Big Data and AI

Mónica-Juliana Pérez

Agenda

- Welcome
- Agenda and Topics of the course
- Rules and grades distribution
- Prerequisites and Recommended Software/Tools:
- Introduction to AI and Big data

Welcome

Learning Objectives:

By the end of the course, students will be able to:

1. Understand the fundamentals of AI and Big Data and their applications in logistics.
2. Analyze large datasets to extract actionable insights for logistics operations.
3. Design AI-driven solutions for supply chain optimization.
4. Evaluate the ethical and practical considerations of using AI in logistics.

Miss. Mónica-Juliana PÉREZ

Monica.PerezMorales@uphf.fr

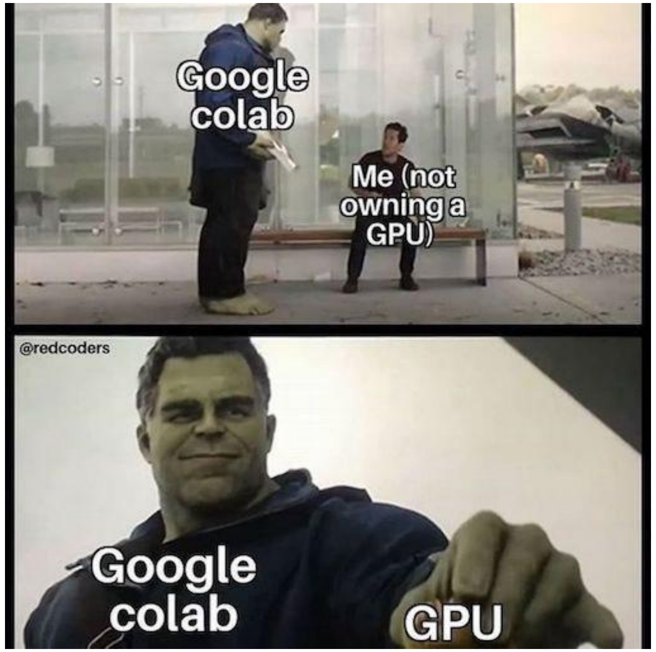
PhD student specializing in Logistics, AI, and Analytics.
Research focuses on AI-driven optimization in logistics.
Dual master’s degrees: Industrial Engineering and Analytics.

Evaluation

- 50% Project
- 40% Exercises (TP)
- 10% Class participation

- **Prerequisites:**
  - Basic programming knowledge (Python preferred) or ChatGPT
  - Familiarity with data analysis tools (Excel, SQL).
- **Recommended Software/Tools:**
  - Python (NumPy, Pandas, Scikit-learn, TensorFlow).

Notebook google Collab



How we take desicions?

“intuition” or “Logic”

simply knowing when something is right or wrong

Learning

Practice

Observation

Learning

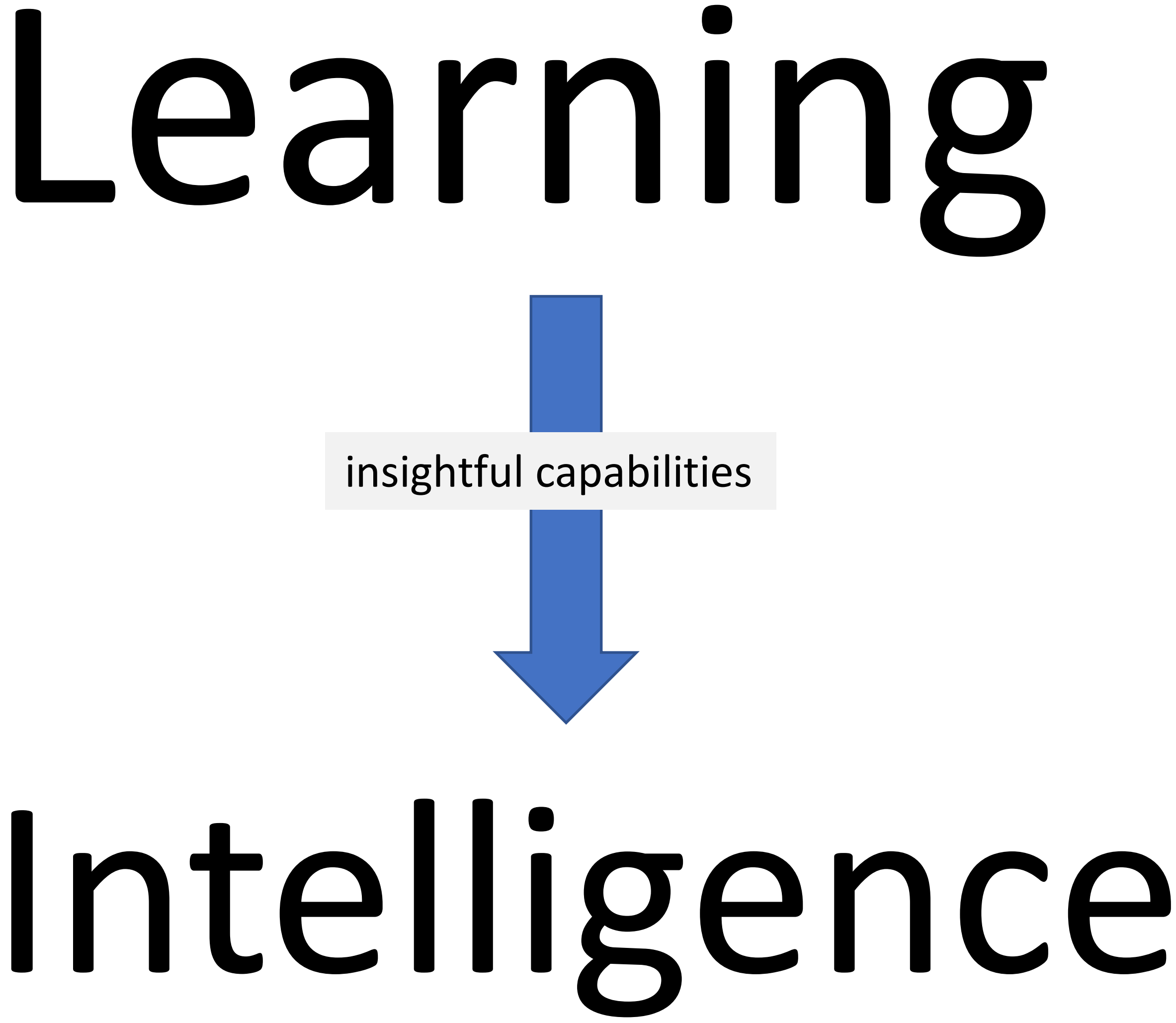
Teaching

Experience

Reflection

Learning

Intelligence



Learning

insightful capabilities

Intelligence

Intelligence

"Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with the environment."

David Wechsler

Intelligence

"Intelligence is the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with the environment."

David Wechsler

spatial

linguistic

logical-mathematical

interpersonal

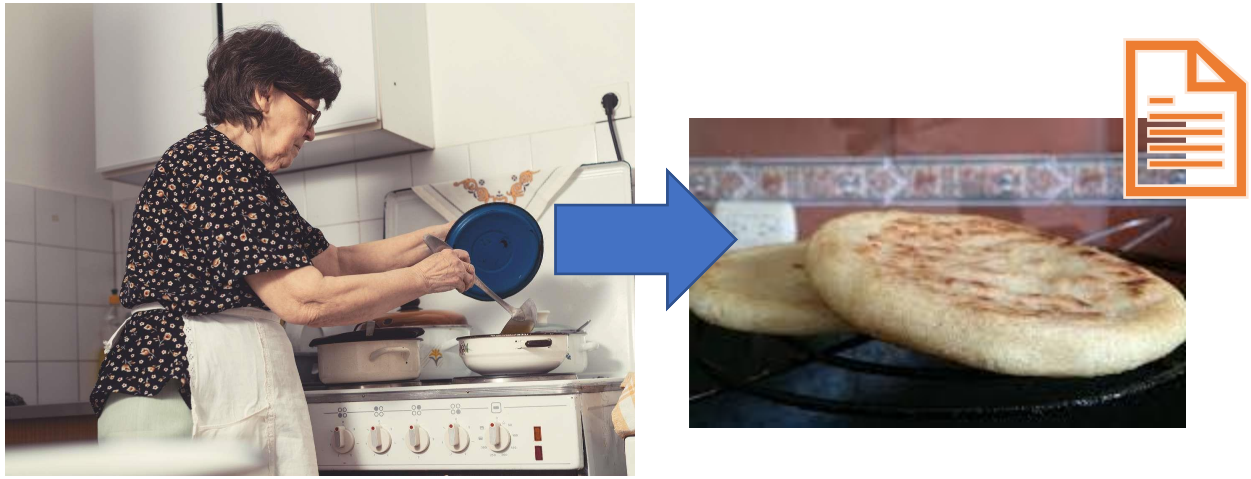
Artificial Intelligence (AI)

The term Artificial Intelligence (AI) was coined by John McCarthy in 1956. Numerous definitions for AI have been proposed by scientists and researchers such as:

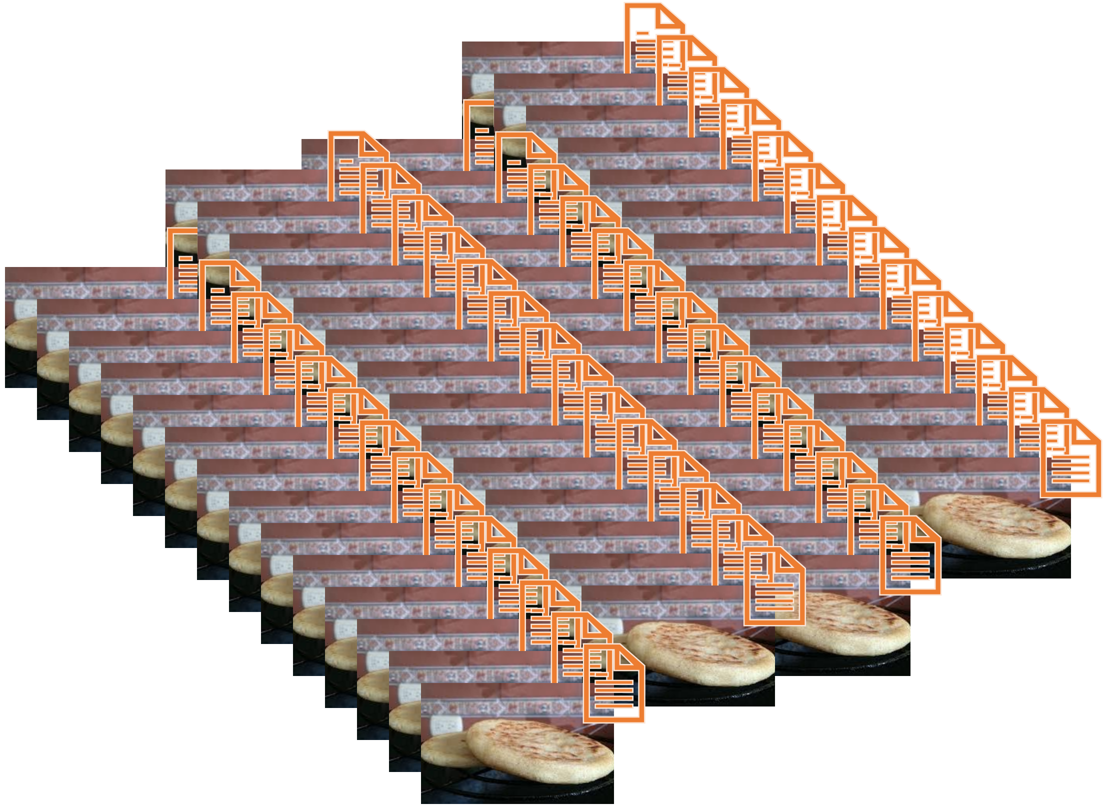
- AI is the field of study focused on creating programs and machines that **can perform tasks that typically require human intelligence**, such as reasoning, learning, perception, and language.

Big Data

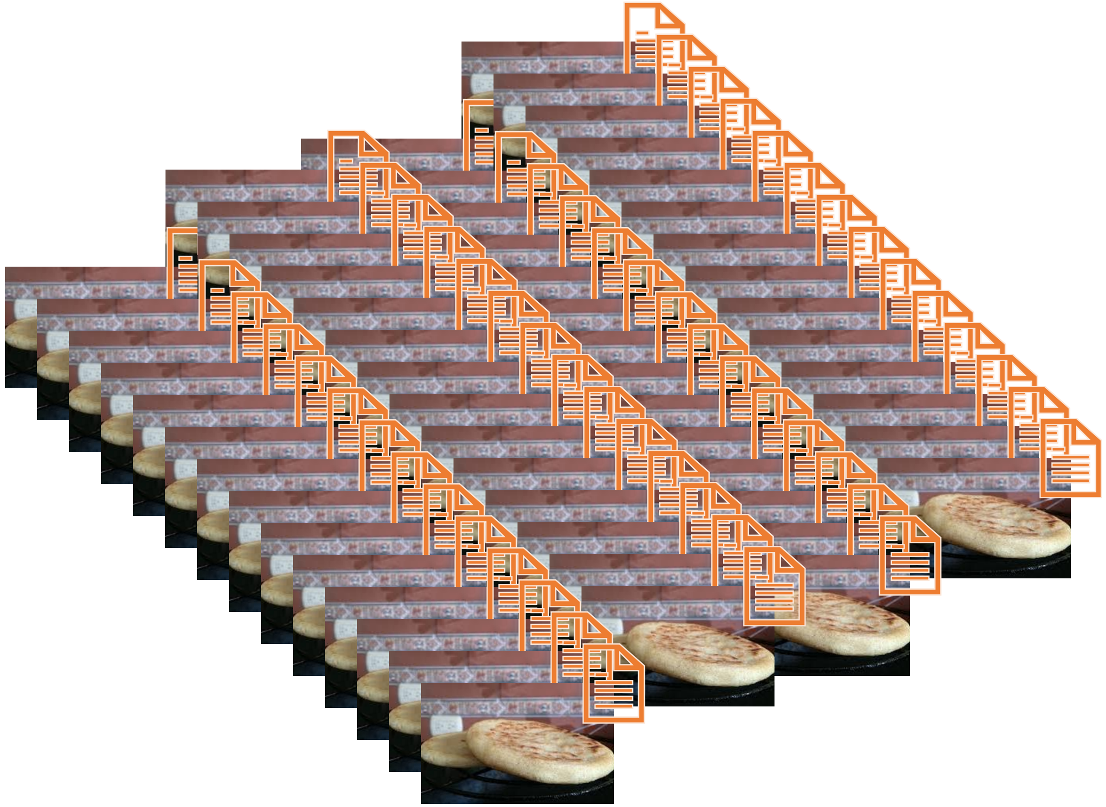
The term big data specifically refers to **large data sets whose size is so large that the quantity can no longer fit into the memory that computers use for processing**. This data can be captured, stored, communicated, aggregated, and analyzed. There is no specific definition of the size of big data, such as the number of terabytes or gigabytes.



DATA



BIG DATA

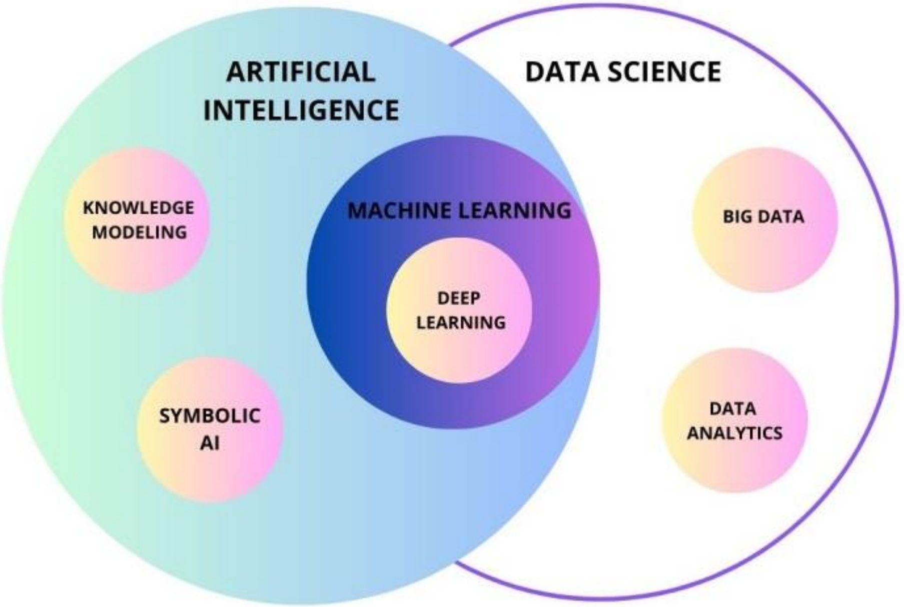


BIG
DATA

ARTIFICIAL
INTELLIGENCE
(AI)

Differences Big data and Artificial Intelligence

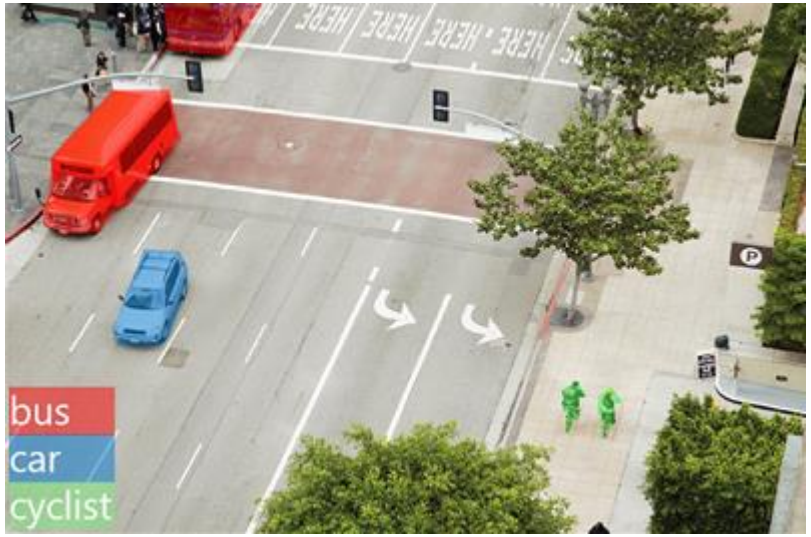
Big data and AI are often used in conjunction with one another, but each fulfill very different roles, one is information and the other is a treatment of that information



https://www.linkedin.com/pulse/unraveling-artificial-intelligence-data-science-barrios-morales-agdof/

ARTIFICIAL INTELLIGENCE (AI)

AI is software that imitates human behaviors and capabilities. Key workloads include



Machine learning - This is often the foundation for an AI system, and is the way we "teach" a computer model to make predictions and draw conclusions from data.

Computer vision - Capabilities within AI to interpret the world visually through cameras, video, and images.

Natural language processing - Capabilities within AI for a computer to interpret written or spoken language, and respond in kind

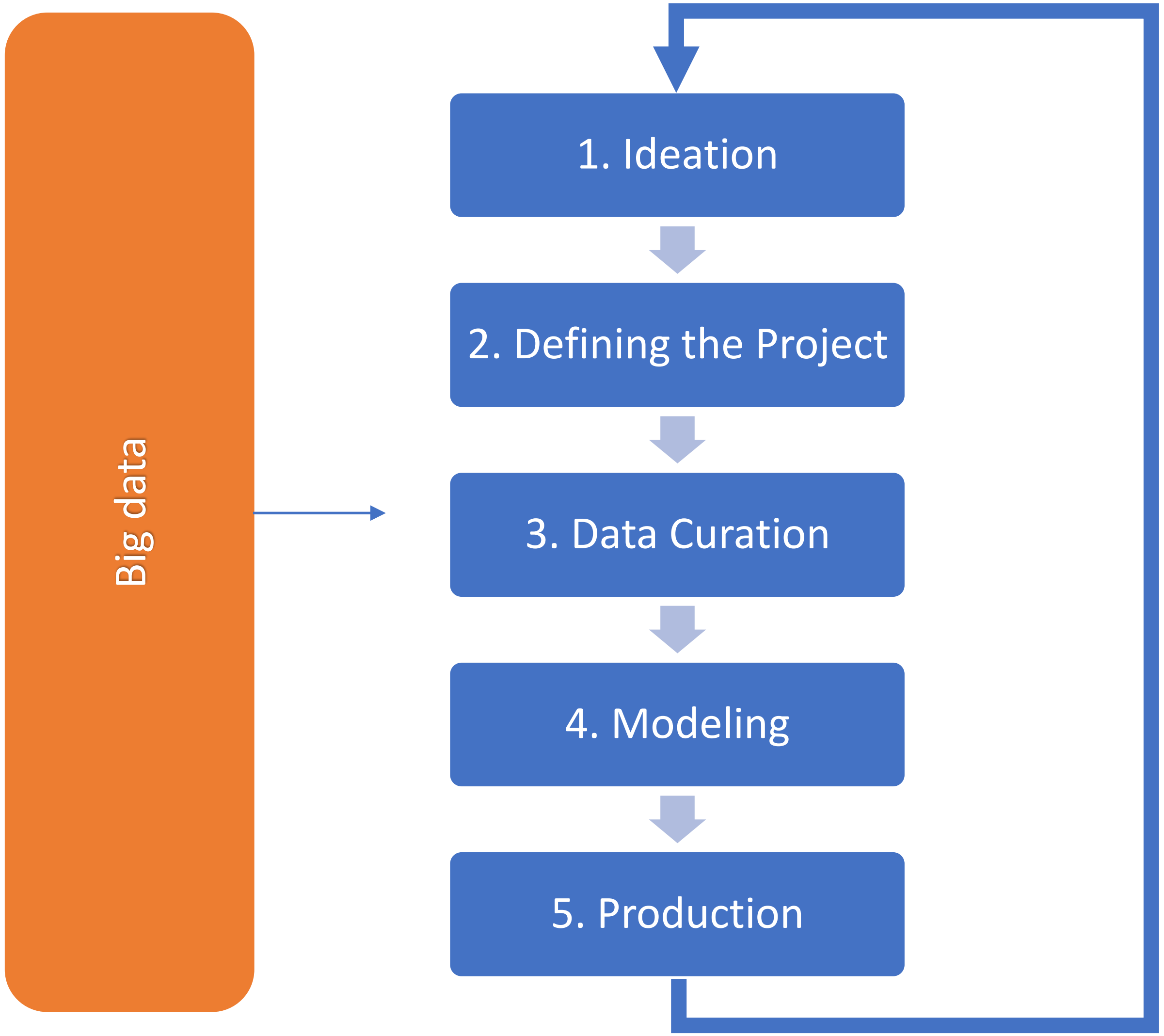
Document intelligence - Capabilities within AI that deal with managing, processing, and using high volumes of data found in forms and documents.

Knowledge mining - Capabilities within AI to extract information from large volumes of often unstructured data to create a searchable knowledge store.

Generative AI - Capabilities within AI that create original content in a variety of formats including natural language, image, code, and more

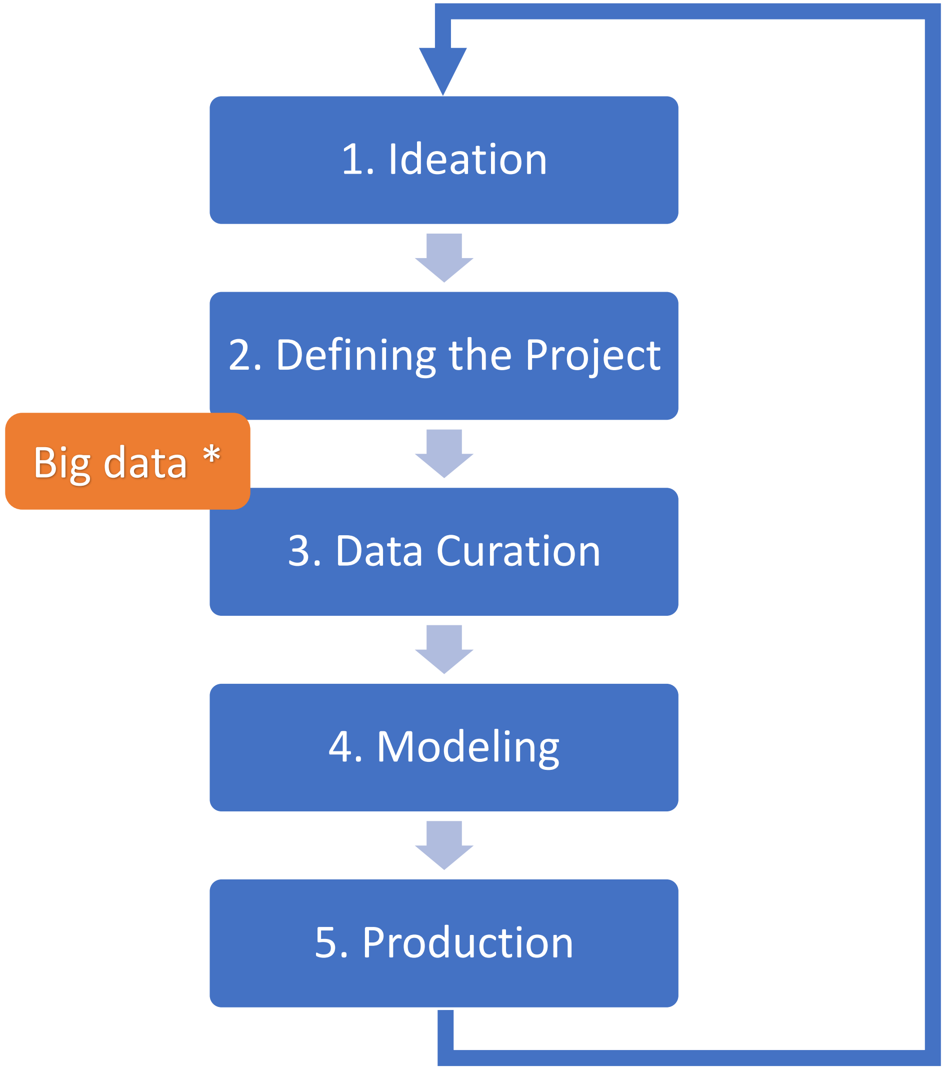
How to Get Started?

AI as a Project NOT as a method



How to Get Started?

AI as a Project NOT as a method



* Not necessary but see it here for reason of this course

1. Ideation

Ideation

This is the initial stage where ideas for the AI project are generated.

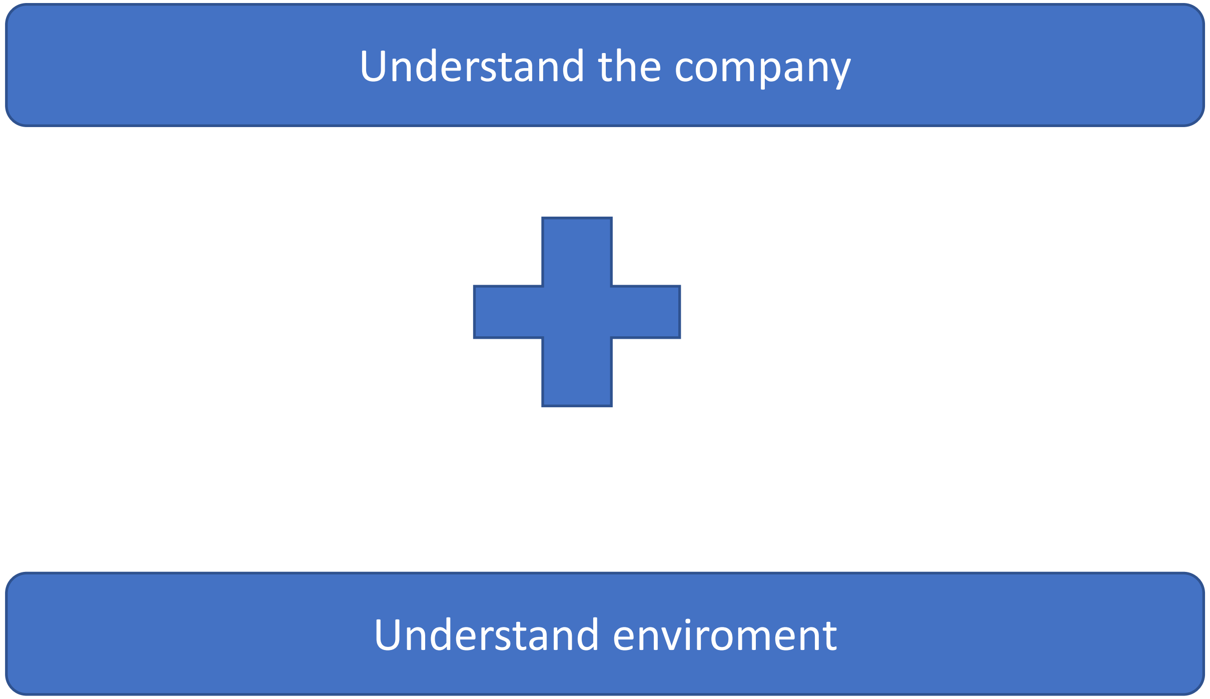
- Identifying the problems to solve
- Analyzing opportunities
- Defining the purpose of the project.

Some methodologies

- Brainstorming
- Problem Framing
- User Persona
- Empathy Mapping
- Pain Points
- Innovation Pipeline
- Idea Validation
- Ideation Workshop

Some methodologies

- Brainstorming
- Problem Framing
- User Persona
- Empathy Mapping
- Pain Points
- Innovation Pipeline
- Idea Validation
- Ideation Workshop



Understand the company

Understand enviroment

Ideation

- Identifying the problems to solve
- Analyzing opportunities
- Defining the purpose of the project.

Example: I have a dinner at home.

Ideation

- Identifying the problems to solve:
  - friend is vegetarian?
  - Any allergies?
- Analyzing opportunities
  - What I can cook?
- Defining the purpose of the project.
  - I will cook…

Example: I have a dinner at home.

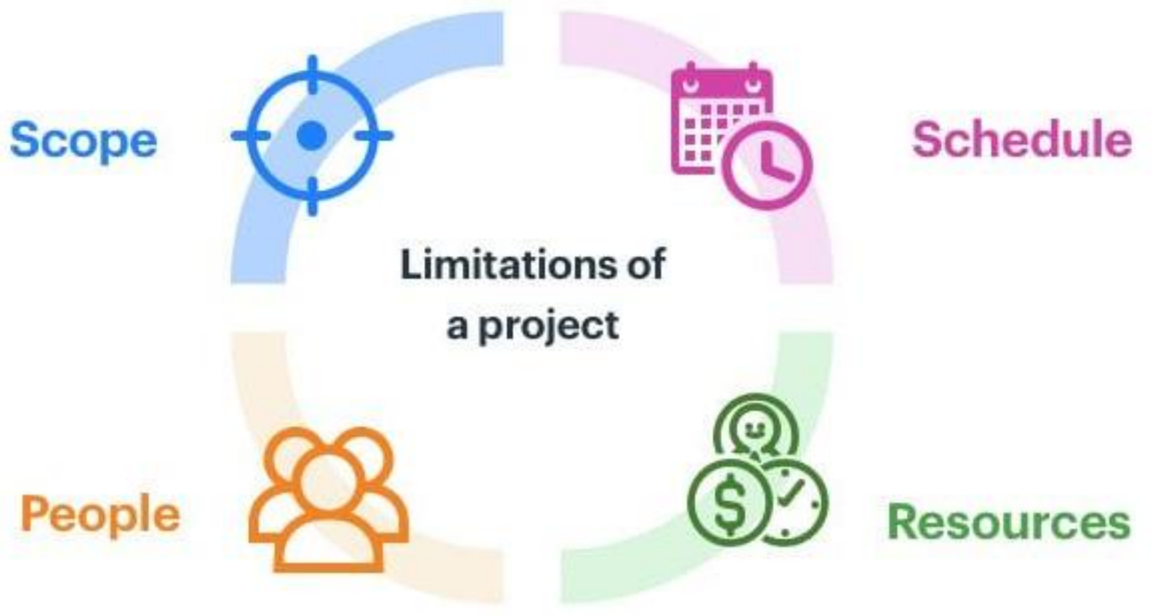
Ideation

- Identifying the problems to solve:
  - friend is vegetarian?
  - Any allergies?
- Analyzing opportunities
  - What I can cook?
- Defining the purpose of the project.
  - I will cook…

Example: I have a dinner at home.

2. Defining the Project

- At this stage, the project scope, objectives, specific goals, and success metrics are clearly defined.
- The necessary resources are also identified.
  - Time
  - Technology
  - team members

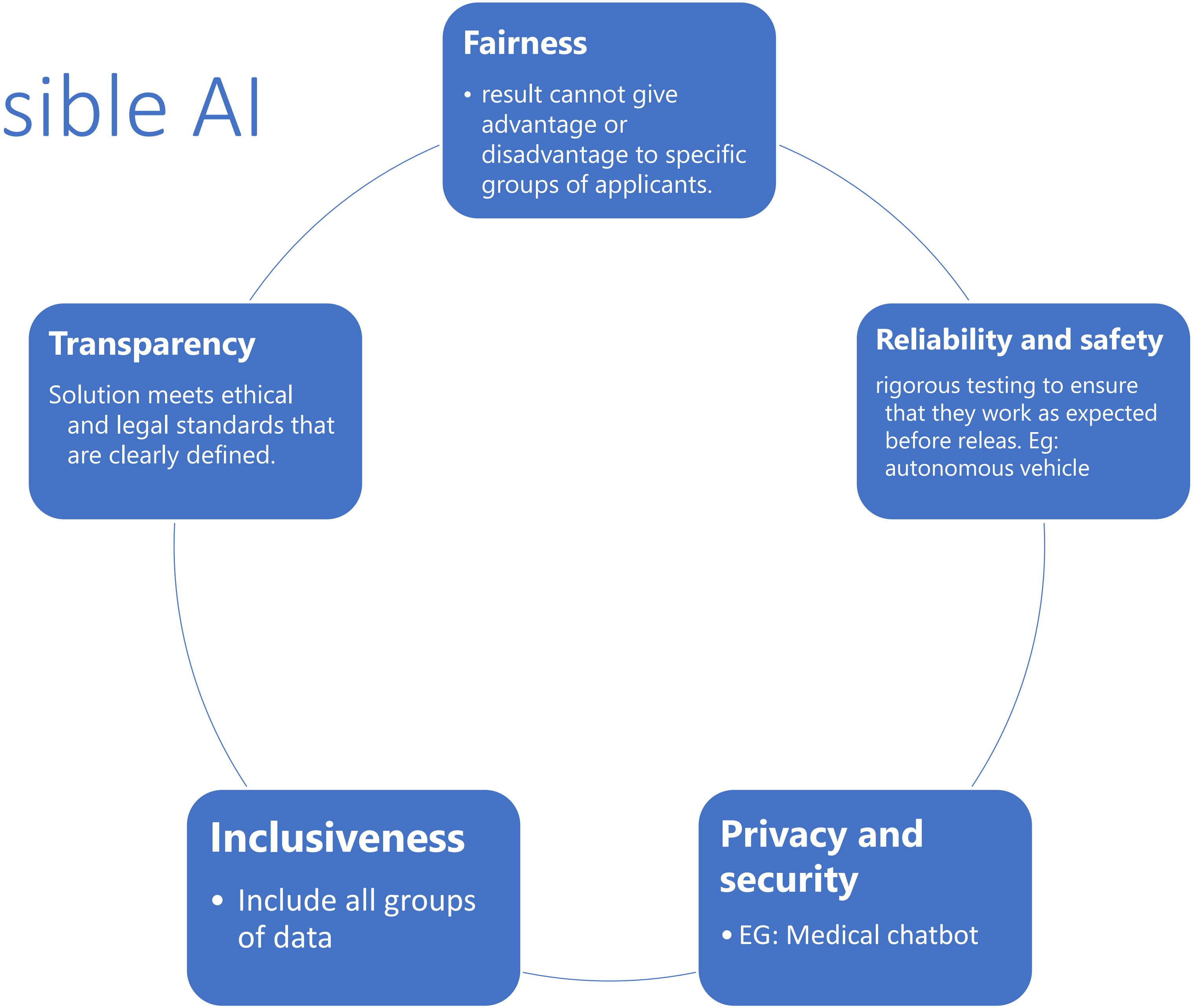


Challenges and risks with AI

Challenge or Risk	Example
Bias can affect results	A loan-approval model discriminates by gender due to bias in the data with which it was trained
Errors may cause harm	An autonomous vehicle experiences a system failure and causes a collision
Data could be exposed	A medical diagnostic bot is trained using sensitive patient data, which is stored insecurely
Solutions may not work for everyone	A home automation assistant provides no audio output for visually impaired users
Users must trust a complex system	An AI-based financial tool makes investment recommendations - what are they based on?
Who's liable for AI-driven decisions?	An innocent person is convicted of a crime based on evidence from facial recognition – who's responsible?

2. Defining the Project

Responsible AI



Fairness

- result cannot give advantage or disadvantage to specific groups of applicants.

Reliability and safety

rigorous testing to ensure that they work as expected before releas. Eg: autonomous vehicle

Privacy and security

- EG: Medical chatbot

Inclusiveness

- Include all groups of data

Transparency

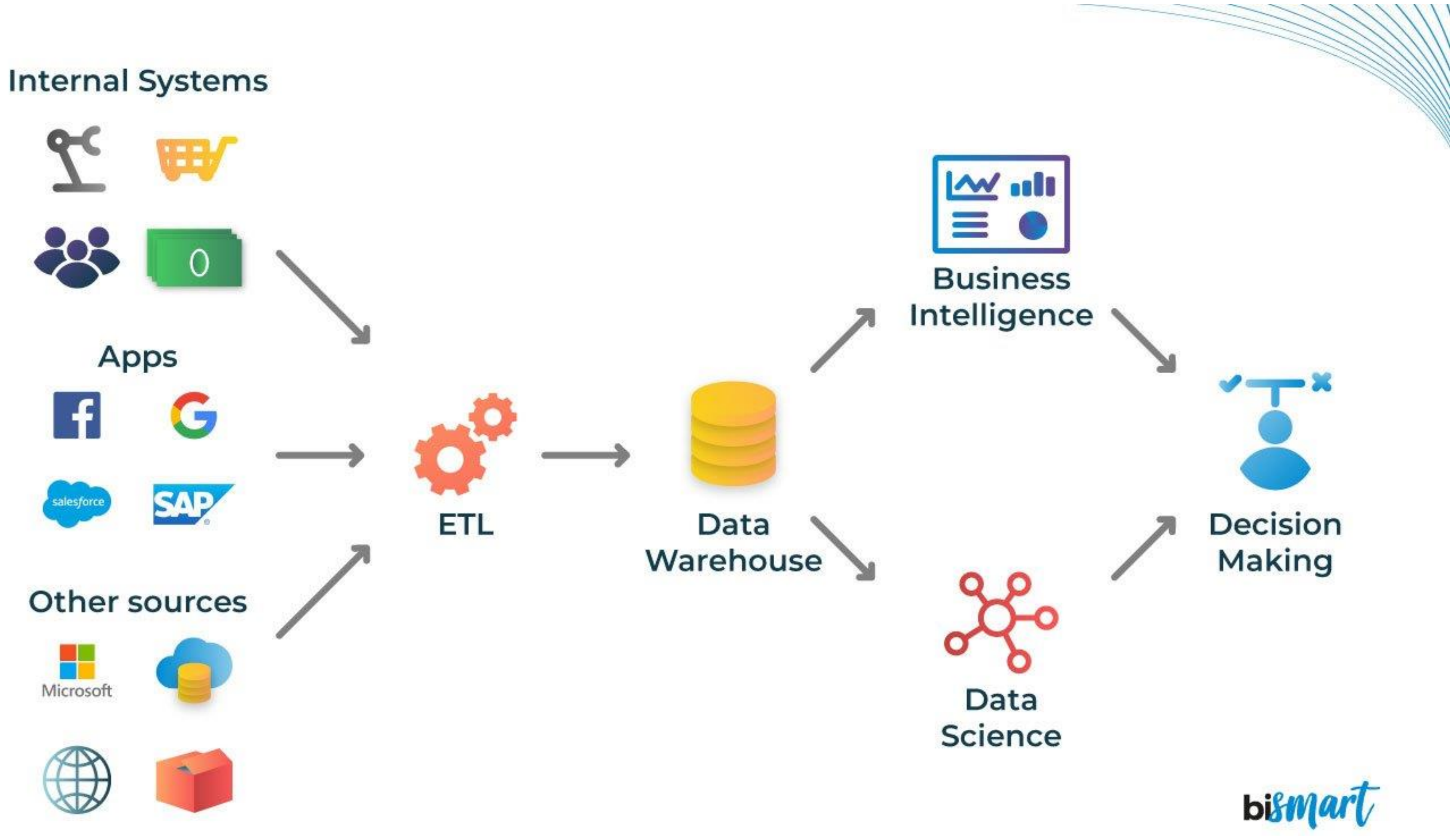
Solution meets ethical and legal standards that are clearly defined.

3. Data Curation

- This step focuses on collecting, cleaning, organizing, and preparing the data to be used in the project.
- It involves:
    - identifying relevant data sources
    - removing duplicates or incorrect data
    - formatting the data to make it usable.

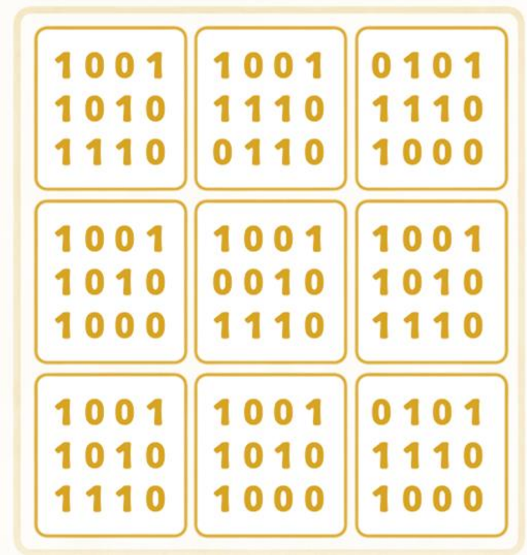
Data refers to raw facts, figures, or pieces of information that are collected and recorded for analysis, processing, or storage.

3. Data Curation



Types of data

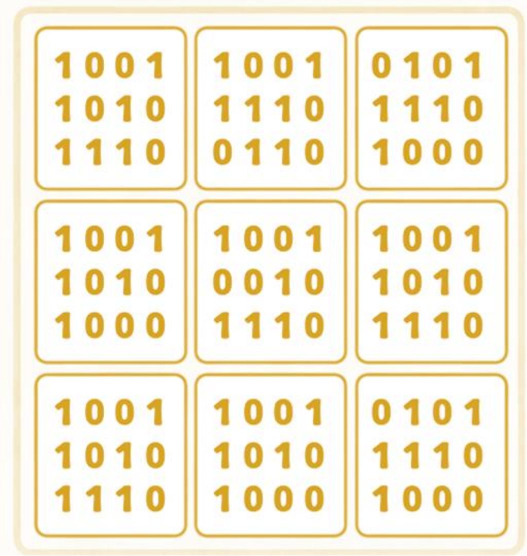
Structured Data



highly organized information that is easily searchable and typically stored in relational databases or spreadsheets

Types of data

Structured Data



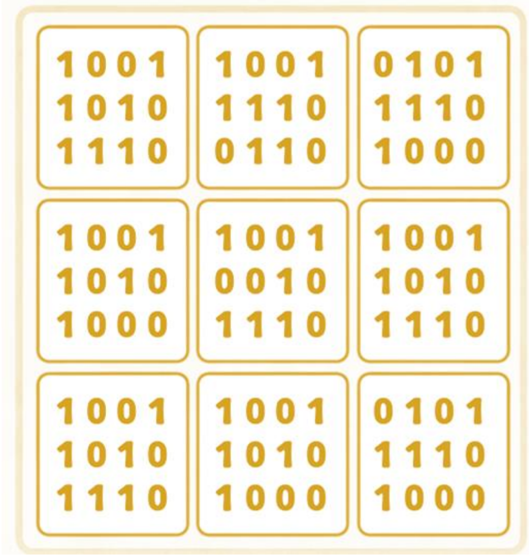
highly organized information that is easily searchable and typically stored in relational databases or spreadsheets

Unstructured Data

Unstructured data lacks a pre-defined data model, making it more difficult to collect, process and analyze

Types of data

Structured Data



highly organized information that is easily searchable and typically stored in relational databases or spreadsheets

Semi- Structured Data

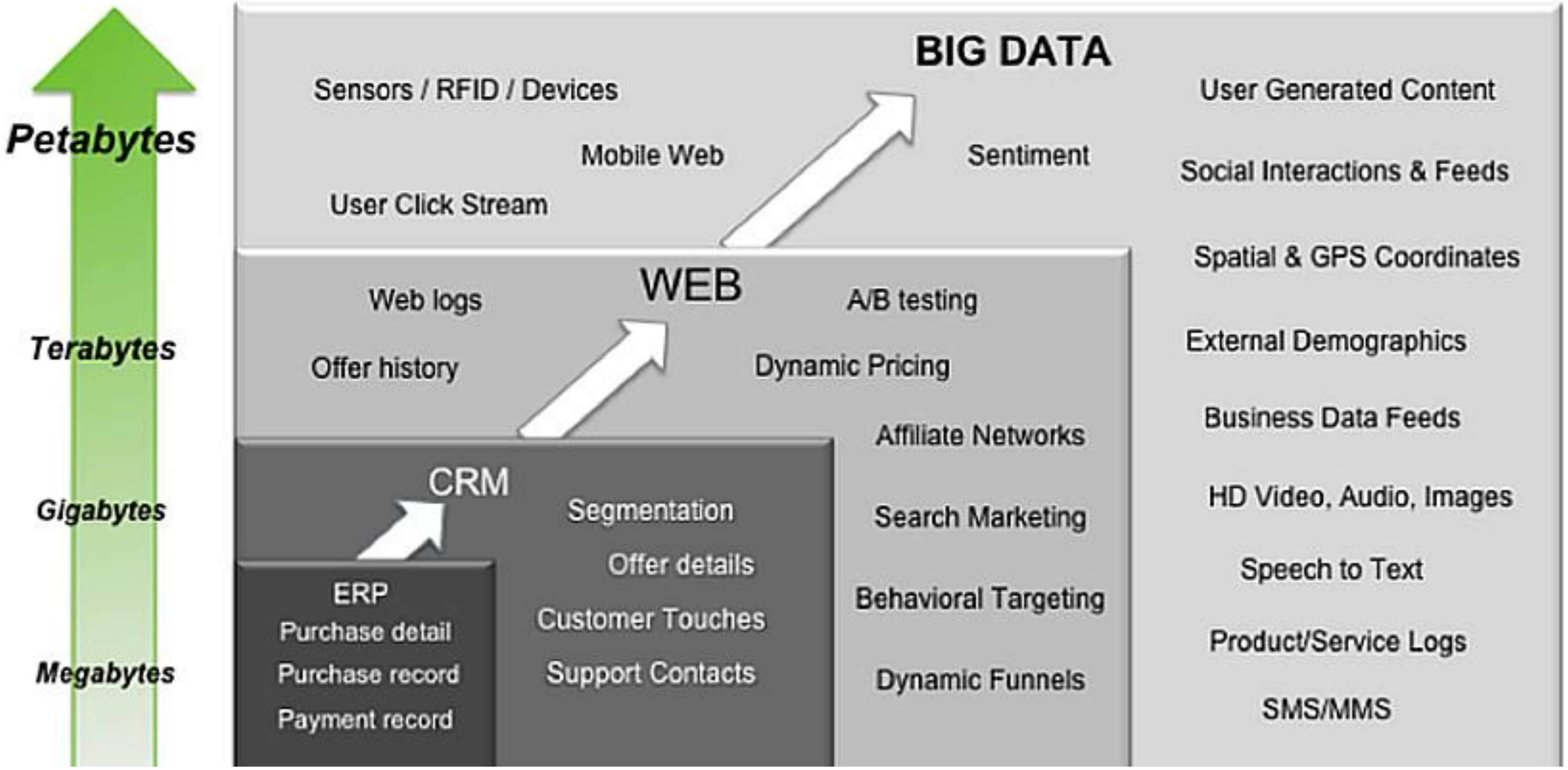
While it does not reside in a relational database, it contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data

Unstructured Data

Unstructured data lacks a pre-defined data model, making it more difficult to collect, process and analyze

Big Data

Big Data = Transactions + Interactions + Observations



Petabytes

Terabytes

Gigabytes

Megabytes

BIG DATA

Sensors / RFID / Devices

Mobile Web

User Click Stream

Sentiment

User Generated Content

Social Interactions & Feeds

Spatial & GPS Coordinates

External Demographics

Business Data Feeds

HD Video, Audio, Images

Speech to Text

Product/Service Logs

SMS/MMS

WEB

Web logs

Offer history

A/B testing

Dynamic Pricing

Affiliate Networks

Search Marketing

Behavioral Targeting

Dynamic Funnels

CRM

Segmentation

Offer details

Customer Touches

Support Contacts

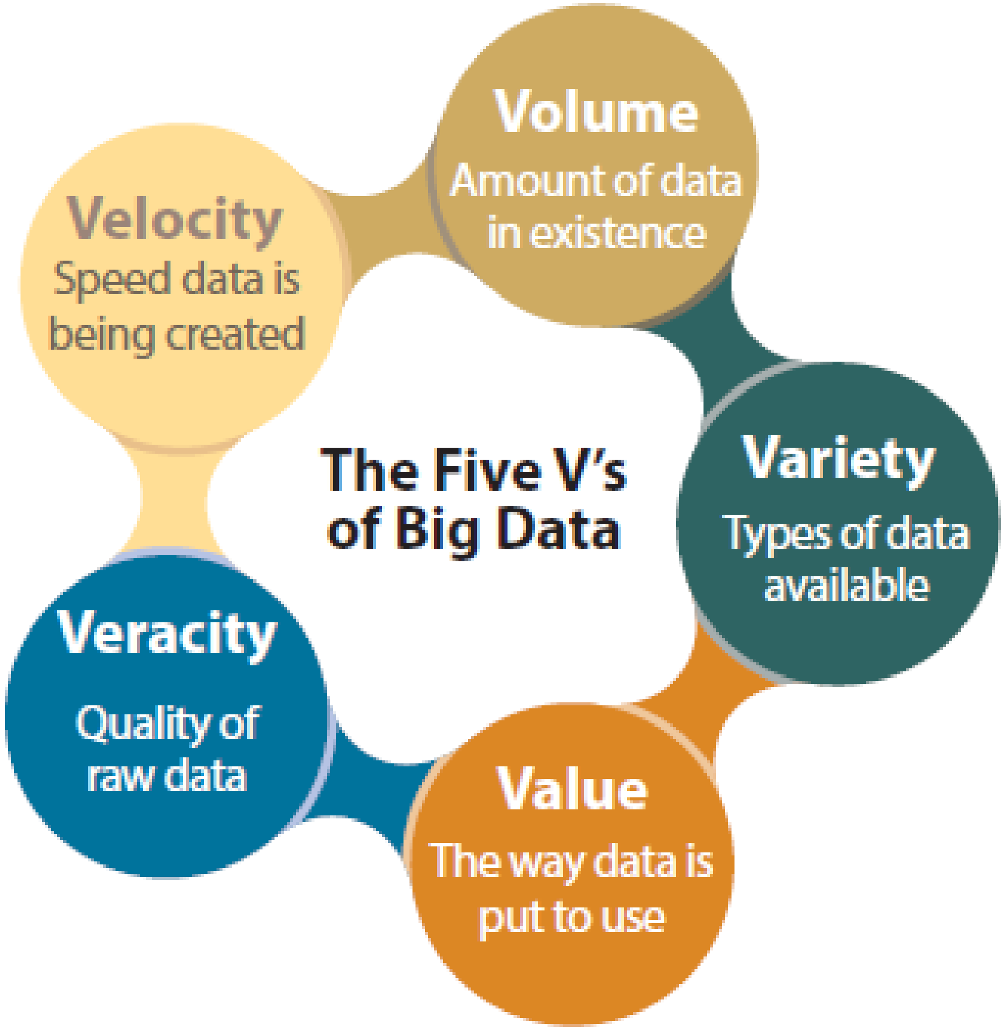
ERP

Purchase detail

Purchase record

Payment record

3. Data Curation



Data cleaning process

- **Understand the Data**: Review structure, key attributes, and metadata.
- **Handle Missing Data**: Identify, remove, or impute missing values.
- **Remove Duplicates**: Detect and drop duplicate records.
- **Resolve Inconsistencies**: Standardize formats, unify text case, and fix typos.
- **Handle Outliers**: Detect, analyze, and decide to retain, transform, or remove.
- **Validate Data Types**: Ensure correct data types and parse dates.
- **Fix Invalid Data**: Identify and correct erroneous or out-of-range values.
- **Normalize and Transform**: Scale numerical data and encode categorical variables.
- **Remove Irrelevant Features**: Drop redundant or irrelevant columns.

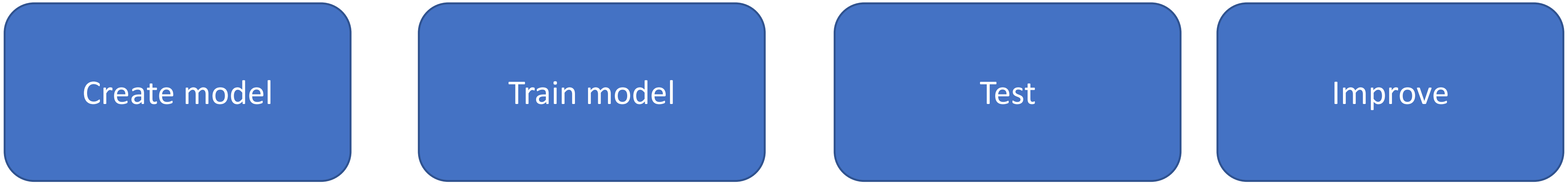
Exploratory Data Analysis

- Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually.

1) EDA Level 0 — Pure Understanding of Original Data
2) EDA Level 1 — Transformation of Original Data
3) EDA Level 2 — Understanding of Transformed Data

Modelling

- During this phase, an initial model or solution is developed to test the project's feasibility.
- This might include training preliminary AI models and conducting tests to validate their performance.



Create model

Train model

Test

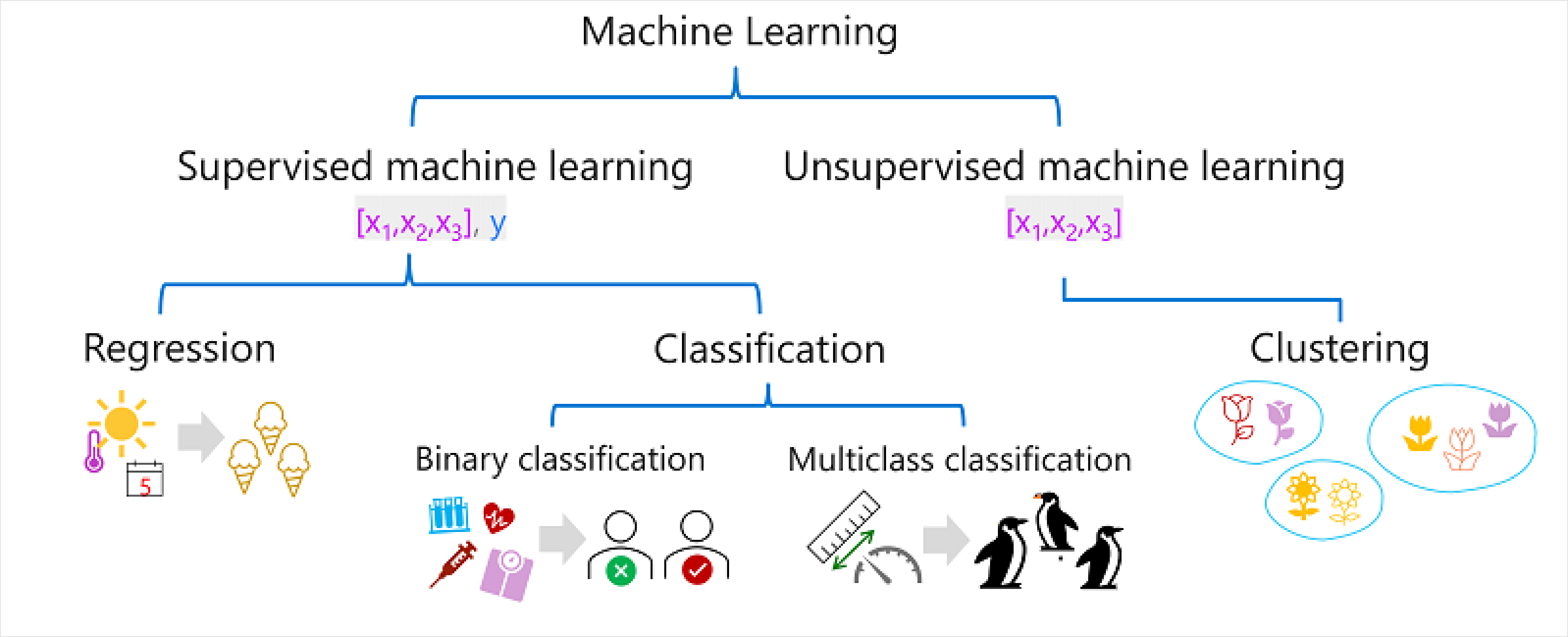
Improve

ARTIFICIAL INTELLIGENCE (AI)

AI is software that imitates human behaviors and capabilities. Key workloads include

Machine learning - This is often the foundation for an AI system, and is the way we "teach" a computer model to make predictions and draw conclusions from data.

4. Modelling


Machine Learning
Supervised machine learning
$[x_1,x_2,x_3]$, y
Unsupervised machine learning
$[x_1,x_2,x_3]$
Regression
5
Classification
Binary classification
Multiclass classification
Clustering

supervised learning

given a data set of input-output pairs, learn a function to map inputs to outputs

4. Modelling

Date	**Humidity** (relative humidity)	**Pressure** (sea level, mb)	**Rain**

Date	**Humidity** (relative humidity)	**Pressure** (sea level, mb)	**Rain**
January 1	93 %	999,7	Rain
January 2	49 %	1015,5	No Rain
January 3	79 %	1031,1	No Rain
January 4	65 %	984,9	Rain
January 5	90 %	975,2	Rain

Date	**Humidity** (relative humidity)	**Pressure** (sea level, mb)	**Rain**
January 1	93 %	999,7	Rain
January 2	49 %	1015,5	No Rain
January 3	79 %	1031,1	No Rain
January 4	65 %	984,9	Rain
January 5	90 %	975,2	Rain
	X1	X2	y

$$f(humidity, pressure)$$

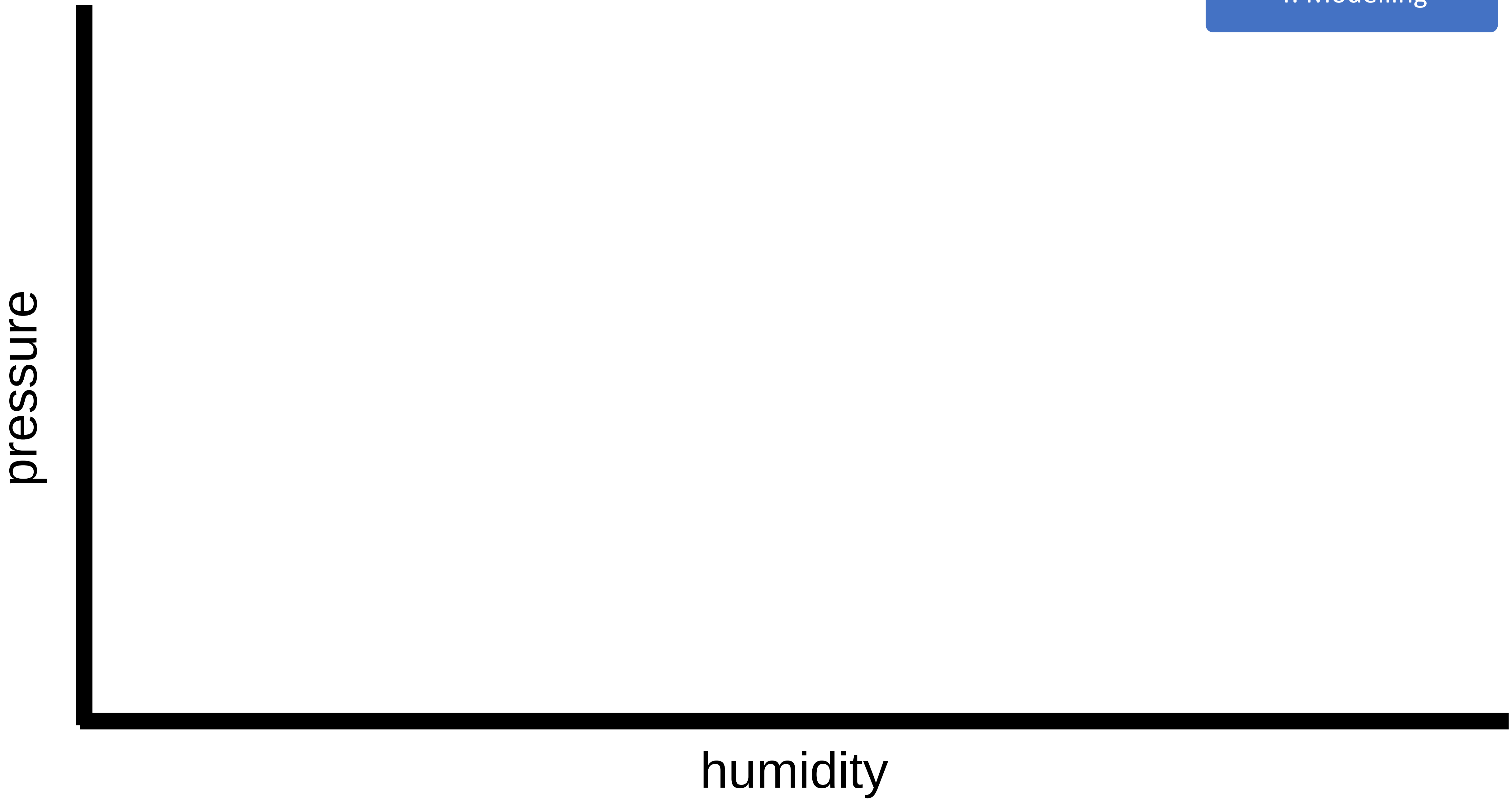
$$f(93, 999.7) = Rain$$

$$f(49, 1015.5) = No\ Rain$$

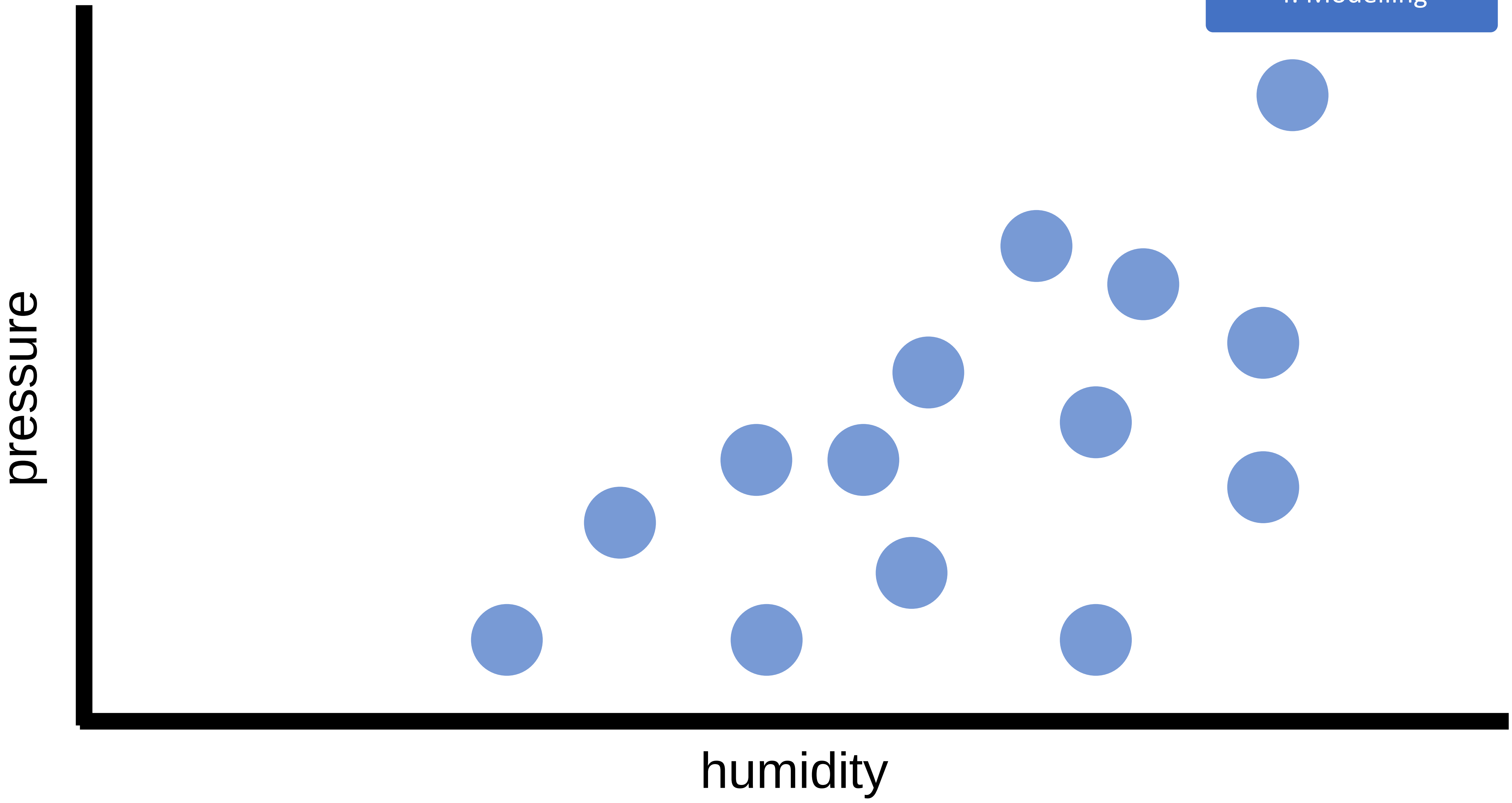
$$f(79, 1031.1) = No\ Rain$$

$h(humidity, pressu$

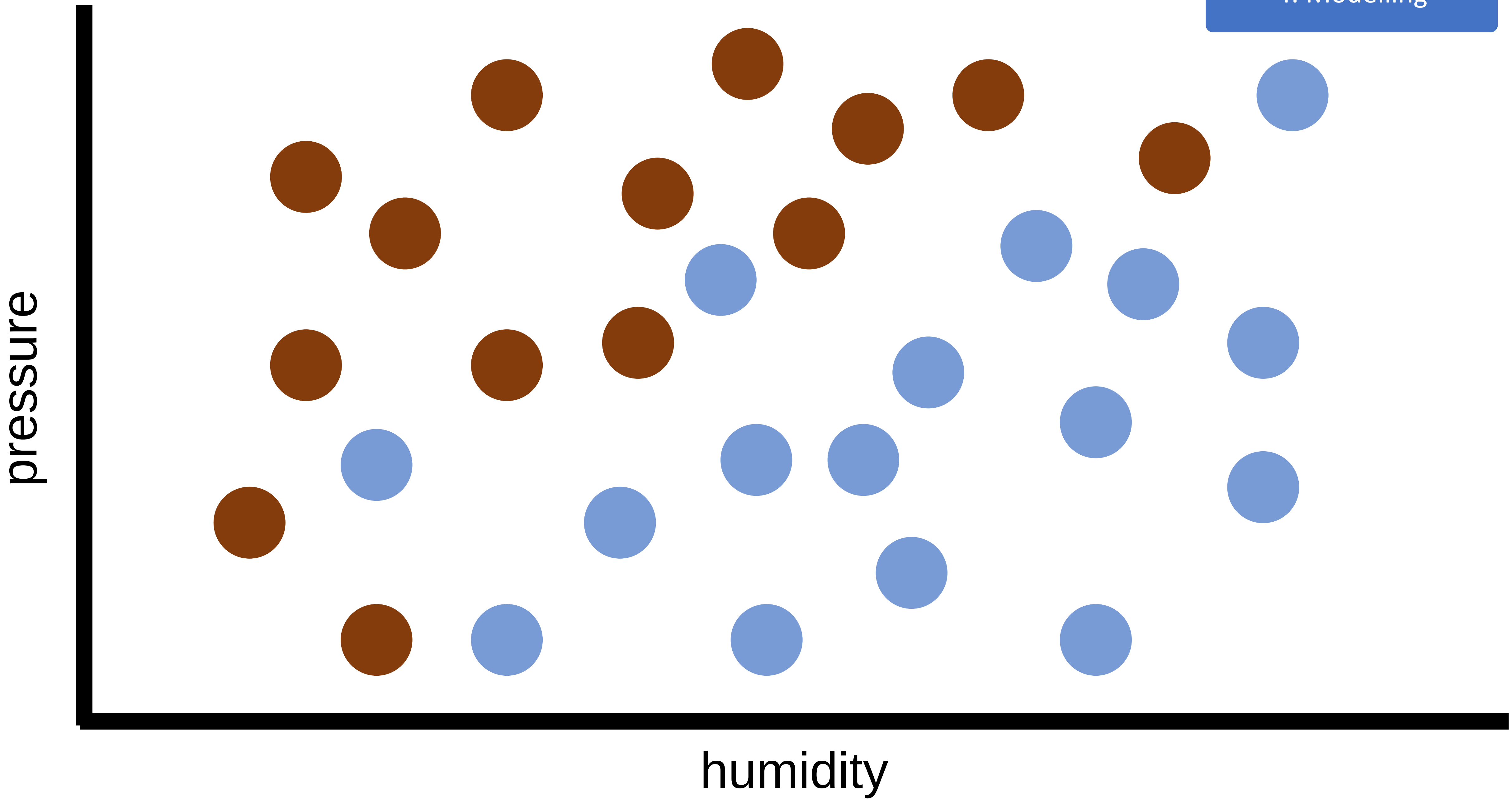
4. Modelling



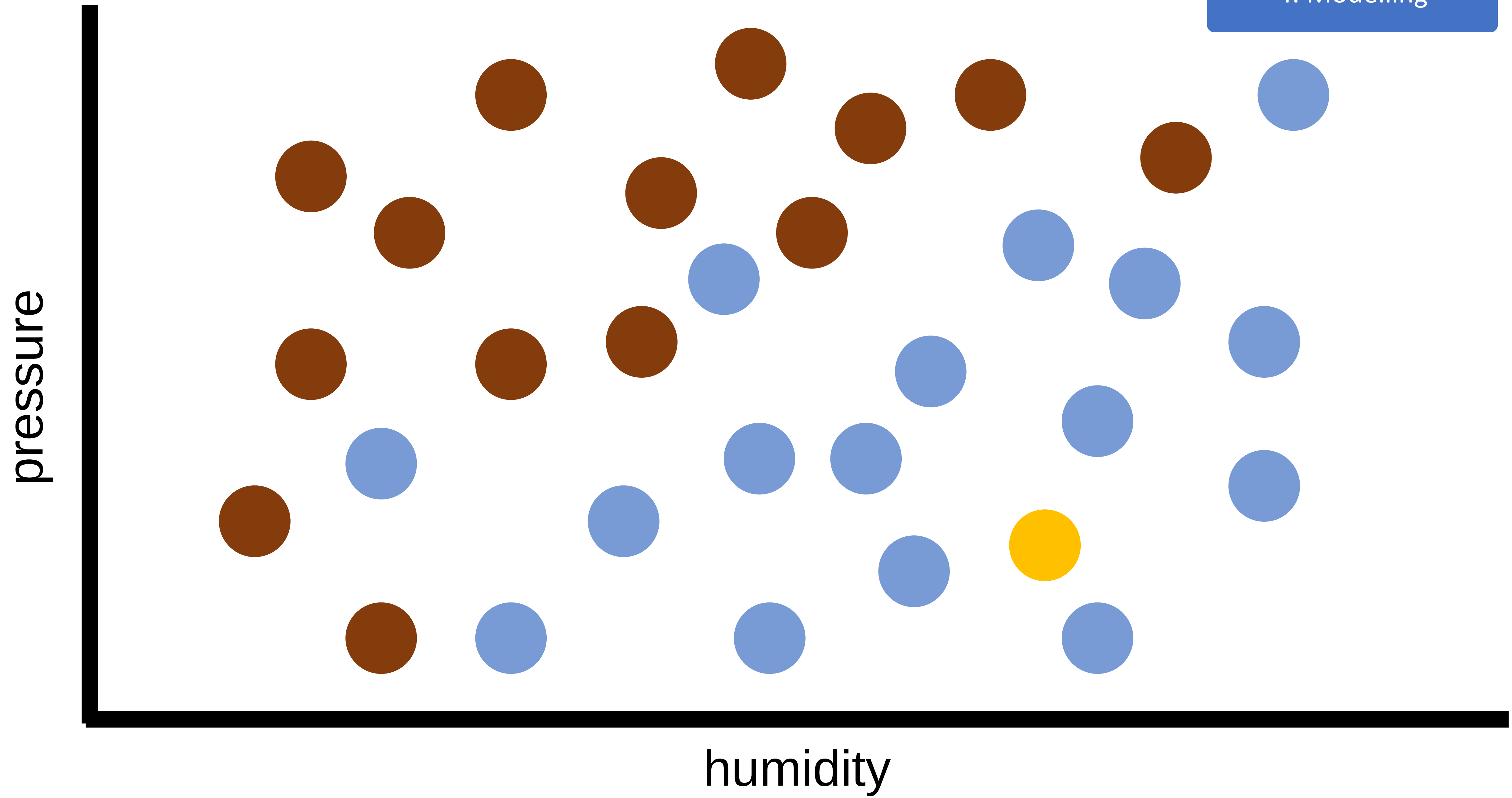
4. Modelling



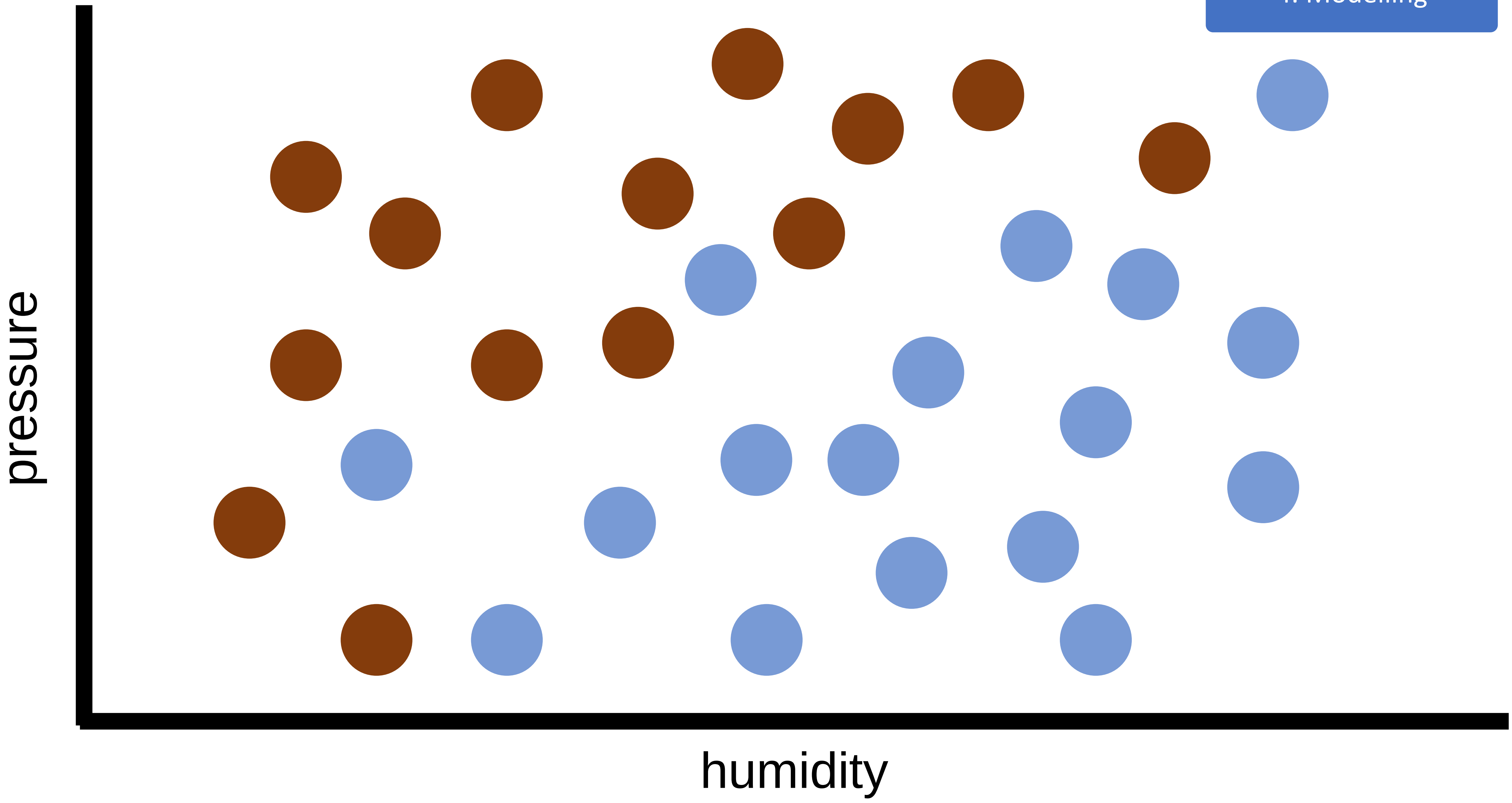
4. Modelling



4. Modelling



4. Modelling



unsupervised learning

given input data without any additional feedback, learn patterns

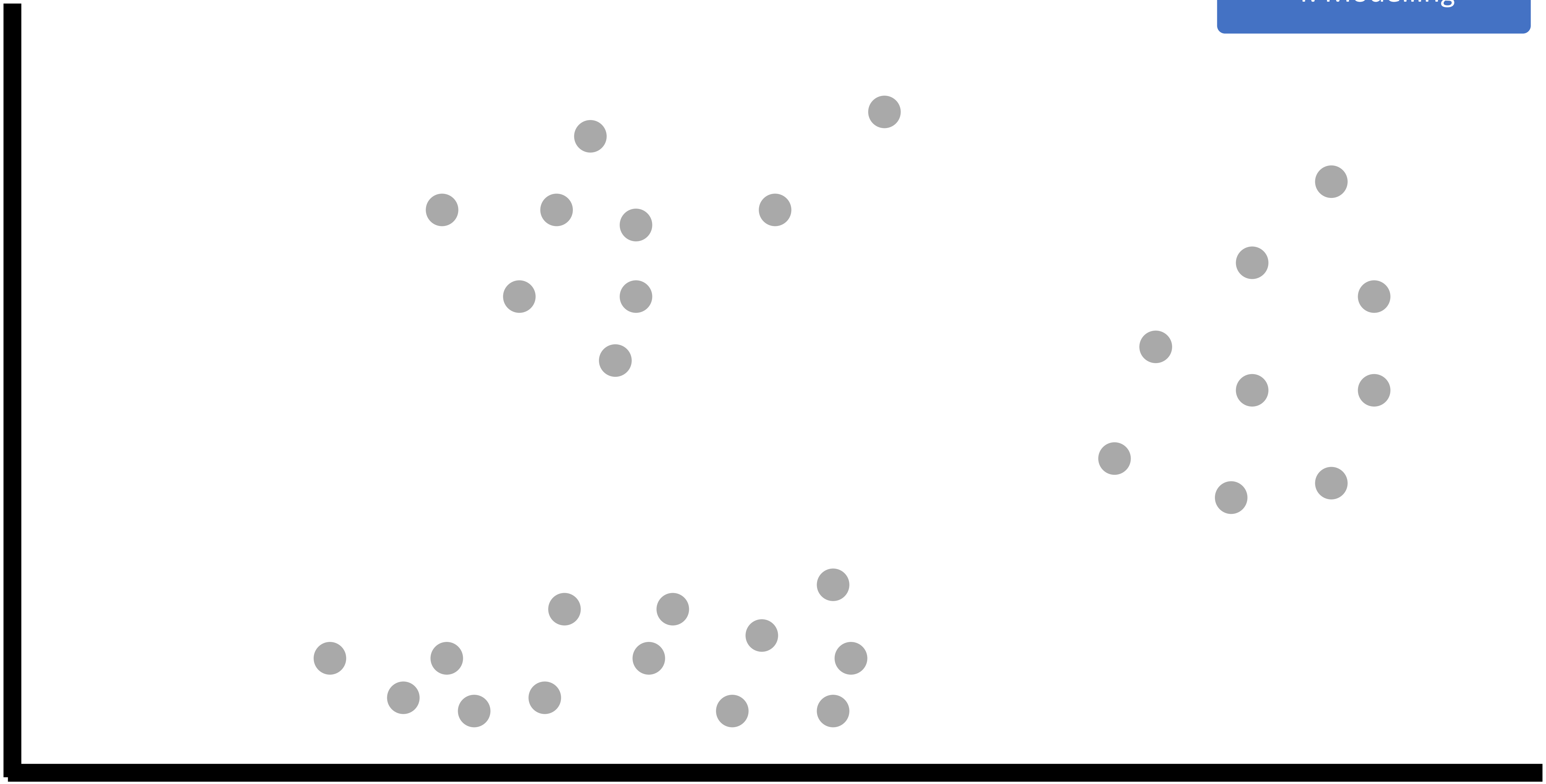
unsupervised learning

given input data without any additional feedback, learn patterns

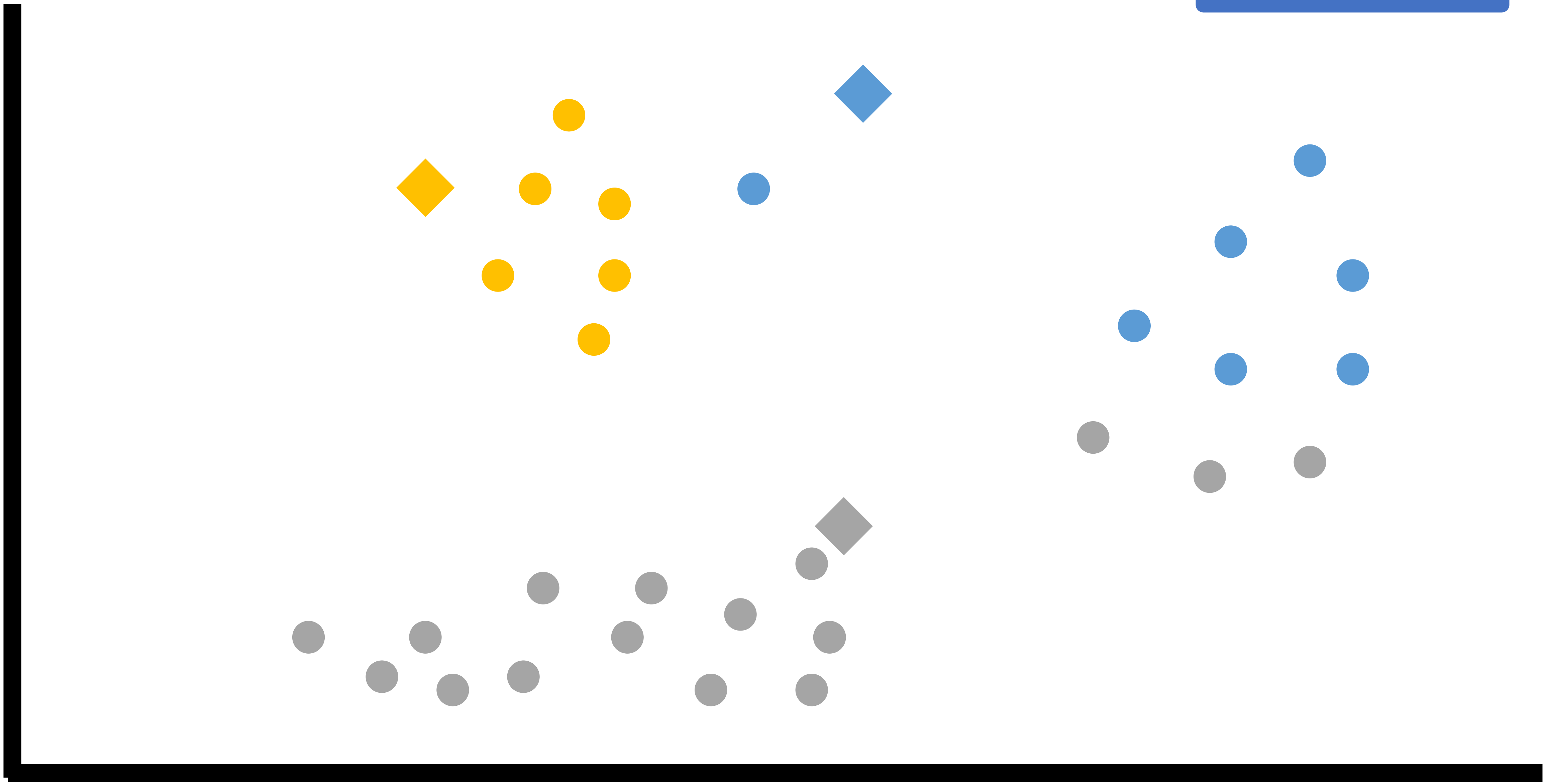
clustering

organizing a set of objects into groups in such a way that similar objects tend to be in the same group

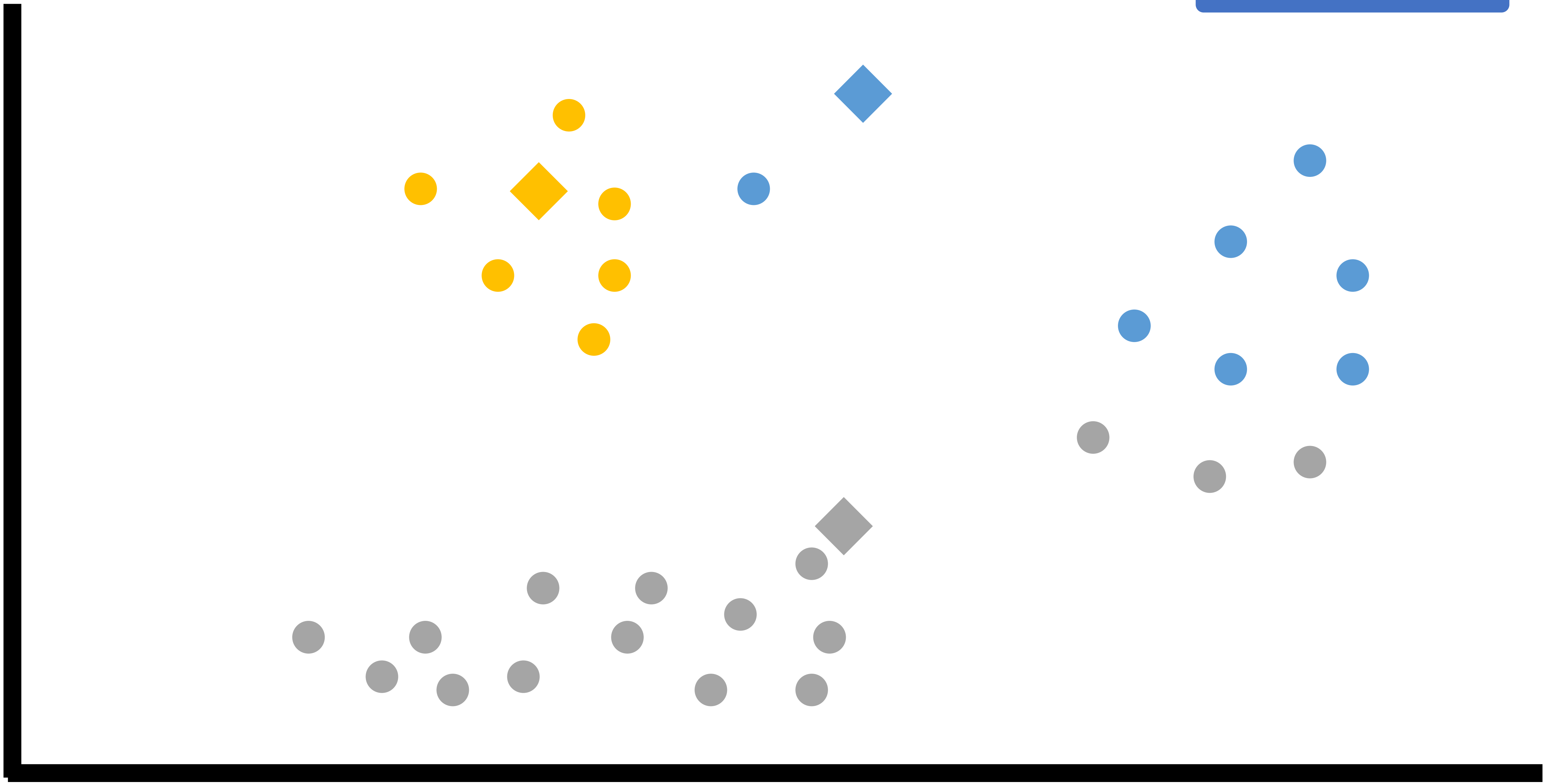
4. Modelling



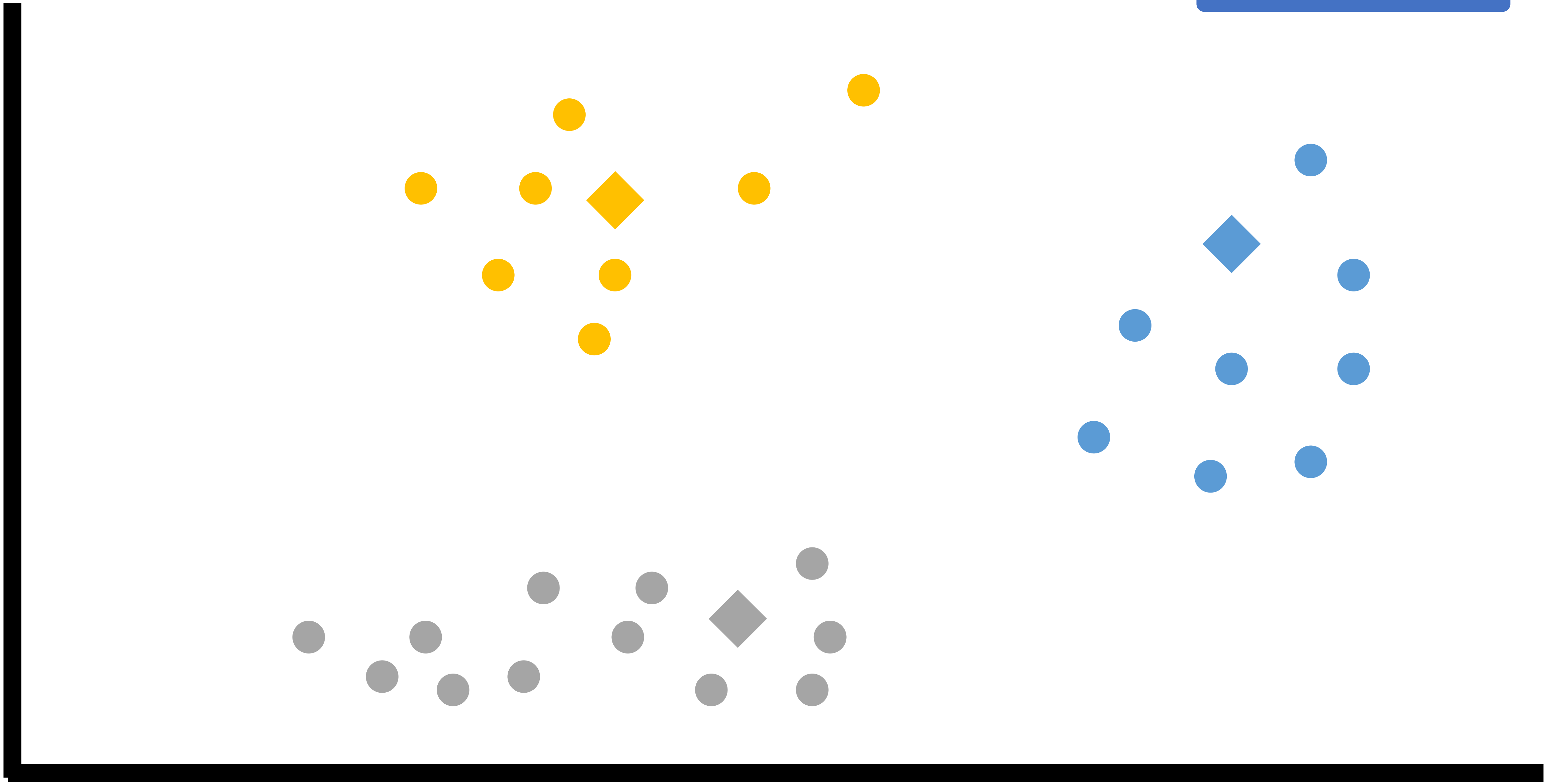
4. Modelling



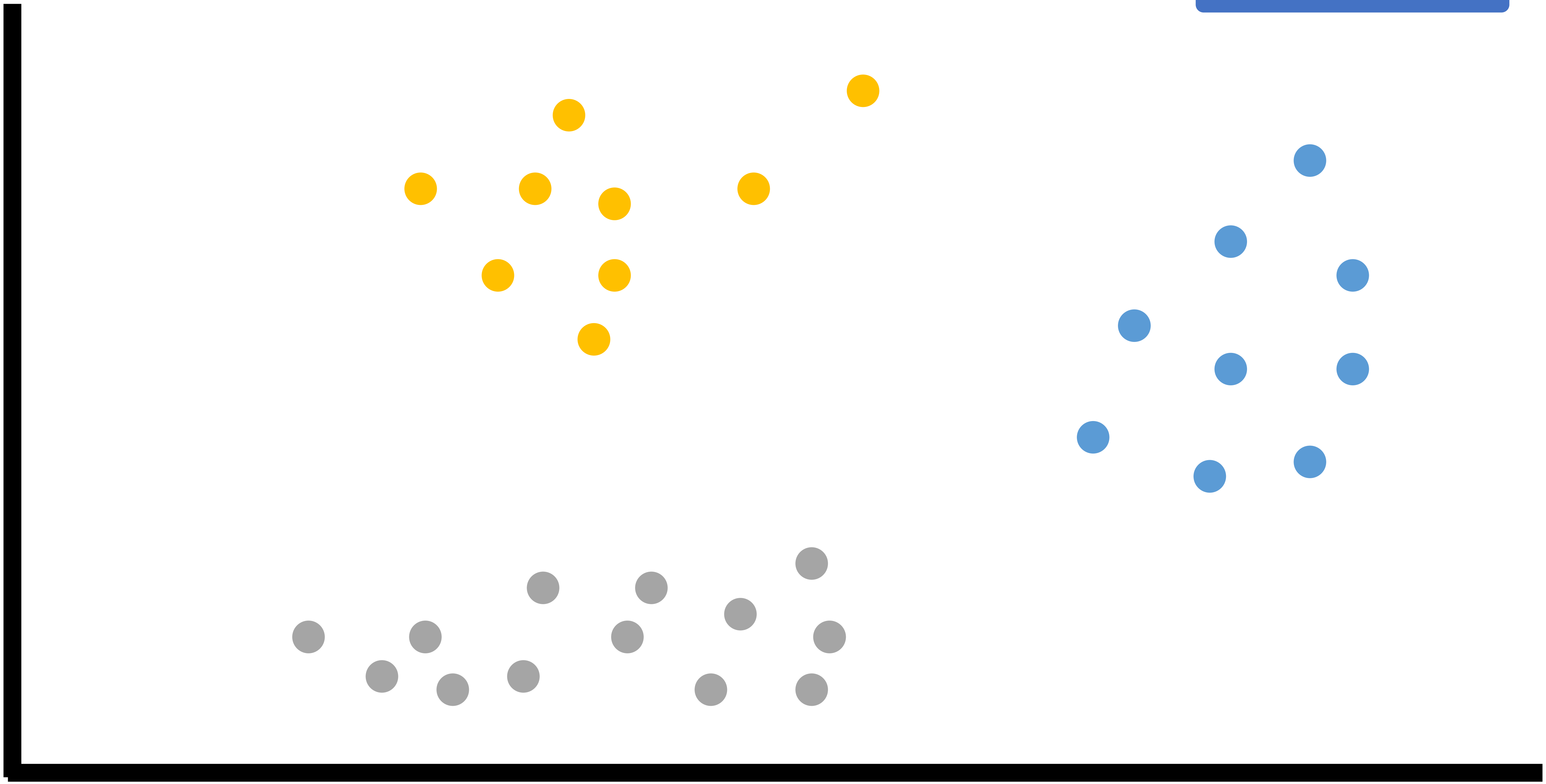
4. Modelling



4. Modelling



4. Modelling



Production

- In this step, the final solution is deployed in a real-world environment. This includes optimizing the model, integrating the system into existing infrastructure, and monitoring its real-time performance.

- https://colab.research.google.com/drive/1Q9aqQJY5oXer2Y5wqXS7TvFgHctr4CFB?usp=sharing