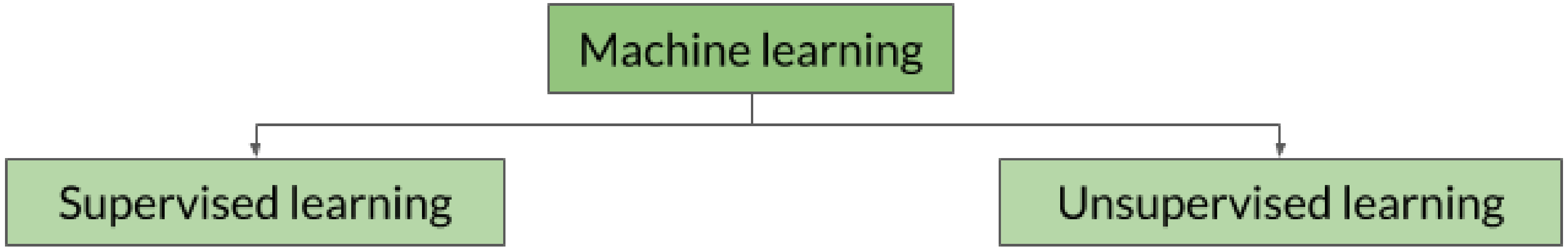
Supervised learning

UNDERSTANDING MACHINE LEARNING

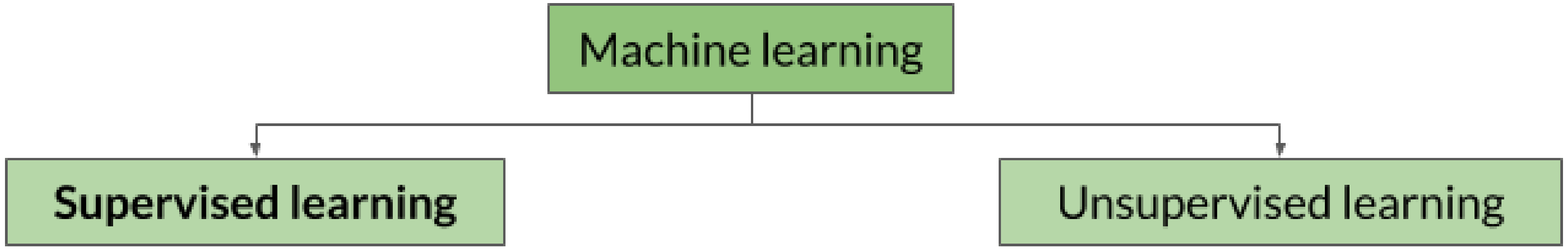
Hadrien Lacroix
Content Developer at DataCamp

Modeling

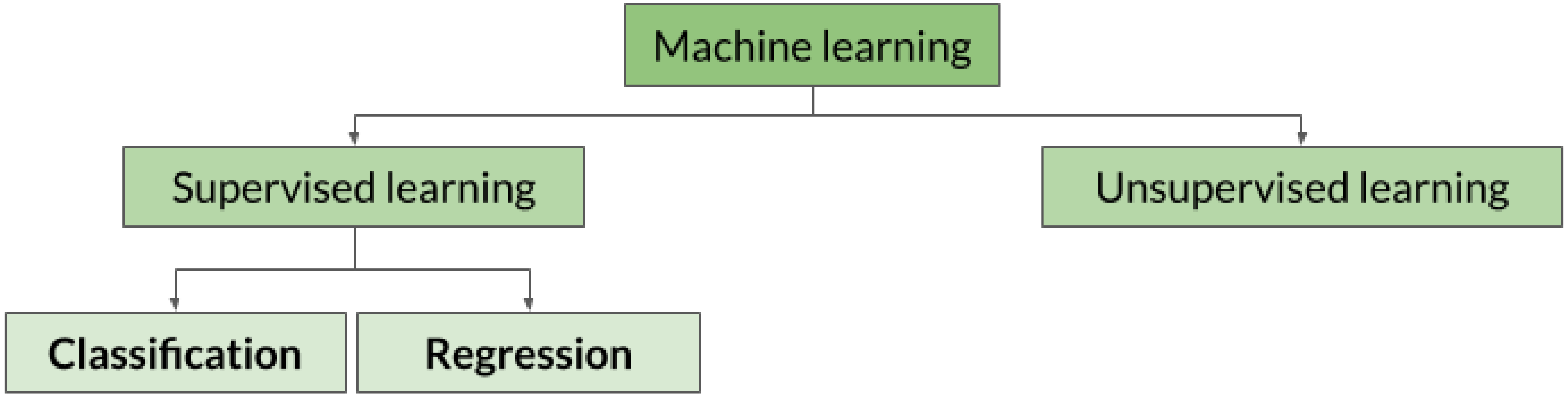
Types


Machine learning
Supervised learning
Unsupervised learning

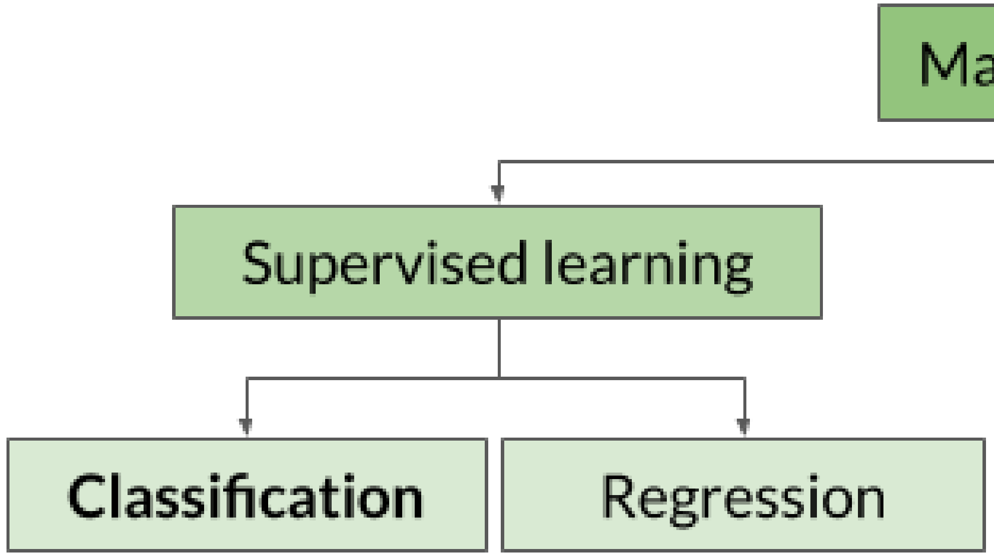
What is supervised learning?



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	...	...	...	...

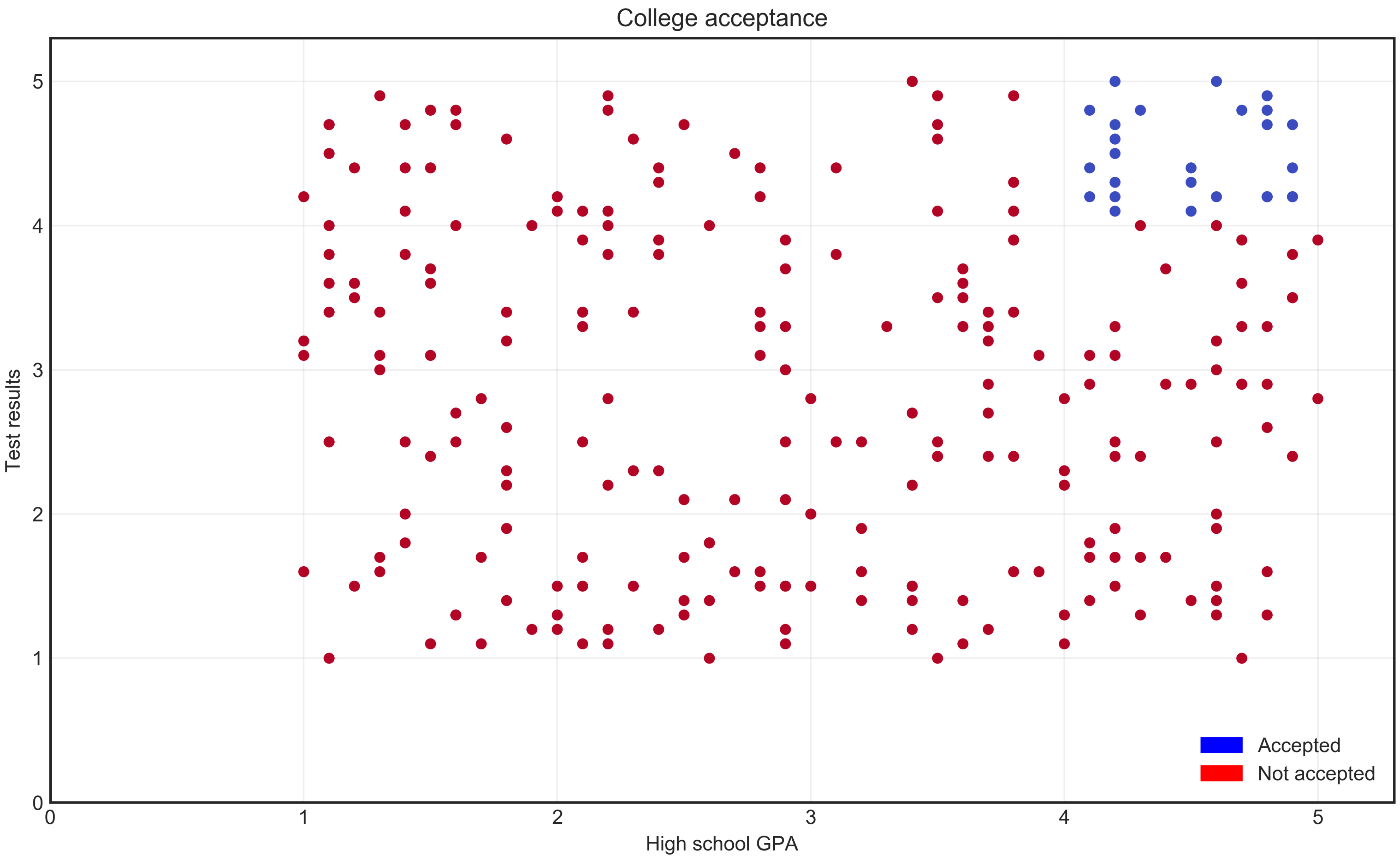
Features

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
...	...	...	...

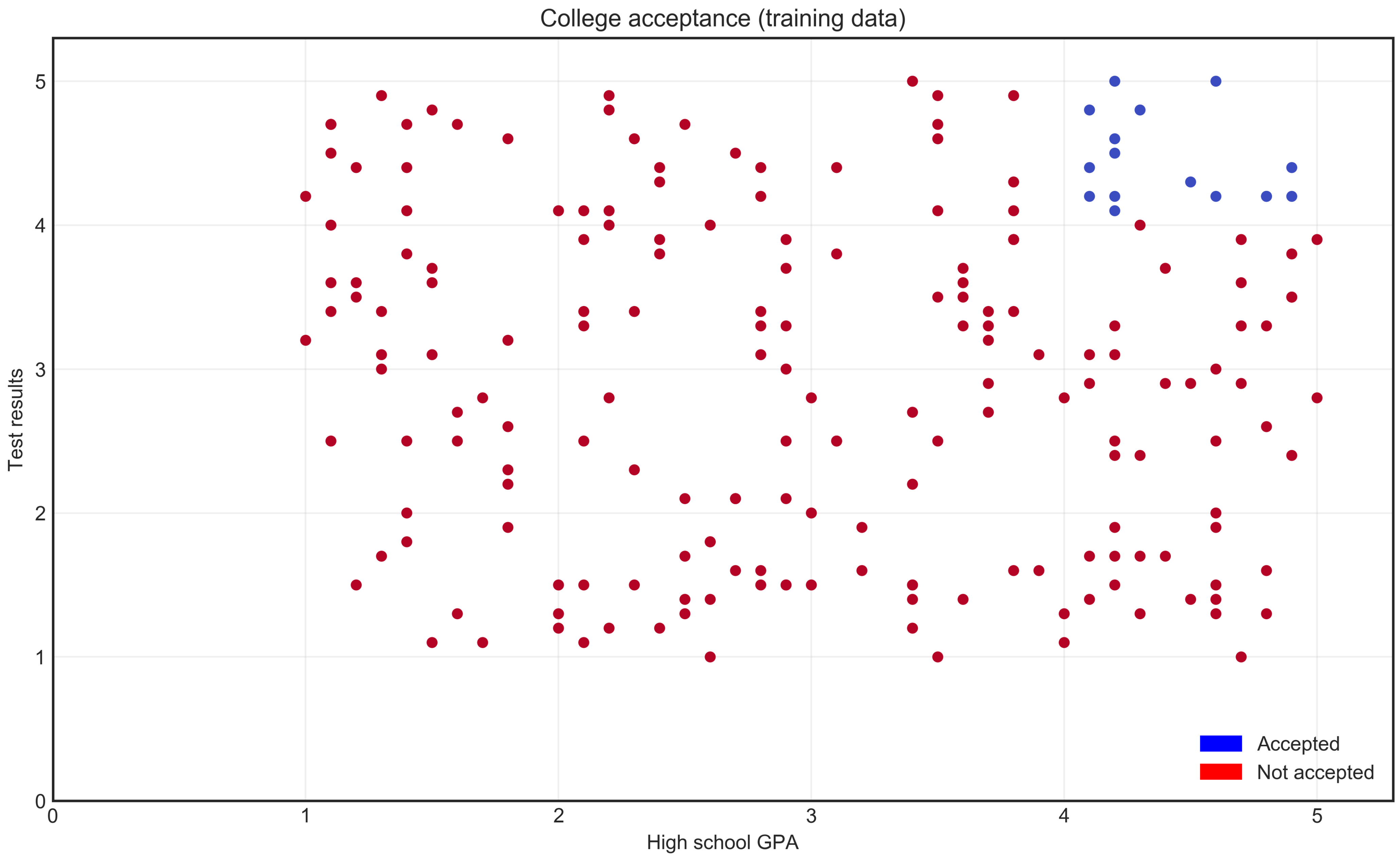
Target

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
...	...	...	...

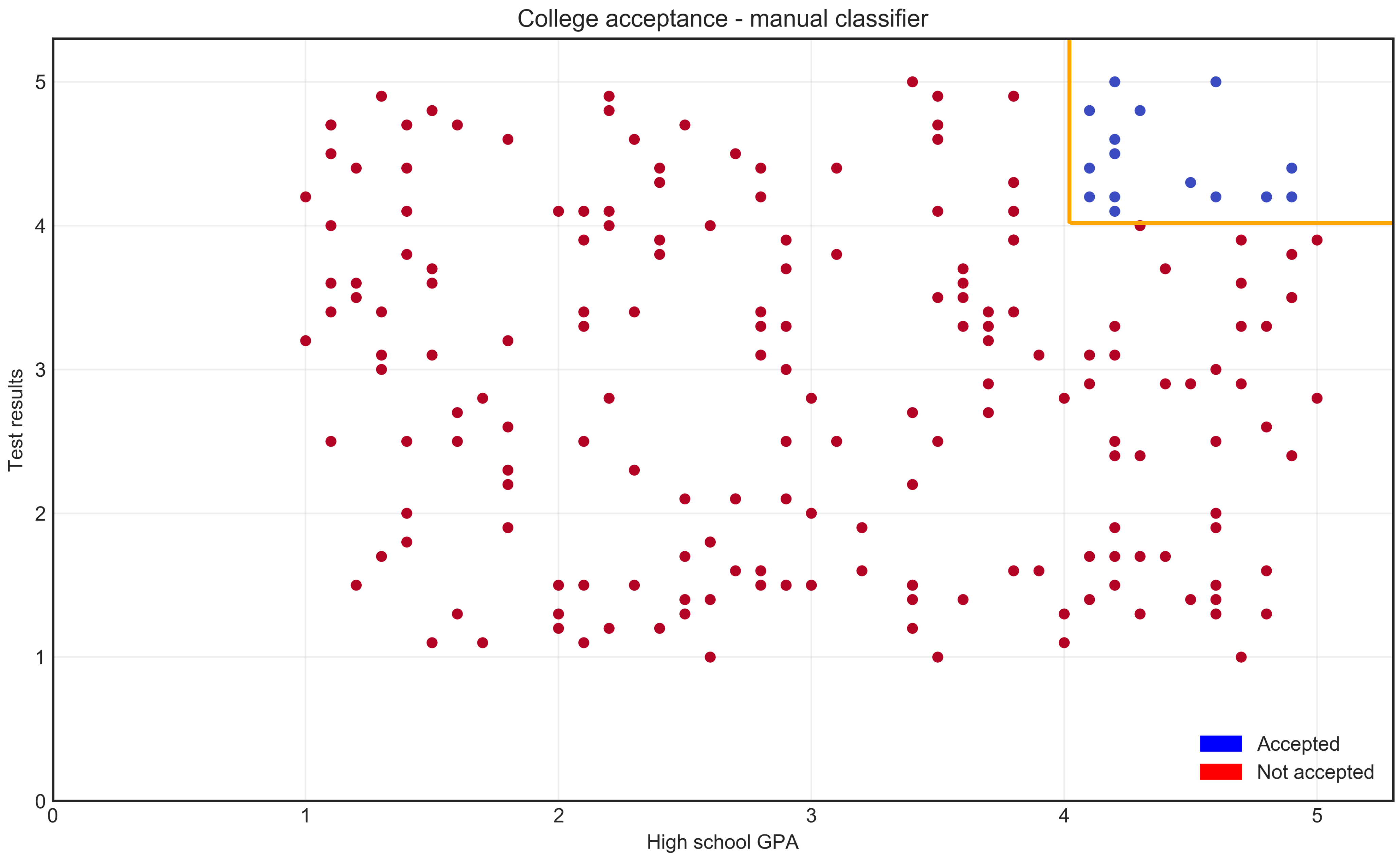
Graphing our data



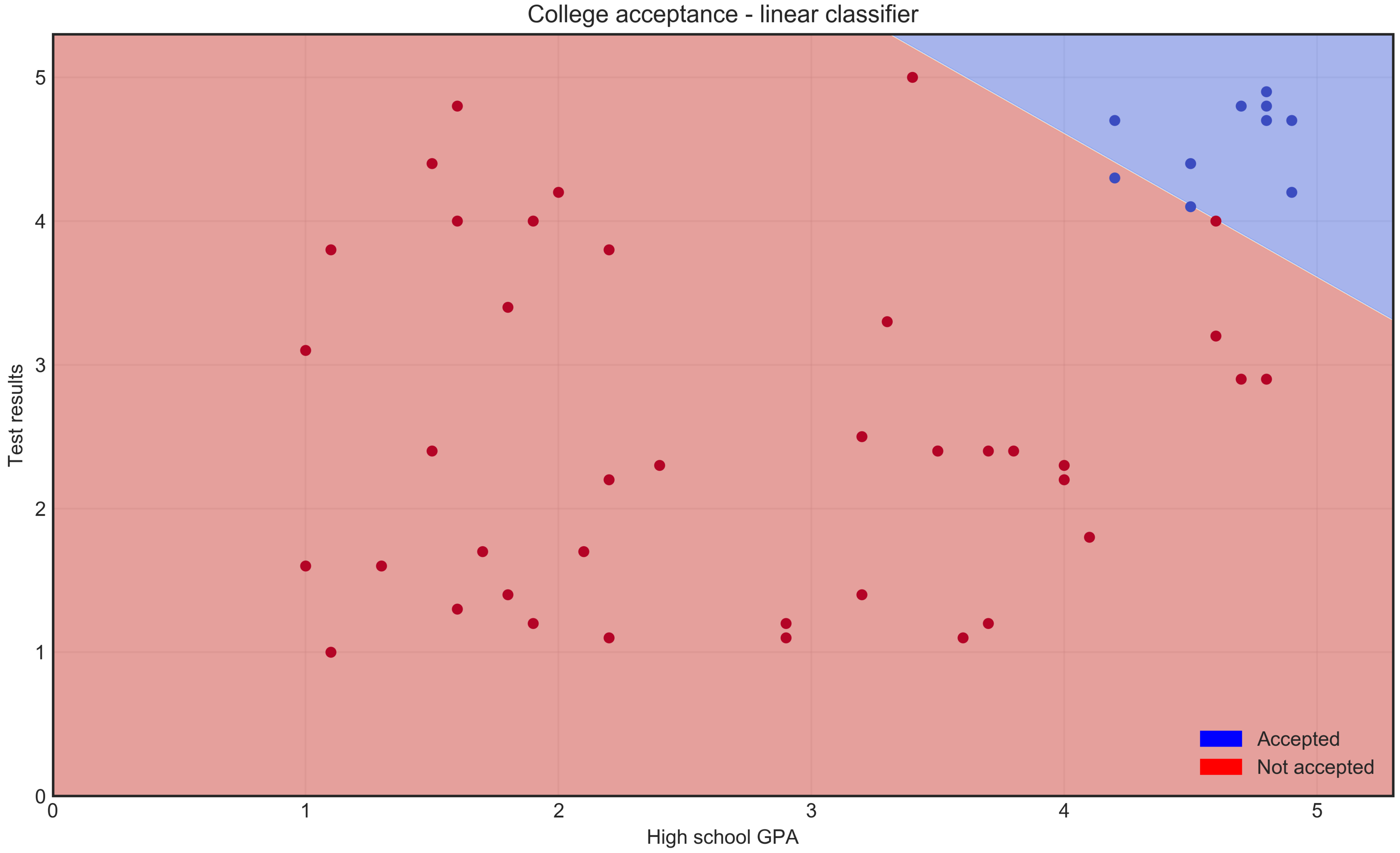
Splitting data



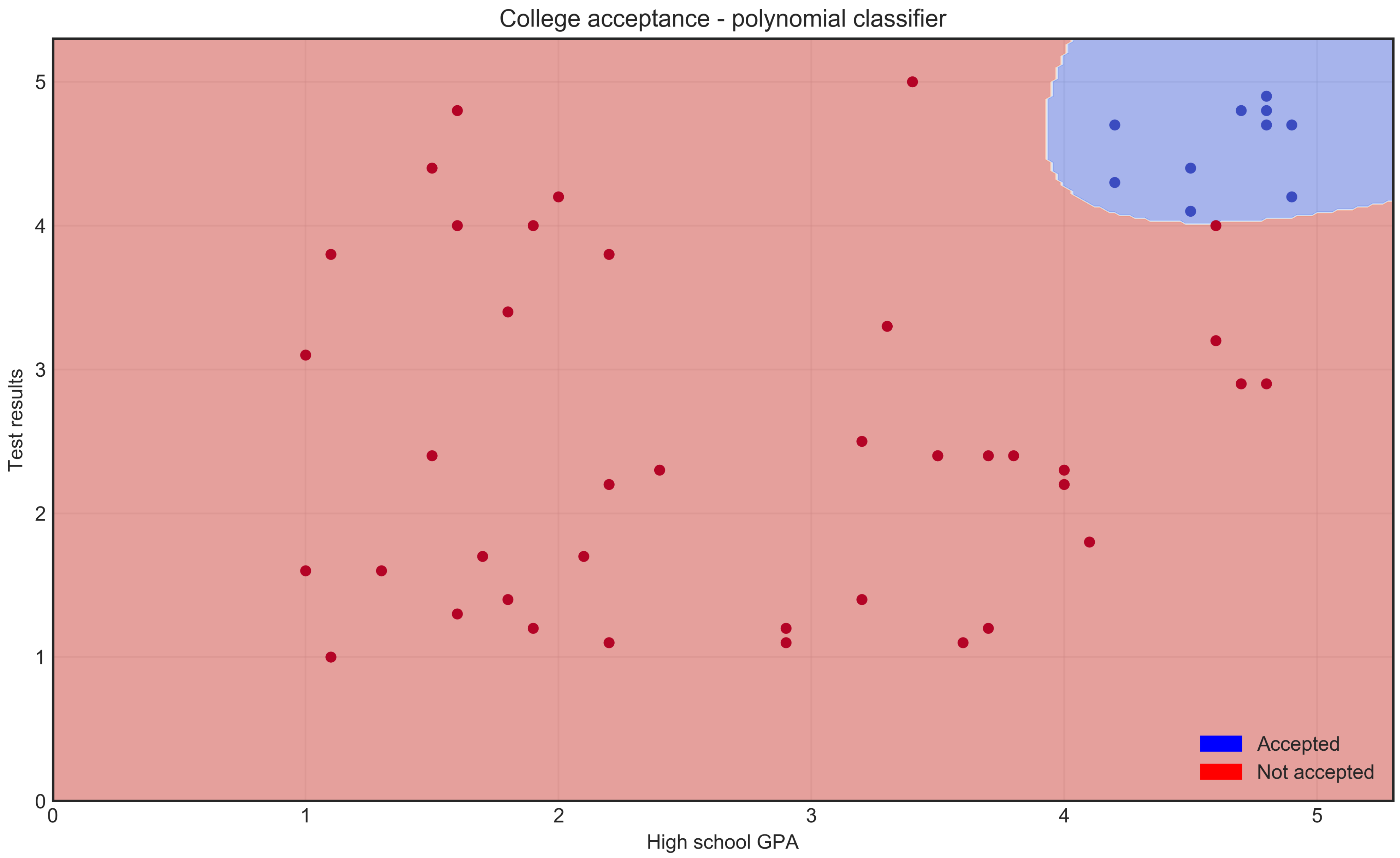
Manual classifier



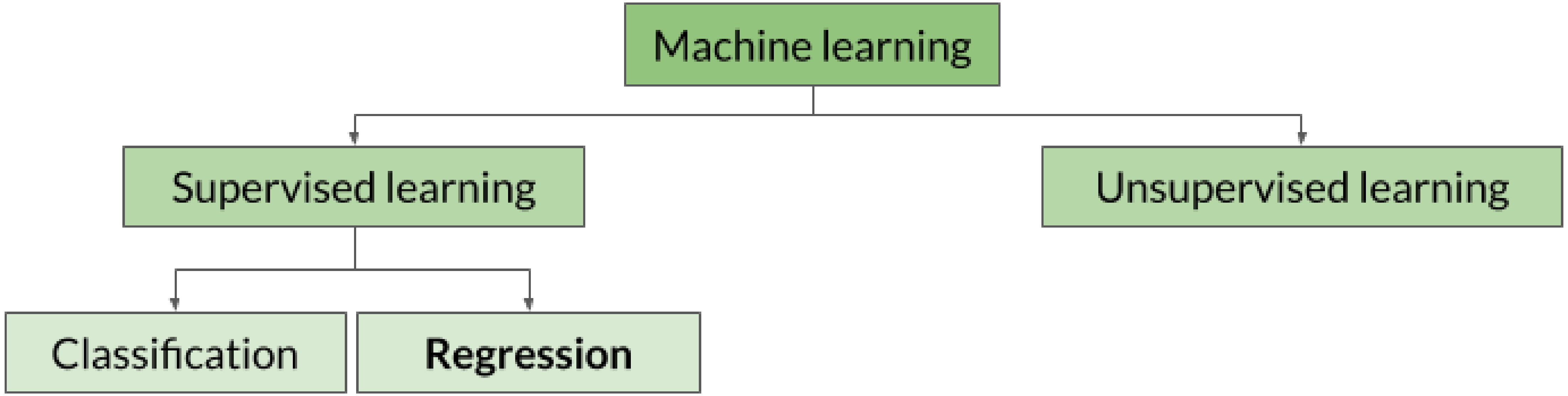
Support vector machine - linear classifier



Support vector machine - polynomial classifier



Regression


Machine learning
Supervised learning
Unsupervised learning
Classification
Regression

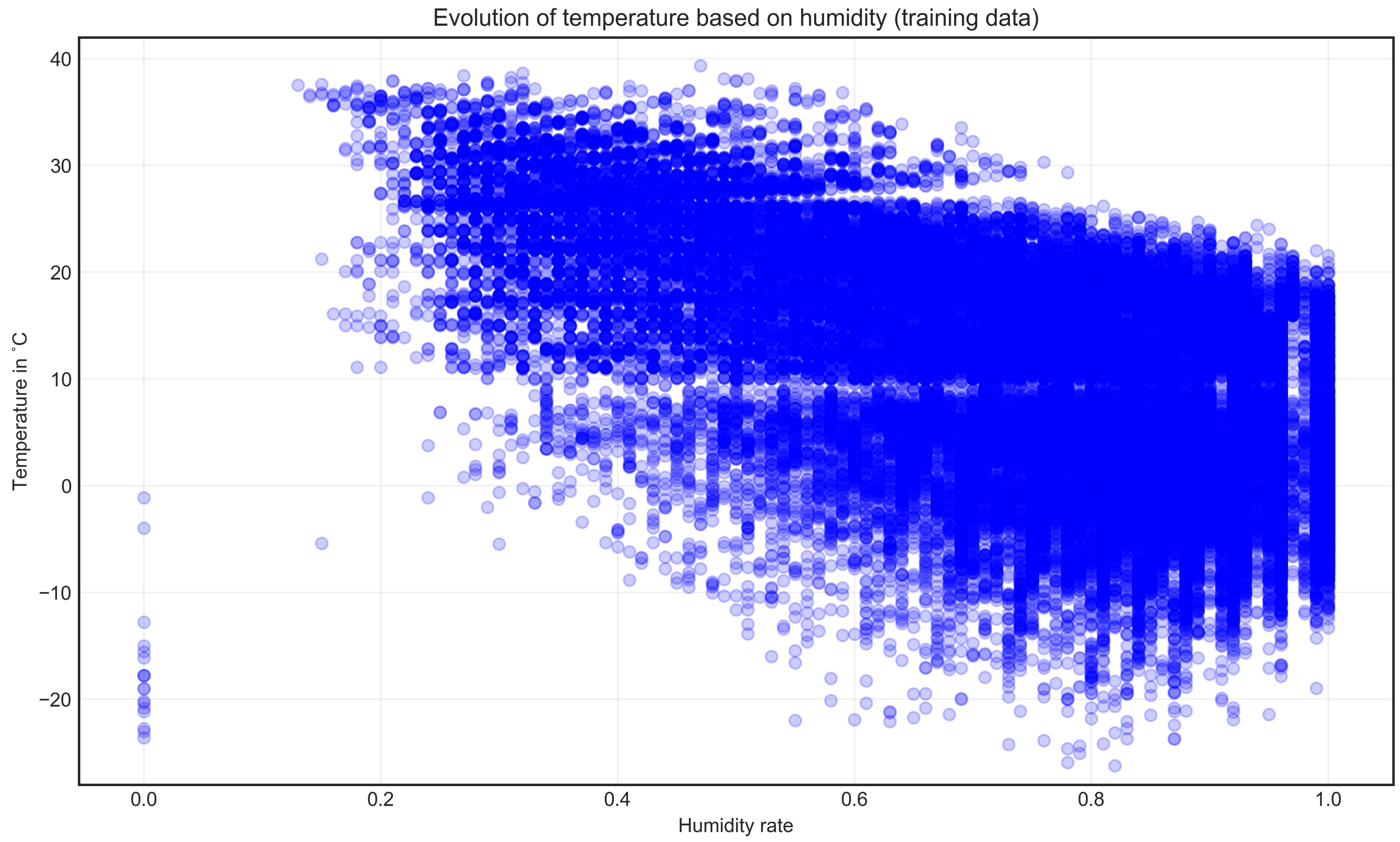
Regression

- **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?

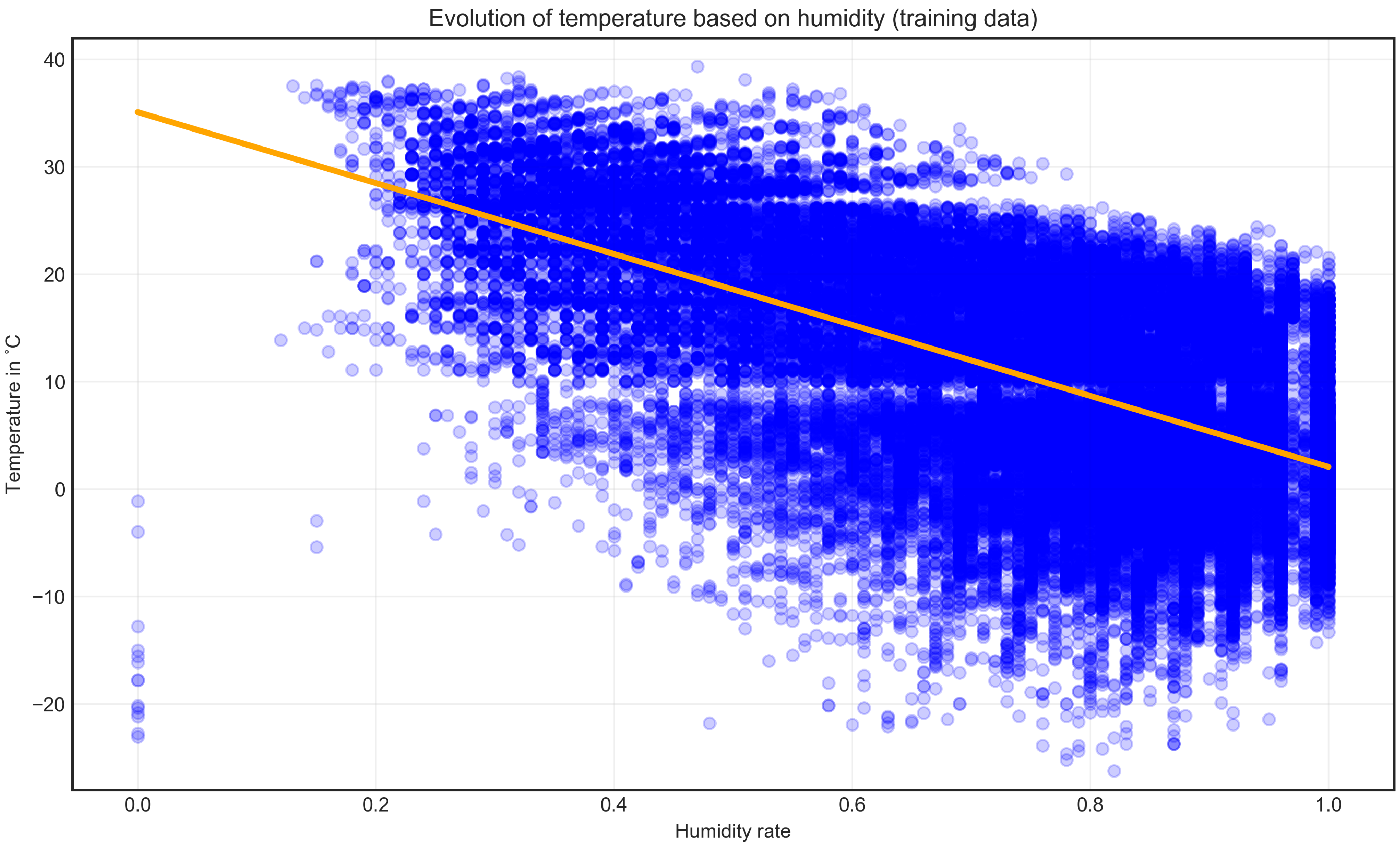
Predicting temperature

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
...	...	...

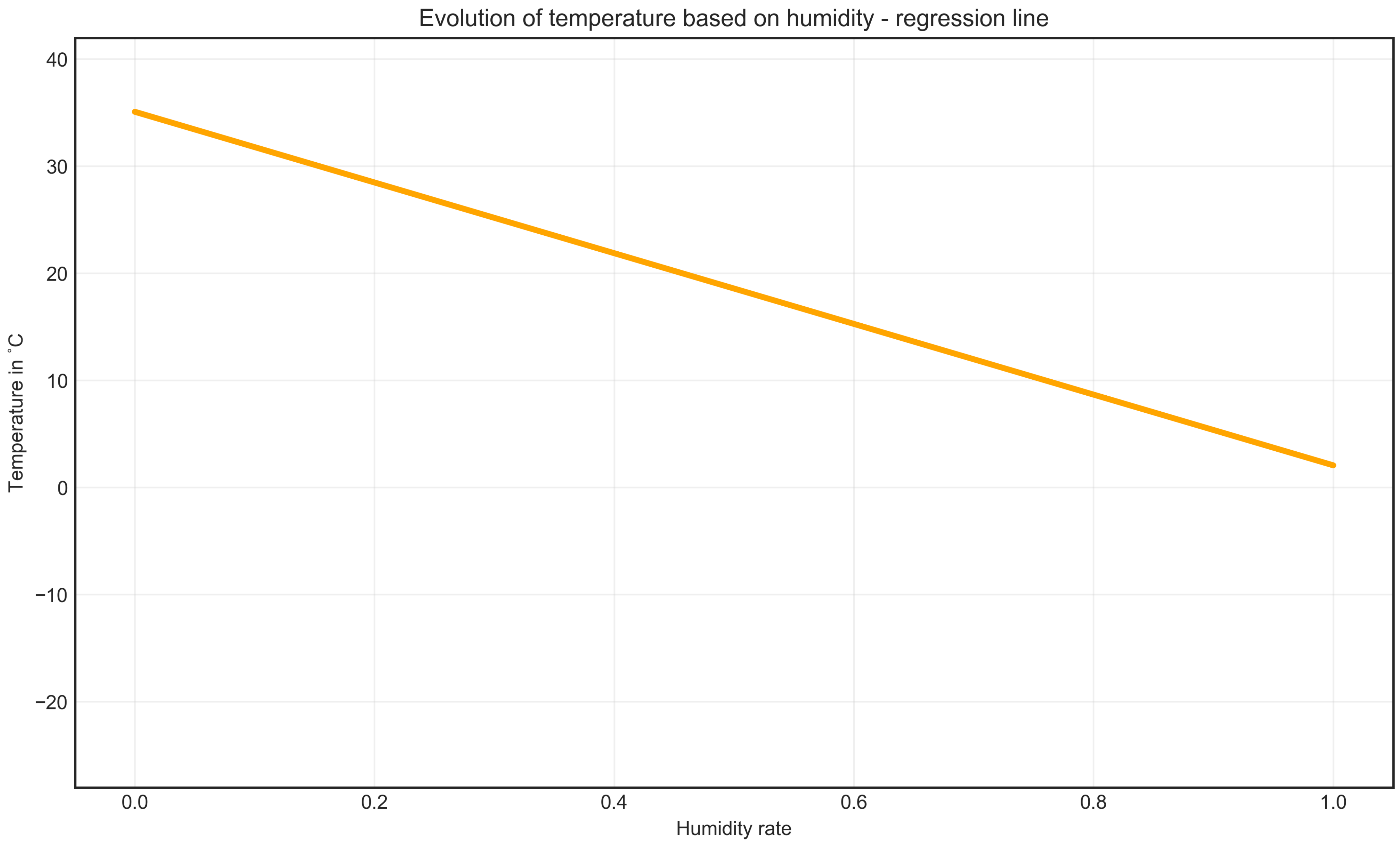
Training data



Linear regression

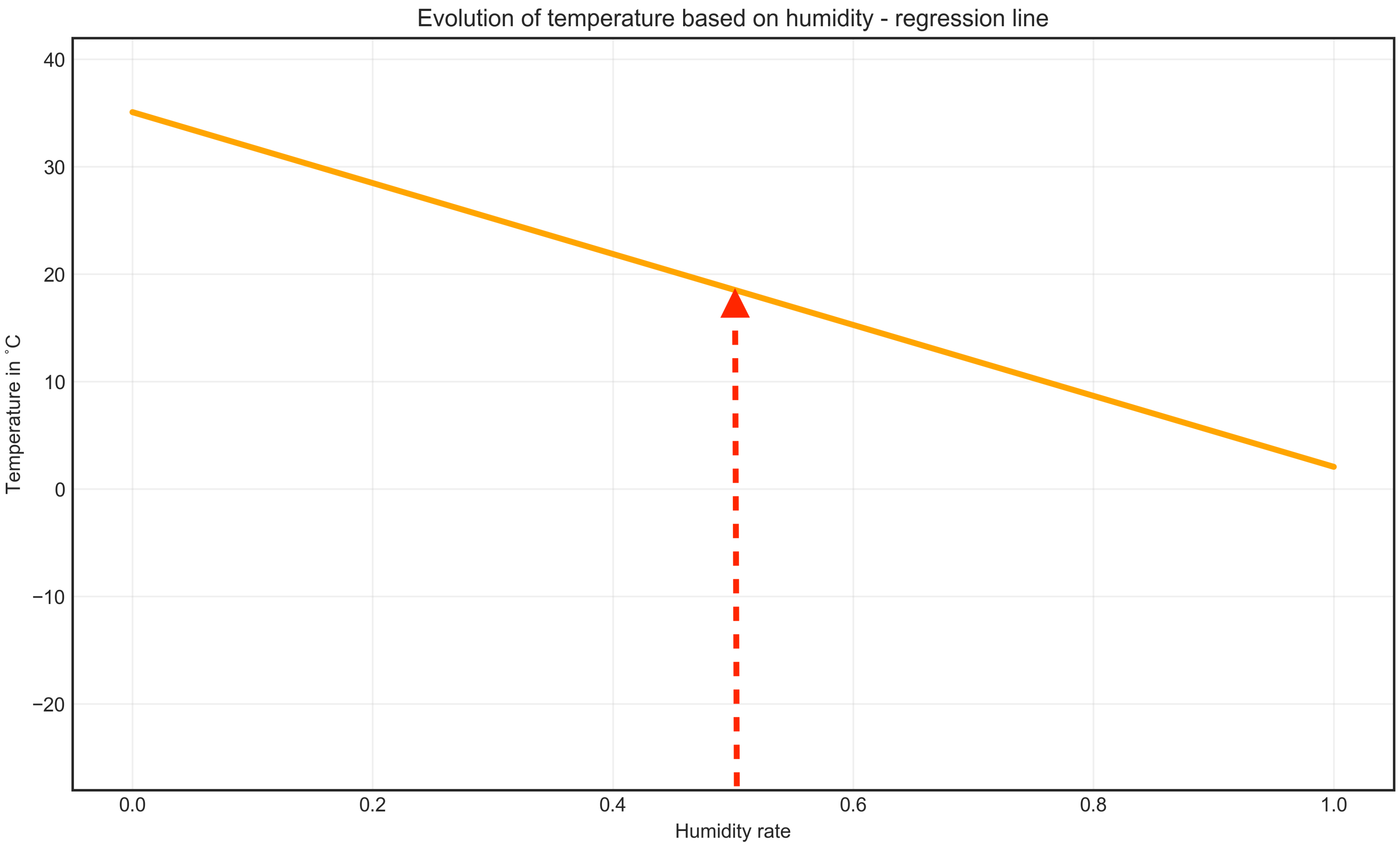


Model

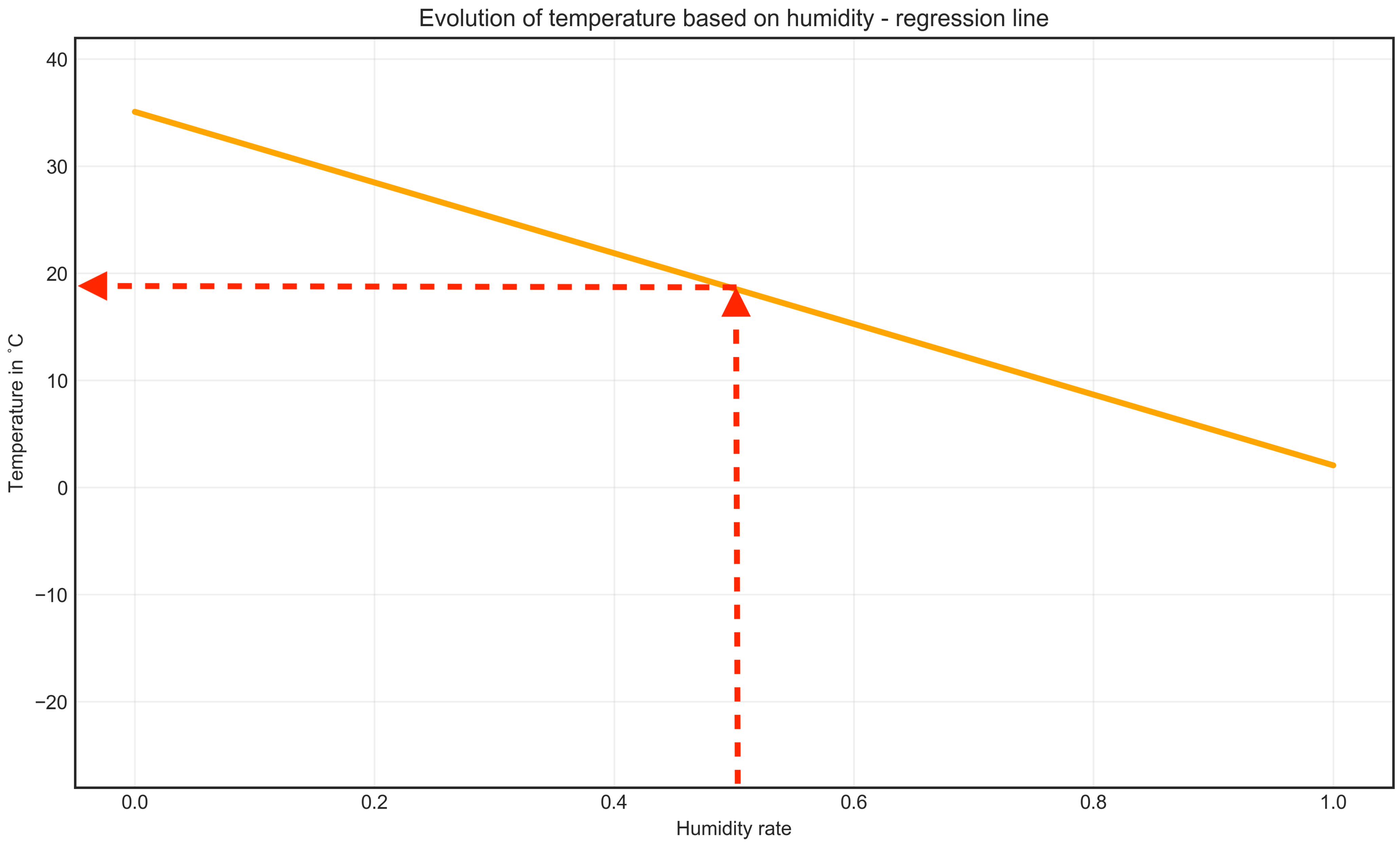


Evolution of temperature based on humidity - regression line
Temperature in °C
40
30
20
10
0
−10
−20
0.0
0.2
0.4
0.6
0.8
1.0
Humidity rate

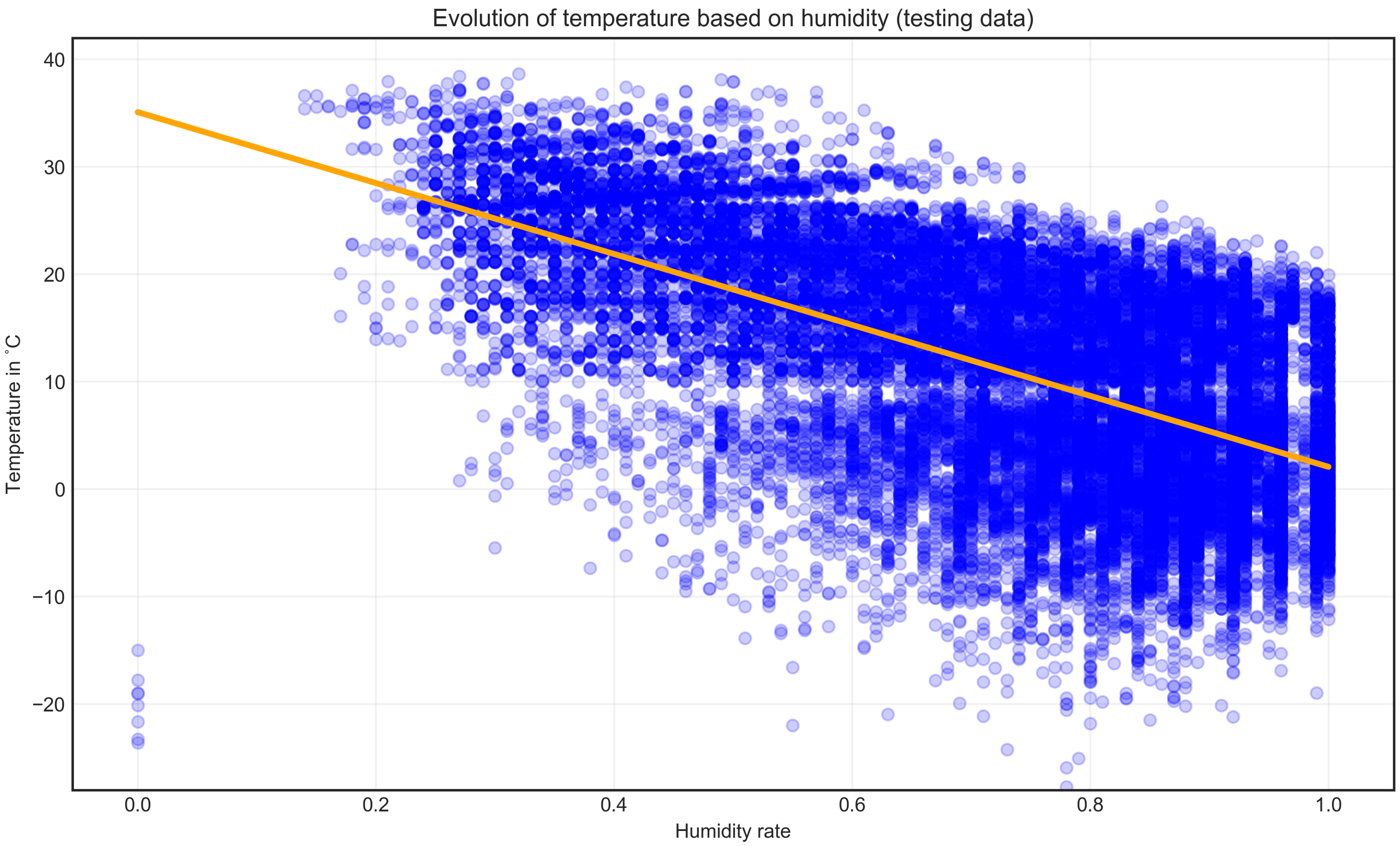
Given humidity...



...find temperature



Testing data



Classification vs regression

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

Let's practice!

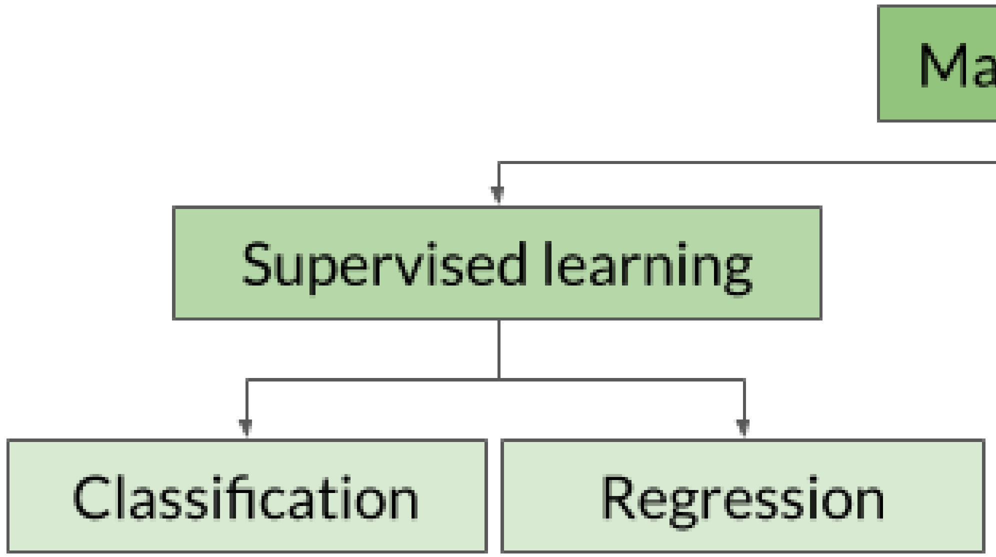
UNDERSTANDING MACHINE LEARNING

Unsupervised learning

UNDERSTANDING MACHINE LEARNING

Hadrien Lacroix
Content Developer at DataCamp

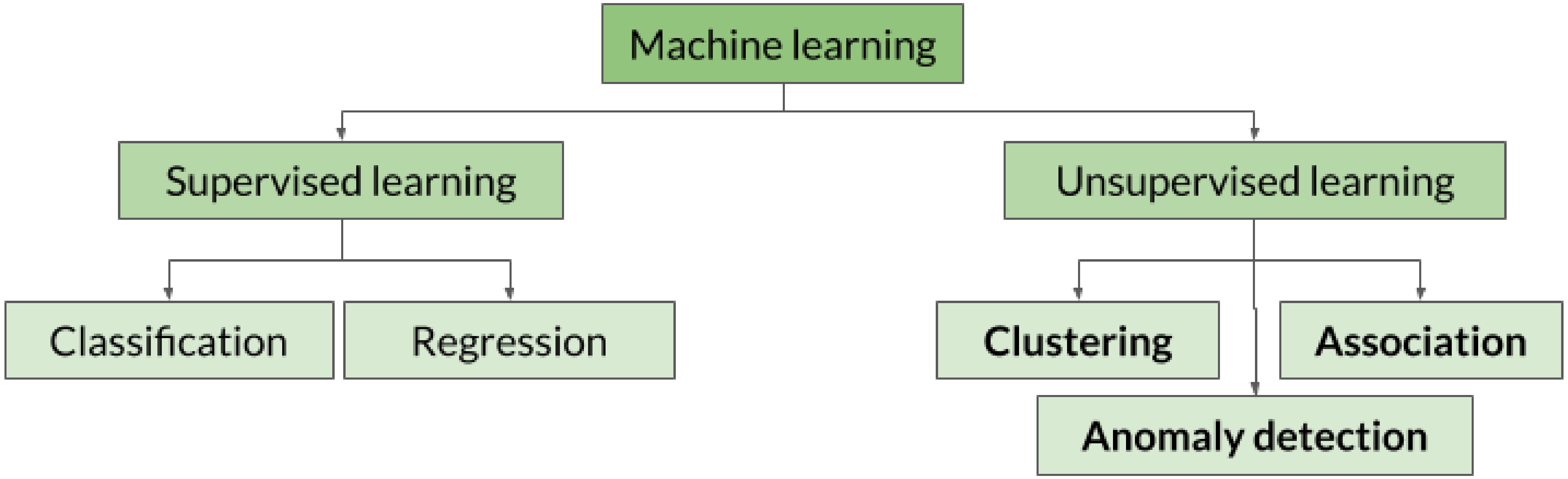
Unsupervised learning



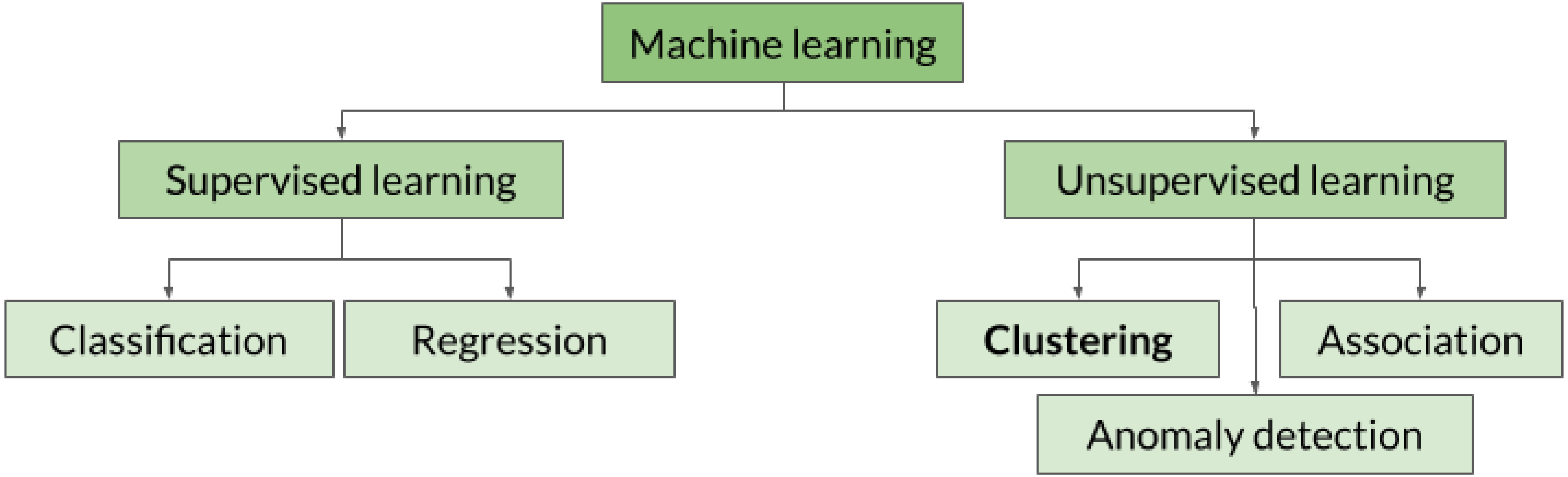
Unsupervised learning

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

Applications

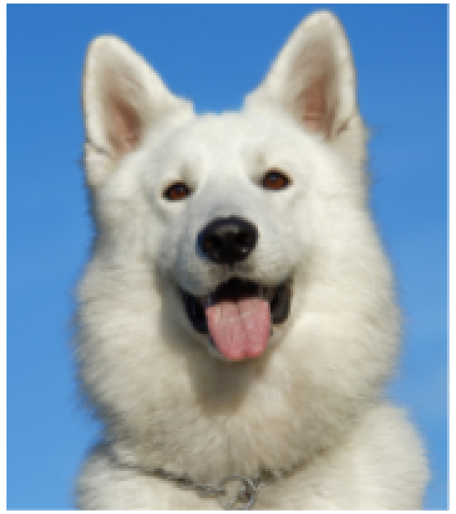


Clustering


Machine learning
Supervised learning
Unsupervised learning
Classification
Regression
Clustering
Association
Anomaly detection

Clustering example

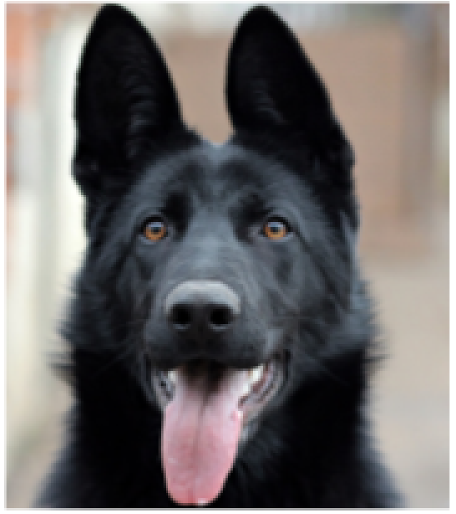
White Swiss Shepherd



Brown Japanese Bobtail

Brown Akita

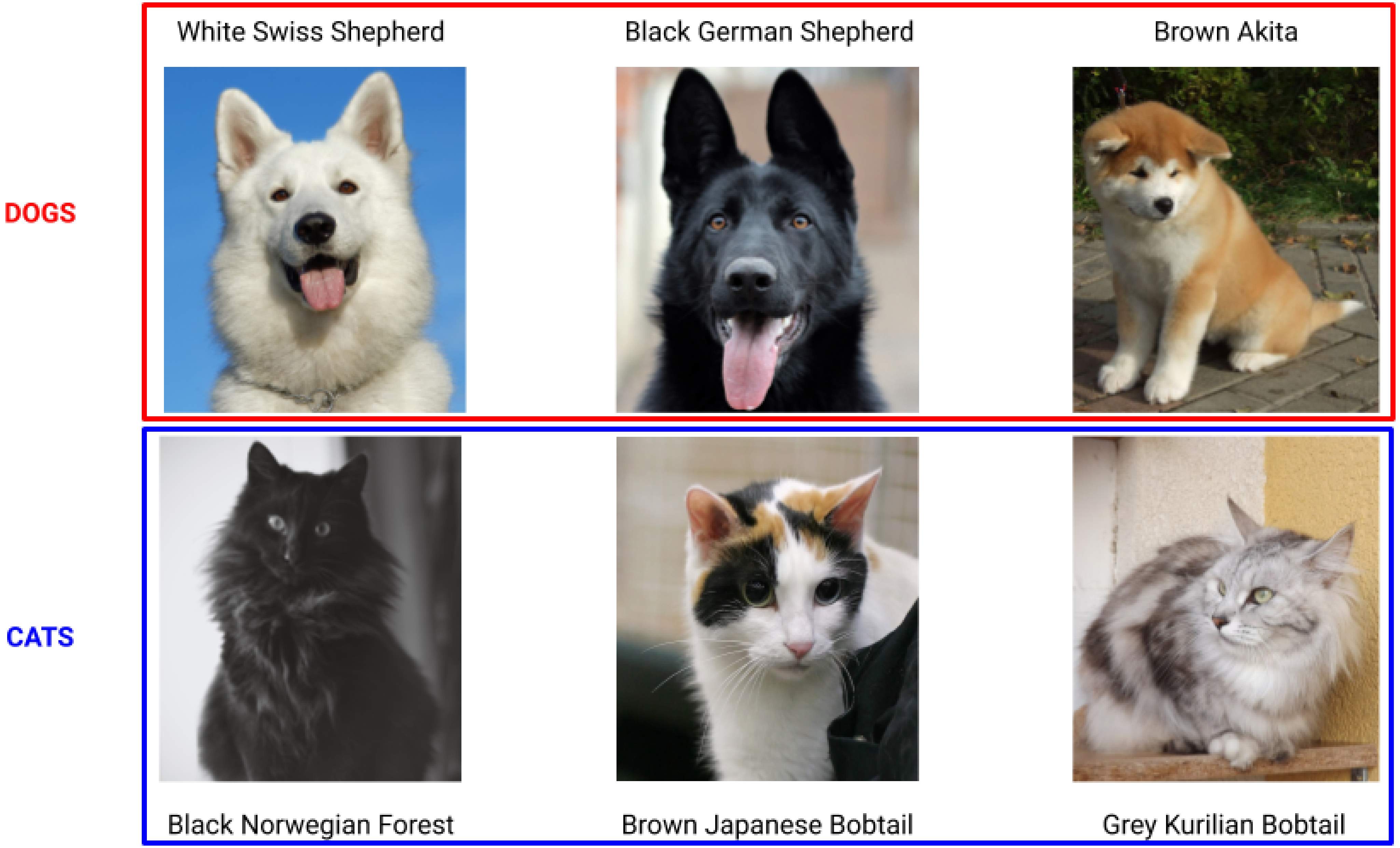
Black Norwegian Forest



Black German Shepherd

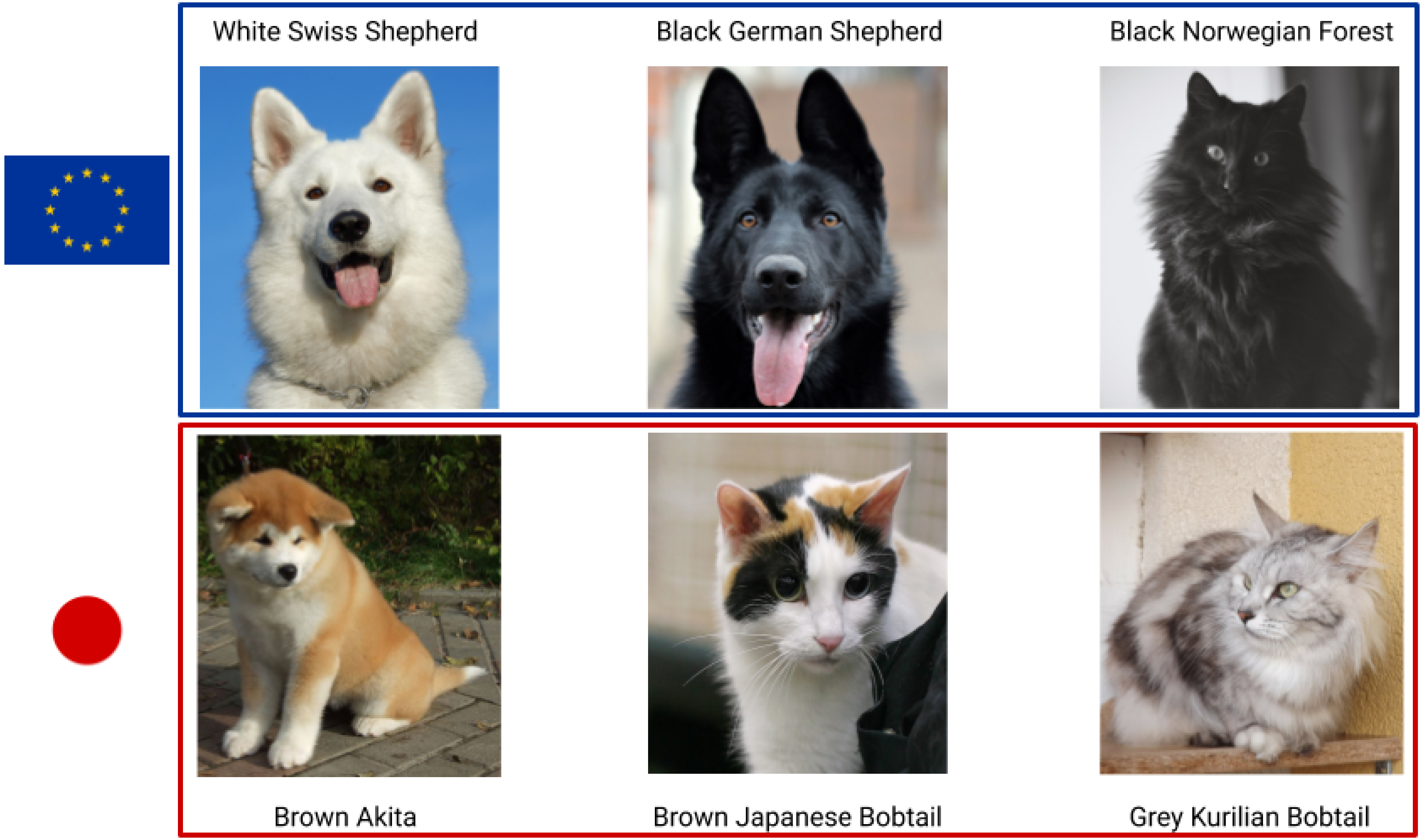
Grey Kurilian Bobtail

Species cluster



Color cluster

Origin cluster



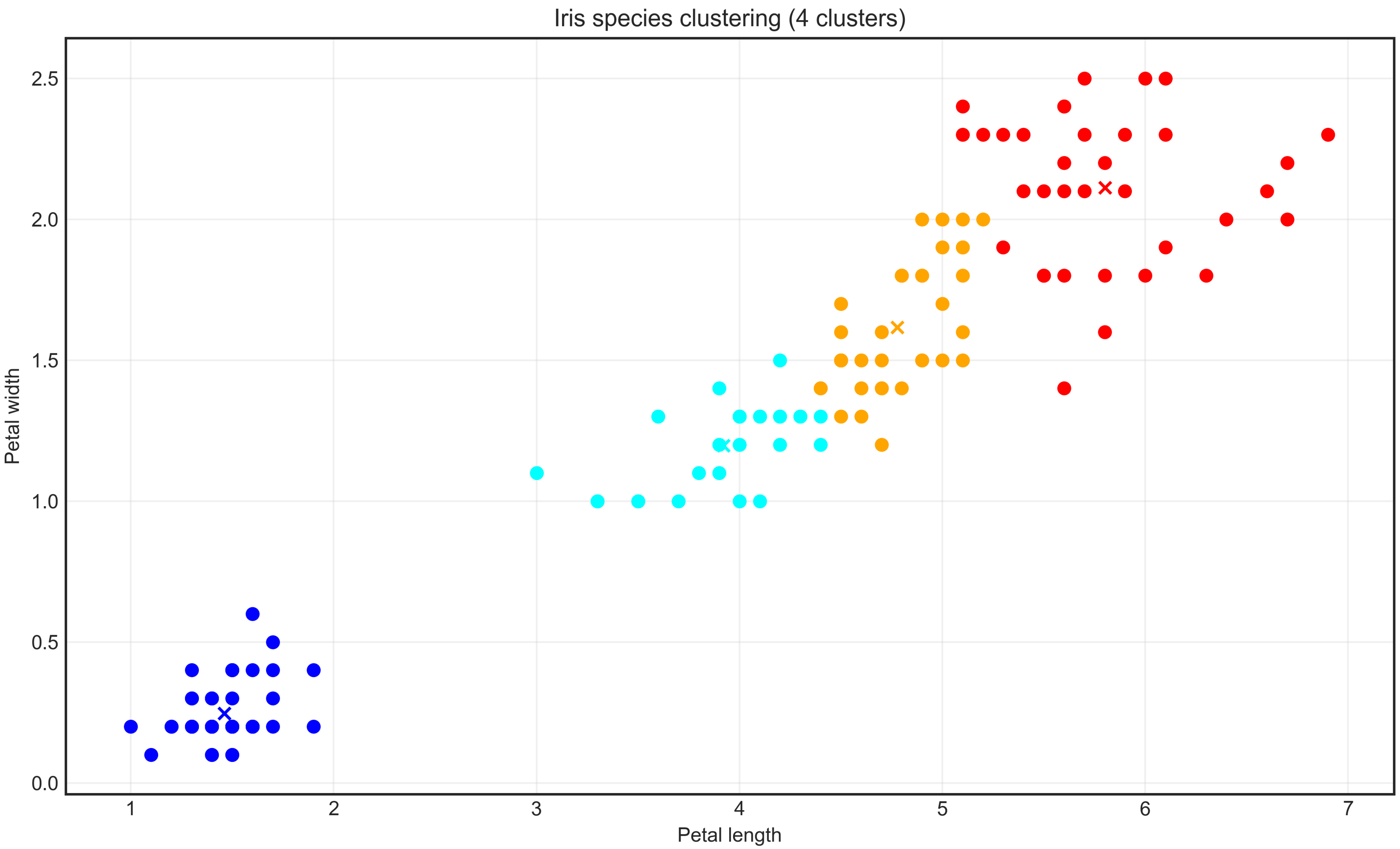
Clustering models

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

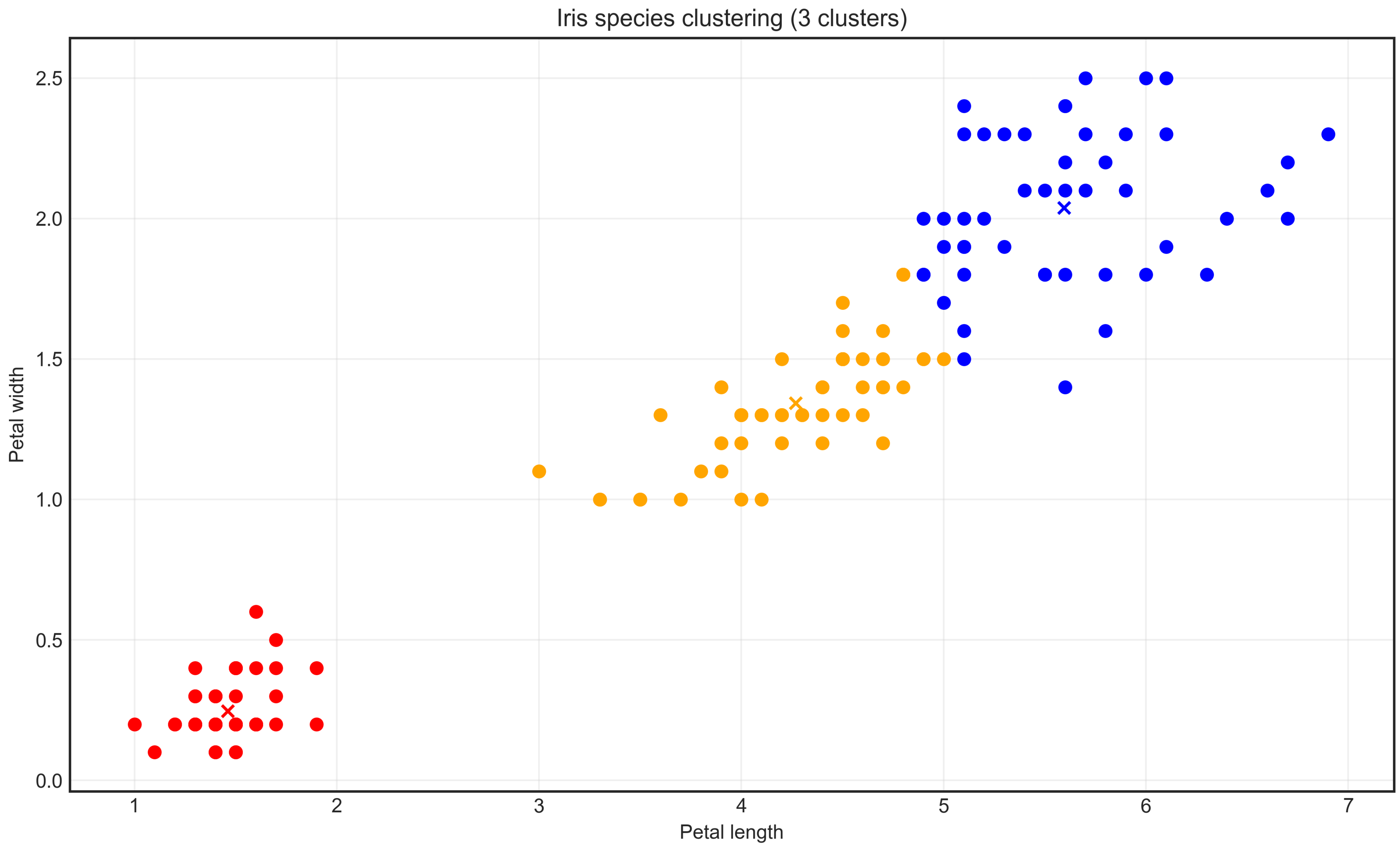
Iris table

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

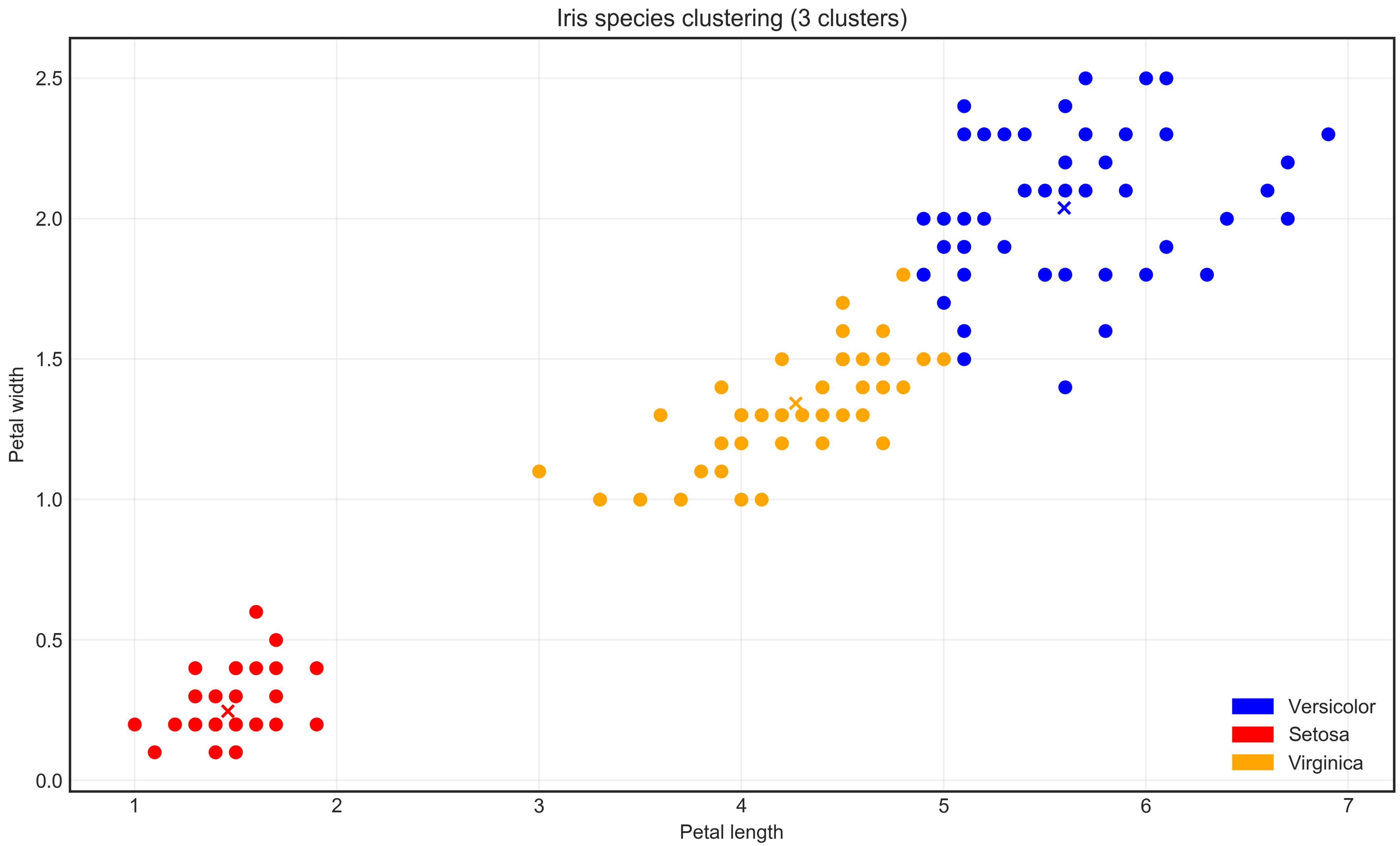
K-Means with 4 clusters



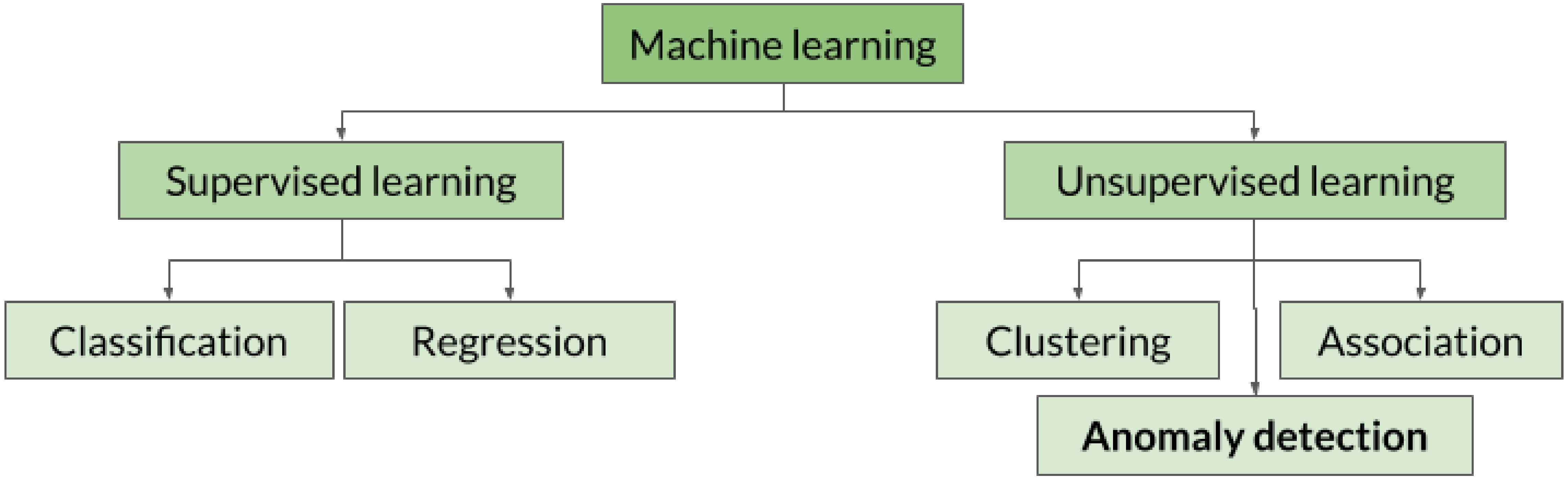
K-Means with 3 clusters



Ground truth



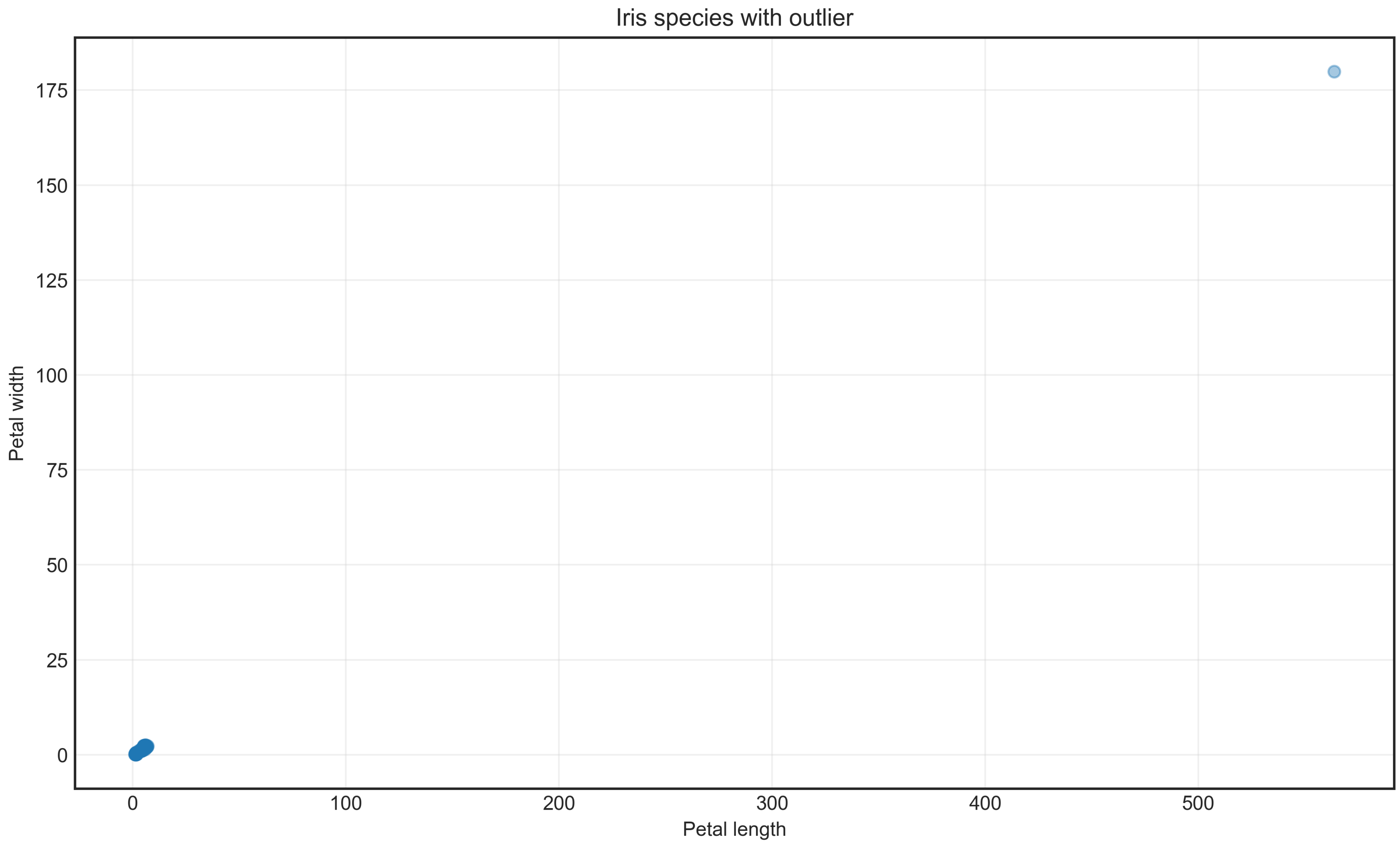
Anomaly detection



Detecting outliers

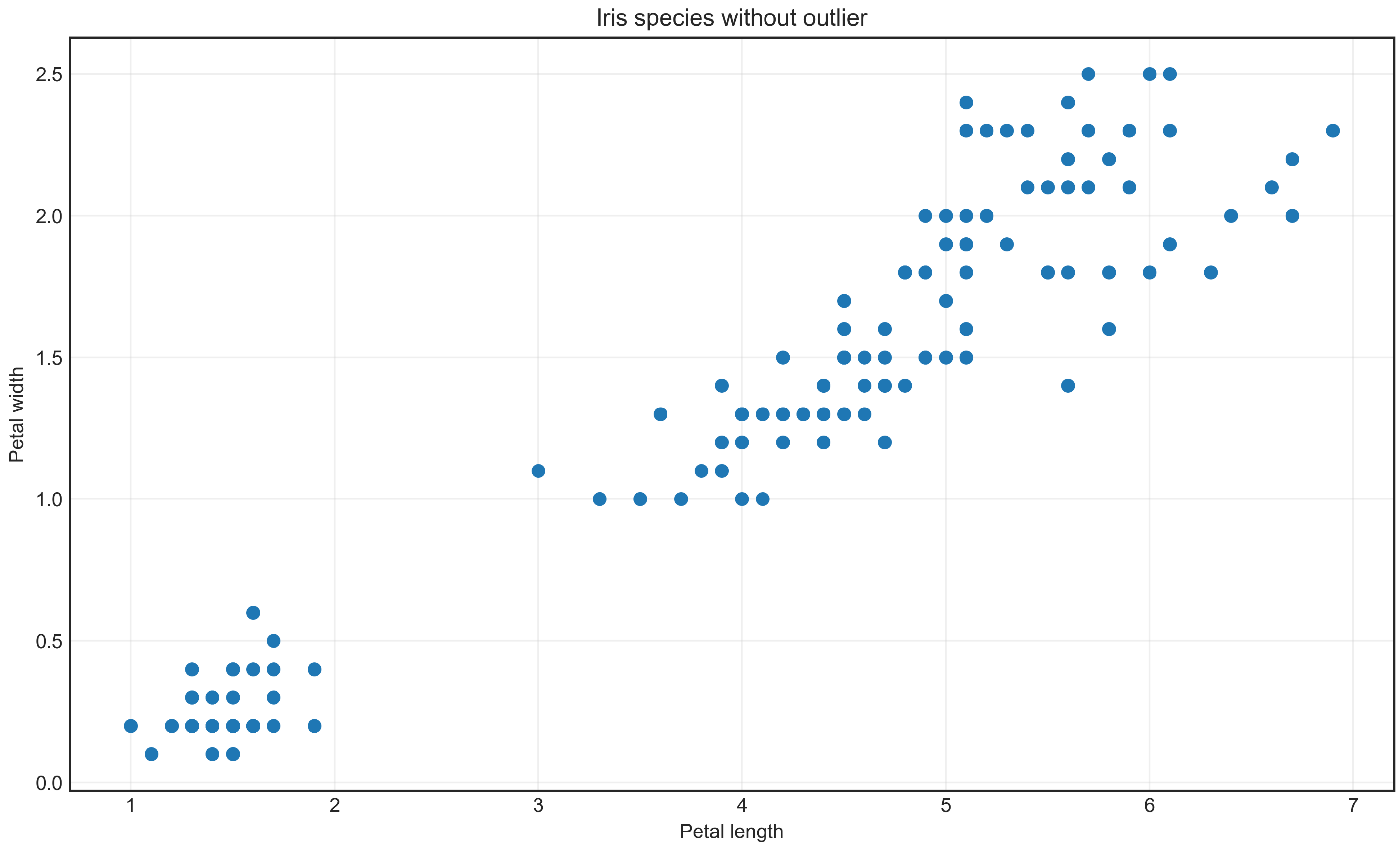
- Anomaly detection = **detecting outliers**
- Outliers = observations that **differ from the rest**

Outliers



Iris species with outlier
Petal width
175
150
125
100
75
50
25
0
0
100
200
300
400
500
Petal length

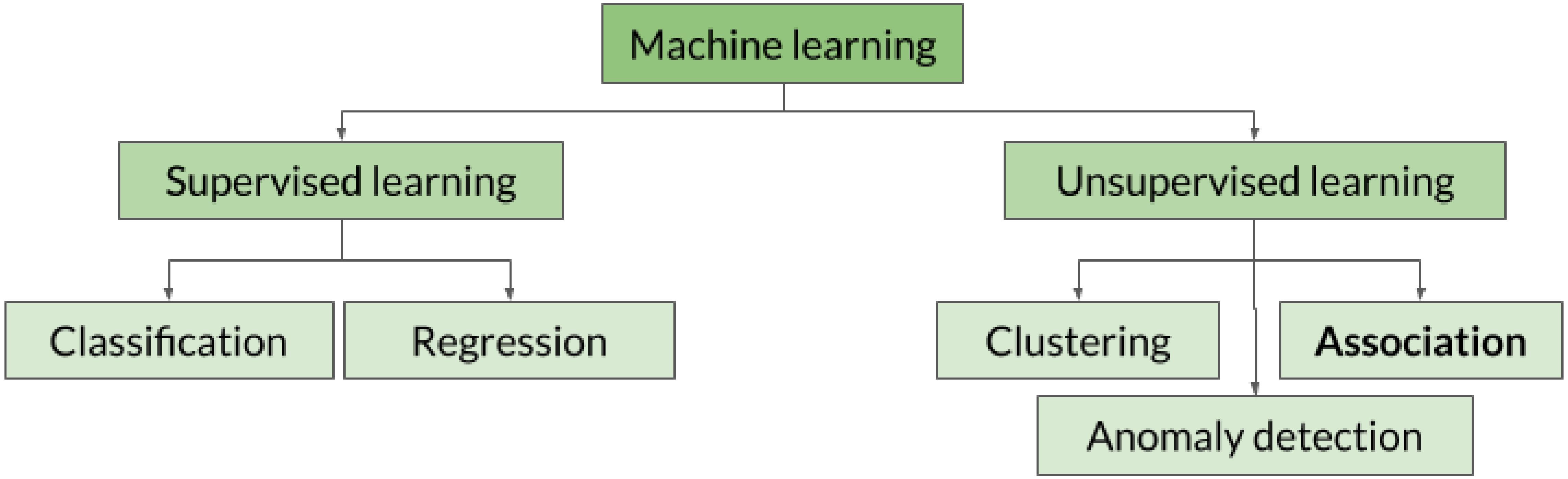
Removing outliers



Some anomaly detection use cases

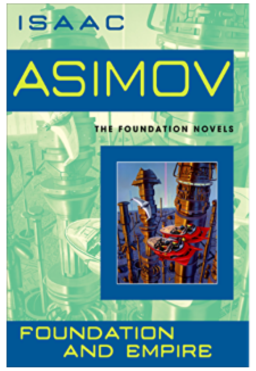
- Discover devices that fail faster or last longer
- Discover fraudsters that manage trick the system
- Discover patients that resist a fatal disease
- ...

Association

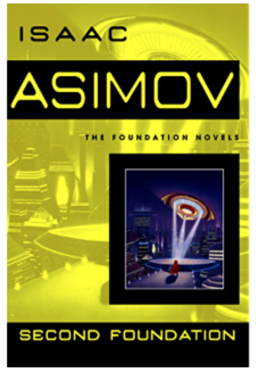


Association

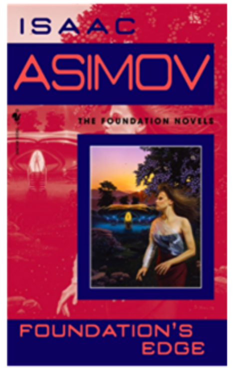
Customers who bought this item also bought



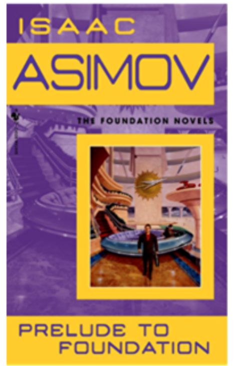
Foundation and Empire
› Isaac Asimov
335
Kindle Edition
1 offer from $5.99



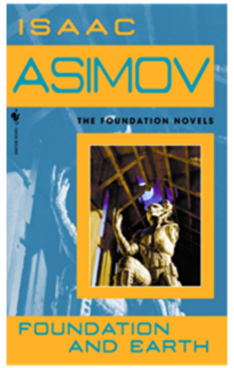
Second Foundation
› Isaac Asimov
253
Kindle Edition
1 offer from $6.99



Foundation's Edge
› Isaac Asimov
277
Kindle Edition
1 offer from $6.99

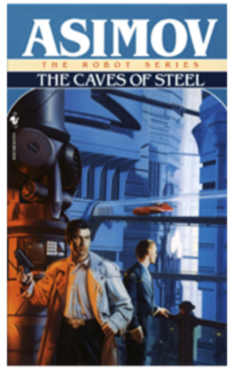


Prelude to Foundation
› Isaac Asimov
371
Kindle Edition
1 offer from $8.99



Foundation and Earth
› Isaac Asimov
322
Kindle Edition
1 offer from $6.99

I, Robot (The Robot Series)
› Isaac Asimov
714
Kindle Edition
1 offer from $7.99



The Caves of Steel (The Robot Series Book 1)
› Isaac Asimov
557
Kindle Edition
1 offer from $7.99

Let's practice!

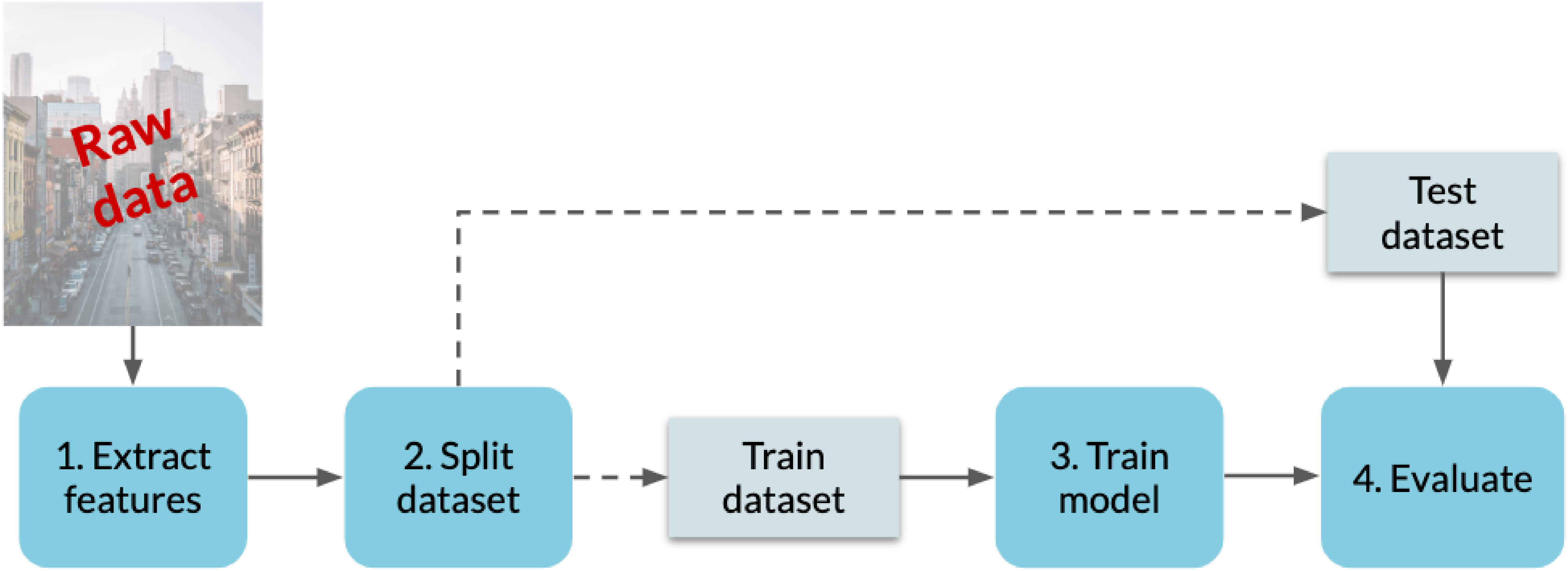
UNDERSTANDING MACHINE LEARNING

Evaluating performance

UNDERSTANDING MACHINE LEARNING

Hadrien Lacroix
Content Developer at DataCamp

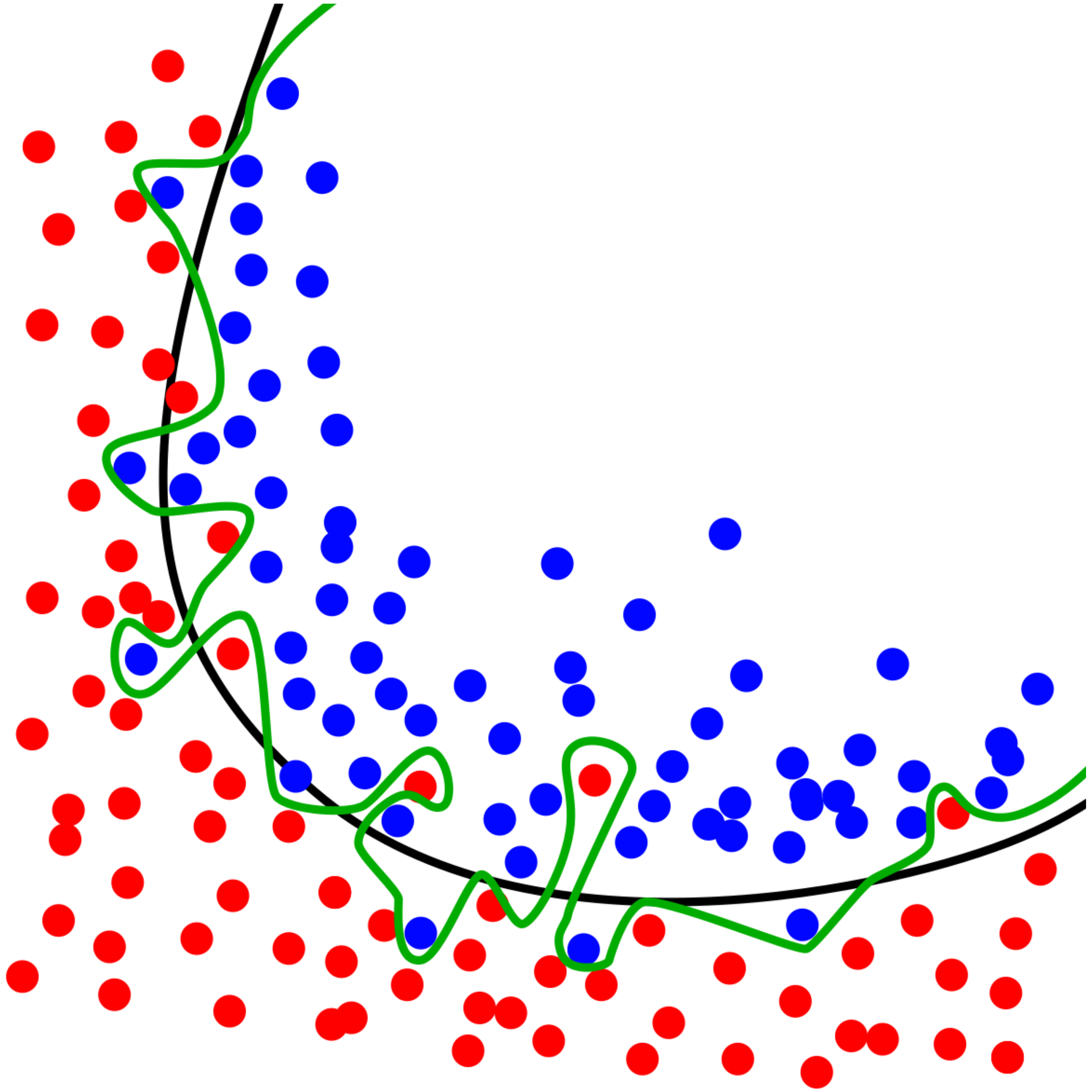
Evaluate step



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**

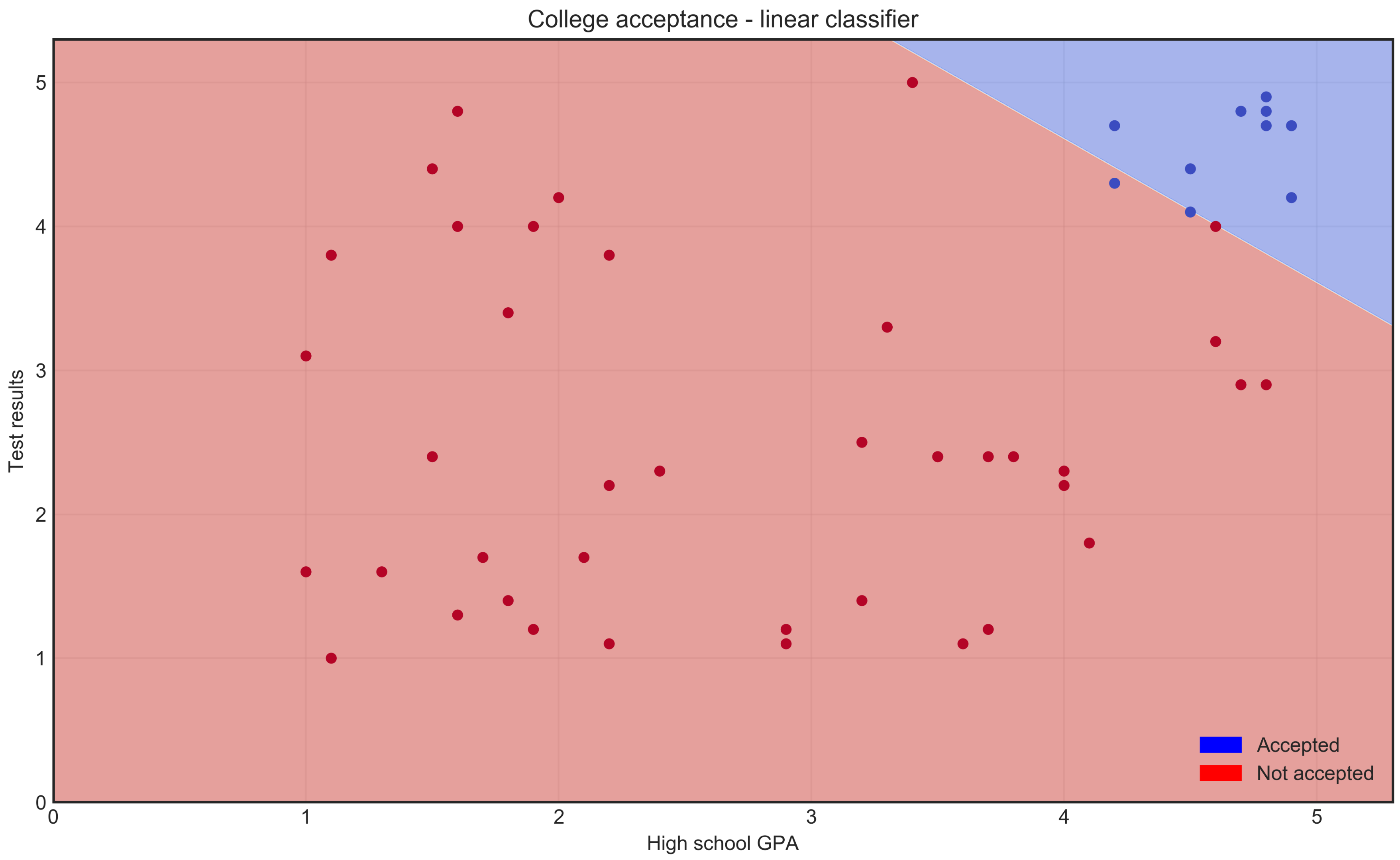
Illustrating overfitting



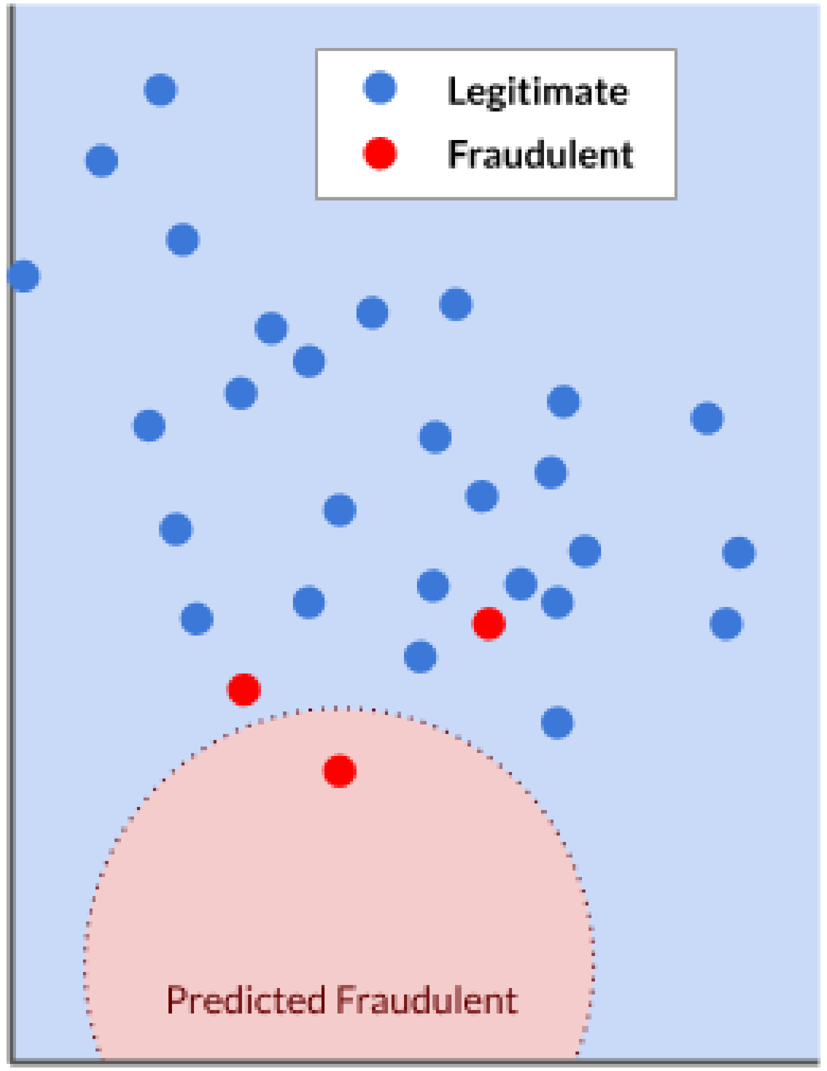
datacamp

Accuracy

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



Limits of accuracy: fraud example

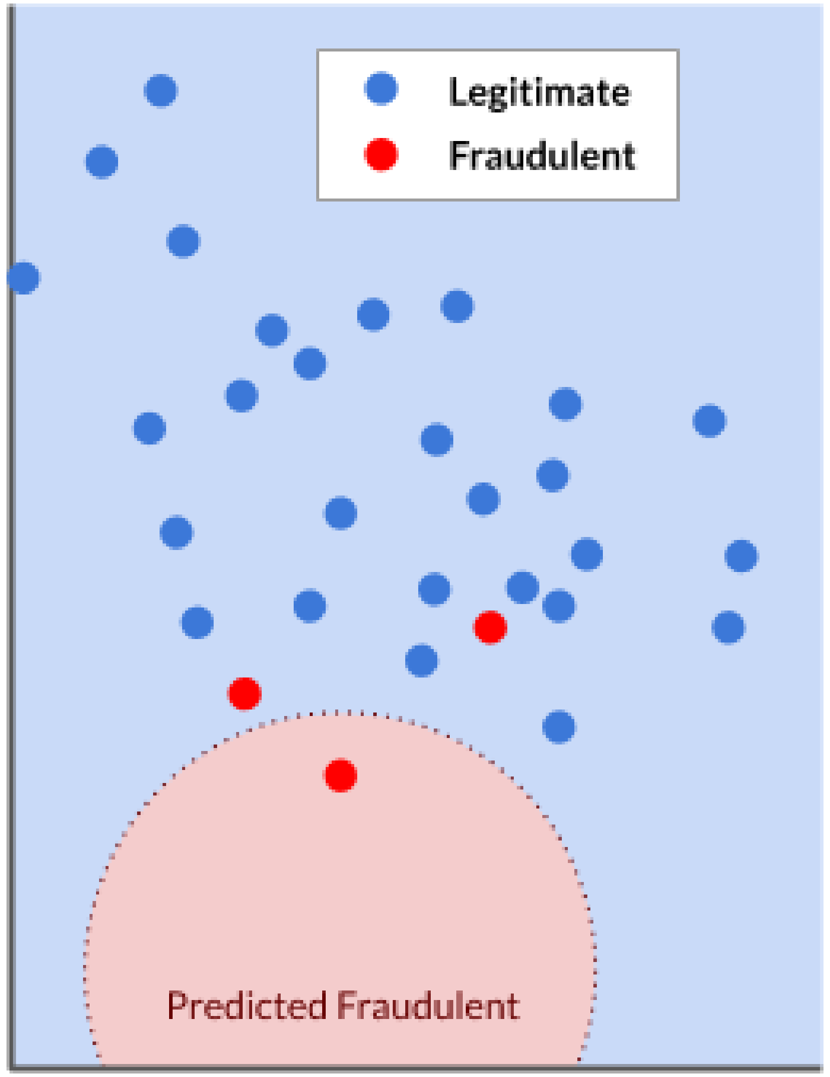


Accuracy of this model:

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

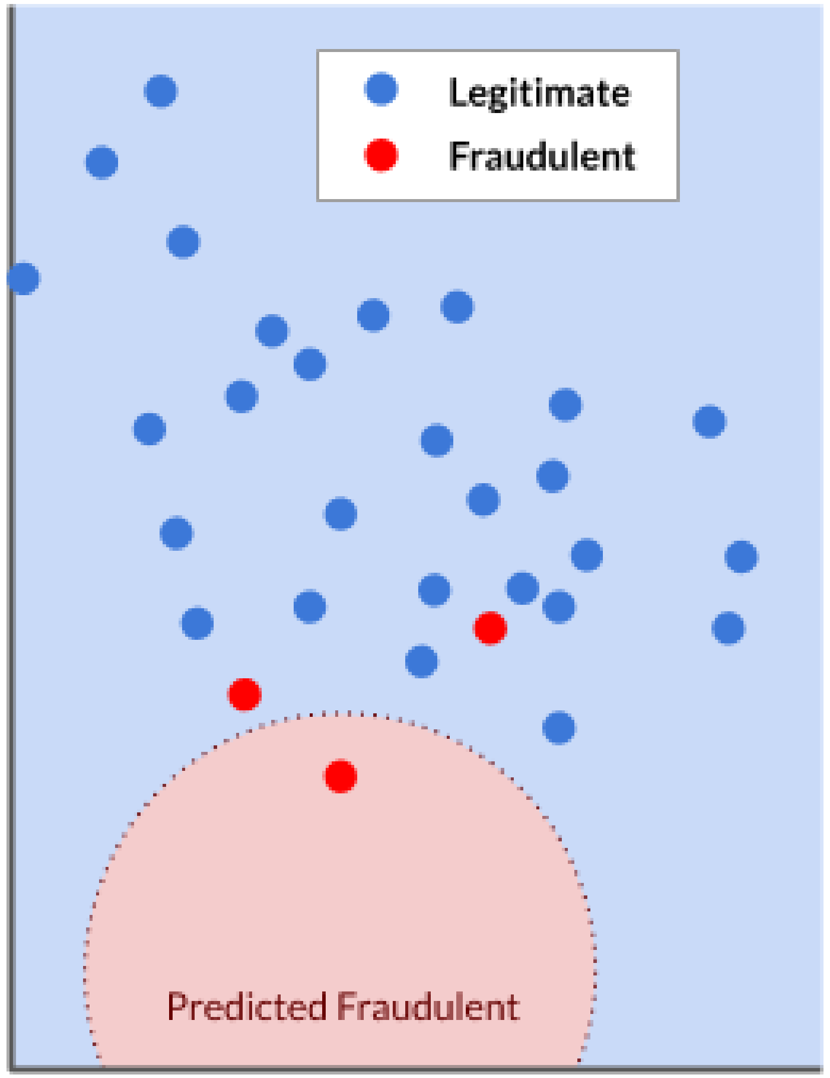
- Misses majority of fraudulent transactions
- Need a better metric

Confusion matrix



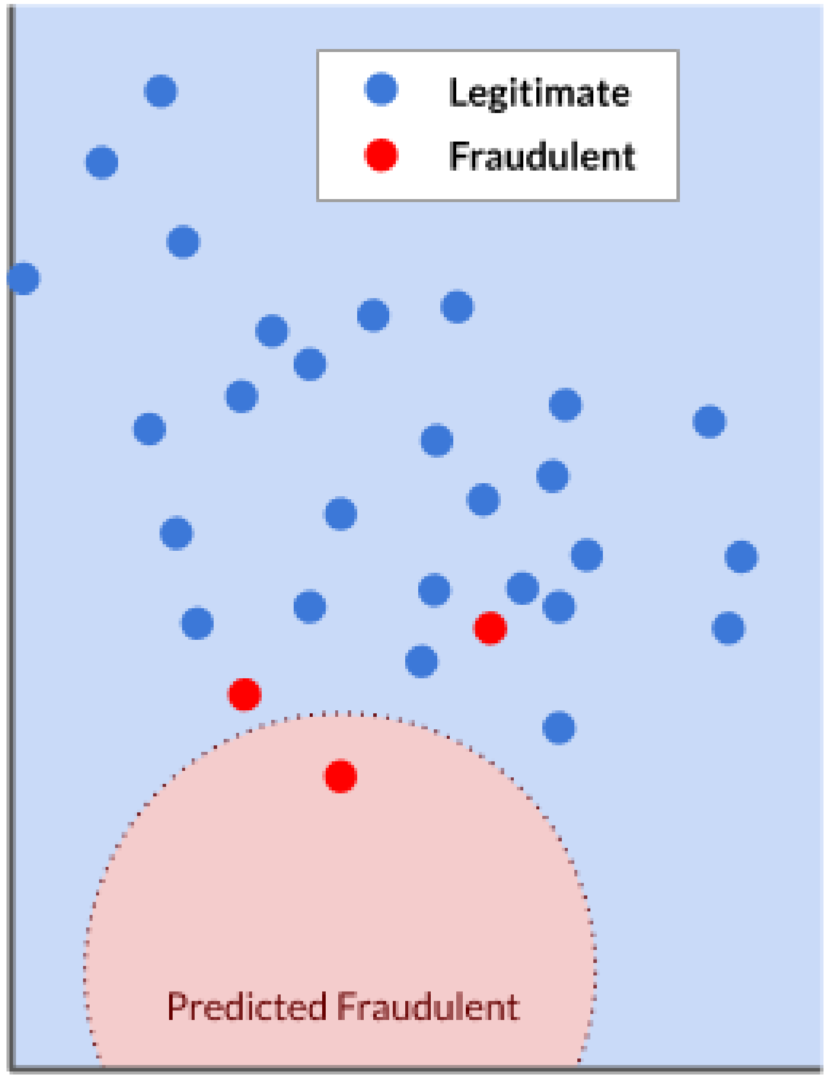
		Actual values	
		Fraudulent	*Not Fraudulent*
Predicted	*Fraudulent*		
	Not Fraudulent		

True positives



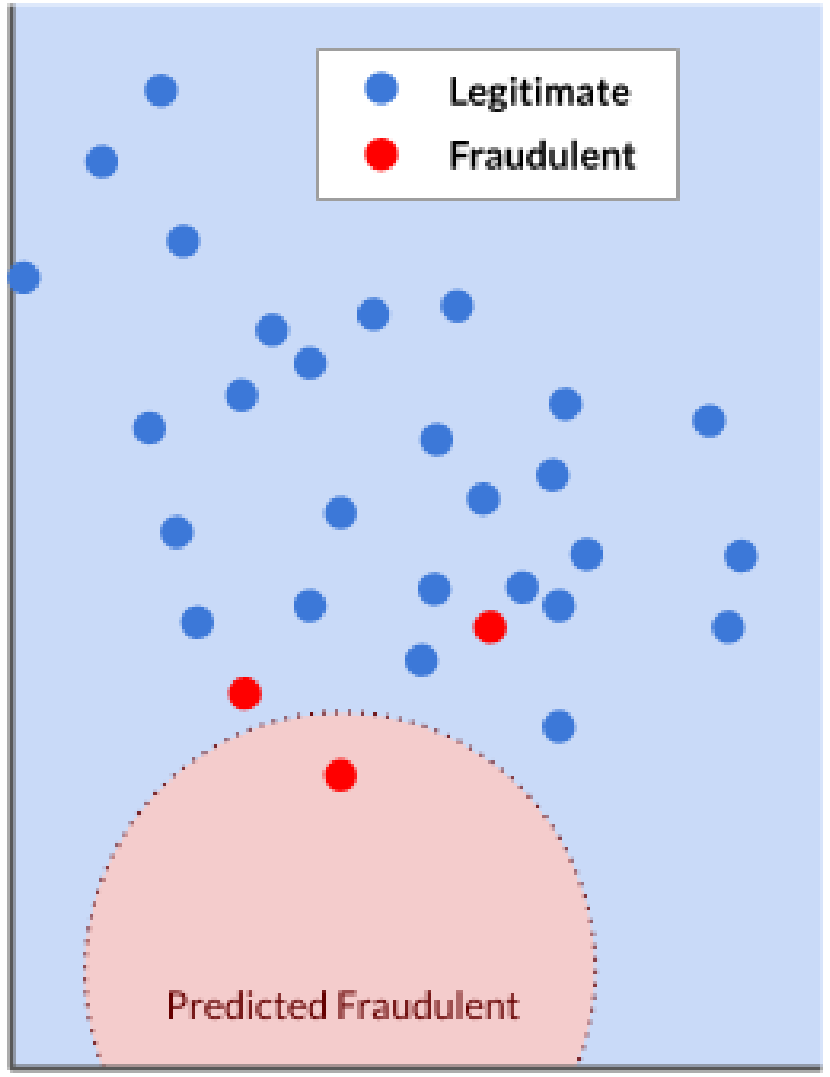
		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent		
	Not Fraudulent		

True positives



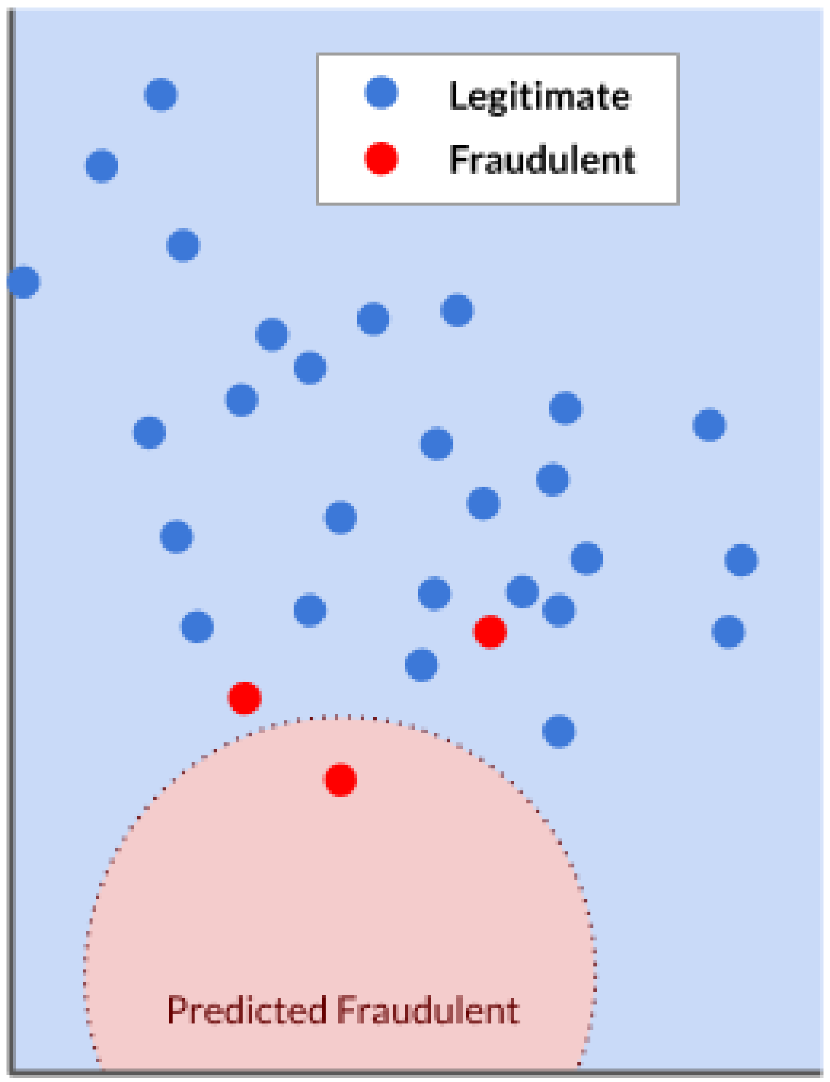
		Actual values	
		Fraudulent	*Not Fraudulent*
Predicted	*Fraudulent*	1 true positives	
	Not Fraudulent		

False negatives



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

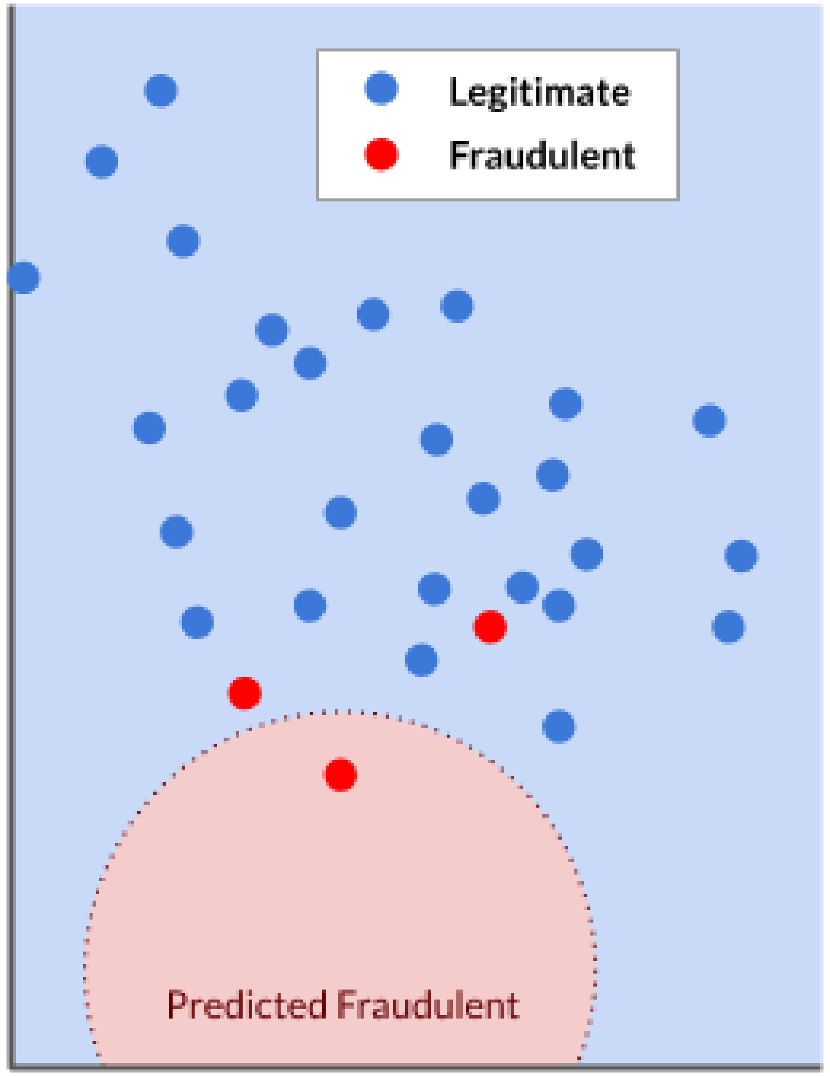
False negatives



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

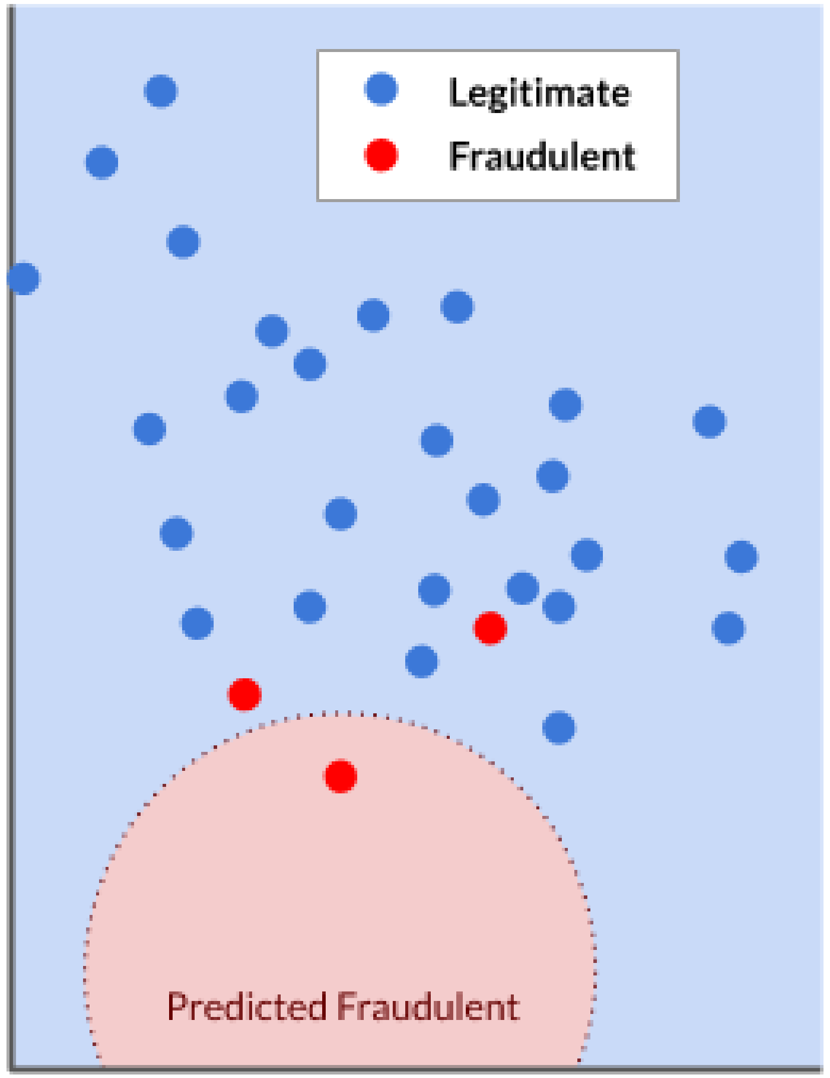
Remembering False Negatives

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

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?

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

- Rather mark legitimate transactions as suspicious than authorize fraudulent transactions

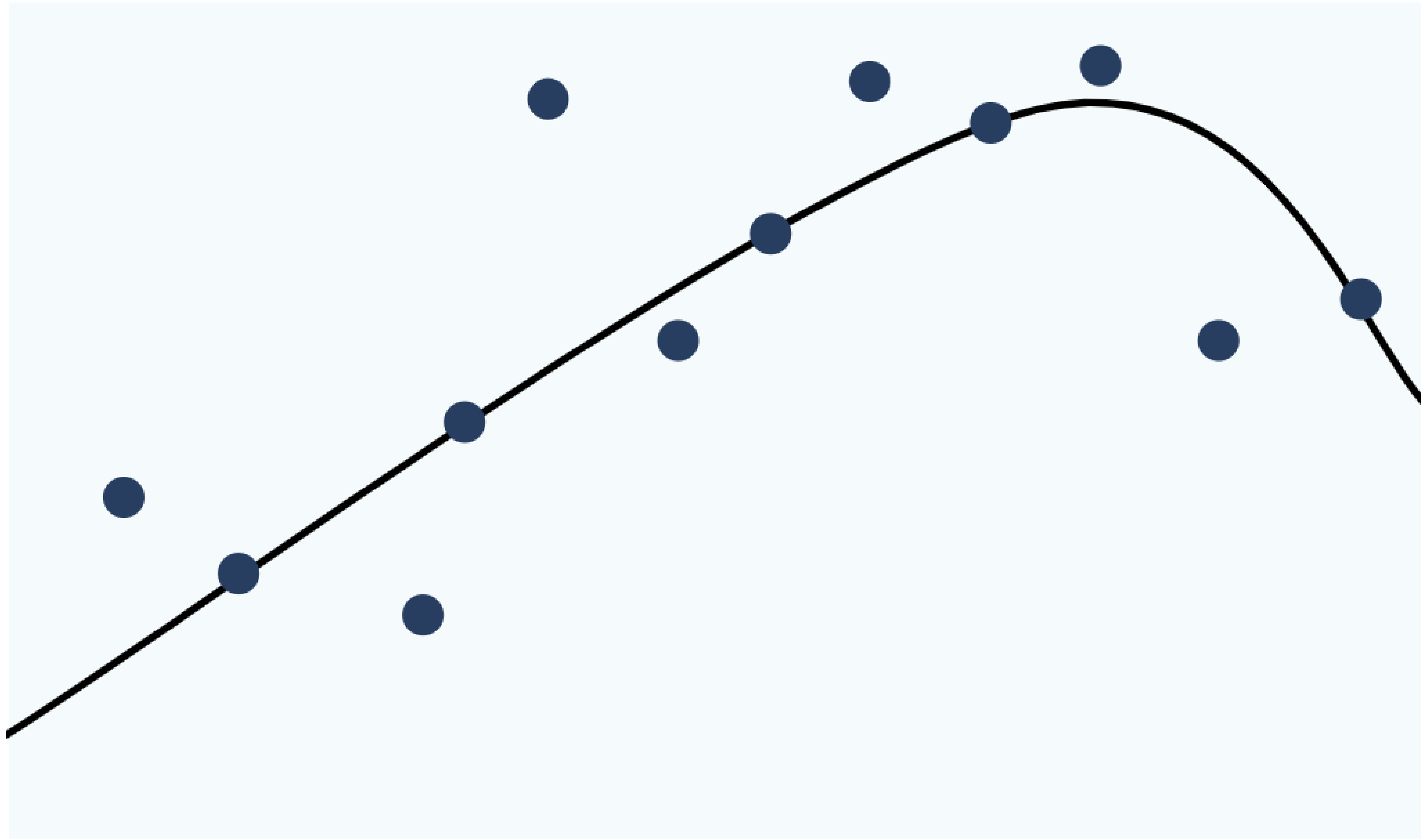
Specificity

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

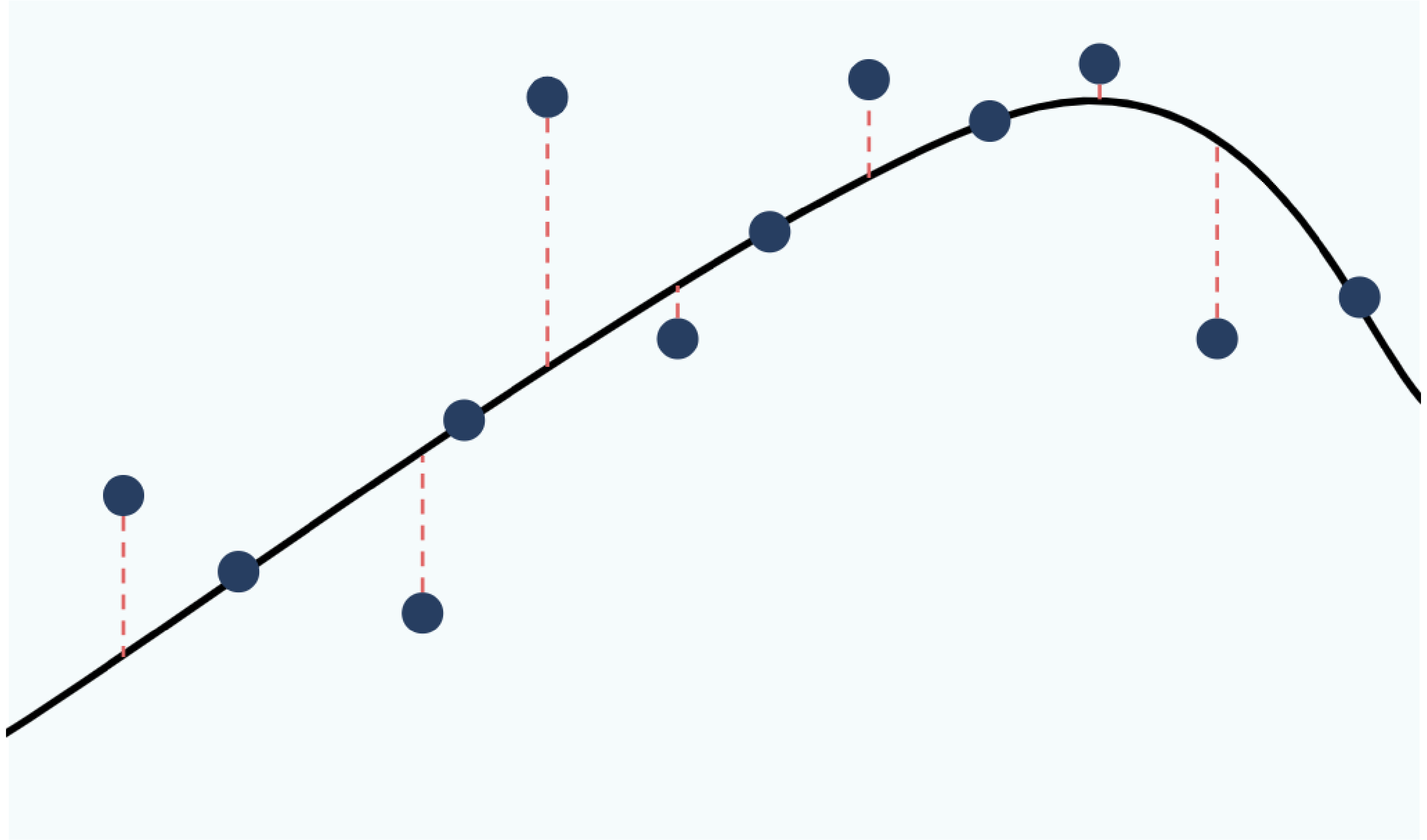
Spam filter:

- 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

[1] https://www.flickr.com/photos/micahdowty/8540188997

Let's practice!

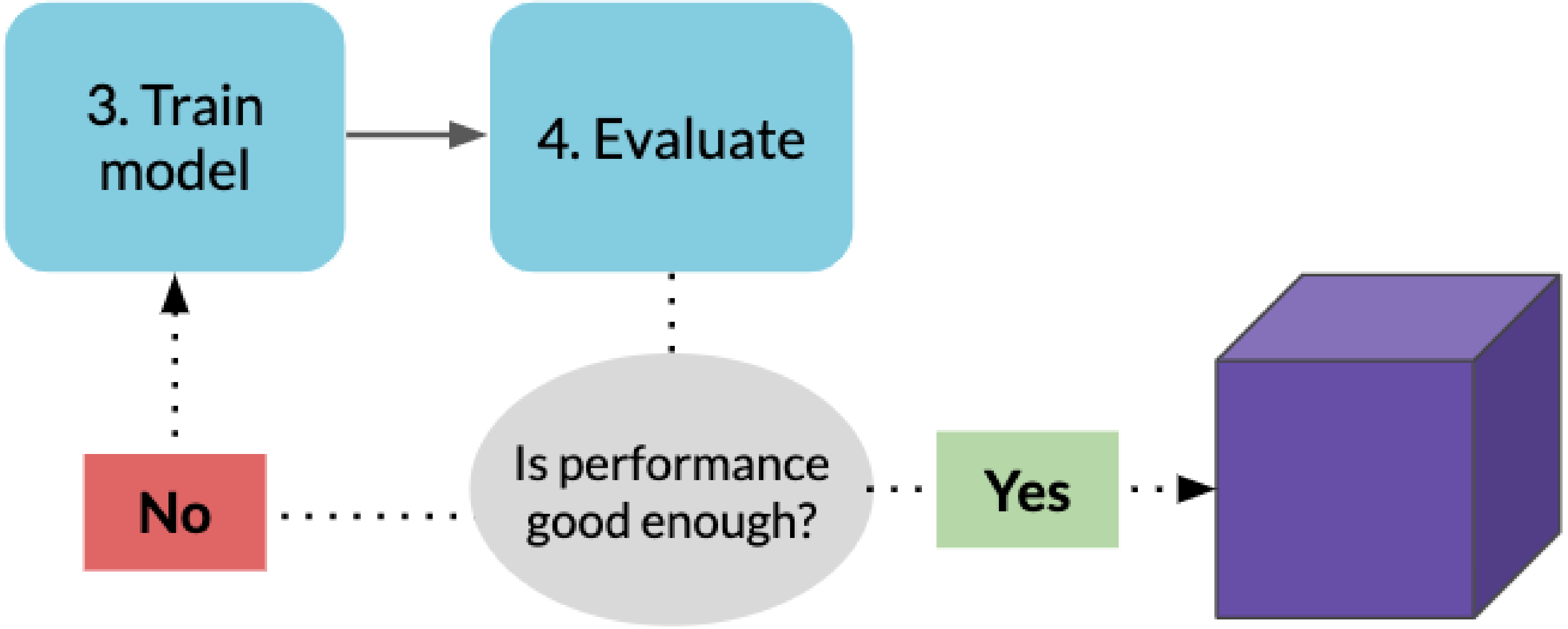
UNDERSTANDING MACHINE LEARNING

Improving performance

UNDERSTANDING MACHINE LEARNING

Hadrien Lacroix
Content Developer at DataCamp

Machine learning workflow



Several options

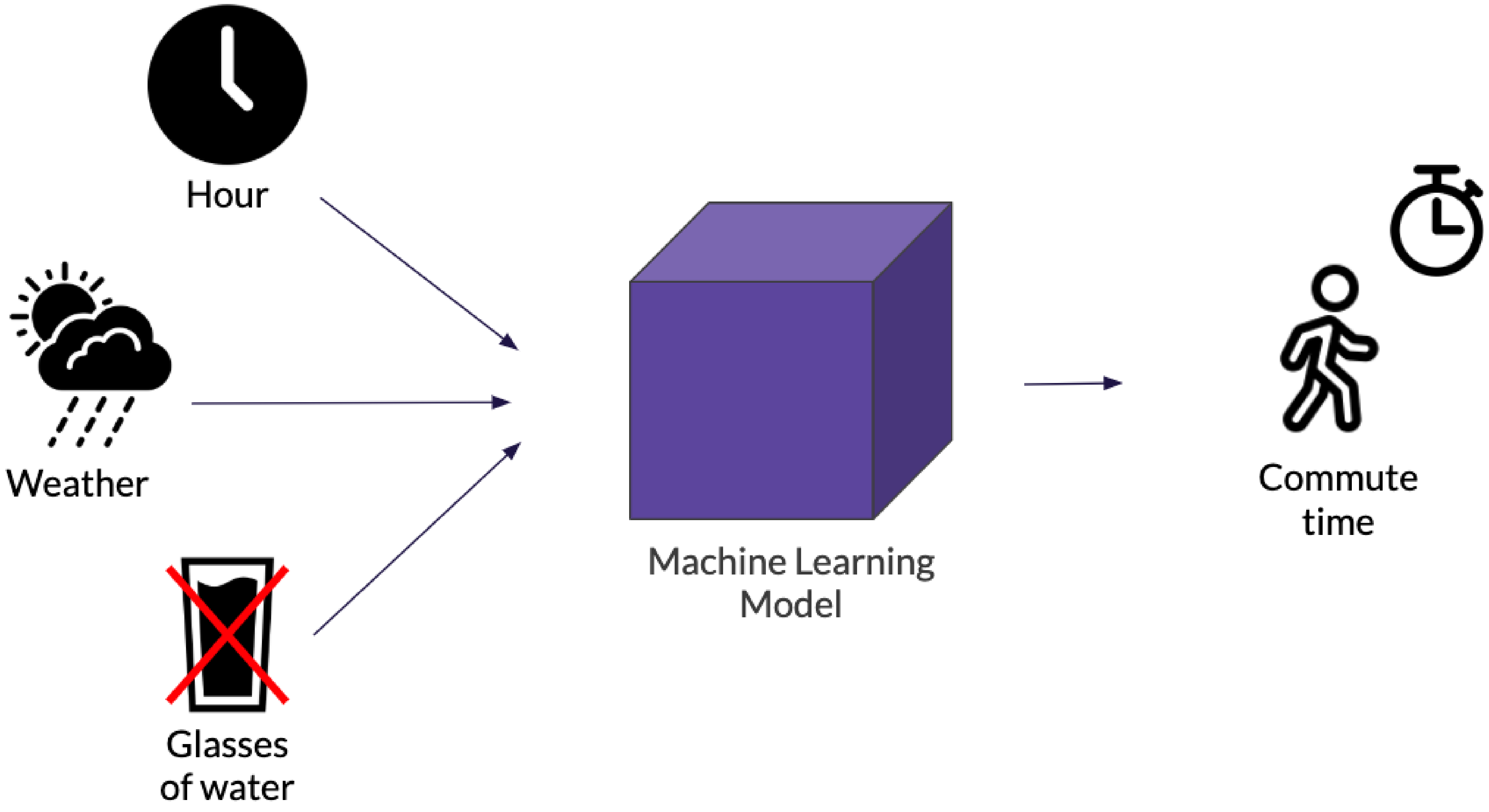
- Dimensionality reduction
- Hyperparameter tuning
- Ensemble methods

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

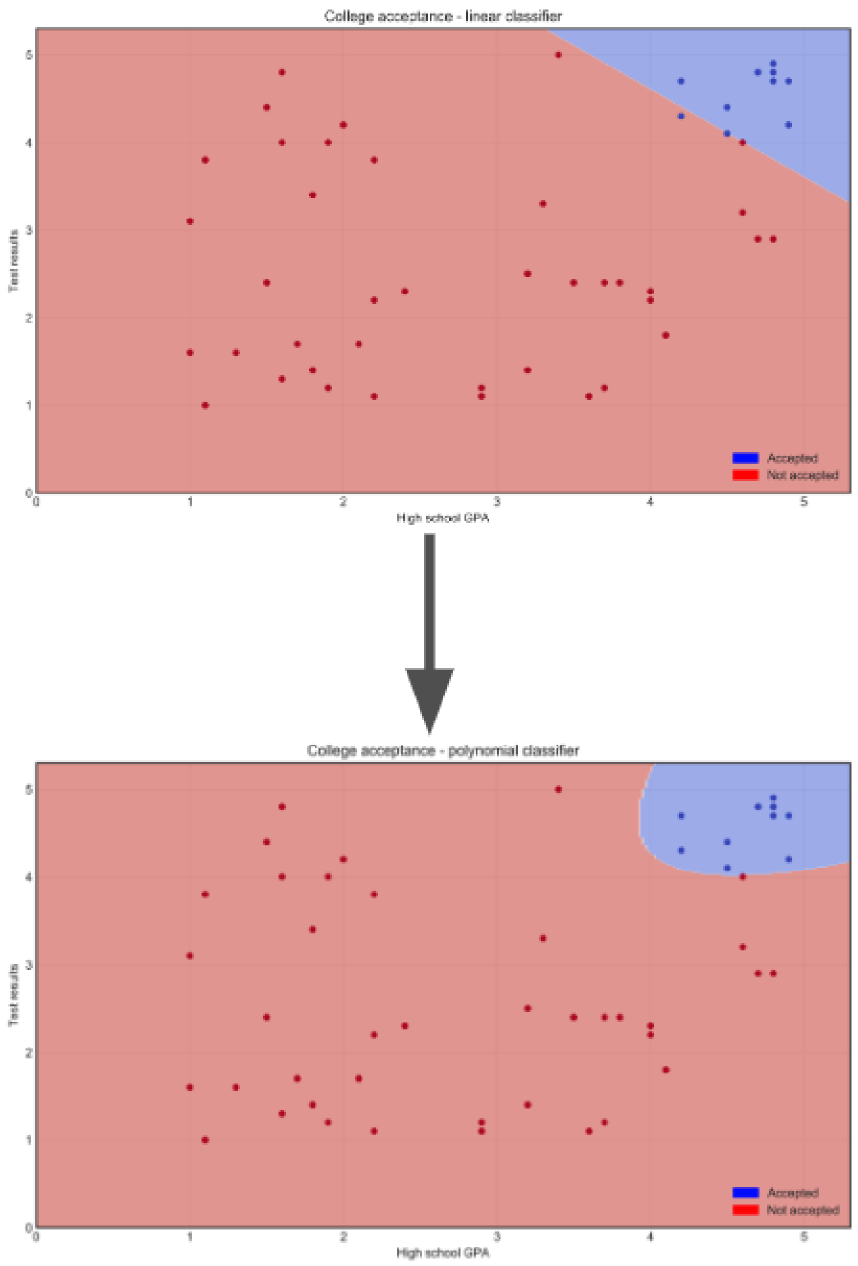
Hyperparameter tuning

Hyperparameter tuning

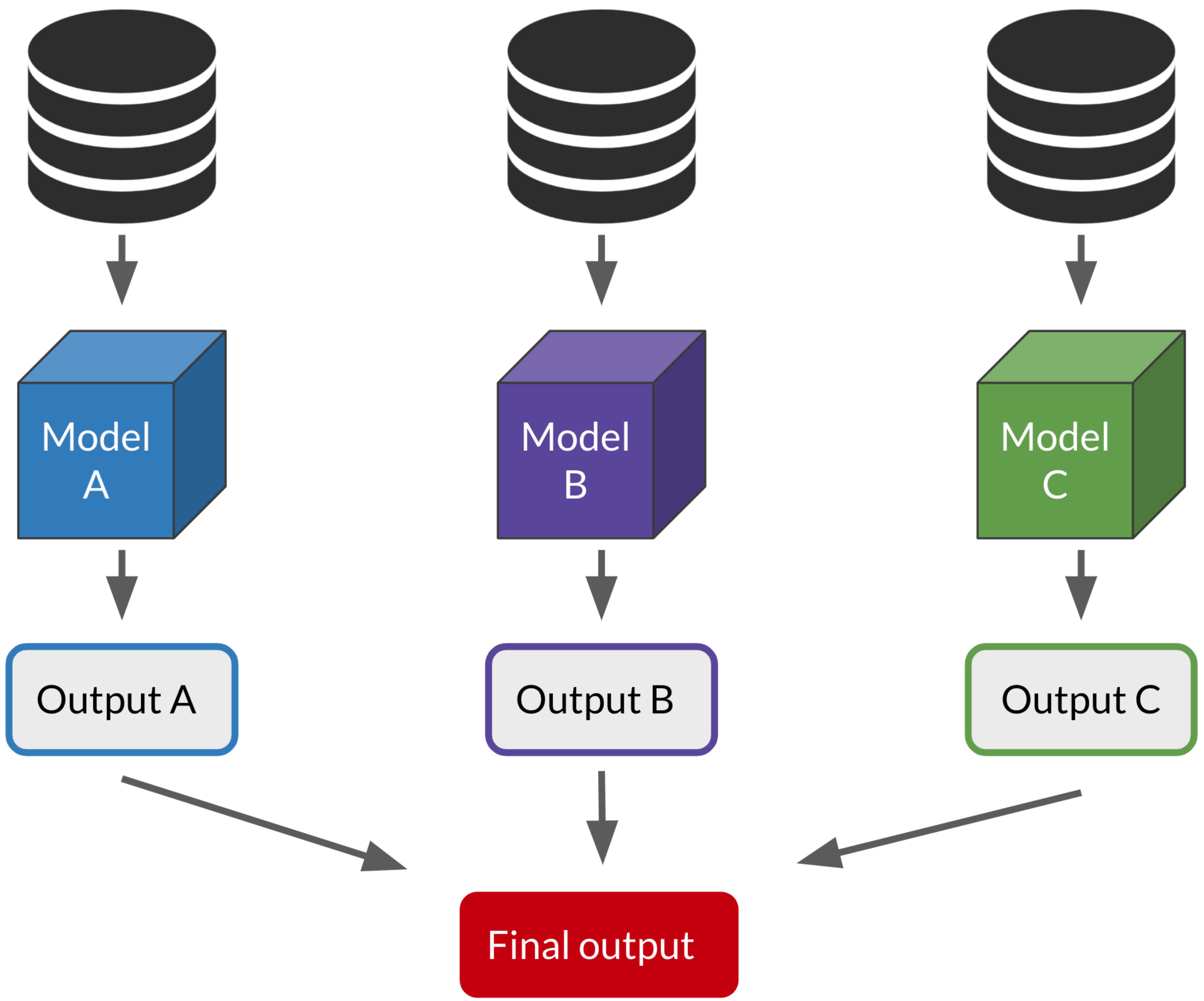
Hyperparameter tuning: example

SVM algorithm hyperparameters:

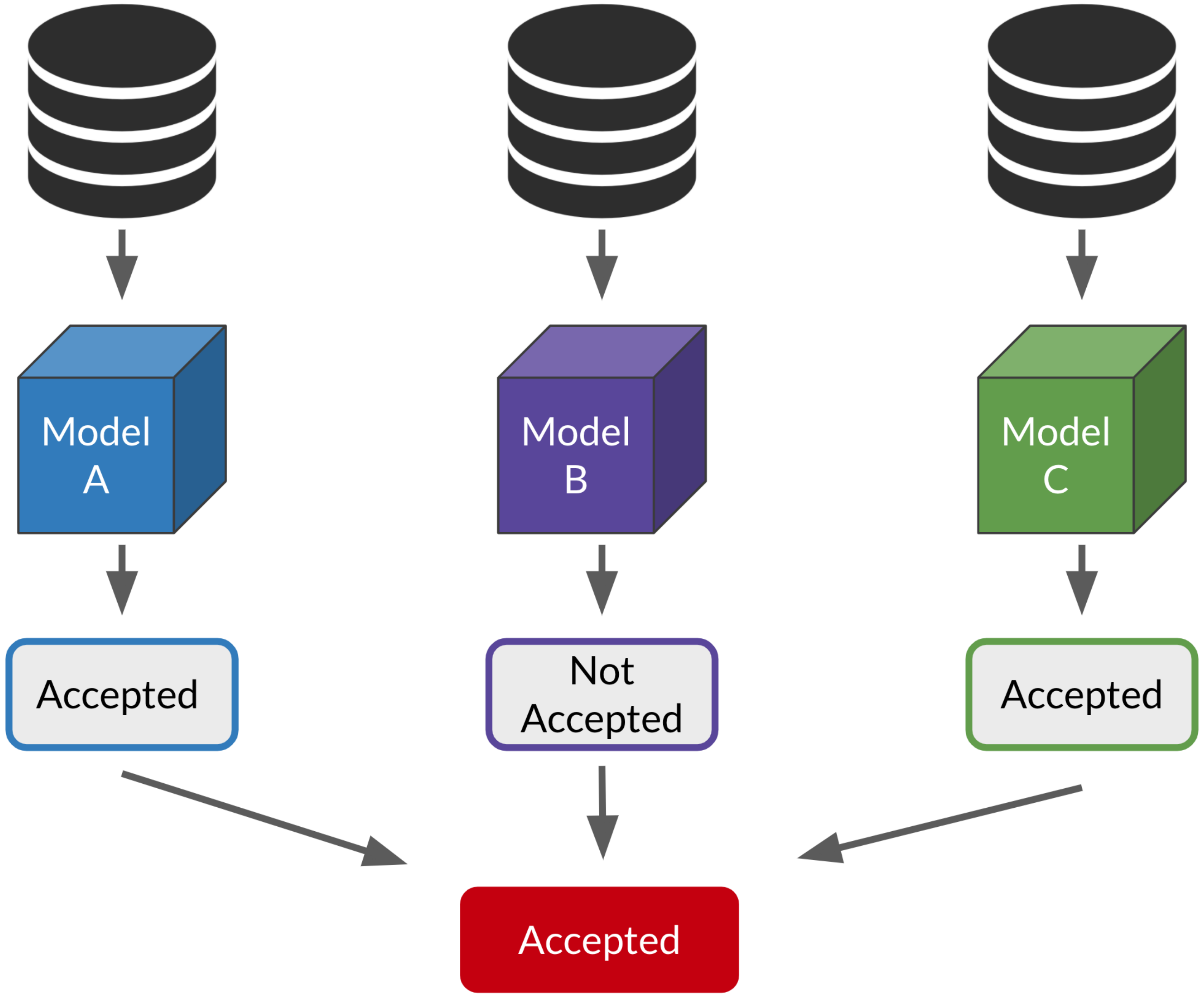
- `kernel` : "linear" --> "poly"
- `C`
- `degree`
- `gamma`
- `shrinking`
- `coef0`
- `tol`
- ...



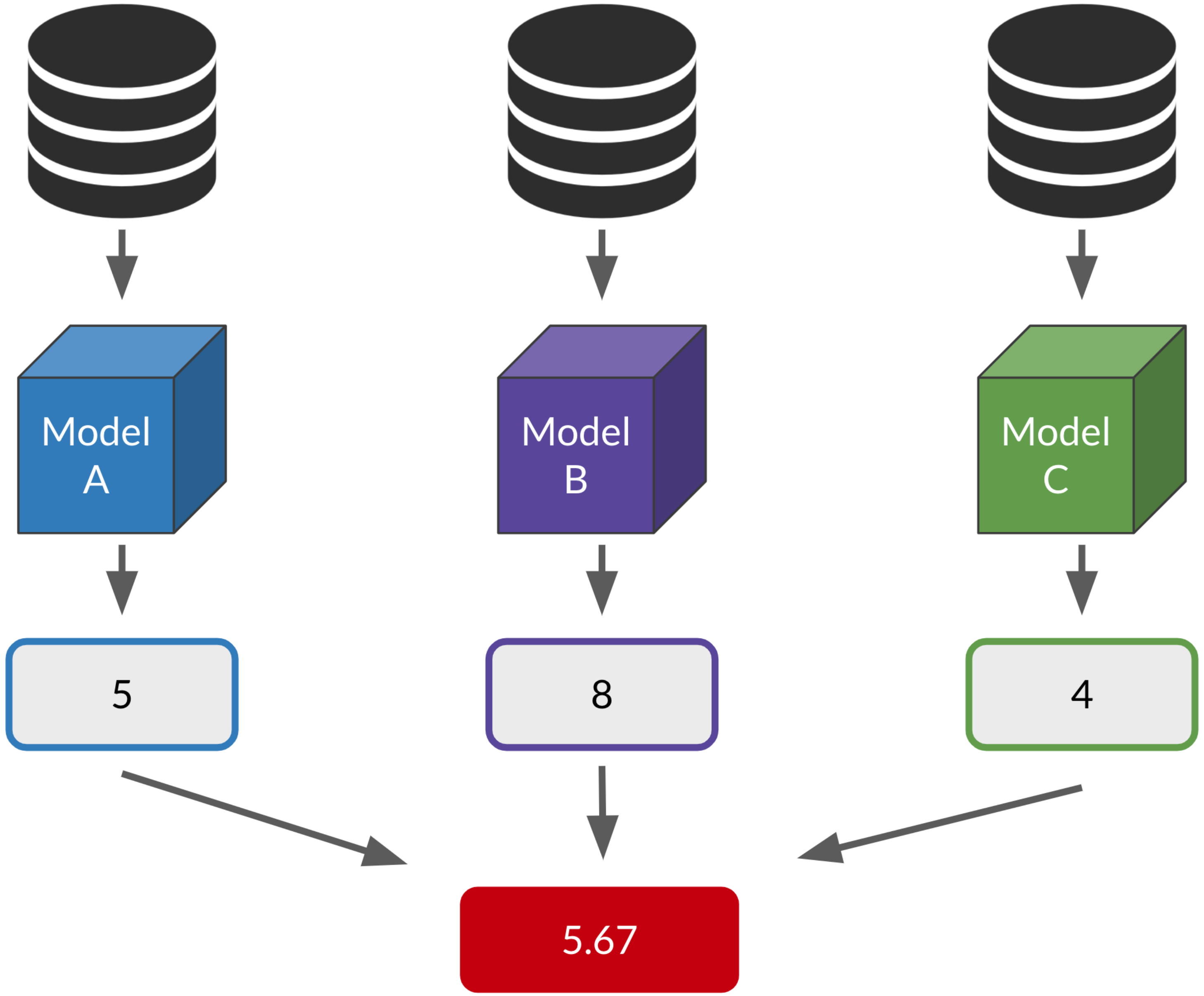
Ensemble methods



Ensemble methods: classification



Ensemble methods: regression



Let's practice!

UNDERSTANDING MACHINE LEARNING