Erasmus+

e-Lab Prof Eng

Innovative Lifelong e-Learning for Professional Engineers

(e-ProfEng)

586391-EPP-1-2017-1-SE-EPPKA2-CBHE-JP

Training in Electrical Engineering Discipline

Modelling and Simulation in Electrical Engineering

Data visualization in data analysis

Data visualization

Josip Job

FERIT Osijek 2.4.2019.

Data visualization

Data visualization in data analysis

Importance of Data visualization
The four sets (Anscombe's quartet)
Why Data visualization
History of Data visualization
Data Wrangling
Mapping Data to Visual Variables
Multidimensional Data
FERIT Osijek experience in Data visualization
Dimensionality Reduction
Visual Encoding Design
Data visualization
Perception
Data visualization tools

Importance of Data visualization

The four sets (Anscombe's quartet)

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

$N = 11$
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 $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$

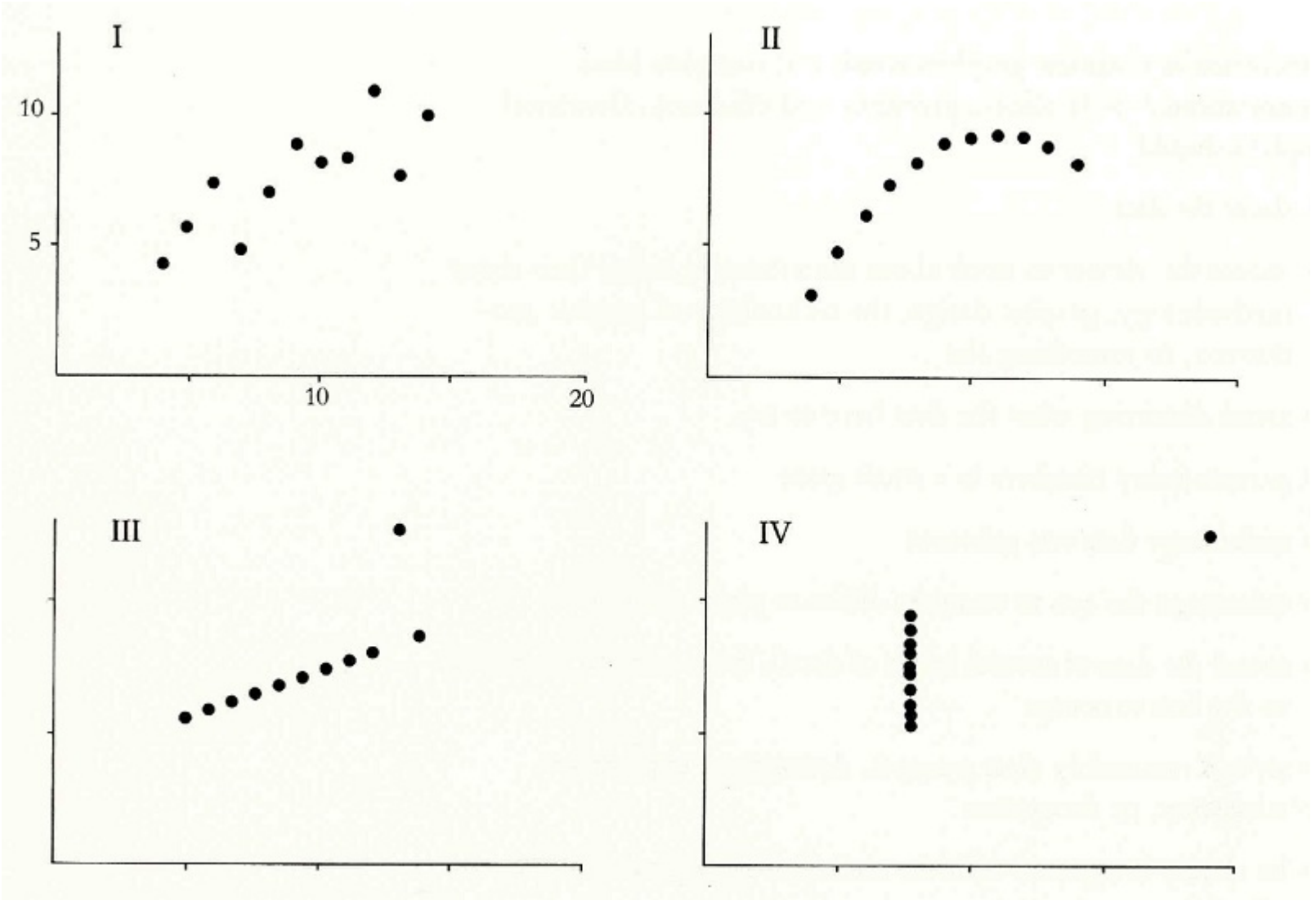
For all four datasets:

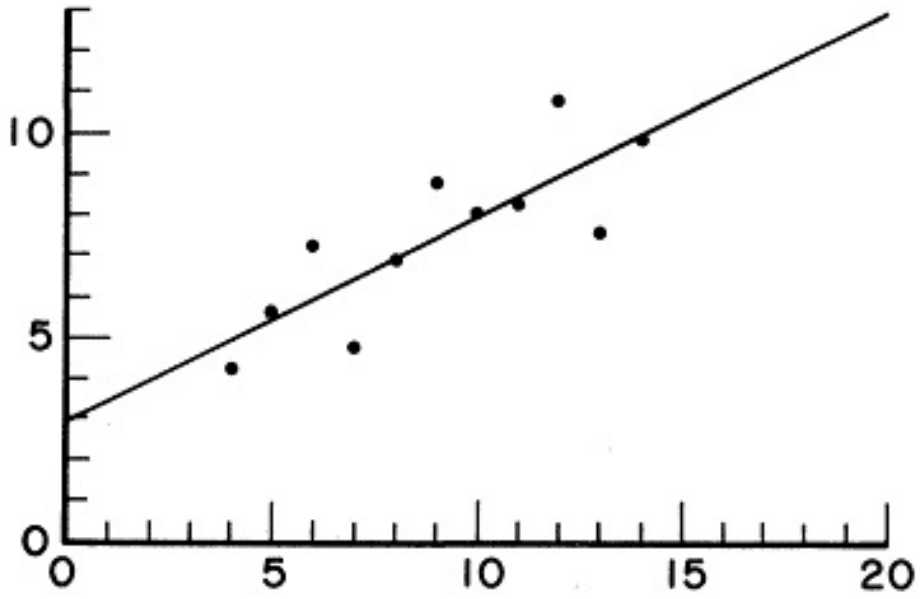
Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0 .500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

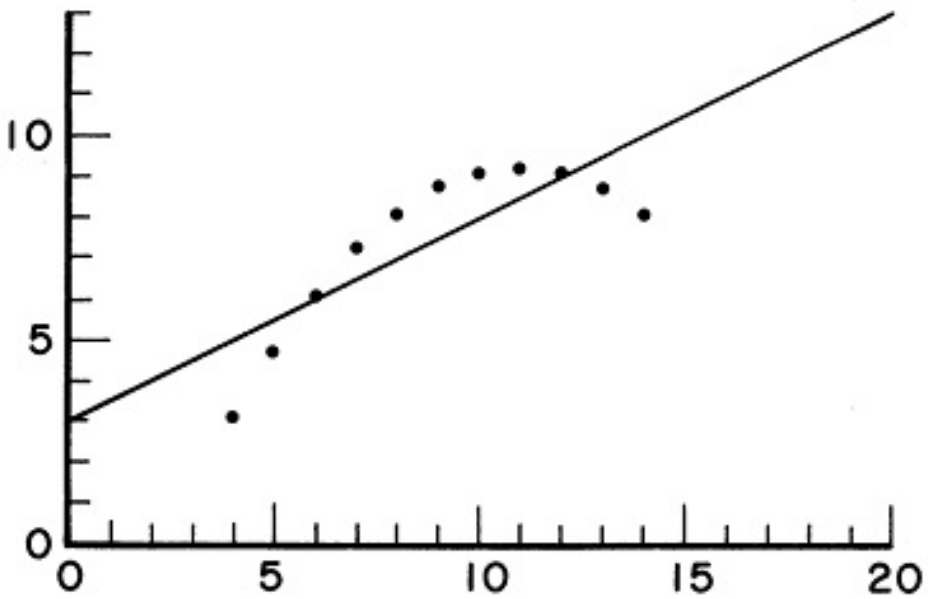
The four sets (Anscombe's quartet)

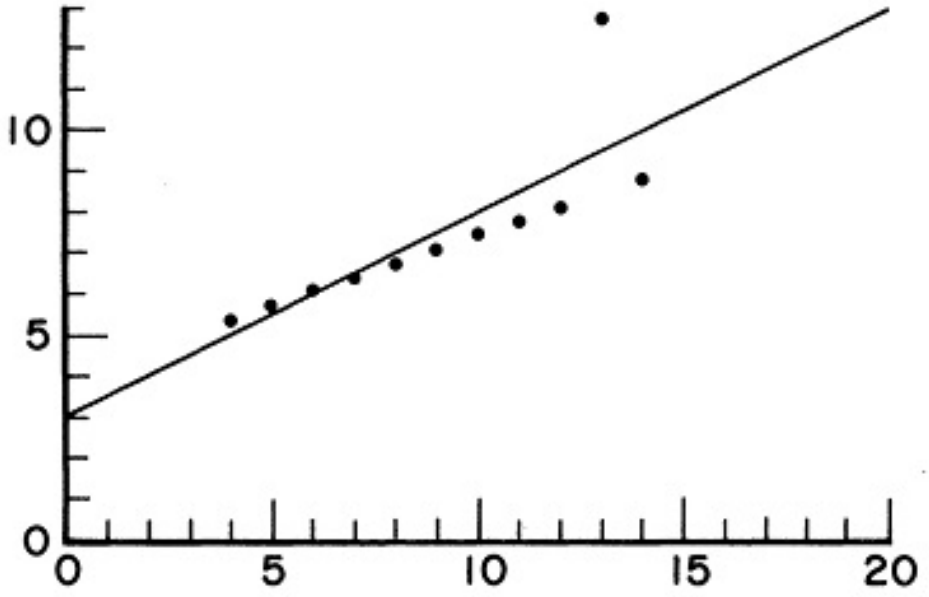
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

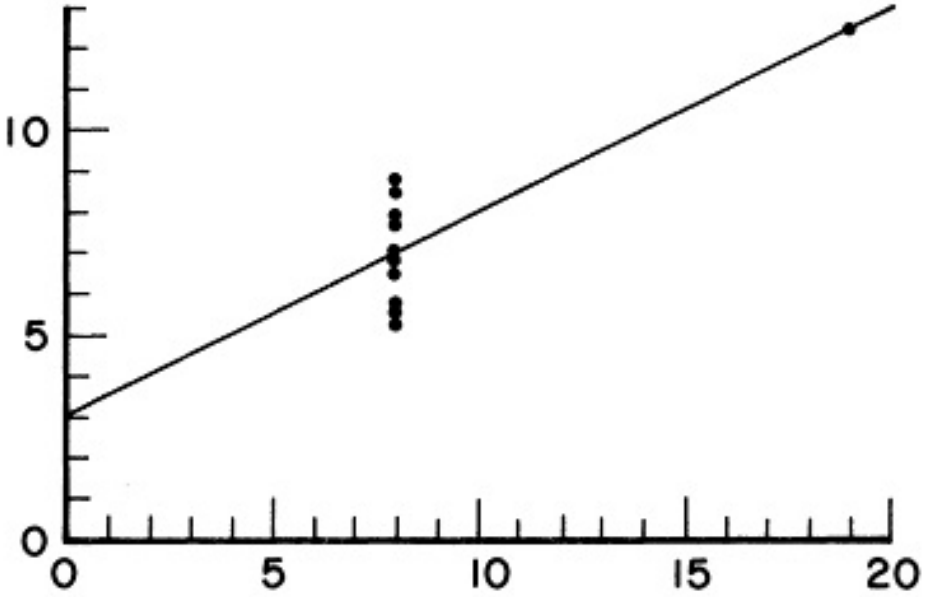
$N = 11$
mean of X's $= 9.0$
mean of Y's $= 7.5$
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope $= 0.118$
$t = 4.24$
sum of squares $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$


I
10
5
10
20
II
III
IV

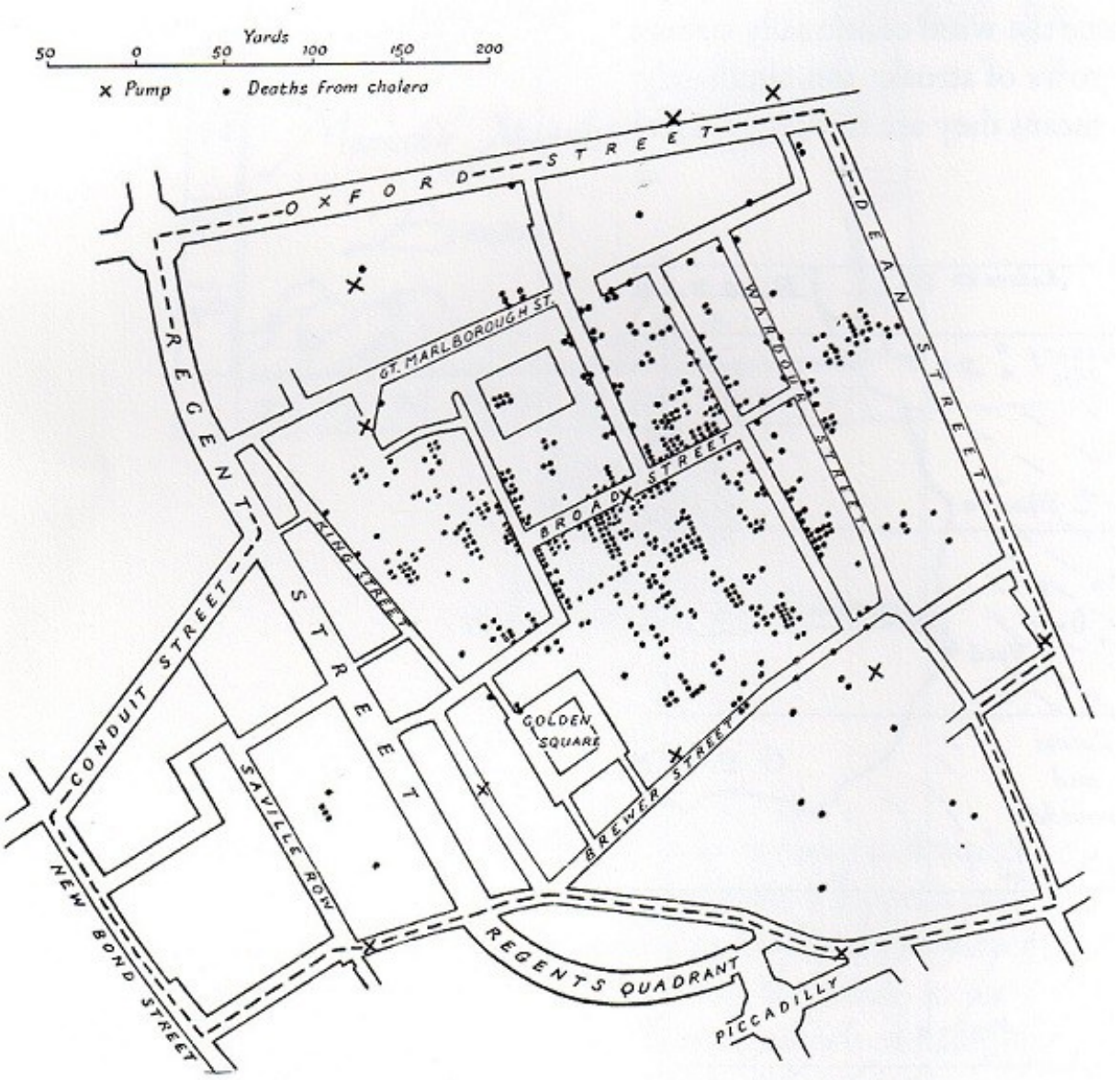

10
5
0
0
5
10
15
20


10
5
0
0
5
10
15
20

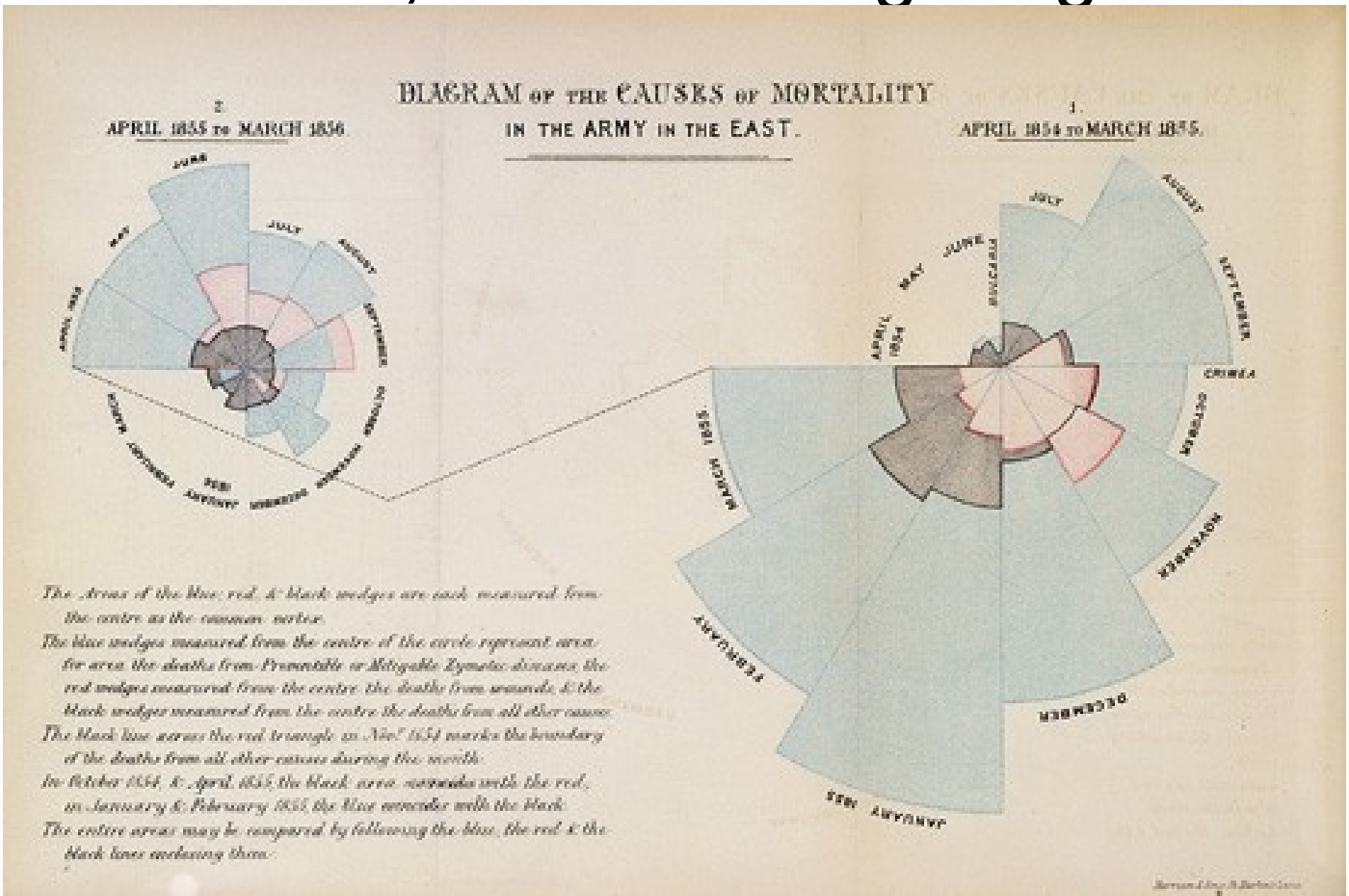

10
5
0
0
5
10
15
20


10
5
0
0
5
10
15
20

Data in Context: Cholera Outbreak



1856 “Coxcomb” of Crimean War Deaths, Florence Nightingale



Global Information Storage Capacity

in optimally compressed bytes

1986
ANALOG
2.6 exabytes

DIGITAL
0.02 exabytes

1993

2000

ANALOG STORAGE

DIGITAL STORAGE

2002:
"beginning of the digital age"
50%

% digital:
1 %
3 %
25 %
94 %

2007 ANALOG
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 %

ANALOG

DIGITAL

- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %

- PC hard disks: 44.5 %
  123 billion gigabytes

- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

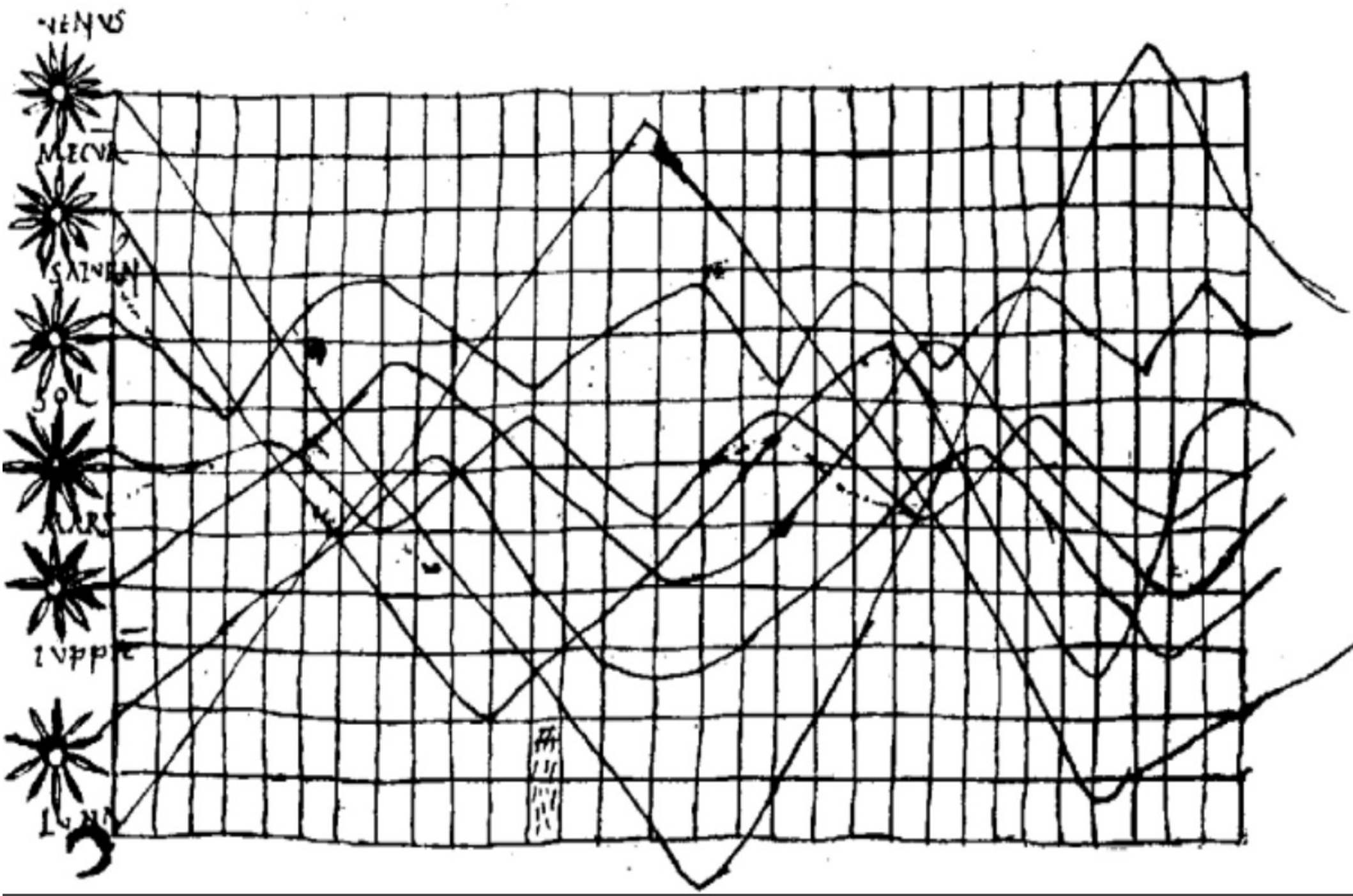
Why Data visualization

Answer questions (or discover them)
Make decisions
See data in context
Expand memory
Support graphical calculation
Find patterns
Present argument or tell a story
Inspire

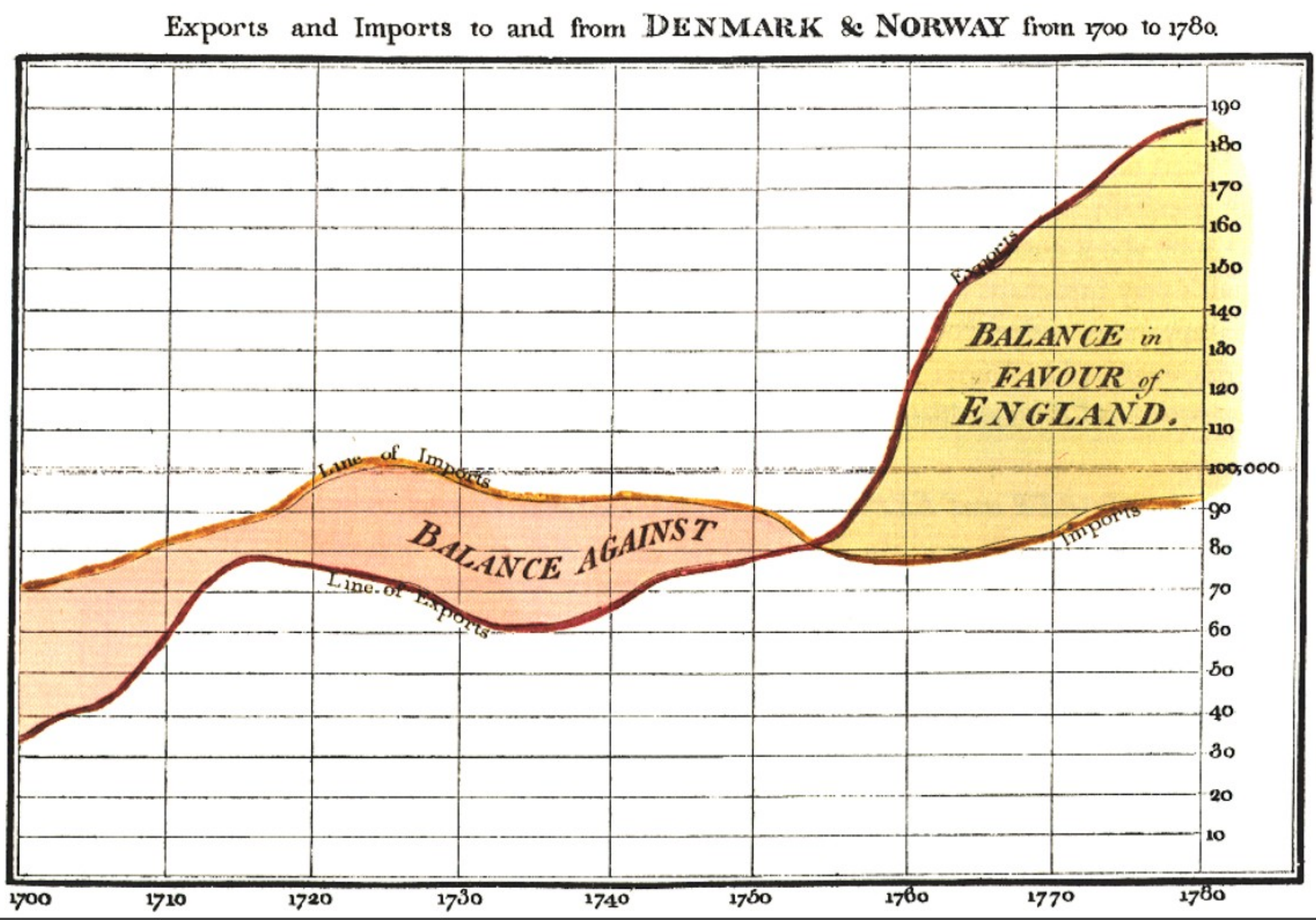
History of Data visualization

~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey

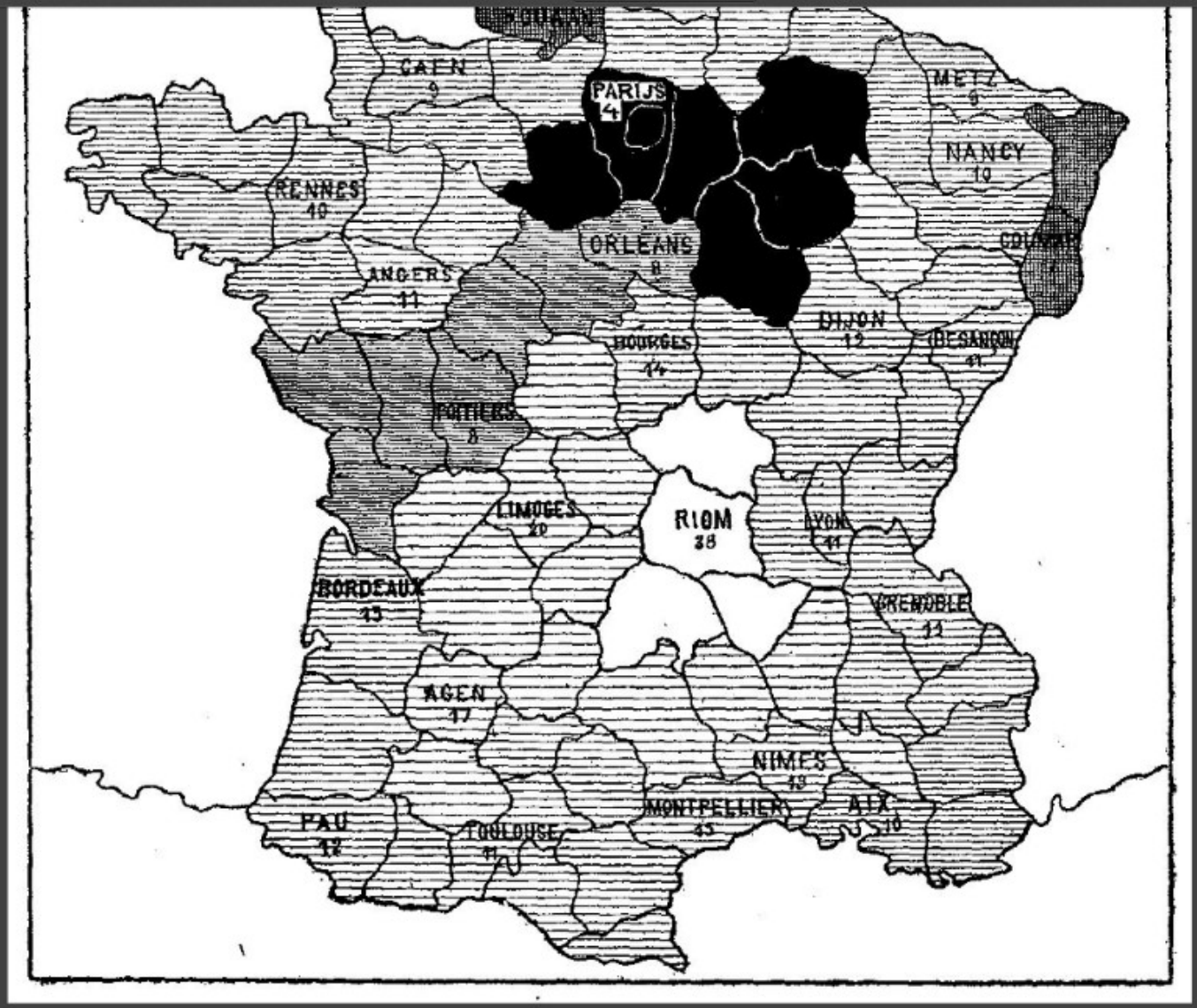
0 BC



~950 AD Position of Sun, Moon and Planets



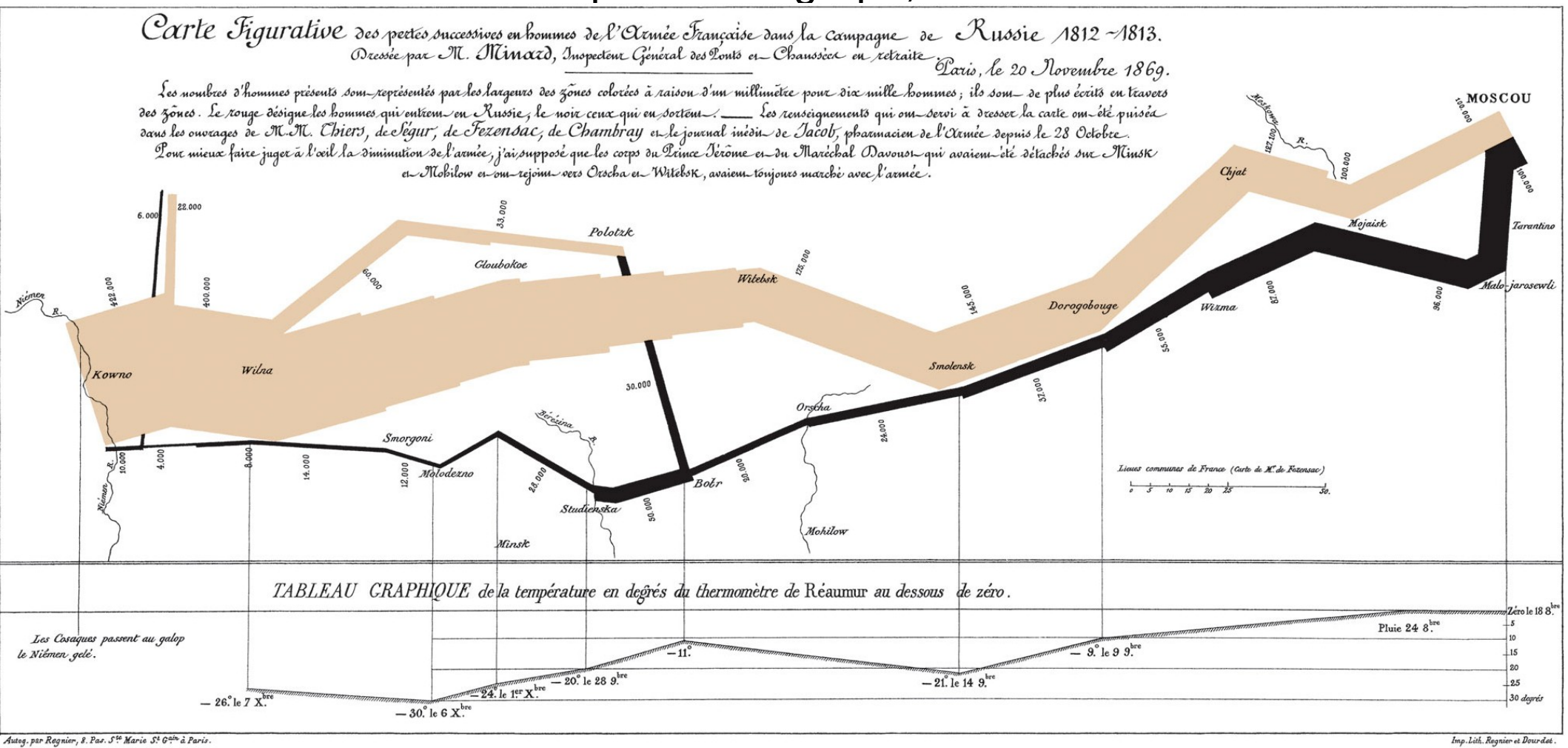
The Commercial and Political Atlas, William Playfair 1786



1786 1826(?) Illiteracy in France, Pierre Charles Dupin

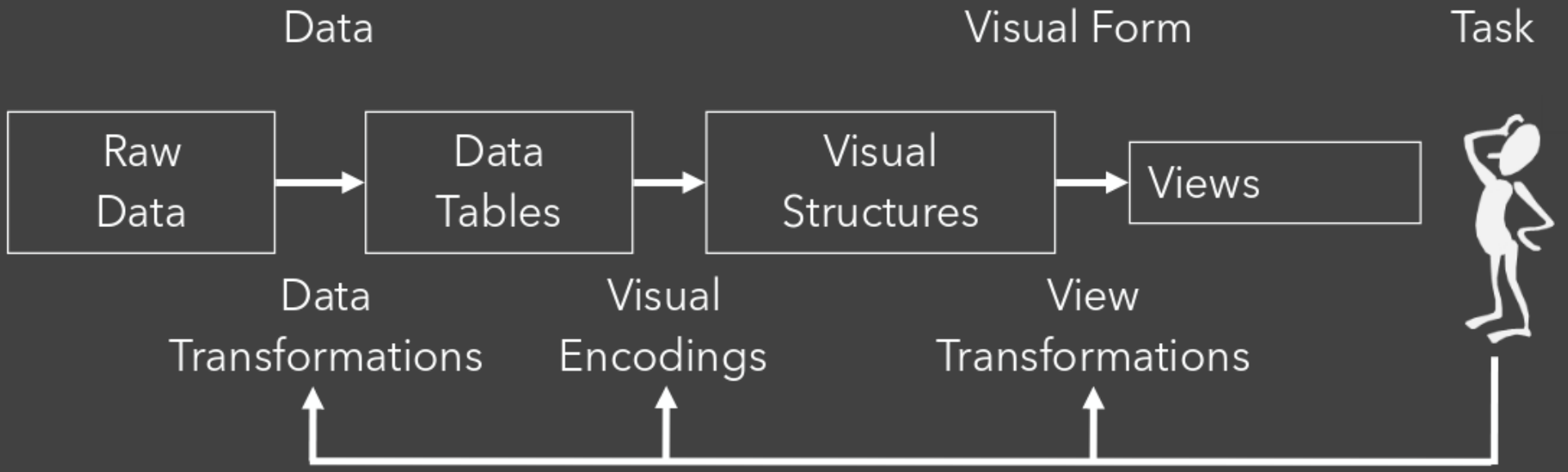
Charles Minard's 1869 – Napoleon’s March

The chart is showing the number of men in Napoleon’s 1812 Russian campaign army, their movements, as well as the temperature they encountered on the return path. Lithograph, 62 × 30 cm.



Data visualization

Data visualization model



Mapping Data to Visual Variables

Mapping Data to Visual Variables

Assign data fields (e.g., with N, O, Q types) to visual channels (x, y, color, shape, size, ...) for a chosen graphical mark type (point, bar, line, ...).

Additional concerns include choosing appropriate encoding parameters (log scale, sorting, ...) and data transformations (bin, group, aggregate, ...).

These options define a large combinatorial space, containing both useful and questionable charts!

Mapping Data to Visual Variables

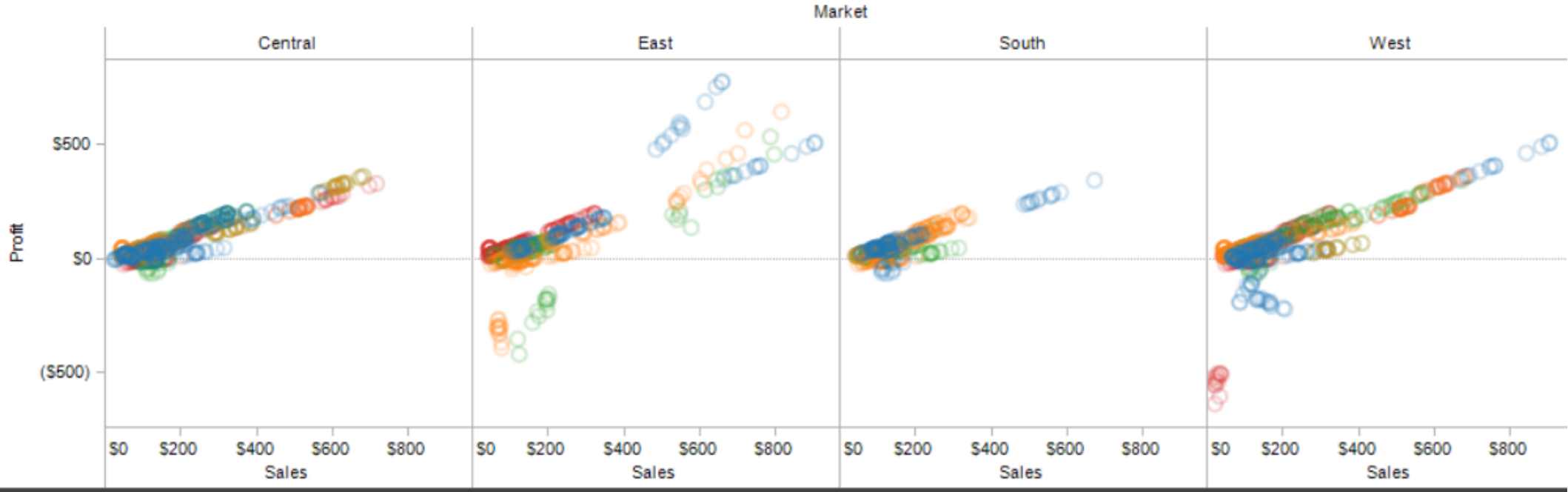
Visual Encoding Variables

1. Position (X)
Position (Y)
2. Size
3. Value
4. Texture
5. Color
6. Orientation
7. Shape

~8 dimensions?

Multidimensional Data

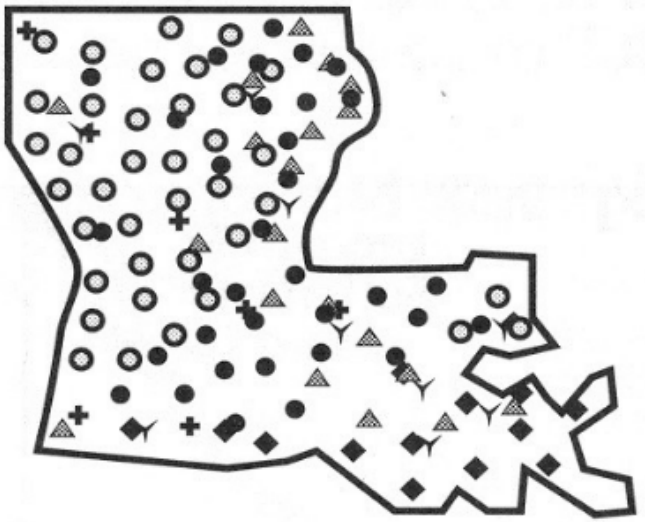
Trellis Plots



A *trellis plot* subdivides space to enable comparison across multiple plots.

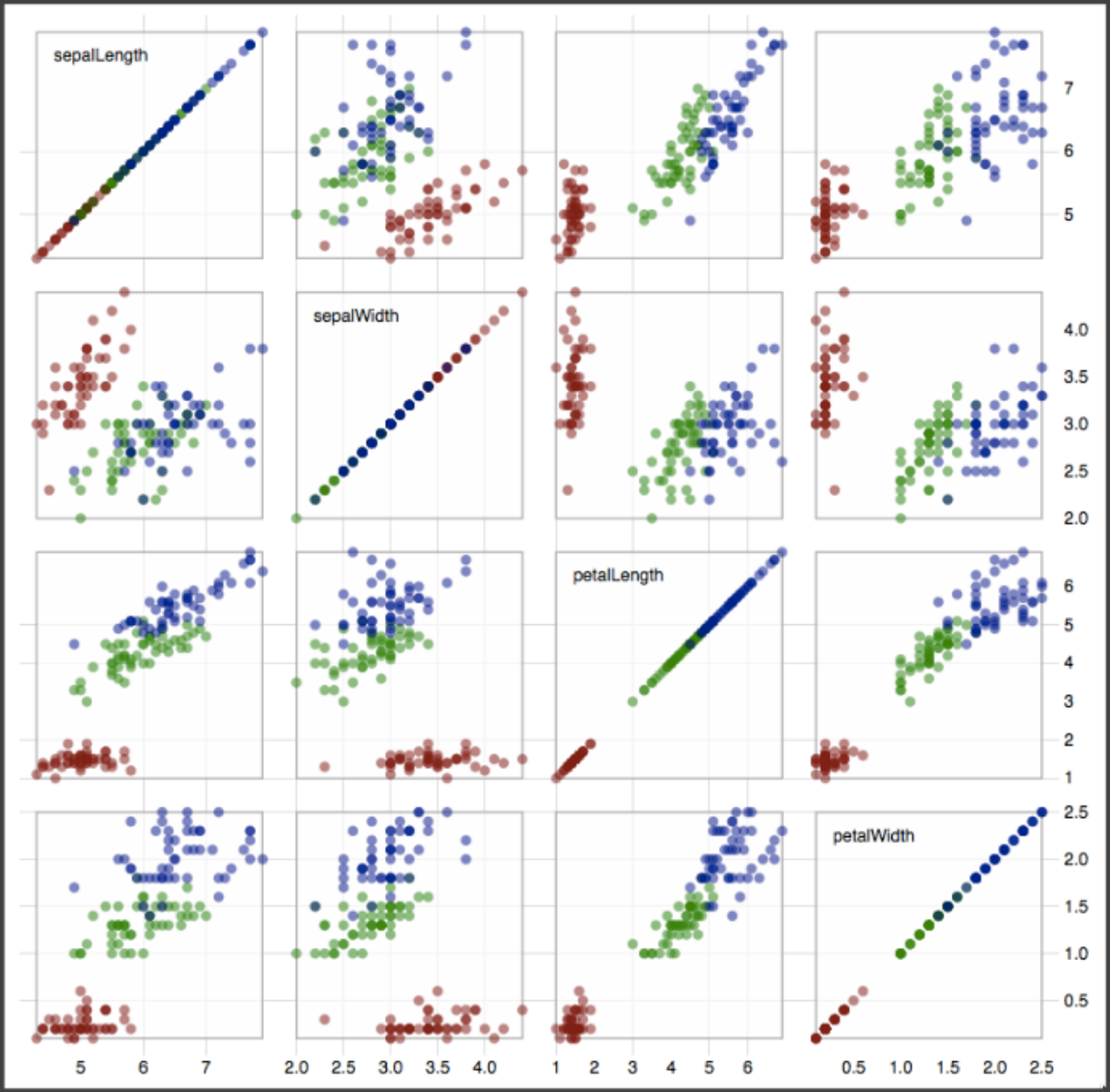
Typically nominal or ordinal variables are used as dimensions for subdivision.

Small Multiples



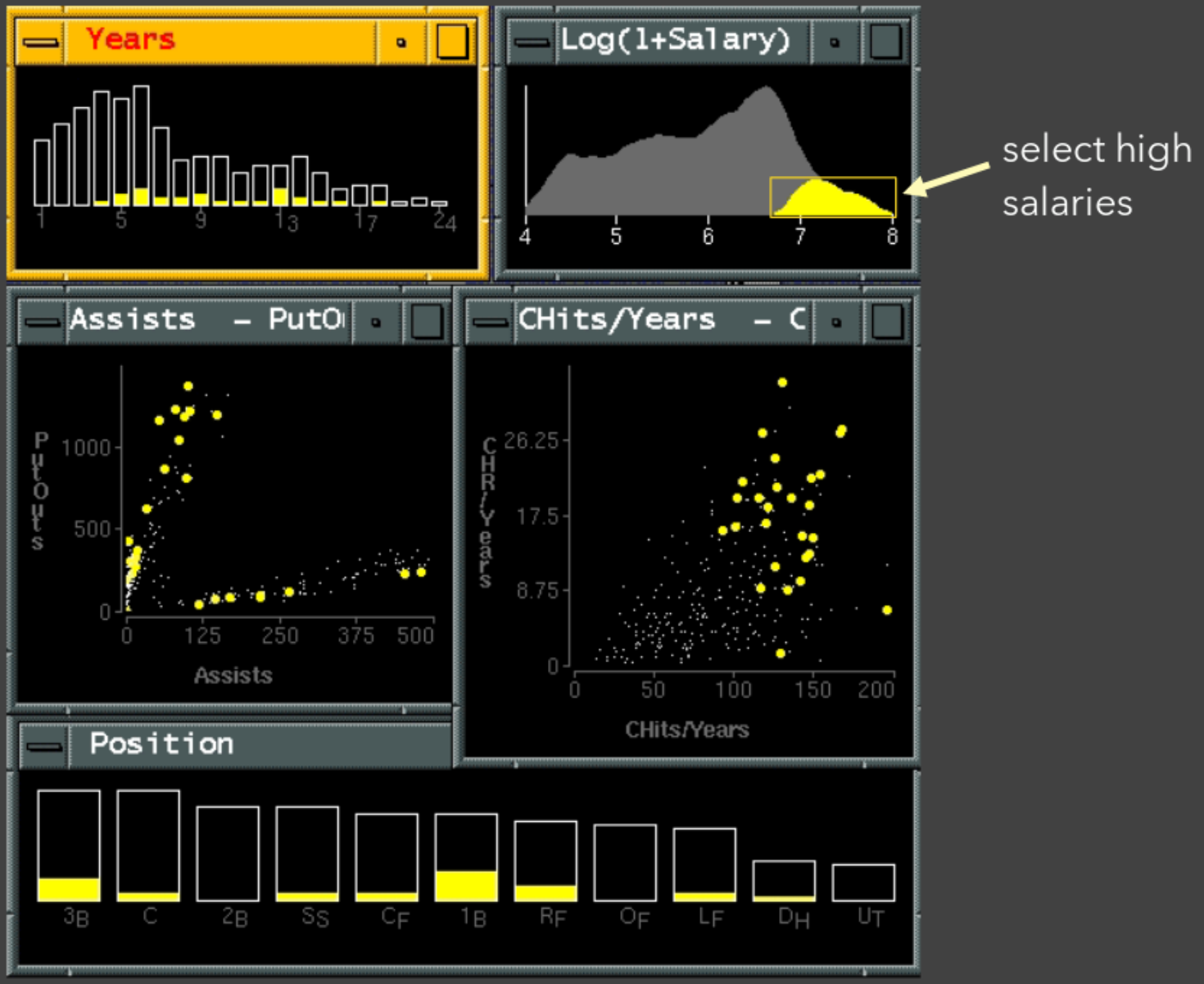
[MacEachren ’95, Figure 2.11, p. 38]

Scatterplot Matrix (SPLOM)

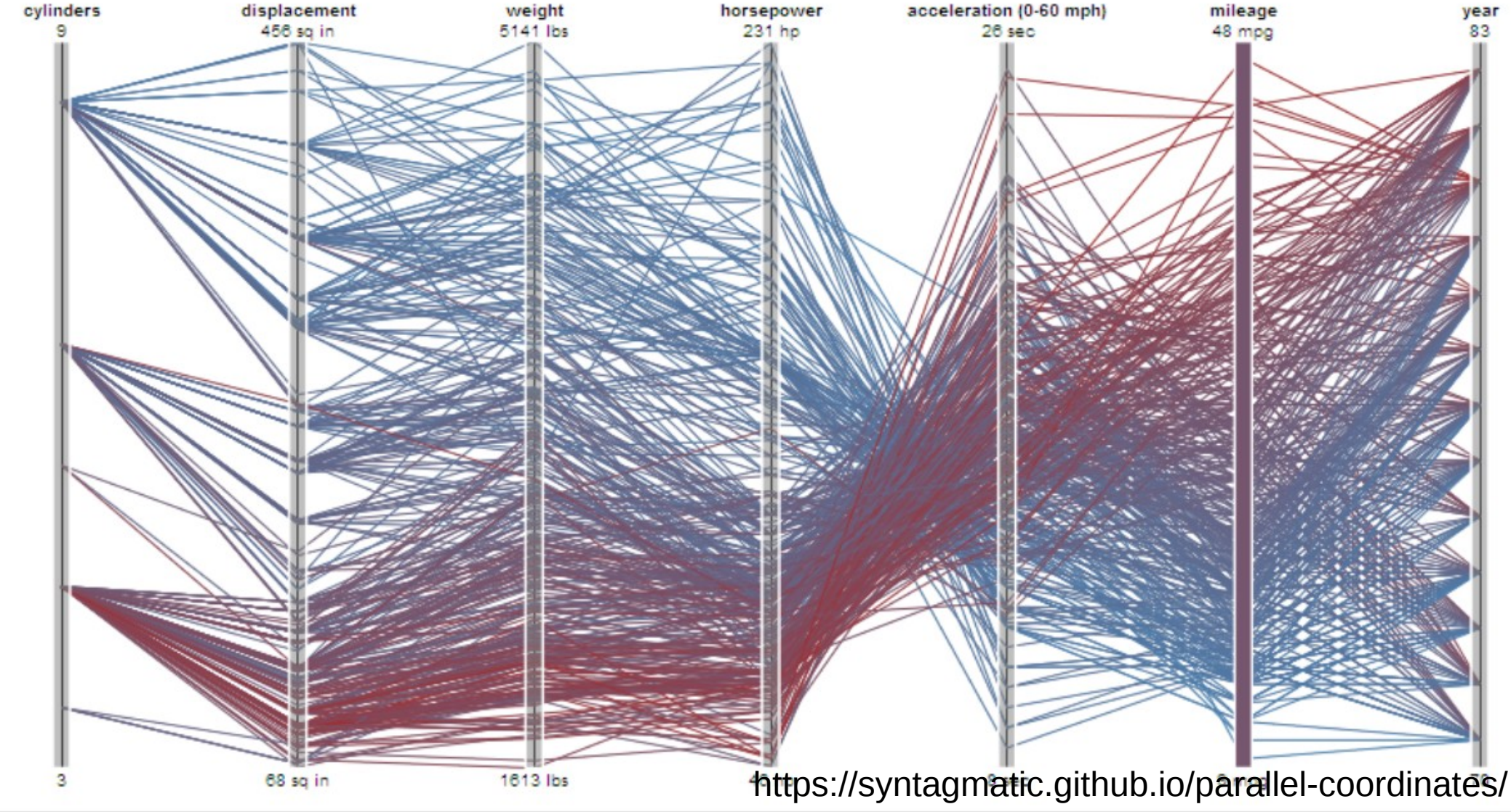


Scatter plots for pairwise comparison of each data dimension.

Multiple Coordinated Views



Parallel Coordinates [Inselberg]



https://syntagmatic.github.io/parallel-coordinates/

- Visualize up to ~two dozen dimensions at once
- 1. Draw parallel axes for each variable
- 2. For each tuple, connect points on each axis

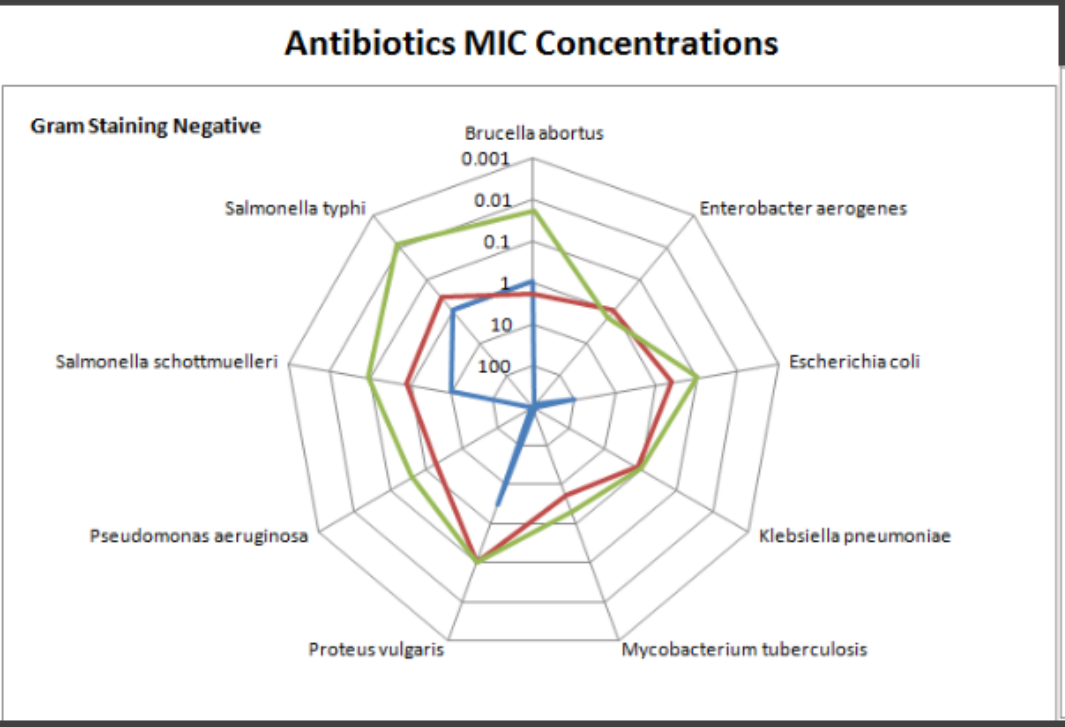
- Between adjacent axes: line crossings imply neg. correlation, shared slopes imply pos. correlation.

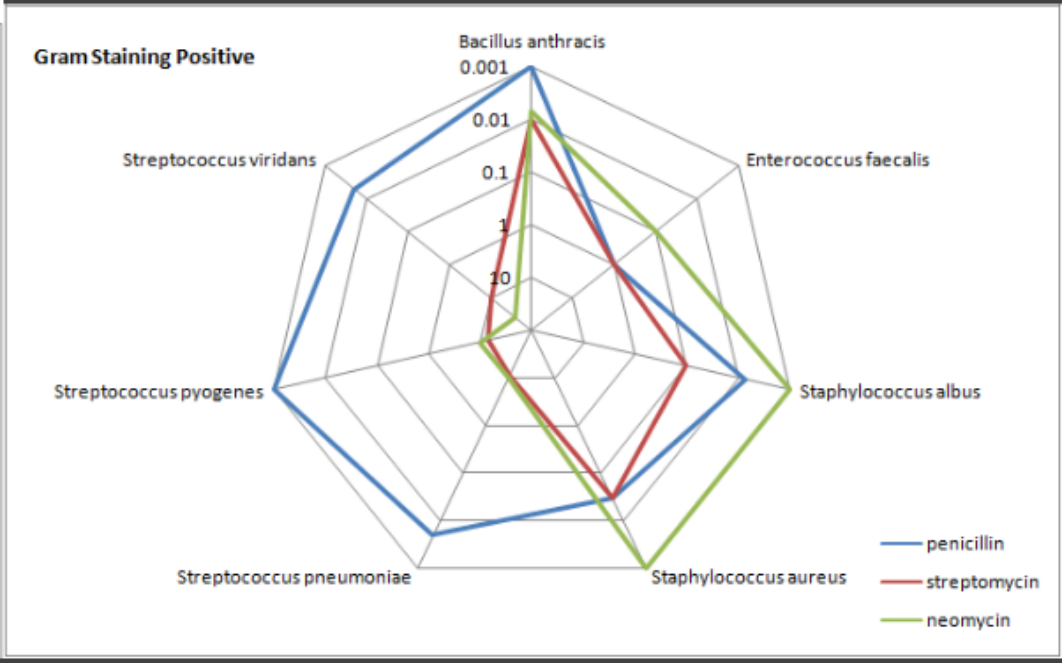
- Full plot can be cluttered. Interactive selection can be used to assess multivariate relationships.

- Highly sensitive to axis scale and ordering.

- Expertise required to use effectively!

Radar Plot / Star Graph





"Parallel" dimensions in polar coordinate space

Best if same units apply to each axis

Visual Encoding Design

- Use expressive and effective encodings
- Avoid over-encoding
- Reduce the problem space
- Use space and small multiples intelligently
- Use interaction to generate relevant views
- Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is critical!