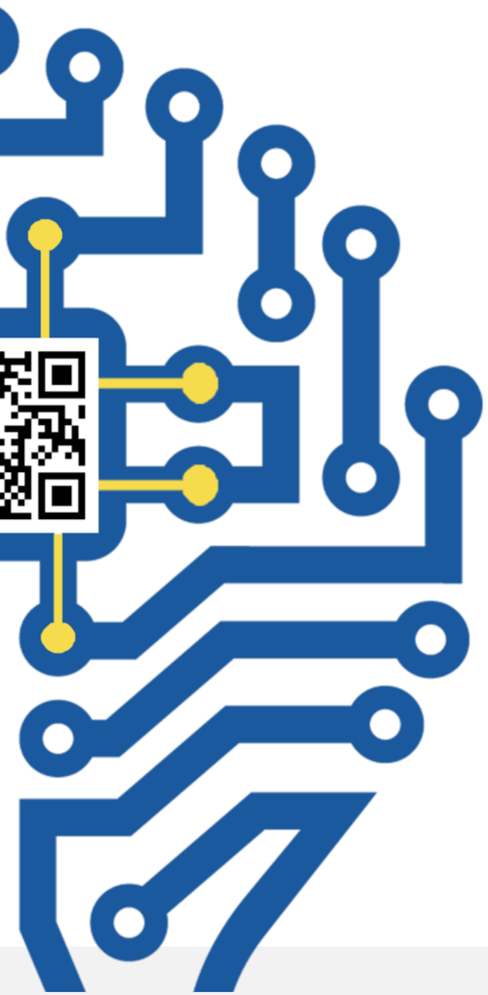
UGANDA CHRISTIAN UNIVERSITY
A Center of Excellence in the Heart of Africa



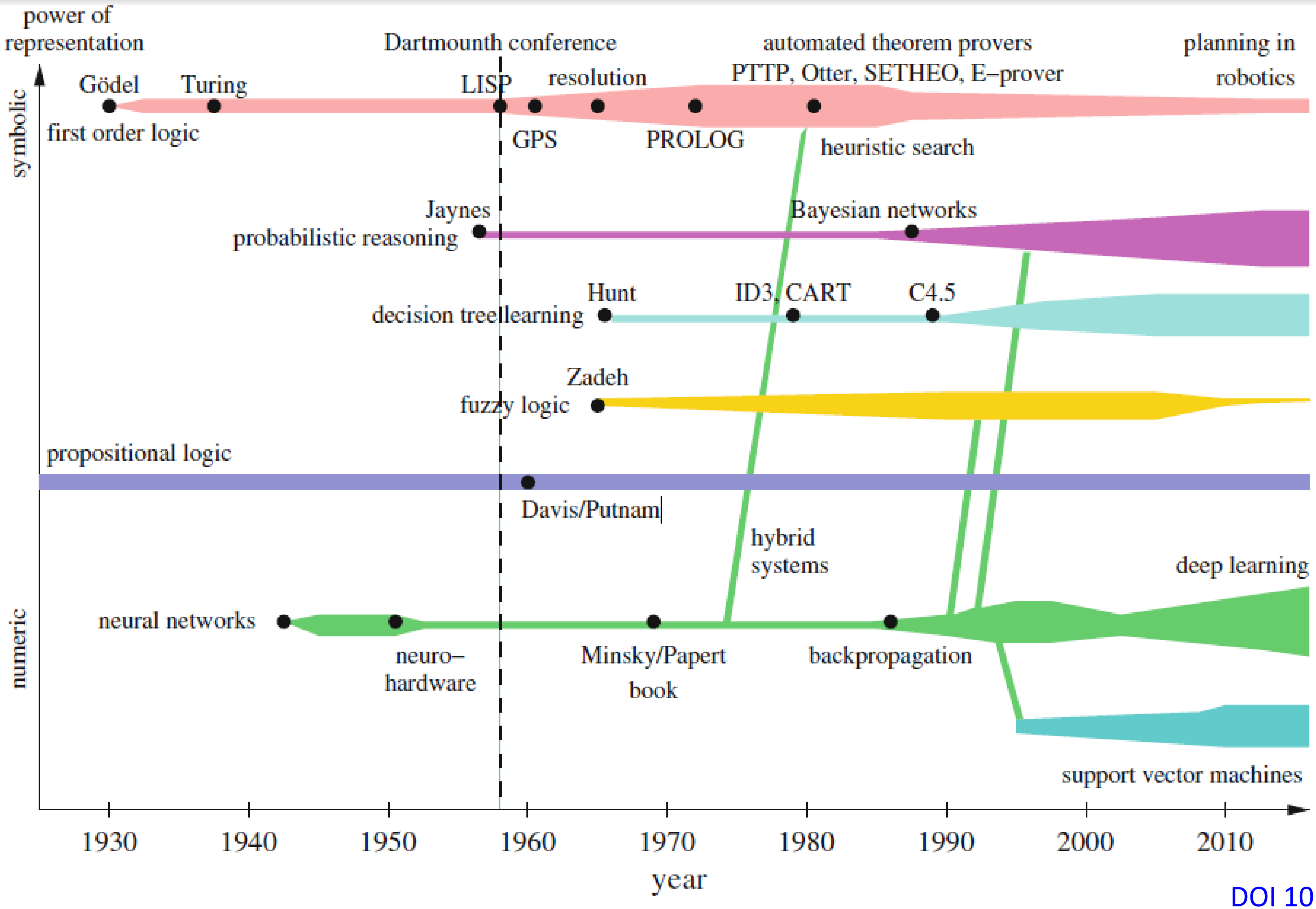
MSDS Data Science Lifecycle

PREDICTIVE MODELLING AND MACHINE LEARNING

Dr. Daphne Nyachaki Bitalo
(PhD Genetics and Bioinformatics)
Department of Computing & Technology
Faculty of Engineering, Design & Technology

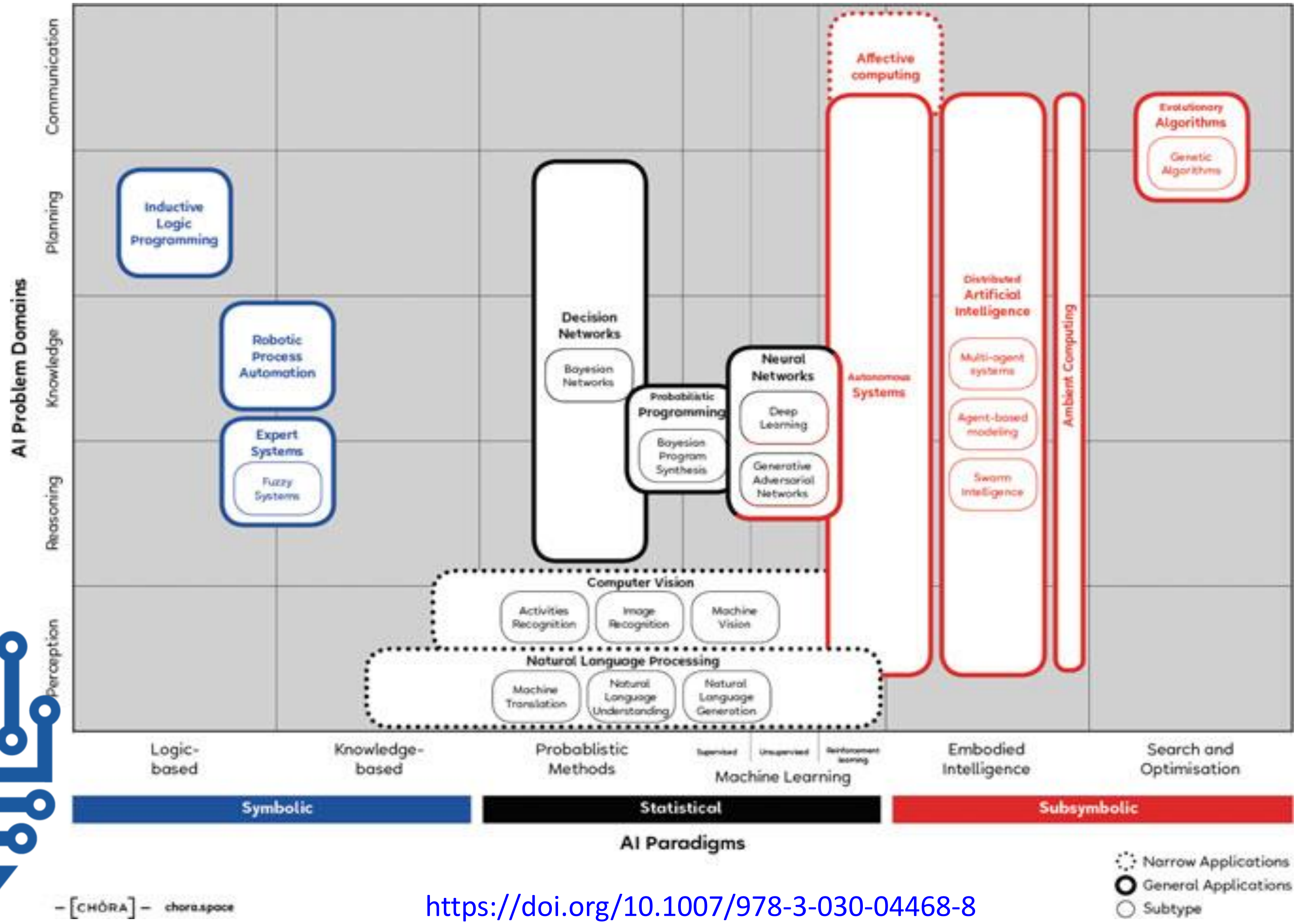
4th March 2023

History Of AI



DOI 10.1007/978-3-319-58487-4

CLASSIFICATION OF AI TECHS



https://doi.org/10.1007/978-3-030-04468-8

– **Logic-based tools**: tools that are used for knowledge representation and problem-solving

– **Knowledge-based tools**: tools based on ontologies and huge databases of notions, information, and rules

– **Probabilistic methods**: tools that allow agents to act in incomplete information scenarios

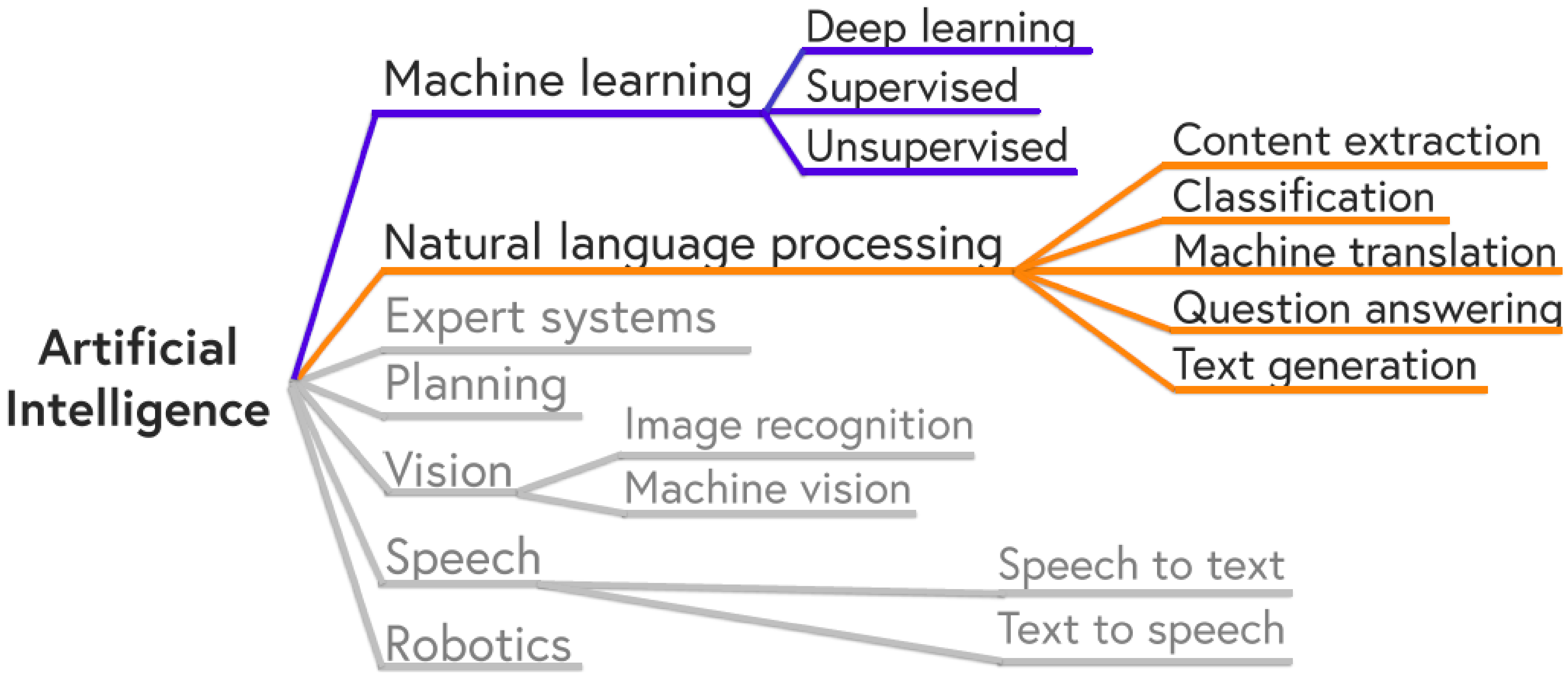
CLASSIFICATION OF AI TECHS

– **Machine learning**: tools that allow computers to learn from data

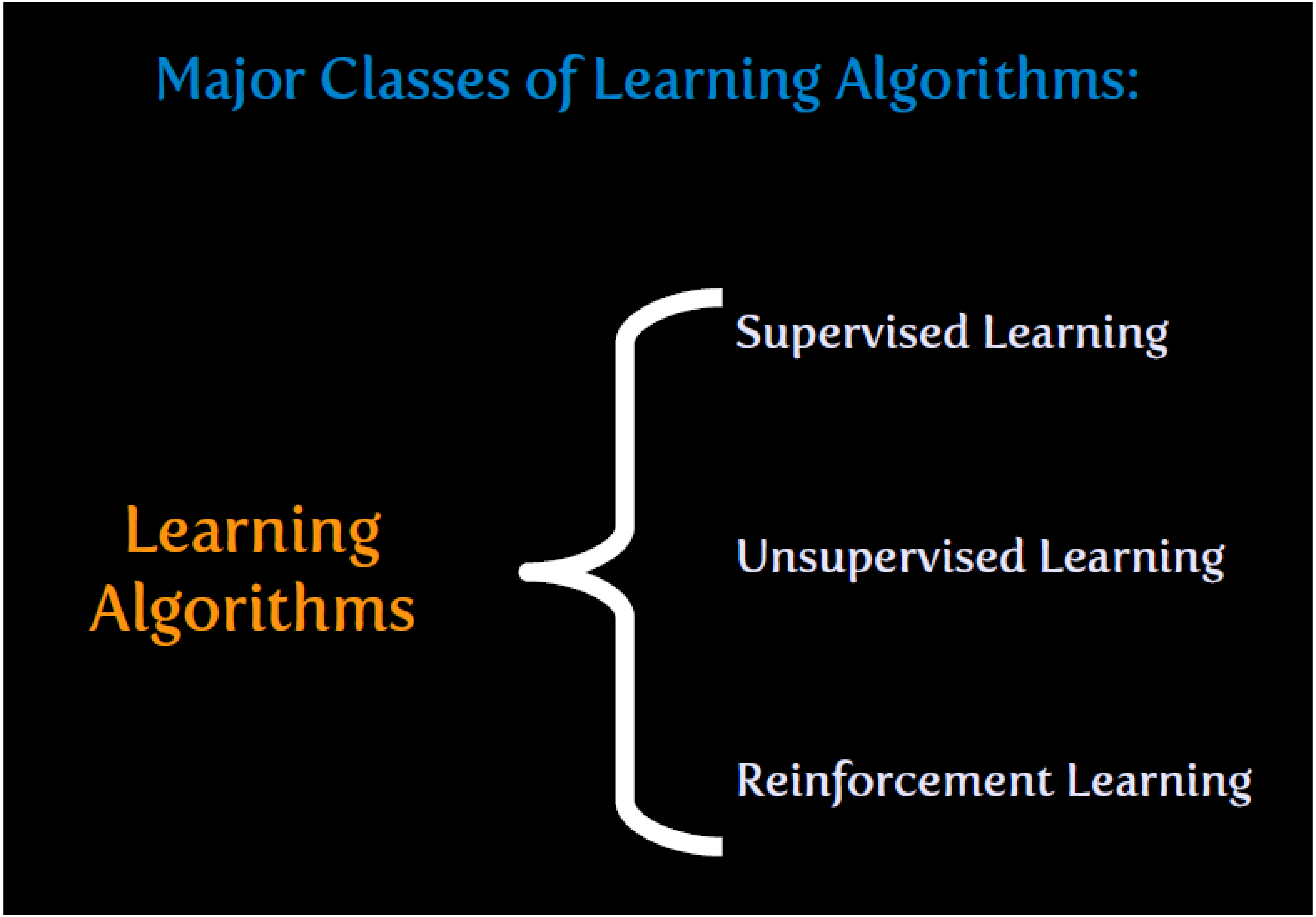
– **Embodied intelligence**: engineering toolbox, which assumes that a body (or atleast a partial set of functions such as movement, perception, interaction, and visualization) is required for higher intelligence

– **Search and optimization**: tools that allow intelligently searching through many possible solutions.

AI Branches



- Artificial Intelligence
  - Machine learning
    - Deep learning
    - Supervised
    - Unsupervised
  - Natural language processing
    - Content extraction
    - Classification
    - Machine translation
    - Question answering
    - Text generation
  - Expert systems
  - Planning
  - Vision
    - Image recognition
    - Machine vision
  - Speech
    - Speech to text
    - Text to speech
  - Robotics



Major Classes of Learning Algorithms:

Learning Algorithms

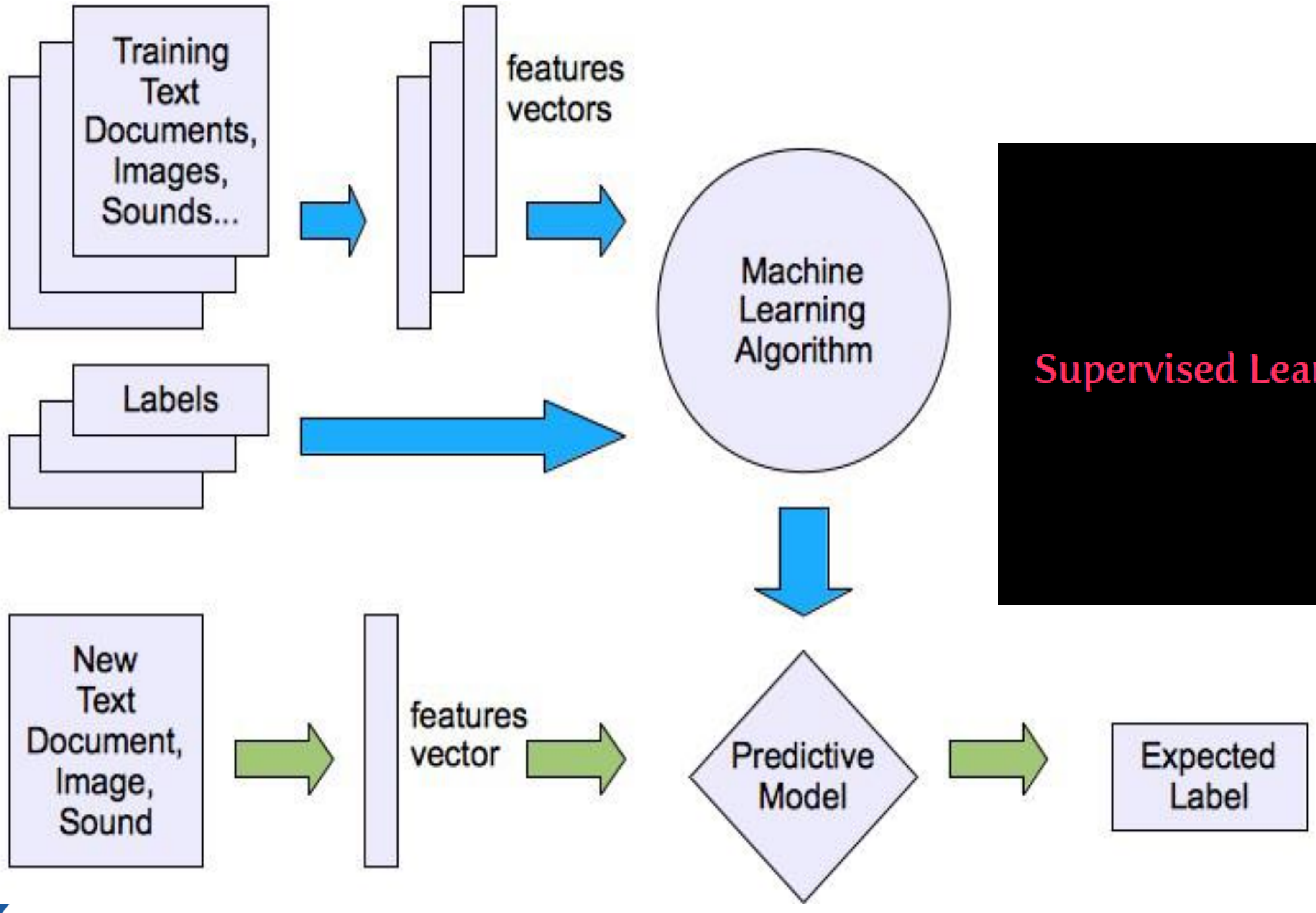
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

- The set of data (training data) consists of a set of input data and correct responses corresponding to every piece of data.
- Based on this training data, the algorithm has to generalize such that it is able to correctly (or with a low margin of error) respond to all possible inputs..

- In essence: The algorithm should produce sensible outputs for inputs that weren't encountered during training.
- Also called learning from exemplars

Supervised Learning


Training Text Documents, Images, Sounds...
features vectors
Labels
Machine Learning Algorithm
New Text Document, Image, Sound
features vector
Predictive Model
Expected Label

Supervised Learning

- Regression Problems
- Classification Problems

Supervised Learning: Regression

- Given some data, you assume that those values come from some sort of function and try to find out what the function is.
- In essence: You try to fit a mathematical function that describes a curve, such that the curve passes as close as possible to all the data points.
- So, regression is essentially a problem of function approximation or interpolation

Supervised Learning: Classification

❑ Consists of taking input vectors and deciding which of the N classes they belong to, based on training from exemplars of each class.

- Is discrete (most of the time). i.e. an example belongs to precisely one class, and the set of classes covers the whole possible output space.

❑How it's done: Find 'decision boundaries' that can be used to separate out the different classes.

- Given the features that are used as inputs to the classifier, we need to identify some values of those features that will enable us to decide which class the current input belongs to

Unsupervised Learning

❑ - Conceptually Different Problem.

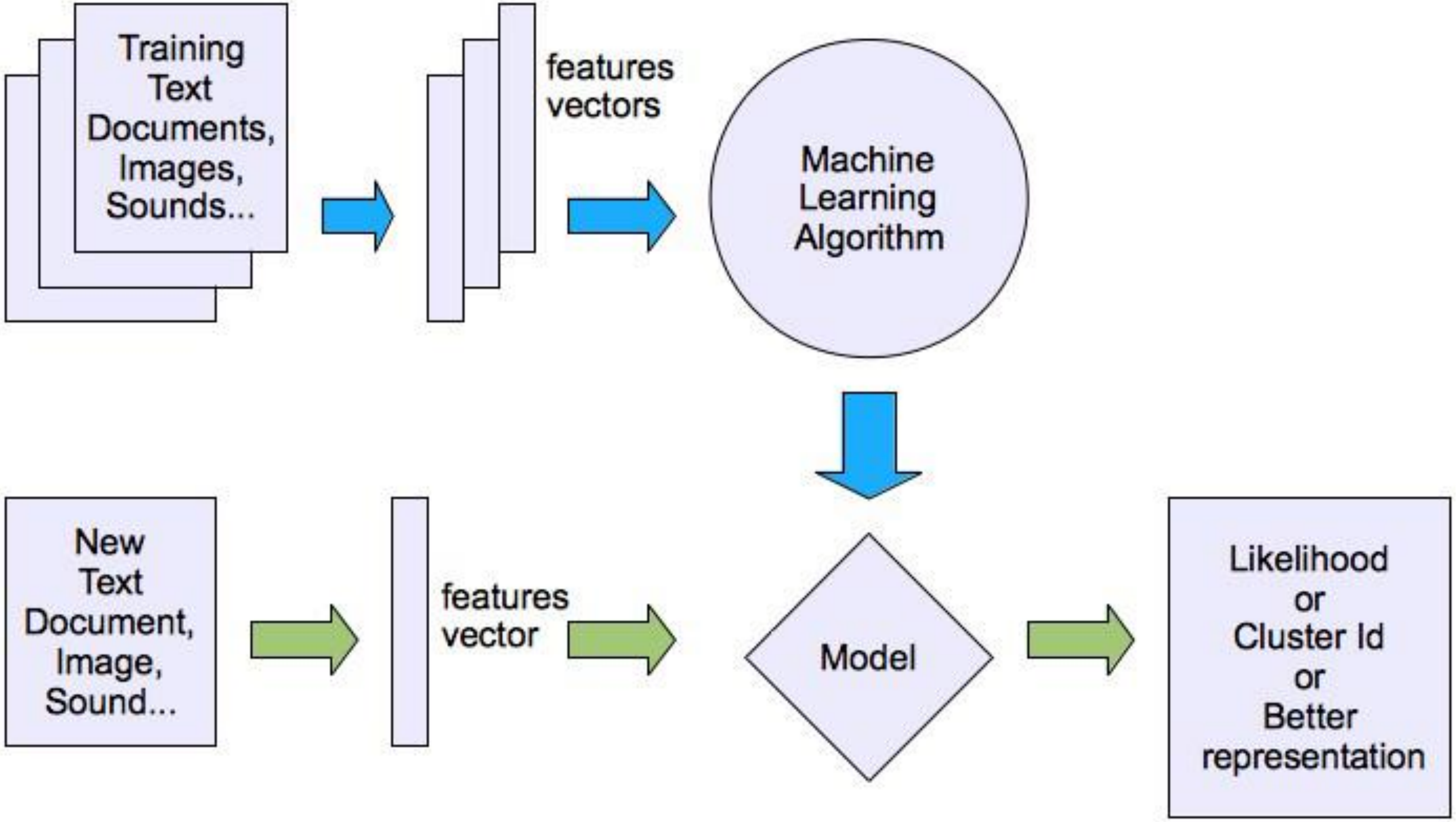
❑- No information about correct outputs are available.

❑- No Regression No guesses about the function can be made

❑-Classification? No information about the correct classes. But if we design our algorithm so that it exploits similarities between inputs so as to cluster inputs that are similar together, this might perform classification automatically

❑In essence: The aim of unsupervised learning is to find clusters of similar inputs in the data without being explicitly told that some datapoints belong to one class and the other in other classes. The algorithm has to discover this similarity by itself

Unsupervised Learning


Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
New Text Document, Image, Sound...
features vector
Model
Likelihood or Cluster Id or Better representation

Reinforcement Learning

- Stands in the middle ground between supervised and unsupervised learning.
- The algorithm is provided information about whether or not the answer is correct but not how to improve it
- The reinforcement learner has to try out different strategies and see which works best
- In essence: The algorithm searches over the state space of possible inputs and outputs in order to maximize a reward

UGANDA CHRISTIAN UNIVERSITY

A Center of Excellence in the Heart of Africa

ML IN ACTION

SUPERVISED LEARNING

(Programme MSDS DATA SCIENCE)
Email: dbitalo@ucu.ac.ug

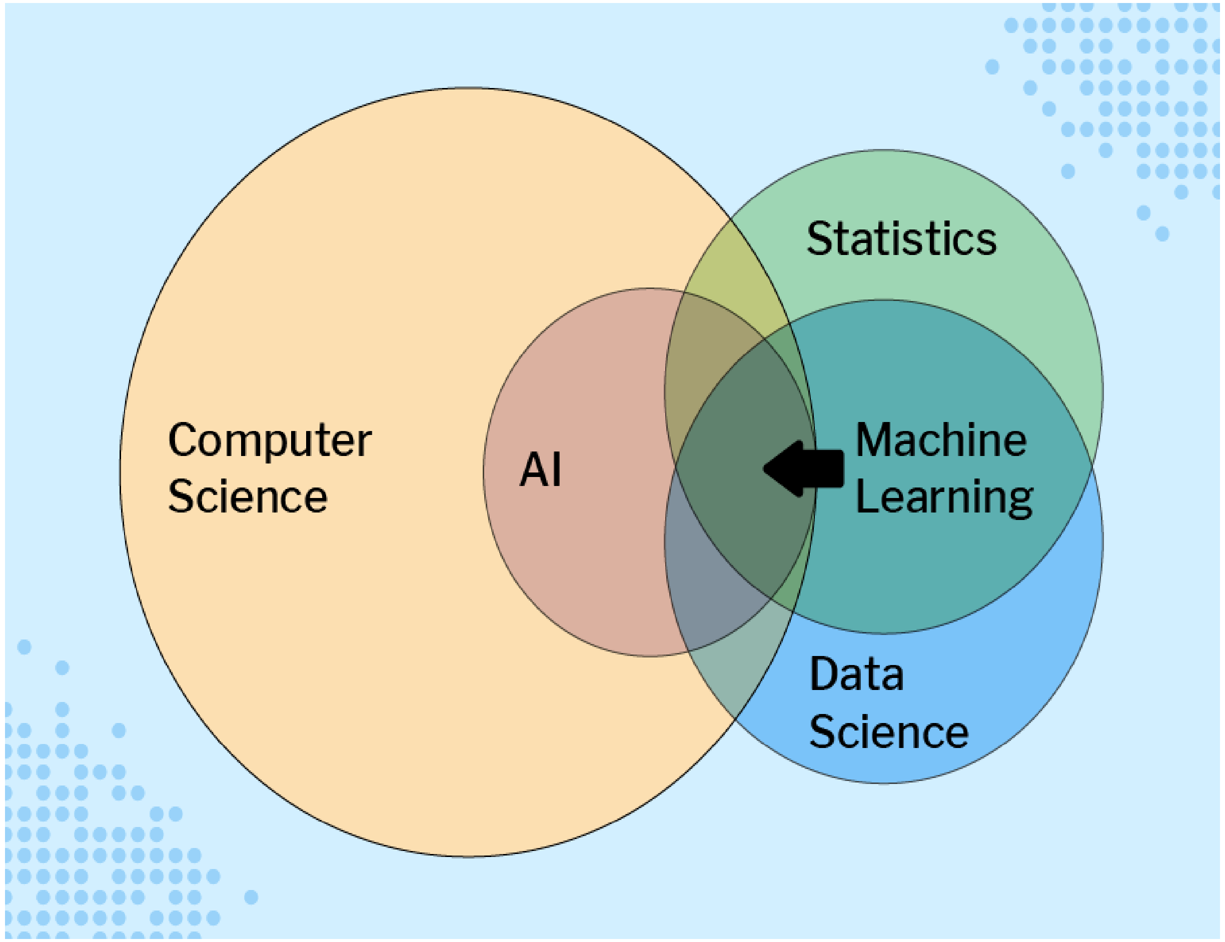
Dr. Daphne Nyachaki Bitalo
(PhD Genetics and Bioinformatics)
Department of Computing & Technology
Faculty of Engineering, Design & Technology

AI PATTERN OF APPROACH

1. Define a problem that requires "intelligence"
2. Work towards and achieve (either in full or close enough) a solution to the problem
3. Change the definition of "intelligent"
4. Repeat

Historically "intelligence" defined using the Turing Test in the 1950 article "Computing machinery and Intelligence"
AI in medical field example: Have a go with the Buoy chatbot here.

AI PATTERN OF APPROACH



Interdisciplinary = Multidisciplinary

Stats: Probability and regression

Data Science: Data preparation and exploration

ML: Trains data to generate AI algorithm

AI EXAMPLE: NAVIGATION

1 SAMUEL 7:12

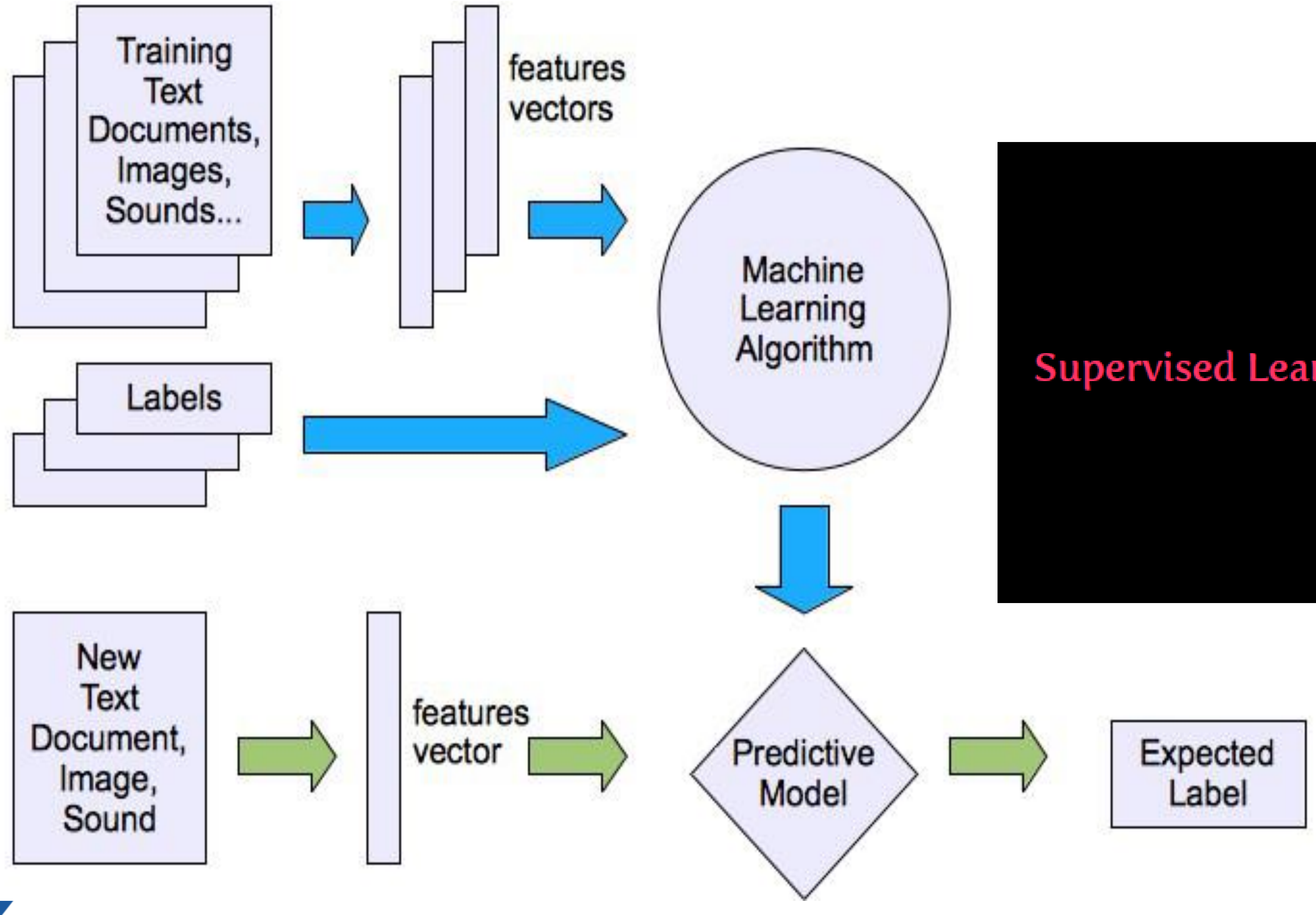
Most people would not think of a satellite navigation system as an example of AI, but it is an intelligent program.

COMPUTER SCIENCE: A navigation system is built on the products of computer science, everything from the satellites that track the system to the phone or other device that displays the maps.

STATISTICS AND DATA SCIENCE: Principles that were integral in creating the map and the program use road data to plan a route. Statistical techniques help aggregate the travel times and detours other passengers are taking to better plan your route. Data scientists will have helped clean and pre-process the data to remove unnecessary information. For example, when planning a route, the algorithm will disregard the number of lanes on a road, as that should not affect the route (although traffic conditions on the road would).

MACHINE LEARNING: Algorithms help a navigation system improve over time. As more routes are planned, the system will feed data back in to help the algorithms plan better routes in the future. For example, this data might include the actual travel time taken to reach a destination, or the speed you travelled on each road.

ML: Supervised Learning



Training Text Documents, Images, Sounds...

features vectors

Labels

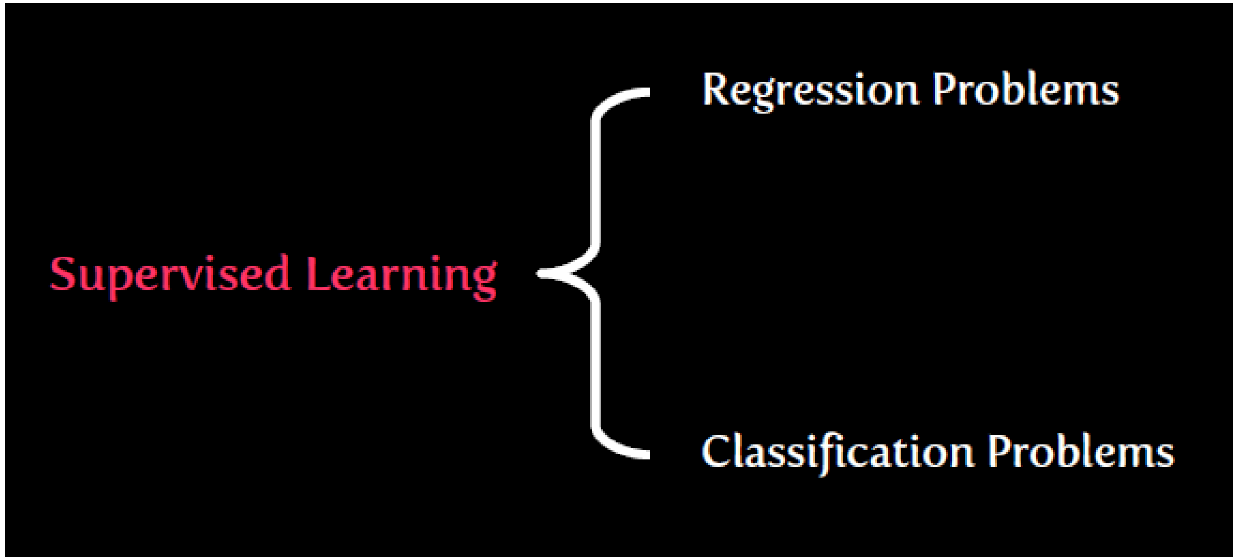
Machine Learning Algorithm

New Text Document, Image, Sound

features vector

Predictive Model

Expected Label



Supervised Learning

Regression Problems

Classification Problems

Eamples of SL using Regression

❑Regression is essentially a problem of function approximation or interpolation

❑Suited to continuous data e.g. temperature, age

❑Regression uses data from the past to attempt to define the relationship between the available data (the inputs) and the value you are trying to predict (output)

❑The output is called the dependent variable and the inputs are called independent variables.

❑For instance the “Arrythymia” dataset, others are independent variables and heart rate is the dependent variable

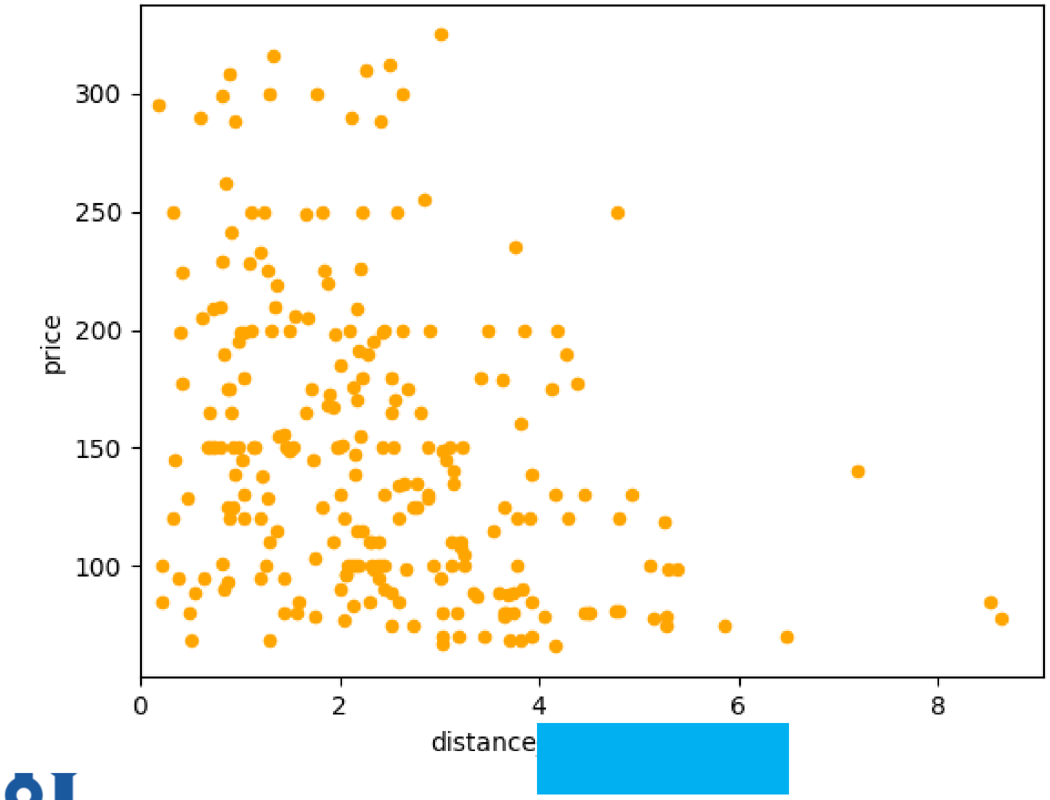
Eamples of SL using Regression

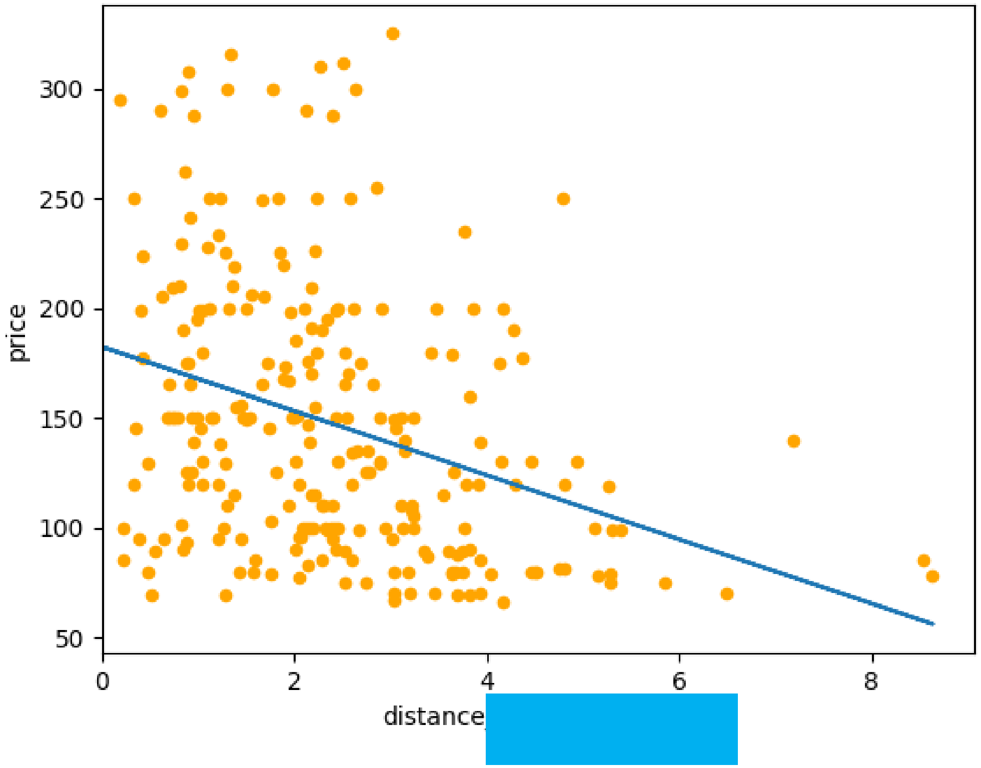
❑Imagine you work for AirBnB and a new host comes to you with an apartment in Jinja. They would like some advice about what price (per night) they should set their rental at.

❑So you decide that you will use the distance to one of the most popular destinations – the Source of the Nile– to analyse the current rentals and estimate a price for the new host.

❑First you clean up the data of all price-listings under AirBnB and get a visual mapping of them with price as the dependent variable and distance to Source of the Nile as the independent variable

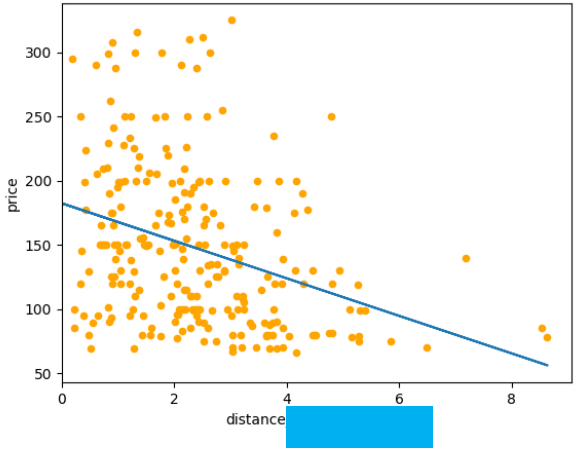
Eamples of SL using Regression





❑Draw a line of “best fit” to see the relationship between the two variables i.e. linear regression

Eamples of SL using Regression



- ❑Equation for the line is y = mx + c
  - ❑y = predicted value
  - ❑m = weight of the input variable
  - ❑x = input variable (distance)
  - ❑c = y-intercept (value the model would predict if input was 0)
- ❑Results shows coffieicent = –14.61, y-intercept = 182.36
- ❑Interpretation: A property that is 0 miles from the Source of the Nile should be rented for $182.36 and for every mile you travel away from the source, the rental drops by $14.61.

Eamples of SL using Regression

❑However, the line of “best fit” can’t pass through all data points

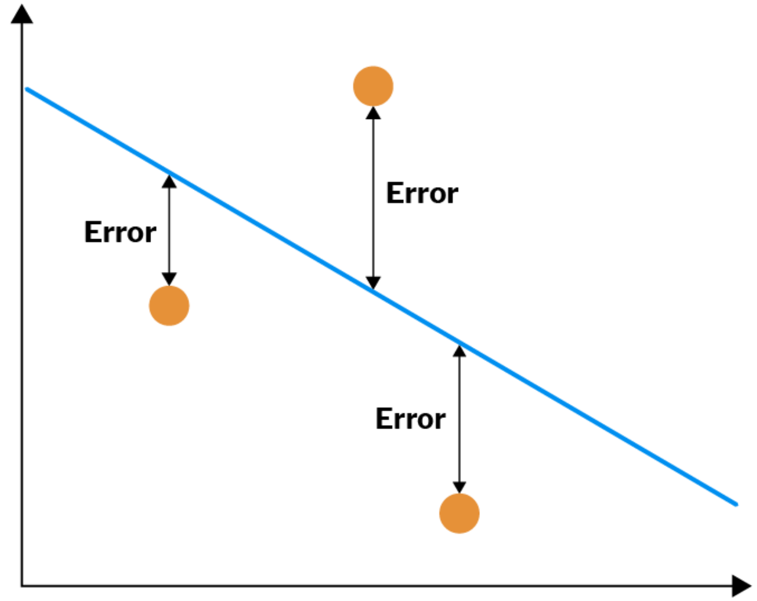
❑Must reduce weight of “errors”

❑Check beforehand if data is normally distributed/remove outliers etc

❑Errors increase with multiple input variables

❑E.g. type of housing, number of rooms etc

❑Regression ML reduces weight of errors



❑Machine learning (SL) is used to “learn” by adjusting weights to reduce the error as much as possible, to get to the optimal weights

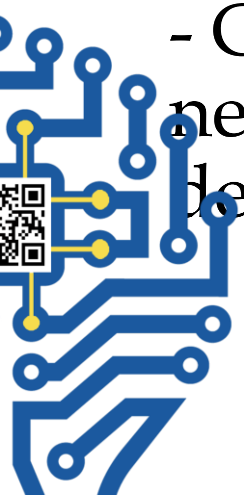
Supervised Learning: Classification

□ Supplying data to a computer for it to then allocate it to a label or a class.

□Use discrete/ categorical data (most of the time). i.e. an example belongs to precisely one class, and the set of classes covers the whole possible output space.

□How it's done: Find 'decision boundaries' that can be used to separate out the different classes.

- Given the features that are used as inputs to the classifier, we need to identify some values of those features that will enable us to decide which class the current input belongs to

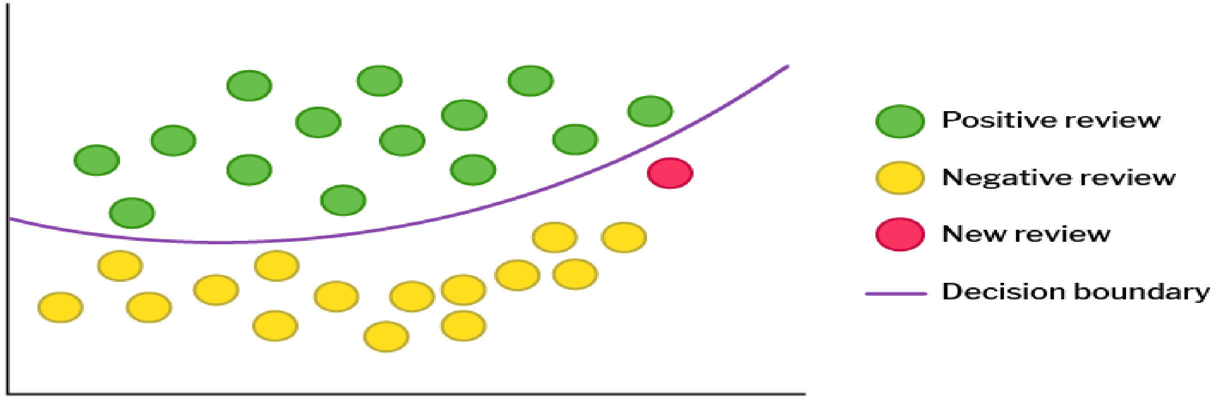


Examples of SL using Classification

❑ For instance, customers are reviewing service provision for an online delivery company

❑Some of the reviews are categorized as “positive” and others “negative” based on the surveys run

❑When a new user submits a review, the ML model must determine if it’s positive or negative



Examples of SL using Classification

1 SAMUEL 7:12

❑ Classifications can be binary (as in case of positive or negative)

❑Classifications multi-class or multi-label

❑E.g. Classifying images (facial recognition) which have multiple aspects

❑An example is the AI app that identifies bird species based on their sounds

❑try this out for yourself on the BirdNet project website

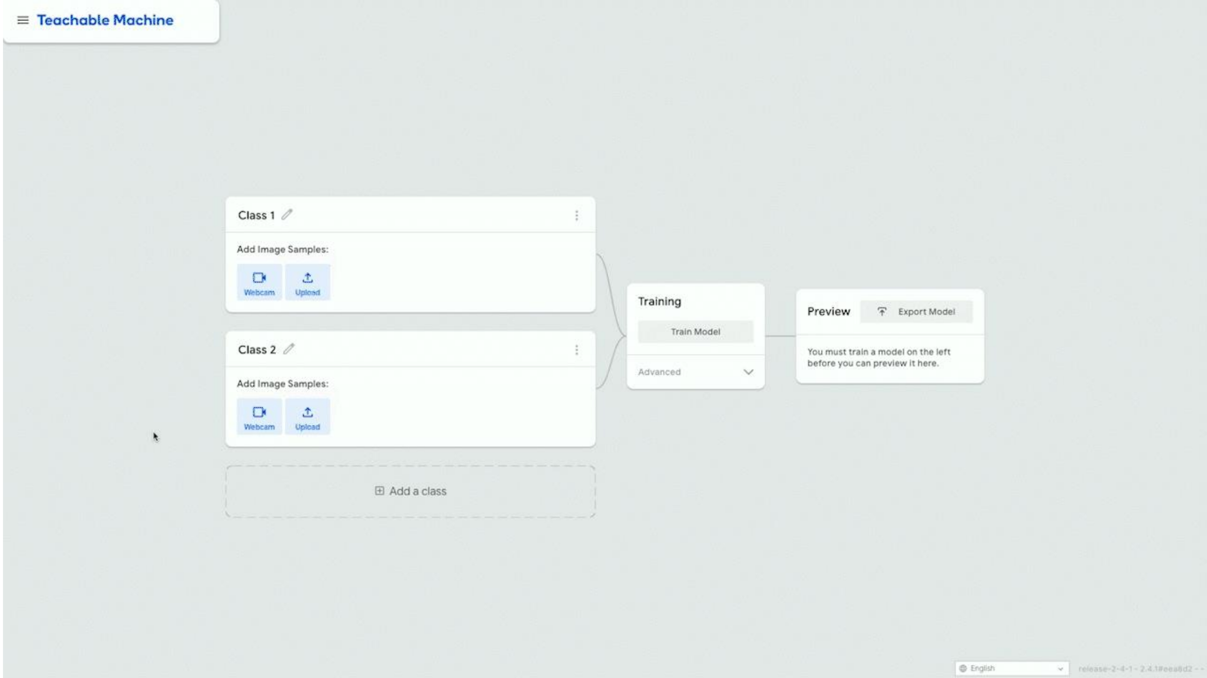
❑This app has limitations in generating a sufficient amount of sound “labels” or categories.

❑So the training data has to be extensive and correctly labelled

Examples of SL using Classification

1 SAMUEL 7:12

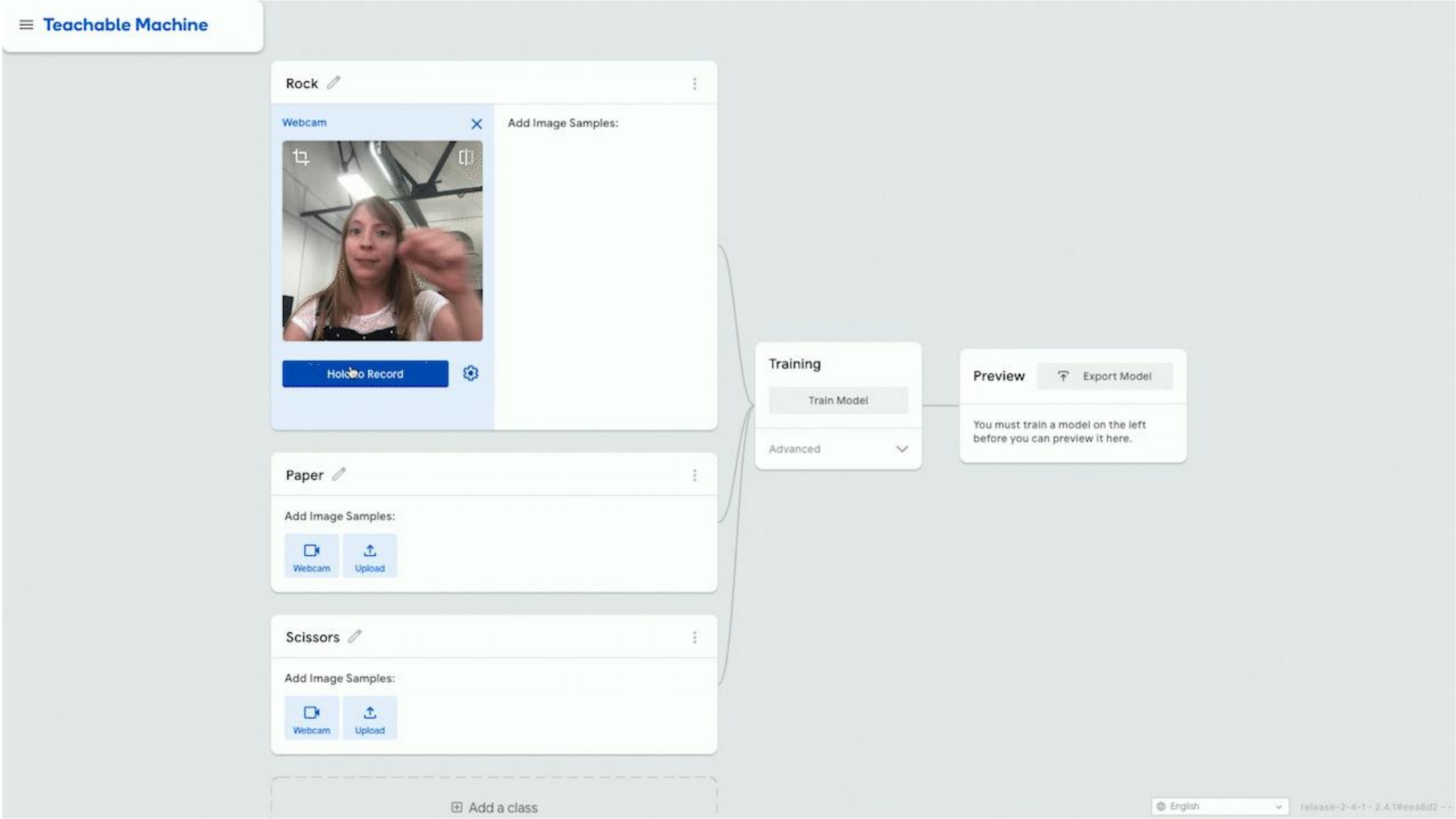
- A game of rock-paper-scissors can also be generated as a classification ML model
- Set up the project. E.g.
- Visit Teachable Machine. Select Image Project.
- Set/define your classes by uploading images (rock, paper, scissors) using your hand gestures



You must train a model on the left before you can preview it here.

Examples of SL using Classification

- ❑ Add image samples of rock-paper-scissors using hand gestures
- ❑ Train the model

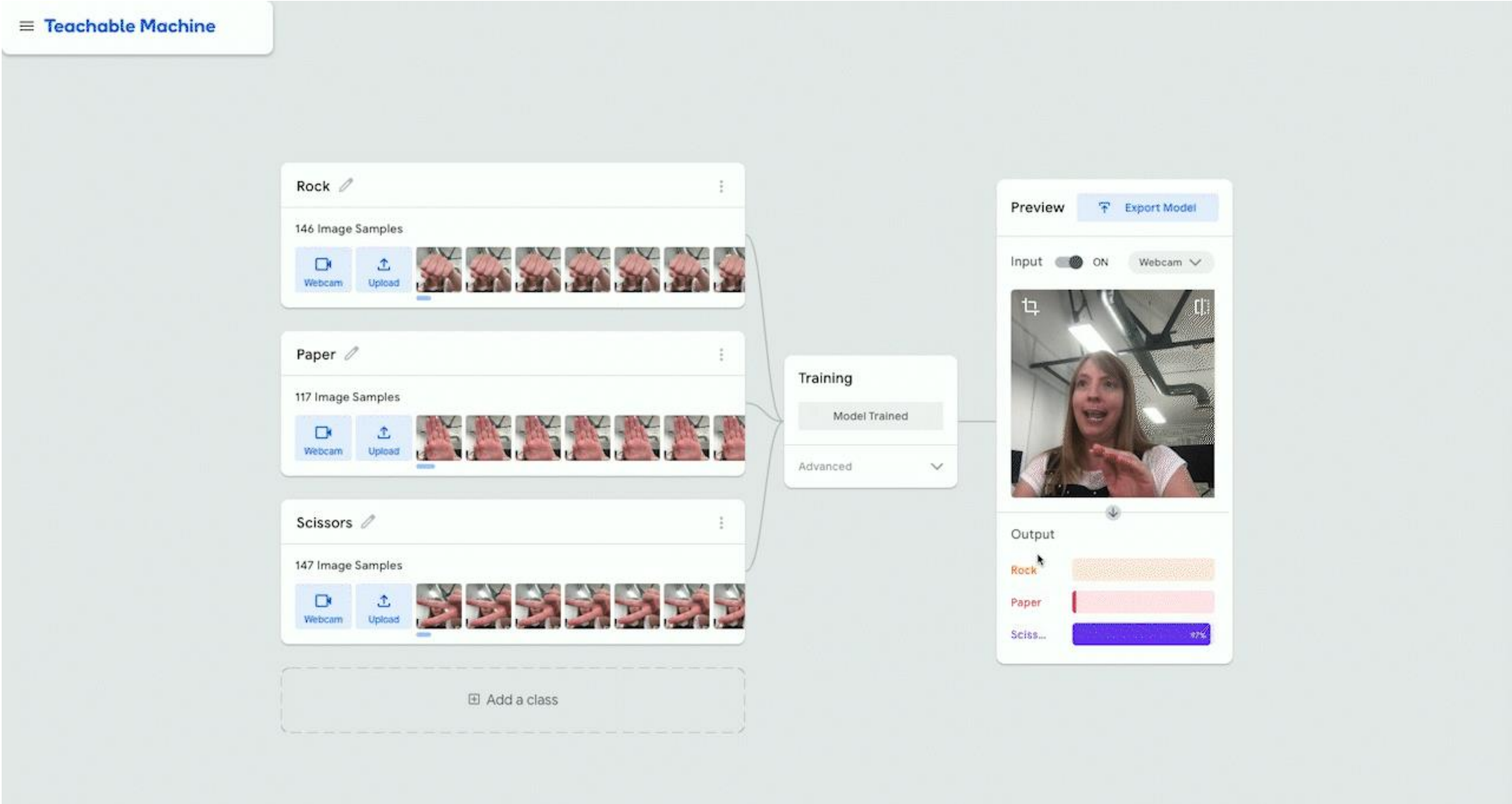


You must train a model on the left before you can preview it here.

Examples of SL using Classification

1 SAMUEL 7:12

- ❑Test the model to see if it correctly classifies your hand gestures
- ❑Save the model



Rock

146 Image Samples

Paper

117 Image Samples

Scissors

147 Image Samples

Training

Model Trained

Advanced

Preview

Input ON

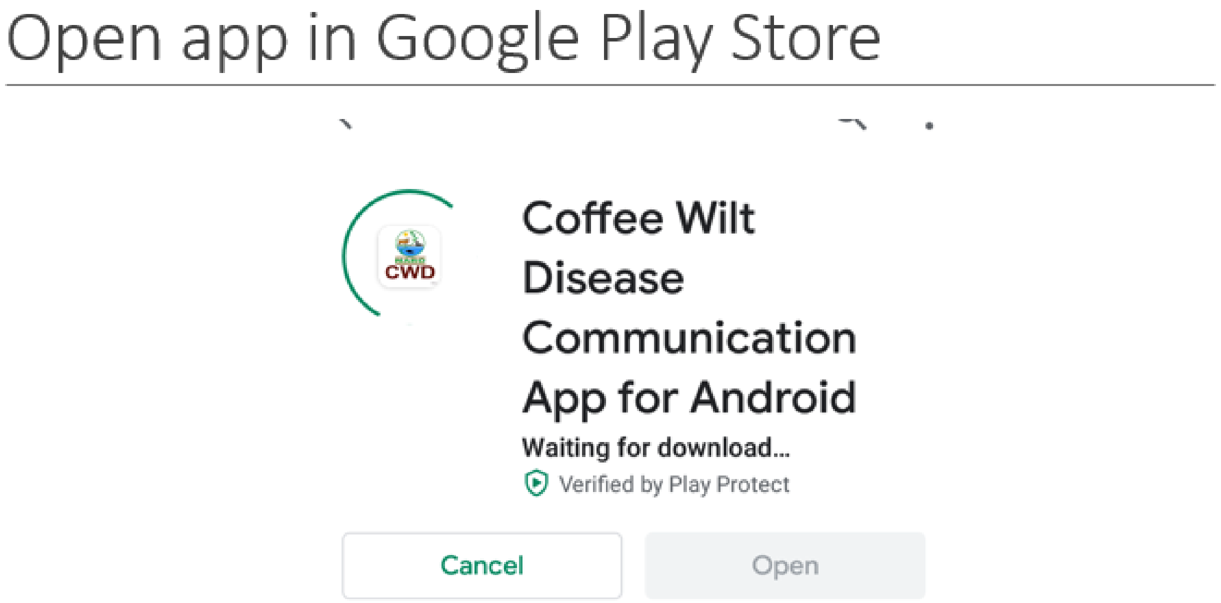
Rock

Paper

Sciss...

Ugandan example of SL Classification

- ❑Generated an app for detection of Coffee Wilt Disease for use by Agricultural Extension Officers and coffee farmers
- ❑Uses a Supervised Learning-Classification model
- ❑Shortcomings of limited training data since there's need for use of Drones to capture several images of coffee plants (still in beta phase)

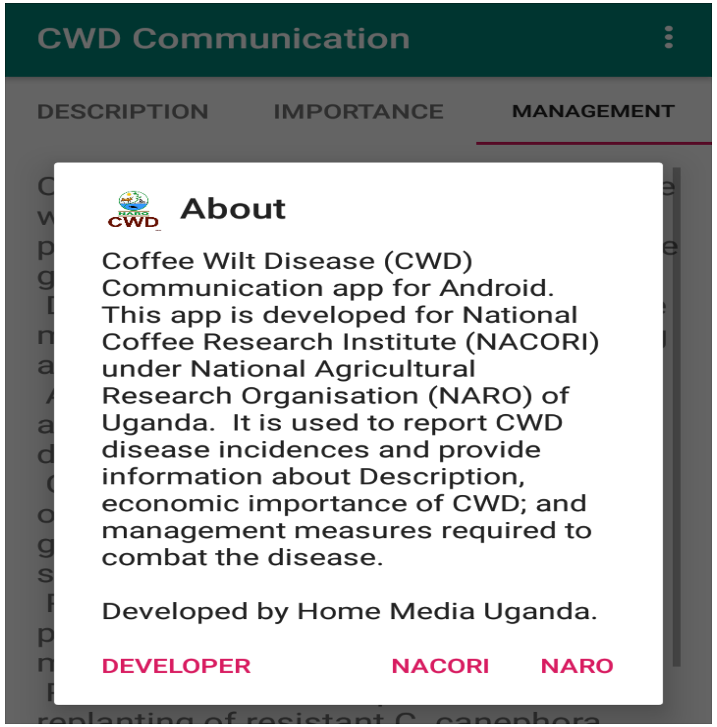


Open app in Google Play Store

Coffee Wilt Disease Communication App for Android

Waiting for download...

Verified by Play Protect



CWD Communication

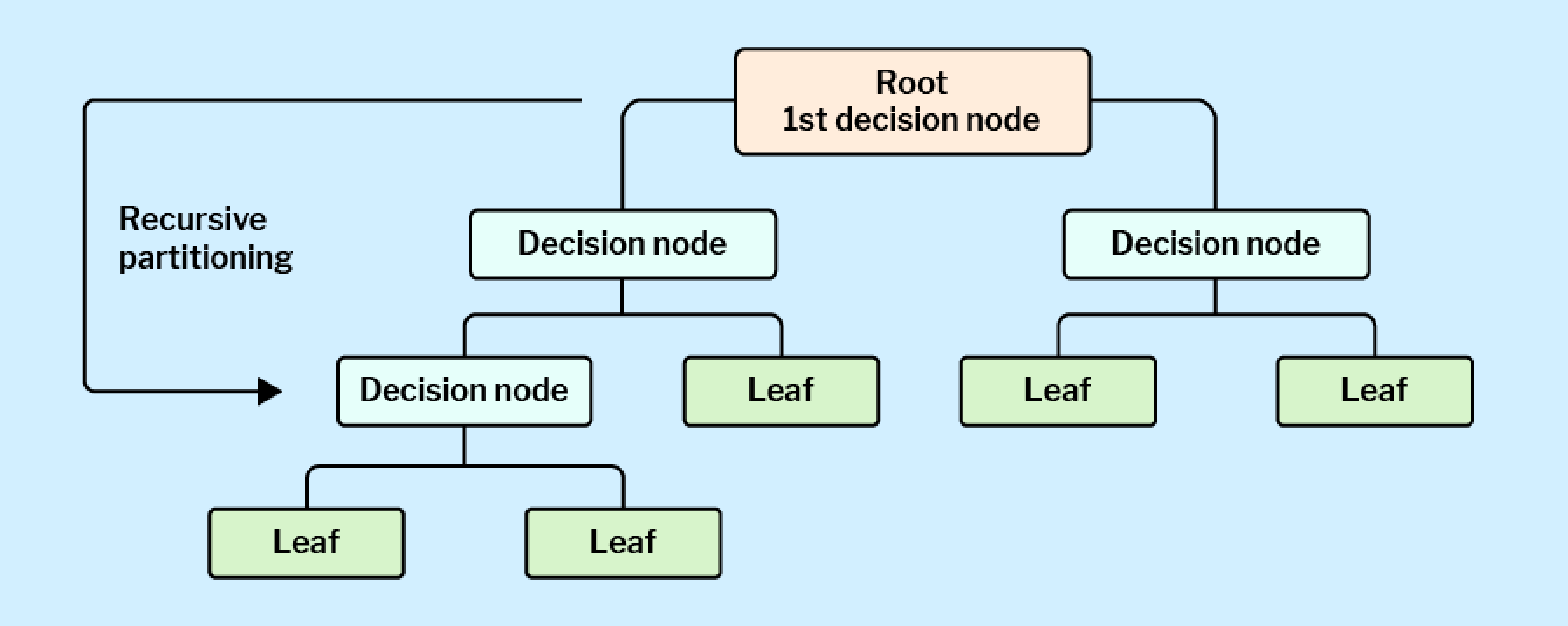
About

Coffee Wilt Disease (CWD) Communication app for Android. This app is developed for National Coffee Research Institute (NACORI) under National Agricultural Research Organisation (NARO) of Uganda. It is used to report CWD disease incidences and provide information about Description, economic importance of CWD; and management measures required to combat the disease.

Developed by Home Media Uganda.

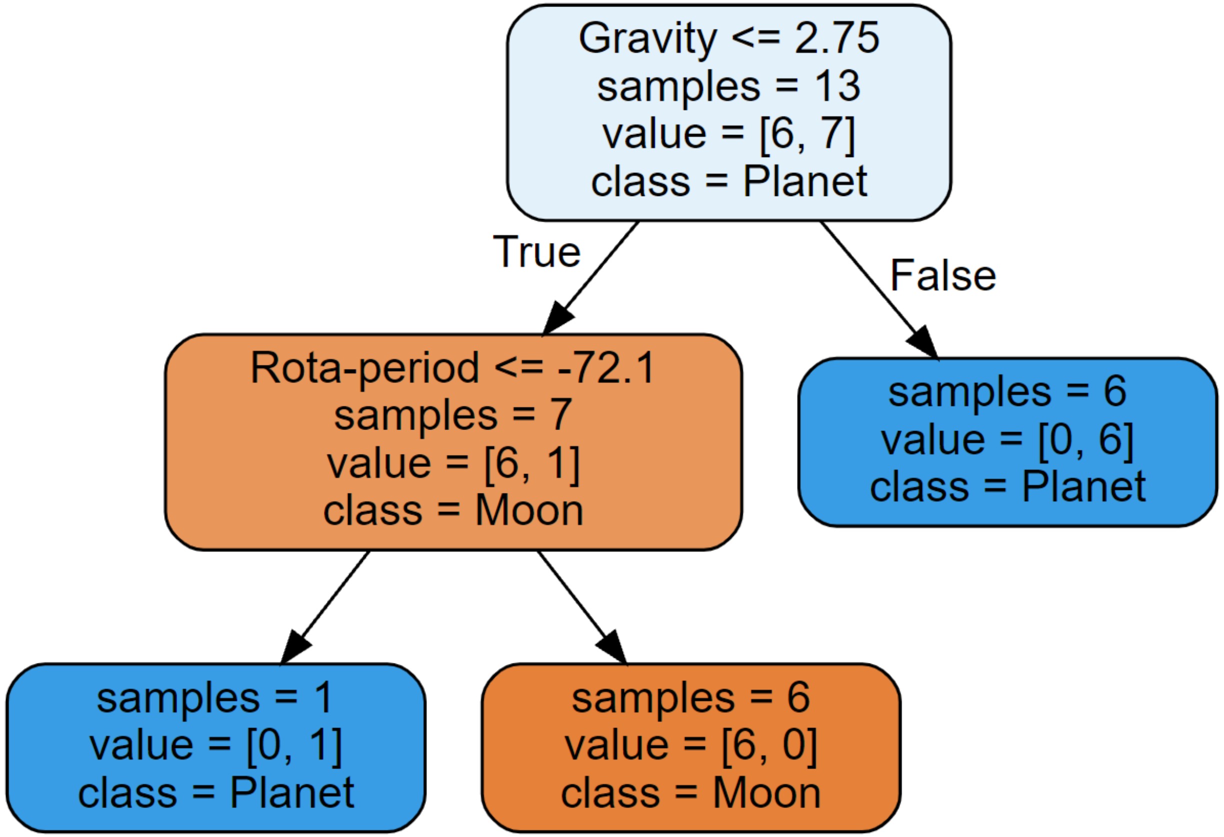
How does the machine learn?
Supervised Learning Algorithms

- ❑ Algorithms for classification and regression problems are available mostly in existent libraries like scikit-learn
- ❑Decision trees
- ❑k-Nearest neighbour (kNN)



Decision trees

- ❑ E.g. predicting whether a visualized celestial body is a moon or planet (Classification problem)
- ❑Data taken from NASA’s planetery factsheet
- ❑13 samples (samples = 13)
- ❑6 are moons and 7 are planets ([6,7]
- ❑Split on Gravity and Rota-period
- ❑An addition of more parameters would generate a complex tree



KNN Algorithm

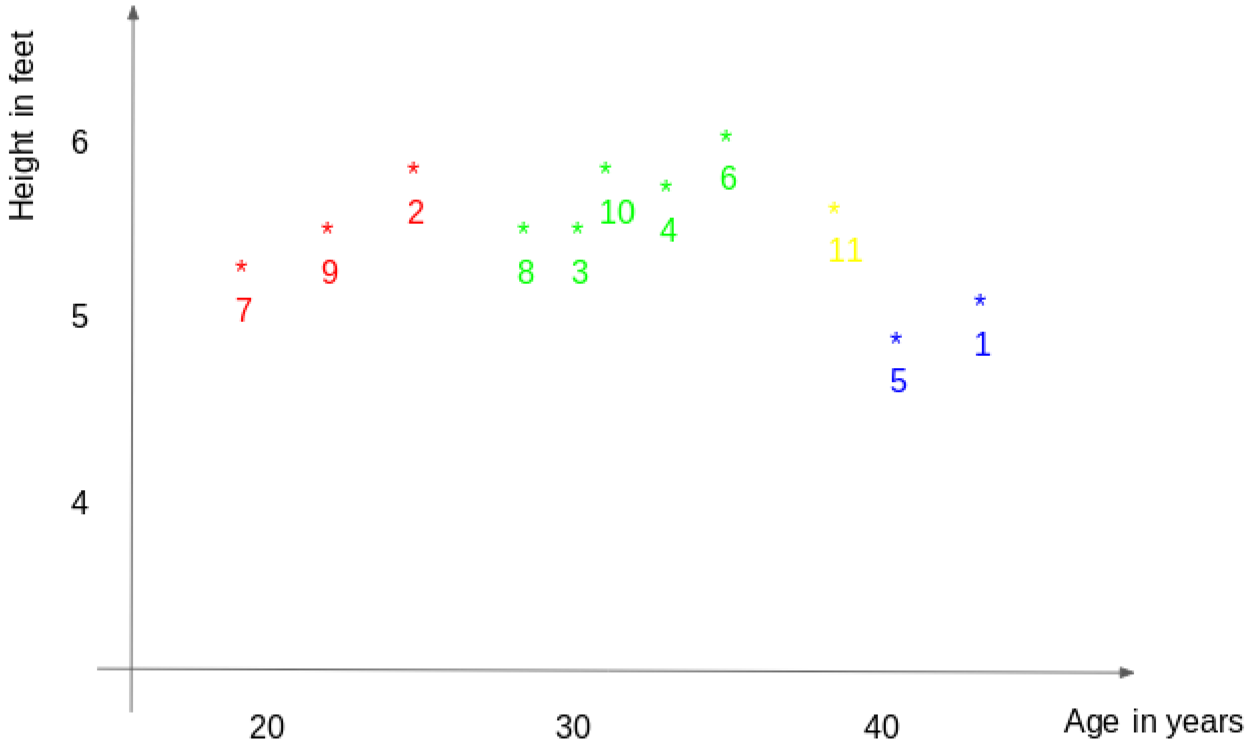
- ❑Is a non-parametric approach (not like linear regression)
- ❑The k in KNN is a variable that is used to determine how many neighbours should be used to make the prediction
- ❑This algorithm applies a distance function to predict the number of neighbours
  - ❑The most commonly the Euclidean distance measure. (read about Euclidean distance on Wikipedia) for continuous data
  - ❑Hamming distance for categorical data
- ❑Applied to both Classification and Regression SL problems
- ❑kNN better suited to our "AirBnB pricing" problem previously discussed

How KNN Works

❑ Predict the weight of sample "11" who is 38 yrs and 5.5 feet tall

❑The assumption would be closest neighbours are 5 and 1 because of age.

ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?



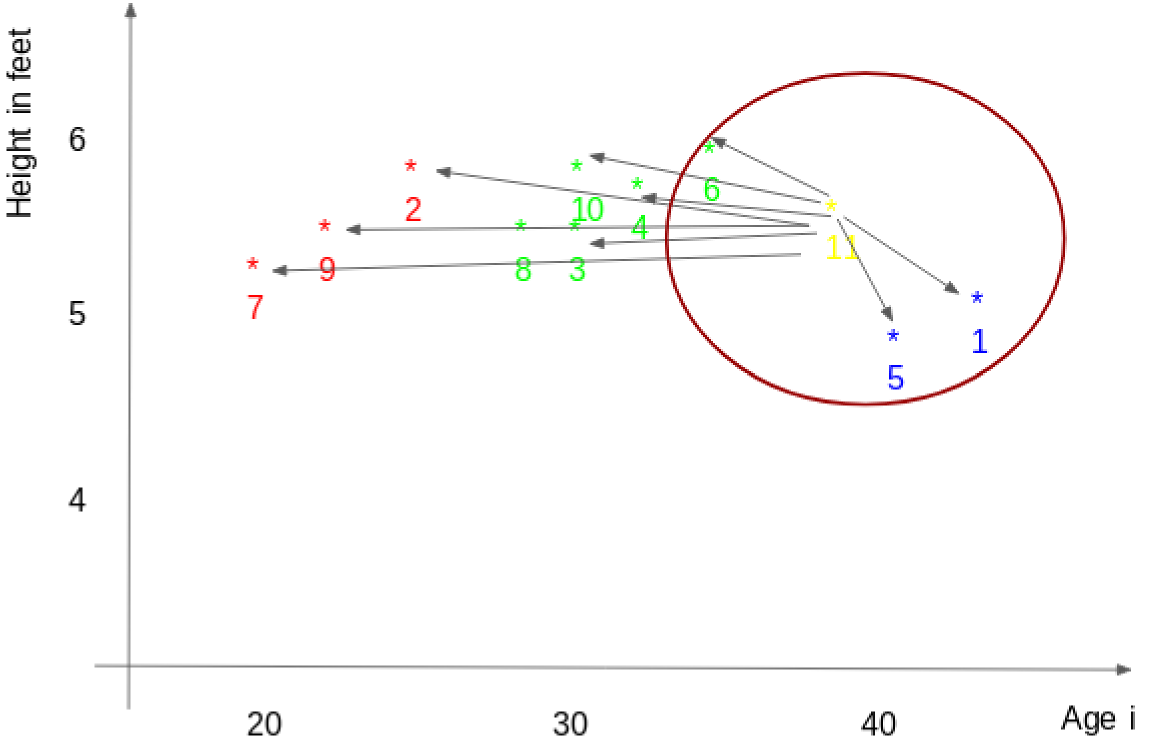
How KNN Works

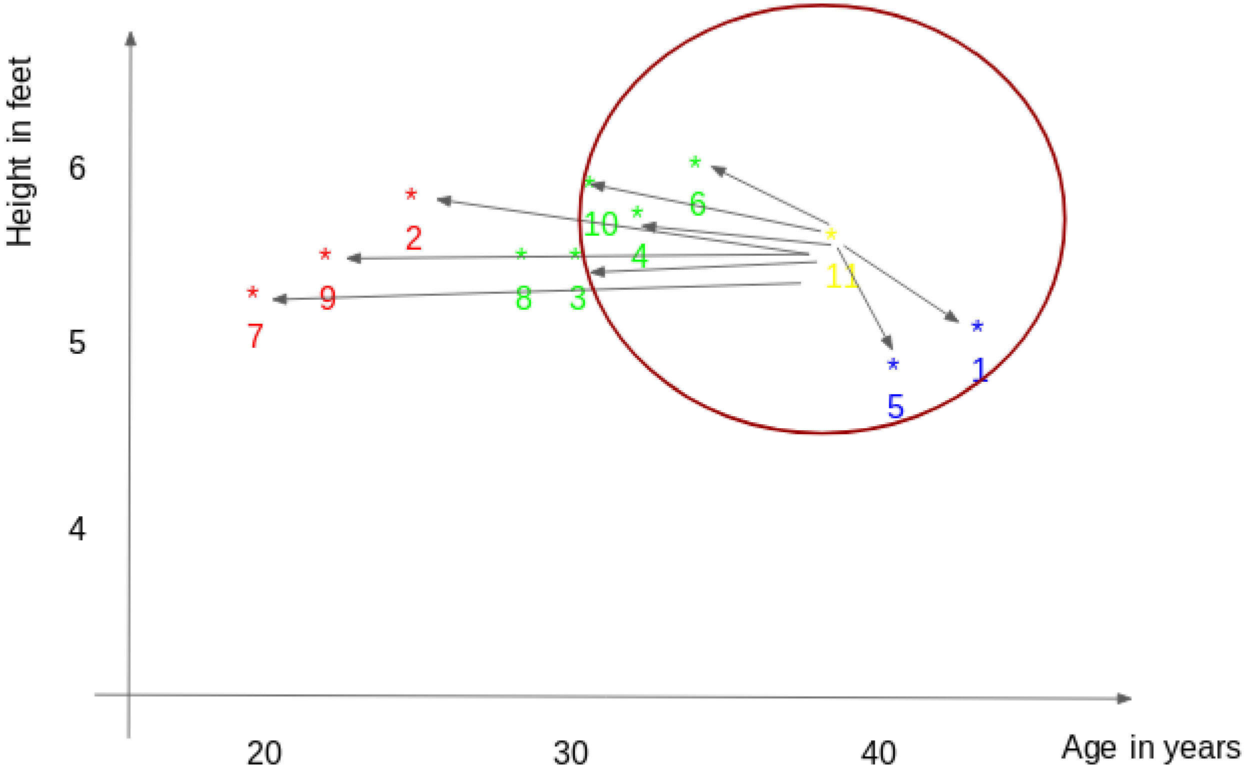
❑ Distance between the test point and other training points is measured

❑The closest neigbours are chosen, k = 3 or even k = 5

❑Predicted weight is either ID11 = (77+72+60)/3 or

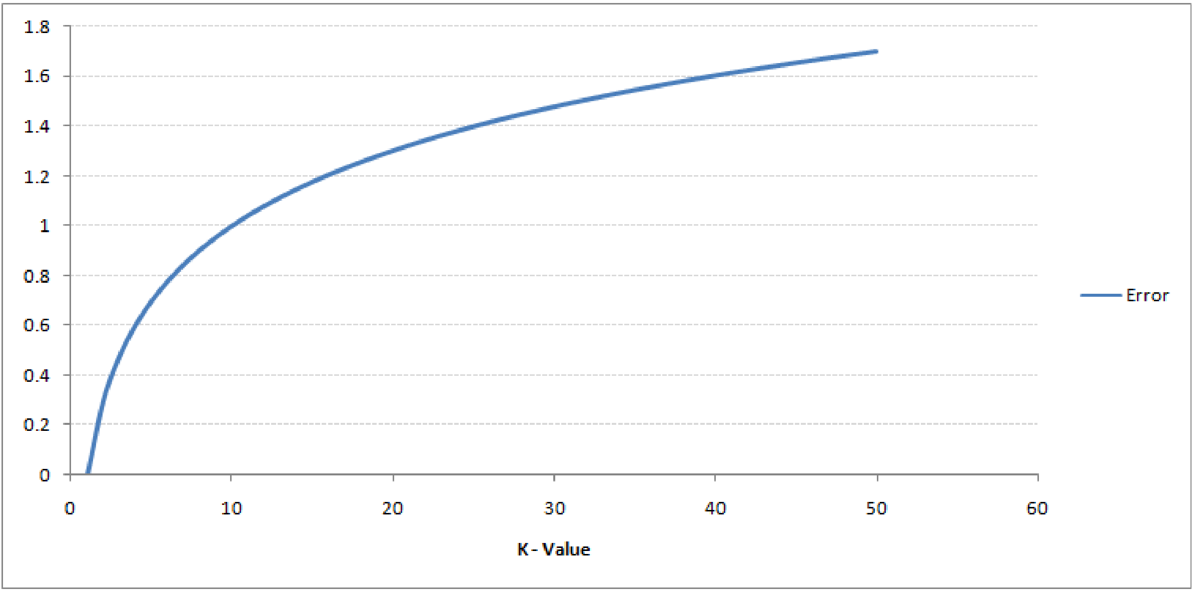
❑ID 11 = (77+59+72+60+58)/5

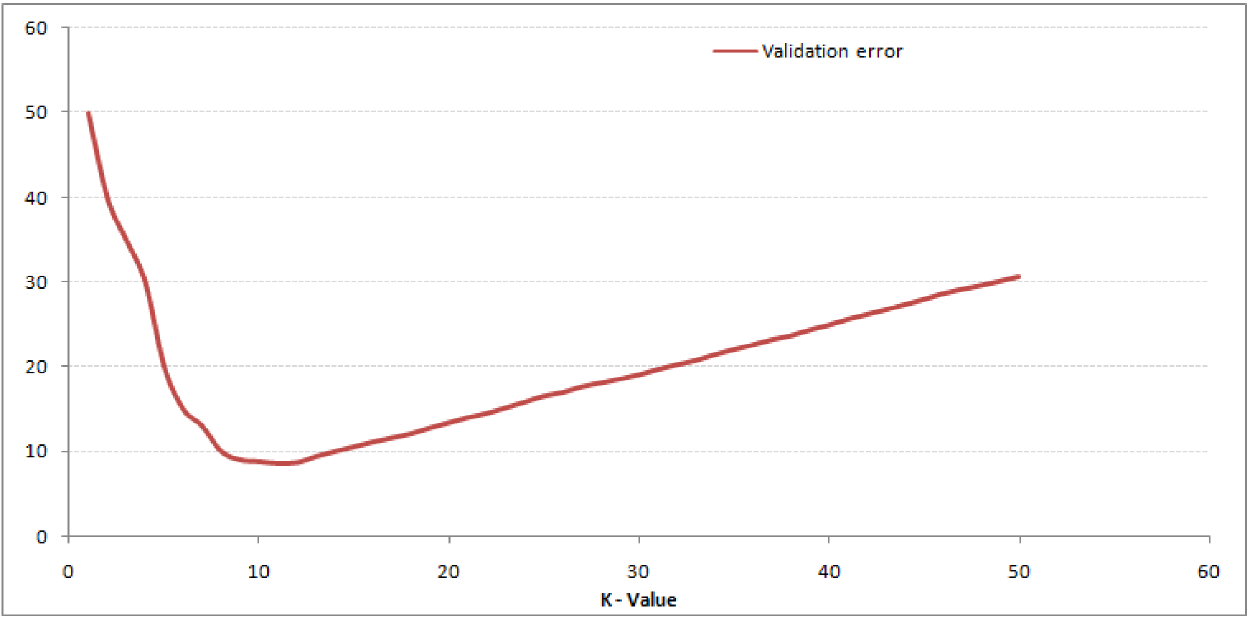




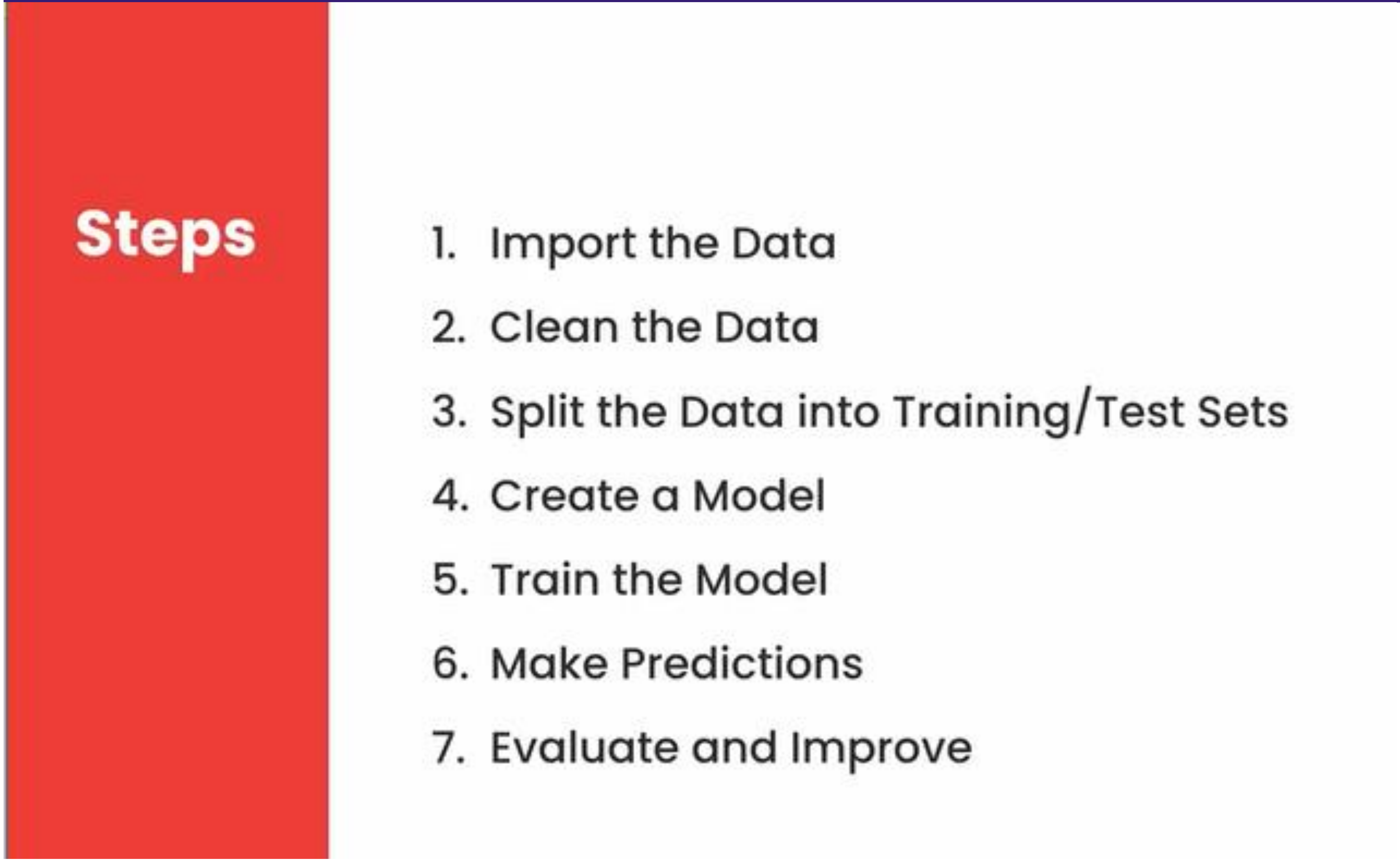
How KNN Works:Determining k

- ❑ Determining the k to select using training error and validation error for different values of k.
- ❑Validation error curve reaches a minima at a value of k = 9 (Elbow curve)





Python Hands-on Example



Steps

1. Import the Data
2. Clean the Data
3. Split the Data into Training/Test Sets
4. Create a Model
5. Train the Model
6. Make Predictions
7. Evaluate and Improve

Python Libraries

np. Multidimensional data arrays

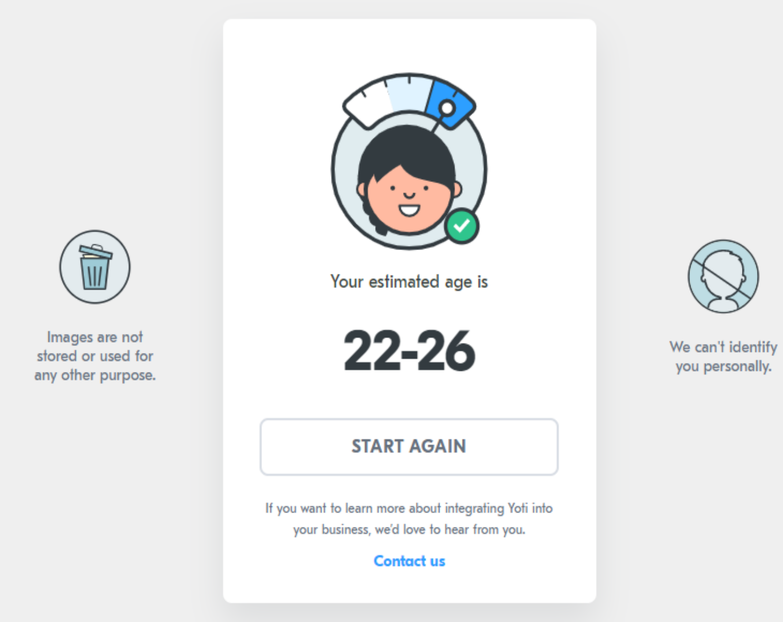
pd. Generates data frames. Used in Data Science

mat. Creates plots and graphs in 2D

sklearn. Has ML algorithms for SL and USL

Limitations and ethical issues

- ❑SL models require extensive training datasets
- ❑SL models (Classification) require proper labelling with no bias(subjectivity)
- ❑The training of AI systems should have strict data consent/protection policies
- ❑Who will store, own, and control data?
- ❑Are all your actions transparent and open to inspection?
  - ❑E.g. the banning of twitter accounts claimed to be "AI bots"
  - ❑The use of ChatGPT in plagiarism



Images are not stored or used for any other purpose.

Your estimated age is

22-26

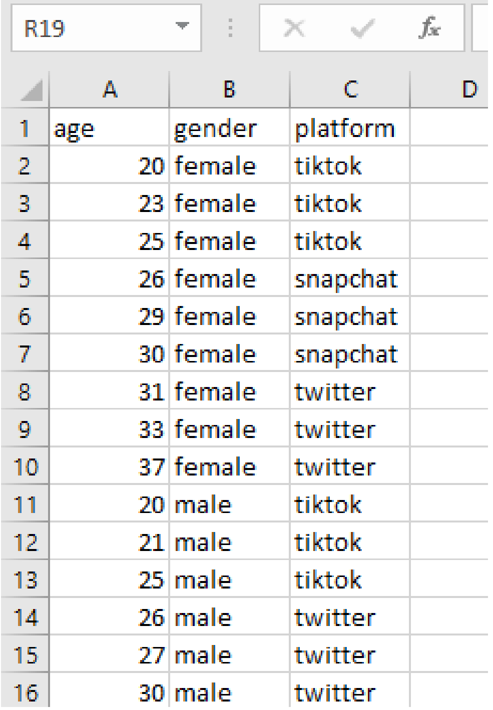
If you want to learn more about integrating Yoti into your business, we'd love to hear from you.

We can't identify you personally.

Practical Session

Python Example of Classification

- ❑Use “Social_Media_Usage.csv” dataset on moodle
- ❑Generate a model that predicts the social media platforms used by a 21-year old female and a 32-year old male
- ❑Test the accuracy of the model
- ❑Demonstration online on Friday



	A	B	C	D
1	age	gender	platform	
2	20	female	tiktok	
3	23	female	tiktok	
4	25	female	tiktok	
5	26	female	snapchat	
6	29	female	snapchat	
7	30	female	snapchat	
8	31	female	twitter	
9	33	female	twitter	
10	37	female	twitter	
11	20	male	tiktok	
12	21	male	tiktok	
13	25	male	tiktok	
14	26	male	twitter	
15	27	male	twitter	
16	30	male	twitter	

Regression Example in Python

❑ Hands-on example using the “Big Mart Sales.zip” dataset on moodle

❑Design a model using kNN algorithm to predict the k-value for test and train data.

❑Test the accuracy of the model

1 SAMUEL 7:12

Questions ????