Визуализация данных: Что лучше работает?

Эдвард Тафте, профессор Йельского университета хорошо известен своей новаторской работой в области информационного дизайна, исследовал картографию для упрощения изучения всех типов информации, а также визуализации данных, которая фокусируется на визуальных картографах для отображения данных.

Classic Readings:

Edward Tufte (Professor Emeritus at Yale University) has written four books on analytical design:

1983: The Visual Display of Quantitative Information
1990: Visual Explanations: Images and Quantities, Evidence and Narrative
1990: Envisioning Information
2006: Beautiful Evidence

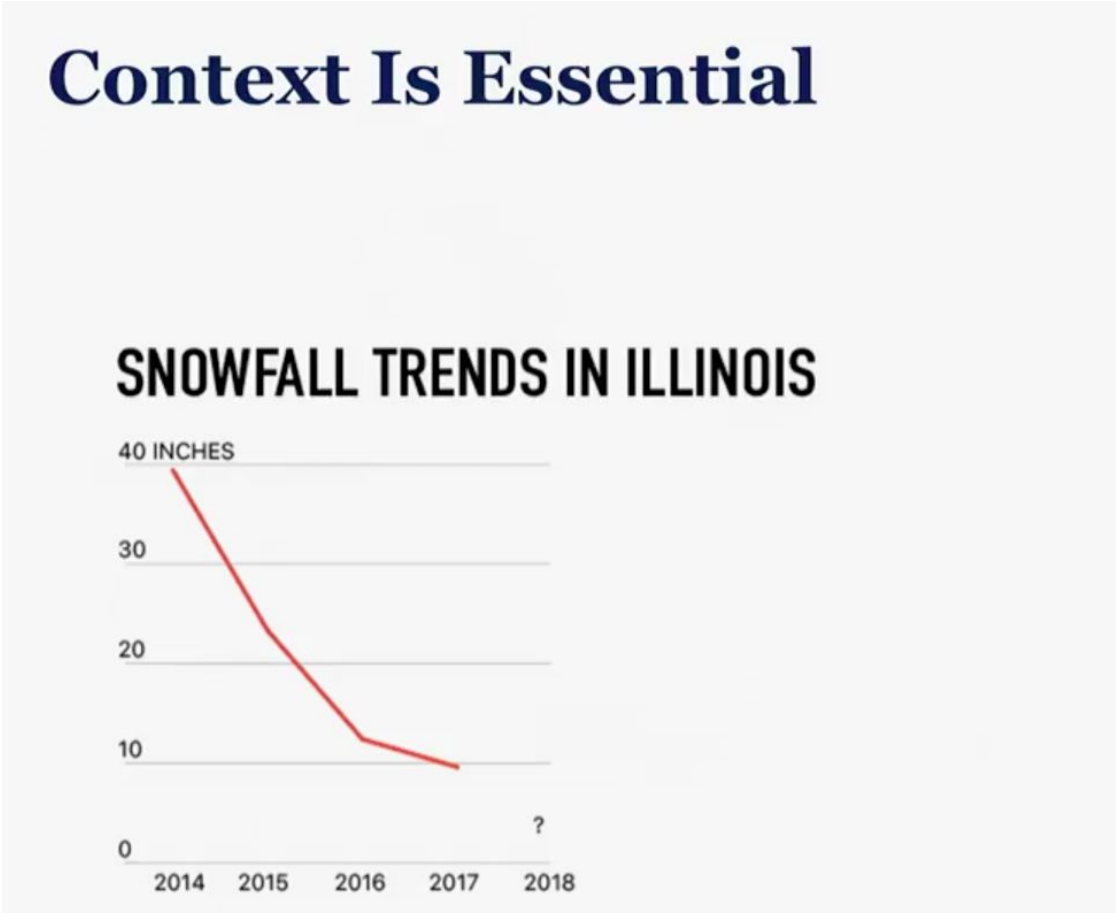
Source: Edward R. Tufte / https://www.edwardtufte.com/tufte/

Целостность графики

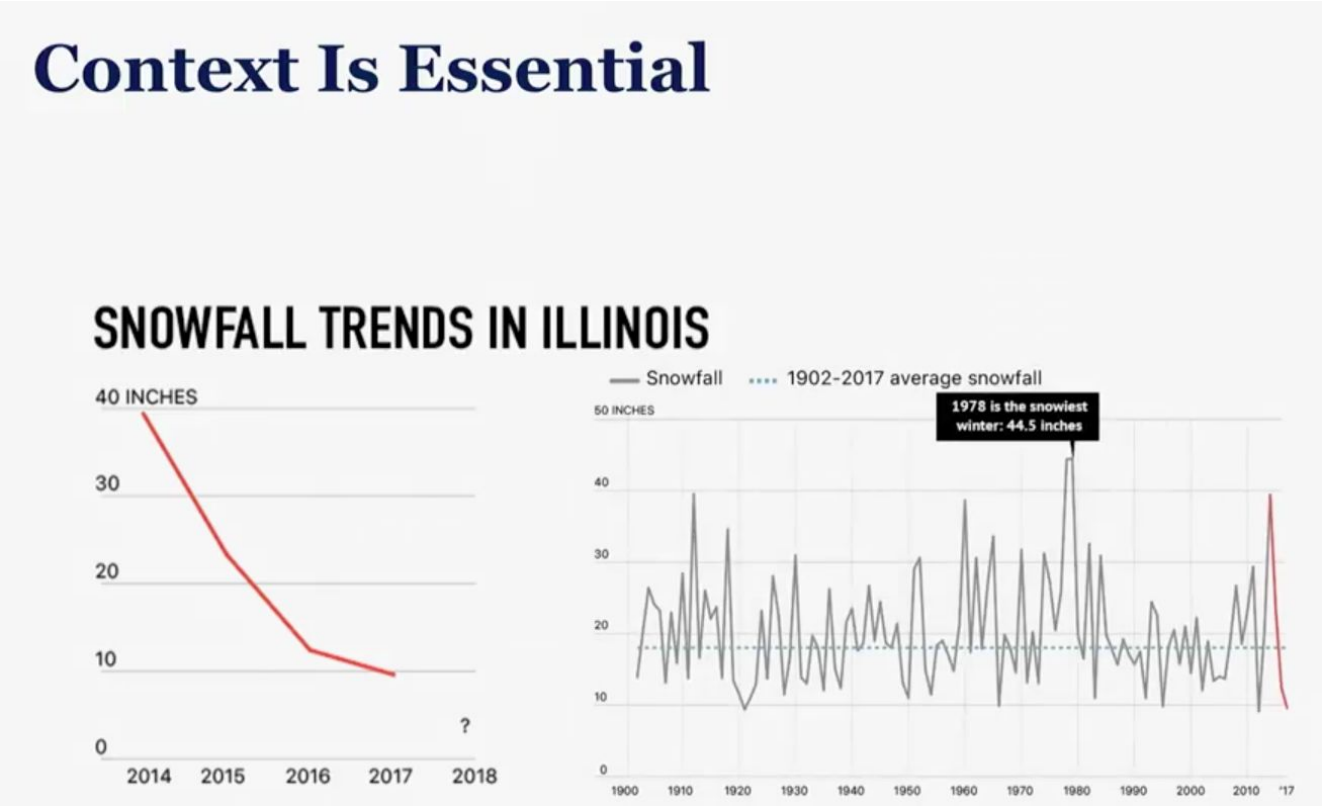
- говорить правду о данных
- графика не должна цитировать данные вне контекста
- избегайте искажения в данных.
- представление чисел должно соответствовать истинным пропорциям.
- маркировка должна быть четкой и подробной.
- для представления финансовых данных лучше всего использовать хорошо известные единицы измерения

ДАННЫЕ ДОЛЖНЫ ОСНОВЫВАТЬСЯ ТОЛЬКО НА ПРАВДЕ!

Снегопад


Context Is Essential
SNOWFALL TRENDS IN ILLINOIS
40 INCHES
30
20
10
0
?
2014
2015
2016
2017
2018

Context Is Essential

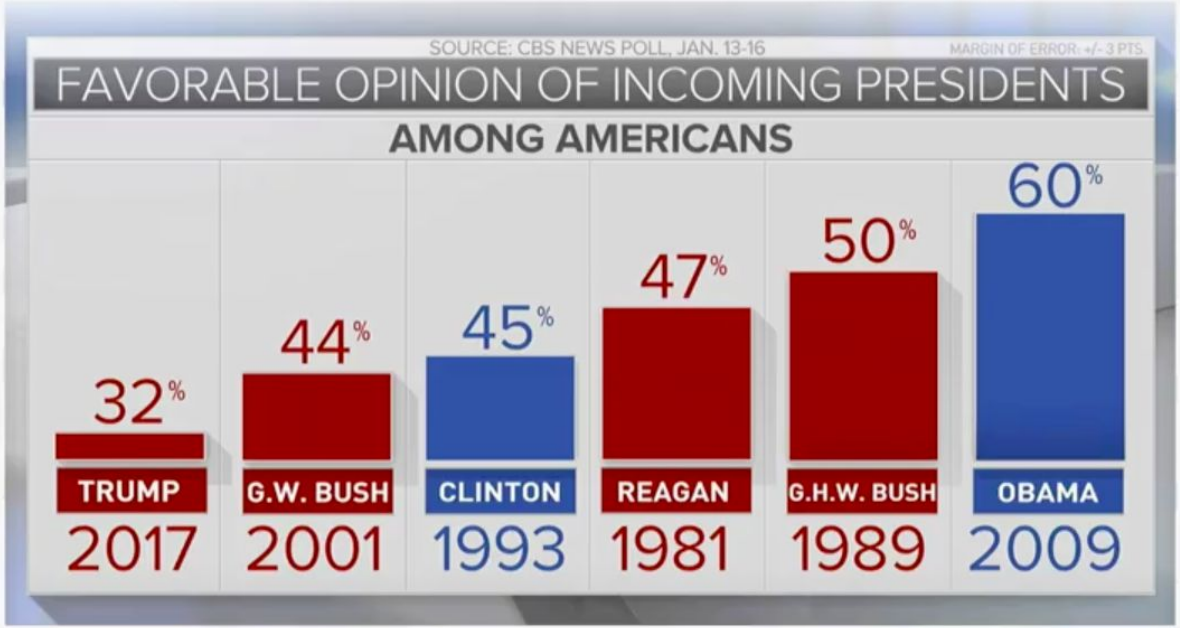


Когда мы видим график, где представлено больше данных за последние десятилетия или даже столетие, получается, что количество снега колеблется, и поэтому количество снега, которое мы ожидаем в 2018 году, не является точным.

Визуализация данных и ложь

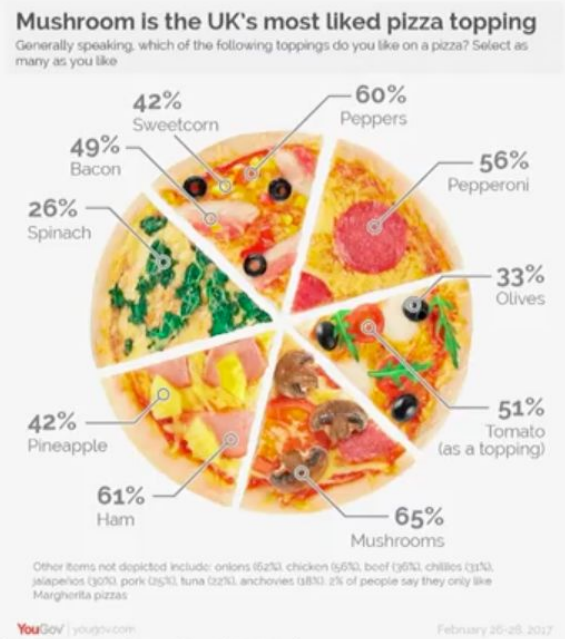
How to Lie with Visualizations

No zero line

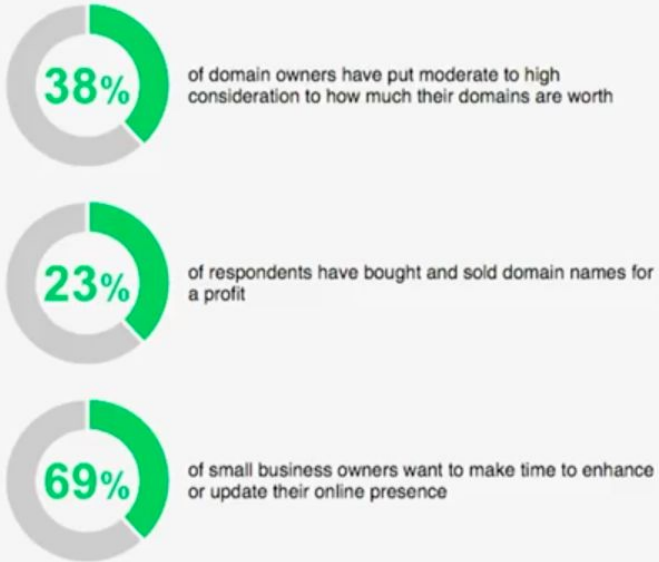


How to Lie with Visualizations

Doesn't add up to 100% and not in right proportion



Source: YouGov / Matthew Smith / http://bit.ly/2ILzNFa /



Source: GoDaddy

Другой способ лгать - это манипулирование областью и размером. Тафте разработал концепцию под названием "Фактор лжи", которая относится к соотношению между размером эффекта, показанного на графике, и размером эффекта, показанного в данных.

Формула.

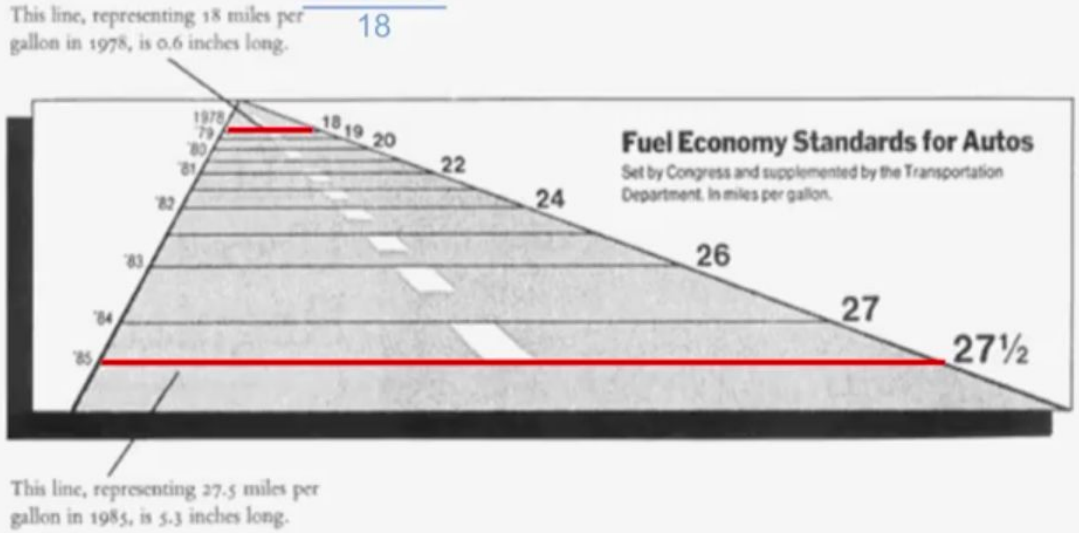
Представление чисел как физической меры на поверхности самого графика должно быть прямо пропорционально представляемой величине.

$$Lie\ Factor = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

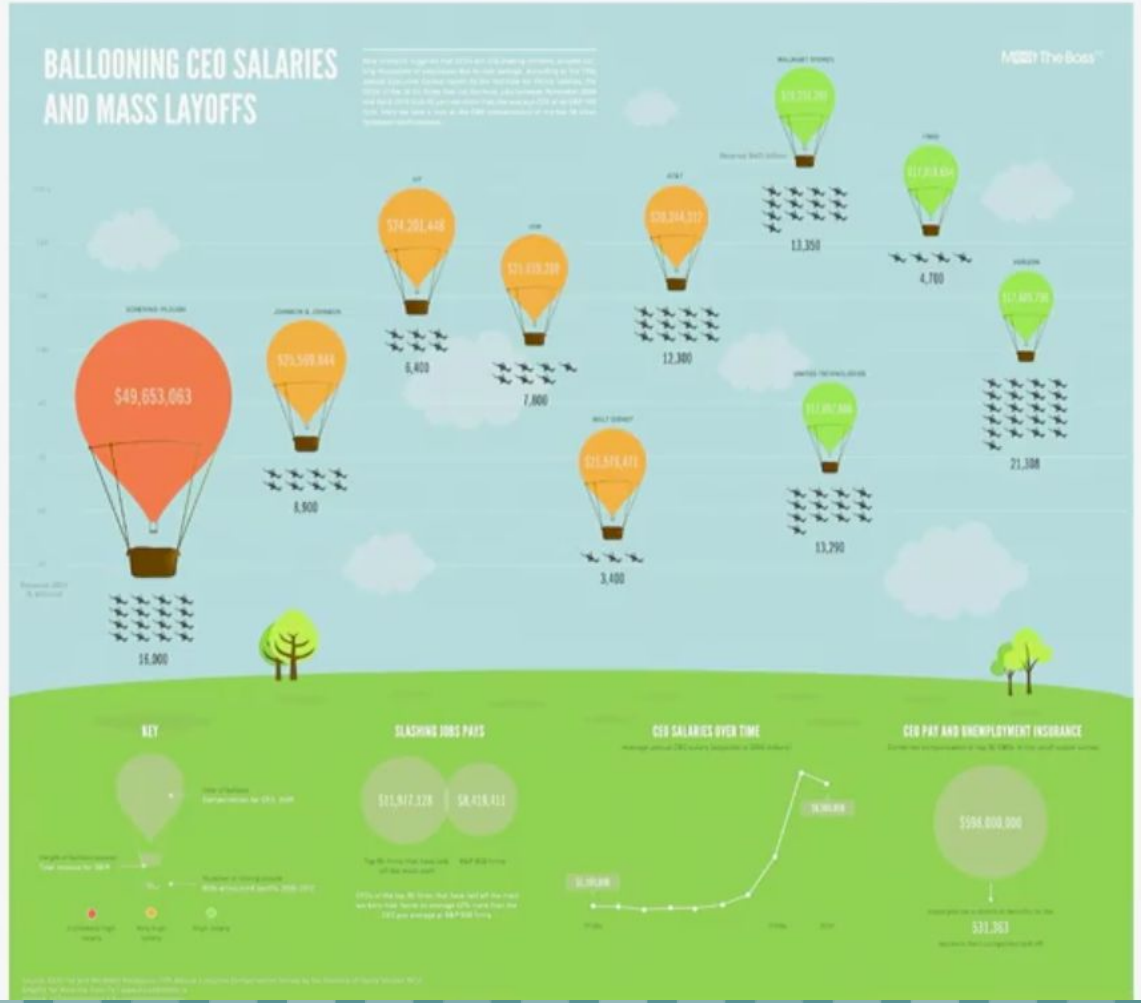
- This graphic (by *The New York Times*) tries to compare the mandated fuel economy standards for autos set by the US Department of Transportation.
- The line representing 18 miles per gallon in 1978, is 0.6 inches long. The line representing 27.5 miles per gallon in 1985, is 5.3 inches long.

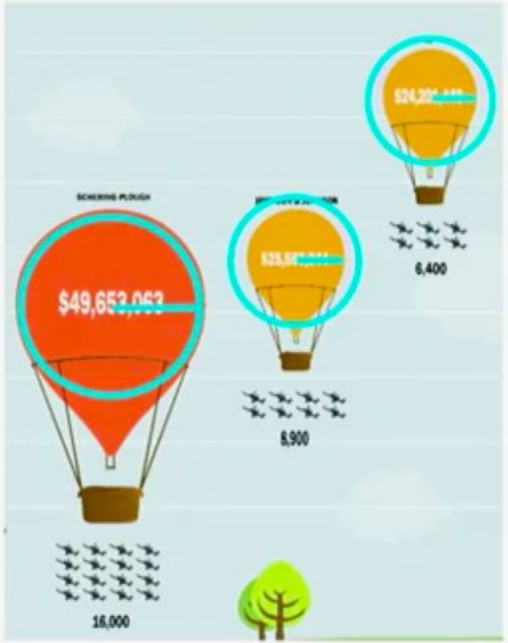
$$\text{Lie Factor} = \frac{\frac{5.3 - 0.6}{0.6}}{\frac{27.5 - 18}{18}} = 14.8$$

Inaccurate visual area and numeric measure!



Source: Edward Tufte / http:/





Inaccurate visual area and numeric measure

тест

Что такое фактор лжи?

A. Значение, подсчитывающее количество неточностей графика.

B. Значение отношения, которое некорректно отображает эффект данных из-за неточного размера графика.

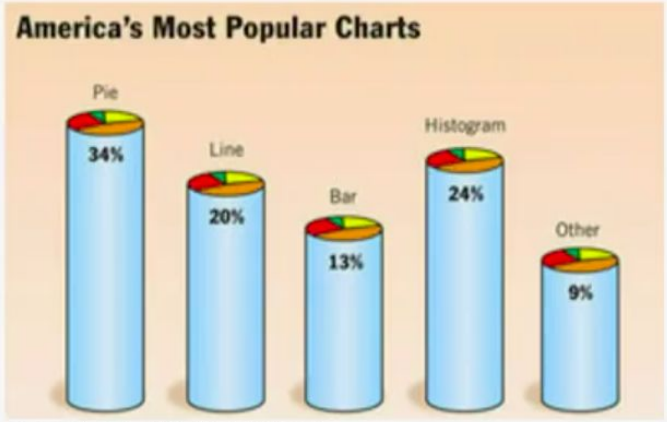
C. Значение коэффициента, отображающего эффект данных в связи с размером графика

D. Значение, подсчитывающее количество ложных данных.

Graphical Integrity (2)

Show data variation, not design variation.

- Induce the viewer to think about the substance rather than about the data.



Source: The Onion / http://bit.ly/2GUQOuO

Source: Junk Charts / http://bit.ly/2UfPohw

Тафте также просит дизайнера не путать вариации дизайна.

Chartjunk

- Chartjunk refers to anything in a chart that does not represent data as not just unnecessary, but harmful.

Common types:

- **Vibrating chartjunk** – which is cross-hatching or other patterns that distract the mind from the information being presented

- **Grids** – over-busy grid lines and excess ticks

- **Self-promoting graphics** ("The Duck chart") – where color schemes and patterns are introduced for artistic appeal rather than information content

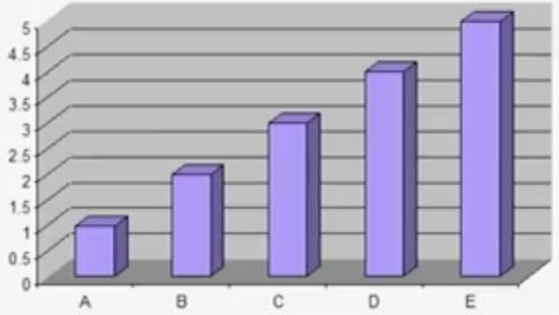


Chart Junk (графический мусор**)**, относится ко всему, что представлено на графике кроме данных.

Соотношение чернильных данных (ink)

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used}}$$

Data-Ink Ratio

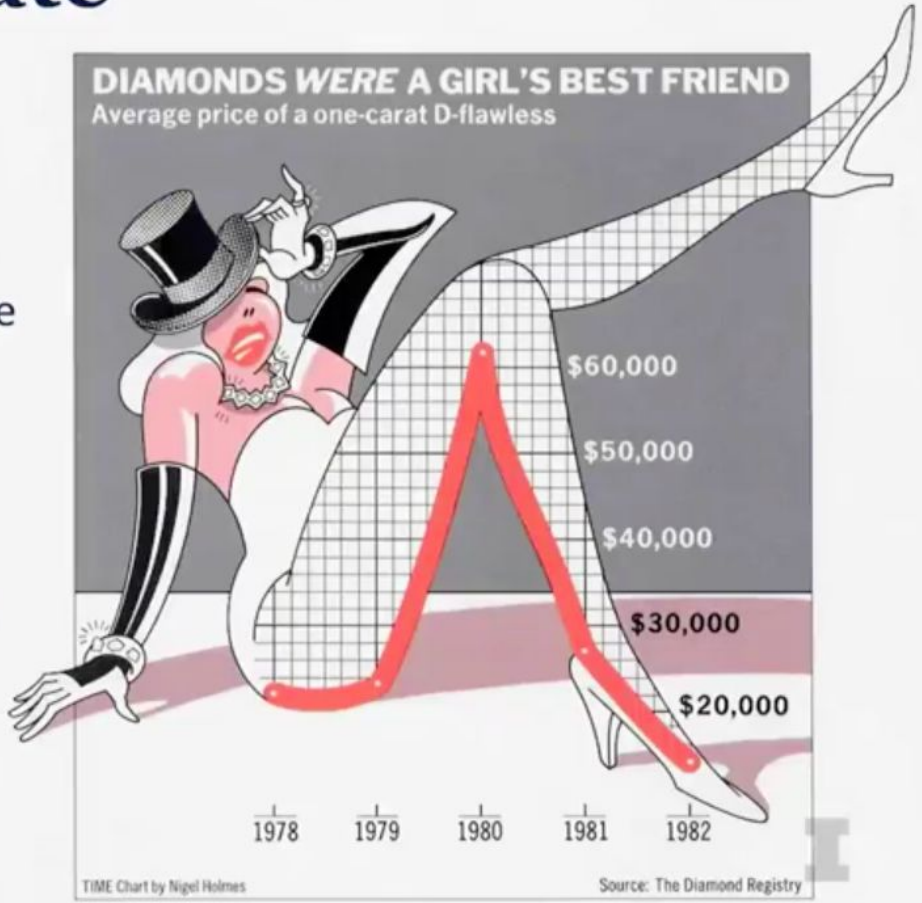
- It is an electroencephalogram – a graph that records the electrical activity from the brain.
- This graph would have a data-ink ratio of 1.

Использование декоративных элементов без цели может привести к искажению данных и неправильной интерпретации данных пользователем.

The Chartjunk Debate

- Holmes: “From ankle to knee, the price changes from lowest to highest over time. By combining image with graph, it made viewers remember the spike shape of the graph more easily.”

- Tufte: “Everything counts, but nothing matters...The data-thin (and thus uncontextual) chart mixes up changes in the value of money with changes in diamond prices...”



Chartjunk Тафте рассматривает как графический мусор

The Chartjunk Debate

Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts

Scott Bateman, Regan L. Mandryk, Carl Gutwin,
Aaron Genest, David McDine, Christopher Brooks
Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
scott.bateman@usask.ca, regan@cs.usask.ca, gutwin@cs.usask.ca,
aaron.genest@usask.ca, dam085@mail.usask.ca, cab938@mail.usask.ca

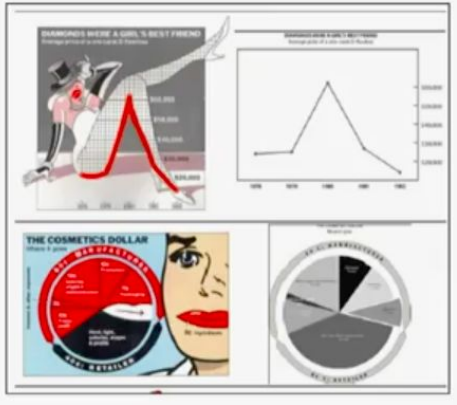
ABSTRACT
Guidelines for designing information charts often state that the presentation should reduce 'chart junk' – visual embellishments that are not essential to understanding the data. In contrast, some popular chart designers wrap the presented data in detailed and elaborate imagery, raising the questions of whether this imagery is really as detrimental to understanding as has been proposed, and whether the visual embellishment may have other benefits. To investigate these issues, we conducted an experiment that compared embellished charts with plain ones, and measured both interpretation accuracy and long-term recall. We found that people's accuracy in describing the embellished charts was no worse than for plain charts, and that their recall after a two-to-three-week gap was significantly better. Although we are cautious about recommending that all charts be produced in this style, our results question some of the premises of the minimalist approach to chart design.

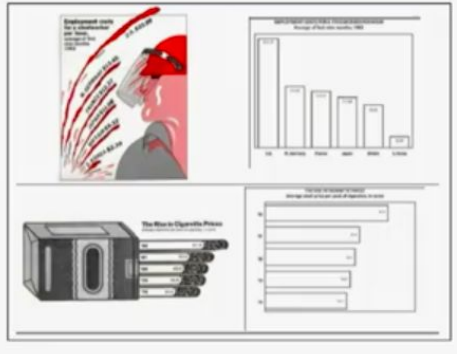
Author Keywords
Charts, information visualization, imagery, memorability.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

data-ink – or the ink in the chart used to represent data.

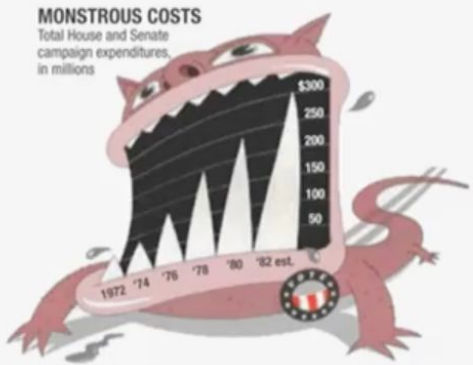
Despite these minimalist guidelines, many designers include a wide variety of visual embellishments in their charts, from small decorations to large images and visual backgrounds. One well-known proponent of visual embellishment in charts is the graphic artist Nigel Holmes, whose work regularly incorporates strong visual imagery into the fabric of the chart [7] (e.g., Figure 1).



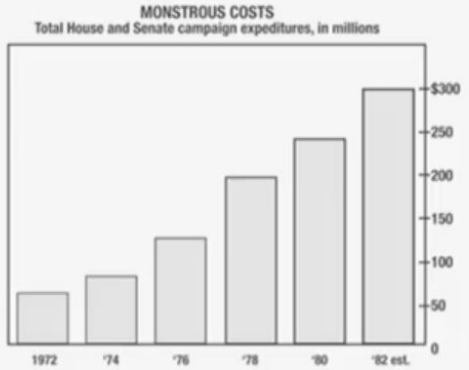


Criticism

- Readers do not find highly embellished charts difficult or easy to comprehend, but they do remember those charts better and find them more attractive and more enjoyable to look at than their plainer alternatives. (Bateman et al., 2010)
- Borkin et al. (2013) also found that visually distinctive infographic style charts were more memorable than more standard statistical charts.



Source: TIME / Nigel Holmes



Читатели считают, что **chartjunk** не влияет на понимание читателя, но на самом деле читатели запоминают, что приукрашенные **chartjunk** смотрятся интересней, чем те же обычные графики.

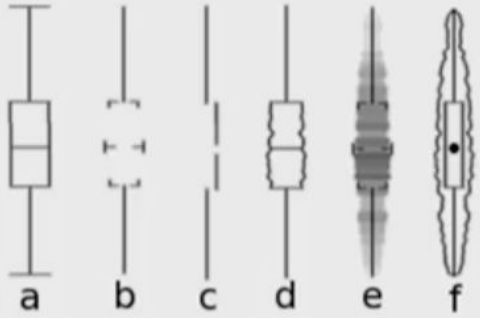
Data-Ink Ratio Debate

- Anderson et al. (2011) asked participants to choose the box plot with the largest range from a set
- Six representations
- Measured cognitive load from EEG brain waves
- The simplest (c) box plot is the hardest to interpret

A User Study of Visualization Effectiveness Using EEG and Cognitive Load

E. W. Anderson[1], K. C. Potter[1], L. E. Matzen[2], J. F. Shepherd[2], G. A. Preston[3], and C. T. Silva[1]

[1]SCI Institute, University of Utah, USA
[2]Sandia National Laboratories, USA
[3]Utah State Hospital, USA



Мы должны помнить о пяти принципах построения графиков данных.

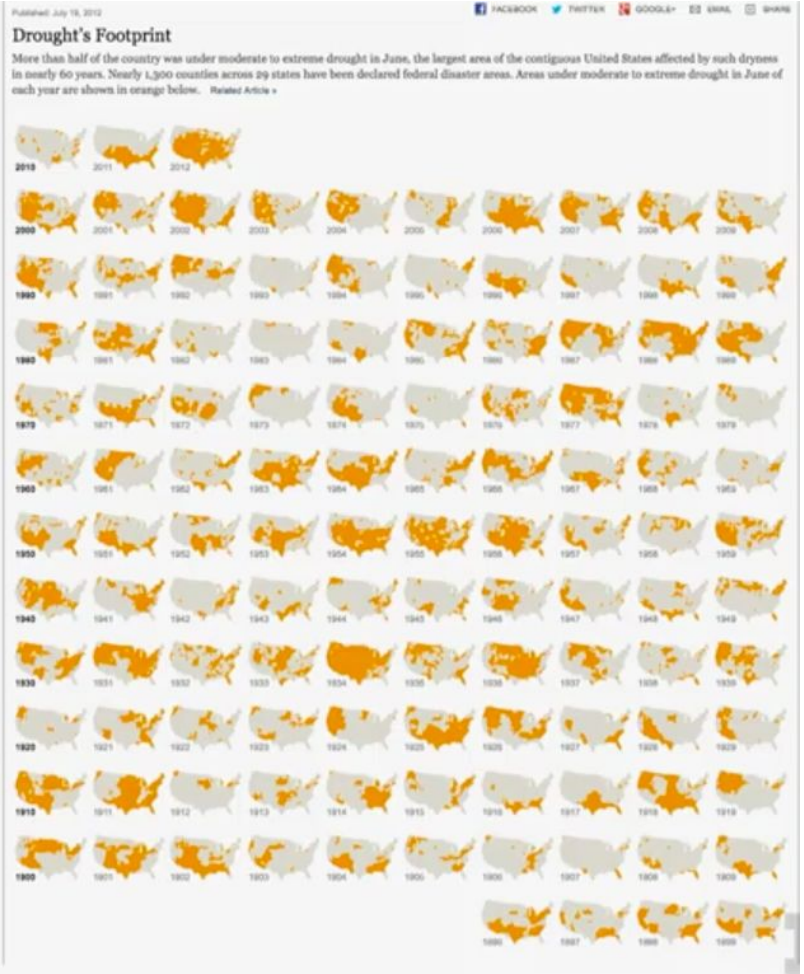
- максимизируйте соотношение "чернил" в пределах разумного
- стирайте не чернил в пределах разумного
- стирайте избыточные чернила в пределах разумного
- пересматривайте и редактируйте

Тафте предлагает сделать большие наборы данных согласованными, ***представив множество данных наименьшим образом.*** Другими словами, ***использовать наименьшее количество чернил в самом маленьком пространстве.***

Small Multiple

Uses a single design repeated several times within the eyespan, each example showing a different value of the independent variable(s)

Allows easy comparison



Source: The New York Times / Haeyoun Park and Kevin Quealy / https://goo

Основные типы маркировки включают в себя:

- заголовок диаграммы
- заголовок осей
- этикетки и шкалу
- некоторые примечания, которые отмечают специфику внутри диаграммы

Совершенство достигается не тогда, когда больше нечего добавить, а там, где больше нечего убрать.