清華大學
Tsinghua University
政治学系

Learning Visualization with Dr. Hu

数据可视化理论与实践

胡悦
清华大学政治学系

提要

内容

- 鉴赏
- 构建
- 实操（基于ggplot）

参考书

Cleveland, William S. 1985. *The Elements of Graphing Data*. 1st Printing edition. Monterey, Cal: Wadsworth, Inc.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd edition edition. Cheshire, Conn: Graphics Press.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. 2nd ed. New York: Springer-Verlag.

张杰. 2019. 《R语言数据可视化之美：专业图表绘制指南》. 第1版. 电子工业出版社.

心理准备

可视化鉴赏

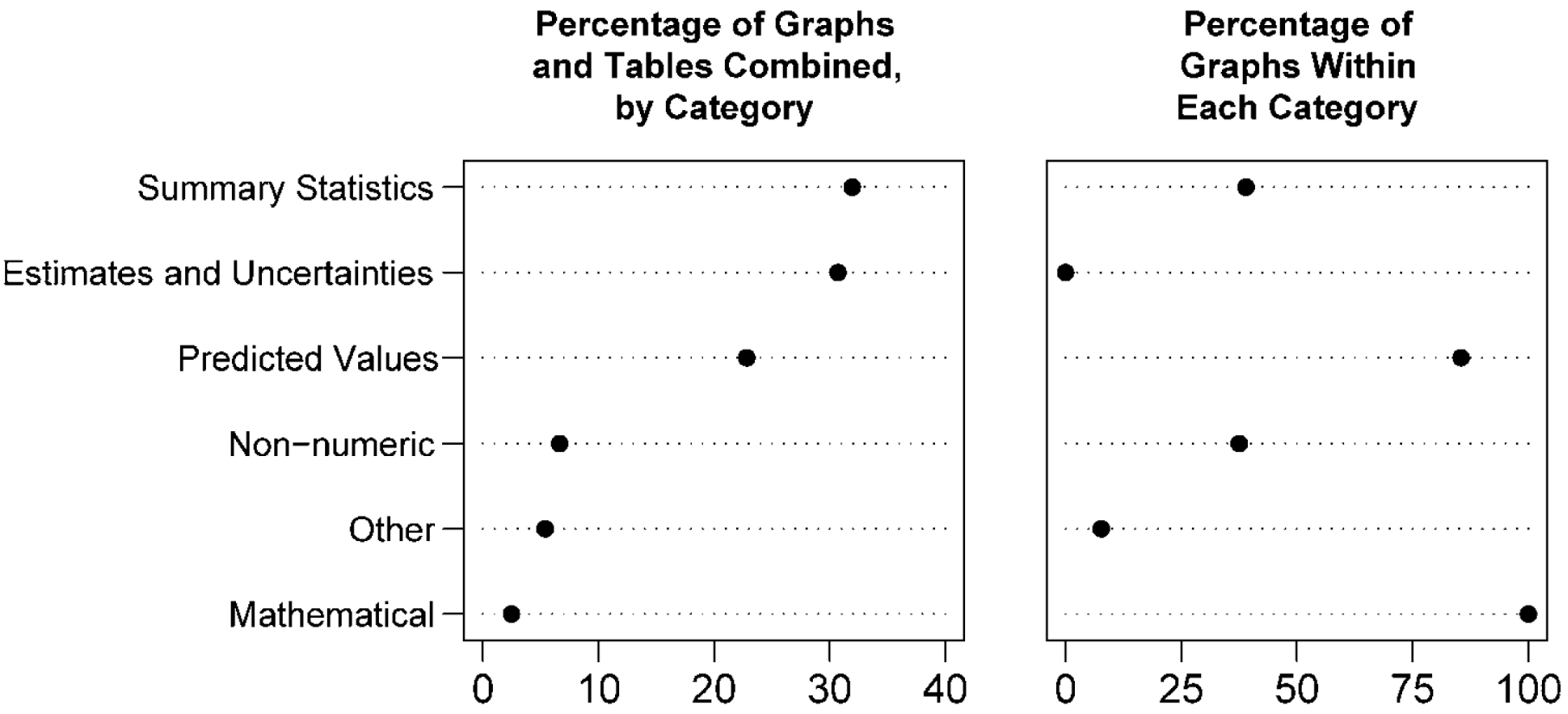
“有学问的人”是怎么做的?

Kastellec, Jonathan P., and Eduardo L. Leoni. 2007. “Using Graphs Instead of Tables in Political Science.” *Perspectives on Politics* 5(4): 755–71.

检验APSR， AJPS和PA 2006年五刊，发现如下结果

	Percentage of Graphs and Tables Combined, by Category	**Percentage of Graphs Within Each Category**
Summary of Statistics	31.95	38.96
Estimates and Uncertainties	30.71	0.00
Predicted Values	22.82	85.45
Non-numeric	6.64	37.50
Other	5.39	7.69
Mathematical	2.49	100.00

如何可视化上表?


Percentage of Graphs and Tables Combined, by Category
Percentage of Graphs Within Each Category
Summary Statistics
Estimates and Uncertainties
Predicted Values
Non−numeric
Other
Mathematical
0
10
20
30
40
0
25
50
75
100

为何学者爱做表?

表

- 容易制作
- 发表常见
- “有助于后续研究”

图

- 制作费事
- 信息不“精确”
- 形式不统一

那……为什么还作图?

A picture is worth a thousand words.

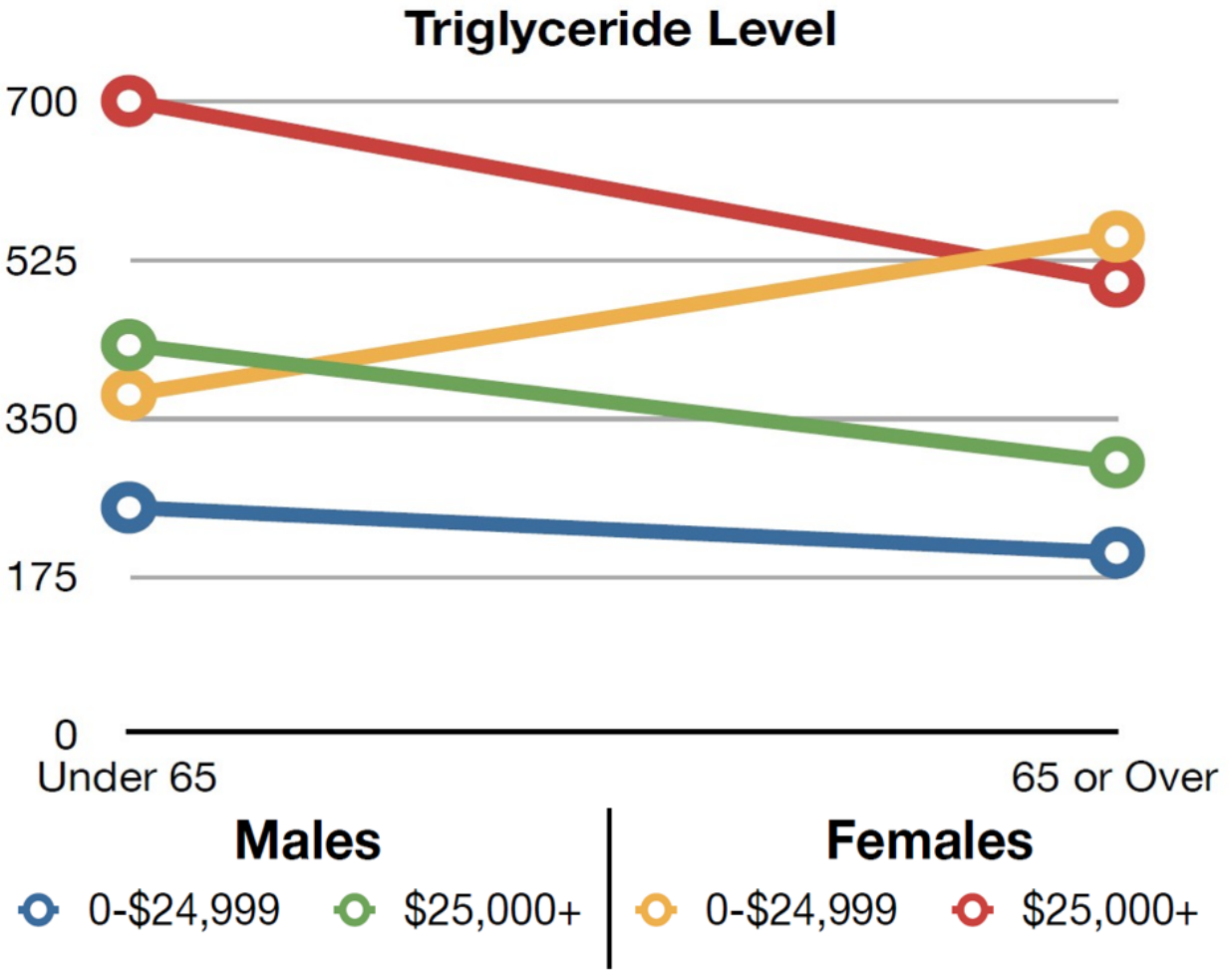
印象更深

便于比较、突出重点

信息传递高效

比表格占据更小版面

哪个性别及收入水平人群的三酸甘油脂水平趋势与其他均不同?

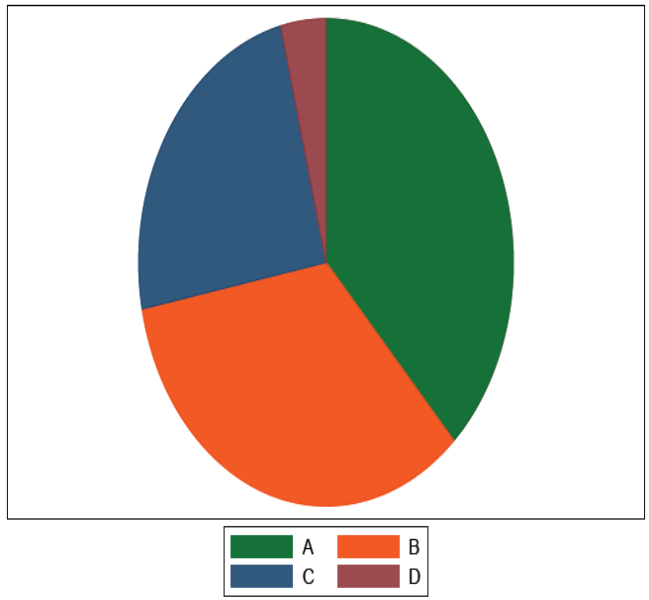


何为好的可视化?

政党	得票率
A	38
B	34
C	24
D	4

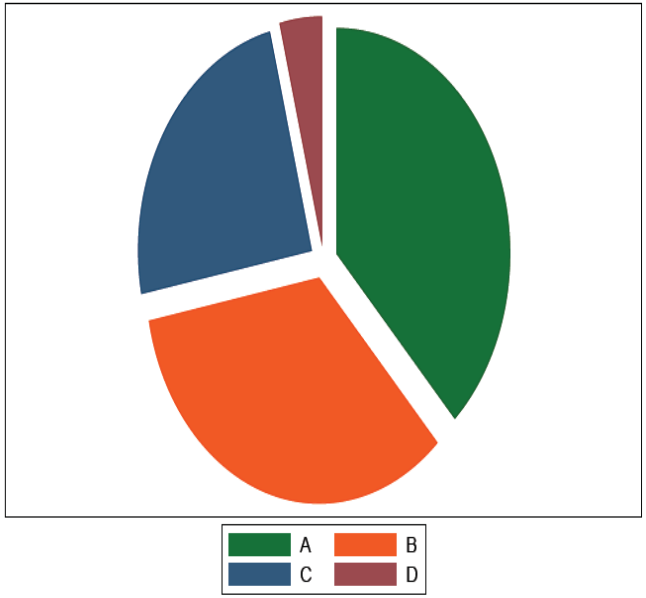
饼状图?

政党	得票率
A	38
B	34
C	24
D	4



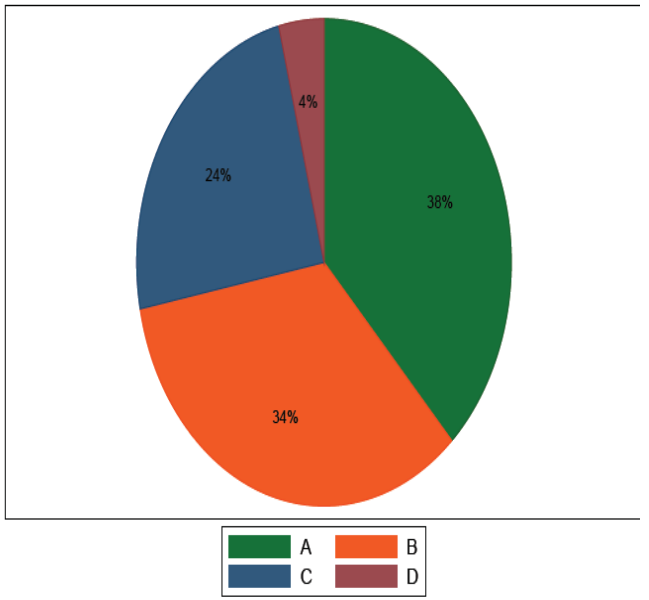
饼状图?

政党	得票率
A	38
B	34
C	24
D	4



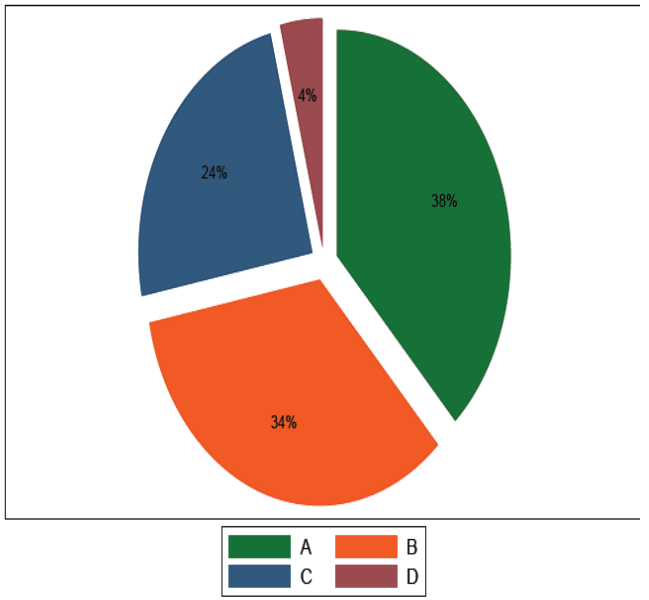
饼状图?

政党	得票率
A	38
B	34
C	24
D	4



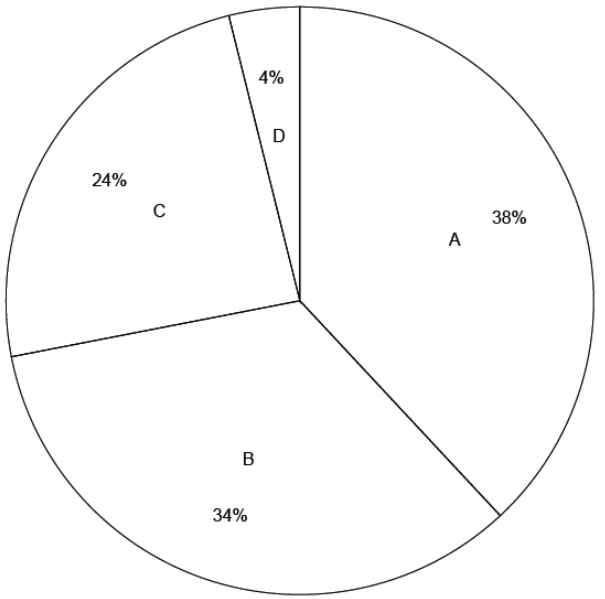
饼状图?

政党	得票率
A	38
B	34
C	24
D	4



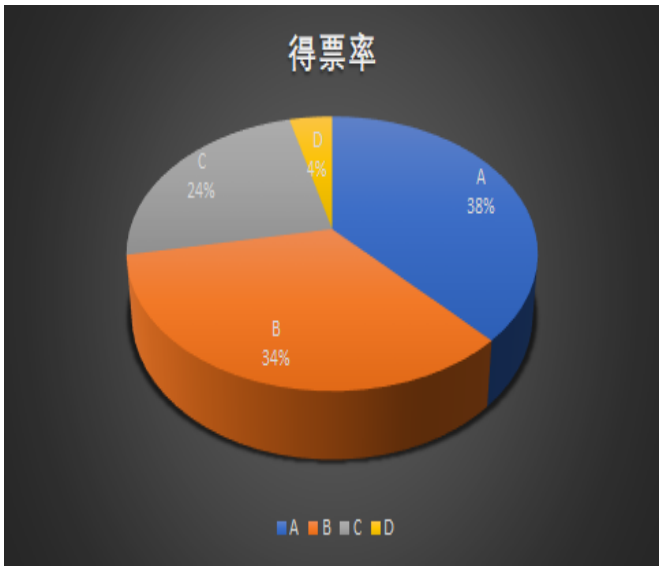
饼状图?

政党	得票率
A	38
B	34
C	24
D	4



饼状图?

政党	得票率
A	38
B	34
C	24
D	4



Why Is Pie Chart a Big No-No?

- 用于呈现百分比（一维）
- 容易误导读者（见后）
- 浪费油墨（Why do we care?）

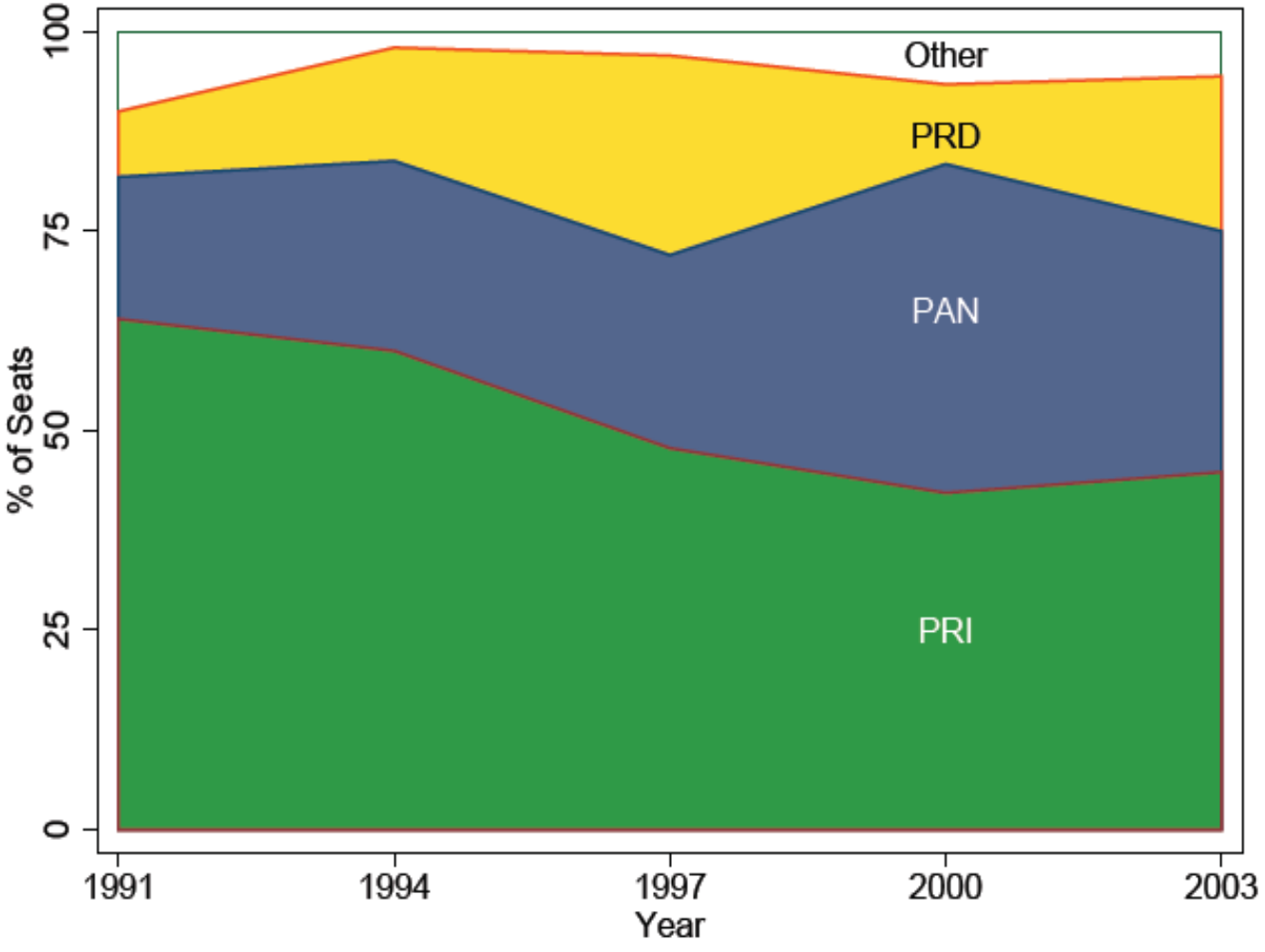
被誉为垃圾图形（junk chart）首选

另一个例子

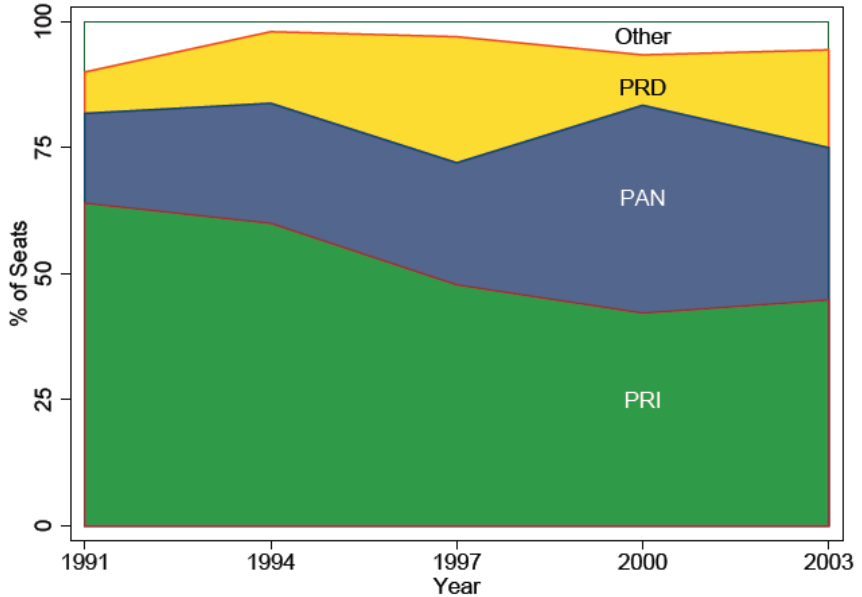
墨西哥各政党在大选中得票比例趋势

Party	PRI	PAN	PRD	Others
1991	64	17.8	8.2	10
1994	60	23.8	14.2	2
1997	47.8	24.2	25	3
2000	42.2	41.2	10	6.6
2003	44.8	30.2	19.4	5.6

怎么可视化这个数据?


Other
PRD
PAN
PRI
% of Seats
0
25
50
75
100
1991
1994
1997
2000
2003
Year

Why No-no?

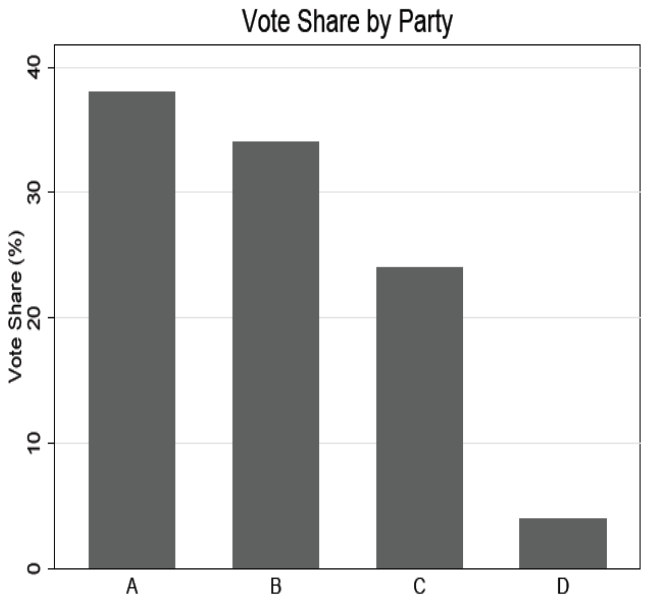


1. 比折线图费更多墨。

2. 各个时间序列间的比较容易误导读者，而且焦点容易被面积所模糊。

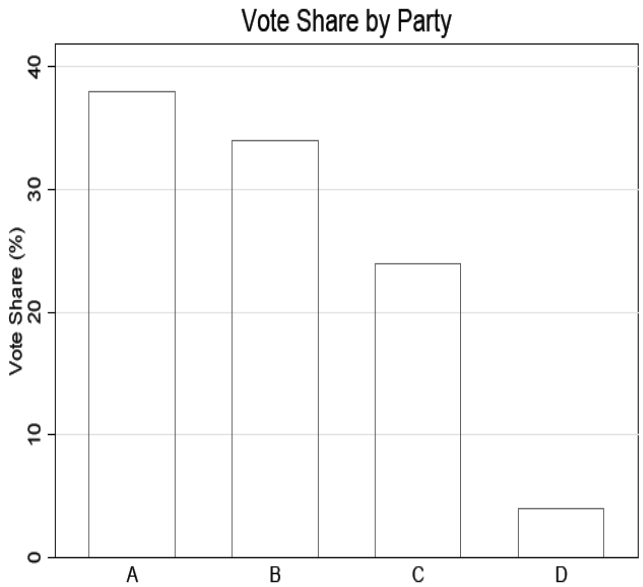
Again, 何为好的可视化?

政党	得票率
A	38
B	34
C	24
D	4



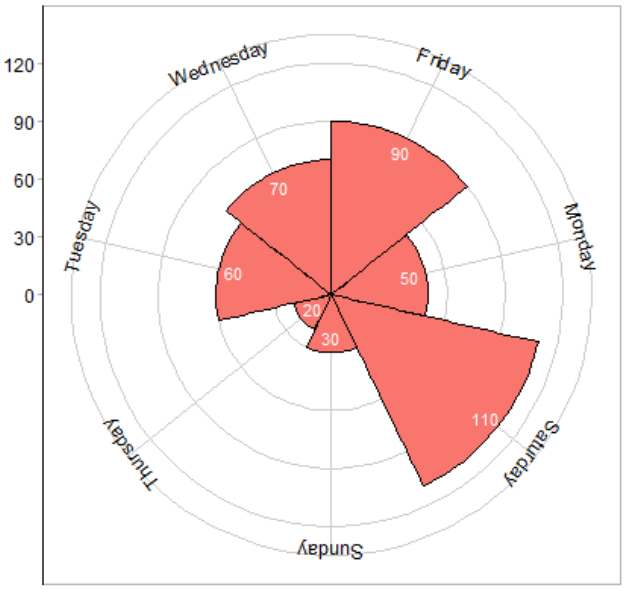
Again, 何为好的可视化?

政党	得票率
A	38
B	34
C	24
D	4



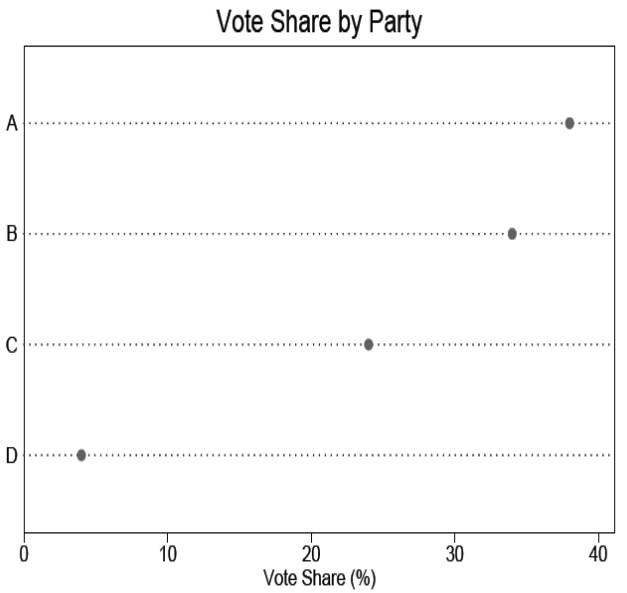
Again, 何为好的可视化?

Day	Browse
Monday	50
Tuesday	60
Wednesday	70
Thursday	20
Friday	90
Saturday	110
Sunday	30

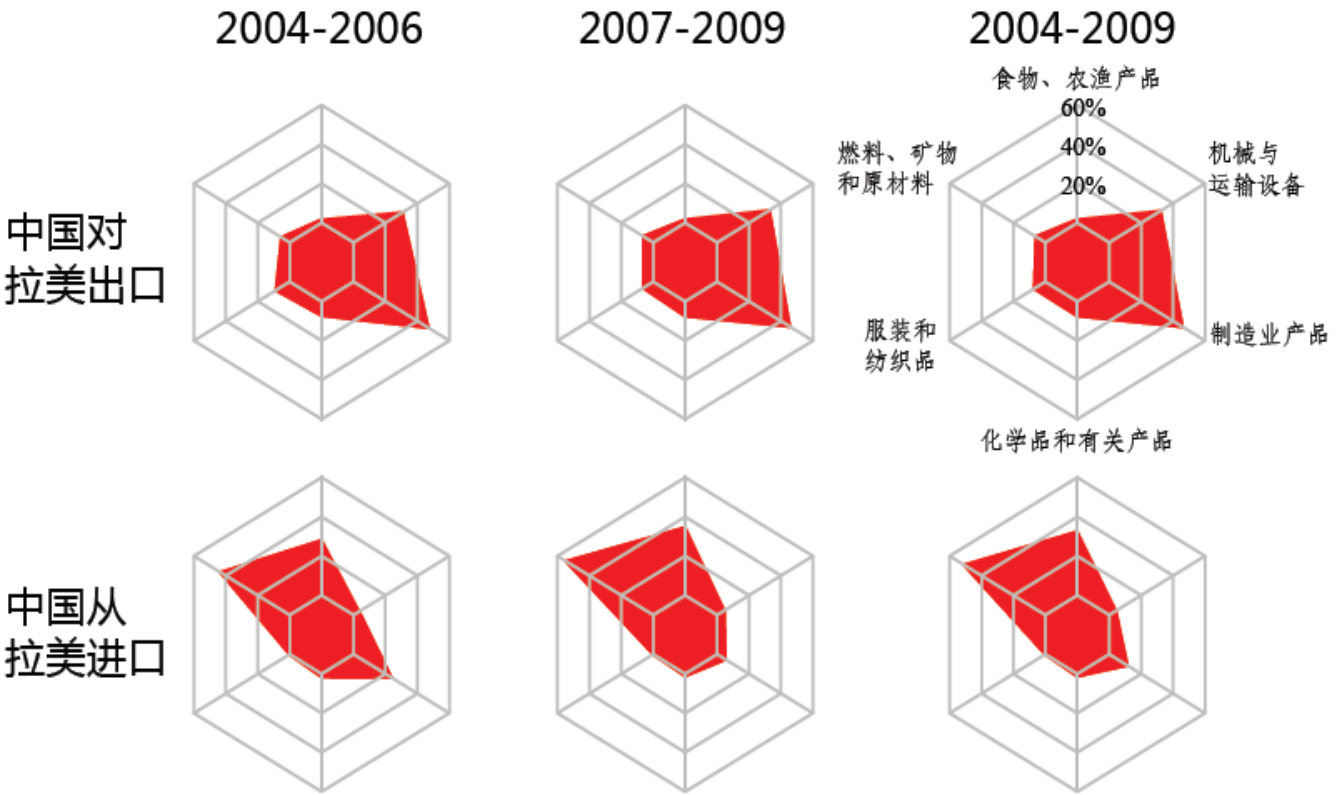


Again, 何为好的可视化?

政党	得票率
A	38
B	34
C	24
D	4

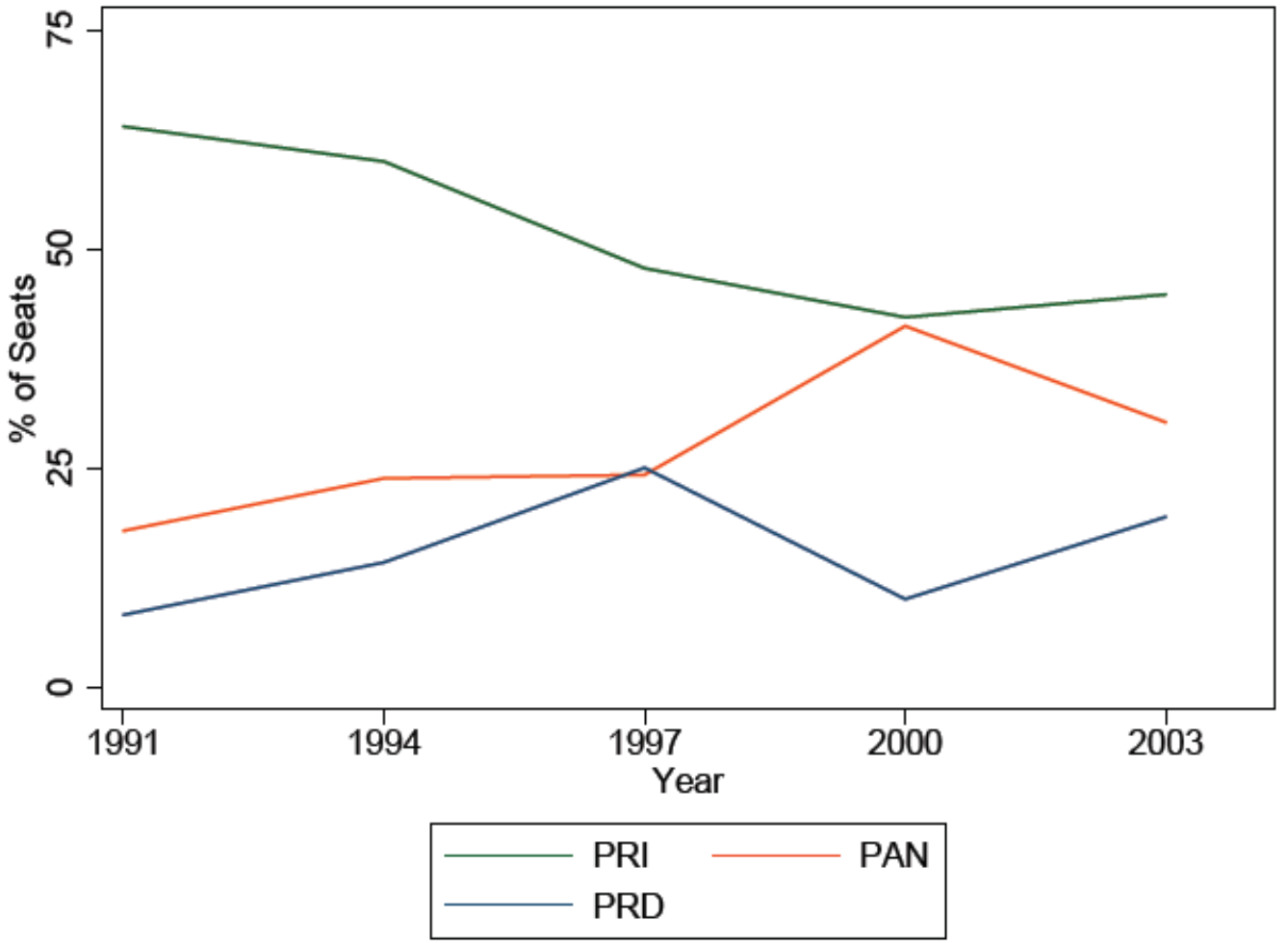


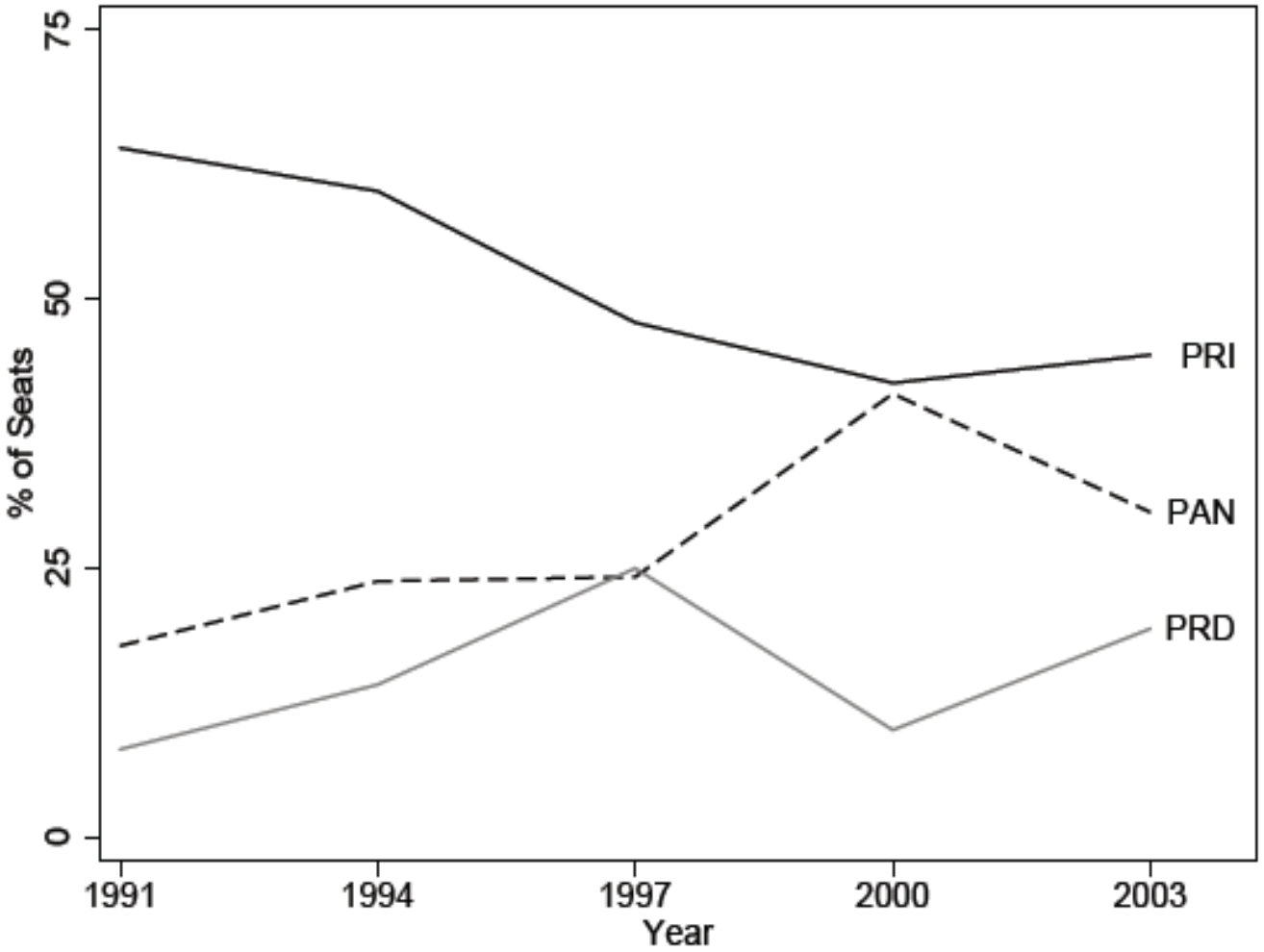
一维图之升维


2004-2006
2007-2009
2004-2009
中国对
拉美出口
中国从
拉美进口
食物、农渔产品
60%
40%
20%
机械与
运输设备
制造业产品
化学品和有关产品
服装和
纺织品
燃料、矿物
和原材料

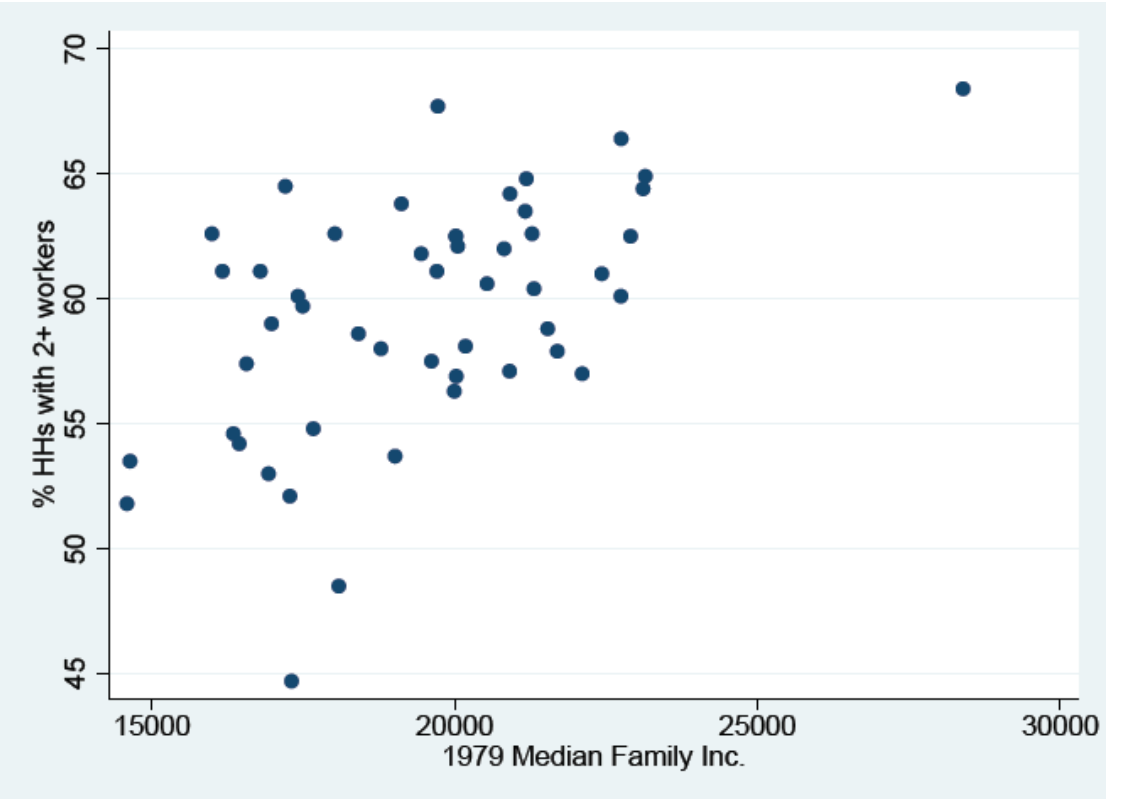
二维数据可视化

Party	PRI	PAN	PRD	Others
1991	64	17.8	8.2	10
1994	60	23.8	14.2	2
1997	47.8	24.2	25	3
2000	42.2	41.2	10	6.6
2003	44.8	30.2	19.4	5.6

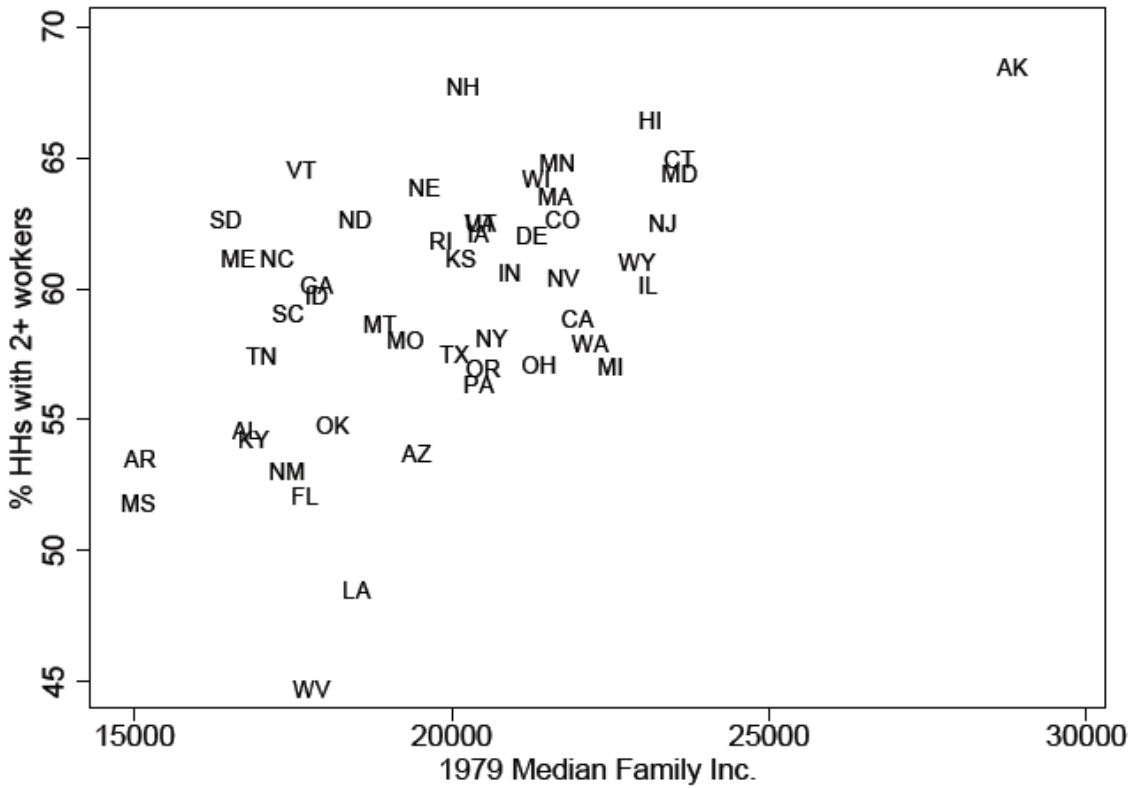

75
50
25
0
% of Seats
1991
1994
1997
2000
2003
Year
PRI
PAN
PRD


75
50
25
0
% of Seats
1991
1994
1997
2000
2003
Year
PRI
PAN
PRD

二维数据的“极简主义”


% HHs with 2+ workers
45
50
55
60
65
70
15000
20000
25000
30000
1979 Median Family Inc.

二维数据的“极简主义”


% HHs with 2+ workers
70
65
60
55
50
45
15000
20000
25000
30000
1979 Median Family Inc.
AK
NH
HI
VT
NE
MN
WI
CT
MD
MA
SD
ND
CO
NJ
RI
IA
DE
ME
NC
KS
IN
NV
WY
IL
GA
ID
SC
MT
CA
MO
NY
WA
TN
TX
OR
OH
MI
PA
AL
OK
KY
AR
AZ
NM
MS
FL
LA
WV

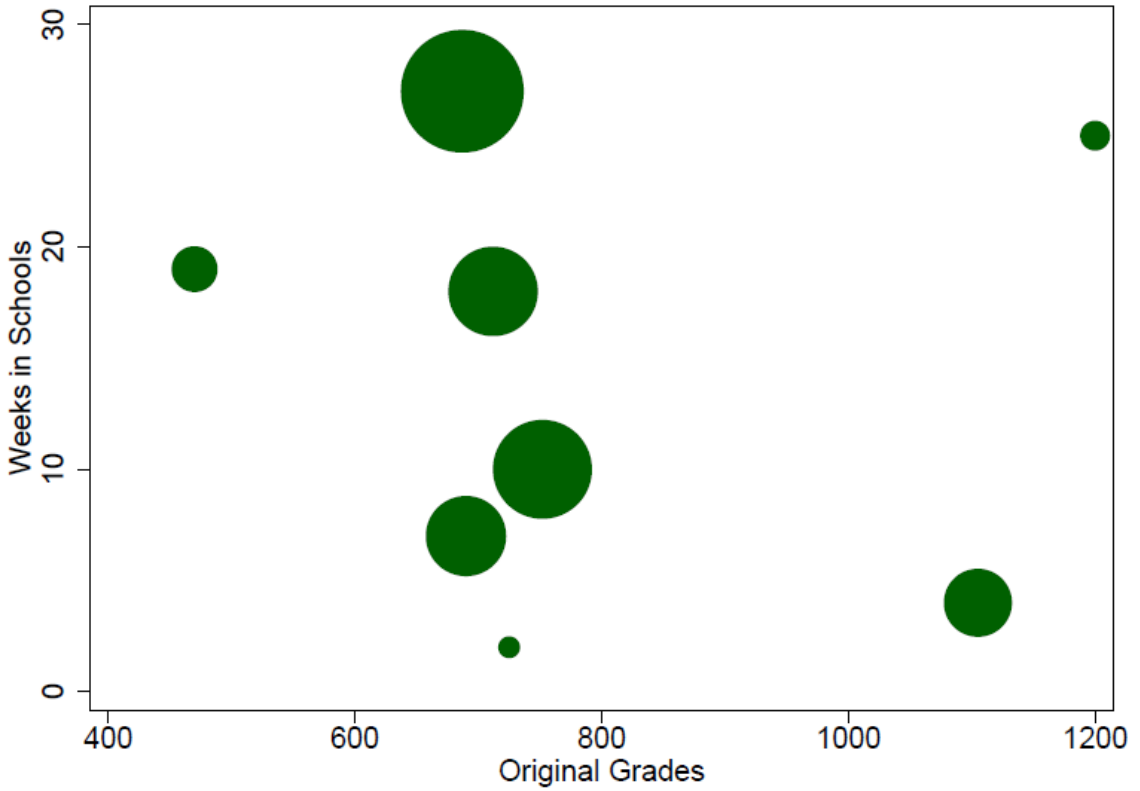
二维数据之升维


S & W
North
Total
% HHs with 2+ workers
1979 Median Family Inc.
40
50
60
70
15000
20000
25000
30000
AK
HI
MD
WY
MT
OK
AZ
LA
WV
NH
VT
ND
NE
CT
NJ
IL
MO
NY
PA
OH
MI
MS
AR
TN

要在一张图显示呢?

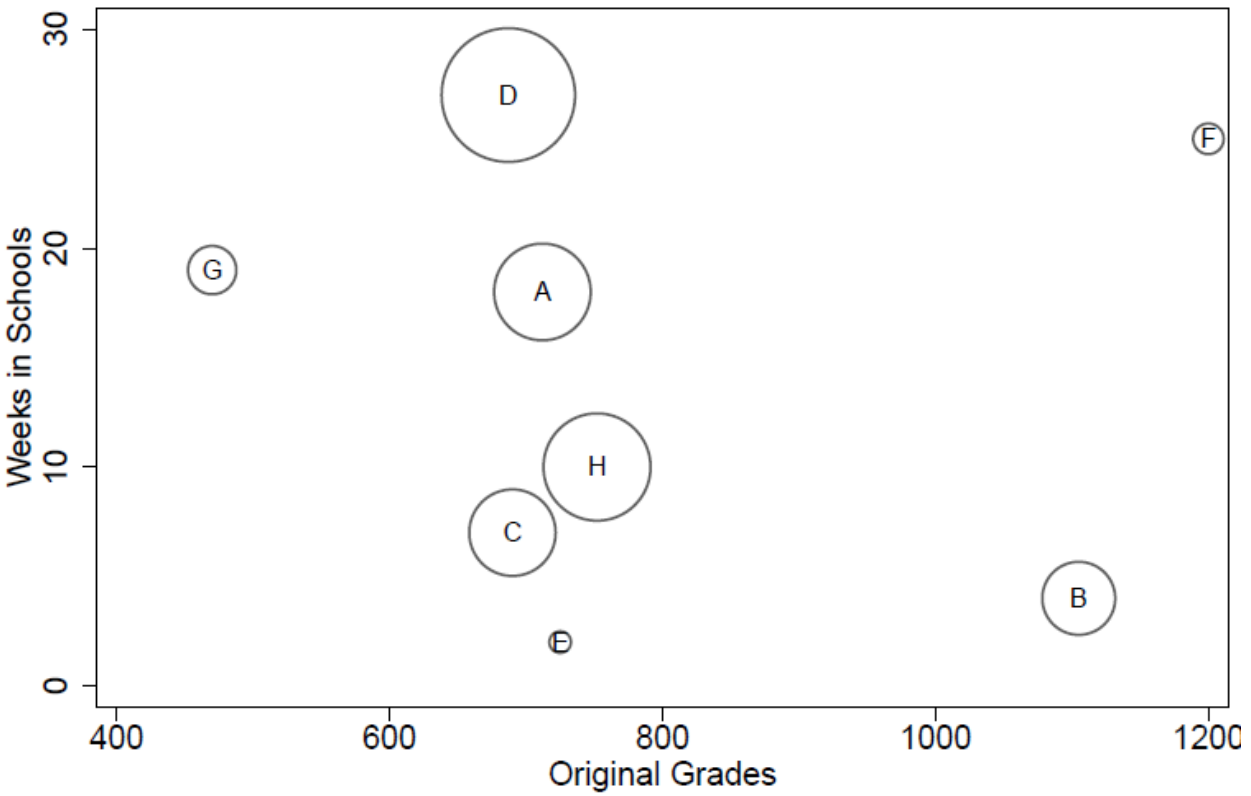
学生	原始成绩	补习班周数	进步分数
A	712	18	100
B	1105	4	57
C	690	7	80
D	687	27	191
E	725	2	5
F	1200	25	10
G	470	19	25
H	752	10	123

Bubble Chart



有什么问题?

Better Bubble Chart



Bubble Chart 优劣

Pro

- 散点图变异
- 通过气泡大小呈现第三维度数据

Con

- 气泡大小当接近时可能造成便是困难，信息无法完全精确地传达

优质可视化的创造

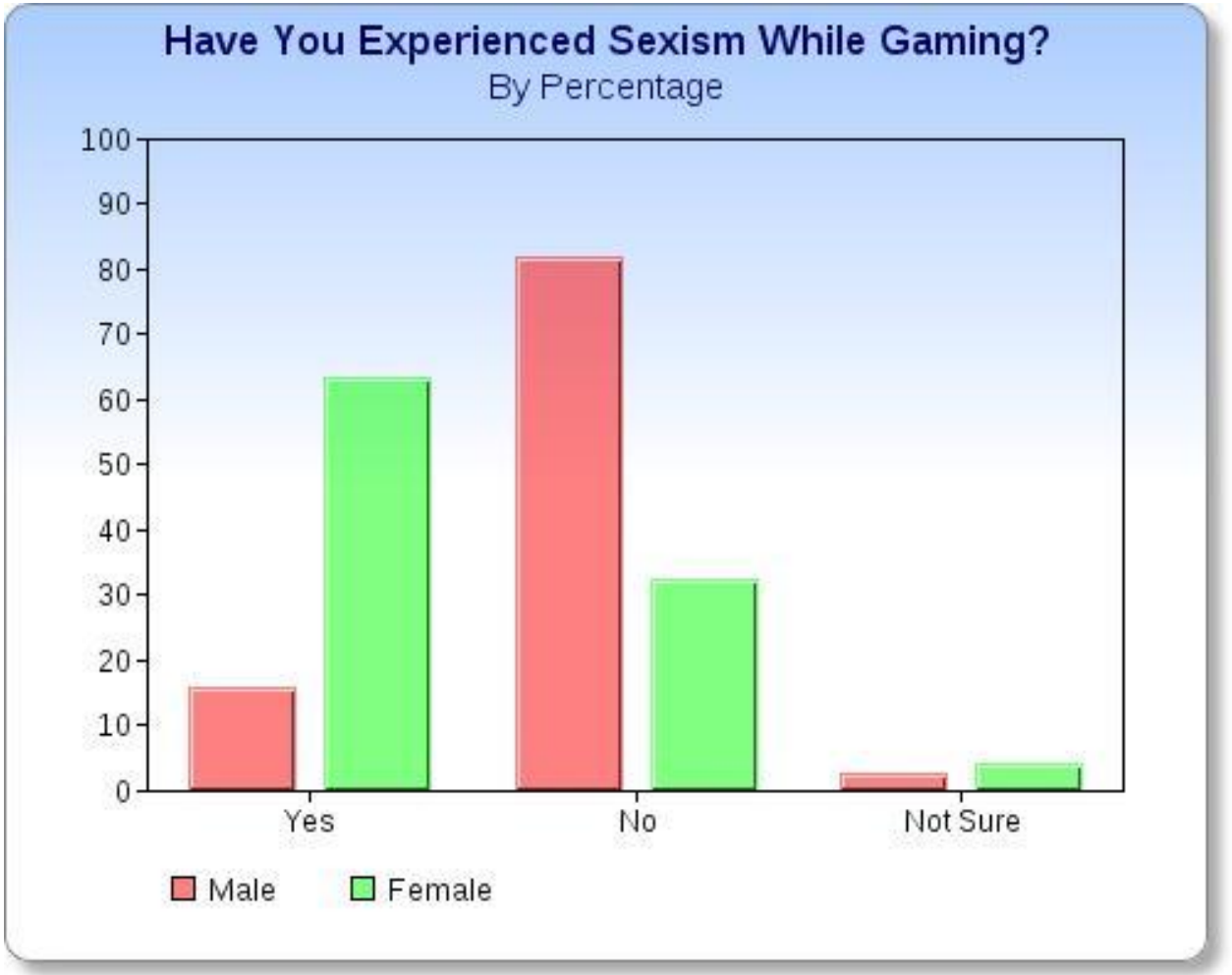
回归本质：可视化是什么？


SOCIAL
ENGINEERING
DIVING
DUMPSTER
PHISHING
TEXTING
PRETEXTING
TAILGATING
SHOULDER
WALLING

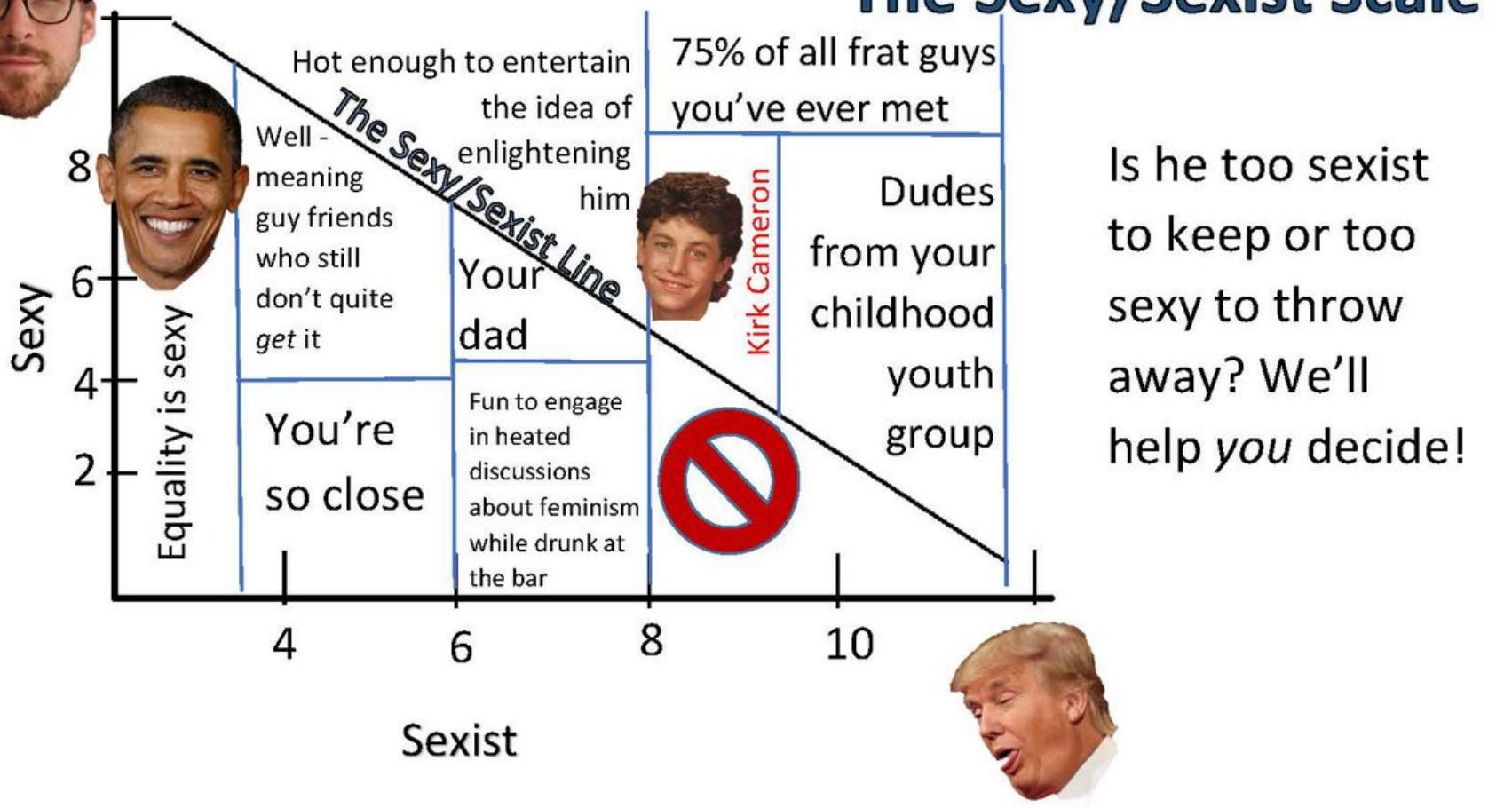
可视化的作用

- 记录信息
- 分析推理
- 证实假设
- 交流思想

“工笔”派：统计可视化

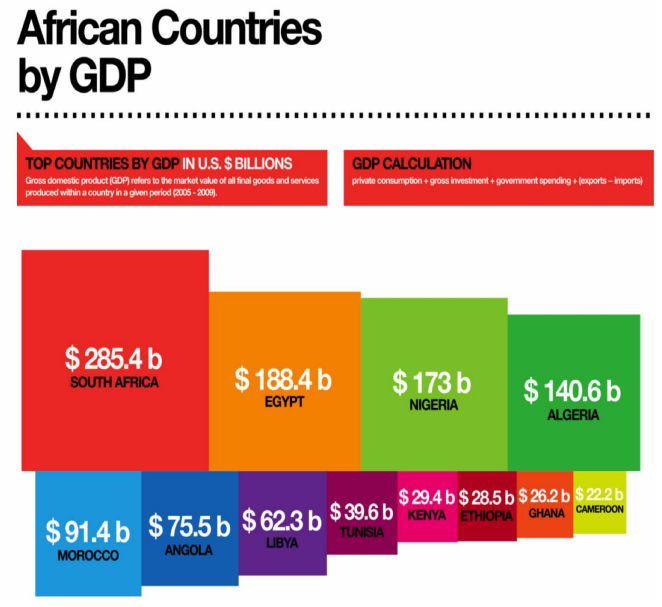

Have You Experienced Sexism While Gaming?
By Percentage
100
90
80
70
60
50
40
30
20
10
0
Yes
No
Not Sure
Male
Female

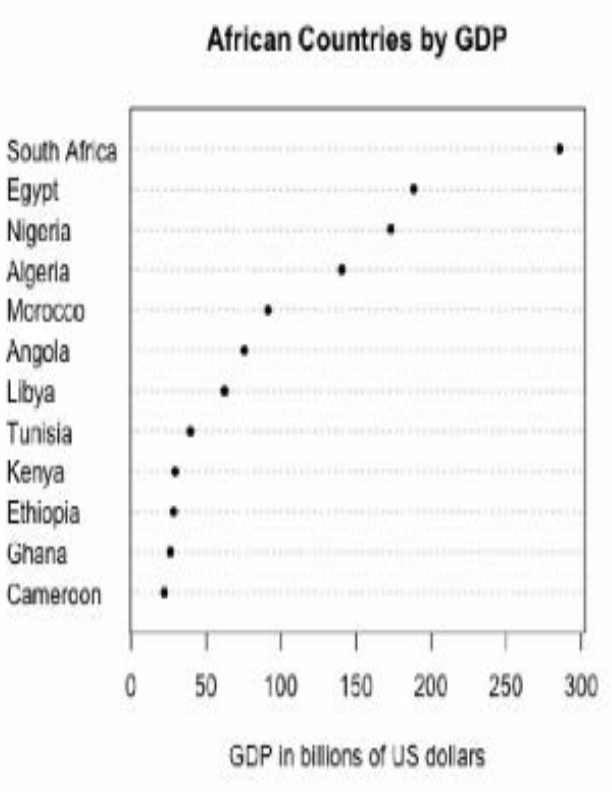
“写意”派：信息可视化



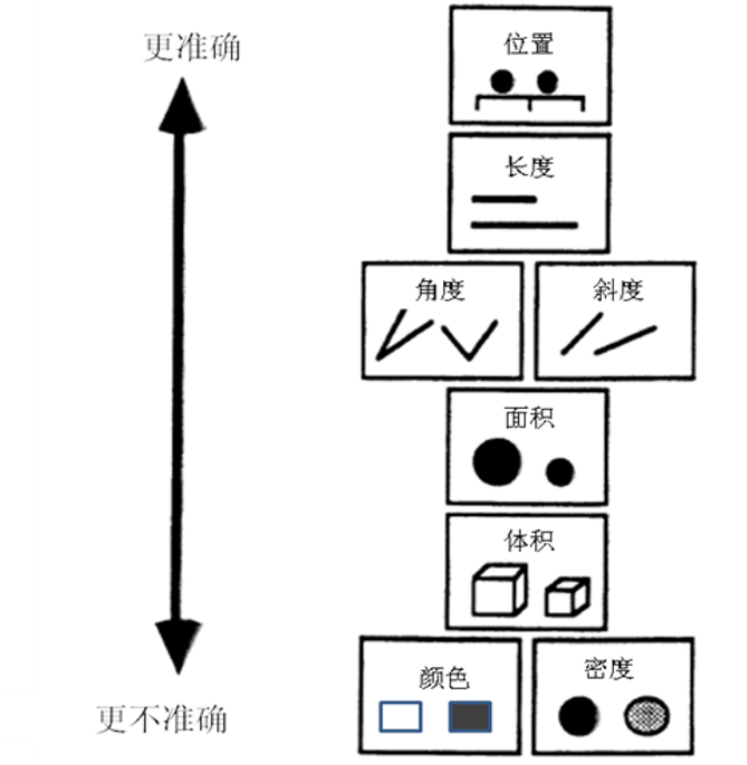
Is he too sexist to keep or too sexy to throw away? We'll help *you* decide!

数据可视化


African Countries by GDP
TOP COUNTRIES BY GDP IN U.S. $ BILLIONS
Gross domestic product (GDP) refers to the market value of all final goods and services produced within a country in a given period (2005 - 2009).
GDP CALCULATION
private consumption + gross investment + government spending + (exports – imports)
$ 285.4 b
SOUTH AFRICA
$ 188.4 b
EGYPT
$ 173 b
NIGERIA
$ 140.6 b
ALGERIA
$ 91.4 b
MOROCCO
$ 75.5 b
ANGOLA
$ 62.3 b
LIBYA
$ 39.6 b
TUNISIA
$ 29.4 b
KENYA
$ 28.5 b
ETHIOPIA
$ 26.2 b
GHANA
$ 22.2 b
CAMEROON


African Countries by GDP
South Africa
Egypt
Nigeria
Algeria
Morocco
Angola
Libya
Tunisia
Kenya
Ethiopia
Ghana
Cameroon
0
50
100
150
200
250
300
GDP in billions of US dollars

可视化编码



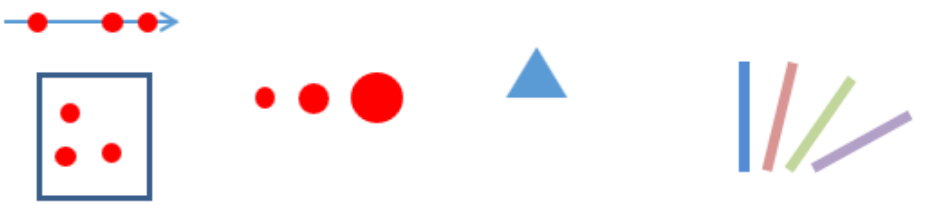
编码类型

- 数值（continuous）
  - 10厘米，17厘米，23厘米
- 定序（ordinal）
  - 小，中，大
  - 周日，周一，周二……
- 定类（nominal）
  - 苹果，桔子，香蕉……

编码要素

标记：点、线、面

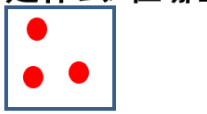
通道：位置、大小、形状、方向、色调、饱和度、亮度……



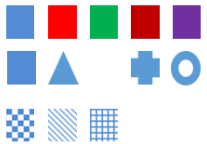
通道

分类的

是什么/在哪里



位置



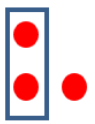
色调

形状

图案

分组的

关系



包含

连接

相似

接近

定量/定序的

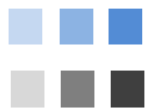
程度

坐标轴位置

长度

角度

面积



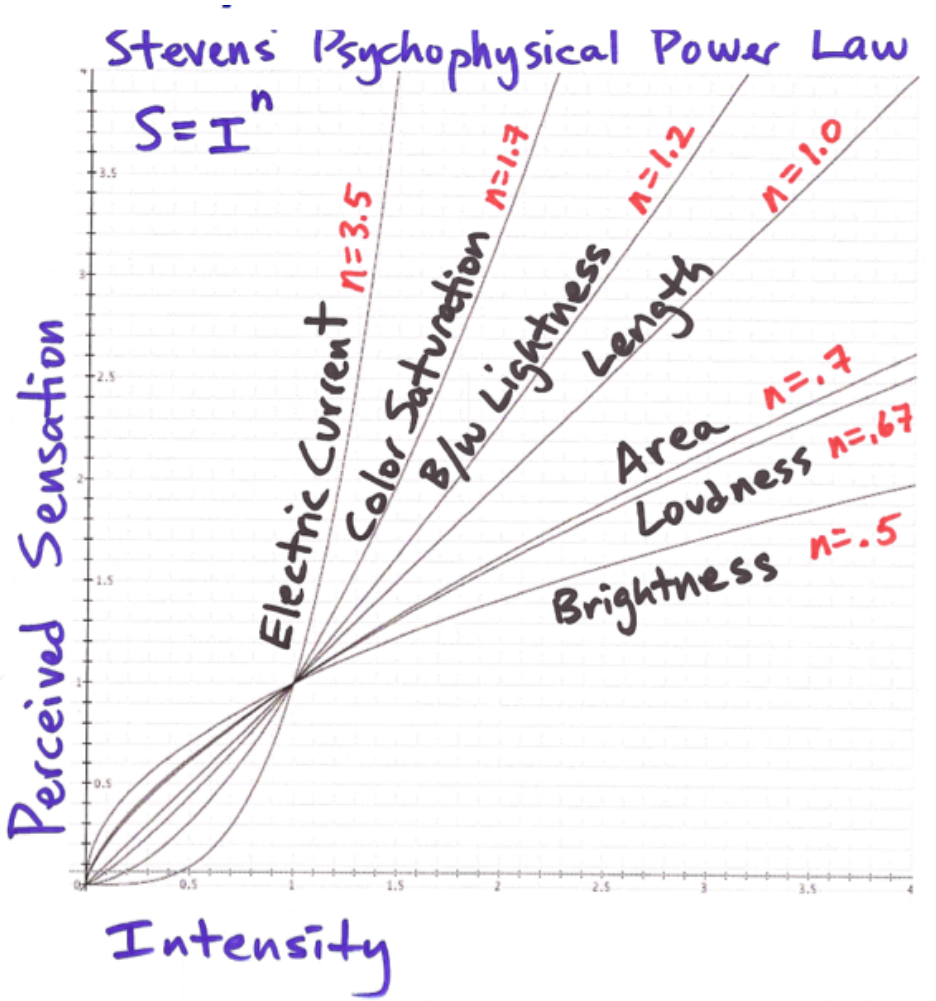
亮度/饱和度

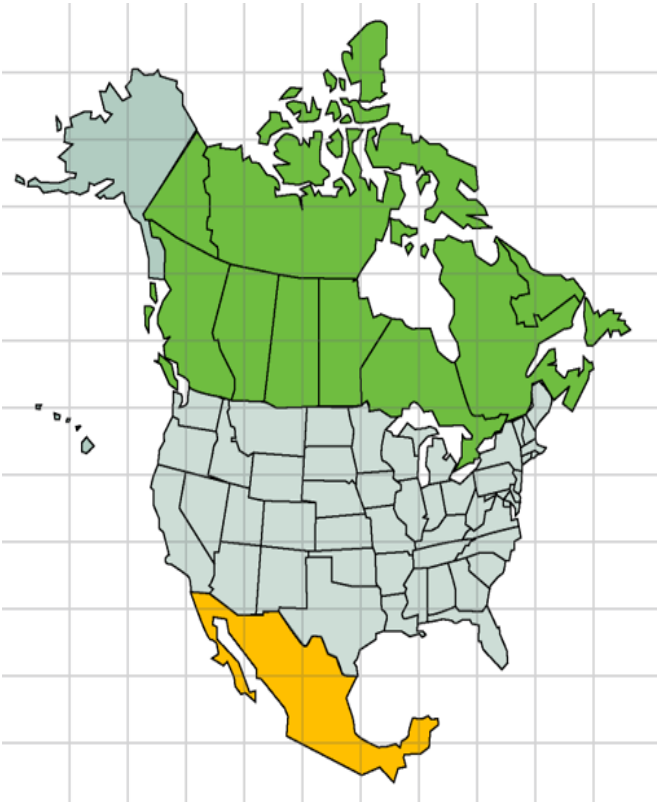
图案密度

如何选择好的通道

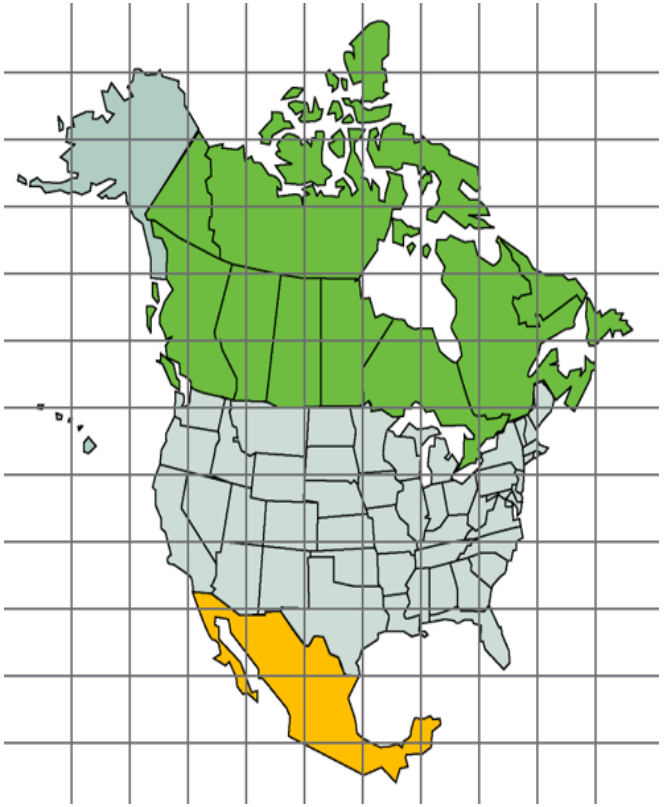
- 类型（type）
  - 是什么/在哪里（what/where）
  - 何种程度（how much）
- 有效性（effectiveness）
  - 通道表现力符合属性的重要性
- 表现力（expressiveness）
  - 表达且仅表达数据的完整属性
  - 判断标准（ADSP）：精确性、可辨性、可分离性、视觉突出

精确性（Accuracy）



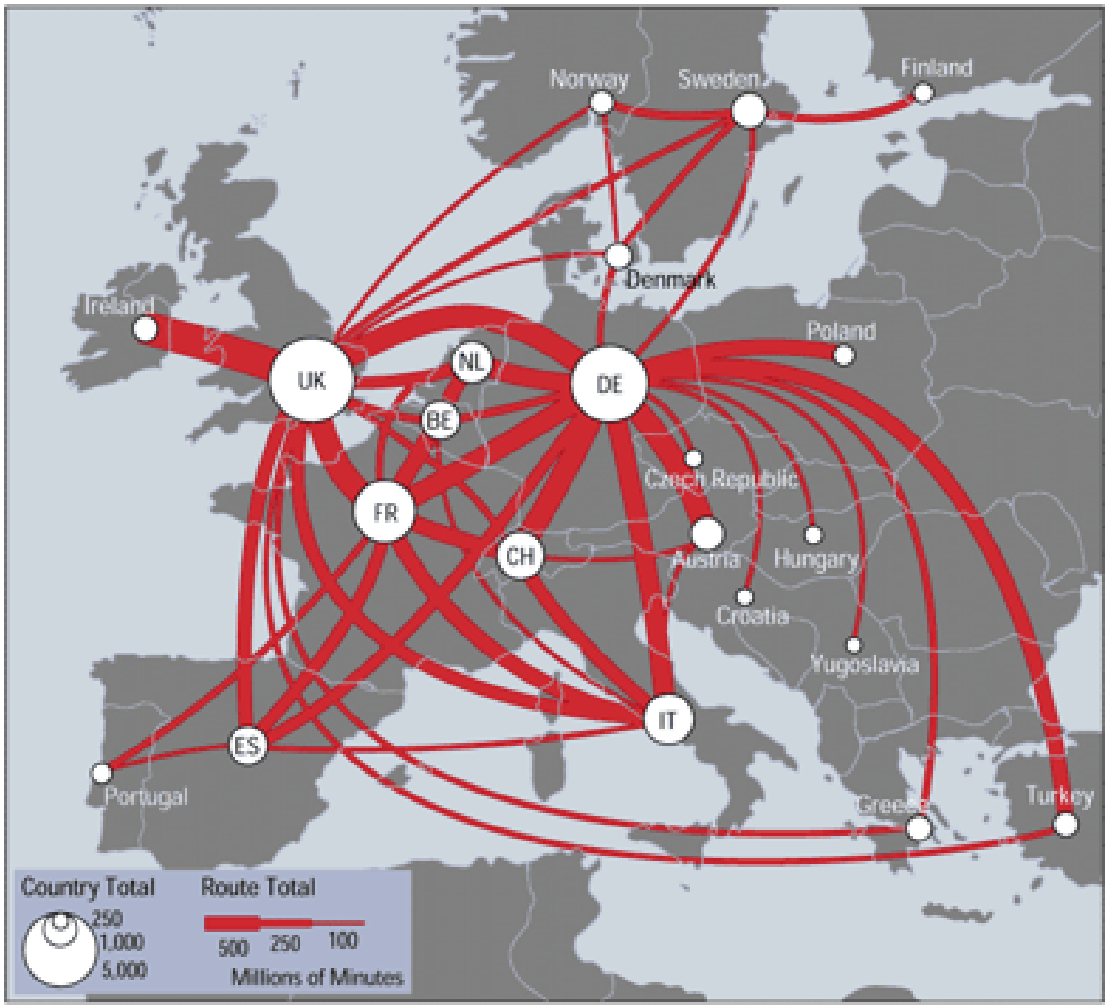


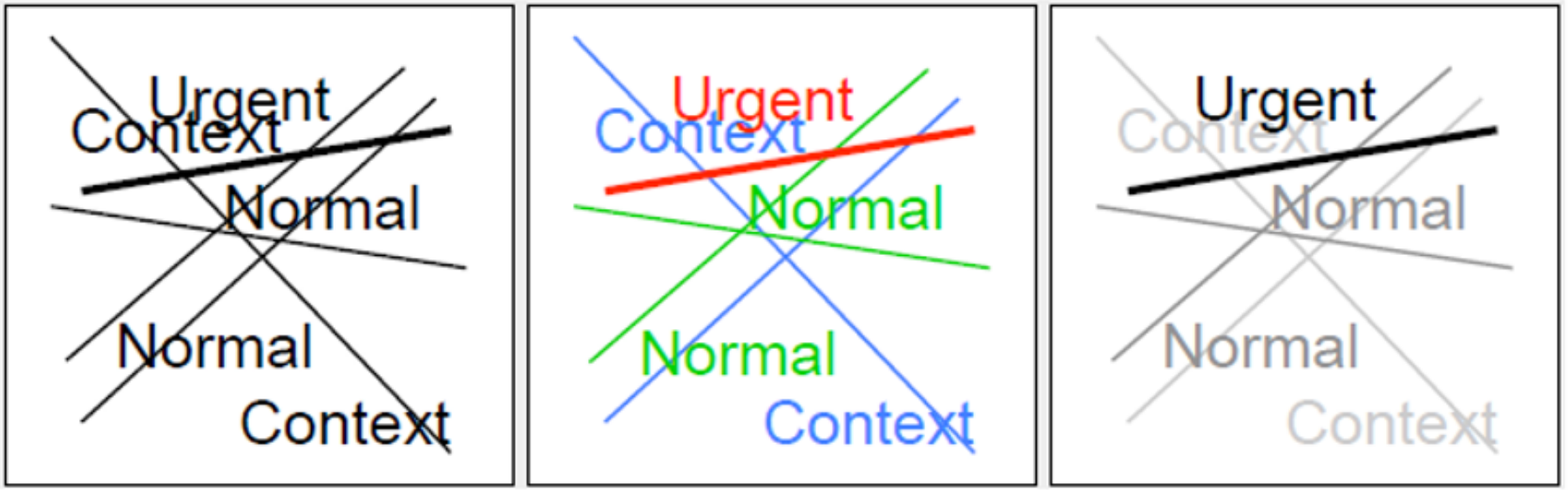
网格处于背景,会不凸显



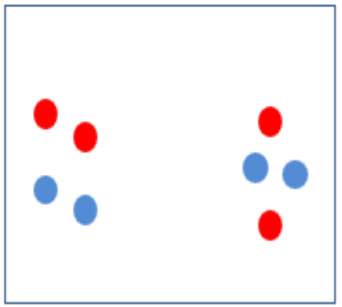
网格处于前景,遮挡地图

可辨性（Discriminability）



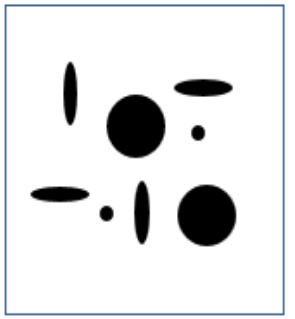

Urgent
Context
Normal
Normal
Context
Urgent
Context
Normal
Normal
Context
Urgent
Context
Normal
Normal
Context

可分离性（Separability）



位置/色调

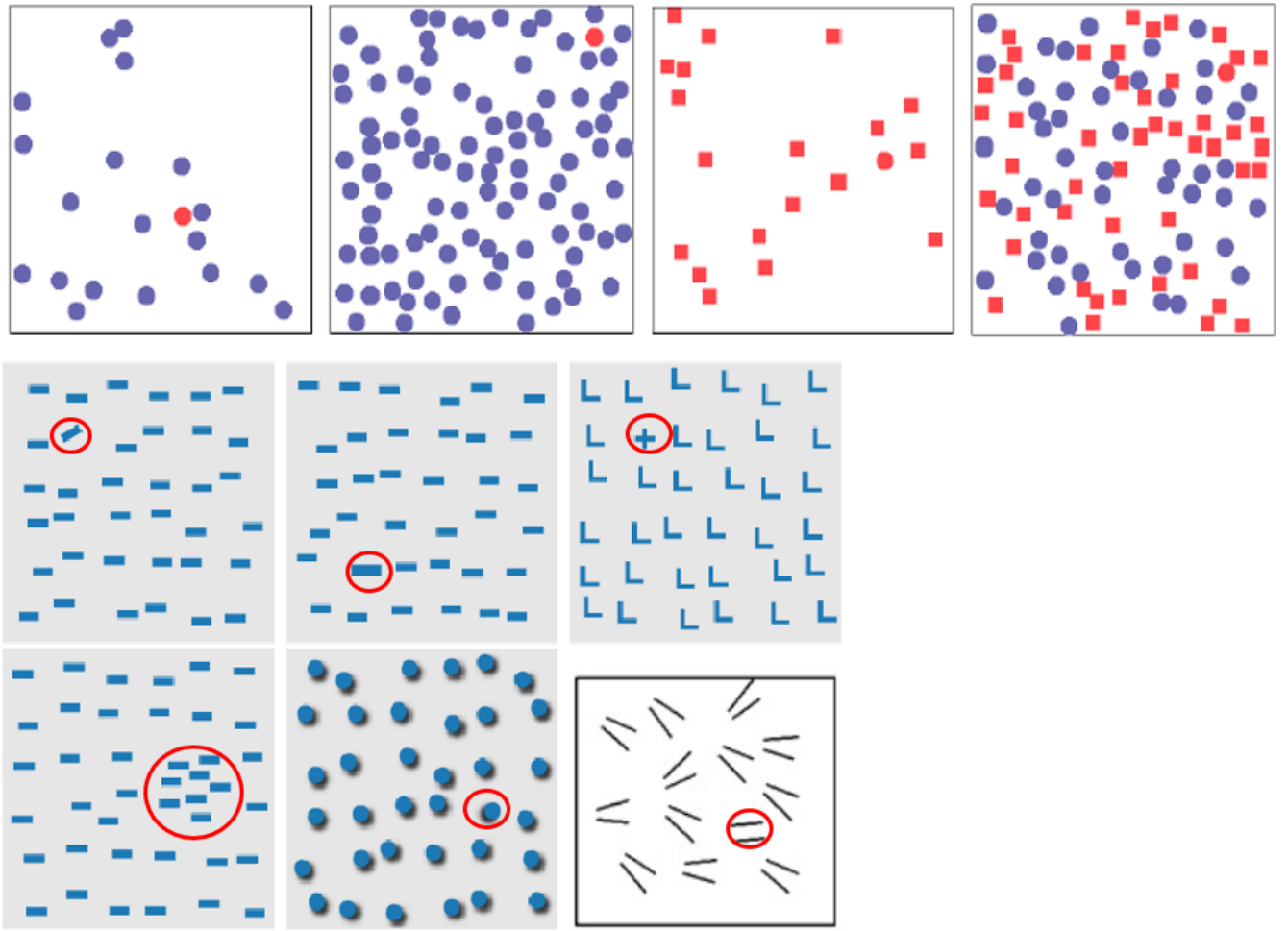
尺寸/色调

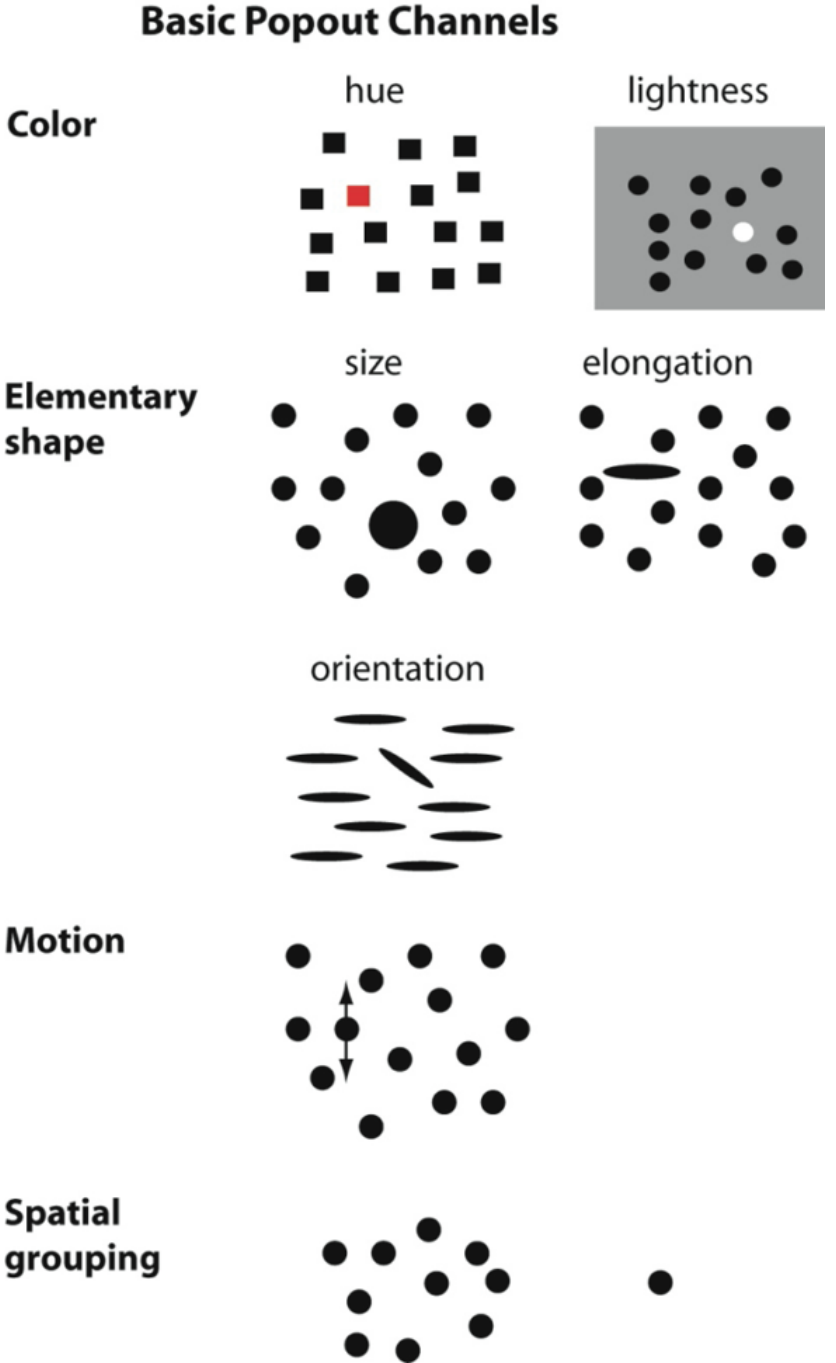


宽度/高度

红/绿

视觉突出（Pop-out）




Basic Popout Channels
Color
hue
lightness
Elementary shape
size
elongation
orientation
Motion
Spatial grouping

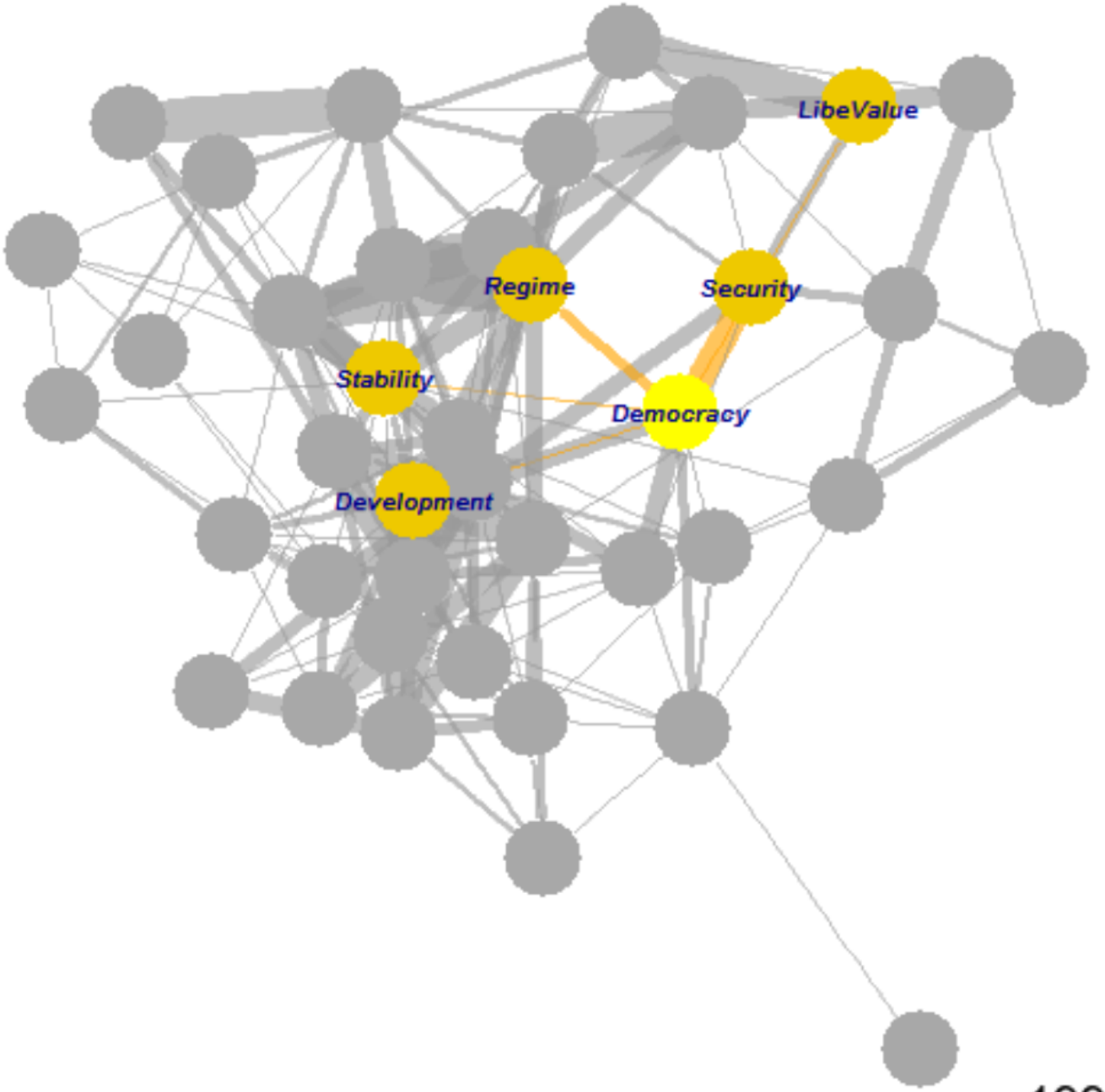
SDFLDKJFSACNMNMVLKJBHUITRUTIOT
PWEPRTPWEHDFBMDZLVNASIKGFDKGN
JKBFDNBKITEWYUTUWPOQRQWPTPYJU
KLBNBZLKVLKJHGSDFPOIQYHERPWOE
NCZLLXHGRSWOGTWPEOTRPWEPWKG
FGCBBQOOPWQIRYTIQOWKXNZCBAKP
FOETIRTUOPEQIOTHGSLQWPOGHWBN


SDFLDKJFSACNMNMVLKJBHUITRUTIOT
PWEPRTPWEHDFBMDZLVNASIKGFDKGN
JKBFDNBKITEWYUTUWPOQRQWPTPYJU
KLBNBZLKVLKJHGSDFPOIQYHERPWOE
NCZLLXHGRSWOGTWPEOTRPWEPWKG
FGCBBQOOPWQIRYTIQOWKXNZCBAKP
FOETIRTUOPEQIOTHGSLQWPOGHWBN

SDFLDKJFSACNMNMVLKJBHUITRUTIOT
PWEPRTPWEHDFBMDZLVNASIKGFDKGN
JKBFDNBKITEWYUTUWPOQRQWPTPYJU
KLBNBZLKVLKJHGSDFPOIQYHERPWOE
NCZLLXHGRSWOGTWPEOTRPWEPWKG
FGCBBQOOPWQIRYTIQOWKXNZCBAKP
FOETIRTUOPEQIOTHGSLQWPOGHWBN

“看家功夫”：分组与分层

- 分组会对大部分任务有效
- 如果不能进行分组，则需要转换任务目标以支持分组
- 避免过多类别
- 对数据的每个维度指定一种或一个阶层的颜色


LibeValue
Regime
Security
Stability
Democracy
Development

色彩信息

- 灰度值可被认为是有序的

- 可用于编码数值型数据

- 色调通常认为是无序的，可用于编码不同维度的值

色彩意义：一种主观感受

颜色	电影观众	财务经理	医护人员	控制工程师
	主观认知			
蓝	温柔	合作、可靠	死亡	冷、水
青	悠闲	冷静、沉着	缺氧	蒸汽
绿	好玩的	有利益的	感染、肝胆疾病	正常、安全
黄	高兴	重要的	黄胆病	警告
红	兴奋	无利益的	健康	危险
洋红	悲伤	富有的	需要观察	热、辐射

色彩设计基本原则

- 使用有限的色调范围
  - 控制低饱和度色彩中的色彩视觉突出
  - 避免过多颜色交错导致的杂乱无章

- 使用中性背景色
  - 控制对全局色彩的影响
  - 最小化“同时对比”（simultaneous contrast）

对于新手而言

辅助软件来选择合适的色阶：

http://colorbrewer2.org/

http://paletton.com/

To Be continued...

Previously in *Learning Visualization with Dr. Hu*

- 鉴赏
  - 表 vs. 图
  - No-no: 饼状图、面积图
  - Alternative: 折线图、散点图、雷达图……
  - BTW, 词云（Hem...）
- 建构
  - 可视化编码：标记、通道（ADSP）
  - 分组与分层
  - 色彩挑选

ggplot可视化举例

RColorBrewer Package

主要code

```
colorRampPalette(brewer.pal(<num.>, <name of pallette>))(n)
```

Alternative

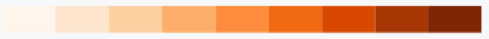
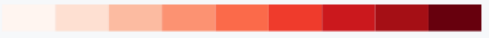
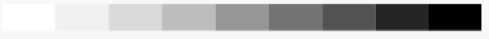
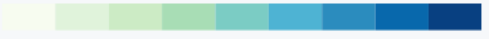
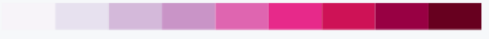
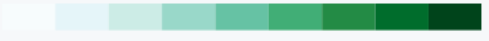
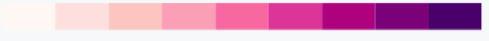
ggsci: 包含Nature、Science等期刊配色（Check "Tron Legacy"）

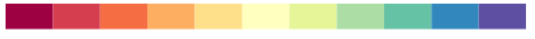
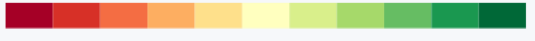
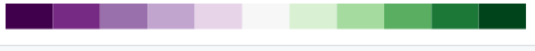
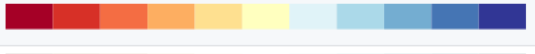
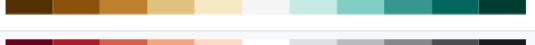
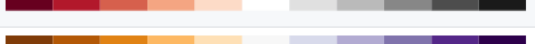
wesanderson

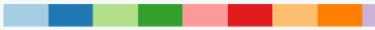
ggthemes

ggtech

Name of Palette

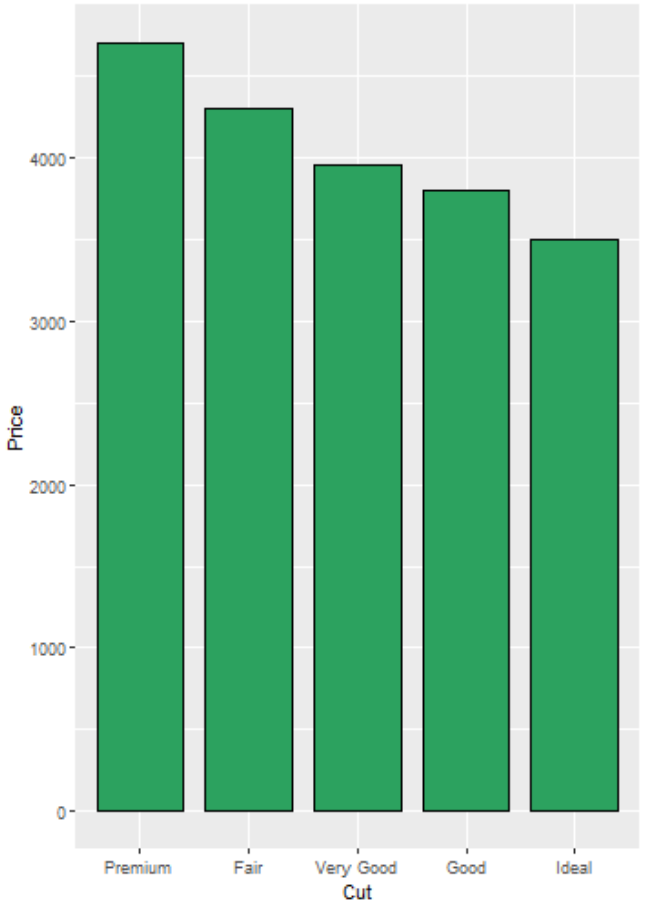
Name	Example
Blues	
Oranges	
Greens	
Reds	
Purples	
Greys	
OrRd	
GnBu	
PuBu	
PuRd	
BuPu	
BuGn	
YlGn	
RdPu	
YlOrBr	
YlGnBu	
YlOrRd	
PuBuGn	

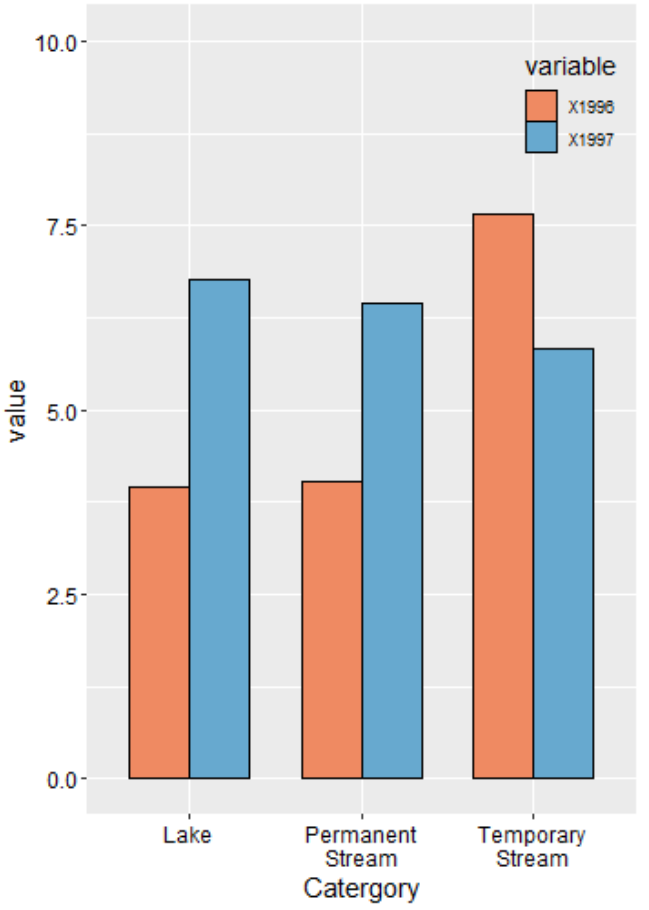
Name	Example
Spectral	
RdYlGn	
RdBu	
PiYG	
PRGn	
RdYlBu	
BrBG	
RdGy	
PuOr	

Name	Example
Set1	
Set2	
Set3	
Accent	
Dark2	
Paired	
Pastel1	
Pastel2	

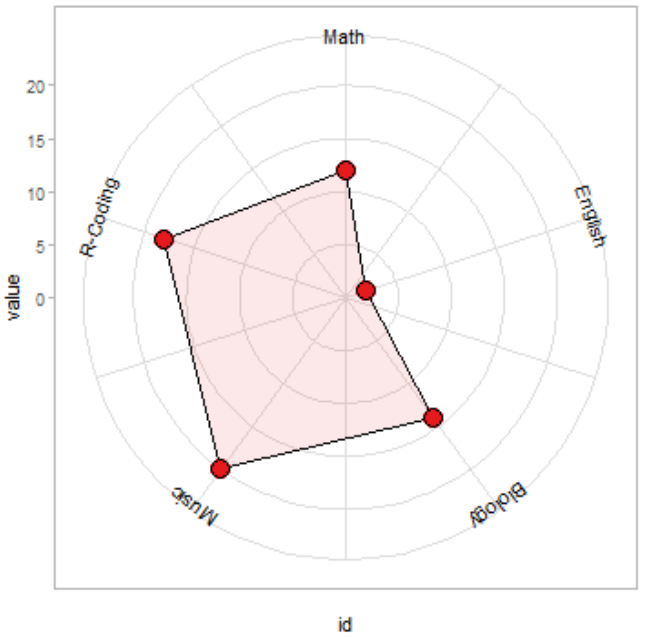
一维图

```
ggplot(data = mydata, aes(Cut, Price)) +
  geom_bar(
    stat = "identity",
    width = 0.8,
    colour = "black",
    size = 0.25,
    fill = "#2ca25f",
    alpha = 1
  )

ggplot(data = mydata2, aes(Catergory, value,
                           fill = variable)) +
  geom_bar(
    stat = "identity",
    color = "black",
    position = position_dodge(),
```


Price
4000
3000
2000
1000
0
Premium
Fair
Very Good
Good
Ideal
Cut


value
10.0
7.5
5.0
2.5
0.0
variable
X1996
X1997
Lake
Permanent Stream
Temporary Stream
Catergory

car	id	value
Math	1	12
English	2	2
Biology	3	14
Music	4	20
R-Coding	5	18
NA	6	12



```
ggplot() +
  geom_polygon(data=mydata,aes(x=id, y=value),color = "black"
  geom_point(data=mydata,aes(x=id, y=value),size=5,shape=21,c
  coord_radar() +
  scale_x_continuous(breaks =label_data$id,labels=label_data$
  ylim(0,22)+
  theme_light()+
  theme(axis.text.x=element_text(size = 11,colour="black",ang
```

二维图

```
ggplot(data = mydata, aes(x, y))
  geom_point(
    fill = "black",
    colour = "black",
    size = 3,
    shape = 21
  ) +
  geom_smooth(
    method = 'loess',
    span = 0.4,
    se = TRUE,
    colour = "#00A5FF",
    fill = "#00A5FF",
    alpha = 0.2
  ) + scale_y_continuous(breaks =
  theme(
```

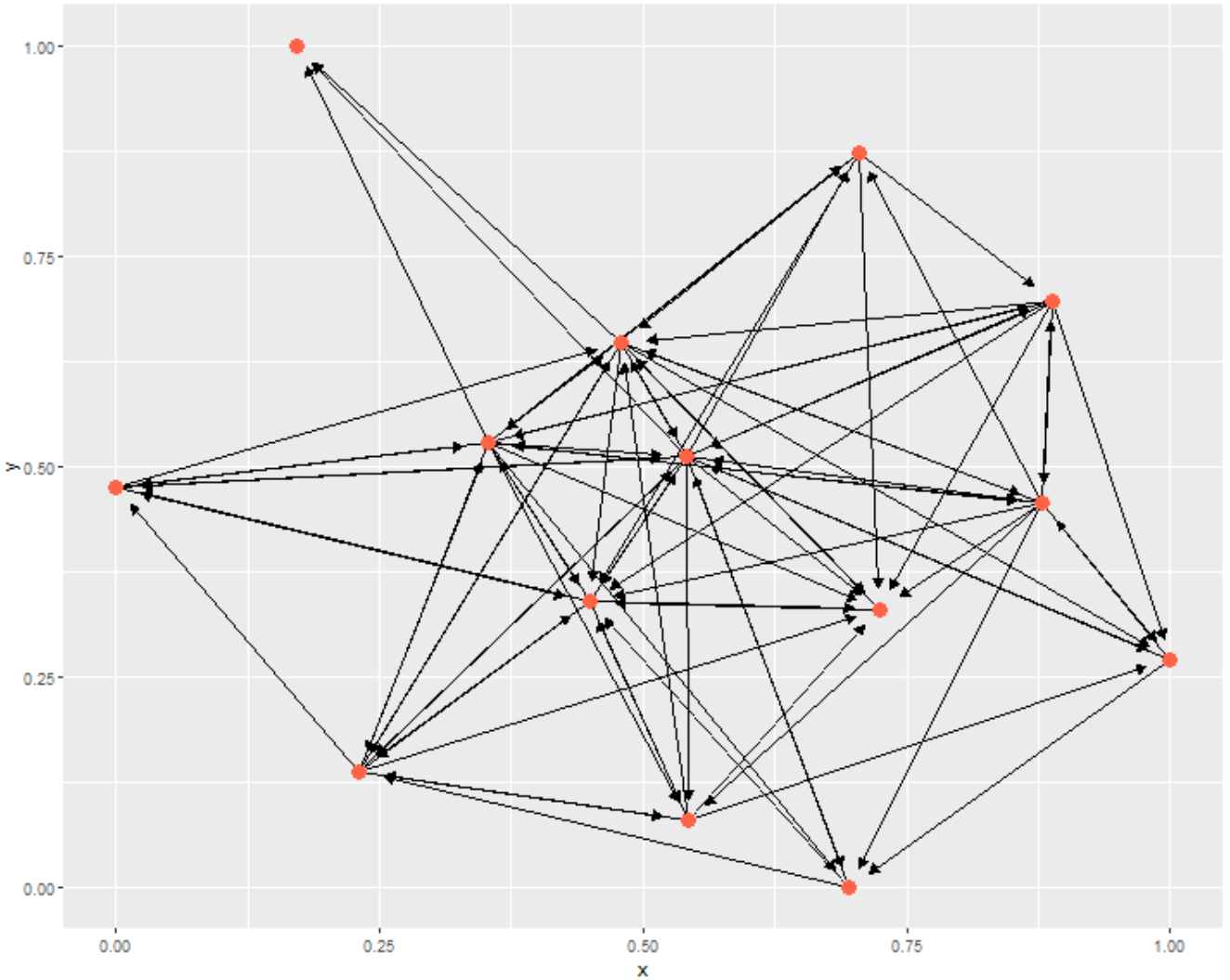
统计图

	Model 1
(Intercept)	41.108***
	(2.842)
cyl	-1.785***
	(0.607)
disp	0.007
	(0.012)
wt	-3.636***
	(1.040)
Num.Obs.	32
R2	0.833
Adj.R2	0.815
* p < 0.1, ** p < 0.05, *** p < 0.01	

```
library(dotwhisker)
dwplot(m1)
```

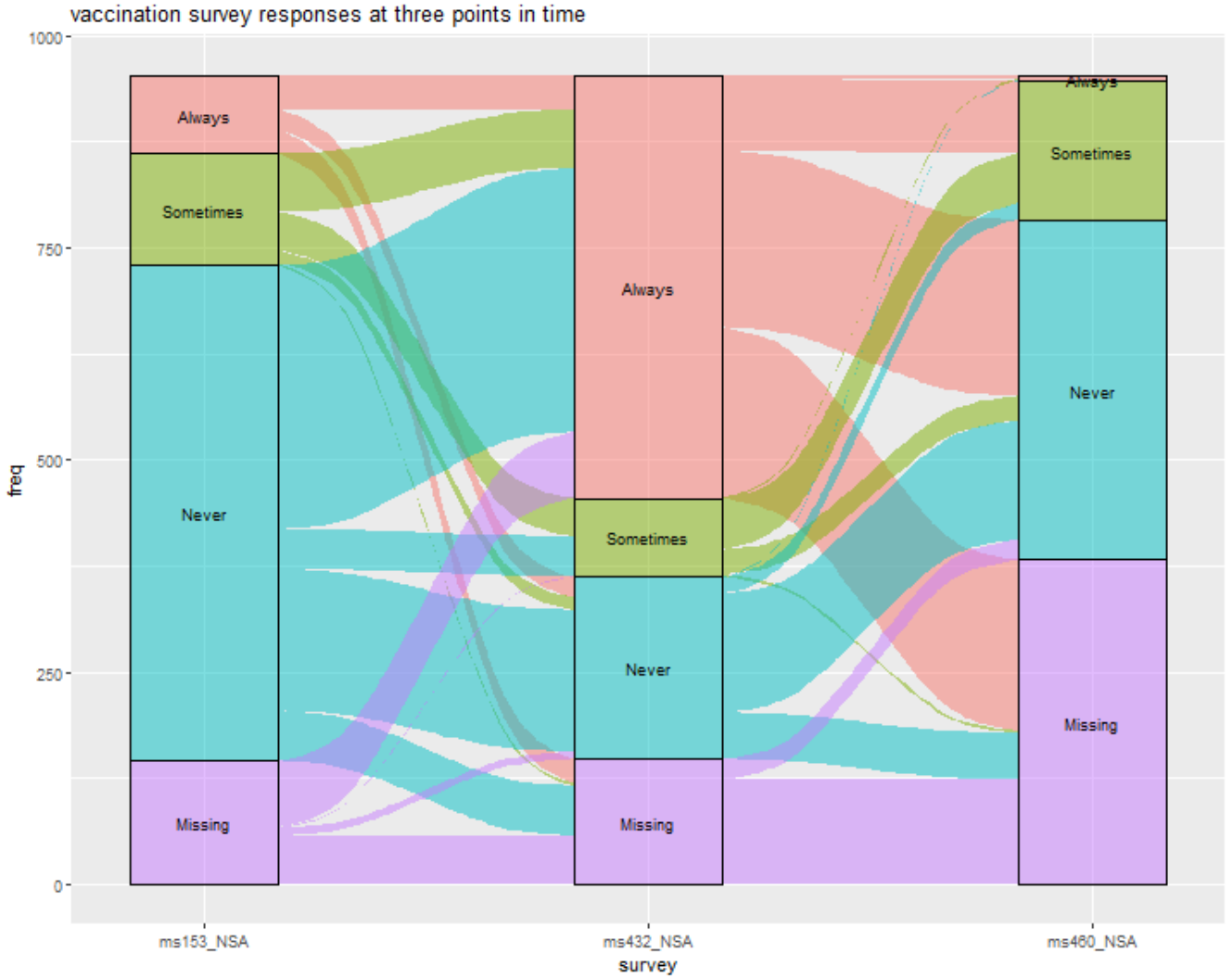
网络图 (Complete Network)

```
library(network) # for data
data(emon)
`library(ggnetwork)`

ggplot(emon[[1]], aes(
  x = x,
  y = y,
  xend = xend,
  yend = yend
)) +
  geom_edges(arrow = arrow(length = unit(6, "pt"), type = "closed")) +
  geom_nodes(color = "tomato", size = 4)
```


1.00
0.75
0.50
0.25
0.00
y
0.00
0.25
0.50
0.75
1.00
x

网络图 (Sankey/Alluvial)

```
`library(ggalluvial)`
data(vaccinations)
levels(vaccinations$response) <- rev(levels(vaccinations$resp

ggplot(vaccinations,
       aes(x = survey, stratum = response, alluvium = subject
           y = freq,
           fill = response, label = response)) +
  scale_x_discrete(expand = c(.1, .1)) +
  geom_flow() +
  geom_stratum(alpha = .5) +
  geom_text(stat = "stratum", size = 3) +
  theme(legend.position = "none") +
  ggtitle("vaccination survey responses at three points in ti
```


vaccination survey responses at three points in time
1000
750
500
250
0
freq
Always
Sometimes
Never
Missing
Always
Sometimes
Never
Missing
Always
Sometimes
Never
Missing
ms153_NSA
ms432_NSA
ms460_NSA
survey

地图

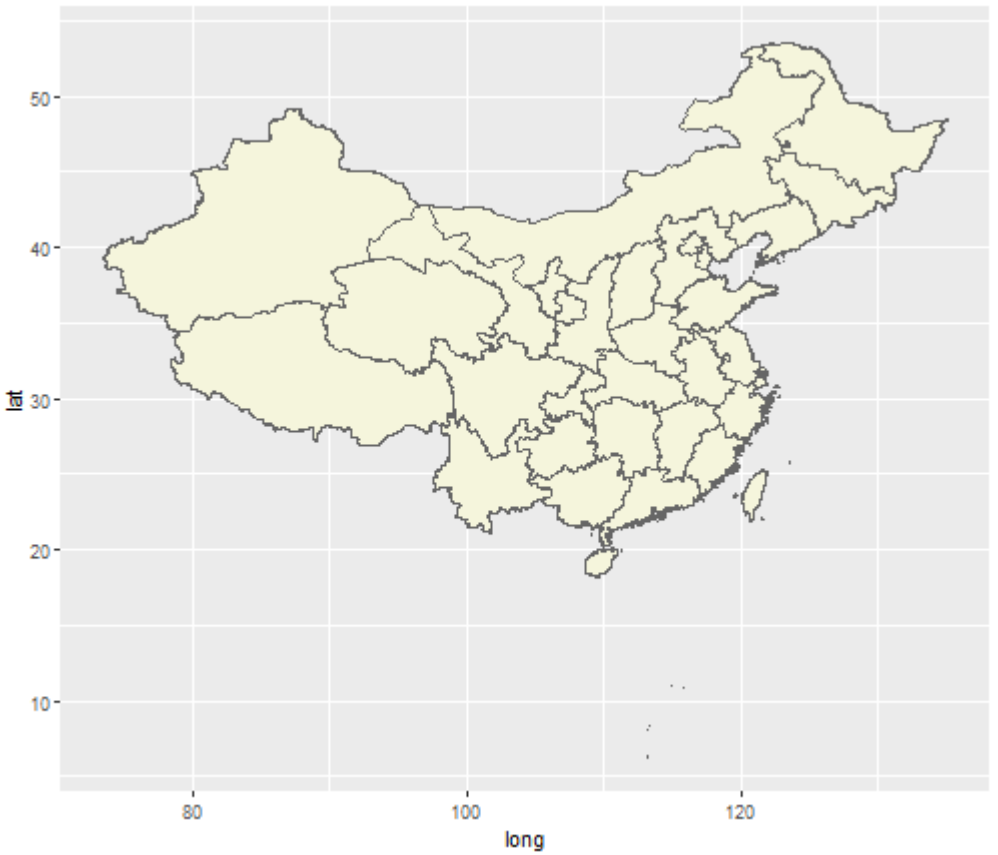
你需要三个ArcGIS文件:

bou2_4p.shp: 多边形要素图形文件

bou2_4p.dbf: 要素属性信息

bou2_4p.shx: 多边形要素索引文件

```
china_map1 <- rgdal::readOGR("data/bou2_4p.shp")

ggplot(china_map1, aes(x = long, y = lat, group = group)) +
    geom_polygon(fill = 'beige') +
  geom_path(color = "grey40")
```



gganimate


Mongolia
China
North Korea
South Korea
Yellow Sea
East China Sea
Shanghai
Beijing
Seoul
Taipei
Taiwan
Hong Kong
Nepal
Bhutan
Bangladesh
India
Myanmar
(Burma)
Laos
Thailand
Vietnam
Cambodia
Bangkok
Bay of Bengal
Andaman Sea
South
China Sea
Luzon
Philippines
Google
Map data ©2016 Google, SK telecom, ZENRIN
60
40
30
20
80
90
100
110
120
130

Take-Home Points

- 好的图形必须进行比较分析。
- 图形不仅可以呈现原始数据，也可以呈现推论结果。
- 善用数据可视化，方能成为沟通研究者与读者的工具。


. . . USE THE POWER WISELY.

Thank you!

yuehu@tsinghua.edu.cn

https://sammo3182.github.io/

sammo3182