Lecture 4 – Machine Learning

THE UNIVERSITY of EDINBURGH
informatics

THE UNIVERSITY of EDINBURGH
School of Social
& Political Science

Dr Clare Llewellyn, Politics and International Relations

Computational Social Science

“the development and application of computational methods to complex, typically large-scale, human (sometimes simulated) behavioral data.”

... to understand society

“Computational Social Science” Lazer et al. 2009

- Every Week:
  Understanding complex social phenomena using big data
- This Week:
  Using computers to learn rules or patters from the data – Machine Learning

What is Machine Learning?

- Who has heard of it before?
- Who has studied ML before?
- Who has used it before?

Format

- Introduction to Machine Learning
- Models and Metrics

Further reading:

- https://www.youtube.com/watch?v=E0Hmnixke2g
- Book: Theobald, O. (2025) *Machine Learning and AI for Absolute Beginners.* S.l: Packt Publishing; Packt Publishing. (https://www.amazon.co.uk/Machine-Learning-Absolute-Beginners-Introduction/dp/B0F2LZ5NP5)
- https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
- https://cs229.stanford.edu/notes2022fall/main_notes.pdf
- https://www.knime.com/getting-started-guide

Part 1

Introduction to Machine Learning

Introduction to Machine Learning

Part 1 (20 min)

1. Definition of Machine Learning
2. History of Machine Learning

Part 2 (50 min)

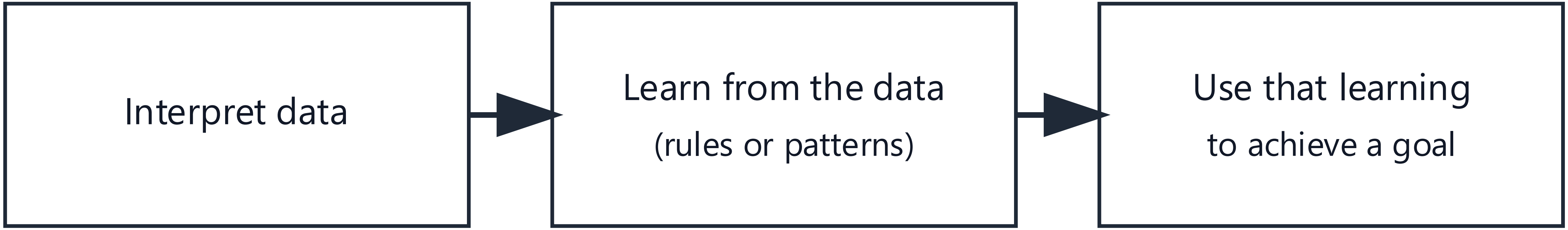
3. The Machine Learning Workflow
4. Ethics
5. A worked example in use

Part 3 (20 min)

6. Whistle stop tour of ML algorithms

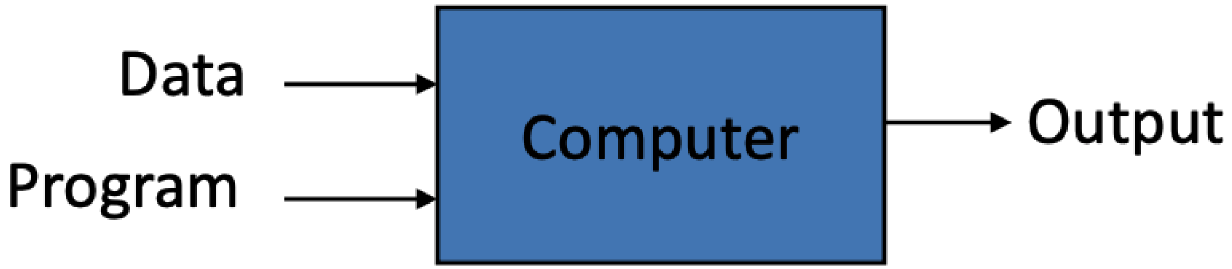
1. Definition of machine learning

Machine Learning

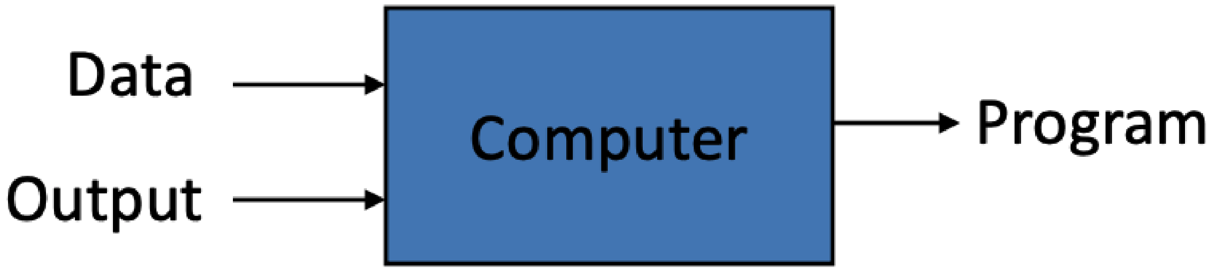


Extending traditional programming

Traditional Programming



Machine Learning



The Machine Learning Program can perform tasks with out specific instructions by using statistical algorithms to learn from data (from labels or patterns) and generalise to new data

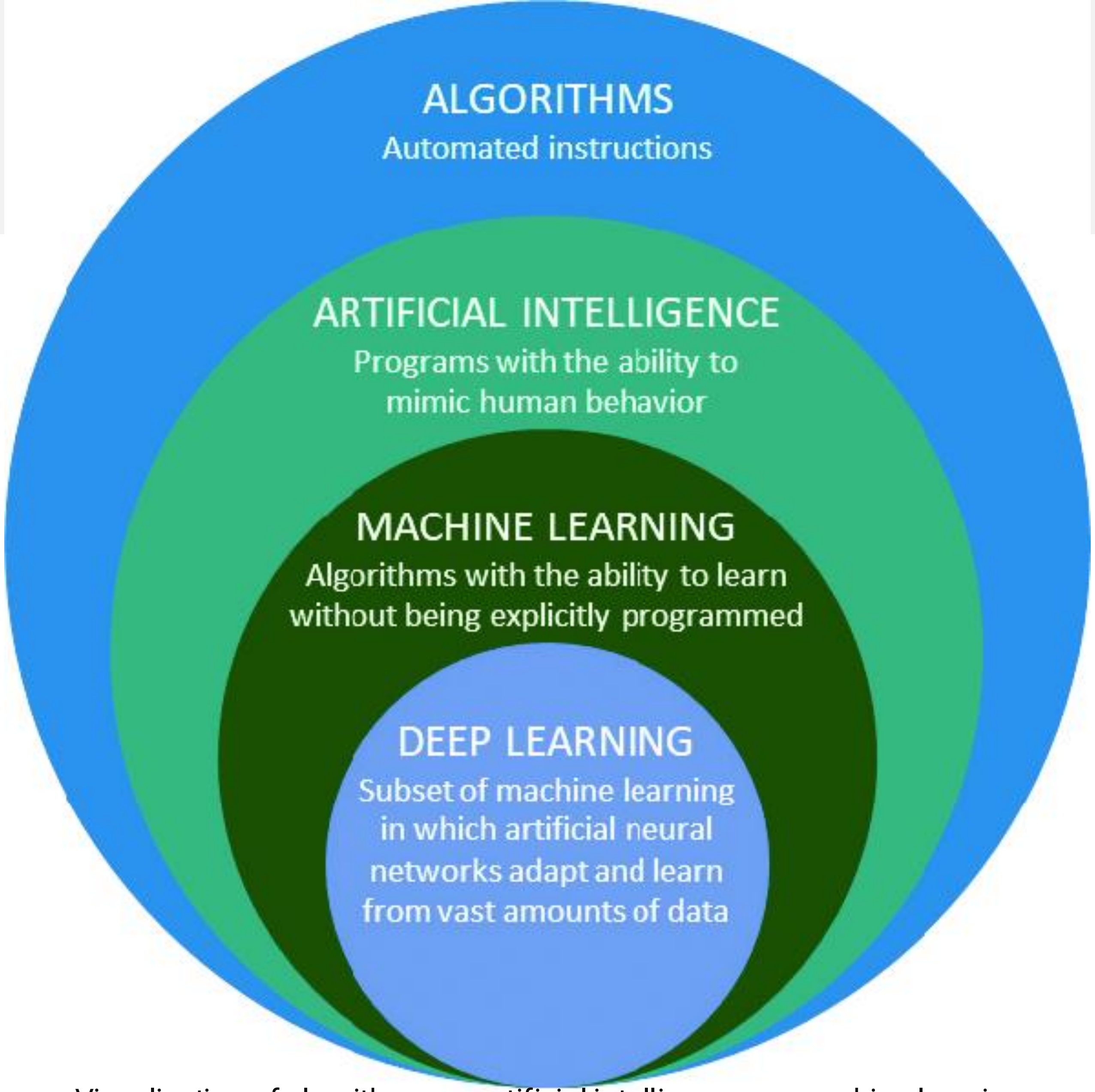
Image taken from: https://www.seas.upenn.edu/~cis5190/fall2017/lectures/01_introduction.pdf

Machine Learning

Artificial Intelligence:
algorithms that do things we think of as human capabilities, **learning, reasoning, problem solving and decision making**.

Machine Learning:
algorithms that process vast amounts of human made data and **learn the rules or the patterns** for these process.

Deep Learning:
uses multi-layer neural networks to automatically learn from large amounts of data. It works well for **complex patterns in unstructured data**, like images, audio, and text.



Visualization of algorithms vs. artificial intelligence vs. machine learning vs. deep learning (Author: Johannes Vrana, Vrana GmbH, Licenses: CC BY-ND 4.0)

Where is Machine Learning used?

- **Recommendation algorithms.** learn our preferences, for example: Netflix and YouTube suggestions, what information appears on your Facebook/Insta/TikTok feed, product recommendations in Amazon.,
- **Image analysis and object detection.** Analysing images for information, for example identify people though facial recognition
- **Fraud detection**. Machines can analyze patterns, for example how we normally spend or where we normally shop, to identify credit fraud.
- **Automatic helplines or chatbots.** Many companies are deploying online chatbots, in which customers or clients don’t speak to humans, but instead interact with a machine. These algorithms use machine learning and natural language processing, with the bots learning from records of past conversations to come up with appropriate responses.
- **Self-driving cars.** Much of the technology behind self-driving cars is based on machine learning
- **Medical imaging and diagnostics.** Machine learning programs can be trained to examine medical text, images, and empirical results, to look for certain markers of illness, like a tools that predict breast cancer risk based on a mammogram.

Adapted from https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Different Requirements

Manufacturing Industry

Efficiency is key to the success of an organization in the manufacturing industry.

- Identifying equipment errors before malfunctions occur, using the internet of things (IoT), analytics, and machine learning
- Using an AI application on a device, located within a factory, that monitors a production machine and predicts when to perform maintenance, so it doesn't fail mid-shift

Banking

Data privacy and security are especially critical within the banking industry.

- detect and prevent fraud and cybersecurity attacks
- Integrating biometrics and computer vision to quickly authenticate user identities and process documents

Health Care

Huge amounts of complex and varied data and increasingly relies on analytics to provide accurate, efficient health services.

- Analyzing data from users' electronic health records through machine learning to provide clinical decision support and automated insights
- Capturing and recording provider-patient interactions in exams or telehealth appointments using natural-language understanding

What can machine learning do?

There are 3 categories of goals that machine learning can achieve. They can be:

Descriptive, explains the data, **customer segmentation data**

"people who buy X also buy Y"

Predictive, can predict unseen data, **credit scores**

using past outcomes to to give a risk score for a new applicant

Prescriptive, helps to decide what action to take, **dynamic pricing**

takes predictions and determines the best action to take

https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

There are 3 Types: Machine Learning

Supervised Machine Learning:
Models are trained with **labelled data sets**, which allow the models to learn and grow more accurate over time. For example, an algorithm would be trained with pictures of dogs and other things, **all labelled by humans**, and the machine would learn ways to identify pictures of dogs on its own. **Supervised machine is the most common type.**

Unsupervised Machine Learning:
A program looks for **patterns in unlabelled data**. Unsupervised machine learning **can find patterns or trends** that people aren't explicitly looking for. For example, an unsupervised machine learning program could look through online sales data and identify different types of clients making purchases.

Reinforcement Machine Learning:
Trains machines through trial and error to take the best action by **establishing a reward system**.
Reinforcement learning can train models to play games or train autonomous vehicles to drive by telling the machine when it made the right decisions, which helps it learn over time what actions it should take.

Types of machine learning

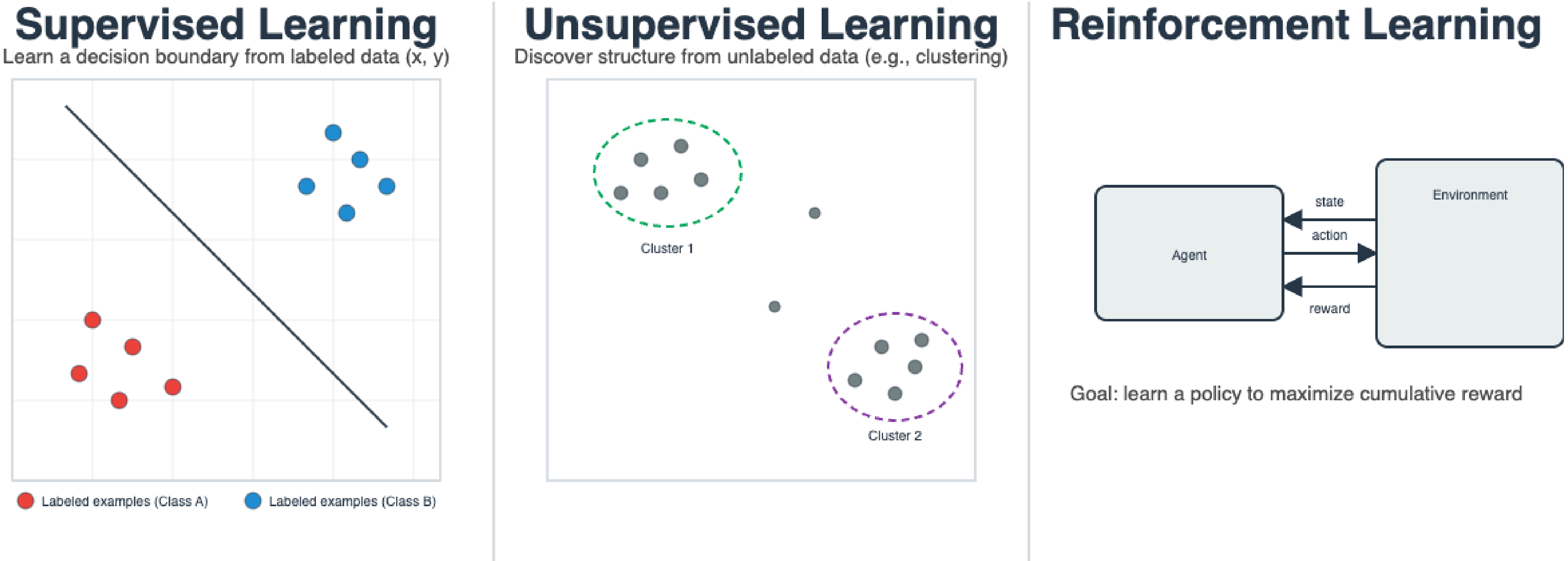
Supervised Learning
- Labeled data
- Direct feedback
- Predict outcome/future

Unsupervised Learning
- No labels/targets
- No feedback
- Find hidden structure in data

Reinforcement Learning
- Decision process
- Reward system
- Learn series of actions

Taken from:https://sebastianraschka.com/pdf/lecture-notes/stat451fs20/01-ml-overview__slides.pdf

Types of Machine Learning



Created by ChatGPT 4.0

2. History of Machine Learning

Nobel Prize in Physics for the 'godfather of AI', Geoffrey Hinton

[08/10/2024] Professor Geoffrey Hinton, who graduated from the University of Edinburgh with a PhD in Artificial Intelligence in 1978 has been awarded a Nobel Prize in Physics for his work on machine learning. He shares the 2024 award with Professor John Hopfield of Princeton University.

Photo used by permission of Geoffrey Hinton

The duo were recognised for their groundbreaking work on artificial neural networks, which underpin many modern applications of artificial intelligence, such as chatbots.

The history of AI at the U of Edinburgh

- **1963:** Department of *Machine Intelligence & Perception* established — early UK centre for AI
- **1970s:** Major contributions to **symbolic AI** (reasoning, search, planning, knowledge representation)
- **1970s–80s:** Edinburgh becomes a leading hub for **NLP / computational linguistics**
- **1998: School of Informatics** formed, consolidating and scaling AI research and teaching
- **2000s:** Machine learning expands across informatics (vision, language, data-driven AI)
- **2010s:** Rapid growth in **deep learning** (neural methods for vision, speech, NLP)
- **2010s–2020s:** Stronger emphasis on **interdisciplinary AI** + **responsible/ethical AI**

History

1950's Alan Turing's paper imagined a machine that could communicate—via an exchange of typed messages—so capably that people conversing with it could not tell whether they were interacting with a machine or another person (Turing 1950).

1955 The term artificial intelligence was proposed by a group of computer scientists, "to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al., 1955).

1957 The perceptron is created (Rosenblatt in 1958) this artificial neuron will become the basis for deep learning

AI WINTER

1980's Advanced decision tree and rule learning

1984 encoding specialized human expertise into rules for the machine to follow (Buchanan and Shortliffe, 1984).

AI WINTER

1990's Data mining, Text learning

2000's: Support vector machines & kernel methods Sequence labelling – Collective classification and structured outputs

2010's Deep learning systems, Learning for big data. "neural networks" and "deep learning"). Together, these advances have created a technological tidal wave.

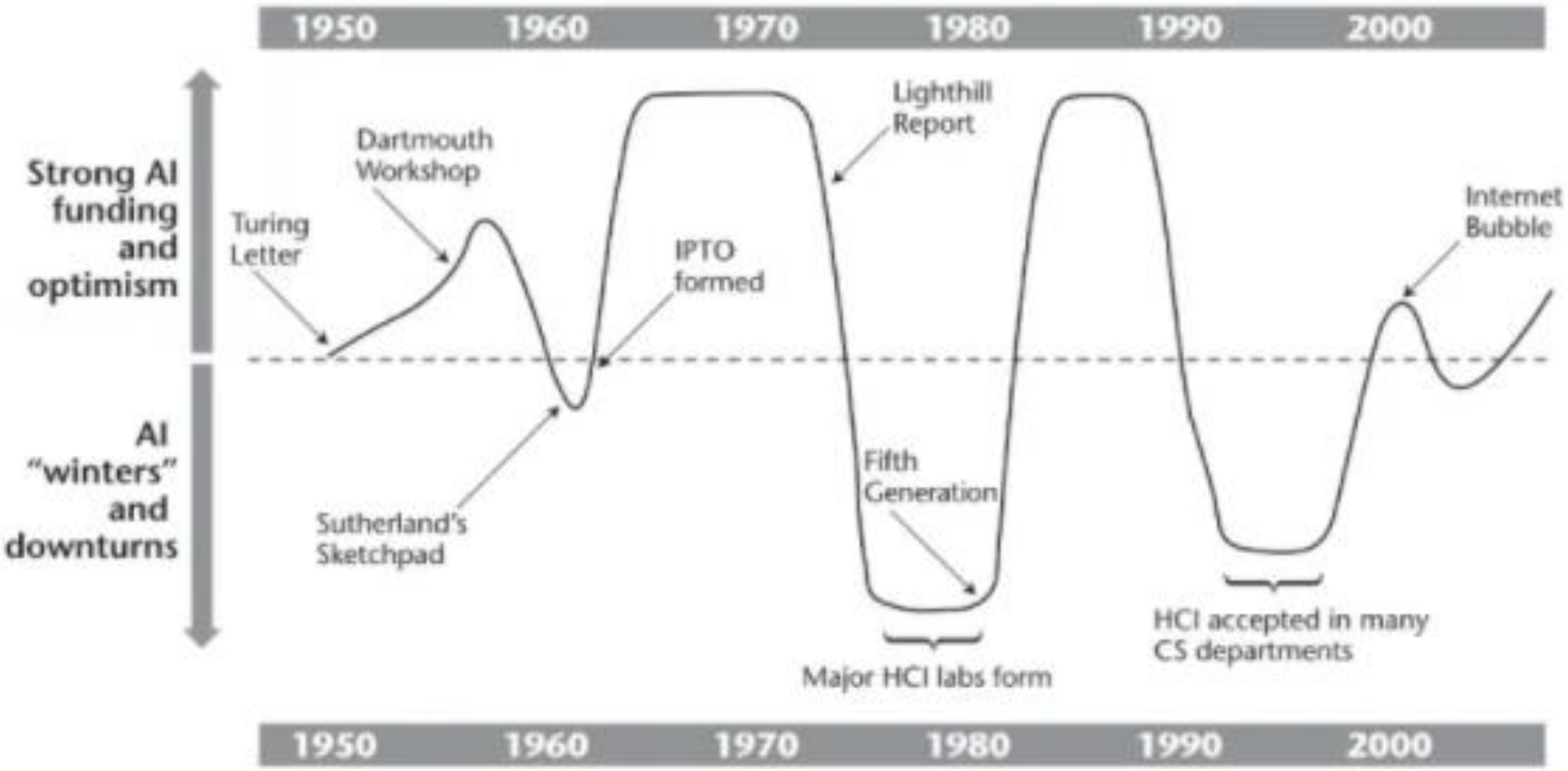
2011 IBM's Watson program beat the best human players of the TV game show Jeopardy.

2015, for example, Google's AlphaGo beat a grand master of the game of Go,

https://workofthefuture-taskforce.mit.edu/wp-content/uploads/2020/12/2020-Research-Brief-Malone-Rus-Laubacher2.pdf

AI in context

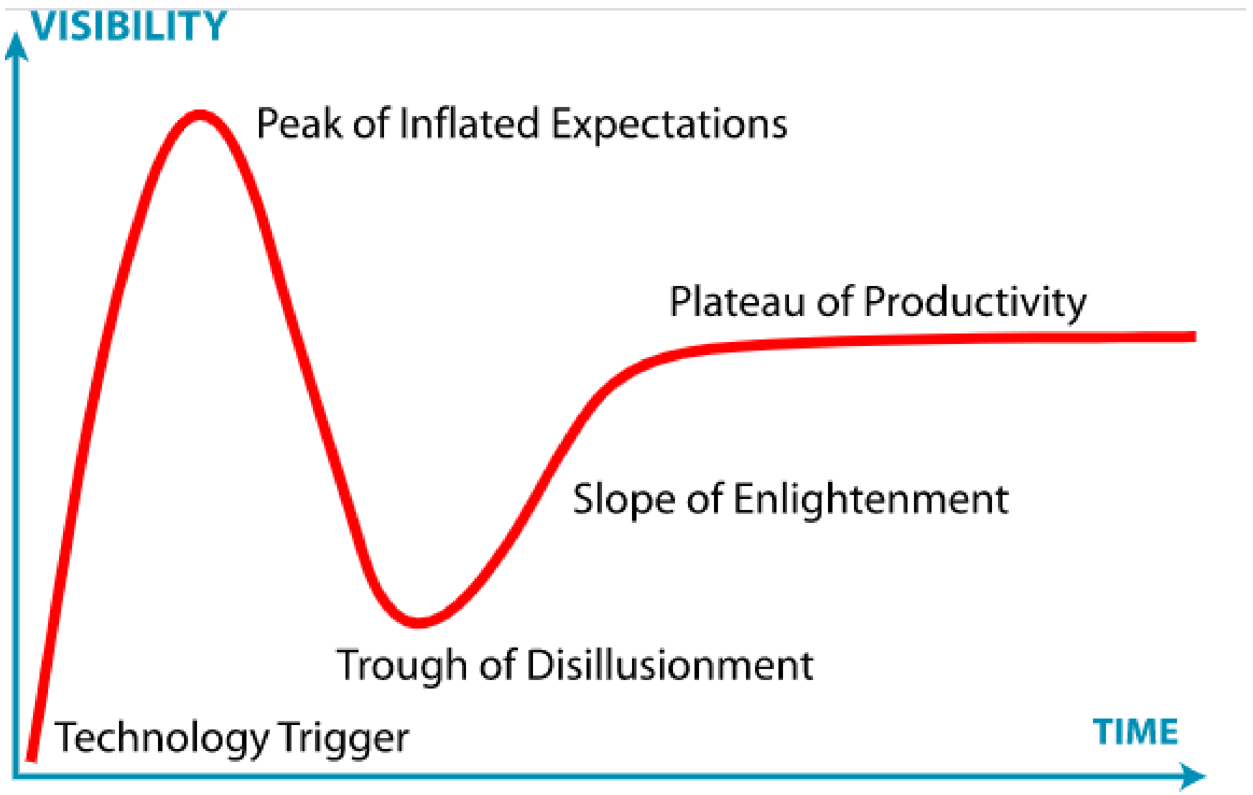
Good Old Fashioned AI



Grudin, J. (2009)
AI and HCI: Two fields divided by a common focus.

Frankish, K. & Ramsey, W.M. (2017)
The Cambridge Handbook of Artificial Intelligence.

Tech Hype Cycle



- Move from academic/angel investor to commercial funding
- Not as exciting to the media
- The problems with it are so bad it becomes rejected by the public “the Acute Crisis Stage / Point of No Return”
- AI becomes a threat to humanity / Fear of what it could become “singularity” when machines surpass human intelligence

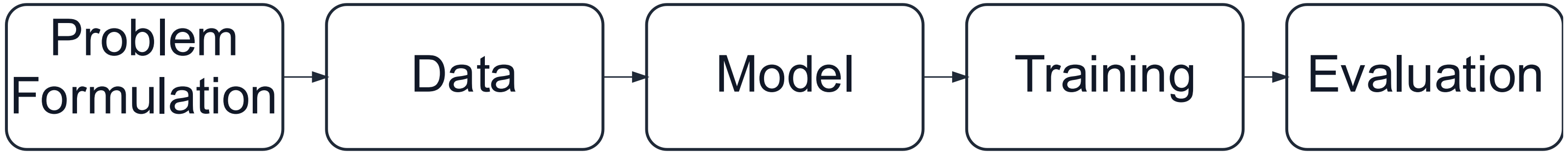
Will there be another AI Winter?

In the 1990’s there was a revival of neural networks (deep learning). This has led to our current AI hype. It is underpinned by 3 factors:

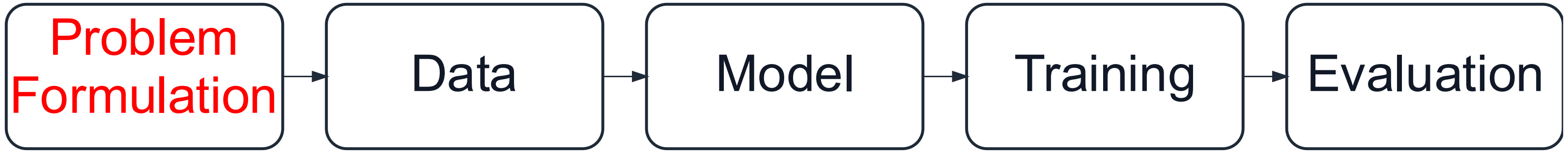
- a rise in computational power (games and phones)
- data availability (digital revolution, internet, social media, cloud computing)
- algorithmic innovations (we harnessed what existed with neural nets and made it better

BREAK

3. The Machine Learning Workflow



3.1 Problem Formulation

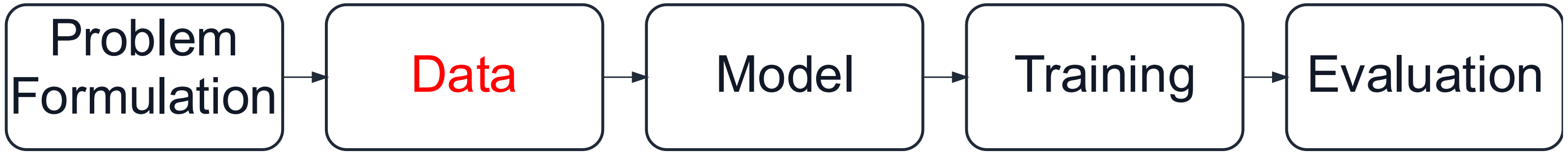


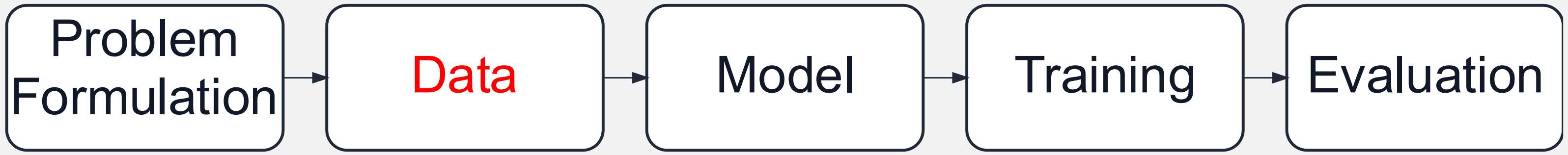
Most ML failures come from ***badly formulated problems*** (wrong label, wrong timeframe, wrong metric, data leakage, wrong optimising) even if the model training is "correct."

Problem Formulation

- **Problem formulation** is the step where you translate a real-world goal into a **precise ML task** with clearly defined **inputs, outputs, constraints, and evaluation.**
- Is the goal **Descriptive**, explains the data, **Predictive**, can predict unseen data, or **Prescriptive?** This helps to decide what action to take
- Choose what the model will learn and **how success will be measured.**
- Next you need to look at your data…

3.2 Data





- Can you visualize the data so you can understand it better?
- What data do I have?
  - How much is enough data?
  - Do I have enough data?
- What are your inputs?
  - Do I know what classes I have?
  - Are the data numerical or text?
- What is the goal?
  - Descriptive, Predicative, Prescriptive
  - Is the data labeled, or can it be labeled?
- How will I split my data for evaluation (train v's test)

Feature Crafting and Analysis

A feature is an input variable taken or derived from your data

- **Features** = what you feed into the model.
- **Feature crafting** = how you create or improve those inputs.
- **Feature analysis** = how you verify the inputs are informative, stable, and appropriate.

Feature Crafting

The process of **creating, transforming, and selecting** features to make the model learn better.

- **Cleaning & preprocessing**
  - handle missing values (do you need an unknown category), deal with outliers, standardise/normalise numeric variables, encode categorical variables, turn text into data
- **Transformations**
  - You may need to transform skewed variables (such as log for income), bucket or bin variables (such as for age or time), or to combine variables
- **Aggregation**
  - You could determine counts or averages “average spend in last 30 days”, “number of late payments in last 12 months”,
- **Domain-derived features**
  - ratios (debt / income), rates (missed_payments / total_payments), behavioural summaries (trend in spending)

Feature Analysis

Checking whether features are **useful, valid, stable, and safe** to use.

- **Predictive usefulness,** correlation (for regression), mutual information, feature importance (tree models), permutation importance, ablation tests: remove a feature and see performance change
- **Relationship with target,** plots: distributions by class, partial dependence, SHAP summaries, monotonicity checks (does risk increase as debt increases?)
- **Data leakage checks,** ensure a feature wouldn't be known at prediction time e.g., "days since last missed payment" might accidentally include future info depending on how it's computed
- **Stability over time / drift,** does the feature's distribution change (e.g., pre/post policy change)?
- **Fairness and compliance,** is the feature a proxy for protected attributes (postcode as proxy for ethnicity/income)? assess disparate impact and document rationale

Data Issues

Main Questions to consider:

- Are labels reliable and consistently defined?
- Is the dataset representative of who/what you’ll predict on?
- Are there signs of drift between training and test data?

What can go wrong with data

1) Data quality problems: Missing values, Outliers / extreme values, Noise and measurement error, Duplicates, Inconsistent formats/units, Invalid values (negative ages, impossible timestamps)

2) Label (target) issues: Incorrect labels, Ambiguous labels, Changing label definition over time, Class imbalance

3) Data leakage: Target leakage: a feature directly or indirectly contains the answer, Train–test contamination: duplicates or near-duplicates across splits, Time leakage: using future data in features for past prediction

4) Data integration and pipeline issues: Incorrect joins (many-to-many, joining across time improperly). Entity resolution problems (same person appears under multiple IDs)

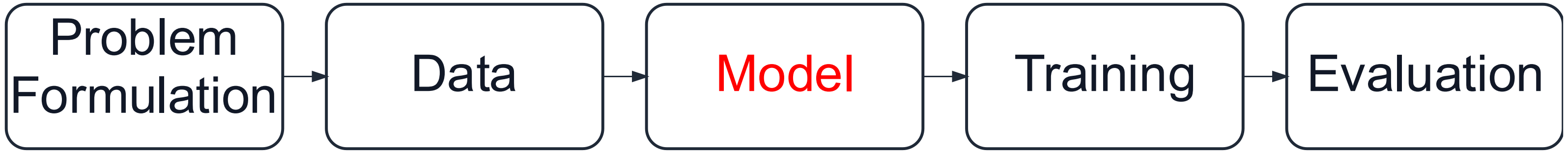
5) Sampling and representativeness issues: Sampling bias (training data not representative of real-world use) Selection bias (you only observe outcomes for a subset; e.g., only approved loans) Coverage gaps (missing certain groups, regions, device types, rare conditions) Non-stationarity (data distribution changes due to seasonality, trends, shocks)

6) Feature problems: High-cardinality categoricals (thousands of categories → sparse/overfit risk), Multicollinearity / redundant features (can destabilise linear models, obscure interpretation),Wrong encoding (ordinal encoding for non-ordered categories) Unscaled features (hurts distance-based models like k-NN, SVM, k-means) Proxy features (features that unintentionally encode sensitive attributes)

7) Dataset size and dimensionality issues. Too little data (especially for complex models; high variance/overfitting) Too many features vs samples (curse of dimensionality), Rare-event sparsity (few positive examples), Unbalanced groups (some subpopulations have too few examples)

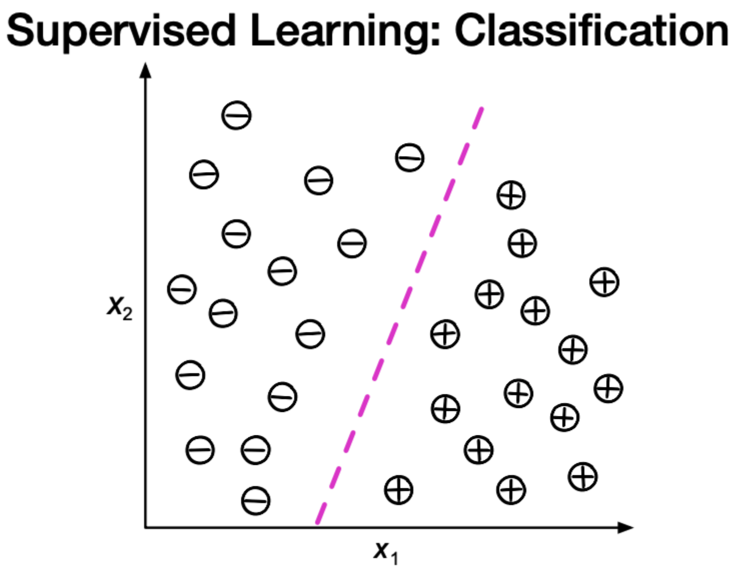
8) Drift and production data mismatch: Training–serving skew: features computed differently in production than in training. Concept drift: the relationship between inputs and target changes (e.g., fraud tactics evolve), Covariate shift: input distribution changes (new user demographics), Instrumentation changes: logging/collection changes break feature meaning

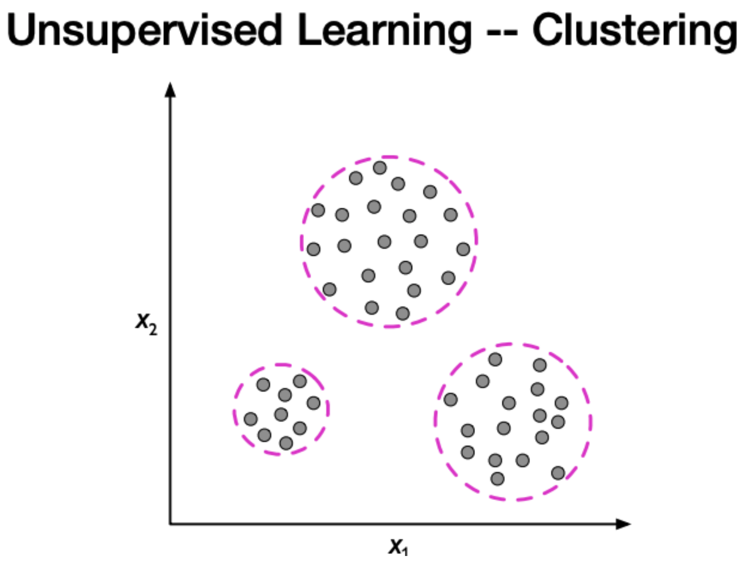
3. Model



What model do I choose?

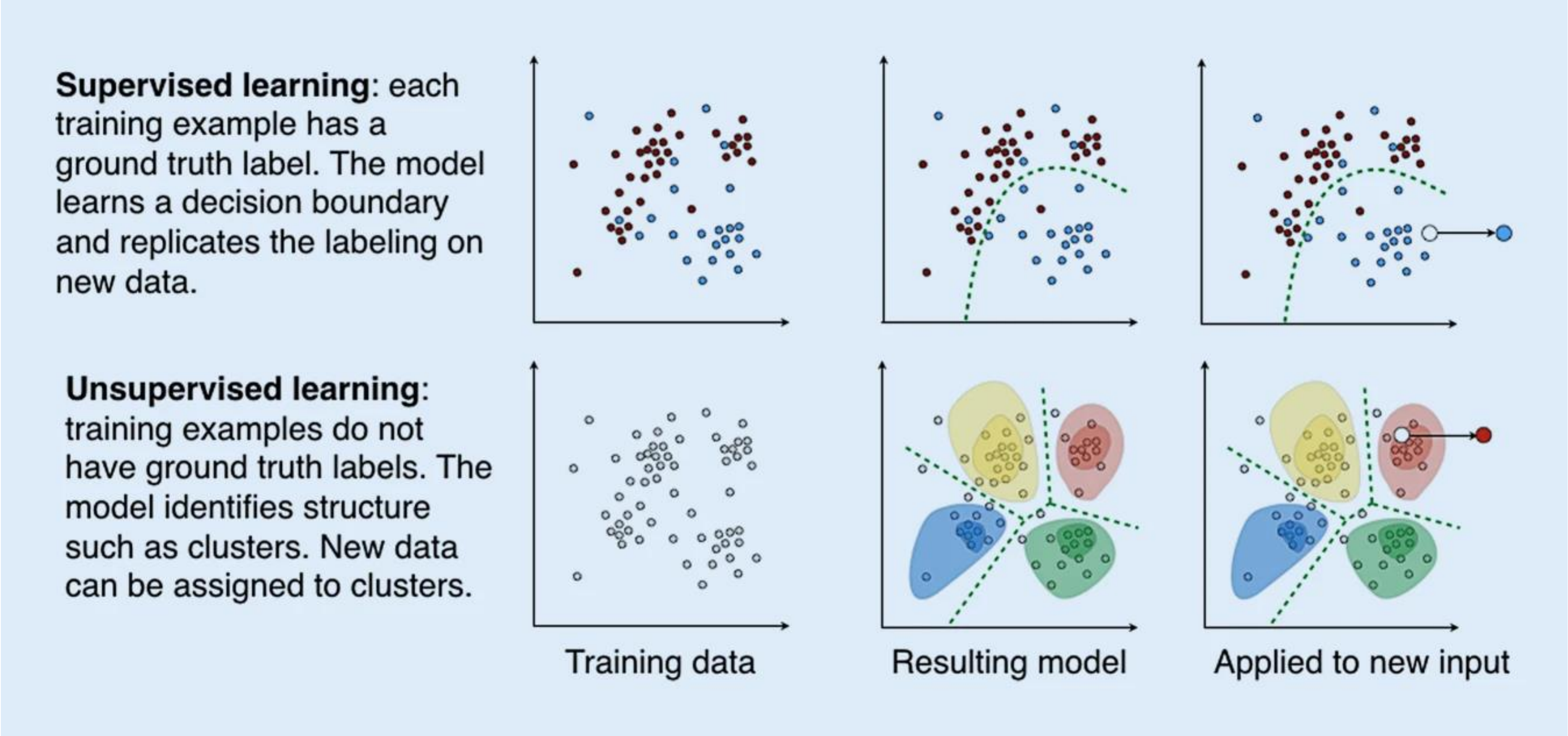
- **Classification** (spam vs not spam)
- **Regression** (predict house price)
- **Clustering / descriptive** (find groups)





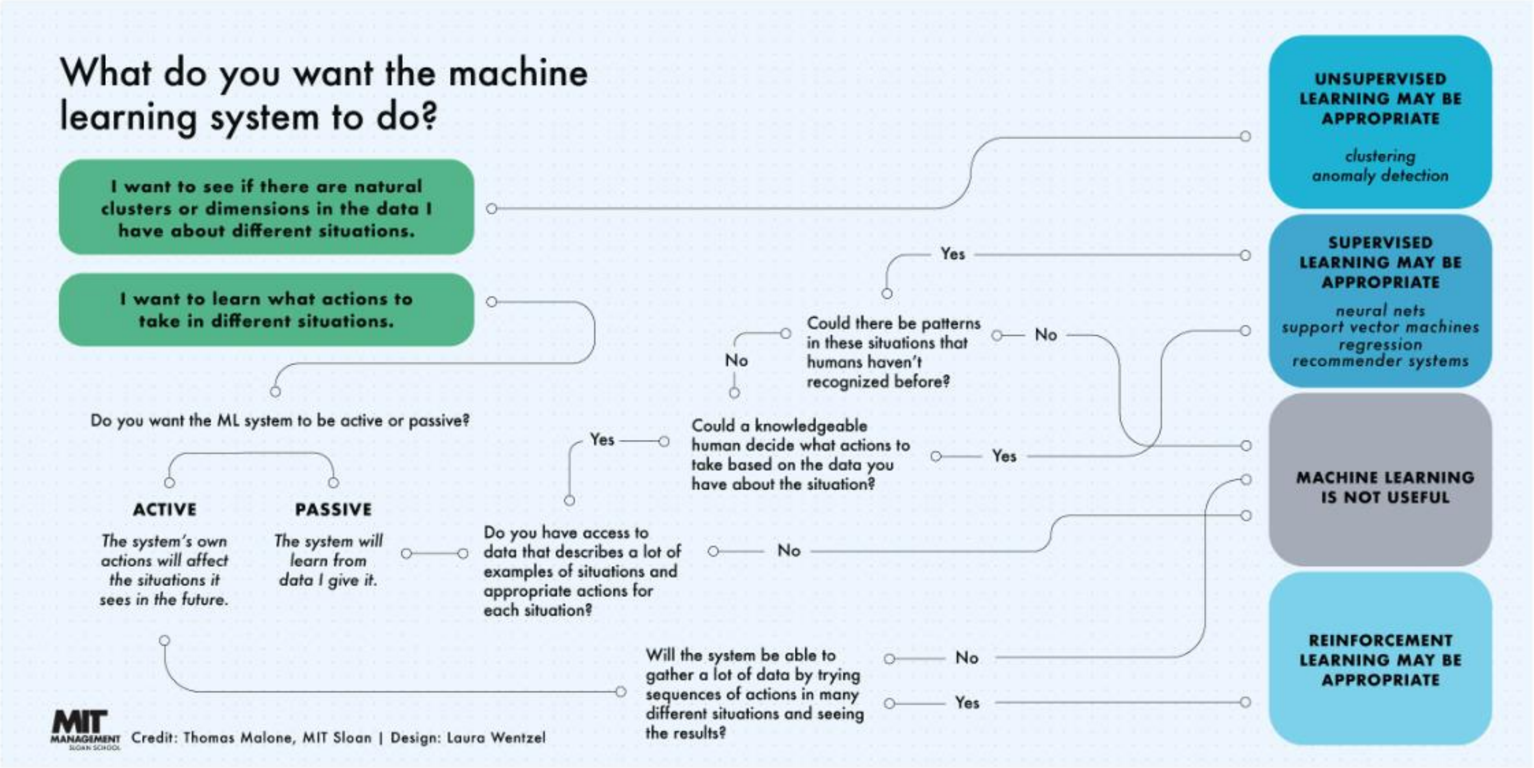
Adapted from: https://sebastianraschka.com/pdf/lecture-notes/stat451fs20/01-ml-overview__slides.pdf

More Realistic Data



Taken from Langs, G., Röhrich, S., Hofmanninger, J. et al. - Langs, G., Röhrich, S., Hofmanninger, J. et al. Machine learning: from radiomics to discovery and routine. Radiologe 58, 1–6 (2018).https://doi.org/10.1007/s00117-018-0407-3, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=101352238

Chosing the type of machine learning



Source: Thomas Malone | MIT Sloan. See: https://bit.ly/3gvRho2, Figure 2.

Taken from: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

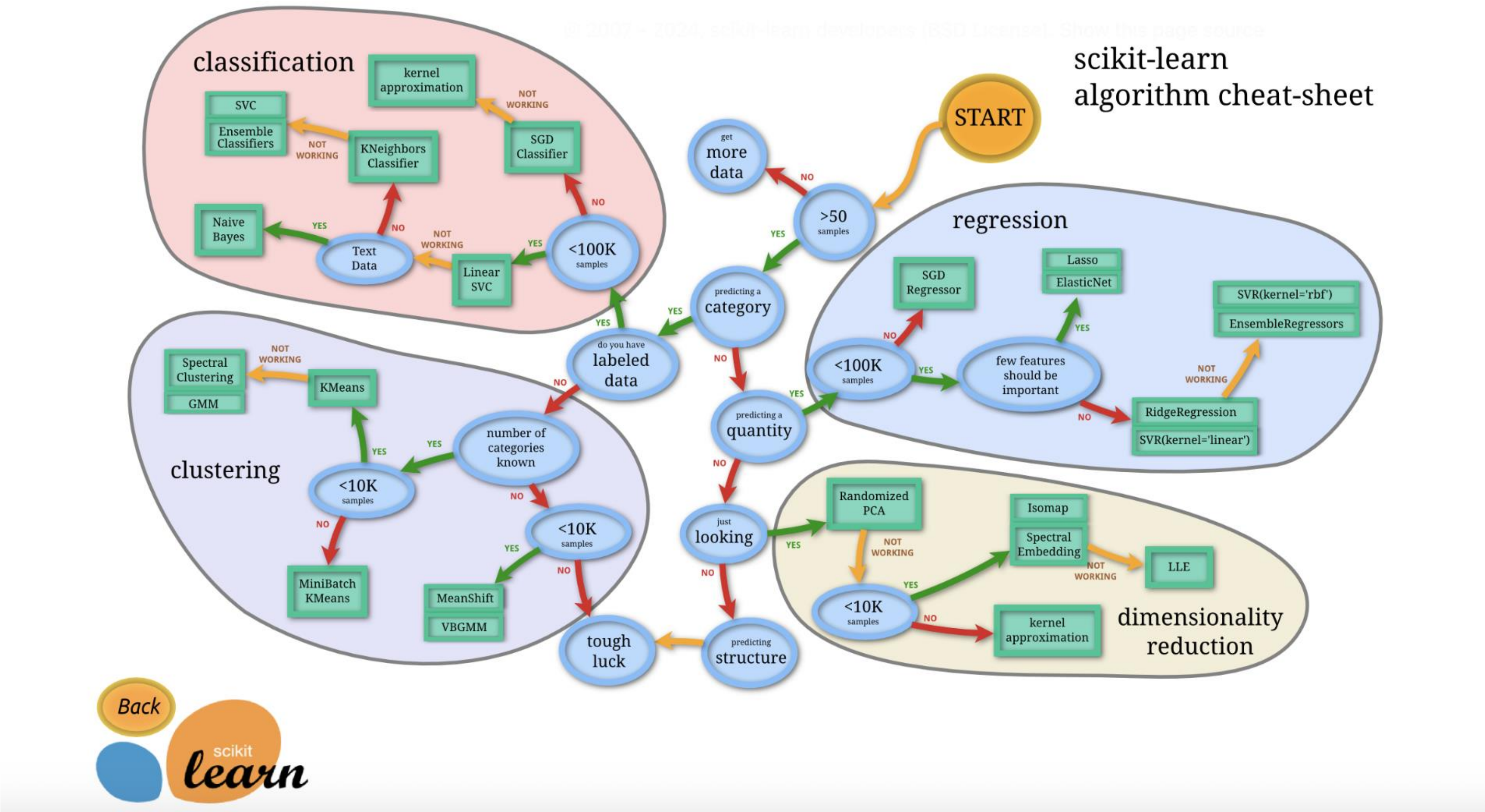
Adapting to the work flow

- **Supervised Learning** is ideal when you have a well-labeled dataset and need to predict outcomes or classify data. It's commonly used in situations where the relationships between inputs and outputs are already known and need to be modeled.
- **Unsupervised Learning** is useful when you have a dataset without labels and want to explore its structure. It's perfect for discovering underlying patterns or grouping data into clusters. It's often used for exploratory data analysis, feature learning, and anomaly detection.
- **Reinforcement Learning** is suitable for scenarios where an agent needs to learn a strategy over time by interacting with an environment. This approach is particularly useful in complex decision-making tasks like robotics, game playing, and autonomous systems

Feature	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Data Requirement	Labeled data	Unlabeled data	Interactive environment
Learning Process	Learn from input-output pairs	Discover patterns in data	Learn from actions and rewards
Model Types	Classification, Regression	Clustering, Dimensionality Reduction	Q-learning, Policy Gradients
Goal	Predict or classify new data points	Identify hidden structures	Maximize cumulative rewards
Examples	Spam detection, Image classification	Customer segmentation, Anomaly detection	Game playing, Robotics
Evaluation	Accuracy, Precision, Recall, MSE	Silhouette Score, Explained Variance	Cumulative reward
Challenges	Requires large labeled datasets, overfitting	Interpretability, evaluation difficulty	Exploration-exploitation trade-off, sample efficient

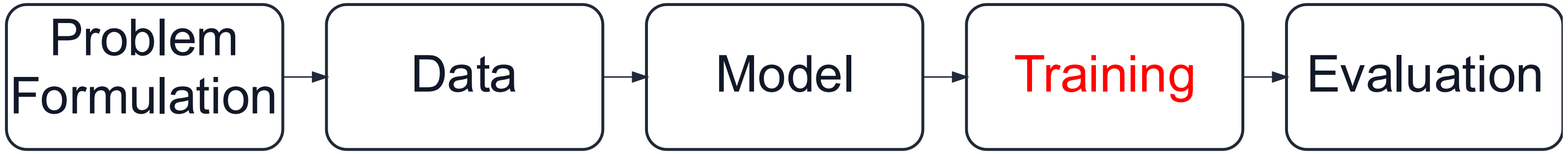
Taken from: https://www.devopsschool.com/blog/machine-learning-compare-supervised-learning-vs-unsupervised-learning-vs-reinforcement-learning/

SciKit Learn Cheat sheet



Taken from: https://scikit-learn.org/stable/machine_learning_map.html

3.4. Training: Model Set Up



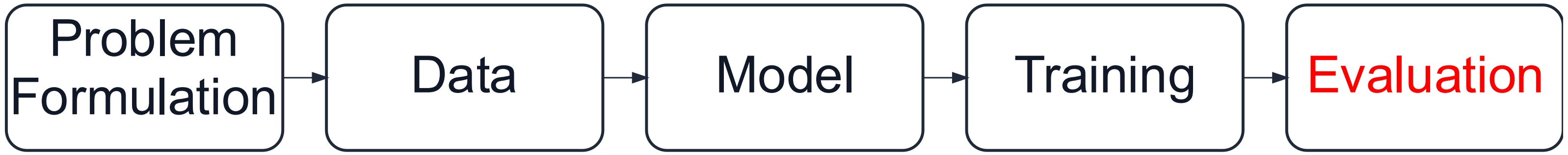
Testing and Training

- **Split** data → Train / Test
  - **Random split** e.g., 80% train / 20% test.
  - Use a **stratified split** for classification to keep class proportions similar in both sets (important with imbalance).
  - If data is time-ordered, use a **time-based split** instead of random.
- You may want a validation set: cross-validate + tune + select thresholds
- **Retrain** best setup on full Train
- **Evaluate once** on Test

Hyperparameter Tuning

- Each model has **specific setting** that you can specify **before training** – these are hyperparameters.
- These can be altered and tested on the data to make sure the model performs well on **unseen data**.
- Trying various settings and measuring performance on **validation data**, ensures these are set correctly.

3.5. Evaluation



Presenting your results

- Results should present a **baseline** and an **evaluation metric**
- A simple baseline is compared with your results to show (hopefully) improvement. They depend on data and model used
  - Classification – majority, random, or rule based, baseline
  - Regression – Mean or median value
  - Clustering – Majority cluster
- Many models are used alongside known evaluation metrics:
  - Classification – Confusion matrix, Precision/Recall/F1, ROC-AUC
  - Regression – MAE (Mean Average Error) or RMSE (Root Mean Average Square Error)
  - Clustering – Silhouette Score, Adjusted Rand Index (need ground truth), NMI (Normalised Mutual Information)

A note on: Reproducability

Someone (including you in the future) must be able to **re-run the same pipeline** and obtain the **same results** given the same data and code

- **Data**: details on the exact dataset used (including any cleaning steps)
- **Features**: how each feature was computed (code + parameters + time windows)
- **Model training**: algorithm, hyperparameters, training method
- **Evaluation**: exact train/test splits, baseline, metrics
- **Environment**: package versions, hardware differences

4. Ethics

Ethical Issues in Data Preparation

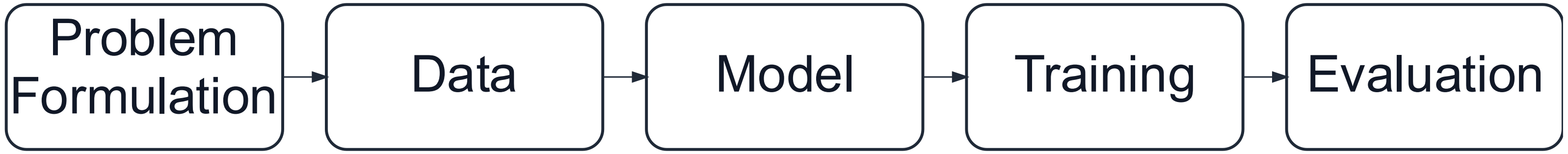
- **Sensitive attributes included** - directly, or via proxies like postcode
- **Historical bias** - labels reflect biased past decisions
- **Consent and legal basis** - GDPR/ethics constraints, data minimisation
- **Re-identification risk** - especially true with small groups and very granular location or time

More general note on Bias

- Machines are trained by humans, and human biases can be incorporated into algorithms — if biased information, or data that reflects existing inequities, is fed to a machine learning program, the program will learn to replicate it and perpetuate forms of discrimination.
- In some cases, machine learning models create or exacerbate social problems. For example, Facebook has used machine learning as a tool to show users ads and content that will interest and engage them — this can mean extreme content is shown which may lead to polarization and the spread of conspiracy theories when people are shown incendiary, partisan, or inaccurate content.

Adapted from: https://workofthefuture-taskforce.mit.edu/wp-content/uploads/2020/12/2020-Research-Brief-Malone-Rus-Laubacher2.pdf

5. Example ML Workflow

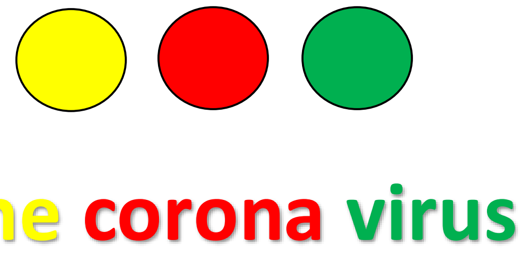


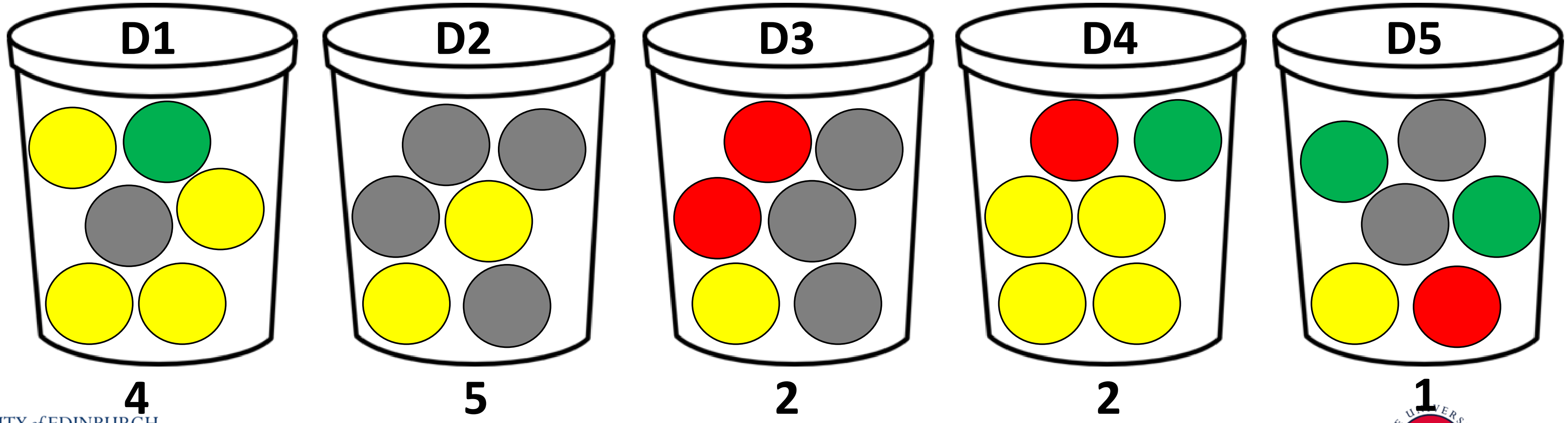
Example

- We are going to highlight the workflow with a text example, much of social communication online is text, text is often available to analyse on large scale
- Data are not always labelled to be analysed on scale
- Some classifiers are available to use
- Sometimes you need to build a specific classifier for a certain task

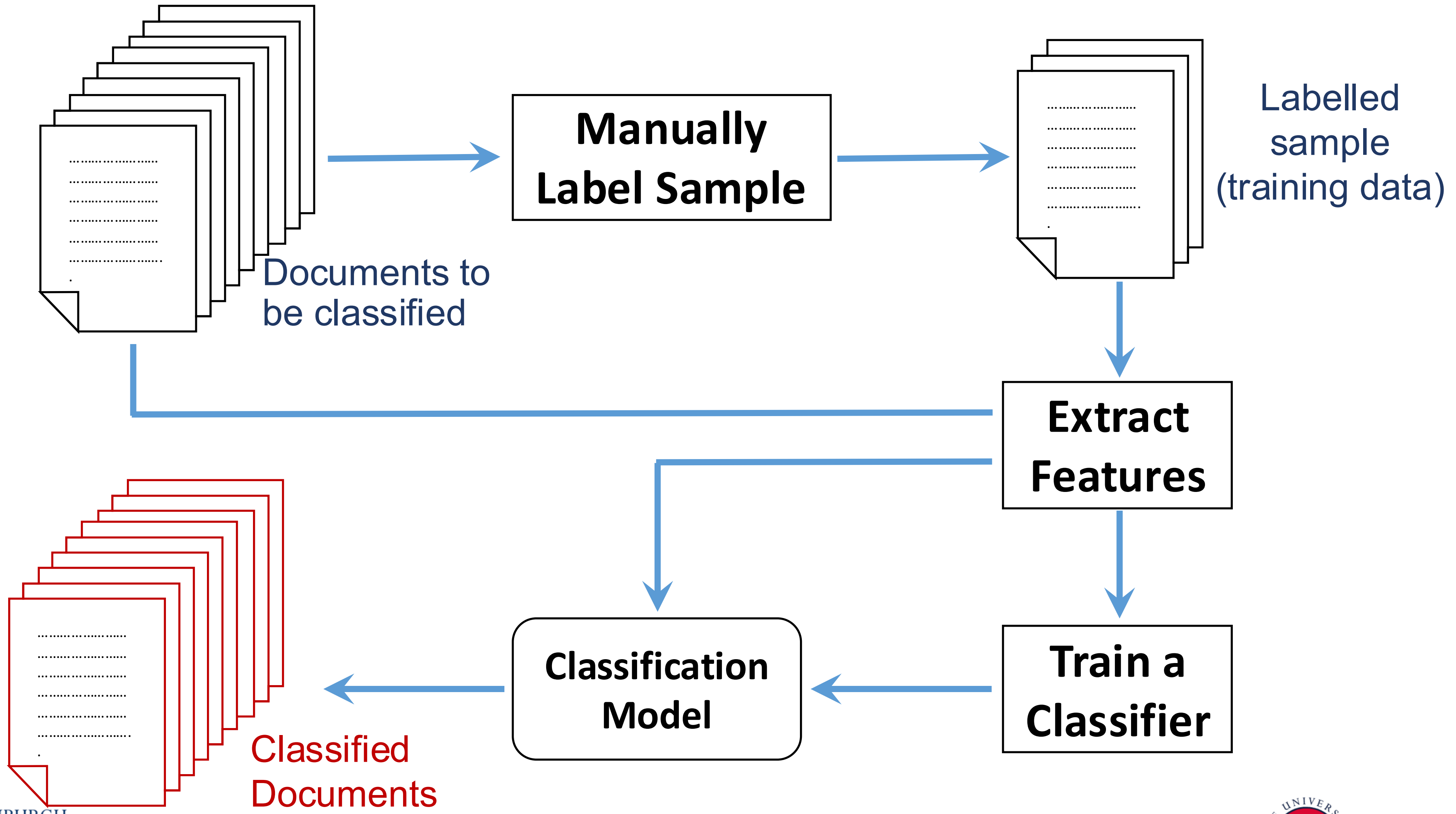
Formulating the problem

- Collection of 5 documents (balls = words)
- We want a system that returns the documents that fit a search query
- We determine relevant documents by how many features they match in the query (words)
- Which is the least relevant document?
- Which is the most relevant document?





Supervised-learning



Choosing a Model

- For binary classification, essentially any supervised learning algorithm can be used for training a classifier;
  classical choices include
  - Support vector machines (SVMs)
  - Random forests
  - Naïve Bayesian methods
  - Lazy learning methods (e.g., k-NN)
  - Logistic Regression
  - ….
- The "No-free-lunch principle" (Wolpert, 1996) → *there is no learning algorithm that can outperform all others in all contexts*
- Implementations need to cater for
  - the very high dimensionality
  - the sparse nature of the representations involved

Choosing a Model

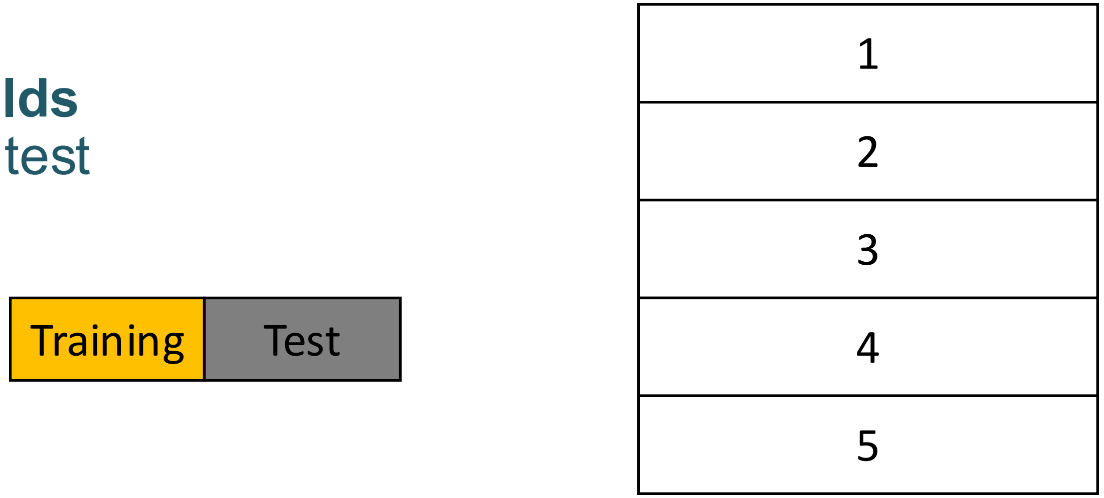
- For Multiclass classification, some learning algorithms for binary classification can be used:
  - Decision trees
  - Random forests
  - Naive Bayesian methods
  - Lazy learning methods (e.g., k-NN)
  - Neural networks
- Some must be used in combinations / cascades of the binary versions
  - e.g. multi-class classification SVM

Hyperparameter Optimisation of a Model

- Most classifiers has some parameters to be optimized:
  (we will usually refer to the ones we set manually as "hyperparameters" to distinguish from the "learned" parameters/weights of the model)
  - The *C* parameter in soft-margin SVMs
  - The *r*, *d* parameters of non-linear kernels
  - Decision threshold for binary SVM
- Optimising the hyperparameters on test data is cheating!
- *Data Split*: Usually labelled data would be split into three parts
  - Training: used to train the classifier (typically **80%** of the data)
  - Validation: used to optimise hyperparameters. Apply the classifier on this data with different values of the hyperparameters and report the one that achieves the highest results (usually **10%** of the data)
  - Test: used to test the performance of the trained classifier with the optimal hyperparameters on these unseen data (usually **10%** of the data)

Training: Cross-Validation

- Sometimes the amount of labelled data in hand is limited (e.g. 200 samples). Having evaluation of a set of 20 samples only might be misleading
- Cross-validation is used to train the classifier with all data and test on all data without being cheating
- Idea:
  - Split the labelled data into ***n* folds**
  - Train classifier on *n*-1 fold and test on the remaining one
  - Repeat *n* times
- 5-fold cross validation

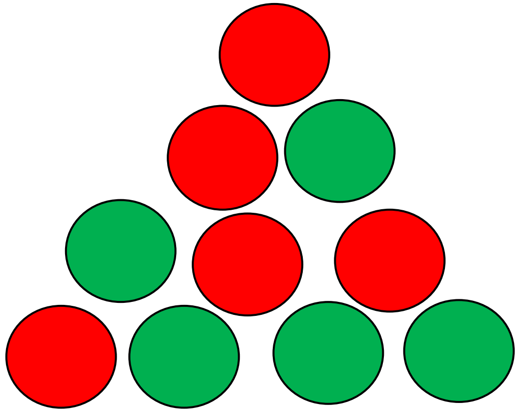


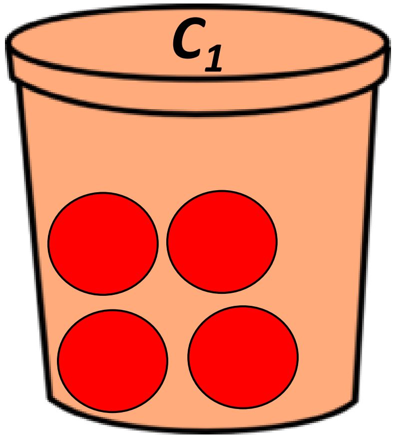
Evaluation: Baselines

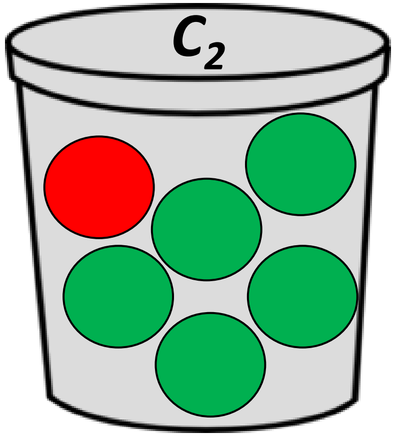
- There are standard methods for creating baselines the most popular/simplest baselines
  - Random classification
    - Classes are assigned randomly
    - How much better is the classifier doing than random?
  - Majority class baseline
    - Assign all elements to the class that appears the most
    - How much better you are doing than if you always picked the same thing output regardless of input?
  - Simple algorithm, e.g. BOW (Bag of words)
    - Usually used when you introduce new interesting features
  - Recently: BERT baseline
  - LLMs: zero-shot / few-shot baselines

Evaluation: Binary Classification

- Accuracy:
    - How many of the samples are classified correctly?
- A = 9/10 = 0.9

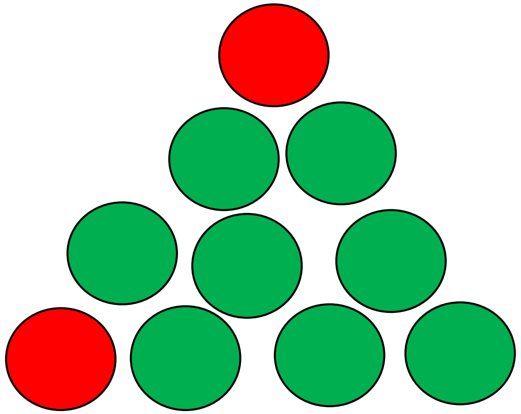


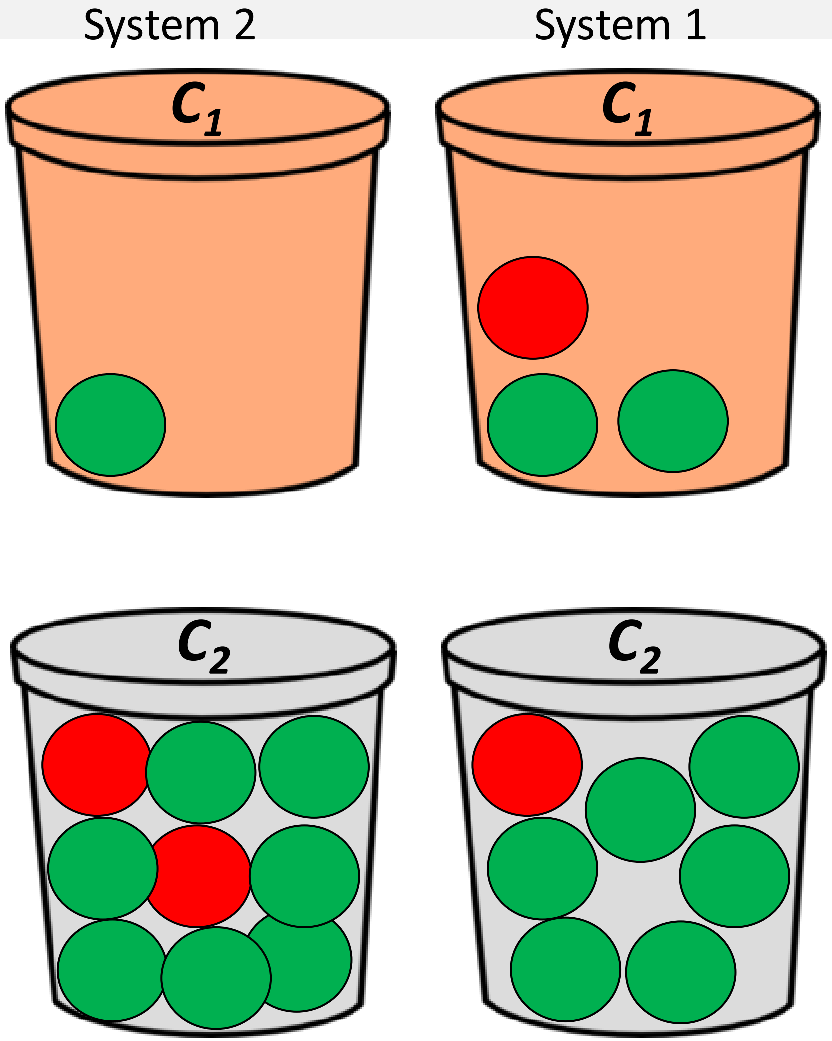




Evaluation: Binary Classification

- A = 7/10 = 0.7 System 1
- A = 7/10 = 0.7 System 2
- When classes are highly unbalanced
  - Precision/recall/F1 for the **rare class**
  - e.g. Spam classification (detection)





Evaluation Metrics: Precision and Recall

- **Precision**:
  What fraction of the classified as X are correct?

$$P = \frac{Classified\ correctly\ as\ X}{All\ samples\ classified\ as\ X}$$

- **Recall**:
  What fraction of the class X has been classified correctly?

$$R = \frac{Classified\ correctly\ as\ X}{Real\ number\ of\ the\ X\ samples}$$

Evaluation Metrics: F-measure

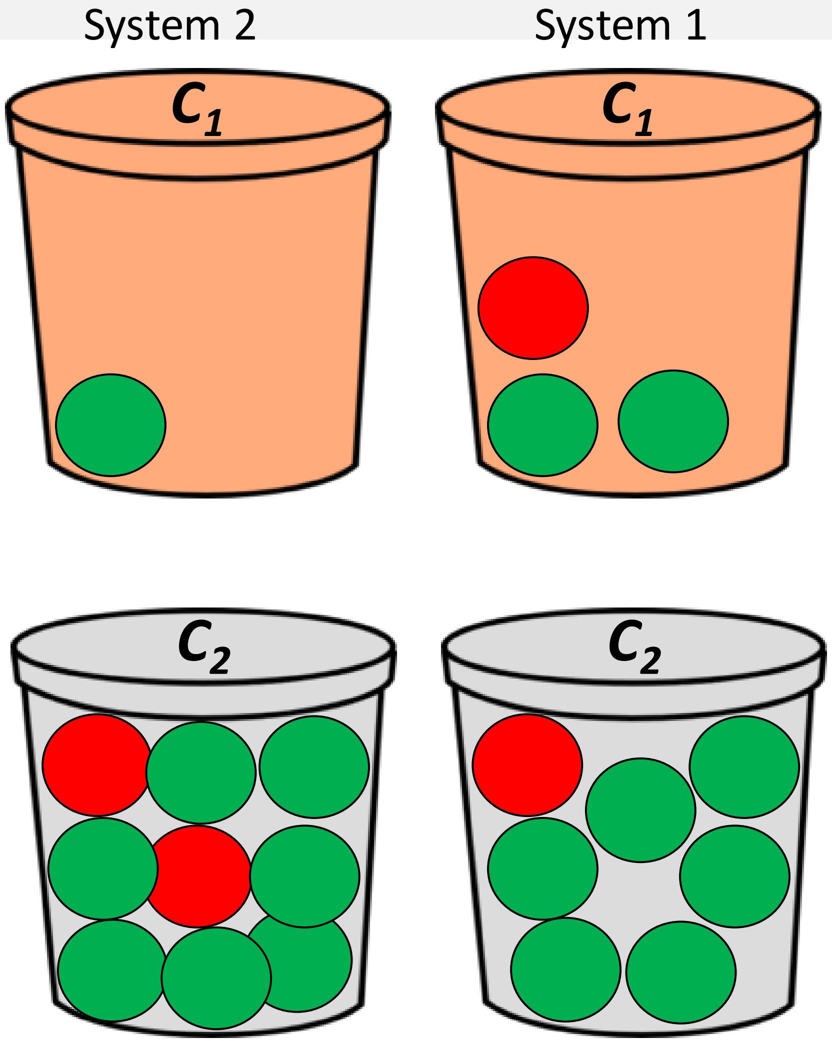
$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

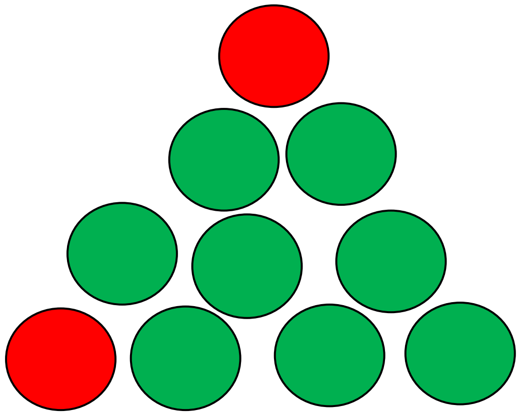
Harmonic mean of recall and precision

- Emphasizes the importance of small values, whereas the arithmetic mean is affected more by outliers that are unusually large

Evaluation: Binary Classification

	System 1	System 2
Precision	1/3 = 0.33	0/1 = 0
Recall	1/2 = 0.5	0/2 = 0
F1	0.4	0



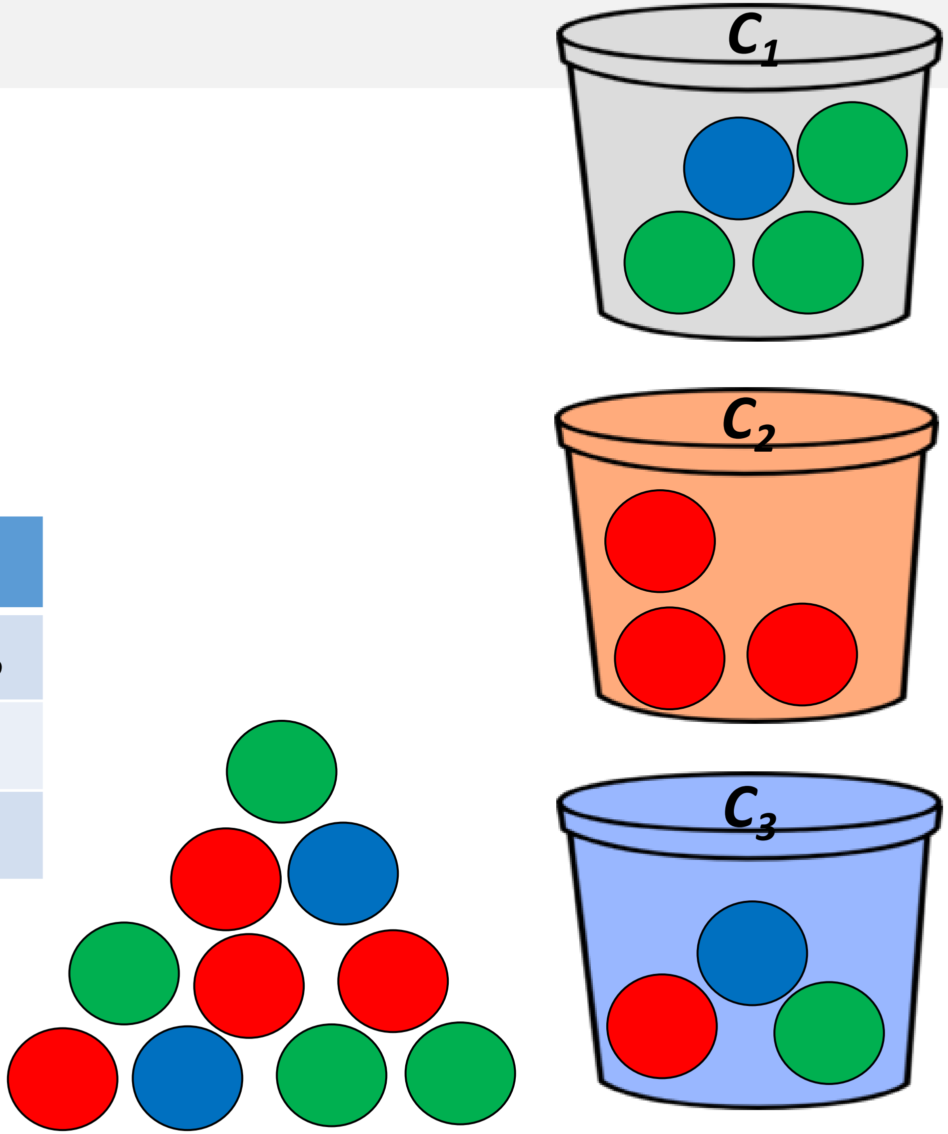


Evaluation: Multi-class

- Accuracy = (3+3+1)/10 = 0.7
- Good measure when
  - Classes are nearly balanced
- Preferred:
  - Precision/recall/F1 for each class

P	0.75	1	0.333
R	0.75	0.75	0.5
F1	0.75	0.86	0.4

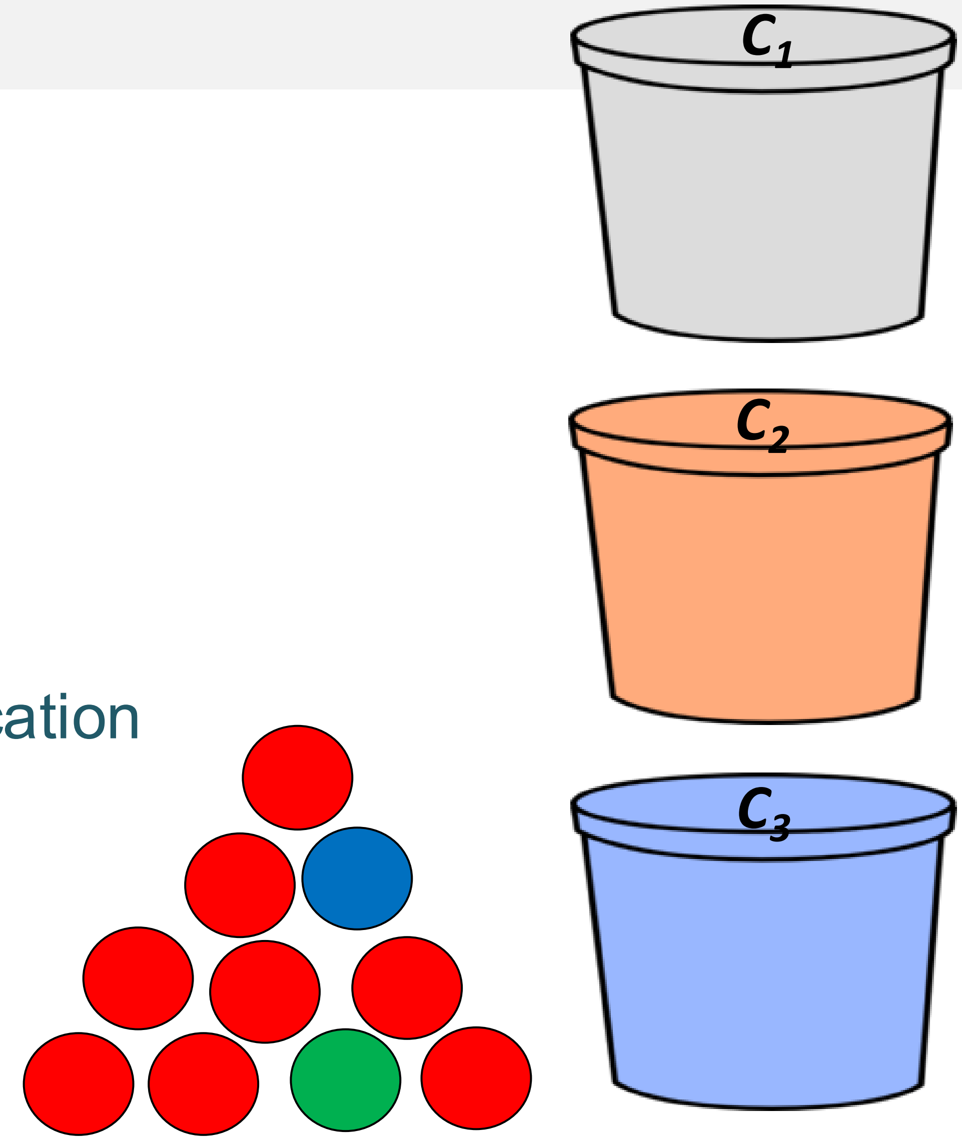
- **Macro-F1**
  = (0.75+0.86+0.4)/3
  = **0.67**



Evaluation: Multi-class

- Majority class baseline
- Accuracy = 0.8
- Macro-F1 = 0.296

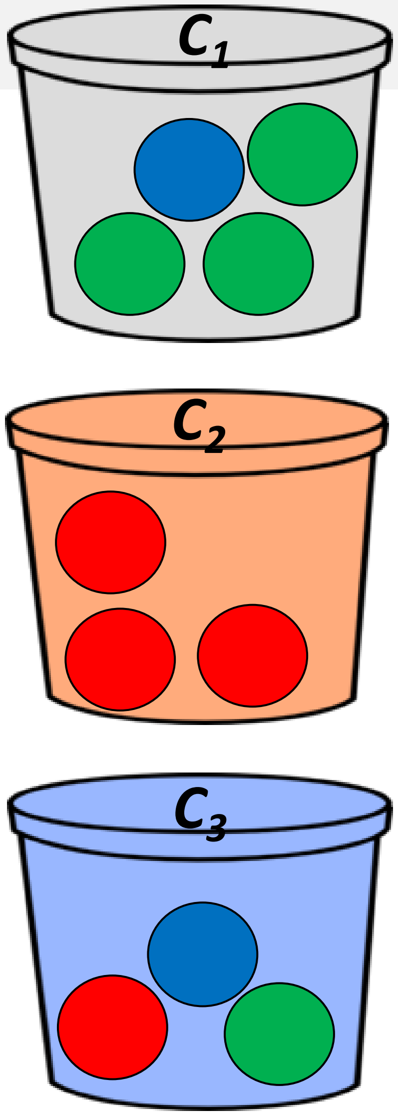
- Macro-F1:
  - Should be used in binary classification when two classes are important
  - e.g.: males/females while distribution is 80/20%



Evaluation Metrics: Error Analysis

- **Confusion Matrix**
  How classes get confused?

	3	0	1
	0	3	1
	1	0	1



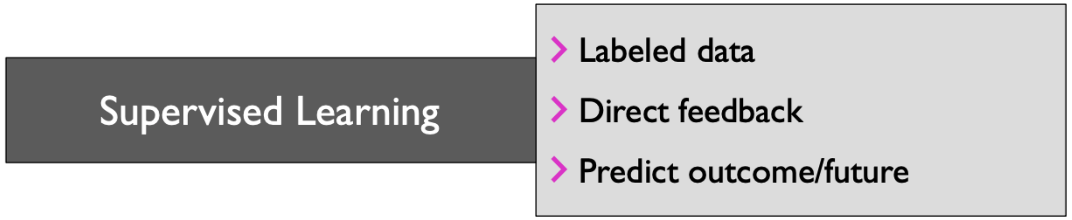
- Useful:
  - Find classes that get confused with others
  - Develop better features to solve the problem

BREAK

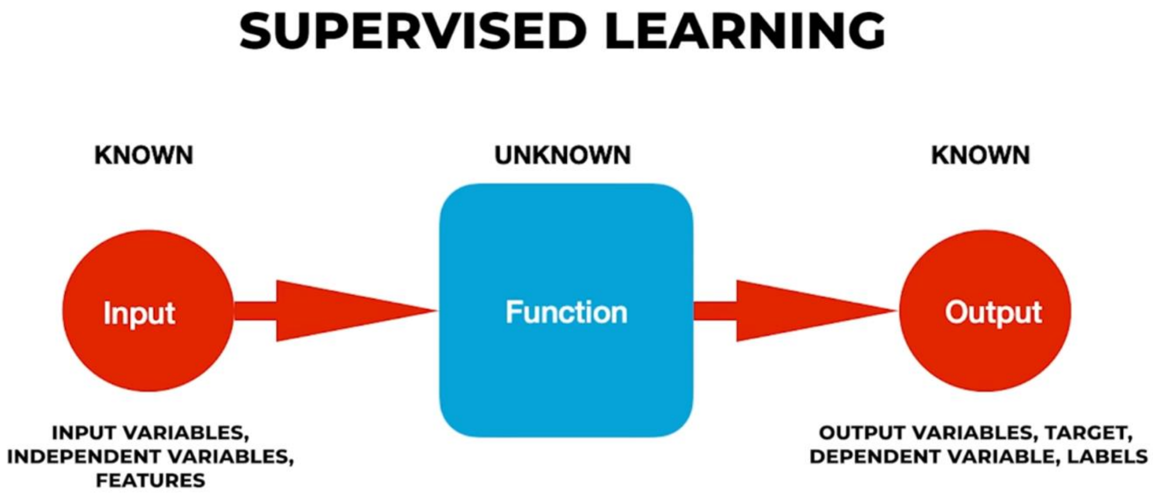
6. Whistle stop tour of ML algorithms

6.1. Supervised Machine Learning

Supervised machine learning



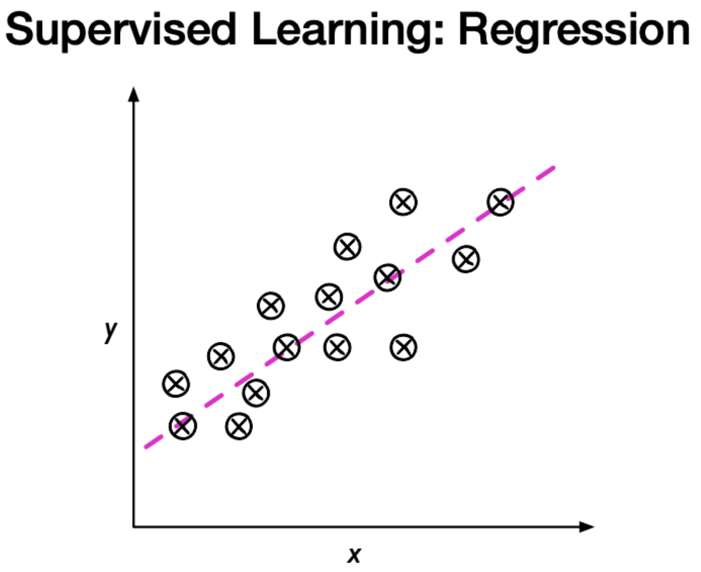
https://sebastianraschka.com/resources/ml-lectures-1/

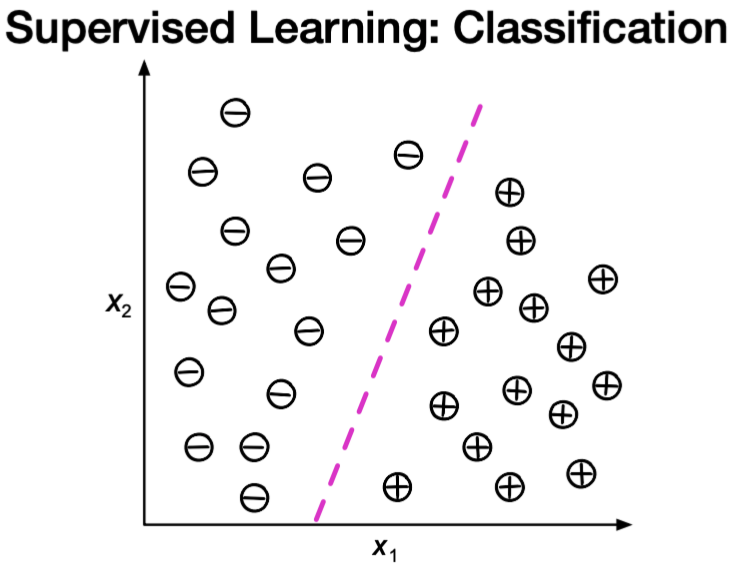


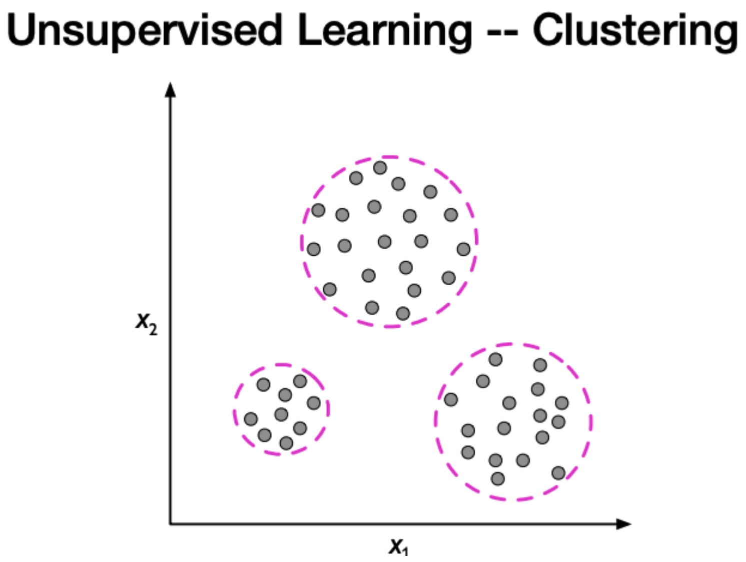
https://www.youtube.com/watch?v=E0Hmnixke2g

What model do I choose?

- **Classification** (spam vs not spam)
- **Regression** (predict house price)
- **Clustering / descriptive** (find groups)



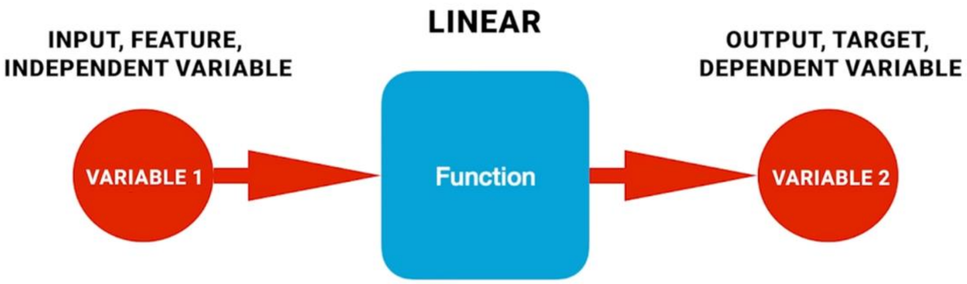




Adapted from: https://sebastianraschka.com/pdf/lecture-notes/stat451fs20/01-ml-overview__slides.pdf

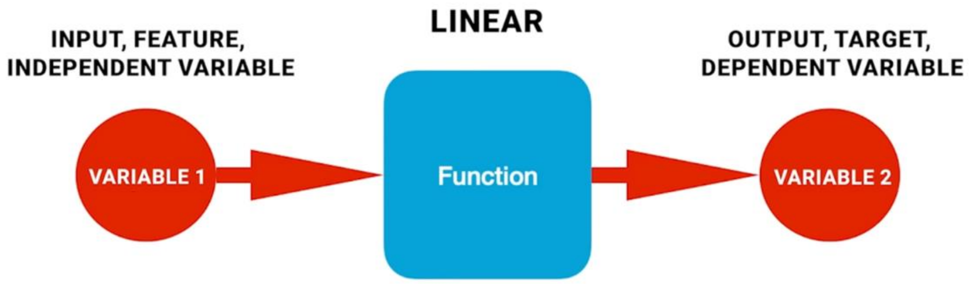
Linear Regression

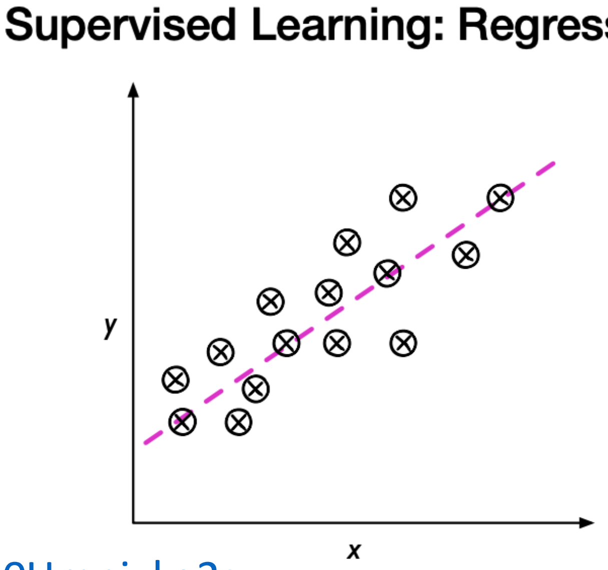
- This is a simple ML method which can be used to predict quantitative outcomes, It is the basis of more complex machine learning algorithms
- You know variable 1 and want to predict variable 2
- If there is a linear relationship between 1 and 2 and they are continuous we can use linear regression
- Note: it can be sensitive to outliers



https://www.youtube.com/watch?v=E0Hmnixke2g

Linear Regression





error

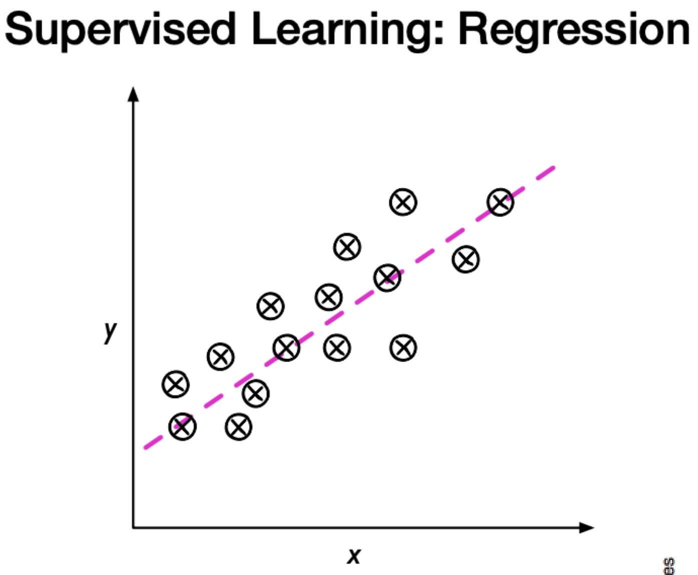
$$Y = \beta_0 + \beta_1 X + \epsilon$$

We are trying to work out the intercept and the slope of the line

https://www.youtube.com/watch?v=E0Hmnixke2g

Linear Regression

We evaluate the model by working out the “fit of the model” using the least squares method we need to know the Residual Sum of Squares (RSS) – higher RSE is worse



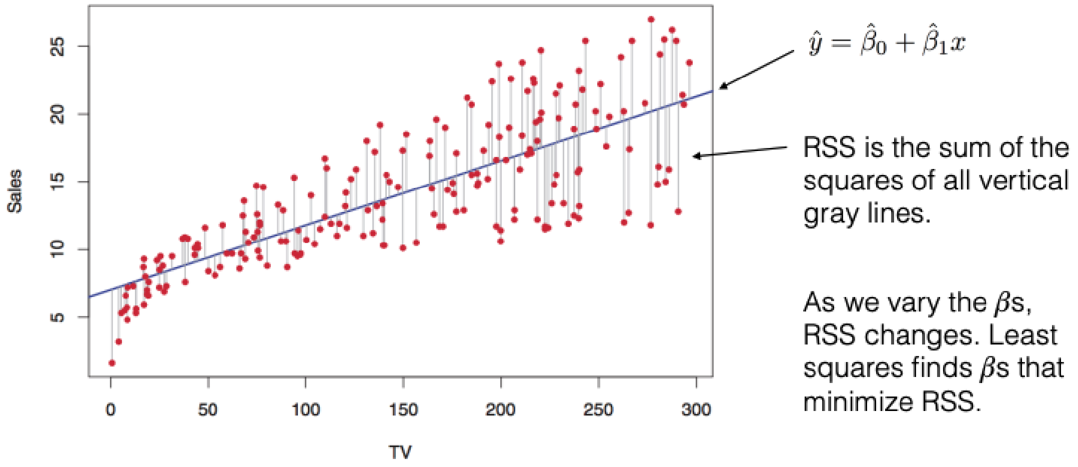


FIGURE 3.1, ISL (8th printing 2017)

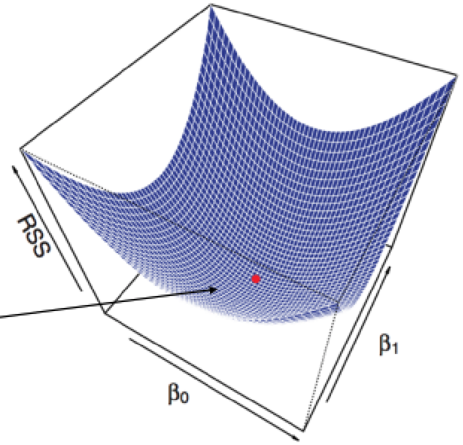
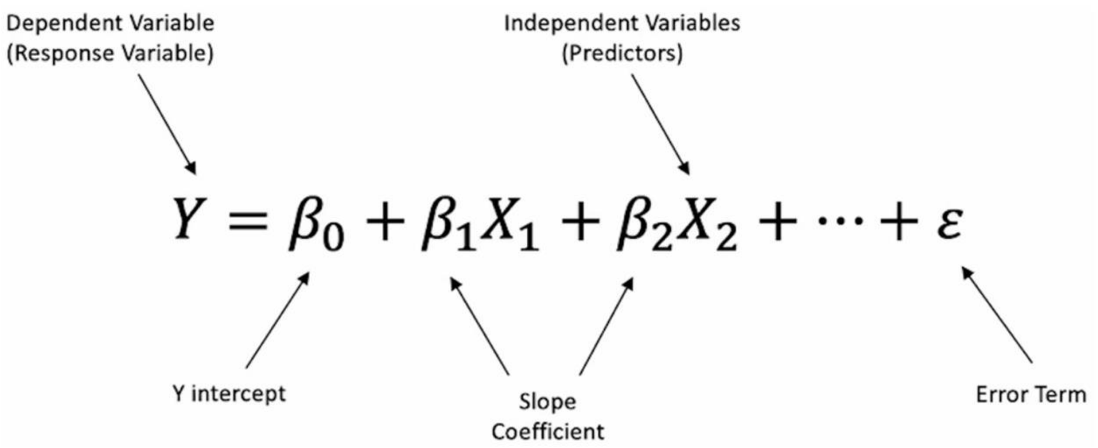
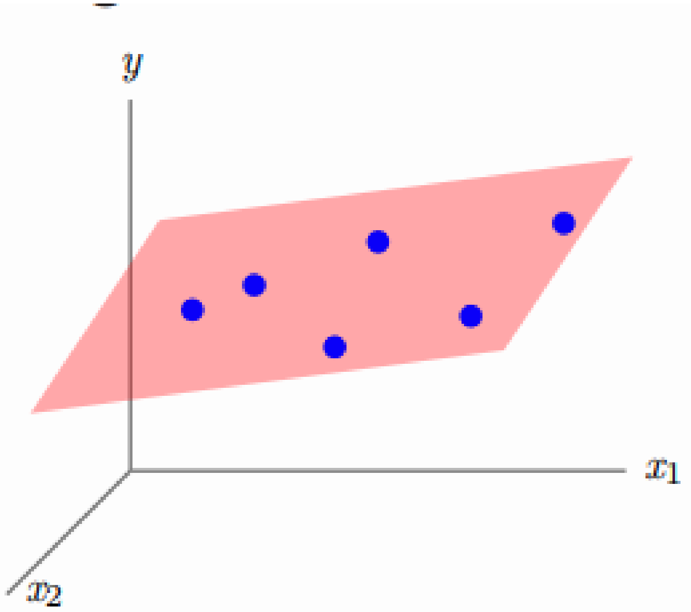


FIGURE 3.2, ISL (8th printing 2017)

Image from:
https://web.stanford.edu/class/cme250/files/cme250_lecture2.pdf

Multiple Linear Regression



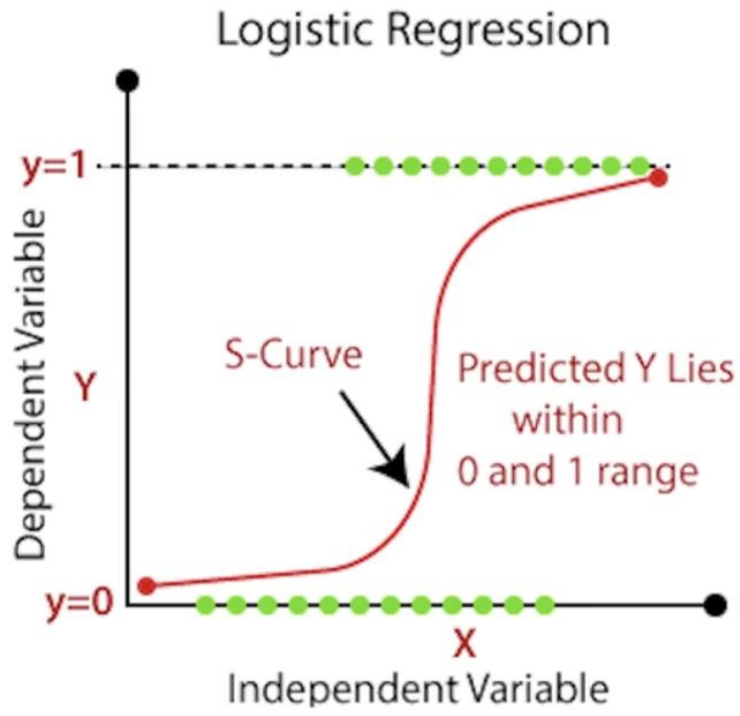


In 2D, instead of a line, we have a **plane**.
In higher dimensions, this would be a **hyperplane**.

[Image from: https://groups.inf.ed.ac.uk/teaching/aml/slides/W03L05-lin_reg.pdf](https://groups.inf.ed.ac.uk/teaching/aml/slides/W03L05-lin_reg.pdf)

Logistic Regression

- If we want to predict a class or categorical variable, we can fit a function to the data
- **It is classification**, most commonly **binary classification** (e.g., spam vs not spam or cat vs dog)
- Set threshold to obtain class decisions
- We use maximum likelihood to obtain the coefficients
- We can extend logistic regression to the case of multiple predictor variables.



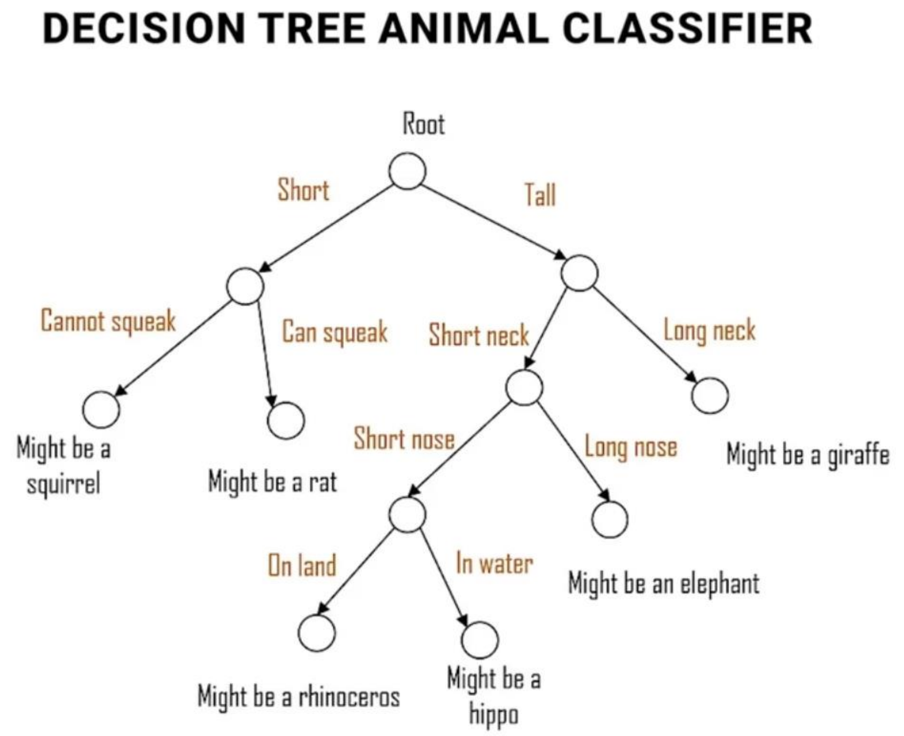
SIGMOID FUNCTION

https://www.youtube.com/watch?v=E0Hmnixke2g

$$Pr(Y = 1|X) = \sigma(\beta_0 + \beta_1 X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

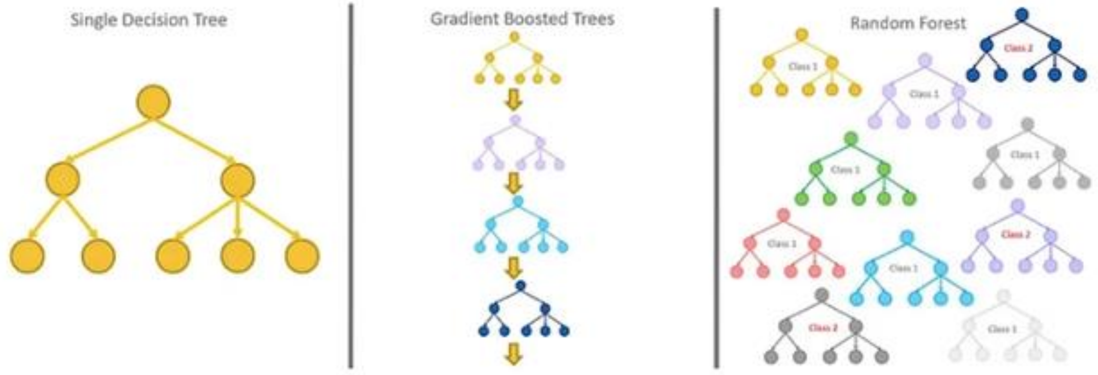
Decision Trees

- Classification
- A series of questions (such as yes or no) on features
- The leaf nodes (at the bottom of the tree are classes)
- Overfitting is a problem – each leaf node is a single example therefore must be pruned
- You can set a minimum number of samples per node.
- The algorithm finds the best feature to use for separating the examples at each split



Decision Trees

- The best split is determined by the information content of the features at each split. The higher the information gain the better a candidate for splitting. this is done through the concept of entropy.
- Hyperparameters can be used to prune the tree
- Can be evaluated through the purity of the clusters Gini index or gain ratio
- They can be grouped together into gradient boost (sequential trees to correct errors) or random forest (many trees averaged)



Naïve Bayes

- Classification
- Uses probability to work out the classes.
- What us the probability of A occurring if B has occurred.

Probability of B occurring given evidence A has already occurred

Probability of A occurring

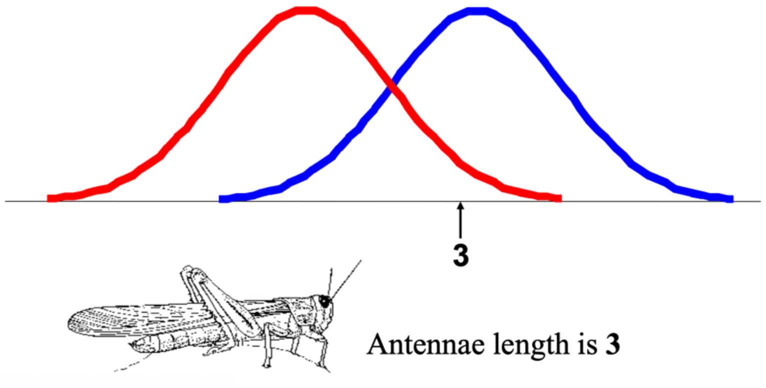
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of A occurring given evidence B has already occurred

Probability of B occurring

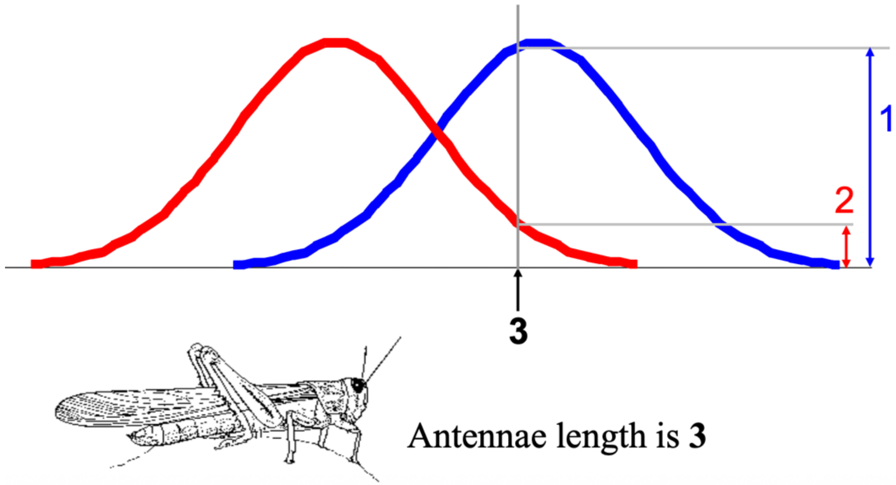
• We can just ask ourselves, give the distributions of antennae lengths we have seen, is it more *probable* that our insect is a **Grasshopper** or a **Katydid**.
• There is a formal way to discuss the most *probable* classification…

$p(c_j | d)$ = probability of class c_j, *given* that we have observed d



P(**Grasshopper** | **3**) = 10 / (10 + 2) = 0.833

P(**Katydid** | **3**) = 2 / (10 + 2) = 0.166



https://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf

Naïve Bayes

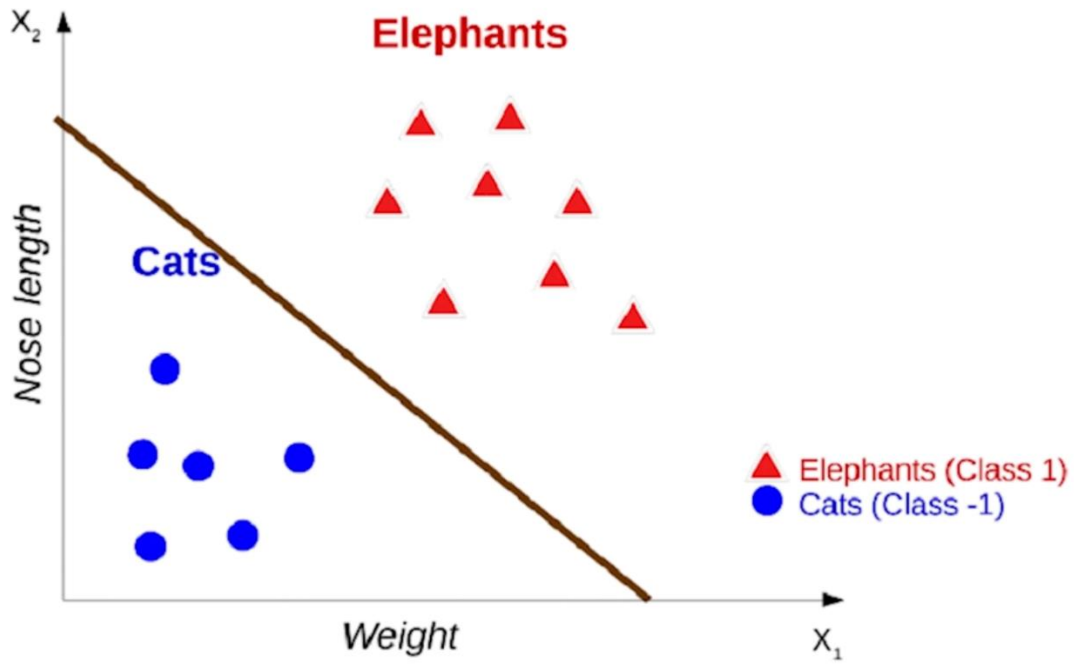
- It is fast and works best with lots of features
- Commonly used in text classification
- Word counts are used as features
- The class can be determined through the probability of a word being present in that class (spam or ham emails)
- It assumes all words are independent features (which they often are not.

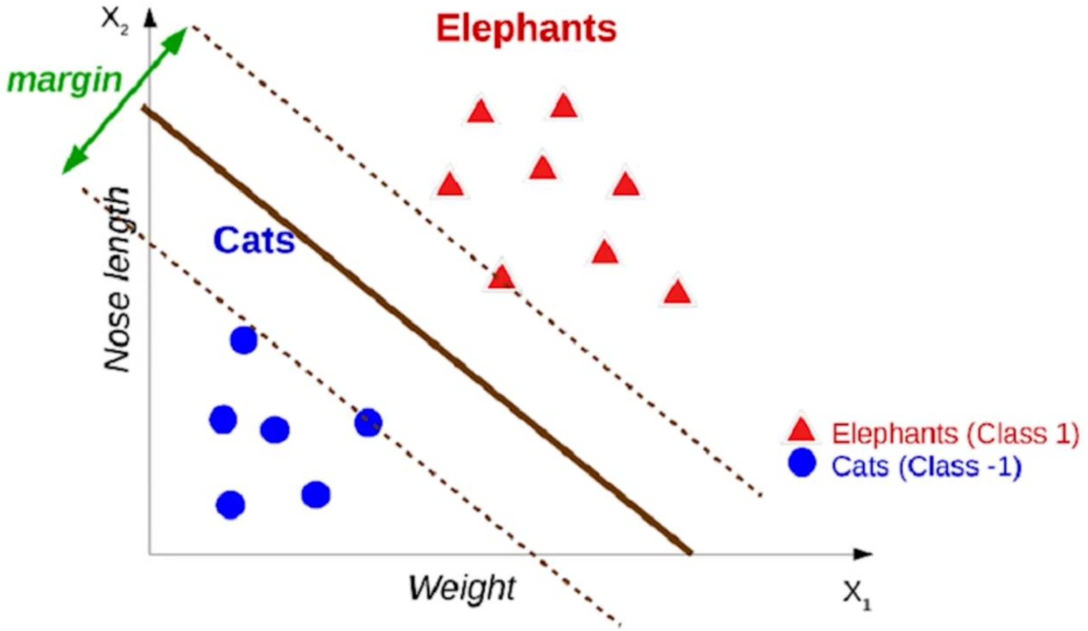
	Label	SMS
0	spam	SECRET PRIZE! CLAIM SECRET PRIZE NOW!!
1	ham	Coming to my secret party?
2	spam	Winner! Claim secret prize now!

↓

	Label	secret	prize	claim	now	coming	to	my	party	winner
0	spam	2	2	1	1	0	0	0	0	0
1	ham	1	0	0	0	1	1	1	1	0
2	spam	1	1	1	1	0	0	0	0	1

Support Vector Machines

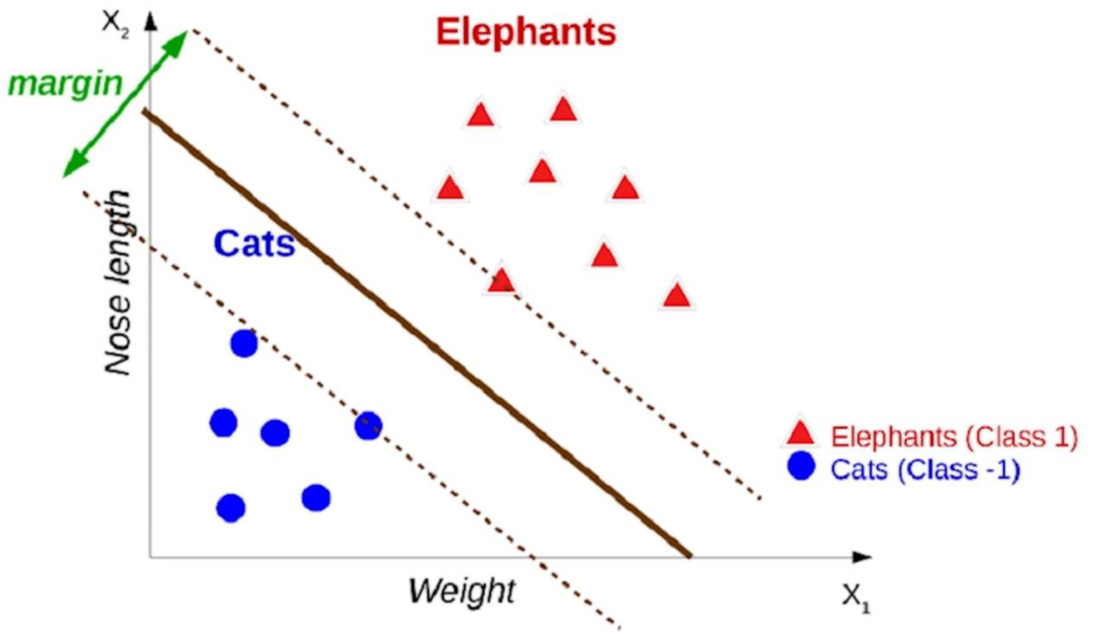




https://www.youtube.com/watch?v=E0Hmnixke2g

Support Vector Machines

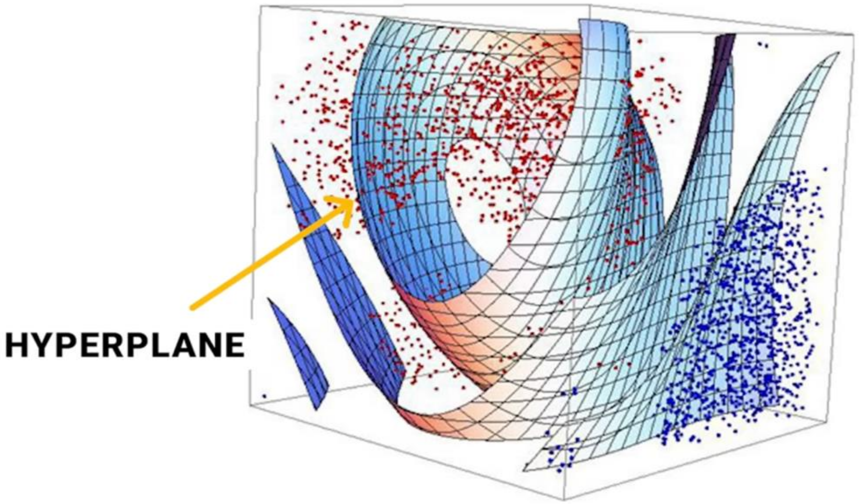
- The support vectors are the points that sit on the line
- The aim is the get the biggest margin
- Uses a subset of training points in the decision function to determine these decision points
- Work well in high dimensional space (lots of features)
- Common metrics: accuracy, F1, ROC
- Can be slow, hyperparameters need choosing (including the kernel)



https://www.youtube.com/watch?v=E0Hmnixke2g

Support Vector Machines

- In high dimensions the boundary is called a hyperplane
- Kernel functions can make the decision boundary nonlinear Common kernels: **Linear:** good when you have lots of features (e.g., text) **RBF (Gaussian):** flexible non-linear boundary (very common) **Polynomial:** non-linear with polynomial interactions
- This allows for implicit feature engineering – taking a feature and creating a new one through a function.



https://www.youtube.com/watch?v=E0Hmnixke2g

6.2. Unsupervised Machine Learning

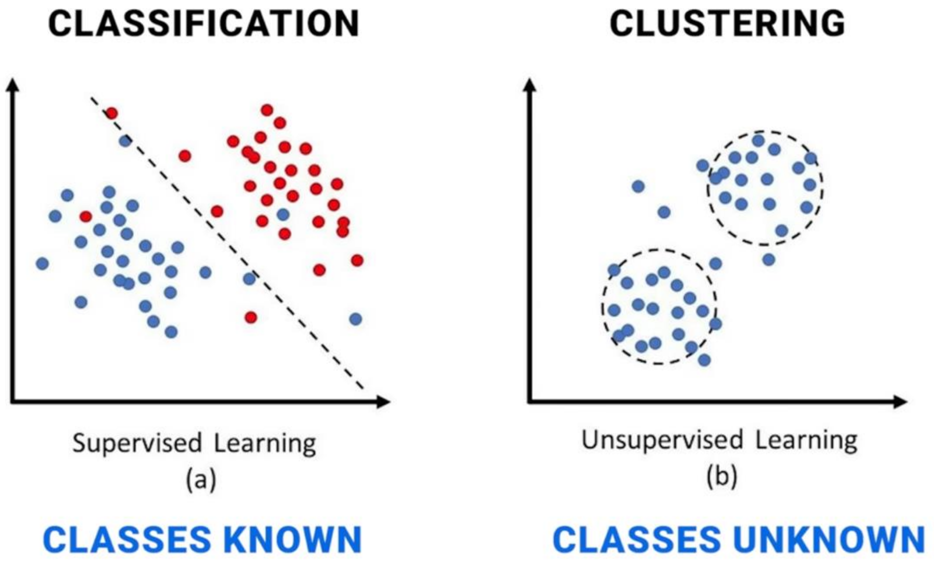
When to use

Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

Unsupervised Learning

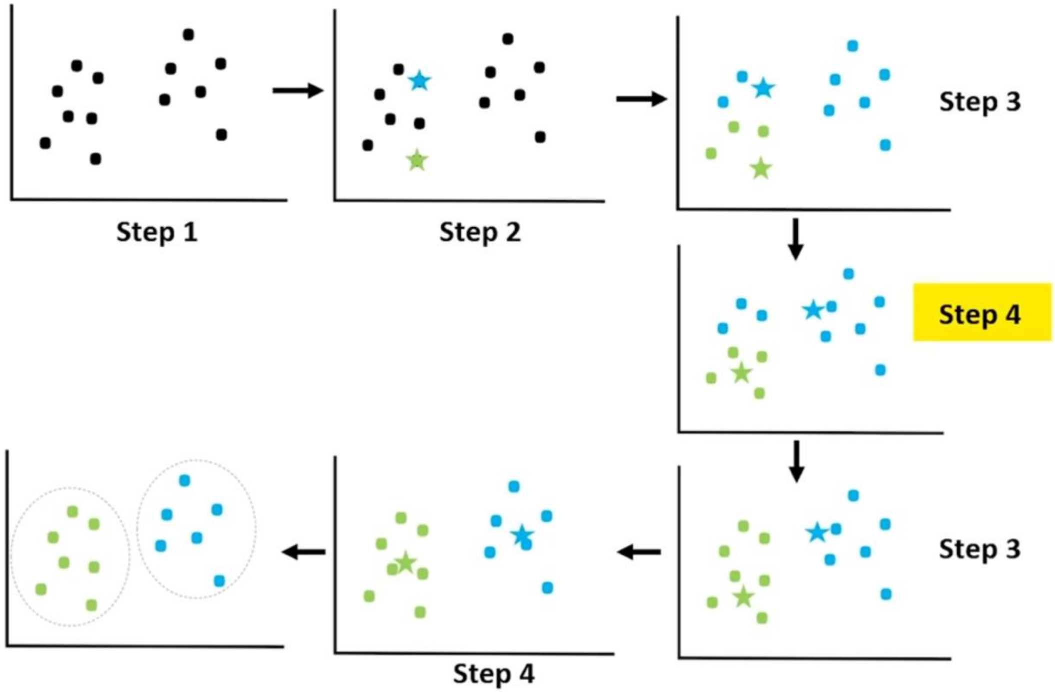
- No labels/targets
- No feedback
- Find hidden structure in data



https://sebastianraschka.com/resources/ml-lectures-1/ https://www.youtube.com/watch?v=E0Hmnixke2g

K means

- Choose 'K': Decide the number of clusters (k) you want.
- Initialize Centroids: Randomly pick 'k' data points as initial cluster centers (centroids).
- Assign Points: Assign each data point to the nearest centroid.
- Recalculate Centroids: Compute the new mean (average) of all points in each cluster to find its new centroid.
- Repeat: Reassign points and recalculate centroids until cluster assignments stop changing or a maximum number of iterations is reached
- Minimise the distortion function, i.e., the sum of the squared distances of each data point to its closest vector
- Evaluate using a Silhouette score (higher is better)



https://www.youtube.com/watch?v=E0Hmnixke2g

Silhouette Score

For a data point i:

•$a(i)$ = average distance from i to all other points in its **own** cluster (cohesion)

•$b(i)$ = smallest average distance from i to points in any **other** cluster (separation)

The silhouette for point i is:

The **overall silhouette score** is the mean of $s(i)$ over all points.

Topic Modelling

- **Topic modelling** discovers **hidden themes ("topics")** in a collection of documents by looking for patterns of word co-occurrence.
- A **topic** is typically represented as a **probability distribution over words**, and each **document** is represented as a **mixture of topics**.
- You need to adjust hyperparameters
- Evaluated using **topic coherence, and human interpretability**

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
world	lol	blog	love	baby
cup	haha	post	#xfactor	kids
england	good	updated	factor	family
#worldcup	dont	comment	big	children
football	yeah	published	cheryl	school
south	hey	entry	amazing	child
spain	love	blogs	show	parents
africa	hope	blogging	live	fun
game	gonna	posts	john	great
germany	time	posting	brother	toys

https://www.cl.cam.ac.uk/teaching/1718/L101/slides5.pdf

For the Lab: KNIME

- Machine Learning: https://www.knime.com/sites/default/files/2021-08/l4-ml-slides.pdf
- L4-ML: https://knime.learnupon.com/content-details/4387303/0 (you need to be logged in)
- General Book: https://www.knime.com/sites/default/files/public/2024-01/knime-press-beginners-luck-5.2-plain.pdf
- List of resources: https://www.knime.com/knimepress

Any questions?

Classifier Evaluation

We can evaluate all classification models (such as logistic regression) using Precision, Recall and F1, Confusion Matrices and ROC

Predicted class

True class	0	1
0	True Positive (TP)	False Negative (FN)
1	False Positive (FP)	True Negative (TN)

Precision:
What fraction of the classified as X are correct?

$$P = \frac{Classified\ correctly\ as\ X}{All\ samples\ classified\ as\ X}$$

Recall:
What fraction of the class X has been classified correctly?

$$R = \frac{Classified\ correctly\ as\ X}{Real\ number\ of\ the\ X\ samples}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$