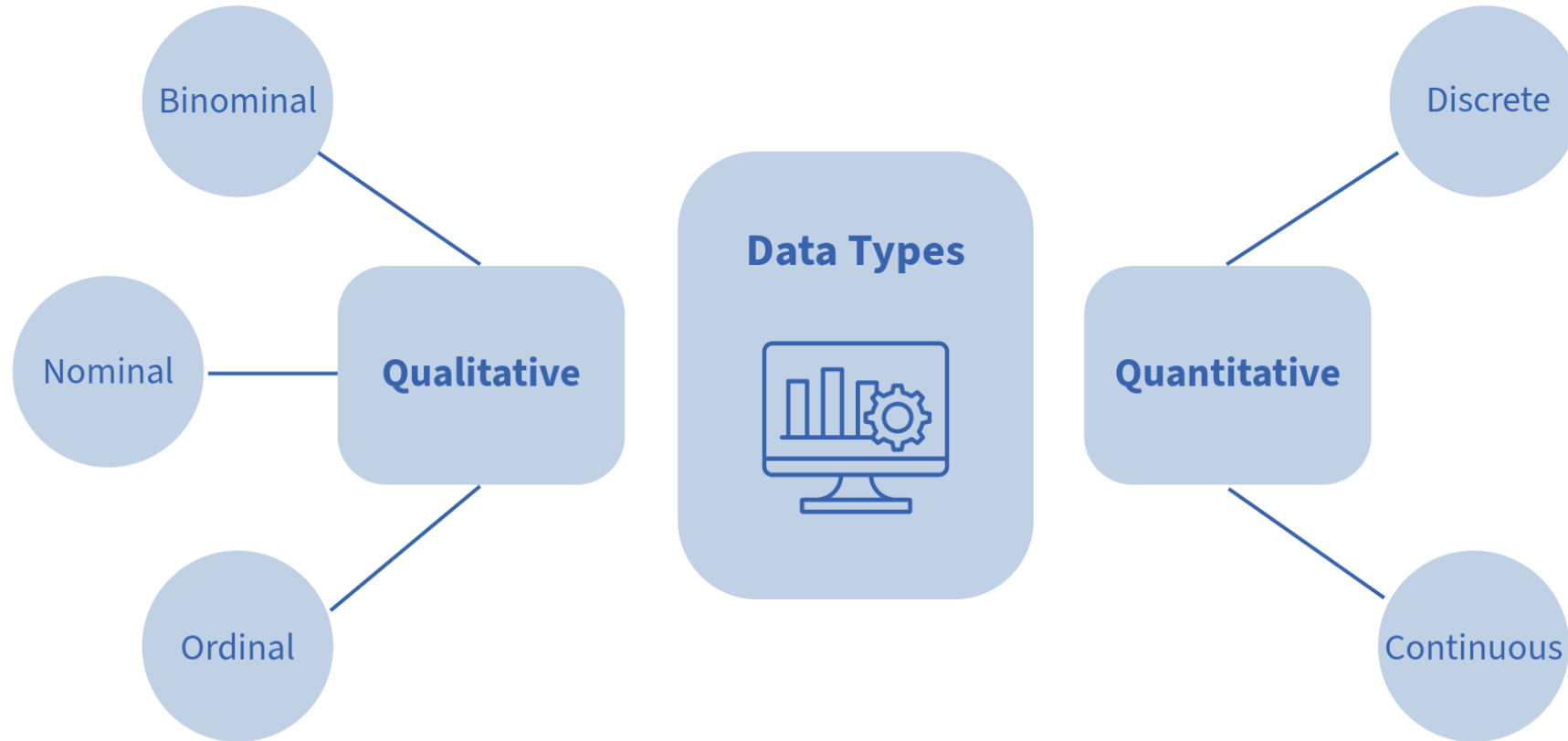


*Data sizes and everyday examples from Byte to Exabyte*

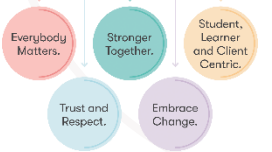


# The Different Types of Data

A hierarchy of two main groups...



Building Careers  
Through Education





# Knowledge Check Poll

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)

**Feedback: B – Terabyte (TB)**

Building Careers  
Through Education



**Submit your responses to  
the chat!**



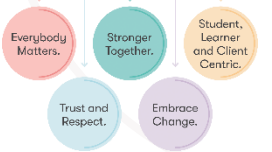
# Knowledge Check Poll

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).

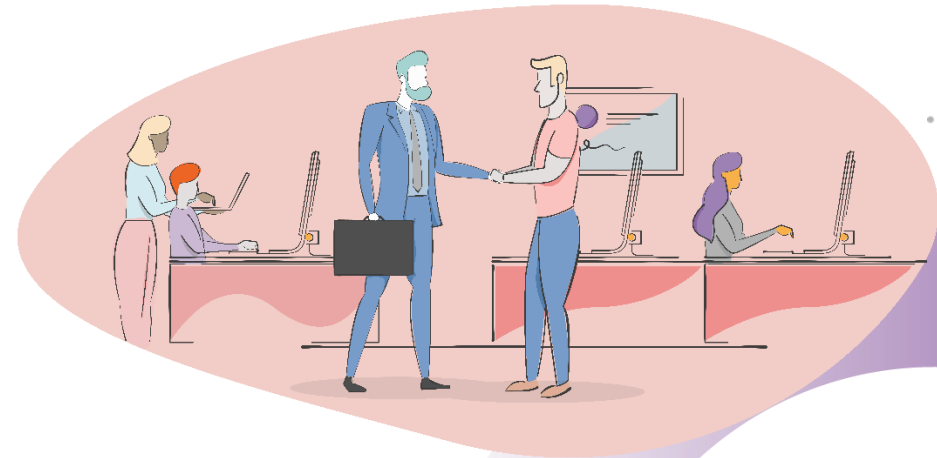
Building Careers  
Through Education



**Submit your responses to  
the chat!**



# Identifying Standards and Engineering Best Practice



# Data Formats

JSON, CSV and XML...



Social Media

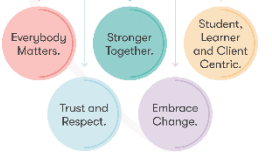


Financial  
Institutions



Healthcare

Building Careers  
Through Education



# 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...

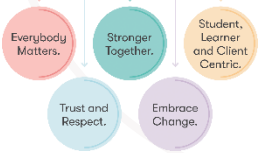


*General Data Protection Regulations  
(GDPR)*



*Information Security Management  
System (ISO) 27001*

Building Careers  
Through Education



# Engineering Best Practices

Vital for developing robust, scalable, and dependable systems...



Scalability



## Engineering Best Practices



Reliability



Security



Performance Optimisation



Data Documentation

Building Careers Through Education



# 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***

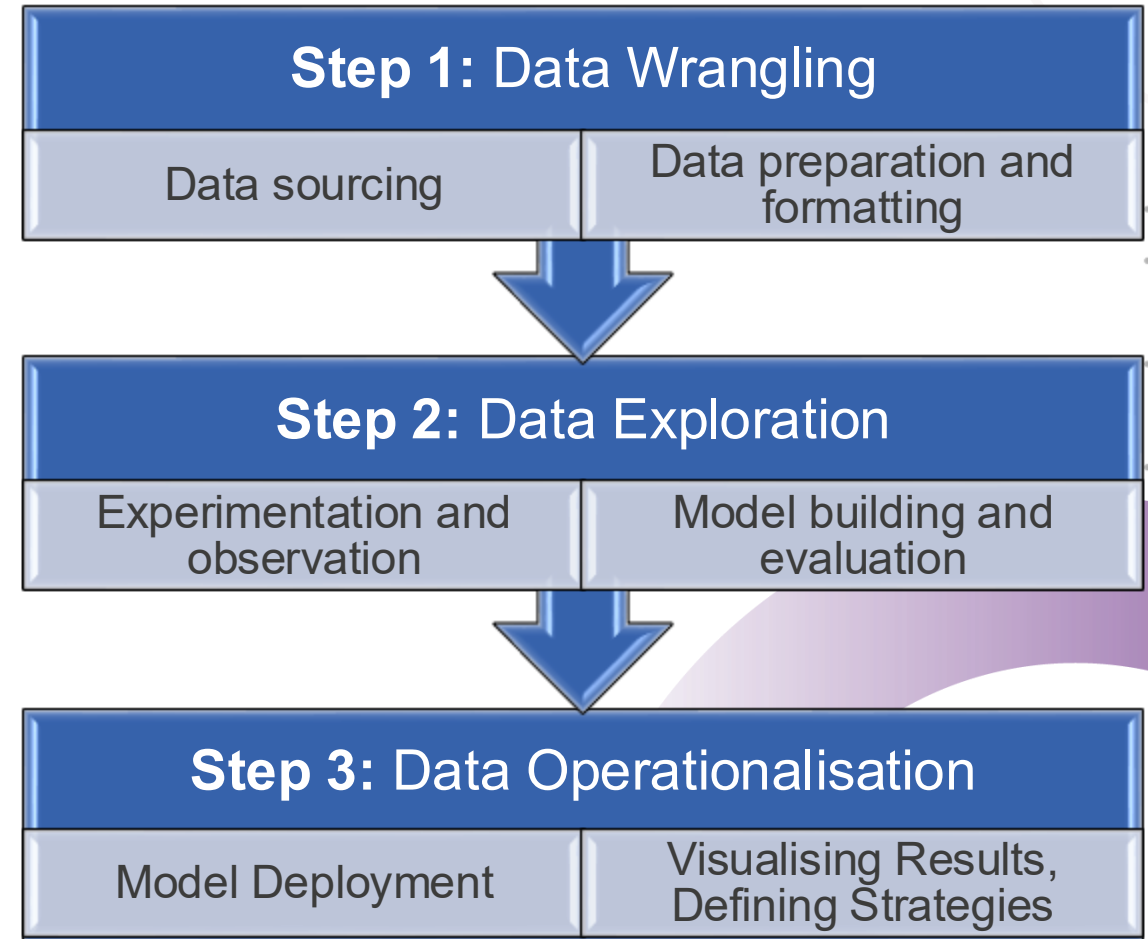




# Adding Value With Data

## 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



*The three stages of adding value with data*

# Data Teams

Who do they include?



Software Engineer



Data Teams



Chief Data Officer



Data Scientist

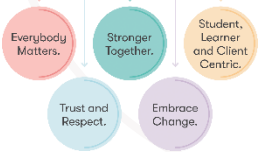


Data Engineer



Data Analyst

Building Careers  
Through Education



# 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

# Staff Retention Example

Will Peter stay with his employer...?

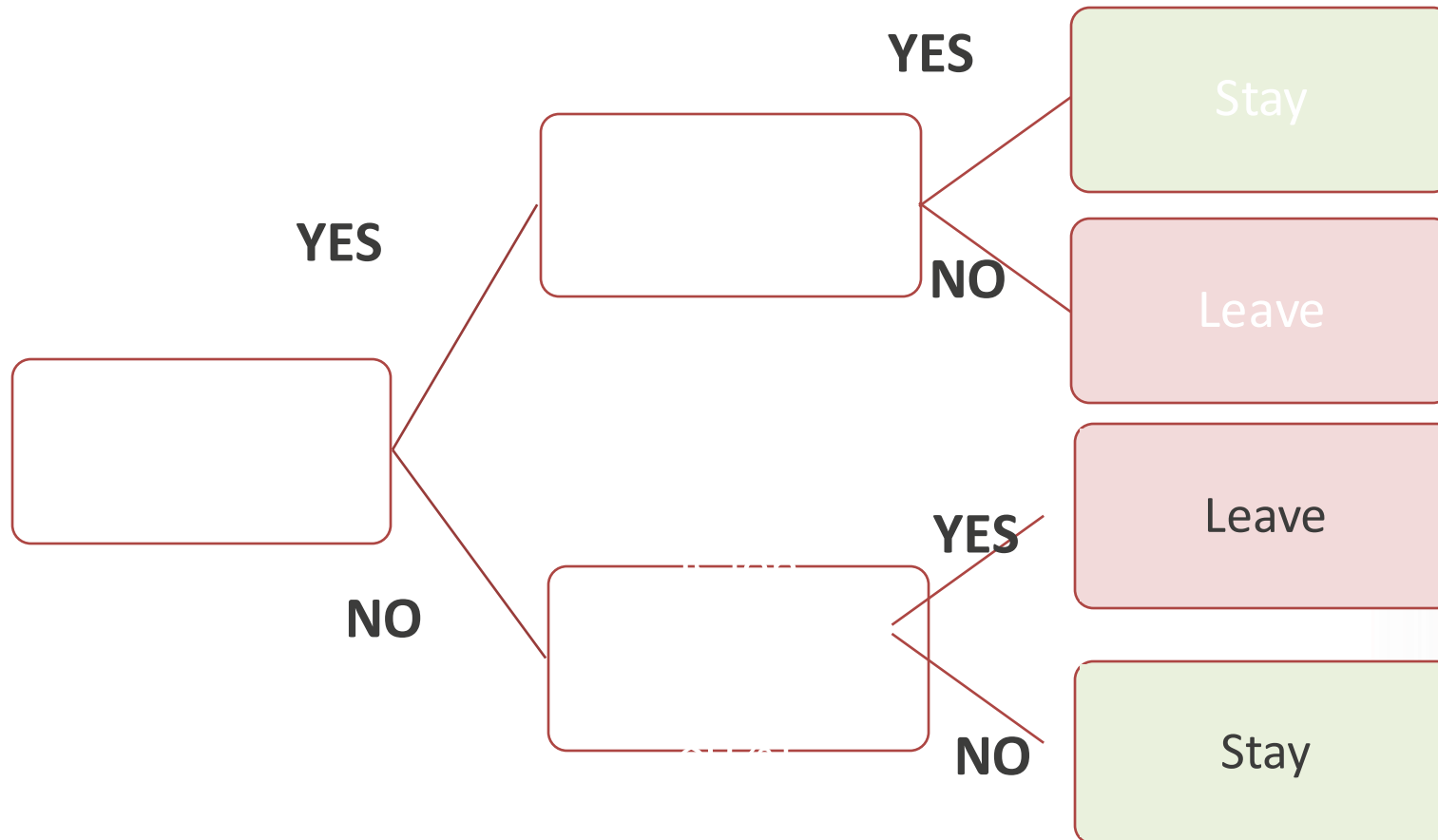
Datapoint	Value
Name:	Peter
Satisfaction Level:	0.80
Last Evaluation Score:	0.86
Number of Projects:	5
Average Monthly Hours	262
Time with Company (Years)	6
Work Accidents	0
Promotions in Last 5 Years	0
Department	Sales
Salary	£45,000
Did he resign?	?



# Decision Trees

## An introduction...

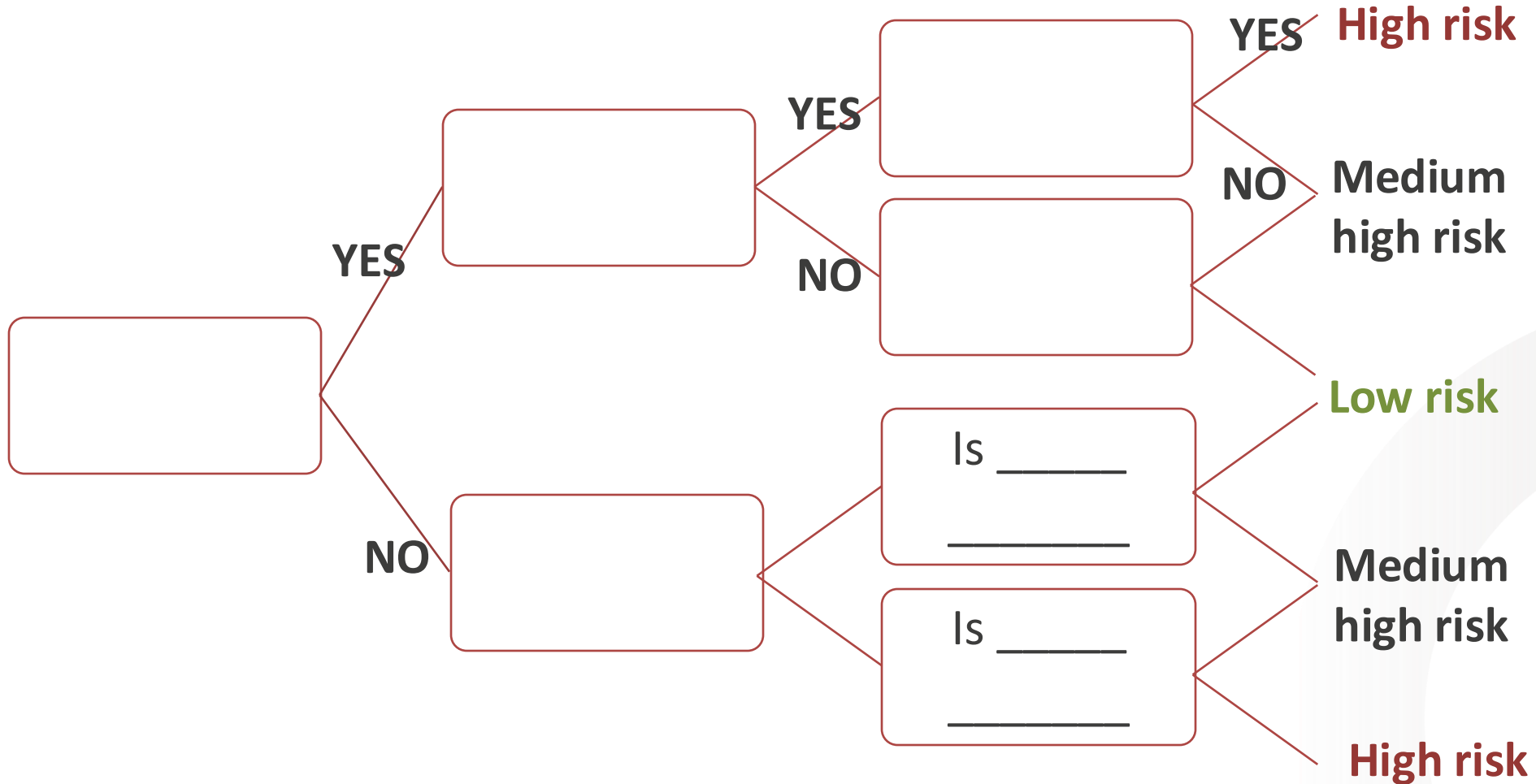
Decision trees are machine learning algorithms that represent decision-making processes in a flowchart-like structure, facilitating clear and interpretable understanding of outcomes.



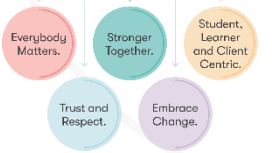
*An example decision tree*

# Decision Trees

The deep decision tree model...



Building Careers  
Through Education



*An example deep decision tree model*

# Building a Decision Tree

## Group discussion

### Building an Employee Churn Decision Tree

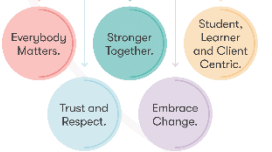
**Objective:** To illustrate how actionable insights are generated from data.

**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

Building Careers  
Through Education



## Group discussion

# Integrating Diverse Data Sources

## Activity

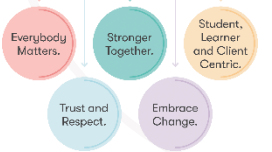
We must now imagine we are part of a team of data engineers within Credit Bank Corporation tasked with integrating employee performance data from various sources into the HR analytics dashboard. The data sources include:

- 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



## Activity

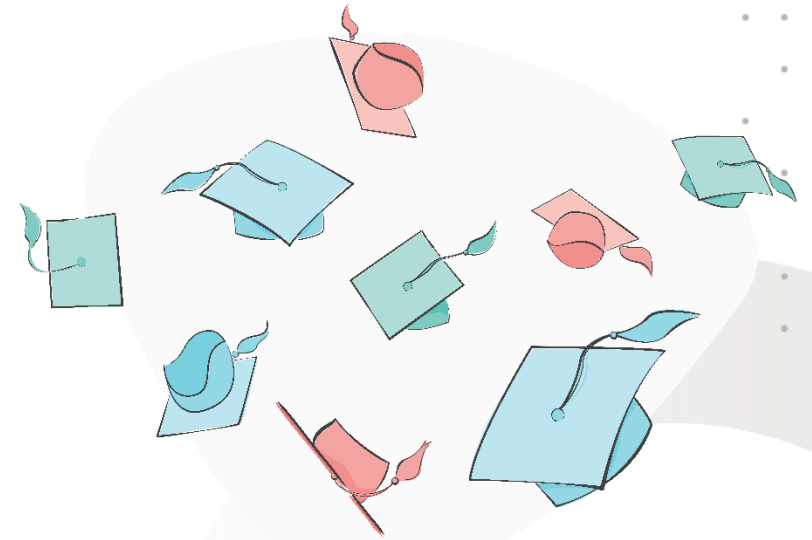
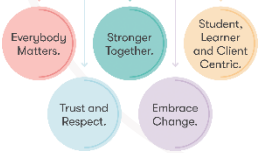


# Key Learning Summary

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
- Appreciating different data size units, from bytes to exabytes, is crucial for quantifying modern data volumes
- Cloud computing standards like the AWS Well-Architected Framework ensure consistent, reliable, and secure cloud implementations

Building Careers  
Through Education

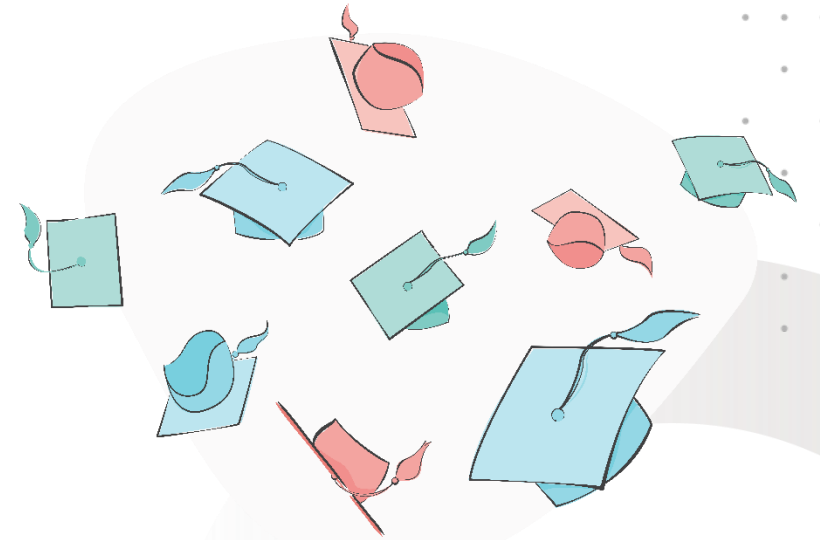
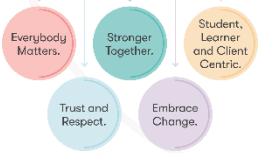


# Key Learning Summary (Cont'd)

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 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

Building Careers  
Through Education





**Thank you**

**Do you have any questions,  
comments, or feedback?**

