

курс «Прикладные задачи анализа данных»

Искусство визуализации

Часть 1. Историческая

Александр Дьяконов

18 сентября 2020 года

План

Зачем смотреть на данные

История визуализации и инфографики

Правила визуализации

Одномерный анализ

Описательные статистики, их визуализации

Первичные действия при анализе признака

Визуализация отдельных признаков

Многомерный анализ

Визуализация пары признаков

Визуализация «алгоритм» – «алгоритм/признак»

3D-визуализации

Dummy-визуализации

Игра «Что изображено?»

Зачем смотреть на данные?

«The greatest value of a picture is when it forces us to notice what we never expected to see»



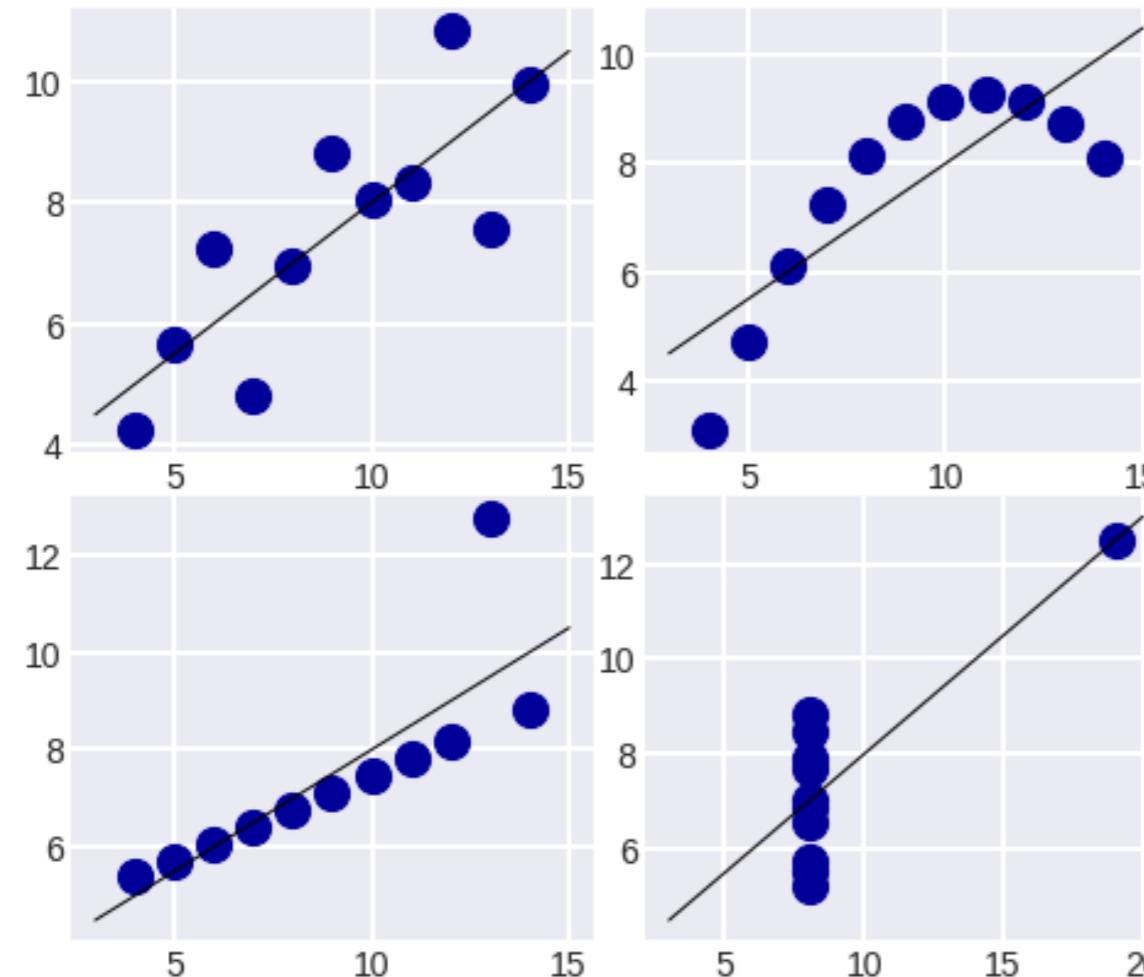
John Tukey

Анализ данных, прежде всего, визуальный анализ!

Визуализация сложнее, чем кажется!

Зачем смотреть на данные?

Наборы данных имеют идентичные статистические характеристики, но их графики существенно различаются.



Характеристика	Значение
Среднее значение переменной X	9
Дисперсия переменной X	10
Среднее значение переменной Y	7.5
Дисперсия переменной Y	3.75
Корреляция между переменными	0.816
Прямая линейной регрессии	$Y=3+X/2$

F.J. Anscombe Graphs in Statistical Analysis // American Statistician, 27 (February 1973), 17-21.

Цели визуализации – где применяется

- «разведочный анализ данных»
(EDA = Exploratory Data Analysis)
понимание данных через их иллюстрации
интересует нас прежде всего
- иллюстрация слов
в широком смысле: доказательства, аргументация допущений,
рассуждения
облегчение восприятия формул и слов
- рассказ (story-telling)
поясняющие иллюстрации для достижения определённой цели
ex: продать продукт

Цели визуализации – для чего используется

- **нахождение закономерностей**
- **детектирование наличия выбросов / аномалий**
- **проверка данных на логичность, полноту и т.п.**
- **нахождение ошибок сбора и предобработки**
- **придумывание признаков**
(деформаций, комбинаций, индикаторов)
- **проверка статистических допущений**
- **понимание смысла задачи, проверка / выдвижение гипотез и предварительный выбор модели**

Надо изучить предметную область!

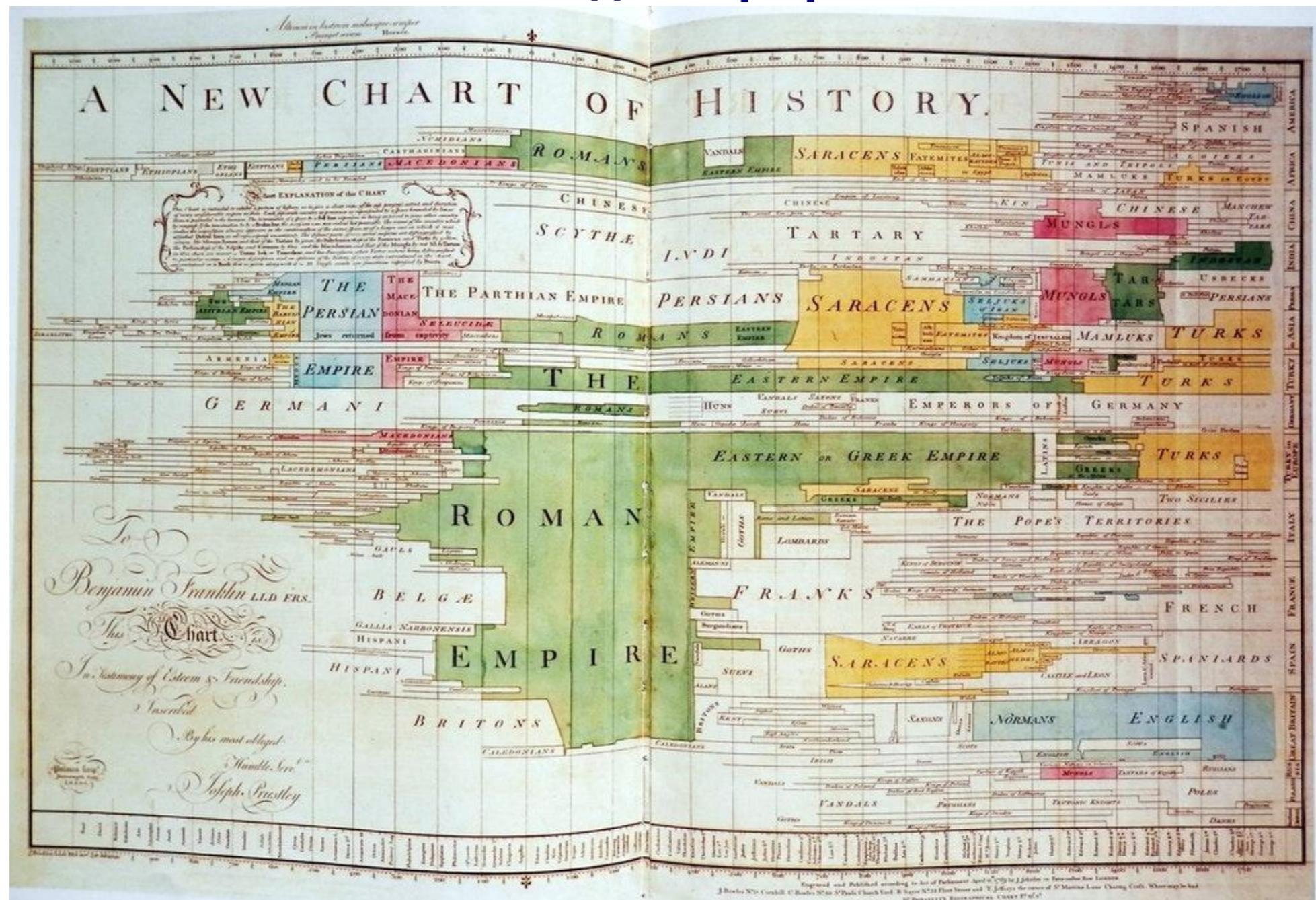
Надо обязательно смотреть на данные / модель!

Немного об истории визуализации

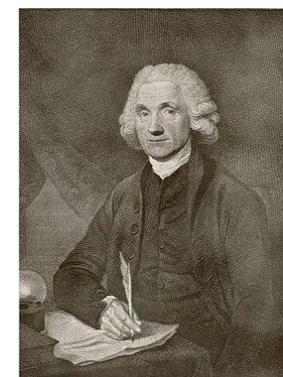
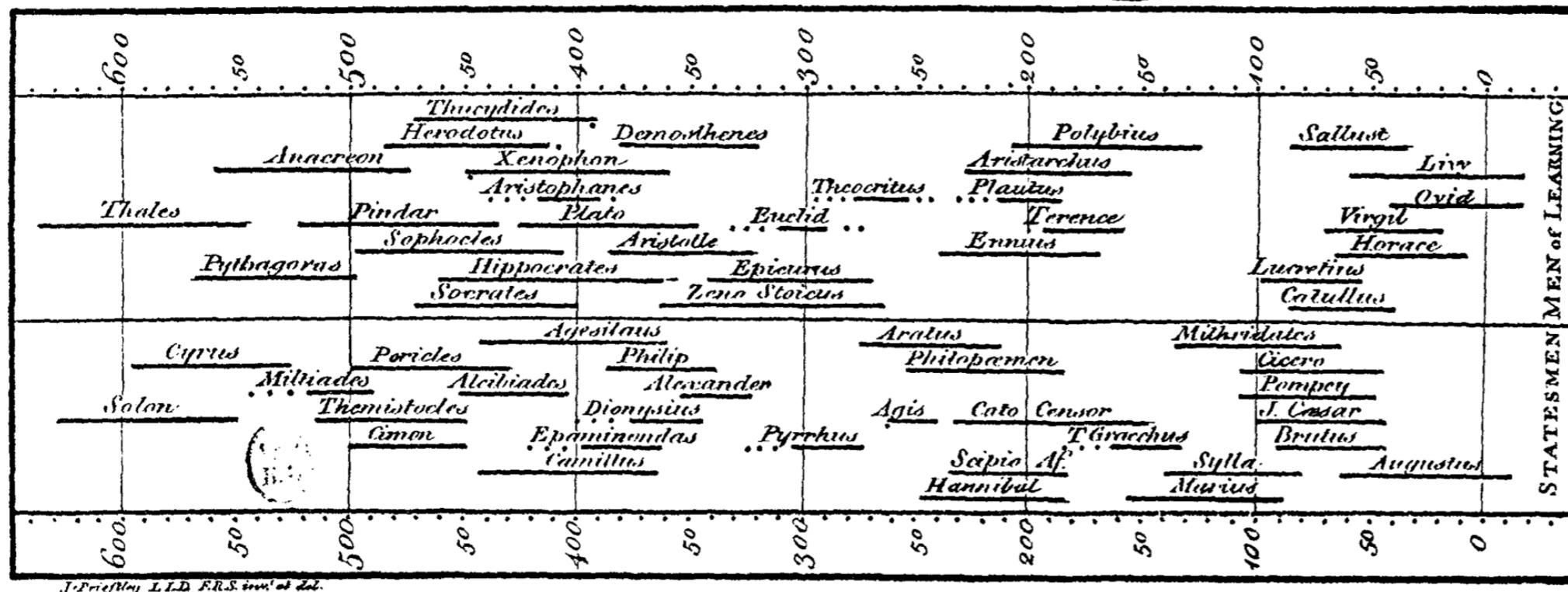
инфографика
виды графиков
графический анализ

18 век – зарождение инфографики
19 век –protoанализ данных

18 век Джозеф Пристли



18 век Джозеф Пристли

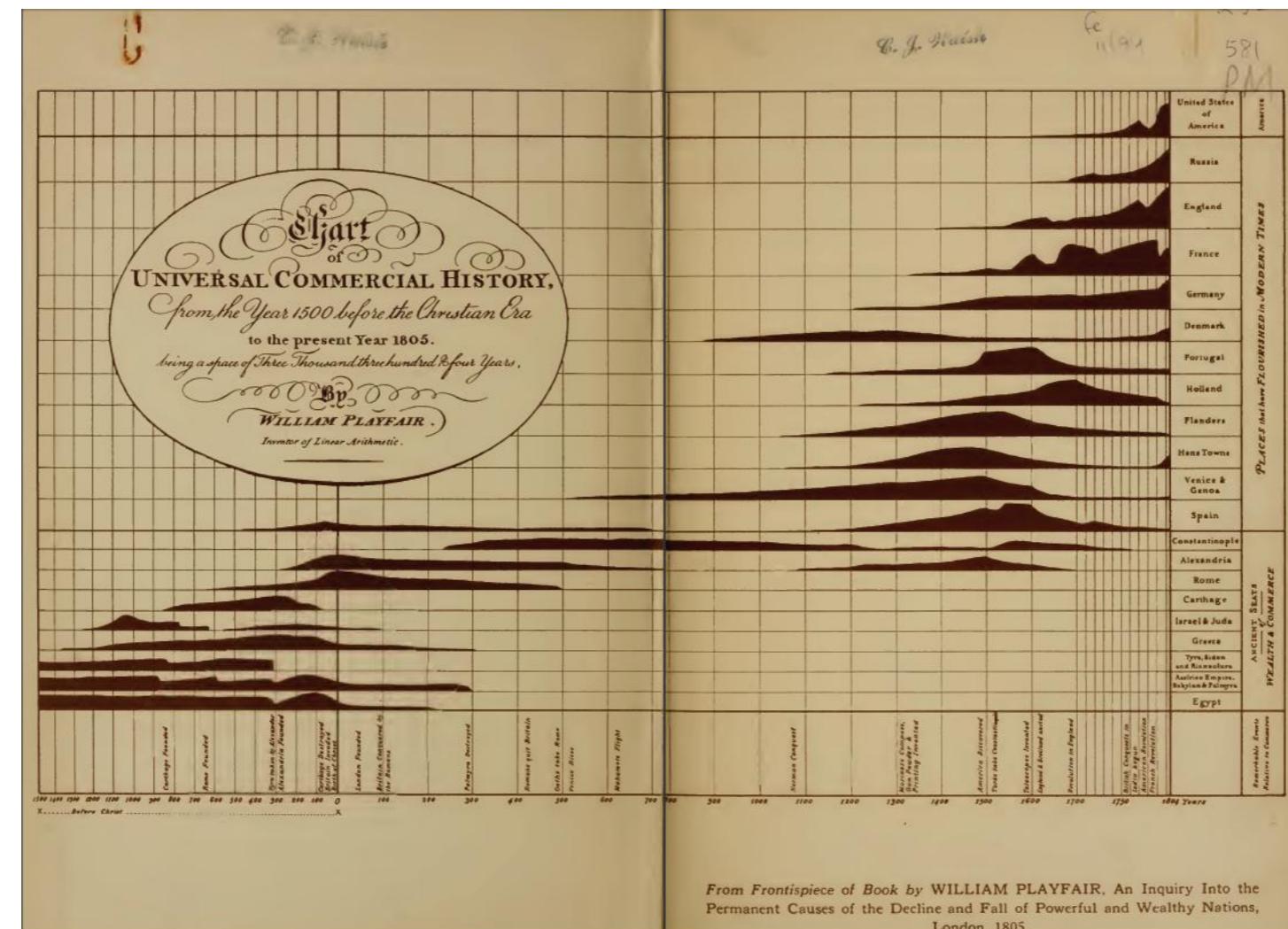
A Specimen of a Chart of Biography.

Joseph Priestley
(13.03.1733 – 6.02.1804)
britанский священник,
естественноиспытатель,
философ. Открыл кислород.

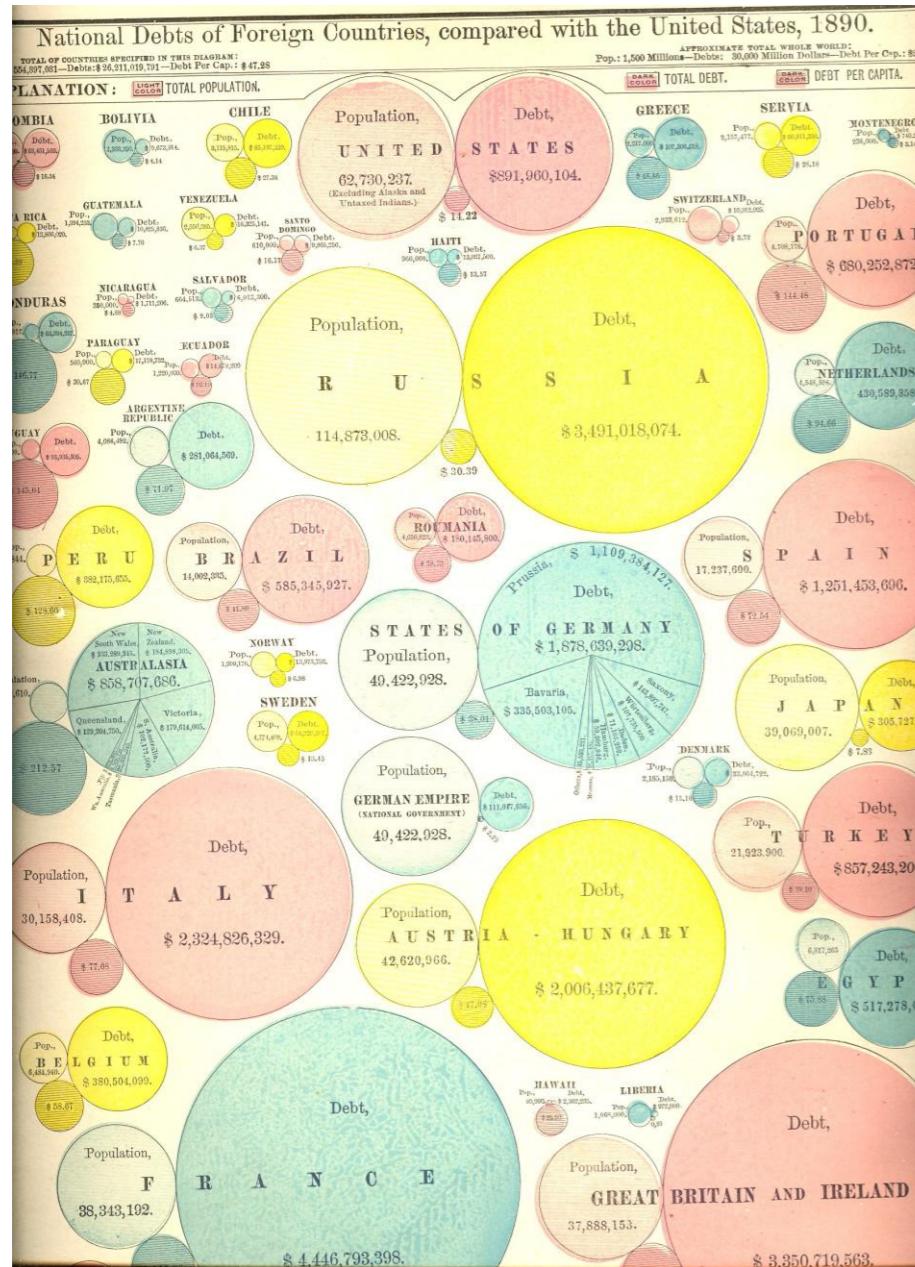
<https://pjodonnell.wordpress.com/2015/11/02/design-history-joseph-priestley/>

18 век Уильям Плейфэр

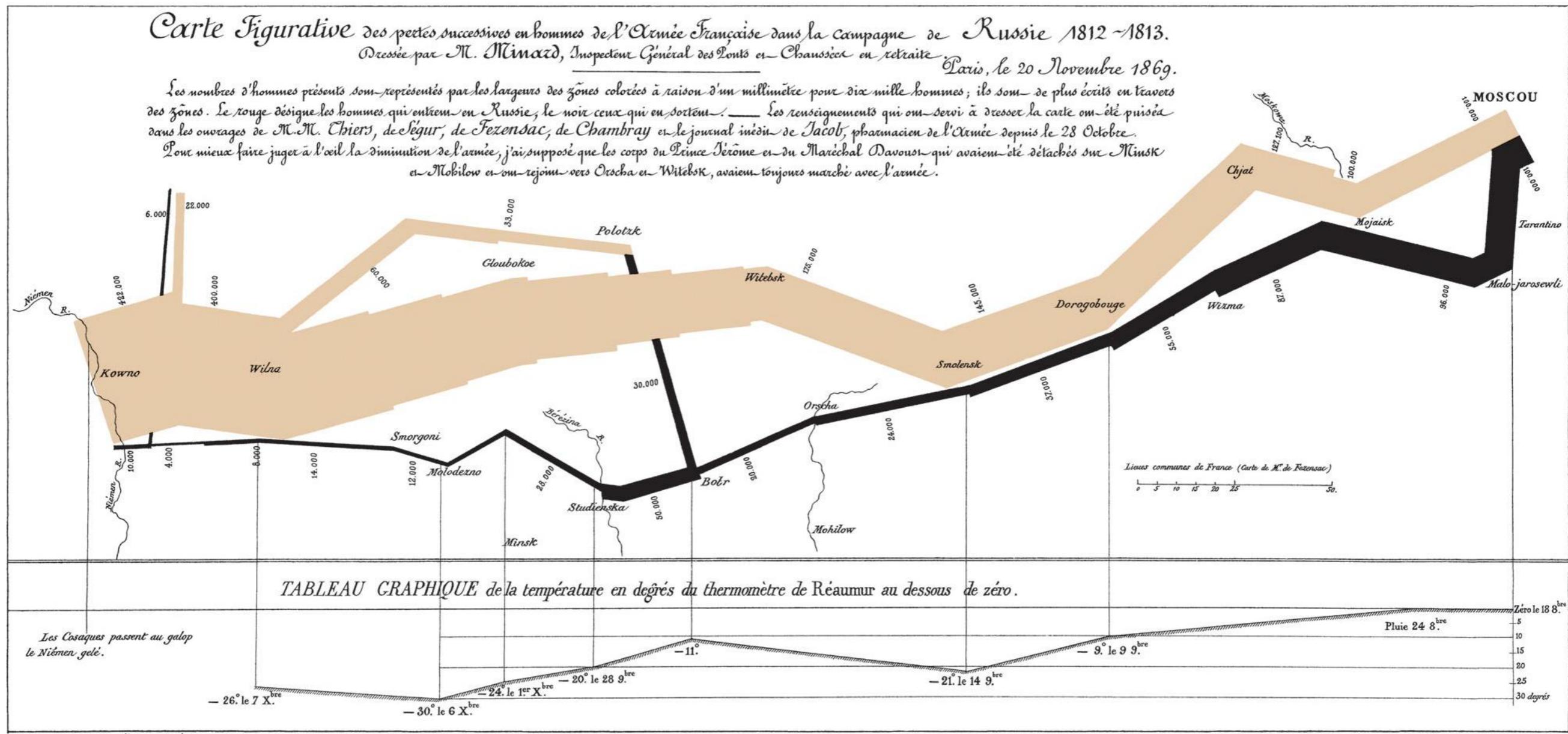
- 1786 – линейчатый график и гистограммы
- 1801 – секторная диаграмма в круге и круговая диаграмма



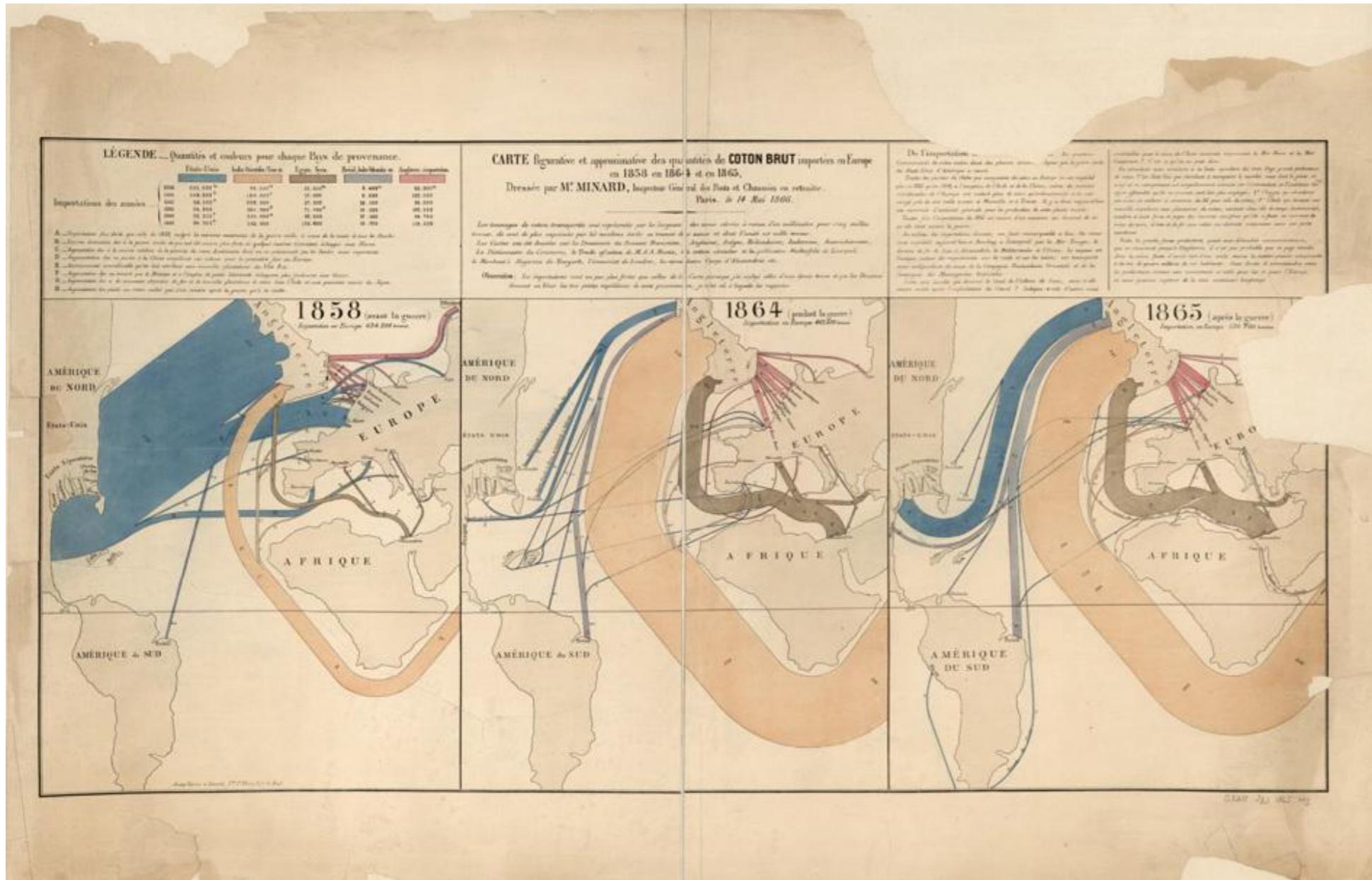
18 век Уильям Плейфэр



19 век Шарль Жозеф Минар

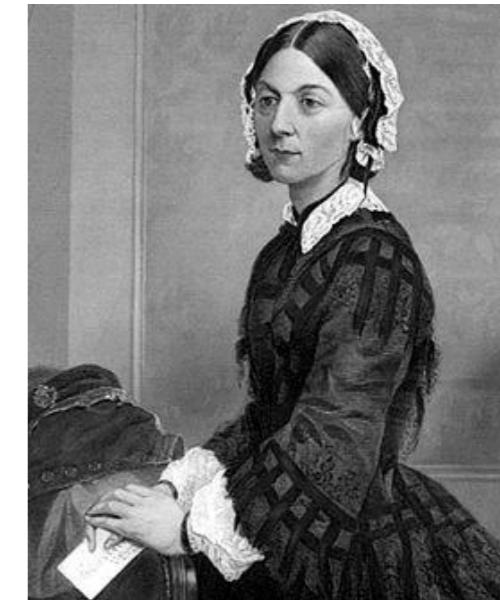
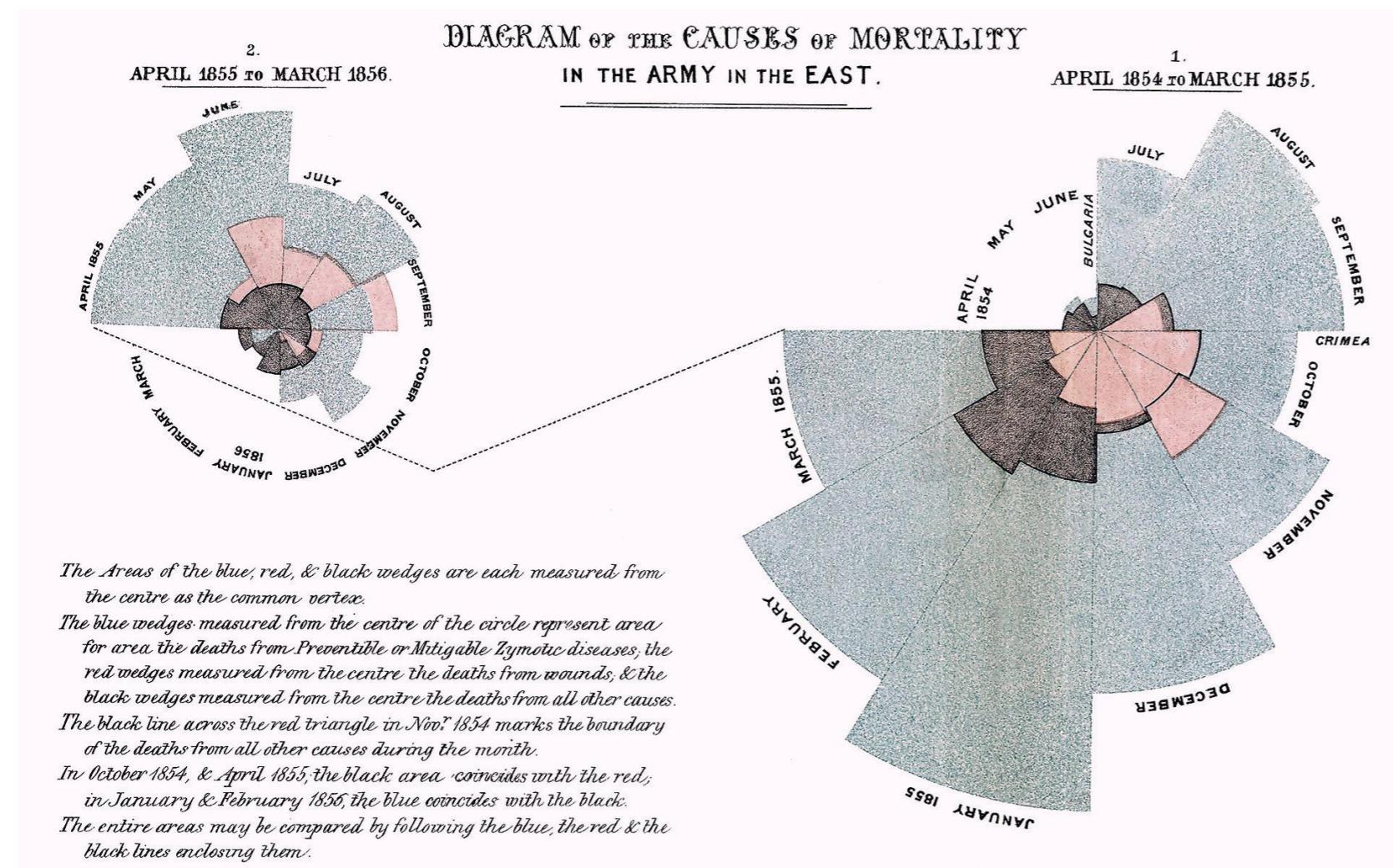


19 век Шарль Жозеф Минар



**Шарль Жозеф Минар
27.03.1781 – 24.10.1870**
**французский инженер,
топограф, пионер в области
графических методов
анализа и представления
информации в области
инженерных наук и
статистики**

19 век – Флоренс Найтингейл



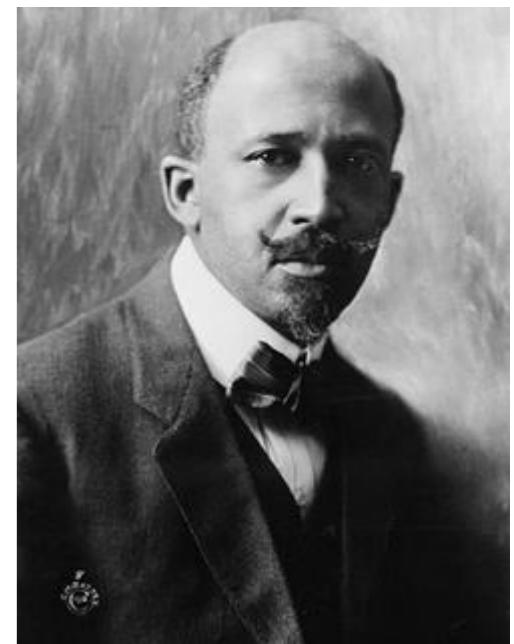
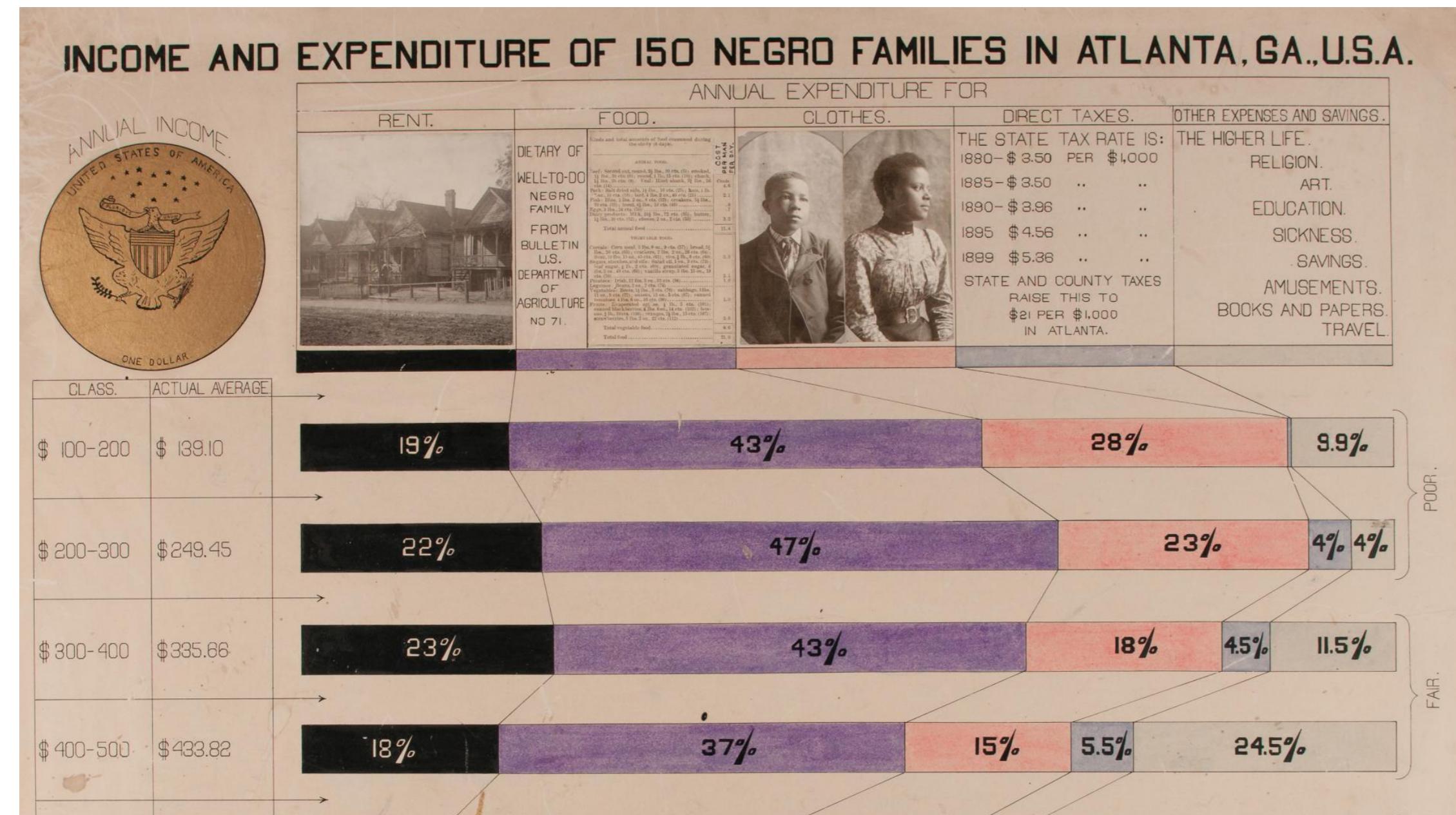
Florence Nightingale
12.05.1820 – 13.09.1910

сестра милосердия и общественная деятельница Великобритании.

«Петушиный гребень» – смертность солдат во время крымской войны.

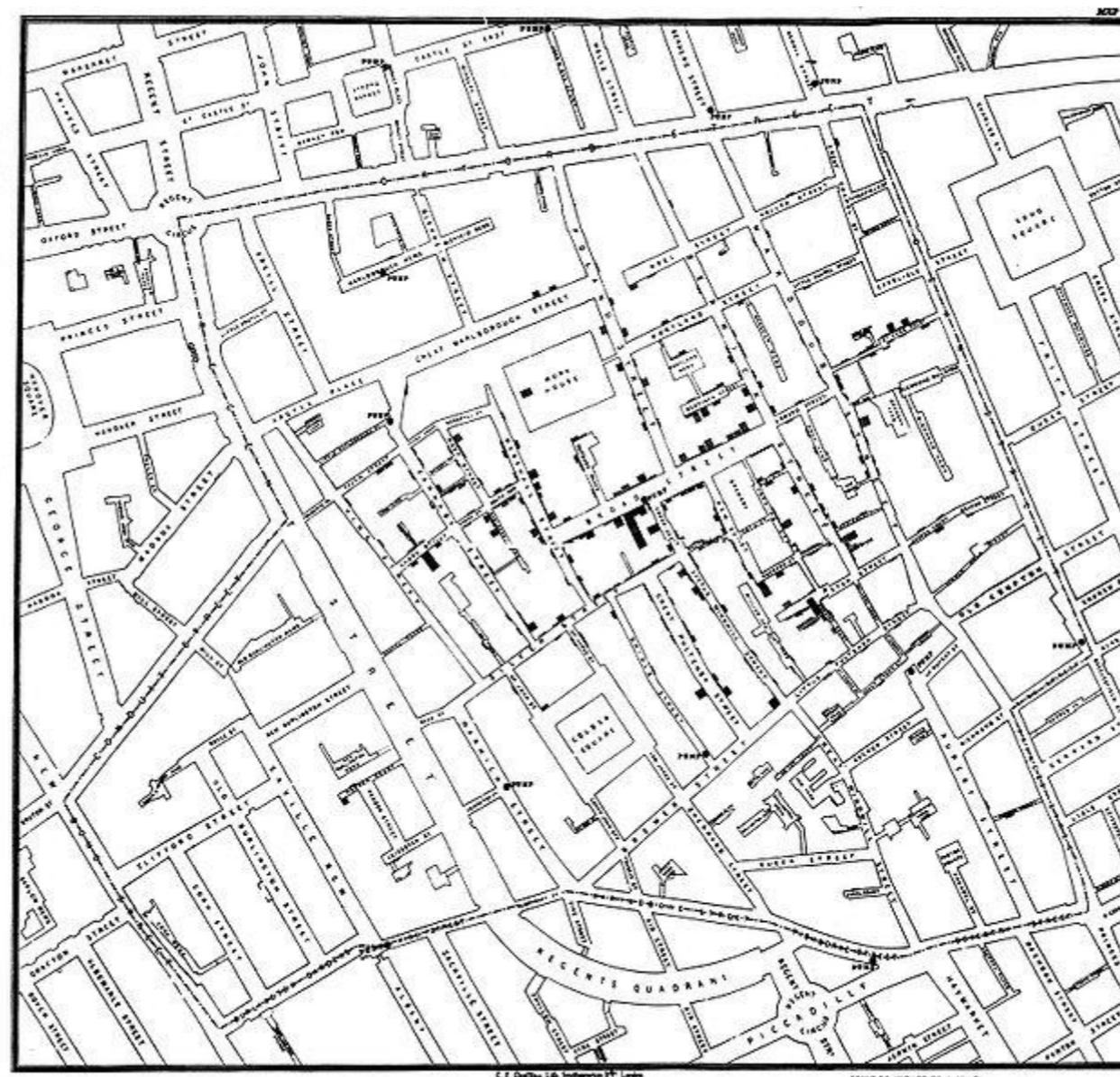
Площадь каждого сектора пропорциональна смертности. Голубой – смертность от болезней, красный – от ран, и коричневый слой – от других причин.

19-20 век, Уильям Дюбуа



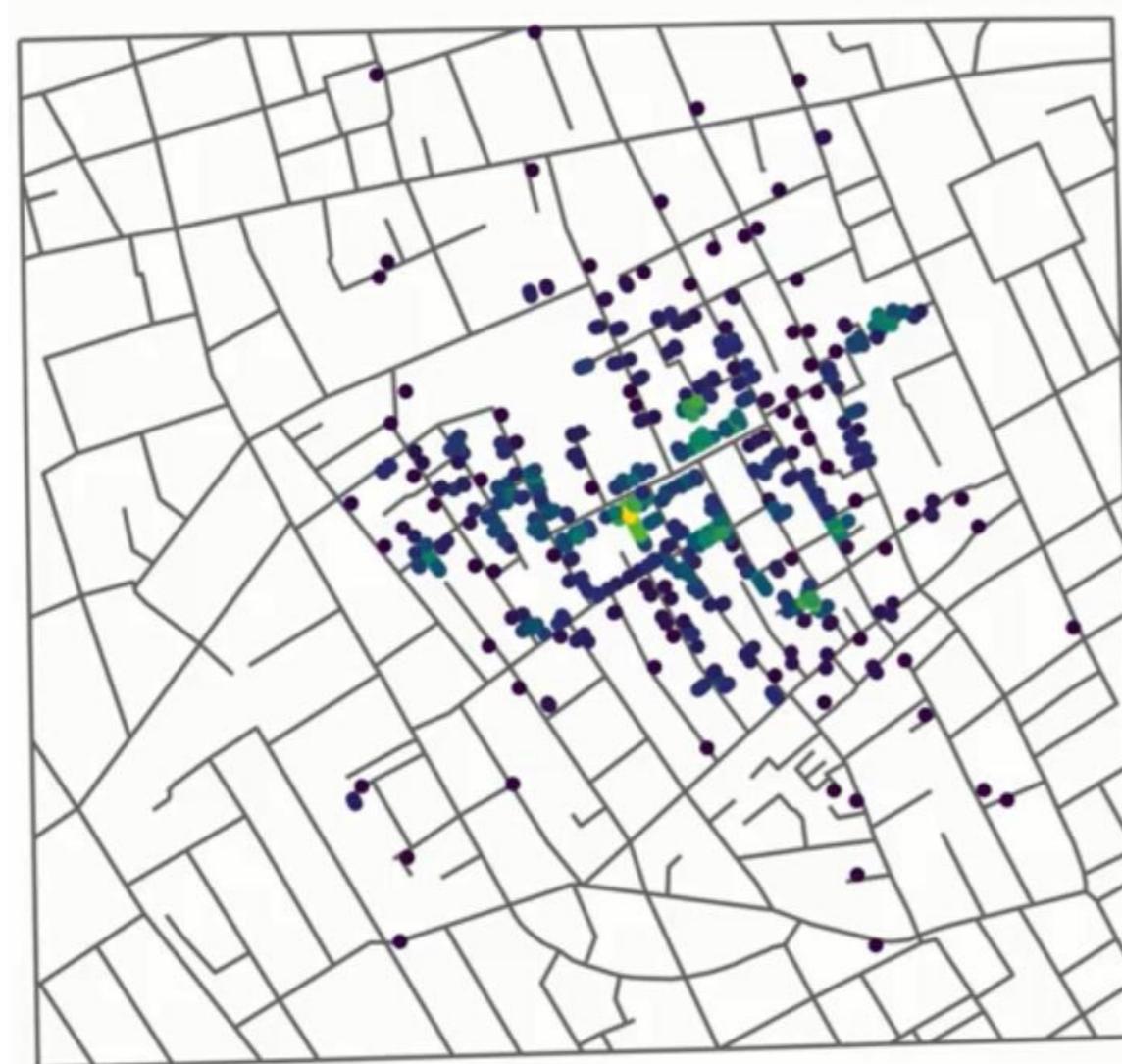
23.02.1868 - 27.08.1963 социолог, публицист (США-Гана)

Вспышка холеры на Брод-стрит в 1854 году



https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Статистика заболевания холерой



n_neighbors
4 8 12 16

Всего умерло 616 человек!
Причина?!
Кто такой Джон Сноу?

Центры эпидемии – колодцы!

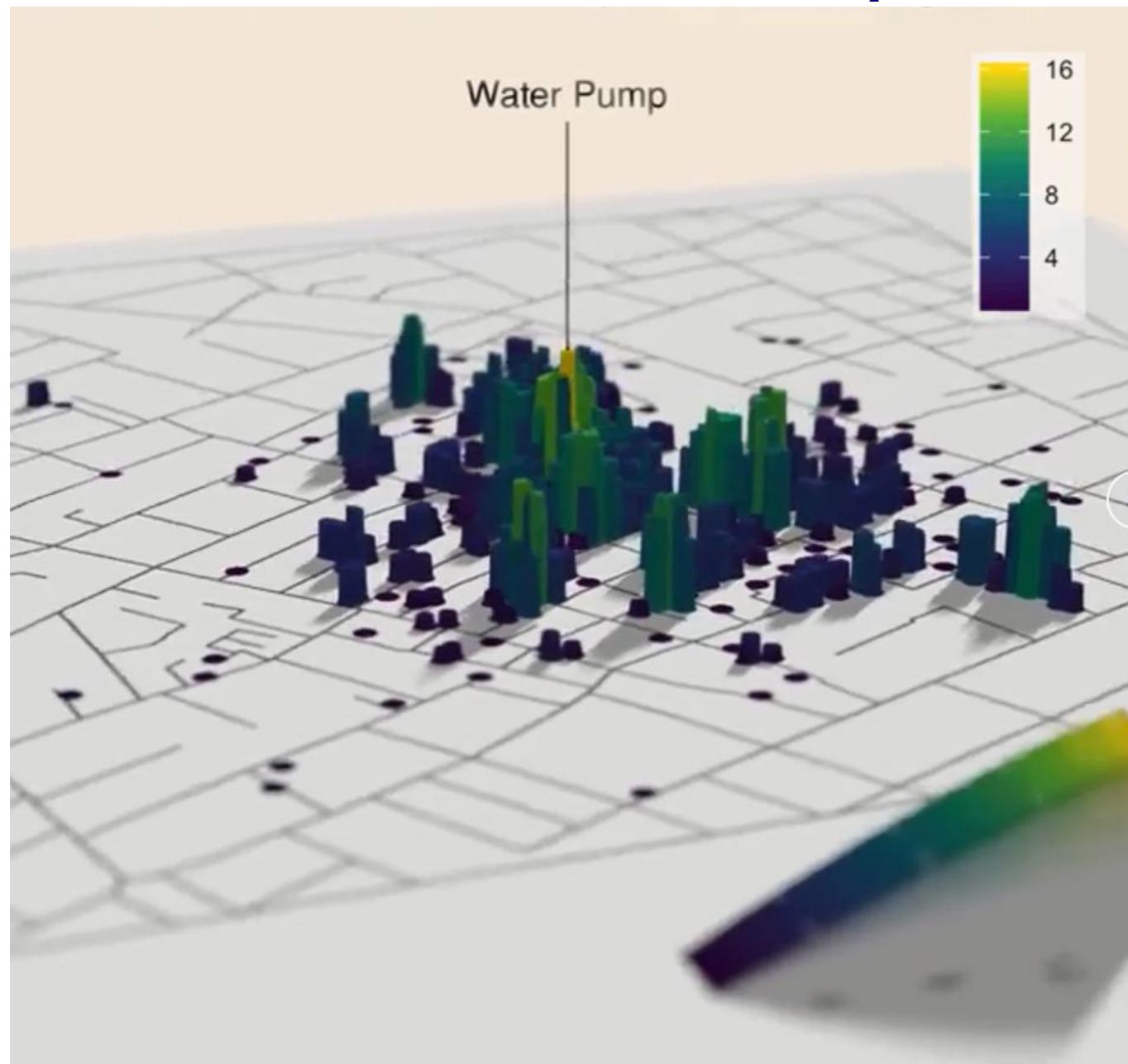
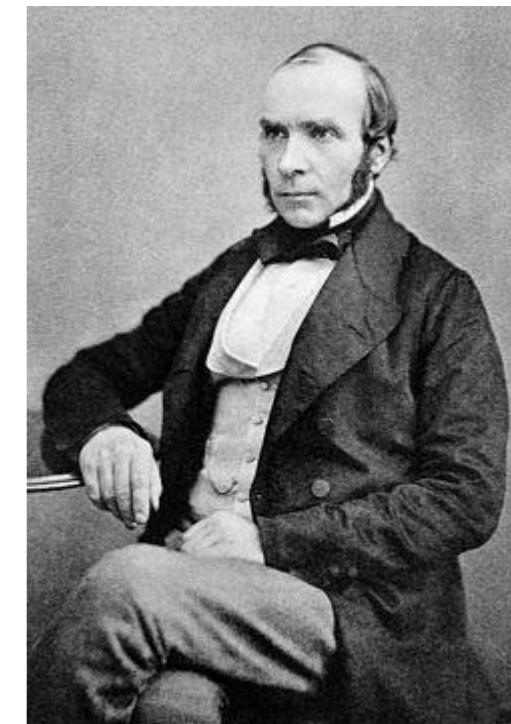


Диаграмма Вороного была видна на карте...

Нечистоты сливались в Темзу, в результате была заражена местная система водоснабжения.

https://www.reddit.com/r/dataisbeautiful/comments/d92sz0/clustering_in_john_snows_classic_1854_london/

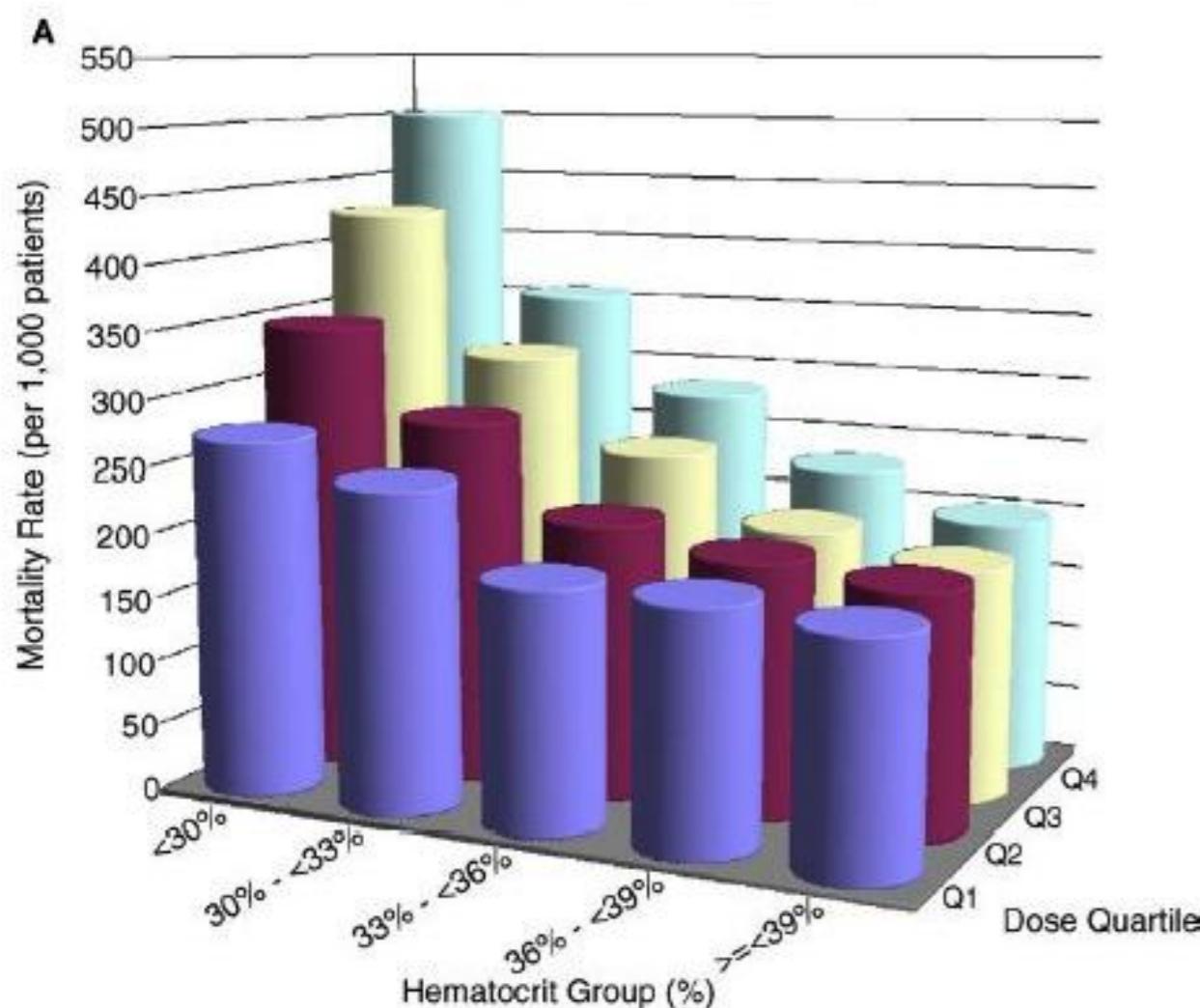
Джон Сноу



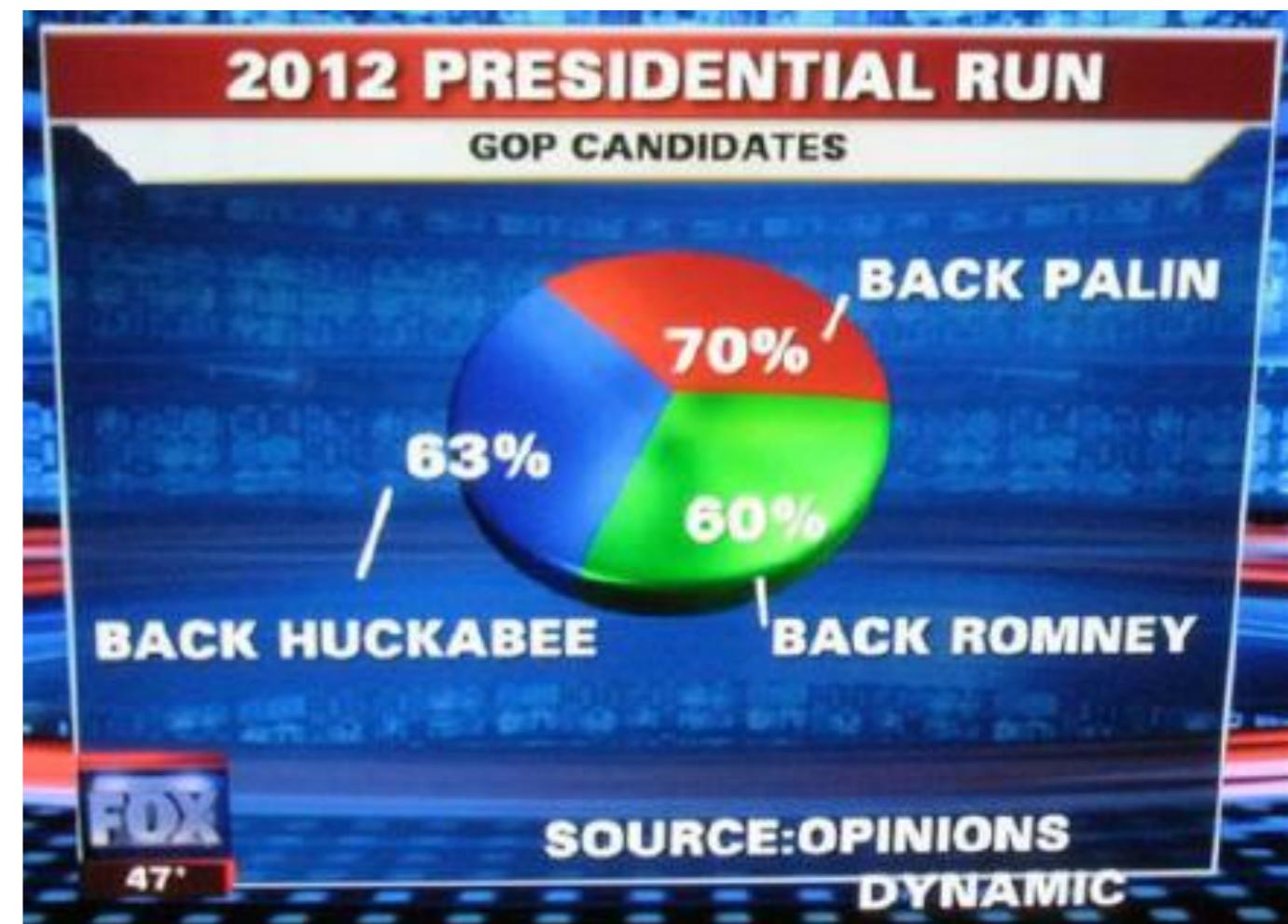
John Snow

(15.03.1813 — 16.06.1858)
britанский врач, один из пионеров массового
внедрения анестезии и медицинской гигиены

Основные правила – 3D – плохо, нелинейные сравнения – плохо!

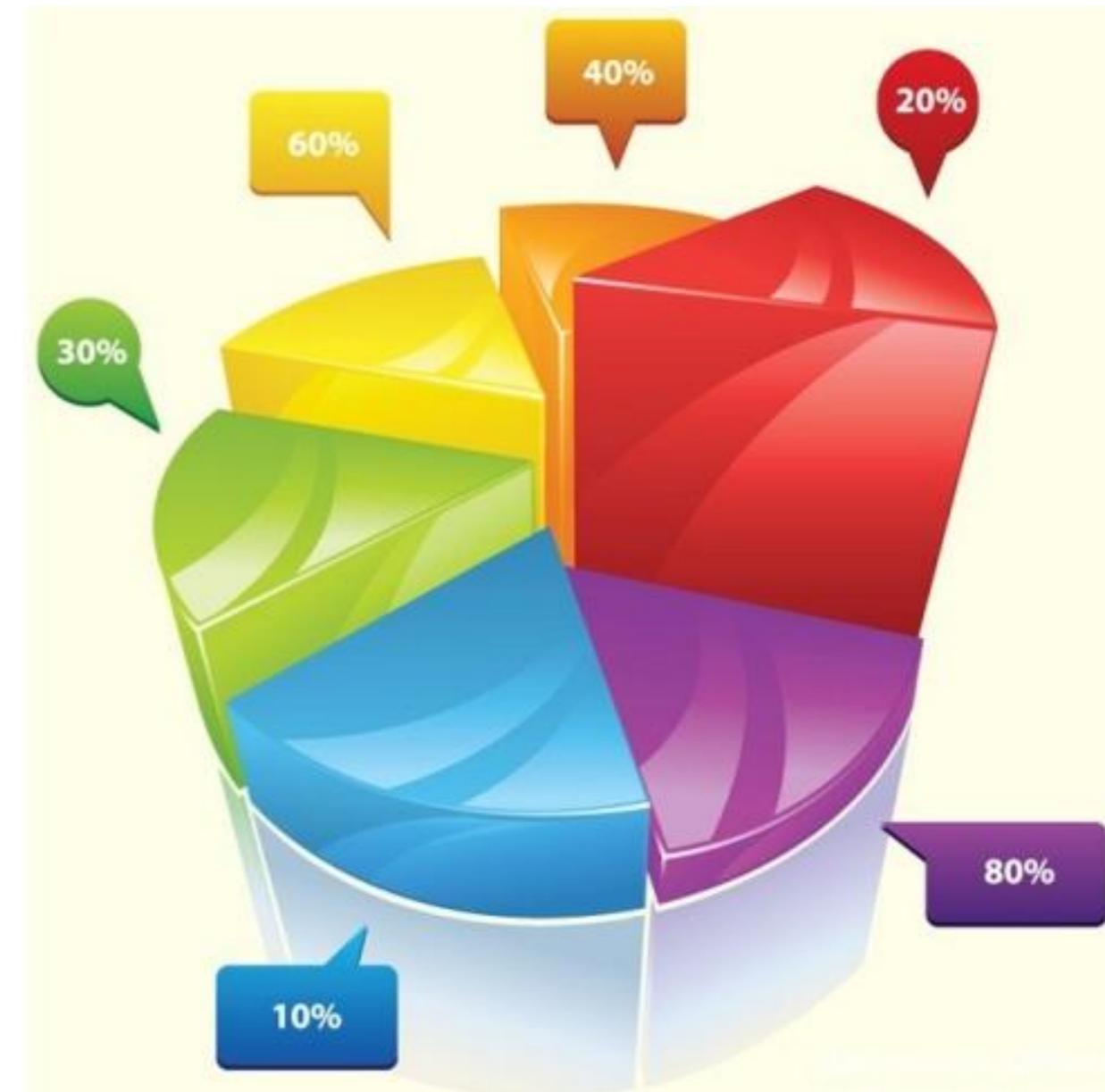
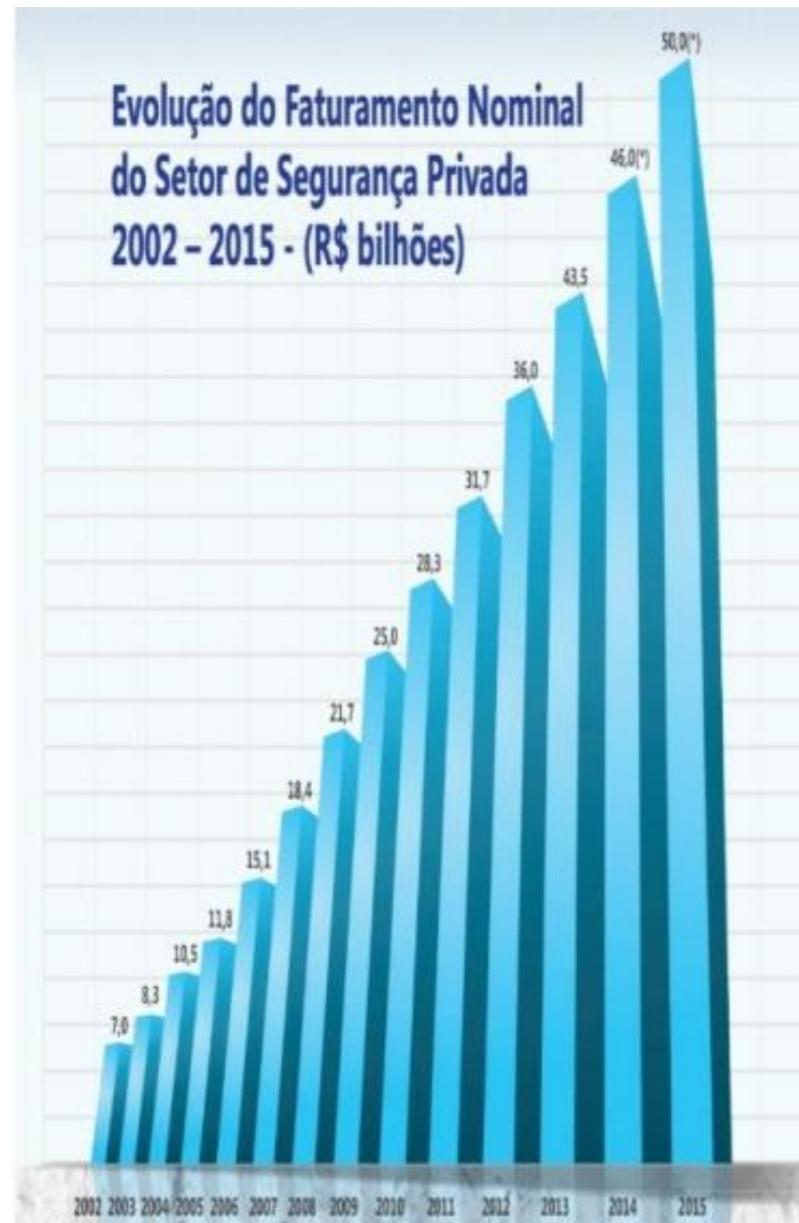


Cotter DJ, et al. (2004)



желательно избегать любой «нелинейности»,
например, диаграмм-пирогов (pie)

Основные правила – 3D – плохо, нелинейные сравнения – плохо!



<https://www.reddit.com/comments/9cql3f>

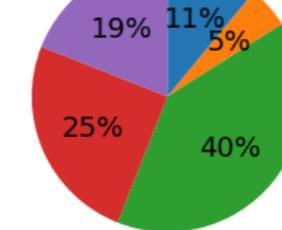
«Прикладные задачи анализа данных»

Пример: «до и после»

Survey Results

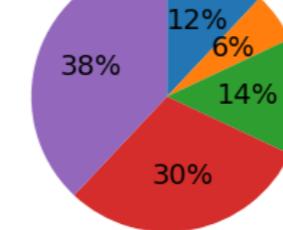
PRE: How do you feel about doing science?

Bored Not great OK Kind of interested Excited



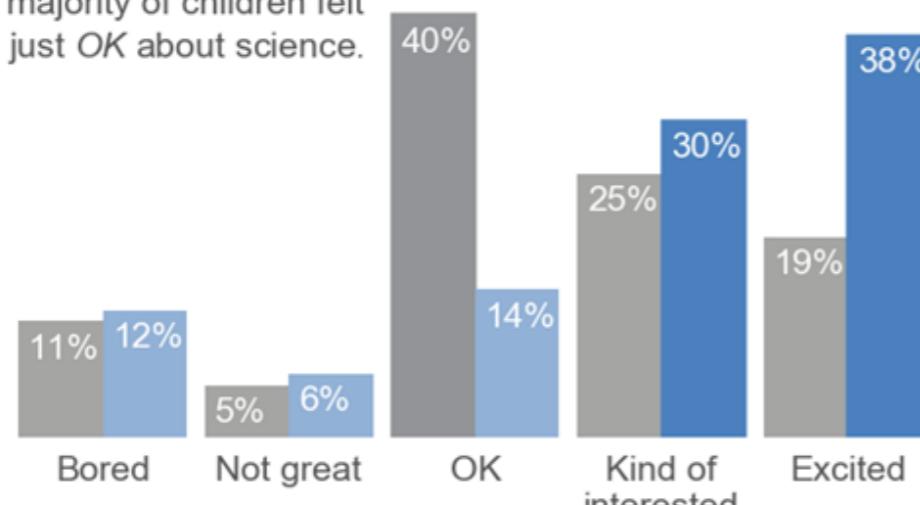
POST: How do you feel about doing science?

Bored Not great OK Kind of interested Excited

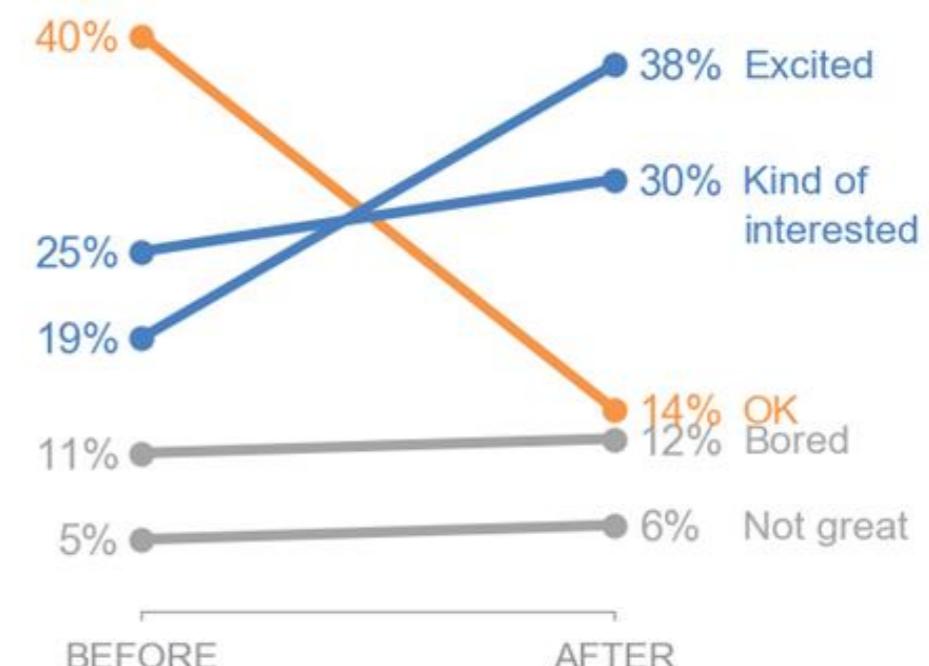


How do you feel about science?

BEFORE program, the majority of children felt just *OK* about science.

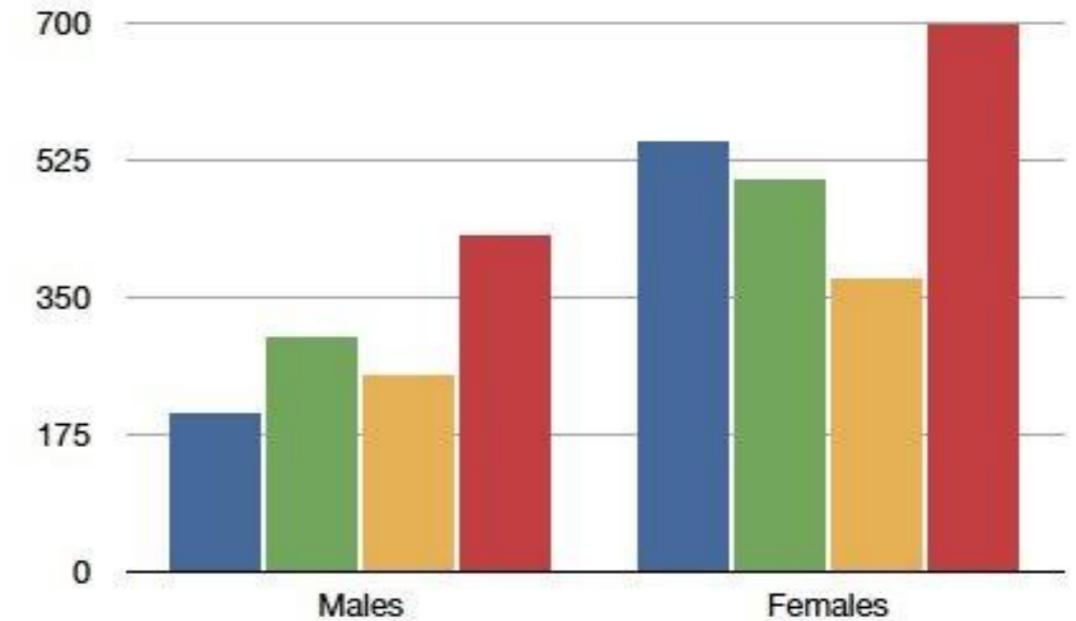
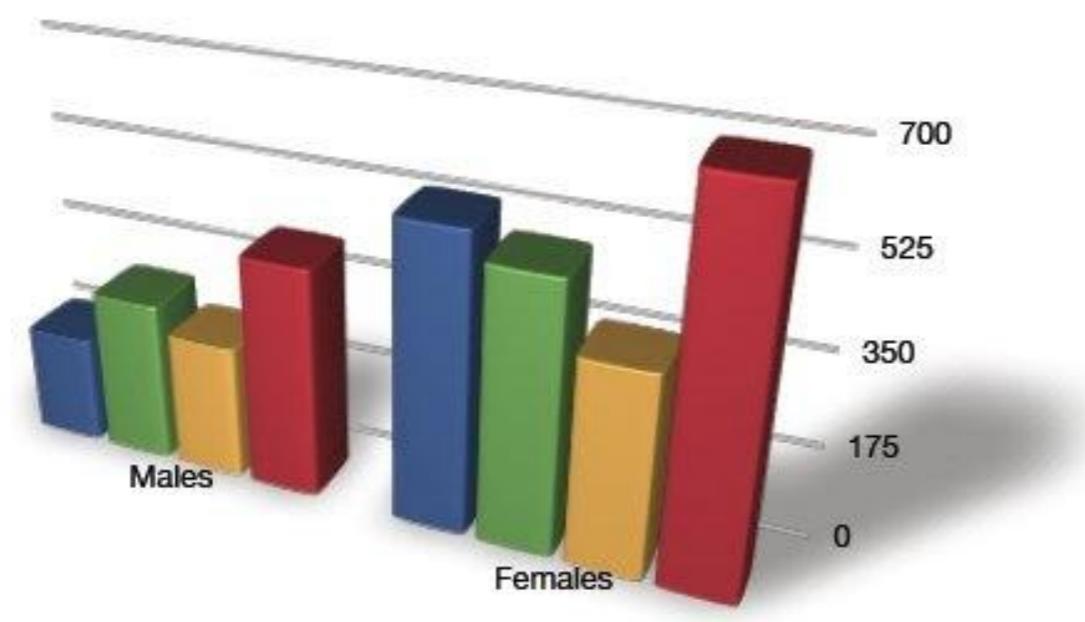


AFTER program, more children were *Kind of interested* & *Excited* about science.



<https://habr.com/company/eastbanctech/blog/422093/>

Основные правила: «минимализм» – не изображать лишнее



■ 0-\$24,999 ■ \$25,000+ ■ 0-\$24,999 ■ \$25,000+

■ 0-\$24,999 ■ \$25,000+ ■ 0-\$24,999 ■ \$25,000+

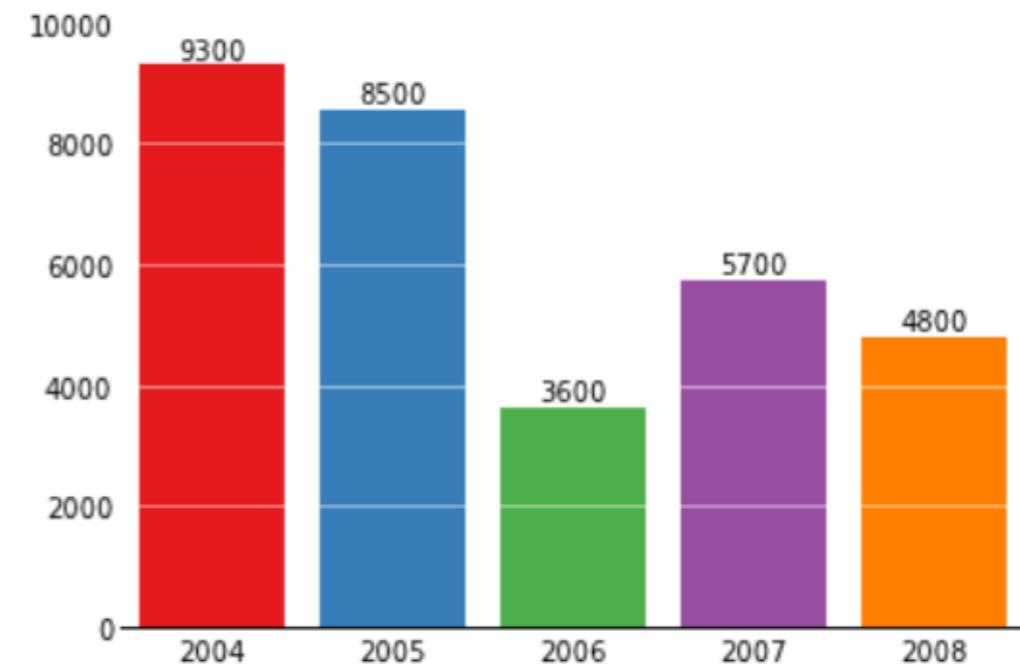
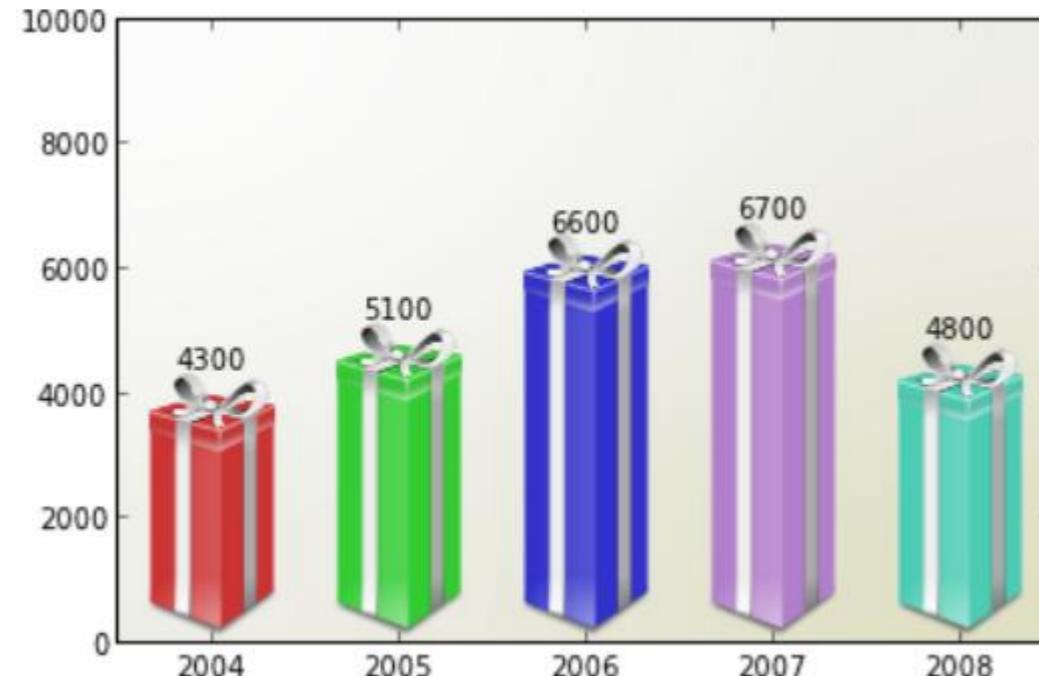
Data-Ink = объём данных / объём чернил → max
Данные важнее изображения!

<http://www.data-manual.com/>

Основные правила: «минимализм» – не изображать лишнее

нет лишних...
информации
цветов
размерностей
графических элементов

Анализируют данные, а не рисунки.
Иллюстрации лишь для облегчения
анализа!



<https://nbviewer.jupyter.org/gist/olgabot/5357268>

Основные правила: «минимализм» – не пишите лишнего

признак	важность	признак	важность
Сфера занятости	0.765176	Сфера занятости	76.5
Т с последнего визита	0.768735	Т с последнего визита	76.9
Запрашиваемая сумма	0.770486	Запрашиваемая сумма	77.0
Размер компании	0.770743	Размер компании	77.1
Сроки старых кредитов	0.772125	Сроки старых кредитов	77.2
Зарплатные поступления	0.772369	Зарплатные поступления	77.2
Доход	0.772609	Доход	77.3
Образование	0.773000	Образование	77.3
Число записей в БКИ	0.774000	Число записей в БКИ	77.4
Должность	0.775000	Должность	77.5
Пол	0.776000	Пол	77.6

Table 5
Simulation results for using full data, CRs only, and proposed method under four missing mechanisms

Method	Bias ^a		Variance ^b		95% CI ^c	
	($\hat{\beta}_W$)	($\hat{\beta}_X$)	($\hat{\beta}_W$)	($\hat{\beta}_X$)	($\hat{\beta}_W$)	($\hat{\beta}_X$)
(M.1) $P(R = 1) = 0.66$						
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.03062	-0.003561	0.1149	0.06732	0.960	0.955
Impu	0.01431	0.021	0.04088	0.05169	0.980	0.975
(M.2) logit $P(R = 1) = 2Y$						
Full	0.007908	-0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01945	0.07096	0.107	0.06581	0.960	0.950
Impu	0.006966	0.01597	0.04227	0.05226	0.975	0.985
(M.3) logit $P(R = 1) = 2X$						
Full	0.007908	-0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01225	0.0589	0.08856	0.06818	0.980	0.975
Impu	0.009563	-0.04699	0.03865	0.04923	0.985	0.970
(M.4) logit $P(R = 1) = X + Y$						
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.02404	1.613	0.1102	0.08202	0.955	0.580
Impu	0.01814	0.08289	0.0578	0.06075	0.955	0.970

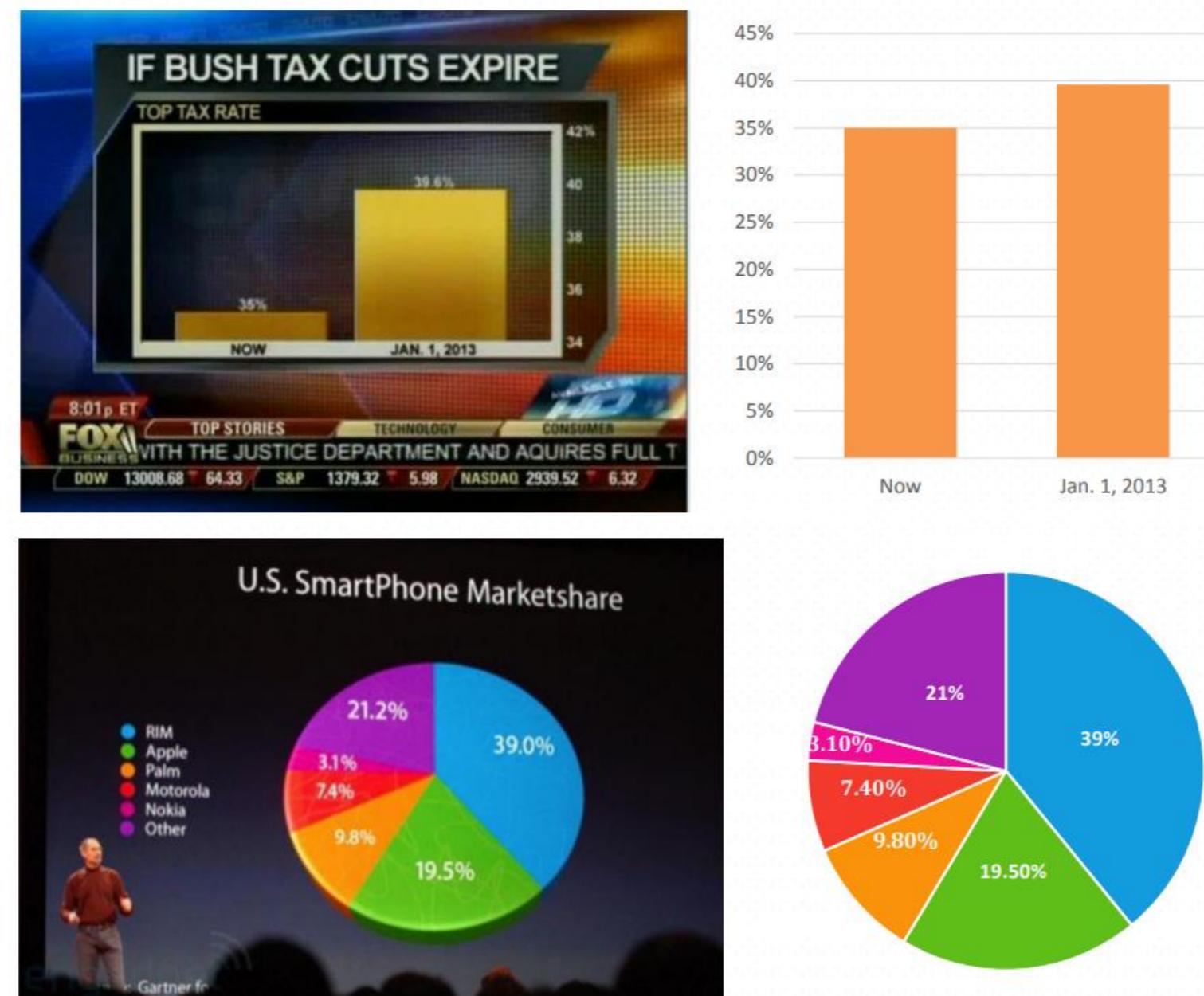
^aBias = ($\hat{\beta} - \beta_0$) / β_0 .

^bSimulation variance.

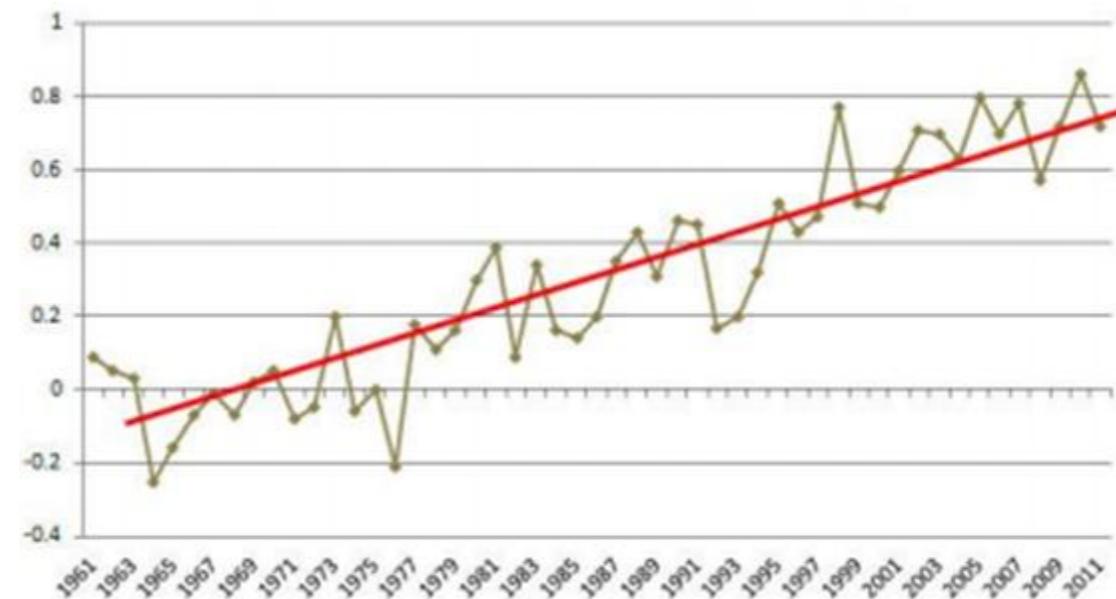
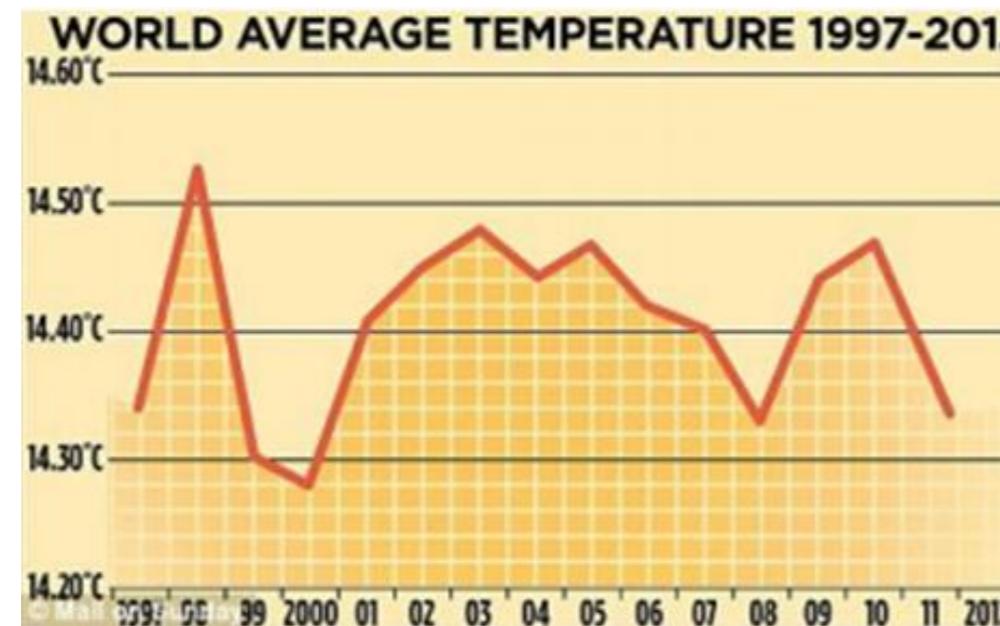
^cConfidence interval using jackknife standard error.

Paik MC (2004)

Основные правила – не обманывать



Основные правила – не обманывать



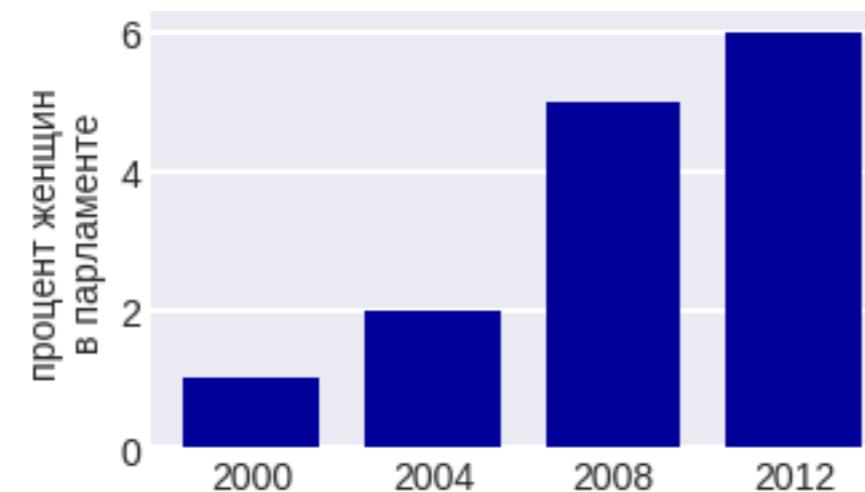
Один из примеров обмана – показать не все данные!

Про рекомендации к визуализации

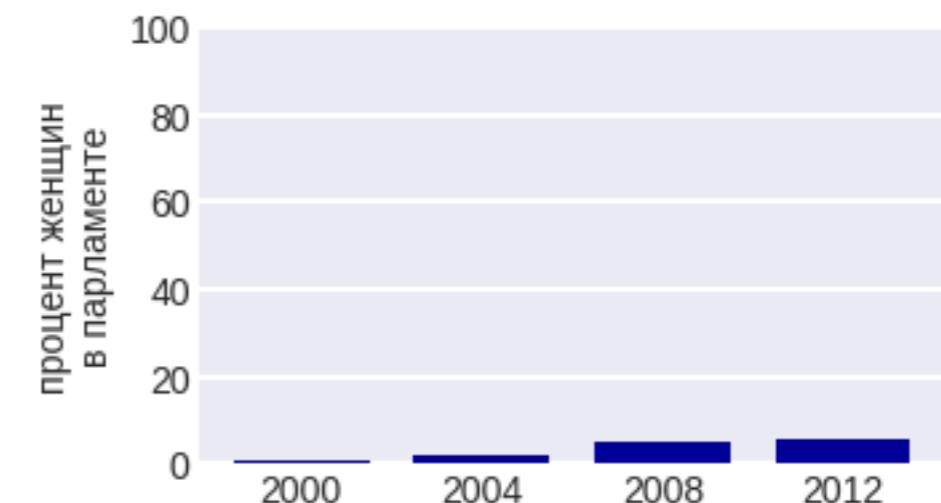
Всегда начинать график «в нуле» – спорное правило!

Процент женщин в парламенте

«неправильно»

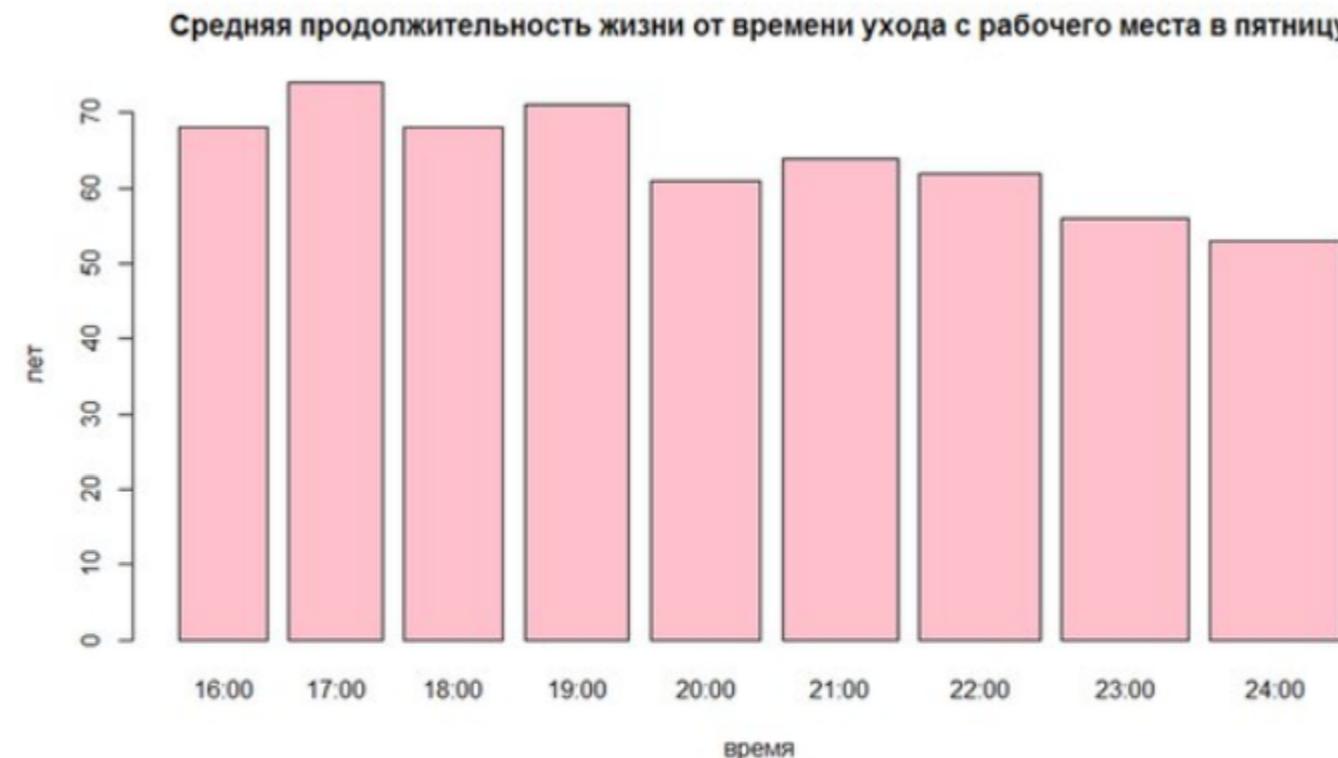


«правильно»



А если это процент убитых в Битцевском парке?

Про рекомендации к визуализации



24 июл в 12:25

Поделиться 🔍 Мне нравится ❤️ 8



масштаб отвратительный

24 июл в 12:43 | Ответить

Визуализация для профессионала

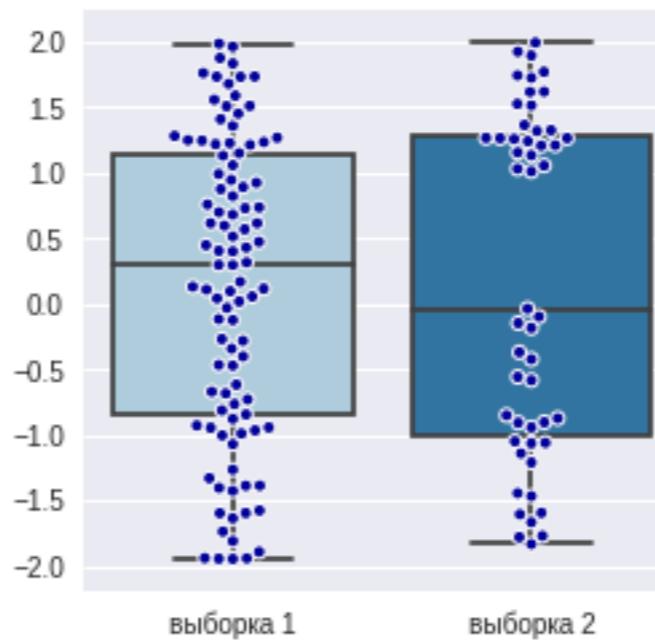
- где **объективный возможный минимум наблюдаемых значений**
- где **объективный возможный максимум наблюдаемых значений**
- какое **ожидаемое среднее у наблюдаемых значений**
- какие **отклонения наблюдаемых значений статистически значимы**

кратко: «ожидания - реальность»



надо отличать визуализацию для профессионалов и любителей!

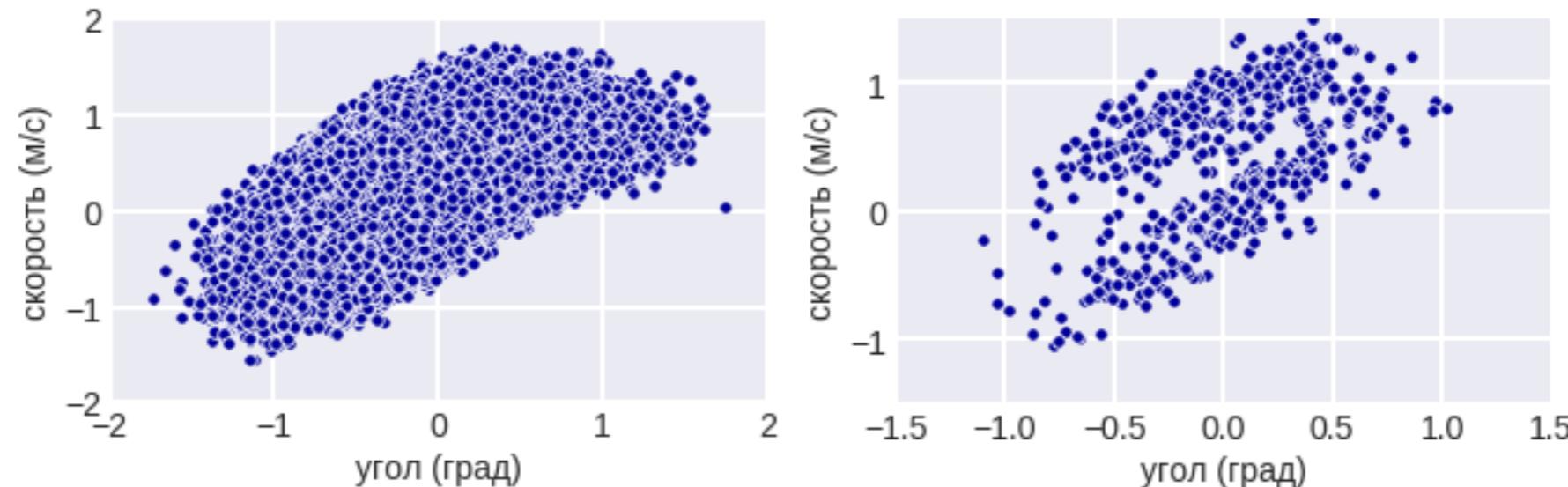
Правило: используйте разные средства



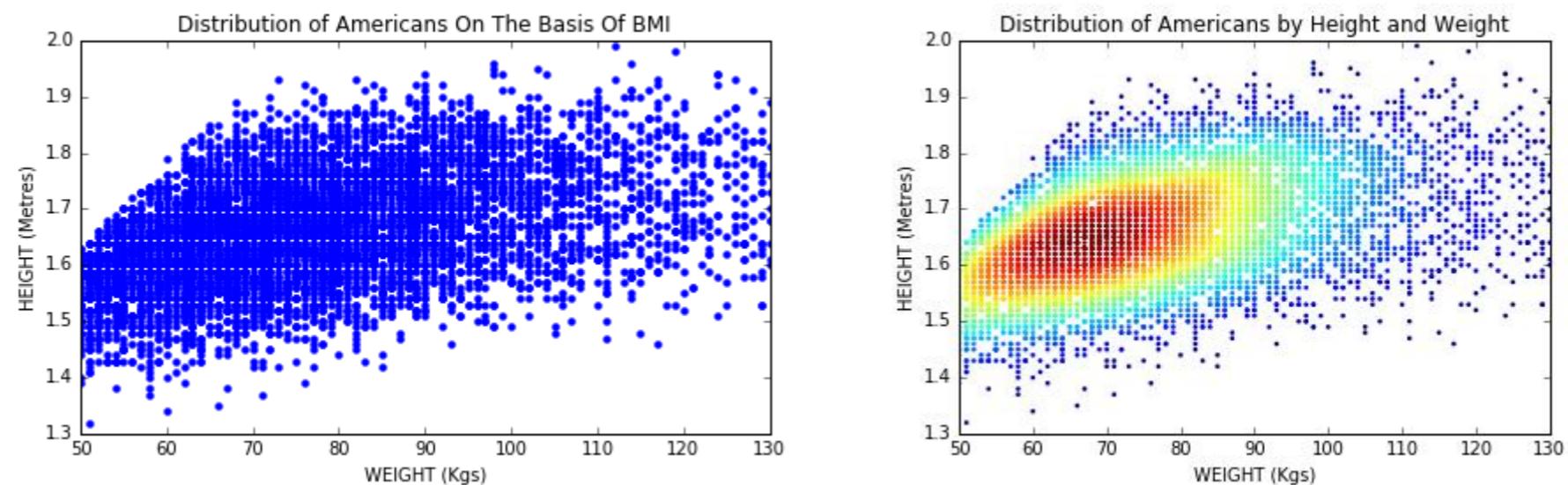
Однаковые диаграммы, а распределения разные

Выборку вообще очень полезно показывать!

Правило: не обязательно смотреть на всё



здесь – подвыборка 500 вместо ~200к: df [name] .sample (frac=0.1)



цвет – плотность <https://www3.cs.stonybrook.edu/~skiena/519/>

Правило: иллюстрация не только рисунок

Выбирайте шкалы

- единицы измерения
- логарифмический / обычный масштаб
- видимая сетка
- общие оси для нескольких графиков
- общая зона для нескольких графиков

Выбирайте текст (должен быть!)

- заголовок
- подписи
- текстовые вставки
- легенду (где и зачем!)

Выбирайте цвет и стиль

- цвет позволит выделить некоторые элементы!
- цвета, которые и в градациях серого будут разными!

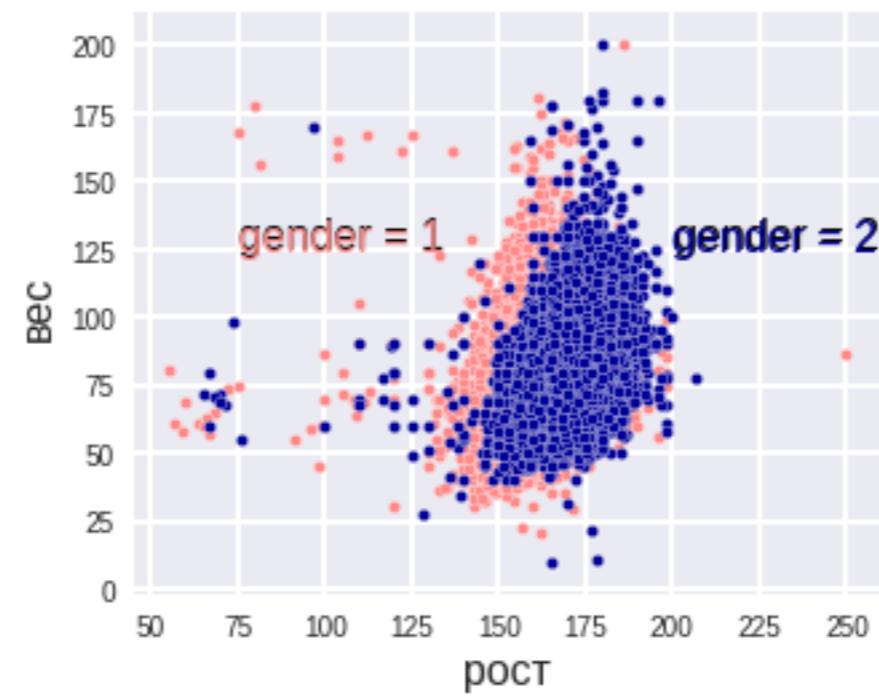
Правило: иллюстрация не только рисунок

достаточно просто изобразить таблицу!
можно подсветить особые элементы
(стиль, шрифт, цвет)

**порядок строк / столбцов
их компактные названия**

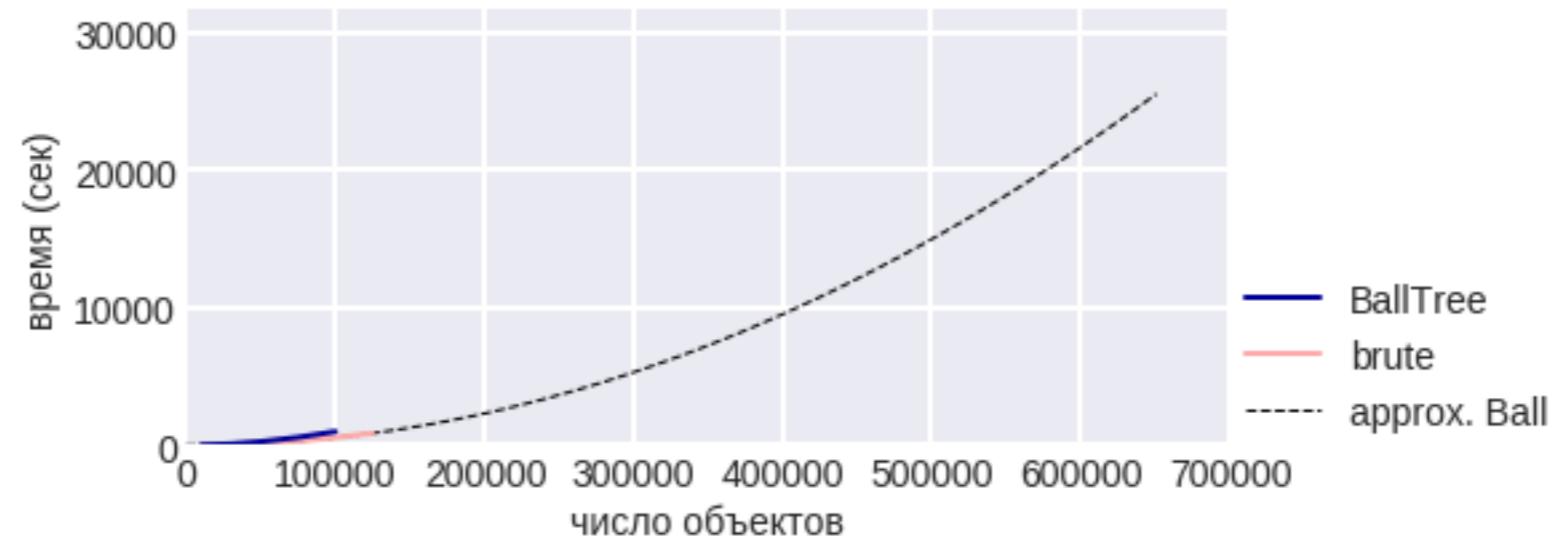
про точность говорили...

Пример, когда можно без стандартной легенды

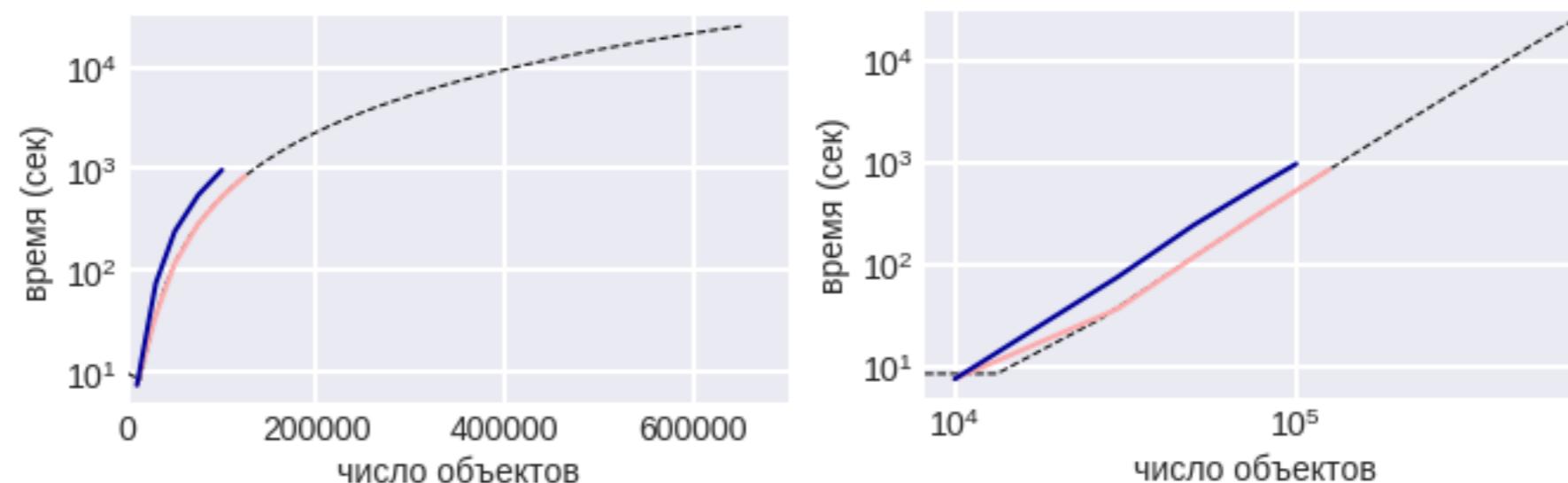


в чём недостаток этой диаграммы рассеивания?

Пример логарифмической шкалы

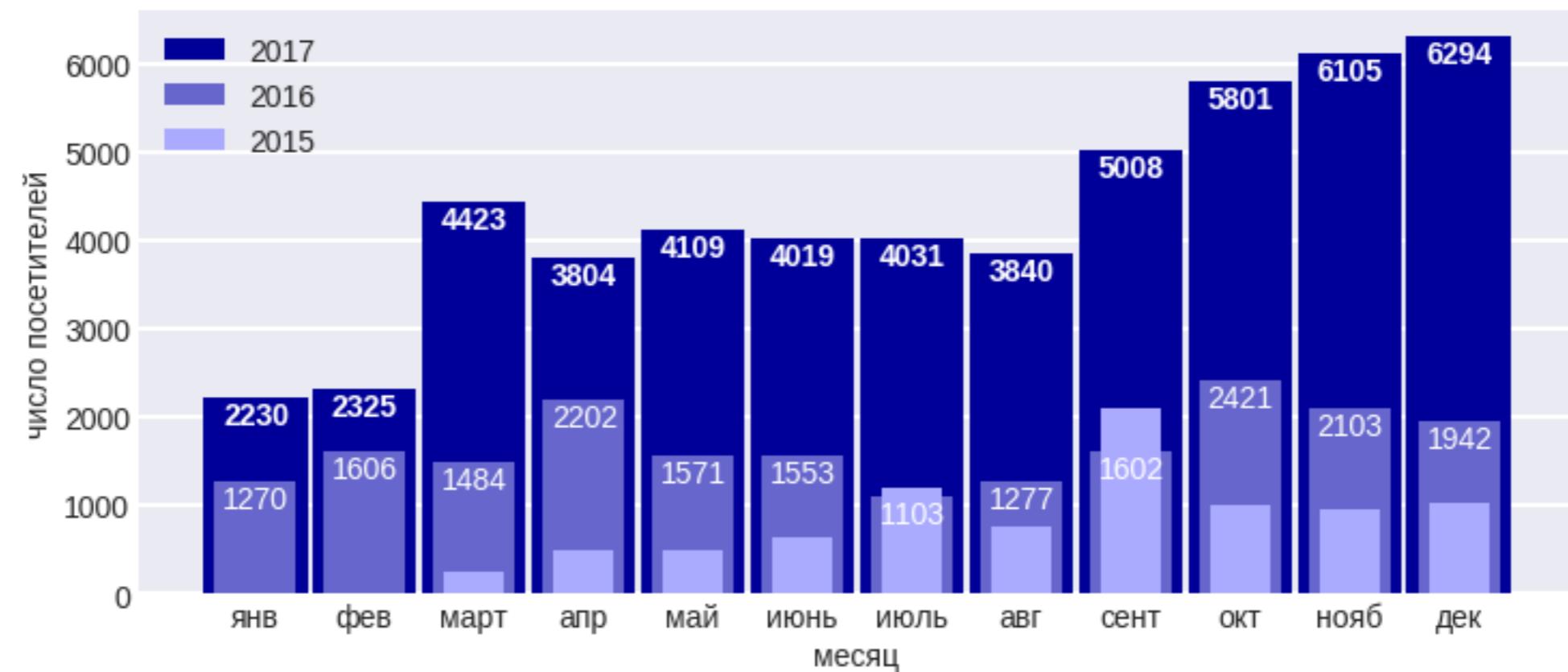


в стандартных шкалах



логарифмируем (слева – Y, справа – X и Y)

Пример иллюстрации



Когда выполнены все советы...

Домашнее задание

ДЗ (необязательное) Найти интересные приёмы для визуализации

**ДЗ (необязательное) Найти другие примеры описанных правил или м.б. примеры других правил, которые Вы считаете ценноыми
(должны быть ссылки на источники)**

**ДЗ (обязательное) Найти интересные нетривиальные визуализации
для игры «Что за данные?»**

**ДЗ (обязательное – большое) Сделать визуализацию данных Kaggle
можно брать свежие (1 год) датасеты
более старые надо согласовать (должно быть малое число кёрнелов)**

Важный совет

Храните данные и код для получения картинки,

а не только само изображение

(в 99 случаях из 100 его придётся немножко переделать)

Итог

Картинка может говорить о распределении больше, чем статистики

Люди научились визуализации сравнительно недавно, и до сих пор учатся...

Вредные советы: 3D, нелинейность, нечестность, лишнее...

Минимализм – максимизация «Data-Ink»

Простыми средствами больше информации

Иллюстрация не только картинка – цвета, стиль, текст, расположение, масштаб и т.п.

Разнообразные средства – не все способы позволяют видеть нужный эффект

Профессионал умеет читать графики!

Профессионал выбирает график: шкалы, текст, стиль

Литература

The Art of Effective Visualization of Multi-dimensional Data

<https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57>

Karl W Broman «More on data visualization»

https://www.biostat.wisc.edu/~kbroman/presentations/more_on_graphs.pdf

Fernanda Viégas, Martin Wattenberg «Visualization for Machine Learning»

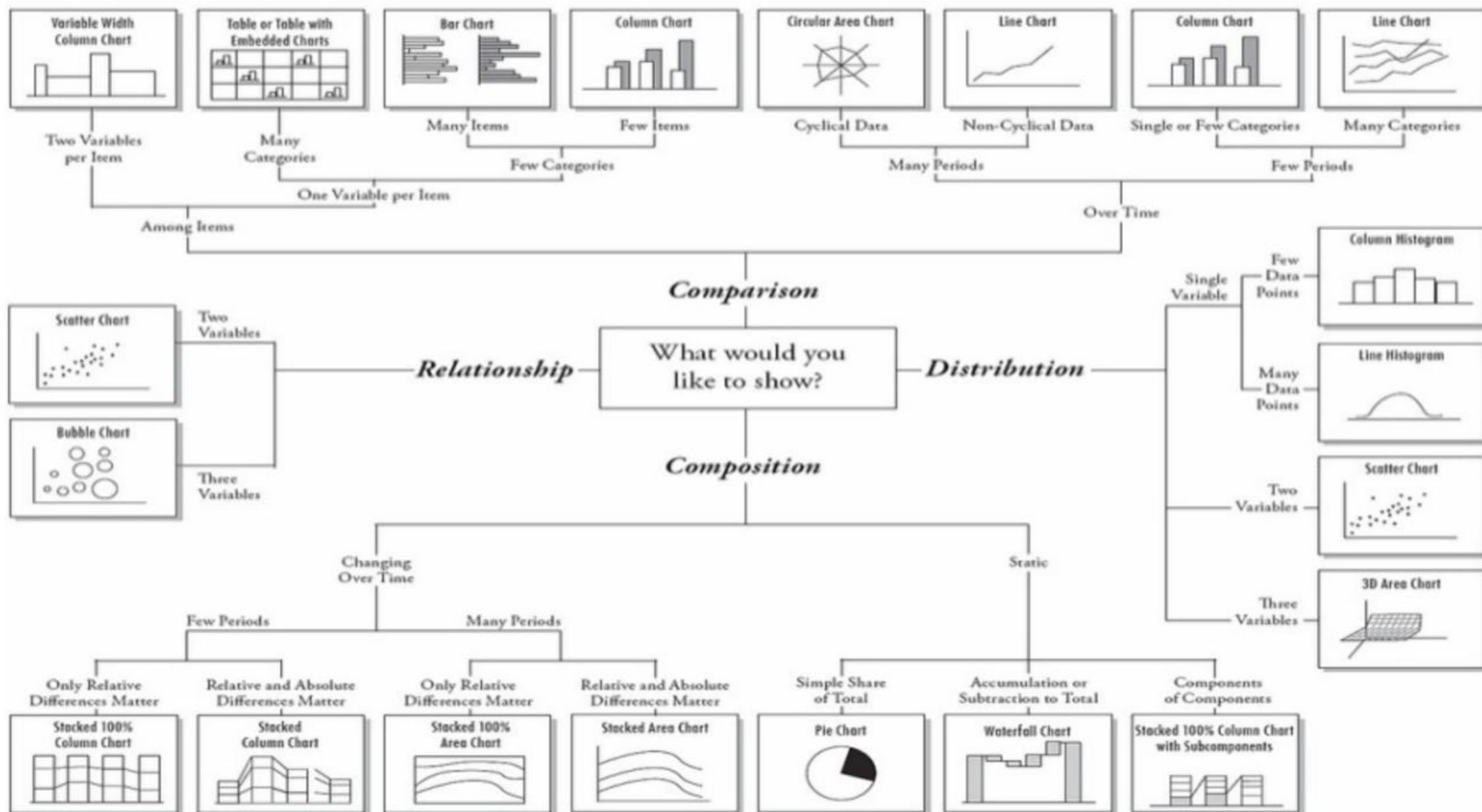
https://media.neurips.cc/Conferences/NIPS2018/Slides/Visualization_for_ML.pdf

Visualizations that make no sense

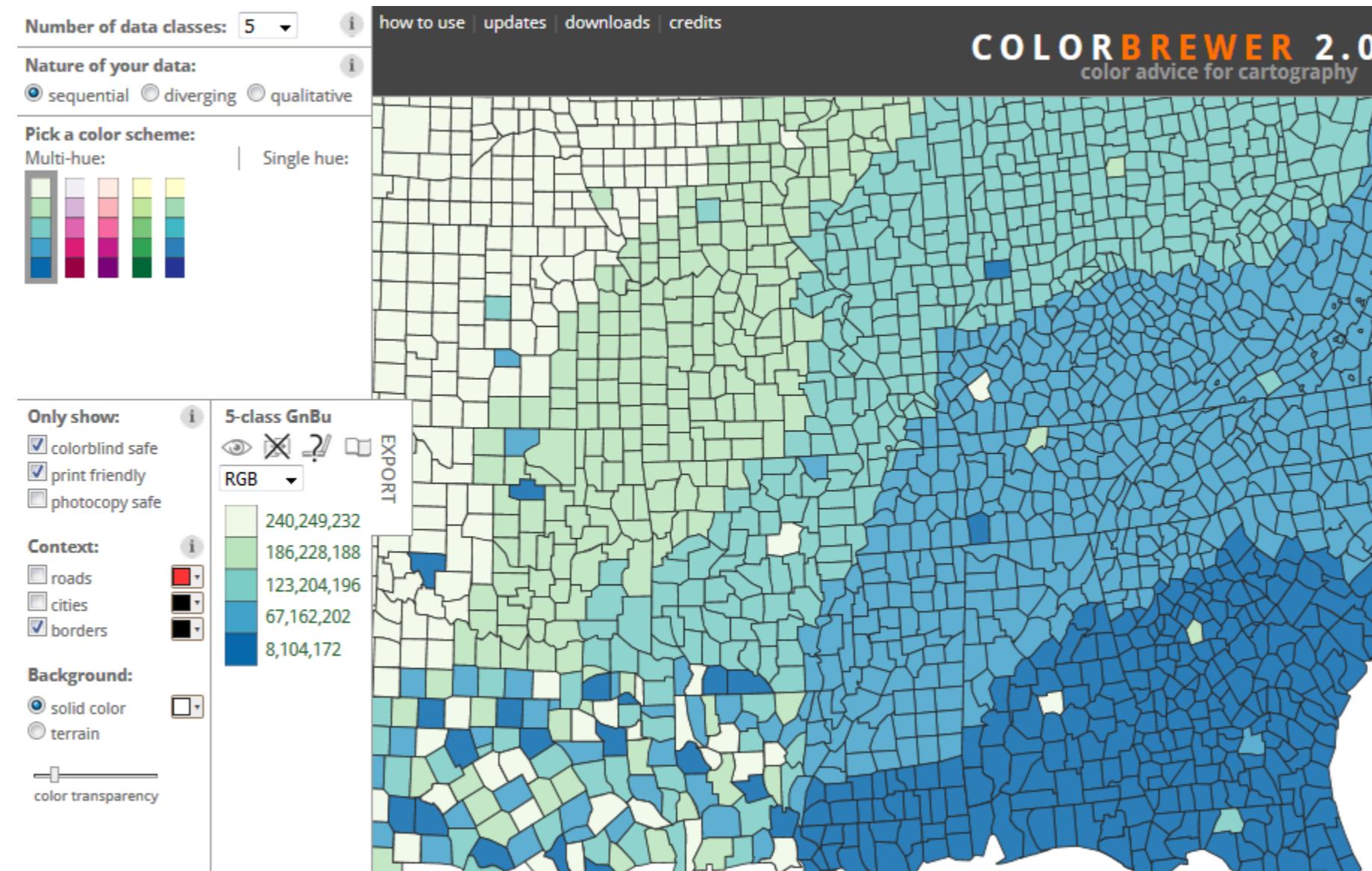
<https://viz.wtf/>

П.С.

Chart Suggestions—A Thought-Starter



Выбор цветов для визуализации



<http://colorbrewer2.org>