[Flux questions]

- Week 1

- Who are the Data Scientists?

person A

person B

person C

person D

Answer: All of them

- Which of the following data science definition you like most?
  Data Science is
  A. machine learning on big data
  B. extraction of knowledge/value from data through the complete data lifecycle process
  C. almost everything that has something to do with data: collecting, analyzing, modeling, etc, yet the most important part is its applications — all sorts of applications
  Answer: Can include all of them

- Which of the following is real world applications of Machine Learning?
  A. Video Games
  B. Self-driving cars
  C. Spam filtering
  D. Predictions
  E. All of the options
  Answer: E

- Using a short phrase or word, which activity in data science process is the most interesting to you.
  Answer: Depends on students' interests

- Week 2

- What is .ipynb?
  A. An illegal file extension.
  B. Interactive Python NoteBook.
  C. Intelligent Python Nota Bene.
  D. Typo, it should be 'pinyin'
  Answer: B

- What is a dataframe?.
  A. An array.
  B. A list.
  C. A theory about data.
  D. A structure that stores tabular data
  Answer: D

- Consider the definition given for data science, is the boundary between data science, data engineering and data analysis fixed?
  A. TRUE
  B. FALSE
  Answer: B

- Referring to the two slides about the car industry:
  First, they underwent a digitization process, followed by a **datafication** process

- Using a word or short phrase, name a non-automotive industry that have had similar developments in recent decades. How do you expect the datafication process to change **retail industry**?
  More effective in **providing customised products**

- Week 3

- Which option is the Mean, Median and Mode of the following set of values respectively?
  1,2,2,3,4,7,9
  A. 4,2,3

B. 5,3,2
C. 4,3,3
D. 4,3,2
Answer: D

Type	Example	Result
Mean	(1+2+2+3+4+7+9) / 7	**4**
Median	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	1, **2**, **2**, 3, 4, 7, 9	**2**

- **Week 4**

- Consider data sources in ”traffic forecasting” task, name a website to access to one of the data
  Answer: For example, we can get weather data from the Melbourne Forecast - Bureau of Meteorology website.

- Name a popular data/information API.
  Answer: For example, Google Maps API

- How many problems can you identify?

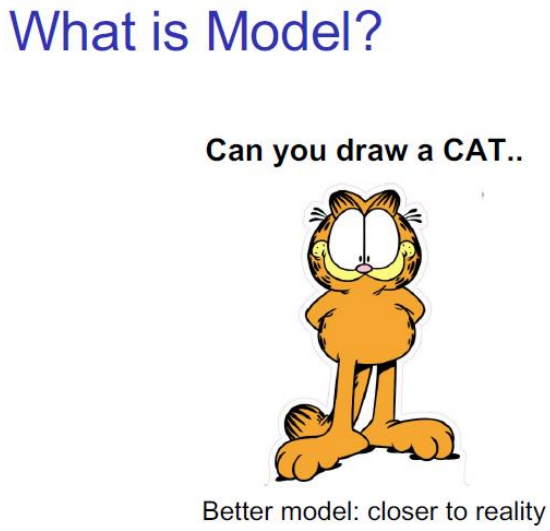
.....	**Suburbs**
.....	burwood.
.....	springvale
....	Burwood
....	Springvae
....	East Melbourne
.....	E. Melbourne
.....	.....

  Answer: For example, inconsistent naming rule and duplication (E. Melbourne and East Melbourne)
- How to deal with missing data?

A. Removing the row or column
B. Replace with a special “unknown” value
C. Replace with an average value
Answer: All can be a correct choice depending on the dataset

- Week 5

- Do you think you drew a perfect model?



A. Yes
B. No
C. Not sure
Answer: Depends on your thought (but the idea is that there is no perfect model, each model you draw, it can miss some details). We then linked this to predictive models.

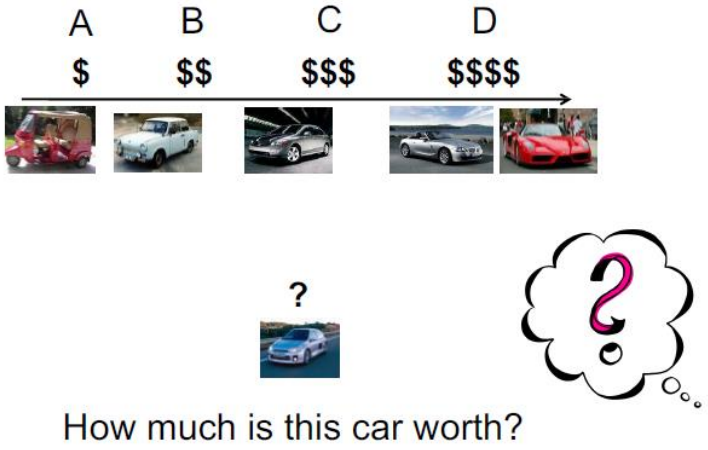
- Which group does this horse belong to?

Answer: Depends on your thought, focusing on different features of the horse, e.g., long leg, tail etc (the idea is to link this to features (variables) in predictive models)

- How can we decide which model is better?
  Answer: We evaluate predictive models based on how well they predict the labels of test instances

- Why is Machine Learning important?
  Answer: Can explain different reasons. For example, machine learning can detect pattern from data and predict the future based the detected patterns.

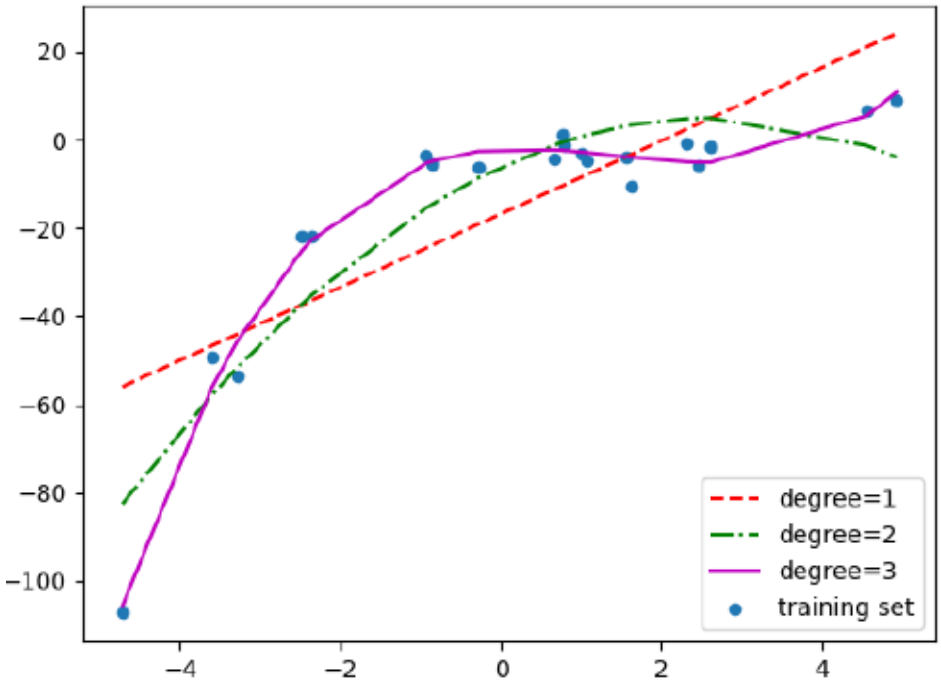
  How much is this car worth?



Answer: Depends on your thought, but seems to be more similar to B or C.

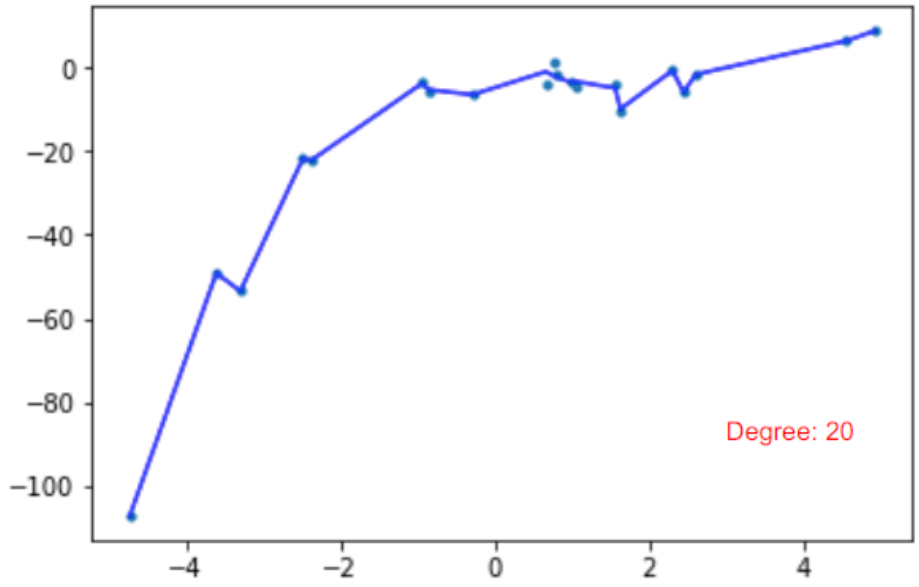
- Week 6

- What is the best degree? 1, 2 or 3?



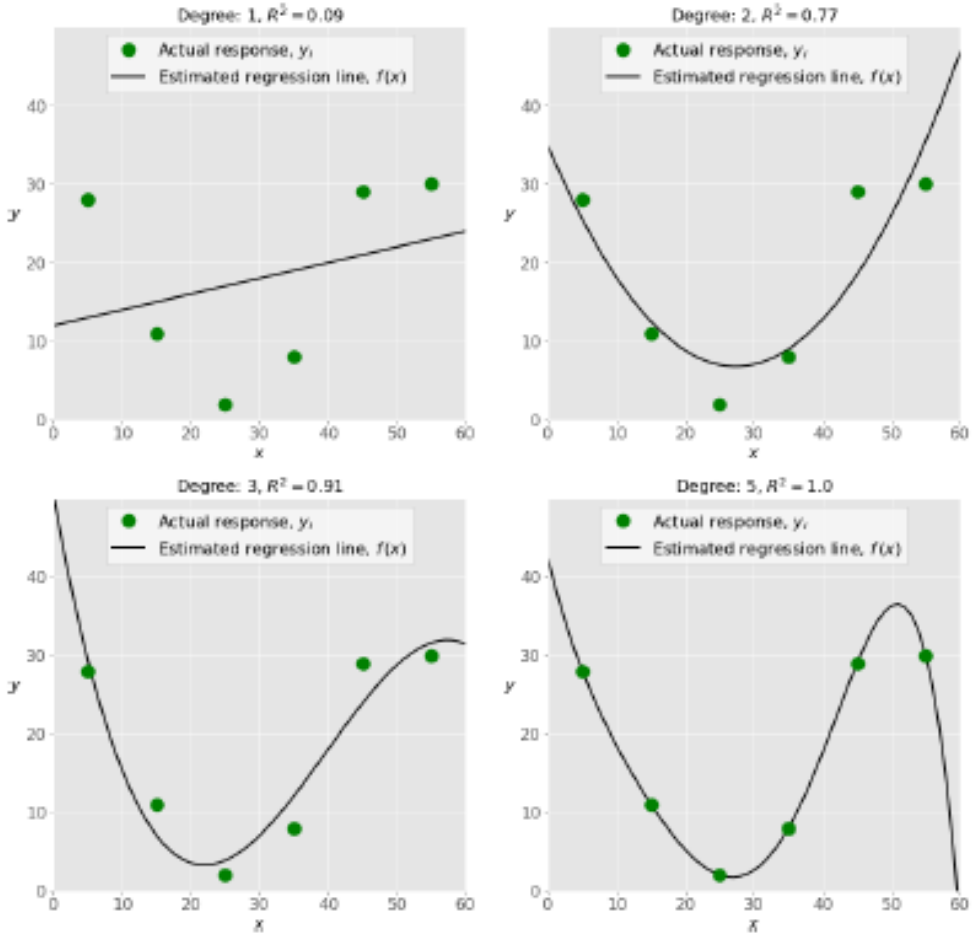
Answer: degree 3

- Is this fit better than previous fits?



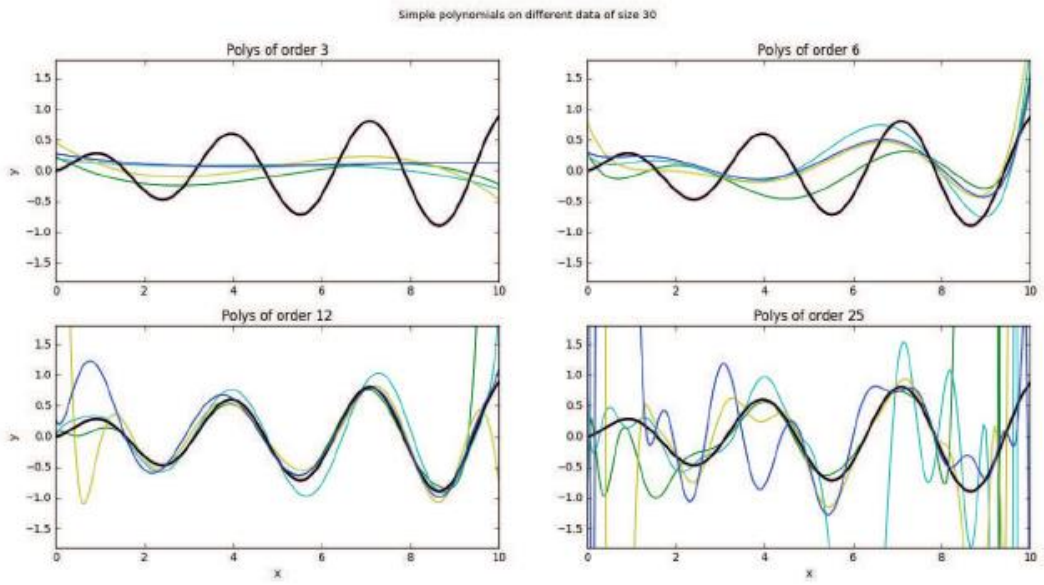
Answer: the error is lower copared to lower degrees, but it is overfitting to noise, degree 20 cannot make accurate prediction for unseen(testing) dataset

- Which model could be well-fitted?



Answer: Degree 3

- Which of the polynomials in the previous slide is a better model?
  A. Order 3
  B. Order 6
  C. Order 12
  D. Order 25



Answer: C

- Week 7

- Determine classification accuracy for the following Confusion Matrix?

n=192	**Predicted: 0**	**Predicted: 1**	
Actual: 0	TN = 118	FP = 12	130
Actual: 1	FN = 47	TP = 15	62
	165	27	

Answer: 0.69 ((118+15)/(118+12+47+15))

- According to the previous slide when is a bad day to play tennis?
  A. When it's sunny and humidity is high
  B. When it's rainy and wind is strong
  C. Both the above options
  Answer: C

- Week 8

- R Or Python?
  Answer: Depends on your thought (the idea is the ability to compare them)

- How to check if a vector(x) contains missing values?
  Answer: is.na(x)

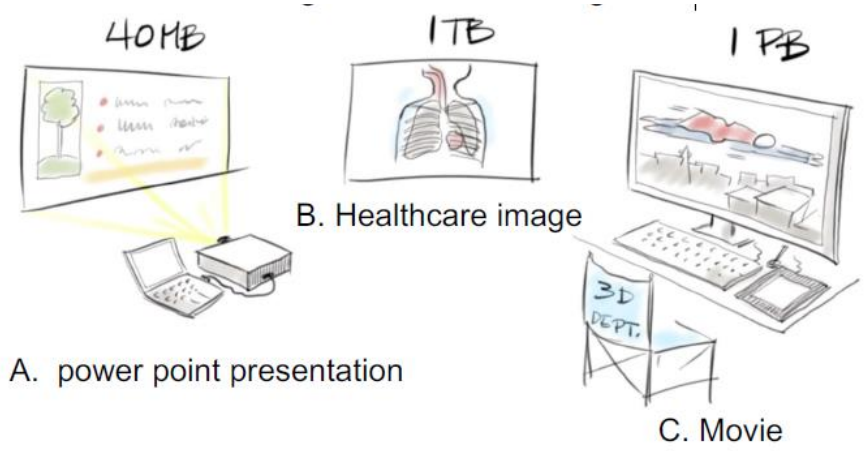
- How to find outliers in R?
  Answer: Use boxplot :

  m<- c(1,5,6,6,6,6,7,10)
  boxplot(m)$out

- Week 9

- How The 3Vs of big data are important because:
  A. They are an industry standard
  B. They are the basis for the development of more Vs (e.g. Value)
  C. They are used to describe in what way a dataset may be too big to handle
  D. They are from the influential Gartner Inc
  Answer: C

- Which of the following is considered as big data?



  Answer: Big data is any attribute that challenges constraints of a system capability or business need so A, B, C can be a big data, depending on how data challenges system capability or business need.

- Name a type of metadata might be associated with an image.
  Answer: For example, created date, owner of image data, size, resolution, etc.

- Unix shell commands like “less” and “grep”:
  A. can be used to manipulate large data files easily
  B. are poorly documented
  C. are examples of technology that is too old to be useful to a modern data scientis
  D. are used to fit regression tree models
  Answer: A

- Week 10

- Remember Bell’s law ... new classes of computing every decade.
  Can you suggest some new classes of computing?
  Answer: For example, mind-control devices, in-body devices, etc.

- Which one of the following tasks is very hard to make data parallel?
  A. Face recognition in 1M images
  B. Invert a large matrix
  C. Looking for common 3-4 word phrases in a collection of documents
  Answer: B

- Which one of the following is suitable for real-time data processing?
  A. Hadoop
  B. Spark
  Answer: B

- Week 11

Guest Lecture

- Week 12

No flux questions