Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Homework
- Visualization: Tufte
- Coming up…Preparing data, more viz, asking questions
- HW for Monday

Visualizing linear relationships

```
%matplotlib inline
```

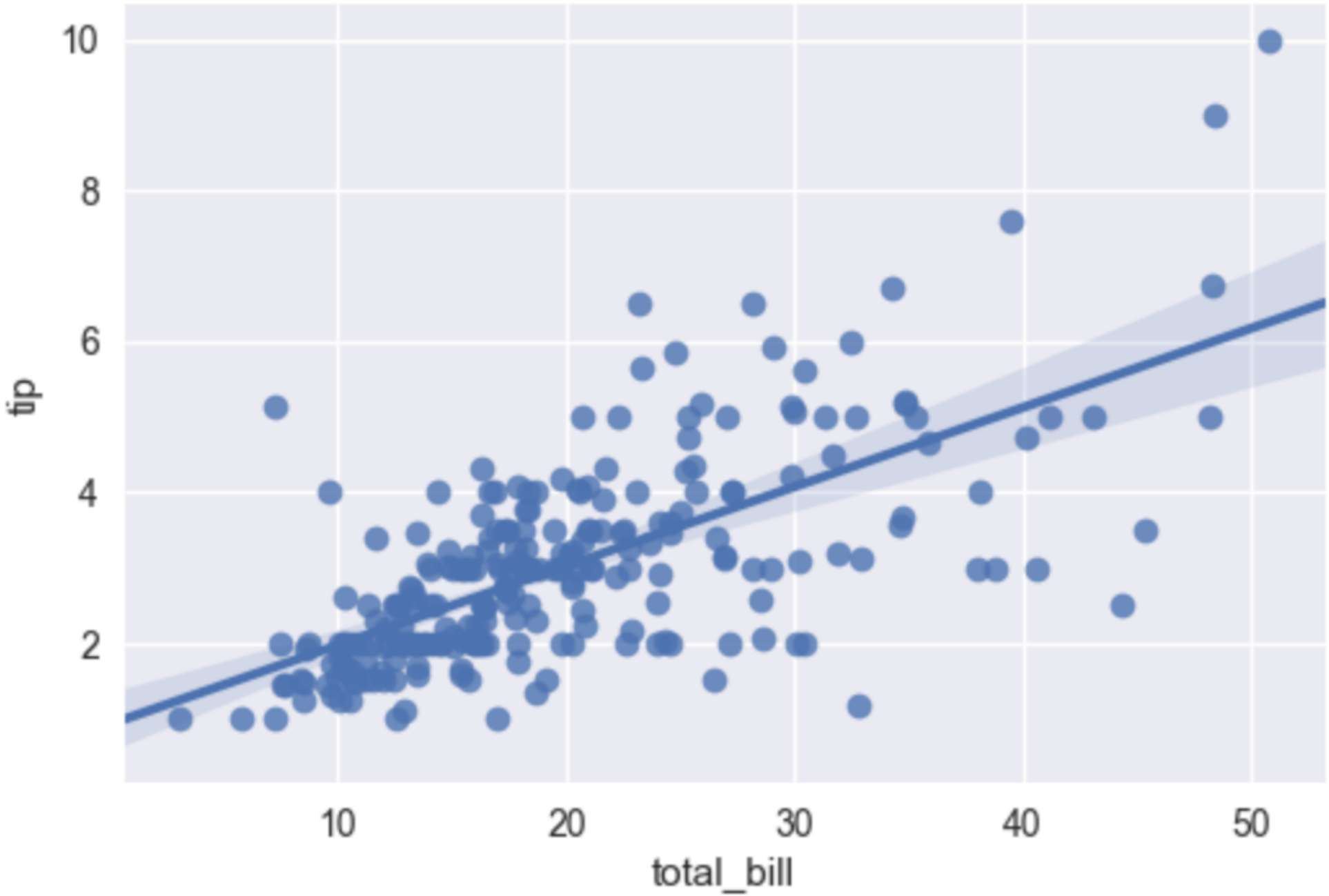
```
import numpy as np
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
```

```
import seaborn as sns
sns.set(color_codes=True)
```

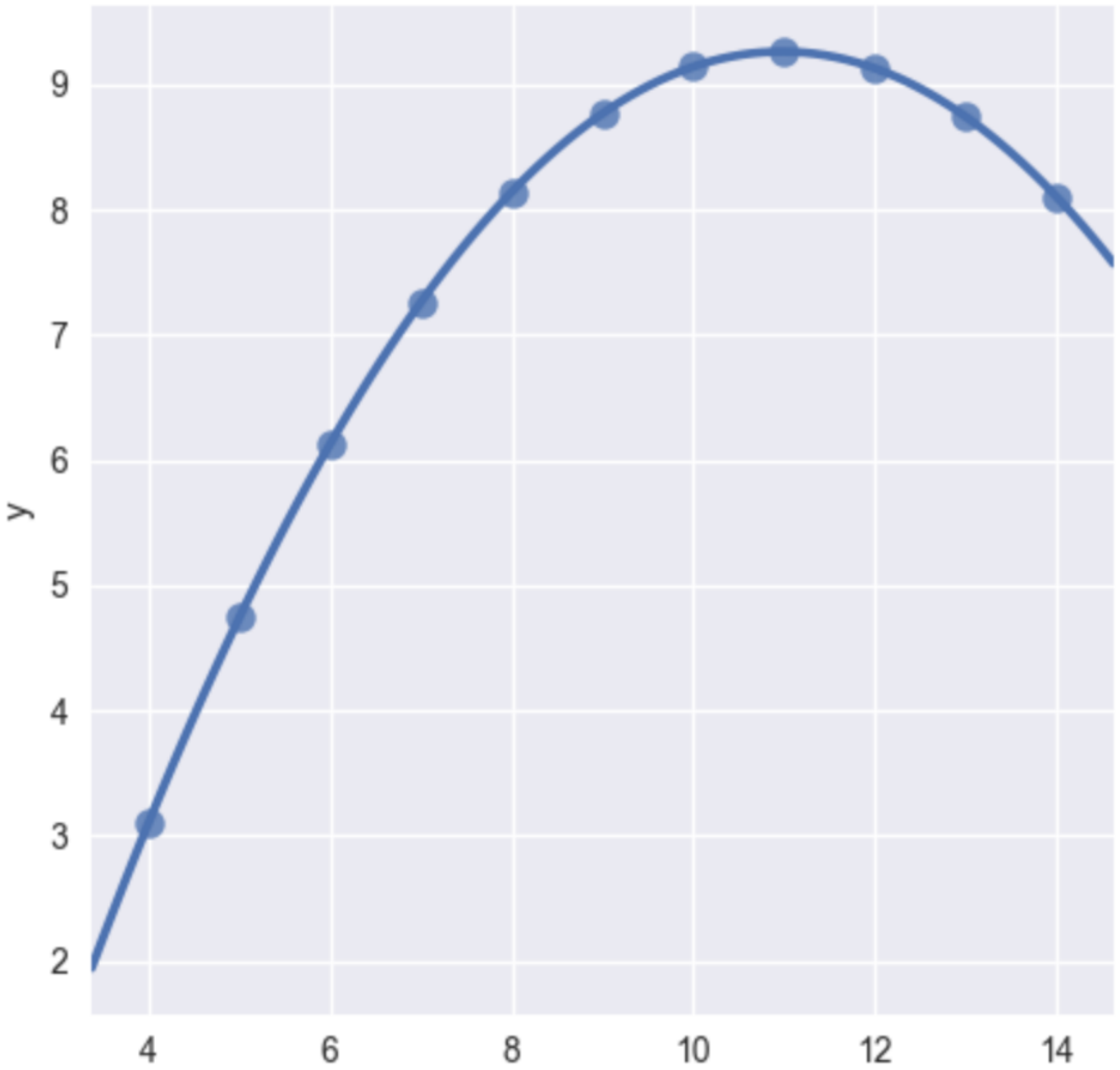
```
np.random.seed(sum(map(ord, "regression")))
```

```
tips = sns.load_dataset("tips")
```

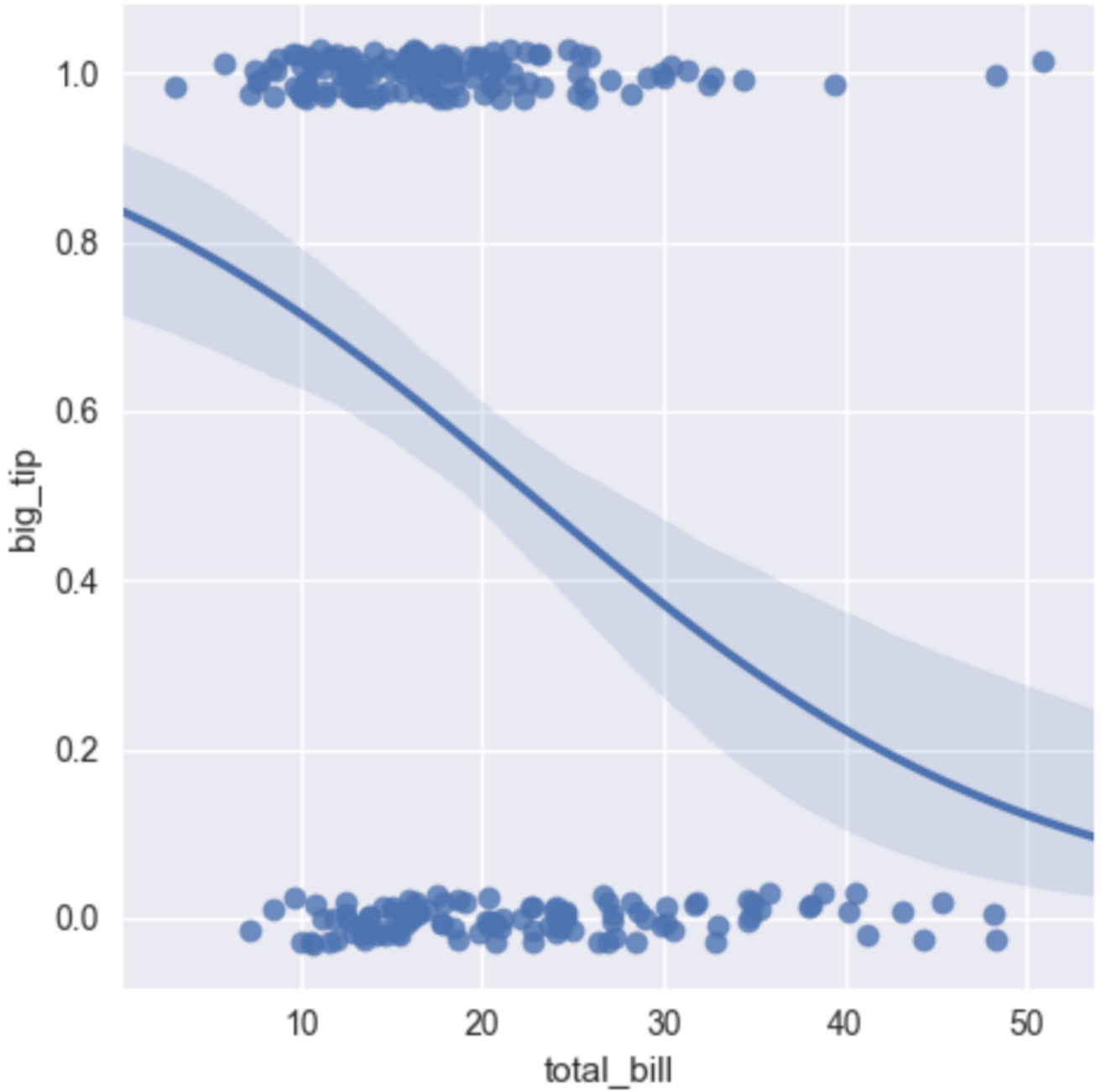
```
sns.regplot(x="total_bill", y="tip", data=tips);
```



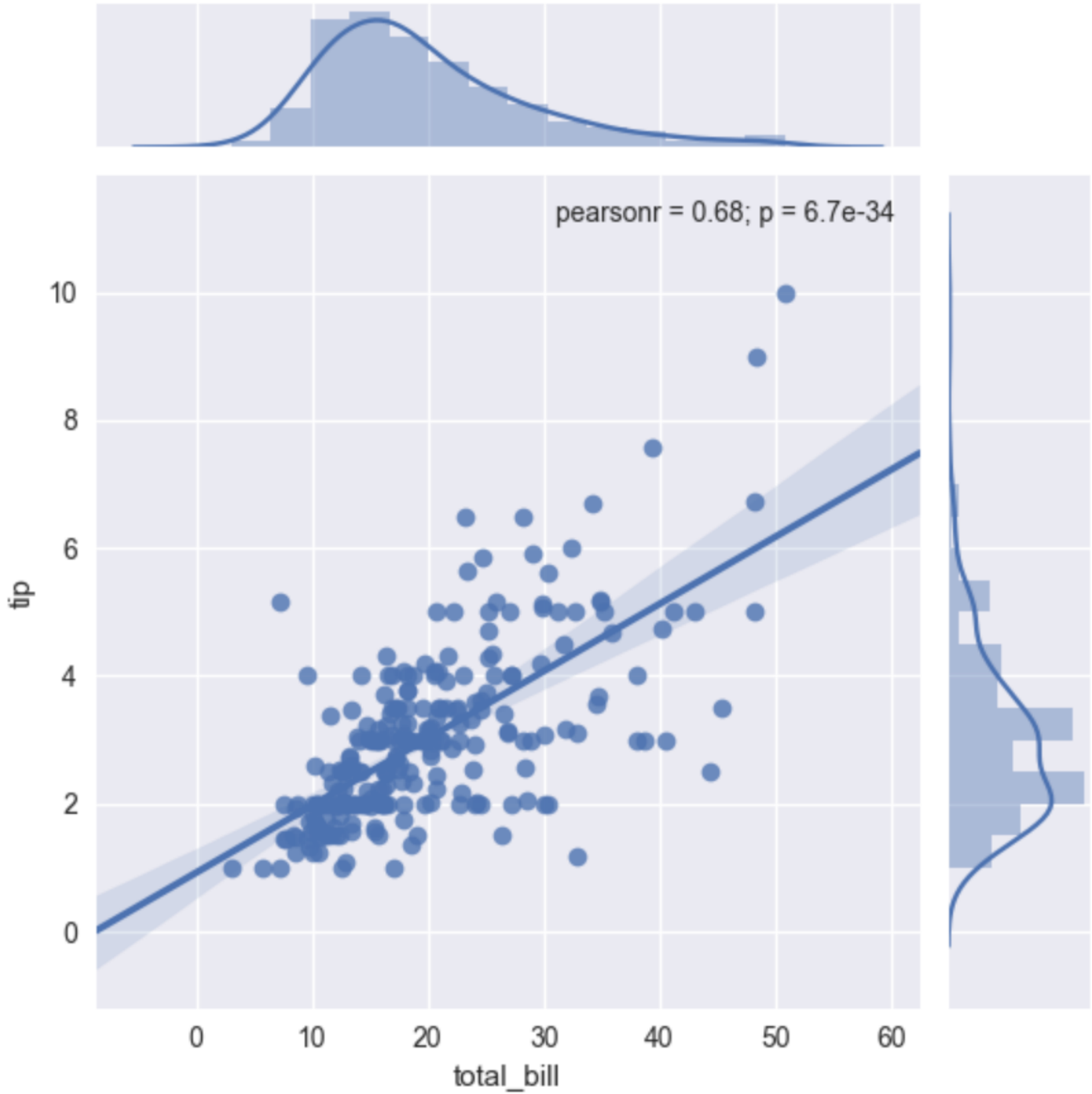
```
sns.lmplot(x="x", y="y", data=anscombe.query("dataset == 'II'"),
           order=2, ci=None, scatter_kws={"s": 80});
```



```
sns.lmplot(x="total_bill", y="big_tip", data=tips,
           logistic=True, y_jitter=.03);
```



```
sns.jointplot(x="total_bill", y="tip", data=tips, kind="reg");
```



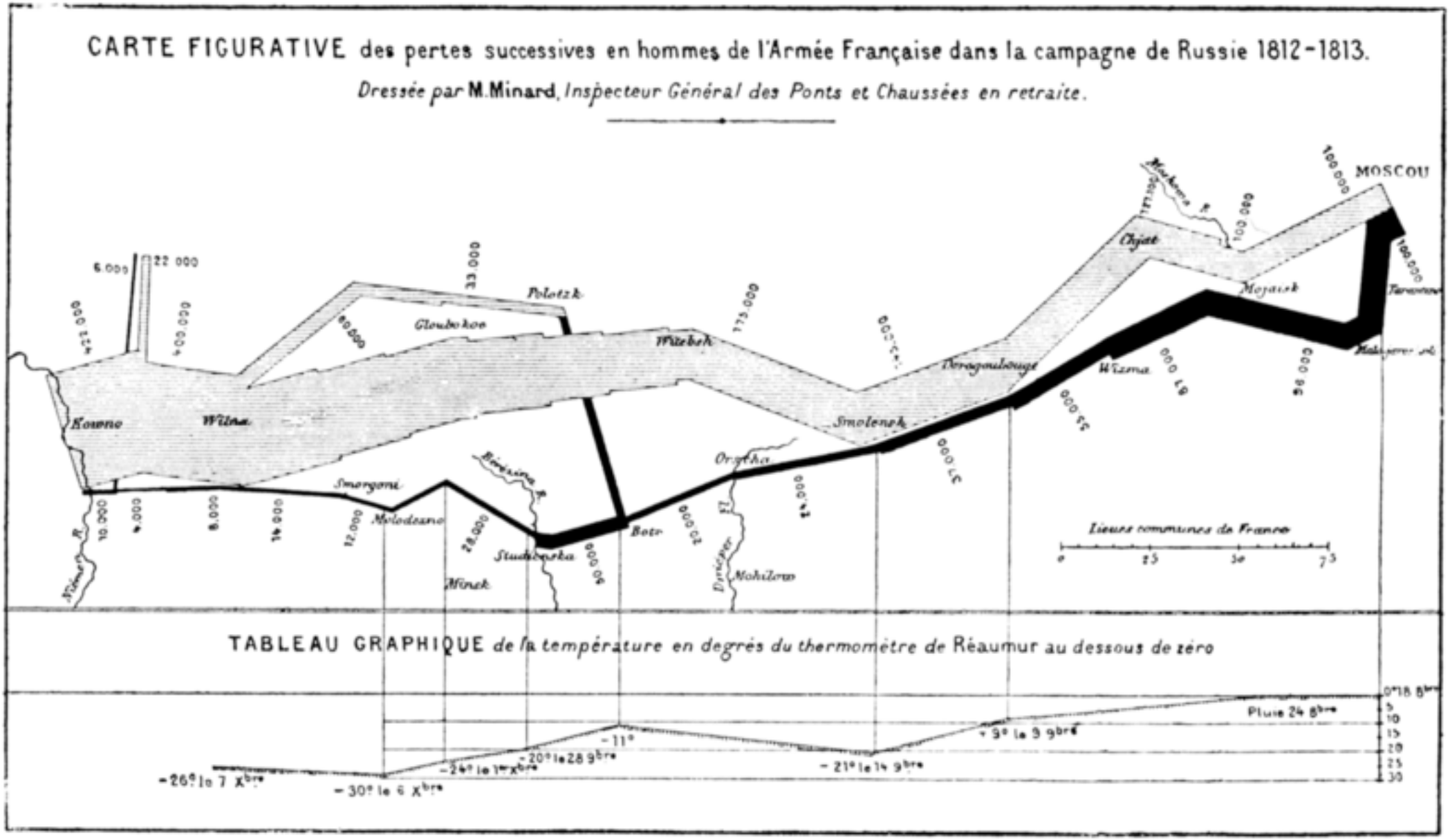
Visualization tools for python

- Seaborn
- Matplotlib
- Bokeh

Principles of visualization

the art of graphing: presenting data clearly

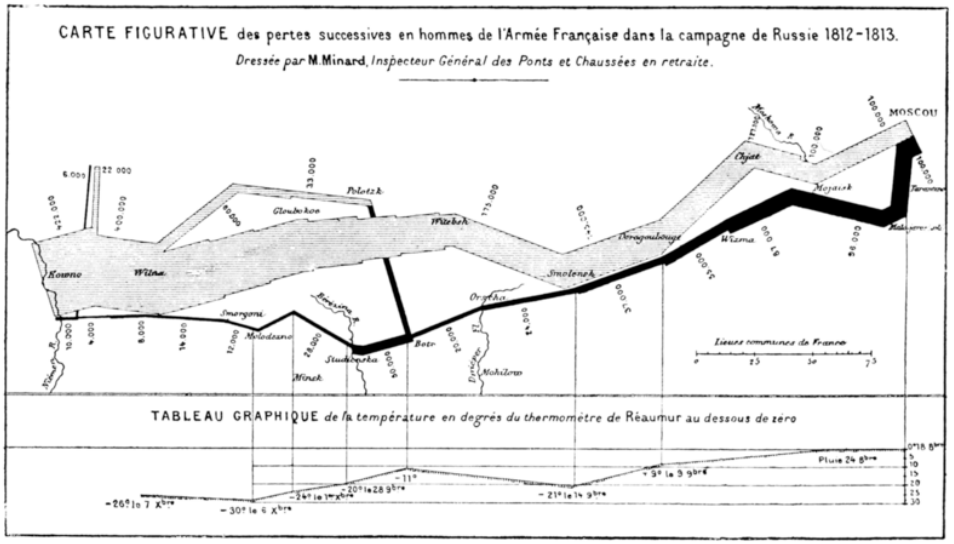
The best statistical graphic ever?



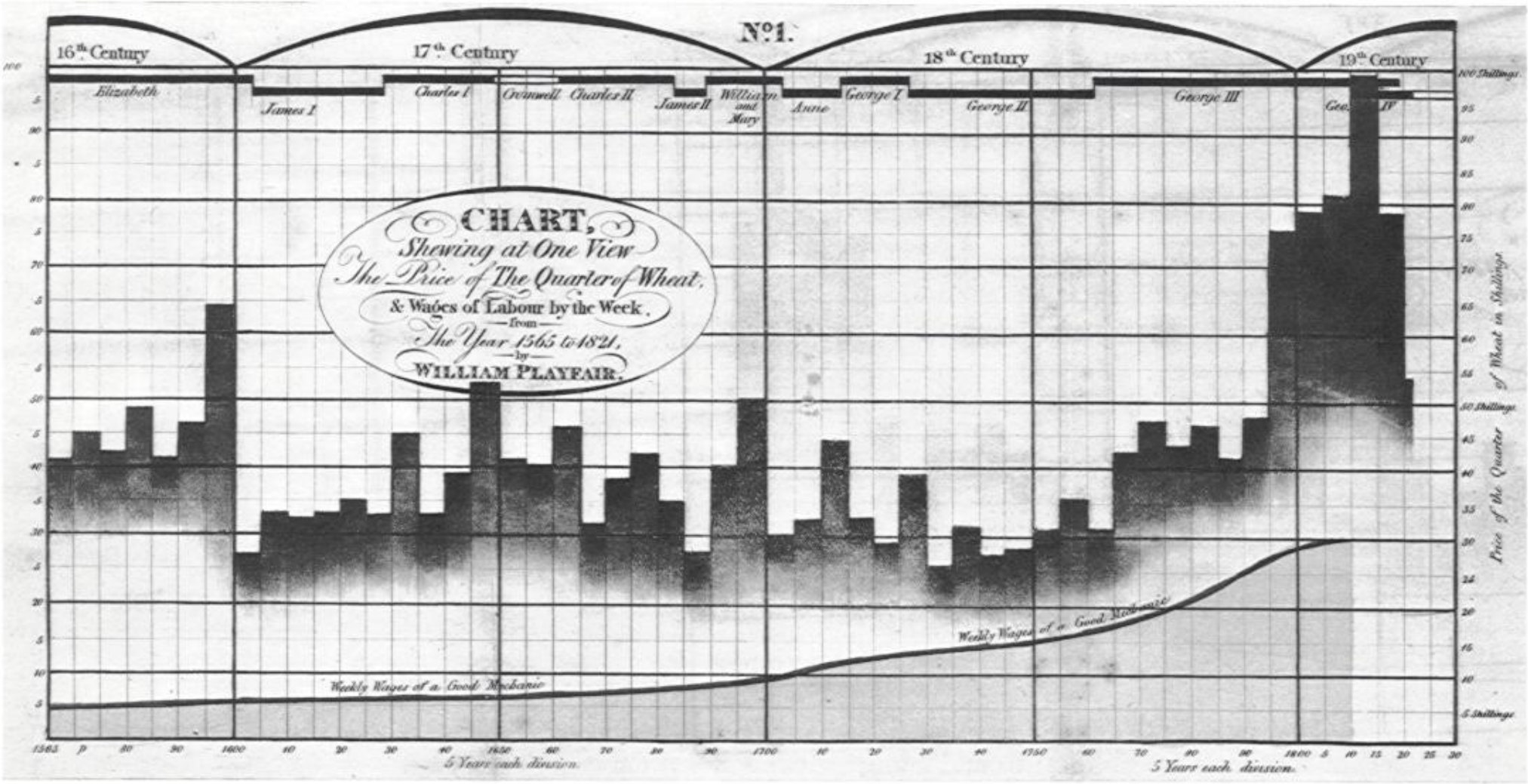
the art of graphing: presenting data clearly

The best statistical graphic ever?

- 6 dimensions of information:
  - Time, size of army, temperature, space (2D), direction of march
- All presented clearly individually, and in a way that it is easy to grasp how they relate.

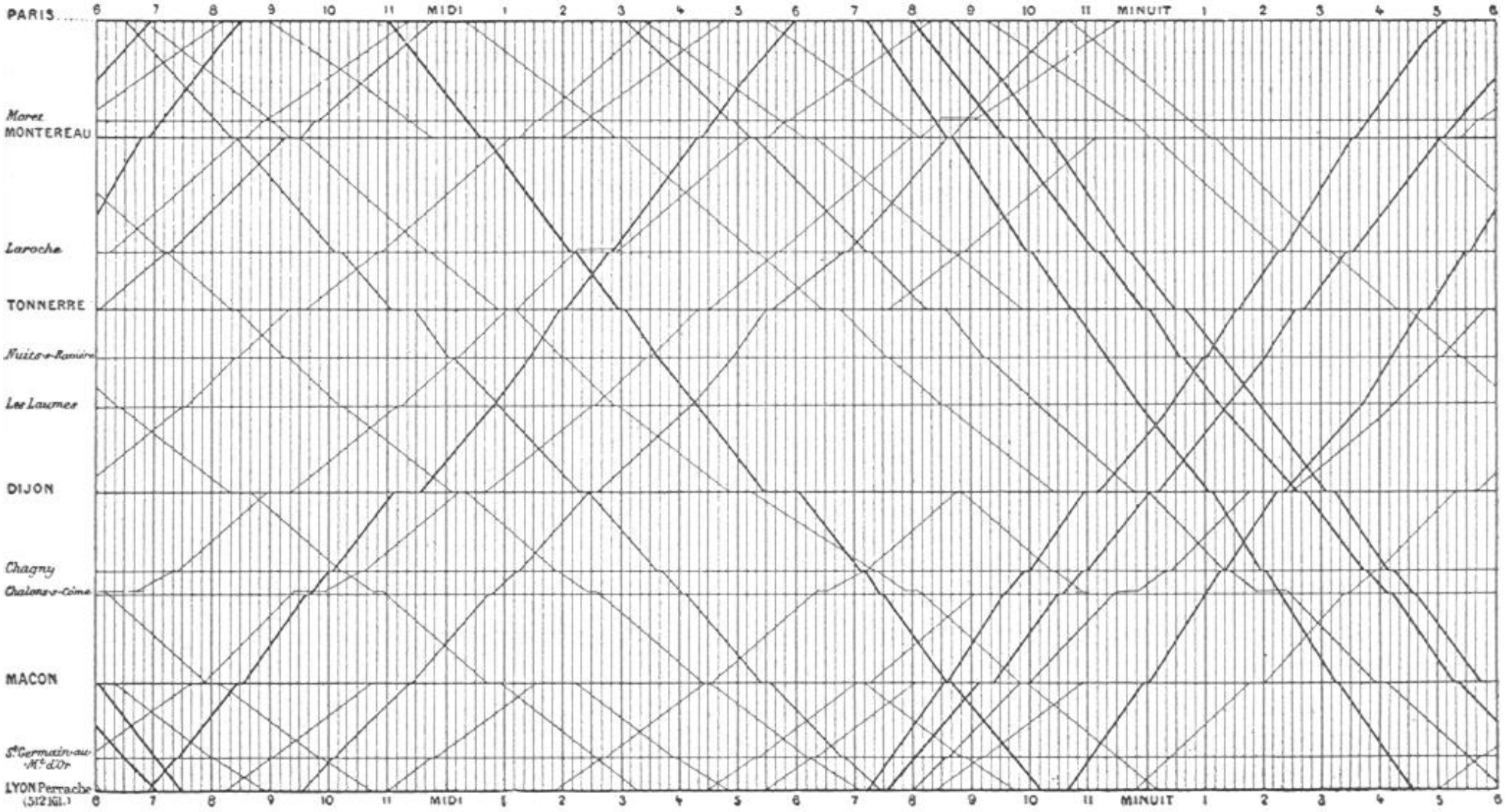


the art of graphing: presenting data clearly



The price of wheat compared to labour wages, William Playfair (1759-1823)

the art of graphing: presenting data clearly

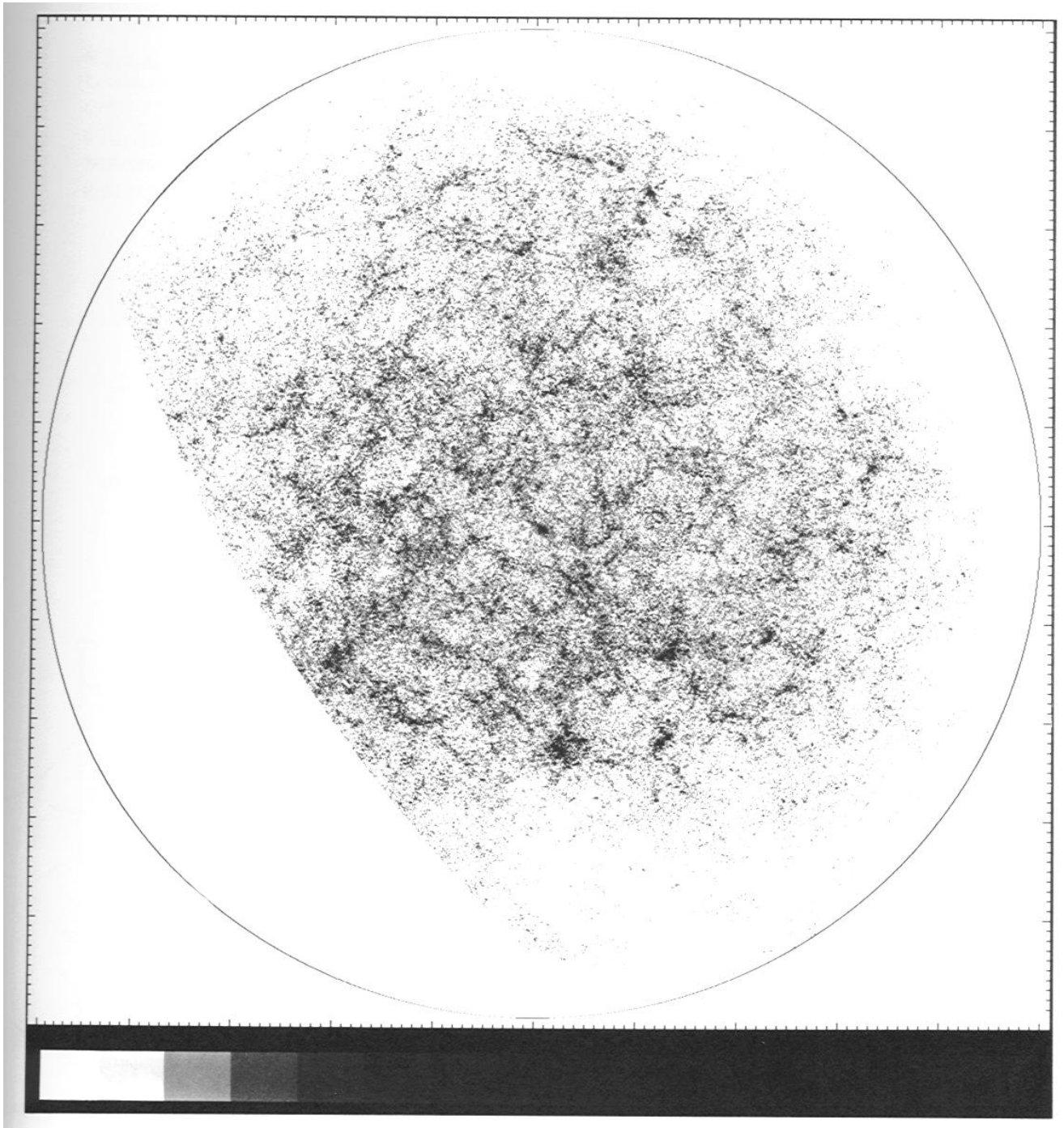


French train schedule, as drawn by E.J. Marey (1830-1904)

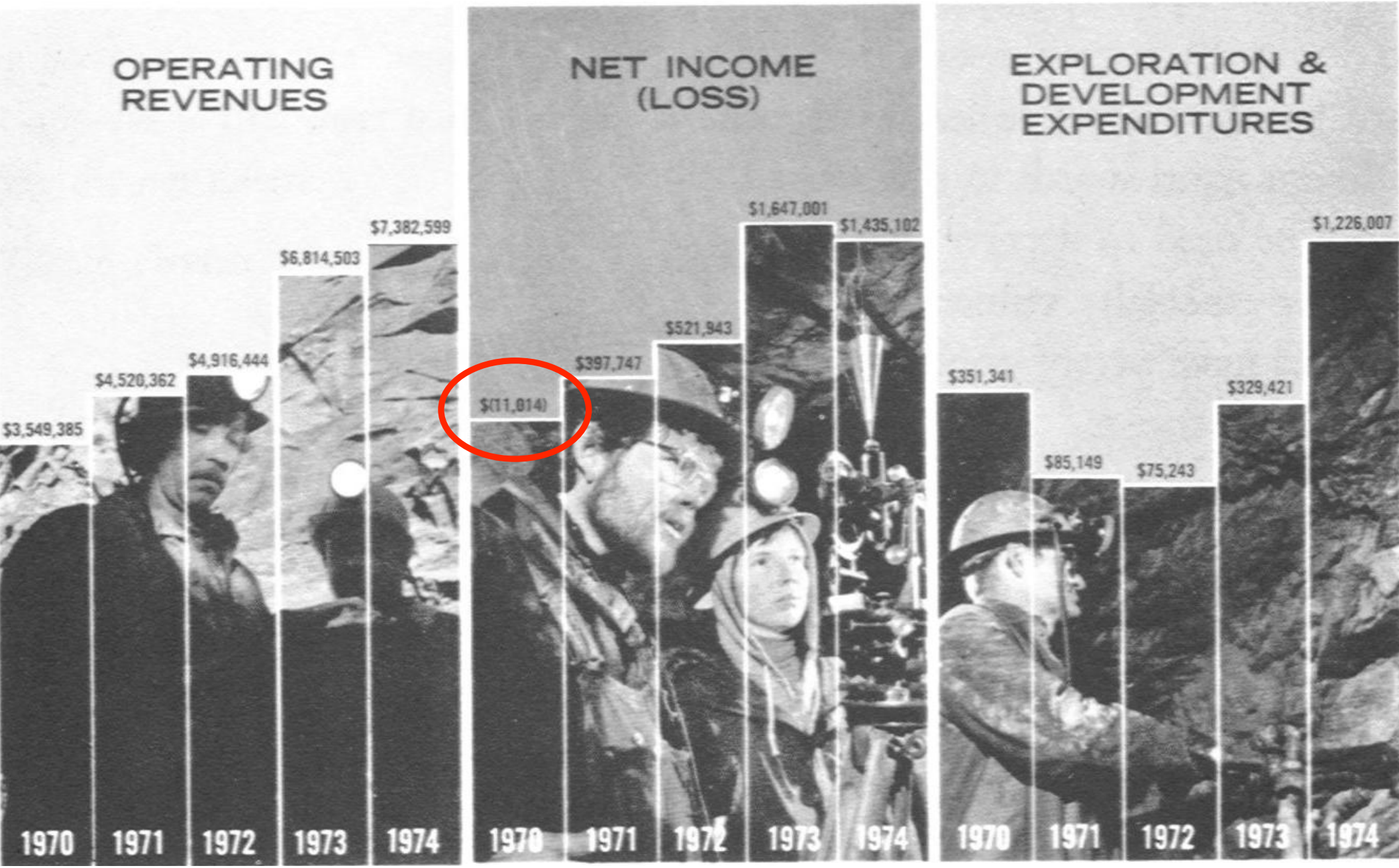
the art of graphing: presenting data clearly

Map of the northern galactic hemisphere

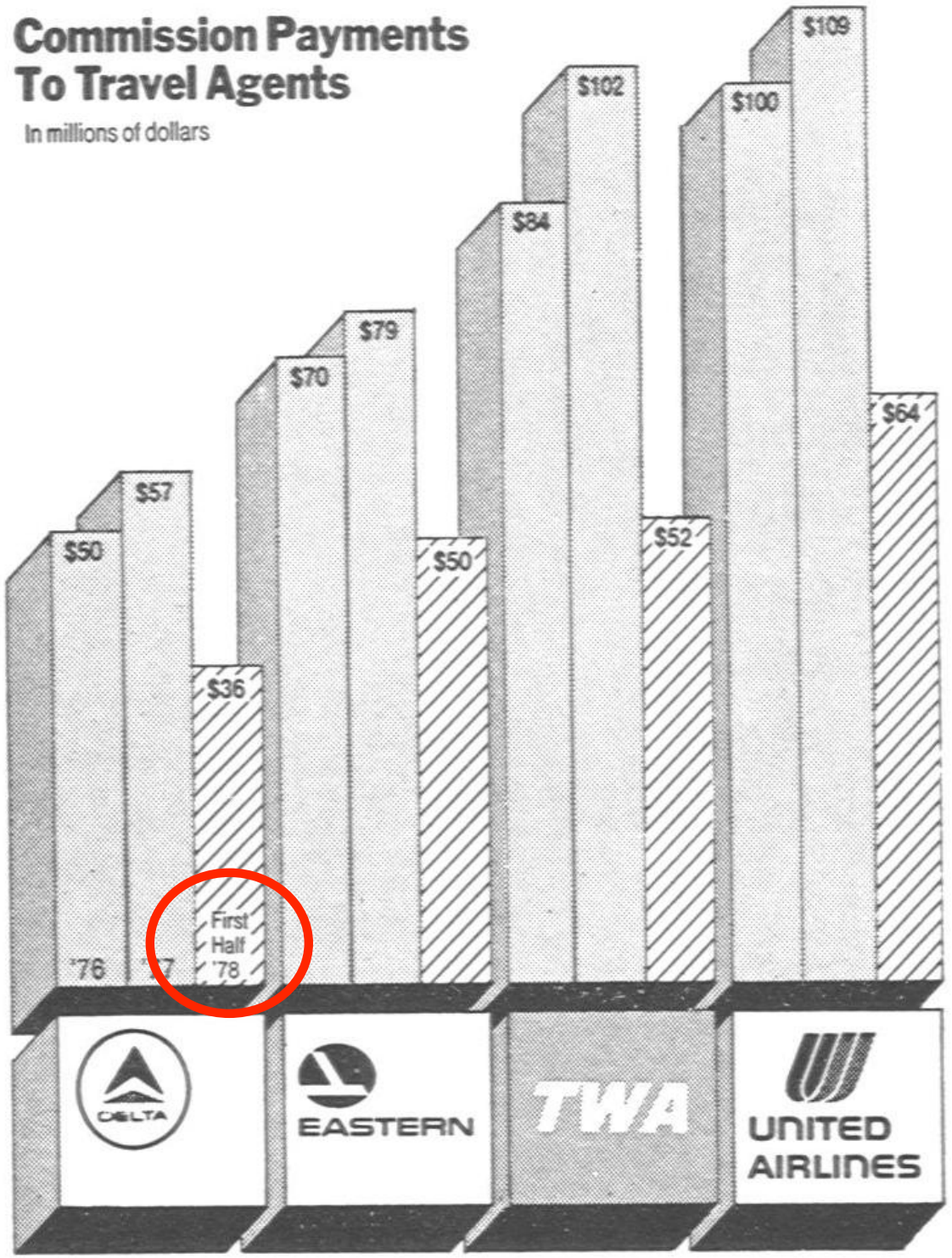
(1.3 million galaxies shown)



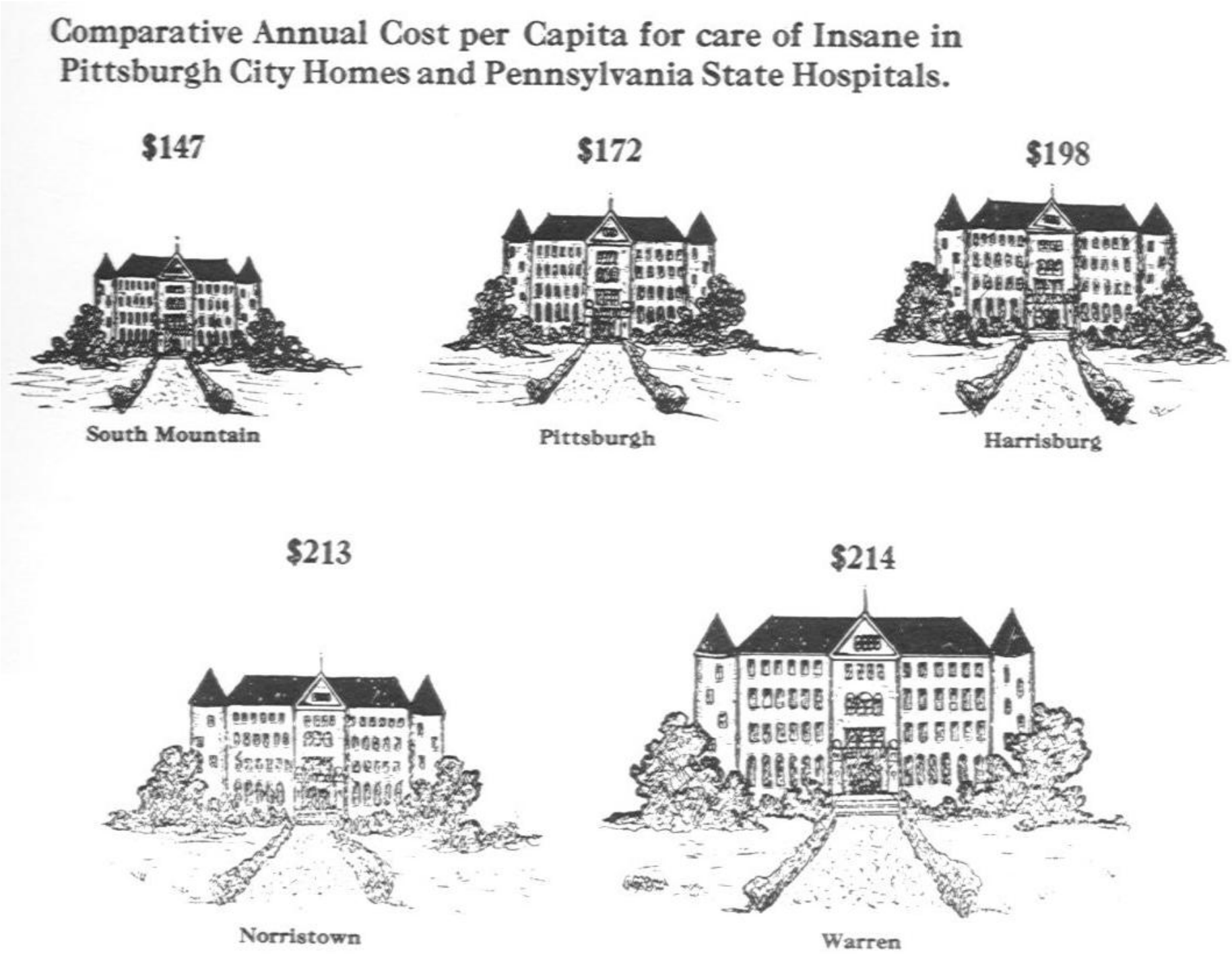
bad graphics!



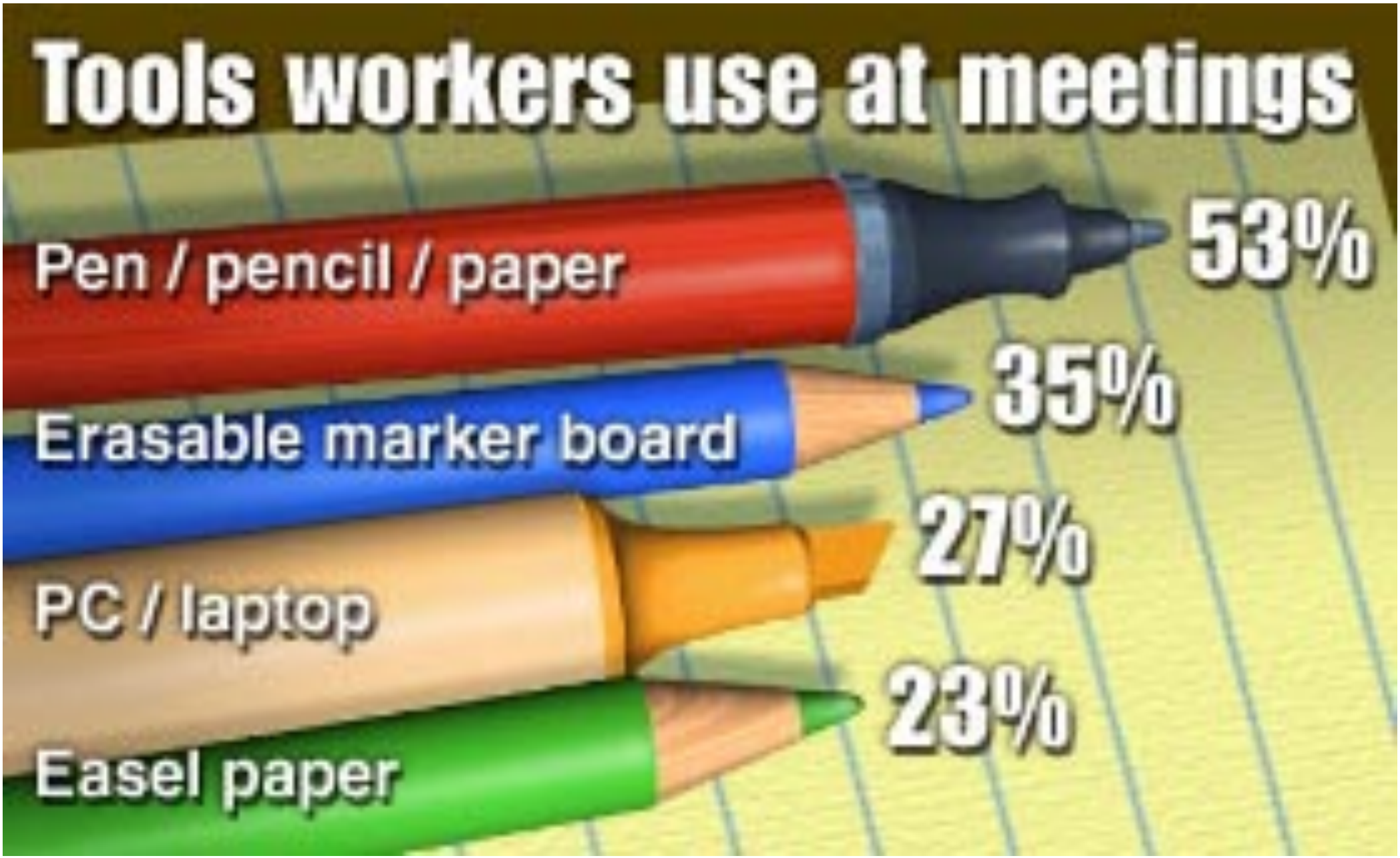
bad graphics!



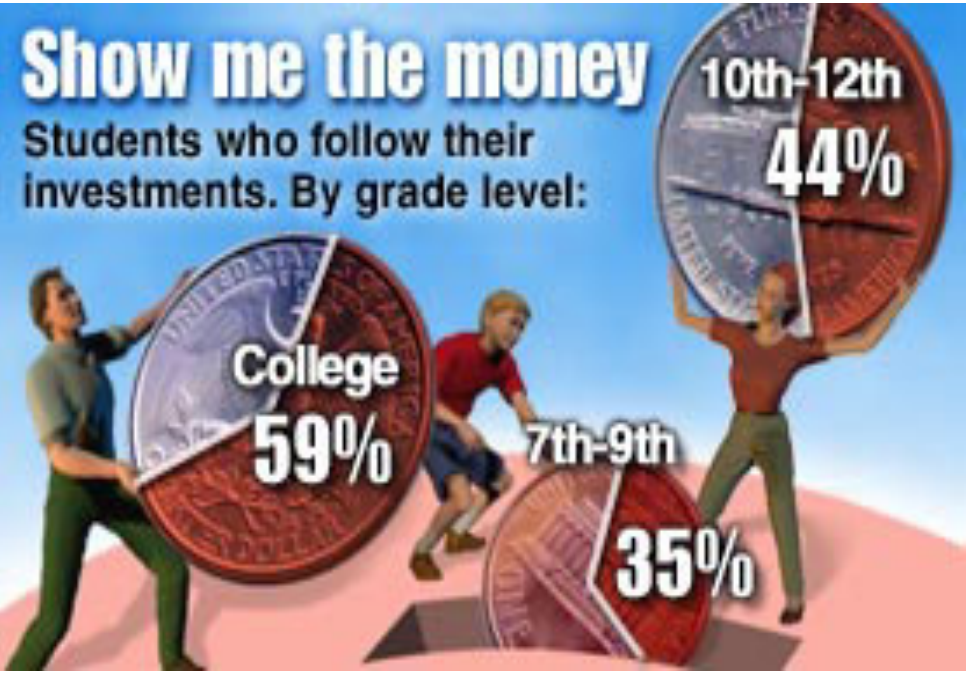
bad graphics!

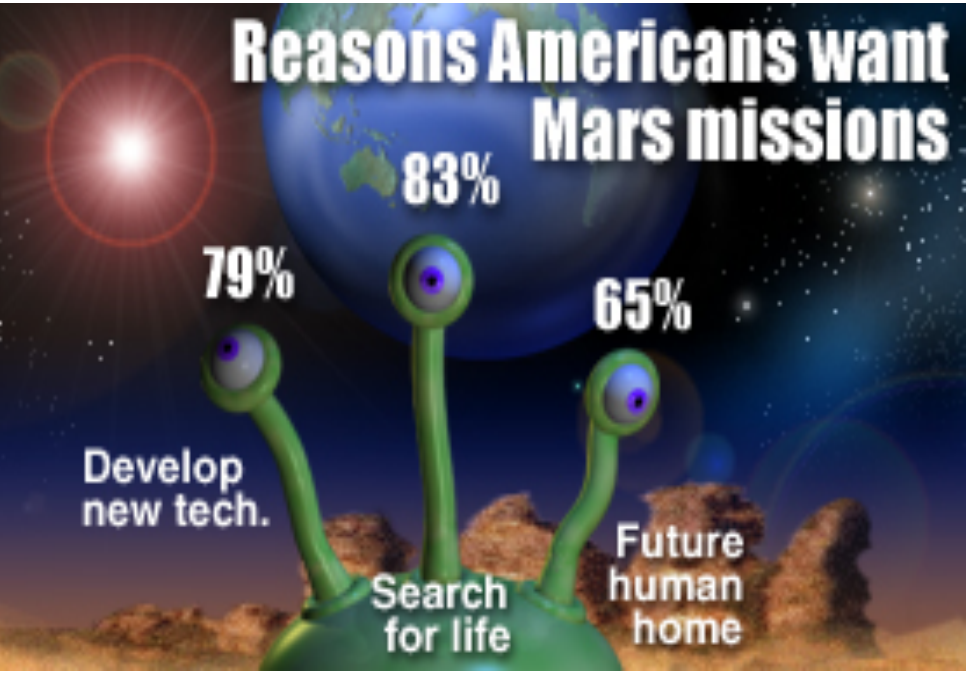


bad graphics!

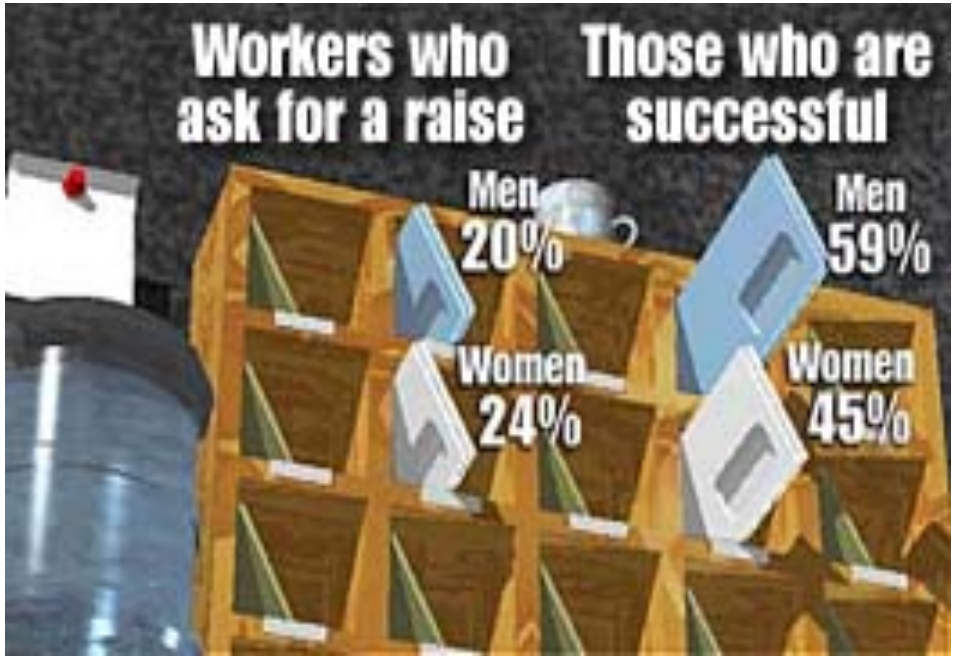


bad graphics!

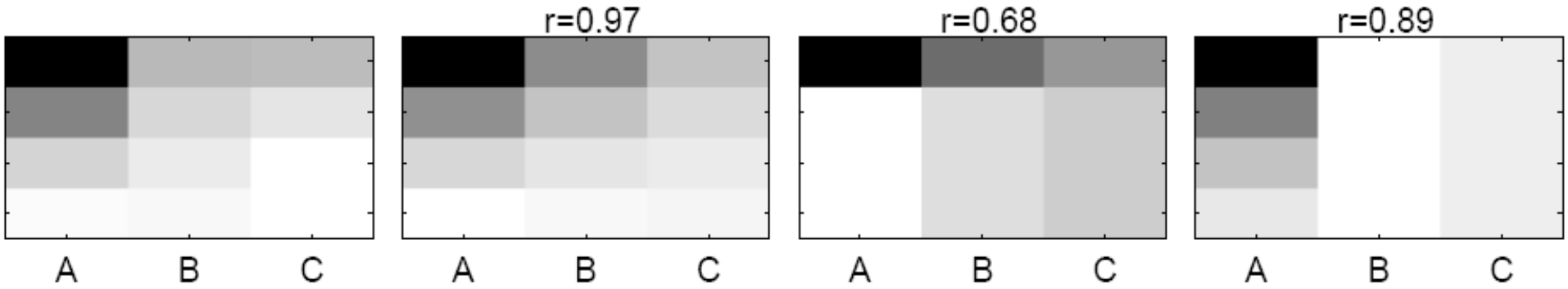


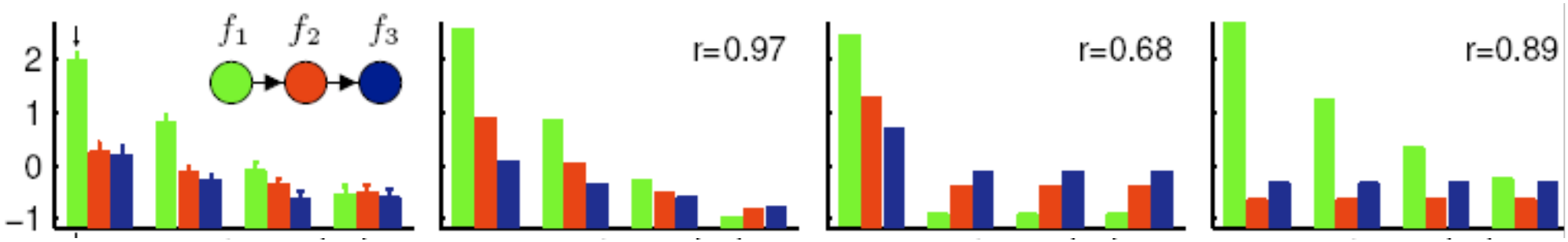


bad graphics!



the art of graphing: presenting data clearly

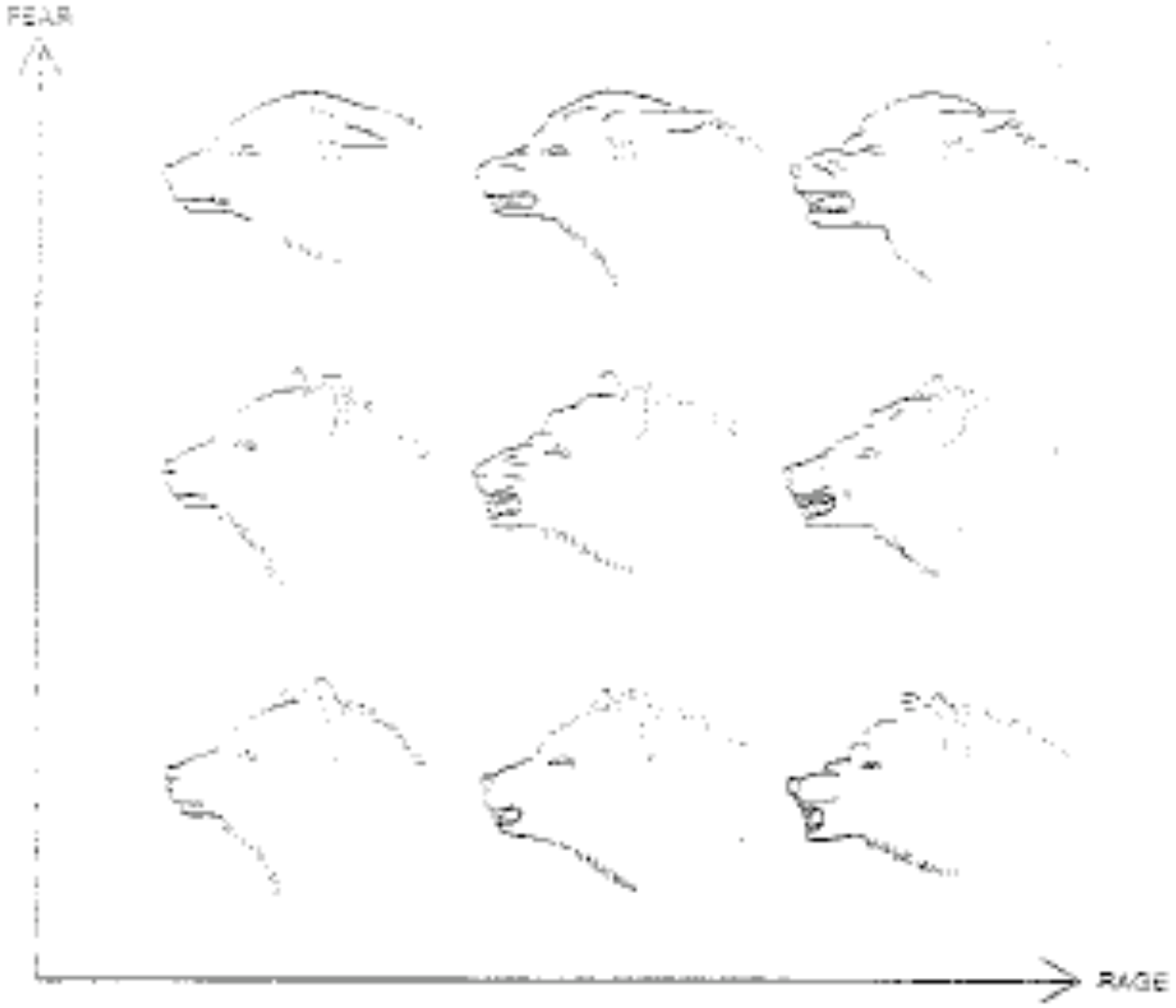




the art of graphing: presenting data clearly

A scatterplot with data points that themselves contain data!

- The pictures here contain more information than could be described otherwise
- Pictures are arranged in an order that shows how they relate



the art of graphing: presenting data clearly

$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{Total ink used to print graphic}}$$

= Proportion of a graphic's ink devoted to the non-redundant display of data-information.

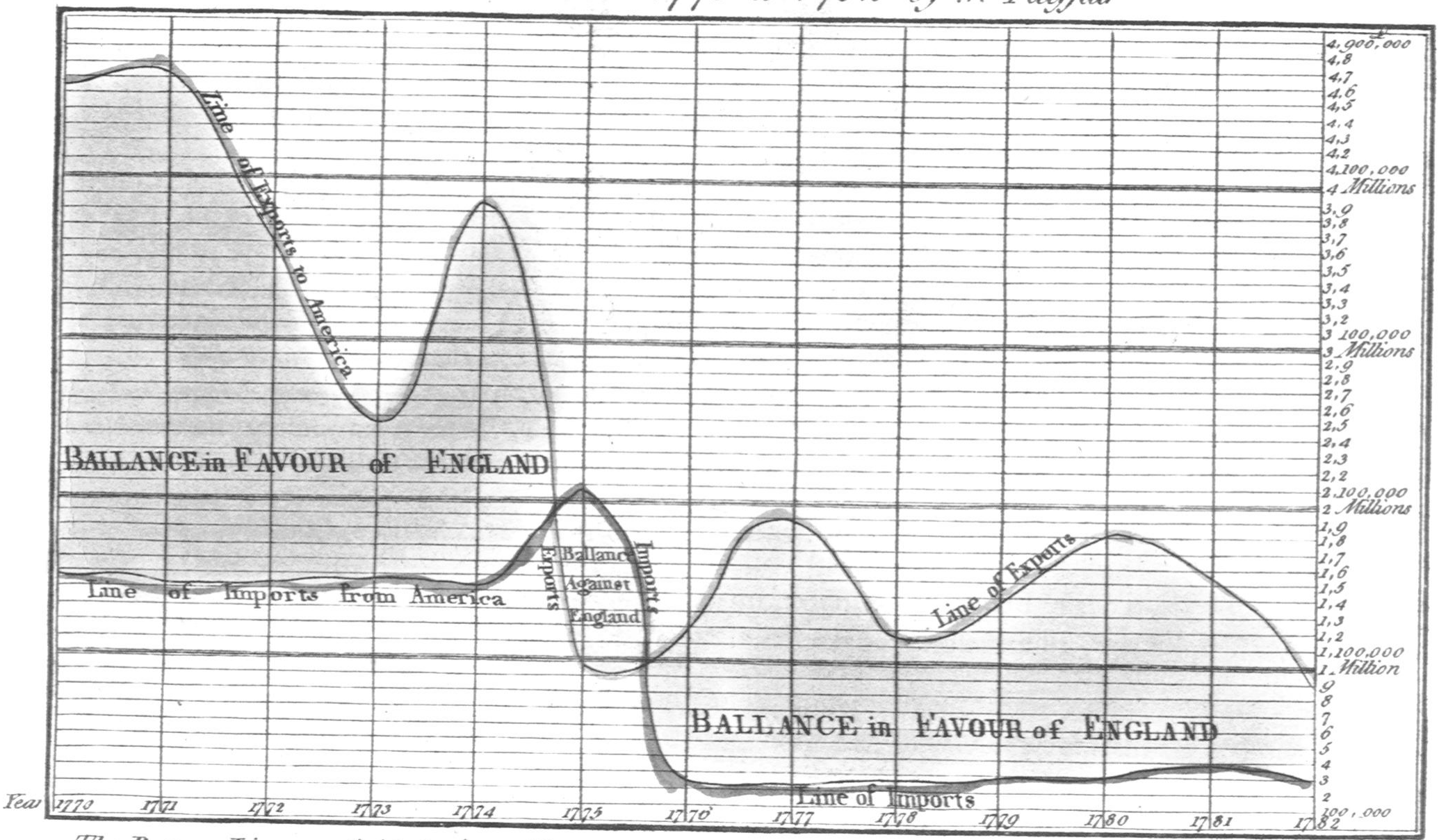
= 1.0 – proportion of graphic that can be *erased* without the loss of information

the art of graphing: presenting data clearly

Tufte presents some principles of data graphics

- Above all else, show the data.
- Maximize the data-ink ratio
- Erase non-data-ink
- Erase redundant data-ink
- Revise and edit

CHART of IMPORTS and EXPORTS of ENGLAND to and from all NORTH AMERICA
From the Year 1770 to 1782 by W. Playfair

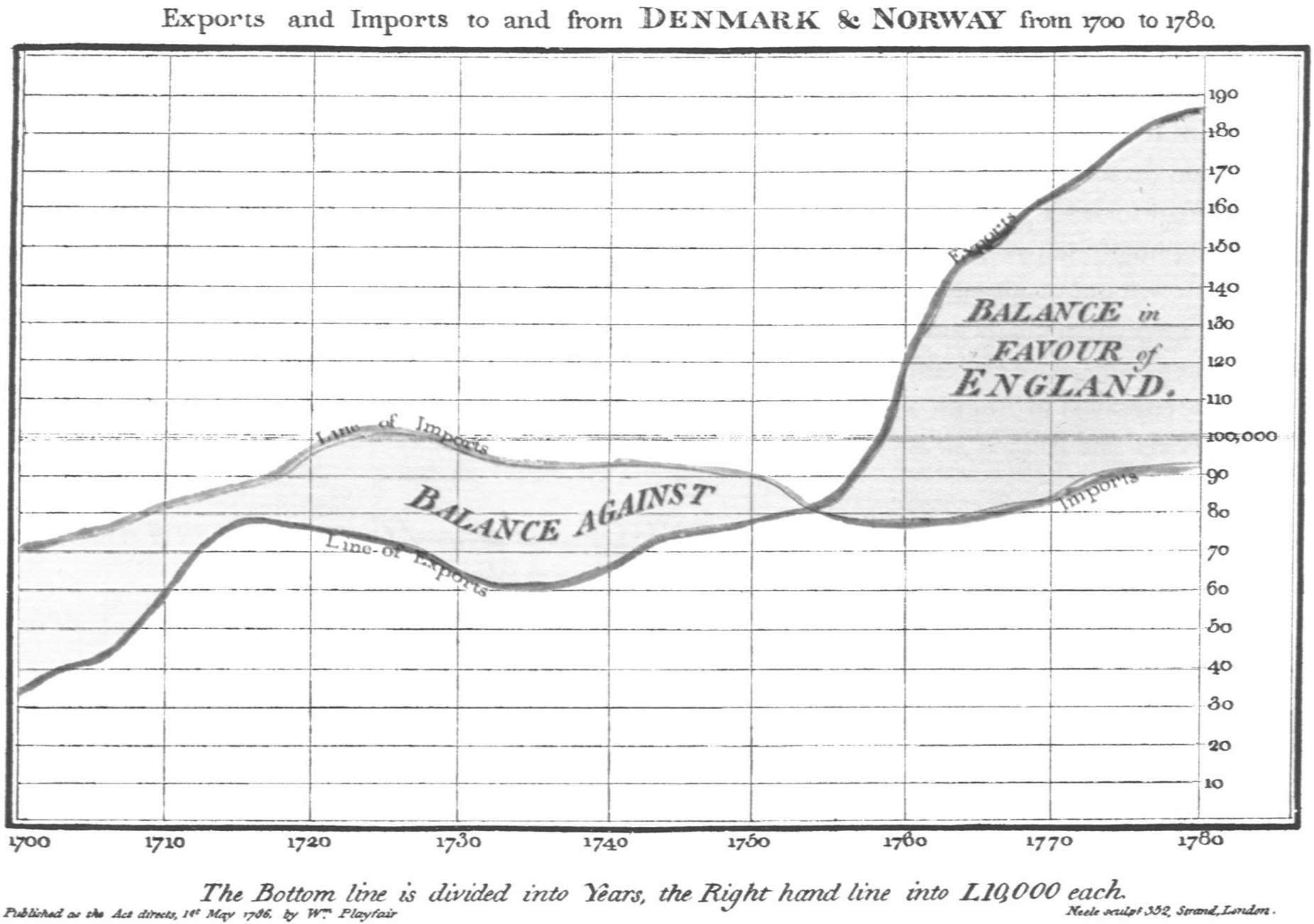


The Bottom Line is divided into Years the right-hand Line into HUNDRED THOUSAND POUNDS

J. Ainslie Sculp.t

Published as the Act directs 20th April 1786

the art of graphing: presenting data clearly



the art of graphing: presenting data clearly

Bad graphing practices:

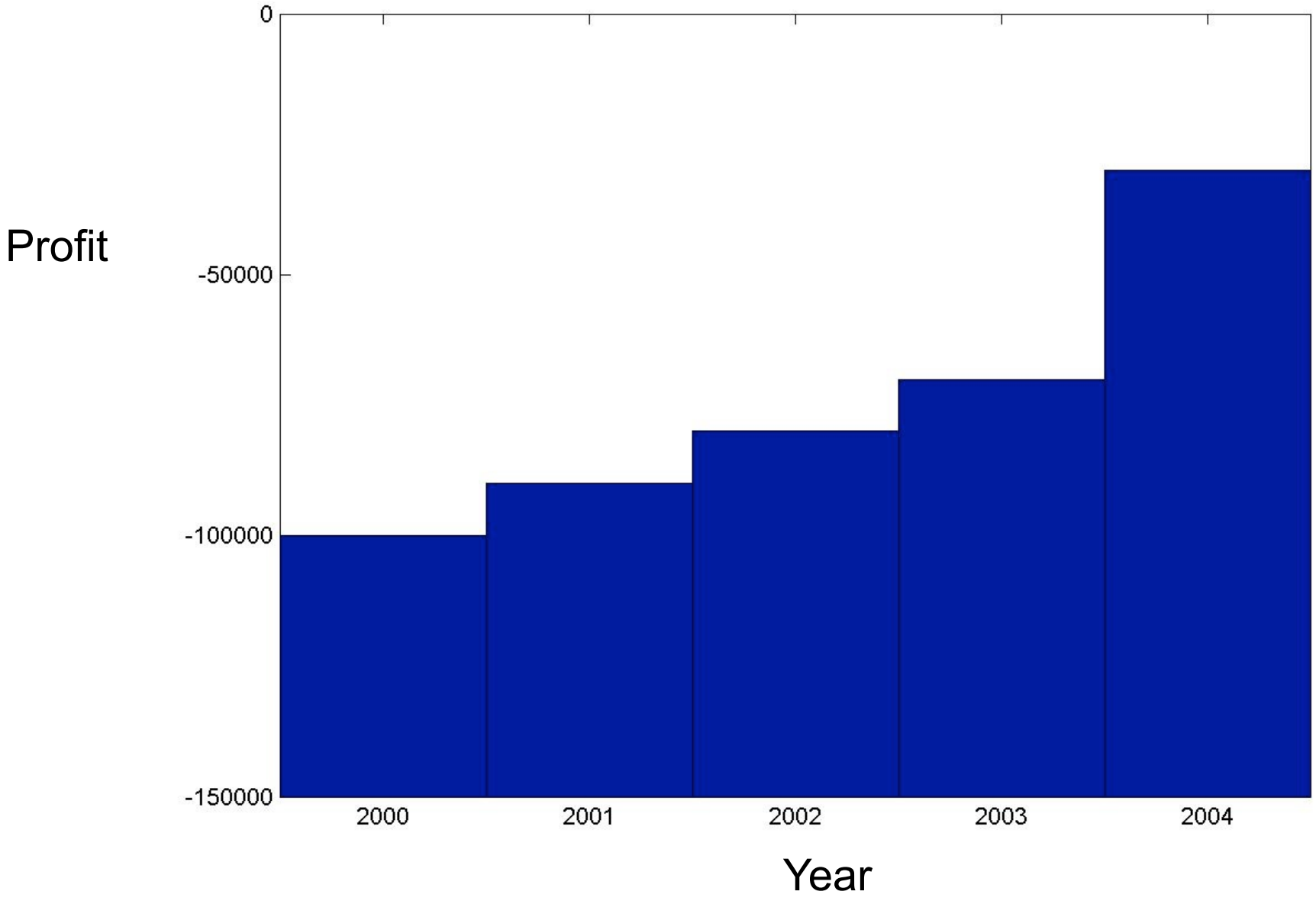
- Setting false baselines
- Comparing apples and oranges
- 2 dimensional graphics representing 1 dimensional data
- Distorting effects
- Ineffective comparisons
- Extra non-information carrying graphics
- Excessive prettiness

from: Tufte, E.R. (1983). The visual display of quantitative information.

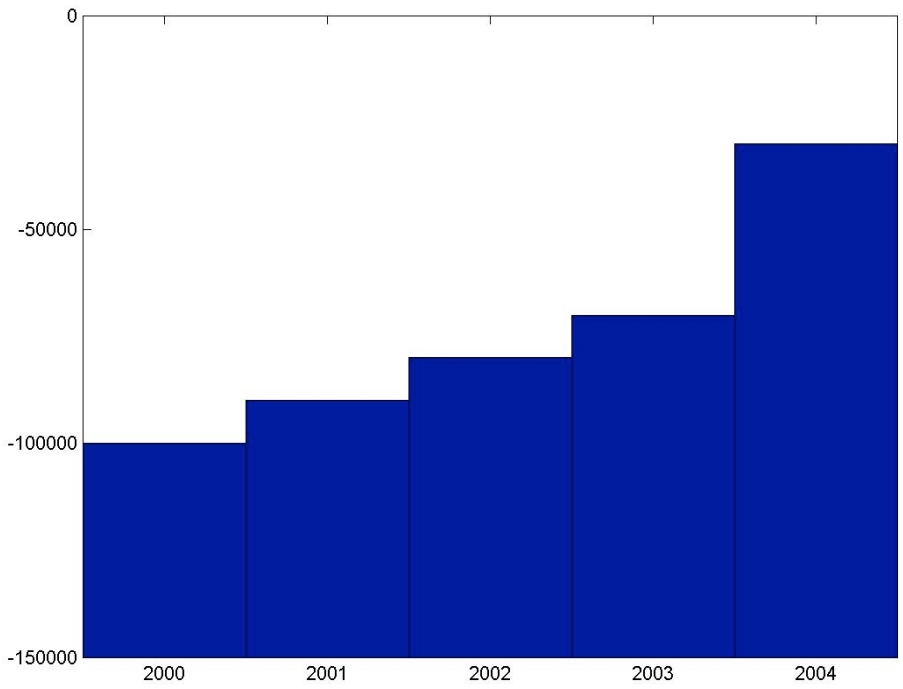
General principles:

- Maximize the amount of data presented for the amount of ink used.

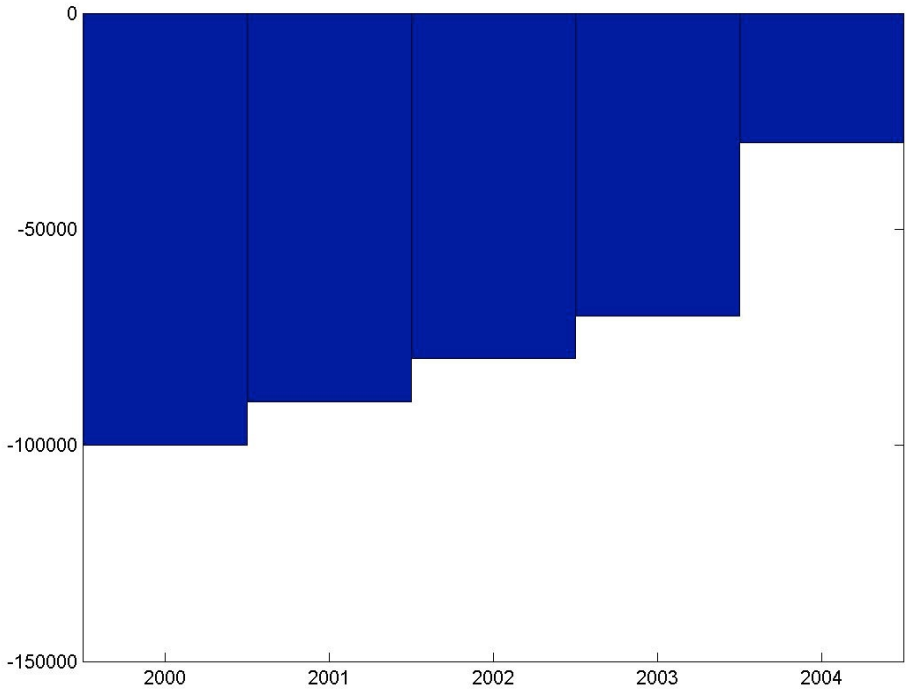
bad practice: false baselines



bad practice: false baselines



Are we making more money?



No! We are just losing less!

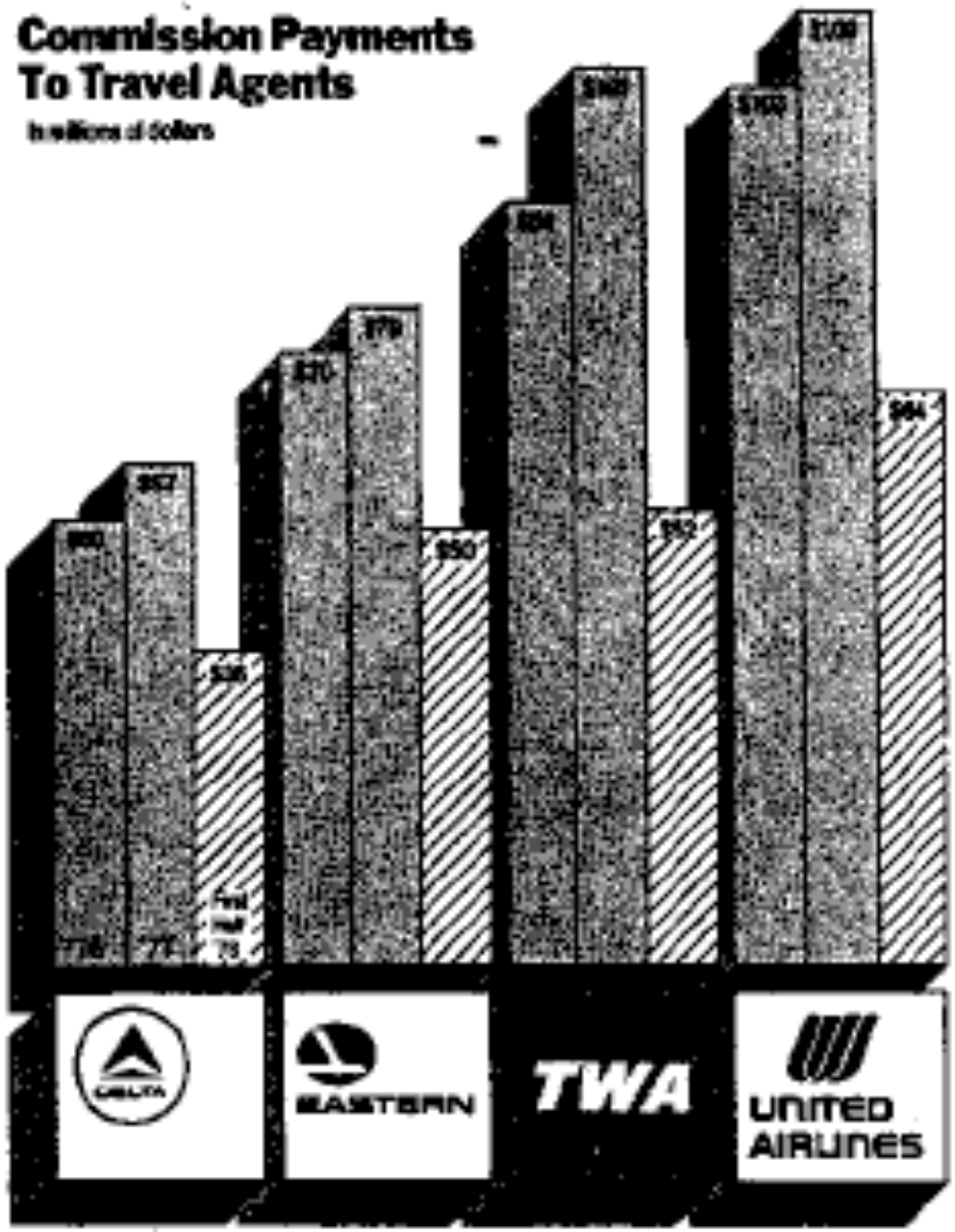
bad practice: comparing apples and oranges

Notice how revenues appear to be sharply declining!

Oh, no! Are the airlines going to fail?

No…the hashed bars are only from the first half of 1978!

Note how the hashing against the dark bars exaggerates the effect.

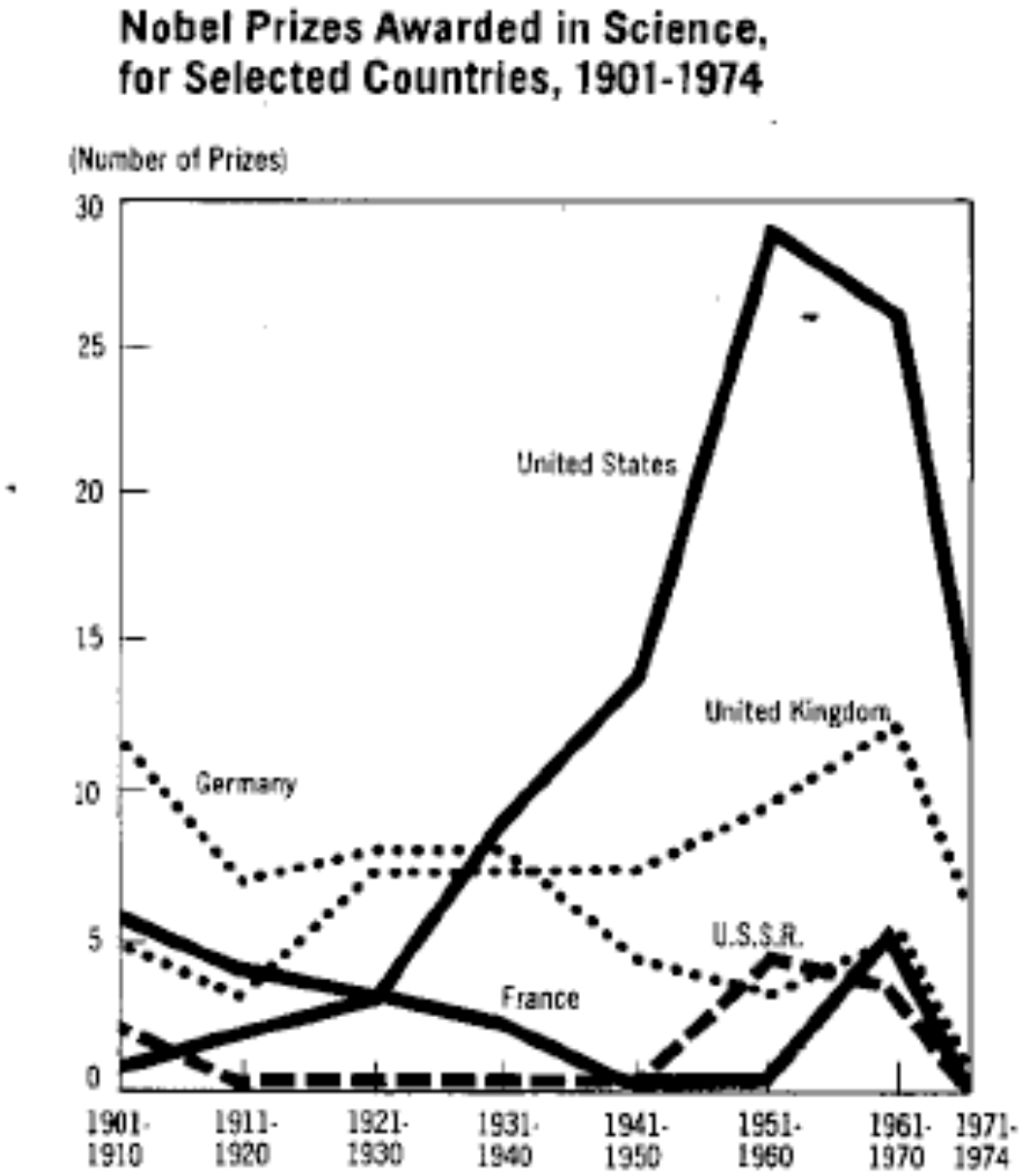


bad practice: comparing apples and oranges

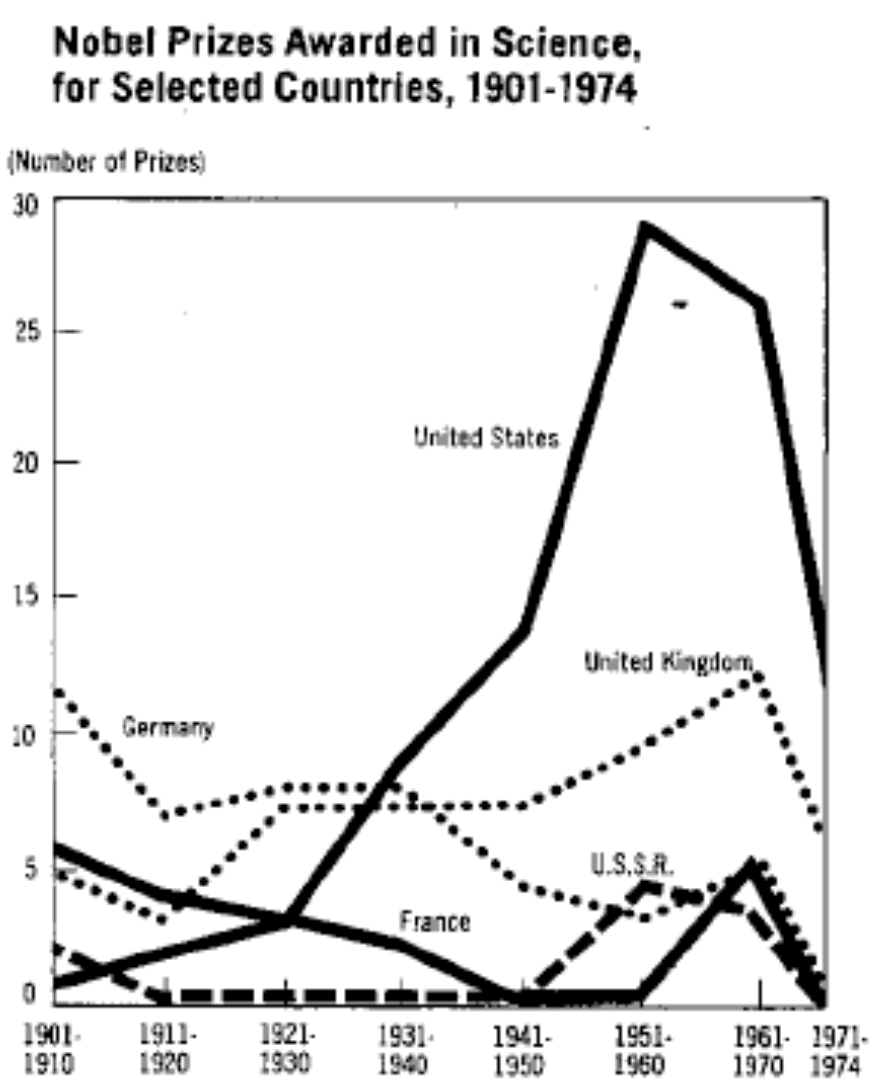
Oh no! Nobel prizes for the US appear to be sharply declining!

Now look at the X axis: the label for the last tick – that data is only for 3 years (not 10).

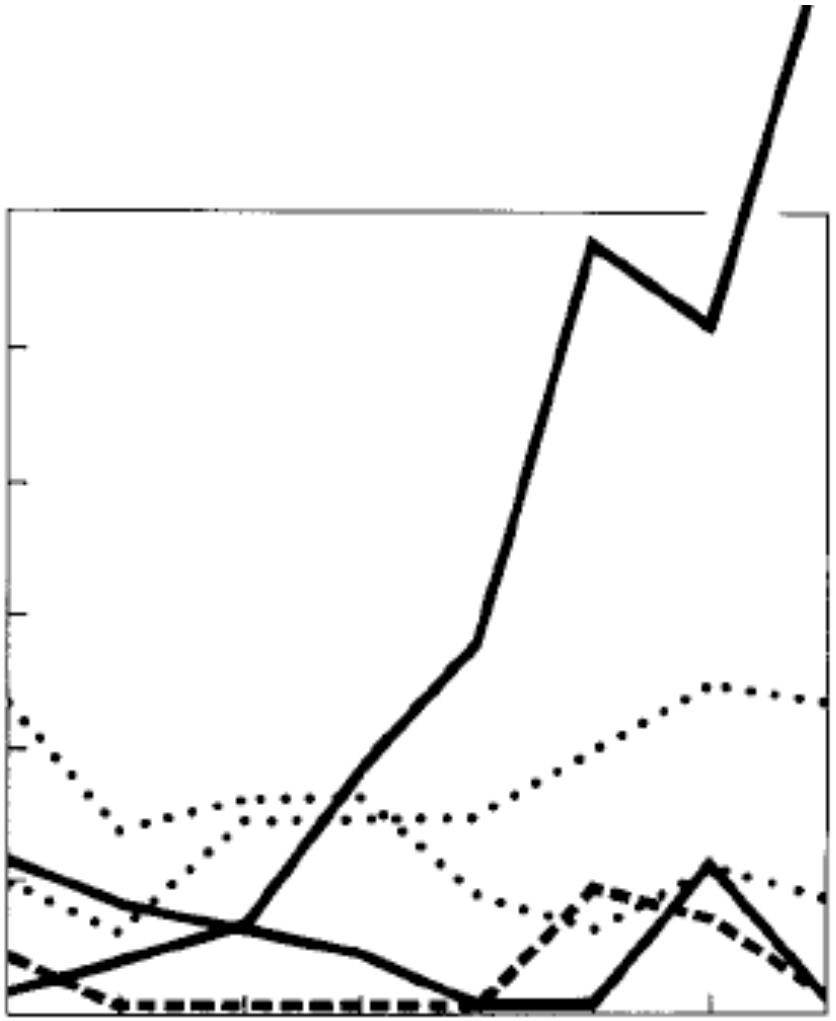
If we extrapolated to a full 10 years, based on this data, the graph would look like this:



bad practice: comparing apples and oranges



Bad graph



Better graph

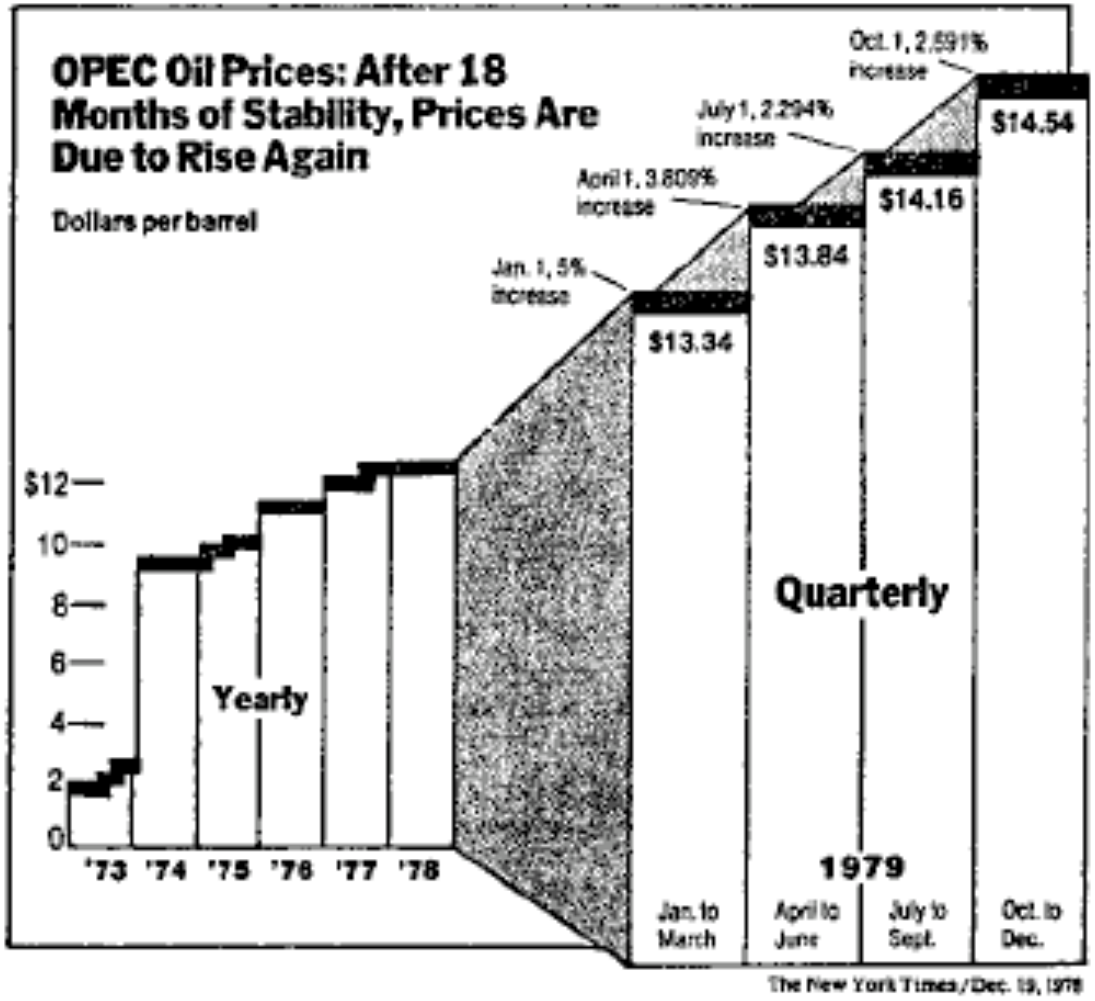
bad practice: comparing apples and oranges

Five different vertical scales show the price:

During this time	one vertical inch equals
1973–1978	$8.00
January–March 1979	$4.73
April–June 1979	$4.37
July–September 1979	$4.16
October–December 1979	$3.92

And two different horizontal scales show the passage of time:

During this time	one horizontal inch equals
1973–1978	3.8 years
1979	0.57 years



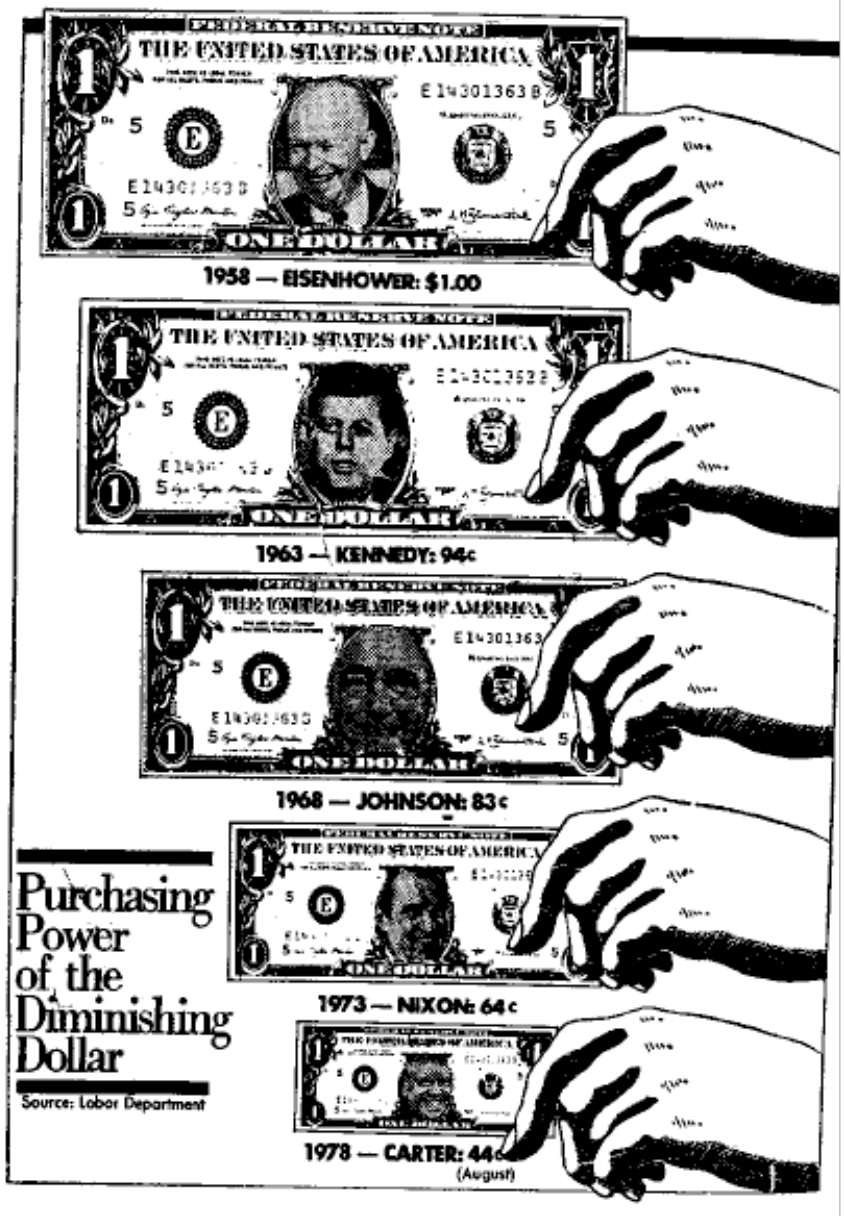
bad practice: 1D data on 2D plot

Price changed 554%.

Area of the barrels changes roughly 1300%!

bad practice: 1D data on 2D plot

If the visual area correctly represented the change, the 1978 dollar should be two times larger!

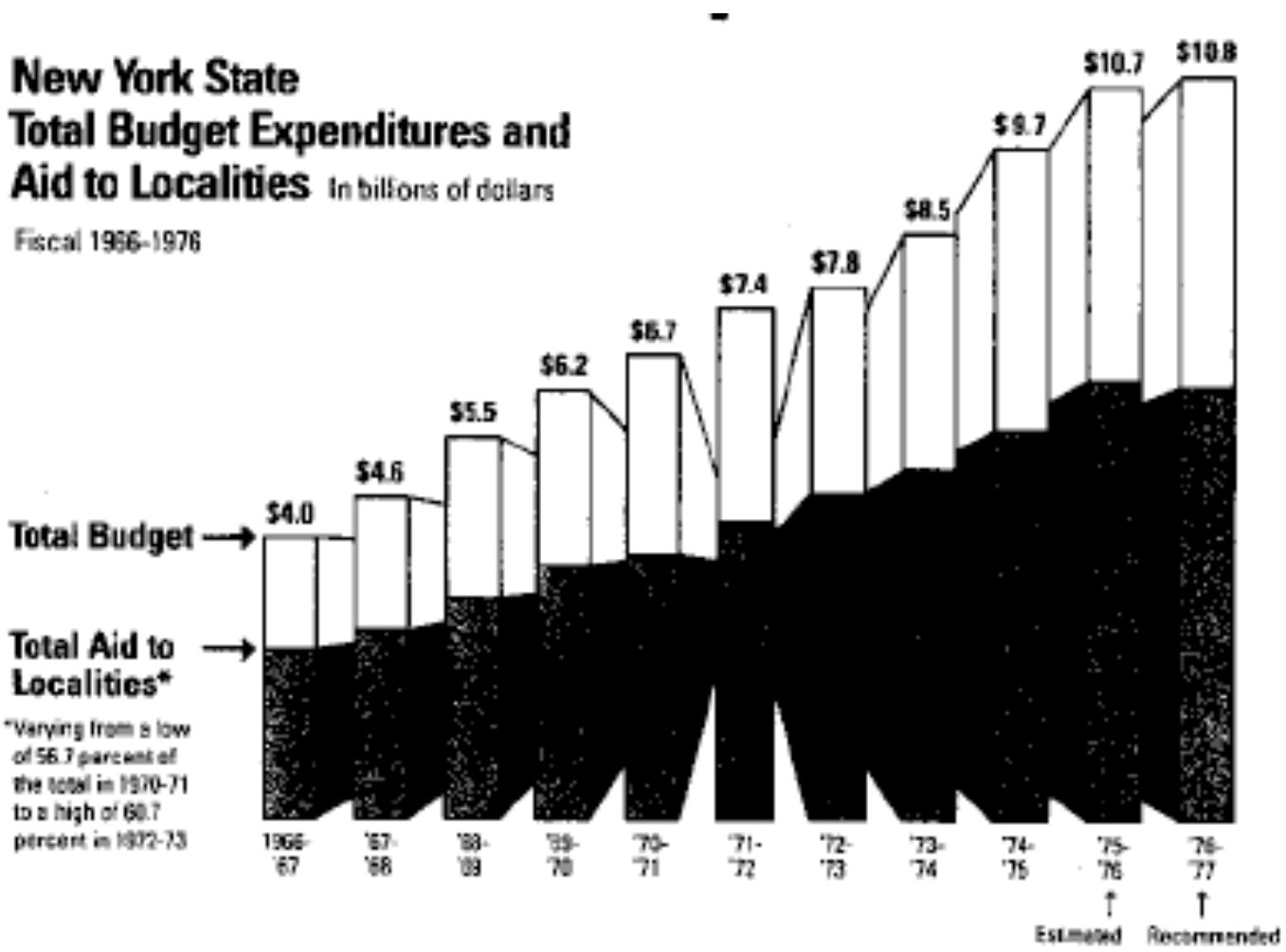


bad practice: distorting effects

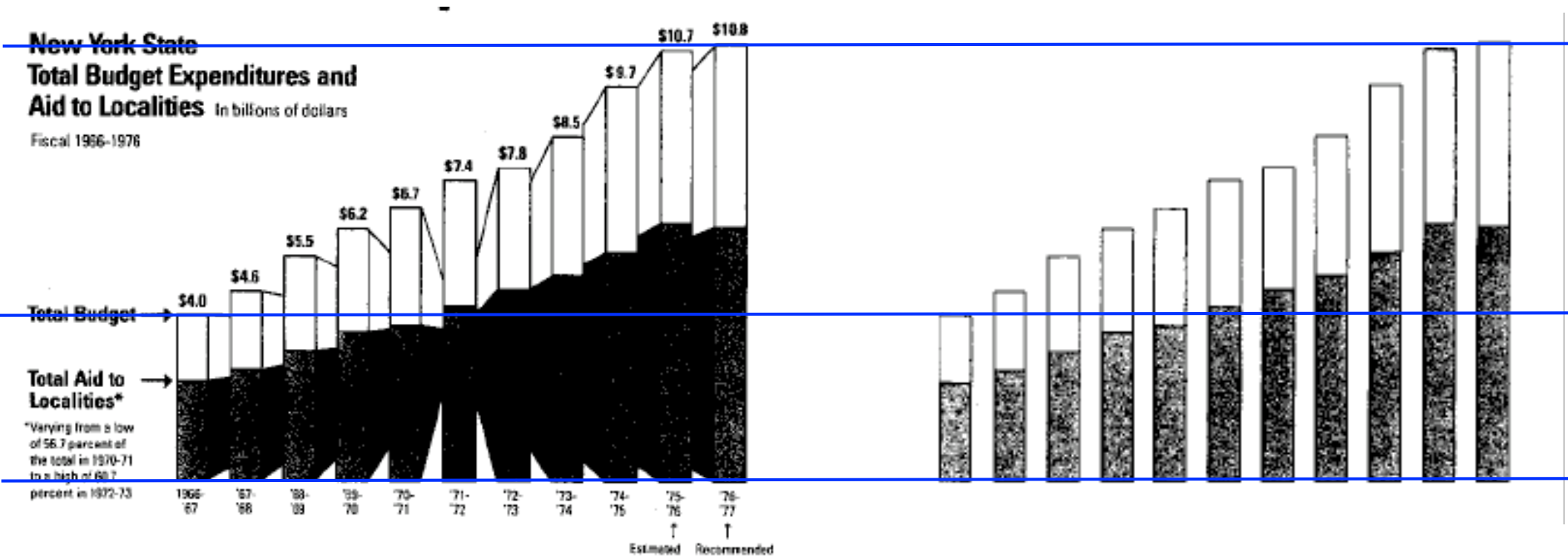
Note:

- The weird 3D thing going on
- Apparent variation in depth
- Use of arrows
- Labels are squeezed next to smallest bar

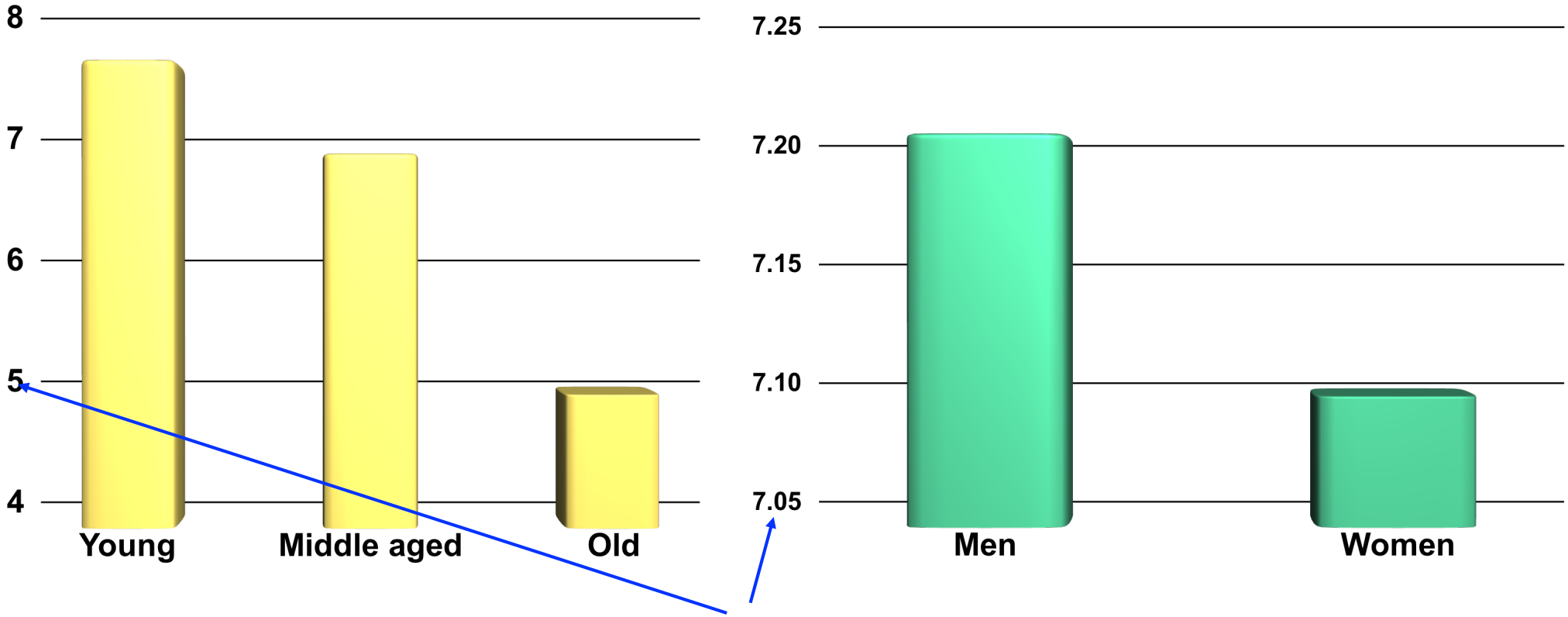
Compare this to when junk is deleted:



bad practice: distorting effects

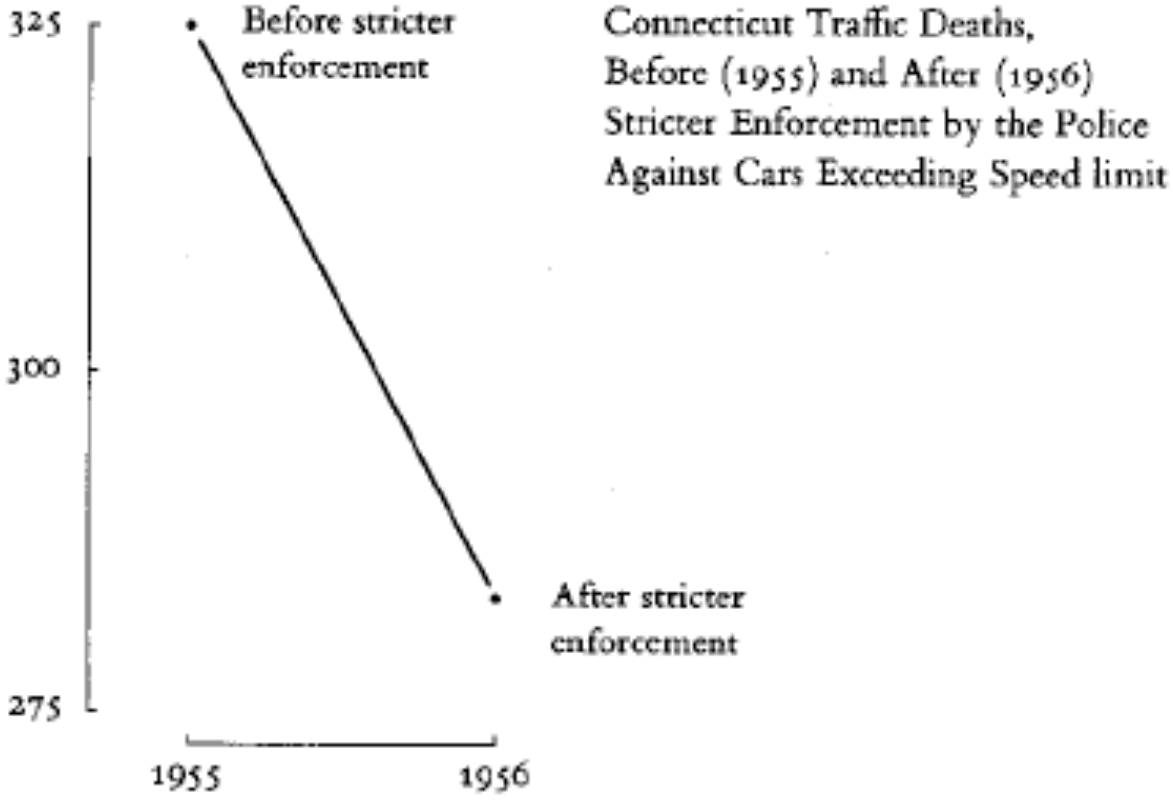


bad practice: ineffective (or no) comparison

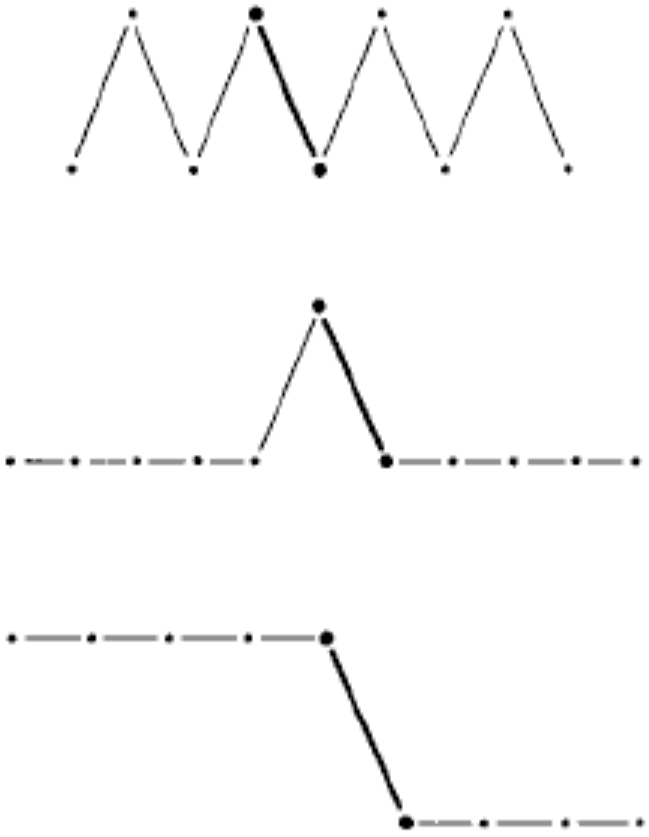


Misleading comparison! Look at the y axis labels!

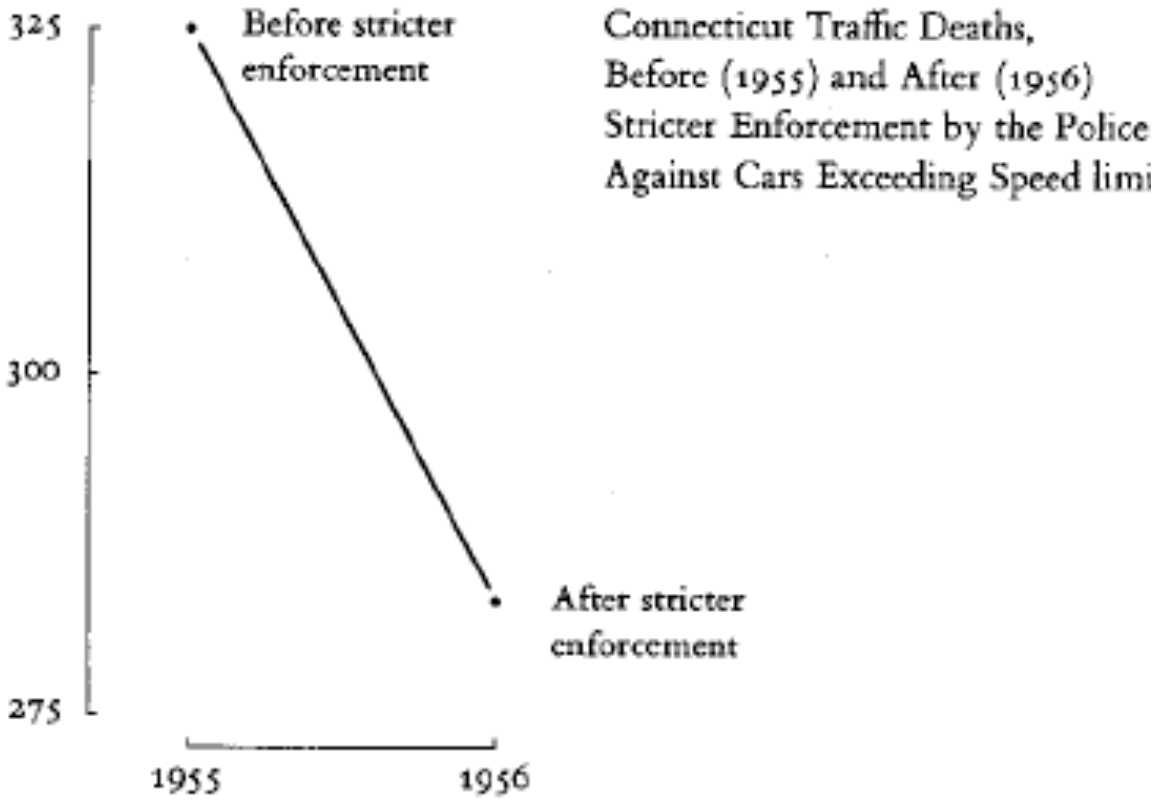
bad practice: ineffective (or no) comparison

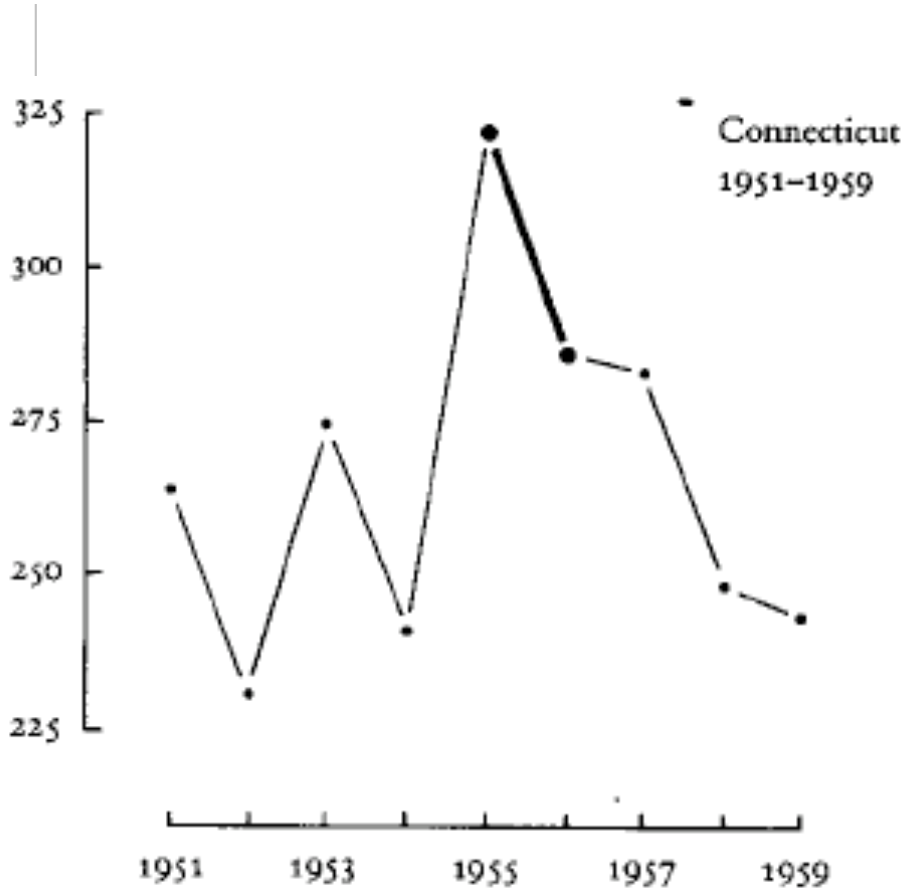


Connecticut Traffic Deaths, Before (1955) and After (1956) Stricter Enforcement by the Police Against Cars Exceeding Speed limit

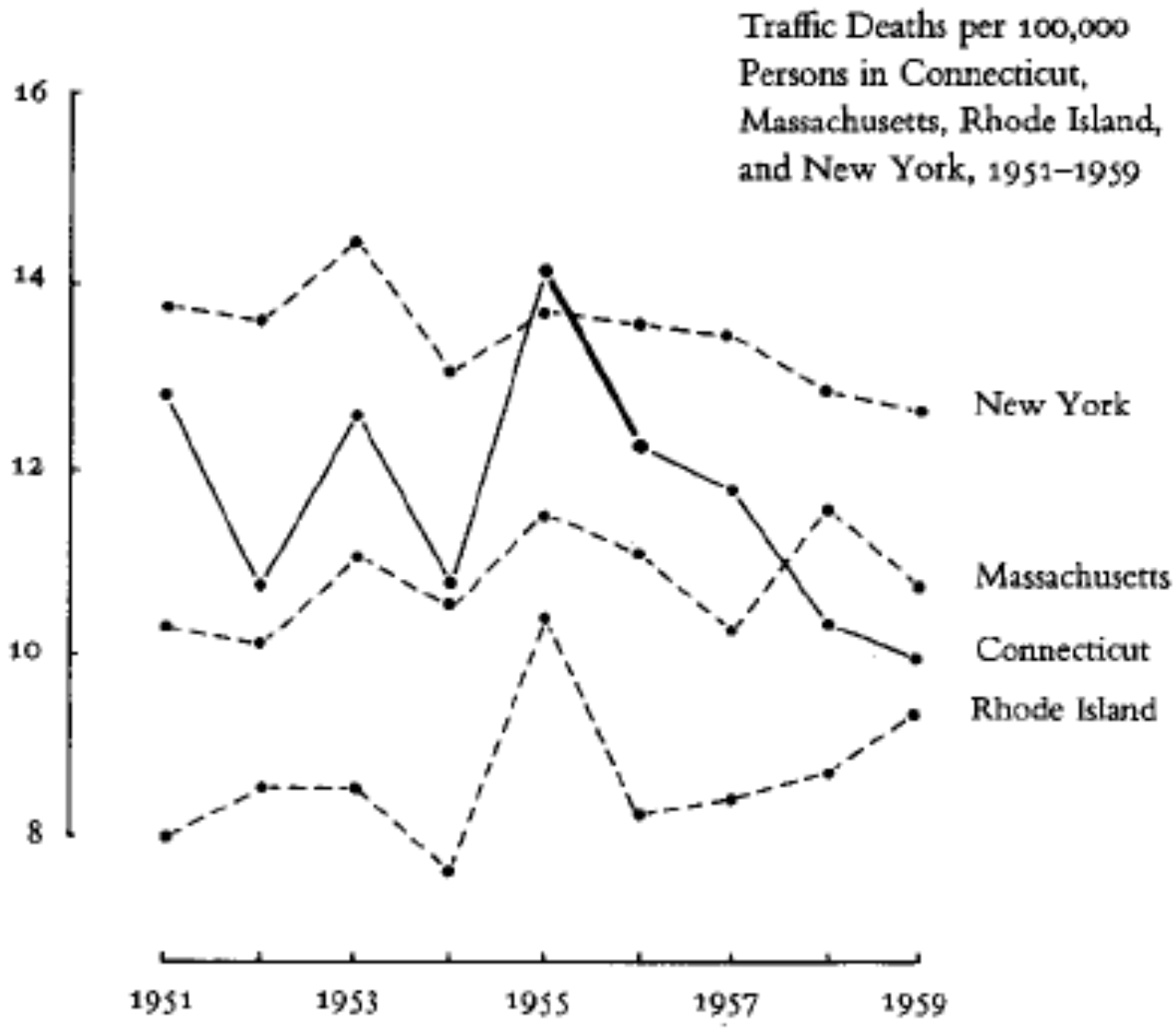


bad practice: ineffective (or no) comparison





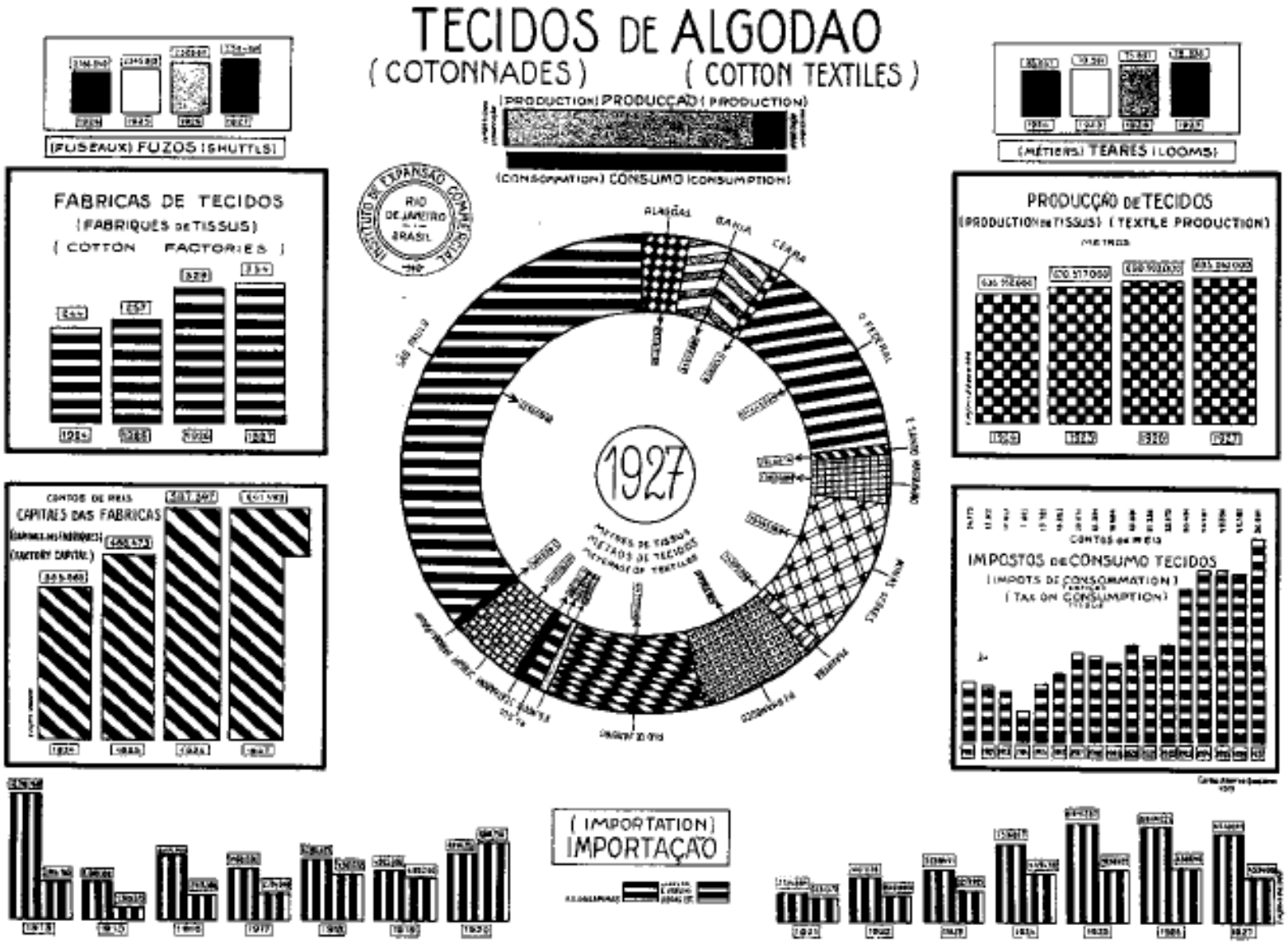
bad practice: ineffective (or no) comparison



bad practice: chart junk!

Cross-hatching is bad.

Stripes are too…

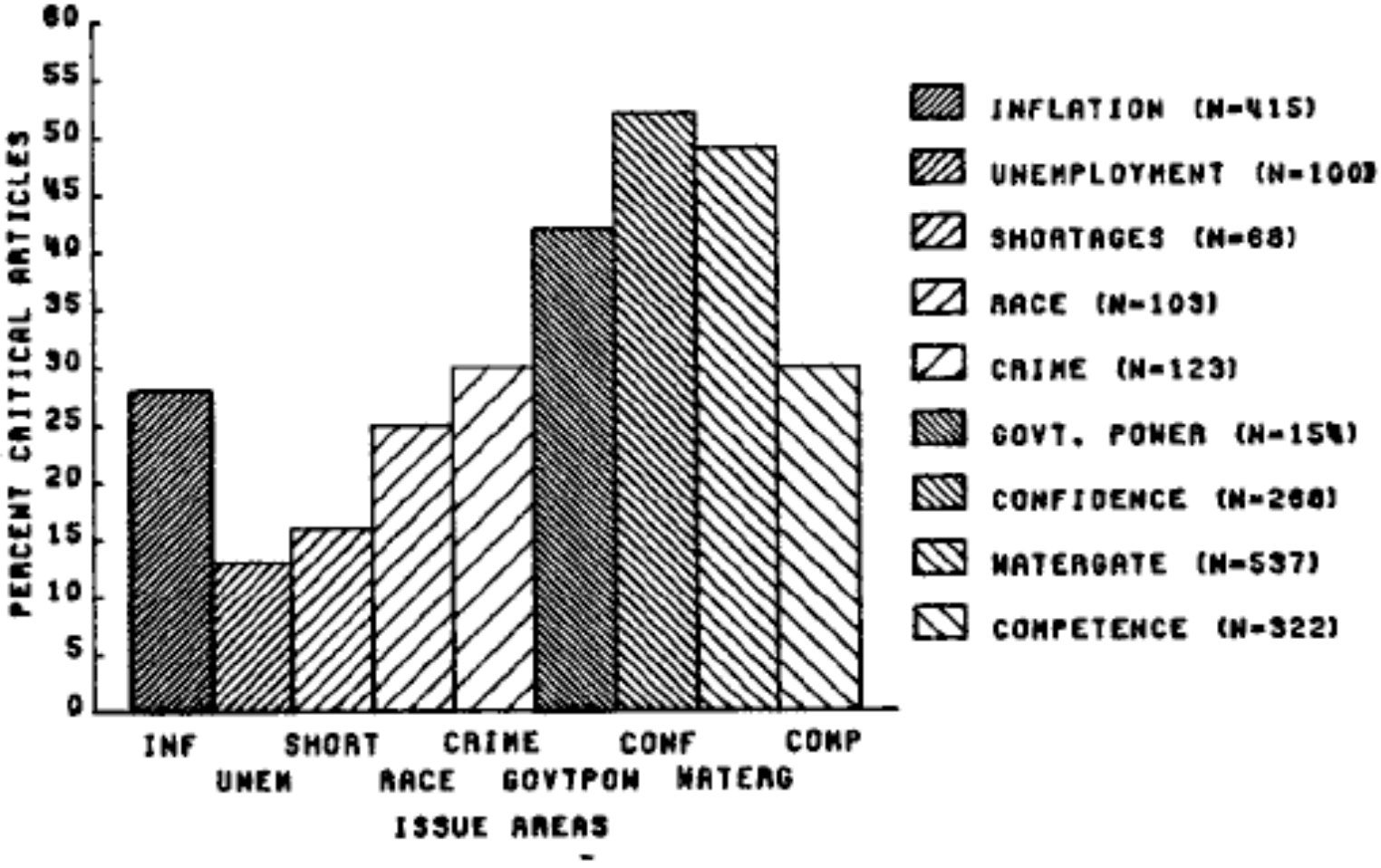


bad practice: chart junk!

Stripes create strange visual effects!

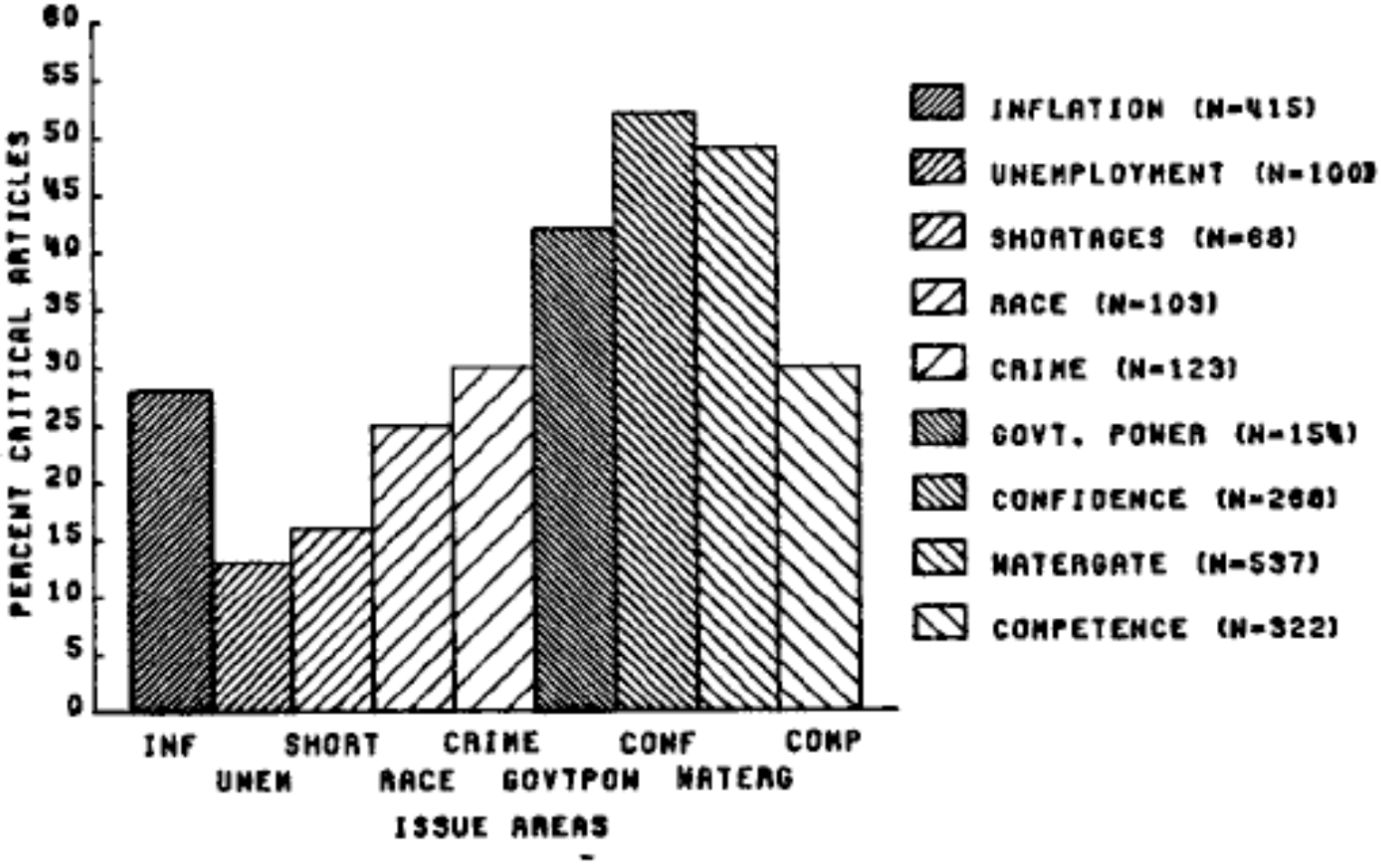
bad practice: chart junk!

Stripes create strange visual effects!



bad practice: chart junk!

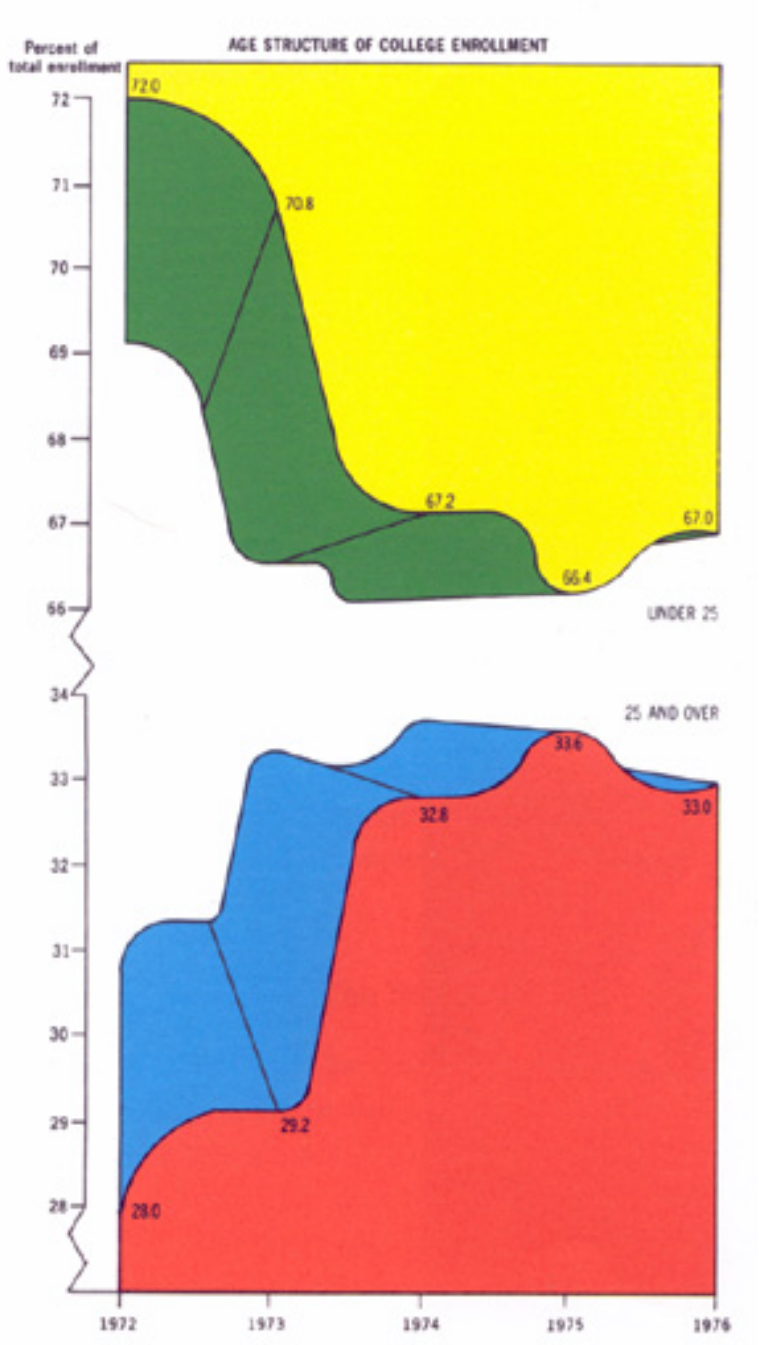
Stripes create strange visual effects!



bad practice: chart junk!

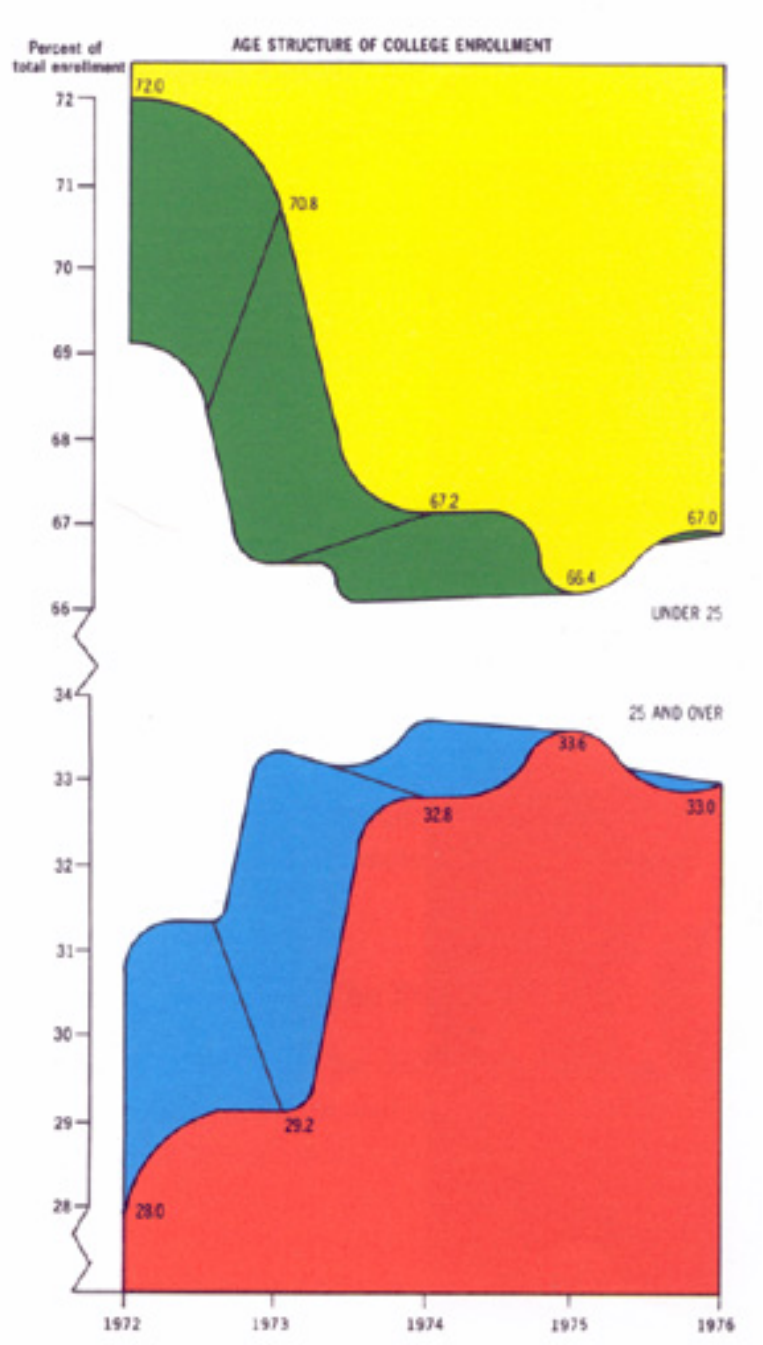
Minimize use of colors!

Never use red/ green to contrast important groups!



bad practice: chart junk!

Data first, pretty second!



Projects!

Pitches start Oct 3rd!

First round due Nov 18th by 11:59 pm

Will then get comments

Give presentation in class

Turn in final assignment Dec 14 11:59 pm

- **HW: Using the data set that you picked for cleaning, visualize 3 variables both individually and pairwise**

- Write a short summary of the conclusions that you can and can't draw from your visualizations