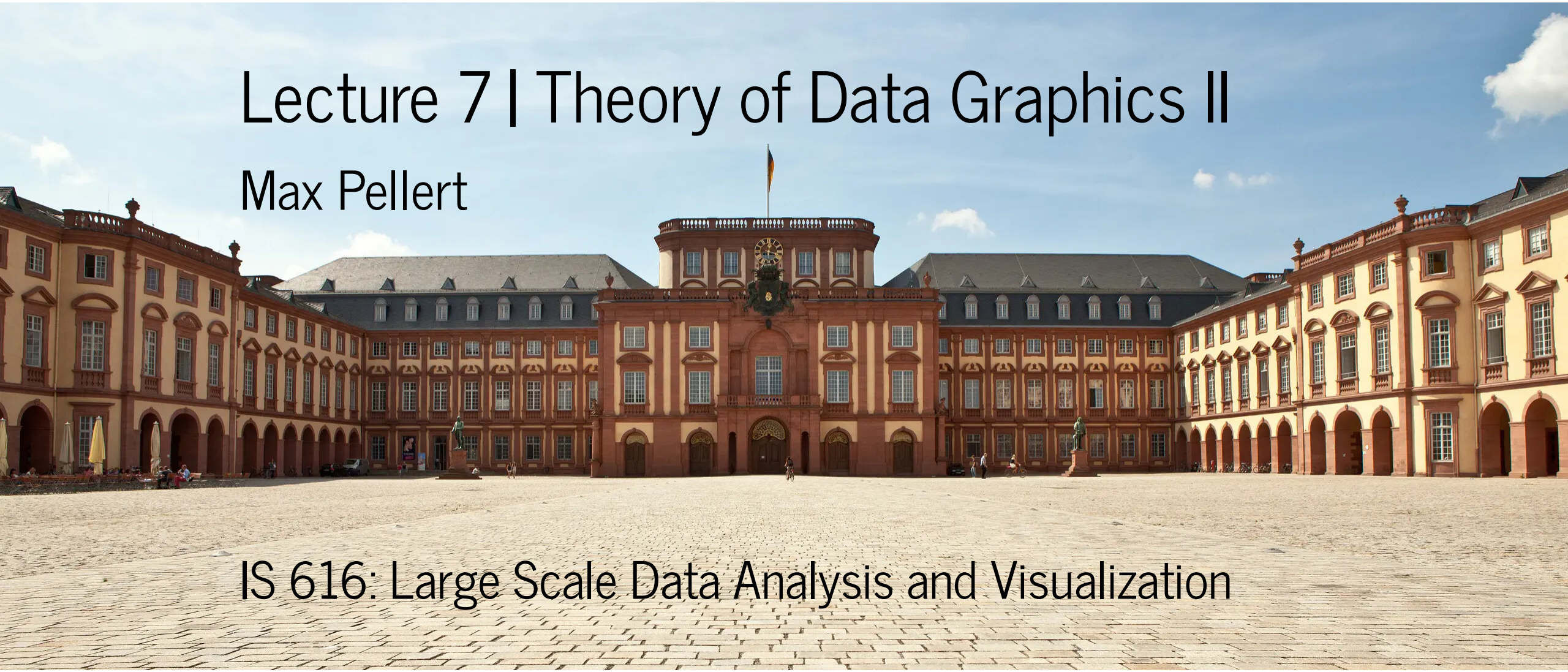
Lecture 7 | Theory of Data Graphics II

Max Pellert



IS 616: Large Scale Data Analysis and Visualization

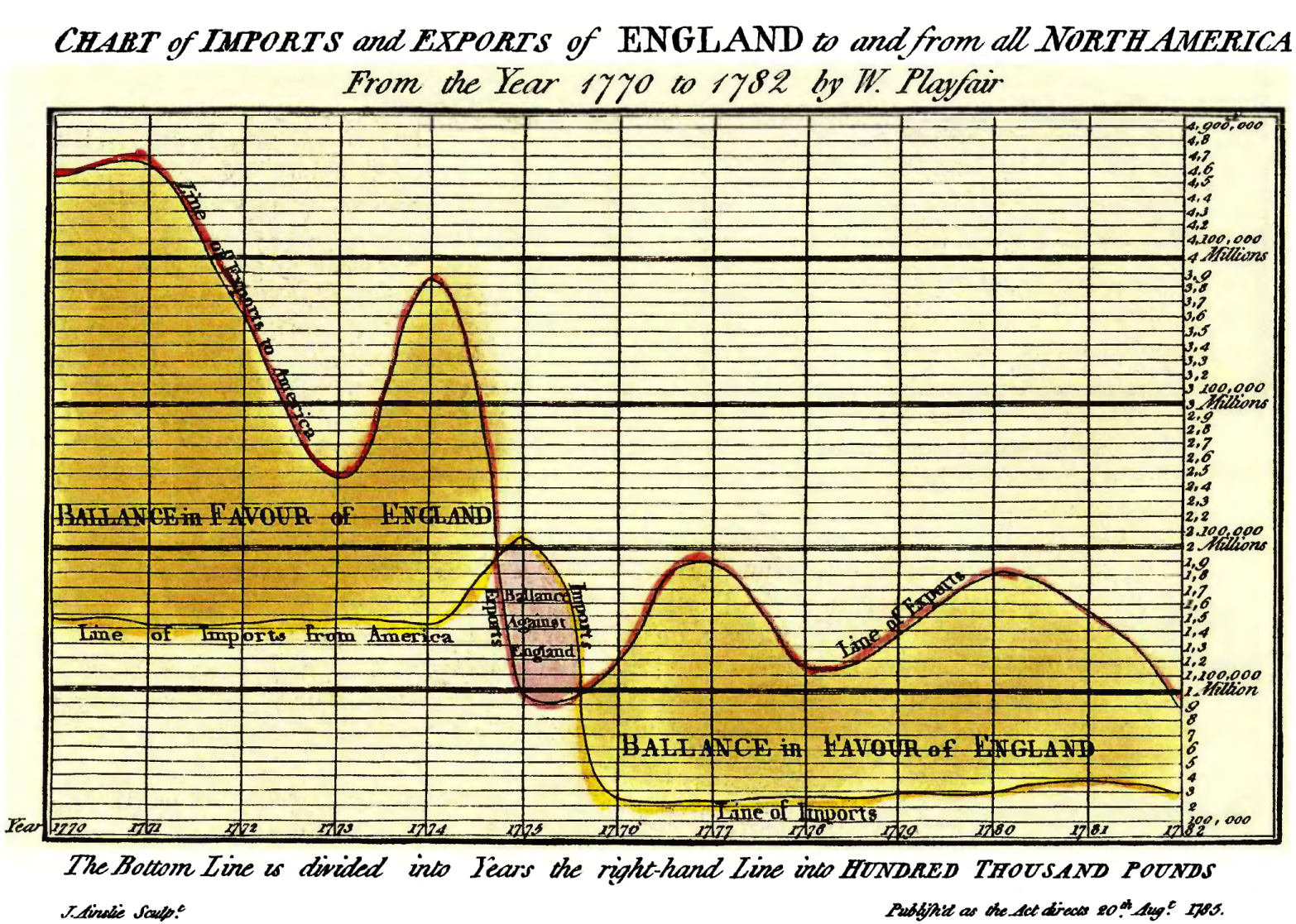
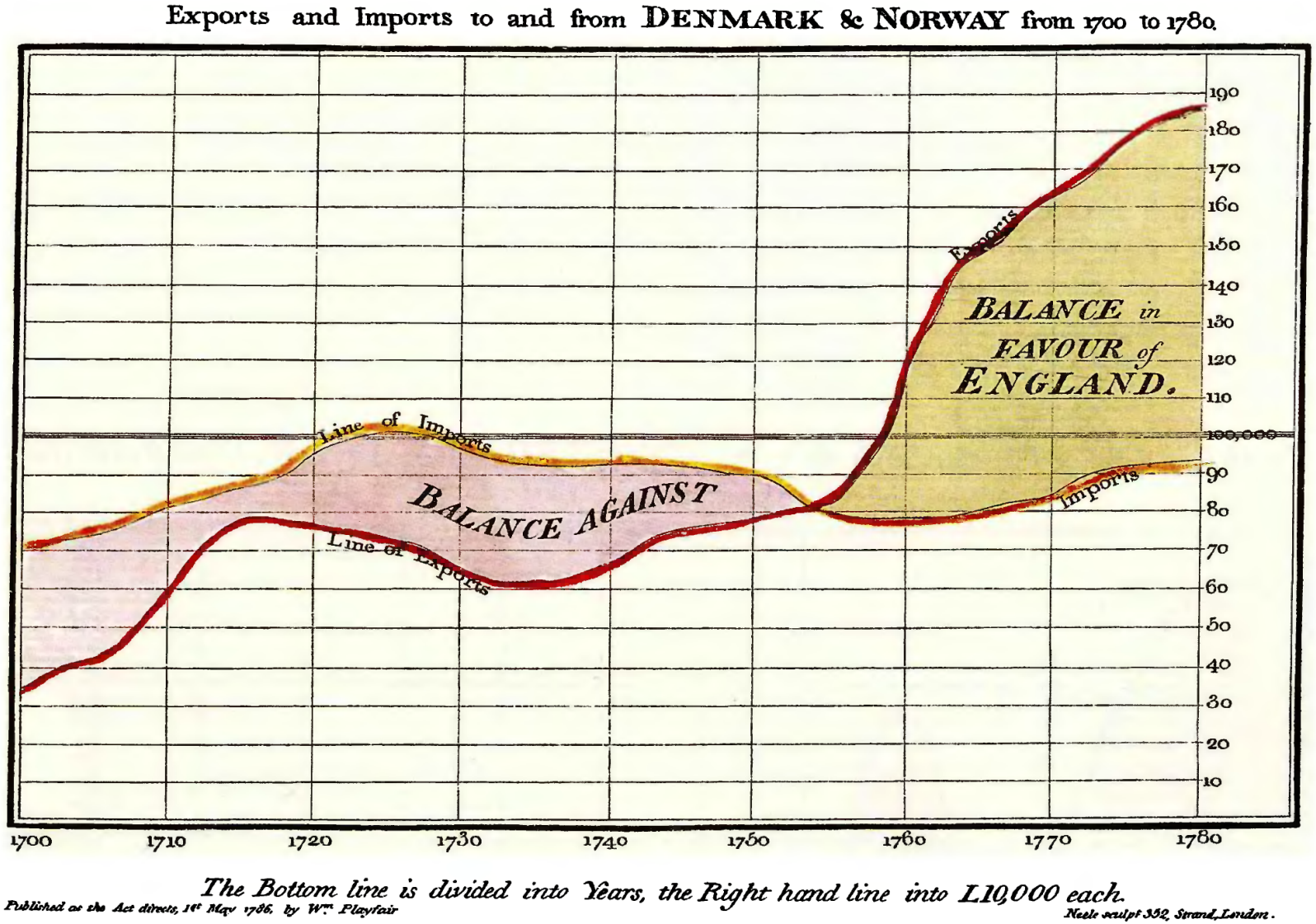


CHART of IMPORTS and EXPORTS of ENGLAND to and from all NORTH AMERICA
From the Year 1770 to 1782 by W. Playfair
Line of Exports to America
BALLANCE in FAVOUR of ENGLAND
Line of Imports from America
Balance Against England
Exports
Imports
Line of Exports
BALLANCE in FAVOUR of ENGLAND
Line of Imports
Year
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
4,900,000
4,800,000
4,700,000
4,600,000
4,500,000
4,400,000
4,300,000
4,200,000
4,100,000
4 Millions
3,900,000
3,800,000
3,700,000
3,600,000
3,500,000
3,400,000
3,300,000
3,200,000
3,100,000
3 Millions
2,900,000
2,800,000
2,700,000
2,600,000
2,500,000
2,400,000
2,300,000
2,200,000
2,100,000
2 Millions
1,900,000
1,800,000
1,700,000
1,600,000
1,500,000
1,400,000
1,300,000
1,200,000
1,100,000
1 Million
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
The Bottom Line is divided into Years the right-hand Line into HUNDRED THOUSAND POUNDS
J. Ainslie Sculp.t
Publish'd as the Act directs 20th Aug.t 1785.


Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.
190
180
170
160
150
140
130
120
110
100,000
90
80
70
60
50
40
30
20
10
Exports
BALANCE in FAVOUR of ENGLAND.
Line of Imports
BALANCE AGAINST
Line of Exports
Imports
1700
1710
1720
1730
1740
1750
1760
1770
1780
The Bottom line is divided into Years, the Right hand line into £10,000 each.
Published as the Act directs, 1st May 1786, by Wm. Playfair
Neele sculpt 352, Strand, London.

The first chart devotes too much of their ink to graphical apparatus, with elaborate grid lines and detailed labels

In the second much of the non-data detail is eliminated

That leads to a cleaner design that focuses attention on the time-series itself

> This improvement in graphical design illustrates the fundamental principle of good statistical graphics:
>
> Above all else show the data.
>
> The principle is the basis for a theory of data graphics.

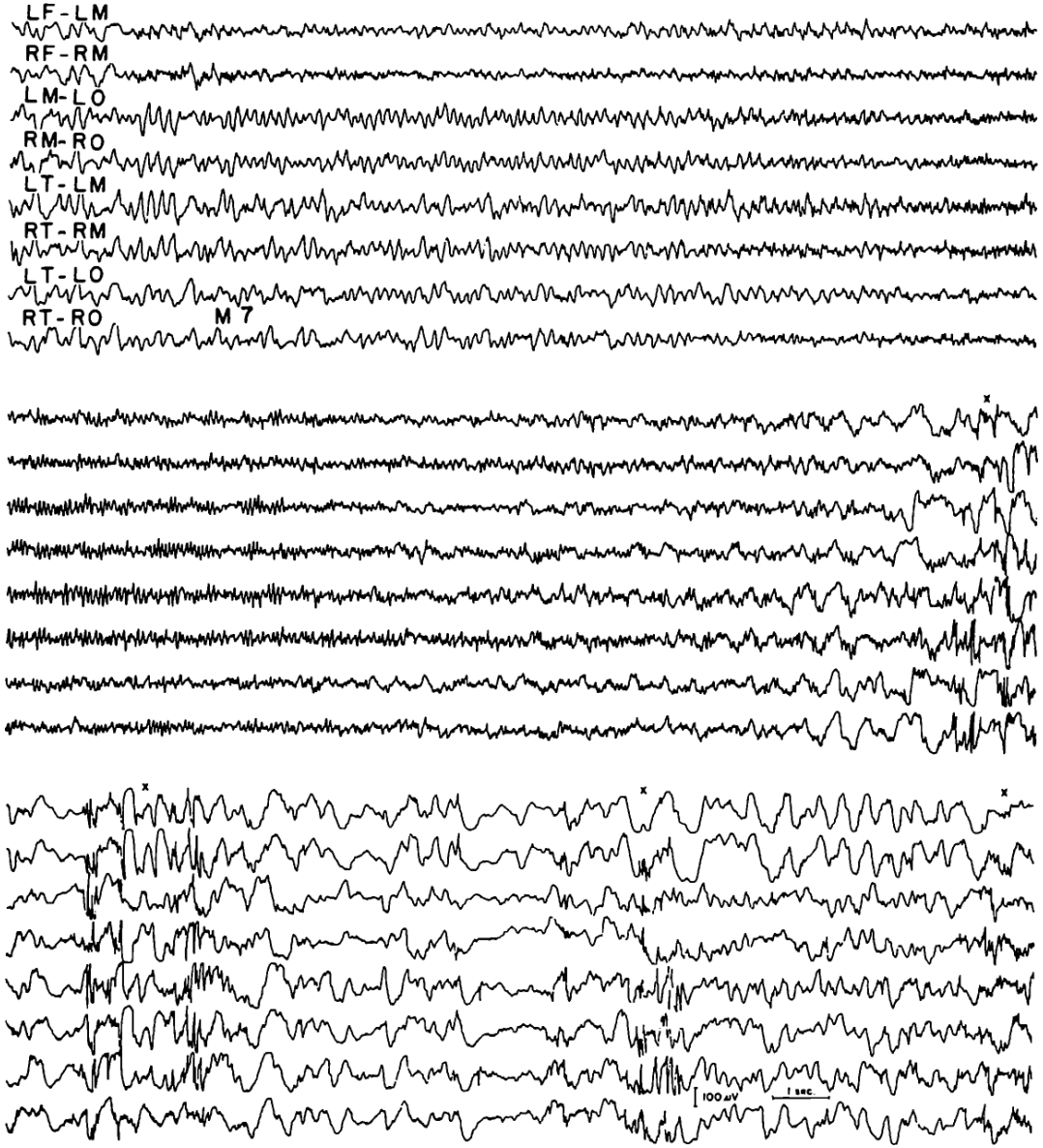
Data-Ink

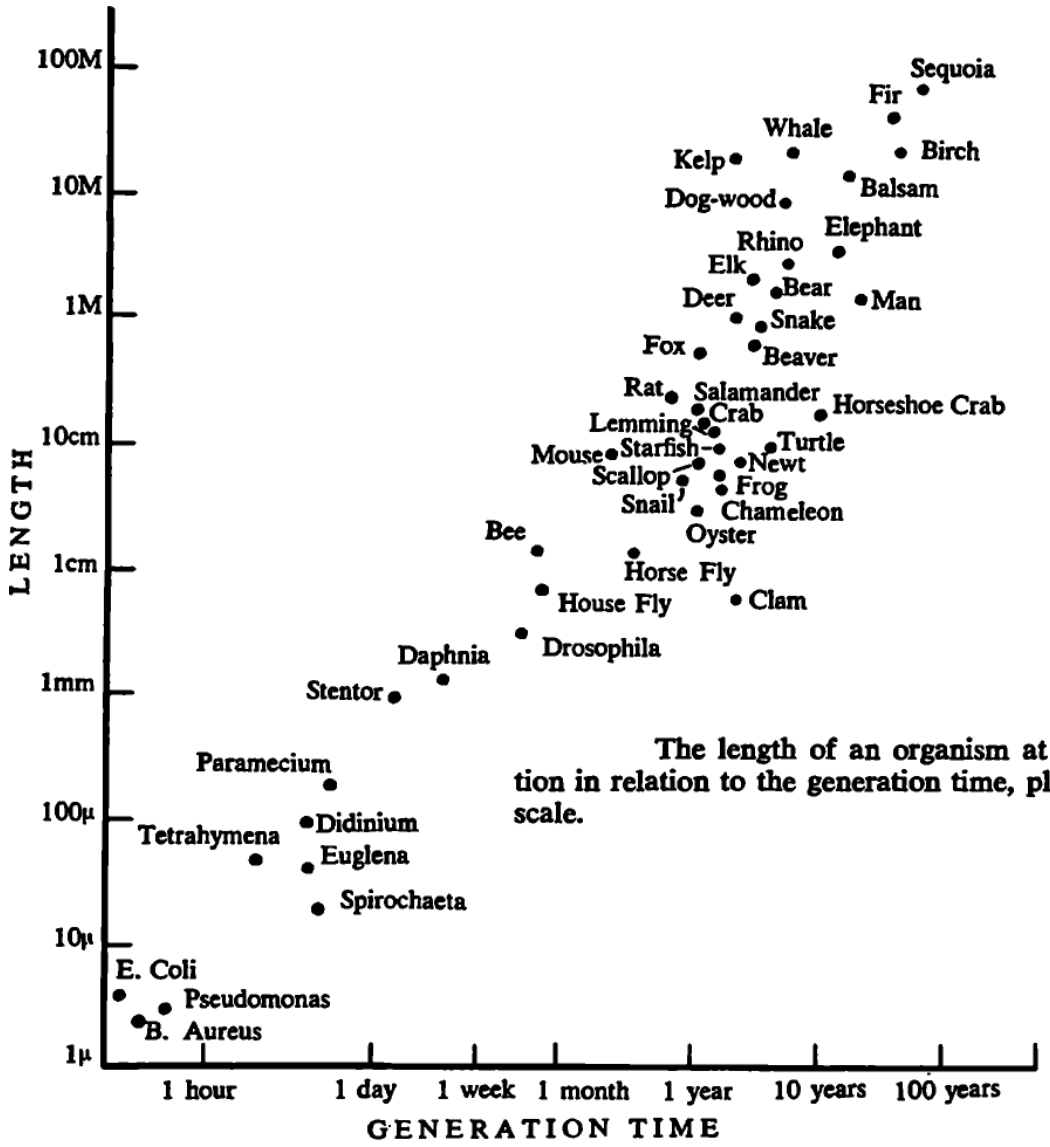
A large share of ink on a graphic should present data-information, the ink changing as the data change. *Data-ink* is the non-erasable core of a graphic, the non-redundant ink arranged in response to variation in the numbers represented.

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

$$= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information}$$

$$= 1.0 - \text{proportion of a graphic that can be erased without loss of data-information.}$$


LF-LM
RF-RM
LM-LO
RM-RO
LT-LM
RT-RM
LT-LO
RT-RO
M 7
100 μV
1 sec.



The length of an organism at the time of reproduction in relation to the generation time, plotted on a logarithmic scale.

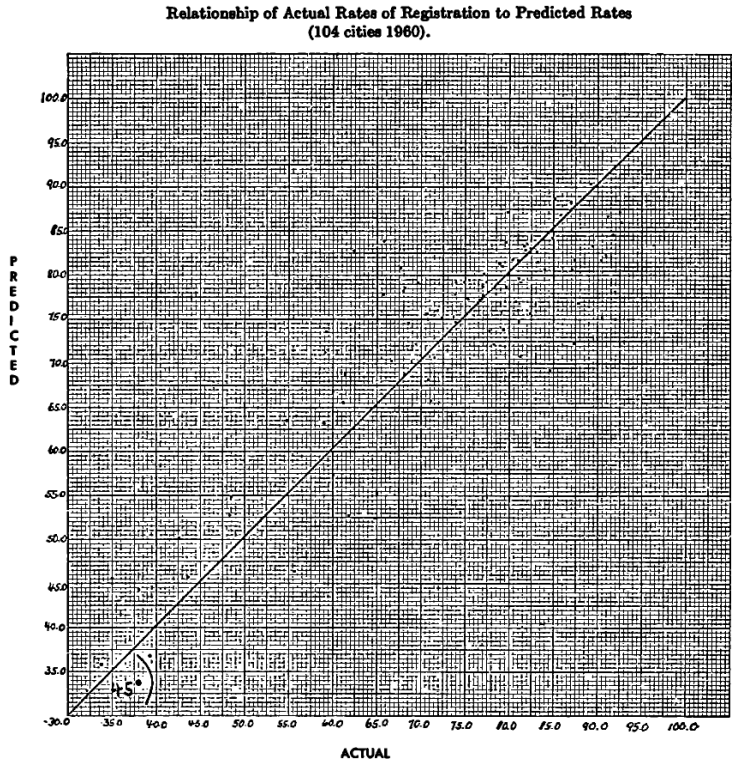
The larger the share of a graphic's ink devoted to data, the better (other relevant matters being equal):

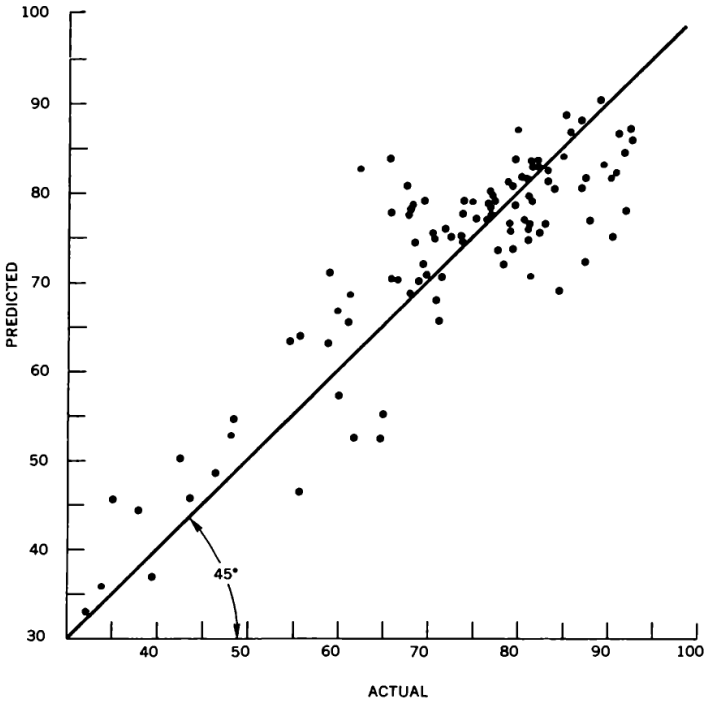
Maximize the data-ink ratio, within reason.

Every bit of ink on a graphic requires a reason. And nearly always that reason should be that the ink presents new information.

The other side of increasing the proportion of data-ink is an erasing principle:

Erase non-data-ink, within reason.





Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

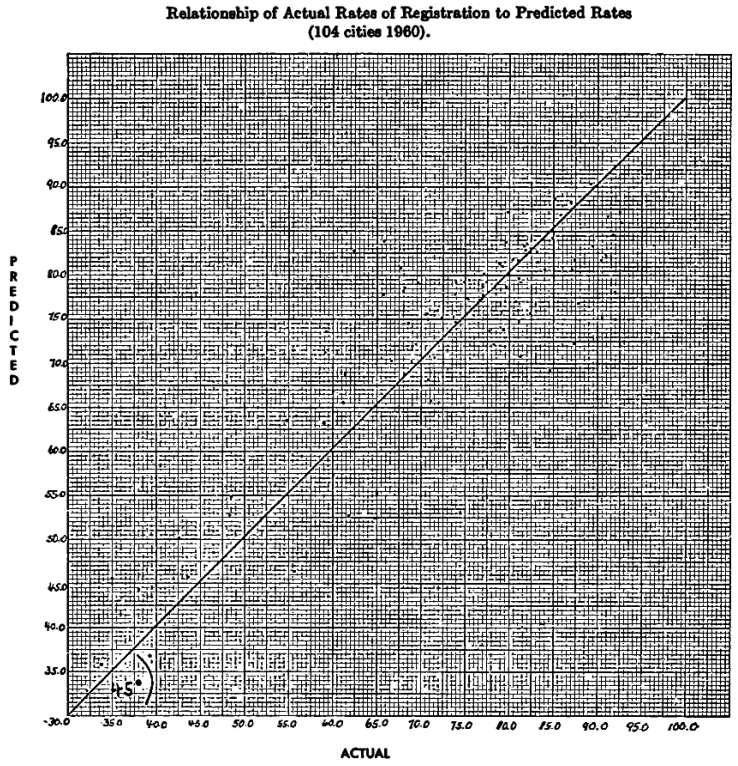
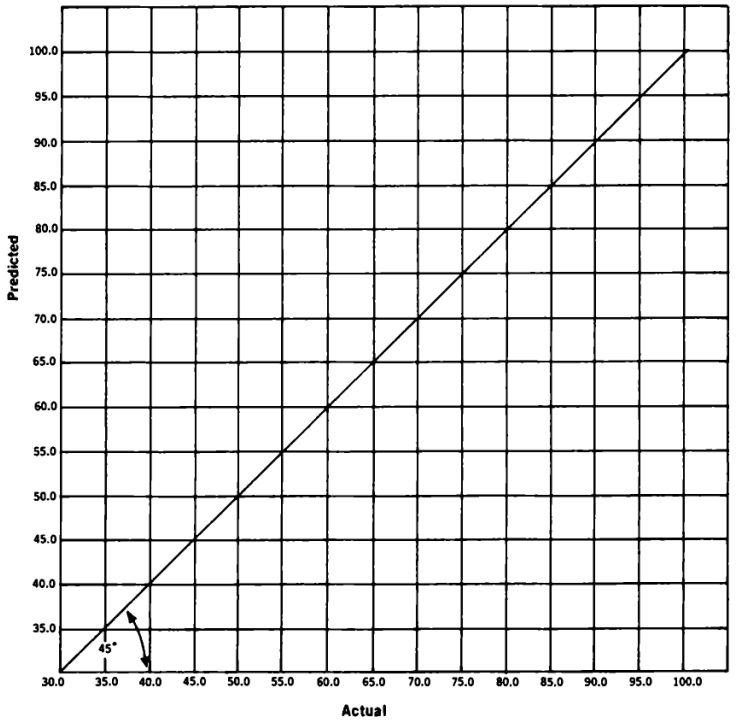
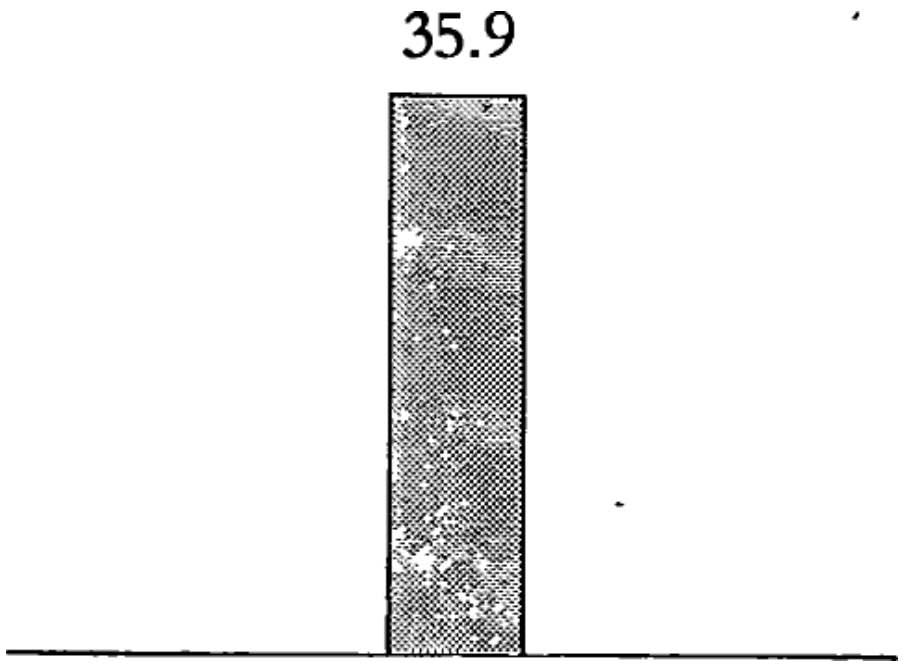
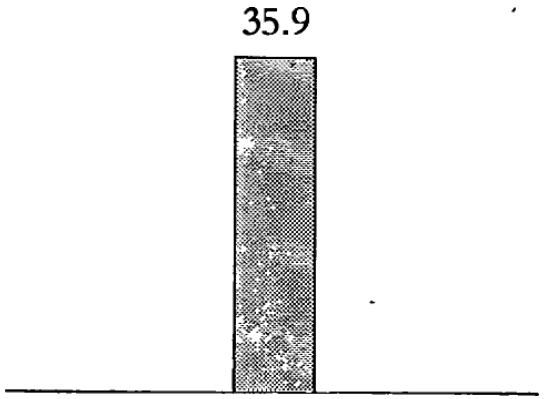


Figure 19.1 Relationship of Actual Rates of Registration to Predicted Rates (104 cities, 1960)





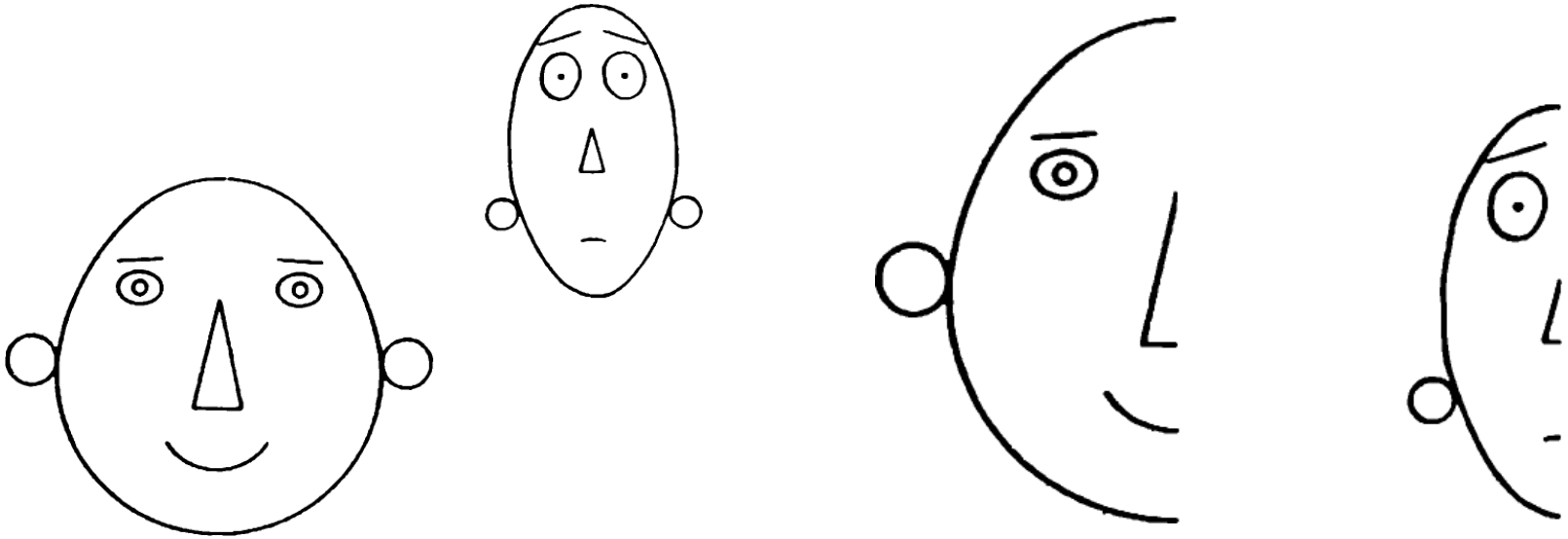
Unambiguously locates the altitude in six separate ways



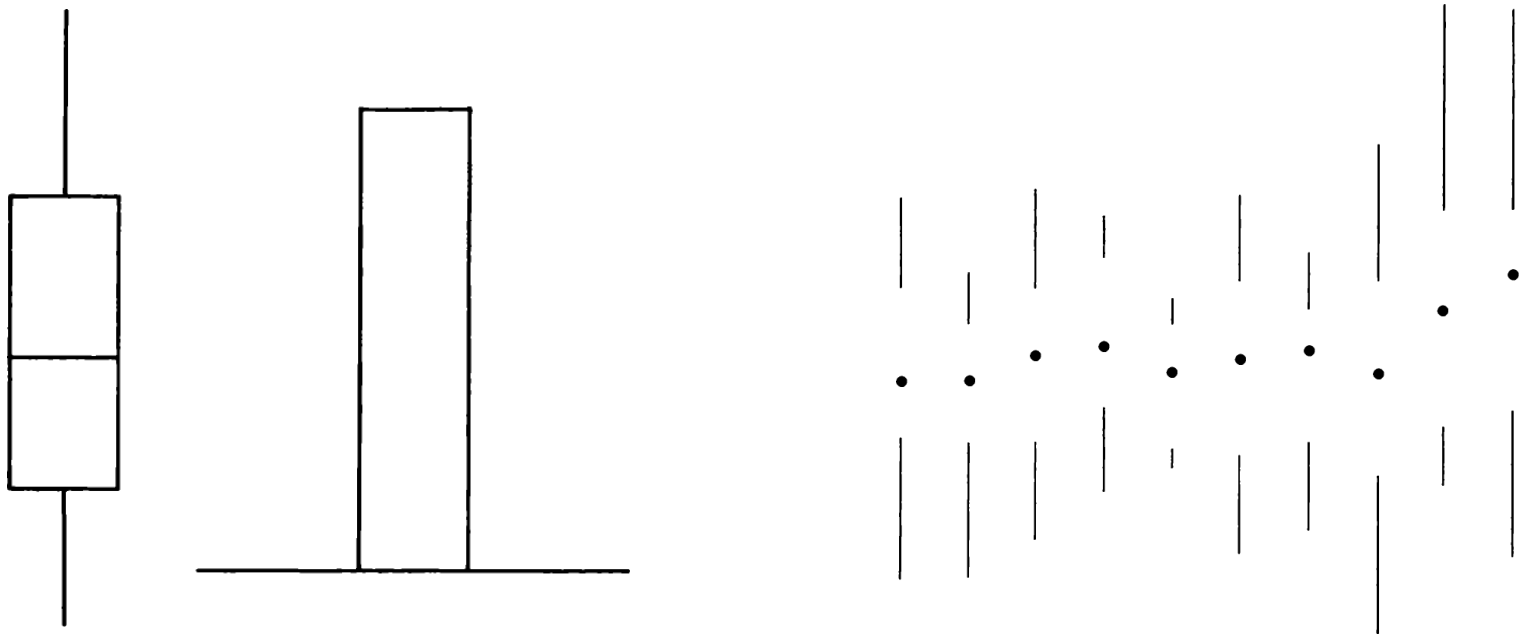
1. Height of the left line
2. Height of shading
3. Height of right line
4. Position of top horizontal line
5. Position (not content) of number at bar's top
6. The number itself

Any five of the six can be erased and the sixth will still indicate the height

Sometimes you don't need symmetry



And you can save space by removing redundant halves



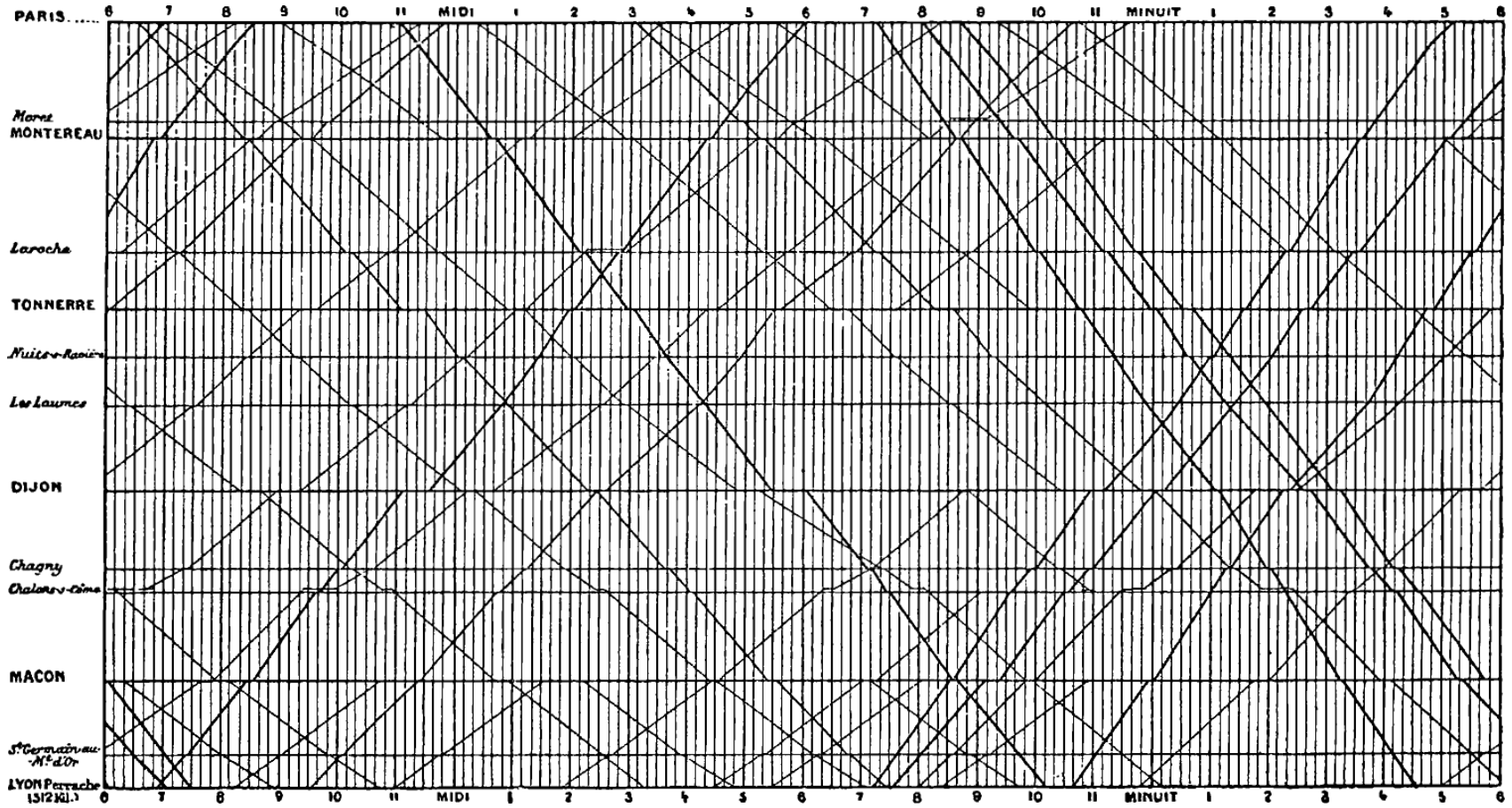
Redundant Data-Ink

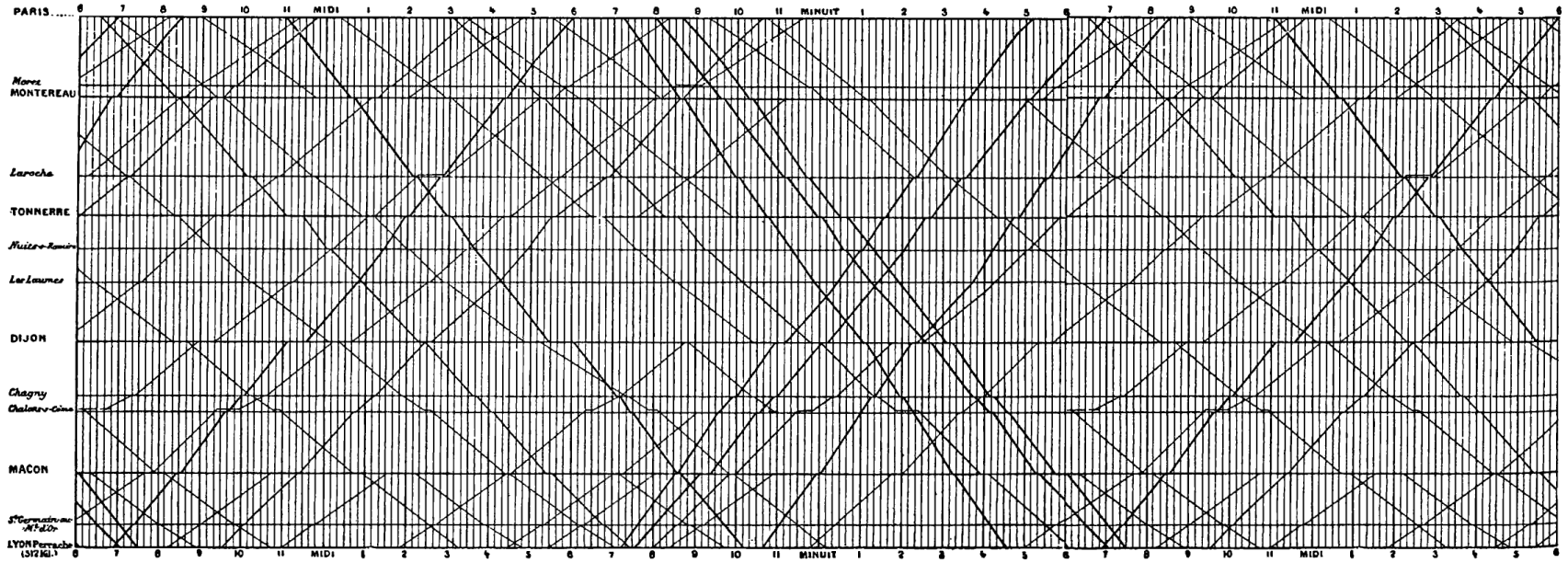
Can also serve a purpose in some cases

If there is a time dimension involved, to show a full circle for example

And, similarly, in map plots to go “once around the world”

> Redundancy, upon occasion, has its uses: giving a context and order to complexity, facilitating comparisons over various parts of the data, perhaps creating an aesthetic balance.


PARIS.....
6 7 8 9 10 11 MIDI 1 2 3 4 5 6 7 8 9 10 11 MINUIT 1 2 3 4 5 6
Morez
MONTEREAU
Laroche
TONNERRE
Nuits-s-Rouvre
Les Laumes
DIJON
Chagny
Chalons-s-Saone
MACON
St Germain-au-Mt d'Or
LYON Perrache
(512 Kl.)
6 7 8 9 10 11 MIDI 1 2 3 4 5 6 7 8 9 10 11 MINUIT 1 2 3 4 5 6


PARIS
Morez
MONTEREAU
Laroche
TONNERRE
Nuits-s-Rouvr
Les Laumes
DIJON
Chagny
Chalons-s-Saone
MACON
LYON Perrache
6 7 8 9 10 11 MIDI 1 2 3 4 5 6 7 8 9 10 11 MINUIT 1 2 3 4 5 6 7 8 9 10 11 MIDI 1 2 3 4 5 6

Most data representations, however, are of a single, uncomplicated number, and little graphical repetition is needed. Unless redundancy has a distinctly worthy purpose, the second erasing principle applies:

Erase redundant data-ink, within reason.

Redesigning

Some example ways of increasing the ink-data ratio:

Doing away with too much grid lines (or making them thinner)

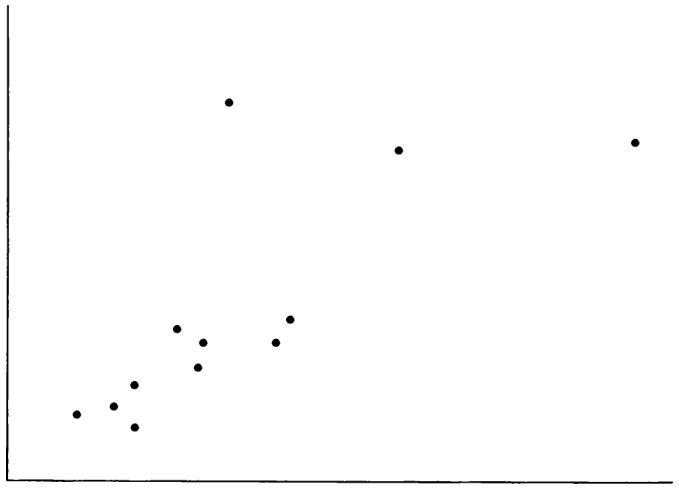
Not plotting axes beyond the data range

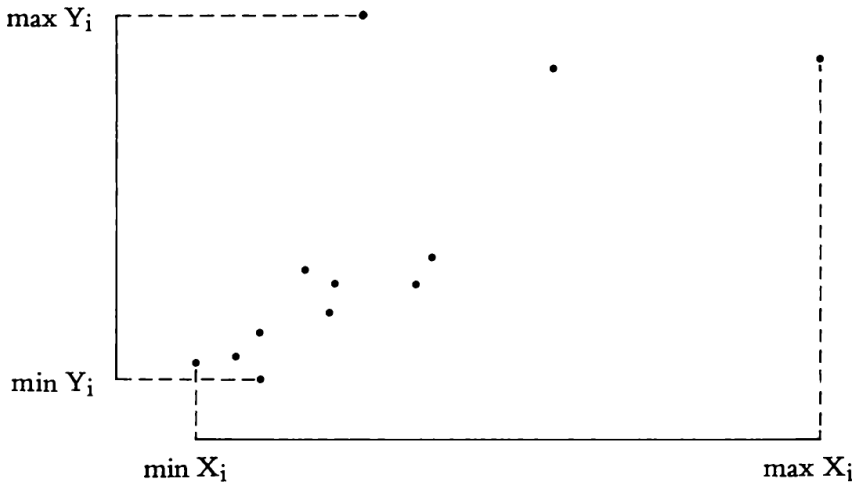
Not plotting unnecessarily many axis ticks

Do you need axes at all?

Can you integrate a legend onto the plot, should you need one?

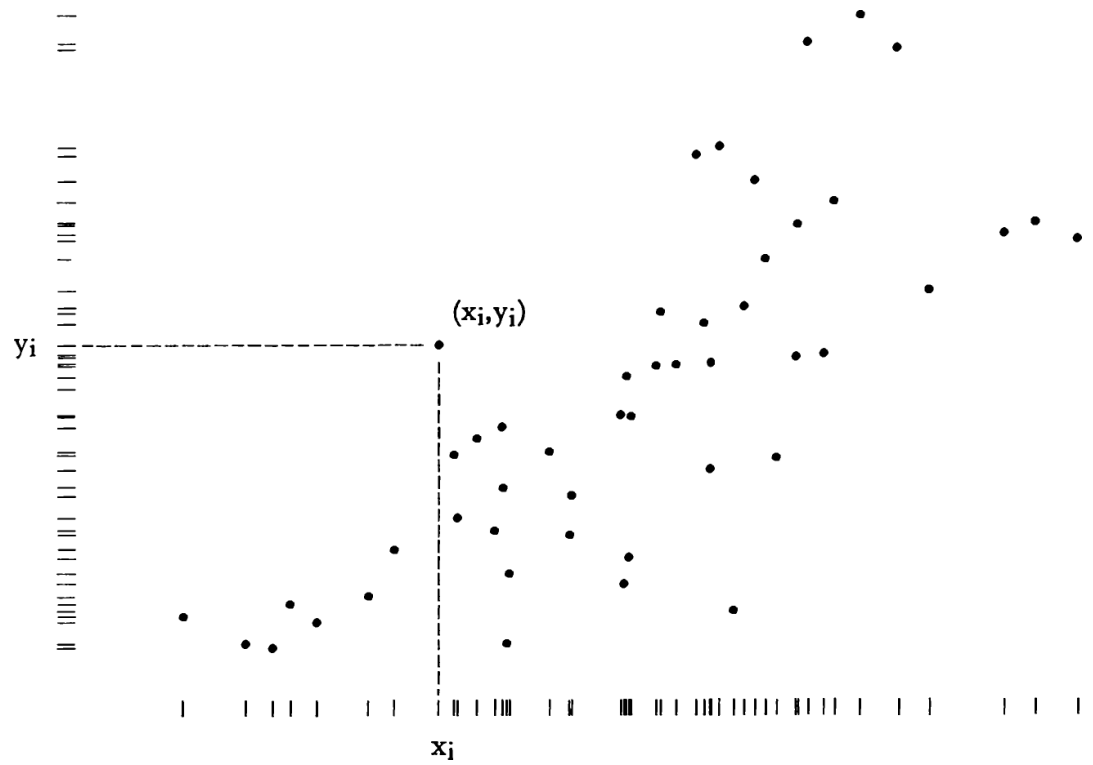
Can you have **multifunctioning graphical elements** in your plot?



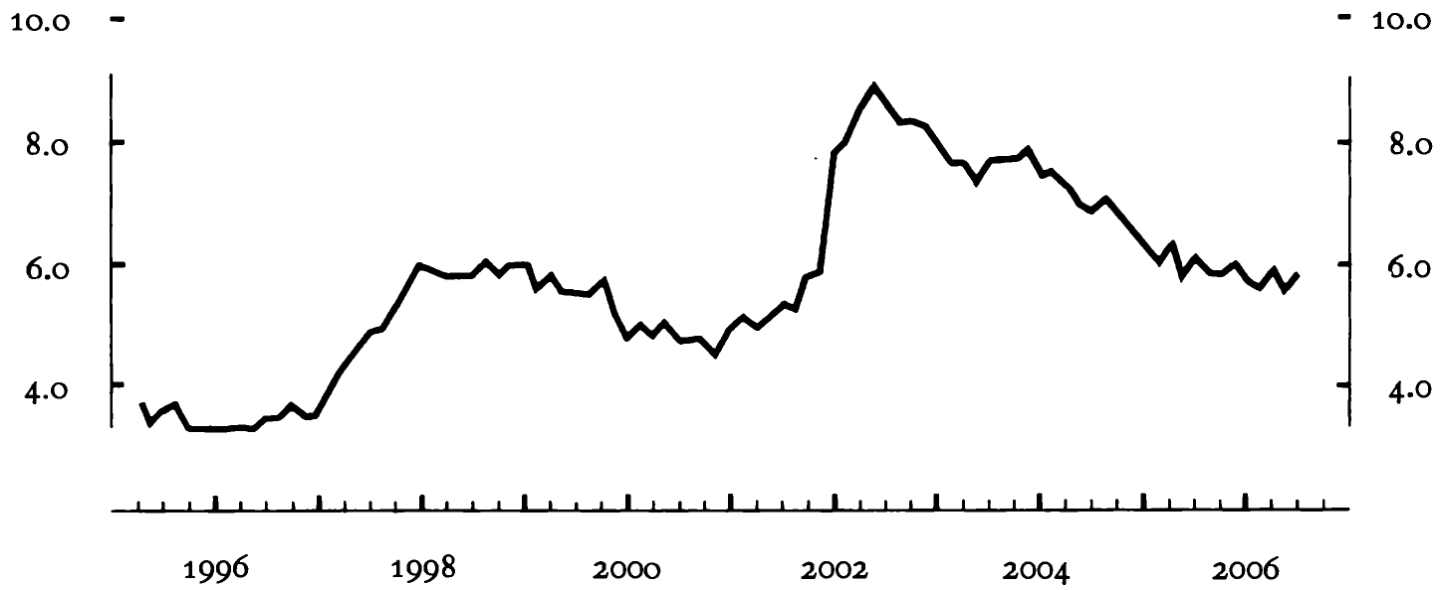

max Y_i
min Y_i
min X_i
max X_i

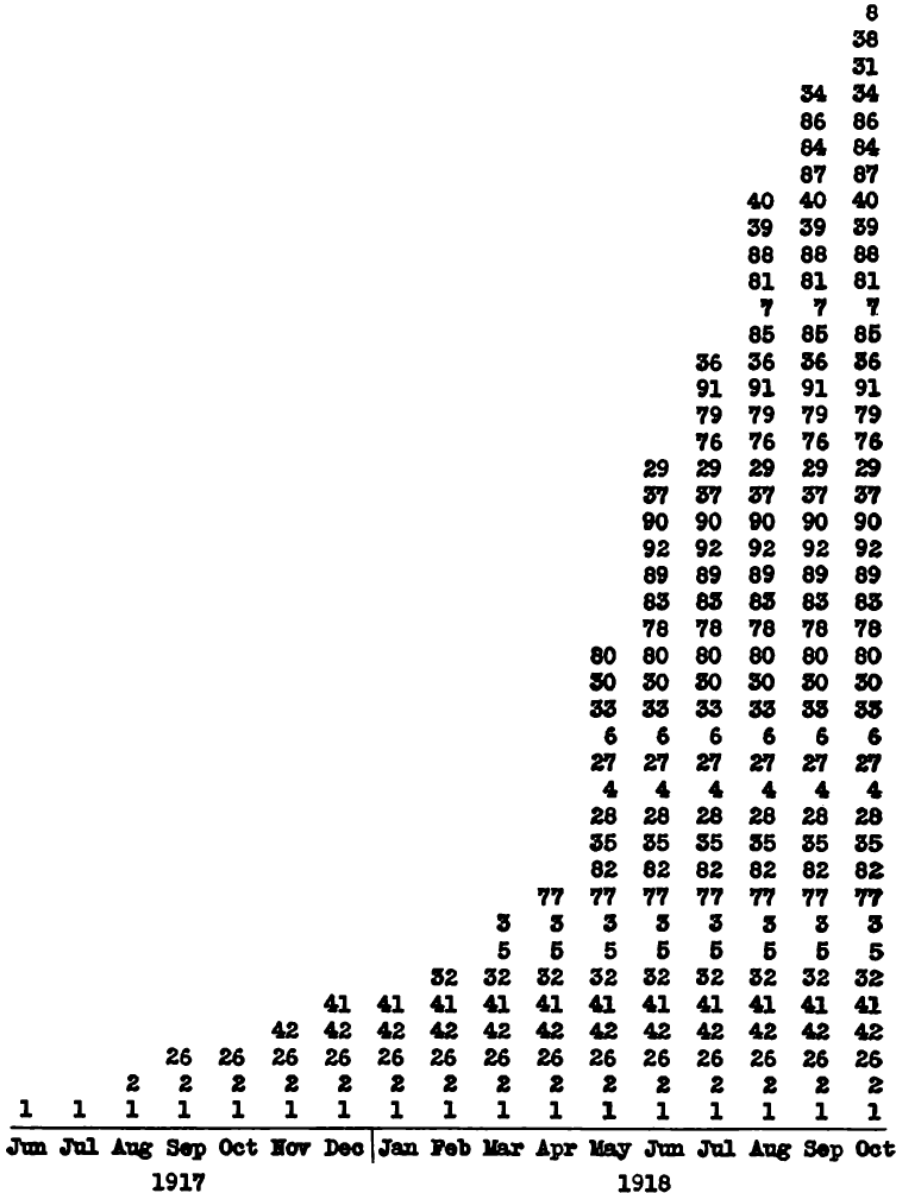
The result, a *range-frame*, explicitly shows the maximum and minimum of both variables plotted (along with the range), information available only by extrapolation and visual estimation in the conventional design.

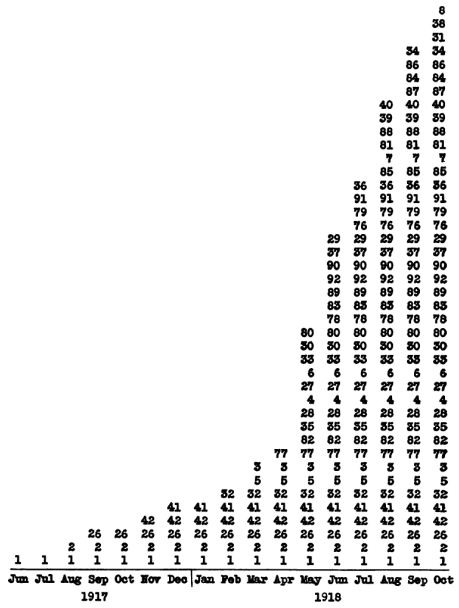
The data-ink ratio has increased: some non-data-ink has been erased, and the remainder of the frame, now carrying information, has gone over to the side of data-ink.



The dot-dash-plot combines the two fundamental graphical designs used in statistical analysis, the marginal frequency distribution and the bivariate distribution. Dot-dash-plots make routine what good data analysts do already—plotting marginal and joint distributions together.


10.0
8.0
6.0
4.0
10.0
8.0
6.0
4.0
1996
1998
2000
2002
2004
2006


Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct
1917
1918



The triple-functioning data measure shows: (1) the number of divisions in France for each month, June 1917 to October 1918; (2) what particular divisions were in France in each month; and (3) the duration of each division's presence in France.

Above all else show the data.

Maximize the data-ink ratio.

Erase non-data-ink.

Erase redundant data-ink.

Revise and edit.

Data Density in Graphical Practice

The numbers that go into a graphic can be organized into a data matrix of observations by variables. Taking into account the size of the graphic in relation to the amount of data displayed yields the *data density*:

$$\text{data density of a graphic} = \frac{\text{number of entries in data matrix}}{\text{area of data graphic}}$$

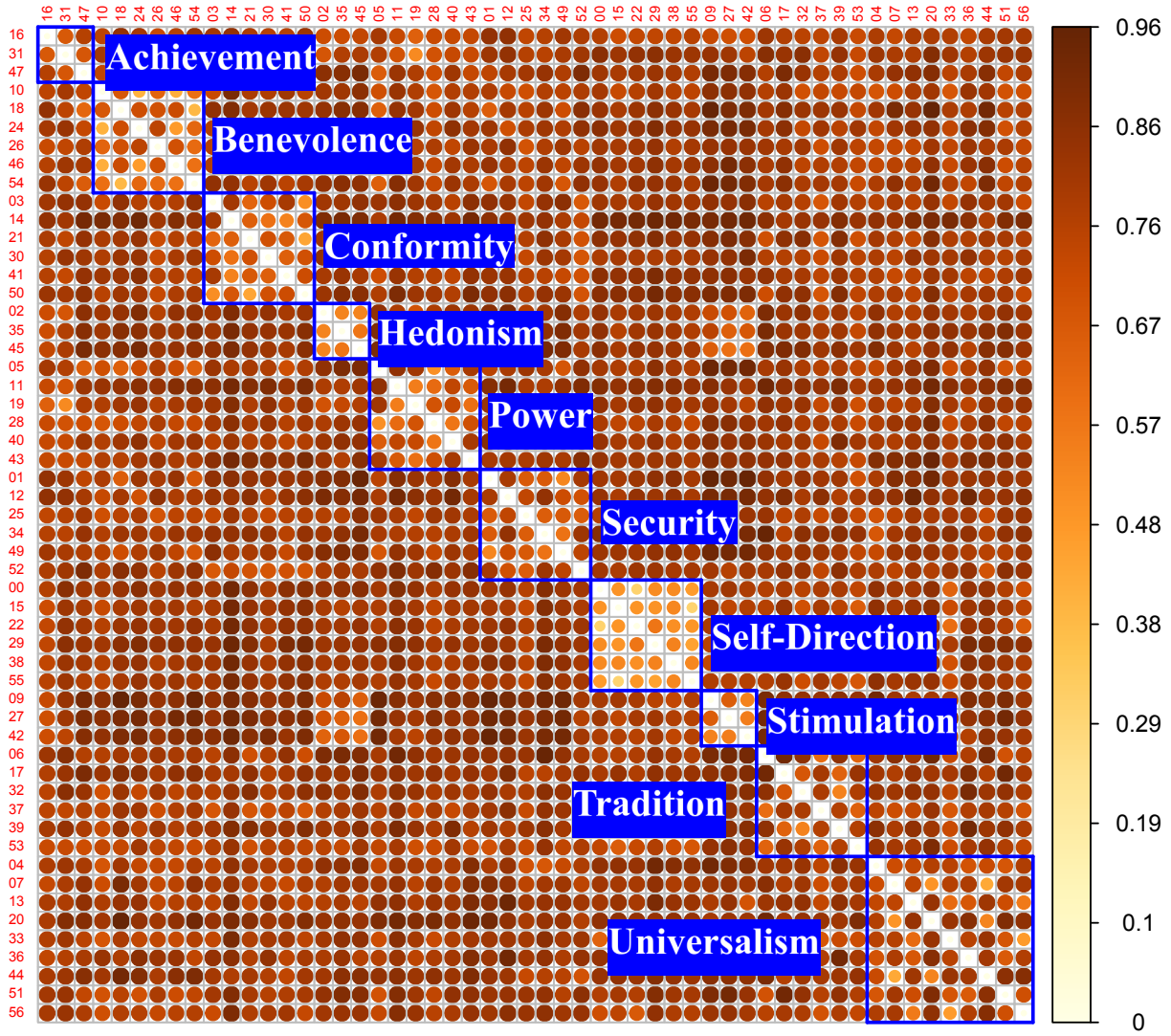
Data Density

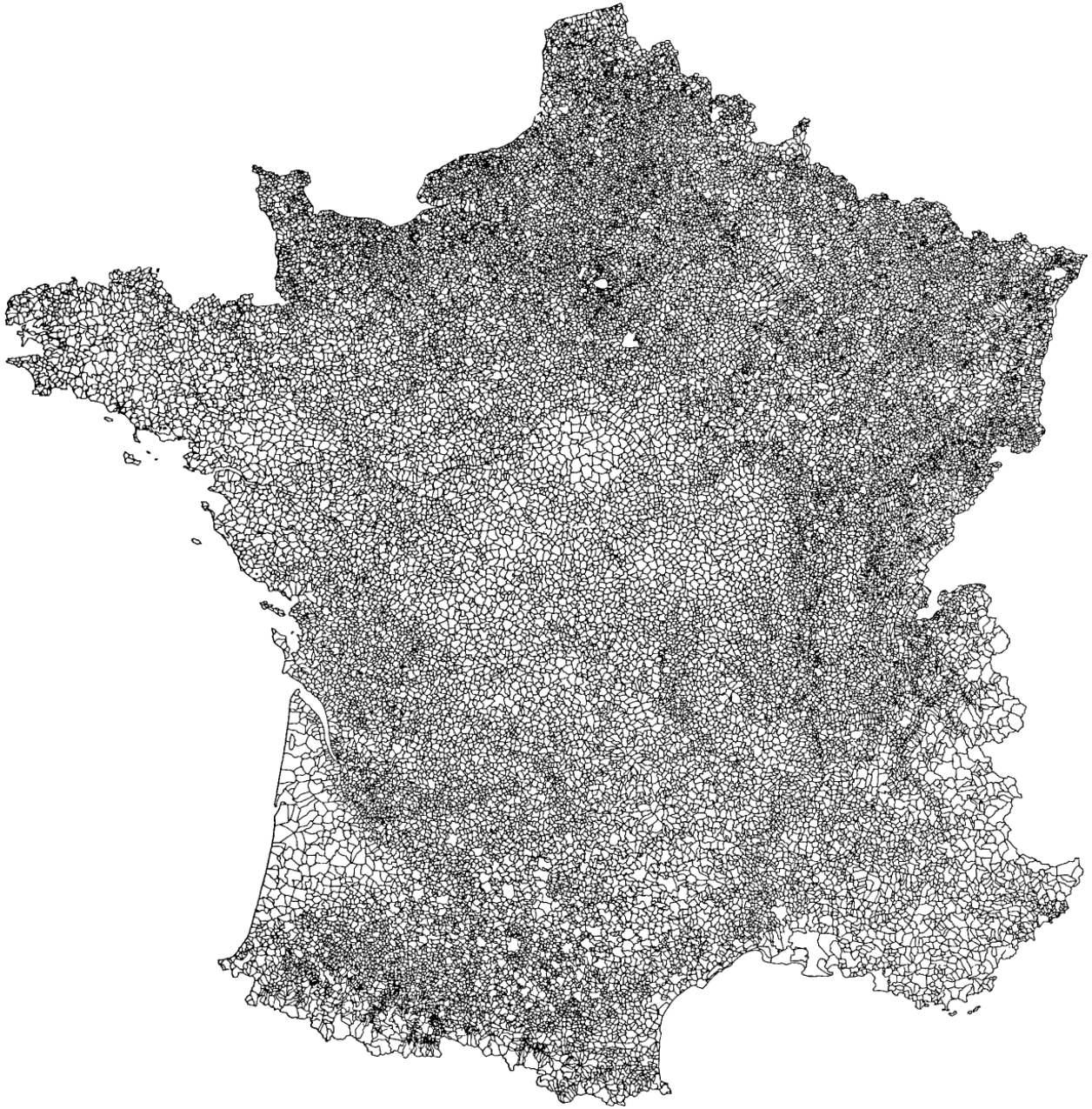
Simple charts often have a (very) low data density

Consider for example a bar chart with only two classes with only one value each (4 entries in total) that takes up considerable space when visualized

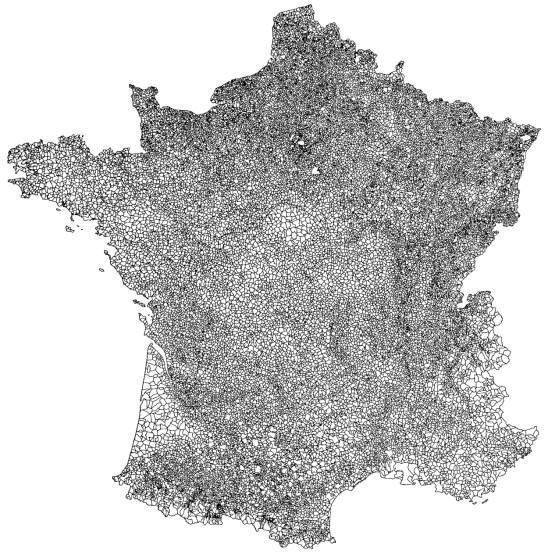
Maps, on the other hand, usually have a very high data density

Or other area plots, for example direct visualizations of matrices


Achievement
Benevolence
Conformity
Hedonism
Power
Security
Self-Direction
Stimulation
Tradition
Universalism
0.96
0.86
0.76
0.67
0.57
0.48
0.38
0.29
0.19
0.1
0



UNIVERSITY OF MANNHEIM

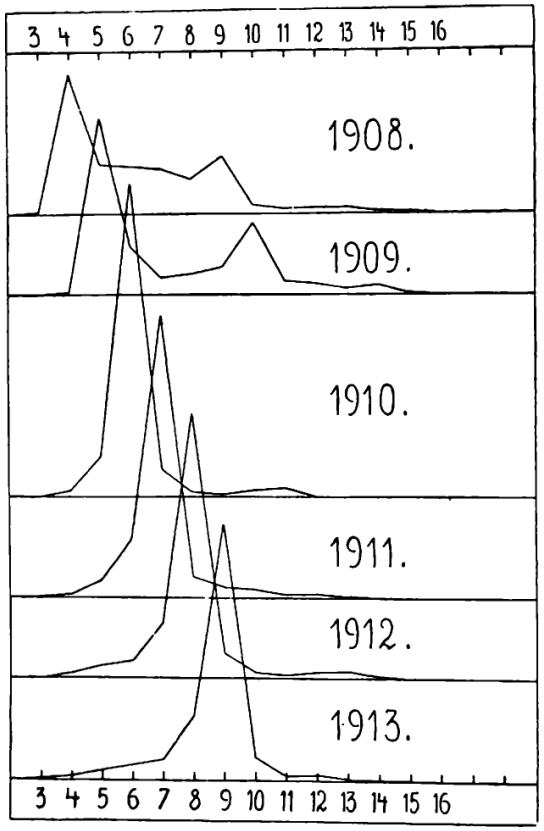


This map (27 square inches, 175 square centimeters) shows the location and boundaries of 30,000 communes of France. It would require at least 240,000 numbers to recreate the data of the map (30,000 latitudes, 30,000 longitudes, and perhaps six numbers describing the shape of each commune). Thus that data density is nearly 9,000 numbers per square inch, or 1,400 numbers per square centimeter.

Maximize data density and the size of the data matrix, within reason.

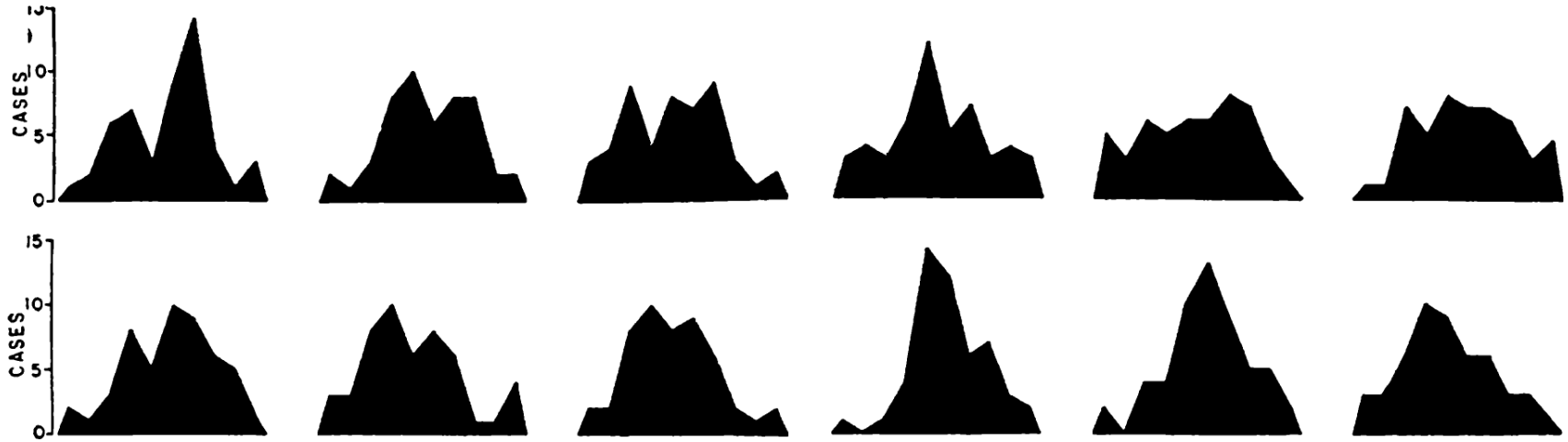
Small Multiples (aka facet, panel, lattice, grid or trellis charts)

Small multiples resemble the frames of a movie: a series of graphics, showing the same combination of variables, indexed by changes in another variable.



These six distributions show the age composition of herring catches each year from 1908 to 1913. A tremendous number of herring were spawned in 1904, and that class began to dominate the 1908 catch as four-year-olds, then the 1909 catch as five-year-olds, and so on.

The effects of sampling errors are shown in these 12 distributions, each based on a sample of 50 random normal deviates:



At the heart of quantitative reasoning is a single question: *Compared to what?* Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives. For a wide range of problems in data presentation, small multiples are the best design solution.

Well-designed small multiples are

- Inevitably comparative
- Deftly multivariate
- Shrunken, high-density graphics
- Usually based on a large data matrix
- Drawn almost entirely with data-ink
- Efficient in interpretation
- Often narrative in content, showing shifts in the relationship between variables as the index variable changes (thereby revealing interaction or multiplicative effects)

Small multiples reflect much of the theory of data graphics:

For non-data-ink, less is more.

For data-ink, less is a bore.