

EE3001 Foundations of Data Engineering

Lecturer: Dr. Shiu Yin YUEN, Kelvin

Email: kelviny.ee Tel: x 7717 Rm: G6359

Web: <http://www.ee.cityu.edu.hk/~syyuen>

Course Aim

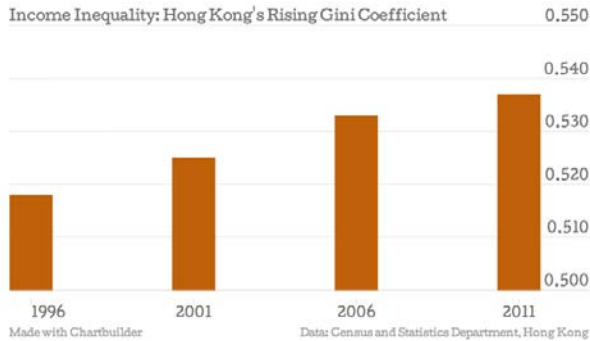
It aims to teach **probability** and **statistics** from a data engineering perspective

It prepares you to take more advanced data engineering courses as well as gives you generic skills to handle other courses that requires knowledge about probability and statistics

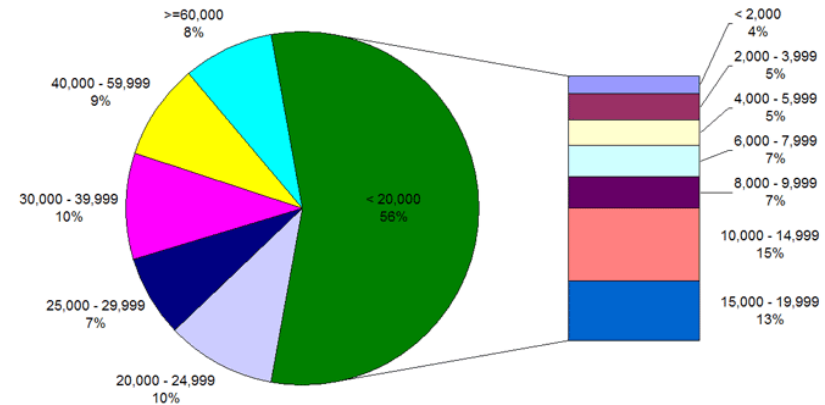
(Caution: Though some of you may have learnt some material before, the course goes far beyond what you have learnt in Mathematics and M1 in secondary school. So not recommend you to miss lessons)

Data helps us to identify problems

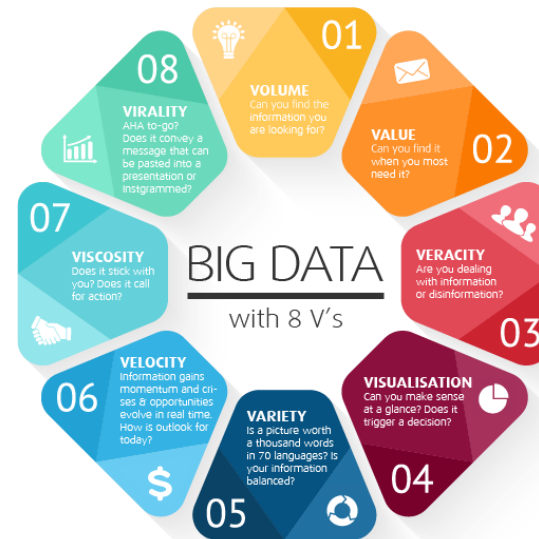
Income Inequality: Hong Kong's Rising Gini Coefficient



Domestic households by monthly income



港鐵沿線屋苑住戶月入中位數



Data helps us to identify correlations and trends

港元匯價 » 美元 7.7534 0.00% | 人民幣 1.1647 +0.09% | 歐羅 8.7842 +0.07% | 日圓 0.0771 +0.04% | 英鎊 10.1800 -0.05% | 澳元 5.9154 0.00% |
紐元 5.6604 +0.16% | 加元 5.9950 +0.06%

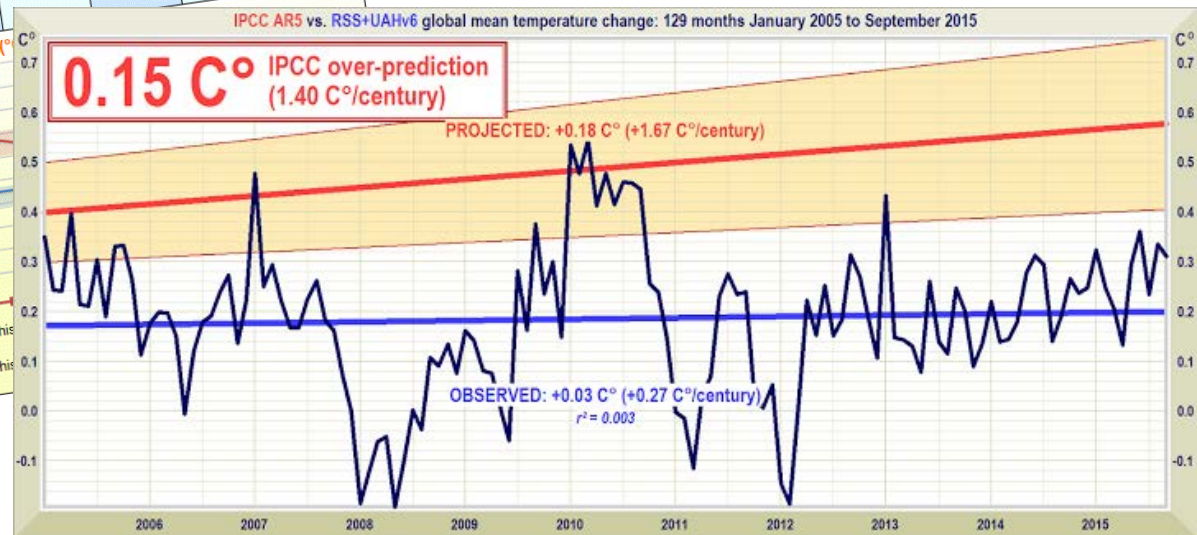
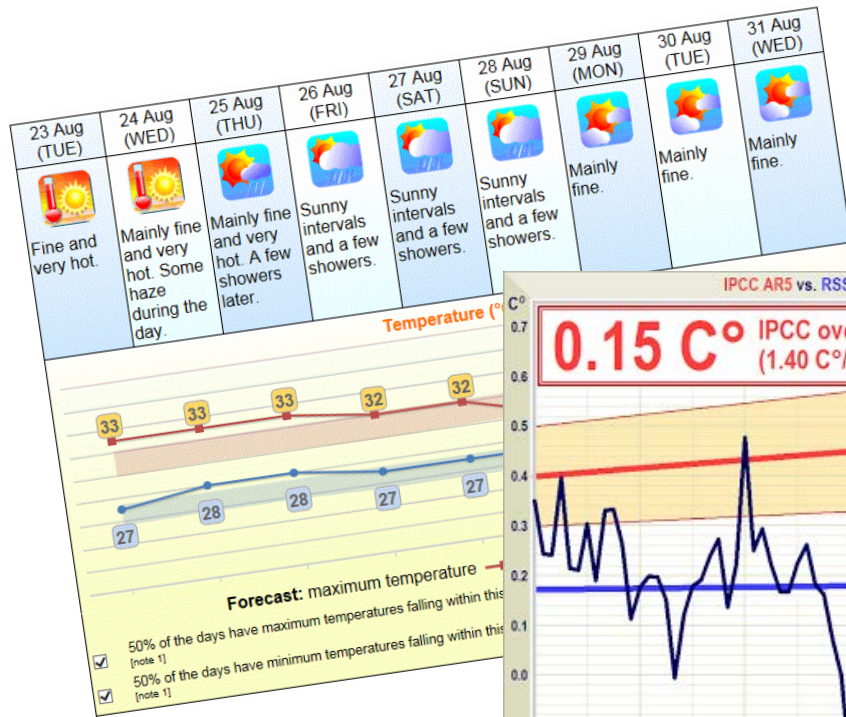


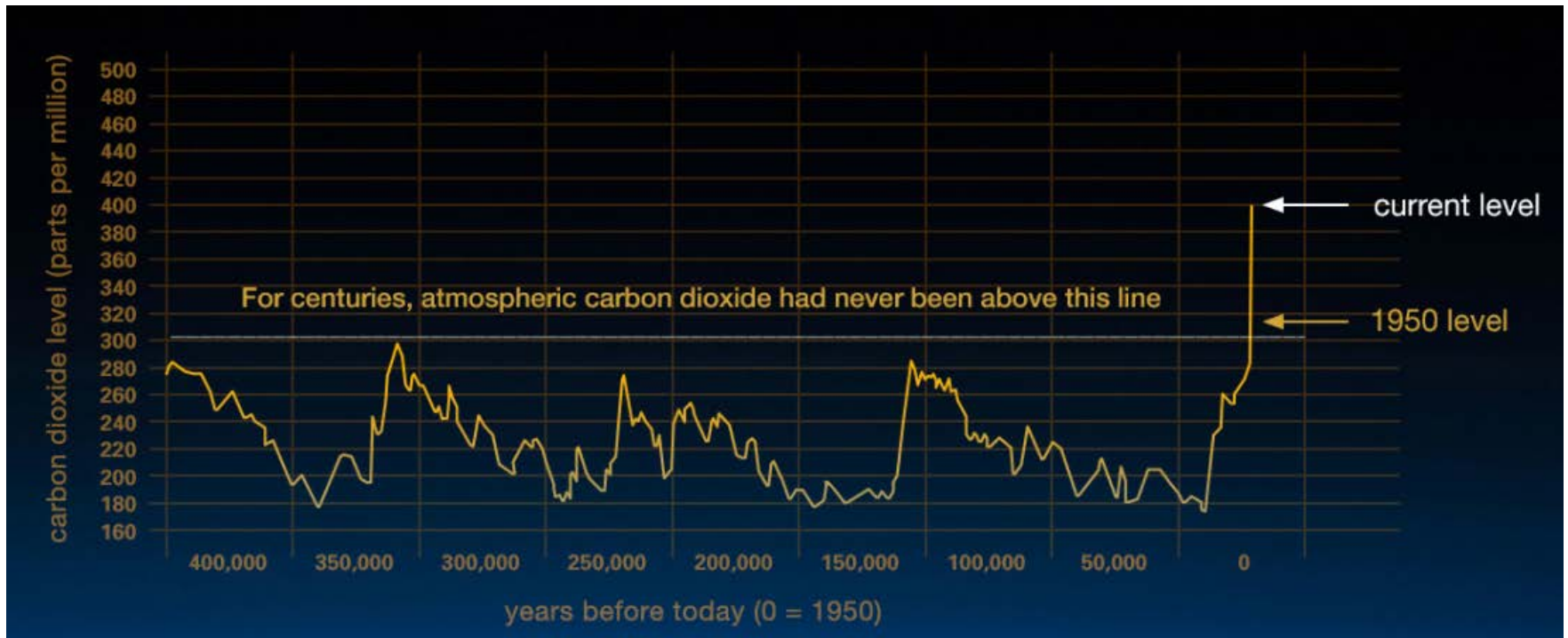
商品期貨 » 紐約期油 46.70 -1.50% | 布蘭特期油 48.56 -1.22% | 金價 1,340.80 -0.19% | 銀價 18.92 +0.35% | 銅價 2.14 +0.23%

Hang Seng Index



Data helps us to make predictions



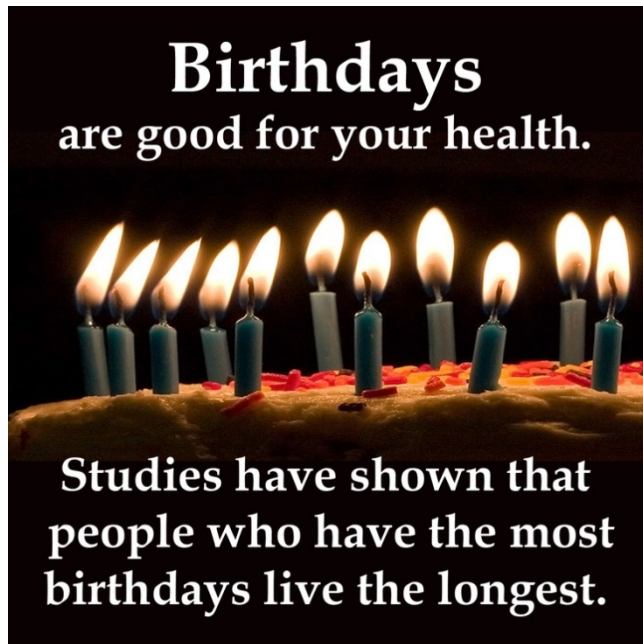


Predictions: global temperature rise, warming oceans, shrinking ice sheets, sea level rise, extreme events, ocean acidification, biodiversity loss, ...

Source: <https://climate.nasa.gov/evidence/>

Beware of incorrect use/pitfalls/limitations in using data

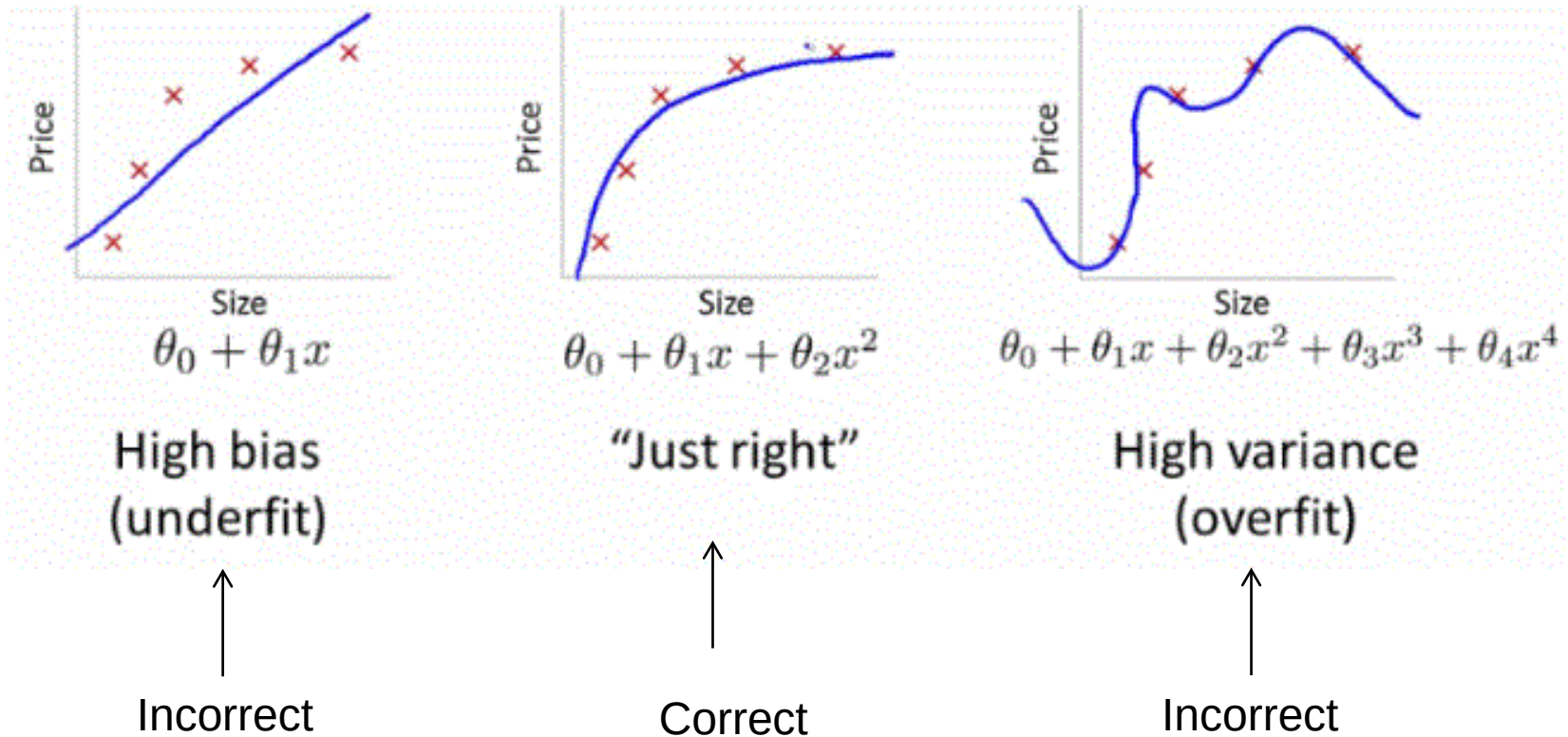
Example 1 Decision making misled by meaningless correlation



Pigeon superstition (Skinner's experiment)



Example 2 Incorrect prediction of trend



Hang Seng Index

INDEXHANGSENG: HSI

26,284.09 -383.18 (1.44%) ↓

14 Jan, 1:29 pm HKT · Disclaimer

1 day

5 days

1 month

6 months

YTD

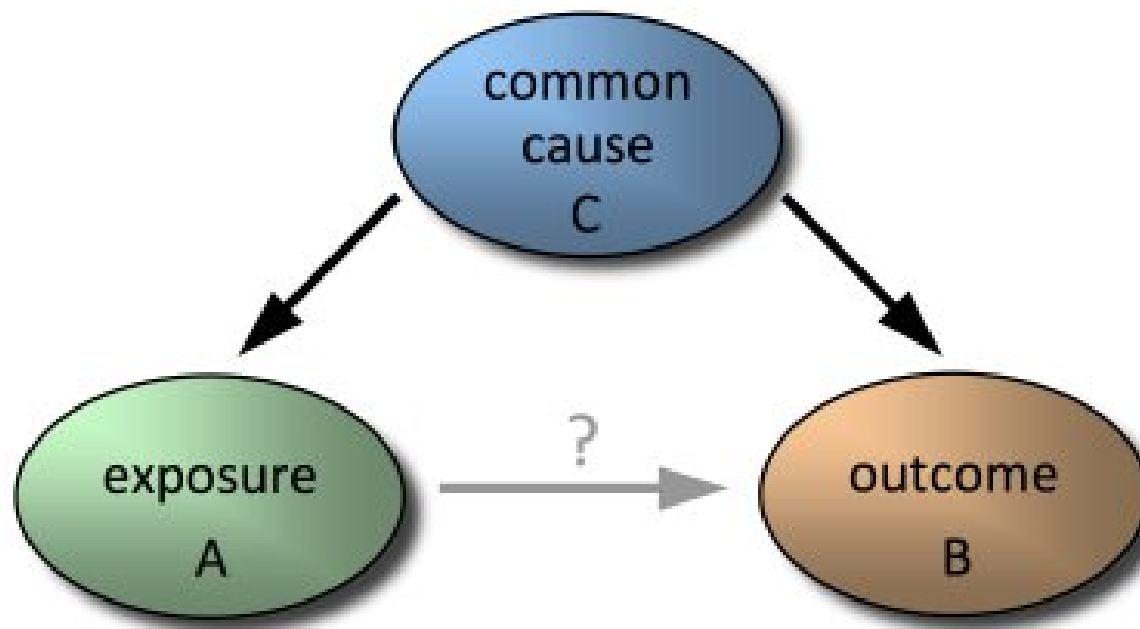
1 year

5 years

Max



Example 3 Identifying meaningful association while failing to find the underlying common cause



Bait Shyness in rats
(Garcia et al.'s experiment)

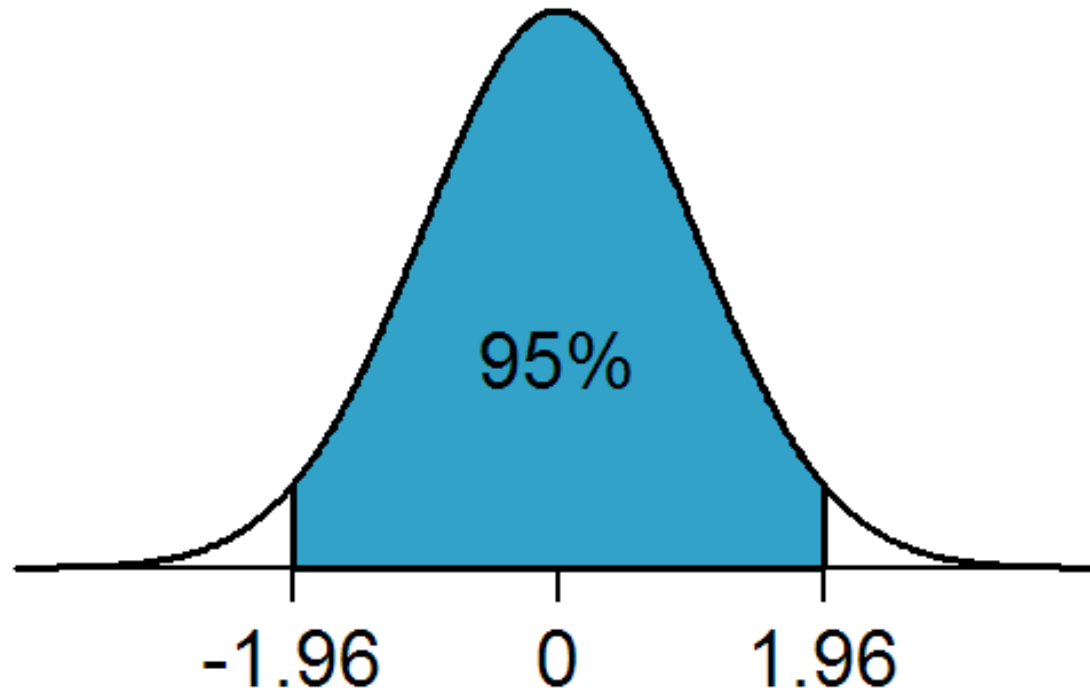


Pigeon superstition
(Skinner's experiment)



For more details, see Machine learning today and tomorrow: a panel discussion, 2017 [youtube](#) (8:20 – 20:05)

Example 4 Misinterpretation of statistics



The 95% confidence interval does not mean that there is a 95% probability that the value of the variable lies within the interval

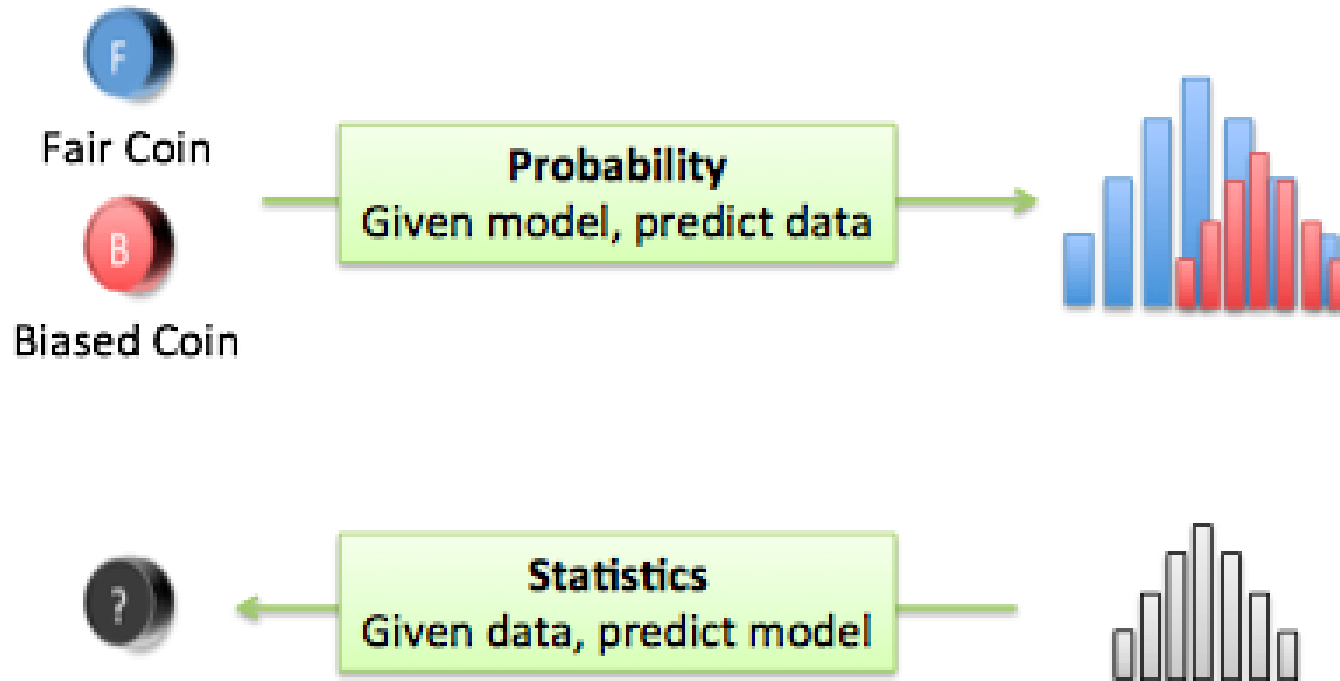
Example 5 Using a normal distribution when the distribution is not normal

Something you would understand well if you pay attention in the course:

An example of normal distribution is the grade distribution in a large class. Do you know what assumptions are behind it?

The number of attempts required to obtain a PASS in this course do not follow the normal distribution. Why?

Probability & Statistics



also discuss the applications in
Data Engineering
as well as uses and misuses

Motivations

- To learn probability
 - It is a general mathematical tool useful in other subjects.
 - It gives you better reasoning power in making decisions that involve risk taking
- To learn statistics
 - There are many statistics in our daily life. It gives you an understanding of how are they arrived at and how they can be used
 - It equips you on how to correctly use the data and avoid misuse of the data

Course Content

- Probability
 1. Probability
 2. Random Variables and Expectation

- Applications in Data Engineering
 3. Statistical Descriptors
 4. Special Random Variables
 5. Parameter Estimation
 6. Hypothesis Testing
 7. Non-parametric Statistics
 8. Finding Correlation or Lack of Correlation
 9. Linear Regression and Other Prediction Methods

Assessment Method

- Continuous Assessment 40%

 - Quiz 1 16 %

 - Quiz 2 16 %

 - Assignments 8 %

- Examination (2 hours) 60%

Students need to obtain at least 30 (out of 100) in both the Continuous Assessment and Examination to pass the course

Coursework Components

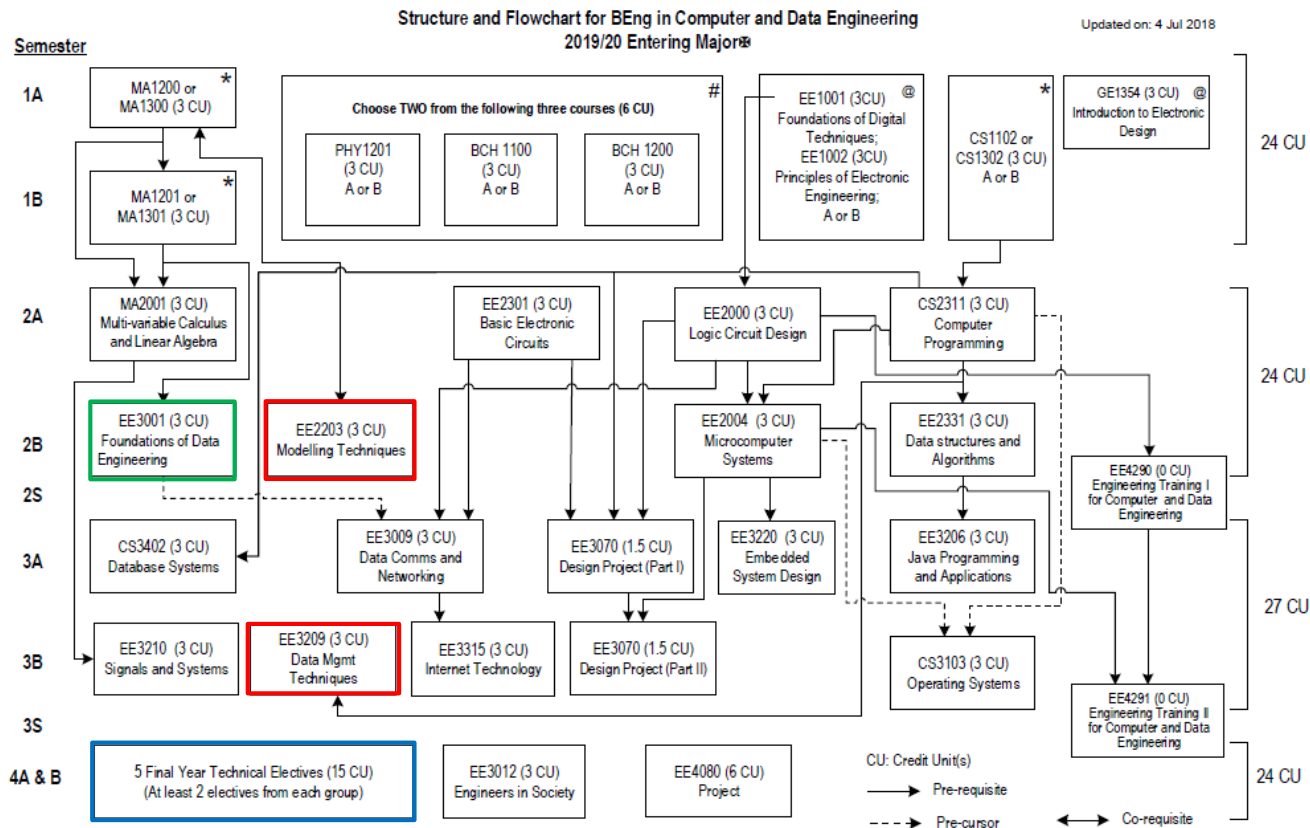
Time	Item	Scope	Percentage
	Assignments		20%
Wk 7 (5th Mar.)	Quiz 1	everything taught in Wk 1-6	40%
Wk 13 (16th Apr.)	Quiz 2	everything taught in Wk 7-12	40%

Late Penalty

- **Students will not be allowed late submission of assessment tasks without the prior approval of the course lecturer.** Late submission without approval will be given zero mark.
- For late submission with genuine and valid reasons, the maximum penalty is 50% per day. The actual penalty will be decided after considering the reasons and communicated to the student.

Related Data Engineering Courses

- EE2203 Modelling Techniques (Core)
- EE3209 Data Management Techniques (Core)
- EE4146 Data Informatics and Learning Systems (Elective)
- EE4222 Digital Forensics (Elective)



Text book and References

- Text book

S.M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, 5th Edition, Elsevier 2014.

- References

J.P. Marques De SÁ, Chance, the life of games & the game of life, Springer 2008.

S.M. Ross, A First Course in Probability, 9th Edition, Pearson 2014.

E-books of the above are available in CityU Library. They can be found in EE3001 Course Reserve.

Copyright Issues

The copyright of some text and images in the notes belong to the authors and should not be used elsewhere without permission. Other images are usually downloaded from Google images or from Wikipedia