NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

SC1015 : Review Lecture

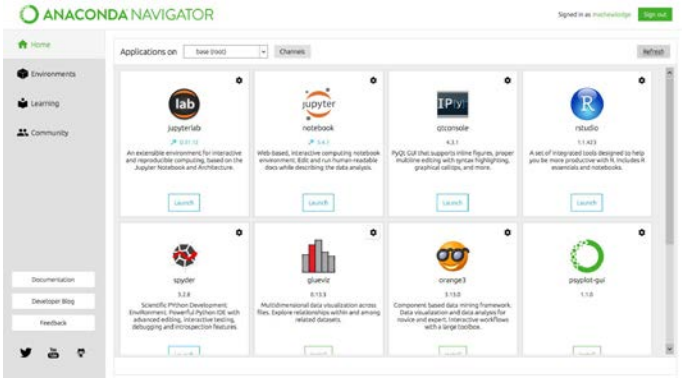
DS Pipeline and Problems

Dr Smitha K G

SC1015

Admin Announcements

Computing Platform



We will use the Anaconda platform.

Python 3.9 within Jupyter Notebook.

1. If you have any issue with the programming aspects of the Exercises, talk to your Lab TA.
2. You DO NOT need to make-up for missing labs in Week 2 or Week 3, as they are not graded. Do complete them as they are the basics for your graded exercises in week 4.
3. Talk to your TA about labs missed for graded Exercises, starting Week 4. If you have a valid explanation for your absence, you are allowed to submit.

Detailed FAQs available on Course Website.

Yes, Exercise 1 is too easy. I know! Just wait.

Important information

Lab on 9th Feb Friday (week4) is shifted to 16th Feb Friday (week 5)

SC1015 : Course Calendar for AY 2023-2024 S2

January			February			March			April		
1	Mo	New Year's Day	1	Th	Ex 2 (Lab)	1	Fr	Ex 5 (Lab)	1	Mo	Project (Lab)
2	Tu		2	Fr	Ex 2 (Lab)	2	Sa		2	Tu	Project (Lab)
3	We		3	Sa		3	Su	LAMS deadline (DS)	3	We	Project (Lab) + Review
4	Th		4	Su	W4 : Graded Lab (5%)	4	Mo	Recess Week	4	Th	Project (Lab)
5	Fr		5	Mo	Ex 3 (Lab)	5	Tu	Recess Week	5	Fr	Ex 7 (Lab)
6	Sa		6	Tu	Ex 3 (Lab)	6	We	Recess Week	6	Sa	
7	Su	W1 : LAMS Released (DS)	7	We	Ex 3 (Lab) + Review Lec	7	Th	Recess Week	7	Su	W12 : Graded Lab (5%)
8	Mo		8	Th	Ex 3 (Lab)	8	Fr	DS Theory Quiz (15%)	8	Mo	Ex 8 (Lab)
9	Tu		9	Fr	No Lab	9	Sa		9	Tu	Ex 8 (Lab)
10	We		10	Sa		10	Su	W8 : LAMS Released (AI)	10	We	No Lab :Hari Raya
11	Th		11	Su	W5 :Free week	11	Mo	Ex 6 (Lab)	11	Th	Ex 8 (Lab)
12	Fr	Course-Site Online	12	Mo	No Lab; CNY	12	Tu	Ex 6 (Lab)	12	Fr	Ex 8 (Lab)
13	Sa		13	Tu	No Lab	13	We	Ex 6 (Lab) + Review	13	Sa	
14	Su	W1 : LAMS Released (DS)	14	We	No Lab + Review Lec	14	Th	Ex 6 (Lab)	14	Su	W13 :Free week
15	Mo	No Lab	15	Th	No Lab	15	Fr	Ex 6 (Lab)	15	Mo	No Lab
16	Tu	No Lab	16	Fr	Ex 3 (Lab)	16	Sa		16	Tu	No Lab
17	We	No Lab + Overview Lec	17	Sa	Project Released	17	Su	W9: Consultation	17	We	Ex 8 (Lab)

Data Science

Data Pipeline

Seven stages, each with two Perspectives.

- You can't draw a Practical Motivation unless you plan the Data Collection exercise too.
- You can't formulate a Data Science problem without considering Data Preparation aspect.
- You can't obtain concise Statistical Description unless you have done Exploratory Analysis.
- You can't Recognize Patterns/Structure in data without proper Analytic Visualization.

Data Science

Data Pipeline

Seven stages, each with two Perspectives.

- You can't blindly use Machine Learning unless you understand the Algorithmic Optimizations.
- You can't draw Inference from the data without being able to Communicate it well enough.
- You can't of course take an Intelligent Decision unless you also consider its effects Ethically.

We will cover most of the stages in this course, and you will have to use (almost all) in the Mini-Project.

Data Science

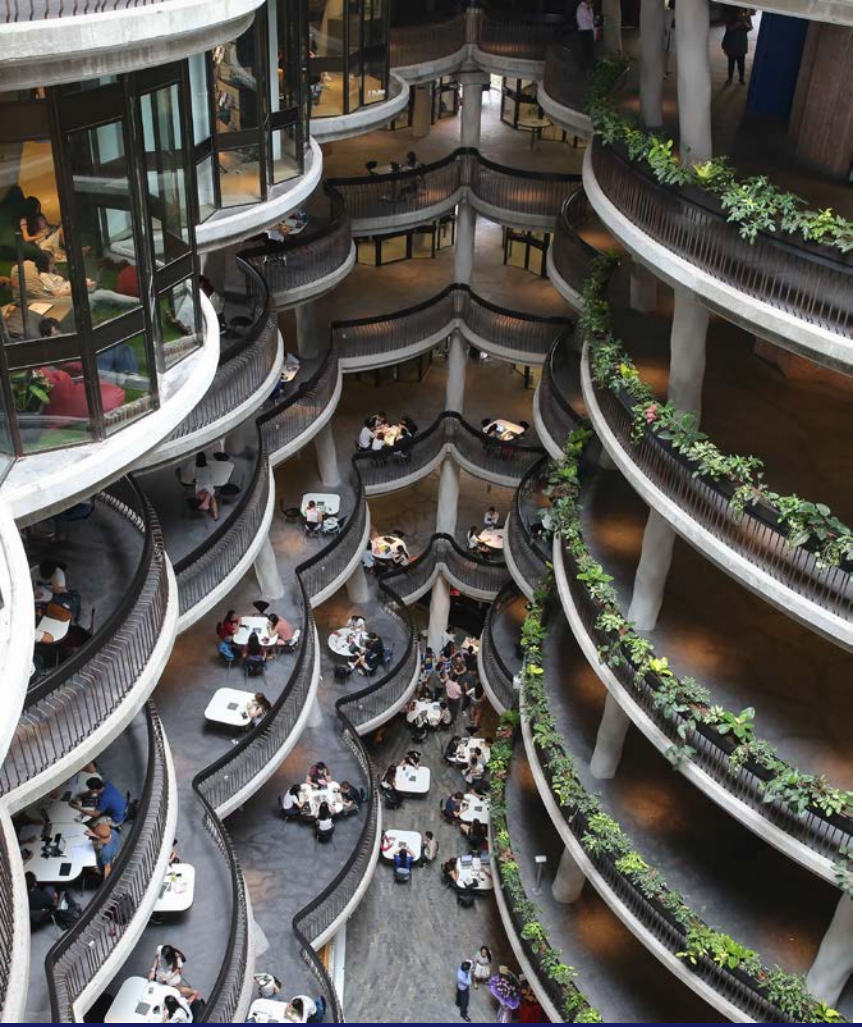
Practical Motivation

Prediction : Numeric

How much?
How many?

Give an example of this type of a Data Science problem in real life.

Justify why the example belongs to this category.



Data Science

Practical Motivation

Prediction : Classes

Is it type A or type B?

Give an example of this type of a Data Science problem in real life.

Justify why the example belongs to this category.

Data Science

Practical Motivation

Detection : Structure

How is this organized?

Give an example of this type of a Data Science problem in real life.

Justify why the example belongs to this category.

Data Science

Practical Motivation

Detection : Anomaly

Is it weird behavior?

Give an example of this type of a Data Science problem in real life.

Justify why the example belongs to this category.

https://en.wikipedia.org/wiki/Anomaly_detection

Data Science

Practical Motivation

Decision : Action

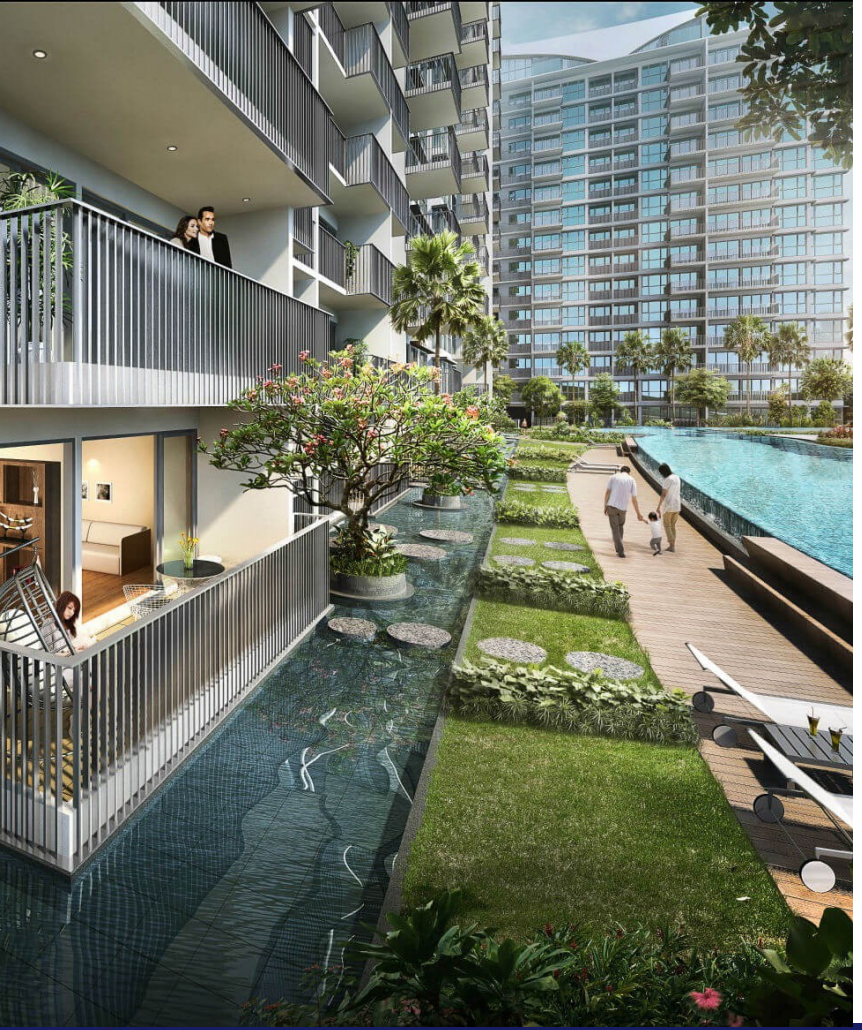
What should be done next?

Give an example of this type of a Data Science problem in real life.

Justify why the example belongs to this category.

Let us take a couple of …

EXAMPLES



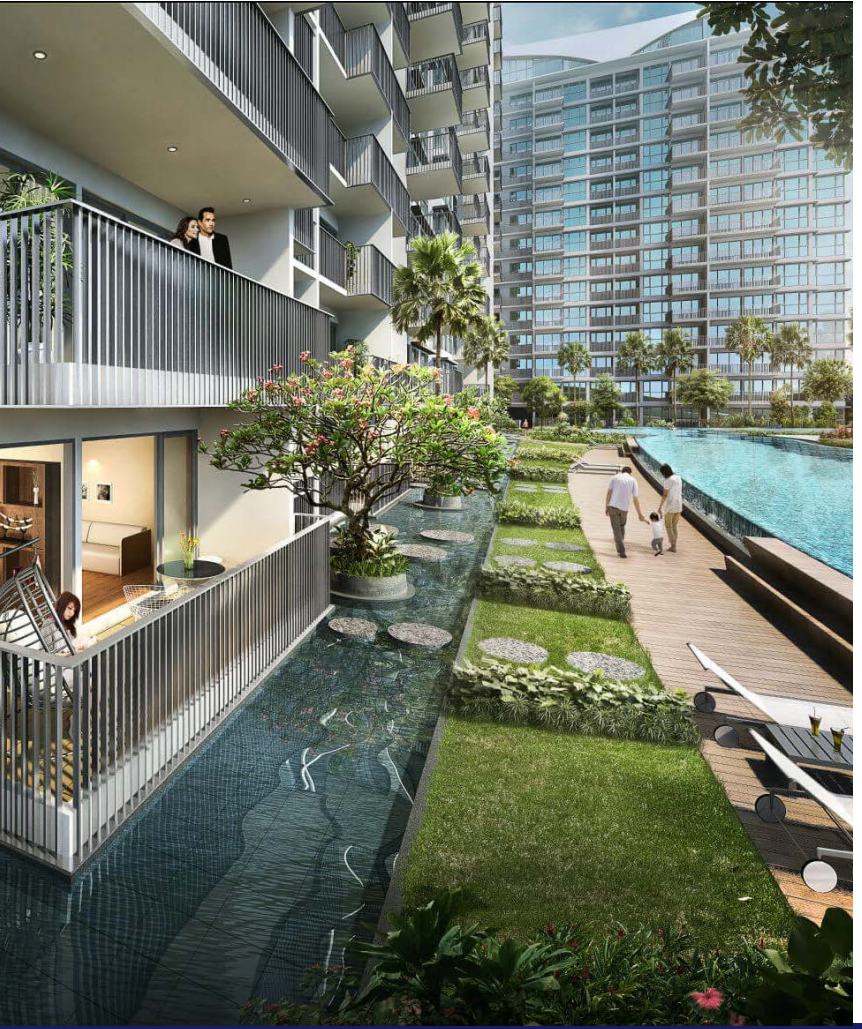
Data Science

Problem Formulation

Suppose that one of your close relatives is planning to buy a Condo in Singapore.

Devise a strategy for them to judge if the price quoted by Seller is expected, higher than expected, or lower than expected, for any given Condo.

How would you perform the Sample Collection?
Is there any requirement for Data Preparation?
How would you finally Formulate the Problem?



Data Science

Estimation Strategy

Which one of the methods do you prefer?

1. **Compare prices of “similar” condos to estimate the price of your target condo.**
2. **Build a “formula” to estimate price of a condo given its features and attributes.**

There are several Model Families in data science.

Choice 1 will imply **Nearest Neighbors**.
Choice 2 will imply **Regression Models**.

https://scikit-learn.org/stable/supervised_learning.html

Data Science

Problem Formulation

Suppose that one of your friends from Biology wants to know how many types of Whales are there in South China Sea.

Devise a strategy to help your friend in identifying the types of Whales that are found generally in the South China Sea.

How would you perform the Sample Collection?
Is there any requirement for Data Preparation?
How would you finally Formulate the Problem?

Data Science

Problem Formulation

Given two Whales and their relevant data, how would you know if they are of similar or dissimilar based on your strategy?

Suppose you have the data on Weight, Length, Girth, Flipper Span, Fin Size, Tail Size, Hump (Y/N) in a DataFrame.

How would you define “similarity” in this context?
Is there a difference between numeric / categorical?
Which kind of a distribution will you consider for each?

https://scikit-learn.org/stable/unsupervised_learning.html

Labs and beyond …