統計計算與模擬

政治大學統計系余清祥

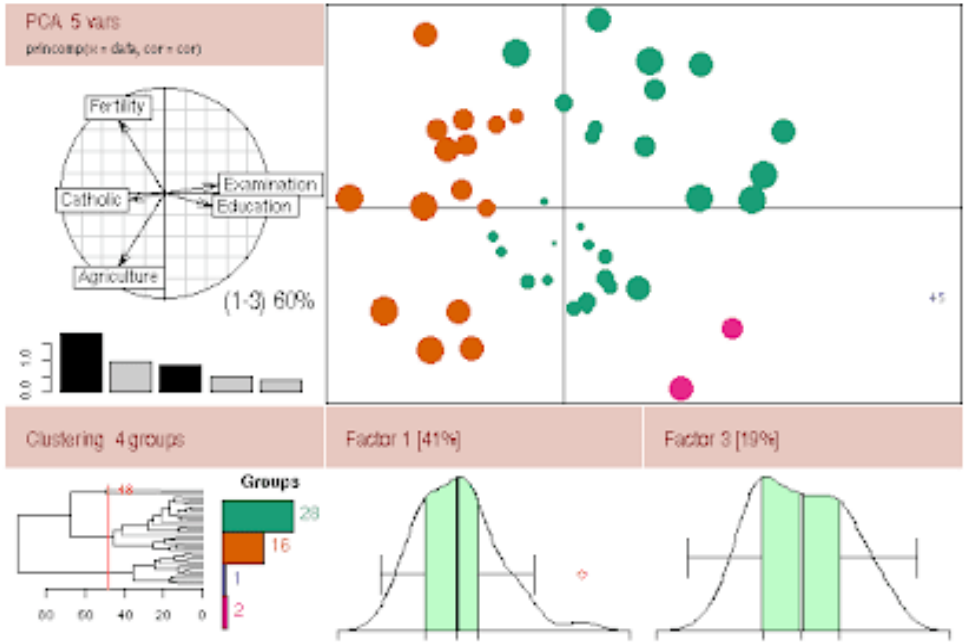
2024年2月20日

第一週：導論

http://csyue.nccu.edu.tw

什麼是「統計計算與模擬」？

- 統計理論與其應用無法自外於統計計算、模擬，尤其在今日巨量資料風行，講究即時分析及判斷的電腦時代。
- 「統計計算與模擬」跨越統計、電腦兩大學門，應用範圍包括各領域，提供專業學門探索新知時不可或缺的工具。



https://sites.google.com/a/alaska.edu/franz-j-mueter/home/teaching?tmpl=%2Fsystem%2Fapp%2Ftemplates%2Fprint%2F&showPrintDialog=1

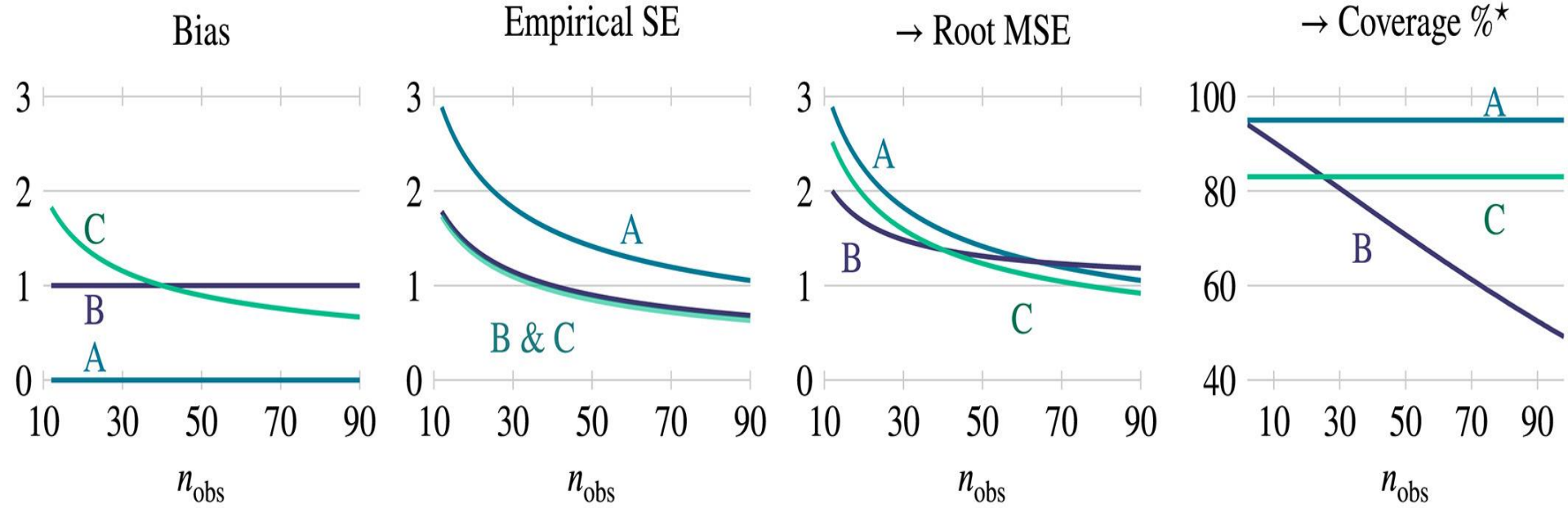
計算模擬的實務考量

- 即使證實理論成立，實際應用時需考慮可行性，包括：時間與金錢成本、市場接受度、穩定性。

→如何產生常態（或特定）分配的亂數，以及如何驗證資料服從常態分配。

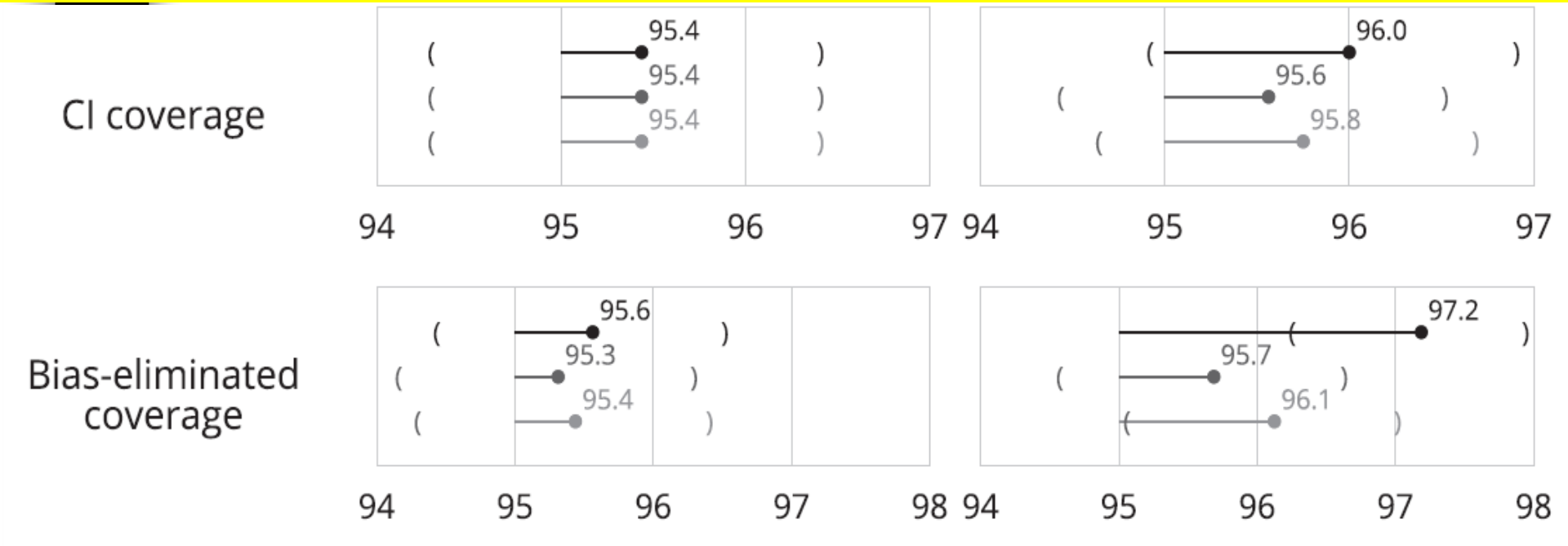
→評估哪一種分析方法較佳，考量執行速度、誤差、敏感度、穩健性(Robustness)。

→探討新工具或想法是否可行，根據測試結果調整執行方法和步驟。



*Coverage calculated for normal-based confidence intervals of correct width so that bias is the source of undercoverage

Using Simulation Studies to Evaluate Statistical Methods (2016) tps://onlinelibrary.wiley.com/doi/10.1002/sim.8086



學術論文與計算模擬

- 美國統計學會期刊(Journal of the American Statistical Association)整理2002年蒙地卡羅及統計方法相關文章，共同特色包括：

→定義問題、過去對此問題的相關研究。

→提出新的研究方法。

→新方法的大樣本理論。

→以蒙地卡羅（電腦模擬）驗證新方法確實具有某些優勢。

註：實證資料未必可得。

計算與模擬的發展歷史

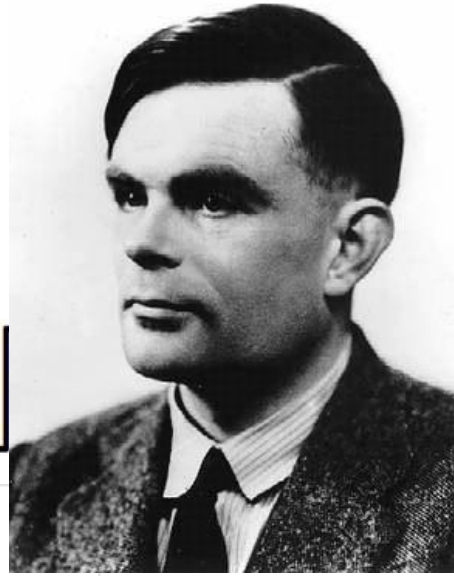
- 電腦在二次大戰後迅速發展，將理論及方法轉化為電腦程式變成顯學，藉以提高人類的生活福祉。

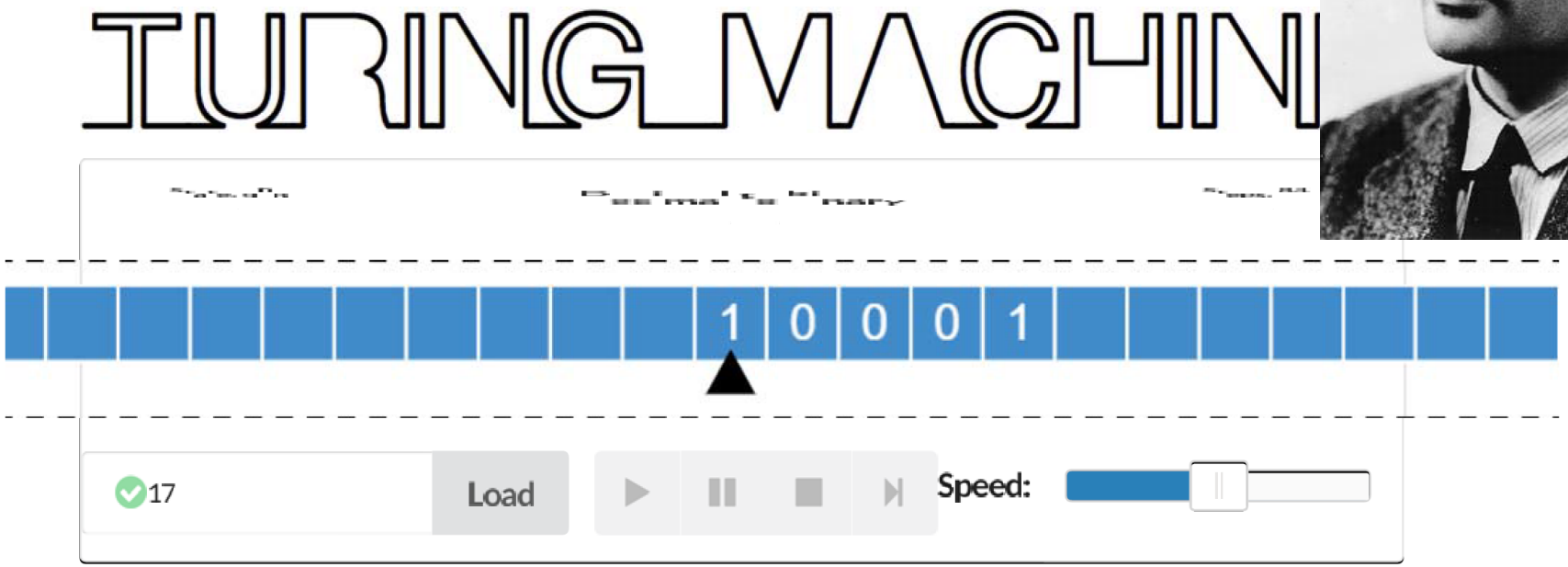
→人工智能之父Alan Turing除了協助盟軍破解密碼，對於人工智能頗多貢獻，對於「機器是否會思考」提出測試方法。

→例如：電腦模擬首次用於二次大戰，在曼哈頓計劃中模擬核爆炸的過程。

https://upload.wikimedia.org/wikipedia/commons/thumb/8/8d/Trinity_shot_color.jpg/800px-Trinity_shot_color.jpg

Turing Machine



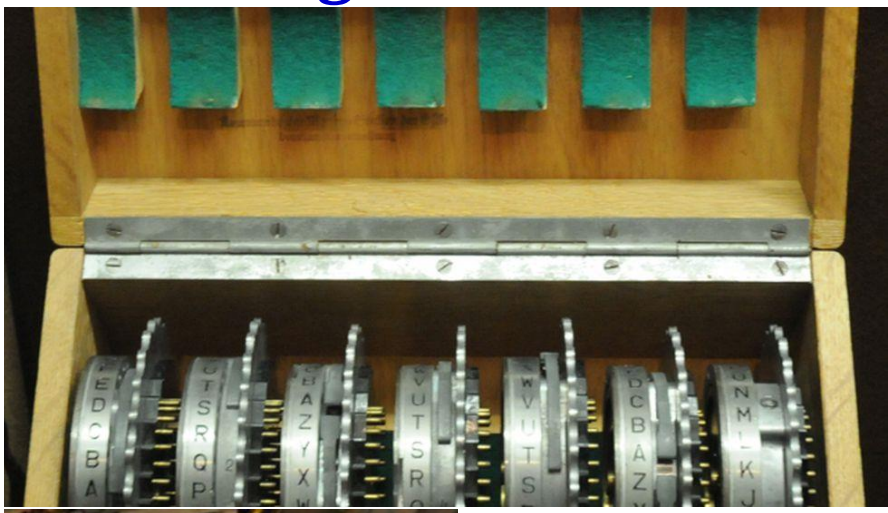


Source: https://turingmachinesimulator.com/

註：誰是Alan Turing、對電腦發展有什麼貢獻？

This is a rebuild of the famous **Colossus Mark 2** machine that finally allowed the code breakers to quickly and efficiently break the high command's ciphers. For decades, since 1918, the Germans had been using Enigma cyphers as the core of their intelligence and military communications system.

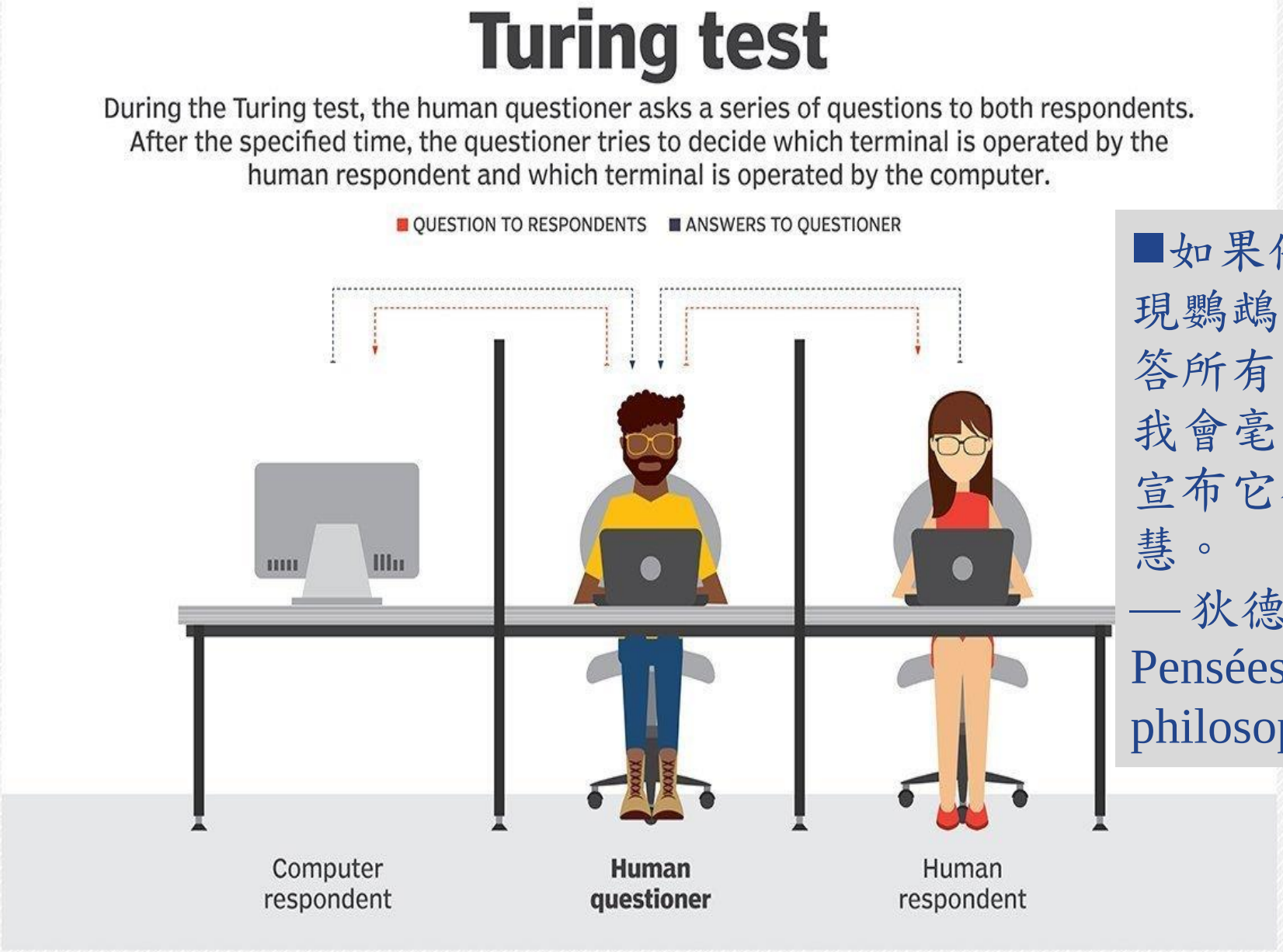
Enigma rotors



https://cnet3.cbsistatic.com/img/f2L_cPTDlz5TH0OwdN5ZK_WX9DI=/980x551/2010/08/05/7e19b914-f0ed-11e2-8c7c-d4ae52e62bcc/Operational_rotors.jpg

https://cnet3.cbsistatic.com/img/YQ0CRrVg9Qz530e7SNDJux7MsN8=/980x551/2011/07/03/7e31eb6c-f0ed-11e2-8c7c-d4ae52e62bcc/Colossus_body_close-up.jpg

機器是否能夠思考的測試？(Chinese Room)



■如果他們發現鸚鵡可以回答所有問題，我會毫不猶豫宣布它存在智慧。
— 狄德羅，Pensées philosophiques

Source: https://cdn.ttgtmedia.com/rms/onlineImages/crm-turing_test.jpg

「統計計算與模擬」的範圍

- 最大概似估計法中函數的最大值求解，一般可用牛頓法求根。
- 迴歸分析中以最小平方法求得參數估計值，估計多藉由矩陣運算、反矩陣求得。
- 只有一組觀察值無法獲得變異數時，可藉由重覆抽樣（如拔靴法）的電腦模擬方法求取近似值。
- 驗證大樣本或某個統計理論。
- 由觀察值估計機率密度函數，可能方法包括核估計法。

統計計算與模擬的幾個例子

- 將統計理念訴諸於實證分析，尤其在資料龐雜時，計算、模擬不可或缺。

→估計某個函數的根或及值(e.g., MLE)

→由巨量資料中找出迴歸方程式的參數估計值，或是最佳參數組合(e.g., stepwise)

→信賴區間的實質詮釋(e.g. 95%信賴區間更具說服力的說明方式）

→如何從資料與模擬估計某參數，或是由模擬確認某個分配(e.g. F分配的定義)

- 範例一、求得某個函數的解，包括聯立方程式、矩陣、MLE、極值(Optimum)等。

→欲求下列方程式的根：(uniroot)

$$f(x) = e - \frac{1}{3.5 + x}$$

$$f(x) = \frac{e^{-x}}{\sqrt{1 + x^2}} - 0.5$$

→求解反矩陣及向量解：

$$A = \begin{bmatrix} 1 & 0.5 & 0.25 & 0.125 \\ 0.5 & 1 & 0.5 & 0.25 \\ 0.25 & 0.5 & 1 & 0.5 \\ 0.125 & 0.25 & 0.5 & 1 \end{bmatrix}$$

- 範例二、迴歸分析的參數估計、及其變異數，以及逐步迴歸等分析。

→如何將正規方程式具體化？

$$Y = X\beta + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2 I_{n\times n})$$

$$\Rightarrow (X'X)\hat{\beta} = X'Y \quad (\textit{Normal equation})$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1}X'Y \equiv AY.$$

註：除了求出反矩陣，是否有其他可能？

→求解下列迴歸方程式：

X	585	1002	472	493	408	690	291
Y	0.1	0.2	0.5	1.0	1.5	2.0	3.0

- 範例三、中央極限定理(CLT)、95%信賴區間的示範及詮釋。

→CLT的假設條件是否絕對必要，到底要多少樣本才能近似常態分配？

註：期望值不存在的變數也適用嗎？*iid*假設的必要性、極端左右偏分配或是離散分配需要較多觀察值？

→重複模擬亂數，再檢驗95%等信賴區間的結果，是否符合95%的涵蓋率。

- 範例四、除了理論推導，使用電腦模擬檢驗大樣本理論等結果。

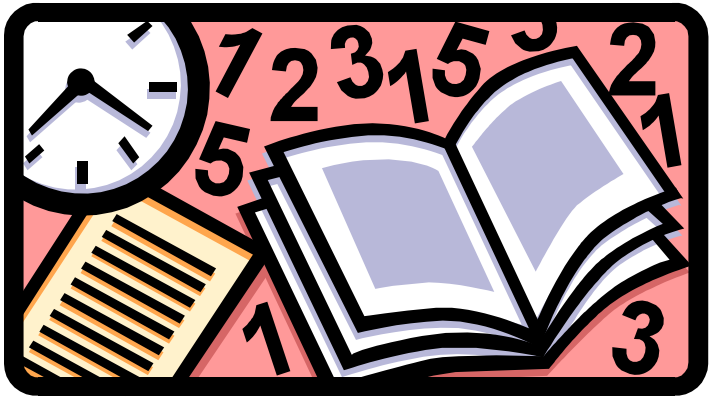
→如何產生服從某個分配函數的亂數（e.g., 常態），或是驗證由已知分配產生新的分配函數（e.g., Cauchy 分配）。

註：相關問題為如何檢查觀察值服從某個分配函數（除了卡方檢定外）。

→電腦模擬也可用於估計某些定值，像是圓周率、自然對數等，透過模擬積分(Simulation Integration) 的結果往往優於數值積分(Numerical Integration)。

本課程將介紹的單元：

- *Simulation and Monte Carlo methods*
- *Optimization methods*
- *Data partition and resampling*
- *Variance reduction*
- *Density estimation*
- *Bayesian computing (MCMC)*



統計學家的價值

- 統計學家注重實驗之邏輯性。

→統計學家承傳科學研究方法，所關心、所寫、所做皆是關於科學研究。(Hooke, 1980)

→統計學家承擔研究發明在邏輯方面的重擔，必須認同此一角色，且做好面對看不出資訊的原始資料的準備。(Price, 1982)

- 思維練習與測試：如何以統計角度說明圓周率是無理數。

→什麼是無理數、如何以統計驗證？

R（S-Plus為商業軟體）

- 本課程將使用 R軟體，R 為免費公用軟體，請至以下網址下載：

  http://www.r-project.org/

  http://cran.r-project.org/

- R的使用介紹可在網路上搜尋，包括中華R軟體學會：

  http://www.r-software.org/

  這個學會也提供教學錄影帶：

https://sites.google.com/site/zhonghuarruantixuehui/movielist

註：R軟體本身也提供基本的使用指南。

R、計算模擬與網路學習

- 除了傳統的紙本及電子書，建議同學也上網參考網路資訊、互動學習資源。

→R 的使用者網頁查詢（如：臉書）。

→參考書籍：《R軟體：應用統計方法》
《 R語言數據操作》

- 本課程網站也有部分使用講義，請同學到 http://csyue.nccu.edu.tw 下載。
- 除非課程需要，建議同學以大眾常用的程式為原則，盡量不自行發展撰寫。

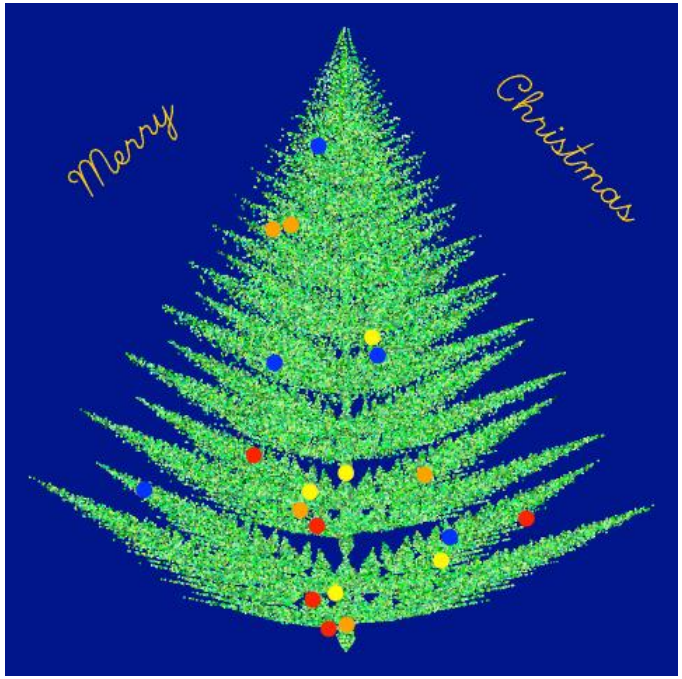
Taiwan R User Group

- 臉書上也能找到學習工具，像是「Taiwan R User Group」，除了分享使用心得，也提供特定主題的課程：

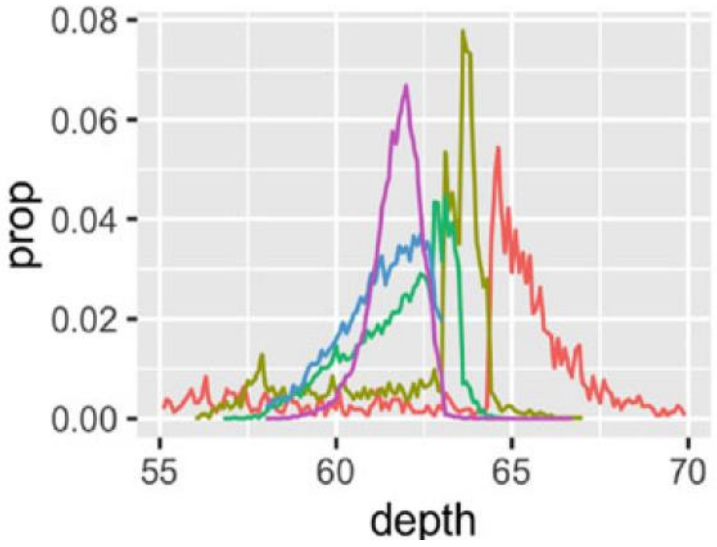
  https://www.facebook.com/Tw.R.User

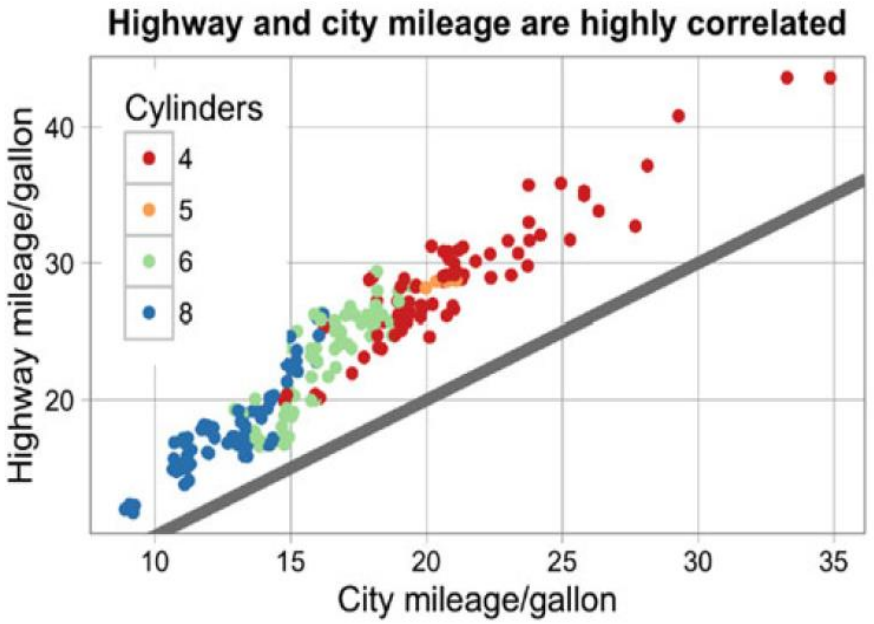
→右圖為這個使用者群組仿製碎形(Fractal)繪製的耶誕樹！

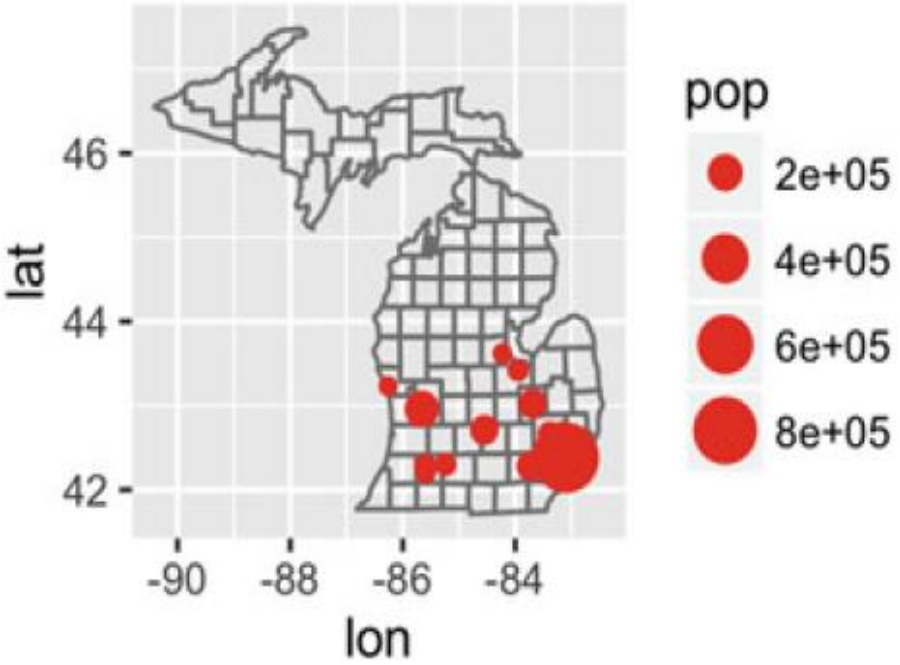
http://wiekvoet.blogspot.tw/2014/12/merry-christmas.html

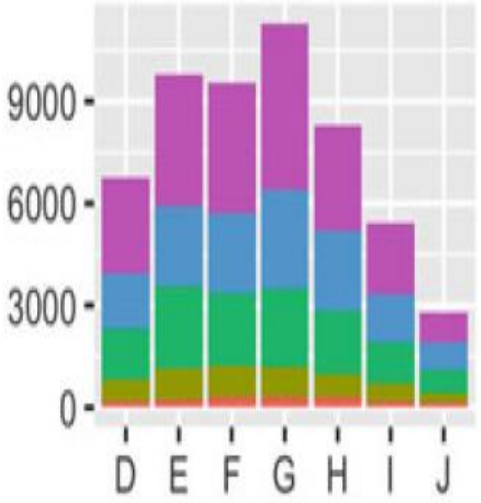


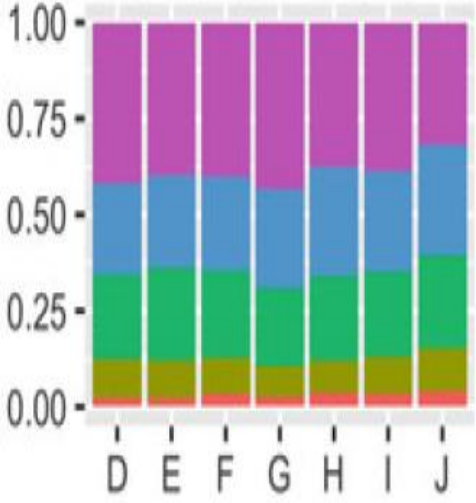
R的繪圖功能非常強大（ggplot2）











充分完備的統計學家

- 統計學家（&資料科學家）所需的專業技能，可視為下列「溝通」能力：

→與人溝通：寫作、口語表達、溝通能力；

→與資料（及統計理論）溝通：data sense、資訊圖像化、特性與趨勢；

→與專業溝通：領域知識、問題定義及結果詮釋、附加價值；

→與電腦（機器）溝通：資料儲存與更新、資訊安全、程式運算。

充分完備的統計學家(續)

- W.G.V. Balchin (The American Cartographer) illustrated in 1976 that humans evolved by first developing keen **visual spatial** skills, then **social** skills, **verbal** skills, and **numerical** skills.

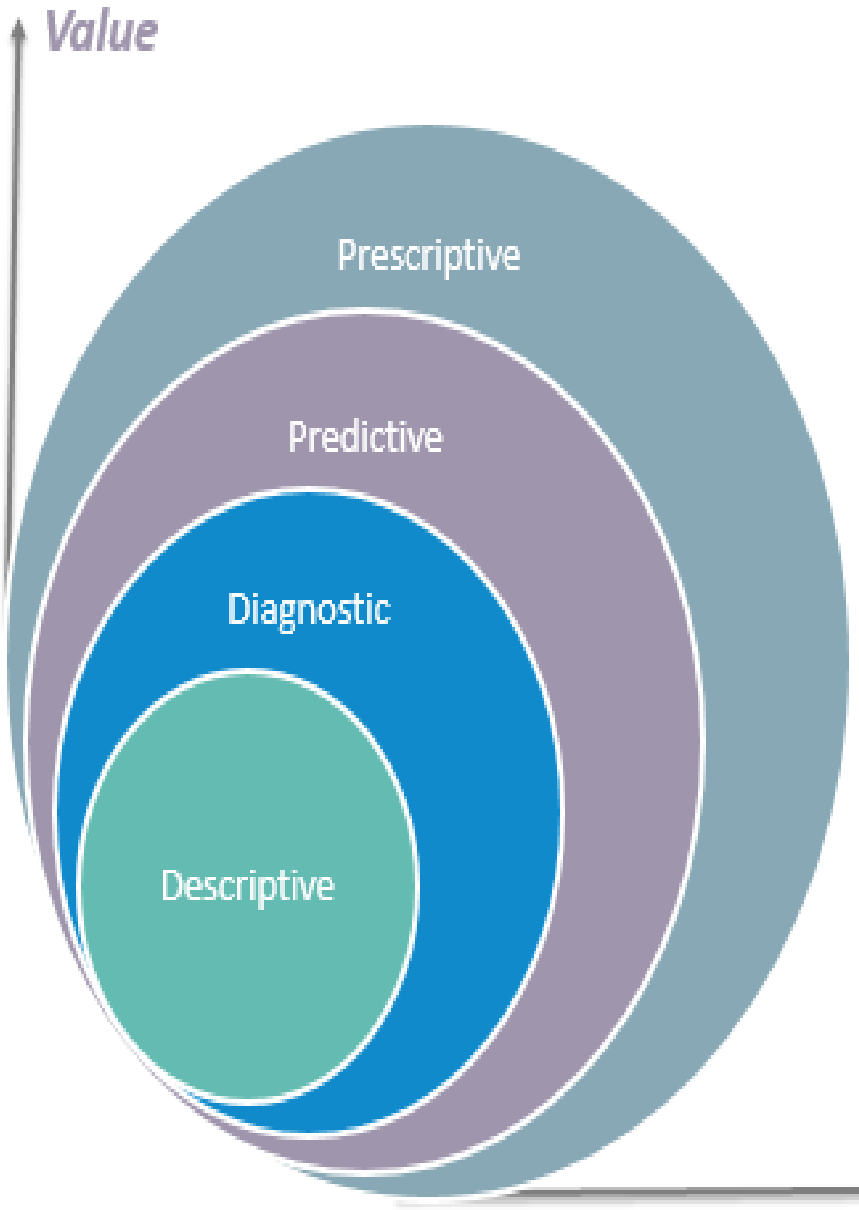
→ Numeracy — formulating & solving problems using mathematics and computing

→Articulacy — speaking & listening (people skills)

→Literacy — writing & reading

→Graphicacy — producing & understanding graphics

4 types of Data Analytics



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa

www.principa.co.za

http://www.kdnuggets.com/2017/07/4-types-data-analytics.html

EDA and CDA

- EDA (or **Descriptive analytics**) tells us what happened up from the data.
- CDA can be separated into two parts:

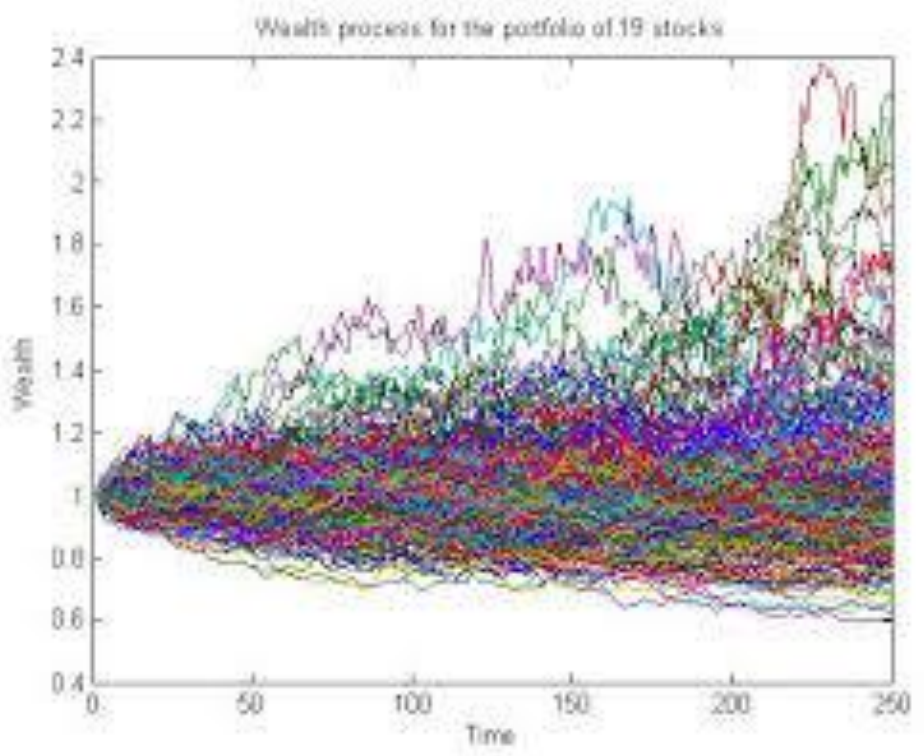
→**Predictive analytics** give us clues about the future, given data and domain knowledge.

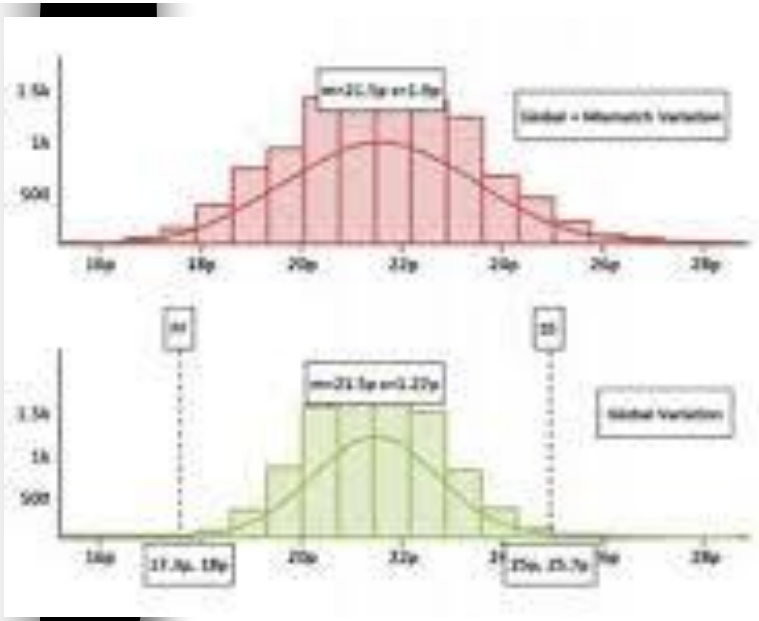
→**Prescriptive analytics** provide suggestions for optimizing the future.

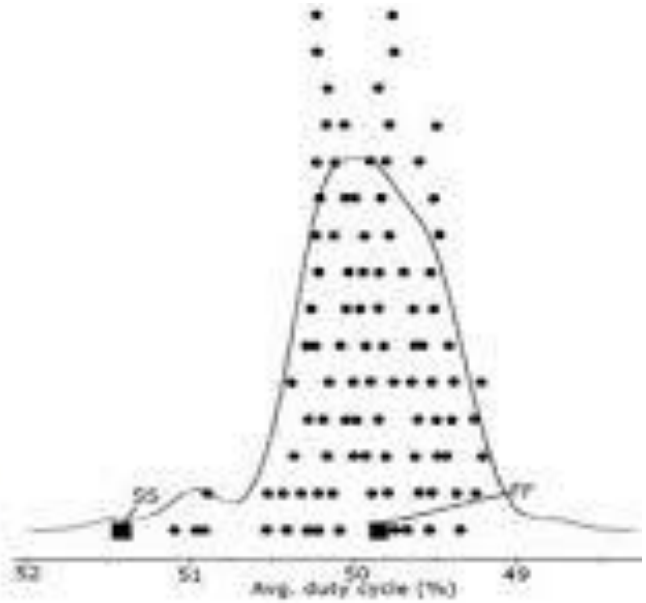
1. **Descriptive**. Traditional HR metrics are largely efficiency metrics (turnover rate, time to fill, cost of hire, number hired and trained, etc.). The primary focus here is on cost reduction and process improvement. Descriptive HR analytics reveal and describe *relationships* and *current and historical data patterns*. This is the foundation of your analytics effort. It includes, for example, dashboards and scorecards; workforce segmentation; data mining for basic patterns; and periodic reports.
2. **Predictive**. Predictive analysis covers a variety of techniques (statistics, modeling, data mining) that use current and historical facts to make predictions about the future. It's about probabilities and potential impact. It involves, for example, models used for increasing the probability of selecting the right people to hire, train, and promote.
3. **Prescriptive**. Prescriptive analytics goes beyond predictions and outlines decision options and workforce optimization. It is used to analyze complex data to predict outcomes, provide decision options, and show alternative business impacts. It involves, for example, models used for understanding how alternative learning investments impact the bottom line (rare in HR).

祝大家有個

豐收的學期！


Wealth process for the portfolio of 19 stocks
Wealth
Time
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0
50
100
150
200
250




55
52
51
50
49
Avg. duty cycle (%)