國立政治大學
國際金融學院
College of Global Banking and Finance, National Chengchi University

國立政治大學
企業永續管理研究中心
Center for Business Sustainability, NCCU

# TCFD報告書 Text Mining評鑑內容探討

政治大學統計系余清祥

2025年12月5日

Email: csyue@nccu.edu.tw

http://csyue.nccu.edu.tw

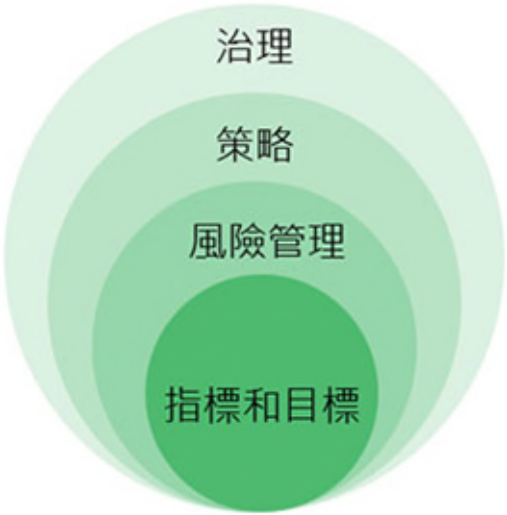
# 研究目的

- 分析氣候相關財務揭露（TCFD）報告書，以數位工具協助人工評鑑作業。

→註記及查詢「治理」、「策略」、「風險管理」、「指標和目標」的相關文字。

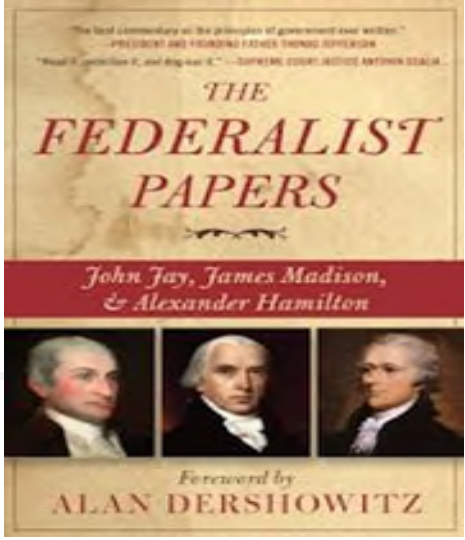
→協助專家判讀TCFD報告、判斷哪些文字為電腦生成(或抄襲)。

→提供文字分析模型及方法，建立量化評鑑機制。




https://www.eman-tw.org.tw/esg-02

https://www.skyhorsepublishing.com/9781631585272/the-federalist-papers/




# 英文分析歷史悠久

- 莎士比亞(Shakespeare)的字彙總數

→ Efron and Thisted (1976) 以Poisson Process 估計莎士比亞字彙，推論1985年在莎翁故居附近發現一首詩應是莎翁所作。

- 《聯邦主義議文集》（The Federal Papers）

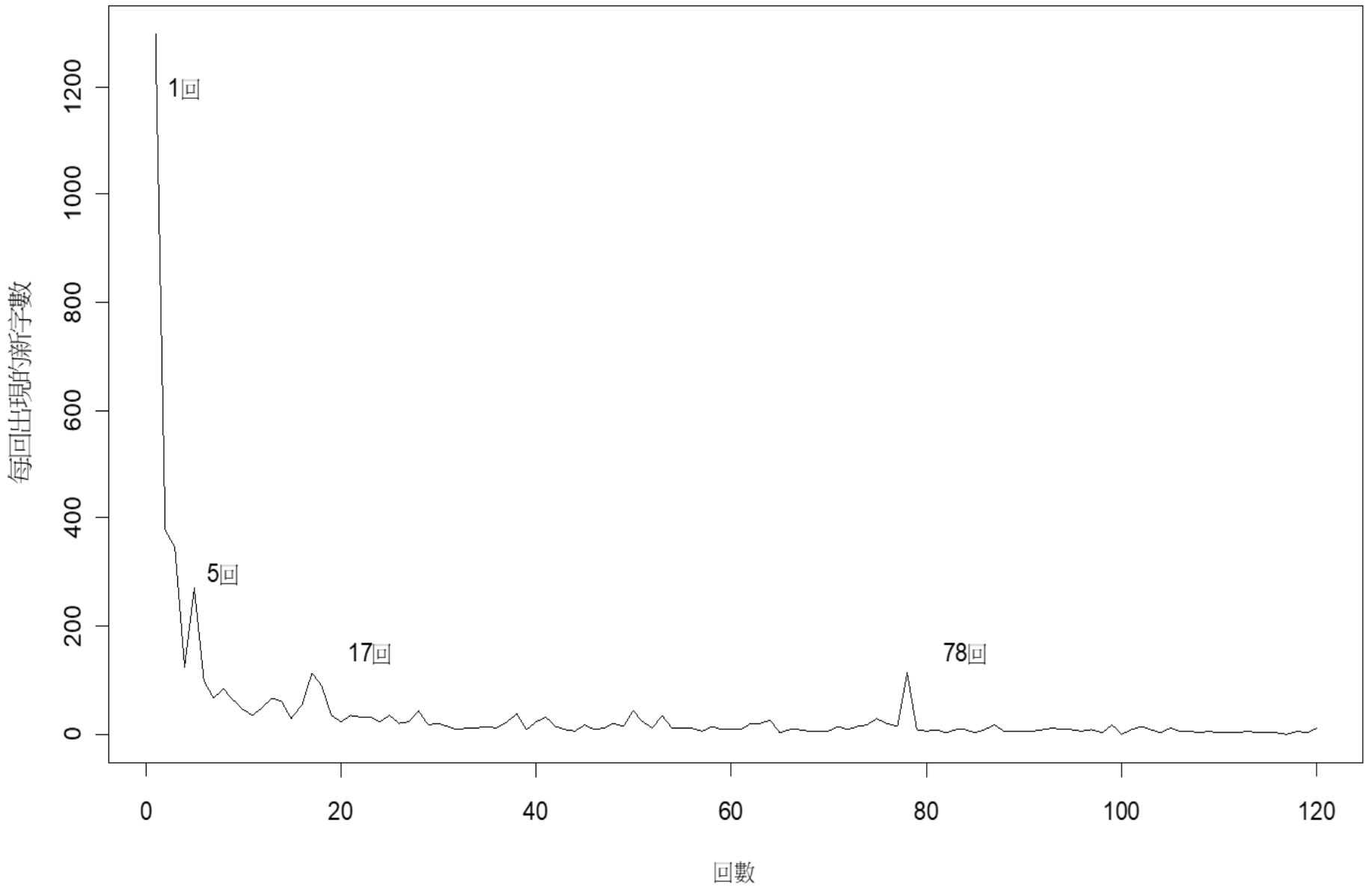
→Mosteller and Wallace (1984)以貝氏方法分析「upon」之類虛字，判斷論文集中未署名的12篇文章應是Madison、而非Hamilton。

# 紅樓夢代入Efron & Thisted估計新字彙

## 前八十回扣除第七十八回之8-fold cross–validation

| Testing Data | 比例 | 預測值 | Variance of Est. | UC.I. | LC.I. | 新字 | 是否在95%C.I. |
|---|---|---|---|---|---|---|---|
| 1~10 | 10/69 | 94.72 | 64.45 | 221.04 | 0 | 157 | Y |
| 11~20 | 10/69 | 96.30 | 64.50 | 222.71 | 0 | 151 | Y |
| 21~30 | 10/69 | 95.03 | 64.92 | 222.26 | 0 | 97 | Y |
| 31~40 | 10/69 | 95.21 | 65.08 | 222.78 | 0 | 75 | Y |
| 41~50 | 10/69 | 91.30 | 64.80 | 218.30 | 0 | 112 | Y |
| 51~60 | 10/69 | 97.11 | 64.94 | 224.39 | 0 | 94 | Y |
| 61~70 | 10/69 | 97.20 | 65.00 | 224.60 | 0 | 86 | Y |
| 71~80 | 9/70 | 81.98 | 64.63 | 208.65 | 0 | 134 | Y |

# 《紅樓夢》各回新出現的字彙



# 紅樓夢代入E&T估計新字彙

## 後四十回視為同一母體，4-fold cross–validation

| Testing Data | 比例 | 預測值 | Variance of Est. | UC.I. | LC.I. | 新字 | 是否在95%C.I. |
|---|---|---|---|---|---|---|---|
| **81~90** | 1/3 | **172.47** | 55.00 | 280.27 | 64.665 | **261** | **Y** |
| **91~100** | 1/3 | **170.36** | 55.38 | 278.91 | 61.817 | **219** | **Y** |
| **101~110** | 1/3 | **175.28** | 55.33 | 283.72 | 66.838 | **225** | **Y** |
| **111~120** | 1/3 | **182.97** | 55.68 | 292.10 | 73.843 | **186** | **Y** |

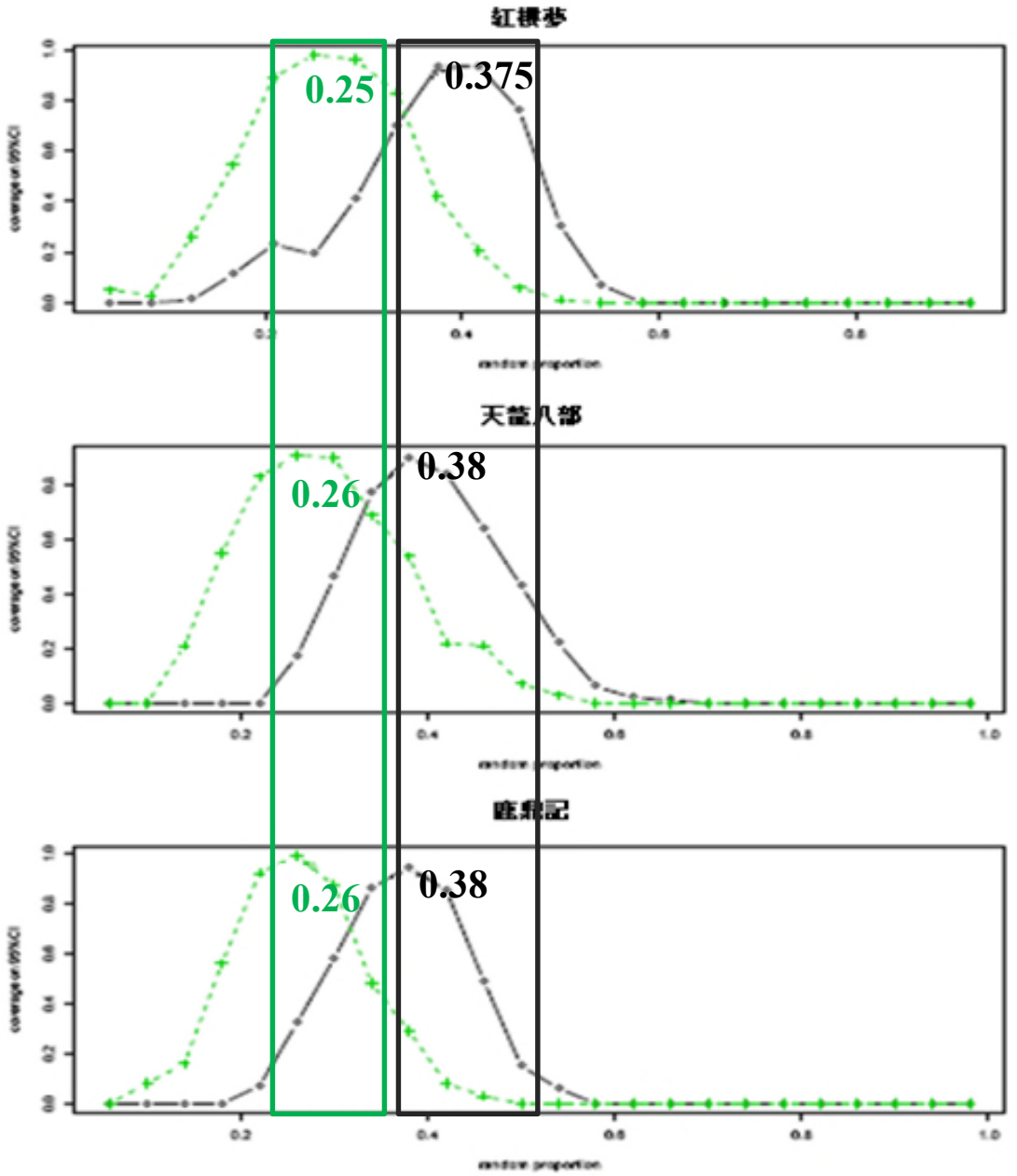
## 紅樓夢前八十回預測後四十回

| Testing Data | 比例 | 預測值 | Variance of Est. | UC.I. | LC.I. | 新字 | 是否在95%C.I. |
|---|---|---|---|---|---|---|---|
| **81~120** | 1/2 | 294.87 | 66.49 | 425.19 | 164.547 | **231** | **Y** |

# 紅樓夢與金庸小說的Jackknife覆蓋機率

＊Magic Number！

→抽出37.5%(25%)即可知道整部小說使用多少種字彙！



# TCFD報告評分模型

http://www.tnooz.com/2012/01/04/how-to/big-data-and-the-infinite-possibilities-for-the-travel-industry/

# 量化模型

透過量化模型描述觀察結果：

**觀察現象＝模型＋誤差**

或是

***y* = *f*(*x*) + error；觀察值＝訊號＋雜訊**

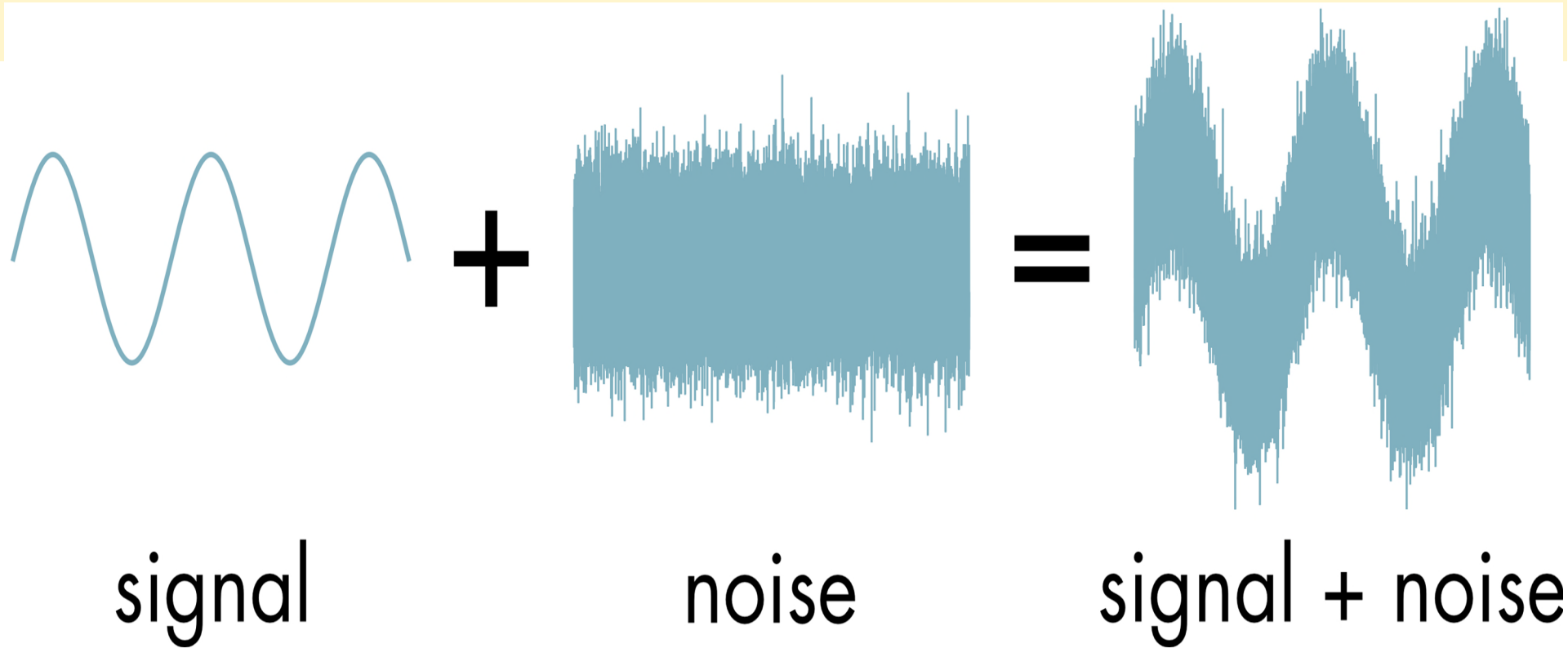
- 數量化模型的關鍵：

→量化目標值$y$：定義問題！

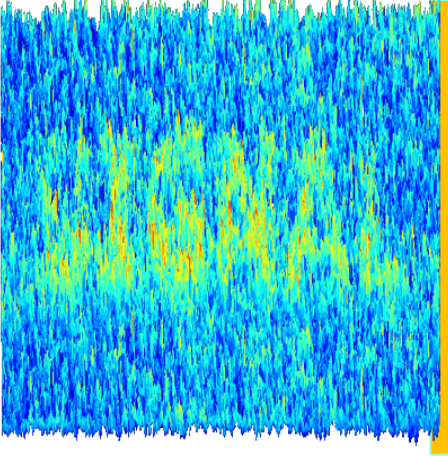
→選取關鍵變數：$x_1, x_2, \ldots, x_p$

→建立量化模型：統計學習、機器學習。

# 如何分辨訊號、雜訊？




https://neurotext.library.stonybrook.edu/C7/C7_1/C7_1.html



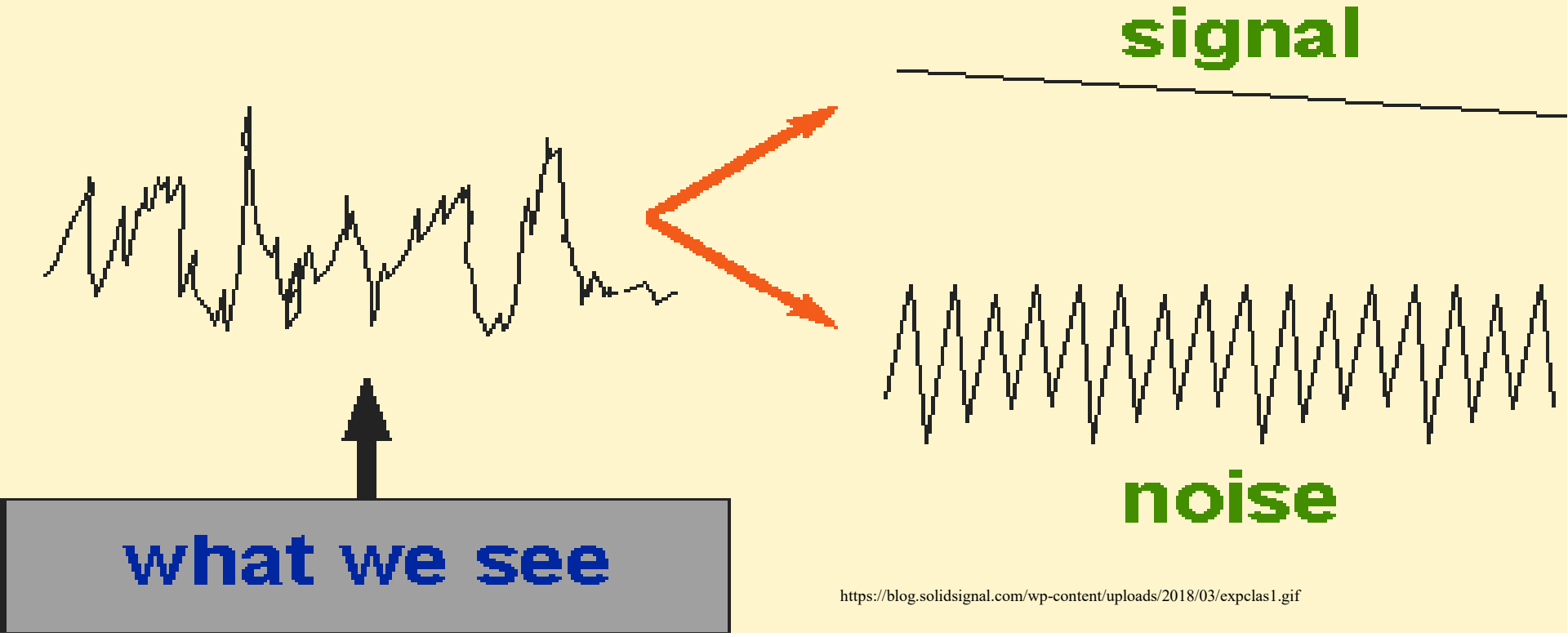
# 觀察現象＝模型＋誤差

# Observation = Signal + Noise


http://www.ljagilamplighter.com/wp-content/uploads/noise.png

http://www.ljagilamplighter.com/wp-content/uploads/noise_signal-mlab1_png_pagespeed_ce_b_GTiE6tAg.png

## What we observe can be divided into:




https://blog.solidsignal.com/wp-content/uploads/2018/03/expclas1.gif

# 如何分析中文？

- 文字屬於非結構(Unstructure)資料，視文字特性、研究目標決定變數。

→中文書寫現在以白話文為主，通常會透過雙字(多字)詞彙表達語意，因此斷詞一般是中文資料分析的關鍵步驟。

→常見的斷詞工具包括中研院研發的CKIP，以及網路常見的jieba (結巴)。

# 斷詞與統計分析

- 根據研究目的及問題定義，挑選適當變數以提高分析的效率和準確性。

→統計量化模型注重**實質解釋**。

- 中文分析先經過斷詞、再挑選重要關鍵詞。

→白話文多以雙字詞（多字詞）表達觀念。

（註：**字**→**詞**→**有意義的詞**）

*我想過過過兒過過的生活*

*用毒毒毒蛇毒蛇會不會被毒毒死*

# 方塊字分析與生物多樣性

大致可由以下三個角度切入：

- 豐富度(Richness)及不均度(Unevenness)：
  - → 字/詞數與種類、相異字比例(TTR)
  - → 常見字詞佔比、Entropy、Simpson Index
- 字詞關聯性(生態系Ecosystem)
  - → 關鍵詞叢、主題模型(Topic Model)
- 趨勢變化(字詞比擬為物種Species)
  - → 共同字彙、新生及滅絕字彙

# 現行分析作法

- 先以字詞及文句結構切入，暫時沒有加入字詞的意義、詞性、關聯性等資訊。

→字詞的TTR、Entropy、Simpson Index

- 此外，也加入各公司的相關資訊：

→資本額、官股、溫室氣體驗證

註：2023～2024年不細分四要素，文字分析以整篇報告為單位。

附表A-1、文字及公司變數

| 序號 | 本國銀行變數名稱 |
|---|---|
| 1 | 銀行狀態 |
| 2 | 有金控母公司 |
| 3 | 銀行上市櫃狀態 |
| 4 | 公股 |
| 5 | 資本額 |
| 6 | TCFD_會計師確信 |
| 7 | TCFD_BSI查核 |
| 8 | 溫室氣體驗證 |
| 9 | 有委任第三方驗證 |
| 10 | 總字數 |
| 11 | 不同字數 |
| 12 | 隨機抽樣500字之平均TTR |
| 13 | 隨機抽樣1000字之平均TTR |
| 14 | 隨機抽樣2000字之平均TTR |
| 15 | 隨機抽樣3000字之平均TTR |
| 16 | 隨機抽樣4000字之平均TTR |
| 17 | 隨機抽樣5000字之平均TTR |
| 18 | 隨機抽樣500字之平均Entropy |
| 19 | 隨機抽樣1000字之平均Entropy |
| 20 | 隨機抽樣2000字之平均Entropy |
| 21 | 隨機抽樣3000字之平均Entropy |
| 22 | 隨機抽樣4000字之平均Entropy |
| 23 | 隨機抽樣5000字之平均Entropy |
| 24 | 隨機抽樣500字之平均Simpson |
| 25 | 隨機抽樣1000字之平均Simpson |
| 26 | 隨機抽樣2000字之平均Simpson |
| 27 | 隨機抽樣3000字之平均Simpson |
| 28 | 隨機抽樣4000字之平均Simpson |
| 29 | 隨機抽樣5000字之平均Simpson |
| 30 | 總詞數 |
| 31 | 不同詞數 |
| 32 | 隨機抽樣500詞之平均TTR |
| 33 | 隨機抽樣1000詞之平均TTR |
| 34 | 隨機抽樣2000詞之平均TTR |
| 35 | 隨機抽樣3000詞之平均TTR |
| 36 | 隨機抽樣4000詞之平均TTR |
| 37 | 隨機抽樣5000詞之平均TTR |
| 38 | 隨機抽樣500詞之平均Entropy |
| 39 | 隨機抽樣1000詞之平均Entropy |
| 40 | 隨機抽樣2000詞之平均Entropy |
| 41 | 隨機抽樣3000詞之平均Entropy |
| 42 | 隨機抽樣4000詞之平均Entropy |
| 43 | 隨機抽樣5000詞之平均Entropy |
| 44 | 隨機抽樣500詞之平均Simpson |
| 45 | 隨機抽樣1000詞之平均Simpson |
| 46 | 隨機抽樣2000詞之平均Simpson |
| 47 | 隨機抽樣3000詞之平均Simpson |
| 48 | 隨機抽樣4000詞之平均Simpson |
| 49 | 隨機抽樣5000詞之平均Simpson |
| 50 | 全部詞彙詞向量 |
| 51 | 全部詞彙詞向量_主成分分析 |
| 52 | 前500大詞彙詞向量 |
| 53 | 前100大詞彙詞向量 |
| 54 | 前10大詞彙詞向量 |
| 55 | 文字源於TCFD報告書 |
| 56 | 文字源於ESG報告書 |

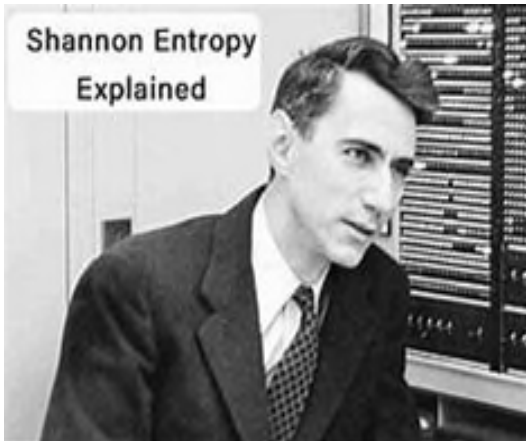
| 序號 | 壽險、產險變數名稱 |
|---|---|
| 1 | 有金控母公司 |
| 2 | 金控母公司是公股 |
| 3 | 資本額 |
| 4 | TCFD_會計師確信 |
| 5 | TCFD_BSI查核 |
| 6 | 溫室氣體驗證 |
| 7 | 有委任第三方驗證 |
| 8 | 總字數 |
| 9 | 不同字數 |
| 10 | 隨機抽樣500字之平均TTR |
| 11 | 隨機抽樣1000字之平均TTR |
| 12 | 隨機抽樣2000字之平均TTR |
| 13 | 隨機抽樣3000字之平均TTR |
| 14 | 隨機抽樣4000字之平均TTR |
| 15 | 隨機抽樣5000字之平均TTR |
| 16 | 隨機抽樣500字之平均Entropy |
| 17 | 隨機抽樣1000字之平均Entropy |
| 18 | 隨機抽樣2000字之平均Entropy |
| 19 | 隨機抽樣3000字之平均Entropy |
| 20 | 隨機抽樣4000字之平均Entropy |
| 21 | 隨機抽樣5000字之平均Entropy |
| 22 | 隨機抽樣500字之平均Simpson |
| 23 | 隨機抽樣1000字之平均Simpson |
| 24 | 隨機抽樣2000字之平均Simpson |
| 25 | 隨機抽樣3000字之平均Simpson |
| 26 | 隨機抽樣4000字之平均Simpson |
| 27 | 隨機抽樣5000字之平均Simpson |
| 28 | 總詞數 |
| 29 | 不同詞數 |
| 30 | 隨機抽樣500詞之平均TTR |
| 31 | 隨機抽樣1000詞之平均TTR |
| 32 | 隨機抽樣2000詞之平均TTR |
| 33 | 隨機抽樣3000詞之平均TTR |
| 34 | 隨機抽樣4000詞之平均TTR |
| 35 | 隨機抽樣5000詞之平均TTR |
| 36 | 隨機抽樣500詞之平均Entropy |
| 37 | 隨機抽樣1000詞之平均Entropy |
| 38 | 隨機抽樣2000詞之平均Entropy |
| 39 | 隨機抽樣3000詞之平均Entropy |
| 40 | 隨機抽樣4000詞之平均Entropy |
| 41 | 隨機抽樣5000詞之平均Entropy |
| 42 | 隨機抽樣500詞之平均Simpson |
| 43 | 隨機抽樣1000詞之平均Simpson |
| 44 | 隨機抽樣2000詞之平均Simpson |
| 45 | 隨機抽樣3000詞之平均Simpson |
| 46 | 隨機抽樣4000詞之平均Simpson |
| 47 | 隨機抽樣5000詞之平均Simpson |
| 48 | 全部詞彙詞向量 |
| 49 | 全部詞彙詞向量_主成分分析 |
| 50 | 前500大詞彙詞向量 |
| 51 | 前100大詞彙詞向量 |
| 52 | 前10大詞彙詞向量 |
| 53 | 文字源於TCFD報告書 |
| 54 | 文字源於ESG報告書 |

# 生物多樣性

- 寫作風格如同生態系各有特色，字詞可比擬為物種(Species)。

→物種豐富度及不均度可描述字詞特性，例如：相異字比(TTR, Type-Token-Ratio)、熵(Entropy, 或稱Shannon Index)。

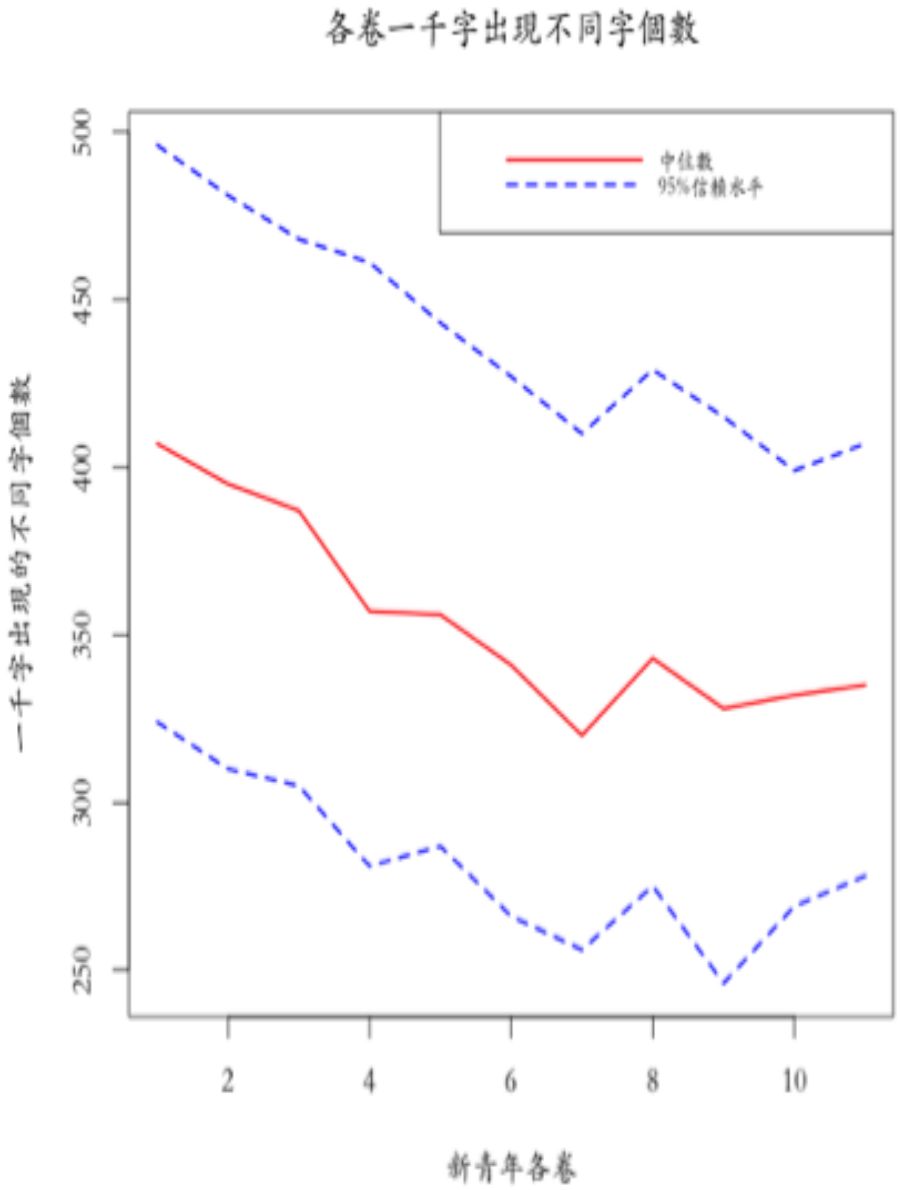
$$Entropy = -\sum_{i=1}^{W} p(w_i) \log_2 p\ (w_i)$$

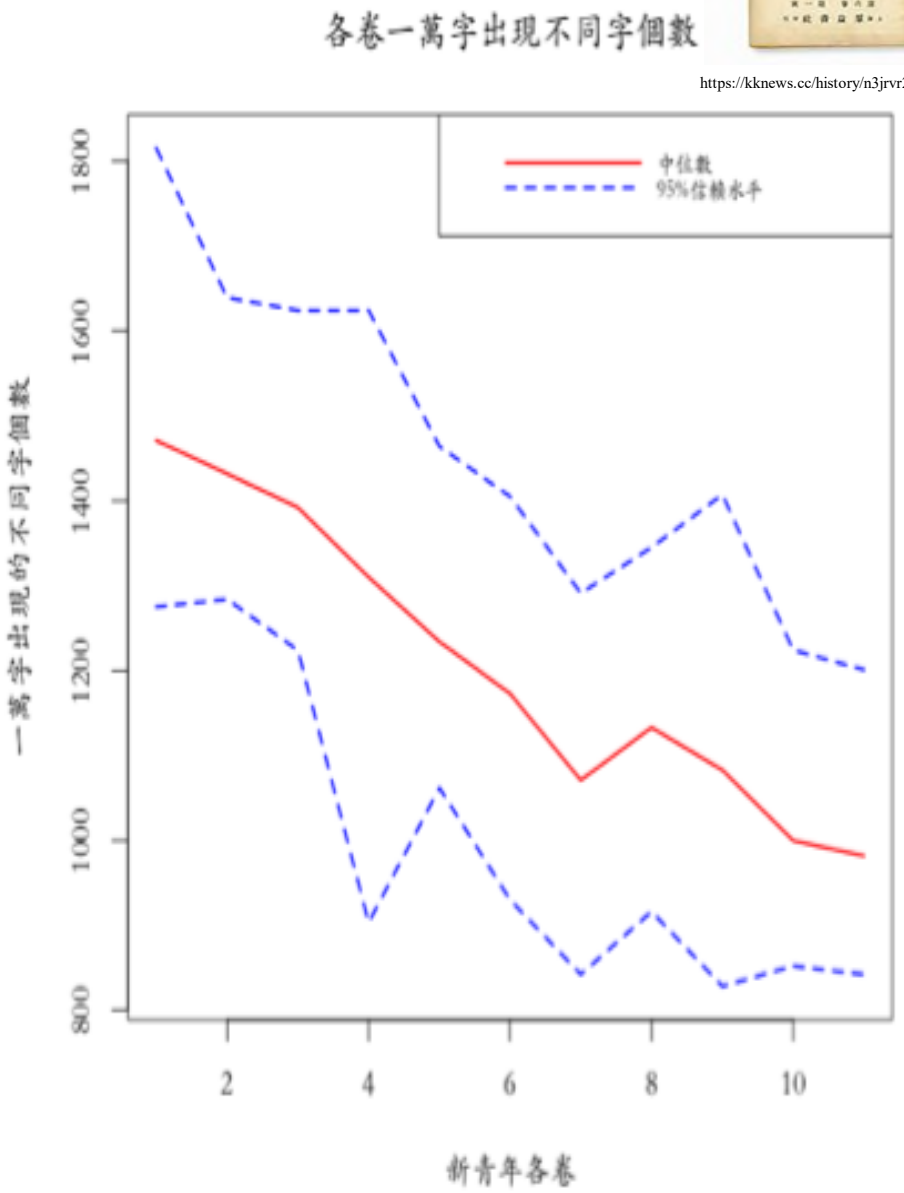



https://atlantisschoolofcommunication.org/communications-foundations/thinkers/claude-shannon/shannon-entropy-explained/

# 《新青年》各卷字彙豐富度(TTR)


https://kknews.cc/history/n3jrvr2.html



# TCFD文字與評分資料

https://www.logicraysacademy.com/blog/wp-content/uploads/2023/08/DA2.jpg

# 氣候相關財務揭露資料統計

| | 年度 | 銀行 | 壽險 | 產險 |
|---|---|---|---|---|
| TCFD報告書 | 2023-24 | 29 | 14 | 11 |
| 永續報告書 | 2023-24 | 9 | 7 | 12 |
| 總計 | 2023-24 | 38 | 21 | 23 |
| 字總數 | 2023 | 263,590 | 117,889 | 57,056 |
| | 2024 | 420,006 | 147,334 | 68,272 |
| 每家平均字數 | 2023 | 6,937 | 5,614 | 3,003 |
| | 2023 | 11,053 | 7,016 | 3,593 |

# 2024年TCFD報告四要素文字統計

| 指標 | 本國銀行業 | | | | 壽險業 | | | | 產險業 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 治理 | 策略 | 風險 | 指標 | 治理 | 策略 | 風險 | 指標 | 治理 | 策略 | 風險 | 指標 |
| 字總數 | 47201 | 161076 | 121821 | 89908 | 14637 | 53545 | 37200 | 41952 | 9220 | 25369 | 17469 | 16214 |
| 每家平均 | 1242 | 4239 | 3206 | 2366 | 697 | 2550 | 1771 | 1998 | 485 | 1335 | 919 | 853 |
| 字種類 | 1105 | 1732 | 1472 | 1442 | 743 | 1310 | 1156 | 1272 | 700 | 1083 | 917 | 958 |
| 詞總數 | 23891 | 80790 | 61415 | 44001 | 7442 | 26557 | 18468 | 20689 | 4609 | 12617 | 8975 | 7827 |
| 詞種類 | 3700 | 10899 | 7981 | 7727 | 1572 | 5309 | 3829 | 4750 | 1244 | 2919 | 2261 | 2280 |
| 句數 | 2109 | 7156 | 5582 | 4170 | 657 | 2584 | 1773 | 2063 | 426 | 1124 | 882 | 778 |
| 句長 | 22.38 | 22.51 | 21.82 | 21.56 | 22.28 | 20.72 | 20.98 | 20.34 | 21.64 | 22.57 | 19.81 | 20.84 |
| 標點總數 | 3952 | 14593 | 10824 | 8116 | 1110 | 4908 | 3264 | 3960 | 753 | 2328 | 1449 | 1459 |
| 標點種類 | 20 | 23 | 22 | 19 | 16 | 19 | 18 | 19 | 17 | 20 | 17 | 20 |

# 2023年TCFD報告四要素文字統計

| 指標 | 本國銀行業 | | | | 壽險業 | | | | 產險業 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 治理 | 策略 | 風險 | 指標 | 治理 | 策略 | 風險 | 指標 | 治理 | 策略 | 風險 | 指標 |
| 字總數 | 29,007 | 105,204 | 82,546 | 46,833 | 15,323 | 50,036 | 36,471 | 16,059 | 6,506 | 22,260 | 19,675 | 8,615 |
| 每家平均 | 763 | 2,769 | 2,172 | 1,232 | 730 | 2,383 | 1,737 | 765 | 283 | 968 | 855 | 375 |
| 字種類 | 838 | 1,396 | 1,229 | 1,203 | 752 | 1,215 | 1,072 | 859 | 583 | 967 | 925 | 731 |
| 詞總數 | 11,995 | 42,114 | 33,464 | 18,365 | 6,245 | 20,158 | 14,777 | 6,246 | 2,637 | 8,983 | 7,911 | 3,310 |
| 詞種類 | 2,021 | 6,560 | 5,125 | 3,978 | 1,376 | 3,951 | 3,019 | 1,724 | 800 | 2,142 | 1,970 | 1,156 |
| 句數 | 1,419 | 5,656 | 4,257 | 2,590 | 710 | 2,421 | 1,814 | 759 | 284 | 977 | 892 | 395 |
| 句長 | 22.52 | 21.58 | 21.98 | 21.45 | 24.19 | 23.91 | 22.86 | 25.27 | 26.28 | 25.91 | 24.76 | 27.53 |

# 2023-24年內外部專家分數變異係數

| 專家分數 | 時間 | 內部人工編碼評鑑 | 外部專家學者評鑑 |
|---|---|---|---|
| 本國銀行業 | 2023年 | 0.25 | **0.08** |
| | 2024年 | 0.35 | **0.06** |
| 壽險業 | 2023年 | 0.24 | **0.07** |
| | 2024年 | 0.27 | **0.06** |
| 產險業 | 2023年 | 0.49 | **0.09** |
| | 2024年 | 0.52 | **0.10** |

# 2023-2024年文字＋公司迴歸分析

| 產業別 | 年度 | 內部分數 | | 外部專家 | |
|---|---|---|---|---|---|
| | | 變數個數 | $R^2$ | 變數個數 | $R^2$ |
| 銀行 | 2023年 | **9(6)** | **0.87** | **9(5)** | **0.90** |
| | 2024年 | **8(4)** | **0.73** | **9(4)** | **0.69** |
| 壽險 | 2023年 | **3(2)** | **0.65** | **7(4)** | **0.77** |
| | 2024年 | **6(4)** | **0.59** | **7(3)** | **0.58** |
| 產險 | 2023年 | **4(3)** | **0.66** | **4(3)** | **0.73** |
| | 2024年 | **5(3)** | **0.64** | **4(1)** | **0.67** |

註：刮號內為**文字變數**個數。

# 迴歸模型與BERT的比較

- 迴歸模型未必不如大語言模型！

→傳統分析除了能確定哪些是關鍵變數，模型也相對準確、穩定。

- 以2023及2024年銀行業外部評審分數為例，計算交叉驗證的RMSE(均方誤差)：

| 銀行業 | 迴歸 | BERT |
|---|---|---|
| 2023年 | 3.51 | 5.62 |
| 2024年 | 4.41 | 4.61 |

# 量化模型的分析結果

- 迴歸等模型比較能捕捉外部評審的想法，以及銀行業TCFD的(內外部)評審結果
- 2023年、2024年模型結果類似

→仍以銀行業估計誤差較小。

→2024年模型誤差略高於2023年，原因：評分標準略有調整、文字篇幅增加(多元)、ChatGPT等AI工具的影響。

- BERT模型未必較佳(樣本數較少)

# 討論與建議

- 文字分析僅置入物種豐富度、不均度

→由文字風格、語意組成關鍵詞叢(Cluster of Keywords)，計算詞彙間的關聯性。

- 樣本數較為不足

→樣本數少使得迴歸及BERT等模型較不穩定，不同年度的結果略有差異。

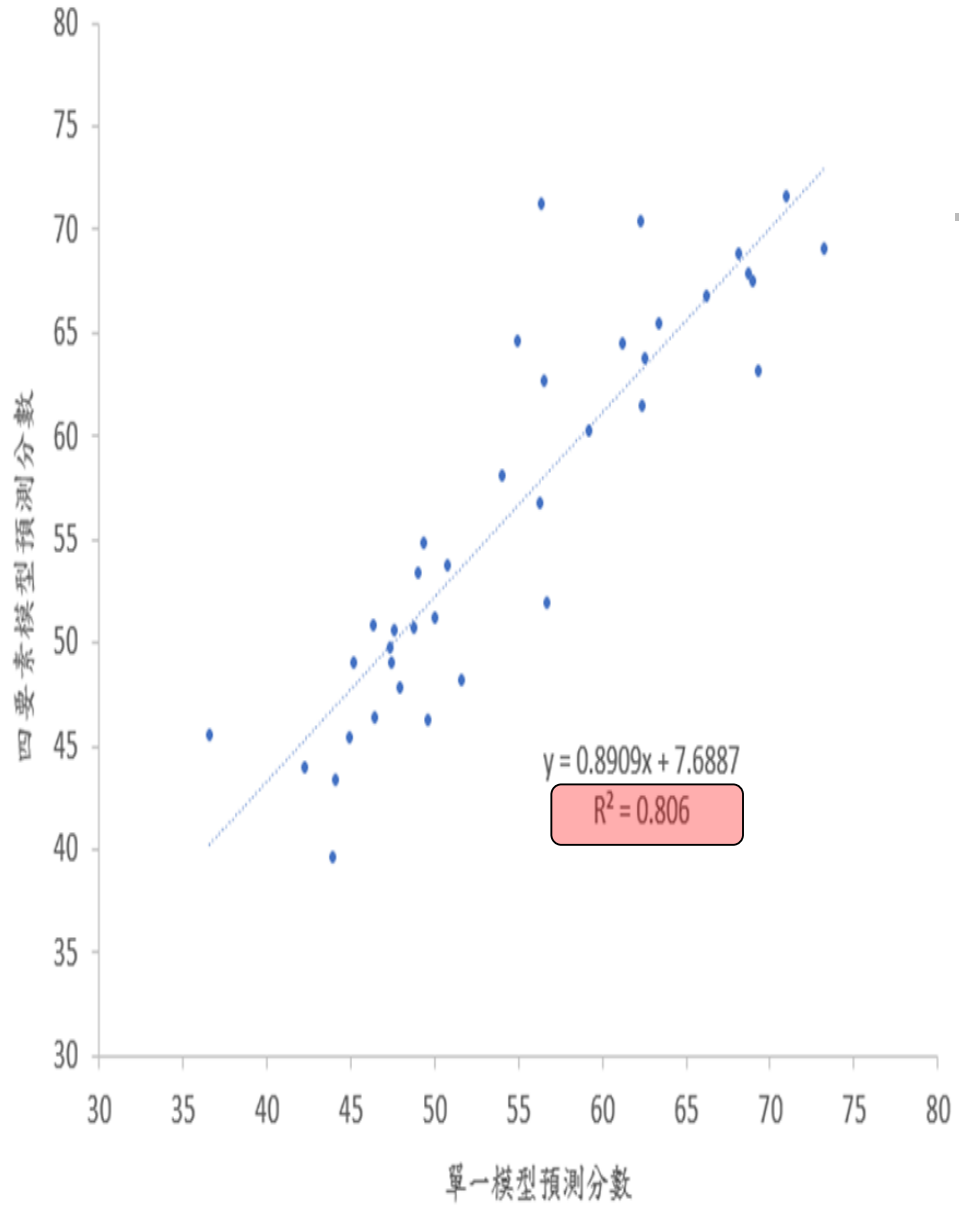
- 四個核心要素的個別模型？

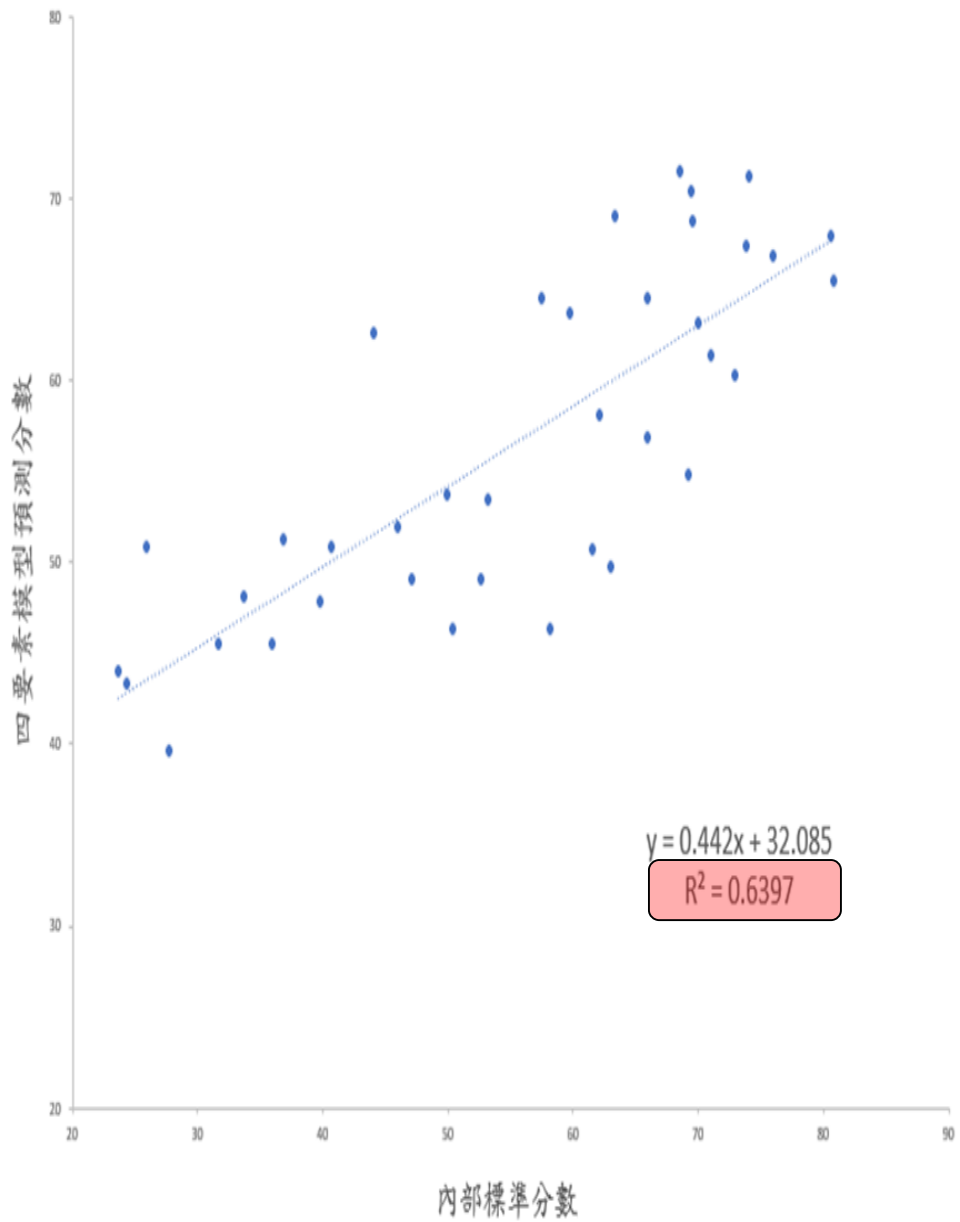
→模型細分為四個子模型、再結合為大模型？

# 2023～2024年銀行業迴歸分析結果

| | | 全文 | 治理 | 策略 | 風險 | 指標 |
|---|---|---|---|---|---|---|
| 2024 | $R^2$ | 73% | 63% | 50% | 38% | 51% |
| | 變數個數 | 8 | 9 | 8 | 2 | 4 |
| 2023 | $R^2$ | 89% | 43% | 73% | 71% | 53% |
| | 變數個數 | 9 | 3 | 9 | 14 | 6 |

# 單一與四要素迴歸模型(2024年銀行業)



# 報告完畢，

# 敬請指教！