

Neural Networks and Interpretability

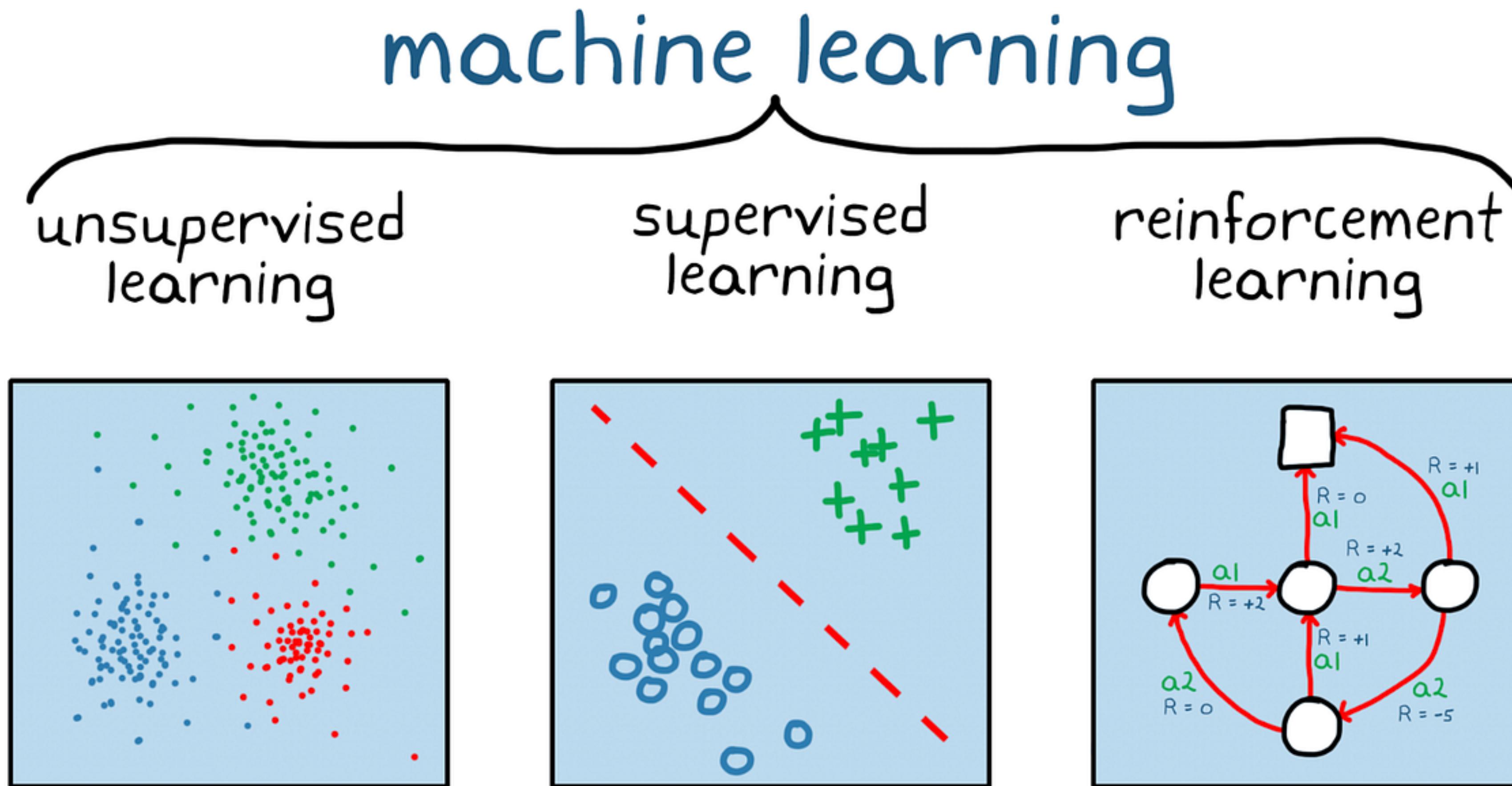
Indira Ocampo,
Instituto de Física Teórica, Madrid

Santiago, 15 - 16 de Enero de 2025

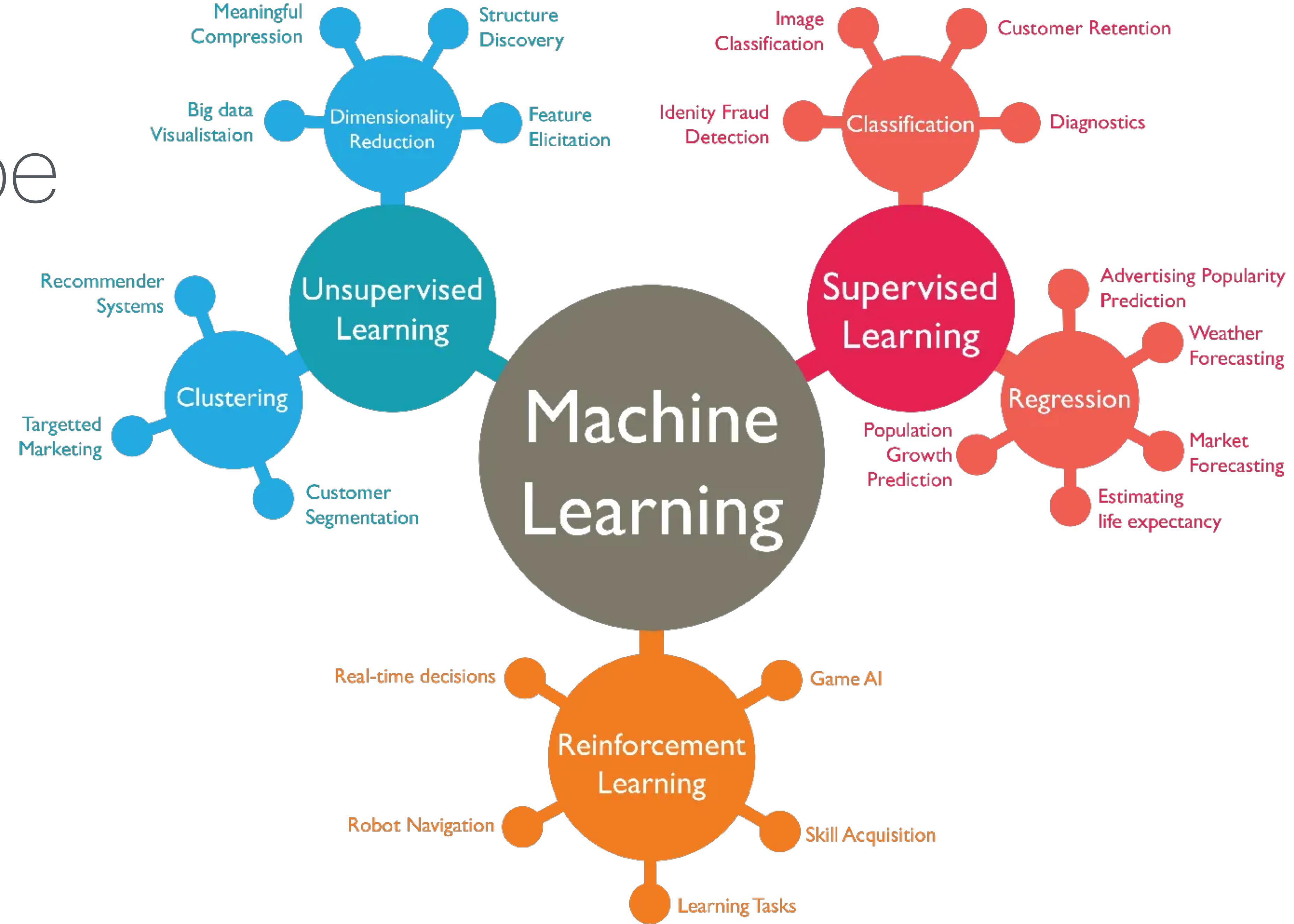
Outline (Part 1)

- Introducción
- Machine Learning landscape
- Linear Regression
- Logistic Regression
- Neural Networks
- Some examples

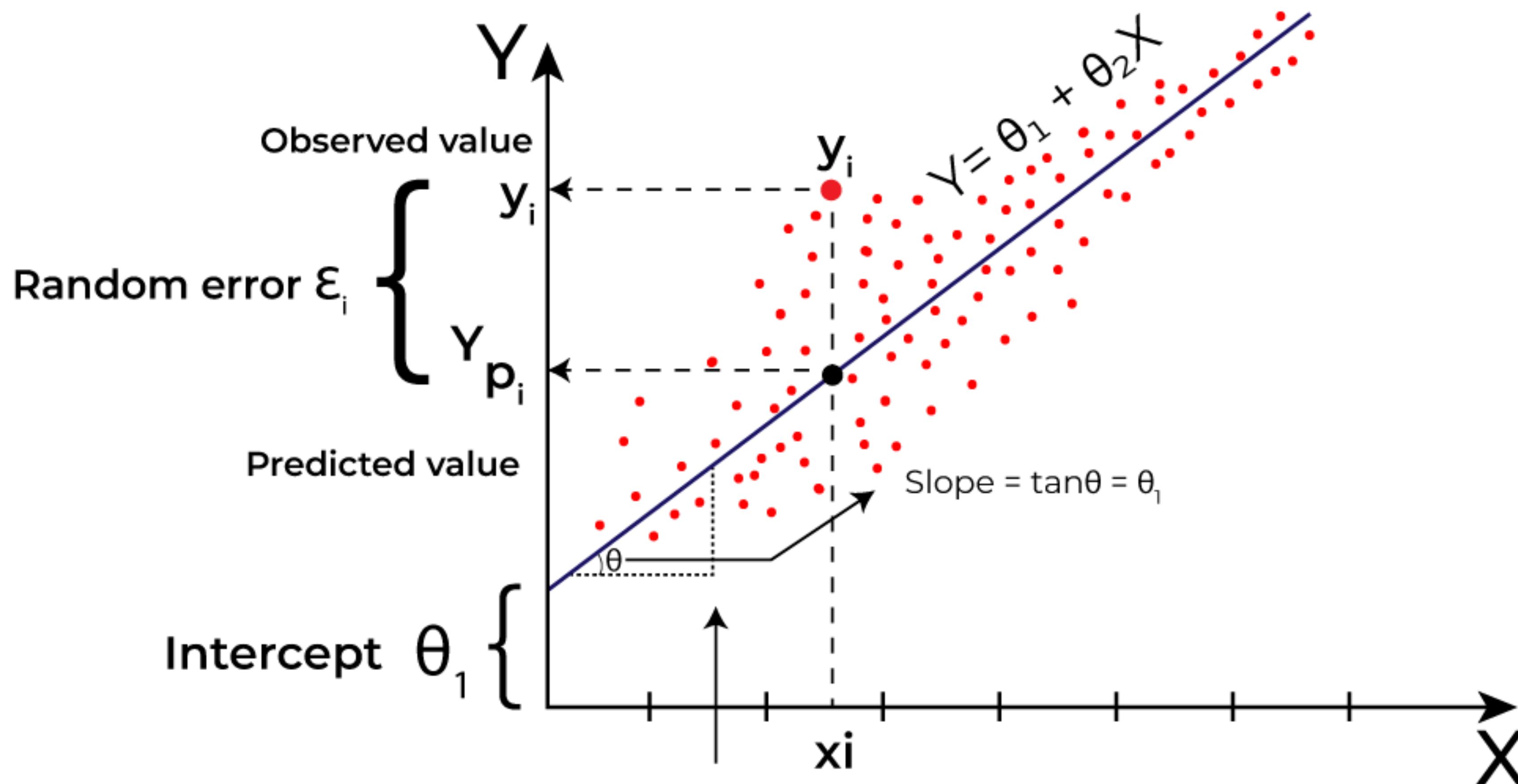
Introducción



ML landscape



Linear regression



$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

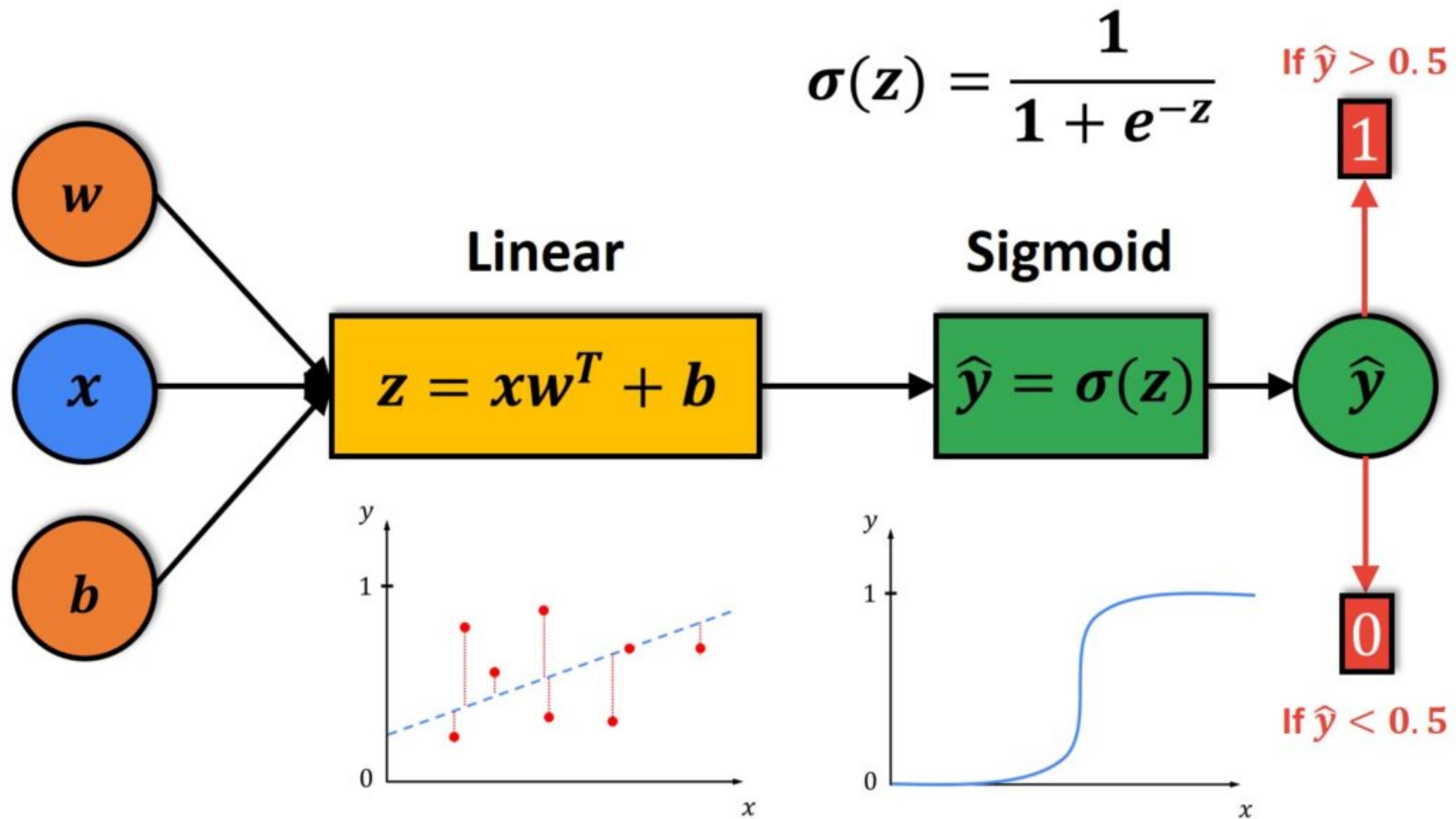
MSE = mean squared error

n = number of data points

Y_i = observed values

\hat{Y}_i = predicted values

Logistic regression



Logistic regression

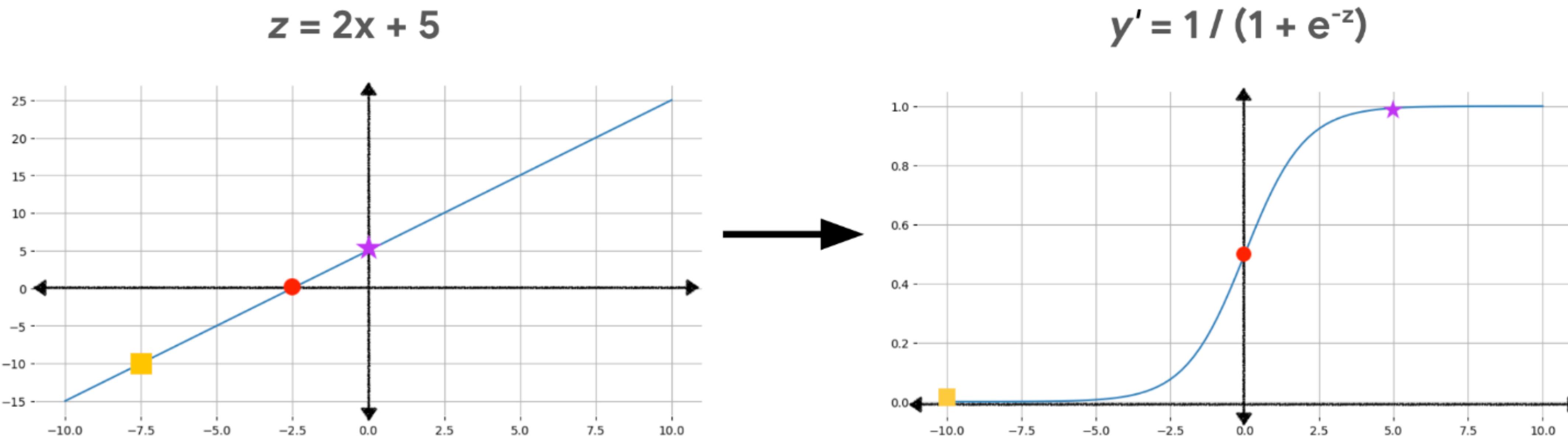
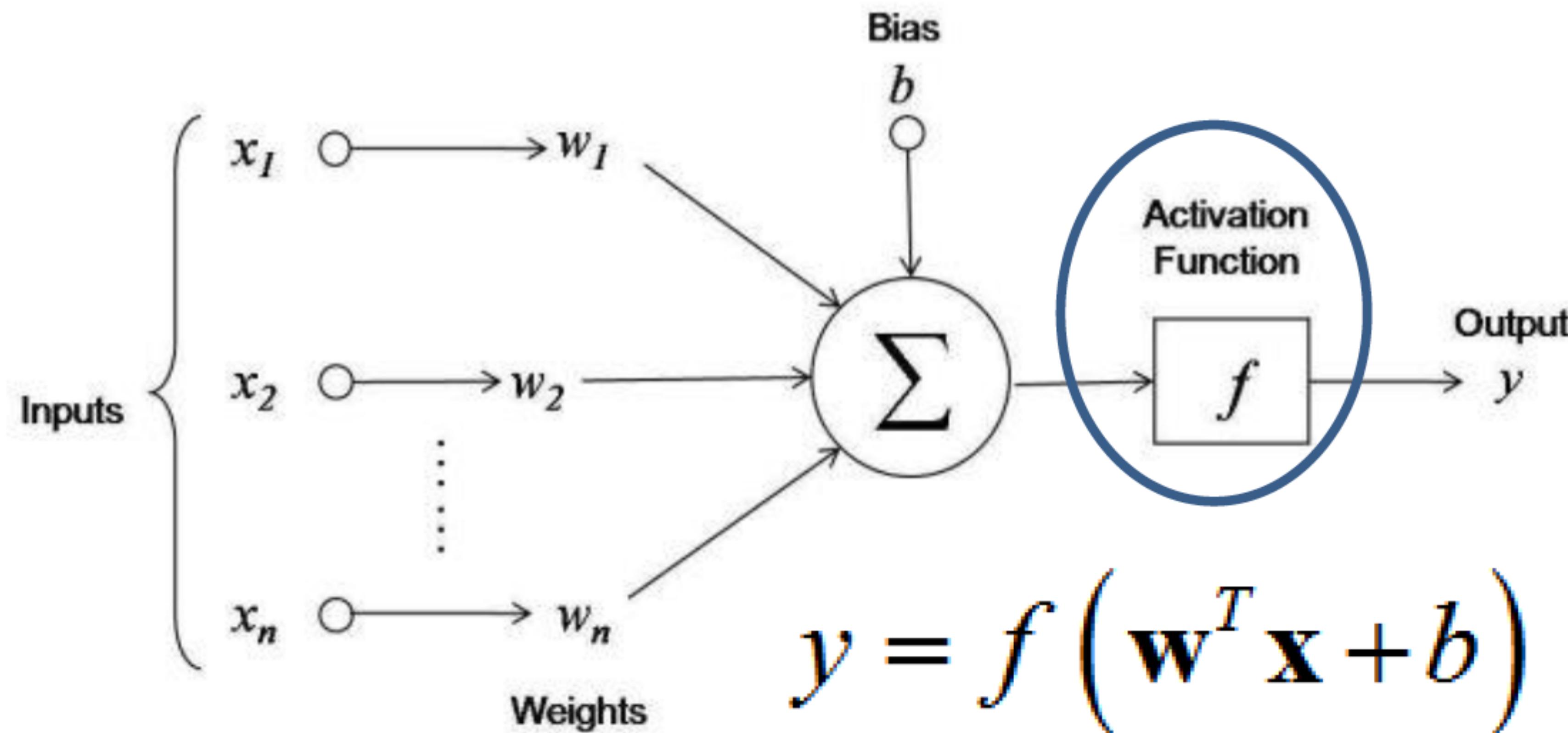


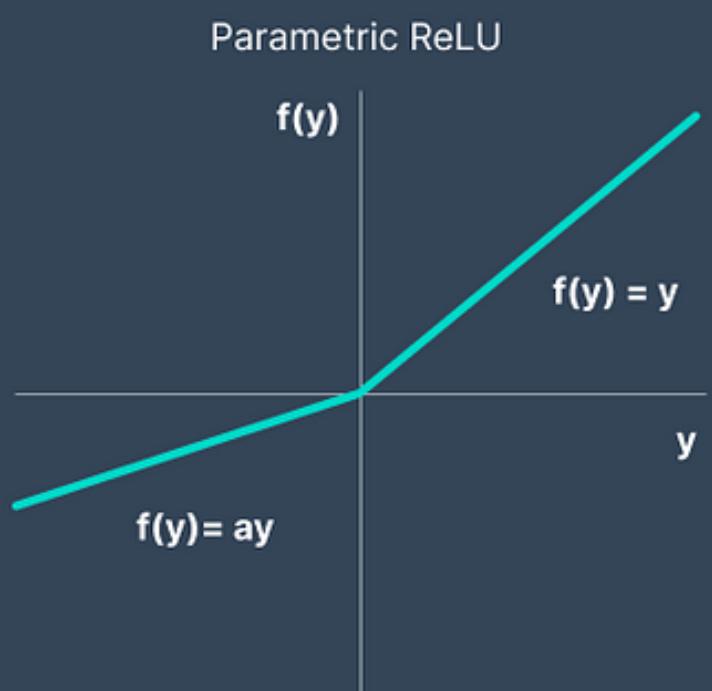
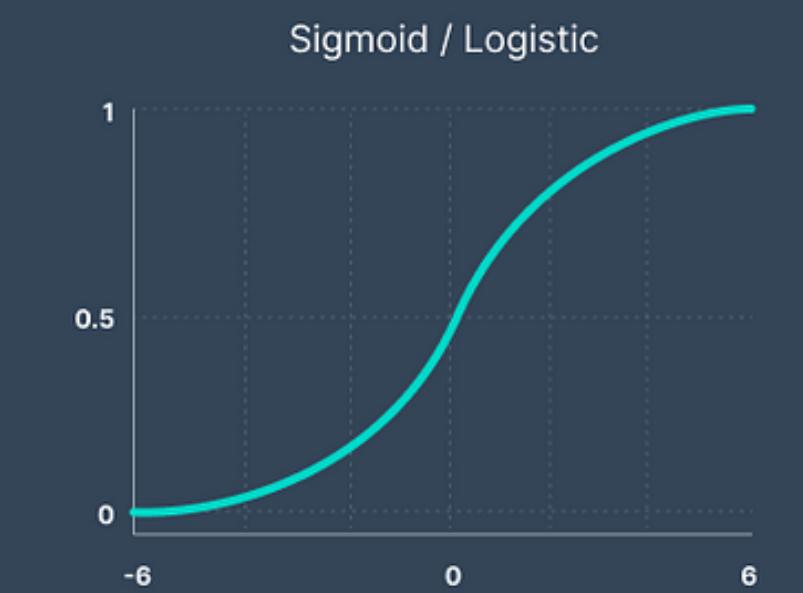
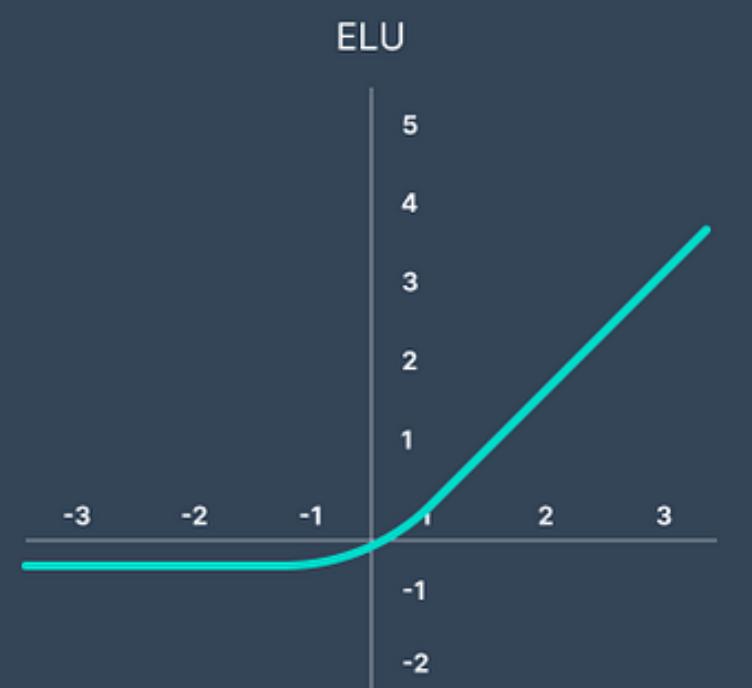
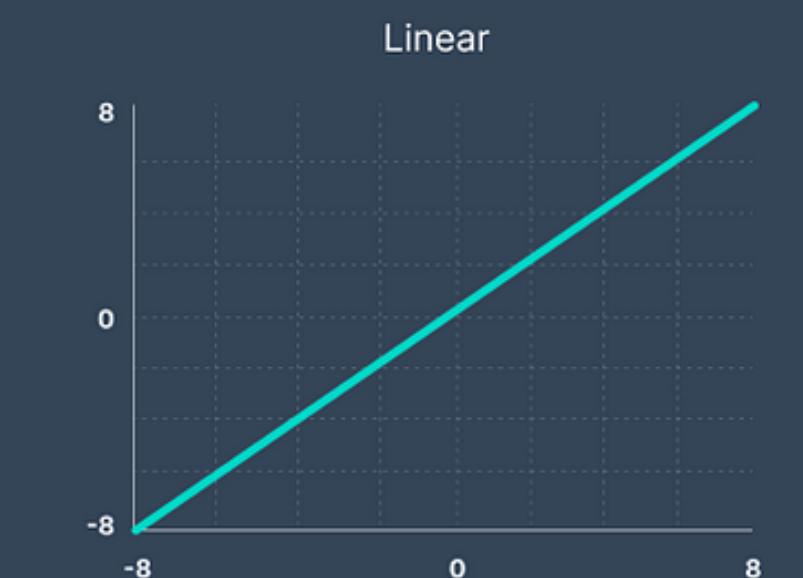
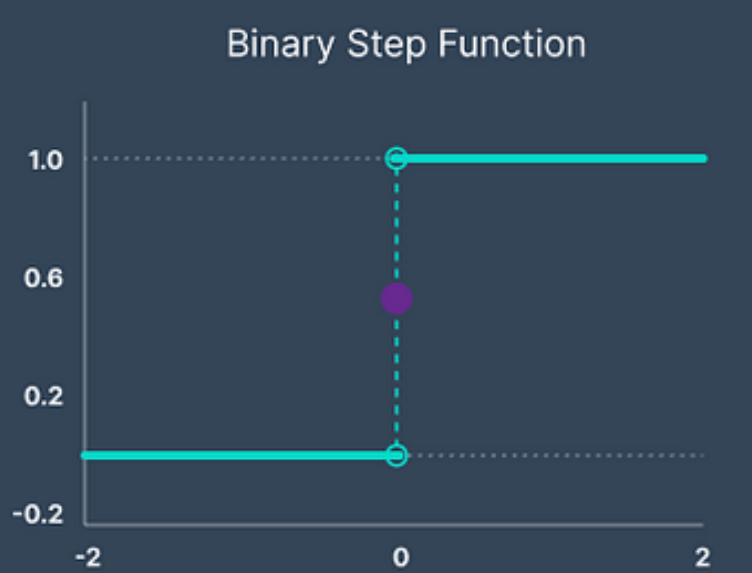
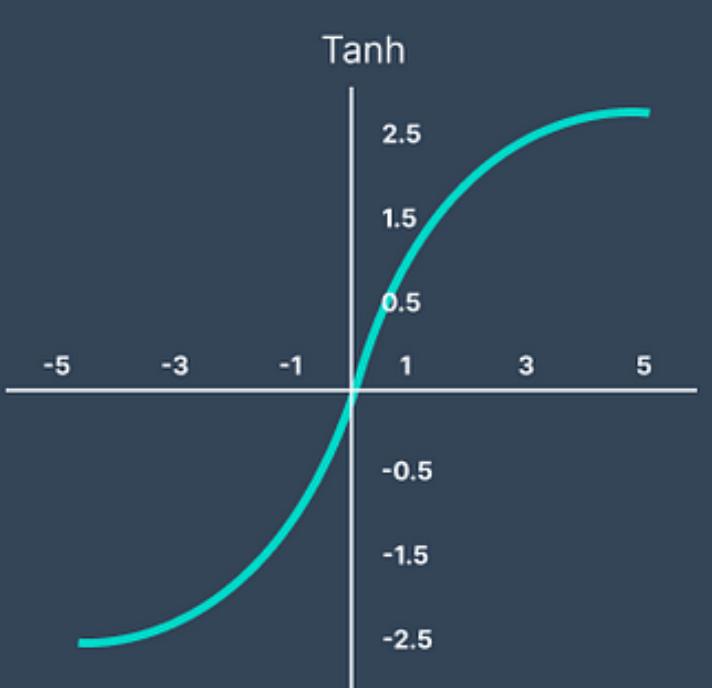
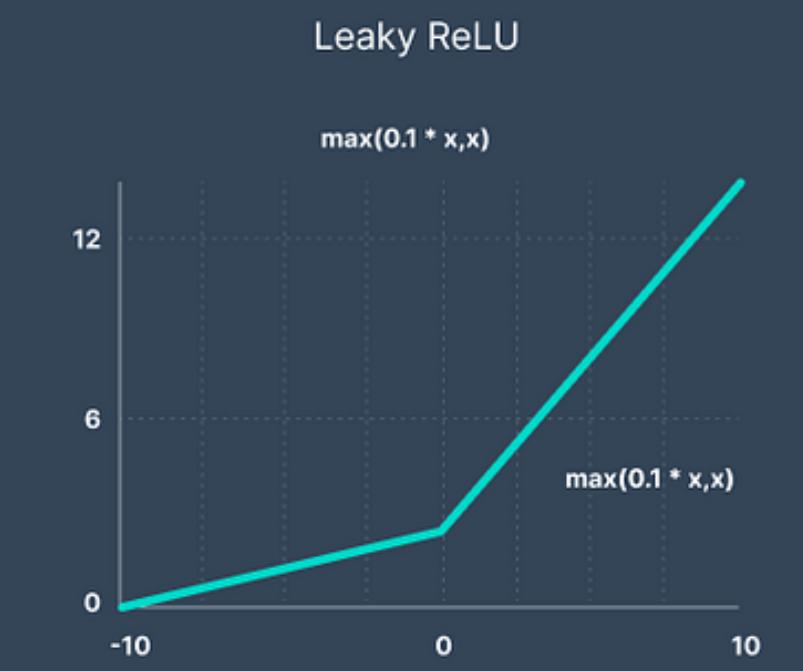
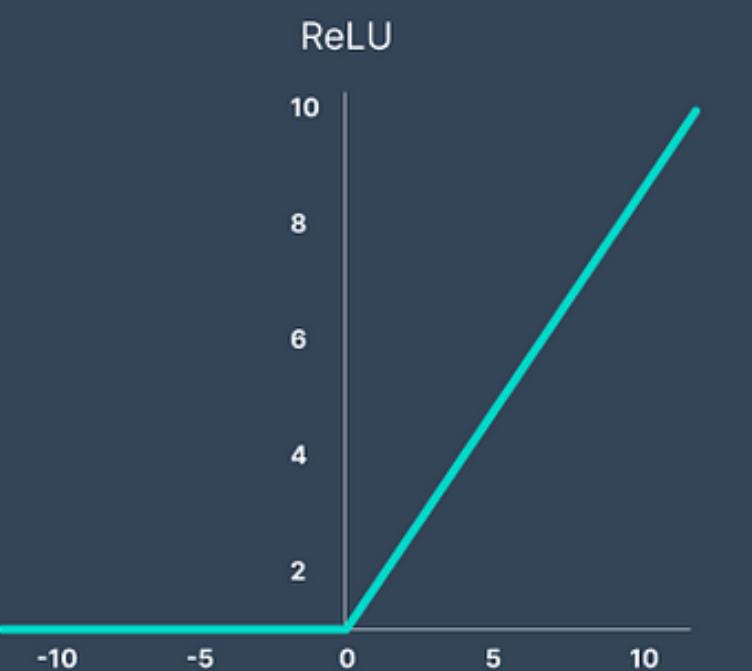
Figure 2. Left: graph of the linear function $z = 2x + 5$, with three points highlighted. Right: Sigmoid curve with the same three points highlighted after being transformed by the sigmoid function.

Neural Networks

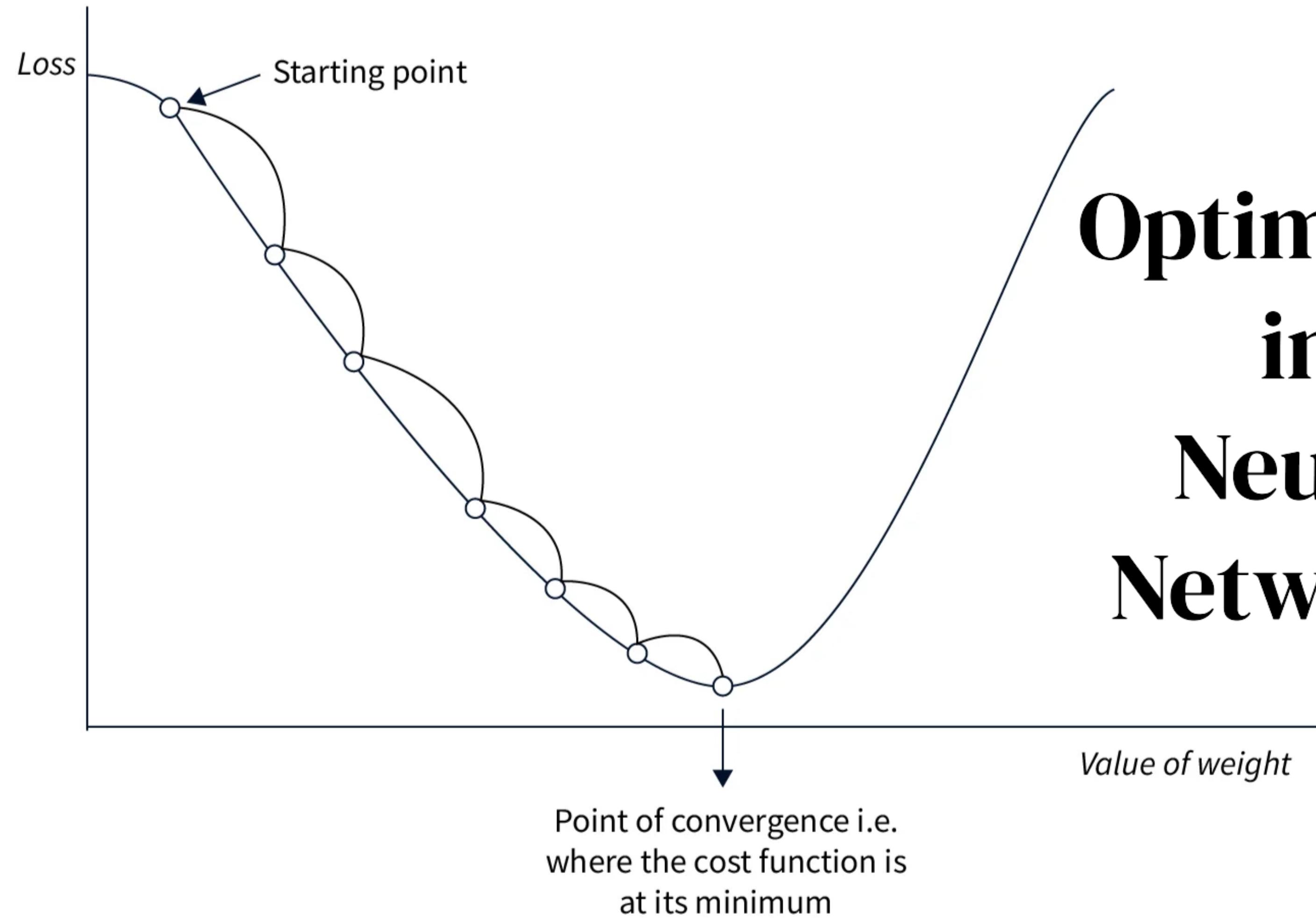


$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Activation functions

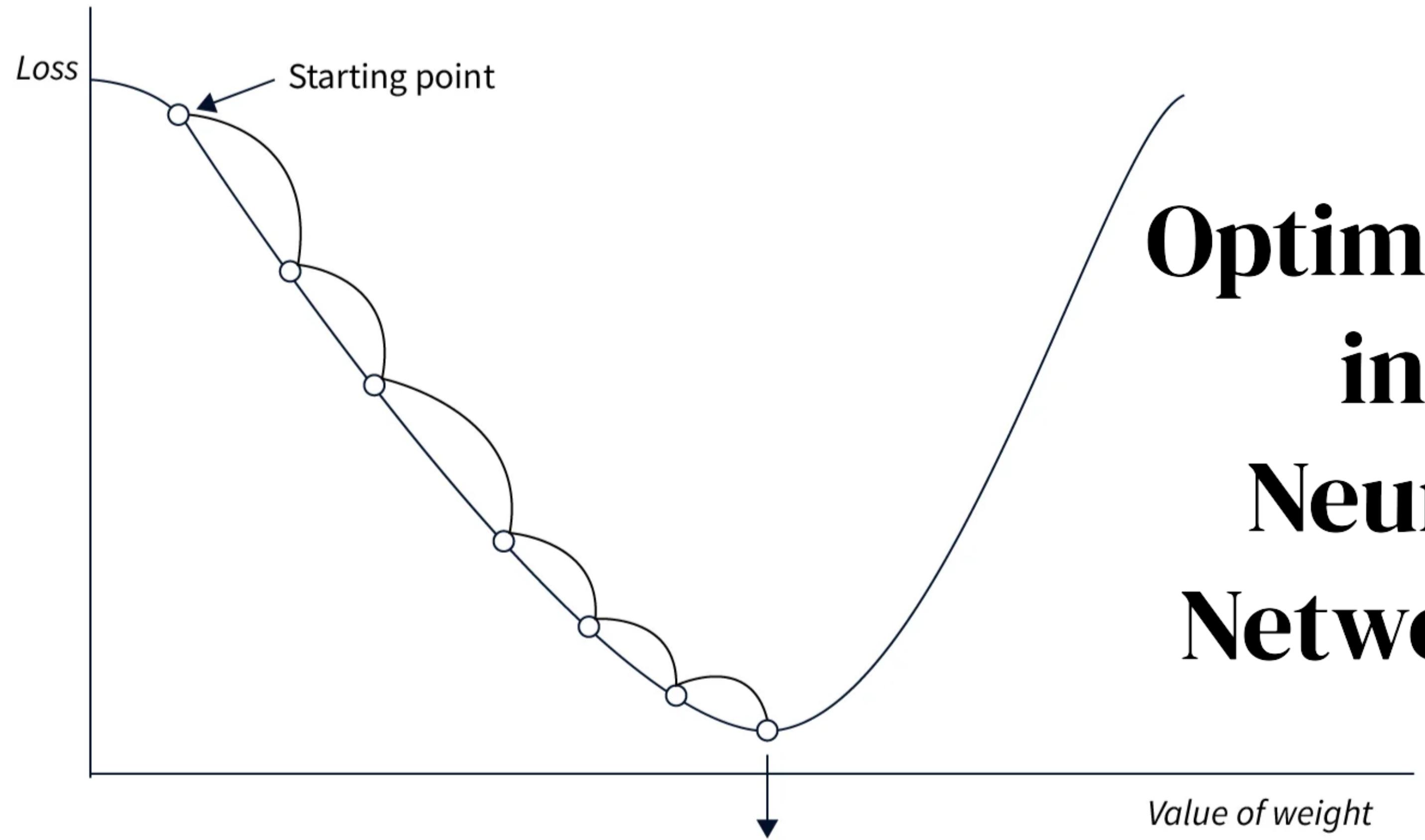


Optimizers



**Optimizers
in
Neural
Networks**

Optimizers

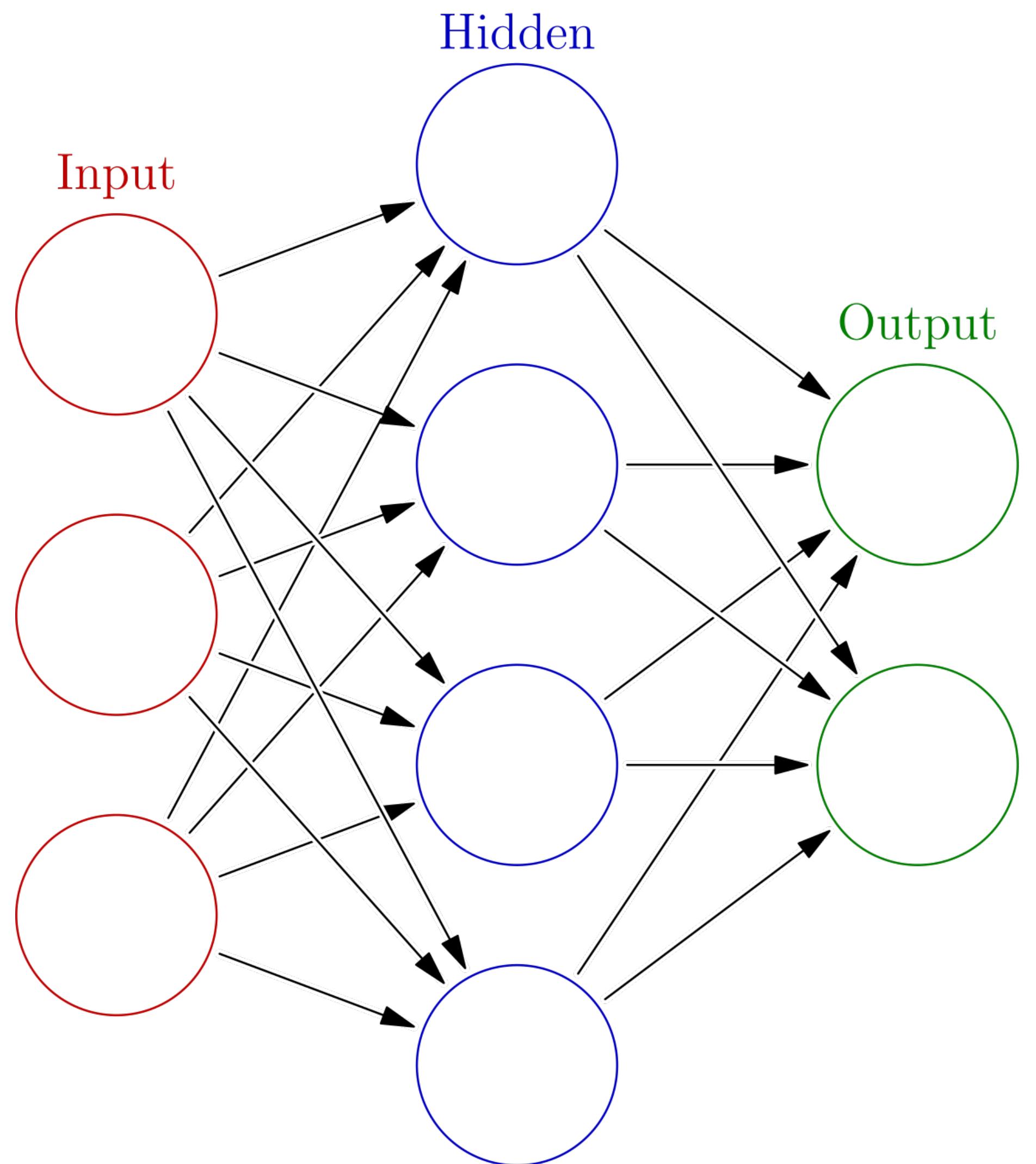


Optimizers in Neural Networks

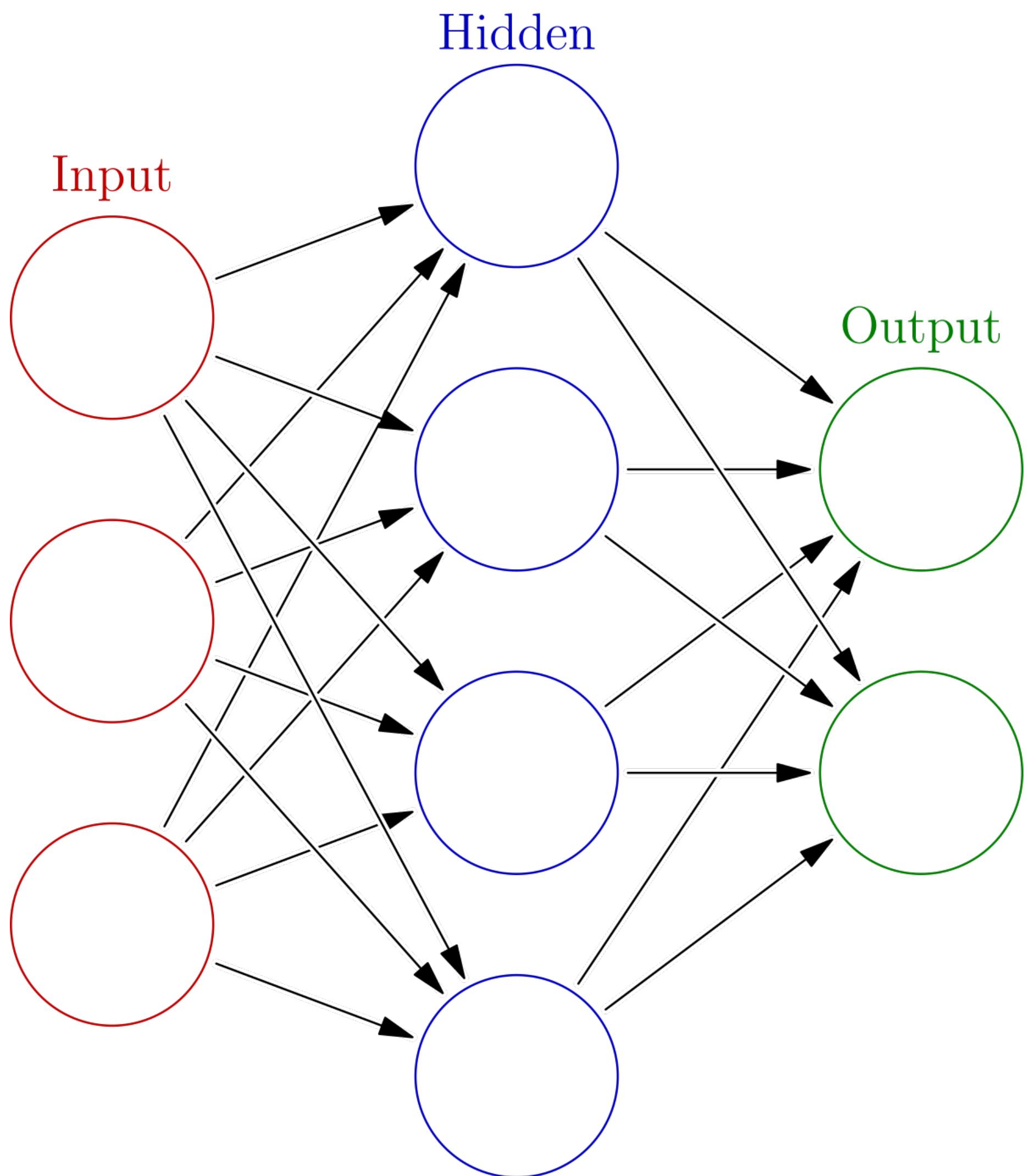
$$w_{\text{new}} = w_{\text{old}} - \alpha \frac{\partial L}{\partial w}$$

$$b_{\text{new}} = b_{\text{old}} - \alpha \frac{\partial L}{\partial b}$$

Neural Networks



Neural Networks



- Forward Pass

$$y = \mathbf{w} \cdot \mathbf{x} + b \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

- Compute the Loss

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- Backpropagation of errors
(gradient descent)

$$\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} - \alpha \frac{\partial L}{\partial w} \quad b_{\text{new}} = b_{\text{old}} - \alpha \frac{\partial L}{\partial b}$$

- Repeat (until minimizing the loss)



epoch

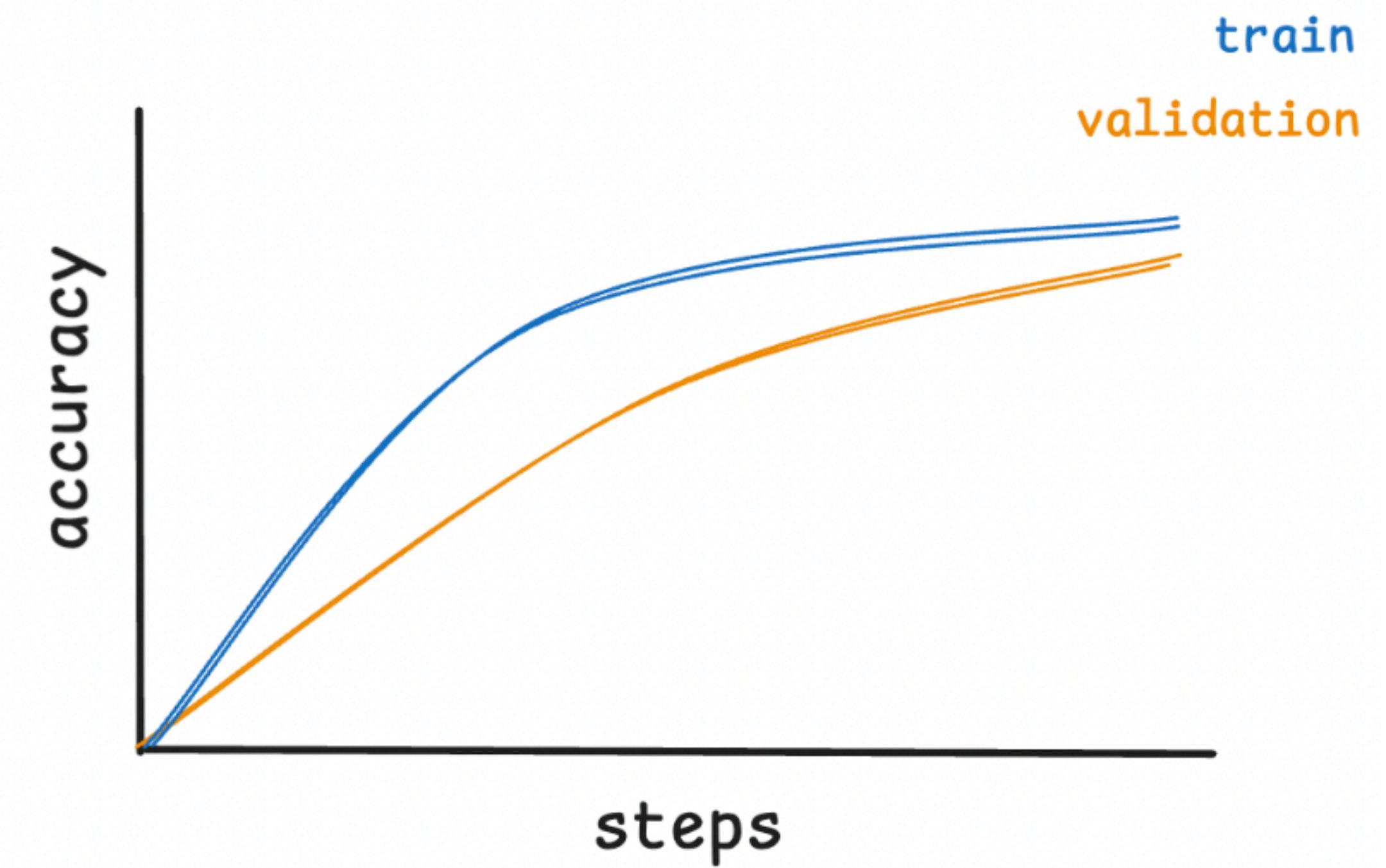
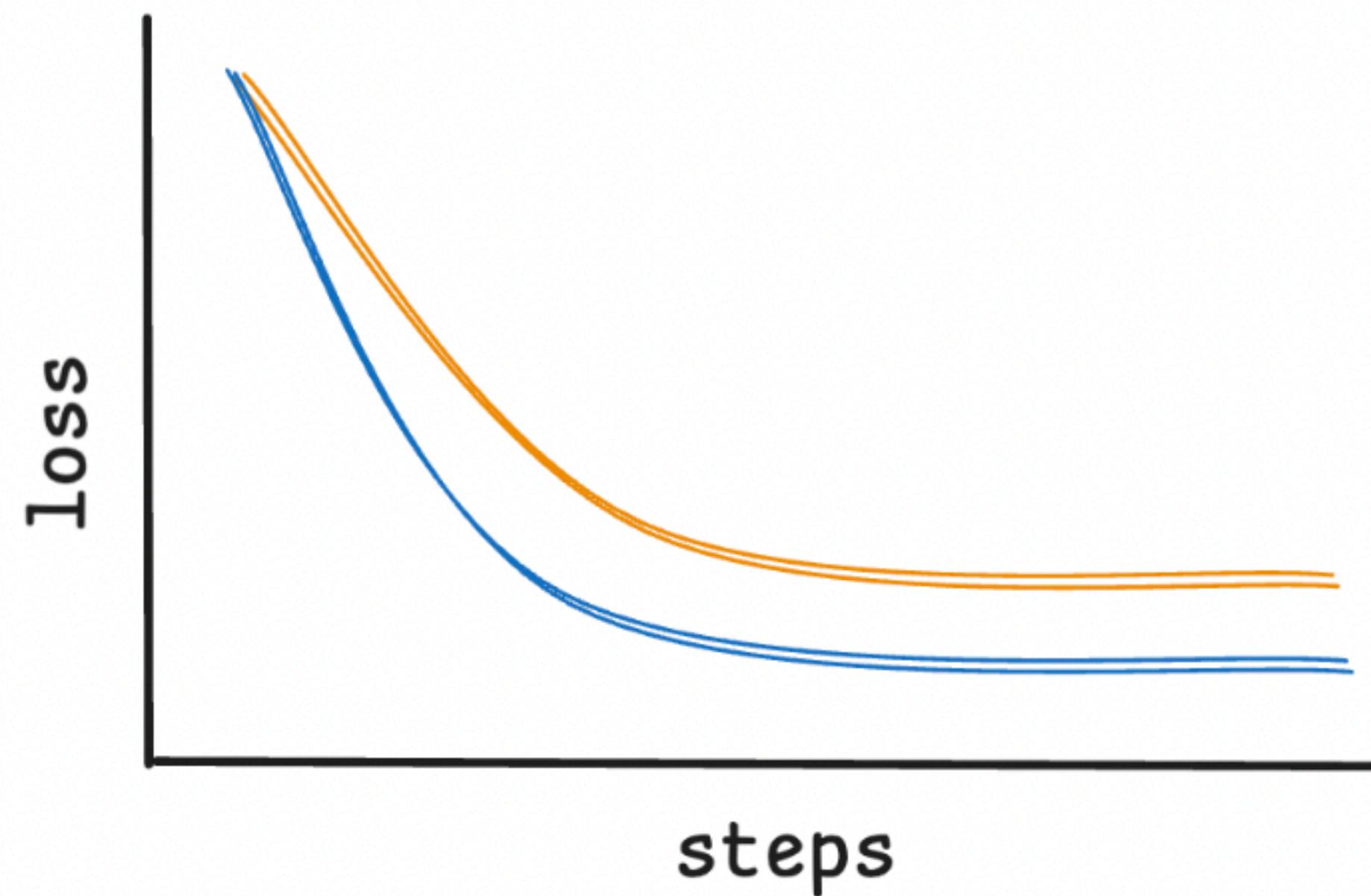
Evaluation metrics

Classification tasks

- Learning curves
- Confusion Matrix
- ROC curve
- AUC

Learning Curves

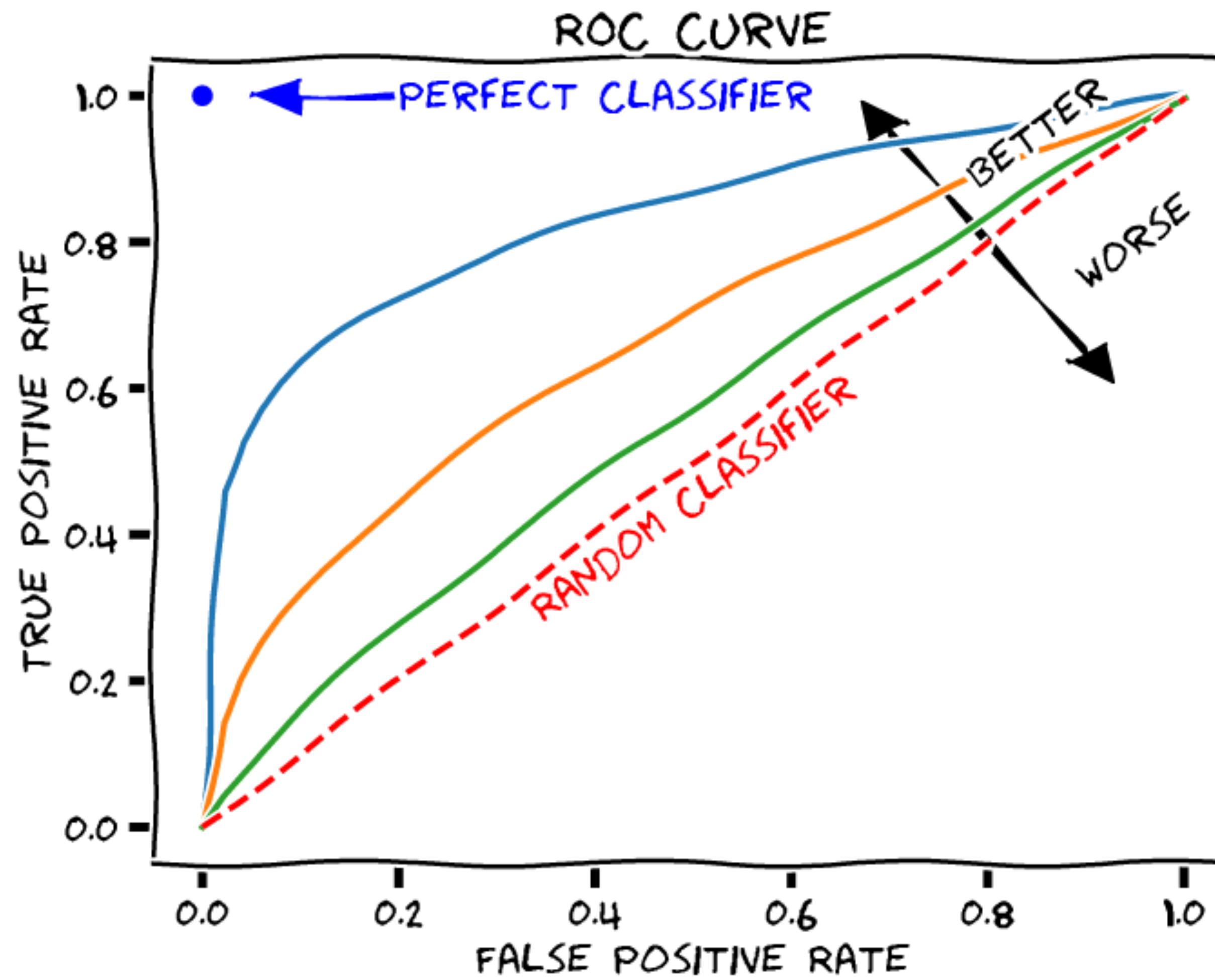
- Accuracy: percentage of correct predictions over time
- Loss: how the model's error decreases over time



Confusion Matrix

		Predicted Class		
		Negative	Positive	
Actual Class	Negative	True Negative (TN)	False Positive (FP) (Type I Error)	Specificity $\frac{TN}{(TN + FP)}$
	Positive	False Negative (FN) (Type II Error)	True Positive (TP)	Recall/Sensitivity $\frac{TP}{(TP + FN)}$
		Negative Predictive Value $\frac{TN}{(TN + FN)}$	Precision $\frac{TP}{(TP + FP)}$	Accuracy $\frac{(TP+TN)}{(TP+FP+TN+FN)}$

Receiver-operating characteristic curve (ROC)



True Class	T	F
Acquired Class	True Positives (TP)	False Positives (FP)
	False Negatives (FN)	True Negatives (TN)

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

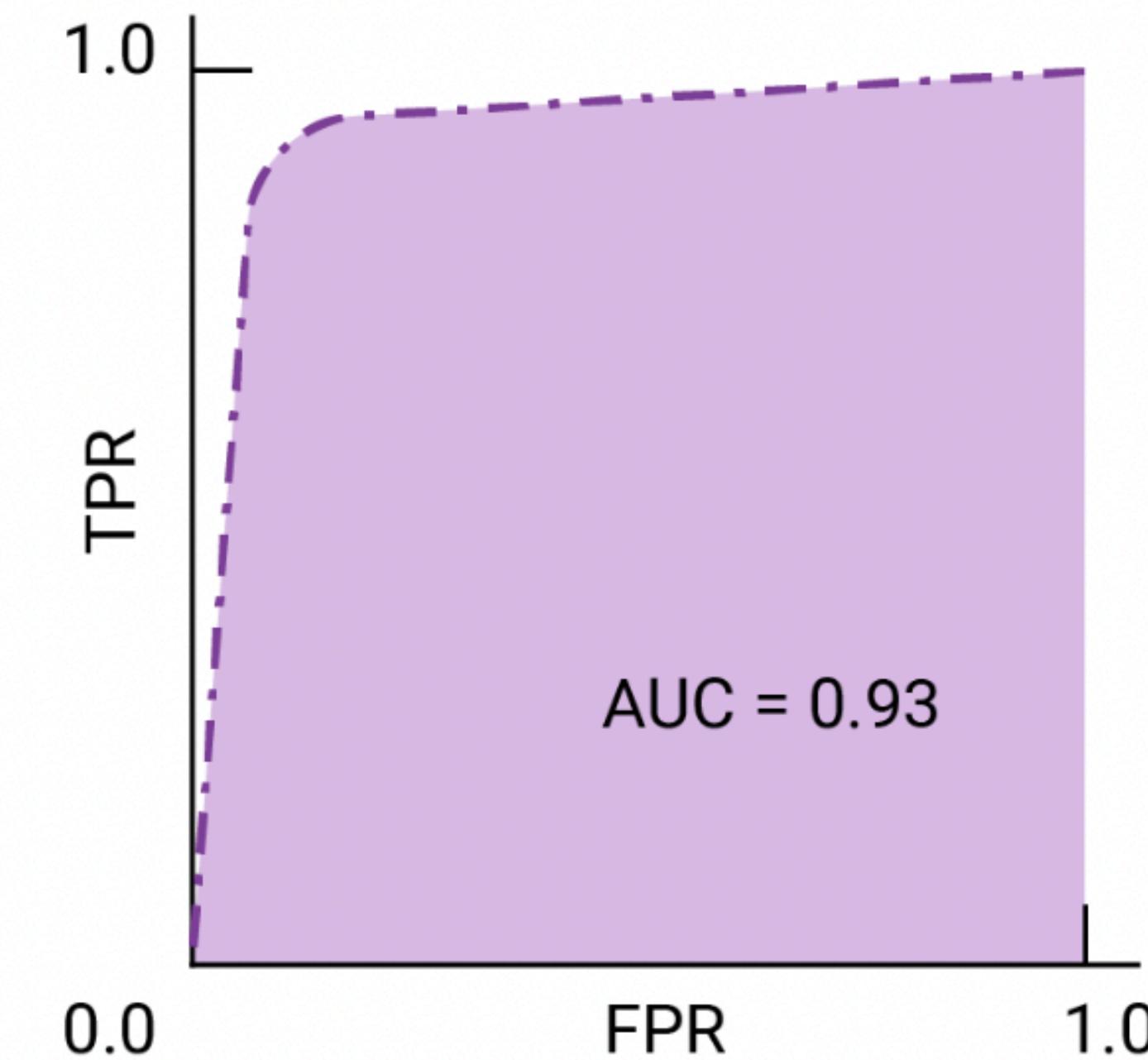
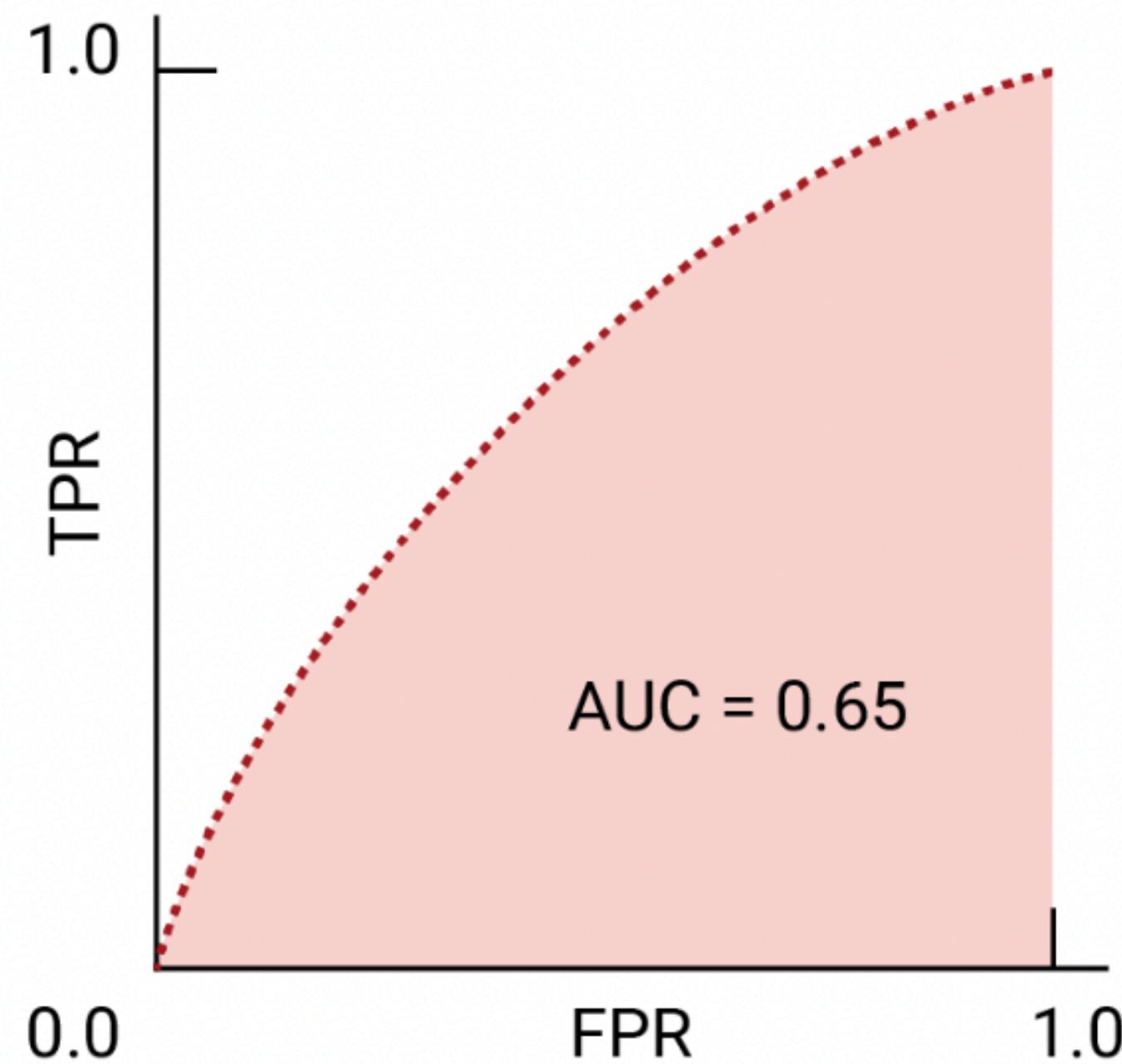
$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Accuracy

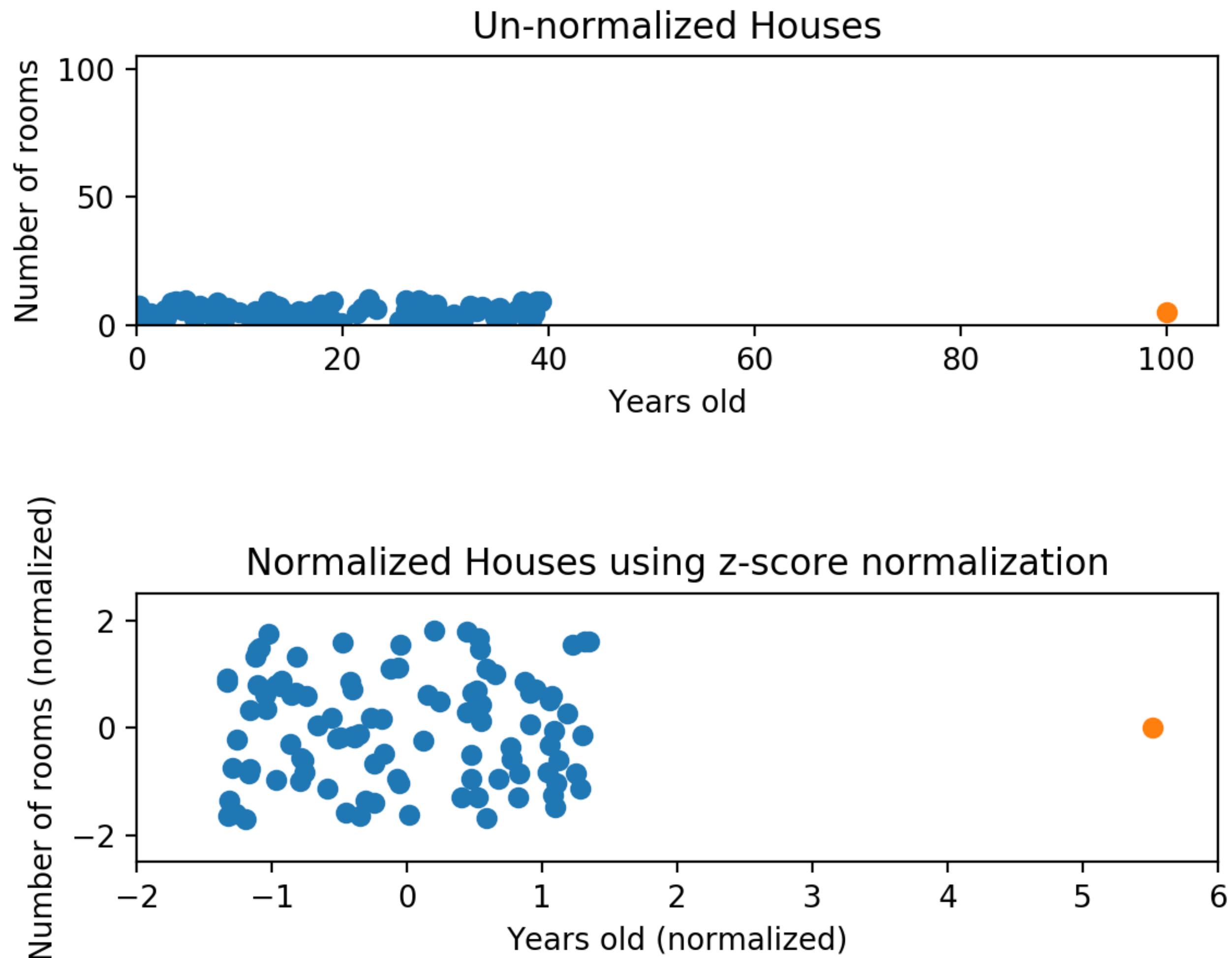
$$(\text{ACC}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Area Under the ROC Curve

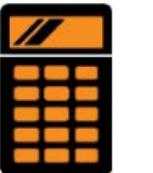
- >> How well the model distinguishes between positive and negative classes across all possible classification thresholds



Normalization

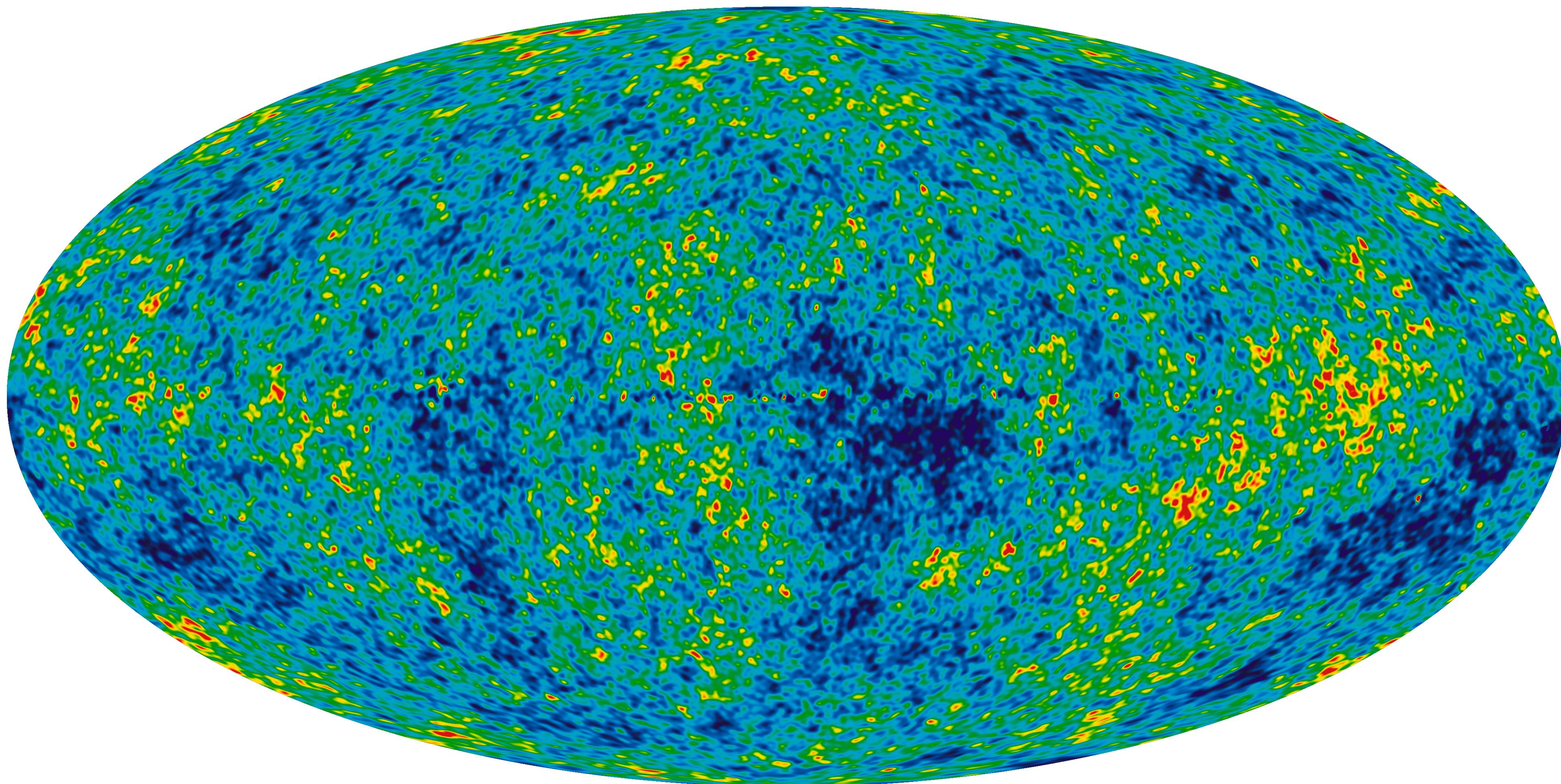


Normalization Formula

$$X_{\text{normalized}} = \frac{(X - X_{\text{minimum}})}{(X_{\text{maximum}} - X_{\text{minimum}})}$$


An example

Cosmic Microwave Background



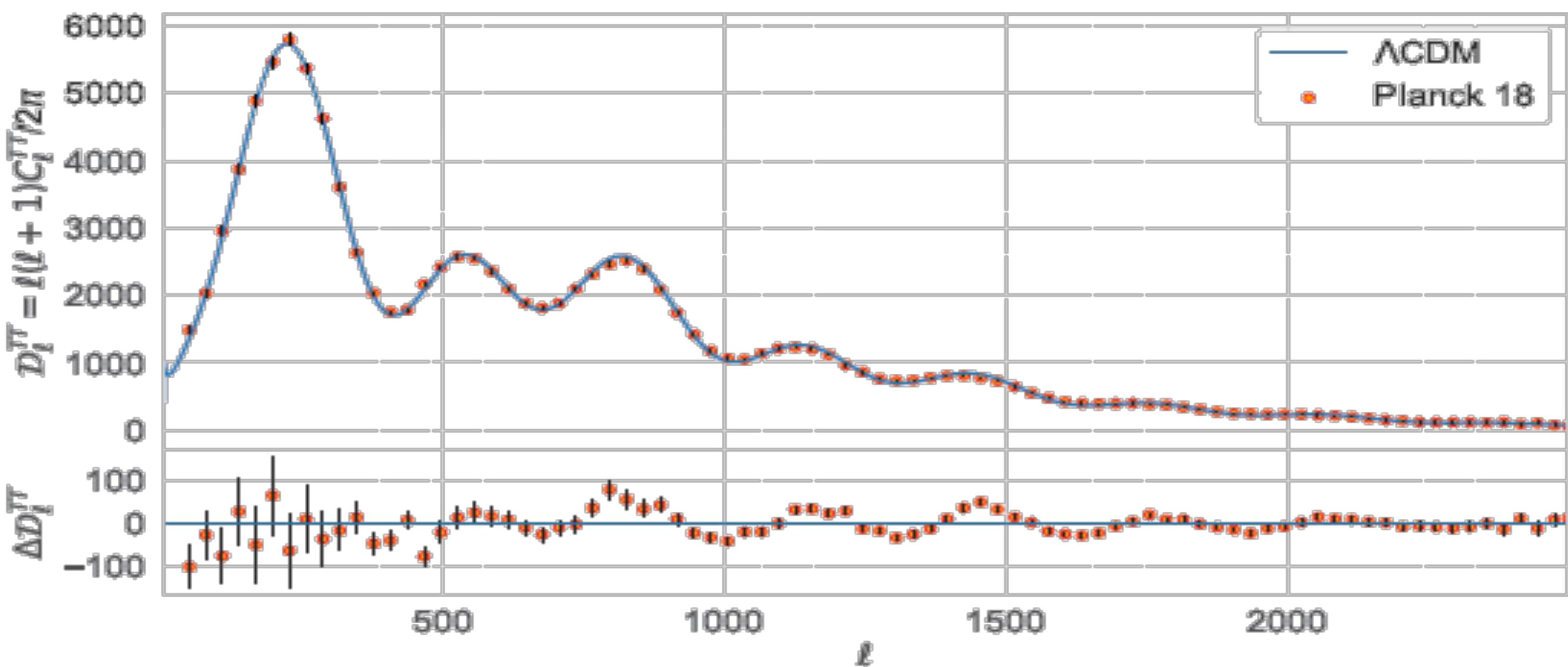
The CMB

Angular Power Spectrum

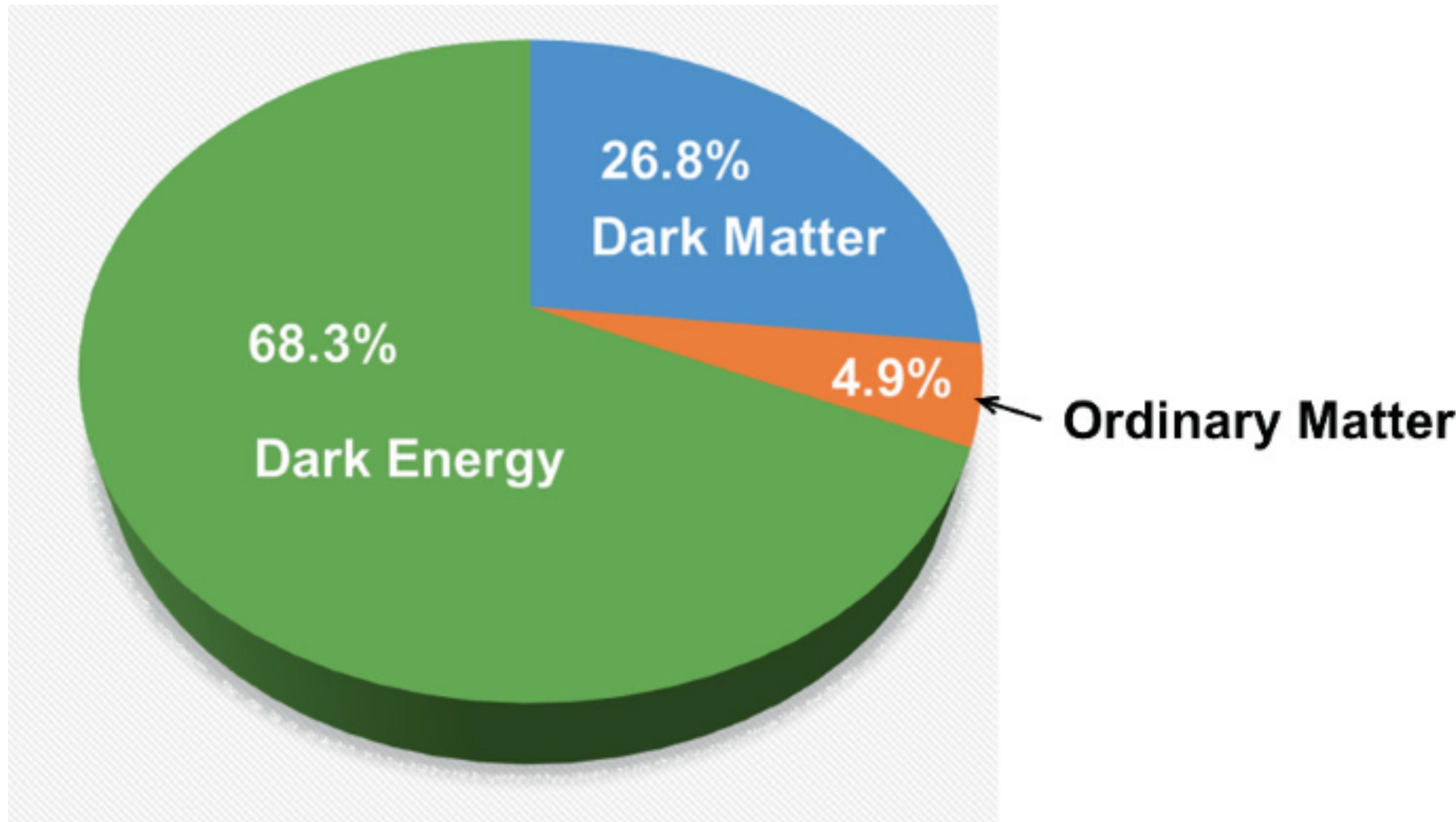
$$\frac{\Delta T}{T} \sim 10^{-5}$$

$$\frac{\Delta T}{T}(\hat{n}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{n})$$

$$C_{\ell} = \langle |a_{\ell m}|^2 \rangle$$

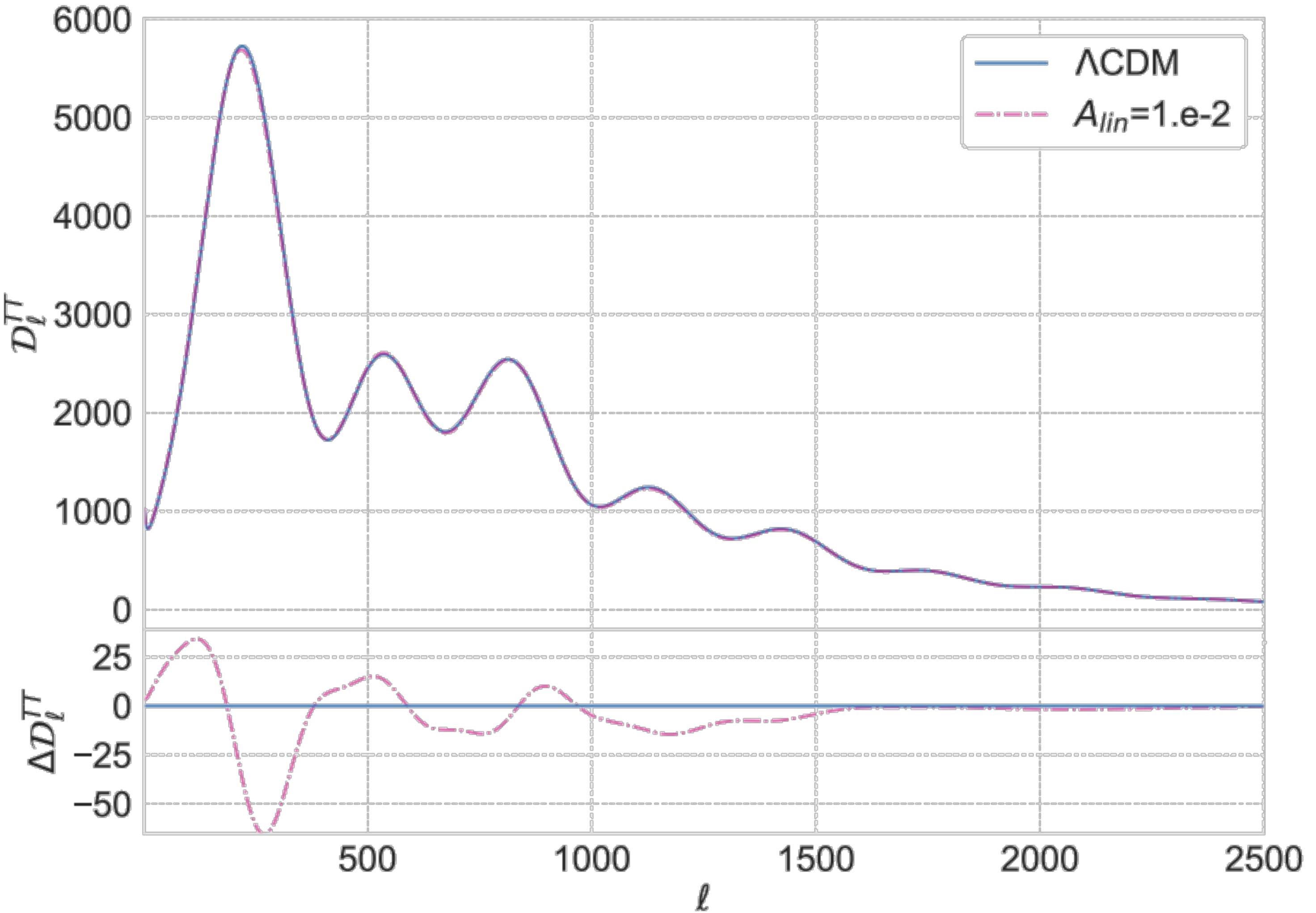


The Λ CDM model

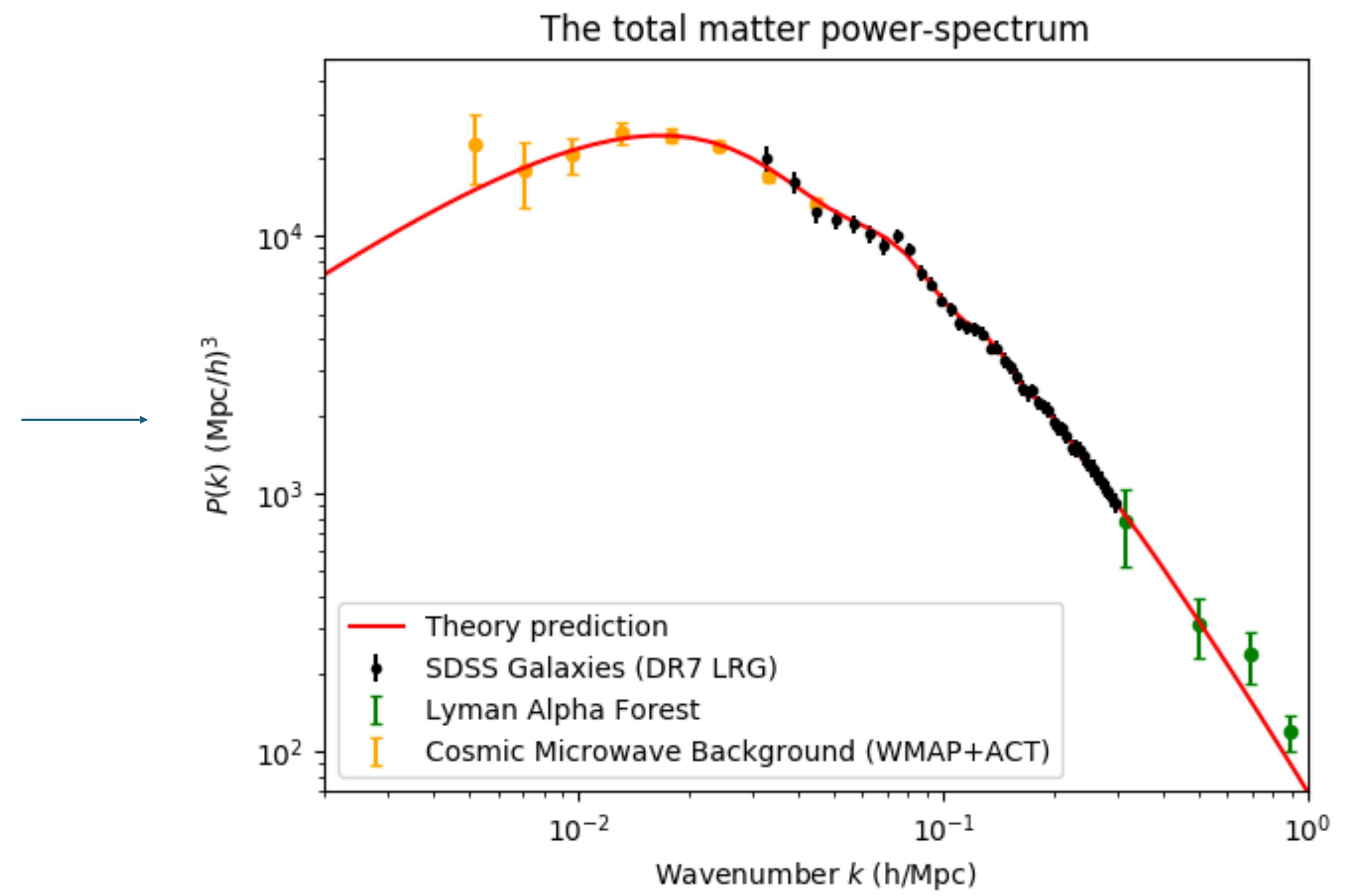
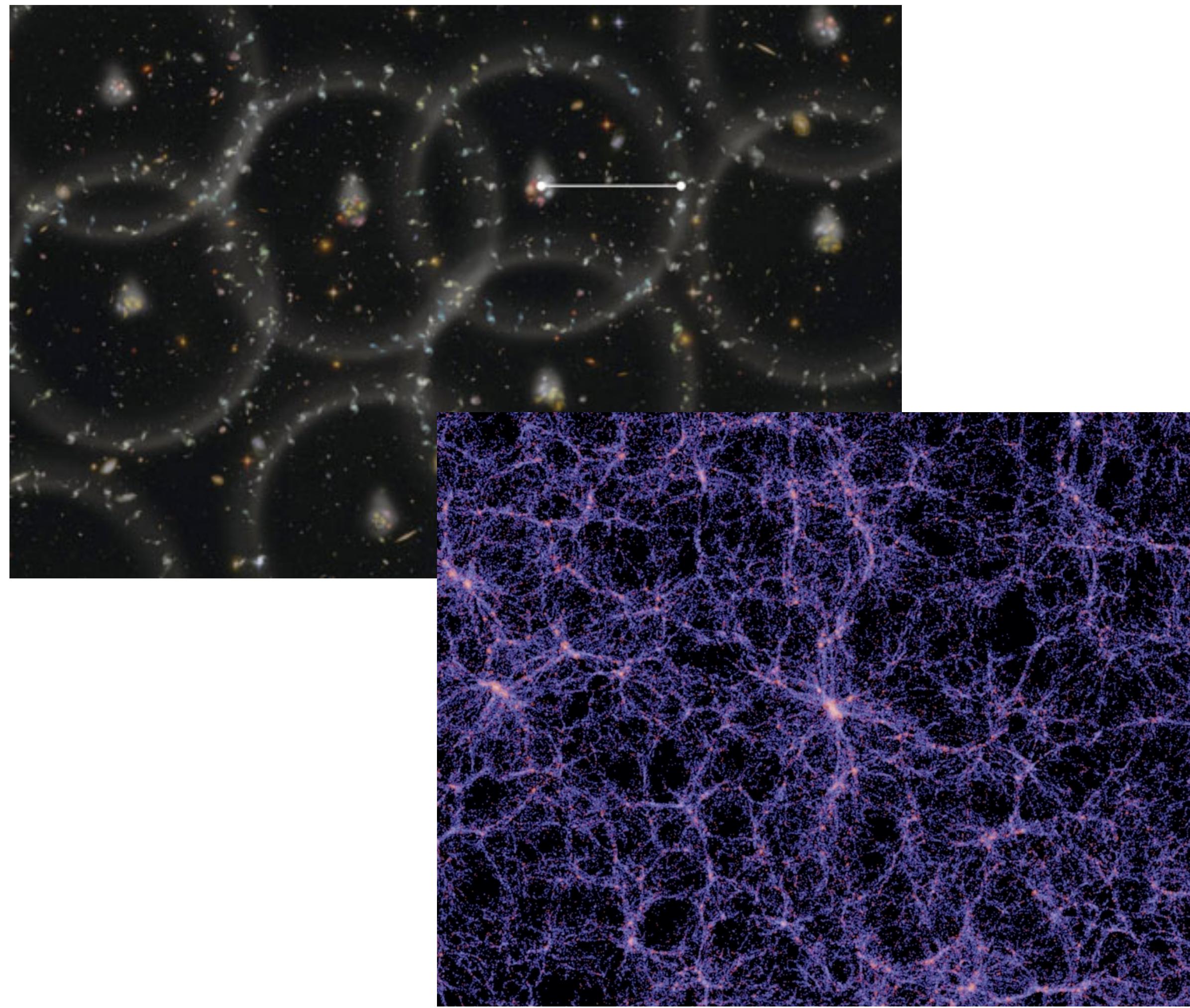


Alternatives?

- Modified gravity
- Non-standard primordial physics
- Dynamical Dark Energy
- ...



Your turn!



Your turn!

2. Exercise Tasks

Task 1:

Split the dataset into train and validation sets.

Task 2:

Build and train a neural network classifier.

Suggested architecture:

- Dense(64, relu)
- Dense(64, relu)
- Dense(1, sigmoid)

Train for 40 epochs.

Task 3:

Evaluate the learning curves, plot ROC, AUC and the confusion matrix.

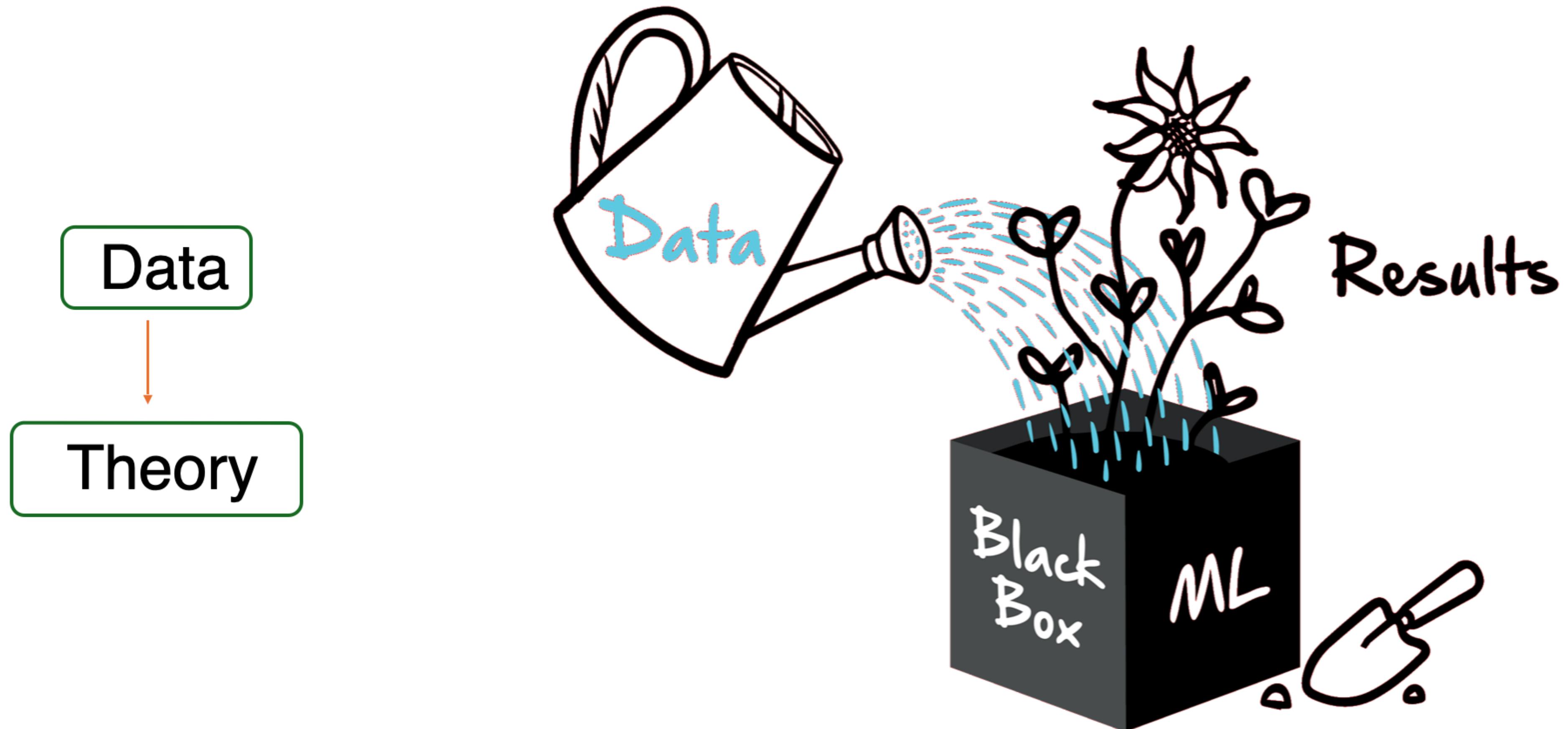
Task 4 (important):

Normalize the inputs using `StandardScaler` and retrain the model, what happens?

Outline (Part 2)

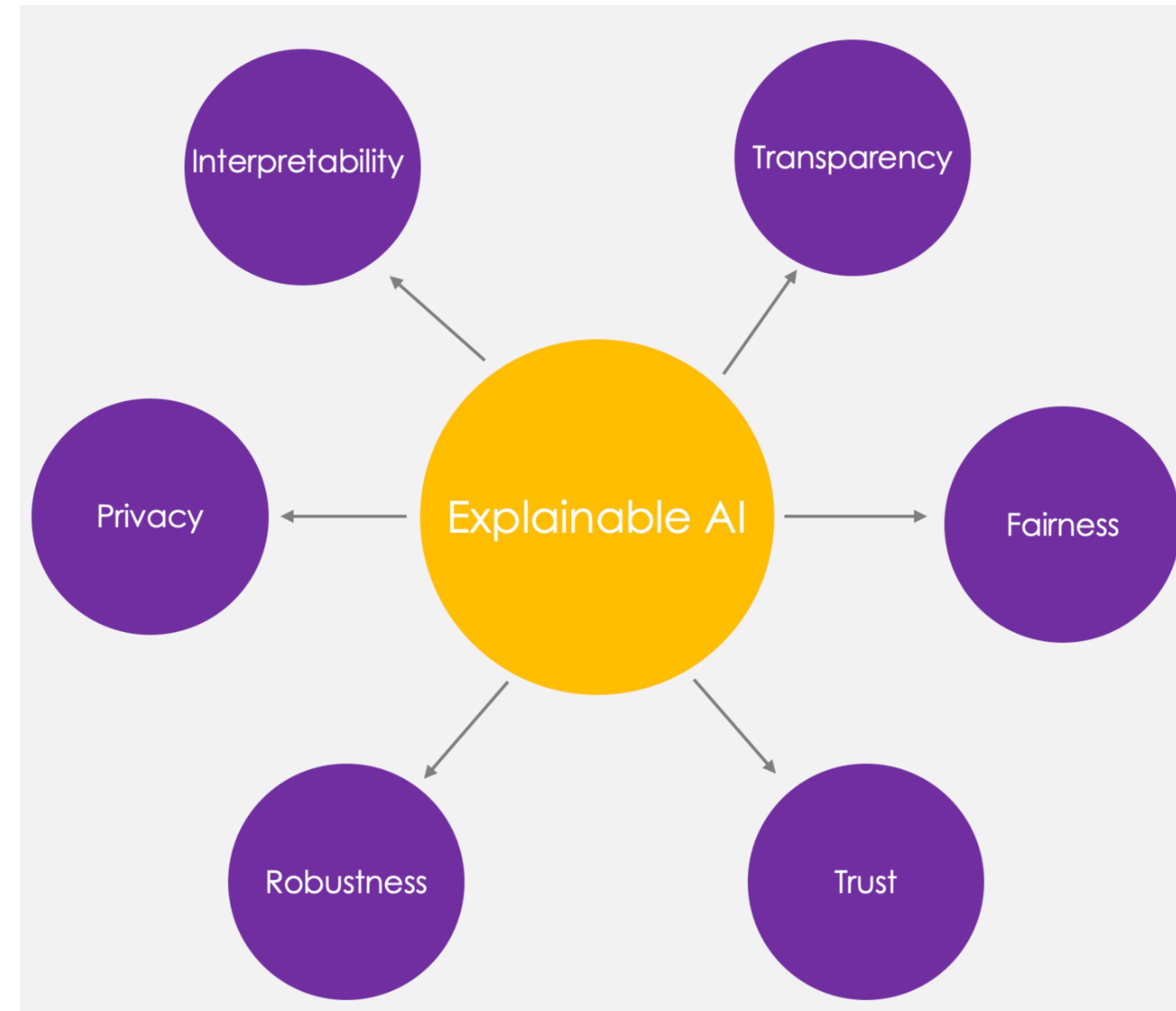
- Introduction
- Machine Learning interpretability
- LIME
- SHAP
- Some examples

Current approach

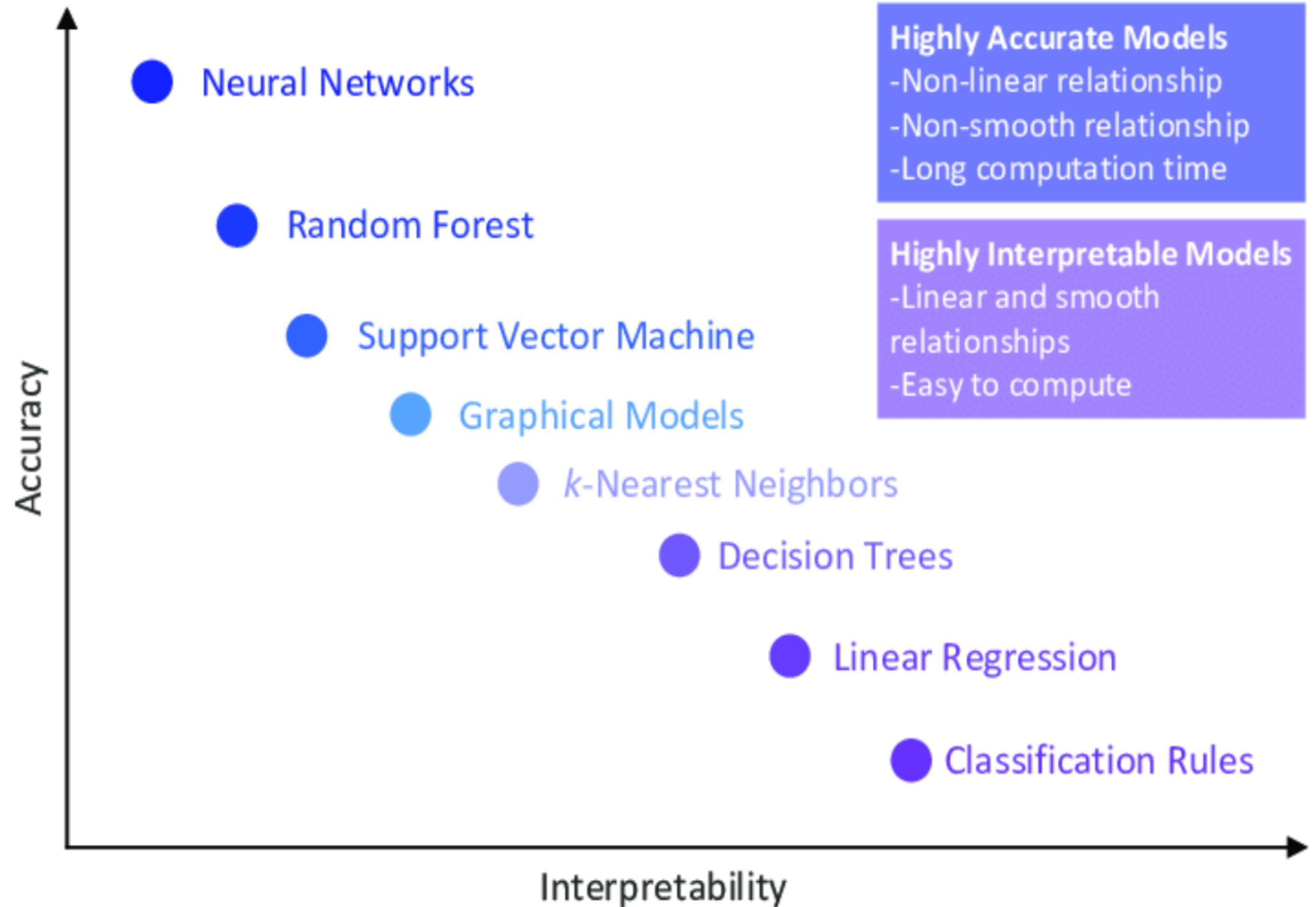


Source: <https://simons.berkeley.edu/>

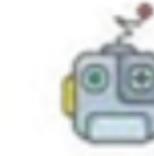
Explainable AI (X)



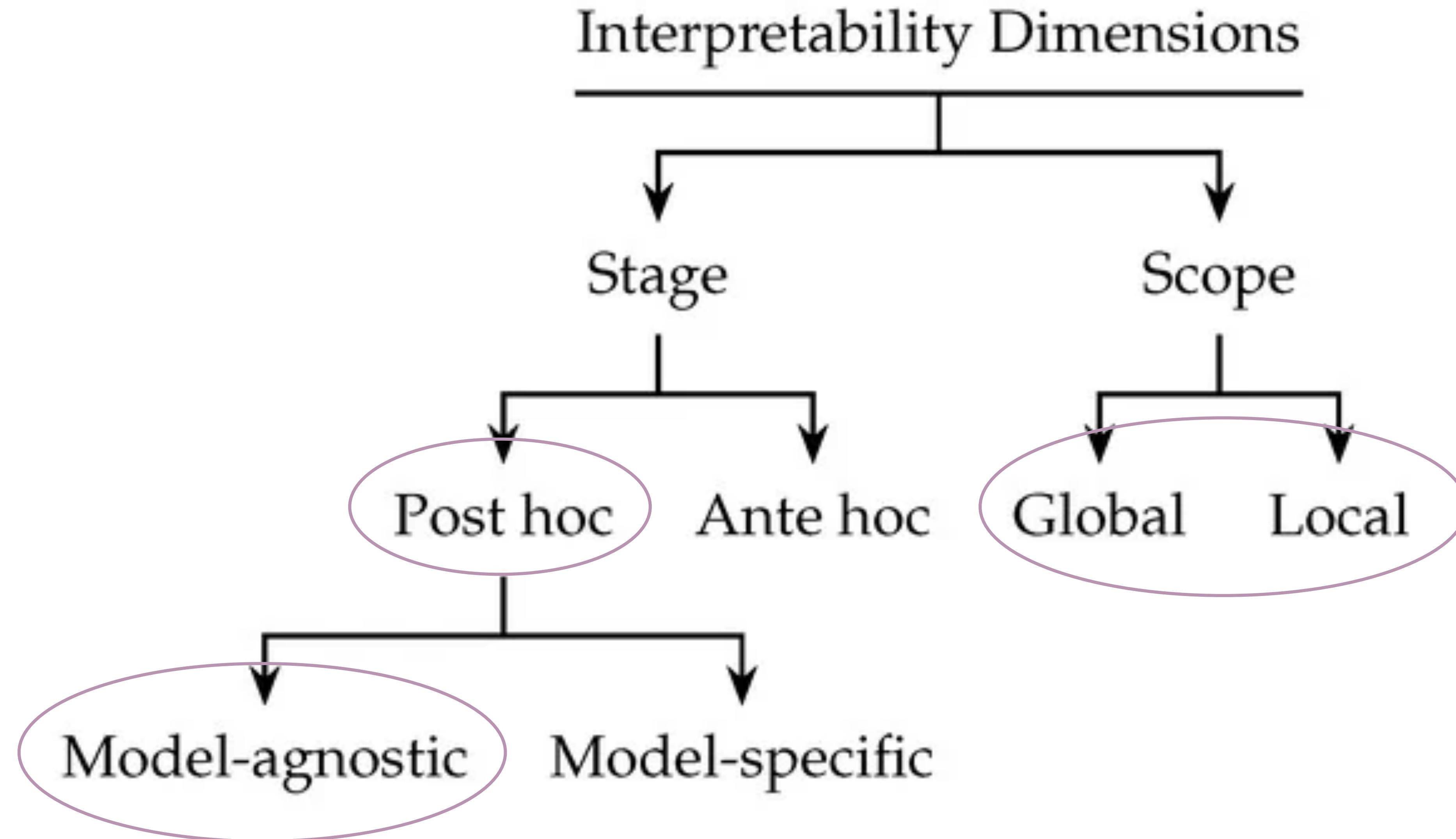
Interpretability in Machine Learning



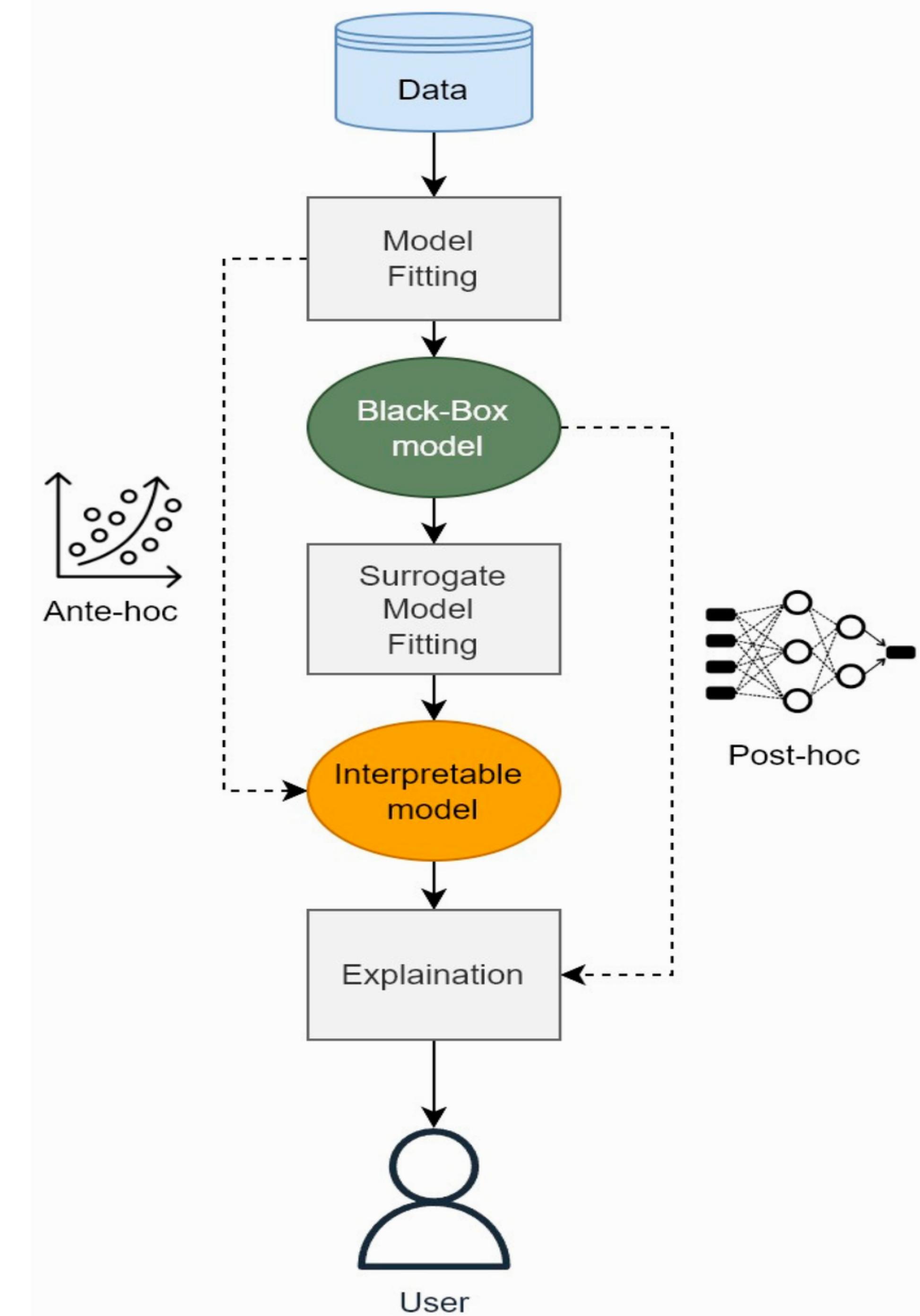
Spurious correlations cause misalignment



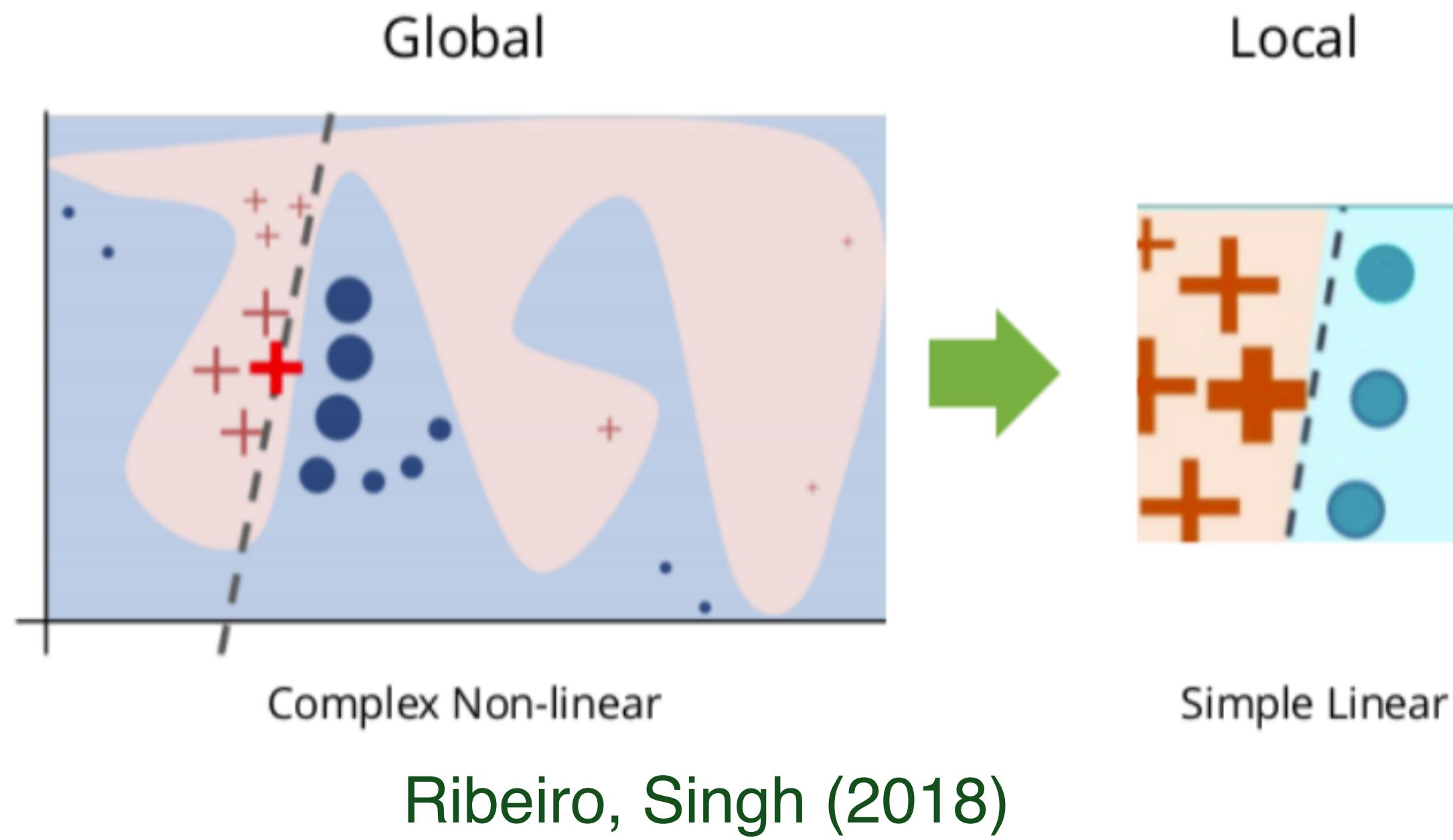
More about interpretability



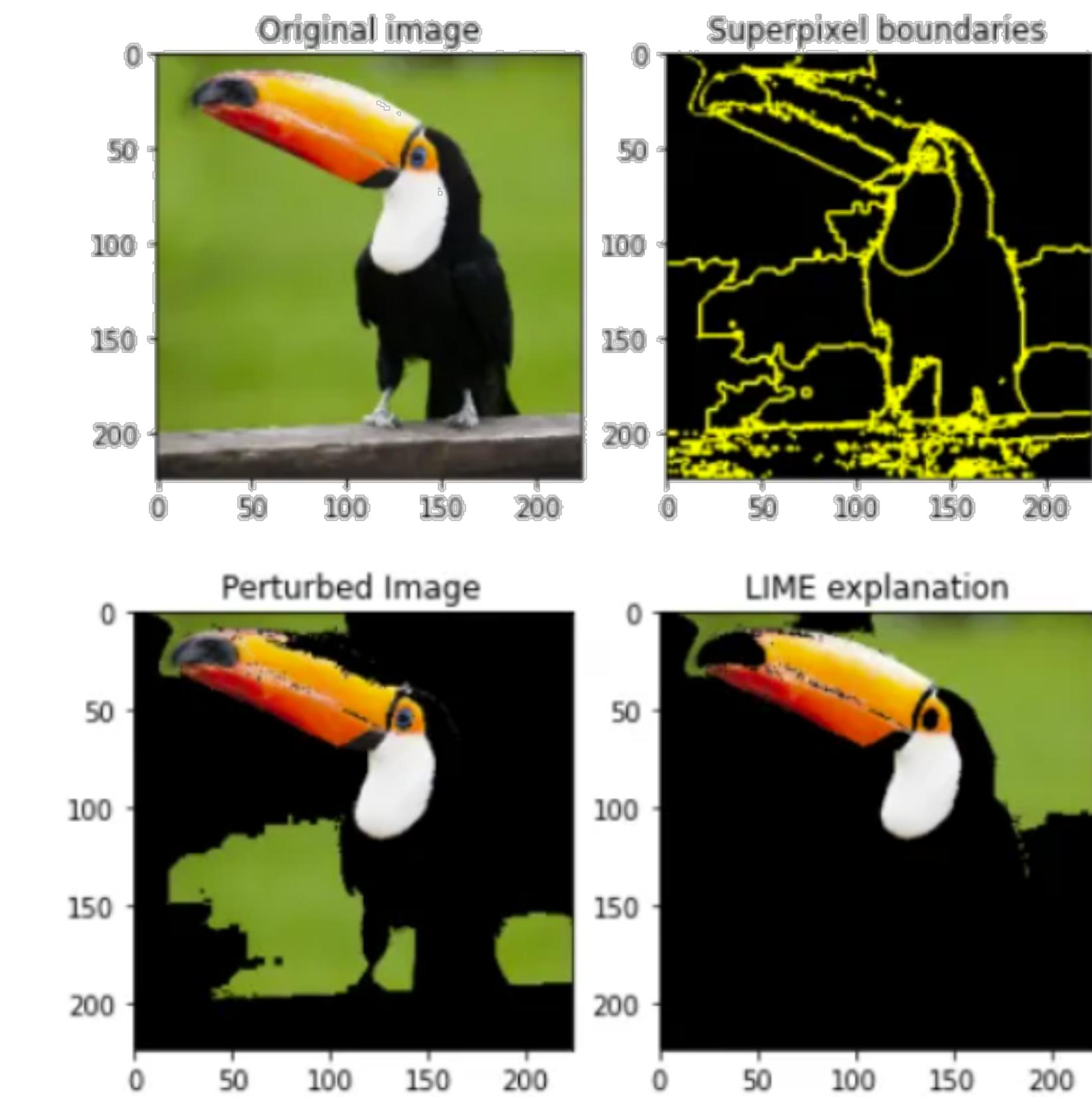
Ante-hoc VS Post-hoc interpretability



LIME (Local Interpretability Model agnostic Explanations)



Source: Explainable AI, <https://bigdatarepublic.nl/>

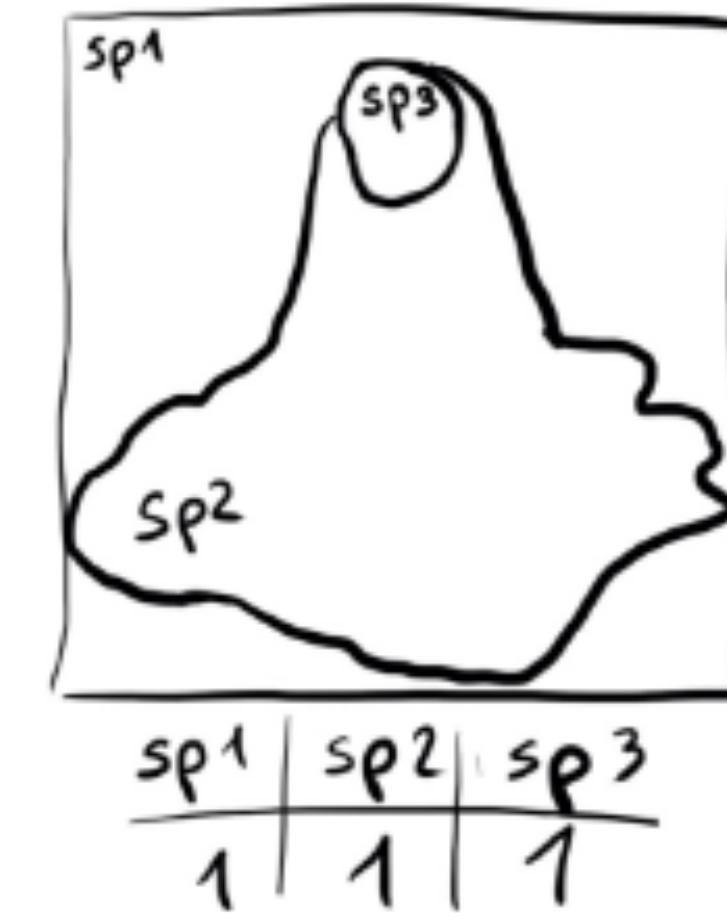


Multiple plots demonstrating a) the original input image, b) the boundaries of all superpixel areas, c) a sample of a perturbed image, and d) the most important super pixel areas to predict the 'toucan' class (Image by Author)

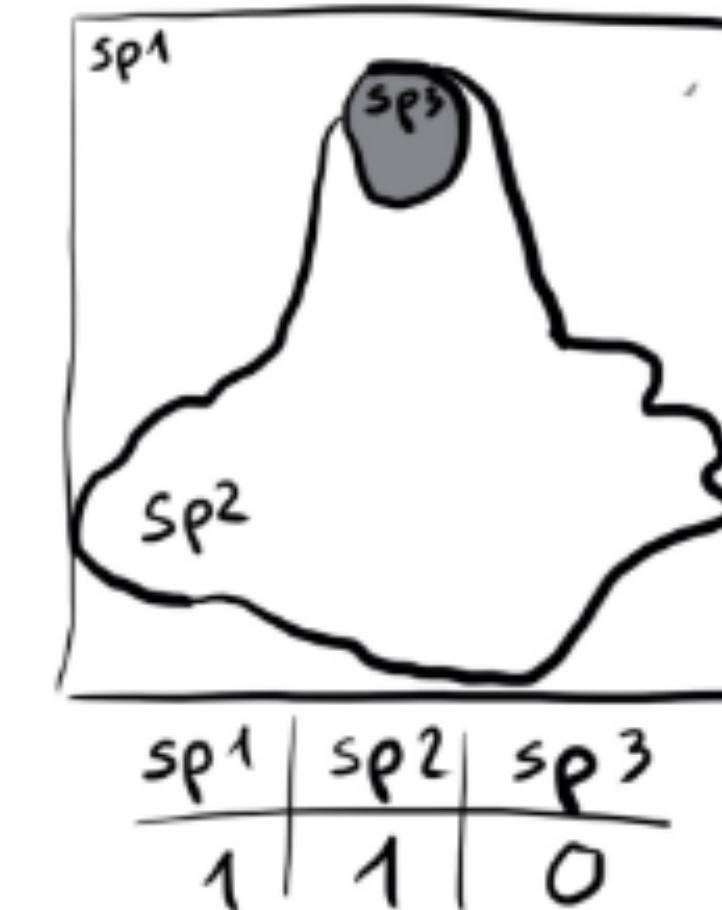
ML interpretability: SHAP (Global)

- Based on collaborative game theory.
- Assigns Shapley values to each input element.
- Shapley values: **introducing each element one by one.**
- **Compared to an average/typical input, which features made this prediction larger or smaller?**

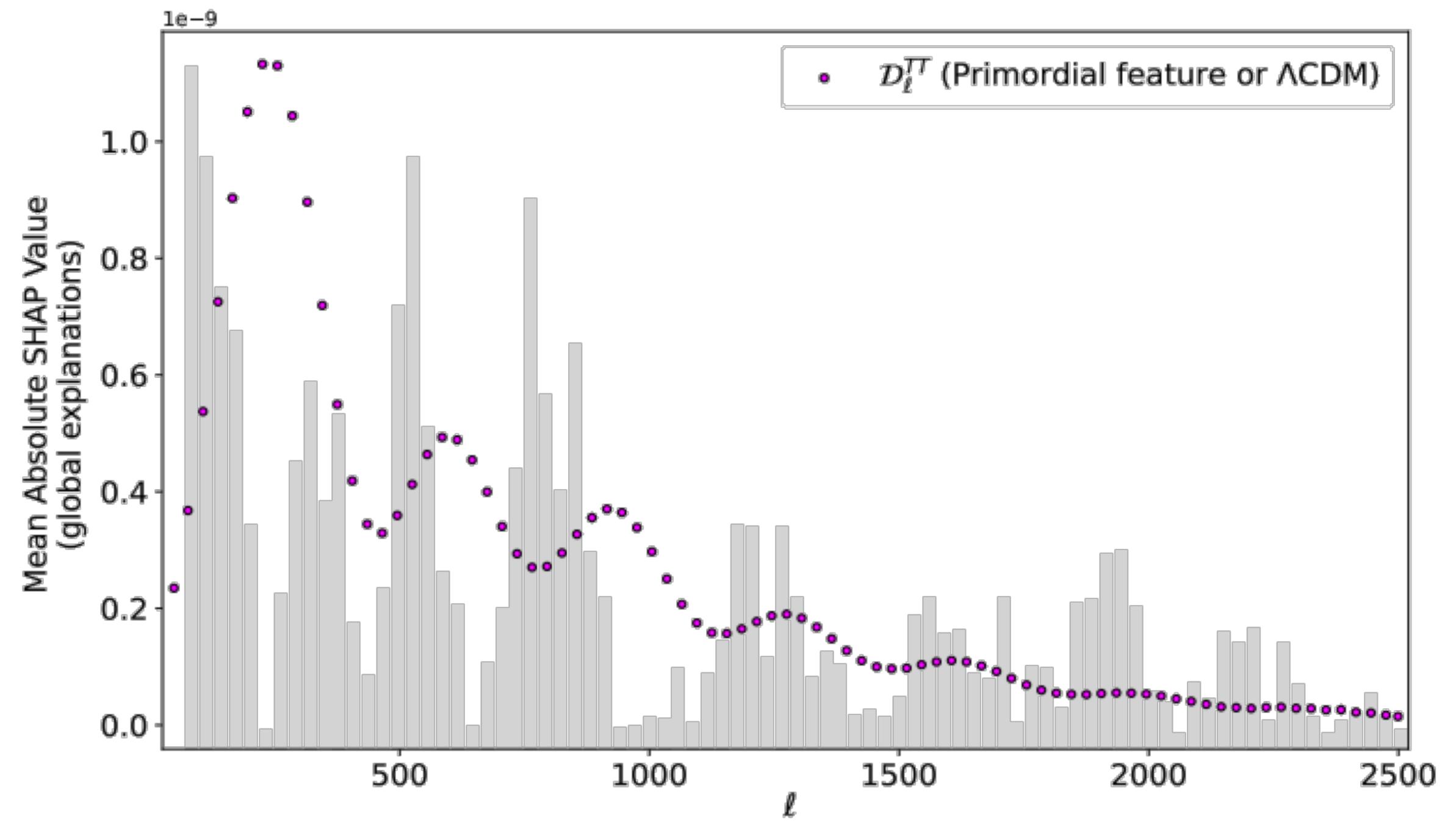
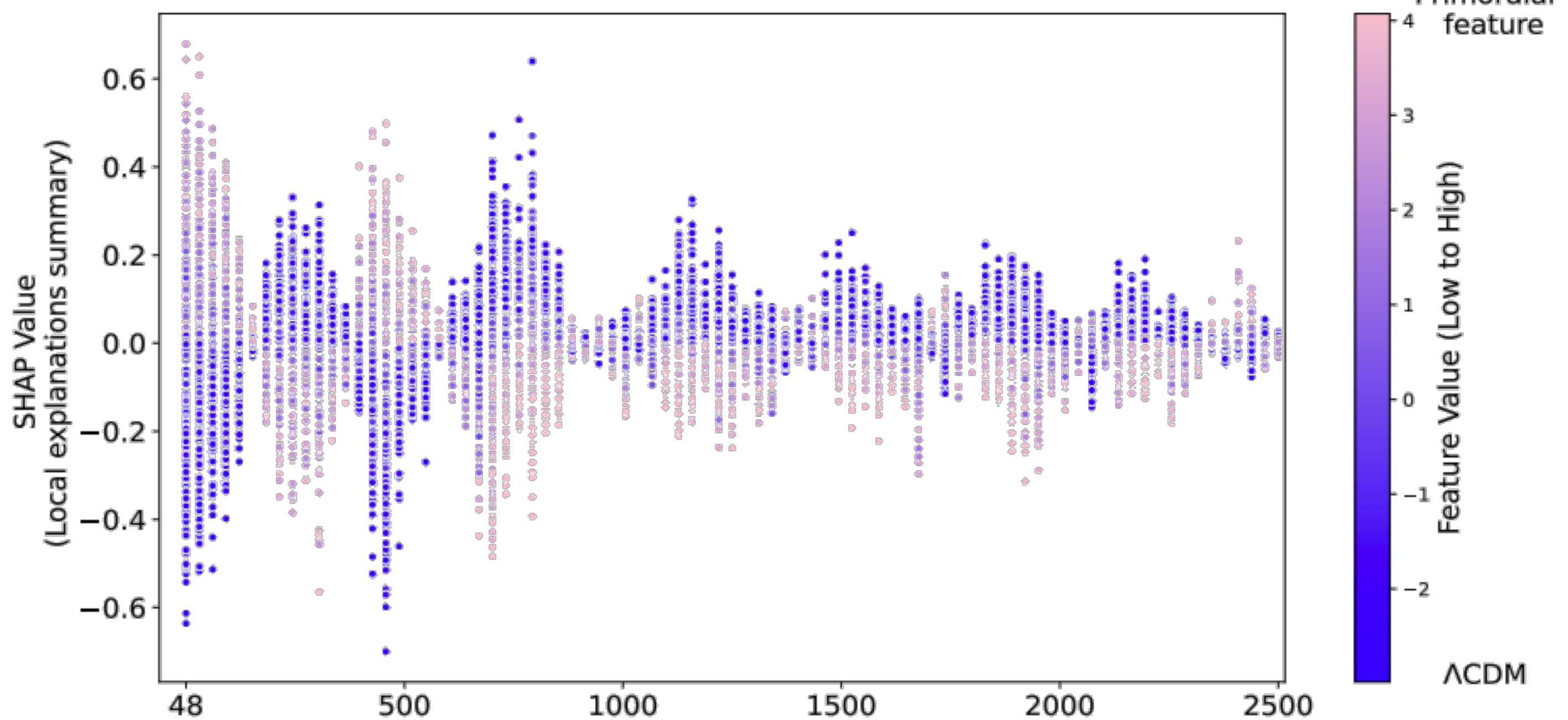
Instance X



Instance X
with absent
features



Match outputs with observables



Gracias!

ML interpretability: SHAP

Step-by-Step Example

Let's say we have a simple model, and we want to compute the SHAP value for x_1 when $x=(1,1,1)$.

We evaluate the model for different subsets:

- **No features used:** $f(\emptyset) =$
- **Only x_1 used:** $f(x_1) = 5$
- **Only x_2 used:** $f(x_2) = 3(1) = 3$
- **Only x_3 used:** $f(x_3) = 2(1) = 2$
- **Using x_1 and x_2 :** $f(x_1, x_2) = 5(1) + 3(1) = 8$
- **Using x_1 and x_3 :** $f(x_1, x_3) = 5(1) + 2(1) = 7$
- **Using all features:** $f(x_1, x_2, x_3) = 5(1) + 3(1) + 2(1) = 10$

We compute contributions across all subsets:

$$f(x_1, x_2, x_3) = 5x_1 + 3x_2 + 2x_3$$

$$\begin{aligned}\phi_1 &= \frac{1}{3}((8 - 3) + (7 - 2) + (10 - 3)) \\ &= \frac{1}{3}(5 + 5 + 7) = \frac{17}{3} \approx 5.67\end{aligned}$$

This means **x_1 contributes 5.67 to the final prediction**, slightly more than its raw weight (5) due to interactions with other features.