Machine learning basics with scikit-learn

A first, introductory lesson, focusing on general concepts rather than coding or maths.

What is machine learning ?

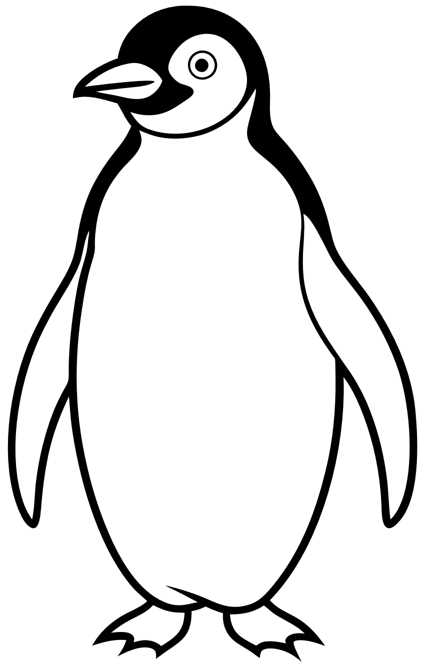
Machine learning deals with building predictive models.

Why and when?

Some examples of machine learning

Which penguin is that?

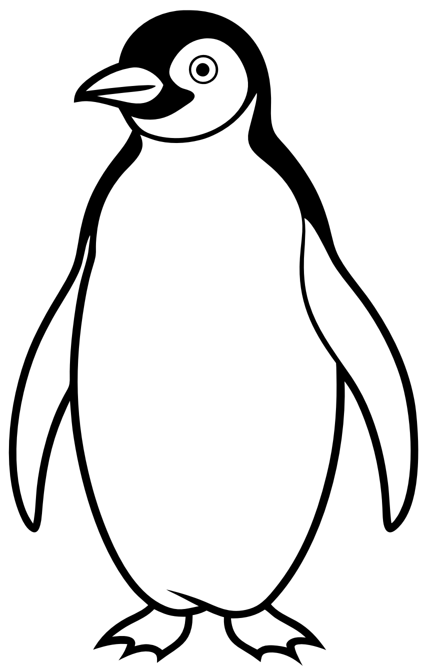
- Adélie
- Chinstrap
- Gentoo



Which penguin is that?

- Adélie
- Chinstrap
- Gentoo

Culmen Length	Culmen Depth	Flipper Length	Body Mass	Species
39.1mm	18.7mm	181.0mm	3.75kg	Adelie
43.5mm	18.1mm	202.0mm	3.40kg	Chinstrap
39.5mm	17.4mm	186.0mm	3.80kg	Adelie
46.1mm	13.2mm	211.0mm	4.50kg	Gentoo



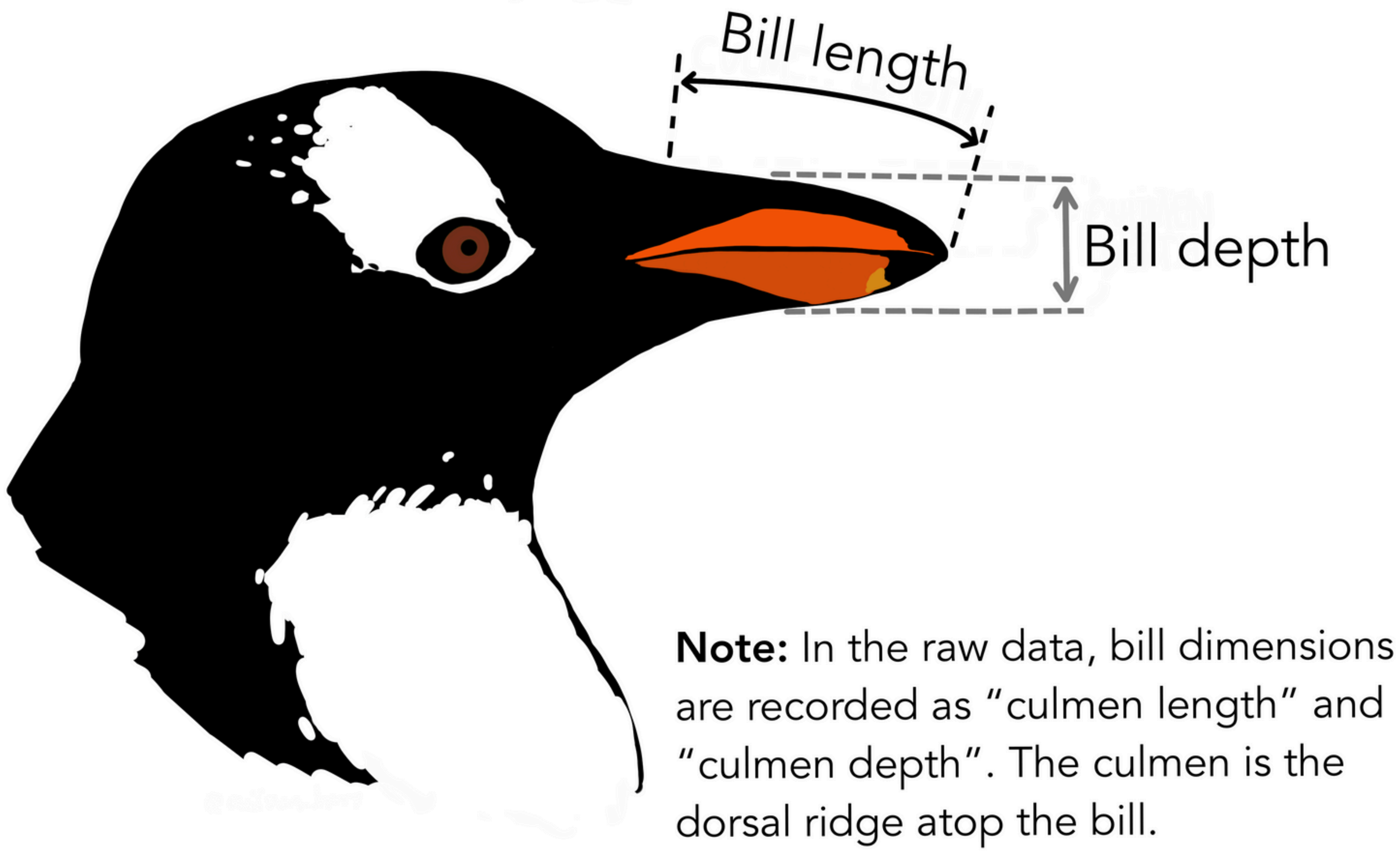
What’s this person’s income?

What's this person's income?

Age	Workclass	Education	Marital-status	Occupation	Relationship	Capital-gain	Hours-per-week	Native-country	Class
25	Private	11th	Never-married	Machine-op-inspct	Own-child	0	40	United-States	<=50K
38	Private	HS-grad	Married-civ-spouse	Farming-fishing	Husband	0	50	United-States	<=50K
28	Local-gov	Assoc-acdm	Married-civ-spouse	Protective-serv	Husband	0	40	United-States	>50K
44	Private	Some-college	Married-civ-spouse	Machine-op-inspct	Husband	7688	40	United-States	>50K

Engineering rules: data versus experts

Expert knowledge: Adélie penguins have shorter bills (shorter culmen)

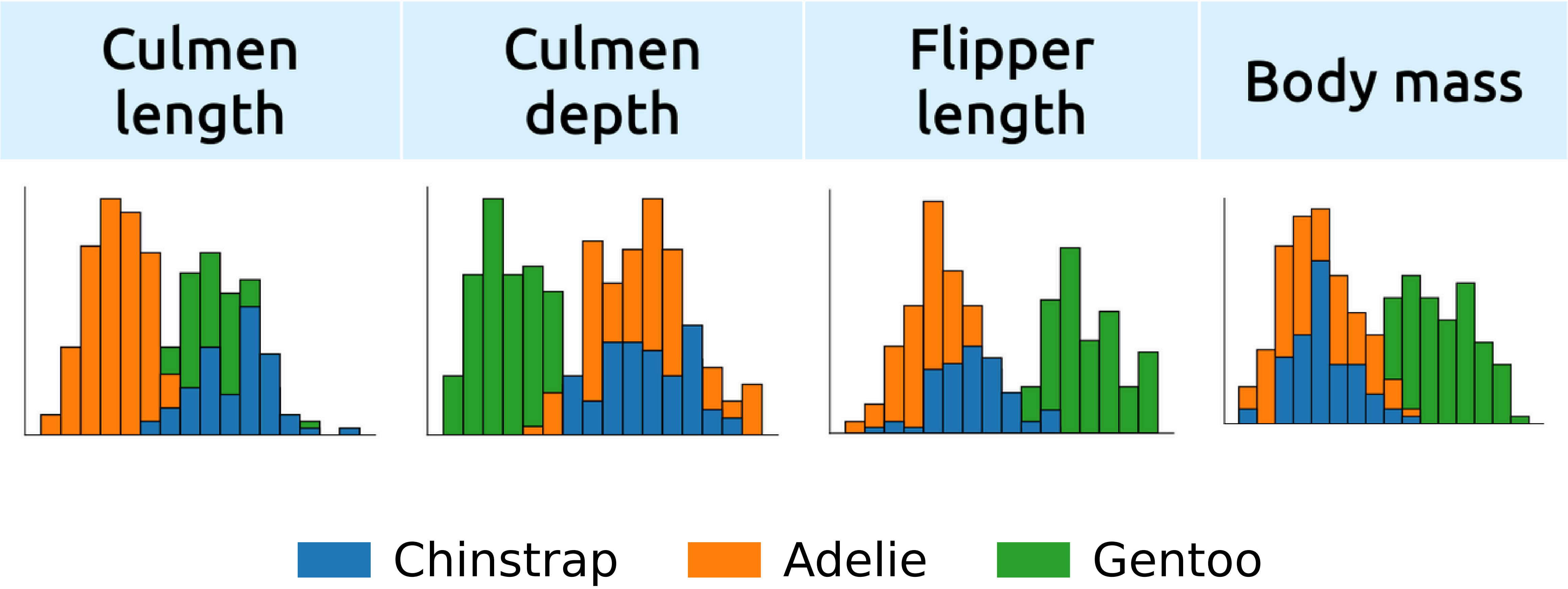


Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Image from palmerpenguinsm
Horst AM, Hill AP, Gorman KB (2020)

Engineering rules: data versus experts

Expert knowledge: Adélie penguins have shorter bills (culmen)



This rule can be inferred from the data

Predictive analysis

Beyond classic statistical tools

Generalizing

Concluding on new instances

Generalizing

Concluding on new instances

Many sources of variability:

- age
- marital status
- education
- hours-per-week
- workclass
- occupation
- relationship
- native-country
- capital-gain
- capital-loss

Generalizing

Concluding on new instances

Many sources of variability:

- age
- marital status
- education
- hours-per-week
- workclass
- occupation
- relationship
- native-country
- capital-gain
- capital-loss

+ Noise: unexplainable variance

Memorizing

- Consider a “nearest neighbors” model
- Store all known individuals (the census)
- Given a new individual, predict the income of its closest match in our database

Memorizing

- Consider a “nearest neighbors” model
- Store all known individuals (the census)
- Given a new individual, predict the income of its closest match in our database

Trying out this strategy on individuals picked from the data we have (the census) what error rate do we expect?

Memorizing

- Consider a “nearest neighbors” model
- Store all known individuals (the census)
- Given a new individual, predict the income of its closest match in our database

Trying out this strategy on individuals picked from the data we have (the census) what error rate do we expect?

0 errors

Memorizing

- Consider a “nearest neighbors” model
- Store all known individuals (the census)
- Given a new individual, predict the income of its closest match in our database

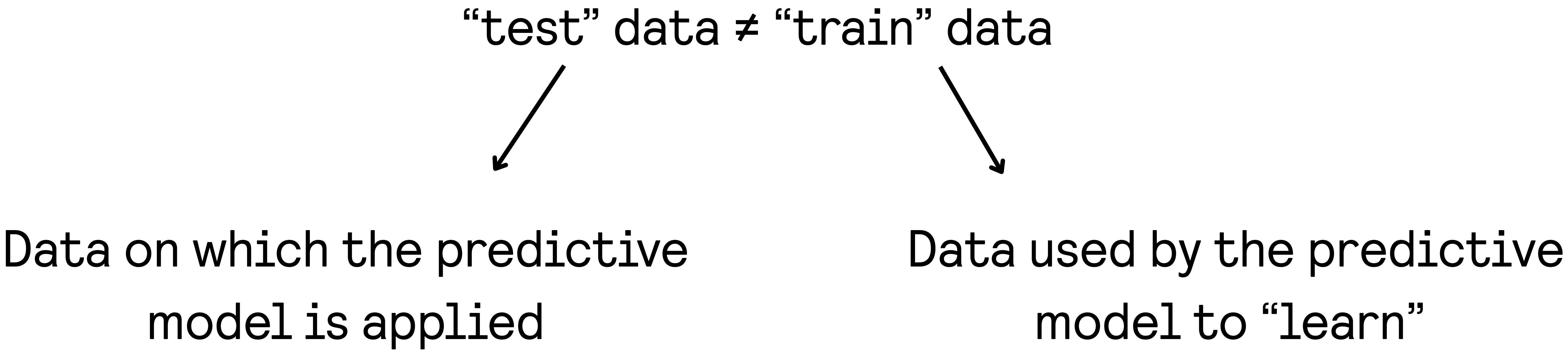
Trying out this strategy on individuals picked from the data we have (the census) what error rate do we expect?

0 errors

Yet, we will make errors on new data

Generalizing ≠ Memorizing

Generalizing ≠ Memorizing



“test” data ≠ “train” data

Data on which the predictive model is applied

Data used by the predictive model to “learn”

- Different sampling of noise
- Unobserved combination of features

The data matrix

We deal with a table of data (figuratively, an spreadsheet):

- Rows are different observations, or samples
- Columns are different descriptors, or features

n_features

n_samples

Culmen Length	Culmen Depth	Flipper Length	Body Mass
39.1mm	18.7mm	181.0mm	3.75kg
43.5mm	18.1mm	202.0mm	3.40kg
39.5mm	17.4mm	186.0mm	3.80kg
46.1mm	13.2mm	211.0mm	4.50kg

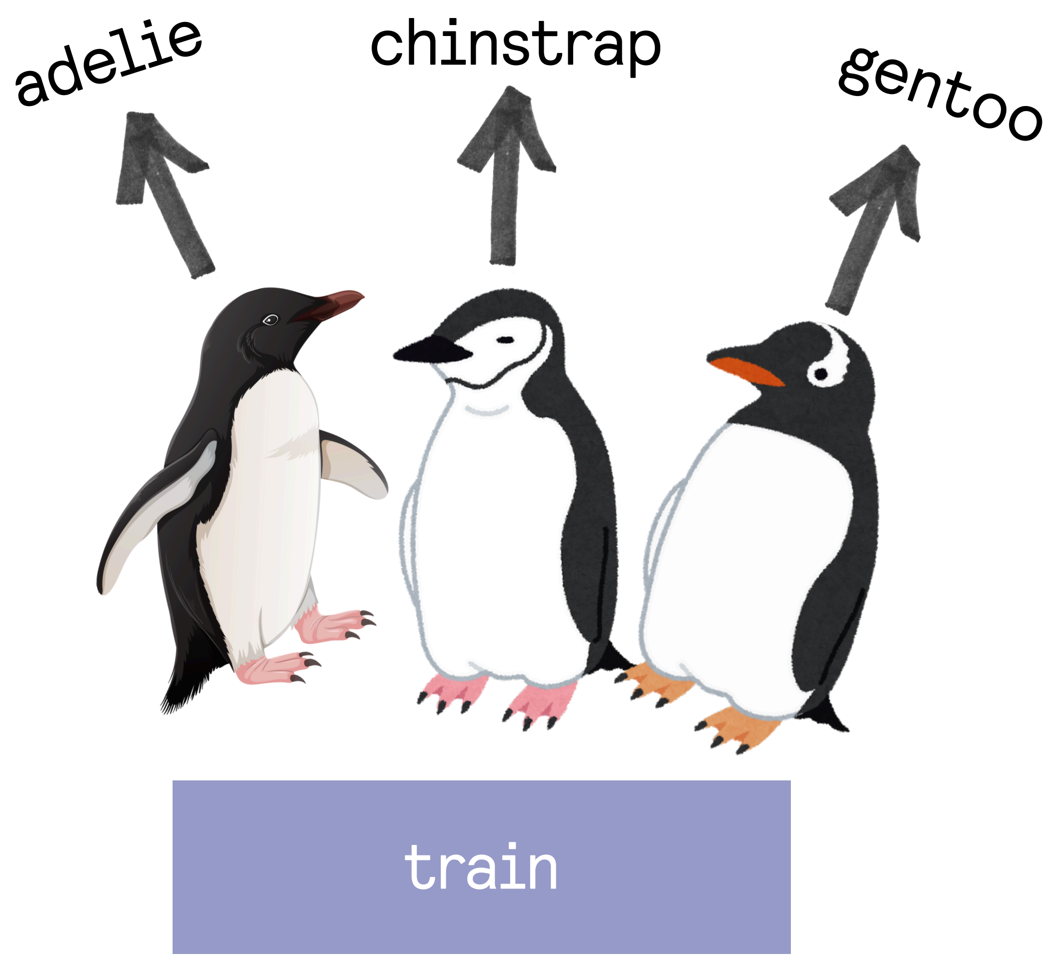
X (data)

Species
Adelie
Chinstrap
Adelie
Gentoo

y (target)

Supervised machine learning

- A data matrix X with n observations
- A target y: a property of each observation



The goal is to predict y

Regression and classification

Supervised learning: predicting a target y

- Classification: y is discrete (qualitative), made of different classes

eg: types of penguins: adelie, gentoo, chinstrap

- Regression: y is continuous (quantitative), a numerical quantity

eg: wage prediction

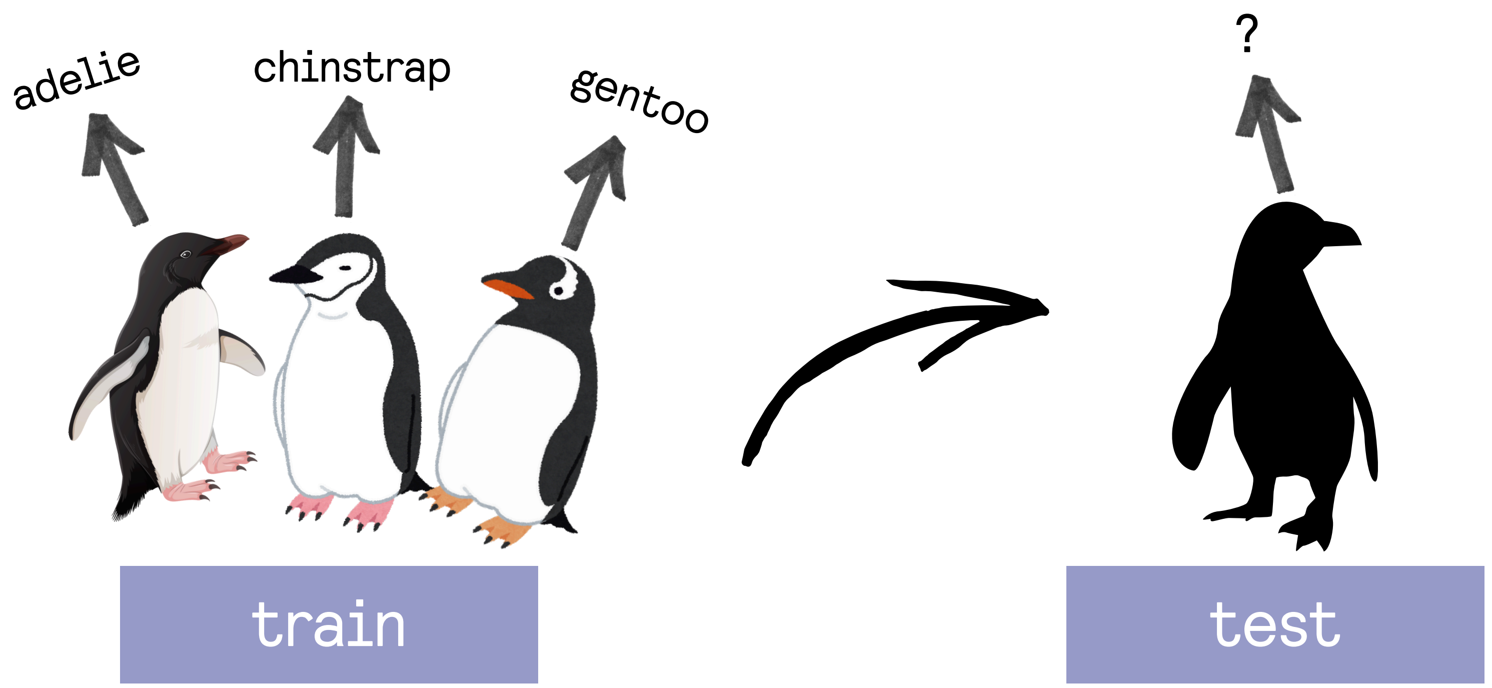
Unsupervised machine learning

- A data matrix X with n observations
- The goal is to extract from X a structure that generalizes.

Very wide variety of different problems.

Main takeaways

- Machine Learning is about extracting rules from data that generalize to new observations
- We work with:
    - a data matrix “X” of shape n_samples x n_features
    - a target “y” of length n_samples for supervised models:
        - continuous numbers for regression
        - discrete classes for classification


adelie
chinstrap
gentoo
?
train
test