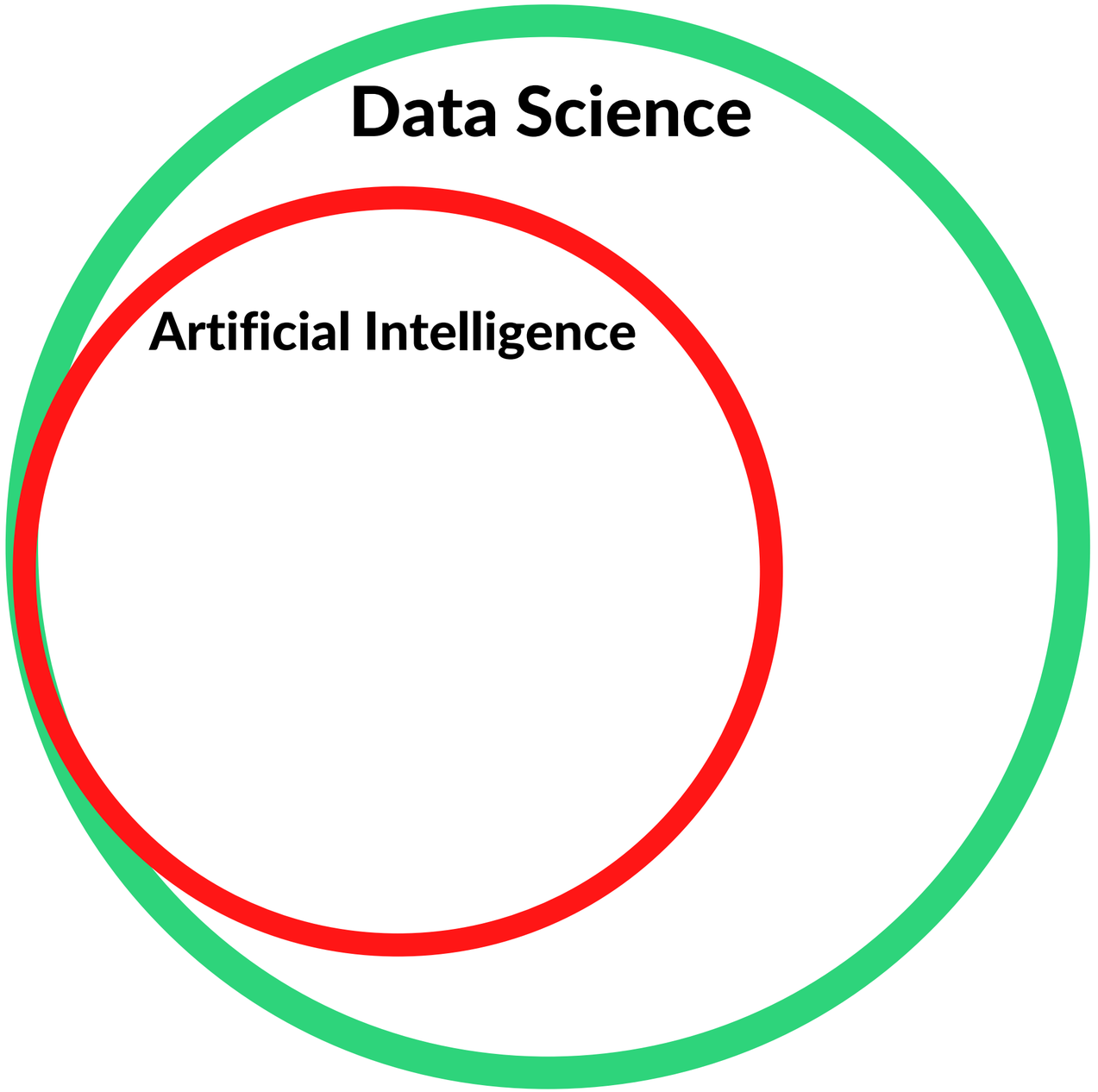
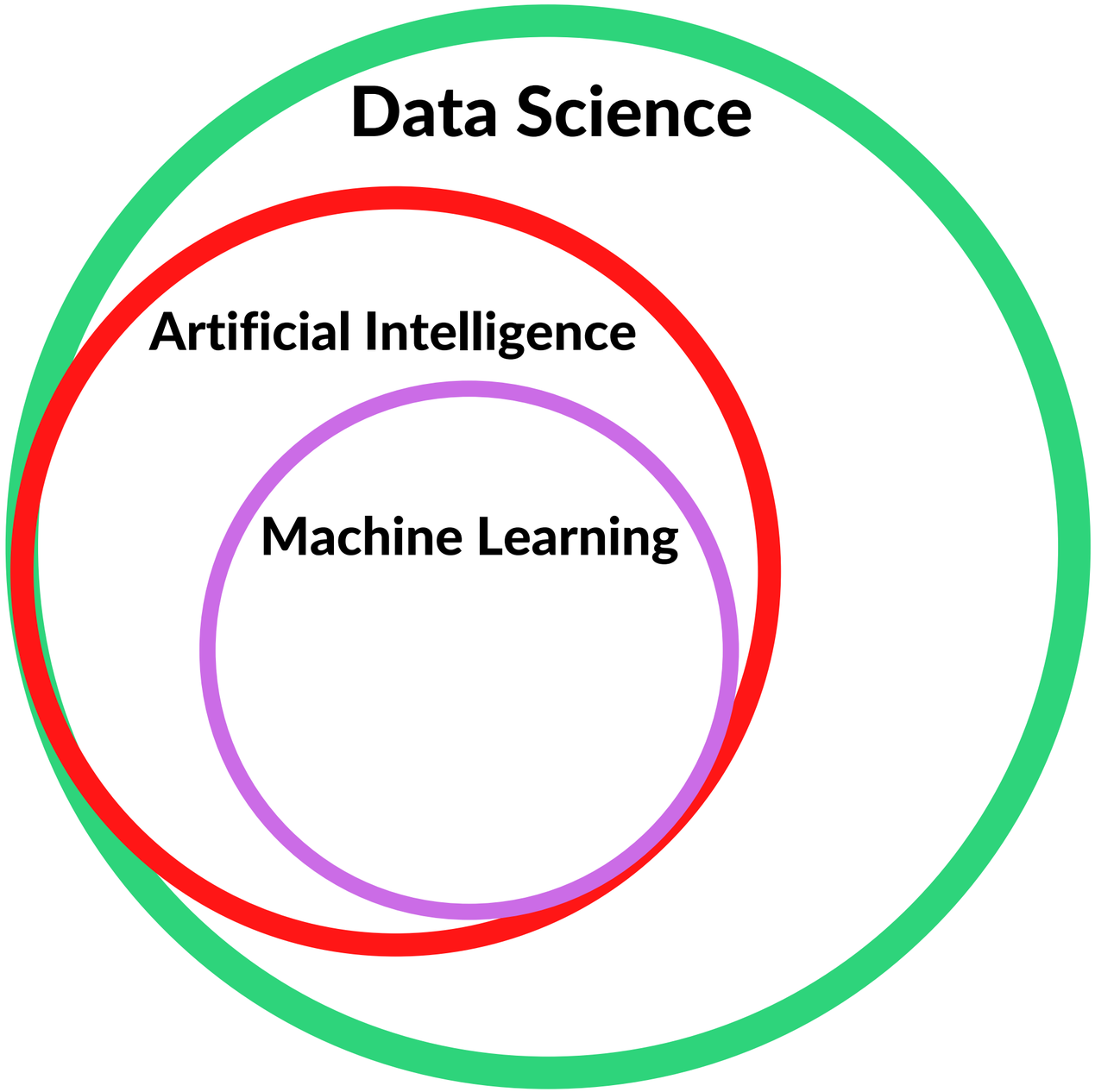
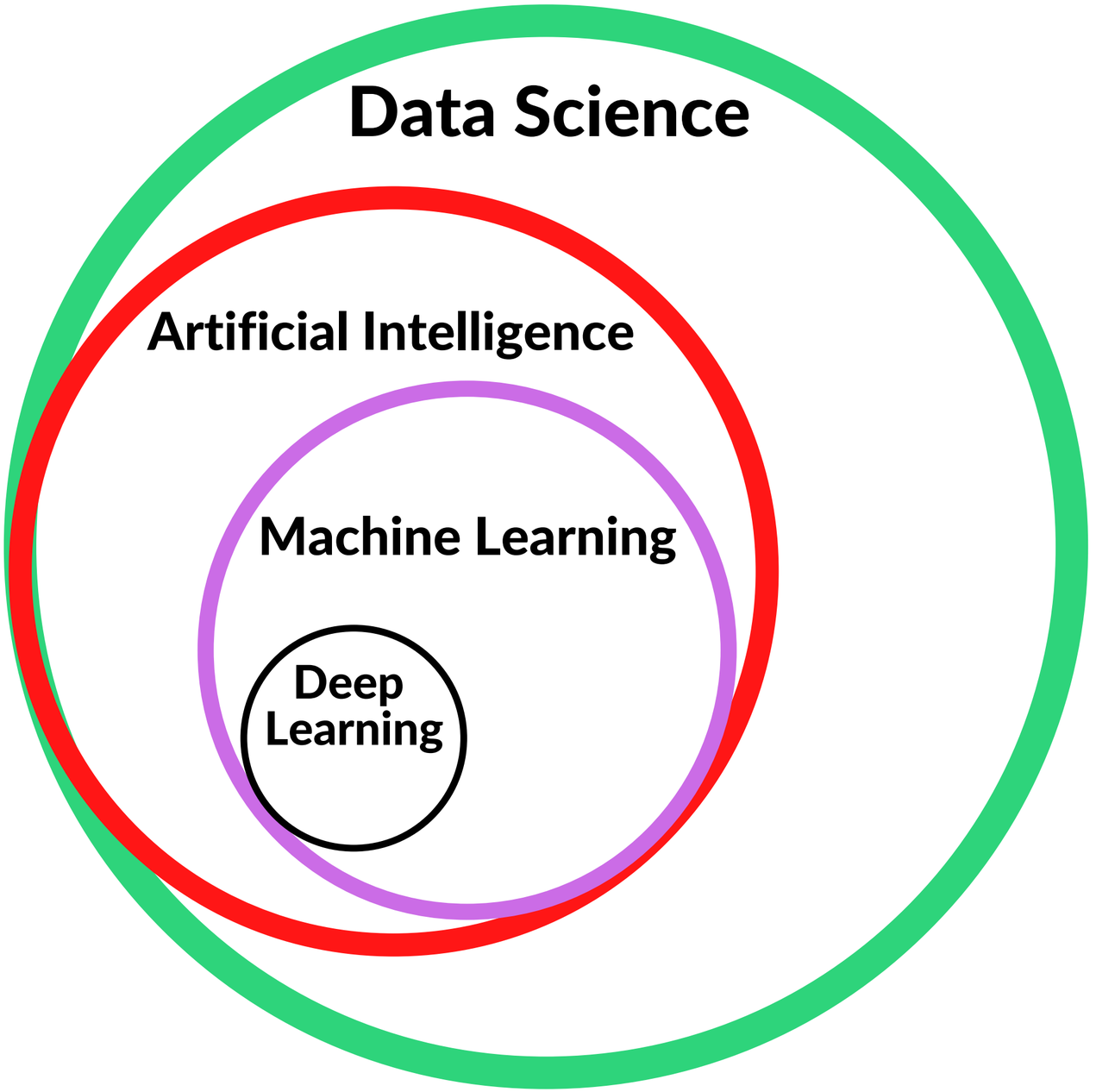
Data Science


Data Science
Artificial Intelligence

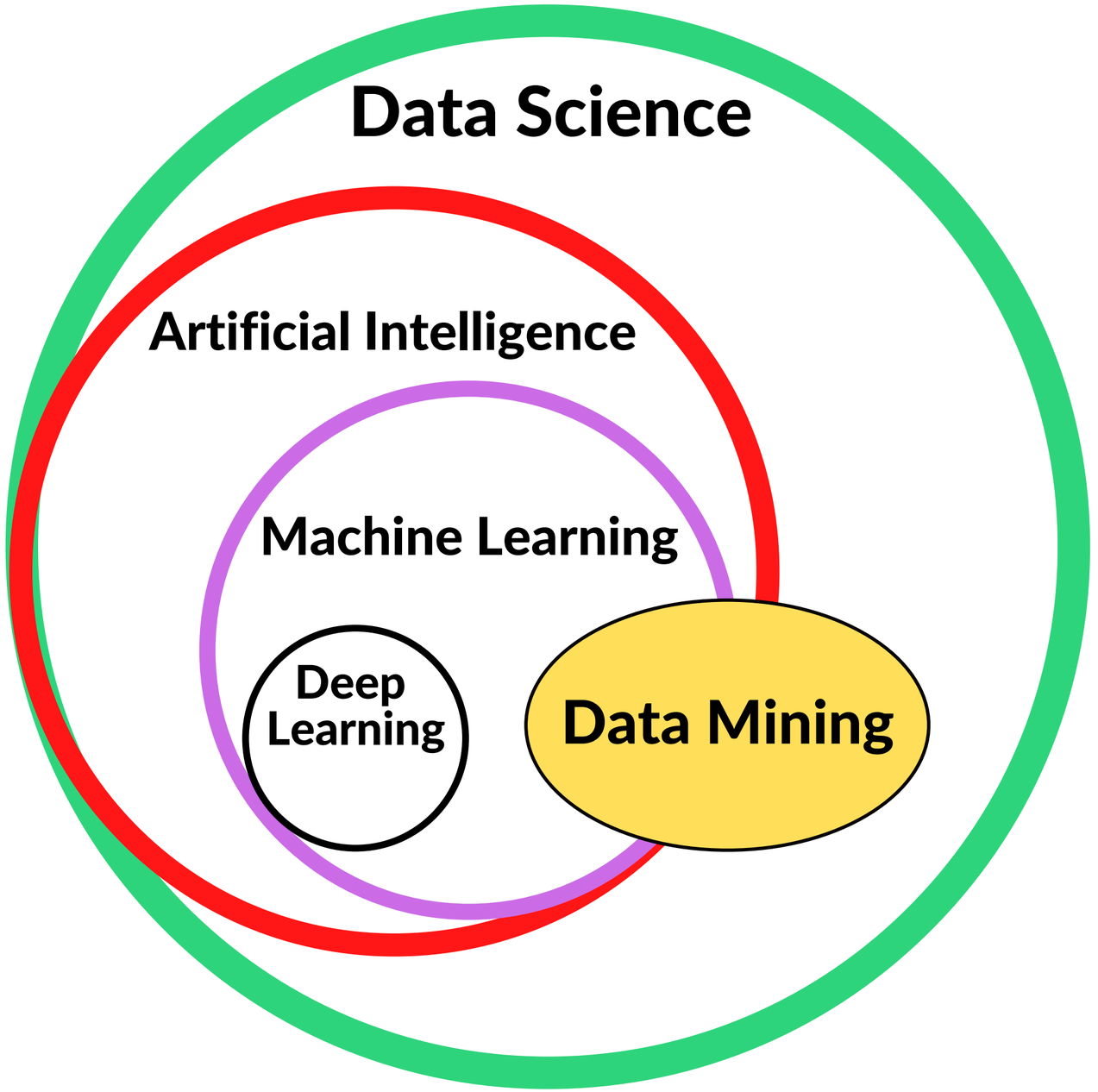
INTRODUCTION TO DATA SCIENCE AND MACHINE LEARNING



INTRODUCTION TO DATA SCIENCE AND MACHINE LEARNING



INTRODUCTION TO DATA SCIENCE AND MACHINE LEARNING



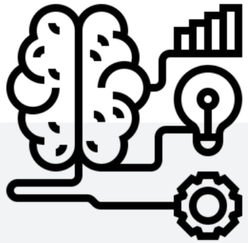
Summary of DS, AI, ML, and DM

DATA SCIENCE

study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

ARTIFICIAL INTELLLIGENCE

ability to learn, understand, imagine the qualities that are naturally found in Humans. Developing a system that has the same or better level of these qualities artificially is termed ad Artificial Intelligence.



MACHINE LEARNING

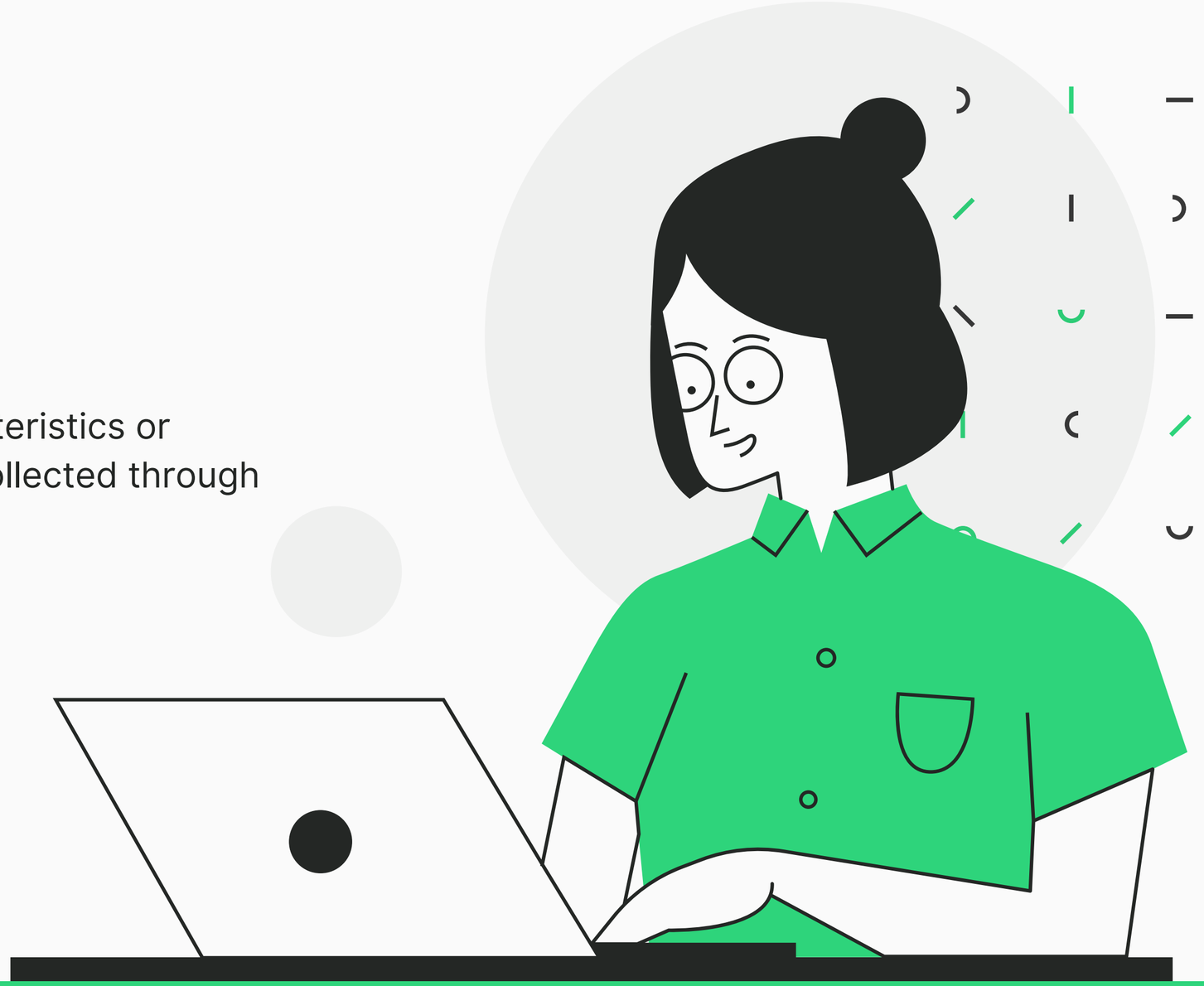
study of computer algorithms that comprises algorithms and statistical models that allow computer programs to automatically improve through experience.

DATA MINING

process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Data

According to Wikipedia, "Data are characteristics or information, usually numerical, that are collected through observation."



Data

A	B	C	D	E	F
Date/Time	**Building Code**	**Power Consumption (KW)**	**Heat Consumption(KW)**	**Power Price($/KW)**	**Head Price ($/KW)**
01/01/21 12:00 PM	6601	450	550	10	4
01/02/21 01:00 PM	6602	480	590	12	5
01/03/21 02:00 PM	6603	500	540	11	7
01/04/21 03:00 PM	6604	550	596	12	3
01/05/21 04:00 PM	6605	670	523	10	4
01/06/21 05:00 PM	6606	-50	488	9	6
01/07/21 06:00 PM	6607	430	610	4	6

Data

Date/Time	Building Code	Power Consumption (KW)	Heat Consumption(KW)	Power Price($/KW)	Head Price ($/KW)
01/01/21 12:00 PM	6601	450	550	10	4
01/02/21 01:00 PM	6602	480	590	12	5
01/03/21 02:00 PM	6603	500	540	11	7
01/04/21 03:00 PM	6604	550	596	12	3
01/05/21 04:00 PM	6605	670	523	10	4
01/06/21 05:00 PM	6606	-50	488	9	6
01/07/21 06:00 PM	6607	430	610	4	6

Row/ Example/
Sample

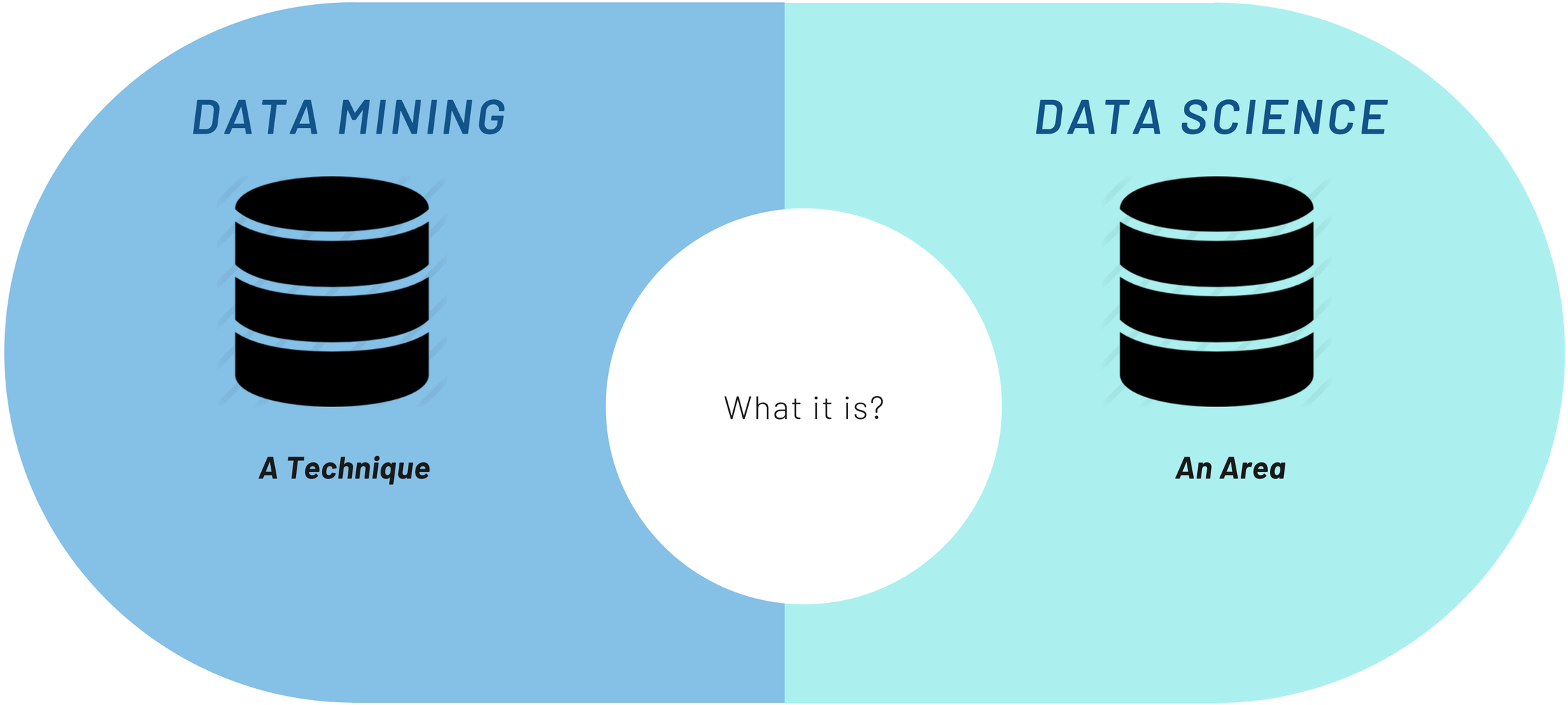
Data

Date/Time	Building Code	Power Consumption (KW)	Heat Consumption(KW)	Power Price($/KW)	Head Price ($/KW)
01/01/21 12:00 PM	6601	450	550	10	4
01/02/21 01:00 PM	6602	480	590	12	5
01/03/21 02:00 PM	6603	500	540	11	7
01/04/21 03:00 PM	6604	550	596	12	3
01/05/21 04:00 PM	6605	670	523	10	4
01/06/21 05:00 PM	6606	-50	488	9	6
01/07/21 06:00 PM	6607	430	610	4	6

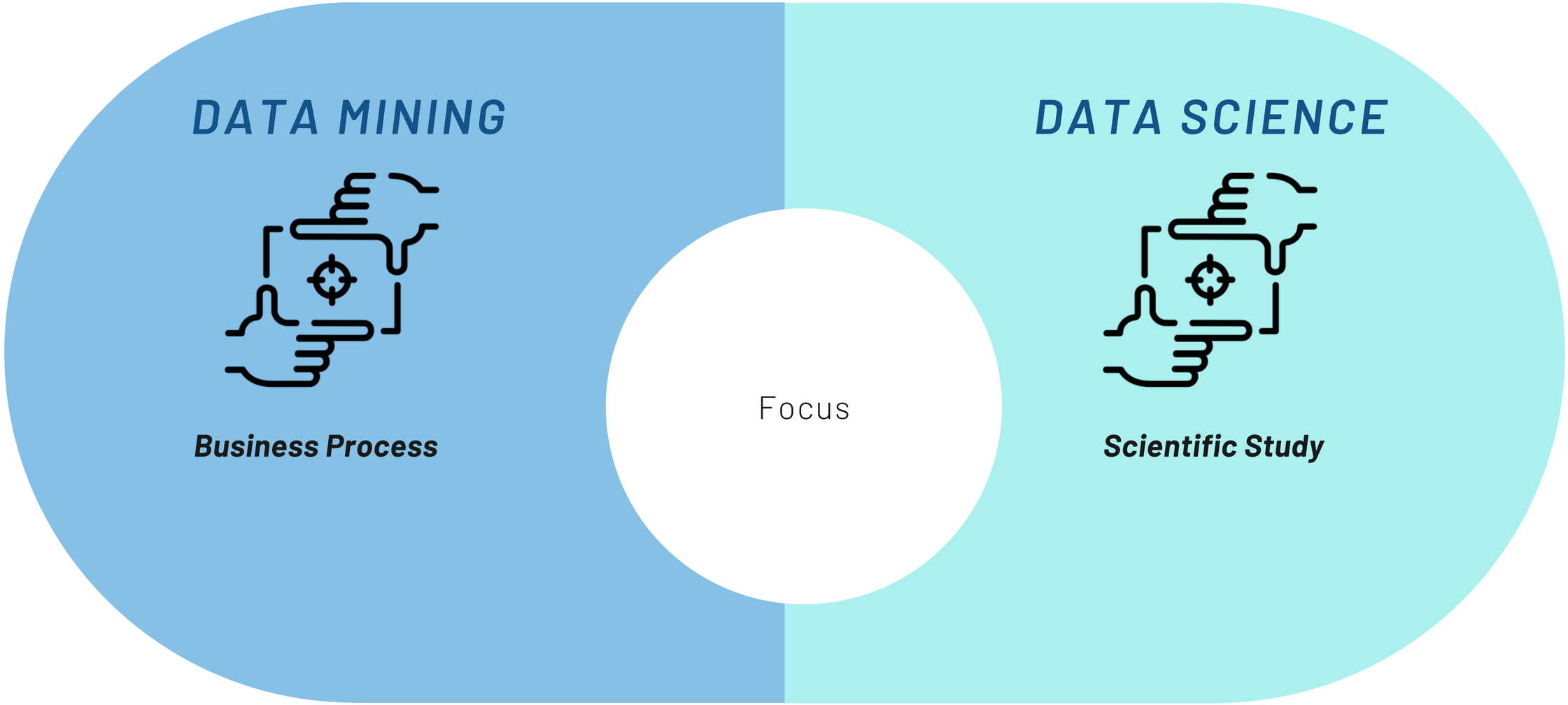
Row/ Example/ Sample

Variable/ Attribute/ Feature

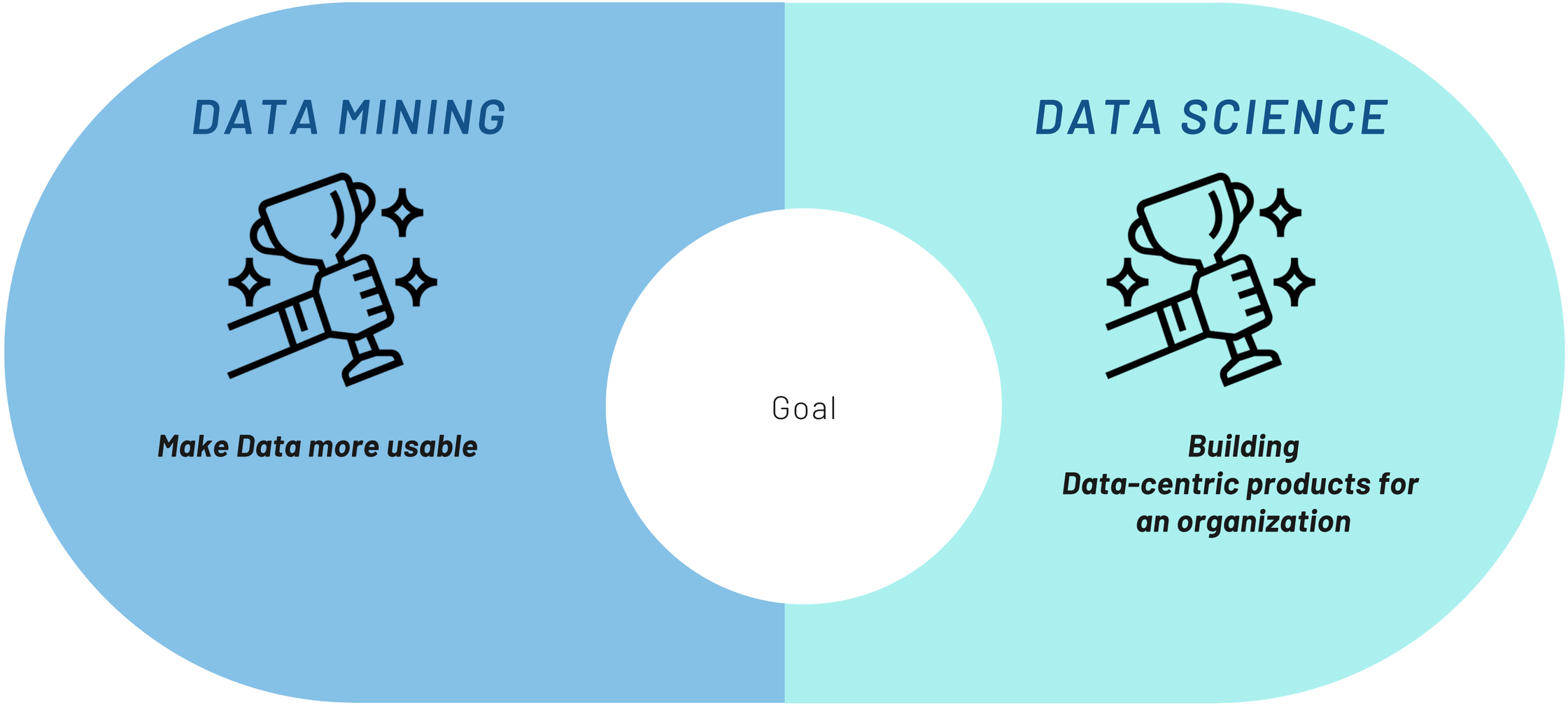
DATA SCIENCE VS DATA MINING (01)



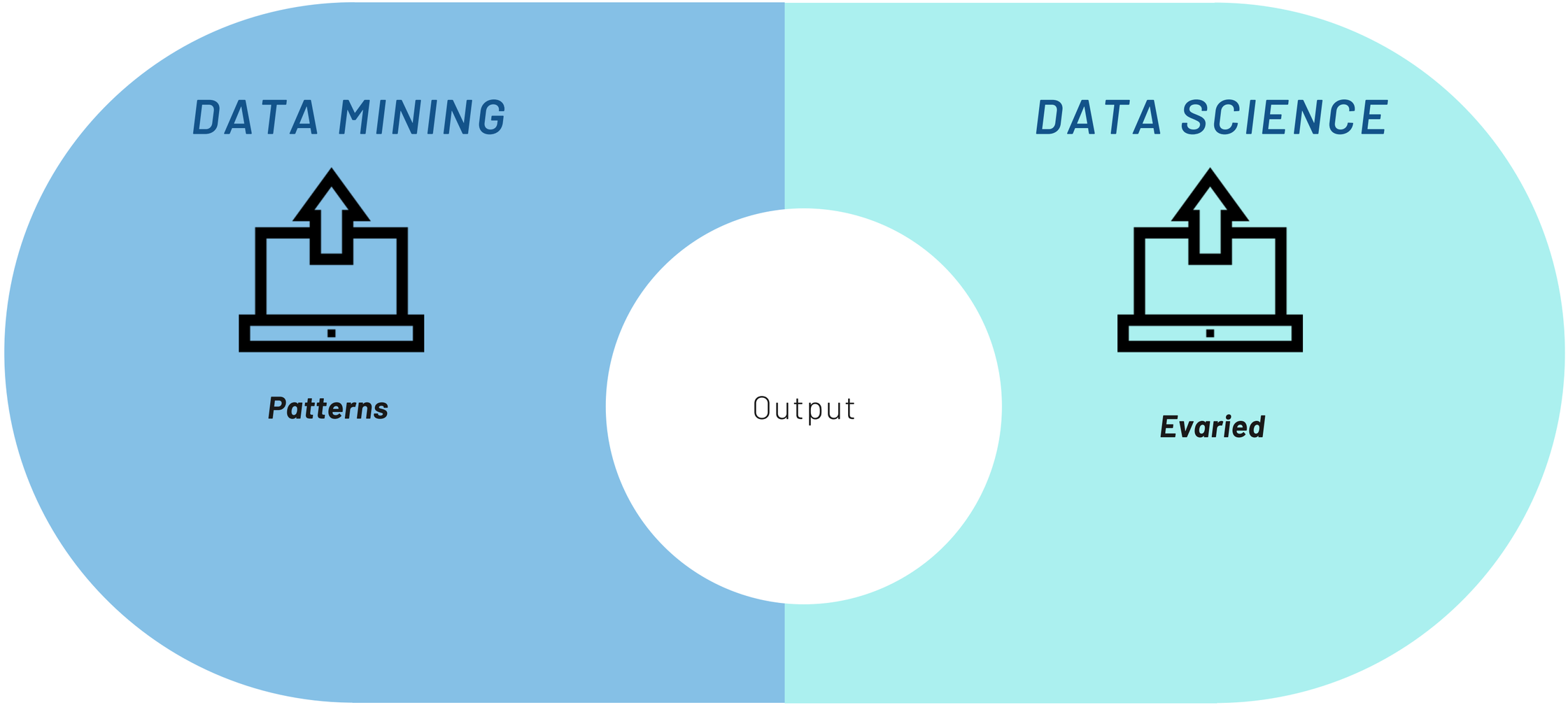
DATA SCIENCE VS DATA MINING (02)



DATA SCIENCE VS DATA MINING (03)



DATA SCIENCE VS DATA MINING (04)



DATA SCIENCE VS DATA MINING (05)

DATA MINING

Finding trends previously not known

Purpose

DATA SCIENCE

Social analysis, building predictive models, unearthing, unknown facts, and more

DATA SCIENCE VS DATA MINING (06)

DATA MINING

Someone with a knowledge of navigating across data and statistical understanding can conduct data mining

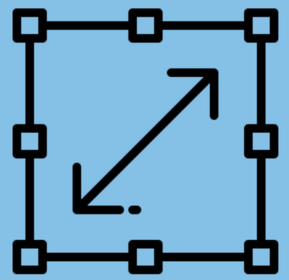
Vocational Perspective

DATA SCIENCE

A person needs to understand ML, Programming, info-graphic techniques and have domain knowledge to become a data scientist.

DATA SCIENCE VS DATA MINING (07)

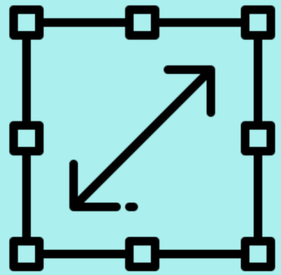
DATA MINING



Data mining can be a subset of Data Science as Mining activities are part of Data Science pipeline.

Extent

DATA SCIENCE



Multidisciplinary - Data Science consists of Data Visualizations, computational Social Sciences, Statistics, Data Mining, Natural Language Processing

DATA SCIENCE VS DATA MINING (08)

DATA MINING

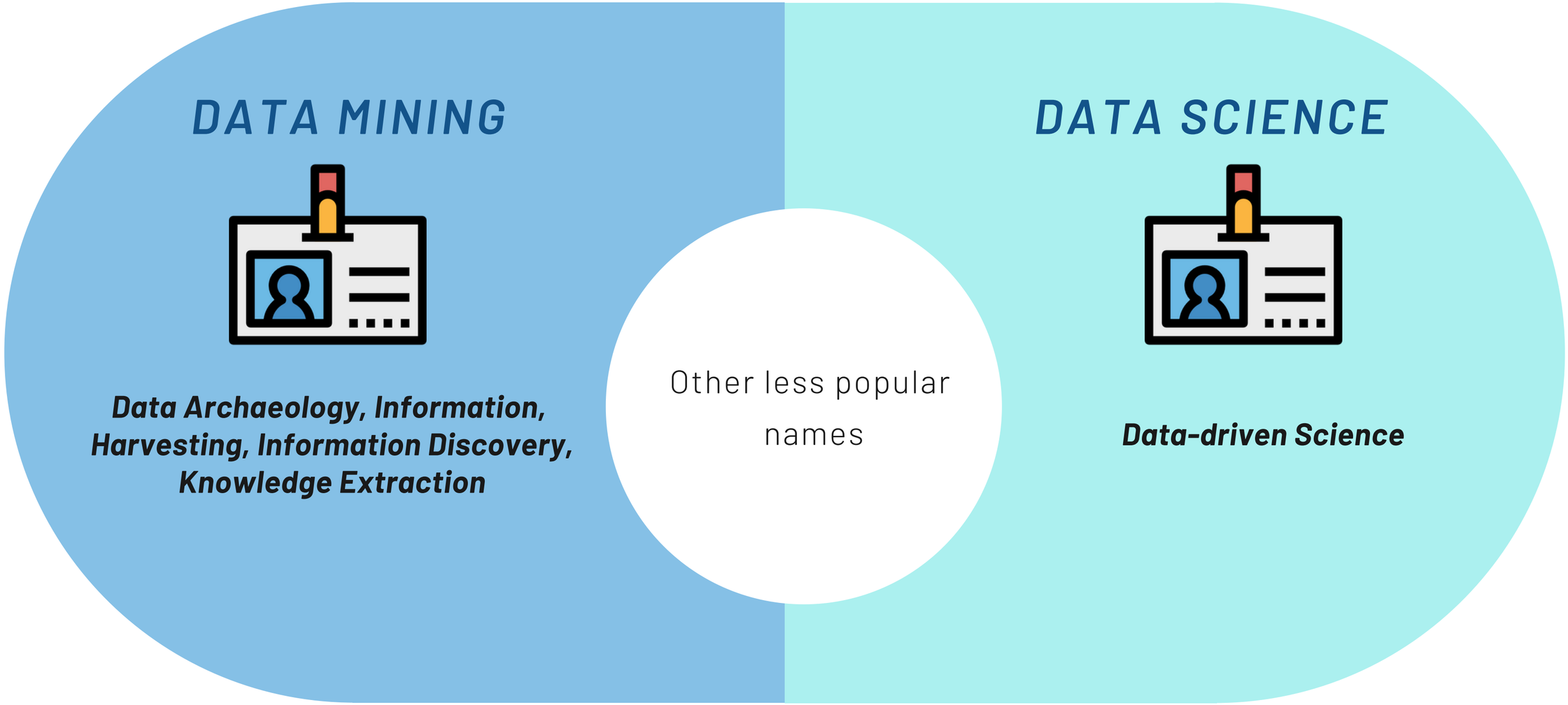
Mostly Structured

Deal with (type of data)

DATA SCIENCE

All forms of data - structured, semi-structured, and unstructured

DATA SCIENCE VS DATA MINING (09)



DATA MINING EXAMPLE USE CASE

Consider a scenario where you are a major retailer in Nepal. You have 120 stores operating in 10 major cities in Nepal and you have been operational for 10 years.

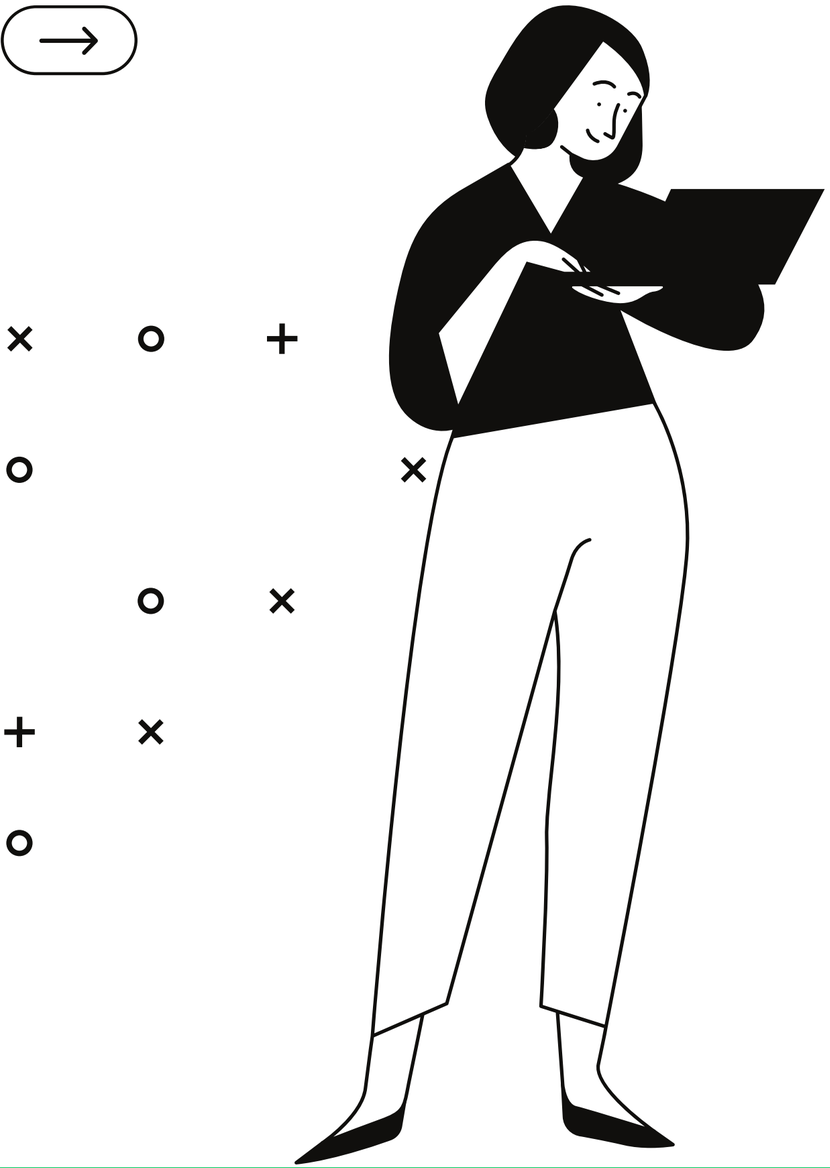
Let's say, you want to study the last 8 years' data to find the number of sales of sweets during festive seasons of 3 cities (Kathmandu, Bhaktapur, and Lalitpur). If that's your objective,

I would recommend you employ a person with Data Mining expertise. A Data Miner would probably go through historical information stored in legacy systems and employ algorithms to extract trends.

DATA SCIENCE EXAMPLE USE CASE

If you want to know which sweets have received more positive reviews. In this case, your sources of data may not be limited to databases, they could extend to social websites or customer feedback messages.

In this case, my suggestion to you would be to employ a Data Scientist. A person employed as a Data Scientist is more suited to apply algorithms and conduct this socio-computational analysis.



What is Machine Learning?

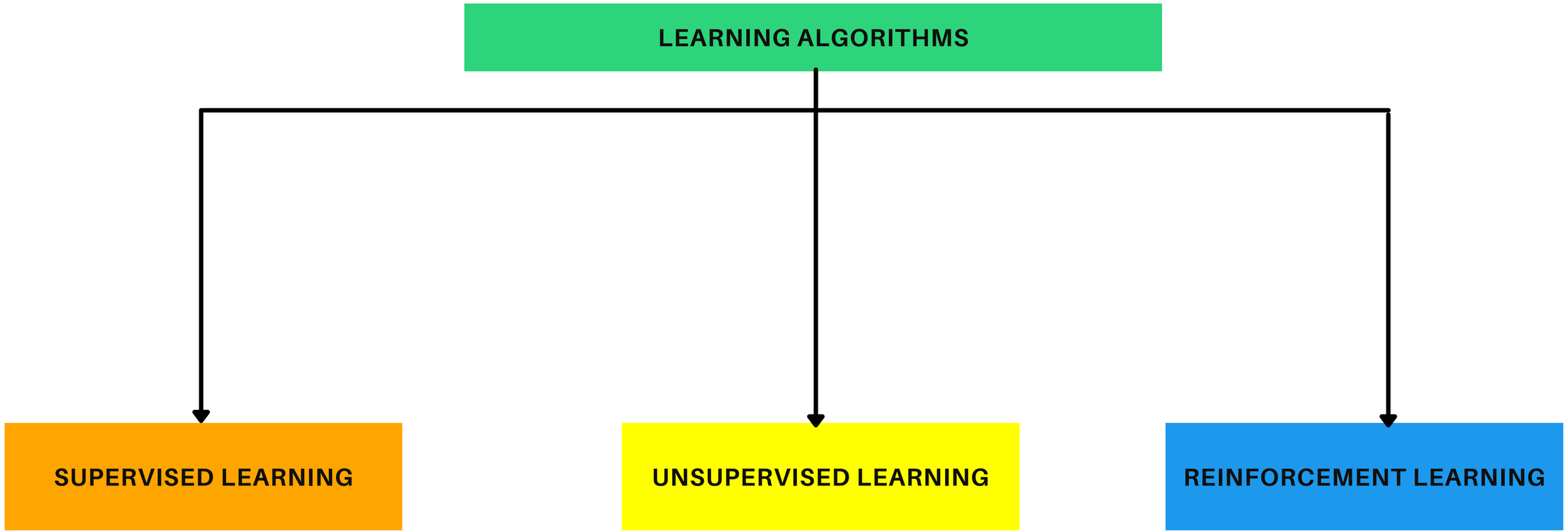
Definition 01

Machine learning is the study of computer algorithms that comprises algorithms and statistical models that allow computer programs to automatically improve through experience

Definition 02

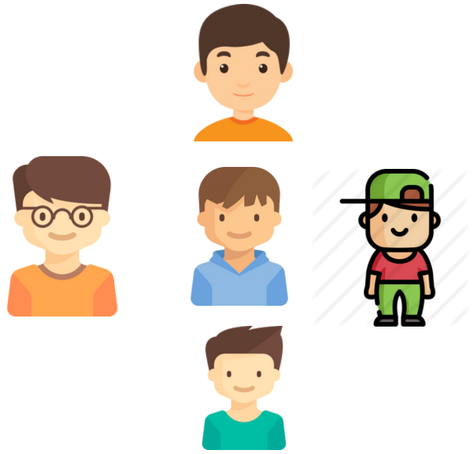
Machine Learning is the science of getting computers to act by feeding them data and letting them learn a few tricks on their own without being explicitly programmed.

MACHINE LEARNING MODELS

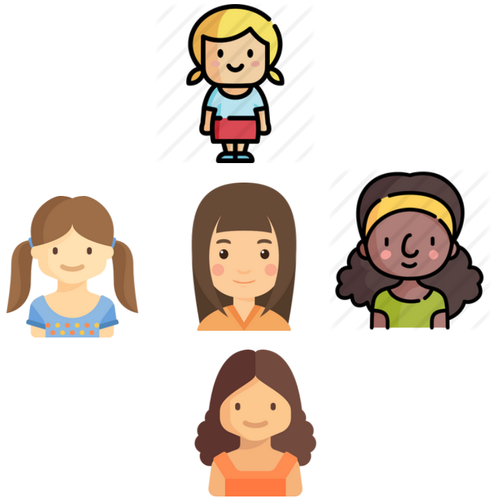


SUPERVISED LEARNING

DATA
Label = Boy and Girl

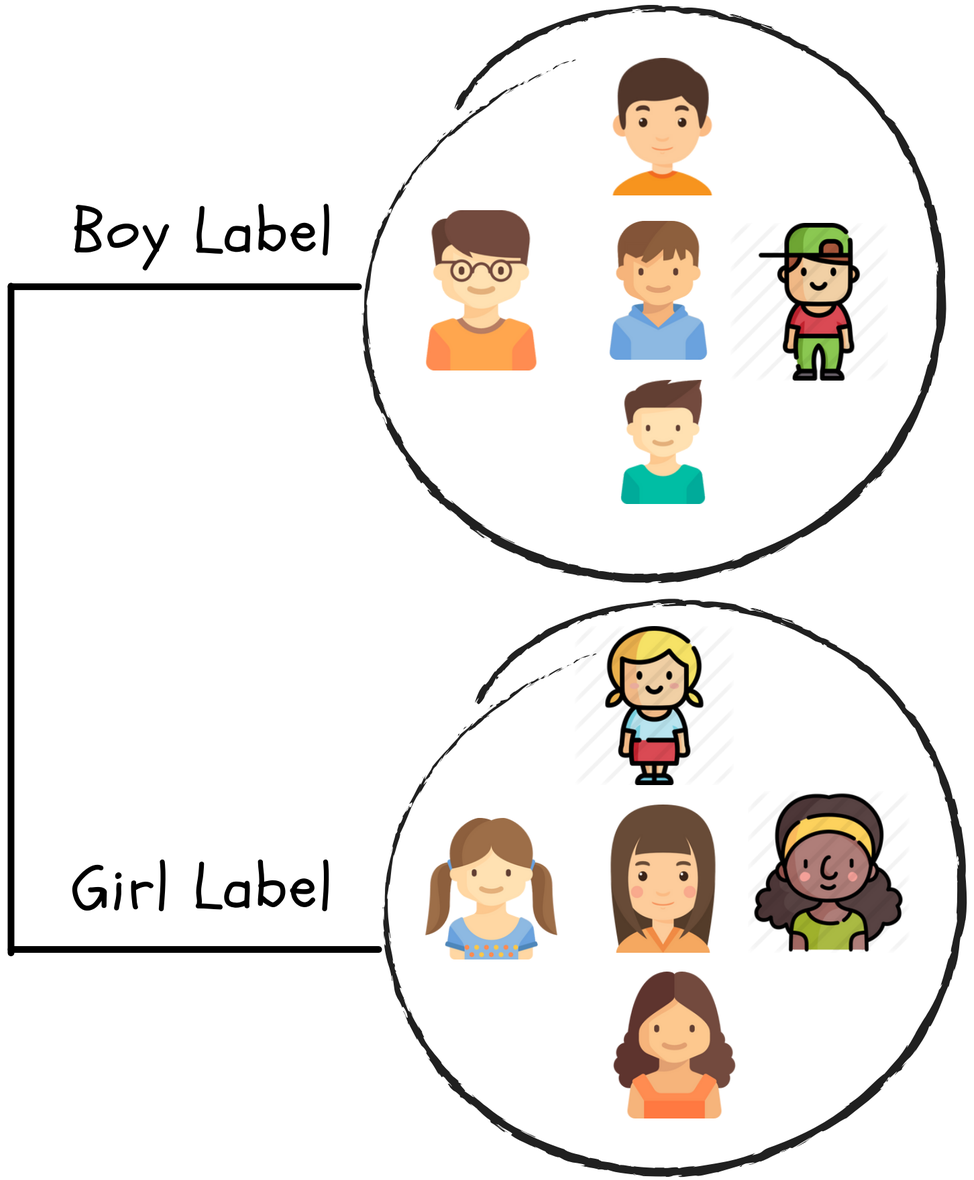


Label = Boy

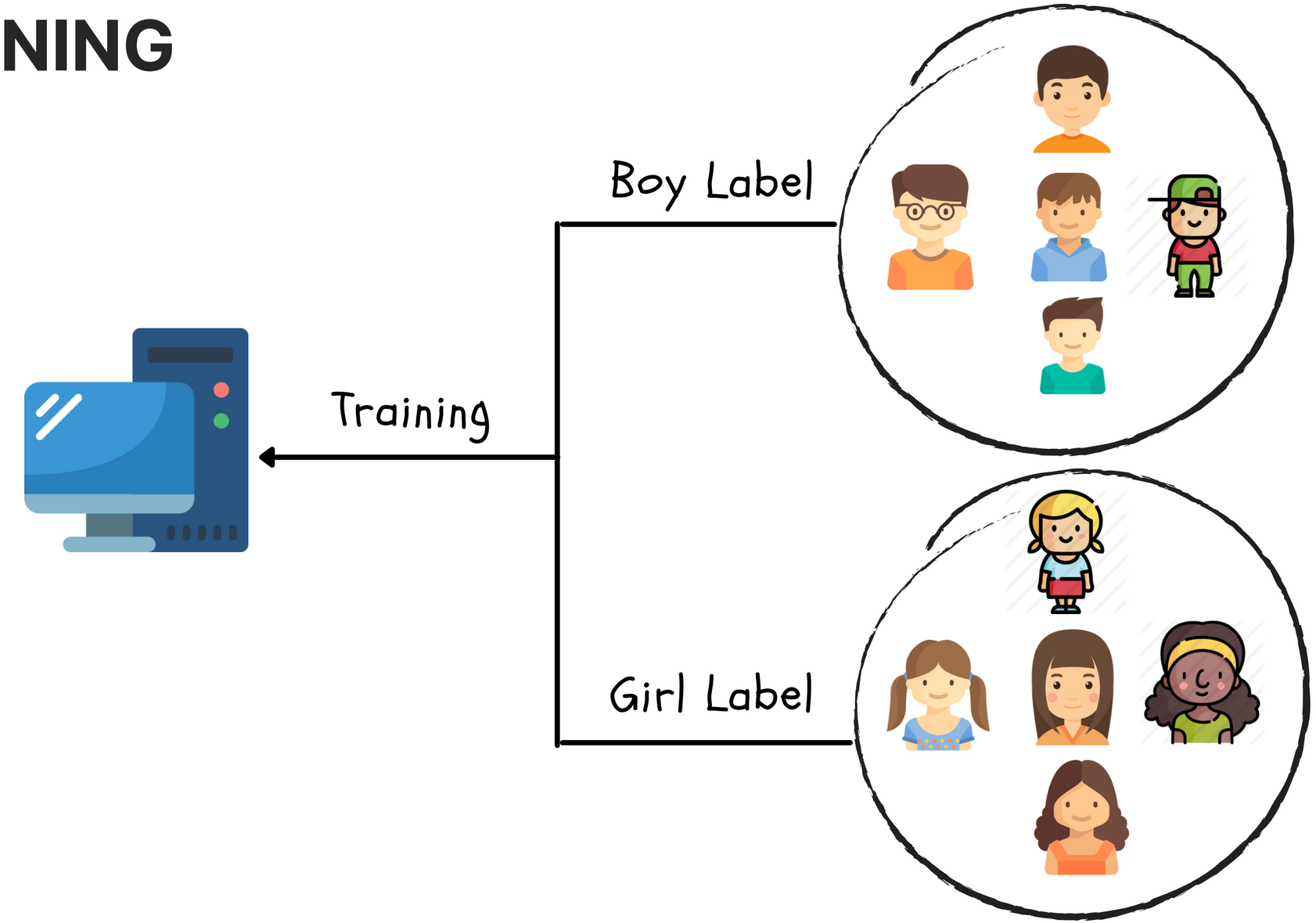


Label = Girl

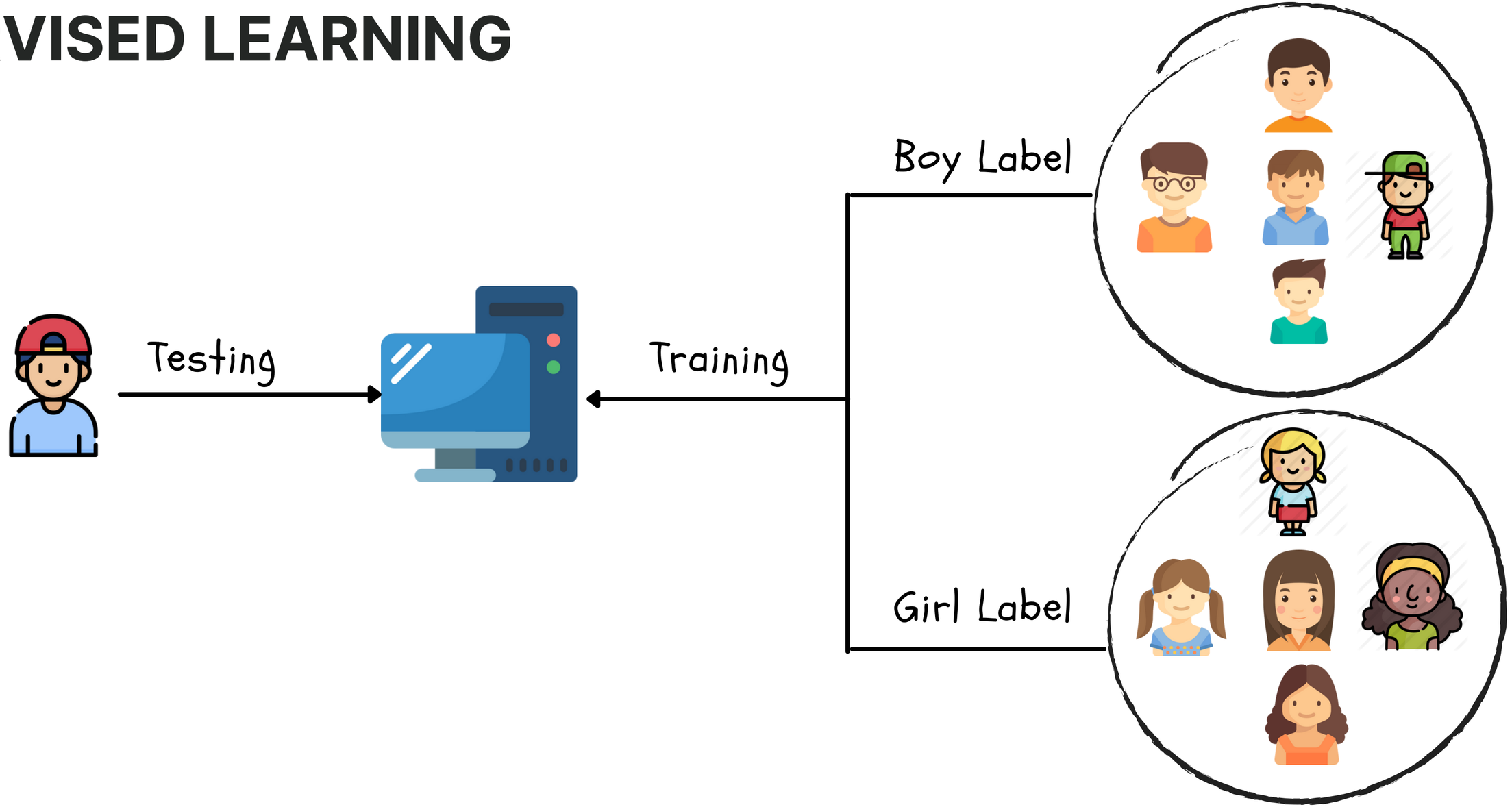
SUPERVISED LEARNING



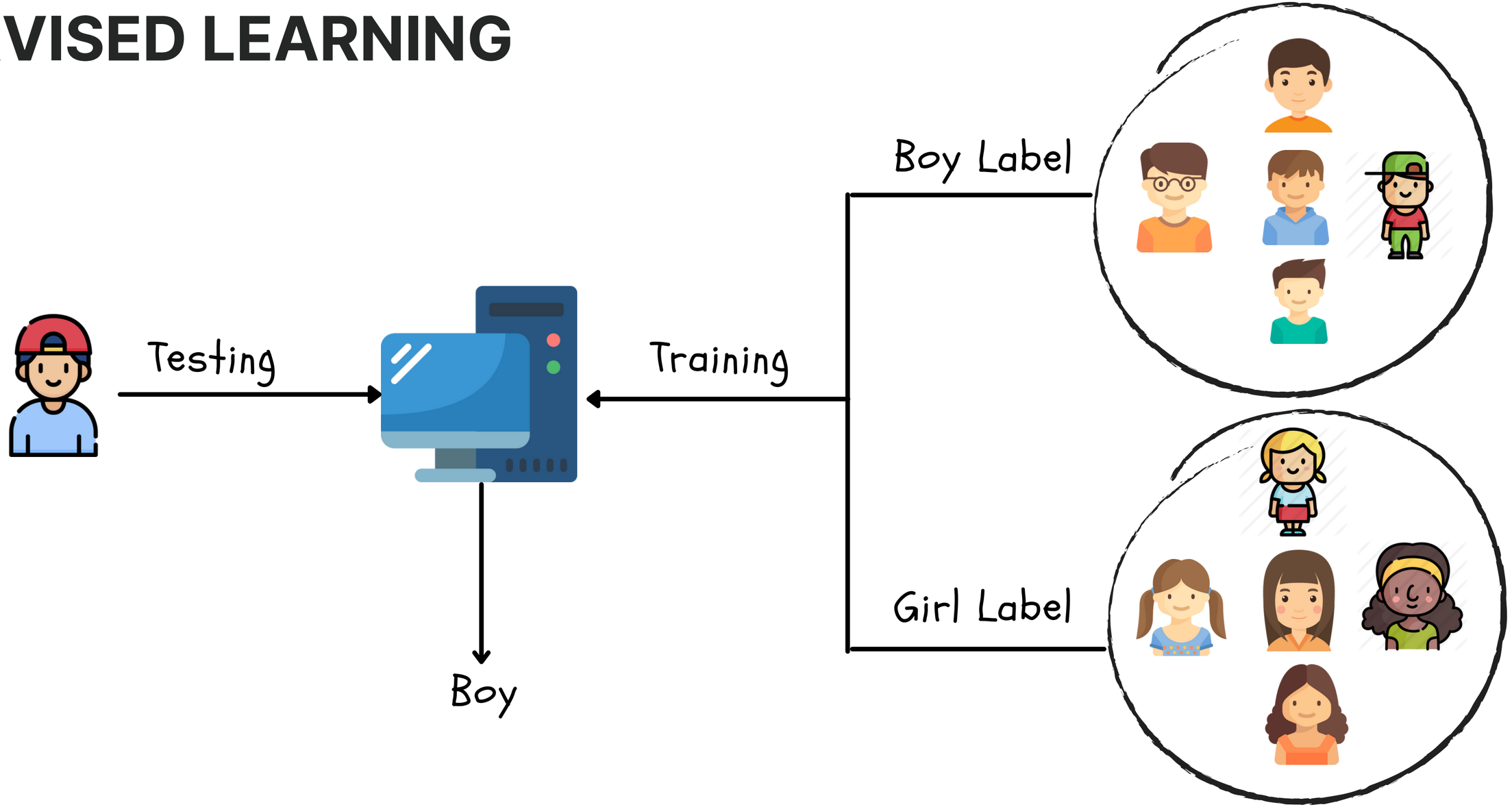
SUPERVISED LEARNING



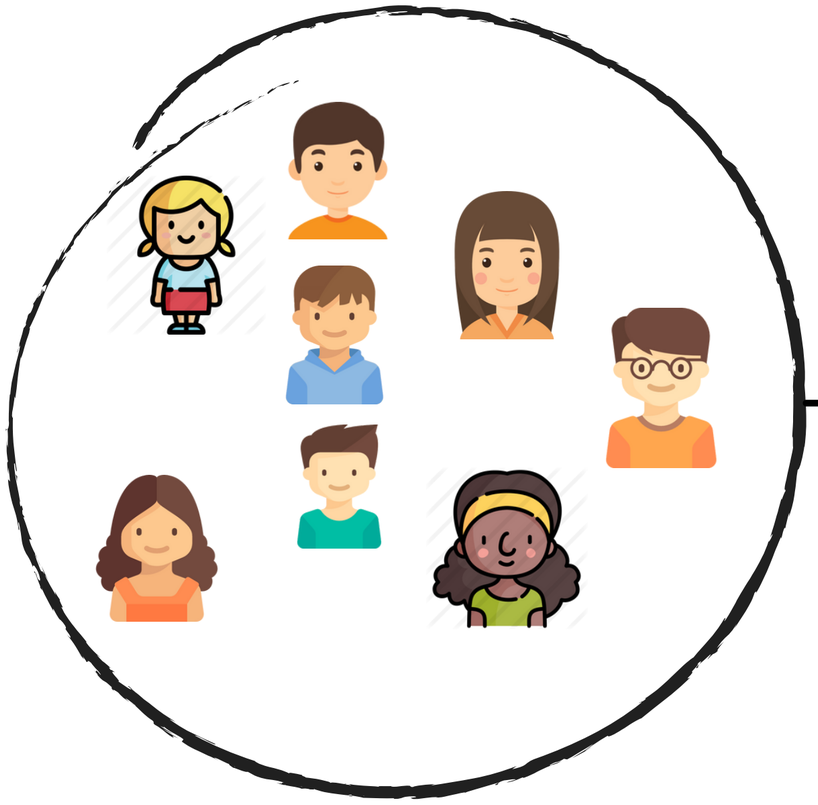
SUPERVISED LEARNING



SUPERVISED LEARNING

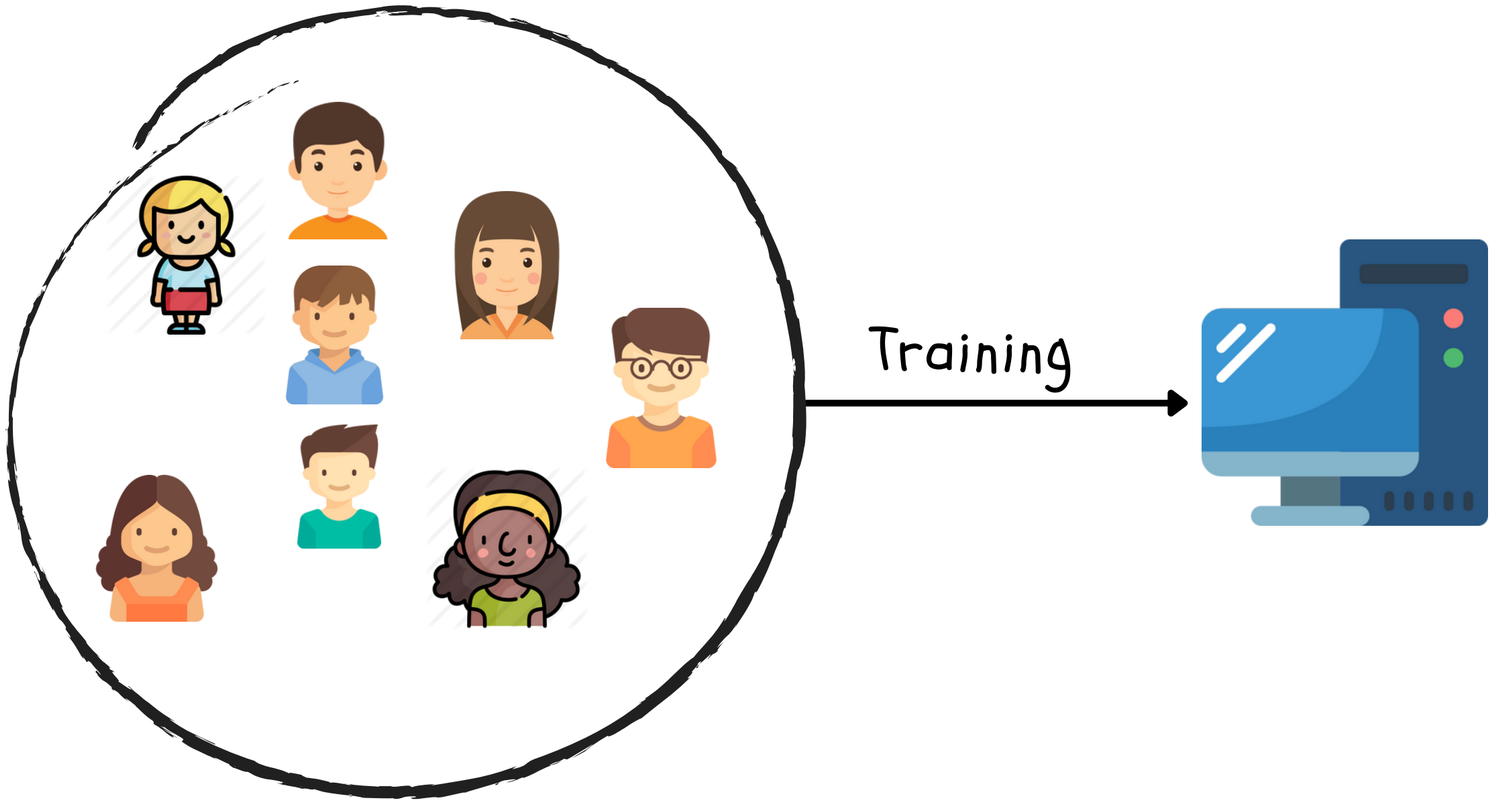


UNSUPERVISED LEARNING

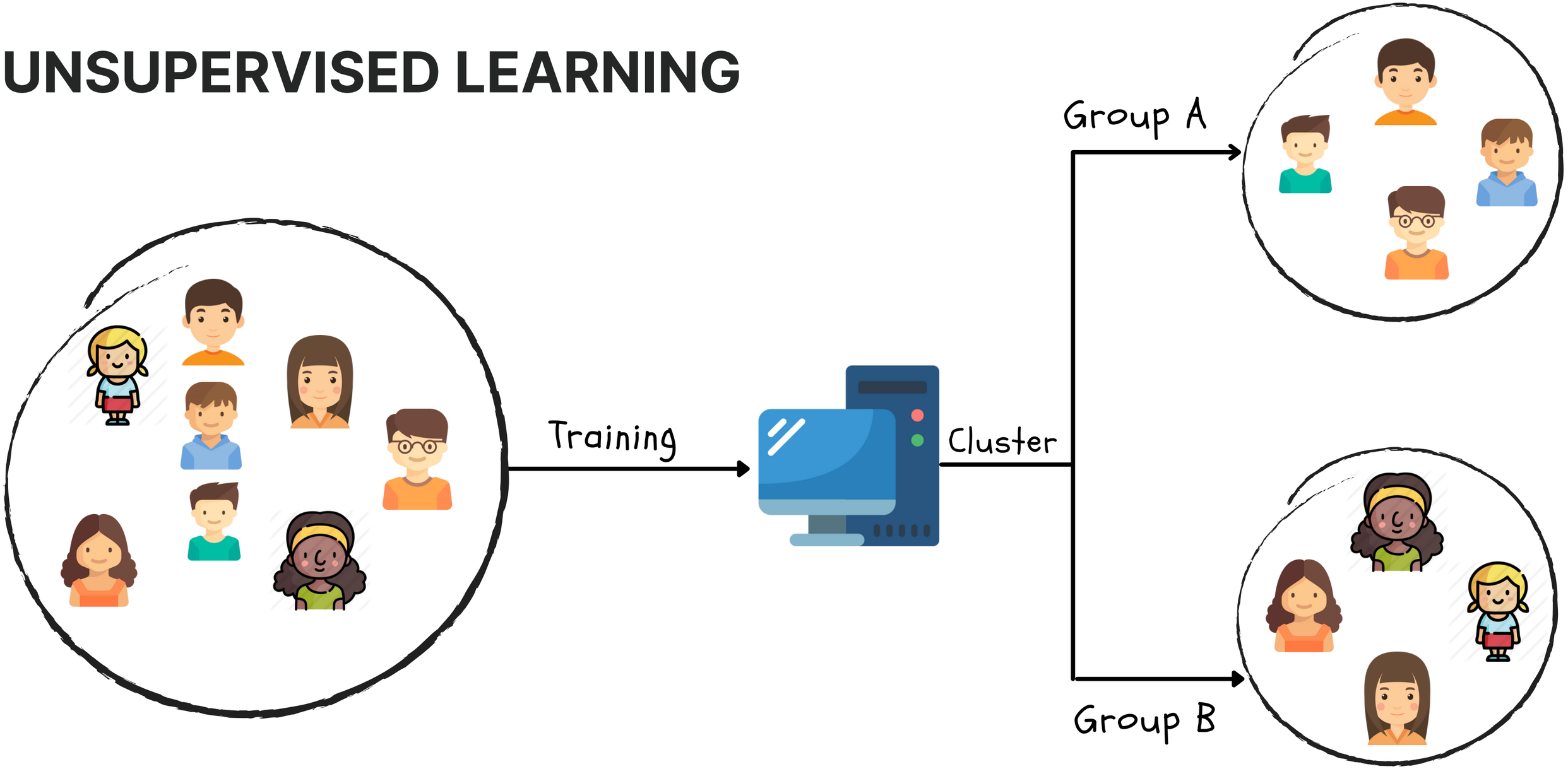


Unlike Supervised Learning, here we don't categorize label the input!

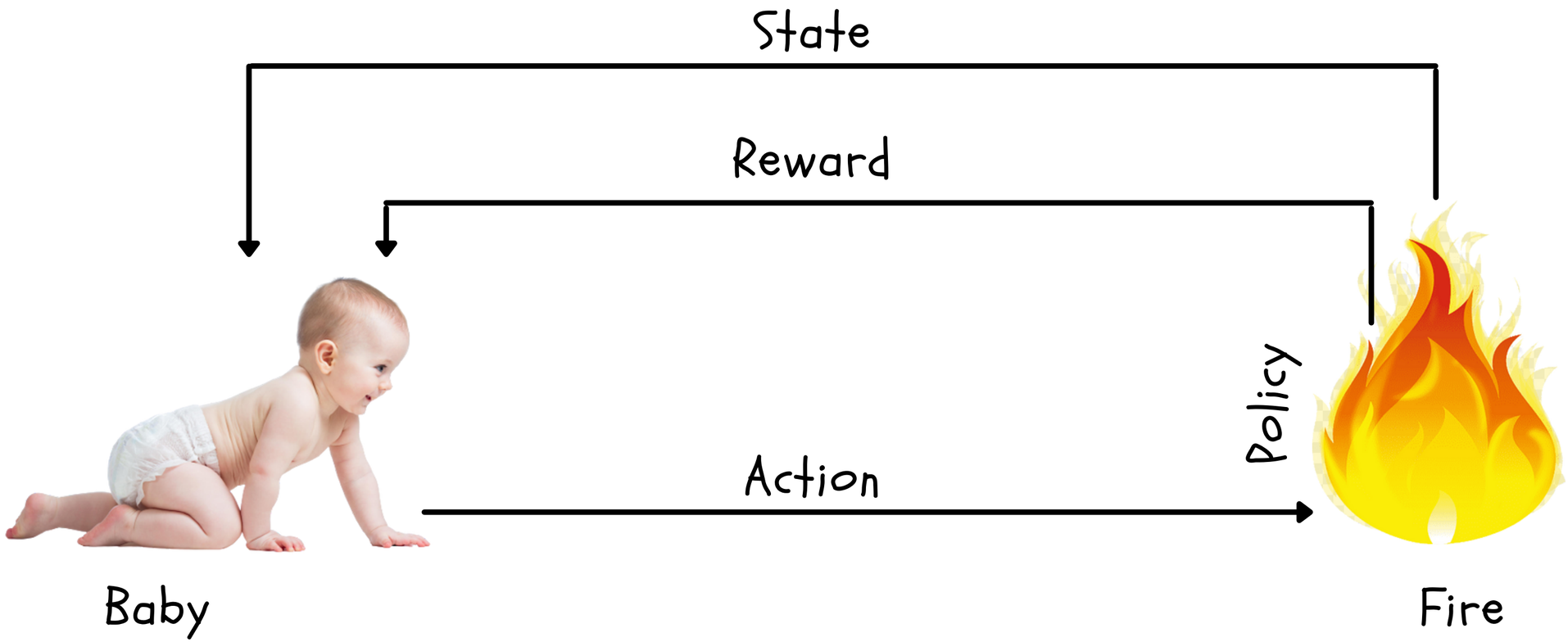
UNSUPERVISED LEARNING



UNSUPERVISED LEARNING



REINFORCEMENT LEARNING



Key terms used in Machine Learning

CLASSIFICATION

REGRESSION

SUPERVISED LEARNING

CLASSIFICATION (DISCRETE DATA)

REGRESSION (CONTINUOUS DATA)

BINARY CLASS

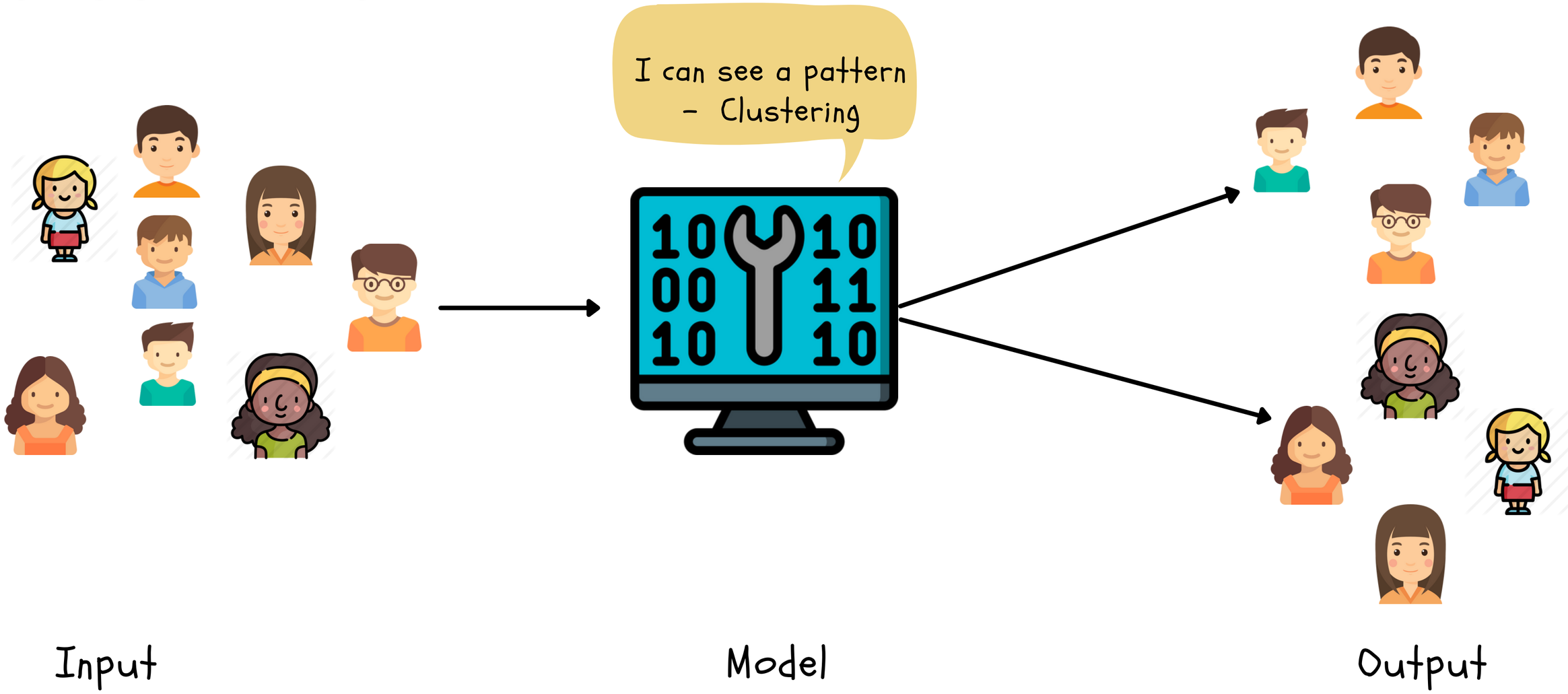
MULTI-CLASS

CORONA - NOT CORONA

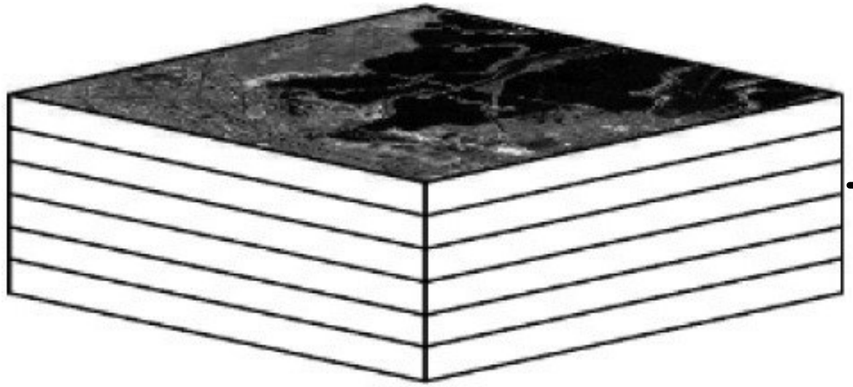
DIAGNOSIS BASED ON SYMPTOMS

HOUSE PRICE BASED ON DIFFERENT FEATURE OF IT

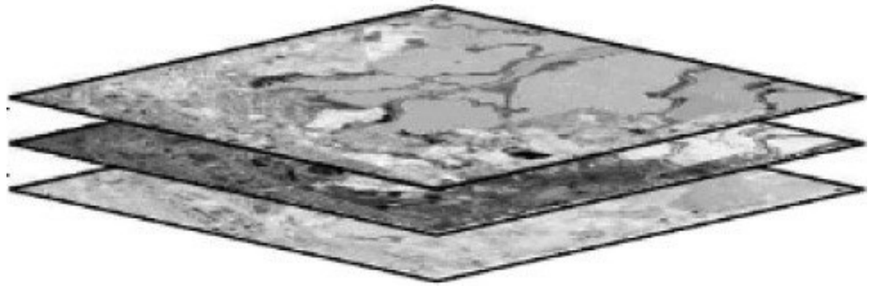
CLUSTERING


I can see a pattern
- Clustering
10 10
00 11
10 10
Input
Model
Output

DIMENSIONALITY REDUCTION AND CURSE OF DIMENSIONALITY



Dimensionality Reduction



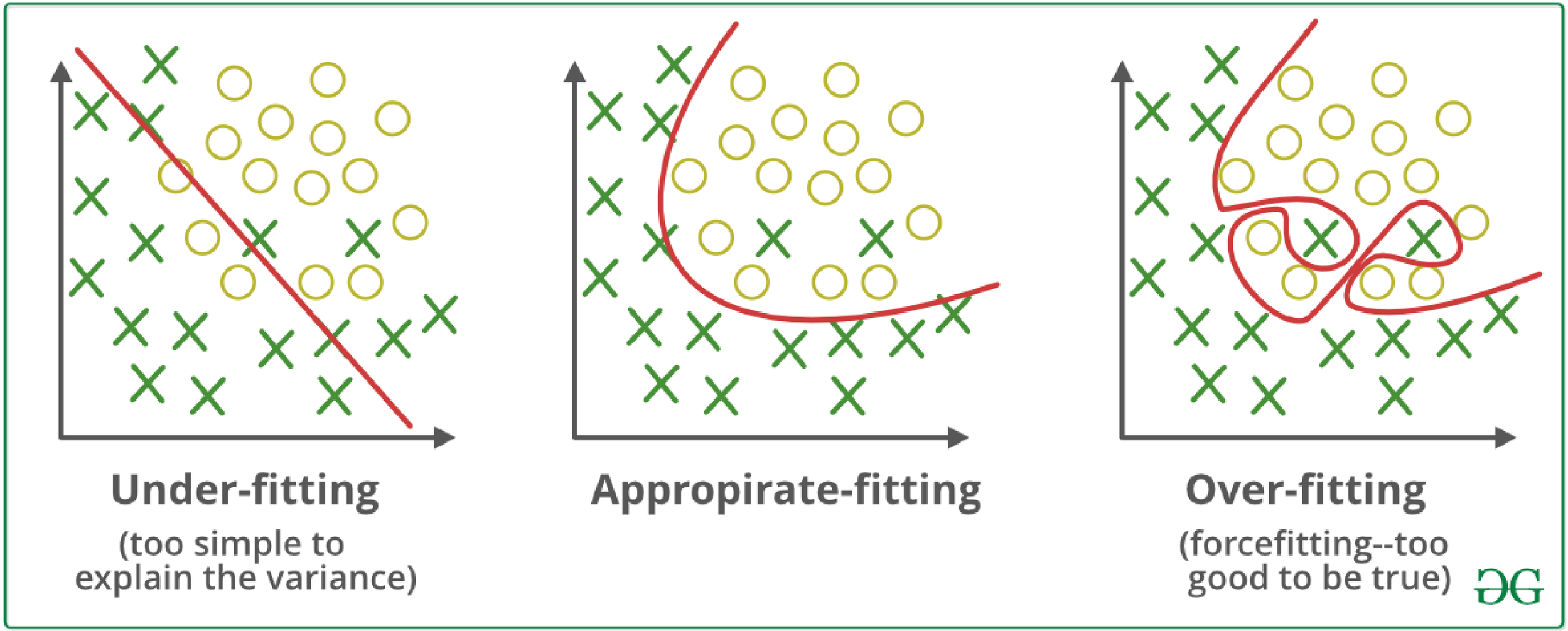
Dimensionality

- large number of attributes or features
- making model slow
- difficult to find solution
- Curse of dimensionality

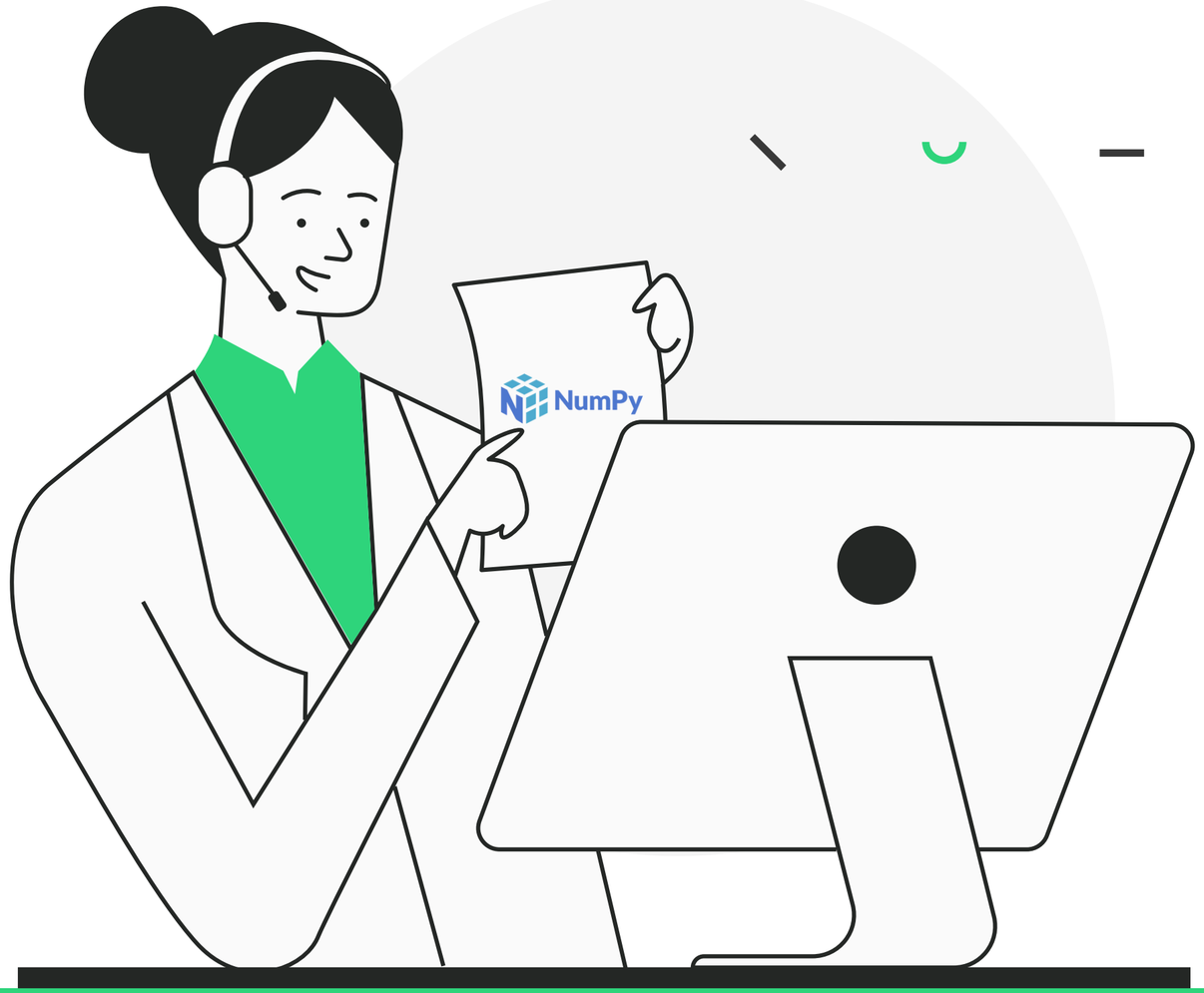
Curse of Dimensionality

- For classification, there will not enough data objects
- For clustering, the density and distance between the points are critical

OVERFITTING AND UNDERFITTING



Machine Learning Useful Package



NUMPY

NumPy is an open-source library that is used for working with arrays

Pandas

Pandas (Panel data) is an open-source a library that is designed for working with Dataframes and series of Data

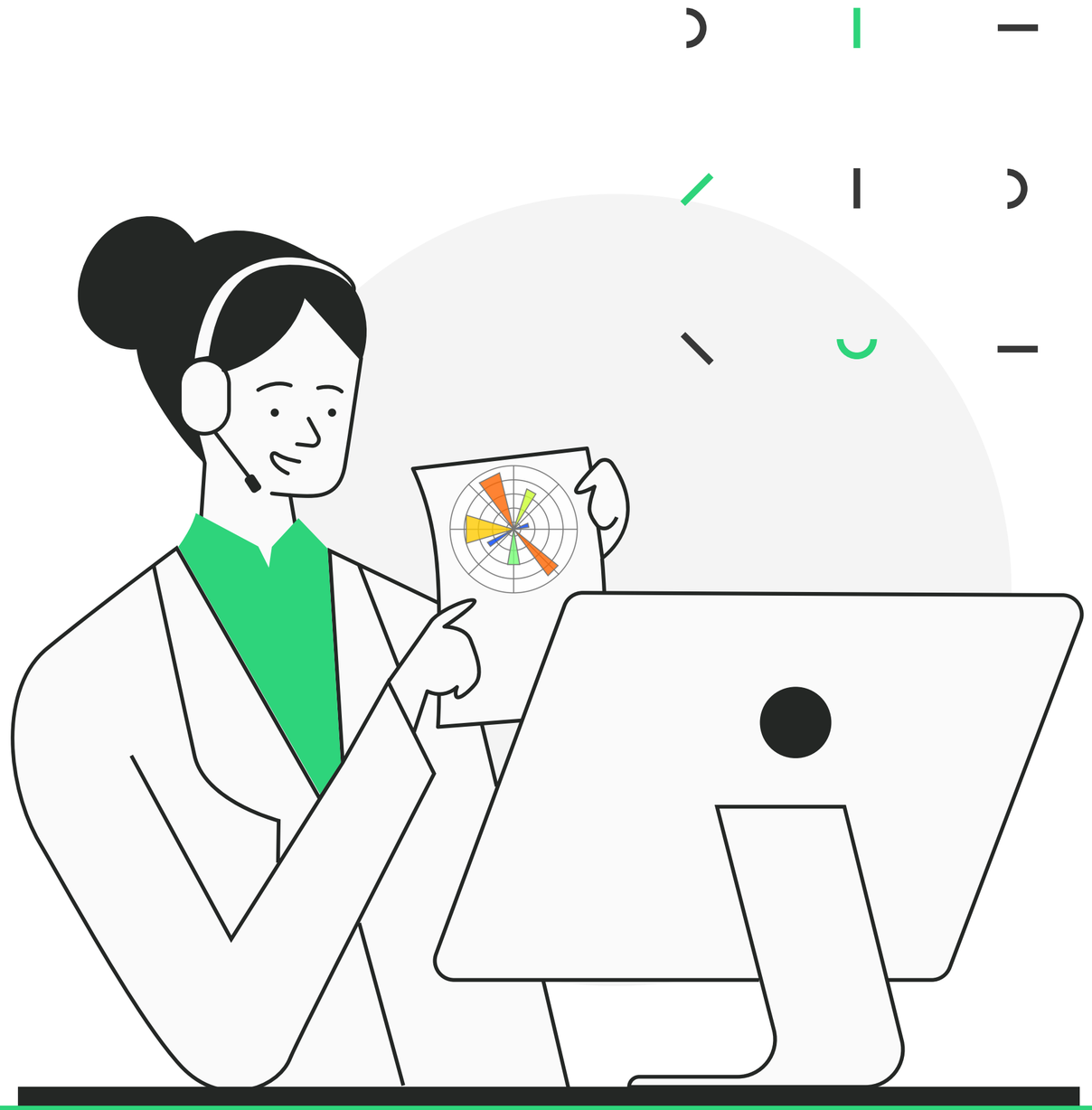
Series are 1D arrays with labeling possibility (Data type could be numerical or string)

Index	Values
age1	10
age2	20
age3	30
age4	40

Several series could create a Dataframe

Pandas Dataframe is 2D labeled data with different type of features

Index	Age	Grade1	Grade2
s1	20	10	8
s2	25	8	10
s3	27	5	3
s4	30	9	7

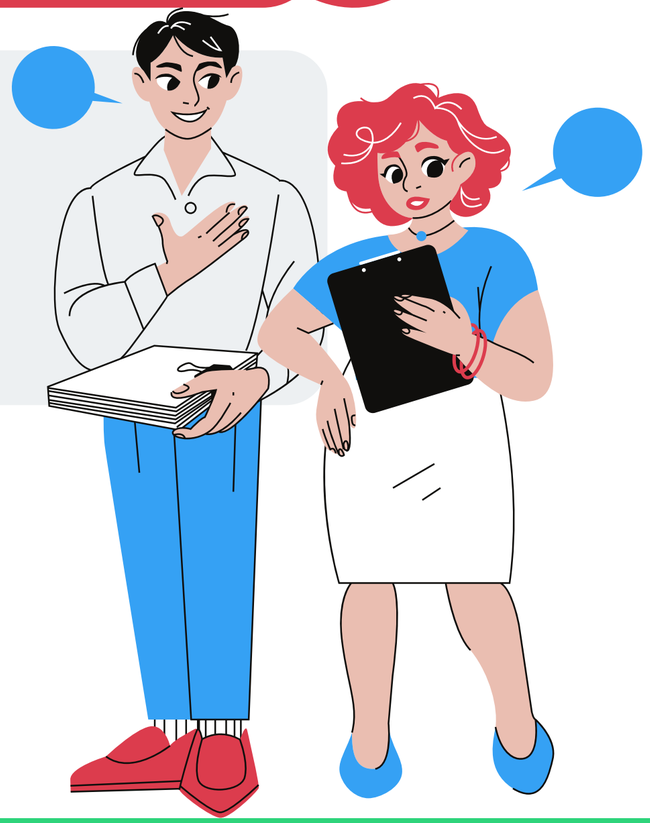


MATPLOTLIB

Matplotlib is an amazing visualization library in Python for 2D plots of arrays

Any Questions?

We are here to support each other



Thank You!