FIT1043 Introduction to Data Science

Week 5

Ian K T Tan

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Week 4 Coverage
Data Sources
Data Wrangling

Week 4 Coverage

Data Sources and Wrangling

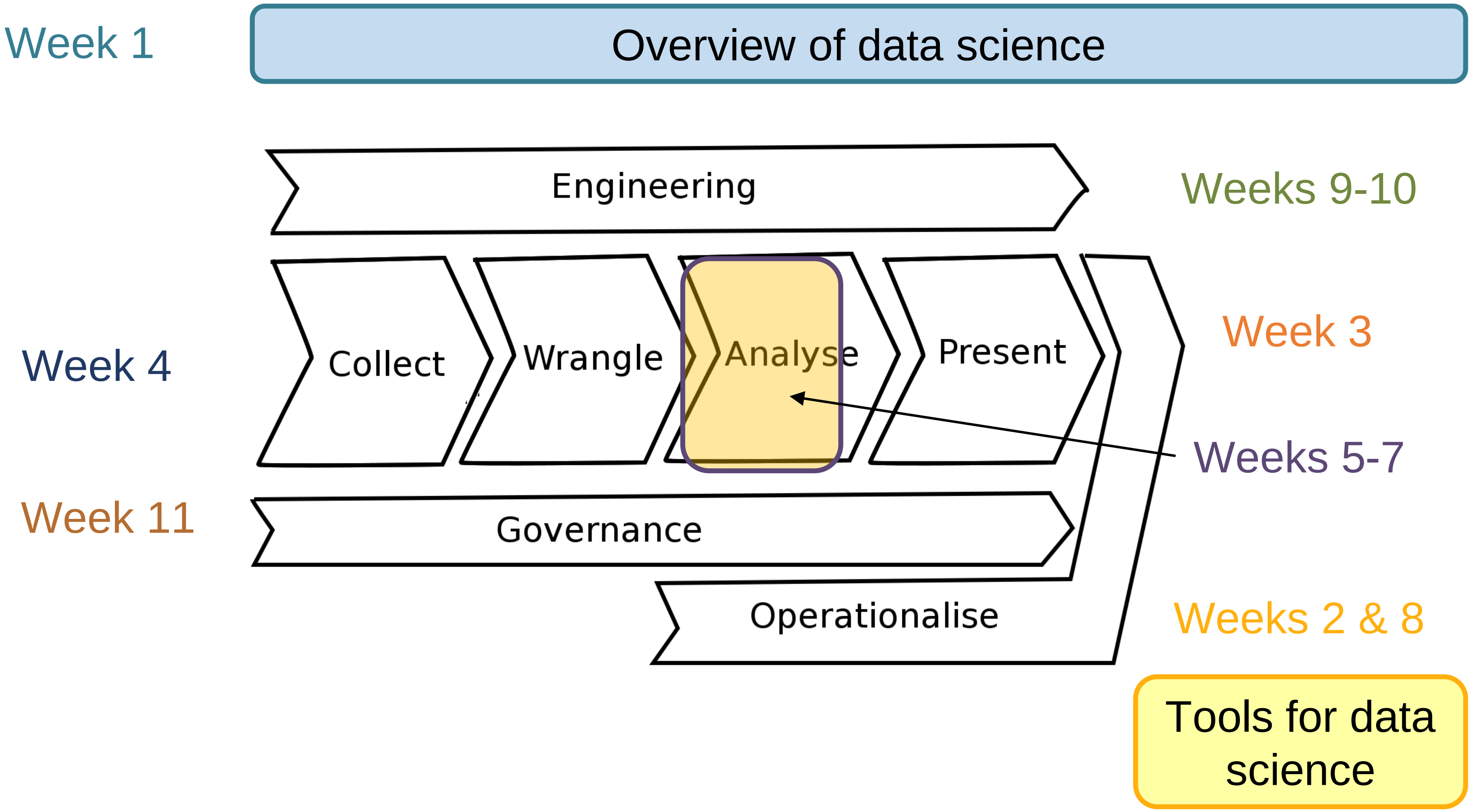
Open Data

API

Data Quality

Data Auditing

Techniques to handle data quality issues



Week 1

Overview of data science

Engineering

Weeks 9-10

Week 4

Collect

Wrangle

Analyse

Present

Week 3

Weeks 5-7

Week 11

Governance

Operationalise

Weeks 2 & 8

Tools for data science

Week	Activities	Assignments
1	Overview of data science	
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	Assignment 2
9	Characterising data and "big" data	
10	Big data processing	
11	Issues in data management	Assignment 3
12	Industry guest lecture (tentative)	

Week 5 Outline

Introduction to Data Analysis

- What is model?
- What are predictive models?
- How to evaluate predictive models?

Overview of Machine Learning

- Machine learning styles
- What is learning theory
- Linear Regression
- Polynomial regression

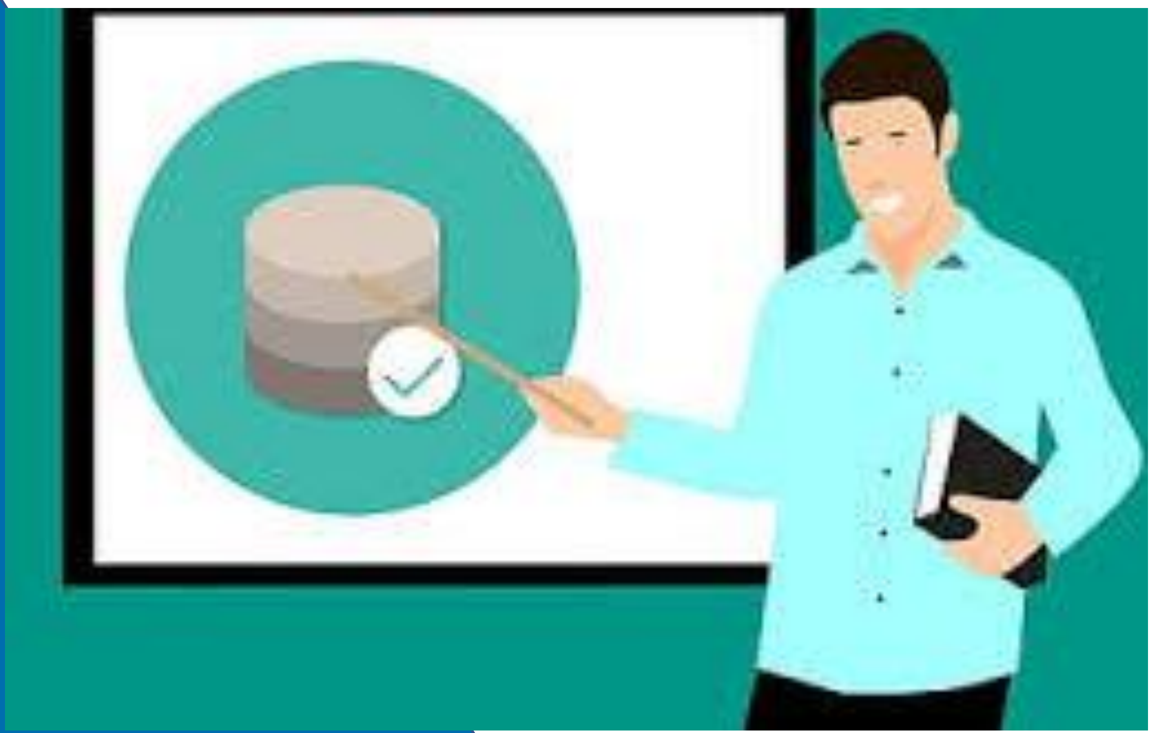
Learning Outcomes

Week 5

By the end of this week you should be able to:

- Explain what are models and predictive models
- Analyse predictive models in different examples
- Understand how to evaluate predictive models
- **Analyse how to estimate linear regression model**
- Apply linear regression and polynomial regression on different data sets using Python

Data Model



MONASH University

What is a Model?

What is a Model?

Can you draw this …

What is a Model?

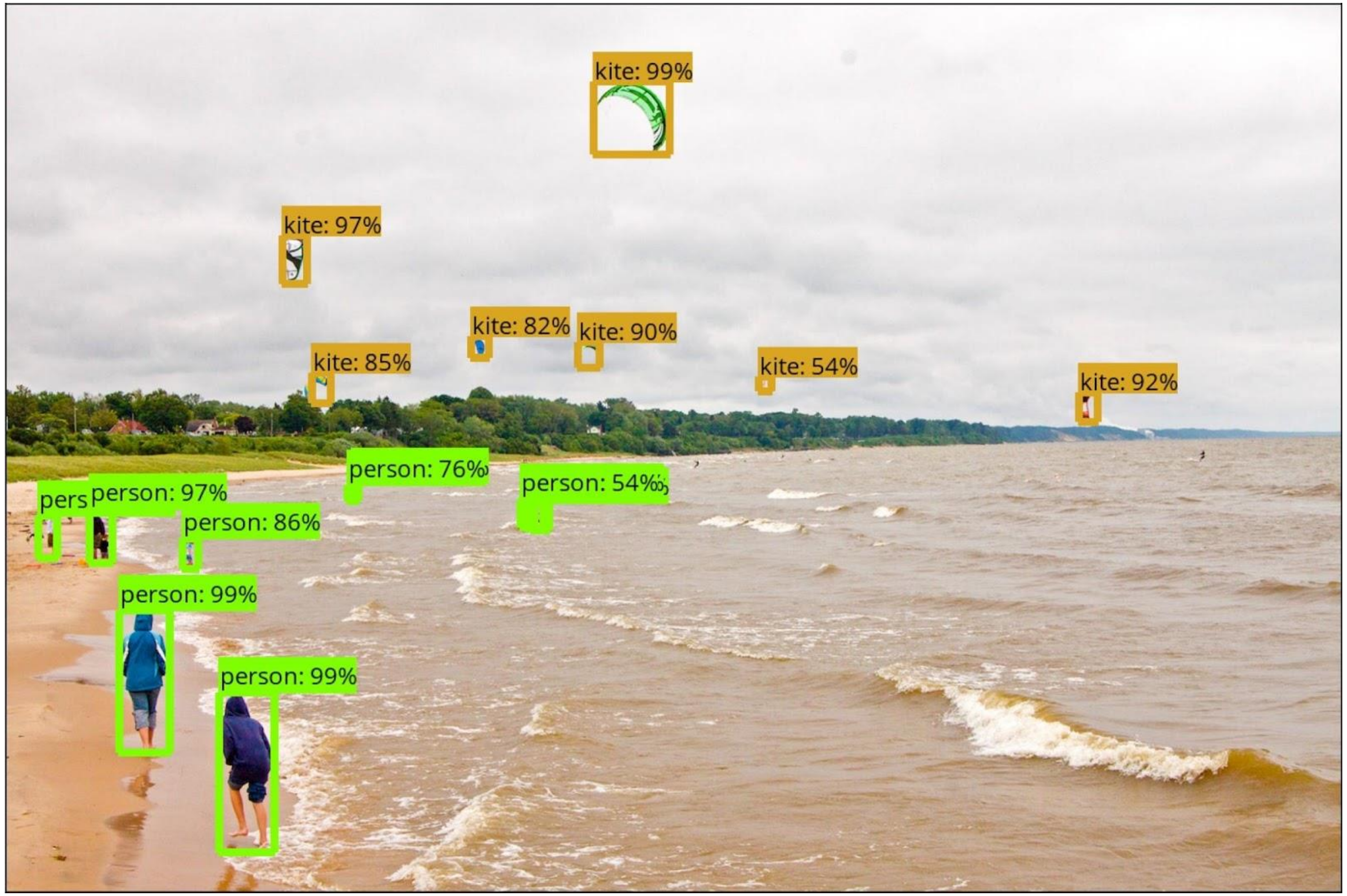
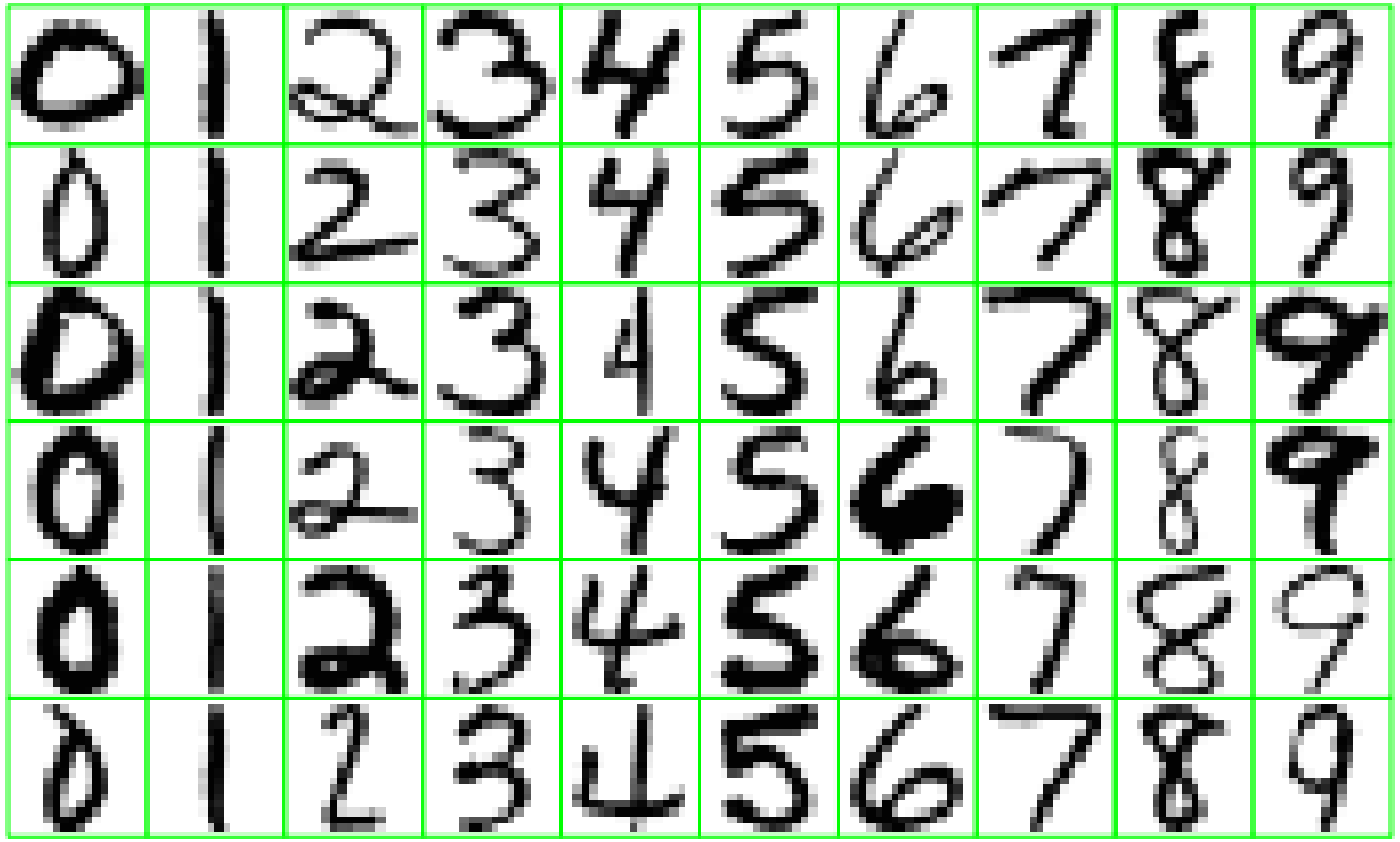


Image Source: https://medium.com/@WuStangDan/step-by-step-tensorflow-object-detection-api-tutorial-part-1-selecting-a-model-a02b6aabe39e

What is a Model?

(1) Help us understand how something works, and
(2) Help us to predict the unknown.

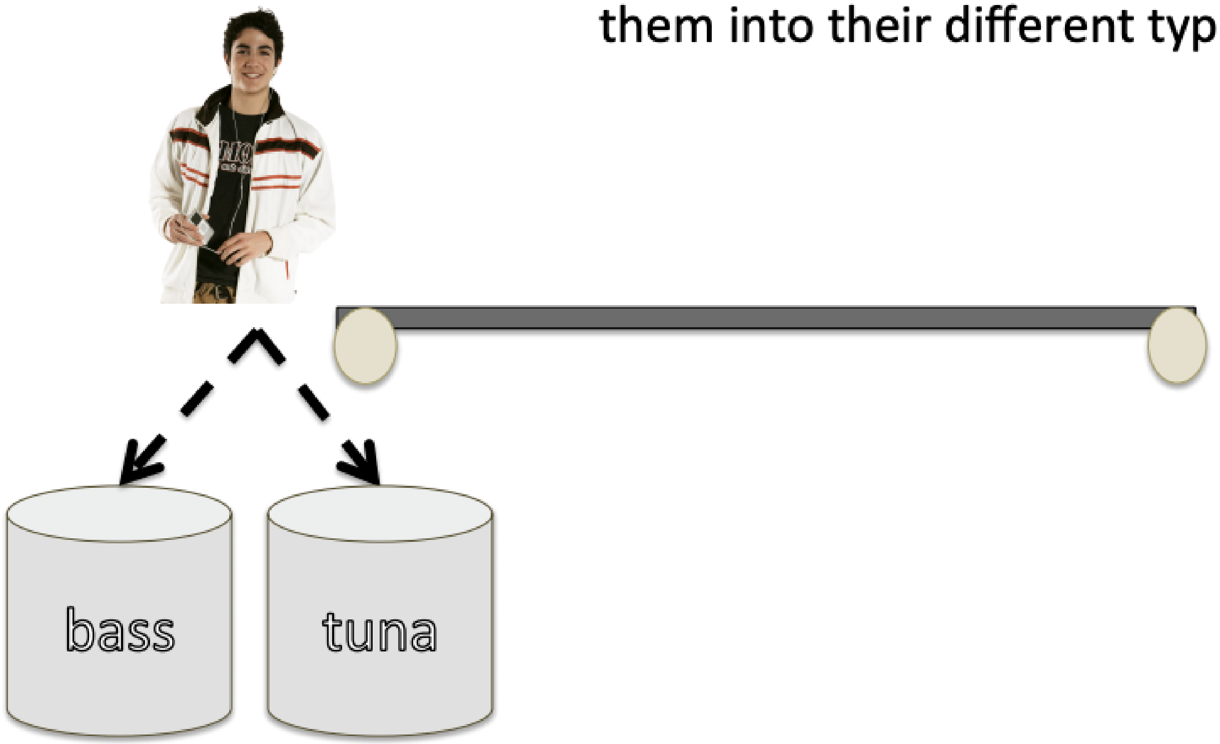


A brief Introduction to Predictive Models For Data Science

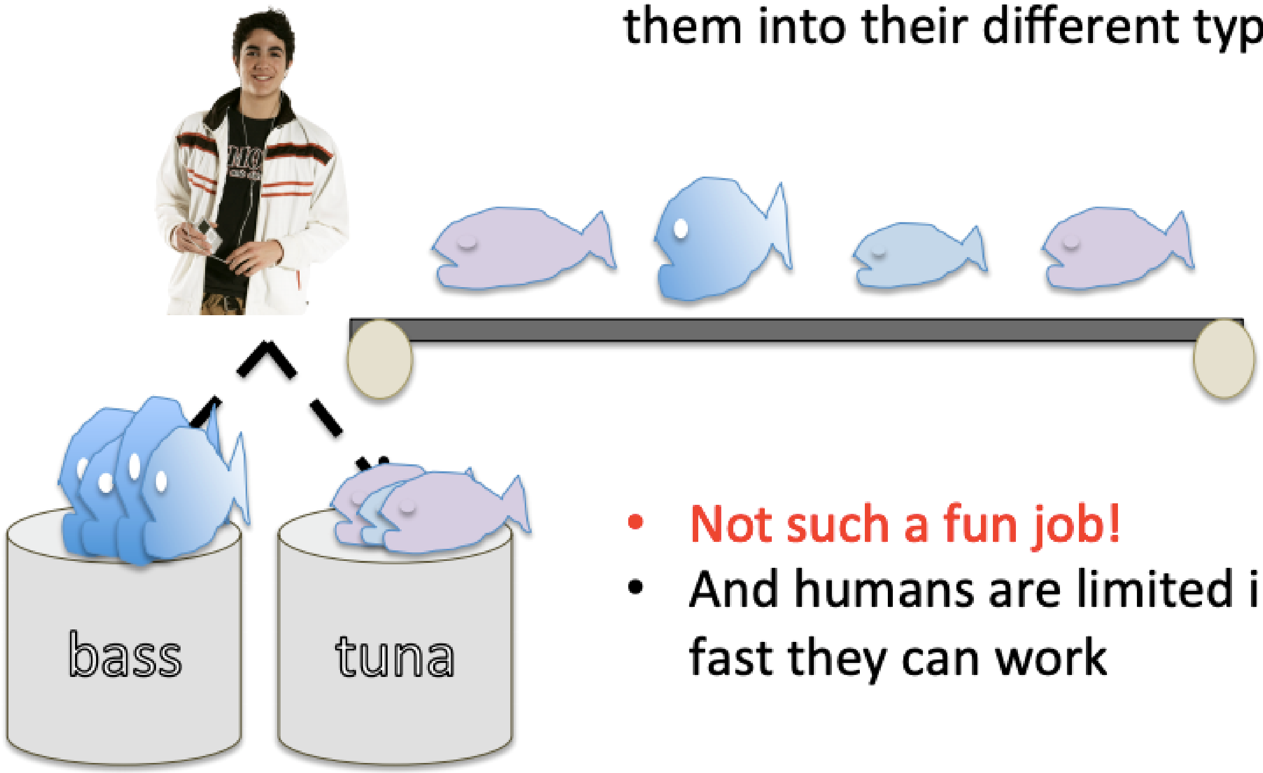
Example from Duda & Hart, Pattern Classification & Scene Analysis, 1973

Motivating Example

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



Motivating Example



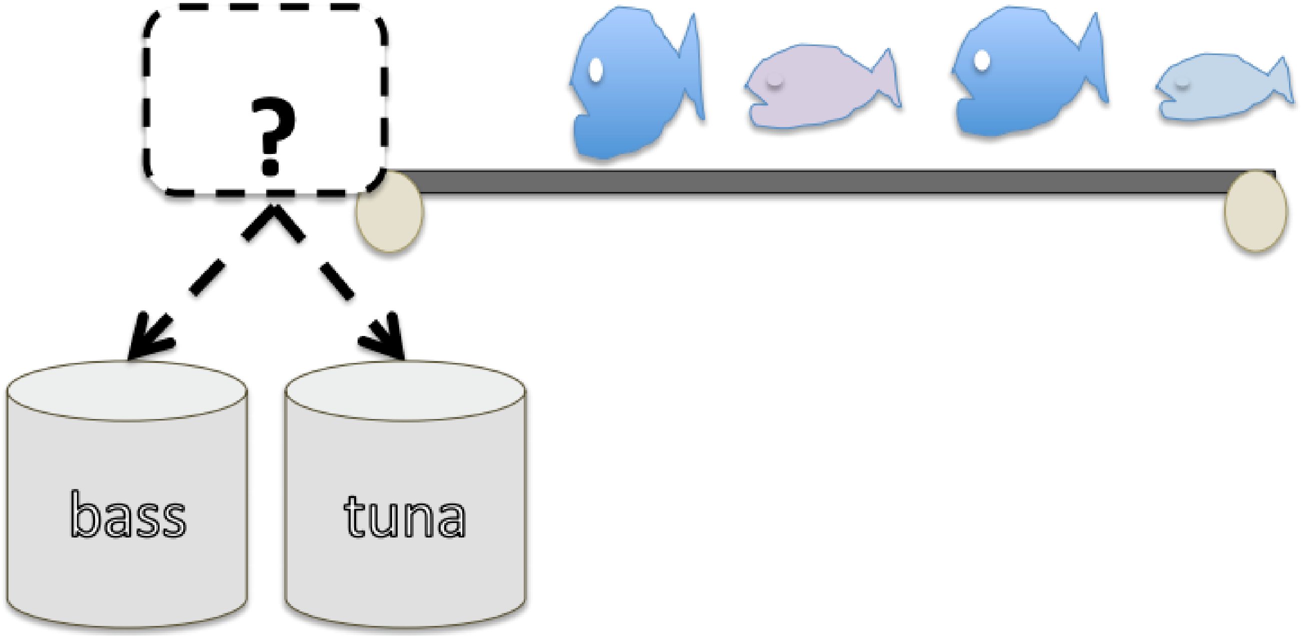
On a fishing boat, a conveyor belt loads fish and a worker separates them into their different types

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

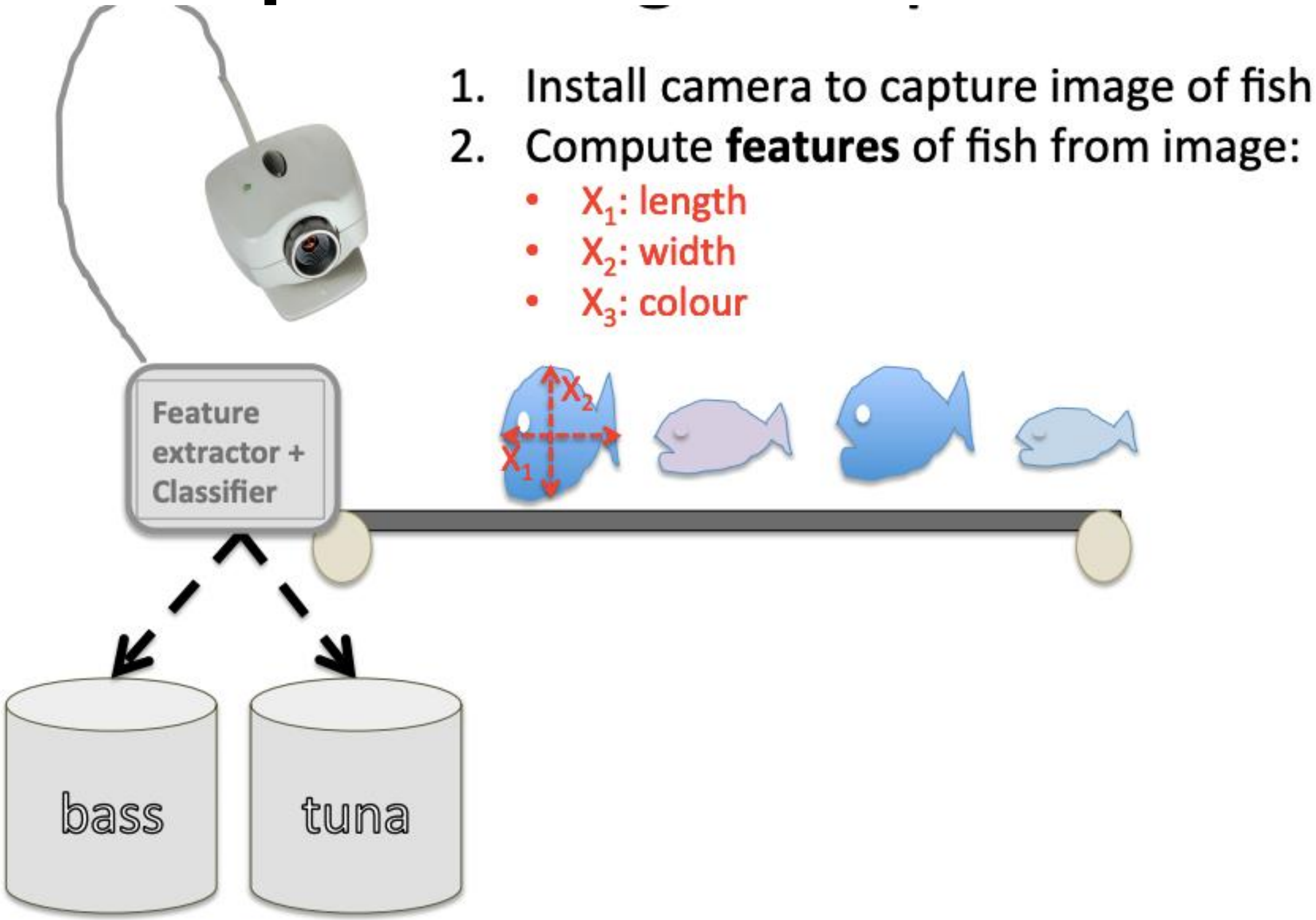
Motivating Example

Motivating Example

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

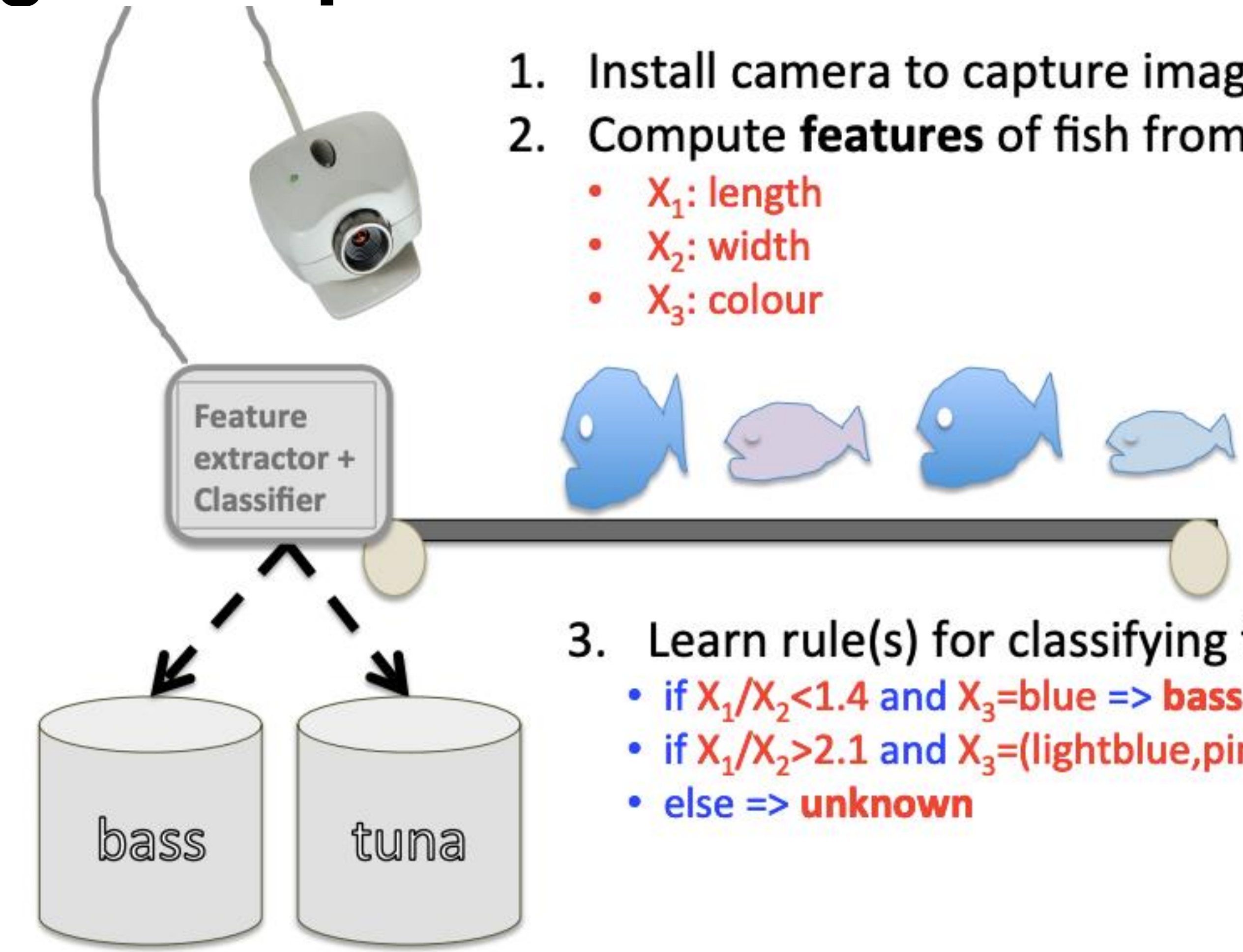


Motivating Example



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

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

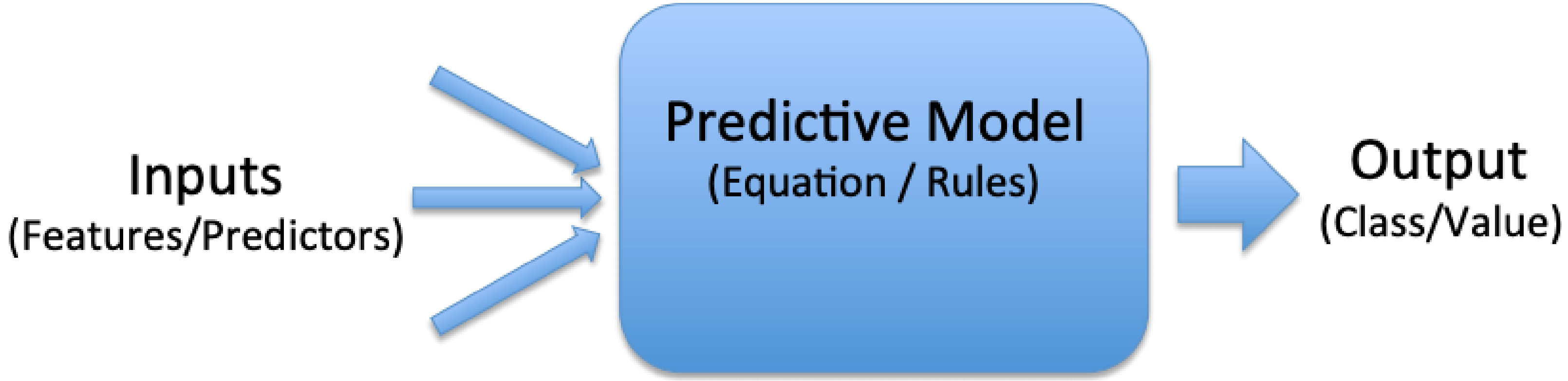
3. Learn rule(s) for classifying fish
   - if $X_1/X_2<1.4$ and X_3=blue => **bass**
   - if $X_1/X_2>2.1$ and X_3=(lightblue,pink) => **tuna**
   - else => **unknown**

Predictive Models

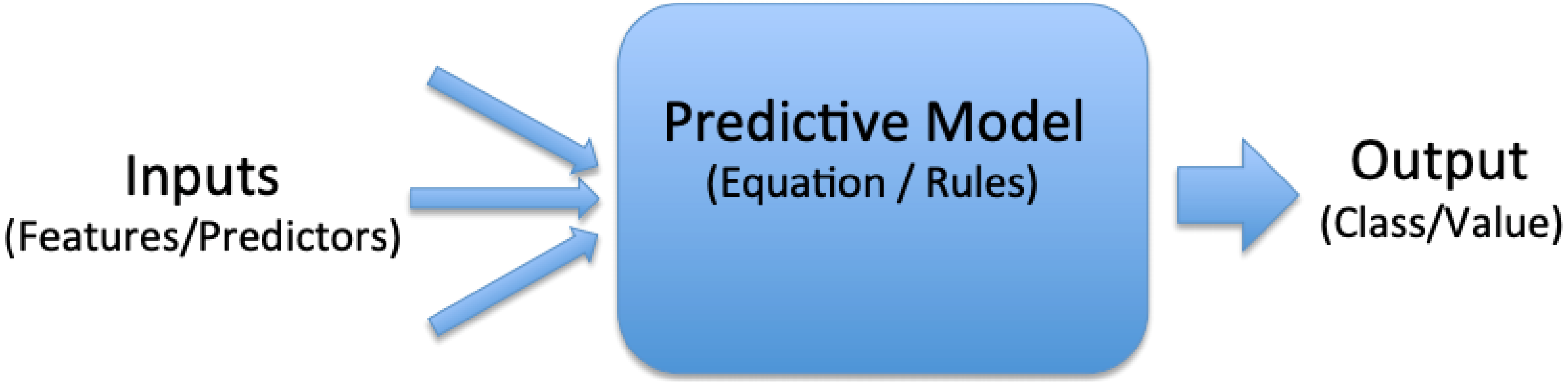
Predictive Models

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.

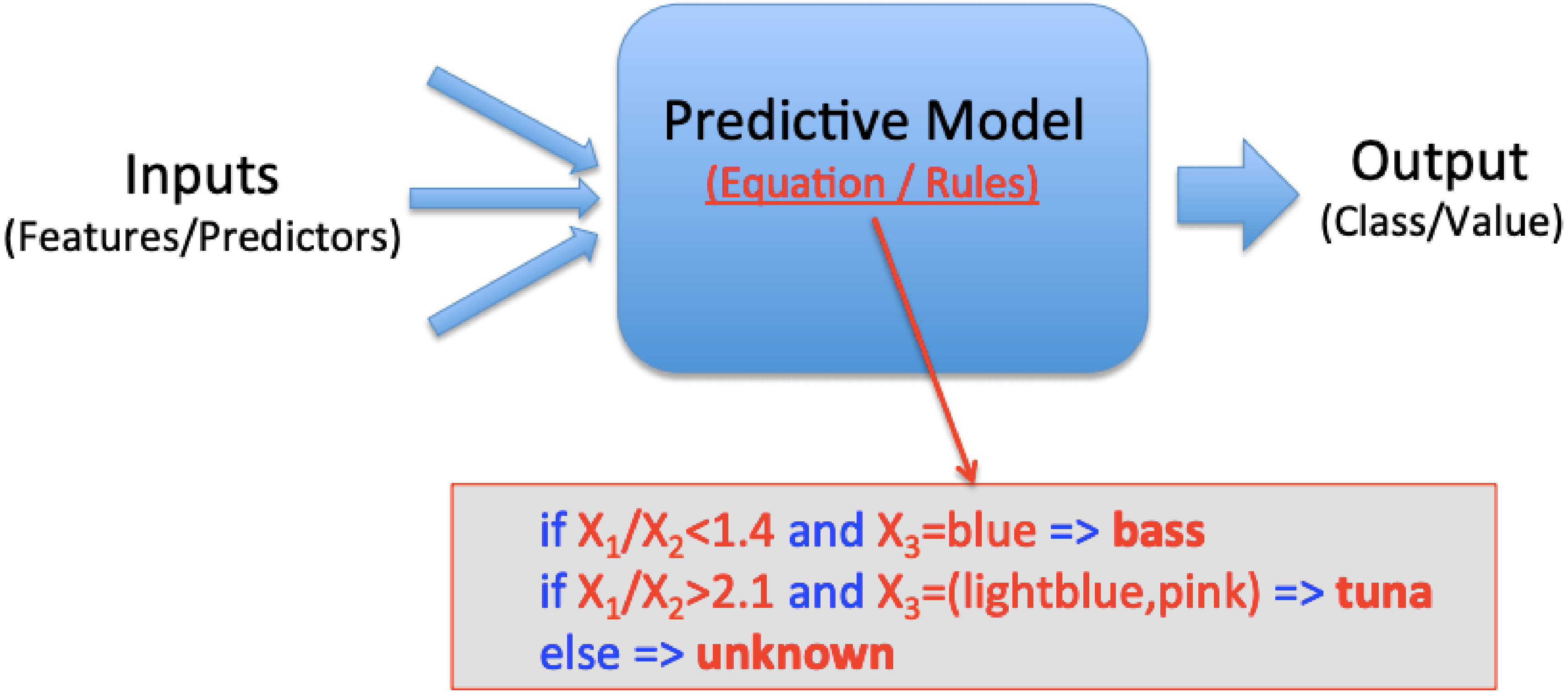


Predictive Models



- 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 models, such as rankings, translation (your predictive words) and so on.

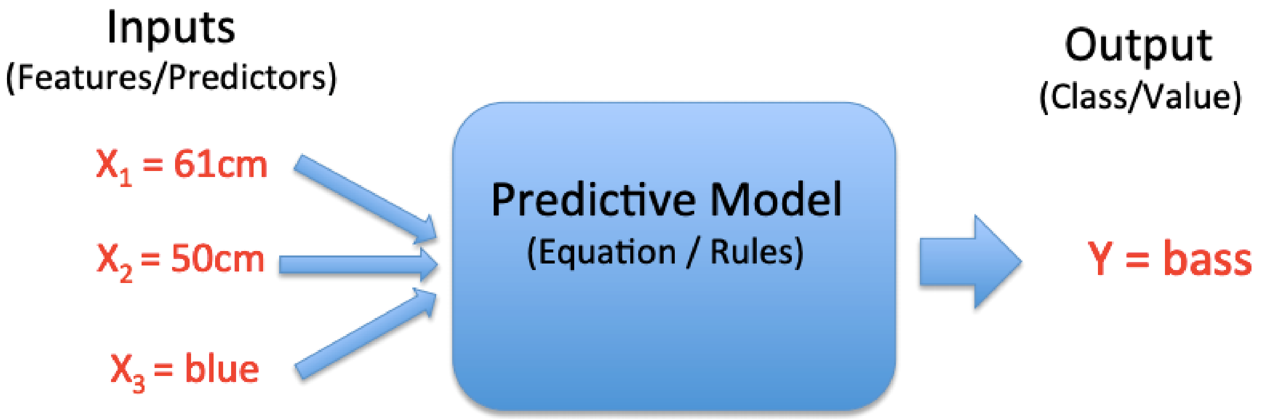
Predictive Models



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**

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

Predictive Models



Inputs
(Features/Predictors)

X_1 = 61cm

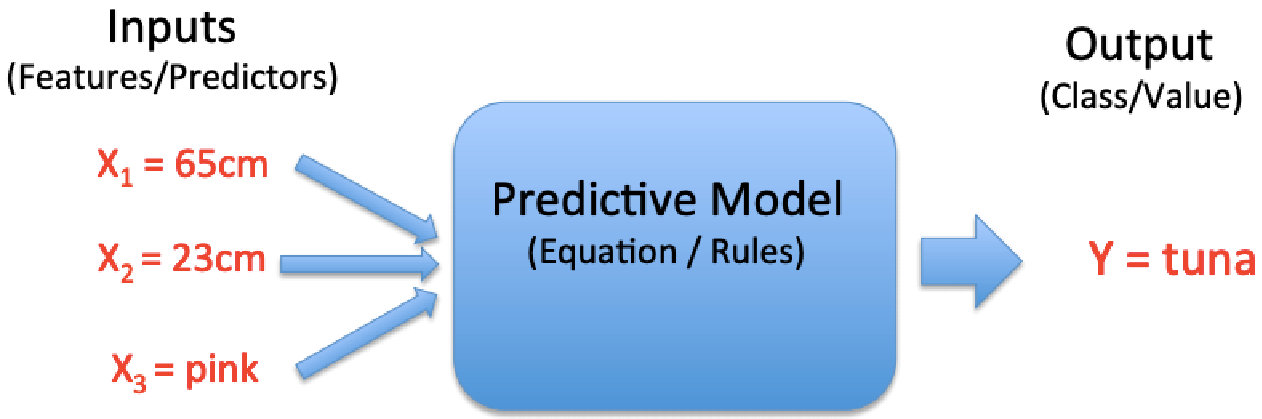
X_2 = 50cm

X_3 = blue

Predictive Model
(Equation / Rules)

Output
(Class/Value)

Y = bass



Inputs
(Features/Predictors)

X_1 = 65cm

X_2 = 23cm

X_3 = pink

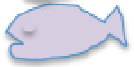
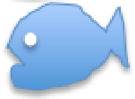
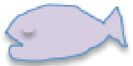
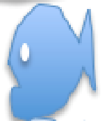
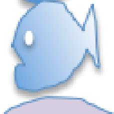
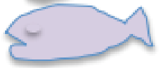
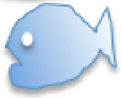
Predictive Model
(Equation / Rules)

Output
(Class/Value)

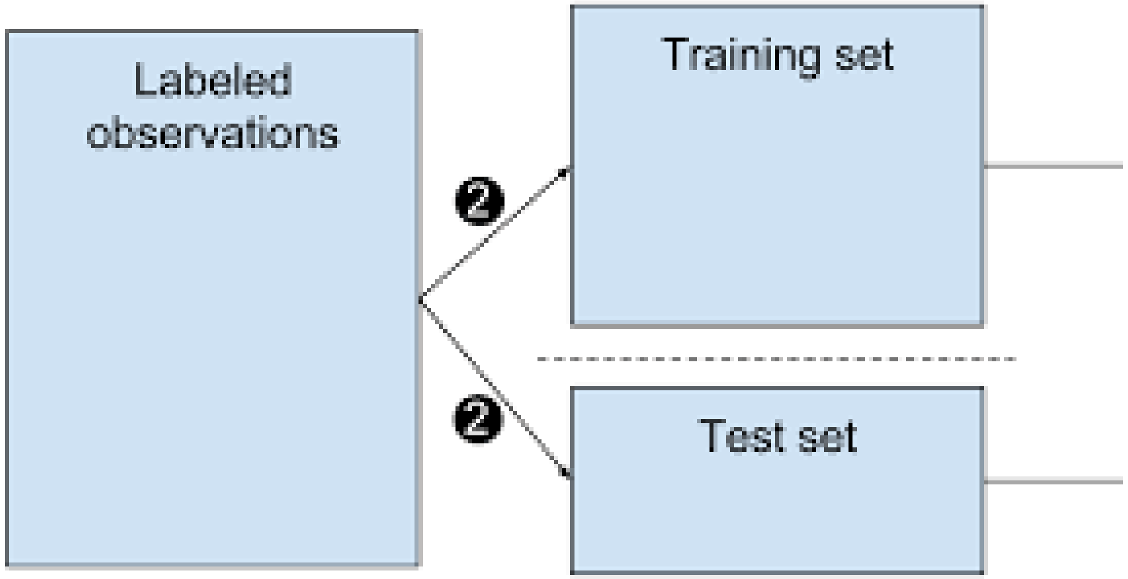
Y = tuna

Models are Built from Examples

Most models are developed through learning from examples

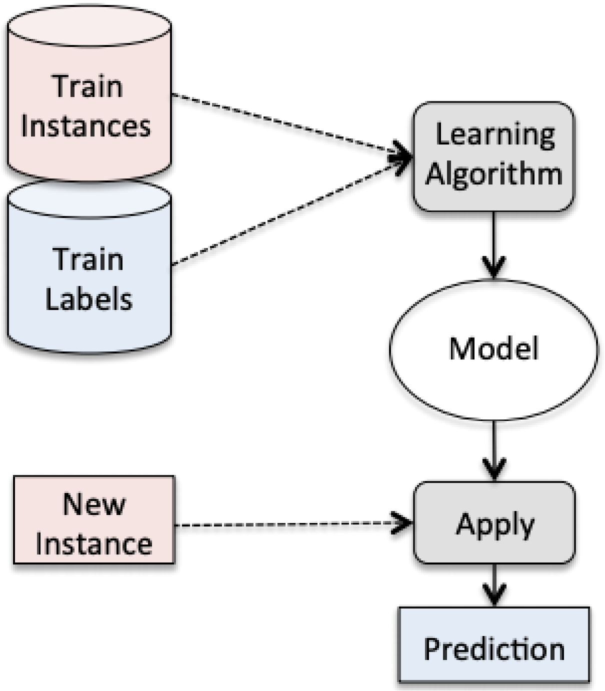
Instance	X1 = length	X2 = width	X3 = colour	Y = class
	55	51	blue	**bass**
	65	23	pink	**tuna**
	67	54	blue	**bass**
	54	20	light-blue	**tuna**
	62	26	pink	**tuna**
	44	62	blue	**bass**
	47	55	light-blue	**bass**
	73	31	pink	**tuna**
	54	48	light-blue	**bass**
	57	23	light-blue	**tuna**

Training and Testing Models



Training a Model

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



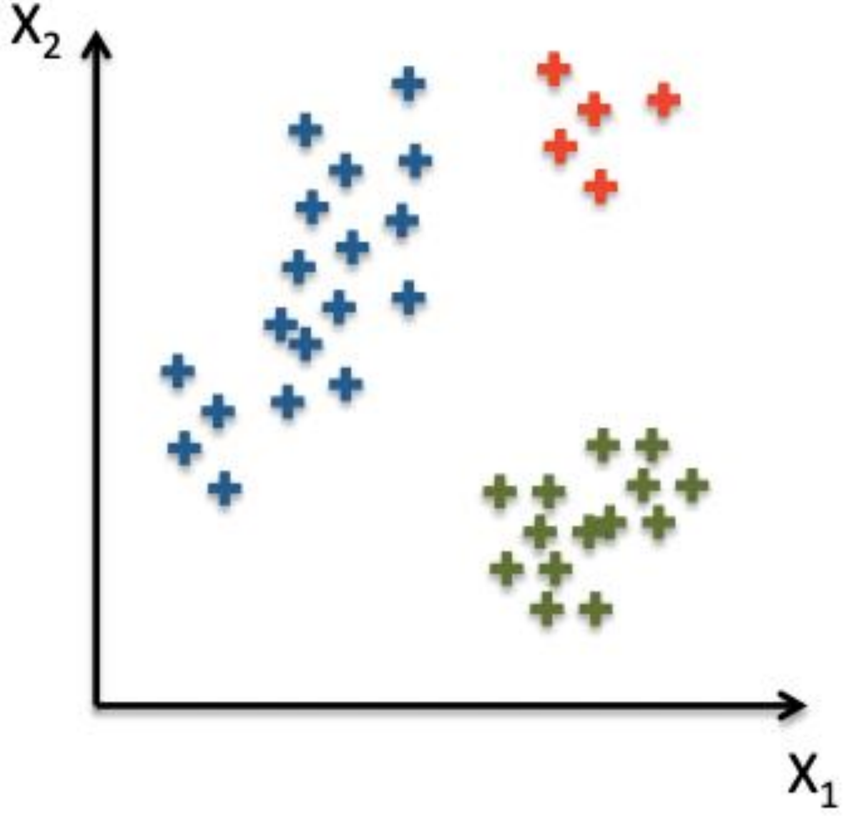
Training a Model

How are models derived (learnt)?

Each training instance (fish in our case) is just a point in some feature space

Here the colour denotes the class

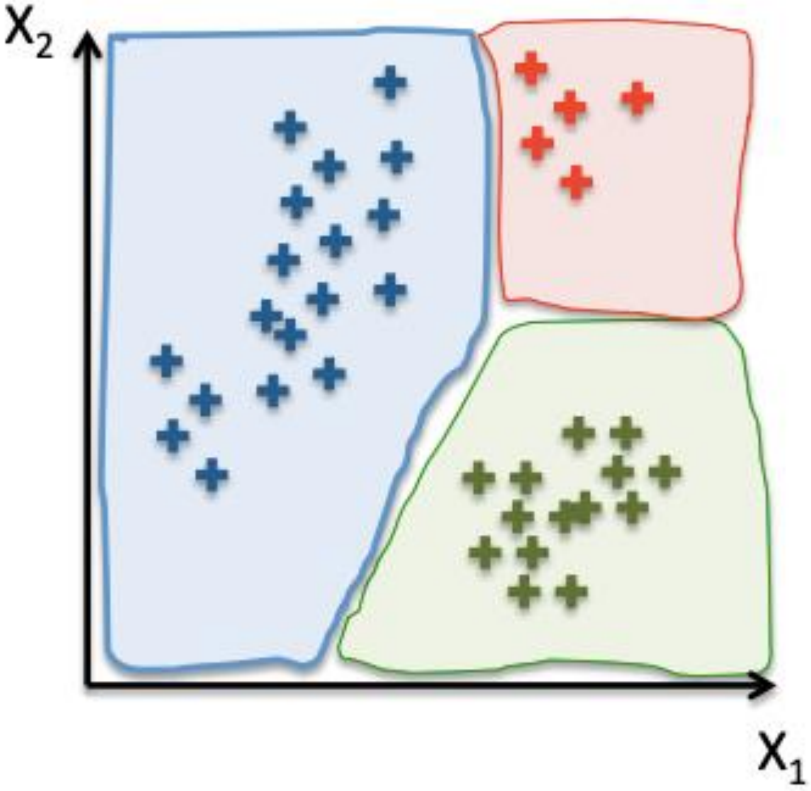
- blue = bass,
- green = tuna,
- red = unknown



Training a Model

How are models derived (learnt)?

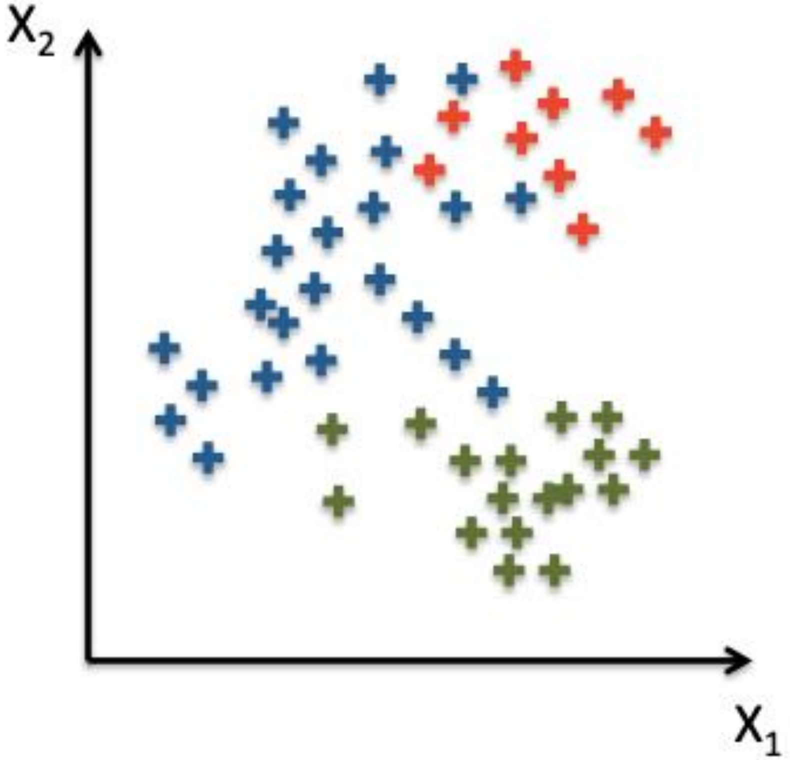
Many (classification) learning algorithms work by **dividing the feature space into regions of the same type**



In Practise

In practice, the data is usually **overlapping**

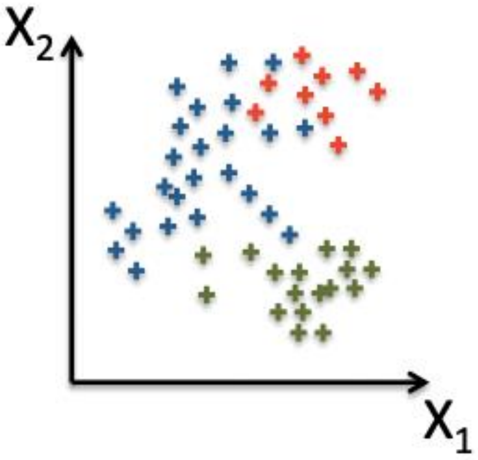
Making it hard to separate the classes

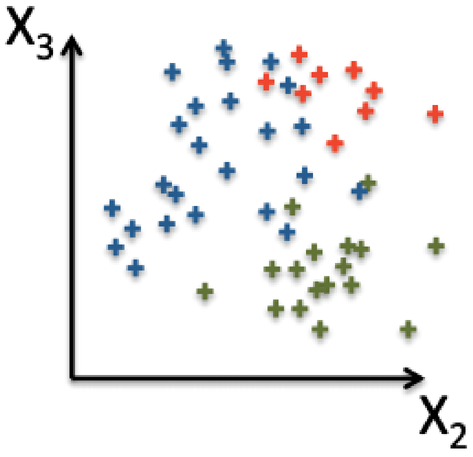


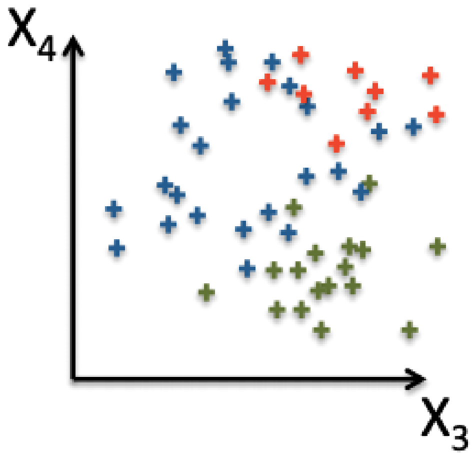
In Practise

And we have many feature dimensions

With **some features more useful than others**

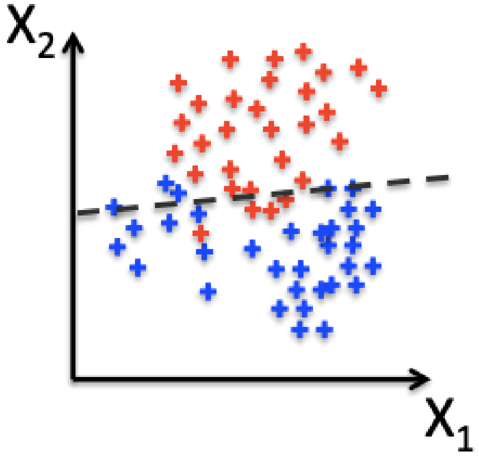






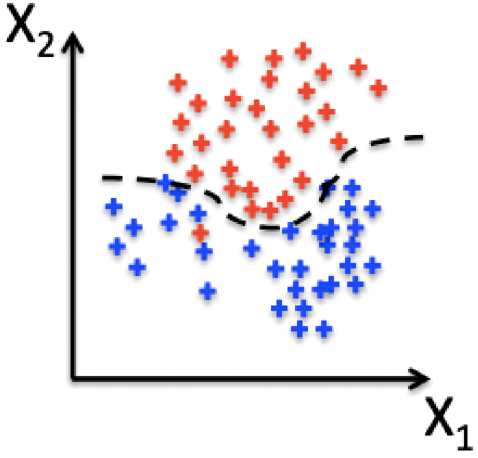
Different Models

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



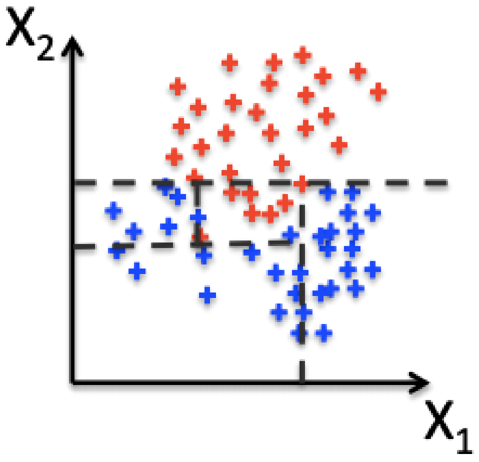
Linear classifiers

e.g. Logistic Regression,
Linear SVMs



Non-linear Classifiers

e.g. Neural Nets,
SVM with RBF kernel

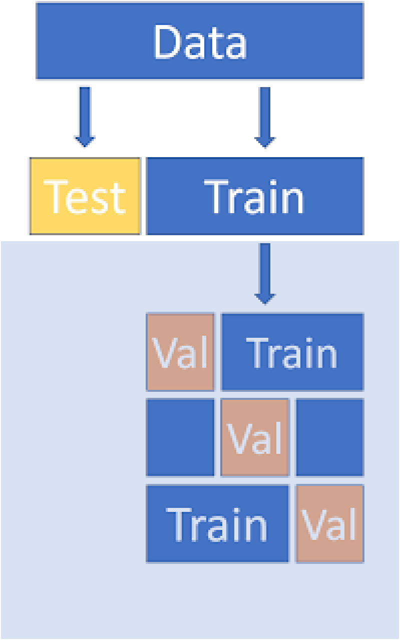


Decision Tree Learners

e.g. Random forests

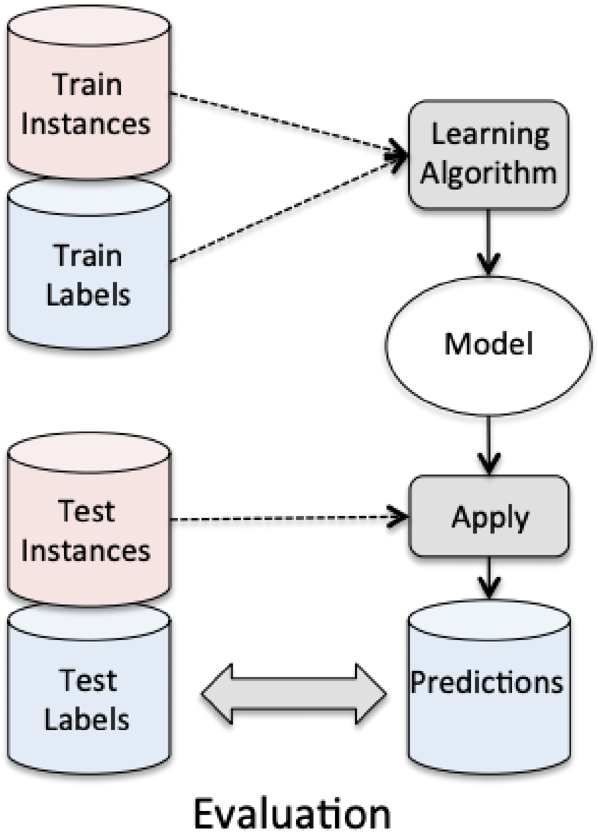
Different Models

How can we decide which model is better?



Testing Models

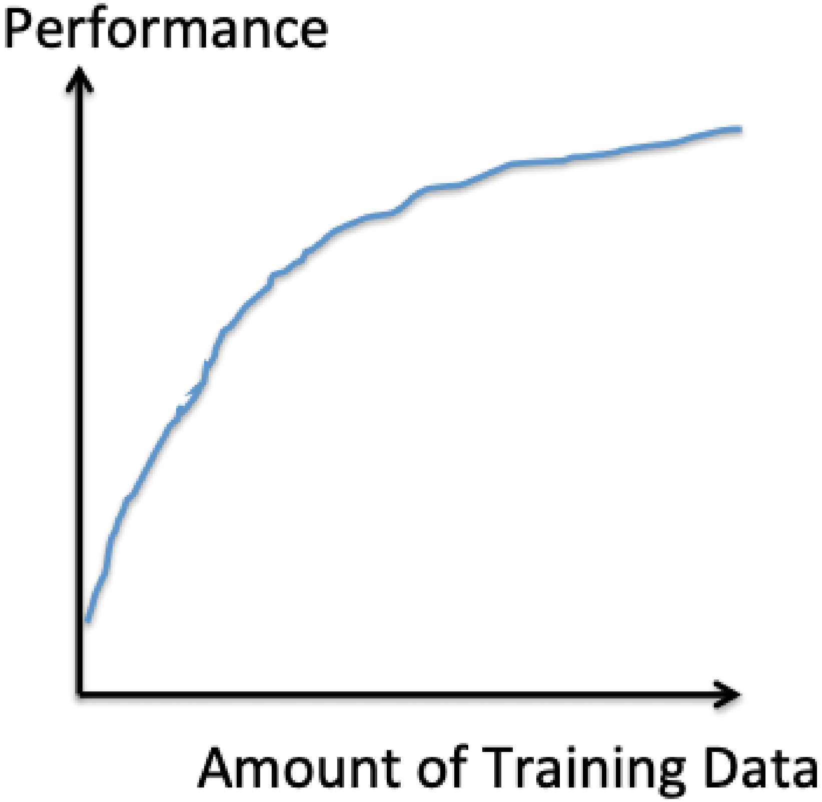
We evaluate predictive models based on how well they predict the labels for test instances (not used in training)



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 will be
    - Caveat: There is a limit to how many features



Home Activities

Suggested Activities for the week

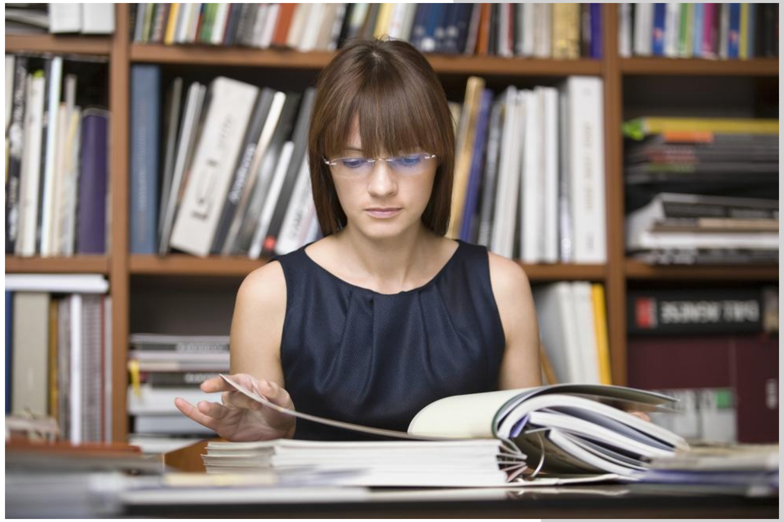
Videos

Watch [Jeremy Howard, “The wonderful and terrifying implications of computers that can learn”, TEDxBrussels]

Watch [Fei-Fei Li, “How we are teaching computers to understand pictures”, TED2015]

Articles

Read [Tarang Shah, “About Train, Validation and Test Sets in Machine Learning”, towards data science]



Recap: Learning Outcomes

Week

By the end of this week you should be able to:

- Explain what are models and predictive models
- Analyse predictive models in different examples
- Understand how to evaluate predictive models
- **Analyse how to estimate linear regression model**
- Apply linear regression and polynomial regression on different data sets using Python

MONASH University