A very brief Introduction to Predictive Models for Data Science

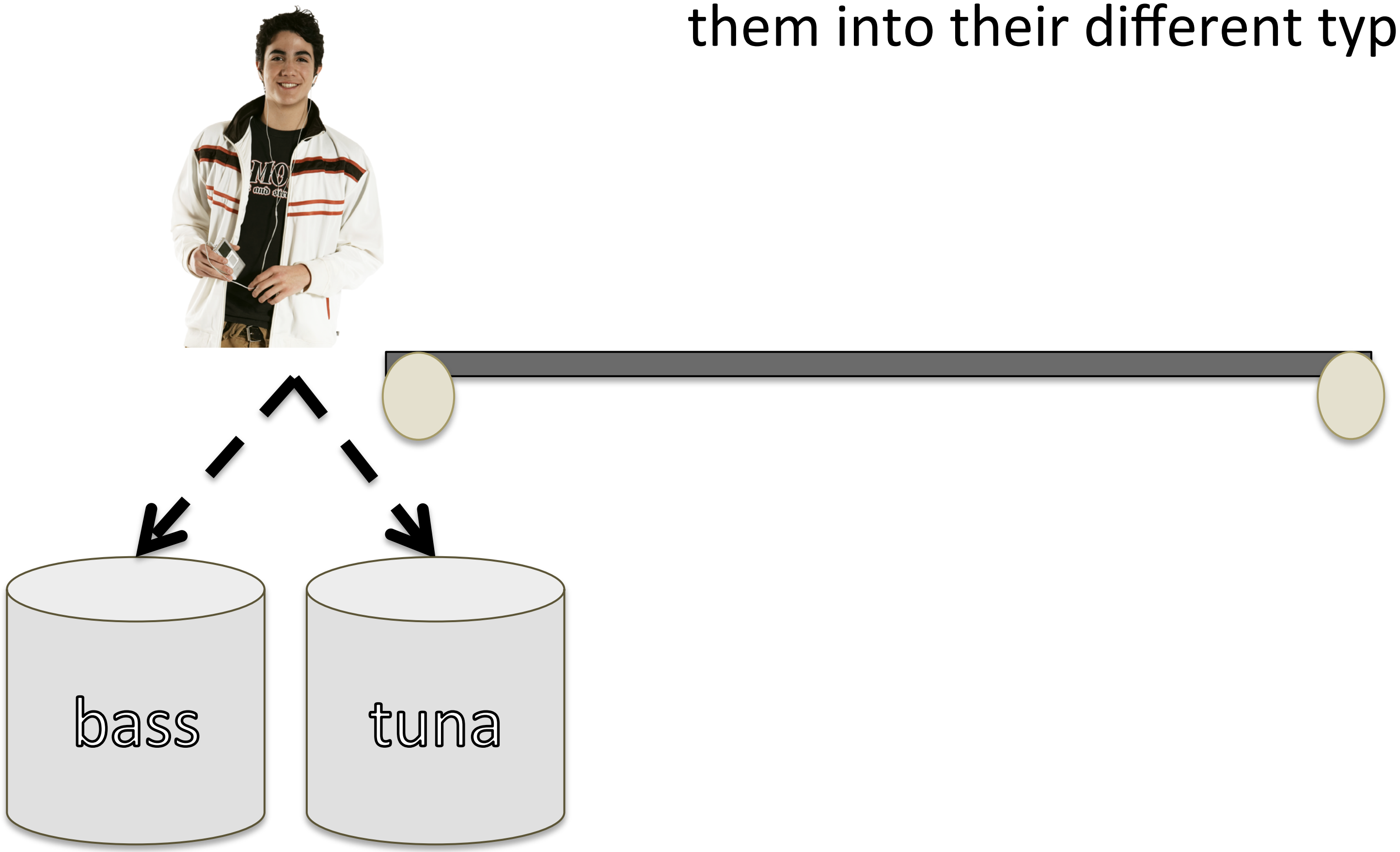
Introduction to Data Science

Slides by Mark Carman

Motivating example

(Example from Duda & Hart, Pattern Classification & Scene Analysis, 1973)

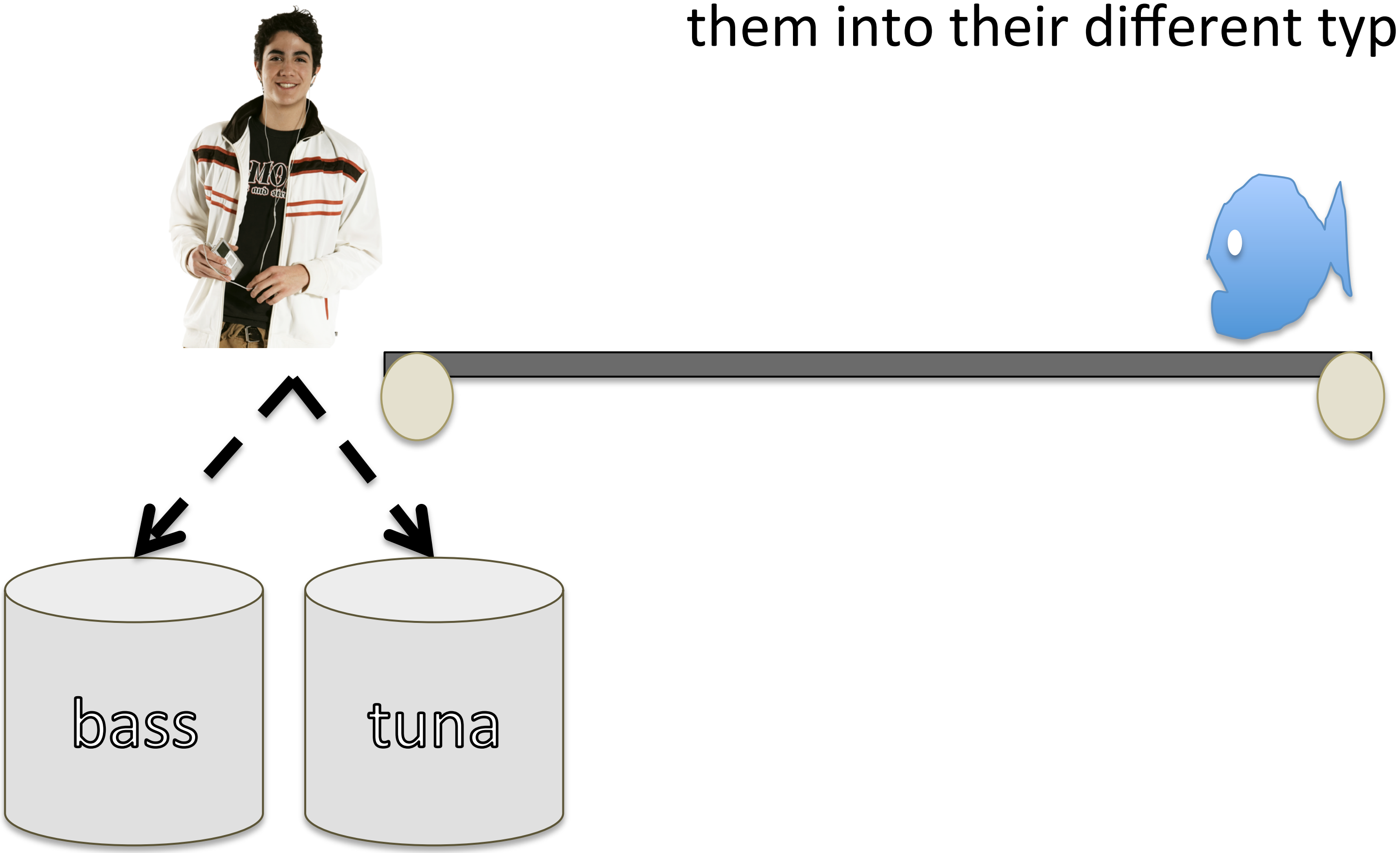
On a fishing boat, a conveyor belt loads fish and a worker separates them into their different types



Motivating example

(Example from Duda & Hart, Pattern Classification & Scene Analysis, 1973)

On a fishing boat, a conveyor belt loads fish and a worker separates them into their different types

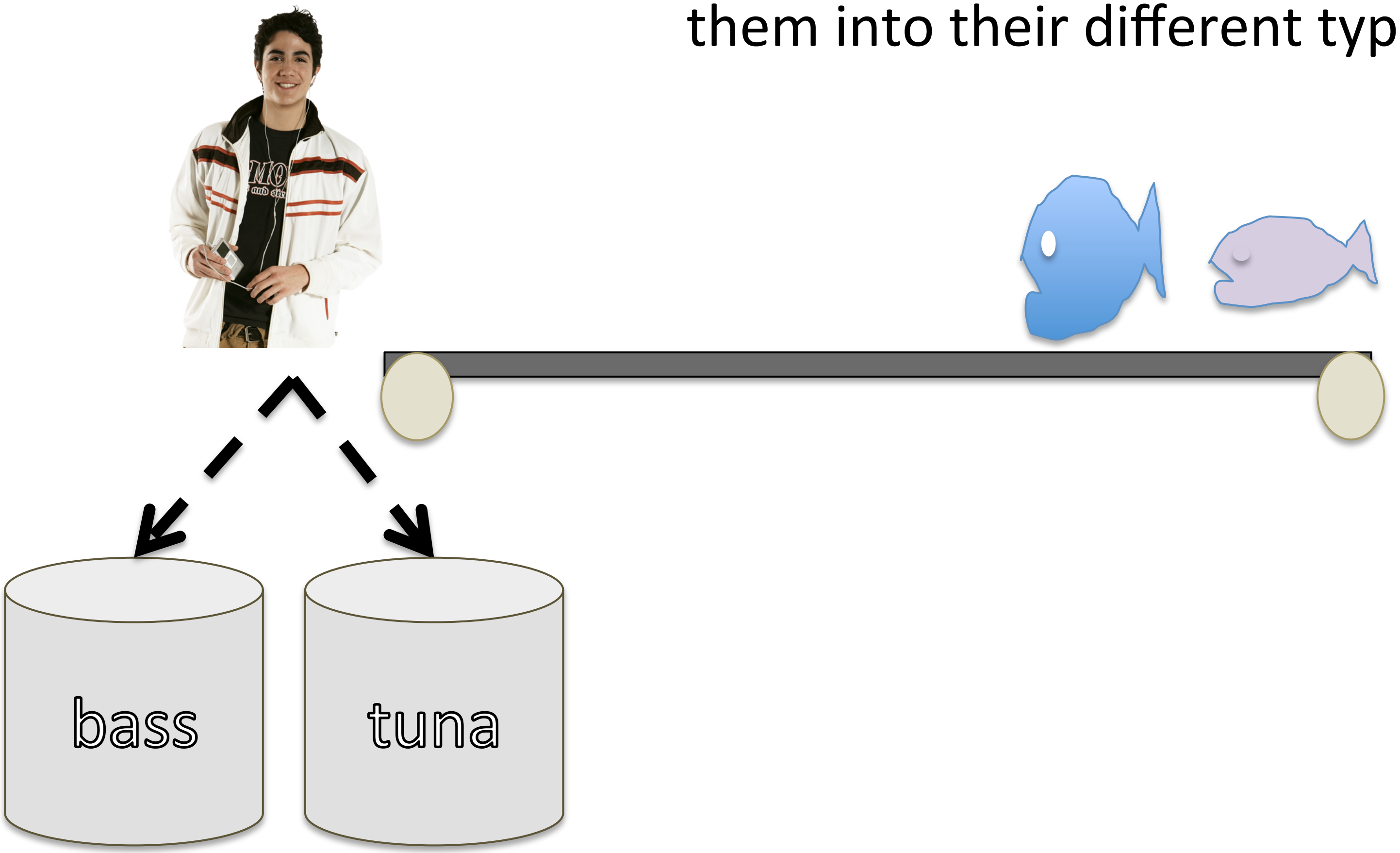


bass

tuna

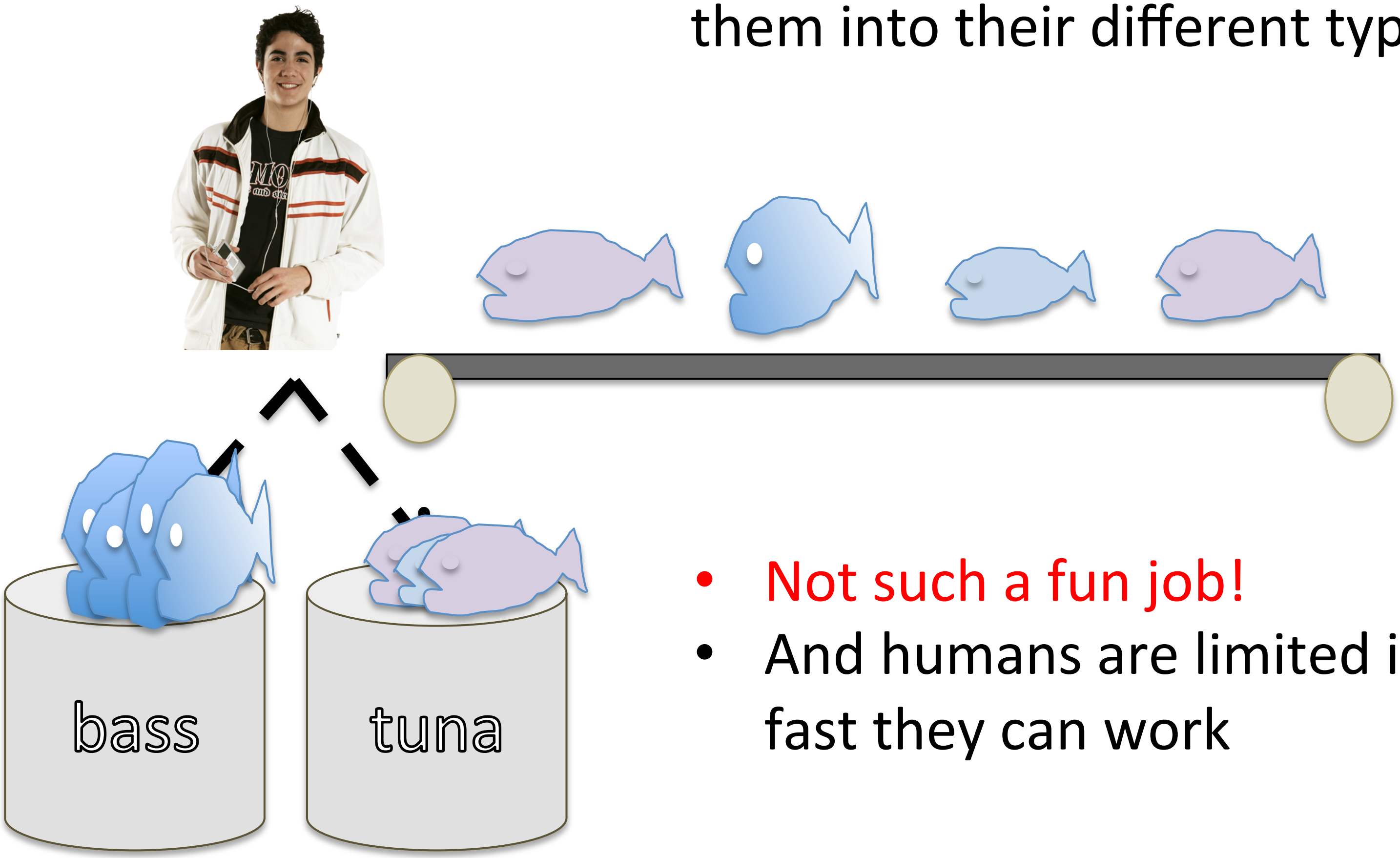
Motivating example

On a fishing boat, a conveyor belt loads fish and a worker separates them into their different types



Motivating example

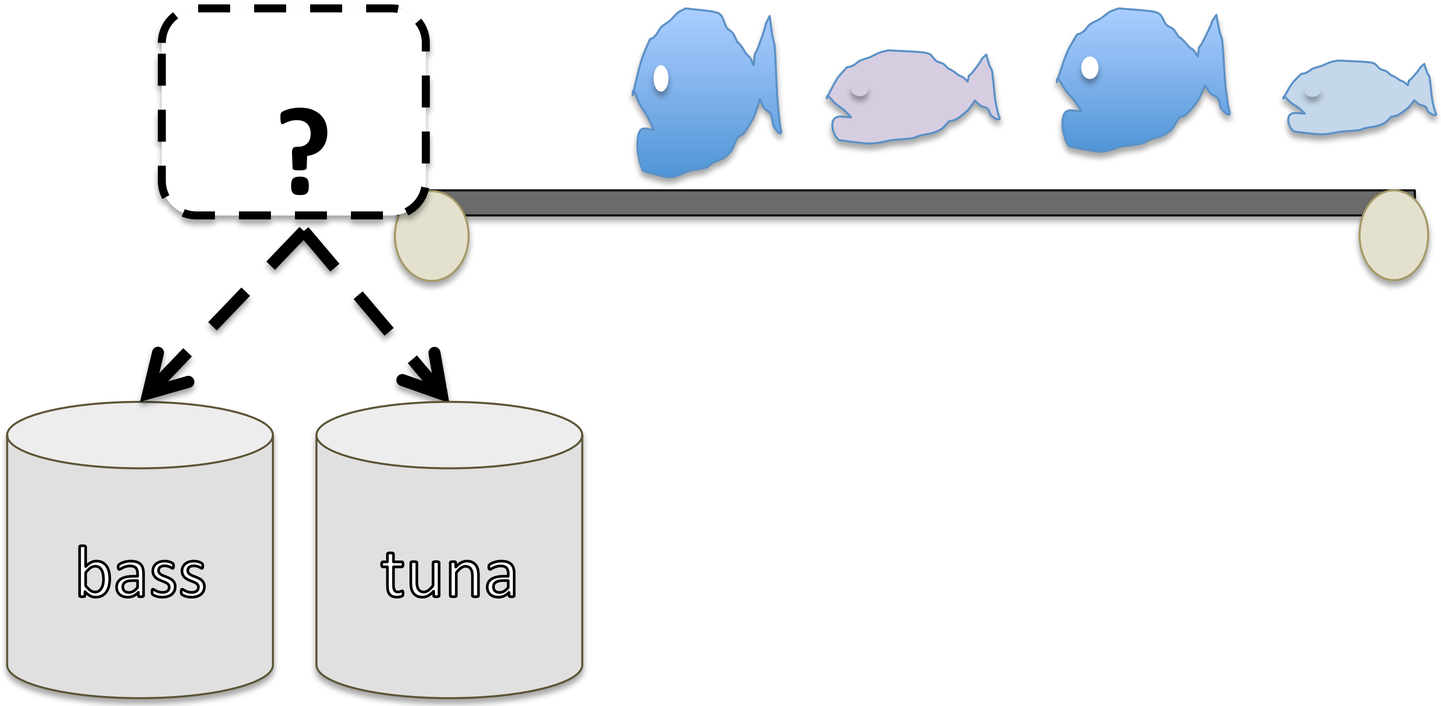
On a fishing boat, a conveyor belt loads fish and a worker separates them into their different types



- Not such a fun job!
- And humans are limited in how fast they can work

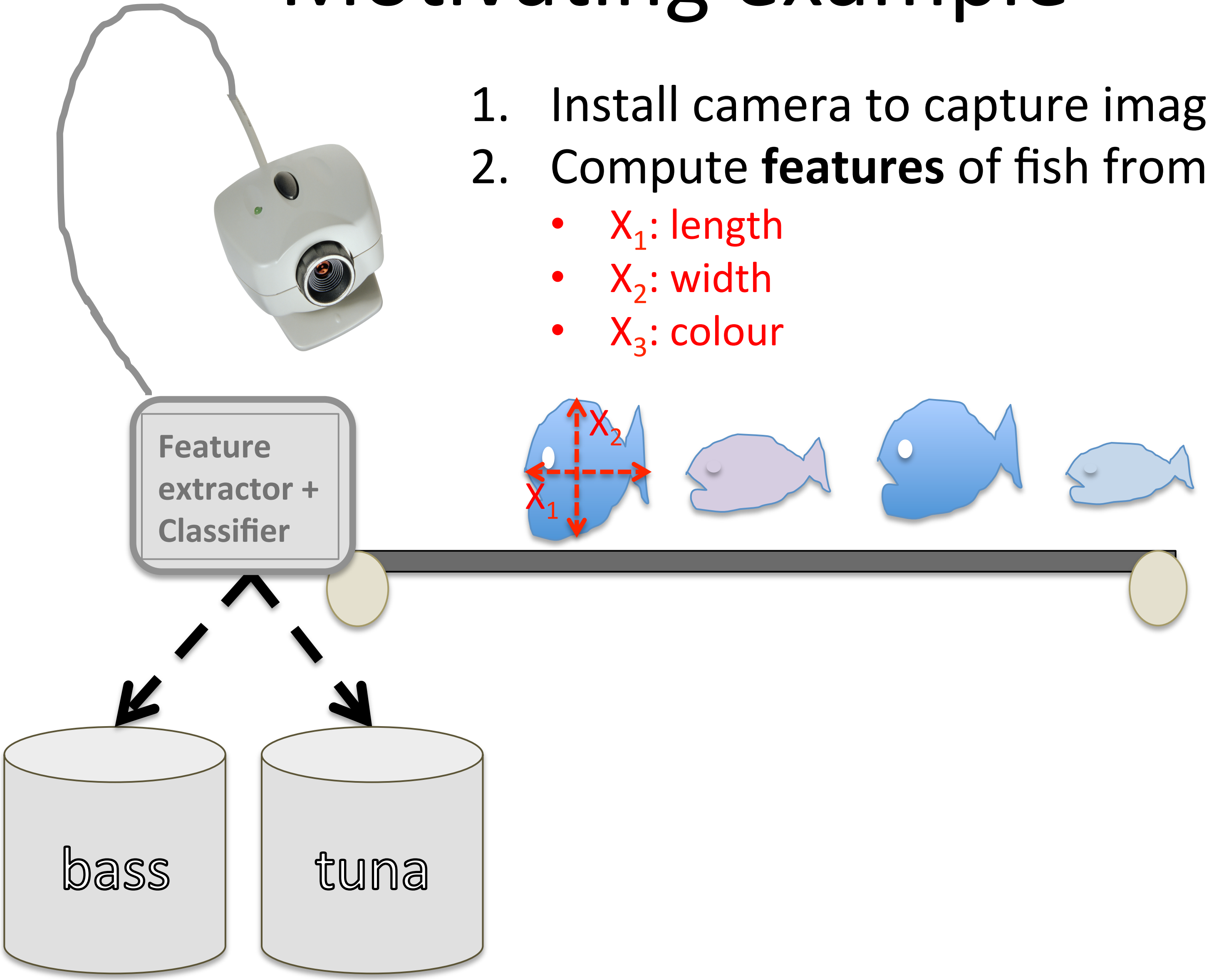
Motivating example

Question: Can we build a system to do the task automatically?



Motivating example

1. Install camera to capture image of fish
2. Compute **features** of fish from image:
   - X_1: length
   - X_2: width
   - X_3: colour



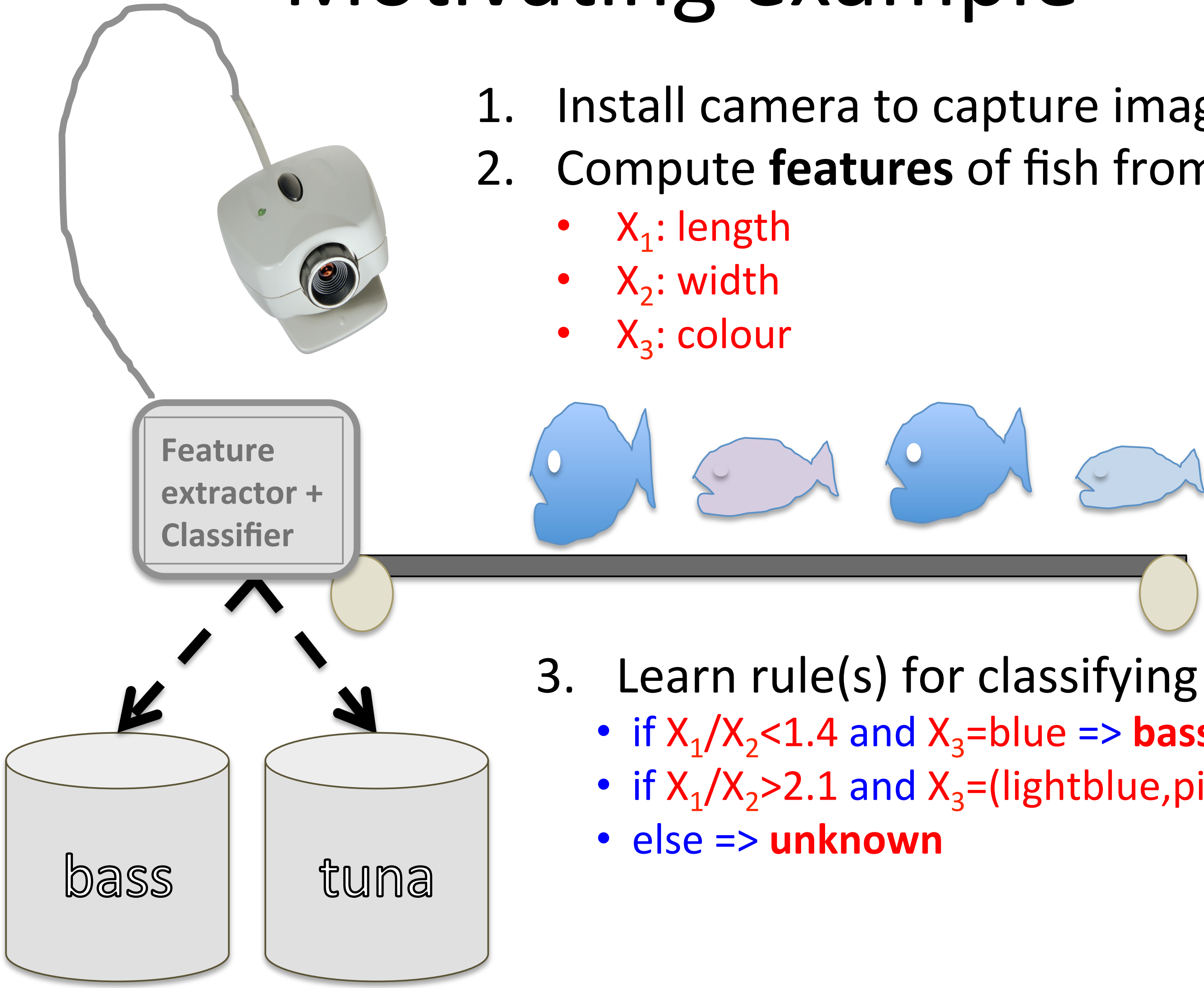
Feature extractor + Classifier

bass

tuna

Motivating example

1. Install camera to capture image of fish
2. Compute **features** of fish from image:
   - X_1: length
   - X_2: width
   - X_3: colour



3. Learn rule(s) for classifying fish
   - if $X_1/X_2<1.4$ and X_3=blue => **bass**
   - if $X_1/X_2>2.1$ and X_3=(lightblue,pink) => **tuna**
   - else => **unknown**

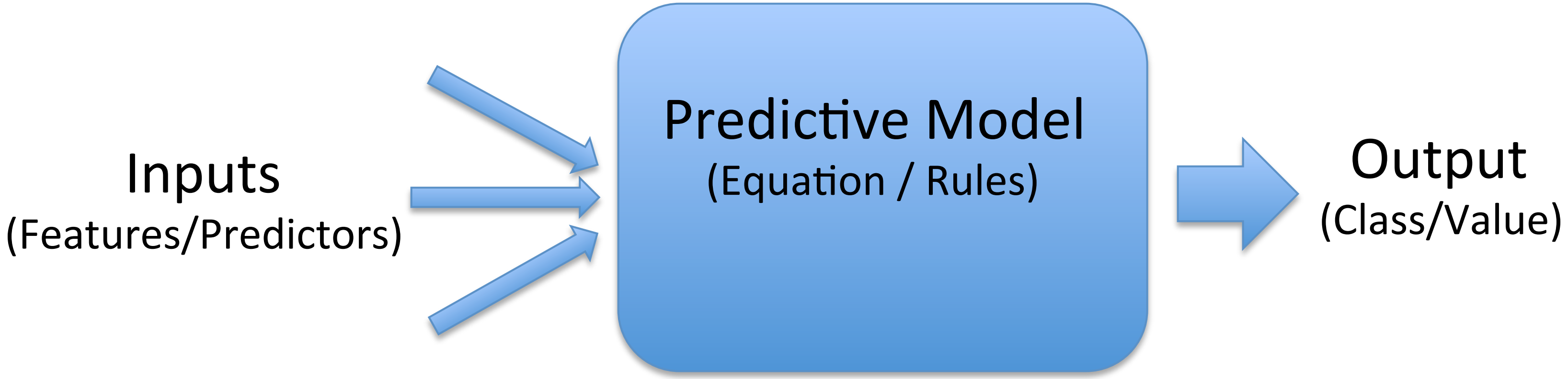
Predictive Models

A predictive model is any model that makes a prediction

- usually based on a set of features describing an object.

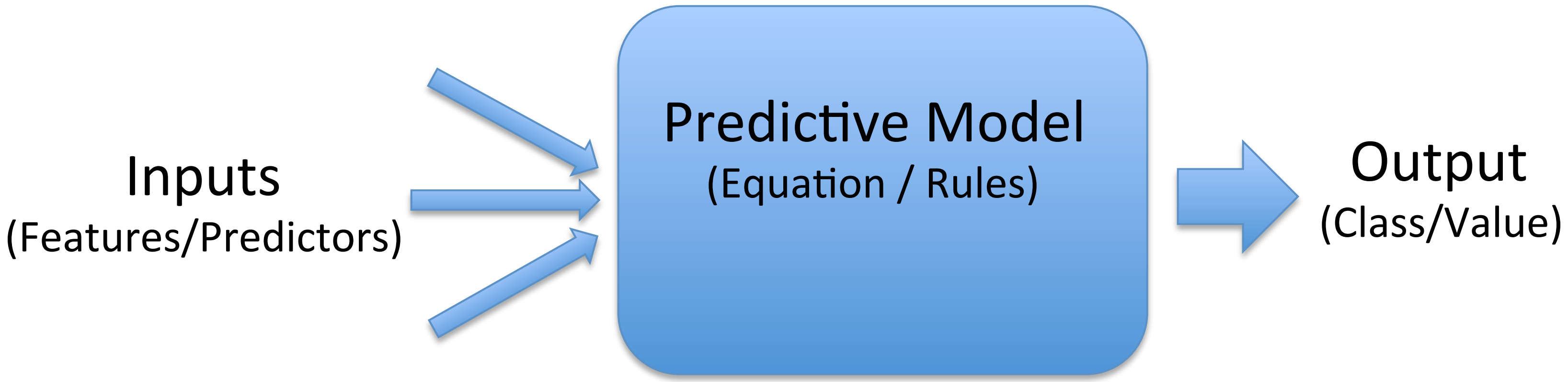
The prediction could be:

- a binary outcome (spam, not-spam)
- categorical (bass, tuna, other)
- a real value (the age of the fish)
- a vector of real values (probability of bass, tuna, etc.)
- Etc.



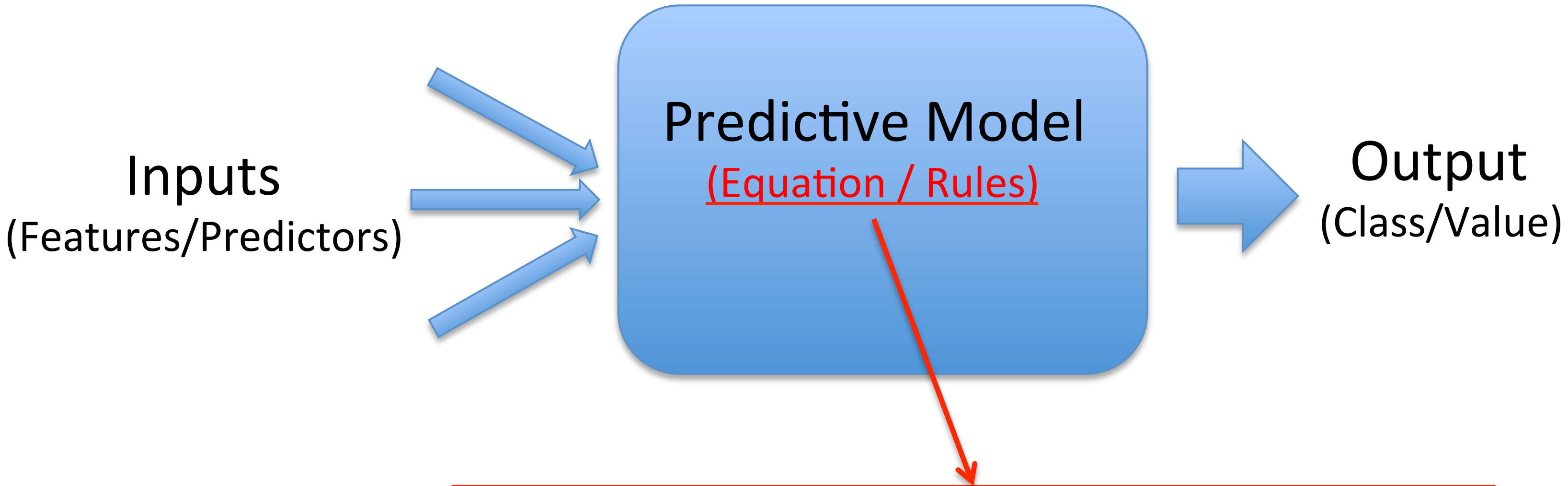
Predictive Models

- If the predicted value is binary/categorical we usually refer to the model as a classifier
- If it predicts real values we refer to it as regression
- Although there are many other types of model (e.g. ranking, translation, etc.)



Predictive Models

The predictive model uses equations/rules to map the input features to output values



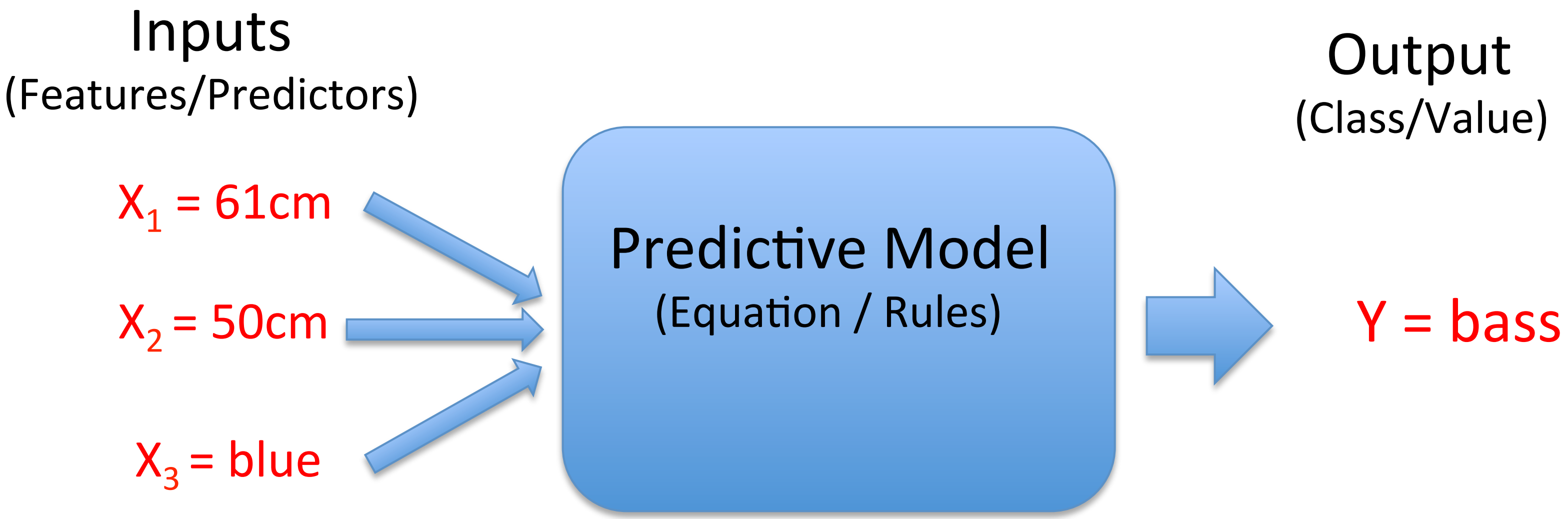
Inputs
(Features/Predictors)

Predictive Model
(Equation / Rules)

Output
(Class/Value)

if $X_1/X_2<1.4$ and X_3=blue => **bass**
if $X_1/X_2>2.1$ and X_3=(lightblue,pink) => **tuna**
else => **unknown**

Predictive Model



Inputs

(Features/Predictors)

X_1 = 61cm

X_2 = 50cm

X_3 = blue

Predictive Model

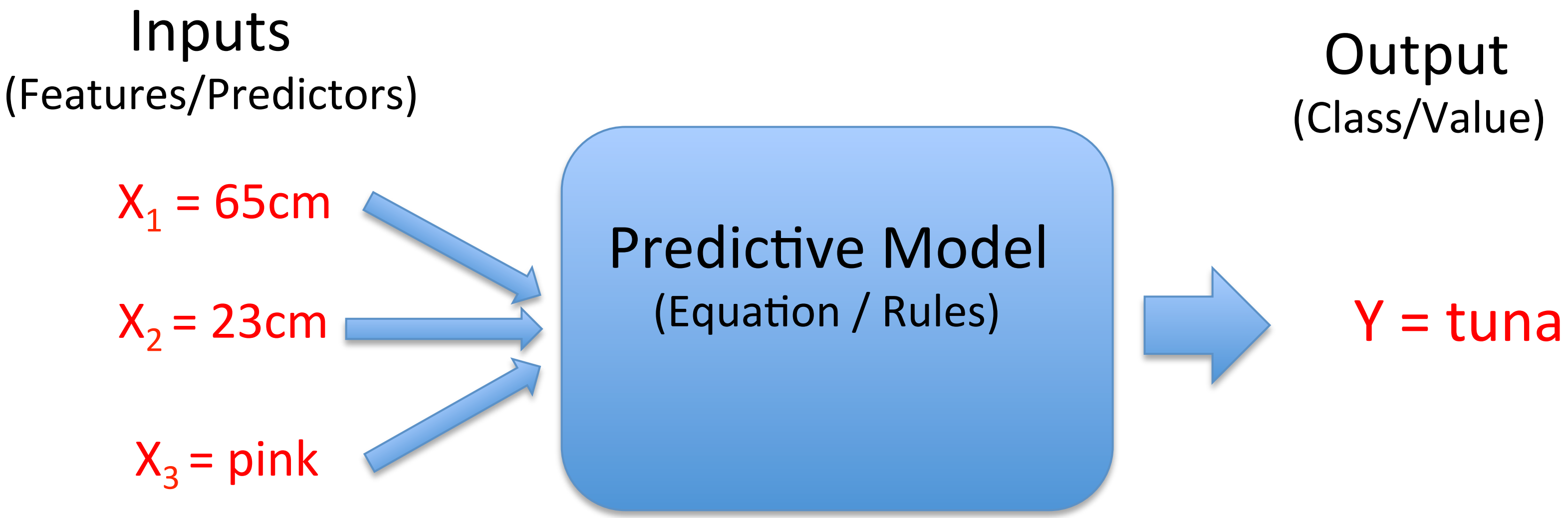
(Equation / Rules)

Output

(Class/Value)

Y = bass

Predictive Model



Inputs
(Features/Predictors)

X_1 = 65cm

X_2 = 23cm

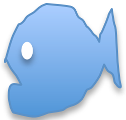
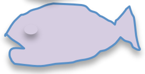
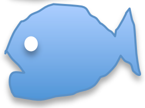
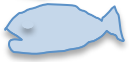
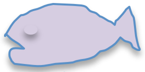
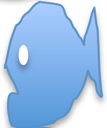
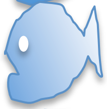
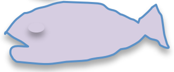
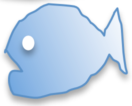
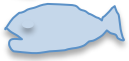
X_3 = pink

Predictive Model
(Equation / Rules)

Output
(Class/Value)

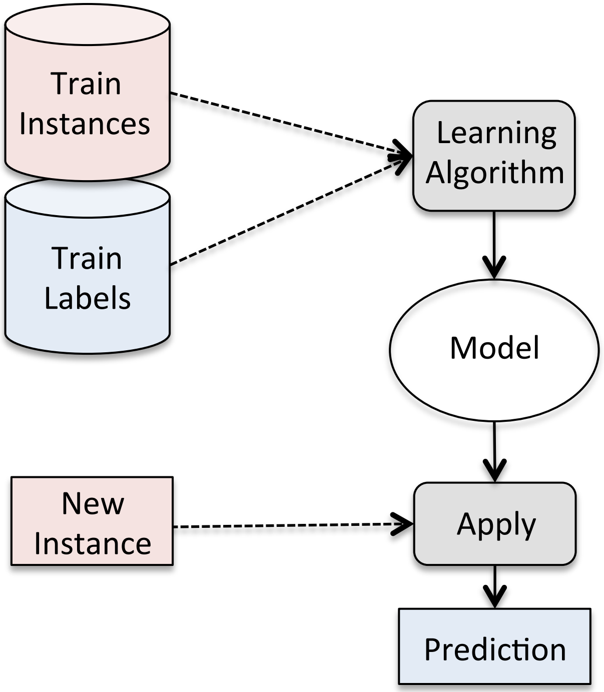
Y = tuna

Models are learnt from Examples

Instance	X1 = length	X2 = width	X3 = colour	Y = class
	55	51	blue	**bass**
	65	23	pink	**tuna**
	67	54	blue	**bass**
	54	20	light-blue	**tuna**
	62	26	pink	**tuna**
	44	62	blue	**bass**
	47	55	light-blue	**bass**
	73	31	pink	**tuna**
	54	48	light-blue	**bass**
	57	23	light-blue	**tuna**

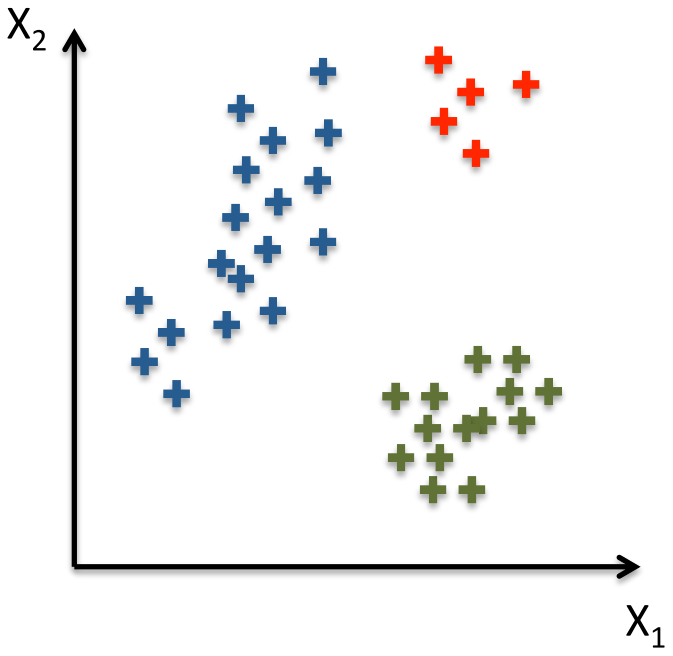
Training a Model

Predictive models are learnt from training data and then applied to make predictions on new instances



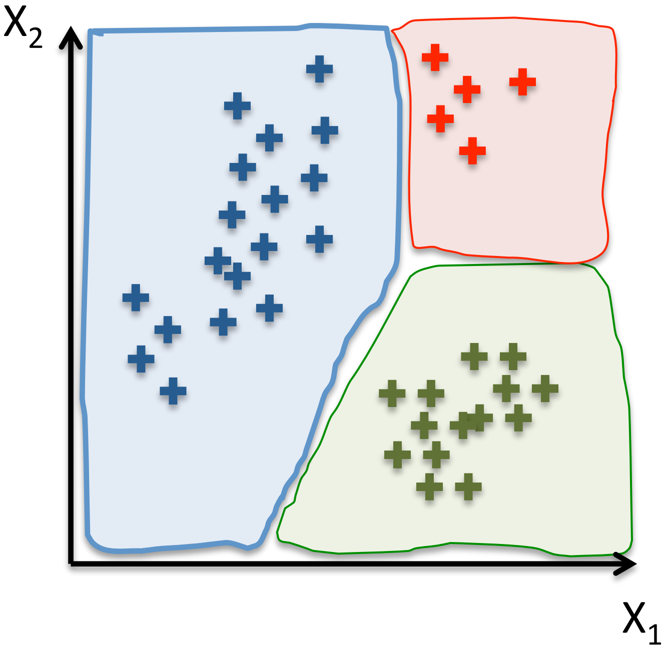
How are models learnt?

- Each training instance (fish in our case) is just a point in some feature space
- Here the colour denotes the class
  - (blue = bass, green = tuna, red = unknown)

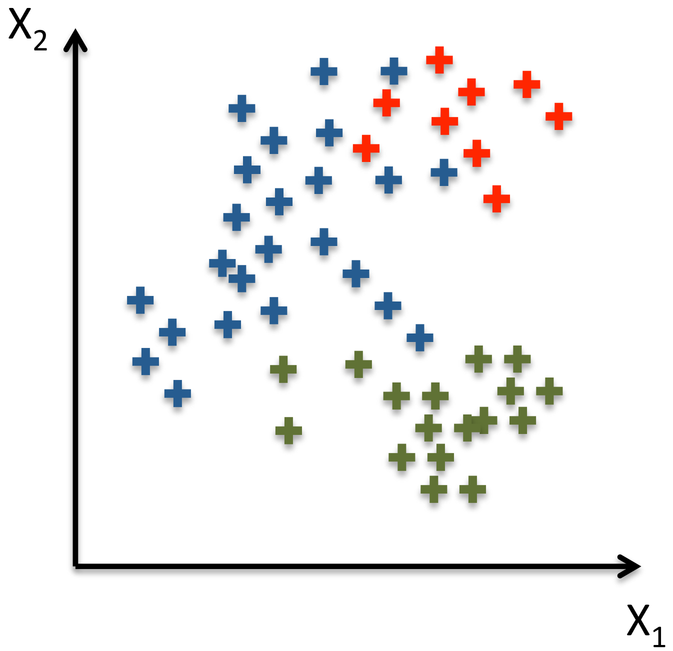


How are models learnt?

Many (classification) learning algorithms work by dividing the feature space into regions of the same type

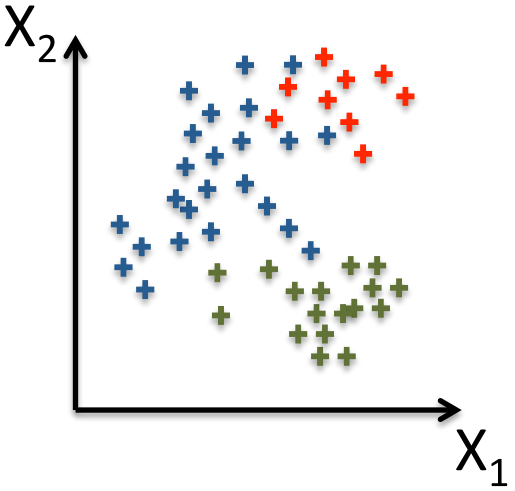


In practice

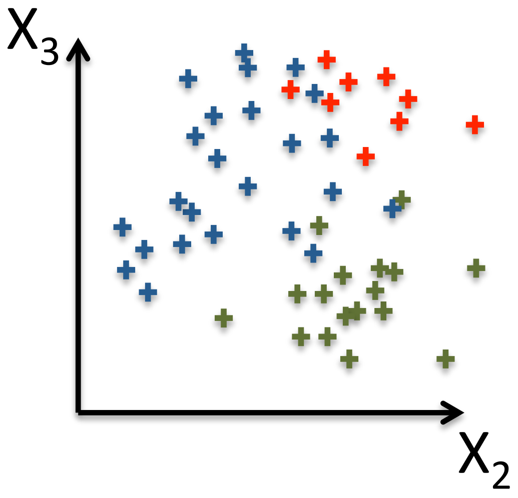


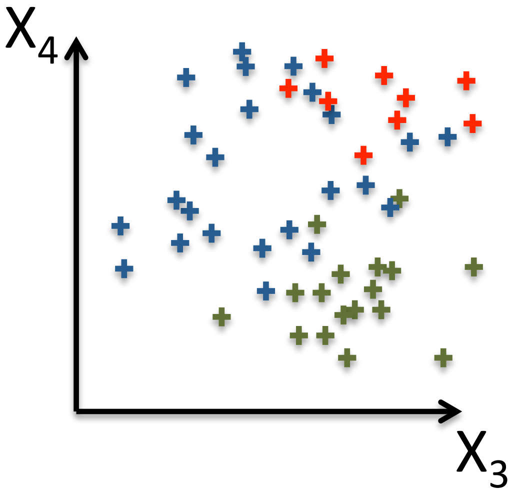
- In practice, the data is usually overlapping
- making it hard to separate the classes

In practice



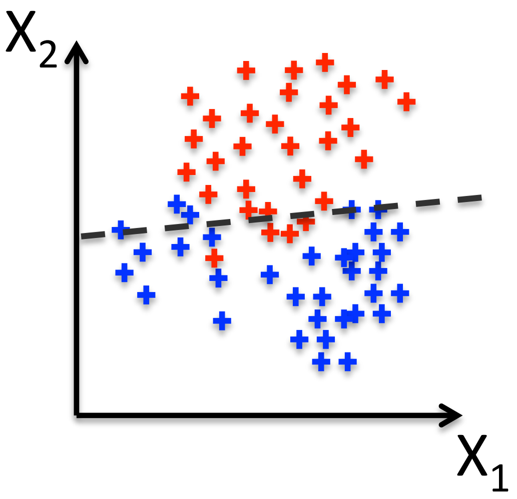
- and we have many feature dimensions
- with some features more useful than others





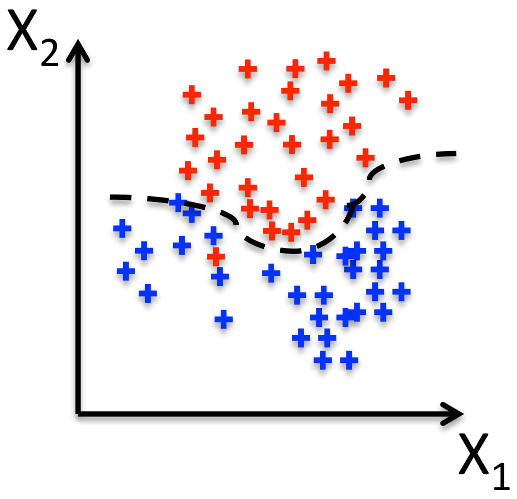
Different Models

There are many different types of models that we can train to classify objects



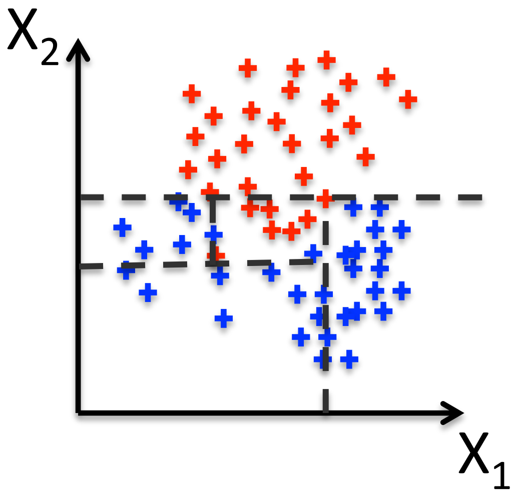
Linear classifiers

e.g. Logistic Regression, Linear SVMs



Non-linear Classifiers

e.g. Neural Nets, SVM with RBF kernel

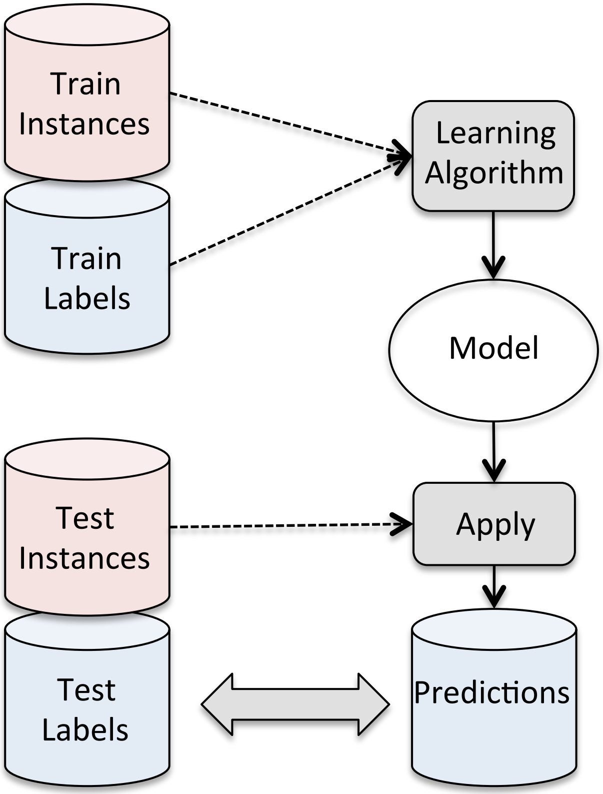


Decision Tree Learners

e.g. Random forests

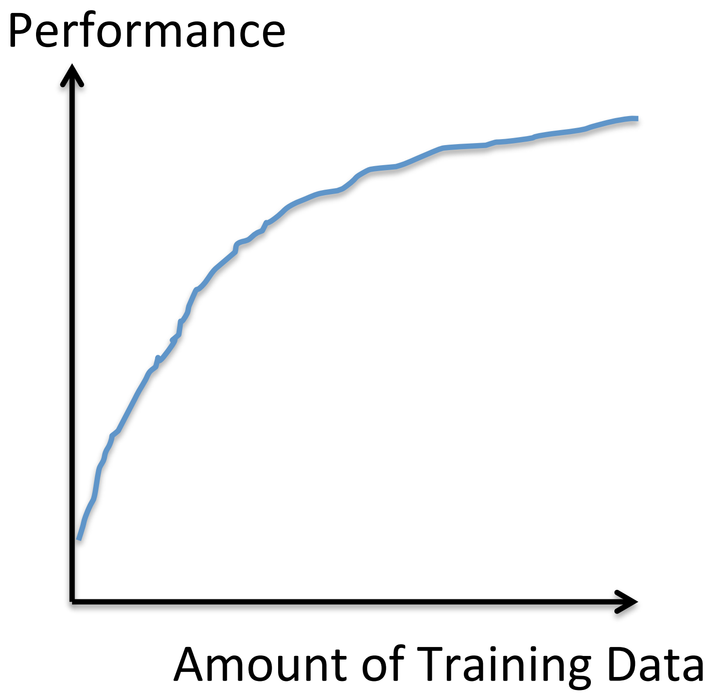
Testing models

We evaluate predictive models based on how well they predict the labels for test instances (not used in training)



Evaluation

Performance of predictive models



Generally:

- The more training data the better the test performance
- And (providing there is sufficient training data) the more features the better performance

End of Introduction!

- We'll talk more about predictive models in the coming weeks, especially in module 5.