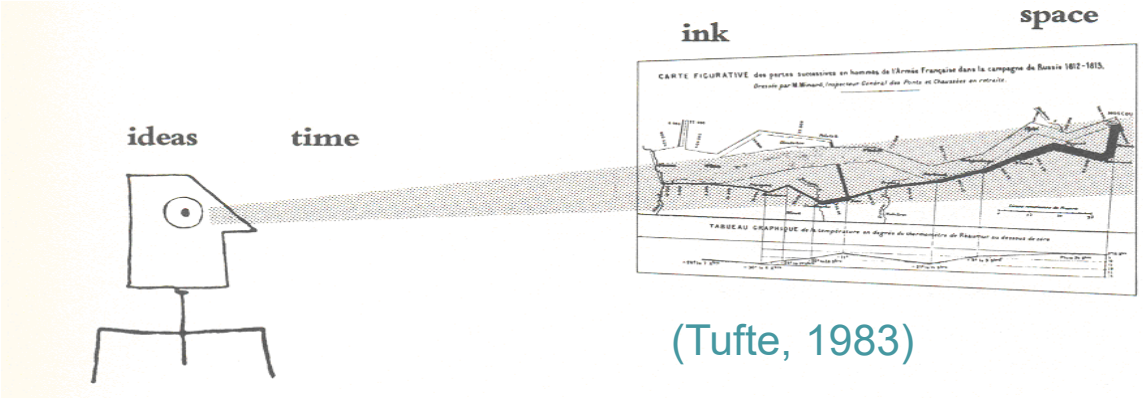
Universidade de Aveiro
Departamento de Electrónica,
Telecomunicações e Informática

Excellence and integrity in graphics and visualization



(Tufte, 1983)

Beatriz Sousa Santos, Universidade de Aveiro

Excellence in (statistical and other) graphics

- Excellence in statistical (and other) graphics consists of complex ideas communicated with clearity, precision and efficiency
- Graphics should:
  - Show the data
  - Induce the user to think about the substance rather than form
  - Avoid distorting what the data have to say
  - Present many numbers in a small space
  - Reveal the data at several levels of detail
  - …

- Graphics can be more precise and revealing than conventional statistics computations

- Consider Anscombe´s quartet: all four datasets are described by the same linear model

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

(Tufte, 1983)

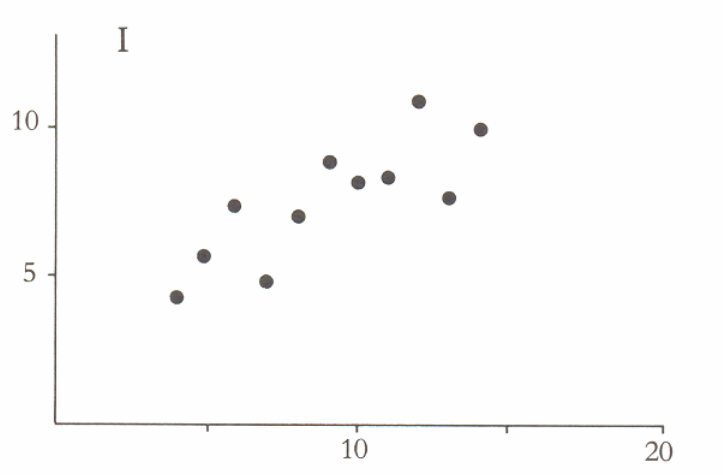
N=11;

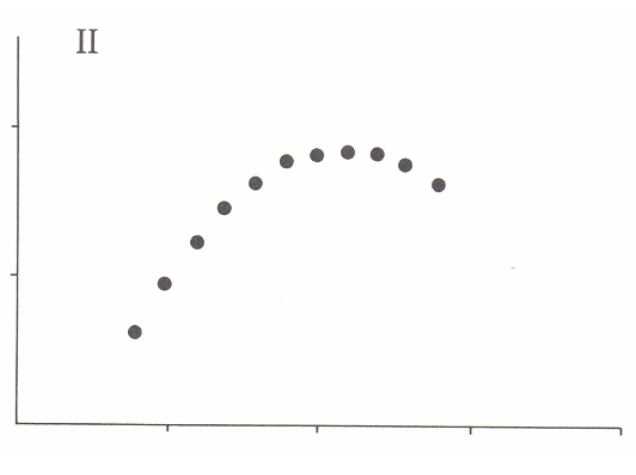
Mean values: X = 9.0; correlation coefficient = 0.82

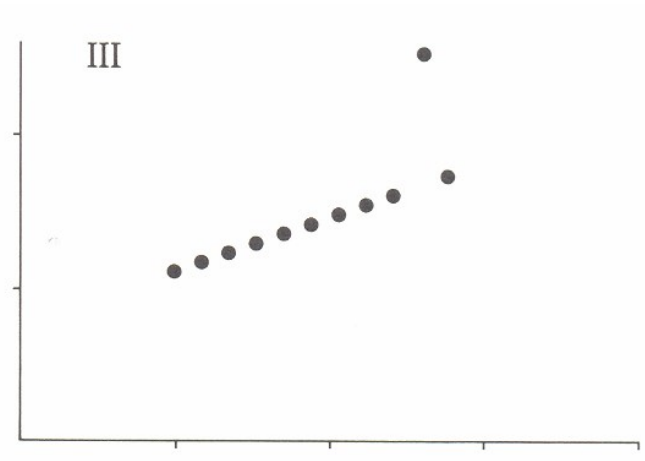
Y = 7.5 equation of regression line: Y = 3 + 0.5 X

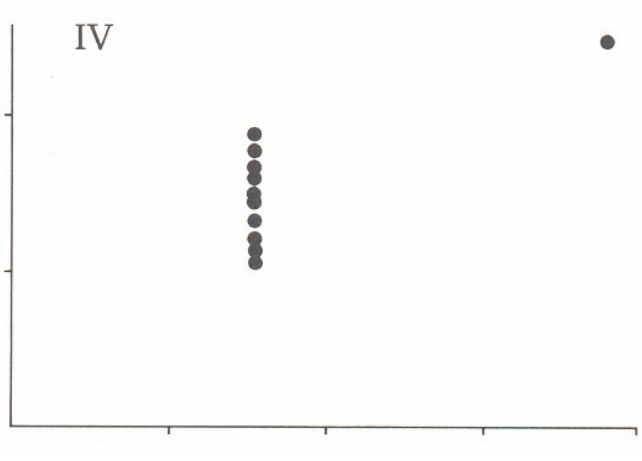
etc.

- Yet the graphical display of the data makes vividly clear how they differ

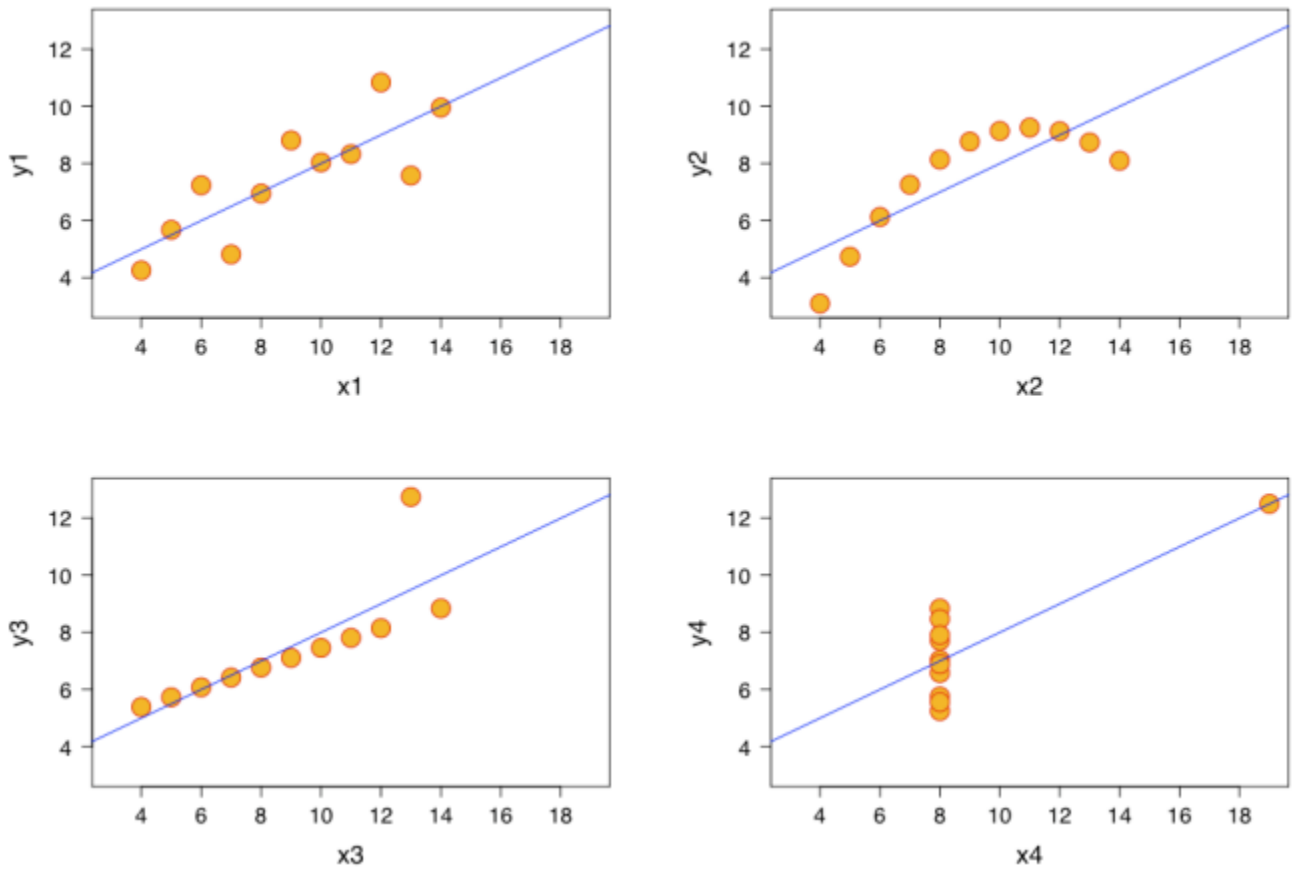








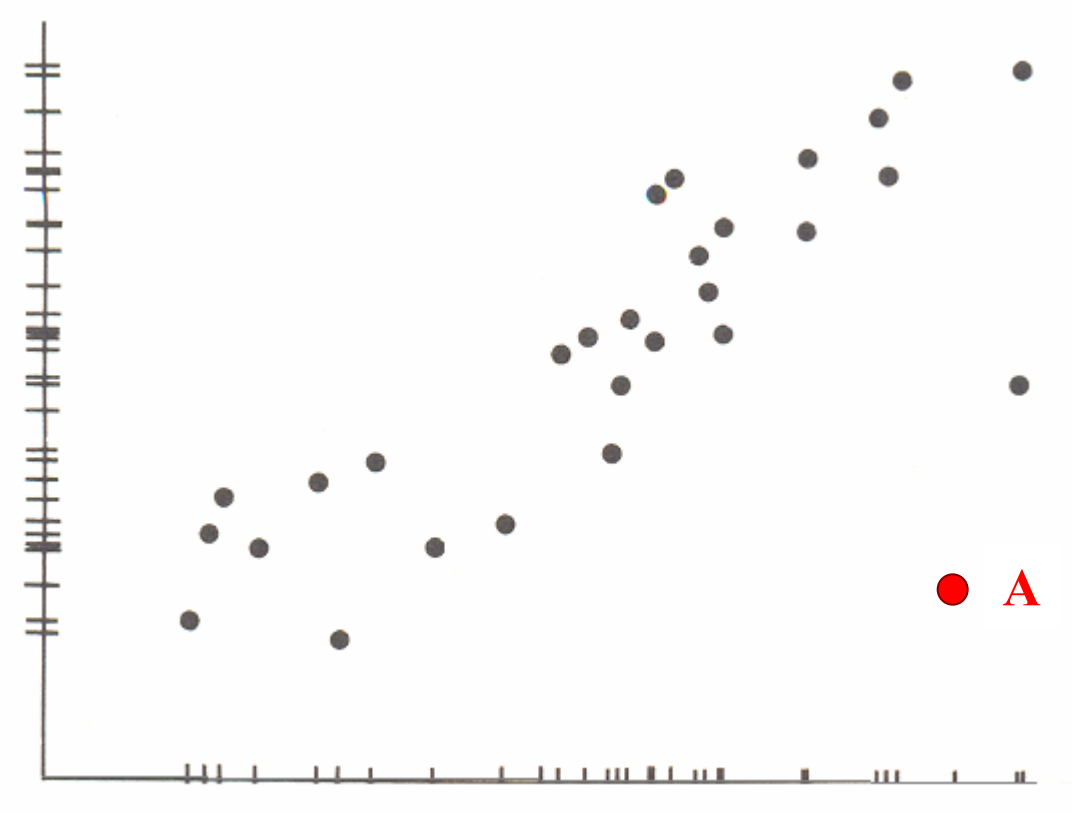
(Tufte, 1983)



Anscombe's quartet and regression line: Visual inspection immediately shows how their structures are quite different

http://en.wikipedia.org/wiki/File:Anscombe.svg

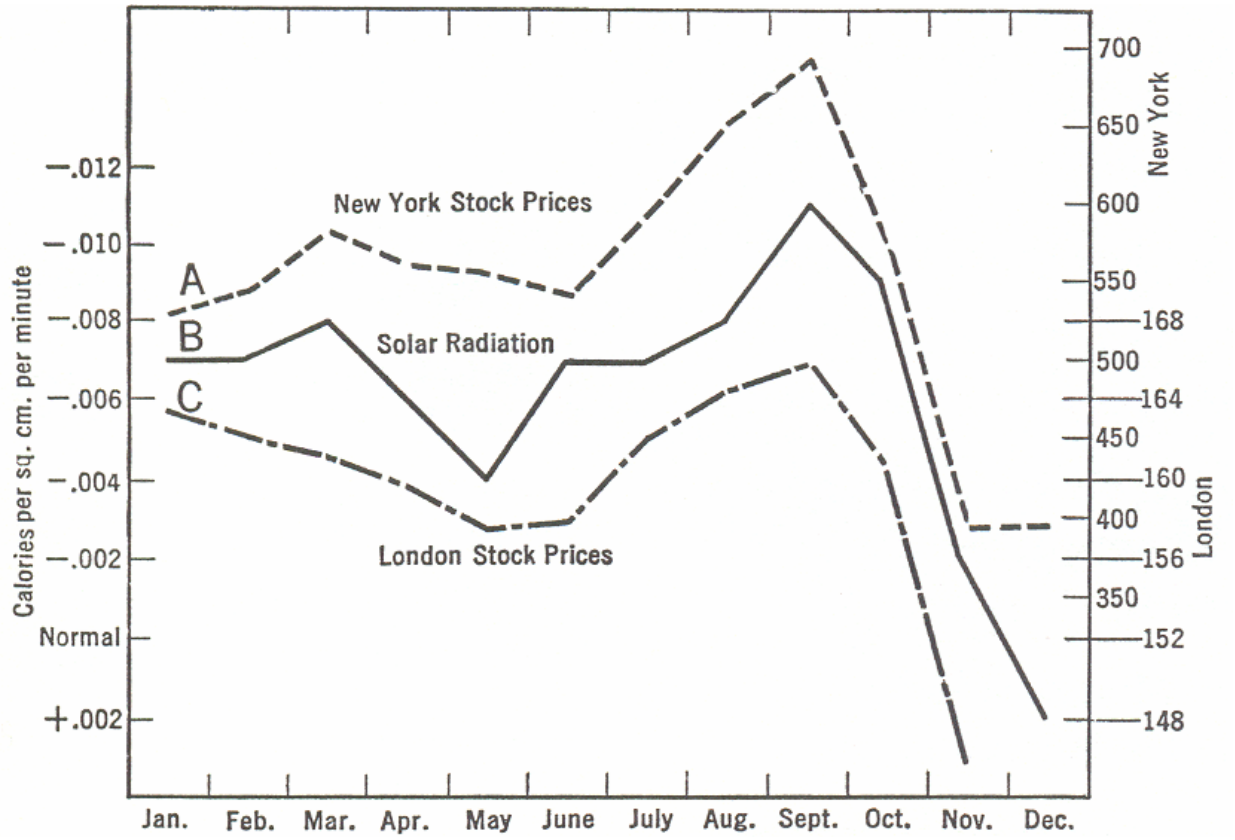
- Point A is easily revealed as a wildshot observation that will dominate standard statistical calculations



(Tufte, 1983)

- Of course, graphics (statistical calculations) <u>are only as good as what goes into them</u>

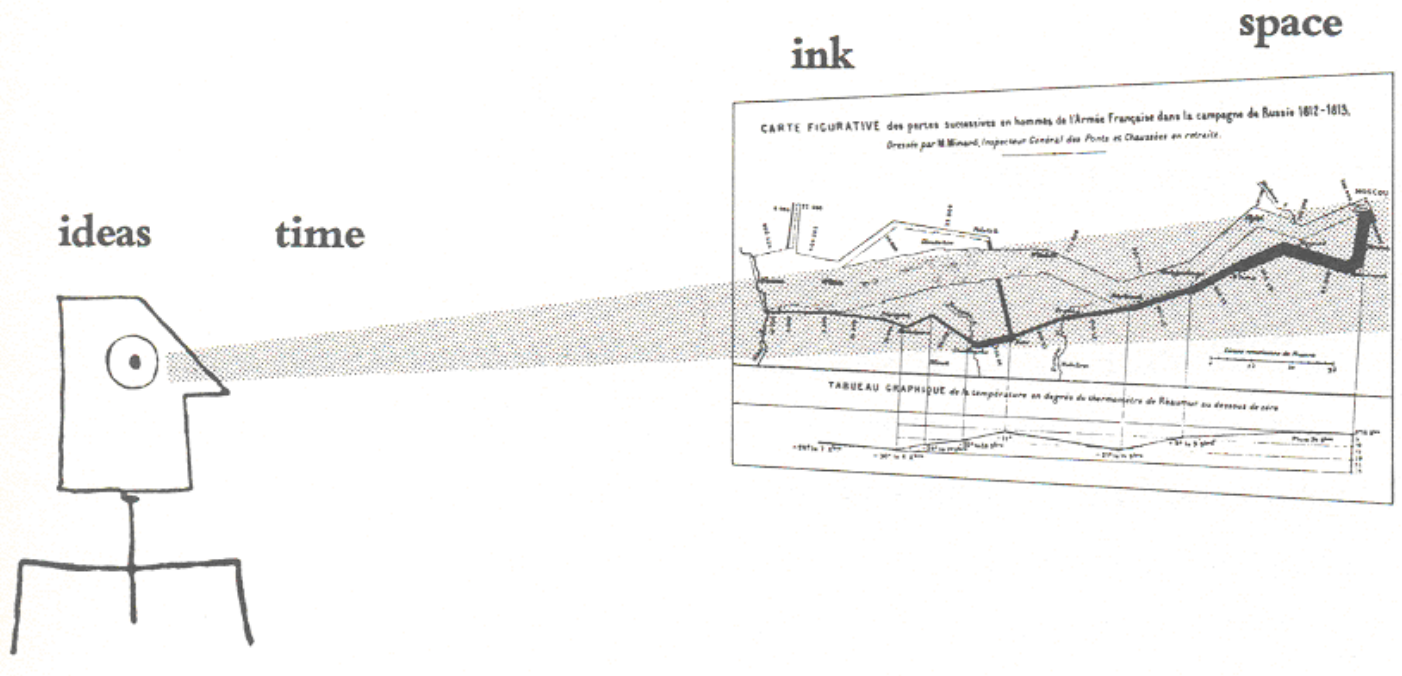
- A hill-specified or preposterous model or a puny data cannot be rescued by a graphic



A silly theory → a silly graphic

(Tufte, 1983)

- Graphical **excellence** is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space



- consists of **complex ideas** communicated with **clarity**, **precision** and **efficiency**
- is nearly always **multivariate**
- requires telling the **truth** about the data

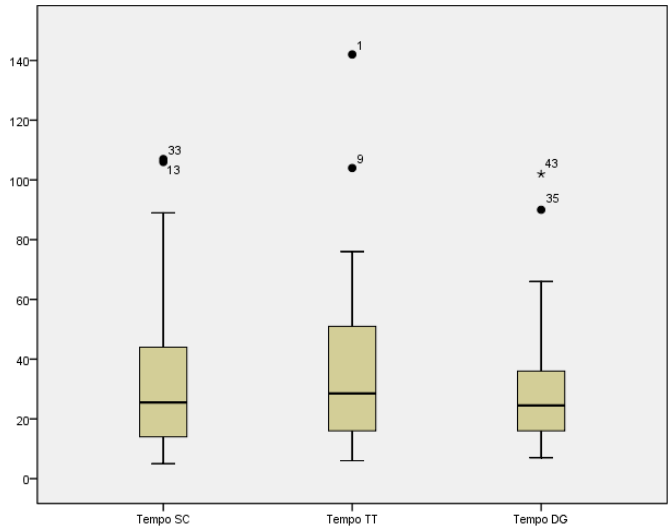
Graphical integrity

“In the late 1960 Jonh Tukey made statistical graphics respectable”

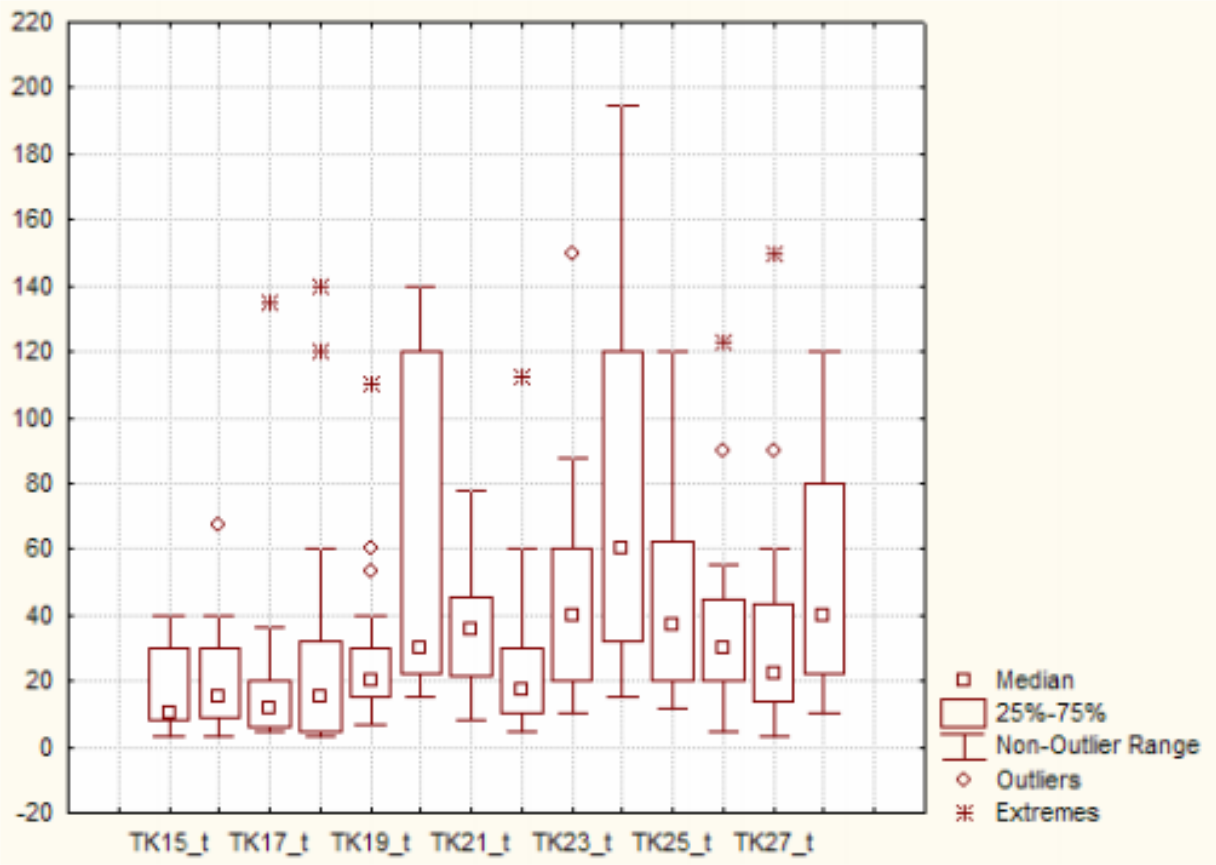
Puting an end on the view that graphics were only for decorating a few numbers

He invented several new designs to explore complex data

boxplot →



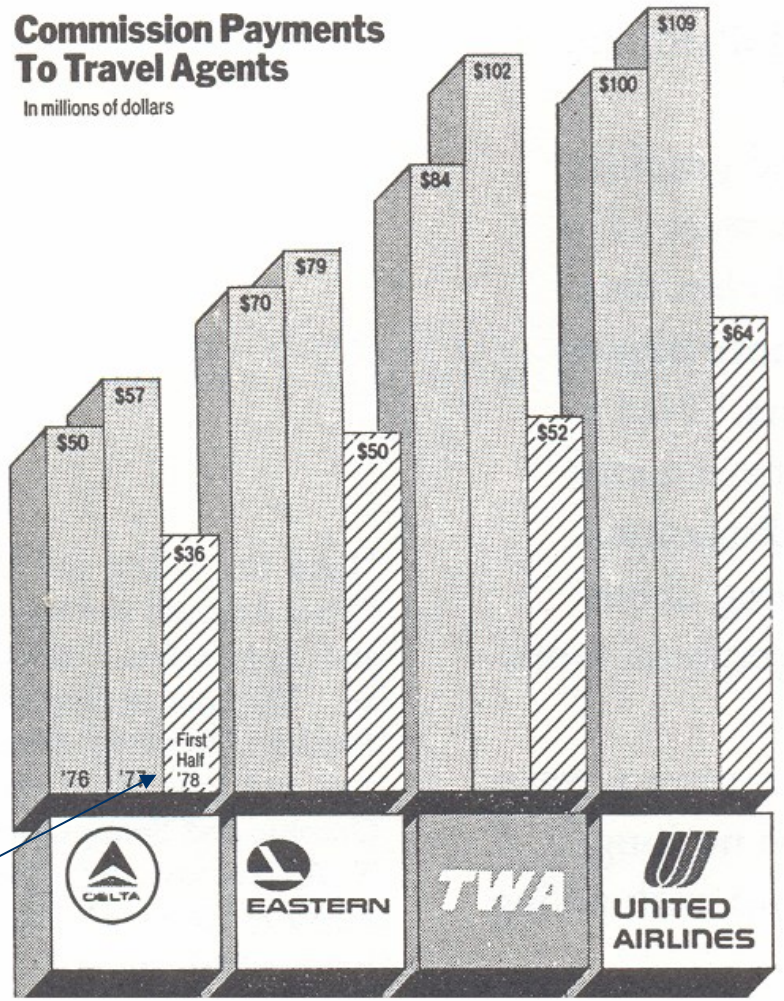
Example:



Box-plot representing times taken by users, by task, at a usability test

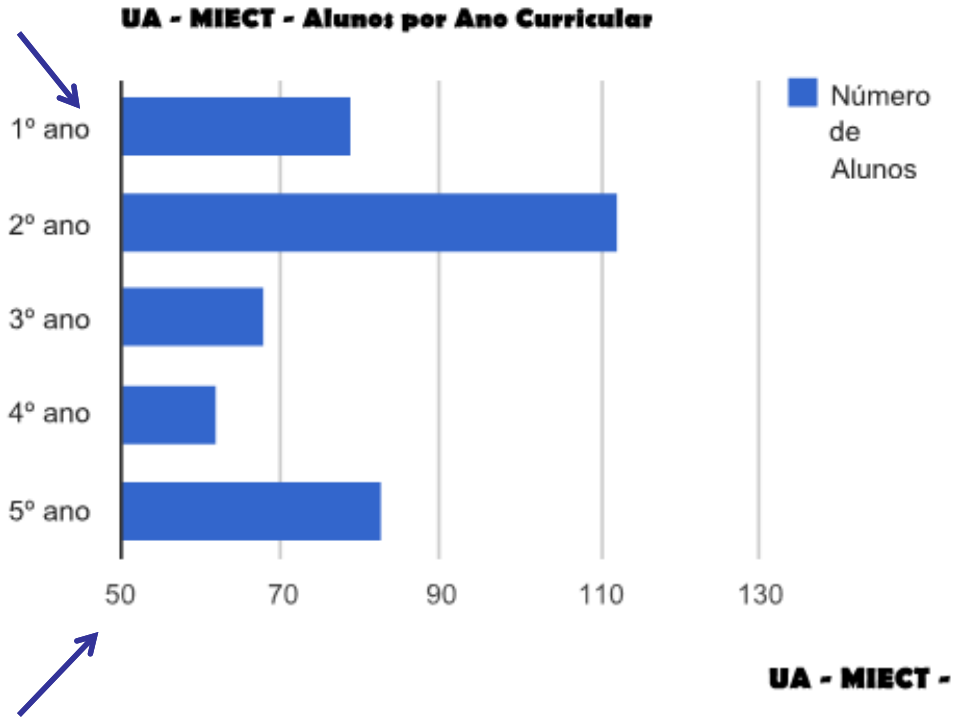
Graphic that fails to tell the truth

- Graphical excellence starts by telling the truth
- This graphic doesn't tell the truth
- The pseudo-decline was created by comparing six months' (1978) to full years (1976, 77)

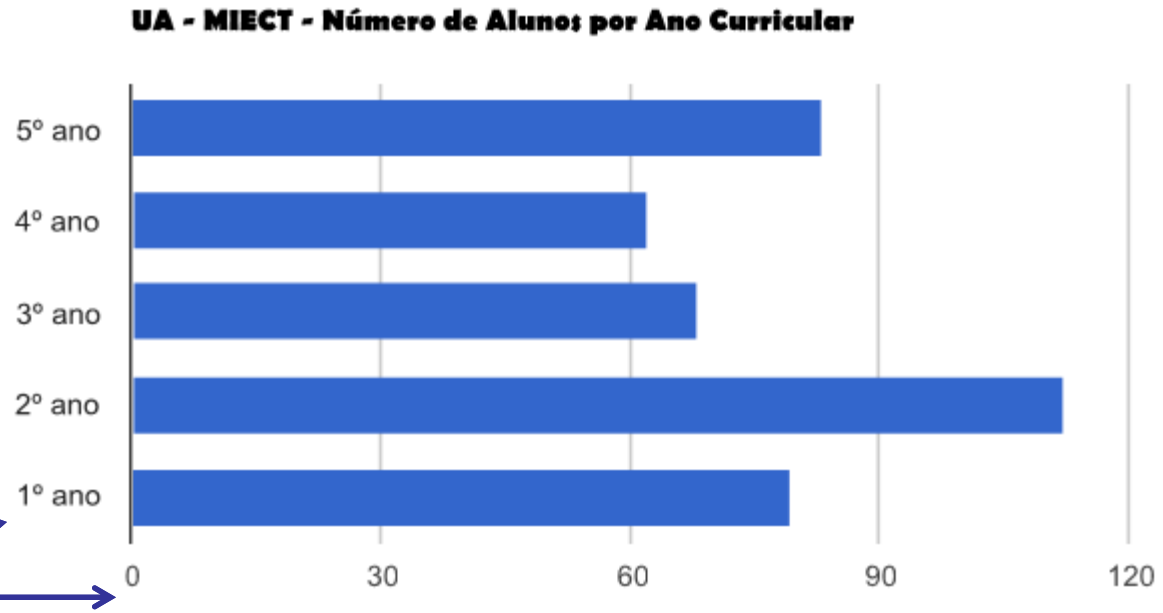


not visible Information !

(Tufte, 1983)



What is wrong with this bar chart?



Better chart

http://www.ieeta.pt/~jmadeira/VI/Exemplos/GCT_Ex_01.htm

What is distortion in a data graphic?

- A graphic does not distort if the visual representation of the data is consistent with the numerical representation
- What is the “visual representation” of the data?

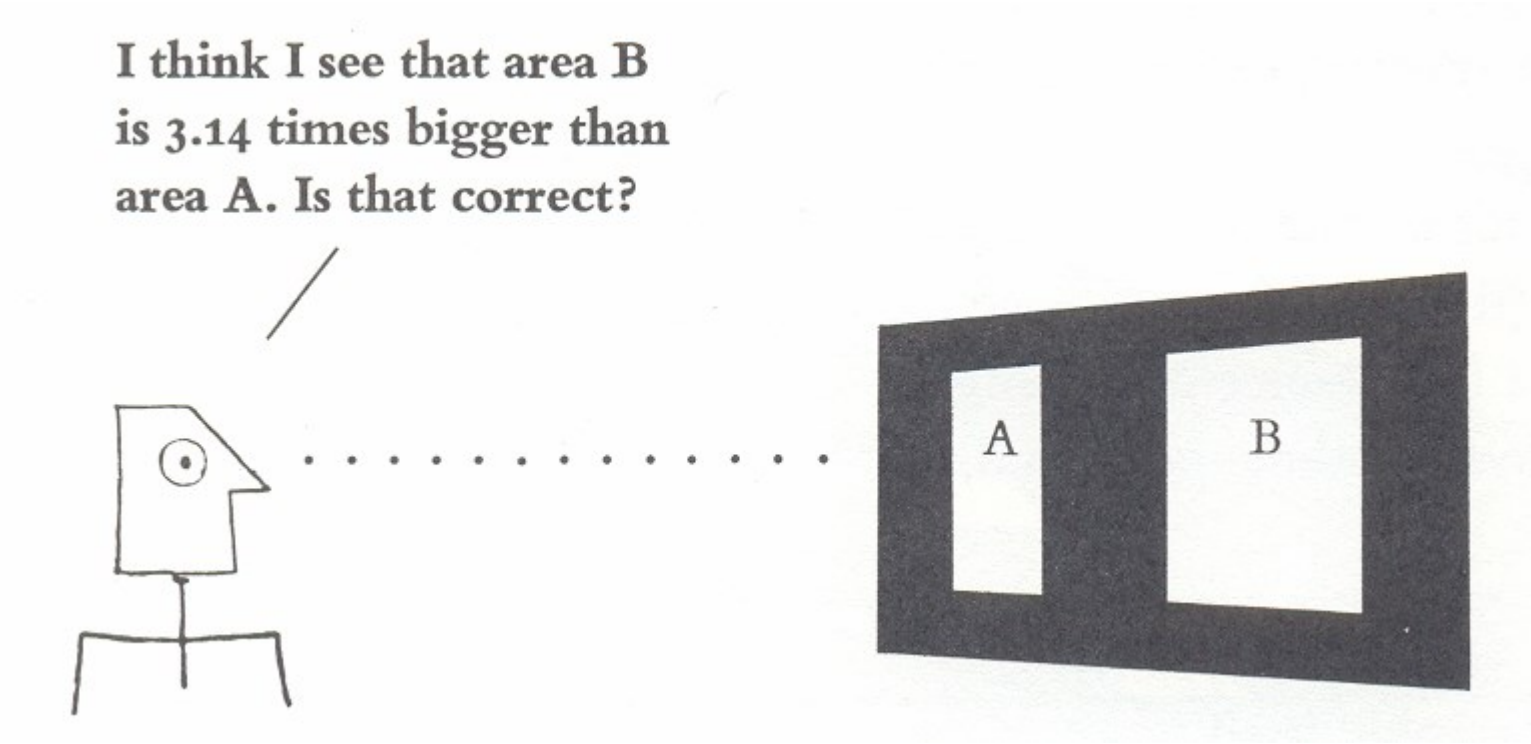
As physically measured on the surface of the graphics?

or

perceived visual effect?

- How do we know that the visual image represents the underlying numbers?

- One way to try to answer these questions is to conduct experiments on the visual perception of graphics



(Tufte, 1983)

- Perception varies with

  - context

  - experience

(Tufte, 1983)

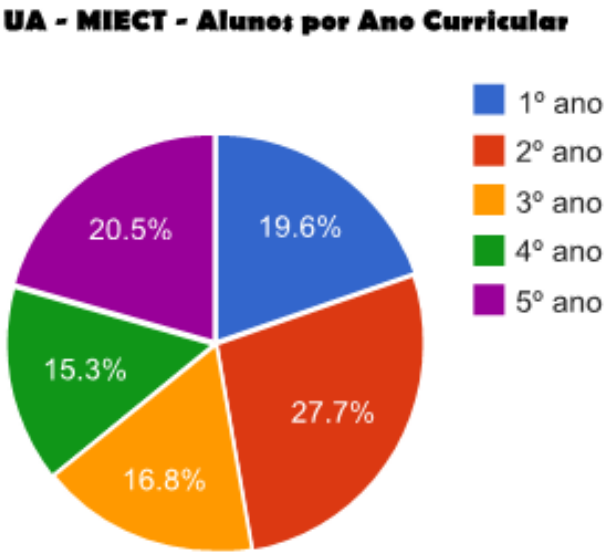
- What to do when we want to represent data in a graphic ?

- According to Tufte there are two fundamental principles to get graphical integrity:

  - represent numbers, as physically measured on the surface of the graphic itself, directly proportional to the numerical quantities represented

  - Clear and thoroughly label to defeat graphical distortion and ambiguity

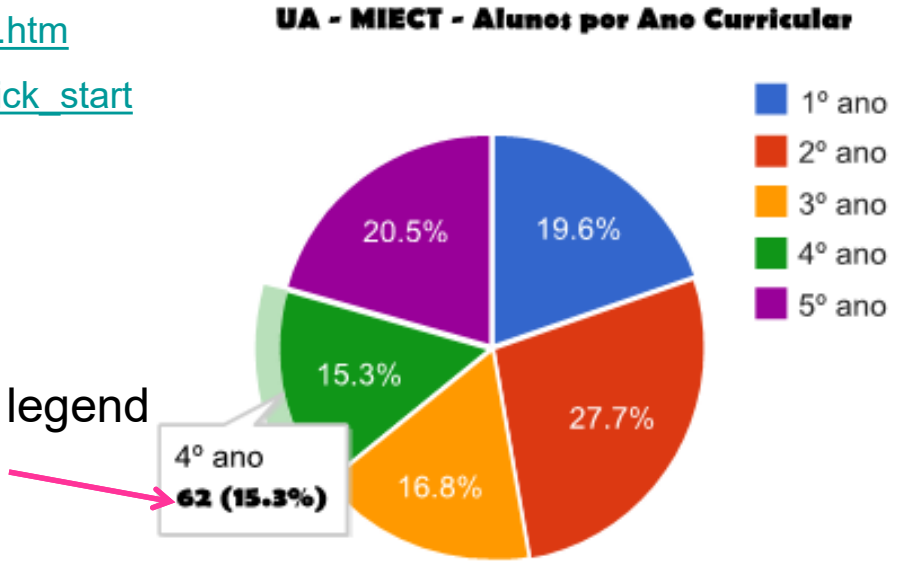
  Note:
  Visual representations must be **tested** as to their efficiency and efficacy for the target users to perform their tasks



- Area of each sector proportional to the represented value
- Values shown on the pie

http://www.ieeta.pt/~jmadeira/VI/Exemplos/GCT_Ex_01.htm

https://developers.google.com/chart/interactive/docs/quick_start



Interactive legend

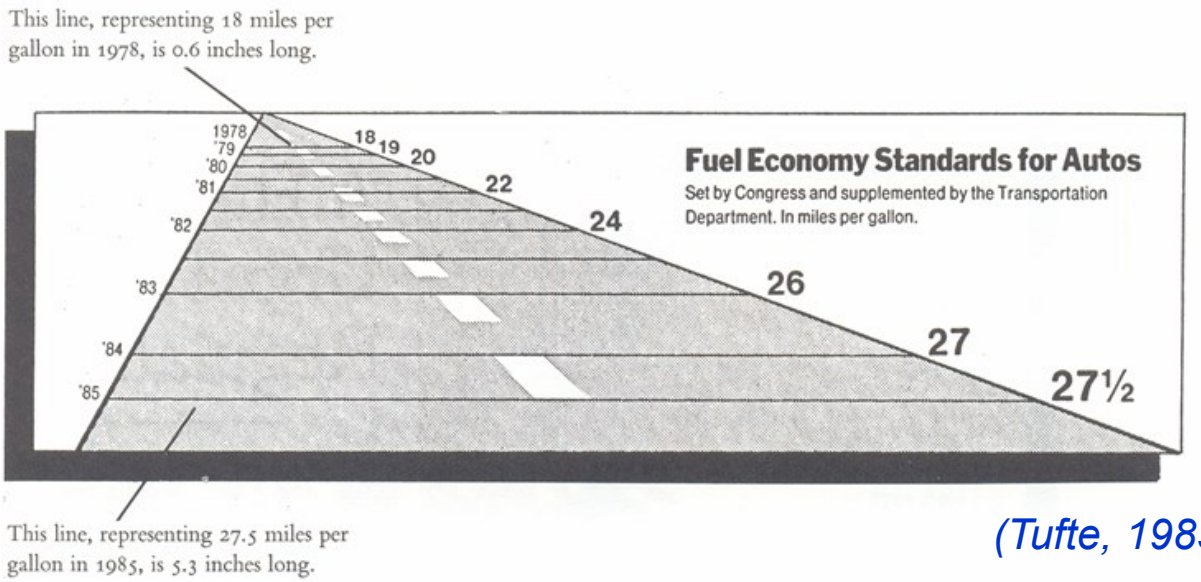
• Violation of the first principle may be measured using the "*Lie Factor*"

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

Lie factor < 0.95

→ significant distorsion

or > 1.05

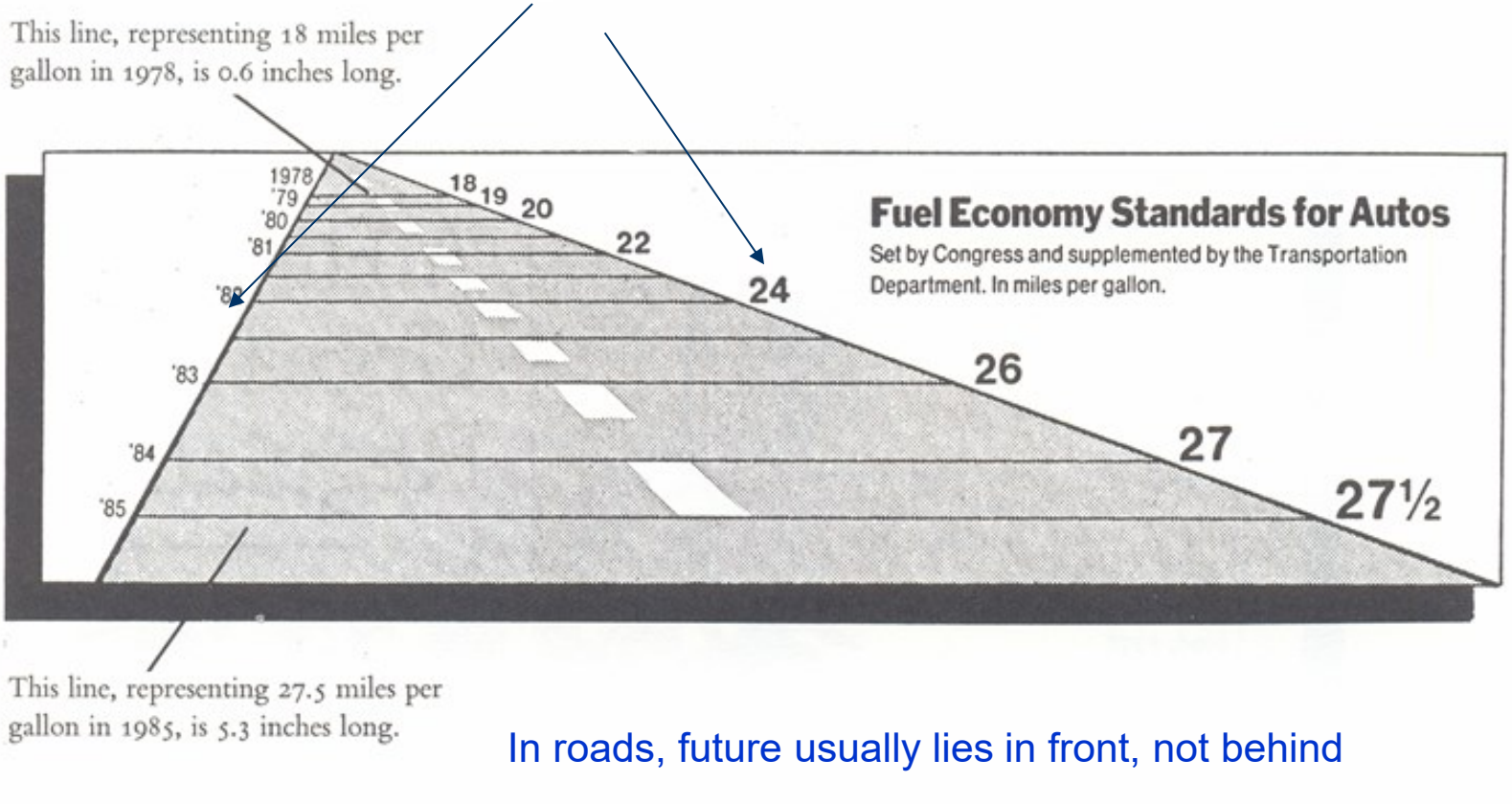


(Tufte, 1983)

this example has several problems:

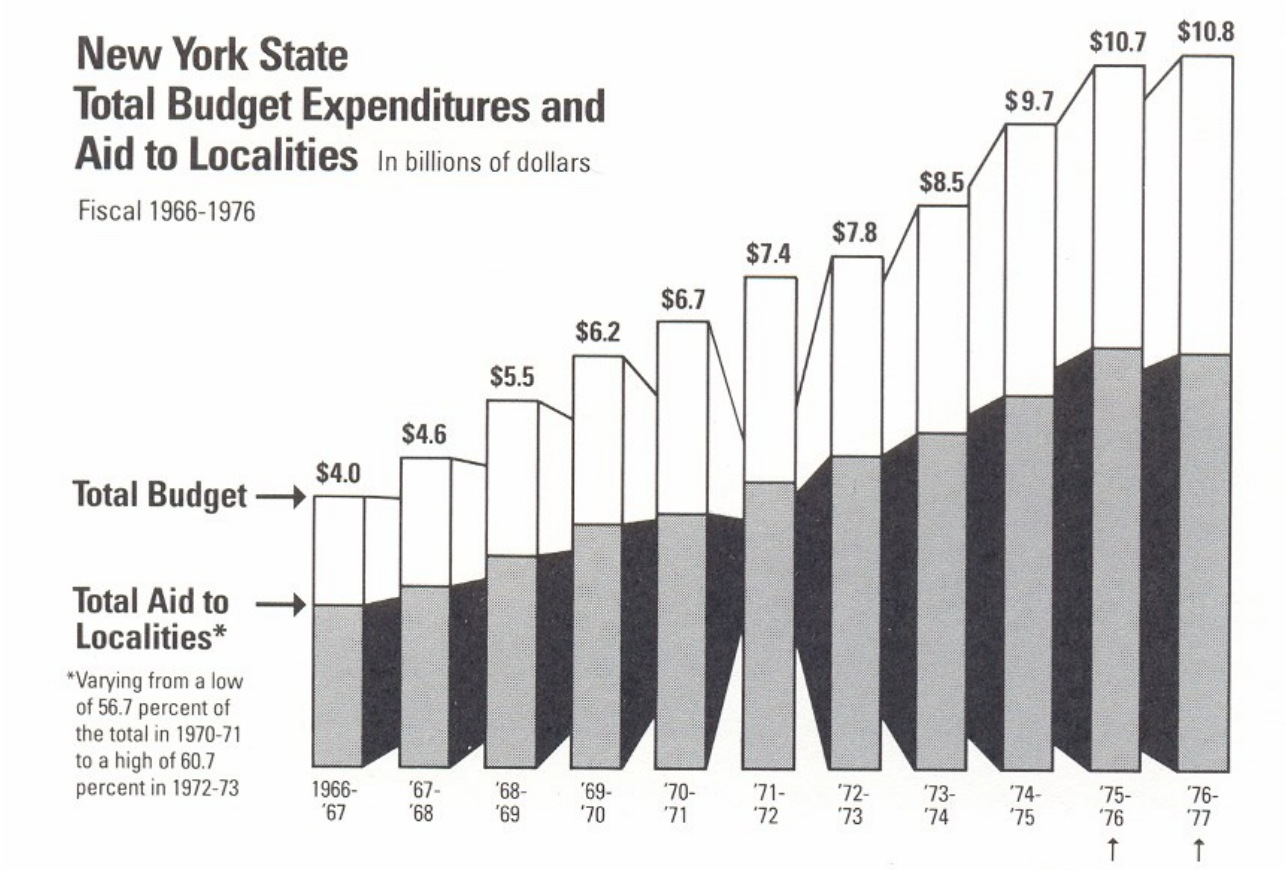
Lie Factor = 14.8

Legends have a constant size in one side and variable in the other



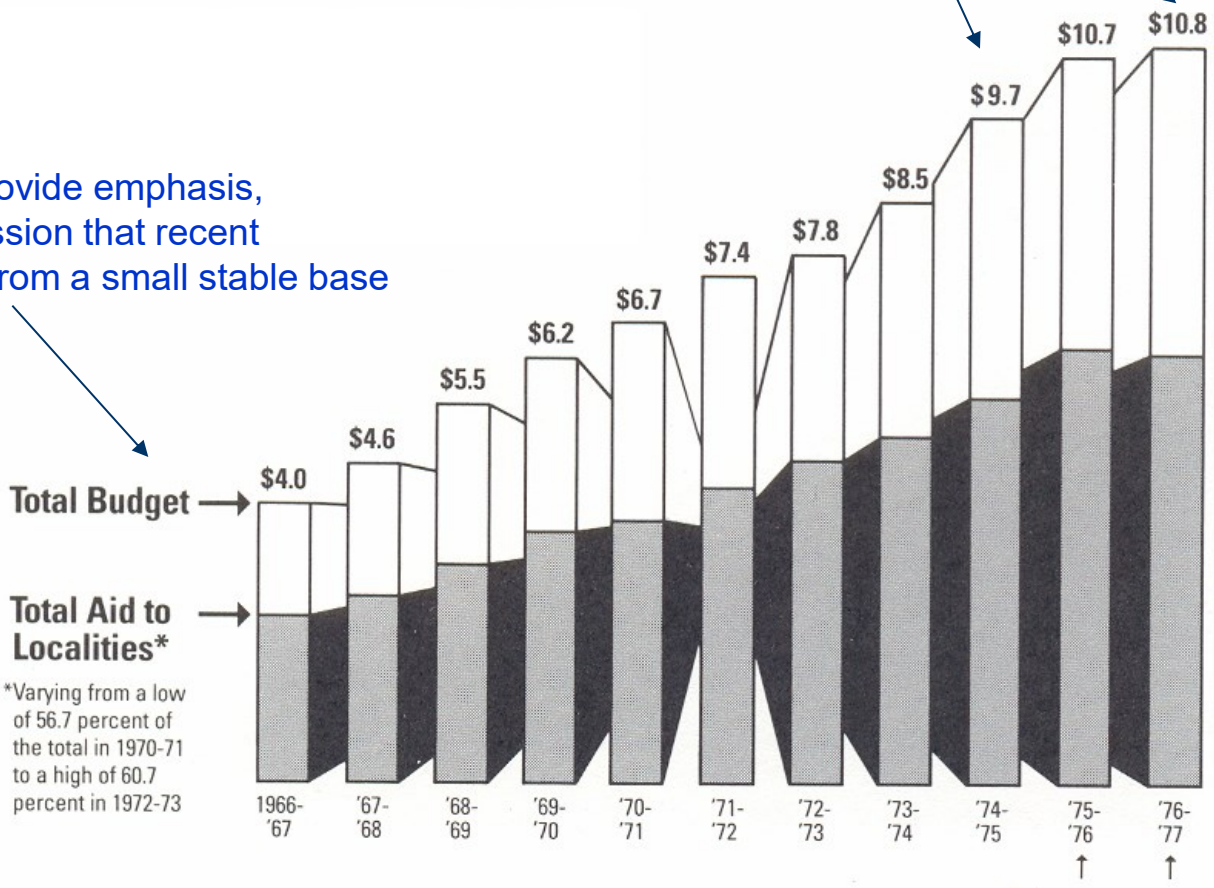
In roads, future usually lies in front, not behind

Another example:



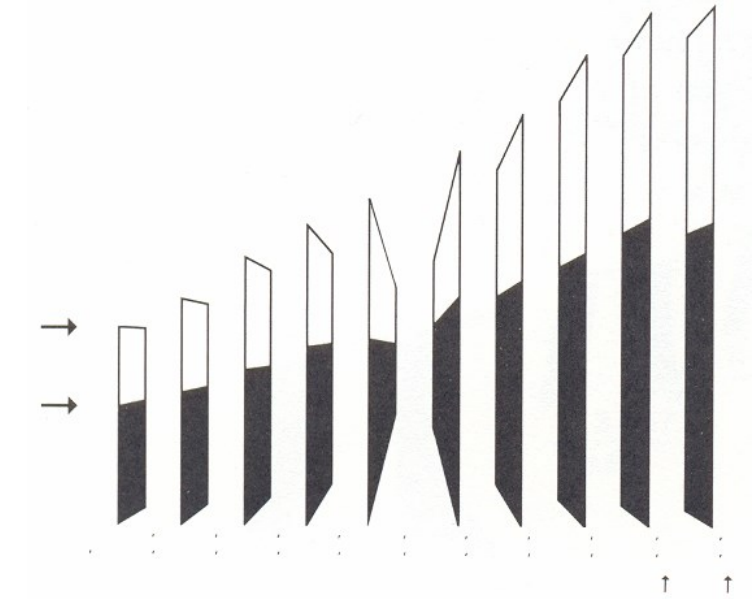
These three parallelipeds have been placed in na optical plane in front of the other eight, creating the image that the newer budgets tower over the older ones

Horizontal arrows provide emphasis, Encorage the impression that recent years have shot up from a small stable base

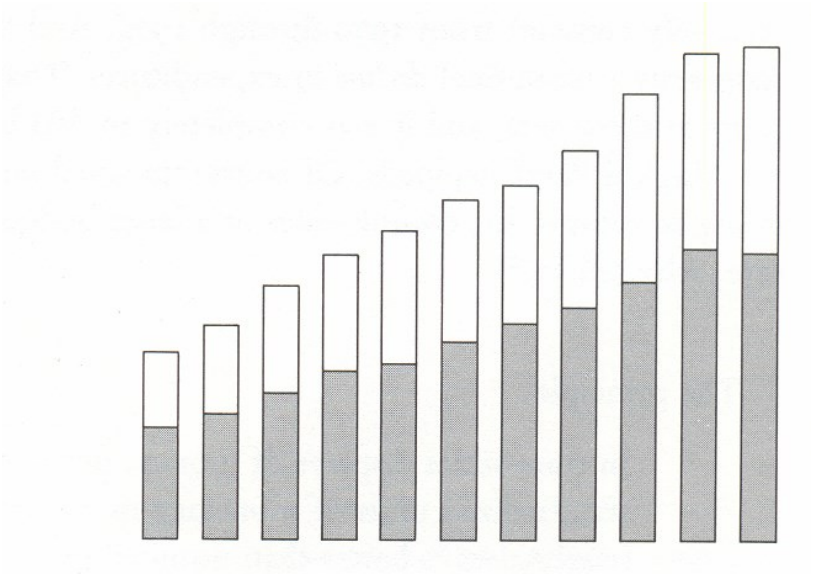


Arrows pointing straight up emphasize recent growth

Leaving behind the distortion



we have a calmer view:

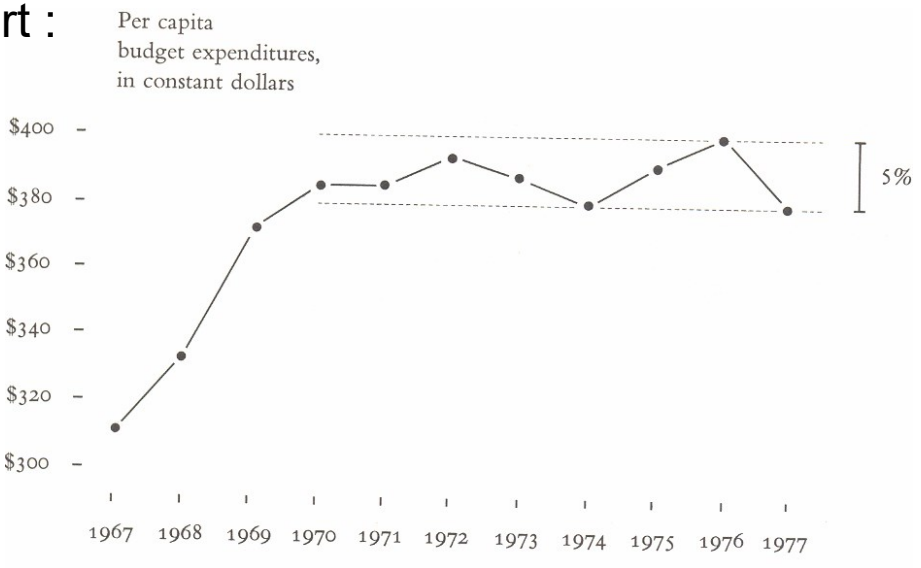


Two statistical lapses also bias the chart :

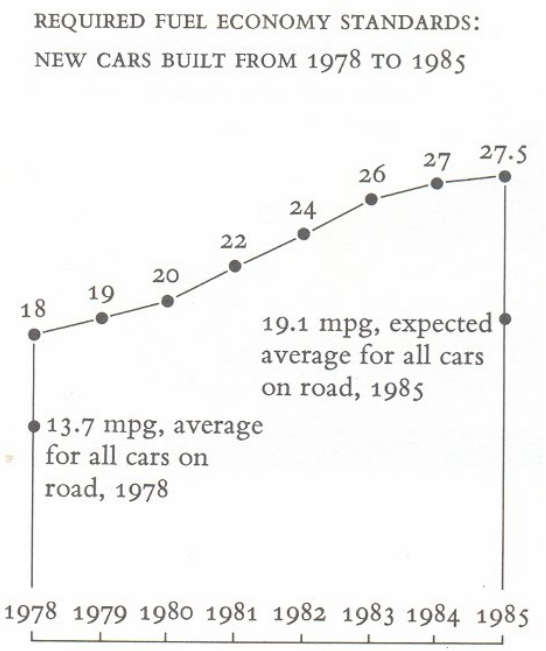
- Population increased10%

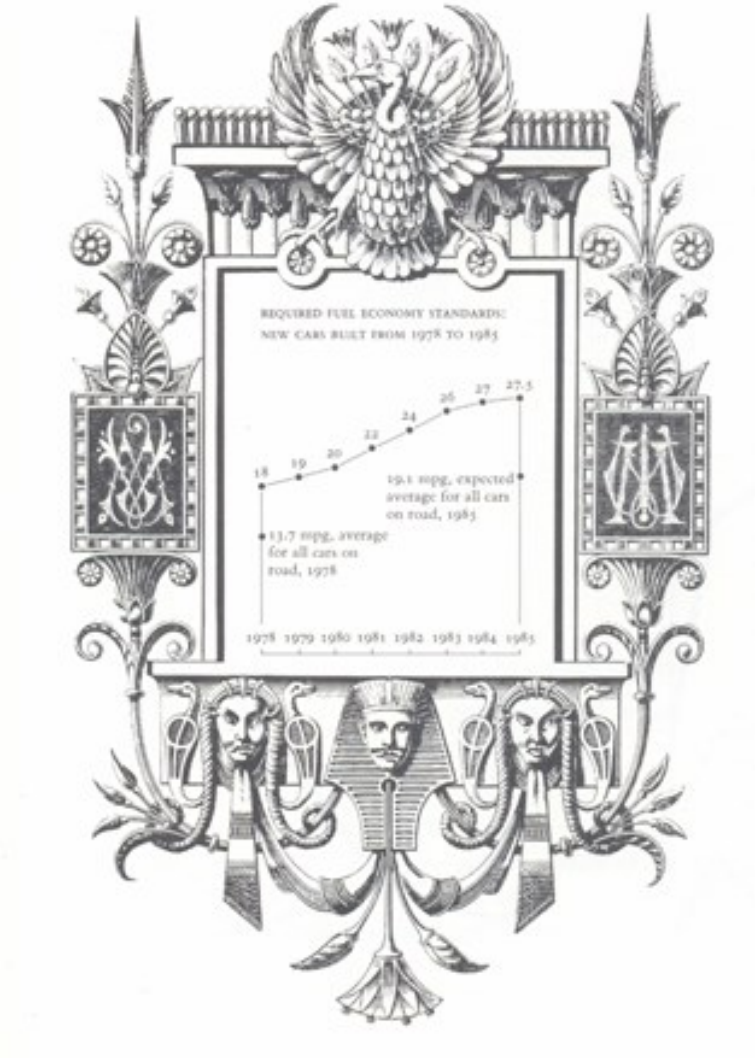
- there was substantial inflation

Final result →



Decorations without lies:





Challenger disaster (1986) investigation by R. Feynman

analysed by E. Tufte

Principles of graphical integrity:

- The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented
- Clear, detailed and thorough labelling should be used to defeat graphical distortion and ambiguity
- Show data variation, not design variation
- In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units
- The number of information carrying (variable) dimensions depicted should not exceed the number of dimensions in the data
- Graphics must not quote data out of context

Main Bibliography

- Tufte, E., *Visual Display of Quantitative Information*, Graphics Press, 1983