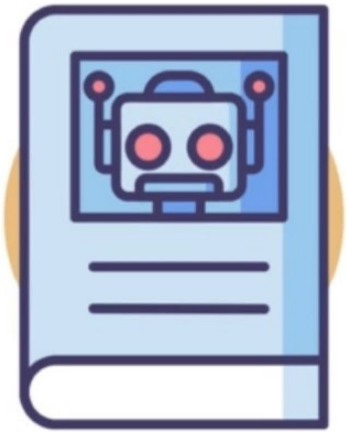


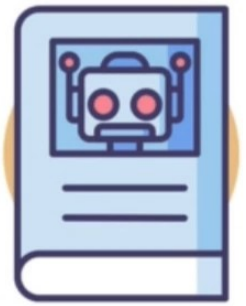
1. Create a framework

2. Match to data science and machine learning tools

3. Learn by doing

Yes	No
Write code	Overthink the process
Make mistakes	Try make things perfect
Build projects	Build things from scratch
Learn what matters	

The framework we'll be using



Steps in a full machine learning project

1. Problem definition

"What problem are we trying to solve?"

Supervised

Unsupervised

Classification

Regression

...and we we learn what these are!

2. Data

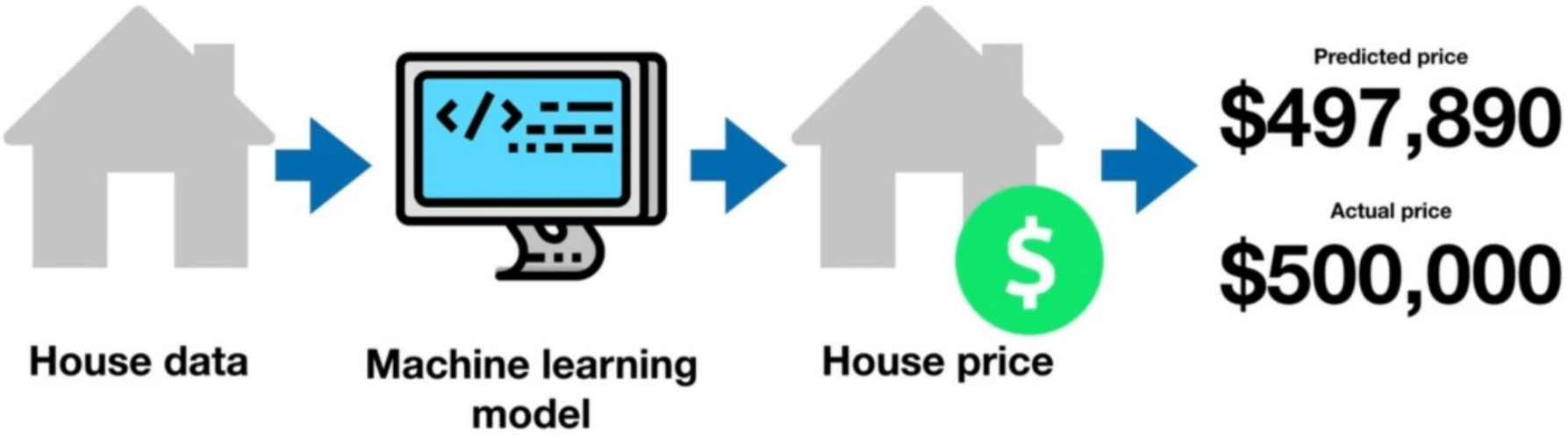
"What kind of data do we have?"

Structured

Unstructured

3. Evaluation

"What defines success for us?"



4. Features

"What do we already know about the data?"

ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4326	110kg	M	120/80	4	yes
5681	64kg	F	130/90	1	NO
7911	81kg	M	130/80	0	NO

Table 1.0: Patient records

5. Modelling

"Based on our problem and data, what model should we use?"

Problem 1

Model 1

Problem 2

Model 2

6. Experimentation

"How could we improve/what can we try next?"

Attempt

1

2

3

Steps in a full machine learning project

1. Problem definition

"What problem are we trying to solve?"

Main types of machine learning

Supervised Learning

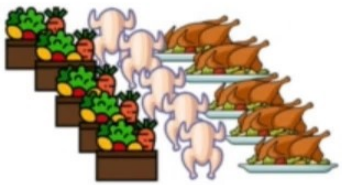
Unsupervised Learning

Transfer Learning

Reinforcement Learning

Supervised learning

Supervised Learning



Examples

Data Label

1. Cut vegetables
2. Season chicken with lots of spice
3. Preheat oven
4. Cook chicken for 30-minutes
5. Add vegetables

Instructions

Data Label

1. Cut vegetables
2. Season chicken
3. Preheat oven
4. Cook chicken for 30-minutes
5. Add vegetables

Instructions

Supervised learning

Classification

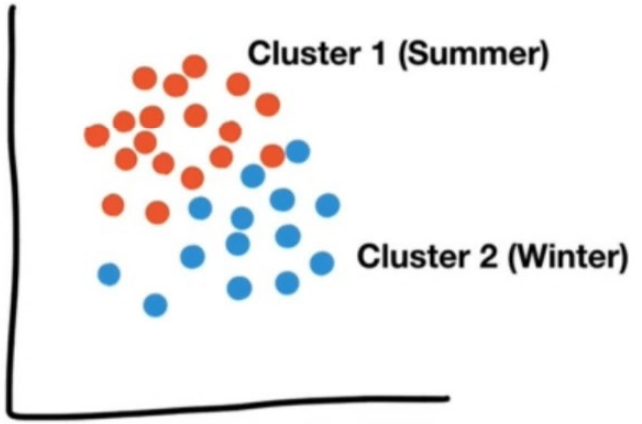
- "Is this example one thing or another?"
- Binary classification = two options
- Multi-class classification = more than two options

Regression

- "How much will this house sell for?"
- "How many people will buy this app?"

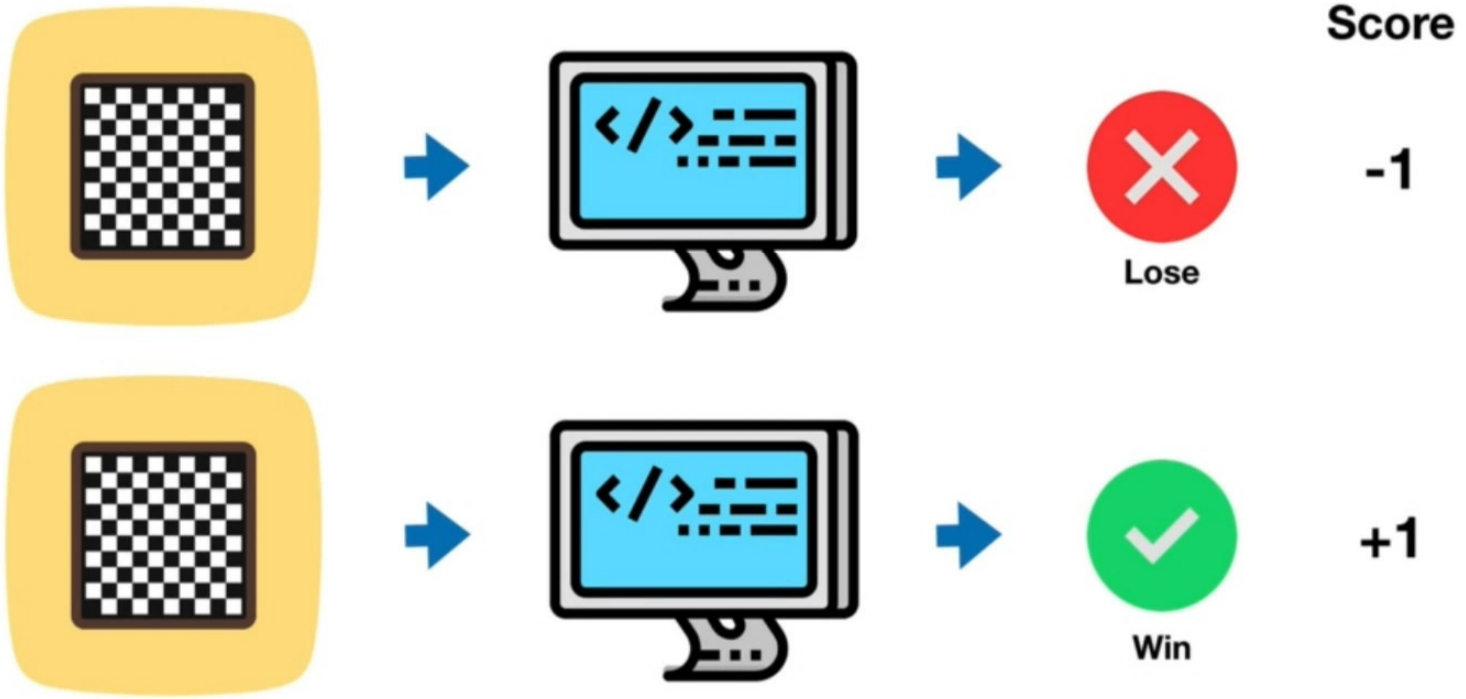
Unsupervised learning

Customer ID	Purchase 1	Purchase2
1	Sunglasses	Singlet
2	Jacket	Snow boots
3	Sunscreen	Beach towel



Transfer learning

Reinforcement learning



Matching your problem

Supervised Learning

"I know my inputs and outputs."

Unsupervised Learning

"I'm not sure of the outputs but I have inputs."

Matching your problem

Transfer Learning

"I think my problem may be similar to something else."

Steps in a full machine learning project

2.Data 

“What kind of data do we have?”

Different types of data

Rows

Columns

ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4326	110kg	M	120/80	4	yes
5681	64kg	F	130/90	1	NO
7911	81kg	M	130/80	0	NO

Table 1.0: Patient records

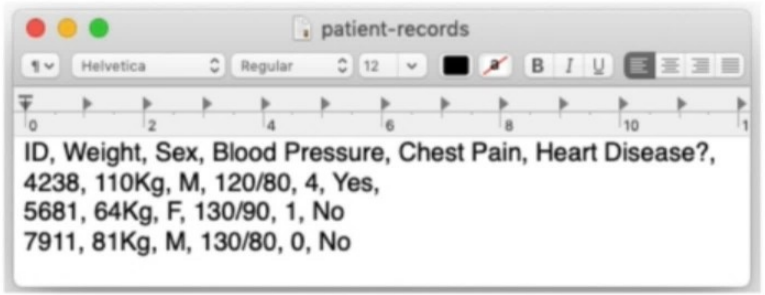
Structured

From: daniel@mrdbourke.com
Hey Daniel,

First of all, thank you for being so amazing.
This machine learning course is incredible.
Thank you for keeping it simple!

Unstructured

Different types of data



ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4326	110kg	M	120/80	4	YES
5681	64kg	F	130/90	1	NO
7911	81kg	M	130/80	0	NO

Table 1.0: Patient records

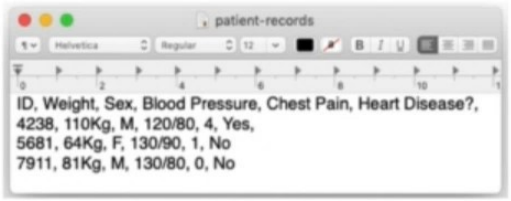
Static

“The more data the better.”

Different types of data

Streaming

A data science workflow



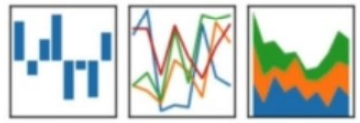
Static data

ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4528	110Kg	M	120/80	4	Yes
5681	64Kg	F	130/90	1	No
7911	81Kg	M	130/80	0	No

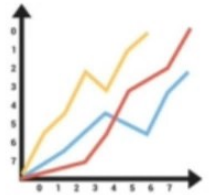
Table 1.0: Patient records

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Data Analysis



Machine learning model

Steps in a full machine learning project

3. Evaluation 

“What defines success for us?”

**“For this project to be worth pursuing further,
we need a machine learning model with over 99% accuracy.”**

ID	Weight	Sex	Blood Pressure	Chest pain	Heart disease?
4326	110kg	M	120/80	4	YES
5681	64kg	F	130/90	1	NO
7911	81kg	M	130/80	0	NO

Table 1.0: Patient records

Machine learning
model

Accuracy
97.8%

Different types of metrics

Classification	Regression	Recommendation
Accuracy	Mean absolute error (MAE)	Precision at K
Precision	Mean squared error (MSE)	
Recall	Root mean squared error (RMSE)	

Classifying car insurance claims

		Data	Label
ID	Img	Text	Result
1		Hi, I crashed into the neighbours letterbox and dented my car.	At fault
2		Someone ran into the back of me while I was at the traffic lights.	Not at fault

Table 2.0: Car insurance claims

(had to try a few of these)

Machine learning model

Minimum accuracy

>95%

What do you measure?

Steps in a full machine learning project

4. Features 

“What do we already know about the data?”

Different features of data

Feature variables | Target variable

ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110kg	M	81	4	YES
5681	64kg	F	61	1	NO
7911	81kg	M	57	0	NO

Table 1.0: Patient records

Different features of data

ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110Kg	M	81	4	YES
5681	64Kg	F	61	1	NO
7911	81Kg	M	57	0	NO

Table 1.0: Patient records

Numerical features

Categorical features

Different features of data

Derived feature

ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?	visit in last year?
4328	110Kg	M	81	4	Yes	Yes
5681	64Kg	F	61	1	NO	Yes
7911	81Kg	M	57	0	NO	NO

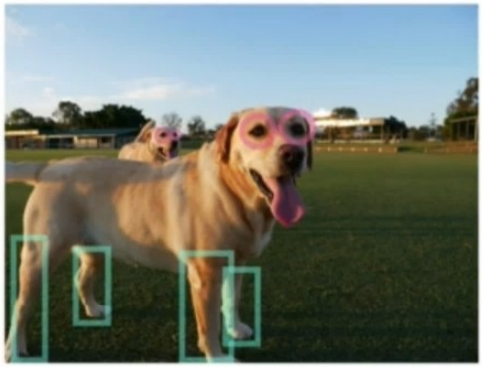
Table 1.0: Patient records

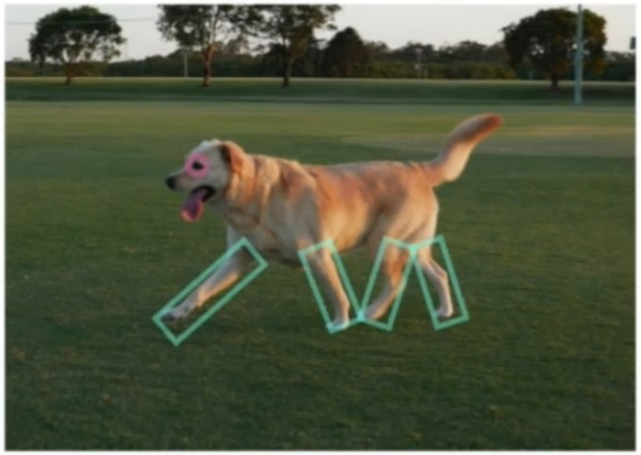
Numerical features

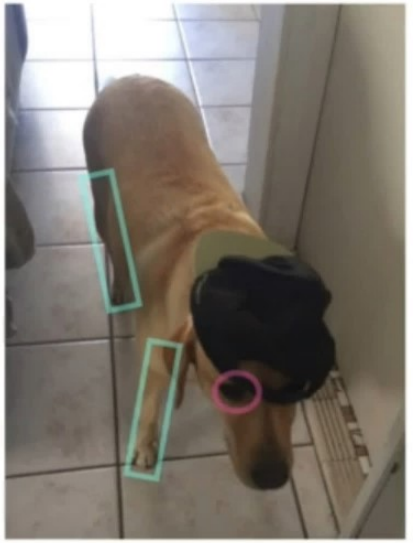
Categorical features

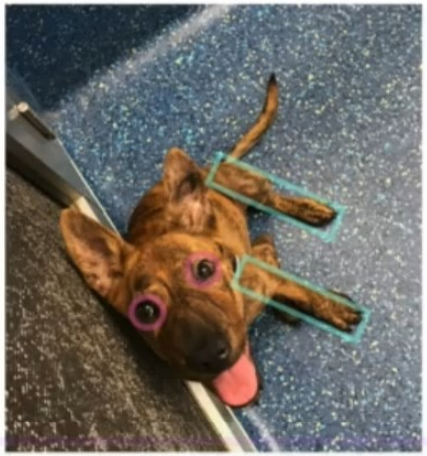
Feature engineering Looking at different features of data and creating new ones/altering existing ones

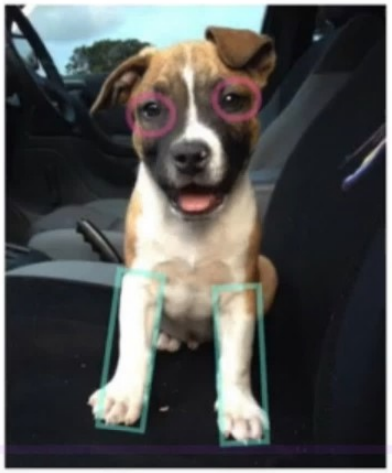
Different features of data



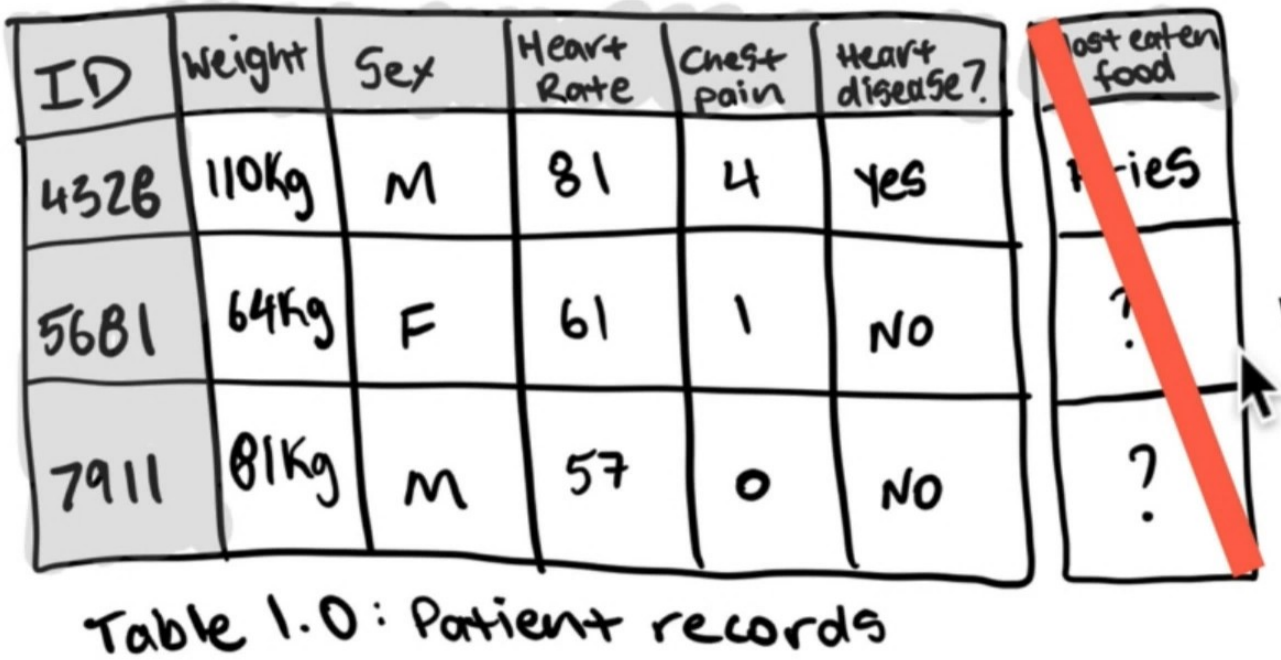








What features should you use?



ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?	Last eaten food
4326	110Kg	M	81	4	YES	Fries
5681	64Kg	F	61	1	NO	?
7911	81Kg	M	57	0	NO	?

Table 1.0: Patient records

Want > 10% coverage

Feature coverage

How many samples have different features? Ideally, every sample has the same features.

What are features of your problems?

Steps in a full machine learning project

5. Modelling Part 1 — 3 sets

"Based on our problem and data, what model should we use?"

3 parts to modelling

1. Choosing and training a model

or

2. Tuning a model

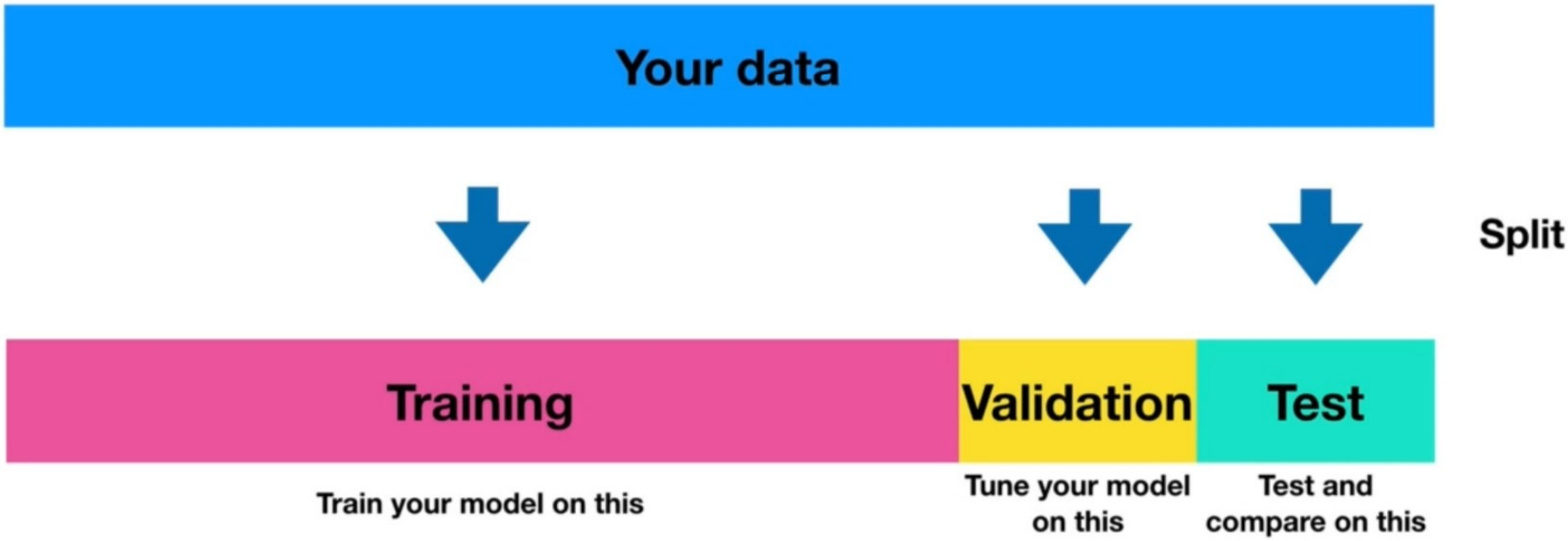
3. Model comparison

vs.

vs.

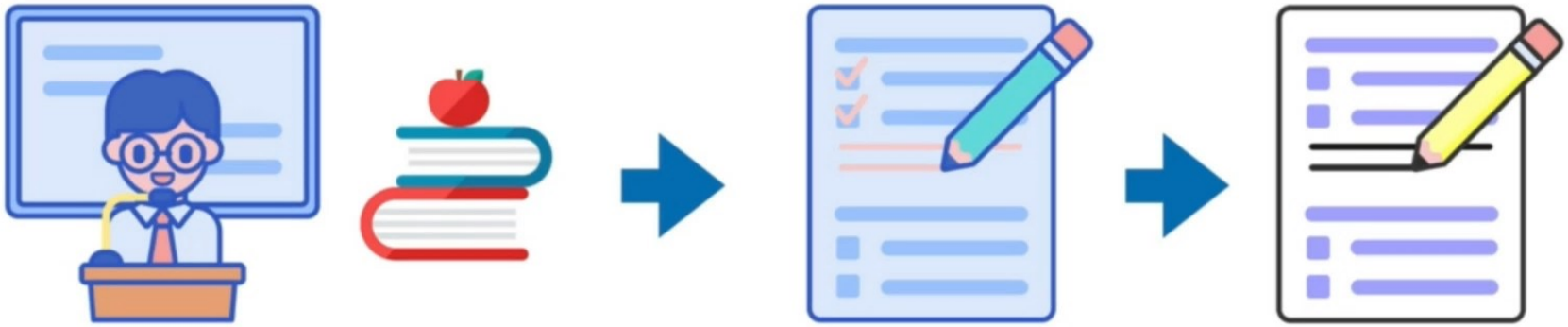
The most important concept in machine learning

(the training, validation and test sets or 3 sets)



The most important concept in machine learning

(the 3 sets)



Course materials
(training set)

Practice exam
(validation set)

Final exam
(test set)

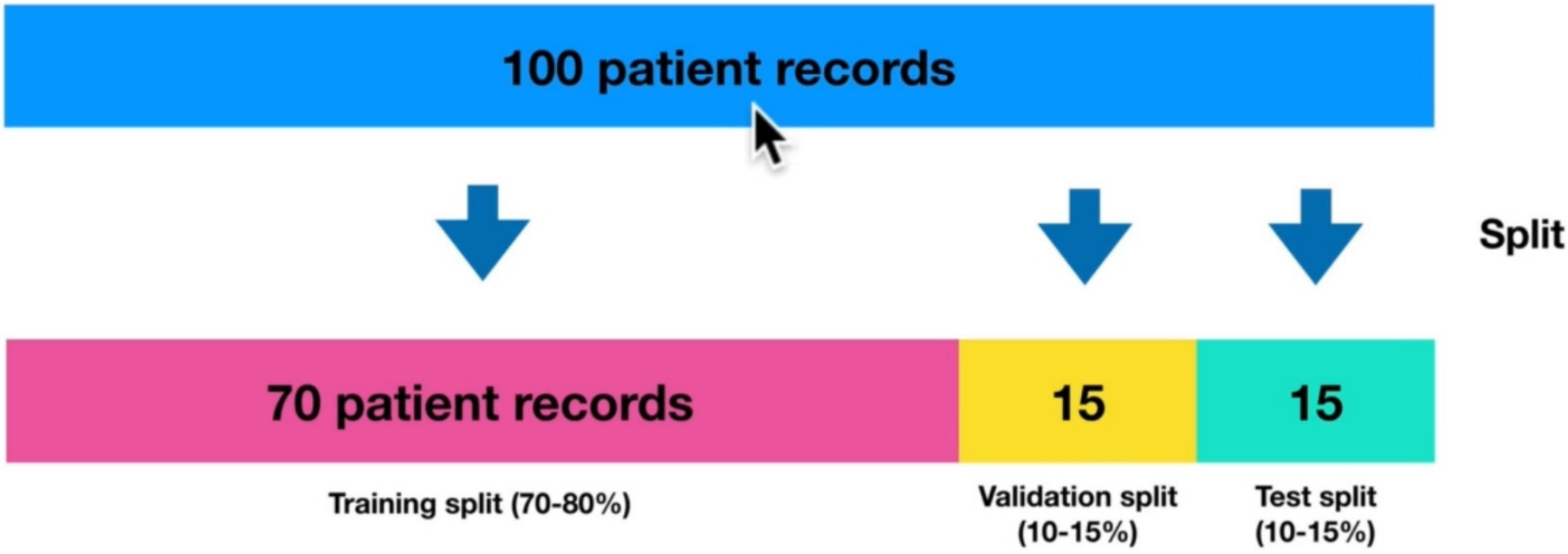
When things go wrong

Practice exam
(same as final exam set)

Final exam
(already seen it)

The most important concept in machine learning

(the 3 sets)

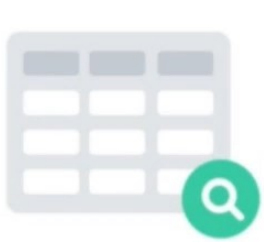


What was the last thing you testing your ability on?

5. Modelling Part 2 — Choosing 

"Based on our problem and data, what model should we use?"

Choosing a model

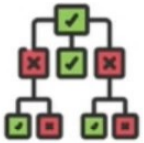


Problem 1

Model 1

Structured Data

CatBoost



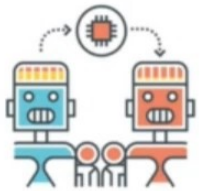
Random Forest

Problem 2

Model 2

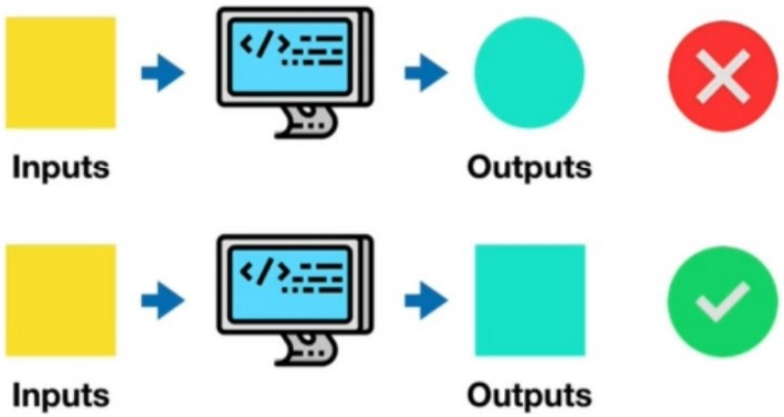
Unstructured Data

Deep Learning



Transfer Learning

Training a model



X (data) y (label)

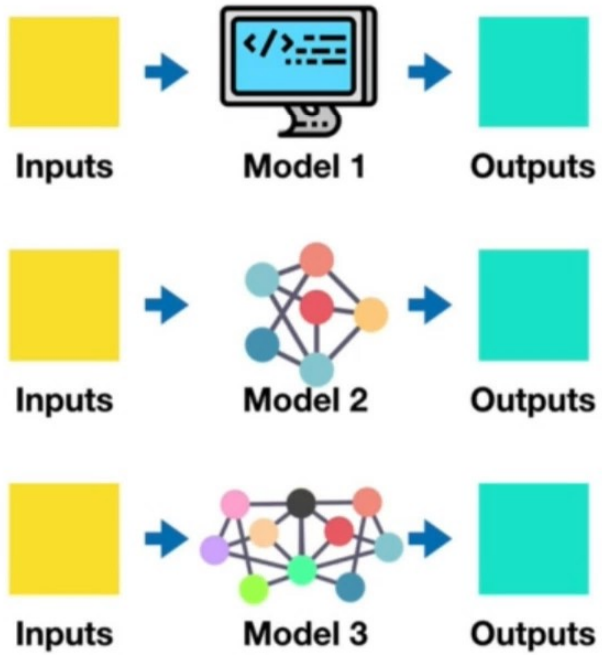
ID	Weight	Sex	Heart Rate	Chest pain	Heart disease?
4326	110kg	M	81	4	YES
5681	64kg	F	61	1	NO
7911	81kg	M	57	0	NO

Table 1.0: Patient records

Training Data

Goal: Minimise time between experiments

Experiment		Accuracy	Training time
1	Inputs → Model 1 → Outputs	87.5%	3 min
2	Inputs → Model 2 → Outputs	91.3%	92 min
3	Inputs → Model 3 → Outputs	94.7%	176 min



Things to remember

- Some models work better than others on different problems
- Don’t be afraid to try things
- Start small and build up (add complexity) as you need

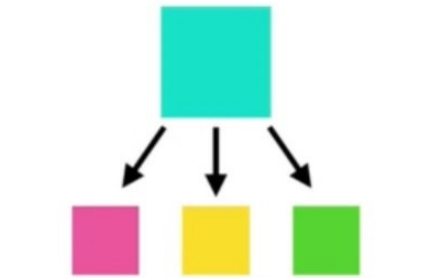
Tuning a model

Cooking time: 1 hour
Temperature: 180°C

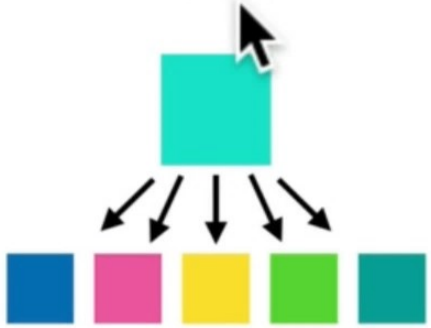
Cooking time: 1 hour
Temperature: 200°C

Tuning a model

Random Forest



3 trees



5 trees

Neural Networks

2 layers

3 layers

Things to remember

- Machine learning models have hyperparameters you can adjust
- A models first results aren't its last
- Tuning can take place on training or validation data sets

Testing a model

Data Set	Performance
Training	98%
Test	96%

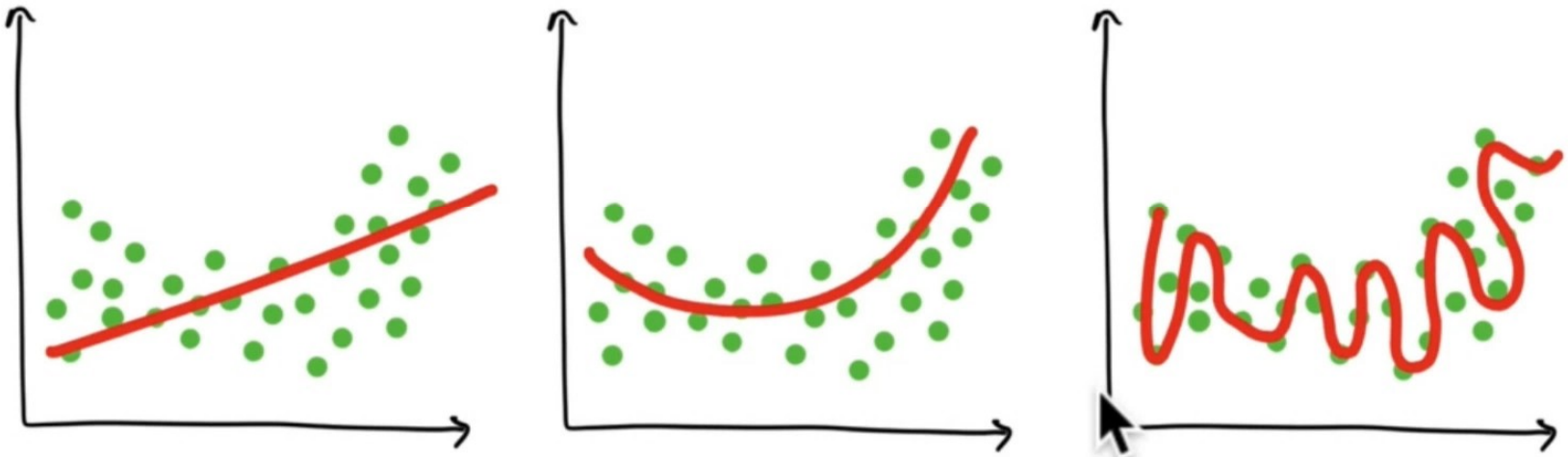
Underfitting (potential)

Data Set	Performance
Training	64%
Test	47%

Overfitting (potential)

Data Set	Performance
Training	93%
Test	99%

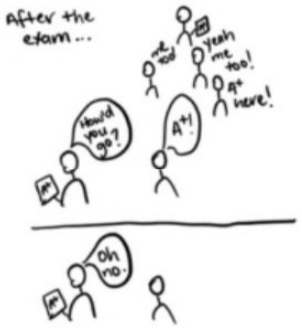
Overfitting and underfitting

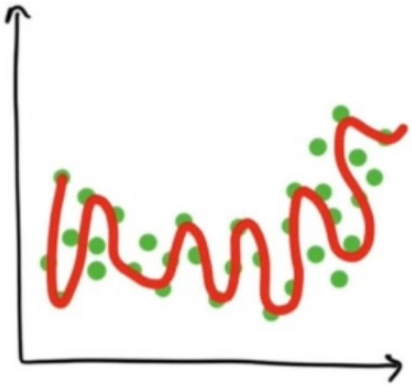


Overfitting and underfitting

Practice exam

Final exam





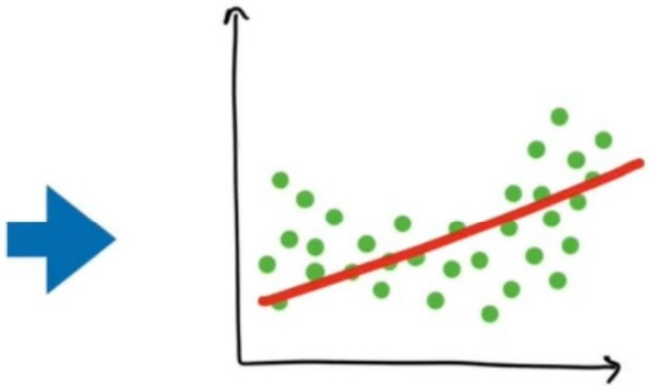
Overfitting

Data leakage

Overfitting and underfitting

Training Data

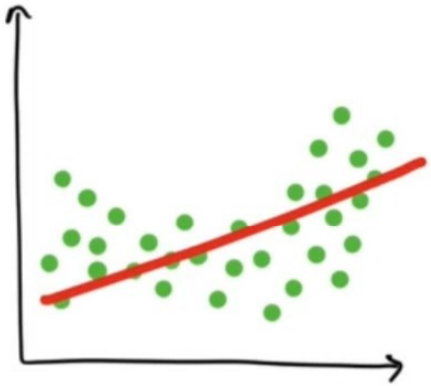
Test Data



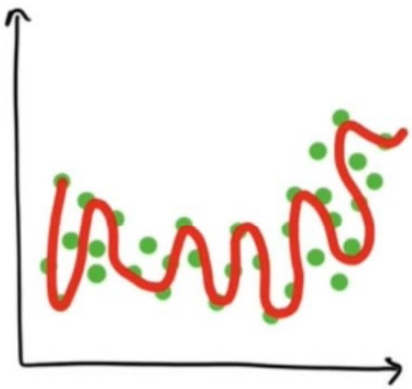
Data mismatch

Underfitting

Fixes for overfitting and underfitting



Underfitting



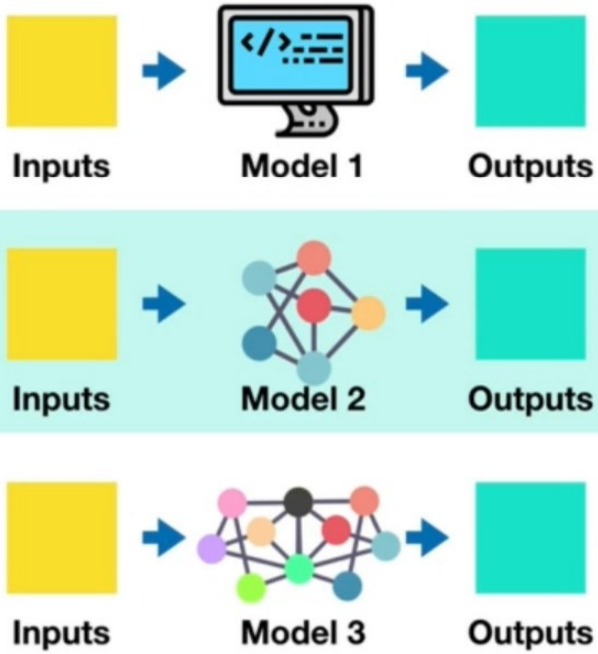
Overfitting

- Try a more advanced model
- Increase model hyperparameters
- Reduce amount of features
- Train longer

- Collect more data
- Try a less advanced model

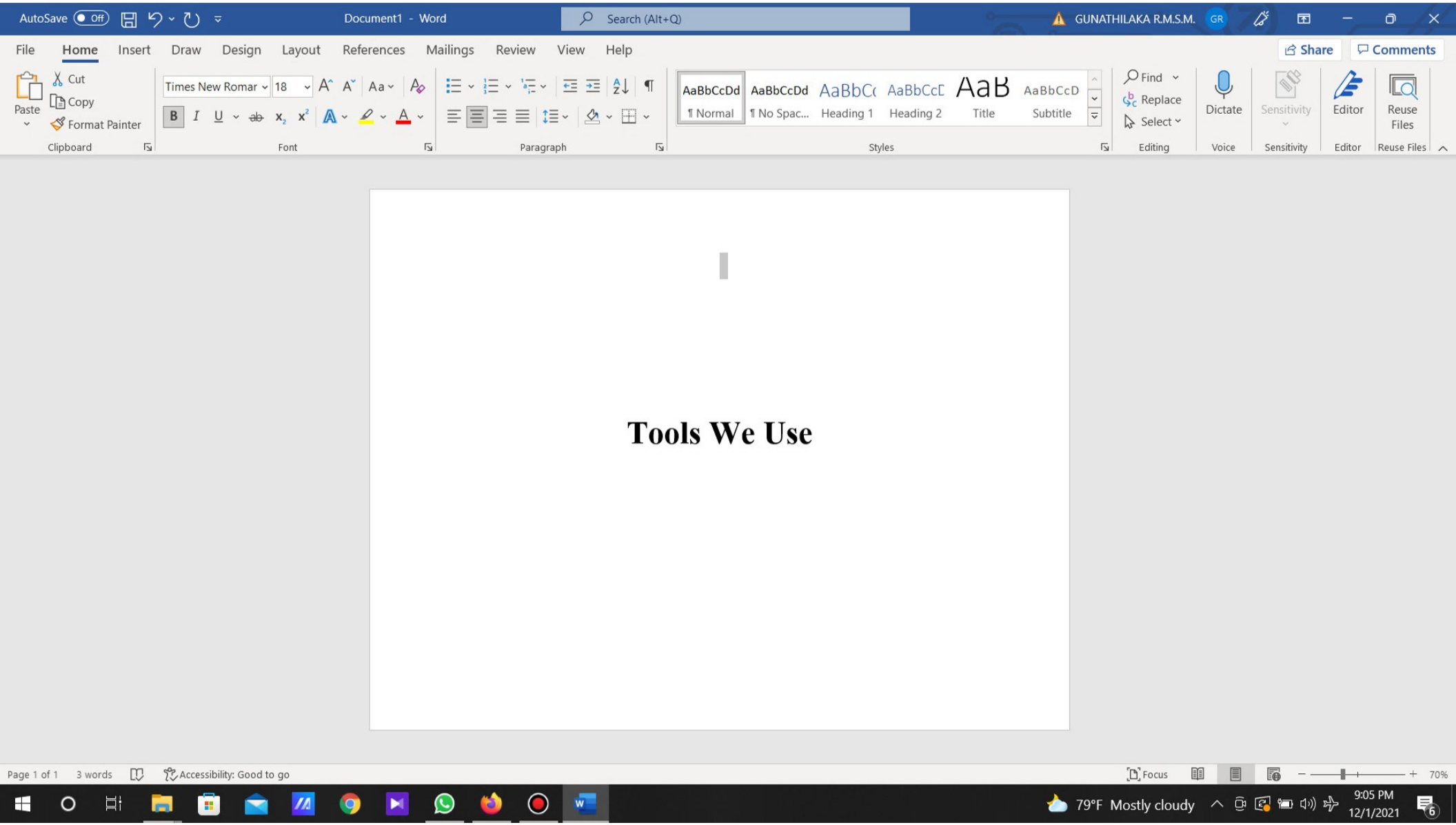
Comparing models

Experiment		Accuracy	Training time	Prediction time
1	Inputs → Model 1 → Outputs	87.5%	3 min	0.5 sec
2	Inputs → Model 2 → Outputs	91.3%	92 min	1 sec
3	Inputs → Model 3 → Outputs	94.7%	176 min	4 sec



Things to remember

- Want to avoid overfitting and underfitting (head towards generality)
- Keep the test set separate at all costs
- Compare apples to apples
- One best performance metric does not equal best model


AutoSave Off
Document1 - Word
Search (Alt+Q)
GUNATHILAKA R.M.S.M.
File Home Insert Draw Design Layout References Mailings Review View Help
Share Comments
Cut Copy Format Painter Paste
Times New Romar 18
Clipboard Font Paragraph Styles Editing Voice Sensitivity Editor Reuse Files
AaBbCcDd ¶ Normal
AaBbCcDd ¶ No Spac...
AaBbC(Heading 1
AaBbCcE Heading 2
AaB Title
AaBbCcD Subtitle
Find Replace Select
Dictate
Sensitivity
Editor
Reuse Files
Tools We Use
Page 1 of 1
3 words
Accessibility: Good to go
Focus
70%
79°F Mostly cloudy
9:05 PM
12/1/2021

Tools We Use


Your computer
ANACONDA
jupyter
matplotlib
NumPy
pandas
scikit learn
dmlc XGBoost

Tools you can use