Visualization of Experimental Data

Paul G. Plöger

SEE

WS 2013/14

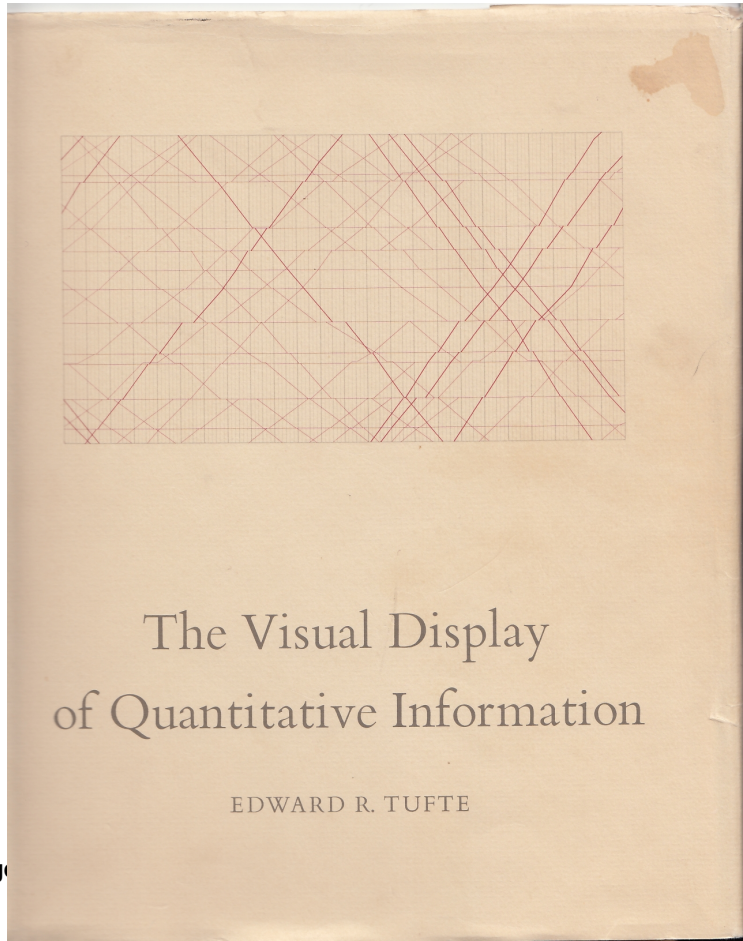
Overview

1d/2d Data, functions, plots, tables, scatter

Time series

Many dimendional data

New diagram styles



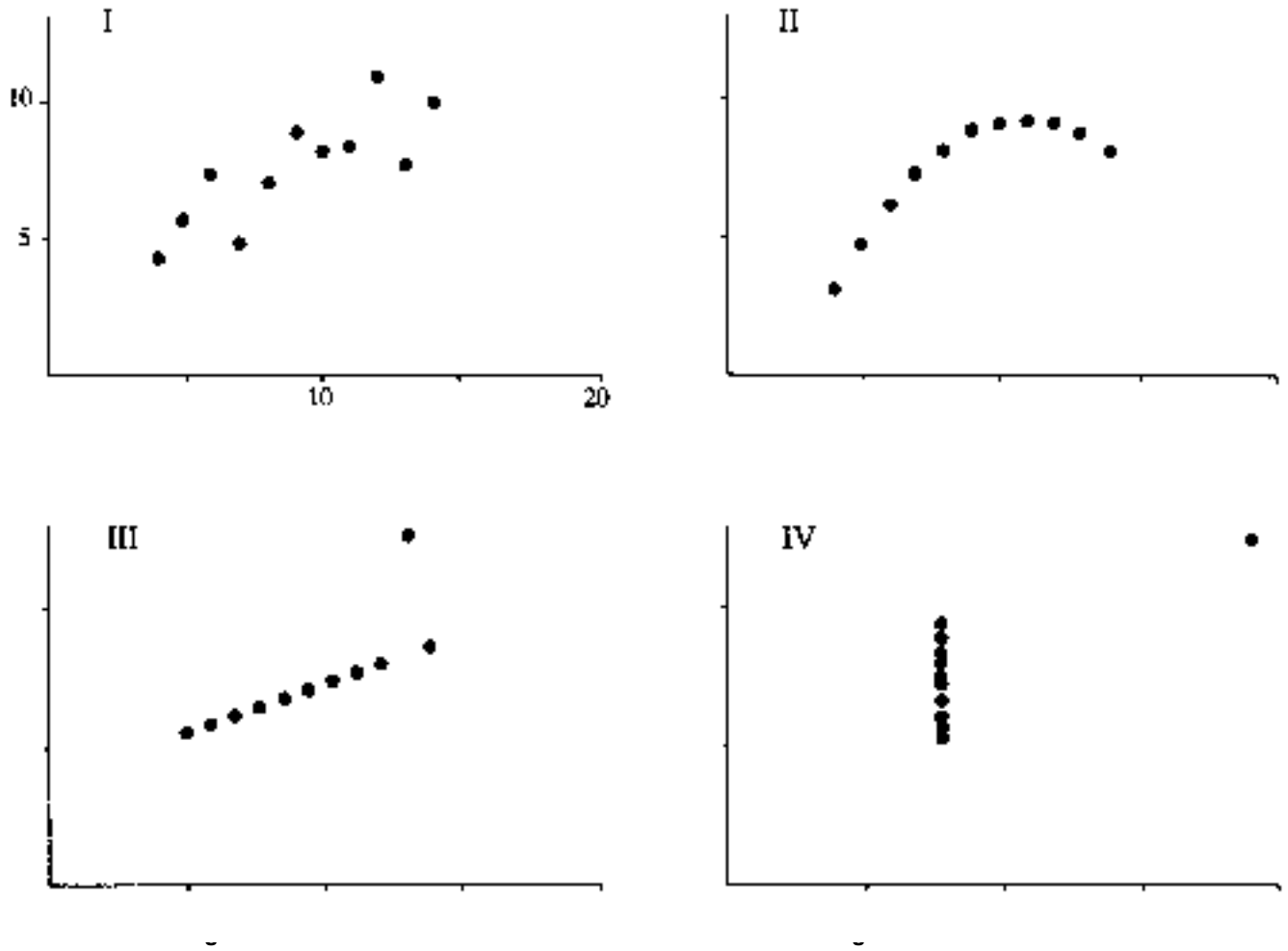
Intro Example

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

A=[
10.0000 8.0400 10.0000 9.1400 10.0000 7.4600 8.0000 6.5800
8.0000 6.9500 8.0000 8.1400 8.0000 6.7700 8.0000 5.7600
13.0000 7.5800 13.0000 8.7400 13.0000 12.7400 8.0000 7.7100
9.0000 8.8100 9.0000 8.7700 9.0000 7.1100 8.0000 8.8400
11.0000 8.3300 11.0000 9.2600 11.0000 7.8100 8.0000 8.4700
14.0000 9.9600 14.0000 8.1000 14.0000 8.8400 8.0000 7.0400
6.0000 7.2400 6.0000 6.1300 6.0000 6.0800 8.0000 5.2500
4.0000 4.2600 4.0000 3.1000 4.0000 5.3900 19.0000 12.5000
12.0000 10.8400 12.0000 9.1300 12.0000 8.1500 8.0000 5.5600
7.0000 4.8200 7.0000 7.2600 7.0000 6.4200 8.0000 7.9100
5.0000 5.6800 5.0000 4.7400 5.0000 5.7300 8.0000 6.8900]

N=11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: Y = 3 +0.5X
standard error of estimate of slope = 0.118
t = 4.24
sum of squares $X - X^{--}$ = 110.0
Regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
r2 = .67

Graphics reveal data!



What Graphical displays should:

show the data

induce the viewer to **think about the substance** rather than about methodology, graphic design, the technology of graphic production, or something else

avoid distorting what the data have to say

present **many numbers** in a small space

make **large data sets coherent**

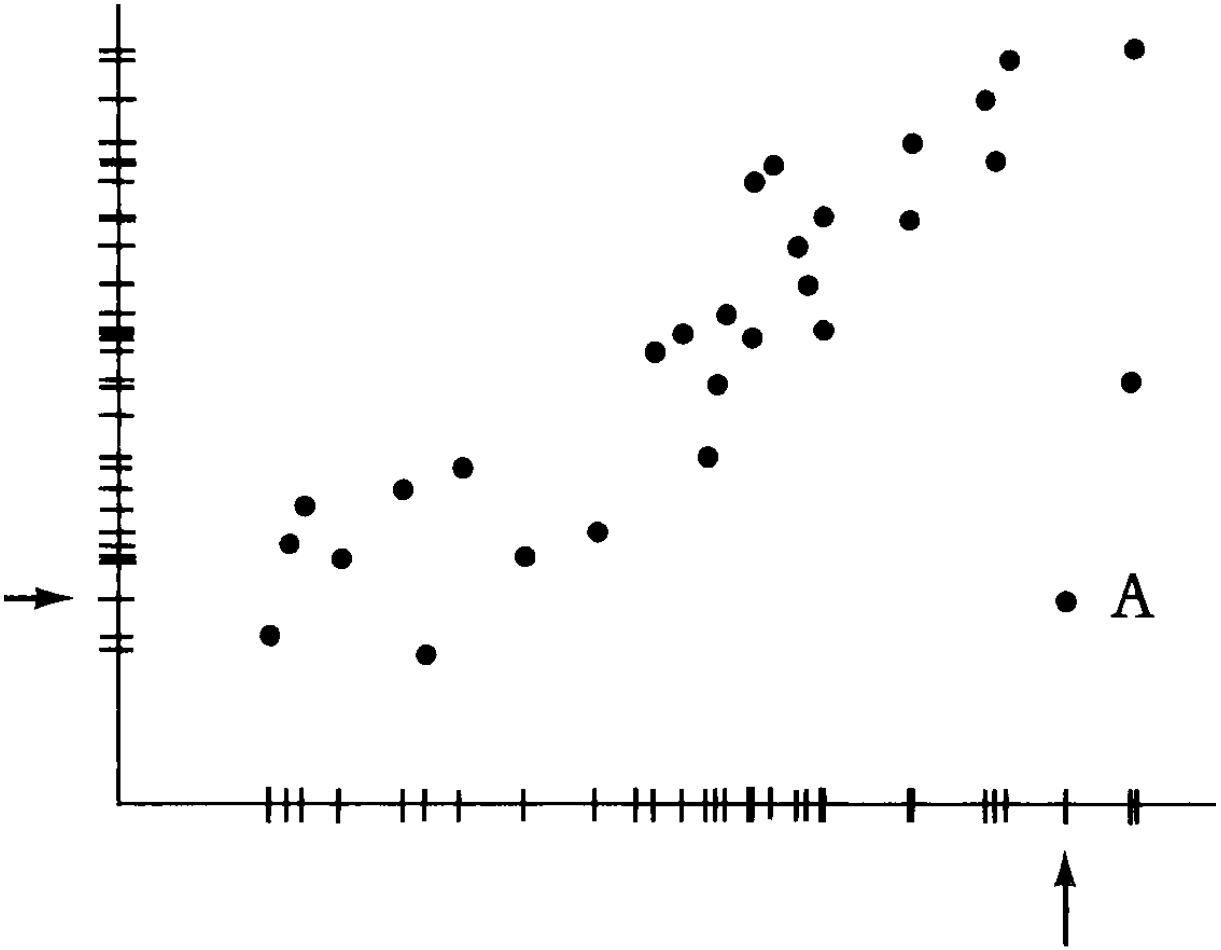
encourage the eye to **compare different pieces** of data

reveal the **data at several levels of detail**, from a broad overview to the fine structure

serve a reasonably **clear purpose**: description, exploration, tabulation, or decoration

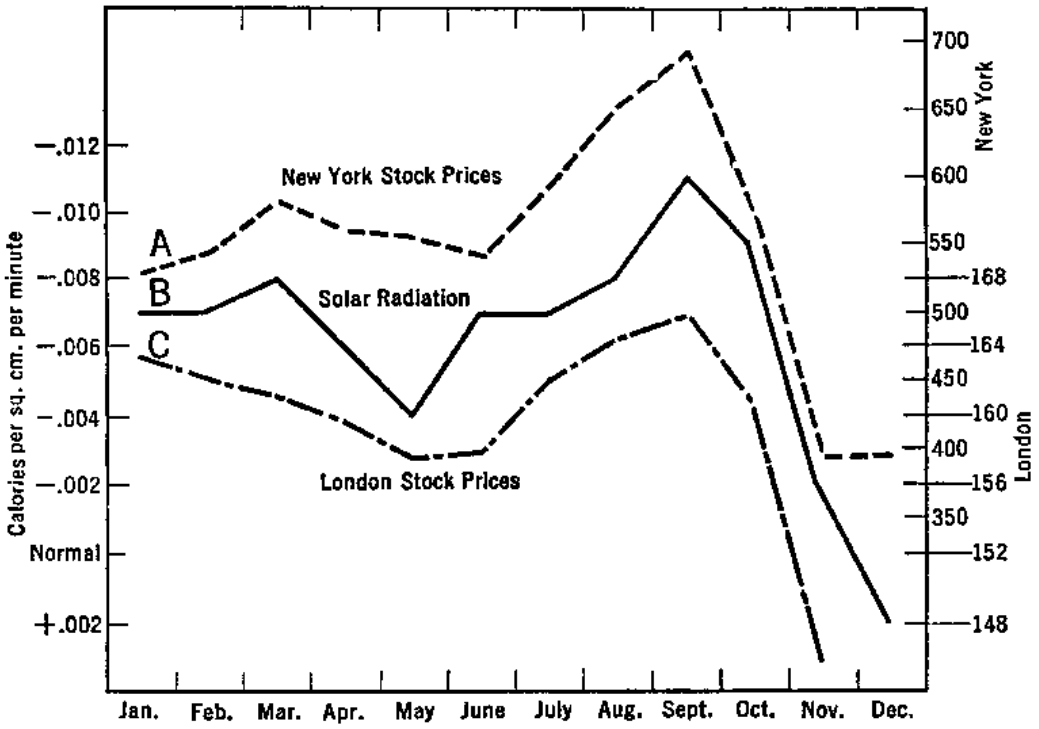
be **closely integrated with the statistical** and **verbal** descriptions of a data set.

Outlier detection



And likewise a graphic easily reveals point A, a wildshot observation that will dominate standard statistical calculations. Note that point A hides in the marginal distribution but shows up as clearly exceptional in the bivariate scatter.

Non sensical:
A silly theory means a silly graphic



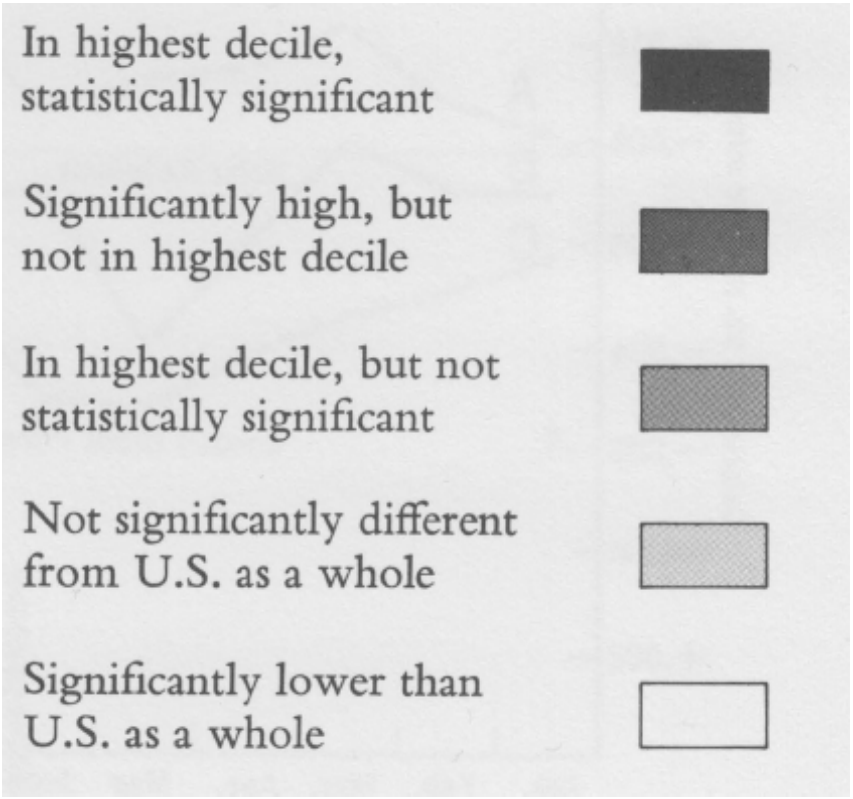
SOLAR RADIATION AND STOCK PRICES

A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

Of course, statistical graphics are only as good as what goes into them.

An ill-specified or preposterous model or a puny data set cannot be rescued by a graphic.

Data Maps

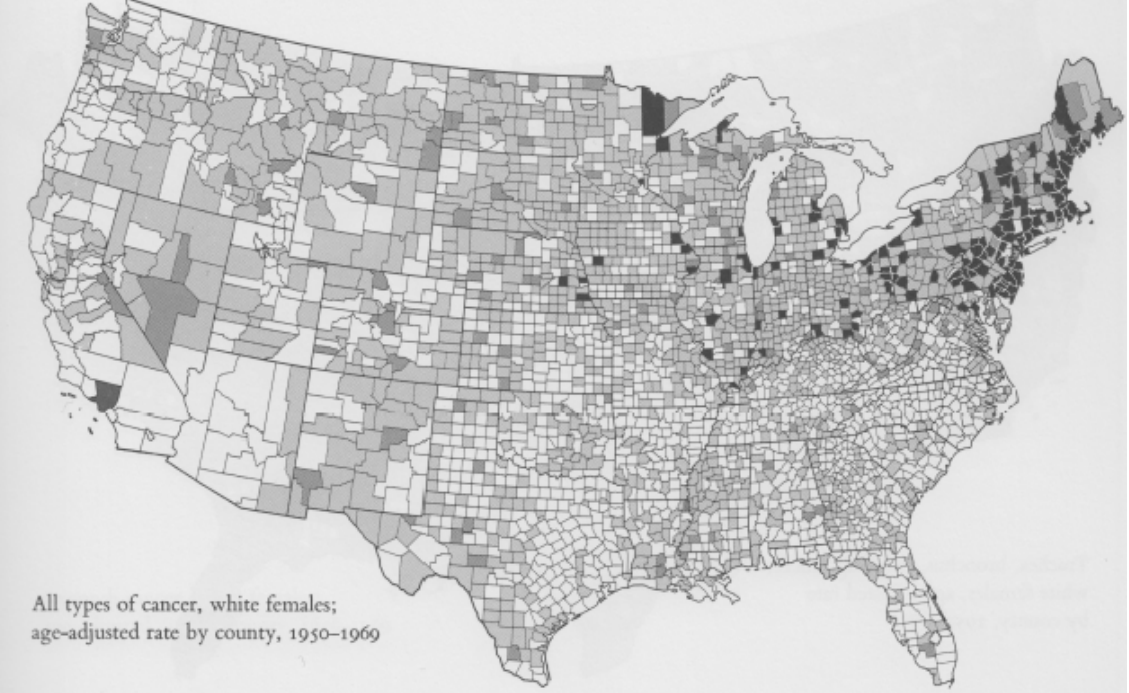


All types of cancer, white **females**; age-adjusted rate by county, 1950-1969

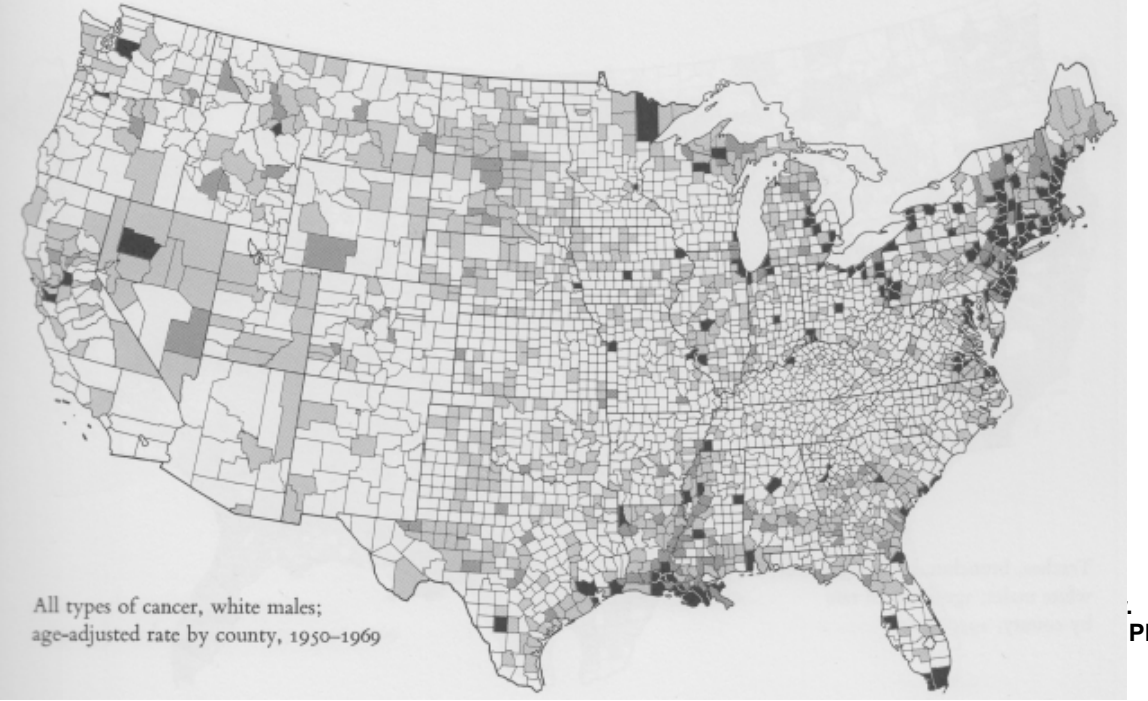
All types of cancer, white **males;** age-adjusted rate by county, 1950-1969

Trachea, bronchus, and lung cancer; white **females** / **males**; age-adjusted rate by county, 1950-1969

Stomach cancer **f**/**m**

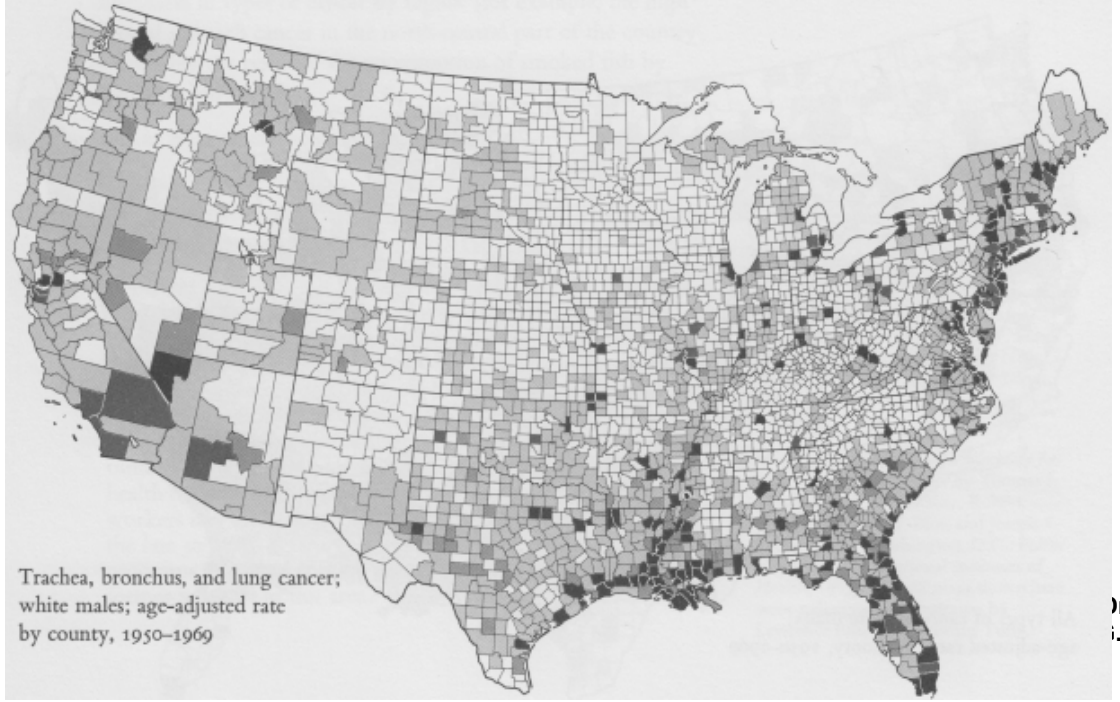


All types of cancer, white females;
age-adjusted rate by county, 1950–1969

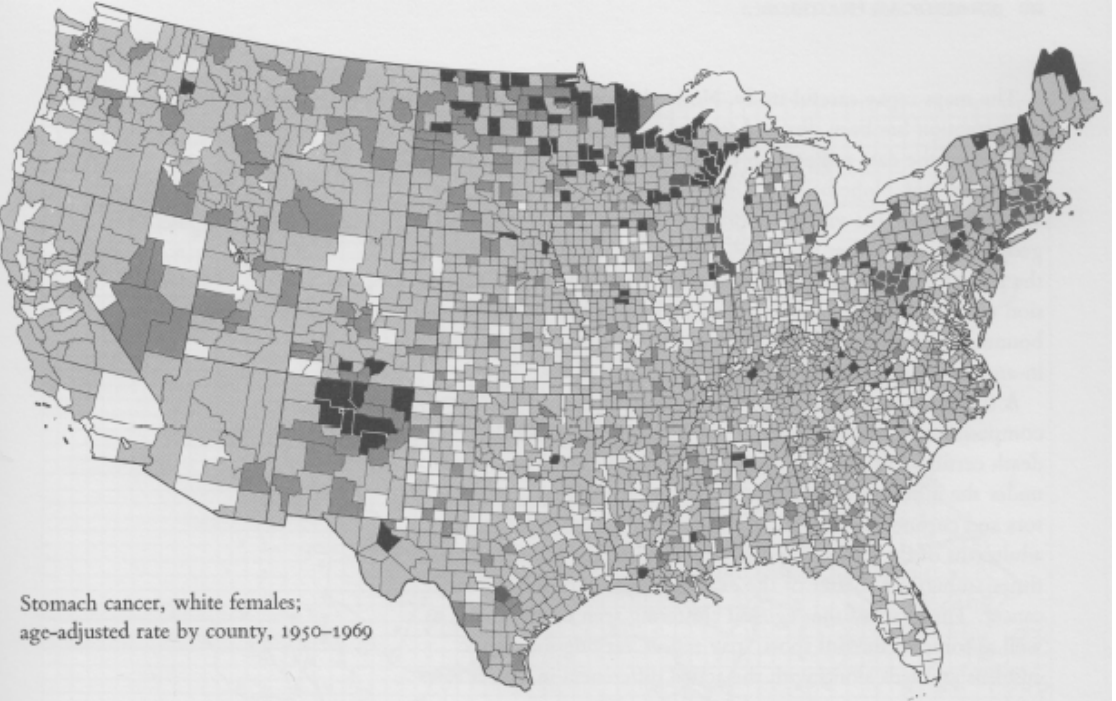


All types of cancer, white males;
age-adjusted rate by county, 1950–1969

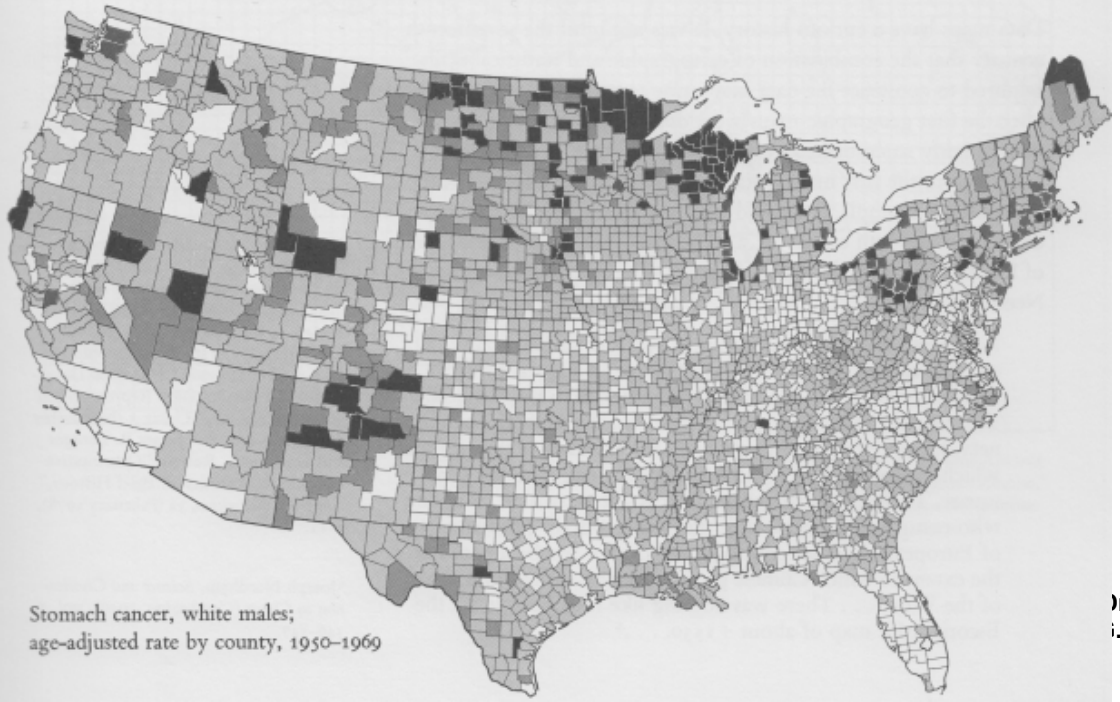
Trachea, bronchus, and lung cancer; white females; age-adjusted rate by county, 1950–1969



Trachea, bronchus, and lung cancer; white males; age-adjusted rate by county, 1950–1969



Stomach cancer, white females;
age-adjusted rate by county, 1950–1969



Stomach cancer, white males;
age-adjusted rate by county, 1950–1969

The Maps provide leads into the causes - and avoidance

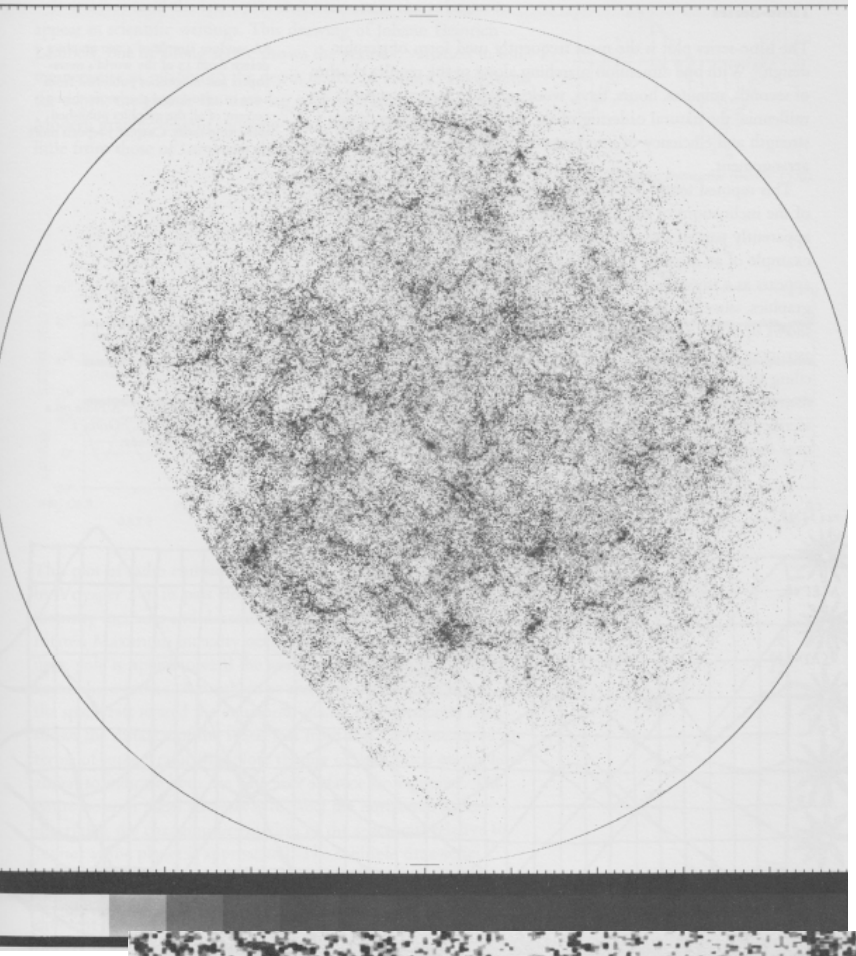
high death rates from cancer in the **northeast part** of the country and **around the Great Lakes**

low rates in an **east-west band across the middle of the country**

higher rates for men than for women in the south, particularly Louisiana (cancers probably caused by occupational exposure, from working with asbestos in shipyards)

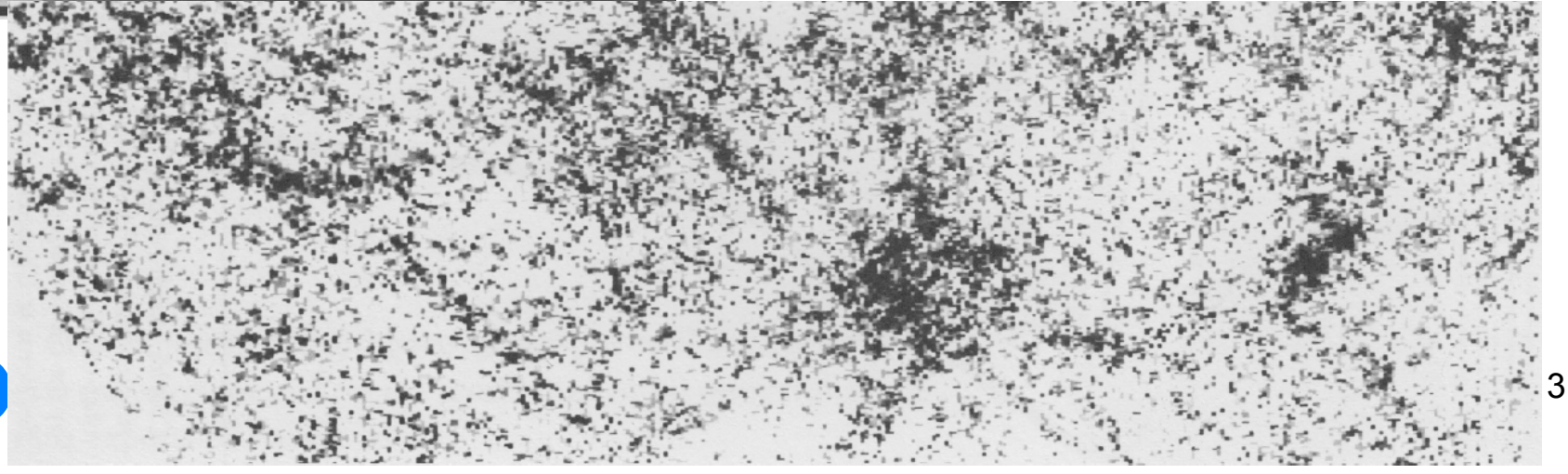
unusual hot spots, including **northern Minnesota** and a few counties **in Iowa and Nebraska** along the Missouri River

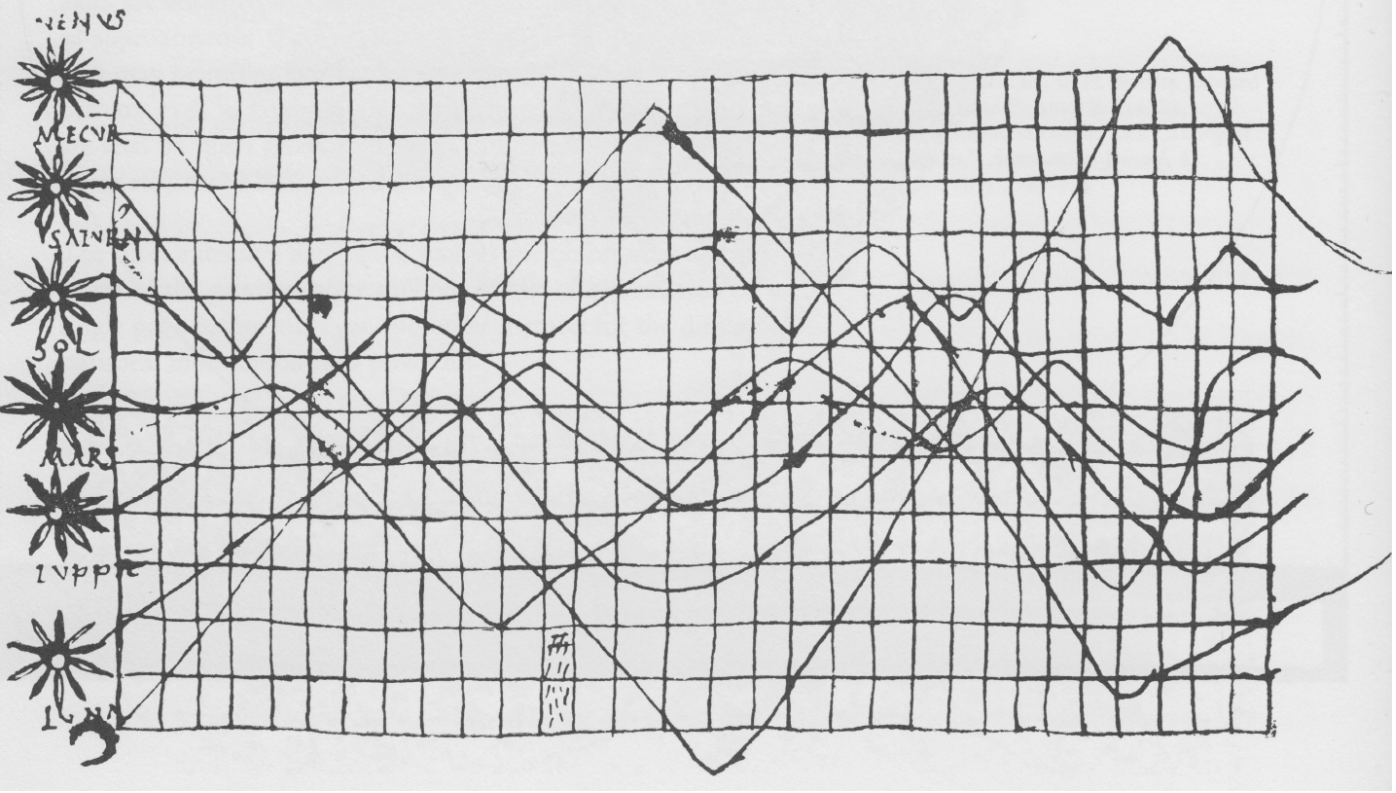
differences in types of cancer by region (for example, the high rates of stomach cancer in the **north-central part of the country** -probably the result of the consumption of smoked fish by Scandinavians)



Even more data in one single map

- distribution of 1.3 million galaxies
- divides the sky into 1,024 X 2,222 rectangles
- The number of galaxies counted in each of the 2,275,328 rectangles is represented by ten gray tones; the darker the tone, the greater the number of galaxies counted.
- The north galactic pole is at the center



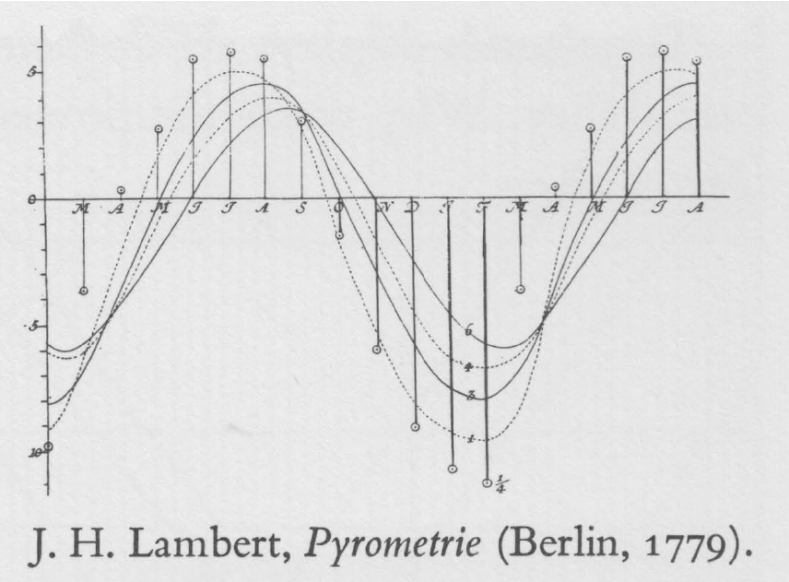


Time Series

A random sample of 4,000 graphics drawn from 15 of the world's newspapers and magazines published from 1974 to 1980 found that more than 75 % of all the graphics published were time-series.

Aboves reputed tenth-century illustration of the inclinations of the planetary orbits as a function of time, apparently part of a text for monastery schools, is the oldest known example of an attempt to show changing values graphically.

Gray Funkhouser, "A Note on a Tenth Century Graph," Osiris, 1 (January 1936), 260-262.

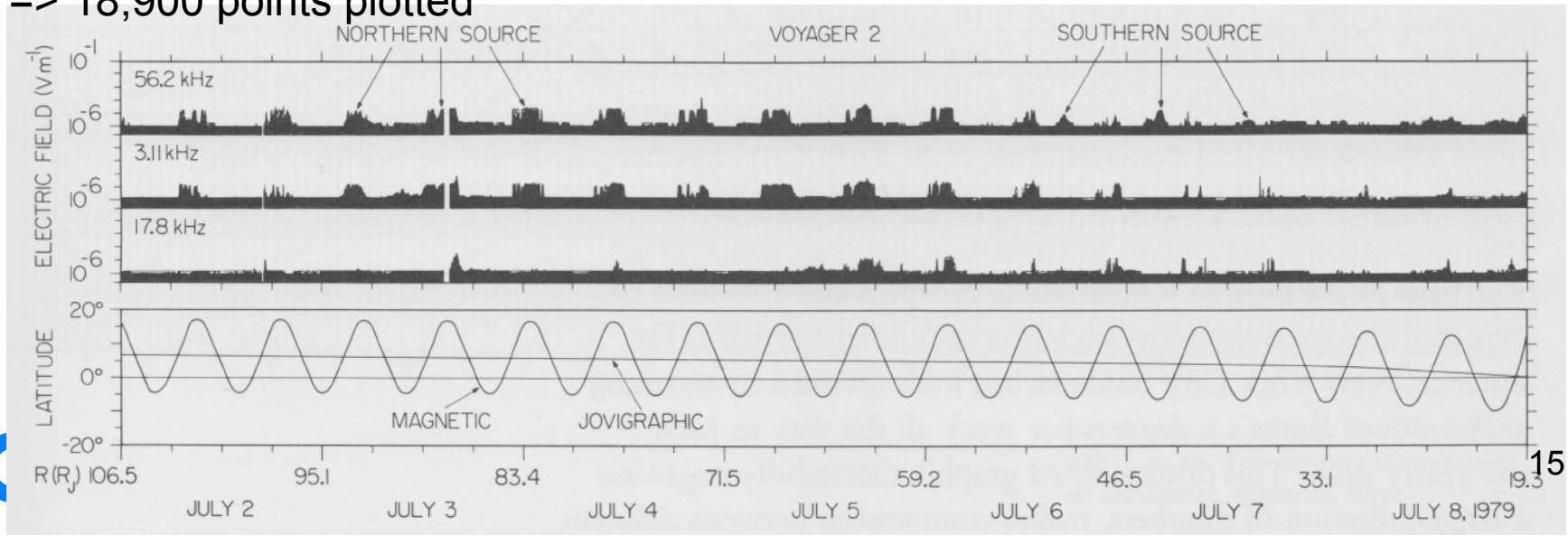


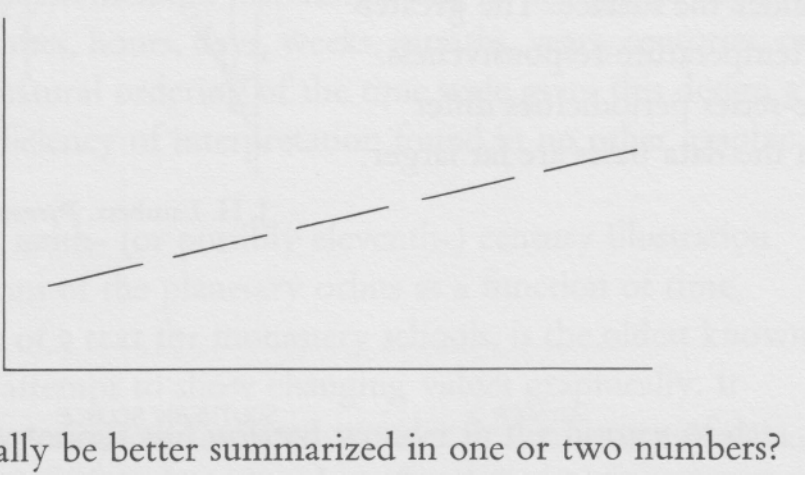
J. H. Lambert, *Pyrometrie* (Berlin, 1779).

More Time Series

This plot of radio emissions from Jupiter is based on data collected by Voyager 2 in its pass close by the planet in July 1979. The radio intensity increases and decreases in a ten-hour cycle as Jupiter rotates. Maximum intensity occurs when the Jovian north magnetic pole is tipped toward the spacecraft

453,600 samples / 8 bits.
=> 3.6 million bits reduced by peak and average processing
=> 18,900 points plotted

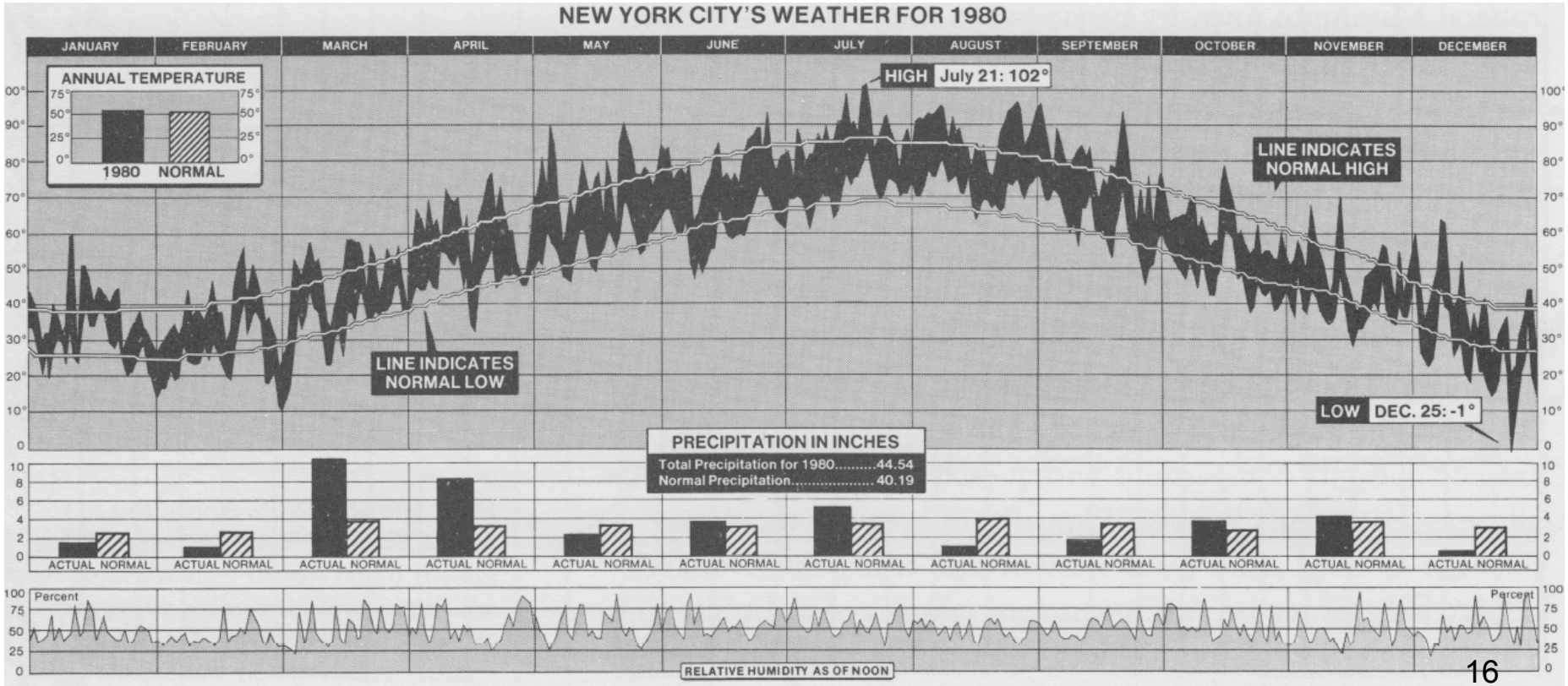




ally be better summarized in one or two numbers?

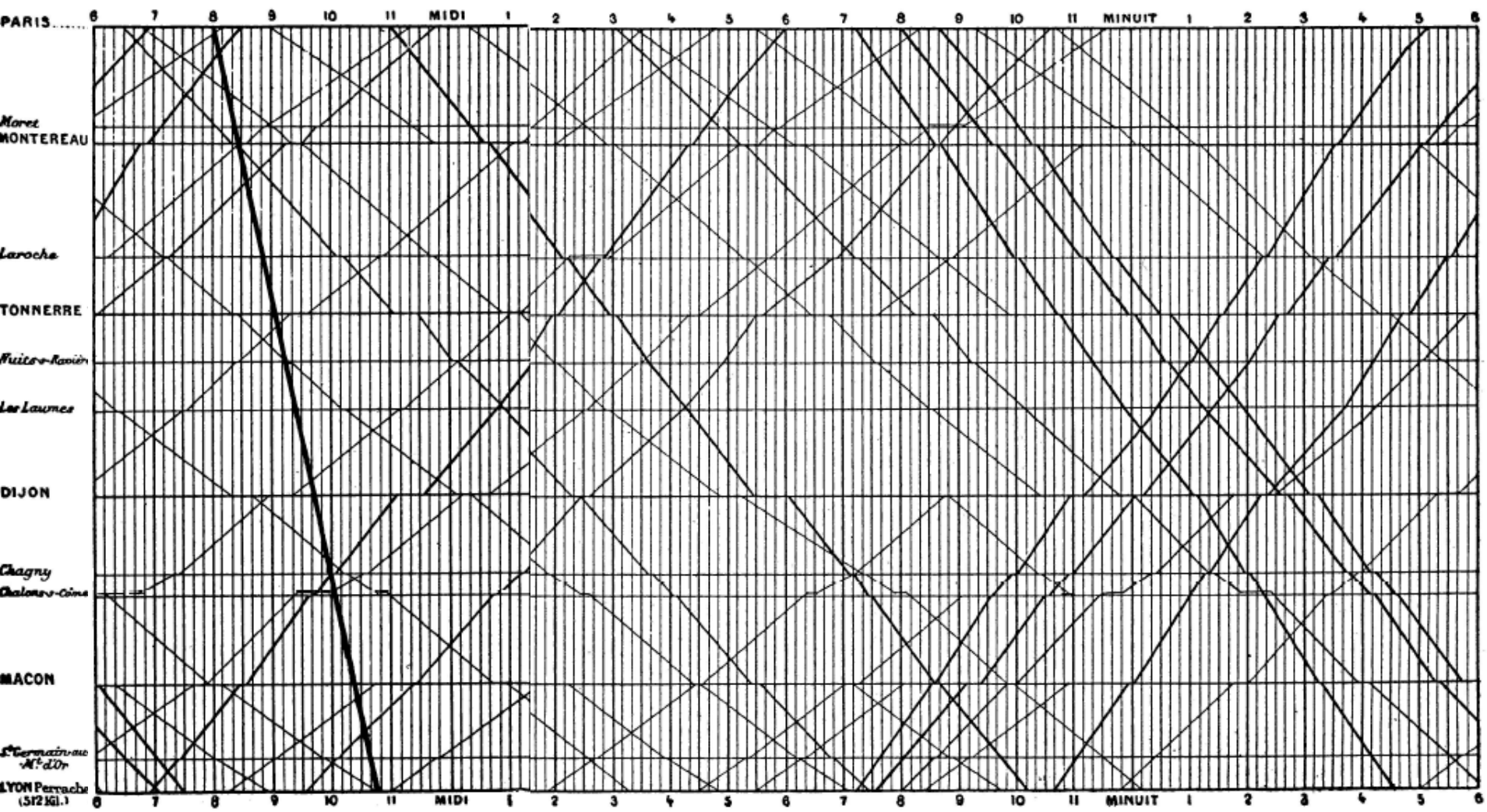
Time-series displays are at their best for big data sets with real variability. Dont waste the power of data graphics on simple linear changes,

No goes versus tell a story



New York Times, January 11, 1981, p. 32.

Train schedule



Principles of Graphical Excellence

Graphical excellence:

- is the **well-designed** presentation of **interesting data** - a matter of substance, of statistics, and of design.
- consists of **complex ideas** communicated with **clarity, precision, and efficiency**.
- is that which gives to the viewer the greatest number of ideas in the shortest time **with the least ink** in the **smallest space**.
- is nearly always **multivariate**.
  And graphical excellence requires telling the truth about the data.

How to reach this goal?

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

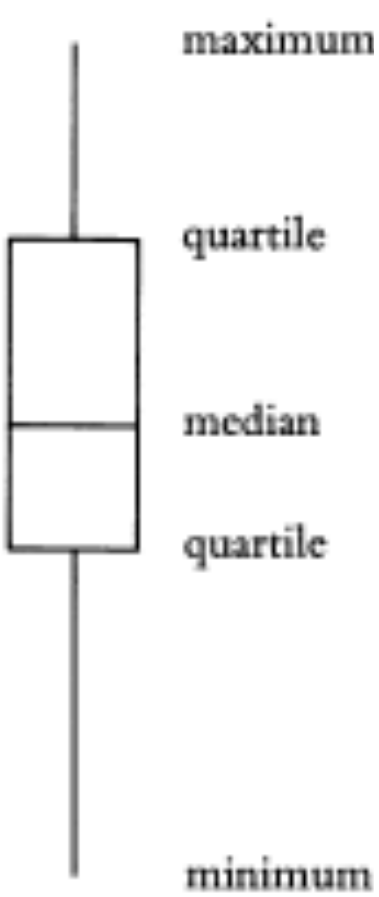
$$= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information}$$

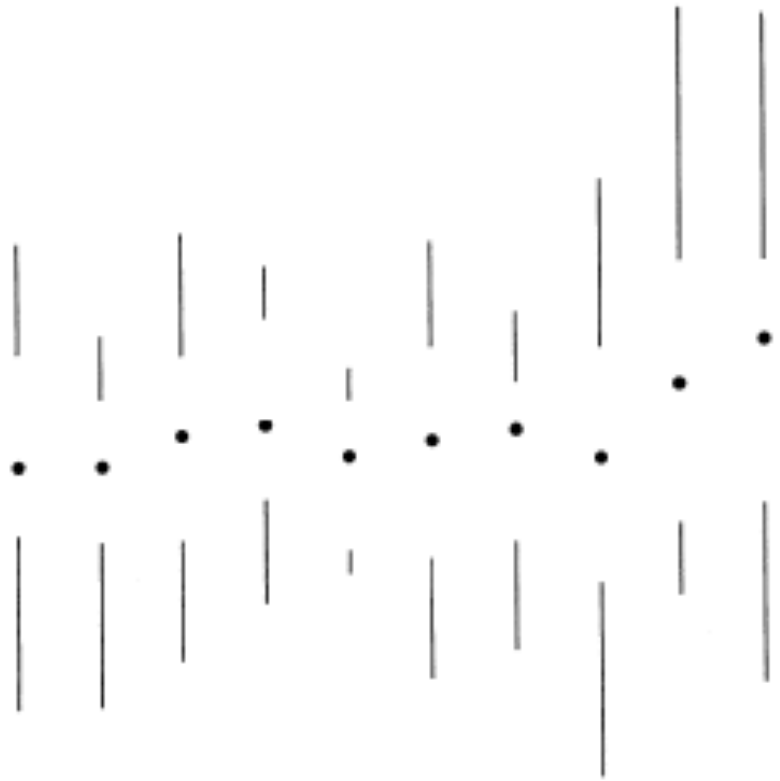
This improvement in graphical design illustrates the fundamental principle of good statistical graphics:

Above all else show the data.

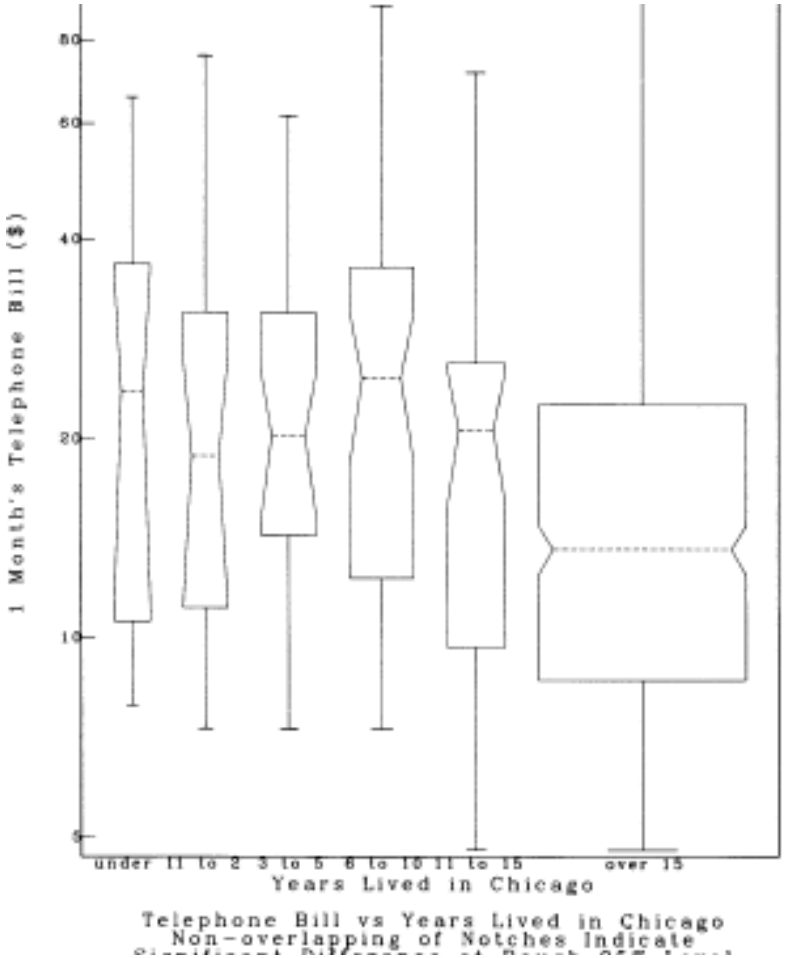
The principle is the basis for a theory of data graphics.

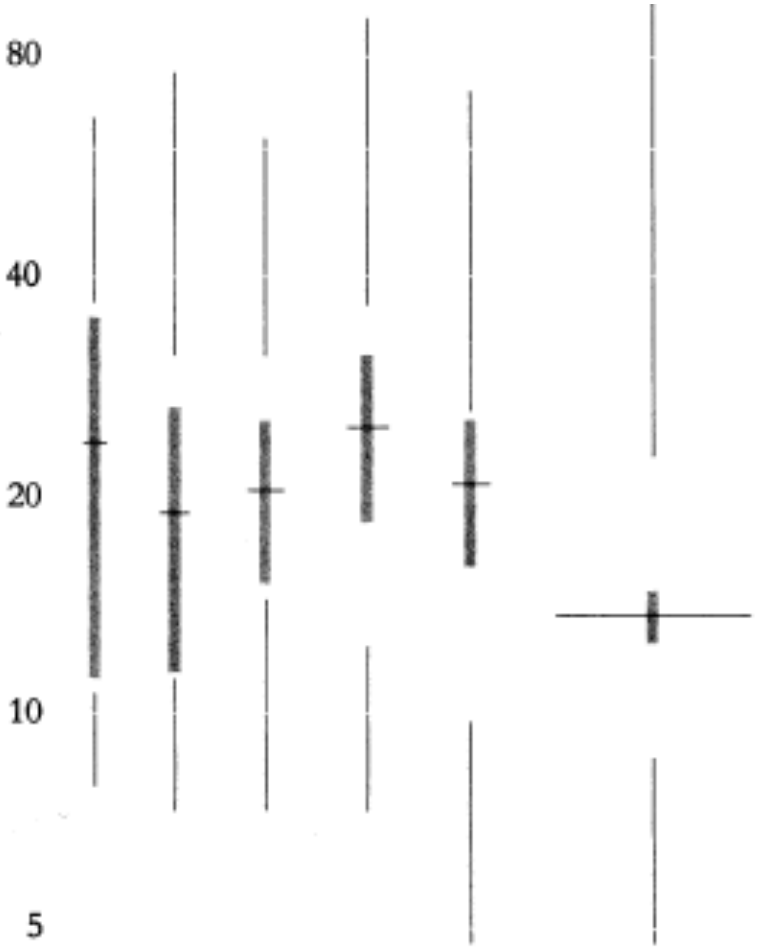
Example



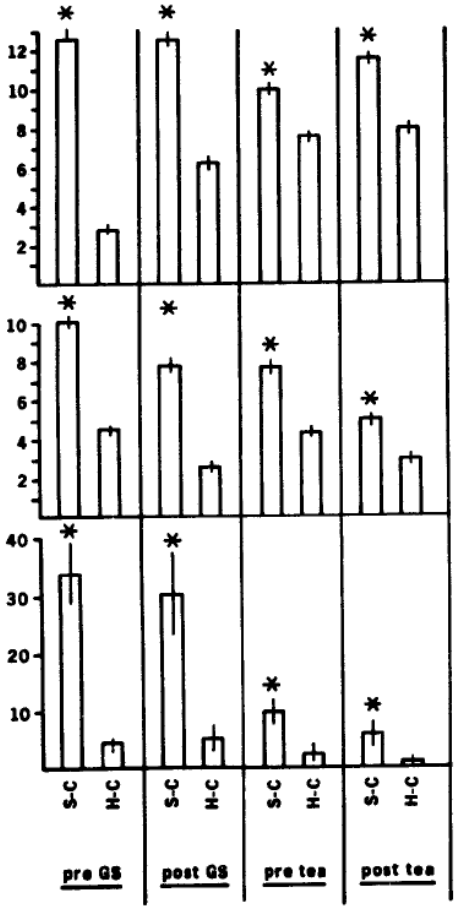


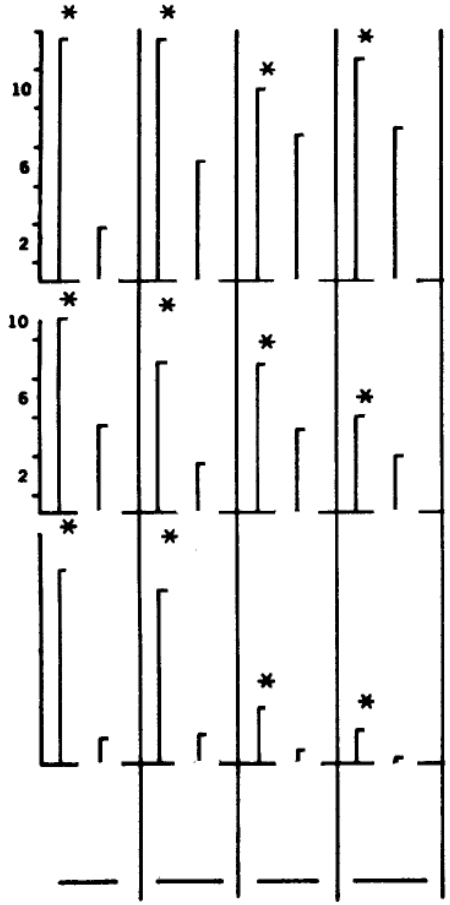
Example

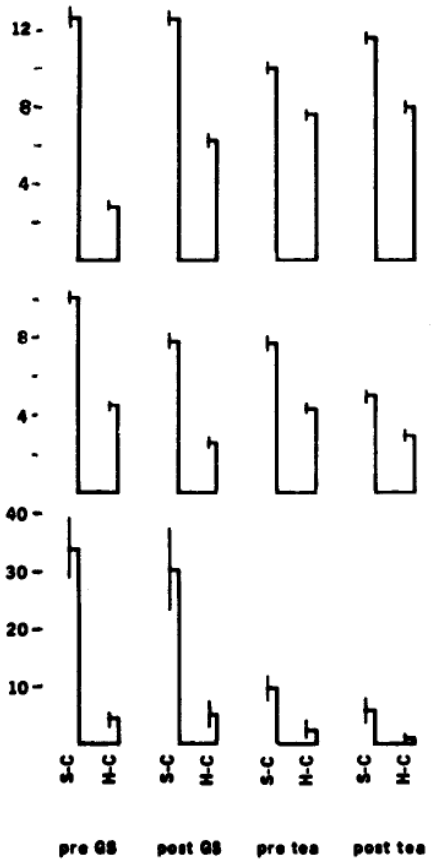

1 Month's Telephone Bill ($)
80
60
40
20
10
5
under 1 1 to 2 3 to 5 6 to 10 11 to 15 over 15
Years Lived in Chicago
Telephone Bill vs Years Lived in Chicago
Non-overlapping of Notches Indicate


80
40
20
10
5

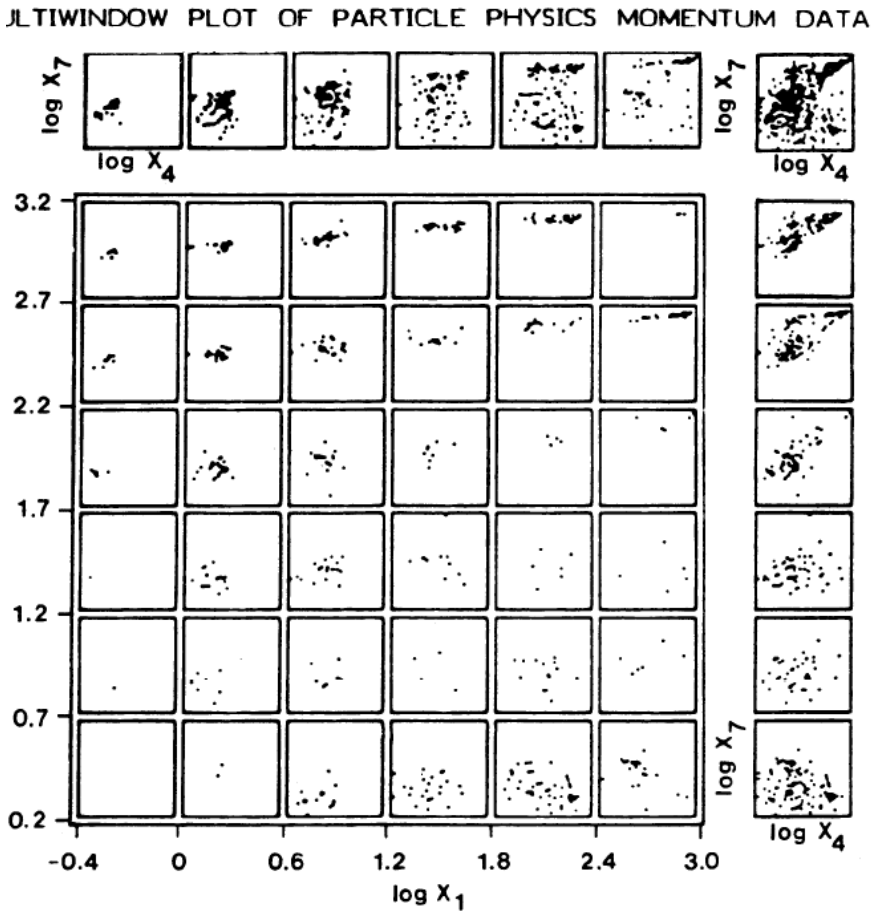
Example

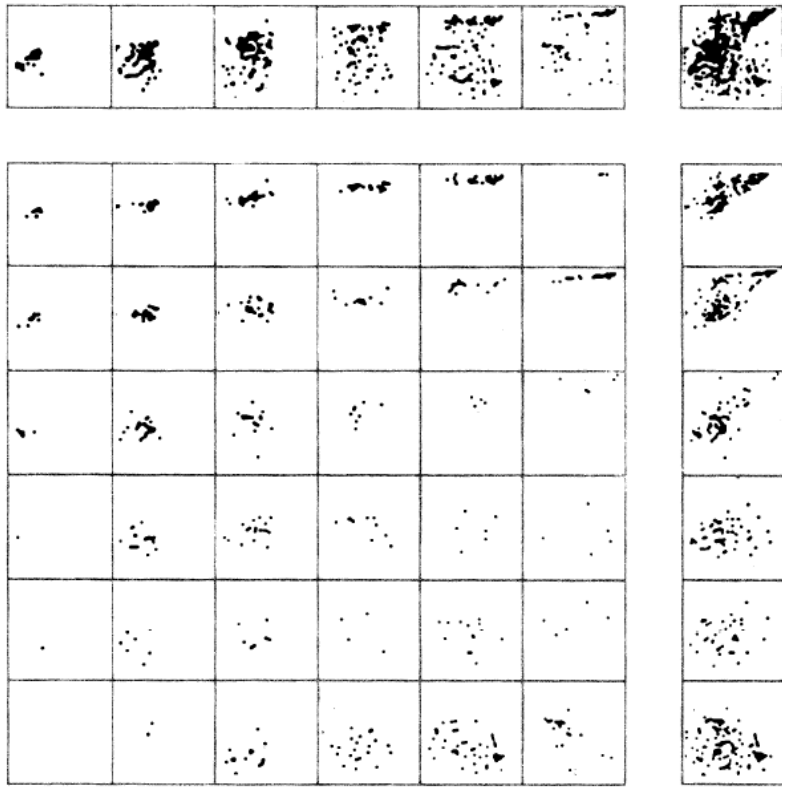

S-C
H-C
pre GS
post GS
pre tea
post tea



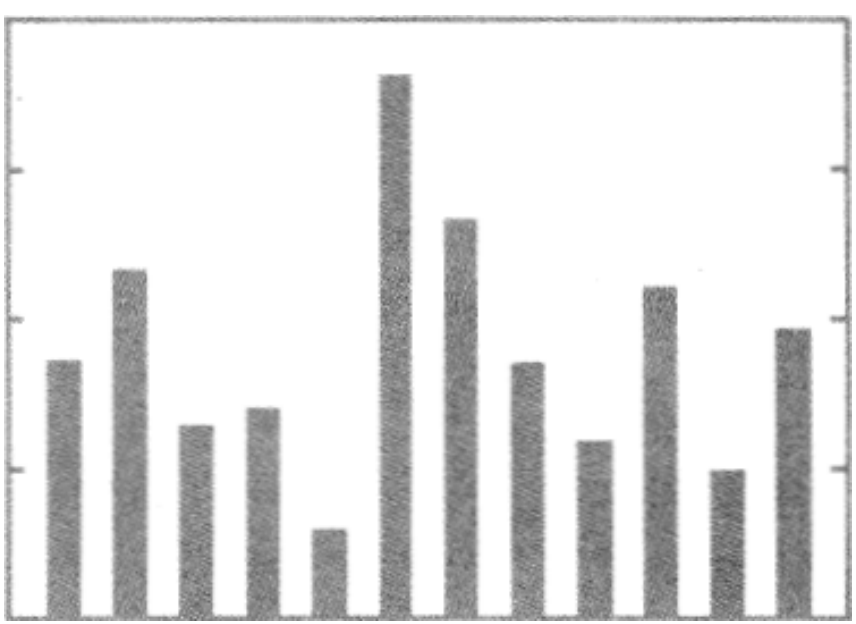

S-C
H-C
pre GS
post GS
pre tea
post tea

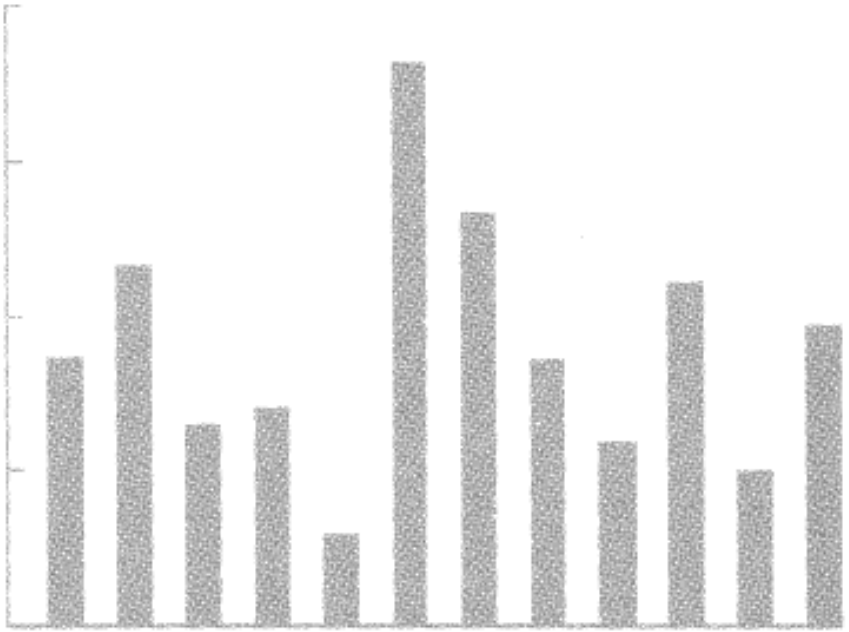
Example


JLTIWINDOW PLOT OF PARTICLE PHYSICS MOMENTUM DATA
log X7
log X4
3.2
2.7
2.2
1.7
1.2
0.7
0.2
-0.4
0
0.6
1.2
1.8
2.4
3.0
log X1

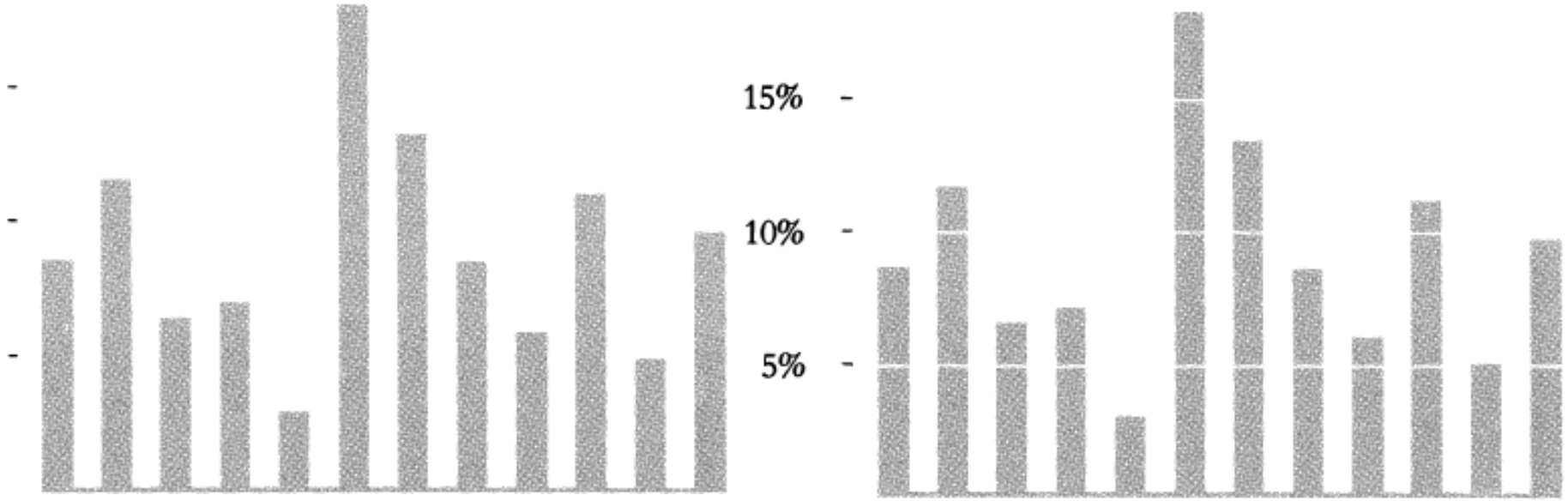


Redesign Bar Cart

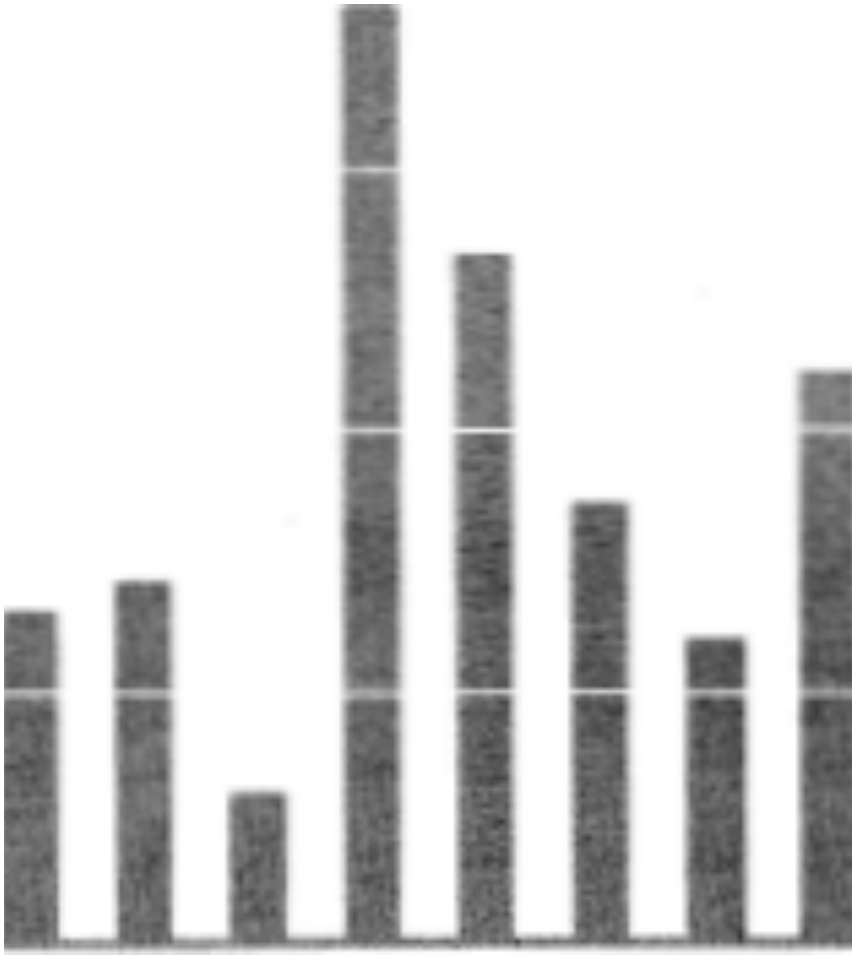




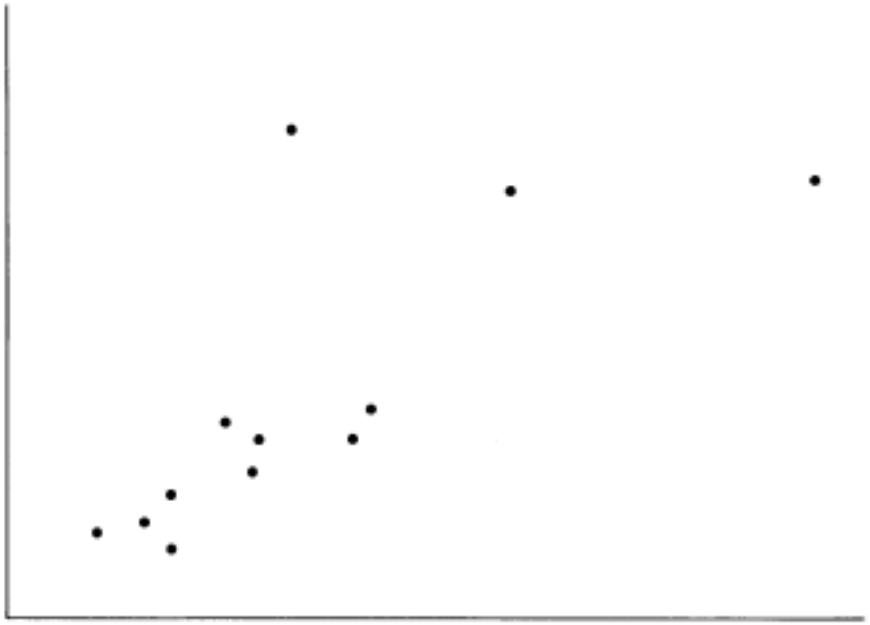
Redesign...

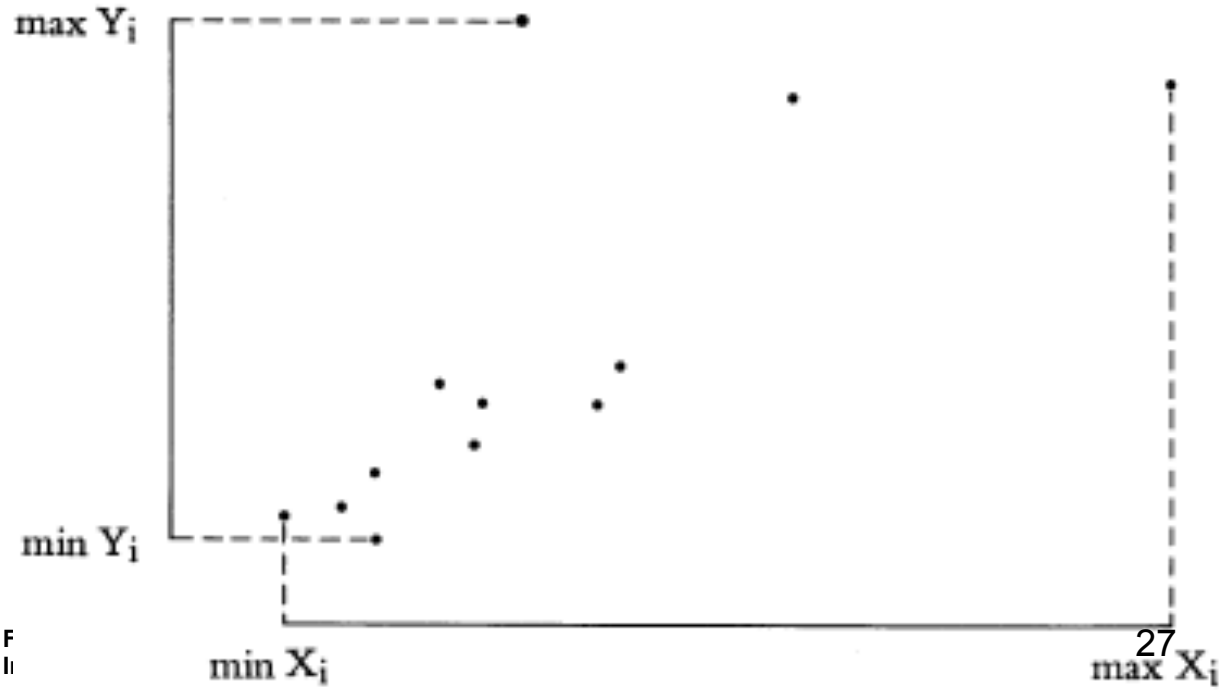


Redesign

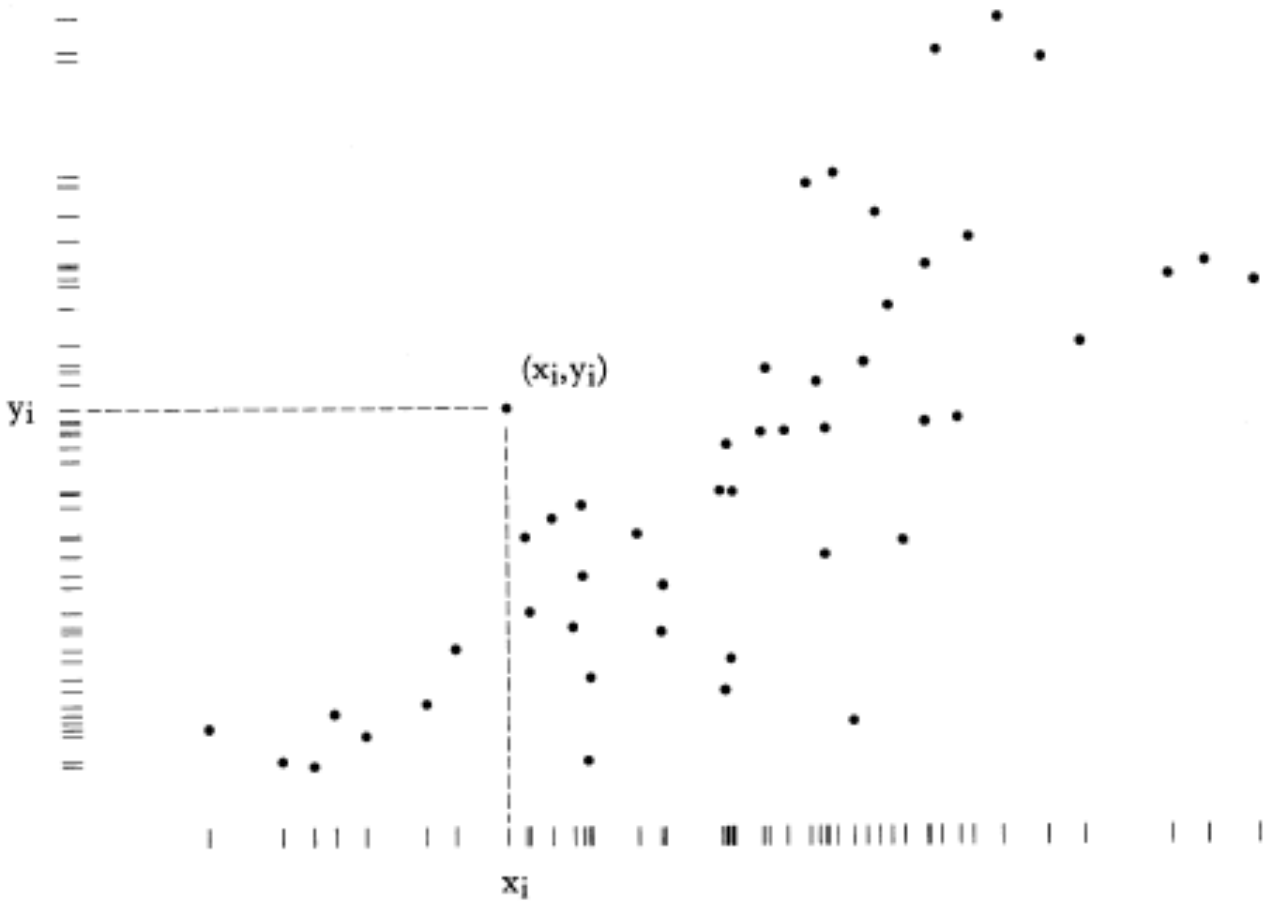


Redesign of Scatter Plot: Range Frame

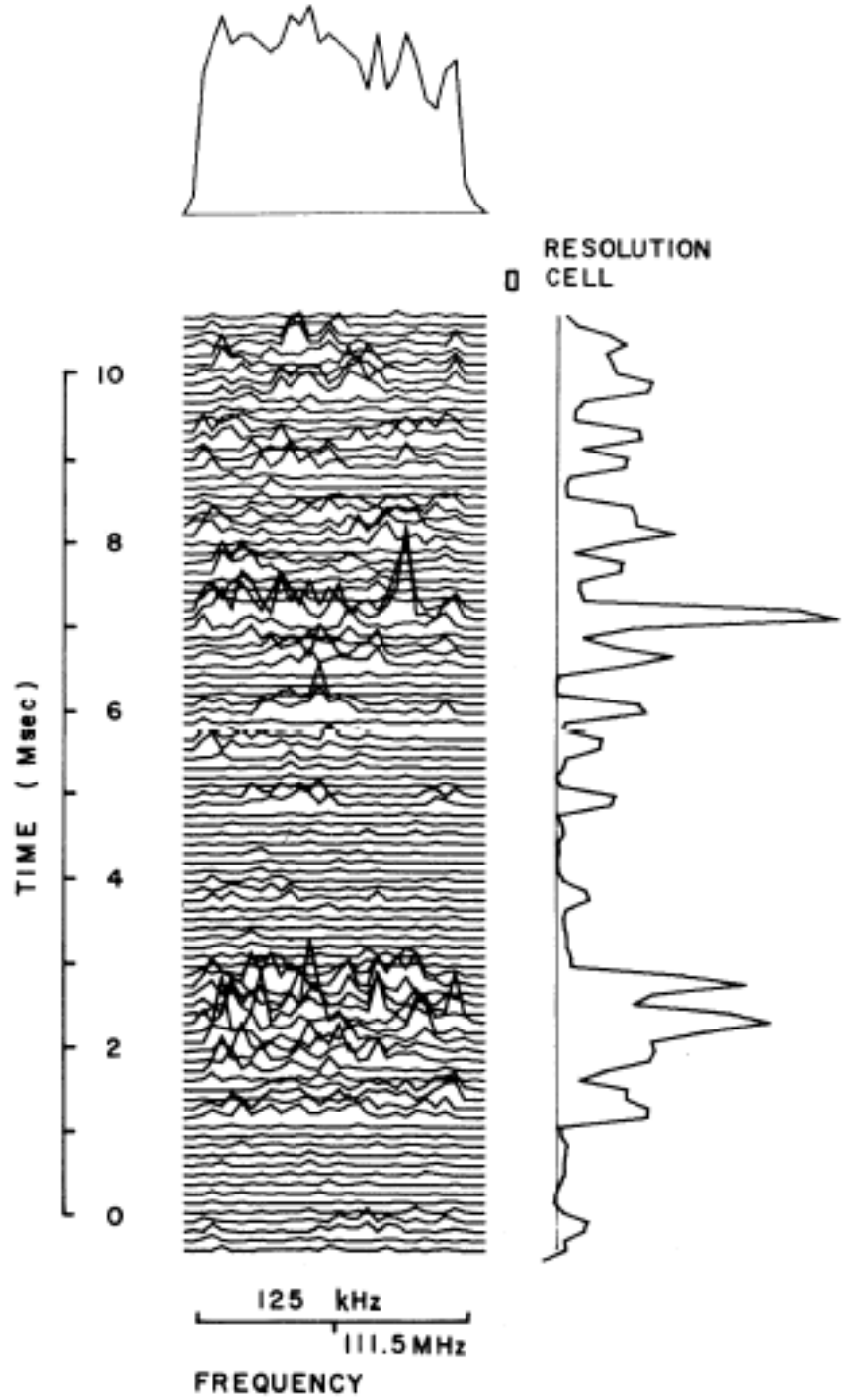




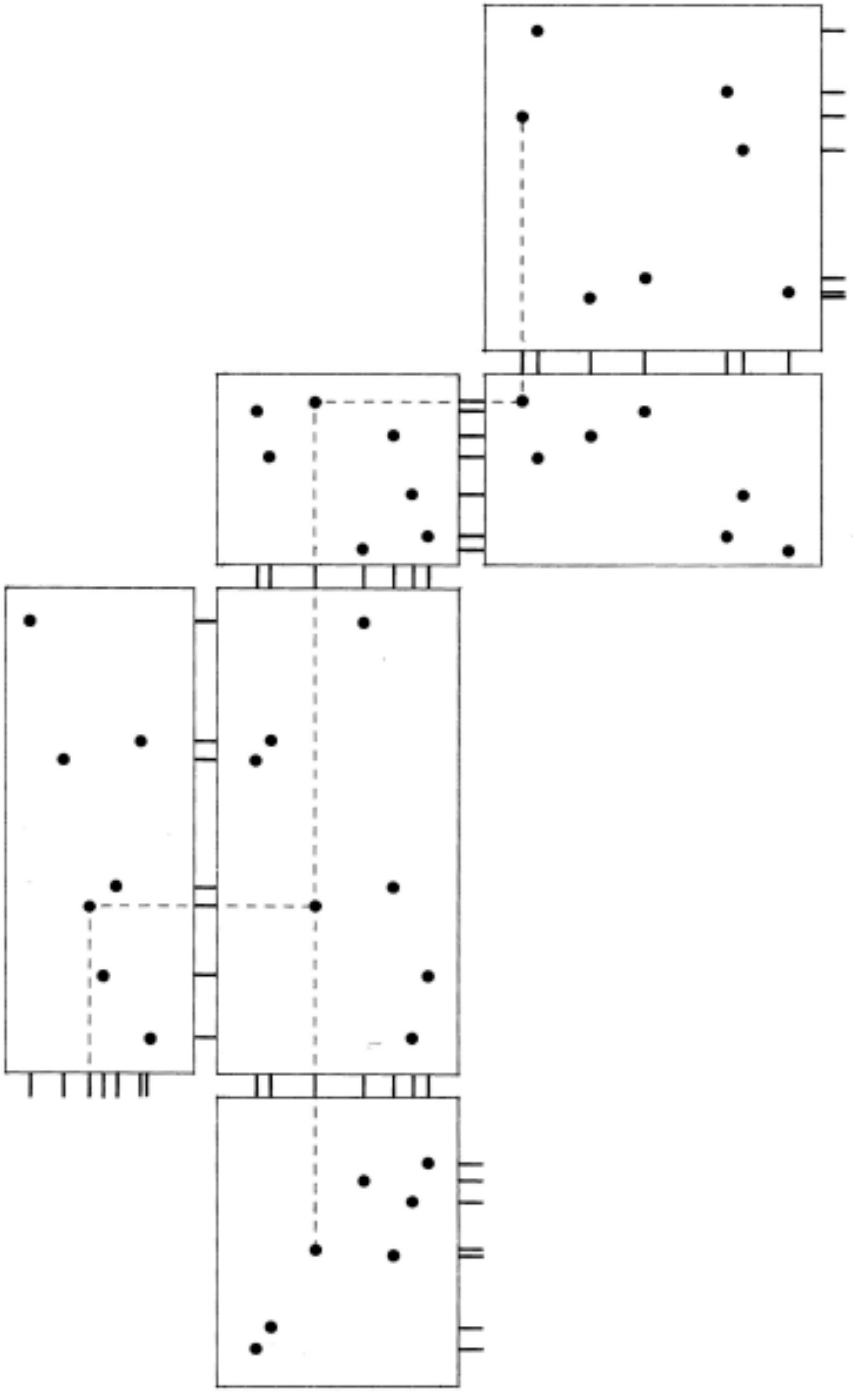
Dot Dash Plots



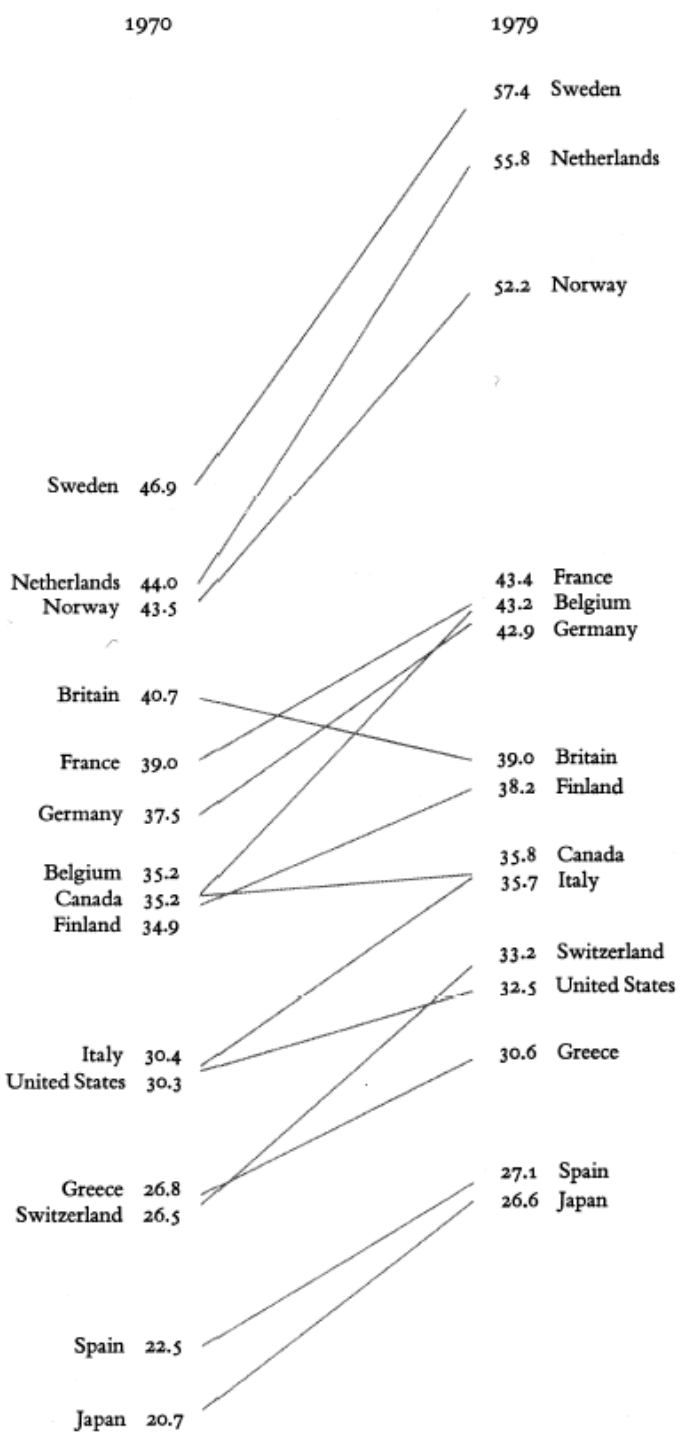
Marginal Distributions



Narrowband spectra of individual sub pulses. Each point of the intensity $I_q(t)$ plotted on the right is the sum of the distribution of intensities across the receiver bandwidth shown in the center. At the top is plotted the spectrum averaged over the pulse. In the limit of many thousands of pulses this would show the receiver bandpass shape.



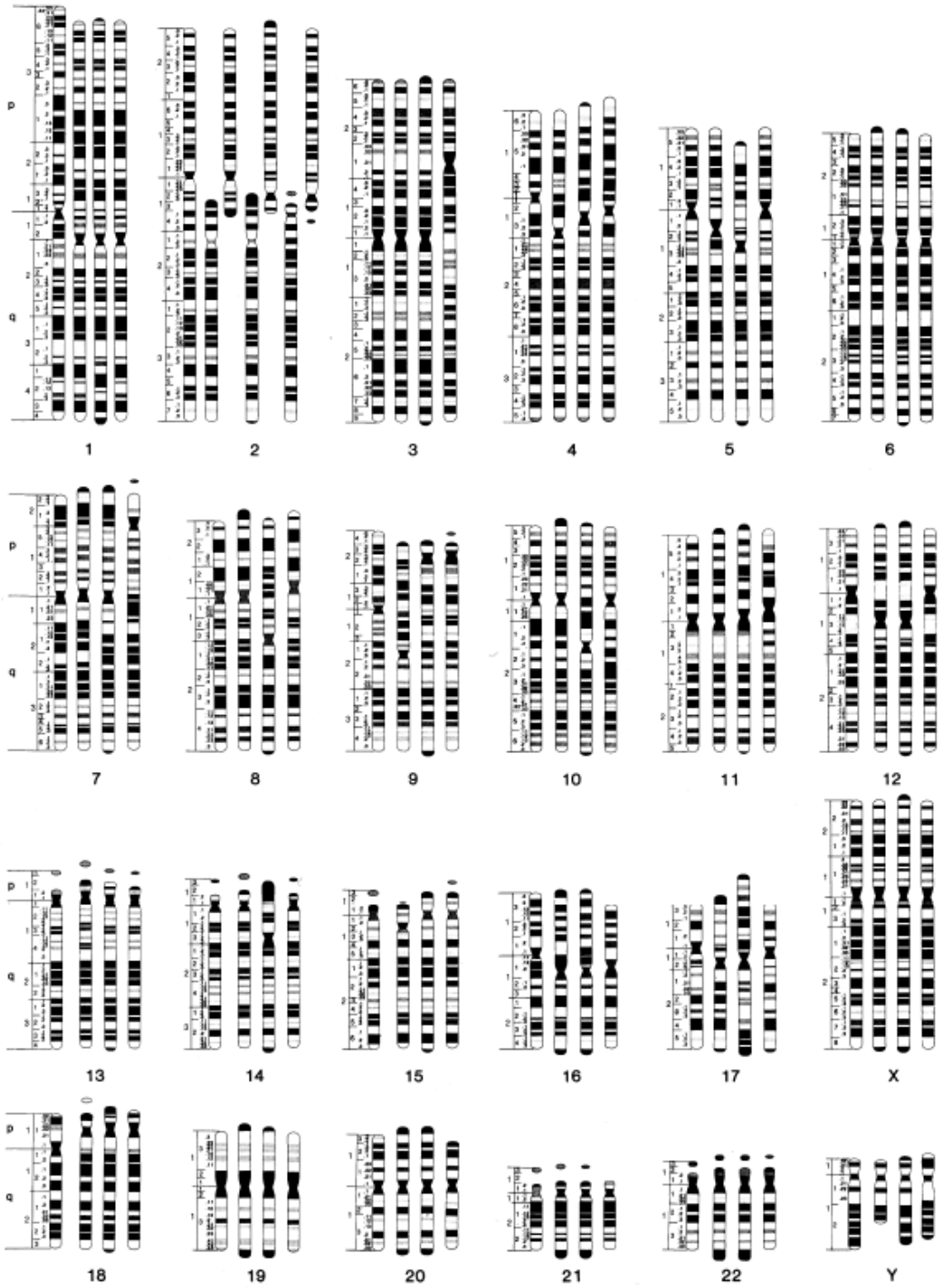
Seq. Bivariate scatters: rugplot



Multi-functional Graphical elements

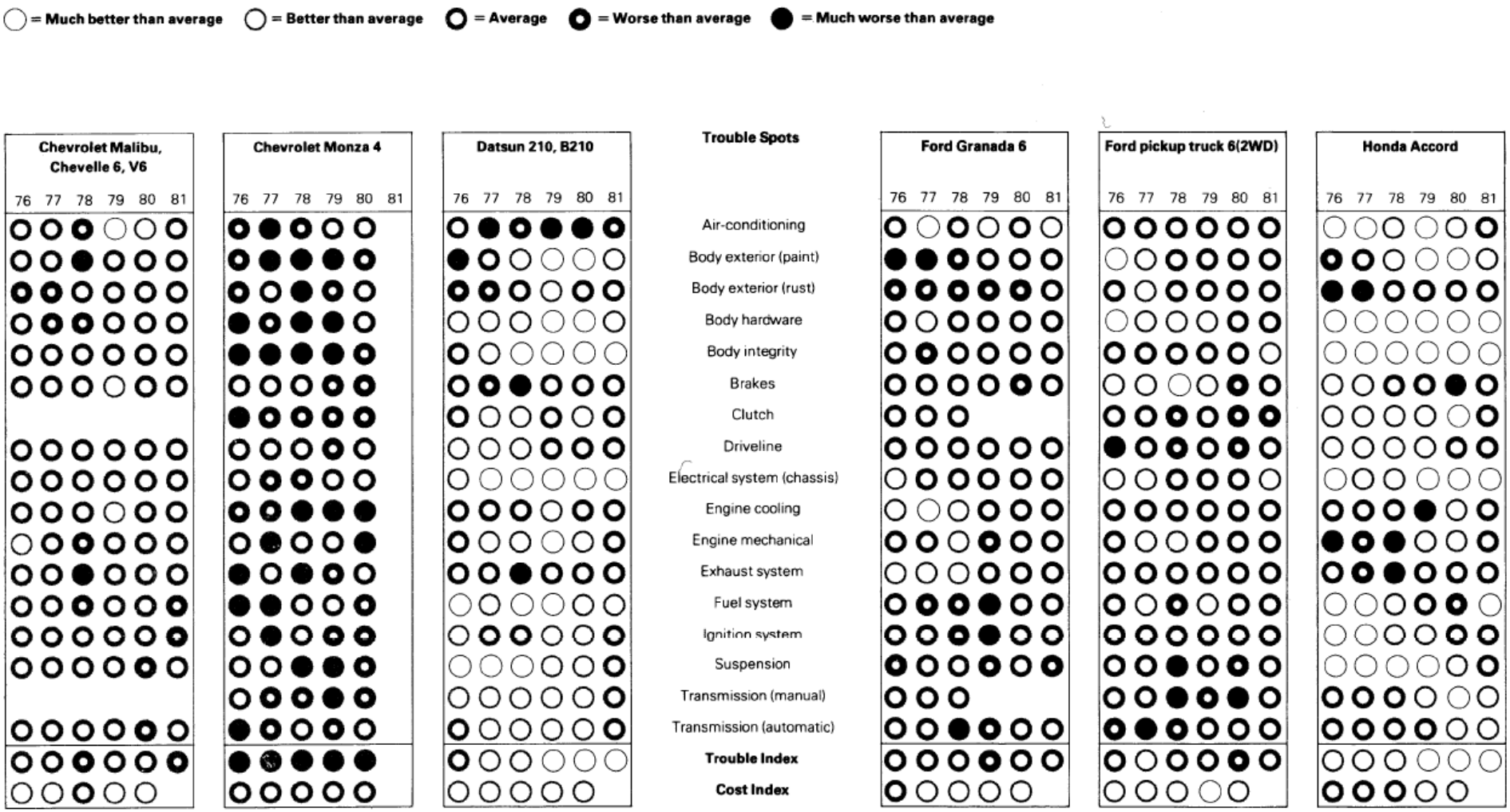
Many (high) dimensions

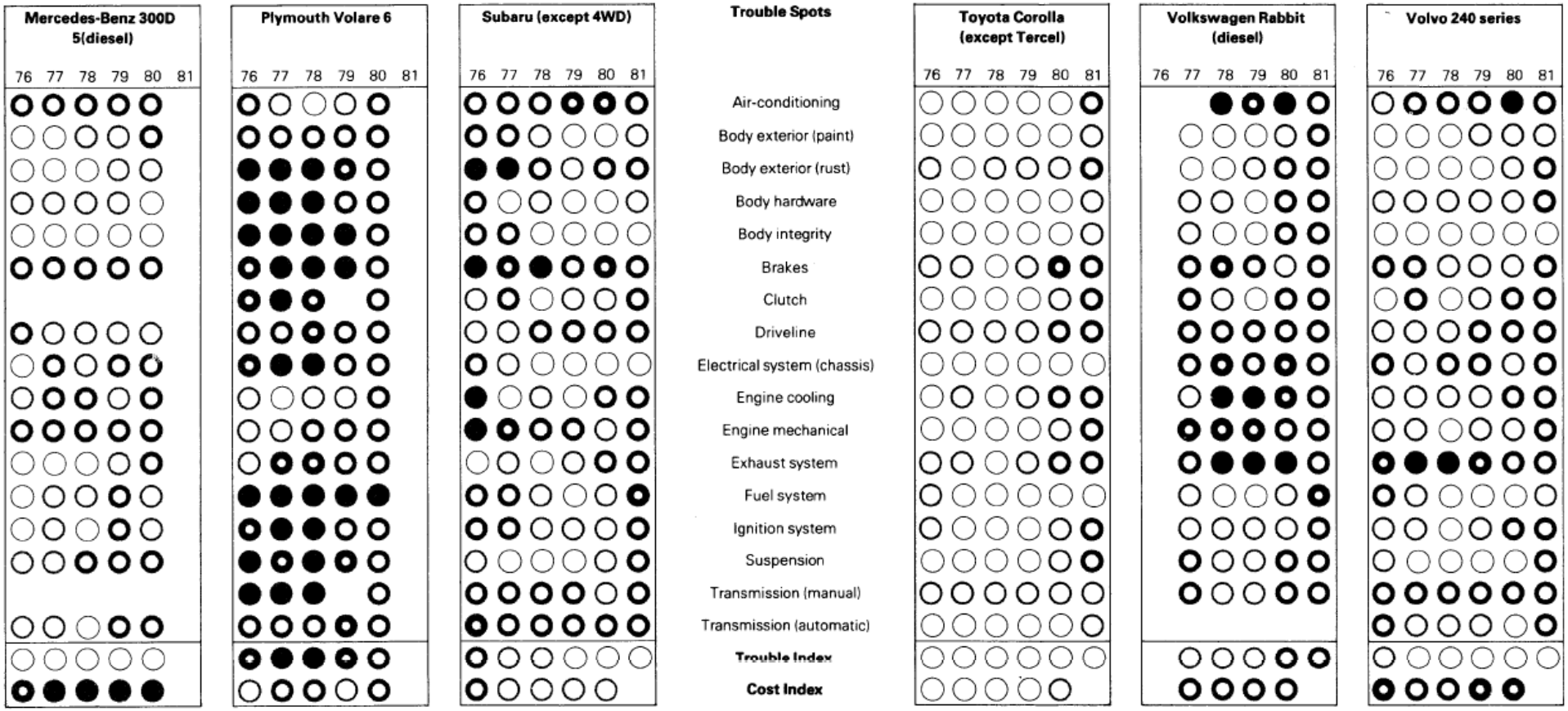
Use Multiples (like sequal of movies stills)



Chromo-somal Pictorial Legacy

Plöger


= Much better than average
= Better than average
= Average
= Worse than average
= Much worse than average
Chevrolet Malibu, Chevelle 6, V6
Chevrolet Monza 4
Datsun 210, B210
Trouble Spots
Ford Granada 6
Ford pickup truck 6(2WD)
Honda Accord
76 77 78 79 80 81
Air-conditioning
Body exterior (paint)
Body exterior (rust)
Body hardware
Body integrity
Brakes
Clutch
Driveline
Electrical system (chassis)
Engine cooling
Engine mechanical
Exhaust system
Fuel system
Ignition system
Suspension
Transmission (manual)
Transmission (automatic)
Trouble Index
Cost Index


Mercedes-Benz 300D 5(diesel)
Plymouth Volare 6
Subaru (except 4WD)
Trouble Spots
Toyota Corolla (except Tercel)
Volkswagen Rabbit (diesel)
Volvo 240 series
76 77 78 79 80 81
Air-conditioning
Body exterior (paint)
Body exterior (rust)
Body hardware
Body integrity
Brakes
Clutch
Driveline
Electrical system (chassis)
Engine cooling
Engine mechanical
Exhaust system
Fuel system
Ignition system
Suspension
Transmission (manual)
Transmission (automatic)
Trouble Index
Cost Index

Conclusion

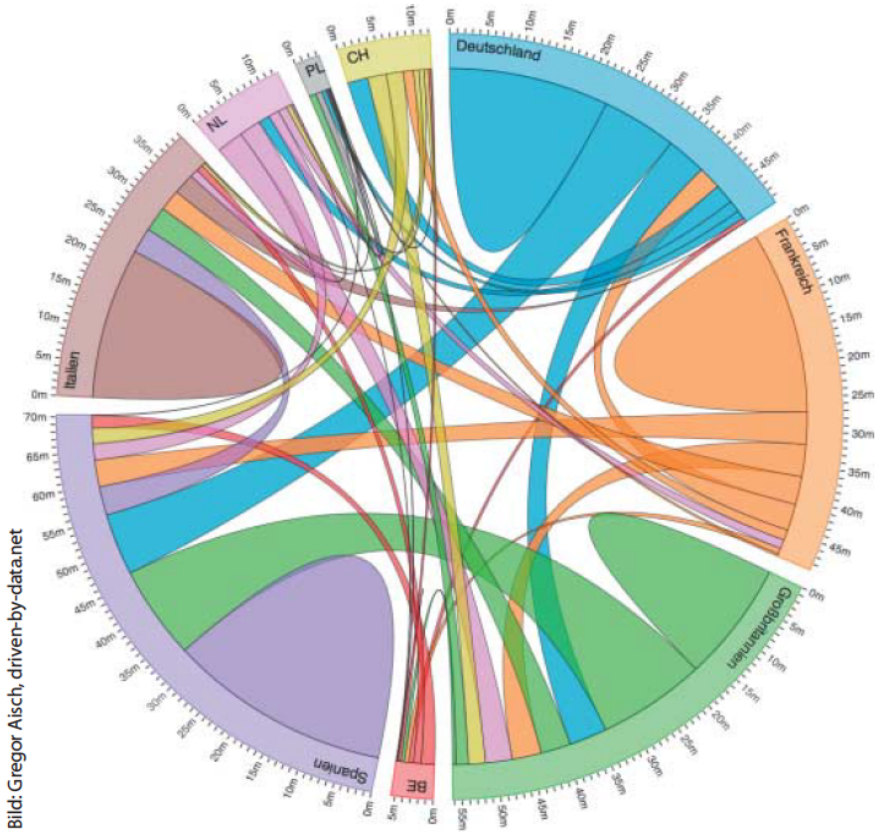
Well-designed small multiples are:

- 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.

Chord Diagram



Visualisiert man die Flugverbindungen in Europa als sogenanntes Chord-Diagramm, springen die wichtigen Verbindungen direkt ins Auge, ohne dass weniger genutzte Routen unter den Tisch fallen.

Parteispenden über 50.000 €

Zeitraum: Juli 2002 bis Januar 2011

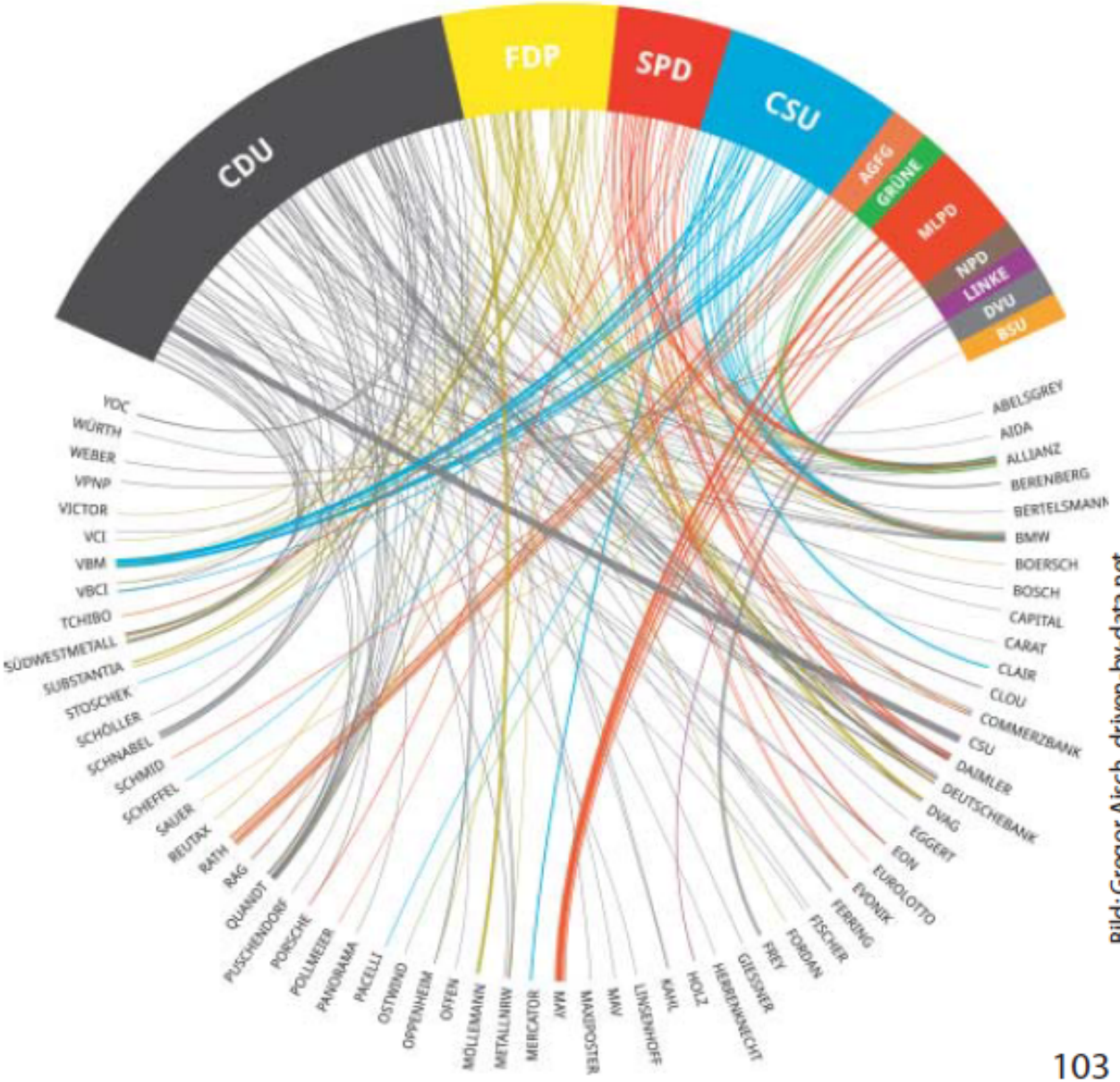


Bild: Gregor Aisch, driven-by-data.net

in:
RoboCup 2011: Robot Soccer World Cup XV
Lecture Notes in Computer Science
Volume 7416, 2012, pp 137-148
Towards Robust Object Categorization for Mobile Robots with Combination of Classifiers
Christian A. Mueller, Nico Hochgeschwender, Paul G. Ploeger

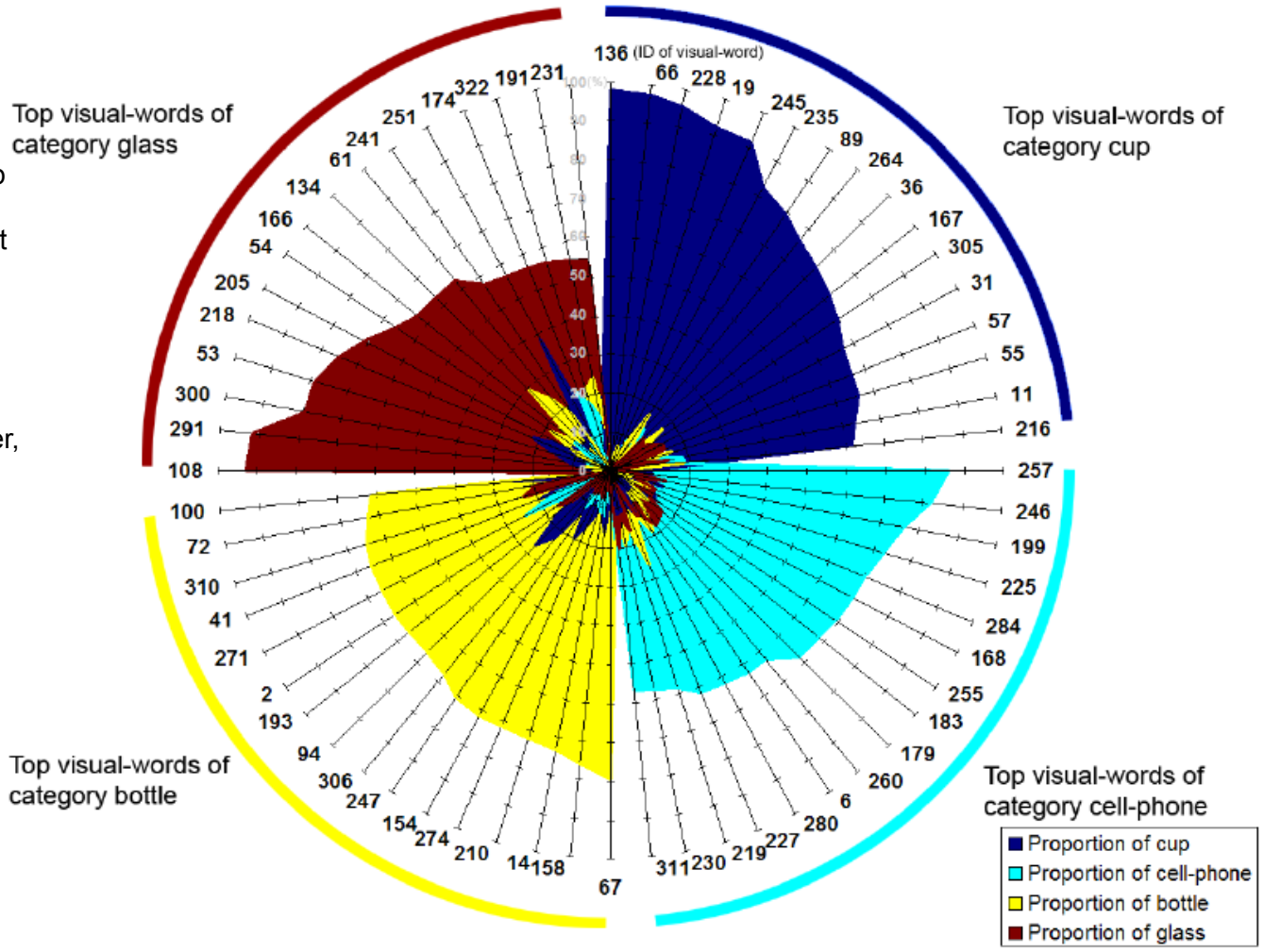
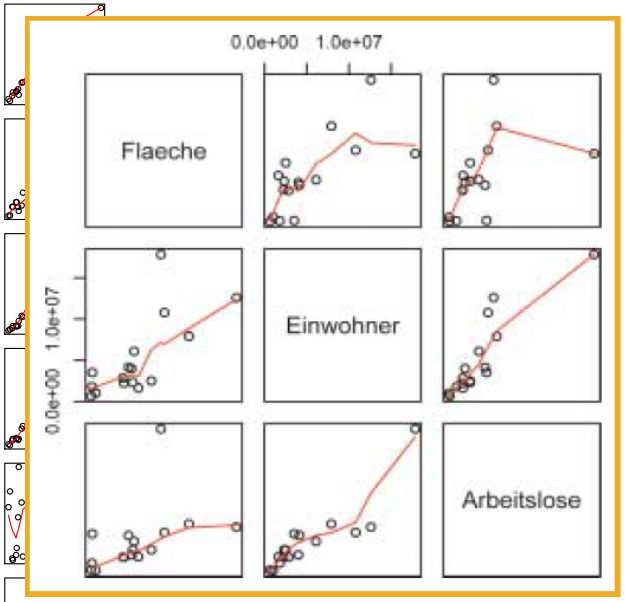


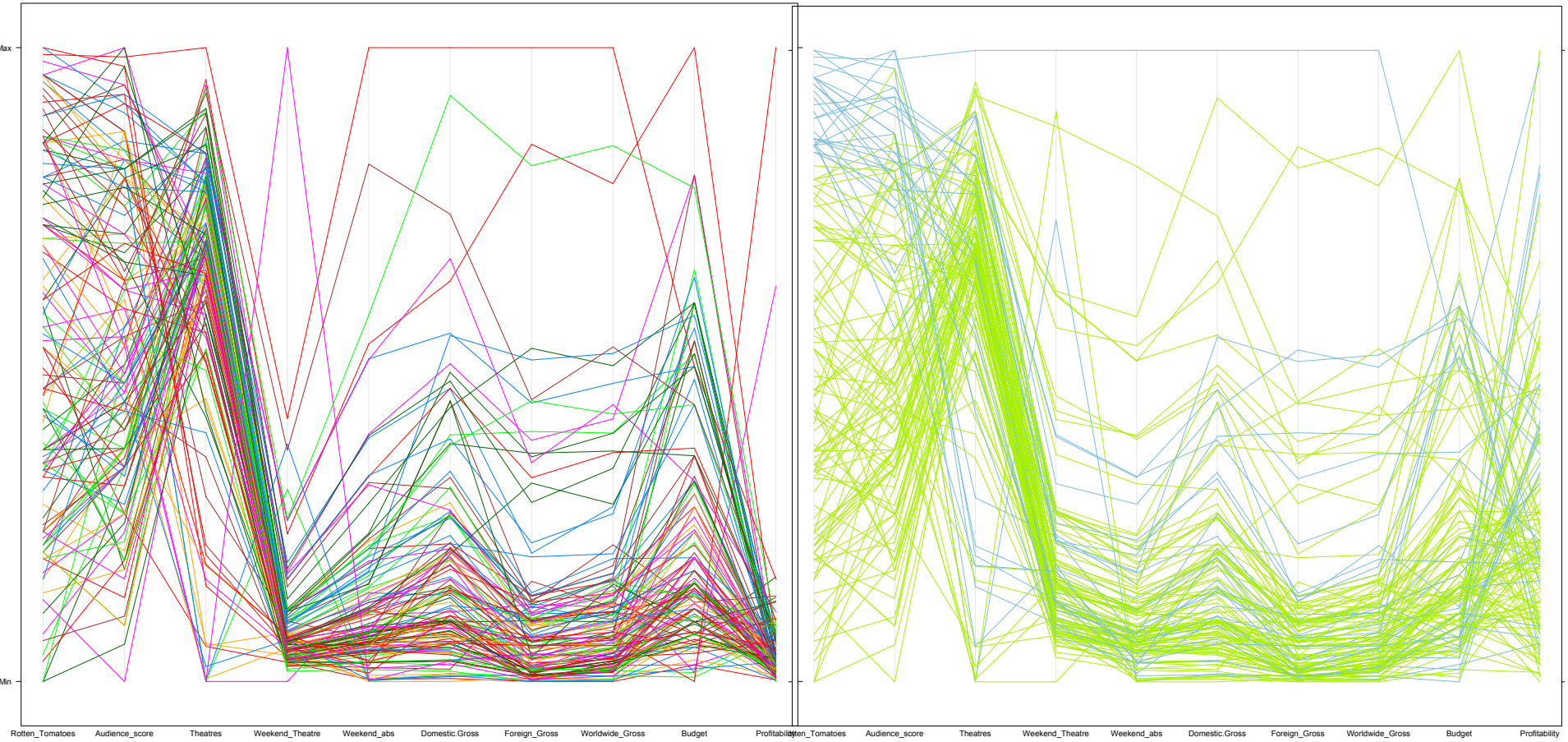
Fig. 3. The top discriminative visual words with their proportions of frequency regarding each category (*16* selected visual words of each category). The visual words are sorted in ascending order of their proportion in the respective dedicated category. E.g. in case of the cup-category: the most discriminative visual word is the one with *ID-136* which has a proportion of *97.7%* in the cup cateogory, *0.8%* in the cell-phone category and *1.4%* in the bottle category.

Scatter Plot

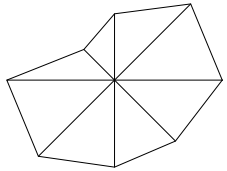


The upper left corner of the scatter plot in detail: In the first row we see that the area of a country is not necessarily correlated with the population and unemployment. The latter two values, however,are closely related to each other, as you can see in bottom row far right or in the middle square in the second row.

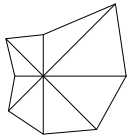
Parallel axises



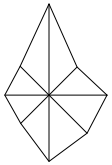
Spider chart



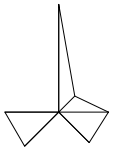
Kießling



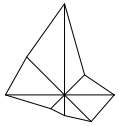
Lewandowski



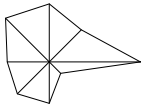
Müller



Ribery

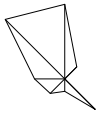


Blaszczykowski

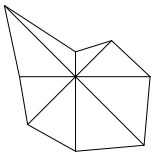


Reus

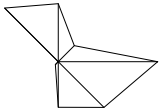
Götze



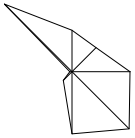
Huszti



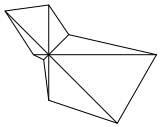
Ibisevic



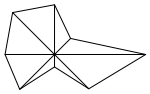
de Bruyne



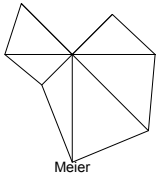
Diouf



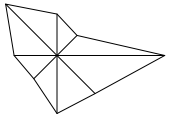
Kruse



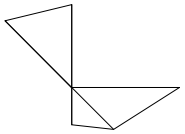
Diego



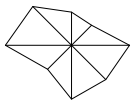
Meier



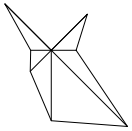
Schürrle



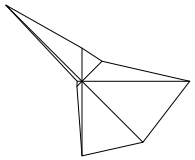
Volland



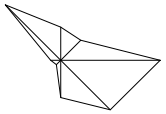
Hunt



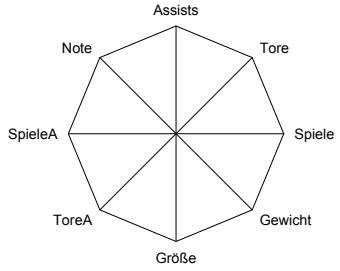
Mandzukic



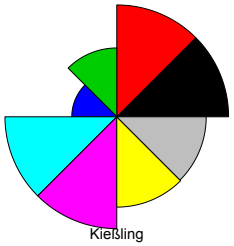
Petersen

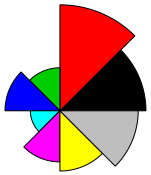


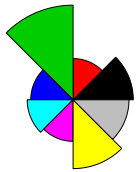
Schmid

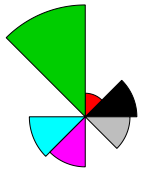


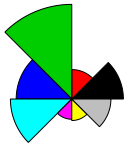
Nightingale -Charts

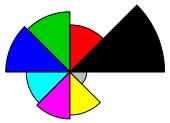

Kießling

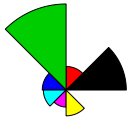

Lewandowski

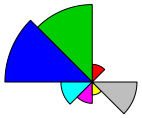

Müller

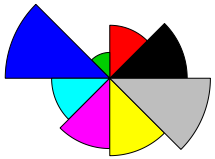

Ribery

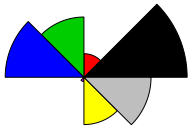

Blaszczykowski

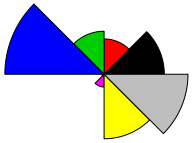

Reus

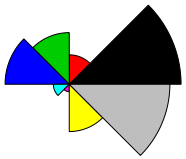

Götze

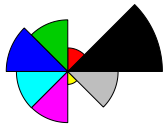

Huszti

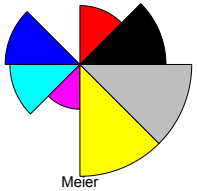

Ibisevic

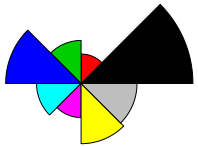

de Bruyne

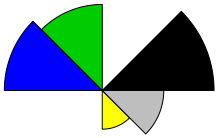

Diouf

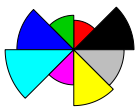

Kruse

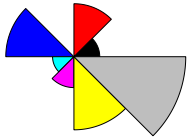

Diego

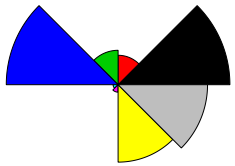

Meier

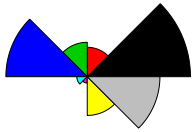

Schürrle

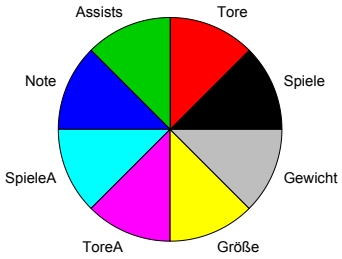

Volland


Hunt

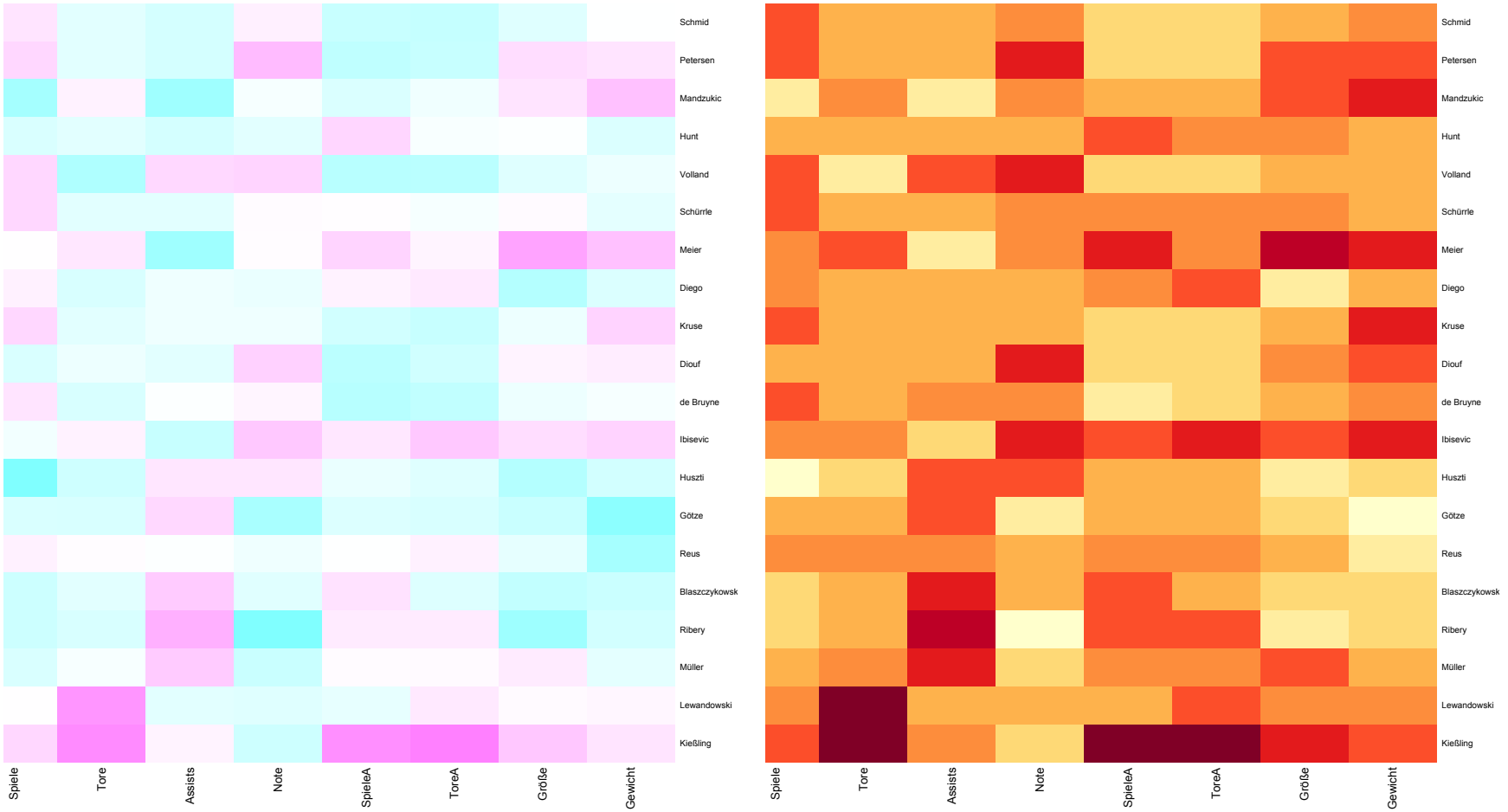

Mandzukic


Petersen


Schmidt



Heatmaps


Schmid
Petersen
Mandzukic
Hunt
Volland
Schürrle
Meier
Diego
Kruse
Diouf
de Bruyne
Ibisevic
Huszti
Götze
Reus
Blaszczykowsk
Ribery
Müller
Lewandowski
Kießling
Spiele
Tore
Assists
Note
SpieleA
ToreA
Größe
Gewicht