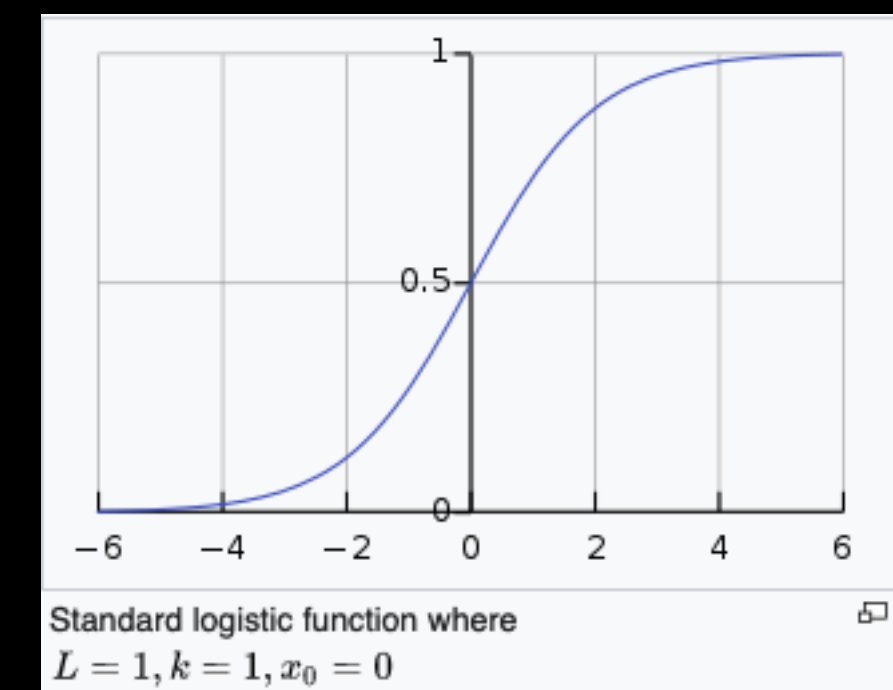
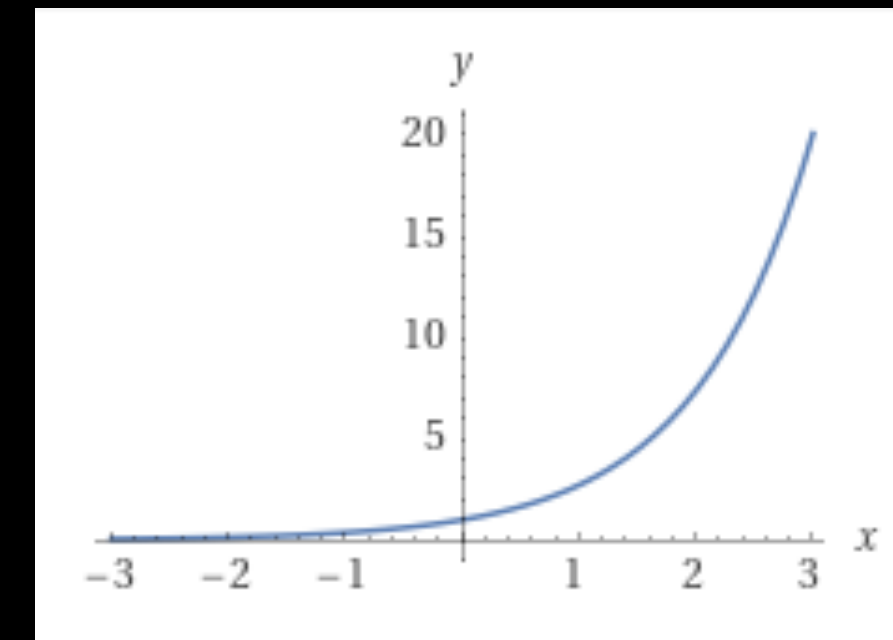
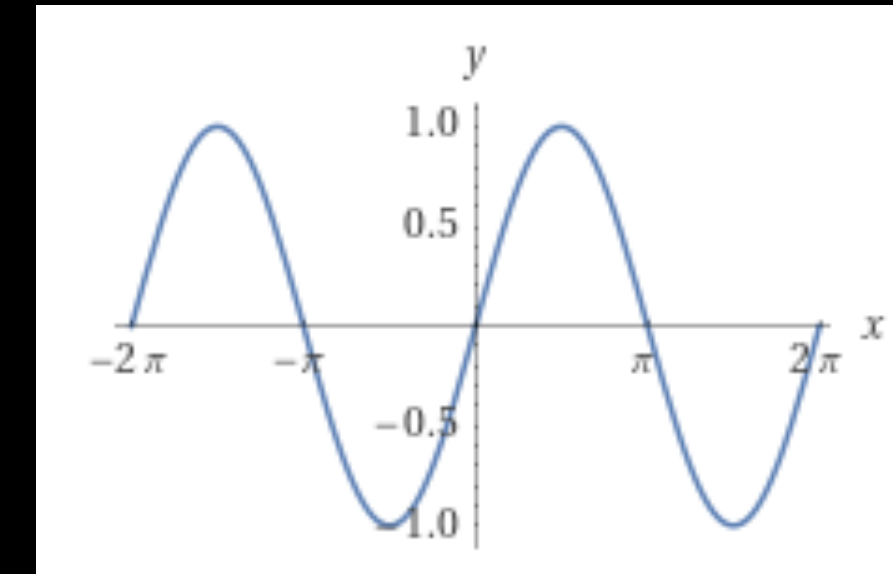


# Modelling Relations

Raoul Grouls, 19-3-2024

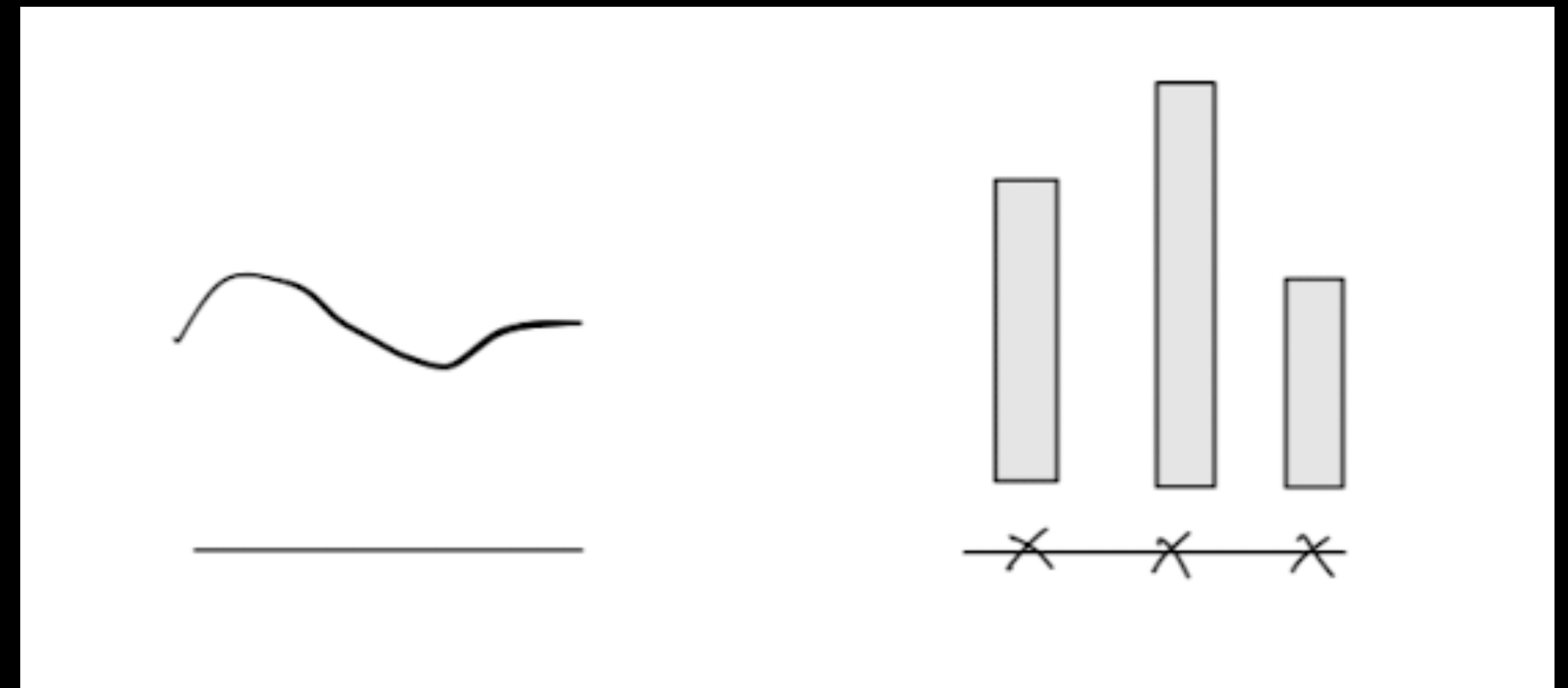
# Recap: The four horsemen of modelling

- Linear
- Sine
- Logistic
- Exponential



# Recap: distributions

- $f(x) = P(X = x)$
- Satisfies:
  - $f(x) \geq 0, \forall x \in X$
  - $\sum_{x \in X} f(x) = 1$
  - $P(\mathcal{A} \subseteq X) = \sum_{x_i \in \mathcal{A}} f(x_i)$









# All machine learning

The summary

$$f: X \rightarrow y$$

# The map is not the territory

- Models are simplifications
- Models are built on assumptions
- The map might be correct, but the territory can change
- The simplicity of the models shows what is otherwise difficult to see



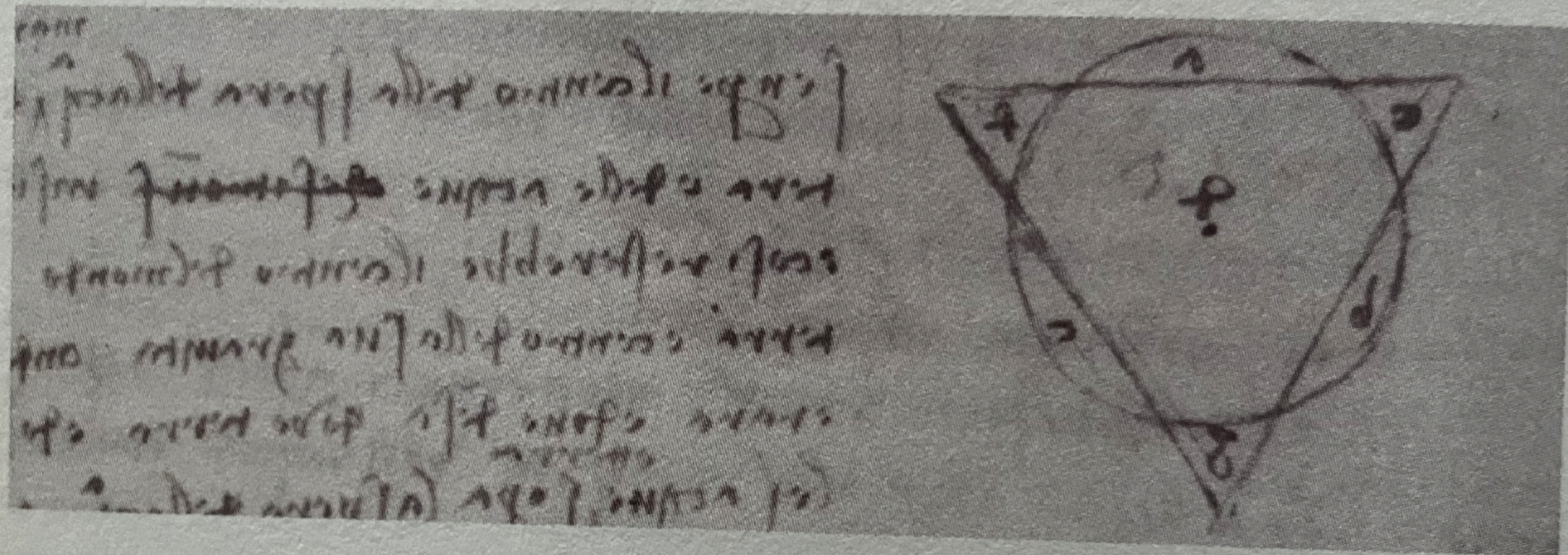
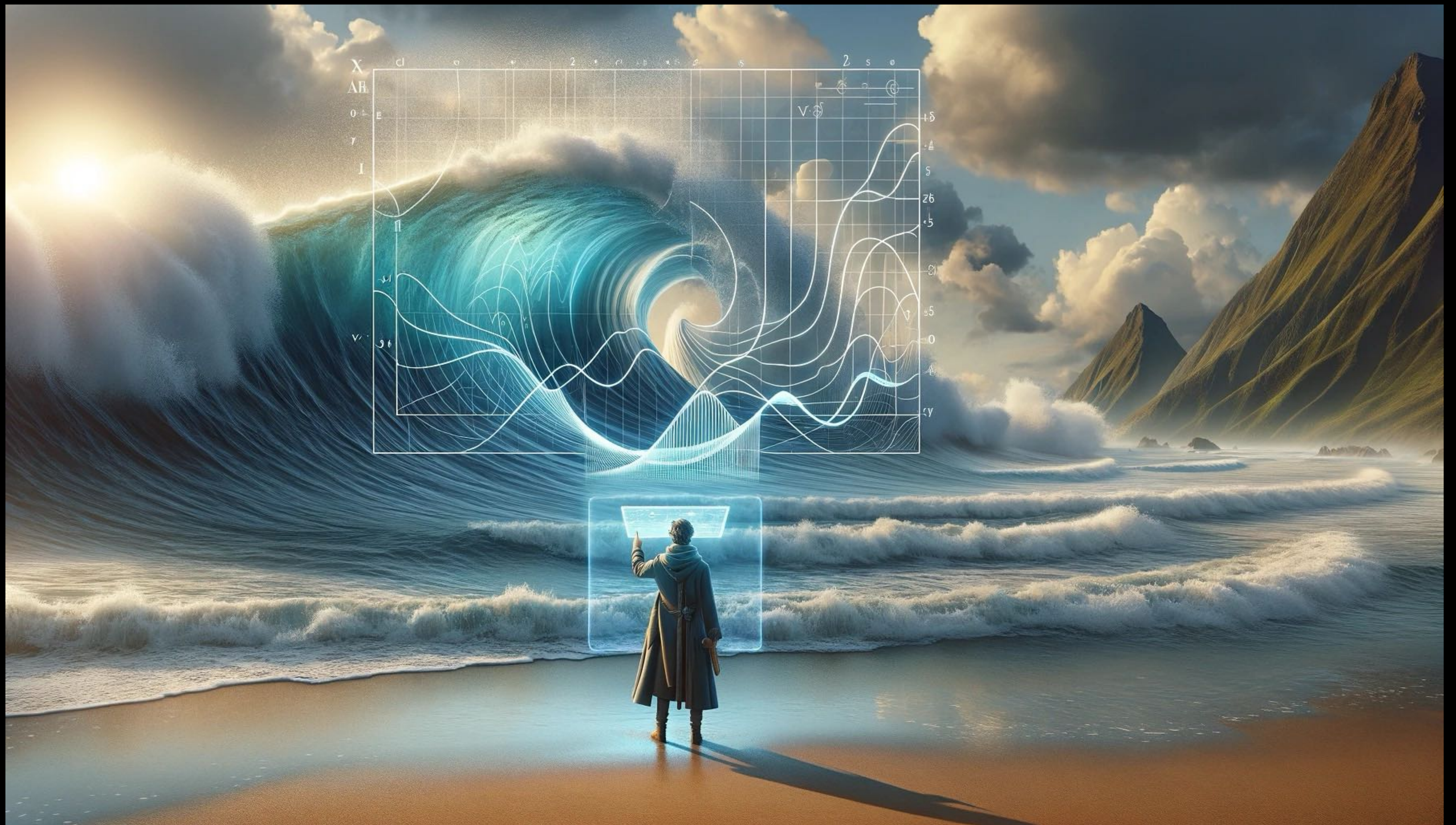


FIG. 2-7. Geometric model of the Earth.  
Codex Leicester, folio 35v (detail).



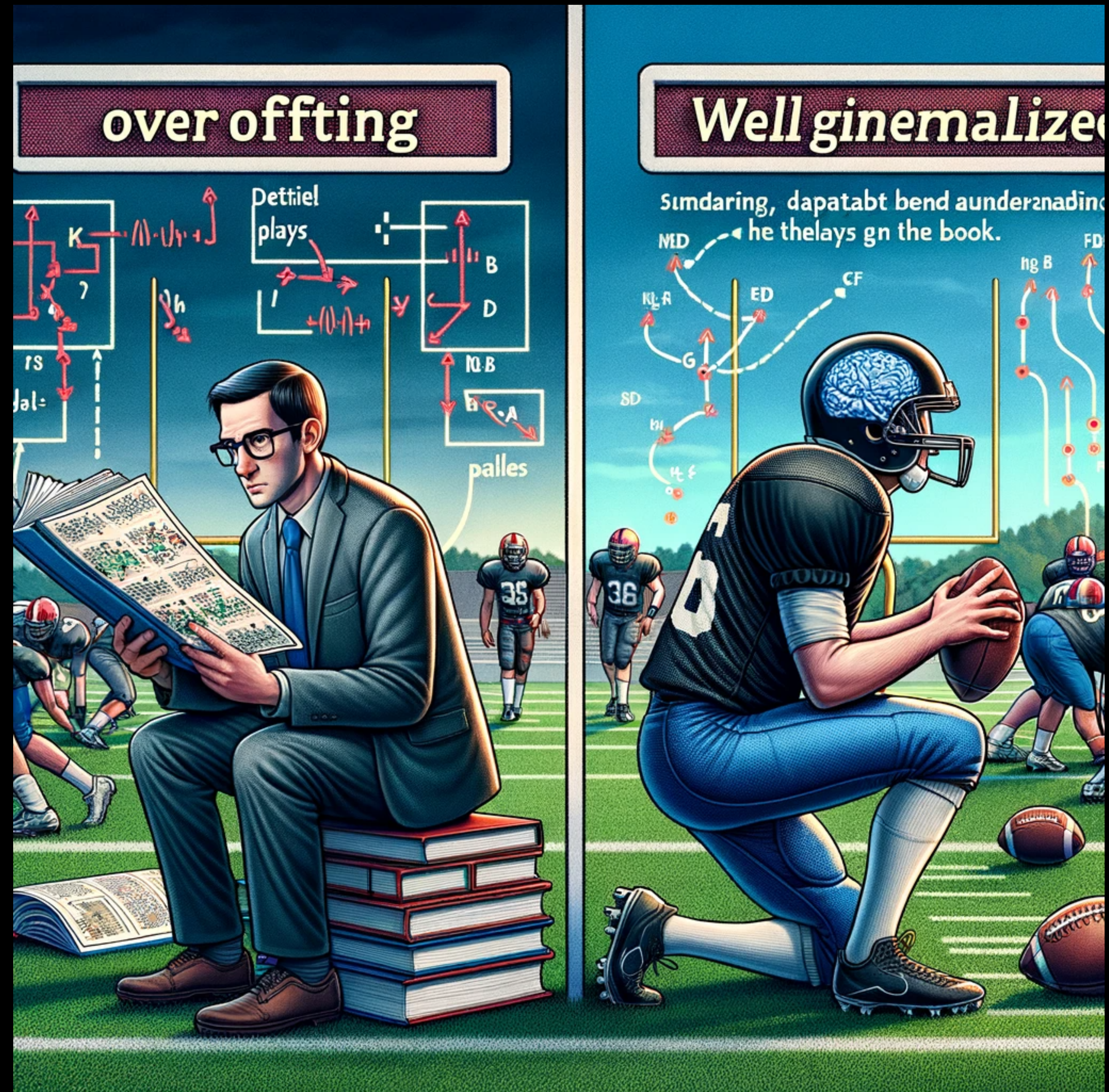




# Overfitting

A model that corresponds too closely to a particular set of data, therefore failing to fit to unseen data.

- the error / performance metrics for the training data are very good, but bad on unseen data.
- Overfit models tend to have higher complexity (number of learnable parameters) than is warranted by the data complexity
- Small changes in the data can lead to large changes in the model.





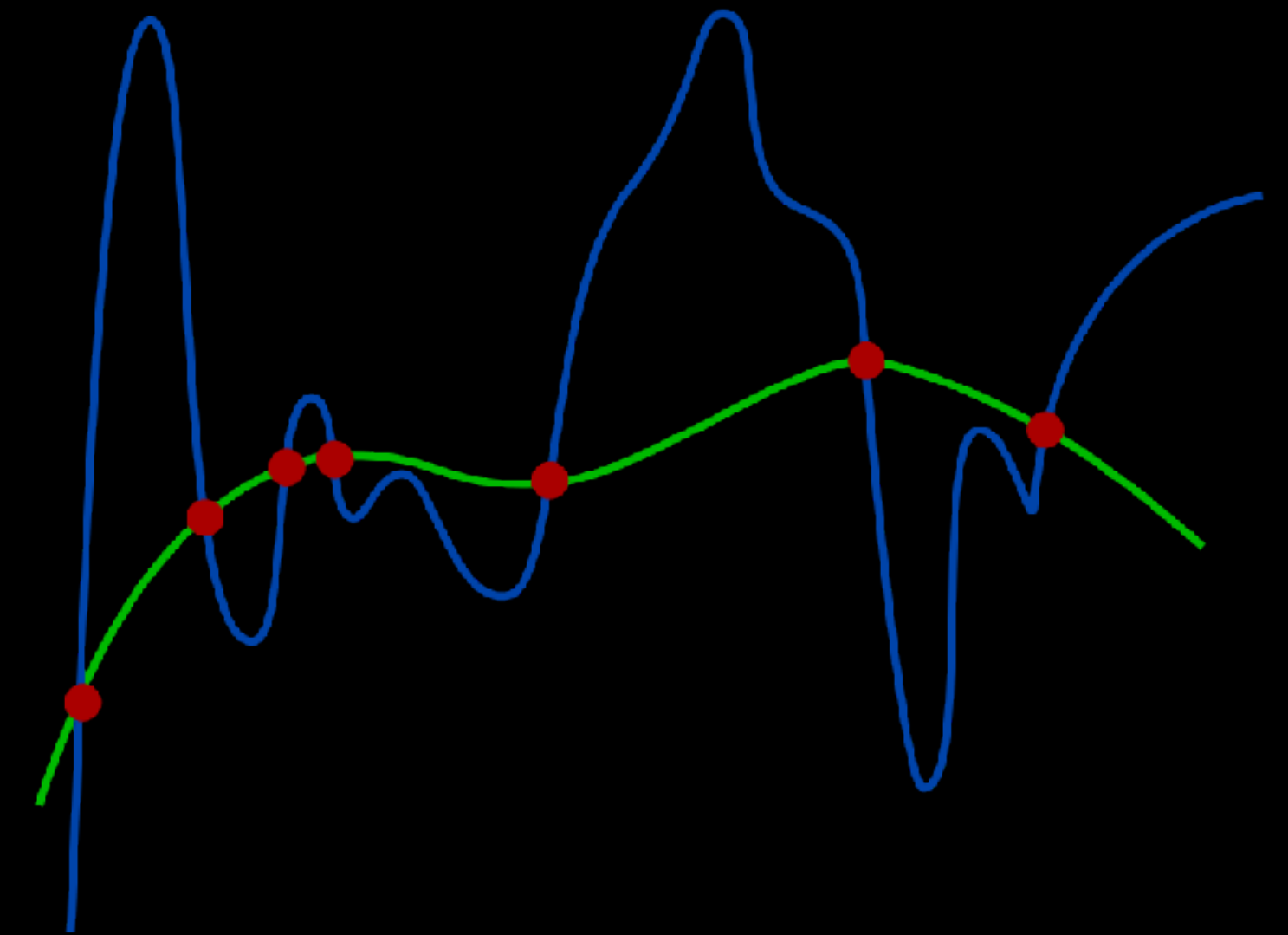
# Regularization

$$J(y, \hat{y} | W) = Loss(y, \hat{y}) + \eta \cdot Penalty(W)$$

A process used to enforce simpler models to prevent overfitting.

## Explicit Regularization:

- Involves directly adding a term to the optimization problem.
- These terms can include priors, penalties, or constraints.
- Imposes a cost on the optimization function.









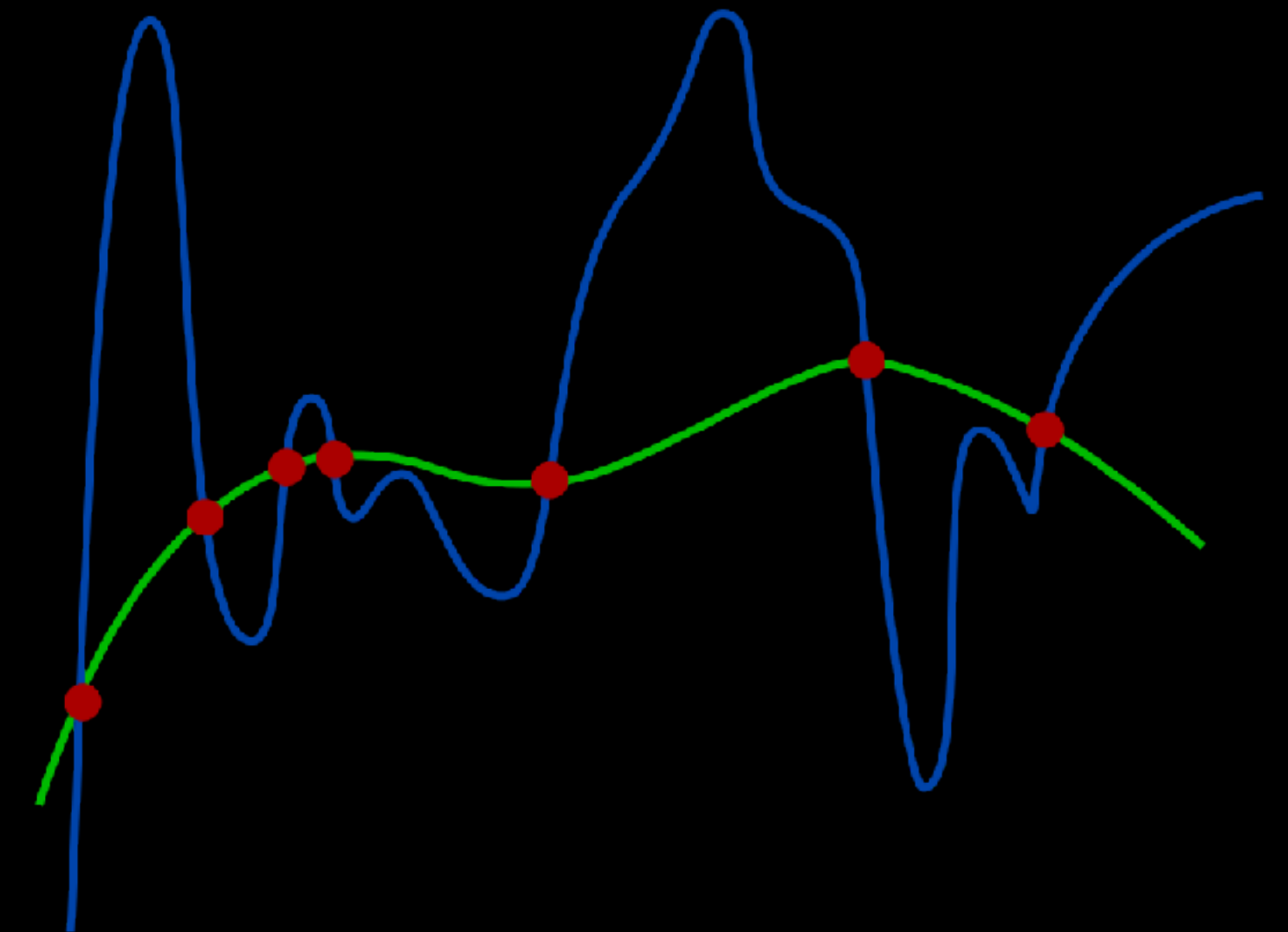
# Regularization

$$J(y, \hat{y} | W) = Loss(y, \hat{y}) + \eta \cdot Penalty(W)$$

A process used to enforce simpler models to prevent overfitting.

## Implicit Regularization:

- All methods that are not directly added to the optimization problem.
- Examples include early stopping, introducing noise (eg with random batches, dropout), and discarding outliers.
- Ubiquitous in machine learning, including batched stochastic gradient descent and ensemble methods like random forests and gradient boosted trees.





# Many Models

## Wisdom of the crowd

In 1907 Francis Galton asked 787 villagers to guess the weight of an ox. None of them got the right answer, but when Galton averaged their guesses, he arrived at a near perfect estimate.







# Many Models

## Condorcet Jury Theorem

- Each member of the group has an independent probability of making a correct decision, which is better than random (greater than 50%).
- The decision is made by majority vote.

Under these conditions, the theorem demonstrates that as the size of the group increases, the probability that the majority decision is correct approaches 100%



# Many Models

## Diversity Prediction Theorem

Many Model Error = Average Model Error - Diversity of model predictions

$$(\bar{M} - V) = \sum \frac{(M_i - V)^2}{N} - \frac{(M_i - \bar{M})^2}{N}$$

$M_i$  is model  $i$ 's prediction

$\bar{M}$  is the average of the models values

$V$  is the true value



# Many Models

## Diversity Prediction Theorem

The logic relies on opposite errors (pluses and minuses) balancing each other.

It assumes:

1. Diverse predictions
2. Independent errors
3. Overestimates are as likely as underestimates
4. Sufficient competence (predictions are better than random guess)