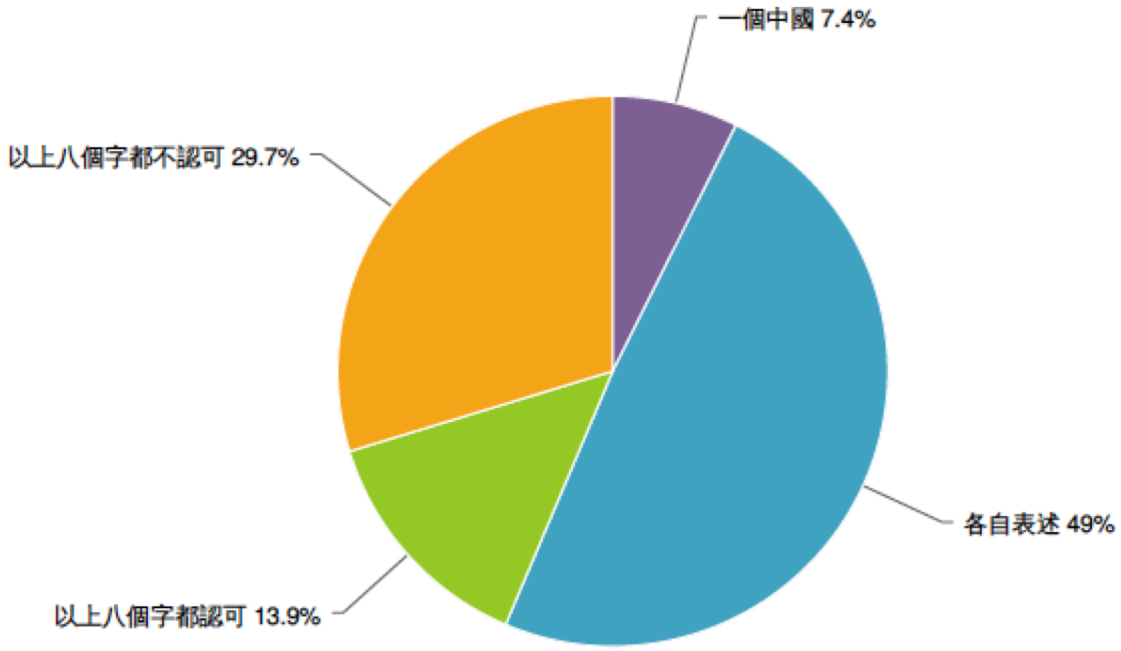
意義探勘怎麼做？
談談如何從小資料看見大格局

劉正山
中山大學 “政治科學”教授
& Director of Smilepoll.tw

2017.11.12 @ 台灣資料科學學會年會

15. 「九二共識」中的「一個中國、各自表述」八個字中您最認可的內涵是什麼？



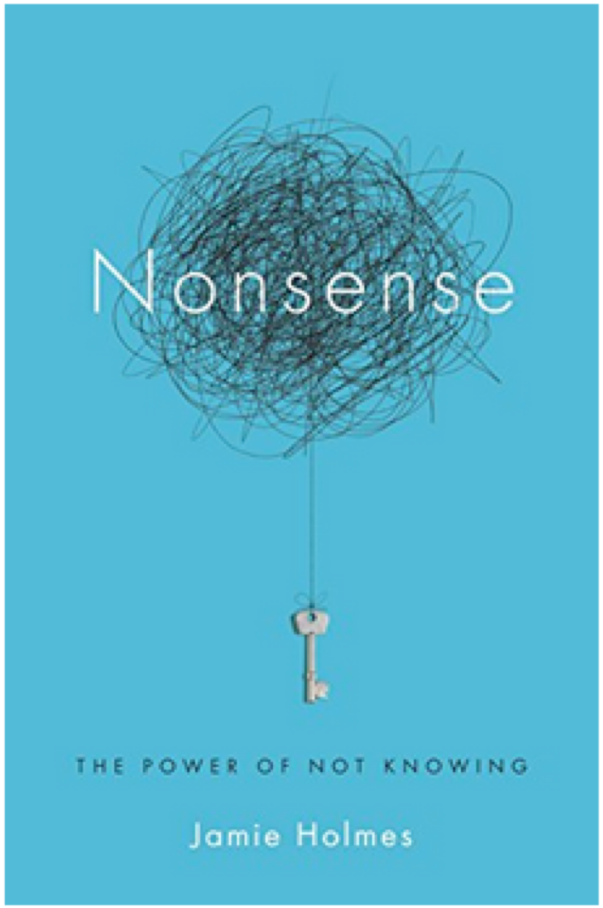
Value	Percent	Count
一個中國	7.4%	61
各自表述	49.0%	404
以上八個字都認可	13.9%	115
以上八個字都不認可	29.7%	245
	Total	825

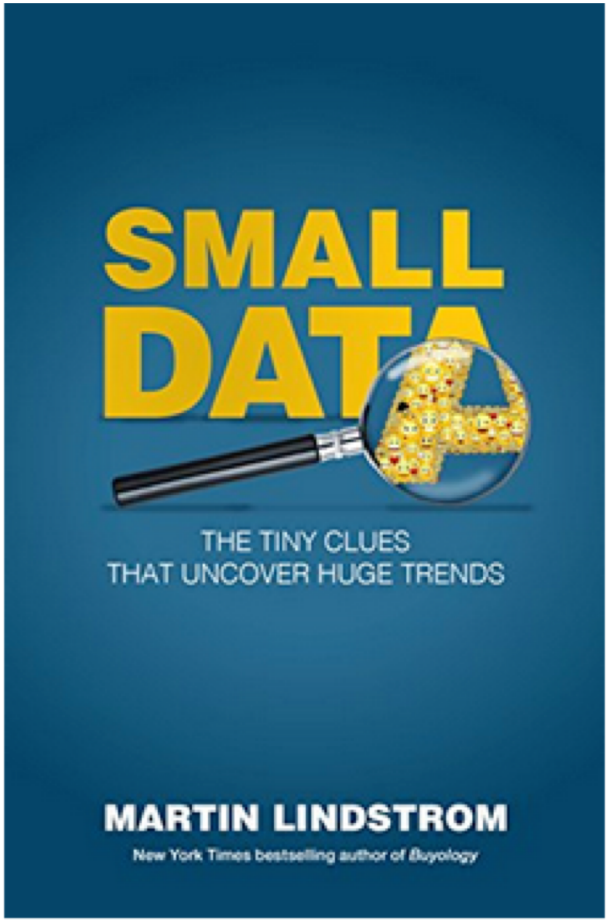
Statistics	
Sum	2,194.0
Average	2.7
StdDev	1.0
Max	4.0

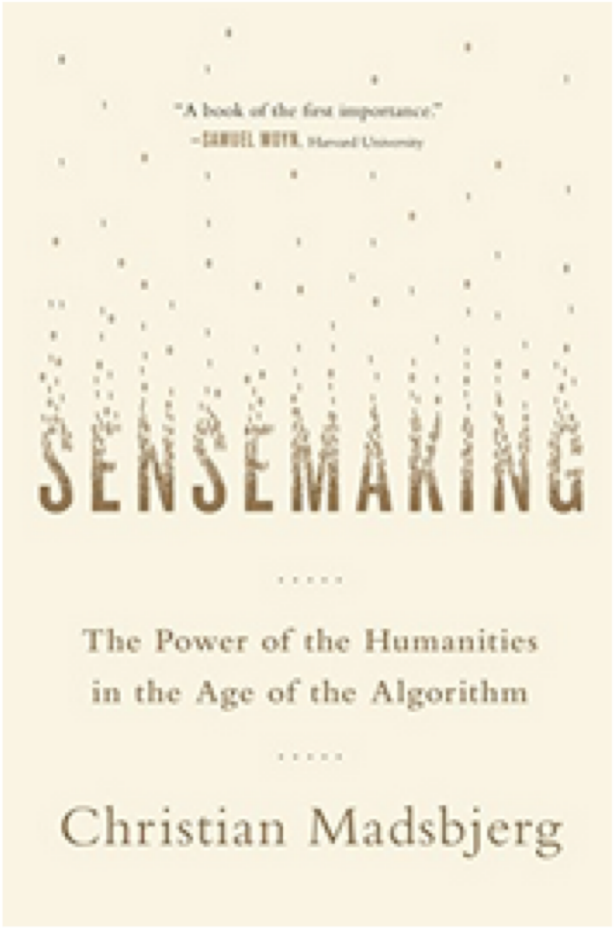
民意調查如何脫困？

- 以預測為目的，卻缺乏理想模型
- 以描述與推論作為目的，勾勒的卻只是粗稿

REFLECTIONS FROM THE HUMANITIES







Holmes, J. (2015). *Nonsense: The Power of Not Knowing* (First Edition). New York: Crown Publishers. 《無知的力量》
Lindstrom, M. (2016). *Small Data: The Tiny Clues That Uncover Huge Trends*. New York City: St. Martin's Press. 《小數據獵人》
Madsbjerg, C. (2017). *Sensemaking: The Power of the Humanities in the Age of the Algorithm*. New York, NY: Hachette Books.

資料輔助的意義織造
DATA-ASSISTED MEANING NETTING

既然是探索，就是在發現，而不是在驗證。

資料數據可用於發現關聯，更可用於探勘意義。

Let's make our exploration DAMN right.

THICK DATA (DAMN) FOR DATA SCIENCE

這是我認為當前「資料科學」在學科定義上忽略的環節

DAMN的方法論

不妨先辨識自己有興趣的概念或面向（什麼價值、什麼行為、什麼態度？），再透過資料進行探索。一面辨識出不同價值、態度、行為之間的可能關係，一面與自己的預期關係進行對話。最後再來進行意義的詮釋。

對民意調查的專業來說，用耙的不如**用問的**直接

文字探勘的大數據 終將與 直接訪問得到的小數據 分進合擊

將網調平台視為意義探勘的平台

- 資料科學家從資料聆聽者（被動爬梳挖來或買來的數據）轉換為資料創造者（主動收集到被研究對象價值和偏好）。
- 降低資料雜訊及更快速的決策。
- 形成社群後可以創造定群追蹤樣本（**panel data**），產生變數的合併帶來的巨大價值。
- 先以小數據作初探（**pilot stud**），之後再啟動隨機電話抽樣，將大幅增加推論力度。
- 初探階段便可以進行隨機分派實驗（**A/A**前測、**A/B**對照），找出意義和印證想法。

微笑小熊 調查小棧

首頁 會員登入 佈告欄 成果筆記 調查資料庫 關於 FB粉絲頁 RSS訂閱

最新消息

[新問卷:最熟悉的人PARTII] 爸爸，是男人最溫柔的...

[新問卷:最熟悉的人PartII] 爸爸，是男人最溫柔的名稱，對每個人來說，爸爸在心中都有一定的地位，但也有媽媽、姐姐..等其他家人也扮演著爸爸的角色。在父親節的前夕，無

進行中的訪問

7/15最熟悉的人PARTII | 博愛座博愛?

最新問卷：最熟悉的人 part II 17

2016-07-15

爸爸，是男人最溫柔的名稱，對每個

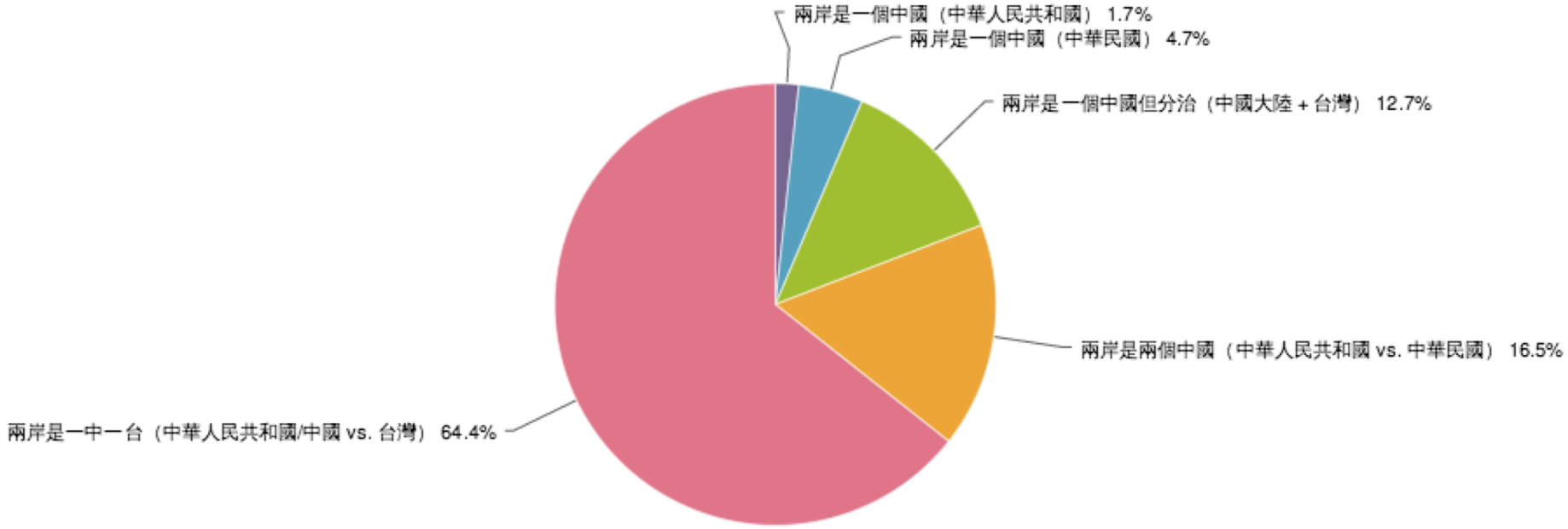
「在小熊身上得到更多其他人對社會的觀點，也改變自己不以自身想法為事件之看法。」 by 石凱云 (2016.05)

贊助單位

微笑小熊調查小棧為立場中立且非營利的網路服務團隊，得到教育部及科技部的支持與肯定。小棧自2012年成立後，已得到眾多網友

一般民調市調只會問出偏重於行為及偏好的問題。從DAMN觀點來看，我們還可以問出更多關於價值觀的問題。

25. 關於兩岸關係的說法很多種，那一個最符合您的想法？

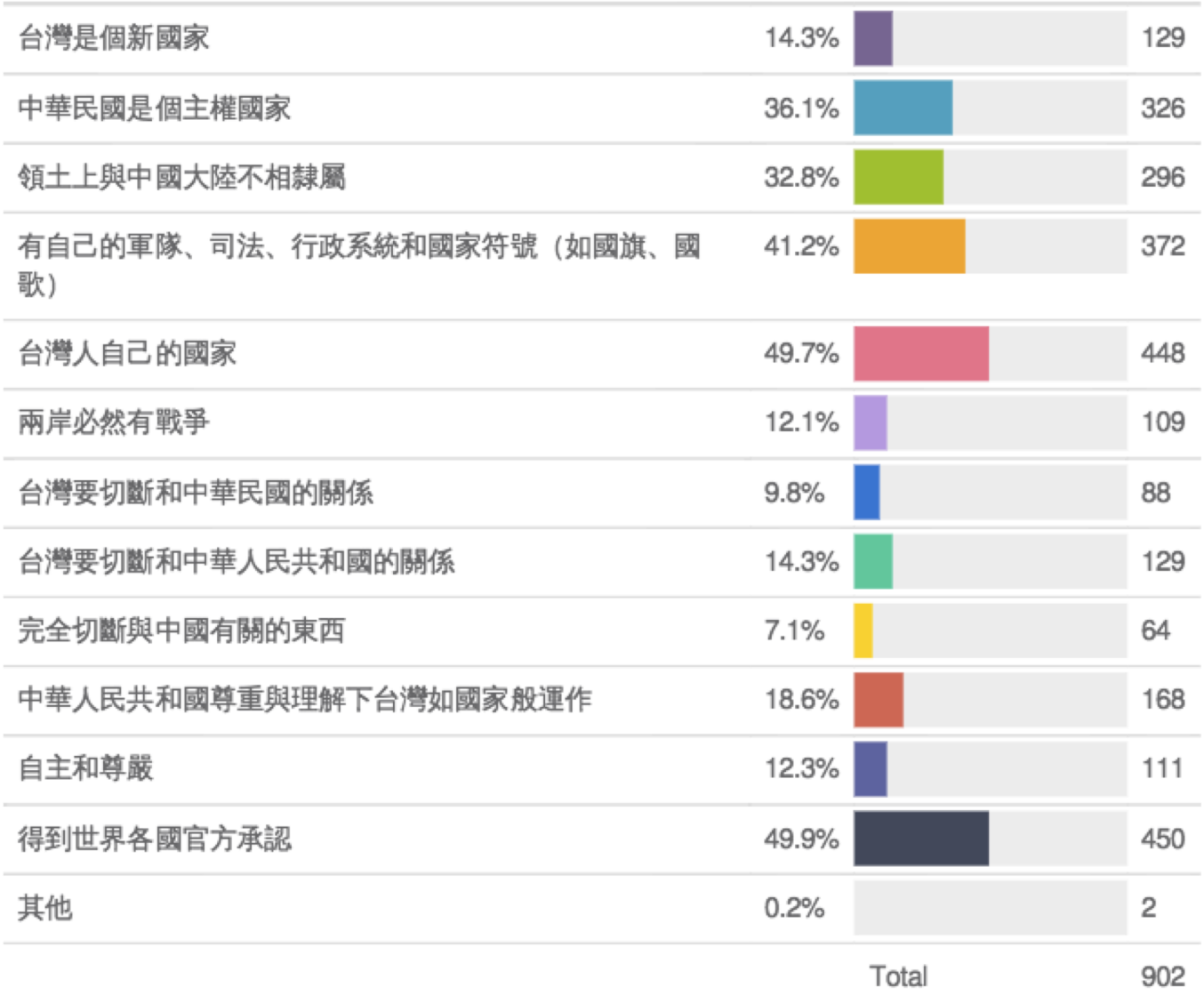


Value	Percent	Count
兩岸是一個中國（中華人民共和國）	1.7%	14
兩岸是一個中國（中華民國）	4.7%	39
兩岸是一個中國但分治（中國大陸 + 台灣）	12.7%	105
兩岸是兩個中國（中華人民共和國 vs. 中華民國）	16.5%	136
兩岸是一中一台（中華人民共和國/中國 vs. 台灣）	64.4%	531
	Total	825

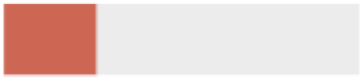
Statistics	
Sum	3,606.0
Average	4.4
StdDev	1.0
Max	5.0

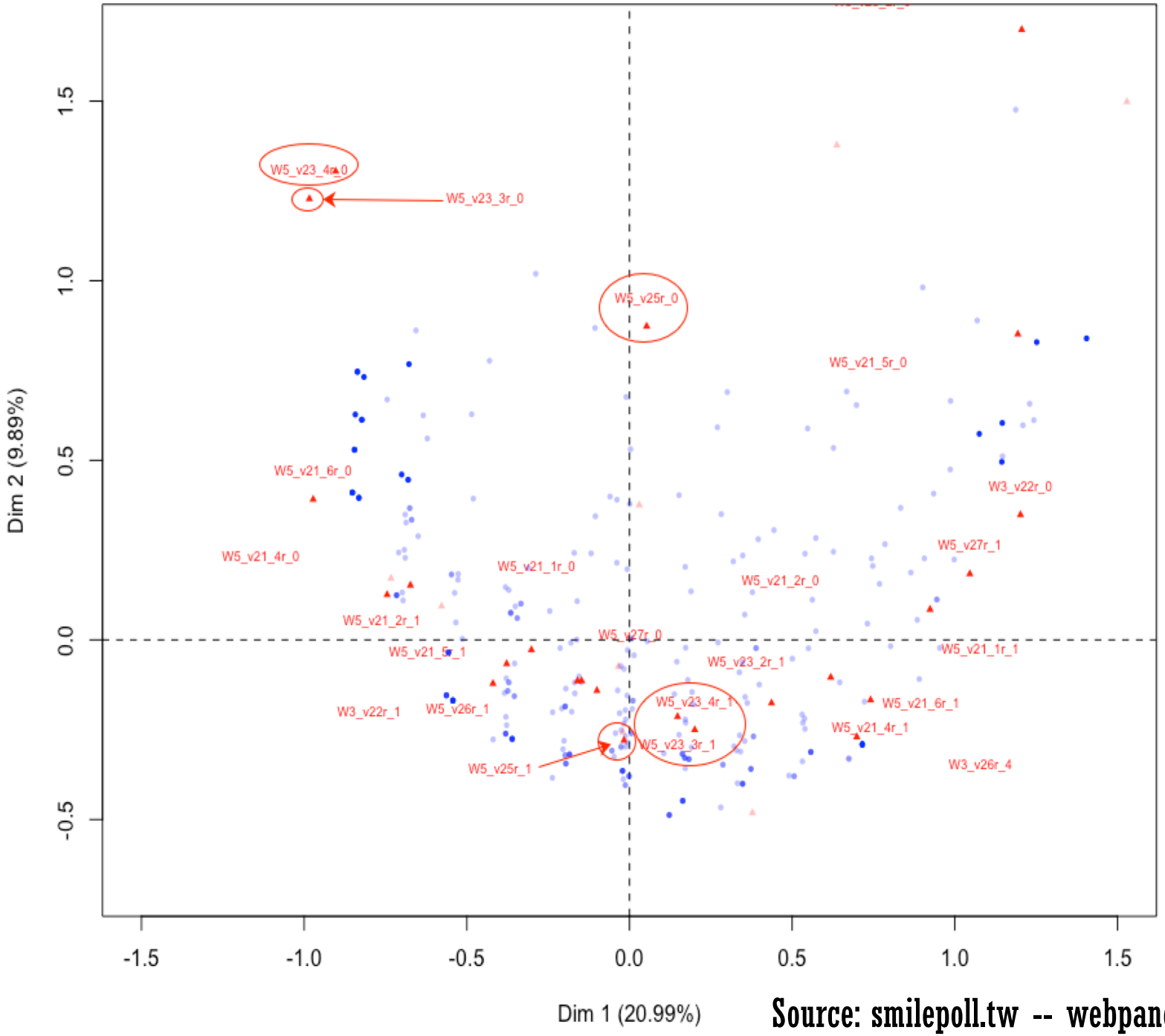
你有想過，台灣民眾對於「獨立」的定義有很多種，而且很可能沒有什麼共識嗎？

19. 說到獨立，您最先想到什麼？請從下面挑出最符合這個想法的選項。（可複選）

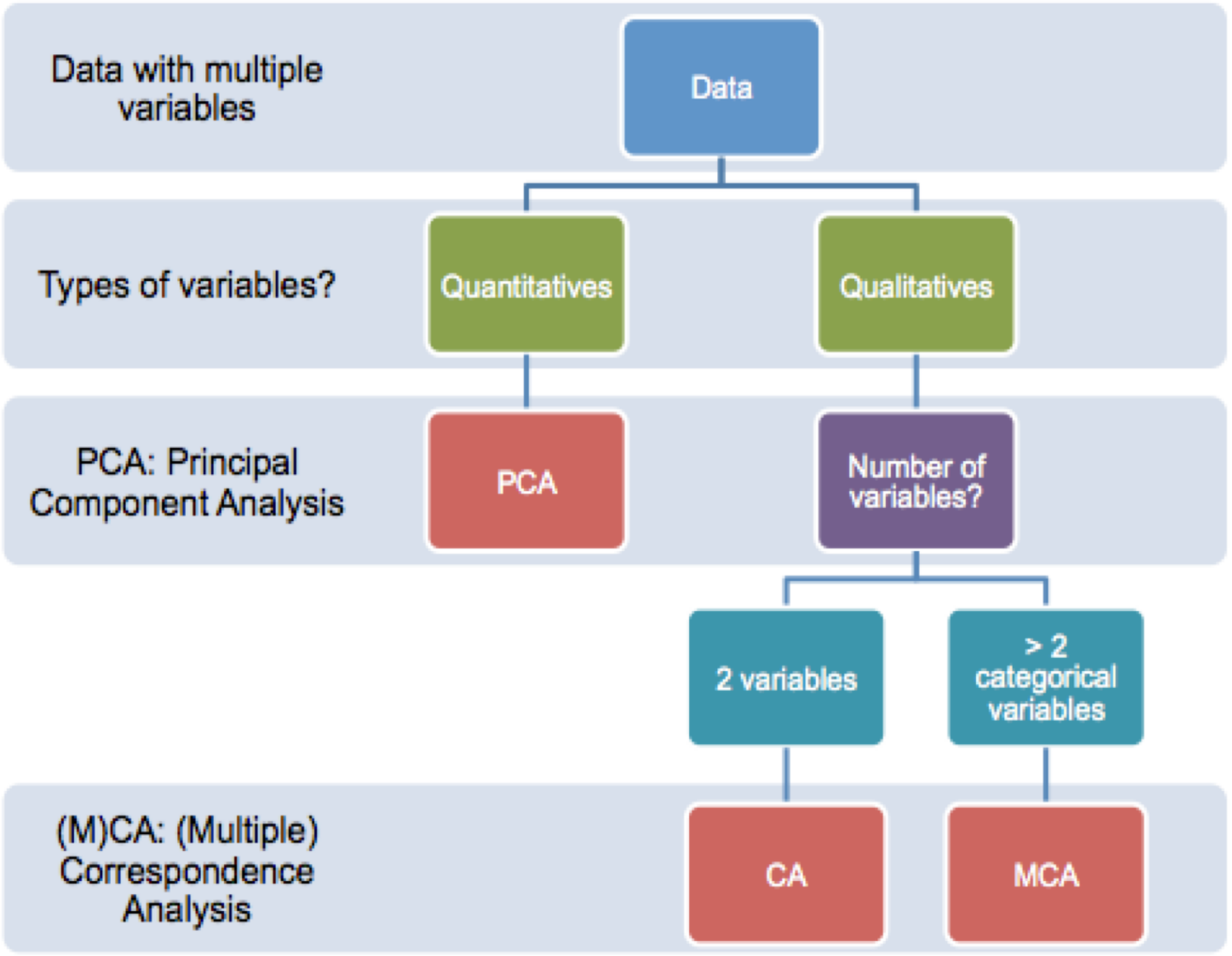


20. 說到統一，您最先想到什麼？請從下面挑出最符合這個想法的選項。（可複選）

中國併吞台灣	49.0%		441
中華人民共和國統一全中國	23.5%		212
中華民國統一全中國	10.8%		97
中華人民共和國消滅中華民國	25.4%		229
中國人的新時代	8.4%		76
中國夢實現	7.9%		71
台灣人最好的選擇	7.0%		63
在一國兩制下，台灣高度自治	16.4%		148
獨立戰爭	11.0%		99
某超商/食品企業	26.5%		239
其他	1.9%		17
		Total	901



Source: smilepoll.tw -- webpanel 2015-2016



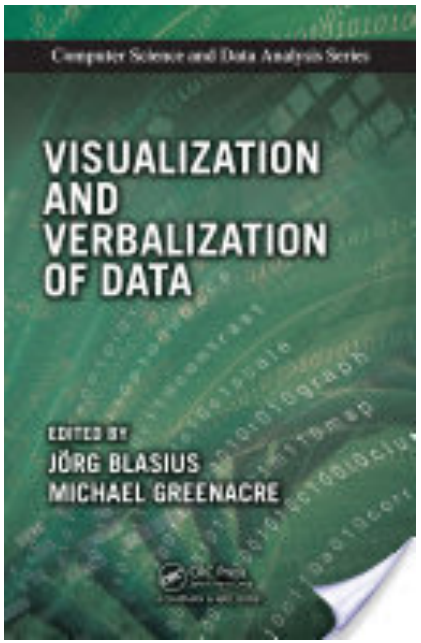
R packages for the analyses: **FactoMineR** (PCA, CA, MCA); **ade4** (PCA, CA, MCA); **stats** (PCA); **ca** (CA); **MASS** (CA)

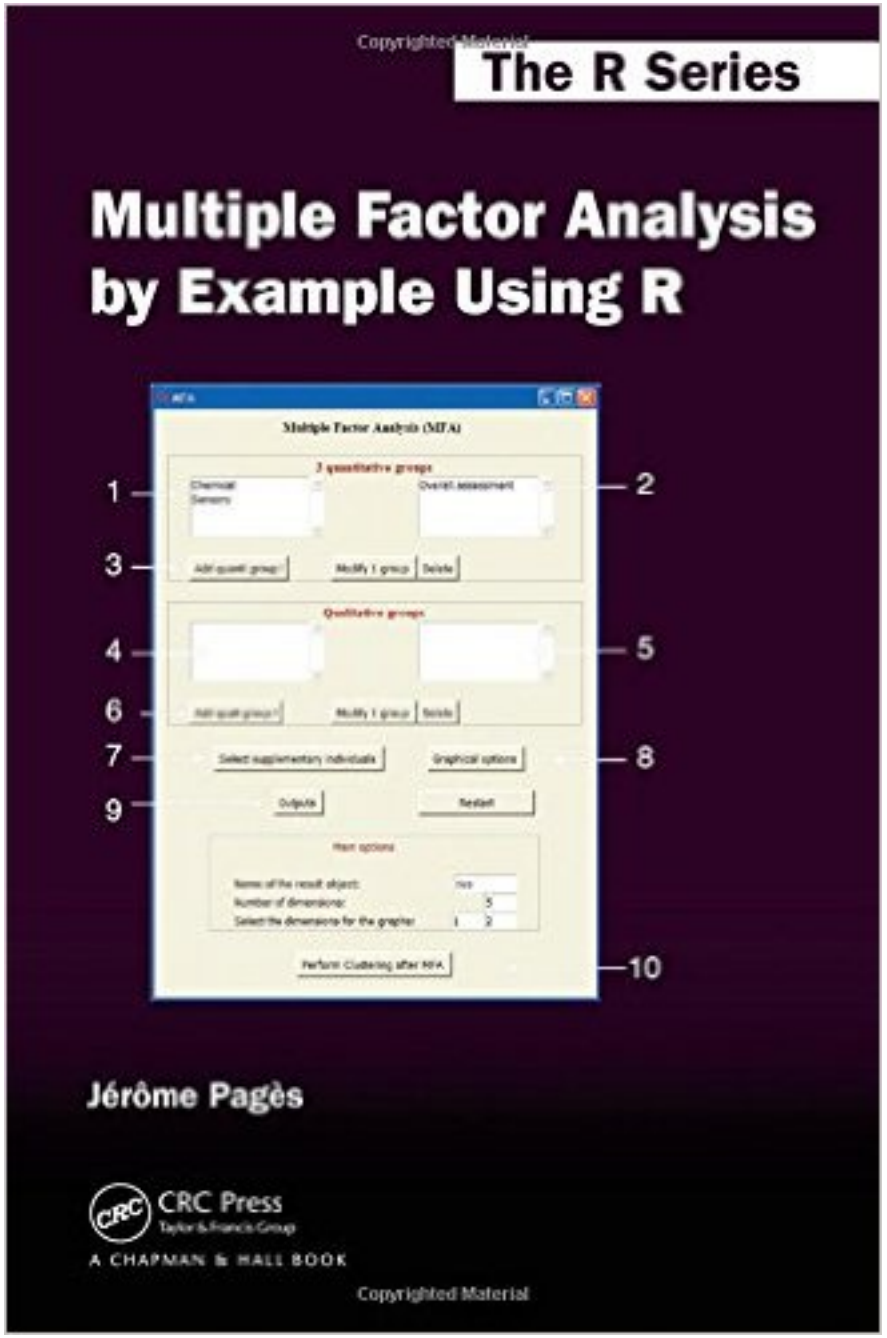
Use **factoextra** to easily extract and visualize the results

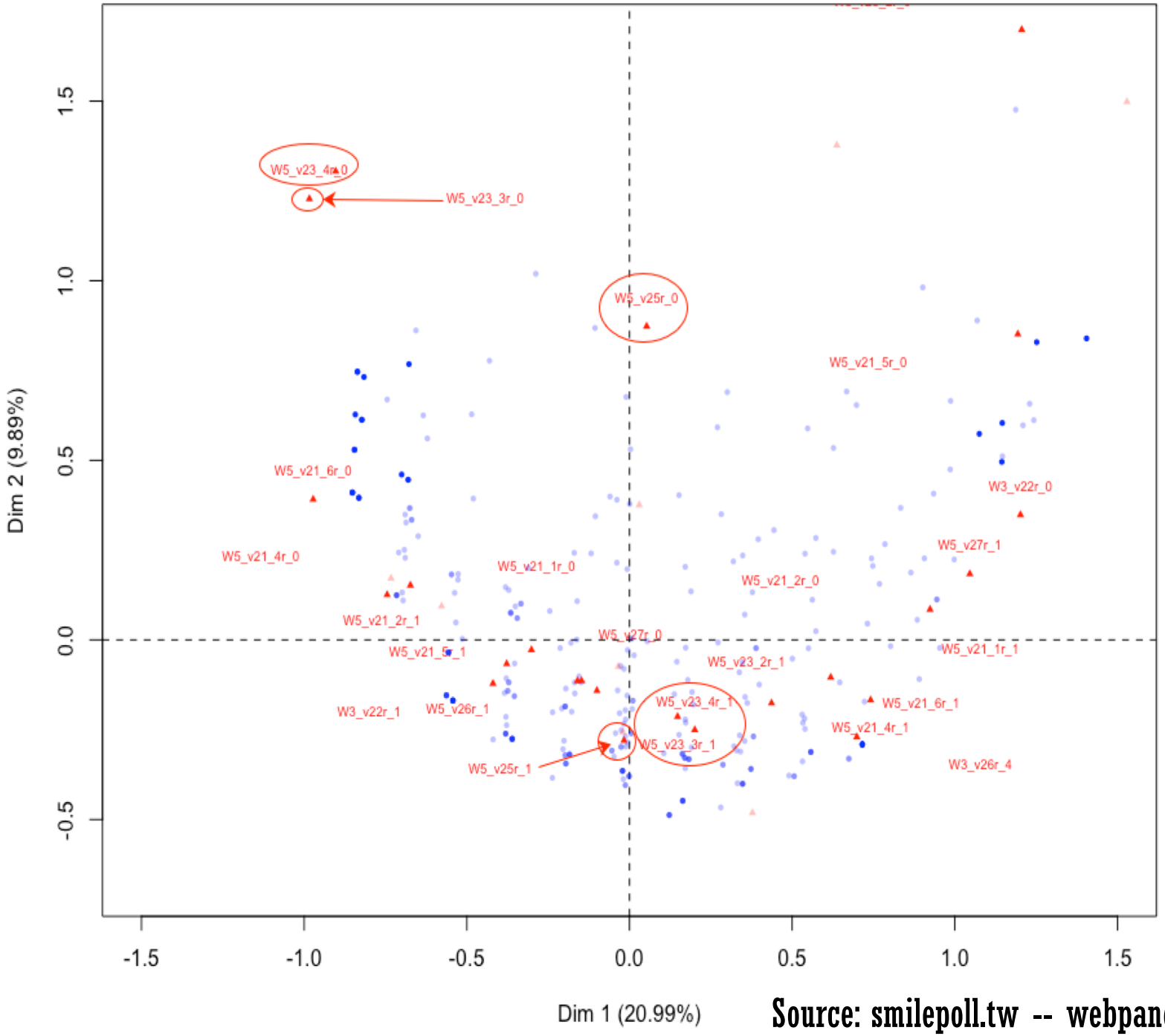
MCA 方法的特性

- **MCA像Principle Component Analysis (PCA)** 一樣，也可視為維次縮減的方法。「組合變數產生概念」，以及「詮釋概念」的過程中同樣充滿了主觀（「人文」？「科學」？）。
- **MCA** 中的**"dimension"** 就是 **PCA**中的 **component**。
- **MCA**特別之處在於：其目的並不只是可以「從多個類別型變數中找出核心概念」而已（要這樣做的話，可以使用**Categorical Principal Components Analysis, CATPCA**），而是可以用於協助探索類別型資料中不同變數的數值（問卷選項）之間的潛在關係。


MULTIPLE
CORRESPONDENCE
ANALYSIS
Brigitte LeRoux
Henry Rouanet
Series: Quantitative Applications
in the Social Sciences
163
SAGE


Computer Science and Data Analysis Series
VISUALIZATION
AND
VERBALIZATION
OF DATA
EDITED BY
JÖRG BLASIUS
MICHAEL GREENACRE


Copyrighted Material
The R Series
Multiple Factor Analysis
by Example Using R
Jérôme Pagès
CRC Press
A CHAPMAN & HALL BOOK
Copyrighted Material



Source: smilepoll.tw -- webpanel 2015-2016

很貴的小數據

- Taiwan Election and Democracy Studies 2016
- Data Collection Period: 2017.1.17 ~ 4.28
- N=1,690
- $$$: > NTD 1,000,000

HTTP://JMP.SH/3XHJ7TZ

本講之R語法及資料檔下載

第一步：變數設計

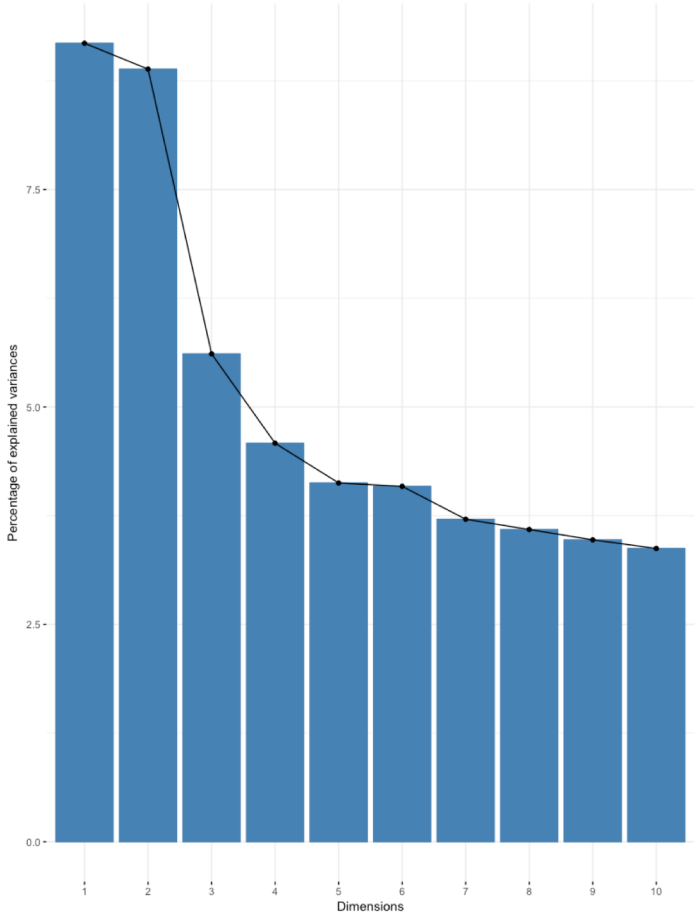
針對感興趣的現象、態度、價值面向，挑選、設計出變數或問卷題。一般民調市調只會問出偏重於行為及偏好的問題。從DAMM觀點來看，我們還可以問出更多關於價值觀的問題。

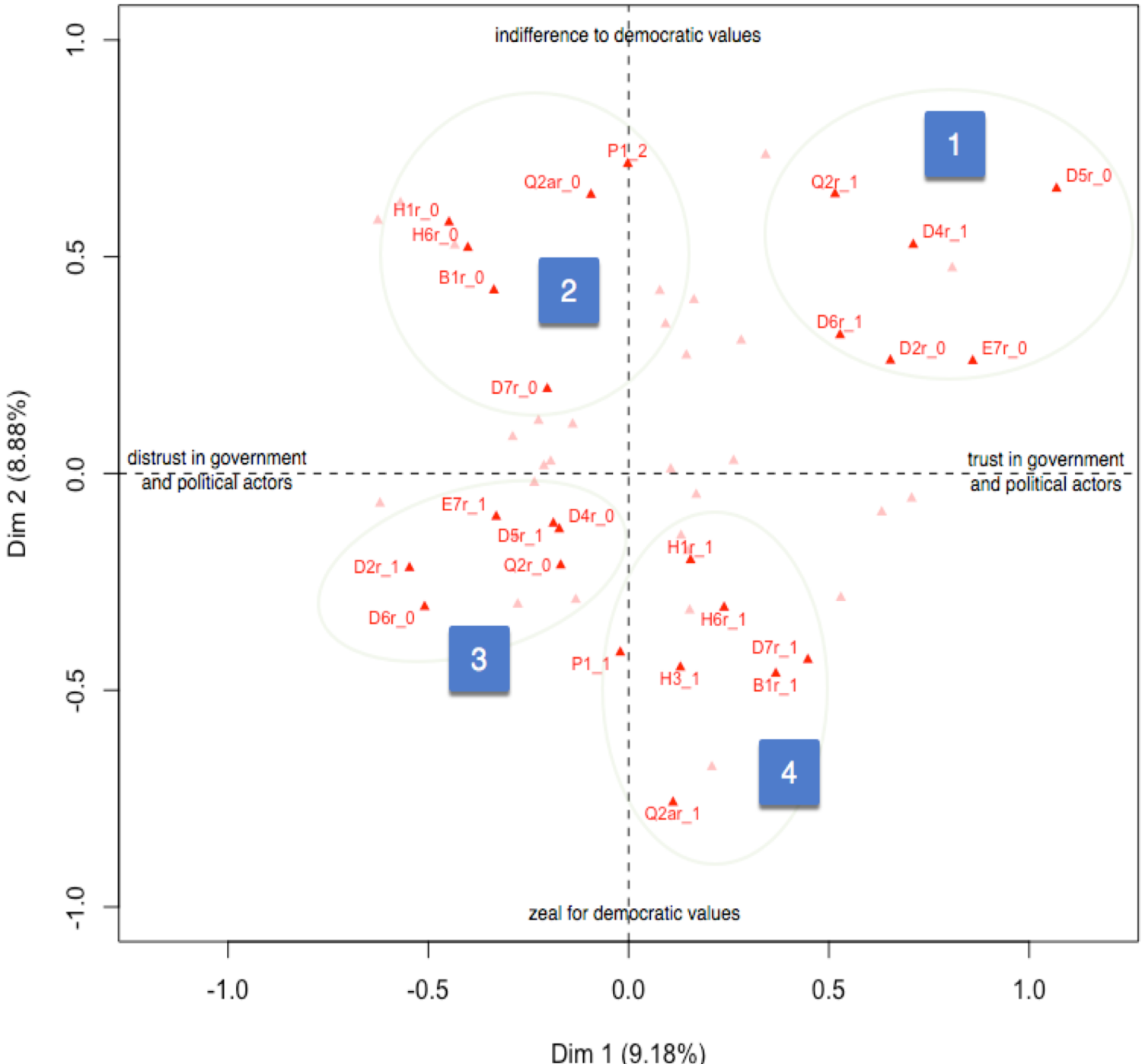
```r
load("teds2016e.rda")
teds2016eforMCA <- select(teds2016e,
                          c(# 核心變數 (core vars)
                            A1r, # 算常接觸媒體=1; otherwise=0
                            B1r, # 對政治算有興趣=1; otherwise=0
                            D1r, # 覺得對政府沒有影響力=1; otherwise=0
                            D3r, # 覺得政治太複雜=1; otherwise=0
                            D2r, # 覺得政治人物不管我們的想法=1; otherwise=0
                            D4r, # 覺得政府作的決定大多時候是對的=1; otherwise=0
                            D5r, # 覺得政府常浪費稅金=1; otherwise=0
                            D6r, # 覺得政府會優先考量公共利益=1; otherwise=0
                            D7r, # 覺得自己算是懂政治=1; otherwise=0
                            E3r, # 覺得大多數政治人物是可信的=1; otherwise=0
                            E4r, # 覺得政治人物(才)是國家的主要問題=1; otherwise=0
                            E5r, # 我們的國家需要政治強人=1; otherwise=0
                            E6r, # 重要政策應該由人民而非政治人物來做=1; otherwise=0
                            E7r, # 政治人物只在乎權貴者的利益=1; otherwise=0
                            G1r, # 覺得過去一年台灣經濟變好=1; 沒變=2; 變差=3
                            G2r, # 覺得未來一年台灣經濟會變好=1; 沒變=2; 會變差=3
                            H1r, # 投票是種責任=1; otherwise=0
                            H3, # 民主應優先於其他型態的政權=1; otherwise=0
                            H5r, # 算是滿意台灣的民主=1; otherwise=0
                            H6r, # 透過投票可以改變現況=1; otherwise=0
                            P1, # 族群認同-台灣人=1; 都是=2; 中國人=3
                            P2r, # 兩岸關係會變和諧=2; 不變=1; 變緊繃=3
                            Q2r, # 11分量尺：表對國民黨算有好感(>=6); otherwise=0
                            Q2ar, # 11分量尺：表對民進黨算有好感(>=6); otherwise=0
                            partisanship # 自覺有(些)政黨傾向=1; otherwise=0
                            ))
```

第二步：樣貌辨識

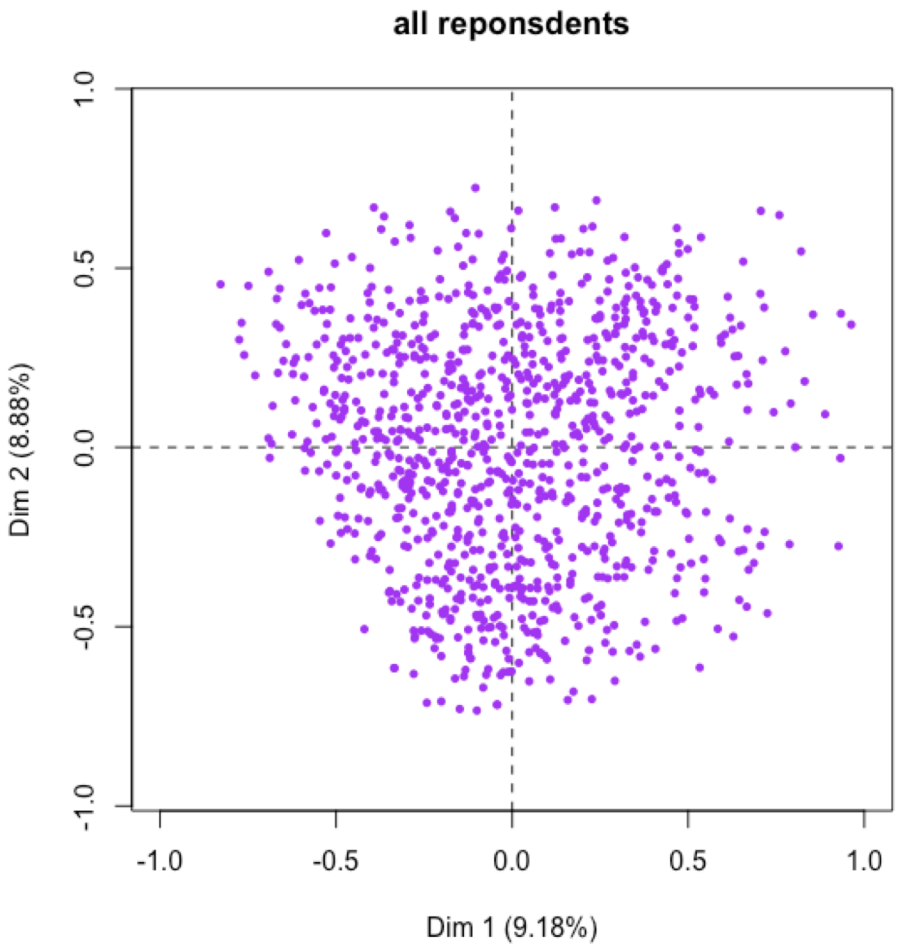
發掘變數及類別選項之間的潛在連結

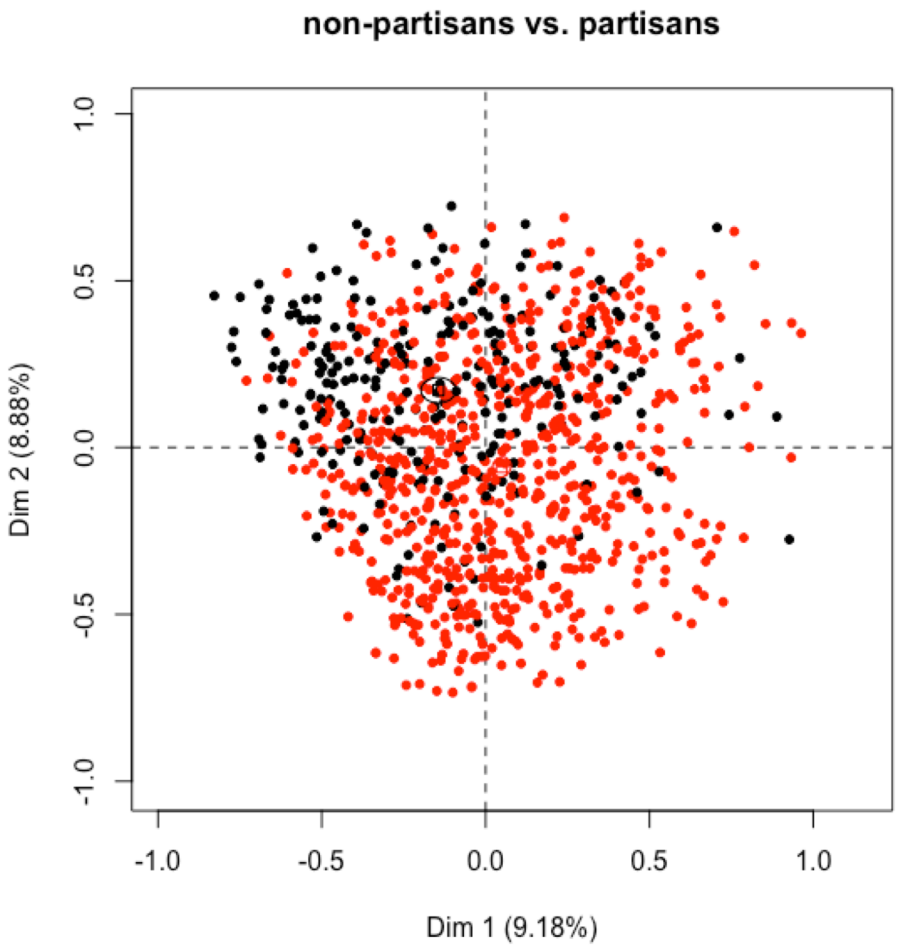
```r
# MCA分析
library(FactoMineR)
library(factoextra)
names(teds2016eforMCA.nona)
res<-MCA(teds2016eforMCA.nona, ncp=10, graph= F) #ncp 10個維次
fviz_screeplot(res, ncp=10) + labs(title="")
```



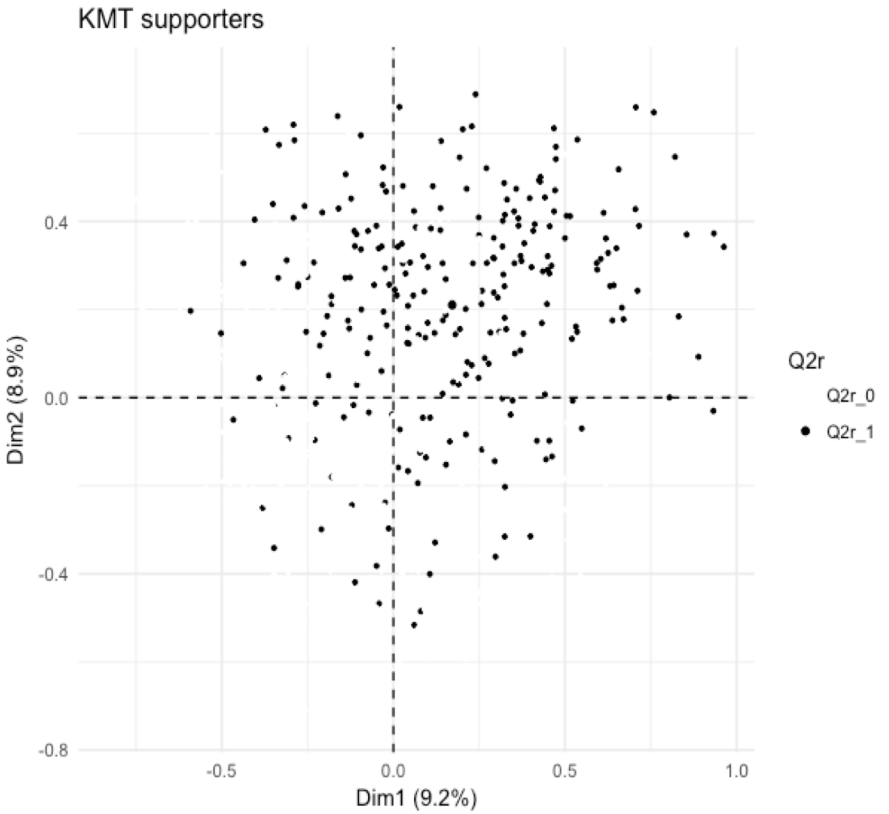

indifference to democratic values
zeal for democratic values
distrust in government and political actors
trust in government and political actors
Dim 2 (8.88%)
Dim 1 (9.18%)
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
1
2
3
4
P1_2
Q2ar_0
H1r_0
H6r_0
B1r_0
D7r_0
Q2r_1
D5r_0
D4r_1
D6r_1
D2r_0
E7r_0
E7r_1
D5r_1
D4r_0
D2r_1
Q2r_0
D6r_0
H1r_1
H6r_1
D7r_1
P1_1
H3_1
B1r_1
Q2ar_1

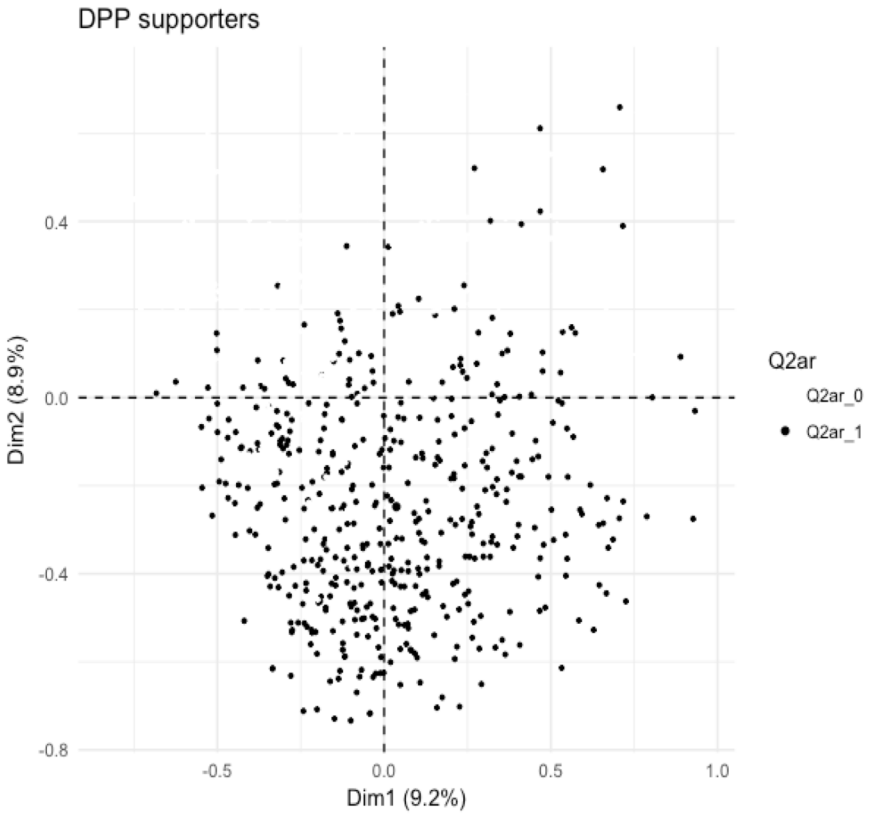
無政黨支持傾向者的樣貌





藍綠支持者的樣貌


KMT supporters
Dim2 (8.9%)
0.4
0.0
-0.4
-0.8
-0.5
0.0
0.5
1.0
Dim1 (9.2%)
Q2r
Q2r_0
Q2r_1


DPP supporters
Dim2 (8.9%)
0.4
0.0
-0.4
-0.8
-0.5
0.0
0.5
1.0
Dim1 (9.2%)
Q2ar
Q2ar_0
Q2ar_1

第三步：意義探勘

用問的可能會比給出牽強的答案或解釋來得重要

WHEN "DEMOCRACY" IS EMPHASIZED BY A POLITICAL PARTY…

Is it possible that the supporters of the opposing political party feel less passionate to democracy?

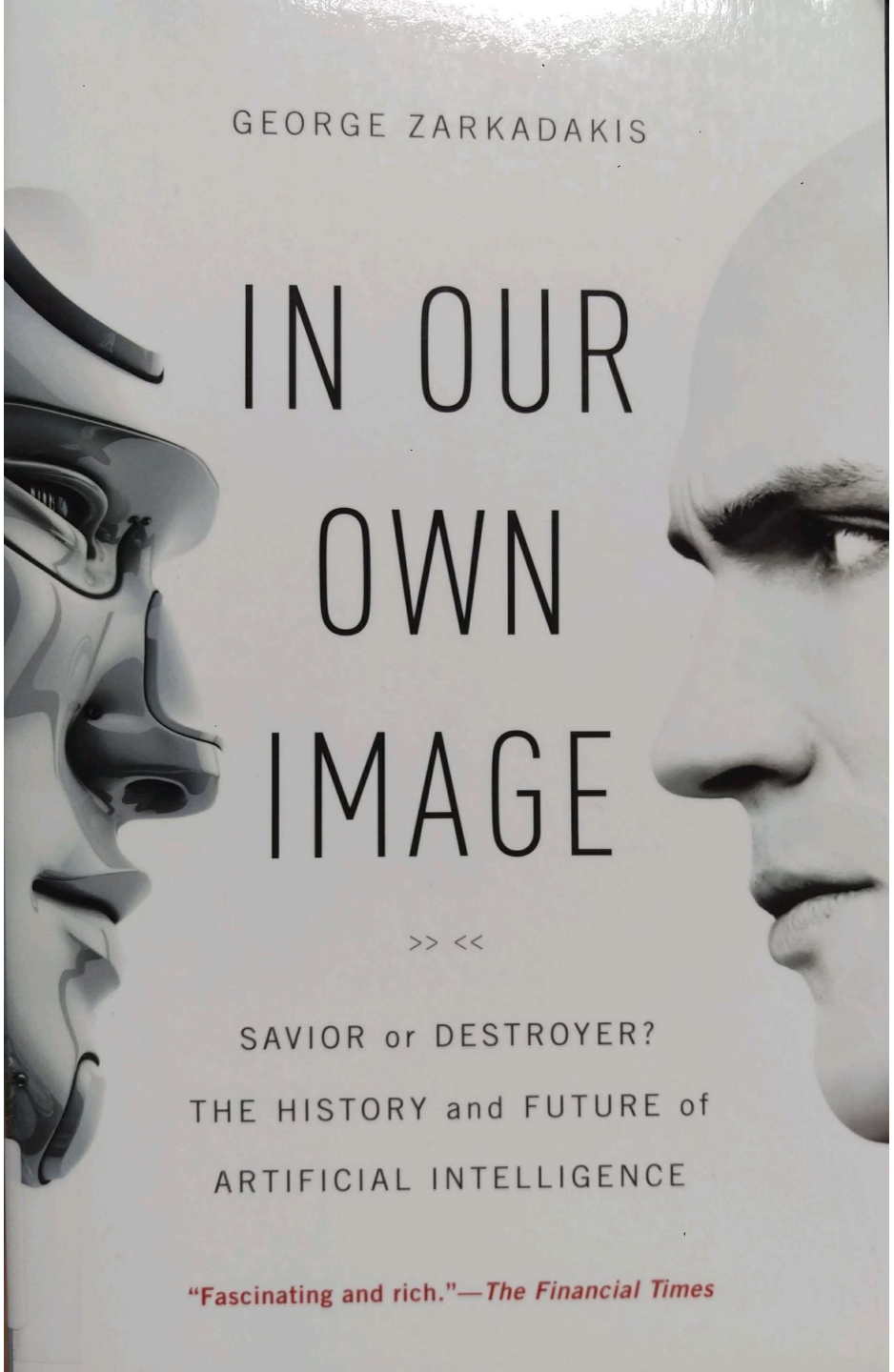
Isn't it too quick to signal and criticize that KMT supporters are not passionate for democracy?

Can't the value of their belief system, "trust in the government and political actors," be as well important for a functioning democratic society?

資料科學 VS. 魔術

經驗本身不見得能被拿來當作詮釋的材料，所以資料科學登場。
但資料科學家所作的詮釋又如何可信？

民眾需要知道提供資訊的人是否已經善盡了他們的責任。


GEORGE ZARKADAKIS
IN OUR OWN IMAGE
>> <<
SAVIOR or DESTROYER?
THE HISTORY and FUTURE of ARTIFICIAL INTELLIGENCE
"Fascinating and rich."—The Financial Times

我們如何抗拒對「確定感」的追求？

科學家、預言家在「幫助我們確定一件事」這一點上是相像的。

民眾相信自己所喜歡的。那麼擁有詮釋權力的資料科學家將具備比一般民眾更可能為惡的本錢。

面對未知，
我們只能不停的用「問」
來逼近真相

資料科學 = 技術成份高？人文社科 = 意義成份高？

探索什麼？做什麼用？

For what?

So what?

展望

- 期待**MCA**技術的突破（**rotation, multidimentional scaling, etc.**）
- 期待更多意義的挖掘（對難得的資料樣貌給予更多具洞察力的詮釋）
- 當資料的意義變得更重要時，期待我們都成為更謙虛的科學家
- 「發問」是種未來極需要的專業，因為所有資料背後的意義都是被問出來的。（國王有穿衣服嗎？）

（值得期待看到的）大格局

- 意義是資料與真實世界的連結點
- 透過對人問更深刻的問題，我們將得以發掘更多關於人的行為、價值、態度等等的之間的可能關聯樣貌
- 但資料科學家必需更勇敢去承認所知的限制，不宜過度自信地宣稱所見就是真相。

DAMN的續篇
FB搜尋: 資料吼

非常感謝全球R社群的奉獻，以及國內資料科學社群的努力！

劉正山 csliu@mail.nsysu.edu.tw

参考資料（1）MCA

- Blasius, J., & Greenacre, M. (Eds.). (2014). *Visualization and Verbalization of Data*. CRC Press.
- Husson, F., Le, S., & Pages, J. (2010). *Exploratory Multivariate Analysis by Example Using R* (1 edition). CRC Press.
- Pagès, J. (2014). *Multiple Factor Analysis by Example Using R* (1 edition). Boca Raton: Chapman and Hall/CRC.
- Pasek, J., Jang, S. M., Cobb, C. L., Dennis, J. M., & Disogra, C. (2014). Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data. *Public Opinion Quarterly*, *78*(4), 889–916.
- Roux, B. L., & Rouanet, H. (2009). *Multiple Correspondence Analysis*. SAGE Publications.

參考資料（2）MEANING MINING

- Blackburn, S. (2012). *What Do We Really Know? The Big Questions in Philosophy*. London: Quercus.
- Cohen, L. H. (2013). *I don't know: In Praise of Admitting Ignorance*. New York: Riverhead Books.
- Holmes, J. (2015). *Nonsense: The Power of Not Knowing* (First Edition). New York: Crown Publishers.
- Madsbjerg, C. (2017). *Sensemaking: The Power of the Humanities in the Age of the Algorithm*. New York, NY: Hachette Books.
- Sesno, F., & Blitzer, W. (2017). *Ask More: The Power of Questions to Open Doors, Uncover Solutions, and Spark Change*. New York: AMACOM.
- Zarkadakis, G. (2016). *In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence* (1 edition). Pegasus Books.