



台灣人工智慧學校

經理人週末研修班

探索式資料分析 統計圖表



吳漢銘

國立臺北大學 統計學系

主要參考書目

<https://www.coursera.org/learn/exploratory-data-analysis>

The screenshot shows the Coursera course page for 'Exploratory Data Analysis'. At the top, there's a navigation bar with the Coursera logo, an 'Explore' dropdown, a search bar, and links for 'For Enterprise', 'Log In', and 'Sign Up'. Below the navigation is a breadcrumb trail: Home > 数据科学 > 数据分析. The main title 'Exploratory Data Analysis' is displayed prominently. On the left, a sidebar contains links for 'Overview', 'Syllabus', 'FAQs', 'Creators', 'Pricing', and 'Ratings and Reviews', along with a large 'Enroll' button and a note that it started on May 14. Below the sidebar, the 'JOHNS HOPKINS UNIVERSITY' logo is shown. The main content area includes a detailed description of the course, a 'Created by' section listing Johns Hopkins University, and a table with basic course information like 'Basic Info', 'Language', 'How To Pass', and 'User Ratings'.

EDA with R: Course Content

- Making exploratory graphs
- Principles of analytic graphics
- Plotting systems and graphics devices in R
- The base, lattice, and ggplot2 plotting systems in R
- Clustering methods (群集分析)
- Dimension reduction techniques (維度縮減)

授課教師

Roger D. Peng, PhD
約翰霍普金斯大學

Jeff Leek, PhD
約翰霍普金斯大學

Brian Caffo, PhD
約翰霍普金斯大學

課程類型

信息、技術和設計
統計和數據分析

Exploratory Data Analysis with R



Roger D. Peng



John Tukey (1915~2000): 統計學界的畢卡索

3/47

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。

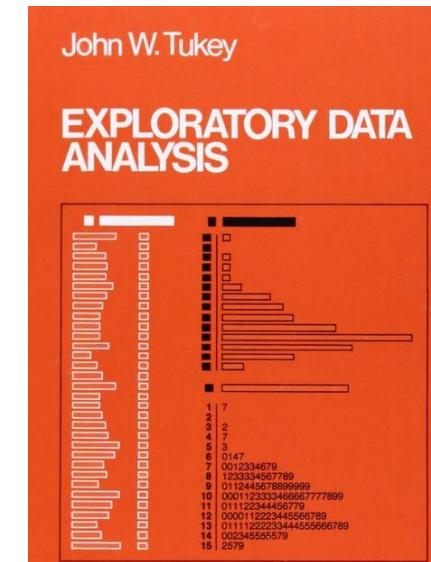
「對正確的問題有個近似的答案，
勝過對錯的問題有精確的答案。」

"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)

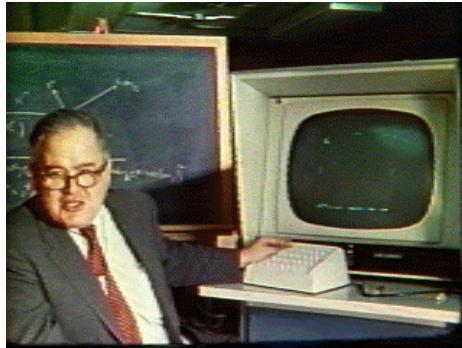


Source: <http://www.unige.ch/ses/sococ/cl/bib/eda/tukey.html>



「統計應該是科學，而非數學！」

4/47



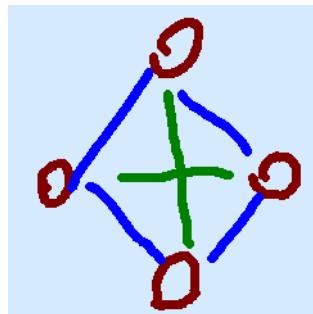
他曾挑戰當時主流的數理統計學家，堅持 data analysis 是統計分析中不可忽視的步驟，數學的假設需要 data 加以驗證才可行。Tukey 說過統計應該是科學，而非數學！

數學思維 vs 統計思維
證明在哪裏? vs 數據在哪裏?

Stanford Linear Accelerator (1973)

"Let the data speak for themselves"

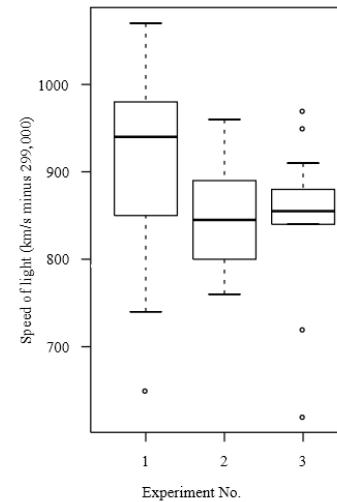
|||| ||| ||||



Stem and Leaf Plot

42 0	
44 0000	
46 000000	
48 000000000000	
50 00000000000000000000	
52 00000	
54 0000000000000000	
56 0000000000000000	
58 000000000000	
60 00000000000000	
62 00000000000000	
64 00000000000000	
66 000000000000	
68 0000000	
70 00	
72 0000	
74 0	
76 00000	
78 0	

Box-and-whisker plot





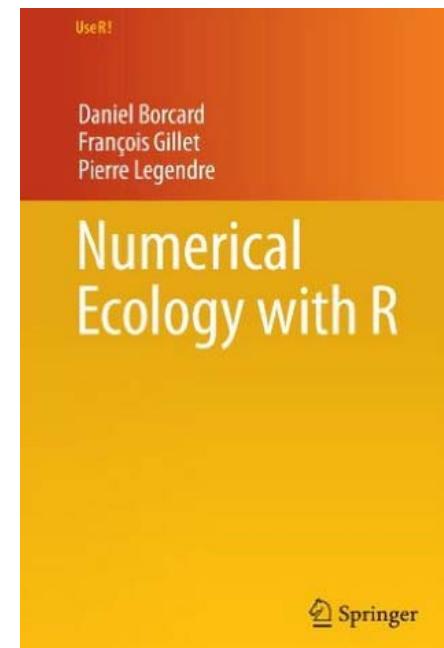
What is EDA?

- Exploratory Data Analysis (EDA) is an **approach/philosophy** for data analysis that employs a variety of techniques (mostly **graphical**) to
 - maximize **insight** into a data set;
 - uncover underlying **structure**;
 - extract important variables;
 - detect **outliers** and anomalies (detection of mistakes);
 - test underlying **assumptions**;
 - develop parsimonious **models** (preliminary selection of appropriate models);
 - determine **optimal** factor settings;
 - determine **relationships** among the explanatory variables; and
 - assess the direction and rough size of relationships between explanatory and **outcome variables**.
- You should always look at every variable - you will learn something!

Source: <http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm>

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.



What Do They Say About EDA?

- Howard J. Seltman (2015), Experimental Design and Analysis.
 - EDA need not be restricted to techniques you have seen before; sometimes you need to **invent a new way** of looking at your data.
 - Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables.
 - **EDA is not an exact science, it is a very important art!**



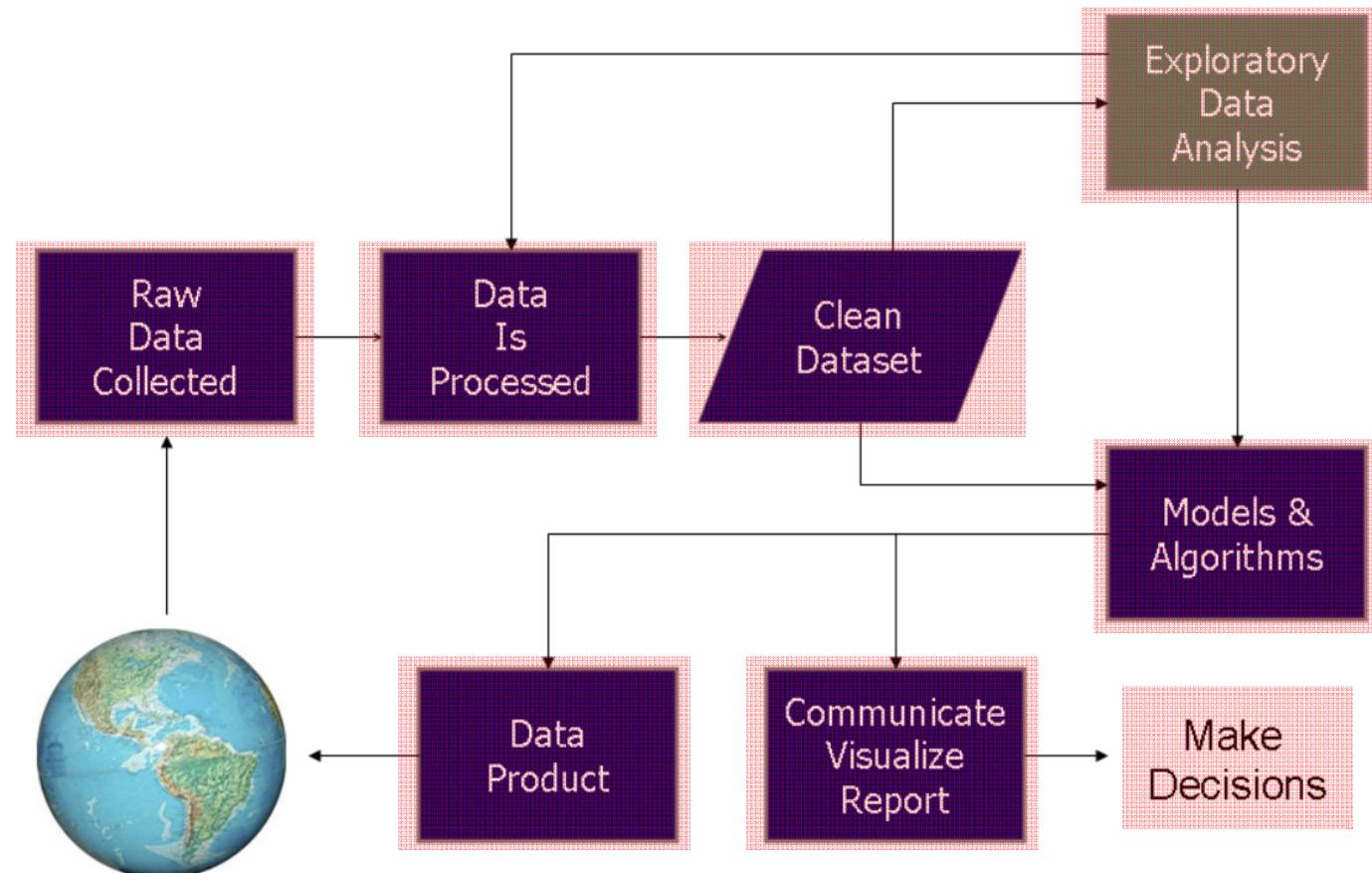
Source: google images



Data Analysis Procedures

- Statistics and data analysis procedures can broadly be split into two parts: (1) **Graphical techniques**. (2) **Quantitative techniques**.

Data Science Process



Source: https://en.wikipedia.org/wiki/Exploratory_data_analysis

例子2：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

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	<input type="text"/>

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



Donald J. Trump

Tweets Tweets & replies Media

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發佈快更安全
- WWDC / watchOS 4來了更聰明貼心
- WWDC / 跑VR輕而易舉 超強iMac登場
- WWDC / 蘋果發表iOS 11 控制中顛換面

[訂閱科](#)

【錯過可惜】Follow me! 權貴一起來

【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？

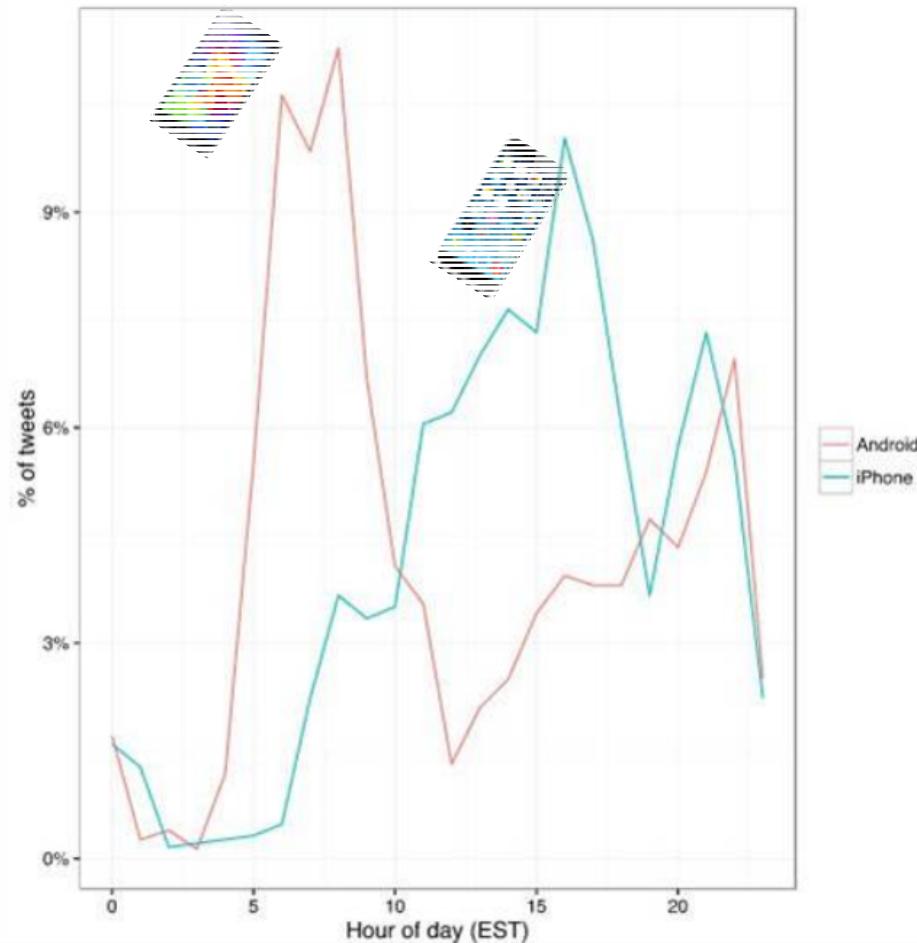
數據分析師David Robinson發現，川普發表祝賀內容時，是透過iPhone；而用來抨擊選戰對手時，則是透過Android手機。到底川普個人推特推文的差異，從何而來？這些推文是不是由他一個人包辦？



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

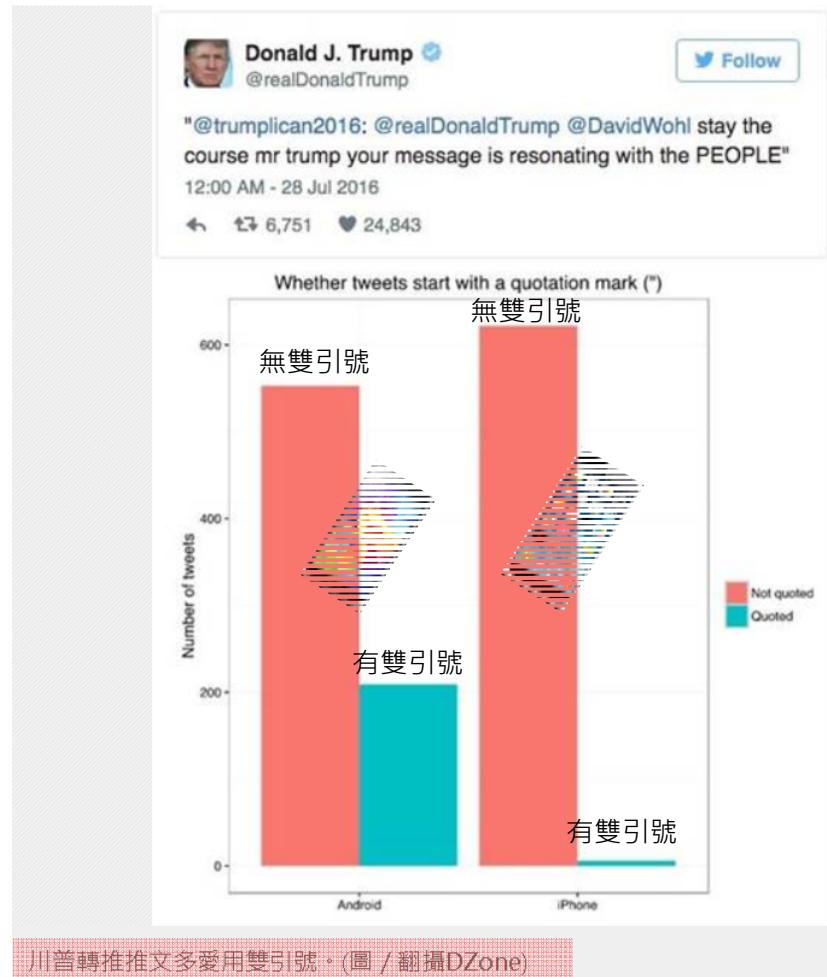
→川普習慣在早上發推文；而他的助理或團隊習慣在下午或晚上發推文



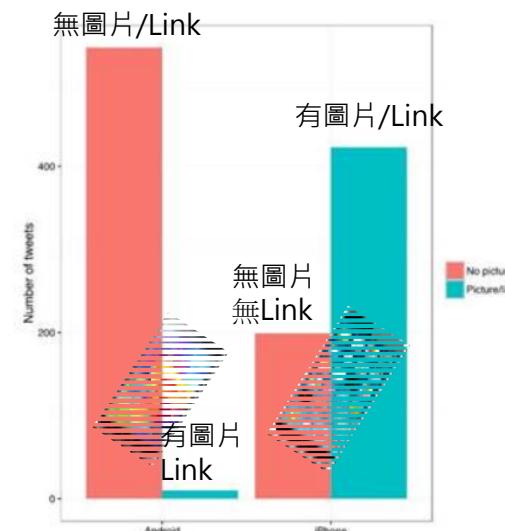
就推文時間分析來看，可看出來自Android手機的推文時間大多落在早上，與來自iPhone端的推文時間區間不同。(圖 / 翻攝DZone)

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



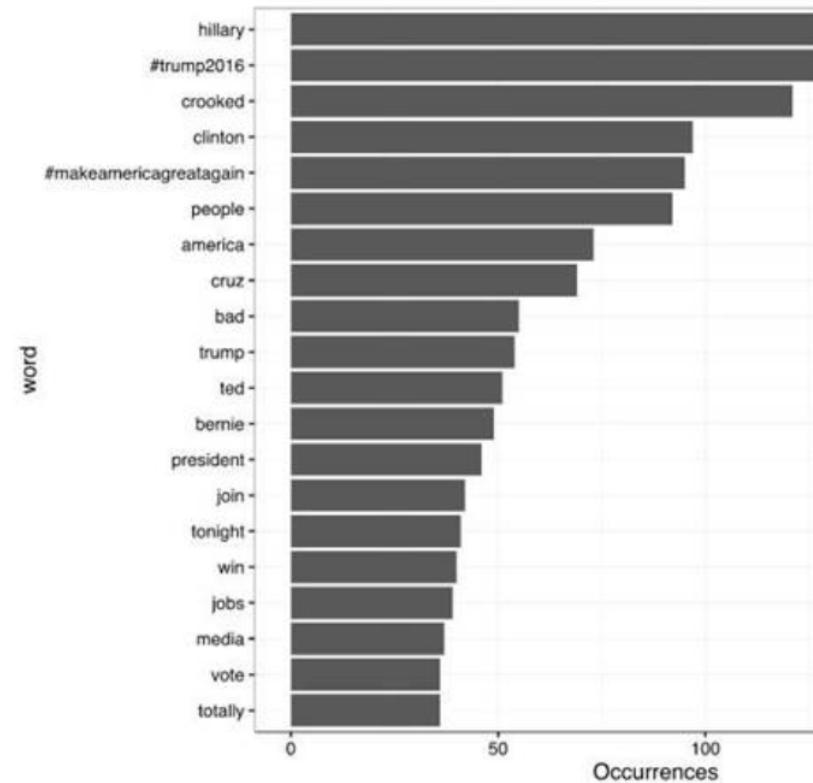
→川普的推文都以文字為主，少附link以及圖片



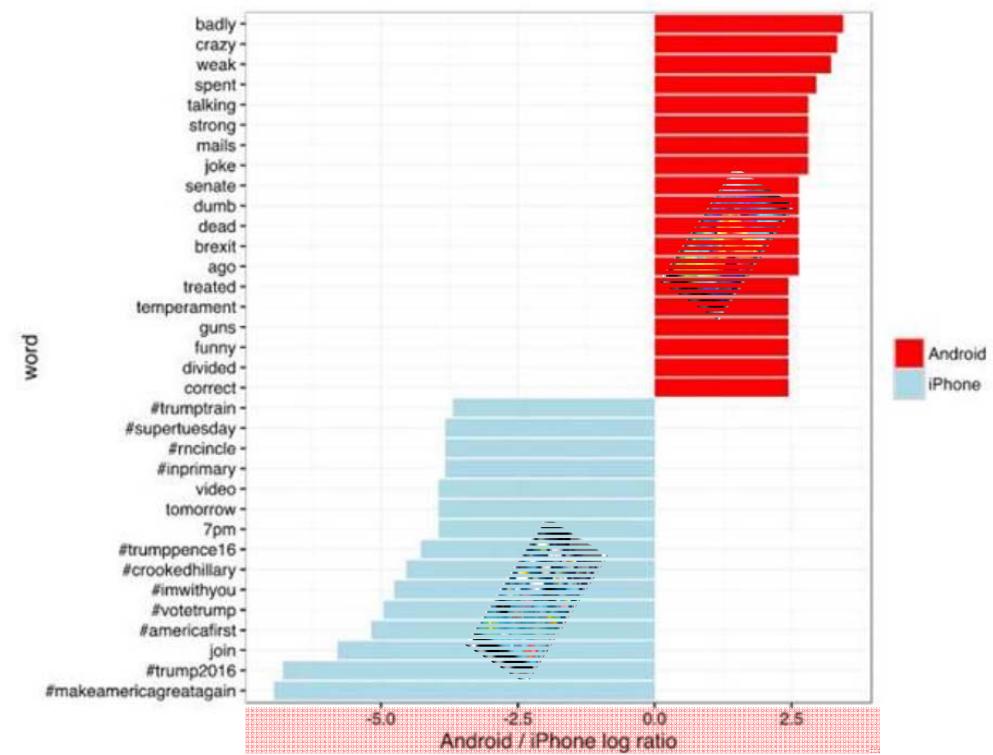
川普的推文很少用link以及圖片(如左下)，來自iPhone的推文習慣不同，常附圖片。(圖 / 翻攝DZone)

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)



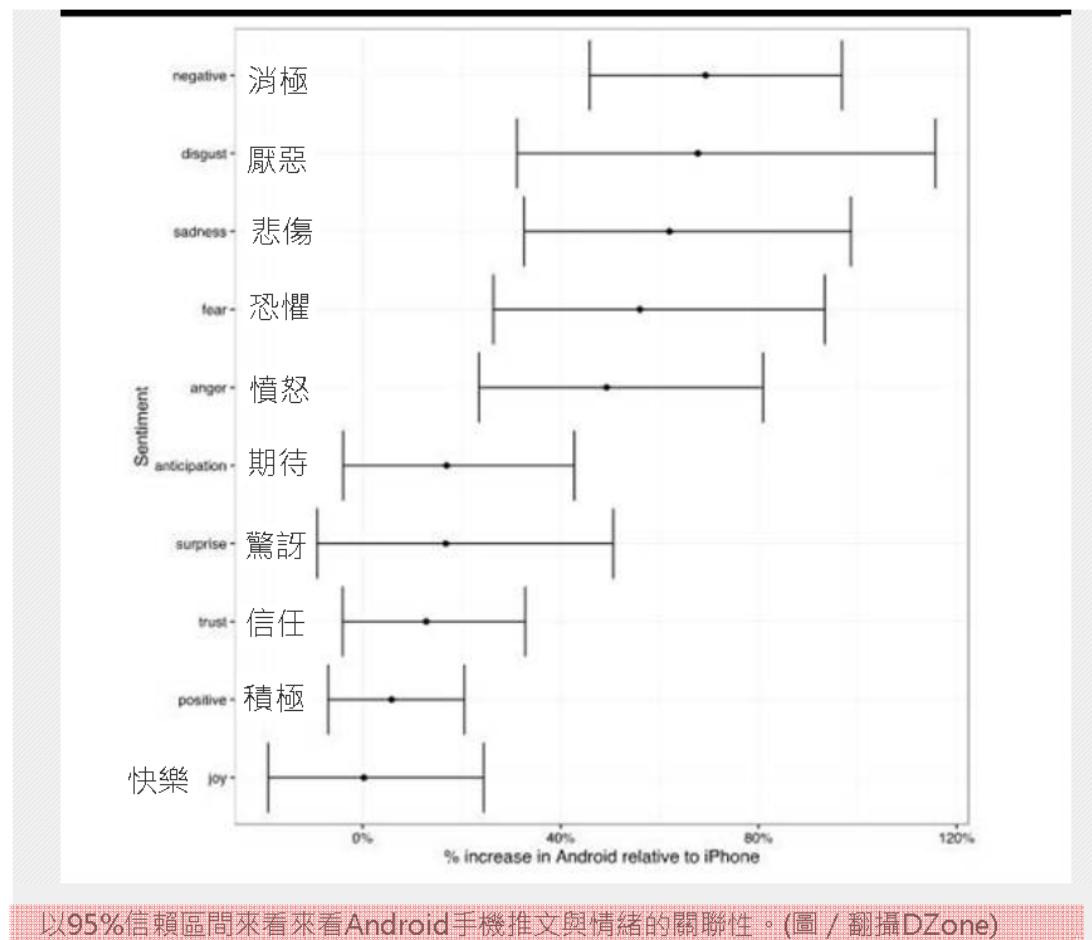
Android帳號推文與iPhone推文常用字的對比。(圖 / 翻攝DZone)

情感分析



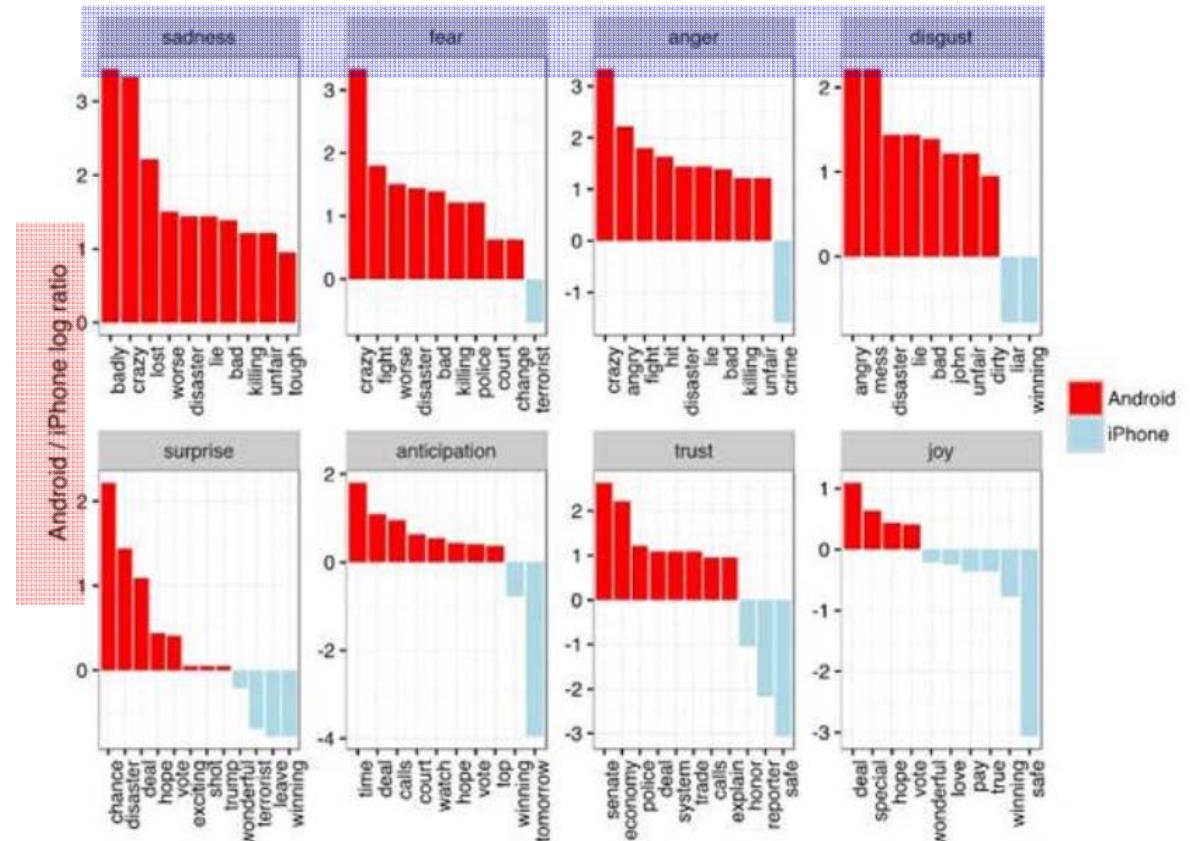
- 用 tidytext 當中的NRC Word-Emotion Association辭典，數據分析師將推文的用詞跟「**積極**、**消極**、**憤怒**、**期待**、**厭惡**、**恐懼**、**快樂**、**悲傷**、**驚訝**、**信任**」這十種情緒進行了**關聯分析**，結果發現：
- Android手機的推文中(共4901個字)，總共有321個字與「**憤怒**」的情感有關、有207個字與「**厭惡**」的情緒有關。
- 而透過**Poisson test** 分析後，更可明顯發現Android手機的推文更喜歡使用強烈情緒性的字眼，若透過**95%信賴區間**來看，就能看出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%。



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)

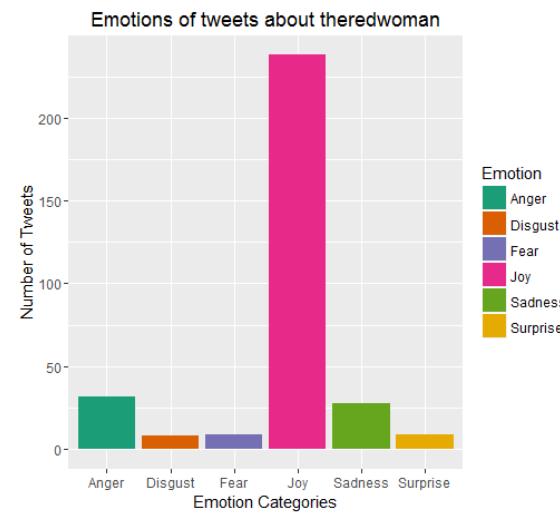


Infovis and Statistical Graphics: Different Goals, Different Looks

16/47

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: A Comment on Gelman and Unwin and an Invitation to Consider the Opportunities, 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>



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	
	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>
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)

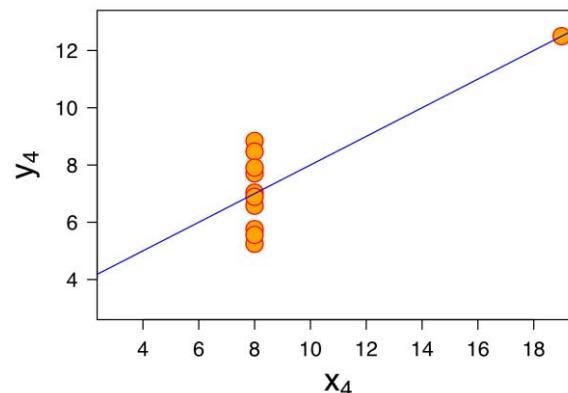
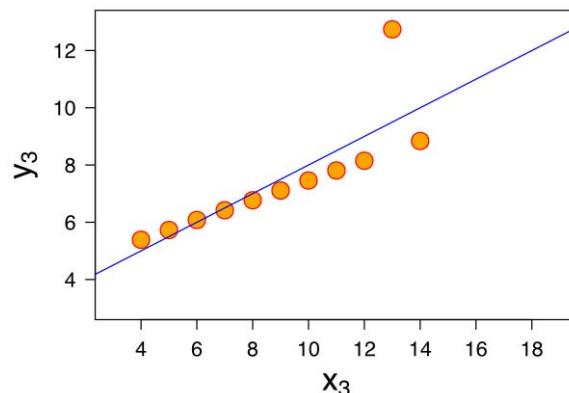
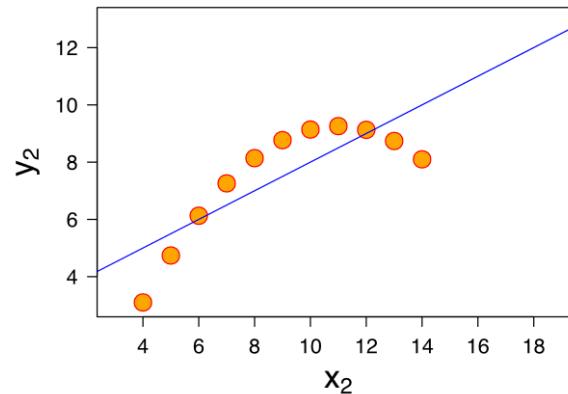
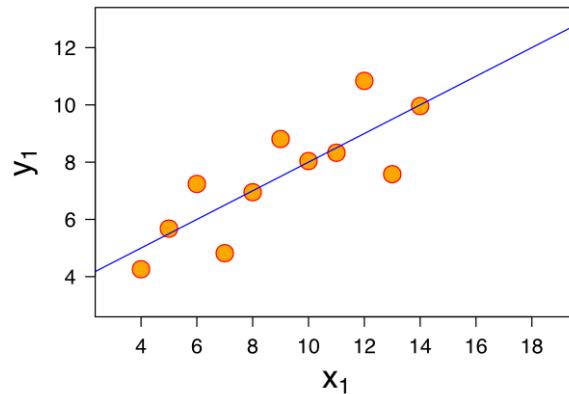
Linear regression line in each case: **$y = 3.00 + 0.500x$** (to 2 and 3 decimal places, respectively)

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)

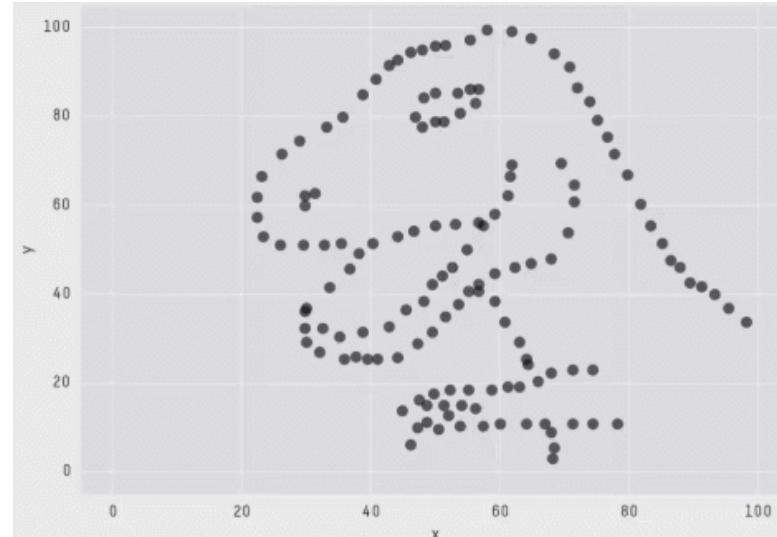




The Datasaurus Dozen

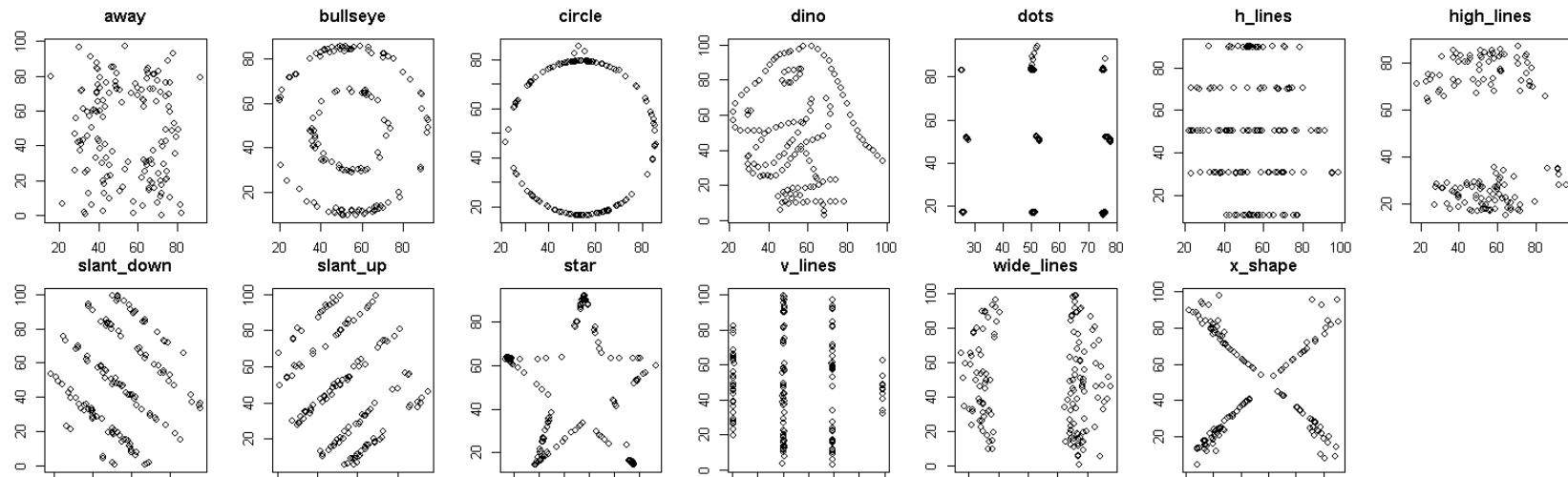
19/47

`install.packages("datasauRus")`



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

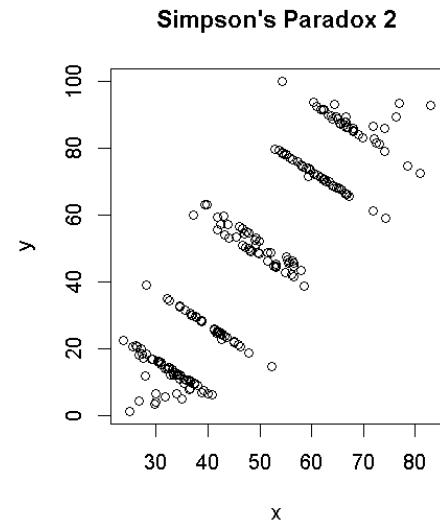
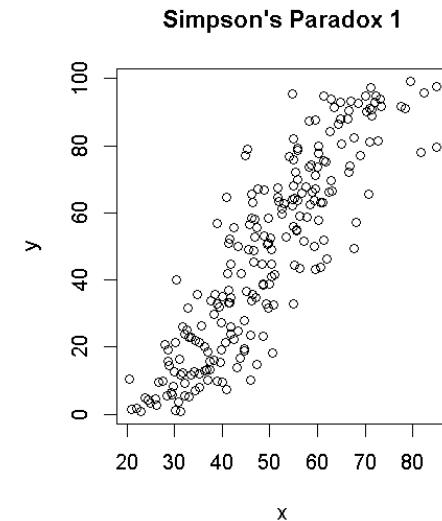
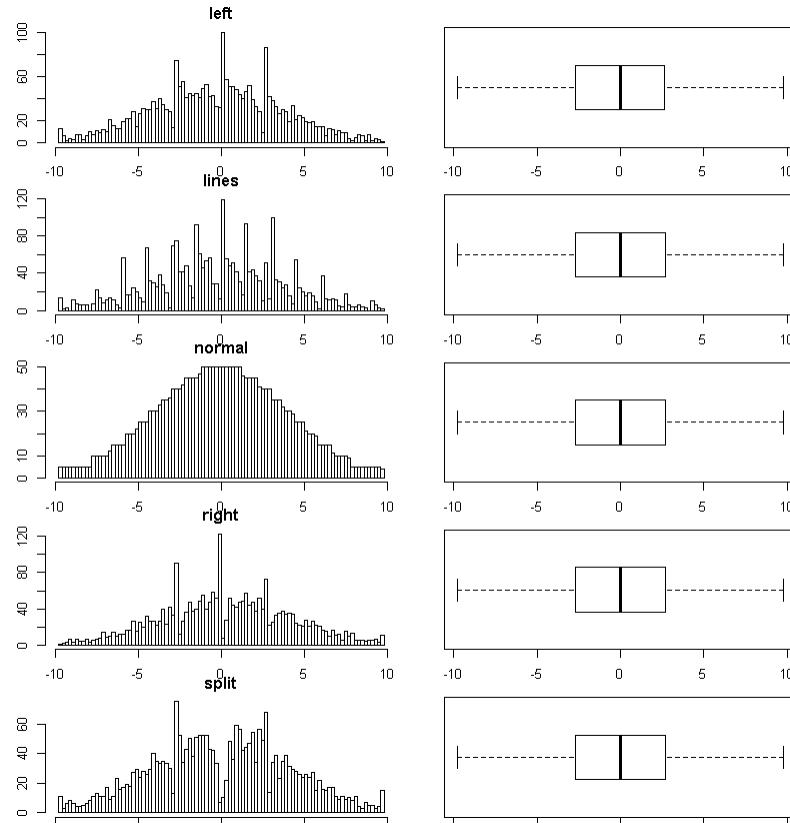
Justin Matejka and George Fitzmaurice, Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. <https://www.autodeskresearch.com/publications/samestats>





The Datasaurus Dozen More examples

20/47





Graphical Perception

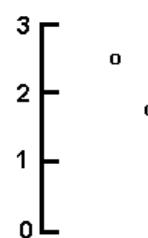
Human reception and comprehension of graphical information involves three fundamental perceptual task:

- **Detection:** the visual recognition of a geometric aspect that encodes a physical value. The basic information from the data must be discernible in the graph.
- **Assembly:** the process of discerning patterned regularities among the discrete elements of a graphical display.
- **Estimation:** the visual assessment of the relative magnitudes of two or more quantitative physical values.

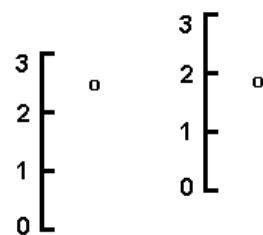
Graphical Perception Tasks.

Ordered from the most accurate to the least accurate (Jacoby, 1997)

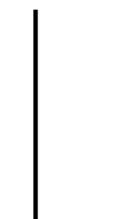
A. Position along a common scale



B. Position along common, nonaligned scales



C. Length



D. Angle



E. Slope, direction

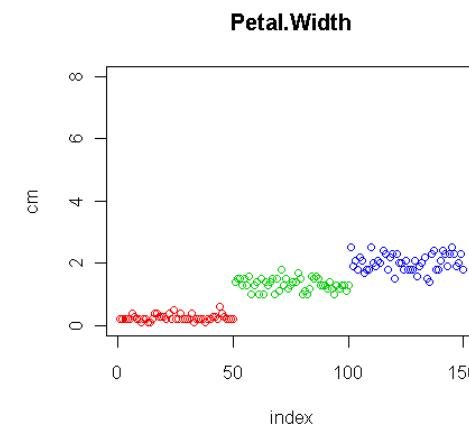
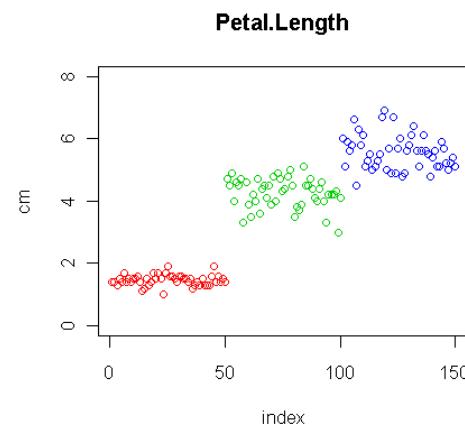
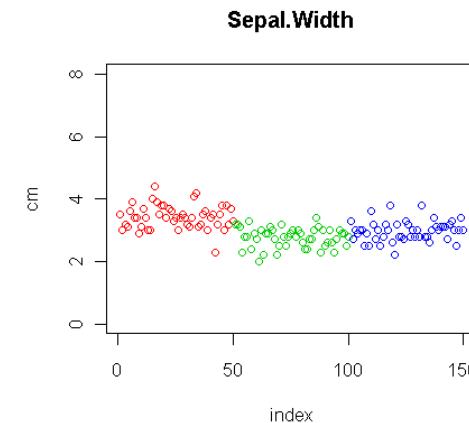
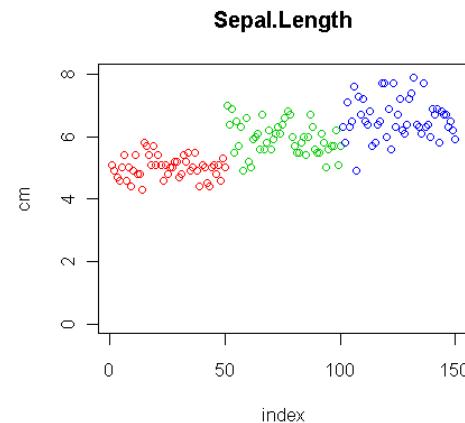
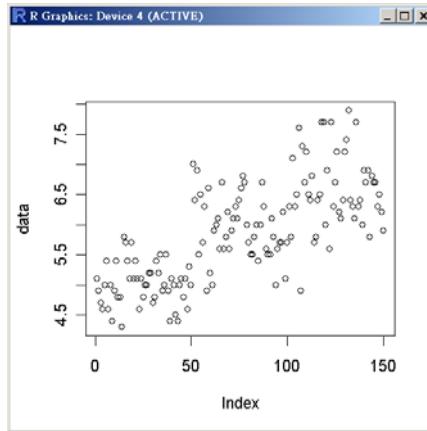
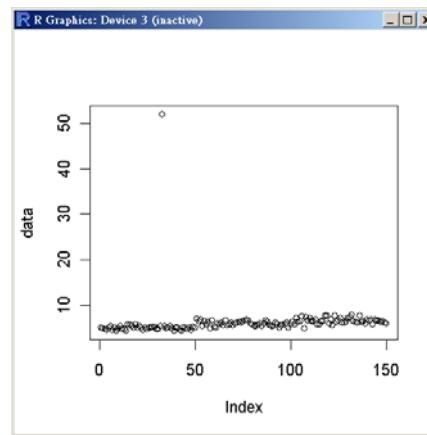


F. Area



Index Plot

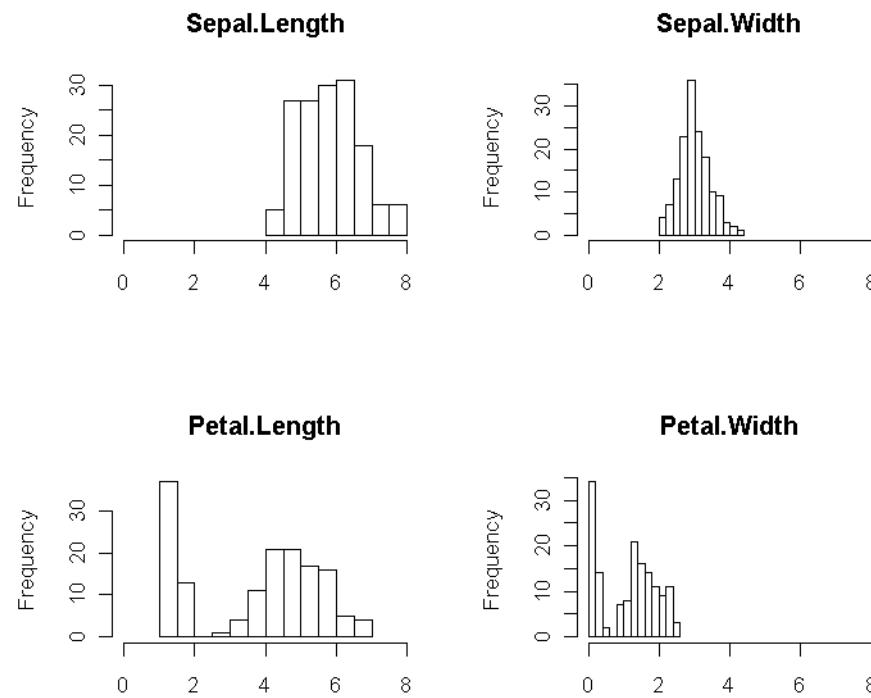
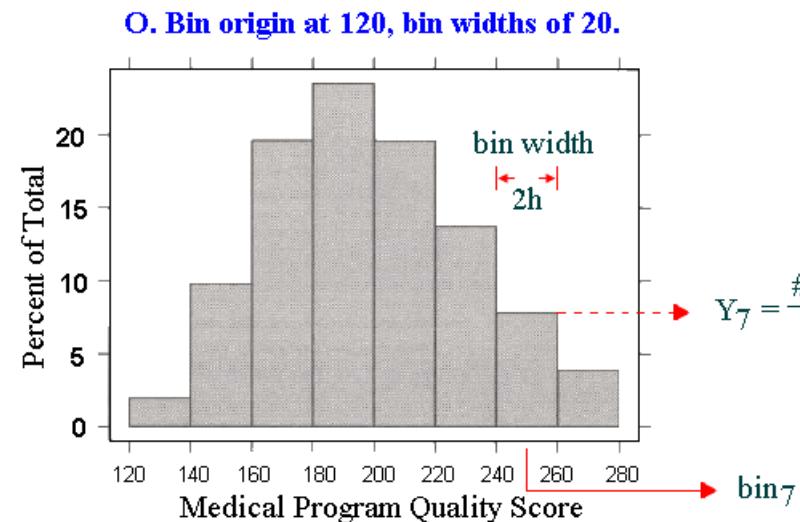
- Index plot takes a single argument which is a **continuous variable** and plots the **values** on the y axis, with the x coordinate determined by the **position** of the number in the vector.
- Useful for error checking.



直方圖 (Histogram) (1/3)

The histogram shows:

1. center of the data (location)
2. spread of the data (scale)
3. skewness of the data
4. presence of outliers
5. presence of multiple modes in the data.



Changes in bin origin and bin widths affect the shape of the histogram

直方圖 (Histogram) (2/3)

- $1/2h$ adjusts the height of each bar so that the total area enclosed by the entire histogram is 1.
- The area covered by each bar can be interpreted as the probability of an observation falling within that bar.

Disadvantage for displaying a variable's distribution:

- selection of **origin** of the bins.
- selection of **bin widths**.
- the very use of the bins is a distortion of information because any data **variability within** the bins cannot be displayed in the histogram.

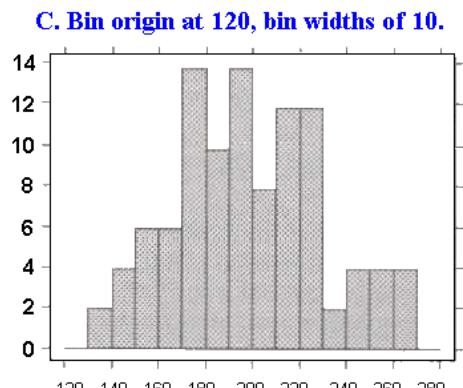
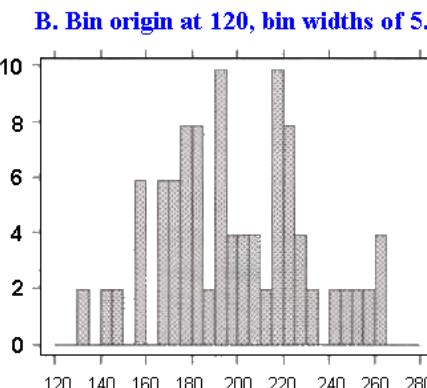
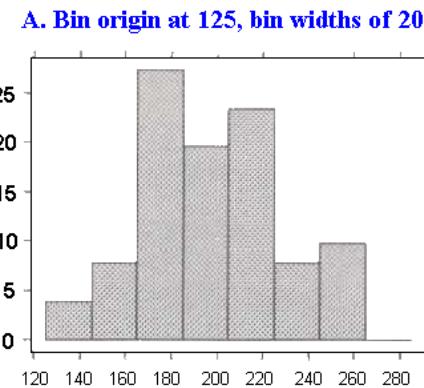
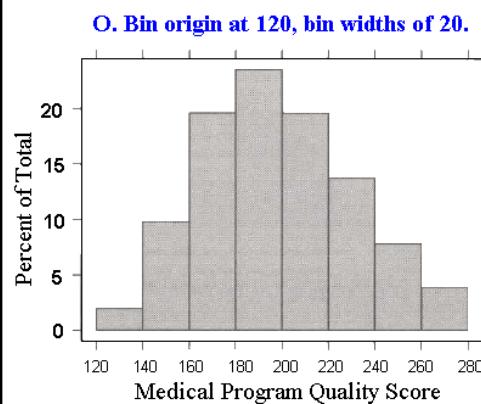
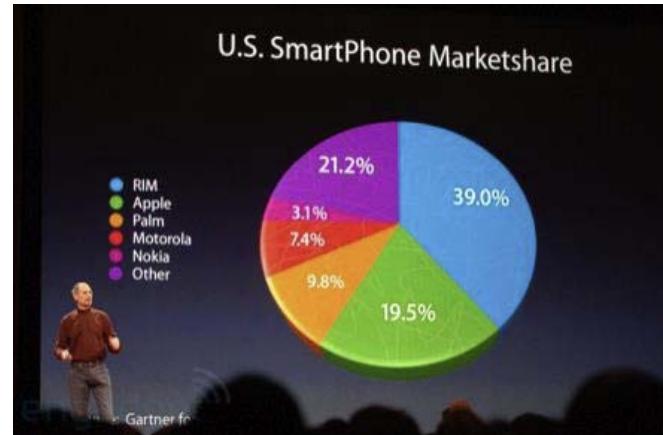
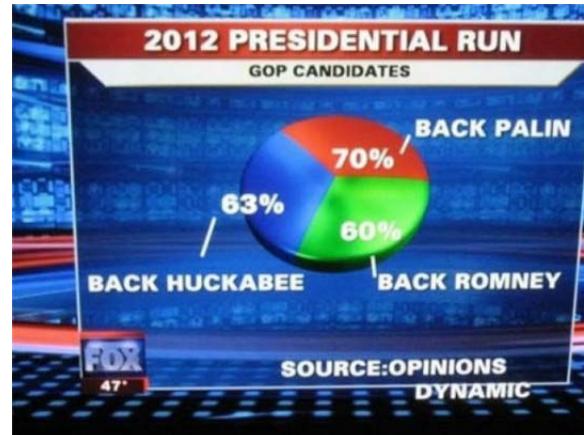


Figure Sources: Jacoby (1997) .

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/

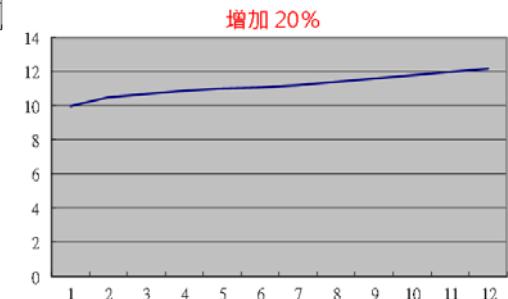
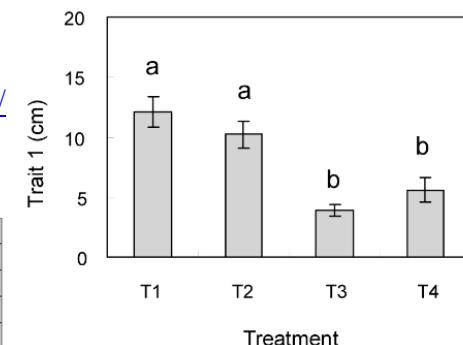
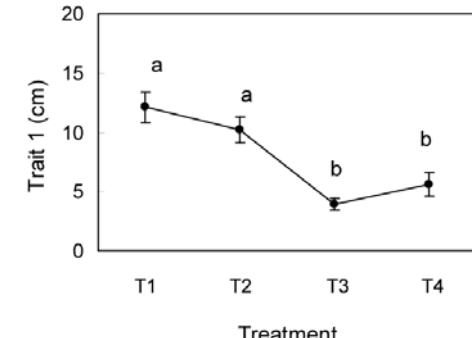
Bad Infographics: 11 Mistakes You Never Want To Make
<http://blog.visme.co/bad-infographics/>

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-int-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



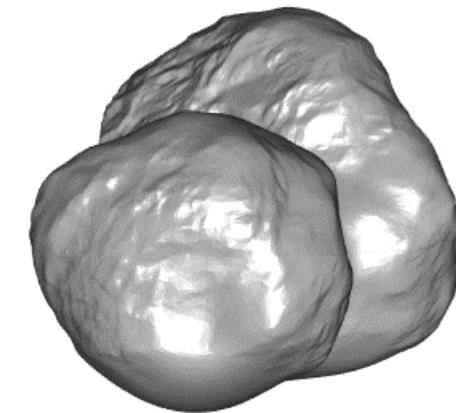
範例: rgl, explore a comet

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>

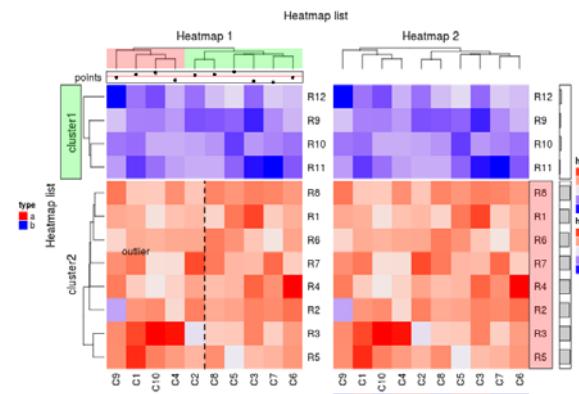
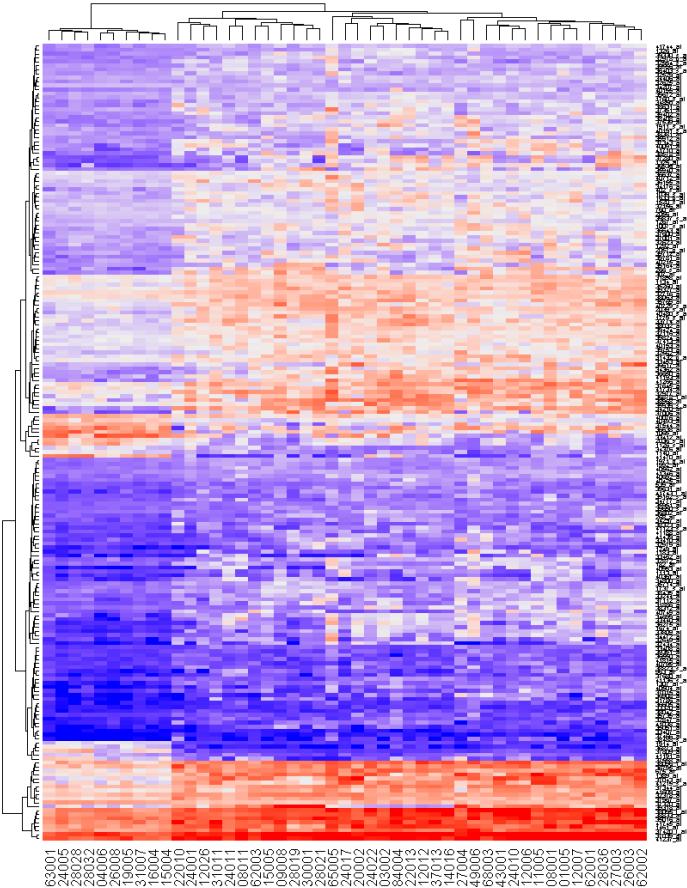
```
> open3d()
> # comet <- readOBJ(url("http://sci.esa.int/science-e/www/object/doc.cfm?fobjectid=54726"))
> comet <- readOBJ("ESA_Rosetta_OSIRIS_67P_SHAP2P.obj")
> class(comet)
[1] "mesh3d"  "shape3d"
> str(comet)
List of 6
 $ vb        : num [1:4, 1:31456] -0.394 0.402 0.443 1 -0.163 ...
 $ it        : 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

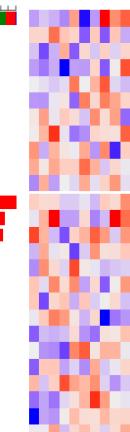
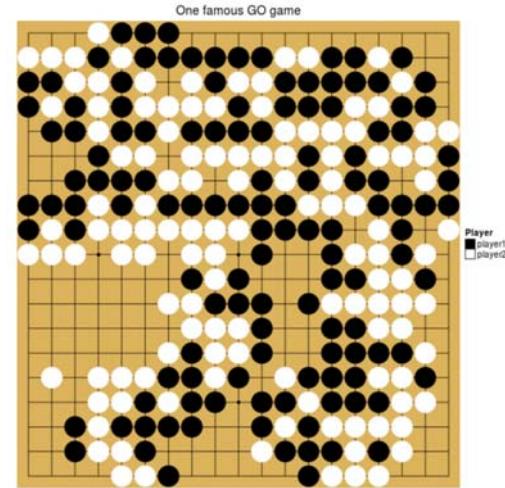
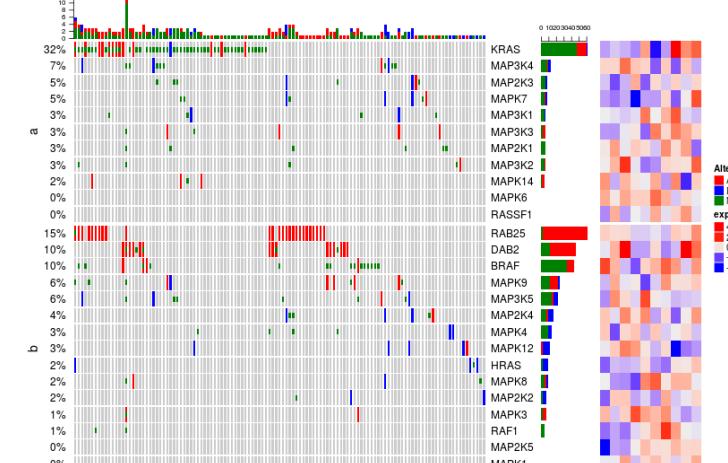
Complex Heatmap

<http://www.bioconductor.org/packages/devel/bioc/html/ComplexHeatmap.html>

```
> source("https://bioconductor.org/biocLite.R")
> biocLite("ComplexHeatmap")
> library(ComplexHeatmap)
> Heatmap(exprs(selected.eset))
```

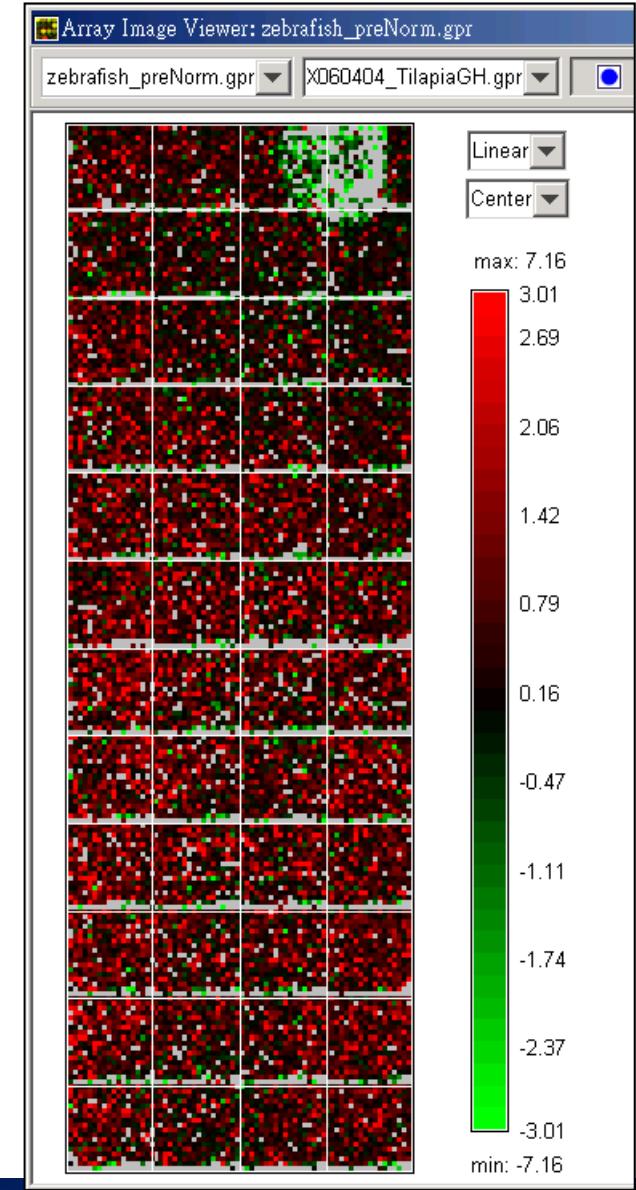
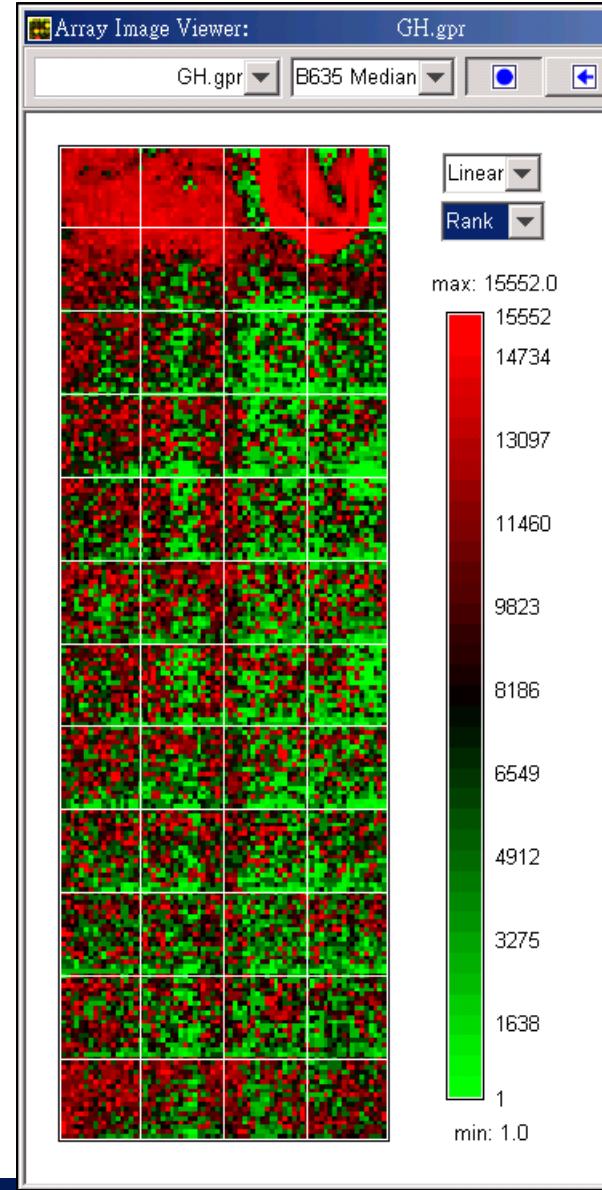
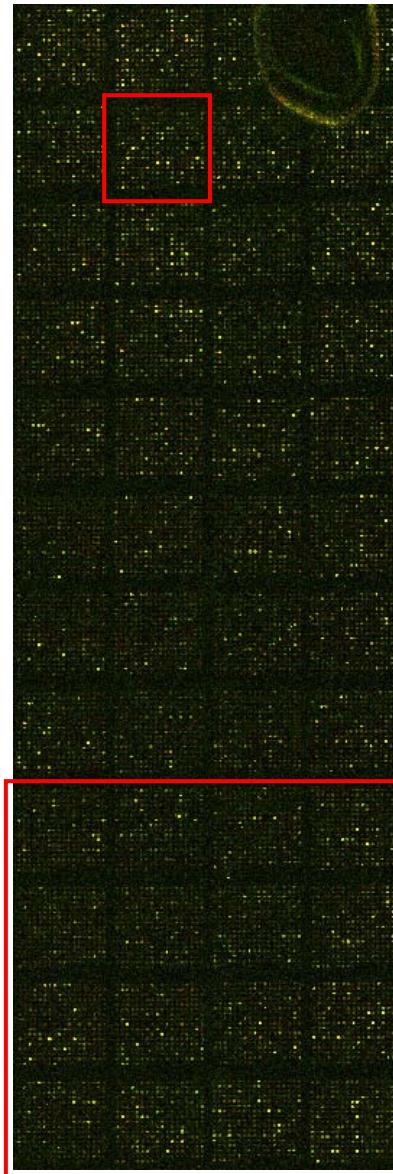


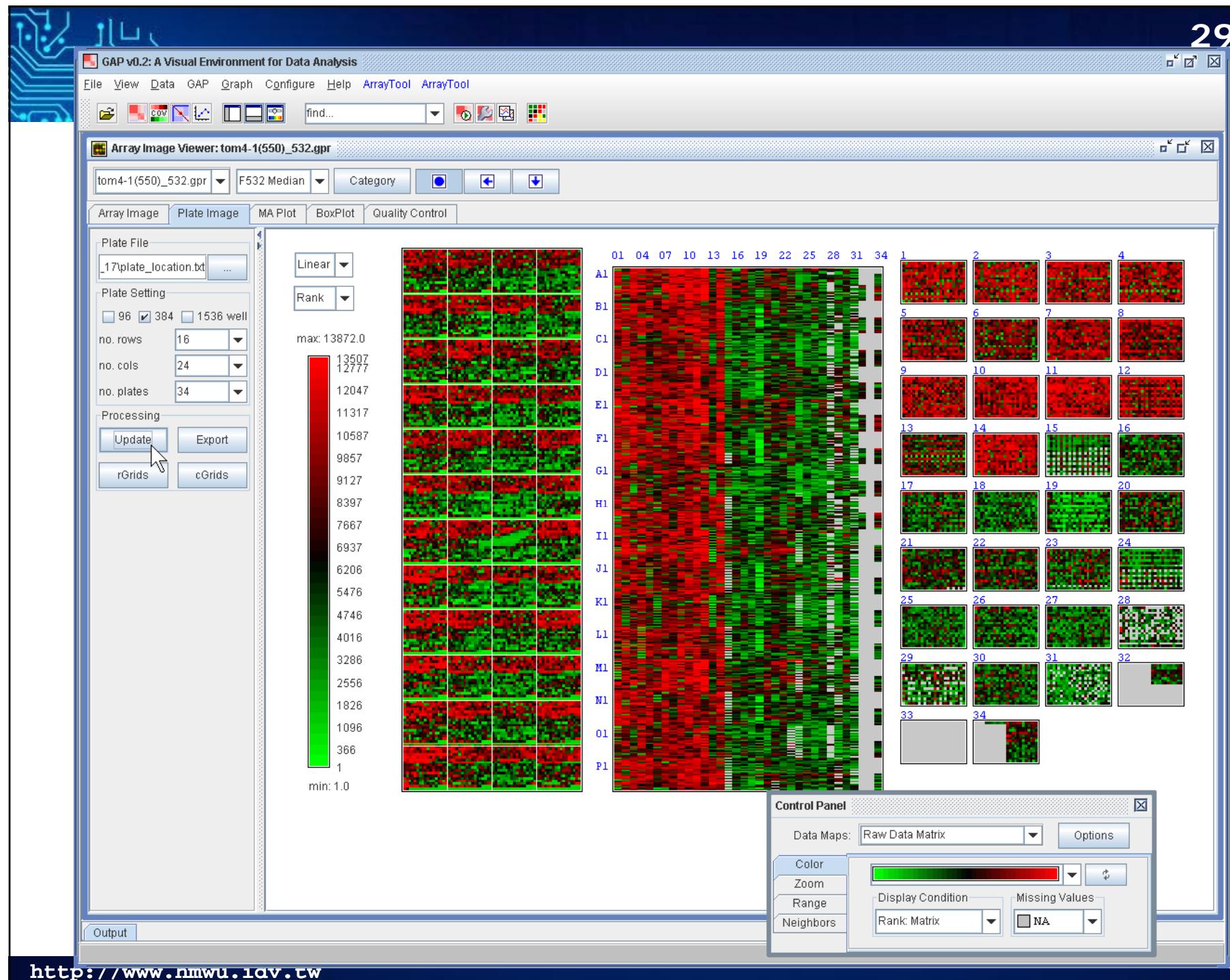
OncoPrint for TCGA Lung Adenocarcinoma, genes in Ras Raf MEK JNK signalling



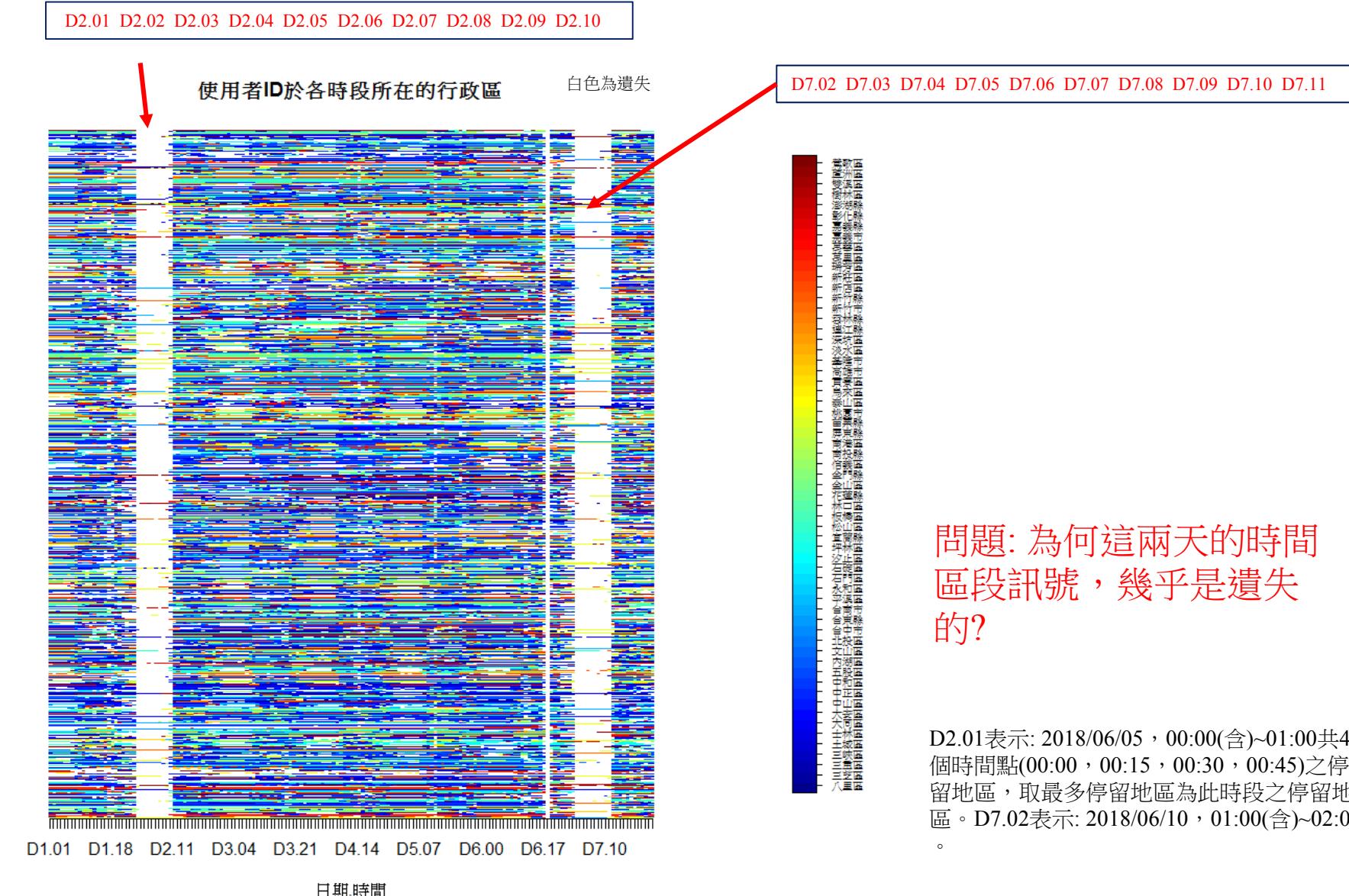
visualize multiple
genomic alteration
events by heatmap

Applications: Array Image





Applications



Applications: Eye-tracking, mouse clicking

eye-tracking research

Beyond the heat map

By Mike Bartels

There is no doubt that eye-tracking has begun. What was once a bit of a novelty has now been used in many diverse fields. Software have improved

and new software has eased the process. This coupled with a growing interest in eye-tracking has inspired a new generation of researchers.

Of all of the industries that have been more enthusiastic than the field of advertising, Web developers, package

Principles for interpreting eye-tracking data

How does this tool get us any closer to understanding our potential customers?



See also: <https://www.tobiipro.com/learn-and-support/learn/steps-in-an-eye-tracking-study/interpret/working-with-heat-maps-and-gaze-plots/>
<http://www.hmwu.idv.tw>

讀取外部影像檔案

```

> install.packages(c("tiff", "jpeg", "png", "fftwtools"),
  repos="http://cran.csie.ntu.edu.tw")
> library(EBImage) # (Repositories: BioC Software)
> Transformers <- readImage("Transformers07.jpg")
> (dims <- dim(Transformers))
[1] 300 421 3
> Transformers
Image
  colorMode : Color
  storage.mode : double
  dim        : 300 421 3
  frames.total : 3
  frames.render: 1

imageData(object)[1:5,1:6,1]
 [,1] [,2] [,3] [,4] [,5] [,6]
 [1,] 0 0 0 0 0 0
 [2,] 0 0 0 0 0 0
 [3,] 0 0 0 0 0 0
 [4,] 0 0 0 0 0 0
 [5,] 0 0 0 0 0 0
> plot(c(0, dims[1]), c(0, dims[2]), type='n',
+ xlab="", ylab="")
> rasterImage(Transformers, 0, 0, dims[1], dims[2])

```

```

> source("https://bioconductor.org/biocLite.R")
> biocLite("EBImage")

```



```

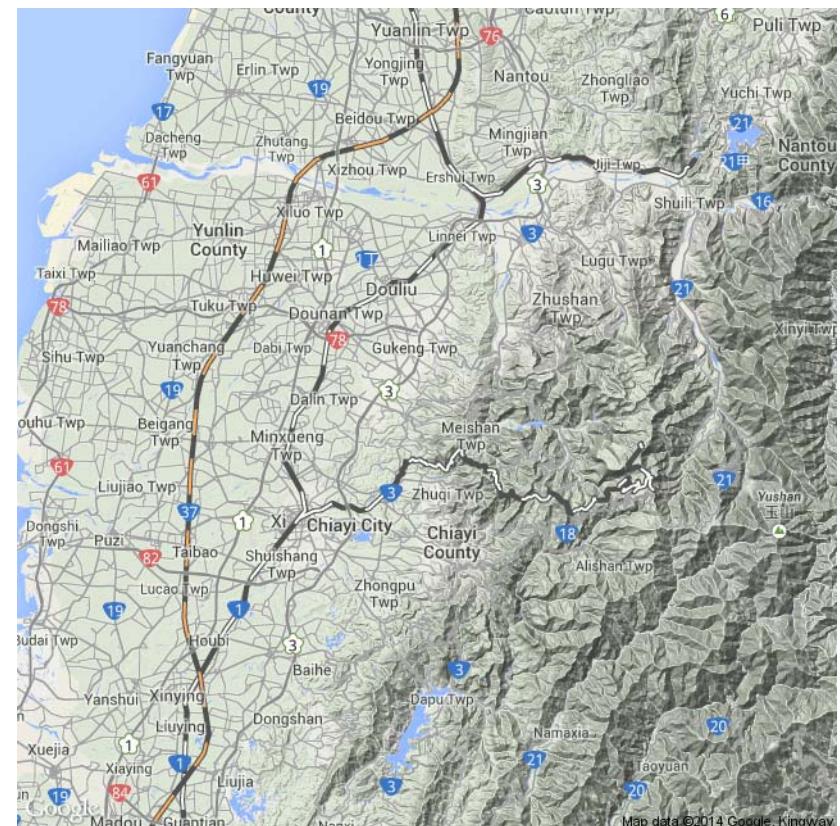
> #install.packages("jpeg")
> library(jpeg)
> Transformers <- readJPEG("Transformers07.jpg")

```

[https://en.wikipedia.org/wiki/Transformers_\(film\)](https://en.wikipedia.org/wiki/Transformers_(film))

台灣地圖

```
TaiwanMap <- GetMap(center=c(lat = 23.58, lon =120.58), zoom =7, destfile =  
"Taiwan1.png")  
TaiwanMap <- GetMap(center=c(lat = 23.58, lon =120.58), zoom = 10, destfile =  
"Taiwan2.png", maptype = "terrain")
```



於地圖上標記

```

my.lat <- c(25.175339, 25.082288, 25.042185, 25.046254)
my.lon <- c(121.450003, 121.565481, 121.614548, 121.517532)
bb = qbbox(my.lat, my.lon)
print(bb)

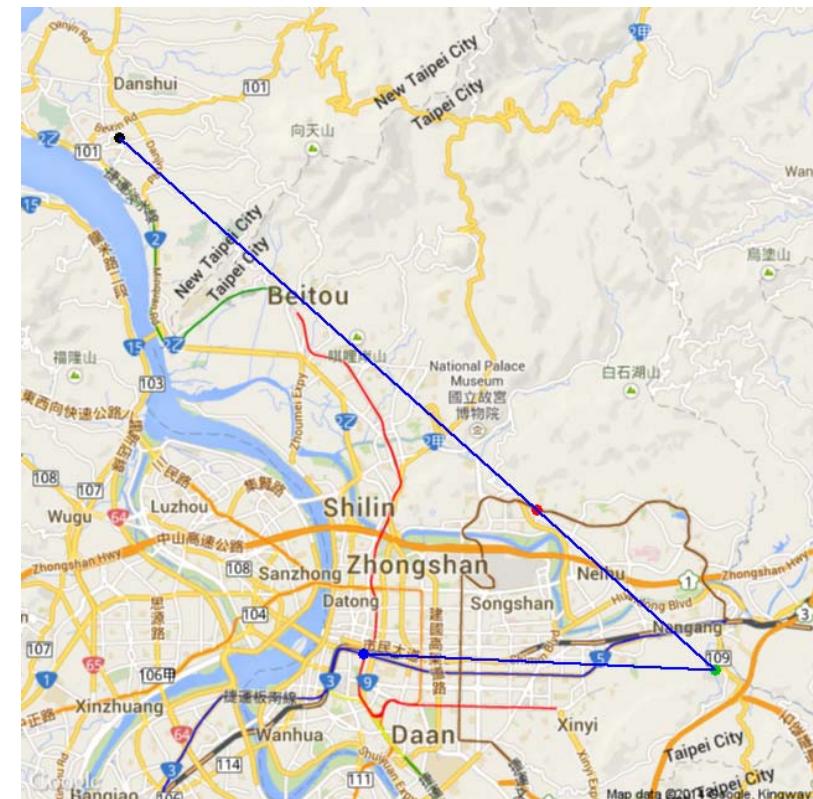
My.markers <- cbind.data.frame(lat = my.lat, lon = my.lon)
tmp <- PlotOnStaticMap(MyMap, lat = My.markers[, "lat"], lon = My.markers[, "lon"], destfile =
"my.png", cex=2.5, pch=20, col=1:4, add=F)

```

查詢經緯度

http://card.url.com.tw/realads/map_latlng.php

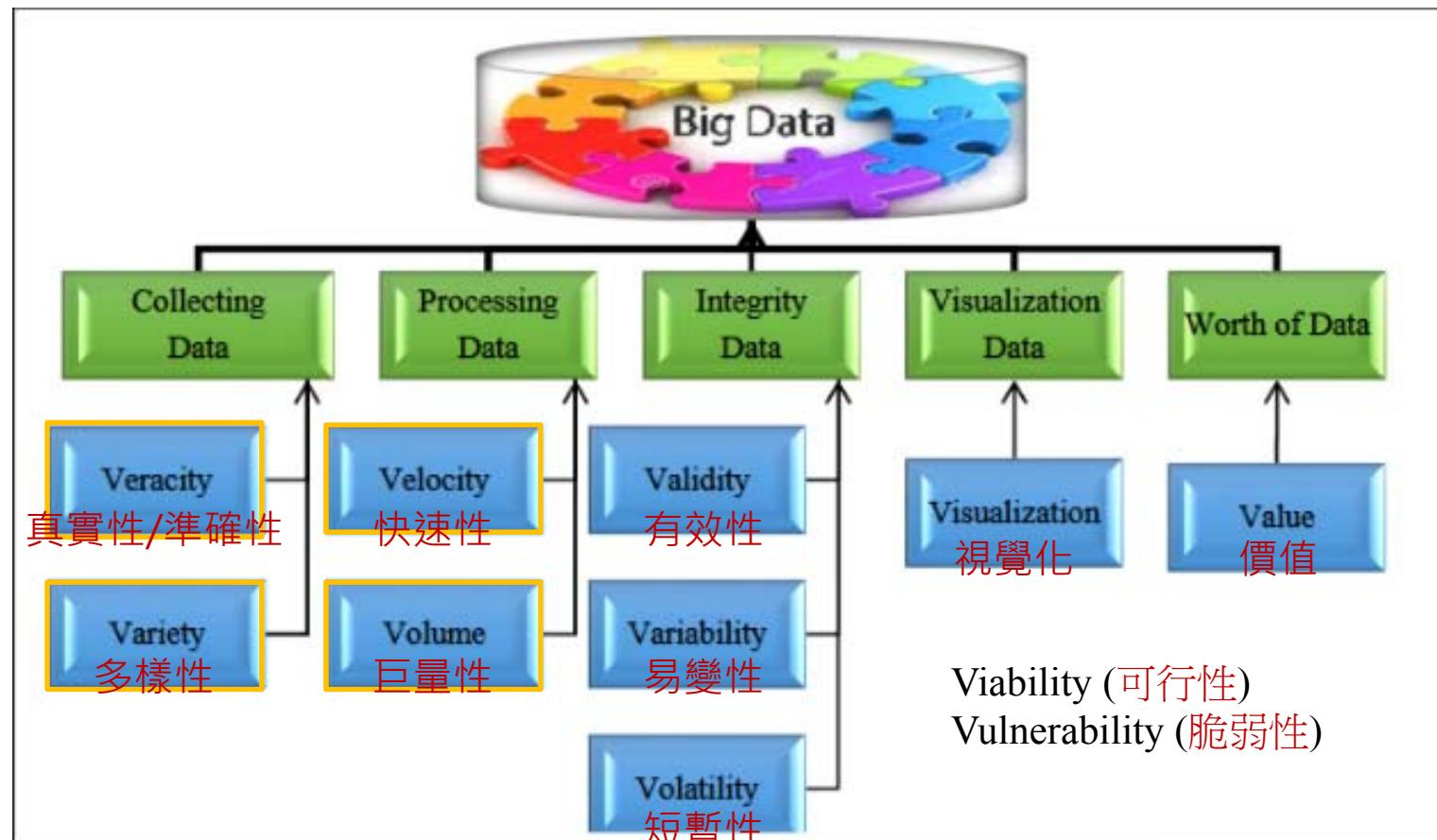
- 淡江大學 25.175339, 121.450003
- 台北市的地理中心位置: 內湖區環山路和內湖路一段
跟基湖路口: 25.082288, 121.565481
- 中研院 25.042185, 121.614548
- 捷運台北站: 25.046254, 121.517532



Big Data: The Era of 9 Vs

- Visualization:

- Visualization will be key to making big data an integral part of decision making.
- Visualization will be the only way to make big data accessible to a large audience.
- Visualization will be essential to the analysis of big data so it can be of highest value.



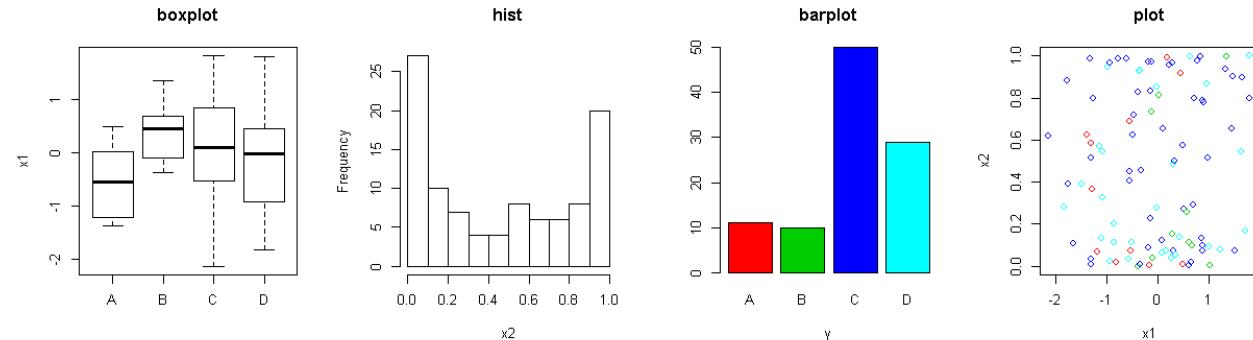
Categorization of Big Data V's

<http://blogs.systweak.com/2017/03/big-data-vs-represents-characteristics-or-challenges-of-big-data/>

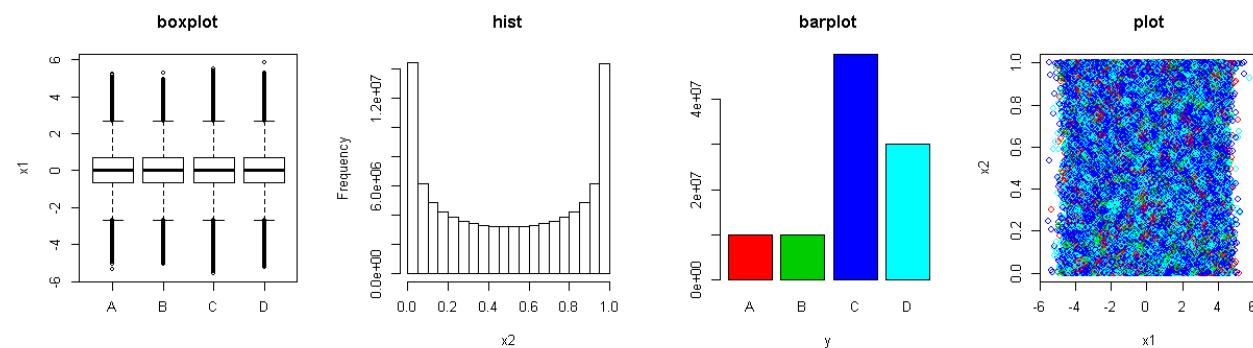
The Challenge of Visualizing Big Data

```
> n <- 1e+02
```

a large p?



```
> n <- 1e+08
```



```
> n <- 1e+02
> y <- as.factor(sample(LETTERS[1:4], n, replace=T, prob=c(0.1, 0.1, 0.5, 0.3)))
> x1 <- rnorm(n)
> x2 <- rbeta(n, 0.5, 0.5)
> xydata <- data.frame(y, x1, x2)
> par(mfrow=c(1,4))
> boxplot(x1~y, data=xydata, ylab="x1", main="boxplot")
> hist(x2, xlab="x2", main="hist")
> barplot(table(y), xlab="y", col = 2:5, main="barplot")
> plot(x1, x2, main="plot", col=as.integer(y)+1)
```

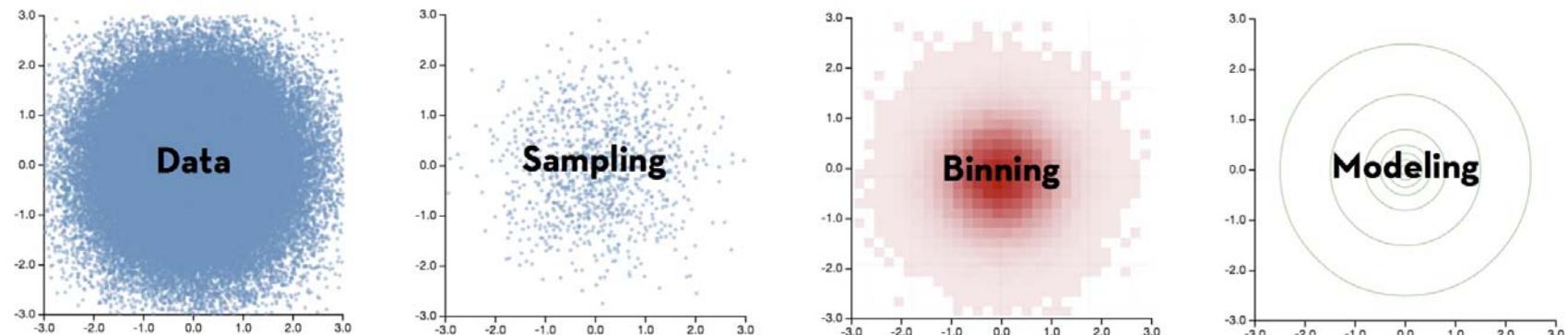
Two principles:
Look at Less Data;
or Look at Data Faster



How Can We Visualize and Interact with Billion+ Record Databases in Real-time?

37/47

- Two central challenges:
 - need to keep visualizations **perceptually effective** regardless of the number of input data points.
 - need to support **real-time interaction** to enable rapid and iterative exploratory analysis.
- Perceptual and interactive scalability should be limited by the chosen **resolution of the visualized data**, not the number of records.



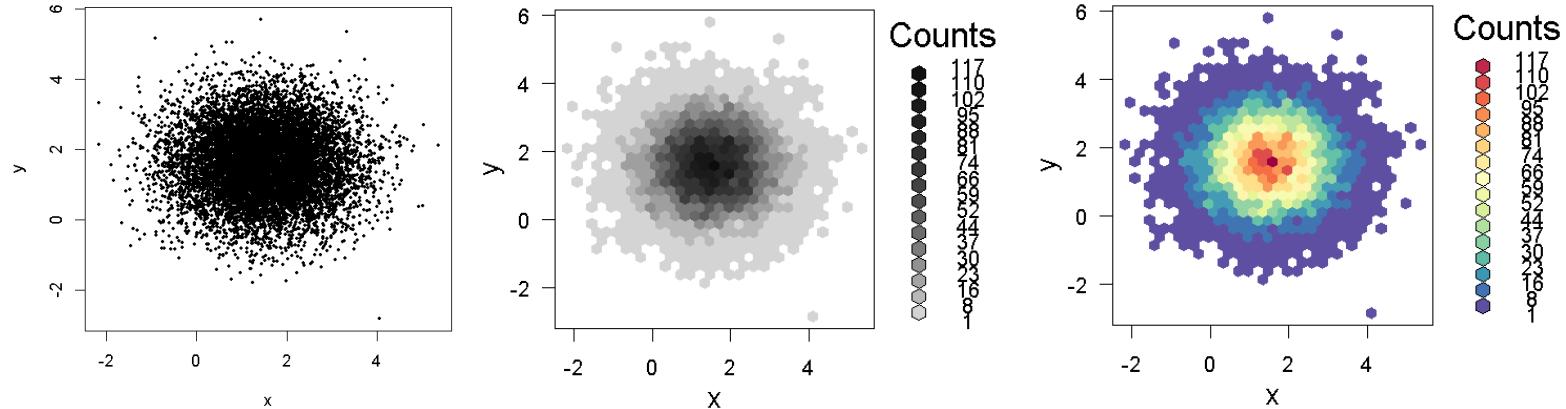
<http://skandel.github.io/slides/strata2013/part1>

<http://www.hmwu.idv.tw>



hexbin Package: Hexagonal Binning Routines

38/47



```
> x <- rnorm(mean=1.5, 10000)
> y <- rnorm(mean=1.6, 10000)
> my.data <- data.frame(x, y)
>
> pk <- c("RColorBrewer", "hexbin", "gplots")
> install.packages(pk, repos="http://cran.csie.ntu.edu.tw")
> library(RColorBrewer)
> # create rainbow color
> col_rb <- colorRampPalette(rev(brewer.pal(11, 'Spectral')))
> # scatterplot
> plot(my.data, pch=16, col='black', cex=0.5)
> library(hexbin)
> h <- hexbin(my.data) # create a hexbin object
> h
'hexbin' object from call: hexbin(x = my.data)
n = 10000  points in      nc = 598  hexagon cells in grid dimensions  36 by 31
> plot(h) # in grey level
> plot(h, colramp=col_rb) # rainbow color
```



tabplot: Tableplot, a Visualization of Large Datasets

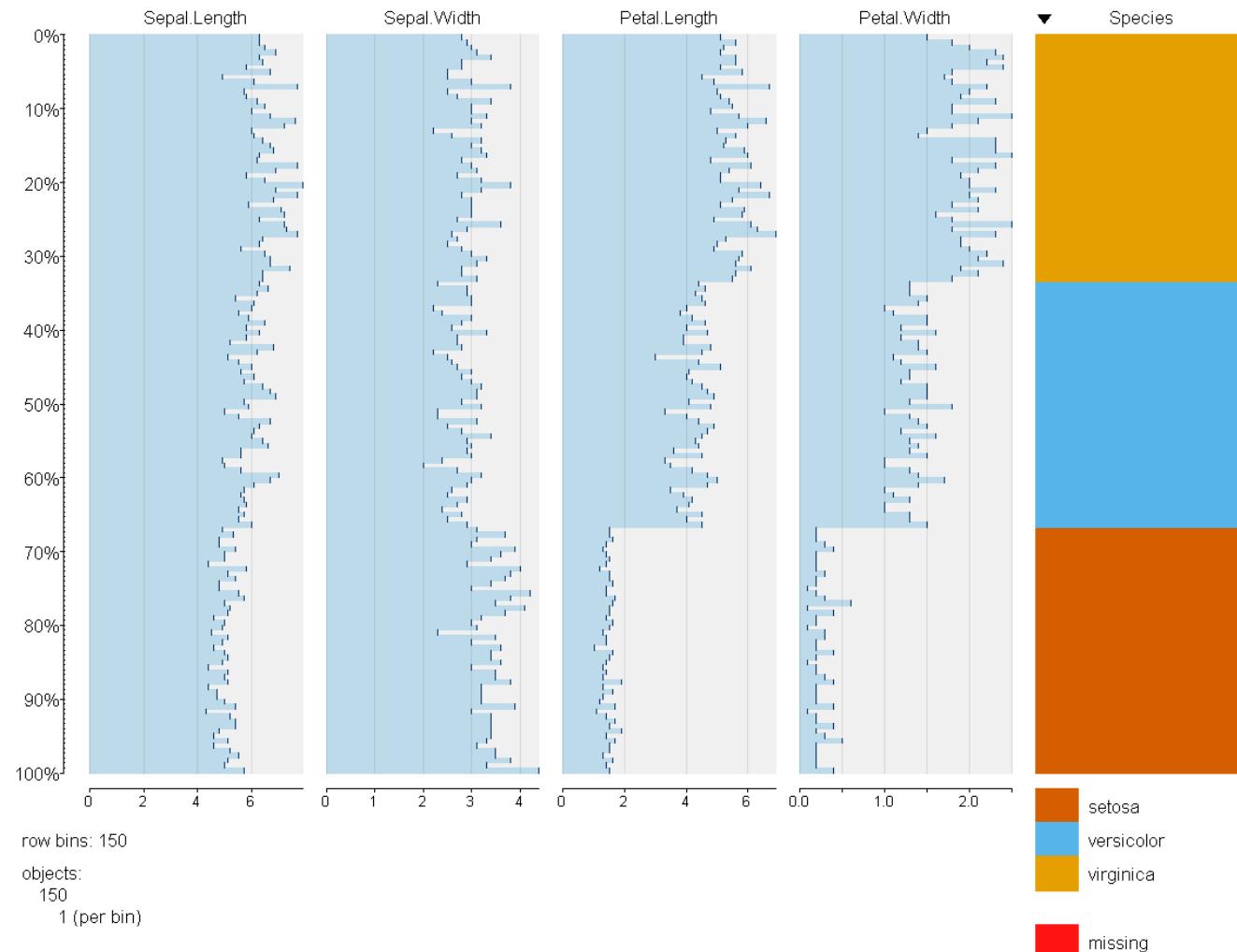
39/47

- A tableplot is a visualisation of a (large) dataset with a dozen of variables, both numeric and categorical.
 - Each column represents a variable and each row bin is an aggregate of a certain number of records.
 - Numeric variables are visualized as bar charts, and
 - categorical variables as stacked bar charts. Missing values are taken into account.
 - Also supports large '**ffdf**' datasets from the '**ff**' package.
 - <https://github.com/mtennekes/tabplot>
 - <https://cran.r-project.org/web/packages/tabplot/vignettes/tabplot-vignette.html>
- Tennekes, M., Jonge, E. de, Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots, Journal of Data Science 11 (1), 43-58.

```
tableplot(dat, select, subset = NULL, sortCol = 1, decreasing = TRUE,
  nBins = 100, from = 0, to = 100, nCols = ncol(dat), sample = FALSE,
  sampleBinSize = 1000, scales = "auto", numMode = "mb-sdb-ml",
  max_levels = 50, pals = list("Set1", "Set2", "Set3", "Set4"),
  change_palette_type_at = 20, rev_legend = FALSE, colorNA = "#FF1414",
  colorNA_num = "gray75", numPals = "OrBu", limitsX = NULL,
  bias_brokenX = 0.8, IQR_bias = 5, select_string = NULL,
  subset_string = NULL, colNames = NULL, filter = NULL, plot = TRUE,
  ...)
```

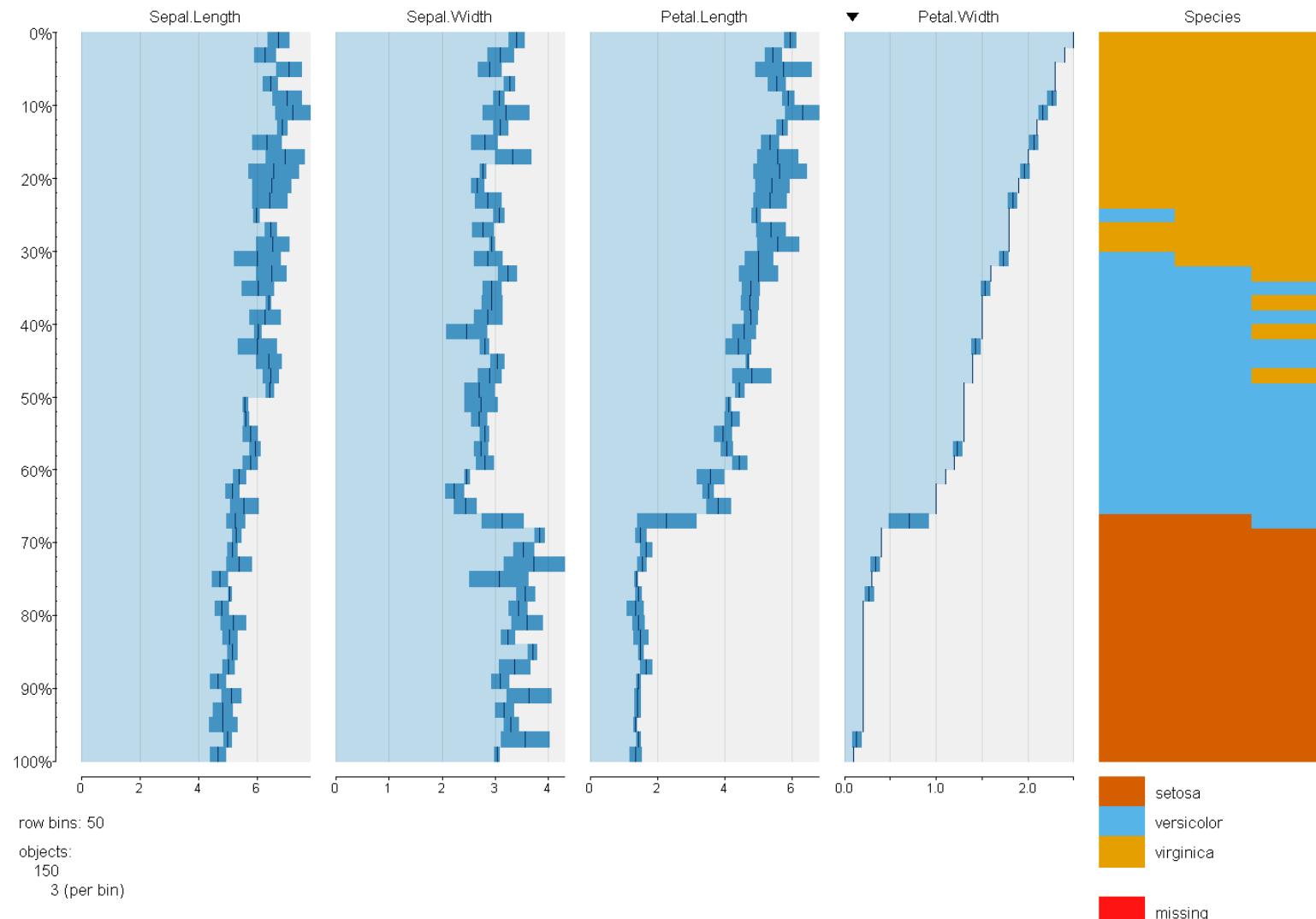
```
tableplot(iris, nBins=150, sortCol=5)
```

```
> install.packages("tabplot")
> library(tabplot)
> tableplot(iris, nBins=150, sortCol=5)
```



```
tableplot(iris, nBins=50, sortCol=4)
```

```
> tableplot(iris, nBins=50, sortCol=4)
```



tableplot(diamonds)

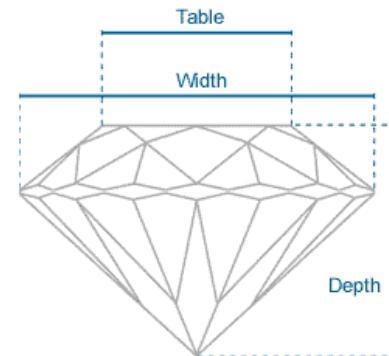
```

> require(ggplot2)
> data(diamonds)
> dim(diamonds)
[1] 53940   10
> head(diamonds)
# A tibble: 6 × 10
  carat      cut color clarity depth table price     x     y     z
  <dbl>     <ord> <ord>    <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
1 0.23     Ideal     E     SI2    61.5    55    326  3.95  3.98  2.43
2 0.21     Premium   E     SI1     59.8    61    326  3.89  3.84  2.31
3 0.23     Good      E     VS1     56.9    65    327  4.05  4.07  2.31
4 0.29     Premium   I     VS2     62.4    58    334  4.20  4.23  2.63
5 0.31     Good      J     SI2     63.3    58    335  4.34  4.35  2.75
6 0.24 Very Good   J     VVS2    62.8    57    336  3.94  3.96  2.48
> tableplot(diamonds)

```

Details

- price. price in US dollars (\$326--\$18,823)
- carat. weight of the diamond (0.2--5.01)
- cut. quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- colour. diamond colour, from J (worst) to D (best)
- clarity. a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))
- x. length in mm (0--10.74)
- y. width in mm (0--58.9)
- z. depth in mm (0--31.8)
- depth. total depth percentage = $z / \text{mean}(x, y) = 2 * z / (x + y)$ (43--79)
- table. width of top of diamond relative to widest point (43--95)



<http://www.lumeradimonds.com/diamond-education/diamond-cut>

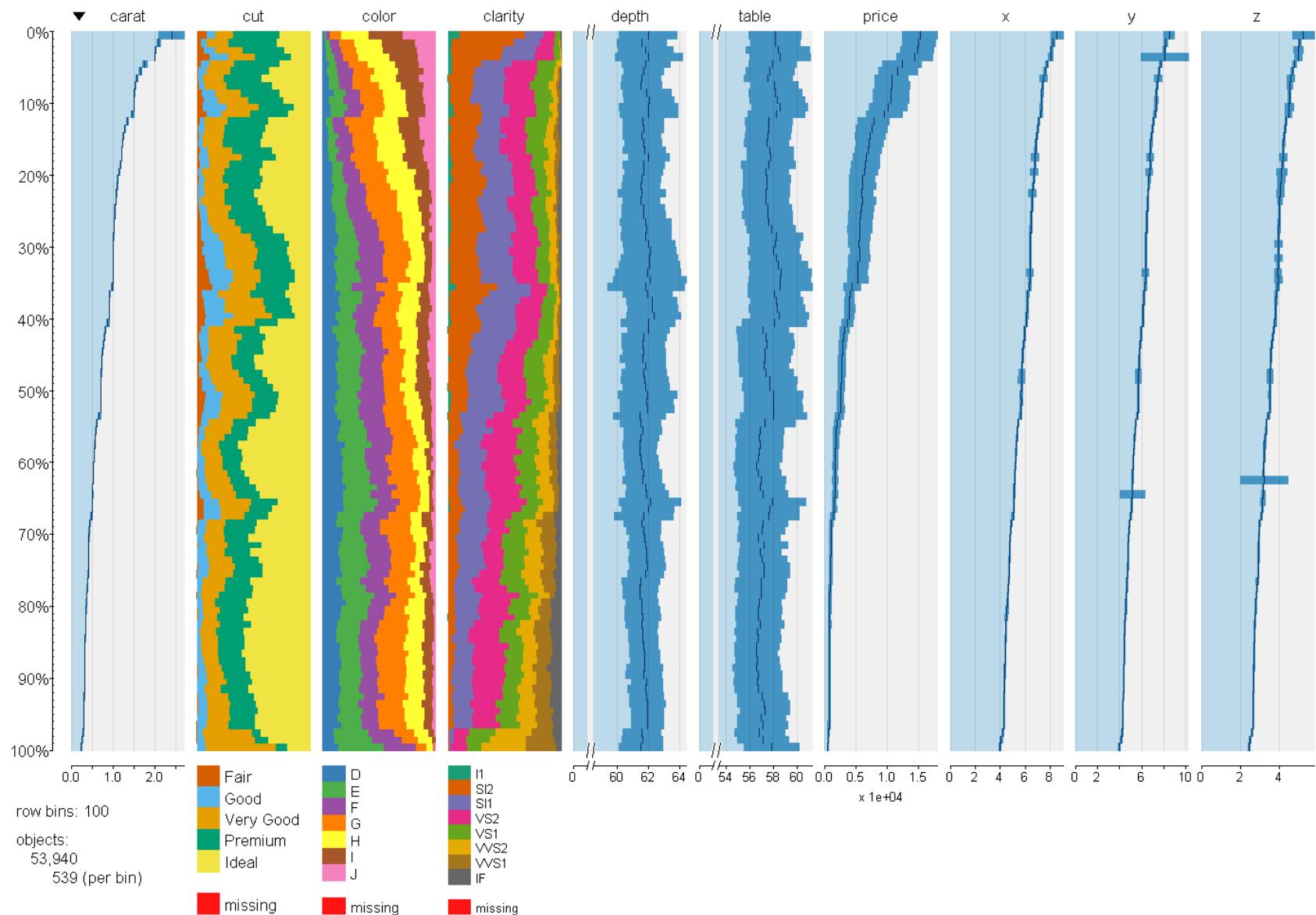
<http://docs.ggplot2.org/0.9.3.1/diamonds.html>

<http://yourdiamondteacher.com/diamond-4cs/cut/>

Excellent Ideal	
Very Good	
Good	
Fair	
Poor	



```
tableplot(diamonds)
```

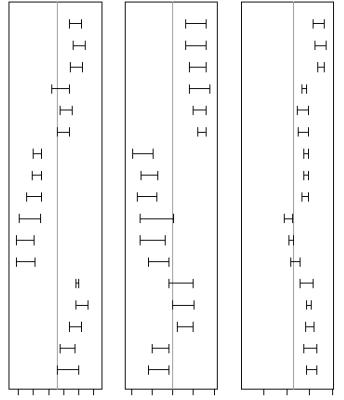




Symbolic Data Analysis (Billard and Diday, JASA 2003)

44/47

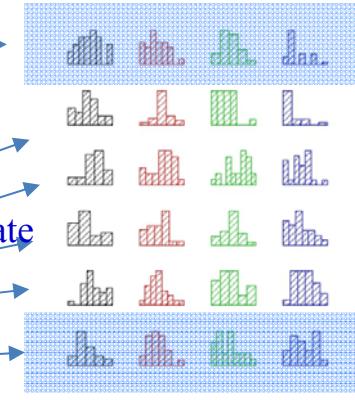
Symbolic data table
(intervals)



The classical data table

	X_1	X_2	...	X_3	...	X_p
s_1	x_{11}	x_{12}	...	x_{13}	...	x_{1p}
s_2	x_{21}	x_{22}	...	x_{23}	...	x_{2p}
:	:	:	...	:	...	:
s_{n_1}	$x_{n_1 1}$	$x_{n_1 2}$...	$x_{n_1 3}$...	$x_{n_1 p}$
s_{n_1+1}	$x_{(n_1+1)1}$	$x_{(n_1+1)2}$...	$x_{(n_1+1)3}$...	$x_{(n_1+1)p}$
:	:	:	...	:	...	:
s_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
:	:	:	...	:	...	:
s_N	x_{N1}	x_{N2}	...	x_{N3}	...	x_{Np}

Symbolic data table
(histograms)



aggregate

aggregate

Symbolic Variable

Numerical

Single-valued
(real or integer)
Multi-valued
(finite set of real)

Interval-valued
 $[a_i, b_i]$

Modal
 (η_i, π_i)
Multi-valued
Modal
 $\{(\eta_{i1}, \pi_{i1}), \dots, (\eta_{ik}, \pi_{ik})\}$
Histogram
 $\{(I_{il}(p_l), \dots, I_{ik}(p_k)\}$

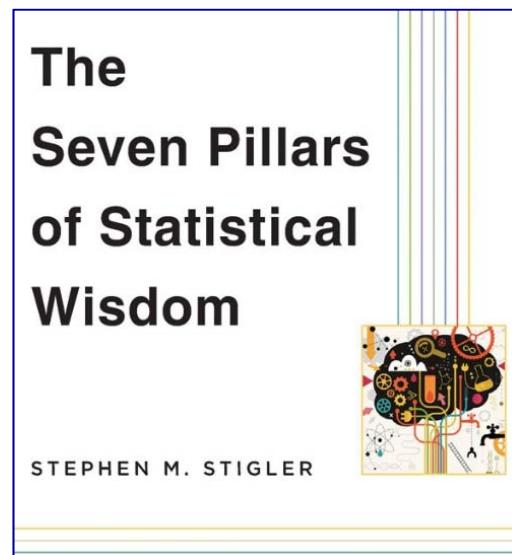
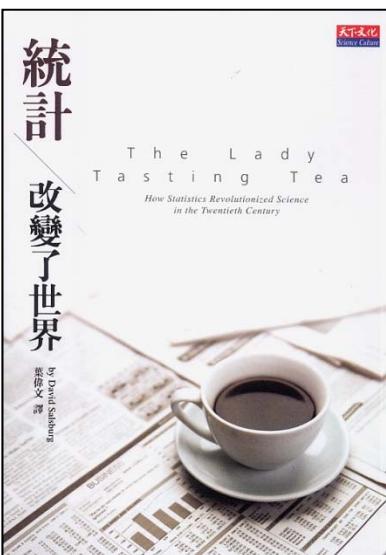
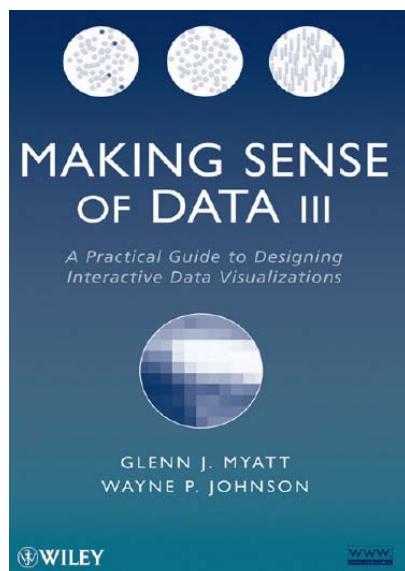
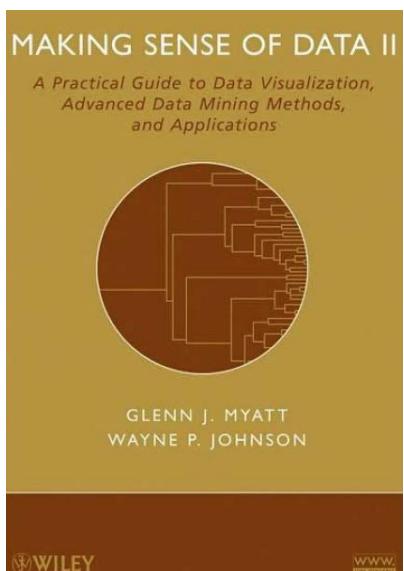
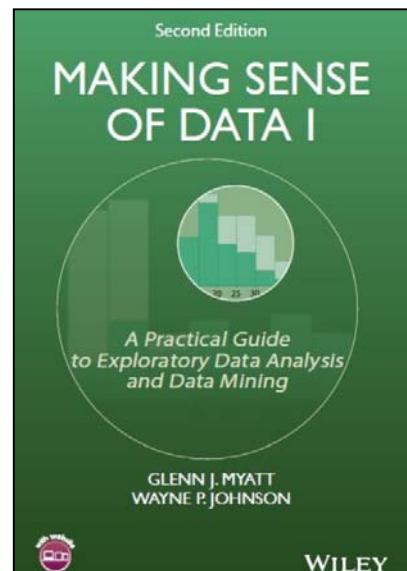
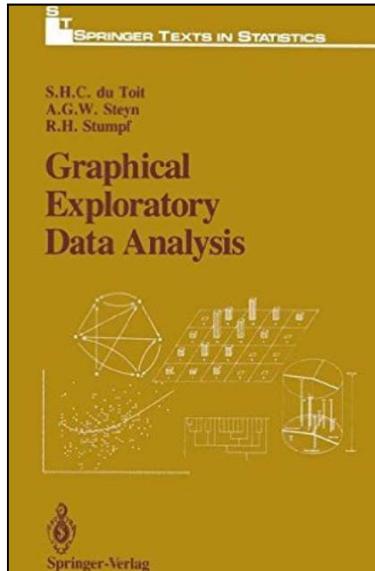
Function
 (pf, cdf)
Time Series
Stochastic
Process
Model

Categorical

Single-valued
(nominal or ordinal)
Multi-valued
(finite set of categories)
Modal
 (ξ_i, π_i)
Multi-valued
Modal
 $\{(\xi_{il}, \pi_{il}), \dots, (\xi_{ik}, \pi_{ik})\}$

π_i : weight, frequency, probability
 η_i : numerical
 ξ_i : categorical

推薦書目



- 1 **AGGREGATION** From Tables and Means to Least Squares
- 2 **INFORMATION** Its Measurement and Rate of Change
- 3 **LIKELIHOOD** Calibration on a Probability Scale
- 4 **INTERCOMPARISON** Within-Sample Variation as a Standard
- 5 **REGRESSION** Multivariate Analysis, Bayesian Inference, and Causal Inference
- 6 **DESIGN** Experimental Planning and the Role of Randomization
- 7 **RESIDUAL** Scientific Logic, Model Comparison, and Diagnostic Display

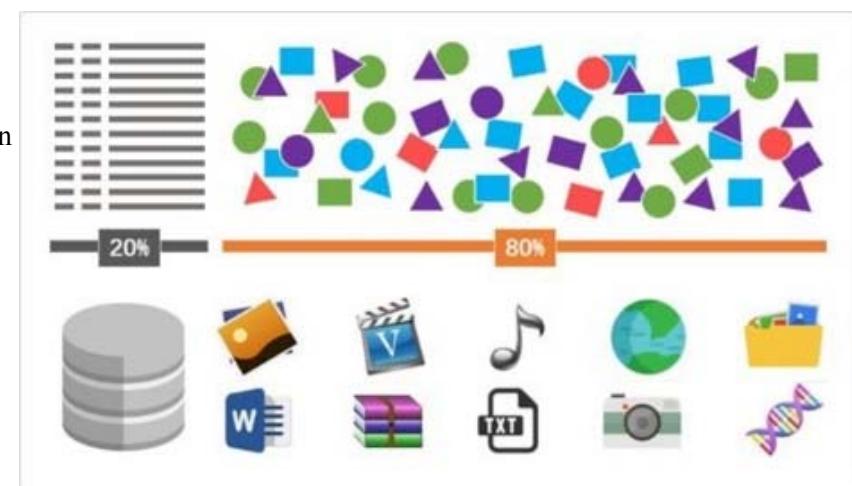
資料無邊界!

Statistician



Image source: <http://marketbusinessnews.com/financial-glossary/what-is-a-statistician>

- 統計教學裡的範例幾乎都是結構性的數據。
- 大數據時代，80%的資料是非結構性的，統計課程如何面對？



≡ Data Science Education

資料科學時代的統計教學

Statistics Education in the Data Science Era

2018.3.9下午4點起至2018.3.10中午12:30 於

台南國立成功大學光復校區統計學系

2018統計教學工作坊

18~19 May, 2018
Lounge (2F),
Institute of Statistical Science, AS



未來方向？

The screenshot shows the homepage of the Academia Sinica Data Science Statistical Cooperation Center. The header features a landscape photograph of a long, straight road receding into the distance under a cloudy sky. Overlaid on the top half of the image is the text "ACADEMIA SINICA DATA SCIENCE STATISTICAL COOPERATION CENTER" and "資料科學統計合作社 - 統計暨資料科學諮詢服務". Below the header is a dark navigation bar with three items: "關於我們" (About Us), "服務與合作" (Services and Cooperation), and "聯絡我們" (Contact Us). A sub-navigation menu titled "Statistics at the Bench: the good, the bad and the ugly" is visible. The main content area contains a sidebar with various research domain categories and a large text block about the cooperation center's mission. On the right side of the page, there is a photograph of a telephone handset on a desk.

趙民德，1999，「統計已死，統計萬歲！」第八屆南區統計研討會演說稿

趙民德
台灣

趙民德，國立台灣大學數學系畢業、美國加州大學柏克萊分校統計博士。在美國求學及工作多年後，1982年回台灣籌設中央研究院統計學研究所，該所於1987年正式成立，並正名為統計科學研究所。國內統計學有今日的發展，以及能在世界佔一席之地，功不可沒。

在文學成就上，名家王鼎鈞以「詩的精緻，劇的張力，散文的鋪陳」肯定其業餘小說家的地位。

“統計有沒有死？會不會萬歲？

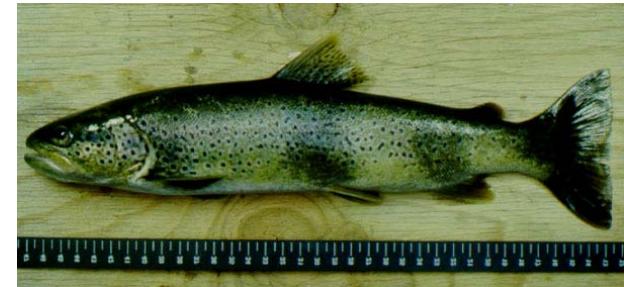
只要有米倉，就會有老鼠；只要有數據，就會發展處理數據的方法。但是不是叫做統計學、或者叫做computer science 的data mining，就要看這一代的統計人如何因應變局。”



進階選讀

例子1: The Doubs Fish Data

- Fish communities were good biological indicators of these water bodies: Verneaux (1973) (Verneaux et al. 2003) proposed to use fish species to characterize ecological zones along European rivers and streams. (River Doubs, 杜河)
- Verneaux proposed a typology in four zones, and he named each one after a characteristic species:
 - the trout (鱒魚 · 鮑鱒魚) zone (from the brown trout *Salmo trutta fario*),
 - the grayling (鱒魚) zone (from *Thymallus*),
 - the barbell (鰻, 有觸鬚的魚) zone (from *Barbus*) and
 - the bream (歐鯿, 鯉科淡水魚) zone (from the common bream *Abramis brama*).
- The two upper zones are considered as the “**Salmonid (鮭魚) region**” and the two lowermost ones constitute the “**Cyprinid (鯉科之魚) region**” .



D. Borcard et al., Numerical Ecology with R, Use R, DOI 10.1007/978-1-4419-7976-6_2, © Springer Science+Business Media, LLC 2011

Image Source:

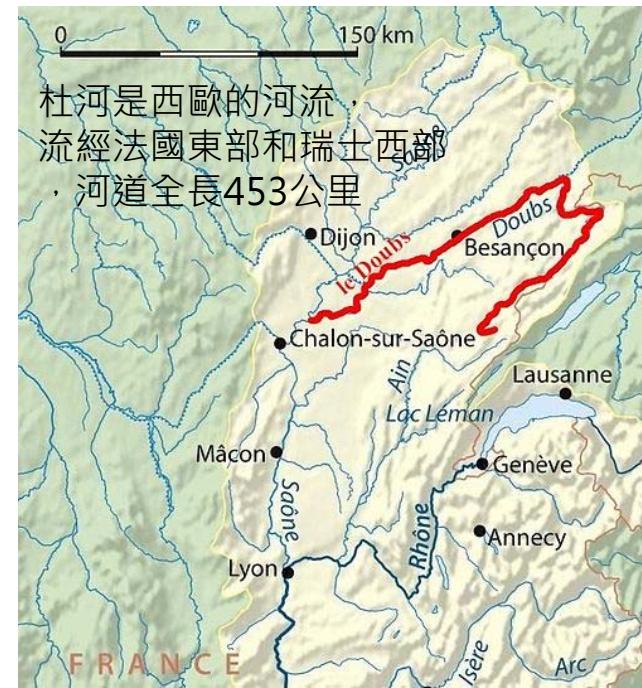
http://www.qub.ac.uk/bb-old/prodohl/TroutConcert/images/gallery/c_lagiader-me07-18-trout.jpg
<http://www.bamboorods.ch/guiding/bilder/grayling2.jpg>
https://en.wikipedia.org/wiki/Barbus_barbus#/media/File:Barbel.jpg
<http://www.ultimateangling.co.za/index.php?topic=15775.0>

River Doubs Map



Source: [https://en.wikipedia.org/wiki/Doubs_\(river\)](https://en.wikipedia.org/wiki/Doubs_(river))

- 背景知識、問題
- 資料收集方式、變數資訊
- 參與人角色 (分析者、廠商、顧主、客戶、居民、...)
- 資料處理、探索(分析)方法
- 呈現: 資料/過程/結果。





The Doubs Fish Data: 檔案

- The Doubs data set have been collected at **30 sites** along the Doubs River (near the France–Switzerland border in the Jura Mountains.)
- **DoubsSpe**: contains coded abundances (豐富充足) of **27 fish species**.
- **DoubsEnv**: contains **11 environmental variables** related to the hydrology, geomorphology and chemistry of the river.
- **DoubsSpa**: contains the **geographical coordinates** (Cartesian, X and Y) of the sites.

The image shows three separate Excel windows side-by-side, each displaying a CSV file:

- DoubsSpe.csv**: This table contains 12 rows of data representing 27 fish species. The columns are labeled CHA, TRU, VAI, LOC, OMB, BLA, and HO'. The data shows various counts or codes for different species across 12 sampling sites.
- DoubsEnv.csv**: This table contains 12 rows of environmental data. The columns are das, alt, pen, deb, pH, and dur. The data includes measurements like water depth (das), altitude (alt), and pH levels.
- DoubsSpa.csv**: This table contains 13 rows of geographical coordinates. The columns are site number (1-13), X, and Y. The data provides Cartesian coordinates for each of the 13 sampling sites.



The Doubs Fish Data: 前置處理

- Verneaux used a semi-quantitative, species-specific, **abundance scale (0–5)** so that comparisons between species abundances make sense. (However, species-specific codes cannot be understood as unbiased estimates of the true abundances (number or density of individuals) or biomasses at the sites.) [你需要一位data domain專家]
- Working with the environmental data available in the R package **ade4** (version 1.4-14), we corrected a mistake in the **das variable** and restored the variables to their original units (Table 1.1.)

Table 1.1 Environmental variables of the Doubs data set used in this book and their units

Variable	Code	Units
Distance from source	das	km
Altitude	alt	m a.s.l.
Slope	pen	%
Mean minimum discharge	deb	$\text{m}^3 \text{s}^{-1}$
pH of water	pH	—
Calcium concentration (hardness)	dur	mg L^{-1}
Phosphate concentration	pho	mg L^{-1}
Nitrate concentration	nit	mg L^{-1}
Ammonium concentration	amm	mg L^{-1}
Dissolved oxygen	oxy	mg L^{-1}
Biological oxygen demand	dbo	mg L^{-1}

Data Extraction: Read Data

- 每一檔案之大小、資料維度、關聯。
- (報告中)列出每一變數之
 - 名稱、所代表意義。
 - 型態(連續、類別、順序、時間等等)、單位
 - 編碼、範圍(五數摘要)、遺失值比例(分佈)。
- 若是類別變數，則列出每一類別之次數分佈、交叉次數表。

```
> # Load the required package, vegan: Community Ecology Package
> library(vegan)

> # Load additionnal functions
> # (files must be in the working directory)
> source("panelutils.R")

> # Import the data from CSV files
> # Species (community) data frame (fish abundances)
> spe <- read.csv("DoubsSpe.csv", row.names=1)
> # Environmental data frame
> env <- read.csv("DoubsEnv.csv", row.names=1)
> # Spatial data frame
> spa <- read.csv("DoubsSpa.csv", row.names=1)
```

```
> library(ade4)
> data(doubs)
> ?doubs
```

Source: Borcard D., Gillet F. & Legendre P. Numerical Ecology with R, Springer, 2011



Species Data: First Contact

Basic functions

54/47

```
> spe    # Display the whole data frame in the console
      CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT
1      0   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
...
> spe[1:5,1:10]    # Display only 5 lines and 10 columns
      CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE
1      0   3   0   0   0   0   0   0   0
...
> head(spe)      # Display only the first few lines
      CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR
1      0   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
...
> nrow(spe)      # Number of rows (sites)
[1] 30
> ncol(spe)      # Number of columns (species)
[1] 27
> dim(spe)        # Dimensions of the data frame (rows, columns)
[1] 30 27
> colnames(spe)   # Column labels (descriptors = species)
[1] "CHA" "TRU" "VAI" "LOC" "OMB" "BLA" "HOT" "TOX" "VAN" "CHE" "BAR" "SPI"
...
> rownames(spe)   # Row labels (objects = sites)
[1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
...
> summary(spe)    # Descriptive statistics for columns
      CHA        TRU        VAI        LOC        OMB
Min. :0.00  Min. :0.00  Min. :0.000  Min. :0.000  Min. :0.00
1st Qu.:0.00 1st Qu.:0.00 1st Qu.:0.000 1st Qu.:1.000 1st Qu.:0.00
Median :0.00 Median :1.00  Median :3.000  Median :2.000  Median :0.00
Mean   :0.50 Mean   :1.90  Mean   :2.267  Mean   :2.433  Mean   :0.50
3rd Qu.:0.75 3rd Qu.:3.75 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:0.75
Max.   :3.00  Max.   :5.00  Max.   :5.000  Max.   :5.000  Max.   :4.00
...
```

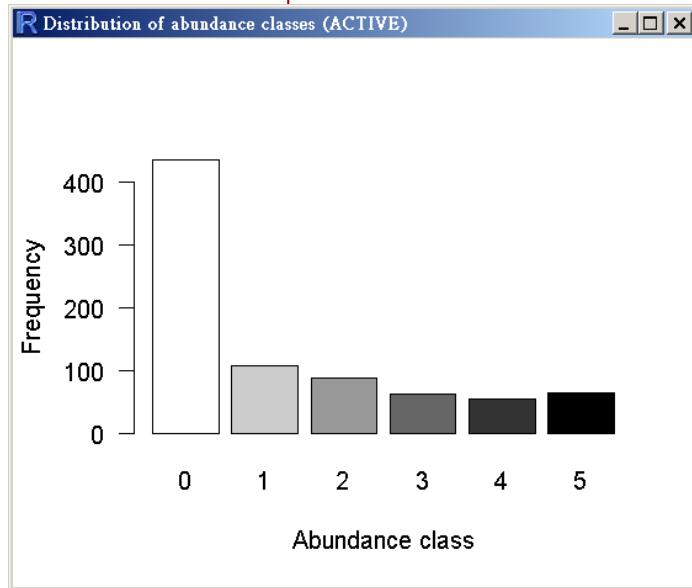


Overall Distribution of Abundances (Dominance Codes)

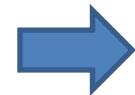
55/47

Compare median and mean abundances. Are most distributions symmetrical?

```
> # Minimum and maximum of abundance values in the whole data set
> range(spe)
[1] 0 5
> # Count cases for each abundance class
> (ab <- table(unlist(spe)))
0   1   2   3   4   5 
435 108  87  62  54  64
> # Create a graphic window with title
> windows(title="Distribution of abundance classes")
>
> # Barplot of the distribution, all species confounded
> barplot(ab, las=1, xlab="Abundance class",
+ ylab="Frequency", col=gray(5:0/5))
> # Number of absences
> sum(spe==0)
[1] 435
> # Proportion of zeros in the community data set
> sum(spe==0)/(nrow(spe)*ncol(spe))
[1] 0.537037
```



How do you interpret the high frequency of zeros (absences) in the data frame?



google "sparse data"



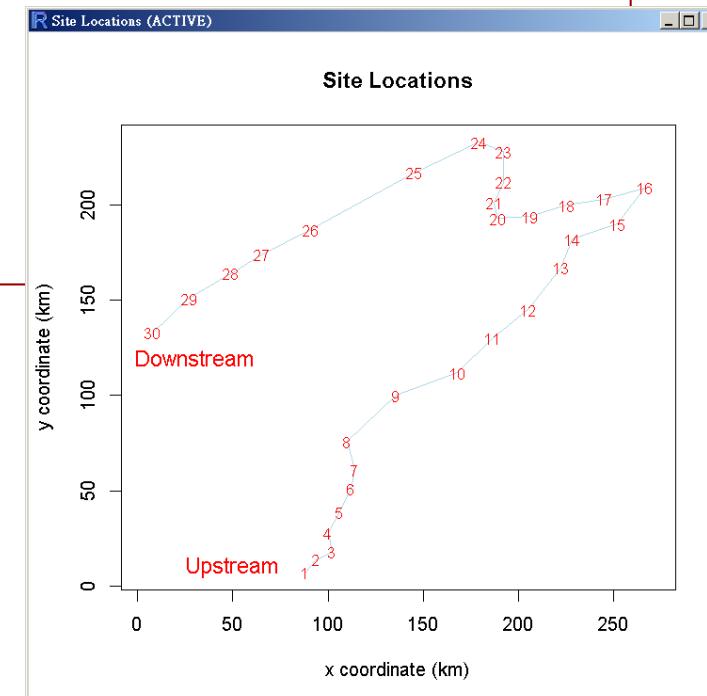
Species Data: A Closer Look

Map of the Locations of the Sites

56/47

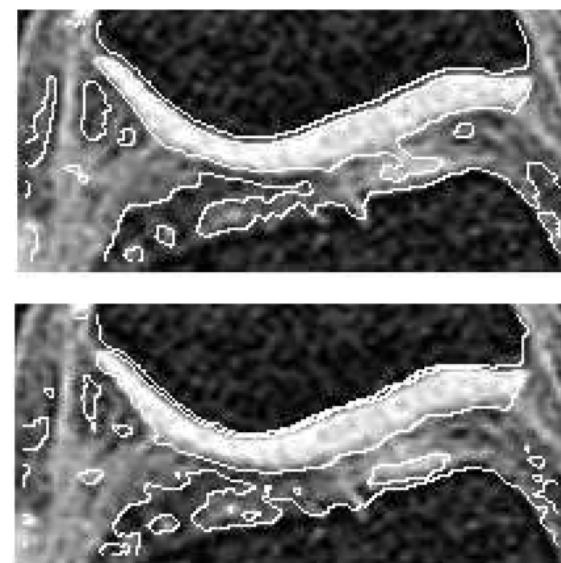
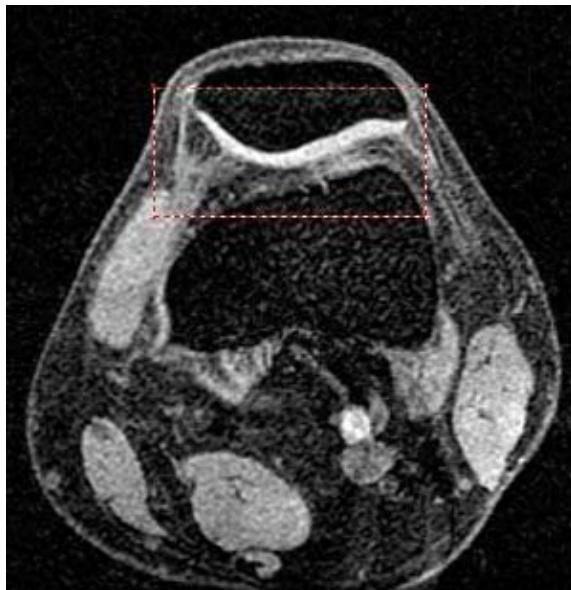
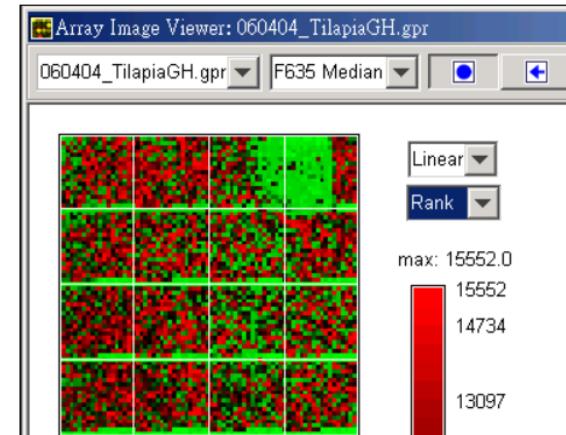
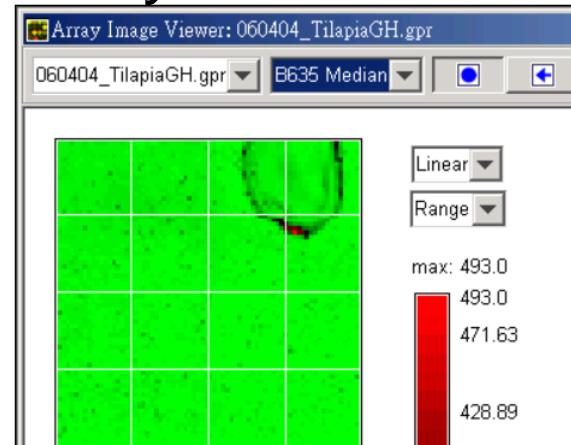
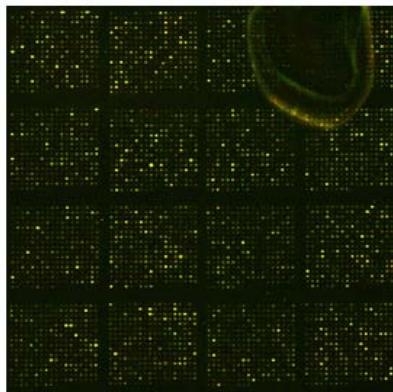
```
> windows(title="Site Locations")
> # Create an empty frame (proportional axes 1:1, with titles)
> # Geographic coordinates x and y from the spa data frame
> plot(spa, asp=1, type="n", main="Site Locations",
+ xlab="x coordinate (km)", ylab="y coordinate (km)")
> # Add a blue line connecting the sites (Doubs river)
> lines(spa, col="light blue")
> # Add site labels
> text(spa, row.names(spa), cex=0.8, col="red")
> # Add text blocks
> text(50, 10, "Upstream", cex=1.2, col="red")
> text(30, 120, "Downstream", cex=1.2, col="red")
```

The river looks more real, but where are the fish?

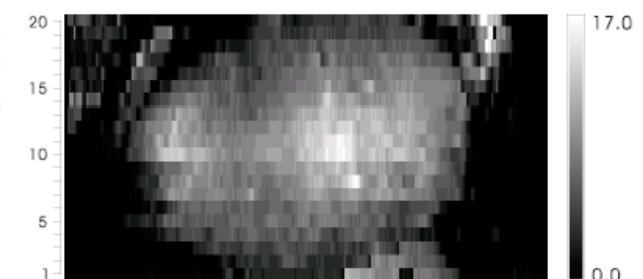


註: 重建 Reconstruction

生物晶片 (Microarray)

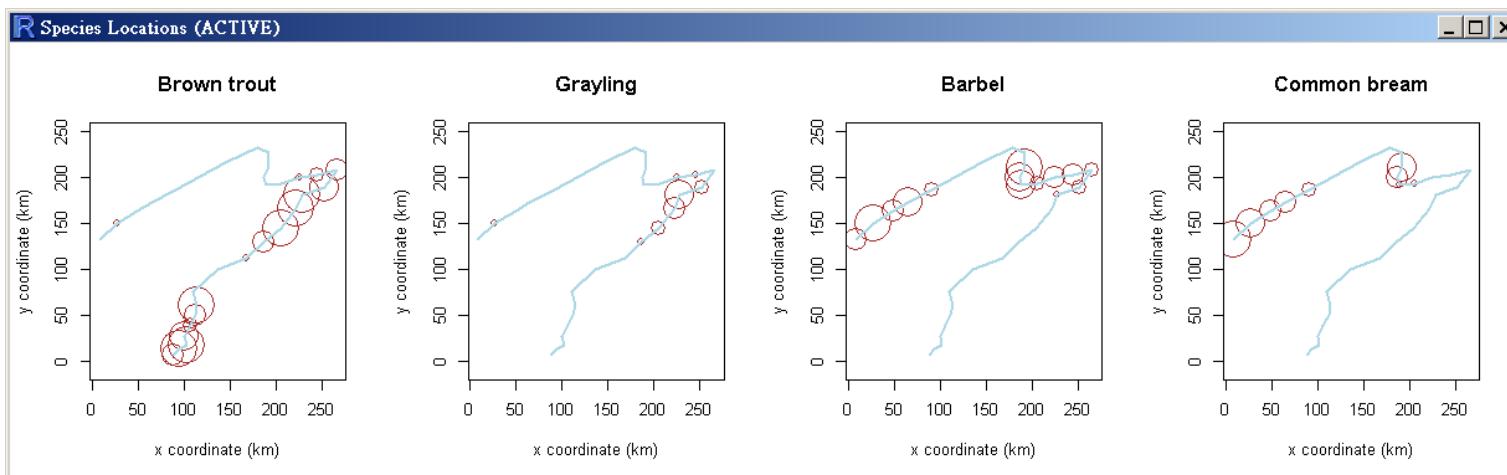


醫學影像 (fMRI)



Maps of Some Fish Species

```
> # New graphic window (size 9x9 inches)
> windows(title="Species Locations", 9, 9)
> par(mfrow=c(1,4))
> # Plot four species
> xl <- "x coordinate (km)",
> yl <- "y coordinate (km)"
> plot(spa, asp=1, col="brown", cex=spe$TRU, main="Brown trout", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$OMB, main="Grayling", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$BAR, main="Barbel", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, col="brown", cex=spe$BCO, main="Common bream", xlab=xl, ylab=yl)
> lines(spa, col="light blue", lwd=2)
```



Bubble maps of the abundance of four fish species

From these graphs you should understand why these four species were chose as ecological indicators.



Compare Species: Number of Occurrences

59/47

At how many sites does each species occur? Calculate the relative frequencies of species (proportion of the number of sites) and plot histograms.

```
> # Compute the number of sites where each species is present
> # To sum by columns, the second argument of apply(), MARGIN, is set to 2
> spe.pres <- apply(spe > 0, 2, sum)
> # Sort the results in increasing order
> sort(spe.pres)

PCH CHA OMB BLA BCO BBO TOX BOU ROT ANG HOT SPI CAR GRE PSO BAR ABL PER TRU TAN
  7   8   8   8   9  10  11  11  11  11  12  12  12  12  12  13  14  14  15  17  17

VAN BRO GAR VAI GOU LOC CHE
  18  18  18  20  20  24  25

> # Compute percentage frequencies
> spe.relf <- 100*spe.pres/nrow(spe)
> # Round the sorted output to 1 digit
> round(sort(spe.relf), 1)

PCH  CHA  OMB  BLA  BCO  BBO  TOX  BOU  ROT  ANG  HOT  SPI  CAR  GRE  PSO  BAR
23.3 26.7 26.7 26.7 30.0 33.3 36.7 36.7 36.7 36.7 40.0 40.0 40.0 40.0 43.3 46.7

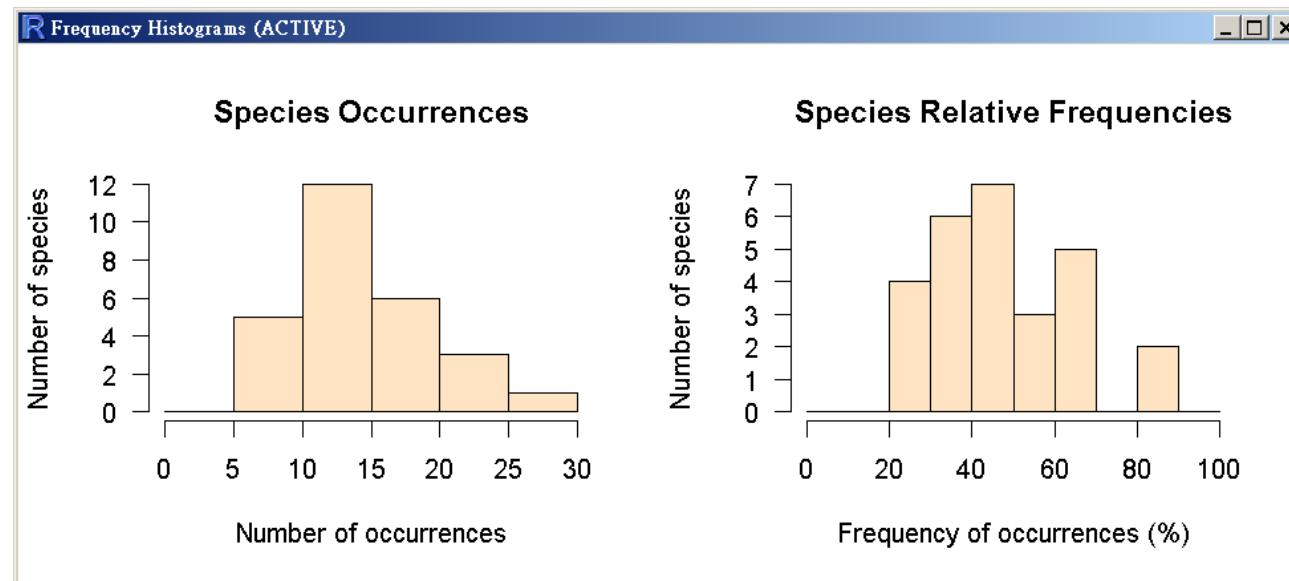
ABL  PER  TRU  TAN  VAN  BRO  GAR  VAI  GOU  LOC  CHE
46.7 50.0 56.7 56.7 60.0 60.0 60.0 66.7 66.7 80.0 83.3
```



Compare Species: Number of Occurrences

60/47

```
> # Plot the histograms
> windows(title="Frequency Histograms", 8, 5)
> # Divide the window horizontally
> par(mfrow=c(1,2))
> hist(spe.pres, main="Species Occurrences", right=FALSE, las=1,
+       xlab="Number of occurrences", ylab="Number of species",
+       breaks=seq(0,30,by=5), col="bisque")
> hist(spe.relf, main="Species Relative Frequencies", right=FALSE,
+       las=1, xlab="Frequency of occurrences (%)", ylab="Number of species",
+       breaks=seq(0, 100, by=10), col="bisque")
```



Compare Sites: Species Richness

Now that we have seen at how many sites each species is present, we may want to know how many species are present at each site (species richness).

```
> # Compute the number of species at each site
> # To sum by rows, the second argument of apply(), MARGIN, is set to 1
> sit.pres <- apply(spe > 0, 1, sum)
> # Sort the results in increasing order
> sort(sit.pres)
 8   1   2   23   3   7   9   10   11   12   13   4   24   25   6   14   5   15   16   26   30   17   20   22   27   28   18   19
 0   1   3   3   4   5   5   6   6   6   6   8   8   8   10   10   11   11   11   17   21   21   22   22   22   22   23   23
21   29
23   26
```

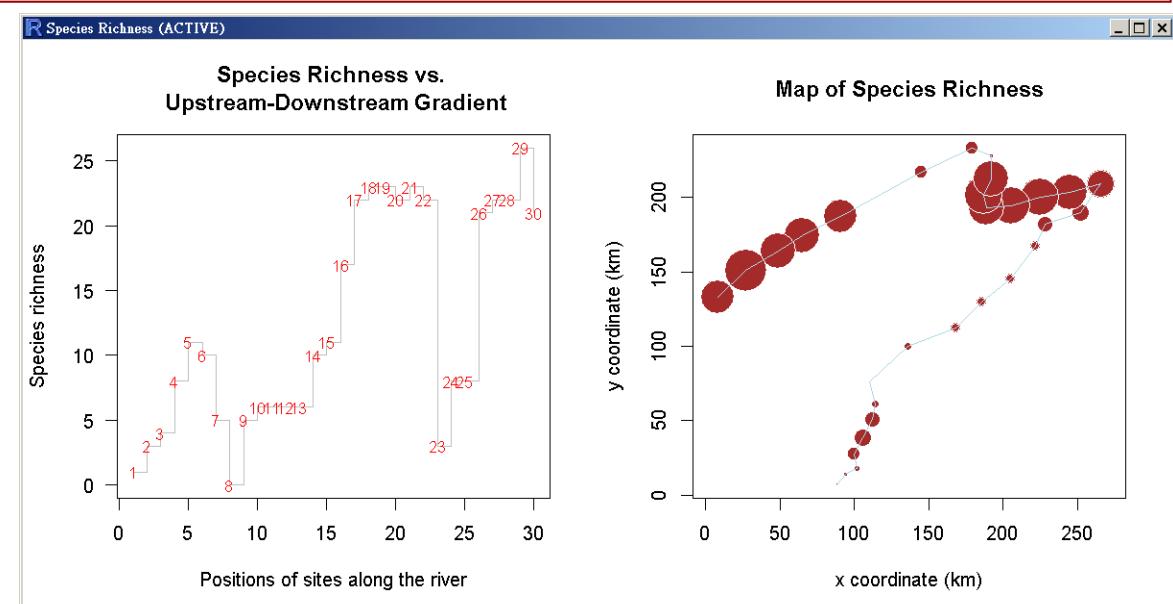
Compare Sites: Species Richness

```

> windows(title="Species Richness", 10, 5)
> par(mfrow=c(1,2))
> # Plot species richness vs. position of the sites along the river
> plot(sit.pres,type="s", las=1, col="gray",
+ main="Species Richness vs. \n Upstream-Downstream Gradient",
+ xlab="Positions of sites along the river", ylab="Species richness")
> text(sit.pres, row.names(spe), cex=.8, col="red")
> # Use geographic coordinates to plot a bubble map
> plot(spa, asp=1, main="Map of Species Richness", pch=21, col="white",
+ bg="brown", cex=5*sit.pres/max(sit.pres), xlab="x coordinate (km)",
+ ylab="y coordinate (km)")
> lines(spa, col="light blue")

```

Can you identify richness hot spots along the river?





Compute Alpha Diversity Indices of the Fish Communities

63/47

Finally, one can easily compute classical diversity indices from the data. Let us do it with the function **diversity()** of the **vegan** package.

生態多樣性指標

diversity {vegan} R Documentation
Ecological Diversity Indices and Rarefaction Species Richness
Description
Shannon, Simpson, and Fisher diversity indices and rarefied species richness for community ecologists.
Usage
diversity(x, index = "shannon", MARGIN = 1, base = exp(1))

```
> # Get help on the diversity() function
> ?diversity
>
> N0 <- rowSums(spe > 0)          # Species richness
> H <- diversity(spe)            # Shannon entropy
> N1 <- exp(H)                  # Shannon diversity (number of abundant species)
> N2 <- diversity(spe, "inv")    # Simpson diversity (number of dominant species)
> J <- H/log(N0)                # Pielou evenness
> E10 <- N1/N0                  # Shannon evenness (Hill's ratio)
> E20 <- N2/N0                  # Simpson evenness (Hill's ratio)
> (div <- data.frame(N0, H, N1, N2, E10, E20, J))
   N0        H        N1        N2        E10        E20        J
1  1 0.000000  1.000000  1.000000 1.0000000 1.0000000      NaN
2  3 1.077556  2.937493  2.880000 0.9791642 0.9600000 0.9808340
3  4 1.263741  3.538634  3.368421 0.8846584 0.8421053 0.9115962
4  8 1.882039  6.566883  5.727273 0.8208604 0.7159091 0.9050696
5 11 2.329070 10.268387  9.633333 0.9334897 0.8757576 0.9712976
6 10 2.108294  8.234184  7.000000 0.8234184 0.7000000 0.9156205
...
...
```



Transformation and Standardization of the Species Data

64/47

- The `decostand()` function of the `vegan` package provides many options for **common standardization of ecological data**.
- In this function, standardization, as contrasted with simple transformation (such as square root, log or presence-absence), means that the values are not transformed individually but relative to other values in the data table.
- Standardization can be done relative to sites (site profiles), species (species profiles), or both (double profiles), depending on the focus of the analysis.

```
> # Get help on the decostand() function
> ?decostand
> ## Simple transformations
> # Partial view of the raw data (abundance codes)
> spe[1:5, 2:4]
  TRU VAI LOC
1   3   0   0
...
> # Transform abundances to presence-absence (1-0)
> spe.pa <- decostand(spe, method="pa")
> spe.pa[1:5, 2:4]
  TRU VAI LOC
1   1   0   0
...
```

decostand {vegan} R Documentation

Standardization Methods for Community Ecology

Description

The function provides some popular (and effective) standardization methods for community ecologists.

Usage

`decostand(x, method, MARGIN, range.global, logbase = 2, na.rm=FALSE, ...)`
`wisconsin(x)`



Transformation and Standardization of the Species Data

65/47

```
> Species profiles: 2 methods: presence-absence or abundance data
> ## Species profiles: standardization by column
> # Scale abundances by dividing them by the maximum value for each species
> # Note: MARGIN=2 (column, default value) for this method
> spe.scal <- decostand(spe, "max")
> spe.scal[1:5,2:4]
  TRU VAI LOC
1 0.6 0.0 0.0
...
> # Display the maximum by column
> apply(spe.scal, 2, max)
CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR TAN
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
BCO PCH GRE GAR BBO ABL ANG
  1   1   1   1   1   1   1
> # Scale abundances by dividing them by the species totals
> # (relative abundance by species)
> # Note: MARGIN=2 for this method
> spe.relsp <- decostand(spe, "total", MARGIN=2)
> spe.relsp[1:5,2:4]
  TRU      VAI      LOC
1 0.05263158 0.00000000 0.00000000
...
> # Display the sum by column
> apply(spe.relsp, 2, sum)
CHA TRU VAI LOC OMB BLA HOT TOX VAN CHE BAR SPI GOU BRO PER BOU PSO ROT CAR TAN BCO
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
PCH GRE GAR BBO ABL ANG
  1   1   1   1   1   1
```

Did the scaling work properly? Keep an eye on the results by a plot or by the use of summary statistics



Scale Abundances by Dividing Them by the Site Totals

66/47

```
> ## Site profiles: 3 methods; presence-absence or abundance data
> ## standardization by row
> # Scale abundances by dividing them by the site totals
> # (relative abundance, or relative frequencies, per site)
> # (relative abundance by site)
> # Note: MARGIN=1 (default value) for this method
> spe.rel <- decostand(spe, "total")
> spe.rel[1:5,2:4]
      TRU        VAI        LOC
1 1.00000000 0.00000000 0.00000000
...
> # Display the sum of row vectors to determine if the scaling worked properly
> apply(spe.rel, 1, sum)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
29 30
 1  1
> # Give a length of 1 to each row vector (Euclidean norm)
> spe.norm <- decostand(spe, "normalize")
> spe.norm[1:5,2:4]
      TRU        VAI        LOC
1 1.0000000 0.0000000 0.0000000
...
> # Verify the norm of row vectors
> norm <- function(x) sqrt(x%*%x)
> apply(spe.norm, 1, norm)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1  1  1  1  1  1  1  0  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
29 30
 1  1
```

The chord transformation: the Euclidean distance function applied to chord-transformed data produces a chord distance matrix. Useful before PCA and K-means.



Compute Relative Frequencies by Rows (Site Profiles)

67/47

- The Hellinger transformation can be also be obtained by applying the chord transformation to square-root-transformed species data.

```
> # Compute relative frequencies by rows (site profiles), then square root
> # Compute square root of relative abundances by site
> spe.hel <- decostand(spe, "hellinger")
> spe.hel[1:5,2:4]
      TRU      VAI      LOC
1 1.0000000 0.0000000 0.0000000
2 0.6454972 0.5773503 0.5000000
3 0.5590170 0.5590170 0.5590170
4 0.4364358 0.4879500 0.4879500
5 0.2425356 0.2970443 0.2425356
> # Check the norm of row vectors
> apply(spe.hel, 1, norm)
   1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28
   1   1   1   1   1   1   1   0   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
  29  30
   1   1
```

<http://artax.karlin.mff.cuni.cz/r-help/library/analogue/html/tran.html>



Standardization by Both Columns and Rows

68/47

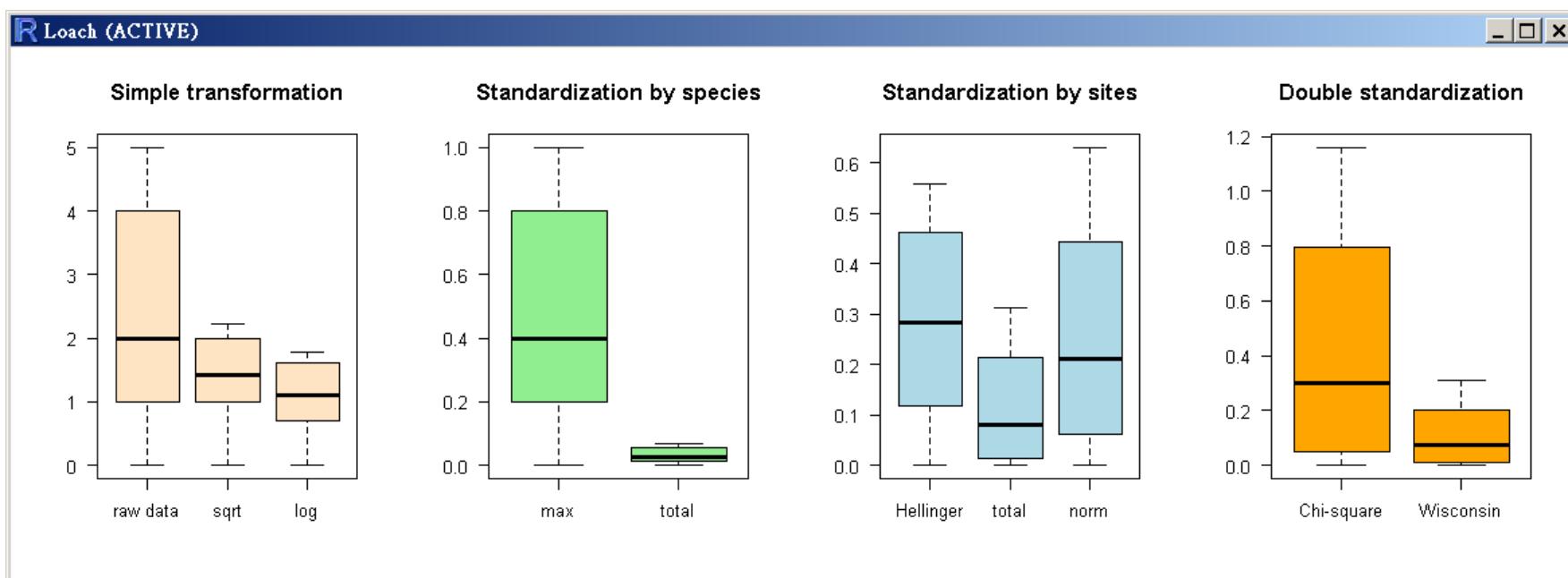
```
> # Chi-square transformation
> spe.chi <- decostand(spe, "chi.square")
> spe.chi[1:5,2:4]
    TRU      VAI      LOC
1 4.1969078 0.0000000 0.0000000
2 1.7487116 1.2808290 0.9271402
3 1.3115337 1.2007772 1.1589253
4 0.7994110 0.9148778 0.8829907
5 0.2468769 0.3390430 0.2181506
> # Check what happened to site 8 where no species was found
> spe.chi[7:9,]
    CHA      TRU      VAI      LOC OMB BLA HOT TOX      VAN      CHE BAR SPI GOU BRO
7 0 1.311534 0.9606217 1.1589253 0 0 0 0 0.302004 0.2646384 0 0 0 0 0
8 0 0.000000 0.0000000 0.0000000 0 0 0 0 0.000000 0.0000000 0 0 0 0 0
9 0 0.000000 0.2744634 0.7946916 0 0 0 0 0.000000 1.5122194 0 0 0 0 0
    PER BOU PSO ROT CAR      TAN BCO PCH GRE      GAR BBO ABL ANG
7 0 0 0 0 0 0.0000000 0 0 0 0.000000 0 0 0
8 0 0 0 0 0 0.0000000 0 0 0 0.000000 0 0 0
9 0 0 0 0 0 0.3373903 0 0 0 1.140587 0 0 0
> # Wisconsin standardization
> # Abundances are first ranged by species maxima and then by site totals
> spe.wis <- wisconsin(spe)
> spe.wis[1:5,2:4]
    TRU      VAI      LOC
1 1.0000000 0.0000000 0.0000000
2 0.41666667 0.33333333 0.25000000
3 0.31250000 0.31250000 0.31250000
4 0.19047619 0.23809524 0.23809524
5 0.05882353 0.08823529 0.05882353
```



Boxplots of Transformed Abundances of a Common Species (Stone Loach)

69/47

```
> windows(title="Loach") # 泥鰍
> par(mfrow=c(1,4))
> boxplot(spe$LOC, sqrt(spe$LOC), log1p(spe$LOC), las=1, main="Simple transformation",
+ names=c("raw data", "sqrt", "log"), col="bisque")
> boxplot(spe.scal$LOC, spe.relsp$LOC, las=1, main="Standardization by species",
+ names=c("max", "total"), col="lightgreen")
> boxplot(spe.hel$LOC, spe.rel$LOC, spe.norm$LOC, las=1, main="Standardization by sites",
+ names=c("Hellinger", "total", "norm"), col="lightblue")
> boxplot(spe.chi$LOC, spe.wiss$LOC, las=1, main="Double standardization",
+ names=c("Chi-square", "Wisconsin"), col="orange")
```



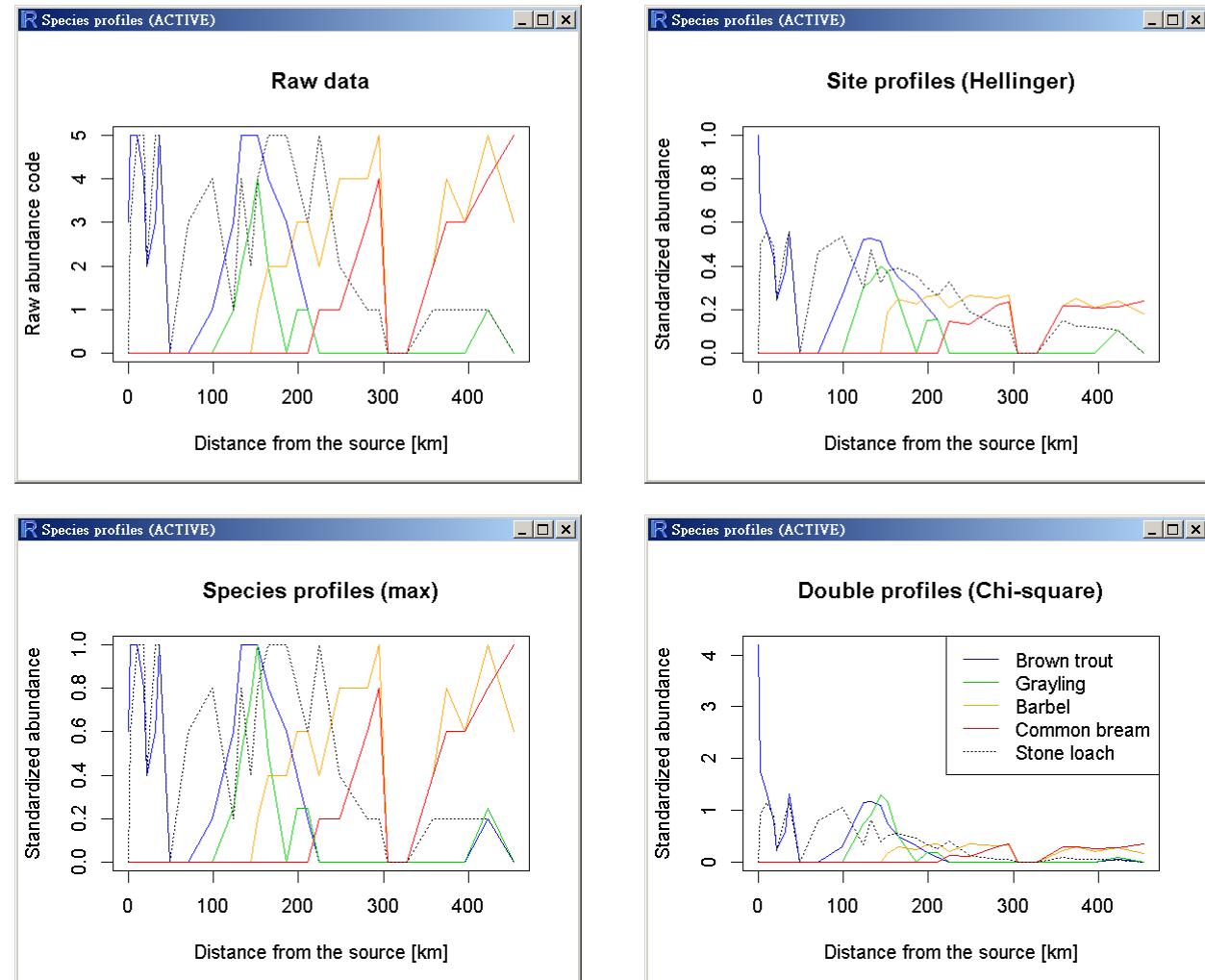
Boxplots of transformed abundances of a common species, *Nemacheilus barbatulus* (**stone loach**)



Plot Profiles Along the Upstream-Downstream Gradient

70/47

Another way to compare the effects of transformations on species profiles is to plot them along the river course.



Compare the profiles and explain the differences.



Plot Profiles Along the Upstream-Downstream Gradient

71/47

```
> windows(title="Species profiles", 9, 9)
> plot(env$das, spe$TRU, type="l", col=4, main="Raw data",
+ xlab="Distance from the source [km]", ylab="Raw abundance code")
> lines(env$das, spe$OMB, col=3); lines(env$das, spe$BAR, col="orange")
> lines(env$das, spe$BCO, col=2); lines(env$das, spe$LOC, col=1, lty="dotted")
>
> plot(env$das, spe.scal$TRU, type="l", col=4, main="Species profiles (max)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.scal$OMB, col=3); lines(env$das, spe.scal$BAR, col="orange")
> lines(env$das, spe.scal$BCO, col=2); lines(env$das, spe.scal$LOC, col=1, lty="dotted")

> plot(env$das, spe.hel$TRU, type="l", col=4, main="Site profiles (Hellinger)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.hel$OMB, col=3); lines(env$das, spe.hel$BAR, col="orange")
> lines(env$das, spe.hel$BCO, col=2); lines(env$das, spe.hel$LOC, col=1, lty="dotted")
>
> plot(env$das, spe.chi$TRU, type="l", col=4, main="Double profiles (Chi-square)",
+ xlab="Distance from the source [km]", ylab="Standardized abundance")
> lines(env$das, spe.chi$OMB, col=3); lines(env$das, spe.chi$BAR, col="orange")
> lines(env$das, spe.chi$BCO, col=2); lines(env$das, spe.chi$LOC, col=1, lty="dotted")
> legend("topright", c("Brown trout", "Grayling", "Barbel", "Common bream", "Stone loach"),
+ col=c(4,3,"orange",2,1), lty=c(rep(1,4),3))
```

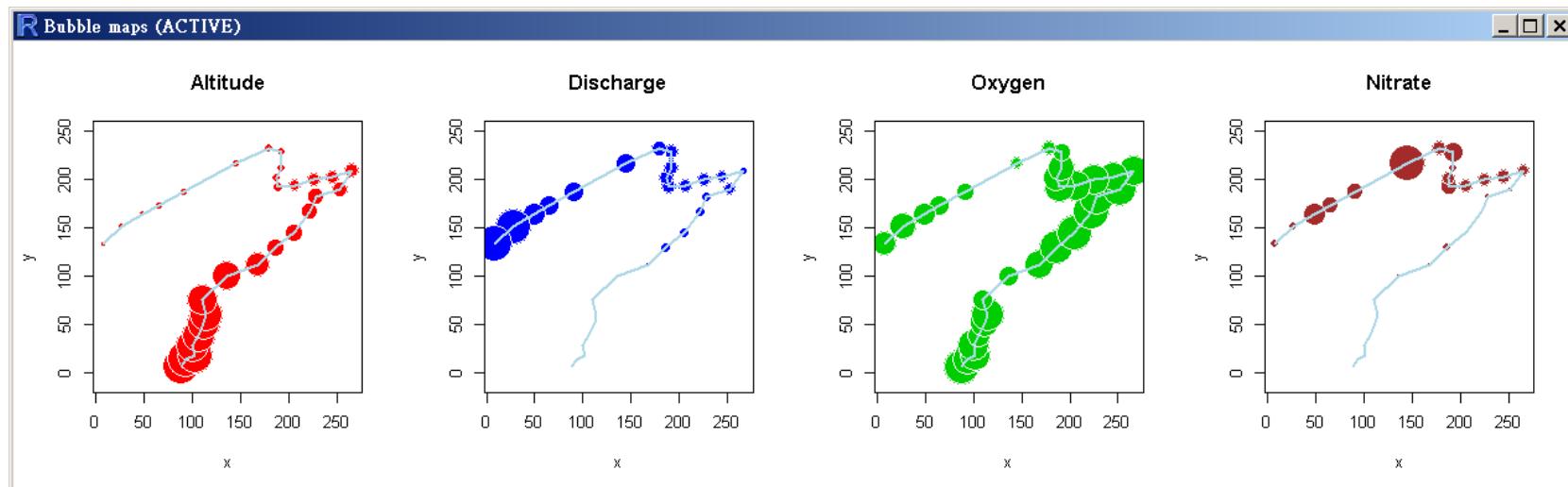


Bubble Maps of Some Environmental Variables

72/47

```
> windows(title="Bubble maps", 9, 9)
> par(mfrow=c(1,4))
> plot(spa, asp=1, main="Altitude", pch=21, col="white",
+ bg="red", cex=5*env$alt/max(env$alt), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Discharge", pch=21, col="white",
+ bg="blue", cex=5*env$deb/max(env$deb), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Oxygen", pch=21, col="white",
+ bg="green3", cex=5*env$oxy/max(env$oxy), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
> plot(spa, asp=1, main="Nitrate", pch=21, col="white",
+ bg="brown", cex=5*env$nit/max(env$nit), xlab="x", ylab="y")
> lines(spa, col="light blue", lwd=2)
```

Apply the basic functions to **env**. While examining the **summary()**, note how the variables differ from the species data in values and spatial distributions. Draw maps of some of the environmental variables.

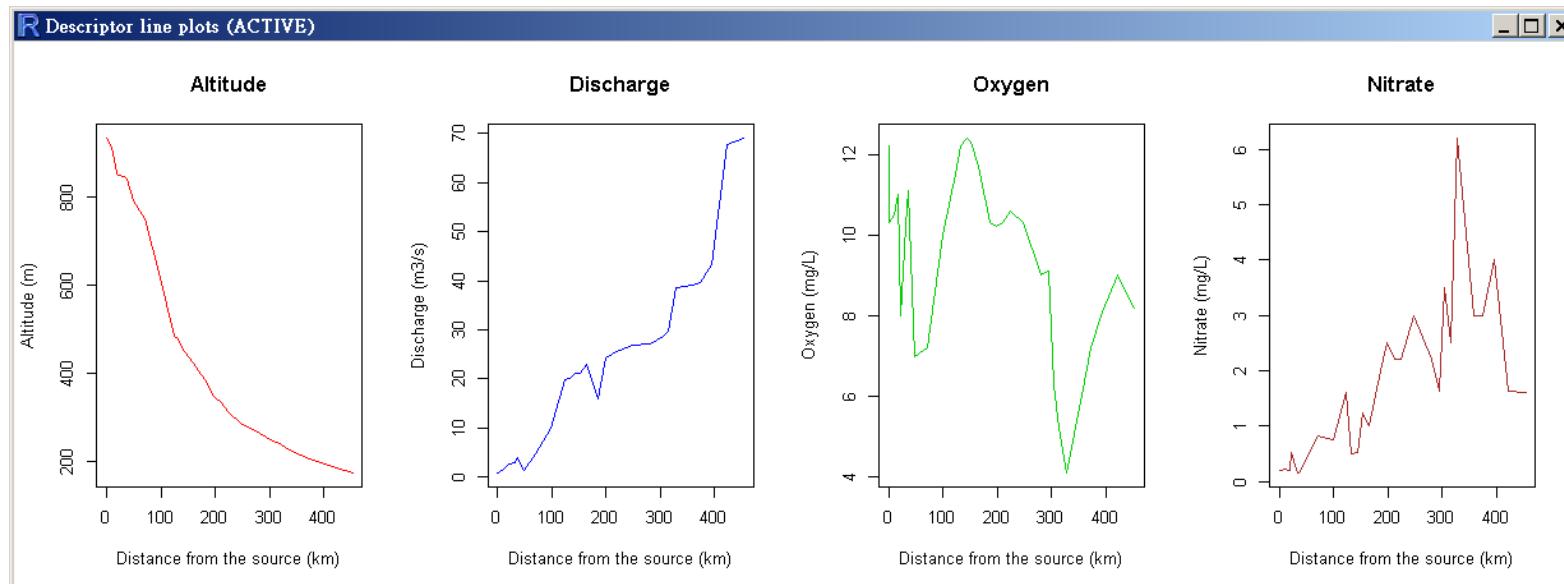


Which ones of these maps display an upstream-downstream gradient? How could you explain the spatial patterns of the other variables?



Examine the Variation of Some Descriptors Along the Stream: Line Plots

```
> windows(title="Descriptor line plots")
> par(mfrow=c(1,4))
> plot(env$das, env$alt, type="l", xlab="Distance from the source (km)",
+ ylab="Altitude (m)", col="red", main="Altitude")
> plot(env$das, env$deb, type="l", xlab="Distance from the source (km)",
+ ylab="Discharge (m³/s)", col="blue", main="Discharge")
> plot(env$das, env$oxy, type="l", xlab="Distance from the source (km)",
+ ylab="Oxygen (mg/L)", col="green3", main="Oxygen")
> plot(env$das, env$nit, type="l", xlab="Distance from the source (km)",
+ ylab="Nitrate (mg/L)", col="brown", main="Nitrate")
```



Note the scalings.

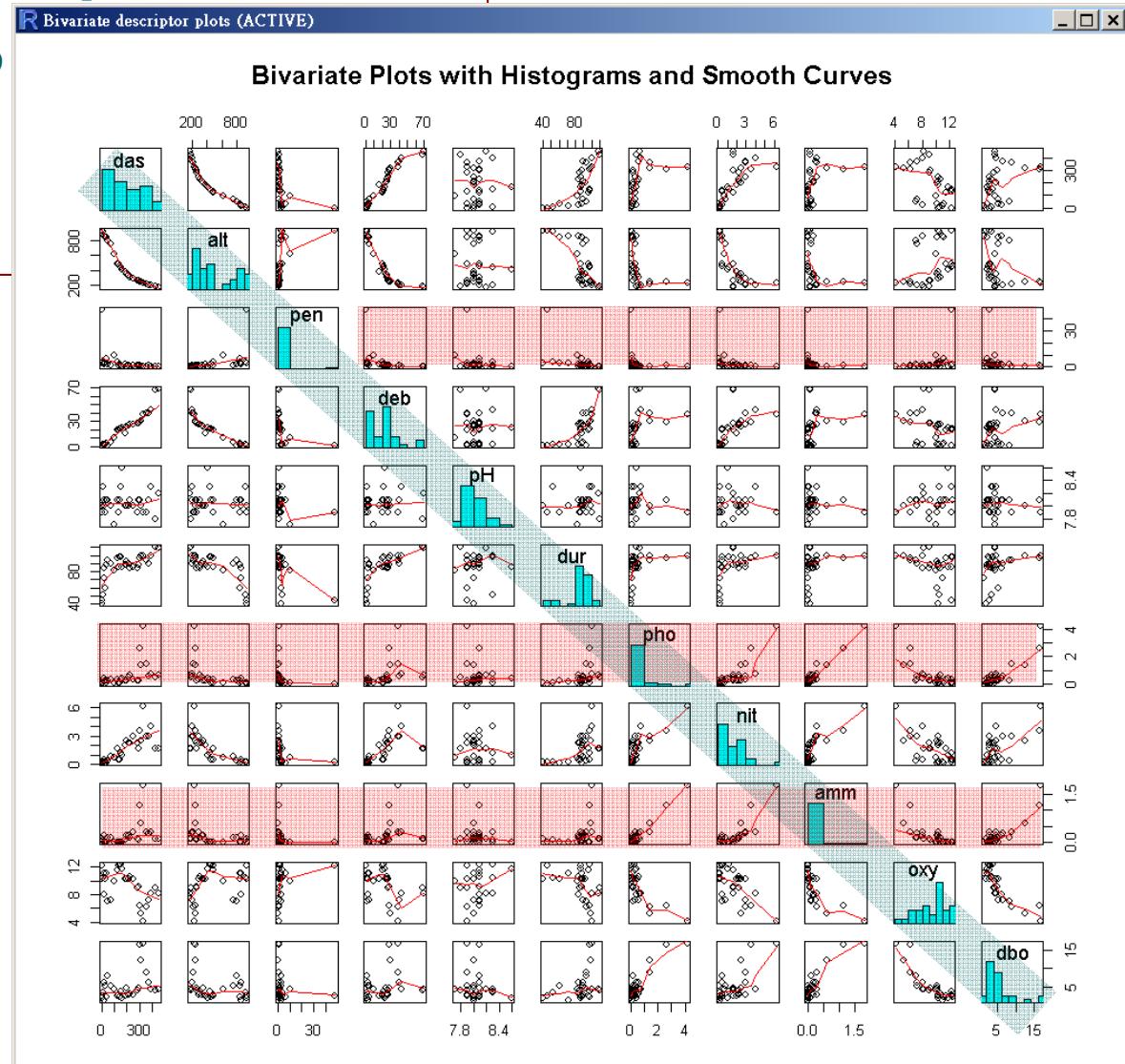


Scatter Plots for All Pairs of Environmental Variables

74/47

```
> windows(title="Bivariate descriptor plots")
> source("panelutils.R")
> op <- par(mfrow=c(1,1), pty="s")
> pairs(env, panel=panel.smooth,
diag.panel=panel.hist,
main="Bivariate Plots with
Histograms and Smooth Curves")
> par(op)
```

- Do many variables seem normally distributed?
- Do many scatter plots show linear or at least monotonic relationships?



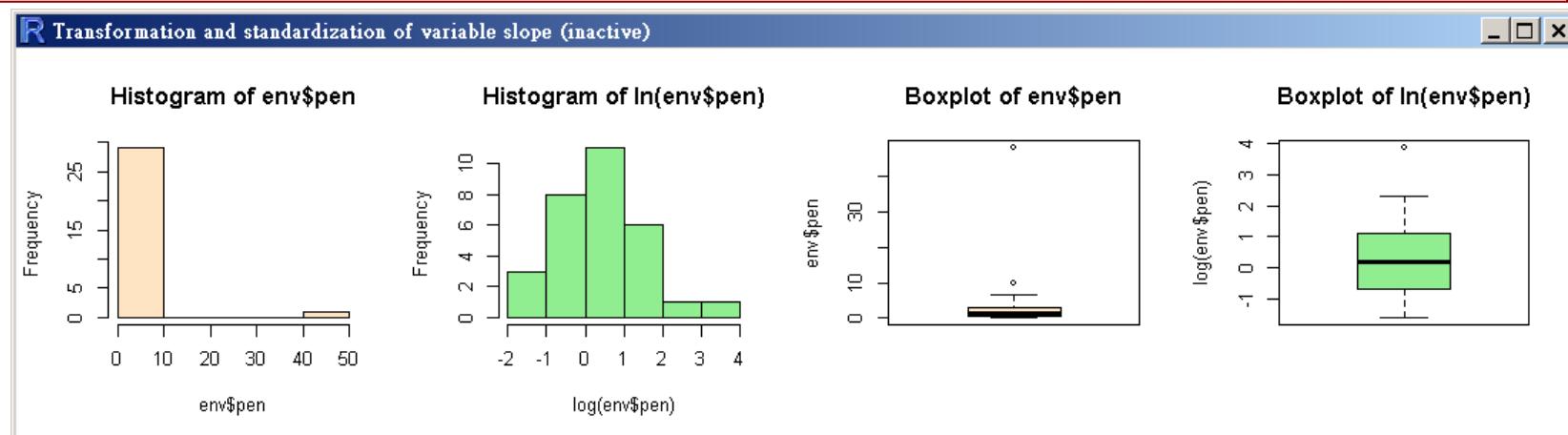


Simple Transformation of An Environmental Variable

75/47

- Simple transformations, such as the log transformation, can be used to improve the distributions of some variables (make it closer to the normal distribution).
- Because environmental variables are dimensionally heterogeneous (expressed in different units and scales), many statistical analyses require their standardization to zero mean and unit variance. These centred and scaled variables are called **z-scores**.

```
> range(env$pen) #河道坡度
[1] 0.2 48.0
> # Log-transformation of the slope variable ( $y = \ln(x)$ )
> # Compare histograms and boxplots of raw and transformed values
> windows(title="Transformation and standardization of variable slope")
> par(mfrow=c(1,4))
> hist(env$pen, col="bisque", right=FALSE)
> hist(log(env$pen), col="light green", right=F, main="Histogram of ln(env$pen)")
> boxplot(env$pen, col="bisque", main="Boxplot of env$pen", ylab="env$pen")
> boxplot(log(env$pen), col="light green", main="Boxplot of ln(env$pen)",
+ ylab="log(env$pen)")
```





Standardization of

76/47

All Environmental Variables

```
> # Center and scale = standardize variables (z-scores)
> env.z <- decostand(env, "standardize")
> apply(env.z, 2, mean)    # means = 0
  das      alt      pen      deb      pH      dur
1.000429e-16 1.814232e-18 -1.659010e-17 1.233099e-17 -4.096709e-15 3.348595e-16
  pho      nit      amm      oxy      dbo
1.327063e-17 -8.925898e-17 -4.289646e-17 -2.886092e-16 7.656545e-17
> apply(env.z, 2, sd)    # standard deviations = 1
das alt pen deb pH dur pho nit amm oxy dbo
  1   1   1   1   1   1   1   1   1   1   1
>
> # Same standardization using the scale() function (which returns a matrix)
> env.z <- as.data.frame(scale(env))
> env.z
  das      alt      pen      deb      pH      dur
1 -1.34949526 1.667360909 5.14106053 -1.18004457 -0.8635475 -2.436958124
2 -1.33585215 1.659991358 -0.05737533 -1.17120570 -0.2878492 -2.733425049
...
```



小結 & 想想看

77/47

- The EDA tools allow researchers to obtain a **general impression** of their data.
- Information about simple parameters and distributions of variables is important to consider in order to choose more **advanced analyses correctly**.
- **Graphical representations** may help generate hypotheses about the processes acting behind the scene. (**try heatmap!**)
- **想想看:** Doubs Fish Data經過這一連串的資料探索，還有哪一些有趣的問題可以提出？(季節？人口、工廠分佈？這些資料可以得到嗎？)

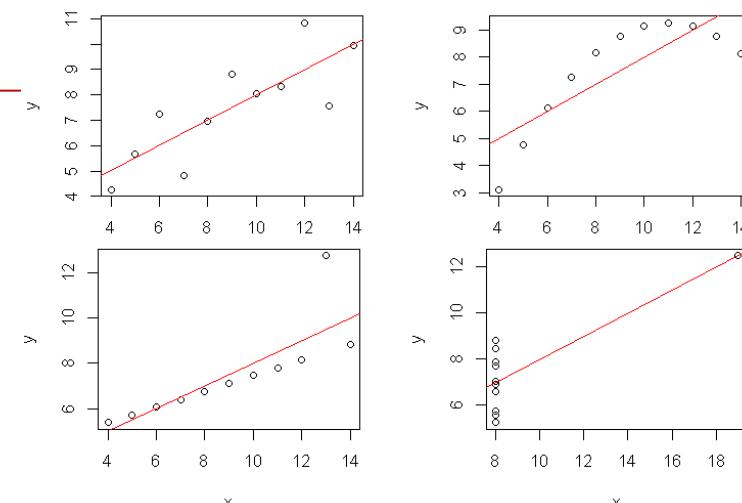
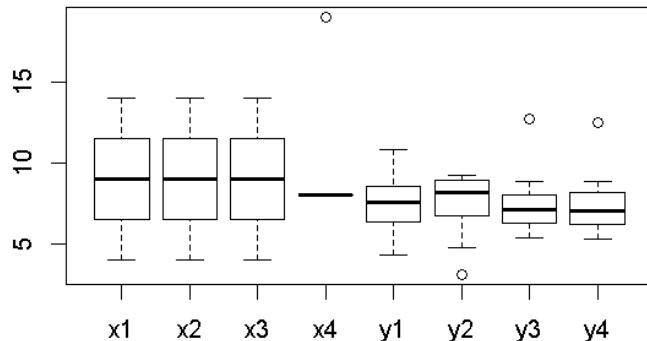


Anscombe's Quartet of 'Identical' Simple Linear Regressions

```
> 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])
```

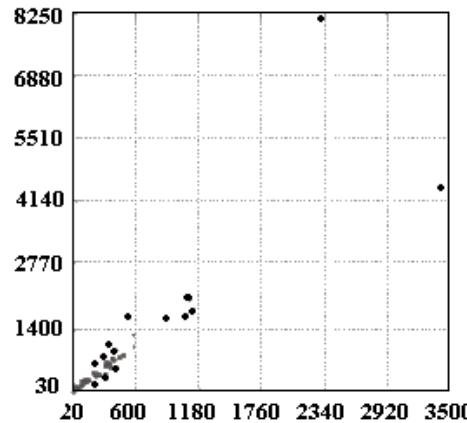
```
boxplot(anscombe)
```



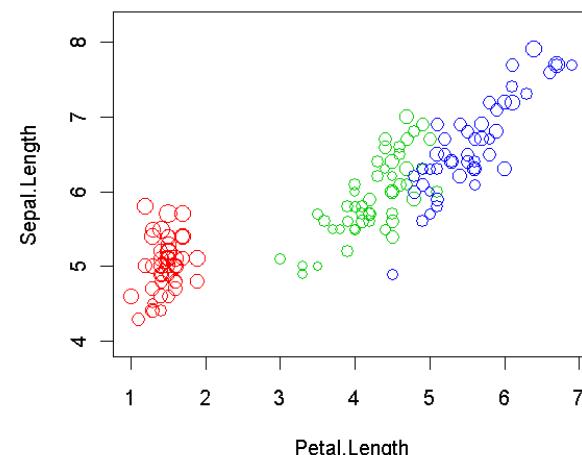
Extensions of Scatterplots

With a smoothing curve

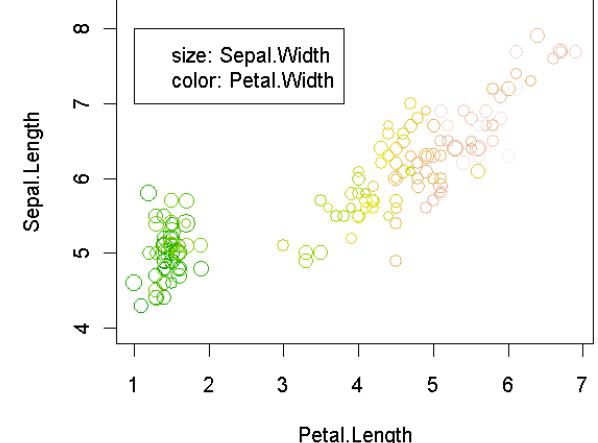
Poorly-Constructed



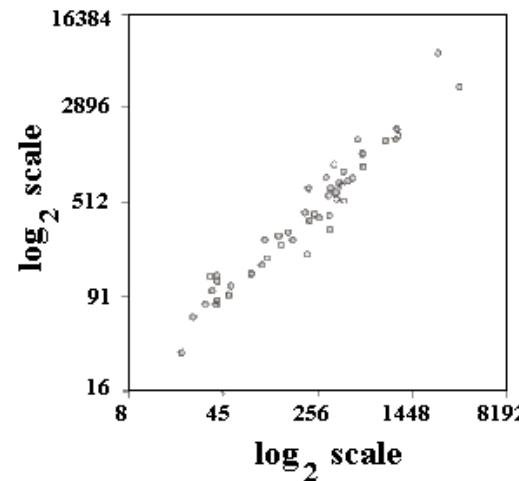
Bubbleplot: Sepal.Width



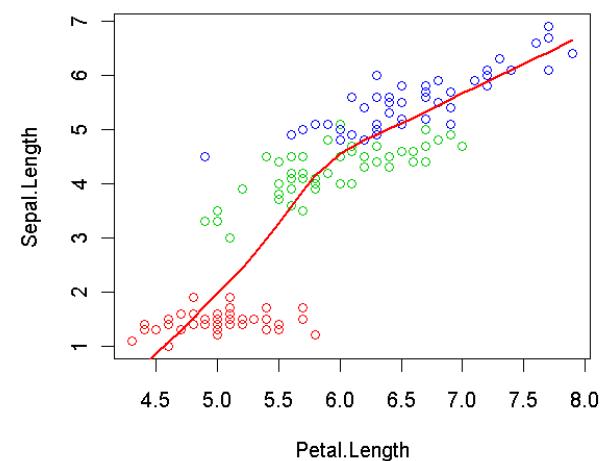
Color Bubbleplot



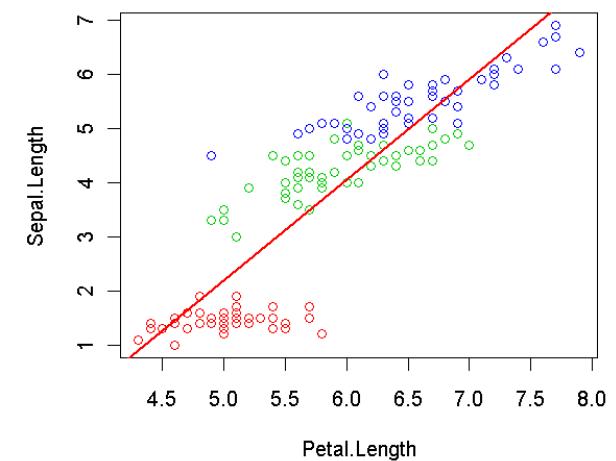
Better



LOWESS



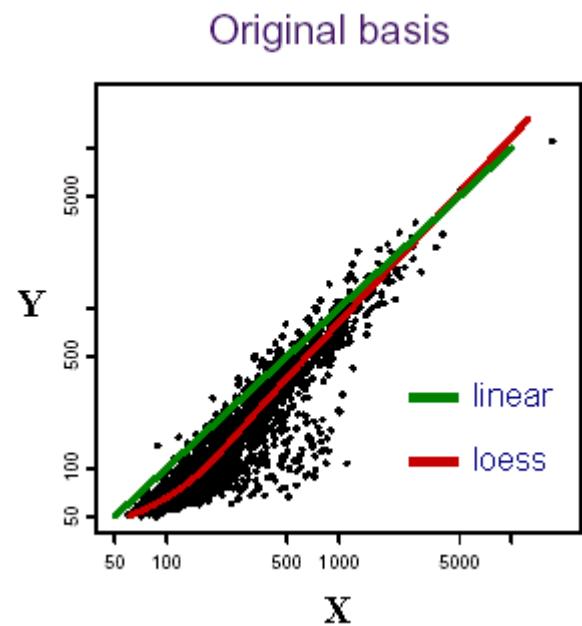
Simple Linear Regression





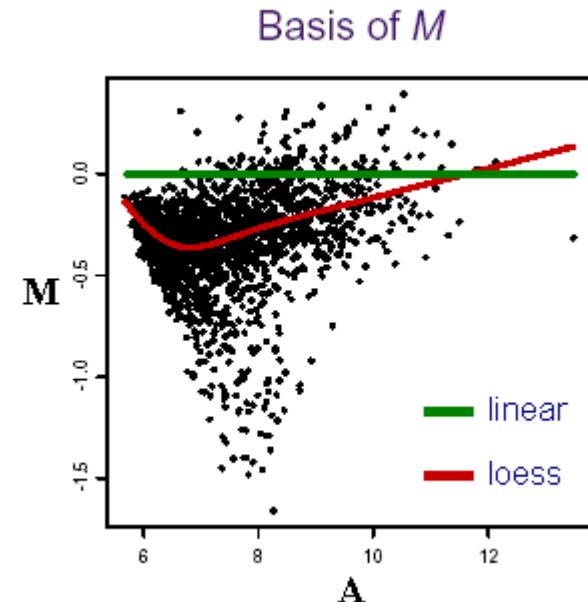
Scatterplot for Gene Expression Data

- **MA plots** can show the intensity-dependent ratio of raw microarray data.



Oligo	cDNA
X = PM ₁ , Y = PM ₂	X= Cy3 Y= Cy5
X = PM ₁ -MM ₁ , Y = PM ₂ -MM ₂	

$$M = \log_2 \left(\frac{Y}{X} \right)$$



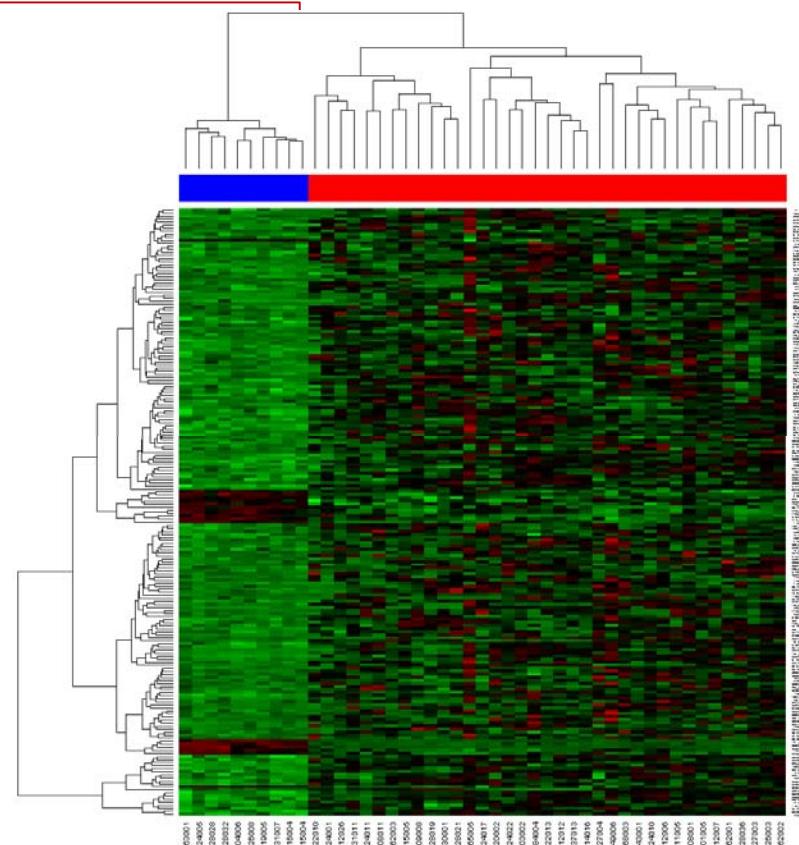
heatmap {stats}

```

> source("https://bioconductor.org/biocLite.R")
> biocLite("ALL")
> library(ALL)
> data(ALL)
> ALL
> str(ALL)
> dim(exprs(ALL))
[1] 12625   128
> exprs(ALL)[1:3, 1:5]
      01005    01010    03002    04006    04007
1000_at  7.597323 7.479445 7.567593 7.384684 7.905312
1001_at  5.046194 4.932537 4.799294 4.922627 4.844565
1002_f_at 3.900466 4.208155 3.886169 4.206798 3.416923
> table(ALL$mol.biol)

ALL1/AF4  BCR/ABL E2A/PBX1      NEG    NUP-98  p15/p16
      10       37       5       74       1       1
> eset <- ALL[, ALL$mol.biol %in%
+               c("BCR/ABL", "ALL1/AF4")]
> dim(exprs(eset))
[1] 12625   47
> f <- factor(as.character(eset$mol.biol))
> eset.p <- apply(exprs(eset), 1, function(x) t.test(x ~ f)$p.value)
> selected.eset <- eset[eset.p < 0.00001, ]
> dim(selected.eset)
Features Samples
      200      47
> ma.col <- colorRampPalette(c("green", "black", "red"))(200)
> var.col <- ifelse(f=="ALL1/AF4", "blue", "red")
> heatmap(exprs(selected.eset), col=ma.col, ColSideColors=var.col,
+          scale="row")

```





Visualizing Categorical Data

Visualizing Categorical Data - Windows Internet Explorer
http://www.math.yorku.ca/SCS/vcd/

檔案(E) 編輯(E) 檢視(V) 我的最愛(A) 工具(I) 說明(H)

Visualizing Categorical Data

Visualizing Categorical Data

by Michael Friendly

SAS Institute (Dec, 2000), Order code 56571, ISBN 1-58025-660-0

Visualizing Categorical Data

Contents

- 1. Introduction [View Ch. 1 in .pdf (264k)]
- 2. Fitting and graphing discrete distributions
- 3. Two-way contingency tables
- 4. Mosaic displays for n-way tables
- 5. Correspondence analysis
- 6. Logistic regression
- 7. Loglinear and logit models
- A. SAS programs and macros
- B. Data sets
- C. Tables

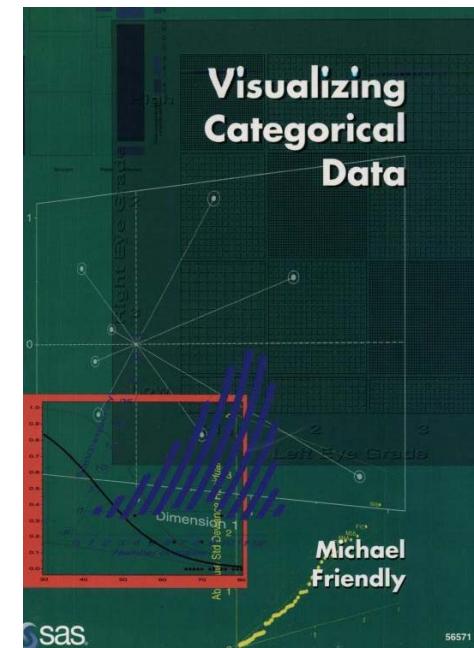
Errata and Updates

See also:

- Reviews of [Visualizing Categorical Data](#)
- [Reader's Guide to Visualizing Categorical Data](#) (SUGI 26 paper, PDF [137k])
- VCD Sampler: [Visualizing Categorical Data: Data, Stories, and Pictures](#) (SUGI 25 paper, PDF [177k])
- [VCD Source Online](#) [550k] (Source code for datasets and macros from the initial release, all in one big file, but no longer maintained)
- [VCD Archive](#) (VCD archive, vcdprog.zip, for purchasers of the book)
- [Mosaic displays web applet](#)
- [Mosaic displays User's Guide](#) (.pdf [318k])
- [Sieve diagrams web applet](#)
- [Gallery of Data Visualization](#)

Next

- Fourfold Display for 2x2 Tables
- Association Plots
- Mosaic Display



```
> library(vcd)
```

vcd: Visualizing Categorical Data

<http://cran.r-project.org/web/packages/vcd/index.html>