Machine Learning Study Jam

Introduction to Machine Learning

Dr. Poo Kuan Hoong

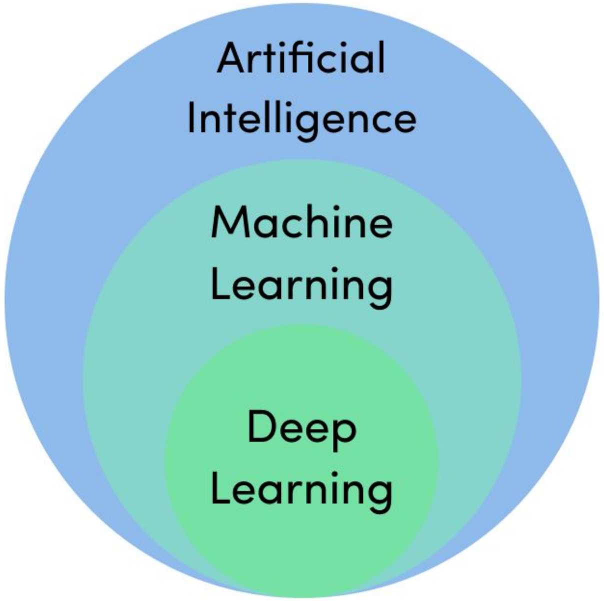
Google Developer Expert (GDE),
Lead Data Scientist

🐦 **@kuanhoong**

Slides: https://bit.ly/3SOxoto

AI vs DL vs ML

- Artificial Intelligence (AI) is the ability of a computer to do tasks that are usually done by humans
- Machine Learning (ML) is one of the methods to “achieve” AI
- Deep Learning (DL) is a method in ML with the use of Neural Networks



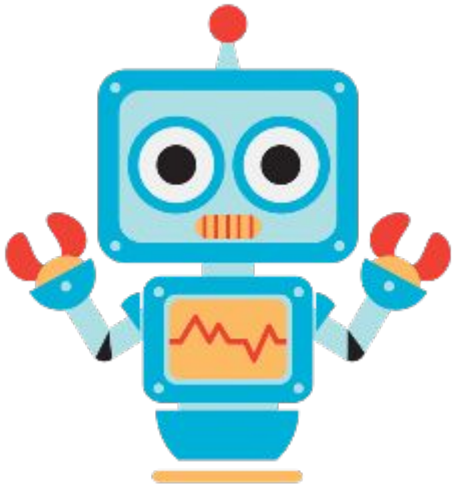
What is Machine Leaning?

What is Machine Learning?

Machine Learning

Learn from experience

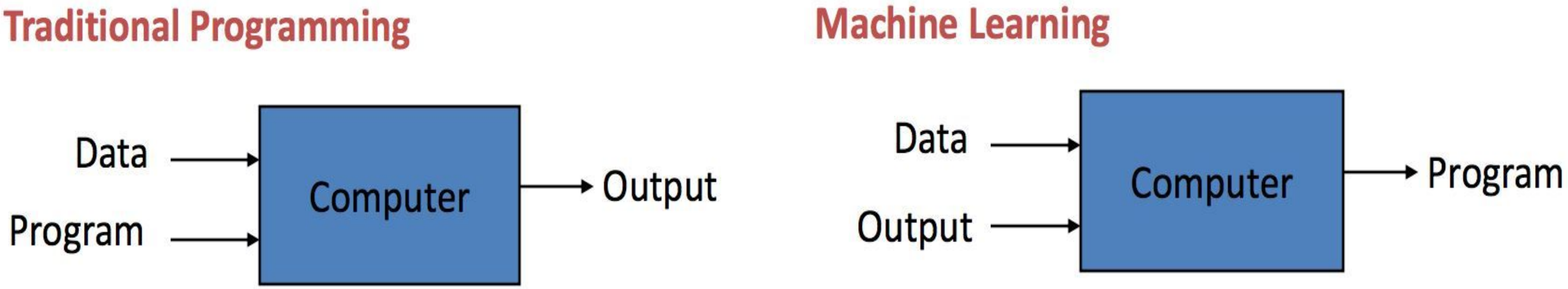
Learn from ? **DATA**



Program
App
Follow instructions

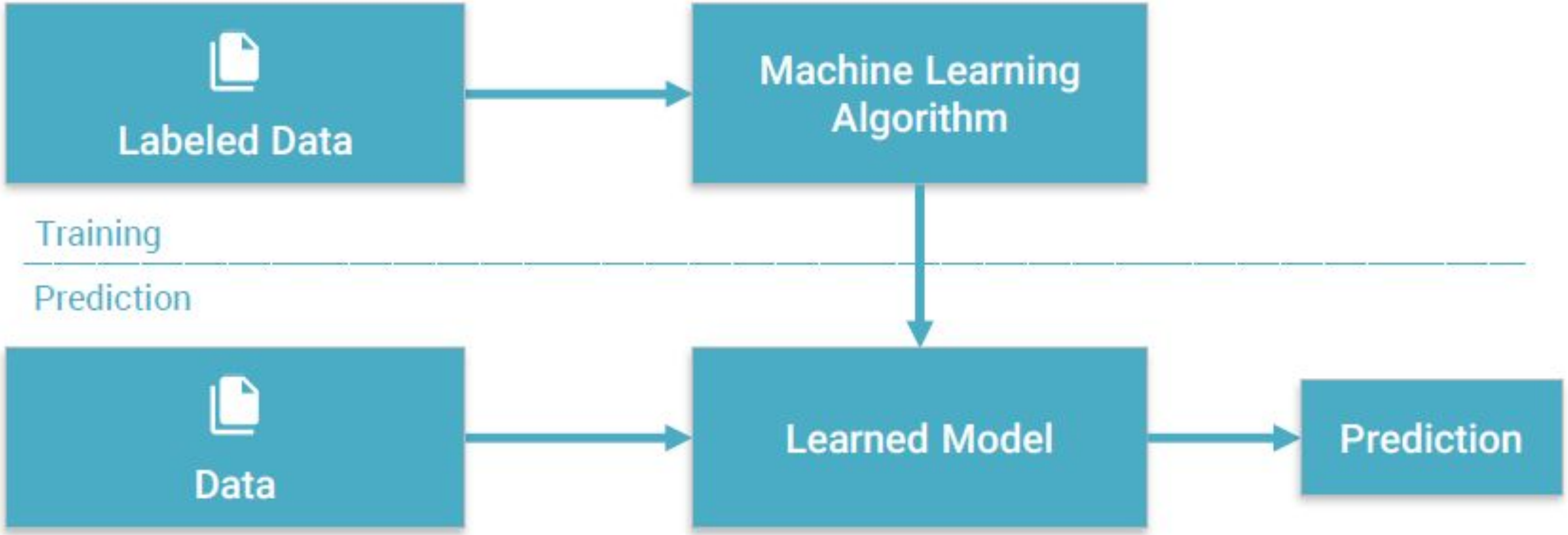
Side-to-side comparison

Traditional Programming vs Machine Learning



Machine Learning

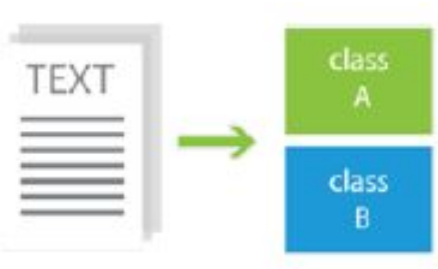
A type of Artificial Intelligence that provides computers with the ability to learn without being explicitly programmed



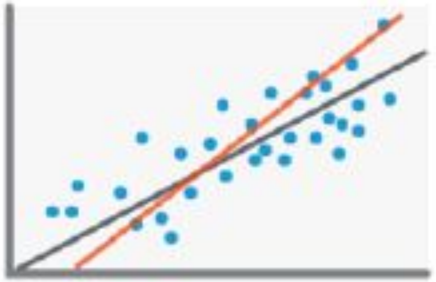
Machine Learning - Approaches

Supervised Learning	Learning from a labeled training set
Unsupervised Learning	Discovering patterns in unlabeled data
Reinforcement Learning	Learning based on feedback or reward

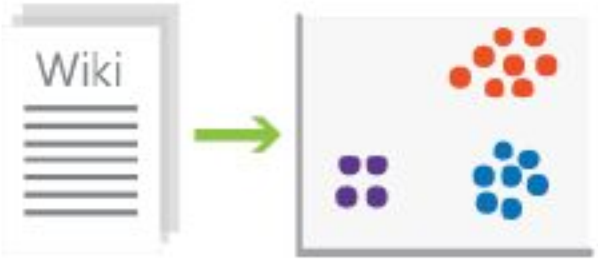
Machine Learning - Types of Problems



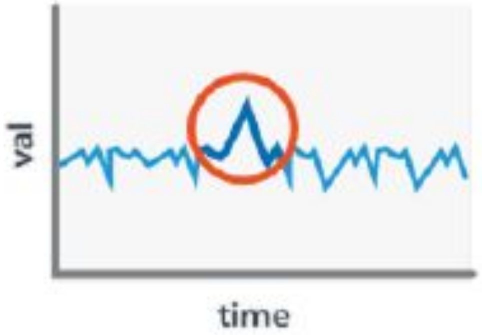
Classification



Regression



Clustering



Anomaly Detection

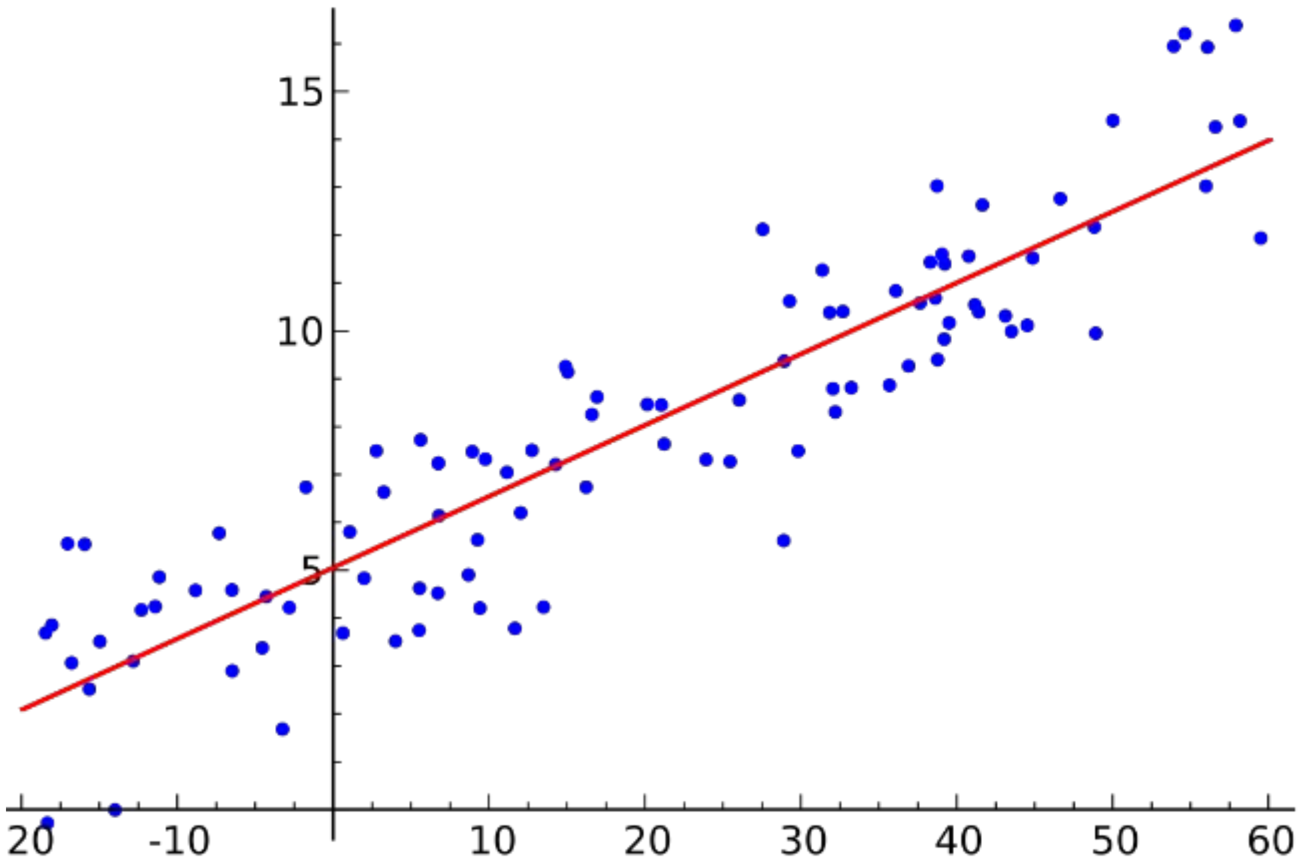
Supervised Learning vs Unsupervised Learning

Supervised Learning	Unsupervised Learning
• Data is **labelled** with class or value	• Data is **unlabeled** or value un-known
• **Goal** **:** predict class or value label	• **Goal** **:** Determine data patterns/groupings
• **Knowledge of output** – learning with the presence of "expert" / teacher	• **No knowledge of output** class or value
• Regression & classification	• Clustering

Regression

Regression

- Regression analysis is a set of statistical processes for estimating the relationships among variables.
- The output is continuous

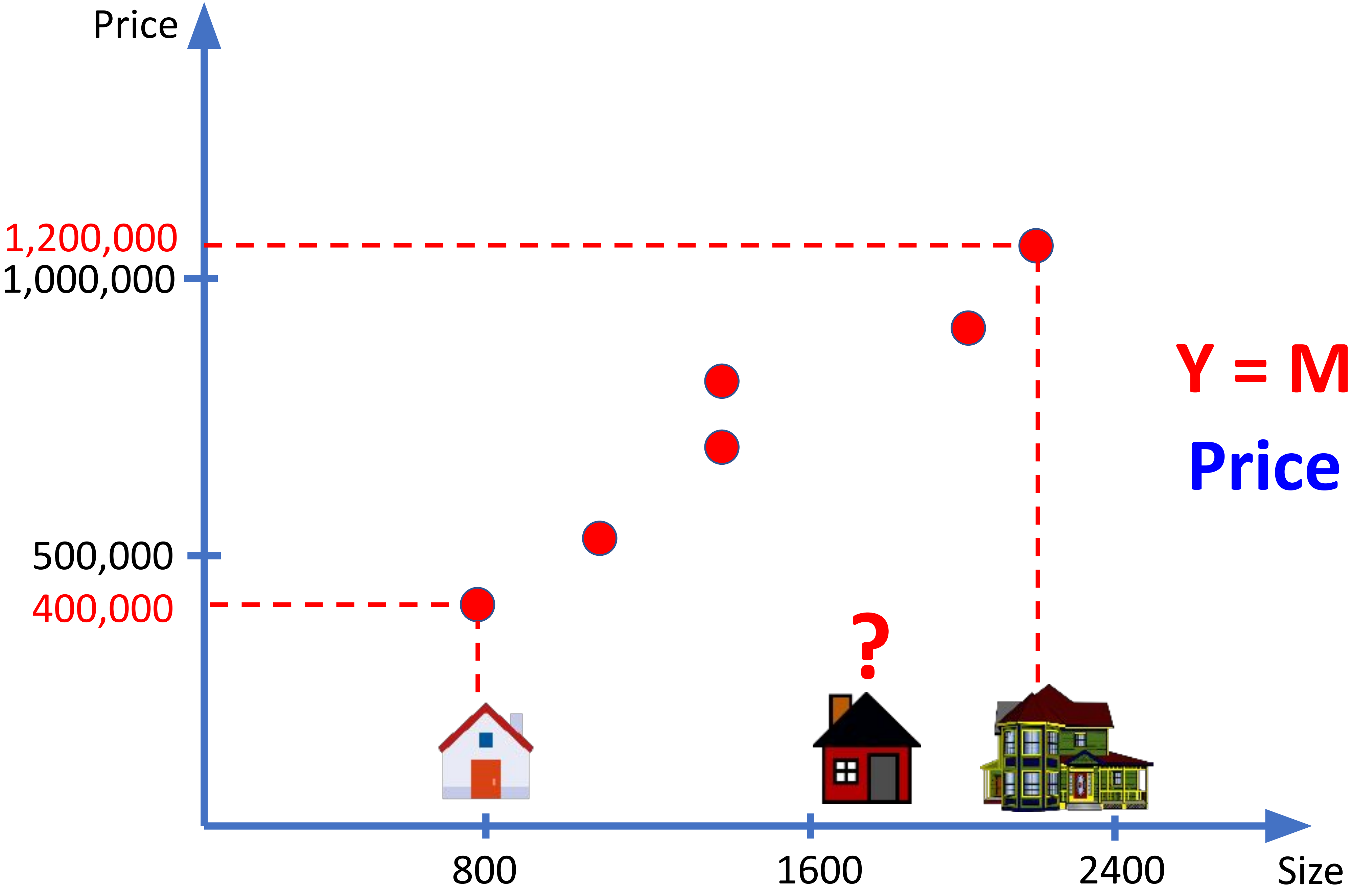


Regression

- Predict House price

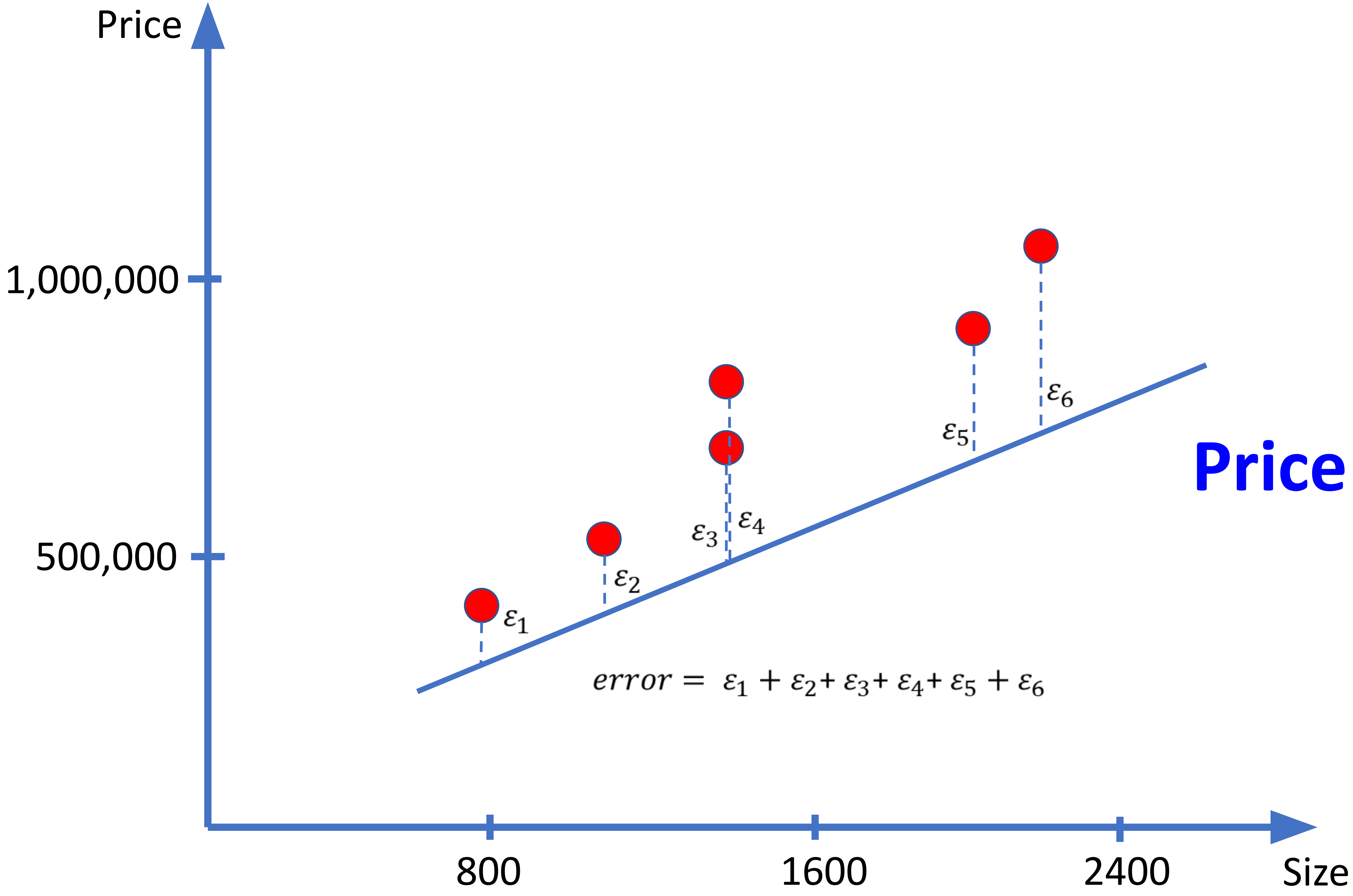
label

Rooms	Size	Price
3	1400	750,000
2	1000	550,000
2	800	400,000
3	2000	900,000
4	2100	1,200,000
3	1400	810,000



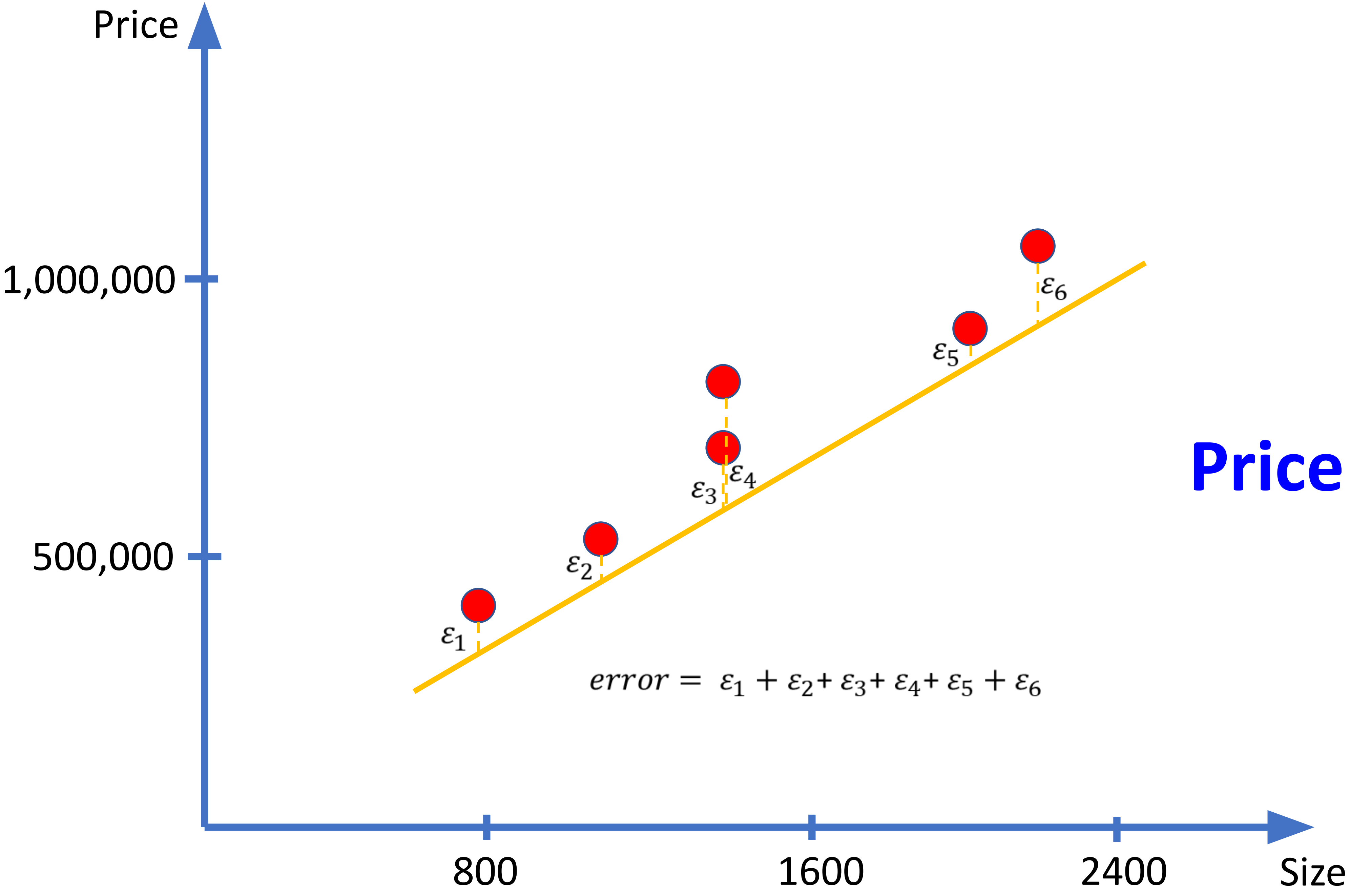
Y = M.X + C

Price ~ Size



$$error = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_6$$

Price ~ Size

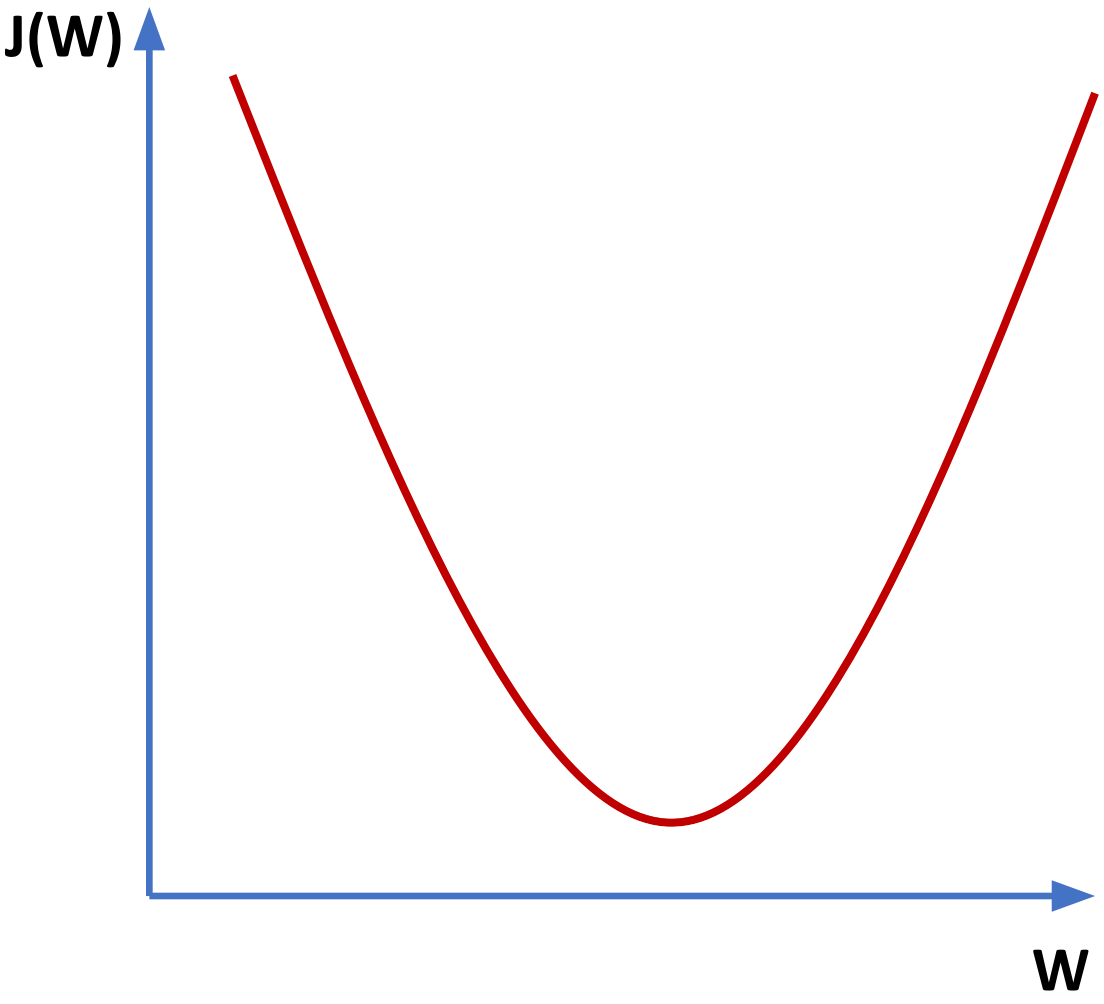


$$error = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_6$$

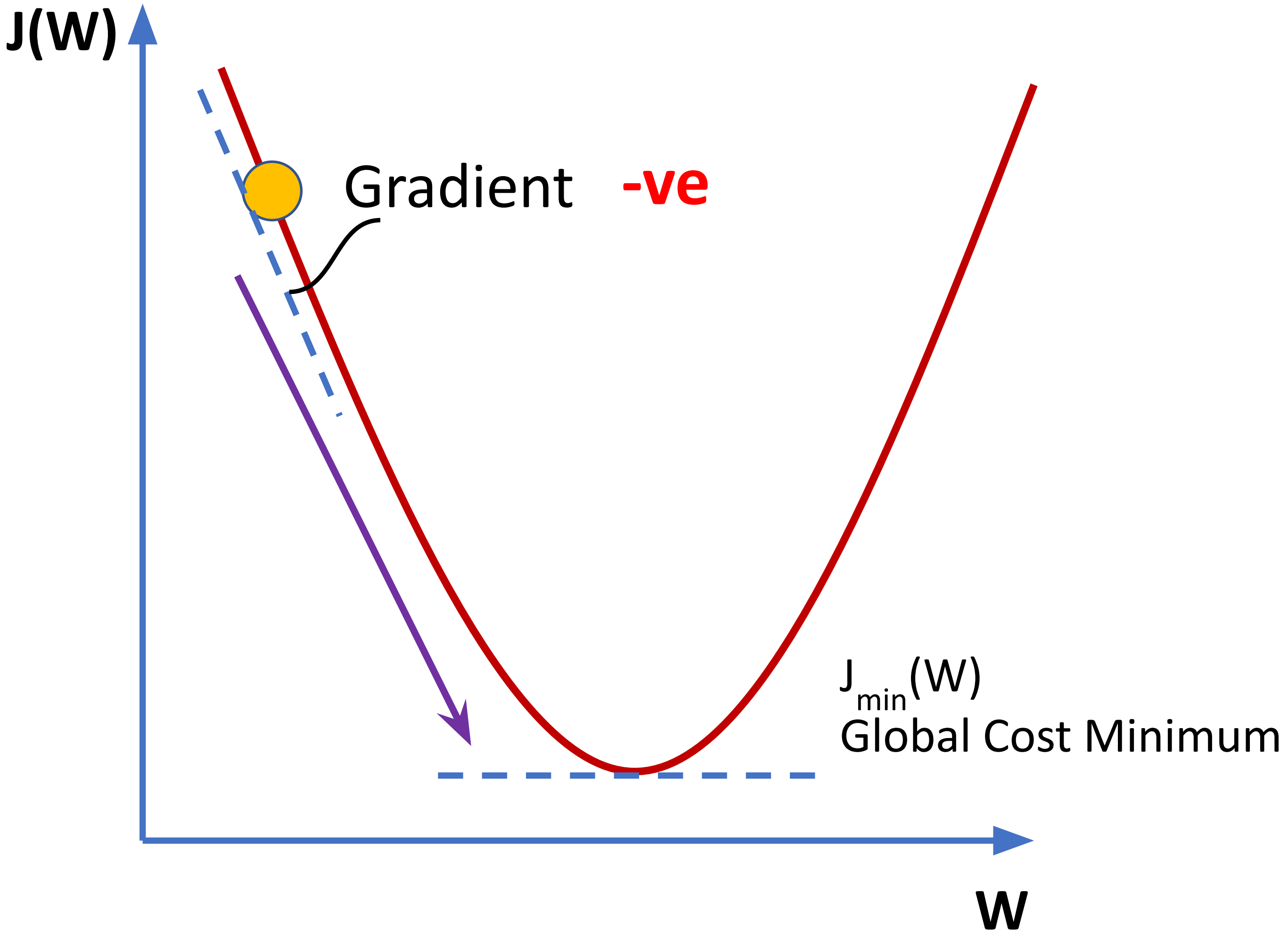
Price ~ Size

Loss/Cost Function – Minimize errors

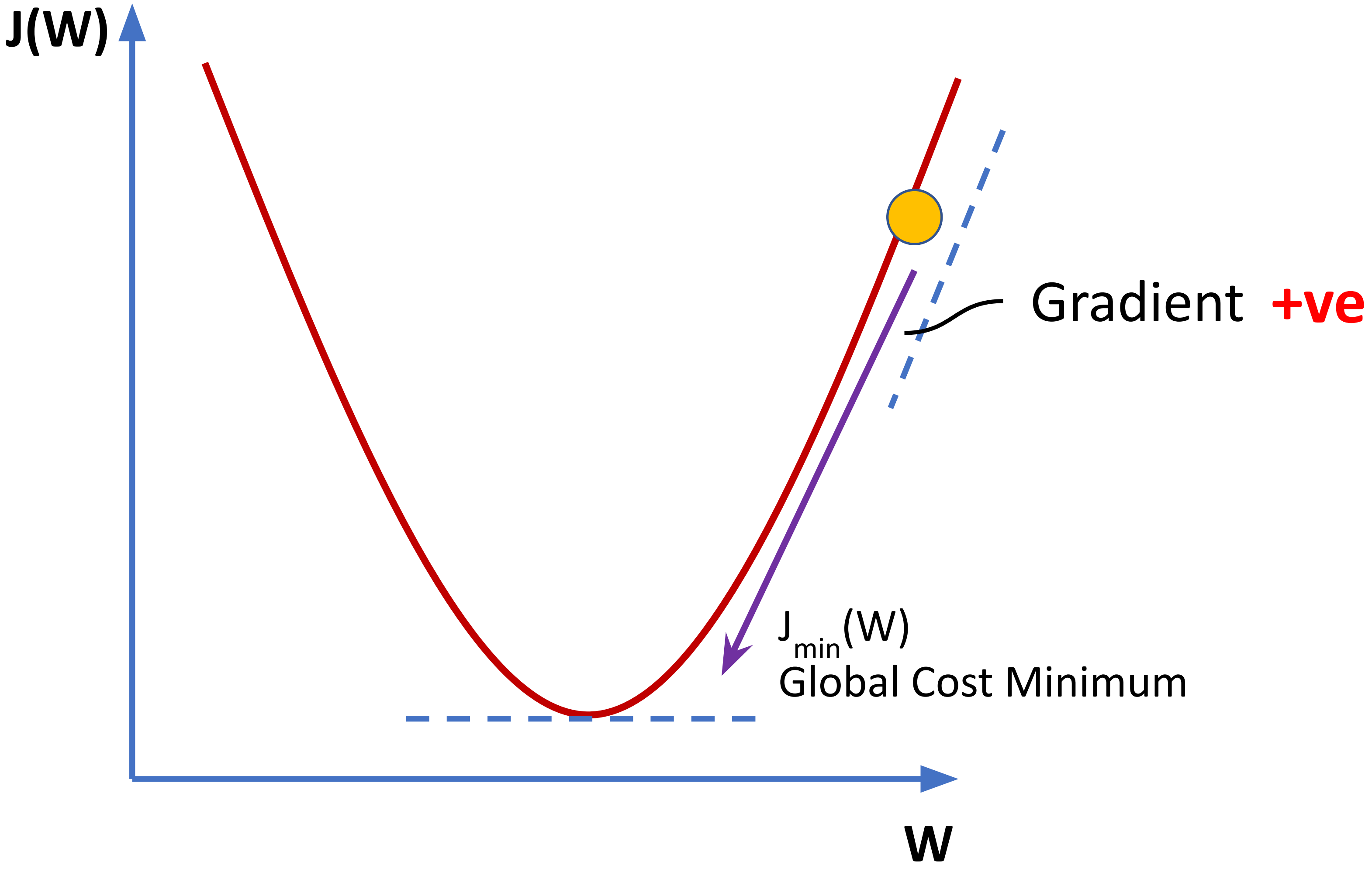
- A Loss function or Cost function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event.
- An optimization problem seeks to minimize a loss function.



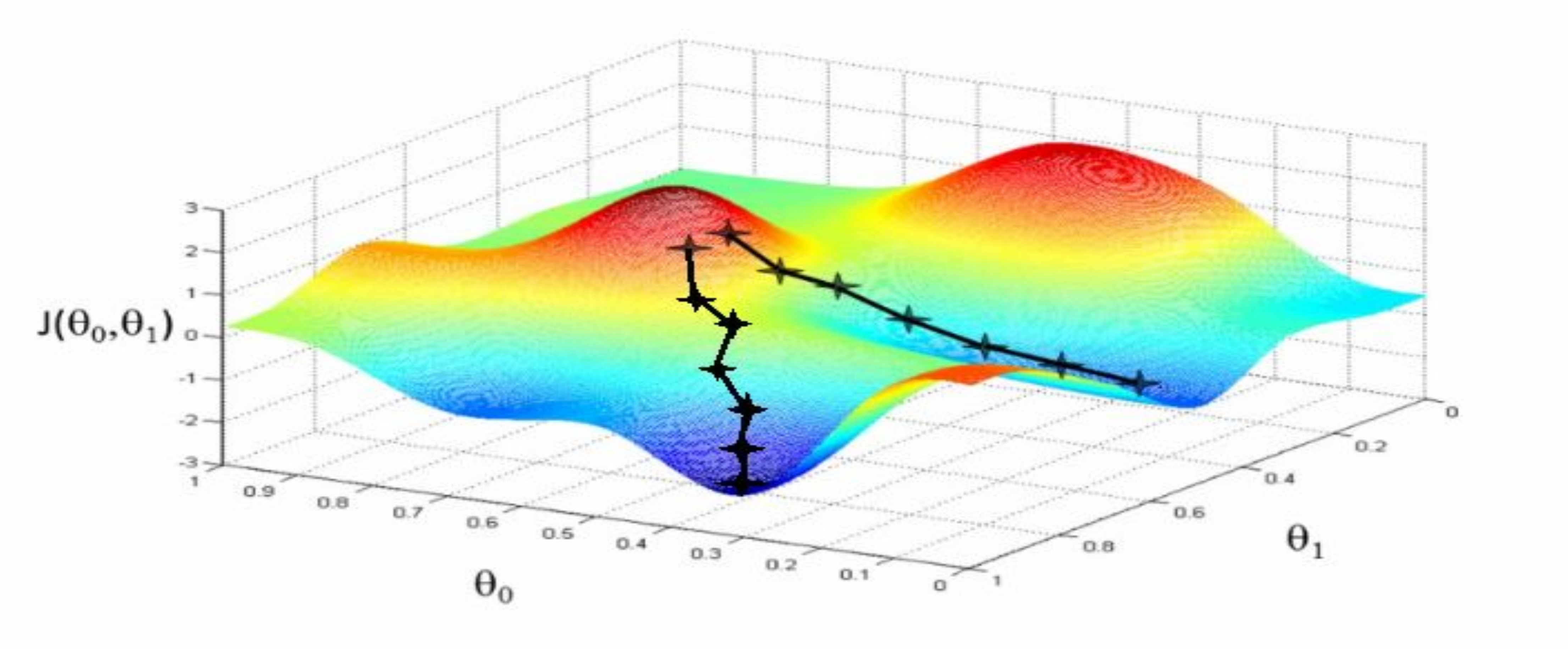
Gradient Descent

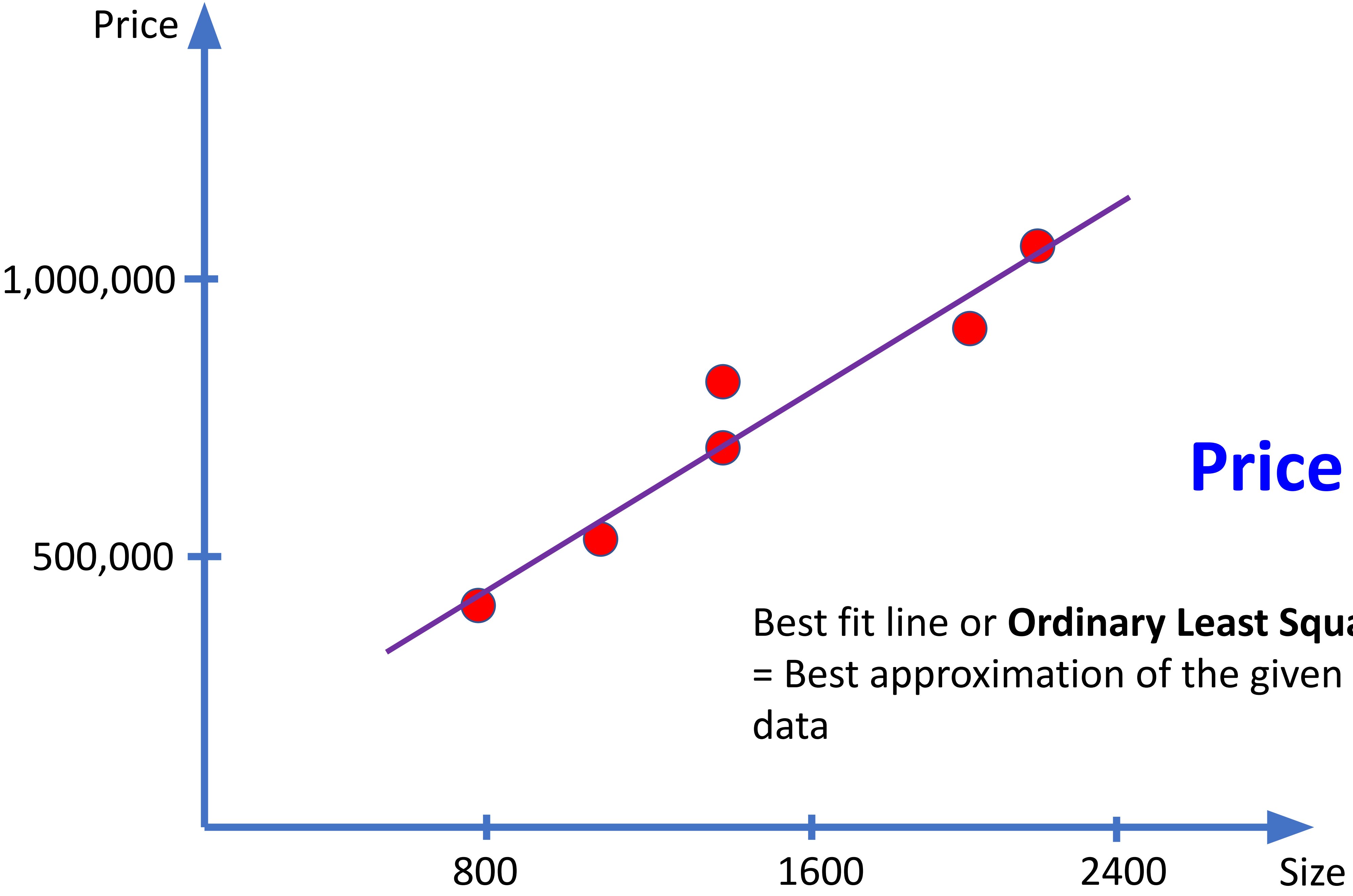


Gradient Descent



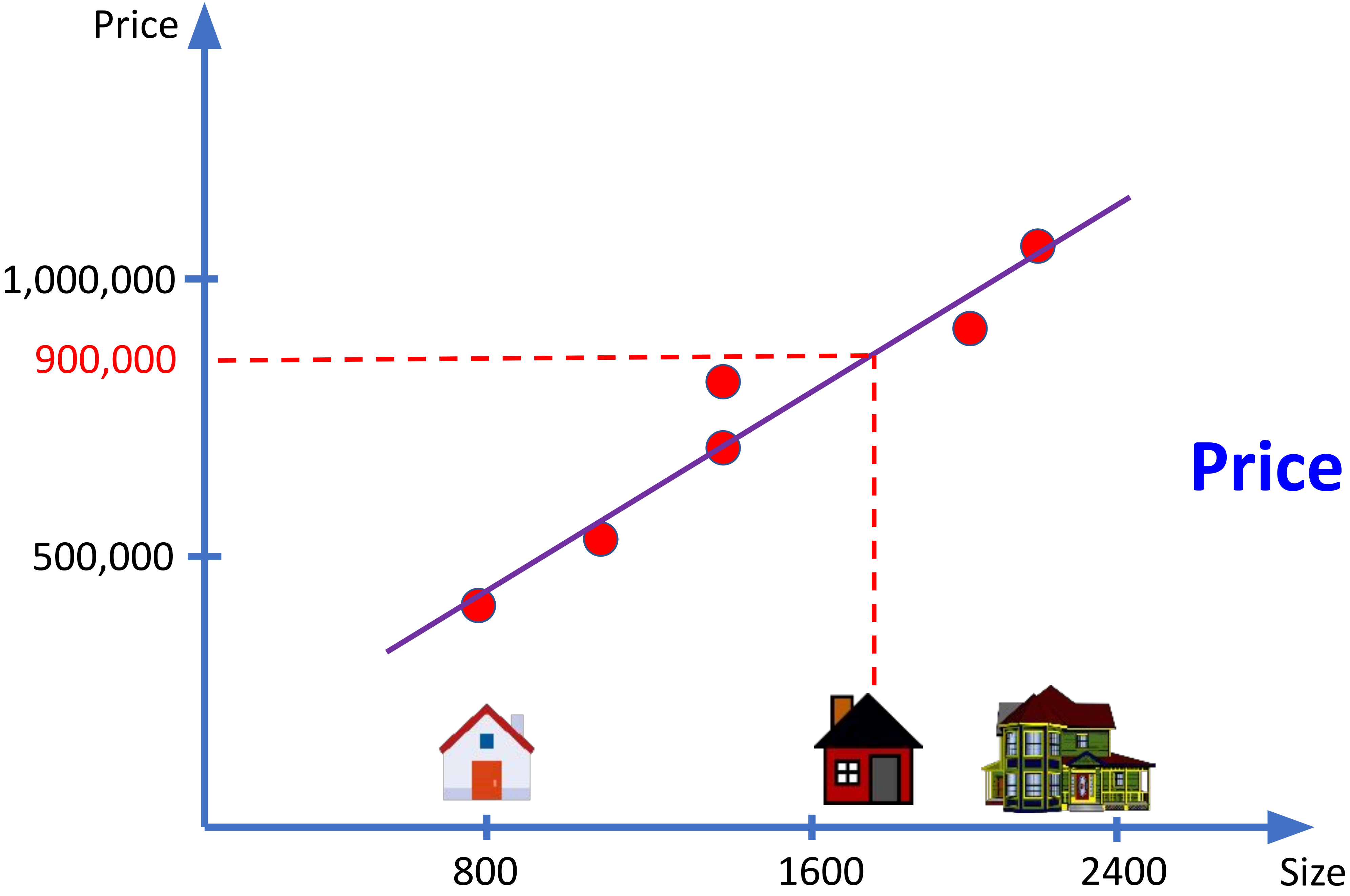
Gradient Descent





Price ~ Size

Best fit line or **Ordinary Least Square (OLS)** = Best approximation of the given set of data

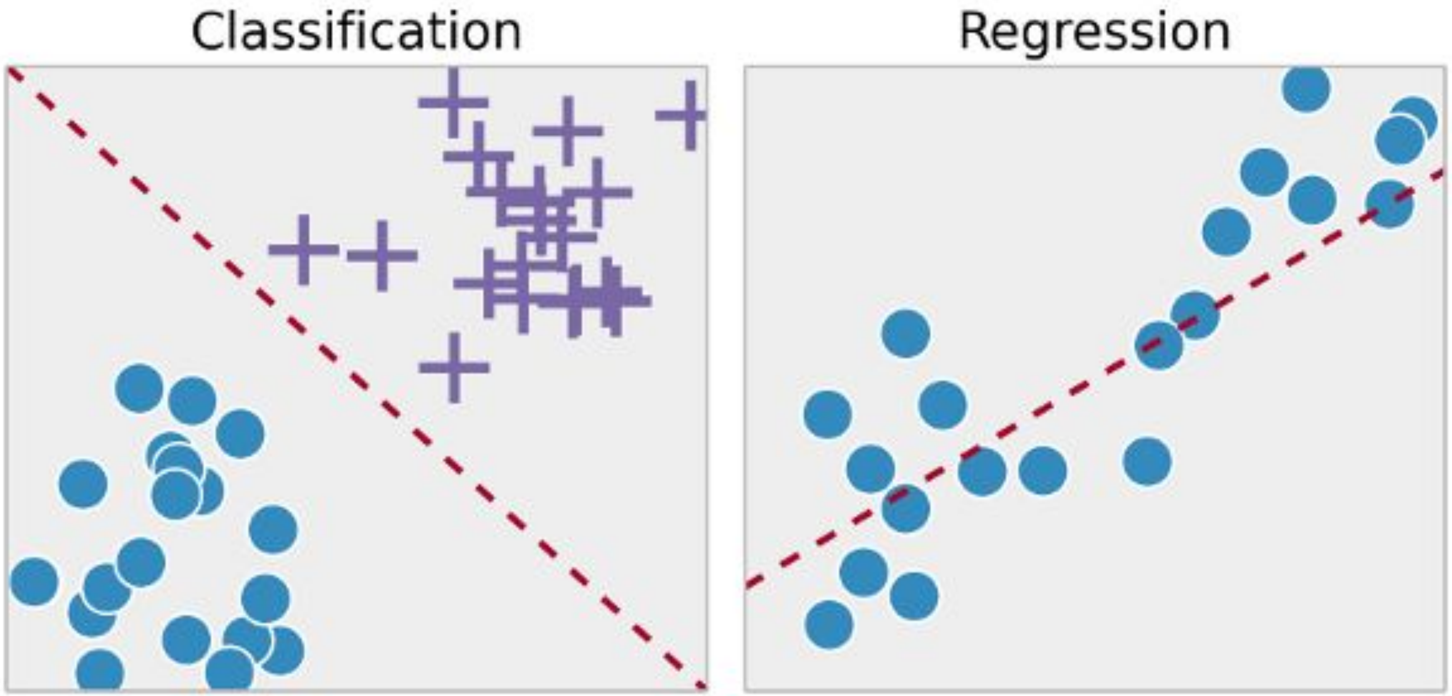


Price ~ Size

Classification

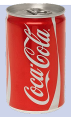
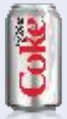
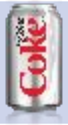
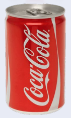
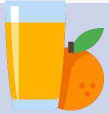
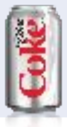
Classification

- Classification is identifying or predict group membership or class
- The output is discrete/categorical variable

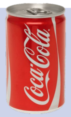
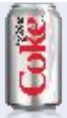
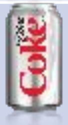
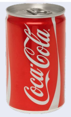
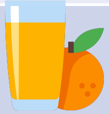
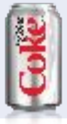


Classification

label

Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	

Classification

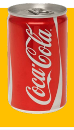
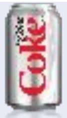
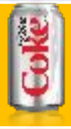
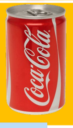
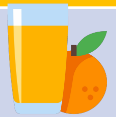
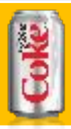
Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	

Quiz: Between Gender and Age, which one seems to be more decisive for predicting which drink will the users choose?

○ Gender

○ Age

Classification

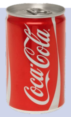
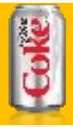
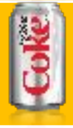
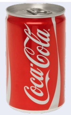
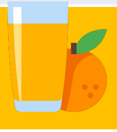
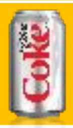
Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	

Quiz: Between Gender and Age, which one seems to be more decisive for predicting which drink will the users choose?

[✓ ○ Gender]

○ Age

Classification

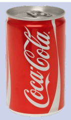
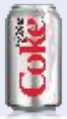
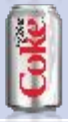
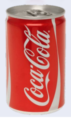
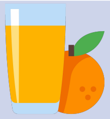
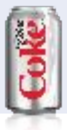
Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	

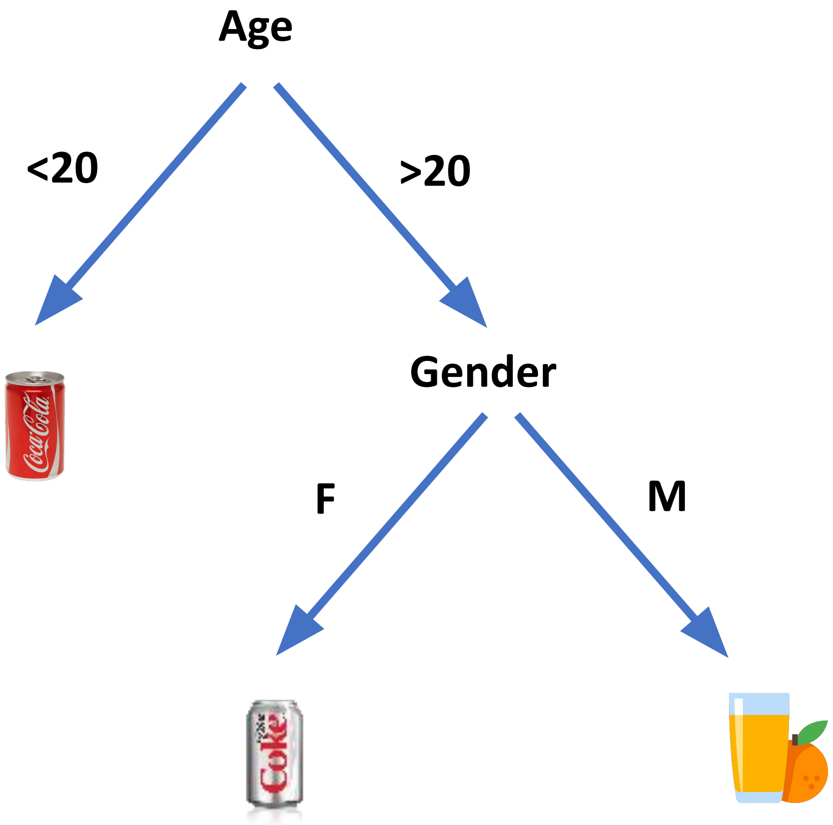
Quiz: Between Gender and Age, which one seems to be more decisive for predicting which drink will the users choose?

○ Gender

[✓ ○ Age]

Decision Tree

Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	



Age

<20

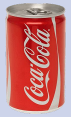
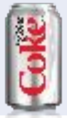
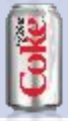
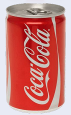
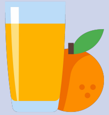
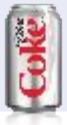
>20

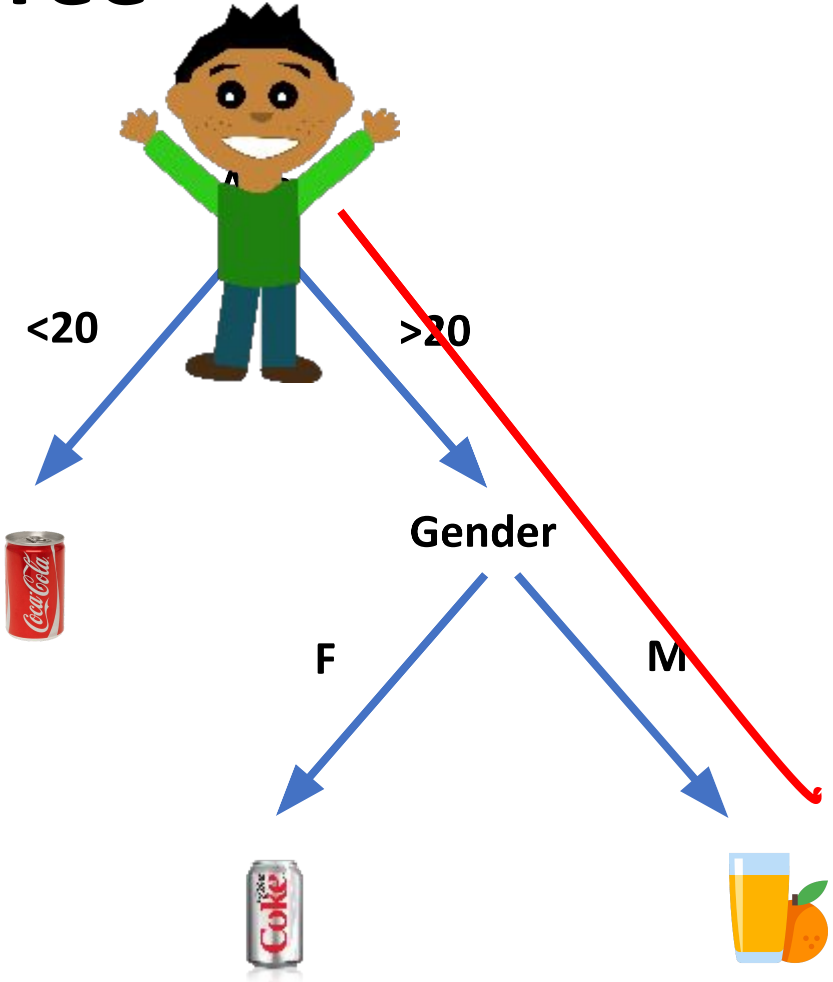
Gender

F

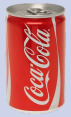
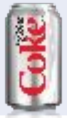
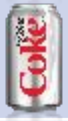
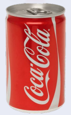
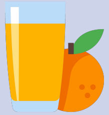
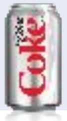
M

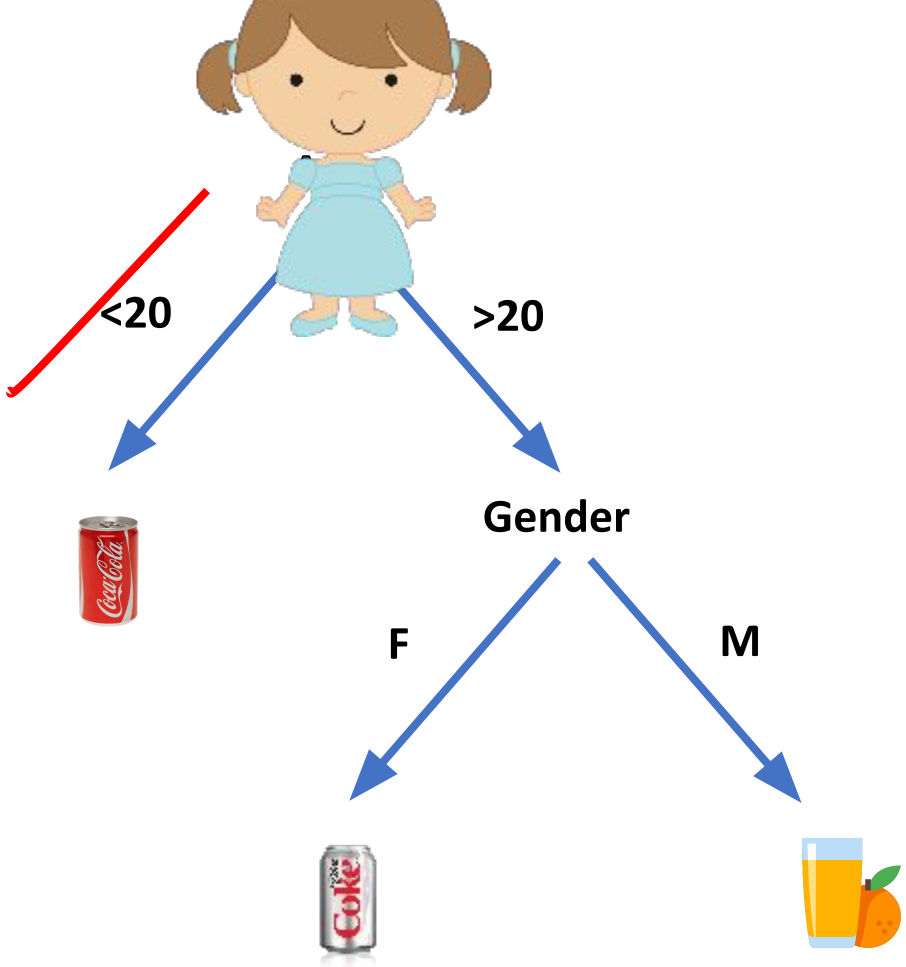
Decision Tree

Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	



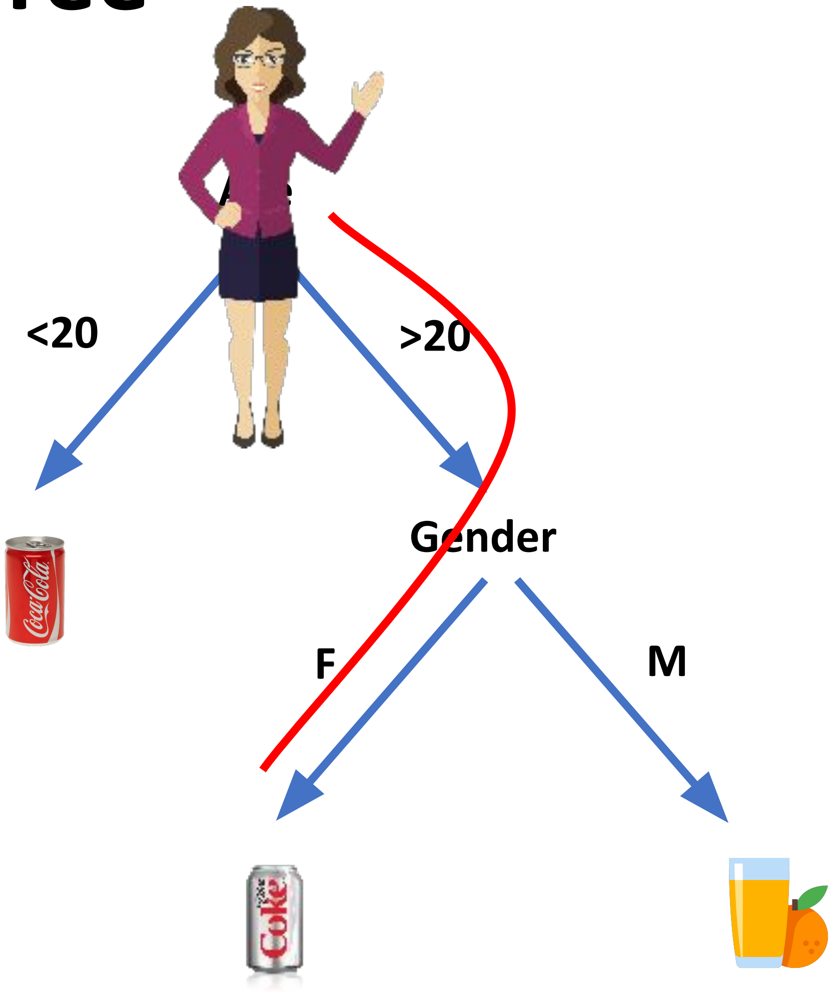
Decision Tree

Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	



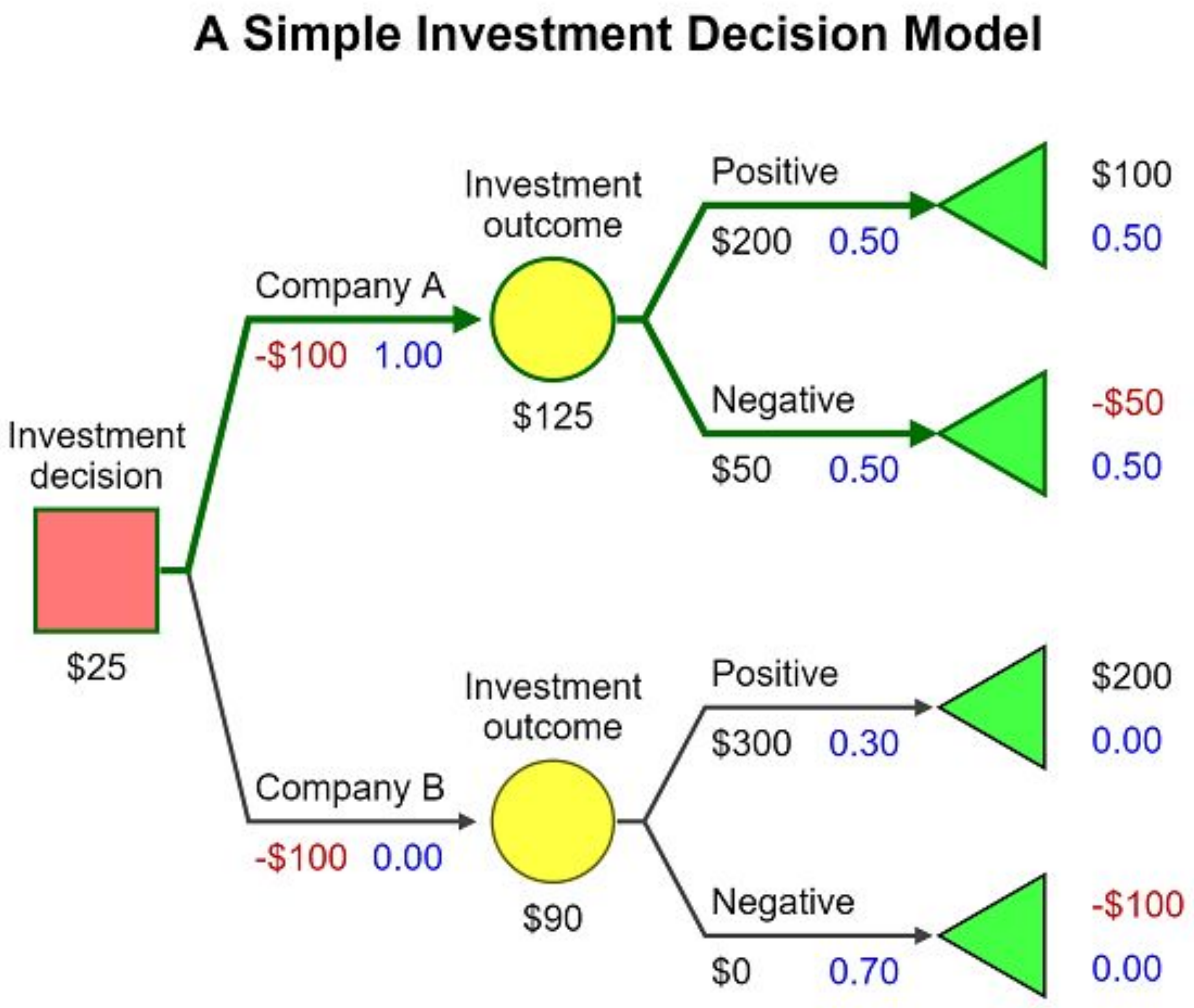
Decision Tree

Gender	Age	Drinks
F	15	
M	20	
F	21	
F	18	
M	23	
F	22	

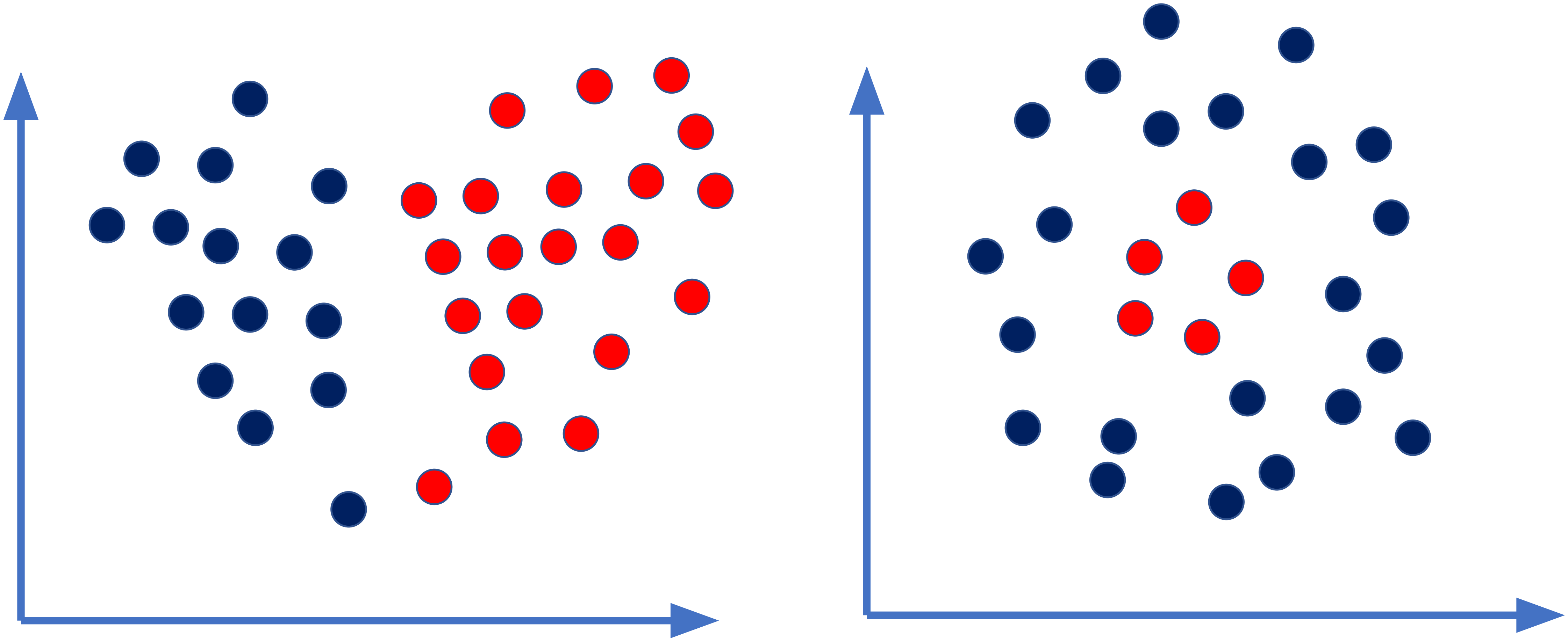


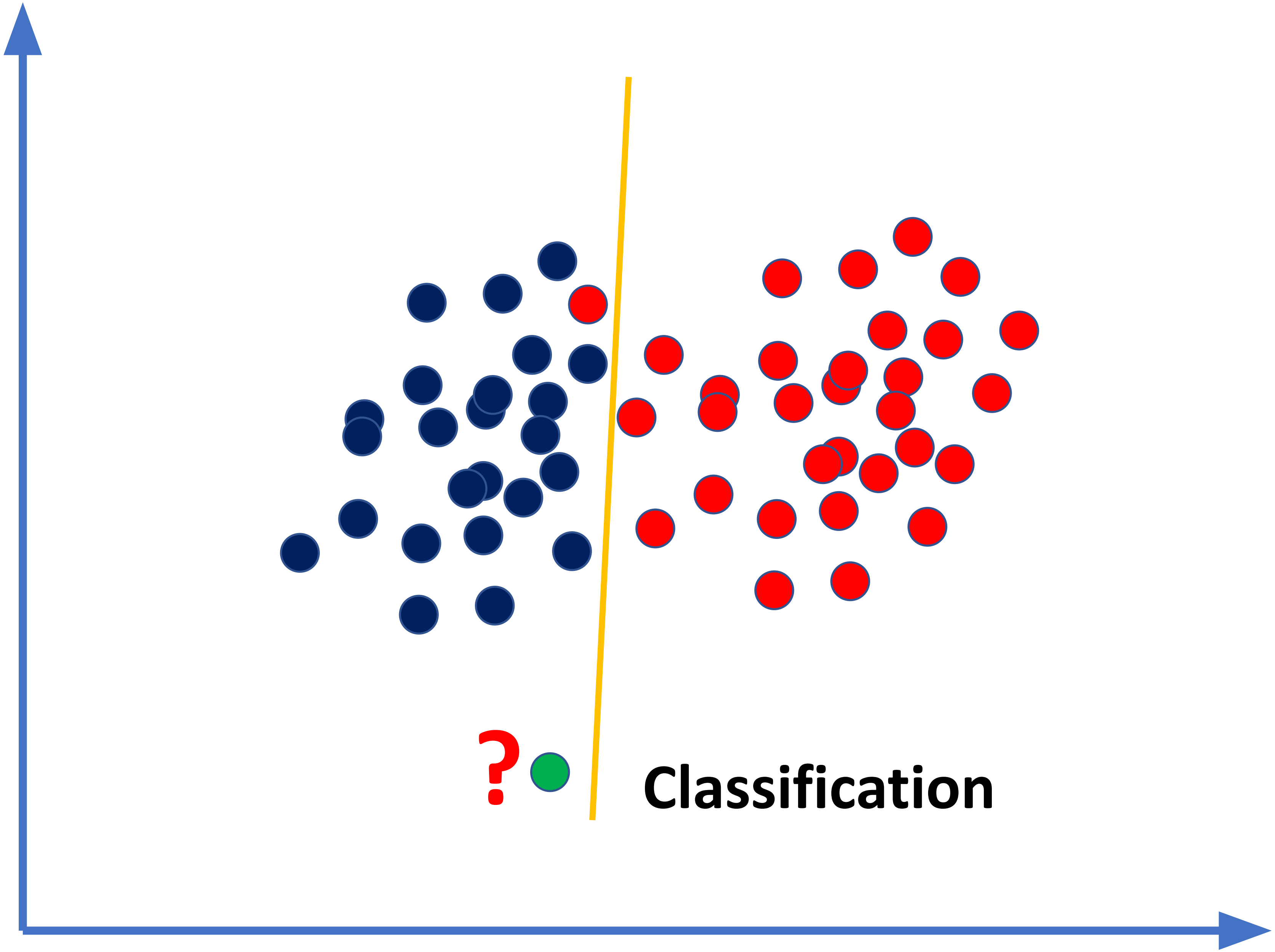
Decision Tree

- Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

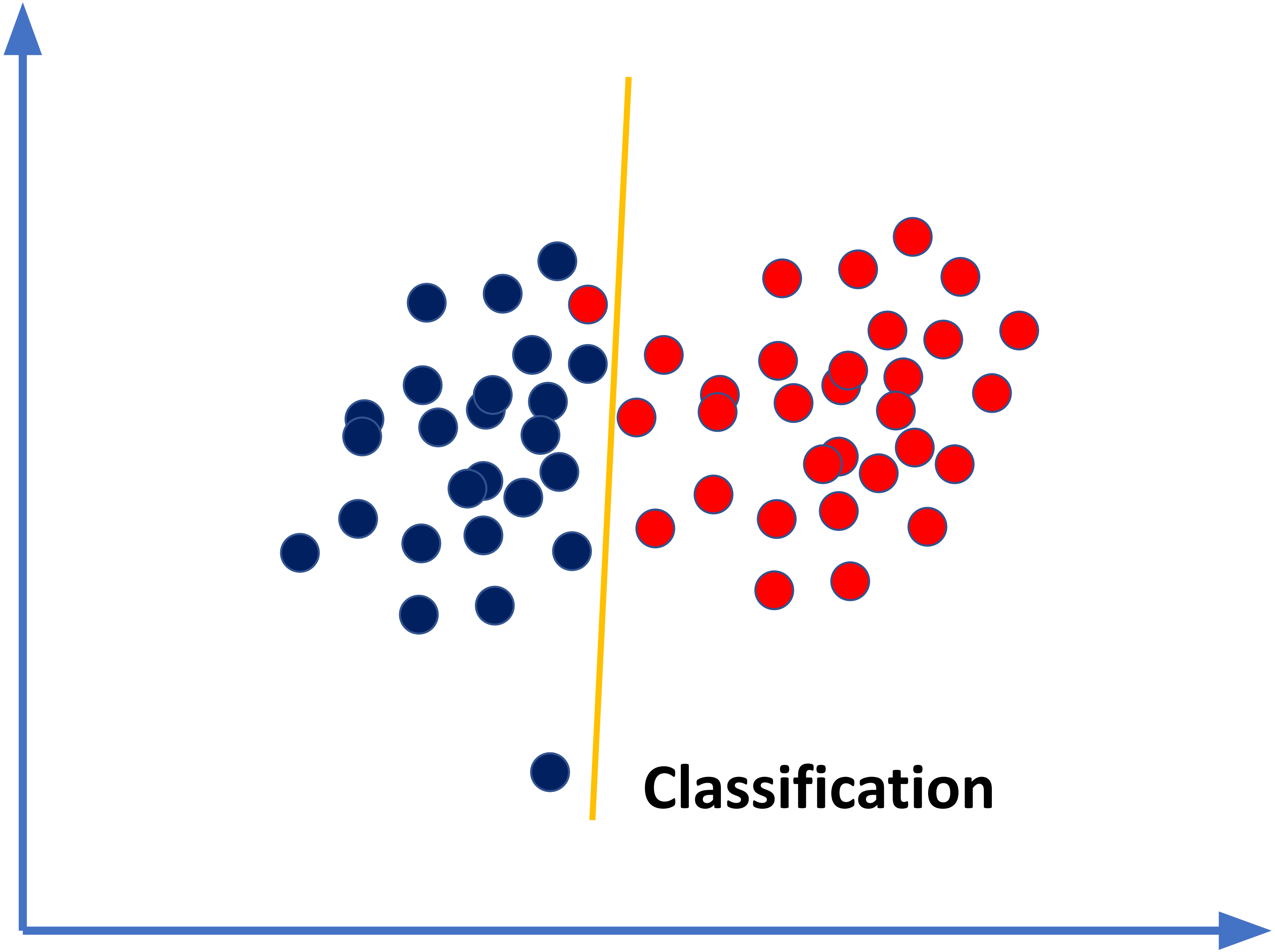


Support Vector Machine

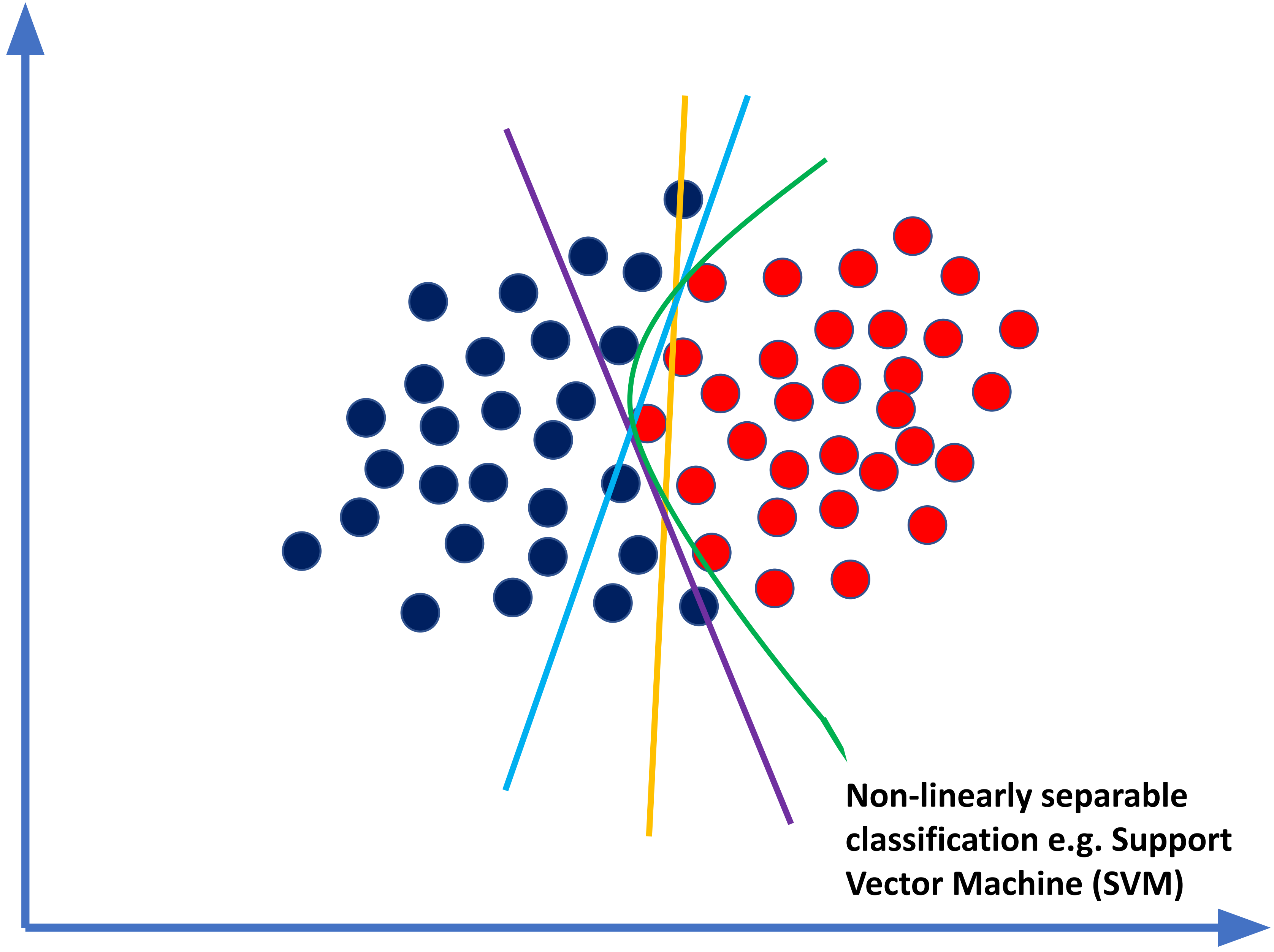




Classification



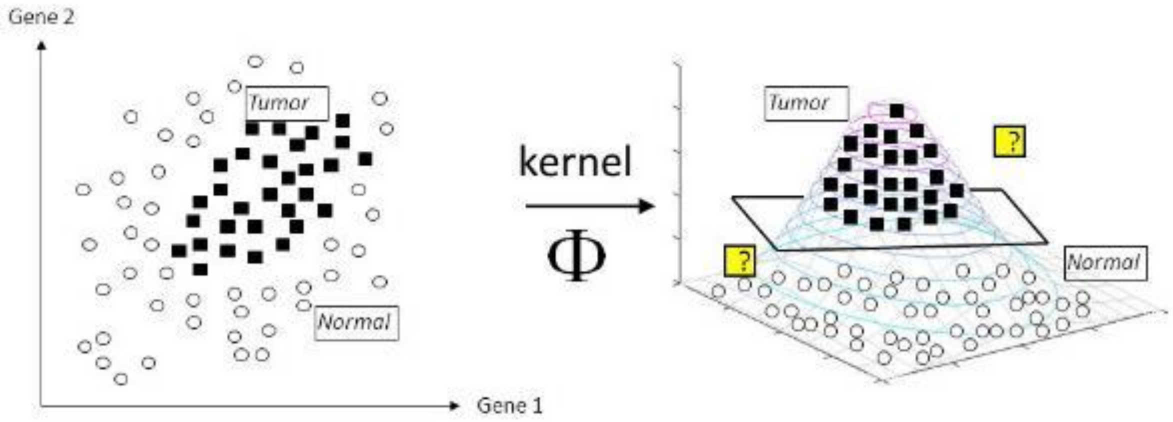
Classification



Non-linearly separable classification e.g. Support Vector Machine (SVM)

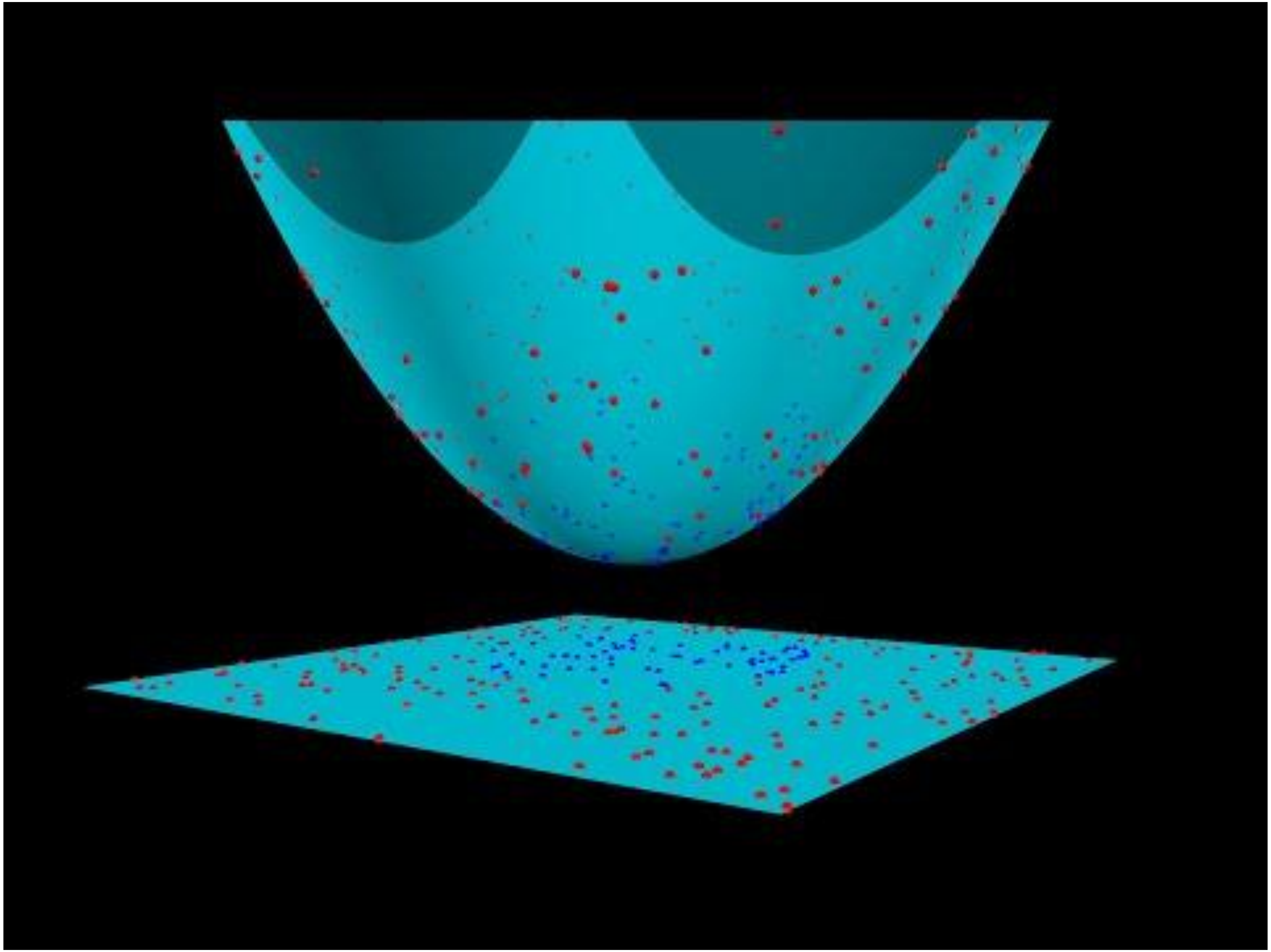
Kernel Trick

- The kernel trick avoids the explicit mapping that is needed to get linear learning algorithms to learn a nonlinear function or decision boundary.
- For all and in the input space , certain functions can be expressed as an inner product in another space .
- The function is often referred to as a kernel or a kernel function.



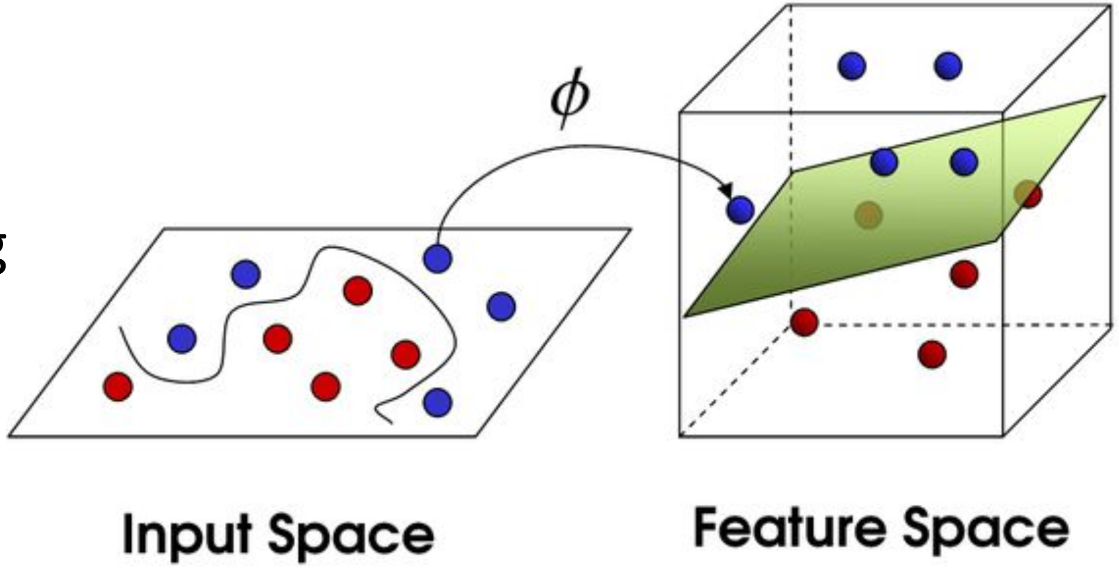
Data is not linearly separable in the input space

Data is linearly separable in the feature space obtained by a kernel



Support Vector Machine (SVM)

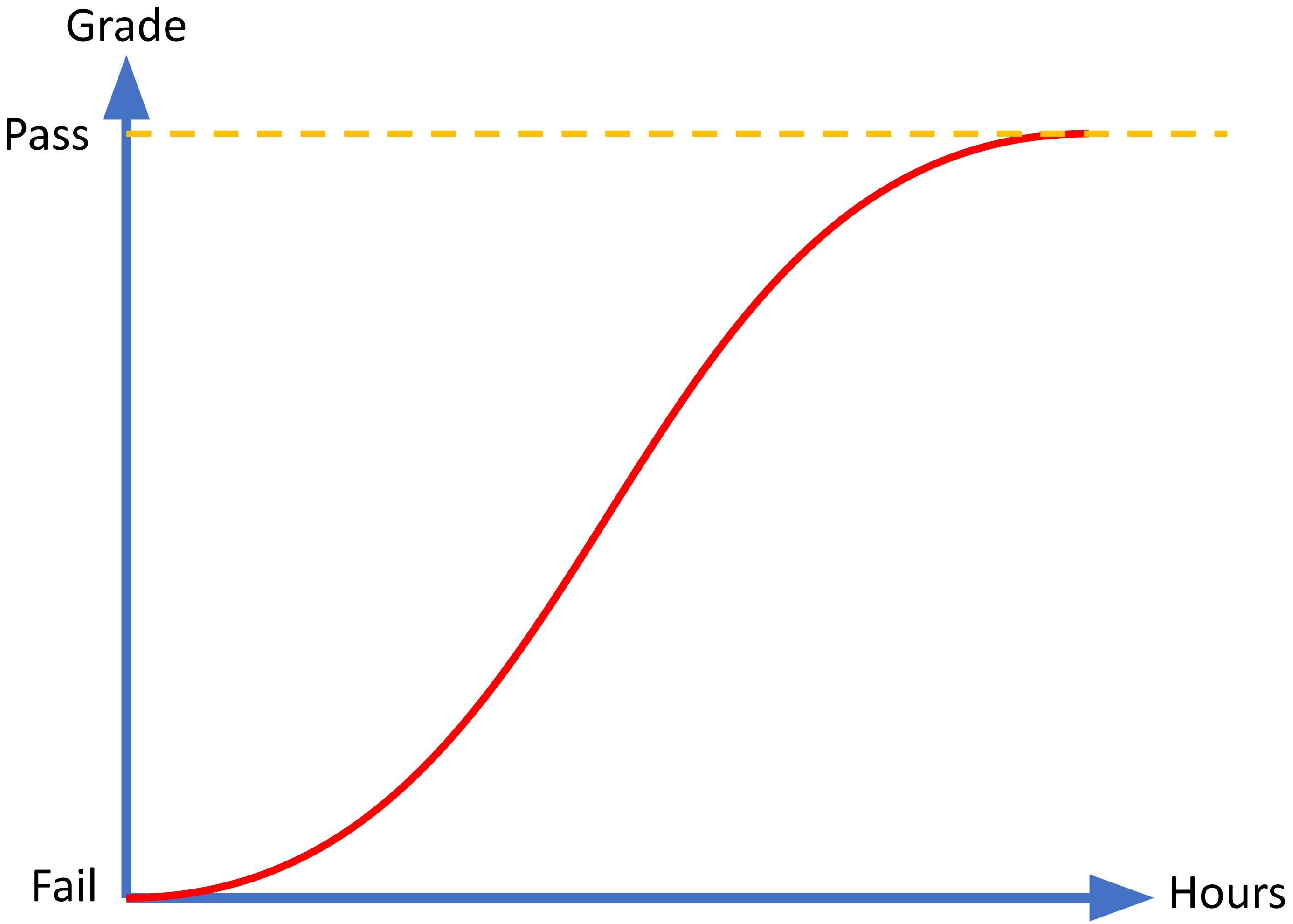
- A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.
- In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.



Logistic Regression

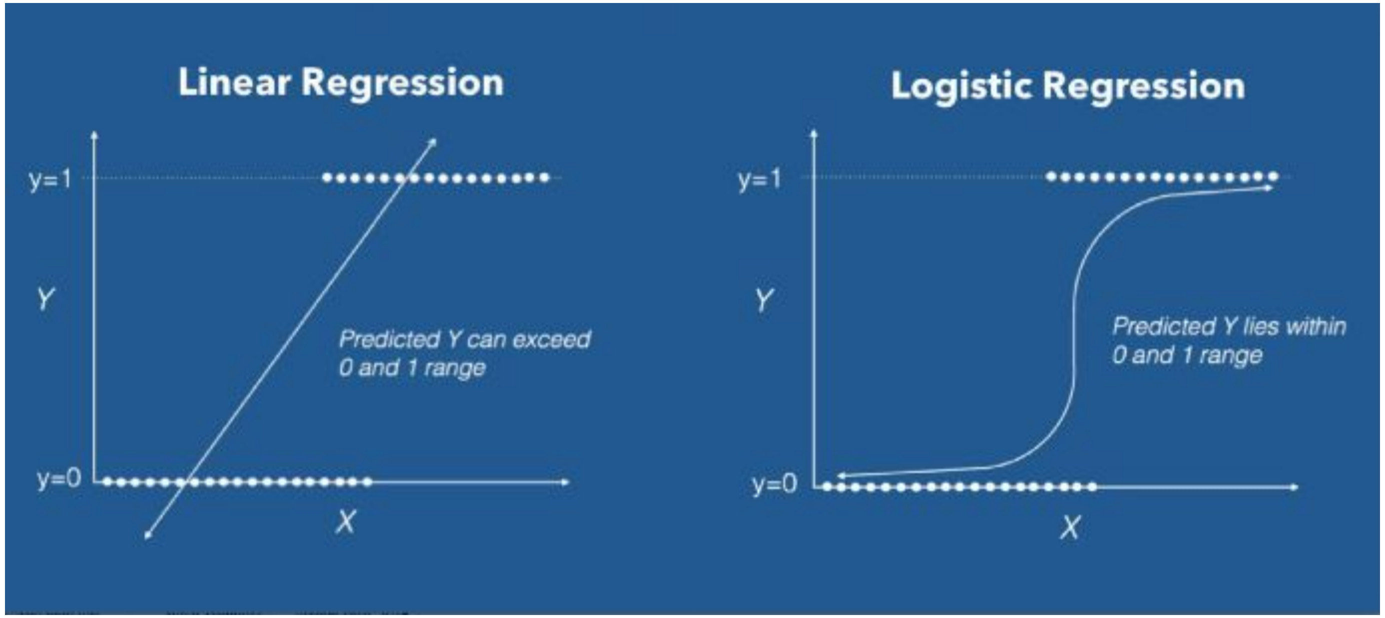
label

Gender	Hours of Study	Pass/Fail
F	4	Pass
M	3	Pass
F	3	Fail
F	2	Fail
M	3	Pass
F	1	Fail



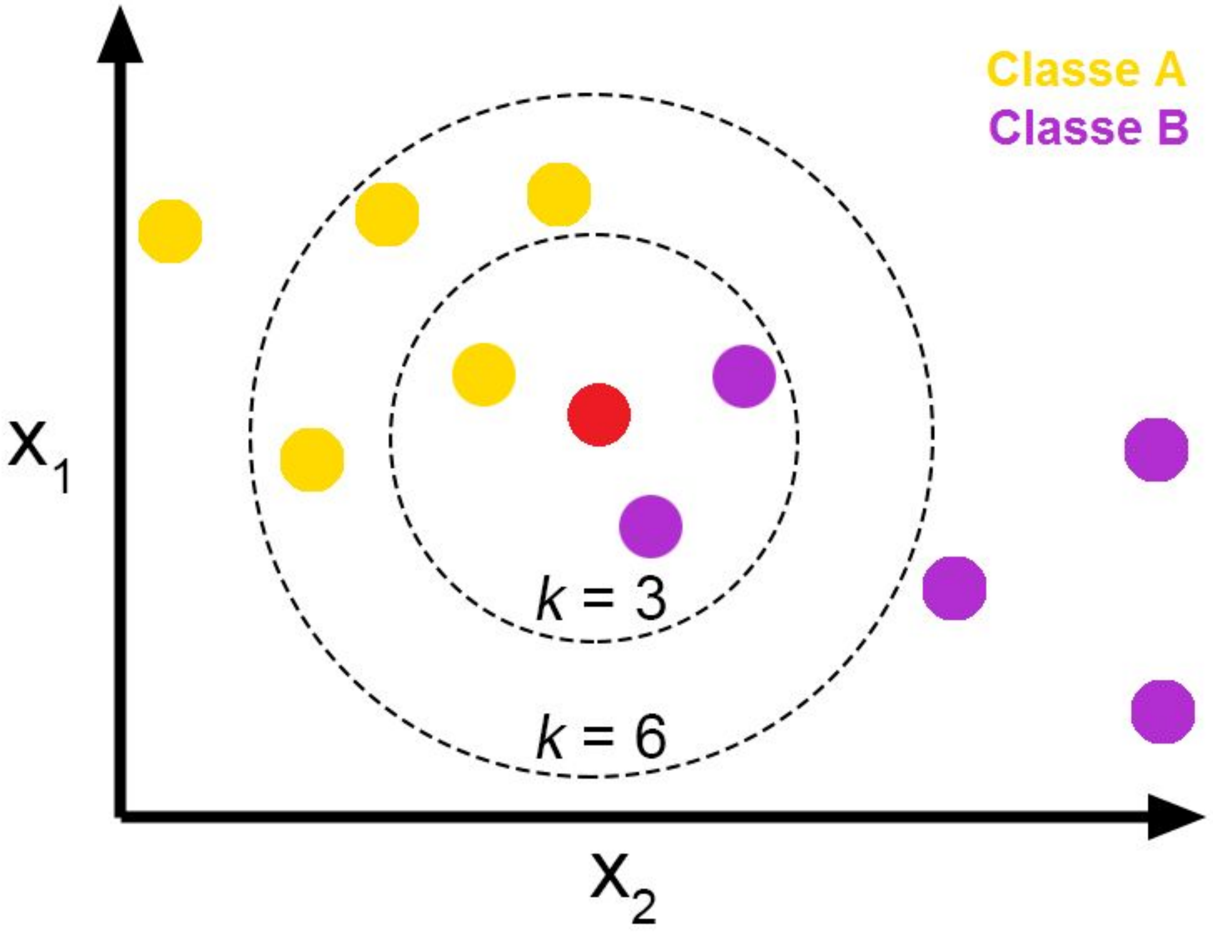
Logistic Regression

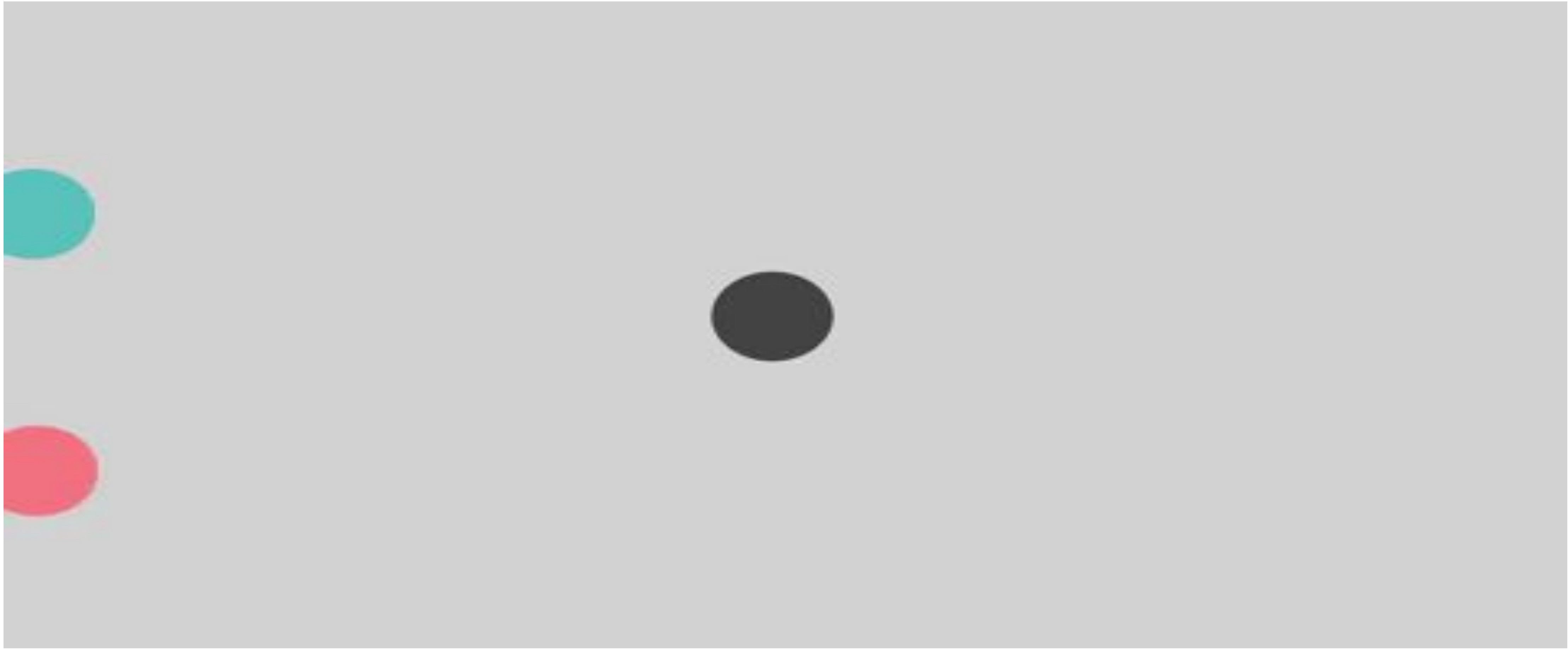
- Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.
- The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).


Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed 0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within 0 and 1 range

K-Nearest Neighbours (KNN)

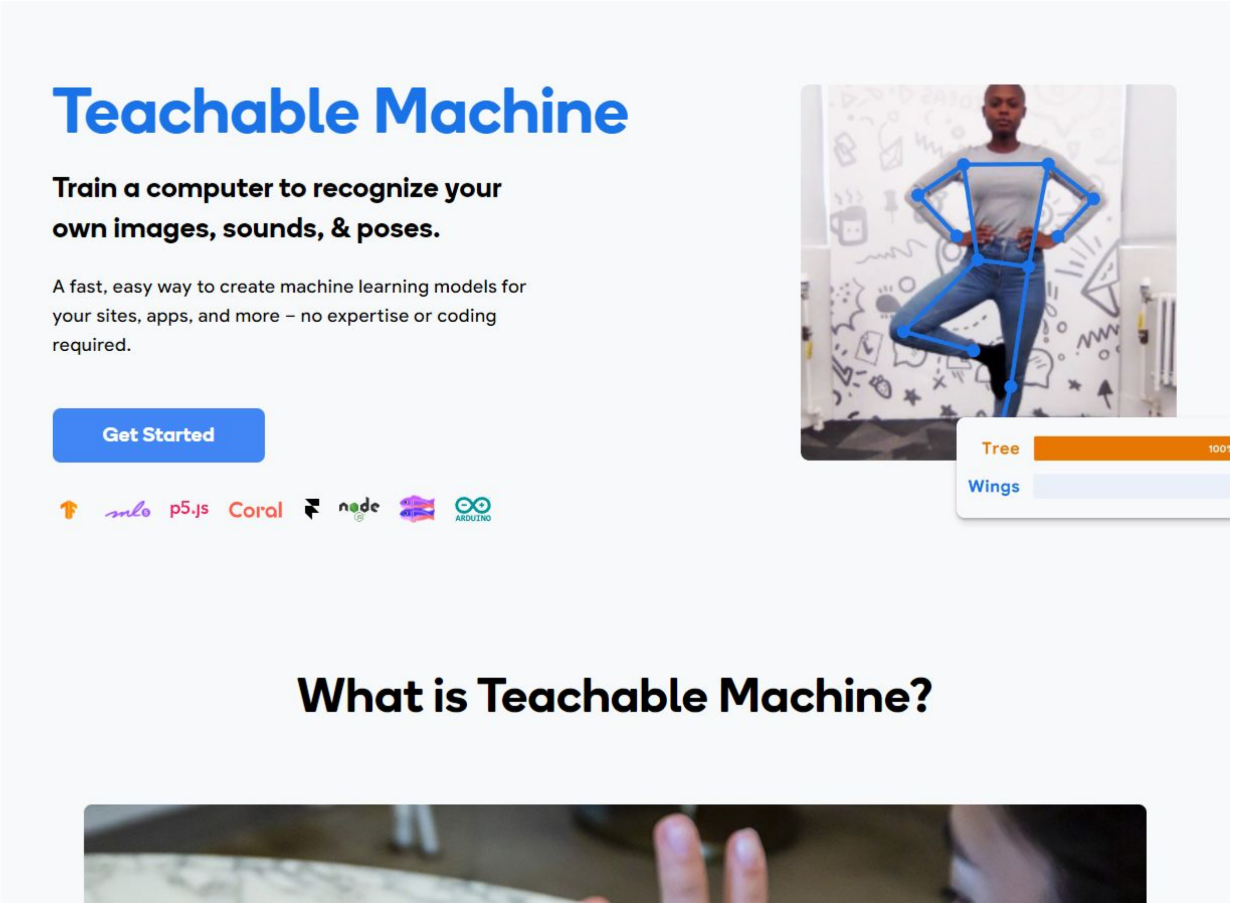
- KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms.
- KNN is a non-parametric, lazy learning algorithm.
- Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.





Fun Activity

https://teachablemachine.withgoogle.com/



Teachable Machine

Train a computer to recognize your own images, sounds, & poses.

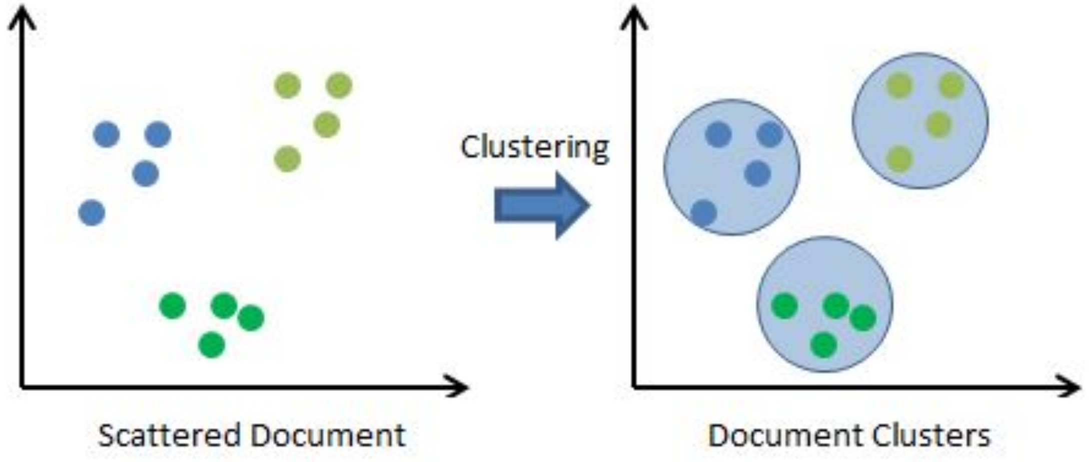
A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.

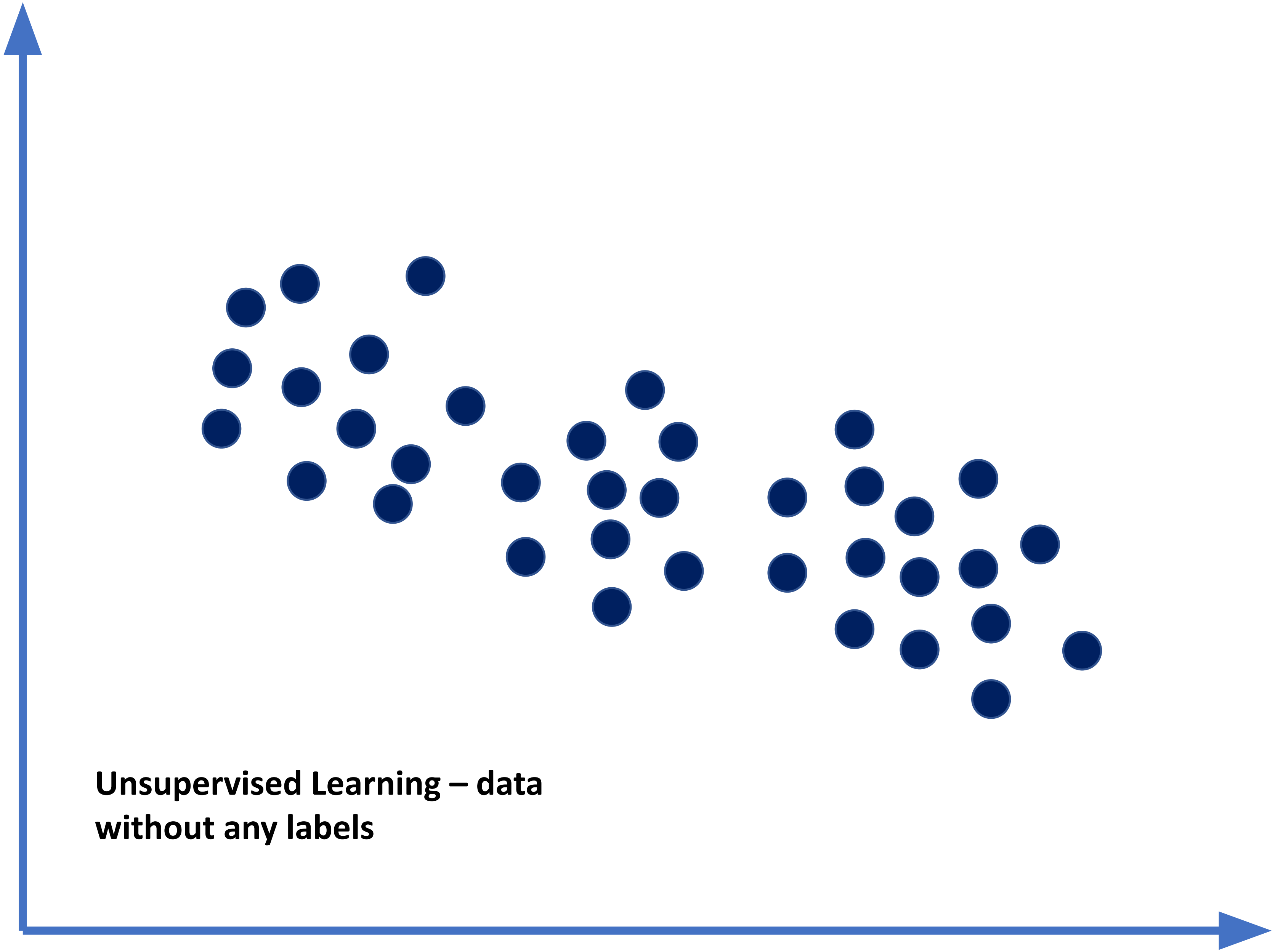
What is Teachable Machine?

Clustering

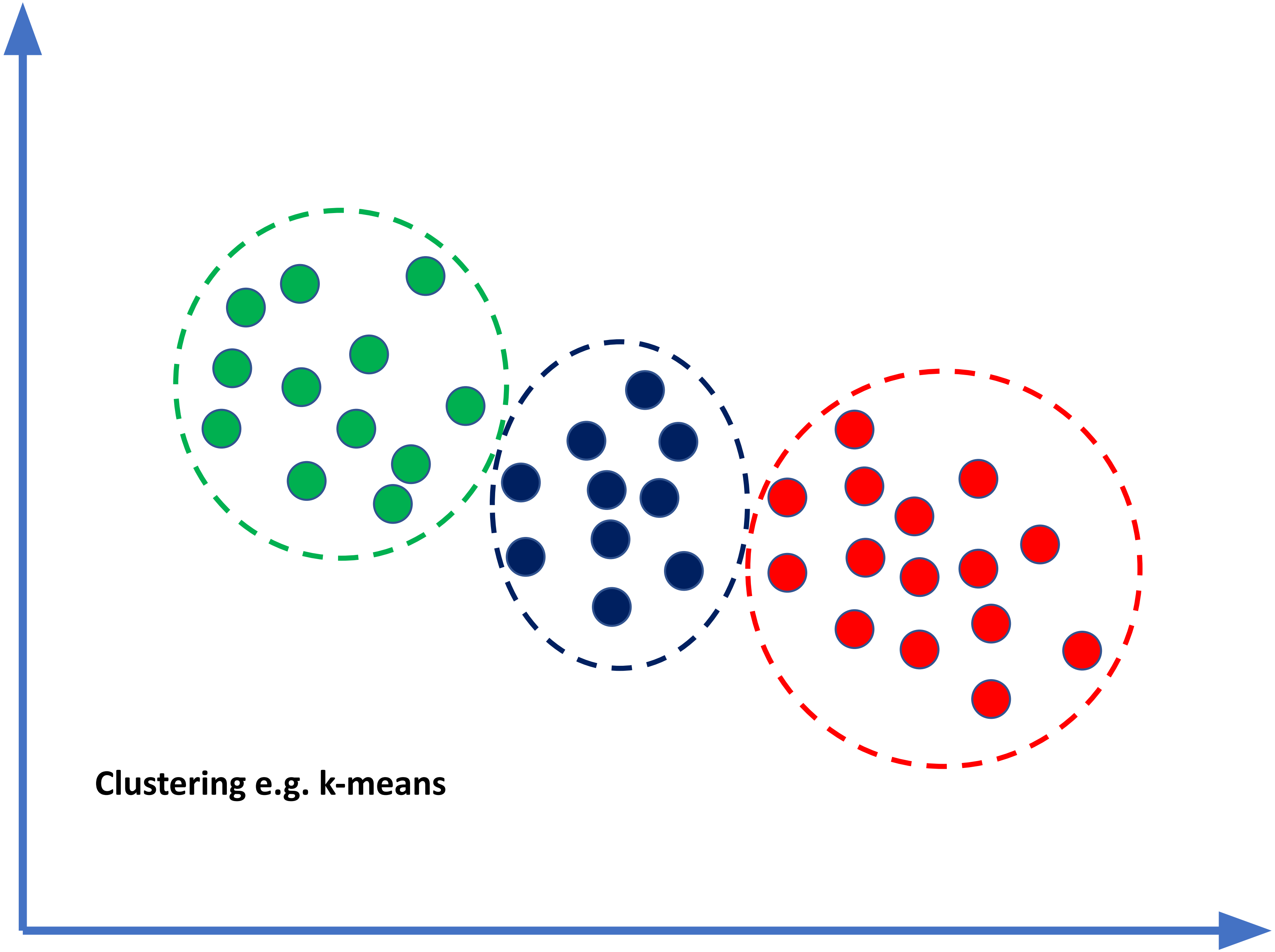
Clustering

- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).





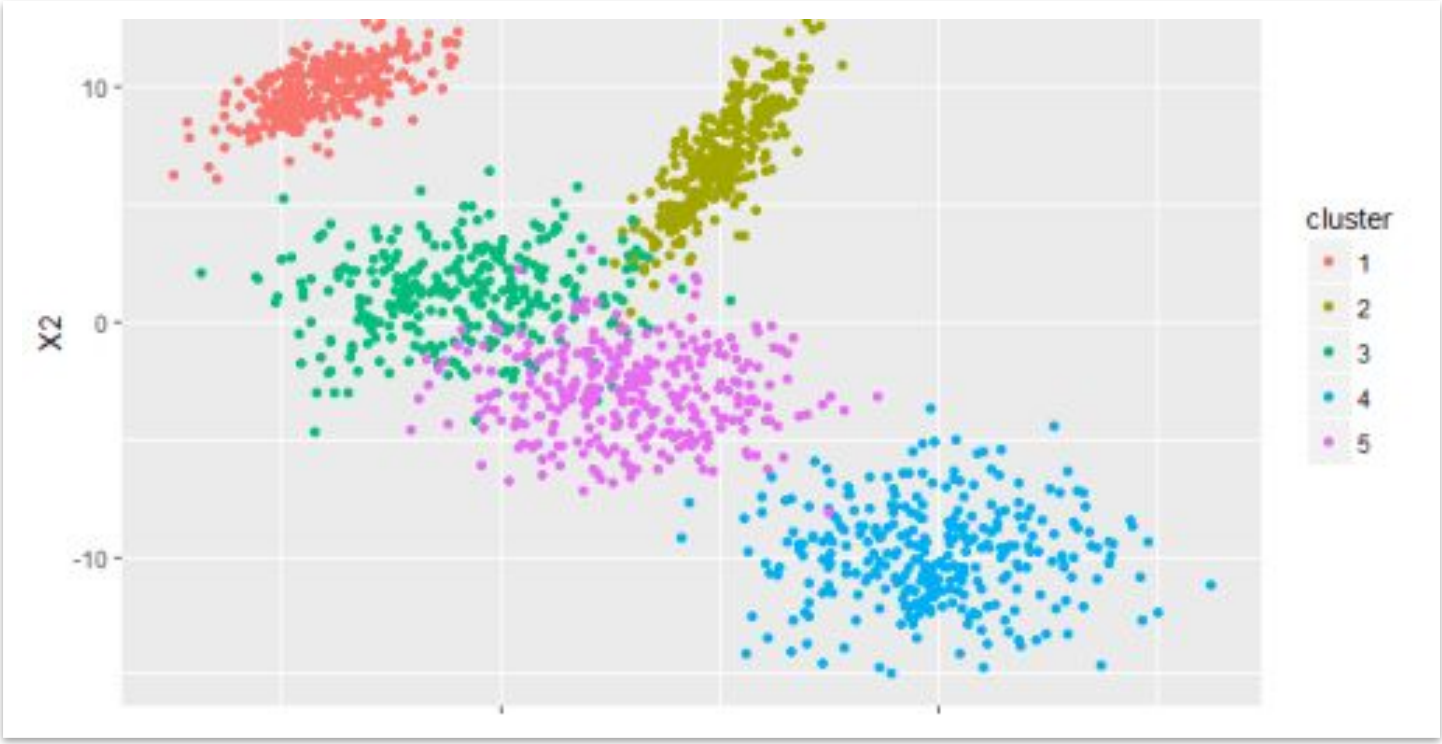
Unsupervised Learning – data without any labels

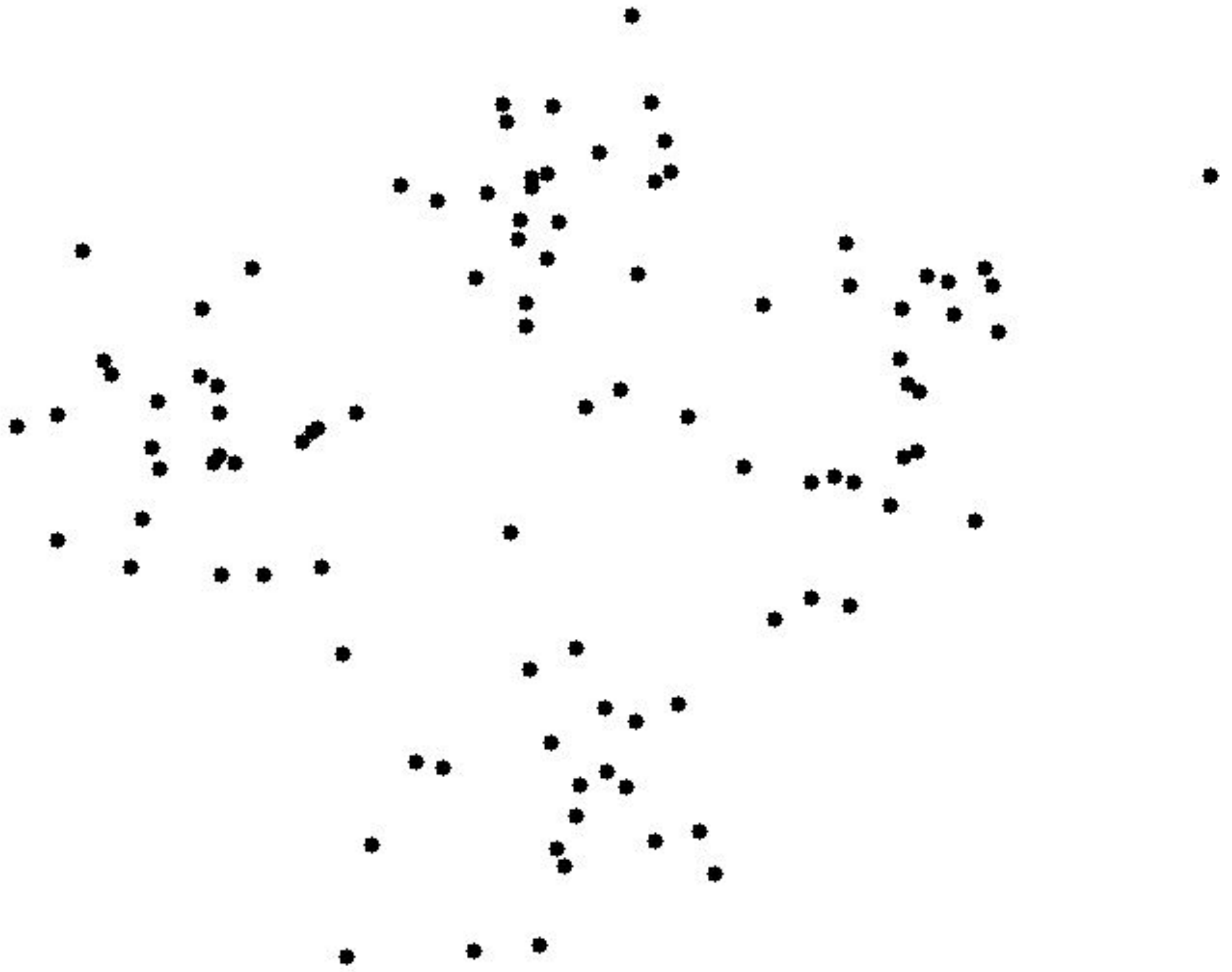


Clustering e.g. k-means

K-means

- k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



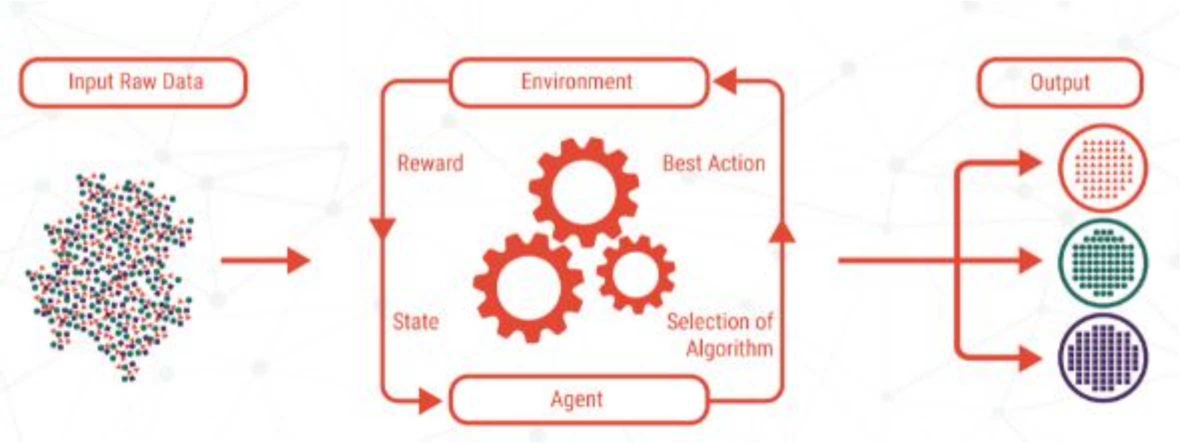


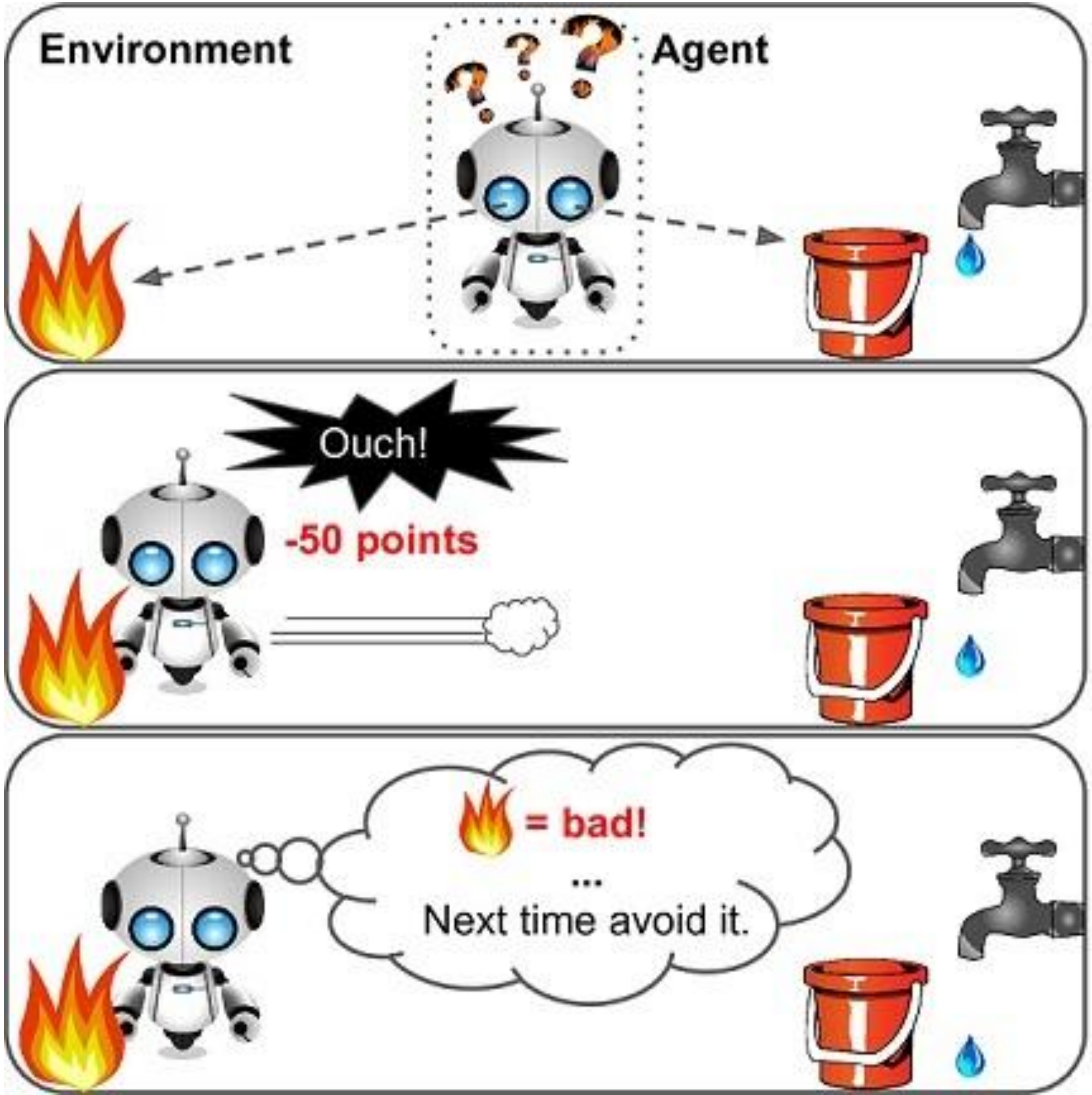
k-means algorithm

Reinforcement Learning

Reinforcement Learning

- Reinforcement learning (RL) is an area of machine learning inspired by behaviourist psychology concerning with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

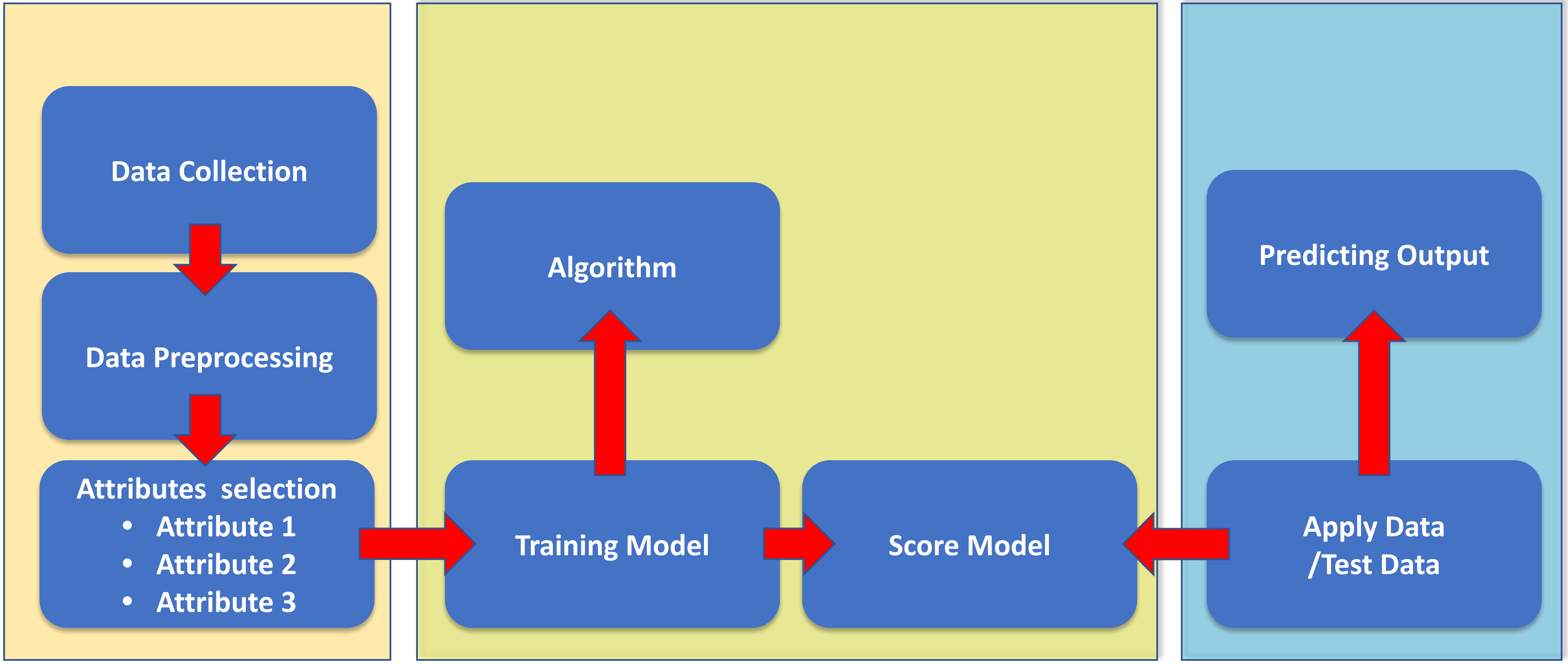




1. Observe
2. Select action using policy
3. Action!
4. Get reward or penalty
5. Update policy (learning step)
6. Iterate until an optimal policy is found

ima...
021 3 1

Machine Learning Framework



Initialization Step

Learn Step

Apply Step

Machine Learning Framework

1. Collect Data

Data is very important as we need to feed it into our model.

Resources:

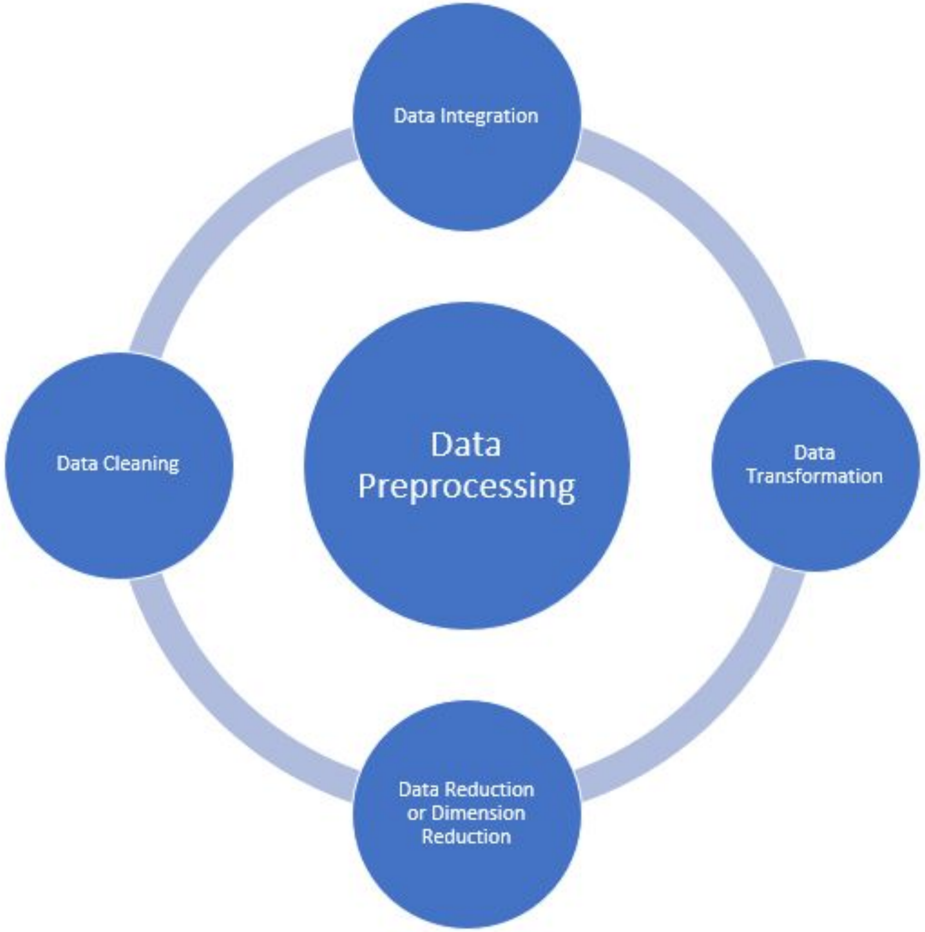
- UCI machine learning dataset
- Kaggle
- Google Dataset Search
- Data.Gov

2. Preprocess Data

- Transforms data into an understandable and readable format.
- Make prediction /result accurate!

Examples of Preprocessing

1. Handling the missing values
2. Deal with outliers
3. Split dataset
4. Feature scaling



3. Choose a model

Different algorithm for different task

Model types:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

What task? Predict house price, filter spam, ...

4.Model training

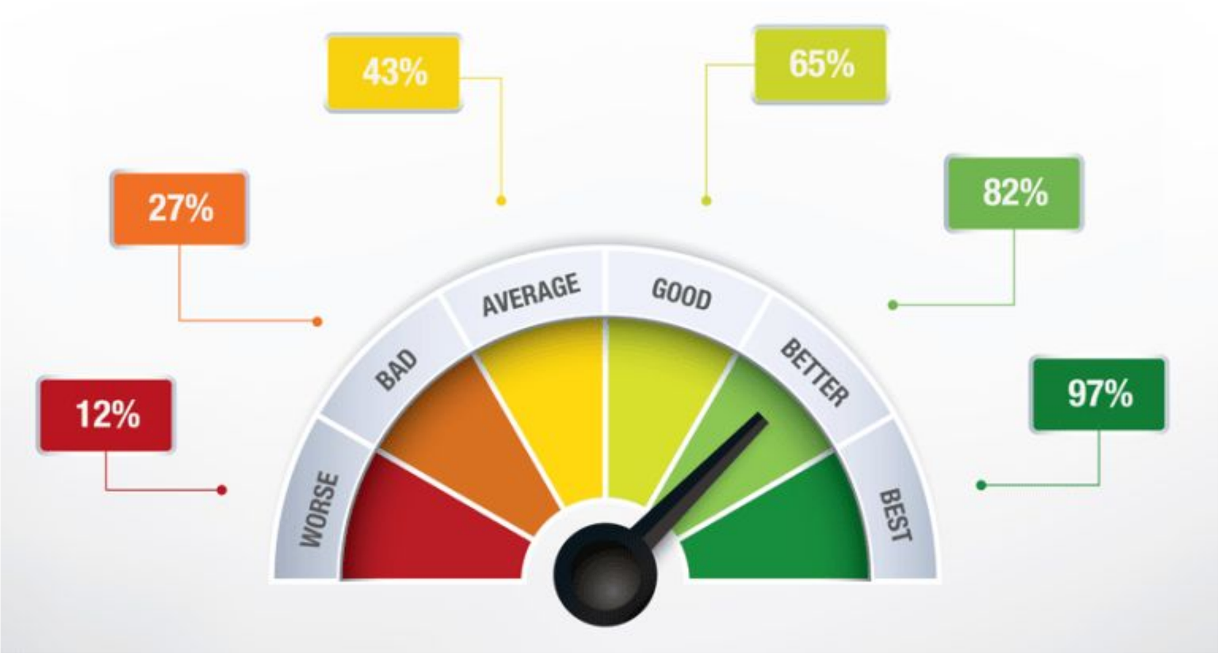
It’s time to feed in your data!

Import the model and train it!

5.Model evaluation

Check the performance of our model

By comparing the prediction result with the test set value



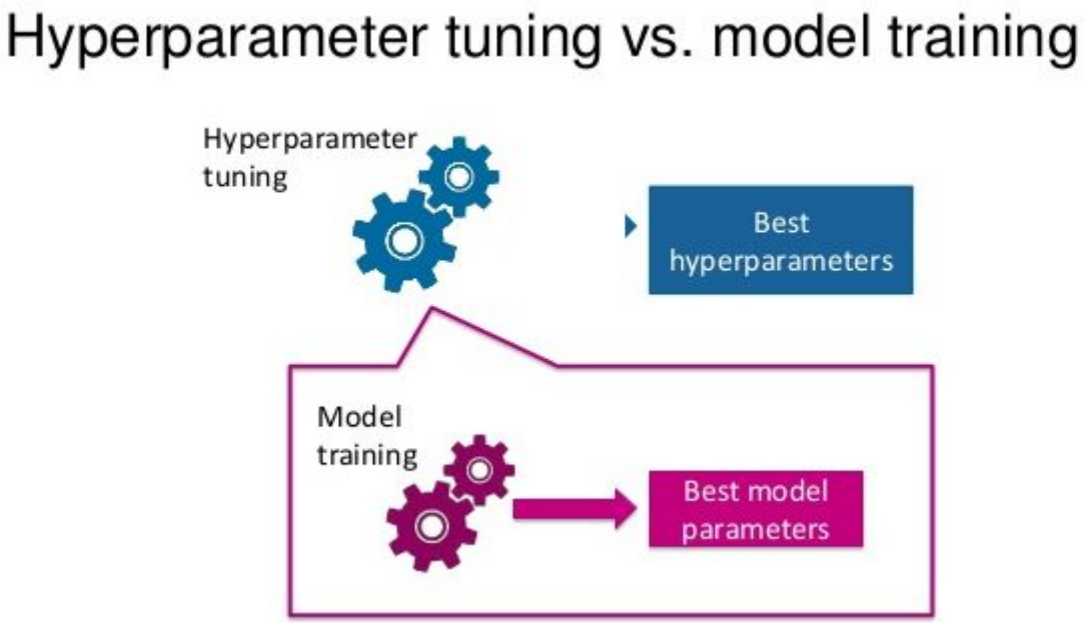
6.Tune model

Improve model performance

Hyperparameter tuning

(control behavior of a machine learning model)

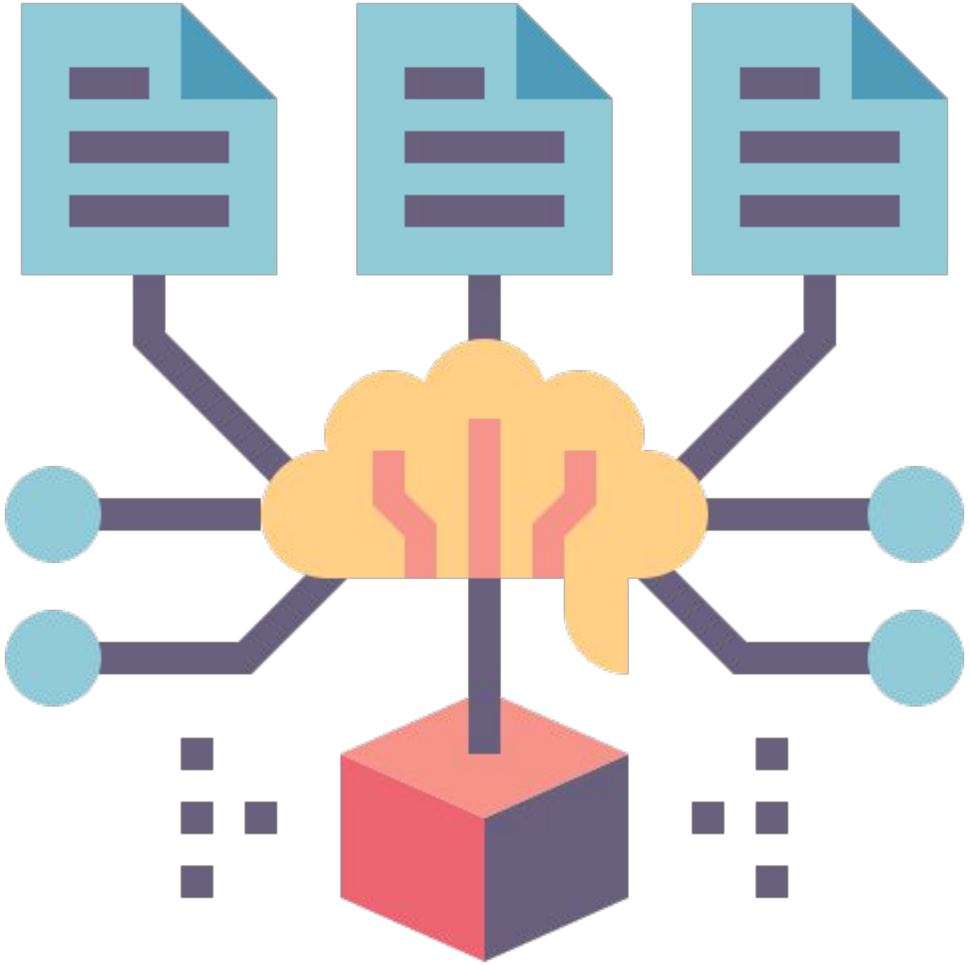
- Number of leaves in decision tree
- Initialization values
- Number of k in k-Nearest Neighbour



7. Make Prediction

Test your freshly built model!

Make prediction using our test set



Python Libraries

A set of useful functions that eliminate the need for writing codes from scratch

NumPy

- Scientific computation, large multi-dimensional array and matrix processing
- Large collection of high-level mathematical functions

SciPy

- Data manipulation
- Contain different modules for optimization, linear algebra, statistics, integration and image manipulation

Pandas

- Data Manipulation, data extraction, and data analysis
- Inbuilt methods for grouping, combining and filtering data

Scikit-learn

- Most popular ML libraries for classical ML algorithms
- Contains efficient tools for machine learning and statistical modeling

Matplotlib

1. Data visualization and graphical plotting
2. Consist of several plots like line graph, bar chart, scatter and histogram

TensorFlow

- High performance numerical computation involving tensors
- Widely used in deep learning research and AI application

Keras

- High-level neural networks API capable of running on top of TensorFlow
- Allows for easy and fast prototyping

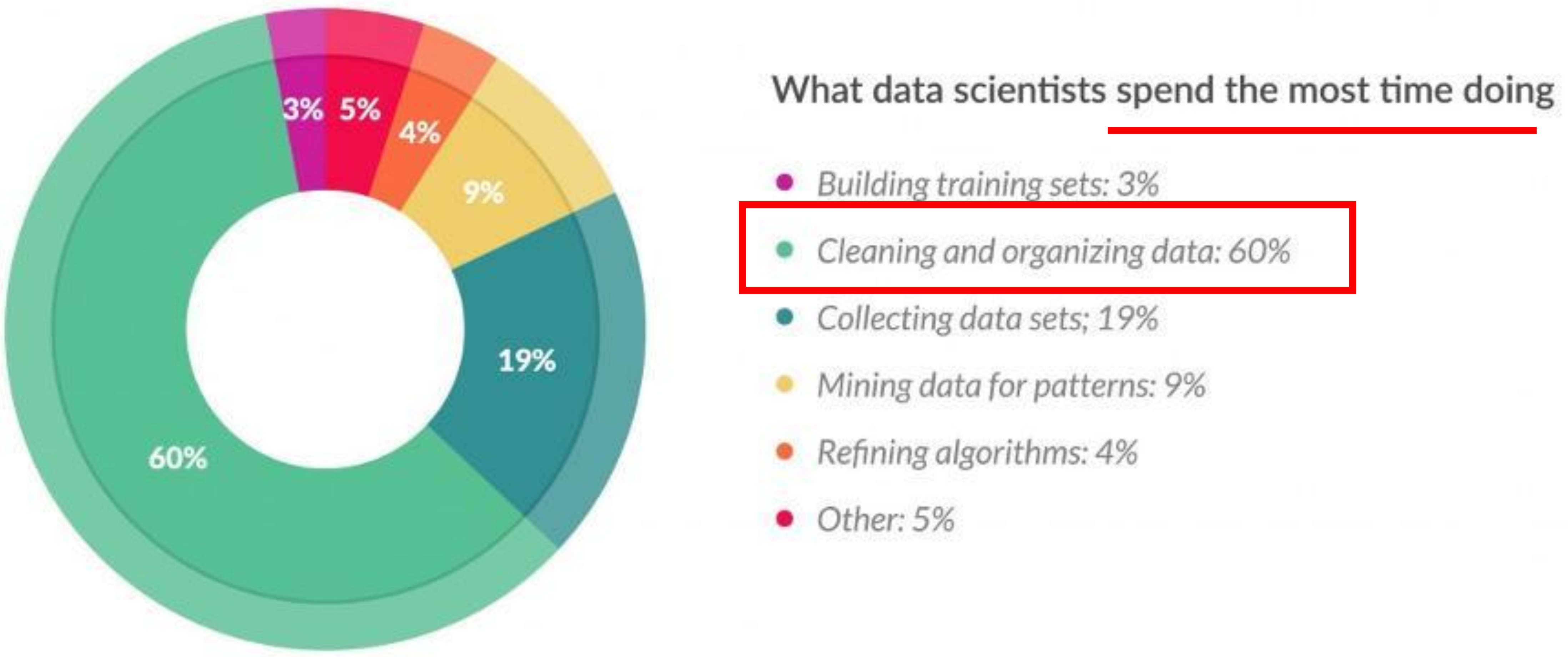
PyTorch

- Supports on Computer Vision, Natural Language Processing(NLP) and many more ML programs
- Helps in creating computational graphs

Congratulations !

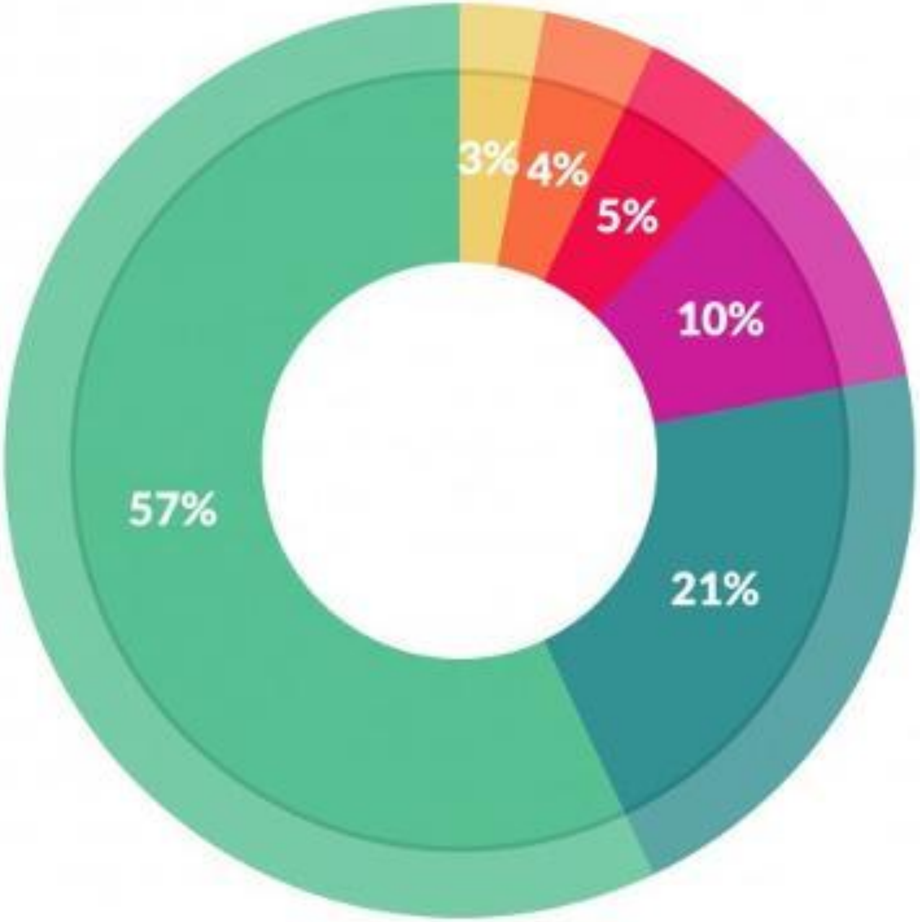
You Had Comprehended The Basic Understanding Of Machine Learning !

General Tips



What data scientists spend the most time doing

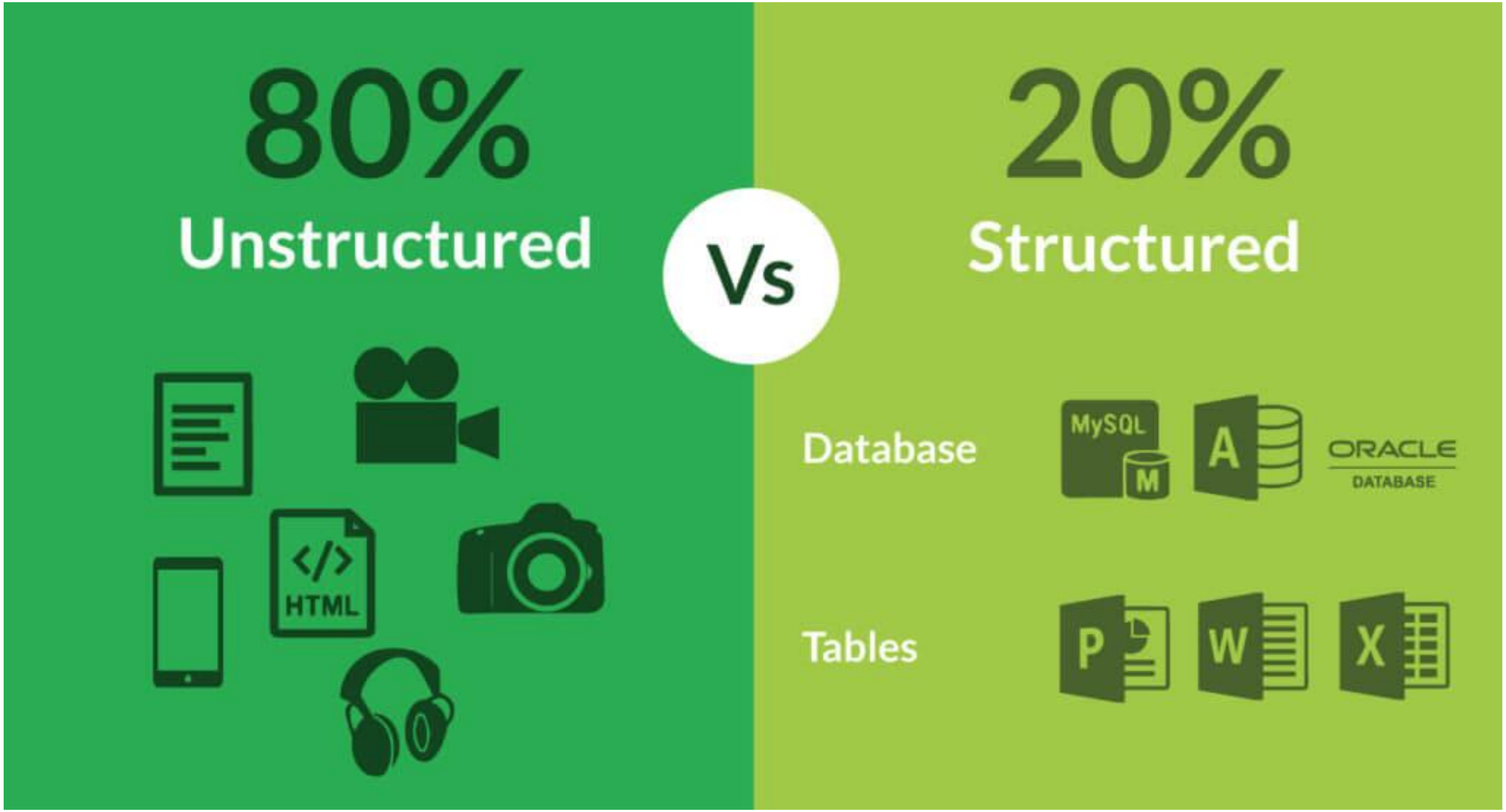
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#f37c7f758459



What's the least enjoyable part of data science?

- ● Building training sets: 10%
- ● Cleaning and organizing data: 57%
- ● Collecting data sets: 21%
- ● Mining data for patterns: 3%
- ● Refining algorithms: 4%
- ● Other: 5%

http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#f37c7f758459



80%
Unstructured

Vs

20%
Structured

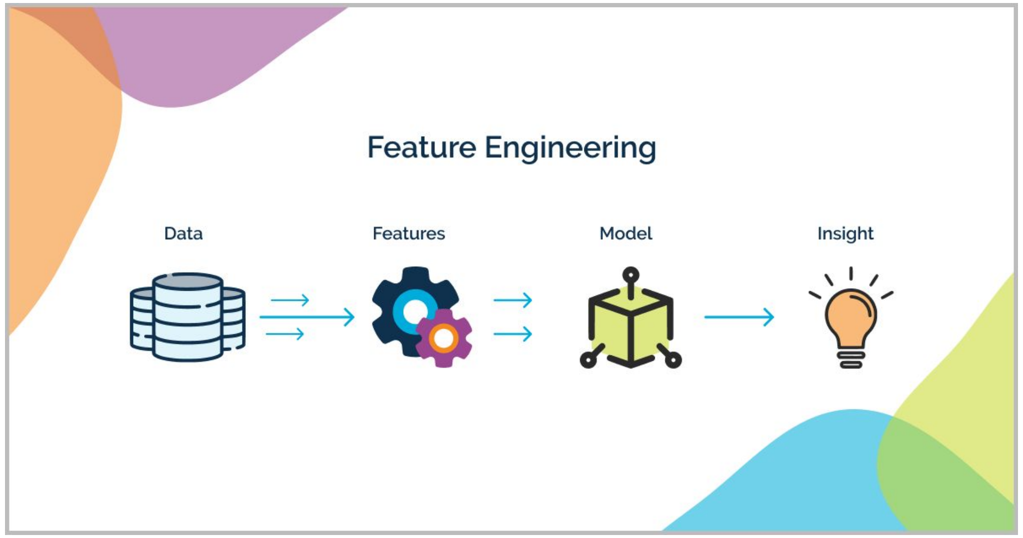
Database

Tables

1. Feature Engineering

Two main goals:

1. Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
2. Improving the performance of machine learning models.



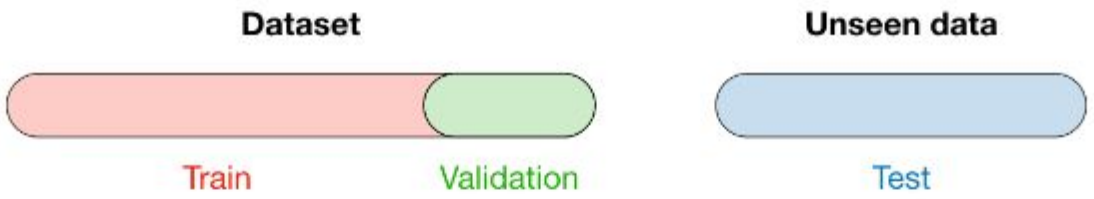
List of Techniques

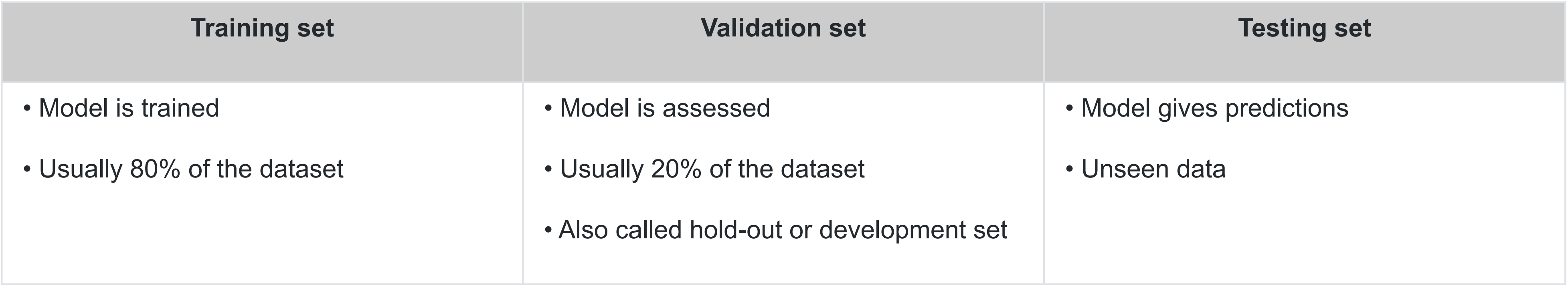
1. Imputation
2. Handling Outliers
3. Binning
4. Log Transform
5. One-Hot Encoding
6. Grouping Operations
7. Feature Split
8. Scaling

https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

2. Model Selection

When selecting a model, we distinguish 3 different parts of the data that we have as follows:





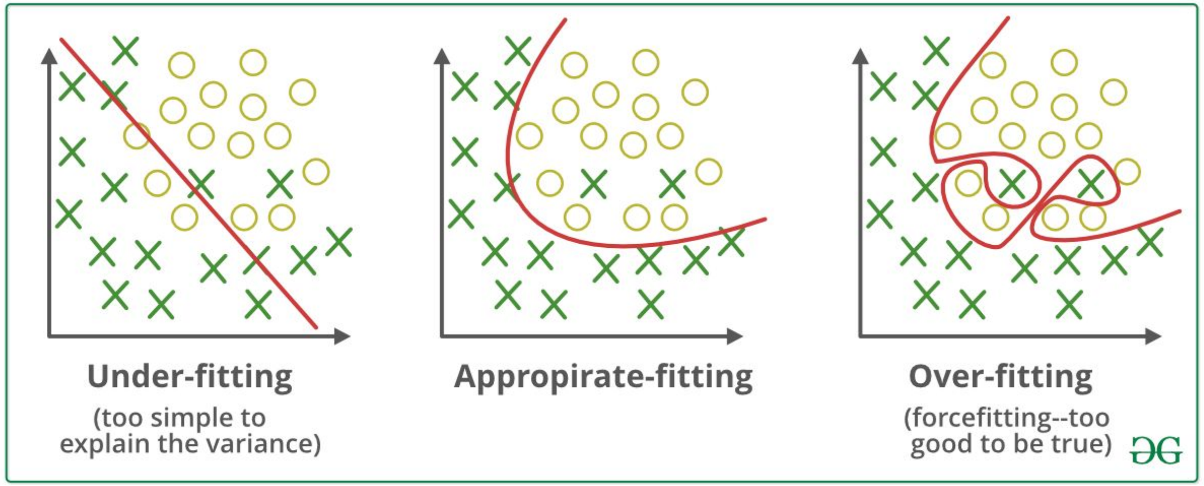
Training set	Validation set	Testing set
• Model is trained • Usually 80% of the dataset	• Model is assessed • Usually 20% of the dataset • Also called hold-out or development set	• Model gives predictions • Unseen data

https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks

3. Overfitting vs Underfitting

- Overfitting refers to a model that models the training data too well.
- Underfitting refers to a model that can neither model the training data nor generalize to new data.

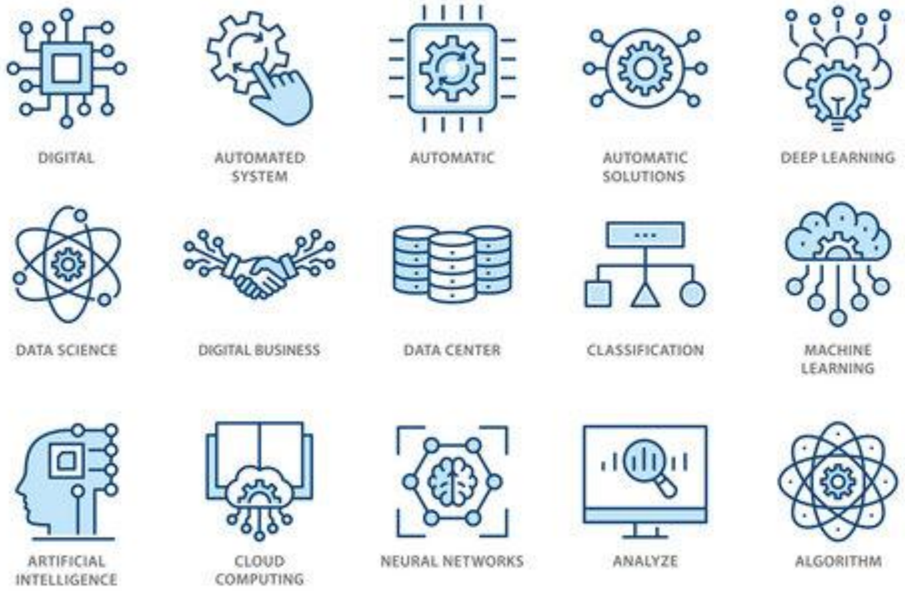
Regularization: procedure aims at avoiding the model to overfit


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance

4. Try simplest model first

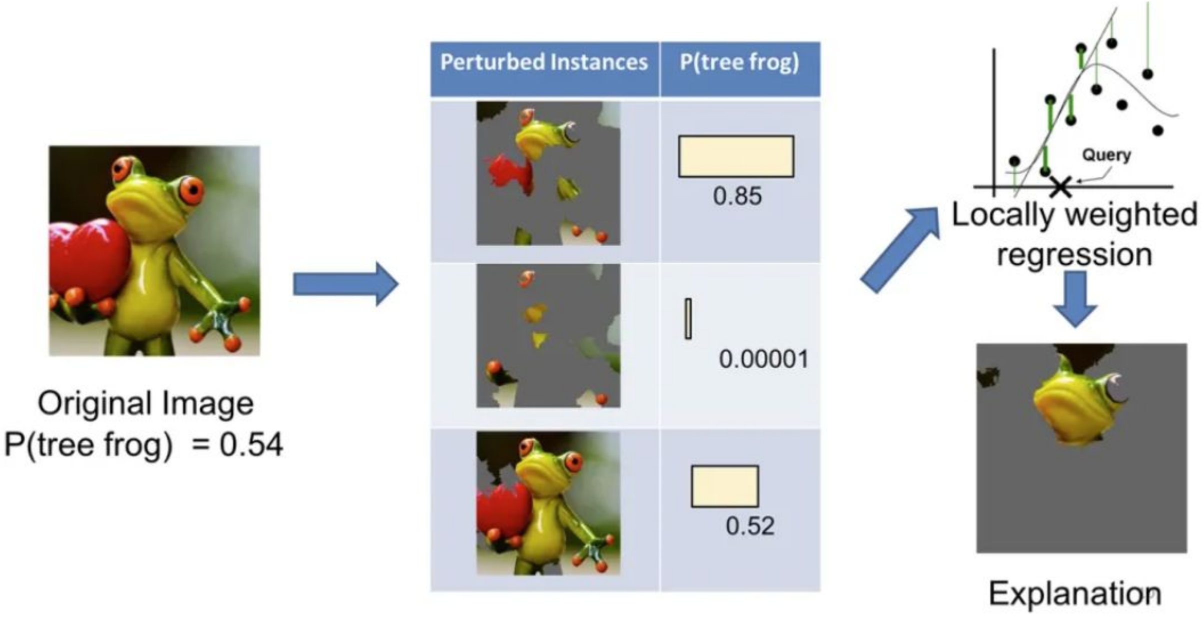
- Always start out with the simplest model as baseline
- Simple model can be executed quickly and provides better estimate
- Identify the trade off between complex models



https://www.forbes.com/sites/forbestechcouncil/2021/08/10/five-reasons-why-simple-models-are-a-data-scientists-best-friend/?sh=6ba857536f89

5. Model explainability

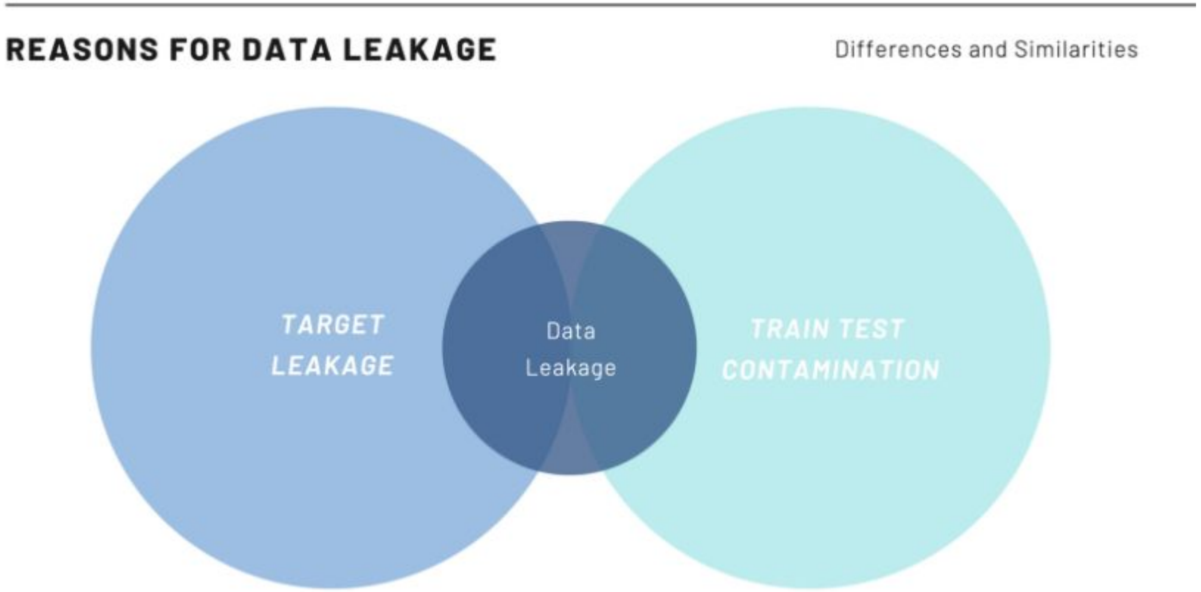
- Choose model that can be understood and easy to be explained to stakeholders
- Blackbox model leads to difficulty in debugging or defining the actual root cause of a problem



https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools

6. Avoid data leakage

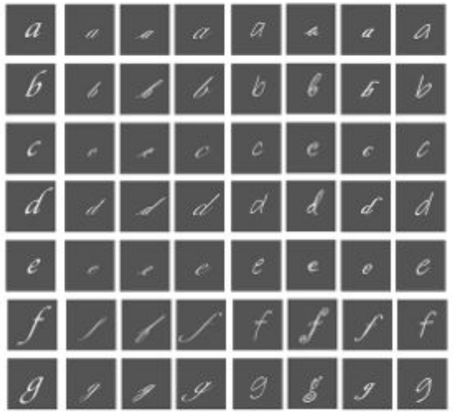
- Data leakage is when information from outside the training dataset is used to create the model.
- eg. usage of certain drugs indicate sickness
- Data leakage can cause you to create overly optimistic if not completely invalid predictive models.



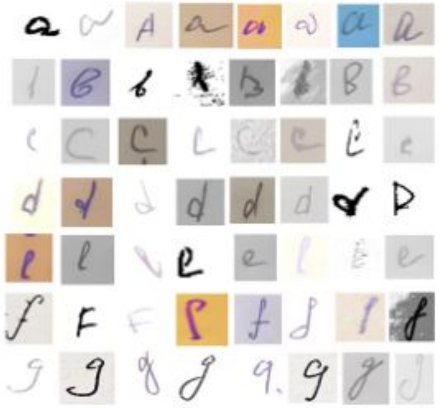
https://towardsdatascience.com/data-leakage-in-machine-learning-how-it-can-be-detected-and-minimize-the-risk-8ef4e3a97562

7. Data drift & Concept Drift

- Data drift occurs when the data a model is trained on changes.
- Data drift is generally a consequence of seasonal changes or changes in consumer preferences over time.
- Concept drift occurs when the model's predicted target or its statistical properties change over time.
- Identify both data drift/concept drift and the need to re-train models



Training data

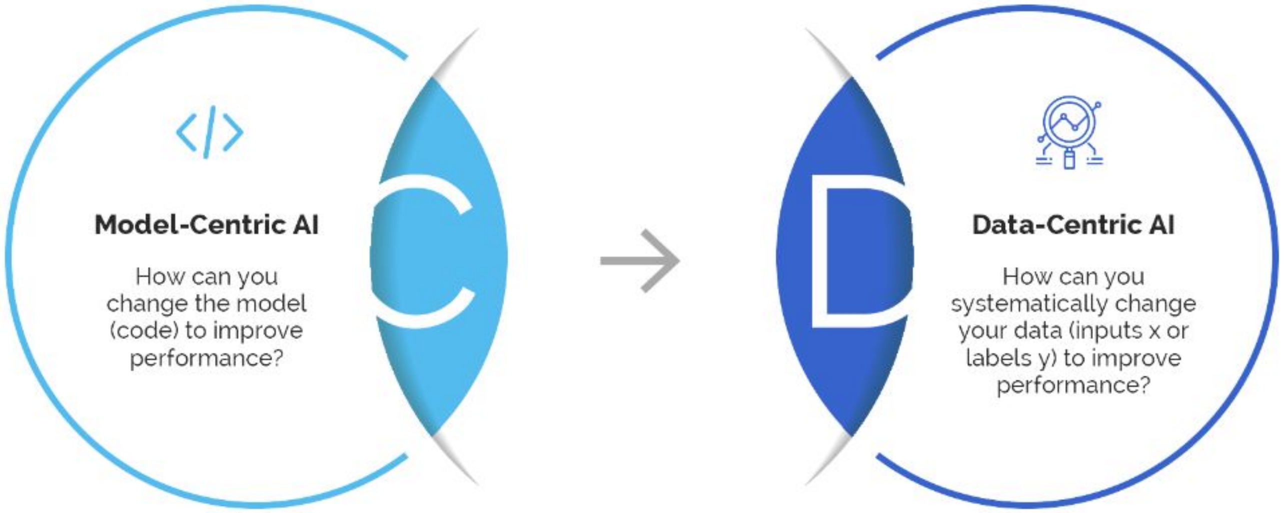


Production data

https://analyticsindiamag.com/concept-drift-vs-data-drift-in-machine-learning/

8. Data Centric vs Model Centric

- Getting the right proper labelled data is more important than choosing the right model
- Good data leads to good result while bad data with extra ordinary model will deliver garbage

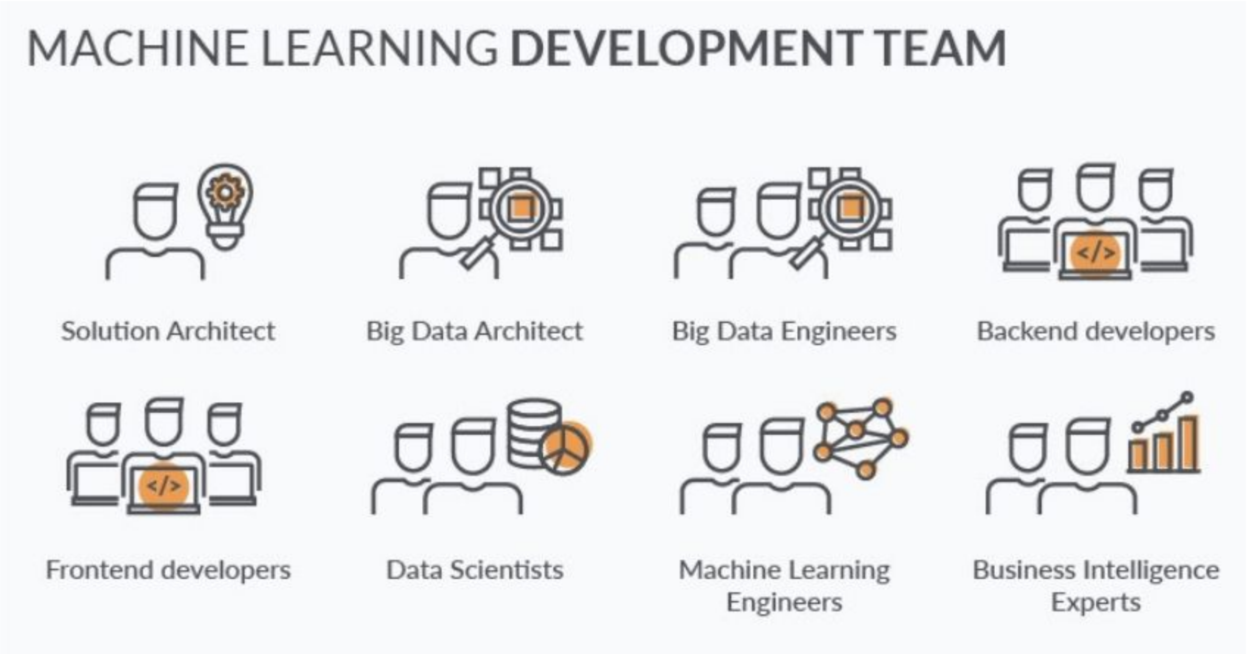


Model-Centric AI
How can you change the model (code) to improve performance?
C
D
Data-Centric AI
How can you systematically change your data (inputs x or labels y) to improve performance?

https://tdan.com/the-data-centric-revolution-data-centric-vs-data-driven/20288

9. ML is team effort

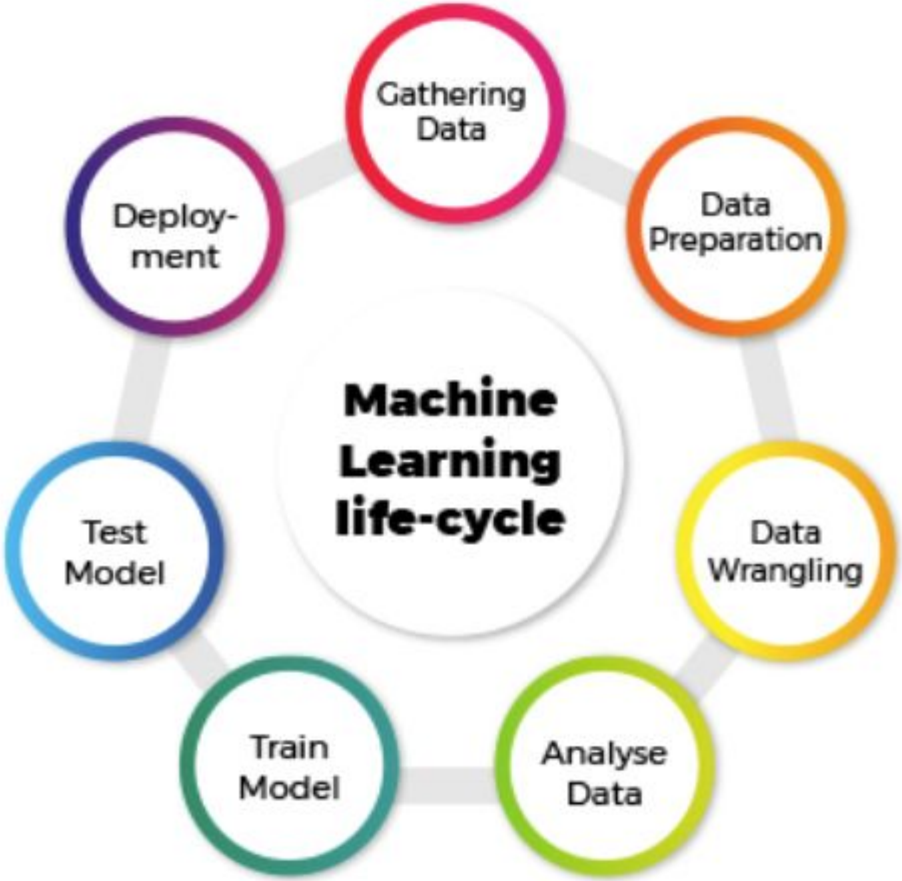
- Machine Learning in commercial world is a team effort
- Consists of Data Scientist, Data Engineer, Machine Learning Engineer, Subject Matter Expert, Dashboard Visualizer etc



https://medium.com/bigdatarepublic/on-machine-learning-team-composition-a9d0d3a3d89

10. Start small

- Machine Learning project is an iterative process
- It takes time to achieve good results
- Start small and measure small success
- Slowly gain experience along the way and expand your ML projects



https://landing.ai/resources/ai-transformation-playbook/

Demo

https://www.kaggle.com/code/pookuanhoong/introduction-to-machine-learning-studyjam