

Основы анализа и визуализации данных для медиа

Лекция 4
9 сентября 2019

Рогович Татьяна
rogovich@gmail.com

Введение в визуализацию данных

Структура лекции

- Посмотрим на одни из первых визуализаций данных.
- Поговорим, зачем мы визуализируем данные.
- Обсудим, что именно называется визуализацией данных и с чем мы не будем ее путать.

Первые визуализации данных

Вспышка холеры в 1854

- Десятки тысяч людей в Англии умерли от холеры между 1831 и 1854 годами
- В то время считалось, что холера передавалась “по воздуху”
- В частных домах не было проточной воды и современной канализации. Люди набирали воду из колонок.

Первые визуализации данных

Вспышка холеры в 1854

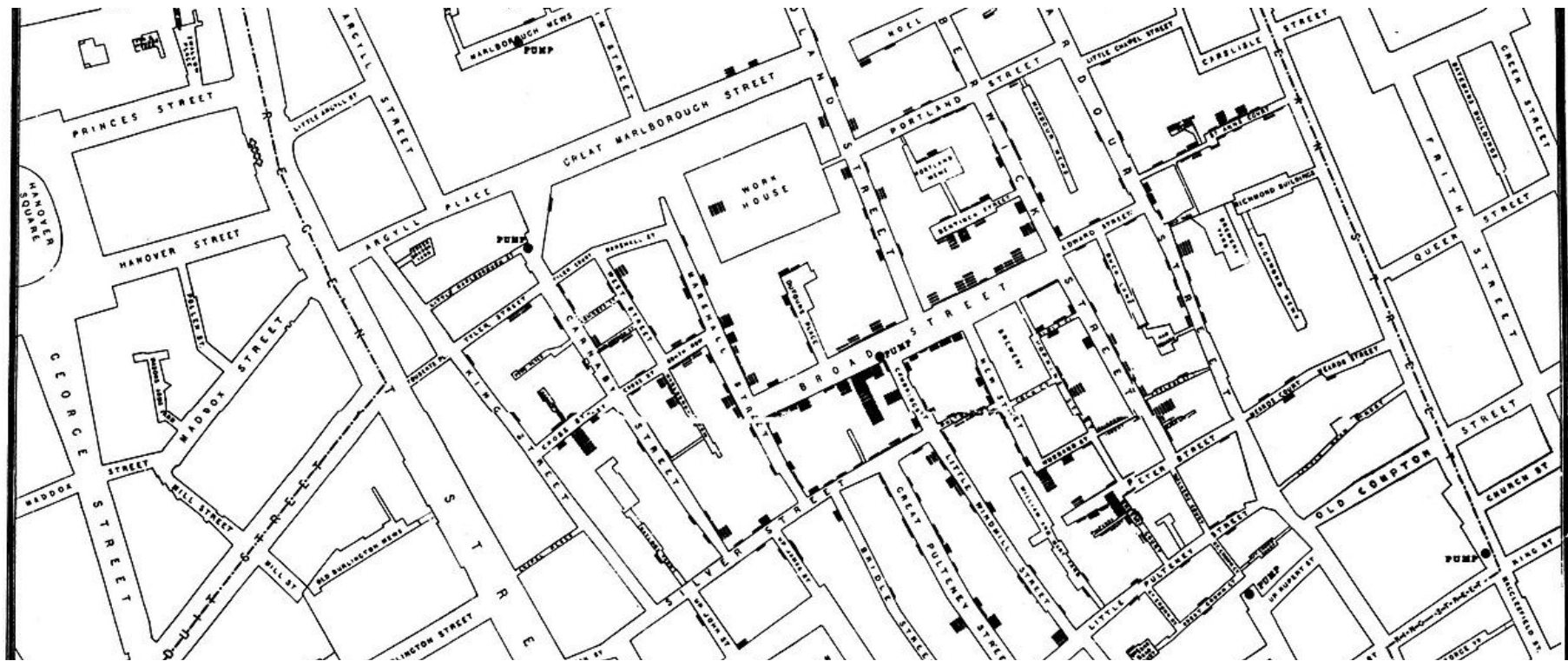
- Ужасная вспышка холеры происходит в Сохо в 1854, там, где живет физик Джон Сноу
- Он собирает данные из госпиталей и открытых источников, о случаях смерти от холеры

Первые визуализации данных

Вспышка холеры в 1854

- Сделал простой график на карте: адреса пострадавших и расположение колонок с водой

<http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html>



<https://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg>



<https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map>

Первые визуализации данных

Вспышка холеры в 1854

- Сноу смог обнаружить зараженные источники воды
- Нашел причину многих случаев в напитке, который продавал один из баров, рядом с зараженной колонкой
- Стал родоначальником эпидемиологии
- Profit

Дело Игнаца Земмельвейса

Иногда отсутствие визуализации
делает данные менее
убедительными и позволяет
пропустить важный тренд

- В 1846 году Земмельвейс работал в университетском госпитале Вены с двумя родильными отделениями, в одном из которых работали врачи, а в другом - акушерки.
- В то время, многие матери умирали после родов из-за загадочных причин.
- Земмельвейс собрал данные и увидел странный тренд, что в клинике, где работали врачи, смертность была выше, чем в клинике, где работали акушерки (9,9 % против 3,9 %).

Данные Игнаца Земмельвейса

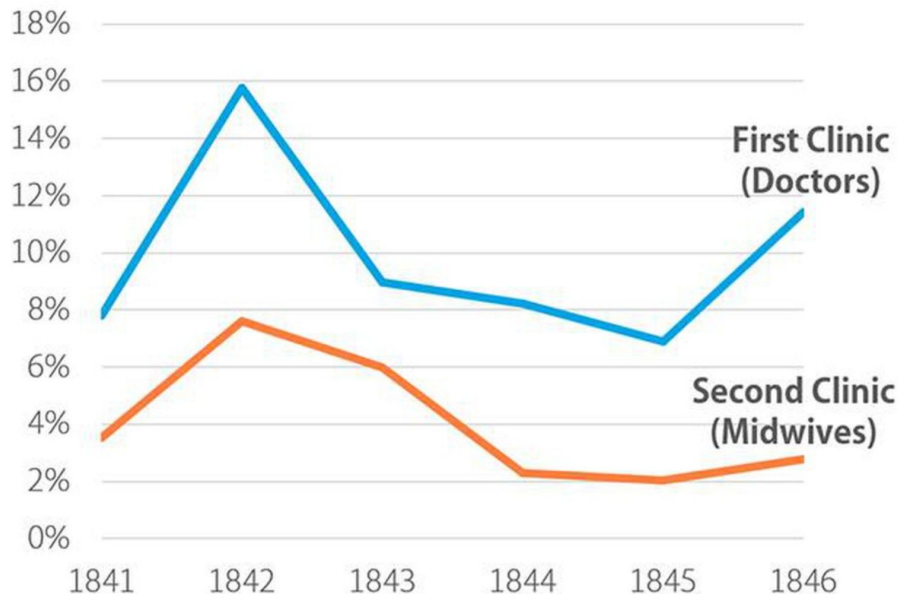
year	births	deaths	clinic
1841	3036	237	clinic 1
1842	3287	518	clinic 1
1843	3060	274	clinic 1
1844	3157	260	clinic 1
1845	3492	241	clinic 1
1846	4010	459	clinic 1
1841	2442	86	clinic 2
1842	2659	202	clinic 2
1843	2739	164	clinic 2
1844	2956	68	clinic 2
1845	3241	66	clinic 2
1846	3754	105	clinic 2

Дело Игнаца Земмельвейса

- Один из врачей клиники умирает от похожих симптомов, что и молодые матери, после того как порезался во время аутопсии
- Акушерки не проводили вскрытий.
- Земмельвейс связал эти события, и предложил ввести в клиниках обязательное мытье рук (1846 год!)
- Немедленное улучшение, но инновация была высмеяна и не прижилась в медицинских кругах (теория получила признание сильно позже).

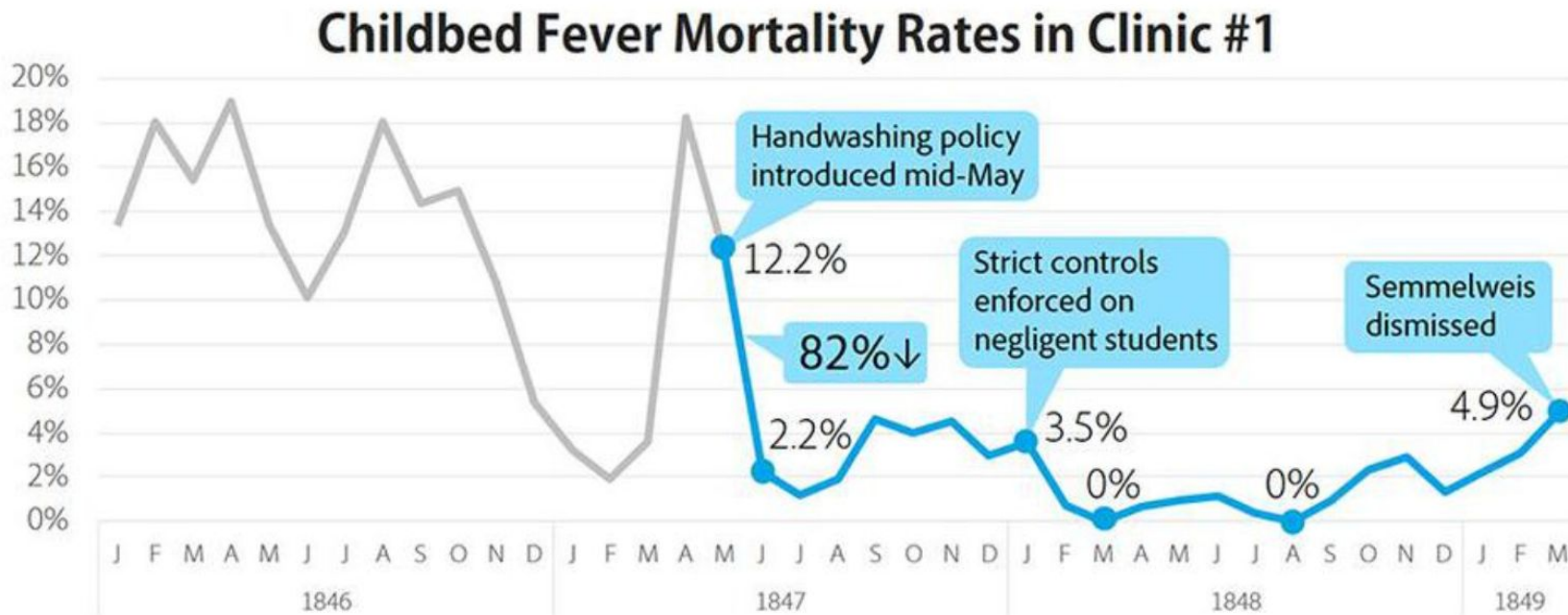
Чего не сделал Земмельвейс?

Childbed Fever Mortality Rates

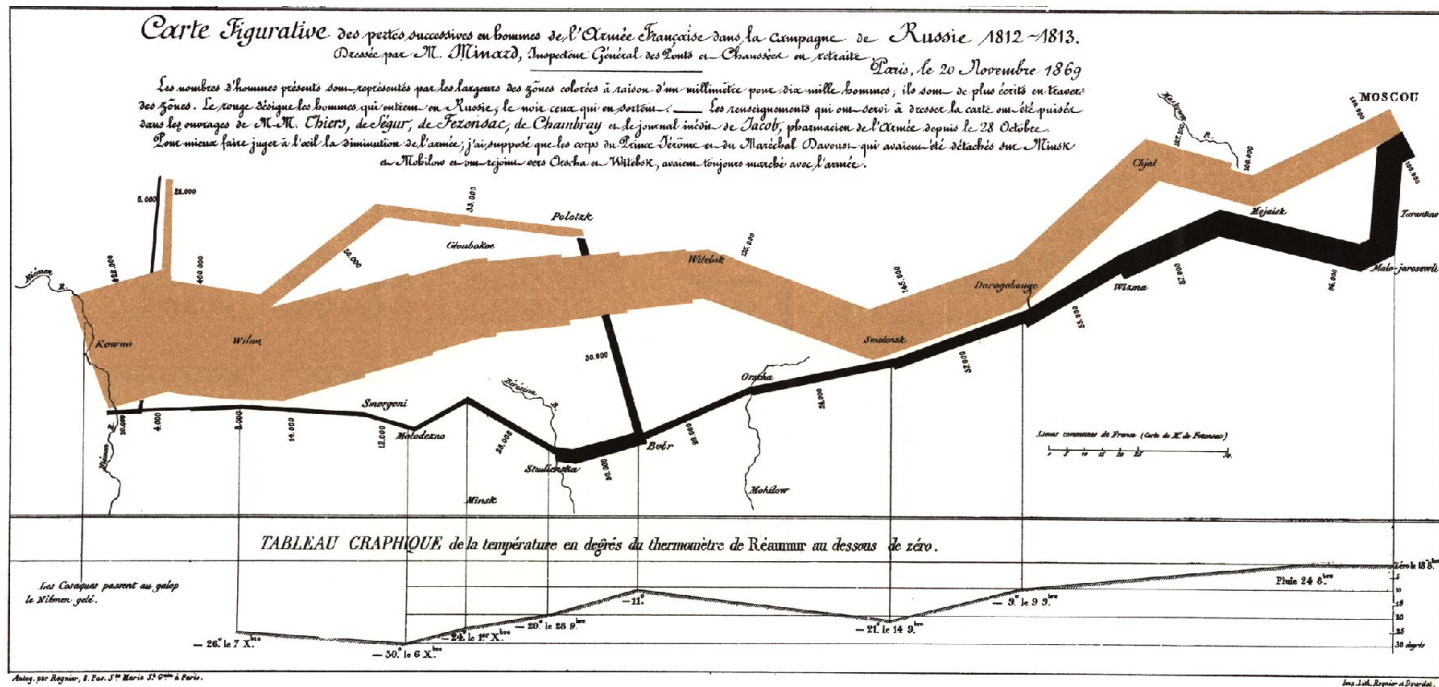


- Пример Земмельвейса часто приводят на фоне истории про Джона Сноу, потому что они произошли примерно в одно и то же время
- Обе теории врачебное сообщество не брало всерьез
- Но Джон Сноу смог визуализировать и убедительно презентовать данные, когда Земмельвейс - нет.

Чего не сделал Земмельвейс?



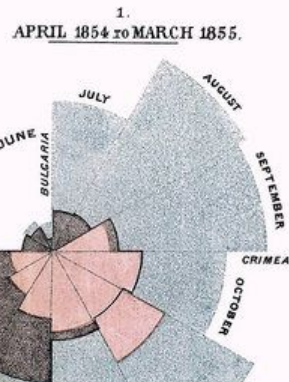
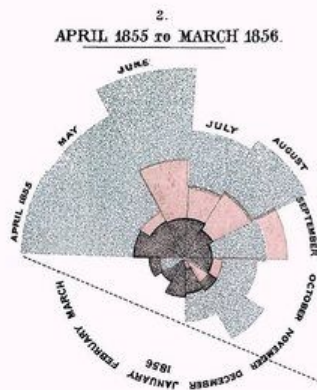
Возможно, лучшая визуализация*: Joseph Minard 1861



*По версии Тафти

Флоренс Найтингел

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.



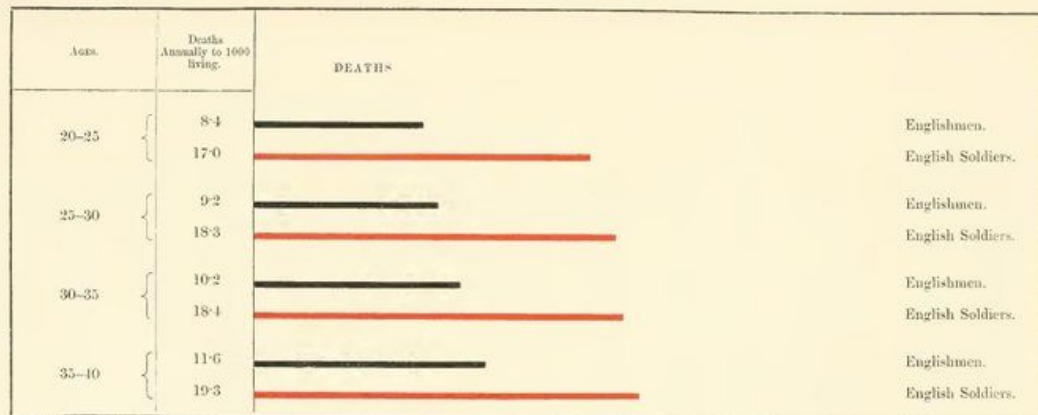
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov^r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red, in January & February 1855, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Флоренс Найтингел

(B.)



Representing the Relative Mortality of the Army at Home and of the English Male Population at corresponding Ages.



JAMES LEWIS, del.

NOTE.—The Mortality of the English Male Population, at the above ages, is taken from English Life Table (1849-53).

<https://www.atlasobscura.com/article/s/florence-nightingale-infographic>

Почему так важна визуализация статистических данных?

Квартет Энскомба (1973)

Четыре набора числовых данных, у которых простые статистические свойства идентичны, но их графики существенно отличаются. Каждый набор состоит из 11 пар чисел. Квартет был составлен математиком Ф. Дж. Энскомбом для иллюстрации важности применения графиков для статистического анализа, и влияния выбросов значений на свойства всего набора данных.

“A computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.”

– **Francis Anscombe, 1973**

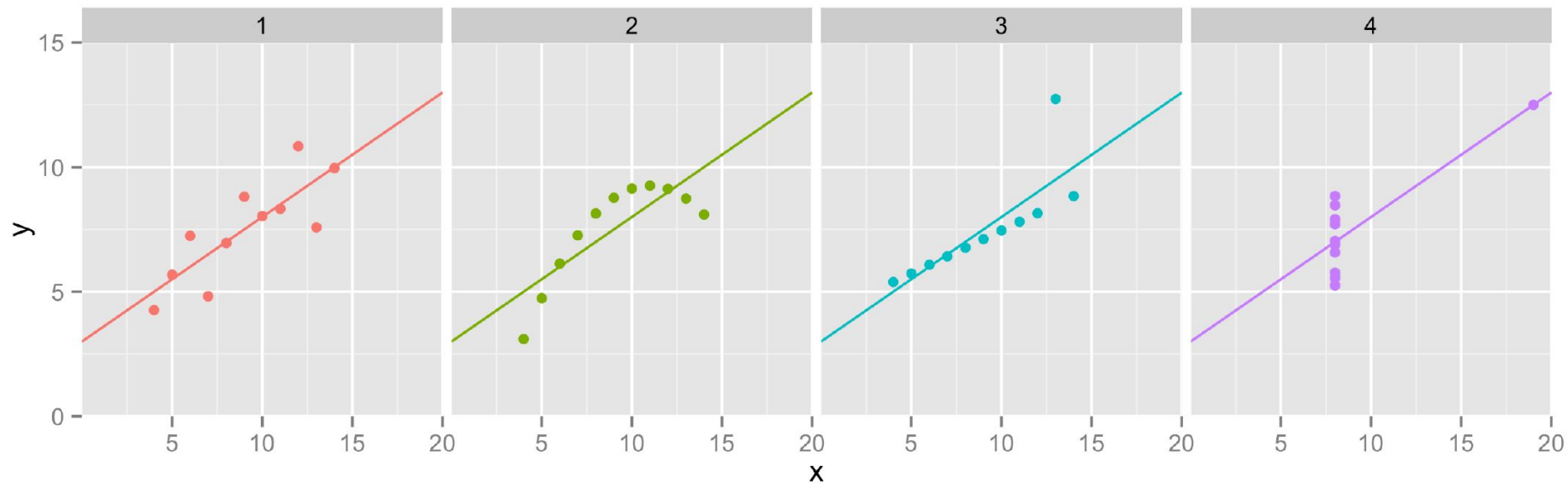
Квартет Энскомба: данные

Group 1		Group 2		Group 3		Group 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.56
9.0	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

Квартет Энскомба: описательные статистики

	Group 1	Group 2	Group 3	Group 4
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

Квартет Энскомба: визуализация



Зачастую,
визуализация –
единственный
способ увидеть,
что происходит
в данных

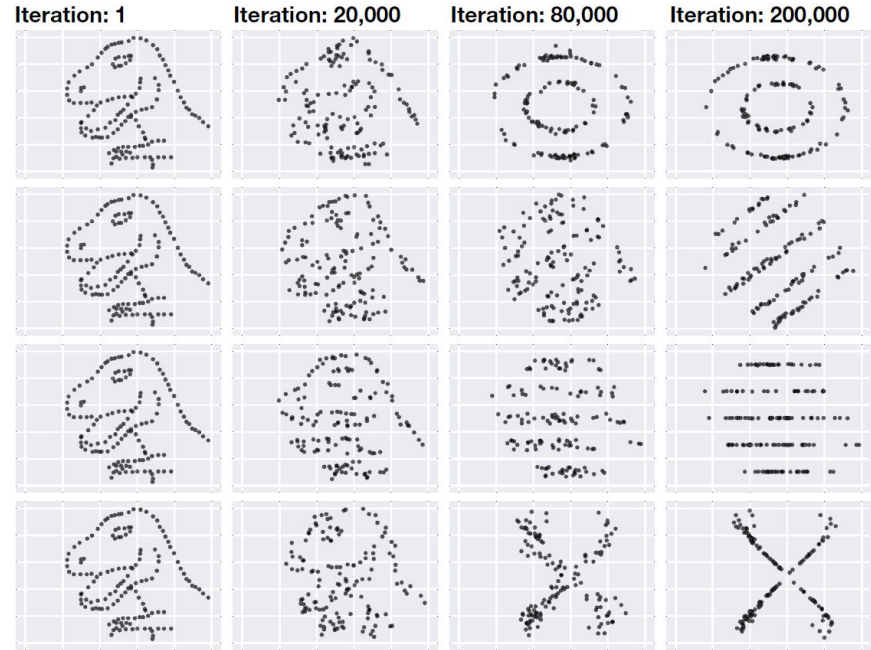


Figure 6. Creating a collection of datasets based on the “dinosaur” dataset. Each dataset has the same summary statistics to two decimal places: ($\bar{x}=54.26$, $y = 47.83$, $sd_x = 16.76$, $sd_y = 26.93$, *Pearson’s* $r = -0.06$).

<https://www.autodeskresearch.com/publications/samestats>

Зачем визуализировать данные?

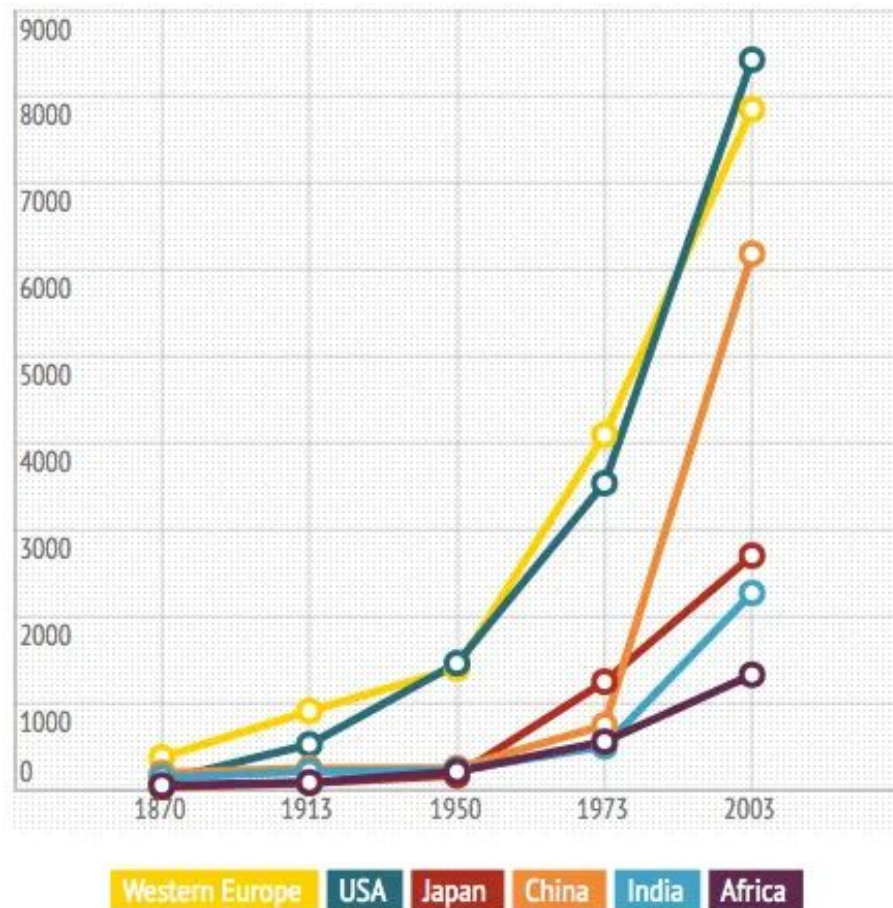
1. Замечать вещи, которые бы иначе не заметили.
2. Исследовать данные.
3. Находить ответы быстрее.

Зачем
визуализировать
данные?

	A	B	C	D	E	F
1	Past GDP	1870	1913	1950	1973	2003
2	Western Europe	367	902	1396	4096	7857
3	USA	98	517	1455	3536	8430
4	Japan	25	71	160	1242	2699
5	China	189	241	244	739	6187
6	India	134	204	222	494	2267
7	Africa	45	79	203	549	1322

Зачем визуализировать данные?

<http://www.mulinblog.com/data-visualization-matters/>

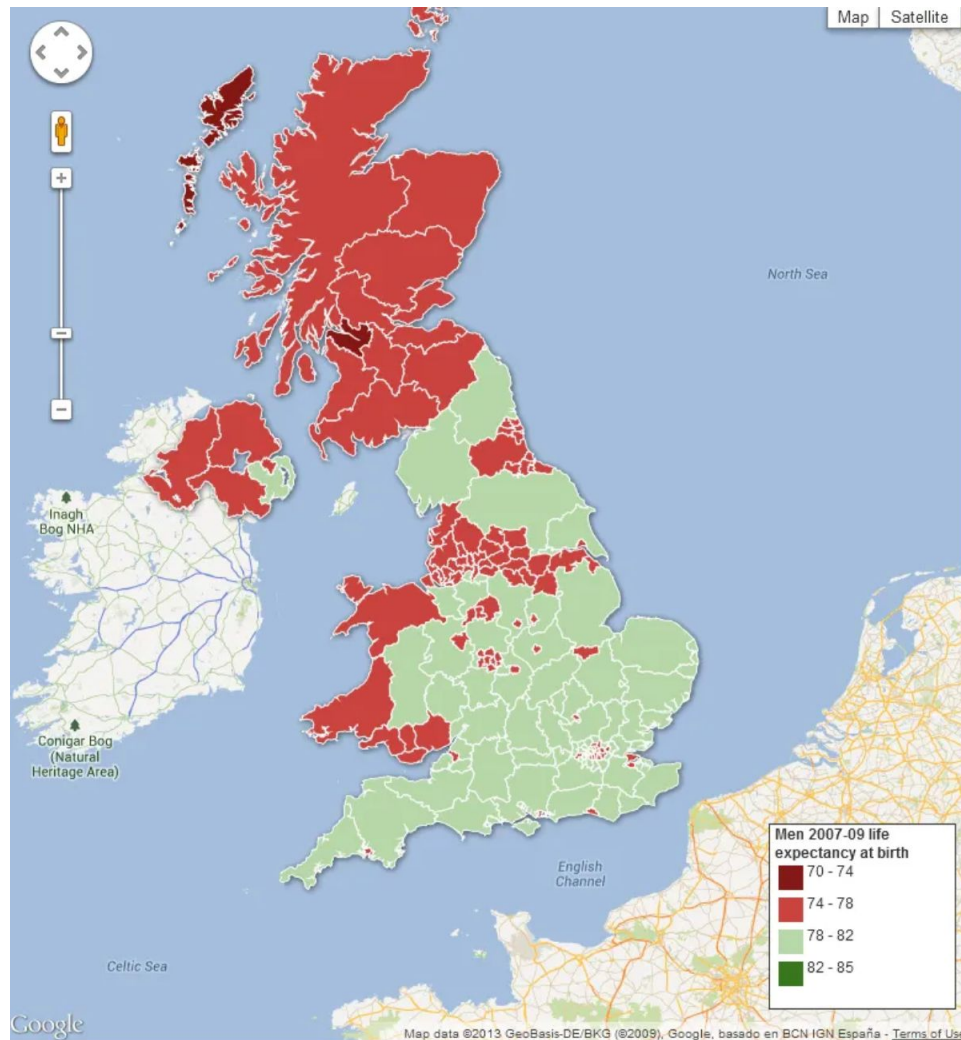


Зачем визуализировать данные?

area	score	std	score	std
Highland	81.0	1.5	76.4	1.9
Hillingdon	83.4	2.0	78.6	2.1
Hounslow	82.1	2.3	77.8	2.5
Hull Teaching	80.0	0.6	75.2	1.5
Hywel Dda	82.1	1.7	77.9	2.2
Isle of Wight National Health Service	83.2	1.5	79.1	2.1
Islington	81.2	1.7	75.4	2.1
Kensington and Chelsea	89.0	4.1	84.4	4.8
Kingston	83.7	2.4	80.7	3.0
Kirdees	80.9	1.6	76.7	1.7
Knowsley	79.8	1.8	75.9	2.5
Lambeth	81.1	1.3	76.4	2.9
Lanarkshire	79.2	1.2	74.4	1.4
Leeds	82.0	1.3	77.7	2.2
Leicester City	80.0	1.0	75.4	1.2
Leicestershire County and Rutland	83.4	1.7	79.7	1.7

Зачем
визуализировать
данные?

<http://www.mulinblog.com/data-visualization-matters/>

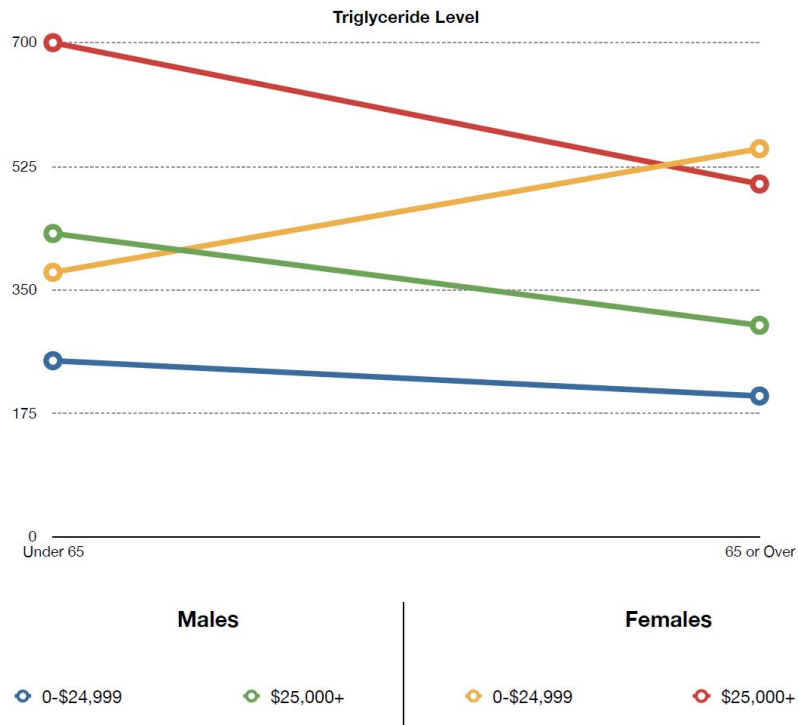


Зачем визуализировать данные?

Основываясь на таблице ниже, скажите, как пол и уровень дохода влияют на уровень холестерина

	Males		Females	
Income Group	Under 65	65 or Over	Under 65	65 or Over
0-\$24,999	250	200	375	550
\$25,000+	430	300	700	500

Зачем визуализировать данные?



Stephen Kosslyn “Clear and to the Point”

Что люди
могут иметь
в виду, когда
говорят
“визуализация
данных”?

- Data Visualization
- Scientific Visualization
- Information Visualization
- Statistical Graphics
- Visual Analytics
- Information Dashboards
- Infographics
- Informative Art

А в чем разница?

Различия обычно сводятся к ответам на вопросы:

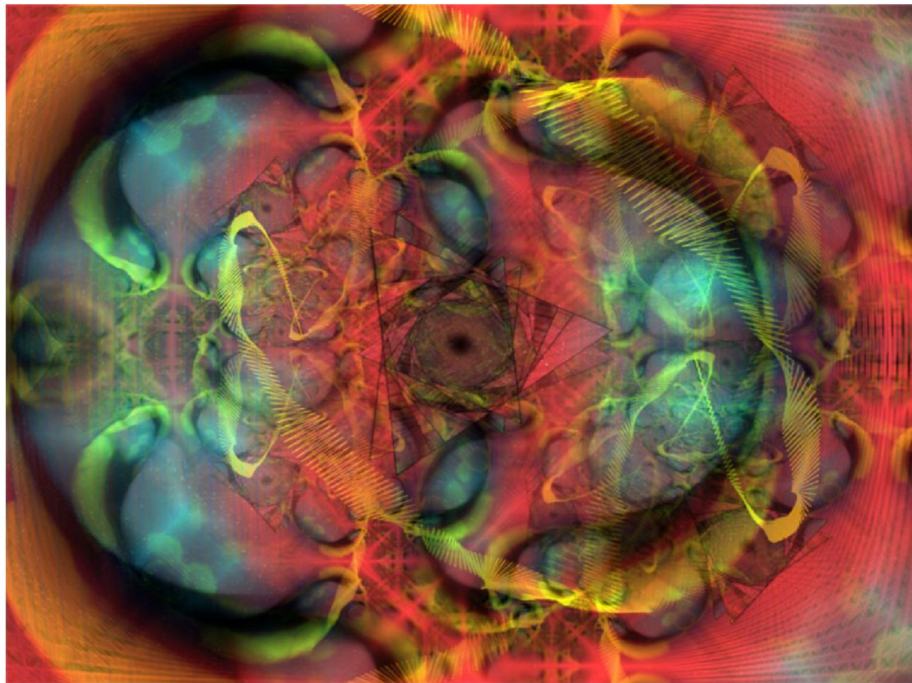
- Какие данные визуализируются?
- Как и с какой целью эти данные визуализируются?

Что мы
называем
визуализацией
данных в
рамках нашего
курса?

Визуализация данных

- Представляет данные визуально
- Результаты должны быть легко читаемы
- “Сырые” данные трансформируются в информацию

Визуализация данных?



- ✓ Представляет данные визуально
- ✗ Результаты должны быть легко читаемы
- ✓ “Сырые” данные трансформируются в информацию

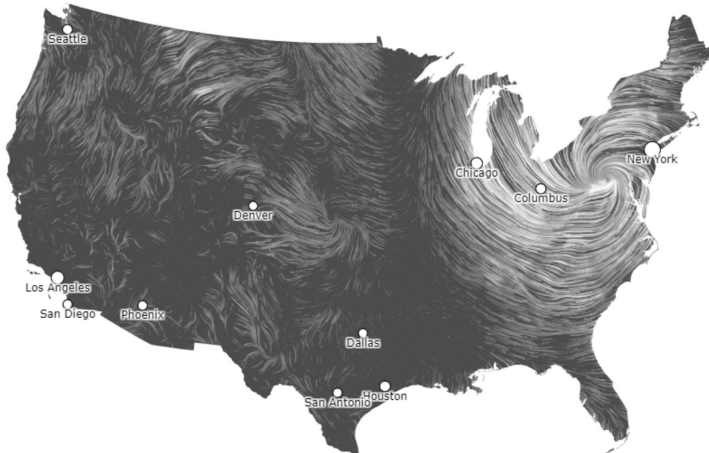
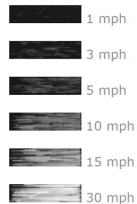
<https://en.wikipedia.org/wiki/MilkDrop>

Визуализация данных?

October 30, 2012

6:59 am EST
(time of forecast download)

top speed: 39.7 mph
average: 8.4 mph



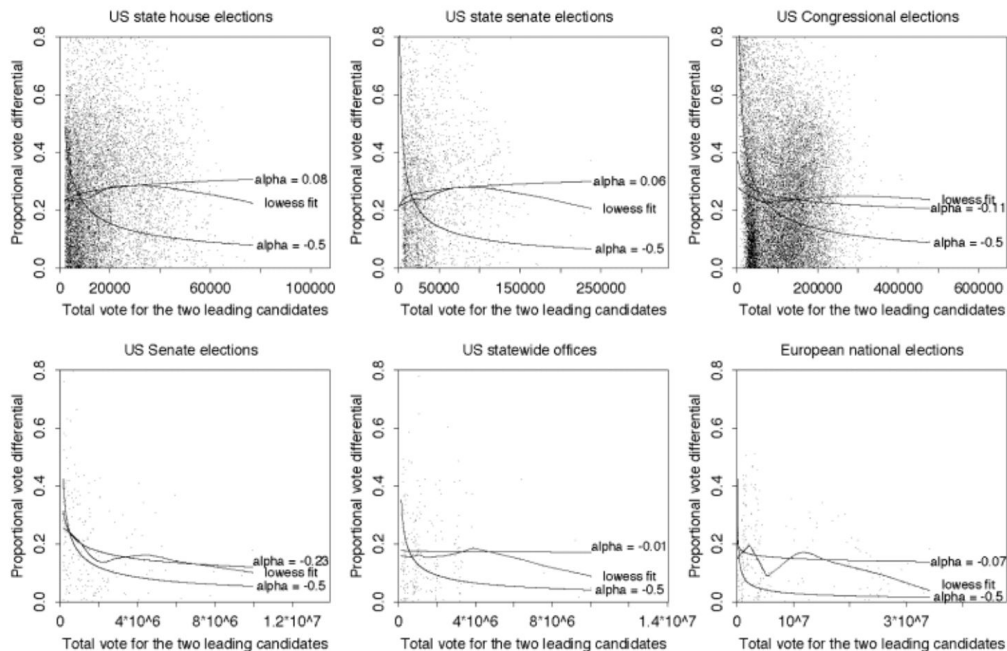
- ✓ Представляет данные визуально
- ✓ Результаты должны быть легко читаемы
- ✓ “Сырые” данные трансформируются в информацию

<http://hint.fm/wind/gallery/oct-30.js.html>

Статистические графики

- Иллюстрация абстрактных статистических вычислений
- На первом месте - точность представления информации.
- Нет интерактивности
- Утилитарны

Статические графики

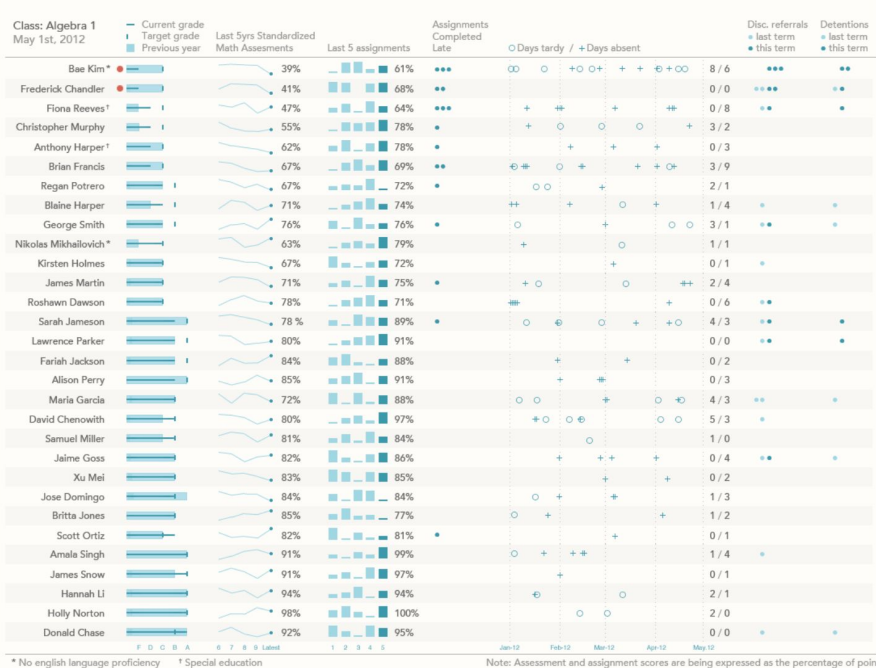


http://andrewgelman.com/2007/10/09/why_the_sq_uarer/

Визуальная аналитика

- Абстрактные данные.
- Отвечает на конкретные вопросы (запрос от бизнеса)
- Подкрепляет процесс принятия решений
- Аудитория: эксперты

Визуальная аналитика



- Оповещения о выбросах
- Тренды
- Перенасыщены информацией

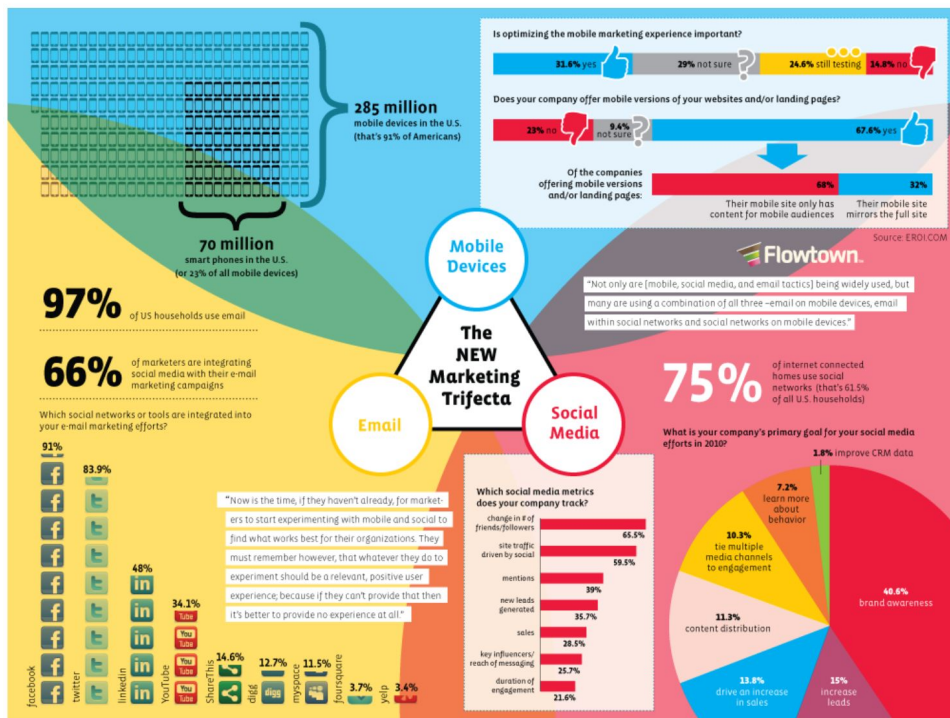
<http://www.perceptualedge.com/blog/?p=1374>

Инфографика

- Цель - быстро донести сложную информацию
- Привлекать внимание
- Использование графических элементов

Периодически употребляется негативно в экспертном сообществе из-за того, что очень много плохих примеров

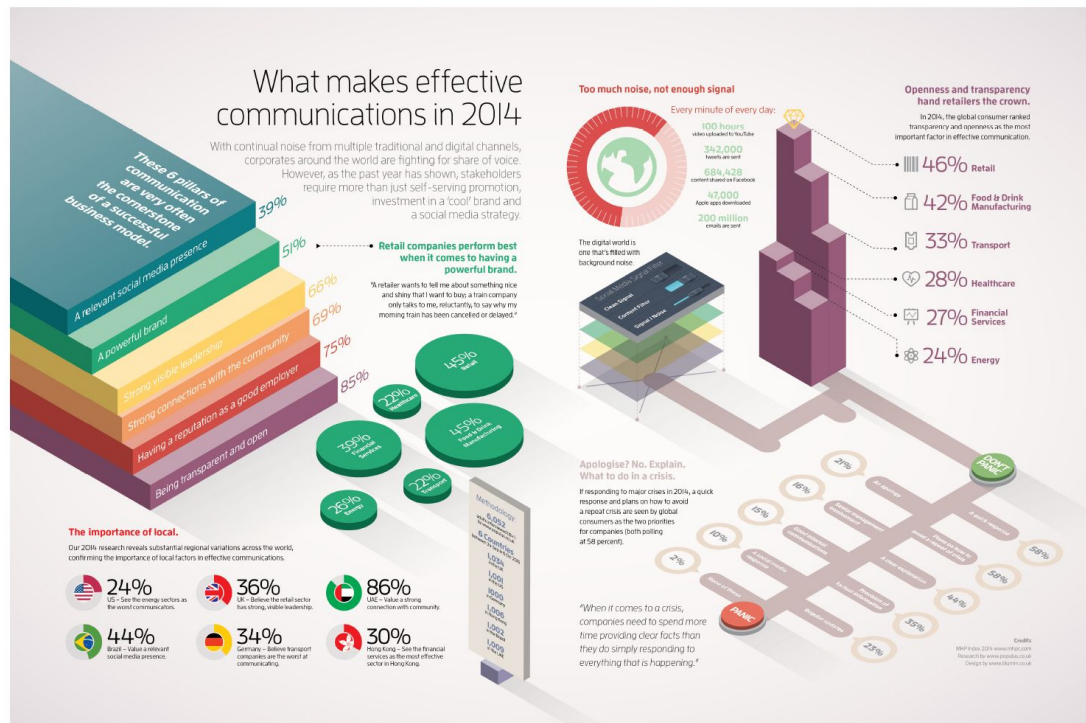
Инфографика: плохо



- Сложно считывать информацию
- Слишком много цветов и элементов
- Нарушены пропорции
- Дизайн ради дизайна

<https://visual.ly/community/infographic/business/new-world-marketing>

Инфографика: плохо



Инфографика: хорошо



- Не нарушаются пропорции визуализаций
- Информация легко считывается
- Элементы дизайна в большинстве своем оправданы тематикой

<http://ripetungi.com/stevengerrard/steven-gerrard-career-infographic-lfc-england.jpg>

https://i.redditlemedia.com/odxOq2_DtTQITkgS3Q6KVx3SW4KaUkpQZ70mlwj1INk.jpg?s=ada746e9540c49d3119ab9ecf94ffd37