AN INITIATIVE OF:

science, technology & innovation
Department:
Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

CSIR 80th anniversary
Touching lives through innovation

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
DIRISA

Introduction to Machine Learning DIRISA

25 June 2025


01
Life without machine learning

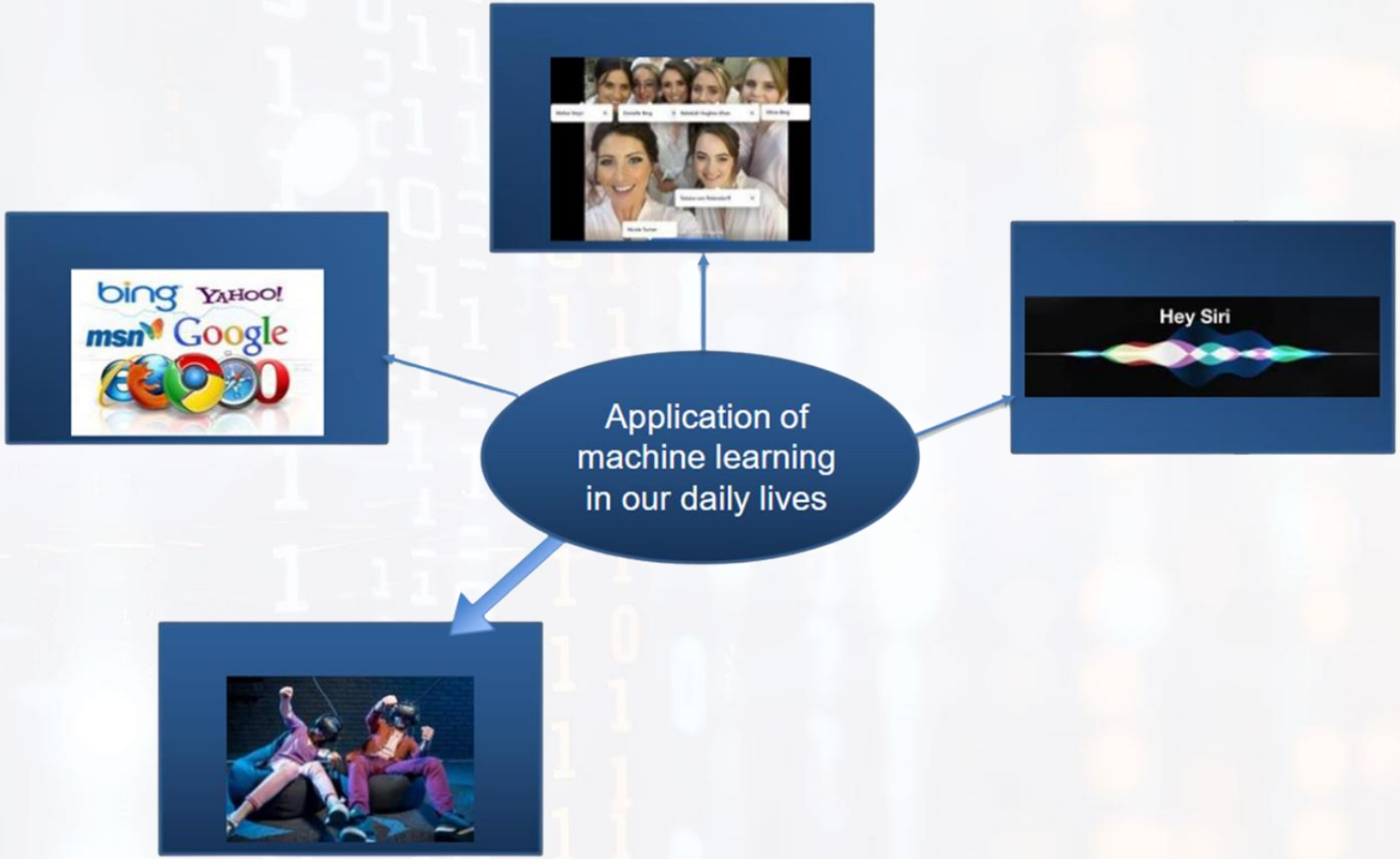
Life without machine learning

Life without machine learning

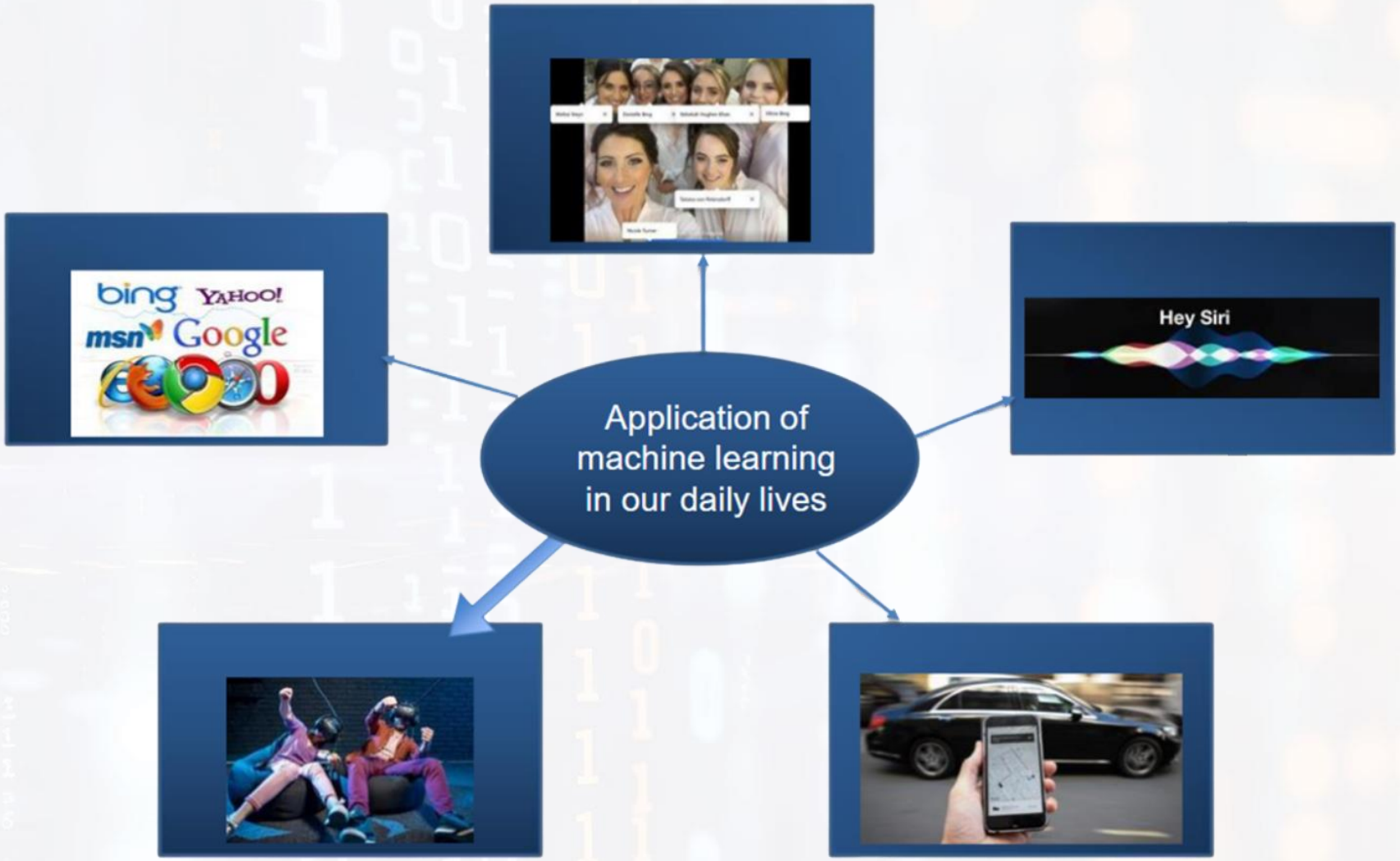
Life without machine learning

Life without machine learning

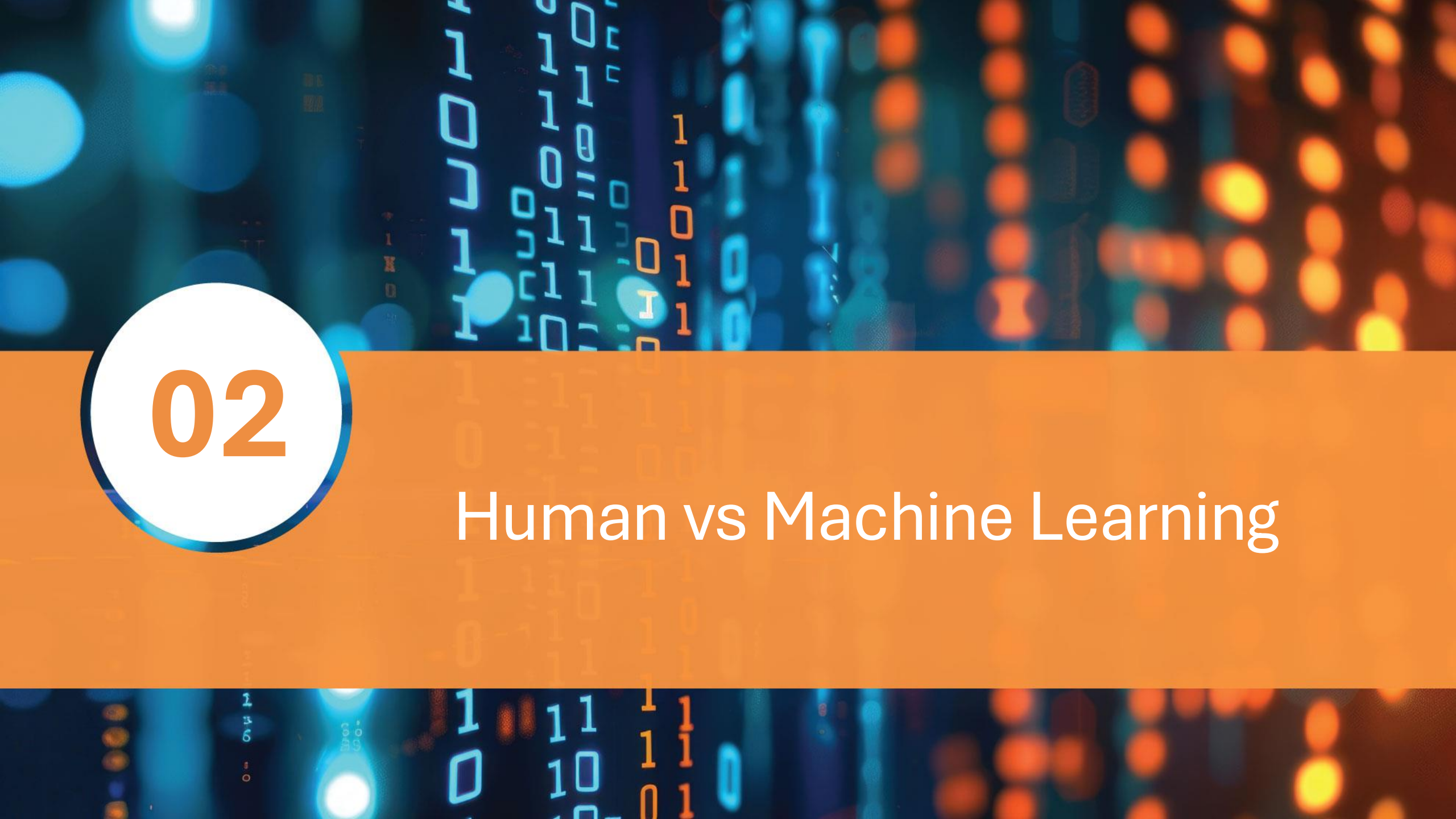
Life without machine learning



Life without machine learning



Life without machine learning

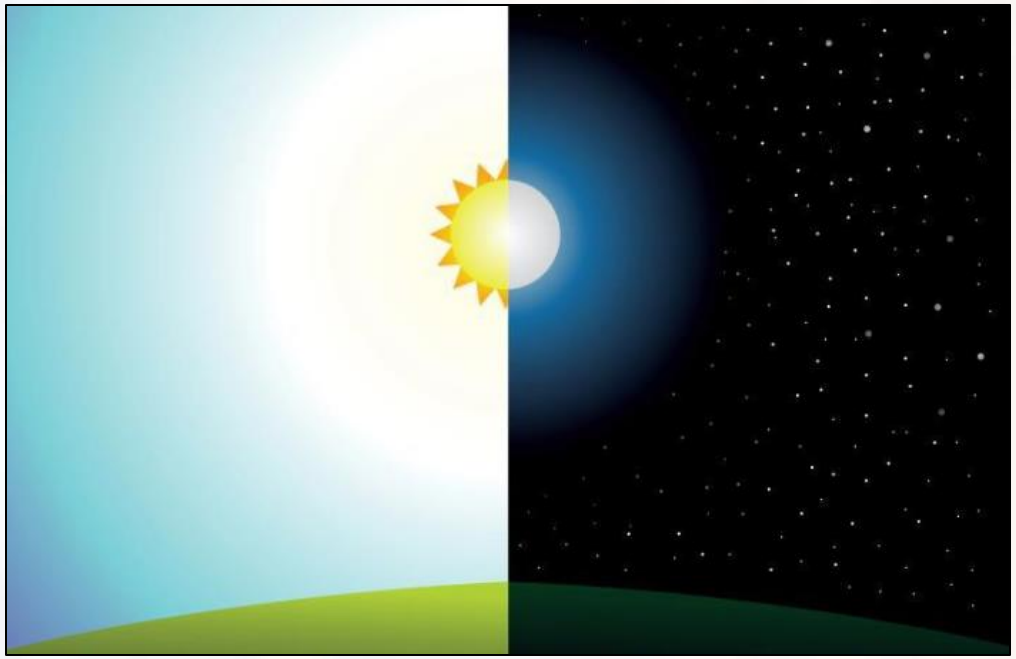


02

Human vs Machine Learning

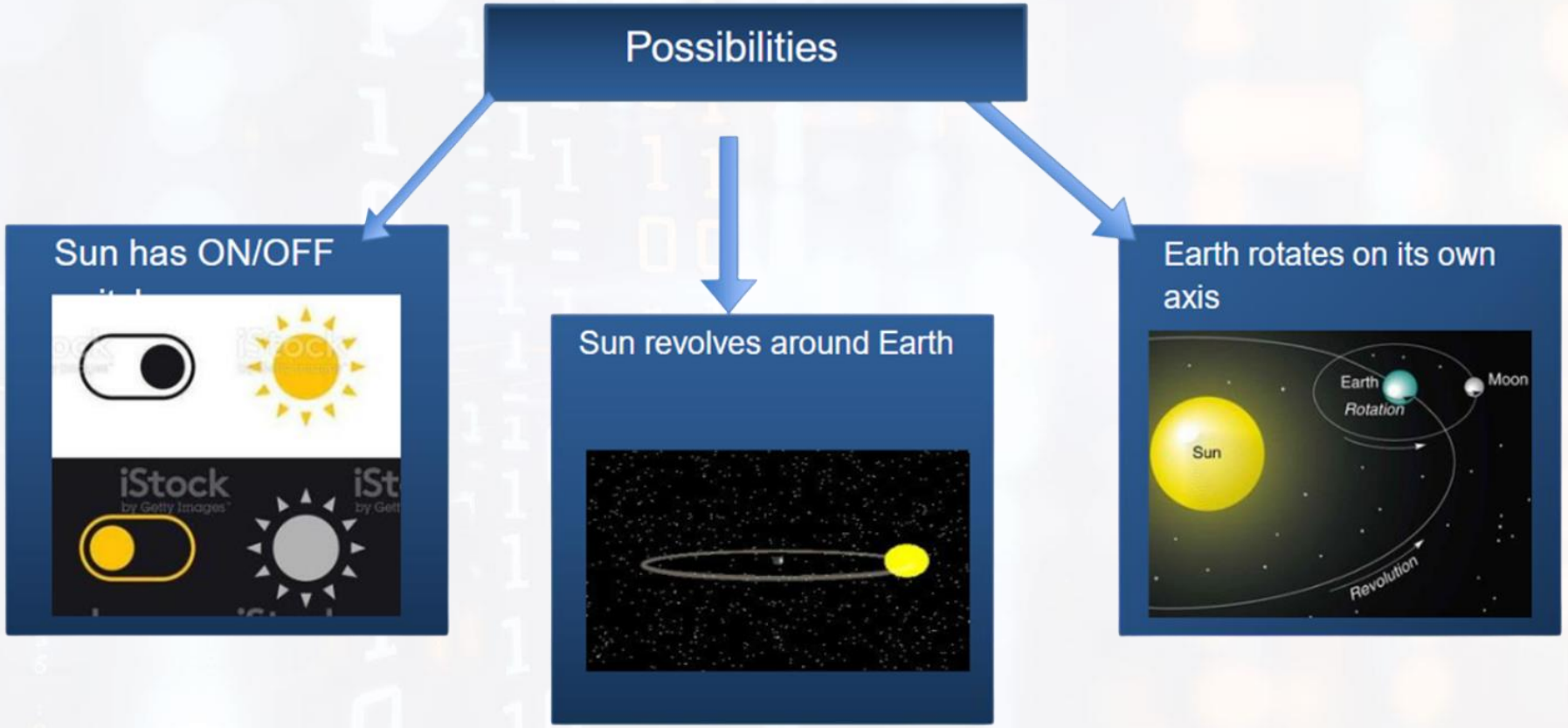
How do humans learn?

- Learning
  - Observe a phenomenon
  - Recognise a pattern
  - Understand how pattern occurs by determining the relationship between the factors involved
- Example: Occurrence of day and night
- [Human Learning vs Machine Learning | by Gaurav Goel | Towards Data Science]
- We **observe a phenomenon**(night and day) and We **recognise a pattern**
  - The pattern suggests that the surface of the Earth receives light from the sun alternately



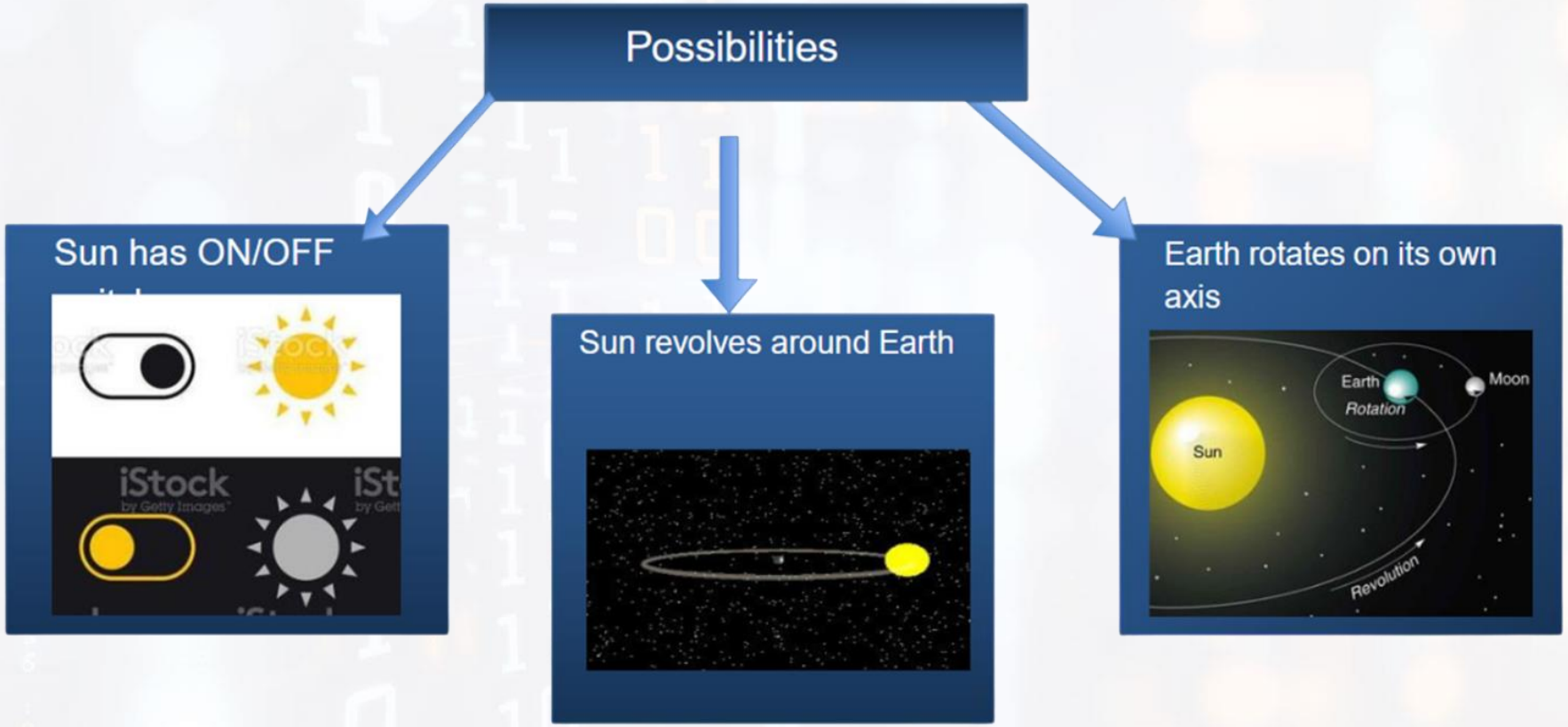
How do Machines learn?

- **Example: Occurrence of day and night**
- Model 1: Day/Night is a function of the ON/OFF switch of the Sun



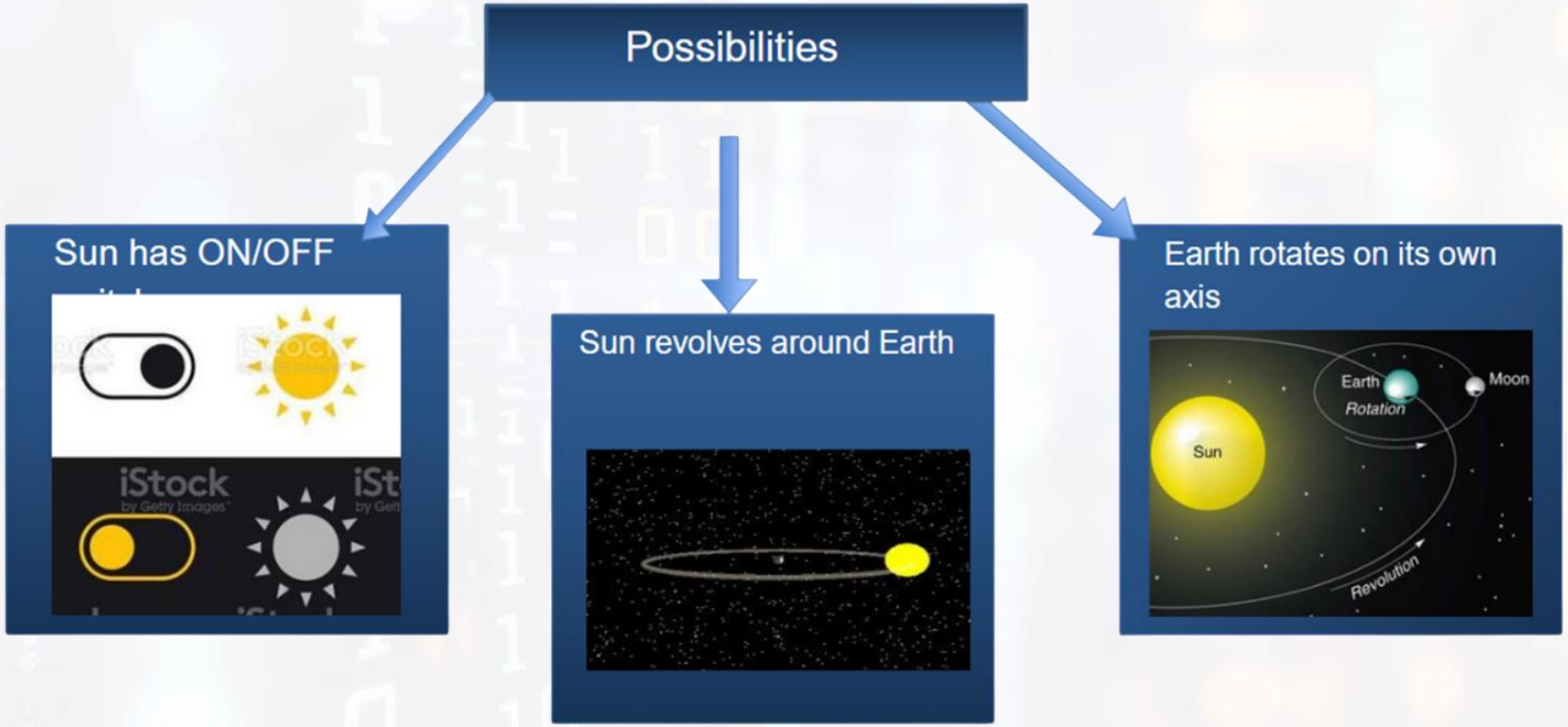
How do Machines learn?

- **Example: Occurrence of day and night**
- Model 2: Day/Night is a function of the Sun revolving around the Earth



How do Machines learn?

- **Example: Occurrence of day and night**
- Model 3: Day/Night is a function of the Earth rotating on its own axis



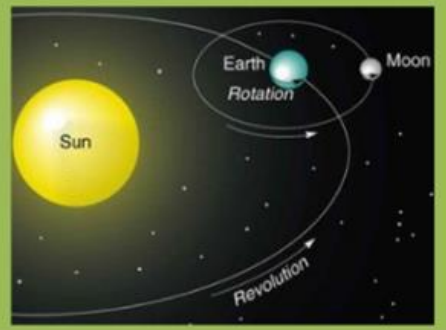
How do Machines learn?

- Example: Occurrence of day and night
- Model 3: Day/Night is a function of the Earth rotating on its own axis
- Model 3 gives most accurate explanation of the Day/Night phenomenon
- →Gives the **best fit** for the observations that explain this phenomenon

- Model can now be used to predict future outcomes for this phenomenon.
  - Model can predict the occurrence of day/night depending on which side/surface of the Earth is facing the Sun

Possibilities

Earth rotates on its own axis



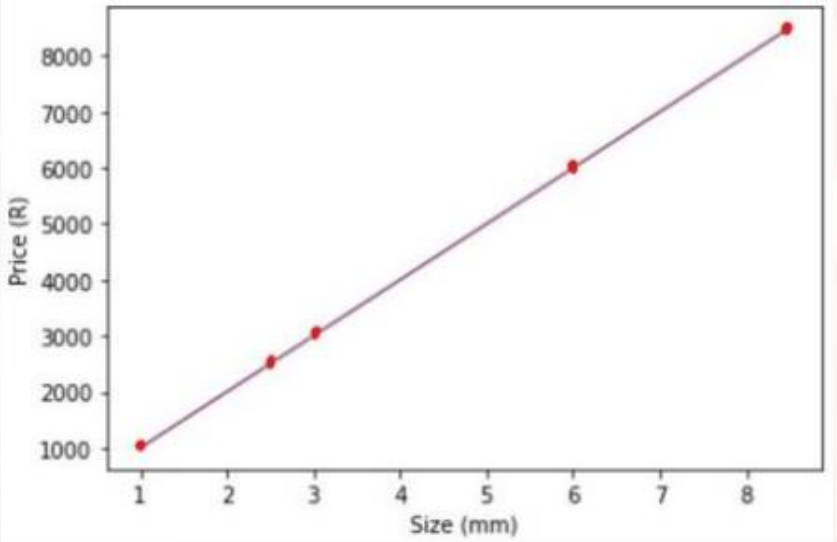
How do humans learn?

- **Human learning is therefore to:**
  - observe something
  - identify a pattern
  - build theory (model)
  - test this theory/model to see whether it fits in most/all cases

- **Is it possible for a machine to mimic this process of human learning?**
- →That is what the field of machine learning / artificial intelligence aims to do!

How do machines learn?

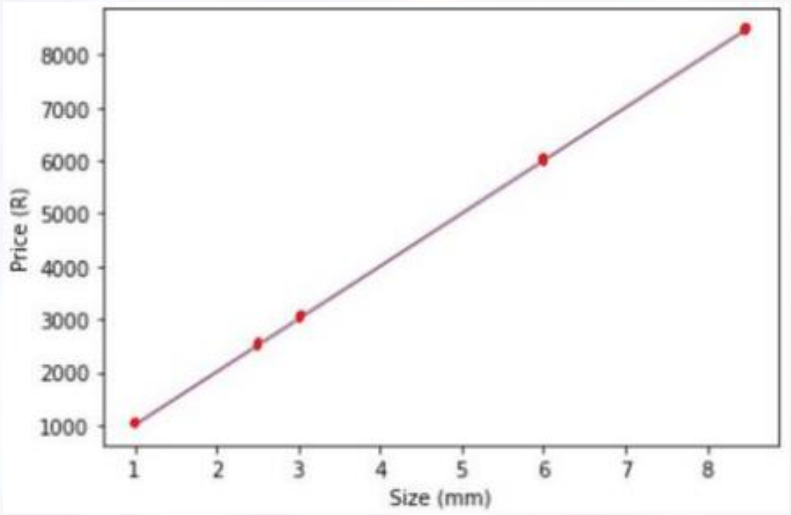
	Price of diamond (R)
	1000
	2500
	3000
	6000
	8500



How do machines learn?

- Consider a fake dataset consisting of the size and price of a diamond

Size (mm)	Price (R)
1	1000
2.5	2500
3	3000
6	6000
8.5	8500



- We can observe a pattern/relationship between the size and the price of the diamond
- →Pattern can be described by the model: **Price = 1000 * Size**

How do machines learn?

How would a machine determine this relationship?

1. It consists of all variables/factors involved (in this case, size and price of diamond) and assumes a relationship of the $Price = w * Size[where\ w = random\ value]$

How do machines learn?

How would a machine determine this relationship?

1. It consists of all variables/factors involved (in this case, size and price of diamond) and assumes a relationship of the $Price = w * Size[where\ w = random\ value]$
2. It assumes a value for $w[let's\ say, w = 950]$

Size of diamond	Actual price of diamond	Calculated price of diamond	Error
1	1000	=1.0*950	
2.5	2500	=2.5*950	
3	3000	=3.0*950	
6	6000	=6.0*950	
8.5	8500	=8.0*950	

How do machines learn?

How would a machine determine this relationship?

1. It consists of all variables/factors involved (in this case, size and price of diamond) and assumes a relationship of the $Price = w * Size[where\ w = random\ value]$
2. It assumes a value for $w[let's\ say, w = 950]$

Size of diamond	Actual price of diamond	Calculated price of diamond	Error
1	(1000	=1.0*950) →	
2.5	(2500	=2.5*950) →	
3	(3000	=3.0*950) →	
6	(6000	=6.0*950) →	
8.5	(8500	=8.0*950) →	

Calculate average error for whole dataset

How do machines learn?

How would a machine determine this relationship?

1. It consists of all variables/factors involved (in this case, size and price of diamond) and assumes a relationship of the $Price = w * Size[where\ w = random\ value]$
2. It assumes a value for $w[let's\ say, w = 980]$

Size of diamond	Actual price of diamond	Calculated price of diamond	Error
1	(1000	=1.0*980) →	
2.5	(2500	=2.5*980) →	
3	(3000	=3.0*980) →	
6	(6000	=6.0*980) →	
8.5	(8500	=8.0*980) →	

Calculate average error for whole dataset

3. Update value for w and repeat step 2

How do machines learn?

How would a machine determine this relationship?

1. It consists of all variables/factors involved (in this case, size and price of diamond) and assumes a relationship of the $Price = w * Size[where\ w = random\ value]$
2. It assumes a value for $w[let's\ say, w = 1000]$

Size of diamond	Actual price of diamond	Calculated price of diamond	Error
1	(1000	=1.0*1000) →	
2.5	(2500	=2.5*1000) →	
3	(3000	=3.0*1000) →	
6	(6000	=6.0*1000) →	
8.5	(8500	=8.0*1000) →	

Calculate average error for whole dataset

3. Update value for w and repeat step 2
4. Continue until a minimum average error is obtained

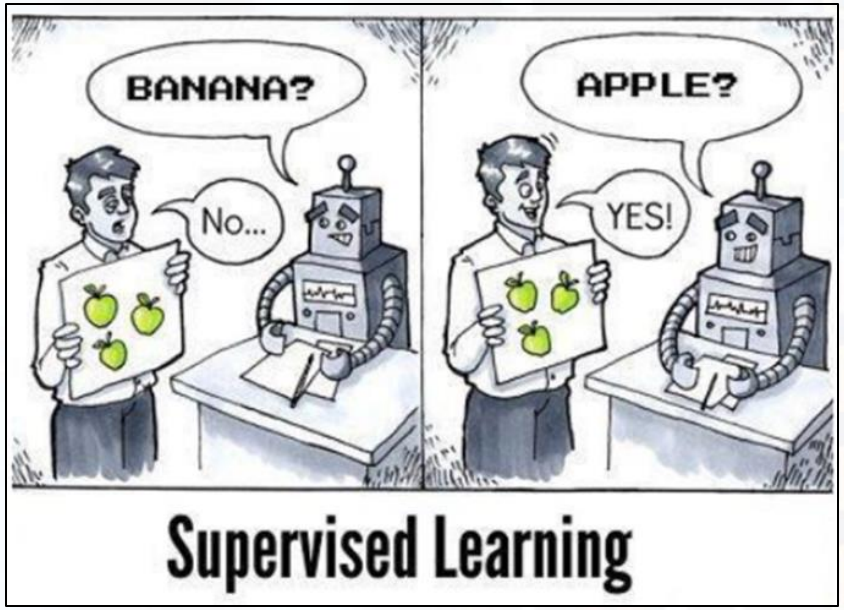
   (-- This model describe a relationship that **best fits the data**)

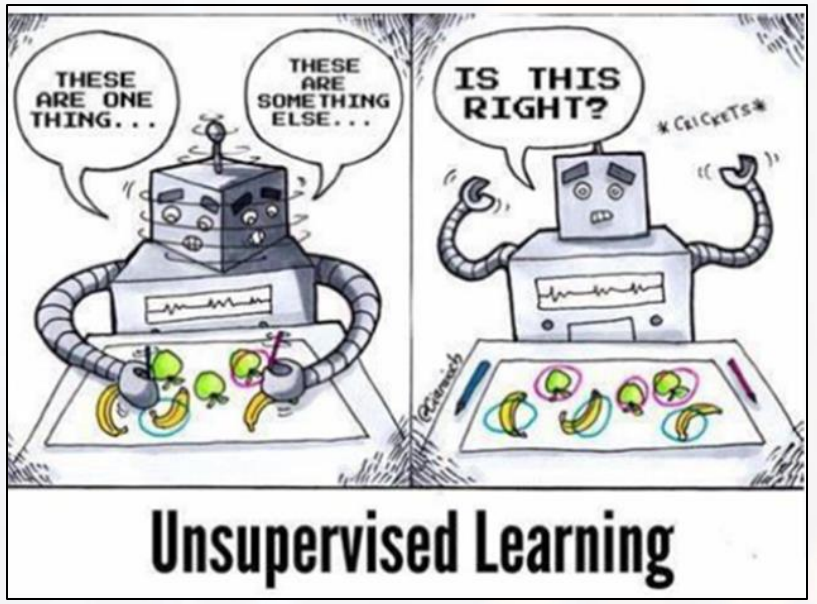
How do machines learn?

- Even though a very simple and perfectly linear relationship was considered for the diamond example, a machine learning algorithm has the **ability to identify complex patterns and highly non-linear relationships** that we cannot identify with our human eye.
- **Machine learning model:**
  - Considers a set of data that describes (a various number of) variables
  - Tries to determine a relationship between these variables and defines/expresses this relationship as a mathematical function / modelprocess of updating the mathematical function until it optimally fits the data = **training**
- The trained model can then be used to make future predictions on previously unseen data that describe the same phenomenon/process/system

How do machines learn?

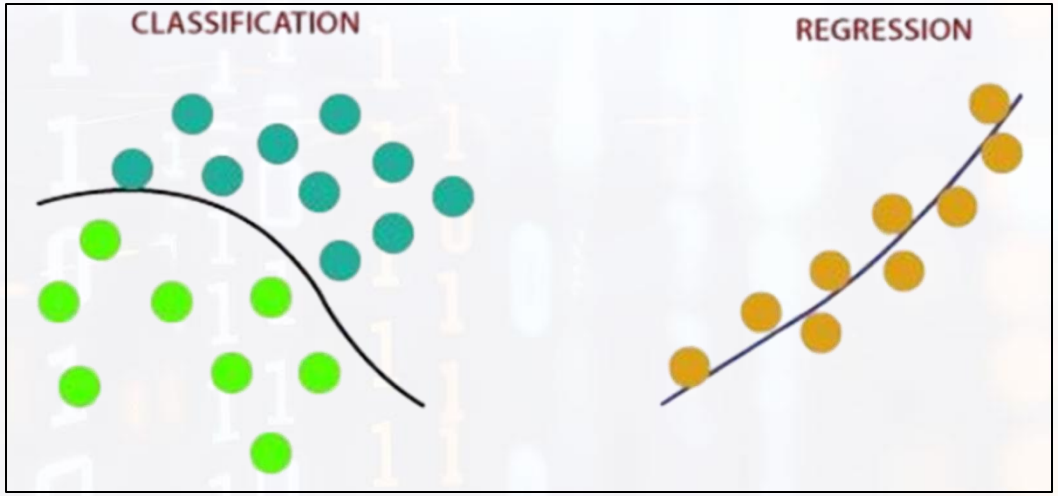
- Machines learn in different ways with various amounts of “supervision”
- Machine learning model/algorithm is mainly classified as **supervised** or **unsupervised**





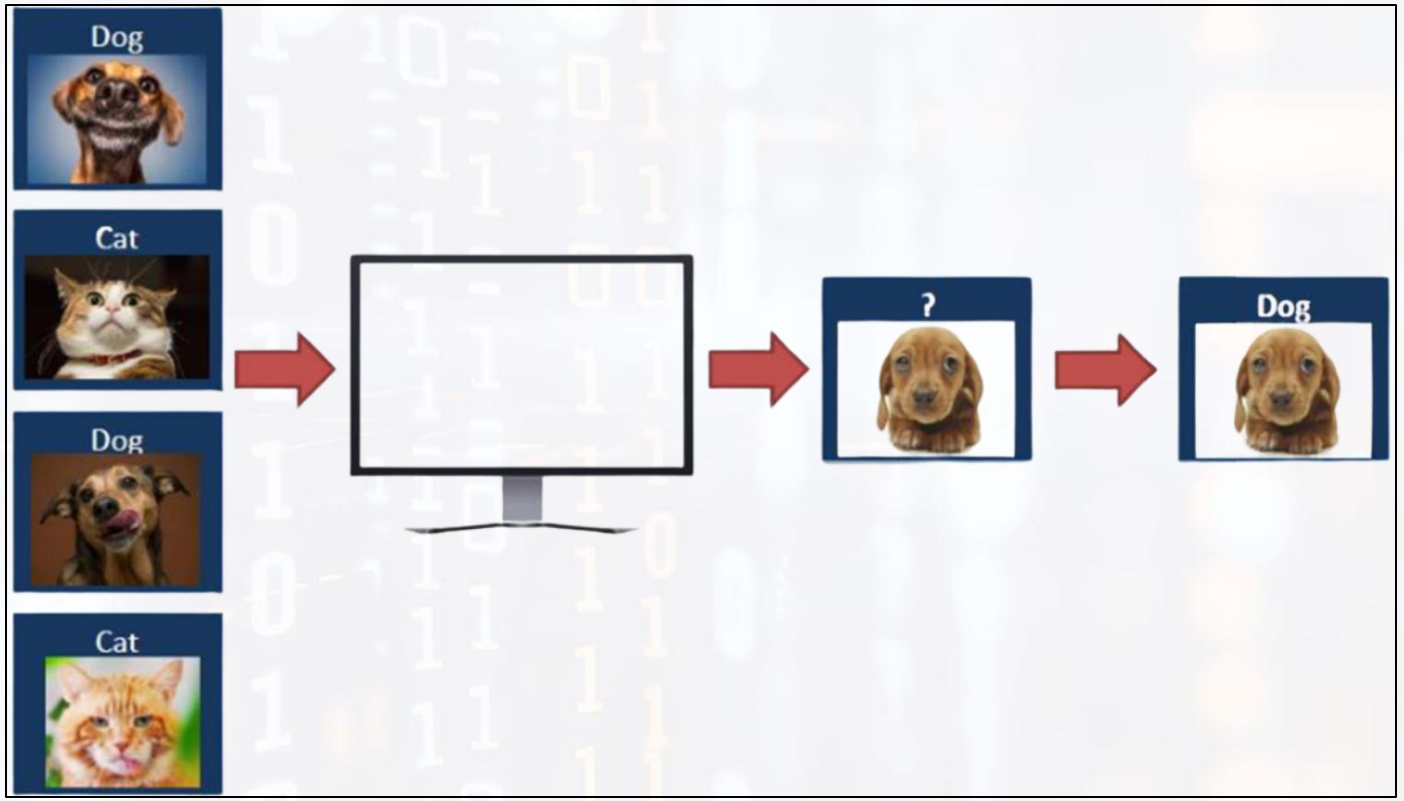
How do machines learn?

- Machine learning model trains on samples that consist of both input and output by determining a function that best describes the relationship between the given input and the output samples.
- Trained model
- Receives new, previously unseen input data and predicts the corresponding output/label.
- Two supervised learning techniques:



How do machines learn?

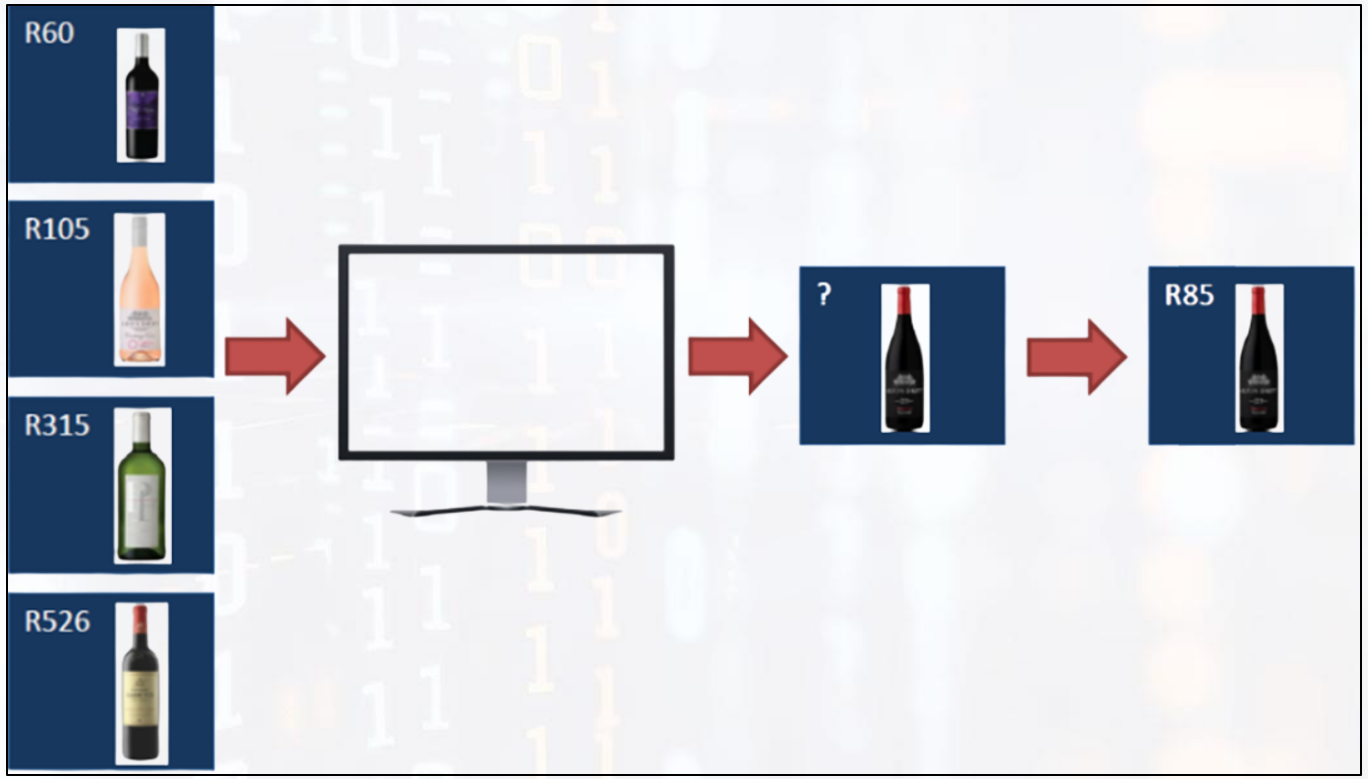
- Classification:
  - Task of machine learning algorithm is to categorize or predict a discrete class label.



How do machines learn?

- Regression:

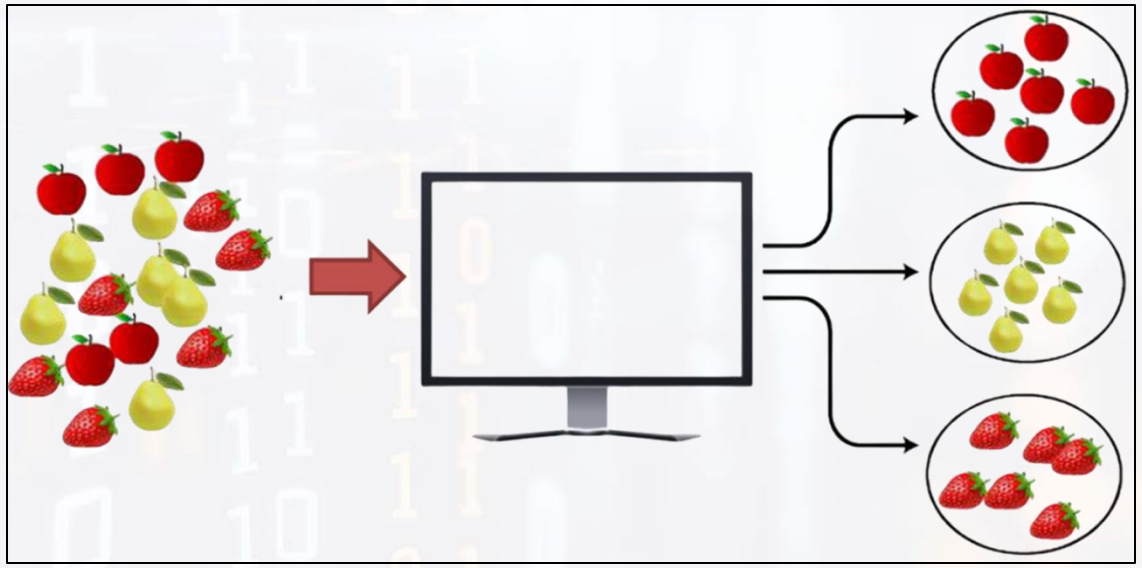
  -Regression is about predicting continuous numerical values (e.g., house prices, temperature) rather than discrete categories.



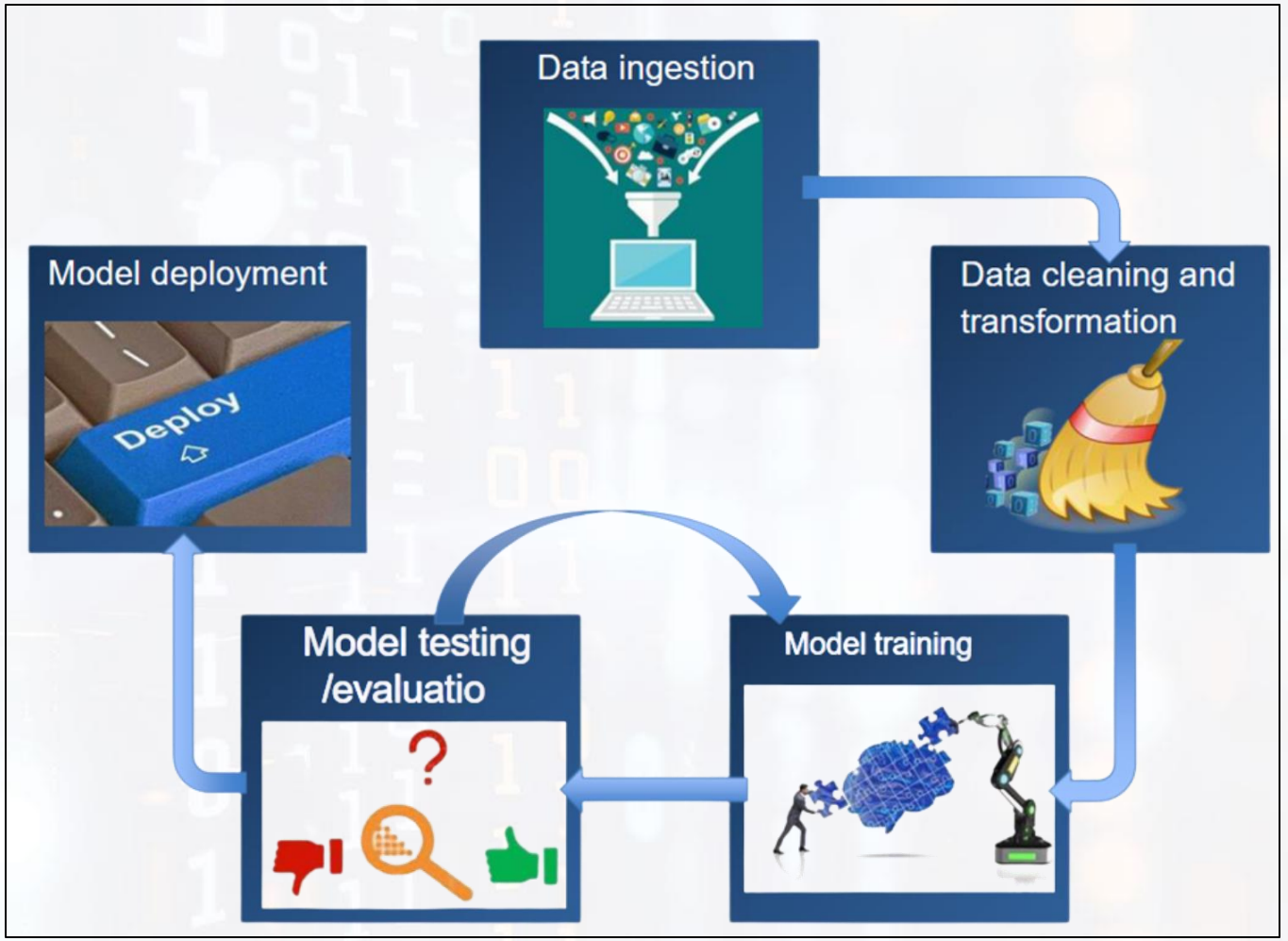
Machine learning types:

Unsupervised learning

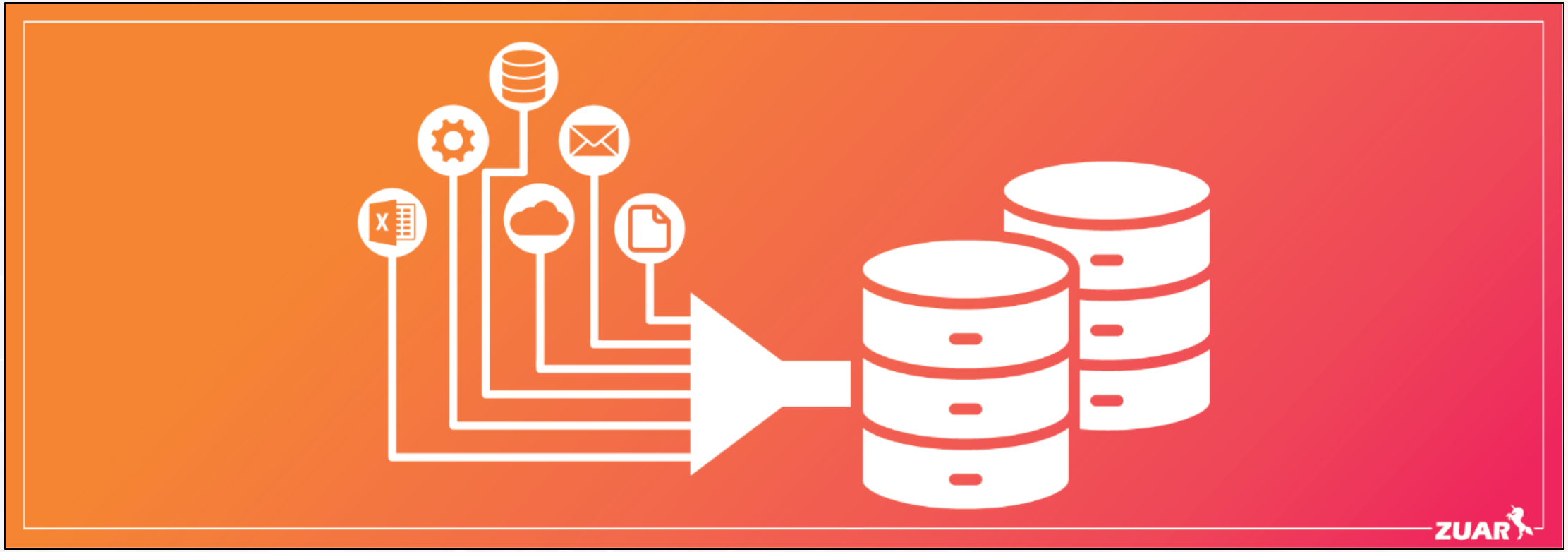
- **Unsupervised learning:**
  - Machine learning model trains on samples that consist of inputs but no corresponding outputs/labels.
  - Goal is to analyse the dataset and cluster the data into different classes based on patterns/similarities that it has found within the data.
- Most common unsupervised learning technique: **clustering**



Machine learning pipeline



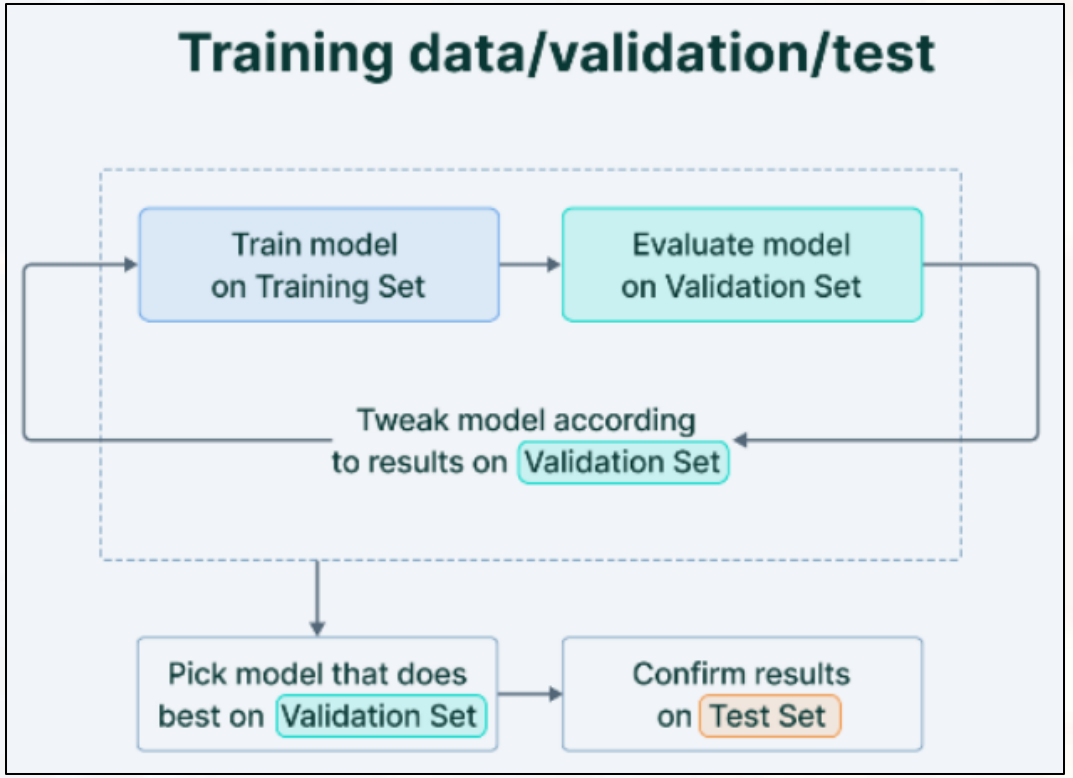
Machine Learning Pipeline – Data Ingestion



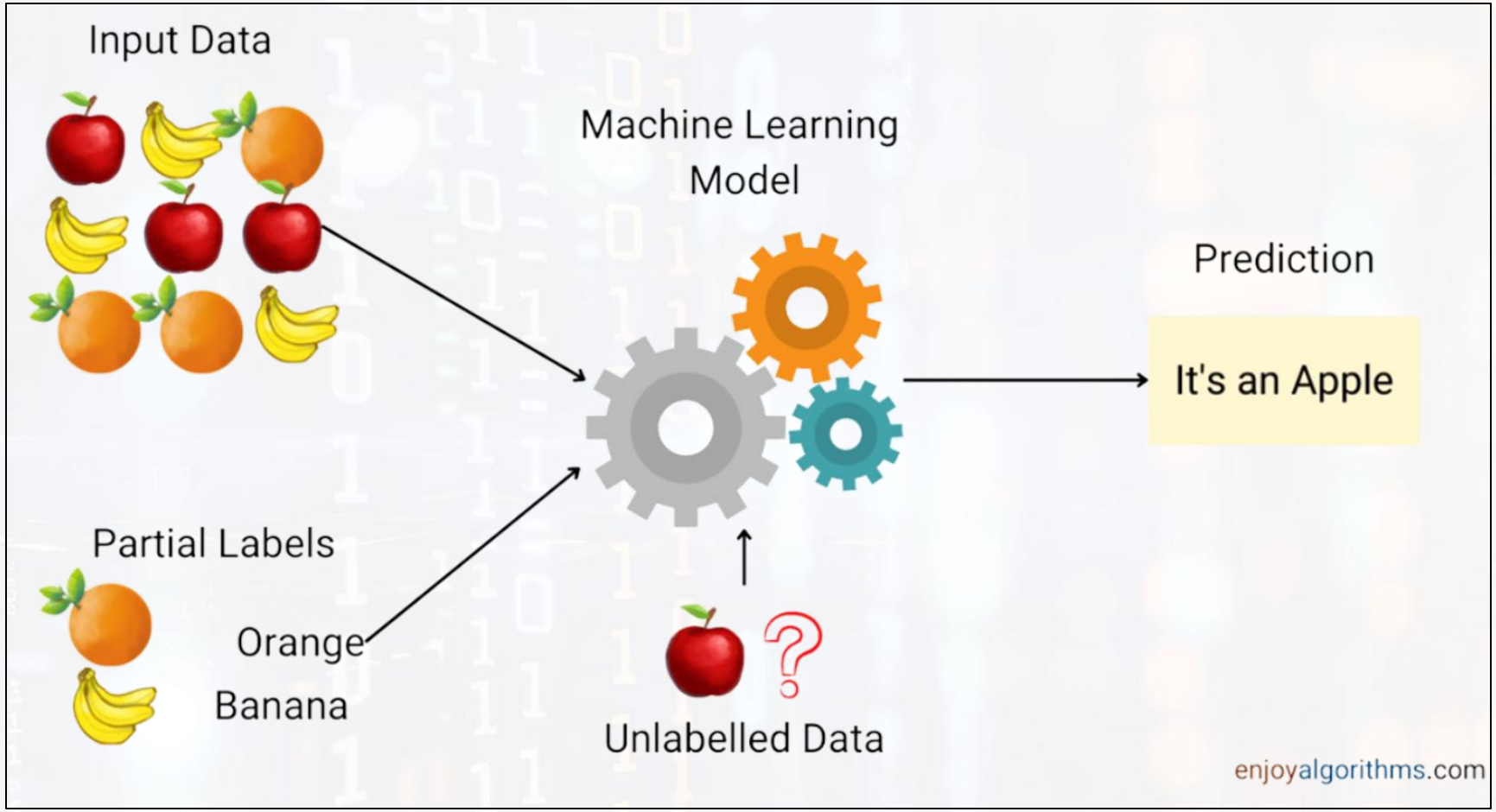
Machine Learning Pipeline – Data Cleaning

Machine Learning Pipeline – Data Splitting

- Splitting the data into training, validation, and testing sets.
  - Training data (70%): Used to train the model
  - Validation data (20%): Used to fine-tune hyperparameters and prevent overfitting.
  - Testing data (10%): Used for final evaluation of the model's generalizability.



Machine Learning Pipeline – Training



Machine Learning Pipeline – Testing and Evaluation

Evaluating the model's performance on the testing data.

- Common metrics: Accuracy, Precision, Recall, F1-Score, AUC-ROC for classification tasks
- Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) for regression tasks.

Machine Learning Pipeline – Deployment

- Deploying the trained model into production.
- **Considerations**: Scalability, efficiency, infrastructure.
- Monitoring the model's performance and retraining as needed.

ML OR NOT?


Questions?
NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM
DIRISA