ALGOSOC
UOB

DATA SCIENCE & MACHINE LEARNING WORKSHOP

Week 4 - Intro to Machine Learning

GOOGLE GEMINI X ALGOSOC TALK ON AI

WE WILL HAVE SOME EXTERNAL SPEAKERS THAT WILL BE DELIVERING AN ENGAGING WORKSHOP FOR STUDENTS ON HOW TO LEVERAGE GEMINI, GOOGLE'S AI TOOL, TO PRACTICALLY AND ETHICALLY ENHANCE—NOT REPLACE—THEIR ACADEMIC STUDIES.

THERE WILL BE 2 SLOTS:

1. 2:30PM - 3:30PM
2. 4:00PM - 5:00PM.

YOU ARE FREE TO ATTEND EITHER ONE - THEY WILL BE THE SAME TALK.

LIMITED AVAILIBILITY!!!
SCAN THE QR CODE TO SECURE YOUR SPOT!!!

LOCATION - Y3-G34
DATE - 20TH NOV 2025

AGENDA

Week 4 topic:

- Do we need modelling and machine learning
- Types of Machine learning (supervised, unsupervised, reinforced)
- Intro to Supervised (Types of Supervised)
- Uncertainties
- Training vs. testing data
- Modelling example

Full agenda this semester:
https://bit.ly/DataScienceAlgosoc

Repository:
https://github.com/AlgoSoc/Data-Science

WHEN DO WE NEED MODELLING

The term machine learning has been thrown around so much, seems like we need to build **a machine learning predictive model** for everything now.

But do we actually need to?

Consider this case:

- Can we train a machine learning predictive model to predict celsius temperature given farrenheit? Yes
  - Should we? No

- Say your company has a promo set up, customers above 60 years old are entitled to discounts, can we make a predictor for it? Yes
  - Should we? No
- Given text data with **information we can't really associate manually**, can we make a sentiment classifier? Yes
  - Should we? Yes

WHEN DO WE NEED MODELLING

ALGOSOC UOB

We should use machine learning when:

- The problem seems unsolveable (intrinsically hard)
- The problem is big (we can extract information but would just be too tedious)
- Information to make a decision are hard to be interpreted

We should **NOT** use machine learning when:

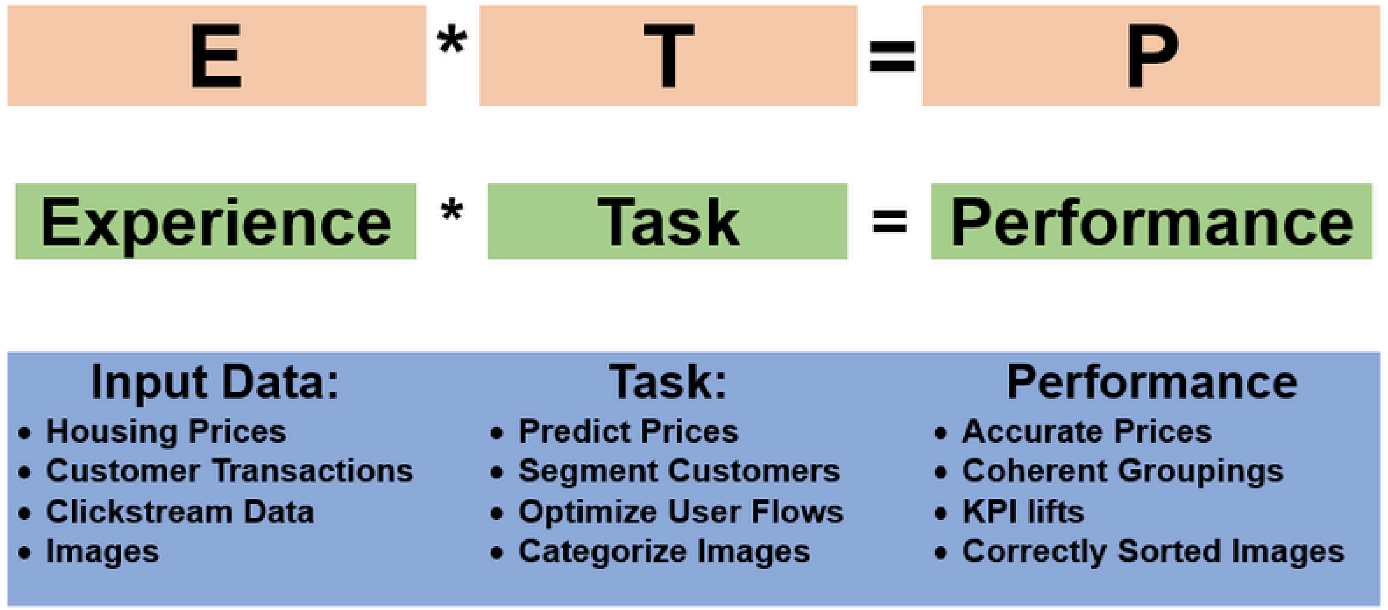
- The problem is straightforward
- We have limited resources
- Supporting information for decision can be determined easily (e.g., setting age threshold for discount)

Keep in mind, when using machine learning, we introduce **uncertainties**

WHAT IS MACHINE LEARNING

What is Machine Learning?

- Arthur Samuel (1959): Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998): A computer programme is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
- Kevin Murphy (2012): The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.
- Oxford Languages Dictionary: the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.



https://www.linkedin.com/pulse/what-machine-learning-ml-mohammad-mehrabani-2qgie

*taken from Leandro Minku's Lecture on Machine Learning

TYPES OF MACHINE LEARNING

Three Types of Machine Learning

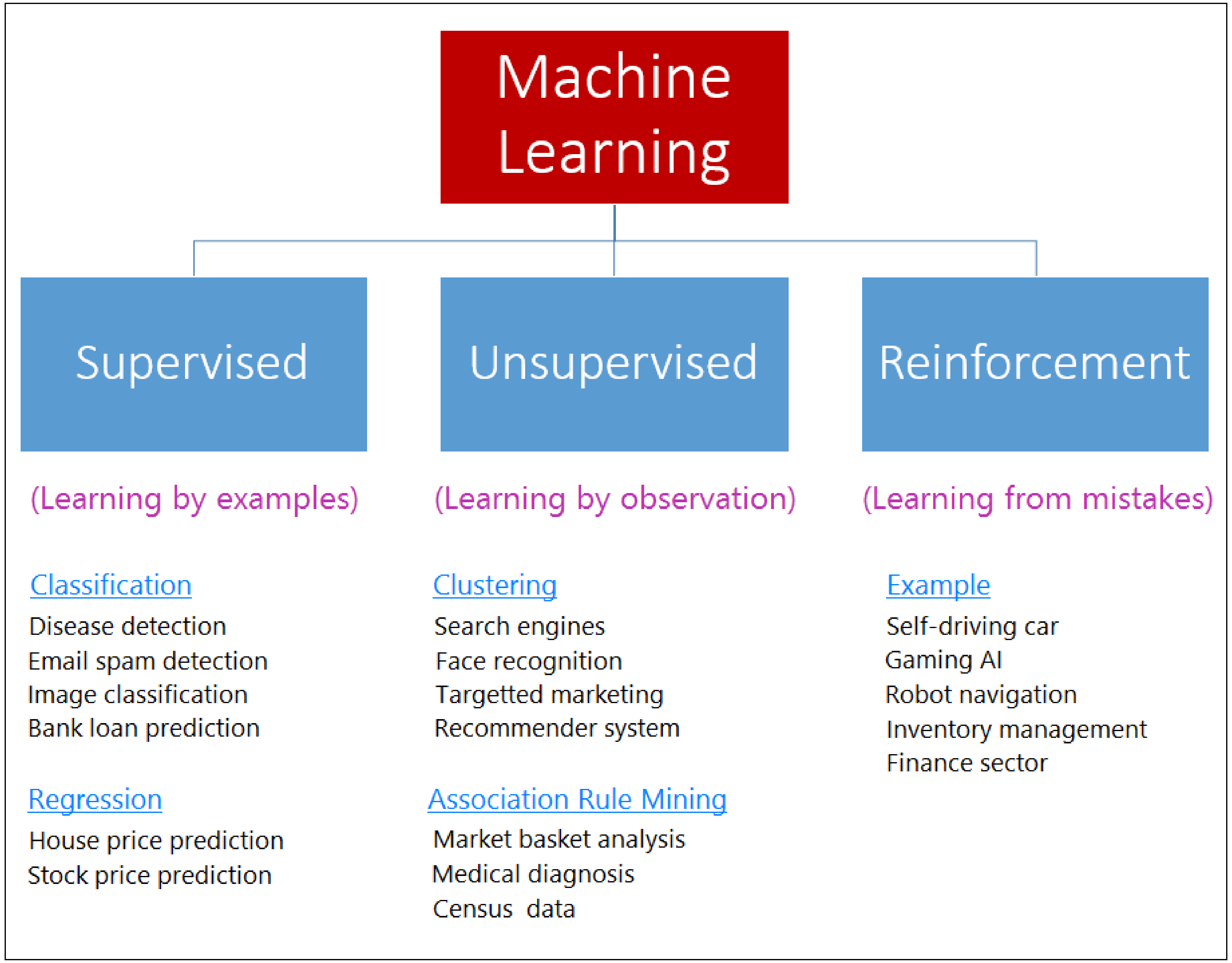
Supervised Learning	Unsupervised Learning	Reinforcement Learning
Has outcome information (“labels”)	No outcome information available	Makes decisions based on trial and error
Finds patterns that relate to those outcomes	Analyzes or identifies groups without labels or human instruction	Decision-making algorithm is constantly refined based on “rewards”
Uses patterns to predict outcomes not yet known	Offers insights into characteristics that define groups	Excels in complex situations

Pecan

https://www.pecan.ai/blog/3-types-of-machine-learning/

TYPES OF MACHINE LEARNING

ALGOSOC
UOB

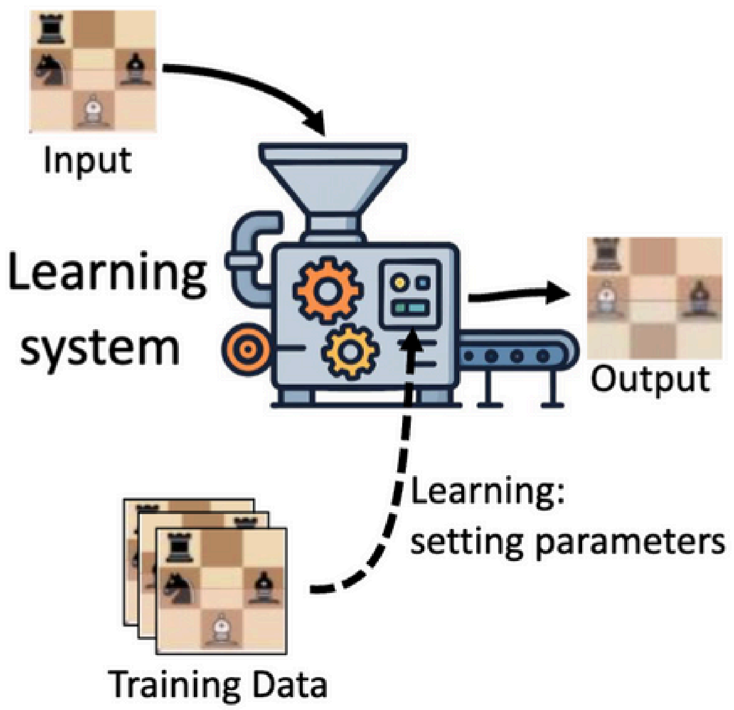


https://www.techguruspeaks.com/types-of-machine-learning/

SUPERVISED LEARNING

In supervised learning we are concerned in letting our machine to learn from our example of inputs (X) and (desired) output (y)

Most of the time, it means we want the computer to ***learn*** the parameters (settings) of our model. By letting the machine learn the appropriate parameters, we are performing supervised learning



Taken from Alexander Krull’s lecture on Neural Computation

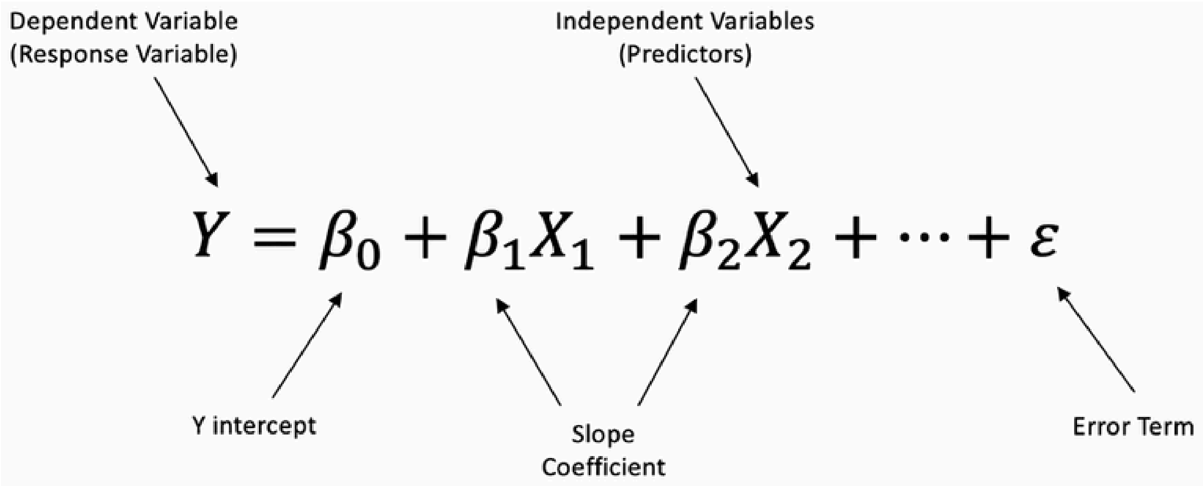


Image Source: https://www.researchgate.net/figure/Linear-regression-equation_fig1_373123252

If our model is a linear function, we would usually want to let the computer learn what are the appropriate **y-intercept and slope coefficient**

While ε is the **underlying uncertainty**

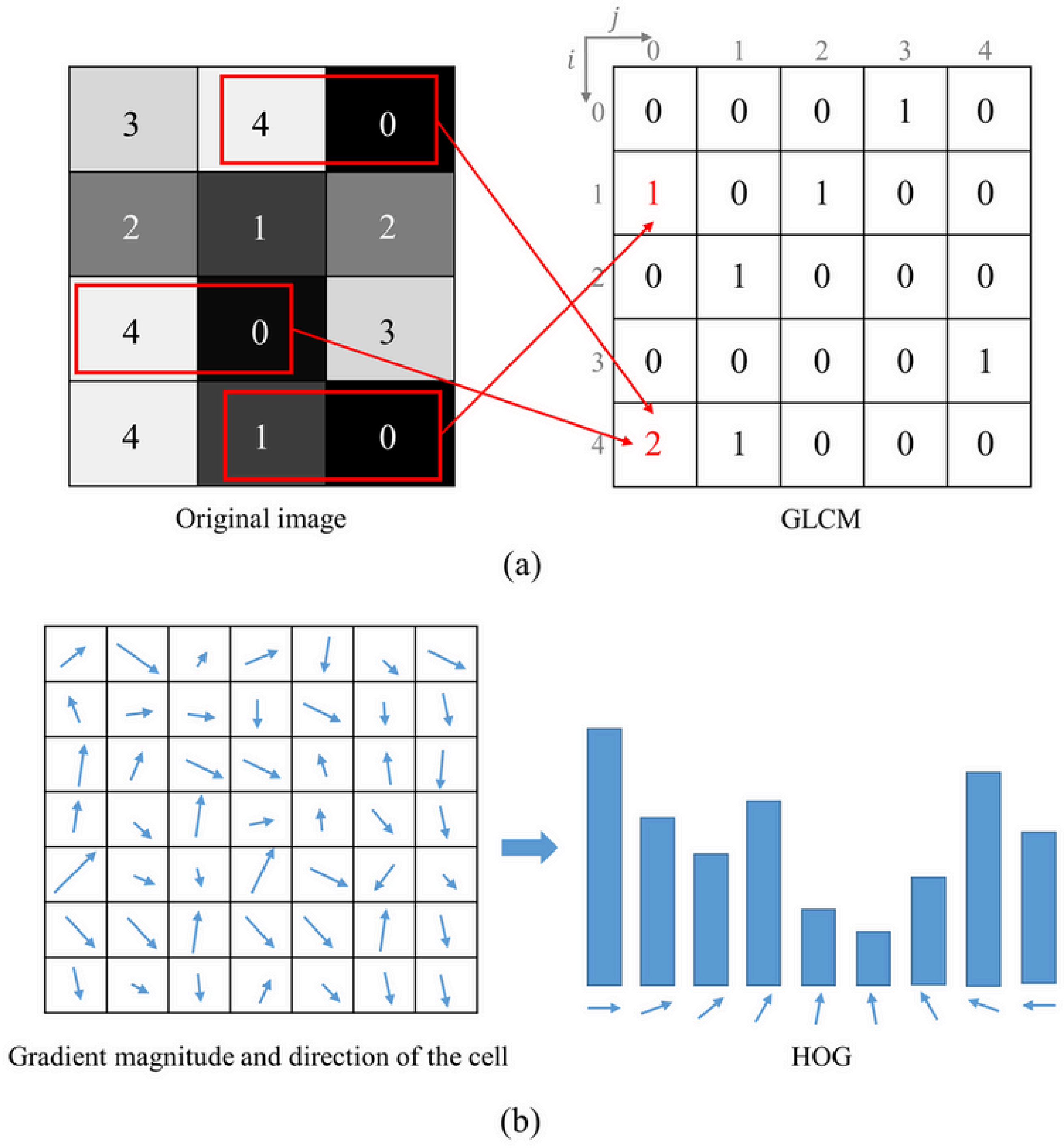
SUPERVISED LEARNING

ALGOSOC
UOB

Task	Task Description	Input (X) Type	Output (y) Type
Regression	Based on the given information, predict a numeric value	Numeric data, categorical data	Numeric values
Classification	Based on the given information, discriminate between two classes of information that the model has known of	Numeric and categorical data	categorical values

What about image and textual data?

SUPERVISED LEARNING



https://www.researchgate.net/figure/Feature-extraction-methods-in-image-processing-GLCM-and-HOG-a-An-example-of_fig1_337940559

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

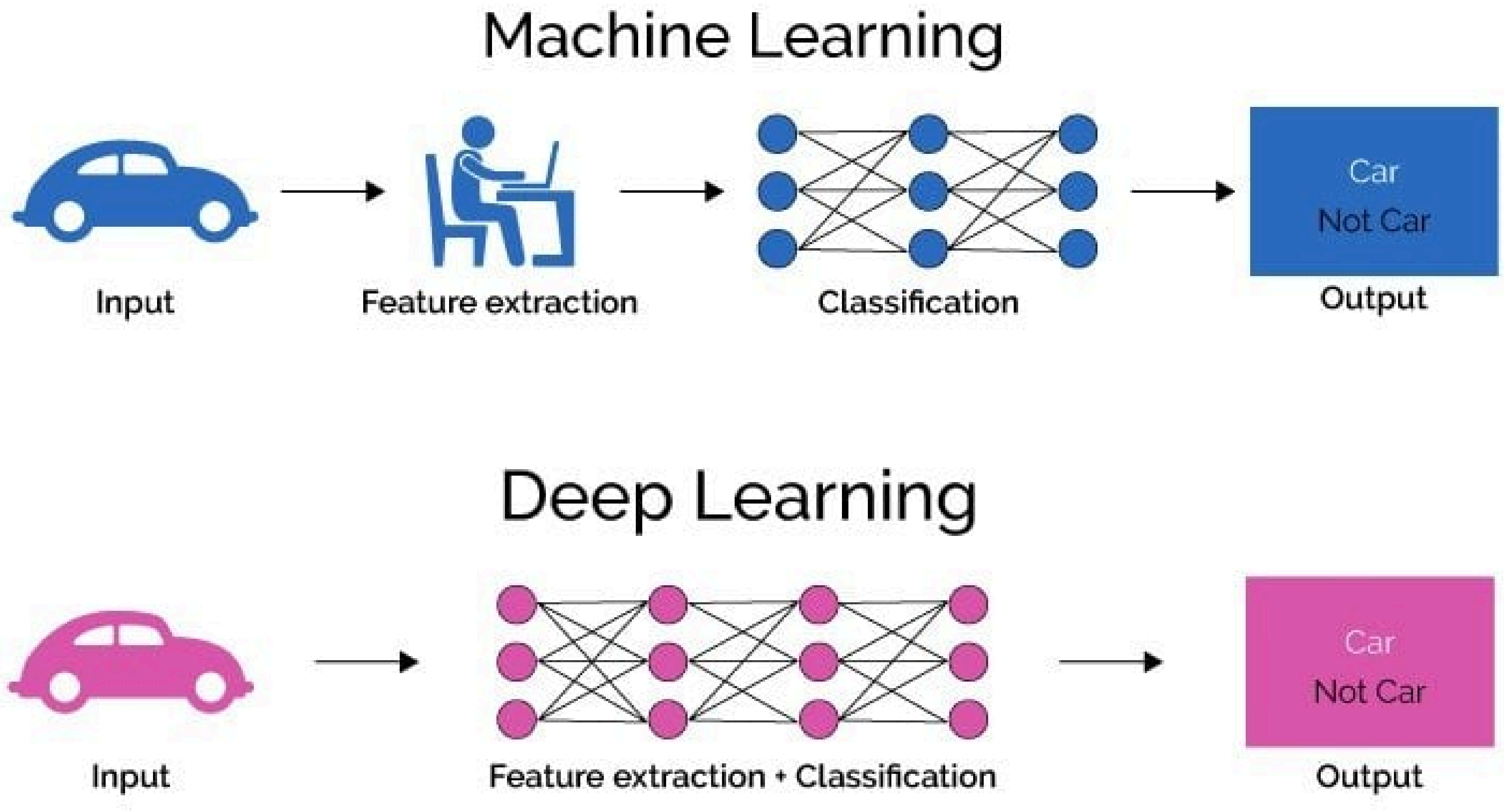
https://www.researchgate.net/figure/Feature-extraction-methods-in-image-processing-GLCM-and-HOG-a-An-example-of_fig1_337940559

Images are essentially matrix, and we can extract numerical feature such as **Histogram of Oriented Gradients** to become numerical inputs for our model

Texts can be represented numerically using features such as **TF-IDF.** These can then become input for our models.

SUPERVISED LEARNING

Calculating Histogram of Oriented Gradients, and TF-IDF can be tedious and more often not representative of the data itself. To solve this, we can use deep learning where (most of the time) we can use the unstructured data in its original form



https://viso.ai/deep-learning/deep-learning-vs-machine-learning/

UNCERTAINTY

To measure how well our model is doing, we can quantify the performance of our models.

f(X): Model prediction

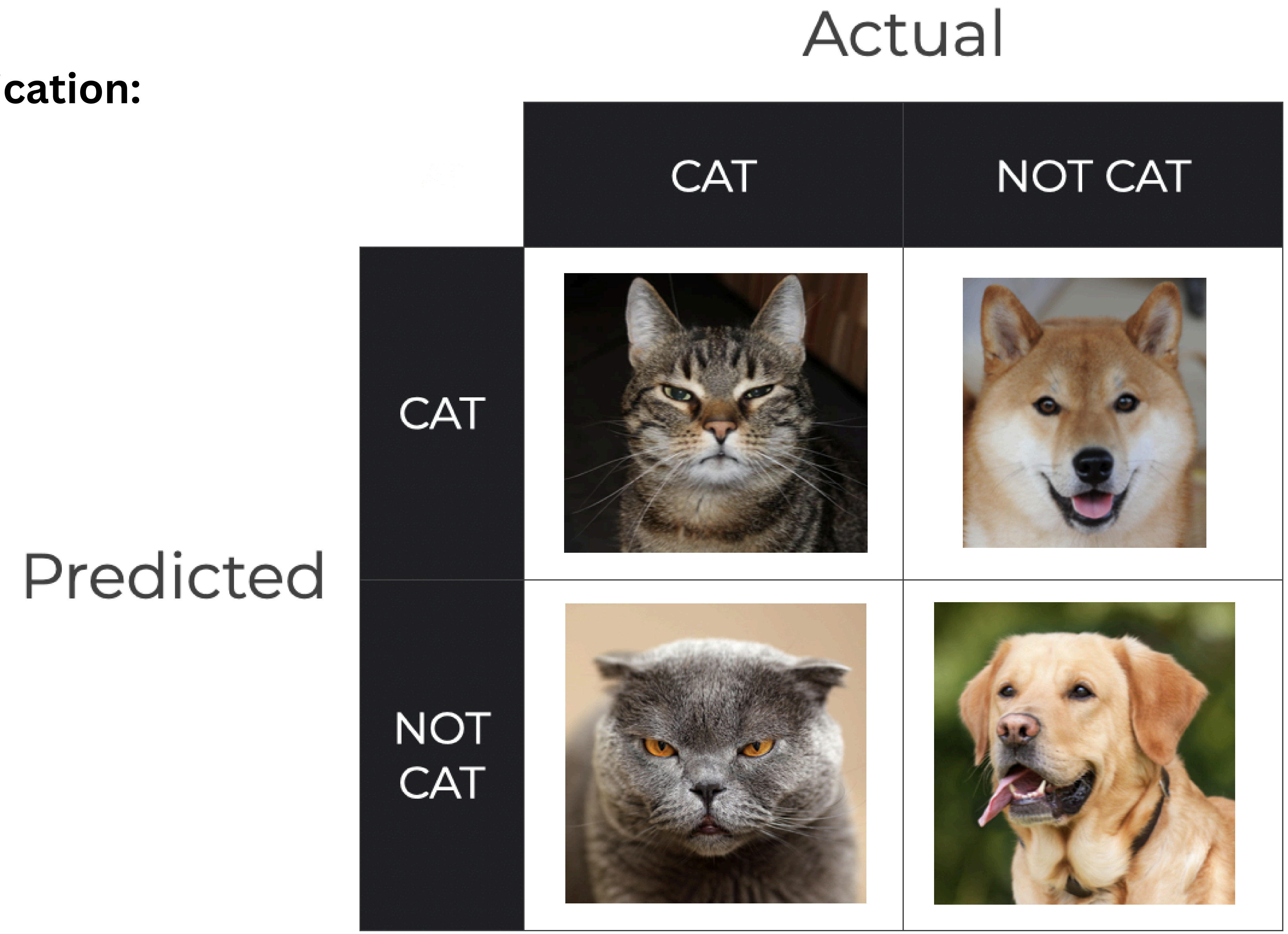
y: Ground truth

Performance metric calculation

Performance metric result

UNCERTAINTY

For classification:



https://sharpsight.ai/blog/confusion-matrix-explained/

UNCERTAINTY

For classification:

		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

https://www.researchgate.net/figure/Calculation-of-Precision-Recall-and-Accuracy-in-the-confusion-matrix_fig3_336402347

$$\boxed{\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}}$$

UNCERTAINTY

For regression:

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

https://medium.com/@chipkarsahil/understanding-regression-performance-metrics-a-simple-guide-4b7f58bd71f8

We would define the error *e* as:

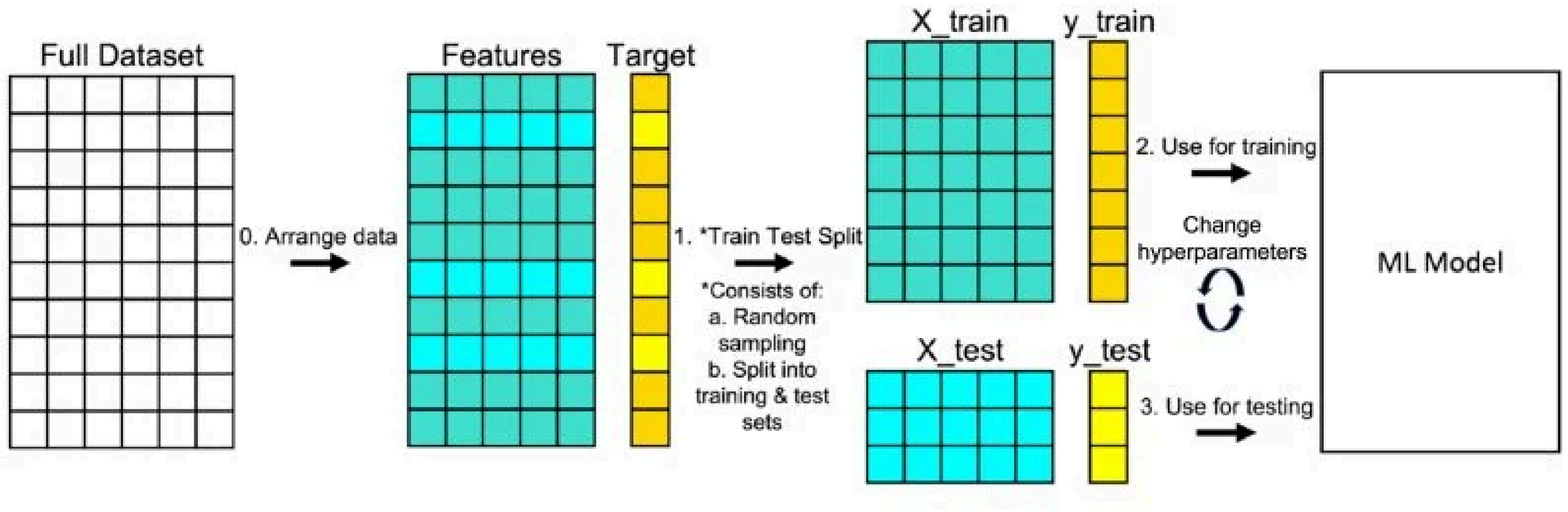
$e = (f(x) - y)$

where:

- f(*x*) : model prediction
- y: true value

TRAIN AND TEST

We want to measure the model's performance on **unseen data (i.e., data not used on training)**



https://builtin.com/data-science/train-test-split

The uncertainty measurment usually would be evaluated against the model's performance with the test data, **NOT** the training data

Recommended further read on:

- Model selection (cross validation)
- Validation set in deep learning

HANDS ON

Link
https://bit.ly/AlgosocWk4

Colab