EE3001 Foundations of Data Engineering

Lecturer: Dr. YUEN Shiu Yin, Kelvin

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

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

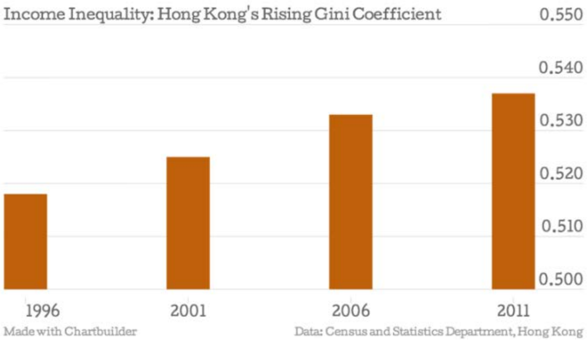
Course Aims

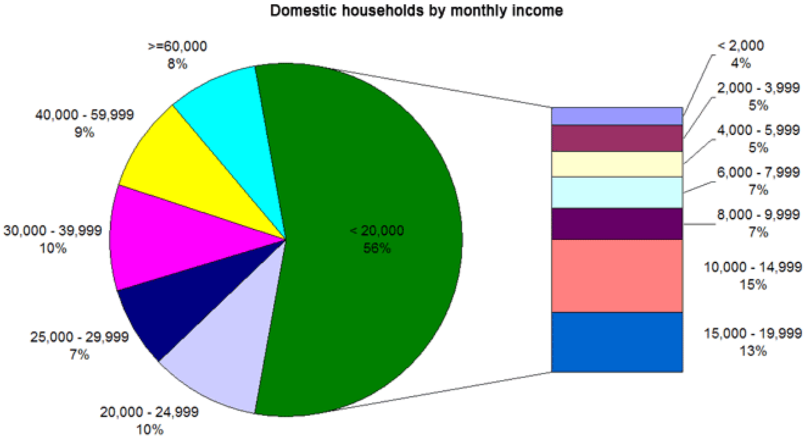
It aims to teach probability and statistics from a data engineering perspective.

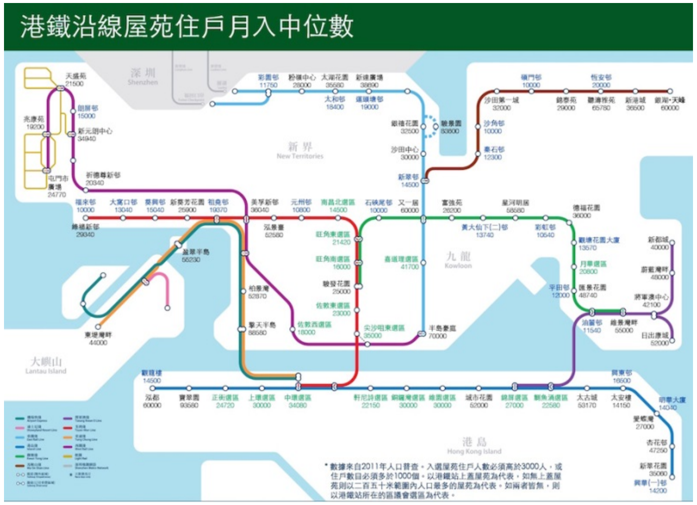
It prepares you to take more advanced data engineering courses as well as gives you generic skills to handle other courses that requires knowledge about probability and statistics

Covers M1 in local secondary school, but goes well beyond it in depth, particularly theoretical derivations. Strongly encourage you to attend ALL lessons irrespective of whether you have taken M1 or not.

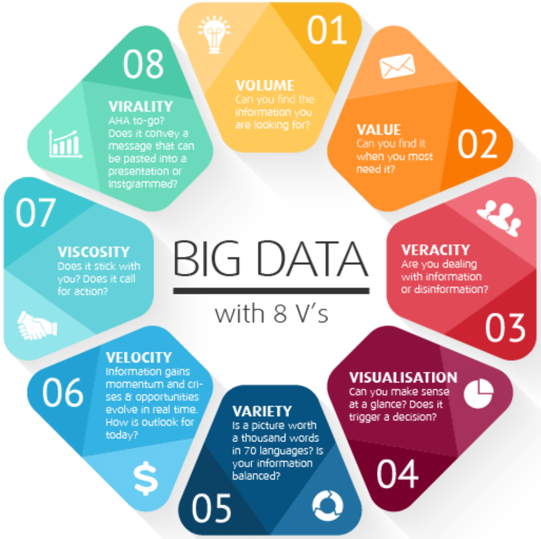
Data helps us to identify problems



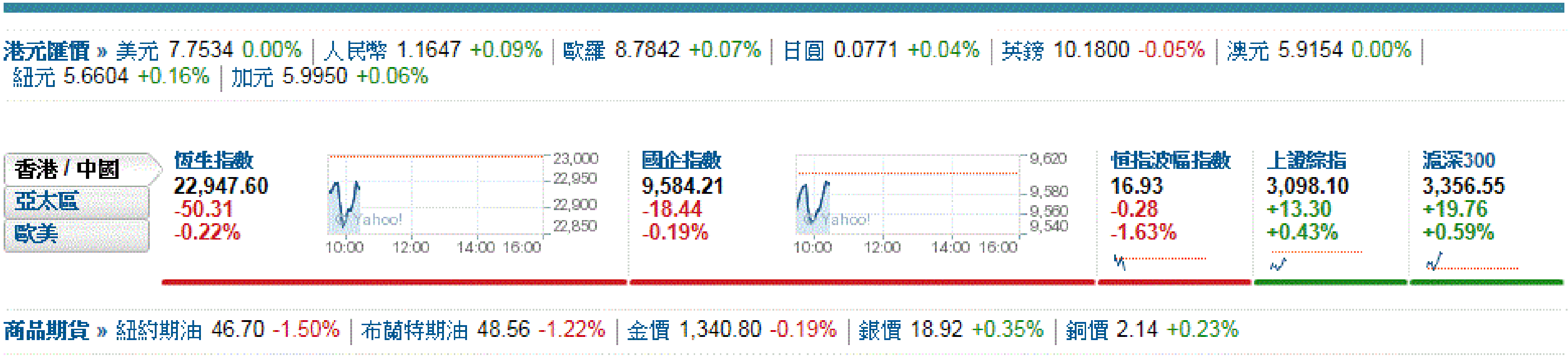




* 數據來自2011年人口普查。入選屋苑住戶人數必須高於3000人，或住戶數目必須多於1000個。以港鐵站上蓋屋苑為代表，如無上蓋屋苑則以二百五十米範圍內人口最多的屋苑為代表。如兩者皆無，則以港鐵站所在的區議會選區為代表。


BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is Outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

Data helps us to identify correlations and trends

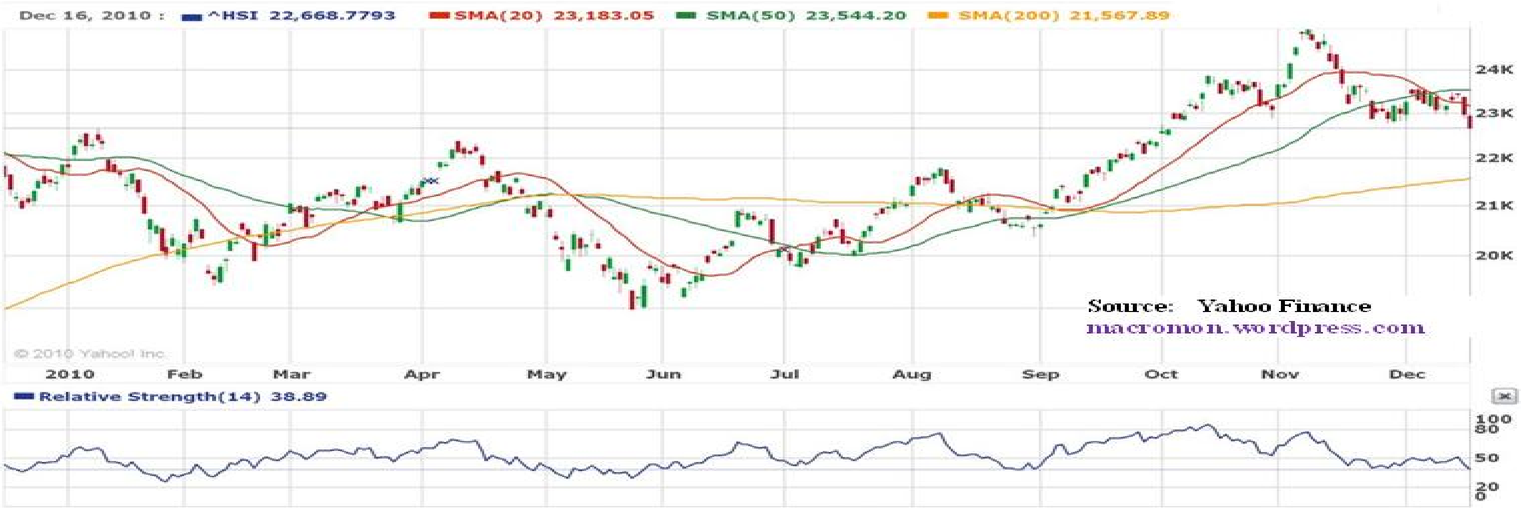


港元匯價 » 美元 7.7534 0.00% | 人民幣 1.1647 +0.09% | 歐羅 8.7842 +0.07% | 日圓 0.0771 +0.04% | 英鎊 10.1800 -0.05% | 澳元 5.9154 0.00% |
紐元 5.6604 +0.16% | 加元 5.9950 +0.06%

香港 / 中國 亞太區 歐美	恆生指數	國企指數	恒指波幅指數	上證綜指	滬深300
	22,947.60	9,584.21	16.93	3,098.10	3,356.55
	-50.31	-18.44	-0.28	+13.30	+19.76
	-0.22%	-0.19%	-1.63%	+0.43%	+0.59%

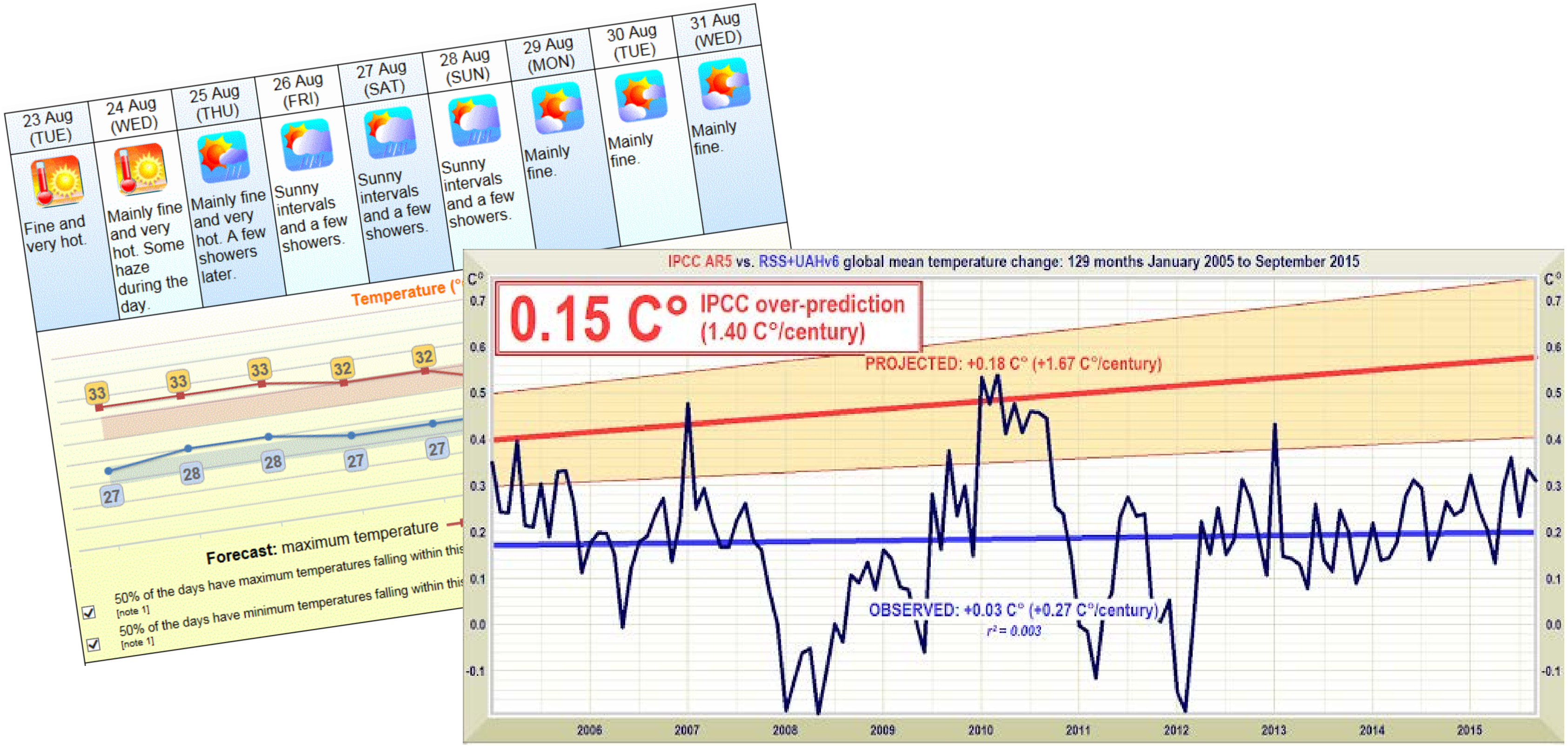
商品期貨 » 紐約期油 46.70 -1.50% | 布蘭特期油 48.56 -1.22% | 金價 1,340.80 -0.19% | 銀價 18.92 +0.35% | 銅價 2.14 +0.23%

Hang Seng Index

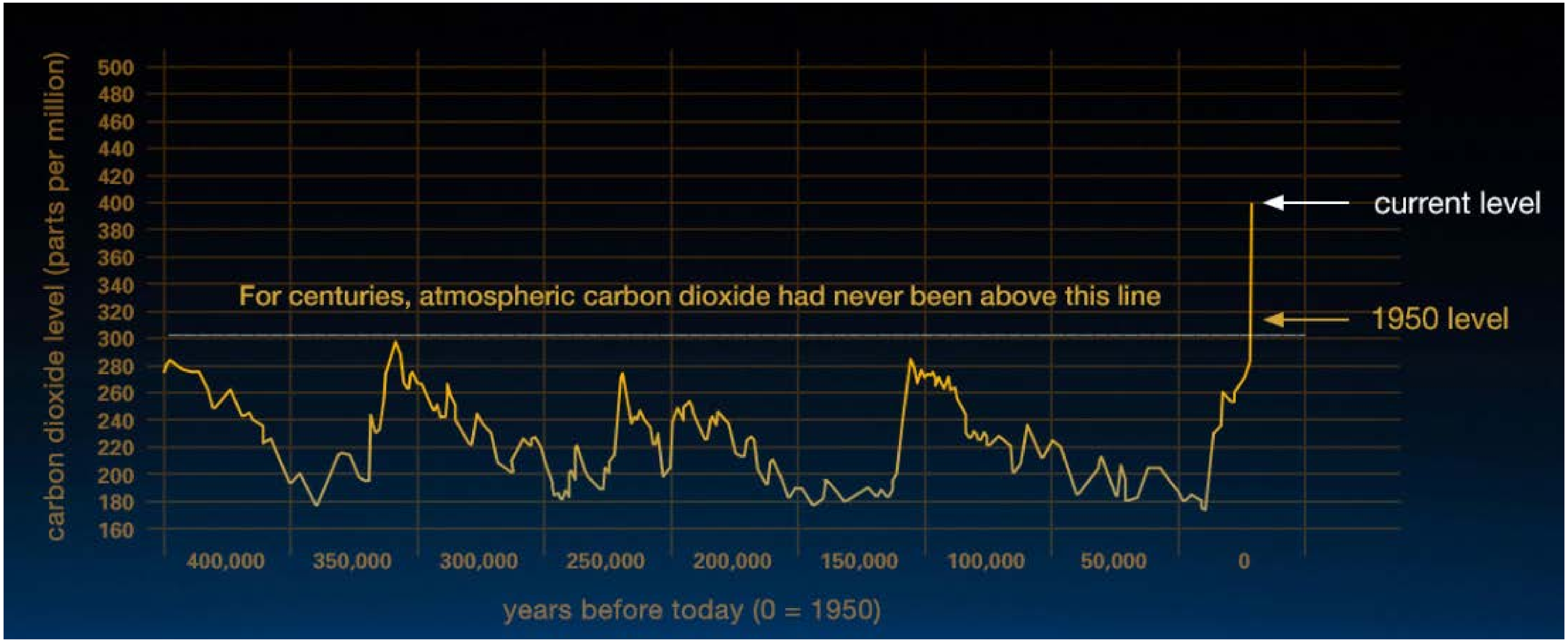


Source: Yahoo Finance
macromon.wordpress.com

Data helps us to make predictions



23 Aug (TUE)	24 Aug (WED)	25 Aug (THU)	26 Aug (FRI)	27 Aug (SAT)	28 Aug (SUN)	29 Aug (MON)	30 Aug (TUE)	31 Aug (WED)
Fine and very hot.	Mainly fine and very hot. Some haze during the day.	Mainly fine and very hot. A few showers later.	Sunny intervals and a few showers.	Sunny intervals and a few showers.	Sunny intervals and a few showers.	Mainly fine.	Mainly fine.	Mainly fine.

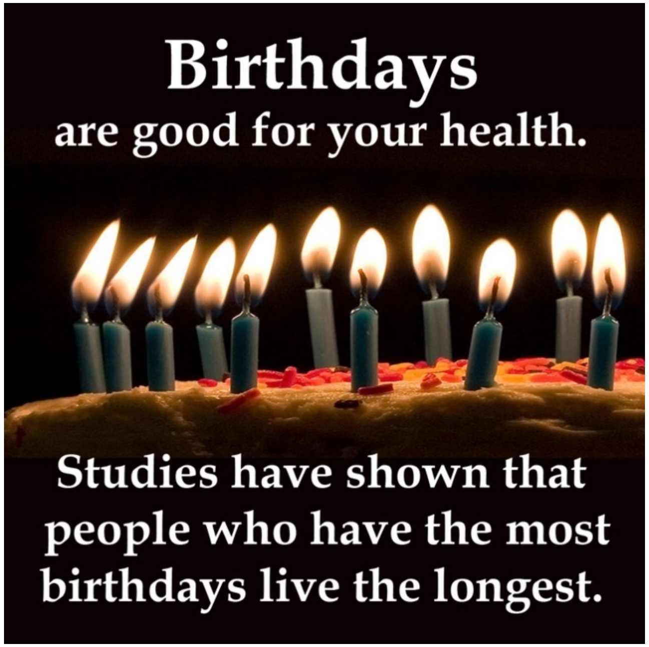


Predictions: global temperature rise, warming oceans, shrinking ice sheets, sea level rise, extreme events, ocean acidification, biodiversity loss, …

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

Beware of incorrect use/pitfalls/limitations in using data

Example 1 Decision making misled by meaningless correlation



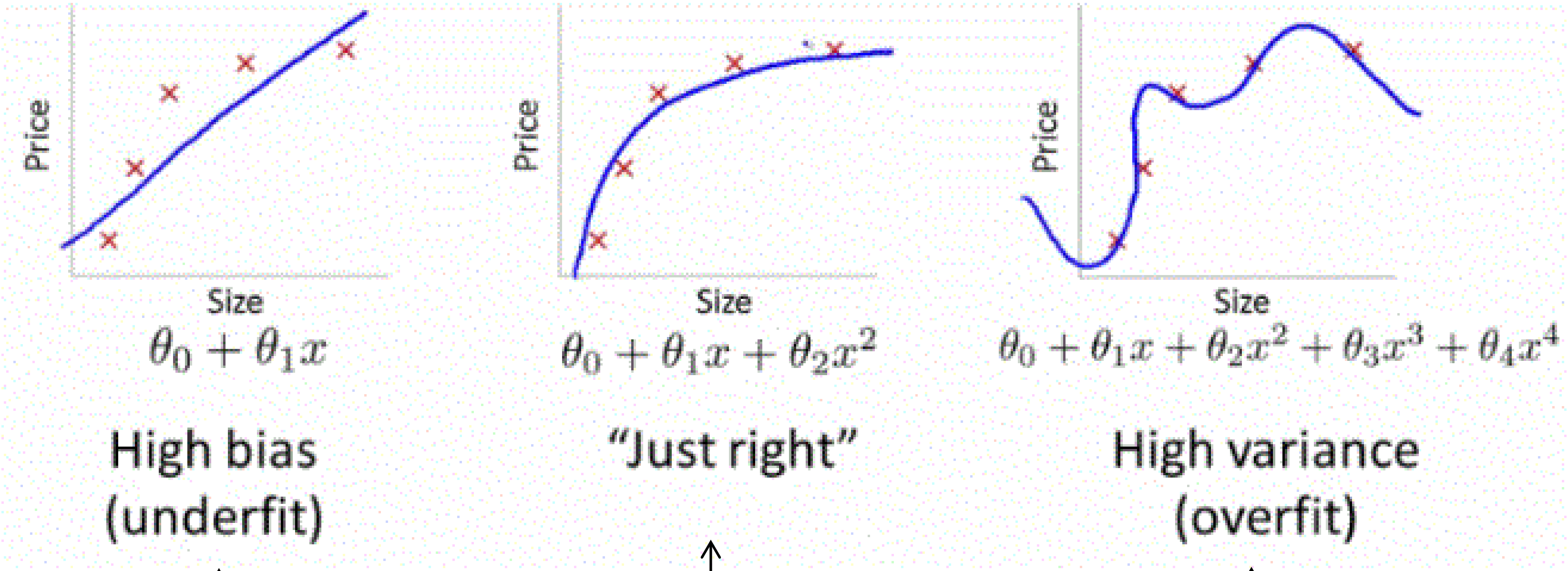
Birthdays

are good for your health.

Studies have shown that people who have the most birthdays live the longest.

Pigeon superstition (Skinner’s experiment)

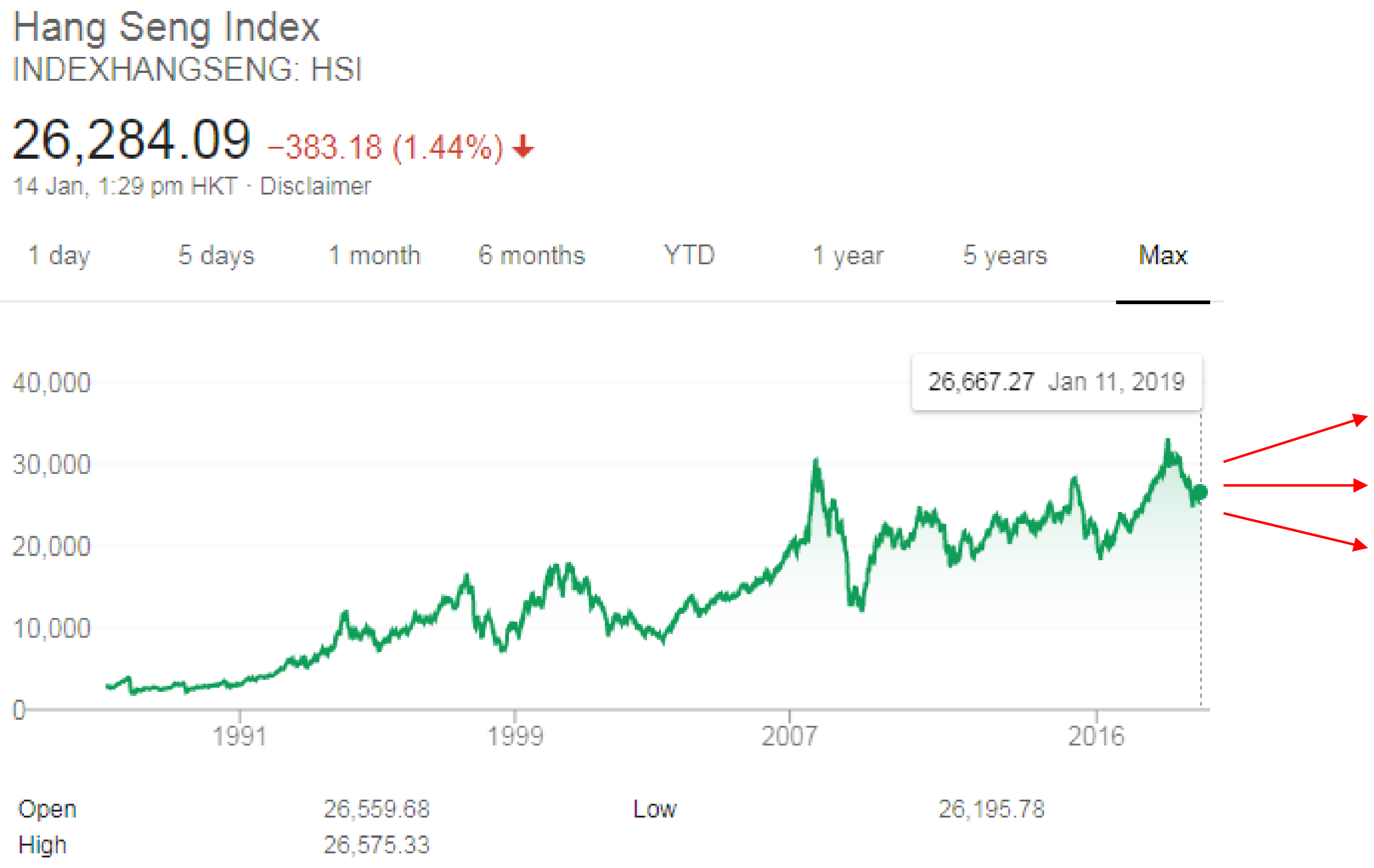
Example 2 Incorrect prediction of trend



Incorrect

Correct

Incorrect

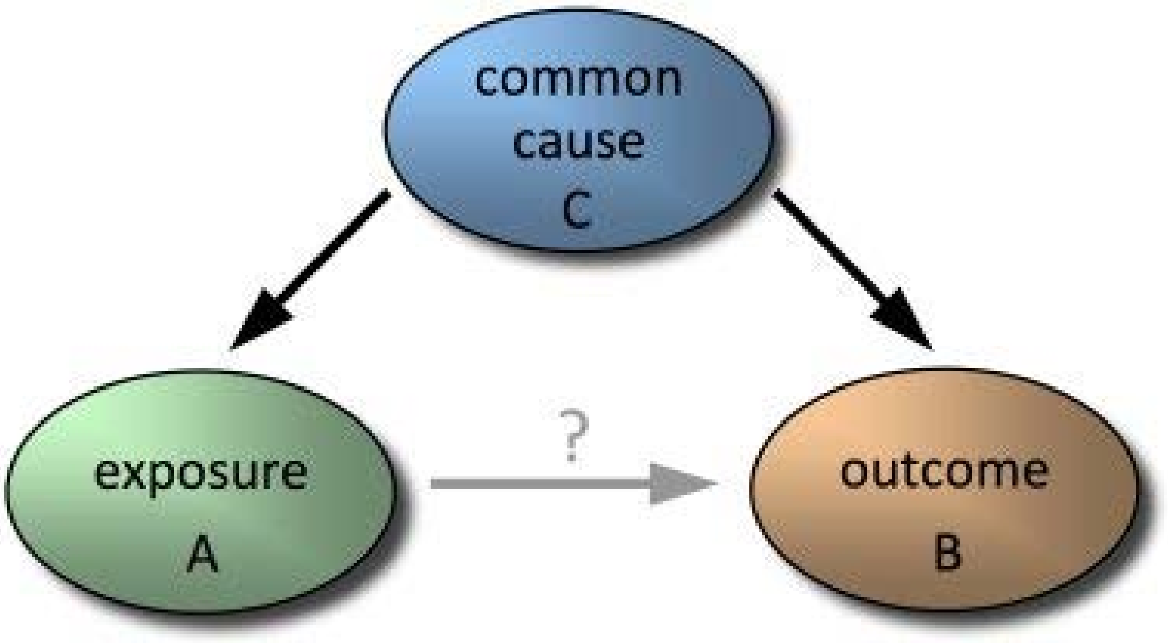


Hang Seng Index
INDEXHANGSENG: HSI

26,284.09 −383.18 (1.44%)
14 Jan, 1:29 pm HKT · Disclaimer

Open	26,559.68	Low	26,195.78
High	26,575.33		

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

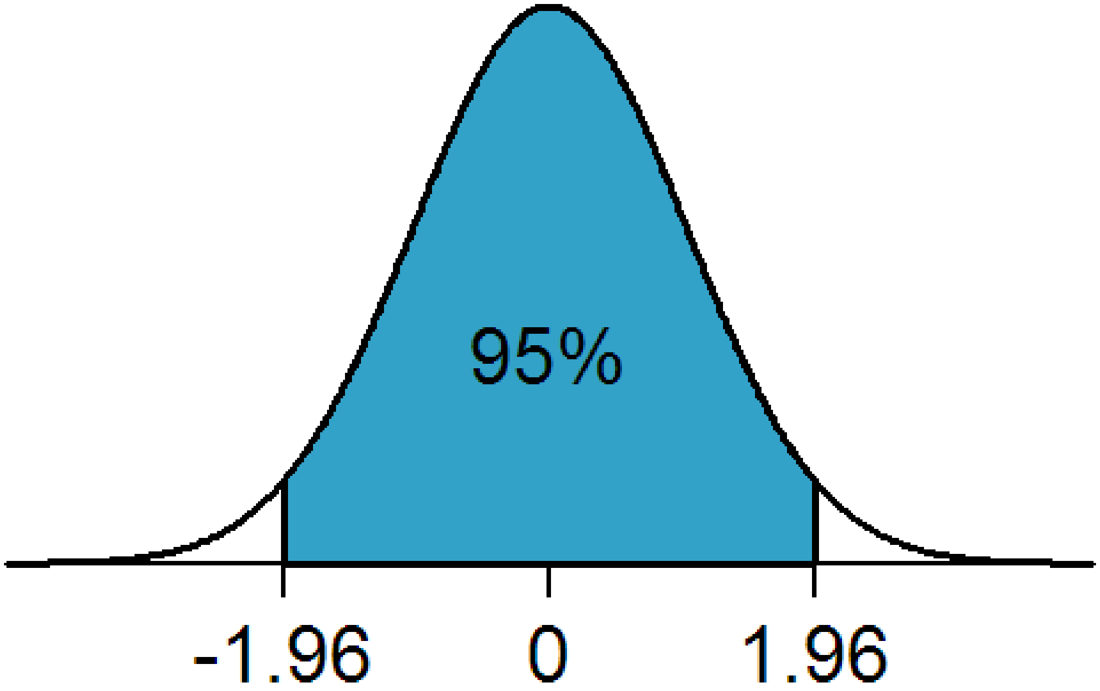


Bait Shyness in rats
(Garcia et al.'s experiment)

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

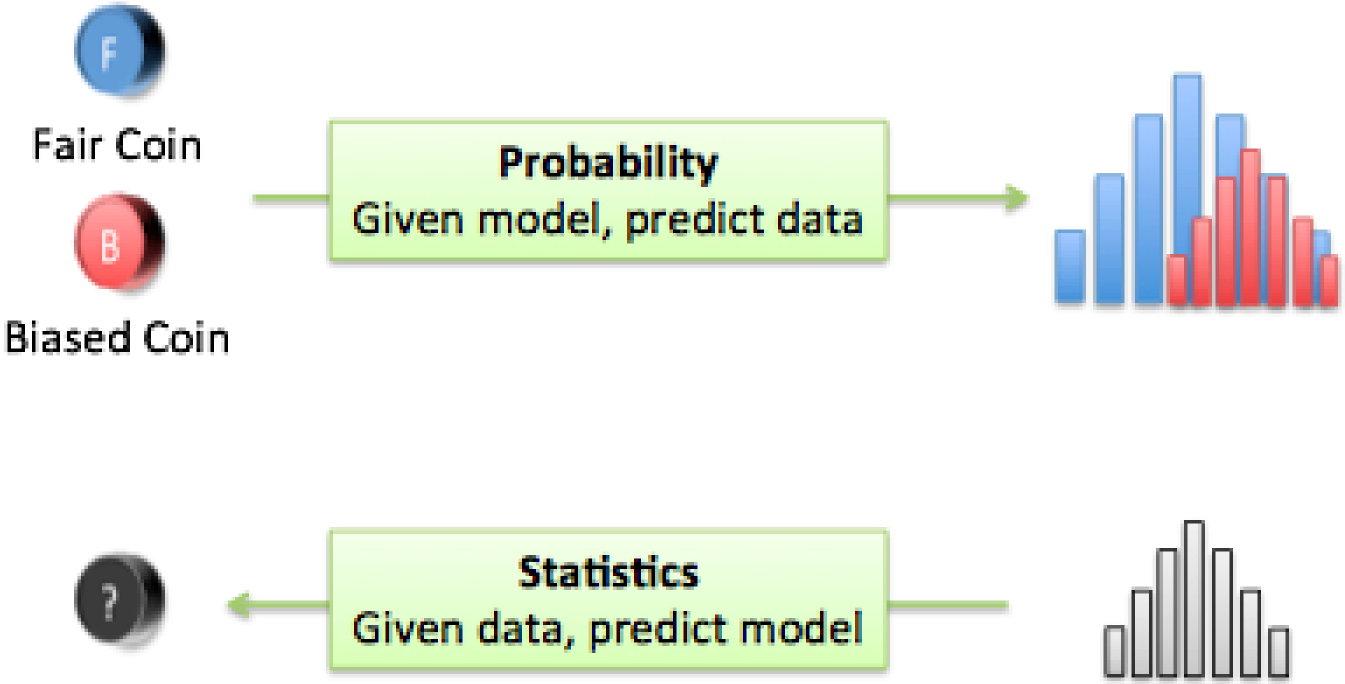
Example 5 Using a normal distribution when the distribution is not normal

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



Fair Coin

Biased Coin

Probability
Given model, predict data

Statistics
Given data, predict model

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

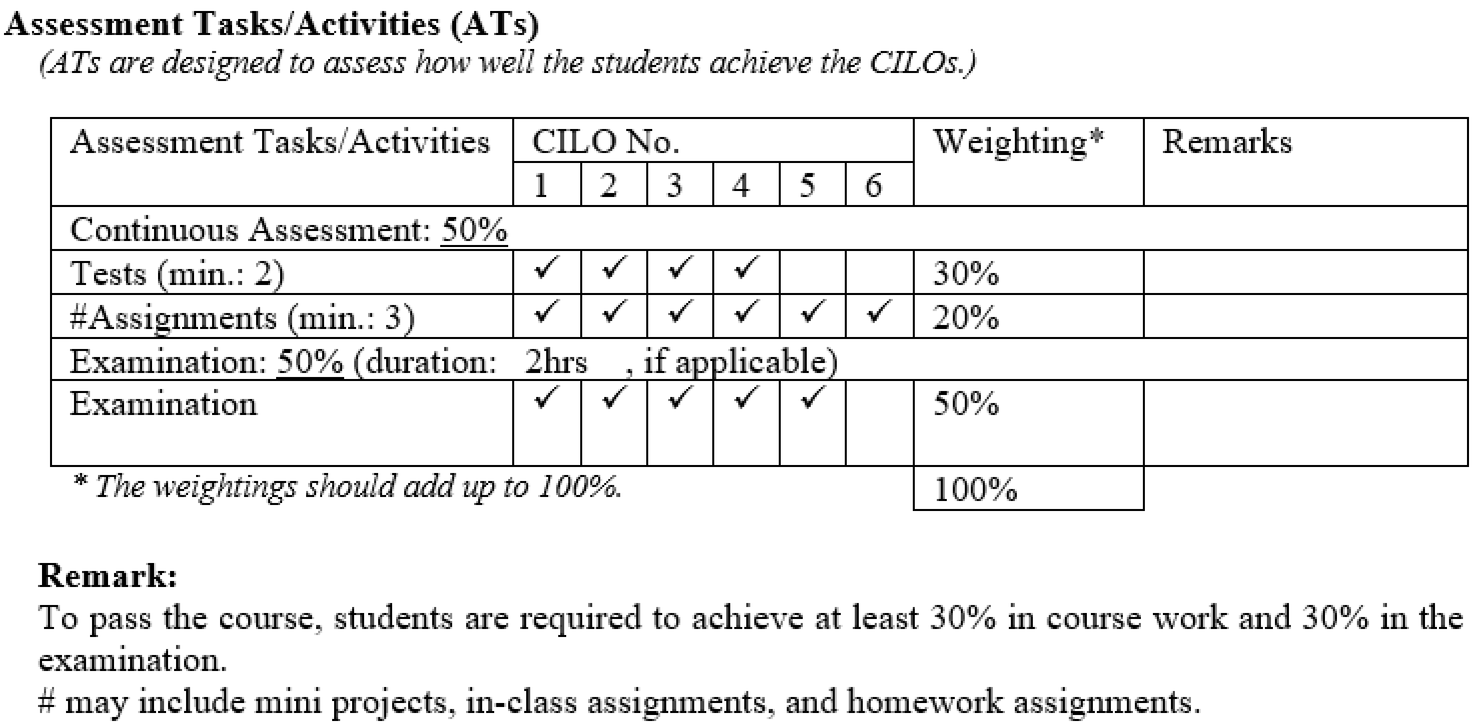
Motivations for studying probability and statistics

- Probability
  - It is a general mathematical tool useful in other subjects.
  - It gives you better reasoning power in making decisions that involve risk taking
- 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

1. Introduction
2. Probability
3. Random Variables and Expectation
4. Statistical Descriptors
5. Special Random Variables
6. Parameter Estimation
7. Hypothesis Testing
8. Non-parametric Statistics
9. Linear Regression and Other Prediction Methods
10. Multiple Comparison Tests

Assessment and Schedule



Assessment Tasks/Activities (ATs)

(ATs are designed to assess how well the students achieve the CILOs.)

Assessment Tasks/Activities	CILO No.						Weighting*	Remarks
	1	2	3	4	5	6		
Continuous Assessment: 50%								
Tests (min.: 2)	✓	✓	✓	✓			30%	
#Assignments (min.: 3)	✓	✓	✓	✓	✓	✓	20%	
Examination: 50% (duration: 2hrs , if applicable)								
Examination	✓	✓	✓	✓	✓		50%	
** The weightings should add up to 100%.*							100%	

Remark:

To pass the course, students are required to achieve at least 30% in course work and 30% in the examination.

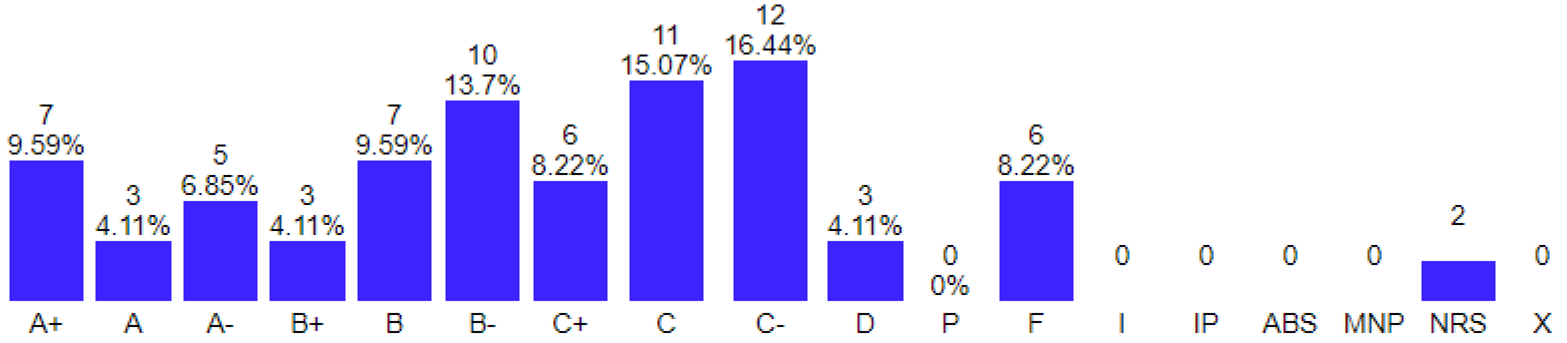
may include mini projects, in-class assignments, and homework assignments.

Coursework Components (50%)

Time	Item	Scope	Percentage
Wk 6	**Quiz 1**	everything taught in Wk 1-5	15%
Wk 12	**Quiz 2**	everything taught in Wk 6-11	15%
	In-Class Assignments and/or Assignments		20%

In-class assignments refer to assignments conducted during lecture

Grade distribution Report in 2018/19



Teaching Assistant 1

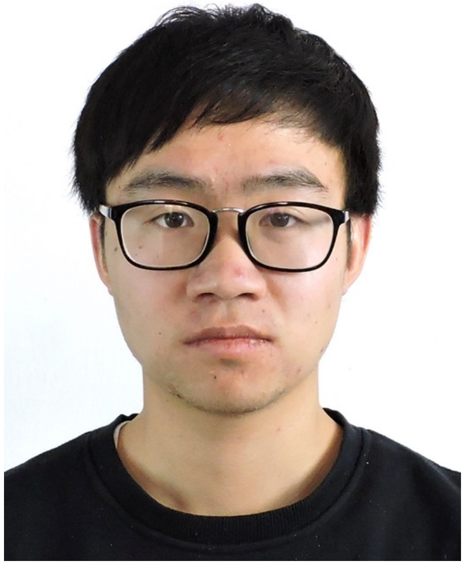
Ms. CHEN Xueli

Office: FYW2386   Tel: 3442 2615

Email: xuelichen3-c@my.cityu.edu.hk

Teaching Assistant 2

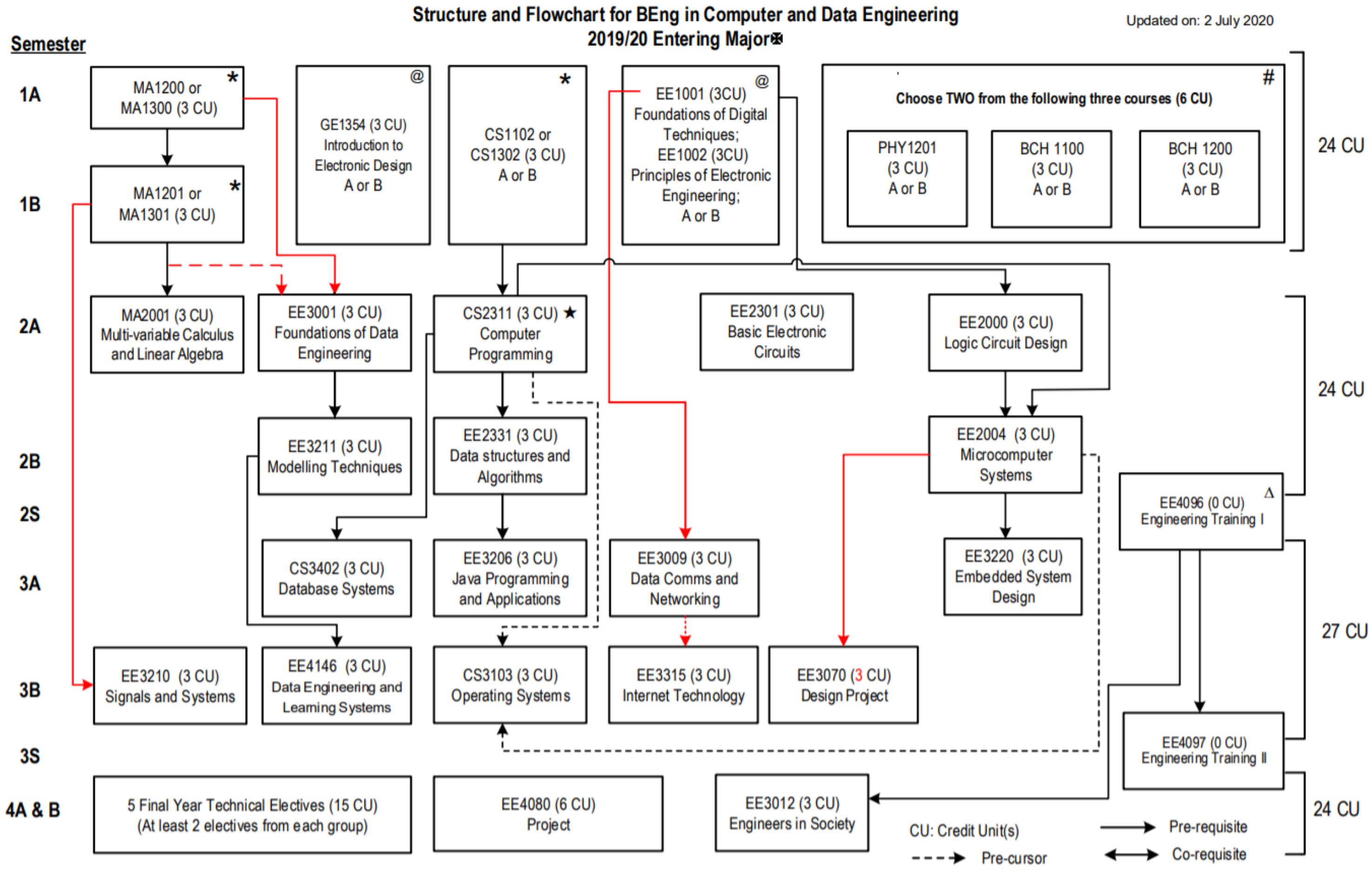
Mr. JIANG Mingjie



Office: FYW2384   Tel: 3442 4119

Email: minjiang5-c@my.cityu.edu.hk

Related Data Engineering Courses



Structure and Flowchart for BEng in Computer and Data Engineering
2019/20 Entering Major⊠

Updated on: 2 July 2020

Semester

1A

MA1200 or MA1300 (3 CU) *

GE1354 (3 CU) Introduction to Electronic Design A or B @

CS1102 or CS1302 (3 CU) A or B *

EE1001 (3CU) Foundations of Digital Techniques; EE1002 (3CU) Principles of Electronic Engineering; A or B @

Choose TWO from the following three courses (6 CU) #

PHY1201 (3 CU) A or B

BCH 1100 (3 CU) A or B

BCH 1200 (3 CU) A or B

24 CU

1B

MA1201 or MA1301 (3 CU) *

2A

MA2001 (3 CU) Multi-variable Calculus and Linear Algebra

EE3001 (3 CU) Foundations of Data Engineering

CS2311 (3 CU) ★ Computer Programming

EE2301 (3 CU) Basic Electronic Circuits

EE2000 (3 CU) Logic Circuit Design

24 CU

2B

EE3211 (3 CU) Modelling Techniques

EE2331 (3 CU) Data structures and Algorithms

EE2004 (3 CU) Microcomputer Systems

2S

EE4096 (0 CU) Engineering Training I Δ

3A

CS3402 (3 CU) Database Systems

EE3206 (3 CU) Java Programming and Applications

EE3009 (3 CU) Data Comms and Networking

EE3220 (3 CU) Embedded System Design

27 CU

3B

EE3210 (3 CU) Signals and Systems

EE4146 (3 CU) Data Engineering and Learning Systems

CS3103 (3 CU) Operating Systems

EE3315 (3 CU) Internet Technology

EE3070 (3 CU) Design Project

3S

EE4097 (0 CU) Engineering Training II

4A & B

5 Final Year Technical Electives (15 CU) (At least 2 electives from each group)

EE4080 (6 CU) Project

EE3012 (3 CU) Engineers in Society

24 CU

CU: Credit Unit(s)

Pre-requisite

Co-requisite

Pre-cursor

College Requirements ★ College-specified GE courses @ Major Requirement (Level 1): applicable to Normative 4-year degree students only

⊠ Excluding Gateway Education Requirements in English, Chinese Civilization and Area Requirements Δ EE2000 or EE2301 or CS2311 are pre-requisites for EE4096

★ Not applicable to ASII students

Relationship with EE3211 Modelling Techniques

- You will learn the basic theory of linear regression and logistic regression in the EE3001 course
- In EE3211, you will learn how to interpret outputs from statistical programs (e.g. R, SAS). They will have hands on experience in the tutorial session to solve two real world problems below

EE3211 Solve the following real world problem using linear regression: Pediatrics Hypertension

- Investigate how the relationship between the blood-pressure levels of newborns and infants relate to subsequent adult blood pressure.

- One problem that arises is that the blood pressure of a newborn is affected by several extraneous factors that make this relationship difficult to study.

- In particular, newborn blood pressures are affected by:
  (1) Birthweight
  (2) the day of life on which blood pressure is measured

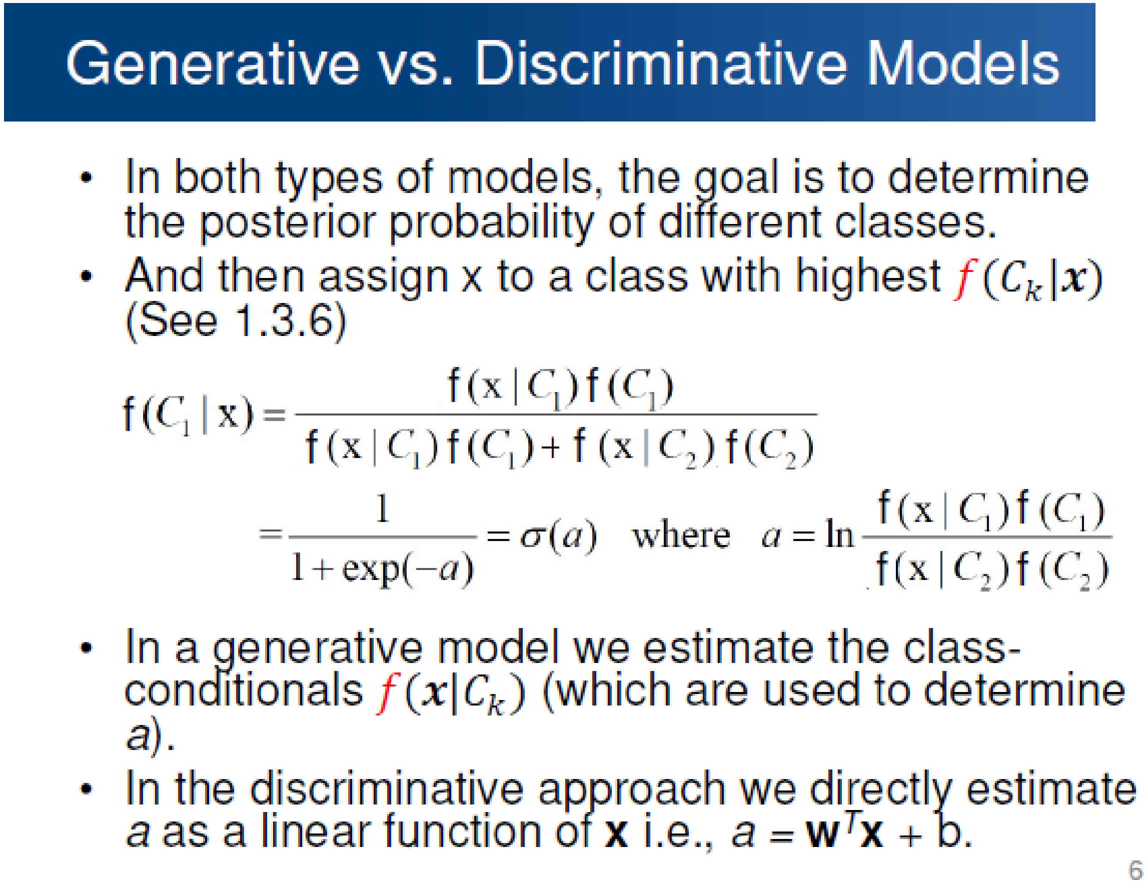
EE3211 Solve the following real world problem using logistic regression: Infectious Disease

- *Chlamydia trachomatis :* microorganism that has been established as an important cause of nongonococcal urethritis, pelvic inflammatory disease
- A study of risk factors for *C. trachomatis in* 431 female college students
- Because multiple risk factors may be involved, several risk factors must be controlled for simultaneously in analyzing variables associated with *C. trachomatis*

Relationship with EE4146 Date Engineering and Learning Systems

- You will learn in details the Bayes’ rule and also briefly learn the concept of maximum likelihood estimation
- In EE4146, you will apply Bayes’ rule to classification problems and learn how to estimate the parameters of the classification model using maximum likelihood estimation

Extract from EE4146 notes: Applying Bayes’ formula for classification



Generative vs. Discriminative Models

- In both types of models, the goal is to determine the posterior probability of different classes.
- And then assign x to a class with highest $f(C_k|\boldsymbol{x})$ (See 1.3.6)

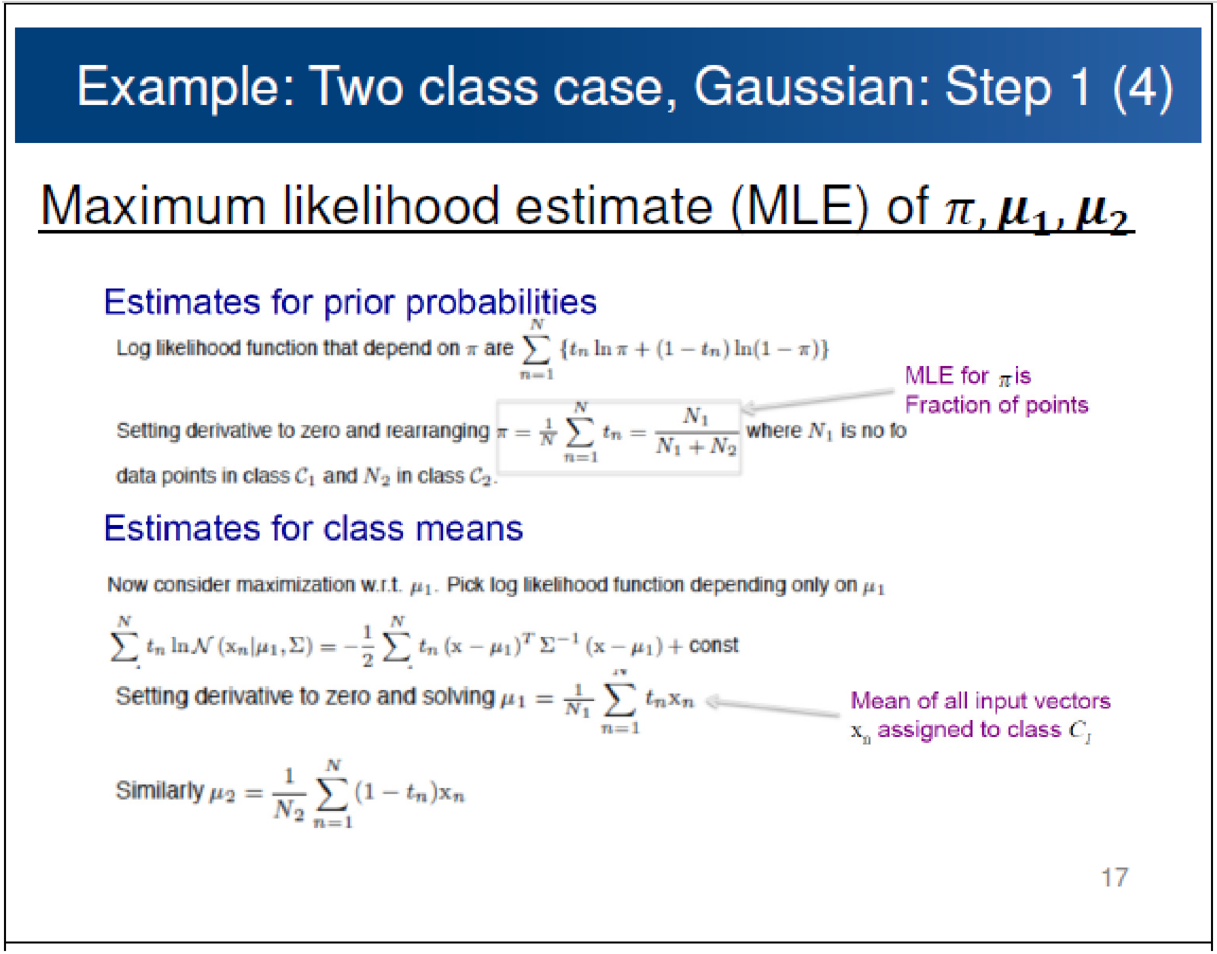
$$f(C_1\,|\,\mathrm{x}) = \frac{f(\mathrm{x}\,|\,C_1)\,f(C_1)}{f(\mathrm{x}\,|\,C_1)\,f(C_1) + f(\mathrm{x}\,|\,C_2)\,f(C_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a) \quad \text{where} \quad a = \ln\frac{f(\mathrm{x}\,|\,C_1)\,f(C_1)}{f(\mathrm{x}\,|\,C_2)\,f(C_2)}$$

- In a generative model we estimate the class-conditionals $f(\boldsymbol{x}|C_k)$ (which are used to determine *a*).
- In the discriminative approach we directly estimate *a* as a linear function of **x** i.e., $a = \mathbf{w}^T\mathbf{x} + b$.

6

Extract from EE4146 notes: maximum likelihood estimation



Example: Two class case, Gaussian: Step 1 (4)

Maximum likelihood estimate (MLE) of $\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$

Estimates for prior probabilities

Log likelihood function that depend on π are $\sum_{n=1}^{N} \{t_n \ln \pi + (1-t_n)\ln(1-\pi)\}$

Setting derivative to zero and rearranging $\boxed{\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N_1+N_2}}$ where N_1 is no fo data points in class C_1 and N_2 in class C_2.

MLE for π is Fraction of points

Estimates for class means

Now consider maximization w.r.t. μ_1. Pick log likelihood function depending only on μ_1

$$\sum_{.}^{N} t_n \ln \mathcal{N}(\mathrm{x}_n|\mu_1, \Sigma) = -\frac{1}{2}\sum_{.}^{N} t_n (\mathrm{x}-\mu_1)^T \Sigma^{-1} (\mathrm{x}-\mu_1) + \text{const}$$

Setting derivative to zero and solving $\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathrm{x}_n$

Mean of all input vectors x_n assigned to class C_j

Similarly $\mu_2 = \frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)\mathrm{x}_n$

17

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