



Erasmus+



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

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

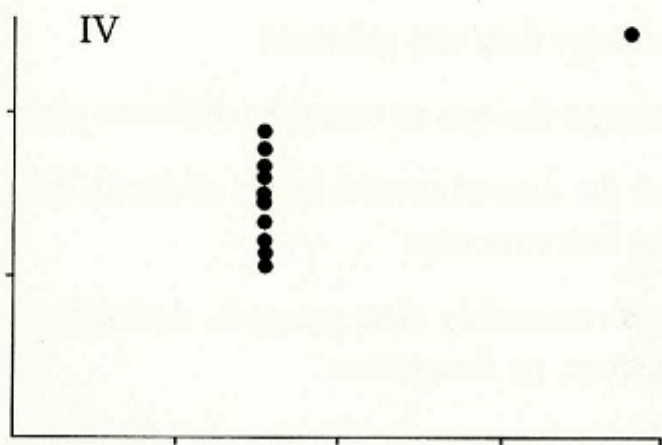
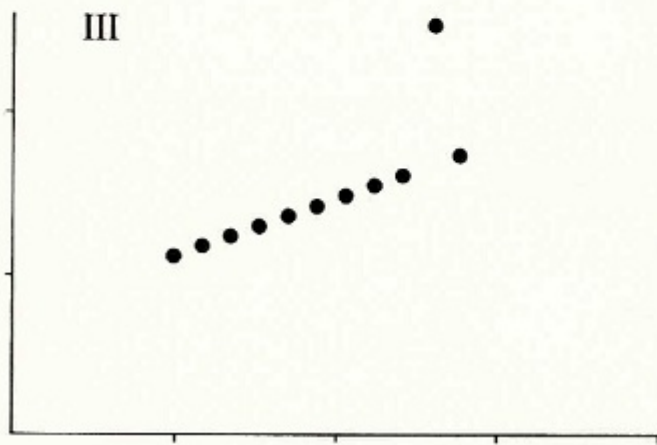
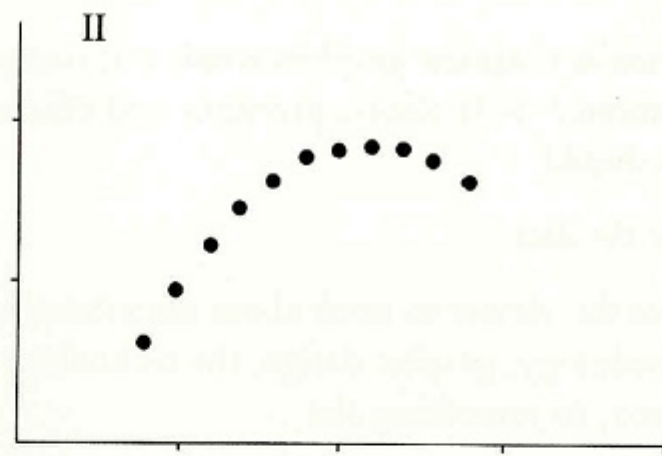
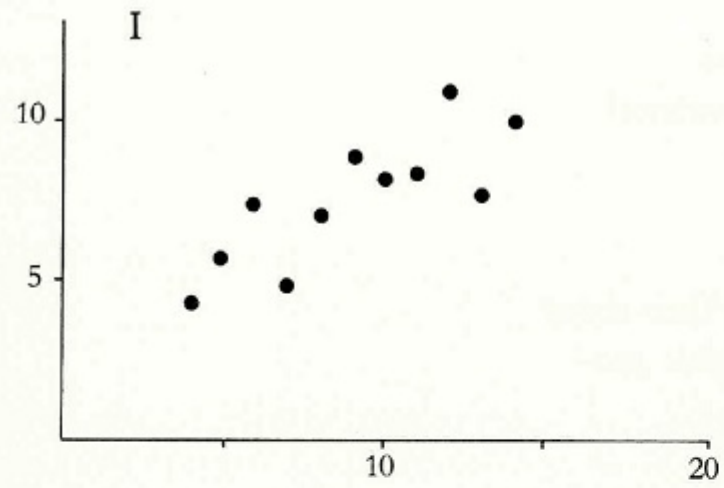
For all four datasets:

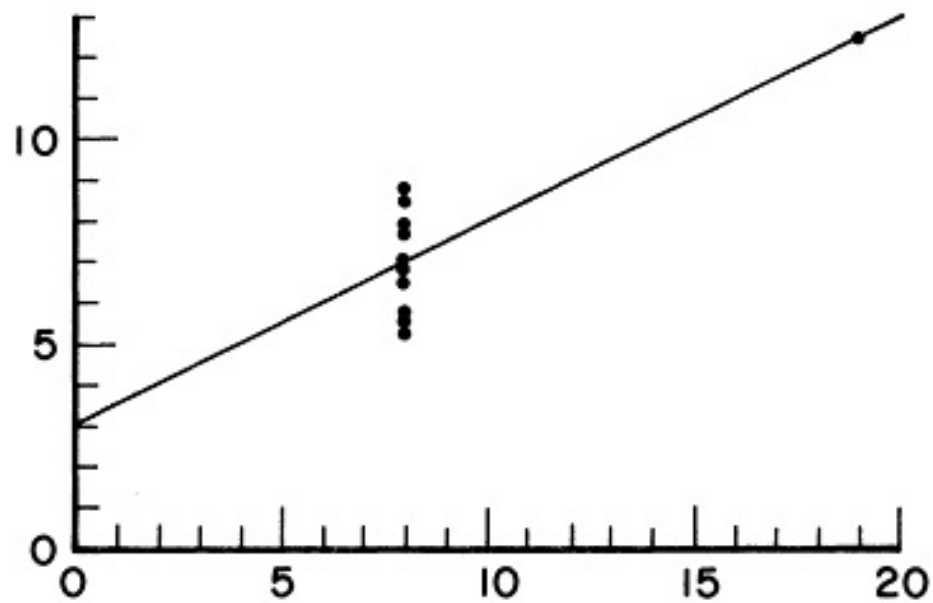
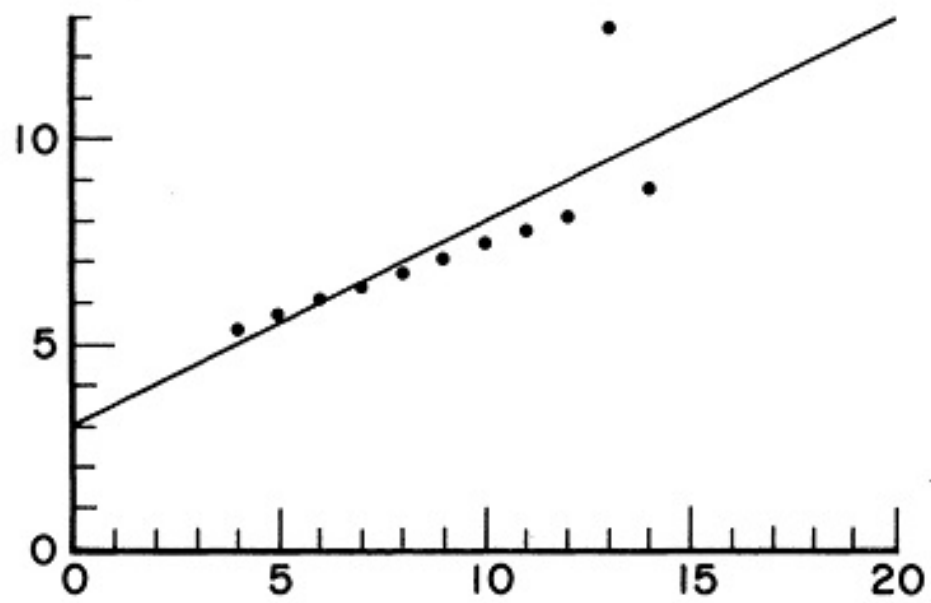
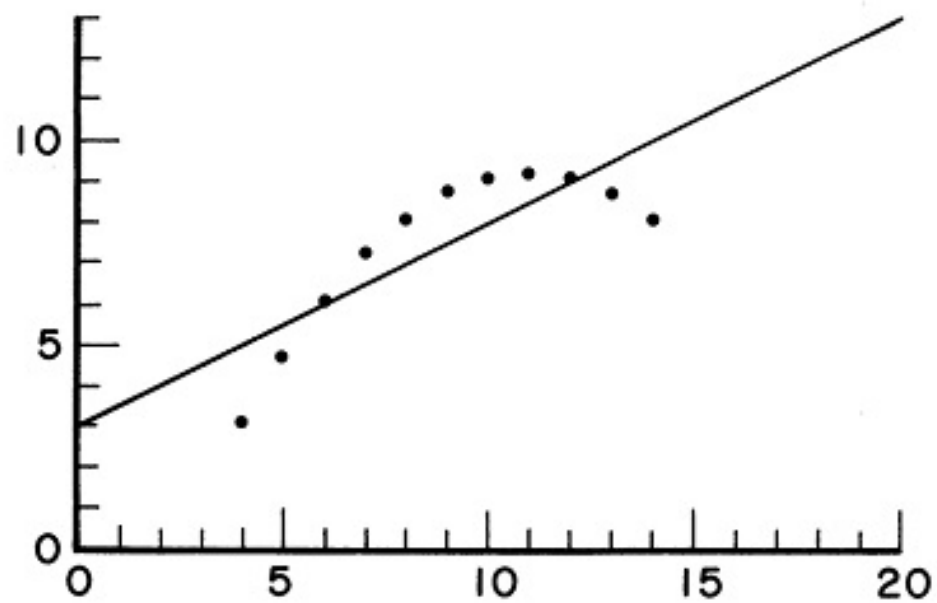
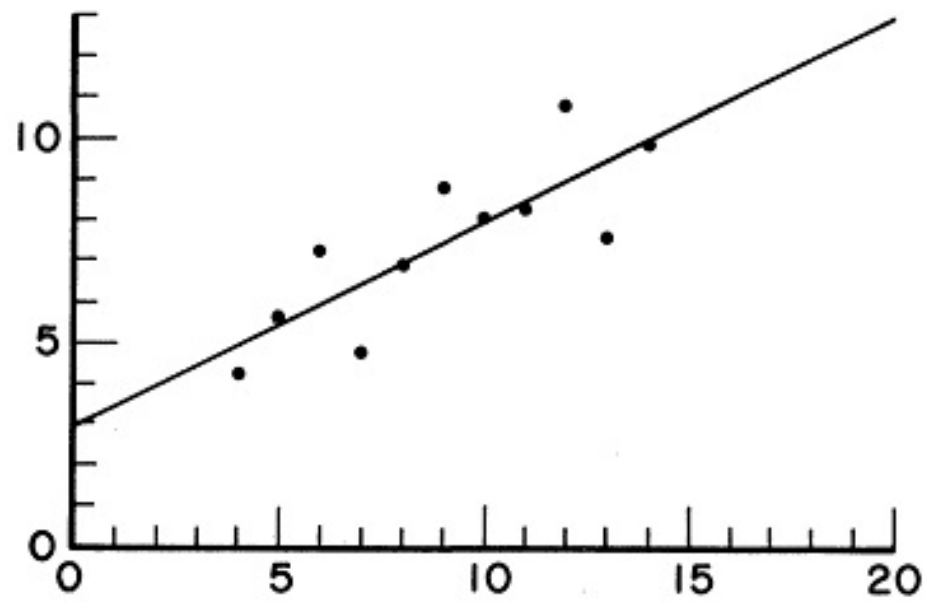
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	± 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination 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 $X - \bar{X} = 110.0$
 regression sum of squares = 27.50
 residual sum of squares of Y = 13.75
 correlation coefficient = .82
 $r^2 = .67$





Data in Context: Cholera Outbreak

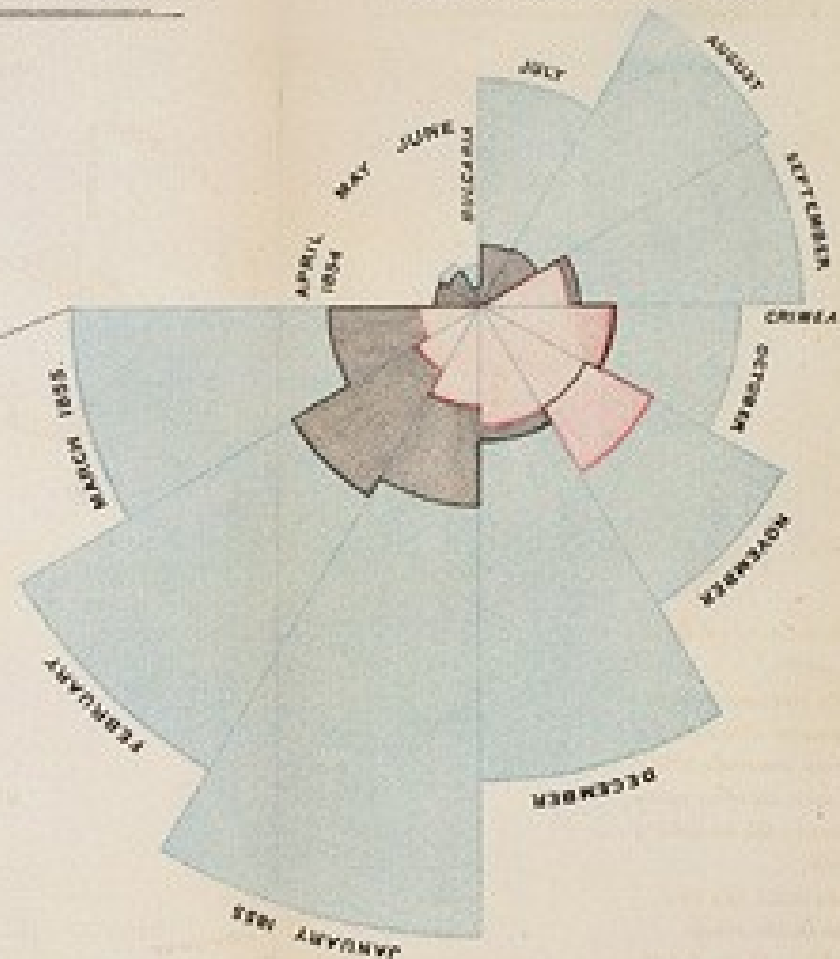
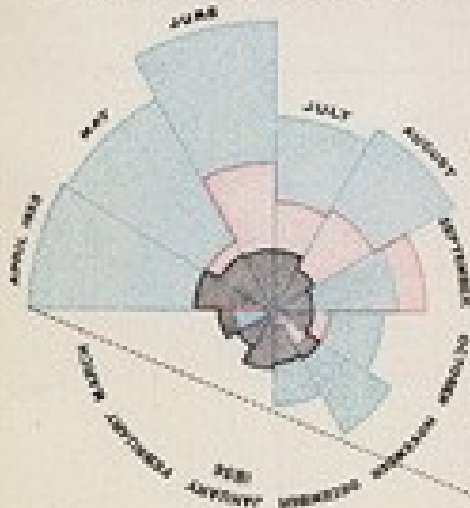


1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

2.
APRIL 1855 TO MARCH 1856.

1.
APRIL 1854 TO MARCH 1855.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

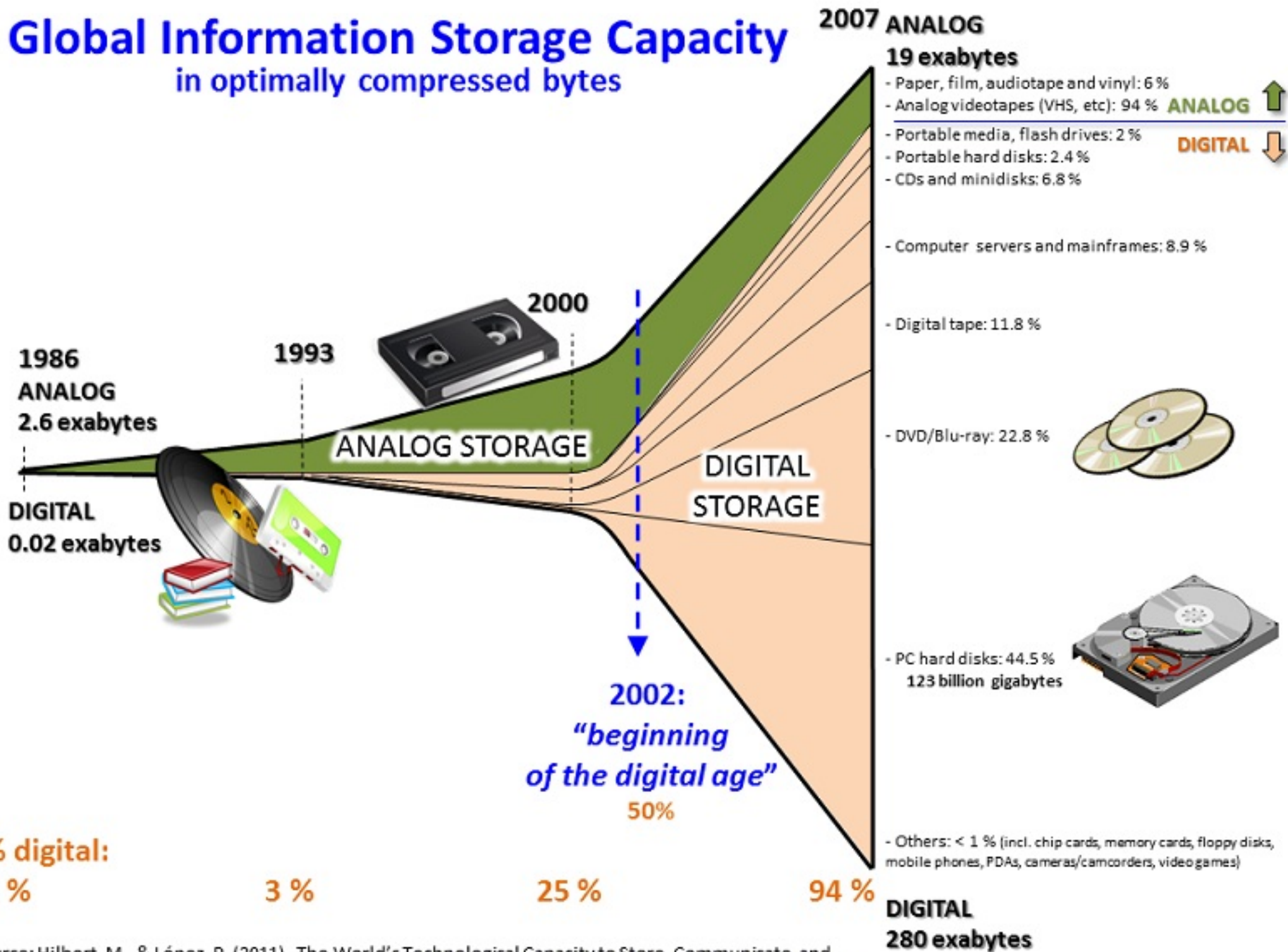
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre, the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red, in January & February 1855, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Global Information Storage Capacity in optimally compressed bytes



Why Data visualization

Answer questions (or discover them)

Make decisions

See data in context

Expand memory

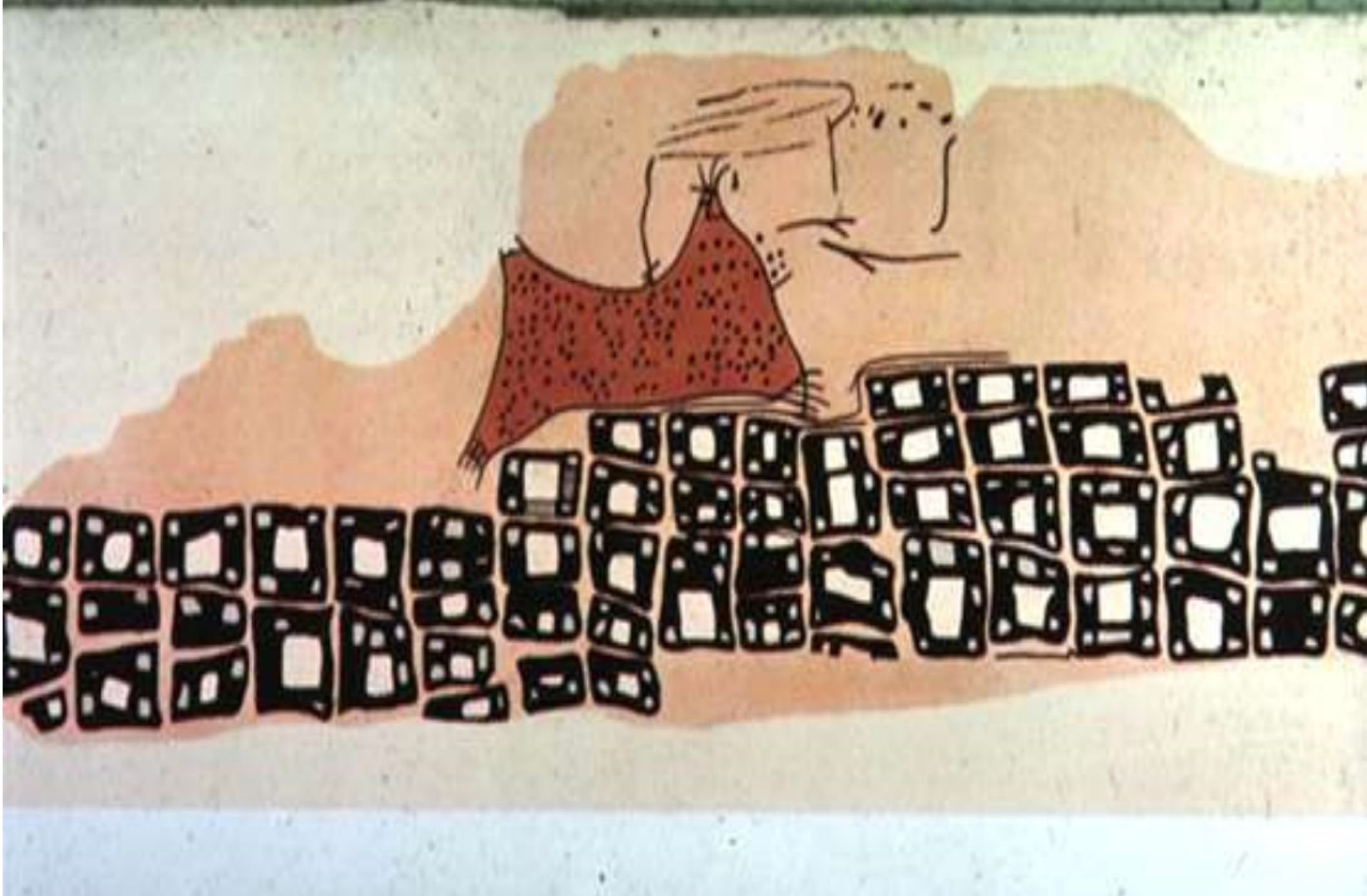
Support graphical calculation

Find patterns

Present argument or tell a story

Inspire

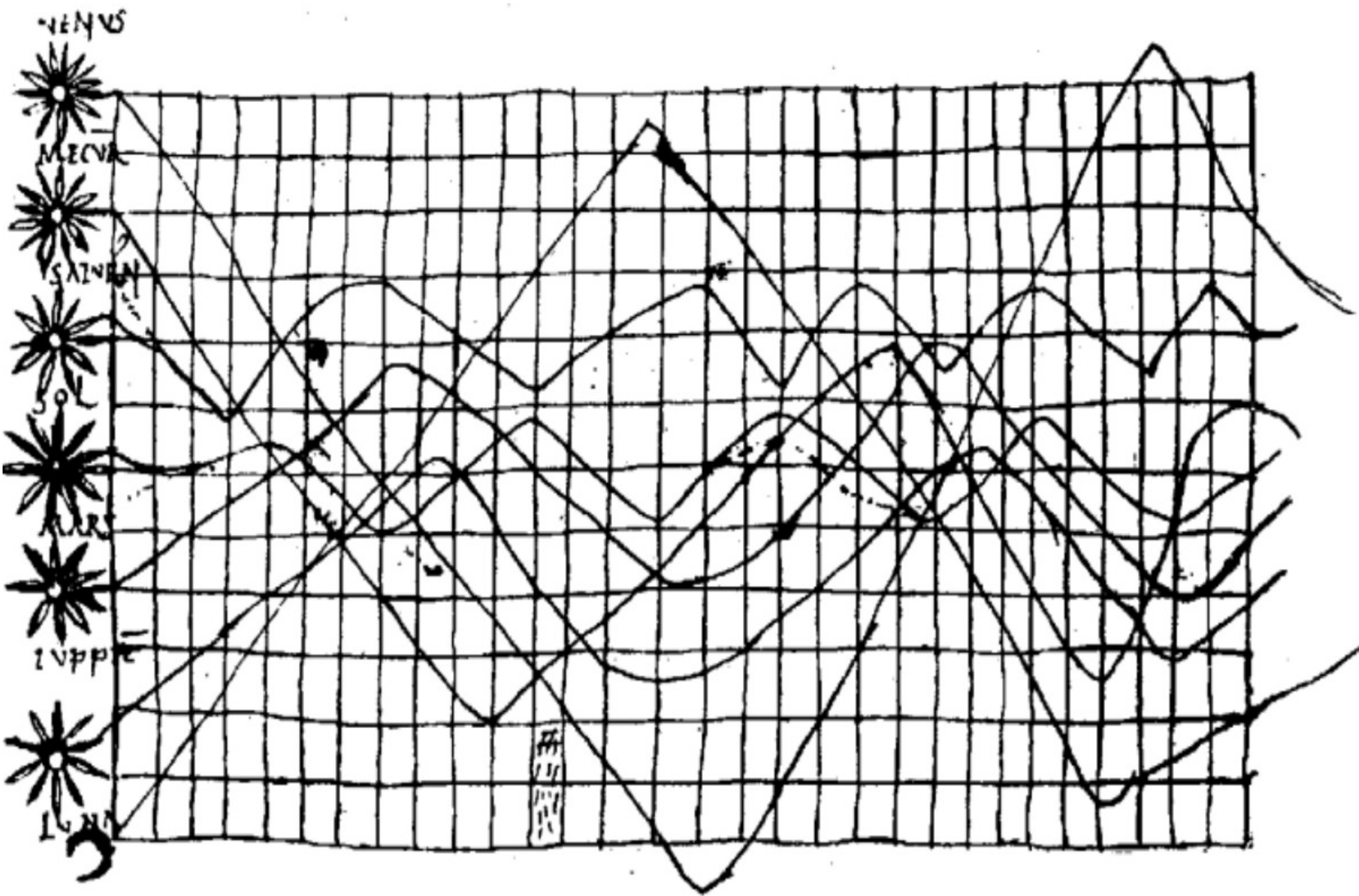
History of Data visualization



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

0 BC

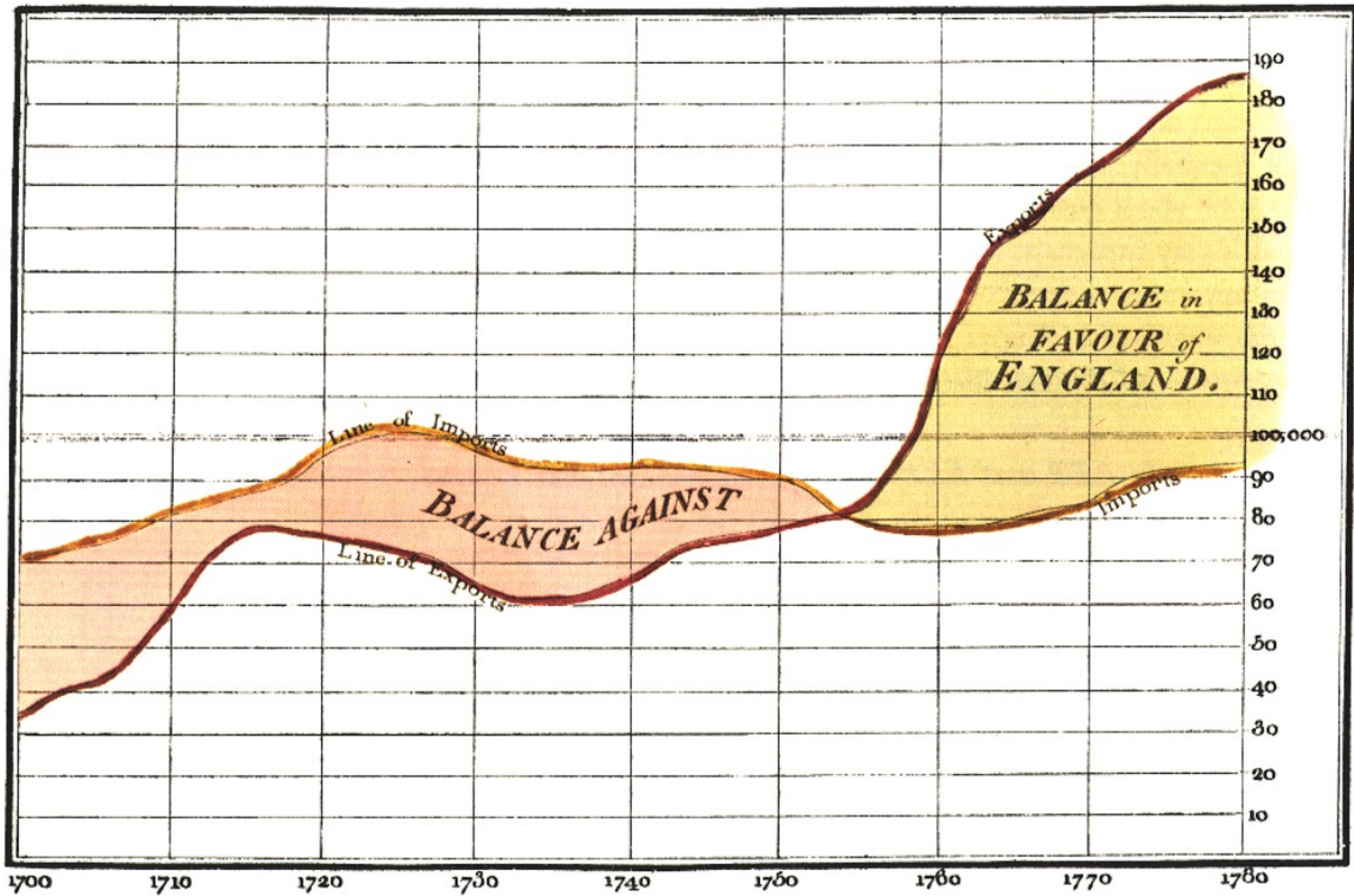




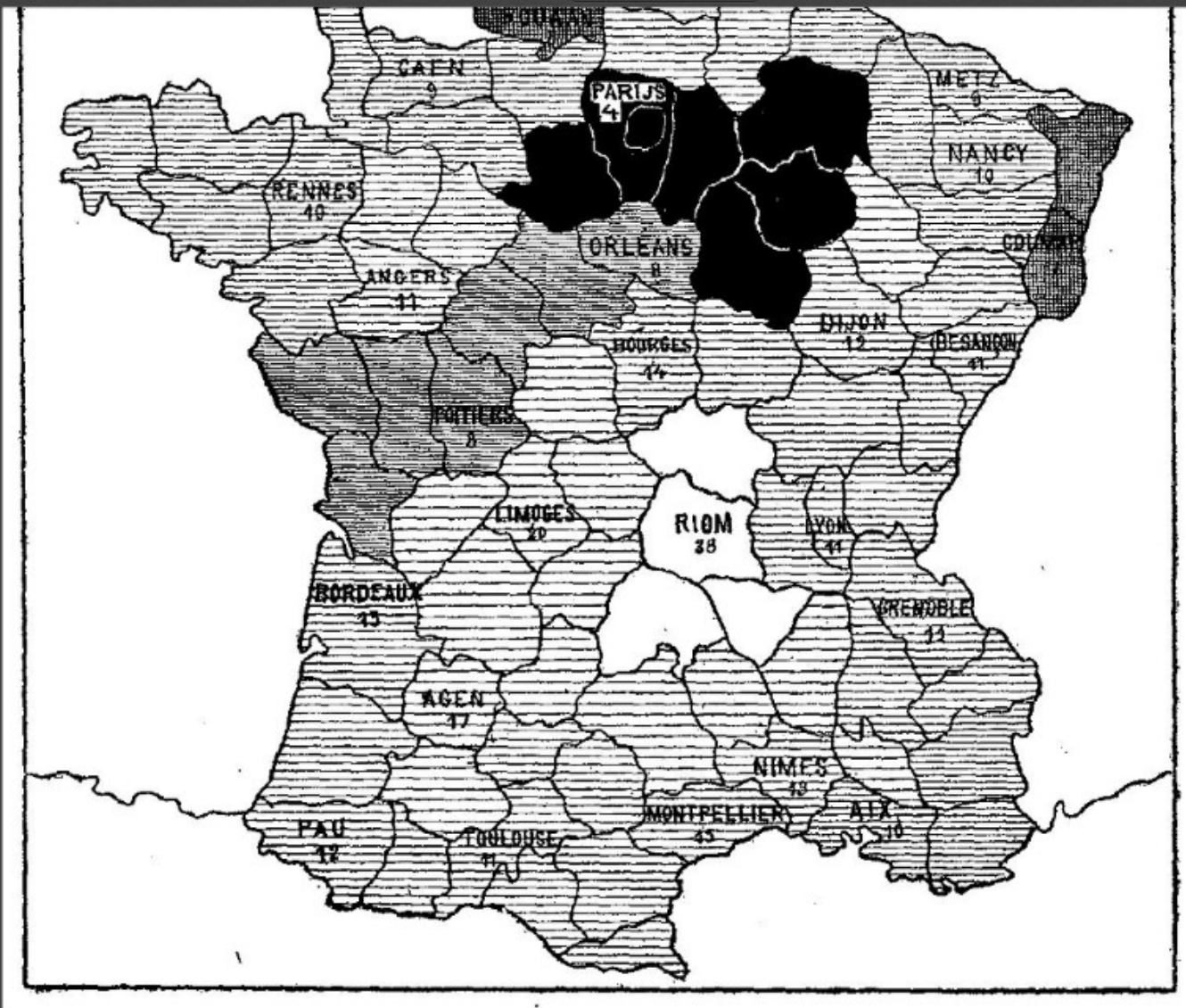
~950 AD Position of Sun, Moon and Planets



Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



- + Automatic Zoom ↕



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.

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
 Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre. Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

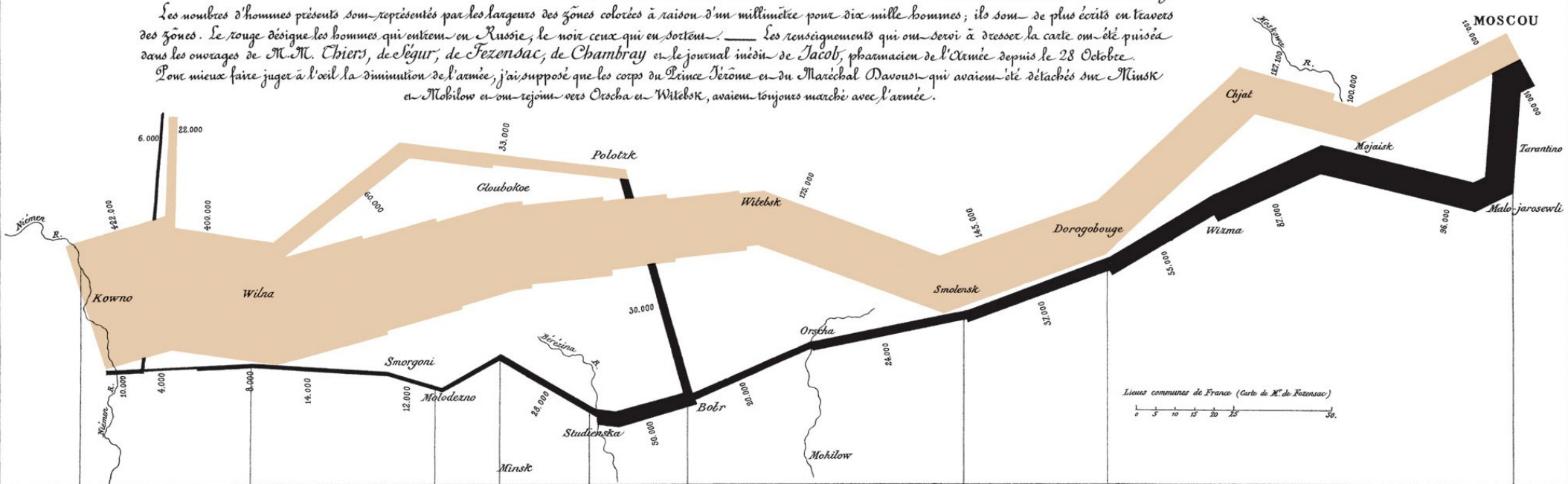
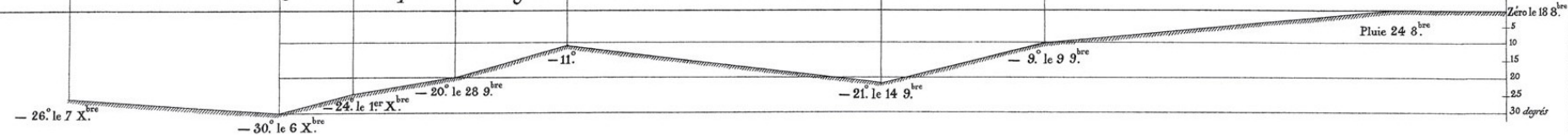
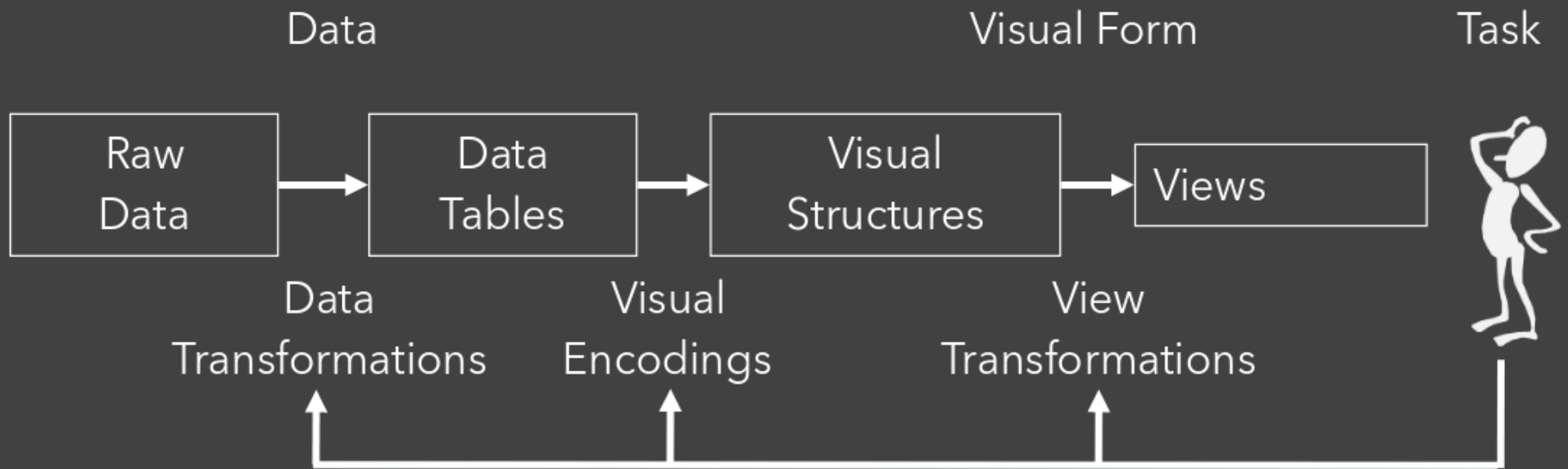


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



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?

		LAS VARIABLES DE LA IMAGEN						12 14			
		PUNTOS			LÍNEAS			ZONAS			
XY	2 DIMENSIONES DEL PLANO	x	x	x	/	~	/	14 15 9 10 21 2 2 14 15 1	2 1 18 2 1 21 15 1 1 2 9	OQ	≠
Z	TAMAÑO	█	█	█	/	~	/	█	█	OQ	≠
	INTENSIDAD	█	█	█	/	~	/	█	█	O	≠
		LAS VARIABLES DE SEPARACIÓN DE LAS IMÁGENES						13			
	GRANO	█	█	█	/	~	/	█	█	≡	≠
	COLOR	█	█	█	/	~	/	█	█	≡	≠
	ORIENTACIÓN	█	█	█	/	~	/	█	█	≡	≠
	FORMA	█	█	█	/	~	/	█	█	≡	≠

Multidimensional Data

Trellis Plots



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

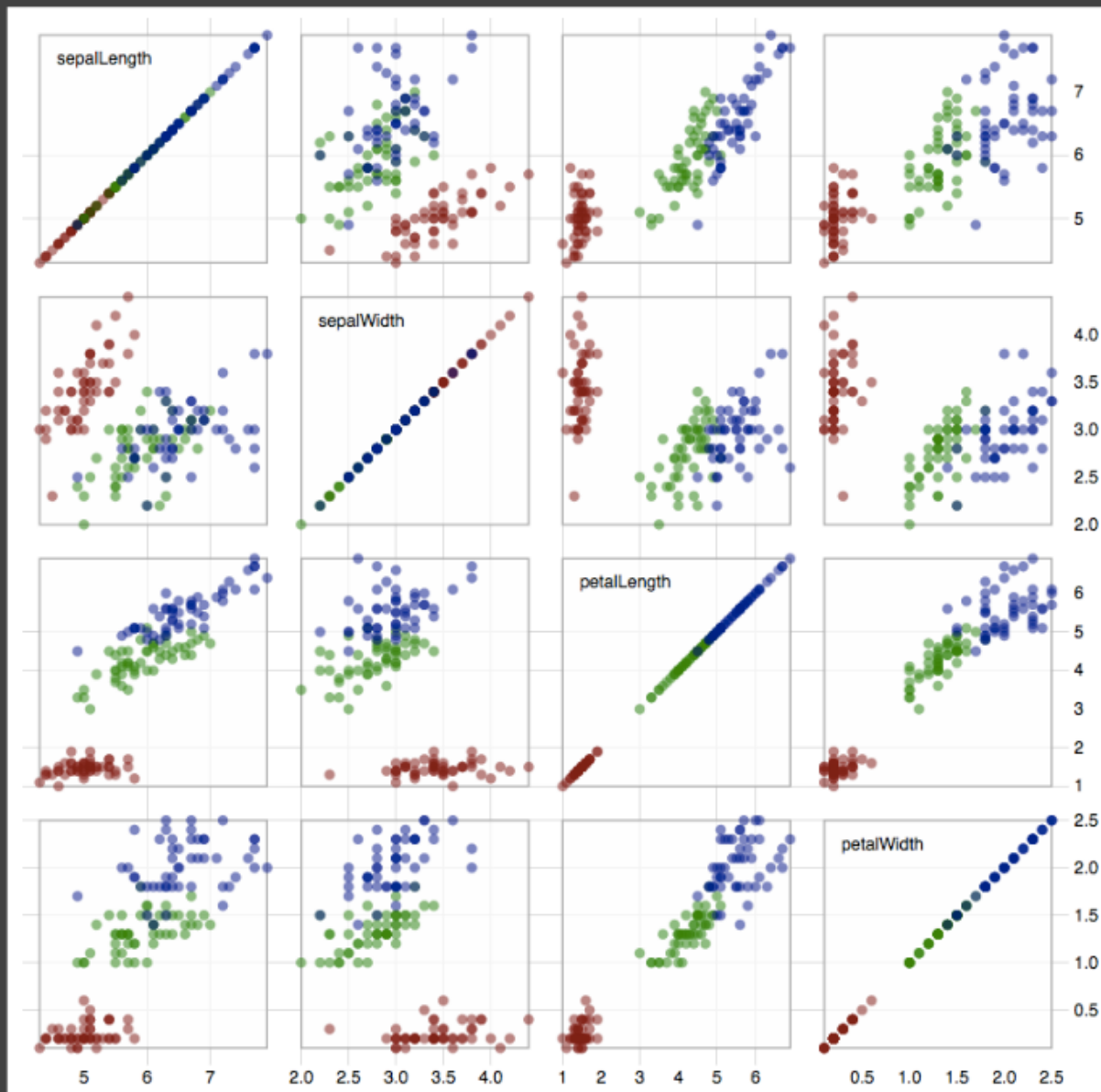
Typically nominal or ordinal variables are used as dimensions for subdivision.

Small Multiples



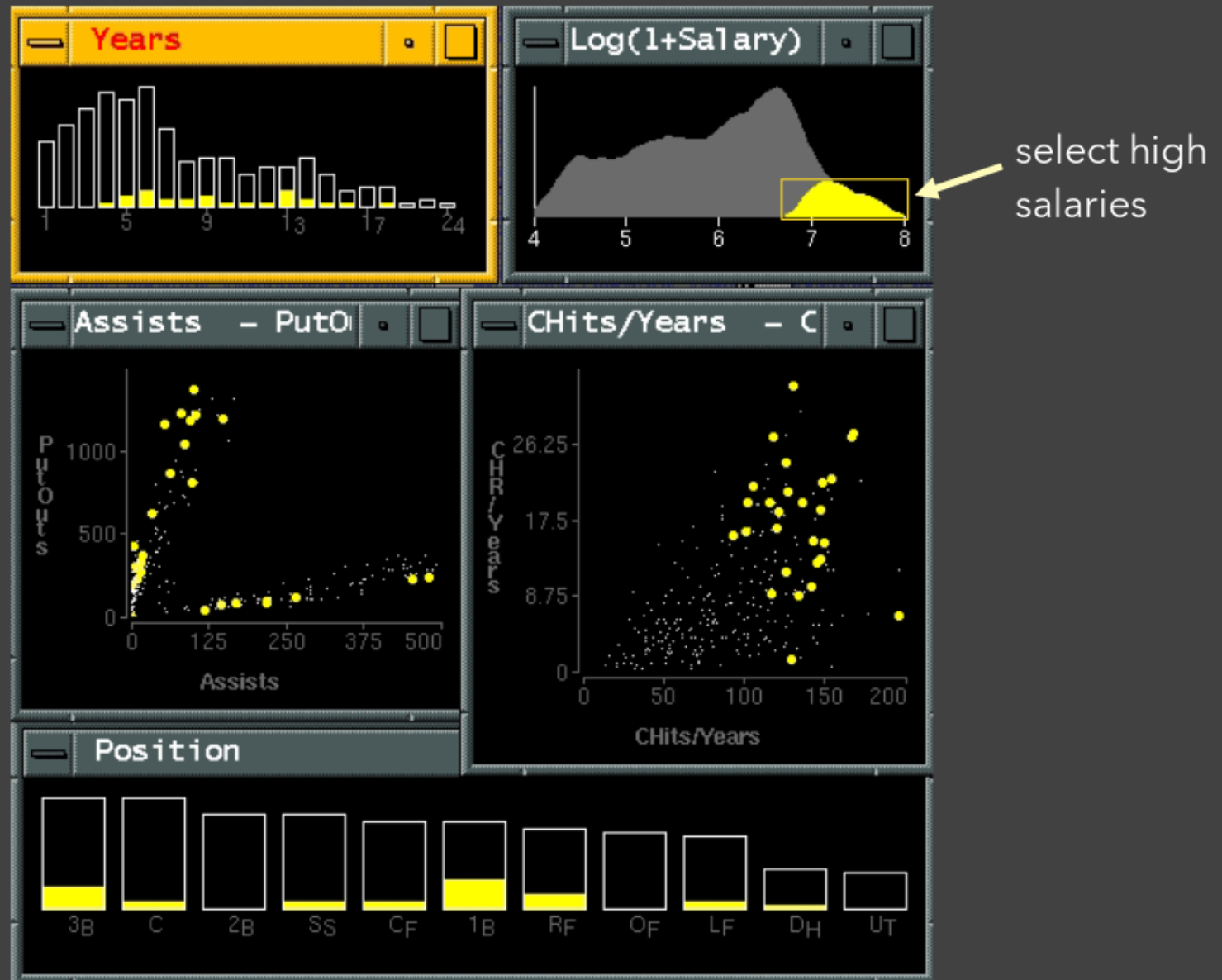
[MacEachren '95, Figure 2.11, p. 38]

Scatterplot Matrix (SPLOM)

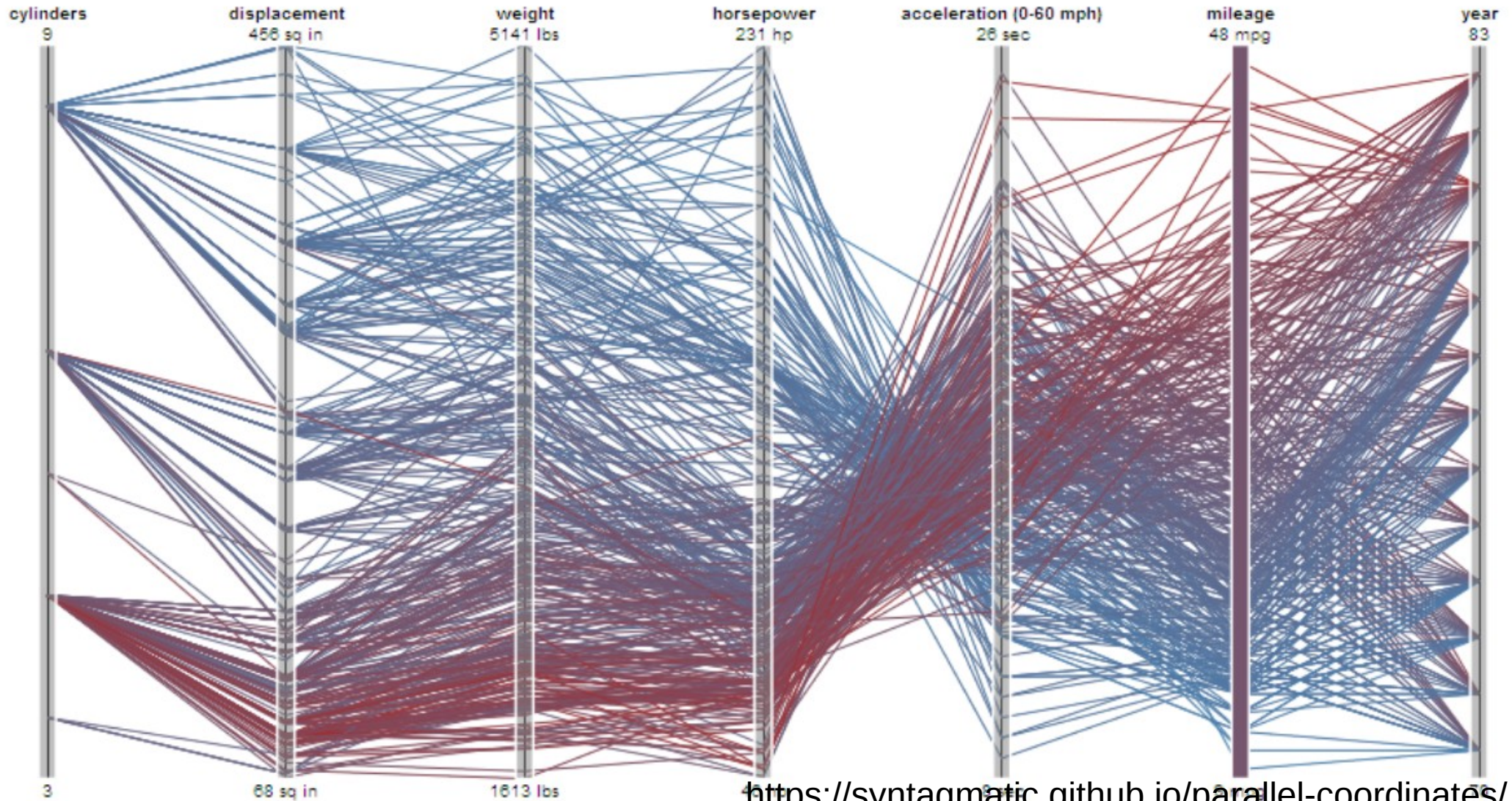


Scatter plots for pairwise comparison of each data dimension.

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.

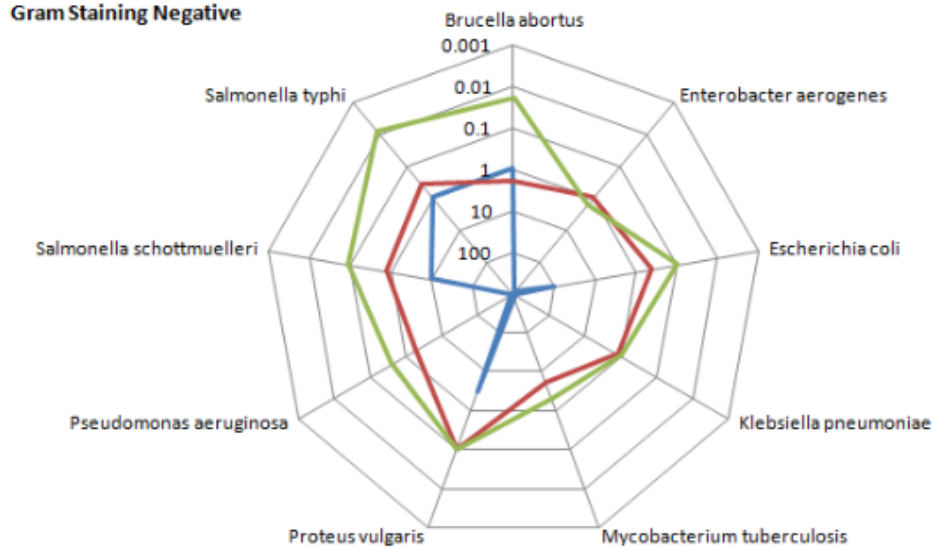
- Highly sensitive to axis scale and ordering.

- Expertise required to use effectively!

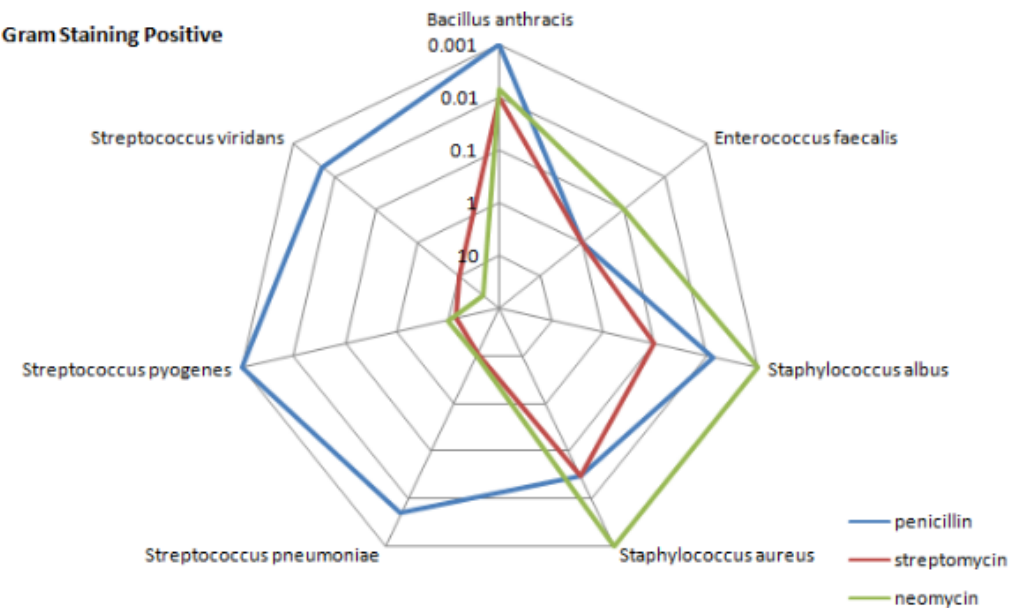
Radar Plot / Star Graph

Antibiotics MIC Concentrations

Gram Staining Negative



Gram Staining Positive



“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!