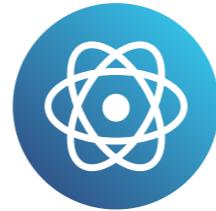


# Supervised learning

MACHINE LEARNING FOR EVERYONE



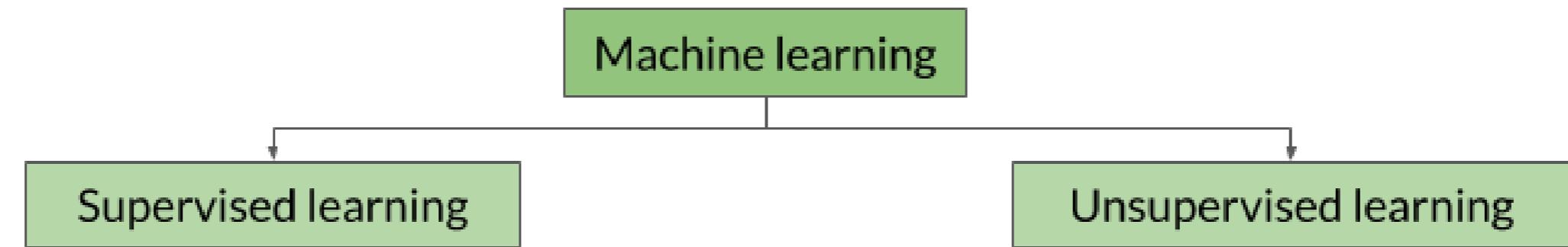
Hadrien Lacroix

Content Developer at DataCamp

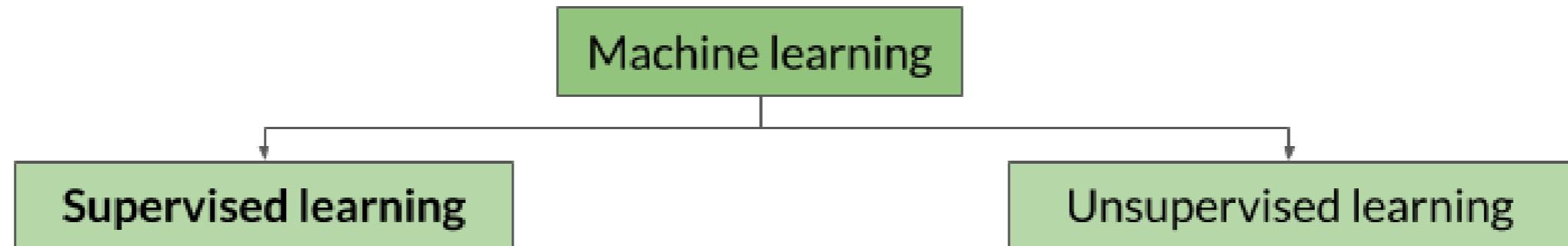
# Modeling



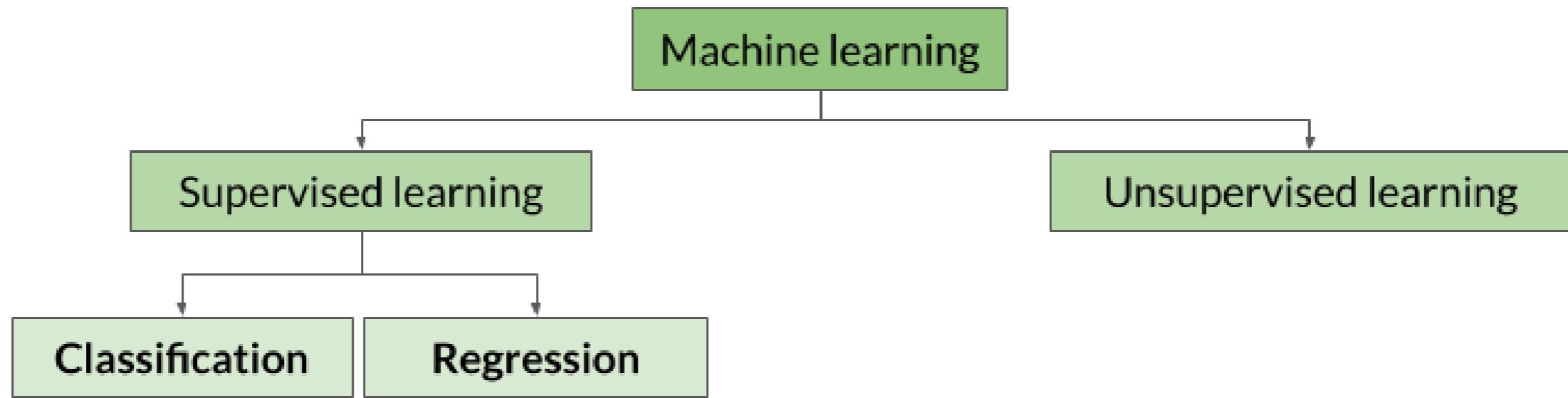
# Types



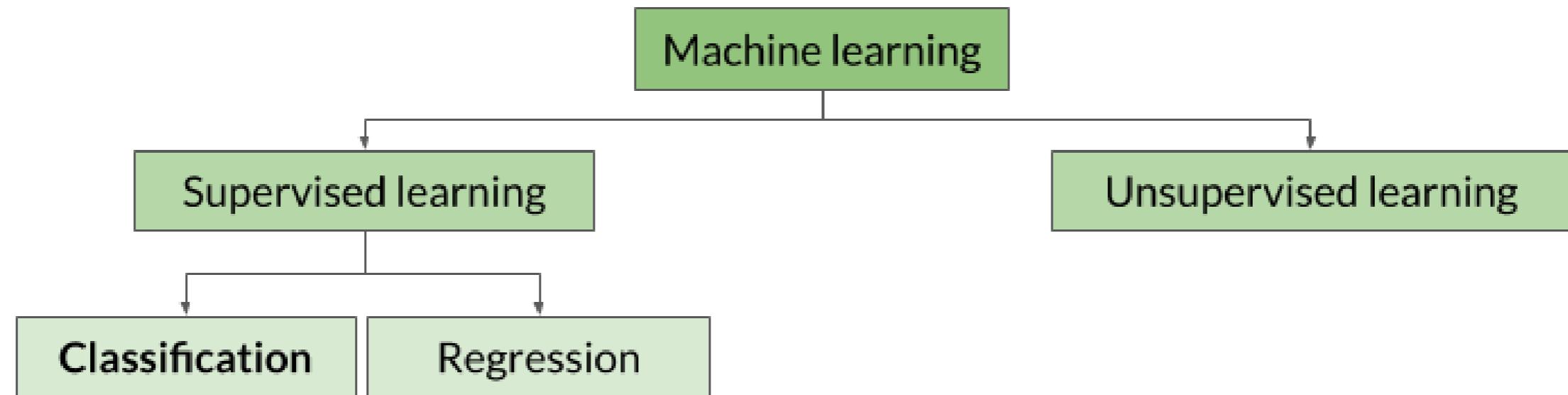
# What is supervised learning?



# Classification and regression



# Classification



# Classification

- **Classification** = assigning a category
  - Will this customer stop its subscription?
    - Yes, No
  - Is this mole cancerous?
    - Yes, No
  - What kind of wine is that?
    - Red, White, Rosé
  - What flower is that?
    - Rose, Tulip, Carnation, Lily

# Observations

	Applicant ID	High school GPA	Test results	Accepted
First observation	0	3.5	2.4	False
Second observation	1	4	2.2	False
Third observation	2	4.2	4.3	True
Fourth observation	3	4.8	2.9	False
n observations	...	...	...	...

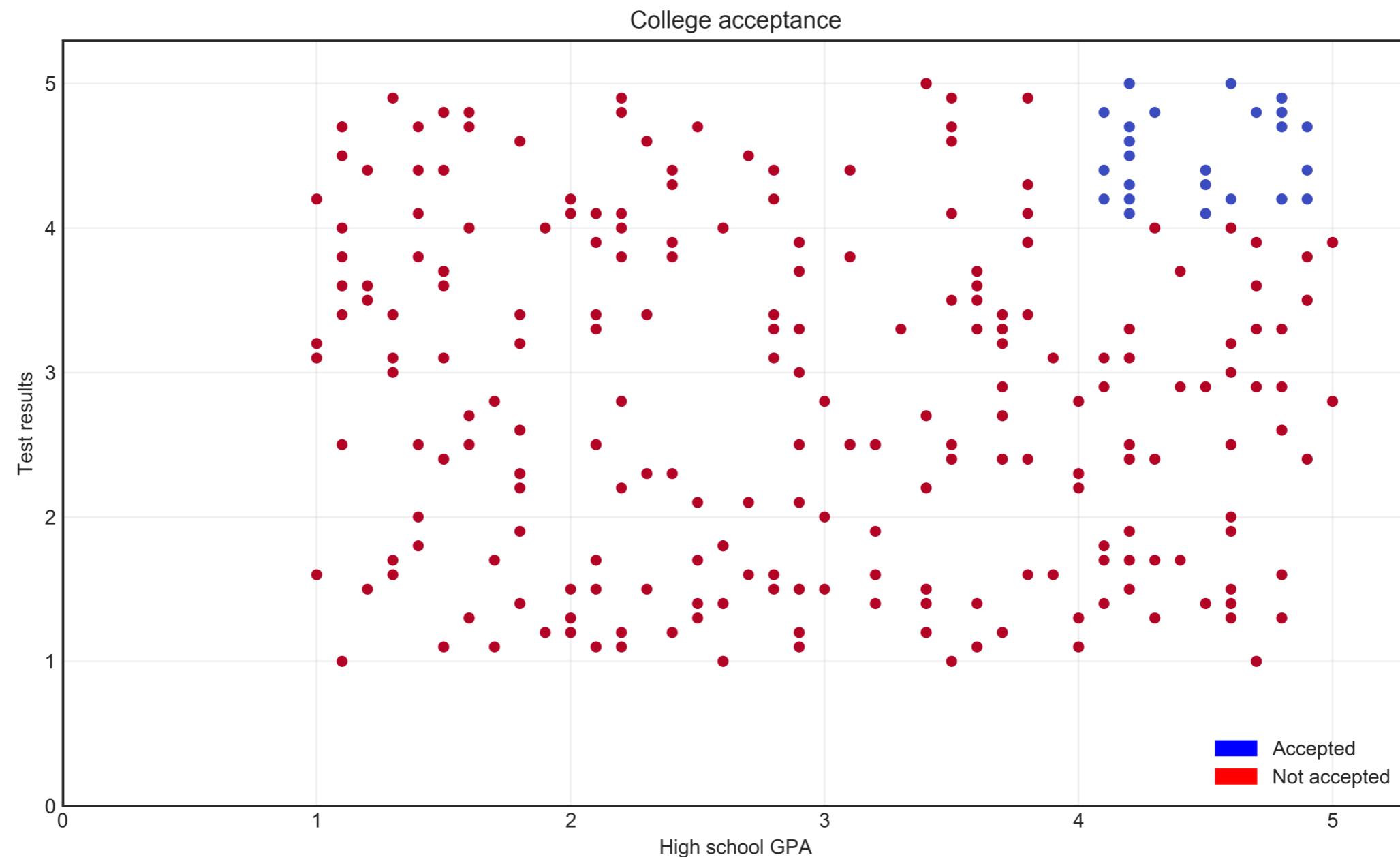
# Features

Applicant ID	High school GPA	Test results	Accepted
0	3.5	2.4	False
1	4	2.2	False
2	4.2	4.3	True
3	4.8	2.9	False
...	...	...	...

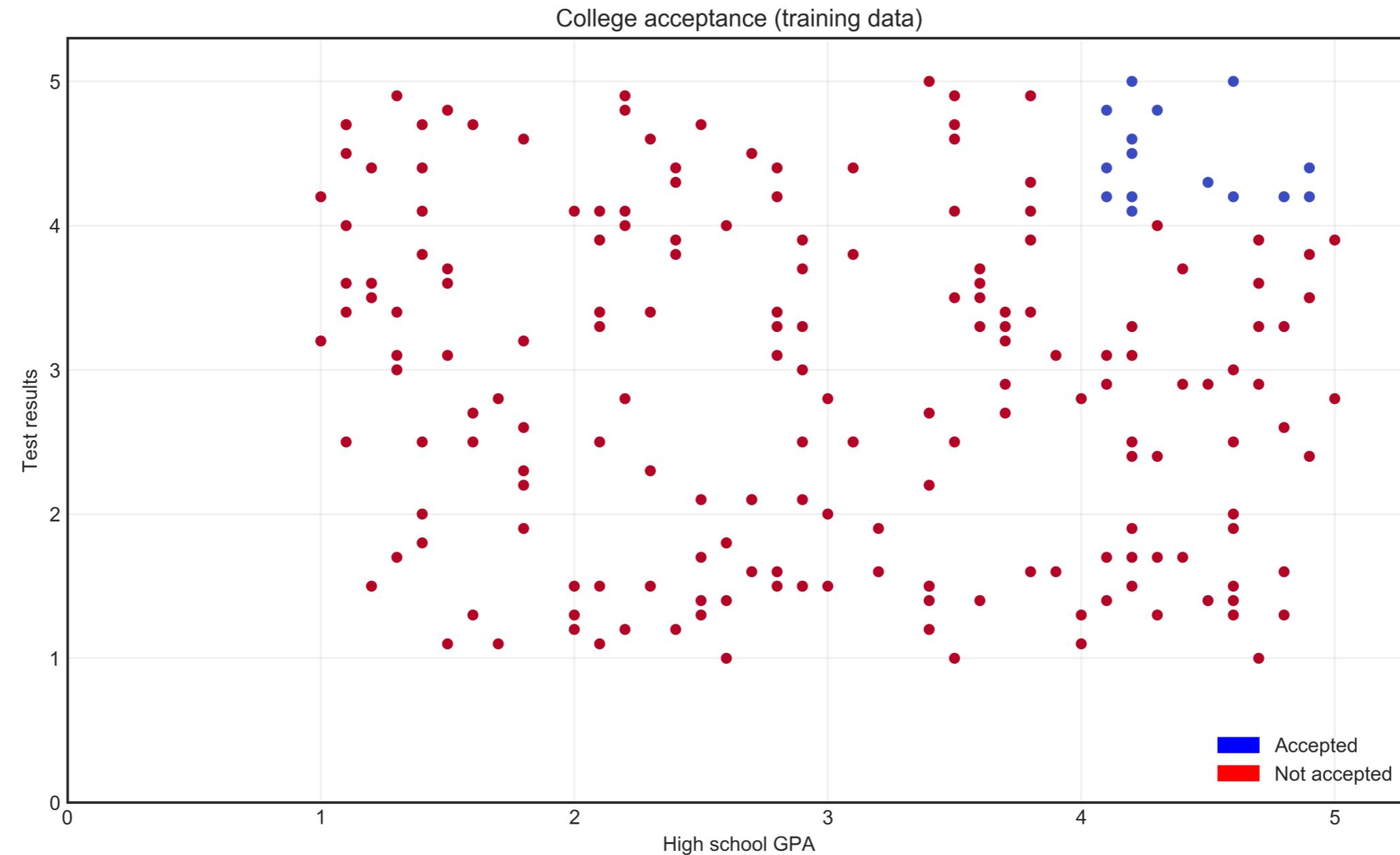
# Target

Applicant ID	High school GPA	Test results	Accepted
0	3.5	2.4	False
1	4	2.2	False
2	4.2	4.3	True
3	4.8	2.9	False
...	...	...	...

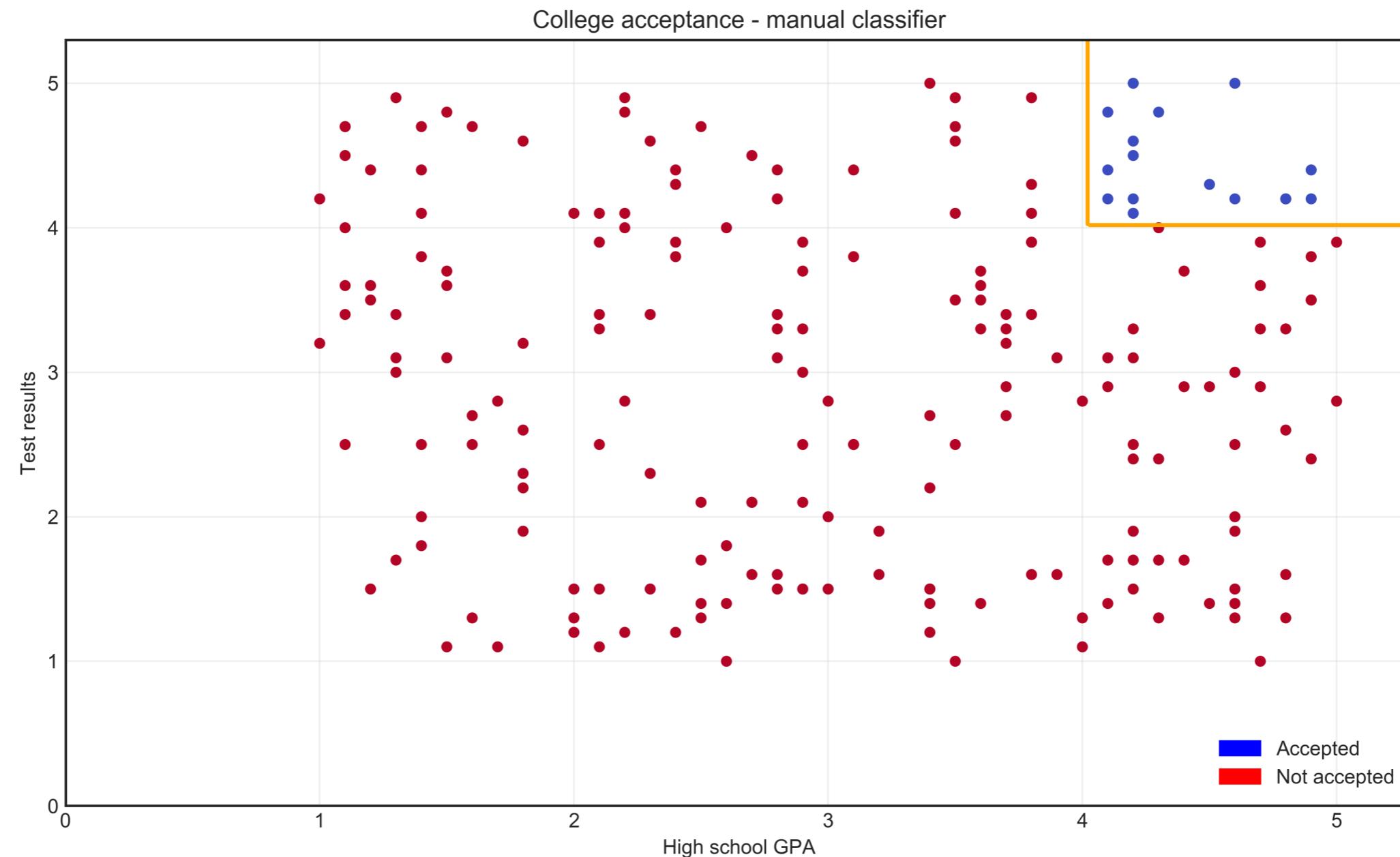
# Graphing our data



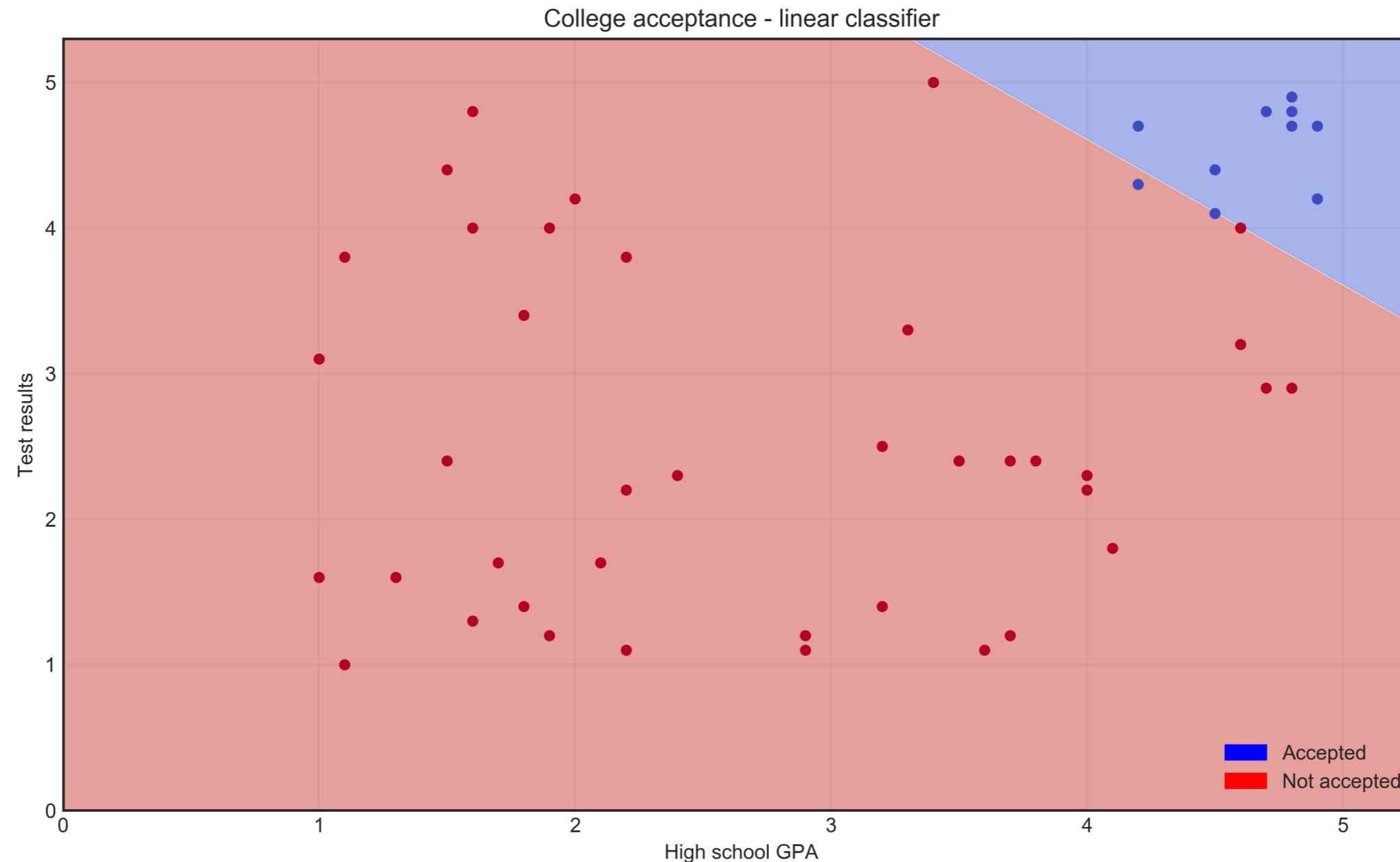
# Splitting data



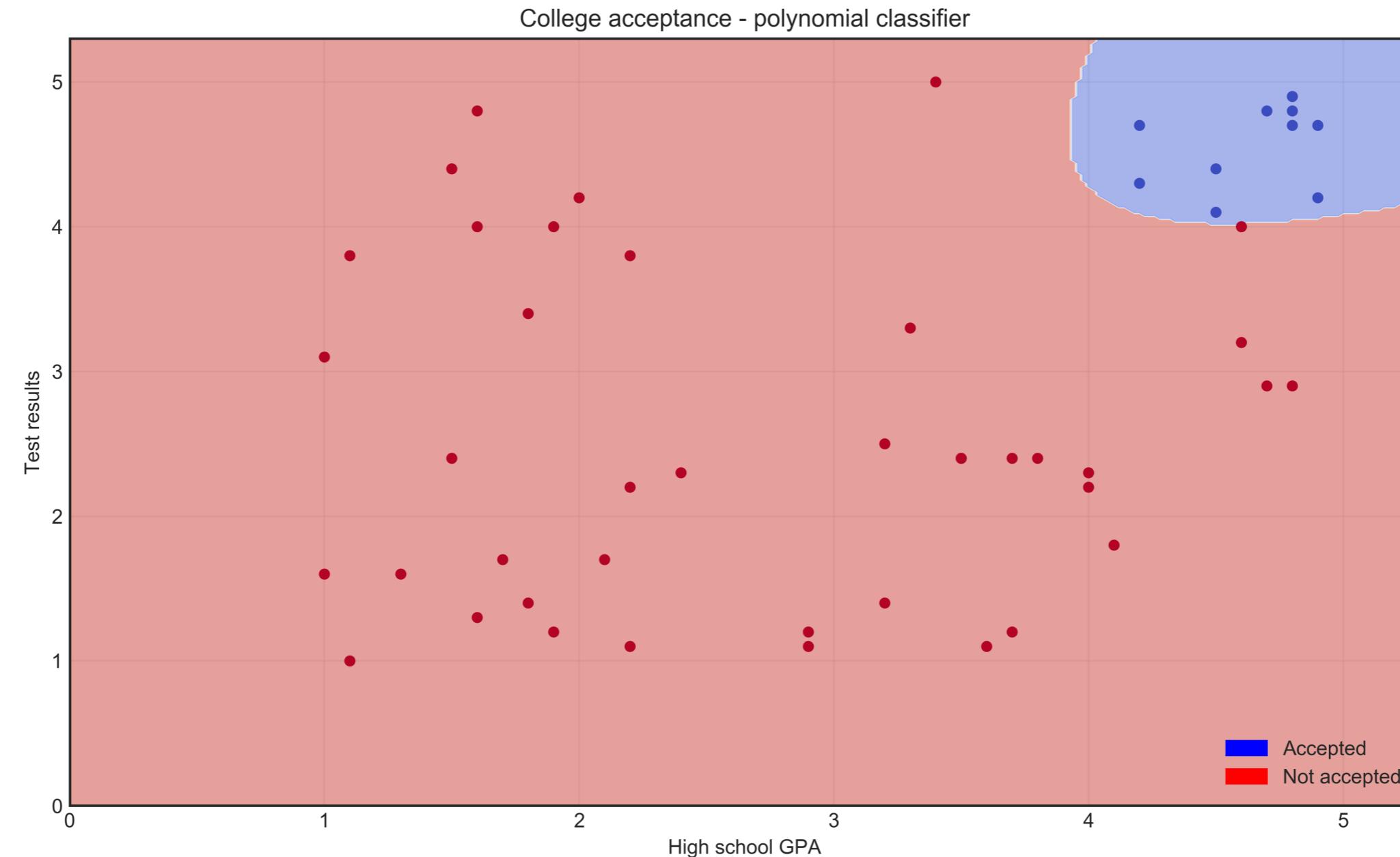
# Manual classifier



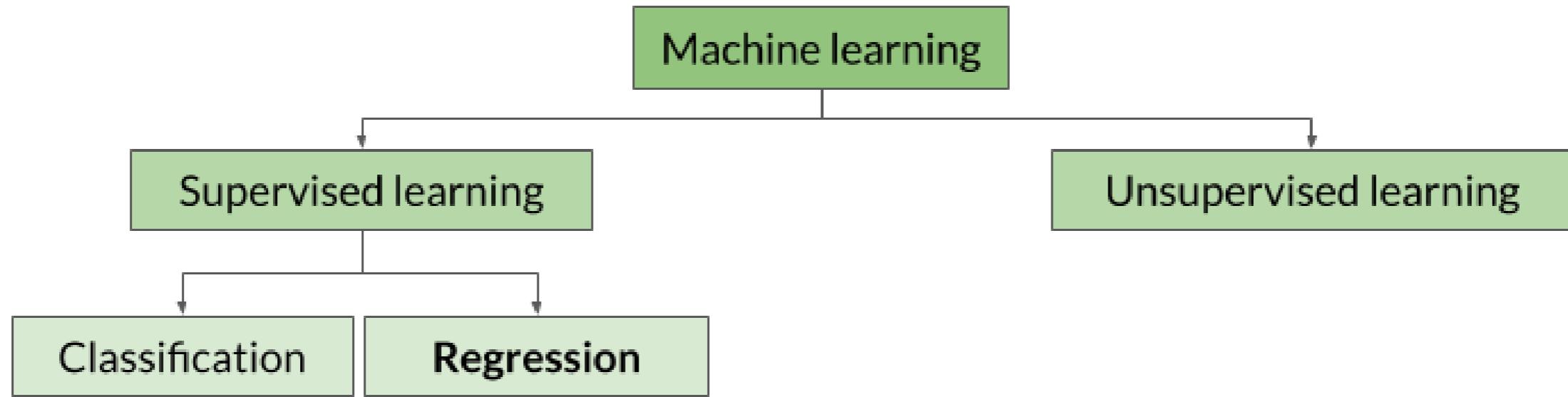
# Support vector machine - linear classifier



# Support vector machine - polynomial classifier



# Regression



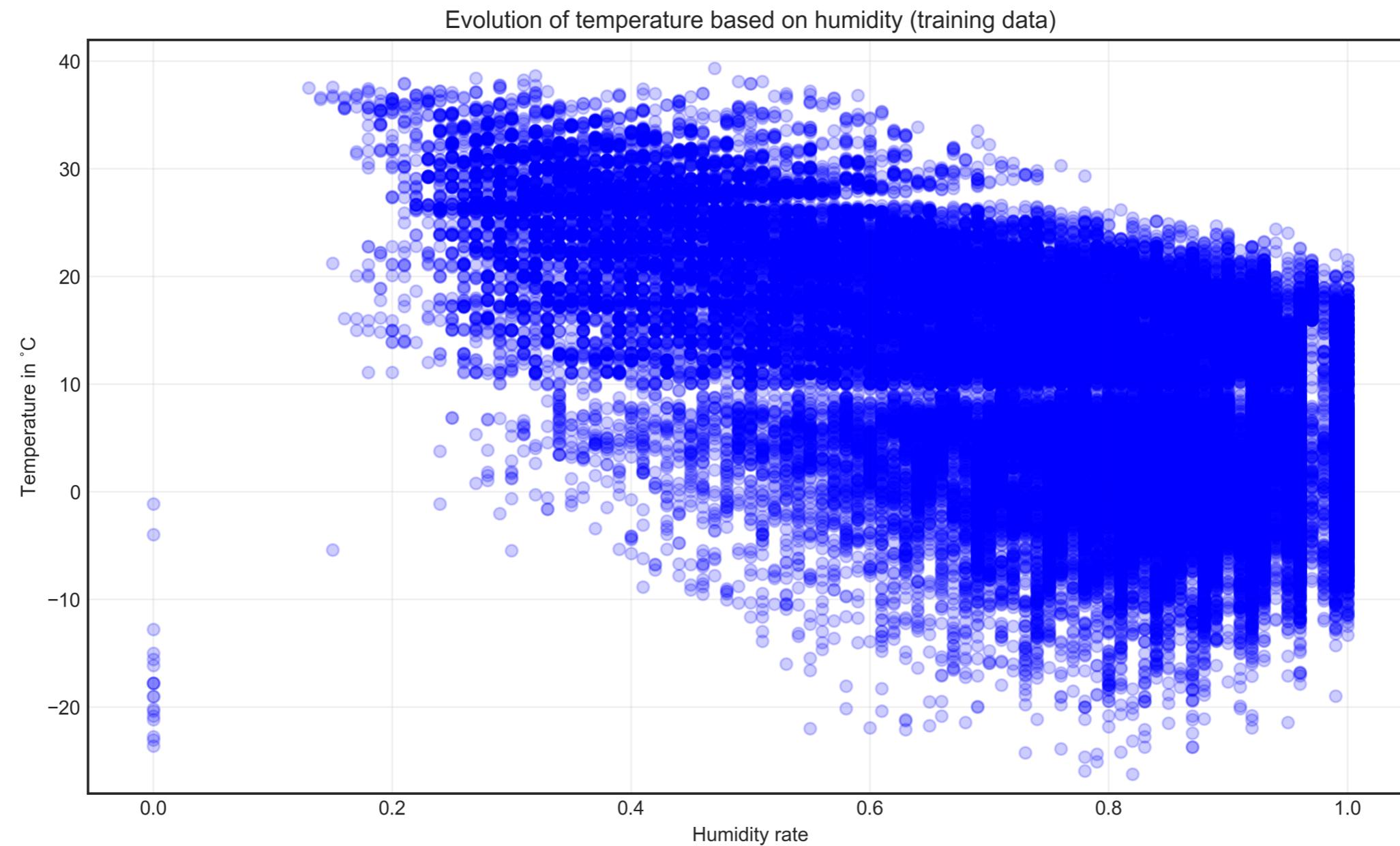
# Regression

- **Regression** = assigning a **continuous** variable
  - How much will this stock be **worth**?
  - What is this exoplanet's **mass**?
  - How **tall** will this child be as an adult?

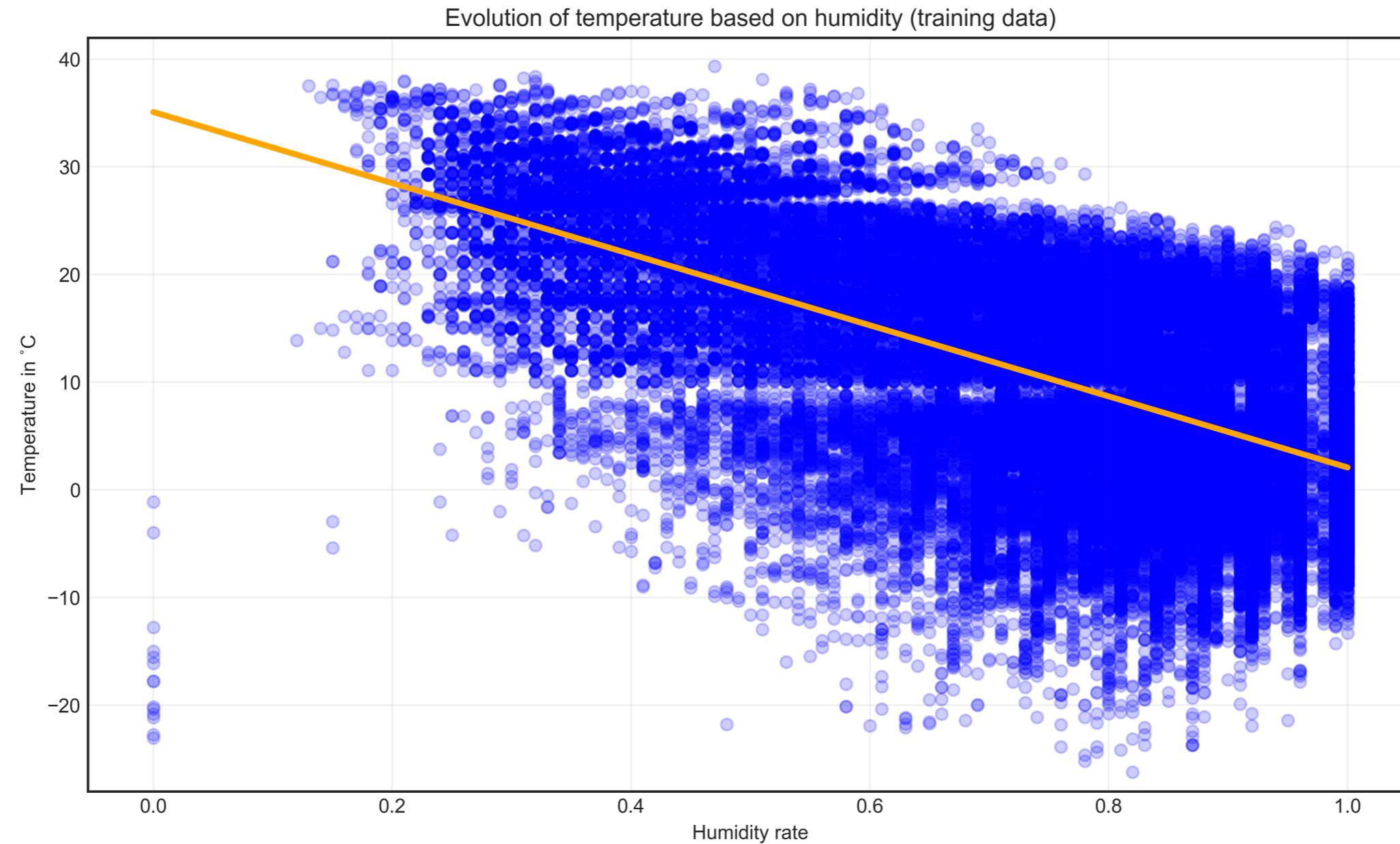
# Predicting temperature

Reading ID	Humidity rate	Temperature in °C
0	0.89	7.388889
1	0.86	7.227778
2	0.89	9.377778
3	0.83	5.944444
...	...	...

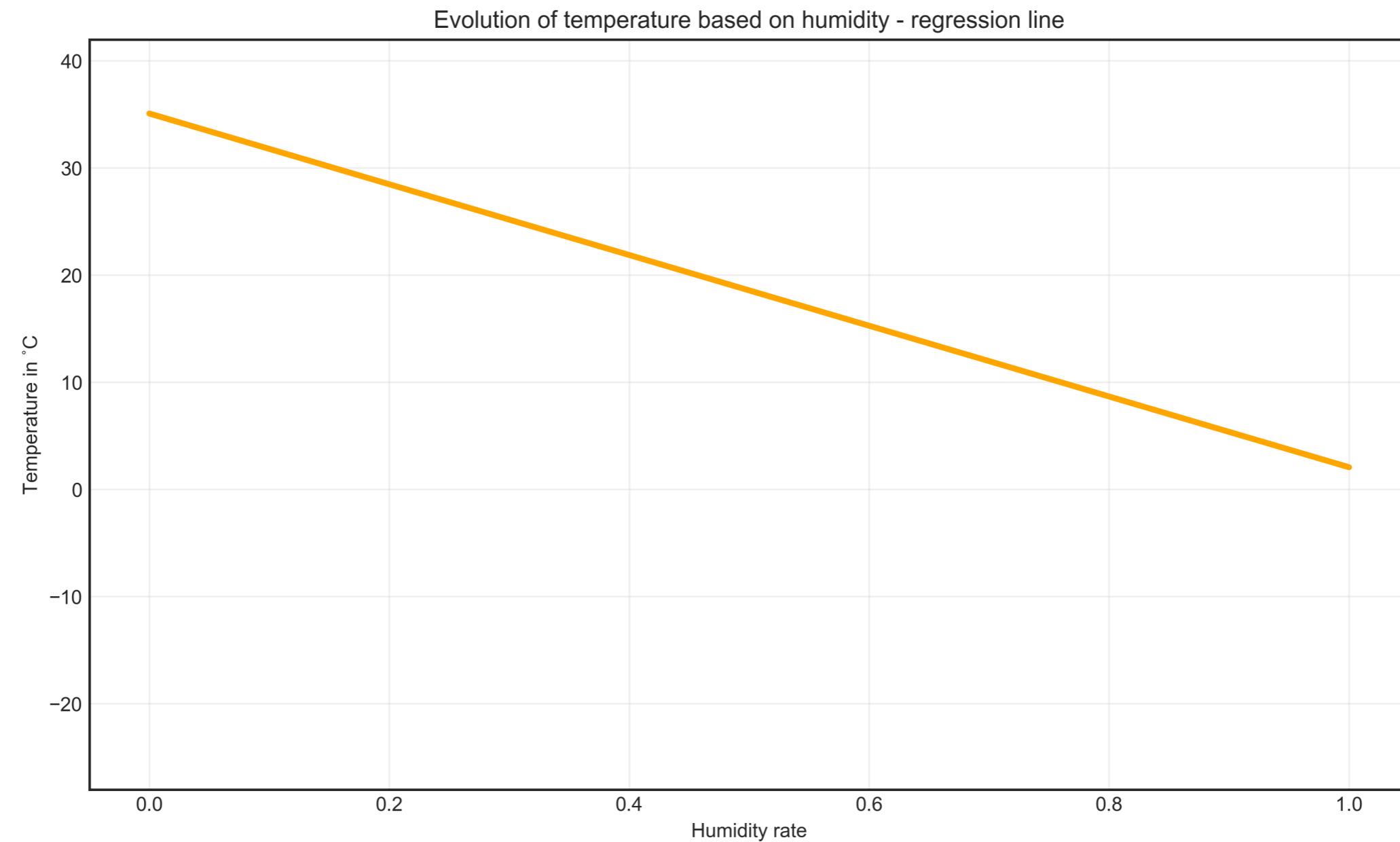
# Training data



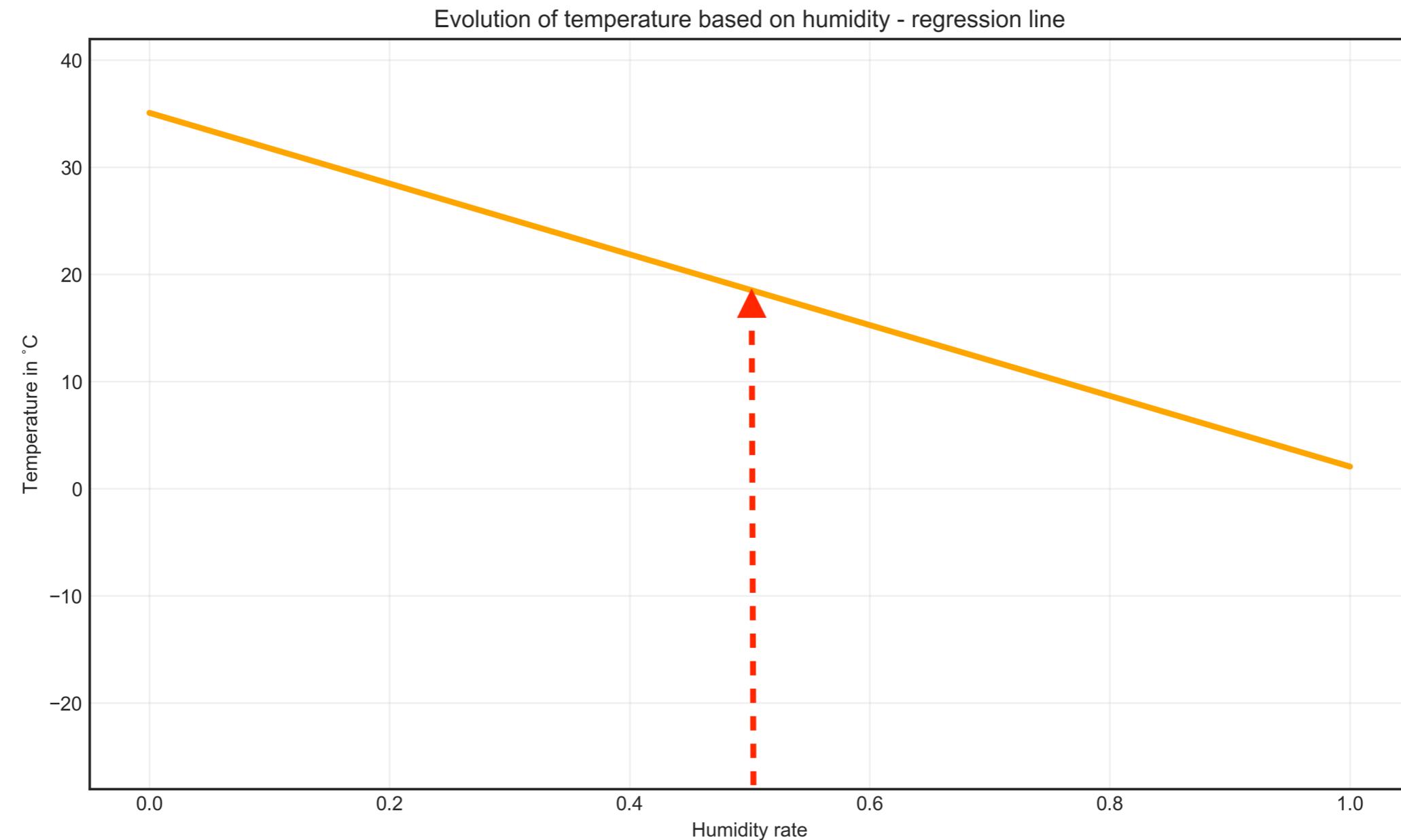
# Linear regression



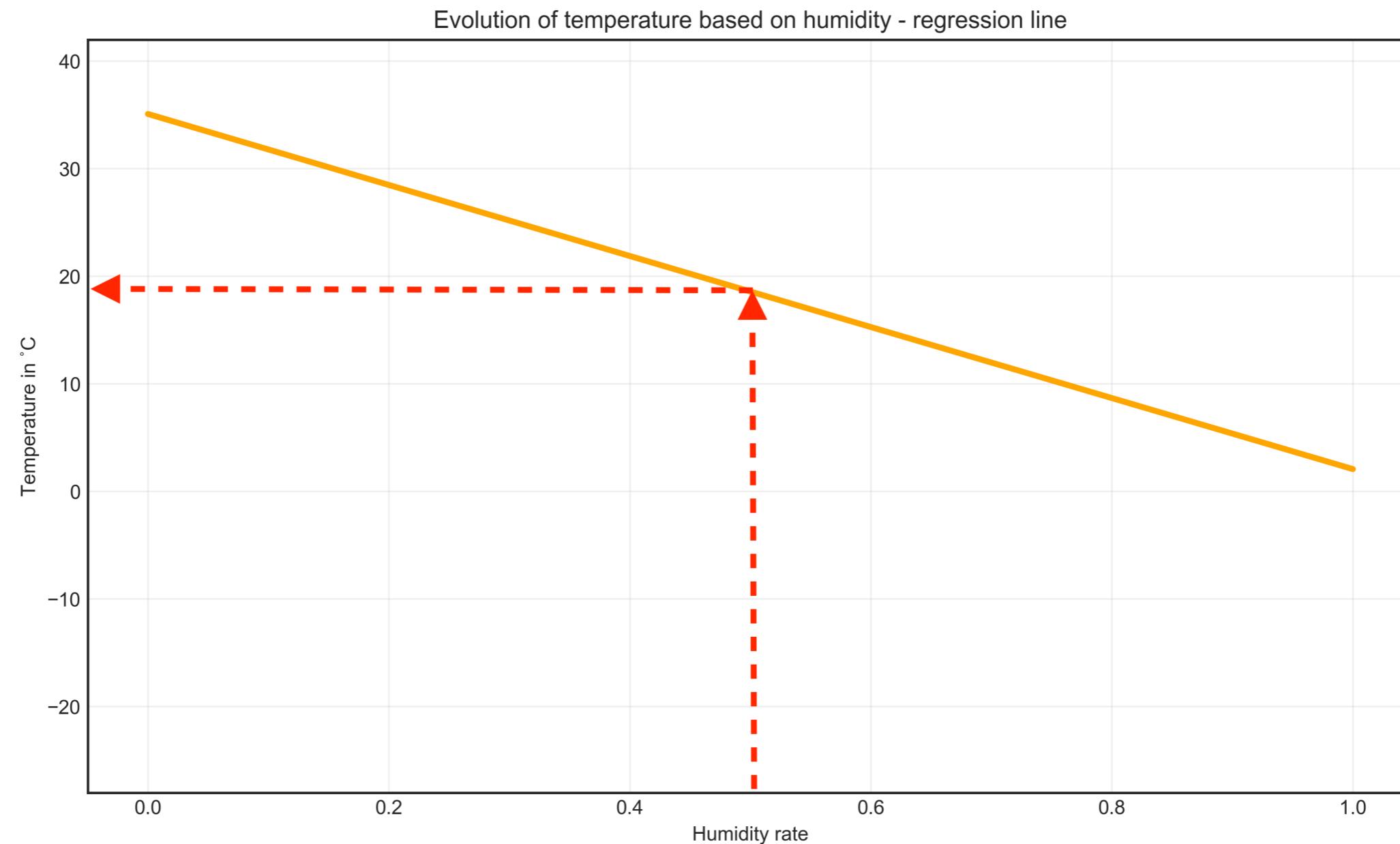
# Model



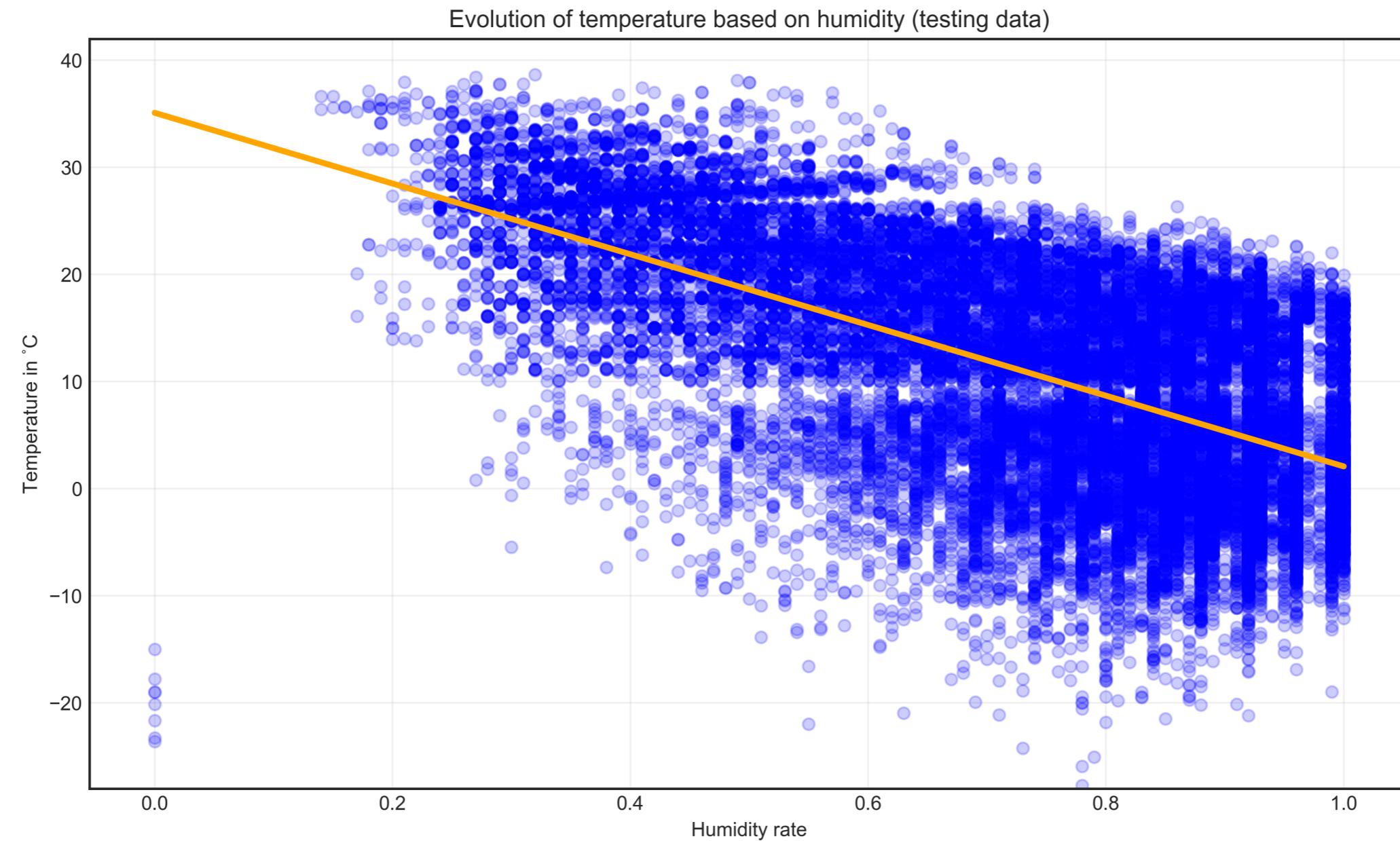
# Given humidity...



# ...find temperature



# Testing data



# Classification vs regression

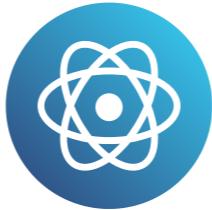
- Regression = **continuous**
  - **Any value** within a finite (*height*) or infinite (*time*) interval
    - $20^{\circ}\text{F}, 20.1^{\circ}\text{F}, 20.01^{\circ}\text{F}...$
- Classification = **category**
  - One of few **specific values**
    - *Cold, Mild, Hot*

# Let's practice!

MACHINE LEARNING FOR EVERYONE

# Unsupervised learning

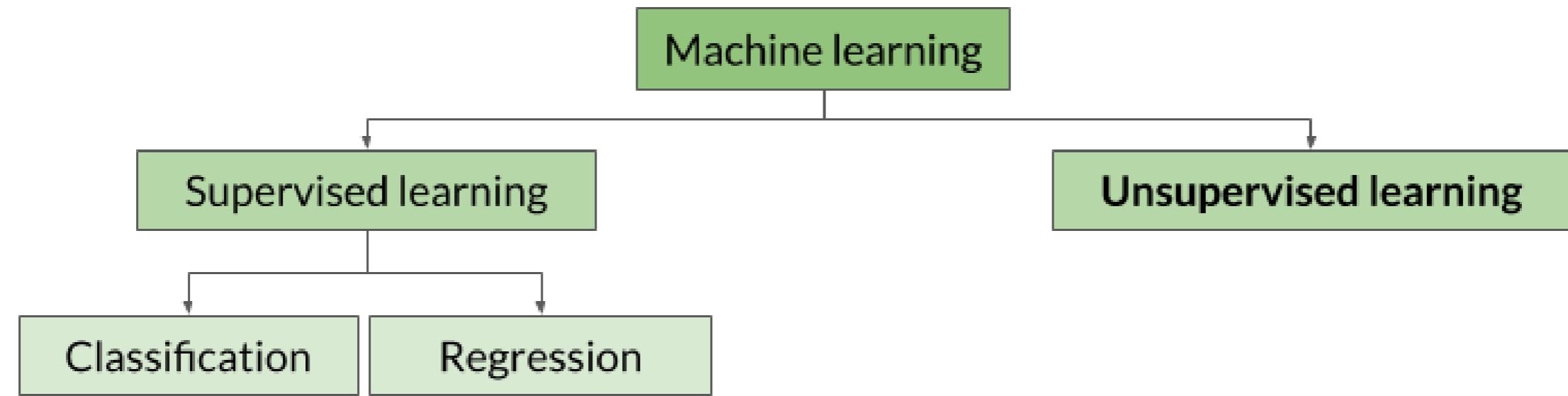
MACHINE LEARNING FOR EVERYONE



**Hadrien Lacroix**

Content Developer at DataCamp

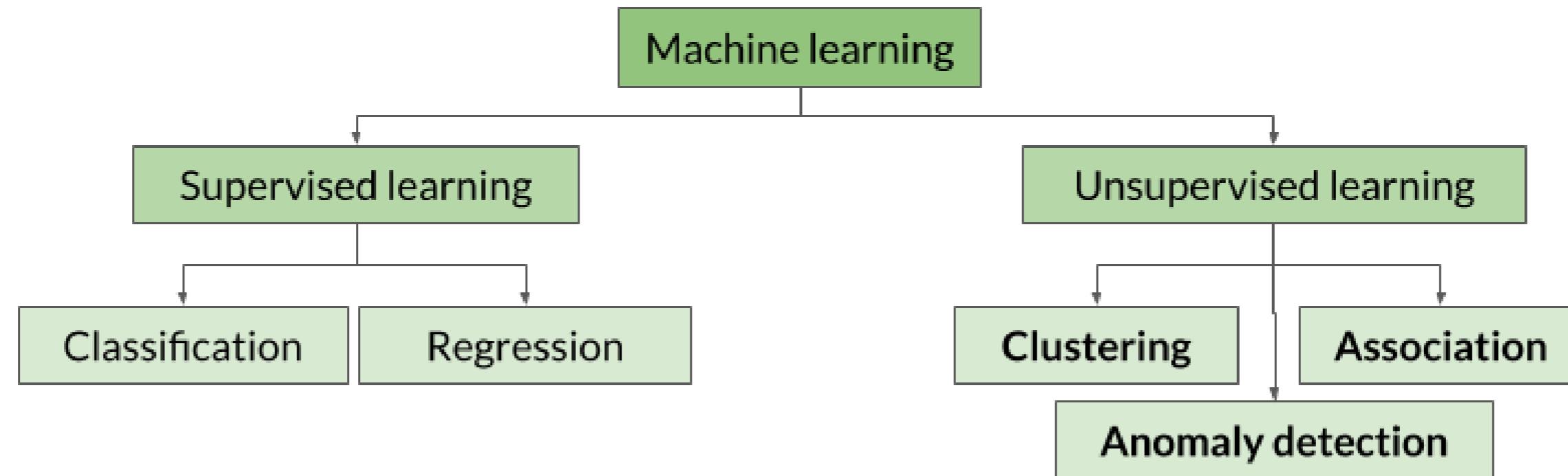
# Supervised learning



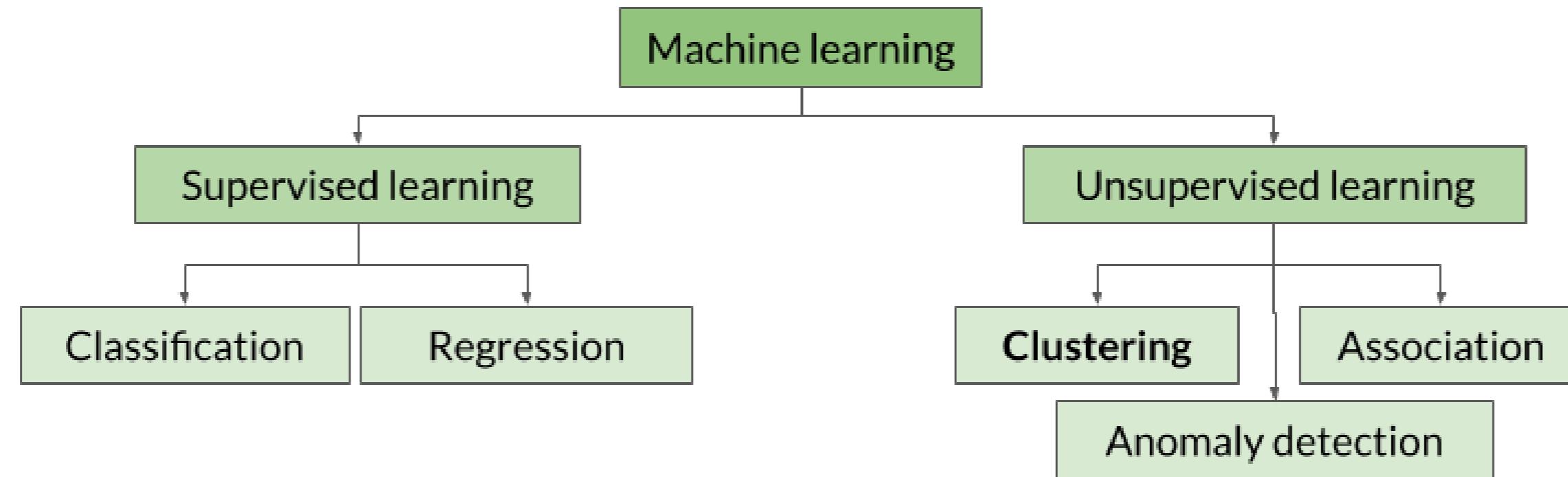
# Unsupervised learning

- Unsupervised learning = **no target column**
  - No guidance
  - Looks at the whole dataset
  - Tries to detect patterns

# Applications



# Clustering



# Clustering example

White Swiss Shepherd



Brown Japanese Bobtail



Brown Akita



Black Norwegian Forest



Black German Shepherd

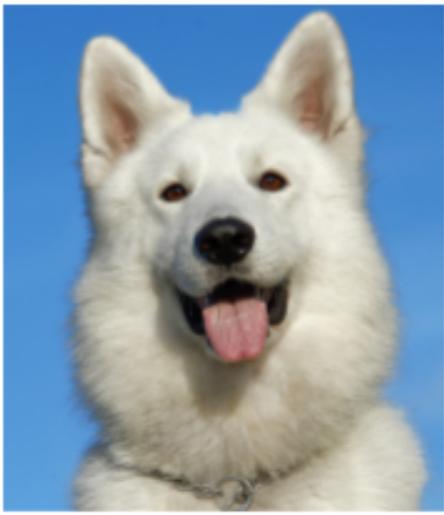


Grey Kurilian Bobtail

# Species cluster

DOGS

White Swiss Shepherd



Black German Shepherd



Brown Akita



CATS

Black Norwegian Forest



Brown Japanese Bobtail



Grey Kurilian Bobtail

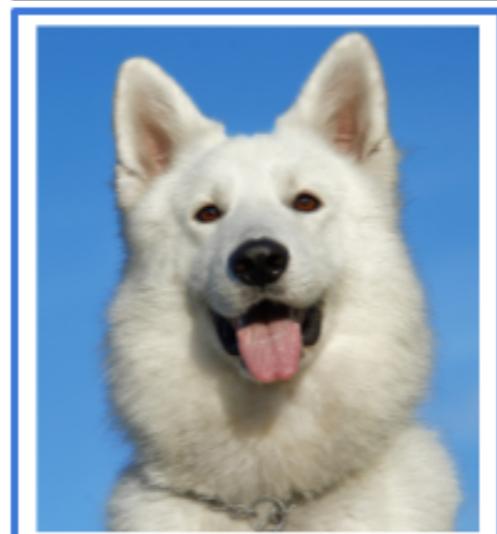


# Color cluster

BLACK

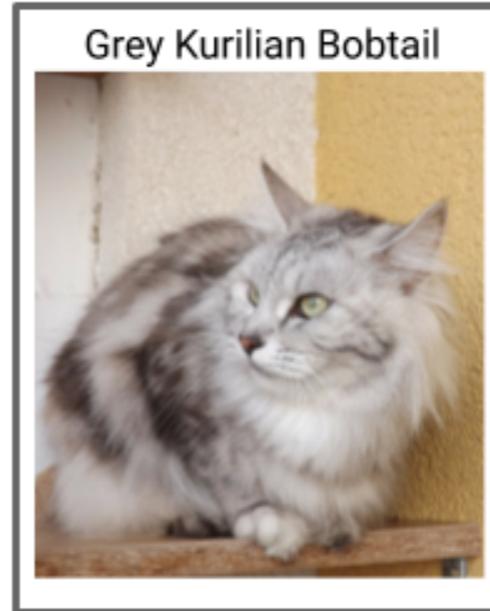


WHITE



White Swiss Shepherd

GREY



BROWN



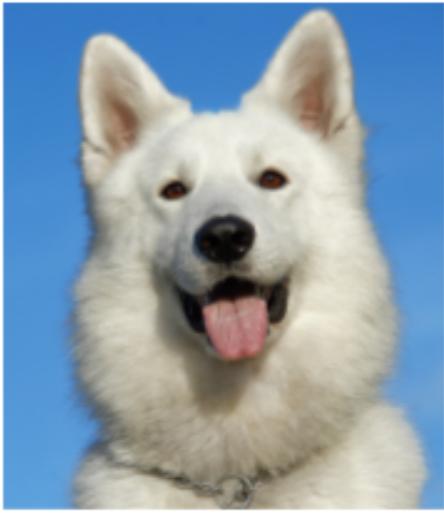
Brown Japanese Bobtail

Brown Akita

# Origin cluster



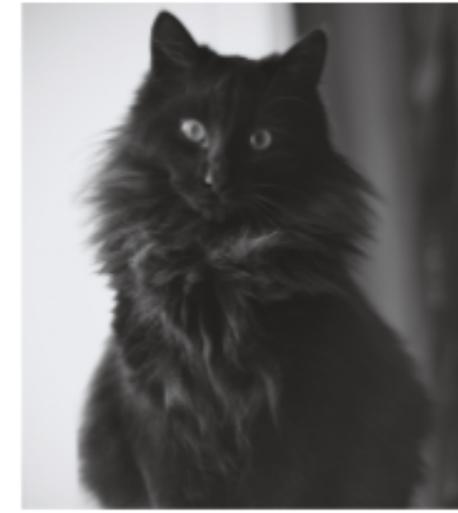
White Swiss Shepherd



Black German Shepherd



Black Norwegian Forest



Brown Akita



Brown Japanese Bobtail



Grey Kurilian Bobtail

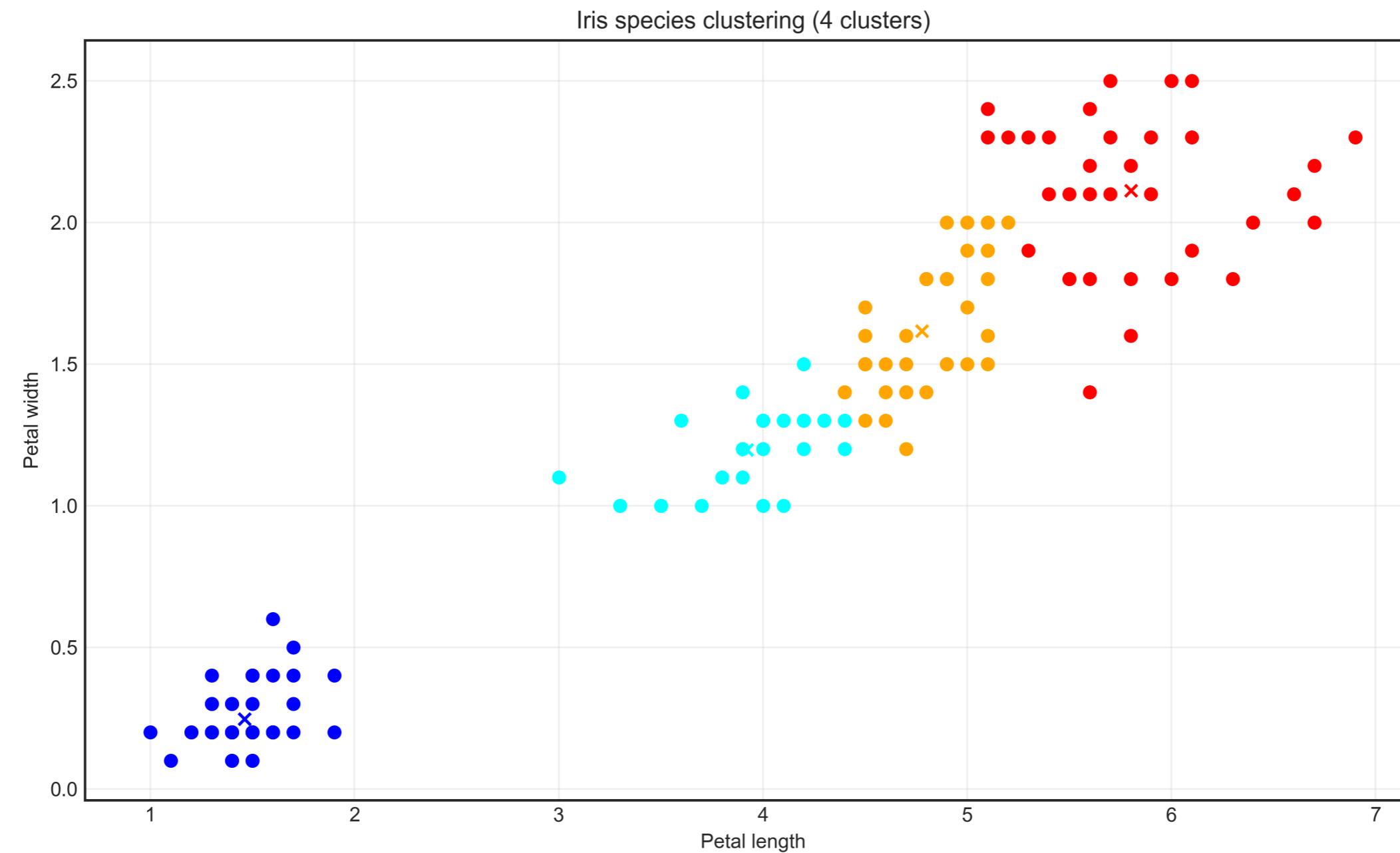
# Clustering models

- K Means:
  - Specify the number of clusters
- DBSCAN (density-based spatial clustering of applications with noise):
  - Specify what constitutes a cluster

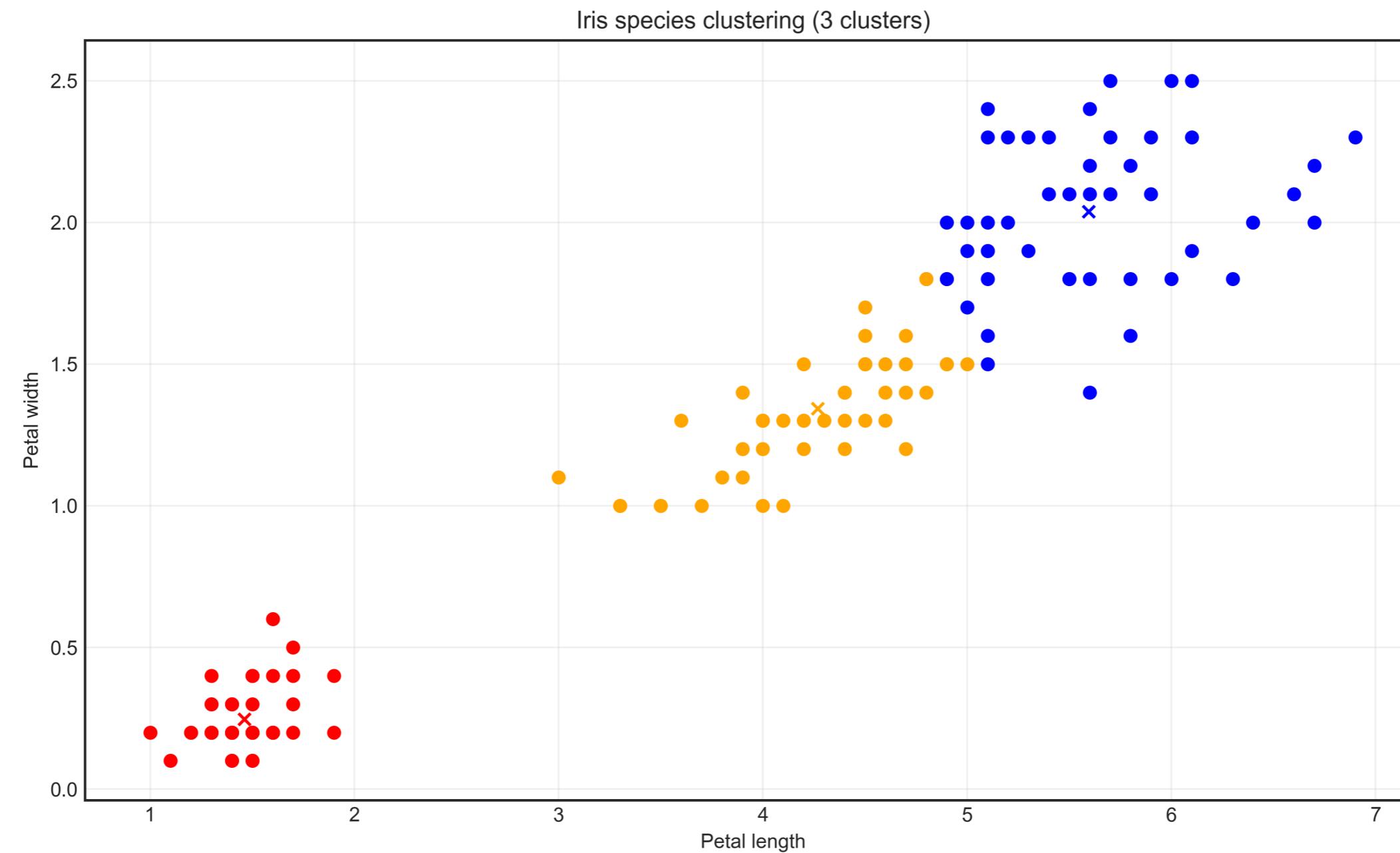
# Iris table

	Petal length	Petal width
0	1.4	0.2
1	1.4	0.2
2	1.3	0.2
3	5.1	1.9
...	...	...

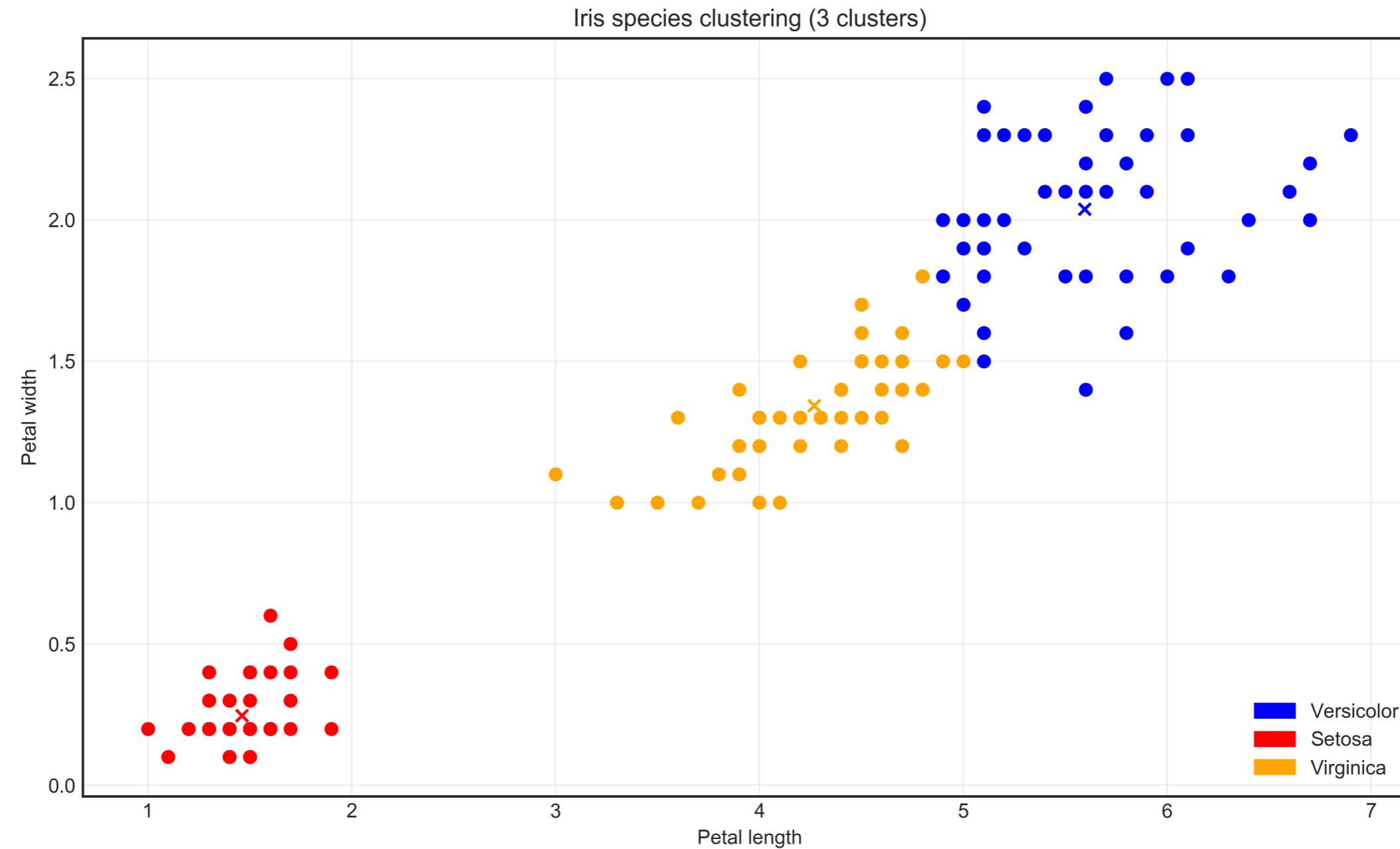
# KNN with 4 clusters



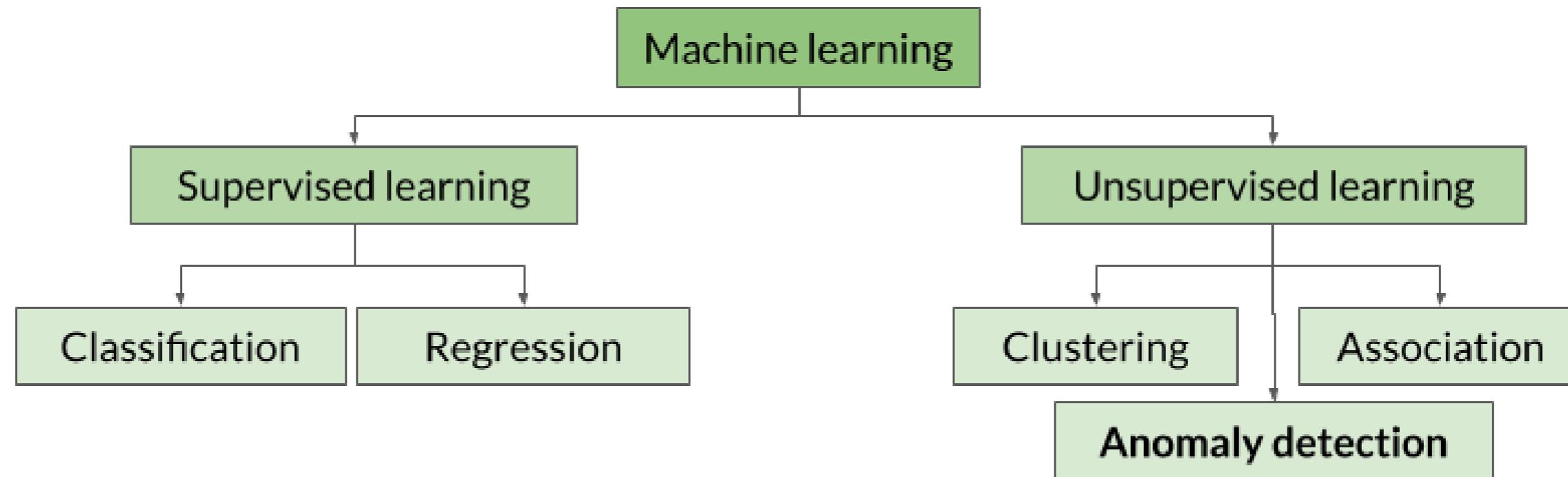
# KNN with 3 clusters



# Ground truth



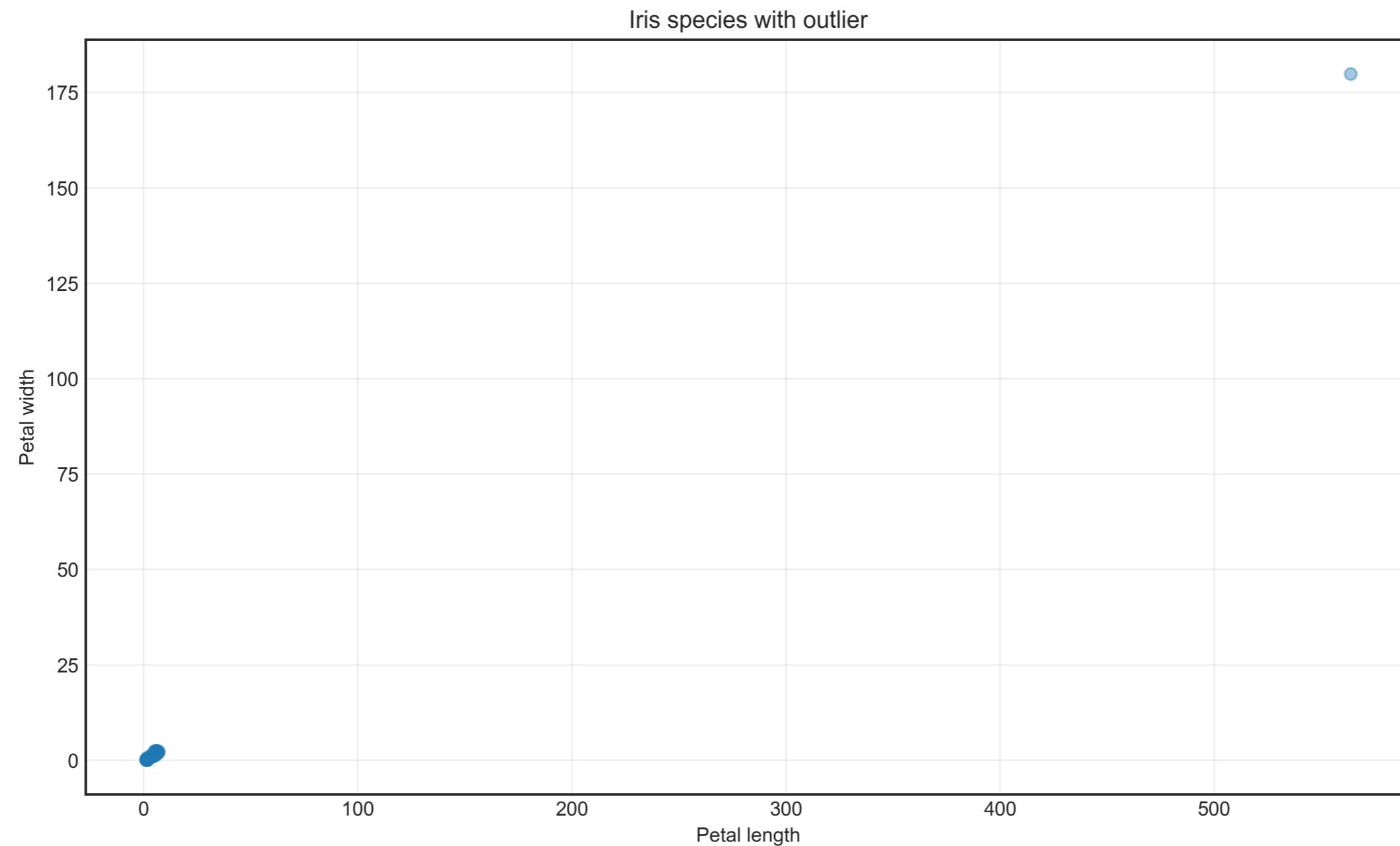
# Anomaly detection



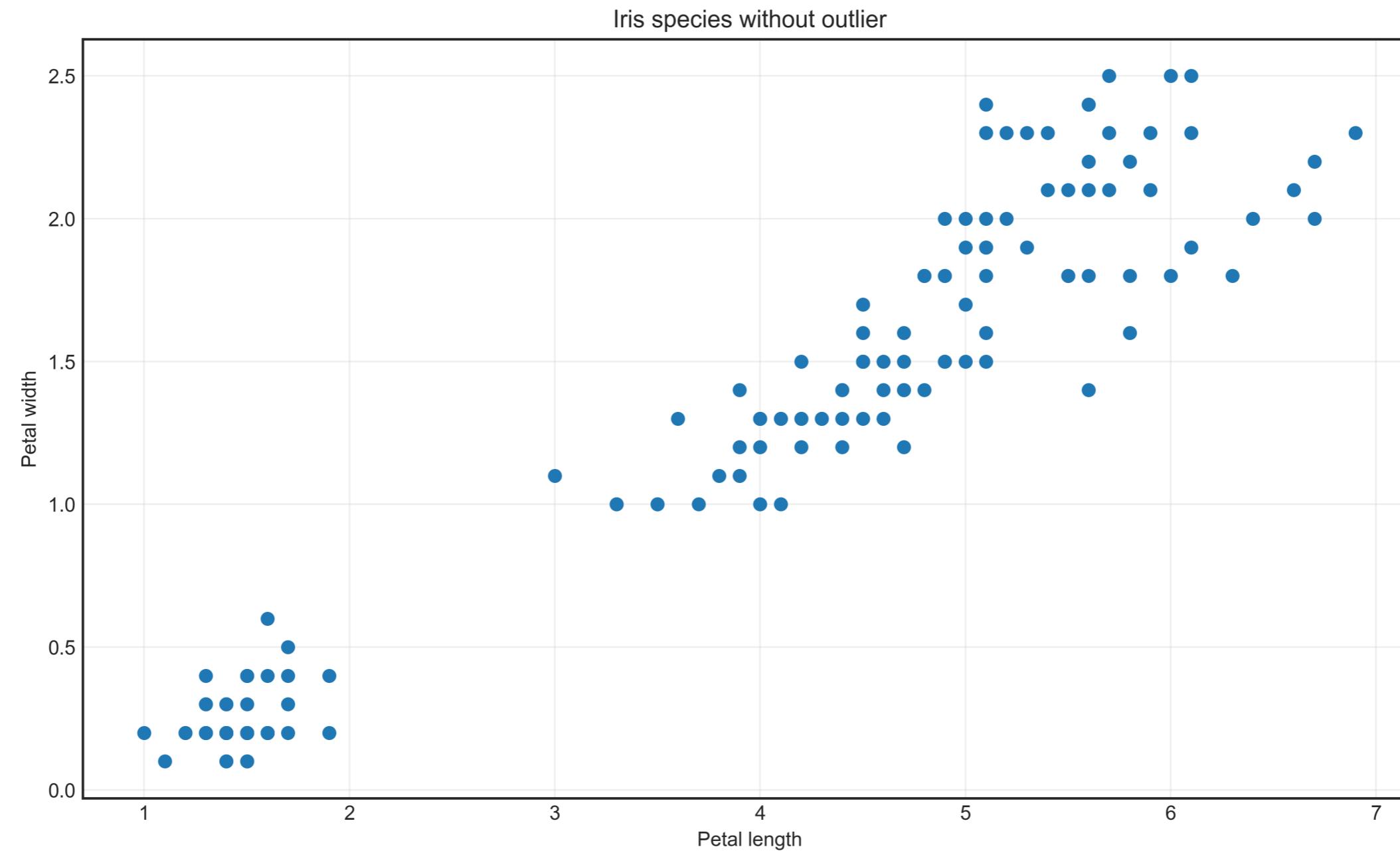
# Detecting outliers

- Anomaly detection = **detecting outliers**
- Outliers = observations that **differ from the rest**

# Outliers



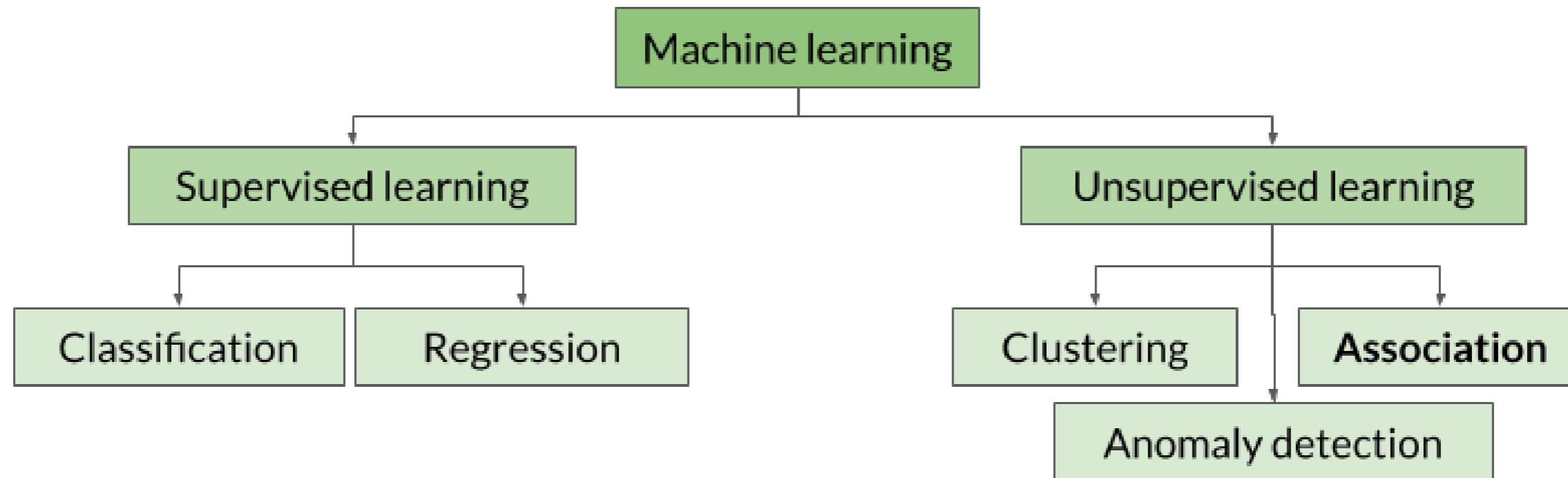
# Removing outliers



# Some anomaly detection use cases

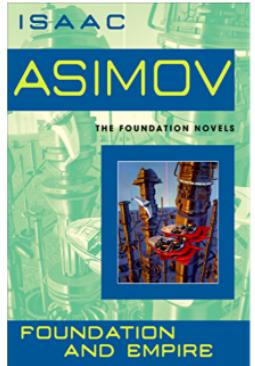
- Discover devices that fail faster or last longer
- Discover fraudsters that manage trick the system
- Discover patients that resist a fatal disease
- ...

# Association

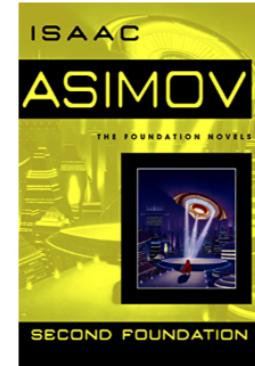


# Association

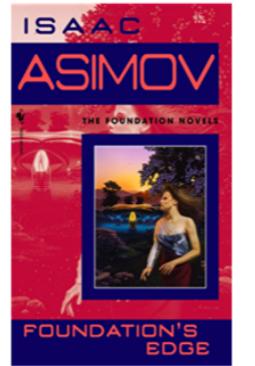
Customers who bought this item also bought



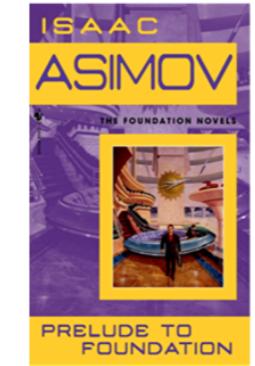
[Foundation and Empire](#)  
by Isaac Asimov  
 335  
Kindle Edition  
1 offer from \$5.99



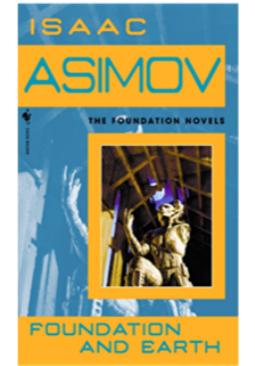
[Second Foundation](#)  
by Isaac Asimov  
 253  
Kindle Edition  
1 offer from \$6.99



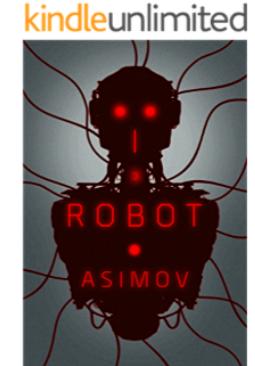
[Foundation's Edge](#)  
by Isaac Asimov  
 277  
Kindle Edition  
1 offer from \$6.99



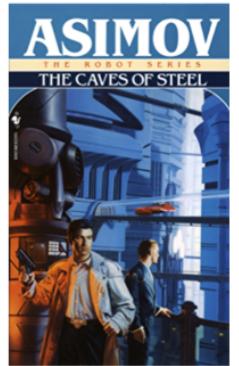
[Prelude to Foundation](#)  
by Isaac Asimov  
 371  
Kindle Edition  
1 offer from \$8.99



[Foundation and Earth](#)  
by Isaac Asimov  
 322  
Kindle Edition  
1 offer from \$6.99



[I, Robot \(The Robot Series\)](#)  
by Isaac Asimov  
 714  
Kindle Edition  
1 offer from \$7.99



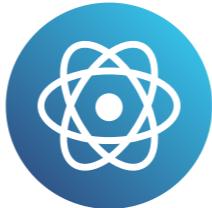
[The Caves of Steel \(The Robot Series Book 1\)](#)  
by Isaac Asimov  
 557  
Kindle Edition  
1 offer from \$7.99

# Let's practice!

MACHINE LEARNING FOR EVERYONE

# Evaluating performance

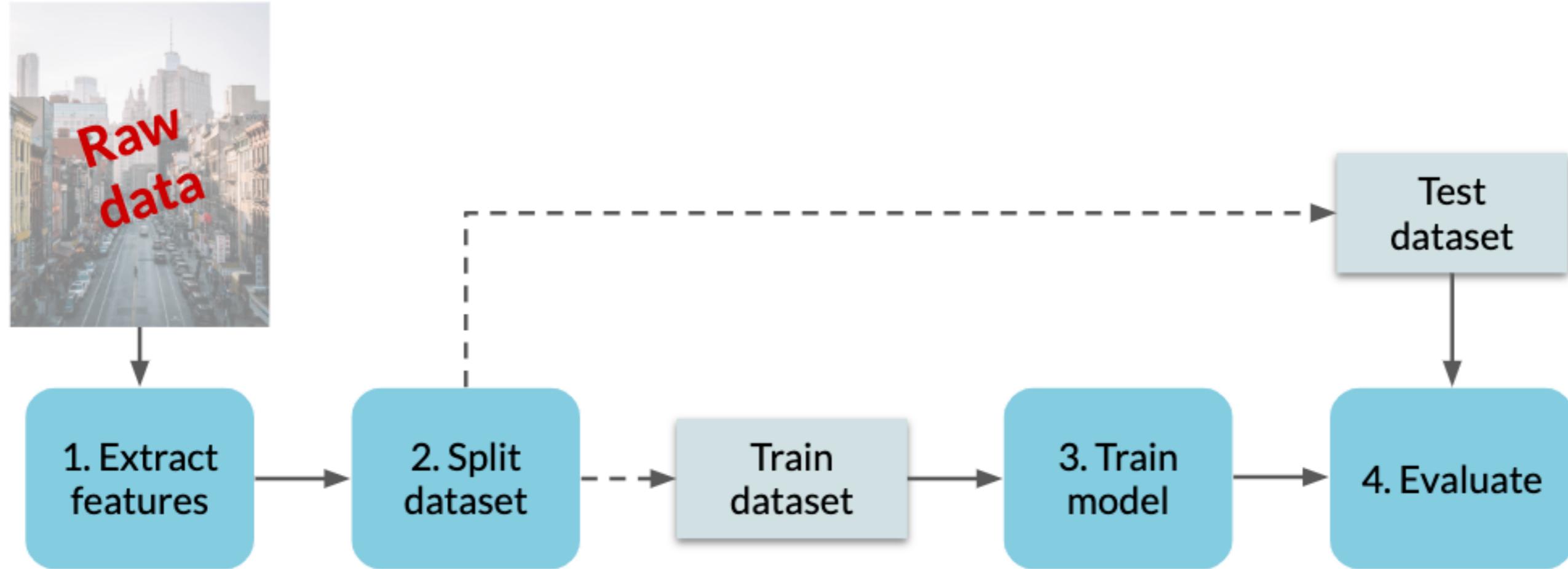
MACHINE LEARNING FOR EVERYONE



**Hadrien Lacroix**

Content Developer at DataCamp

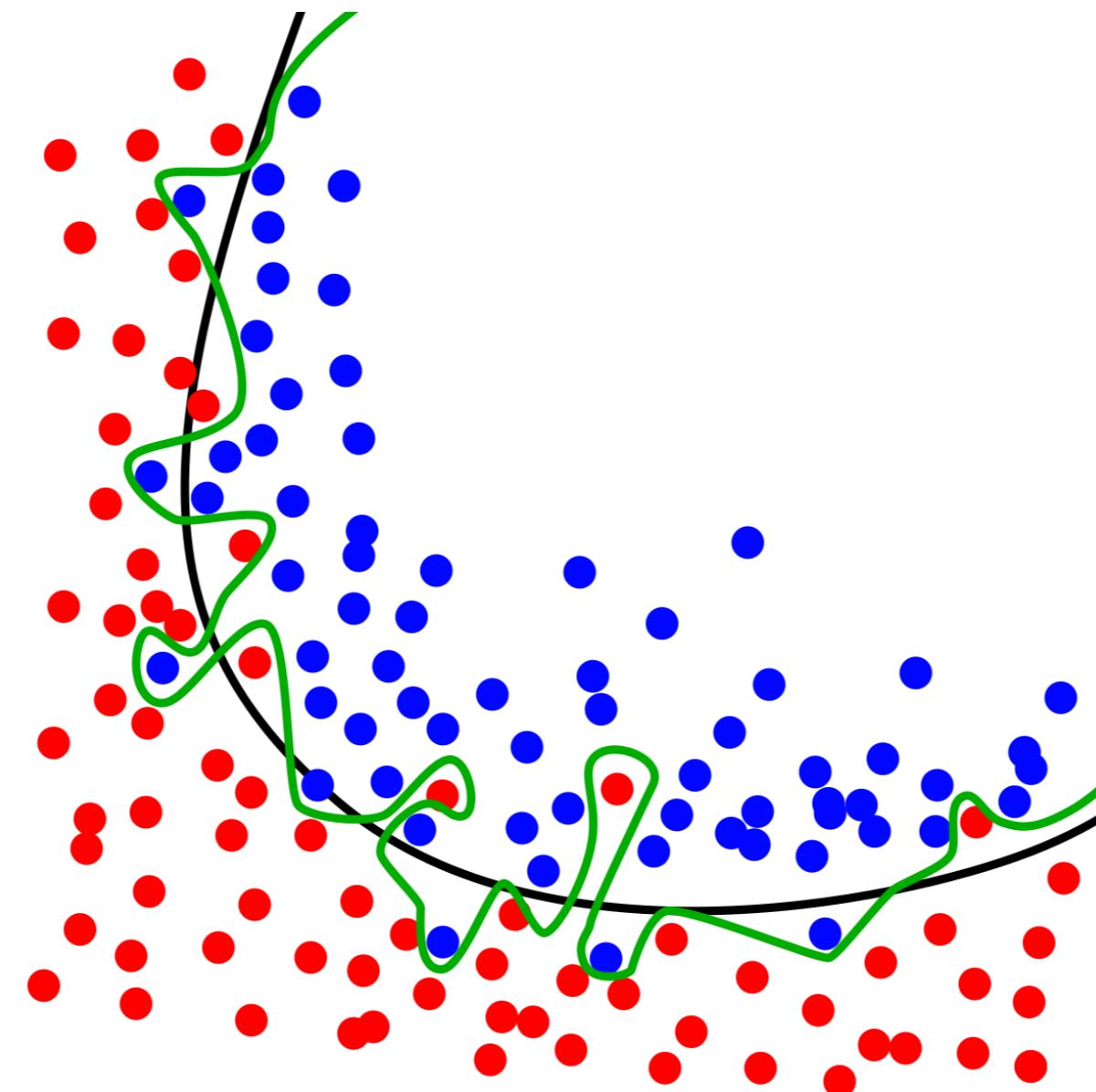
# Evaluate step



# Overfitting

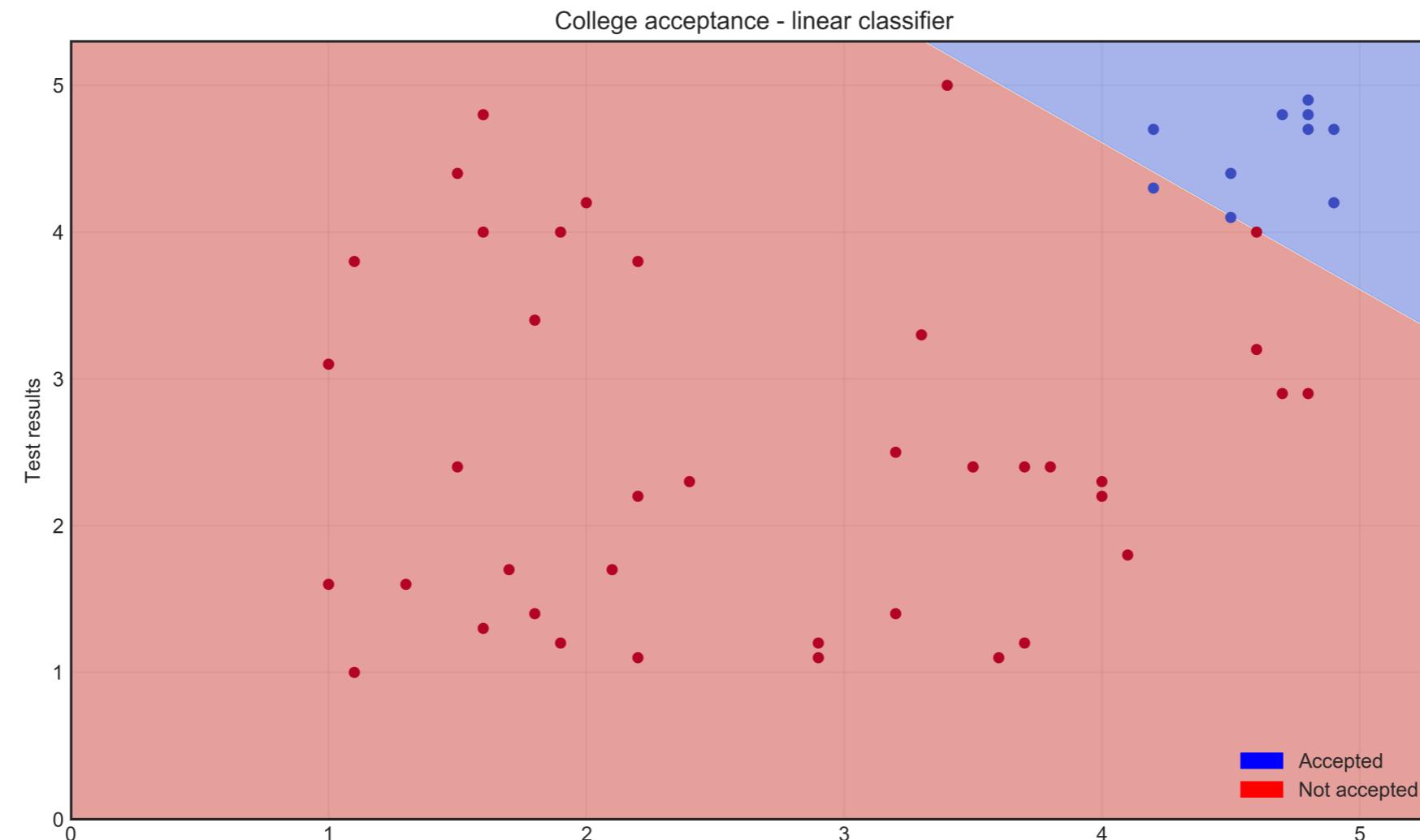
- Performs **great** on **training** data
- Performs **poorly** on **testing** data
- Model memorized **training** data and can't generalize learnings to new data
- Use **testing** set to check model **performance**

# Illustrating overfitting

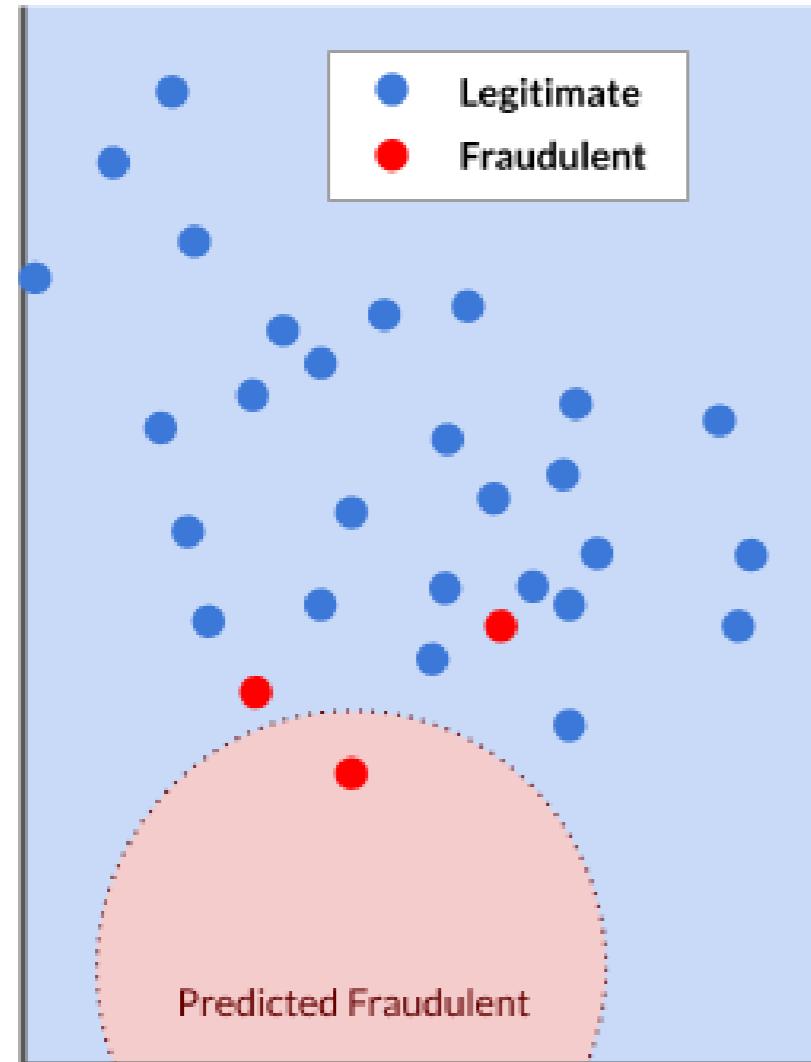


# Accuracy

- Accuracy = **correctly classified observations / all observations**
- $48 / 50 = 96\%$



# Limits of accuracy: fraud example

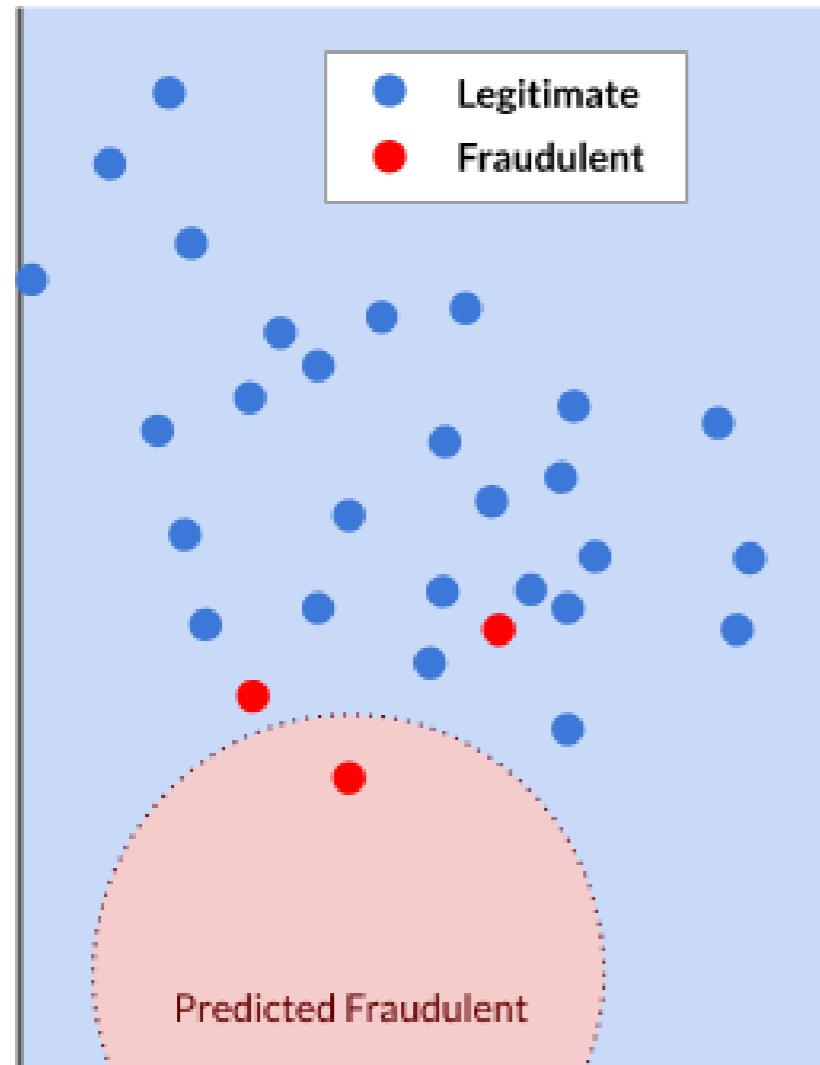


Accuracy of this model:

$$\frac{28 \text{ correctly classified}}{30 \text{ total points}} = 93.33\%$$

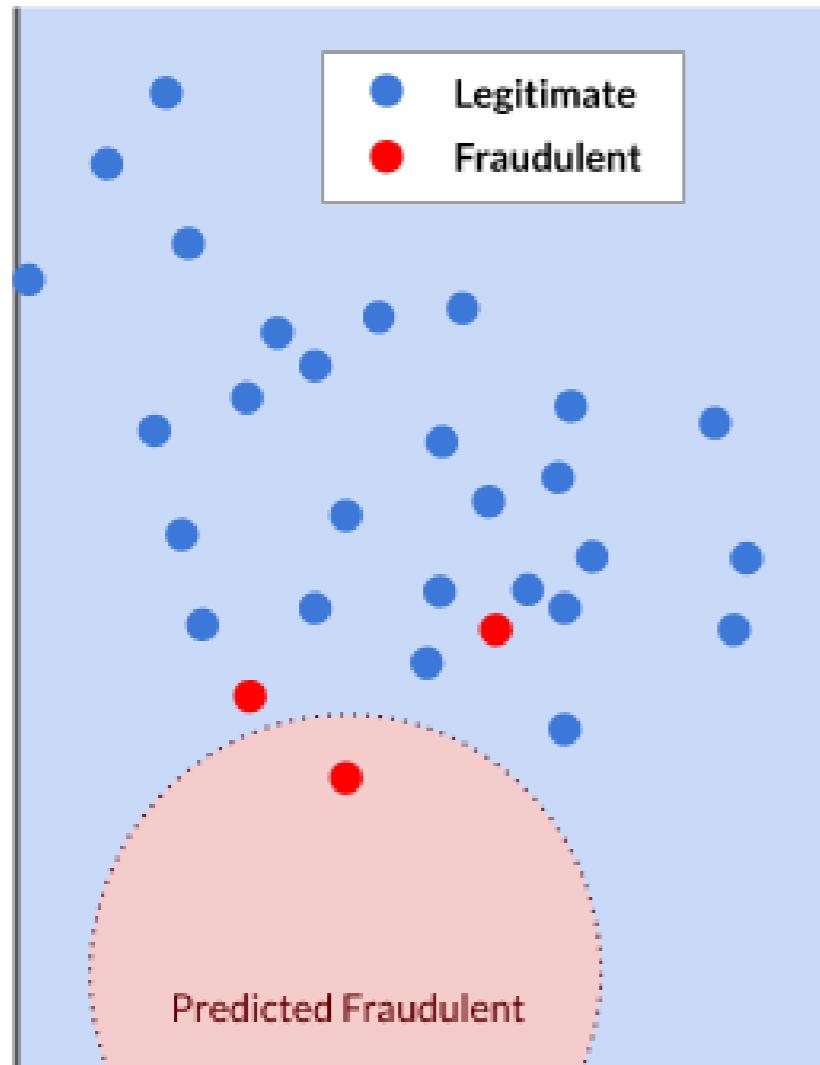
- Misses majority of fraudulent transactions
- Need a better metric

# Confusion matrix



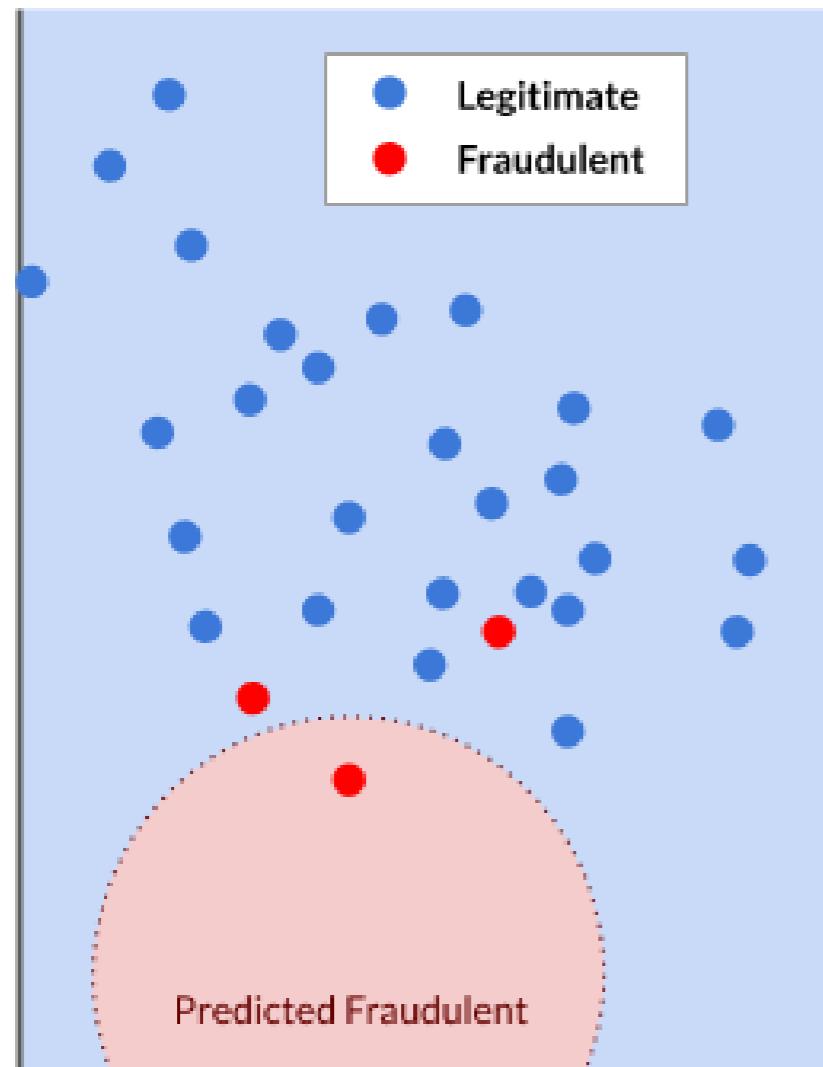
		Actual values	
		<i>Fraudulent</i>	<i>Not Fraudulent</i>
<b>Predicted</b>	<i>Fraudulent</i>		
	<i>Not Fraudulent</i>		

# True positives



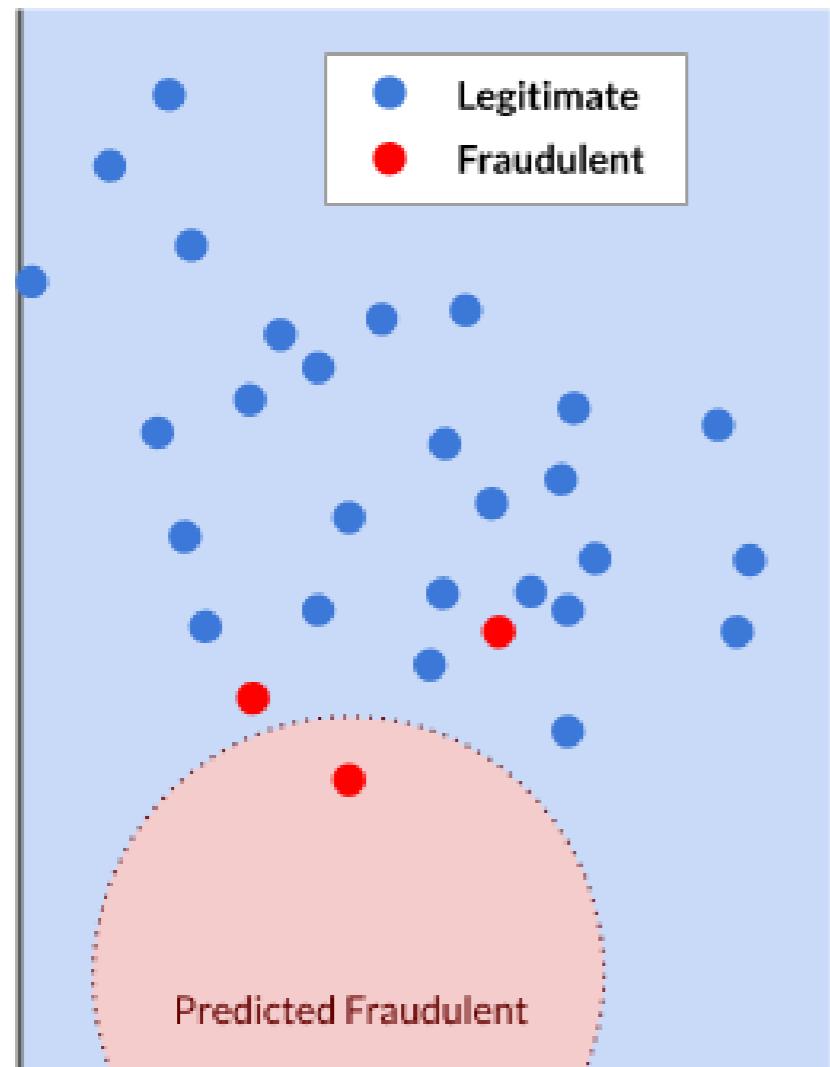
		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent		
	Not Fraudulent		

# True positives



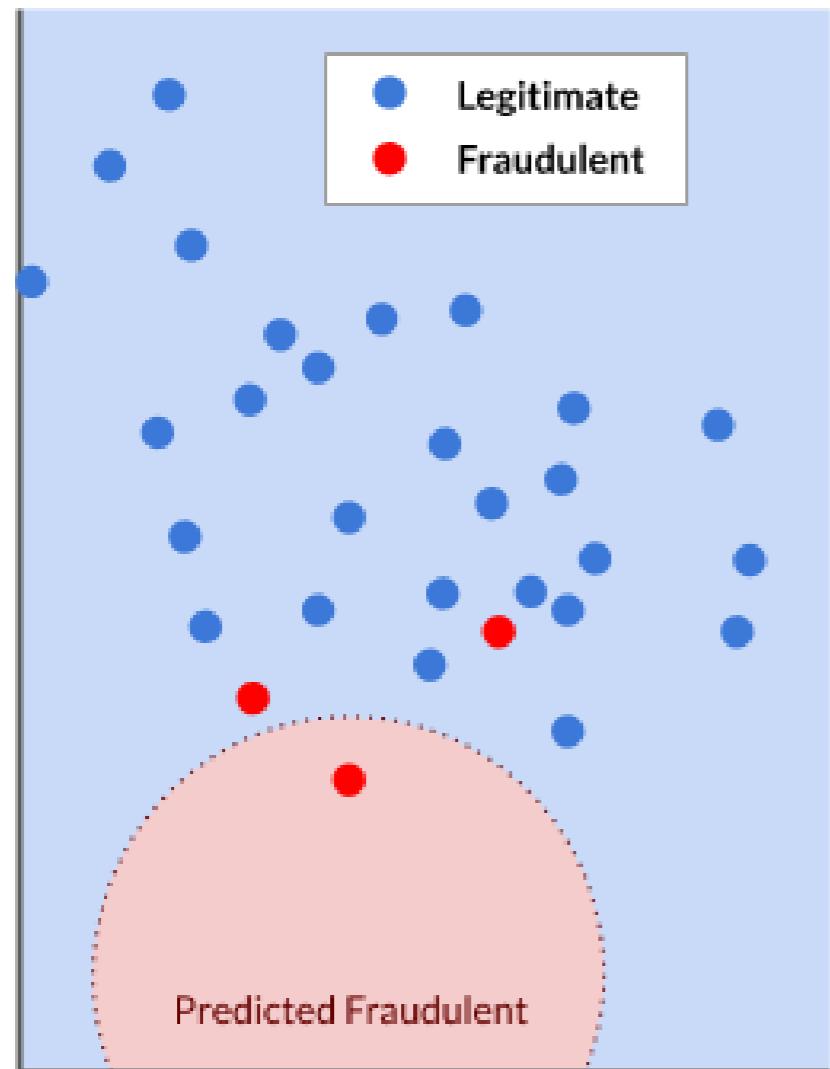
		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent	1 true positives	
	Not Fraudulent		

# False negatives



		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent	1 true positives	
	Not Fraudulent		

# False negatives

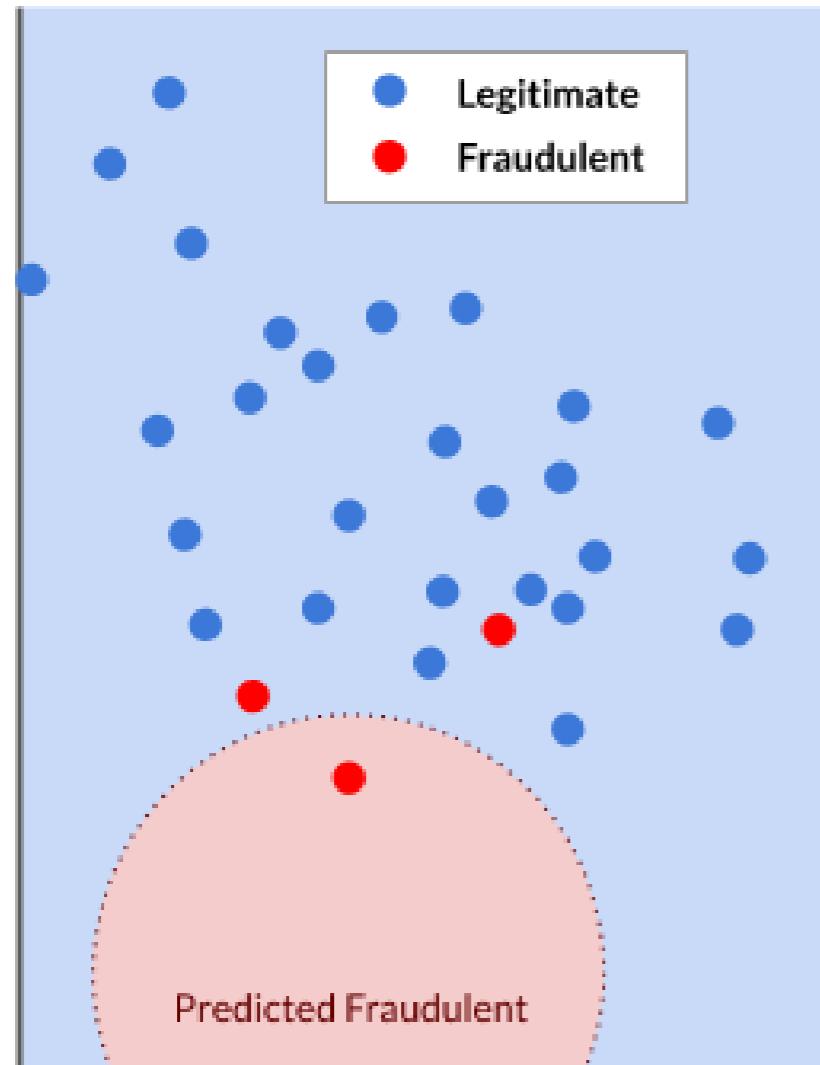


		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent	1 true positives	
	Not Fraudulent	2 false negatives	

# Remembering False Negatives

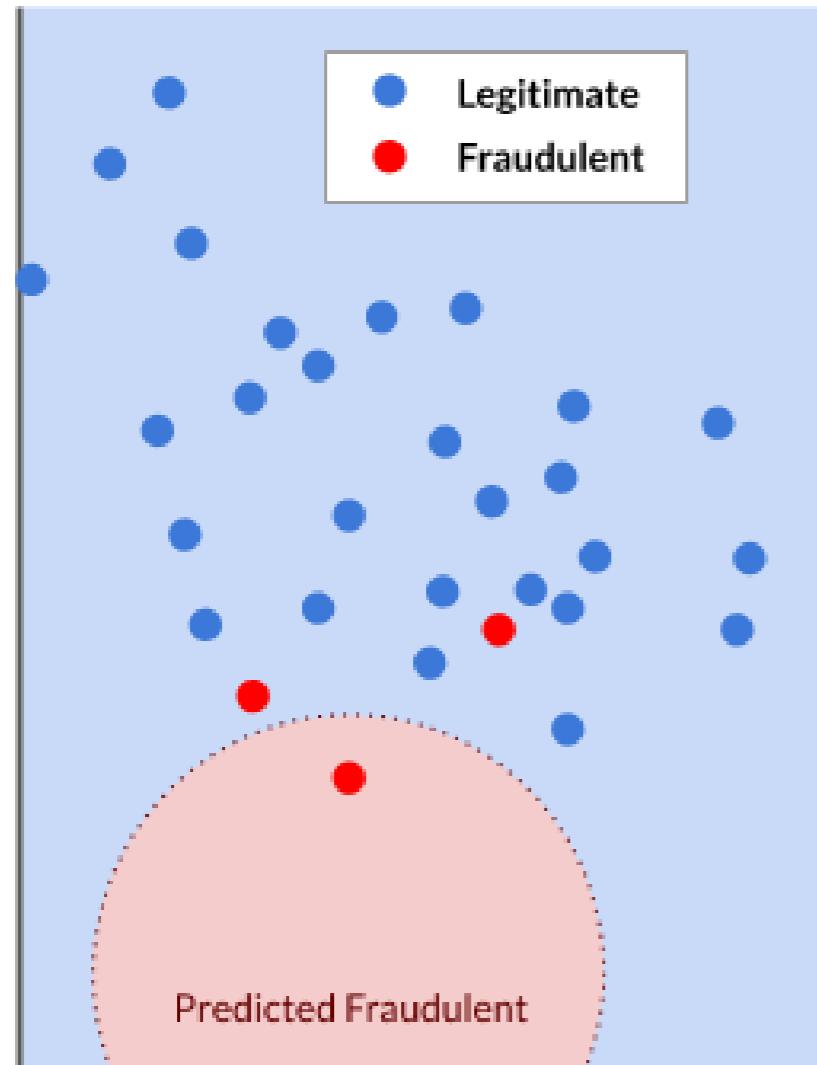


# Fill out the rest...



		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent	1 true positives	
	Not Fraudulent	2 false negatives	

# False positives, true negatives



		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent	1 true positives	0 false positives
	Not Fraudulent	2 false negatives	27 true negatives

# Remembering False Positives



<sup>1</sup> <https://www.flickr.com/photos/59632563@N04/6104068209>

# Sensitivity

		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent	1 true positives	0 false positives
	Not Fraudulent	2 false negatives	27 true negatives

How many fraudulent transactions did we classify correctly?

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = 1/3 = 33.33\%$$

- Rather mark legitimate transactions as suspicious than authorize fraudulent transactions

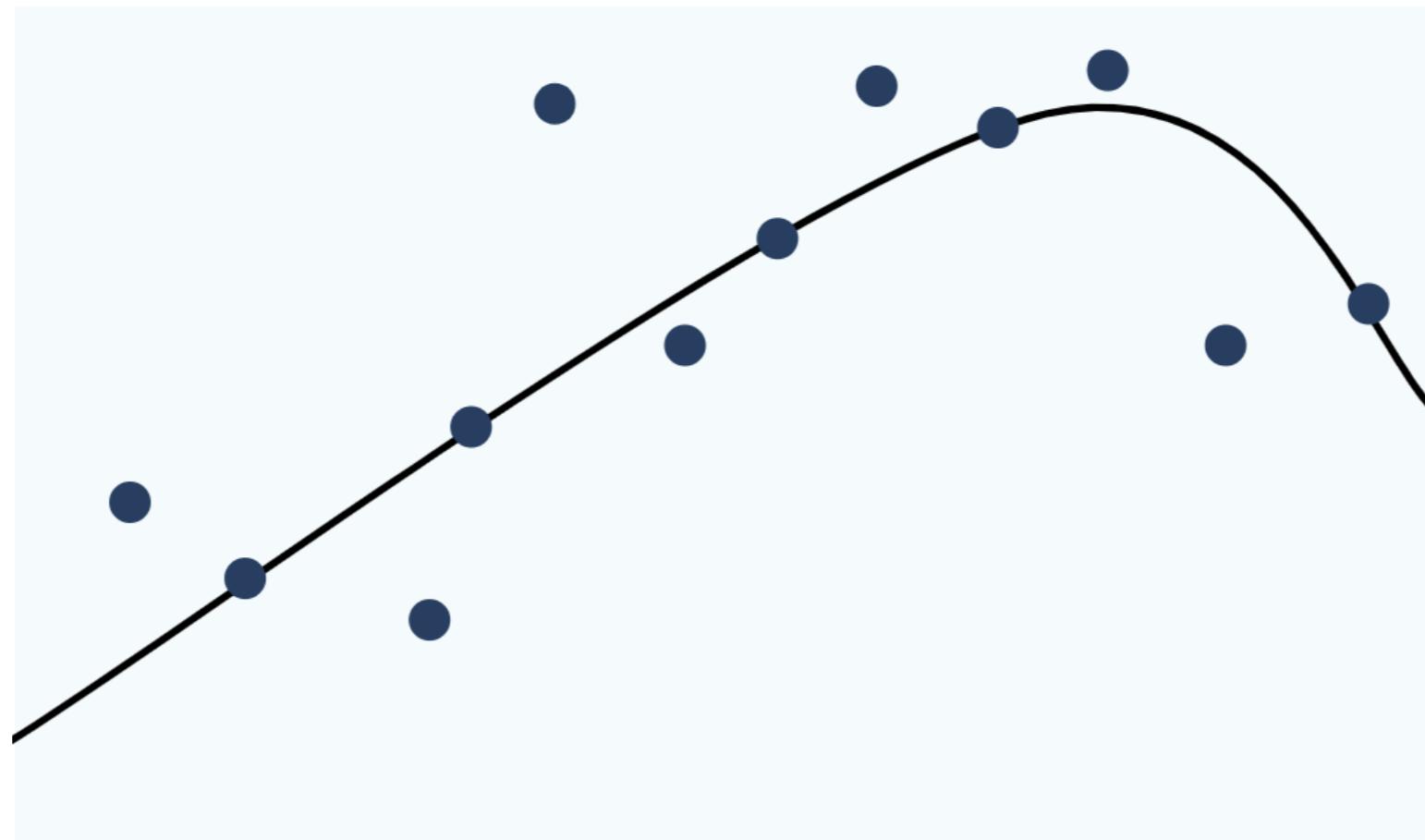
# Specificity

$$Specificity = \frac{true\ negatives}{true\ negatives + false\ positives}$$

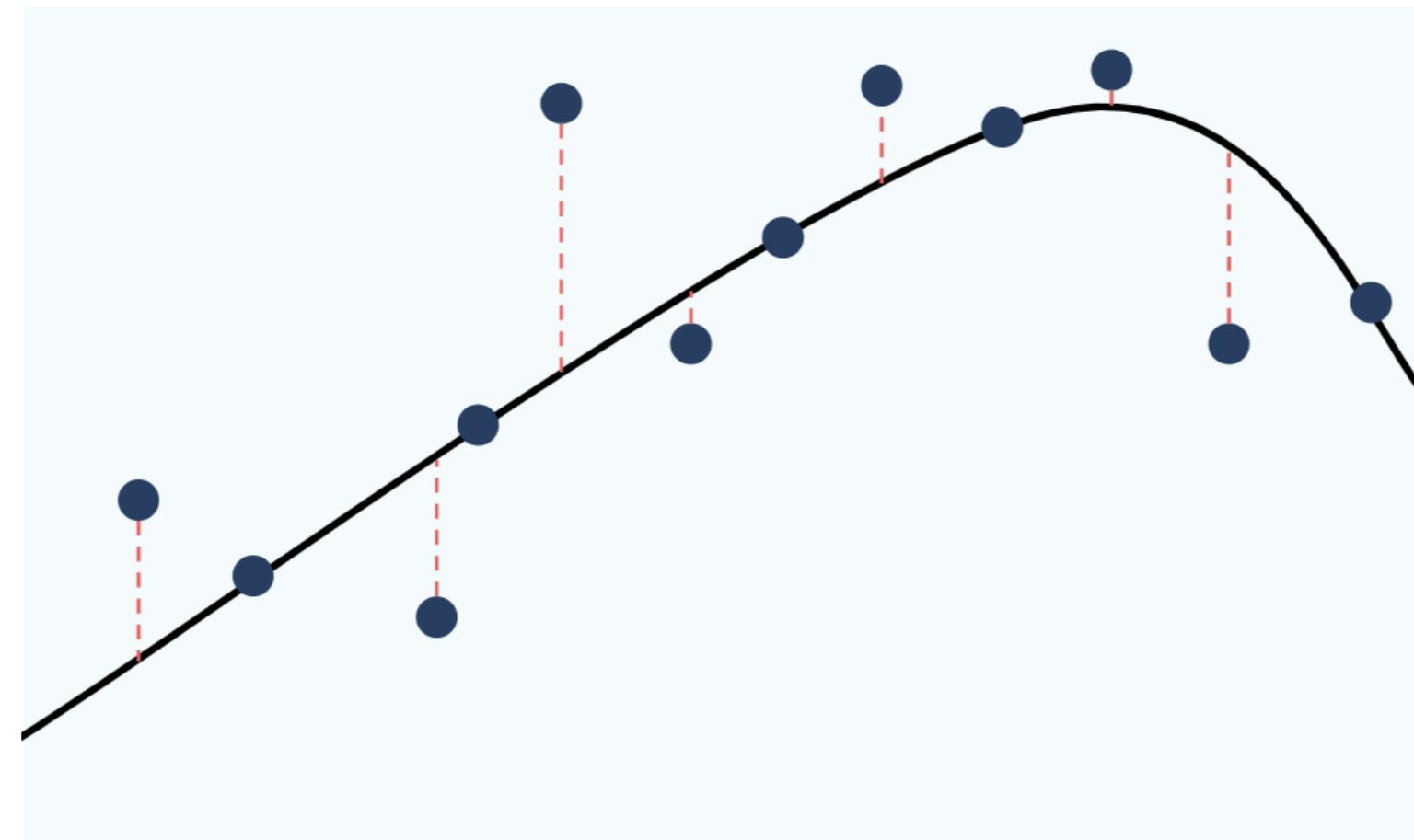
Spam filter:

- Rather send spam to inbox than send real emails to the spam folder

# Evaluating regression



# Evaluating regression



- Error = distance between point (actual value) and line (predicted value)
- Many ways calculate this. e.g, root mean square error

# Unsupervised learning



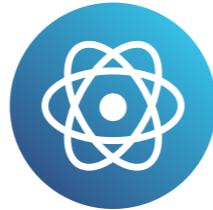
<sup>1</sup> <https://www.flickr.com/photos/micahdowty/8540188997>

# Let's practice!

MACHINE LEARNING FOR EVERYONE

# Improving performance

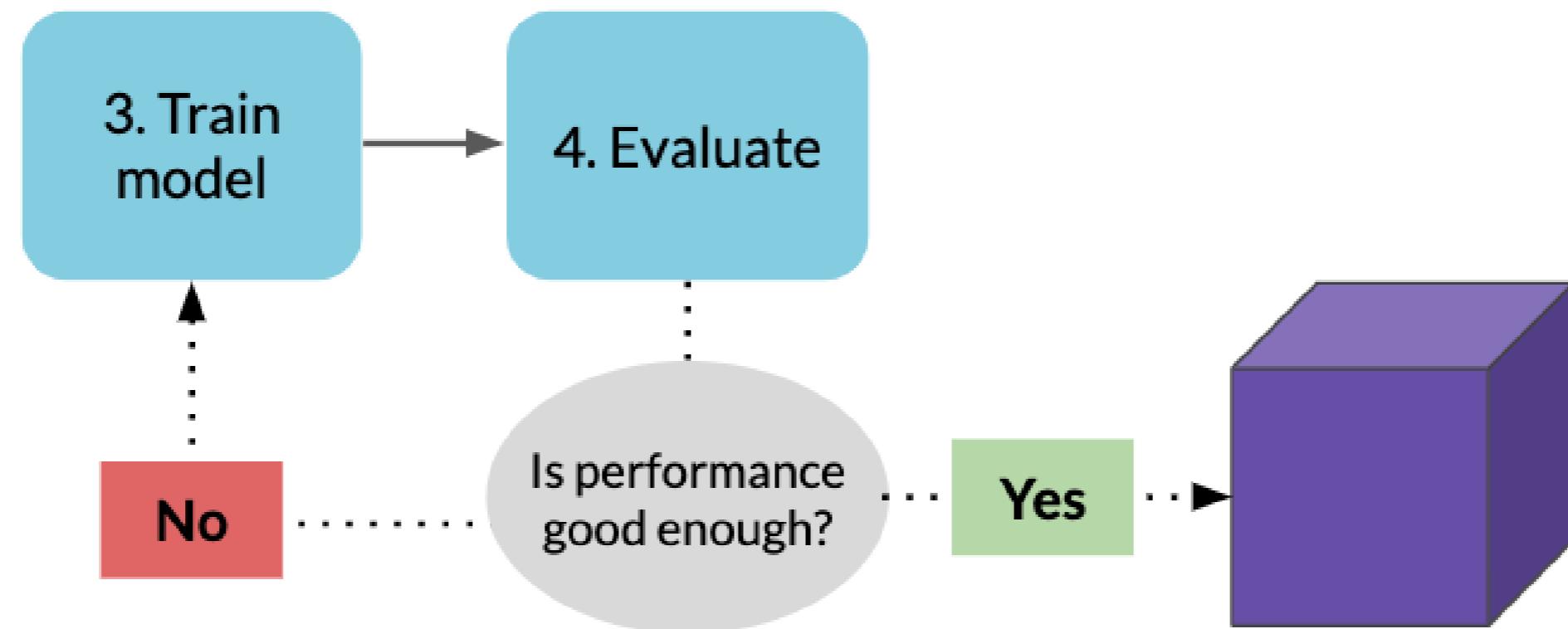
MACHINE LEARNING FOR EVERYONE



**Hadrien Lacroix**

Content Developer at DataCamp

# Machine learning workflow



# Several options

- Dimensionality reduction
- Hyperparameter tuning
- Ensemble methods

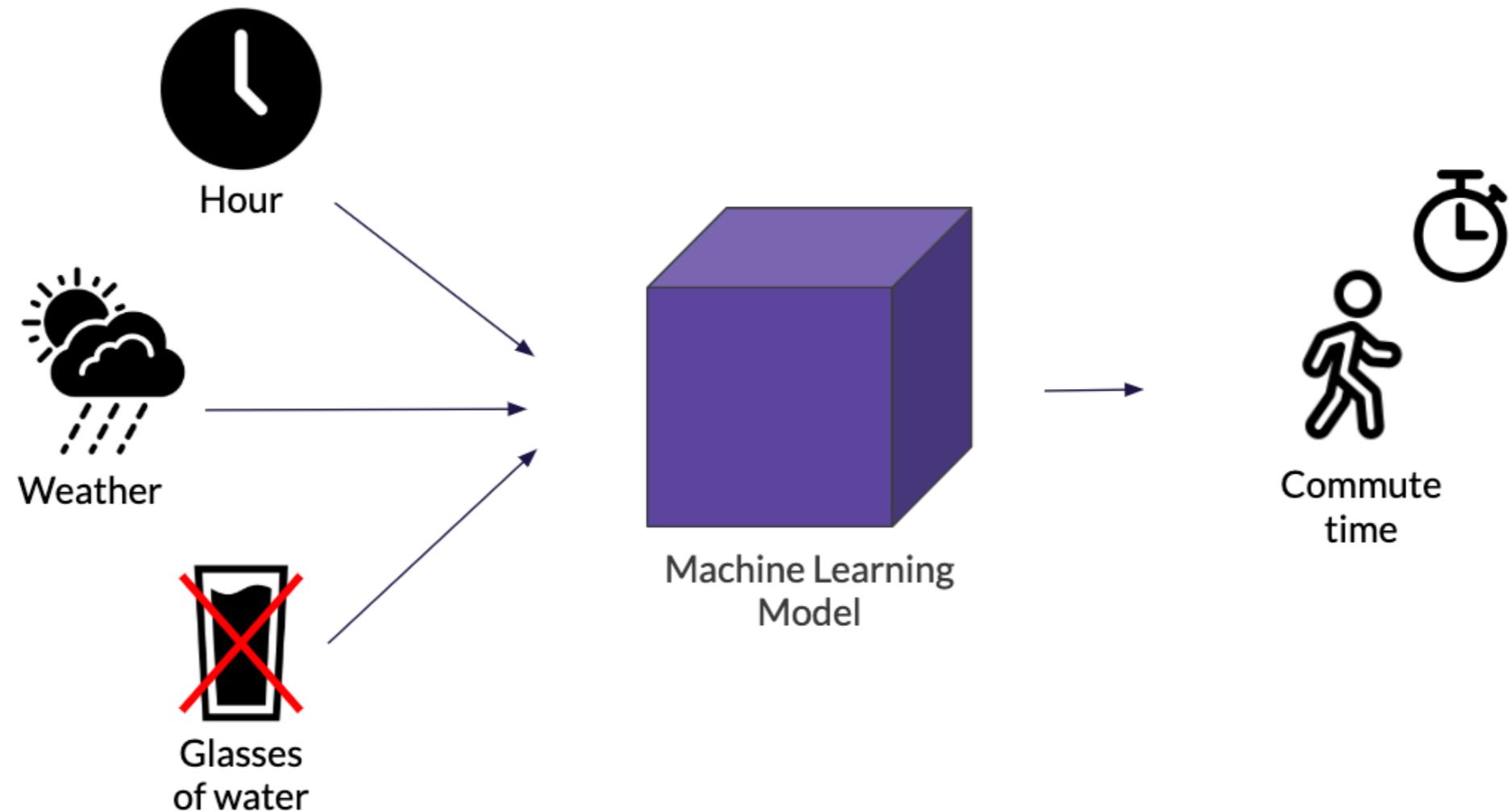
# Dimensionality reduction

Reducing the number of features



# Dimensionality reduction: example

Irrelevance: some features don't carry useful information



# Dimensionality reduction: example

Correlation: some features carry similar information

- Keep only one feature
  - e.g. *height* and *shoe size* --> *height*
- Collapse multiple features into one underlying feature
  - e.g. *height* and *weight* --> *Body Mass Index*

# Hyperparameter tuning



# Hyperparameter tuning



# Hyperparameter tuning



# Hyperparameter tuning



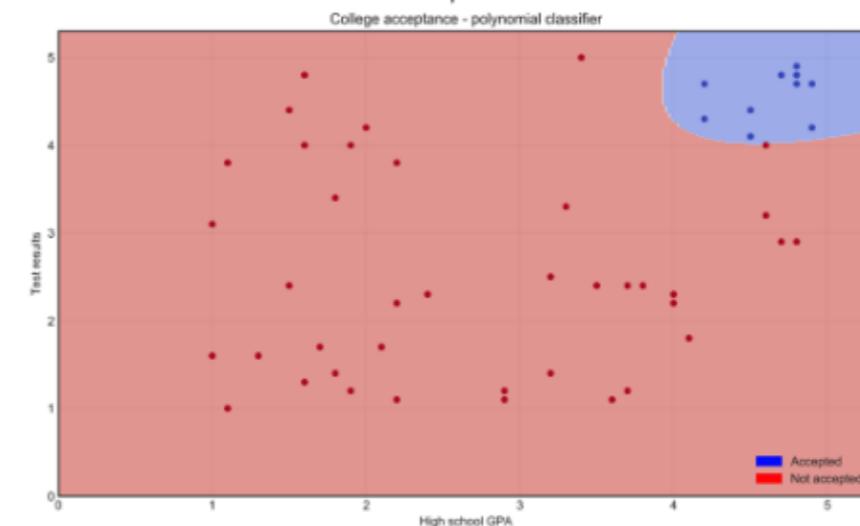
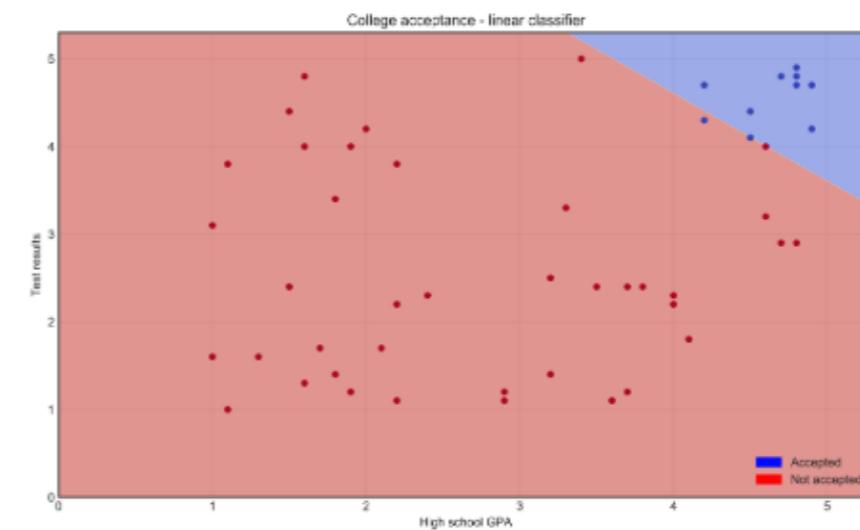
# Hyperparameter tuning



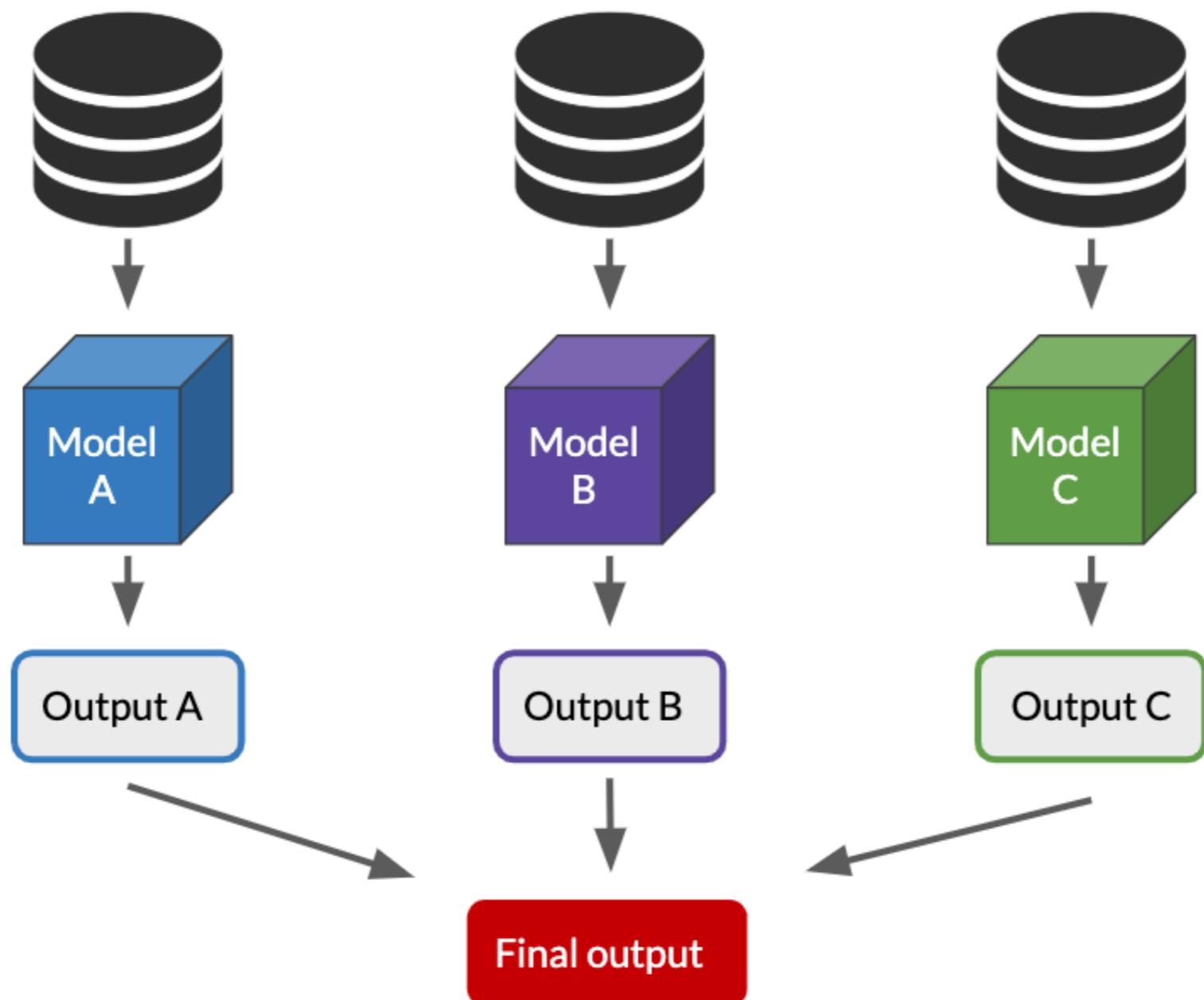
# Hyperparameter tuning: example

SVM algorithm hyperparameters:

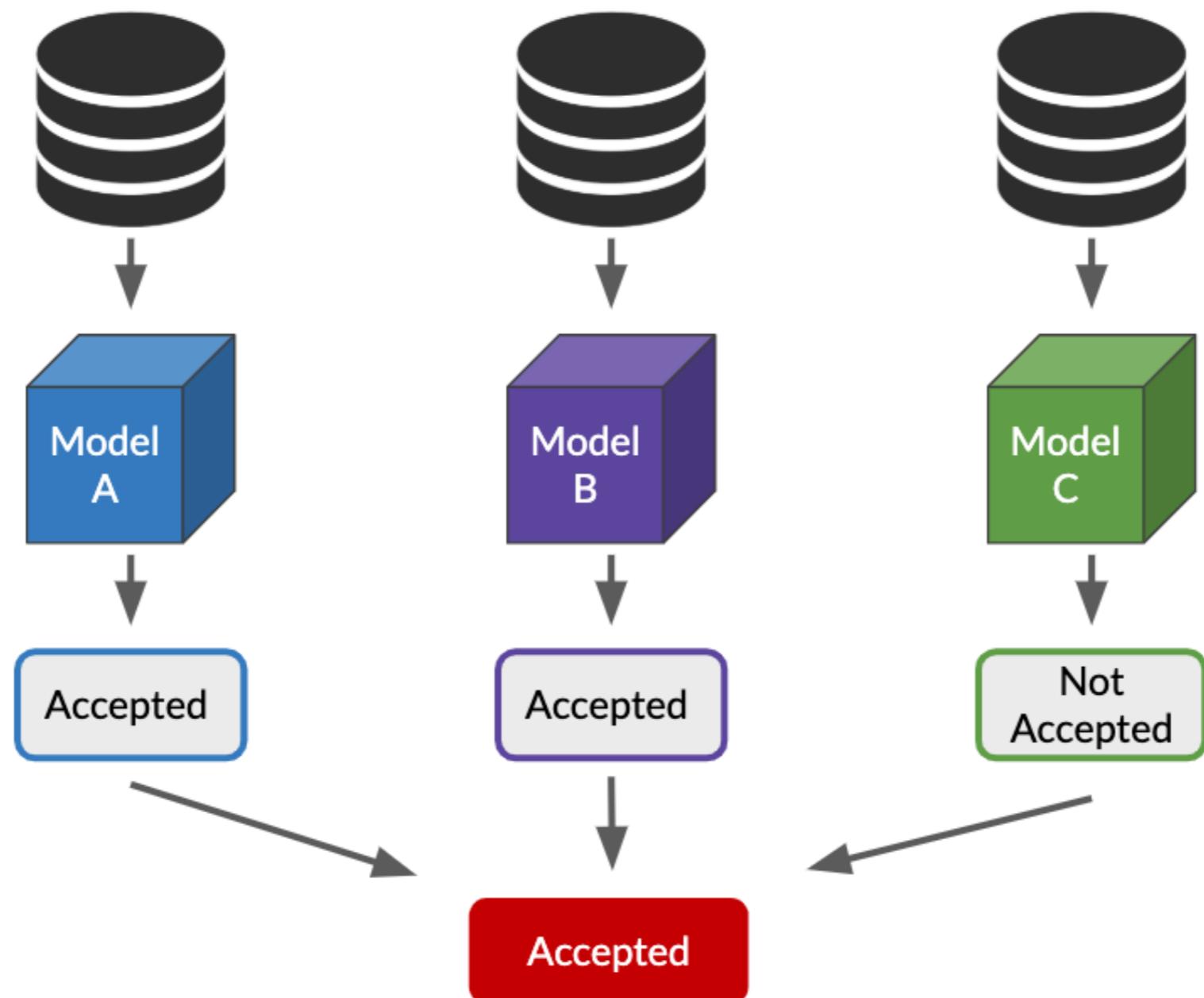
- kernel : "linear" --> "poly"
- C
- degree
- gamma
- shrinking
- coef0
- tol
- ...



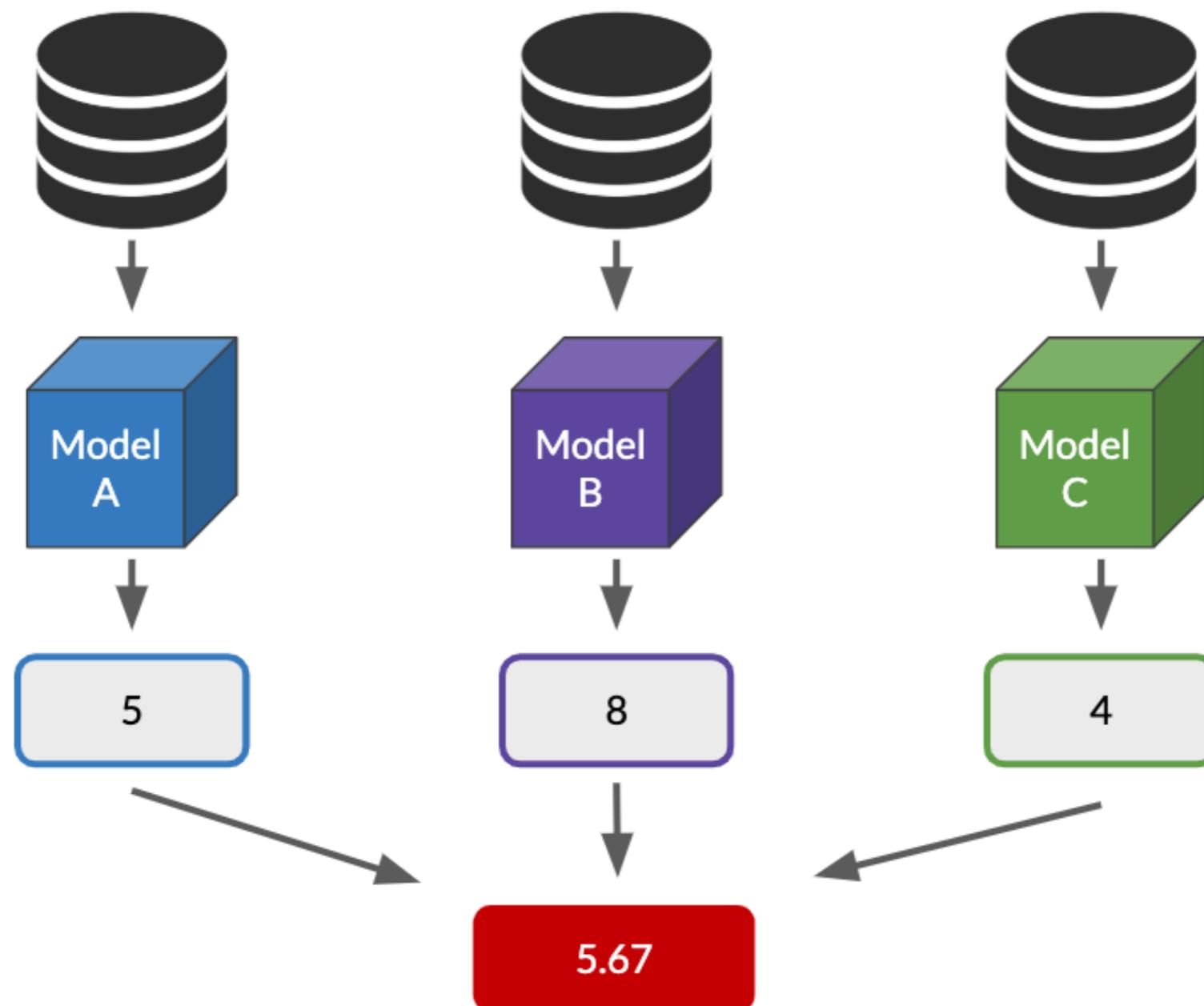
# Ensemble methods



# Ensemble methods: classification



# Ensemble methods: regression



# Let's practice!

MACHINE LEARNING FOR EVERYONE