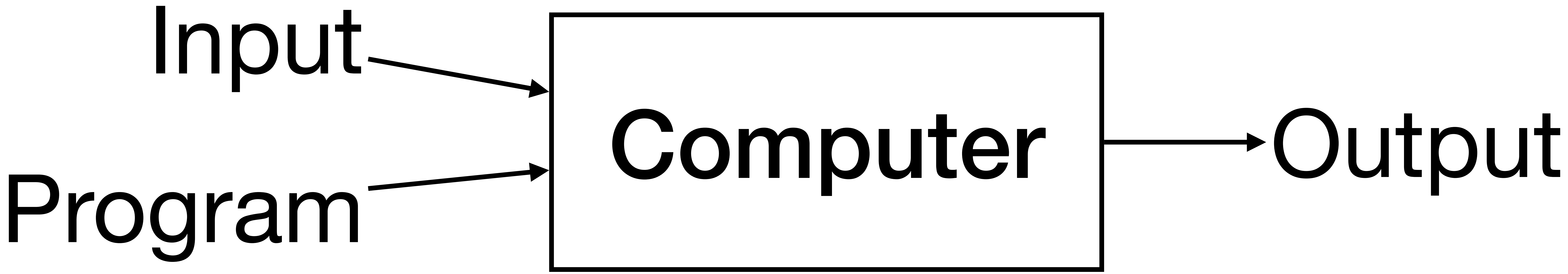
Introduction to Machine Learning

Lachlan Astfalck
School of Physics, Mathematics and Computing & School of Earth and Oceans
The University of Western Australia

The von Neumann model of computing



1943: theoretical model for neural networks

BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

I. *Introduction*

1958: first hardware implementation

Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN[1]

F. ROSENBLATT

Cornell Aeronautical Laboratory

If we are eventually to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking, we must first have answers to three fundamental questions:

1. How is information about the physical world sensed, or detected, by the biological system?
2. In what form is information stored, or remembered?
3. How does information contained in storage, or in memory, influence recognition and behavior?

and the stored pattern. According to this hypothesis, if one understood the code or "wiring diagram" of the nervous system, one should, in principle, be able to discover exactly what an organism remembers by reconstructing the original sensory patterns from the "memory traces" which they have left, much as we might develop a photographic negative, or translate the pattern of electrical charges in the "memory" of a digital computer. This hypothesis is appealing in its simplicity and ready intelligibility, and a large family of theoretical brain

1967: first deep learning implementation

IEEE TRANSACTIONS ON ELECTRONIC COMPUTERS, VOL. EC-16, NO. 3, JUNE 1967 299

A Theory of Adaptive Pattern Classifiers

SHUNICHI AMARI

Abstract—This paper describes error-correction adjustment procedures for determining the weight vector of linear pattern classifiers under general pattern distribution. It is mainly aimed at clarifying theoretically the performance of adaptive pattern classifiers. In the case where the loss depends on the distance between a pattern vector and a decision boundary and where the average risk function is unimodal, it is proved that, by the procedures proposed here, the weight vector converges to the optimal one even under nonseparable pattern distributions. The speed and the accuracy of convergence are analyzed, and it is shown that there is an important tradeoff between speed and accuracy of convergence. Dynamical behaviors, when the probability distributions of patterns are changing, are also shown. The theory is generalized and made applicable to the case with general discriminant functions, including piecewise-linear discriminant functions.

Index Terms—Accuracy of learning, adaptive pattern classifier, convergence of learning, learning under nonseparable pattern distribution, linear decision function, piecewise-linear decision function, rapidity of learning.

needs a parametric treatment, that is, the distributions must be limited to those of a certain known kind whose distributions can be specified by a finite number of parameters. Moreover, the discriminant functions thus obtained depend directly on all of the past patterns so that they are not able to quickly follow the sudden change of the distributions. In order to avoid these shortcomings, we shall propose nonparametric learning procedures, by which the present discriminant function is modified according only to the present misclassified pattern.

The steepest-descent method is often used in order to minimize a known function. However, in our learning situation, we cannot obtain the descending directions of the average risk which we intend to minimize, because the probability distributions of the patterns are unknown. What we can utilize is the present pattern only,

1980: first convolutional neural network

Biol. Cybernetics 36, 193–202 (1980)

Biological Cybernetics

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

Abstract. A neural network model for a mechanism of visual pattern recognition is proposed in this paper. The network is self-organized by "learning without a teacher", and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their positions. This network is given a nickname "neocognitron". After completion of self-organization, the network has a structure similar to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel. The

reveal it only by conventional physiological experiments. So, we take a slightly different approach to this problem. If we could make a neural network model which has the same capability for pattern recognition as a human being, it would give us a powerful clue to the understanding of the neural mechanism in the brain. In this paper, we discuss how to synthesize a neural network model in order to endow it an ability of pattern recognition like a human being.

Several models were proposed with this intention

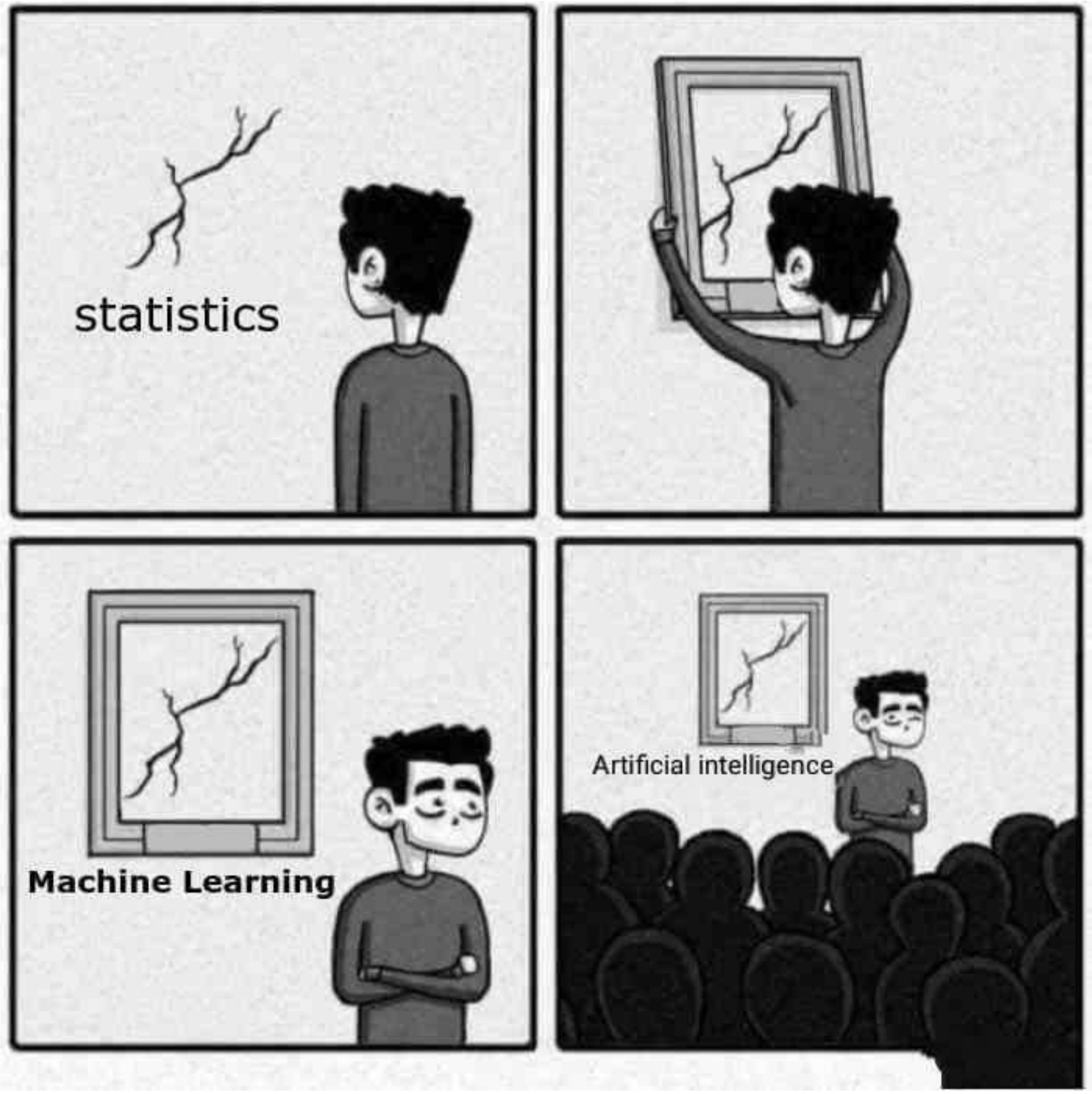
2011: deep learning is superhuman

2011: DanNet triggers deep CNN revolution

Abstract. In 2021, we are celebrating the 10-year anniversary of DanNet, named after my outstanding Romanian postdoc Dan Claudiu Cireșan (*aka* Dan Ciresan). In 2011, DanNet was the first pure deep convolutional neural network (CNN) to win computer vision contests. For a while, it enjoyed a monopoly. From 2011 to 2012 it won every contest it entered, winning four of them in a row (15 May 2011, 6 Aug 2011, 1 Mar 2012, 10 Sep 2012), driven by a very fast implementation based on graphics processing units (GPUs). Remarkably, already in 2011, DanNet achieved the first superhuman performance in a vision challenge, although compute was still 100 times more expensive than today. In July 2012, our CVPR paper on DanNet hit the computer vision community. The similar AlexNet joined the party in Dec 2012. Our even much deeper Highway Net (May 2015) and its variant ResNet (Dec 2015) further improved performance (a ResNet is a Highway Net whose gates are always open). Today, a decade after DanNet, everybody is using fast deep CNNs for computer vision.

What took so long? Why now?

1. Training data (data explosion and open source)
2. Optimisation algorithms (e.g. back-propagation)
3. Computation (e.g. GPUs, TPUs, cloud computing)
4. Funding and investment


statistics
Machine Learning
Artificial intelligence


HEY ML WHY DO YOU ALWAYS WEAR THAT MASK?
Machine Learning
Machine Learning
Statistics
LET'S KEEP THIS ON.
Machine Learning
imgflip.com

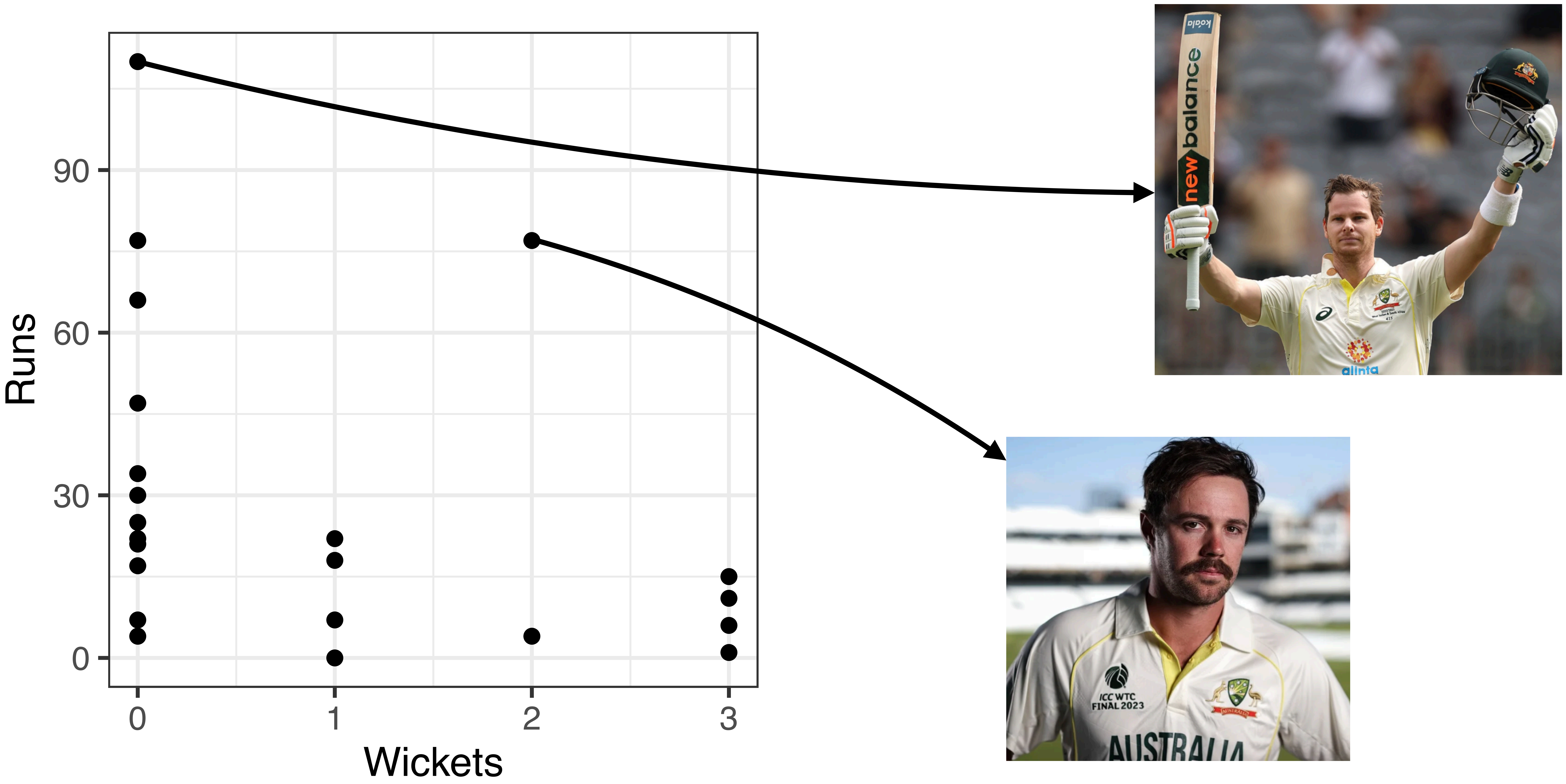
Statistics vs ML vs AI

- **Statistics** aims to infer conclusions from data, provide estimates/inference, tests hypotheses and causality.
- **Machine learning** aims to create predictive models that can generalise well to new unseen data. Does not necessarily require interpretability.
- **Artificial intelligence** is a broader field encompassing systems capable of performing tasks that normally require human intelligence. Includes ML, natural language processing, robotics, computer vision.

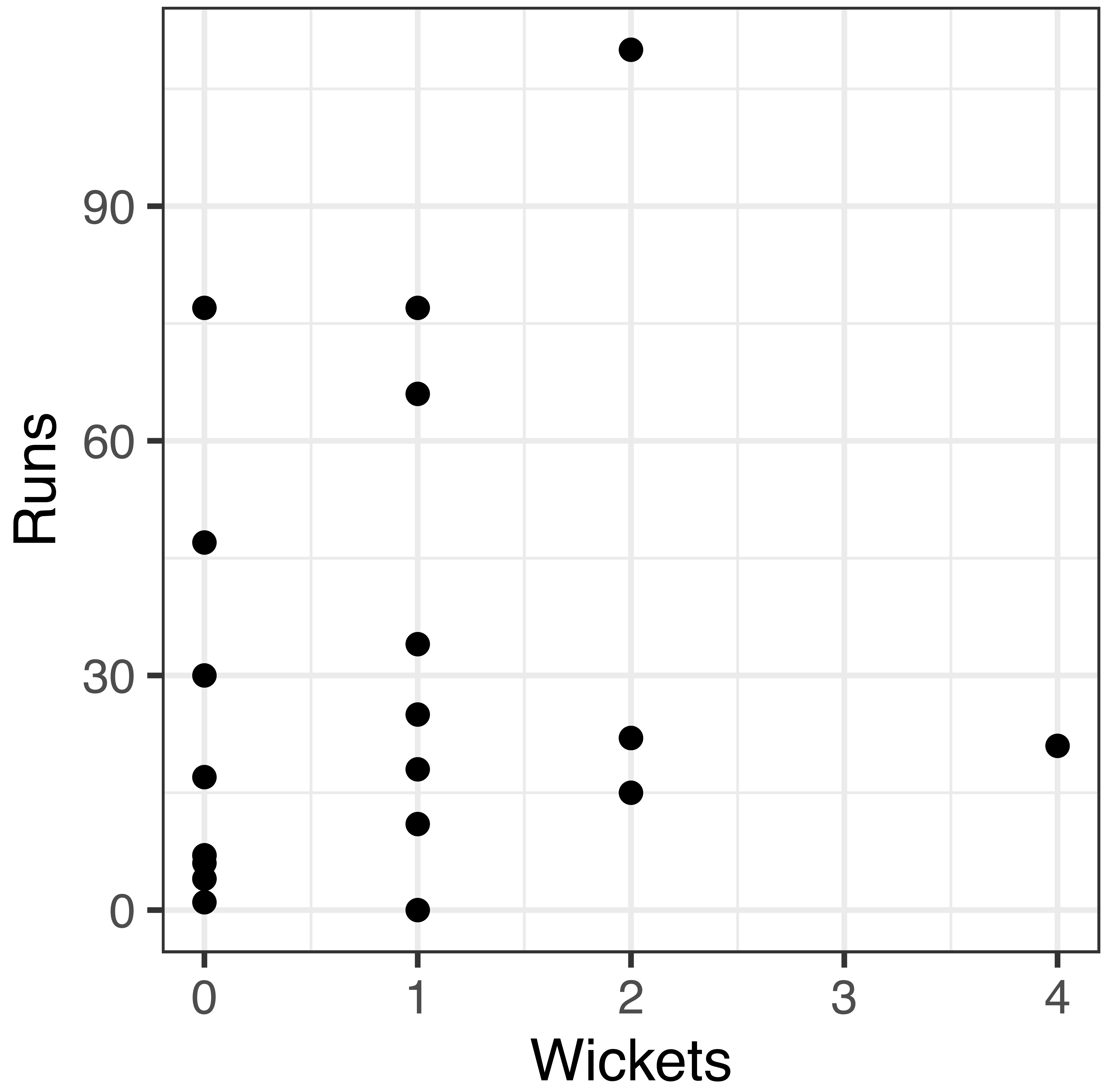
Components of any ML analysis

1. A Goal
2. Training data
3. Features
4. The Model Architecture
5. Inference/Optimisation
6. Validation

2023 Ashes, 2nd Test, Lord’s



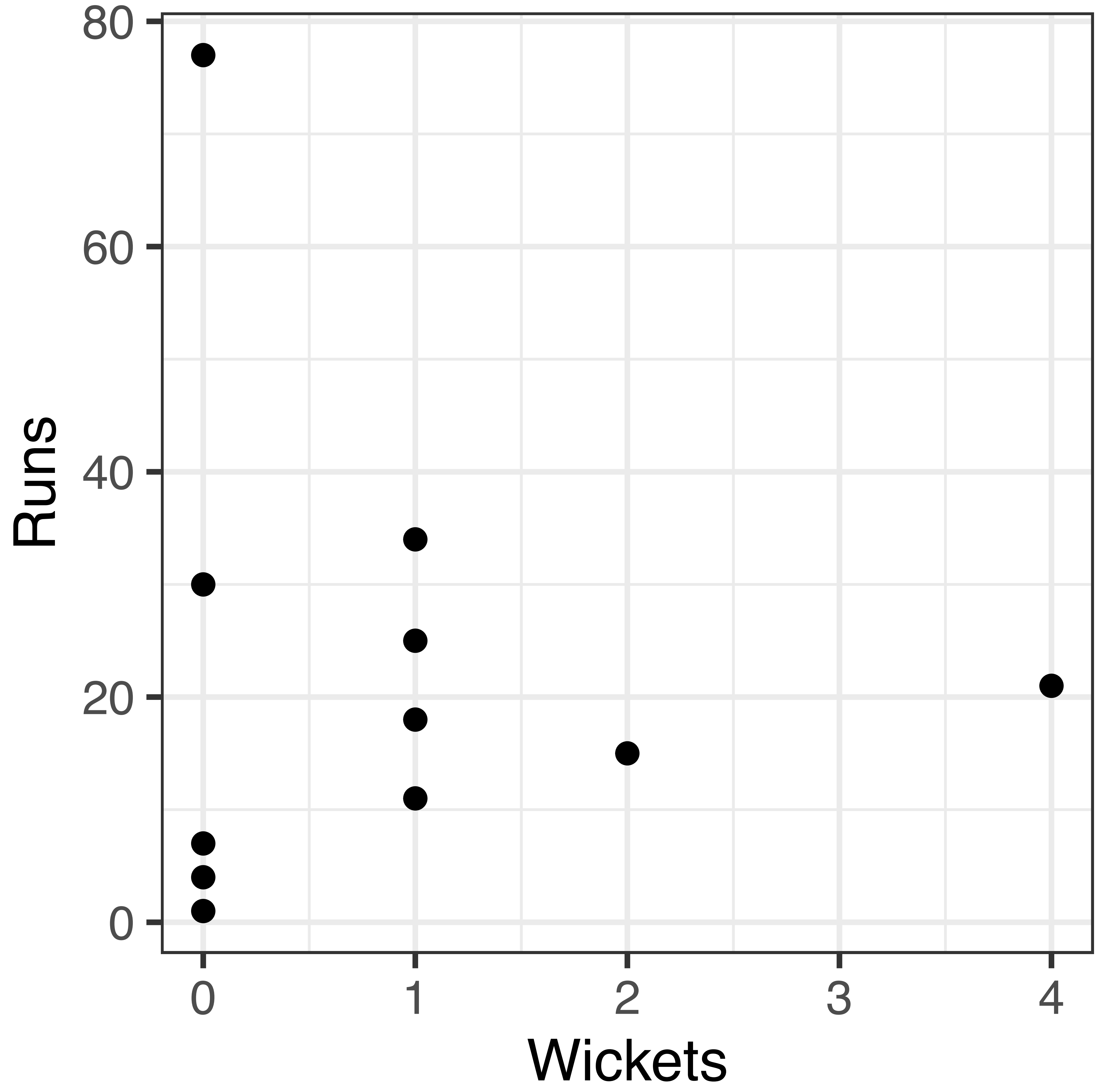
2023 Ashes, 2nd Test, Lord's



1. The Goal

Can we classify each player as either a batsman or a bowler?

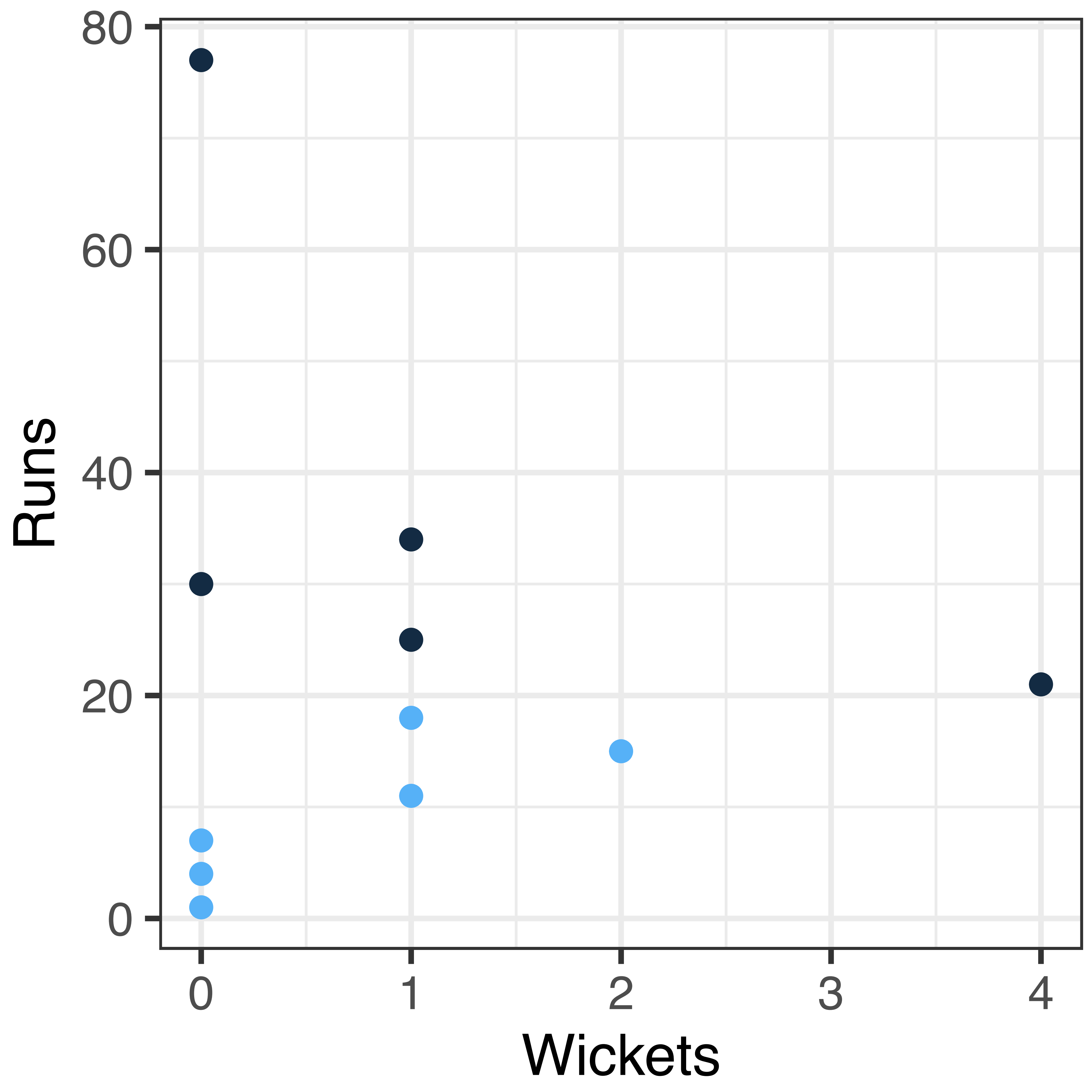
2023 Ashes, 2nd Test, Lord's



2. Training data

Divide our training data by innings. Let's train on the first innings and validate on the second.

2023 Ashes, 2nd Test, Lord's

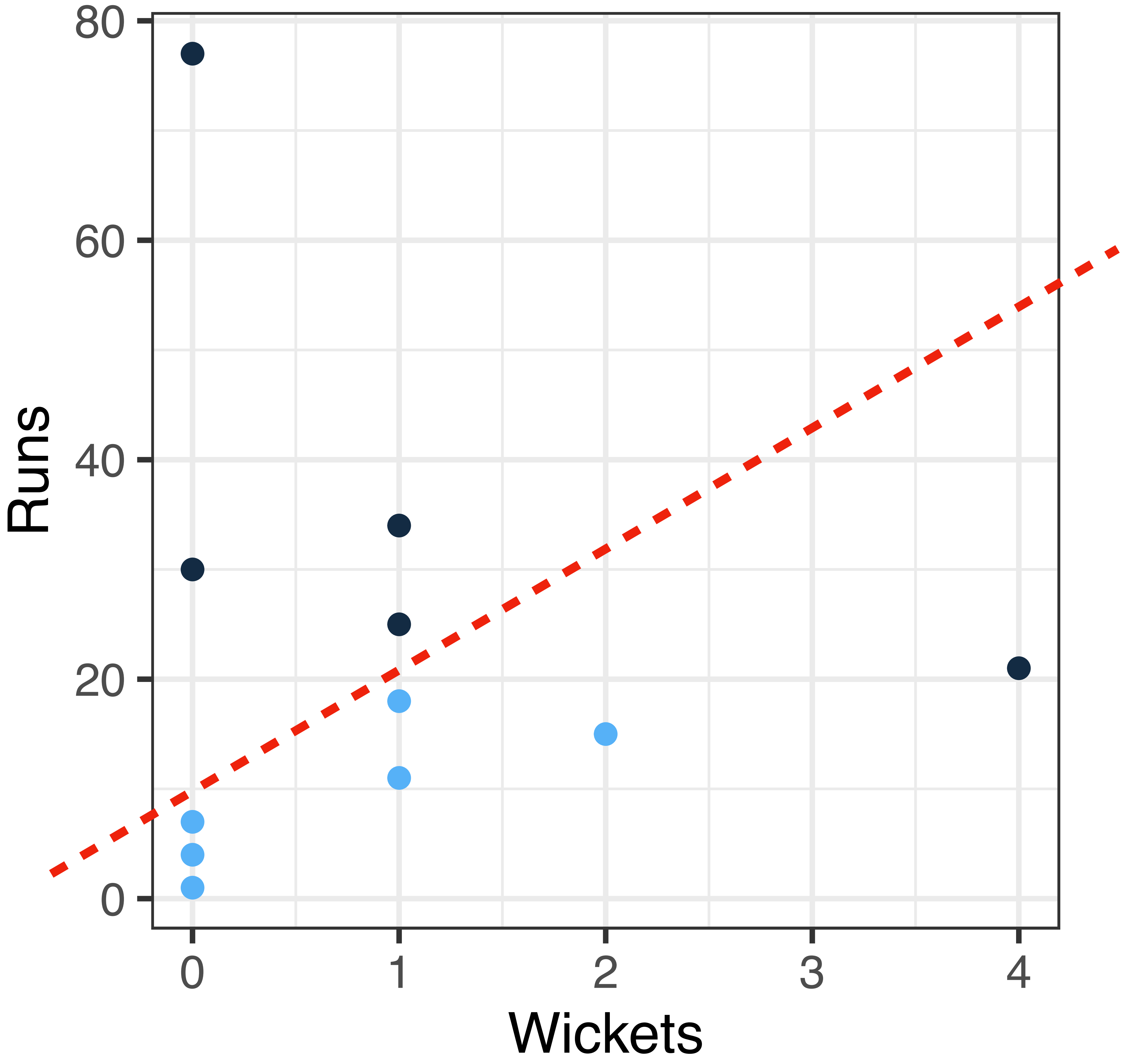


2. Training data

Divide our training data by innings. Let's train on the first innings and validate on the second.

Let's also assume our data are labelled.

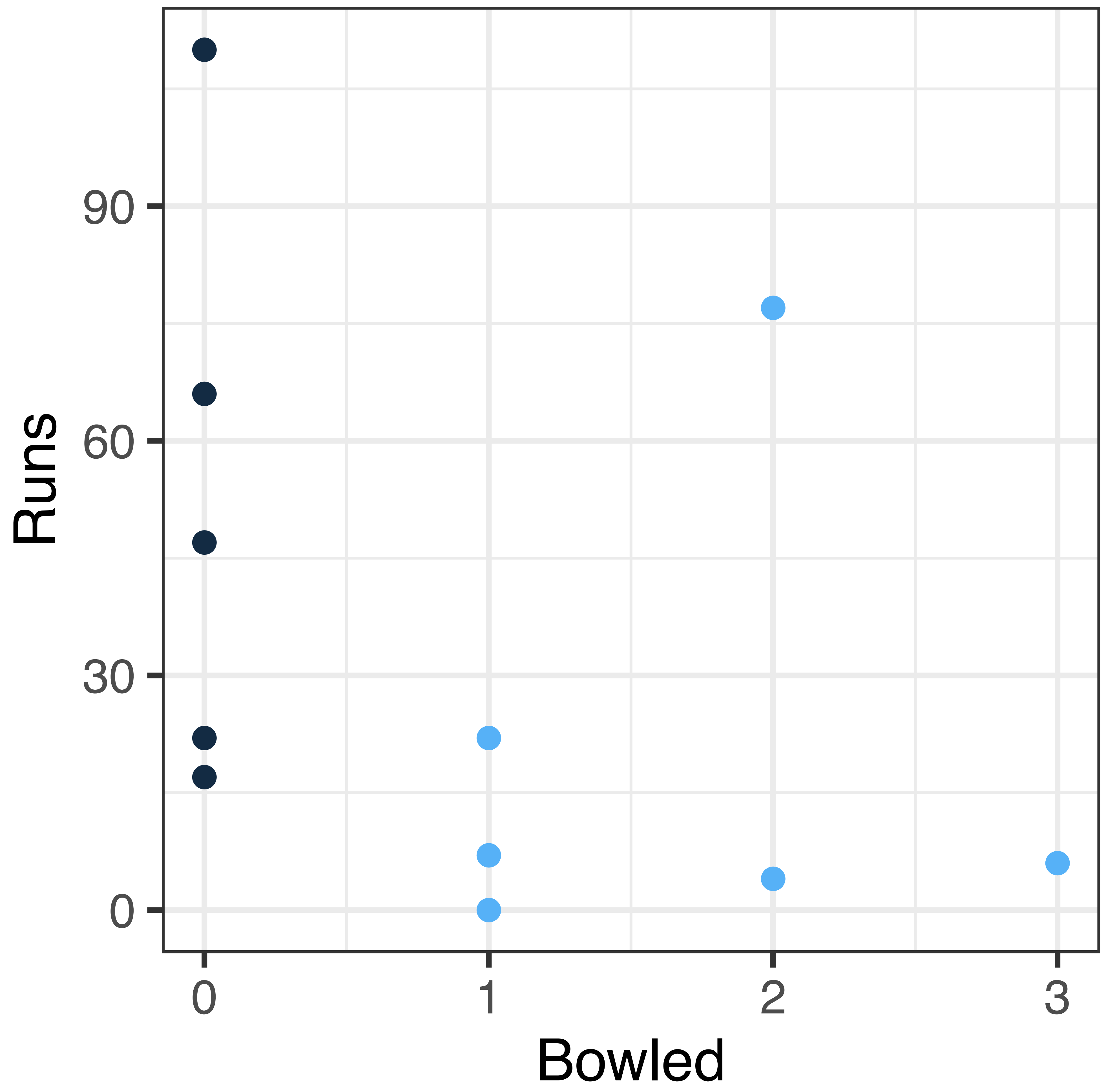
2023 Ashes, 2nd Test, Lord's



3. Features

Should we look at who the wickets are attributed to?

2023 Ashes, 2nd Test, Lord's

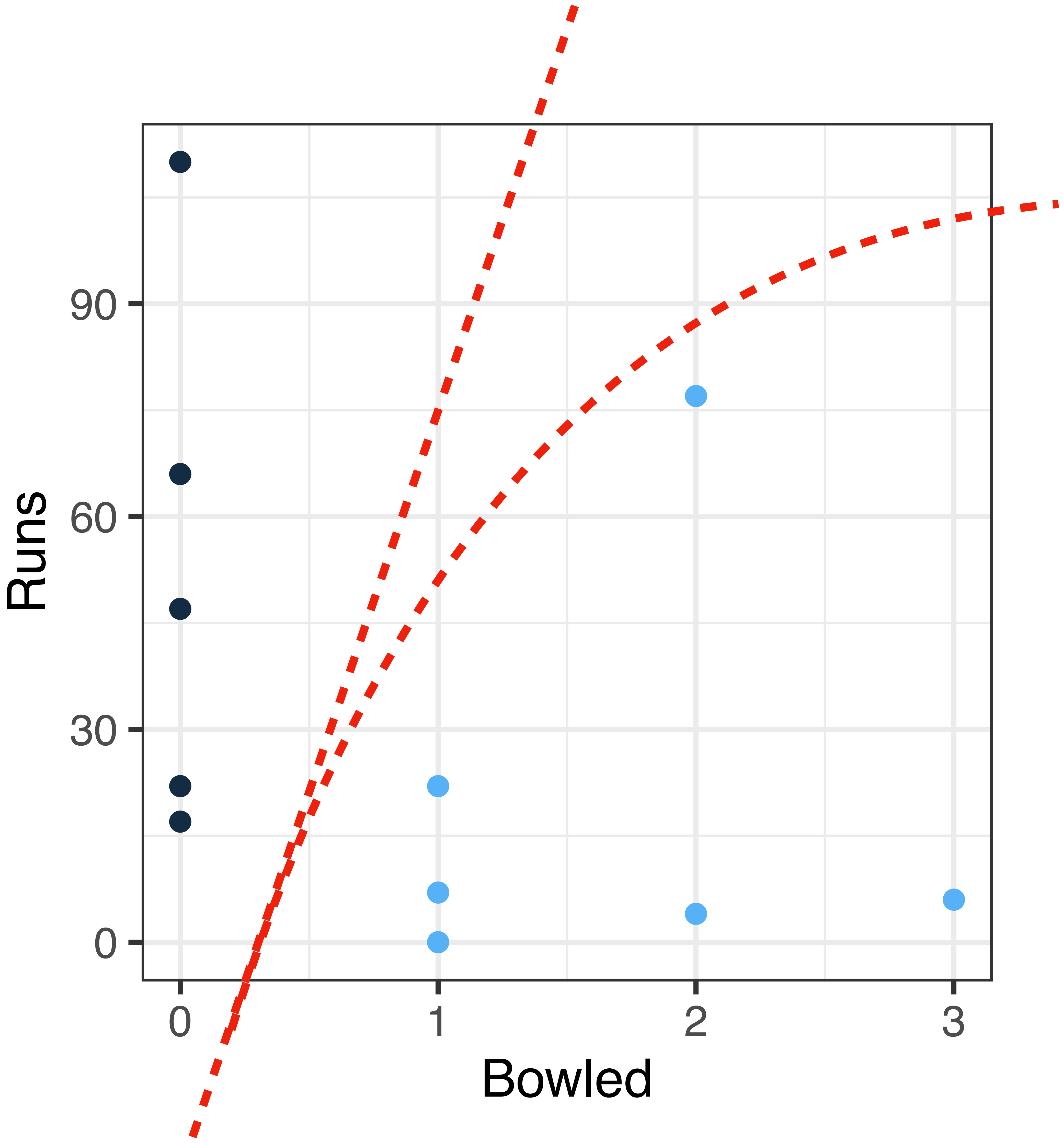


3. Features

Should we look at who the wickets are attributed to?

Or who bowled the ball?

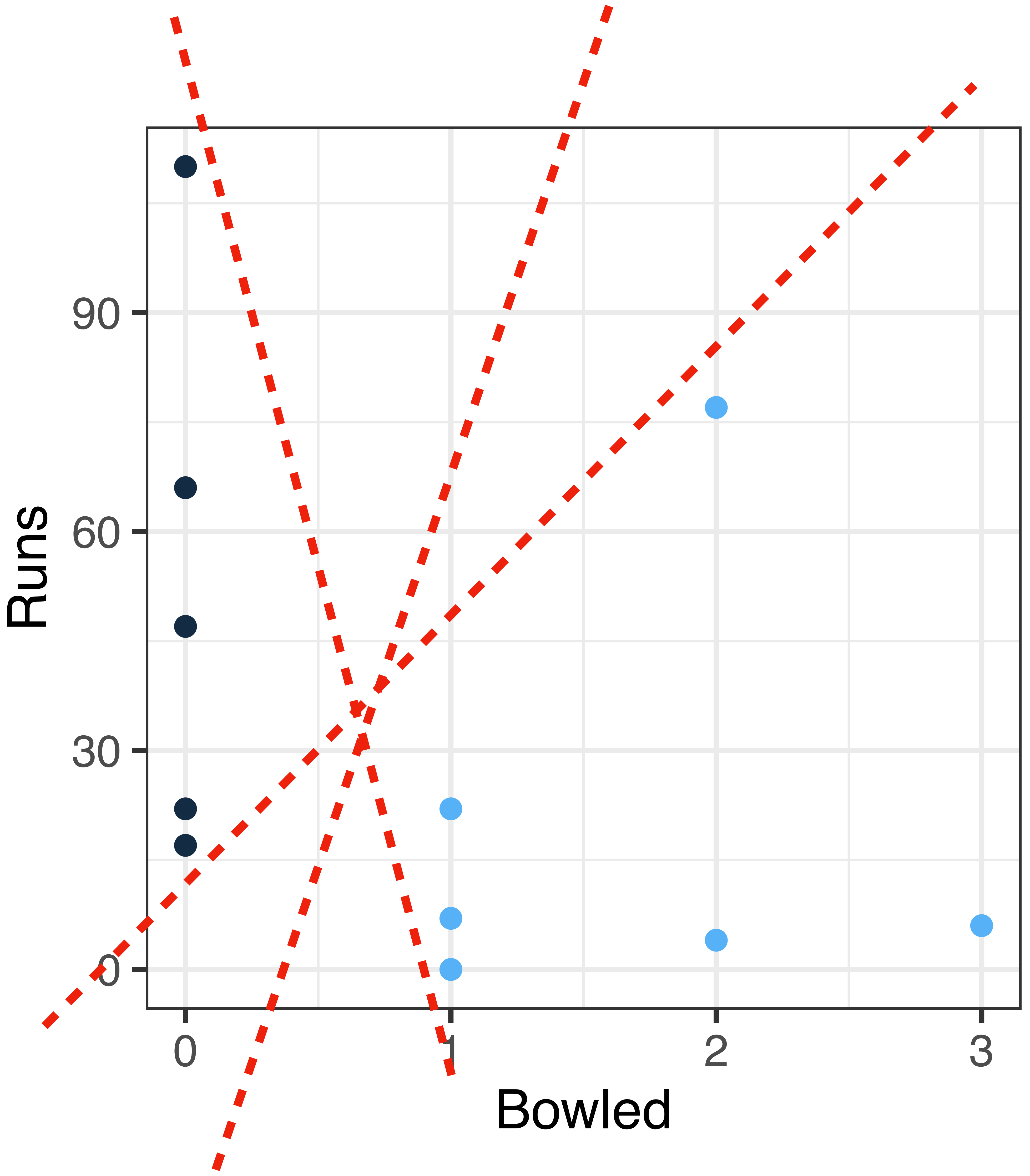
2023 Ashes, 2nd Test, Lord's



4. Model Architecture

Via what function should we classify these points?

2023 Ashes, 2nd Test, Lord's



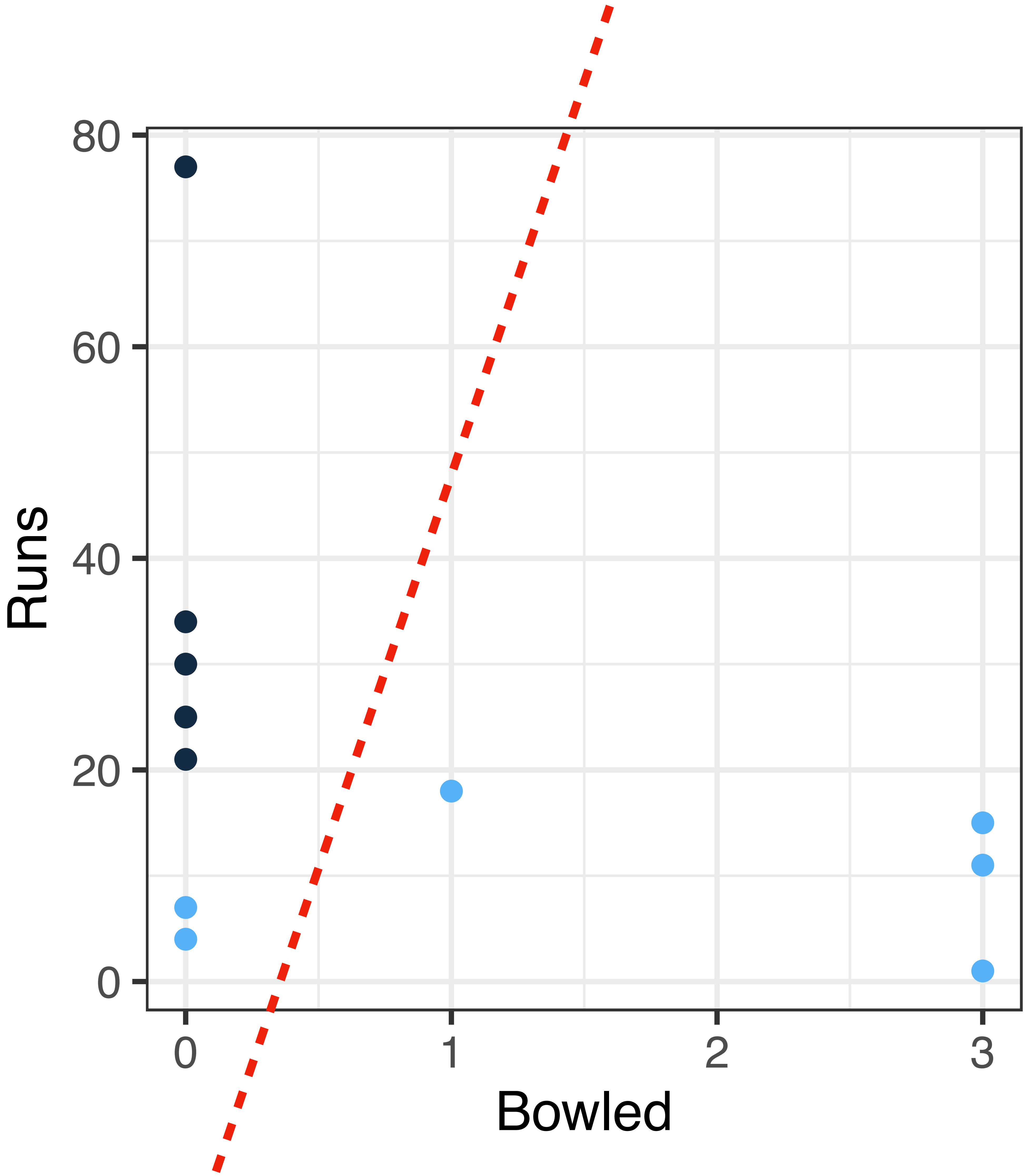
5. Inference/Optimisation

What is our loss function? What are we optimising?

$$\sum_i \mathbb{I}\{\text{label} \neq \text{modelled}\}$$

Do we need any regularisation?

2023 Ashes, 2nd Test, Lord's



6. Validation

Let's evaluate out model based on out of sample predictions.

What happened here?

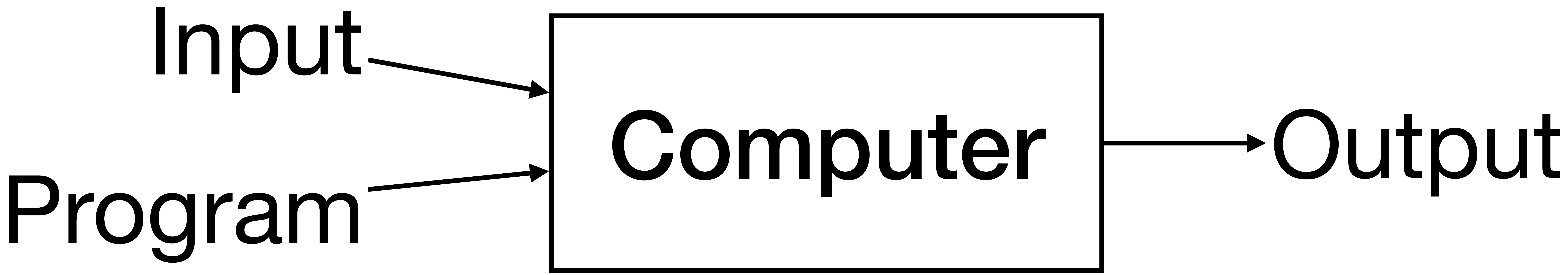
Postmortem: 2023 Ashes, 2nd Test, Lord's

- Are there only batsmen and bowlers? Were we correct in asserting k = 2?
- Is there labelling errors? What defines a bowler? Smith bowled for 1 over, is he a bowler?
- Are there any other features? What is the effect of including too many?
- How much data are enough data?

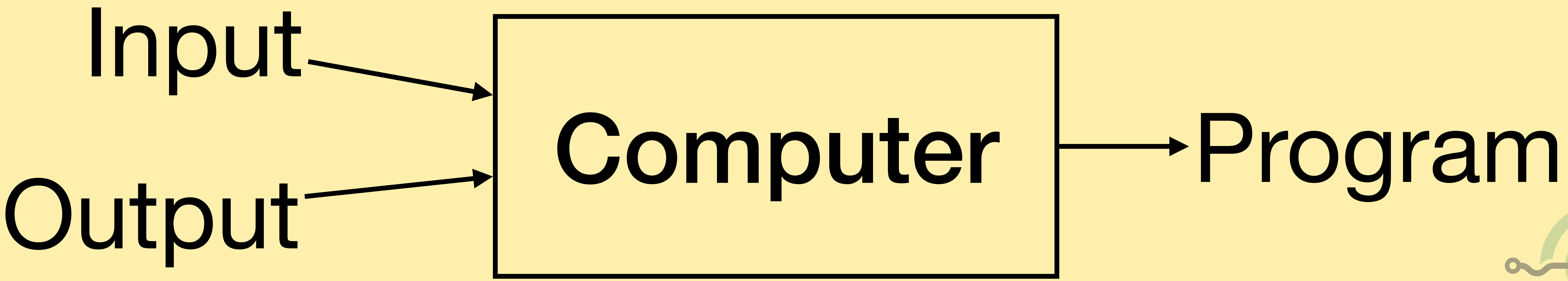
Types of Machine Learning

- **Supervised Learning**
  - Regression and classification
- **Unsupervised Learning**
  - Clustering, dimension reduction, feature analysis
- **Reinforcement Learning**
  - Robotics, game AI

Von Neumann Model



Machine Learning Model



Supervised Learning

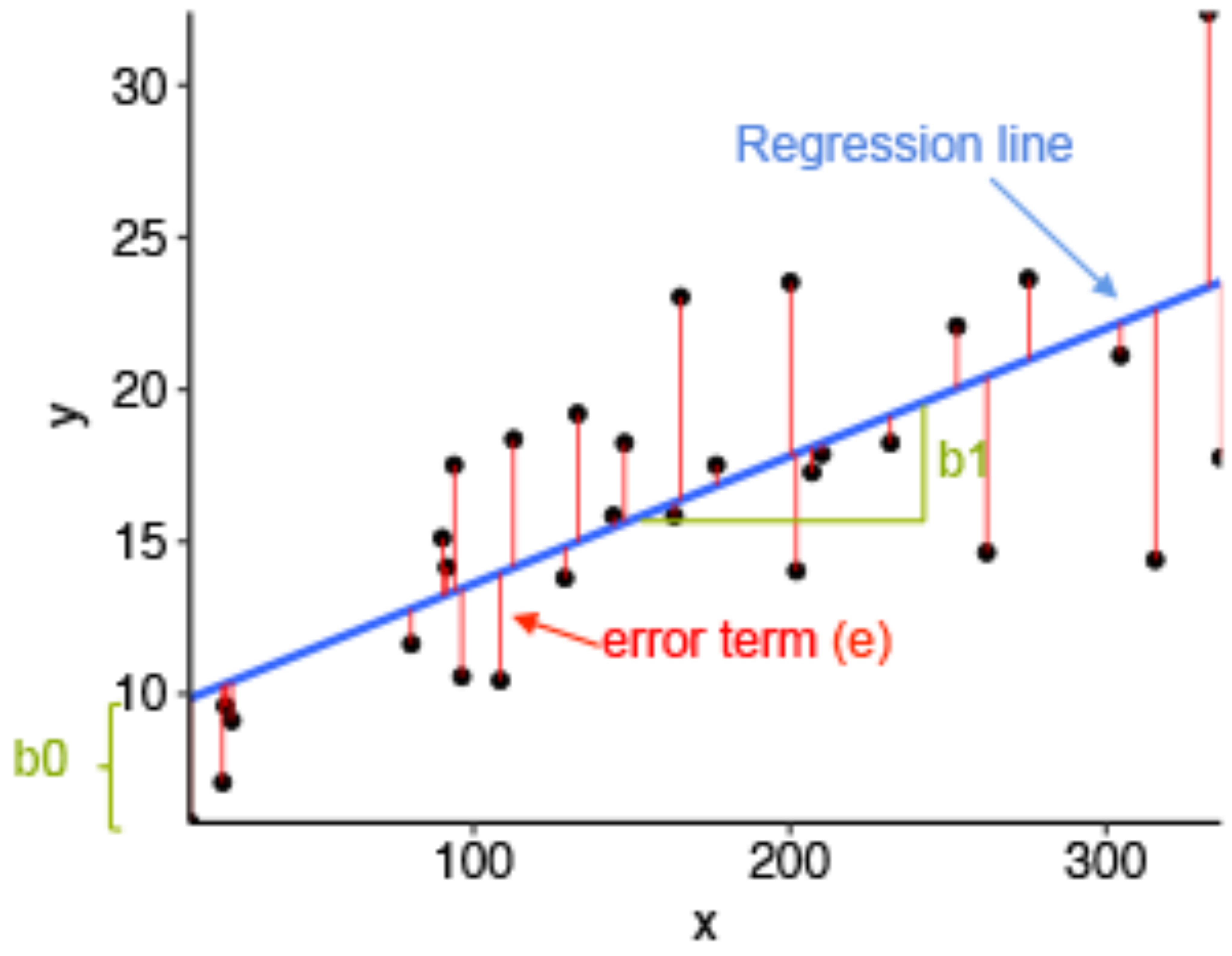
- A type of machine learning where the model learns from **labelled** data
- The model is provided with input-output pairs (X, Y) where
    - X are the input features
    - Y is the target labels
- **Common supervised learning tasks:**
    - **Classification:** predict a discrete label (email spam, medical diagnosis)
    - **Regression:** predict a continuous label (housing prices, remaining useful life)
- **Training phase:** The model is trained so as to minimise some distance/loss of the model predictions from the observed Y

Linear Regression

- Fits a linear relationship between the input features X and a continuous output variable Y

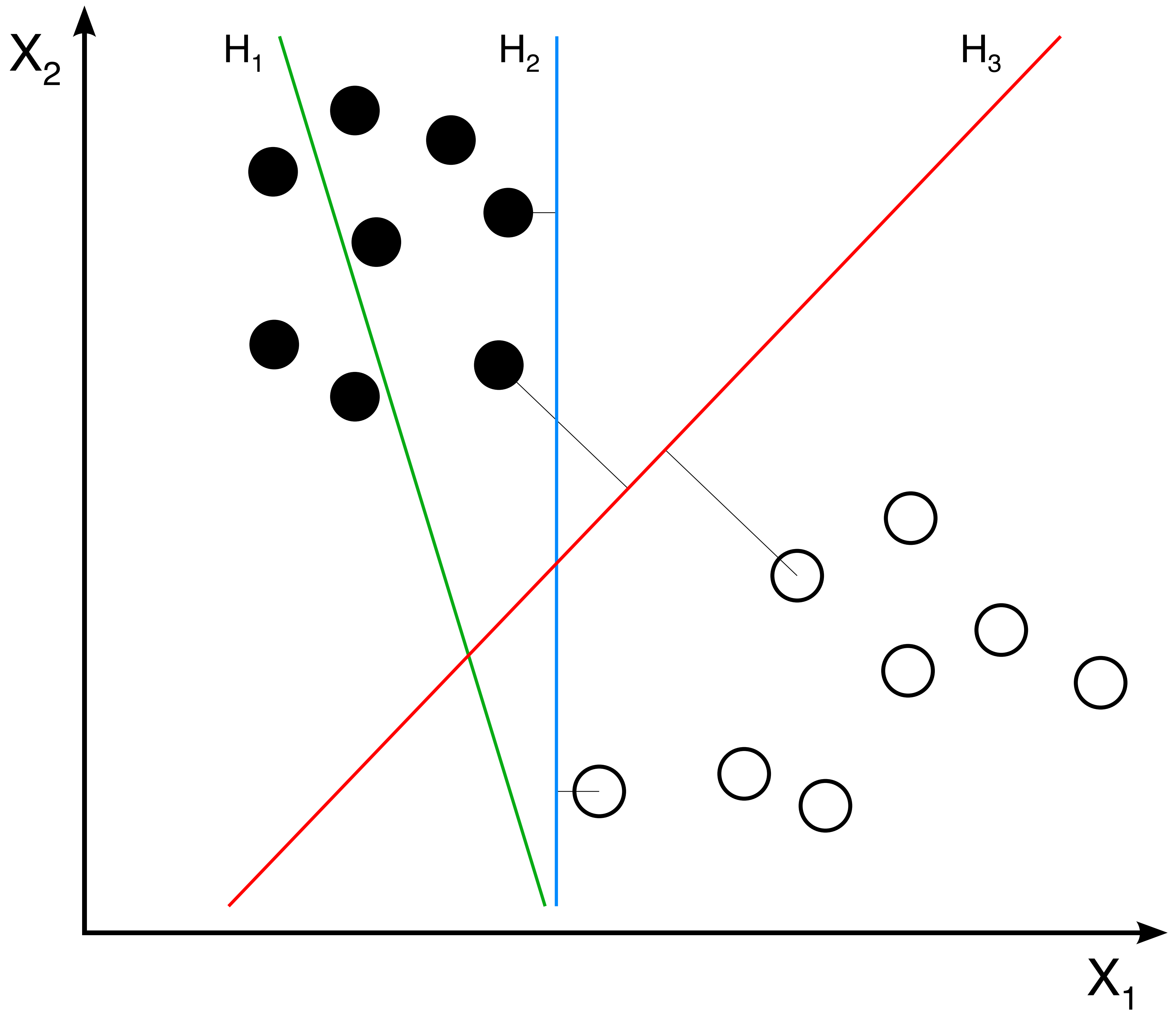
$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m + \epsilon$$

- **Goal:** the β_i are trained to minimise some loss on ϵ
- **Advantage:** simple, interpretable, easy to fit
- **Disadvantage:** isn't necessarily a good model (too simple)

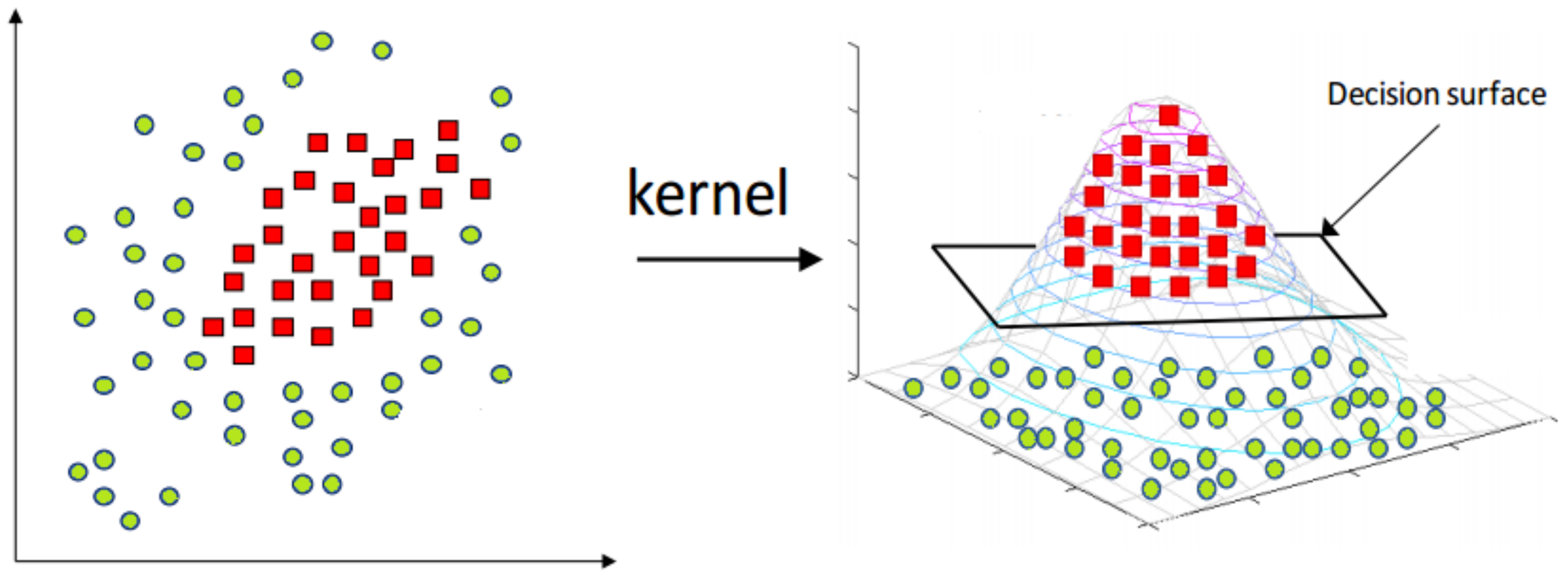


Support Vector Machine

- Finds the best **hyperplane** that maximises the margin between different classes in the data
- **Goal:** maximise the margin, defined at the distance between the support vectors and the hyperplane
- For **non-linear** data, there are some nice tricks to project the data into a higher dimensional feature space
- **Advantage:** can capture non-linear, works with smaller data
- **Disadvantage:** can be sensitive to parameterisation, needs well defined margin



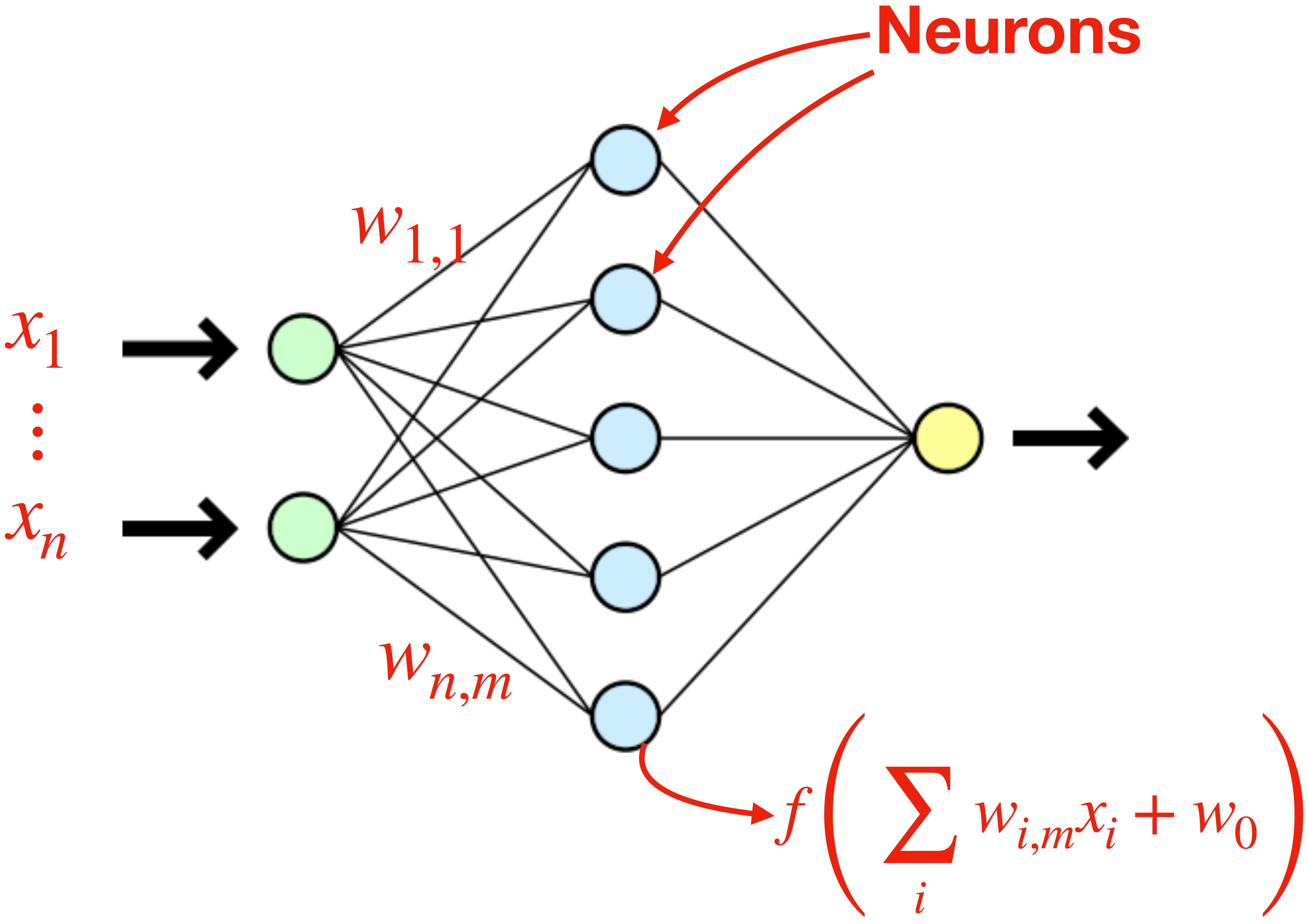
Support Vector Machine



Neural Networks

- Inspired by the human brain for both classification and regression
- Composed of layers of interconnected nodes that transform input data to meaningful outputs
- **Components:** input layer (raw features), hidden layer (transformations), output layer (final predictions)
- Each **neuron** in the hidden layer applies a weighted sum of inputs, passes through an **activation function**, and sends that output to the next layer
- The network learns by adjusting the weights through **backpropagation**, which minimises the prediction error
- **Advantages:** can model highly non-linear/complex relationships
- **Disadvantages:** requires lots of data and computational power

Neural Networks

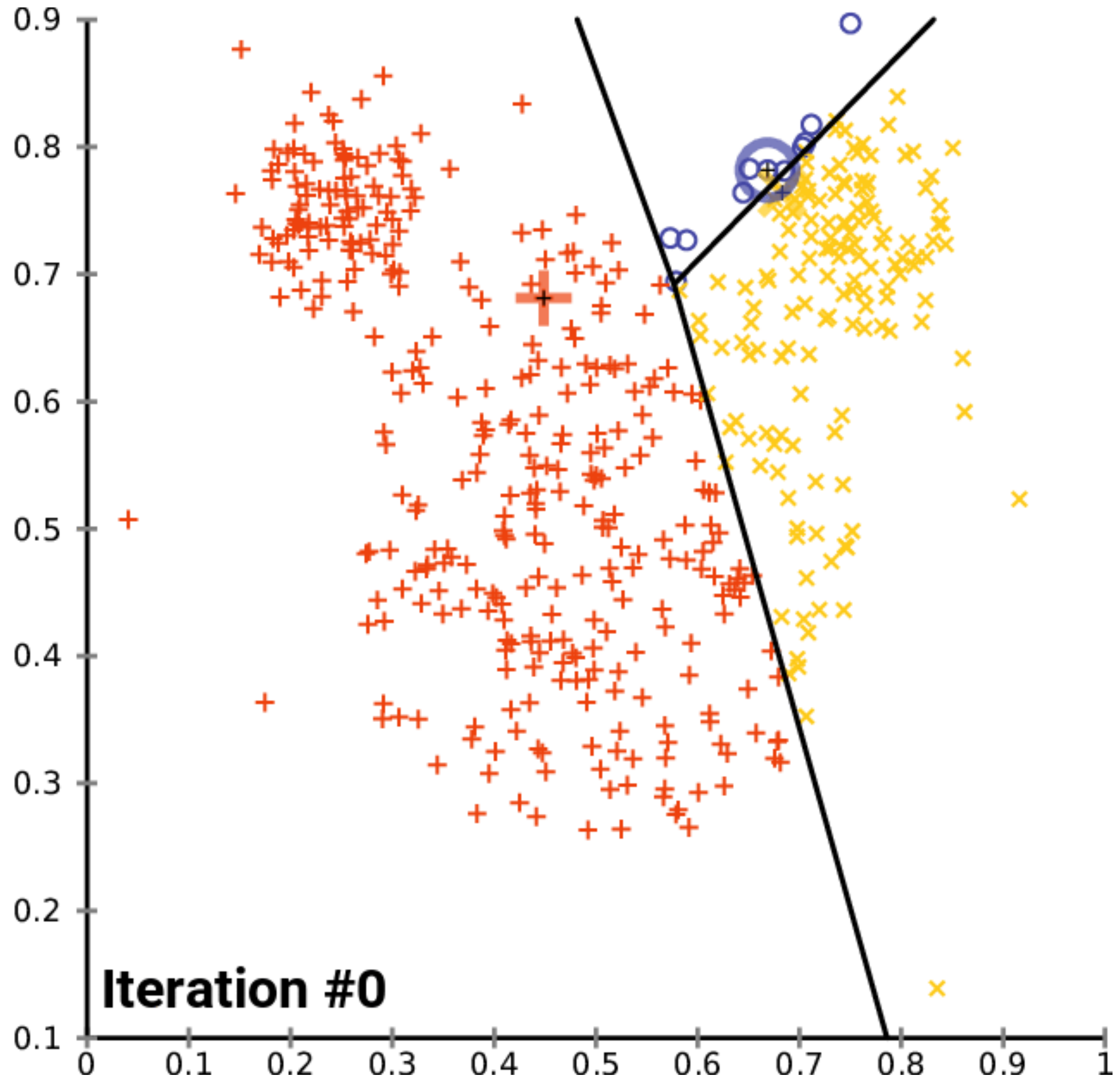


Unsupervised Learning

- A type of machine learning where the model learns from **unlabelled** data
- The goal is to find patterns, groupings or structures in the data without predefined output labels
- **Common supervised learning tasks:**
    - **Clustering:** grouping similar data together (anomaly detection, market segmentation)
    - **Dimension reduction:** reduce the number of features whilst preserving important information (compression, trend analysis, customer preference)
- Common algorithms include k-means clustering, hierarchical clustering, principal component analysis, autoencoders

K-means clustering

- Divide data into K distinct clusters. Each data point is assigned to the cluster with the nearest centroid.
- Typically uses Euclidean distance as the measure.
- **Advantages**: easy, works for large data, works well when separation is large
- **Disadvantages:** requires specification of K, struggles with irregularly shaped clusters

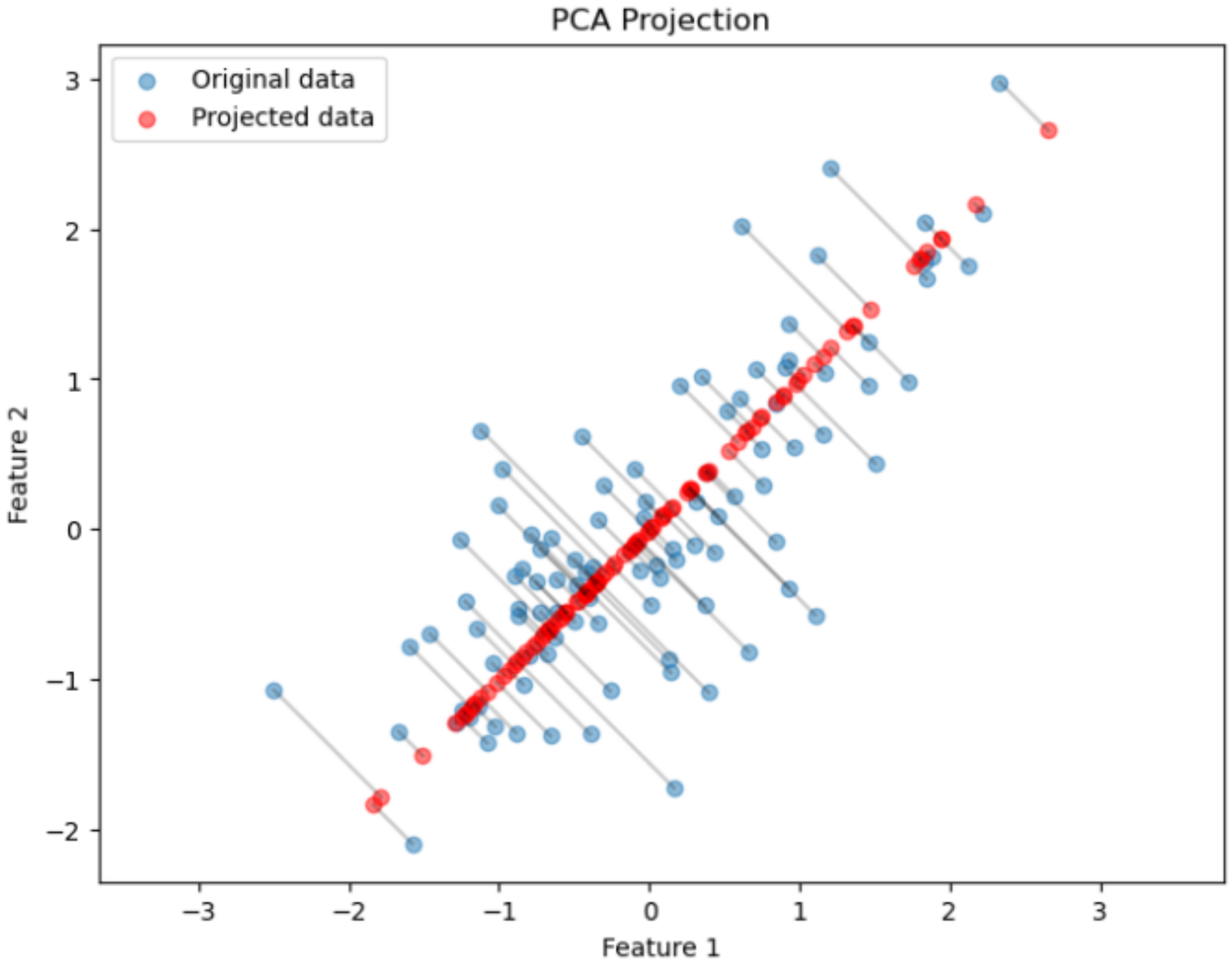


Principal Component Analysis

- Transforms high dimensional data into fewer dimensions whilst preserving as much variability as possible

- PCA finds new features called **principal components** which are linear combinations of the new features

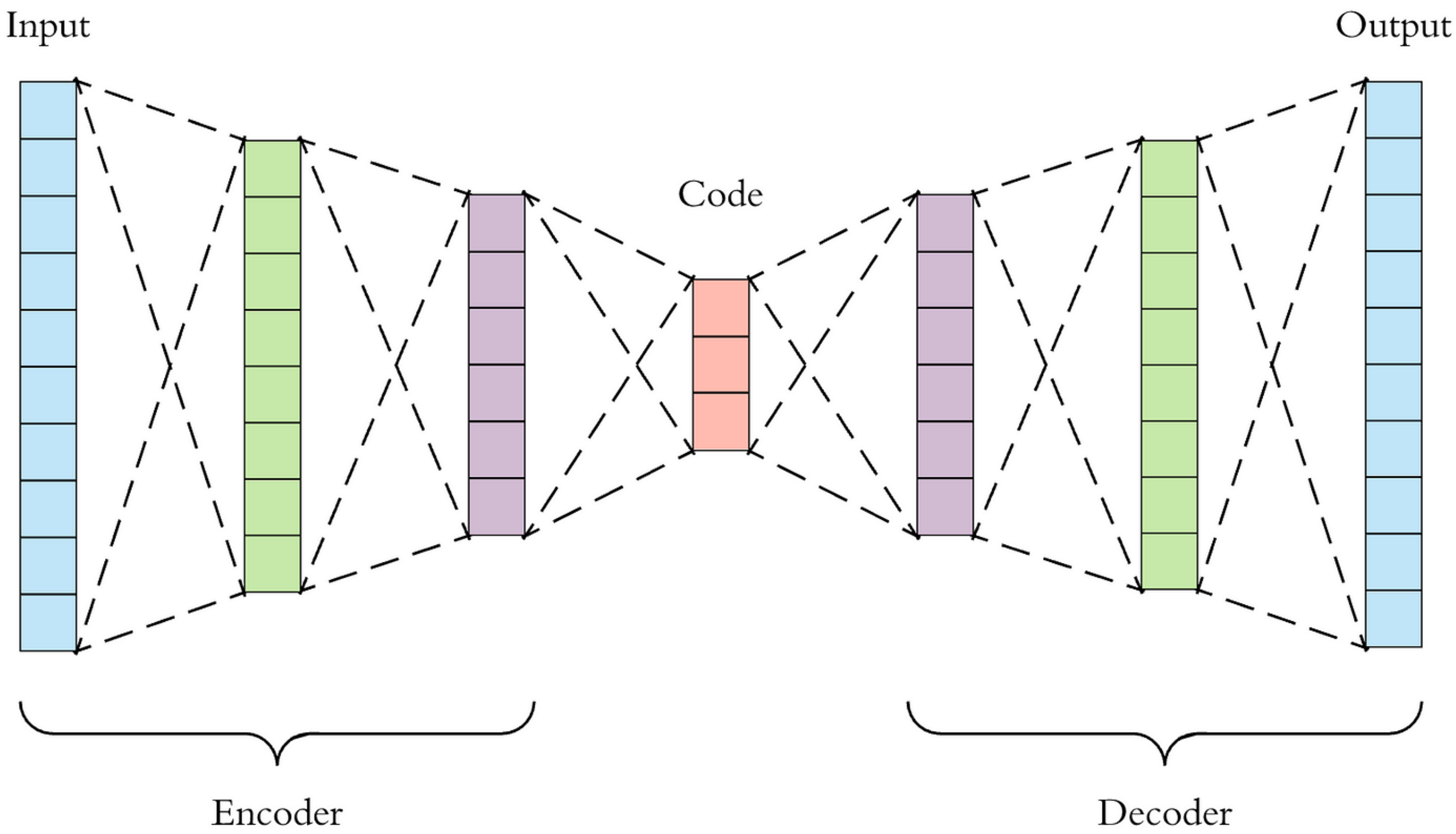
- The **principal components** are orthogonal and capture the directions of maximum variance



Autoencoders

- A type of neural network for unsupervised learning and dimensionality reduction
- Compresses the inputs and reconstructs them as accurately as possible
  - **Encoder:** compresses the input into a smaller representation
  - **Latent space:** compressed version of the input
  - **Decoder:** Reconstructs the input from the latent space
- **Advantages:** good for complex high-dimensional data, can capture nonlinearities
- **Disadvantages:** needs lots of data, careful tuning, can overfit

Autoencoders



Reinforcement Learning

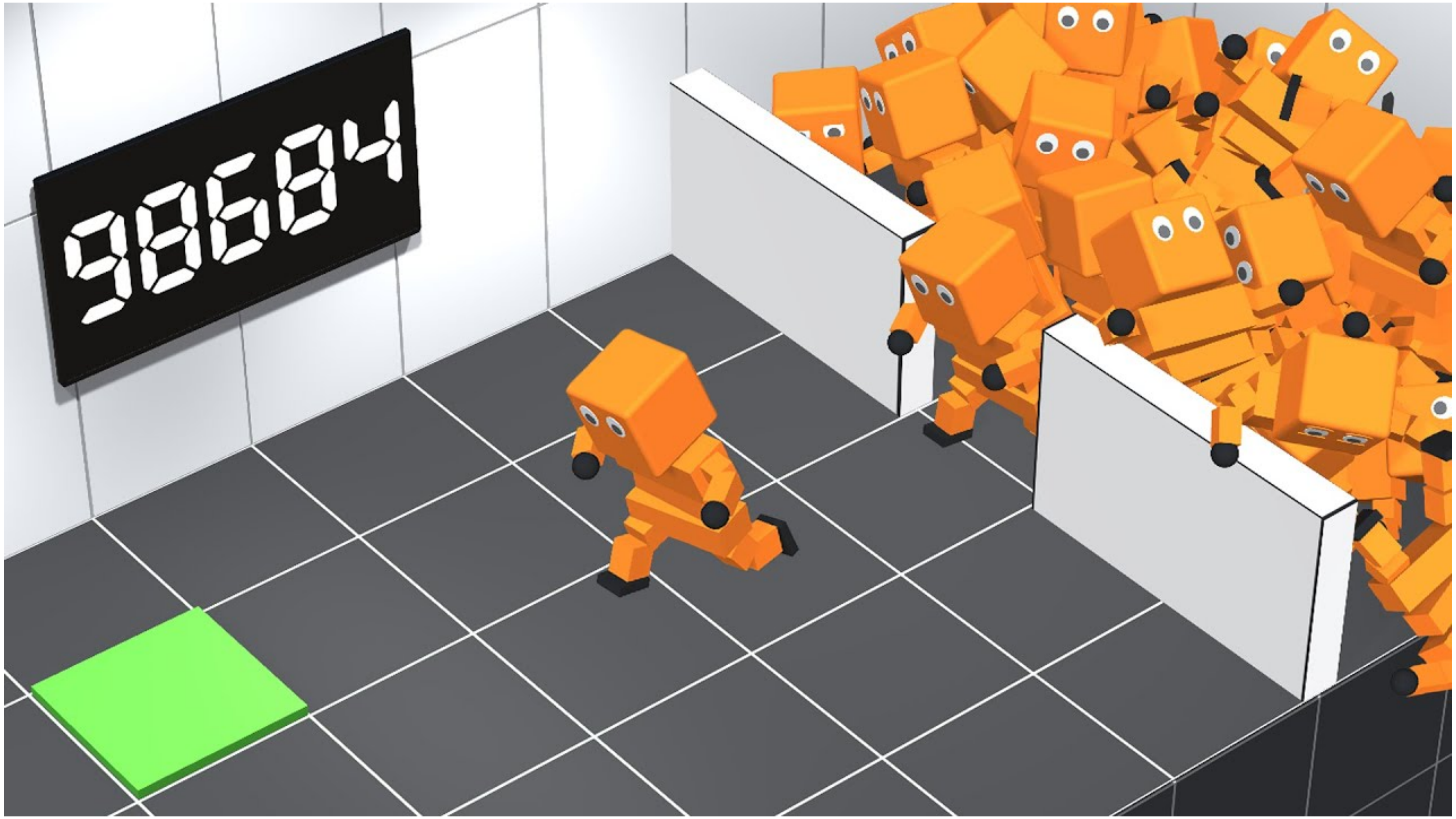
- A type of machine learning where an agent learns to make decisions by interacting with an environment so as to maximise some reward

  1. The agent takes an action in a given state
  2. The environment responds with a new state and reward
  3. The agent updates its strategy based on the reward, aiming to maximise

- **Advantages:** suitable for problems where correct action is not obvious, and rewards are complex and delayed

- **Disadvantages:** requires large amount of training data, must be able to simulate environment, can be very difficult to tune

Reinforcement Learning



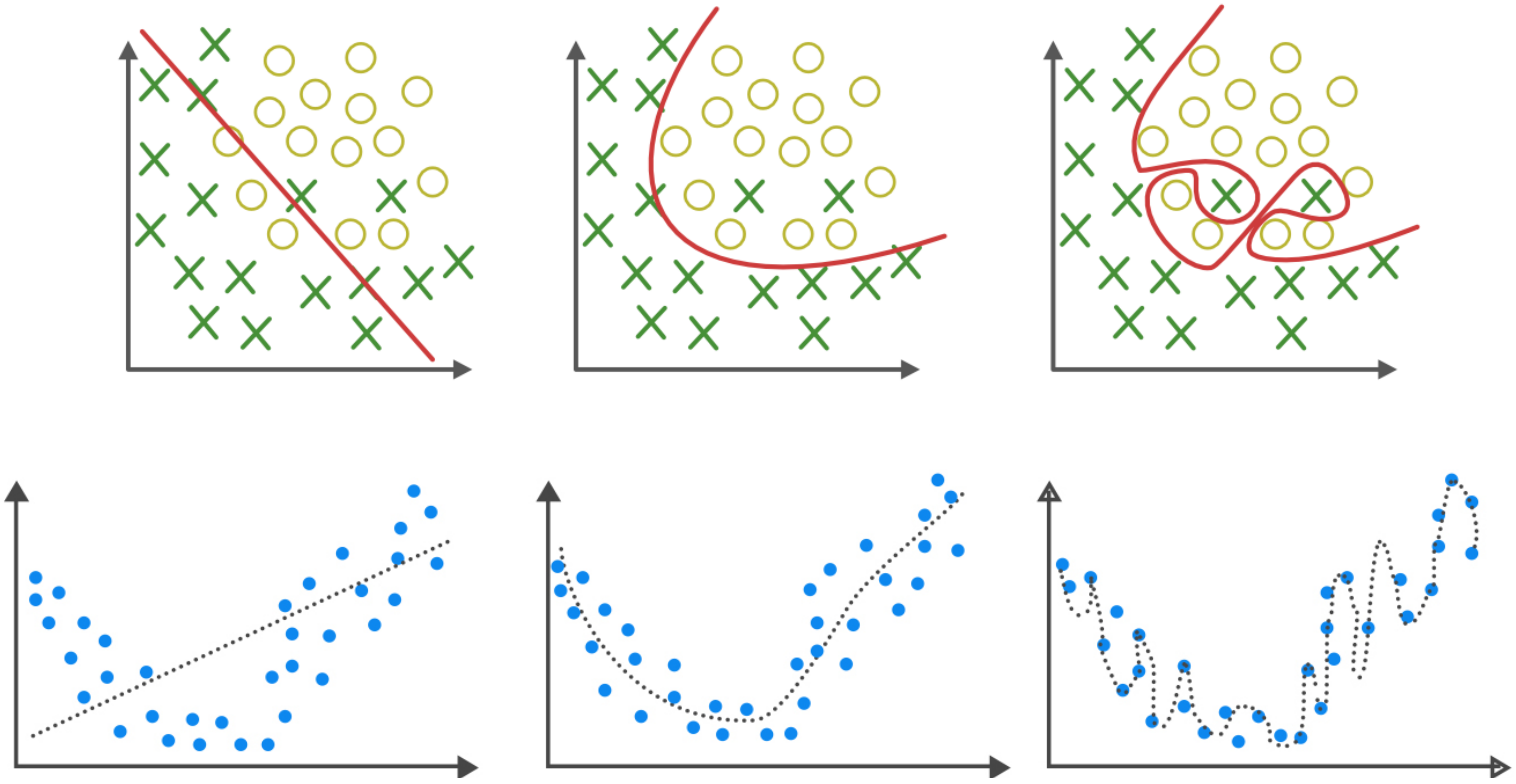
The Machine Learning Workflow

- **Data collection and pre-processing**
    - Cleaning, normalisation, missing values
- **Model training and hyper parameter selection**
    - Model selection, algorithm selection, computational resources
- **Model evaluation and cross-validation**
    - Accuracy, precision, recall

Overfitting and Underfitting

- **Overfitting**: the model learns not only the underlying patterns, but also the noise in the training data
    - High performance on training data, low performance on validation data
    - Model is too complex
- **Underfitting:** The model is too simplistic and fails to capture the underlying patterns in the data
    - Poor performances on both training and validation data
    - Model is too simple
- **Remedies:** Regularisation, cross-validation, more data, feature engineering

Overfitting and Underfitting



Real-world Use Cases

- Smart-phone face recognition, voice recognition, speech-to-text
- Netflix movie recommendations
- Google advertisement recommendations and basket analysis
- Predictive analytics for healthcare
- Financial fraud detection
- Self-driving cars
- Natural language processing, e.g. ChatGPT

Challenges and Future Trends

- **Model interpretability**
  - ML models are commonly considered ‘black-boxes’
  - Difficult to communicate the why; e.g. in engineering, healthcare, finance
- **Computational resources**
  - More complicated models require non-linear increase in compute time
  - Environmental concerns, inequality barrier
- **Ethics and bias**
  - Models can reflect biases in the training data leading to unintentional consequences in decision making (e.g. racial profiling)

Where and how to learn more

- *An Introduction to Statistical Learning.* James, Whitten, Hastie, Tibshirani
- *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* Géron
- Learn a language other than MATLAB (probably python, maybe Julia or R)
- Andrew Ng's Coursera modules: *Machine Learning* and *Deep Learning*
- Youtube: 3Blue1Brown, StatQuest, Corey Schafer, MIT, sentdex
- **Mathematics, statistics, machine learning and coding are best learnt by doing and not by passive learning**