Bocconi
AI&Neuroscience
Student
Association

AI in MEDICINE.

Cancer diagnosis using nuclei from breast masses

Omer, Samuele, Velina

INTRODUCTION

INTRODUCTION

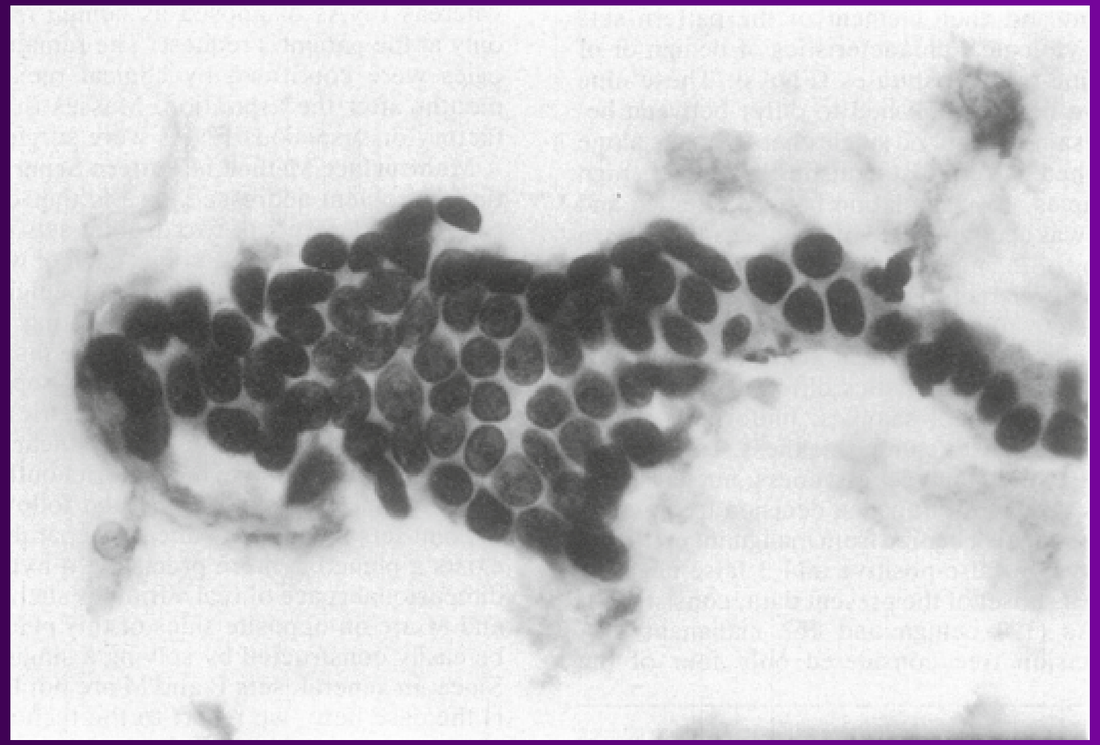
- improve diagnosis
- predict patient outcomes
- insights from comments from clinical incident reports, social media activity, doctor performance feedback, and patient reports after successful cancer treatments

INTRODUCTION

- **Psychosis** (2015) - **automated speech analysis** predicted transition to psychosis with perfect accuracy
- **Skin Cancer** (2017) - **dermatologist-level accuracy**
- **Diabetes** (2016) - predict the progression of **pre-diabetes** to **type 2 diabetes**
- **Breast Cancer** (2019)

DATASET

- Breast Cancer Wisconsin Diagnostic Data Set
- **Fine-needle aspiration** (FNA), a common diagnostic procedure in oncology
- January 1989 to November 1991



An example of an image of a breast mass

DATASET

		Features									Outcome
Instance No.	Sample I.D.	Thickness	Cell Size	Cell Shape	Adhesion	Epithelial Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class (Diagnosis)
1	1000025	5	1	1	1	2	1	3	1	1	2
2	1002945	5	4	4	5	7	10	3	2	1	2
3	1015425	3	1	1	1	2	2	3	1	1	2
.											
.											
.											
699	897471	4	8	8	5	4	5	10	4	1	4

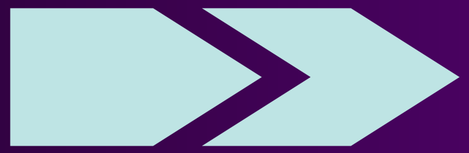
01

Generalised Linear Model (GLMs)

Logistic Regression

- **supervised** machine learning algorithm used for binary classification tasks

linear combination of input features

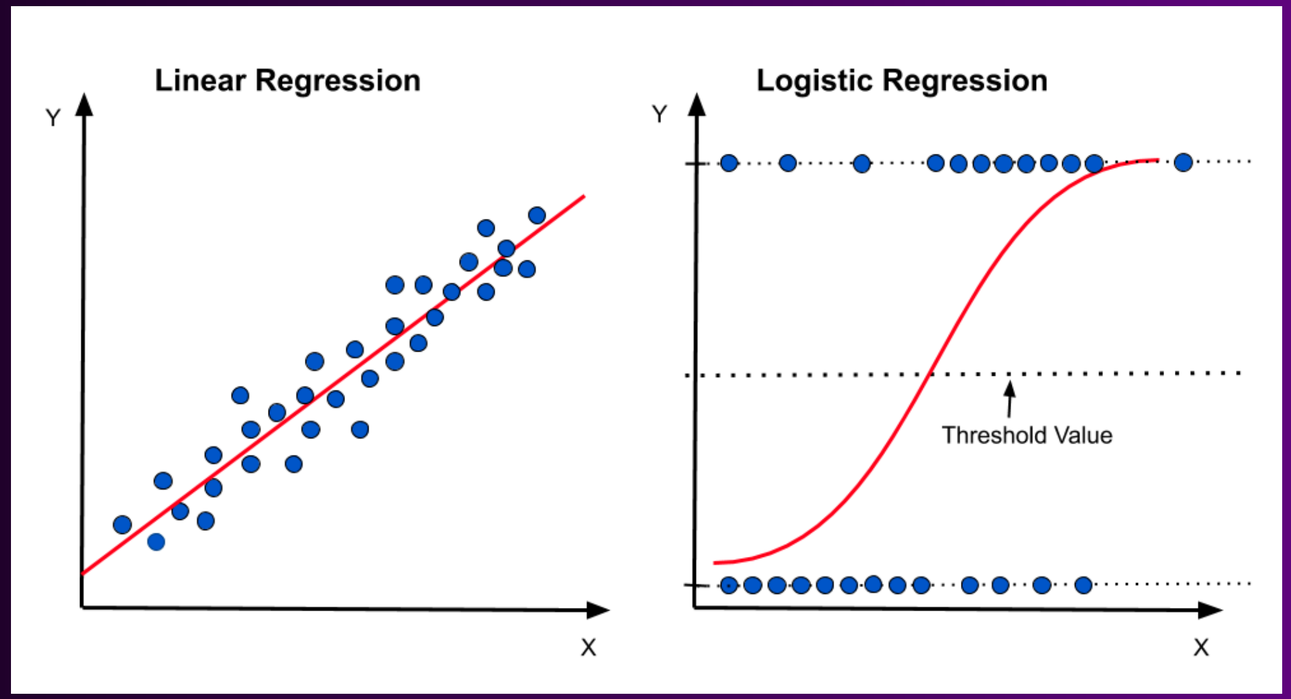


probability value between 0 and 1

01

Generalised Linear Model (GLMs)

Logistic Regression

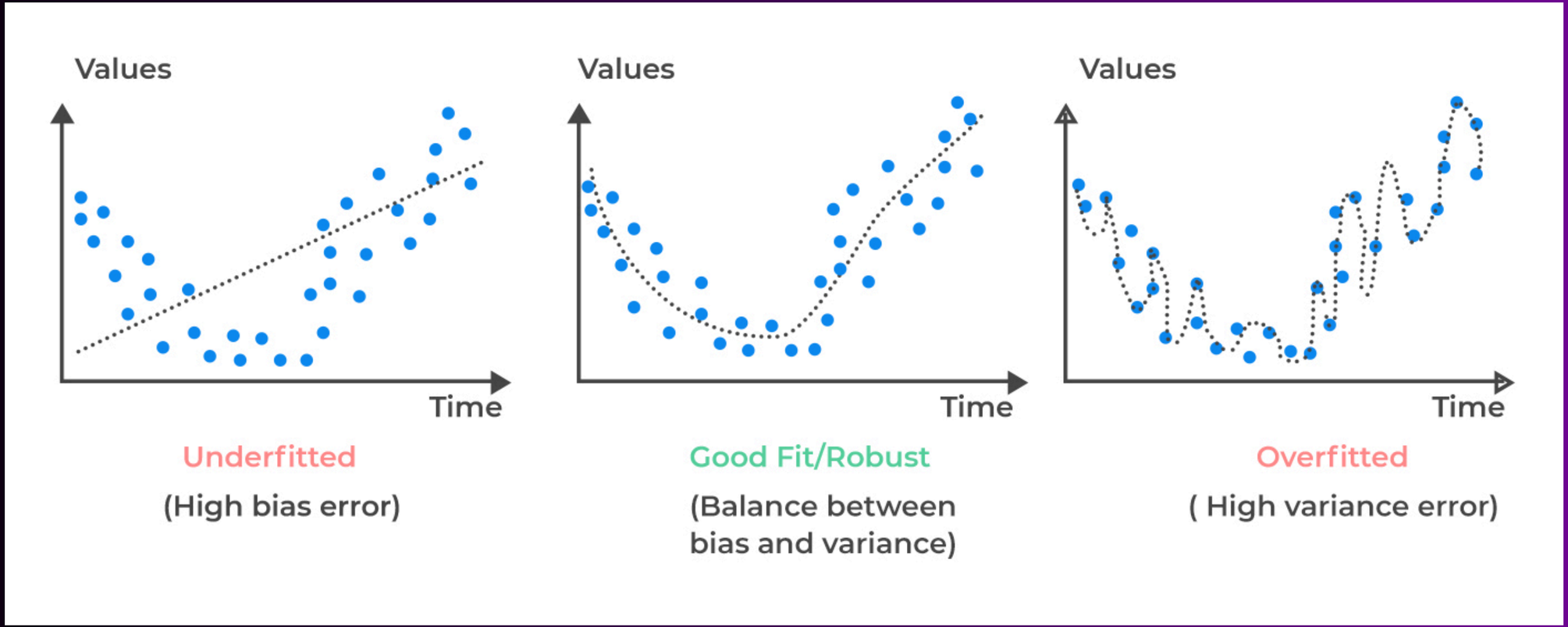


Regularisation of Dataset

LASSO vs RIDGE

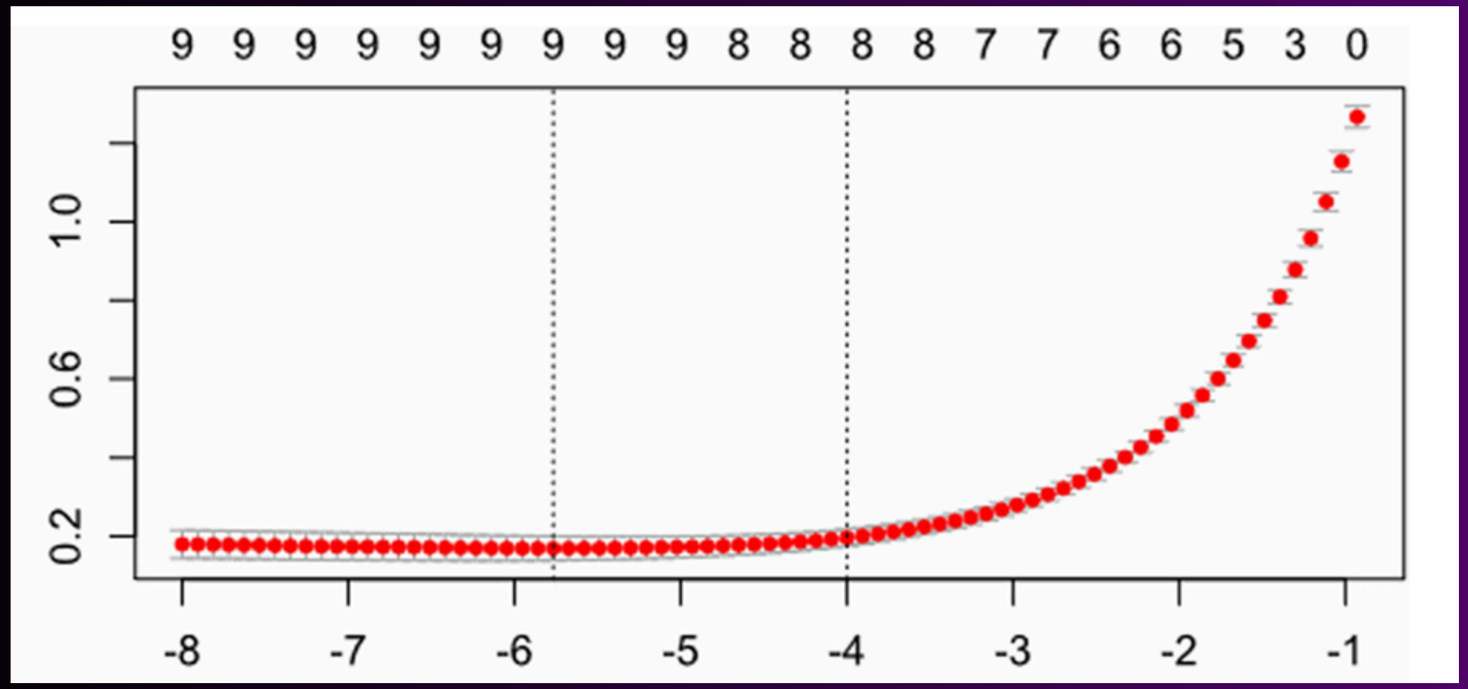
- **overfitting**
- **multicollinearity**

We need dataset with the right ratio between **DATA** and **FEATURES**



Cross Validation Curve

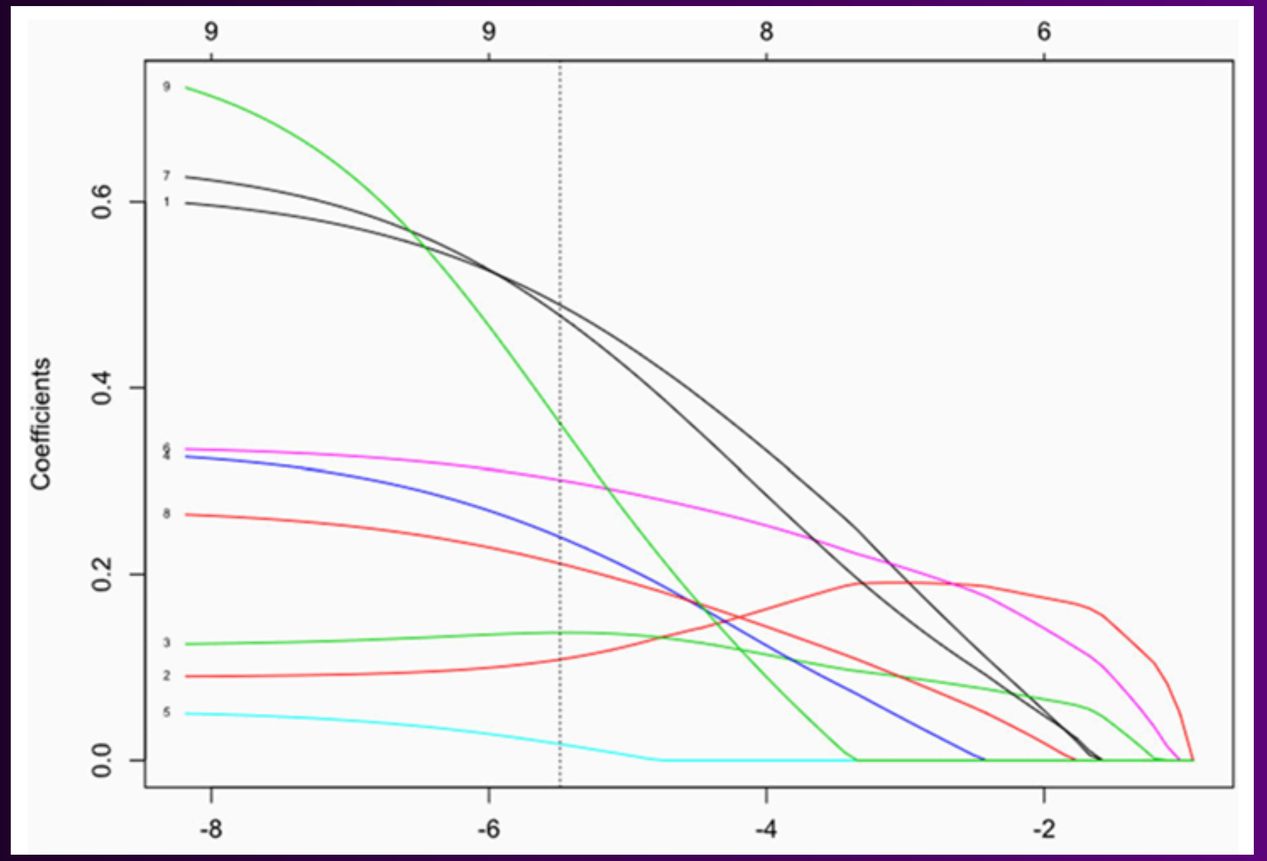
- the relation between the **STANDARD DEVIANCE** and the **LAMBDA-COEFFICIENT**




```
library(glmnet)

glm_model =
cv.glmnet(x_train, y_train,
alpha = 1, nfolds = 10)

lambda.min =
glm_model$lambda.min

glm_coef =
round(coef(glm_model,
s = lambda.min), 2)
```

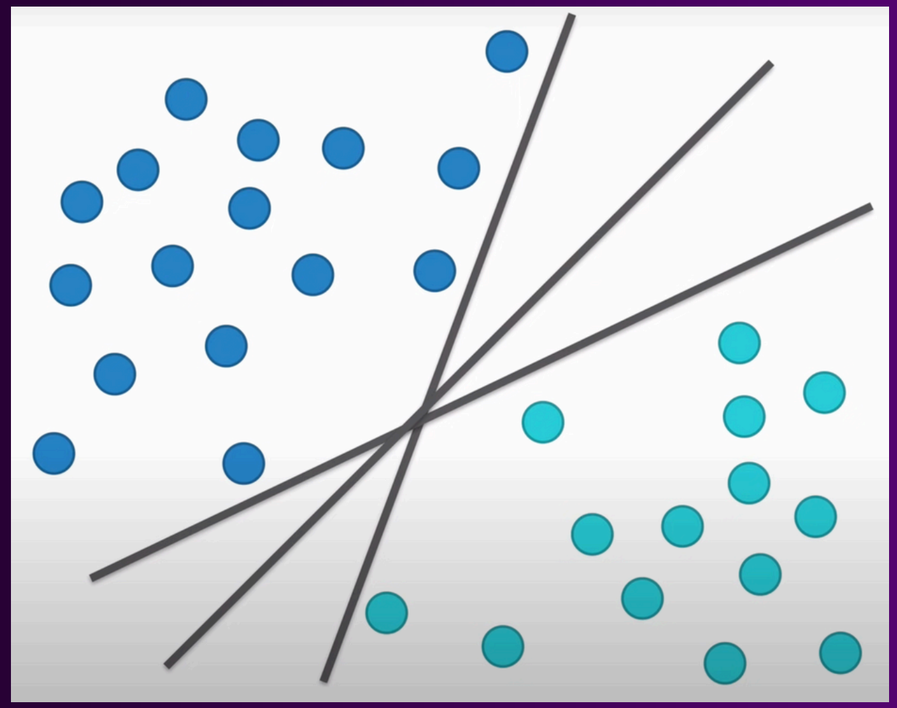
Regression Coefficient Curve

- the **QUANTITY** and the **MAGNITUDE** of each feature in relation to the **LAMBDA-COEFFICIENT**

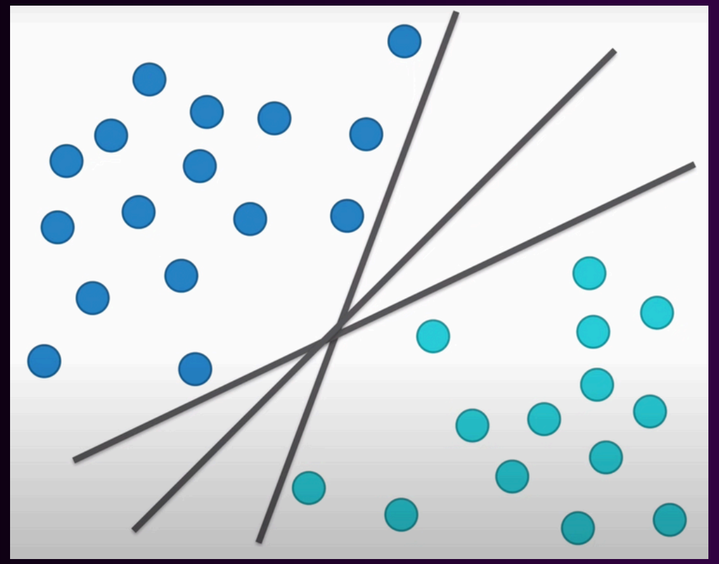


02

Support Vector Machine (SVMs)



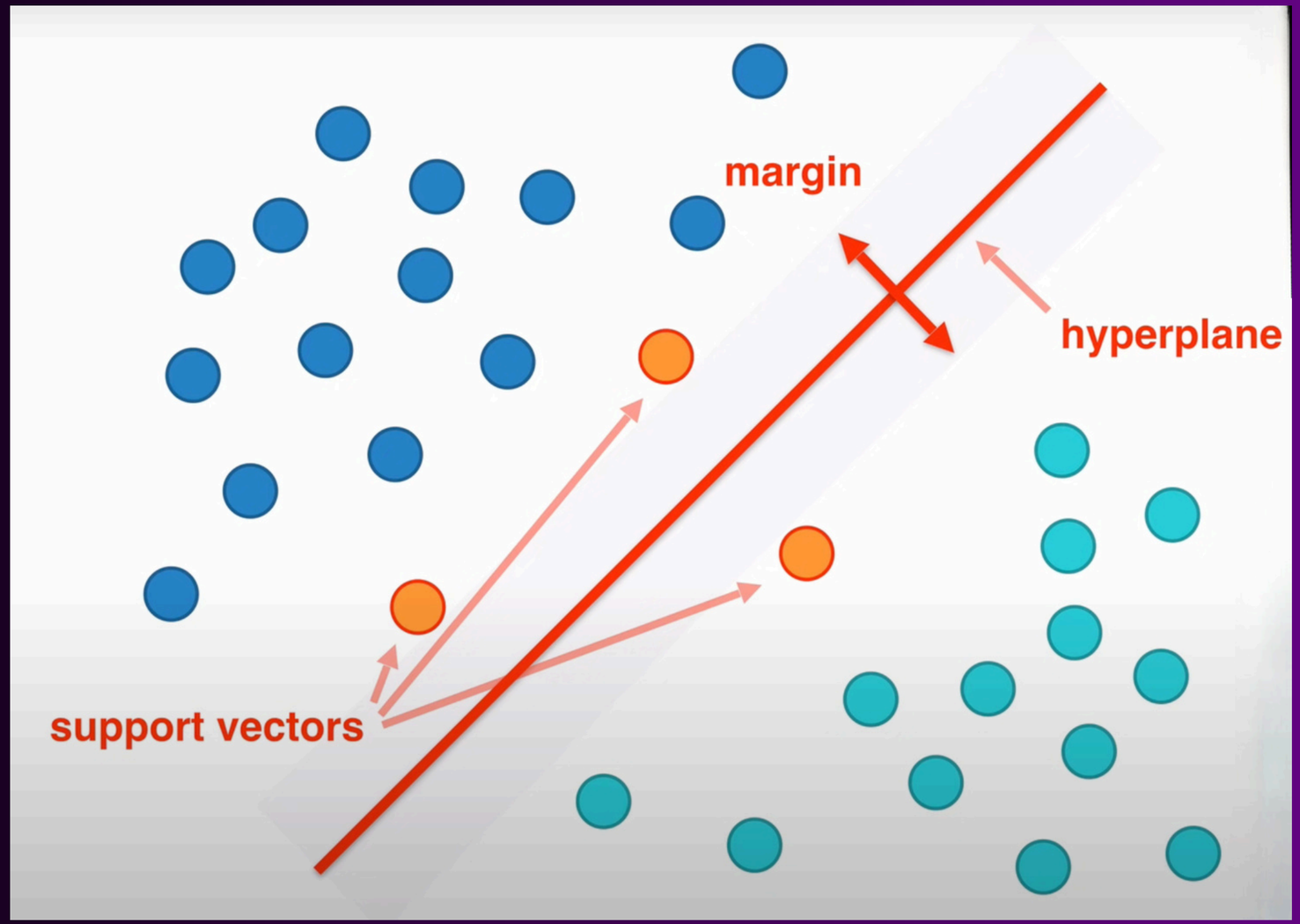
Which Line Should I Choose?



SVMs distinguish between two classes by finding the **optimal line (hyperplane).**

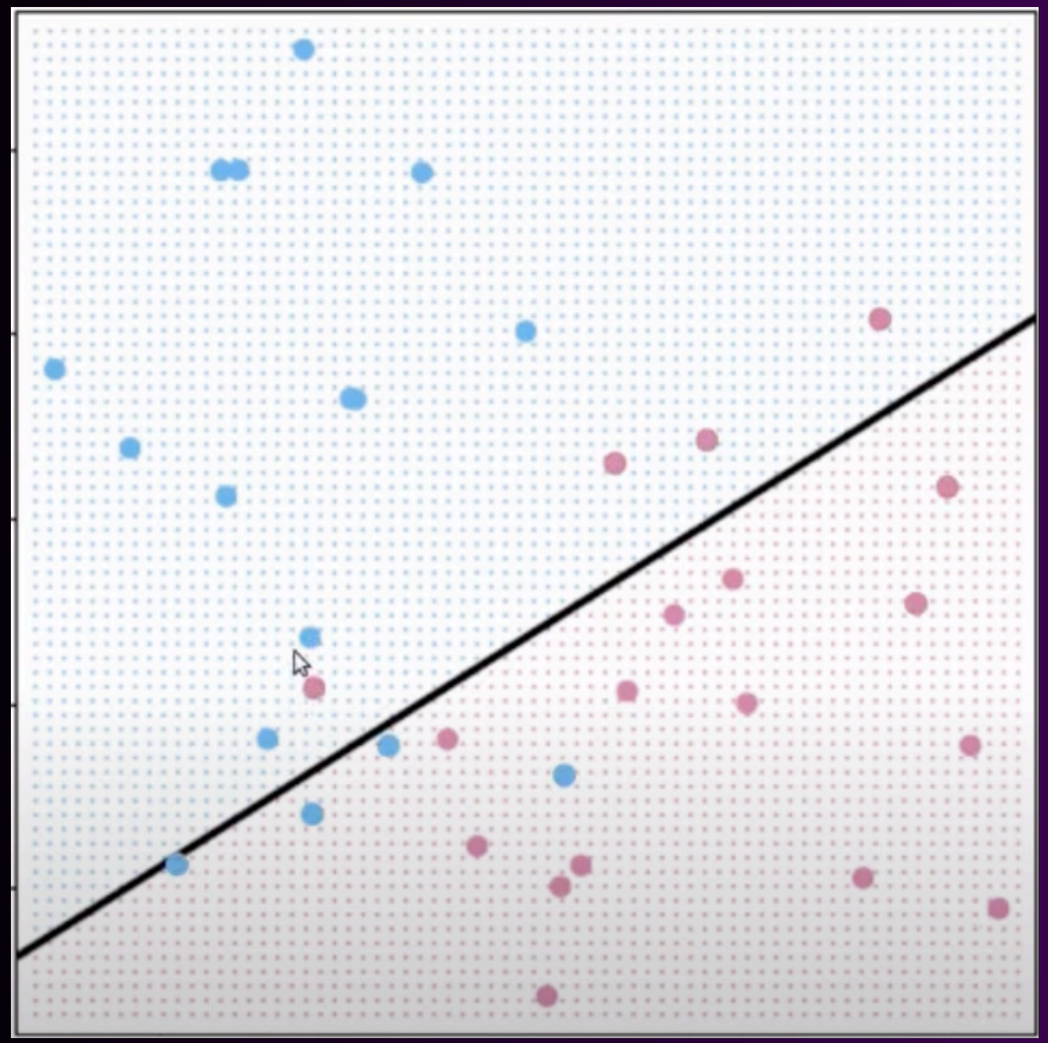
This line **maximizes the distance** between the **closest data points** of opposite classes.

The number of features in the input data determine if the hyperplane is a line in a 2-D space or a plane in a n-dimensional space.



BUT IS LIFE ALWAYS THIS EASY?

The C parameter



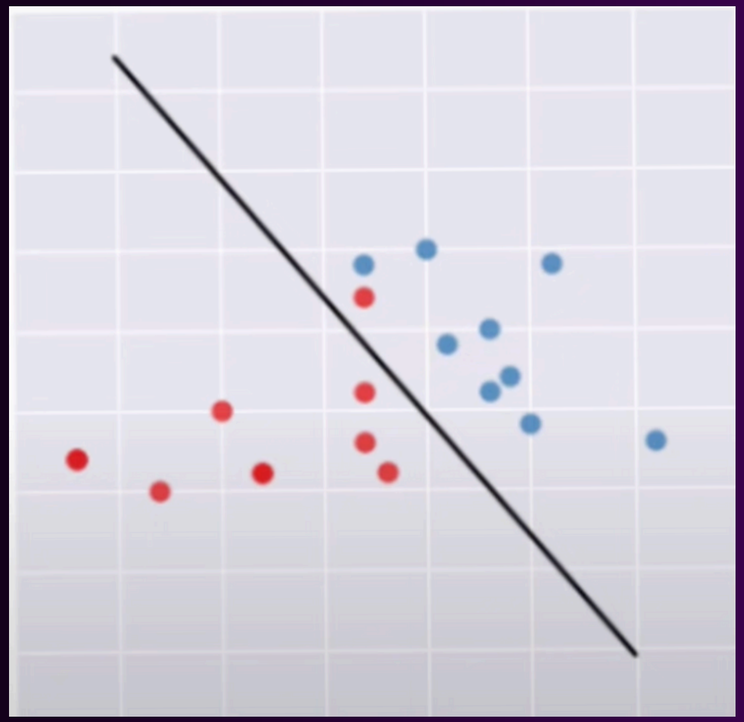
Sometimes you will have some data that is not perfectly seperated, and you will allow for some “little” **misclassifications** in the short run for a higher accuracy in the long run.

At those times you will make a decision:

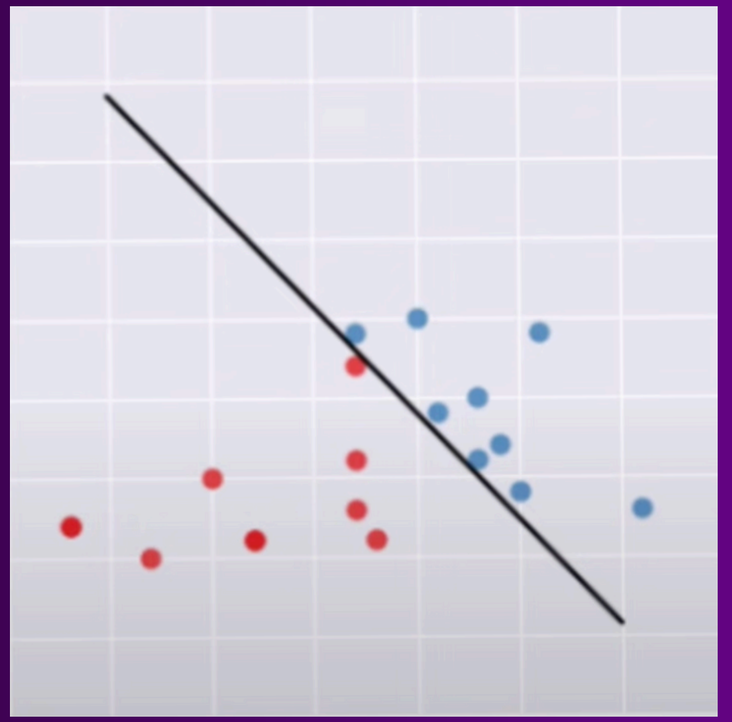
You will choose a **C parameter**, which allows you to decide how much you want to penalize each misclassification.

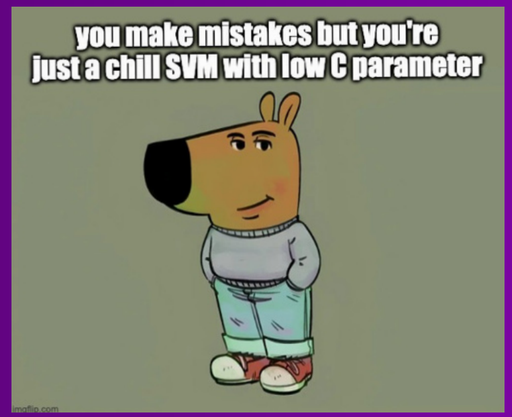
We can use cross validation to determine the optimal C

Low C parameter (Soft Margin)



High C parameter (Hard Margin)

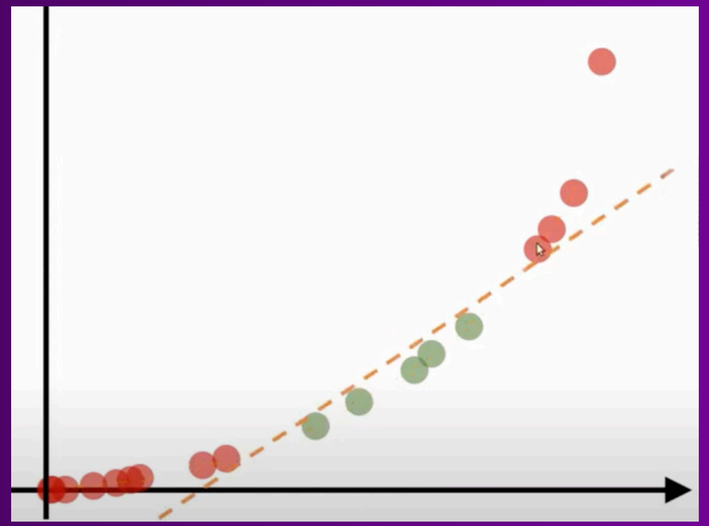


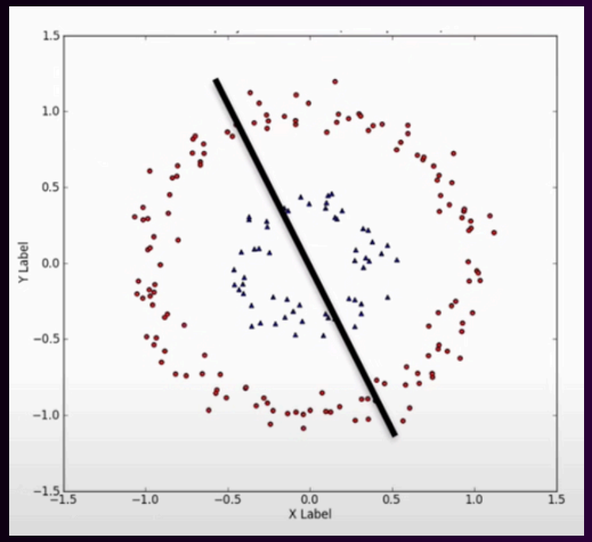


Kernels and Kernel Trick

The kernel function is just a mathematical function that converts a low-dimensional input space into a higher-dimensional space

Kernel

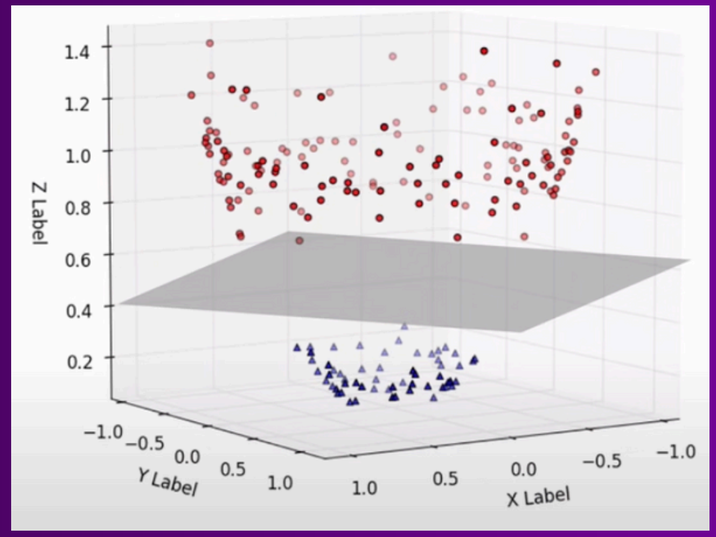




Low Dimensions

Kernel Options

- Linear
- RBF
- Polynomial
- Sigmoid



High Dimensions

RBF Kernel and Gamma

CLOSE points get a **STRONG** signal -> they're "**similar**"
FAR points get a **WEAK** signal -> they're "**not similar**"

$$k(x, y) = e^{(-\gamma \|x-y\|^2)}$$

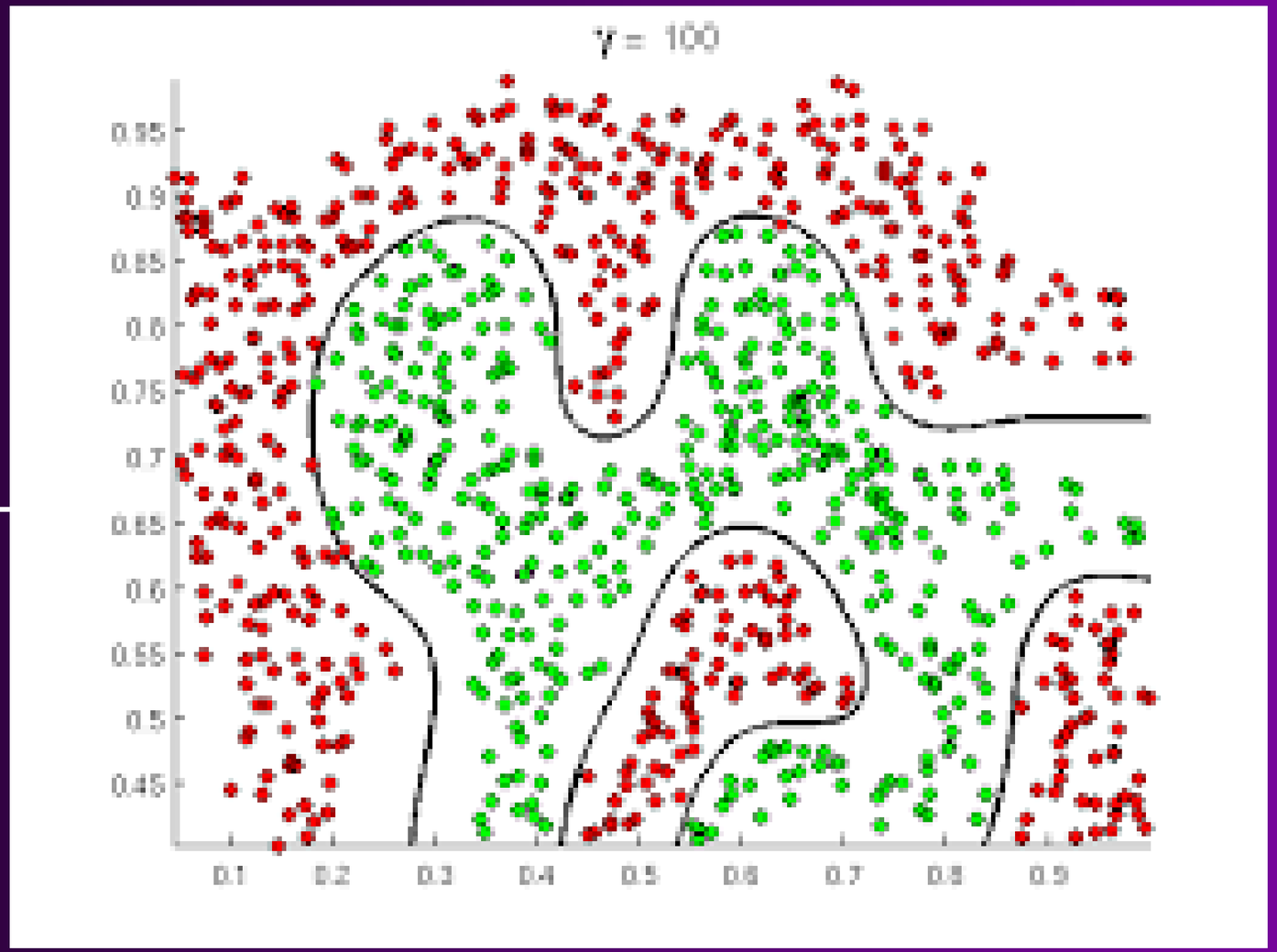
The RBF kernel calculates this similarity using an **exponential function**. It ensures the signal **weakens rapidly** with distance, creating localized zones of influence around each point.

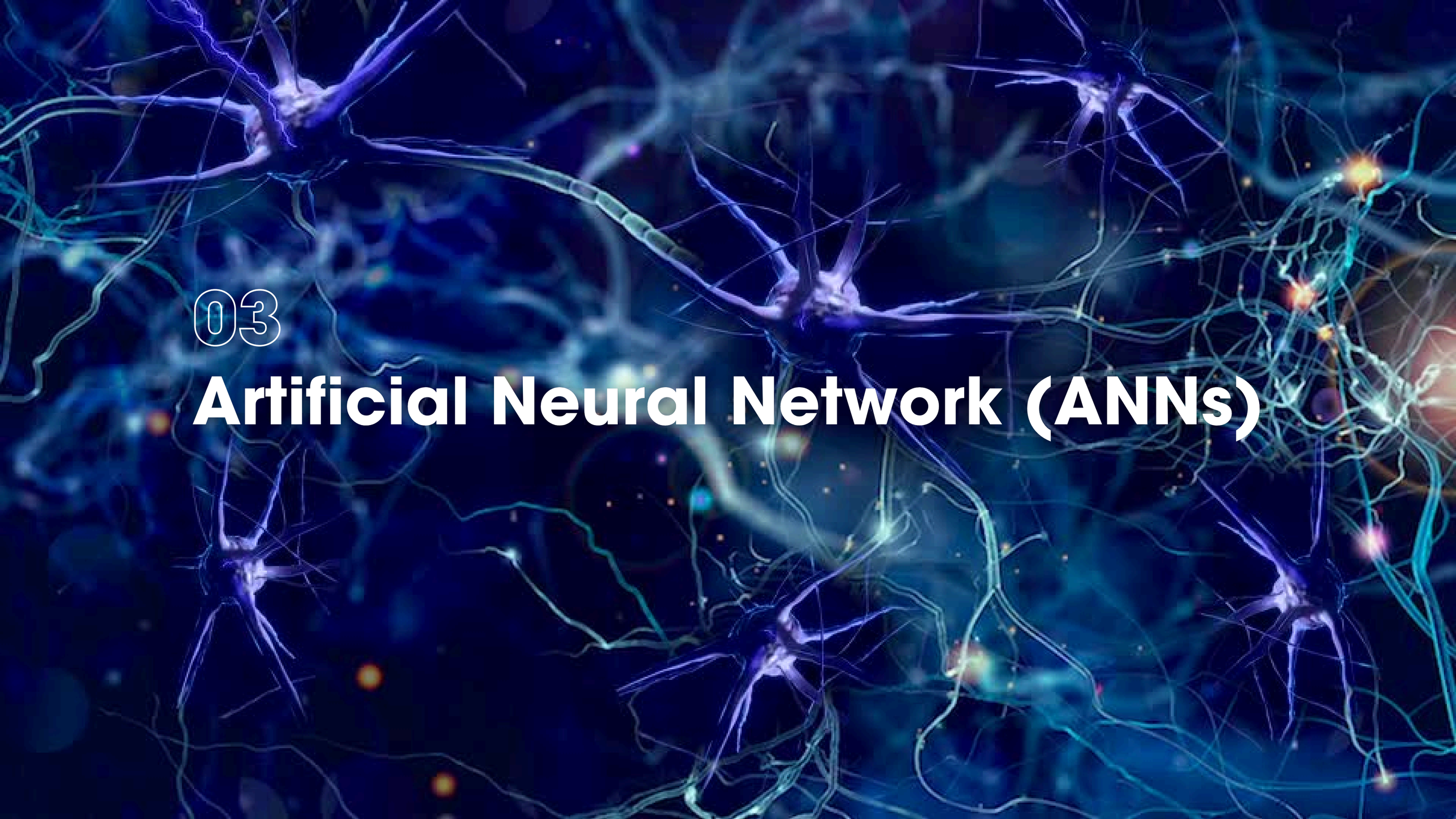
HIGH GAMMA(Y)
points only "see" their very close neighbors

The decision boundary becomes very wiggly and tightly hugs the training data, possibly overfitting.

LOW GAMMA(Y)
points can "talk" to faraway neighbors

The decision boundary is smoother, but it might miss capturing small patterns.



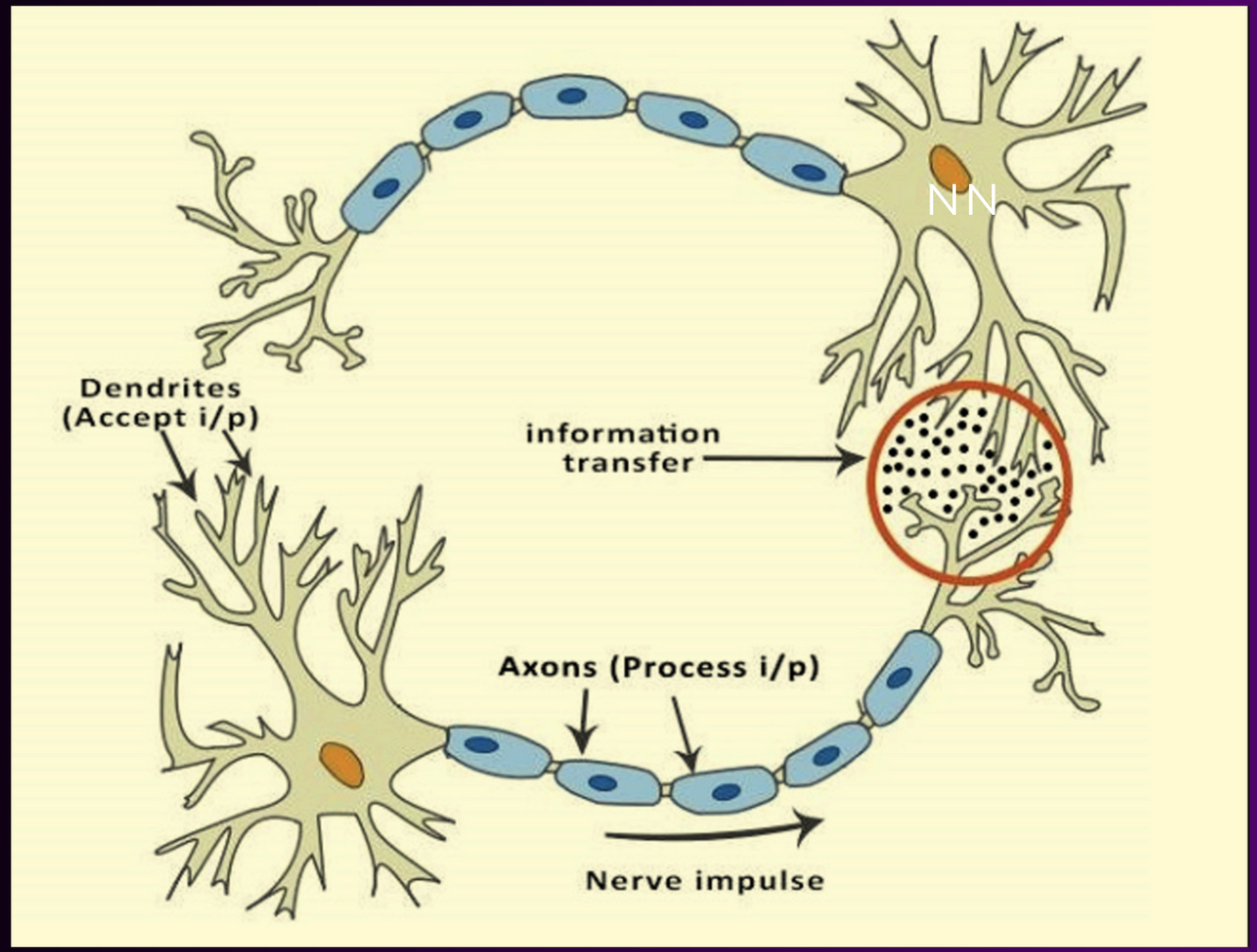


03

Artificial Neural Network (ANNs)

Structure of ANNs

Structure of ANNs:

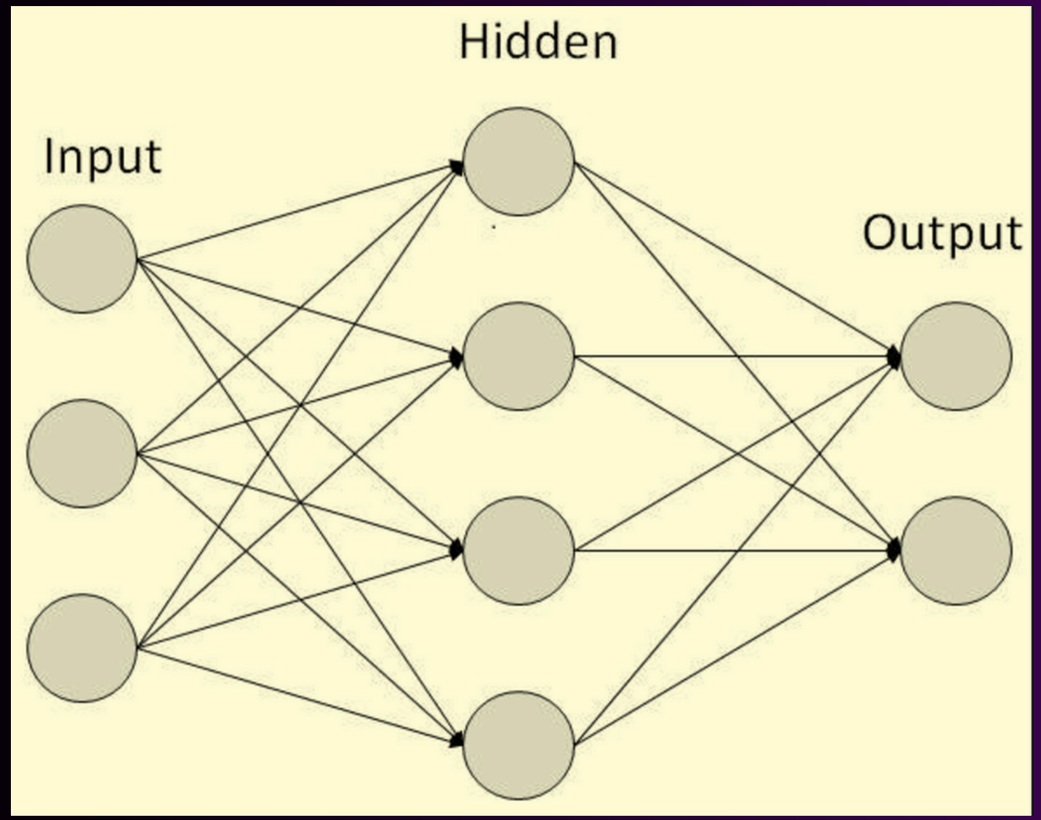


- **Dendrites**, which receive signals from other neurons
- **Cell Body**, which processes these signals
- **Axon**

- **HEBBIAN LEARNING**

"CELLS THAT FIRE TOGETHER, WIRE TOGETHER."

Structure of ANNs:



1.INPUT LAYER
- the input data fed into the network, with each neuron corresponding to one feature or variable in the dataset

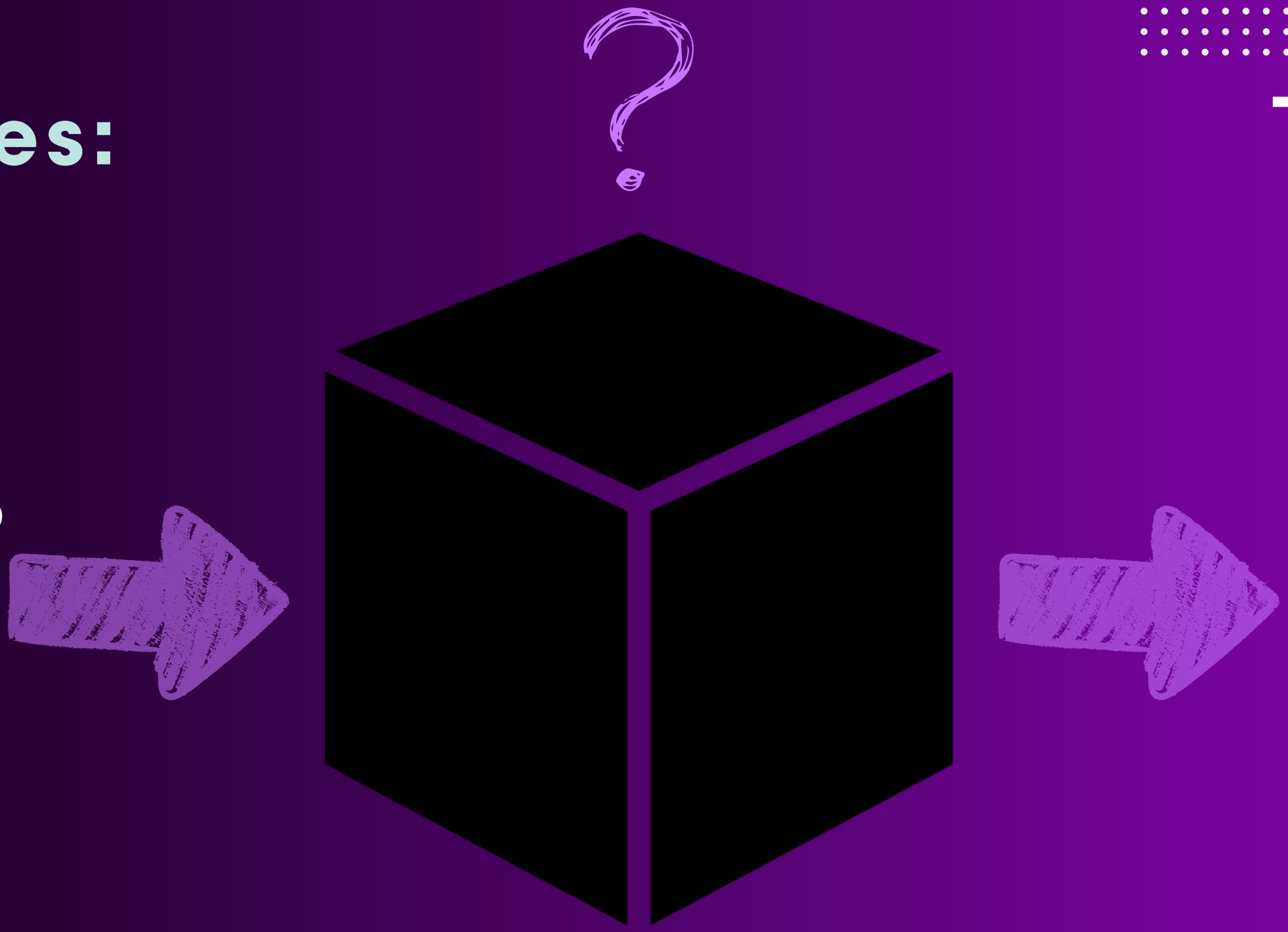
2.HIDDEN LAYERS
- perform intermediate computations

3.OUTPUT LAYER
- final predictions or classifications

- each connection between nodes in these layers is assigned a **weight:** how strongly one node influences the next
- each node has a **bias:** threshold value that determines if a signal should pass through

Challenges:

- Overfitting
- Computational Requirements
- Interpretability (“**Black boxes**”)



Challenges:

COMPLEMENT

~~replace~~ human expertise

Performance Metrics - confusionMatrix

			Actual outcome				
			Benign (0)	Malignant (1)	Sensitivity	Specificity	Accuracy
Predicted Outcomes	GLM	Benign (0)	148	10	.99	.87	.95
		Malignant (1)	2	67			
	SVM	Benign (0)	146	5	.97	.94	.96
		Malignant (1)	4	72			
	ANN	Benign (0)	148	11	.99	.86	.94
		Malignant (1)	2	66			

$$\frac{\text{True Positive}}{\text{Actual Positive}}$$

SENSITIVITY

$$\frac{\text{True Negative}}{\text{Actual Negative}}$$

SPECIFICITY

$$\frac{\text{True Positive + True Negative}}{\text{Total Prediction}}$$

ACCURACY

CONCLUSIONS

This data demonstrates an important principle of ML: *More complex algorithms **do not** necessarily get more useful predictions.*

CONCLUSIONS

STRENGHTS

- Accessibility and Familiarity
- Simple, Reusable Code
- Computational Efficiency: the small size of the dataset ensures that computational times are short
- Demystifying Machine Learning for Medical Practitioners
- Demonstration of Effectiveness

CONCLUSIONS

WEAKNESSES

- Limited Data Complexity
- Risk of Oversimplification: the model’s performance could degrade when faced with real-world data that includes noise, higher dimensionality, and unbalanced classes
- Black Box Nature of Some Algorithms
- Ethical and Bias Risks

Thank You

For All Your Attention