商务智能之数据可视化

第一讲：课程介绍

课程信息

- 参考书籍
  - “数据可视化”，陈为，沈则潜，陶煜波著，2013（2019第二版）
  - “Python数据分析与可视化，从入门到精通”高博，刘冰，李力，2020
  - “*The Visual Display of Quantitative Information*,” Edward Tufte, 2001
  - “*Visualization Analysis & Design*,” Tamara Munzner, 2014
  - *“Interactive Data Visualization for the Web”,* Scott Murry, 2013

课程大纲

- 可视化基本介绍
- 可视化基础
  - 视觉感知
  - 数据
- 可视化应用
  - 多维数据可视化
  - 文本数据可视化
  - 时空数据可视化
  - 层次及网络数据可视化

第一讲：可视化基本介绍

大纲

- 什么是可视化
- 为什么要可视化
- 可视化研究挑战

数据探索

- 难点：不确定我们在找什么（直到我们找到为止）
- **数据探索案例**：Antibiotics, Will Burtin, 1951

https://gist.github.com/borgar/cd32f1d804951034b224

数据

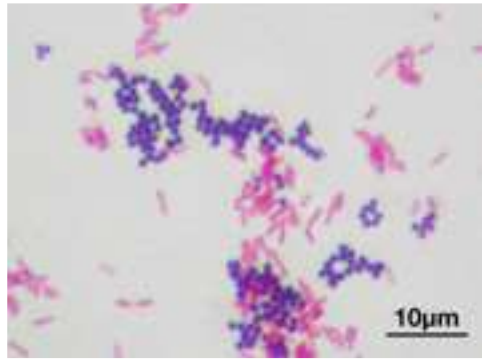


Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

基于这个数据，可以提出什么问题？

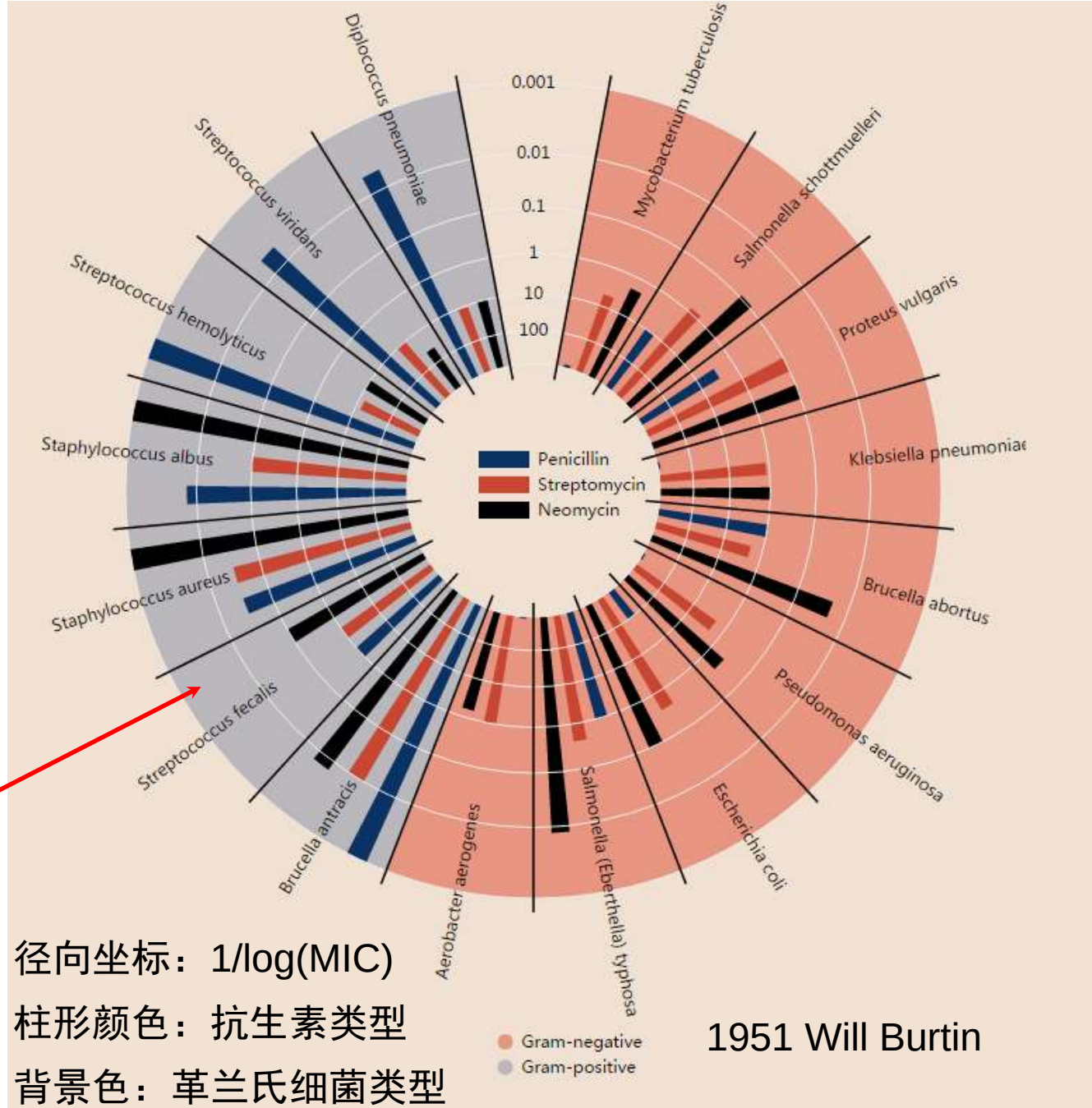
Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

How effective are the drugs?

- If bacteria is gram positive, Penicillin（青霉素） & Neomycin（新霉素） are most effective
- If bacteria is gram negative, Neomycin is most effective

玫瑰图



径向坐标：1/log(MIC)

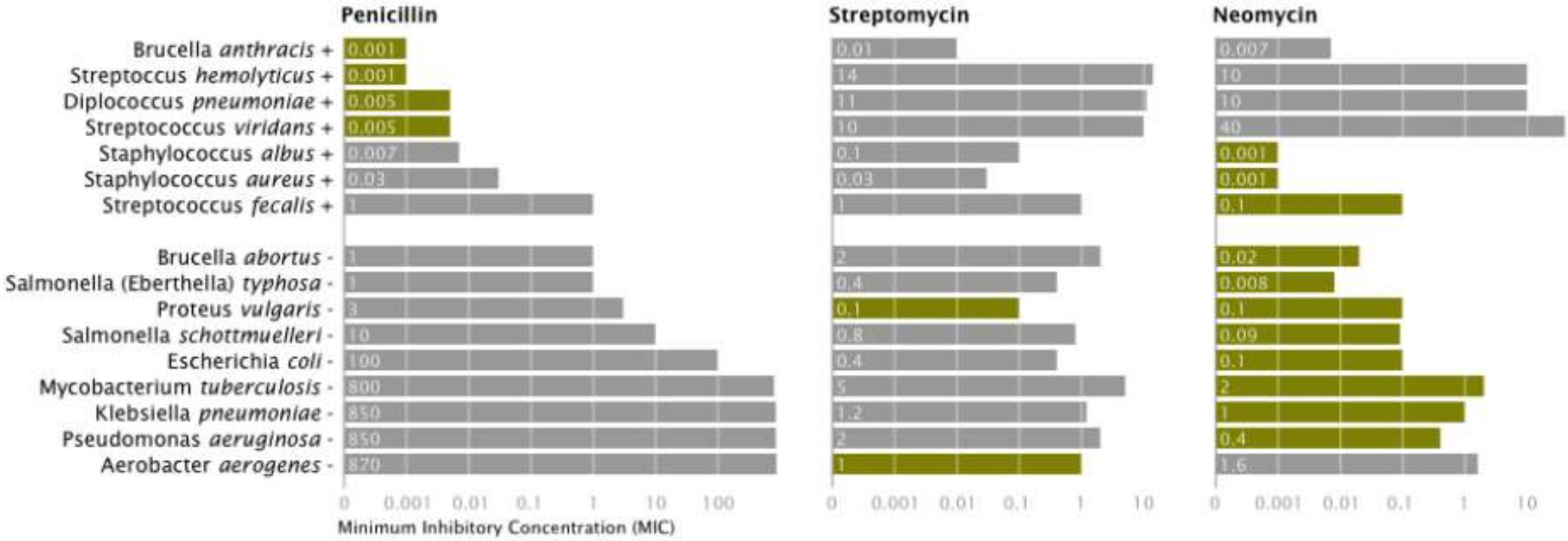
柱形颜色：抗生素类型

背景色：革兰氏细菌类型

1951 Will Burtin

http://mbostock.github.io/protovis/ex/antibiotics-burtin.html

How do the drugs compare?

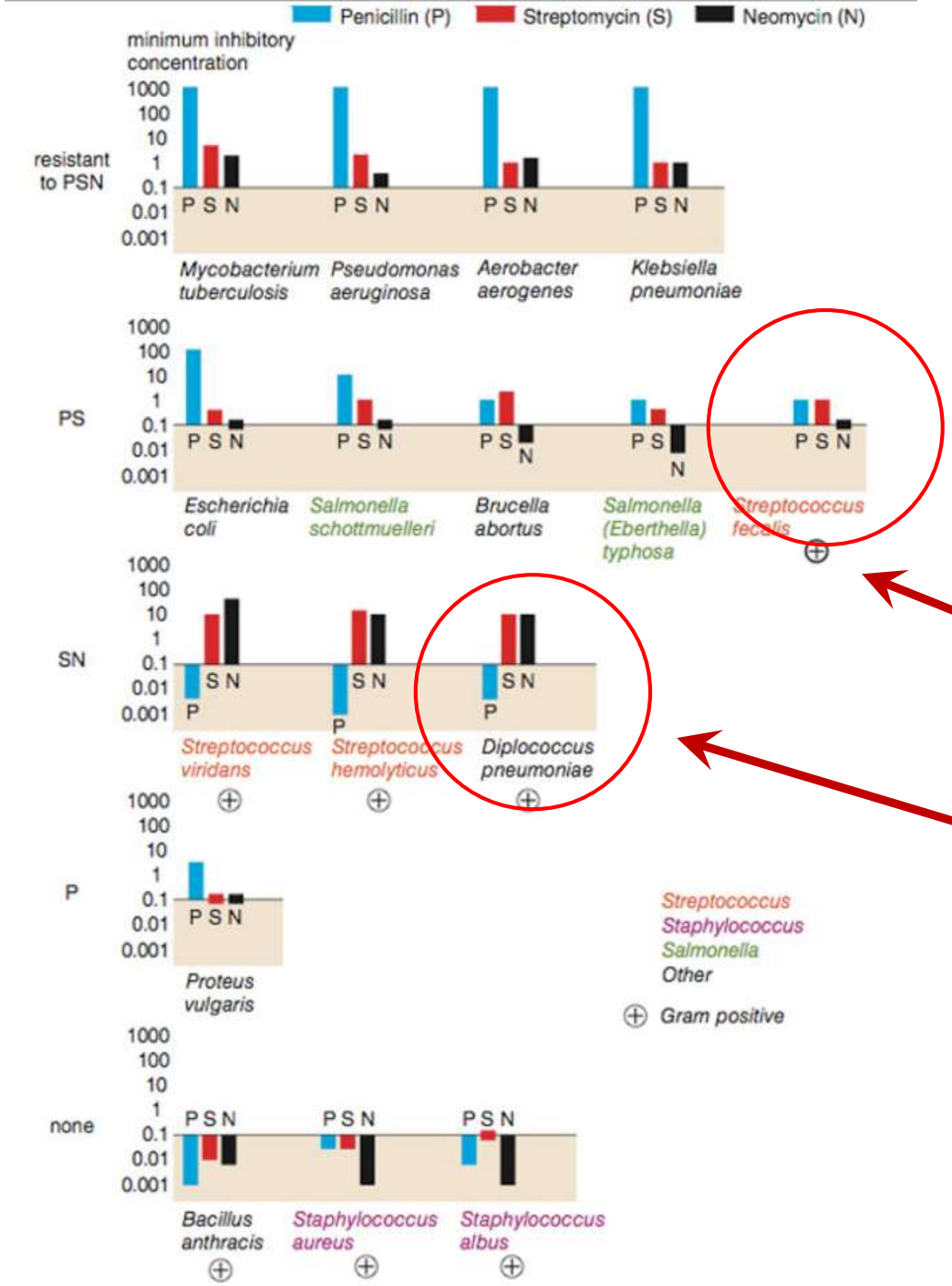


横坐标：log(MIC)

纵坐标：细菌名称+革兰氏细菌类型

颜色：突出抗生素有效性

Mike Bostock
Stanford CS448B, Winter 2009

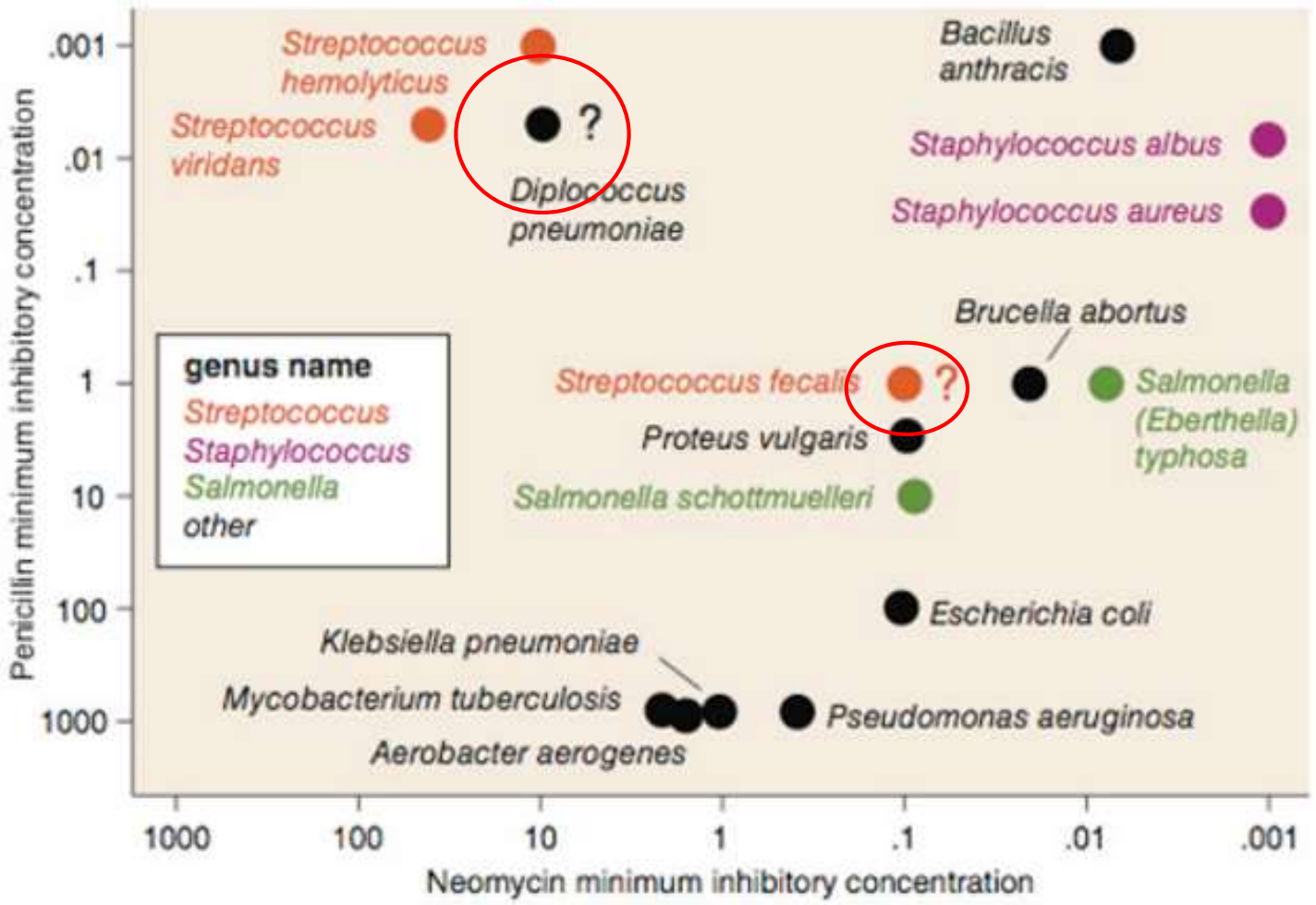


How do the bacteria compare?

Not a streptococcus! (realized ~30 years later, 1984), actually enterococcus faecalis.

Really a streptococcus! (realized ~20 years later, 1974)

How do the bacteria compare?



Wainer & Lysen, “That’s funny...” American Scientist, 2009

https://www.americanscientist.org/article/thats-funny

The most exciting phrase to hear in science, the one that heralds new discoveries, is not “Eureka” but “That’s funny...”

- Isaac Asimov (1920–1992)

什么是可视化

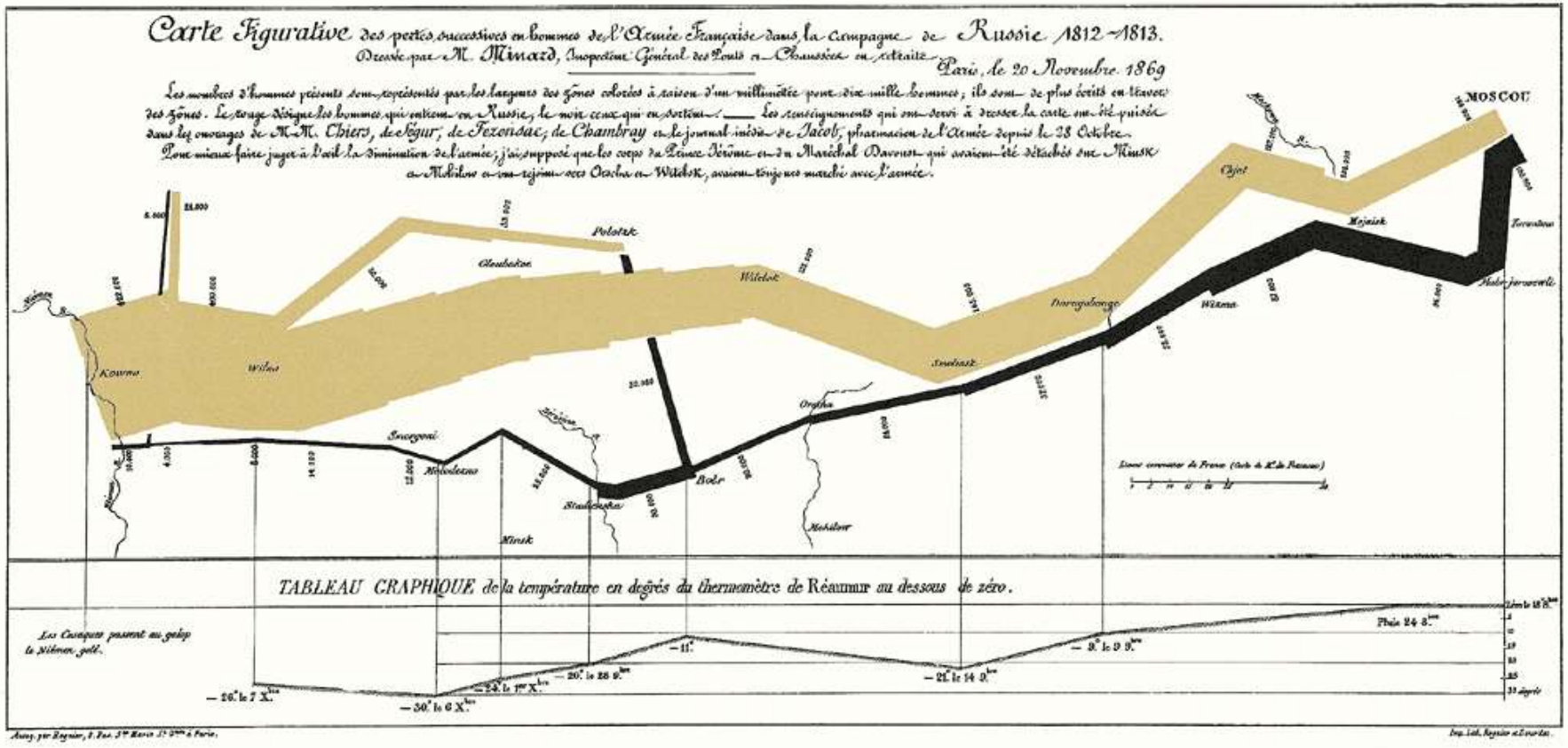
- 将数据以图形的方式呈现，从而更好地传达信息（帮助人们理解）

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

- John Tukey

什么是可视化

- 利用人眼感知能力对数据进行交互的可视表达以增强认知的技术



流型图：1812–1813年进军莫斯科的历史事件

Charles Minard, 1869

数据可视化

- 创建并研究数据的视觉表达（visual Representation）
  - 输入：数据
  - 输出：视觉形式
  - 目标：深入理解
- 主要任务
  - 表示数据
  - 分析数据
  - 交流数据

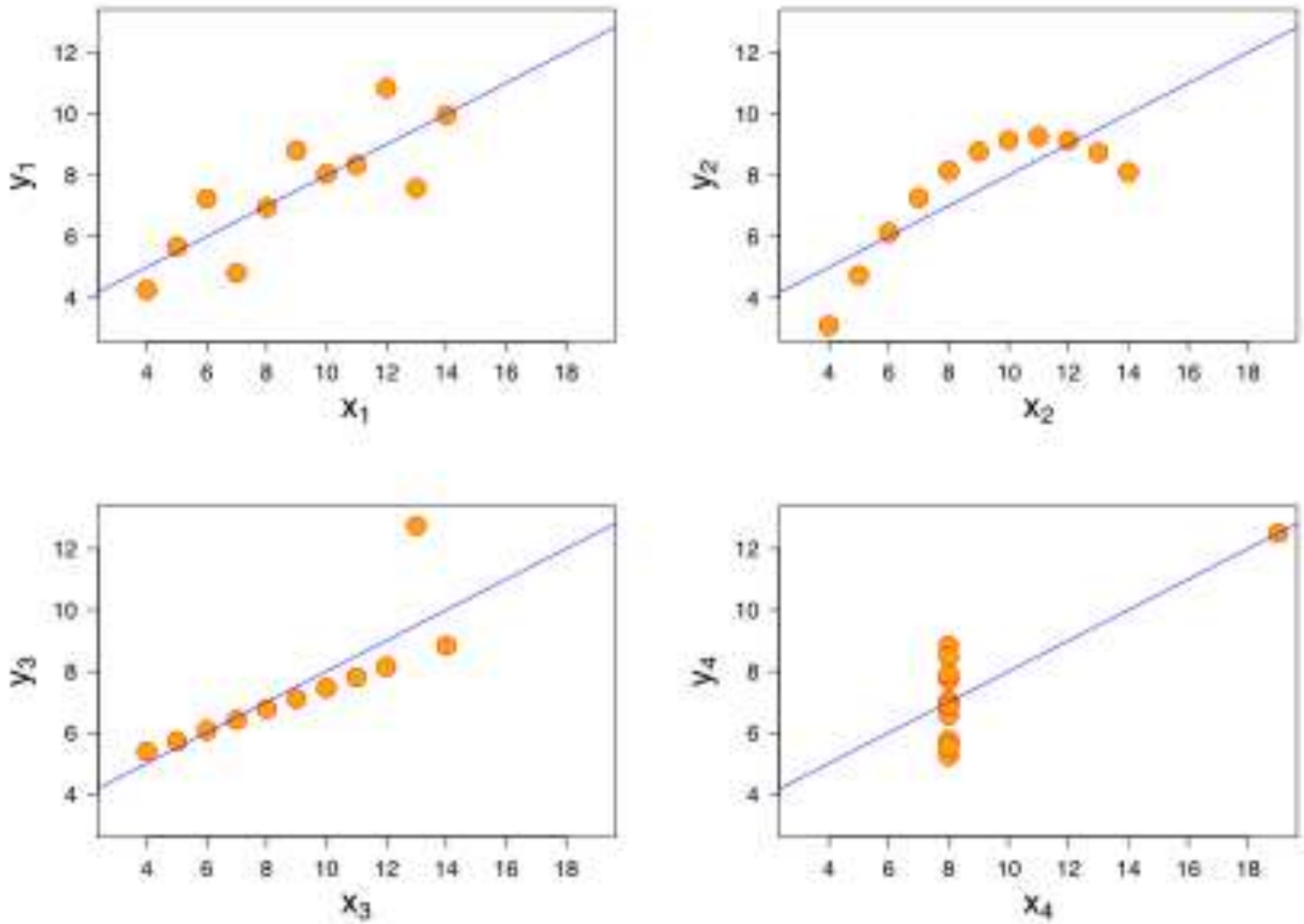
为什么要可视化：安斯库姆四重奏

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

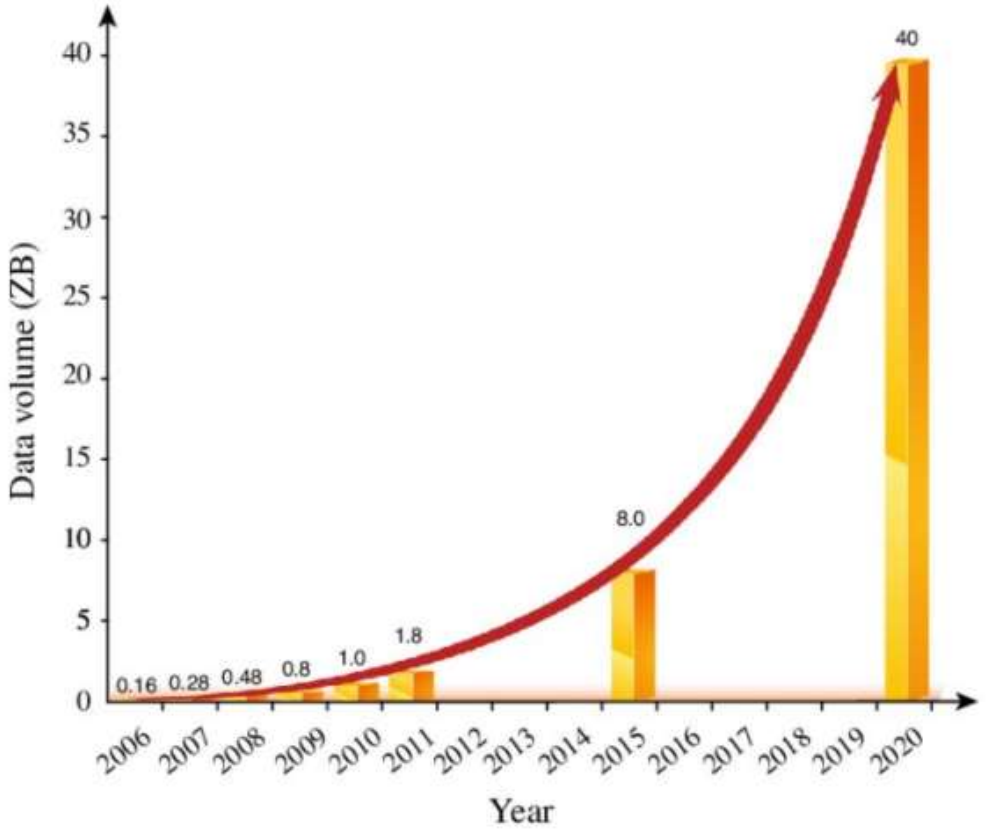
https://en.wikipedia.org/wiki/Anscombe_‘s_quartet

描述统计

	I	II	III	IV
Mean(x)	9.00	9.00	9.00	9.00
Mean(y)	7.50	7.50	7.50	7.50
Var(x)	11.00	11.00	11.00	11.00
Var(y)	4.13	4.13	4.12	4.12
Correlation	0.82	0.82	0.82	0.82
Lm intercept	3.00	3.00	3.00	3.00
Lm x effect	0.50	0.50	0.50	0.50


y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18

为什么需要可视化



Pics copied from Guo, H., Wang, L., Chen, F., & Liang, D. (2014). Scientific big data and digital earth. Chinese Science Bulletin, 59(35), 5066-5073.

为什么要可视化

- 信息爆炸
- 科学数据爆炸

将来几十年中，处理数据的能力将会成为至关重要的技术——理解数据、加工数据、提取数据价值、可视化数据、与数据交流。…因为现在我们的确拥有无处不在的、可自由获取的数据。”

“The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it —that’s going to be a hugely important skill in the next decades,... because now we really do have **essentially free and ubiquitous data**.”

Hal Varian,Google’s Chief Economist
The McKinsey Quarterly, Jan 2009

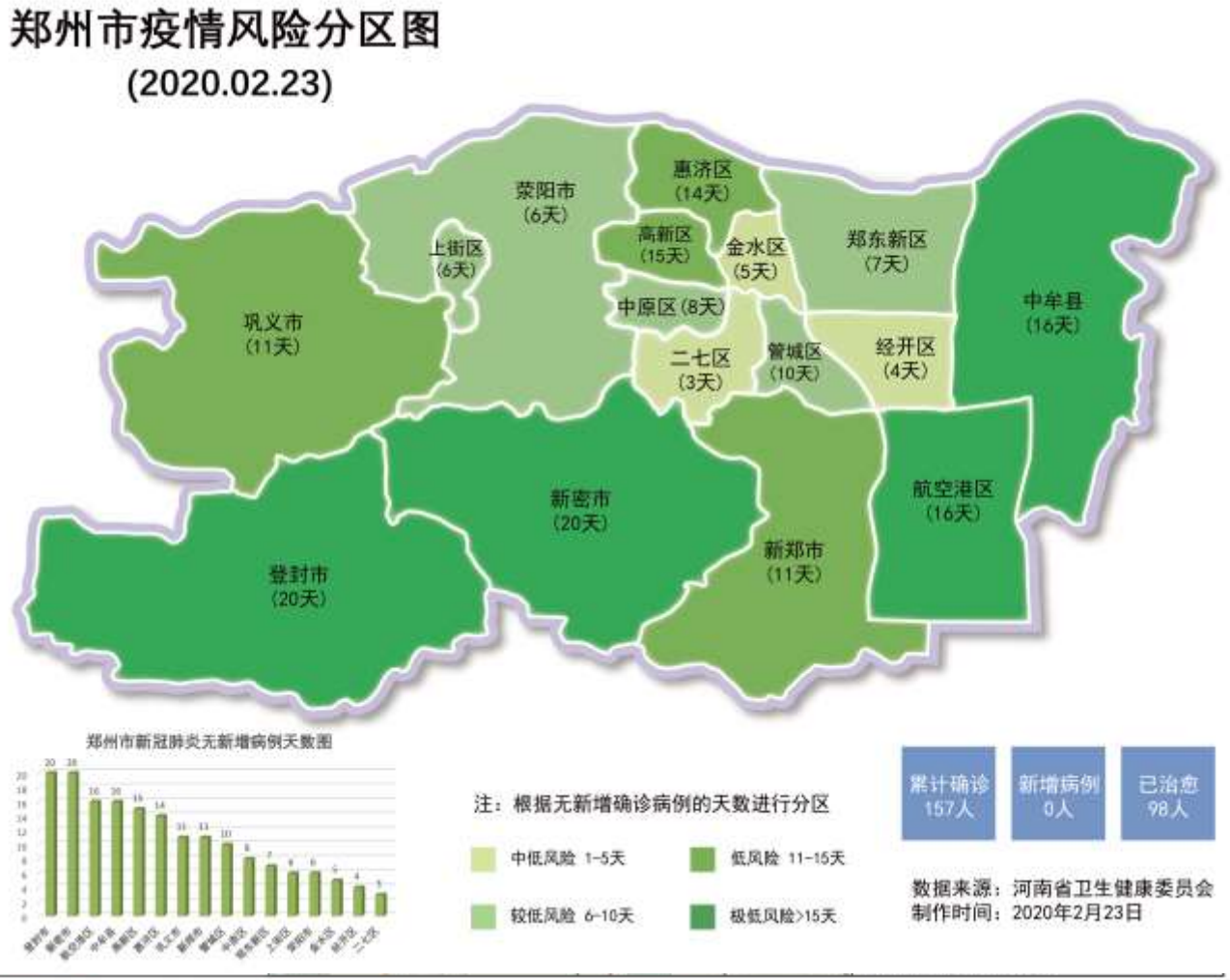
为什么需要可视化

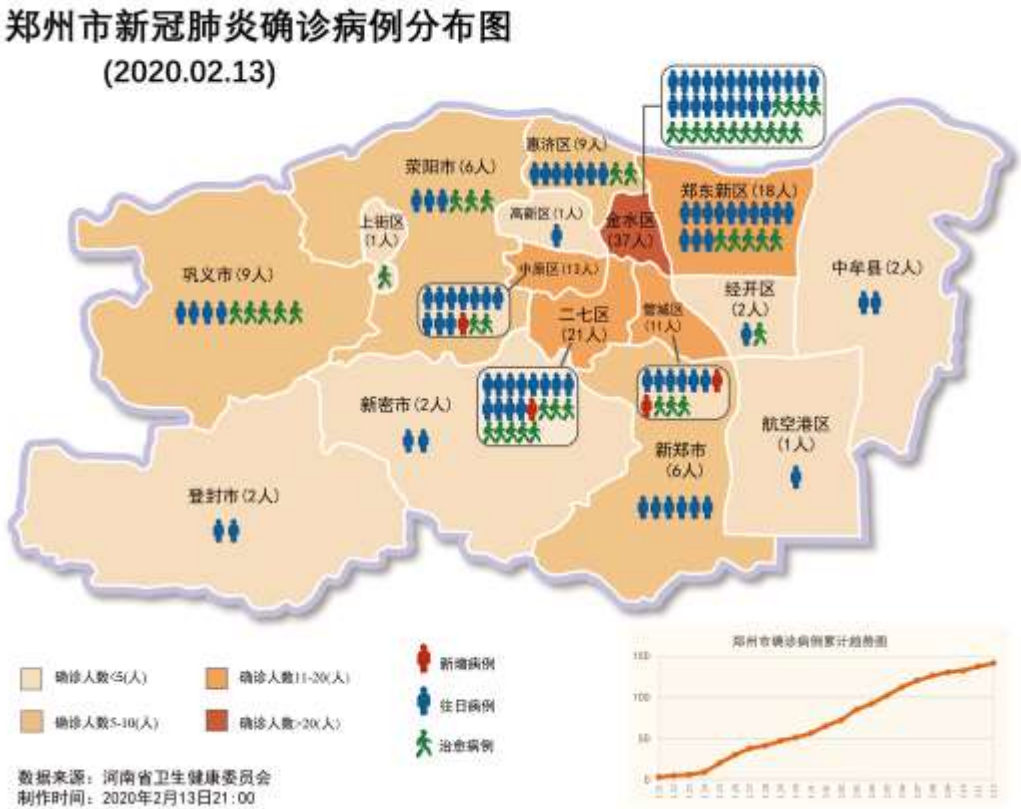
- 中国科技创新2030“新一代人工智能”和 “大数据”专项均将可视化和可视分析列为大数据智能急需突破的关键共性技术
- 2008年后，美、欧盟、日均成立国家可视分析研究中心。国内外著名企业成立独立部门，研发新兴可视化与可视分析技术
  - 包括阿里、百度、amazon、google、360等
- 2018年11月19日，美国商务部工业与安全局（BIS）拟定14项针对中国的技术管制，可视化技术（数据分析）为其中之一

- ***可视化成为基础自然科学研究的必要手段，是科学大数据发展的必需***

http://zchlisc.net/a/xinwenzhongxin/xingyexinwen/20190413/37.html, 2021-03-01

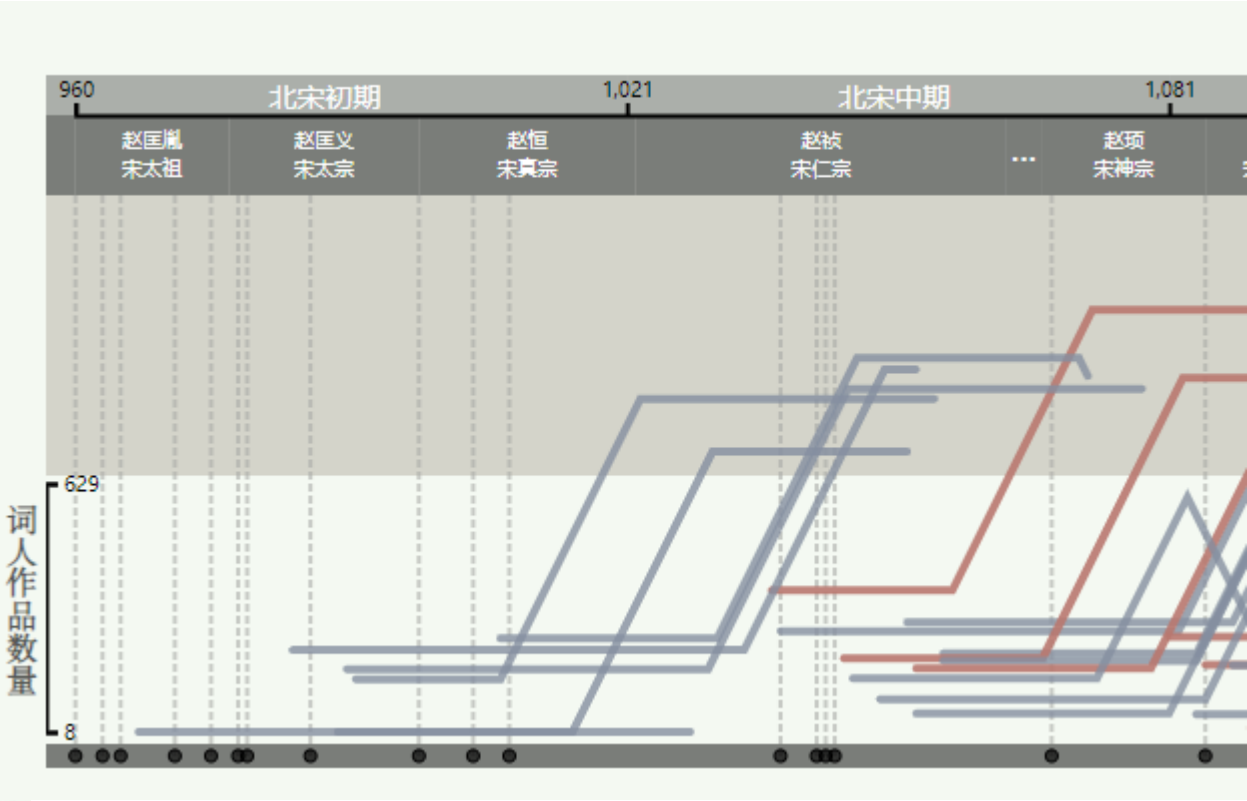
2020年新冠肺炎疫情可视化





https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.54.4df556b6ANuWu6&postId=98818, 2022-02-21

宋词文化可视化

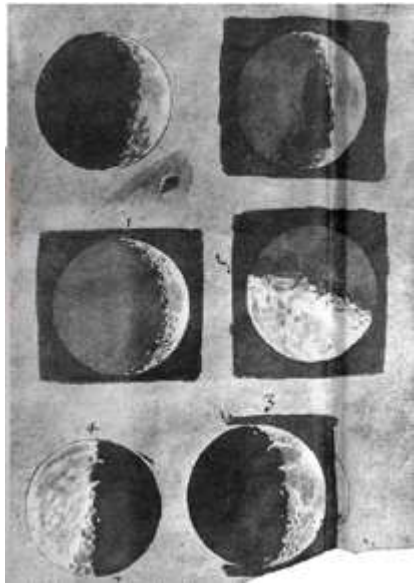


宋代词人生平及所处年代图谱

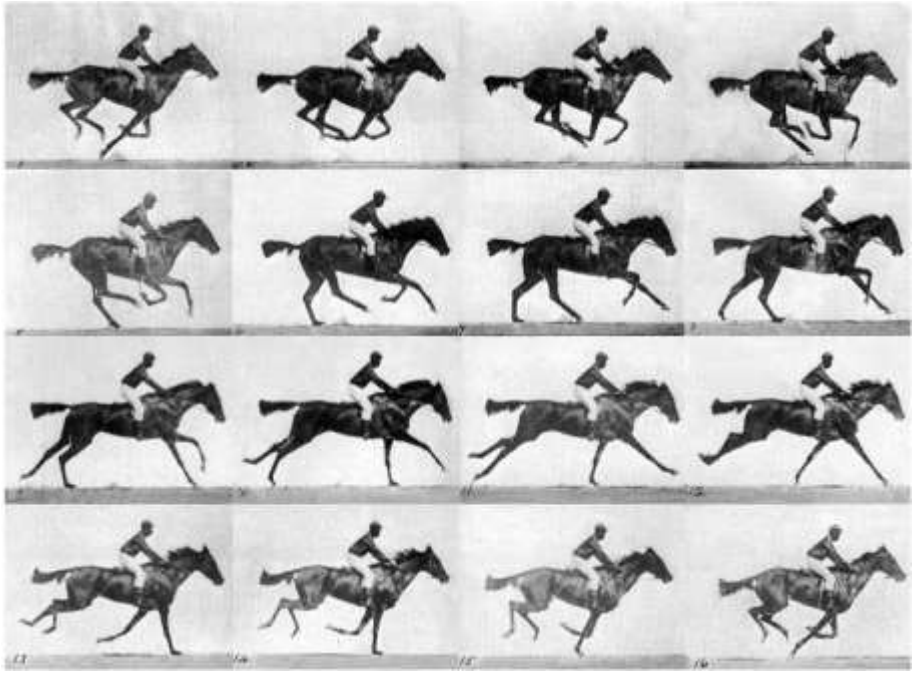
http://fms.news.cn/swf/2018_sjxw/quansongci/index.html#/, 2021-02-21

可视化的作用

- 记录信息
  - 成像、蓝图设计…
- 分析推理
  - 证实猜想
  - 反馈与交互
  - 过程与计算
- 信息传播与协同
  - 共享与说服
  - 协作与修正
  - 突出数据的重要部分



记录信息：Galileo Galilei,1616

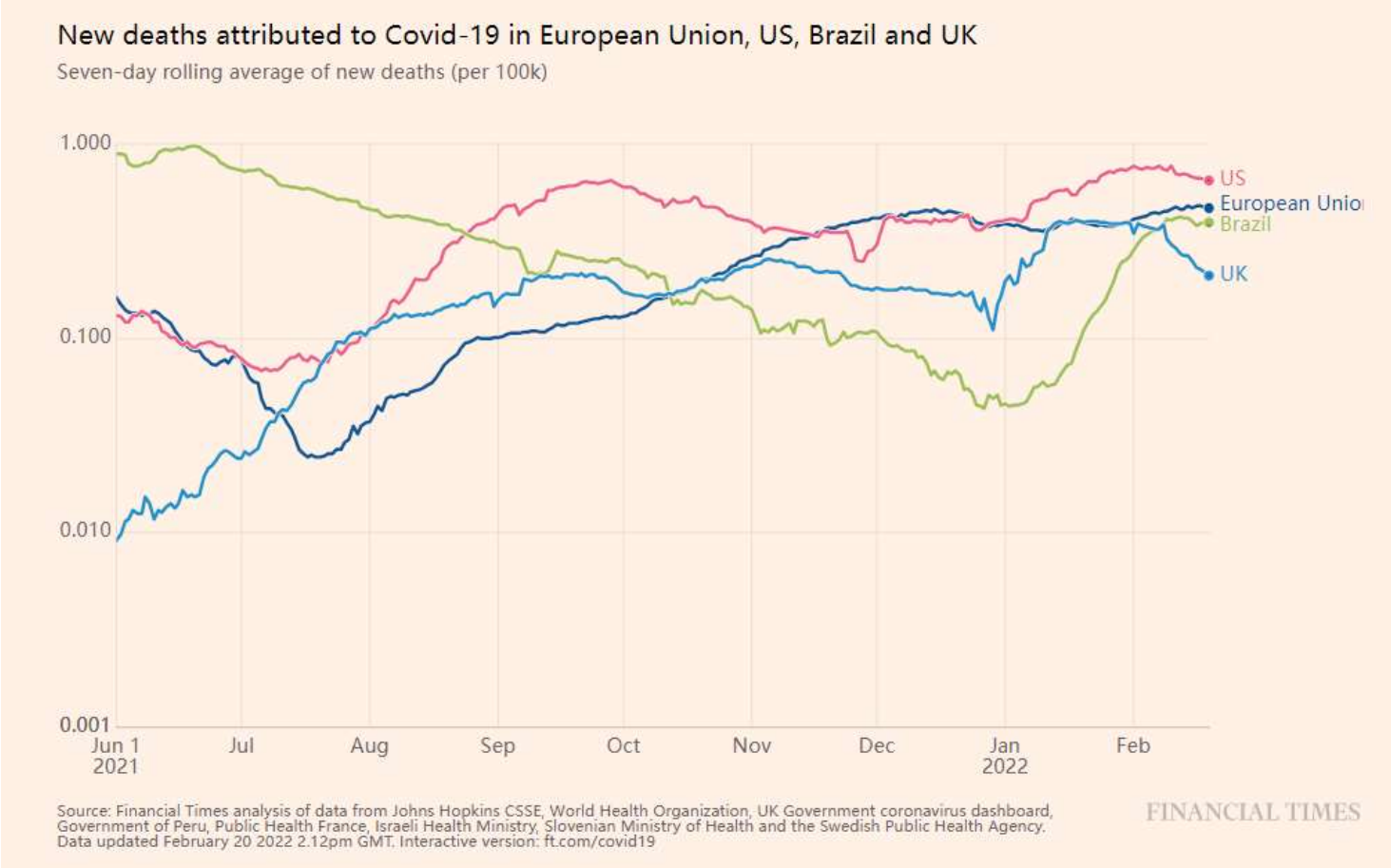


记录信息：Gallop, Bay Horse “Daisy” [Muybridge,1884-86]

可视化的作用：证实假设

分析推理：Ghost map, John Snow 1854
欧洲霍乱

可视化的作用：信息传播与协同



可视化的作用：信息传播与协同

浙大可视化团队与央视合作，民众关于脱欧观点分析，2019年

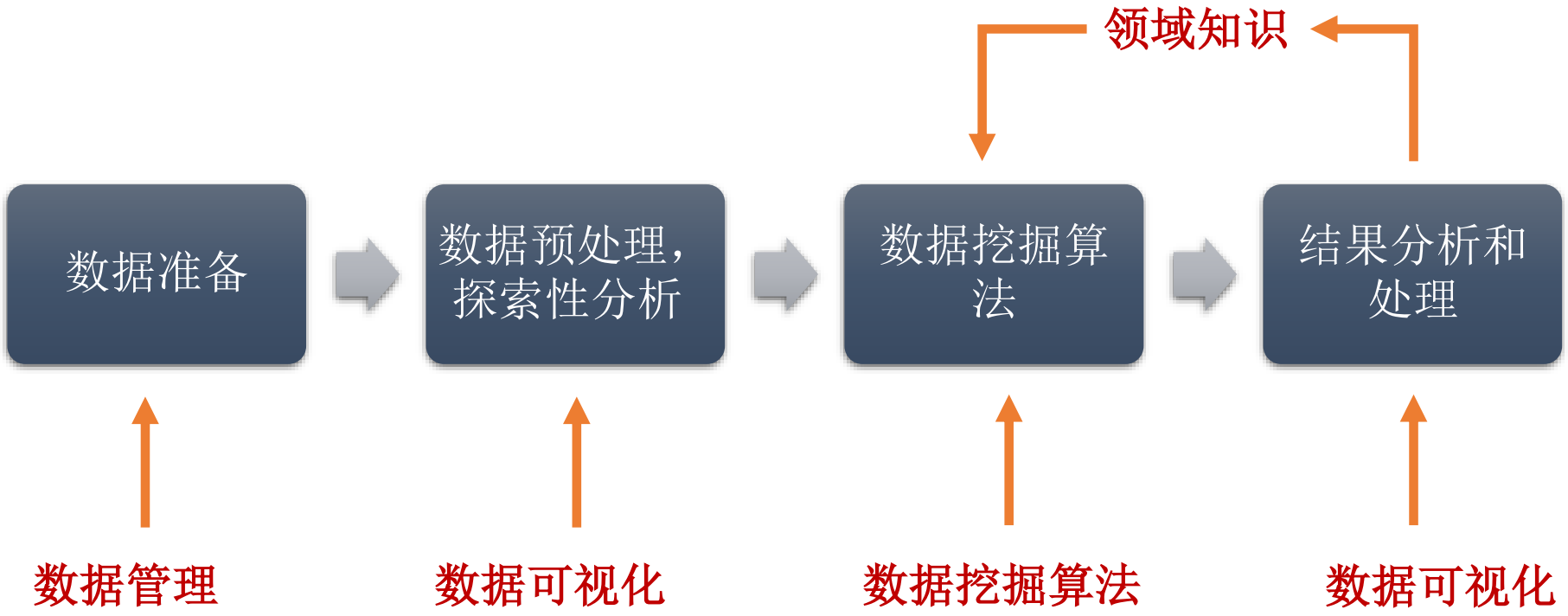
可视化研究挑战

- 可视化的三个约束
  - 计算能力的可拓展性
  - 感知和认知能力的局限性
  - 显示能力的局限性
- 核心挑战：如何利用人的感知能力，增强有限的认知能力，以应对理解和分析复杂数据的迫切需求
  - 大数据可视化
  - 以人为中心的探索式可视分析

可视化+

- 大数据的每个环节与可视化融合

数据分析与可视化基本流程



课程结构

Part 1
什么是可视化
可视化的重要性

Part 2
可视化的理论基础

- 视觉感知
  - 格式塔理论八大原则
  - 可视化编码：视觉通道
- 数据基础
  - 数据与可视化密不可分
  - 可视化贯穿每一个环节
- 图形完整性六原则
  图形重新设计：Lie factor，Data-ink ratio，Data density

Part 3
可视化应用及基于python的数据可视化实操

Python可视化工具介绍
基本图形及复杂图形的可视化
时空数据可视化
文本数据可视化
网络和层次数据可视化