



Прикладные задачи анализа данных

искусство визуализации

Часть 1. Историческая

Дьяконов А.Г.

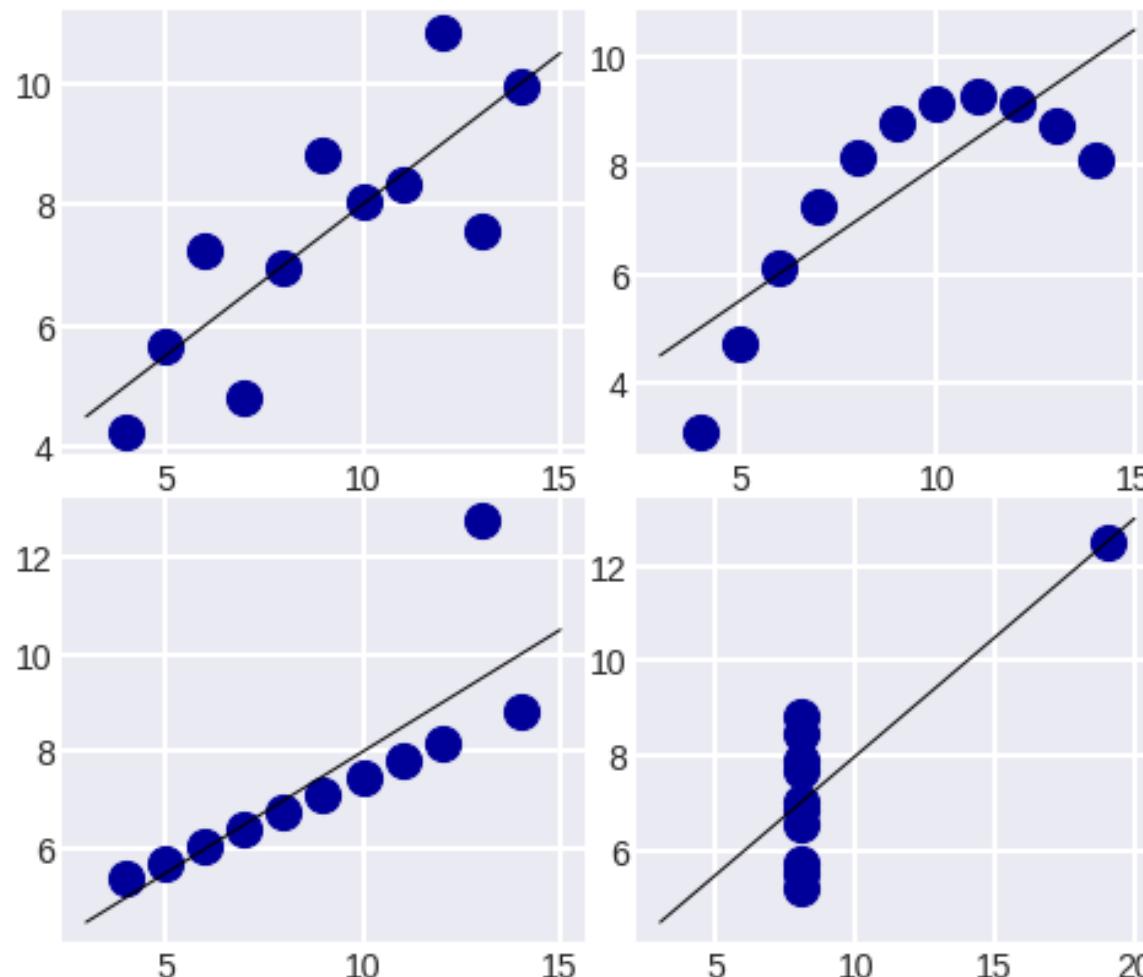
**Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)**

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

John Tukey

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	$Y=3+X/2$

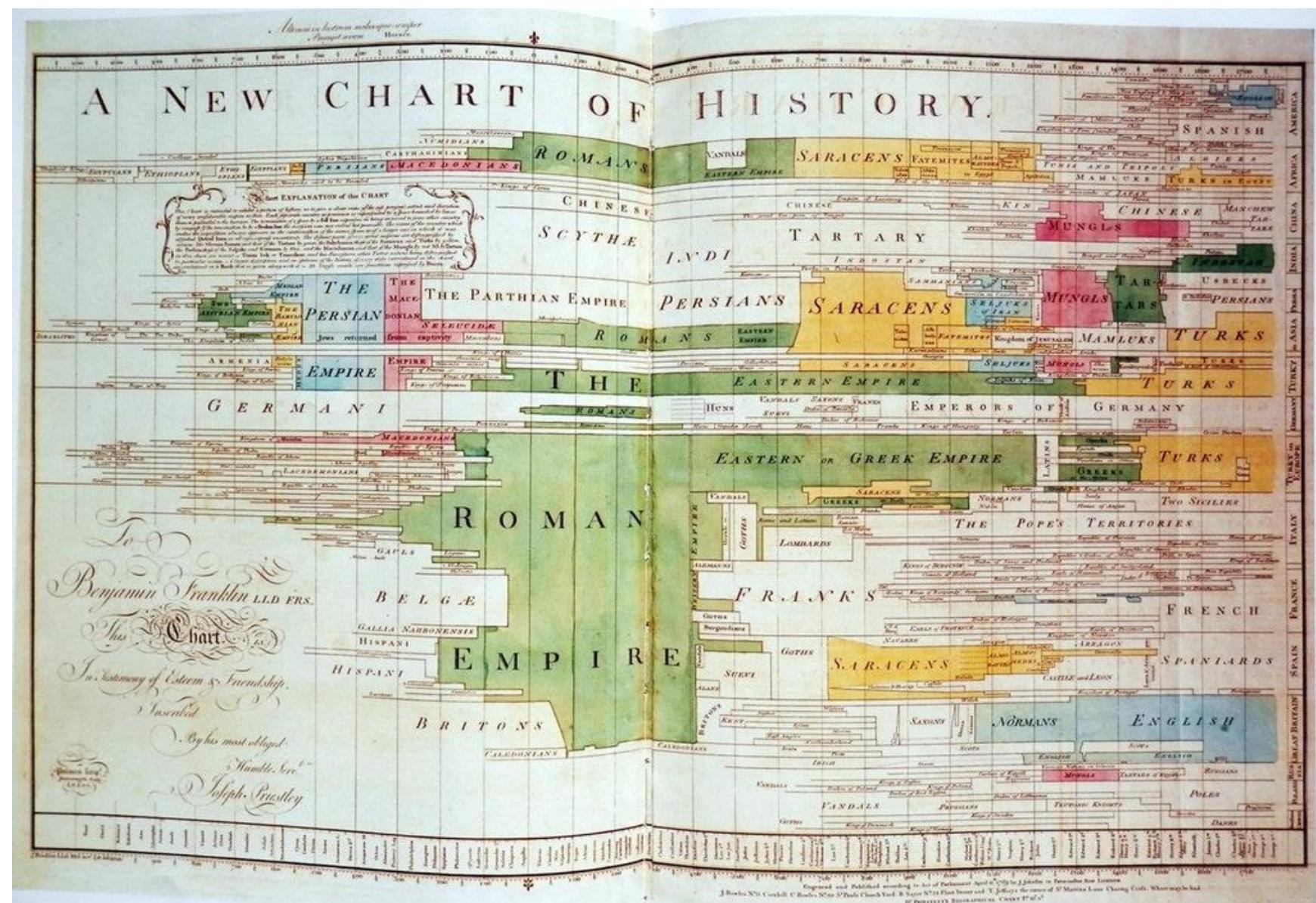
F.J. Anscombe Graphs in Statistical Analysis // American Statistician, 27 (February 1973), 17-21.

Немного об истории визуализации

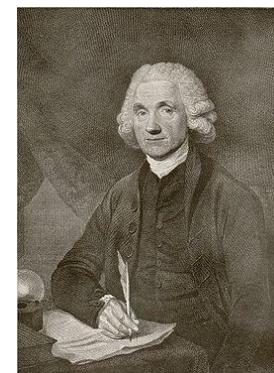
**инфографика
виды графиков
графический анализ**

**18 век – зарождение инфографики
19 век –protoанализ данных**

18 век Джозеф Пристли



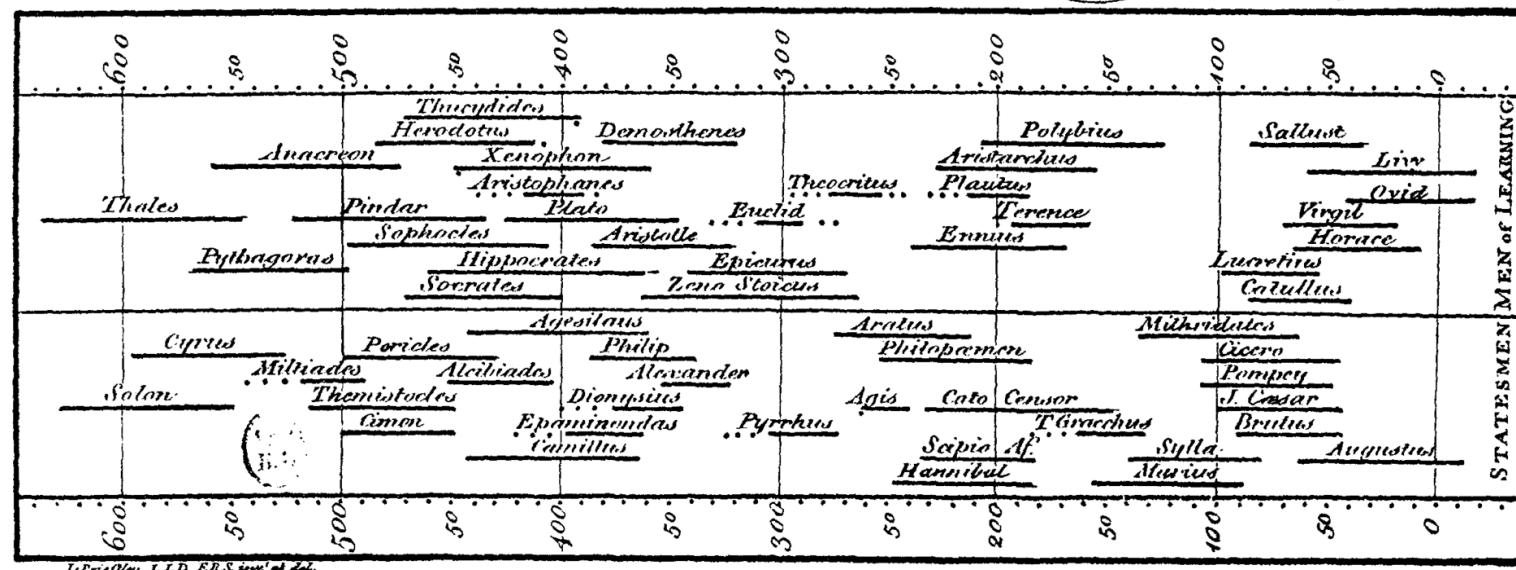
18 век Джозеф Пристли



Joseph Priestley

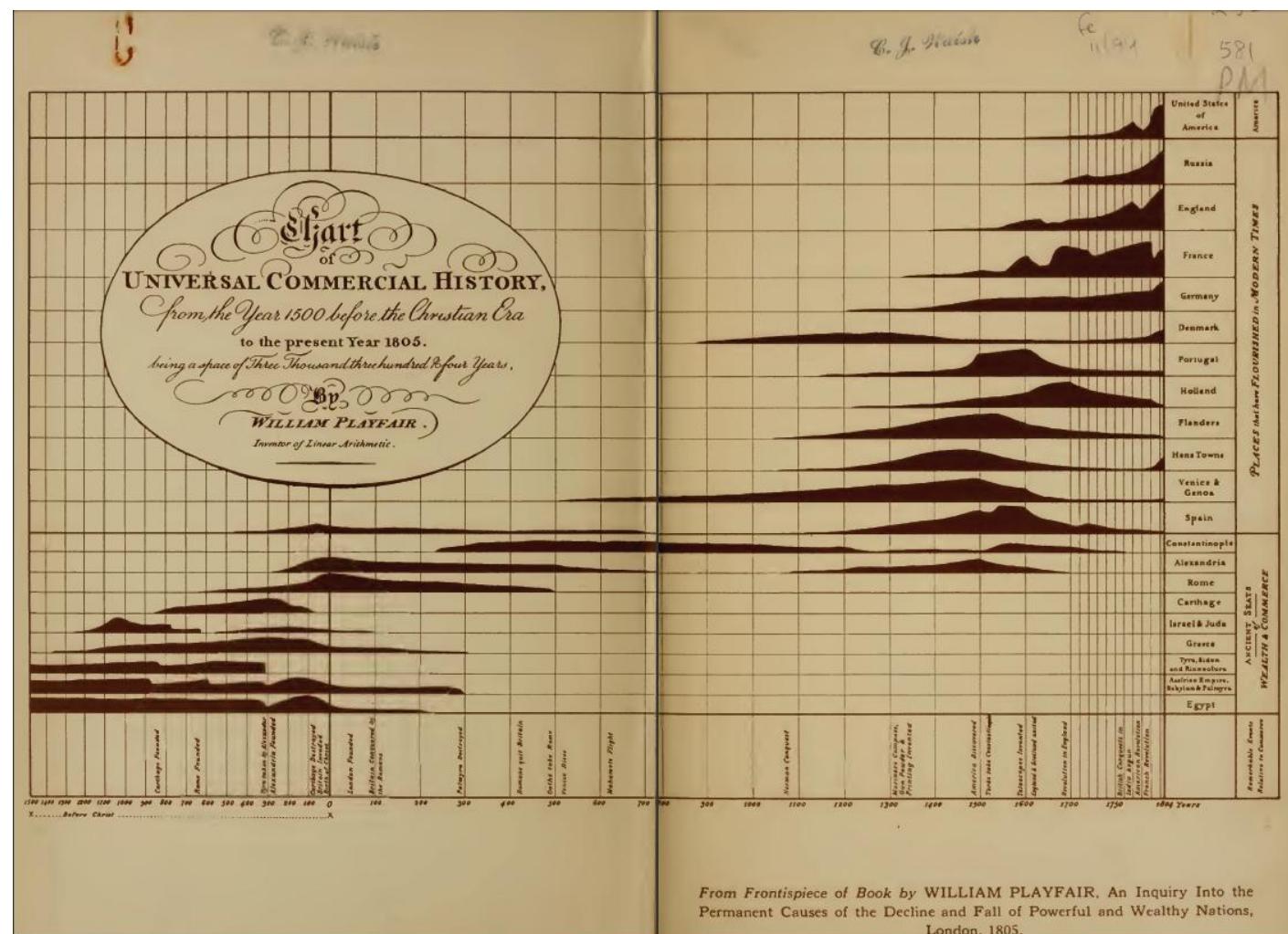
(13.03.1733 – 6.02.1804) британский
священник, естествоиспытатель,
философ. Открыл кислород.

A Specimen of a Chart of Biography.

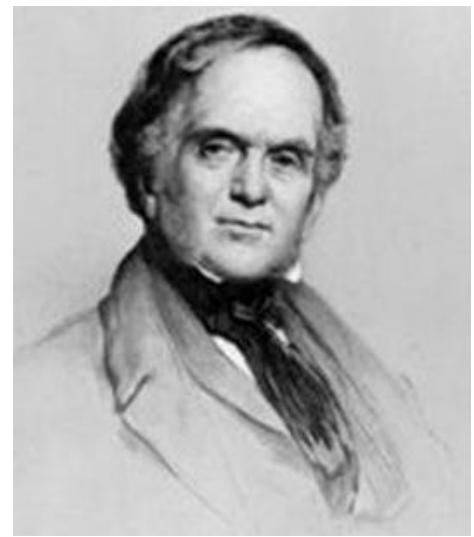
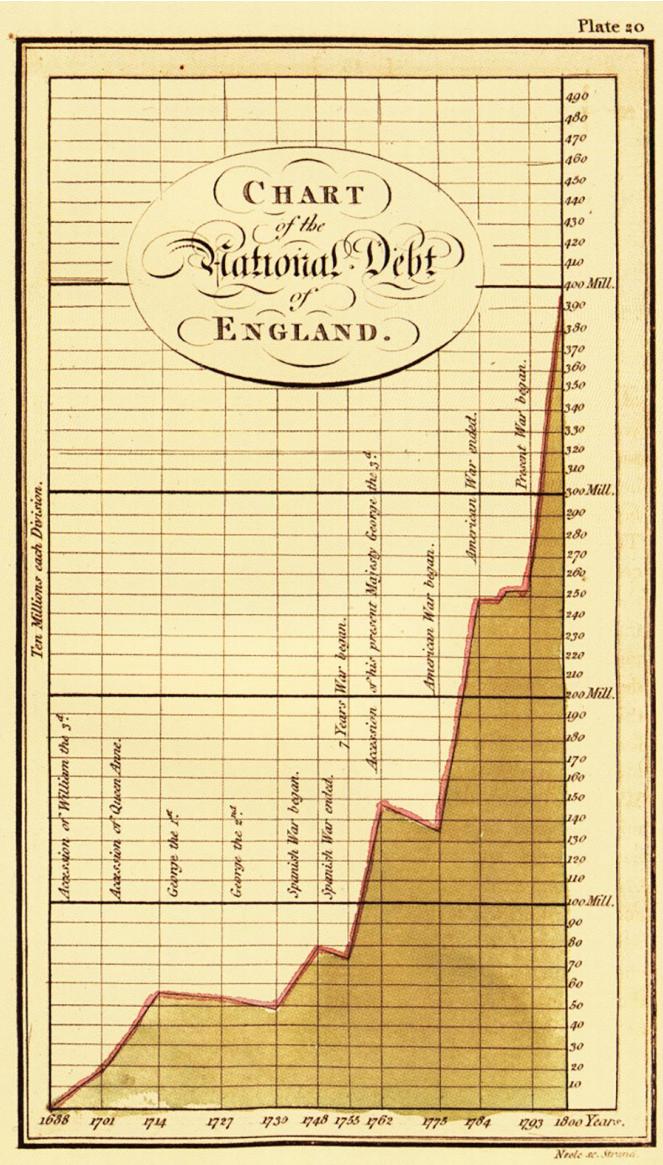
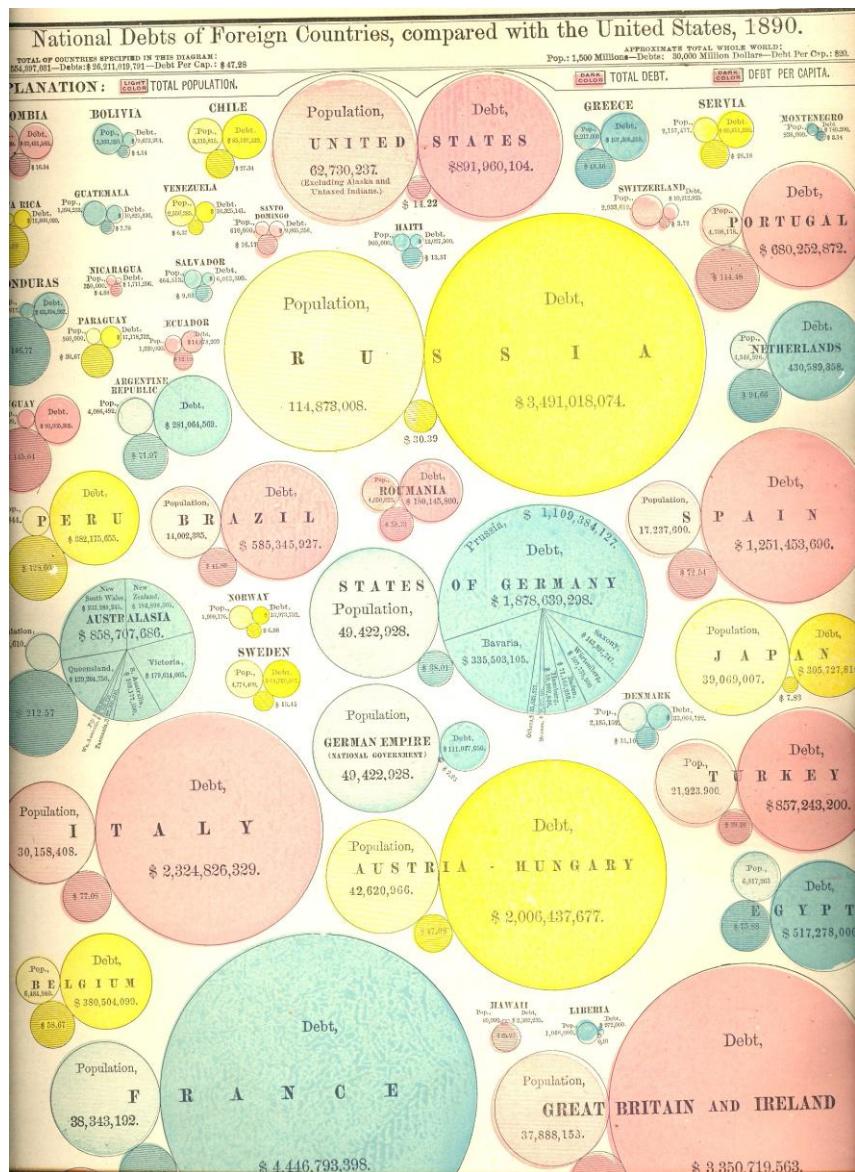


18 век Уильям Плейфэр

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

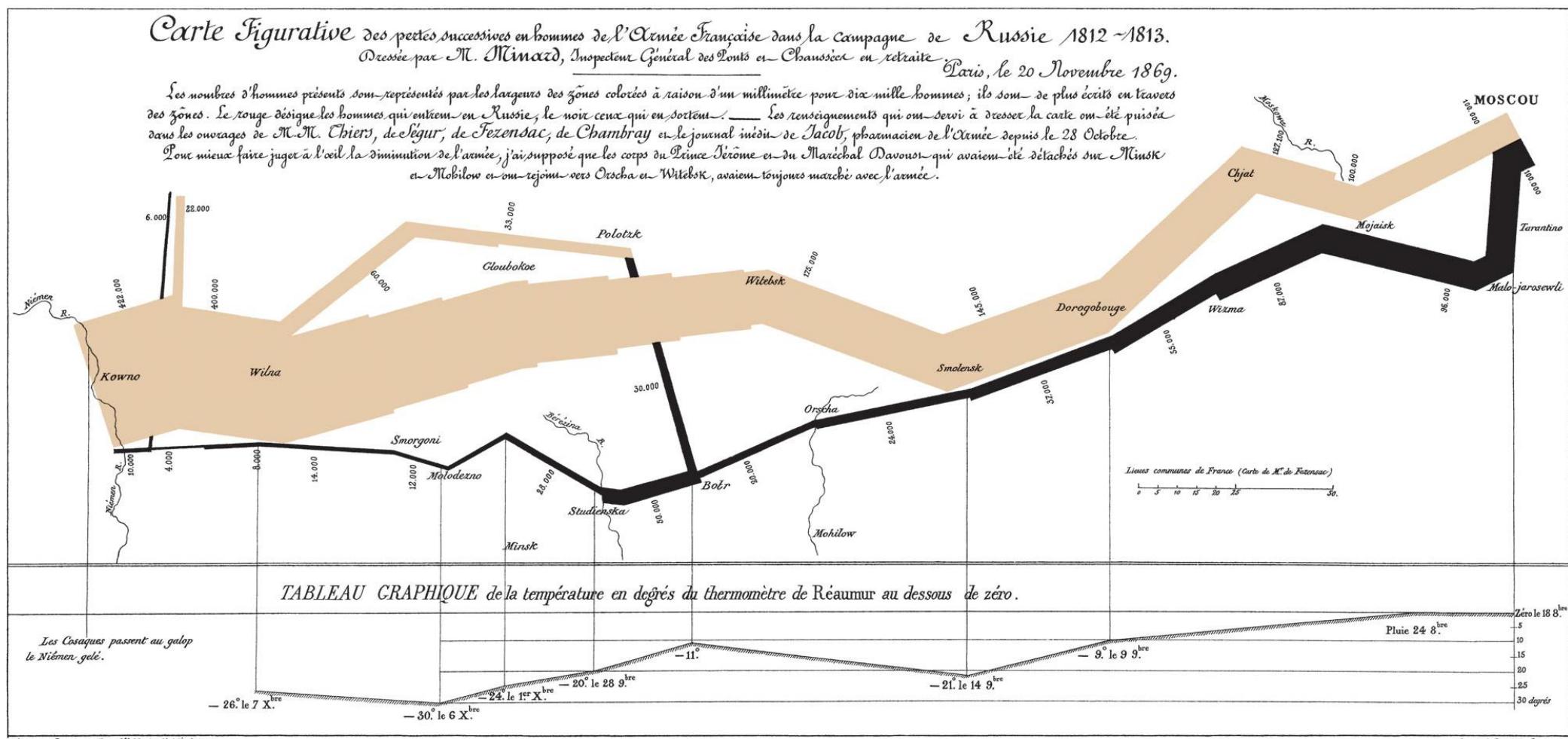


18 век Уильям Плейфэр

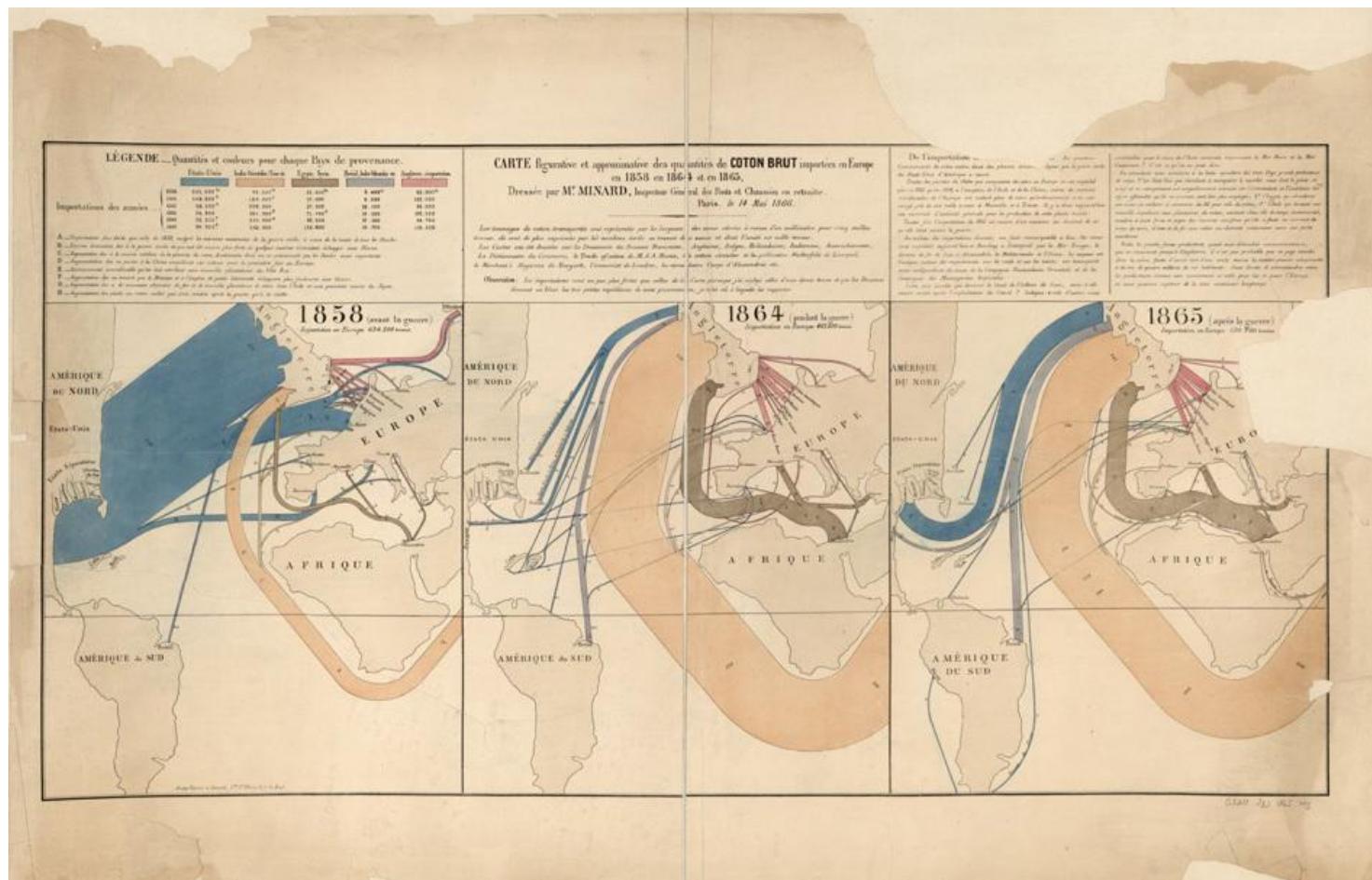


22.09.1759 – 11.02.1823
шотландский инженер,
основатель
графических методов
статистики.

19 век Шарль Жозеф Минар



19 век Шарль Жозеф Минар

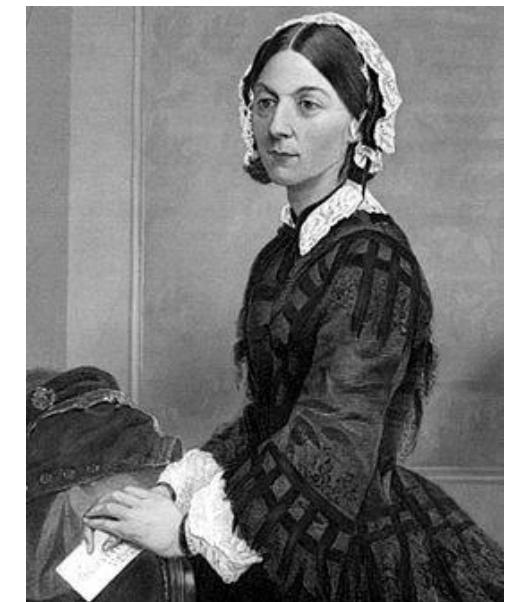
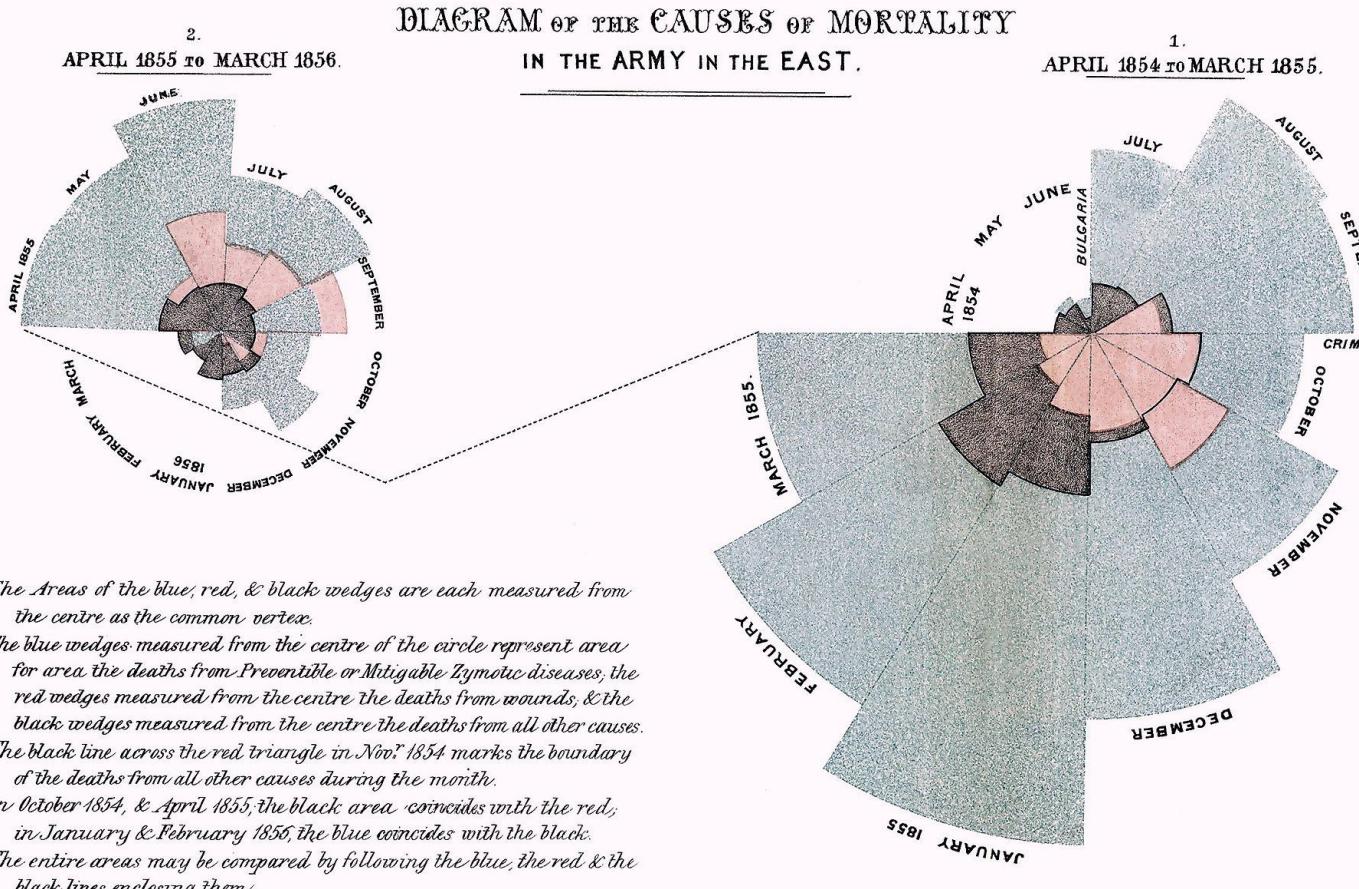


**Шарль Жозеф Минар
27.03.1781 – 24.10.1870**

**французский инженер,
топограф, пионер в
области графических
методов анализа и
представления
информации в области
инженерных наук и
статистики**

19 век – Флоренс Найтингейл

«Петушиный гребень»



Florence Nightingale

12.05.1820 –

13.09.1910

**сестра милосердия
и общественная
деятельница
Великобритании.**

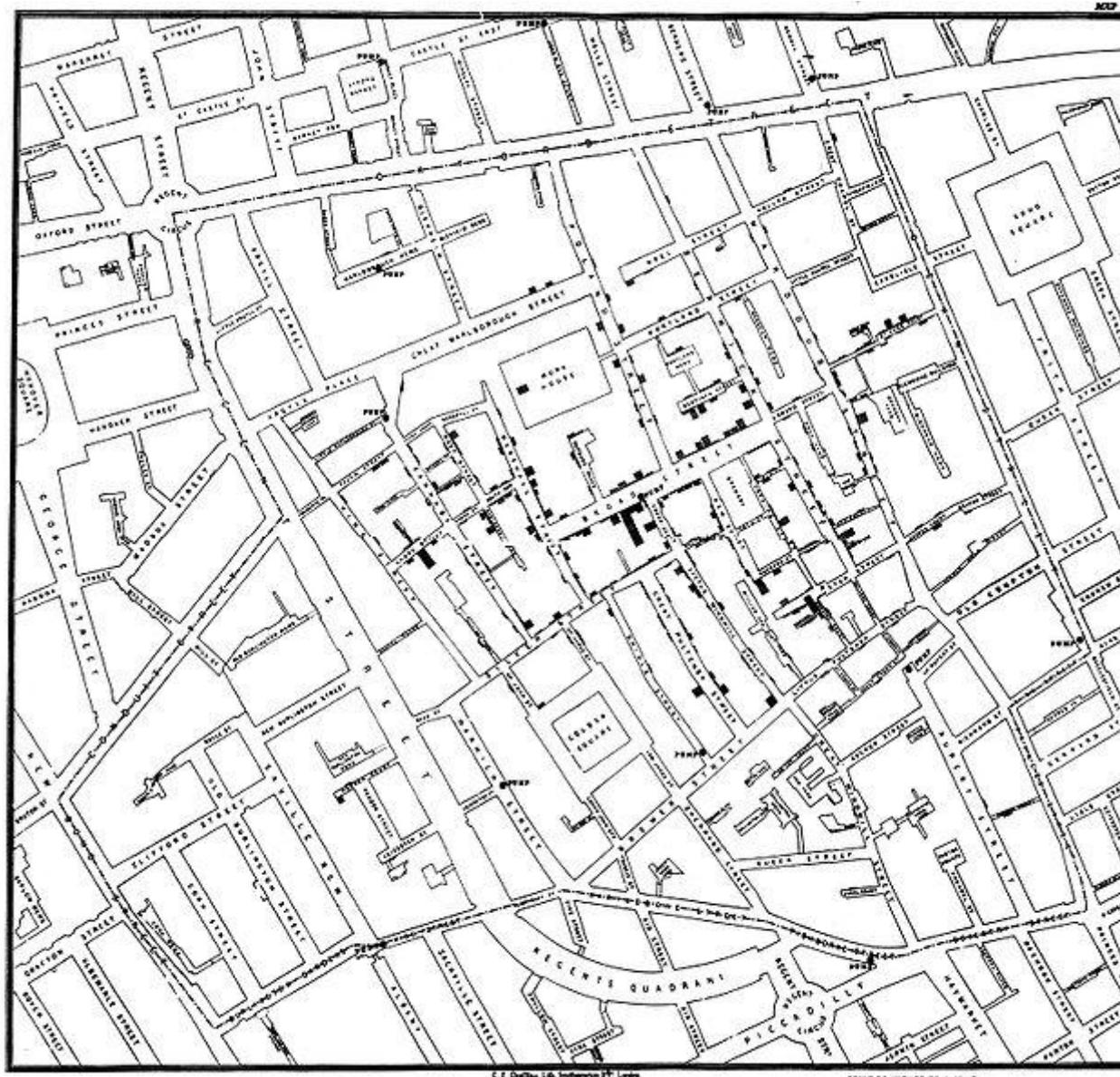
Смертность солдат во время крымской войны.

Площадь каждого сектора пропорциональна смертности.

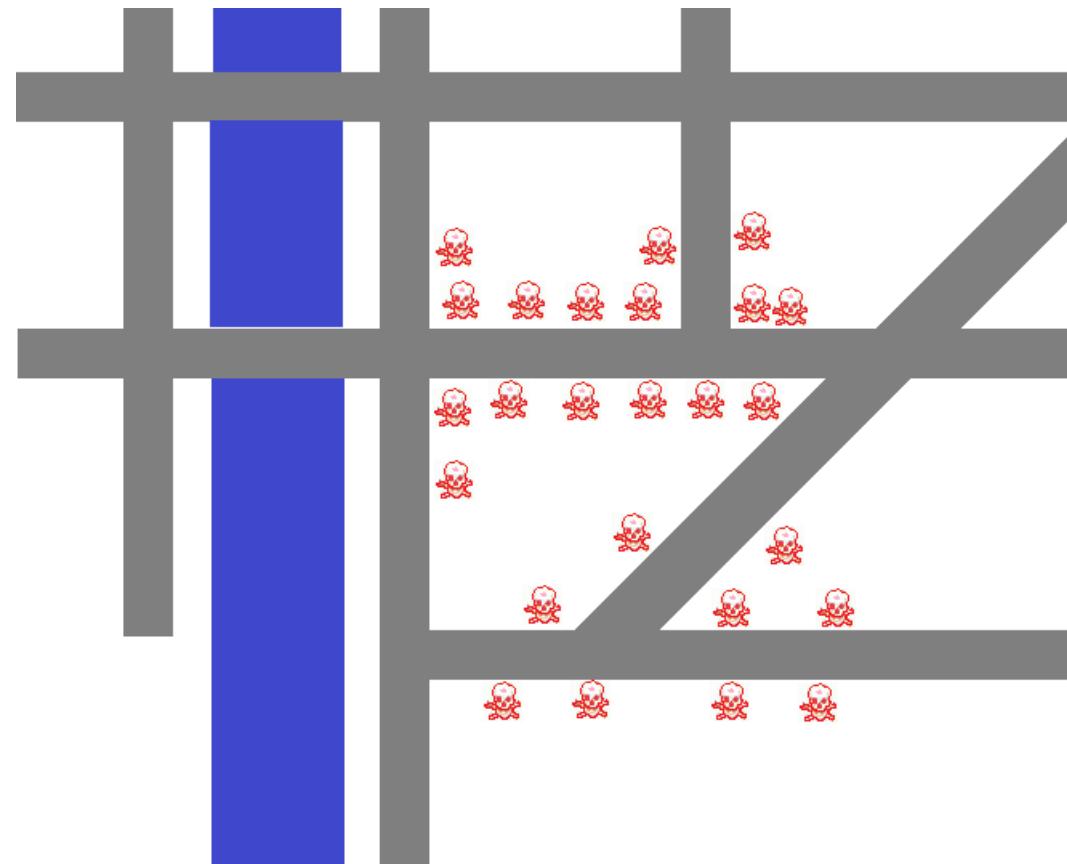
Голубой –смертность от болезней, красный – от ран, и коричневый слой – от других причин.

Вспышка холеры на Брод-стрит в 1854 году

См. Википедию



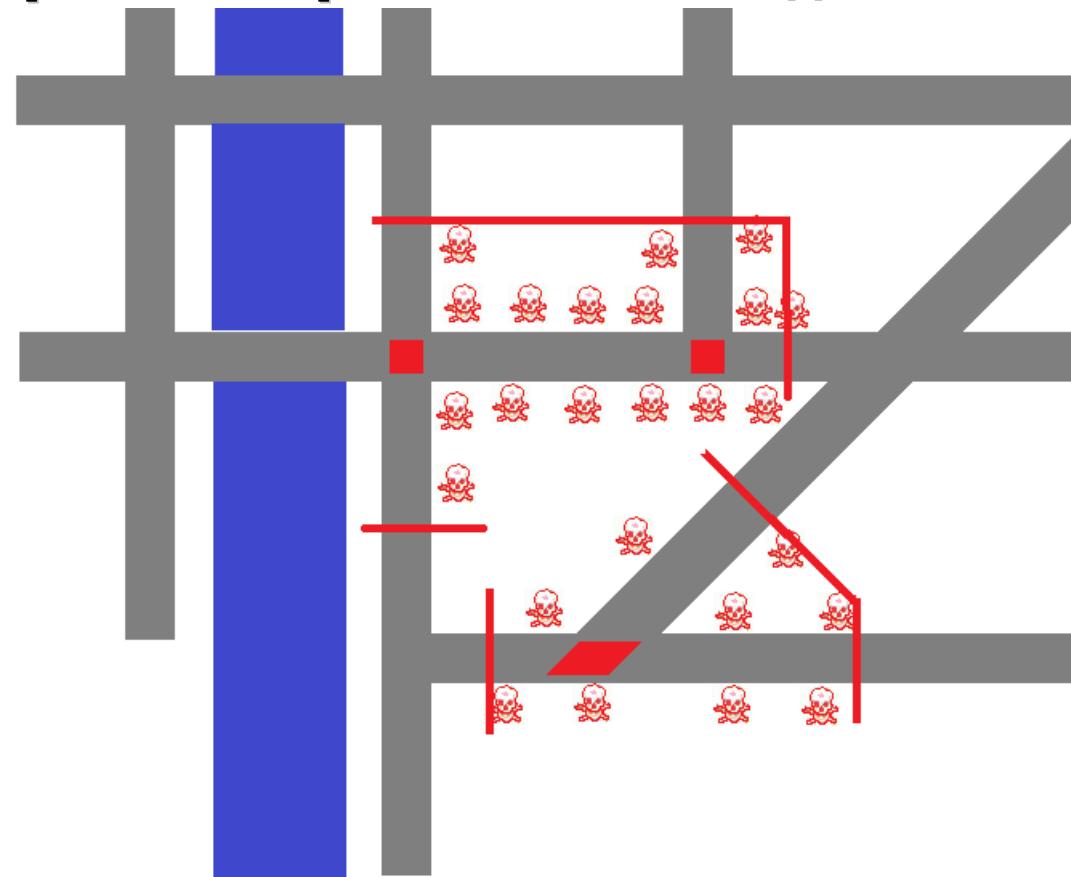
Статистика заболевания холерой



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

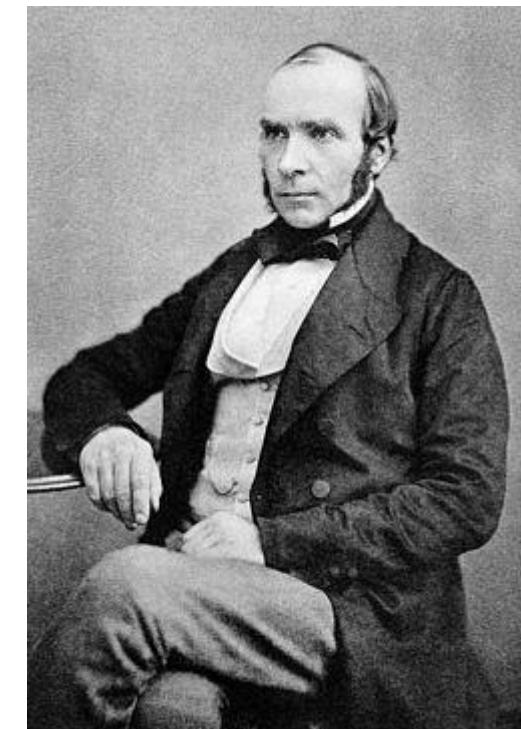
Центры эпидемии – колодцы!

Диаграмма Вороного была видна на карте...



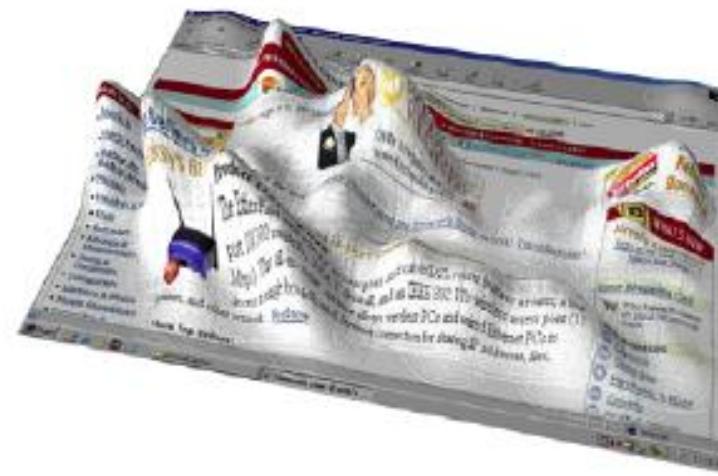
**Нечистоты сливались в Темзу,
в результате была заражена
местная система водоснабжения.**

Джон Сноу



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

Что это за данные?



Что это за данные?



Ответ

Плотность внимания на Интернет-странице
Измеряется по числу и продолжительности фиксаций

Viewer 1:



Viewer 2:



POR data

Gaze clusters with $\sigma_s = 100$, $\sigma_t = \frac{1}{2}$ Regions-of-interest with $\sigma_s = 100$

**L. A. Granka, H. A. Hembrooke, G. Gay, M. K. Feusner Correlates of Visual Salience and Disconnect:
An Eye-tracking Evaluation**

**A. Santella D. DeCarlo Robust Clustering of Eye Movement Recordings for Quantification of Visual
Interest**

Что за данные?

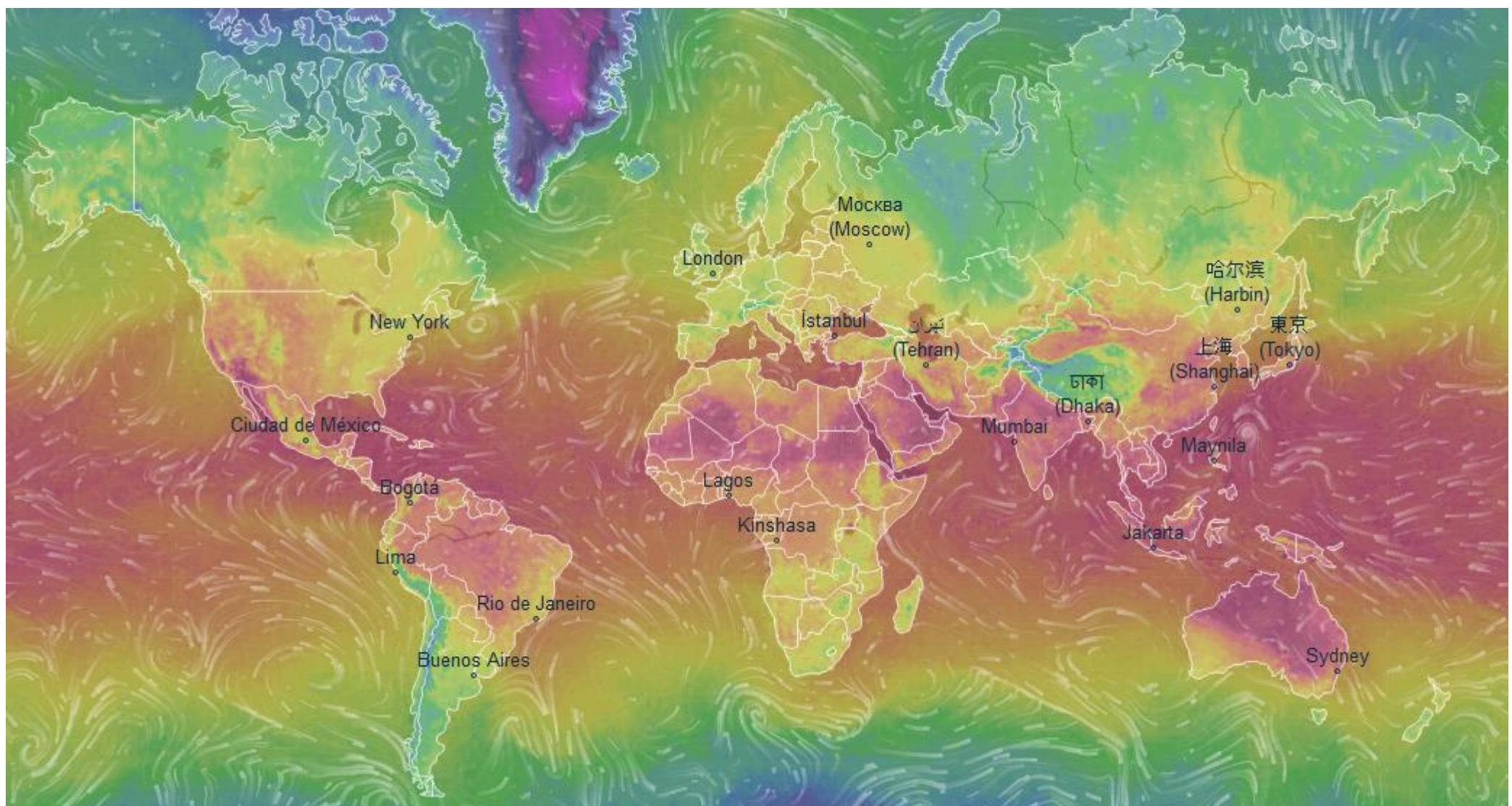


Что за данные?

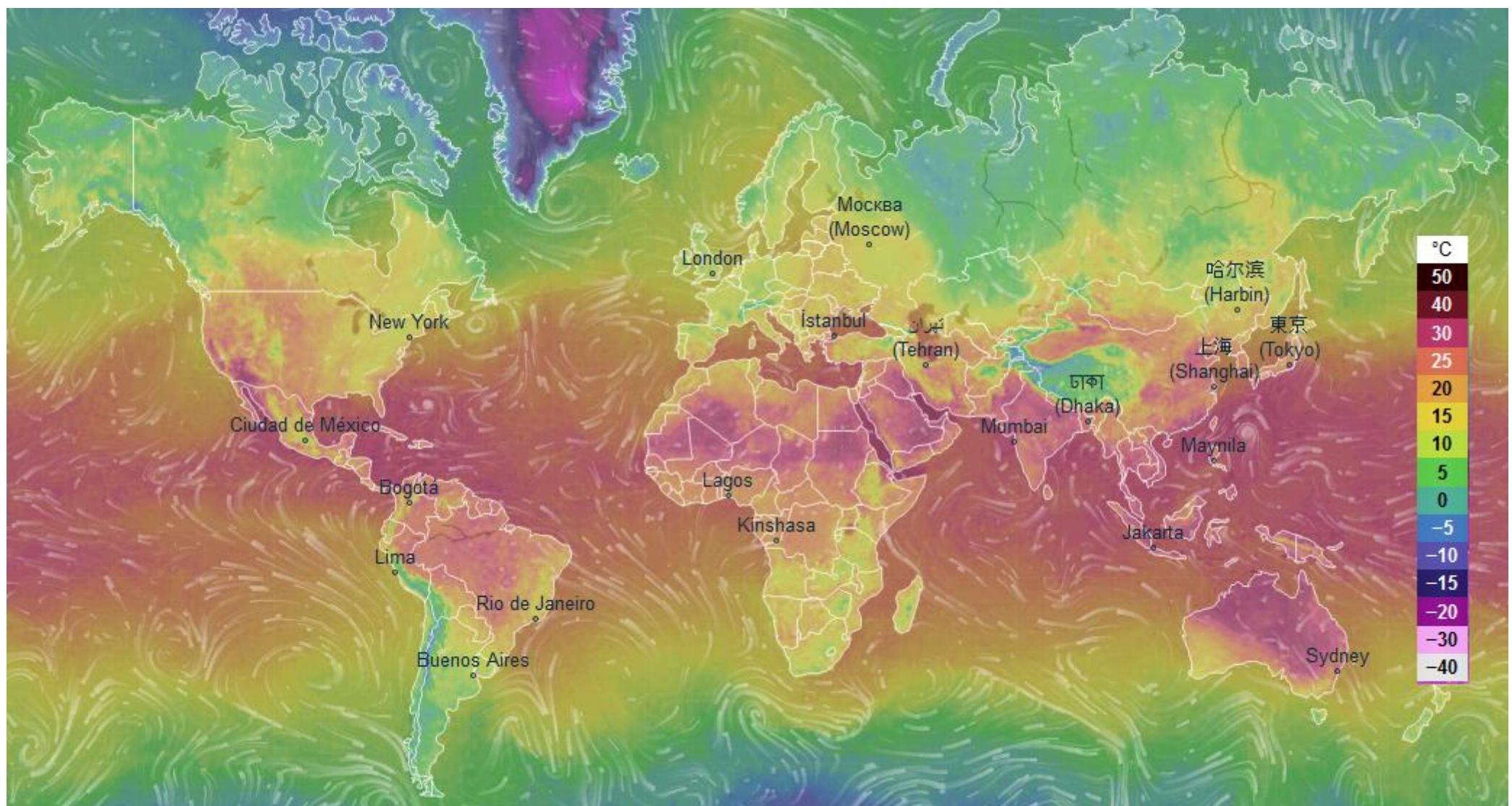


Трафик авиарейсов США

Что за данные?

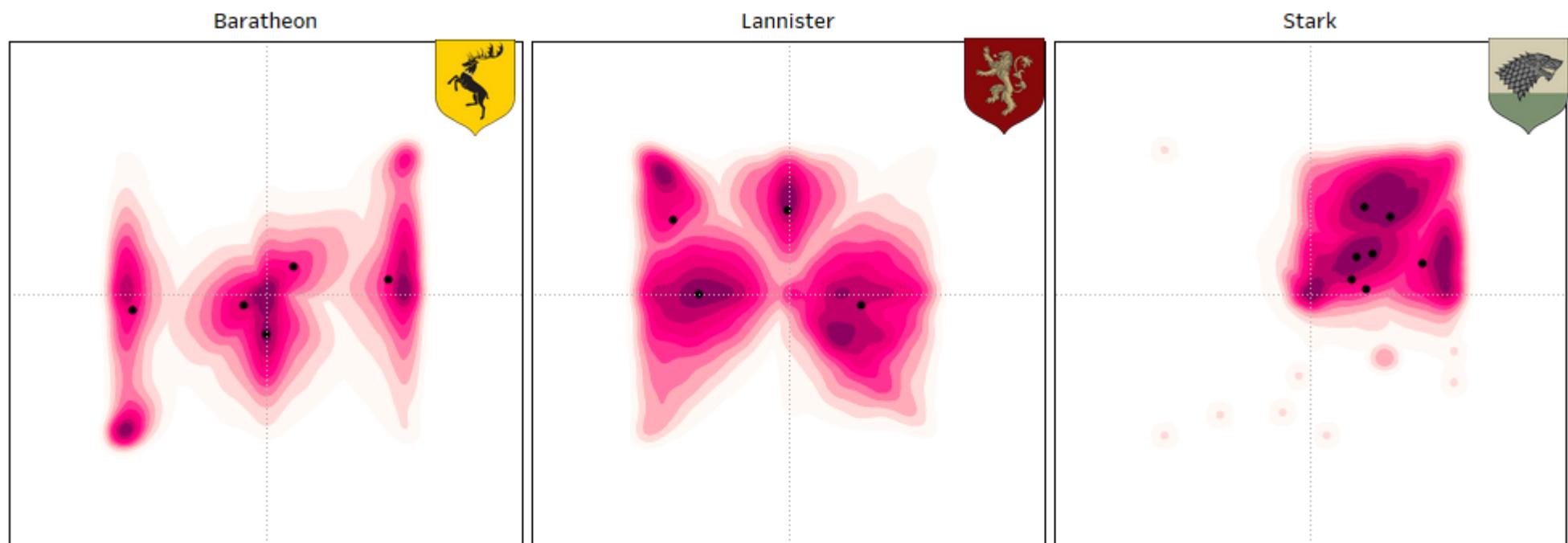


Что за данные?

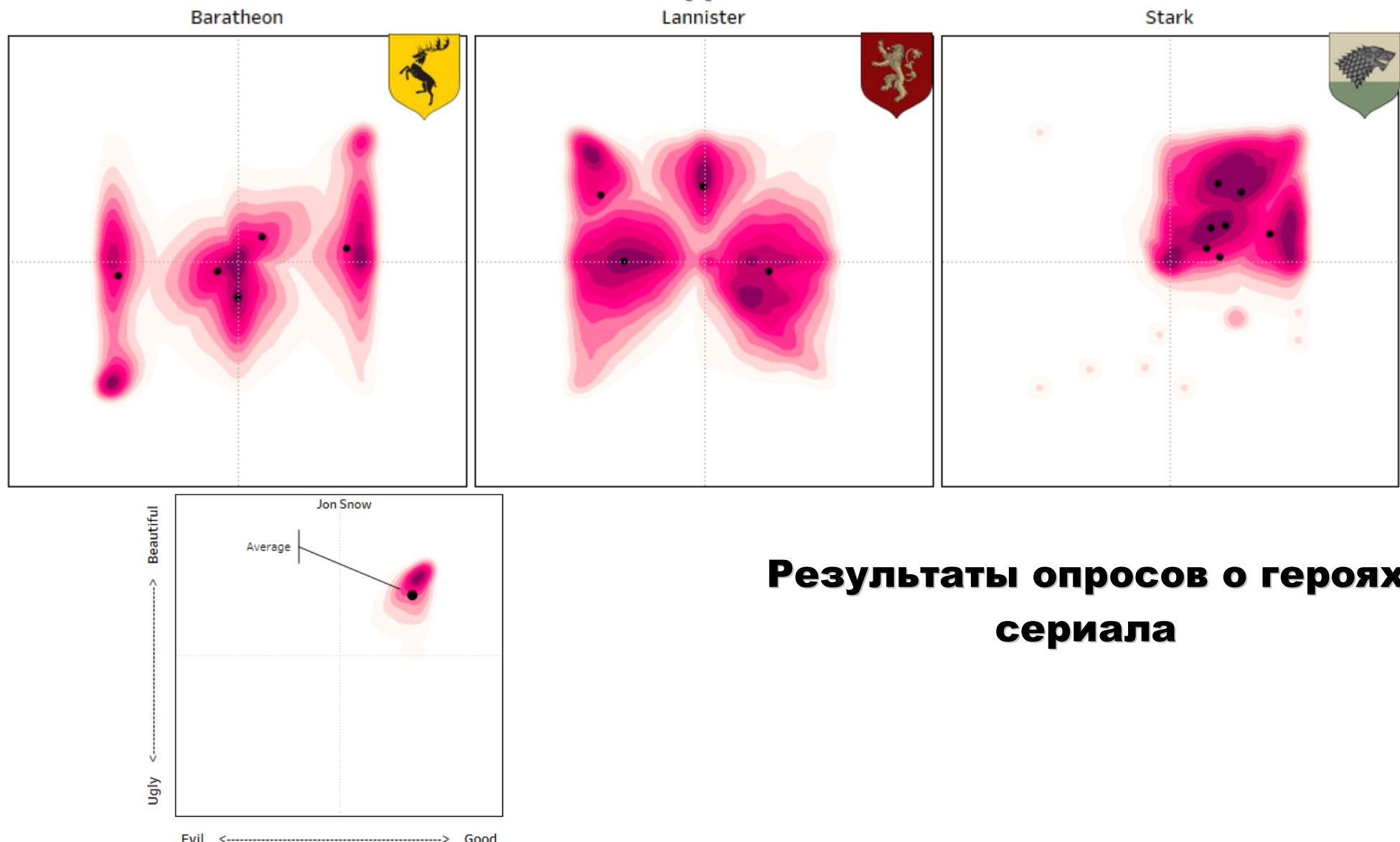


Температура (2м над землёй) <https://www.ventusky.com/>

Что за данные?



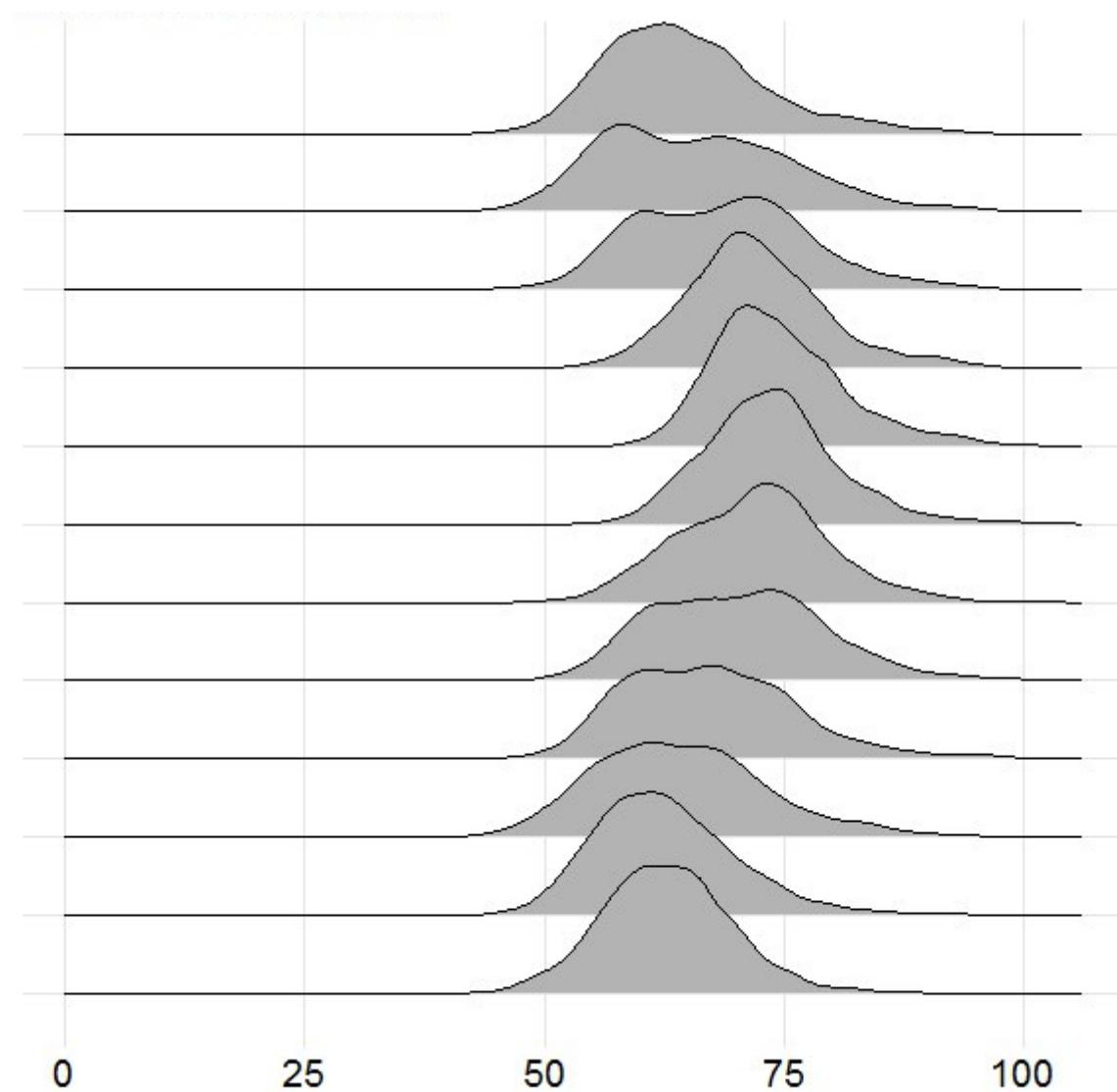
Что за данные?



**Результаты опросов о героях
сериала**

<https://public.tableau.com/profile/nicco.cirone#/vizhome/GameofThronesastudyofthemaincharacters/GameOfThronesAstudyoncharacters>

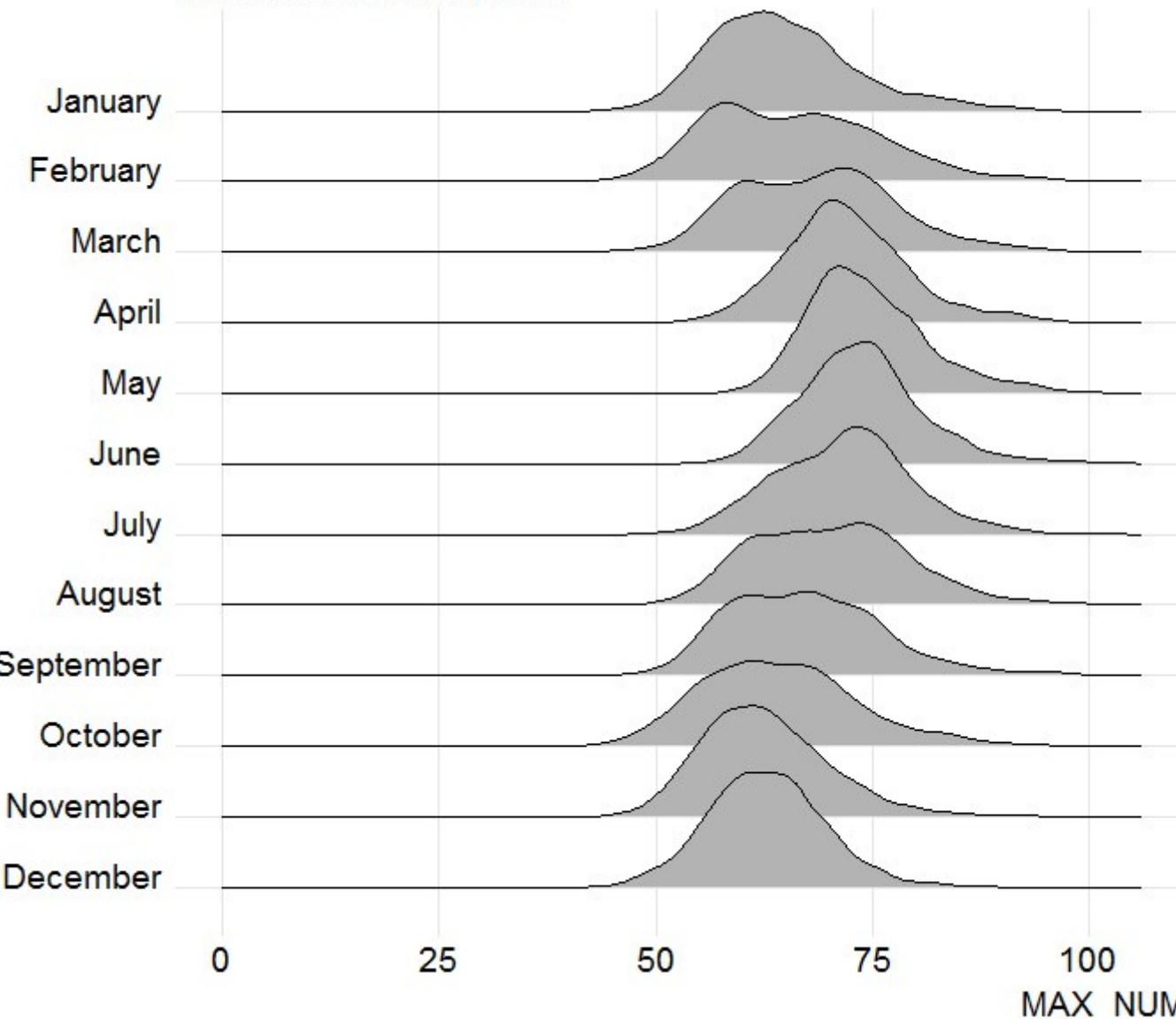
Что это за данные?



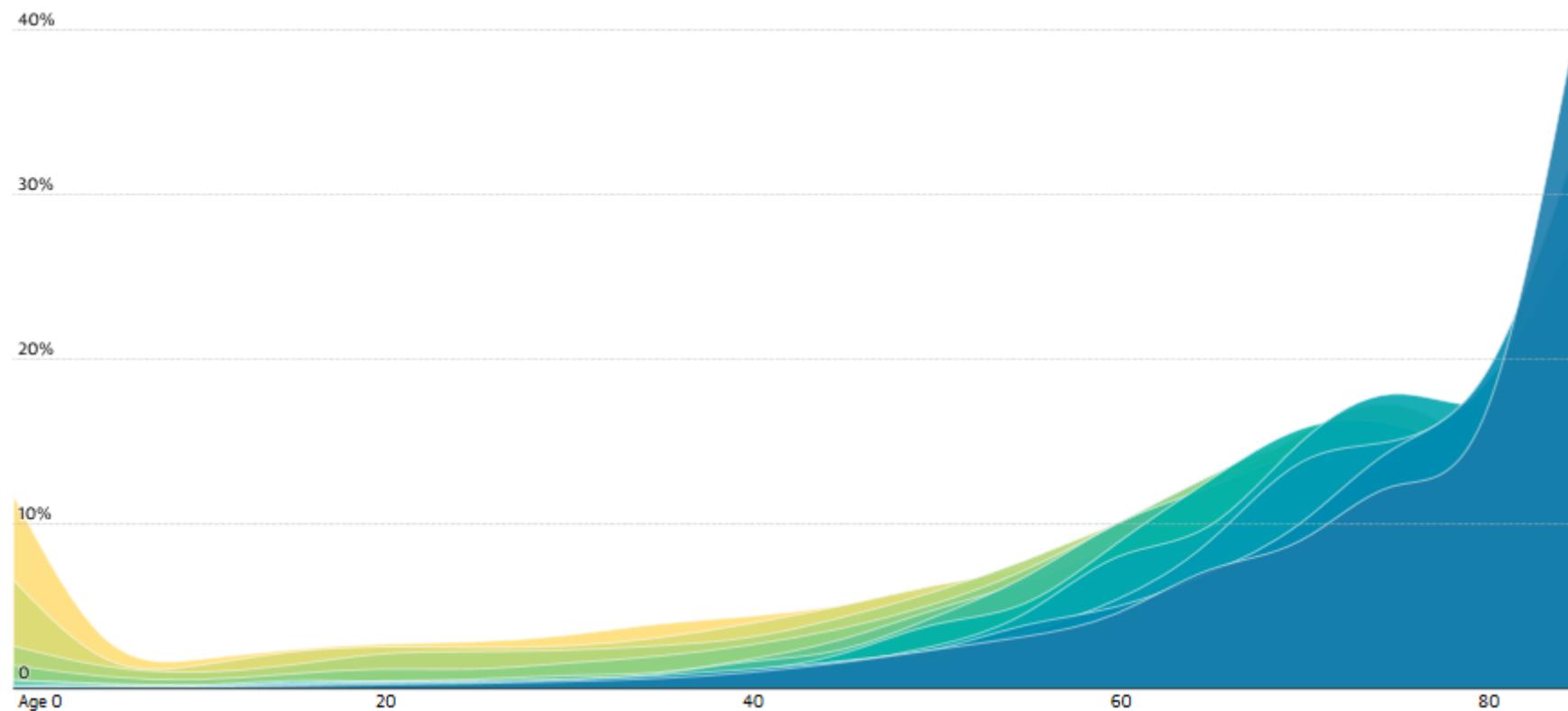
Что это за данные?

MAX Daily Temperatures in Mountain View, CA (1950-2016)

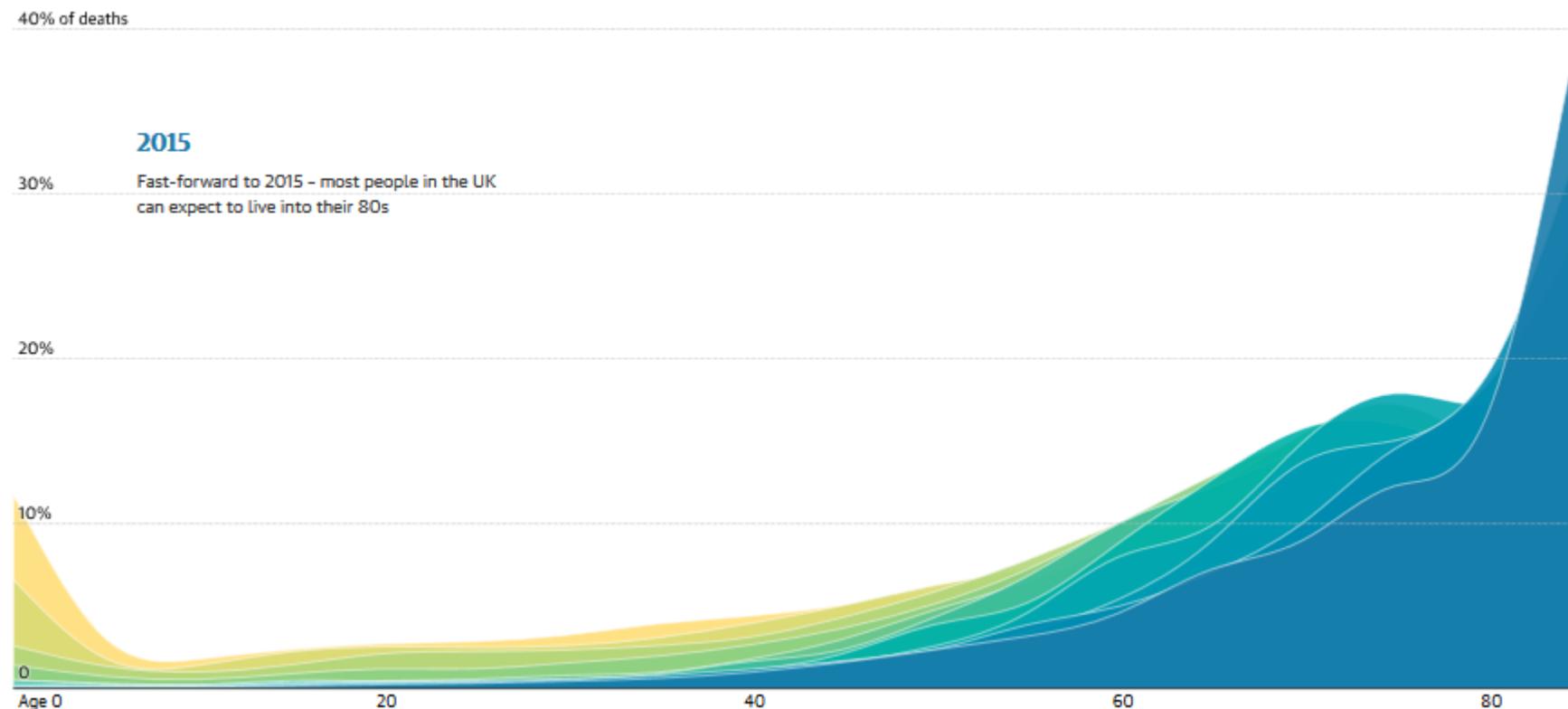
Thanks to austinwehrwein.com



Что за данные?



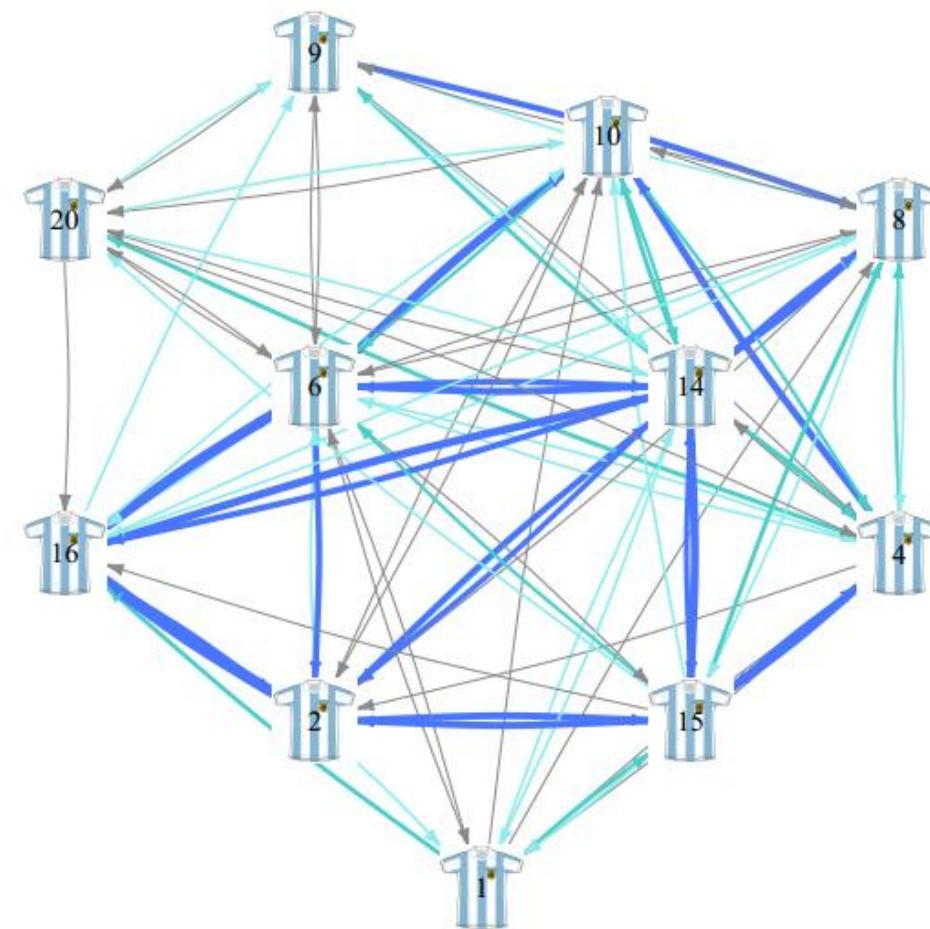
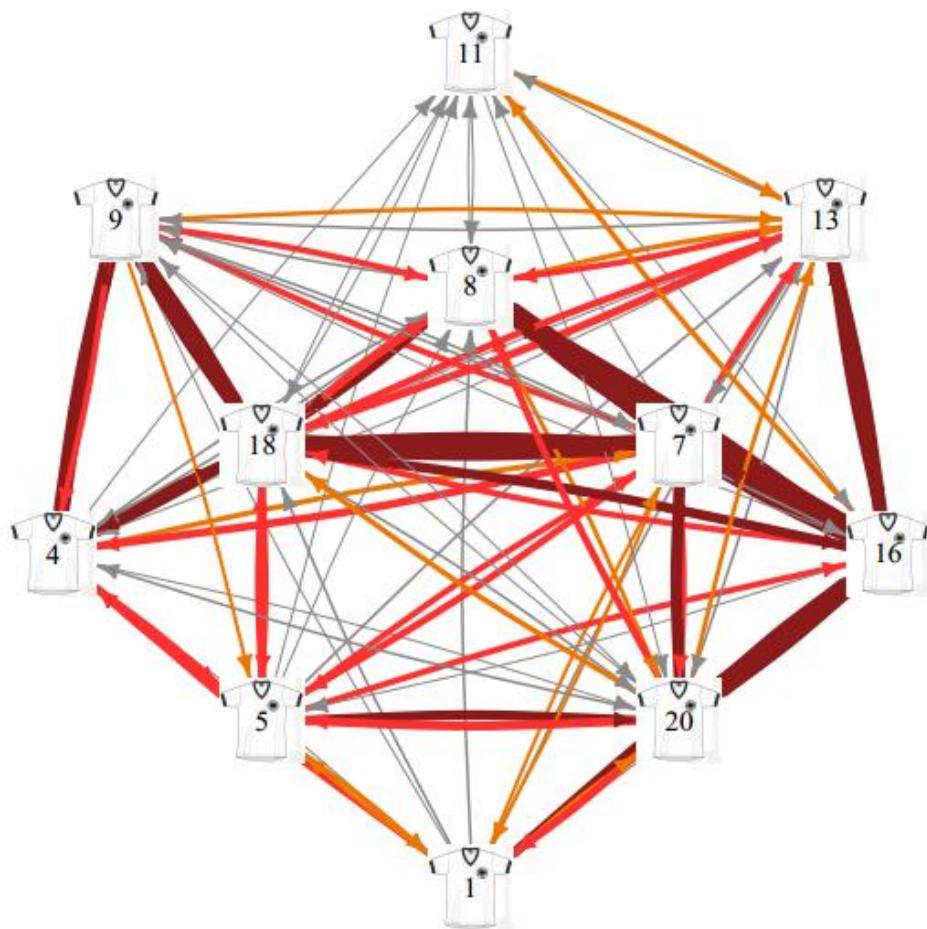
Что за данные?



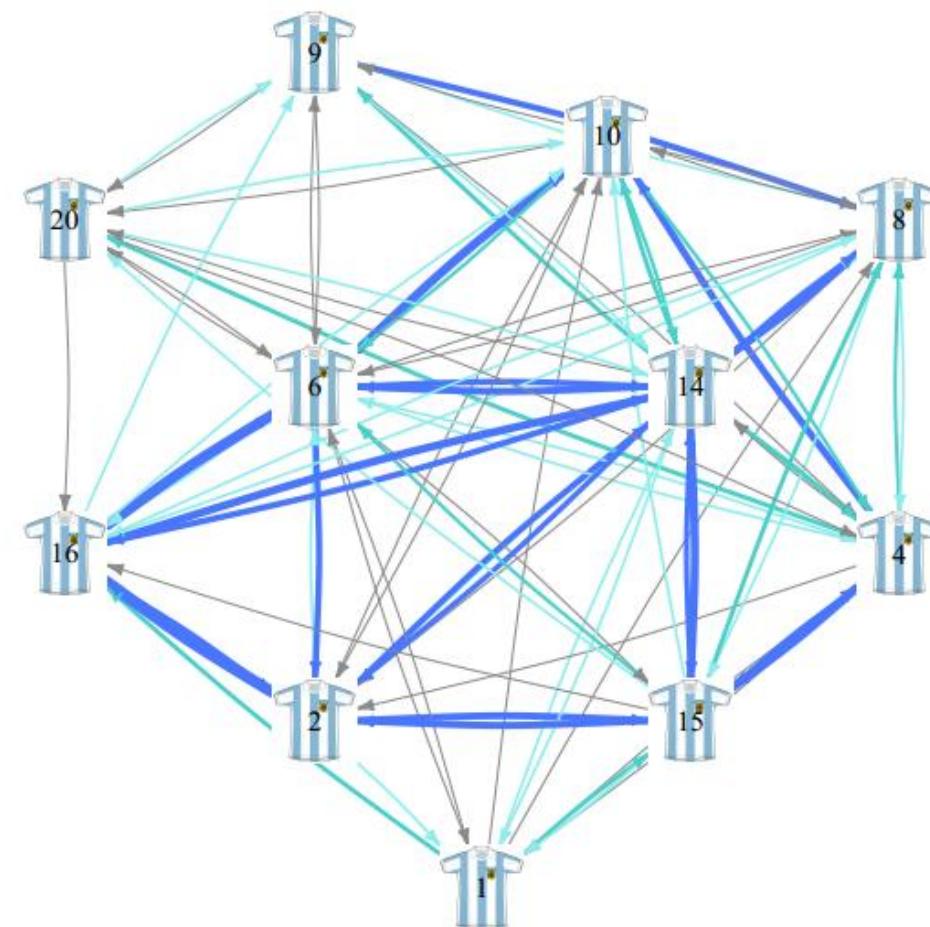
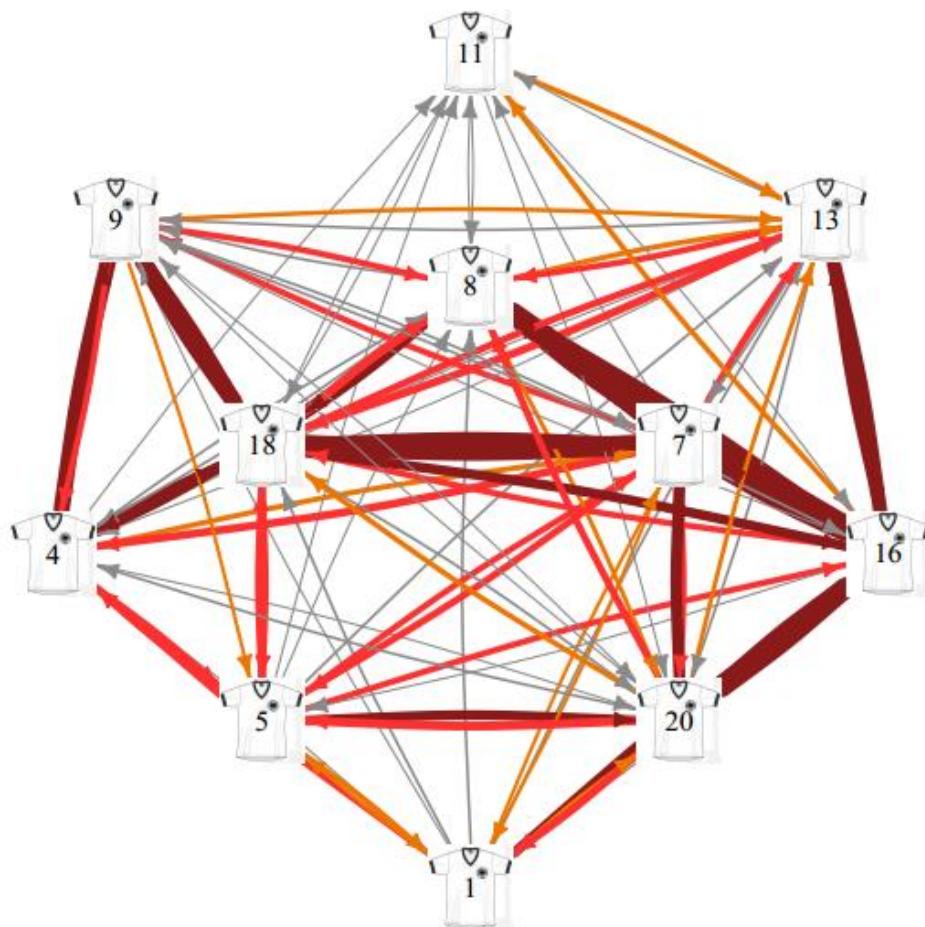
**Плотность распределения смертности по возрасту
в разные годы в Великобритании**

<https://www.theguardian.com/lifeandstyle/ng-interactive/2017/sep/18/how-death-has-changed-over-100-years-in-britain>

Что за данные?

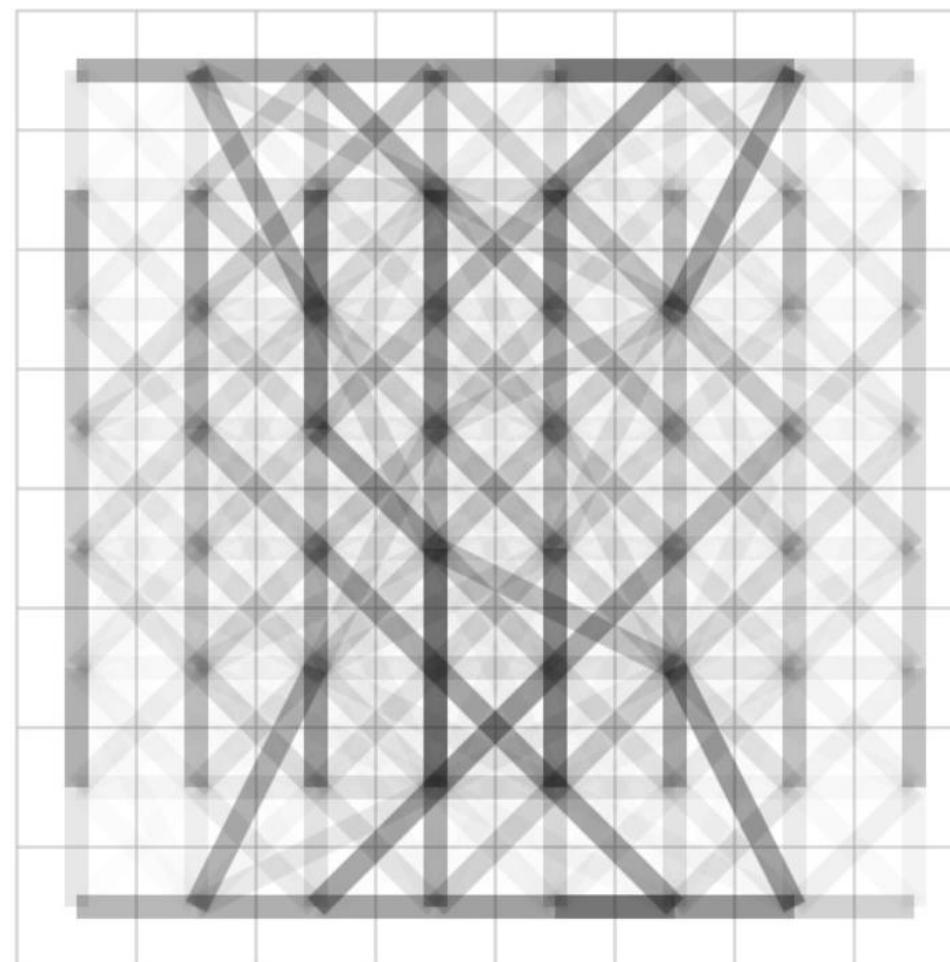


Что за данные?

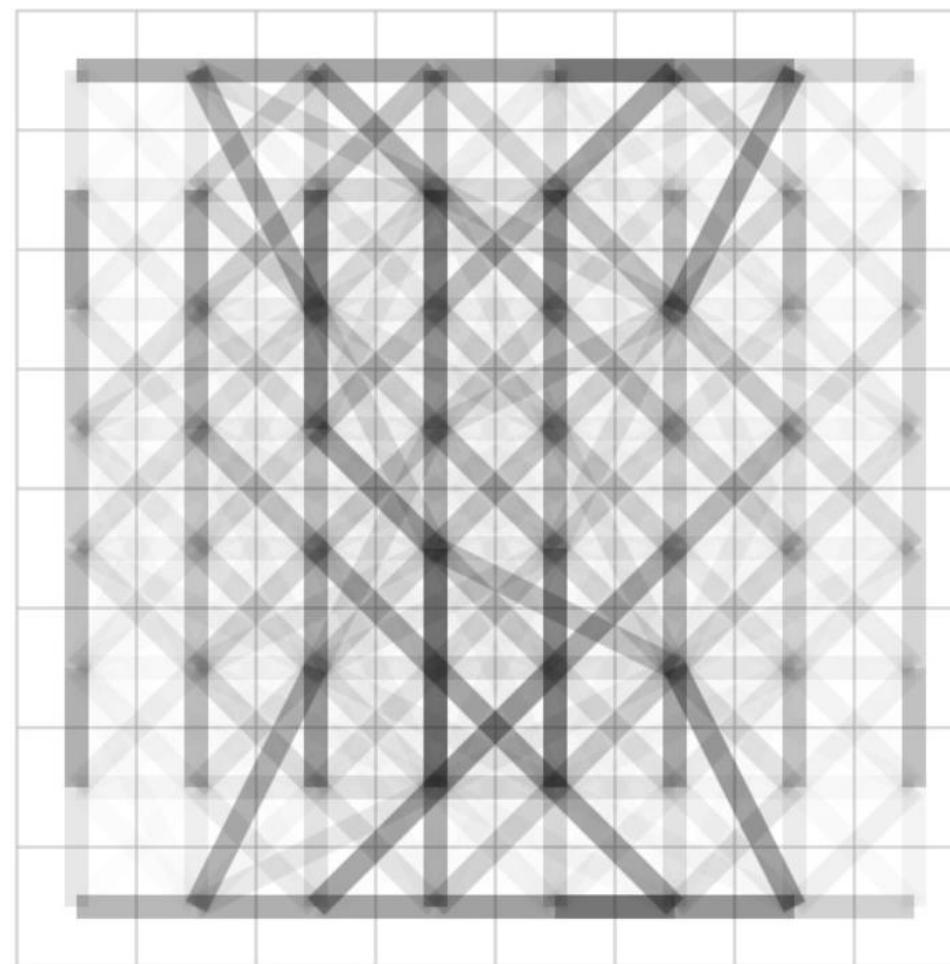


Передачи в финале ЧМ-2014 Германия – Аргентина

Что за данные?



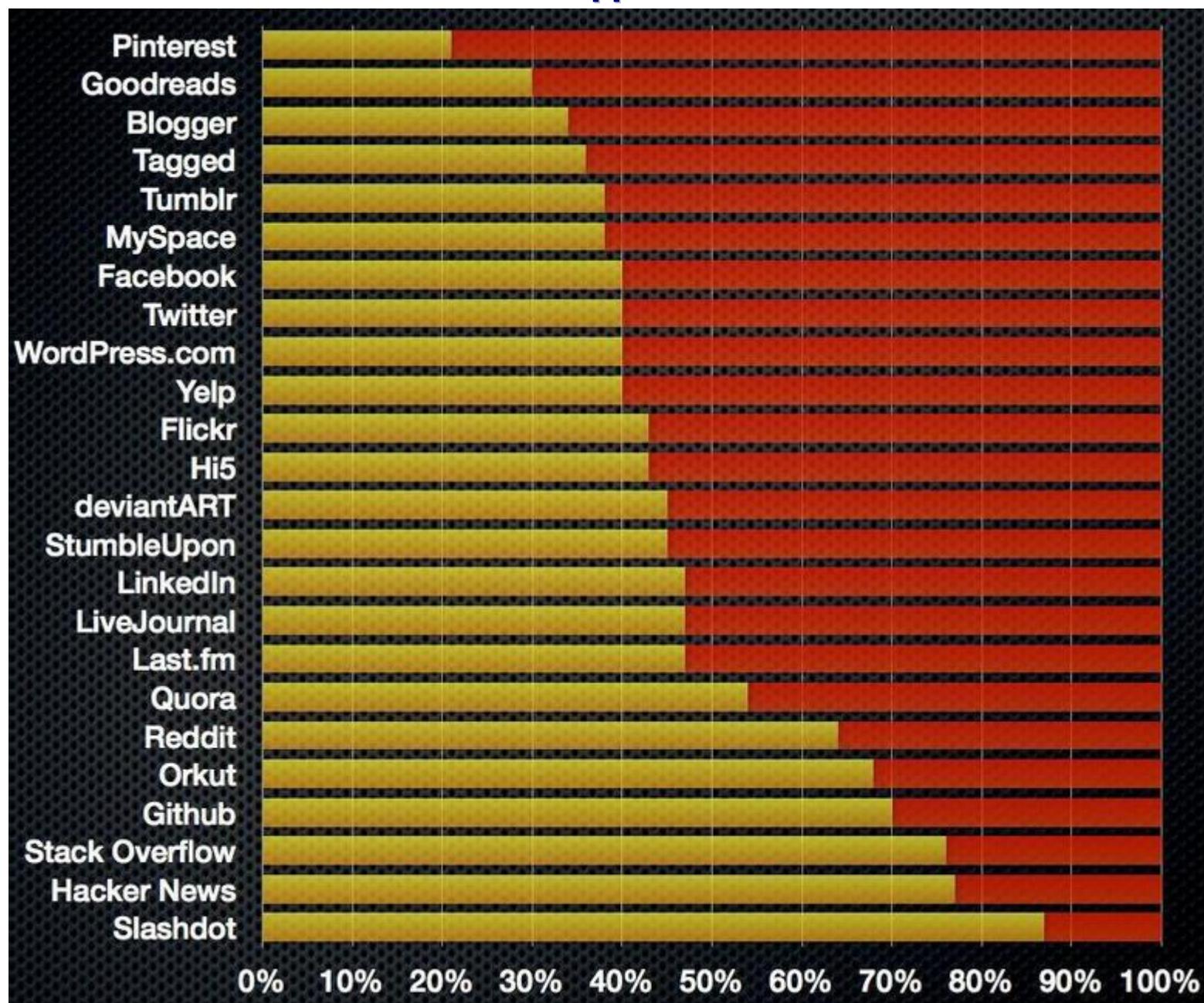
Что за данные?

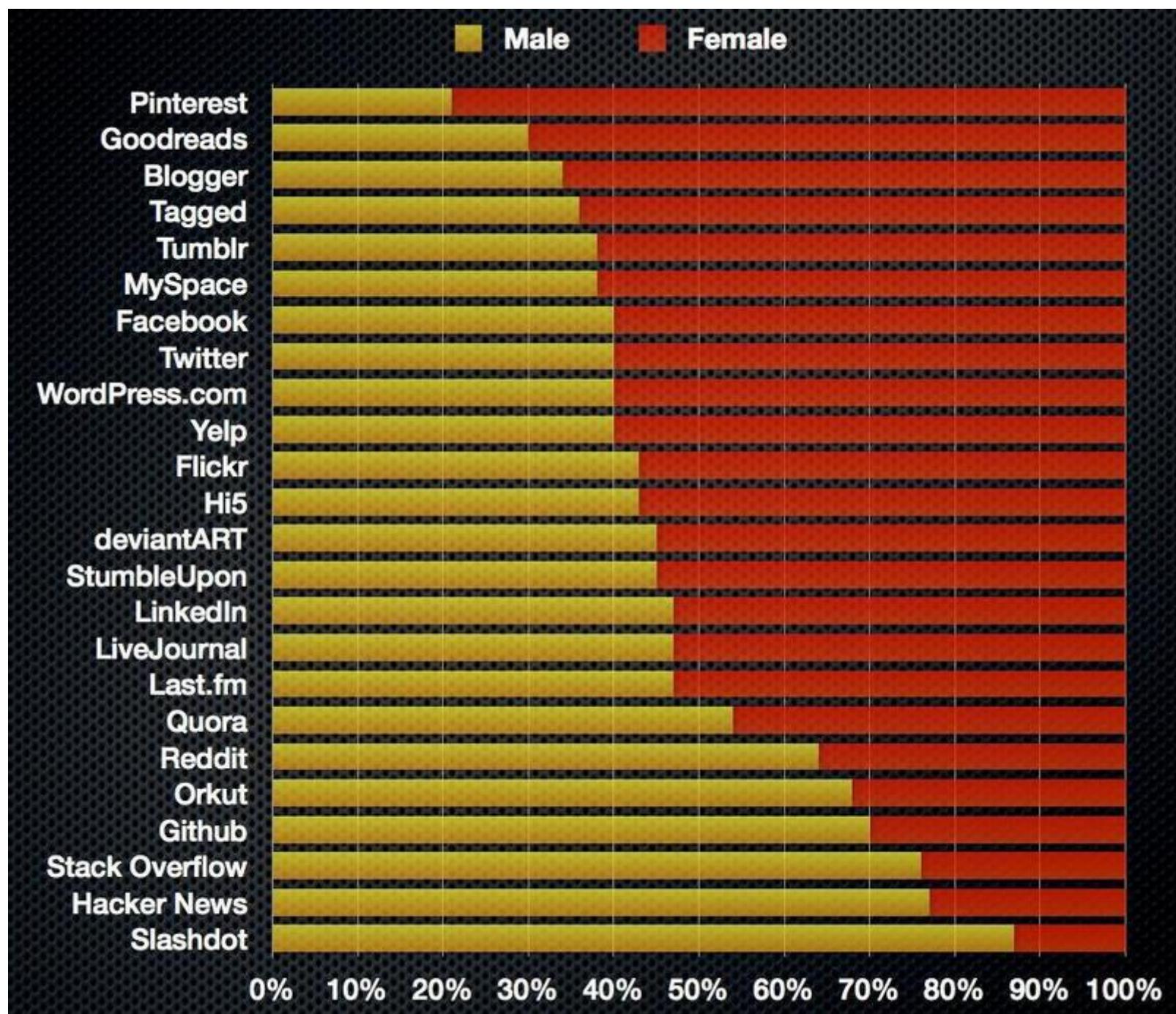


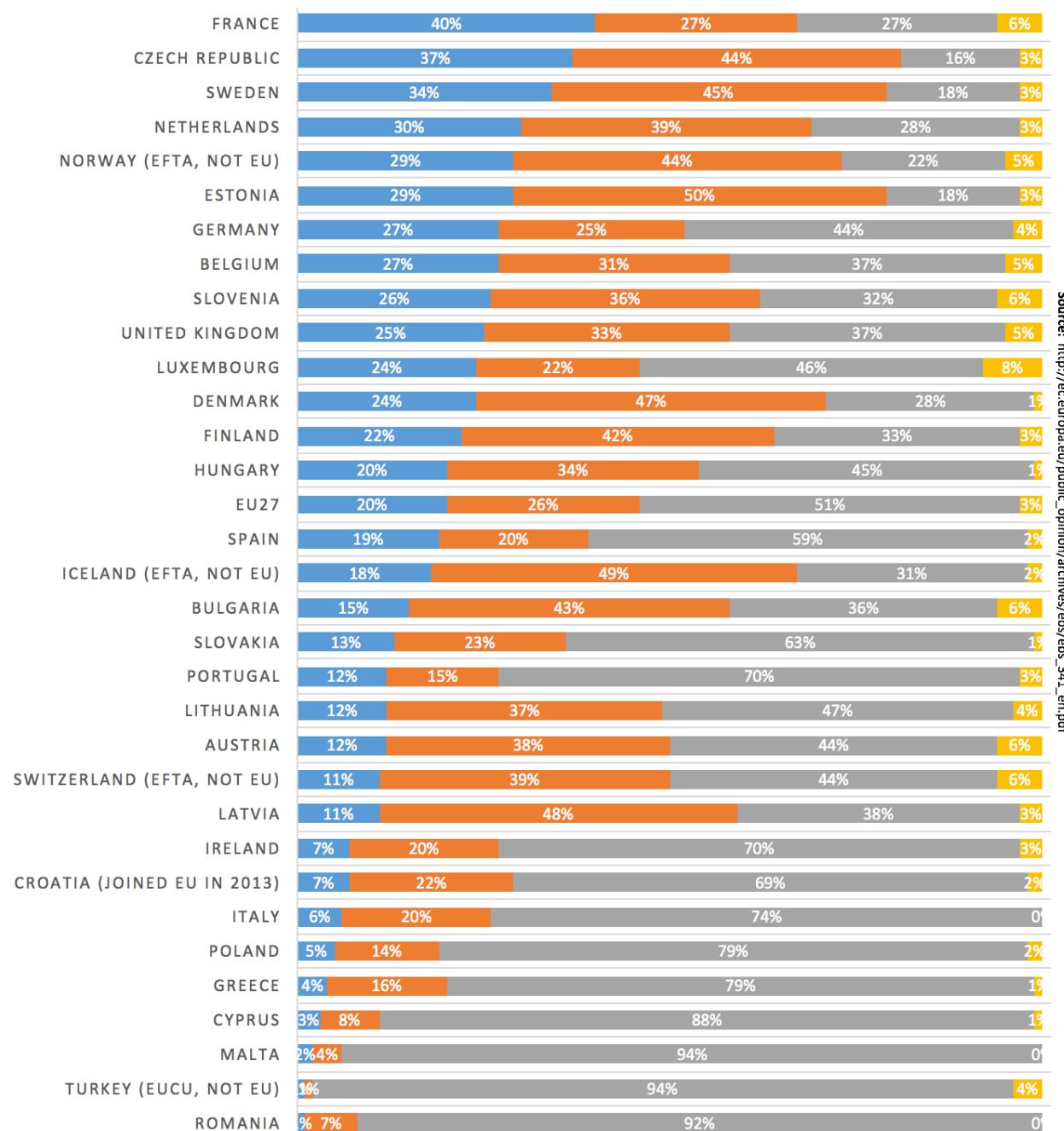
Перемещение фигур в шахматных партиях

<https://github.com/timhutton/chessviz>

Что за данные?



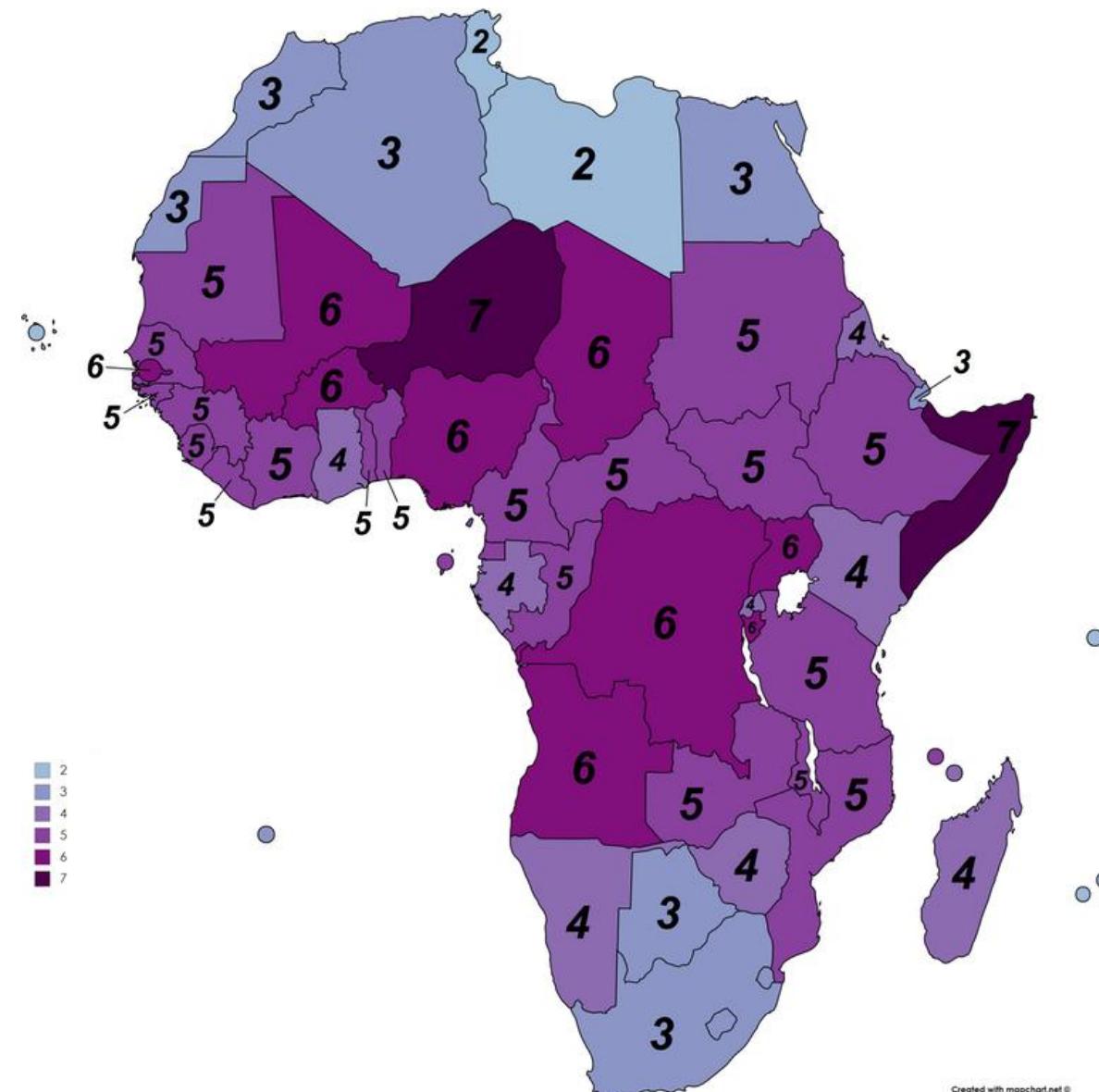




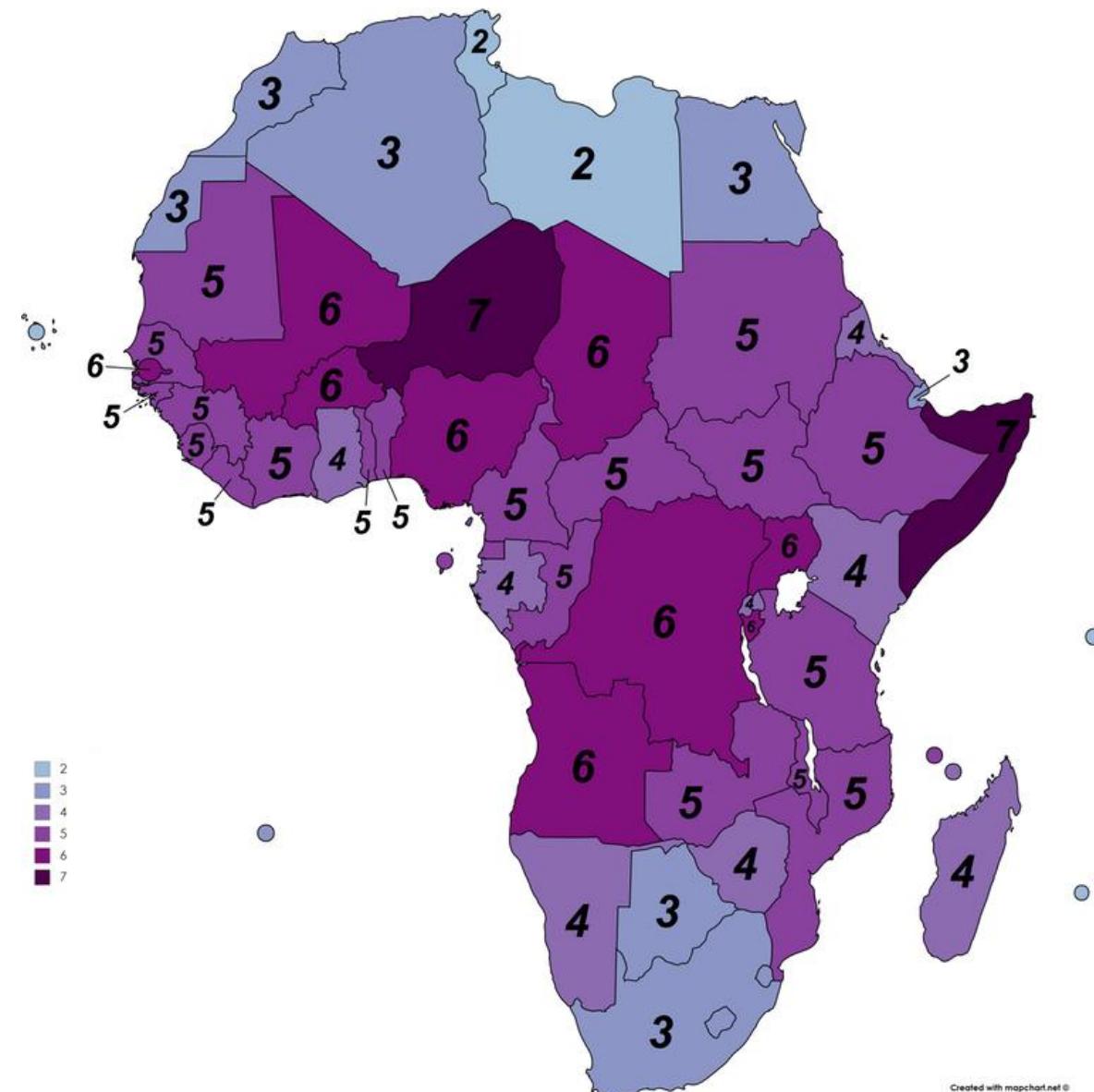
Source: http://ec.europa.eu/public_opinion/archives/ebs/ebs_341_en.pdf

■ "I don't believe there is any sort of spirit, God or life force" ■ "I believe there is some sort of spirit or life force" ■ "I believe there is a God" ■ Doesn't know

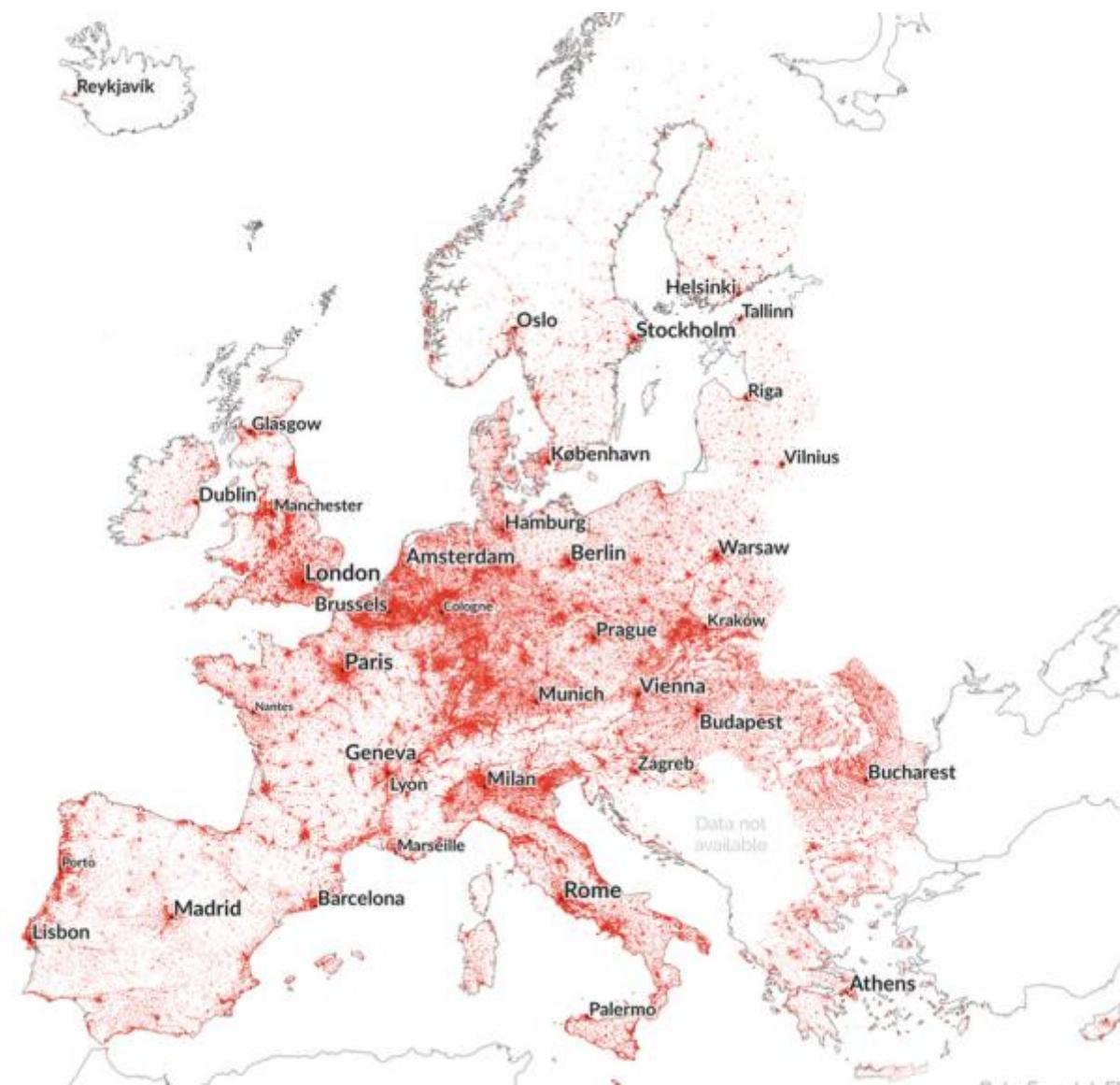
Что за данные?



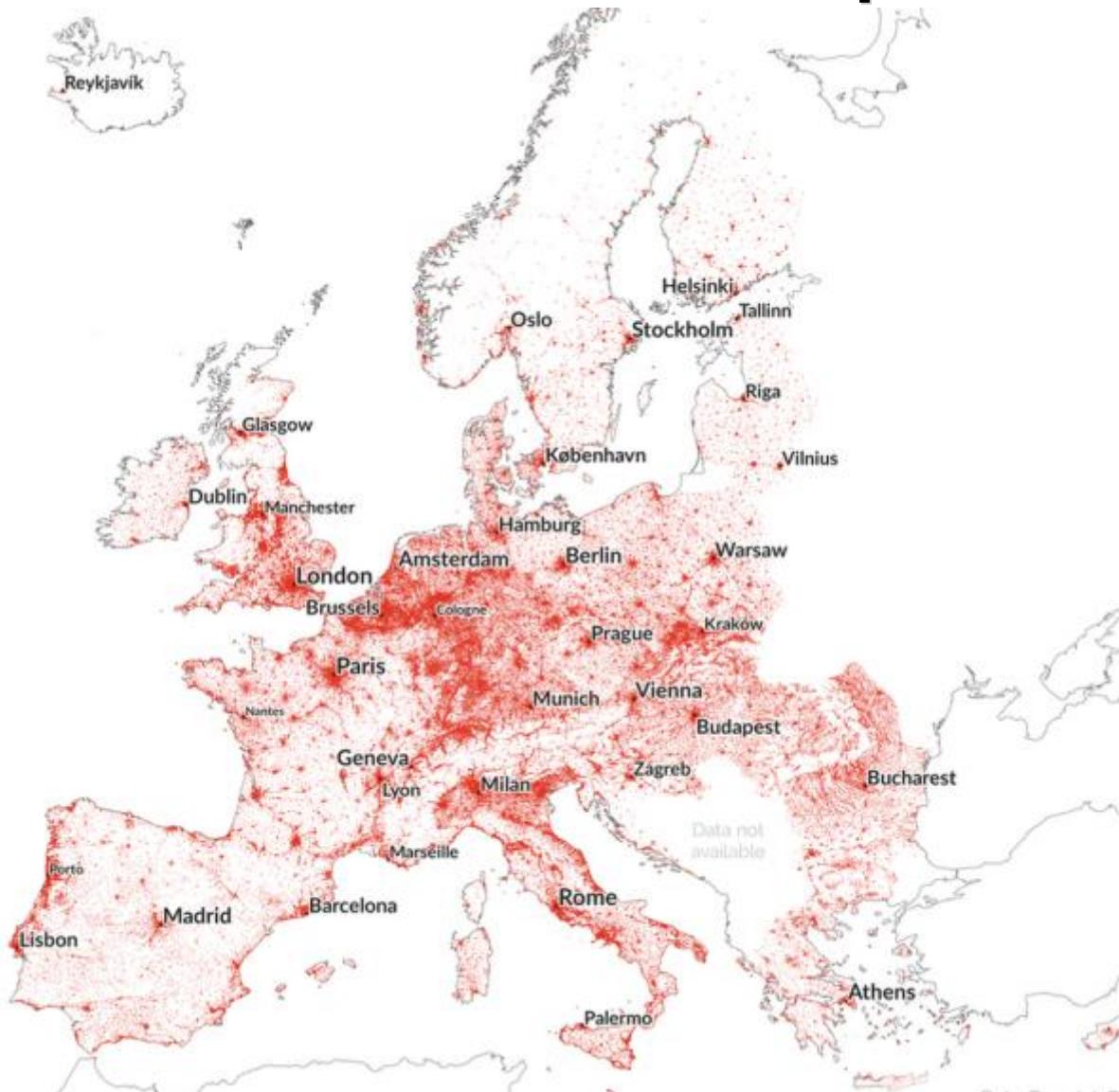
Среднее число детей на одну женщину



Что за данные?

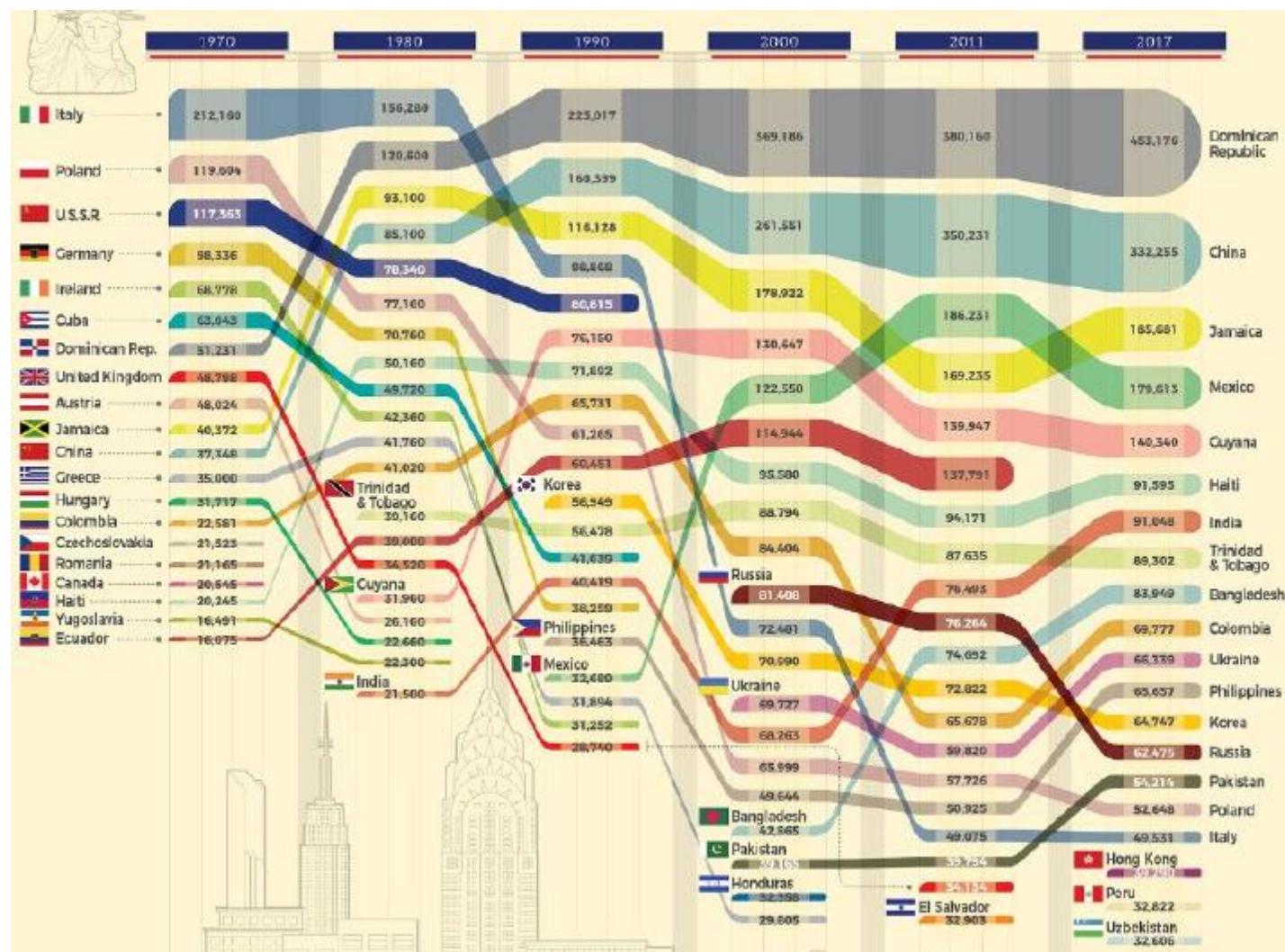


Что за данные? Плотность населения в Европе

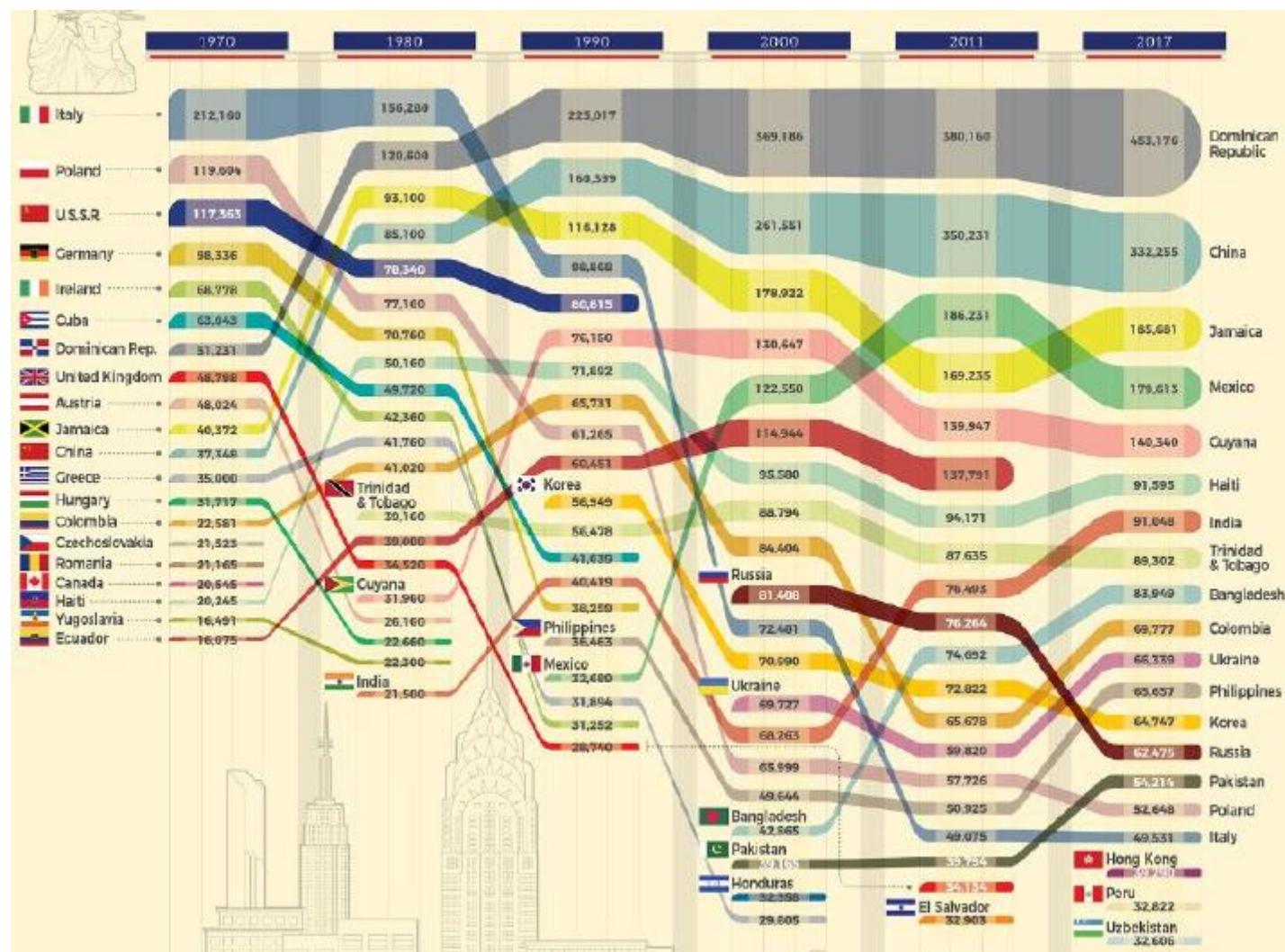


https://www.reddit.com/r/europe/comments/9ektgr/population_density_in_europe/

Что за данные?

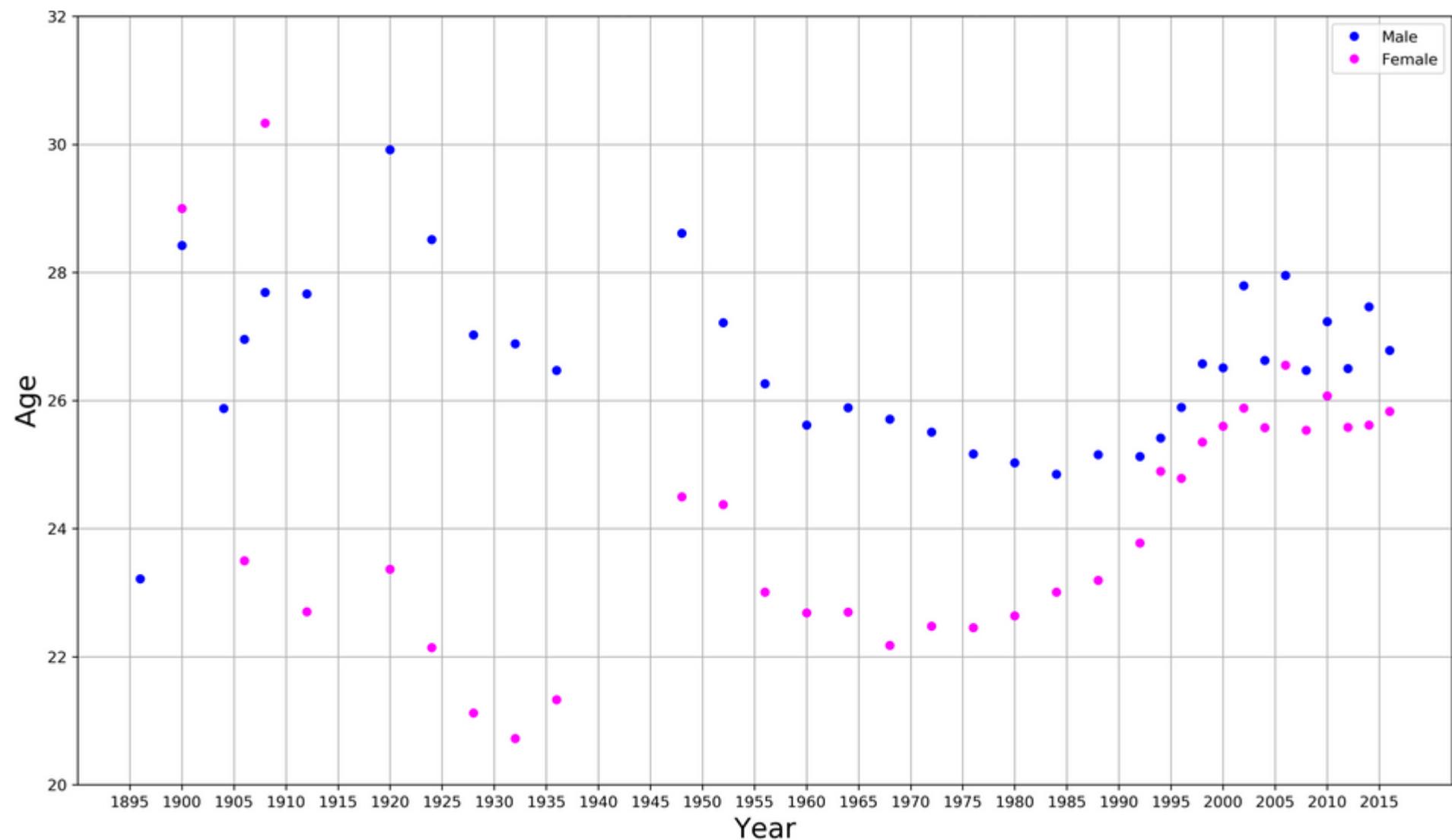


Что за данные?

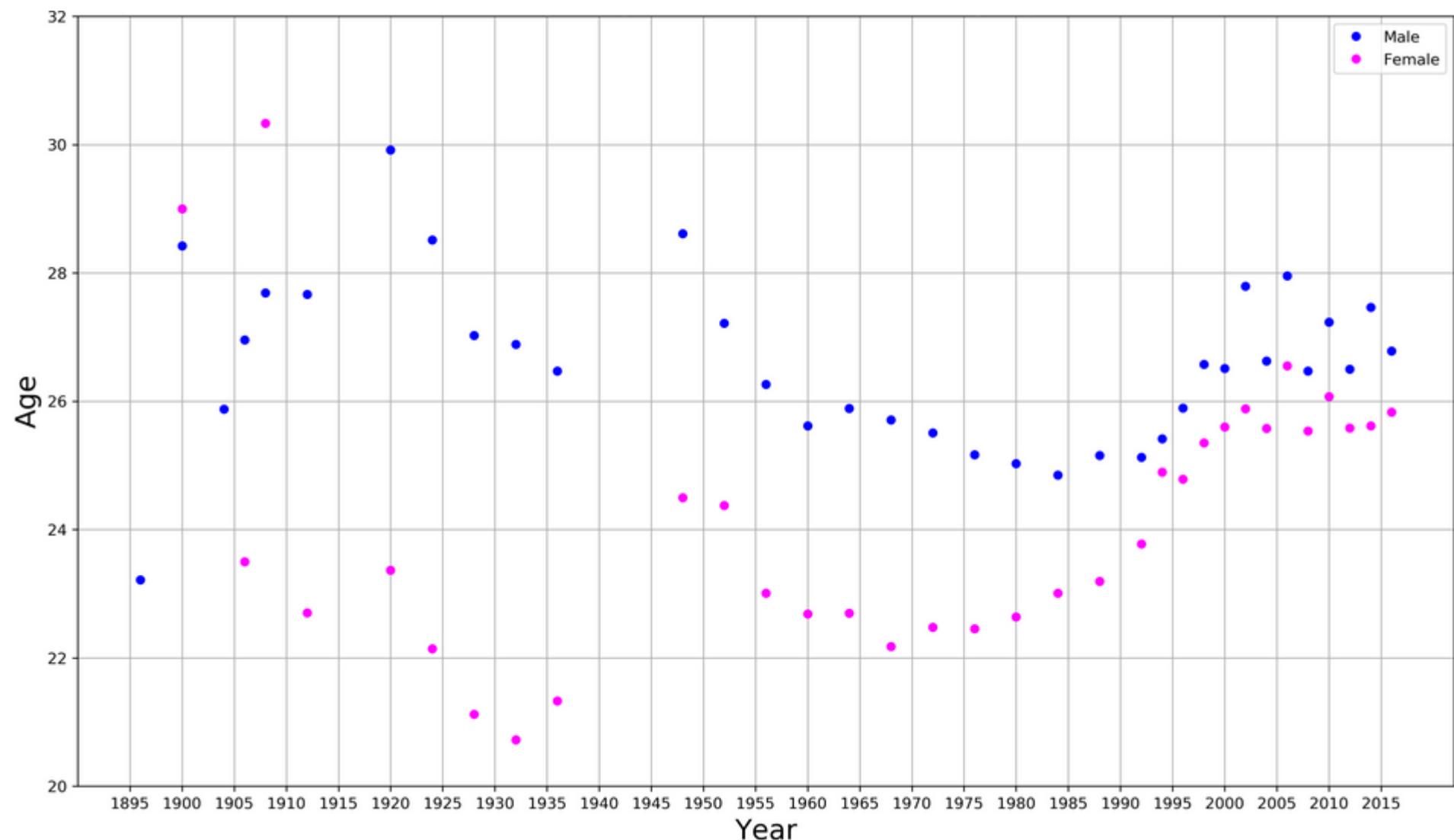


Иммигранты в Нью-Йорке

Что за данные?

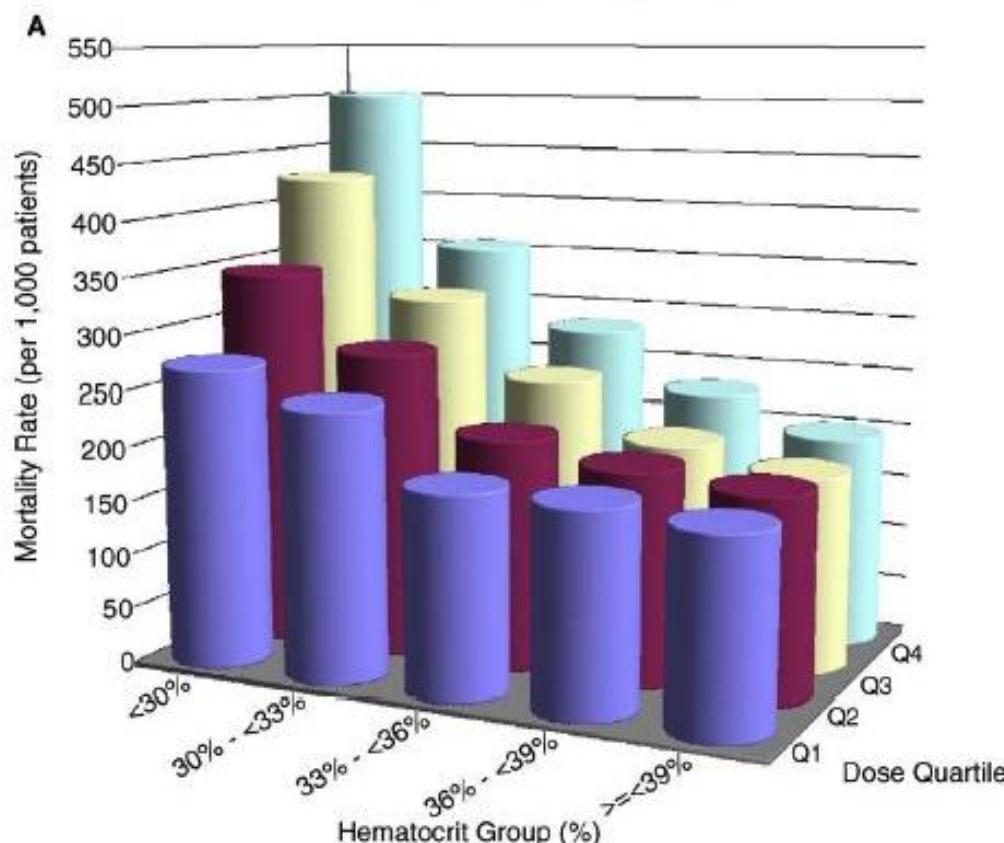


Что за данные?

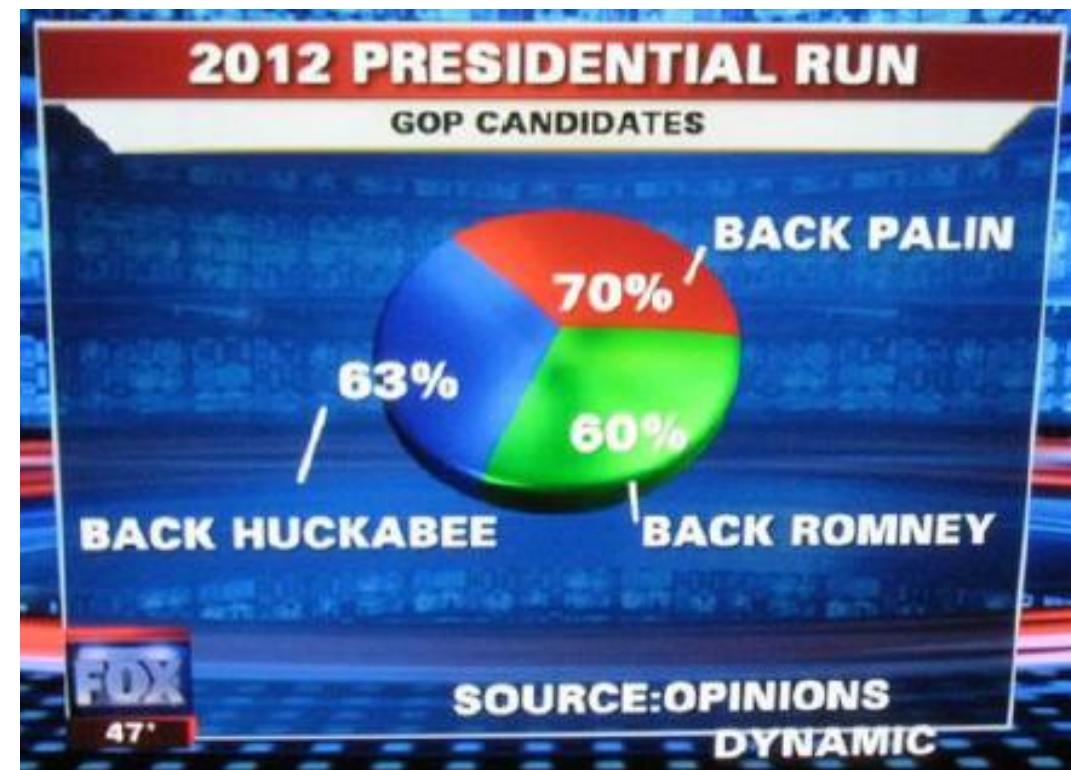


Средний возраст олимпийских медалистов

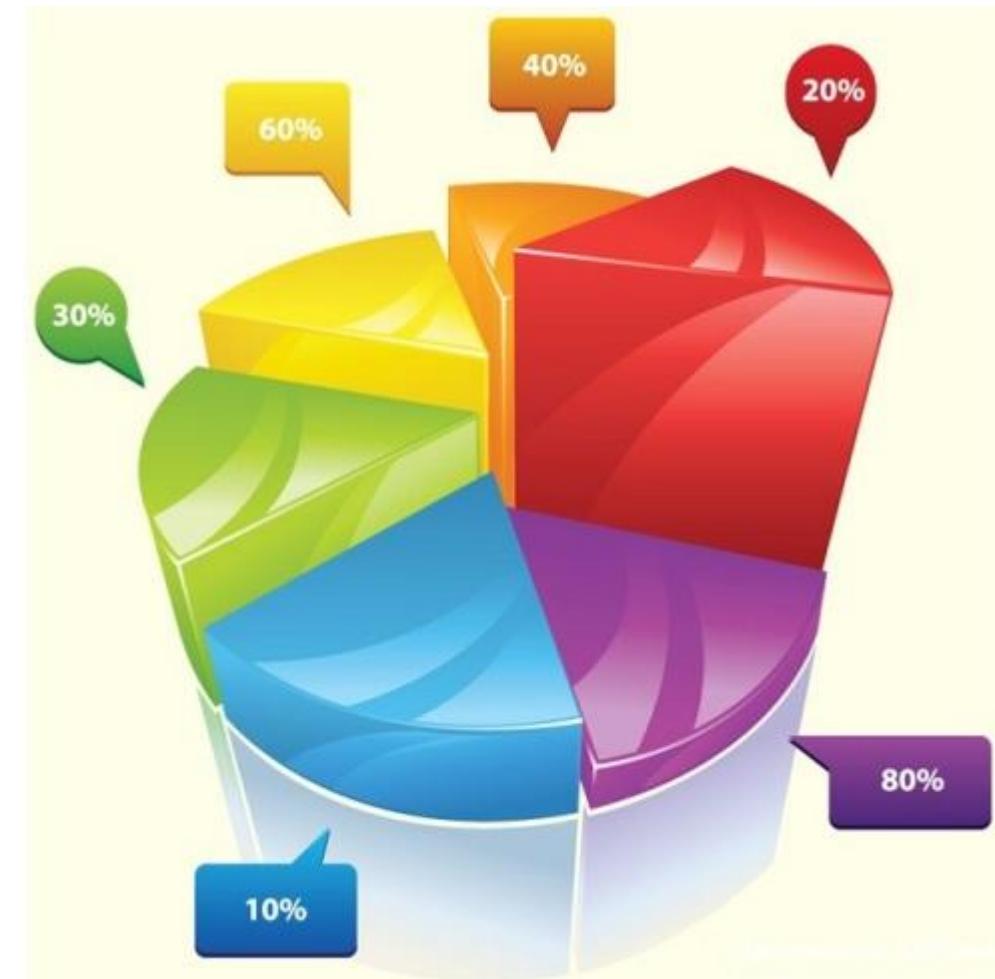
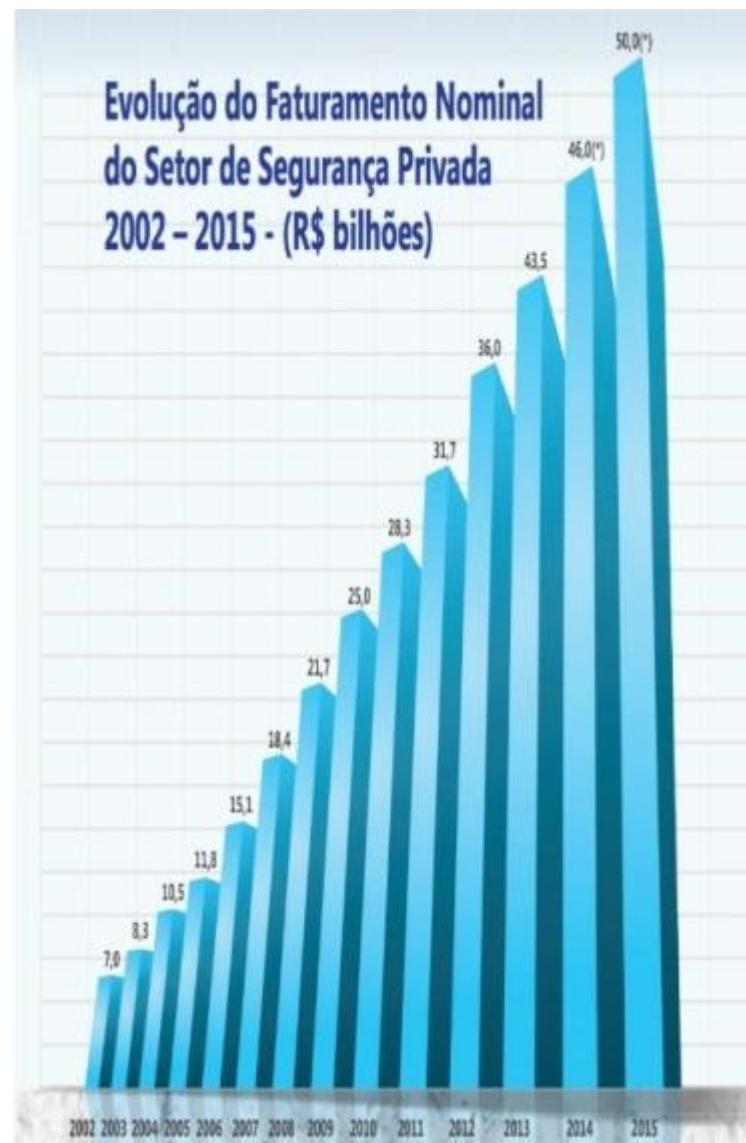
Основные правила – 3D – плохо, нелинейные сравнения – плохо!



Cotter DJ, et al. (2004)



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



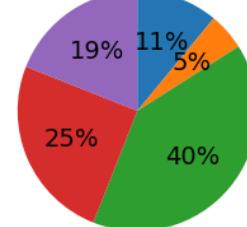
<https://www.reddit.com/comments/9cql3f>

Пример: «до и после»

Survey Results

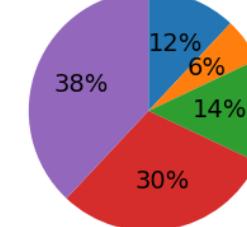
PRE: How do you feel about doing science?

Bored Not great OK Kind of interested Excited



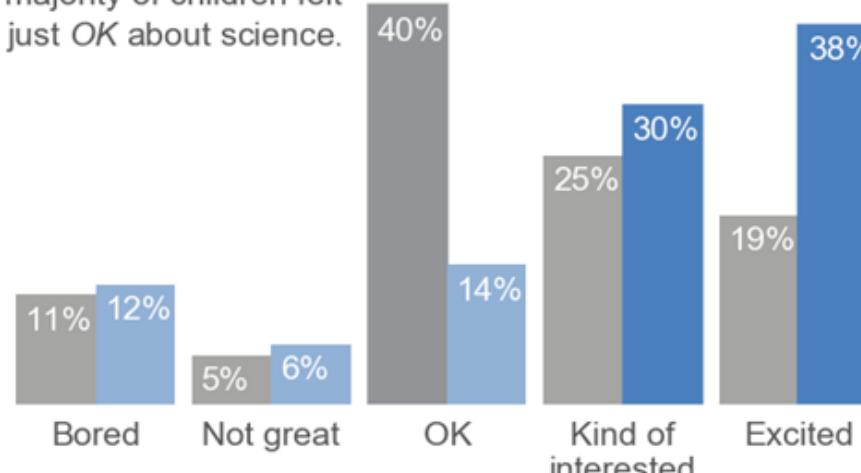
POST: How do you feel about doing science?

Bored Not great OK Kind of interested Excited

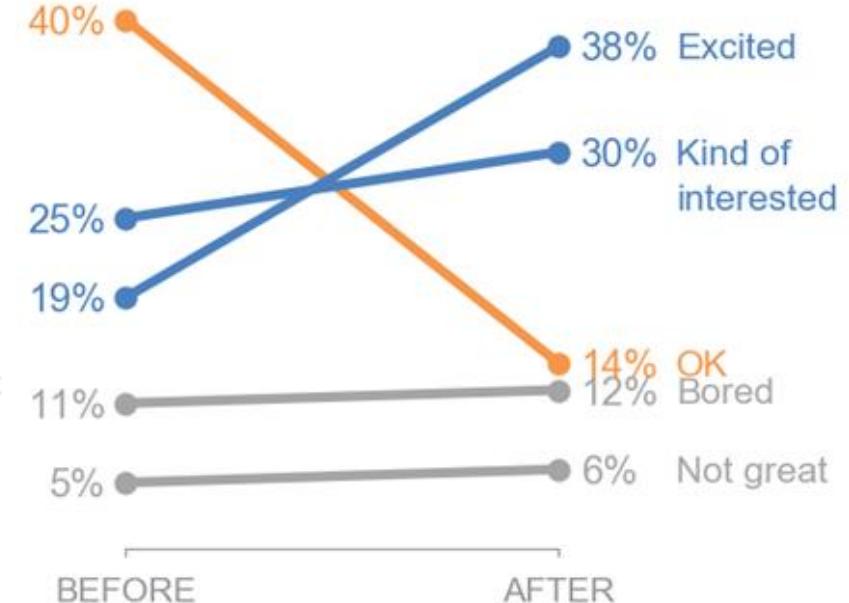


How do you feel about science?

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



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

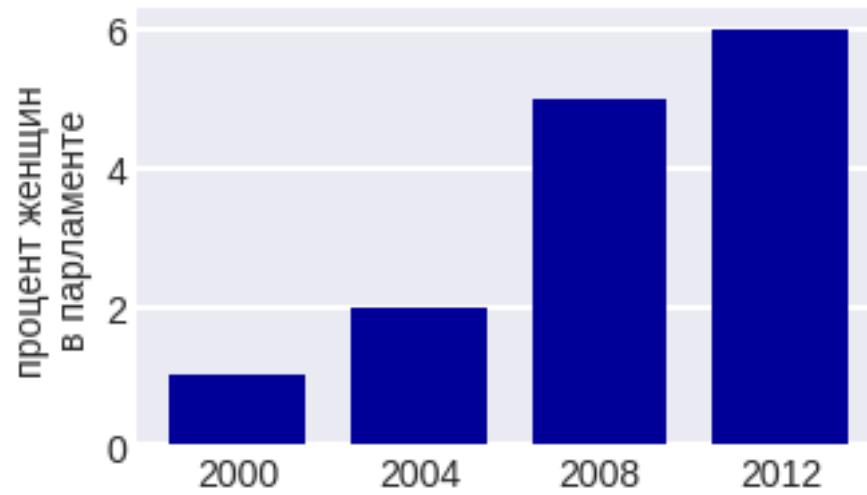


<https://habr.com/company/eastbanctech/blog/422093/>

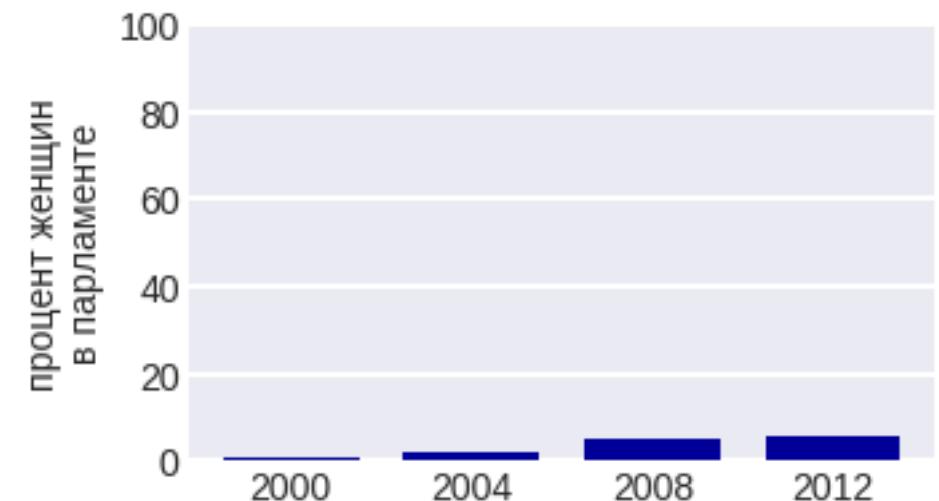
Про рекомендации к визуализации

Процент женщин в парламенте

«неправильно»



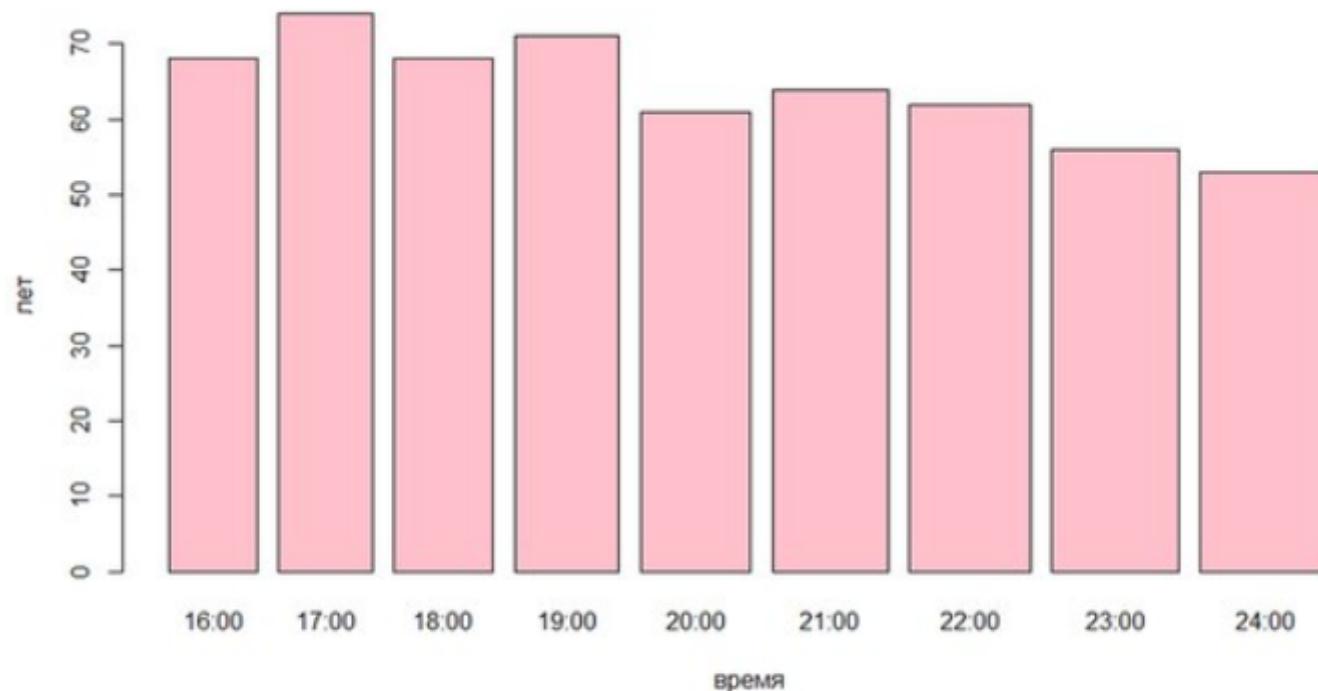
«правильно»



А если это процент убитых в Битцевском парке?

Про рекомендации к визуализации

Средняя продолжительность жизни от времени ухода с рабочего места в пятницу



24 июл в 12:25

Поделиться 🔍 Мне нравится ❤ 8

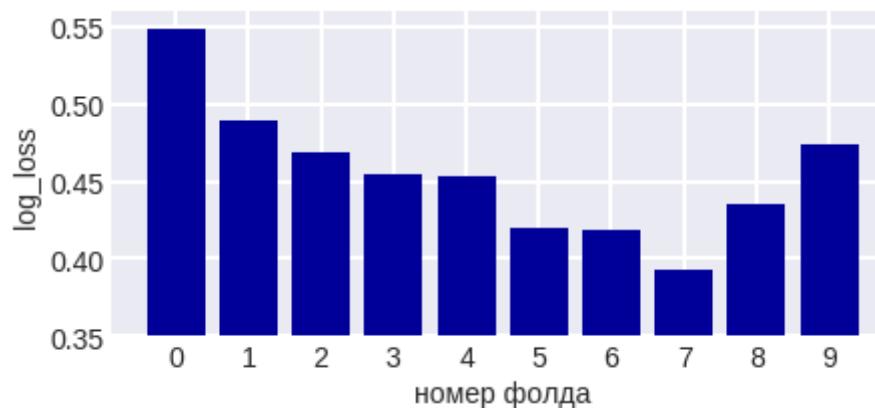


масштаб отвратительный

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.



Правило: 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

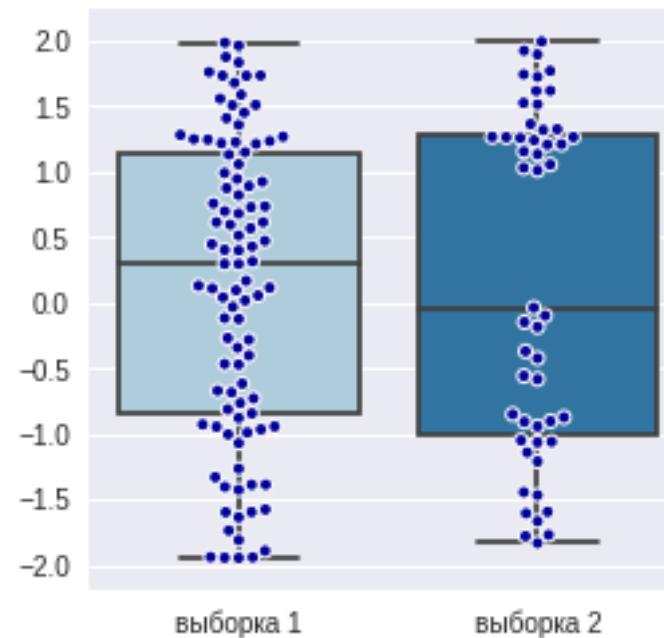
^aBias = ($\hat{\beta} - \beta_0$)/ β_0 .

^bSimulation variance.

^cConfidence interval using jackknife standard error.

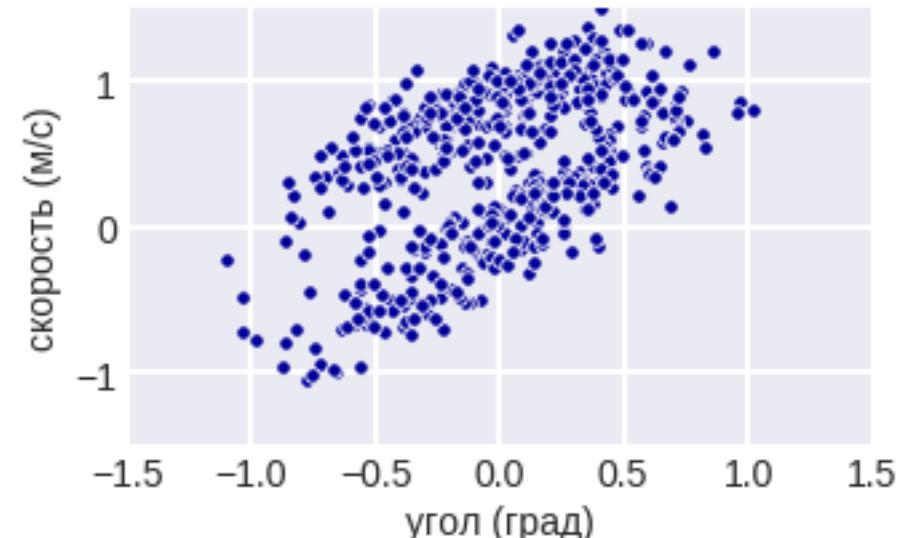
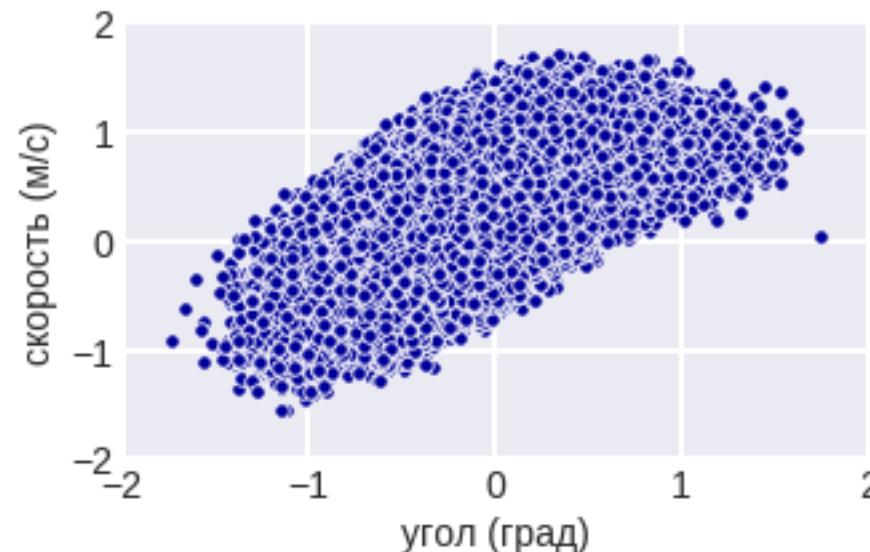
Paik MC (2004)

Правило: используйте разные средства



Однаковые диаграммы, а распределения разные

Правило: используйте разные средства



**Не обязательно смотреть на все данные
(здесь – подвыборка 500 вместо ~200к)**

`df [name] . sample (frac=0.1)`

Правило: иллюстрация не только рисунок

Выбирайте шкалы

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

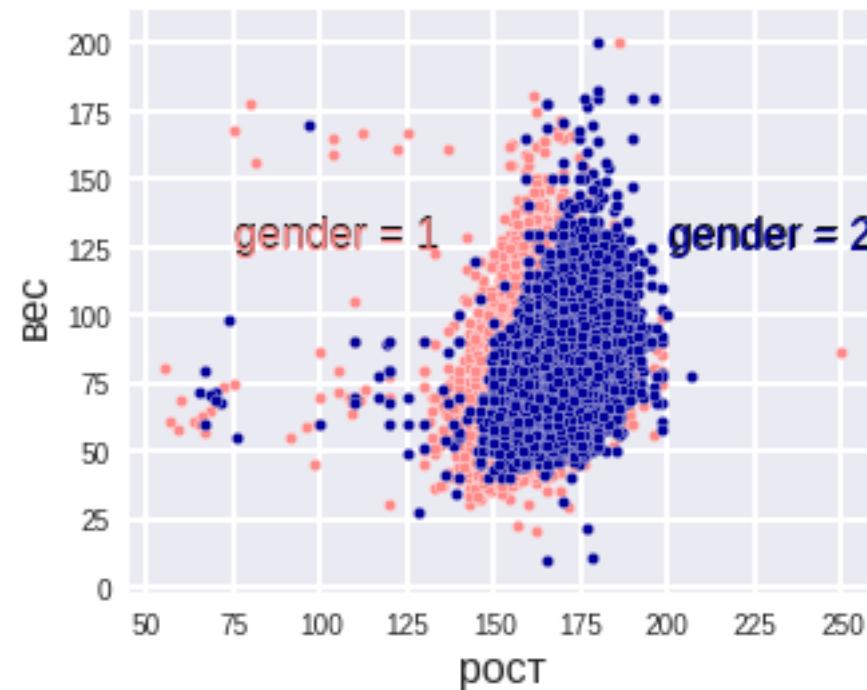
Выбирайте текст

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

Выбирайте цвет и стиль

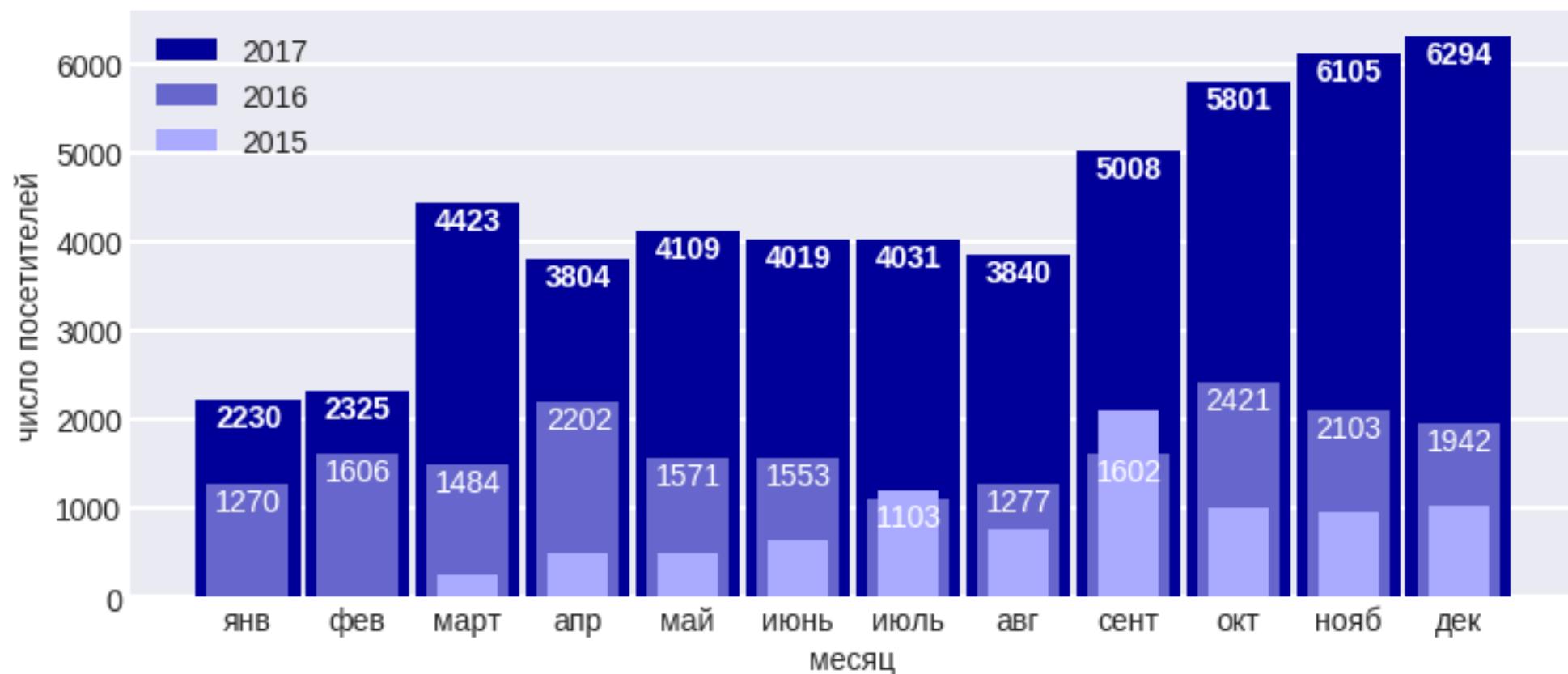
цвет позволит выделить некоторые элементы!
цвета, которые и в градациях серого будут разными!

Пример, когда можно без стандартной легенды



в чём недостаток этой диаграммы рассеивания?

Пример иллюстрации



Когда выполнены все советы...

Домашнее задание

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

ДЗ (необязательное) Найти интересные приёмы для визуализации

ДЗ (большое) Сделать визуализацию данных Kaggle

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

П.С.

Chart Suggestions—A Thought-Starter

