AI Algorithms Fundamentals & Applications
CA6001
Chapter 1

Dr Zhang Jiehuang

College of Computing and Data Science

Nanyang Technological University

email: *jiehuang.zhang@ntu.edu.sg*

Course Learning Outcomes

ILO 1	Explain the principles behind core AI algorithms such as search, optimization, decision-making, and machine learning.
ILO 2	Compare and contrast different AI algorithms in terms of their use cases, strengths, and limitations.
ILO 3	Apply AI algorithms to solve practical problems in domains such as finance, technology, and healthcare.
ILO 4	Analyze real-world datasets to select and justify appropriate AI techniques for specific tasks.
ILO 5	Implement key AI algorithms using programming tools and libraries in hands-on projects.
ILO 6	Evaluate the performance of AI models using appropriate metrics and testing strategies.

Course Structure

10 Nov – 20 Dec 2025, 6 weeks

One pre-recorded lecture to watch before Thursday class

Thursday: Consultative format where you come with questions/doubts

Saturday: Lectures with hands on coding exercises (Bring your laptops and chargers) We will be using the google collab platform

Assessment: 2 Quizzes on week 3 and 6 and 1 assignment

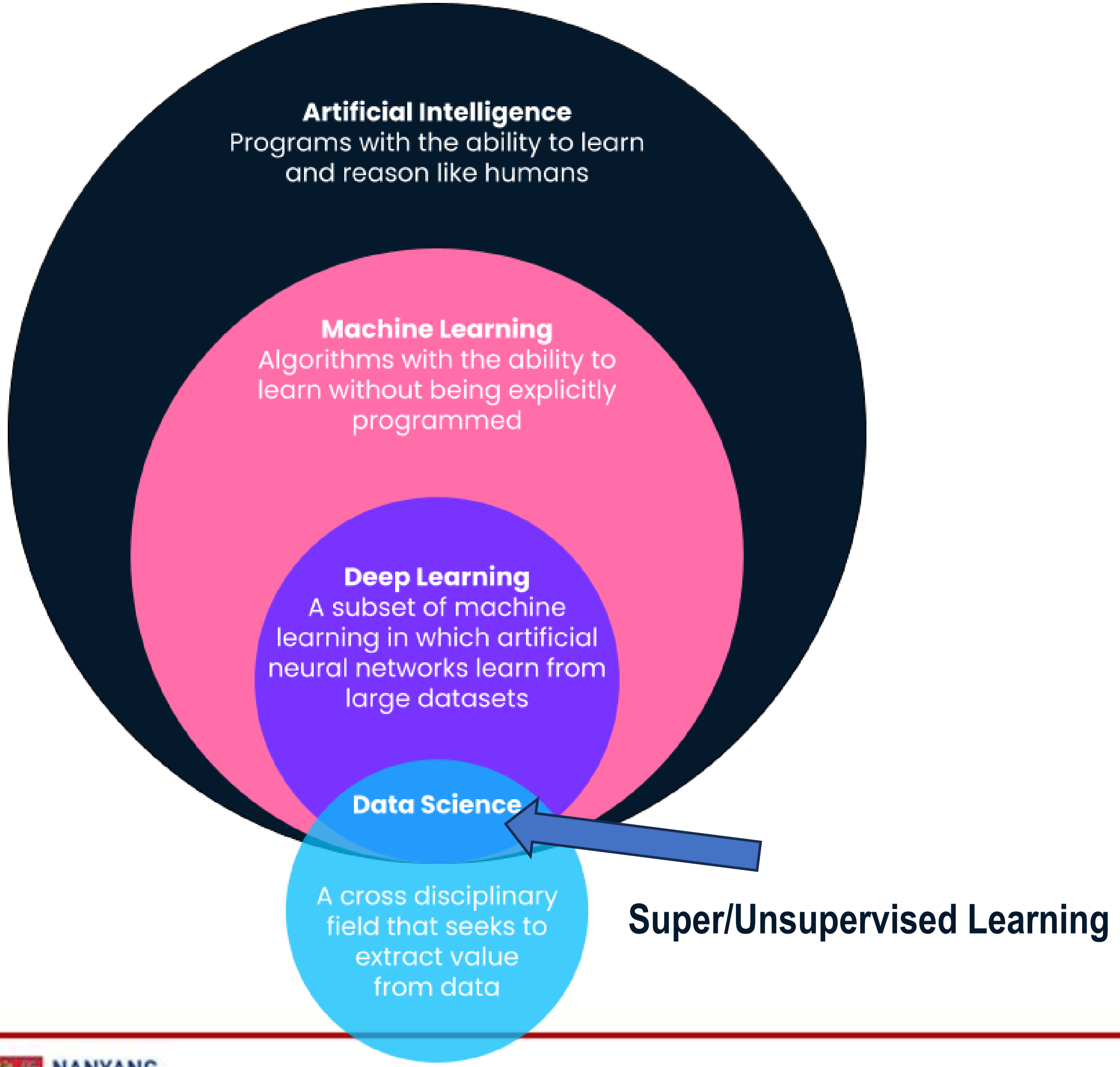
Quiz 1 on 29/11/25 Sat, 1230pm-200pm

Quiz 2 on 20/12/25 Sat, 1230pm-200pm

Chapter 1 – Supervised Vs Unsupervised Learning

1. What is AI/ML and Data Science?
2. Why is Supervised/Unsupervised Learning Important?
3. Supervised Learning: Overview and Examples
4. Linear Regression and Classification
5. Mathematical Model of Linear Regression
6. Unsupervised Learning: Overview and Examples
7. Supervised VS Unsupervised Learning Comparison

What is AI and Data Science?

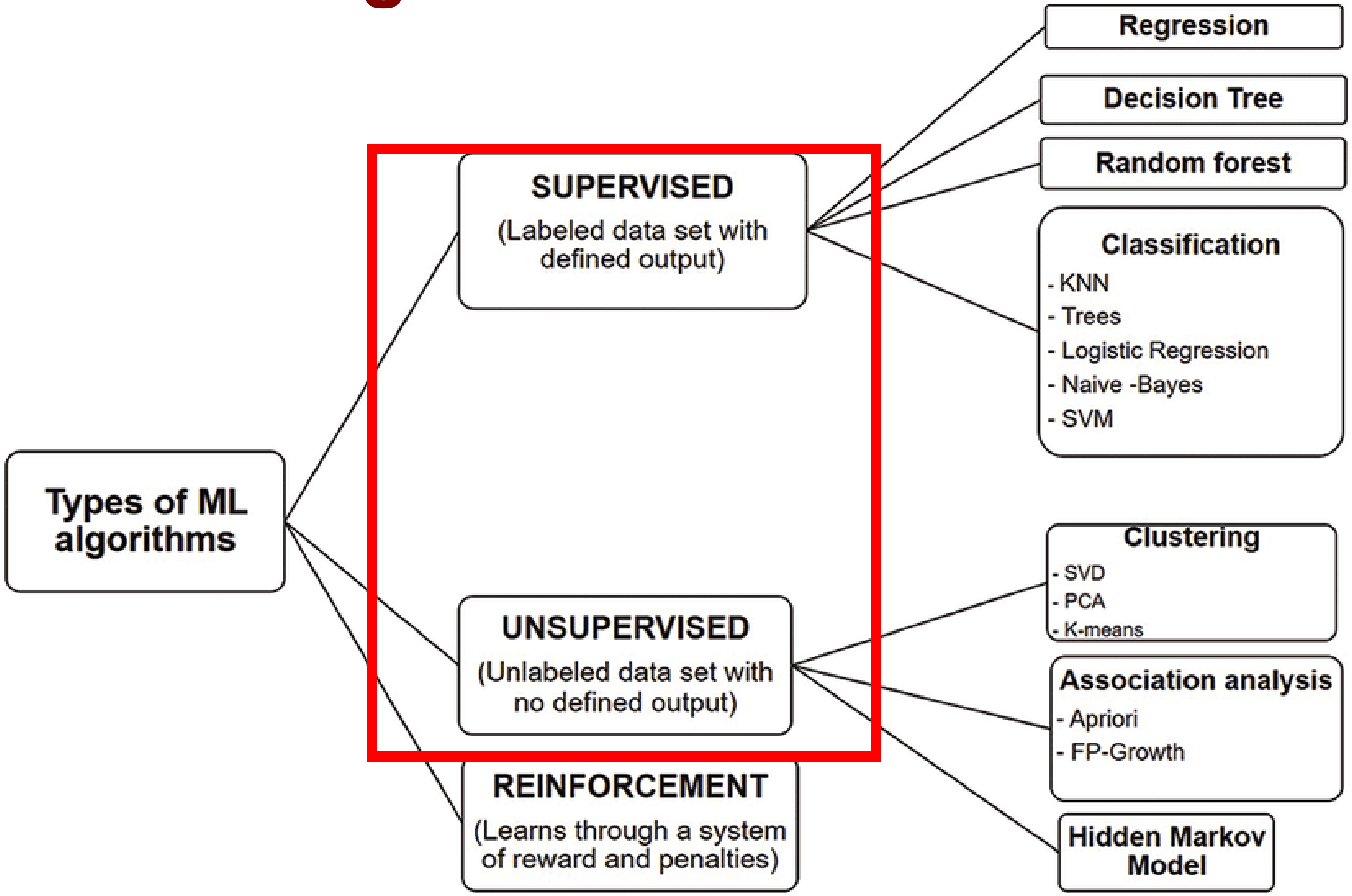


Artificial Intelligence
Programs with the ability to learn and reason like humans
Machine Learning
Algorithms with the ability to learn without being explicitly programmed
Deep Learning
A subset of machine learning in which artificial neural networks learn from large datasets
Data Science
A cross disciplinary field that seeks to extract value from data
Super/Unsupervised Learning

1. Data Science is an interdisciplinary field that aims to use data to create value and insights
2. Data Science overlaps with AI, ML, and Deep Learning
3. Domain where data science and AI is applied to forms the context of any AI project. Domain specific knowledge is usually required
4. That value could be in the form of predictive models that use machine learning,
5. It could also mean surfacing insights with a dashboard or report
6. Super/unsupervised learning is foundational for data science

https://www.datacamp.com/blog/top-machine-learning-use-cases-and-algorithms

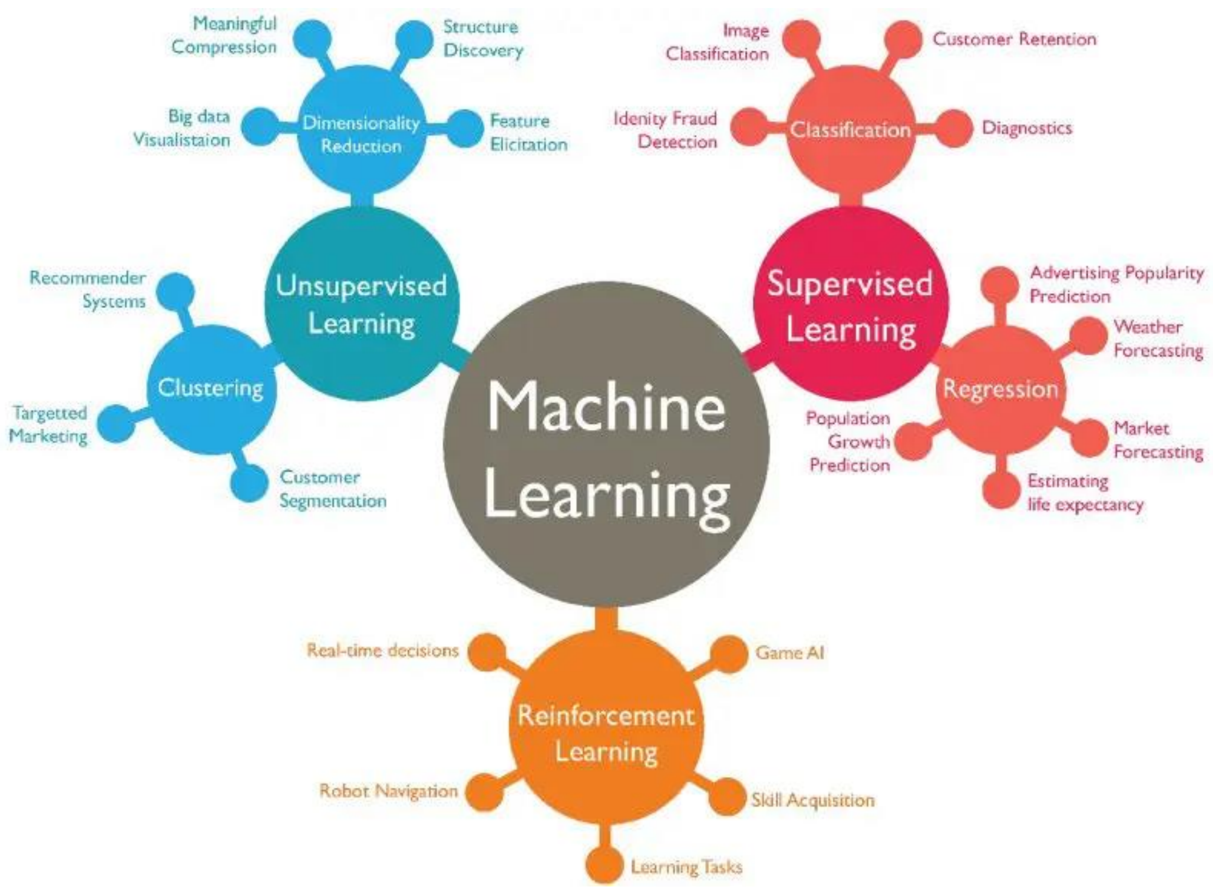
What Types of ML Algorithms are there?



Abbreviations: KNN: k-nearest neighbour; SVM: Support Vector Machine; SVD: Singular Value Decomposition; PCA: Principal Component Analysis; FP: Frequent pattern

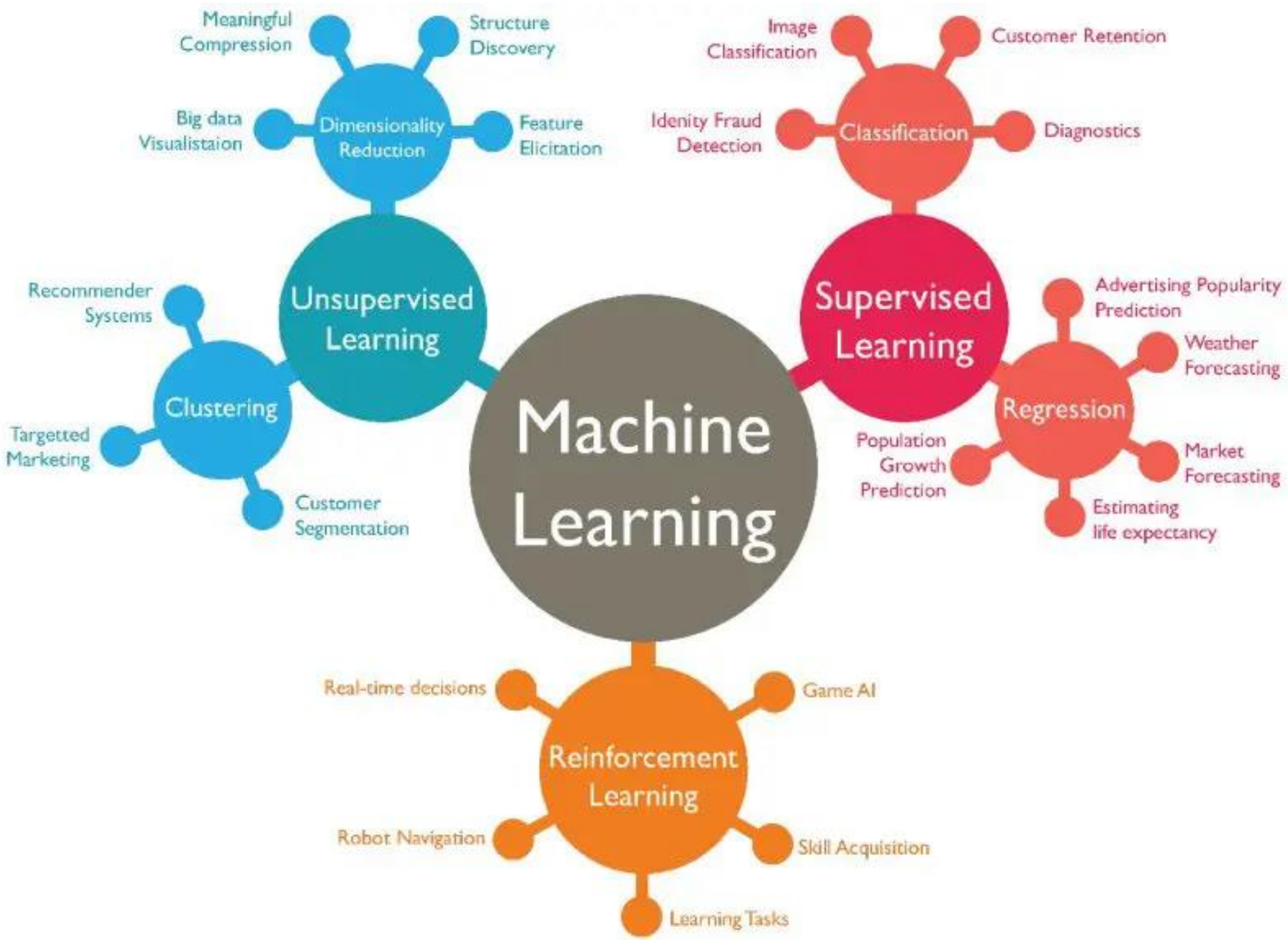
https://www.researchgate.net/publication/351021675_Artificial_intelligence_in_cancer_diagnostics_and_therapy_Current_perspectives

Why Is Super/Unsupervised Learning Important

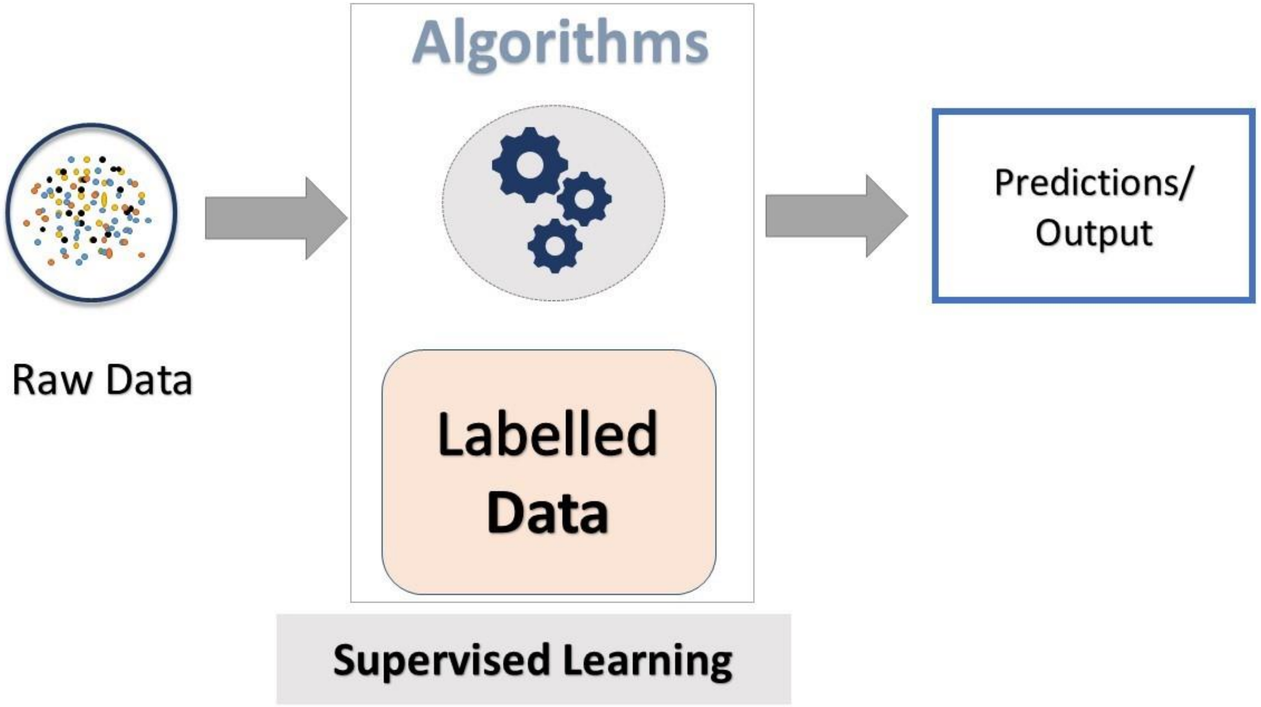


1. Essential Learning for new Data Scientist/AI Engineer as they form the foundations of ML
2. Super/unsupervised learning is the foundation for many more advanced techniques in AI
3. Allows you to solve a wide range of problems
4. Real World Relevance where ML is applied in many fields
5. Essential to guide business and research decision making at senior levels
6. Flexibility to apply across difference industries and domains from
   - finance
   - healthcare
   - marketing

https://www.datacamp.com/blog/top-machine-learning-use-cases-and-algorithms

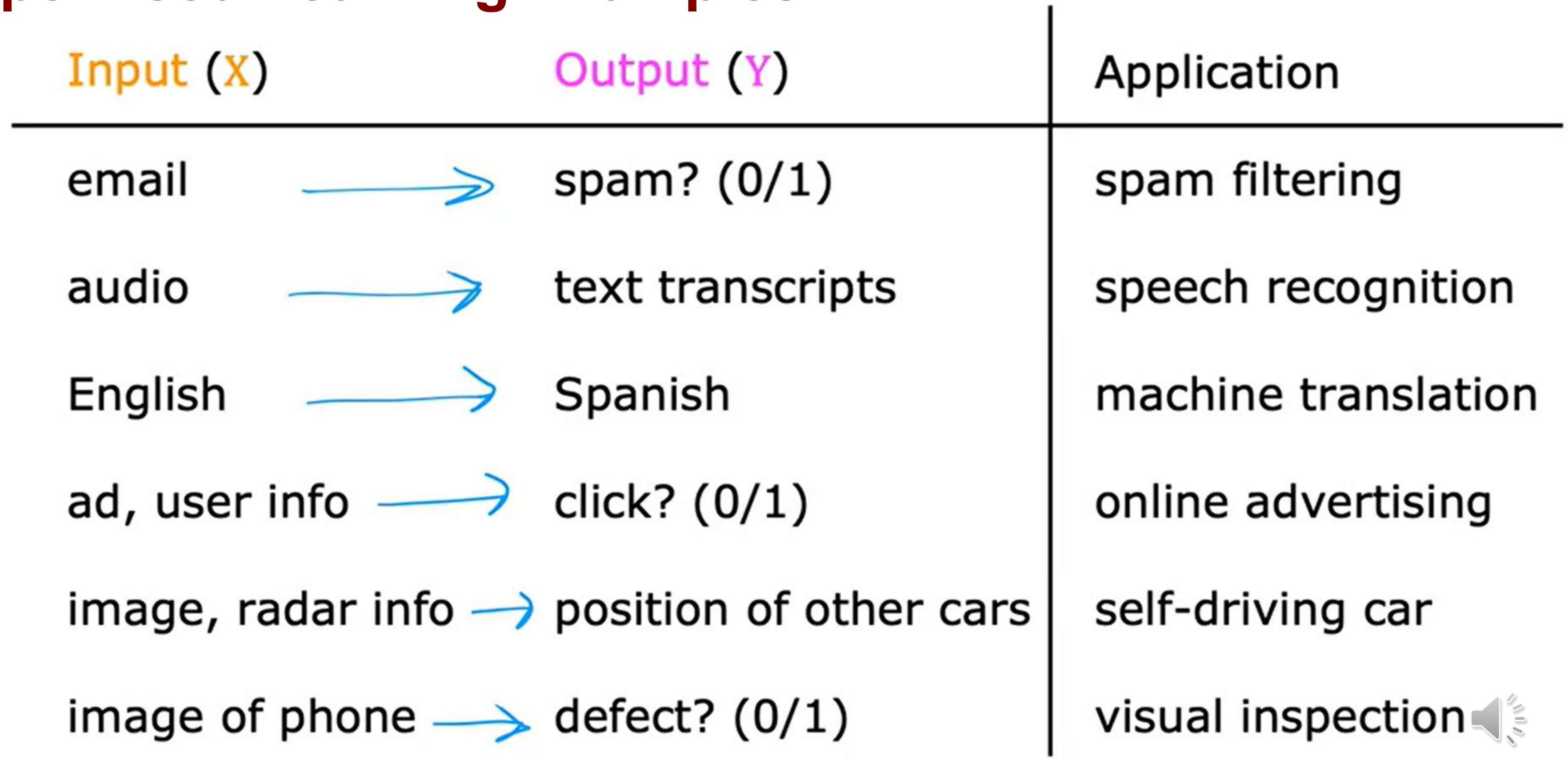

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Supervised Learning: Overview



- We have a dataset with features X and labelled target Y
- Goal: to create a mathematical model that estimates the relationship between X and Y
- Supervised Learning is the AI model learning from ‘correct’ examples in data (ground truth)
- Raw data consist of features X, and predictions Y
- AI model learns how X arrives at Y and gives the ‘best fit’ model
- In other words, model learns the relationship between X and Y

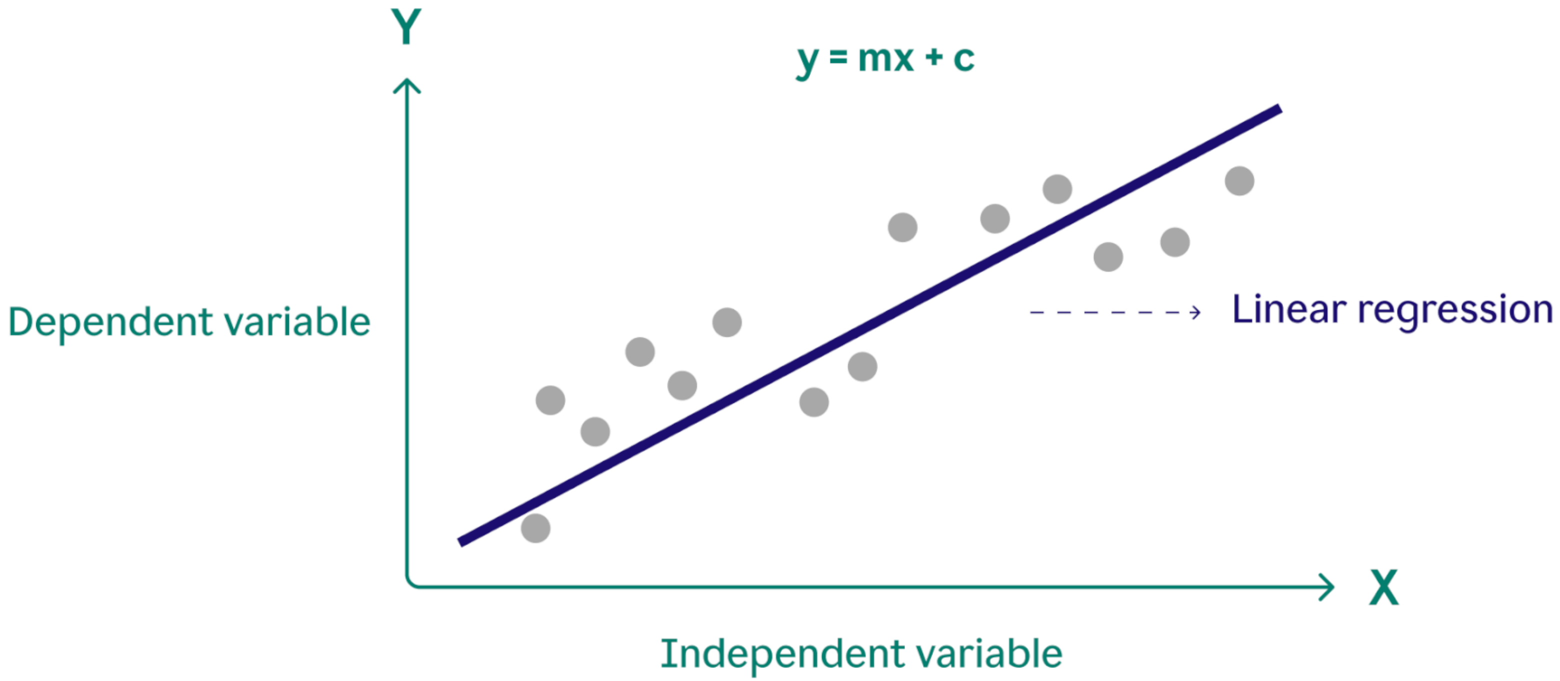
Supervised Learning Examples



Input (X)		Output (Y)	Application
email	→	spam? (0/1)	spam filtering
audio	→	text transcripts	speech recognition
English	→	Spanish	machine translation
ad, user info	→	click? (0/1)	online advertising
image, radar info	→	position of other cars	self-driving car
image of phone	→	defect? (0/1)	visual inspection

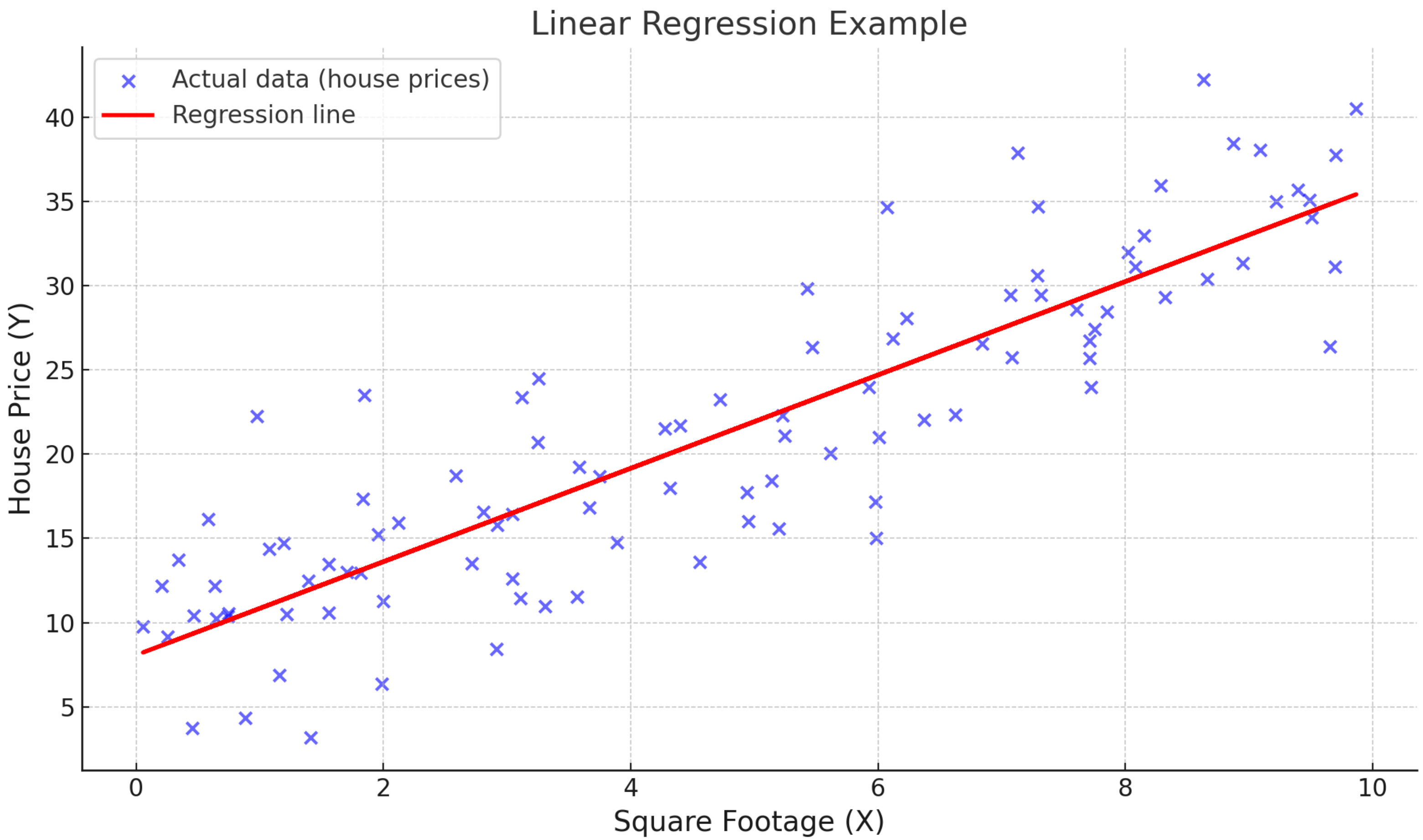
https://www.coursera.org/learn/machine-learning/lecture/Q8Vvp/supervised-learning-part-2

Supervised Learning: Linear Regression Overview



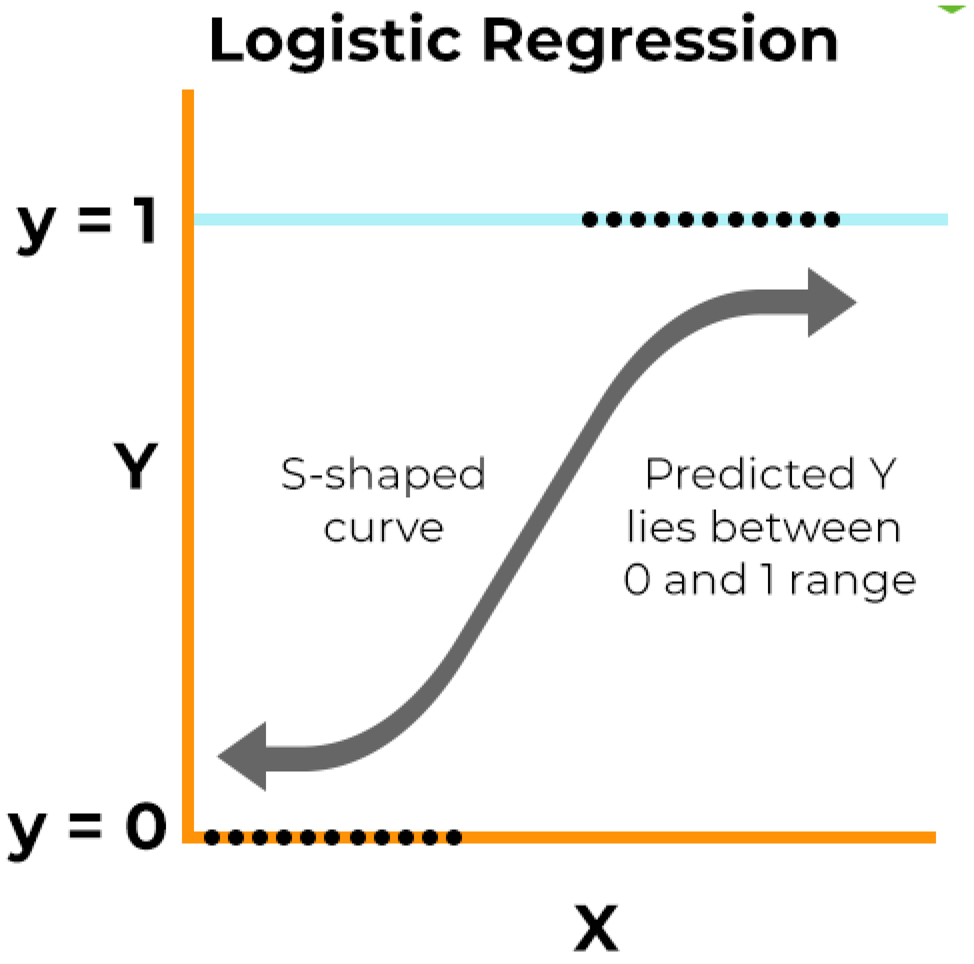
- Linear Regression is finding the ‘best fit line’ for our dataset
- Independent variable X is the input feature, while dependent variable Y is the output
- X can be house size (sqft), Y is price $
- X is a continuous variable that has a linear relationship with Y
- Using the best fit line, we can determine Y using the X values

Supervised Learning: Linear Regression Example



- Blue Points: Actual data points representing houses with varying square footage and their corresponding prices.
- Red Line: The regression line, which shows the predicted relationship between square footage (X) and house price (Y).
- The line minimizes the vertical distances (errors) between the actual data points and the line. It represents the best-fit line that predicts house price based on square footage.

Supervised Learning: Logistic Regression



Similar to linear regression, logistics regression needs to find the relationship that represents x and y. However in this case, the relationship is non linear

Logistic Regression uses a logistic function called sigmoid function to convert a value to a range between 0 and 1

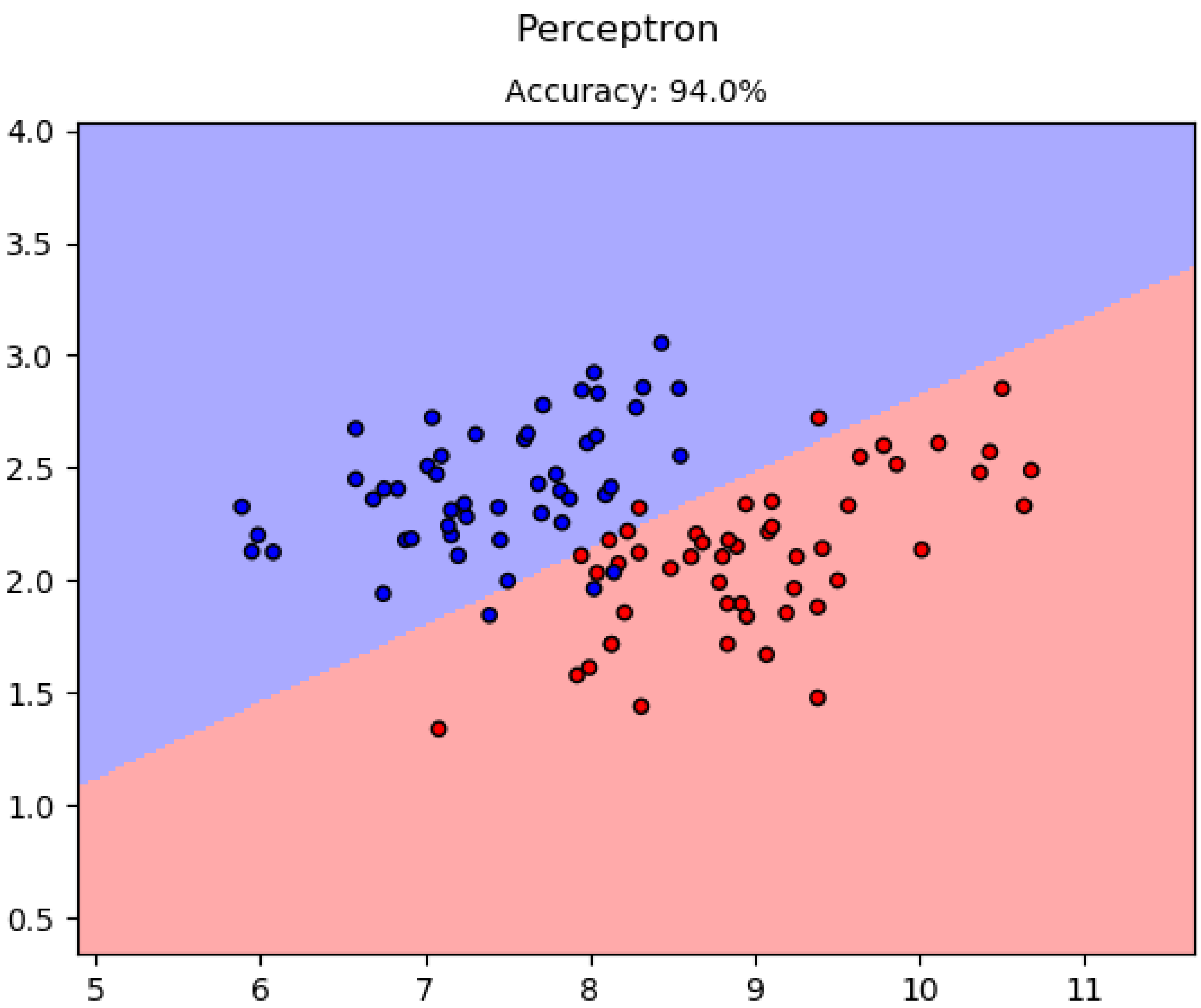
If we set the threshold to 0.5, then any y value above 0.5 is classified as 1. Any value below 0.5 is classified as 0.

$$p(x) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}}$$

w and b are the weights that the model needs to learn to represent the relationship between x and y

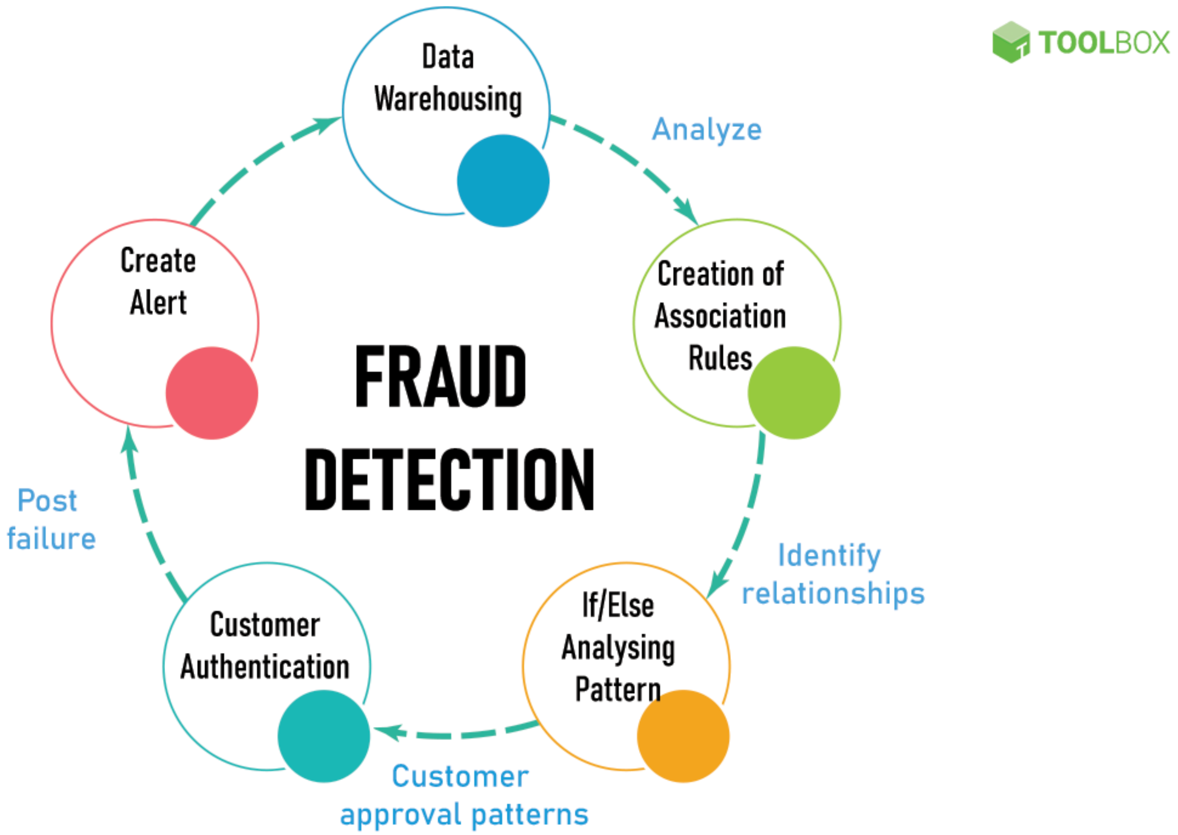
https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

Supervised Learning: Classification Example



- We have a dataset of emails that are labelled spam or not spam
- Goal: classify incoming email, using number of words in email and if spam keyword appears regularly
- Blue Points: Class 1 (Not Spam)
- Red Line: Class 2 (Spam)
- X: number of words in email
- Y: number of spam keywords
- Model learns the decision boundary shown as the border between blue and red regions

Supervised Learning: Identifying Financial Fraud



- Training a supervised machine learning model to detect financial fraud is difficult due to the low number of actual confirmed examples of fraud
- However, the presence of a known set of rules that identify types of fraud can help create a set of synthetic labels and an initial set of features.
- The output of the detection pattern that has been developed by the domain experts in the field has likely gone through the appropriate approval process to be put in production.
- It produces the expected fraudulent behavior flags and therefore, be used as a starting point to train a machine learning model.
- This shows the importance of intuition in AI

https://www.heavy.ai/technical-glossary/fraud-detection-and-prevention

Mathematical Model of Linear Regression

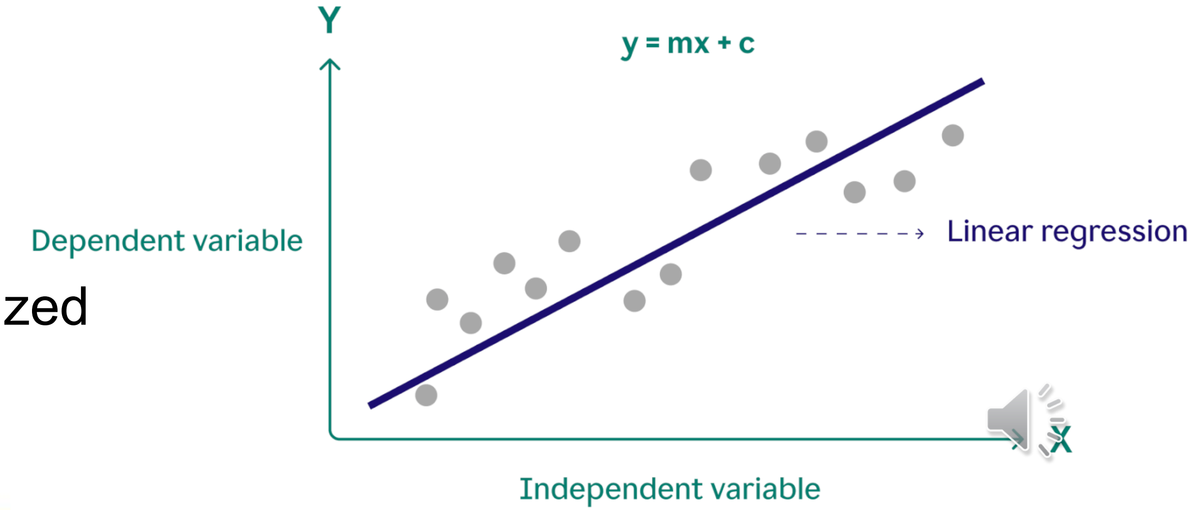
Function of model for linear regression: $f_{w,b}$ = wx + b

We need to find the optimal values of parameters w, b to model $f_{w,b}$ (best fit line)

Use cost function to calculate the error of model: average square error

$$J(w,b)=\frac{1}{2m}\sum_{i=1}^{m}\Big(\underbrace{\hat{y}^{(i)}-y^{(i)}}_{\text{error}}\Big)^2$$

m = number of training examples



Find parameters w, b such that J(w,b) is minimized

w = gradient of line, b = y intercept

Unsupervised Learning: Overview

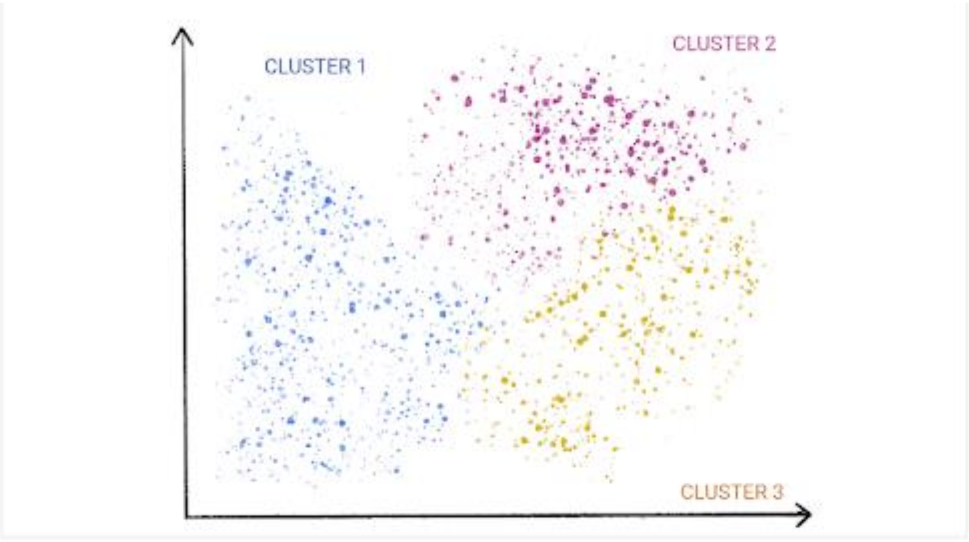


Figure 1. An ML model clustering similar data points.

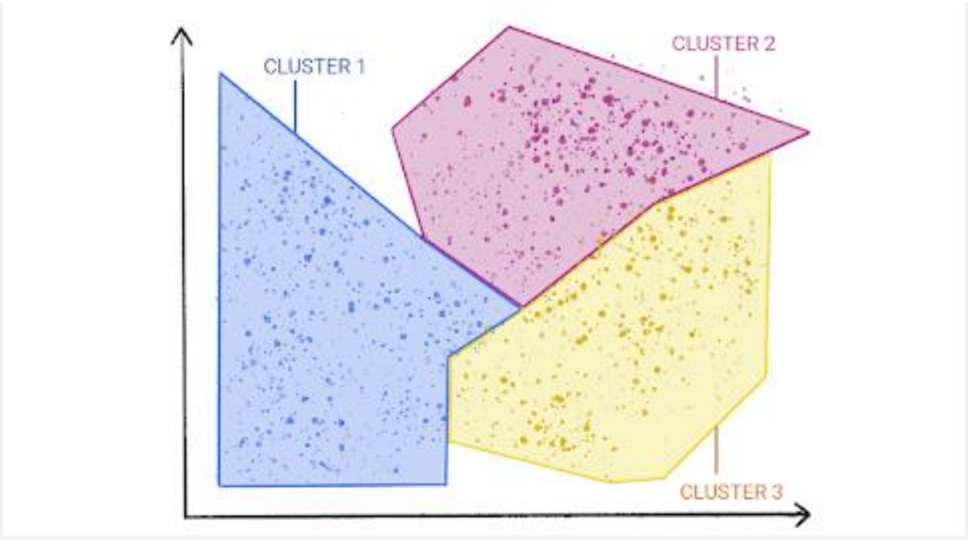


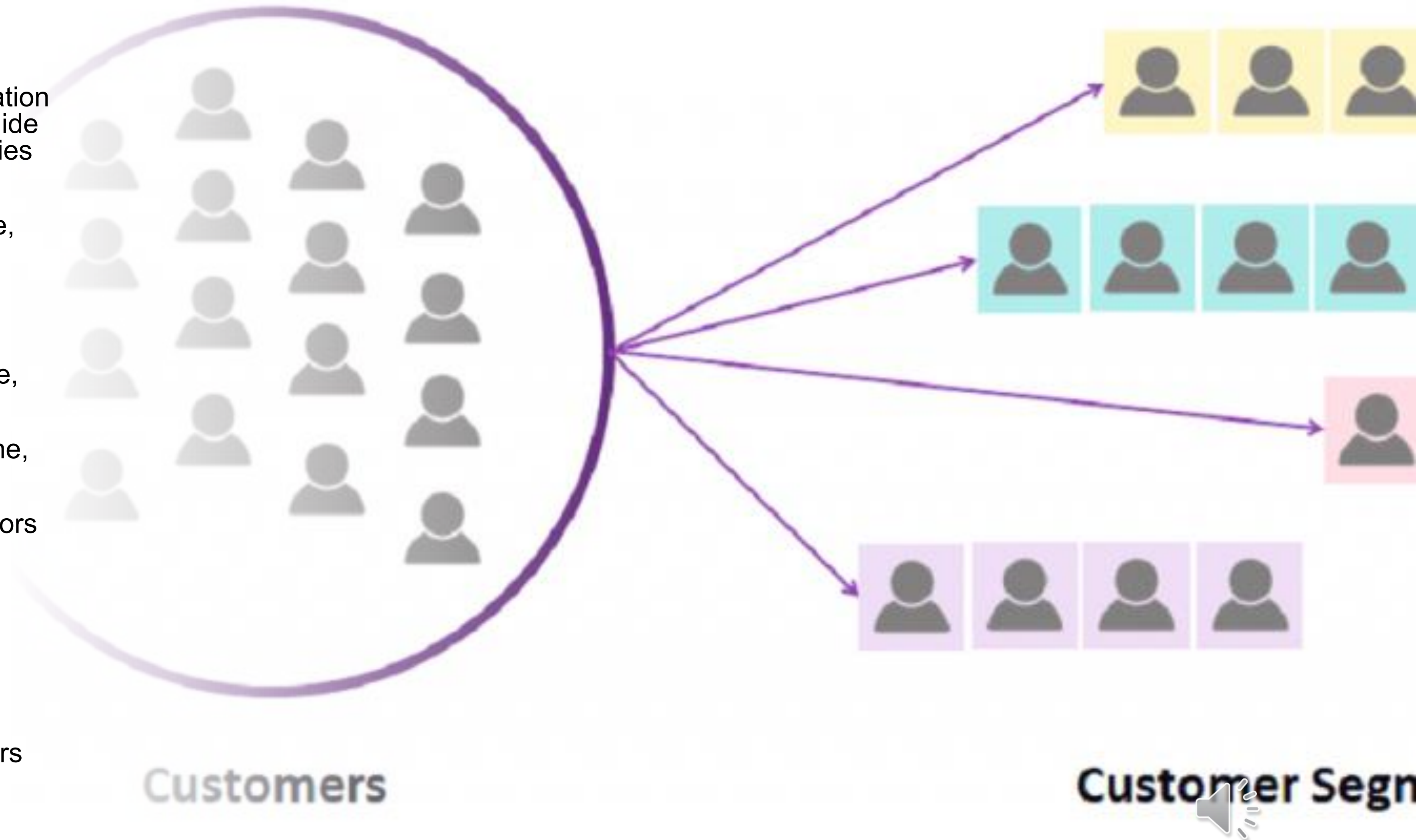
Figure 2. Groups of clusters with natural demarcations.

- Unsupervised learning is given unlabelled data and aims to find patterns and insights in data
- Unlike supervised learning, unsupervised learning trains on unlabelled data without ground truth Y, so they must predict the target Y
- Unsupervised learning algorithms are better suited for more complex processing tasks, such as organizing large datasets into clusters. They are useful for identifying previously undetected patterns in data and can help identify features useful for categorizing data.
- In above chart, the unsupervised learning algorithm aims to find the 3 separate clusters' decision boundaries

https://cloud.google.com/discover/what-is-unsupervised-learning

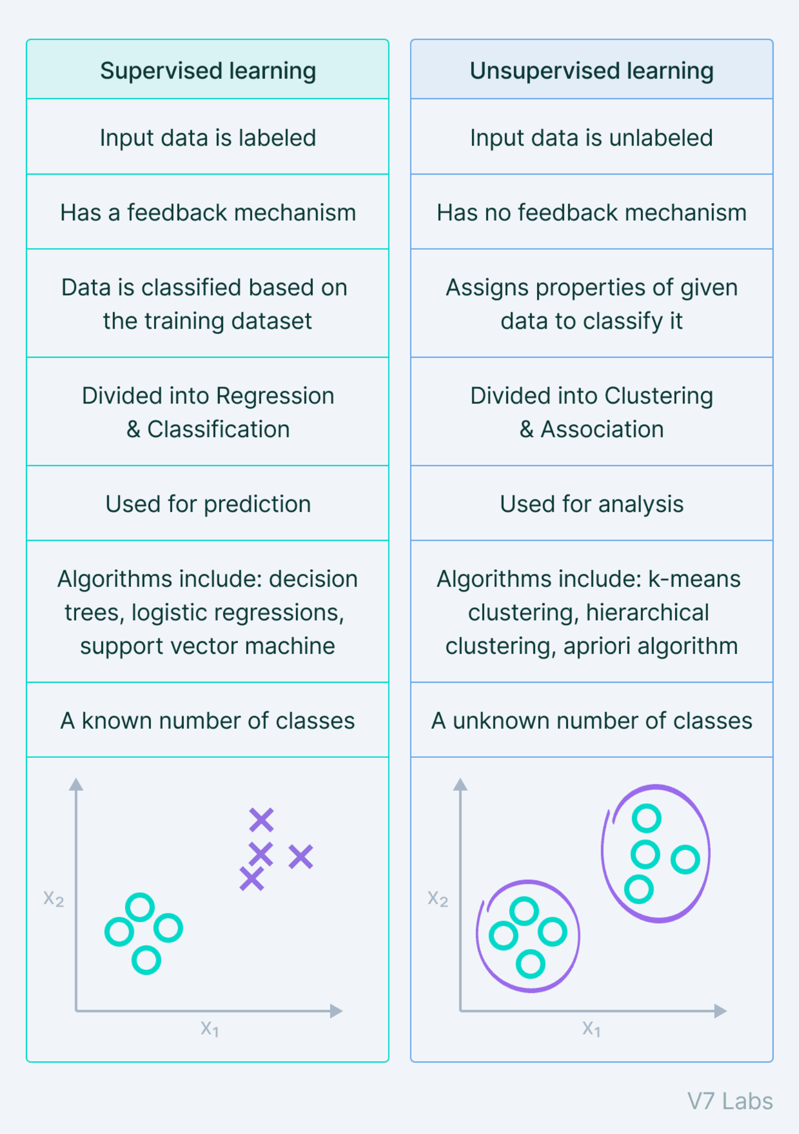
Unsupervised Learning: Customer Segmenation

- Companies want to do customer segmentation to achieve effective personalization and guide decisions on marketing and pricing strategies
- Customer segmenation means grouping customers into categories according to age, location, gender, income etc.
- There are four types of parameters for customer segmentation:
  1. Geographic – country, city, zip code, etc.
  2. Demographic – age, gender, income, occupation, etc.
  3. Behavioral – past observed behaviors of customers such as products purchased, peak spending and purchase times, etc.
  4. Psychological – personality traits, attitudes, beliefs, etc.
- Main algorithms used are clustering algorithms, ie to cluster groups of customers together



https://blog.aspiresys.com/data-and-analytics/customer-segmentation-empowered-by-machine-learning-reap-the-benefits-of-ai-to-serve-your-customers-better/

Supervised VS Unsupervised Learning



Supervised learning	Unsupervised learning
Input data is labeled	Input data is unlabeled
Has a feedback mechanism	Has no feedback mechanism
Data is classified based on the training dataset	Assigns properties of given data to classify it
Divided into Regression & Classification	Divided into Clustering & Association
Used for prediction	Used for analysis
Algorithms include: decision trees, logistic regressions, support vector machine	Algorithms include: k-means clustering, hierarchical clustering, apriori algorithm
A known number of classes	A unknown number of classes

V7 Labs

Supervised Learning

- Supervised Learning works with the help of a well-labelled dataset, in which the target output is well known.
- Supervised Learning has a feedback mechanism.
- Supervised Learning can be further divided into Classification problems and Regression problems.
- In Classification, the output variable is categorical, whereas, for Regression, the output variable is a real or continuous value.

Unsupervised Learning

- In Unsupervised Learning, the algorithm is trained using data that is unlabelled, finding the patterns in data
- Unsupervised Learning can be further grouped into Clustering and Association.
- Unsupervised Learning areas of application include market basket analysis, semantic clustering, recommender systems, etc

https://www.v7labs.com/blog/supervised-vs-unsupervised-learning

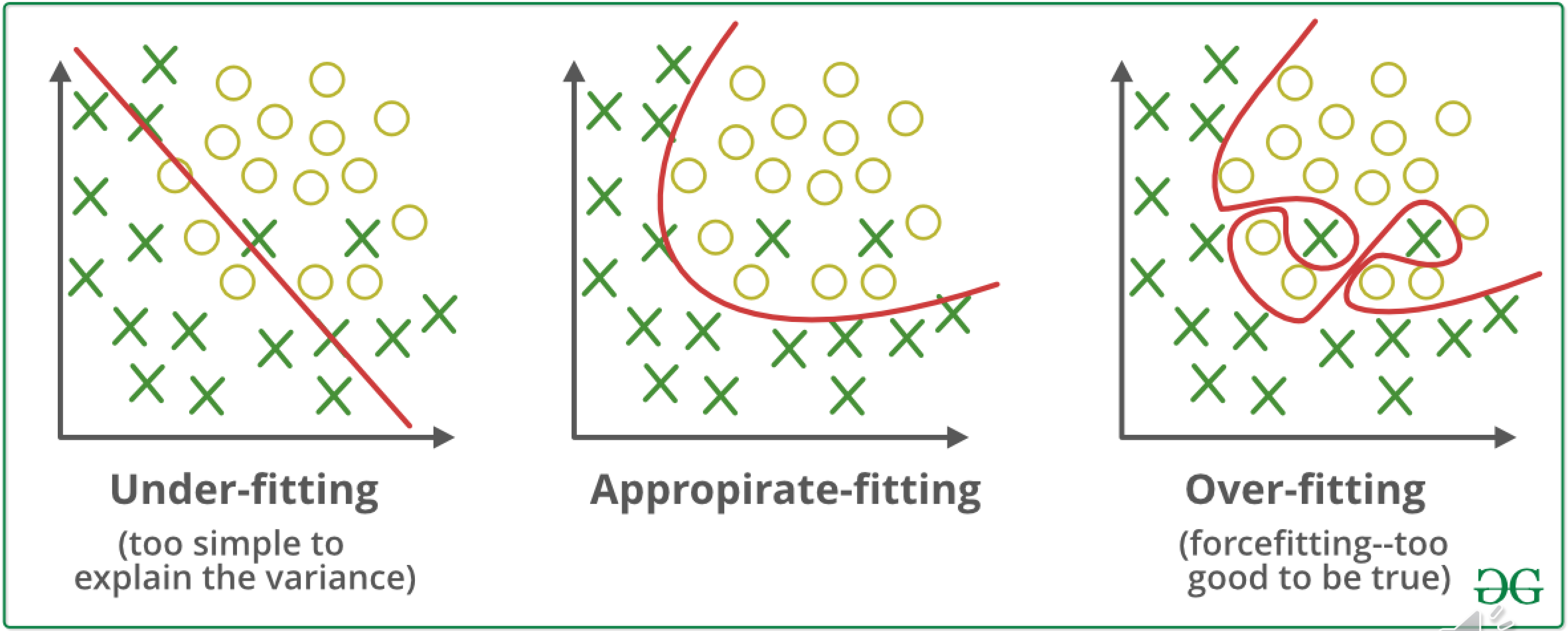
How to deal with small amounts of labelled data

Labelling data is know to be labour intensive, consuming many man hours and money
However supervised learning requires a lot of data to train the models

How do we deal with this problem? Where there is a lot of data, but only a small subset is labelled

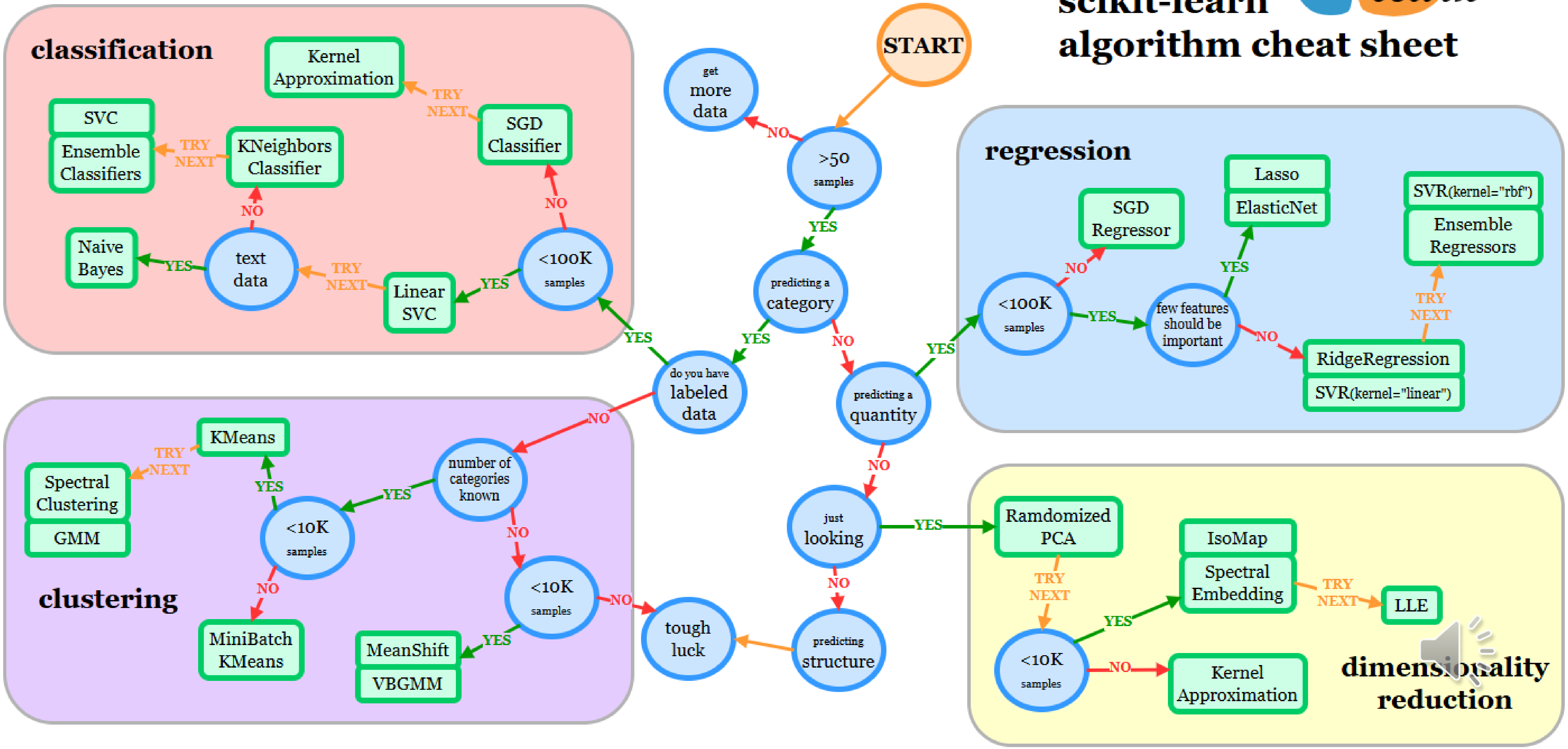
Think about it and share your ideas on wooclap

Overfitting (High Variance) VS Underfitting (High Bias)


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

https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

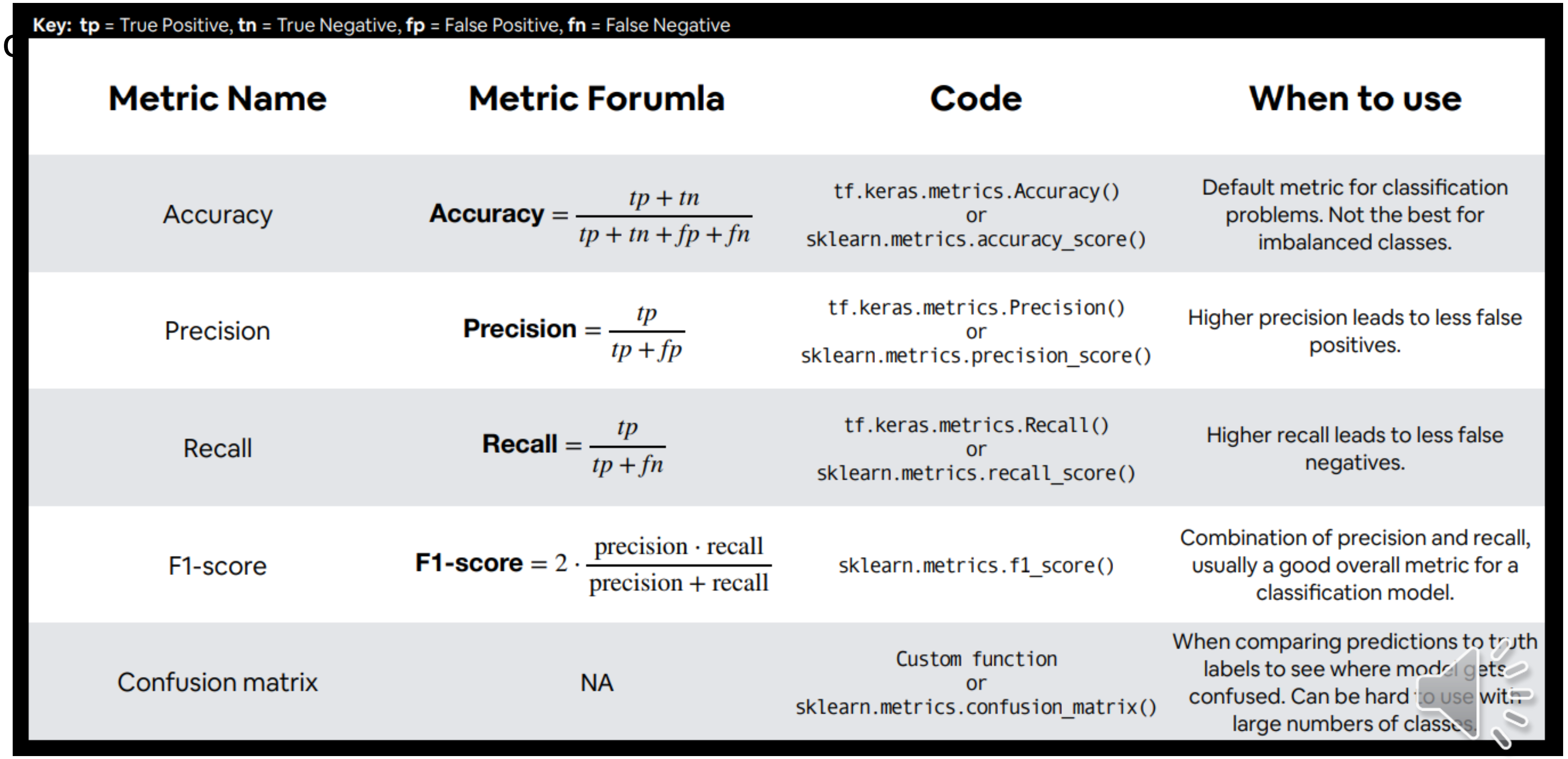
scikit-learn algorithm cheat sheet



Learning Evaluation: How is My Algo Performing?

Loss function is already minimized after gradient descent converges, next steps:

1. Train Test Validate Split: Train set for training, validation set to tune hyperparameters, test set to evaluate final model
2. Cross-Validation: Use k-fold cross validation to ensure robust performance evaluation and prevent overfitting
   - Data is split into k-subsets and model is trained and validated k times
3. Precision, Recall, F1-score, Confusio



Key: tp = True Positive, **tn** = True Negative, **fp** = False Positive, **fn** = False Negative

Metric Name	Metric Forumla	Code	When to use

Accuracy	**Accuracy** $= \frac{tp + tn}{tp + tn + fp + fn}$	`tf.keras.metrics.Accuracy()`
		or
		`sklearn.metrics.accuracy_score()`
			Default metric for classification problems. Not the best for imbalanced classes.

Precision	**Precision** $= \frac{tp}{tp + fp}$	`tf.keras.metrics.Precision()`
		or
		`sklearn.metrics.precision_score()`
			Higher precision leads to less false positives.

Recall	**Recall** $= \frac{tp}{tp + fn}$	`tf.keras.metrics.Recall()`
		or
		`sklearn.metrics.recall_score()`
			Higher recall leads to less false negatives.

F1-score	**F1-score** $= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$	`sklearn.metrics.f1_score()`	Combination of precision and recall, usually a good overall metric for a classification model.
Confusion matrix	NA	Custom function or `sklearn.metrics.confusion_matrix()`	When comparing predictions to truth labels to see where model gets confused. Can be hard to use with large numbers of classes.

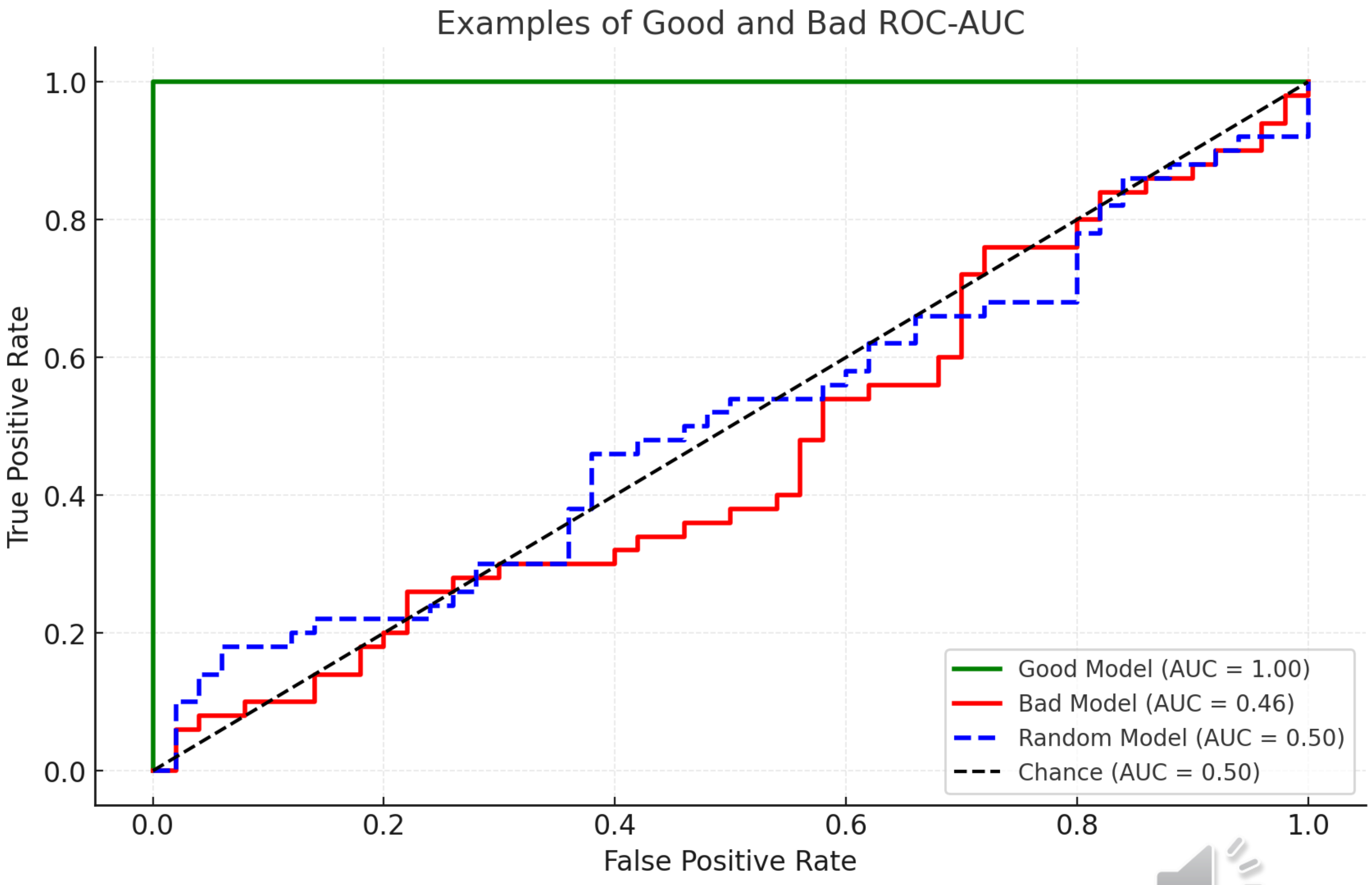
https://www.kaggle.com/discussions/getting-started/351680

Learning Evaluation: Area Under Curve

Good Model (Green Curve): A high AUC value (close to 1) indicates excellent performance, with the curve hugging the top-left corner.

Bad Model (Red Curve): A low AUC value (close to 0.5) signifies poor performance, almost equivalent to random guessing.

Random Model (Blue Dashed Line): The diagonal line represents a model with no predictive power (AUC = 0.50).



References for Chapter 1 – Supervised vs Unsupervised Learning

1. https://www.datacamp.com/blog/top-machine-learning-use-cases-and-algorithms
2. https://www.databricks.com/resources/ebook/big-book-of-machine-learning-use-cases/thank-you?scid=7018Y000001Fi19QAC&utm_source=google&utm_adgroup=141597893652&utm_offer=big-book-of-machine-learning-use-cases&utm_term=machine+learning+use+cases&gad_source=1&gclid=CjwKCAiAxqC6BhBcEiwAlXp45zG
3. https://www.researchgate.net/publication/351021675_Artificial_intelligence_in_cancer_diagnostics_and_therapy_Current_perspectives-G9y0tvxwNF2eskPqGlVAsxxtPXDibjGQBobW-_5A4ZhFFsDKTRoCWT8QAvD_BwE
4. https://www.heavy.ai/technical-glossary/fraud-detection-and-prevention
5. Andrew Ng’s Machine Learning course https://www.coursera.org/learn/machine-learning/lecture/Q8Vvp/supervised-learning-part-2
6. https://cloud.google.com/discover/what-is-unsupervised-learning
7. https://blog.aspiresys.com/data-and-analytics/customer-segmentation-empowered-by-machine-learning-reap-the-benefits-of-ai-to-serve-your-customers-better/
8. https://www.v7labs.com/blog/supervised-vs-unsupervised-learning

Chapter 1 – Supervised vs Unsupervised Learning

The End

Questions?