F01

探索式資料分析與維度縮減

吳漢銘
國立政治大學 統計學系

http://www.hmwu.idv.tw

本章大綱&學習目標

- **探索式資料分析 (EDA)**
  - EDA簡介，川普推特發文例子
  - 資料視覺化的重要性
  - 圖表的誤用
  - 3D動態圖 (rgl套件)
  - 用R畫地圖
- **維度縮減方法**
  - 奇異值分解 (Singular Value Decomposition, SVD)
  - 主成份分析 (Principal Component Analysis, PCA)
  - 多維尺度法 (Multidimensional Scaling, MDS)
  - 等軸距特徵映射 (Isometric Feature Mapping, ISOMAP)
- **高維度低樣本數資料** (High-dimension Low-sample Size, HDLSS)
- **維度縮減評估指標**

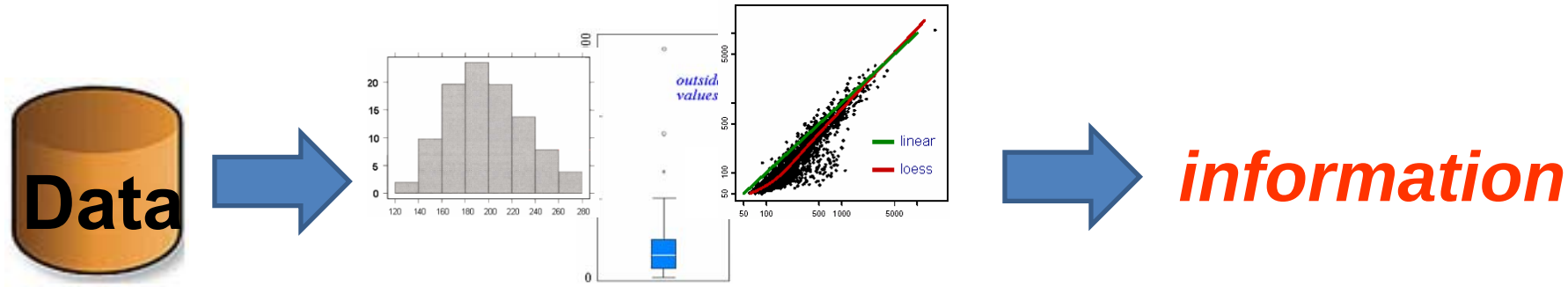
什麼是探索性資料分析?
(Exploratory Data Analysis, EDA)

Exploratory data analysis

From Wikipedia, the free encyclopedia

In statistics, **exploratory data analysis** is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA),[1] which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

- EDA method is generally cross-classified in two ways:
  - either non-graphical or graphical.
  - either univariate or multivariate (usually just bivariate).
- EDA is an iterative cycle.

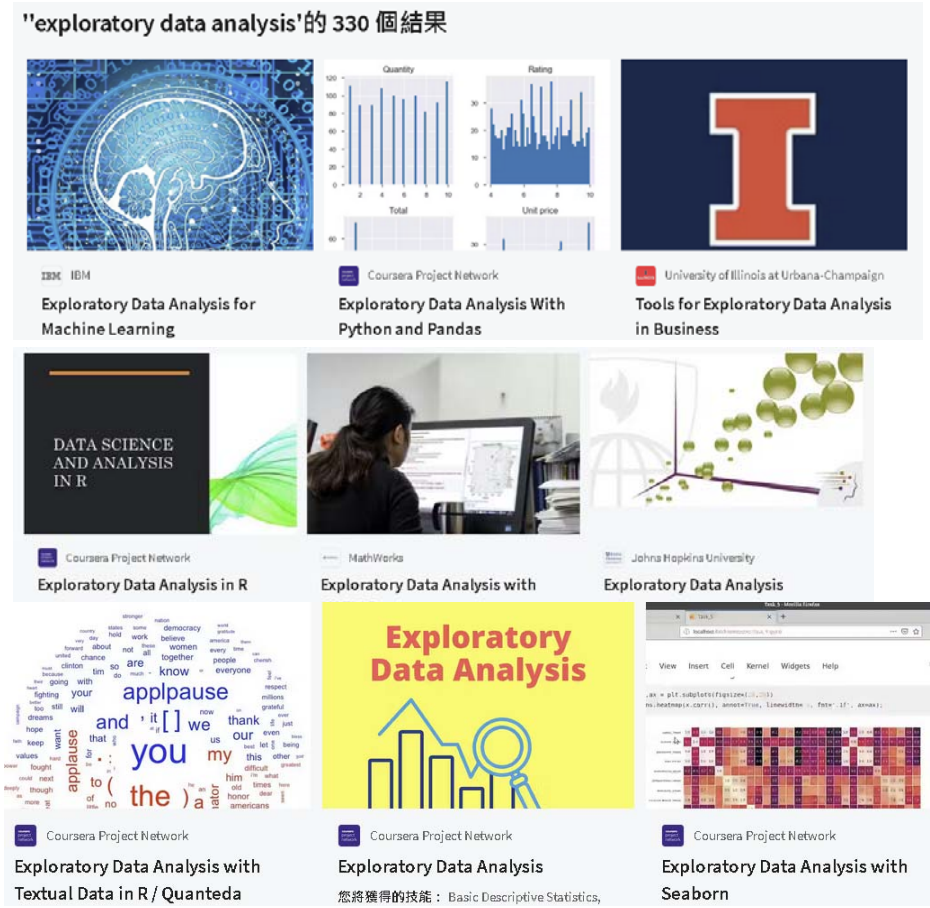


Exploratory Data Analysis (EDA) Tool

學習資源

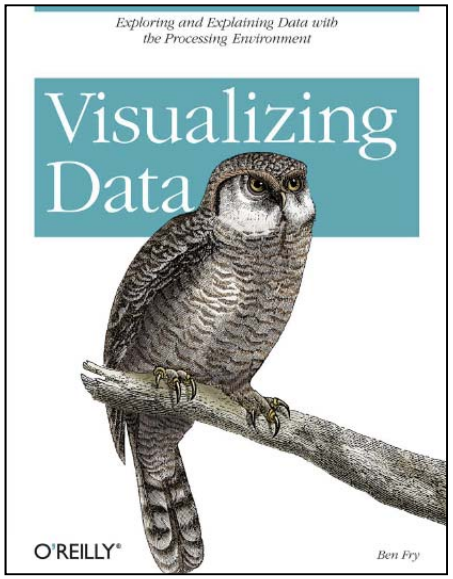
https://www.coursera.org/courses?query=exploratory%20data%20analysis

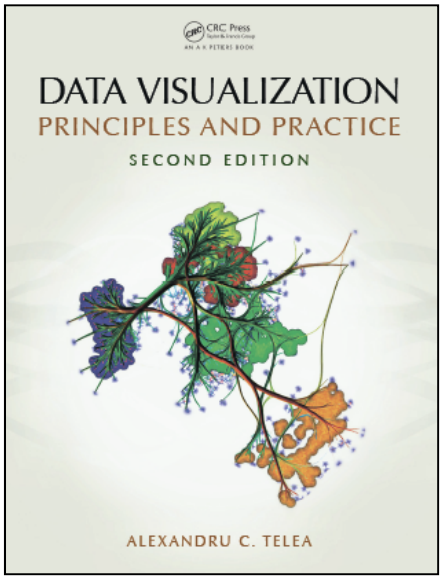
- UDACITY:
- edX:

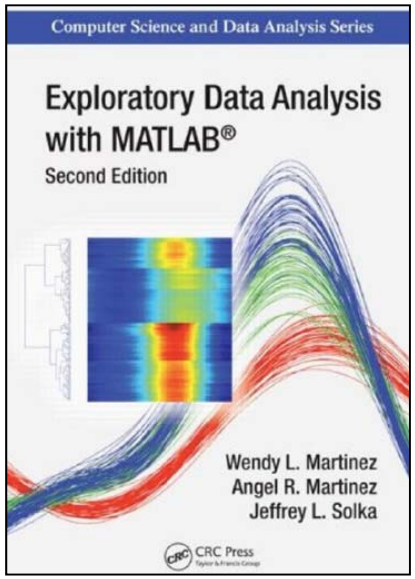


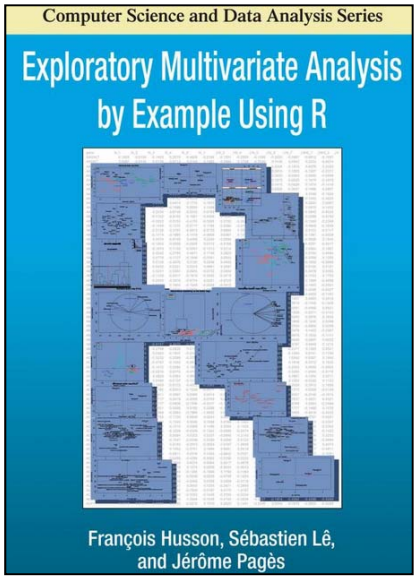
https://www.youtube.com/playlist?list=PLjTlxb-wKvXPhZ7tQwlROtFjorSj9tUyZ

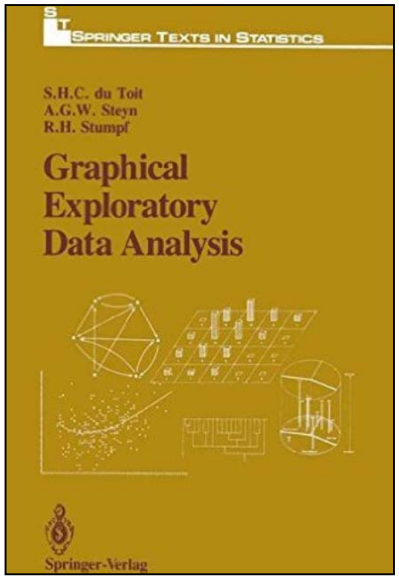
學習資源

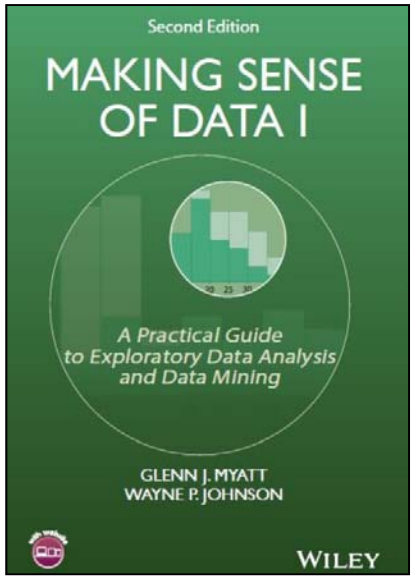

Exploring and Explaining Data with the Processing Environment
Visualizing Data
O'REILLY
Ben Fry

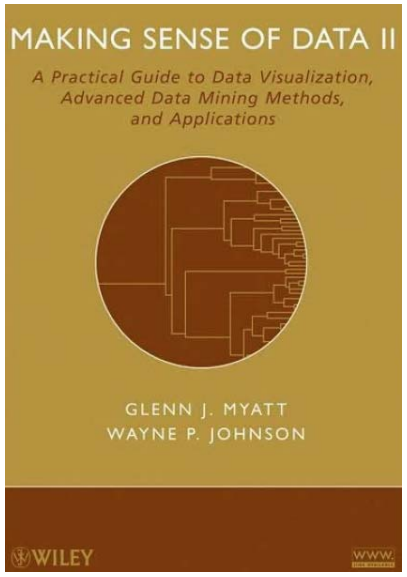

CRC Press
DATA VISUALIZATION
PRINCIPLES AND PRACTICE
SECOND EDITION
ALEXANDRU C. TELEA

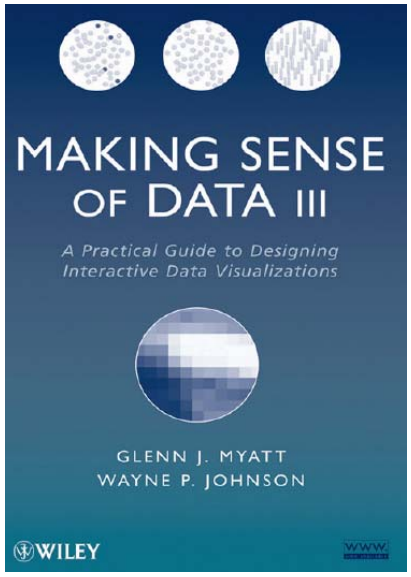

Computer Science and Data Analysis Series
Exploratory Data Analysis with MATLAB®
Second Edition
Wendy L. Martinez
Angel R. Martinez
Jeffrey L. Solka
CRC Press


Computer Science and Data Analysis Series
Exploratory Multivariate Analysis by Example Using R
François Husson, Sébastien Lê, and Jérôme Pagès


Springer Texts in Statistics
S.H.C. du Toit
A.G.W. Steyn
R.H. Stumpf
Graphical Exploratory Data Analysis
Springer-Verlag


Second Edition
MAKING SENSE OF DATA I
A Practical Guide to Exploratory Data Analysis and Data Mining
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY


MAKING SENSE OF DATA II
A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY


MAKING SENSE OF DATA III
A Practical Guide to Designing Interactive Data Visualizations
GLENN J. MYATT
WAYNE P. JOHNSON
WILEY

EDA在資料科學中的重要性

- Exploratory Data Analysis (EDA) is an **approach/philosophy** for data analysis that employs a variety of techniques (mostly **graphical**) to
  - summaries **for large, complicated data sets.**
  - maximize insight into a data set,
  - discover underlying **structure, patterns, features, trends, outliers, anomalies, and relationships in data .**
  - refine your questions and/or generate new questions about your data.
  - helps determine how best to **manipulate data sources** to get the answers you need,
  - extract **important variables**,
  - investigate the **quality** of your data: detect **outliers** and anomalies (detection of mistakes),
  - determine if the **statistical techniques** are appropriate,
  - test a hypothesis, or **check assumptions** in statistical models,
  - develop parsimonious **models**,,
  - determine **optimal** factor settings,
  - determine **relationships** among the explanatory variables, and **outcome variables**
  - **Interaction between the researcher and the data.**
  - **Identifying the areas of interest.**

- **You should always look at every variable - you will learn something!**

John Tukey (1915~2000): 統計學界的畢卡索

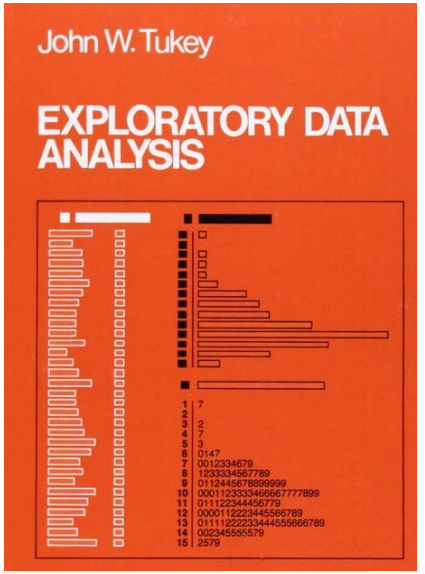
生平

- 布朗大學**化學**學士及碩士。
- 1939年: 普林斯頓大學**數學**博士。(數理統計)
- 二次大戰加入火砲控制研究室，以及後來加入**AT&T**貝爾實驗室(**創立統計組**)，接觸統計上的實際問題。

「對正確的問題有個近似的答案，

勝過對錯的問題有精確的答案。」

"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question."



對後世的貢獻

- 發明快速傅立葉轉換(FFT)。
- 創造bit (位元)及 software(軟體) 。
- 探索性的資料分析 (Exploratory Data Analysis, EDA, 1977)

Source: http://www.unige.ch/ses/sococ/cl/bib/eda/tukey.html

「統計應該是科學，而非數學！」

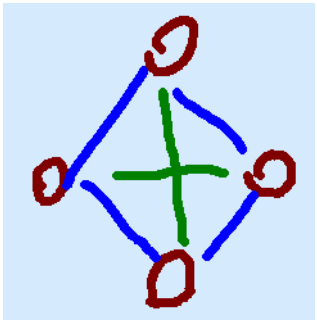
Stanford Linear Accelerator (1973)

他曾挑戰當時主流的數理統計學家，堅持 data analysis 是統計分析中不可忽視的步驟，數學的假設需要 data 加以驗證才可行。 Tukey 說過統計應該是科學，而非數學！

數學思維 vs 統計思維

證明在哪裏? vs 數據在哪裏?

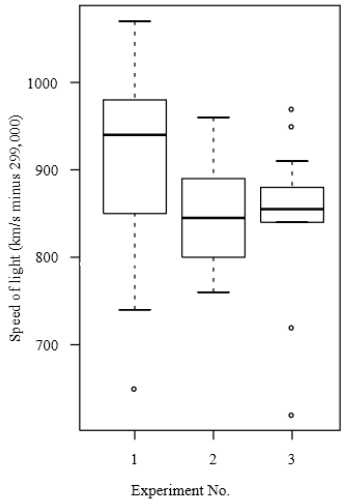
"Let the data speak for themselves"



Stem and Leaf Plot

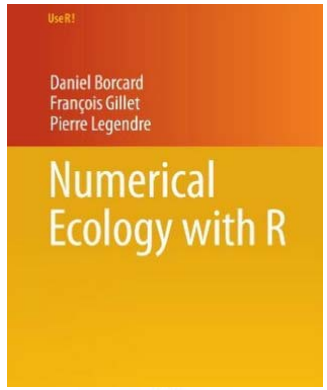
```
42 | 0
44 | 0000
46 | 000000
48 | 00000000000
50 | 00000000000000000000
52 | 00000
54 | 00000000000000
56 | 000000000000000
58 | 0000000000
60 | 000000000000
62 | 0000000000000
64 | 000000000000
66 | 0000000000
68 | 0000000
70 | 00
72 | 0000
74 | 0
76 | 00000
78 | 0
```

Box-and-whisker plot



What Do They Say About EDA?

- **Daniel Borcard, Francois Gillet, Pierre Legendre (2011):**
  - A first exploratory look at the data can tell much about them.
  - Information about simple parameters and distributions of variables is important to consider in order to choose more advanced analyses correctly.
  - EDA is often neglected by people who are eager to jump to more sophisticated analyses. It should have an important place.

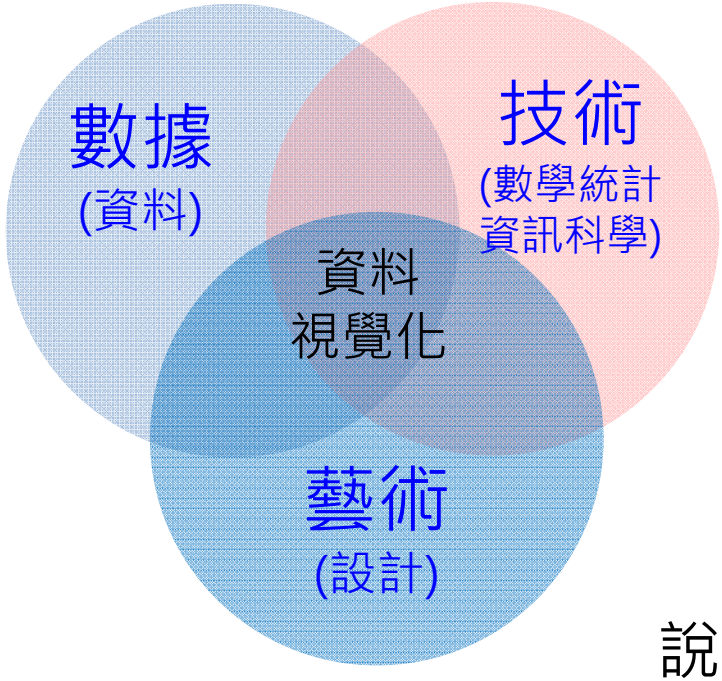


Source: google images

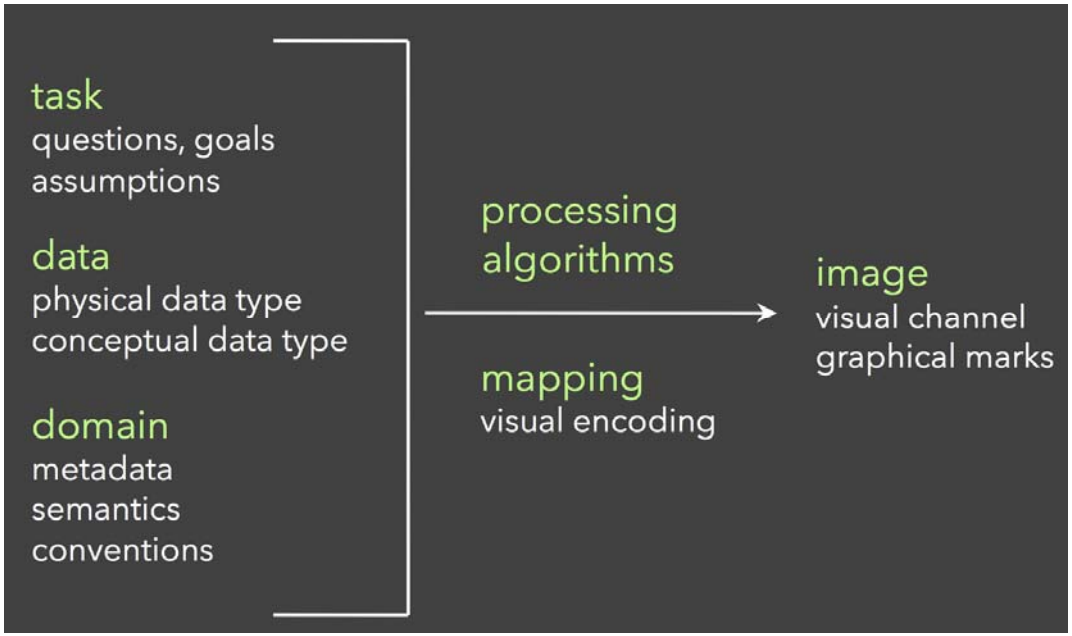
資料視覺化

Visualization: Making things/processes/abstractions visible (to transform into pictures) that are not directly accessible by the human eye.

Visualization = Graphing for Data + Fitting + Graphing for Model



通用的視覺化流程 (分析-處理-生成)



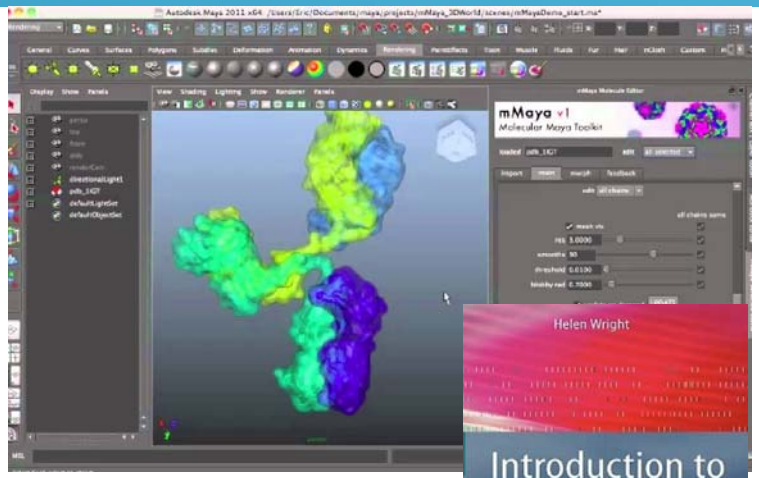
https://geekplux.com/2017/01/01/basics-of-data-visualization-the-process-model.html

說出數據的故事：設計資料視覺化的要點

https://www.inside.com.tw/2015/07/10/telling-the-story-of-your-data

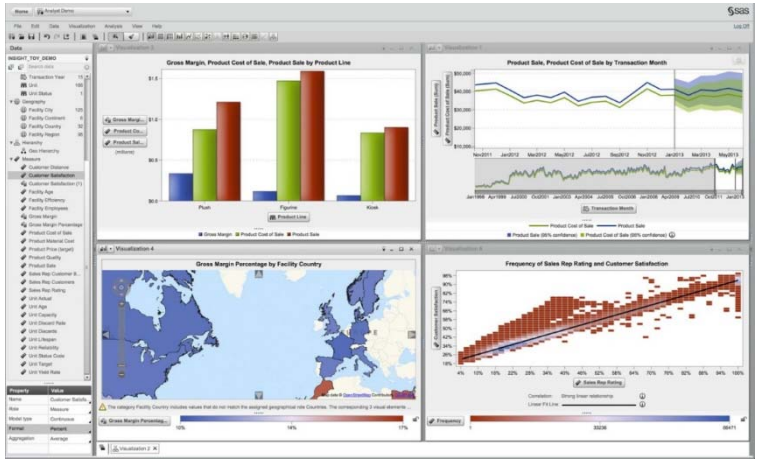
資料視覺化 (Data Visualization)

Scientific Visualization, Information Visualization, Visual Analytics



- **Scientific Visualization**:
  - 以圖像方式說明科學資料(天然幾何結構: 如磁感線、流體分佈等)，使科學家能夠從資料中瞭解、說明和收集規律。
  - 觀察基於物理的數據(人體、地球、分子等幾何結構)。
  - 通常是三維資料現象的視覺化。例如: 建築學、氣象學、醫學或生物學方面的各種系統。
- **Information Visualization**:
  - 處理抽象概念、非結構化的資料集合，轉化成為視覺化資訊。(包括數位和非數位資料，例如金融數據、地理資訊、層次結構、文本)
  - 互動式視覺呈現以加強人類認知。例如: 柱狀圖、趨勢圖、流程圖、樹狀圖等。
- **Visual Analytics**:
  - 利用互動式視覺介面進行分析推理的科學。
  - 有效結合人腦智慧和機器智慧，以視覺感知為通道，通過視覺化互動介面，降低資料的複雜度到人腦與機器智慧均可處理的範圍。

SAS Visual Analytics



例子: 川普推特誰寫的?

真道理性 真愛台灣
中時電子報
chinatimes.com

樂公益 | 廣播 | 旺車 | 開卷 | 樂購物 | 爆料

六月 6

文史百科 您知道，世界上最早釀造啤酒和飲用啤酒的民族是？

速覽 | 政治 | 生活 | 社會 | 財經 | 國際 | 兩岸 | 軍事 | 熱門 | 旅遊 | 娛樂 | 體育

即時 | 日報 | 言論 | 時周 | 周刊王 | 樂時尚 | 有影 | 話題 | 秒懂圖 | 精選 | CAMPUS

首頁 › 中時電子報 › 科技

即時首頁 | 政治 | 生活 | 社會 | 旅遊 | 娛樂 | 體育 | 財經 | 國際 | 兩岸 | 科技 | 軍事 | 熱門 | 人物

川普推特都自己寫的嗎？大數據揭密

2017年02月03日 10:55 黃慧雯 / 綜合報導

A A A

分享至Facebook 分享至Google+ 分享至Twitter 分享至Weibo

透過大數據分析川普個人推特的推文，結果十分驚人。(圖 / 翻攝川普個人推特)

若要形容甫就任美國第45任總統的川普(Donald Trump)，「推特狂人」肯定是個不會被遺忘的說法。川普靠著他的Tweets(推文)，在總統選戰中餵養著成千上萬

黃慧雯

黃慧雯的最新文章

- WWDC / 向開發者釋出善意 蘋果ARkit等開發工具
- WWDC / macOS High Sierra發表快更安全
- WWDC / watchOS 4來了 更聰明貼心
- WWDC / 跑VR輕而易舉 超強iMa登場
- WWDC / 蘋果發表iOS 11 控制中頭換面

訂閱科

【錯過可惜】Follow me! 權員一起來

【魅力城市】魅力海南 美味文昌

【魅力城市】魔鬼城 鬼斧神工

【台味餐盒】央行責便當 文化野餐「綠光」

真道理性 真愛台灣
中時電子報
chinatimes.com

樂公益 | 廣播

台灣地理知識 台灣最長的河川「濁水溪」沒有流經

速覽 | 政治 | 生活 | 社會 | 財經 | 國際 | 兩岸 | 軍事

即時 | 日報 | 言論 | 時周 | 周刊王 | 樂時尚 | 有影 | 話題

首頁 › 中時電子報 › 科技

即時首頁 | 政治 | 生活 | 社會 | 旅遊 | 娛樂 | 體育 | 財經 | 國際 | 兩岸 | 科技

民調已死！美大選川普勝出 大數據神預測

2016年11月09日 14:57 黃慧雯 / 綜合報導

A A A

共和黨候選人川普正式贏得2016美國總統選舉，跌破一票專家眼鏡，也打臉各家民調。(圖 / 美聯社)

2016年美國總統大選結果已經出爐，共和黨候選人川普(Donald Trump)至截稿

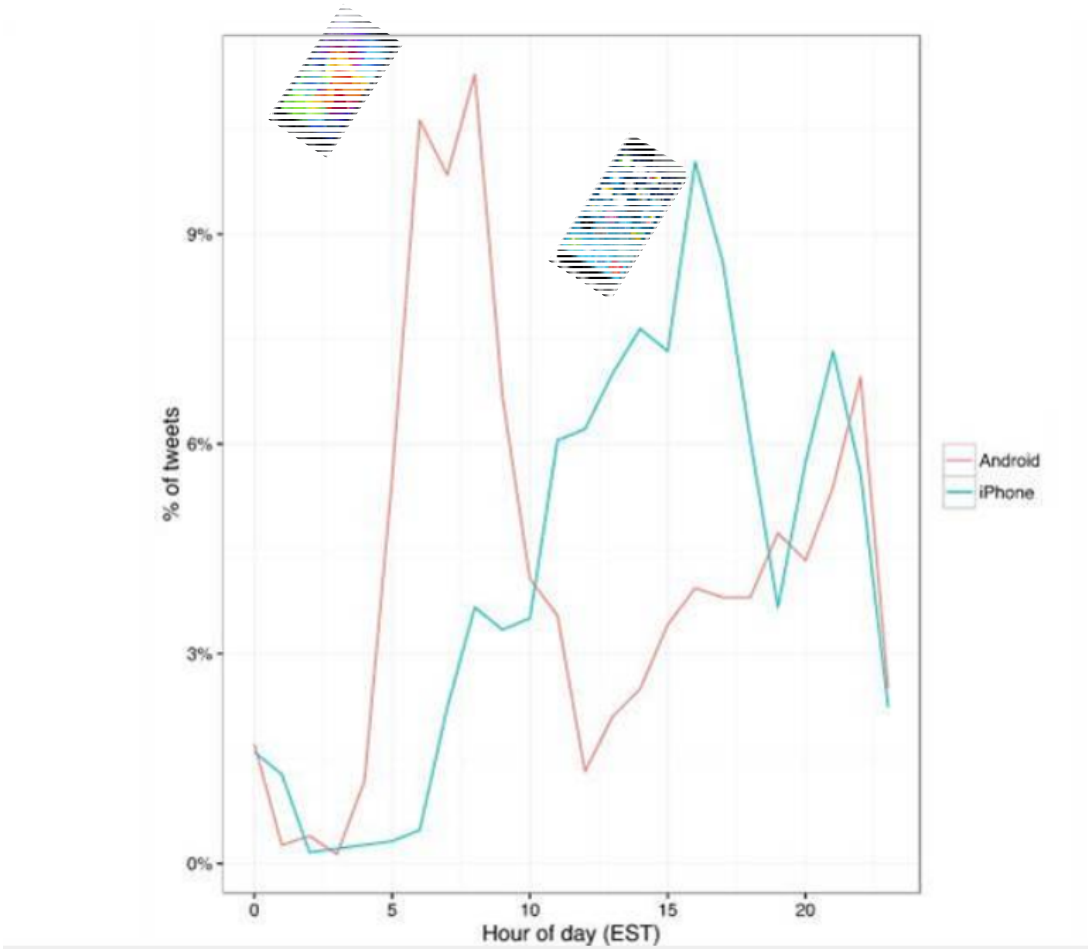
有疑問?

數據分析師David Robinson發現，川普發表祝賀內容時，是透過iPhone；而用來抨擊選戰對手時，則是透過Android手機。到底川普個人推特推文的差異，從何而來？這些推文是不是由他一個人包辦，

Twitter網友發現川普推文分別來自iPhone與Android手機端，且發文內容風格迥異。(圖 / 翻攝DZone)

發文時間對比

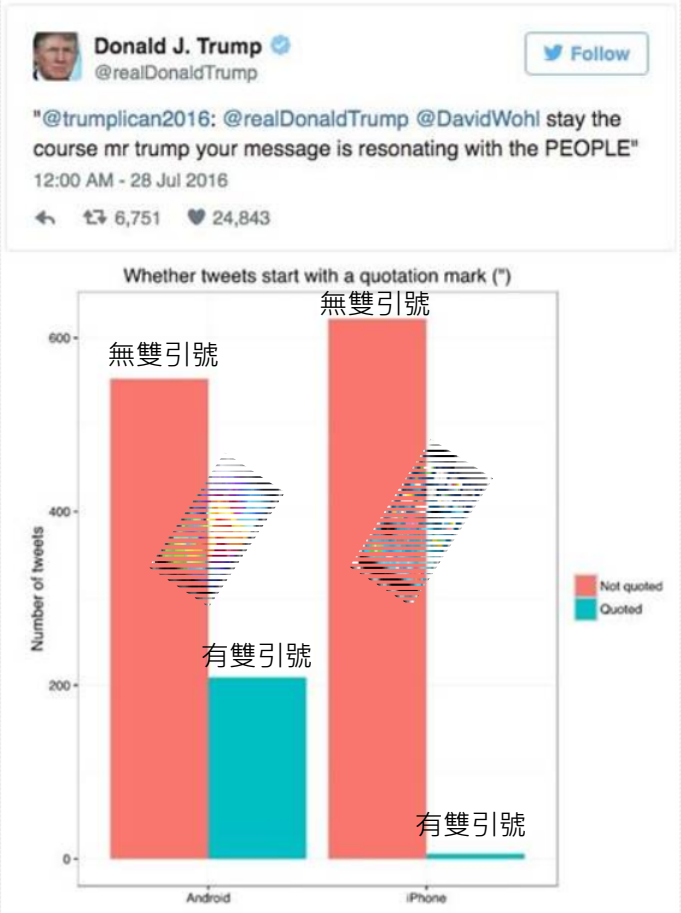
→川普習慣在早上發推文；而他的助理或團隊習慣在下午或晚上發推文



就推文時間分析來看，可看出來自Android手機的推文時間大多落在早上，與來自iPhone端的推文時間區間不同。(圖 / 翻攝DZone)

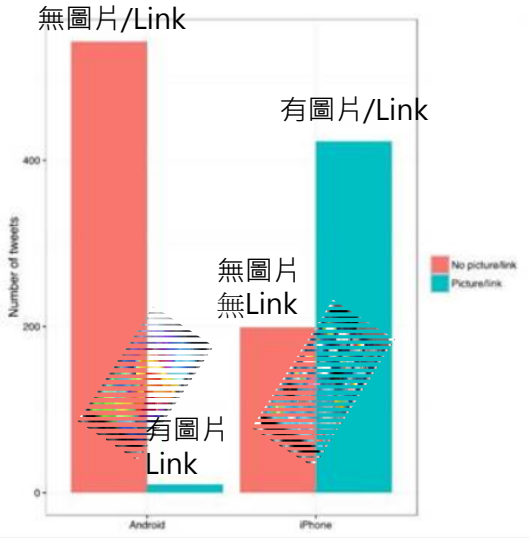
發文習慣對比

→川普轉推慣用雙引號，他的團隊則沒有這個習慣



川普轉推推文多愛用雙引號。(圖 / 翻攝DZone)

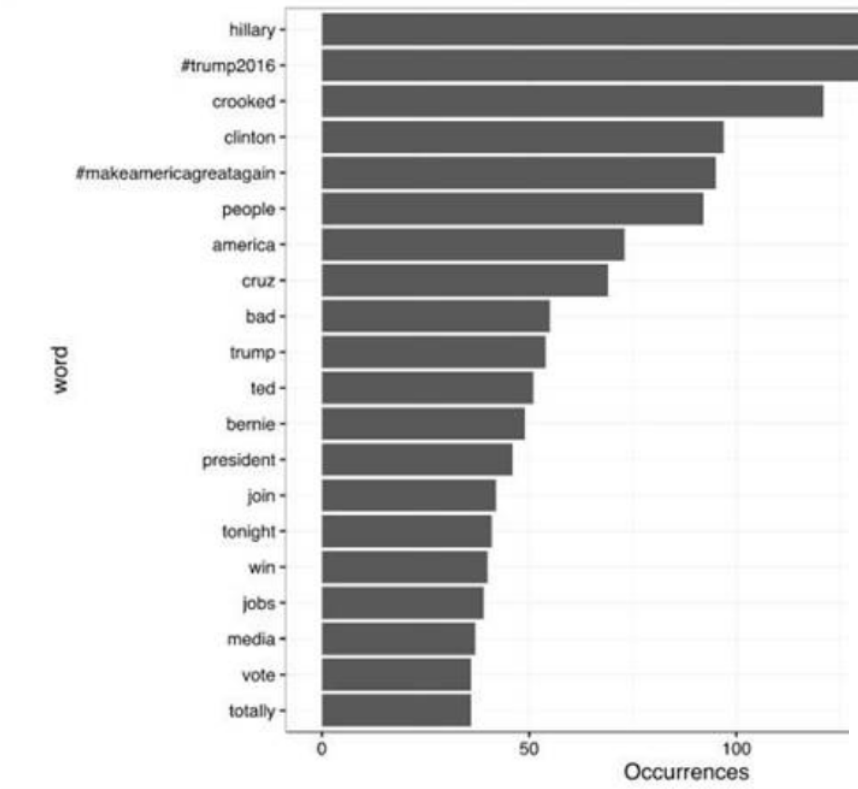
→川普的推文都以文字為主，少附link以及圖片



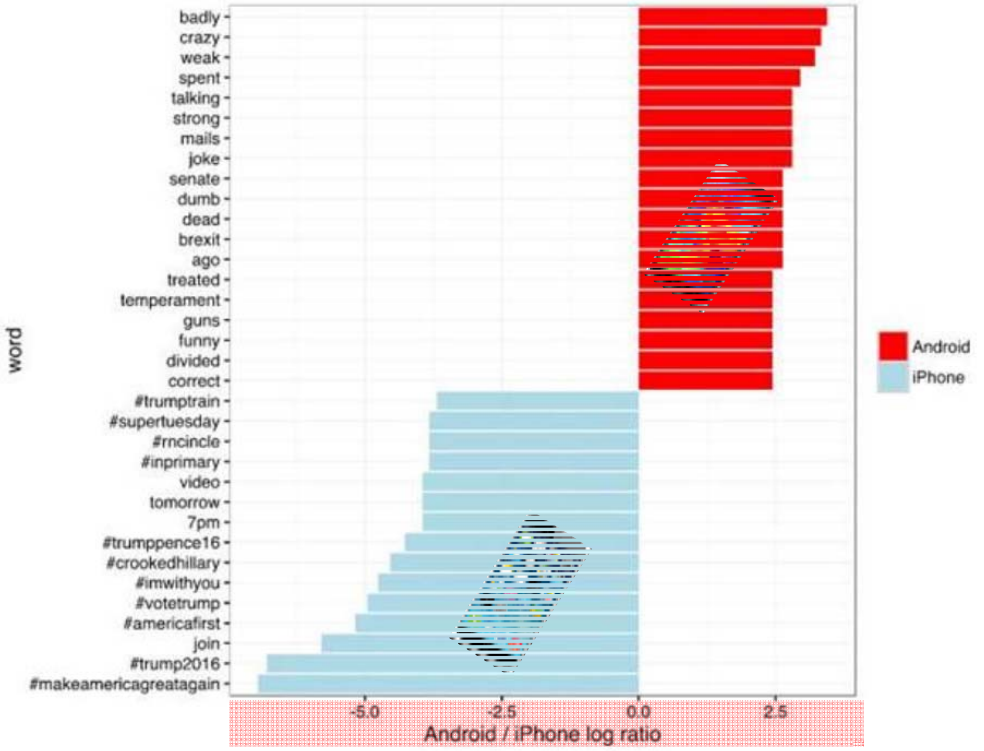
川普的推文很少用link以及圖片(如左下)，來自iPhone的推文習慣不同，常附圖片。(圖 / 翻攝DZone)

發推文文字對比

就發推文時使用的文字來看，以下是來自Android手機的推文常見字



川普推文常用字。(圖 / 翻攝DZone)



Android帳號推文與iPhone推文常用字的對比。(圖 / 翻攝DZone)

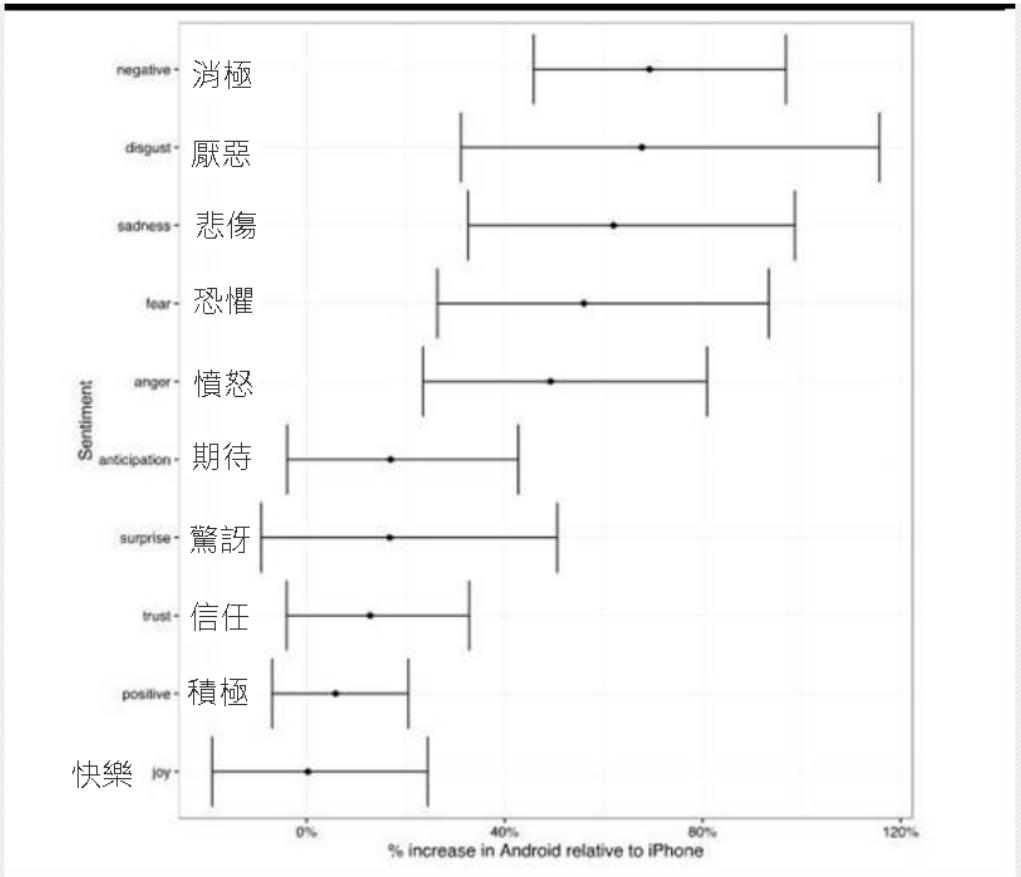
情感分析

- 用 tidytext 當中的NRC Word-Emotion Association辭典，數據分析師將推文的用詞跟「積極、消極、憤怒、期待、厭惡、恐懼、快樂、悲傷、驚訝、信任」這十種情緒進行了**關聯分析**，結果發現：

- Android手機的推文中(共4901個字)，總共有321個字與「**憤怒**」的情感有關、有207個字與「**厭惡**」的情緒有關。

- 而透過**Poisson test** 分析後，更可明顯發現Android手機的推文更喜歡使用強烈情緒性的字眼，若透過**95%信賴區間**來看，就能看出Android手機推文與iPhone推文的不同。

→從結果來看，Android手機端的推文，使用「厭惡、悲傷、恐懼、憤怒」等消極情緒字眼的比例比iPhone的推文高出40%~80%。

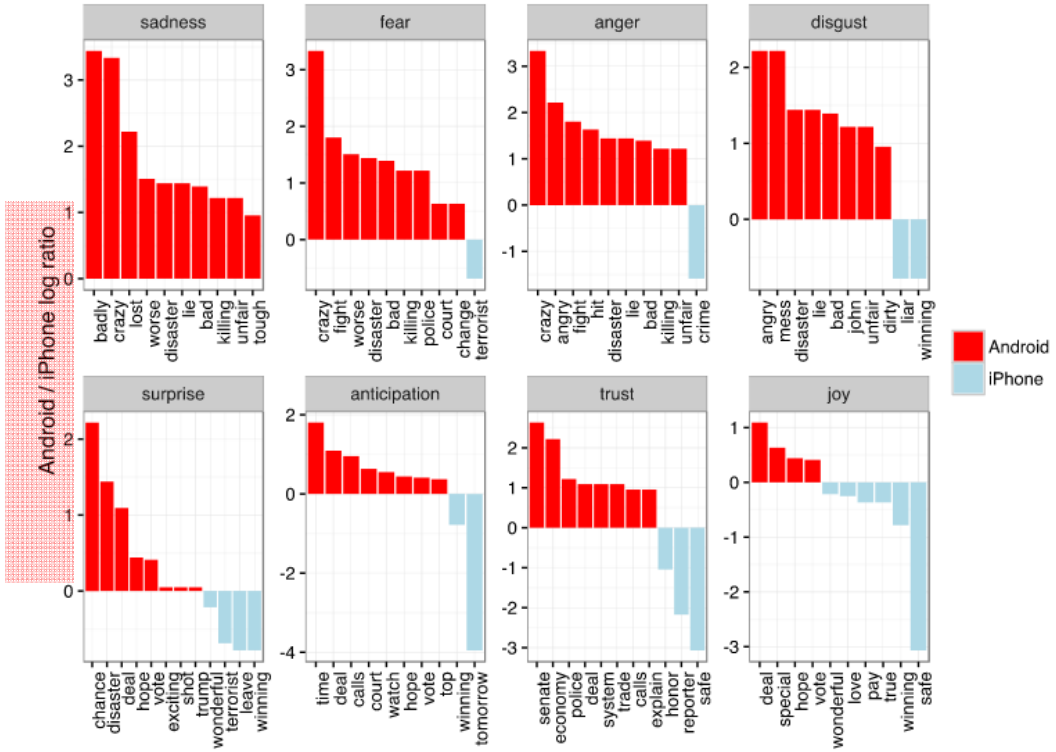


以95%信賴區間來看來看Android手機推文與情緒的關聯性。(圖 / 翻攝DZone)

總結: 川普推特誰寫的?

- 從川普個人推特帳號的**單則推文**中，可能看不出個所以然。然而在**大數據的分析下**，卻能很清楚看出脈絡。

- 川普個人推特的推文，來自Android手機的發文與來自iPhone的發文，明顯是由不同人所寫，因為發推時間、推文內容、標籤使用率、轉發方式都截然不同。且**來自Android手機的推文也顯得更為激烈與消極**。
- 川普個人用來發推的行動裝置，就是三星的Galaxy系列手機。基於上述分析，幾乎可以確定來自Android手機的推文是由川普本人所發；而來自iPhone的推文，則應該是出於他助理團隊之手。



Android手機推文愛用情緒性字詞的比例比iPhone推文高出很多。(圖 / 翻攝DZone)

Text Analysis of Trump's Tweets with R Code

https://dzone.com/articles/text-analysis-of-trumps-tweets-confirms-he-writes

the majority of the tweets from the iphone are fairly benign declarations. but consider cases like these, both posted from an iphone:

the dataset

first, we'll retrieve the content of donald trump's timeline using the `usertimeline` function in the twitter package:

```
library(dplyr)
library(purrr)
library(twitter)
```

```
# you'd need to set global options with an authenticated app
setup_twitter_oauth(getoption("twitter_consumer_key"),
                    getoption("twitter_consumer_secret"),
                    getoption("twitter_access_token"),
                    getoption("twitter_access_token_secret"))

# we can request only 3200 tweets at a time; it will return fewer
# depending on the api
trump_tweets <- usertimeline("realdonaldtrump", n = 3200)
trump_tweets_df <- tbl_df(map_df(trump_tweets, as.data.frame))
```

```
# if you want to follow along without setting up twitter authentication,
# just use my dataset:
load(url("http://varianceexplained.org/files/trump_tweets_df.rda"))
```

Why Data Visualization?

- It is not about "**infographics**", the beautiful, heavily customized products of expert graphic designers.
- Data visualization can provide clear understanding of patterns in data, detect hidden structures in data, condense information.
- **Anscombe's quartet** comprises four datasets. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
- Four datasets have nearly identical simple statistical properties, yet appear very different when graphed.

	I		II		III		IV	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

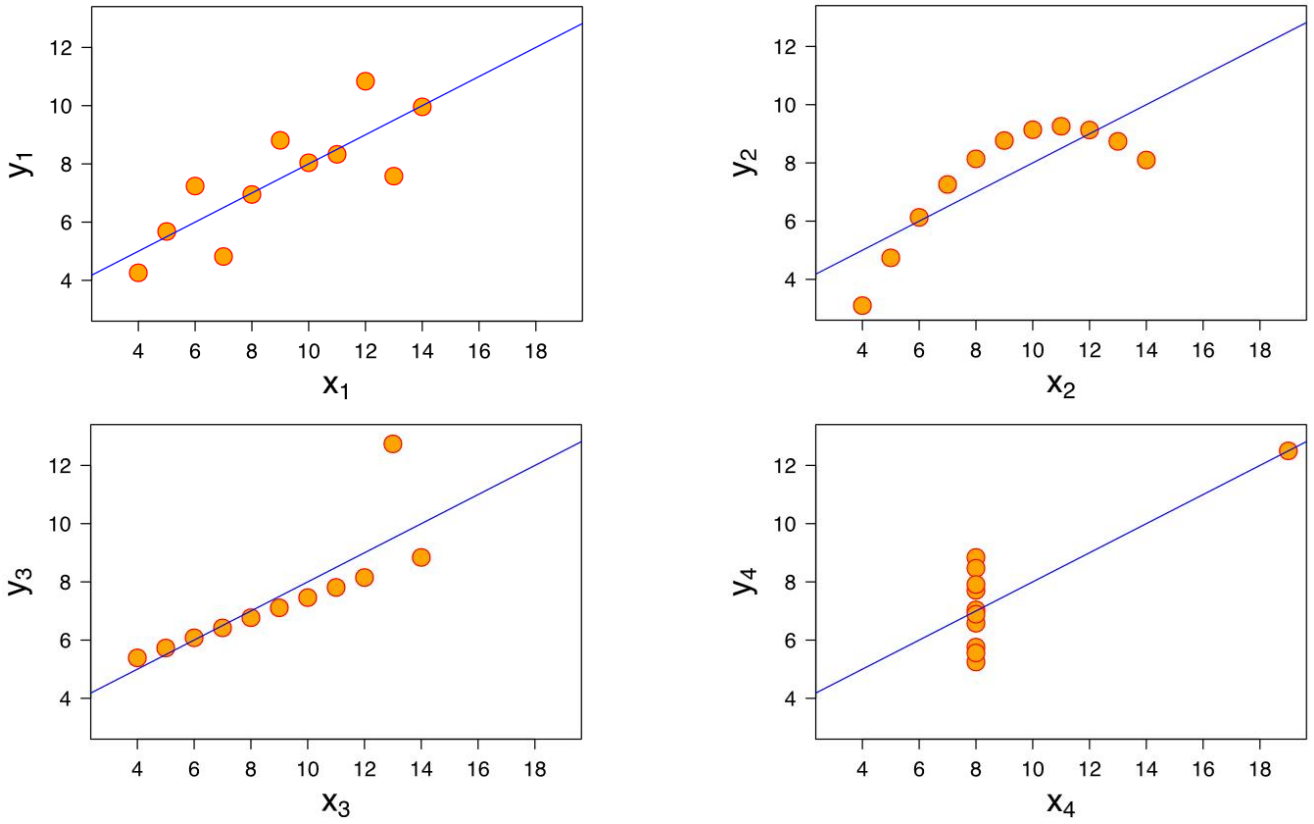
Mean of x in each case: 9 (exact)
Sample variance of x in each case: 11 (exact)
Mean of y in each case: 7.50 (to 2 decimal places)
Sample variance of y in each case: 4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case: 0.816 (to 3 decimal places)
Linear regression line in each case: y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)

https://en.wikipedia.org/wiki/Anscombe's_quartet

http://ryanwomack.com/IASSIST/DataViz/

Anscombe's Quartet

- **Mean of x** in each case: 9 (exact)
- **Sample variance of x** in each case: 11 (exact)
- **Mean of y** in each case: 7.50 (to 2 decimal places)
- **Sample variance of y** in each case: 4.122 or 4.127 (to 3 decimal places)
- **Correlation** between x and y in each case: 0.816 (to 3 decimal places)
- **Linear regression line** in each case: y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)



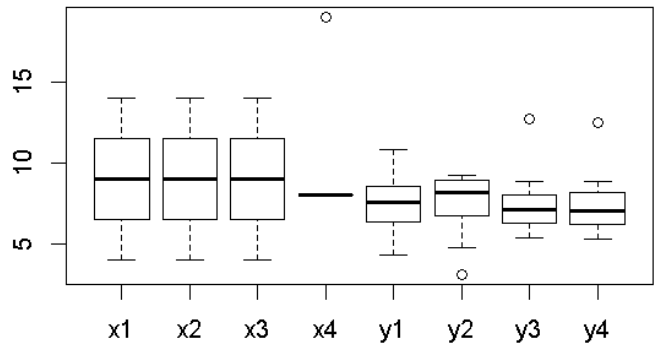
anscombe {datasets}

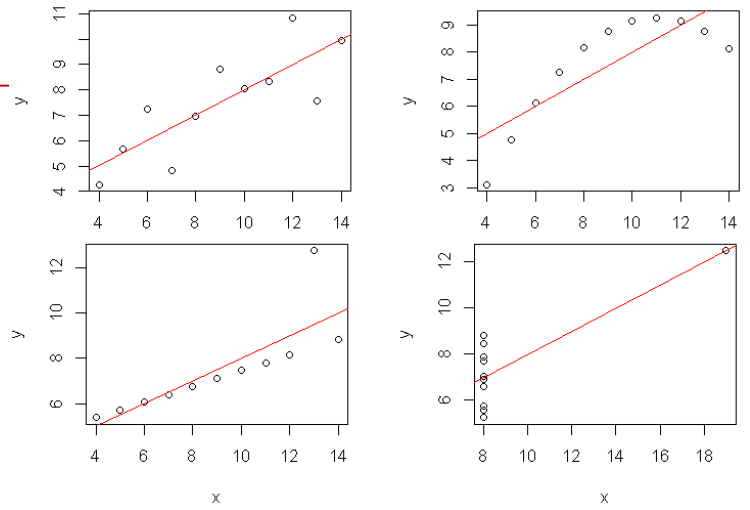
Anscombe's Quartet of 'Identical' Simple Linear Regressions

```
> head(anscombe, 3)
   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
> apply(anscombe, 2, mean)
      x1       x2       x3       x4       y1       y2       y3       y4
9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909
> apply(anscombe, 2, sd)
      x1       x2       x3       x4       y1       y2       y3       y4
3.316625 3.316625 3.316625 3.316625 2.031568 2.031657 2.030424 2.030579
> mapply(cor, anscombe[,1:4], anscombe[,5:8])
       x1        x2        x3        x4
0.8164205 0.8162365 0.8162867 0.8165214
> mapply(function(x, y) lm(y~x)$coefficients, anscombe[, 1:4], anscombe[, 5:8])
                 x1       x2        x3        x4
(Intercept) 3.0000909 3.000909 3.0024545 3.0017273
x           0.5000909 0.500000 0.4997273 0.4999091
```

```
par(mfrow=c(2, 2))
regplot <- function(x, y){
  plot(y~x)
  abline(lm(y~x), col="red")
}
mapply(regplot, anscombe[, 1:4], anscombe[, 5:8])
```

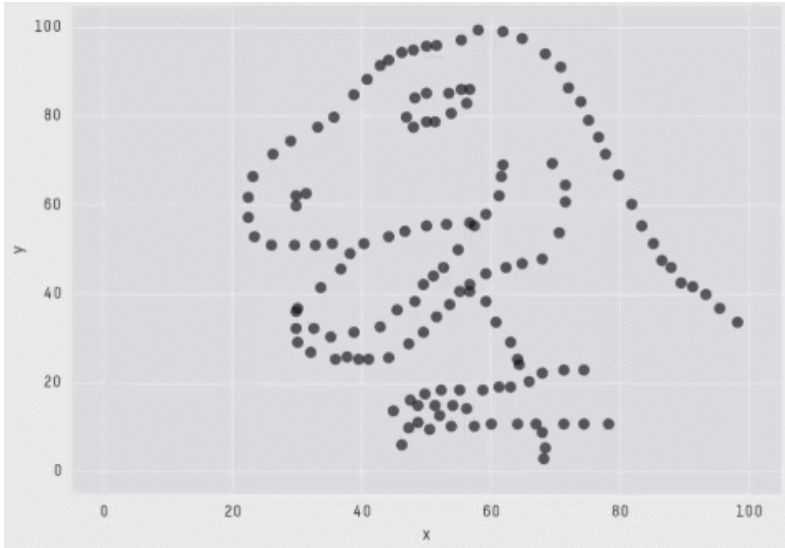
boxplot(anscombe)





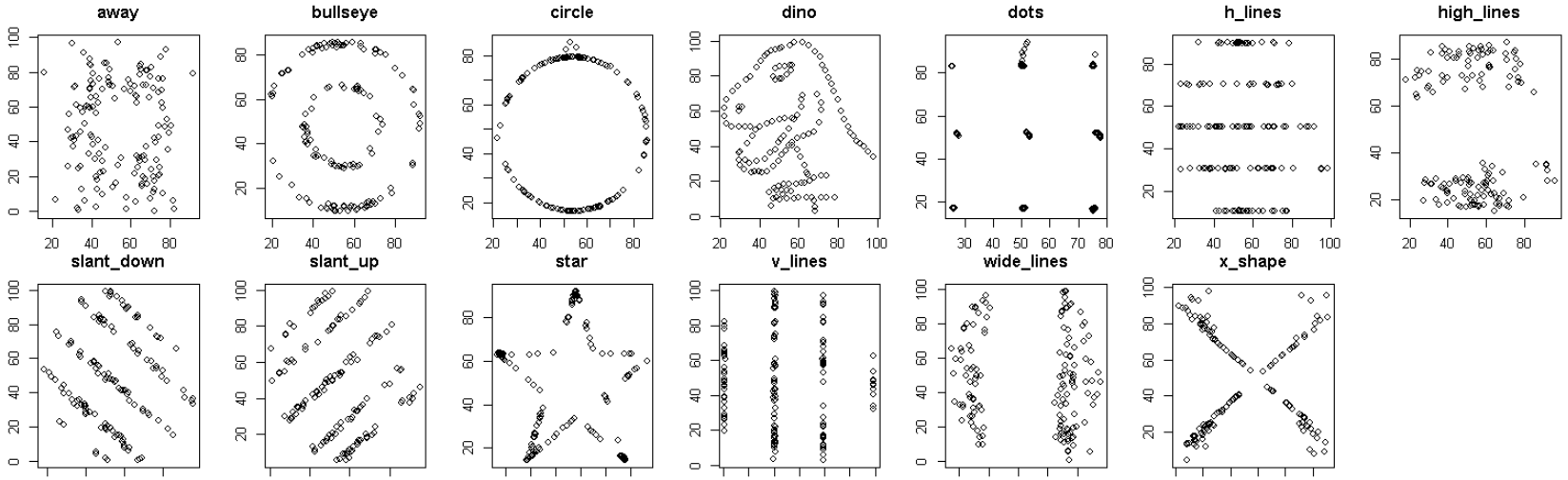
The Datasaurus Dozen

```
install.packages("datasauRus")
```



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

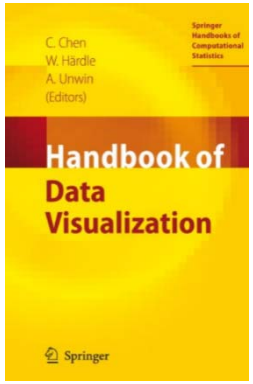
Justin Matejka and George Fitzmaurice, Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. https://dl.acm.org/doi/10.1145/3025453.3025912

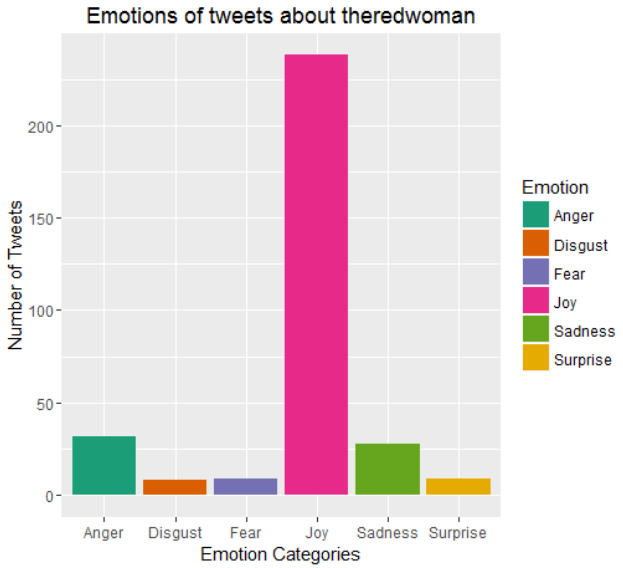


Infovis and Statistical Graphics: Different Goals, Different Looks

Journal of Computational and Graphical Statistics, Volume 22, 2013 - Issue 1

- **Infovis and Statistical Graphics: Different Goals, Different Looks**
  Andrew Gelman & Antony Unwin, Pages: 2-28
- **InfoVis Is So Much More**, Robert Kosara, Pages: 29-32
- **InfoVis and Statistical Graphics: Comment**
  Paul Murrell, Pages: 33-37
- **Graphical Criticism: Some Historical Notes**
  Hadley Wickham , Pages: 38-44
- **Tradeoffs in Information Graphics**
  Andrew Gelman & Antony Unwin , Pages: 45-49





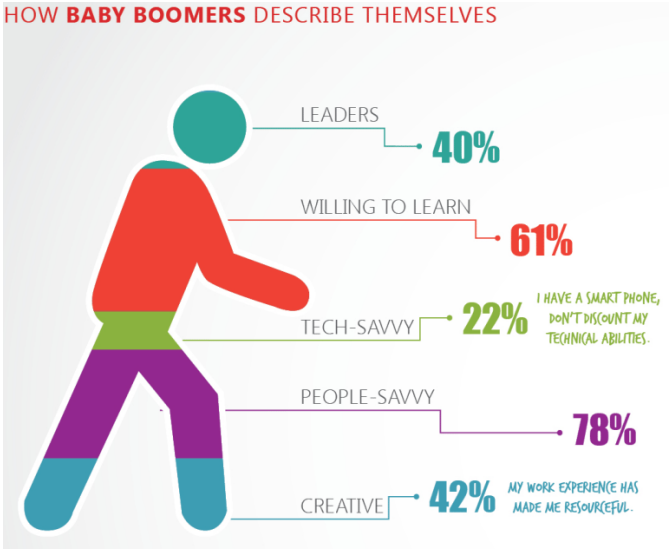
http://emarketingwall.com/how-twitter-responded-to-the-latest-episode-of-game-of-thrones

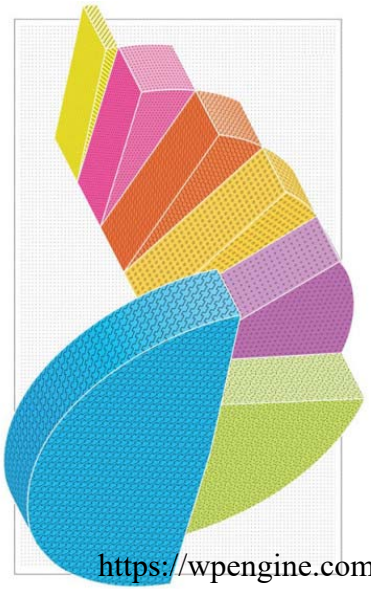
https://www.r-bloggers.com/words-in-politics-some-extensions-of-the-word-cloud/

不好的圖形~

- **Friedman (2008):**
  - The main goal of data visualization is to communicate information clearly and effectively through graphical means.
  - (X) look boring to be functional or extremely sophisticated to look beautiful.
  - Designers often fail to achieve a balance between (aesthetic) form and functionality, creating gorgeous data visualizations which fail to serve their main purpose — to communicate information.

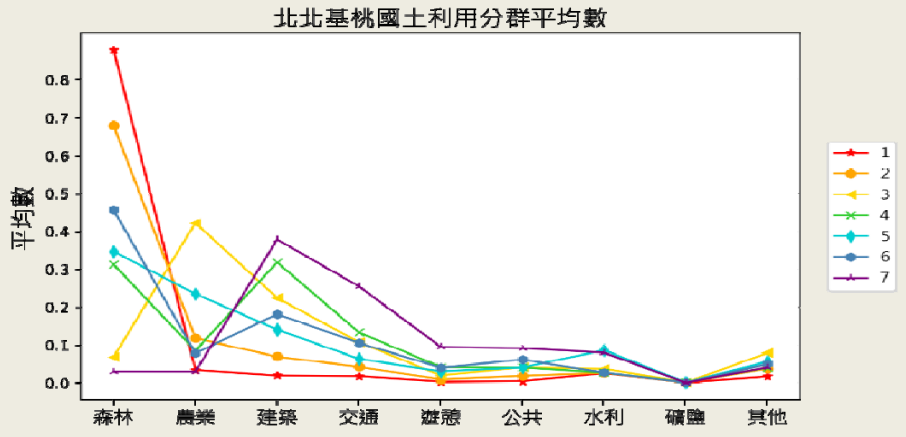
https://www.smashingmagazine.com/

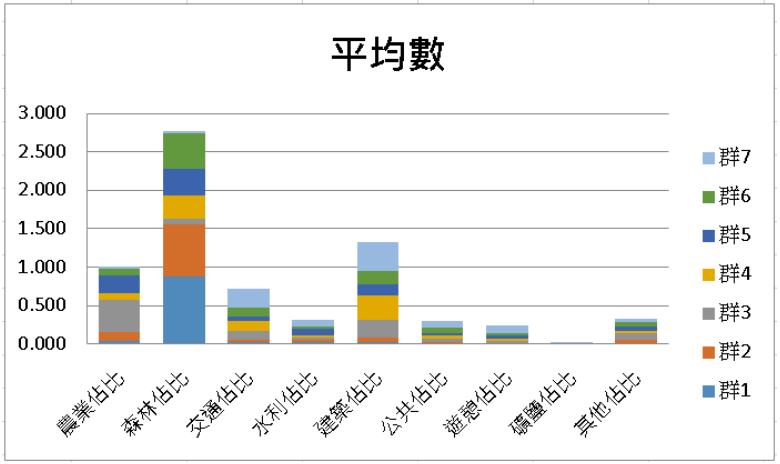


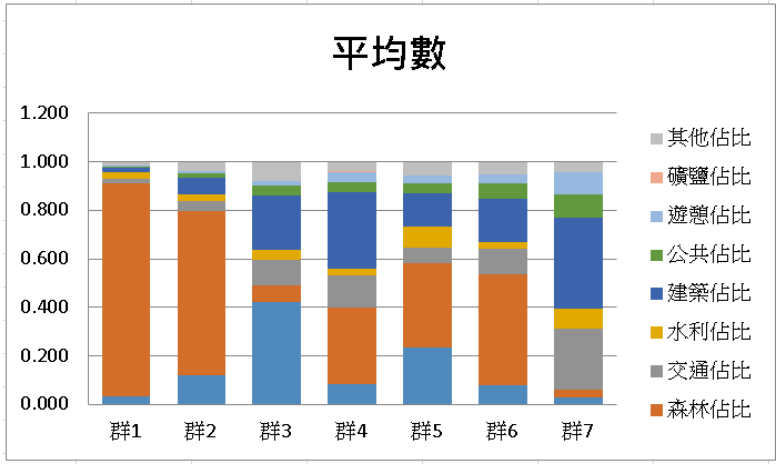


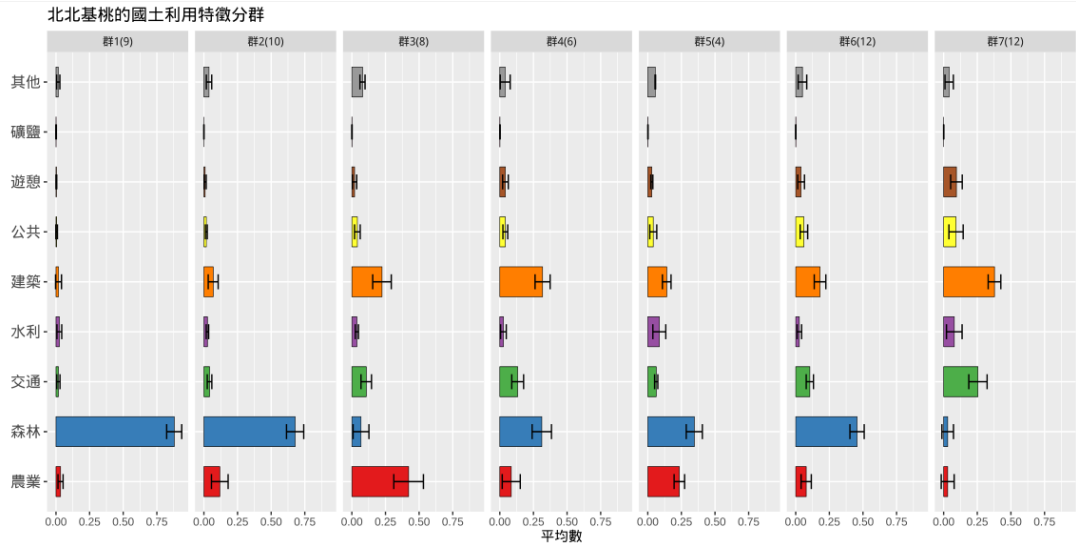
https://wpengine.com/blog/ugly-side-data-visualization/

圖表的修正與改進

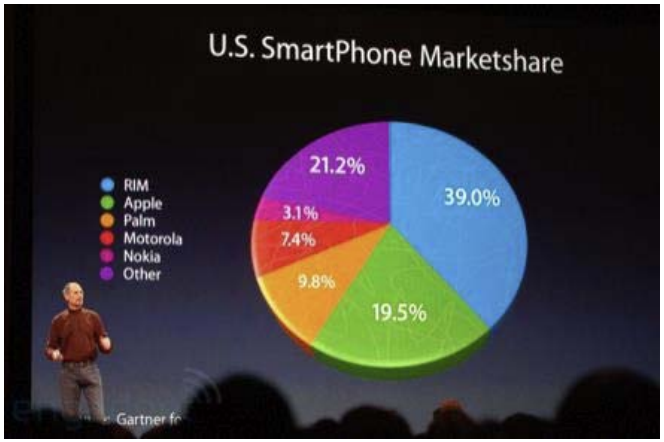

北北基桃國土利用分群平均數
平均數
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
森林
農業
建築
交通
遊憩
公共
水利
礦鹽
其他
1
2
3
4
5
6
7

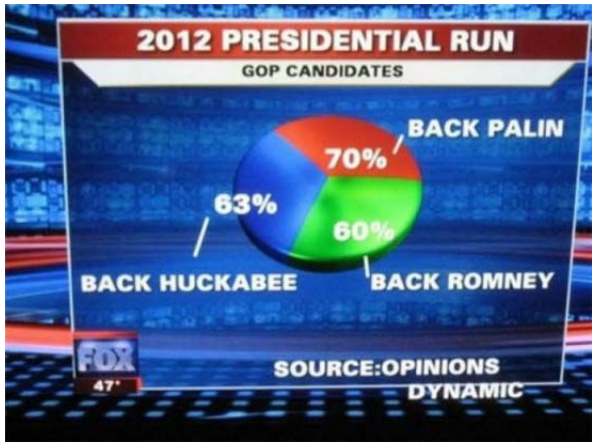

平均數
3.000
2.500
2.000
1.500
1.000
0.500
0.000
農業佔比
森林佔比
交通佔比
水利佔比
建築佔比
公共佔比
遊憩佔比
礦鹽佔比
其他佔比
群7
群6
群5
群4
群3
群2
群1


平均數
1.200
1.000
0.800
0.600
0.400
0.200
0.000
群1
群2
群3
群4
群5
群6
群7
其他佔比
礦鹽佔比
遊憩佔比
公共佔比
建築佔比
水利佔比
交通佔比
森林佔比

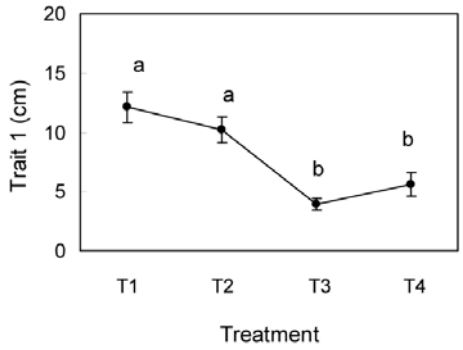

北北基桃的國土利用特徵分群
群1(9)
群2(10)
群3(8)
群4(6)
群5(4)
群6(12)
群7(12)
其他
礦鹽
遊憩
公共
建築
水利
交通
森林
農業
0.00
0.25
0.50
0.75
平均數

圖表的誤用

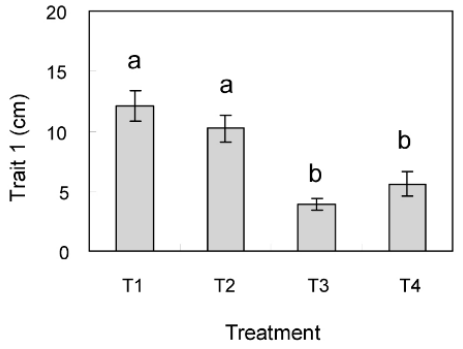




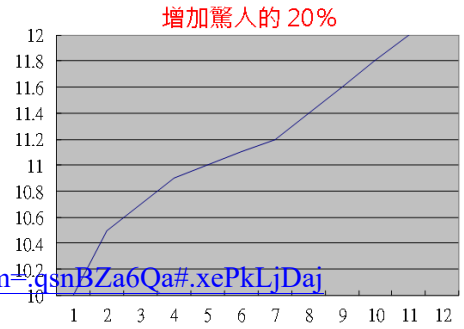
Source: https://www.managertoday.com.tw/articles/view/51480

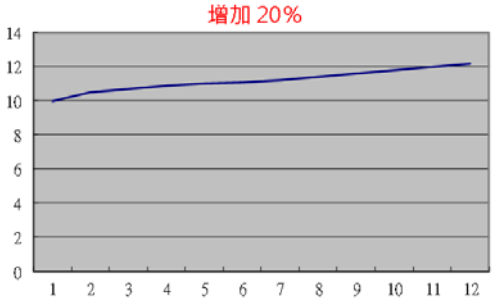


Source: http://ir.tari.gov.tw:8080/bitstream/345210000/3094/1/journal_arc_60-1-6.pdf



Misleading Graphs: Real Life Examples
http://www.statisticshowto.com/misleading-graphs/
The top ten worst graphs
https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/
Bad Infographics: 11 Mistakes You Never Want to Make
http://blog.visme.co/bad-infographics/
13 Graphs That Are Clearly Lying
https://www.buzzfeed.com/katienotopoulos/graphs-that-lied-to-us?utm_term=.qsnBZa6Qa#.xePkLjDaj
11 Most Useless And Misleading Infographics On The Internet
https://io9.gizmodo.com/11-most-useless-and-misleading-infographics-on-the-inte-1688239674
The most misleading charts of 2015, fixed
https://qz.com/580859/the-most-misleading-charts-of-2015-fixed/
Misleading graph
https://en.wikipedia.org/wiki/Misleading_graph





Example:「蔡政府執政後出生數連年下滑」

圖表的誤導?

yahoo! 新聞 Search 搜尋新聞 搜尋網頁

熱門話題： 總統候選人2020 12/21遊行 季汴寫真集圖片 年終獎金計算 電動牙刷 地震分級增至10級

首頁 政治 2020大選 論壇 財經 娛樂 運動 社會地方 國際 生活 健康 科技 天氣 影音 立委o

揭張善政「玩弄數據」！四叉貓狠打臉

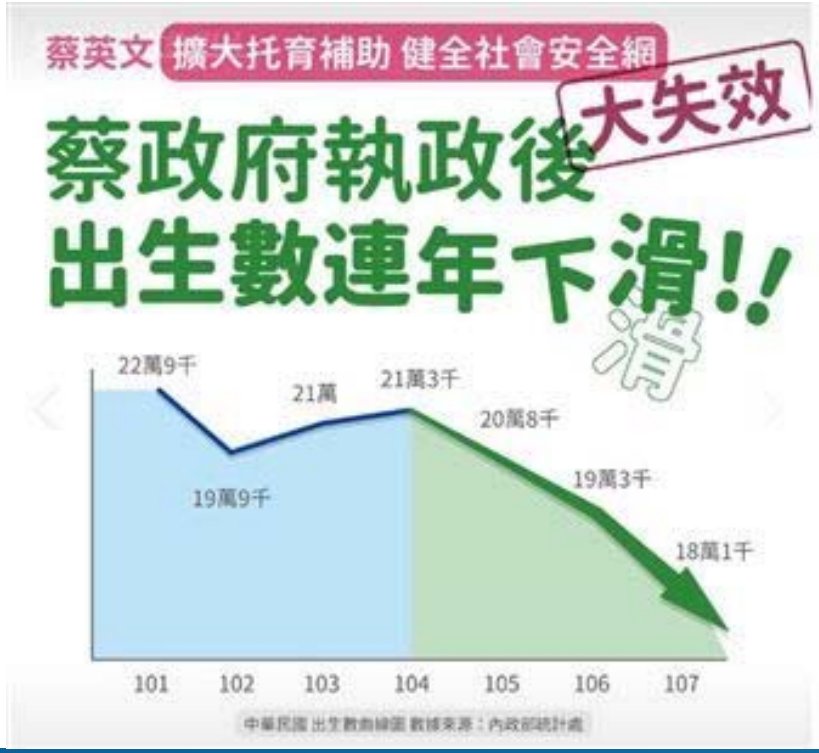
三立新聞網 setn.com | 44.3k 人追蹤 追蹤

三立新聞網 2019年12月14日 下午2:50

4 則留言

政治中心／綜合報導

國民黨副總統候選人張善政近日以歧視言論攻擊總統蔡英文，一句「她沒生過小孩，不知道這個（為人父母）心」引發砲轟，之後張善政雖然表示願意收回，卻又自稱「我這樣的理工男」沒有適當包裝文字，引發誤會非常抱歉。今（14）日張善政再度猛攻蔡政府生育政策，同時PO出數據予以佐證，卻被網紅「四叉貓」劉宇提出3點狠打臉，直言張善政根本是「玩弄數據」，更反嗆：「這樣也敢自稱理工男？」




張善政

#用數據客觀檢視蔡政府的政策
#蔡總統執政後新生兒人數年年下滑
#台灣不需要恐嚇人民的政黨

台灣面臨少子化的危機早已不是新聞，但攤開統計數據仍令人目驚心。蔡總統的生育政策做得好不好，不用我們多費唇舌，數據會說話。

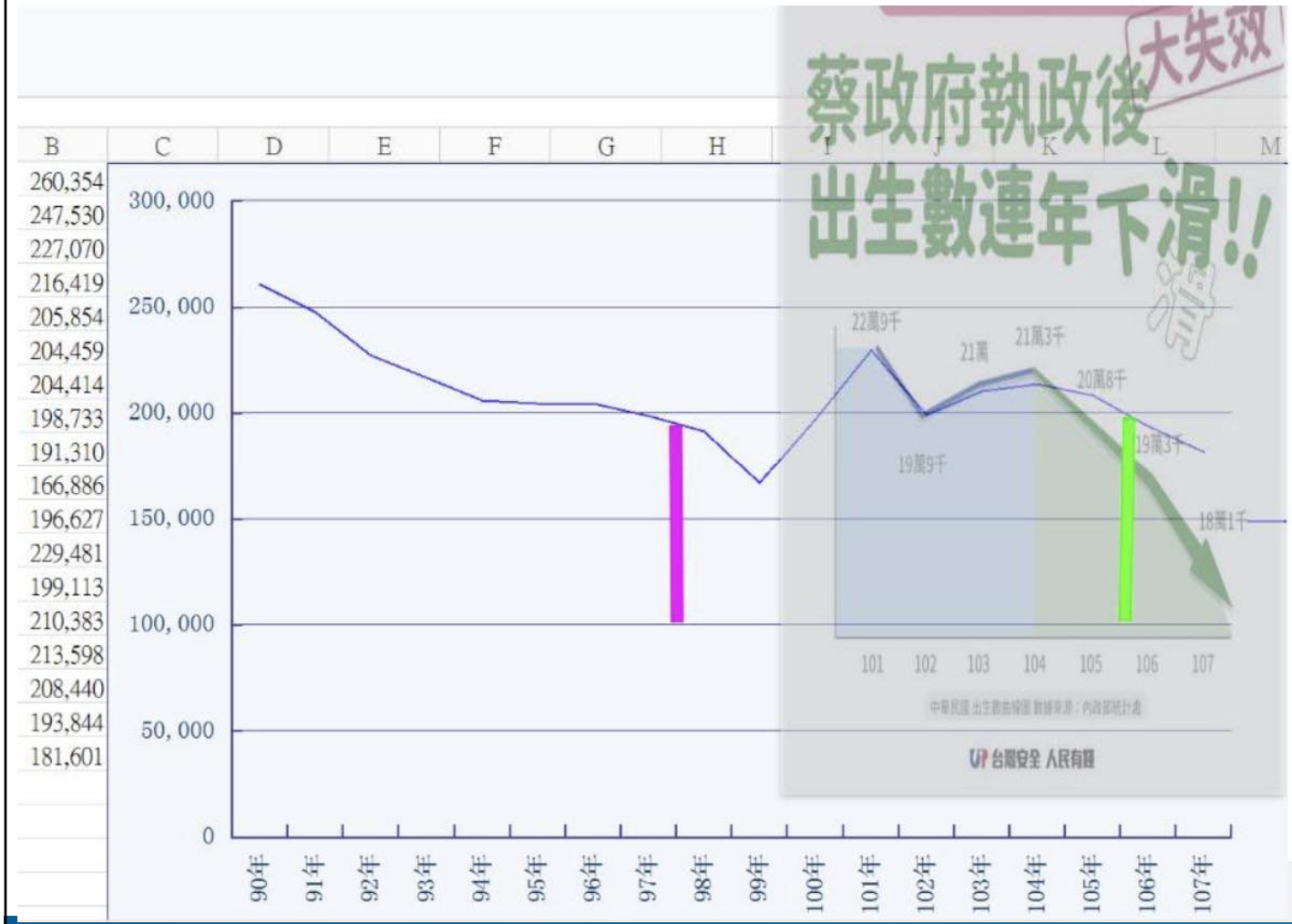
根據內政部的統計，台灣新生兒人數在蔡政府上任後斷斷下滑，去年僅18萬1千餘人，今年的情況更加嚴峻，截至11月只有16萬零8百人。

蔡政府的生育政策是否對症下藥，還是讓台灣人愈來愈不敢生小孩，經過統計後，很遺憾的是後者。姑且不論蔡政府對少子化毫無作為，也提不出有效的生育政策，只能抄襲國政顧問團的「0到6歲國家幫忙養」，蔡政府在其他方面的政策也是導致少子化的幫兇。

蔡政府上任後提出一例一休，不僅沒有解決結構性的問題，反而讓勞、資雙方互不信任，勞工無法多加班賺錢，雇主不敢多聘員工，隨後的多次修法更讓國人無所適從，甚至加重民眾的經濟負擔，在這樣的窘境中，民眾連下個月的薪水都不能確定，更別提生養下一代。

不僅如此，民進黨為了總統大選不斷操作反送中、芒果乾等議題，持續恐嚇國人，煽惹人心惶惶，如此讓還敢在台灣養育小孩，拳勸蔡政府別再操弄意識形態恐嚇國人，專心解決少子化危

圖表的誤導?



劉宇
3小時

張善政 副總統候選人你好，

你既然自稱『理工男』就不要玩弄數據好嘛！

你製作的圖表有以下幾個錯誤：

1.蔡英文總統上任日期是105年5月20日，我用綠色直線標示，你的圖表卻從104年1月開始計算？任期整整多了一年半是怎樣 XD

2.你故意把圖表後面的傾斜度畫大，意圖製造生育率大幅降低的假象，但我用**excel**把內政部數據的圖表和你畫的圖表重疊在一起(張善政的圖調成半透明)，再壓縮你的圖讓年份對齊，兩者比較之下就能看出真實數據的後段根本沒那麼傾斜。

3.馬英九總統上任日期是民國97年5月20日，我用紫色直線標示，目前蔡英文任期和馬英九任期兩個區塊比較起來，台灣並沒有顯著的生育率降低。

這樣也敢自稱理工男？自稱數據客觀？嘆ㄘ～

3,722

170則留言
604次分享

讚 留言 分享

留言.....

修正過後的圖表?

張善政
12月14日

#用數據客觀檢視蔡政府的政策
#蔡總統執政後新生兒人數年年下滑
#台灣不需要恐嚇人民的政黨

非常抱歉因製圖錯誤造成紛擾，感謝各界指正，錯誤需要更正是進步的力量，非常感謝各界對於國家出生數的重視及探討，人口是國家國力的象徵，也是我們非常重視的政策，將原圖錯誤更正蔡英文為105年5月20日上任，並補上截至今年十一月統計數字，針對此圖我們想呼籲的是政策需要對症下藥，並不是靠選前丟補助及抄襲政策來修正連年下滑的出生數，國家整體環境需要遠見性的規劃，才能讓國力長遠發展，政策需要傾聽民意、苦民所苦才能讓台灣往更好的方向邁進。

台灣面臨少子化的危機早已不是新聞，但攤開統計數據仍怵目驚心。蔡總統的生育政策做得好不好，不用我們多費唇舌，數據會說話。

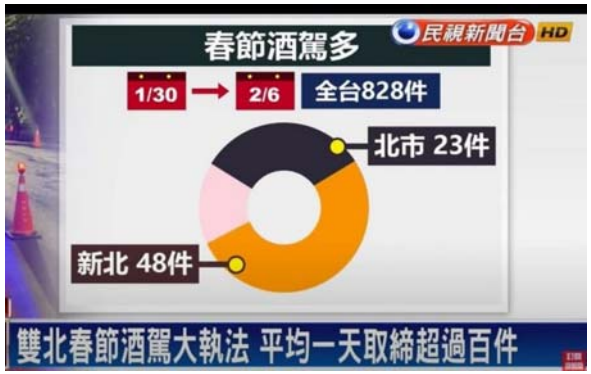
根據內政部的統計，台灣新生兒人數在蔡政府上任後節節下滑，去年僅18萬1千餘人，今年的情況更加嚴峻，截至11月只有16萬零8百人。

蔡政府的生育政策是否對症下藥，還是讓台灣人愈來愈不敢生小孩，經過統計後，很遺憾的是後者。姑且不論蔡政府對少子化毫無作為，也提不出有效的生育政策，只能抄襲國政顧問團的「0到6歲國家幫忙養」，蔡政府在其他方面的政策也是導致少子化的幫兇。

蔡政府上任後提出一例一休，不僅沒有解決結構性的問題，反而讓勞、資雙方互不信任，勞工無法多加班賺錢，雇主不敢多聘員工，隨後的多次修法更讓國人無所適從，甚至加重民眾的經濟負擔，在這樣的窘境中，民眾連下個月的薪水都不

留言.....

更多範例

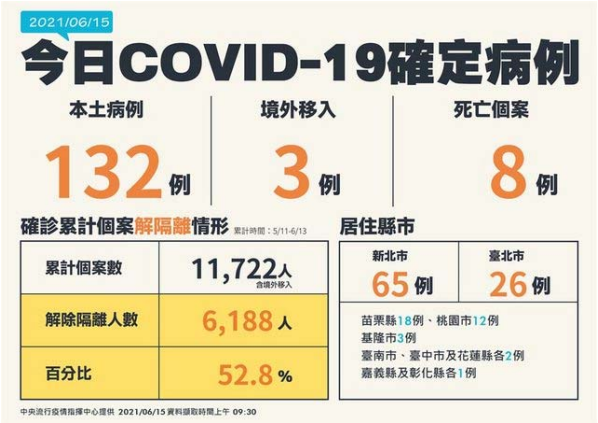


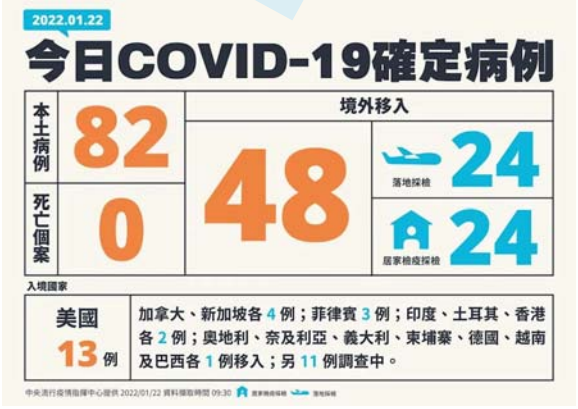
CDC是作圖之鬼嗎?
https://www.ptt.cc/bbs/Gossiping/M.1642833713.A.D40.html
CDC 製圖技術之演進
https://www.ptt.cc/bbs/Gossiping/M.1642920918.A.BD5.html

28.2k 人追蹤 ☆ 追蹤

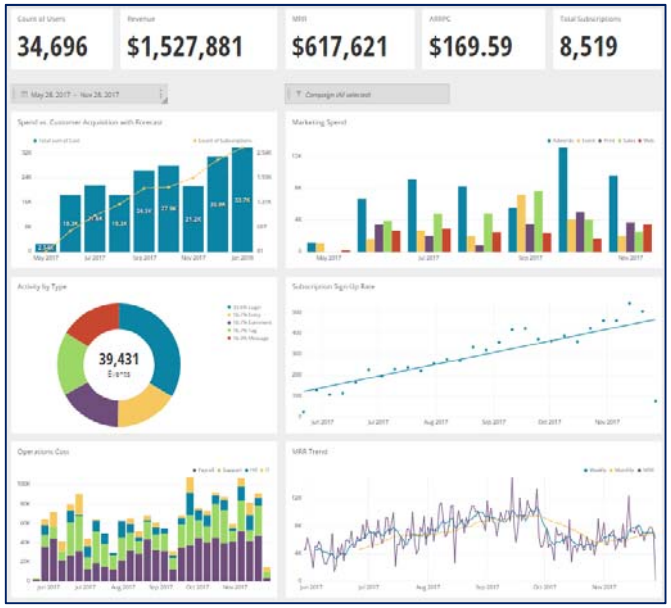
批指揮中心確診圖卡標示不清 鄉民自製被推爆：一目瞭然

吳妍
2022年1月24日 · 3 分鐘 (閱讀時間)

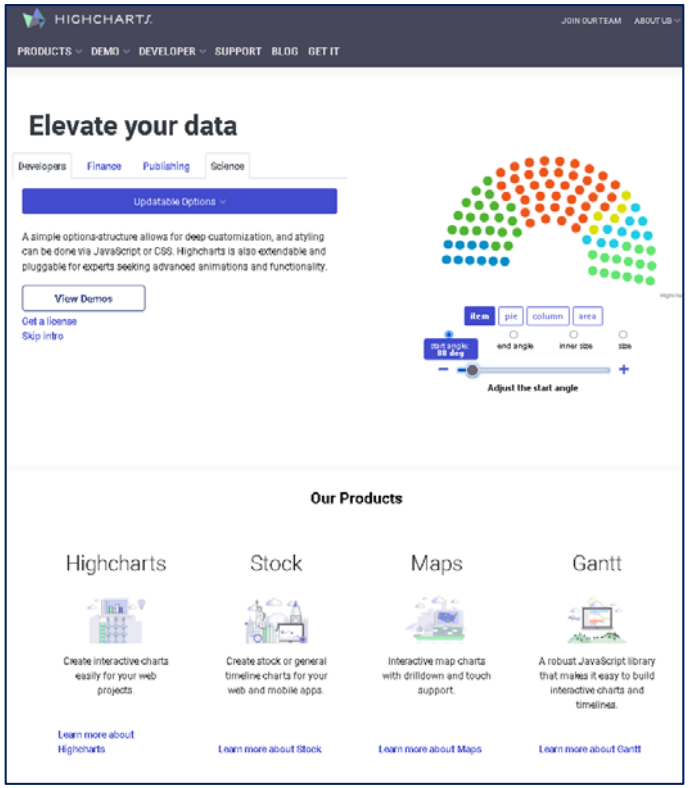


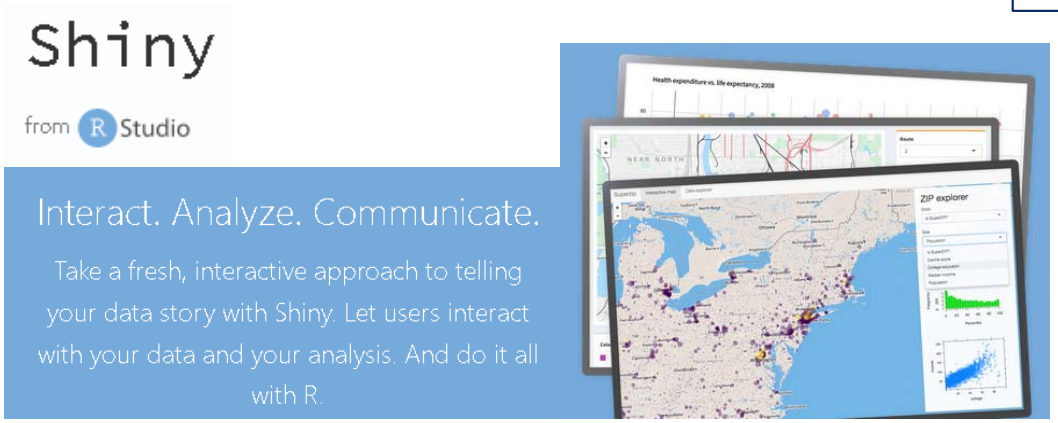


圖表的應用:儀表板 (Dashboard)



sample dashboard from CHARTIO
https://chartio.com/docs/dashboards



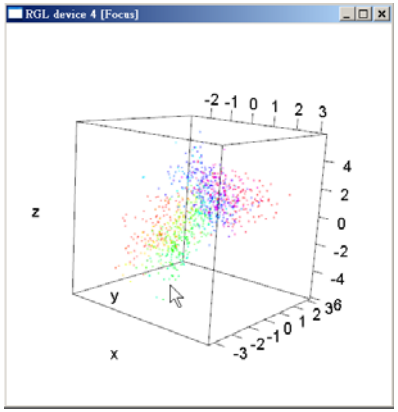


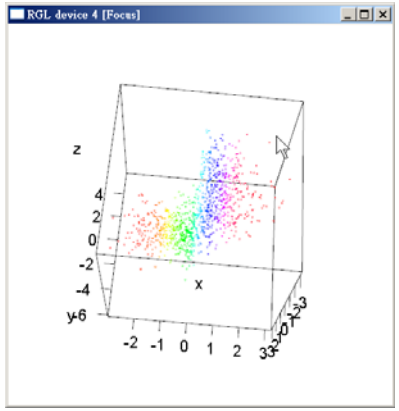
Highcharts - Interactive javascript charts library
https://www.highcharts.com/

3D visualization device system (OpenGL)

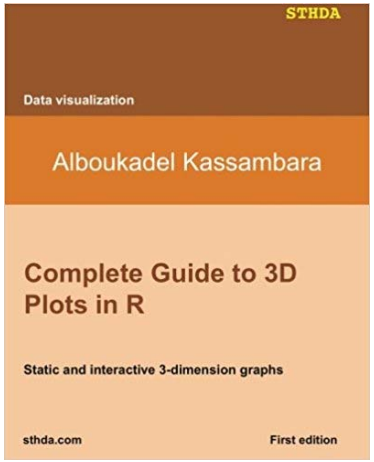
```
> library(rgl)
> demo(rgl)
>
> open3d()
> x <- sort(rnorm(1000))
> y <- rnorm(1000)
> z <- rnorm(1000) + atan2(x,y)
> plot3d(x, y, z, col=rainbow(1000), size=2)
>
> M <- par3d("userMatrix")                                  # 90 degree rotation about the x axis
> play3d(par3dinterp(userMatrix=list(M, rotate3d(M, pi/2, 1, 0, 0),
+                                       rotate3d(M, pi/2, 0, 1, 0))),
+                                       duration=4)
```

```
> M
             [,1]       [,2]       [,3] [,4]
[1,] -0.98849827 -0.1478247 -0.0319229    0
[2,] -0.01036166 -0.1443882  0.9894670    0
[3,] -0.15087697  0.9784170  0.1411958    0
[4,]  0.00000000  0.0000000  0.0000000    1
```





- 截取靜態2d圖: `rgl.postscript {rgl}`, 或 `rgl.snapshot {rgl}`.
- 存出動態圖: `writeWebGL {rgl}`.



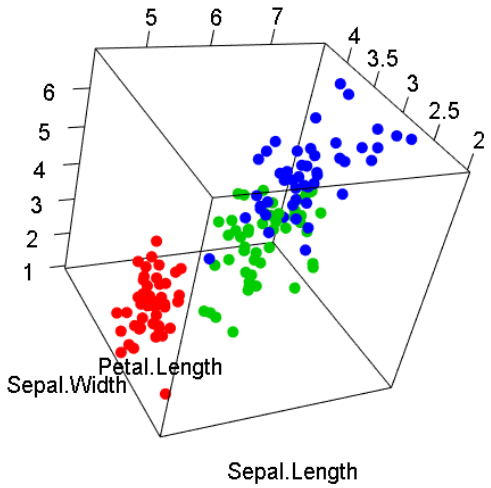
STHDA: rgl

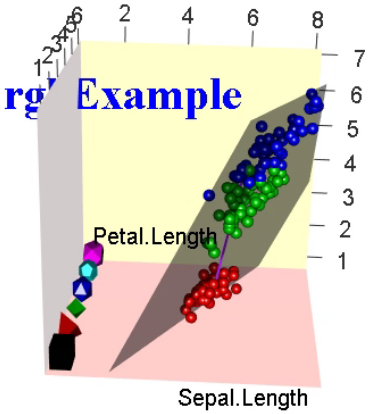
http://www.sthda.com/english/wiki/a-complete-guide-to-3d-visualization-device-system-in-r-r-software-and-data-visualization

http://www.sthda.com/english/download/3-ebooks/6-complete-guide-to-3d-plots-in-r/

範例: rgl

```
> library("rgl")
> open3d()
> plot3d(iris[,1:3], col=as.integer(iris[,5])+1, type ="p", size=10)
```






```
> plot3d(iris[,1:3], col=as.integer(iris[,5])+1, type ="s",
+        radius=0.15)
> bbox3d(color=c("red", "black"), emission="gray",
+        specular="yellow", shininess=5, alpha=0.8, nticks = 3)
> aspect3d(1,1,1)
>
> lines3d(iris[c(1, 150), 1:3], col="purple", lwd=2)
> # points3d, lines3d,segments3d,triangles3d, quads3d.
>
> shapes <- list(cube3d(), tetrahedron3d(), octahedron3d(),
+           icosahedron3d(), dodecahedron3d(), cuboctahedron3d())
> shapelist3d(shapes, x=1, y=1:6, z=1, size=0.3, col=1:6)
> aspect3d(1,1,1)
>
> texts3d(x=2, y=6, z=6, texts="rgl Example", font=2,
+         color="blue", cex=2, family="serif")
>
> # Show regression plane with z as dependent variable
> fit <- lm(iris[,3] ~ iris[,1] + iris[,2])
> coefs <- coef(fit)
> planes3d(a=coefs[2], b=coefs[3], c=-1, d= coefs["(Intercept)"],
+          alpha = 0.5)
> # planes3d draws planes using ax + by + cz + d = 0.
>
> play3d(spin3d(axis = c(0, 0, 1), rpm = 20), duration = 4)
```

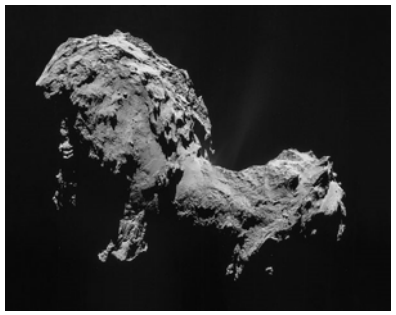
範例: rgl, explore a comet

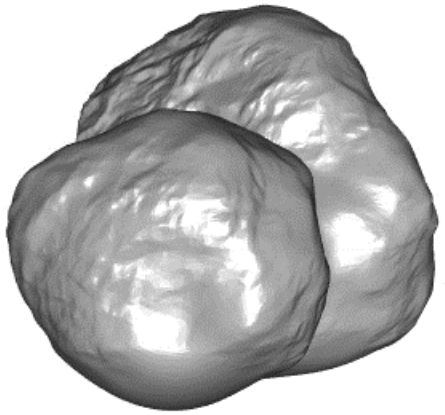
Explore a comet with R's "rgl" package

December 24, 2014

http://blog.revolutionanalytics.com/2014/12/explore-a-comet-with-rs-rgl-package.html

"Last month, the Philae lander touched down on comet Churyumov–Gerasimenko. In the process, the lander and the orbiting Rosetta probe captured detailed data on the geometry of the comet, which the ESA published as a shape file. ..."



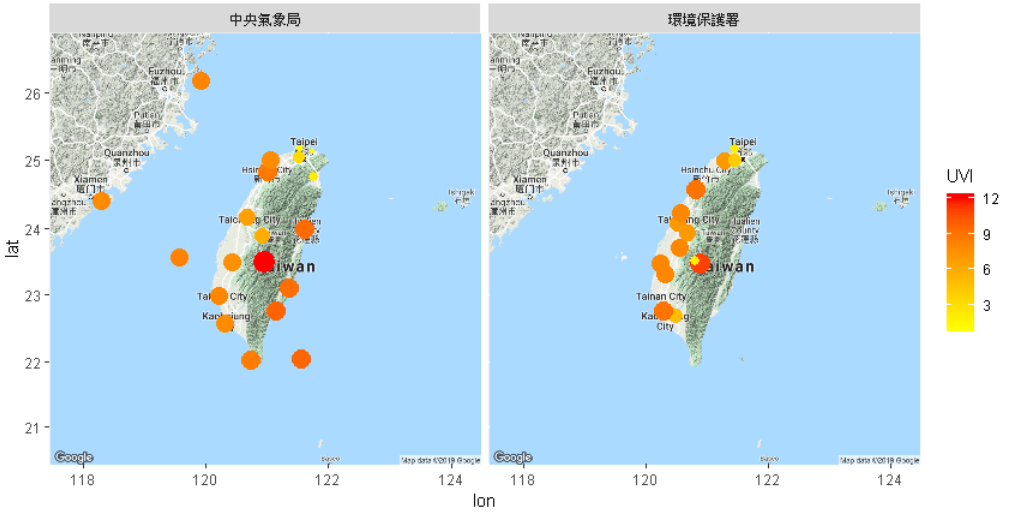


https://en.wikipedia.org/wiki/67P/Churyumov%E2%80%93Gerasimenko

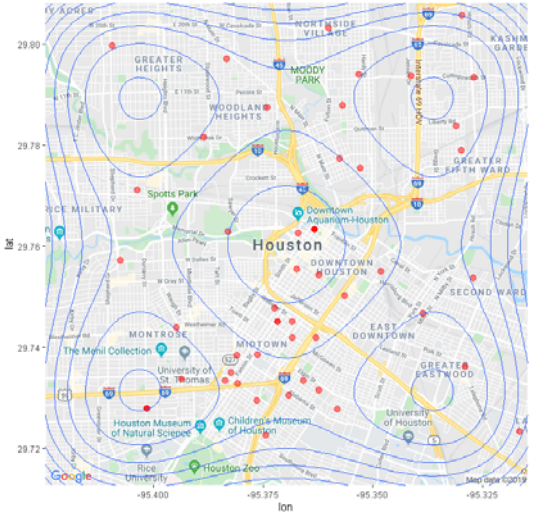
```
> open3d()
> # comet <- readOBJ(url("http://sci.esa.int/science-e/www/object/doc.cfm?fobjectid=54726"))
> comet <- readOBJ("ESA_Rosetta_OSIRIS_67P_SHAP2P.obj")
> class(comet)
[1] "mesh3d"  "shape3d"
> str(comet)
List of 6
 $ vb           : num [1:4, 1:31456] -0.394 0.402 0.443 1 -0.163 ...
 $ it           : num [1:3, 1:62908] 14327 6959 18747 8258 15598 ...
 $ primitivetype: chr "triangle"
 $ material     : NULL
 $ normals      : NULL
 $ texcoords    : NULL
 - attr(*, "class")= chr [1:2] "mesh3d" "shape3d"
> shade3d(comet, col="gray")
```

```
# it: indices for triangular faces
# ib: indices for quad faces
# vb: matrix of vertices: 4xn matrix (rows
x, y, z, h) or equivalent vector, where h
indicates scaling of each plotted quad
```

用R語言畫地圖

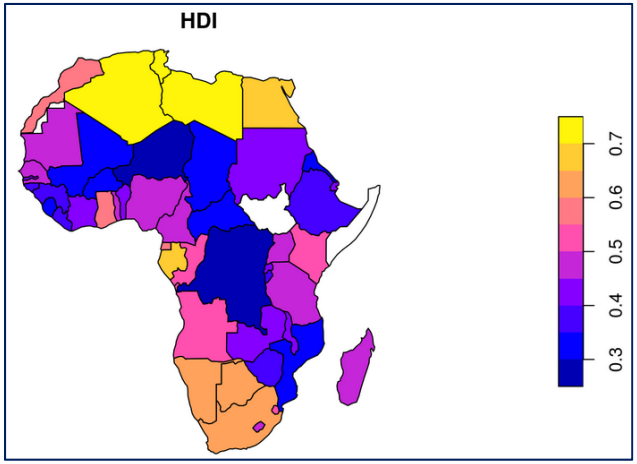


- 若要使用Google 地圖:
  - 需註冊Google Cloud Platform雲端服務
  - 從RStudio測試是否已註冊GCP成功
- 以下僅說明如何畫面量圖(Choropleth Maps)。

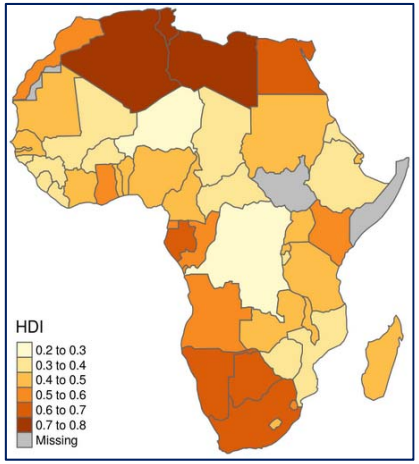


Making maps with R packages

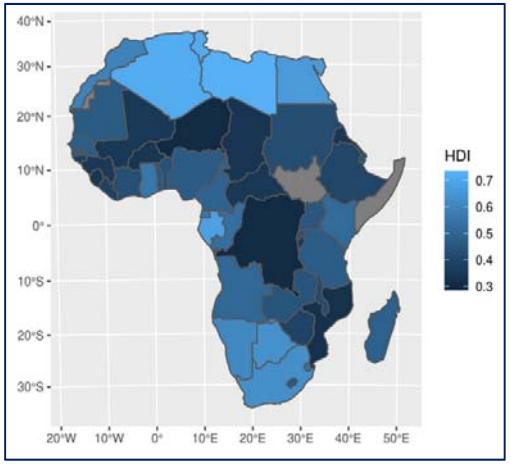
plot



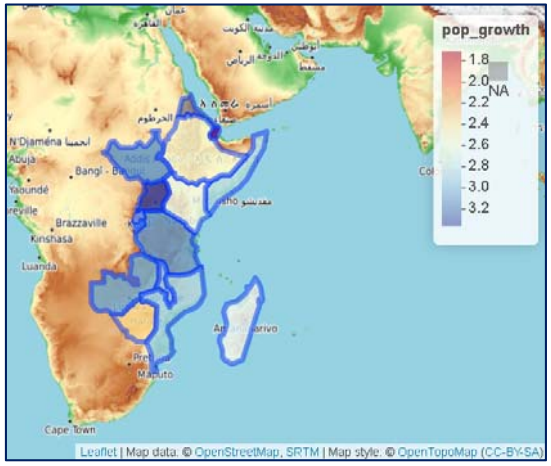
tmap



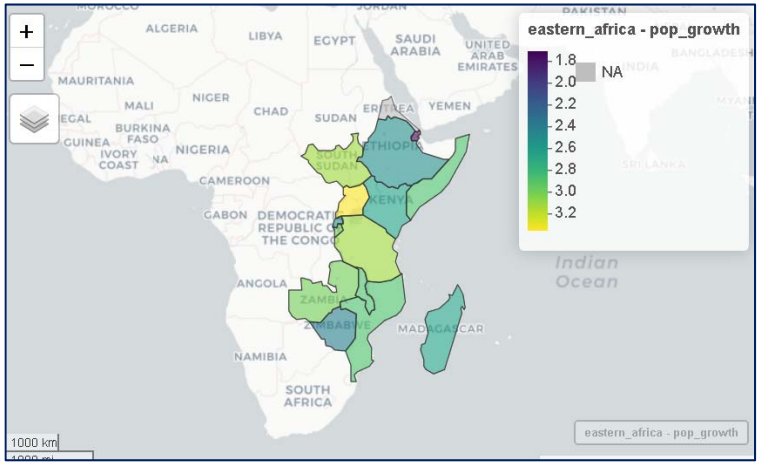
ggplot2



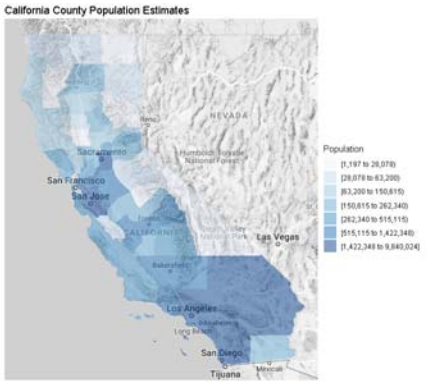
leaflet



mapview



cartography

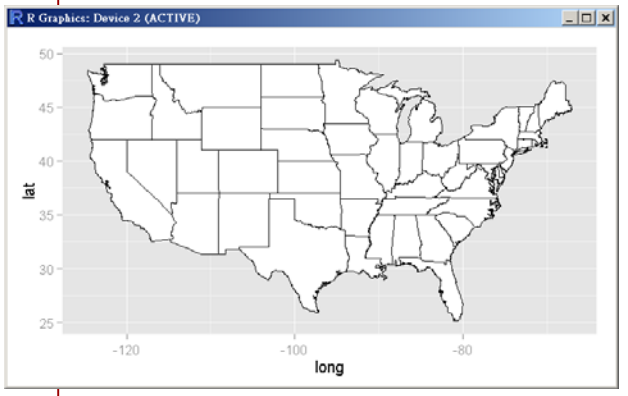


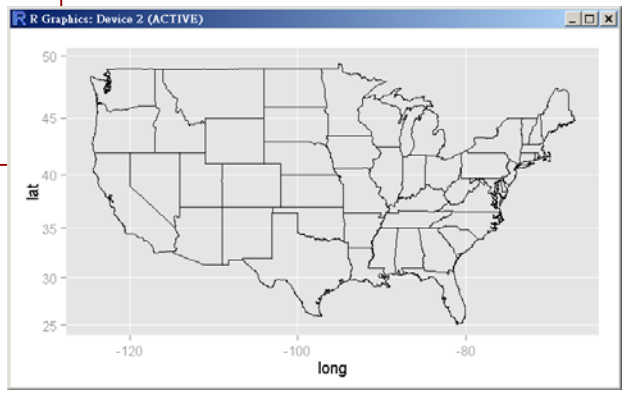
Chapter 8: Making maps with R, Robin Lovelace, Jakub Nowosad, Jannes Muenchow, 2019-08-30
https://geocompr.github.io/geocompkg/articles/solutions08.html

Creating a Map

選讀

```
> library(ggplot2)
> library(maps)
> library(mapproj)
> states.map <- map_data("state")
> head(states.map, 3)
       long      lat group order  region subregion
1 -87.46201 30.38968     1     1 alabama      <NA>
2 -87.48493 30.37249     1     2 alabama      <NA>
3 -87.52503 30.37249     1     3 alabama      <NA>
> tail(states.map, 3)
           long      lat group order  region subregion
15597 -107.9223 41.01805    63 15597 wyoming      <NA>
15598 -109.0568 40.98940    63 15598 wyoming      <NA>
15599 -109.0511 40.99513    63 15599 wyoming      <NA>

> ggplot(states.map, aes(x=long, y=lat, group=group)) +
    geom_polygon(fill="white", colour="black")

> ggplot(states.map, aes(x=long, y=lat, group=group)) +
    geom_path() + coord_map("mercator")
```





mercator: equally spaced straight meridians, conformal, straight compass courses

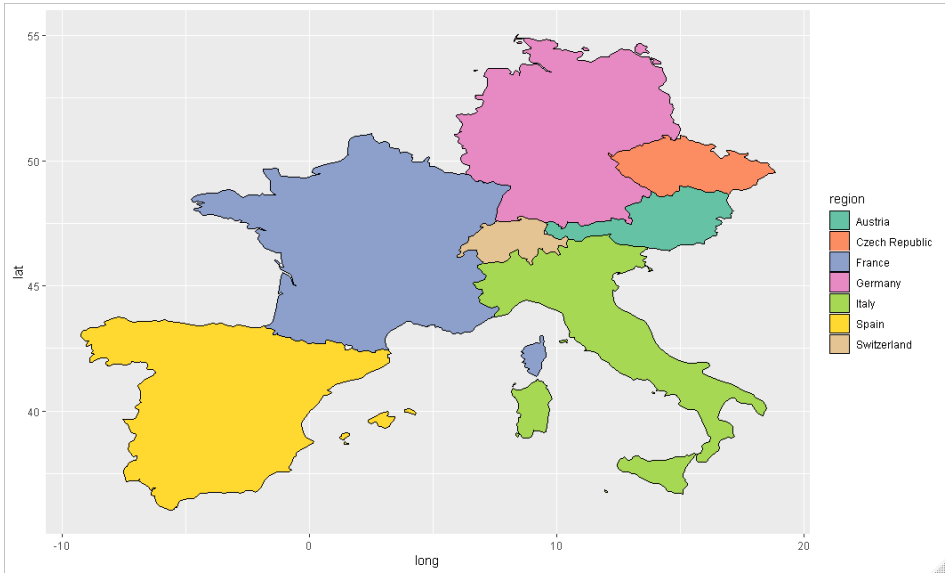
Source: 13.17. Creating a Map, R Graphics Cookbook 2nd

分級著色圖 /面量圖(Choropleth Maps)

選讀

```
> world.map <- map_data("world")
> sort(unique(world.map$region))
  [1] "Afghanistan"                "Albania"
  [3] "Algeria"                    "American Samoa"
  [5] "Andaman Islands"            "Andorra"
   ...
> country <- c("France", "Austria", "Italy", "Switzerland", "Germany", "Spain", "Czech
Republic")
> mymapdata <- map_data("world", region = country)
> ggplot(mymapdata, aes(x = long, y = lat, group = group, fill = region)) +
   geom_polygon(colour = "black") +
   scale_fill_brewer(palette = "Set2")
```

```
> head(mymapdata)
      long      lat group order  region subregion
1 16.95312 48.59883     1     1 Austria      <NA>
2 16.94883 48.58858     1     2 Austria      <NA>
3 16.94336 48.55093     1     3 Austria      <NA>
4 16.90449 48.50352     1     4 Austria      <NA>
5 16.86270 48.44141     1     5 Austria      <NA>
6 16.86543 48.38691     1     6 Austria      <NA>
> tail(mymapdata)
          long      lat group order  region subregion
2941 6.734766 53.58252    26  2941 Germany    Borkum
2942 6.642090 53.57920    26  2942 Germany    Borkum
2943 6.668555 53.60566    26  2943 Germany    Borkum
2944 6.754590 53.62549    26  2944 Germany    Borkum
2945 6.800879 53.62549    26  2945 Germany    Borkum
2946 6.734766 53.58252    26  2946 Germany    Borkum
```



NOTE:

See the **mapdata** package for more map data sets. It includes maps of China and Japan, as well as a high-resolution world map, **worldHires**.

See Also: **mapproject, map**

Violent Crime Rates by US State (USArrests)

選讀

Violent Crime Rates by US State (USArrests): the data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

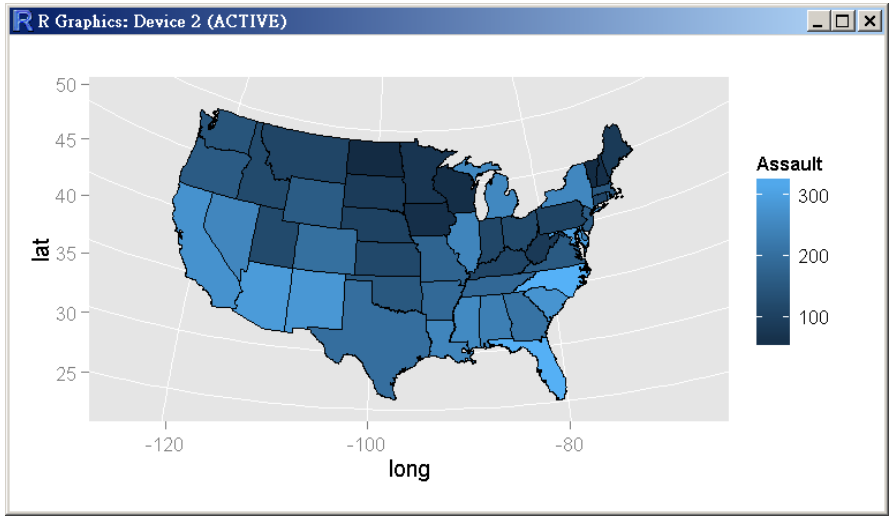
```
> head(USArrests, 3)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
> crimes <- data.frame(state = tolower(rownames(USArrests)), USArrests)
> head(crimes, 3)
               state Murder Assault UrbanPop Rape
Alabama      alabama   13.2     236       58 21.2
Alaska        alaska   10.0     263       48 44.5
Arizona      arizona    8.1     294       80 31.0

> library(maps); library(ggmap)
> states.map <- map_data("state")
> head(states.map, 3)
       long      lat group order  region subregion
1 -87.46201 30.38968     1     1 alabama      <NA>
2 -87.48493 30.37249     1     2 alabama      <NA>
3 -87.52503 30.37249     1     3 alabama      <NA>
> crime.map <- merge(states.map, crimes, by.x="region", by.y="state")
> head(crime.map, 3)
   region      long      lat group order subregion Murder Assault UrbanPop Rape
1 alabama -87.46201 30.38968     1     1      <NA>   13.2     236       58 21.2
2 alabama -87.48493 30.37249     1     2      <NA>   13.2     236       58 21.2
3 alabama -87.95475 30.24644     1    13      <NA>   13.2     236       58 21.2
```

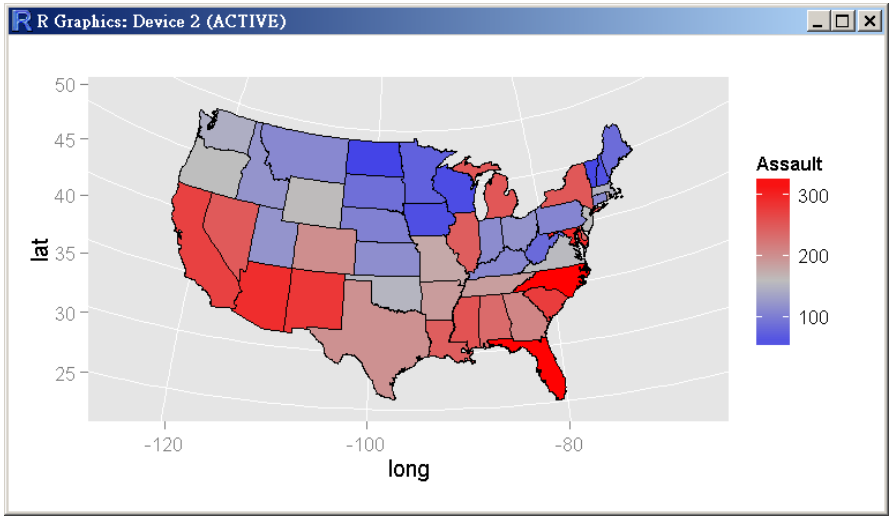
分級著色圖 /面量圖(Choropleth Maps)

選讀

```
> library(plyr)
> crime.map <- arrange(crime.map, group, order)
> head(crime.map, 3)
   region      long      lat group order subregion Murder Assault UrbanPop Rape
1 alabama -87.46201 30.38968     1     1      <NA>   13.2     236       58 21.2
2 alabama -87.48493 30.37249     1     2      <NA>   13.2     236       58 21.2
3 alabama -87.52503 30.37249     1     3      <NA>   13.2     236       58 21.2
> ggplot(crime.map, aes(x=long, y=lat, group=group, fill=Assault)) +
        geom_polygon(colour="black") +
        coord_map("polyconic")
```

```
ggplot(crime.map, aes(x=long, y=lat, group=group, fill=Assault)) +
    geom_polygon(colour="black") +
    coord_map("polyconic") +
    scale_fill_gradient2(low="blue", mid="grey", high="red",
                         midpoint=median(crimes$Assault))
```





維度縮減
Dimension Reduction

keep information as much as possible without loss of information.

input data matrix: X

Group	Data	X1	X2	X3	•••	Xp
1	subject01	0.81	-1.29	-0.50		1.13
1	subject02	0.64	2.16	-1.51		0.00
2	subject03	0.13	0.60	1.10		0.11
2	subject04	-0.17	0.31	-0.37		-0.50
3	subject05	-1.01	0.99	0.70		-0.08
3	subject06	0.95	0.75	-0.83		0.60
3	subject07	0.72	1.12	-1.35		-1.22
1	subject08	0.77	1.24	-0.04		1.03
1	subject09	-0.49	0.02	-1.73		1.61
1	subject10	1.93	0.45	-0.01		0.03
3	subject11	-0.15	-1.36	1.05		0.50
3	subject12	-1.16	0.11	-0.57		-0.80
3	subject13	-0.02	2.05	-1.18		0.45
3	subject14	-0.05	0.79	1.33		0.81
2	subject15	-0.21	-0.38	0.72		-0.61
1	subject16	-0.28	0.57	1.02		-0.01
⋮	⋮					
2	subjectN	0.33	0.01	1.19		-0.33

PCA
FA
MDS

transformed data matrix: Z

Data	Z1	Z2	•••	Zk
subject01	-1.55	-0.66		0.60
subject02	0.57	-0.51		-1.03
subject03	1.99	1.44		-0.60
subject04	0.10	0.20		-1.21
subject05	-0.20	-0.64		0.24
subject06	2.85	1.32		-0.61
subject07	-0.34	-0.35		0.15
subject08	0.66	0.44		0.28
subject09	8.44	1.66		2.12
subject10	0.37	-0.17		-1.73
subject11	-1.14	0.01		0.61
subject12	-1.73	-1.13		0.81
subject13	1.53	0.67		0.48
subject14	-0.21	-0.14		-0.29
subject15	-0.03	0.66		0.17
subject16	2.56	-2.25		0.26
⋮				
subjectN	2.04	0.71		0.76

- Visualization
- Clustering
- Classification
- ....

Y

Methods using additional information *y:* LDA, Sufficient Dimension Reduction (SIR, SAVE, pHd, IRE,...)

Why Reduce Dimensionality?

- **Reduces time complexity**:
  less computation
- **Reduces space complexity**:
  Less parameters
- **Saves the cost** of observing the features:
  input is unnecessary.
- **Simpler models are more robust on small datasets**:
  simpler models vary less depending on the particulars of a sample.
- **More interpretable; simpler explanation**
- **Data visualization** (structure, groups, outliers, etc) if plotted in 2 or 3 dimensions. (Dimension reduction visualization is often adopted for presenting grouping structure for methods such as K-means.)

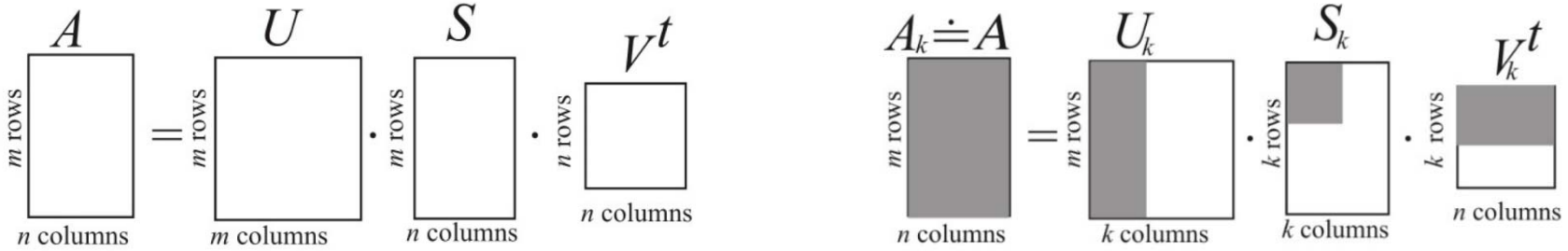
Singular Value Decomposition

奇異值分解

Singular Value Decomposition (SVD) factorizes a general $m \times n$ (assume $m \geq n$) matrix in the form

$$A = USV^t,$$

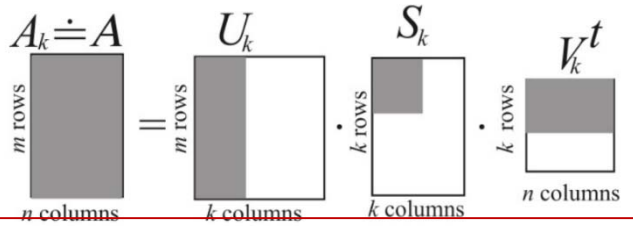
where U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, and S is an $m \times n$ matrix whose only nonzero elements lie along the main diagonal.



- The m columns of $\boldsymbol{U}$ are called the left singular vectors. The n columns of $\boldsymbol{V}$ are called the right singular vectors.
  - The right singular vectors are eigenvectors of $\boldsymbol{A^TA}$.
  - The left singular vectors are eigenvectors of $\boldsymbol{AA^T}$.
  - The non-zero values of $\boldsymbol{S}$ are the square root of the eigenvalues of $\boldsymbol{AA^T}$ and $\boldsymbol{A^TA}$ are called the singular values
- $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrixes, i.e. $\boldsymbol{UU^T=I}$, $\boldsymbol{VV^T=I}$.

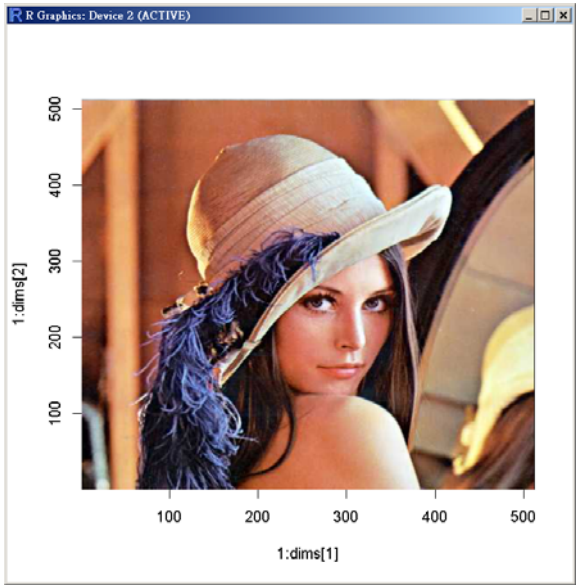
SVD: Making Approximations

```
> iris.sub <- iris[sample(1:150, 8),1:4]
> iris.sub
   Sepal.Length Sepal.Width Petal.Length Petal.Width
58          4.9         2.4          3.3         1.0
59          6.6         2.9          4.6         1.3
77          6.8         2.8          4.8         1.4
84          6.0         2.7          5.1         1.6
54          5.5         2.3          4.0         1.3
39          4.4         3.0          1.3         0.2
71          5.9         3.2          4.8         1.8
87          6.7         3.1          4.7         1.5
> M.svd <- svd(iris.sub)
> M.svd
$d
[1] 22.2553197  2.2360595  0.7611307  0.1773813

$u
           [,1]         [,2]         [,3]        [,4]
[1,] -0.2898553 -0.096733080 -0.007319459 -0.05013322
[2,] -0.3885172 -0.006908608  0.305429166 -0.28018114
[3,] -0.3992219  0.066823762  0.450619905  0.06954991
[4,] -0.3793864  0.327663490 -0.174978375 -0.69121320
[5,] -0.3275406  0.106859076  0.218964907  0.50428246
[6,] -0.2285027 -0.918407065 -0.136688043 -0.13248141
[7,] -0.3782377  0.154232634 -0.776519488  0.27425130
[8,] -0.3989561 -0.009386179  0.058068514  0.29884180

$v
           [,1]       [,2]       [,3]       [,4]
[1,] -0.7497994 -0.3154600  0.5318218  0.2354811
[2,] -0.3527468 -0.5480151 -0.7276317 -0.2140122
[3,] -0.5343751  0.7023730 -0.1376906 -0.4496184
[4,] -0.1667746  0.3268587 -0.4108028  0.8346201
```




```
> M.svd$u %*% (diag(M.svd$d) %*% t(M.svd$v))
     [,1] [,2] [,3] [,4]
[1,]  4.9  2.4  3.3  1.0
[2,]  6.6  2.9  4.6  1.3
[3,]  6.8  2.8  4.8  1.4
[4,]  6.0  2.7  5.1  1.6
[5,]  5.5  2.3  4.0  1.3
[6,]  4.4  3.0  1.3  0.2
[7,]  5.9  3.2  4.8  1.8
[8,]  6.7  3.1  4.7  1.5
>
> # use the first two values to approximate
> d.sub <- diag(M.svd$d[1:2])
> u.sub <- as.matrix(M.svd$u[, 1:2])
> v.sub <- as.matrix(M.svd$v[, 1:2])
> iris.sub.approx <- u.sub %*% d.sub %*% t(v.sub)
> iris.sub.approx
         [,1]     [,2]     [,3]      [,4]
[1,] 4.905057 2.394043 3.295235 1.0051334
[2,] 6.488070 3.058517 4.609664 1.4369796
[3,] 6.614690 3.052204 4.852772 1.5306008
[4,] 6.099701 2.576853 5.026535 1.6476201
[5,] 5.390302 2.440411 4.063166 1.2938078
[6,] 4.460863 2.919270 1.275109 0.1768745
[7,] 6.202869 2.780357 4.740493 1.5166003
[8,] 6.664012 3.143504 4.729919 1.4739142
>
> # compute the sum of squared errors
> sum((iris.sub - iris.sub.approx)^2)
[1] 0.610784
```

SVD: Image Compression (1/3)



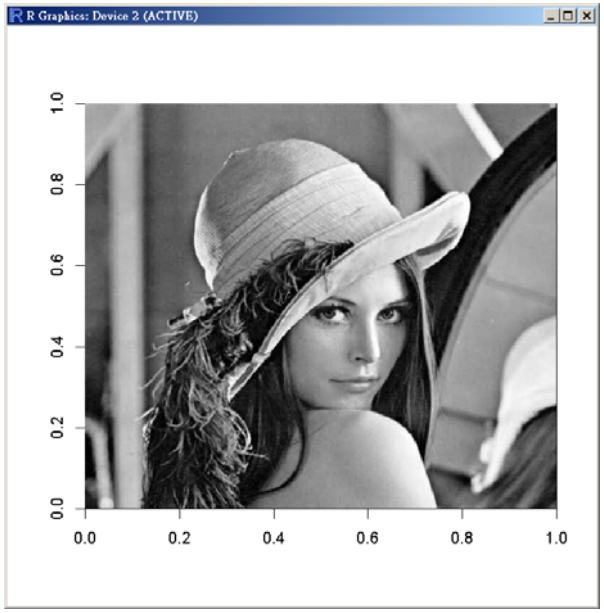
Lenna 97: A Complete Story of Lenna
http://www.ee.cityu.edu.hk/~lmpo/lenna/Lenna97.html
https://en.wikipedia.org/wiki/Lenna
This scan became one of the most used images in computer history.[4] In a 1999 issue of *IEEE Transactions on Image Processing* "Lena" was used in three separate articles,[5] and the picture continued to appear in scientific journals throughout the beginning of the 21st century. ... To explain Lenna's popularity, David C. Munson, editor-in-chief of IEEE Transactions on Image Processing, noted that it was a good test image because of its detail, flat regions, shading, and texture. However, he also noted that its popularity was largely because an image of an attractive woman appealed to the males in a male-dominated field

```
> # require packages: locfit, tiff, fftwtools
> library(EBImage) # (Repositories: BioC Software)
> lena <- readImage("lena.jpg")
> dims <- dim(lena)
> dims
[1] 512 512   3
>
> plot(c(0, dims[1]), c(0, dims[2]), type='n', xlab="", ylab="")
> rasterImage(lena, 0, 0, dims[1], dims[2])
```

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("EBImage")
```

```
> library(jpeg) # install.packages("jpeg")
> lena <- readJPEG("lena.jpg")
```

SVD: Image Compression (2/3)




```
> lena.flip <- Image(flip(lena))
> # convert RGB to grayscale
> red.weight   <- .2989
> green.weight <- .587
> blue.weight  <- 0.114
>
> lena.gray <- red.weight   * imageData(lena.flip)[,,1] +
+              green.weight * imageData(lena.flip)[,,2] +
+              blue.weight  * imageData(lena.flip)[,,3]
> dim(lena.gray)
[1] 512 512
> lena.gray[1:5, 1:5]
           [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.07128863 0.06344627 0.05952510 0.06736745 0.07913098
[2,] 0.07913098 0.06736745 0.06344627 0.07128863 0.08305216
[3,] 0.08697333 0.07520980 0.07128863 0.07913098 0.09089451
[4,] 0.09089451 0.08305216 0.07913098 0.08305216 0.09873686
[5,] 0.09570980 0.08786745 0.08002510 0.08786745 0.10355216
> image(lena.gray, col = grey(seq(0, 1, length = 256)))
```

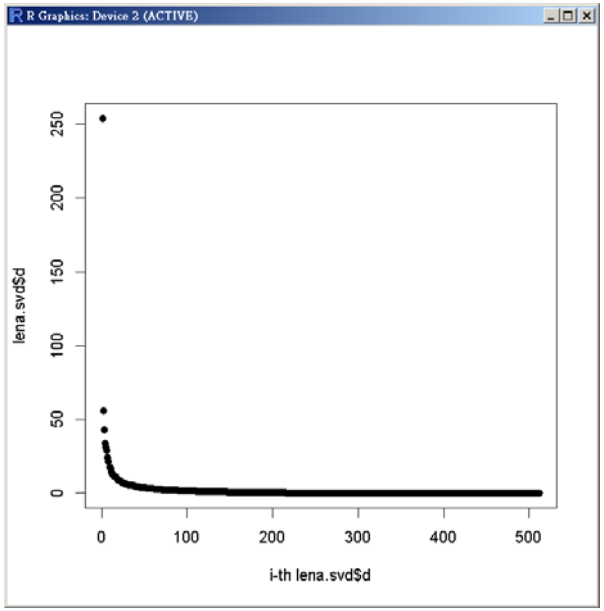
In 1972, at the age of 21. 67-year-old

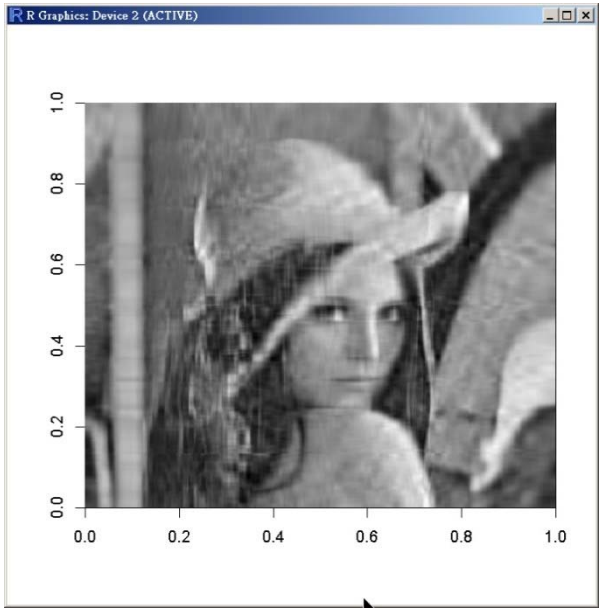
Converting RGB to grayscale/intensity
http://stackoverflow.com/questions/687261/converting-rgb-to-grayscale-intensity

工程師都愛的萊娜小姐　究竟是誰？(林一平 2019-12-06)
https://www.digitimes.com.tw/col/article.asp?id=1129

SVD: Image Compression (3/3)

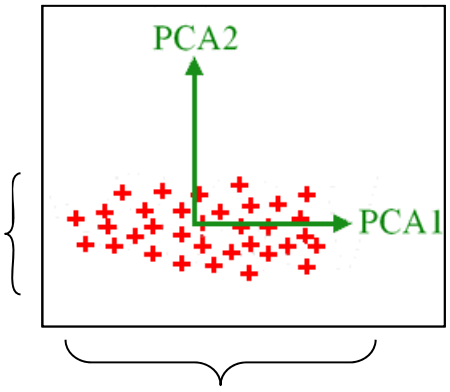
```
> lena.svd <- svd(lena.gray)
> d <- diag(lena.svd$d)
> dim(d)
[1] 512 512
> u <- lena.svd$u
> v <- lena.svd$v
> plot(1:length(lena.svd$d), lena.svd$d, pch=19, xlab="i-th lena.svd$d", ylab="lena.svd$d")
>
> used.no <- 20
> u.sub <- as.matrix(u[, 1:used.no])
> v.sub <- as.matrix(v[, 1:used.no])
> d.sub <- as.matrix(d[1:used.no, 1:used.no])
> lena.approx <- u.sub %*% d.sub %*% t(v.sub)
> image(lena.approx, col = grey(seq(0, 1, length = 256)))
```





Principal Component Analysis

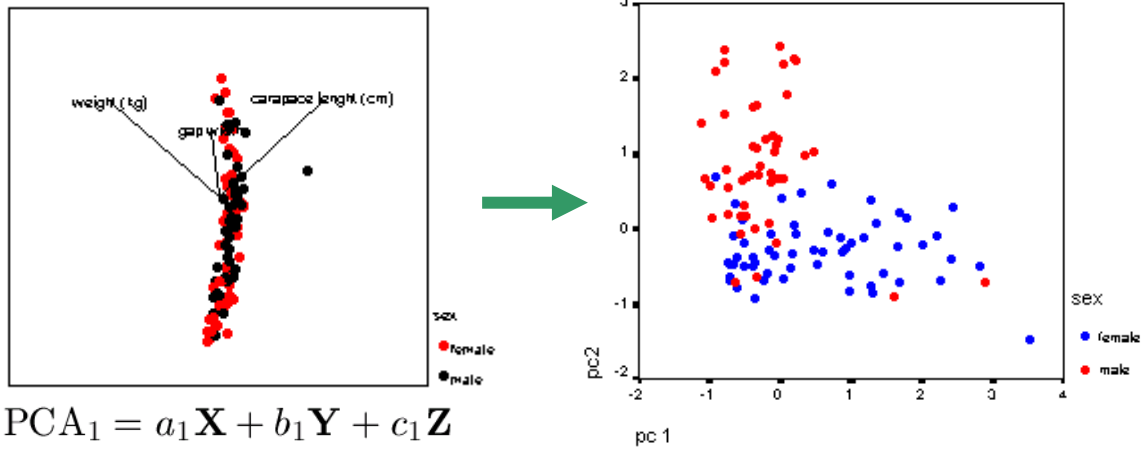
PCA is a method that reduces data dimensionality by finding the new variables (major axes, principal components).



$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y}$$

Image source: 61BL4165 Multivariate Statistics, Department of Biological Sciences, Manchester Metropolitan University



$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y} + c_1\mathbf{Z}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y} + c_2\mathbf{Z}$$

$$\mathrm{PCA}_1 = a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + \cdots + a_{1p}\mathbf{X}_p$$

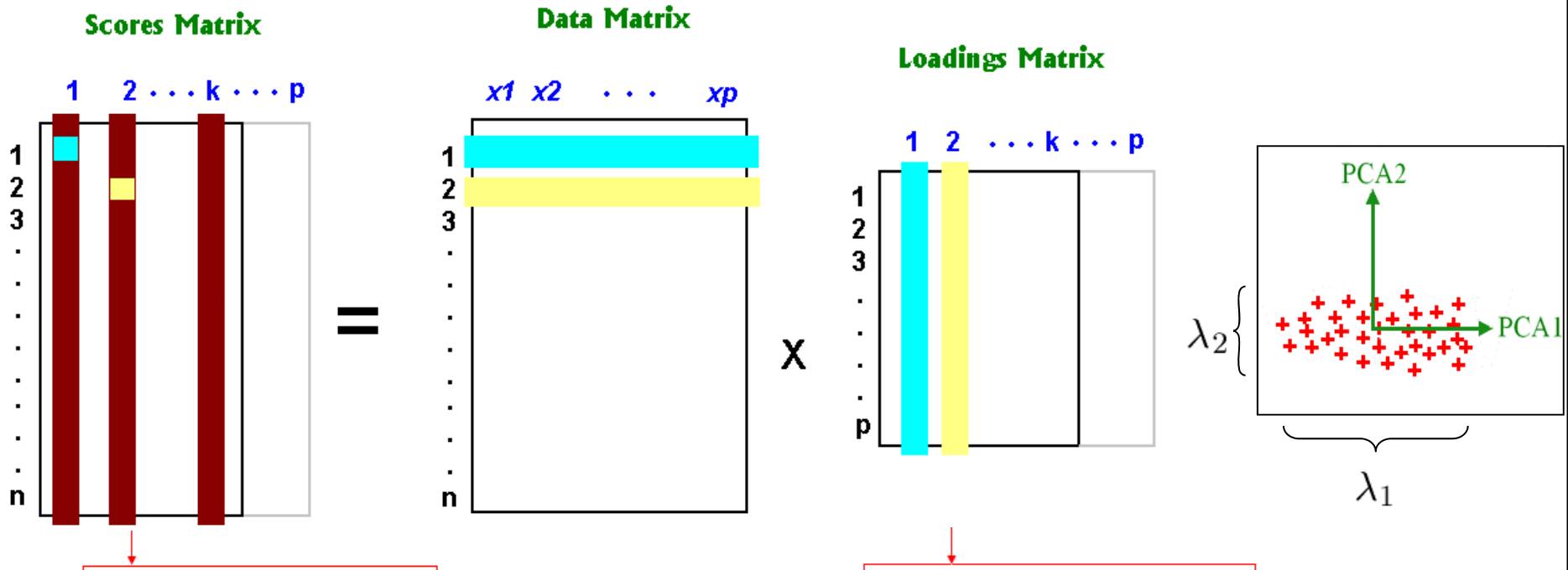
$$\mathrm{PCA}_2 = a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \cdots + a_{2p}\mathbf{X}_p$$

Karl Pearson 1901; Hotelling 1933; Jolliffe 2002

Amongst all possible projections, PCA finds the projections so that the maximum amount of information, measured in terms of variability, is retained in the smallest number of dimensions.

PCA: Loadings and Scores

$$Z = X \quad W$$



The *i*th principal component of $\mathbf{X}$ is $\mathbf{X}\mathbf{w}_i$, where $\mathbf{w}_i$ is the *i*th normalized eigenvector of $\Sigma_{\mathbf{x}}$ corresponding to the *i*th largest eigenvaules.

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

$$\text{proportion} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

PCA is the spectral decomposition of the variance covariance matrix, ***S***, of the data matrix, ***X***. we can write $S=CDC^T$.

PCA: General Methodology

- Find a low-dimensional space such that when x is projected there, information loss is minimized.
- PCA centers the sample and then rotates the axes to line up with the directions of highest variance.

The projection of $\boldsymbol{x}$ on the direction of $\boldsymbol{w}$ is: $z = \boldsymbol{w}^T\boldsymbol{x}$

Find $\boldsymbol{w}$ such that var(z) is maximized

$$\begin{aligned}\text{var}(z) &= \text{var}(\boldsymbol{w}^T\boldsymbol{x}) \\ &= E[(\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{\mu})^2] \\ &= E[\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})^2\boldsymbol{w}] \\ &= E[\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{w}] \\ &= \boldsymbol{w}^T E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]\boldsymbol{w}\end{aligned}$$

$$\begin{aligned}\text{var}(\boldsymbol{x}) &= E[(\boldsymbol{x} - \boldsymbol{\mu})^2] \\ &= E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] \\ &= \boldsymbol{\Sigma}\end{aligned}$$

```
?cov
x <- iris[, 1:4]
cov(x)
```

PCA: General Methodology

$\text{var}(z) = w^T \Sigma w$ Maximize var(z) subject to $||w||=1$

$\mathcal{L}(\lambda, w_1) = w_1^T \Sigma w_1 - \lambda(w_1^T w_1 - 1)$ **Lagrangian method**

拉格朗日

$$\max_{w_1} \mathcal{L}(\lambda, w_1) \rightarrow \begin{cases} \frac{\partial \mathcal{L}(\lambda, w_1)}{\partial \lambda} = 0 \rightarrow w_1^T w_1 - 1 = 0 \\ \frac{\partial \mathcal{L}(\alpha, w_1)}{\partial w_1} = 0 \rightarrow 2\Sigma w_1 - 2\lambda w_1 = 0 \end{cases}$$

$\rightarrow \Sigma w_1 = \lambda w_1$

$\rightarrow$ w_1 is an eigenvector of Σ

λ is an eigenvalue associated with w_1

```
x <- iris[, 1:4]
(covx <- cov(x))
e <- eigen(covx)
V <- e$vectors
V.inverse <- solve(e$vectors)
covx.hat <- V %*% diag(e$values) %*% V.inverse
covx.hat # same with covx
```

$$\text{var}(z) = \text{var}(w_1^T x) = w_1^T \Sigma w_1 = w_1^T \lambda w_1 = \lambda w_1^T w_1 = \lambda$$

$\rightarrow \max \text{var}(z) = \max \lambda$

How Many Components to Use?

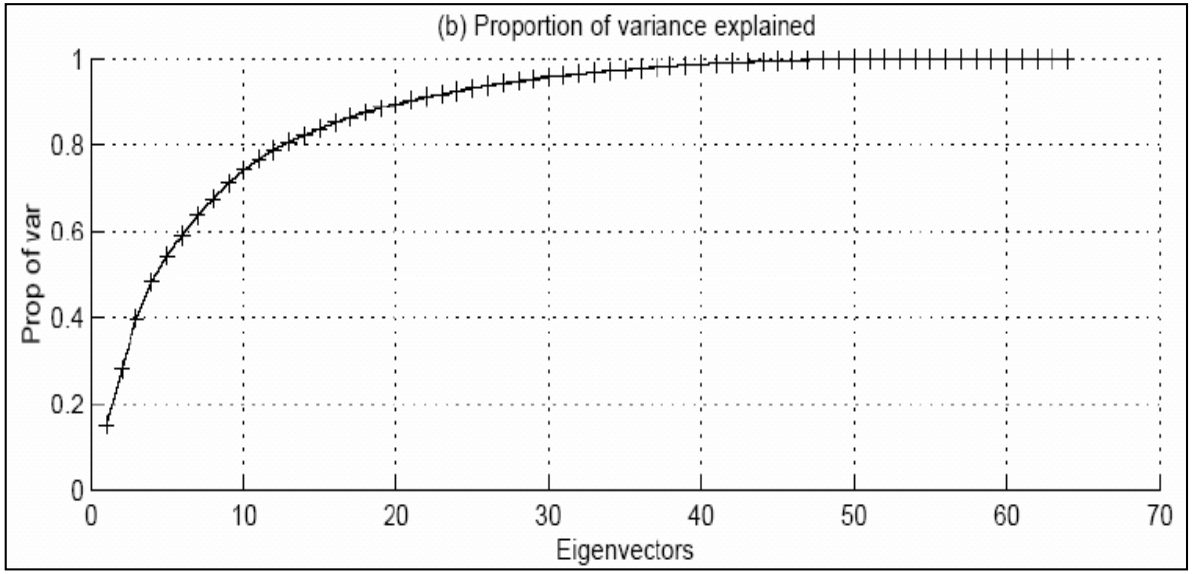
- The proportion of Variance explained by the first k principal components

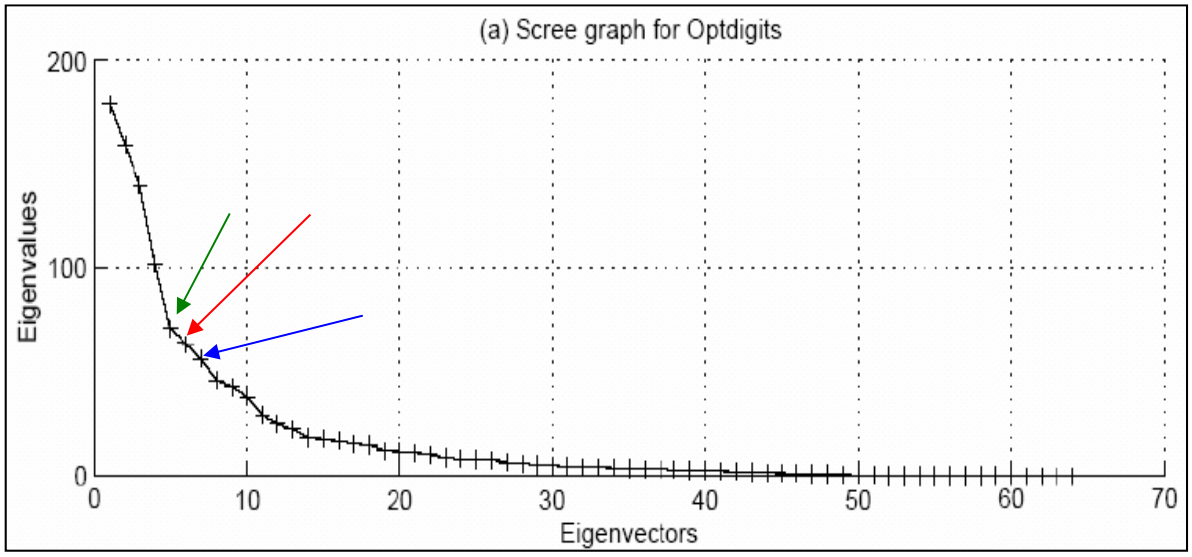
$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_d} \qquad \frac{\lambda_k}{\sum_{j=1}^{r} \lambda_j} = \frac{\lambda_k}{\mathrm{tr}(\Sigma)}.$$

- If $\lambda_j < 1$ then component explains less variance than original variable (correlation matrix).
- Use 2 components (or 3) for visual ease.
- Keep only the eigenvectors with eigenvalues greater than **average eigenvalue**.
- Keep only those which have higher than the **average input variance**.

Scree Diagram

At the elbow, adding another eigenvector does not significantly increase the variance explained.





PCA: Potential Problems

- Sensitive to outliers
  - Robust estimation
  - Discarding the isolated data points that are far away.
- Lack of Independence
  - NO PROBLEM
- Lack of Normality
  - Normality desirable but not essential
- Lack of Precision
  - Precision desirable but not essential
- Many Zeroes in Data Matrix
  - Problem (use Correspondence Analysis)

This works better than taking the average because we take into account correlations and differences in variances.

Example:

average

	A	B	C	D	E	F	平均	PCA1
1	學號	小考1	期中考	小考2	期末考	報告		
2	A01	69	92	85	45	62		
3	A02	66	90	83	36	90		
4	A03	72	92	80	62	70		
5	A04	68	90	60	37	95		
6	A05	74	60	86	54	70		
7	A06	77	90	88	88	95		
8	A07	73	88	77	51	95		
9	A08	61	90	84	40	82		
10	A09	66	88	82	39	80		
11	A10	76	75	87	72	80		
12	A11	64	90	90	26	95		
13	A12	75	90	60	55	70		
14	A13	92	90	83	90	95		

Circle of Correlations

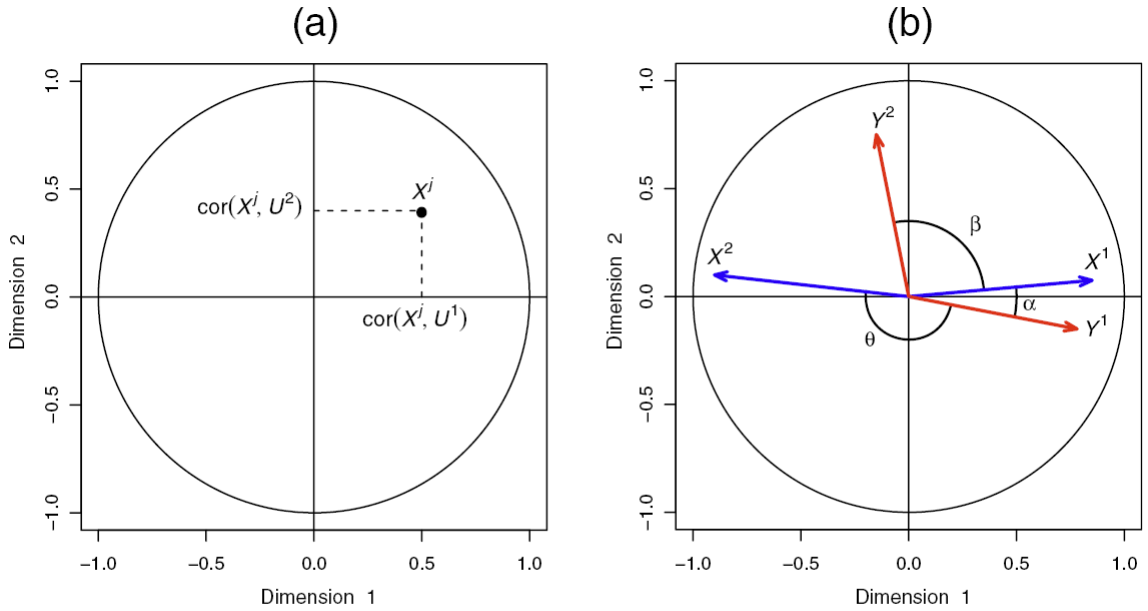


Figure 1 Correlation Circle plot. a) Coordinates of the X-variables on the plane defined by the first two variates U^1 and U^2. **b)** The correlation between two variables is positive if the angle is sharp $\cos(\alpha) > 0$, negative if the angle is obtuse $\cos(\theta) < 0$, and null if the vectors are perpendicular $\cos(\beta) \approx 0$.

González *et al. BioData Mining* 2012, **5**:19
http://www.biodatamining.org/content/5/1/19

the input variables, X_j,

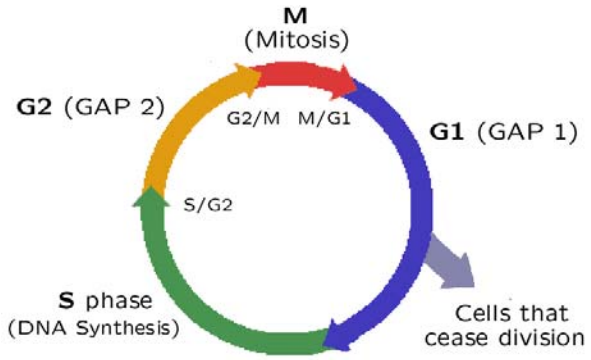
the DR components, Z_k,

$$\rho_{jk} = \mathrm{Corr}(X_j; Z_k) = \frac{\mathrm{Cov}(X_j, Z_k)}{\sqrt{\mathrm{Var}(X_j)\mathrm{Var}(Z_k)}}.$$

Circles of correlations associated with the first two DR components based on the applied vector and path point approaches applied to the microarray data of the yeast cell cycle.

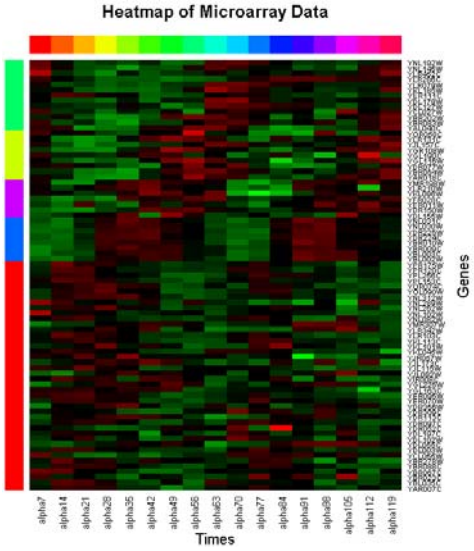
範例: Microarray Data of Yeast Cell Cycle

- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)
- 103 known genes: every 7 minutes and totally 18 time points. (remove NA's: 79 genes)



Microarray Data Matrix

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

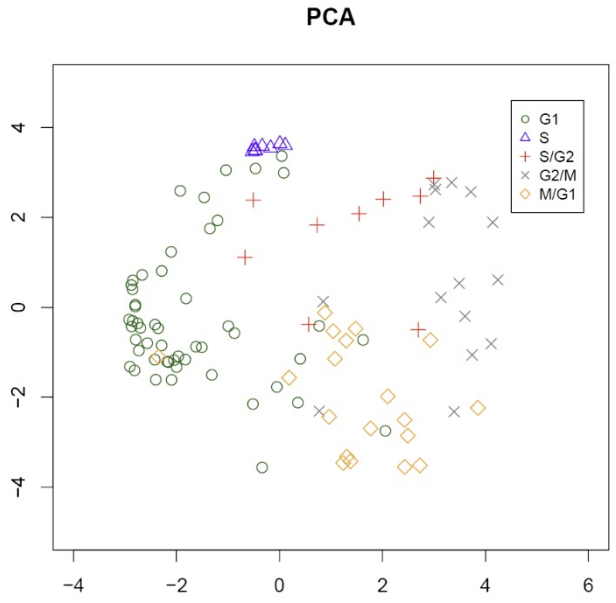


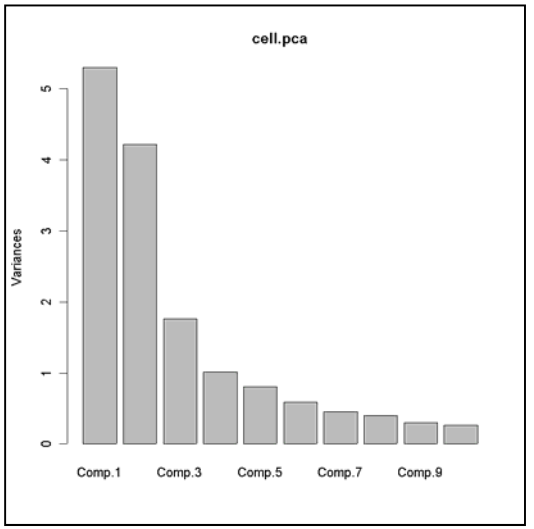

```
cell.matrix  <- read.table("YeastCellCycle_alpha.txt", header=TRUE, row.names=1)
n <- dim(cell.matrix)[1]
p <- dim(cell.matrix)[2]-1
cell.data <- cell.matrix[,2:p+1]
gene.phase <- cell.matrix[,1]
phase <- unique(gene.phase)
phase.name <- c("G1", "S", "S/G2", "G2/M", "M/G1")
cell.sdata <- t(scale(t(cell.data)))
rc <- rainbow(5)[as.integer(gene.phase)]
cc <- rainbow(ncol(cell.sdata))
hv <- heatmap(cell.sdata, col = GBRcol, scale = "column", Colv=NA, Rowv=NA,
              RowSideColors = rc, ColSideColors = cc, margins = c(5,10),
              xlab = "Times", ylab =  "Genes", main = "Heatmap of Microarray Data")
```

PCA on Conditions

```
cell.pca <- princomp(cell.sdata, cor=TRUE, scores=TRUE)
# 2D plot for first two components
pca.dim1 <- cell.pca$scores[,1]
pca.dim2 <- cell.pca$scores[,2]
plot(pca.dim1, pca.dim2, main="PCA for Cell Cycle Data on Genes", xlab="1st PCA
  Component", ylab="2nd PCA Component",col=c(phase), pch=c(phase))
legend(3, 4, phase.name, pch=c(phase), col=c(phase))

# shows a screeplot.
plot(cell.pca)
biplot(cell.pca)
```

MA Table	PCA-1	PCA-2	PCA-3
gene001	-0.18	-0.11	-0.03
gene002	0.51	-0.53	0.54
gene003	-0.35	-0.39	0.26
gene004	-0.18	-1.08	0.41
gene005	-0.62	-0.8	0.13
gene006	-0.09	-0.23	0.77
gene007	-0.38	-0.32	1.08
gene008	-0.88	-0.55	1.03
gene009	-1.26	0.45	0.41
gene010	0.12	-0.36	-0.16
gene011	-0.28	-0.44	2.13
gene012	-0.45	-0.23	0.82
gene013	-0.2	-0.43	0.44
gene014	0.03	-0.26	-0.68
gene015	-0.7	-0.76	0.5
gene016	-0.61	0.07	-0.04
gene017	-0.23	-0.71	0.01
gene018	0.1	0.1	0.11
gene019	-0.94	-0.97	0.24
gene020	-0.55	-0.53	0.86
gene021	-0.47	-0.87	-0.02
gene022	-0.34	-1.1	0.51
gene•••	-0.49	-0.2	0.91
gene n	-0.15	-1.04	-0.01



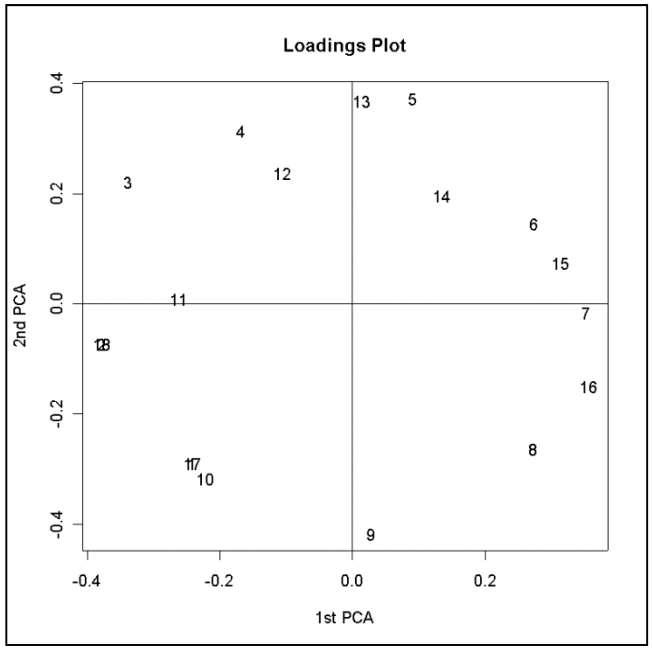


Loadings Plot

```
# loadings plot
plot(loadings(cell.pca)[,1], loadings(cell.pca)[,2], xlab="1st PCA",
ylab="2nd PCA", main="Loadings Plot", type="n")
text(loadings(cell.pca)[,1], loadings(cell.pca)[,2], labels=paste(1:p))
abline(h=0)
abline(v=0)
```

```
> summary(cell.pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5        Comp.15
Standard deviation     2.3012110 2.0542795 1.3300507 1.00895544 0.90053289    0.308577283
Proportion of Variance 0.3309732 0.2637540 0.1105647 0.06362444 0.05068497 ... 0.005951246
Cumulative Proportion  0.3309732 0.5947272 0.7052919 0.76891637 0.81960134    1.000000000
```

```
# print loadings
loadings(cell.pca)
summary(cell.pca)
```



Loadings:	Comp.1	Comp.2	Comp.3	Comp.4
alpha14	-0.283	-0.21	0.283	0.136
alpha21	-0.374	0.211	-0.135	-0.16
alpha28	-0.26	0.298	0.161	-0.168
alpha35	-0.102	0.372	0.165	-0.321
alpha42	0.161	0.355	0.2	-0.317
alpha49	0.287	0.167	0.116	-0.515
alpha56	0.35	0.172	-0.274	-0.115
alpha63	0.251	-0.258	-0.275	-0.37
alpha70	-0.372	-0.217	-0.382	-0.159
alpha77	-0.253	-0.221	-0.321	-0.32
alpha84	-0.249	-0.437	-0.309	-0.256
alpha91	-0.115	0.279	-0.436	0.114
alpha98	0.36	-0.284	0.186	-0.138
alpha105	0.16	0.257	-0.283	-0.125
alpha112	0.347	0.319	-0.178	-0.276
alpha119	0.348	-0.164	-0.201	0.11

Multidimensional Scaling (MDS)

(Torgerson 1952; Cox and Cox 2001)

http://www.lib.utexas.edu/maps/united_states.html

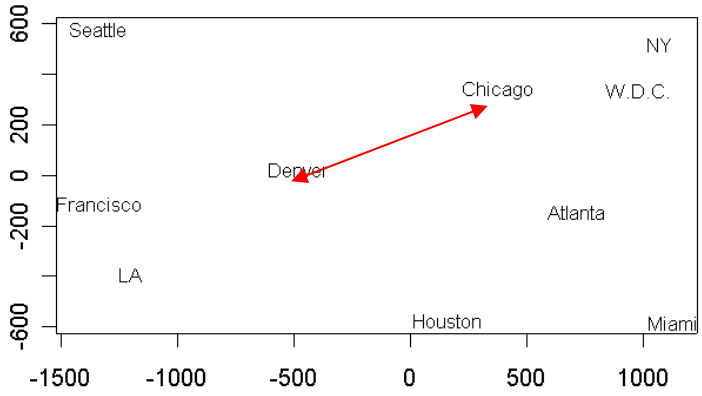
- Classical MDS takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities.
- projection from some unknown dimensional space to 2-d dimension.

Flying Mileages Between Ten U.S. Cities

0										Atlanta
587	0									Chicago
1212	920	0								Denver
701	940	879	0							Houston
1936	1745	831	1374	0						Los Angeles
604	1188	1726	968	2339	0					Miami
748	713	1631	1420	2451	1092	0				New York
2139	1858	949	1645	347	2594	2571	0			San Francisco
2182	1737	1021	1891	959	2734	2408	678	0		Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington D.C.

?

MDS



MDS: Metric and Non-Metric Scaling

- Given the distance between pairs of points, we don't know the exact coordinates of the points, or their dimensionality, or how the distances are calculated.
- MDS is the method for placing these points in a low space such that the Euclidean distance between them in the two-dimensional space is as close as possible to the given distances in the original space.

Question: Given a *dissimilarity matrix* D of certain objects, can we construct points in k-dimensional (often 2-dimensional) space such that

Goal of metric scaling
the Euclidean distances between these points approximate the entries in the dissimilarity matrix?

$$S = \sum_{i,j} (\hat{d}_{ij} - d_{ij})^2$$

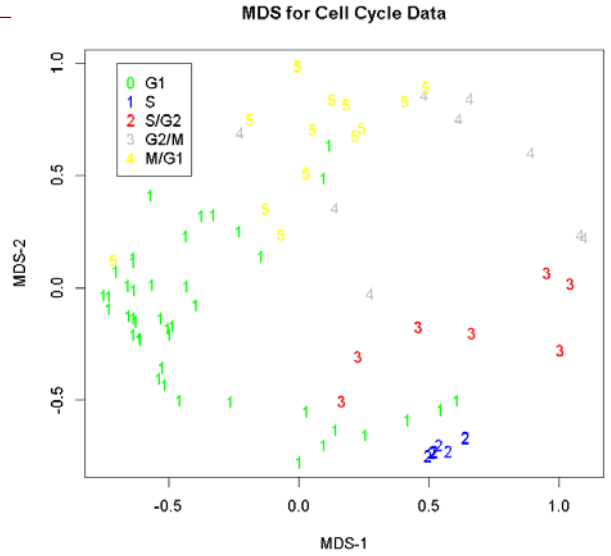
Goal of non-metric scaling
the order in distances coincides with the order in the entries of the dissimilarity matrix approximately?

$$Stress = \sqrt{\frac{\sum_{i,j} (\hat{d}_{ij} - d_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

Mathematically: for given $\boldsymbol{k}$, compute points $\boldsymbol{x_1,\ldots,x_n}$ in $\boldsymbol{k}$-dimensional space such that the object function is minimized.

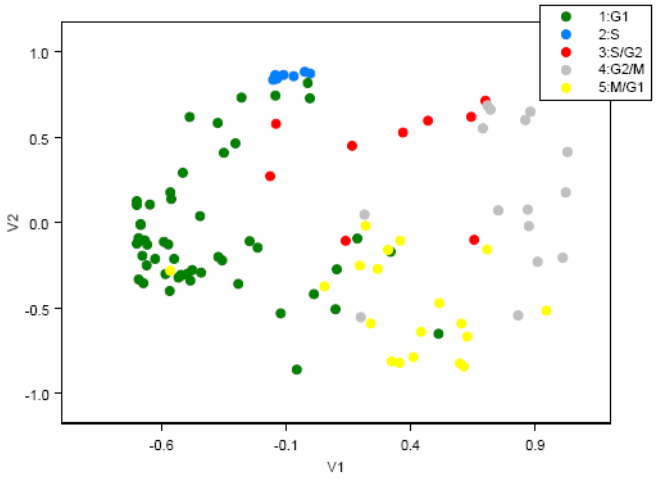
MDS to Microarray Data

```
cell.matrix  <- read.table("YeastCellCycle_alpha.txt", header=TRUE, row.names=1)
n <- dim(cell.matrix)[1]
p <- dim(cell.matrix)[2]-1
cell.data <- cell.matrix[,2:p+1]
gene.phase <- cell.matrix[,1]
phase.name <- c("G1", "S", "S/G2", "G2/M", "M/G1")
cell.sdata <- t(scale(t(cell.data)))
cell.cor <- cor(t(cell.sdata))
cell.dist <- sqrt(2*(1-cell.cor))
cell.mds <- cmdscale(cell.dist)
plot(cell.mds[,1], cell.mds[,2], type="n", xlab="MDS-1", ylab="MDS-2", main="MDS for
Cell Cycle Data")
number <- c(1, 4, 5, 2, 3)[as.integer(gene.phase)]
phase.color <- c("green", "blue", "red", "gray", "yellow")
text(cell.mds[,1], cell.mds[,2], number, col= phase.color[number])
legend(-0.7, 1.0, phase.name, pch="01234", col=phase.color)
```

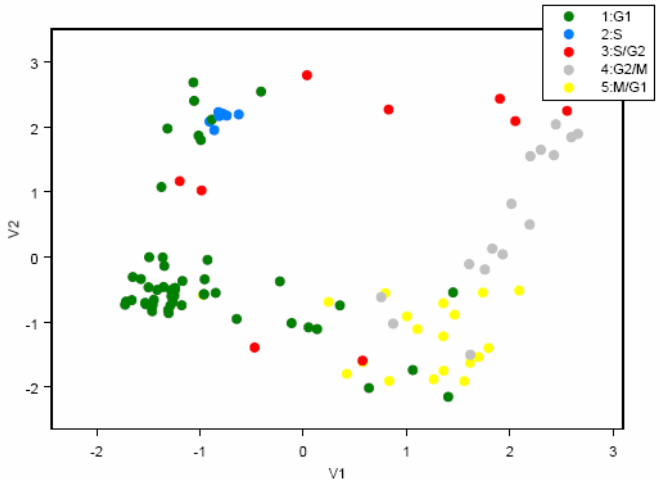


Using Different Features

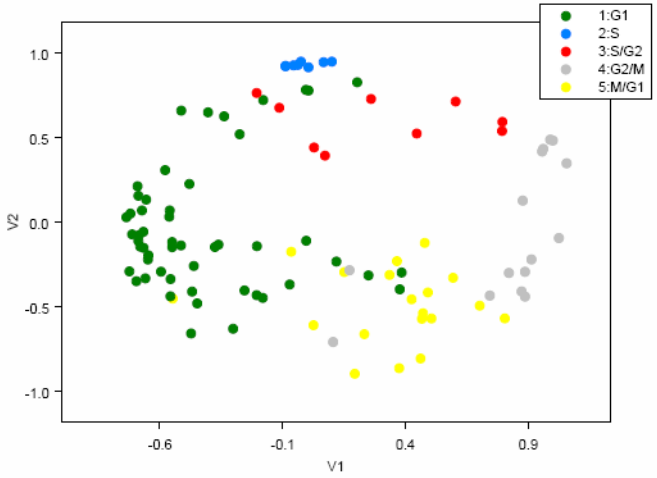
(a) MDS



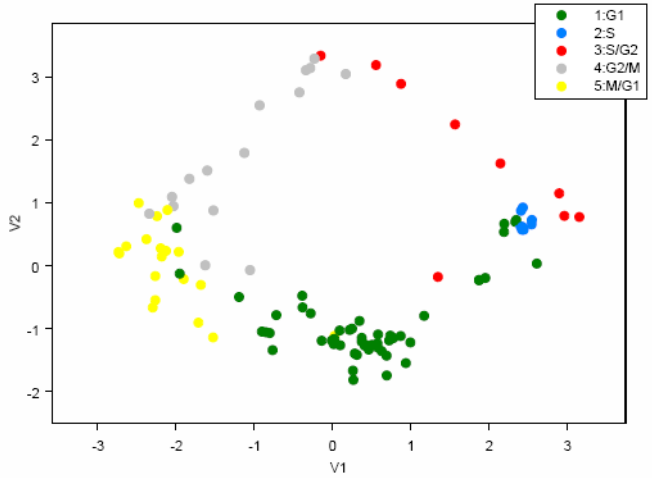
(c) Isomap



(b) MDS+DWT



(d) Isomap+DWT



Isometric Mapping (Isomap)

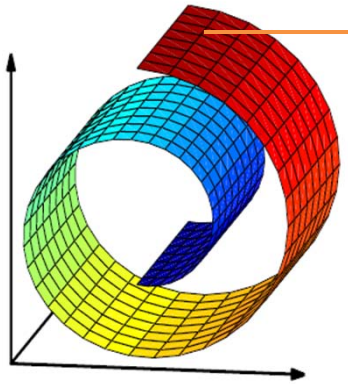
Manifolds (流形) and Nonlinearity:

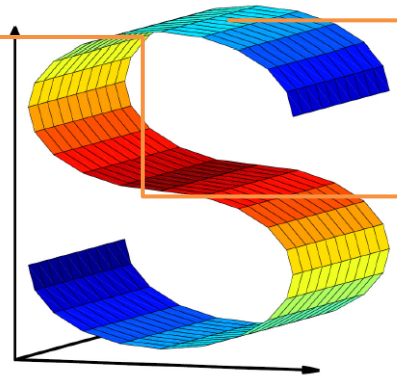
A manifold is a mathematical geometric space and in which the local space is Euclidean space.

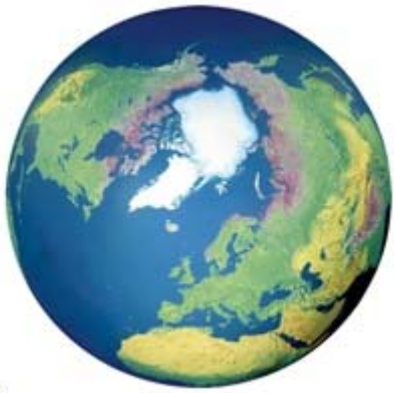
Swissroll

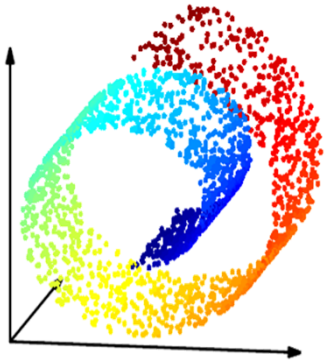
S-curve

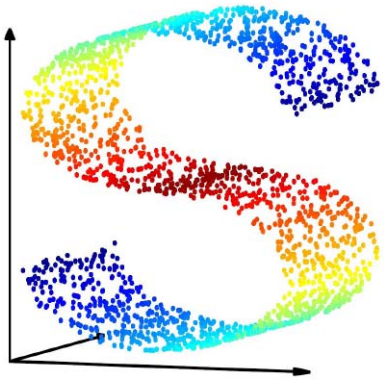
Locally Euclidean





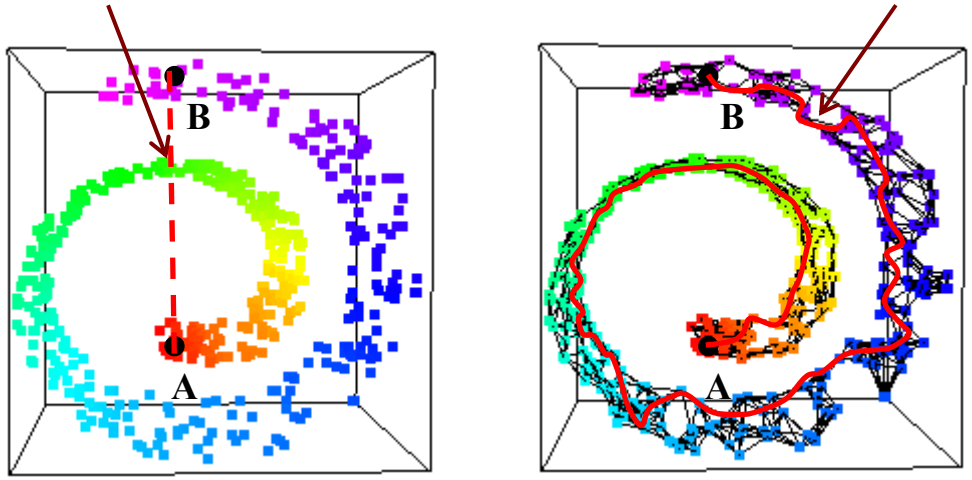






Nonlinear Structure

linear distance nonlinear path



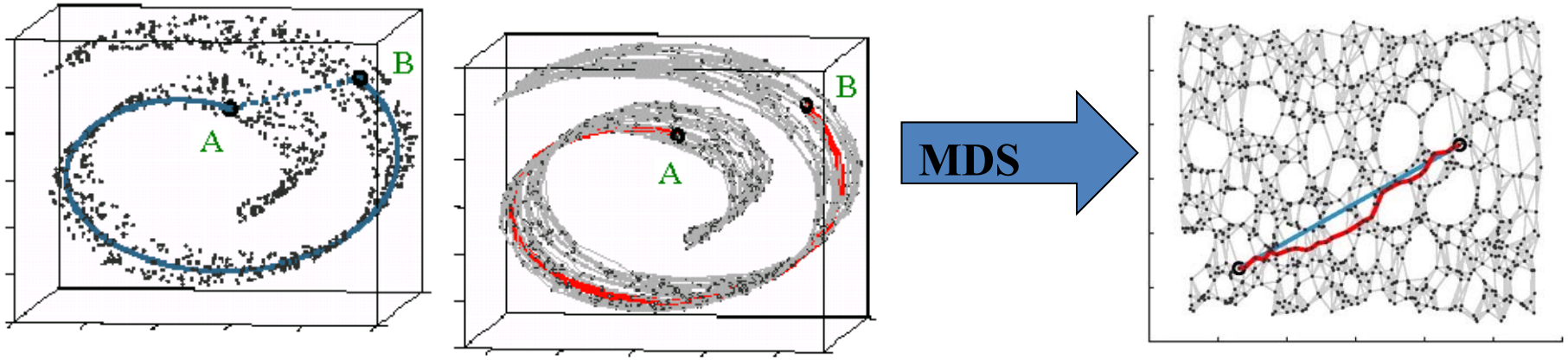
Machine learning for nonlinear dimension reduction (NLDR) and manifold learning

- **ISOMAP** (Tenenbaum, de Silva and Langford, 2000)
- Local linear embedding (**LLE**)(Roweis and Saul, 2000)
- **Hessian eigenmaps** (Donoho and Grimes, 2003)
- **Laplacian eigenmaps** (Belkin and Niyogi, 2003)
- **Diffusion maps** (Coifman et al., 2005) and their variants.

Isometric Mapping (Isomap)

Tenenbaum , J. B., Silva, V. de, and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science* 290, 2319-2323.

Isomap finds the projection that preserves the global, nonlinear geometry of the data by preserving the geodesic manifold interpoint distances.

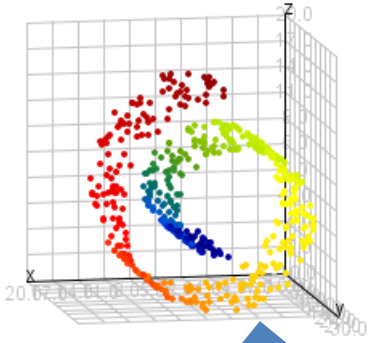


What is important is the geodesic distance!

Isometric Feature Mapping (ISOMAP)

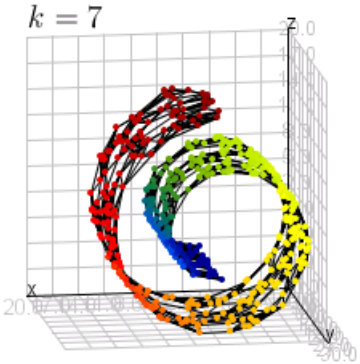
Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science **290** (2000) 2319–2323

Geodesic Distance Approximation

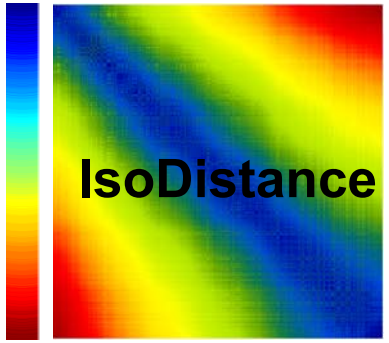


nx3

K-nearest neighbors (KNN)

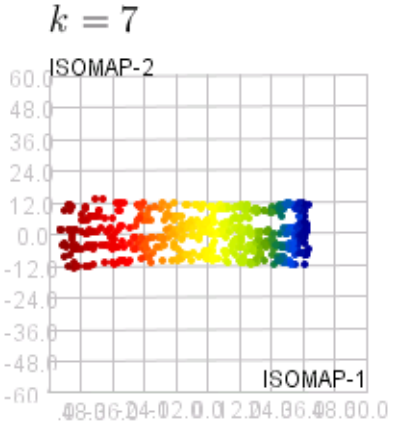


nx3



nxn

Nonlinear Dimension Reduction



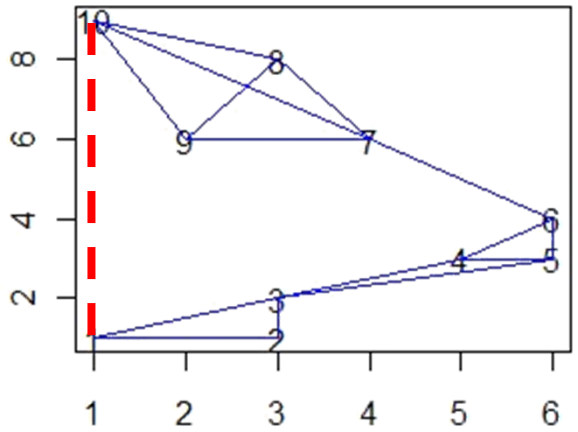
nx2

Multidimensional Scaling (MDS)

Geodesic Distance

Euclidean Distance

samples = 10, neighbors = 3



$$d_G(i,j) = \min\{\ d_G(i,j),\ d_G(i,l) + d_G(l,j)\ \}$$

IsoDistance

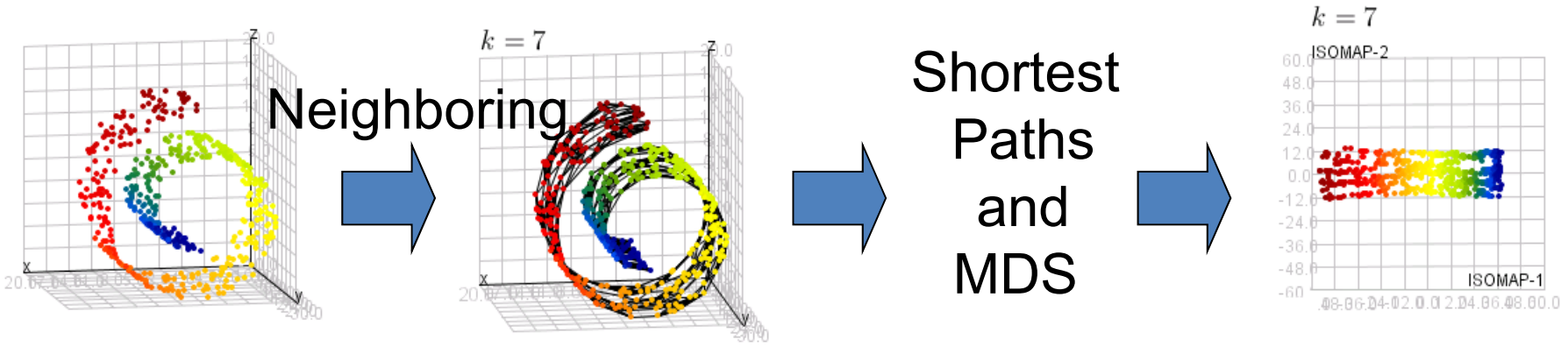
	1	2	3	4	5	6	7	8	9	10
1	0	2	2.236	4.472						?
2	2	0	1	2.828						
3	2.236	1	0	2.236						
4			2.236	0	1	1.414				
5			3.162	1	0	1				
6				1.414	1	0	2.828			
7						2.828	0	2.236	2	
8							2.236	0	2.236	2.236
9							2	2.236	0	3.162
10							4.243	2.236	3.162	0

12.957

The shortest path by point 1 to point 10: 1 →3 →4 →6 →7 →8 →10

	1	2	3	4	5	6	7	8	9	10
1	0	2	2.236	4.472	5.398	5.886	8.715	10.951	10.715	12.957
2	2	0	1	2.828	3.828	4.243	7.071	9.307	9.071	11.314
3	2.236	1	0	2.236	3.162	3.65	6.479	8.715	8.479	10.721
4	4.472	2.828	2.236	0	1	1.414	4.243	6.479	6.243	8.485
5	5.398	3.828	3.162	1	0	1	3.828	6.064	5.828	8.071
6	5.886	4.243	3.65	1.414	1	0	2.828	5.064	4.828	7.071
7	8.715	7.071	6.479	4.243	3.828	2.828	0	2.236	2	4.243
8	10.951	9.307	8.715	6.479	6.064	5.064	2.236	0	2.236	2.236
9	10.715	9.071	8.479	6.243	5.828	4.828	2	2.236	0	3.162
10	12.957	11.314	10.721	8.485	8.071	7.071	4.243	2.236	3.162	0

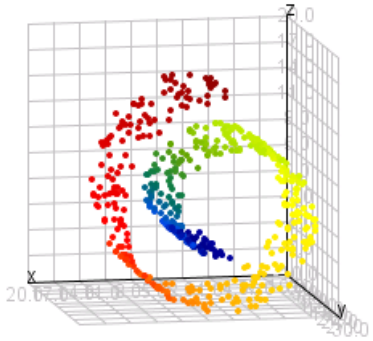
Algorithm of Isomap

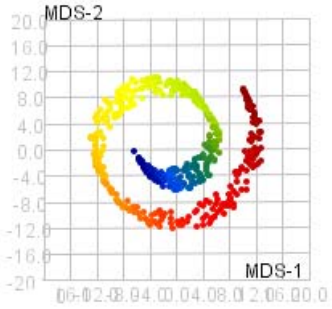


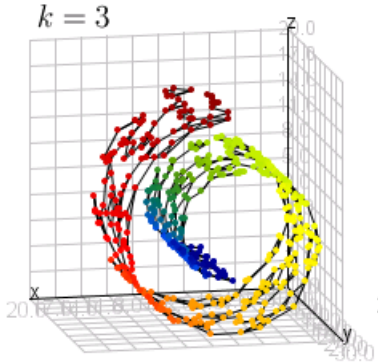
Algorithm of Isomap (Tenenbaum *et al.*, 2000)

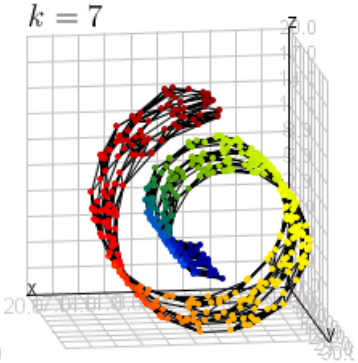
1. Calculate the distance $d_X(i,j)$ between all pairs i,j from n data points in the p-dimensional input space.
2. Construct the graph by determining the neighbors for each data point with ϵ-Isomap or k-Isomap.
3. Pursue the shortest paths in the graph G.
   Initialize $d_G(i,j) = d_X(i,j)$ if i,j are neighbors; otherwise, set $d_G(i,j) = \infty$. For each value of $l = 1, 2, \cdots, n$ and for all i,j, $d_G(i,j)$ are replaced by $\min\{d_G(i,j), d_G(i,l) + d_G(l,j)\}$.
4. Apply classical MDS to D_G.

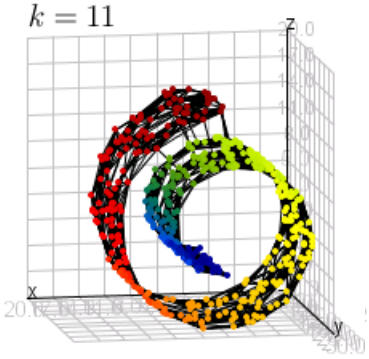
The Pinch and Short-Circuit Problem

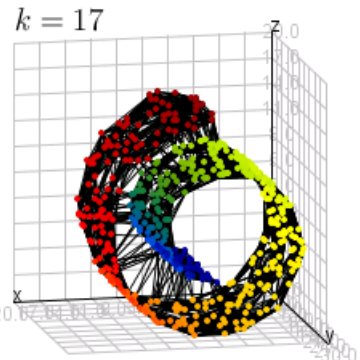


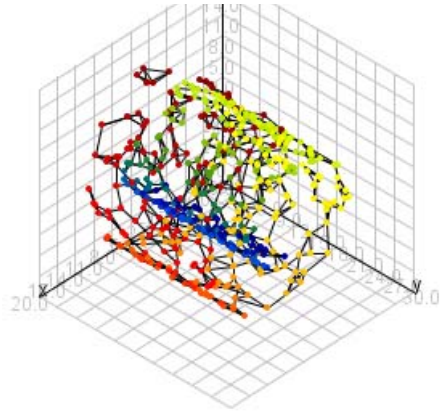


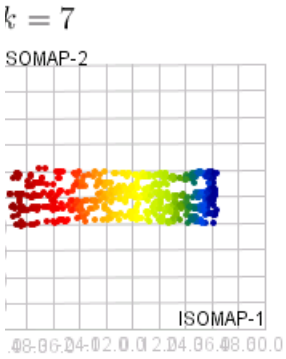


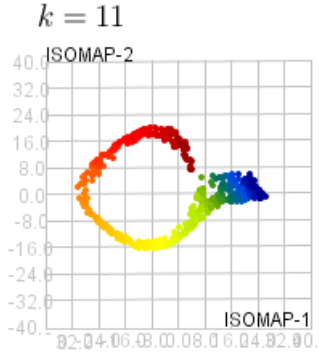


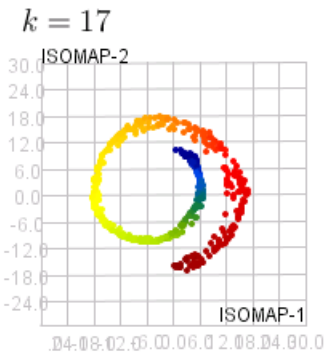


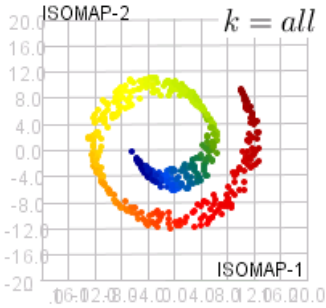










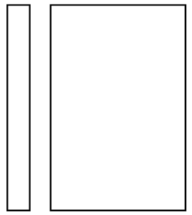


Example

lymphoma dataset

Alizadeh *et al.* (2000)

96 samples



854 genes

9 diagnostic classes

defined by Alizadeh *et al.* (2000).

- DLBCL
- Germinal Centre B
- NI Lymph Node/Tonsil
- Activated blood B
- Resting/activated T
- Transformed cell lines
- Follicular lymphoma
- Resting blood B
- CLL

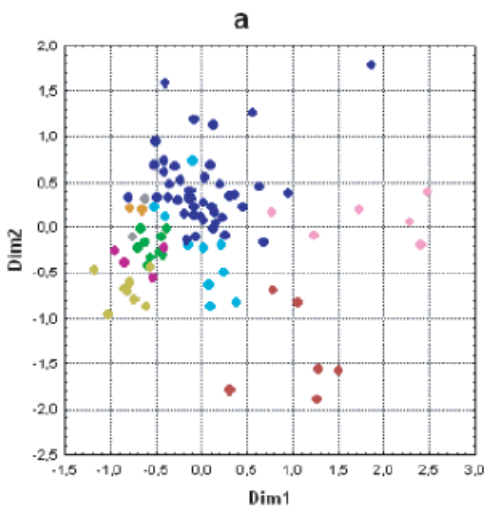
BIOINFORMATICS

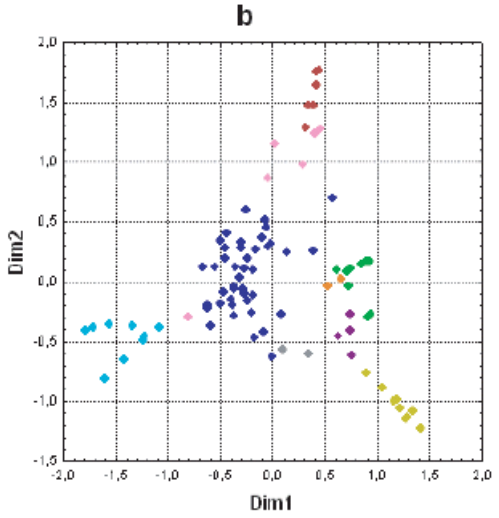
Vol. 20 no. 6 2004, pages 874–880
DOI: 10.1093/bioinformatics/btg496

Approximate geodesic distances reveal biologically relevant structures in microarray data

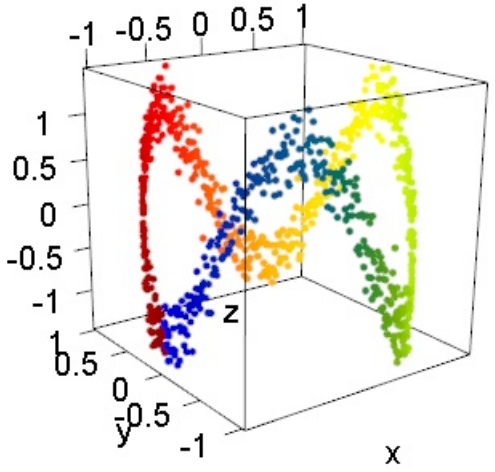
Jens Nilsson[1,], Thoas Fioretos[2], Mattias Höglund[2] and Magnus Fontes[1]*

[1]Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden and [2]Department of Clinical Genetics, Lund University Hospital, SE-221 85 Lund, Sweden

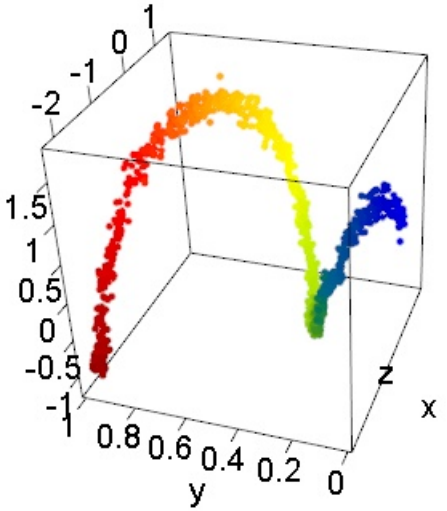




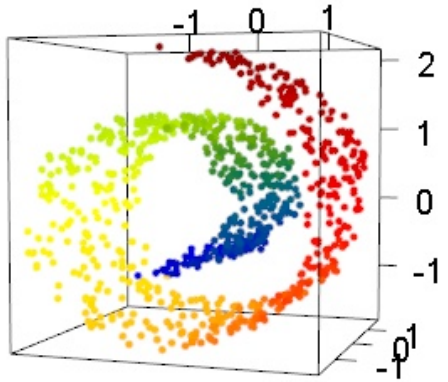
Some Nonlinear Manifold Data Sets



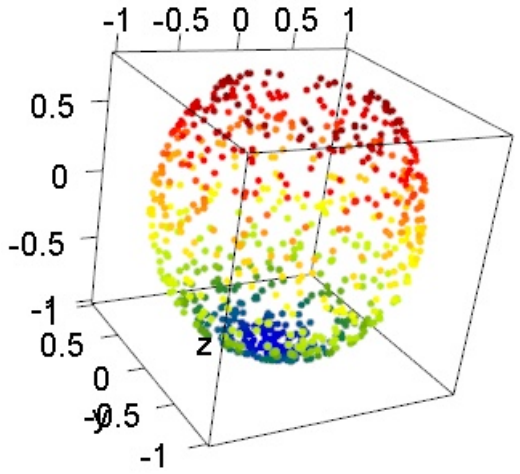
Trigcircle



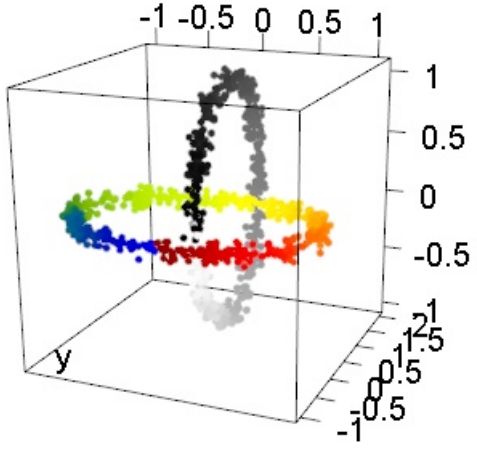
Spiral



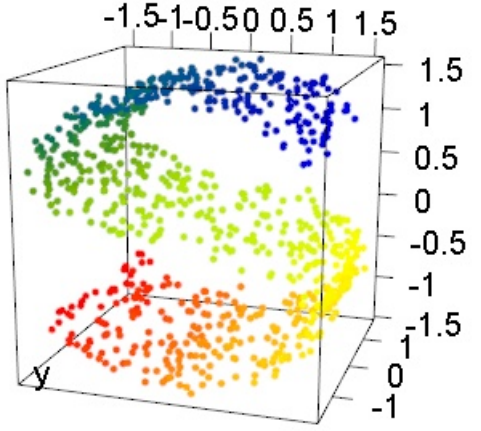
Swiss roll



Fishbowl



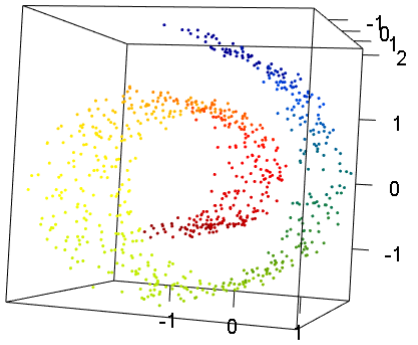
Chainlink

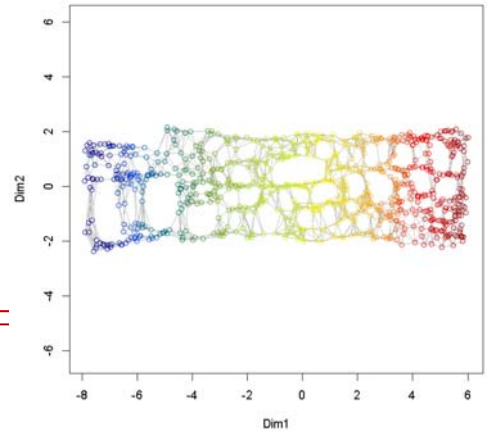


S curve

Simulation of Swiss Roll Data Set

```
swissroll <- function(n, sigma=0.05){

          angle <- (3*pi/2)*(1+2*runif(n));
          height <- runif(n);
          xdata <-  cbind(angle*cos(angle), height, angle*sin(angle))

          # angle <-  seq((1.5*pi): (4.5*pi), length.out=n);
          # height <- sample(0:21, n, replace = TRUE);
          # xdata <-  cbind(angle*cos(angle), height, angle*sin(angle))

          xdata <- scale(xdata) + matrix(rnorm(n*3, 0, sigma), n, 3)

          order.angle <- order(angle)
          sort.angle <- sort(order.angle, index.return=TRUE)
          col.id <- rainbow130(n)
          my.color <- col.id[sort.angle$ix]

          colnames(xdata) <- paste("x", 1:3, sep="")

    return(list(xdata=xdata, angle=angle, color=my.color))
}
```

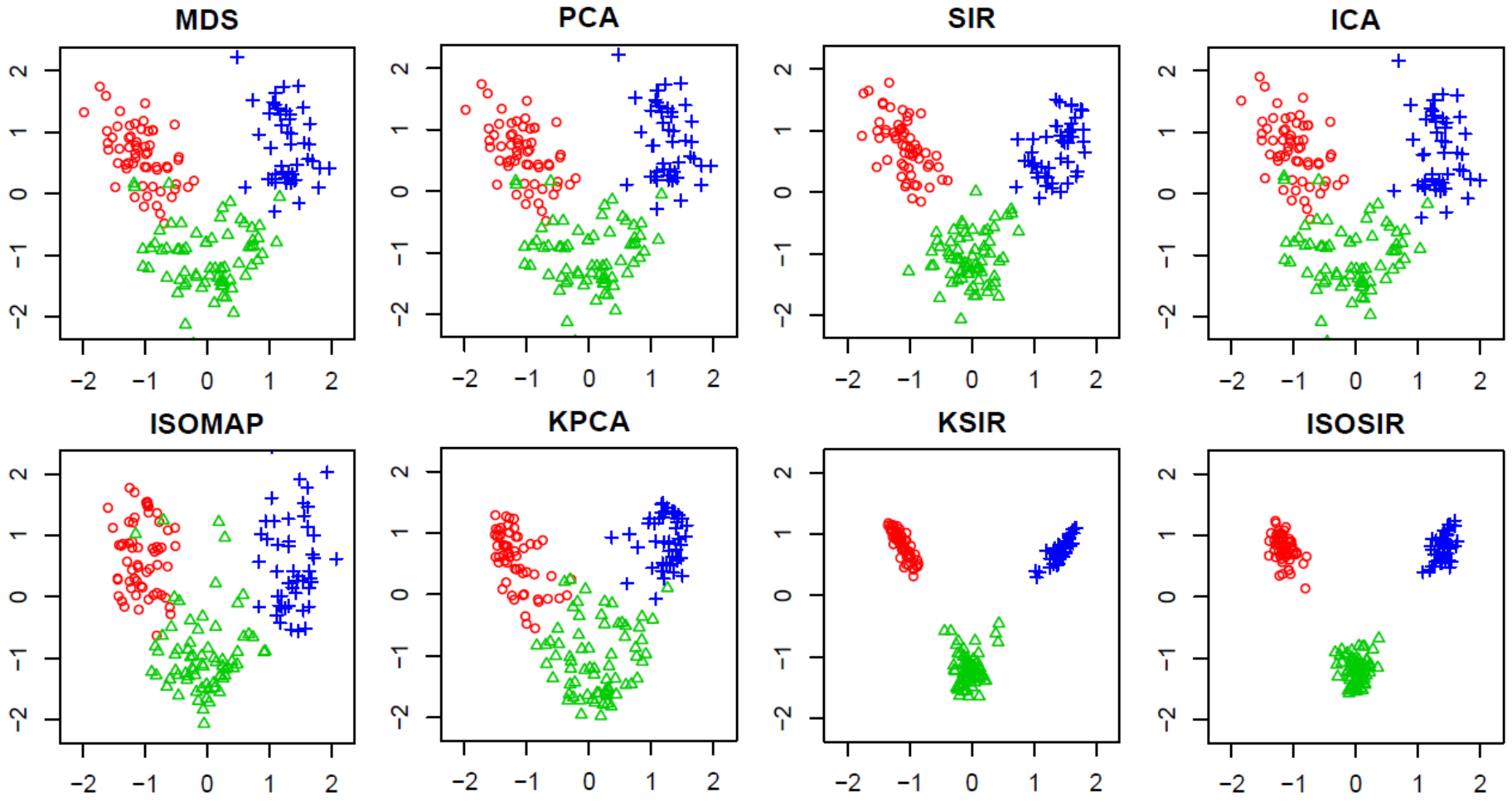




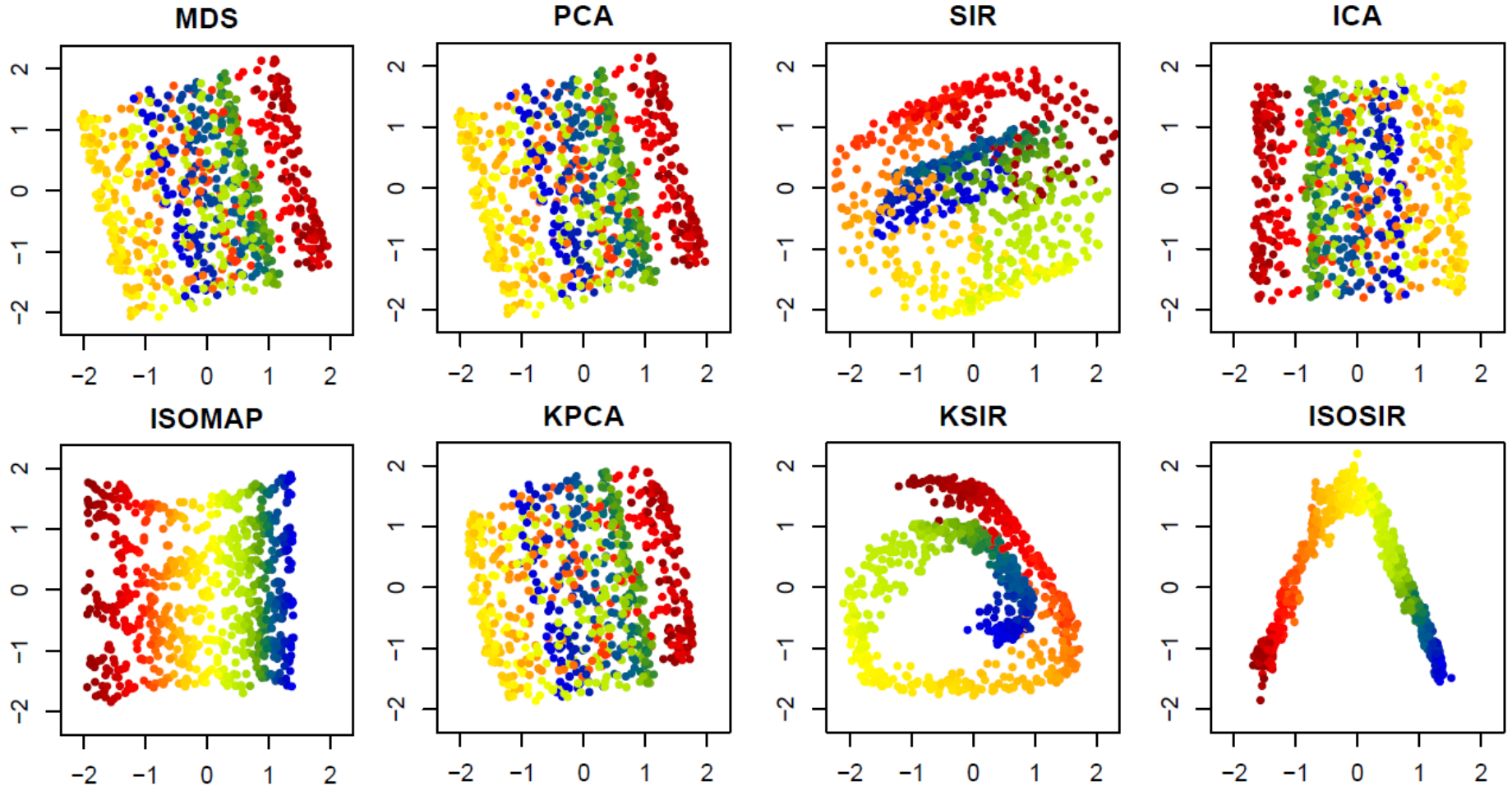

```
source("rainbow130.r")
sr <-  swissroll(800)
library(rgl); open3d()
plot3d(sr$xdata[,1], sr$xdata[,2], sr$xdata[,3], col=sr$color, size=3,
  xlab="", ylab="", zlab="", axes = T)
library(vegan)
sr.isomap <-  isomap(dist(sr$xdata), ndim=2, k=7) # try different k
plot(sr.isomap, col=sr$color)
```

ISOSIR for Nonlinear Dimension Reduction and Data Visualization: Wine Data (178x13)

https://archive.ics.uci.edu/ml/datasets/Wine



ISOSIR for Nonlinear Dimension Reduction and Data Visualization: Swissroll Data (800x3)

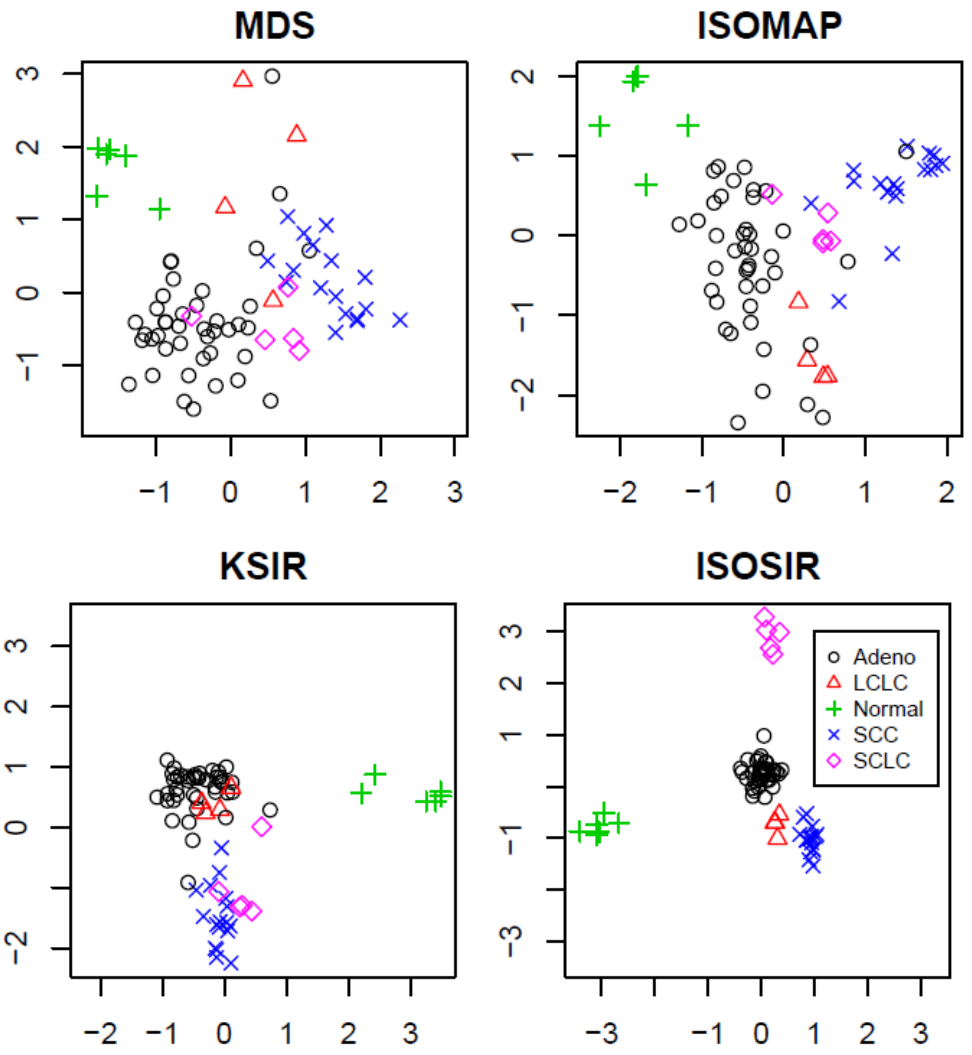


ISOSIR for Nonlinear Dimension Reduction and Data Visualization: Lung Cancer Microarray Data (73x831)

Aim: find differentially expressed genes that can provide a basis for the classification of lung cancer subtypes (Garber et al. 2001).

Samples: 73 samples were divided into five diagnostic classes: adenocarcinoma (Adeno, 41), large-cell lung cancer (LCLC, 4), Normal (6), small-cell lung cancer (SCLC, 5), and squamous cell carcinoma (SCC, 17).

Genes: Nilsson et al. (2004) selected a subset of 831 genes and performed MDS and ISOMAP to visualize these samples.



小細胞肺癌 (SCLC) | 非小細胞肺癌(NSCLC): 包括腺癌(adenocarcinoma)、鱗狀細胞癌(SCC)、大細胞癌(LCLC)。

DR Quality Assessment: local continuity meta-criterion

$$\mathcal{N}^D_{K'}(i) = \{j_1, \ldots, j_{K'}\} \qquad K'\text{-NNs with regard to } D_{i,j}$$

$$\mathcal{N}^X_{K'}(i) = \{k_1, \ldots, k_{K'}\} \qquad K'\text{-NNs with regard to } \|\mathbf{x}_i - \mathbf{x}_k\| \text{ (excluding } i\text{)}.$$

$$N_{K'}(i) = \left|\mathcal{N}^D_{K'}(i) \cap \mathcal{N}^X_{K'}(i)\right|, \quad N_{K'} = \frac{1}{N}\sum_{i=1}^{N} N_{K'}(i).$$

normalize overlap to the [0, 1] interval: $M_{K'} = \frac{1}{K'} N_{K'}$

adjusted LC meta-criteria (Chen 2006) $M^{\text{adj}}_{K'} = M_{K'} - \frac{K'}{N-1}$.

Lee, J.A., Lee, J.A., Verleysen, M., 2009. Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 72.
Chen, L., Buja, A., 2009. Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Layout and Proximity Analysis，JASA 104(485), 209-219.

Co-ranking Matrix and LCMC

```
my.coranking <- function(X.high, X.low){
  # X.high <- iris[1:10,1:4]
  # X.low <- princomp(X.high)$score[,1:2]

  D.high <- as.matrix(dist(X.high))
  D.low <- as.matrix(dist(X.low))
  f <- function(x){
     rank(x, ties.method = 'first', na.last = FALSE)
  }
  diag(D.high) <- NA # NA is rank 1
  diag(D.low) <- NA
  R.high <- apply(D.high, 1, f)
  R.low <- apply(D.low, 1, f)
  table(R.high, R.low)[-1, -1]

}
```

```
> LCMC
function (Q, K = 1:nrow(Q))
{
...
    nQ <- nrow(Q)
    nK <- length(K)
    N <- nQ + 1
    if (nK < 0.2 * nQ) {
        lcmc <- numeric(nK)
        for (i in 1:nK) {
            k <- K[i]
            lcmc[i] <- k/(1 - N) + sum(Q[cm.UL_K(k, nQ)])/N/k
        }
    }
    else {
        lcmc_ <- diag(apply(apply(Q, 2, cumsum), 1, cumsum))/(1:nQ)/N - (1:nQ)/nQ
        lcmc <- lcmc_[K]
    }
    lcmc
}
<environment: namespace:coRanking>
```

The *co-ranking matrix* [23] can then be defined as

$$\mathbf{Q} = [q_{kl}]_{1 \leqslant k,l \leqslant N-1} \quad \text{with} \quad q_{kl} = |\{(i,j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|.$$

coRanking: Co-Ranking Matrix

Calculate the co-ranking matrix to assess the quality of a diminsionality reduction.

```
coranking(Xi, X, input = c("data", "dist", "rank"), use = "C")
```

Arguments

`Xi`: high dimensional data

`X`: low dimensional data

`input`: type of input

Plots the co-ranking matrix nicely

```
imageplot(Q, lwd = 2, bty = "n", main = "co-ranking matrix",
xlab = expression(R), ylab = expression(Ro),
col = colorRampPalette(colors = c("gray85", "red", "yellow",
"green", "blue"))(100), axes = FALSE, legend = TRUE, ...)
```

Calculate the local continuity meta-criterion from a co-ranking matrix.

```
LCMC(Q, K = 1:nrow(Q))
```

Arguments

`Q`: a co-ranking matrix

`K`: vector of integers describing neighborhood size

The *co-ranking matrix* [23] can then be defined as

$$\mathbf{Q} = [q_{kl}]_{1 \leqslant k,l \leqslant N-1} \quad \text{with} \quad q_{kl} = |\{(i,j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|.$$

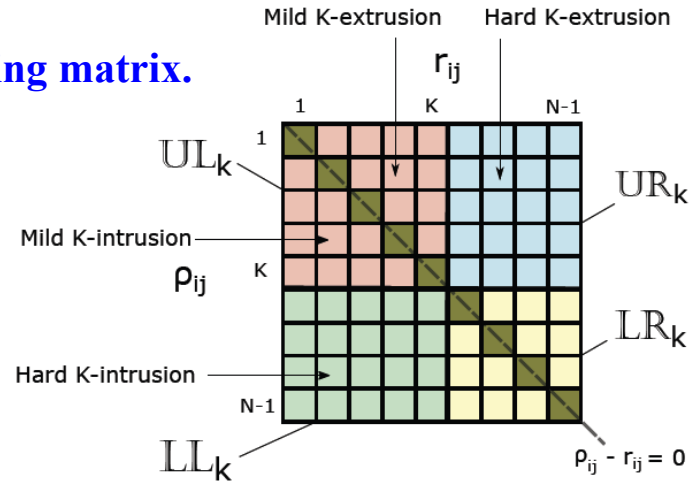
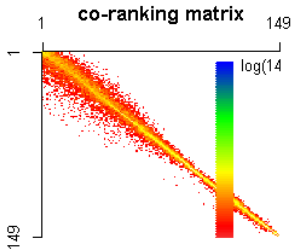
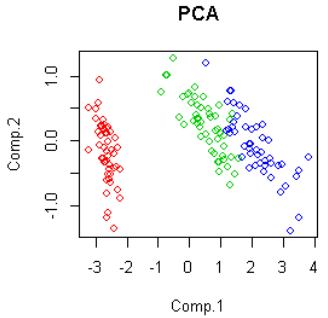


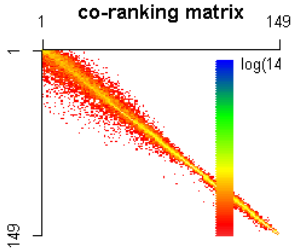
Figure 2: The co-ranking matrix with four blocks.

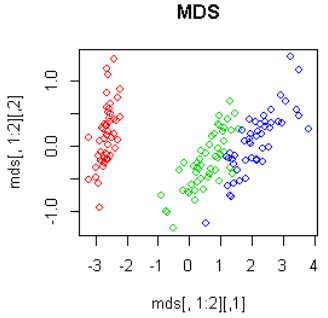
Example

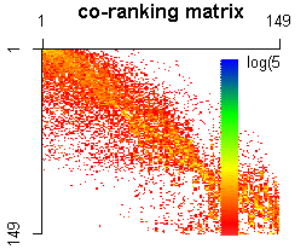
```
> library(coRanking) # install.packages("coRanking")
> library(MASS)
> par(mfrow=c(2, 3))
> x <- iris[,1:4]
> y <- iris[,5]
> pca <- princomp(x)
> Q.pca <- coranking(x, pca$score[,1:2], input = "data")
> imageplot(Q.pca)
> lcmc.pca <- LCMC(Q.pca, K = 5:10)
>
> mds <- cmdscale(dist(x))
> Q.mds <- coranking(x, mds[,1:2], input = "data")
> imageplot(Q.mds)
> lcmc.mds <- LCMC(Q.pca, K = 5:10)
>
> mylda <- lda(x, grouping=y)
> lda.dim <- as.matrix(x)%*%mylda$scaling[,1:2]
> Q.lda <- coranking(x, lda.dim, input = "data")
> imageplot(Q.lda)
> lcmc.lda <- LCMC(Q.lda, K = 5:10)
>
> names(lcmc.pca) <- paste0("K=", 5:10)
> rbind(lcmc.pca, lcmc.mds, lcmc.lda)
              K=5       K=6       K=7       K=8       K=9
lcmc.pca 0.6077763 0.6108427 0.6292106 0.6521421 0.6581158
lcmc.mds 0.6077763 0.6108427 0.6292106 0.6521421 0.6581158
lcmc.lda 0.3171096 0.3286204 0.3396868 0.3638087 0.3781158
> plot(pca$score[,1:2], col=as.integer(y)+1, main="PCA")
> plot(mds[,1:2], col=as.integer(y)+1, main="MDS")
> plot(lda.dim[,1:2], col=as.integer(y)+1, main="LDA")
```

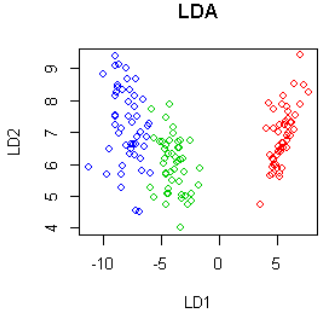












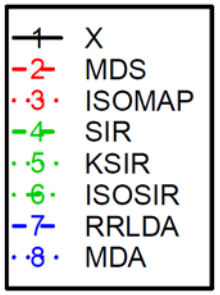
Evaluate DR by Classification ($n >> p$)
UCI Datasets

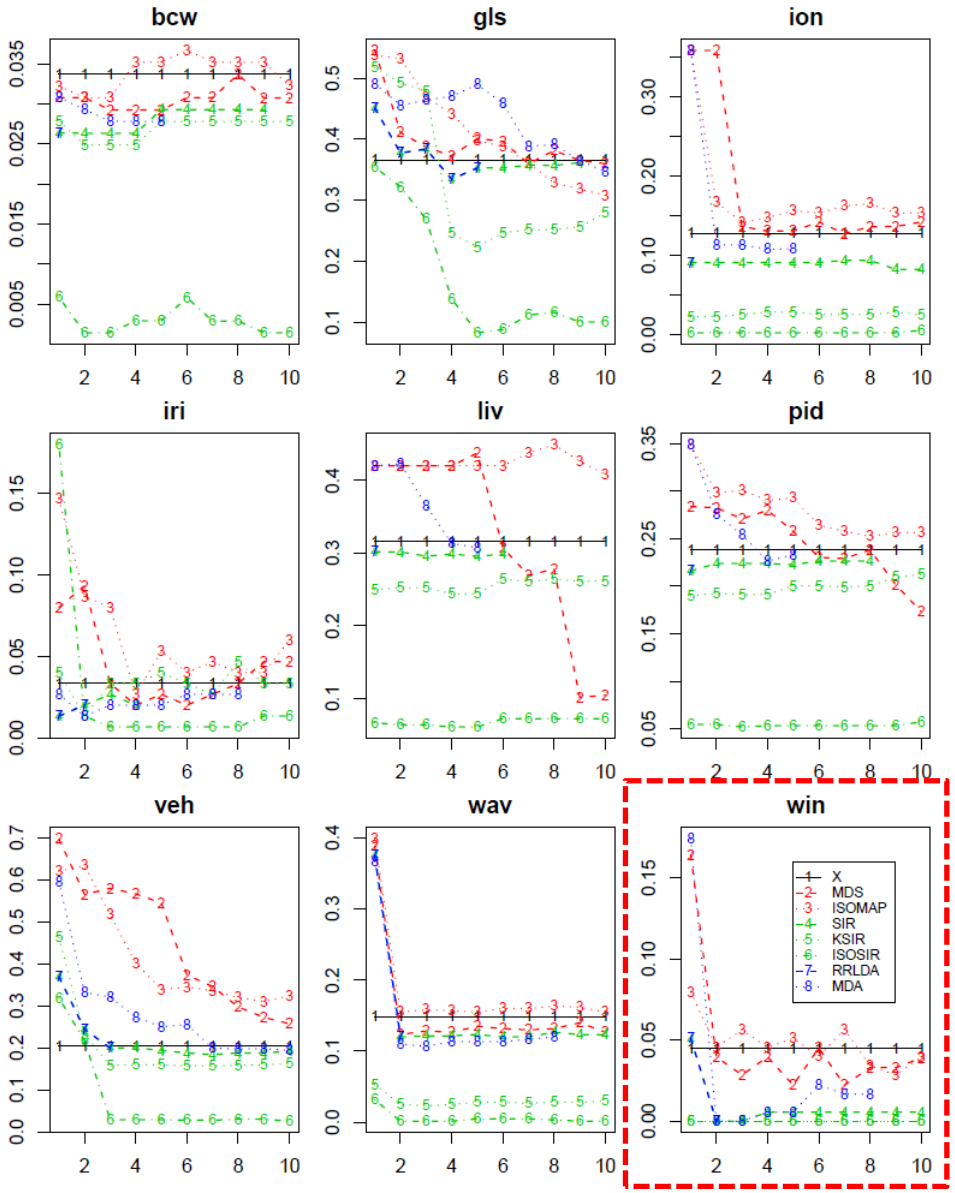
DR as a tool of feature extraction.

UCI machine learning repository datasets

Data set	n	p	$G(n_h)$
Wisconsin breast cancer (`bcw`)	683	9	2 (444, 239)
Glass identification (`gls`)	214	9	6 (70, 76, 17, 13, 9, 29)
Ionosphere (`ion`)	351	33	2 (225, 126)
Iris plants (`iri`)	150	4	3 (50×3)
BUPA liver disorders (`liv`)	345	6	2 (145, 200)
Pima Indians diabetes (`pid`)	768	8	2 (500, 268)
StatLog vehicle silhouettes (`veh`)	846	18	4 (212, 217, 218, 199)
Waveform database generator (`wav`)	600	21	3 (200×3)
Wine recognition data (`win`)	178	13	3 (59, 71, 48)

- Classifier : linear SVM
- 10-fold cross-validation error rate
- Gaussion kernel with scale 0.05





Evaluate DR by Classification ($n << p$) Microarray Datasets

Datasets	Publication	n	p	$G(n_h)$	Response
Brain	Pomeroy et al. 2002	42	5597	5(10, 10, 10, 4, 8)	Different tumor types
Colon	Alon et al. 1999	62	2000	2(22, 40)	Tumor/normal tissue
Leukemia	Golub et al. 1999	72	3571	2(47, 25)	Subtypes of leukemia
Lymphoma	Alizadeh et al. 2000	62	4026	3(42, 9, 11)	Subtypes of lymphoma
Prostate	Singh et al. 2002	102	6033	2(50, 52)	Tumor/normal tissue
SRBCT	Khan et al. 2001	63	2308	4(23, 20, 12, 8)	Different tumor types

n* x *p*; *n* << *p

$$\Sigma_{XX} \rightarrow p \times p$$

$$\text{IsoDistance} \rightarrow n \times n$$

- Classifier : linear SVM
- leave-one-out cross-validation error rate
- Gaussion kernel with scale 0.05