经济数据解读 02（丽湖校区）

第二次线下课

黄嘉平

中国经济特区研究中心

huangjp.com

2024.10.17

小组汇报

- **汇报内容**：**围绕一个经济、金融、民生或其他社会领域的热门话题，获取与其紧密相关的变量的历史数据，结合图表展示数据特征，并讲述隐藏在数据背后的故事。**

- 最终选课人数为 **83** 人，总可用时间为 **240** 分钟

  ⇒ **4** 人组共 **20** 组，**3** 人组共 **1** 组。

  每次课 **7** 组汇报，每组可用时间约 **11** 分钟

关于小组汇报的分组

- **分组第一阶段**：自由组队。原则上只接受 4 人一队，在 **10月27日** 前报给助教姜宏卓（可以在QQ里私聊）。如果上报 3 人队伍，最后会面临被调整的风险（包括拆分队伍和补充成员）。
- **分组第二阶段**：第一阶段没有组队的同学，或者上报的 3 人队伍，我们会按随机原则进行调整。
- 最终**分组结果**和**汇报顺序**将在 **10月31日** 第三次线下课上公布。
- 在最终分组公布后原则上不再接受人员调整，确实需要调整的小组可先在课下和其他组自行协调，双方同意后将调整结果报给助教。

数据的可视化

统计数据的可视化

Edward R. Tufte 在其著作 *The Visual Display of Quantitative Information* 中写道：

优秀的统计数据可视化应当将复杂的内容通过清晰、准确、有效的方式呈现。
统计图表需要

- **展示数据**
- 使读者关注事实，而不是方法、图形设计、绘图技术等其他方面
- **避免扭曲数据包含的信息**
- 在小空间内展示很多数值
- 在大数据集上体现一致性
- 鼓励读者比较数据的不同部分
- 从整体到细节展示数据的不同层面
- 有清晰合理的目的：解释、探索、展示、或装饰
- 与数据的统计学或语言描述机密结合

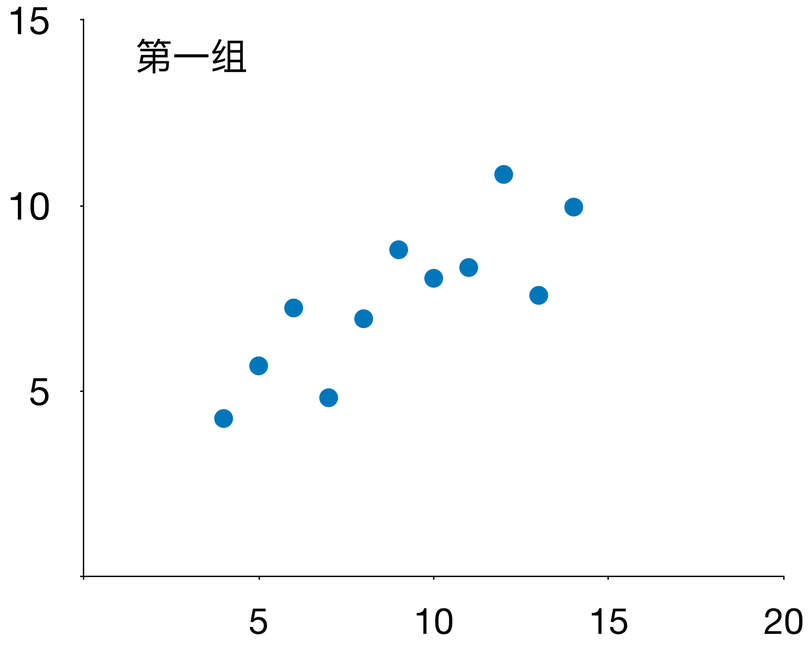
图表有时候比表格更有效

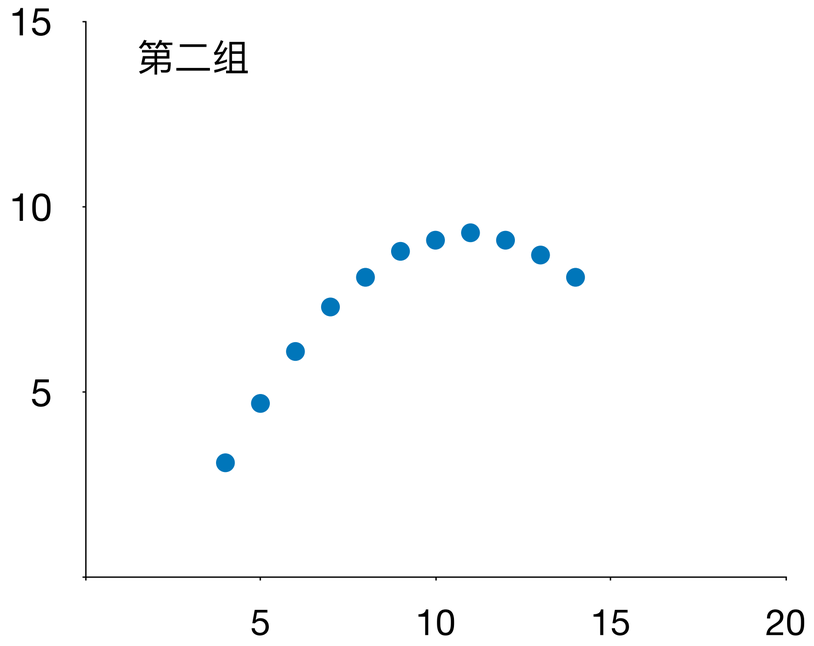
Anscombe, F. J. (1973). Graphics in Statistical Analysis. *American Statistician*, 27:17-21.

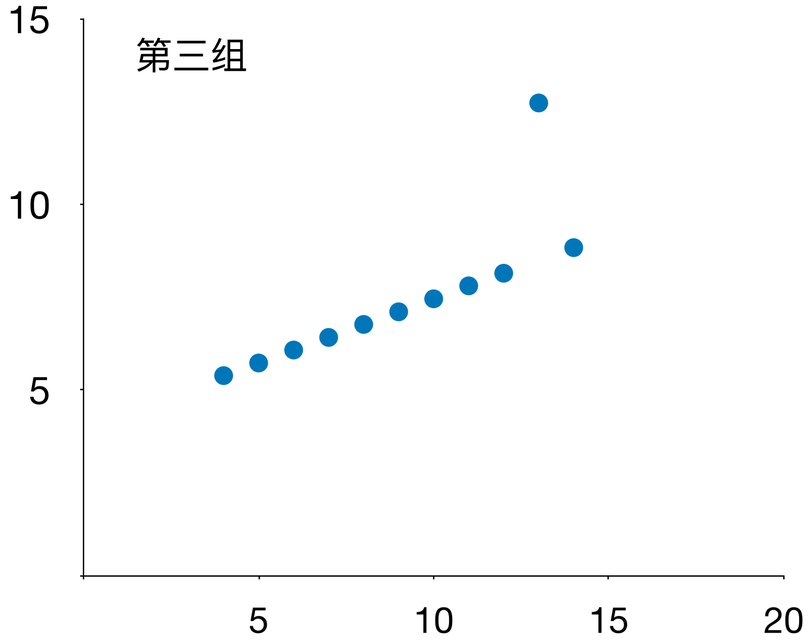
第一组		第二组		第三组		第四组	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10.0	8.04	10.0	9.1	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.1	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.7	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.8	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.3	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.1	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.1	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.1	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.1	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.3	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.7	5.0	5.73	8.0	6.89

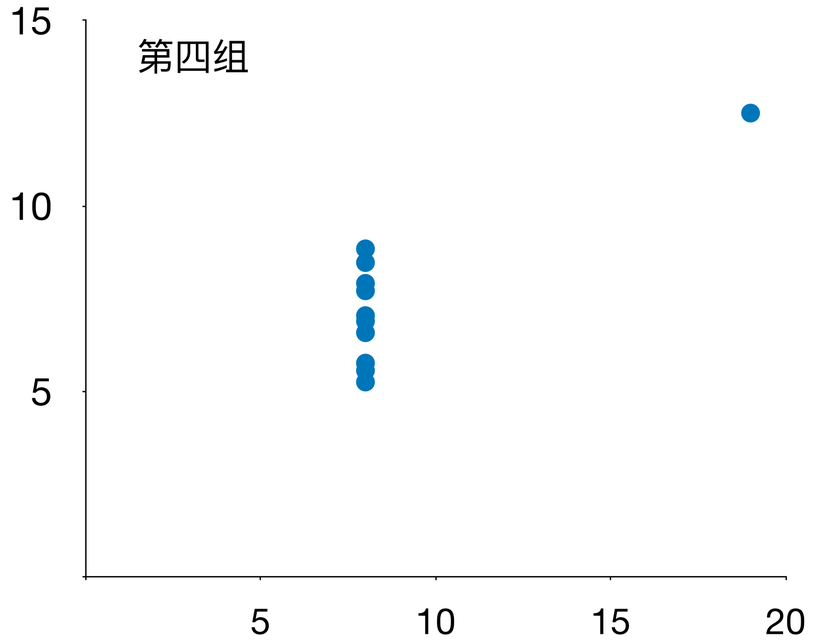
四组数据的描述性统计一致：

- 样本量 = 11
- X 的均值 = 9.0， Y 的均值 = 7.5
- 线性回归函数 $Y = 3 + 0.5X$
- 斜率的估计标准误 = 0.118 (t = 4.24)
- $X - \overline{X}$ 的平方和 = 110.0
- 回归平方和 = 27.50
- 残差平方和 = 13.75
- 相关系数 = 0.82
- R^2 = 0.67

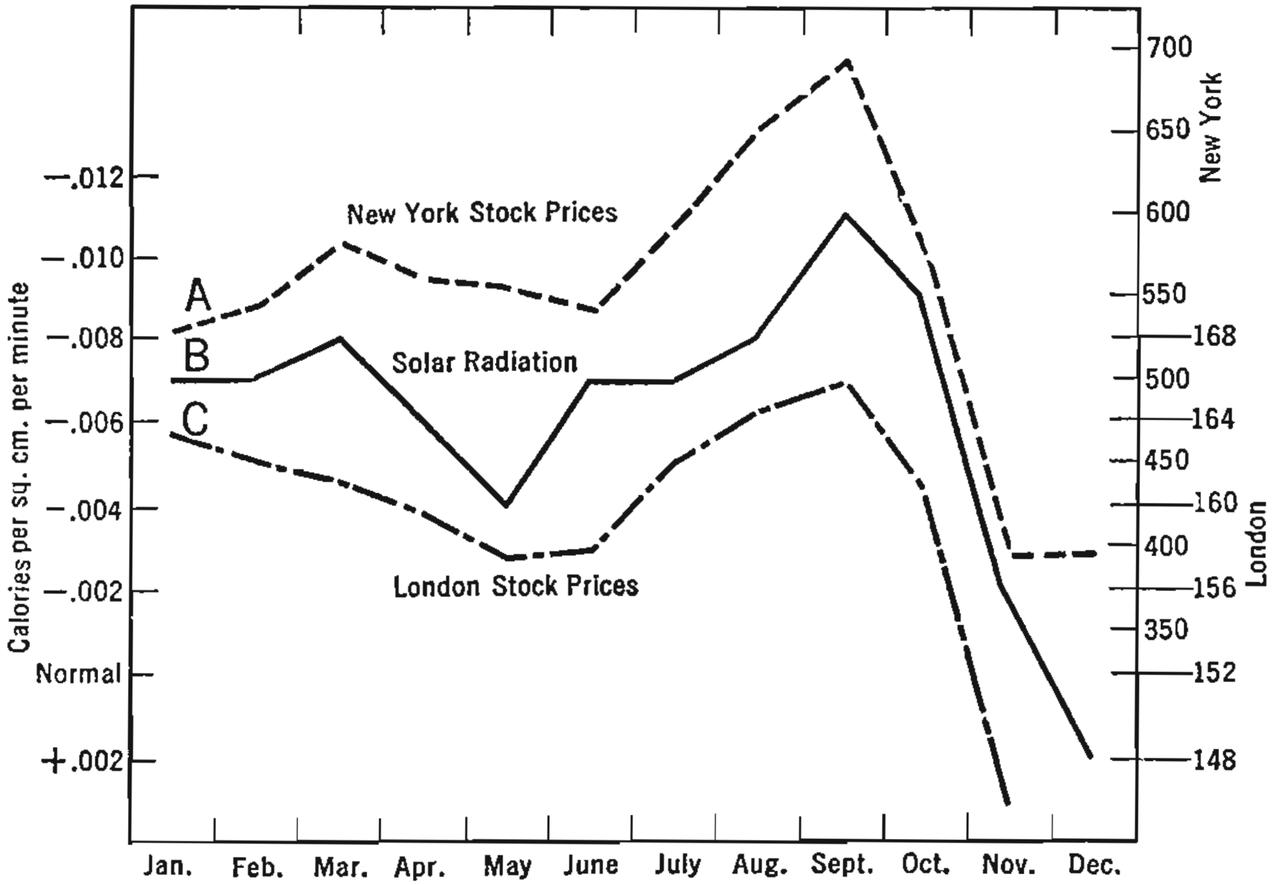

第一组
15
10
5
5
10
15
20


第二组
15
10
5
5
10
15
20


第三组
15
10
5
5
10
15
20


第四组
15
10
5
5
10
15
20

图表无法填补数据或模型本身的不足



SOLAR RADIATION AND STOCK PRICES

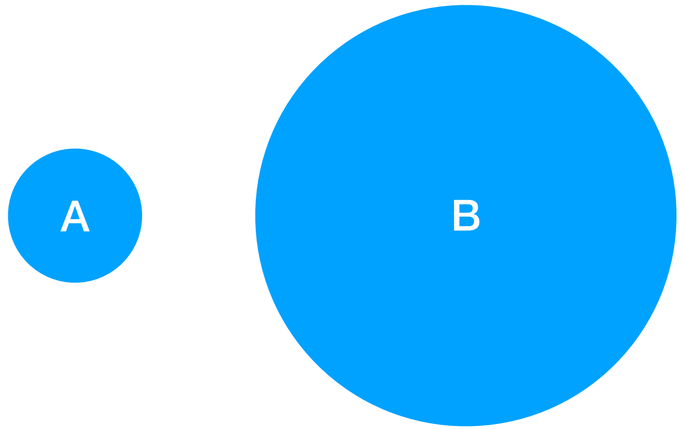
A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

Dewey, E. R. & Dakin, E. F. (1947). *Cycles: The Science of Prediction*, p144. New York.

图片取自 Tufte, E. R. (2001). *The Visual Display of Quantitative Information*, 2nd Edition, p15.

不能用数据骗人

- 我们都知道数据会“骗人”，可视化后的数据同样会“骗人”。
- 图表不应当扭曲数据中的信息，但是不同的人对同一图形的感知可能存在差异。



B 的面积是 A 的几倍?

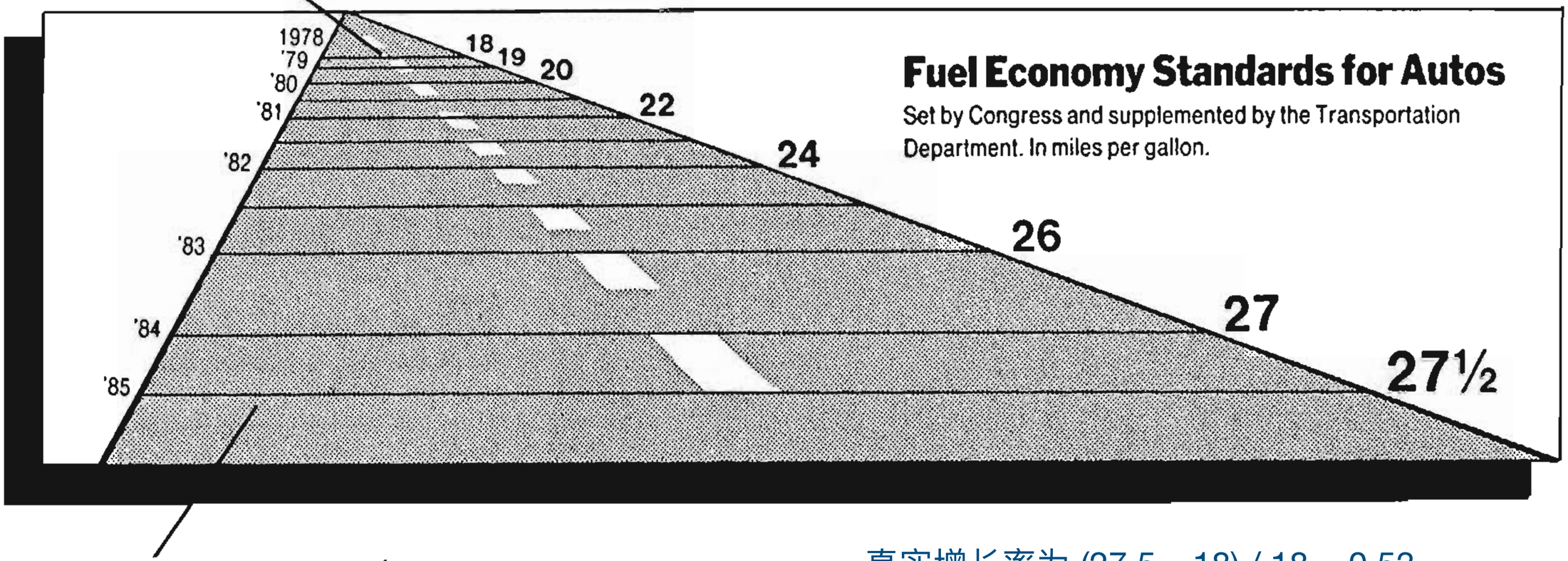
- 8 倍
- 9 倍
- 10 倍
- 11 倍

- 那么我们应该为每一个人量身定做一幅图吗?
  或是以读者的平均感知水平为标准?
- Tufte 认为，可视化应该满足两个原则：
  1. 图中的物理测度（长度、面积等）应当和它所展示的数值成正比
  2. 应当有清晰详细全面的标注，以减少读者对数据的误解。
     例如在图中加入对数据的说明，或标注重要的事件等。
- “欺骗系数” = 图中展示的效果大小 / 数据中的效果大小

New York Times, August 9, 1978, p. D-2.

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

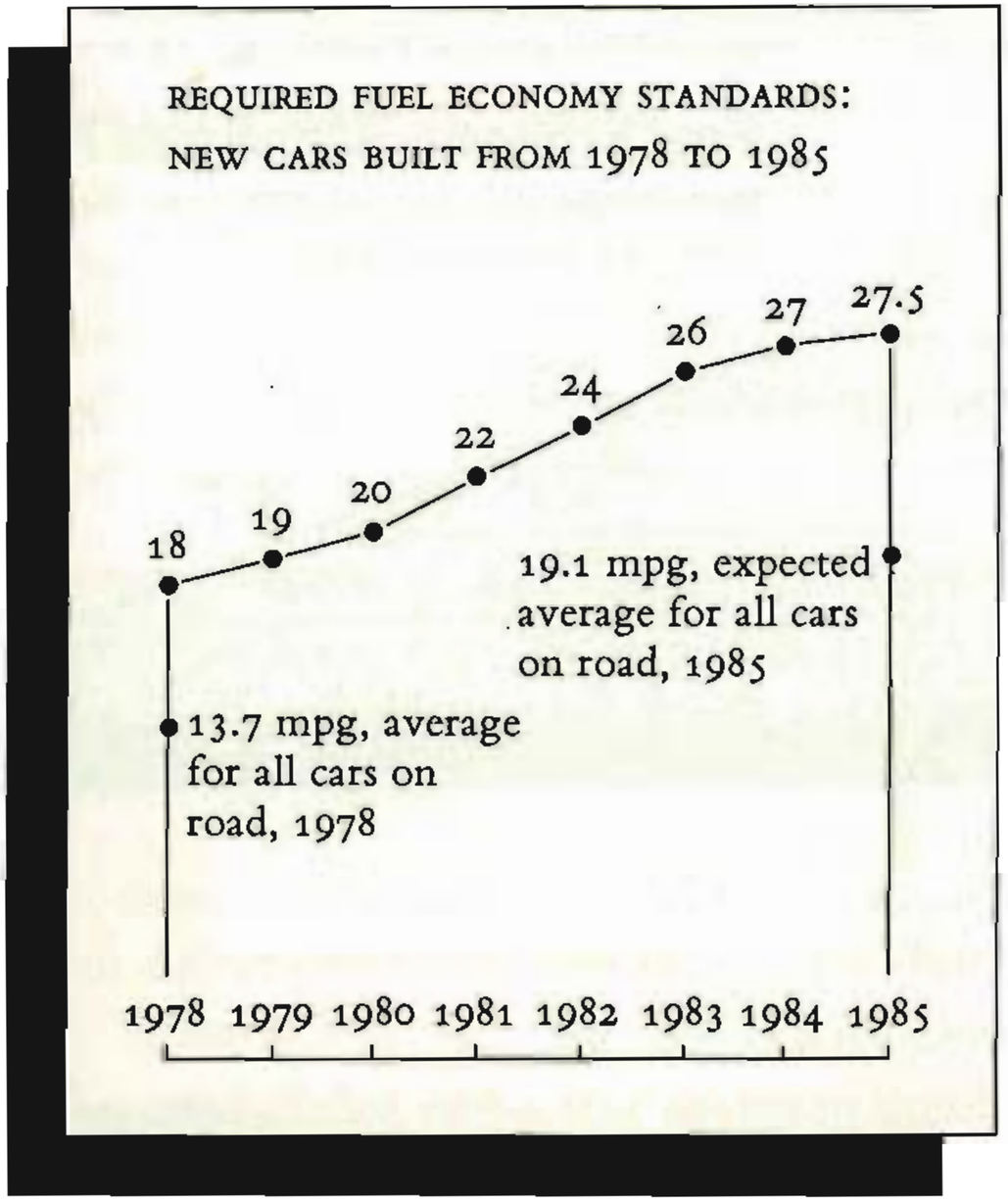
美国国会和交通部对新车的燃油经济性曾作出规定，从1978年的 18mpg 起应逐年改善，最终于1985年达到 27.5mpg 的目标



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

真实增长率为 (27.5 – 18) / 18 = 0.53
图中展示的增长率为 (5.3 – 0.6) / 0.6 =7.83
欺骗系数 = 7.83 / 0.53 = 14.8

图片取自 Tufte, E. R. (2001). p57.



Tufte 对前图的改善（无欺骗版）

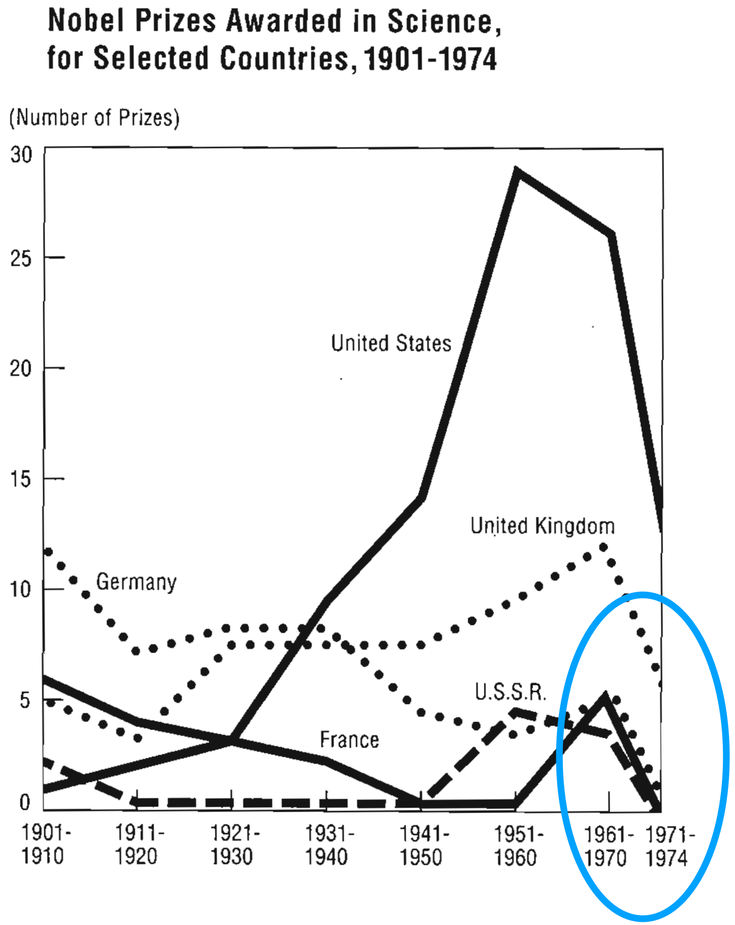
除了准确展示了数字，此图的改进点还包括

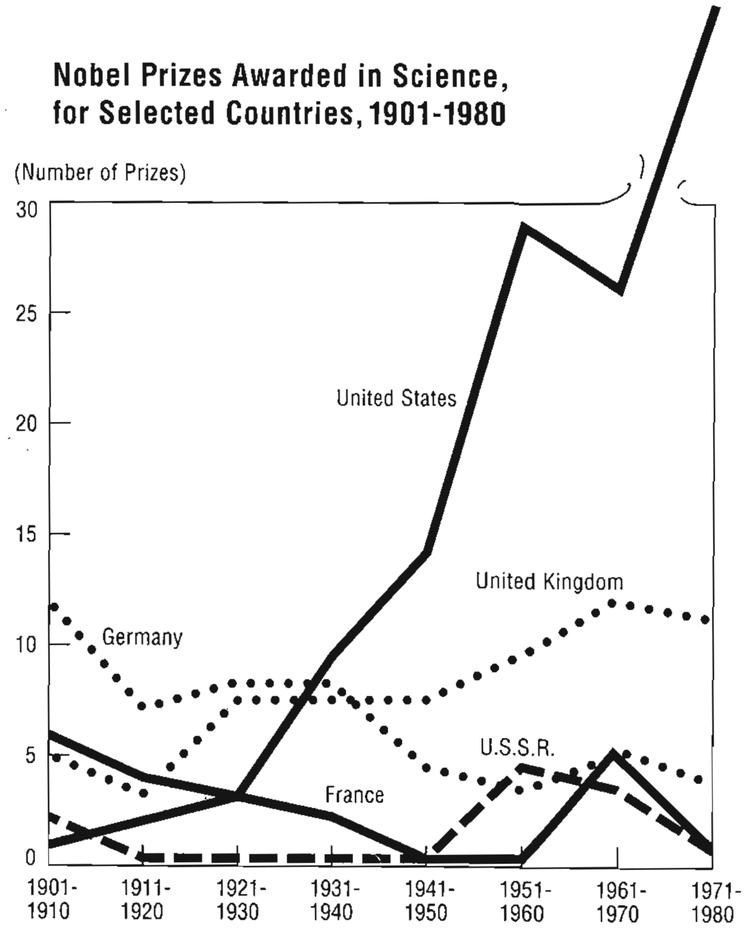
- 加入了对所有车辆平均燃油经济性的估计值（标准只针对新车）
- 时间发展从左到右，符合阅读习惯
- 年份间隔相同
- 折线图更好的展示了1980-1983年间标准变化率的增大

展示数据的变化，而不是图形设计的变化

- 人们会根据图的一部分预测其他部分，因此图表的各部份之间应该保持一致
- 左图的横轴最右侧的跨度为4年，与其他部分的10年不一致
- 右图是将数据补充完整后展示的正确信息（即使没有准确的数据，也应将4年间的数据乘以2.5进行估计）
- 左图的问题完全来自图形设计上的变化，而不是数据本身的变化

National Science Foundation, *Science Indicators, 1974* (Washington, D.C., 1976), p. 15.





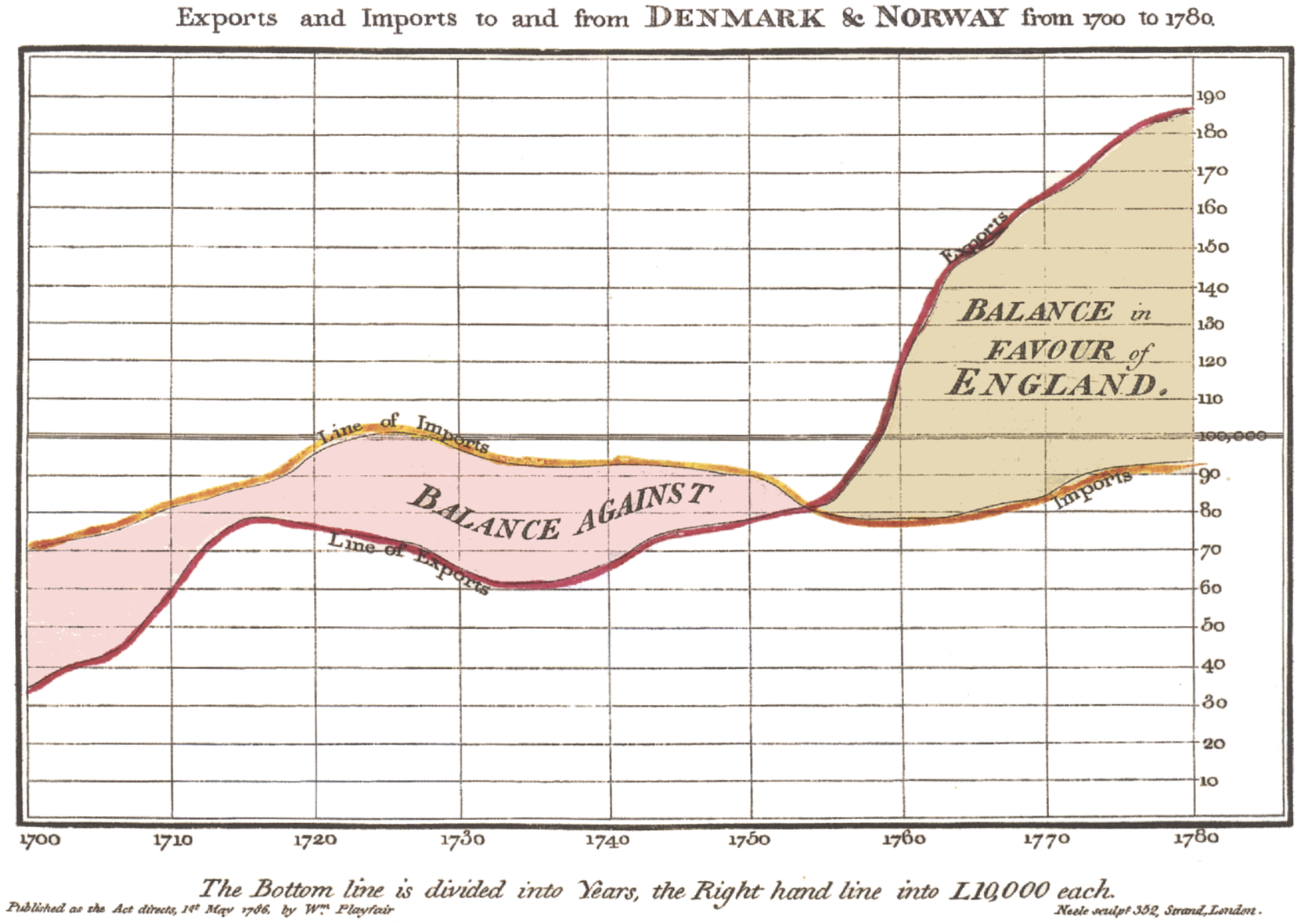
常用统计图表的起源

折线图、柱形图、饼图

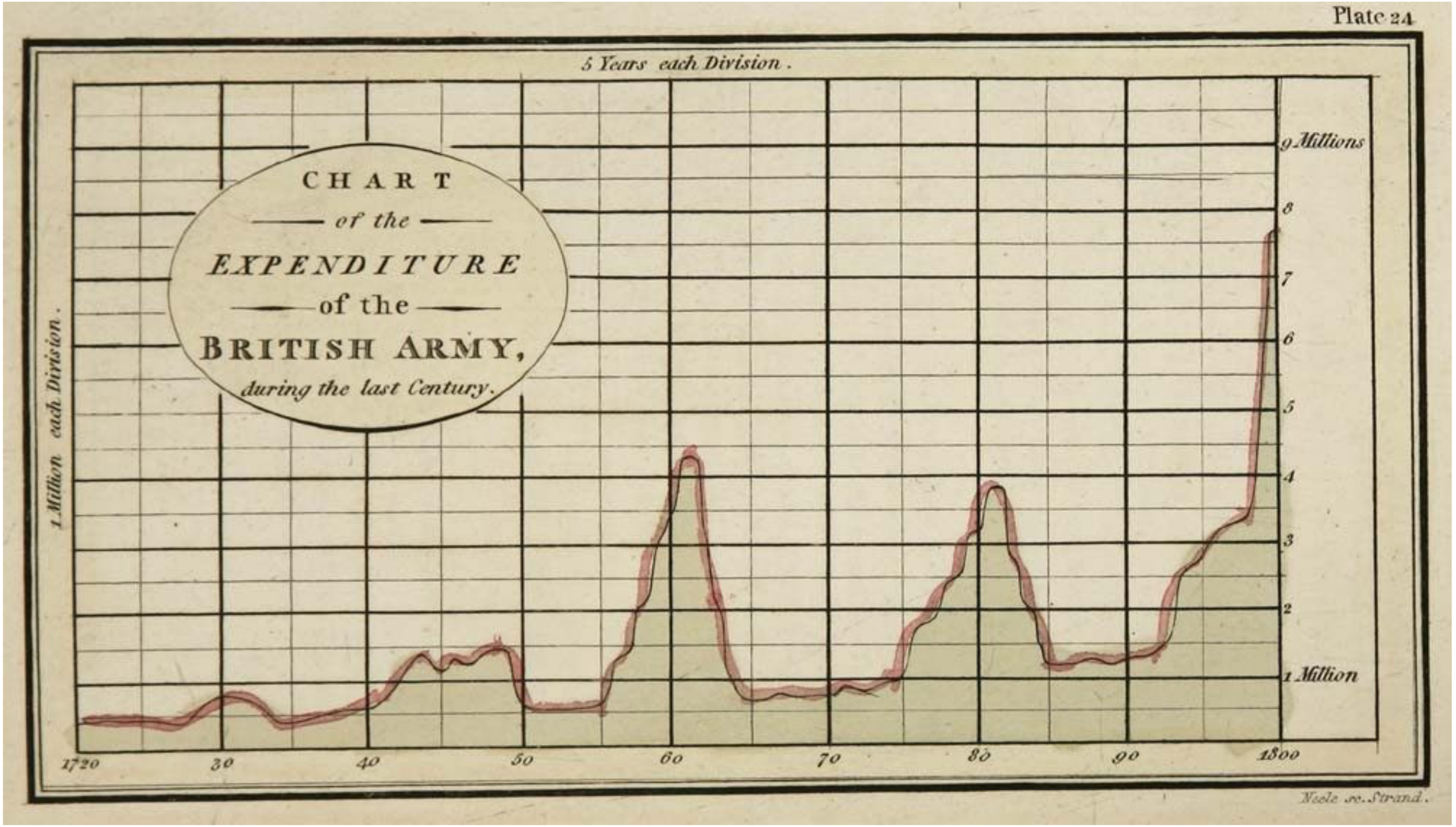
- 第一个将统计数据可视化付诸实践的是18世纪苏格兰工程师 William Playfair (1759 – 1823)，被誉为现代统计图表之父，在他的著作中可以发现许多折线图、柱形图和饼图的应用。
- Playfair, W. (1786). *The Commercial and Political Atlas*.
  关于英格兰的国际贸易和其他经济数据，共包含 44 幅图，包括折线图和柱形图。在此之前，学者们通常只使用表格展示数据。
- Playfair, W. (1801). *Statistical Breviary*.
  首次使用饼图。

William Playfair

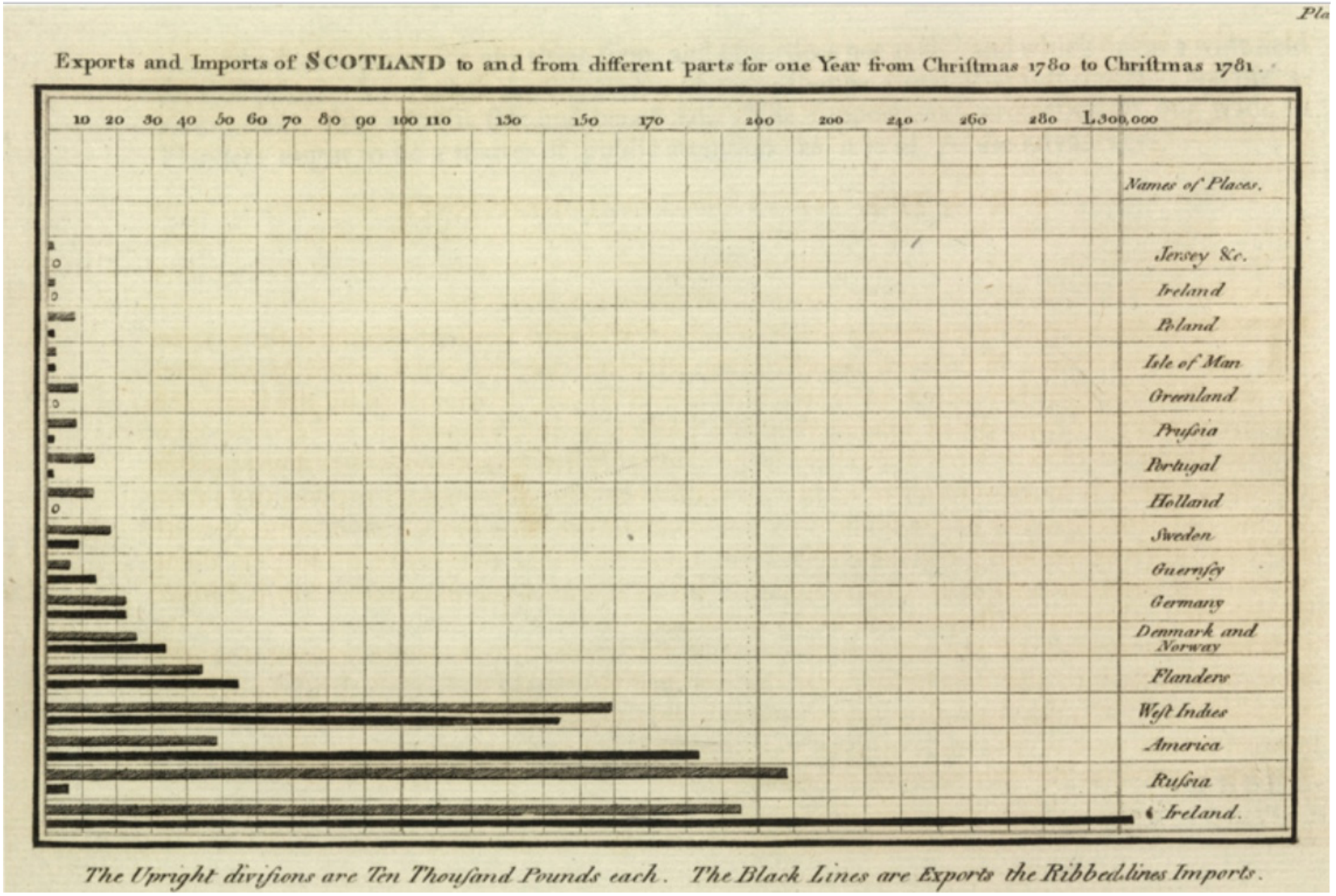
来源：https://www.playfairprize.com/william-playfair



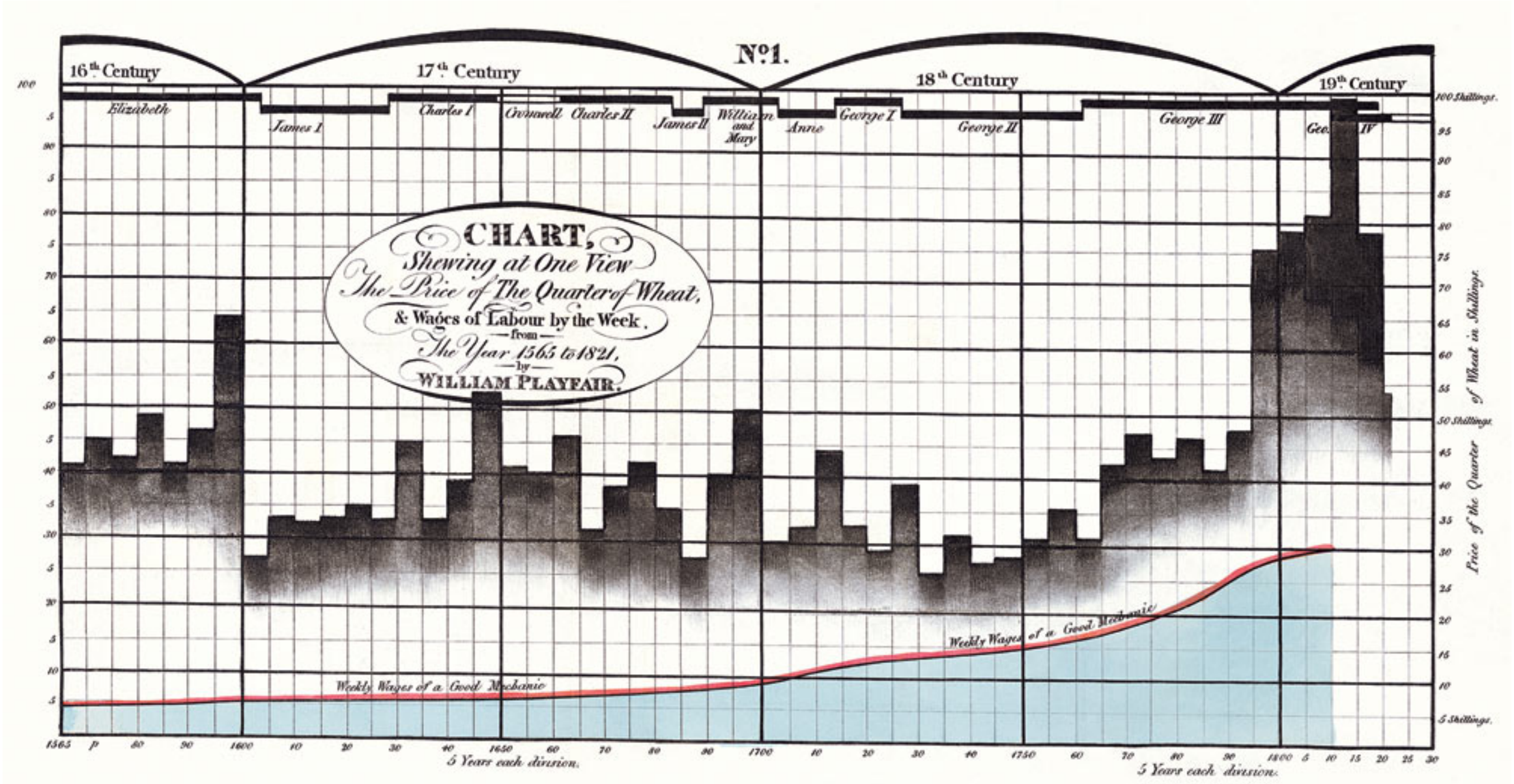
Playfair, W. (1786). *The Commercial and Political Atlas*. 英格兰在1700-1780年间对丹麦和挪威的进出口额。



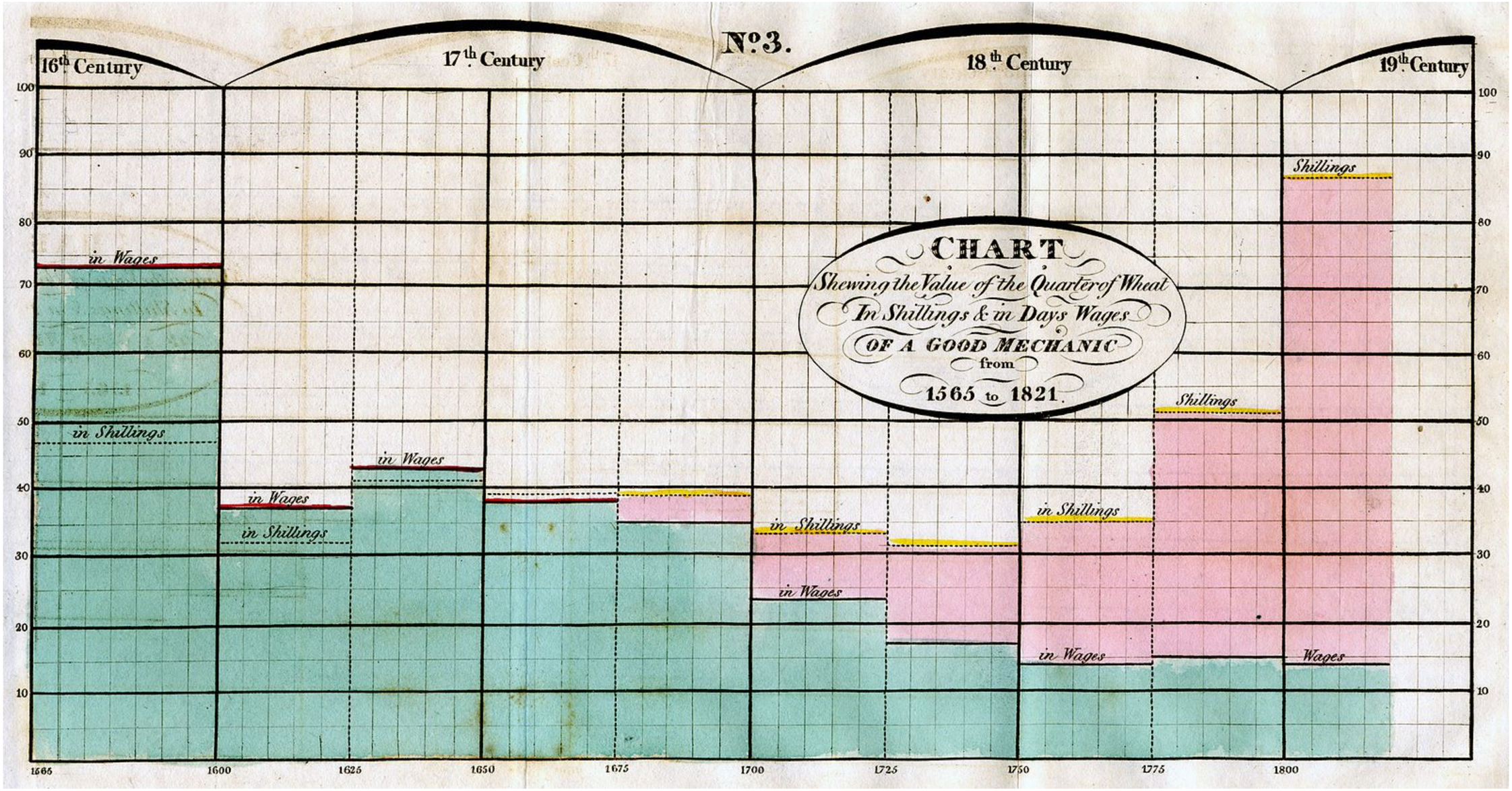
Playfair, W. (1801). *The Commercial and Political Atlas*, 3rd. 1720-1800年间英国的陆军支出。



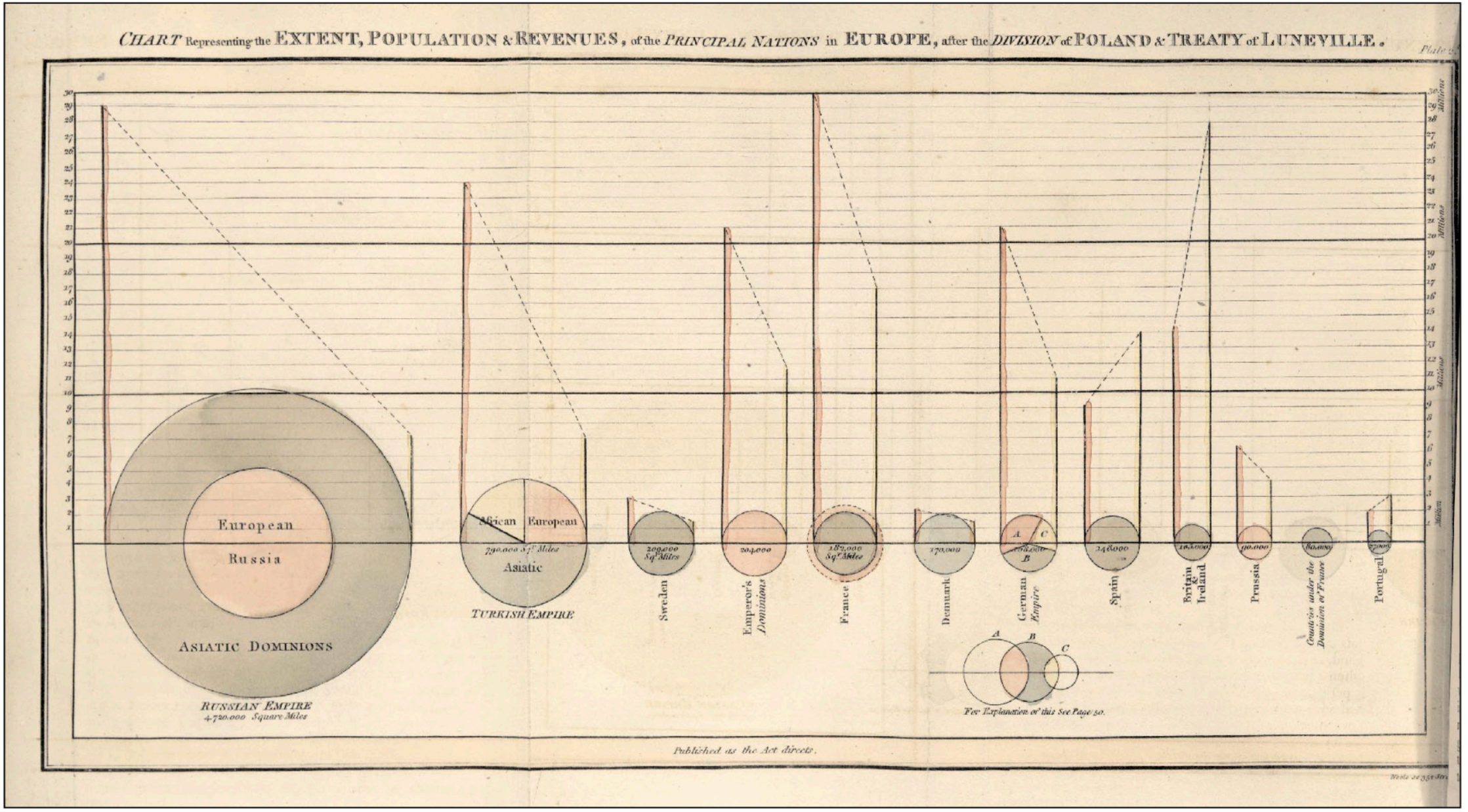
Playfair, W. (1786). *The Commercial and Political Atlas*. 1780圣诞-1781圣诞间苏格兰对其他国家或地区的进出口额。



Playfair, W. (1821). *A letter on our agricultural distresses, their causes and remedies.* 小麦（1/4磅）的货币价值（先令），一个优秀技工的周薪变化，以及英国的朝代更替。



Playfair, W. (1821). *A letter on our agricultural distresses, their causes and remedies.* 小麦（1/4磅）的货币价值（先令），以及一个优秀技工需要花多少天的工资才可以买到它（小麦价格/日薪）。



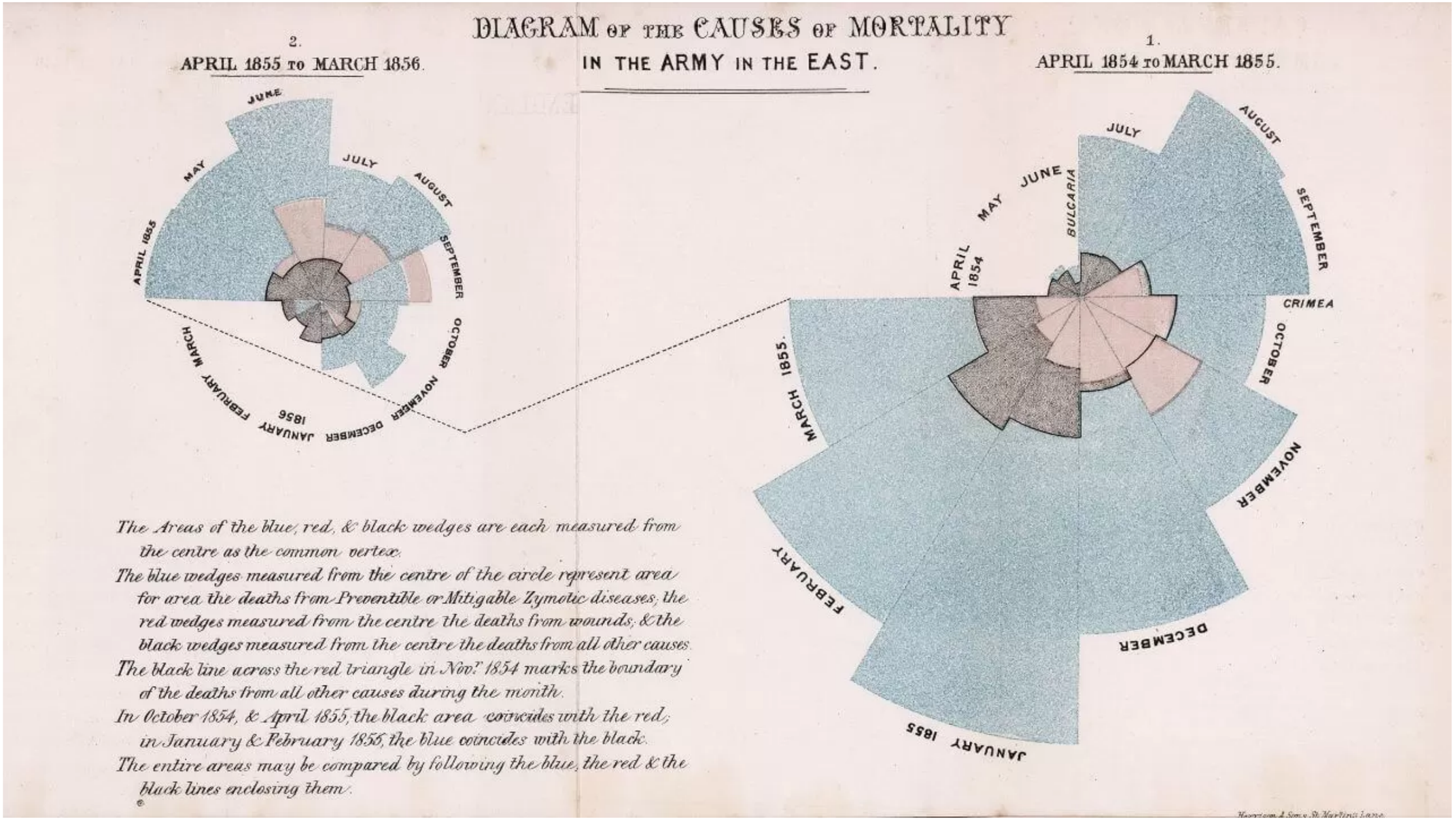
Playfair, W. (1801). *Statistical Breviary*. 圆的直径代表国土面积，颜色区分海洋国家和陆地国家，圆左侧的竖线代表人口，右侧竖线代表税收收入。同一国家内的不同区域分别以同心圆（俄罗斯）和扇形（土耳其、德意志）的方式展示。

南丁格尔与玫瑰图

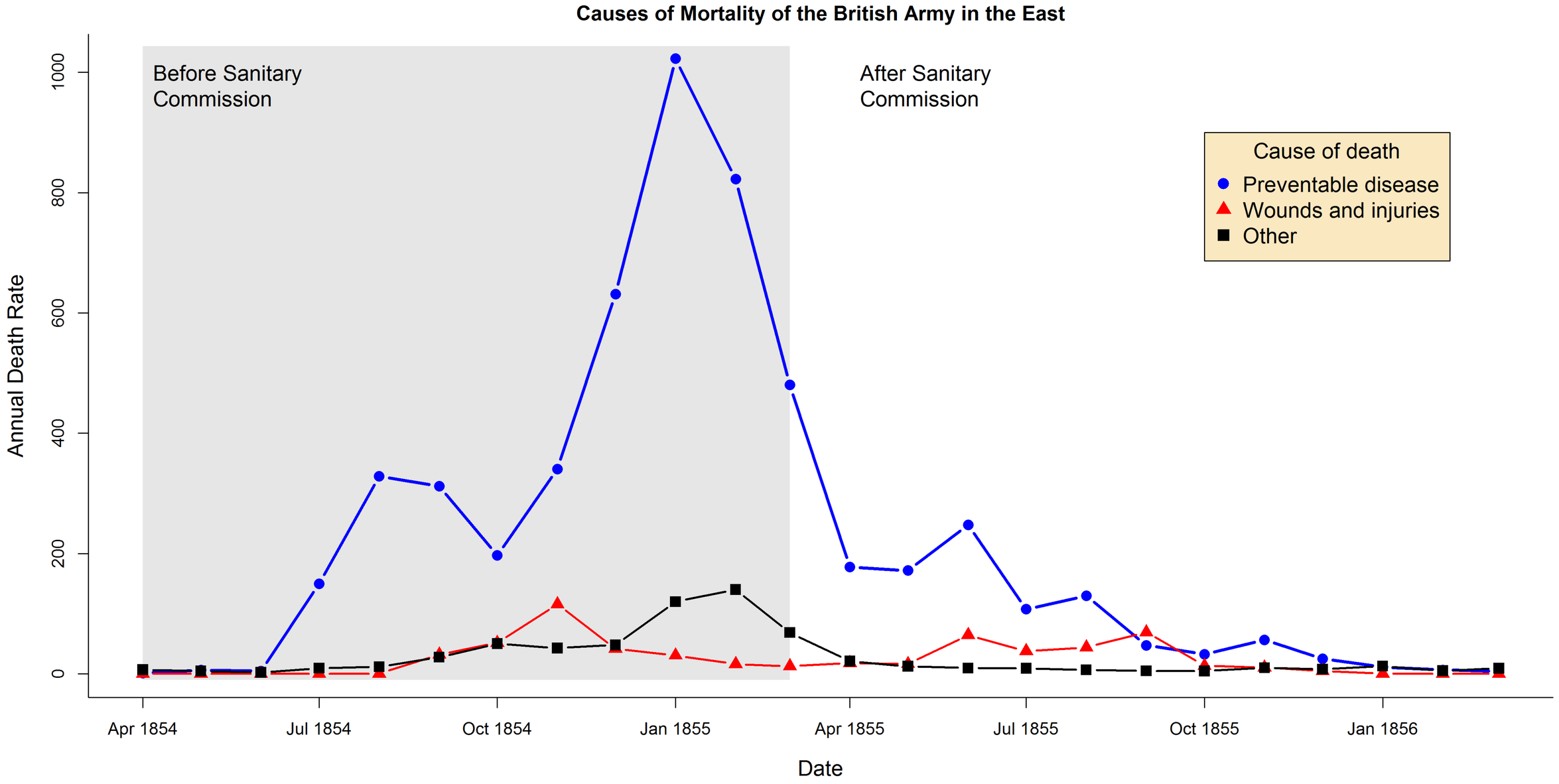
- Florence Nightingale (1820 – 1910) 是我们熟知的现代护理工作的奠基人，同时她也是一名应用统计学家和社会活动家。
- 南丁格尔最有影响力的统计可视化作品是她的玫瑰图，也称为极坐标区域图（polar area diagram）或鸡冠花图（coxcombs，南丁格尔自己的命名），用来描述在克里米亚战争期中后期 (1854 – 1856) 英军士兵死亡人数和死亡原因的分布及演化。
- 南丁格尔极具视觉冲击效果的玫瑰图帮助她说服了英国政府建设战地医院以提高士兵的存活率。

Florence Nightingale

来源：https://www.playfairprize.com/florence-nightingale



Nightingale, F. (1859). *A contribution to the sanitary history of the British army during the late war with Russia*. 南丁格尔玫瑰图。每一个扇形代表一个自然月，**面积**代表死亡人数，蓝红黑三色重叠。蓝色部分为传染病导致的可以避免的死亡，红色部分为受伤导致的死亡，黑色部分为其他原因导致的死亡。1855年4月（左图），英国政府成立卫生委员会改善战地卫生条件，死亡人数也因此下降。

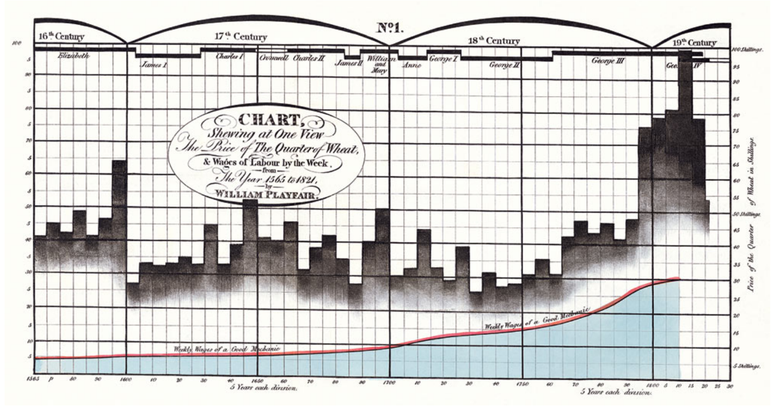


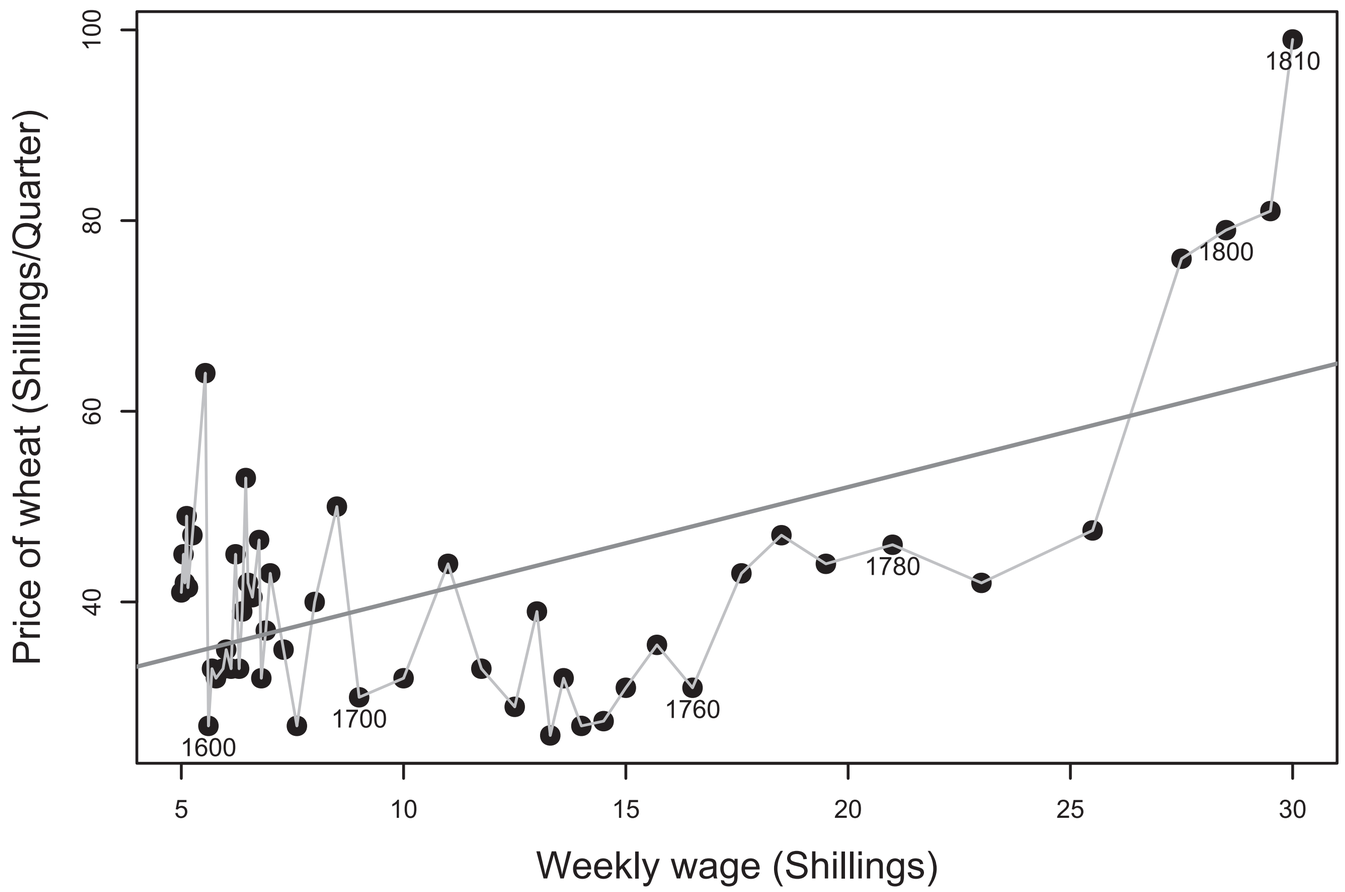
Friendly, M. & Wainer, H. (2021). *A History of Data Visualization and Graphic Communication*. Harvard University Press.

作者将南丁格尔玫瑰图中的数据以折线图的形式重新描绘。

散点图

- 散点图可以清晰地展示两个变量之间的相关关系。
- 右图是根据 Playfair 小麦价格和工资水平的时序图重新绘制的散点图，横轴为周薪，纵轴为小麦价格。

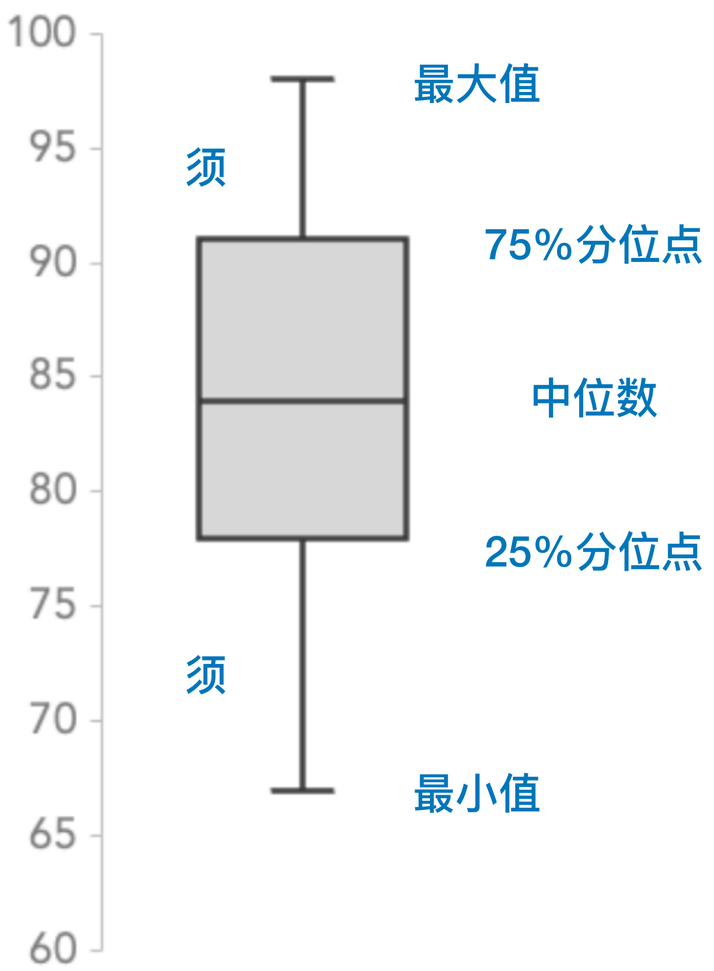




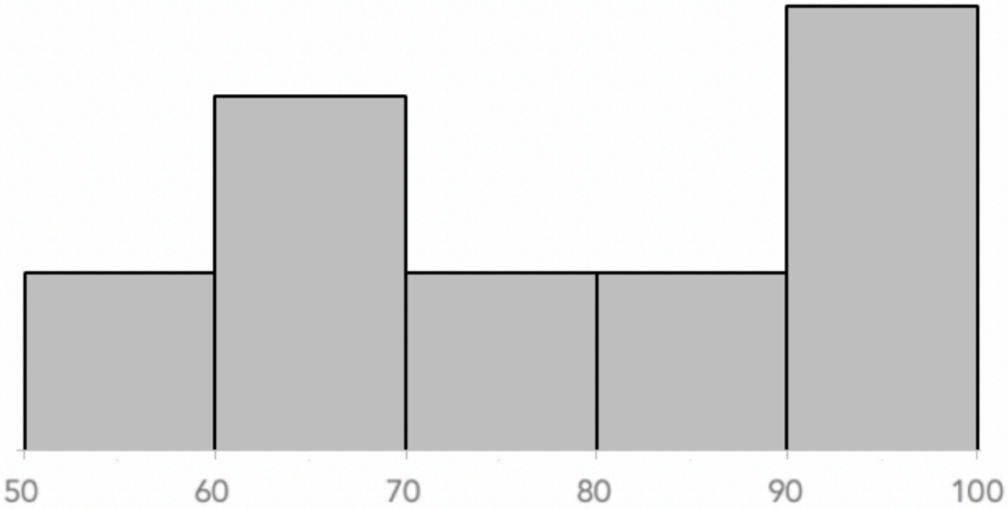
Friendly, M. & Wainer, H. (2021). *A History of Data Visualization and Graphic Communication*. Harvard University Press.

展示数据的分布

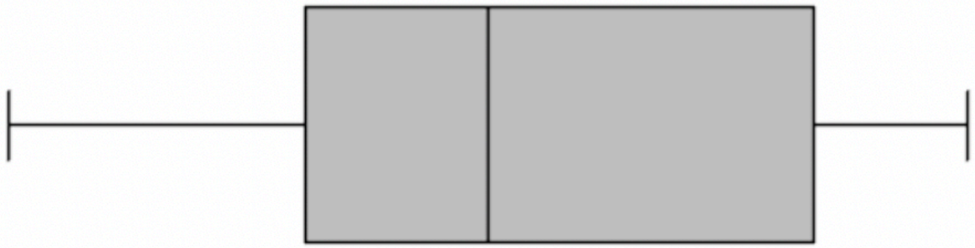
箱线图

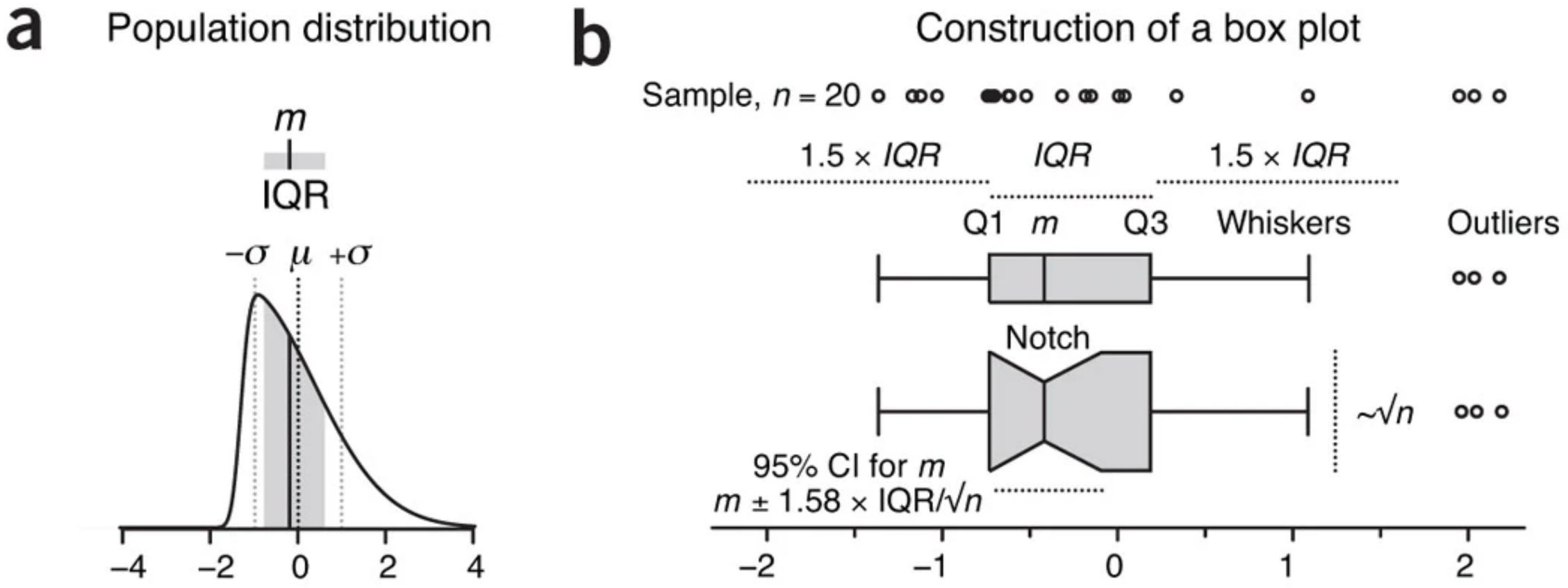

100
95
90
85
80
75
70
65
60
最大值
须
75%分位点
中位数
25%分位点
须
最小值

直方图


50
60
70
80
90
100

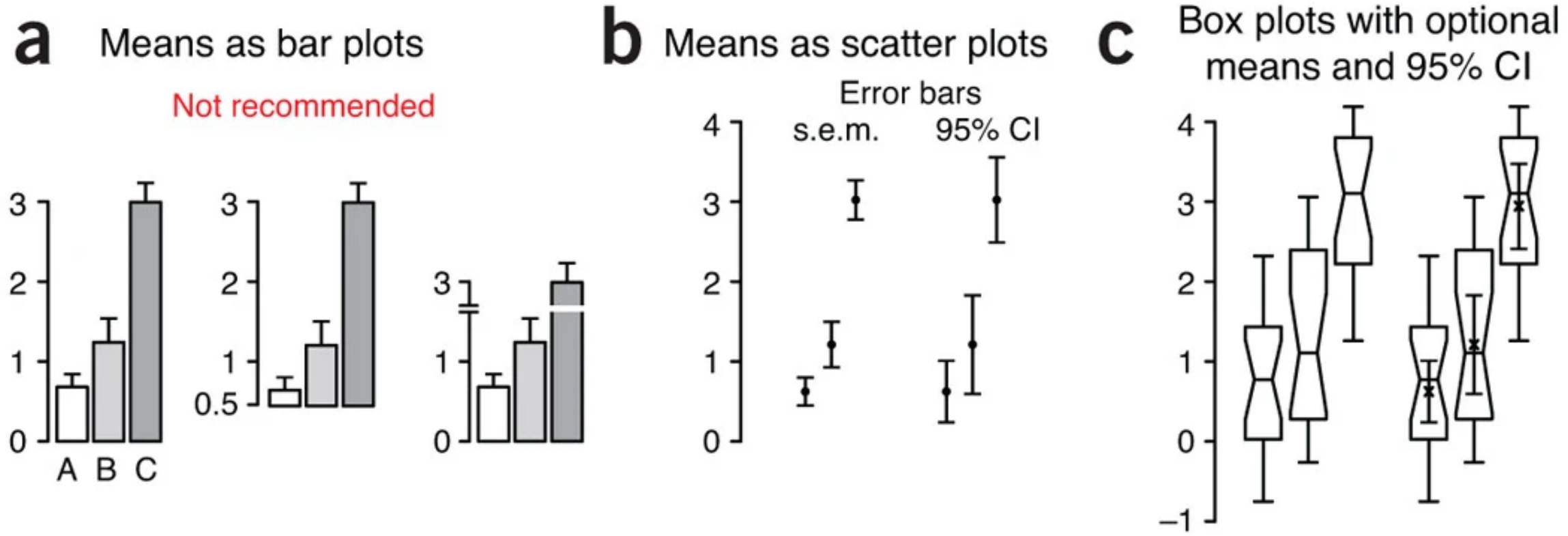
箱线图





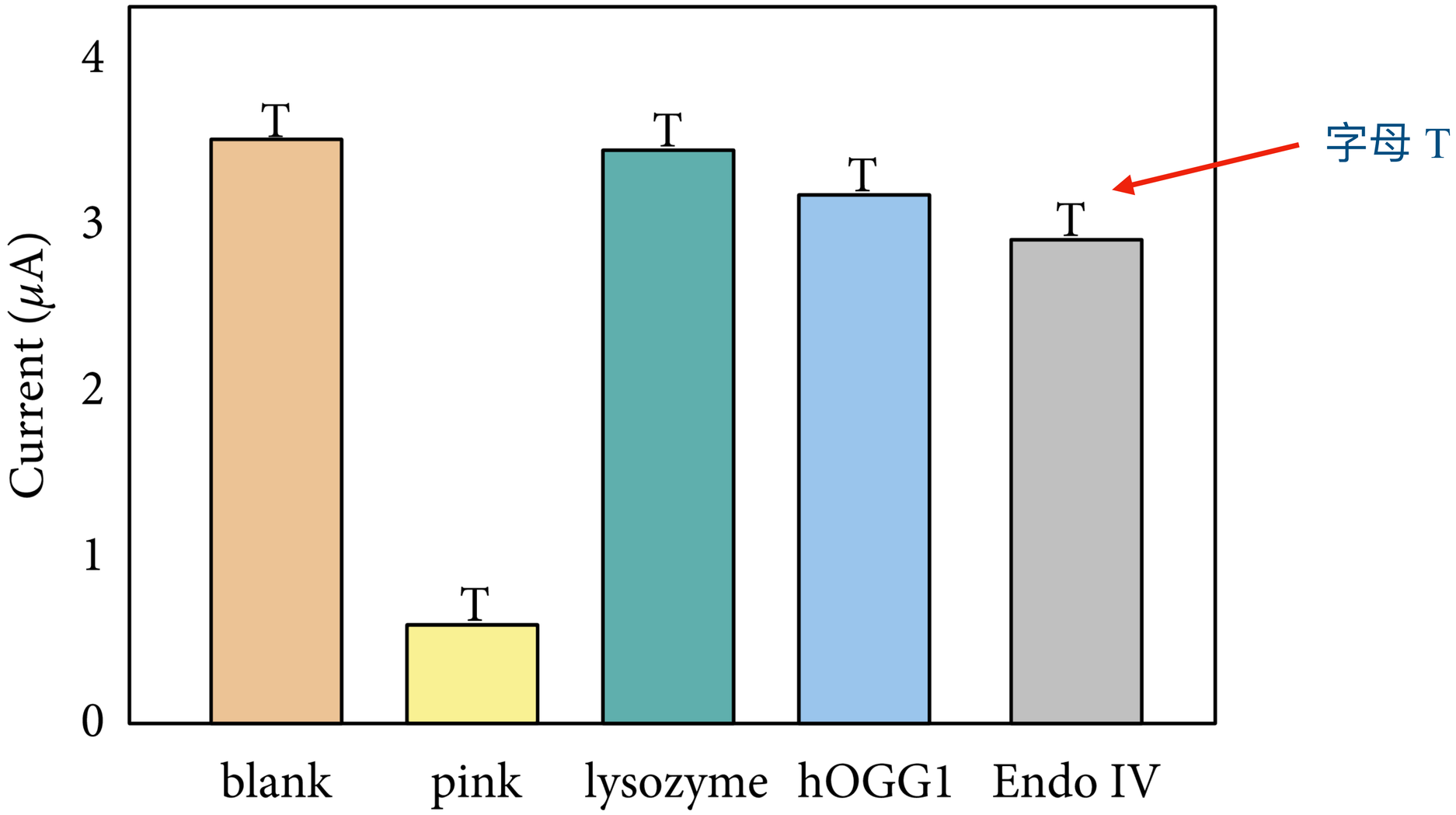
Krzywinski, M. & Altman, N. (2014). Visualizing samples with box plots. *Nature Methods*, 11:119-120.

另一种常见的箱线图将须的两端绘制在箱的两端正负1.5箱长范围内的最大和最小观测值处，并将剩余的观测值作为极值绘制。如果数据服从正态分布，则须的两端之间约等于 $\mu \pm 2.7\sigma$，大约包含总数据的 99.3%。



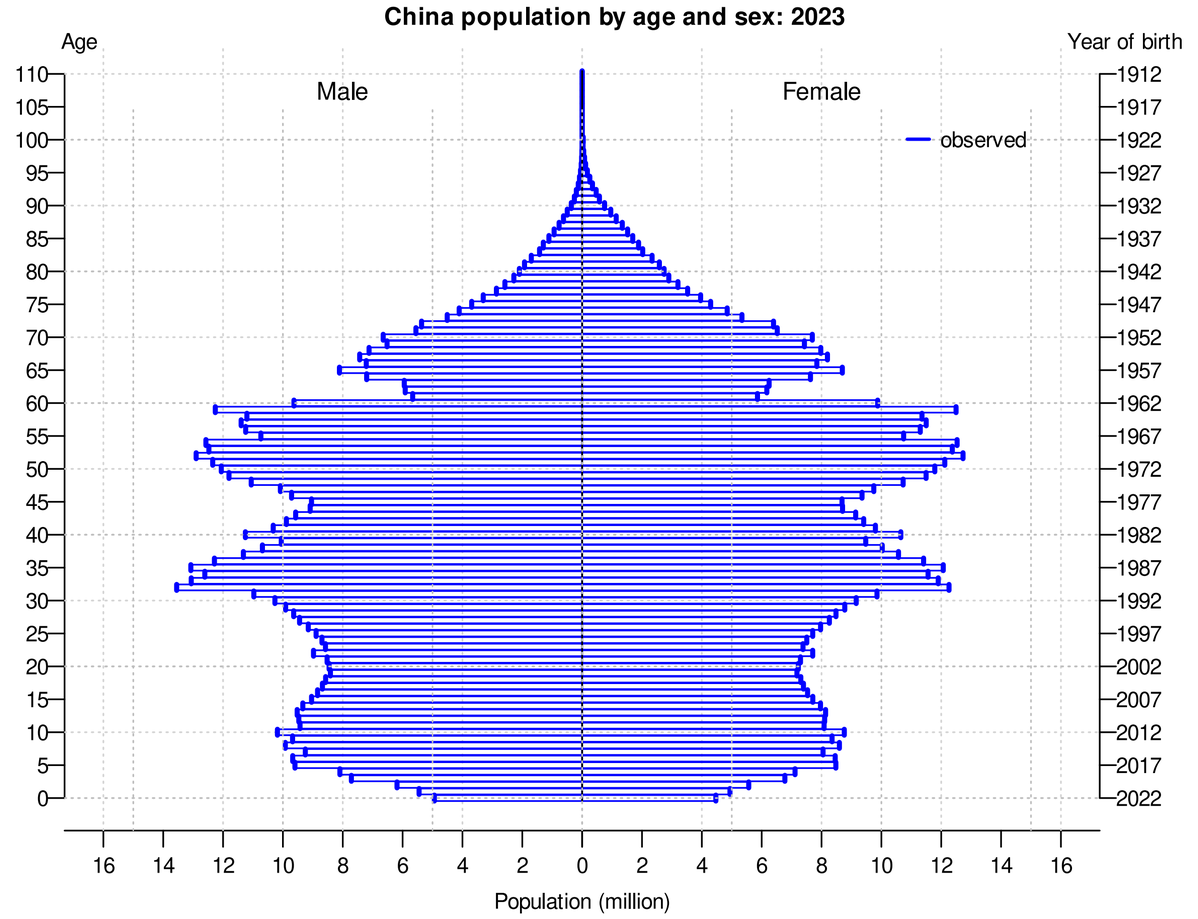
Krzywinski, M. & Altman, N. (2014). Visualizing samples with box plots. *Nature Methods*, 11:119-120.

(a) 常见的以柱形图的形式绘制样本均值和标准误。**不推荐以这种形式描绘样本的分布特征**。(b) 样本均值 ± 标准误或样本均值的95%置信区间更加合适。(c) 箱线图与置信区间相结合。

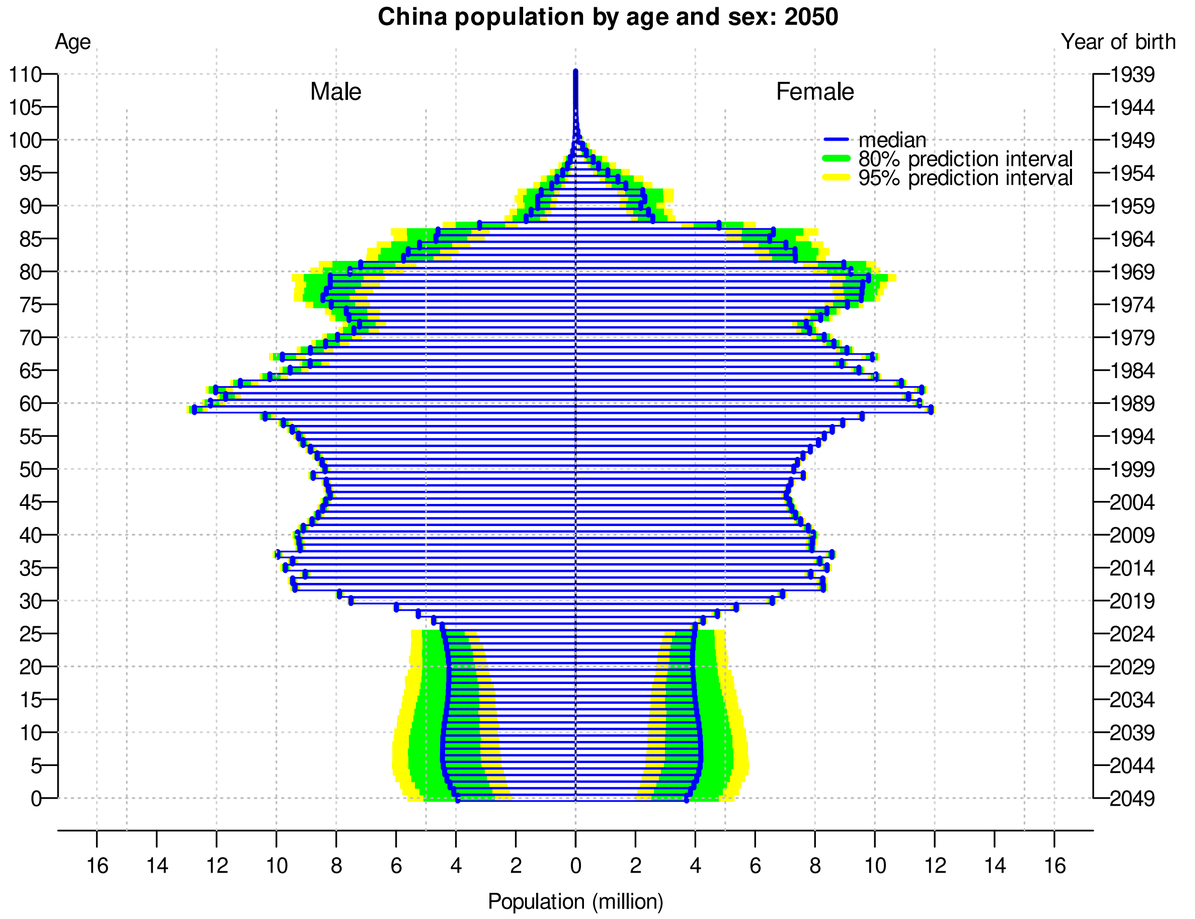


Gong, R. & Liu, B. (2022). Monitoring of Sports Health Indicators Based on Wearable Nanobiosensors, *Advances in Materials Science and Engineering*, vol. 2022, Article ID 3802603, 11 pages, [Retracted]. Figure 9.

人口金字塔



United Nations, DESA, Population Division. *World Population Prospects 2024*. http://population.un.org/wpp/

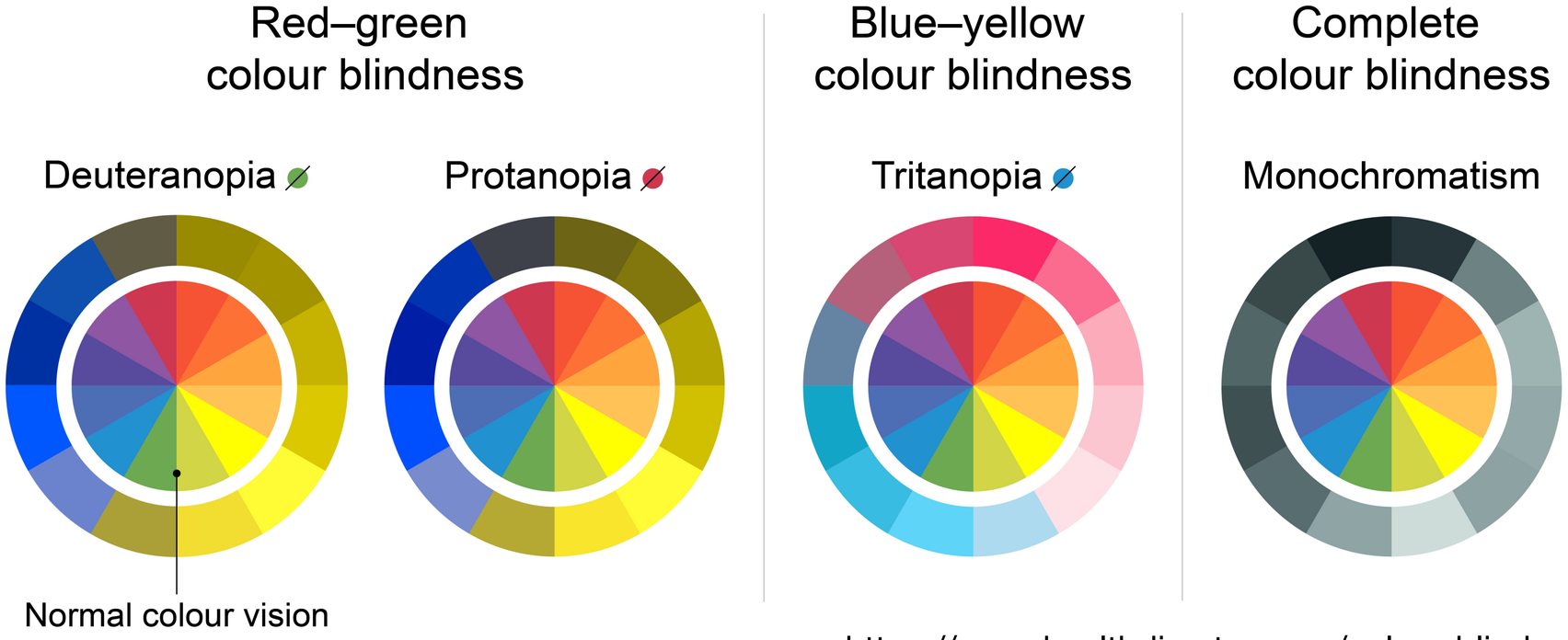


United Nations, DESA, Population Division. *World Population Prospects 2024*. http://population.un.org/wpp/

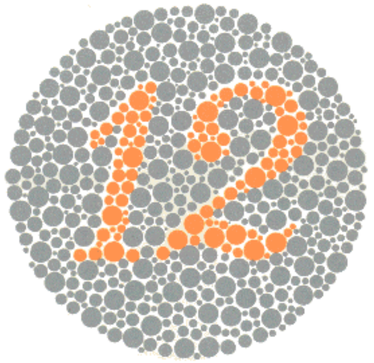
颜色感知偏差

并不是所有人都能正确感知颜色

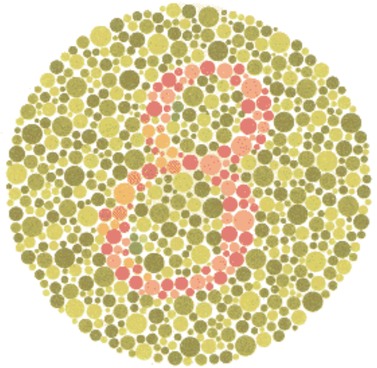
- 男性中有近 8% 为先天性色弱或色盲，女性中的比例约为 0.5%。
- 大多数色弱为红绿系色弱，极少数为蓝黄系色弱或色盲。



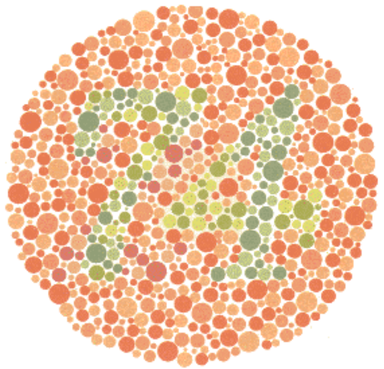
https://www.healthdirect.gov.au/colour-blindness



所有人都应该能够看到数字 12，包括色弱和色盲。

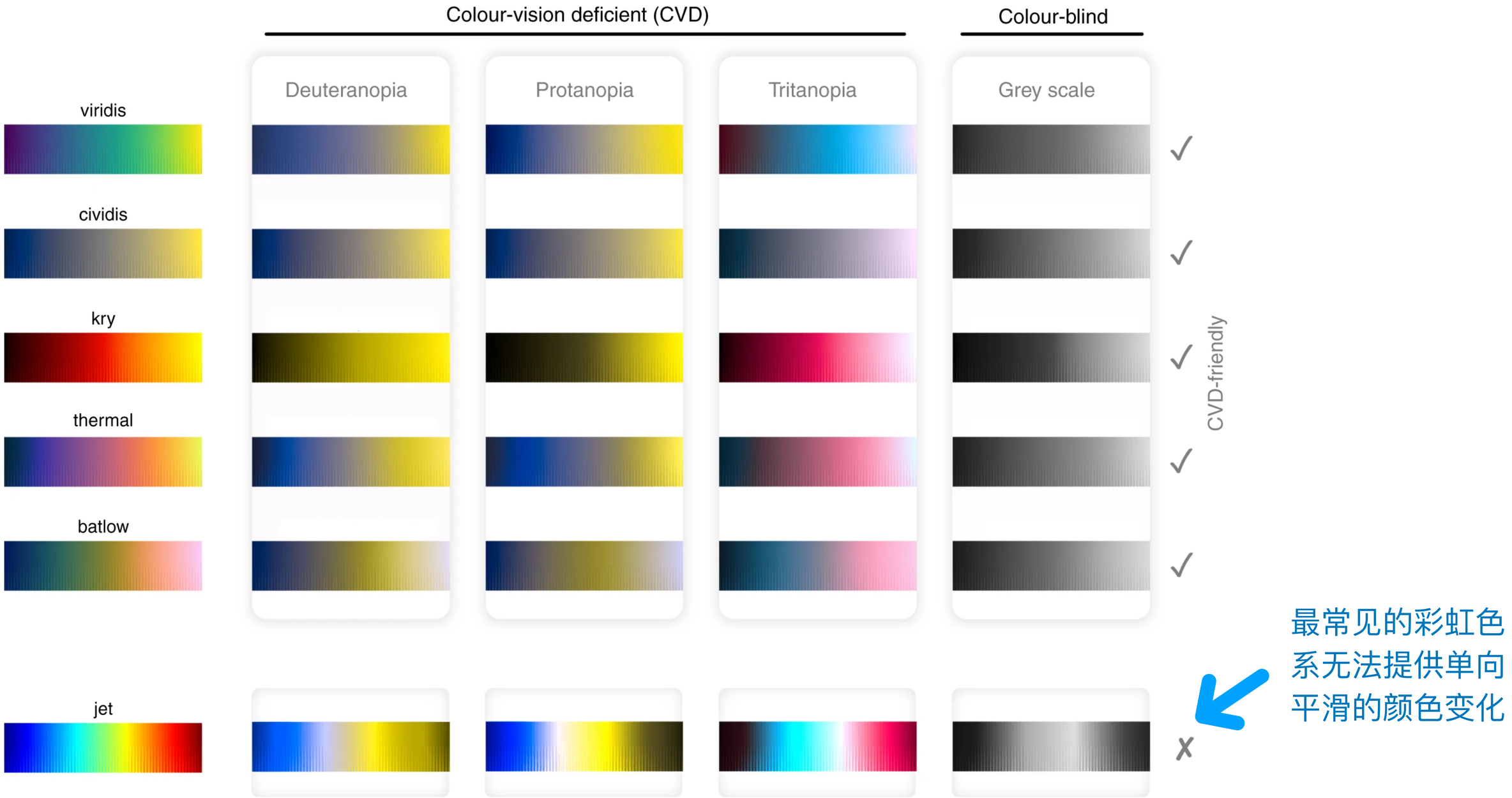


视觉正常者能看到数字 8，红绿系色弱者能看到 3，色盲者看不到数字。

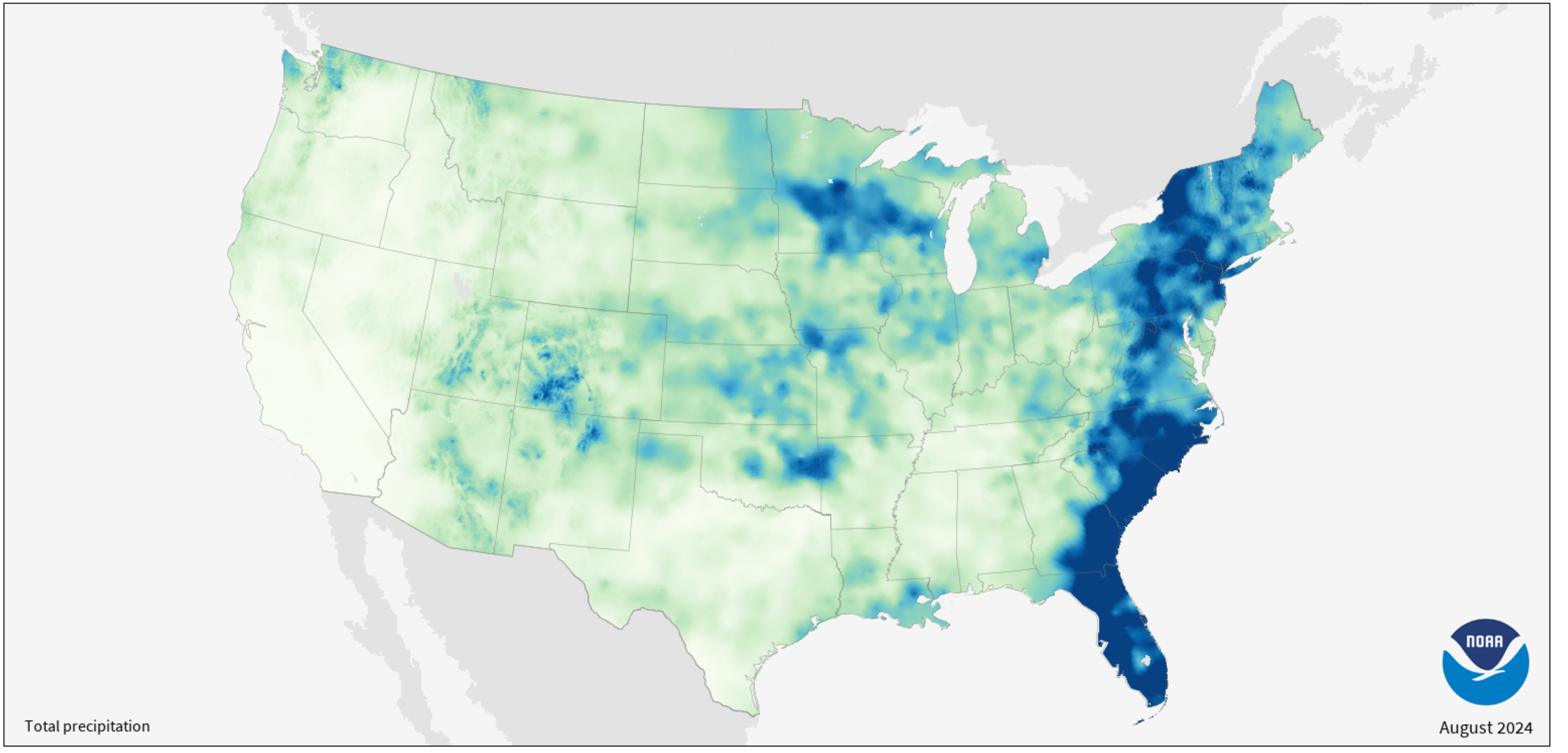


视觉正常者能看到数字 74，红绿系色弱者能看到 21，色盲者看不到数字。

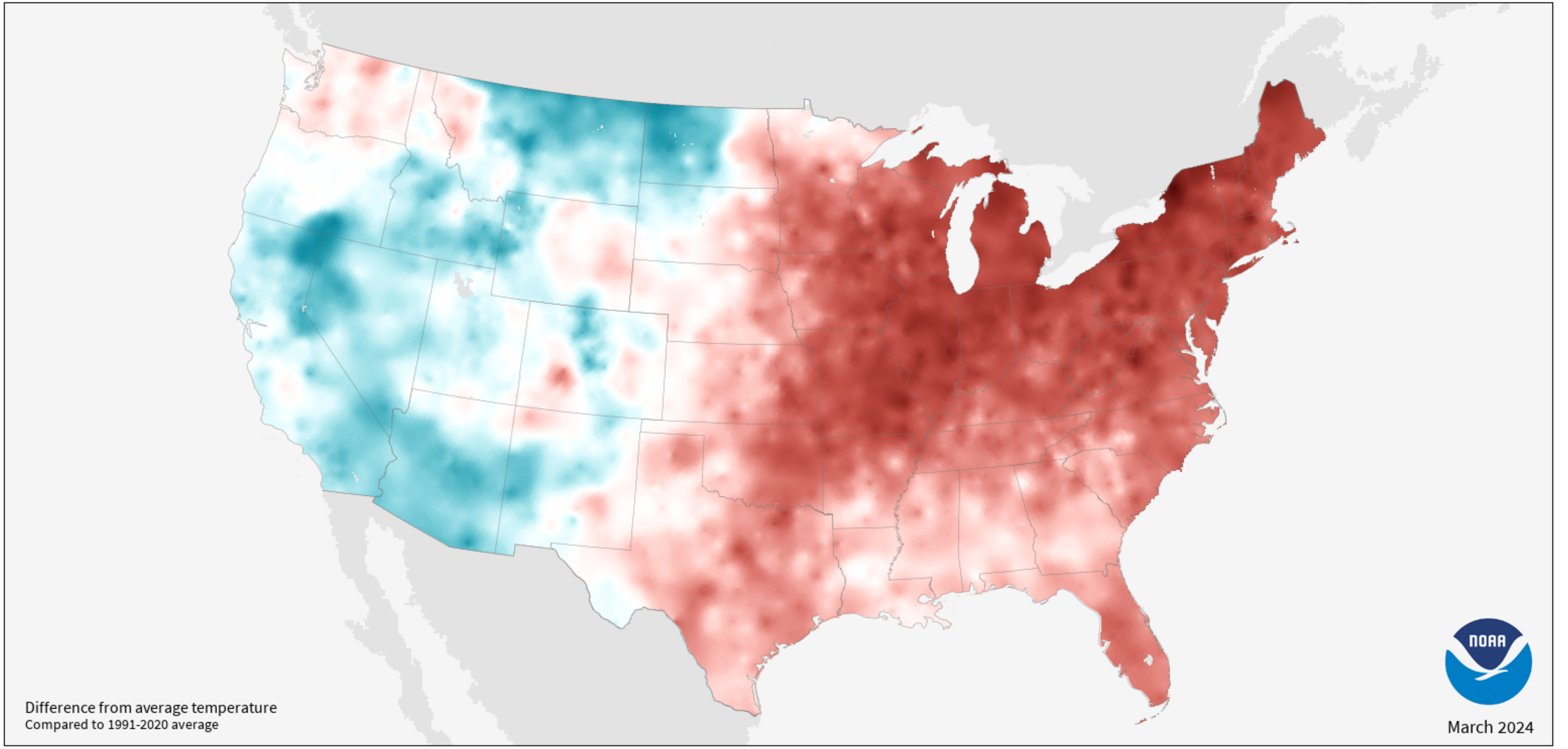
https://www.colour-blindness.com/colour-blindness-tests/ishihara-colour-test-plates/#google_vignette



Crameri, F., Shephard, G. E., Heron, P. J. (2020). The misuse of colour in science communication. *Nature Communications*, 11:5444.



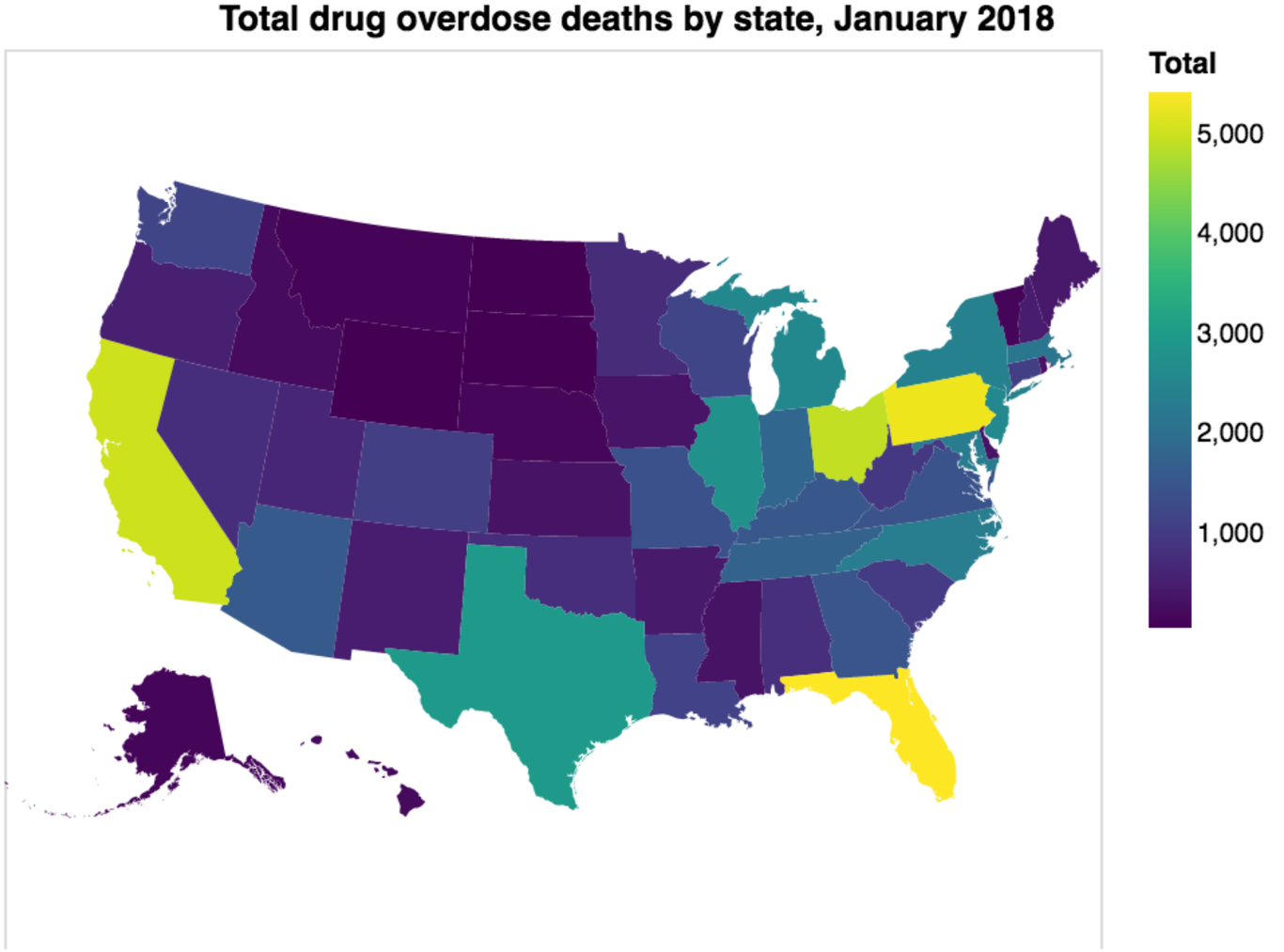
https://www.climate.gov/



https://www.climate.gov/

A growing drug crisis

Drug overdose deaths by US state, Jan 2018



Source: *CDC*

https://www.playfairprize.com/worked-examples

关于颜色的选择：更多信息

- Picking A Colour Scale For Scientific Graphics
  https://betterfigures.org/2015/06/23/picking-a-colour-scale-for-scientific-graphics/
- Scientific colour maps
  https://www.fabiocrameri.ch/colourmaps/
- ColorCET
  https://colorcet.com/
- Beautiful colormaps for oceanography: cmocean
  https://matplotlib.org/cmocean/#why-jet-is-a-bad-colormap-and-how-to-choose-better
- mpl colormaps
  https://bids.github.io/colormap/
- ColorBrewer 2.0
  https://colorbrewer2.org/

应该避免的错误
从《每日经济新闻》汲取教训

媒体中的可视化实践

- *Economist*, *Financial Times*, *Wall Street Journal* 等报纸/杂志在数据可视化实践上走在前列。
- 《每日经济新闻》是成都传媒集团主管主办的报刊，在数据可视化方面也非常积极，但缺乏对细节的追求，存在很多具有代表性的错误。

每日经济新闻

NATIONAL BUSINESS DAILY

中共中央政治局召开会议

分析研究当前经济形势和经济工作

中共中央总书记习近平主持会议

涨3.61%！沪指突破3000点

本周内沪指已上涨近10%，距离5月高点还差约6%

股市持续上涨有两方面原因

以市值管理"硬约束"推动上市公司治理变革

小微企业续贷政策优化调整

奋进强国路 阔步新征程

破净板块昨日掀起涨停潮

02

焦点

中央定调促进房地产市场止跌回稳

对商品房建设要严控增量、优化存量、提高质量

前8个月电信业务收入同比增2.7%

去年全国地表水优良水质断面比例达到89.4%

生态环境部答每经问：已超出"十四五"目标4.4个百分点

奋进强国路 阔步新征程

丰巢需探索更丰富盈利模式

Rokid AR眼镜面临"淘汰赛"

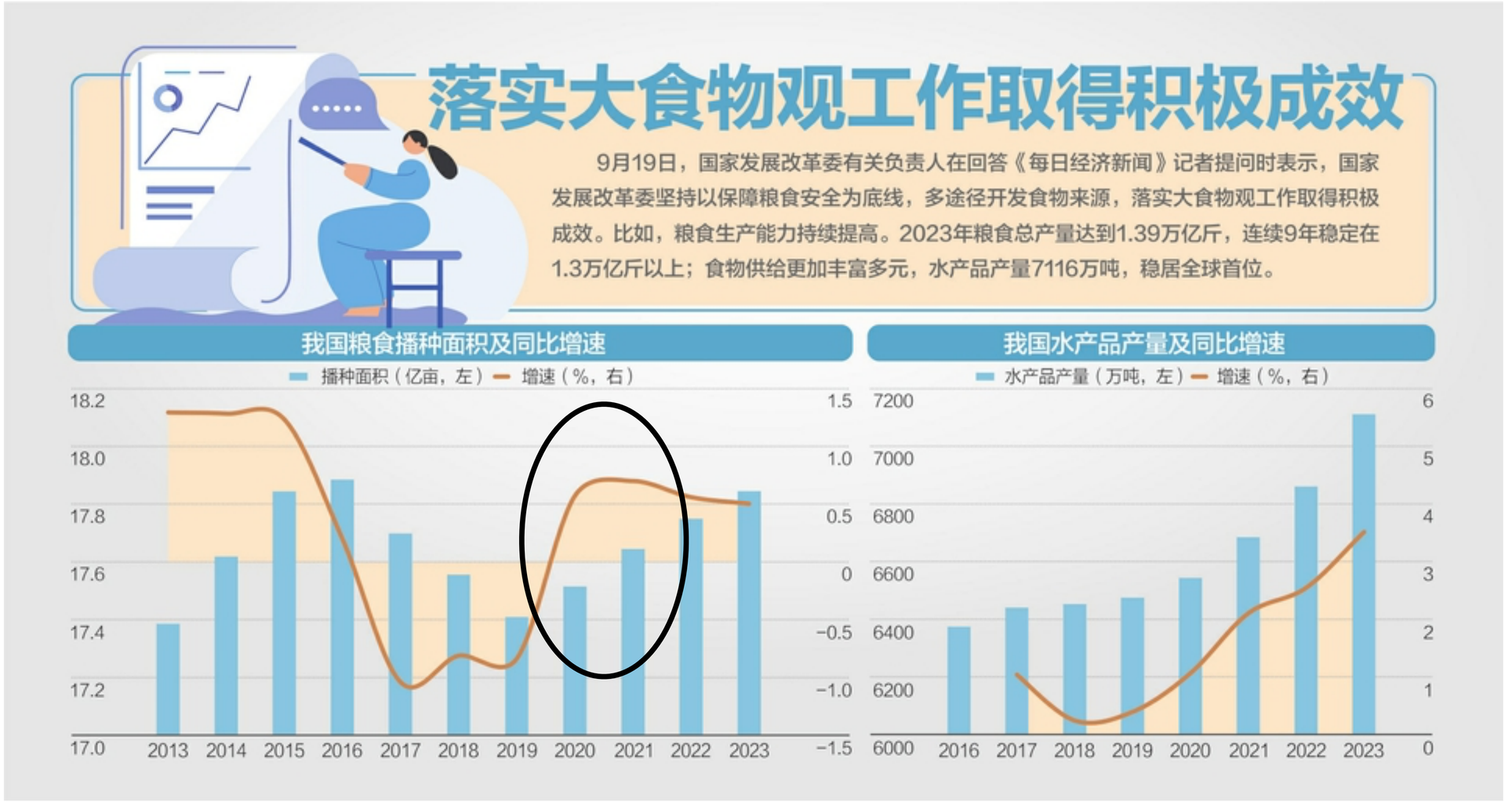
银行信用卡业务，何时改观？

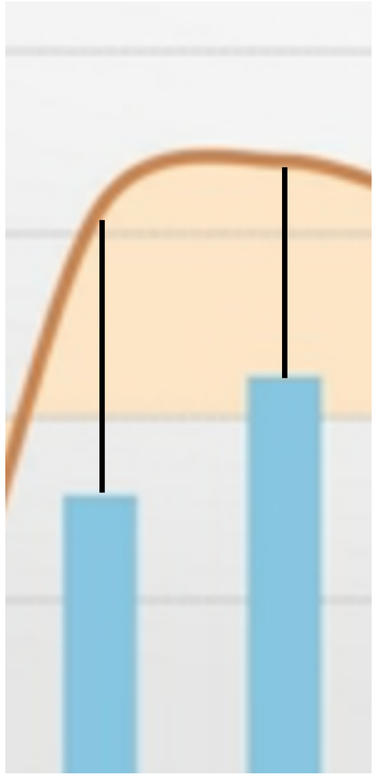
和府捞面门店总数不增反减

2024年9月30日

没有标注数据值

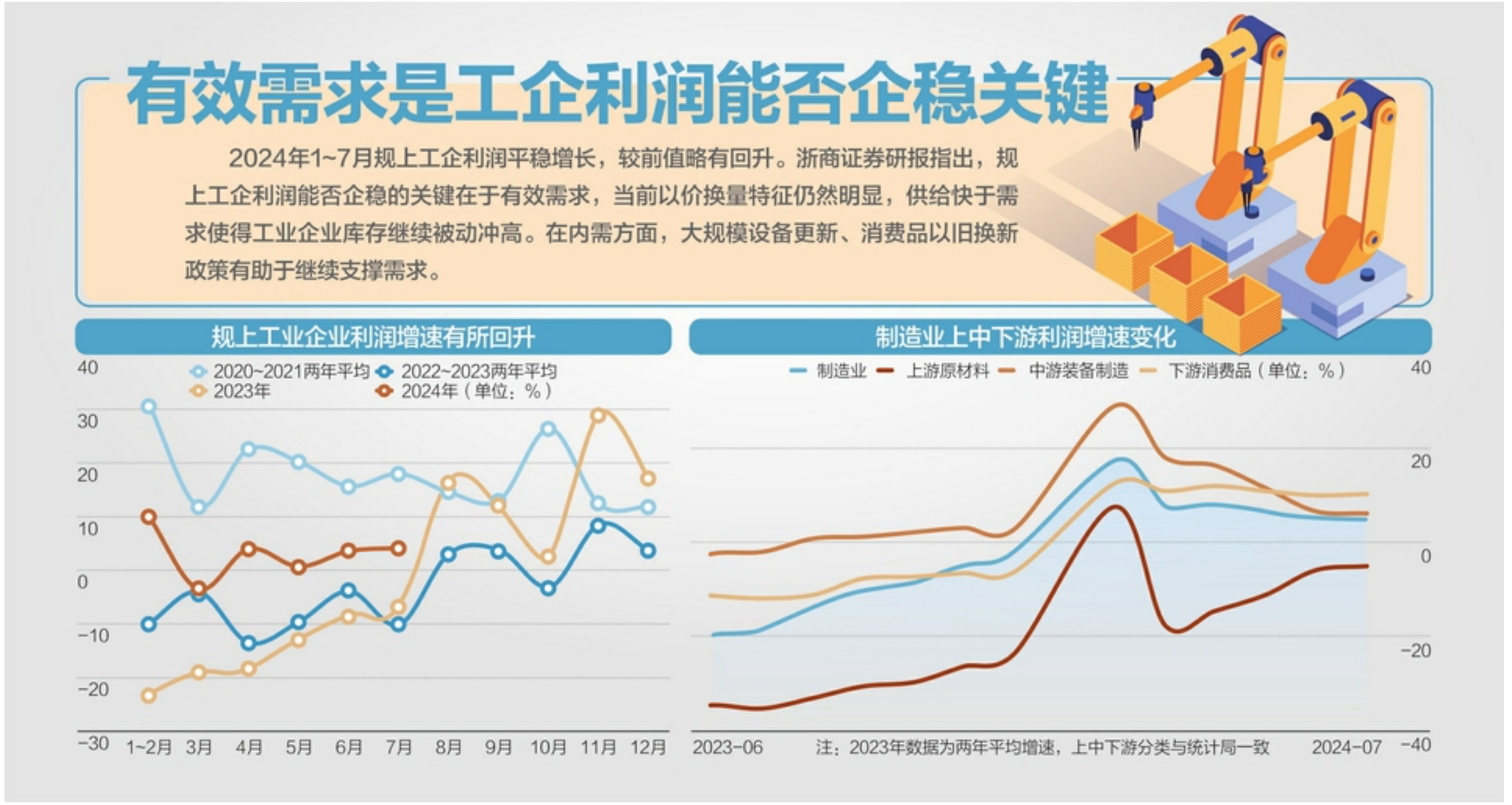
2024年9月20日





1. 平滑的折线图只会起到误导的作用
2. 柱形图的纵坐标应从 0 点开始

2024年8月28日



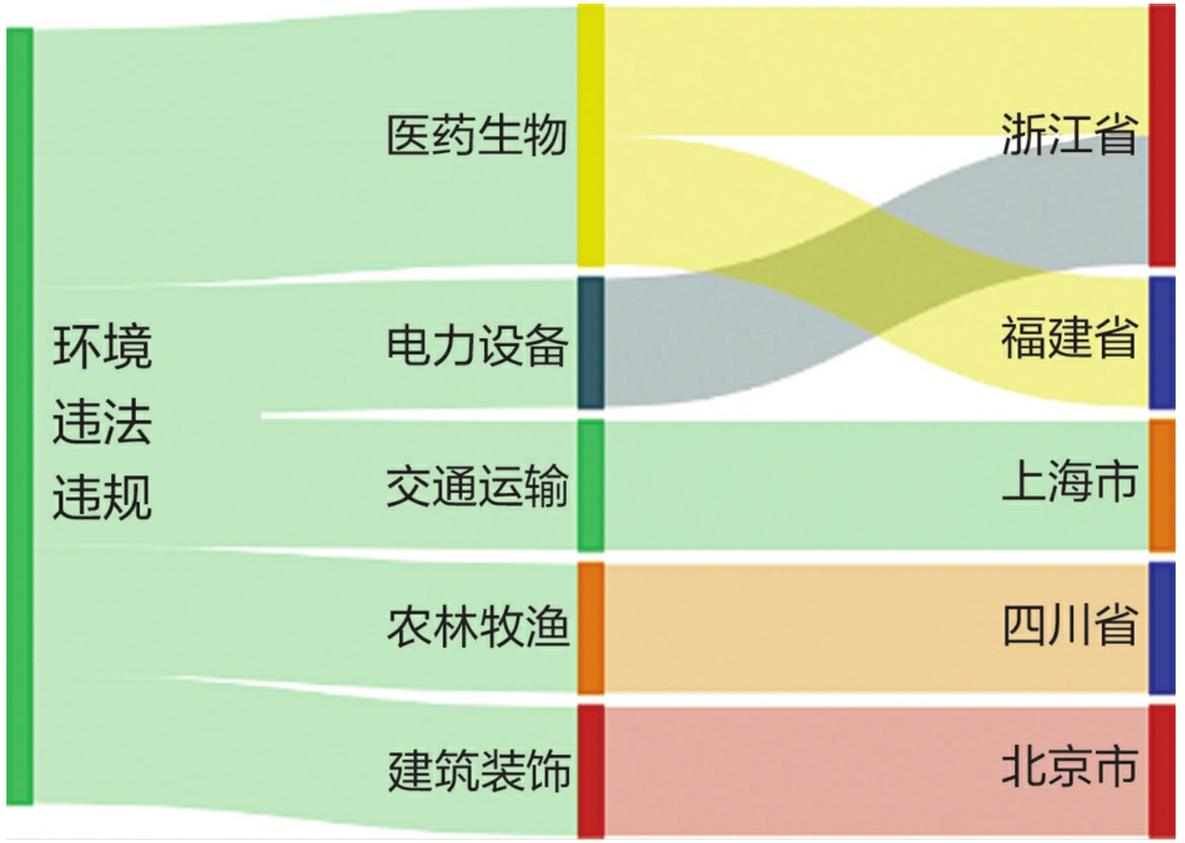
另一个平滑折线图的例子　　横坐标缺失（这是月度数据吗？）

2024年8月19日

每日经济新闻联合环保领域知名NGO（非政府组织）公众环境研究中心（IPE），自2020年9月起，基于31个省区市、337个地级市政府发布的环境质量、环境排放和污染源监管记录等权威数据来源，每周收集剖析中国数千家上市公司及其旗下数万家公司（包括分公司、参股公司和控股公司）的环境信息数据，发布“A股绿色周报”，旨在借助环境数据库及专业解析、传播能力，让上市公司经营活动中的环境信息更加阳光透明。　　根据8月第三周收集到的数据，记者发现，共有6家上市公司在近期暴露了环境风险。

根据8月第三周收集到的数据，记者发现，共有6家上市公司在近期暴露了环境风险。

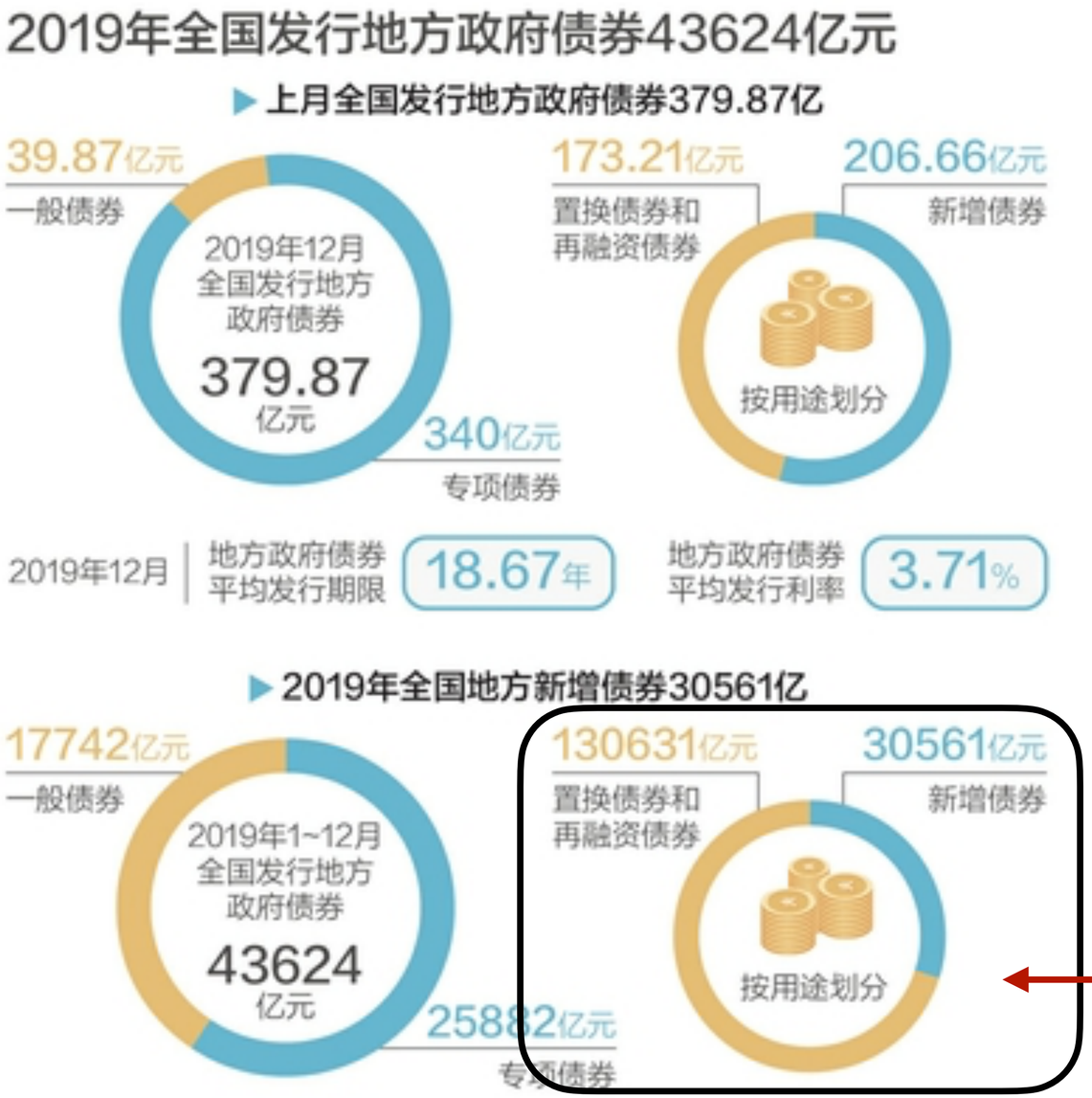
8月第三周环境风险榜涉及公司分布情况

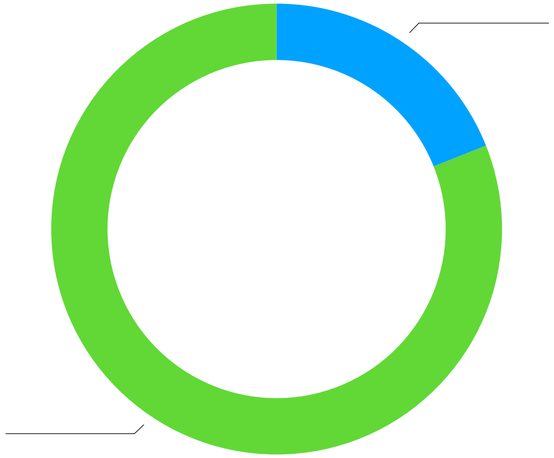


为了画图而画图?

只有六家公司，为什么不用列表?

2020年1月23日



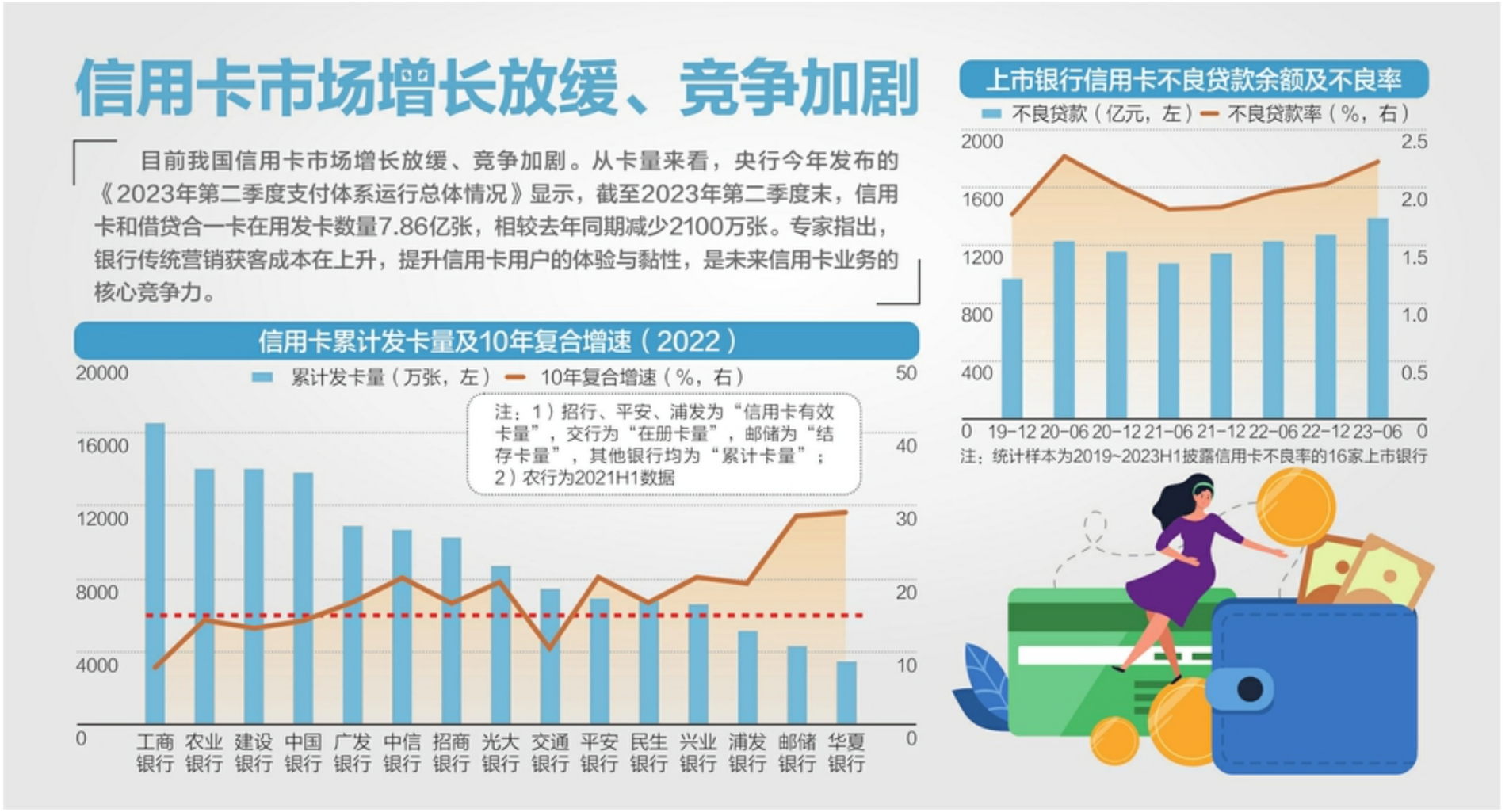


图中的数字和角度不匹配
（饼图的常见错误）

2023年7月18日

1. 横坐标缺失，且季度数据的时间格式应为 2019Q1
2. 左图中的堆叠柱形图应该如何理解?

2023年11月15日



1. 左图为分类数据（不是时序数据），不宜用折线图
2. 没有给出水平虚线的含义