ALTRON | KARABINA

# Introduction to Data Science

## Bronze – Day 1

MTN

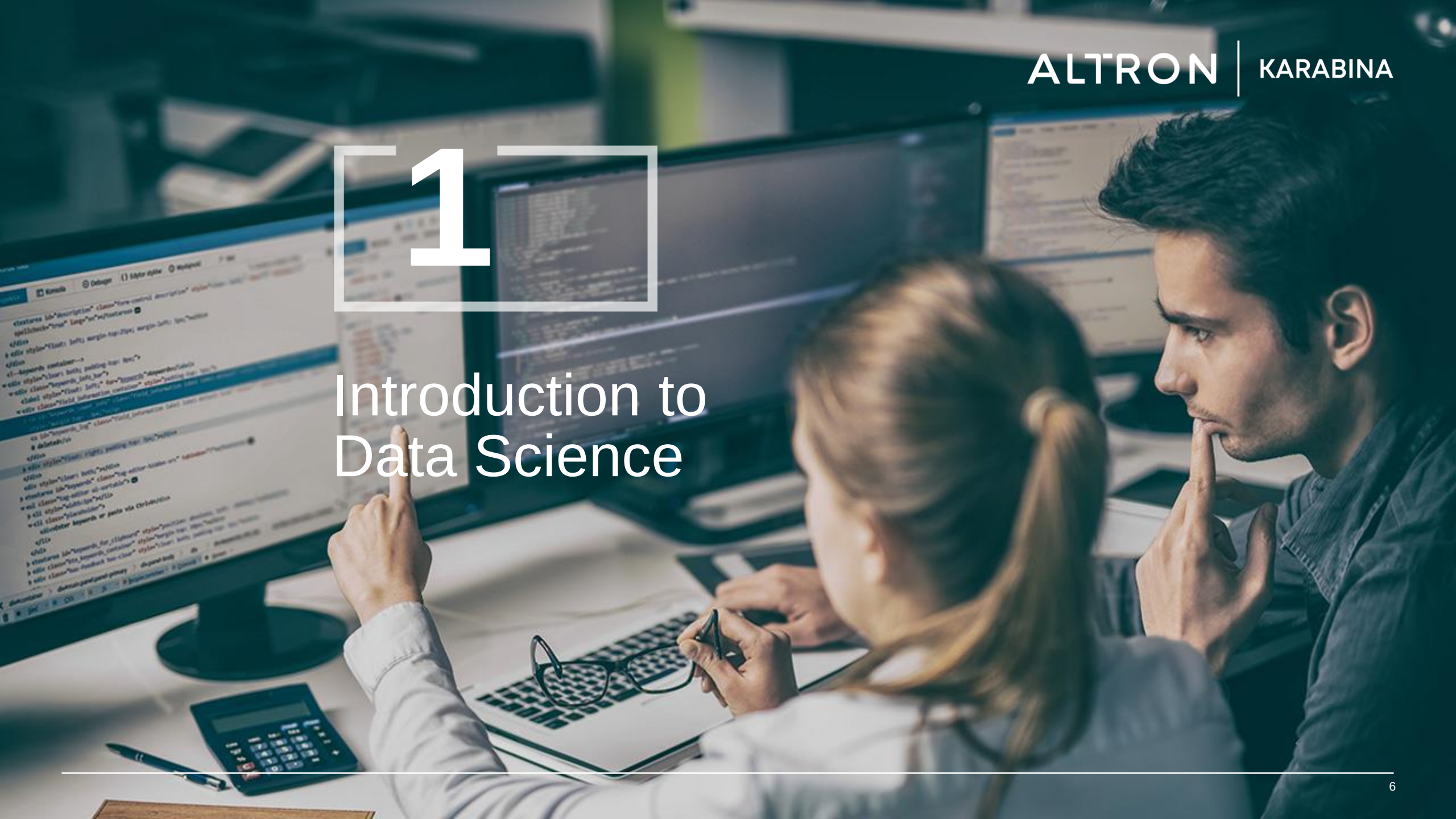
# Introduction to Data Science

ALTRON | KARABINA

## Course Outline

- Introduction and Welcome
- Data Science vs Artificial Intelligence vs Machine Learning
- What is Machine Learning and why do we care?
- Questions we can ask
- Practical lab: Employee Churn
- Appendices


PURPOSE



ALTRON | KARABINA

# 1

# Introduction to Data Science

# What is Data Science?

Field that uses the **science method** to extract **knowledge** and **insights** from data, enabling companies to make **smarter decisions.**

ALTRON | KARABINA

# What is Artificial Intelligence?

The theory and development of **computer systems** able to perform tasks normally requiring **human intelligence**, such as visual perception, speech recognition, decision-making, and translation between languages.

ALTRON | KARABINA

# What is Machine Learning?

Machine learning is the **scientific study**

of algorithms that **computer systems** use to effectively

perform a **specific task** without using explicit instructions,

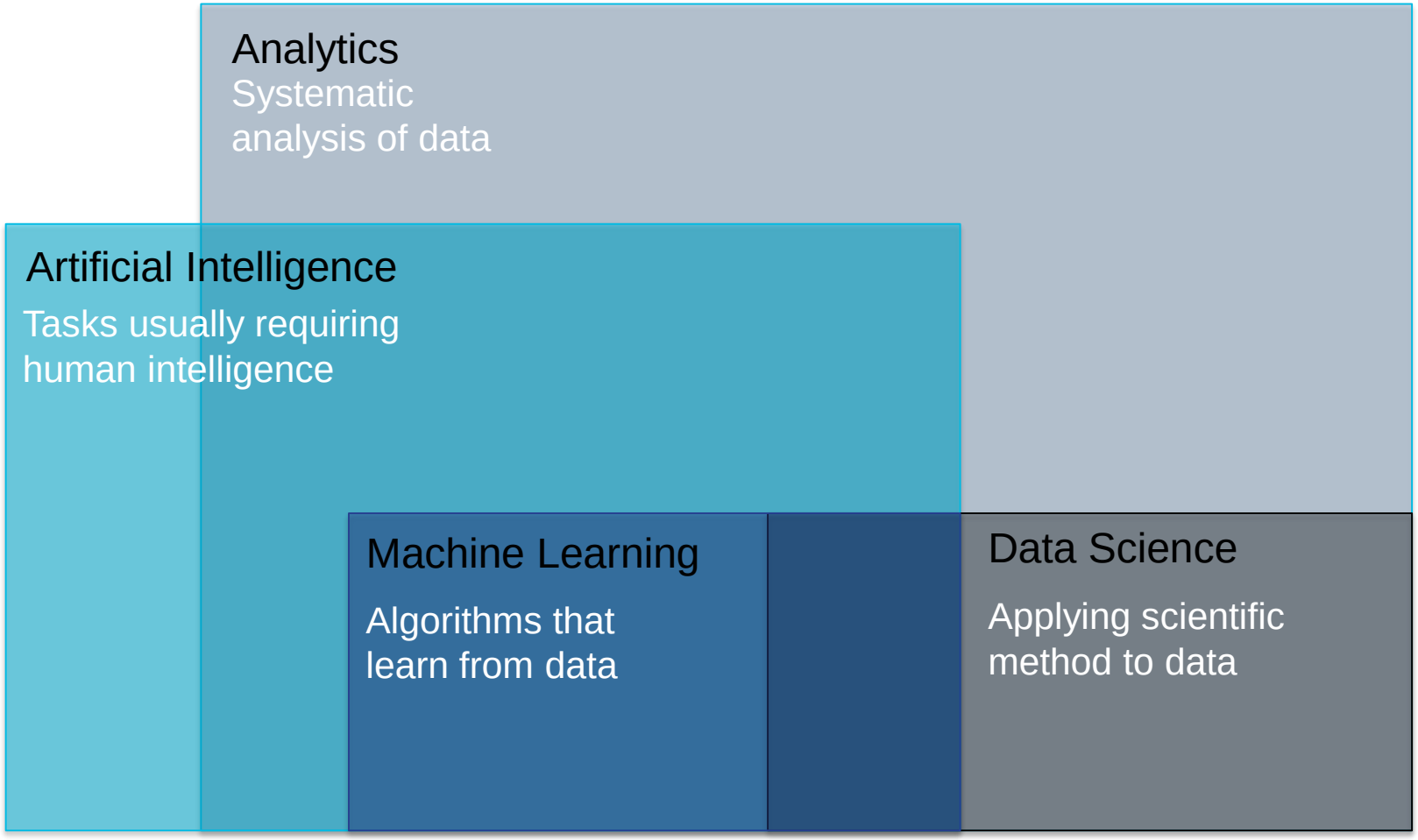
relying on **models** instead.

ALTRON | KARABINA

# Introduction to Data Science

ALTRON | KARABINA

## Clearing up the confusion



# Why do we care?

Make **better decisions**

**Directing action** based on trends

Identifying **opportunities**

Recruiting the **right talent**

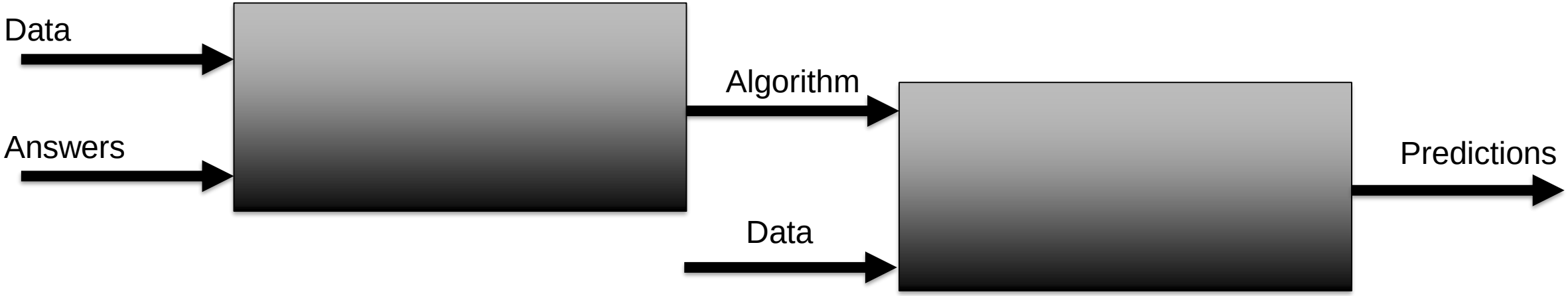
ALTRON | KARABINA

# Programming vs Machine Learning

## Traditional Programming

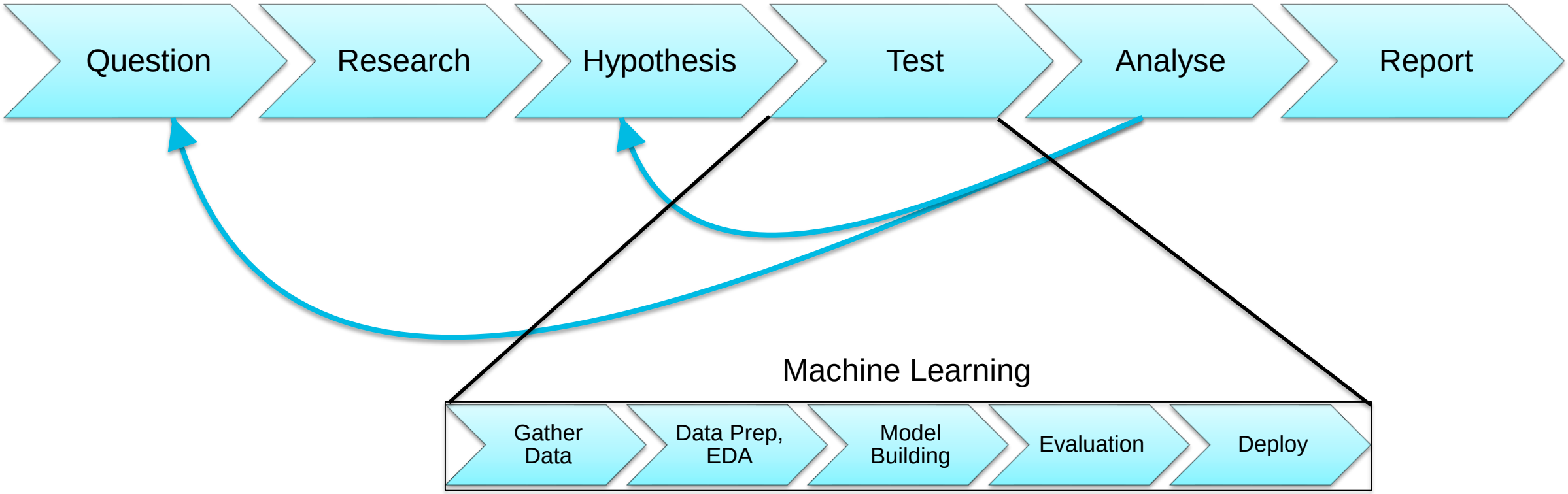
# Programming vs Machine Learning

## Machine Learning

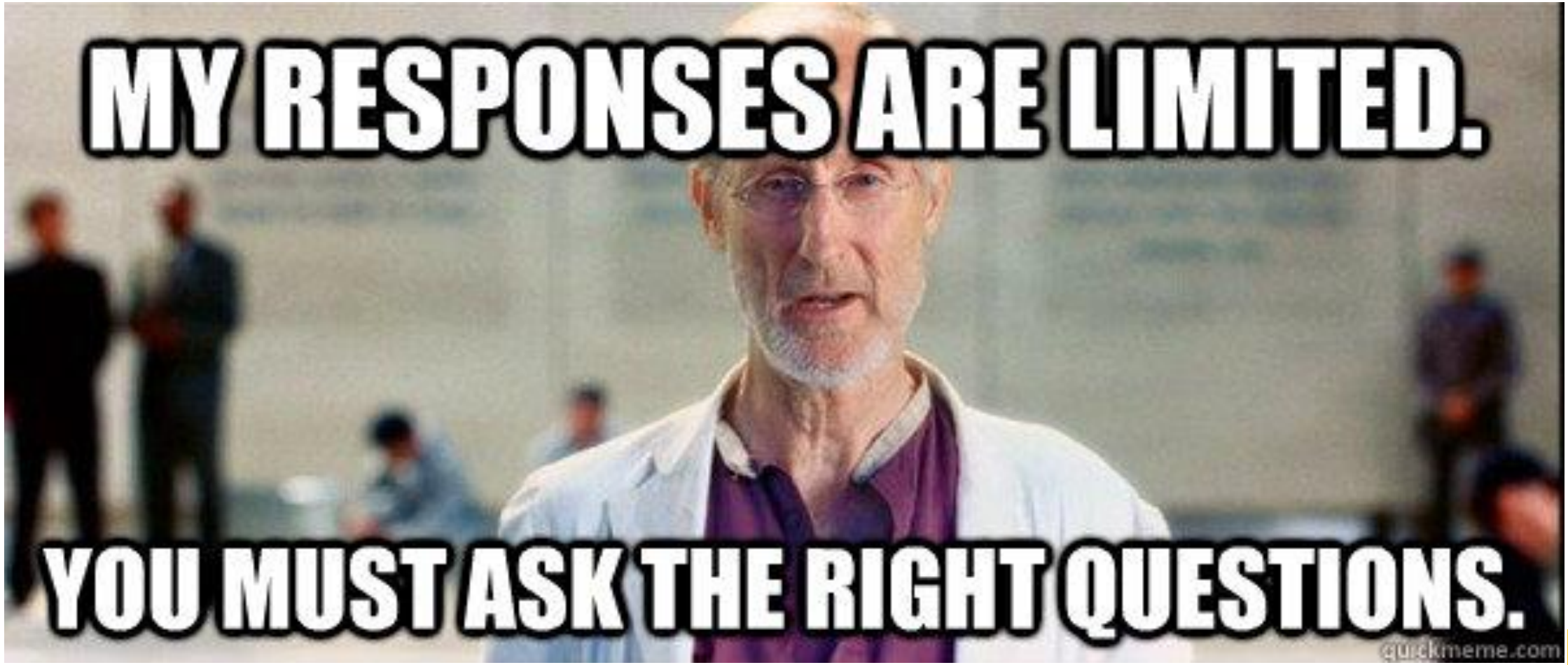


# Data Science

## Scientific Method



# Starting point of ML projects



# 2

# Machine Learning

# Am I ready for machine learning?

ALTRON | KARABINA

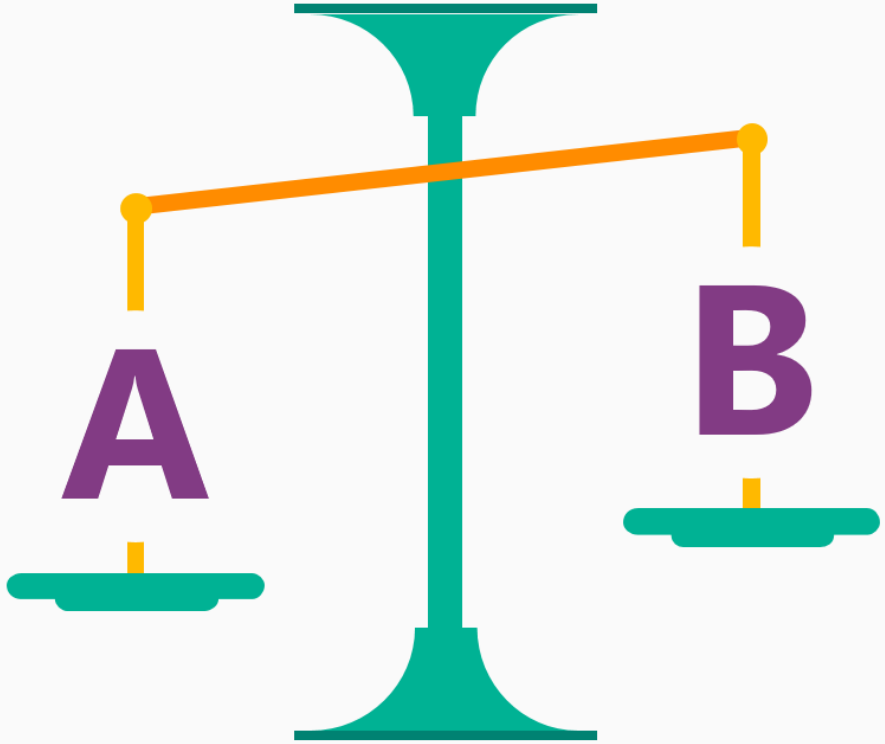
## What you need

- Sharp questions
- Data measures what is cared about
- Data is accurate
- Data is connected
- A lot of data

# Is this A or B?

## Classification algorithms

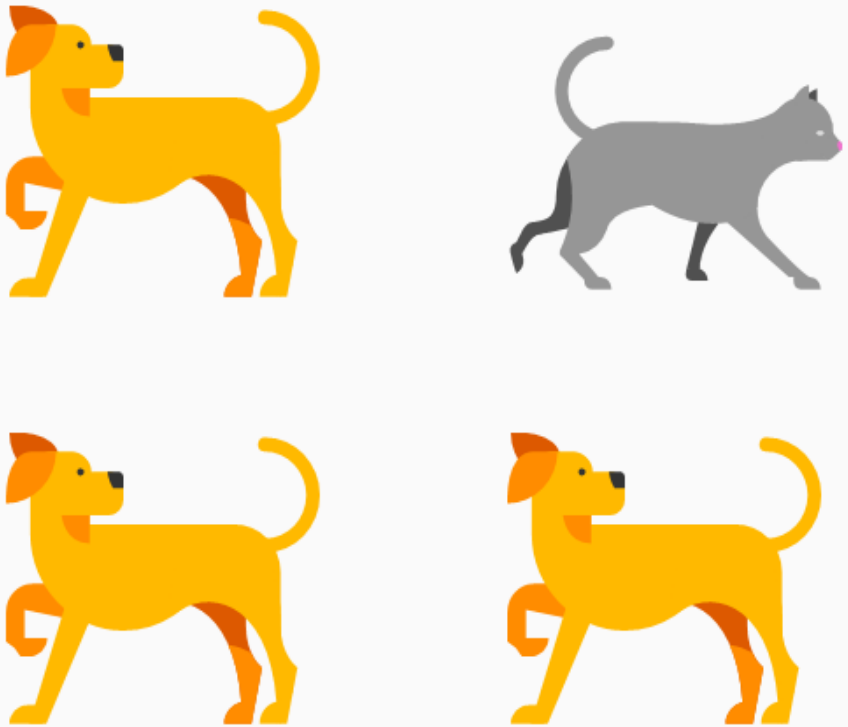
- Will this customer renew their subscription?
- Will this tyre fail in the next thousand km?
- Does the R5 coupon or the 25% off coupon result in more return customers?



# Is this weird?

## Anomaly detection algorithms

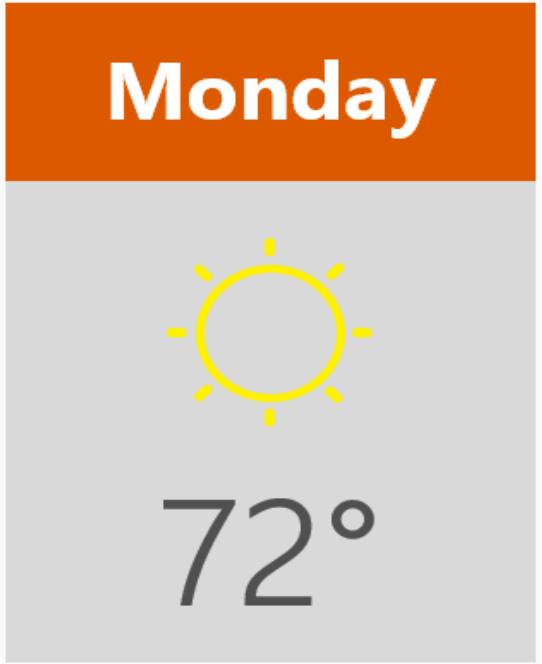
- Is this pressure reading unusual?
- Is this combination of purchases very different from what this customer has made in the past?

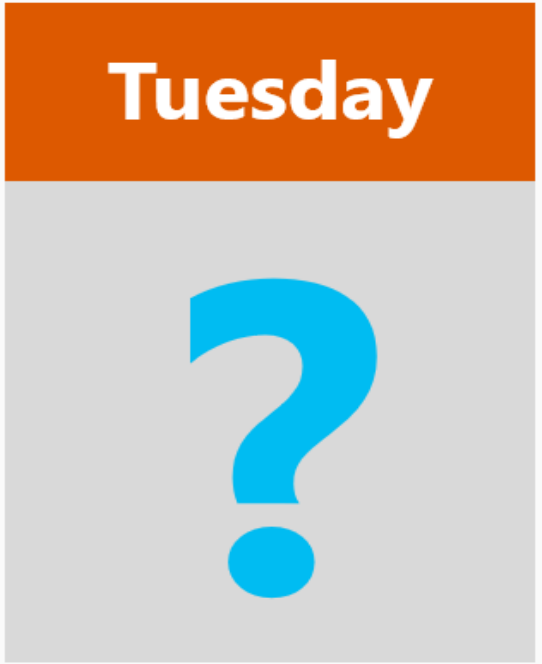


# How much? How many?

## Regression algorithms

- What will the temperature be next Tuesday?
- What will my fourth quarter sales in Portugal be?
- Out of a thousand units, how many will survive 10,000 hours of use?

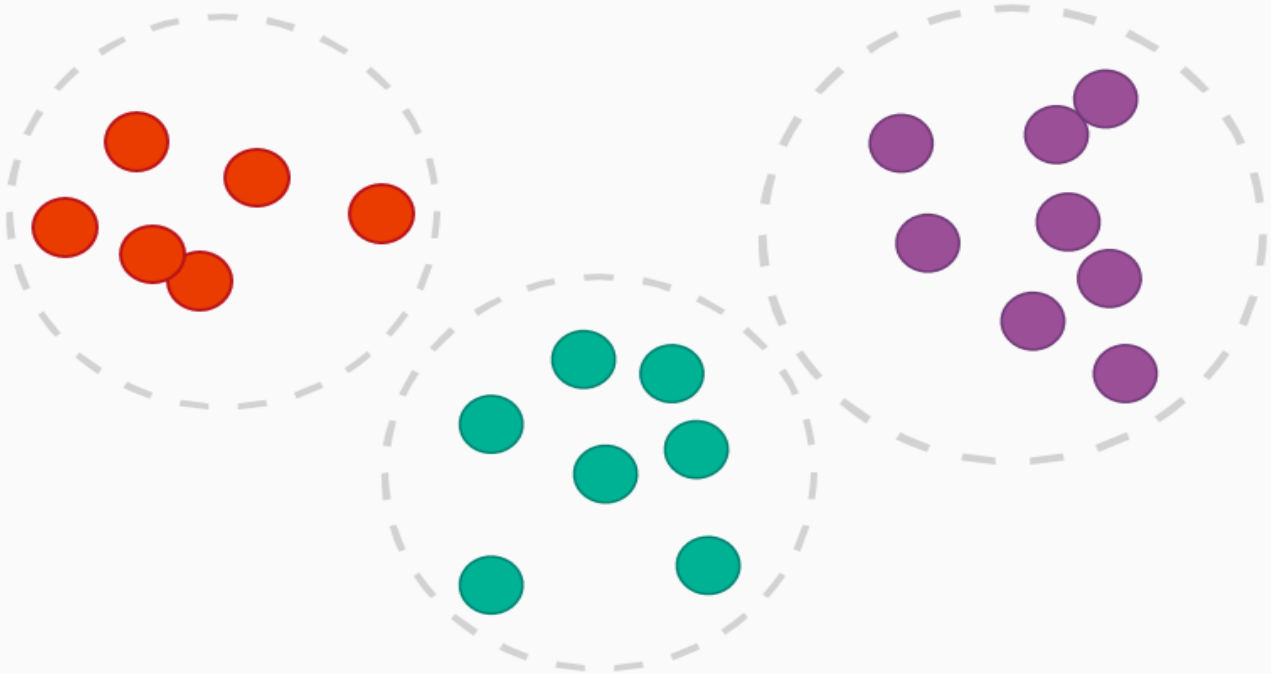


# How is this organized?

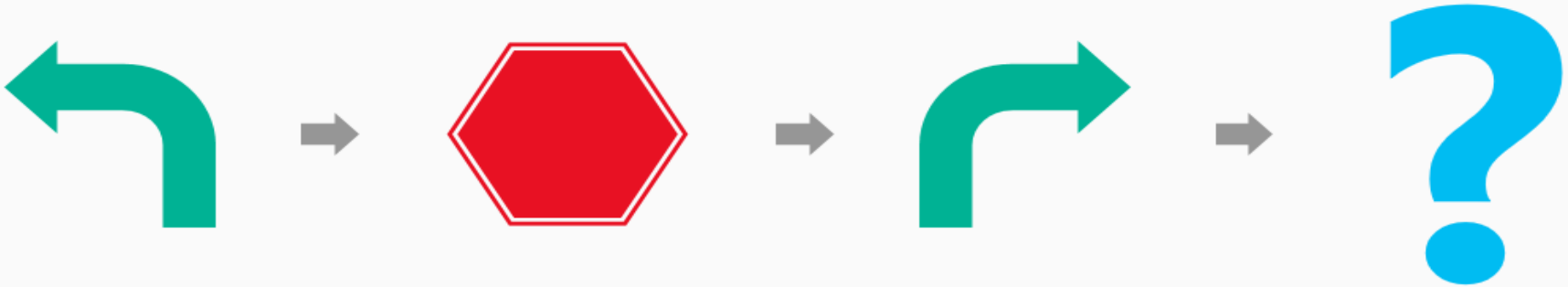
## Clustering Algorithms

- Which shoppers have similar tastes in produce?
- Which viewers like the same kind of movies?
- Which printer models fail the same way?



# What should I do now?

## Reinforcement Learning Algorithms



- How many shares of this stock should I buy right now?
- Should I continue driving at the same speed, brake, or accelerate in response to that yellow light?

# Traditional BI vs Machine Learning

Traditional BI reports aim to answer **many questions,**

an ML project aims to answer a **single, sharp question**.

ALTRON | KARABINA

# Traditional BI

ALTRON | KARABINA

## Like the army

- Large, overarching, requires effort to maintain.
- Absolutely necessary for the operations of the business.
- Needs to be coordinated and managed to meet the objectives of the business.

# Machine Learning Team

ALTRON | KARABINA

## Like a SWAT team

- Small team, highly trained and skilled.
- Given the space to work without much red tape.
- Delivers a prototype solution quickly.
- A development team could integrate the ML solution into production.

# 3

## Practical Lab

# Context

ALTRON | KARABINA

## HR churn/attrition

- You will train and evaluate a binary classification model to predict employee churn.
- You will be given a dataset with known labels for whether an employee stayed or left the company.
- The dataset contains various features/attributes about the employees.
  - Age, Daily Rate, Department, Education level, etc.
- You will build the solution in the Azure Machine Learning Studio.
- You will deploy this to be consumed by Excel.

MTN

ALTRON | KARABINA

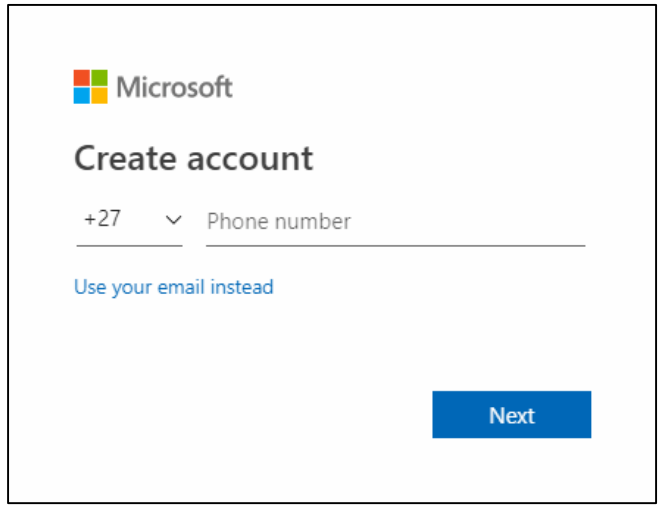
# 4

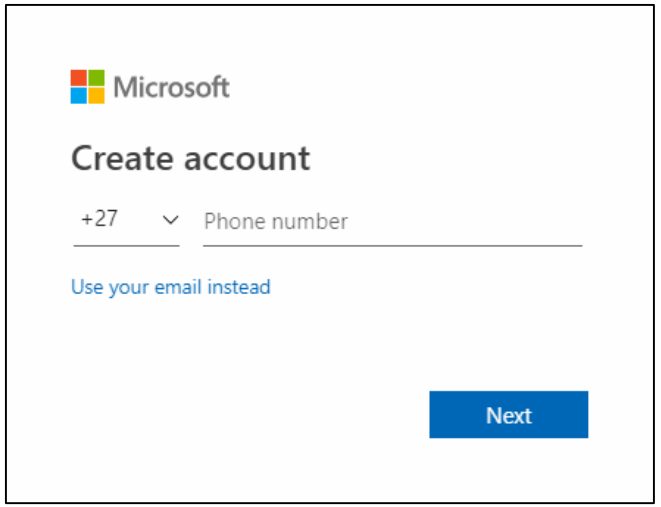
# Environment Setup

ALTRON | KARABINA

## Getting Started – Create Microsoft Account

- Step 1: Go to the Microsoft account sign-up page and select **No account? Create one!**
  - https://login.live.com/login.srf?lw=1
- Step 2: Fill out the form with your information and create a password.

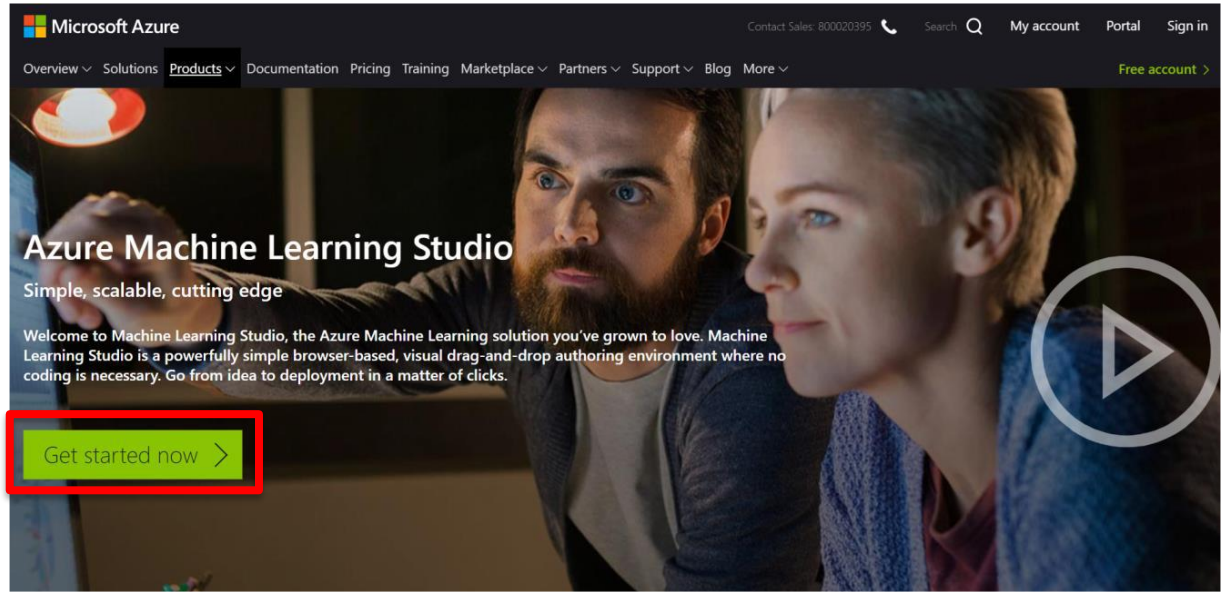


# Introduction to Data Science

## Getting Started – Azure Machine Learning Studio

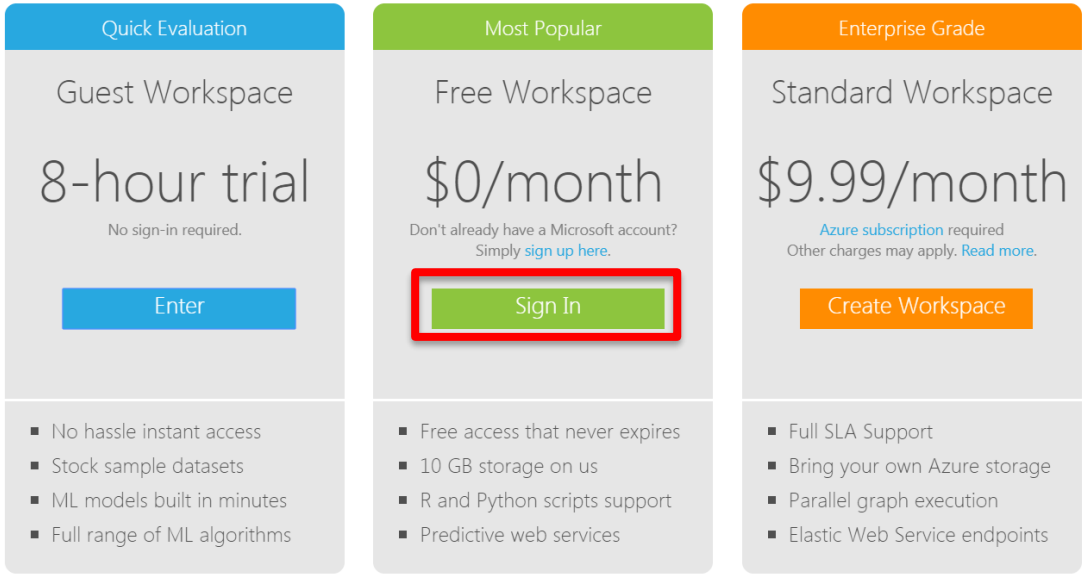
- Step 1: Go to Azure Machine Learning Home
  - https://azure.microsoft.com/en-us/services/machine-learning-studio/
- Step 2: Click on the start now button



ALTRON | KARABINA

## Getting Started – Azure Machine Learning Studio

- Step 3: Click on sign in on the “Free Workspace” option
- Step 4: Sign in with your Microsoft account
- A workspace in the Free tier will be created for you and you can start to explore Machine Learning experiments.

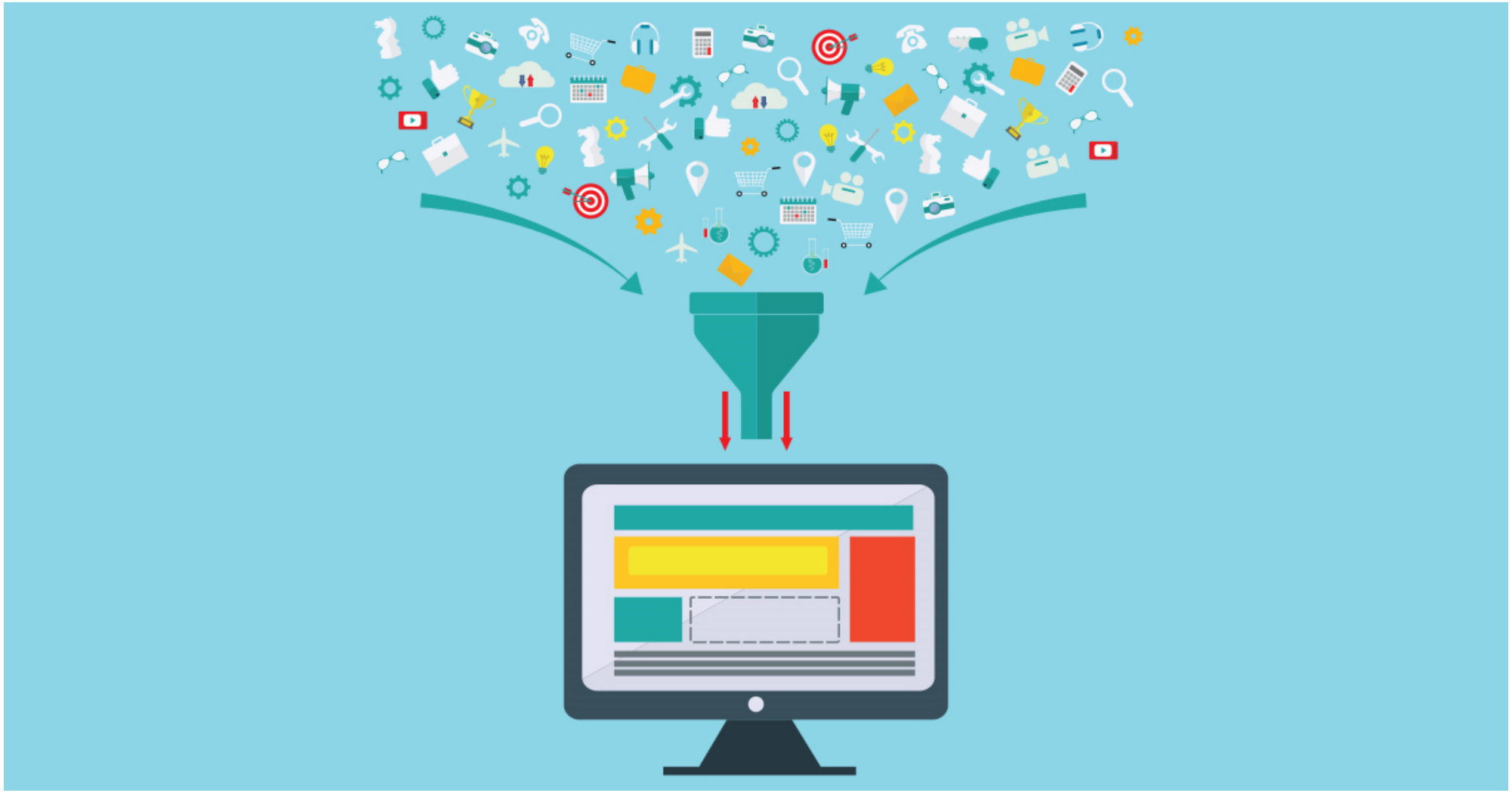


# 5

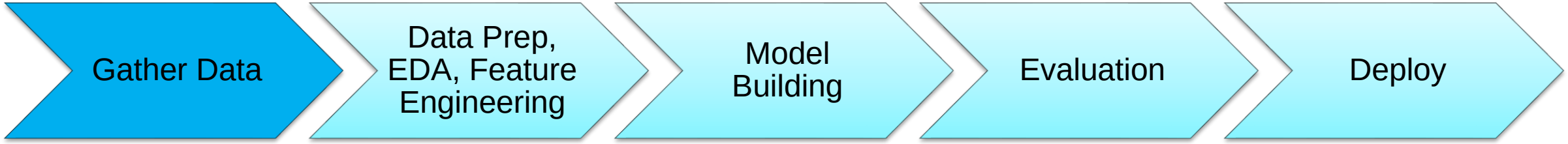
# Practical Lab

# Gather Data

ALTRON | KARABINA



# The machine learning process



# Gather Data

ALTRON | KARABINA

## Some terminology

- Features, Attributes, Dimensions:
  - These refer to the columns of a table.
- Record, Observation, Sample:
  - These refer to the rows.
- Null:
  - Missing data; unknown. This does not mean ‘zero’.
- Data type:
  - Type of data which is handled differently by the computer.
    - Numeric: Can do sums, averages, standard deviation, etc.
    - Text/String: Cannot do sums and averages.
    - Date/DateTime: Also numeric, used for date logic.
    - Boolean: Also numeric, True/false.
- Ordinal vs Nominal Data (Categorical Data)
  - Generally string data types.
  - Ordinal Data: Has inherent order, e.g., **Small, Medium, Large.**
  - Nominal Data: Has no inherent order, e.g., **Mangos, Apples, Bananas.**
- Algorithm
  - The computation ‘formula’ which is used to solve the problem.

MTN

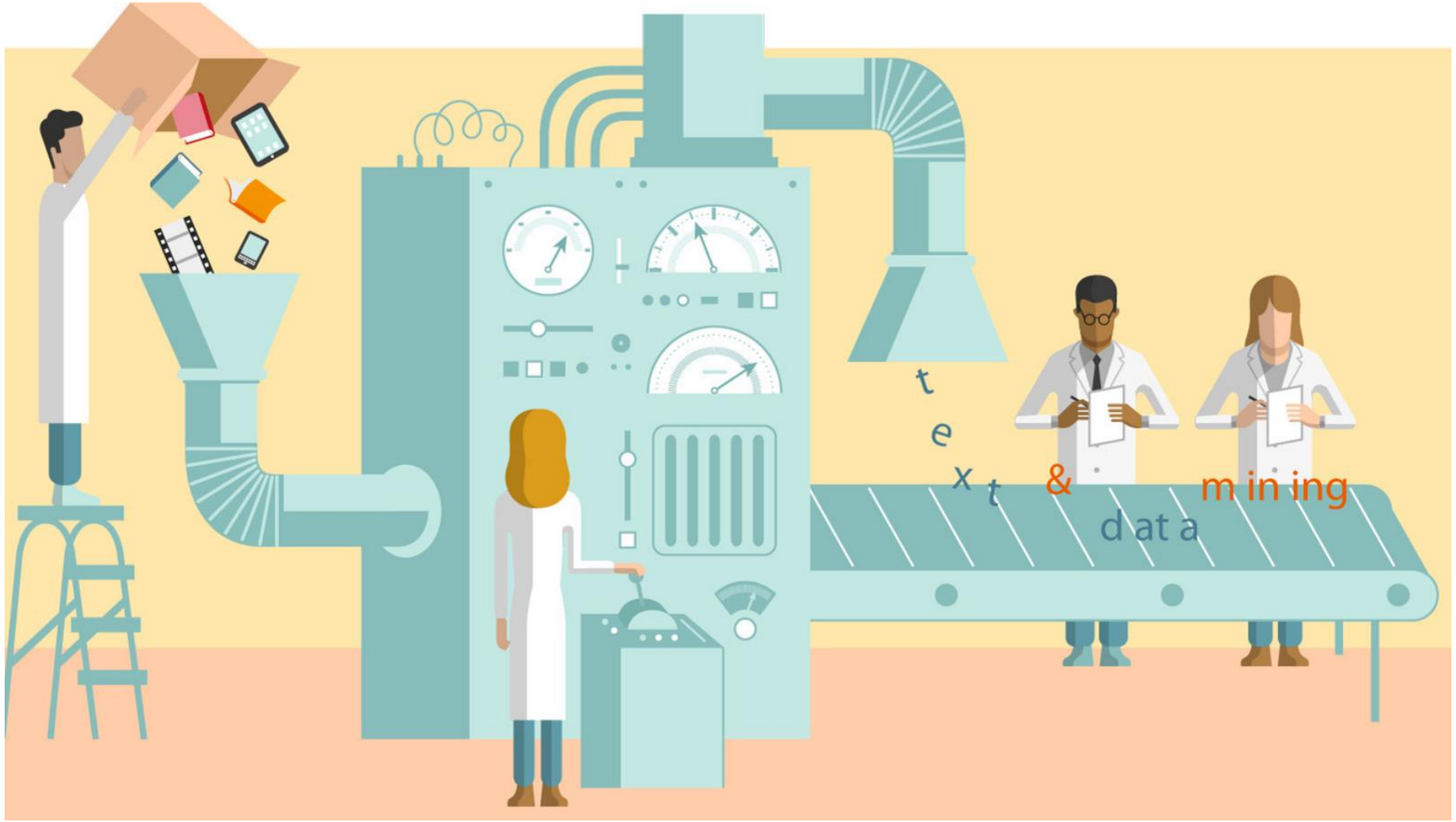
# Gather Data

ALTRON | KARABINA

## Commonly used data sources

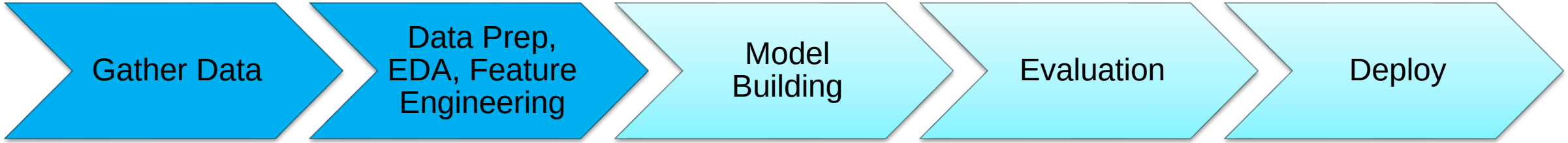
- Relational databases, like SQL Server, MySQL, PostgreSQL
- Unstructured data, like COSMOS DB, MongoDB, Data Lake
- CSV, Excel
- Streaming data from IOT devices, sensors, cell phones
- Social Media, like Twitter and Facebook
- Images
- Audio

# Data Preparation



# Data Preparation and Exploration

ALTRON | KARABINA



# Data Preparation

ALTRON | KARABINA

## Data cleaning and exploratory data analysis

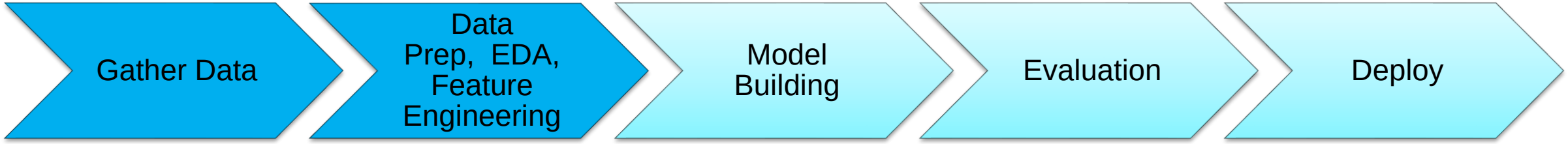
- Data is never sterile. Data errors include:
  - Missing values
    - Fill nulls with mean/median/mode; fill null with value from the previous row; delete record
  - Incorrect data types
    - Parse data as the correct type
  - Spelling mistakes
    - Have a look-up table which replaces incorrect spelling with correct spelling, delete the record
  - Outliers
    - Clip values above/below a threshold (replace value with mean/median/mode); delete the record
  - Duplicate records
    - Determine rules for finding duplicates; keep first entry, delete duplicates

MTN

# Data Visualisation

ALTRON | KARABINA

# Data Visualisation

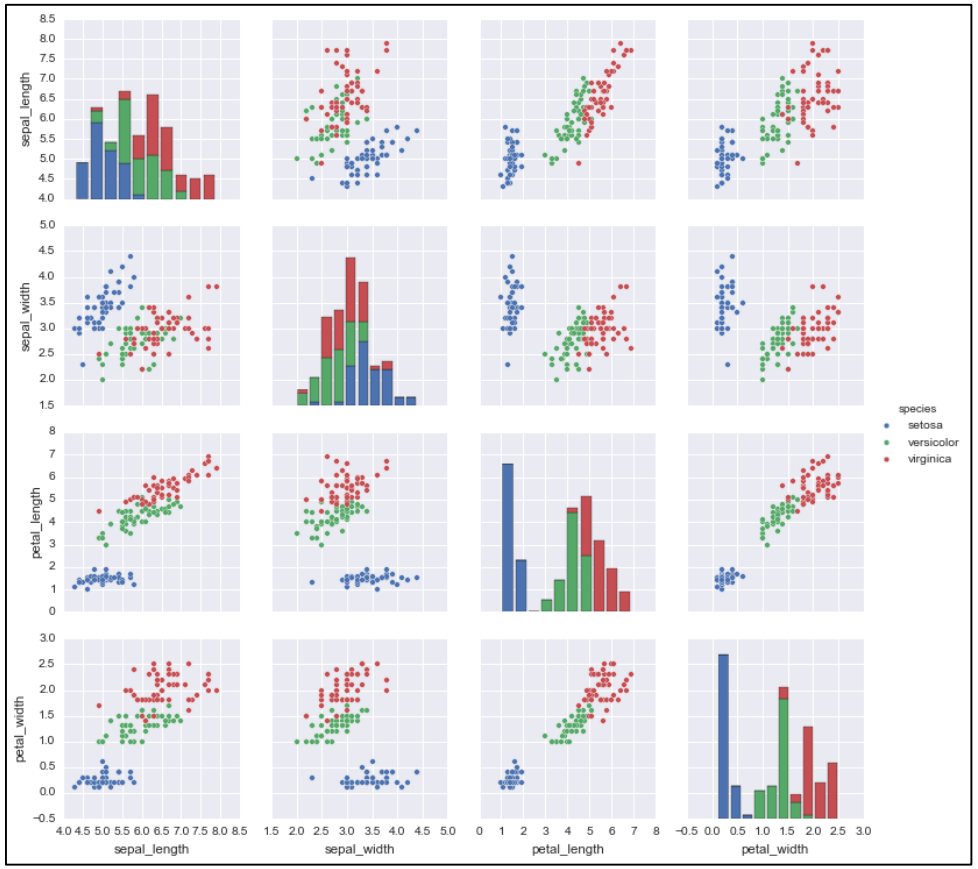


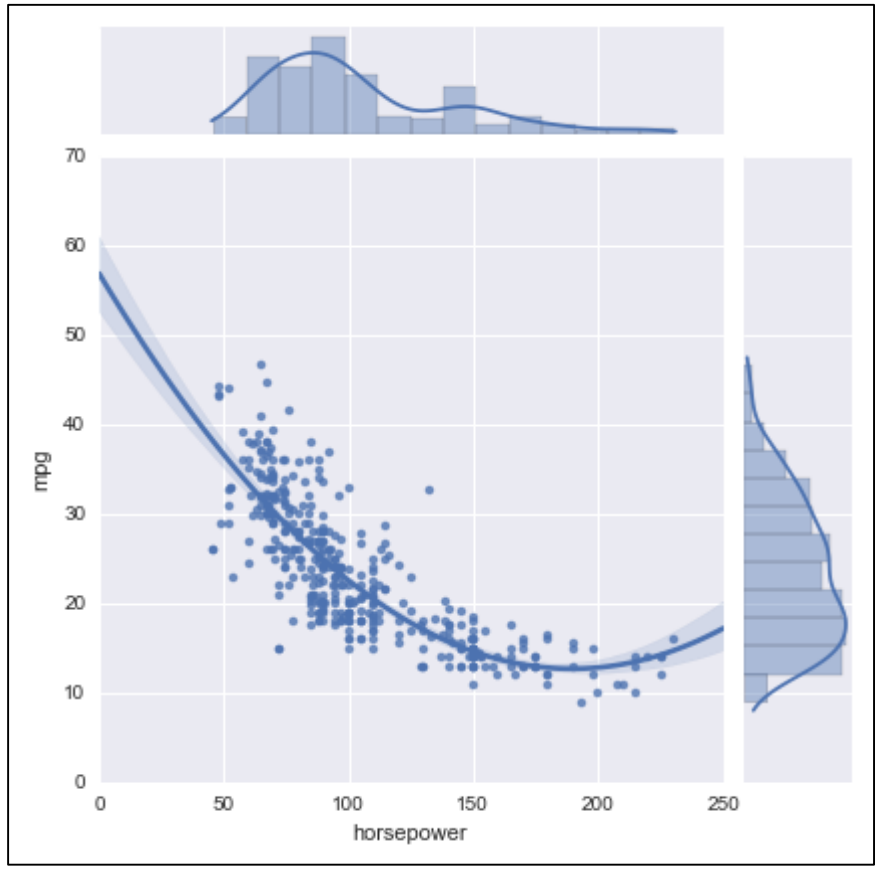
# Data Visualisation

ALTRON | KARABINA

## Data cleaning and exploratory data analysis

- Visualise the data. This is different to the reporting dashboards.



# Data Visualisation

ALTRON | KARABINA

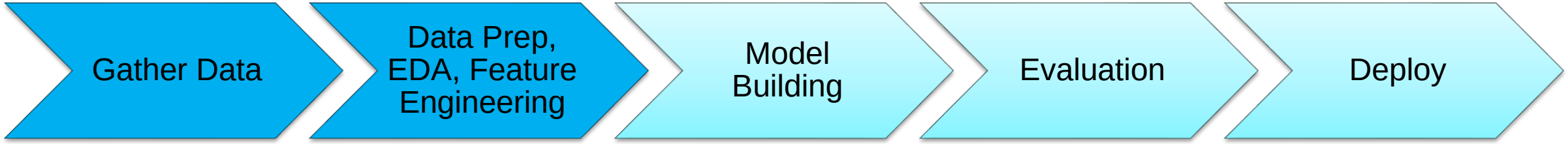
## Data cleaning and exploratory data analysis

- Visualise the data.
  - Use packages like Matplotlib, Seaborn and ggplot in Python and R.
  - This gives the data scientist a good understanding of the data, and correlations between fields.
  - This will generally guide the data scientist as to which algorithm to use.
  - Can be used to expose errors and outliers in the data.

MTN

# Feature Engineering

# Feature Engineering



# Feature Engineering

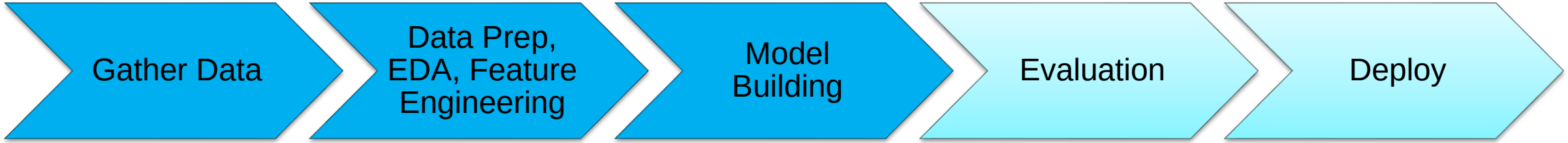
ALTRON | KARABINA

## Reducing the complexity of the data

- Represent categorical data as numeric
  - Example, **Small, Medium, Large** could be encoded as **1, 2, 3**
- Reduce the number of columns (dimensionality reduction)
  - If one column perfectly predicts another
- Removing features that do not contribute to the predictive power of the algorithm
  - Example, an employee ID number
- Normalisation
  - Normalise the data such that numeric columns' data are all within the same range.
- Binning
  - Place numeric values into bins, e.g., [0-4), [4-10), [10, 14),…

# Model Building

# Model Building



# Model Building

ALTRON | KARABINA

## Create a training and test set

- Split the data into a training and test set. Do a random split.
- The training set will be used to train the algorithm.
- The test set will be used to evaluate the accuracy of the algorithm.
  - Answers predicted by the machine learning algorithm will be compared to the actual answers of the test dataset.

Dataset

Training – 70%

Testing – 30%

# Model Building

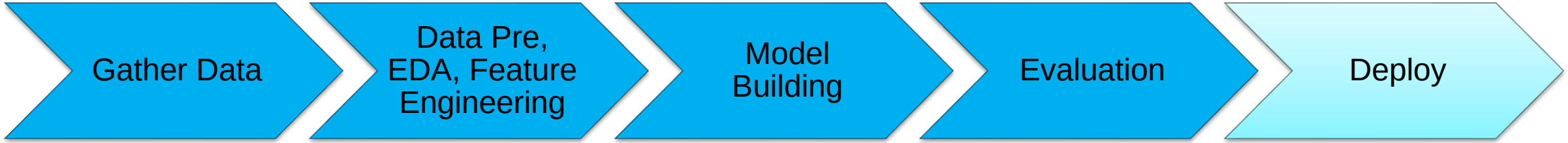
ALTRON | KARABINA

## Training the algorithm

- Input the training data with the answers.
- The chosen algorithm will find the relationship between data and the answers.
- Exactly how this is done depends on the type of machine learning algorithm, and the type of question being answered.

# Model Evaluation

# Model Evaluation



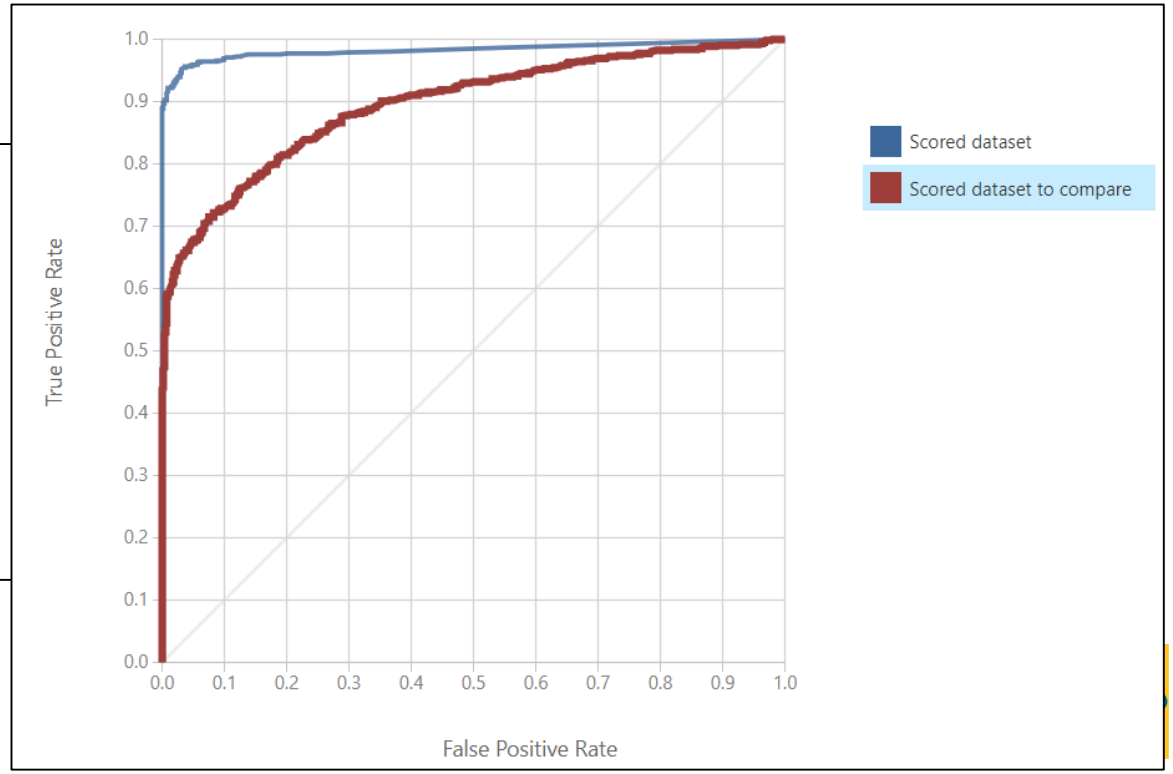
# Model Evaluation

ALTRON | KARABINA

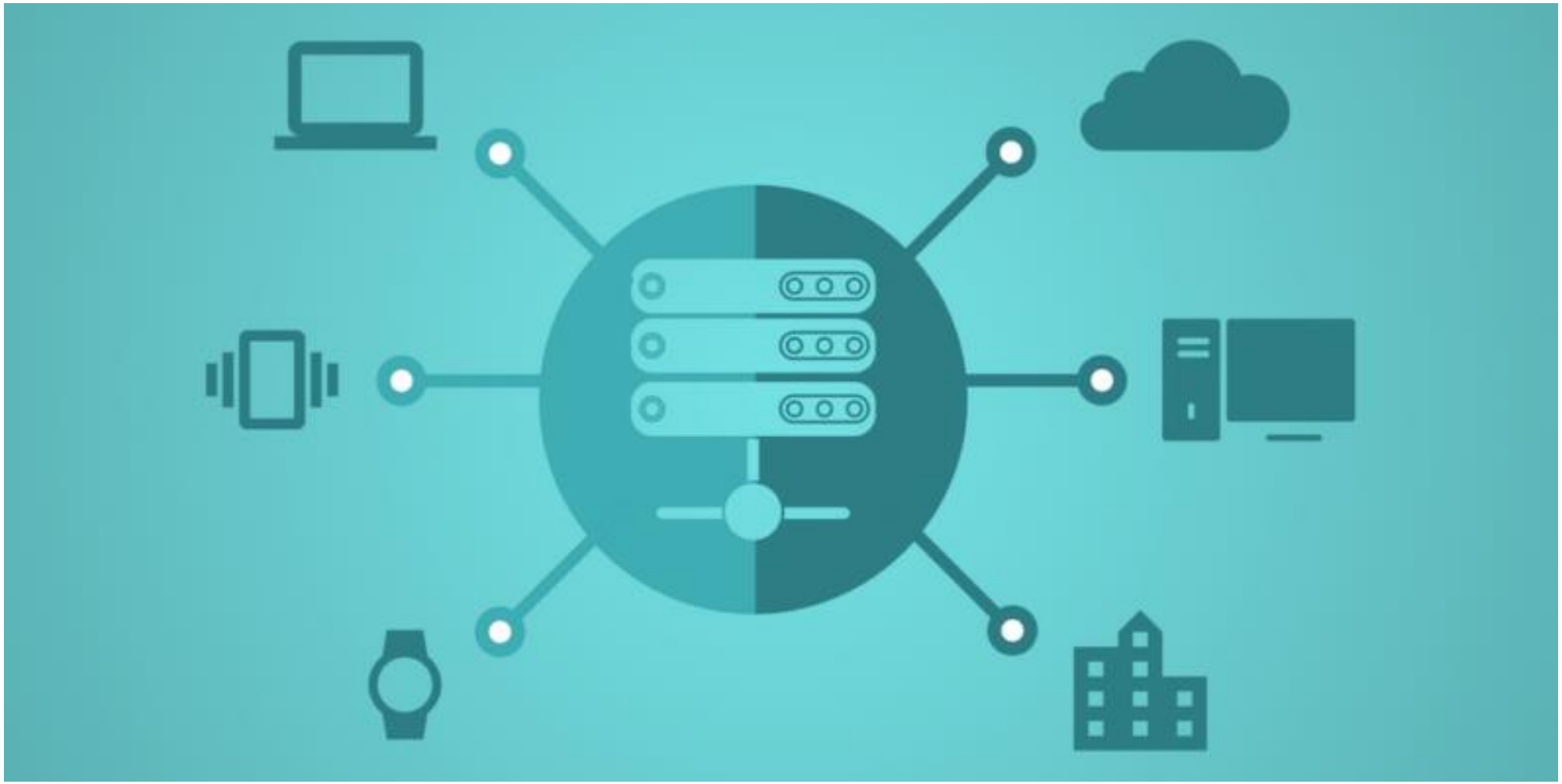
## Classification

- You want to determine how many observations were misclassified.
  - How many True observations were classified as False?
  - How many False observations were classified as True?

| Actual Answer | Predicated Answer: Positive | Predicated Answer: Negative | | |
|---|---|---|---|---|
| Positive | True Positive: 475 | False Negative: 139 | Accuracy: 0.811 | Precision: 0.856 |
| Negative | False Positive: 80 | True Negative: 465 | Recall: 0.774 | F1 Score: 0.813 |
| | Positive Label: 1 | Negative Label: -1 | | |

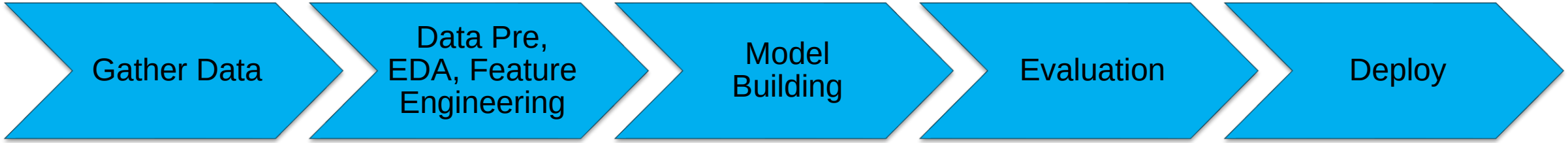


# Model Deployment



# Model Deployment

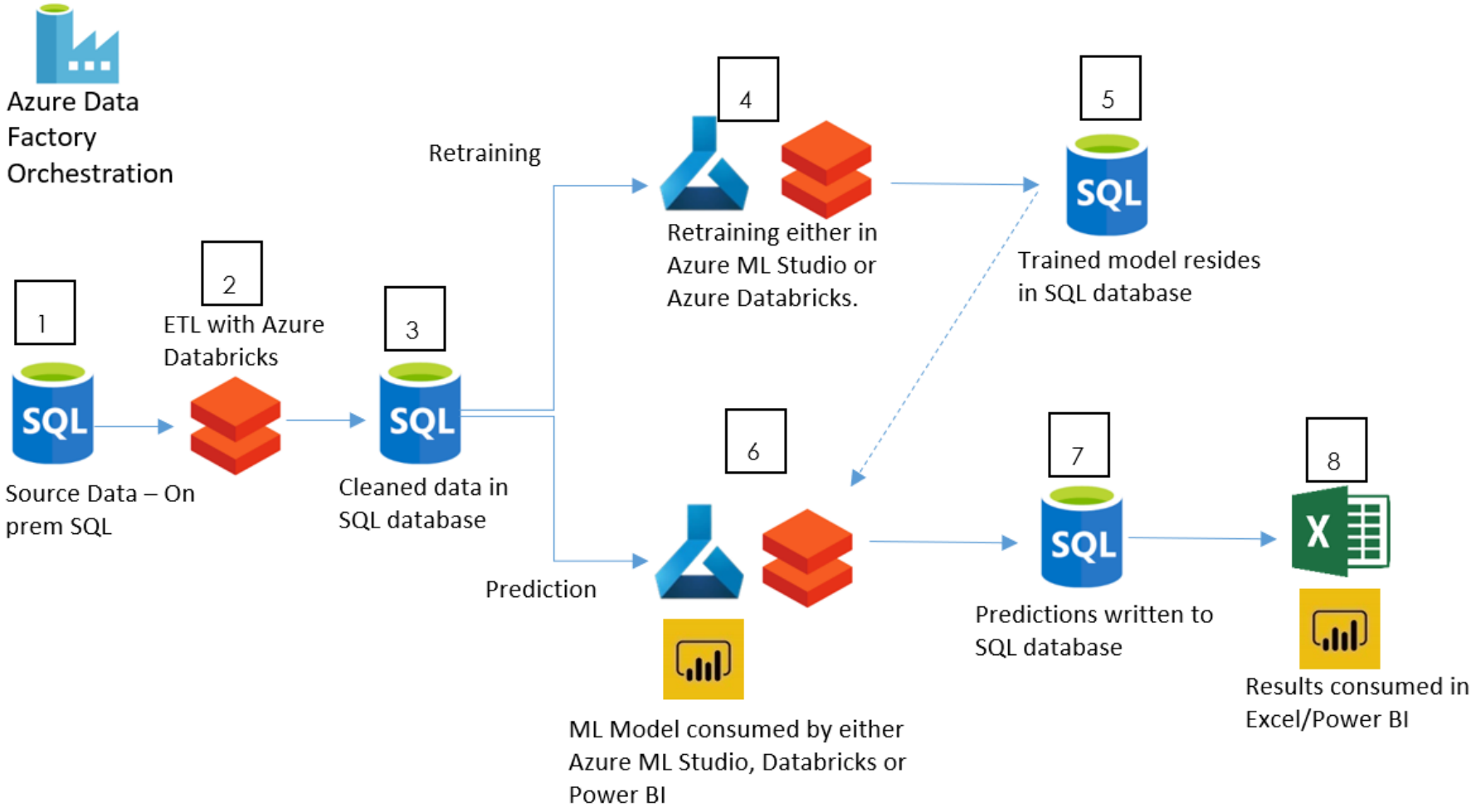
ALTRON | KARABINA



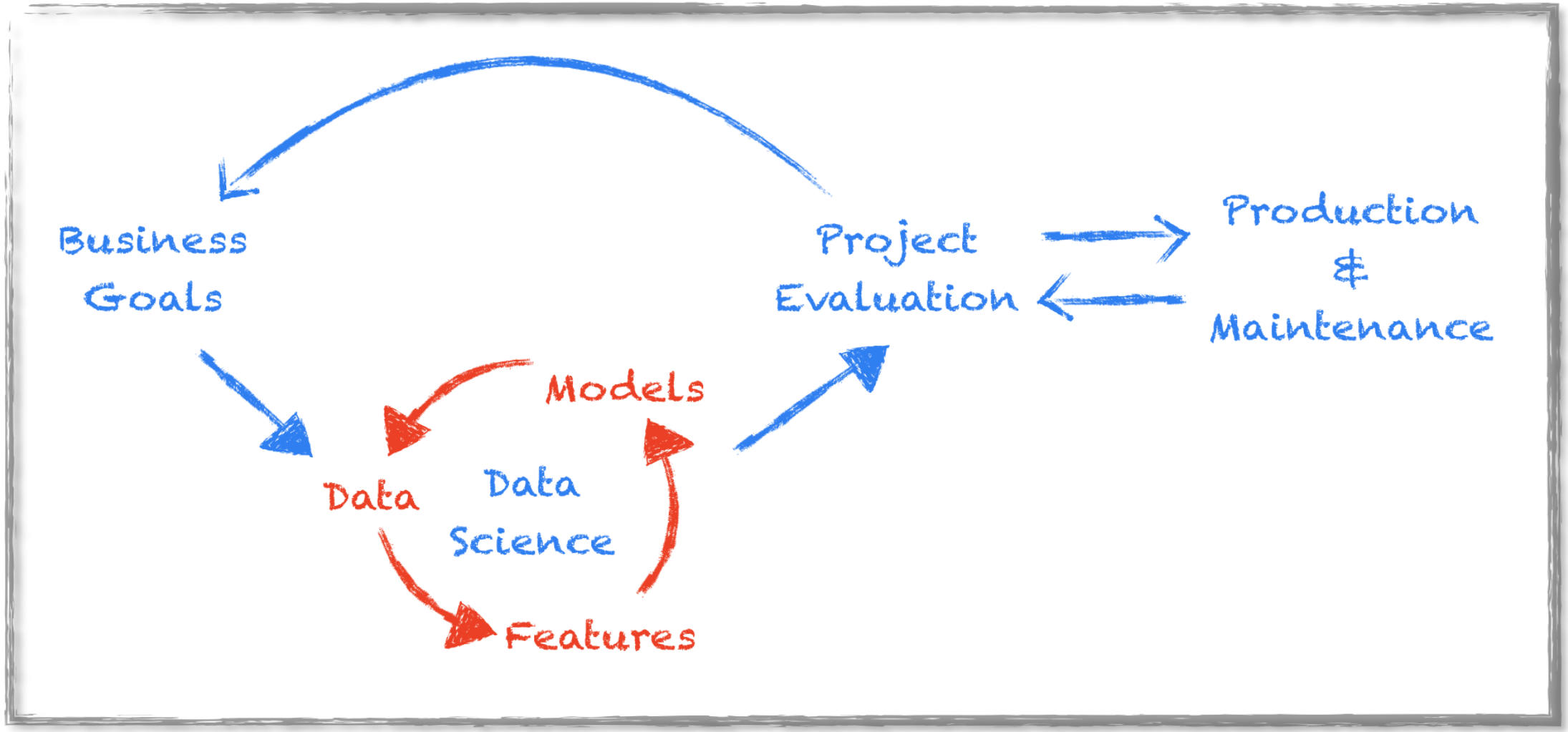
# Deployment

ALTRON | KARABINA

## Data Factory Deployment Strategy



# Next Steps


Business
Goals
Project
Evaluation
Production
&
Maintenance
Models
Data
Data
Science
Features

ALTRON | KARABINA

# Thank you

connect@altronkarabina.com

www.altronkarabina.com

JHB +27 11 463 8155