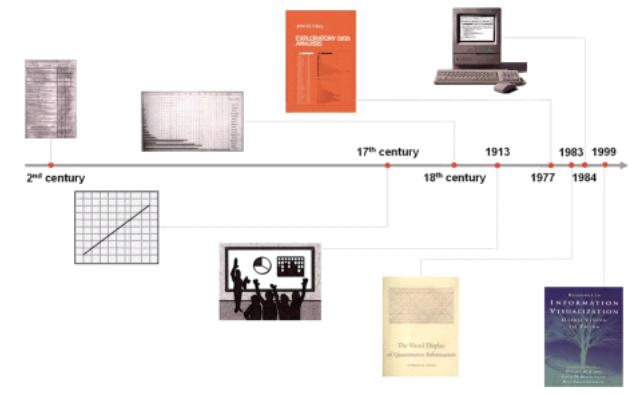


# Lesson 3

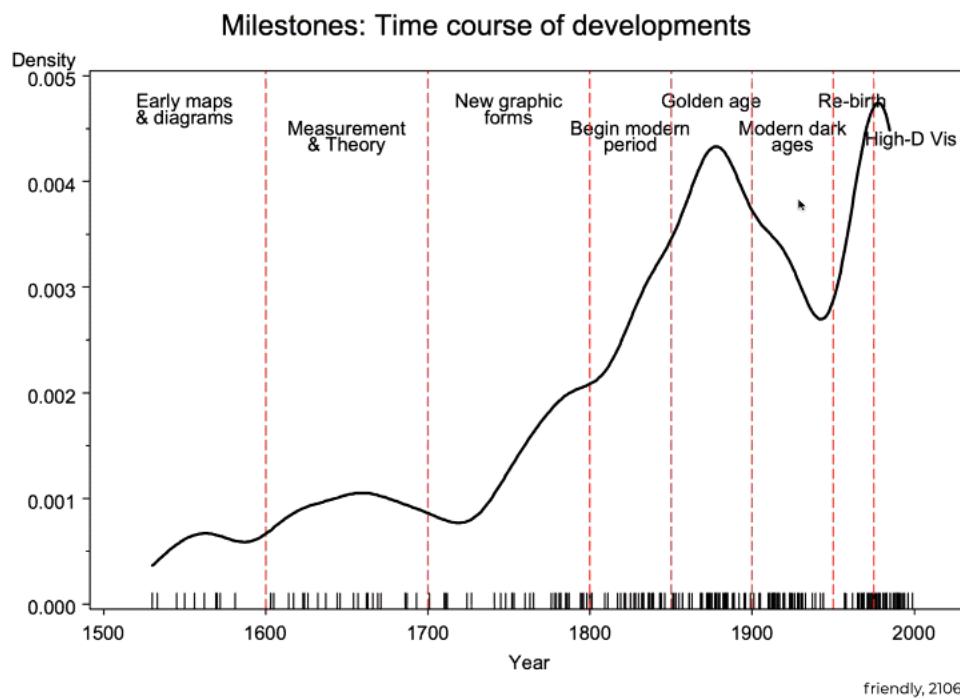
martedì 2 marzo 2021 11:30

A bit of history in order to understand what is data visualization.



Ritaglio schermata acquisito: 02/03/2021 11:31

Data visualization advance is characterised by some important steps.



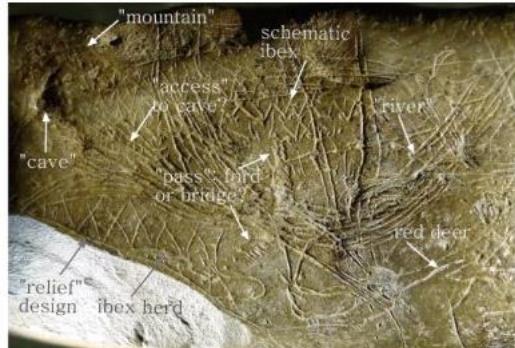
The above image shows different ages and the most important event that characterized the data visualization of that period.

## PRE-1600

We treat all this period as a single unit.

- The characteristics of this period are :
- Abundant of geometric diagrams
- Maps as practical tool for navigation and exploration
- Tables of the positions of stars and other celestial bodies

Ritaglio schermata acquisito: 02/03/2021 11:36



engraved stone blocks from the late magdalenian in abauntz cave (13660 calbp\*)



Journal of Human Evolution  
Volume 57, Number 1, August 2010, Pages 99–111



Review  
A palaeolithic map from 13,660 calBP: engraved stone blocks from the Late Magdalenian in Abauntz Cave (Navarra, Spain)

F. Liria, A.R. C. Alay, M.C. Segura, M. Martínez-Bon, R. Díaz-Orive



the meandering course of a river crossing the upper part of side a of the block, joined by two tributaries near two mountains

\*calibrated years before the present

This is a map: a course of a river. Local map, probably engraved by people living in the cave itself to show other people how was the surrounding territory.  
It is a map because the need for that people was to show how was the world outside. We can say that this is an accurate and detailed map. It is strange but there are also decorations. Although it is a practical tool the graphical aspect is still important.



Ritaglio schermata acquisito: 02/03/2021 11:40

This is another map. This shows a town. In the landscape we can also see a volcano. Archaeologists thanks to the representation of the volcano were able to localize the town..

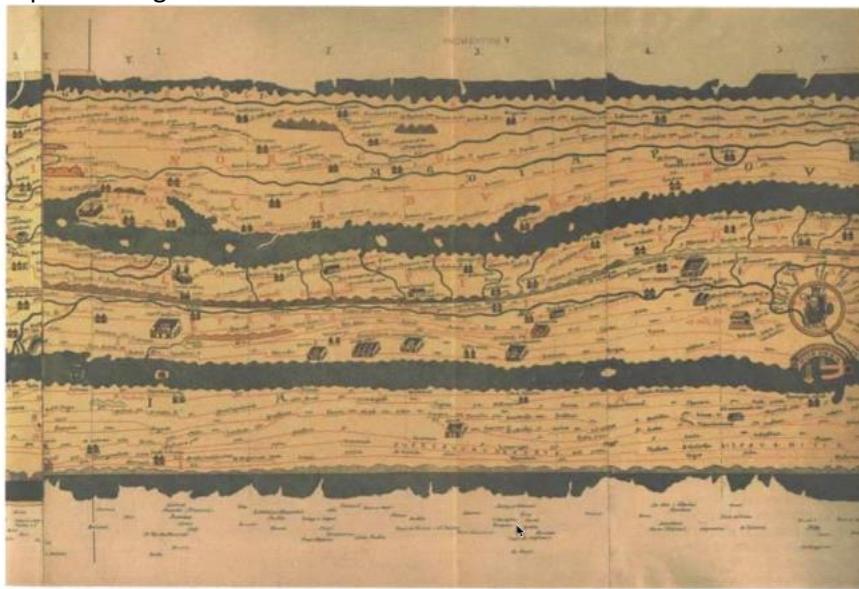
Local maps but also global maps. This is a quite accurate map for that period.





Ritaglio schermata acquisito: 02/03/2021 11:41

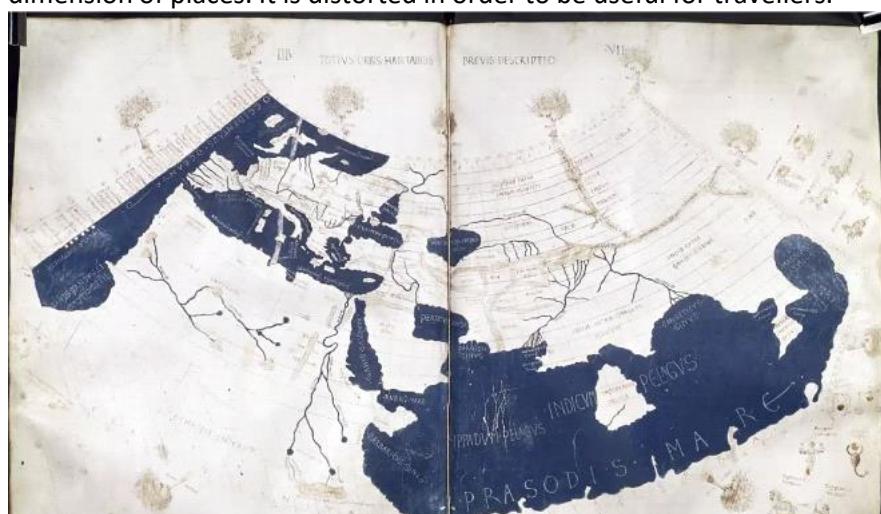
The most well represented territories are Italy, Greece and Turkey so we can guess that this map was designed in one of those territories.



Ritaglio schermata acquisito: 02/03/2021 11:44

Much more detailed map. This map is colored, it is incredibly detailed. In two hundred years there was a big development of topography. This map should describe all the roman world.

However, it is more a tool than a representation. It gives information about distances, dimension of places. It is distorted in order to be useful for travellers.

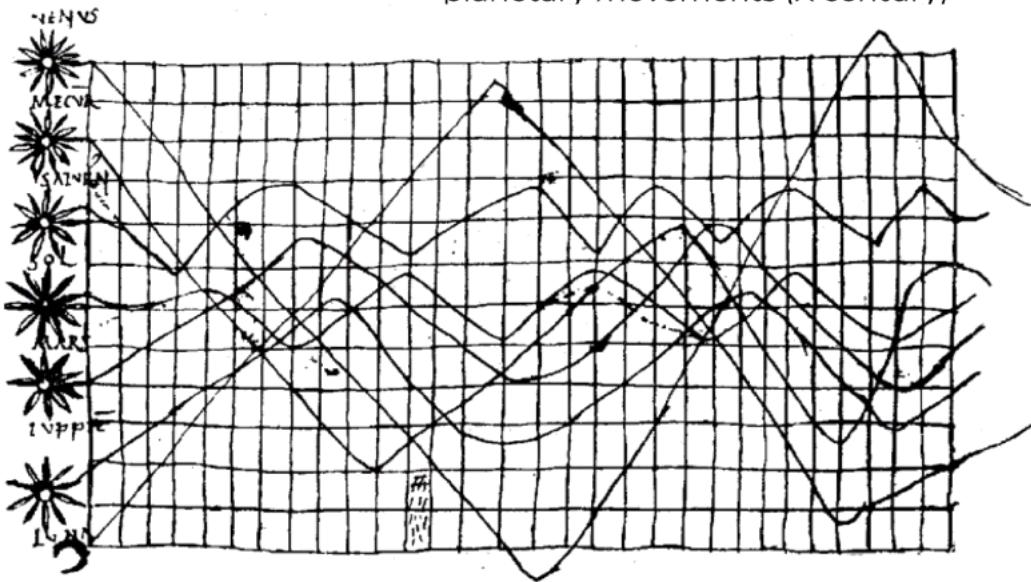




Ritaglio schermata acquisito: 02/03/2021 11:47

This is a world map. It is a big achievement for the time. The level of details is incredible not only for europe but also for eastern Asia. In this map we can see the first use of longitudinal and latitudinal lines.

planetary movements (X century)



Ritaglio schermata acquisito: 02/03/2021 11:50

This is the first example of data representation. It is not a map !

If we have to interpret it we can say that it is a multiple time series graph of changing position of the seven most prominent heavenly bodies over space and time.

Horizontal scale is independent for each planet. We can see also the use of annotations.

## artifitium electionis personarum (r. Hull, 1280)

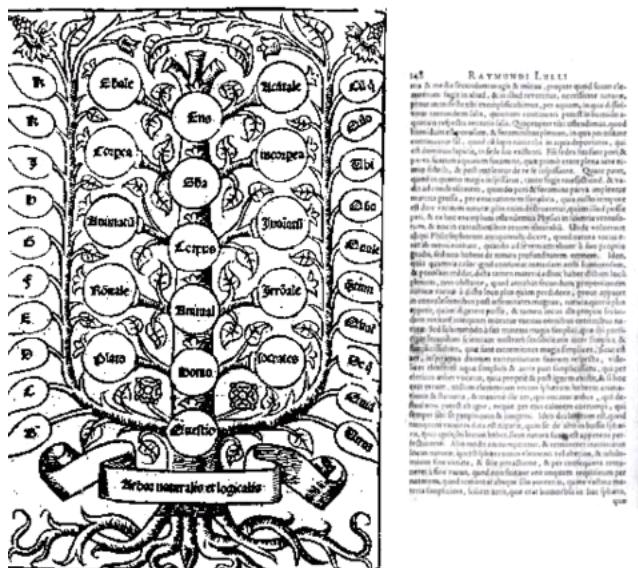
De est figura recte tibi his constraintis ac ratiocinationis regi  
apposita quod sentiuntur electio. Letatimque hinc hoc de  
finitus est. Lumen non potest esse nisi sit manifestum adiutorium  
rationis quod sentiuntur repetitio et ratiocinationis regi  
regi quod sentiuntur operatio. Invenimus igitur hanc ipsam rationem  
et ratione illius sentiuntur operatio. Propter

185  
re quidam fig. sigl. mtr. pqr. lmp. poterit dlig. h[ab]et  
j[ur]is allegant. pl[ac]ita ut p[ro]m[iss]io q[ui]t i[st] electione q[ui] n[on]  
fuit fons u[er]o s[ecundu]s poterit c[on]trafigi.

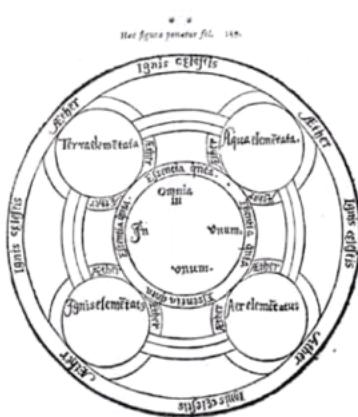
Ritaglio schermata acquisito: 02/03/2021 11:53

For the first time in a book there is a semi-graphic table. It is a reference table that summarizes the results of an election.

mechanical diagrams of knowledge (r. Hull, 1305)



## tree of knowledge



mechanical disks

Ritaglio schermata acquisito: 02/03/2021 11:54



bar graph (n. oresme, ~1350)

proto-bar graph



logical relation between tabulating values, and graphing them

"if a pioneering contemporary had collected some data and presented oresme with actual figures to work upon, we might have had statistical graphs four hundred years before playfair." [funkhauser, 1936]

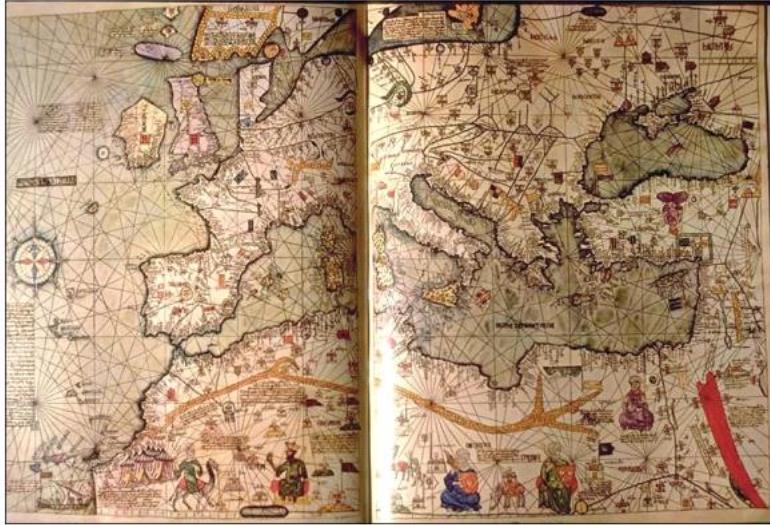
1450: n. da cusa plotted graph distance ~ speed (theoretical relation)

Ritaglio schermata acquisito: 02/03/2021 11:54

For the first time in history Erasmus introduced graphs in his work. We can see a prototype of bar graph. This graph are in the middle of the description.

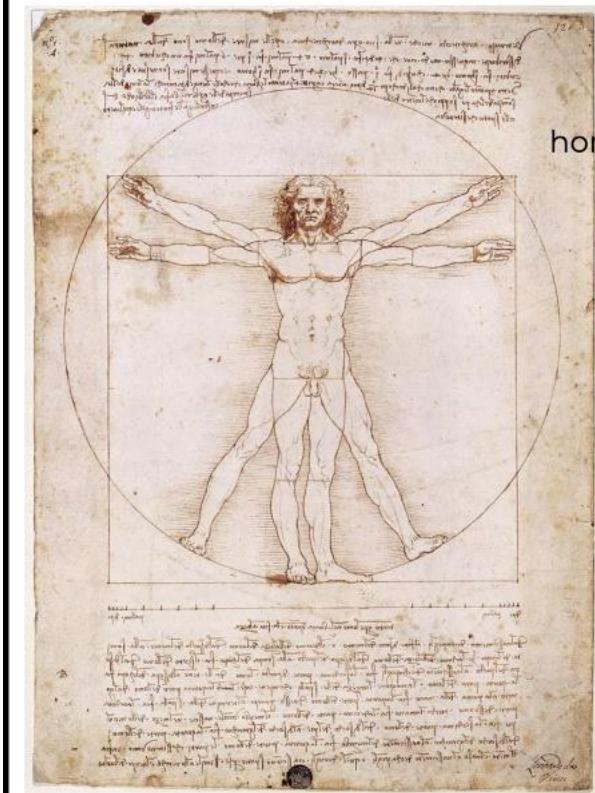
Middle age precursor of modern data visualization.

catalan atlas (a. cresques, 1375)



Ritaglio schermata acquisito: 02/03/2021 11:57

Cartography stil went on. There is also a tematic representation. This map is decorated so it went behind being a practical tool. It is an illustration of the world. There is also a different pourpose for the use of the map.



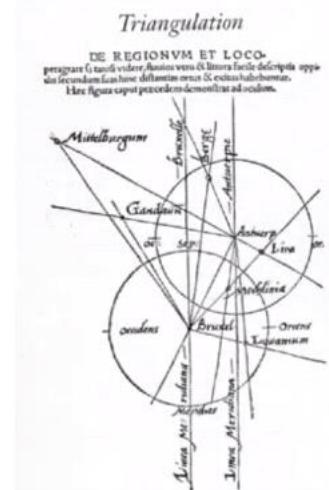
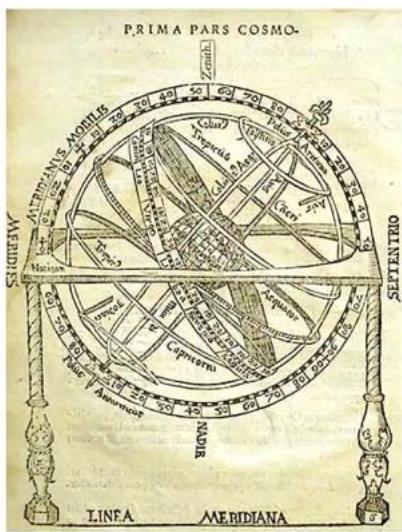
homo vitruvianus (l. da vinci, 1487)

correlations of ideal human body proportions with geometry described by the ancient roman architect vitruvius in book III of his treatise *de architectura*.

Ritaglio schermata acquisito: 02/03/2021 11:58

Data can be of every nature. Here we have anatomy, geometry, proportion. This is the ancestor of anatomy.

triangulation (r. gemma-frisius, 1533)



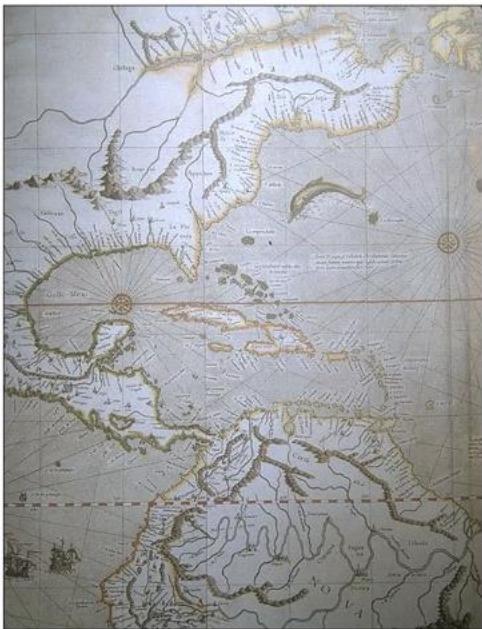
mapping locations by triangulation, from similar triangles, and with use of angles w.r.t meridians

Ritaglio schermata acquisito: 02/03/2021 12:00

1600 modern way of exploring the world. So it is important to understand where you are in a precise moment.

This are two representation which use triangulations ( angles and meridians)

cylindrical projection (g. mercator, 1569)



cylindrical projection for portraying the globe on maps, to preserve straightness of rhumb lines

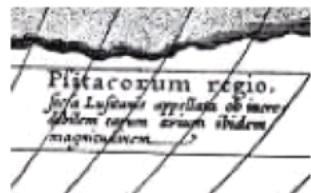
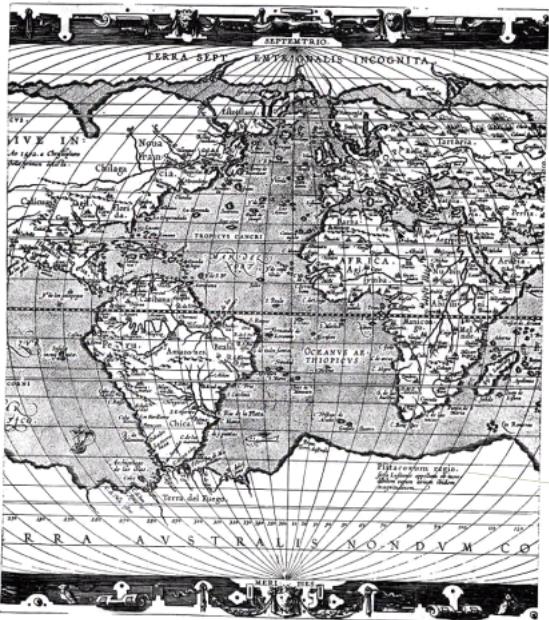
World Mercator projection with true country size added



Ritaglio schermata acquisito: 02/03/2021 12:01

A big step occurred when geographer understood that you can represent the world on a plane. So they developed the concept of projection which is used still today. Mercator projection for northern Europe and America distorts the real size.

first modern atlas (a. ortelius, 1570)



Ritaglio schermata acquisito: 02/03/2021 12:04

Atlases is a collection of maps: global map and extraction of local map from the bigger one.

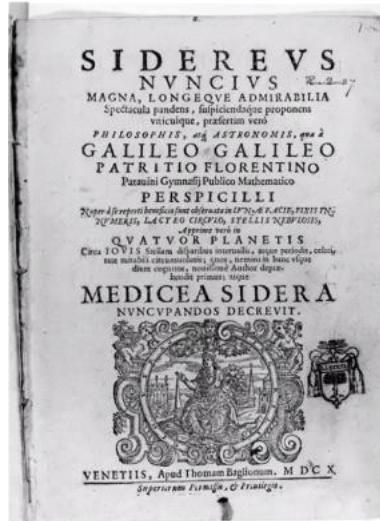
1600 was characterized by huge differences:

beginnings of visual thinking: emergence of quantitative tools (distances and measurement)

Physical measurement

Astronomy map making, navigation and territorial expansion  
 Growth in theory and dawn practice  
 Analytic geometry, theory of errors, probability theories  
 Demographic statistics

printed astronomical images (g. galilei, 1610)



telescope observations



craters on the moon, the 4 satellites of jupiter and a vast number of unseen stars

Ritaglio schermata acquisito: 02/03/2021 12:08

Galileus in his several works inserted really interesting pieces of representation.  
 He drawn what he saw through his telescope.

first logarithmic [gunter] scale (e. gunter & w. oughred, 1620/1628)



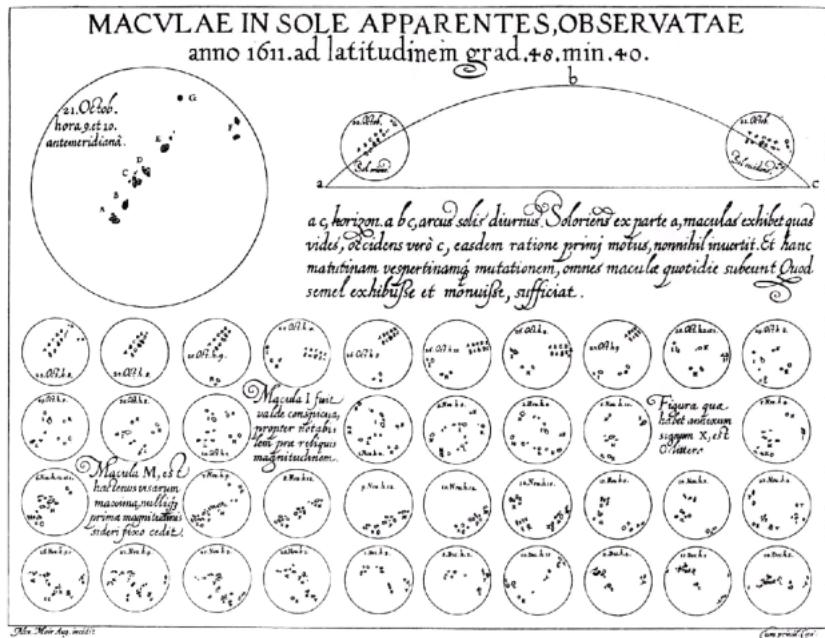
1 2 3 4 5 6 7 8 9 1



Ritaglio schermata acquisito: 02/03/2021 12:10

For the first time we have a reference scale.  
 Here we have logarithmic scale. This should have helped people in their operation of calculating logarithms.

## sunspots (c. scheiner, 1626)

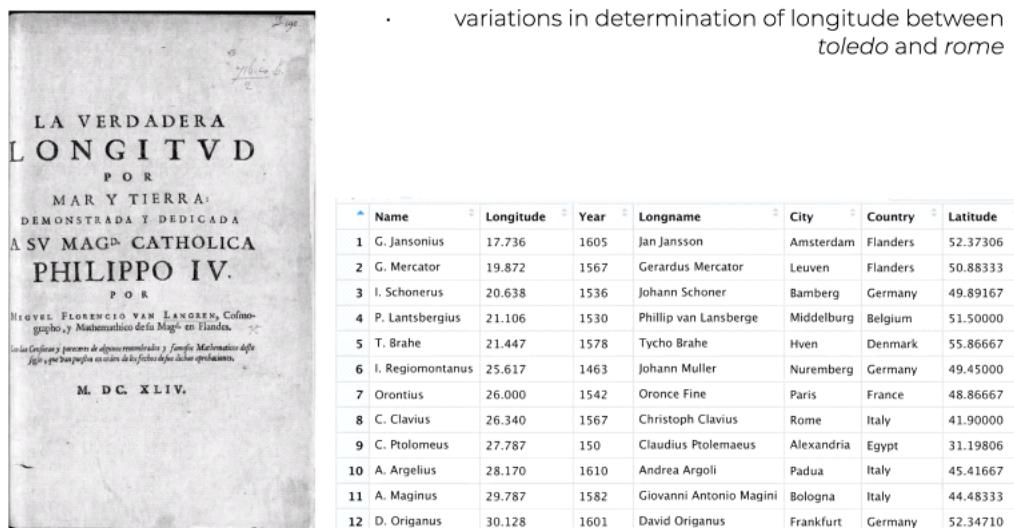


Ritaglio schermata acquisito: 02/03/2021 12:11

Here we have a masterpiece.

This representation is devoted to sunspots. This representation has three panels; first introduction of small multiples, repetition of same framework; 7 groups of sunspots in a large subplot are later repeated in the following 37 panels.

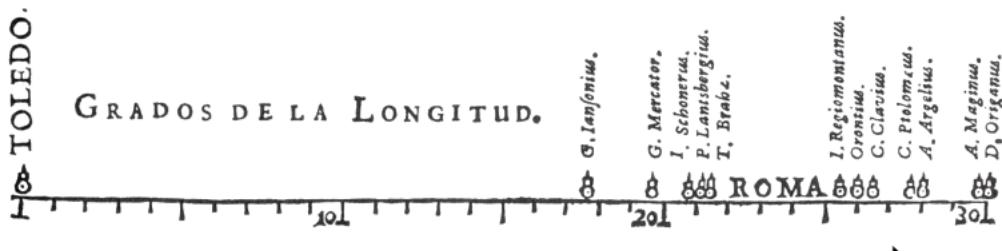
## first data graph (m.f. van langren, 1644)



Ritaglio schermata acquisito: 02/03/2021 12:14

### first data graph (m.f. van langren, 1644)

- variations in determination of longitude between toledo and rome



Ritaglio schermata acquisito: 02/03/2021 12:15

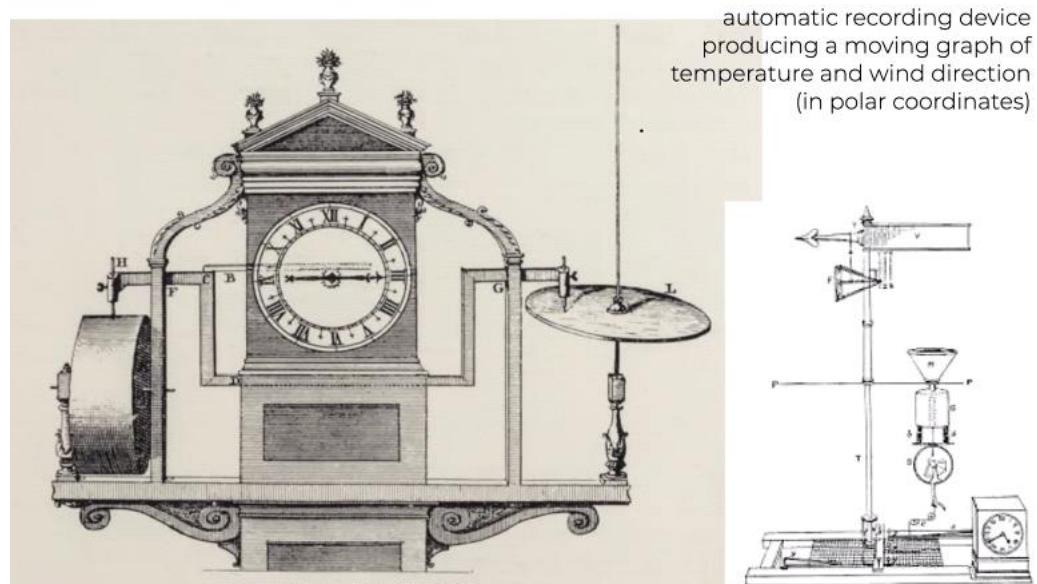
### first data graph (m.f. van langren, 1644)

- variations in determination of longitude between toledo and rome



Ritaglio schermata acquisito: 02/03/2021 12:16

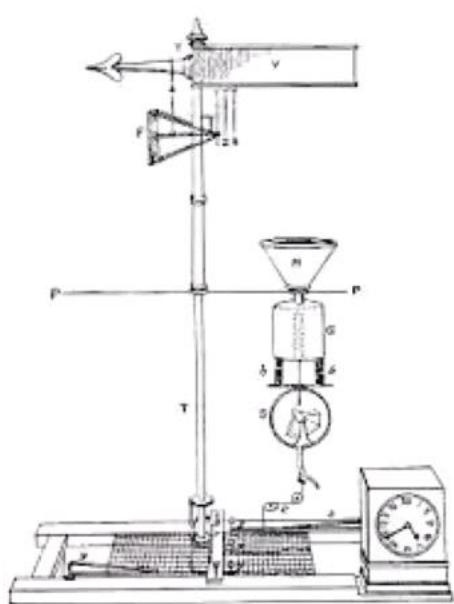
### weather clock (c. wren, 1663)



The design was submitted by Christopher Wren to the Royal Society, 9 December 1663.

Ritaglio schermata acquisito: 02/03/2021 12:16

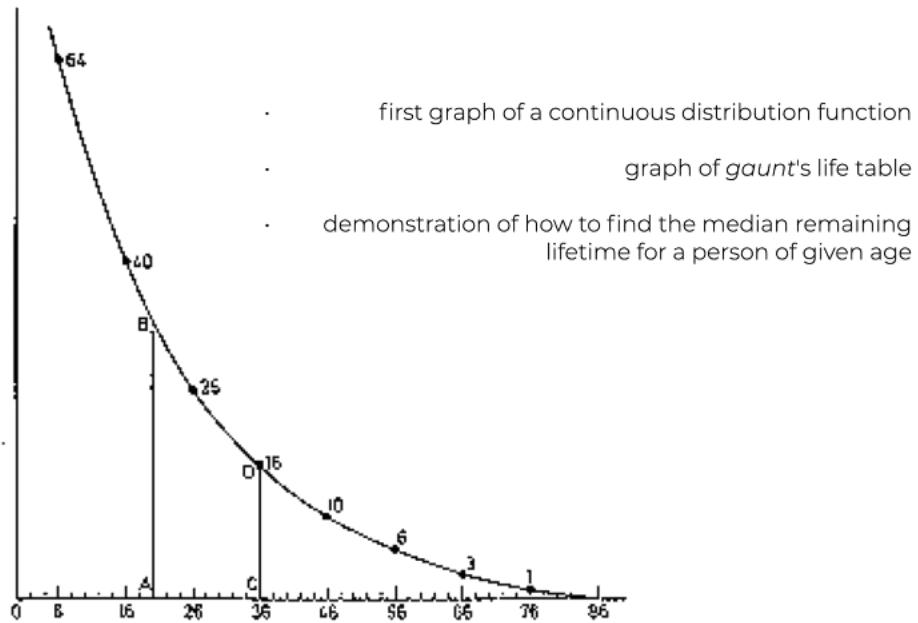
The weather clock. For the first time in history we have a moving graph.



We have a full representation of the tool that describe what was going on.  
Everything was done in polar coordinated instead of longitudine and latitudine.

Ritaglio schermata acquisito: 02/03/2021 12:17

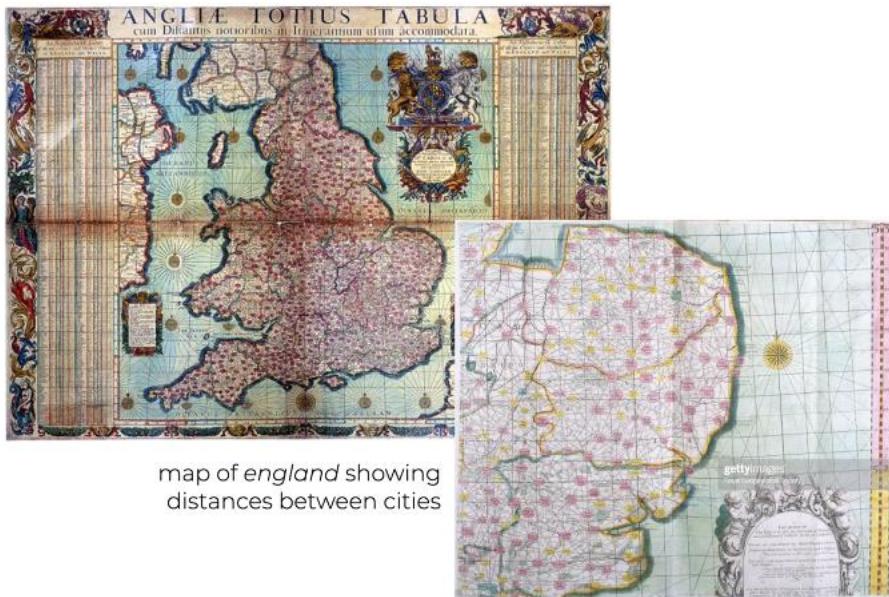
life table (c. huygens, 1669)



Ritaglio schermata acquisito: 02/03/2021 12:18

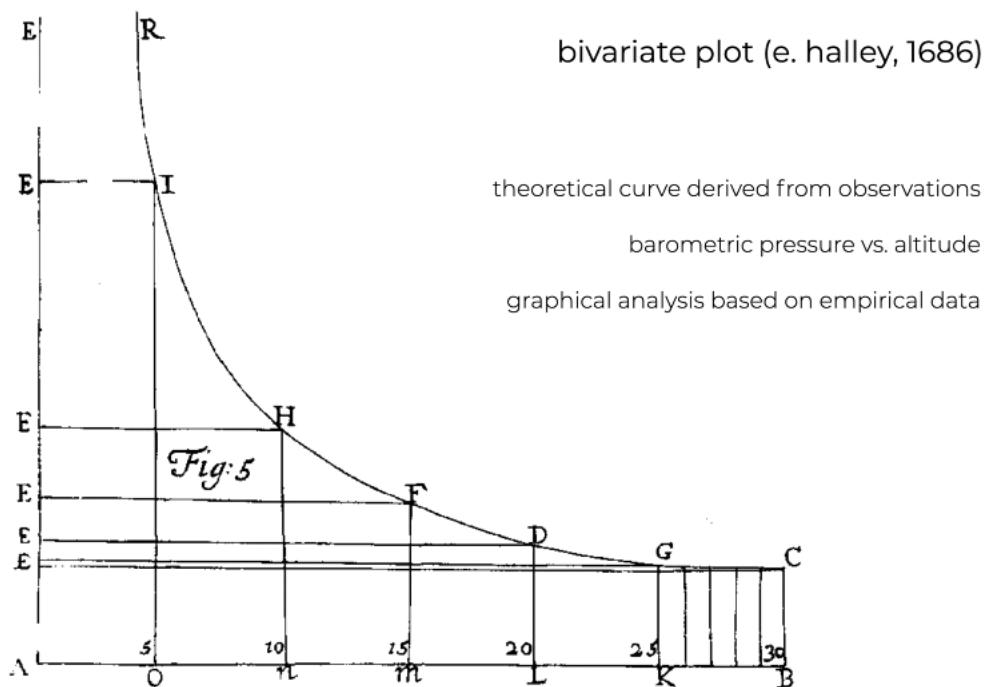
Curve with different reference points. We have a statistical relation. For the first time the representation of curve (continuous distribution function) is believed to be more representative than a table.

network diagram on a map (j. adams, 1679)



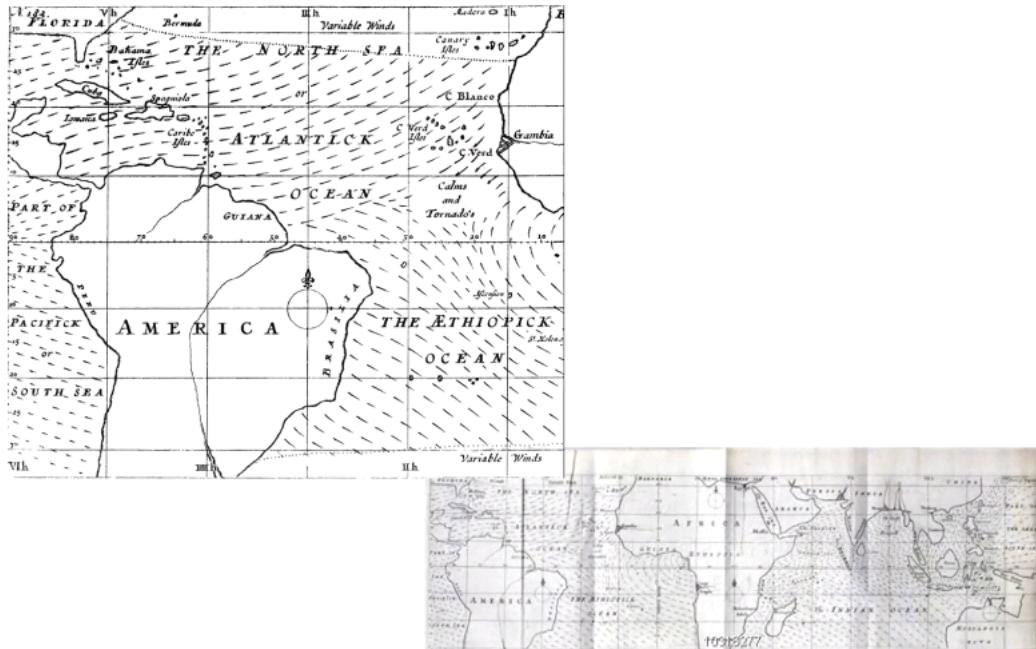
Ritaglio schermata acquisito: 02/03/2021 12:20

In the same period we have for the first time in cartography, we have an additional layer of information represented as a network. A network of towns connected with lines annotated with numbers (distances). Cartographic information with additional information.



Ritaglio schermata acquisito: 02/03/2021 12:22

first weather map (e. halley, 1686)

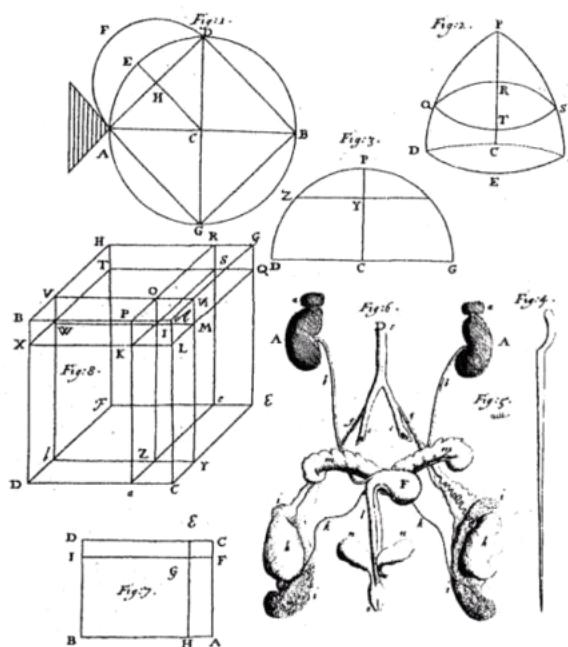


Ritaglio schermata acquisito: 02/03/2021 12:23

Halley had the idea of plotting the current flows of the ocean between the ocean: we have a wheather map. We have fields of forces, we have a complitly new kind of data.

*Philos. Transact. N°. 196.*

probs (e. halley, 1686)



first use of areas of rectangles to display probabilities of independent binary events

Ritaglio schermata acquisito: 02/03/2021 12:25

Halley used also many new kind of portraits for data visualization such as the use of rectangles for the probabilities of indipendent binary events.

With halley we finish the 1600

The 1700 is more modern era. This century was characterized ny:

- Initial germination of the seeds of visualization
- More than just geographical position(isolines and contours)

- Thematic mapping of physical quantities, geologic economic and medical data.
- Early beginnings of statistical theory ( measurement error)
- Novel visual forms
- Reproduction of data images (color printing, lithography)

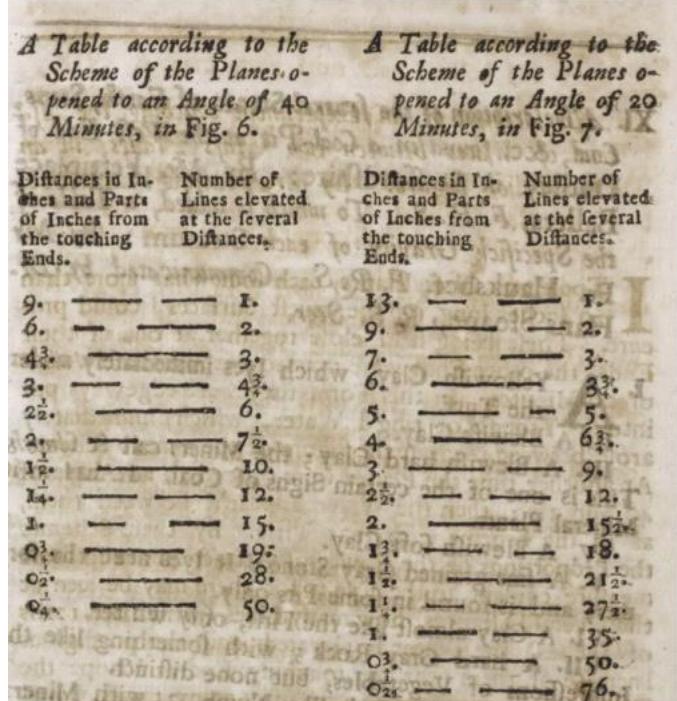
first contour map (e. halley, 1701)



Ritaglio schermata acquisito: 02/03/2021 12:29

- contour map showing curves of equal value
- isogonic map; lines of equal magnetic declination for the world
- possibly the first contour map of a data-based variable

literal line graph (f. hauksbee, 1712)



X. An Account of an Experiment touching the Ascent of Water between two Glass Plates, in an Hyperbolick Figure. By Mr. Francis Hauksbee, F. R. S.

- literal line graph
- inspired by observation of nature
- section of hyperbola
- capillary action of colored water between two glass plates

Ritaglio schermata acquisito: 02/03/2021 12:30

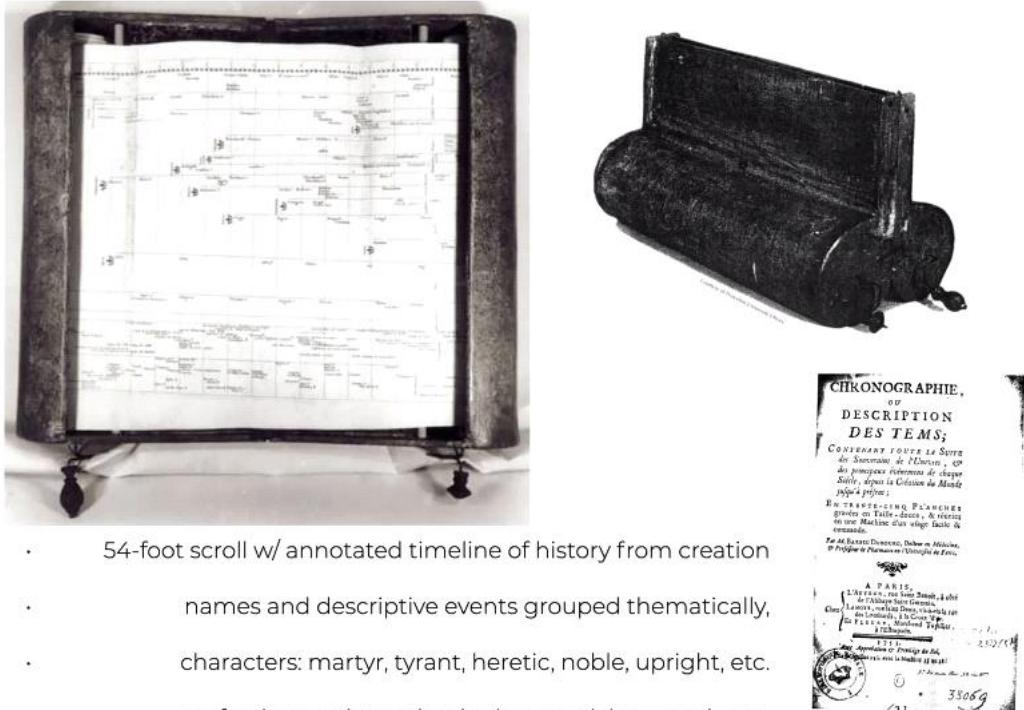
first modern contour map (p. buache, 1752)



Ritaglio schermata acquisito: 02/03/2021 12:32

Here we have a level of precision quite high. This is a reliable map.

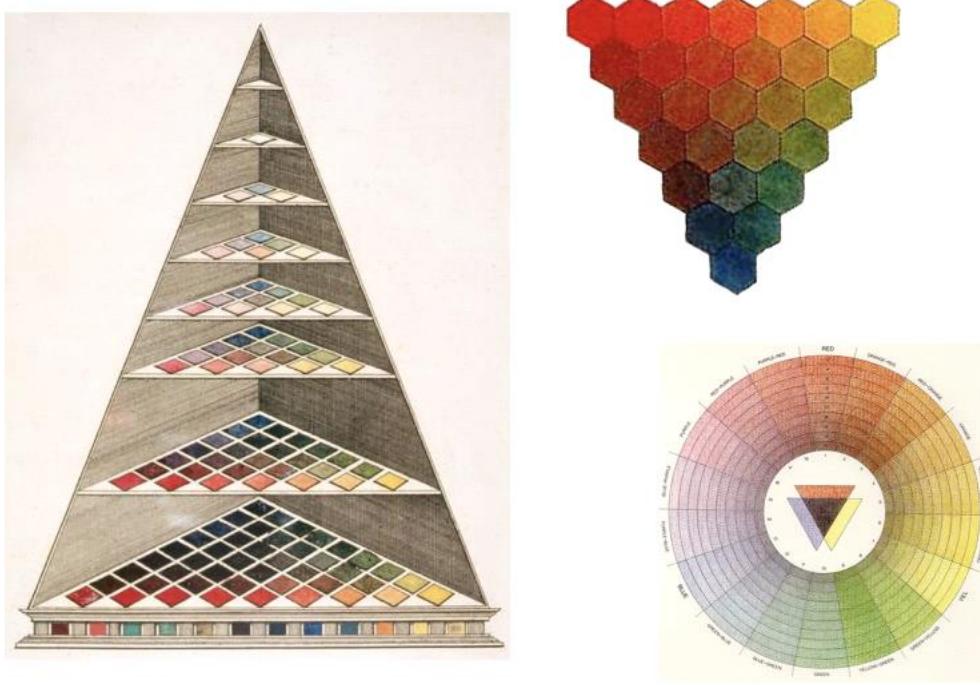
carte chronologique (j. barbeu-dubourg, 1753)



- 54-foot scroll w/ annotated timeline of history from creation
- names and descriptive events grouped thematically,
- characters: martyr, tyrant, heretic, noble, upright, etc.
- professions: painter, theologian, musician, monk, etc.

Ritaglio schermata acquisito: 02/03/2021 12:33

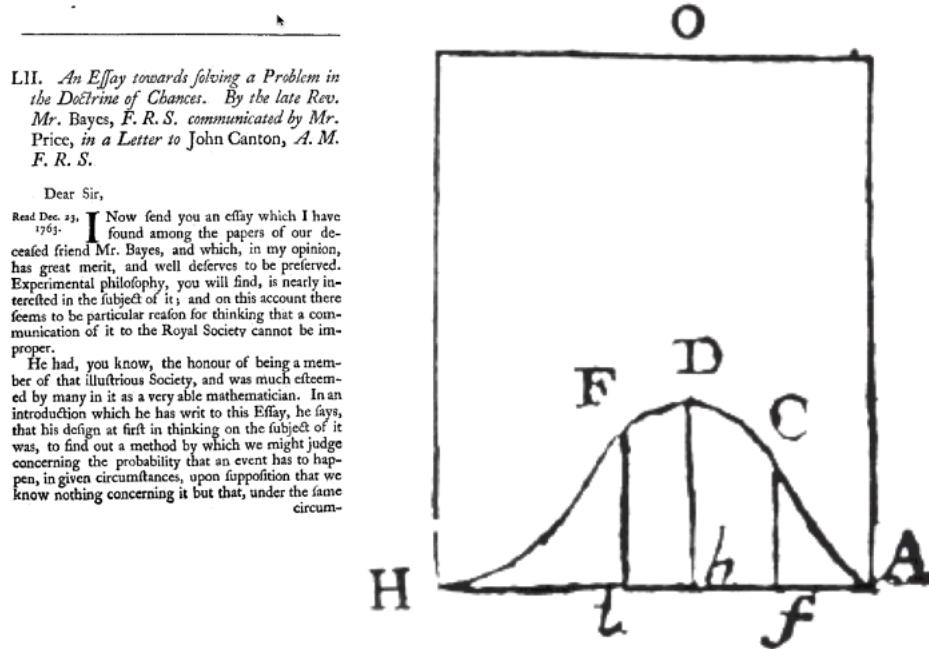
## color system (m. harris/j.h. lambert/j.t. mayer, 1758-1772)



Ritaglio schermata acquisito: 02/03/2021 12:35

Systematic theory of colours. Theory of colour become rounded. This is a representation of how to structure color space, pyramidal form, triangular form and circular form. This is also a practical tool for the data visualizer to pick colour for his representation.

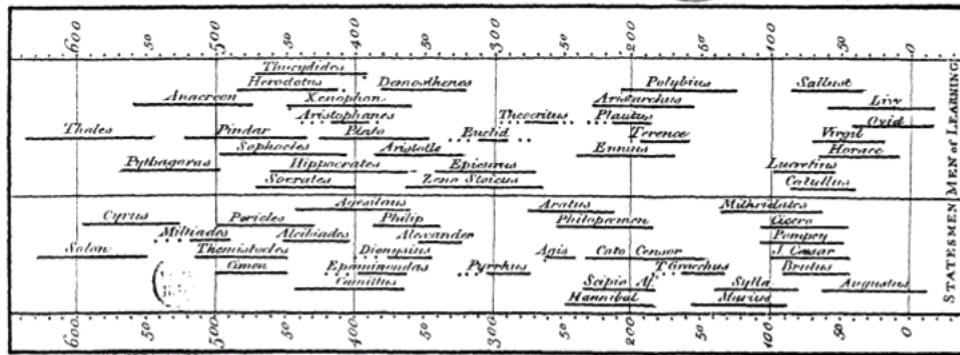
## beta density graph (t. bayes, 1763)



Ritaglio schermata acquisito: 02/03/2021 12:37

The founder of modern statistics, we have a full plotting of the density distribution. For the first time the full distribution is portrayed.

# *A Specimen of a Chart of Biography.*



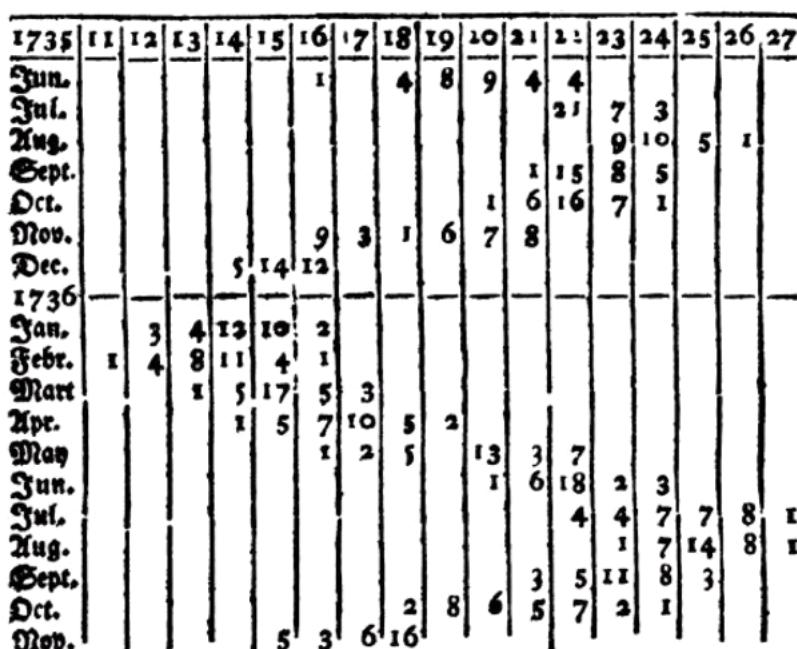
life spans of 2,000 famous people, 1200 b.c. to 1750 a.d.

quantitative comparison by means of bars

Ritaglio schermata acquisito: 02/03/2021 12:38

We have the annotation of the name and also a quantitative design line plot. This gives the possibility to the reader to visually compare the life span of different characters through the years.

→ first semi-graphic display (j.h. lambert, 1779)



combining tabular and graphical formats

Ritaglio schermata acquisito: 02/03/2021 12:40

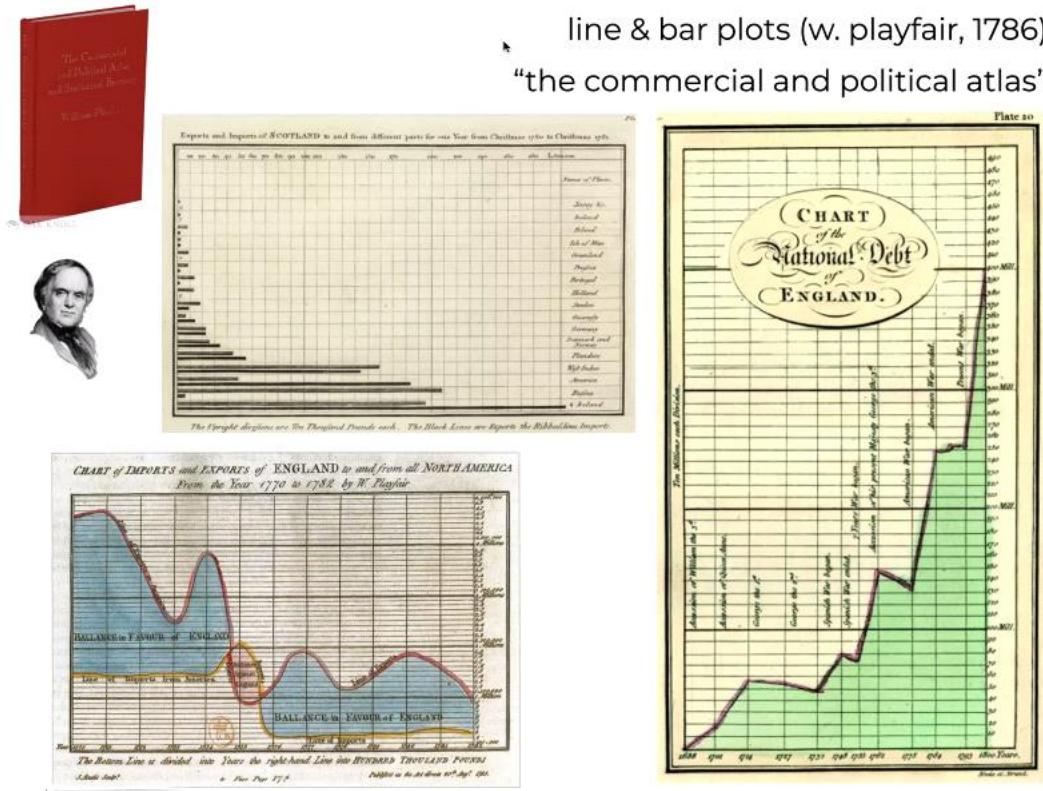
It is a sort of calendar combining the tabular data and the graphical forms using the columns.

first thematic & comparison maps (a.f.w. crome, 1782/1786)



Ritaglio schermata acquisito: 02/03/2021 12:41

A new figure emerges, the combination of maps with statistics or demography.

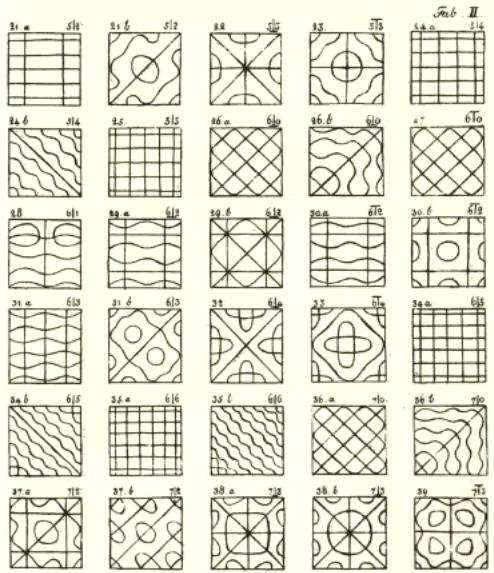


Ritaglio schermata acquisito: 02/03/2021 12:42

This is the father of modern data visualization. We have information about the economy of the places e.g import and export flows.

We have all the information that characterized the modern graphs such as color, grid, annotations, title. This all all the characteristic to rate a plot.

## vibrational patterns (e.f.f. chiadni, 1787)

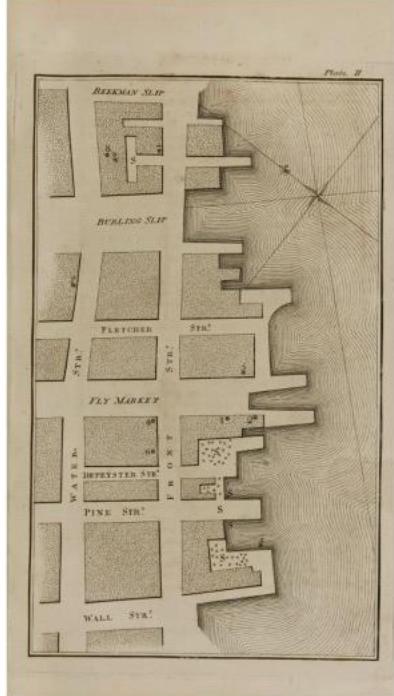


visualization of vibration patterns by spreading a uniform layer of sand on a disk, and observing displacement when vibration is applied

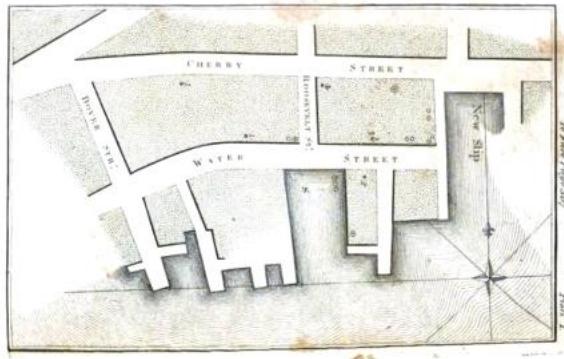
Ritaglio schermata acquisito: 02/03/2021 12:45

Also other subjects started to use data visualization.

## epidemiological map (v. seaman, 1797)



mapping yellow fever in nyc



Ritaglio schermata acquisito: 02/03/2021 12:46

Displaying with points on a map the patients that where affectyed by yellow fever. This is a first map reporting the map of a disease.

1800

# Use of R

mercoledì 3 marzo 2021 11:31

## Array

An array is multiple subscripted collection of data entries.

# History of data visualization part 2

martedì 9 marzo 2021 11:30

1800

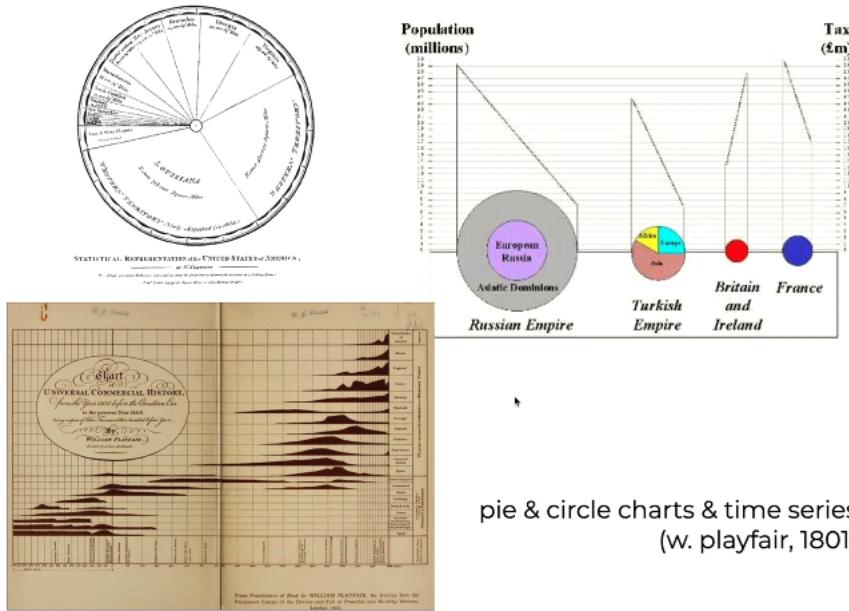
- "age of enthusiasm" in graphics and thematic cartography
- explosive growth in statistical graphics and thematic mapping
- bar & pie charts, histograms, line graphs, time-series, contour plots
- from single maps to comprehensive atlases,
- depicting data on a wide variety of topics (economic, social, moral, medical, physical, etc.)
- a wide range of novel forms of symbolism

Ritaglio schermata acquisito: 09/03/2021 11:31

A number of elements of element of data visualization are now daily basis.

We move from single maps to comprehensive atlases.

Different fields approach to data viz.



pie & circle charts & time series  
(w. playfair, 1801)

Ritaglio schermata acquisito: 09/03/2021 11:33

Modern version of pie chart and even the time series chart.



first uk geological map (w. smith, 1805)



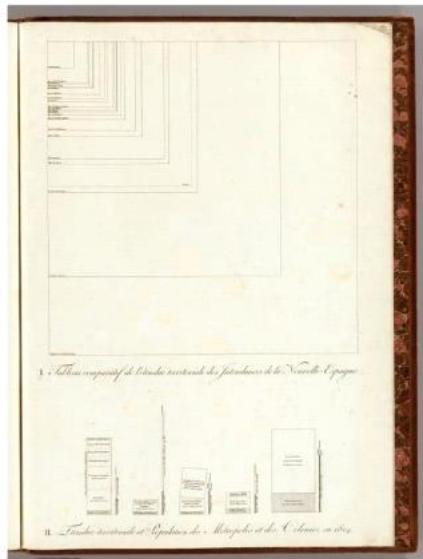
"the map that changed the world" [winchester,2001]



Ritaglio schermata acquisito: 09/03/2021 11:34

Many different sciences were taking the benefits of the innovation in data visualization such as geology. For the first time we have a map with orography (structure of the mountain) we can see from the examples the use of colors and isoline. In the bottom graph we also see differnt color to indicate the material of the montains.

subdivided bar graph (a. von humboldt, 1811)

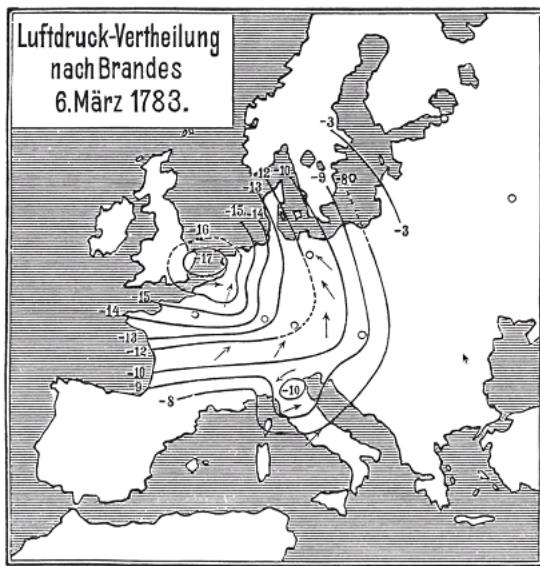


subdivided bar graphs & superimposed squares, showing the relative size of mexican territories and populations in the colonies

Ritaglio schermata acquisito: 09/03/2021 11:36

New ways of depicting rather common concepts -> subdivided and superimposed bar graph. New way of presenting data (new borns in colonies).

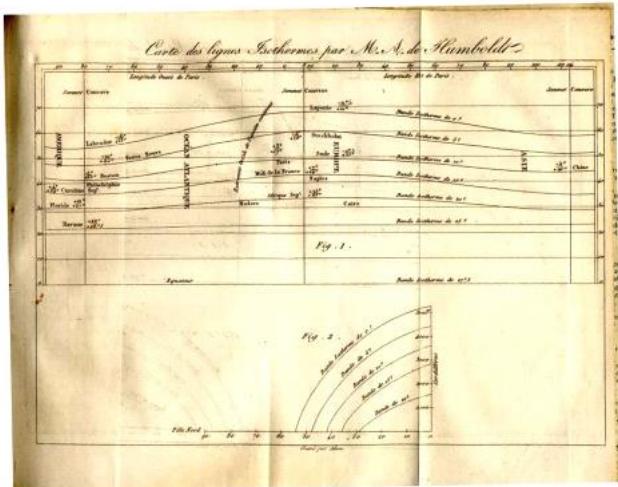
first weather map (h.w. brandes, 1816)



Ritaglio schermata acquisito: 09/03/2021 11:37

Here we have an example of a weather map with isolines as we are used to see them now. Metherology and clime sciences benefits also from differnt kind of plots.

first graph of isotherms (a. von humboldt, 1817)

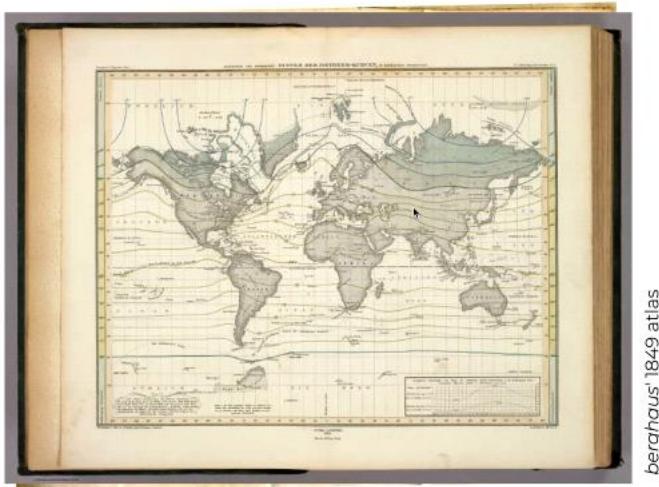


mean temperature around the world by latitude and longitude.

Ritaglio schermata acquisito: 09/03/2021 11:38

Plot of mean temperature arounf the word according to a global cordinate systems. We can appreciate the different temperature by time and space.

first graph of isotherms (a. von humboldt, 1817)



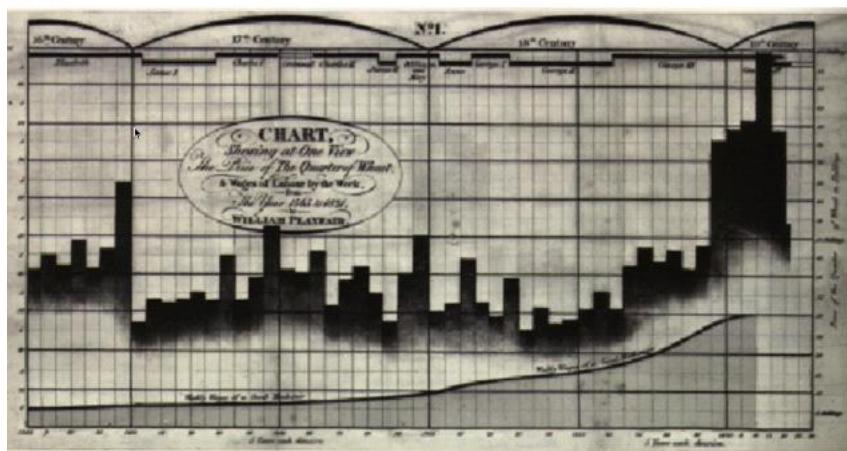
berghaus 1849 atlas

mean temperature around the world by latitude and longitude.

Ritaglio schermata acquisito: 09/03/2021 11:39

Here we have the same graph as before plotted together with a global map.

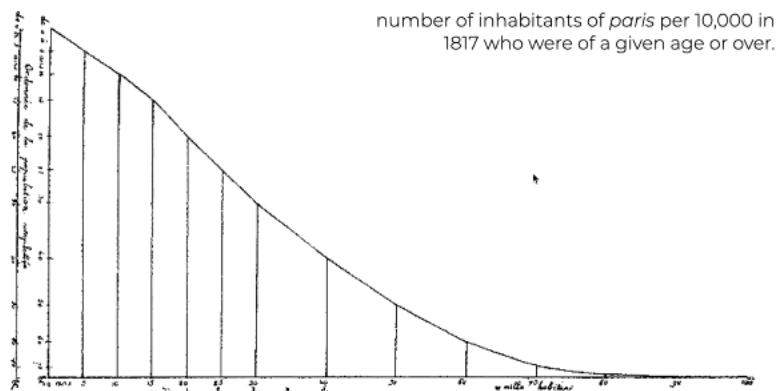
multiple time series (w. playfair, 1821)



time series graph of prices, wages, and ruling monarch over a 250 year period

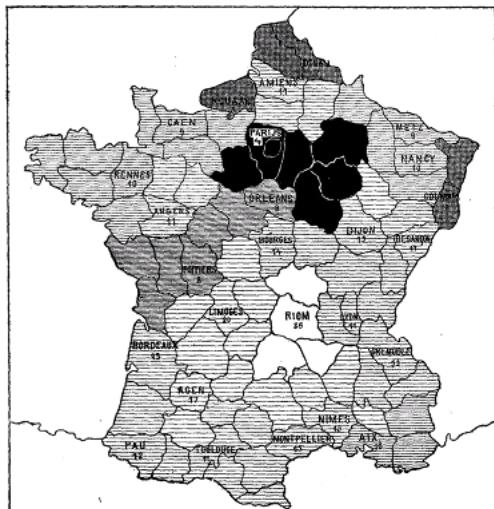
Ritaglio schermata acquisito: 09/03/2021 11:40

first cumulative frequency curve (j.b.j. fourier, 1821)



There is a blooming of statistic and demographic. This is the first cumulative curve. This is a way of plotting an age pyramid.

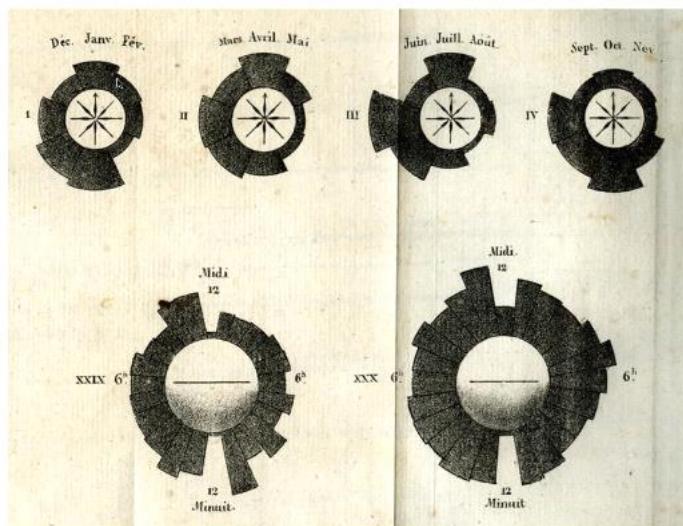
first chloroplet map (p.c. dupin, 1826)



- shadings from black to white
- distribution and intensity of illiteracy in france
- first modern statistical map

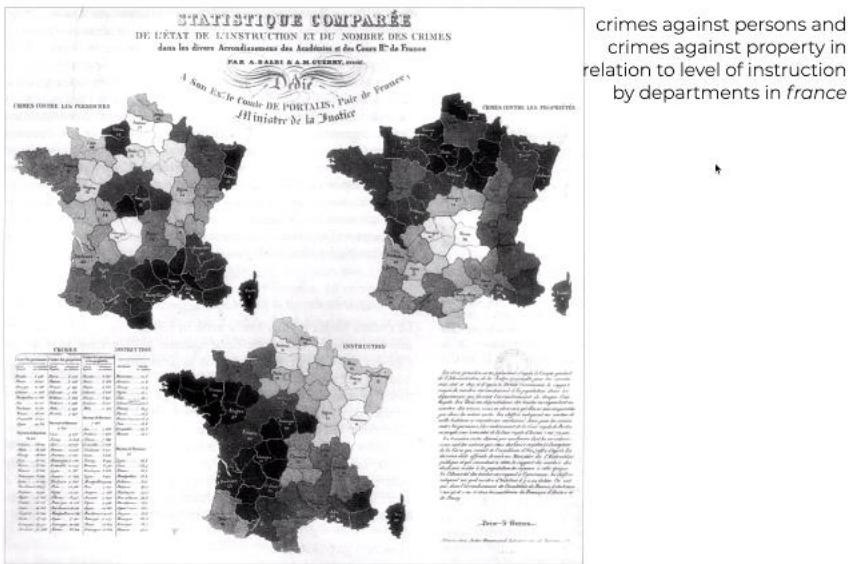
This is a real milestone. First cholroplet map -> different layer represented with different revolution. The map shows the intensity of illiteracy. France is divided in regions, they are charachterized by color and name. This influenced a lot the way through which statistics and data representation interact between each other.

first polar area chart (a.m. guerry, 1829)



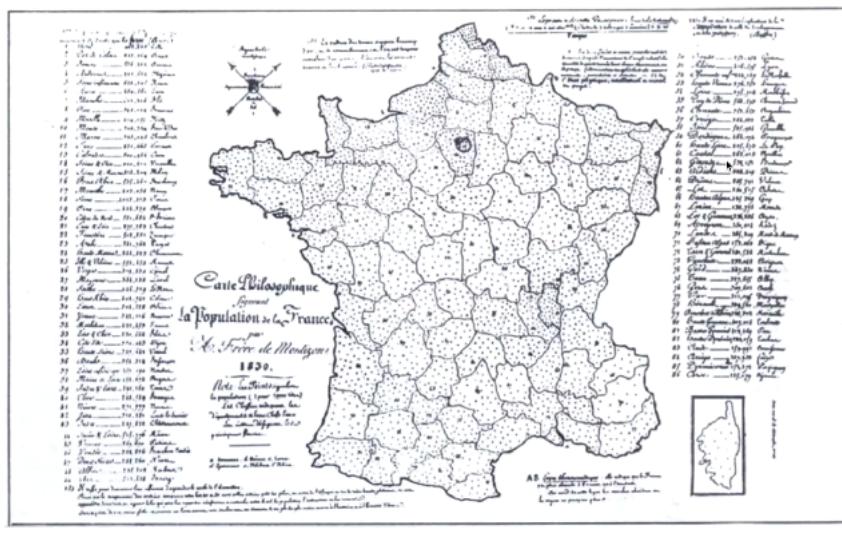
daily phenomena: direction of the wind in 8 sectors, births and deaths by hour of the day.

first comparative chloroplet maps (a.m. guerry & a. balbi, 1829)



Ritaglio schermata acquisito: 09/03/2021 11:46

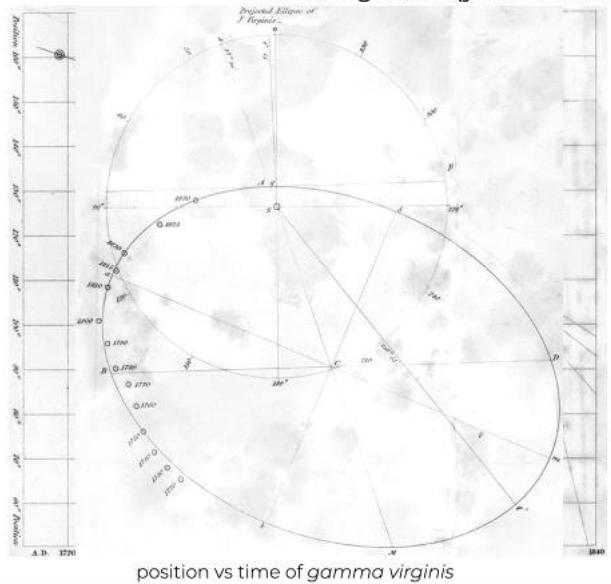
first dot map (a.j.f. de montizon, 1830)



Ritaglio schermata acquisito: 09/03/2021 11:46

Data visualizers understood that they can represent village by village according to the density of the dots.

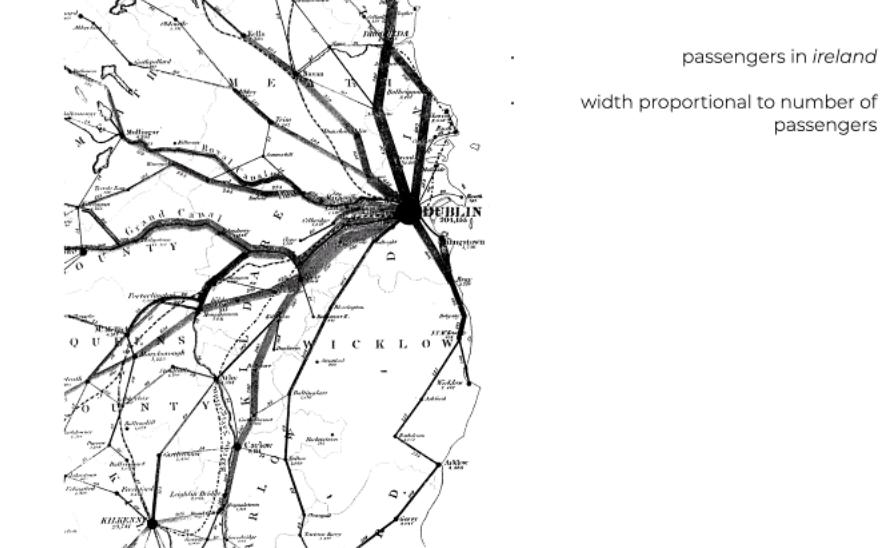
first fitting curve (j.f.w. herschel, 1832)



Ritaglio schermata acquisito: 09/03/2021 11:48

Using data viz he was able to plot the root of the star.

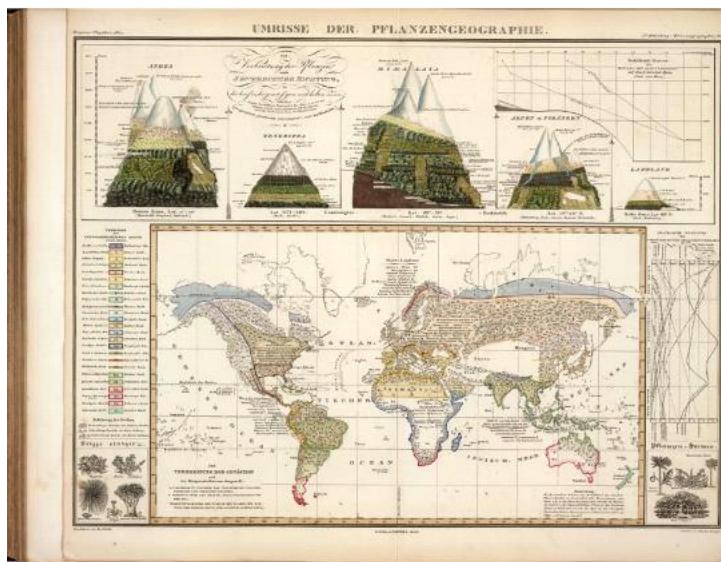
flow map (h.d. harness, 1837)



Ritaglio schermata acquisito: 09/03/2021 11:49

Flow map -> network of rows and the dimension, color, size of the line connecting places indicates a new statistics: the number of people who cross that street.

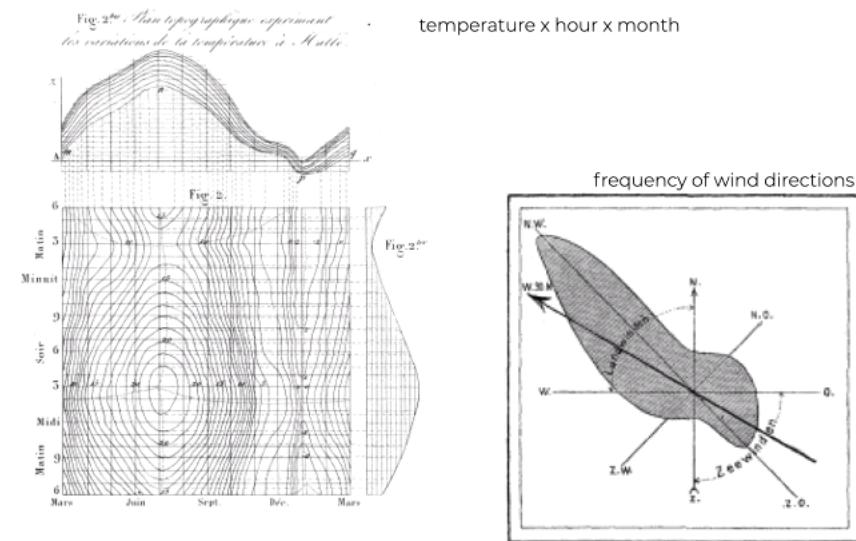
This is a kind of river flow.



Ritaglio schermata acquisito: 09/03/2021 11:50

Plants scientists benefits as well from data viz. In this map it is shown the distribution of plants around the world. This is a complex plot representing the phenomenon of distribution of plants around the world.

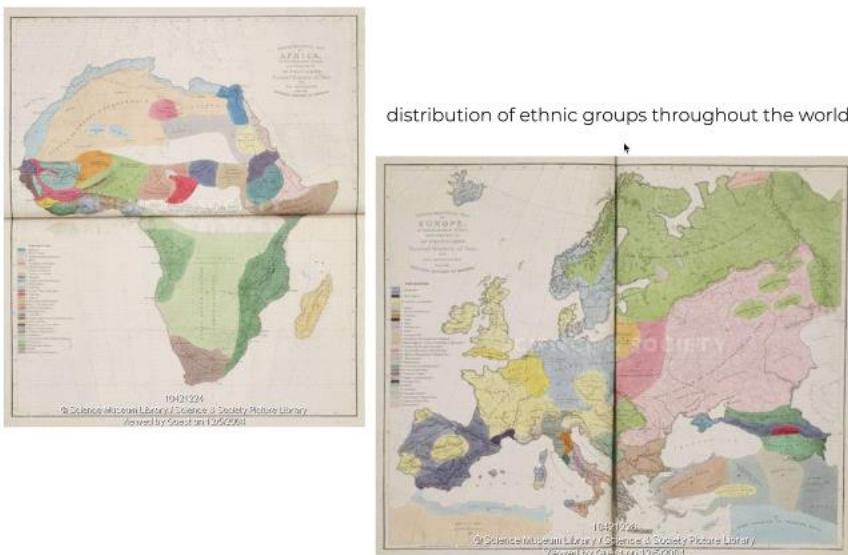
first 3d contour map & polar coordinates (l. lalanne, 1843)



Ritaglio schermata acquisito: 09/03/2021 11:51

First 3-D countour map.

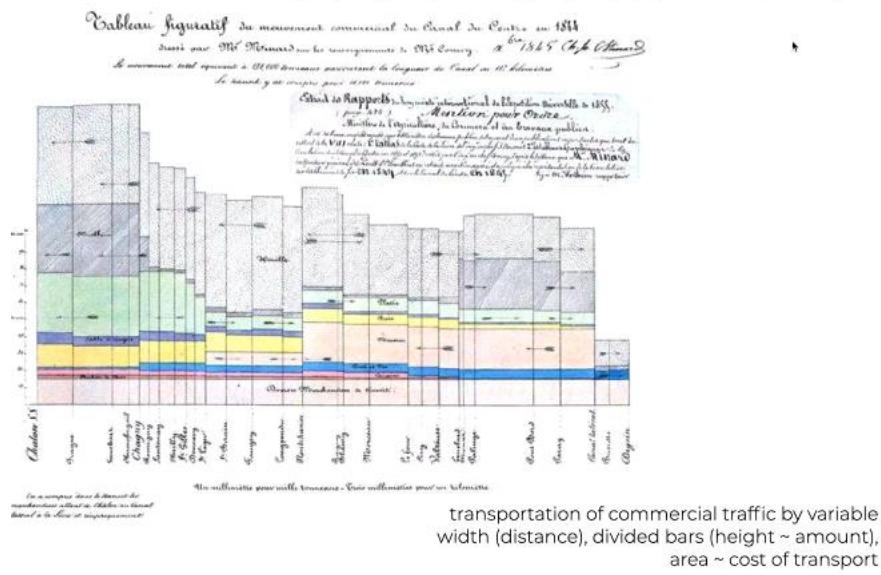
## first ethnographic maps (a.k. johnstone & j.c. pritchard, 1843)



Ritaglio schermata acquisito: 09/03/2021 11:53

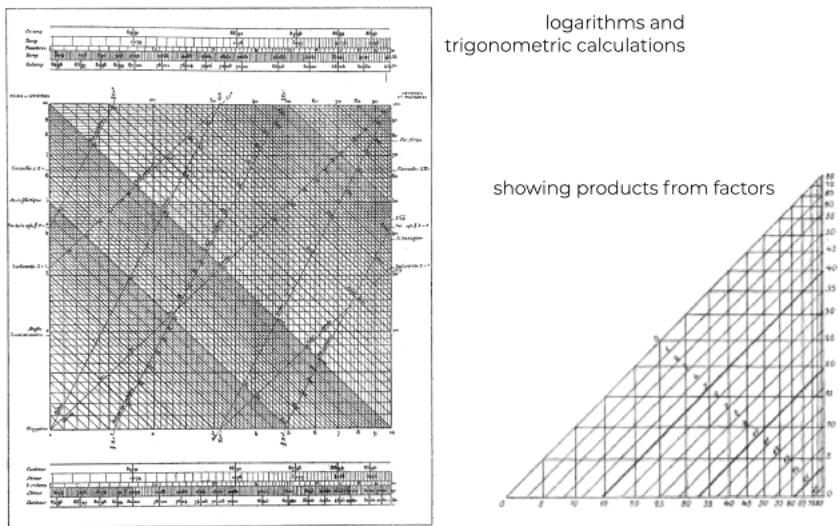
This last two graph are a novelty we have many different categorical elements -> languages it is a more general concept related to culture.

## tableau-graphique [mosaique plot] (c.j. minard, 1844)



Ritaglio schermata acquisito: 09/03/2021 11:54

first log-log nomogram & universal calculator (I. Lalanne, 1846)

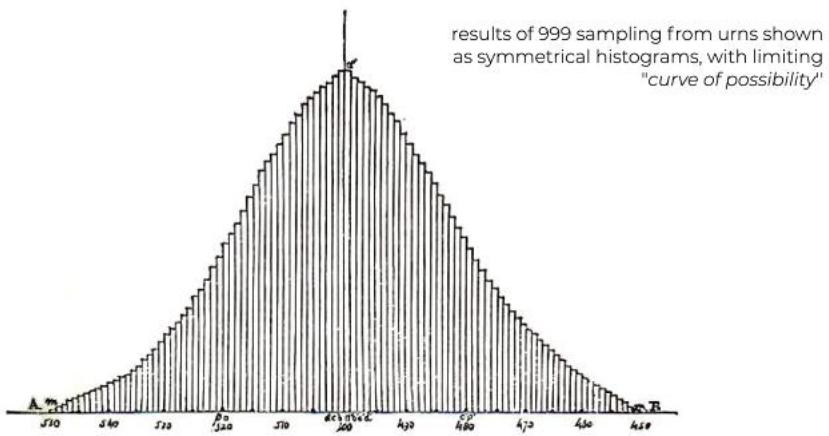


Ritaglio schermata acquisito: 09/03/2021 11:56

This is the first visualization that helped people to understand easily trigonometric calculations.

New elements of mathematics needed to be wrote fro the practitioners in a very smooth way.

first empirical binomial distribution (a. quatelet, 1846)



Ritaglio schermata acquisito: 09/03/2021 11:57

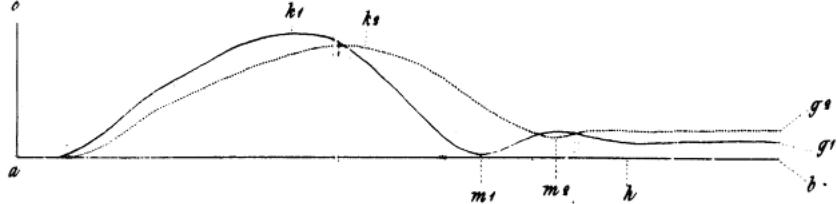
For the first time we have an emporical binomial distribution -> bell curve showing simmetrical histogram. The first time for Guassian distribution to be used for experiment.

## muscle action (h. helmoltz, 1850-1852)



graphical representation of muscle action after stimulation

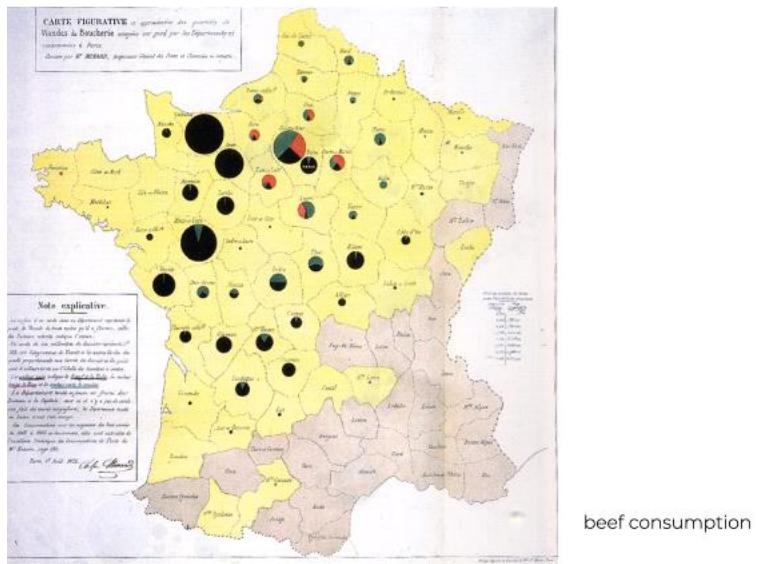
first example of deduction from graph



Ritaglio schermata acquisito: 09/03/2021 11:59

This is another milestone. This is related to anatomy. This graph is new because instead of representing a phenomenon by data, he used the representation to make a rule. The plot/ the data visualization part of the process was the engine to derive an empirical law.

## first map including statistical charts (j.c. minard, 1852)

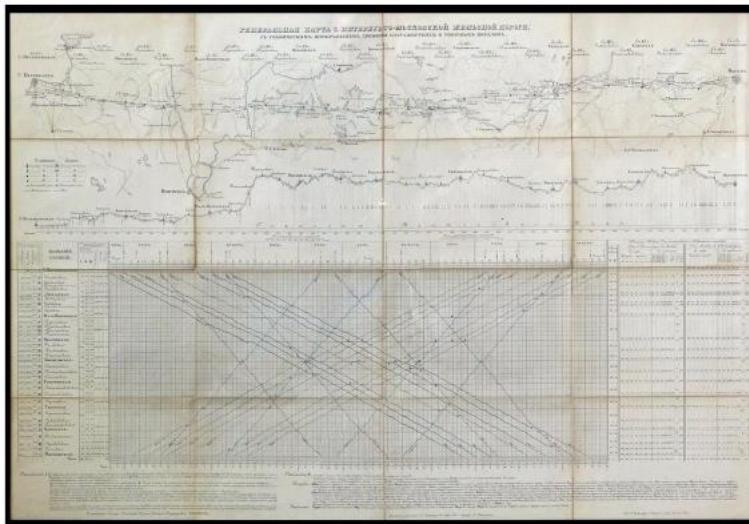


beef consumption

Ritaglio schermata acquisito: 09/03/2021 12:00

Here we have a chloroplet map but also statistical chart on some given department of interest. We have a graph of beef consumption in France. For some department we have data which indicates which type of beef we are dealing with.

### train schedule (serjev, 1854)

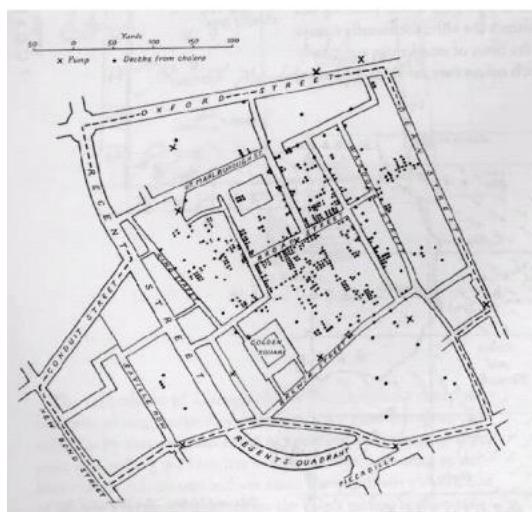


train schedule for 35 railways stations, between st. petersburg and moscow

Ritaglio schermata acquisito: 09/03/2021 12:02

Using data visualization for providing data for the people. Instead of just using a table here we have lines connecting station and showing travel time and other useful information. This panel is quite modern in its philosophy. This is a quite complex map but it is thought to be a good support for travellers.

### true epidemiological success w/ graphics (j. snow, 1855)



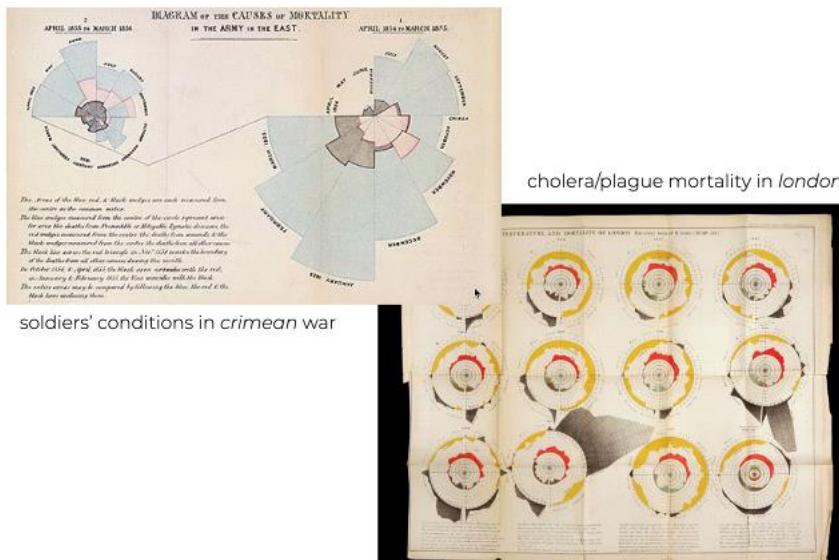
discovering a cholera outbreak in london using a map

Ritaglio schermata acquisito: 09/03/2021 12:04

Now data viz is not just a representation tool but it is also a fundamental tool for understanding the analyzed phenomenon. This is the most common example -> the statistician John Snow who studied the driving reason for the epidemics of cholera.

He put a dot which indicates the place where infected people live. He noted that there was a concentration in a certain part. He understood that there was a correlation with the epidemic and the pump.

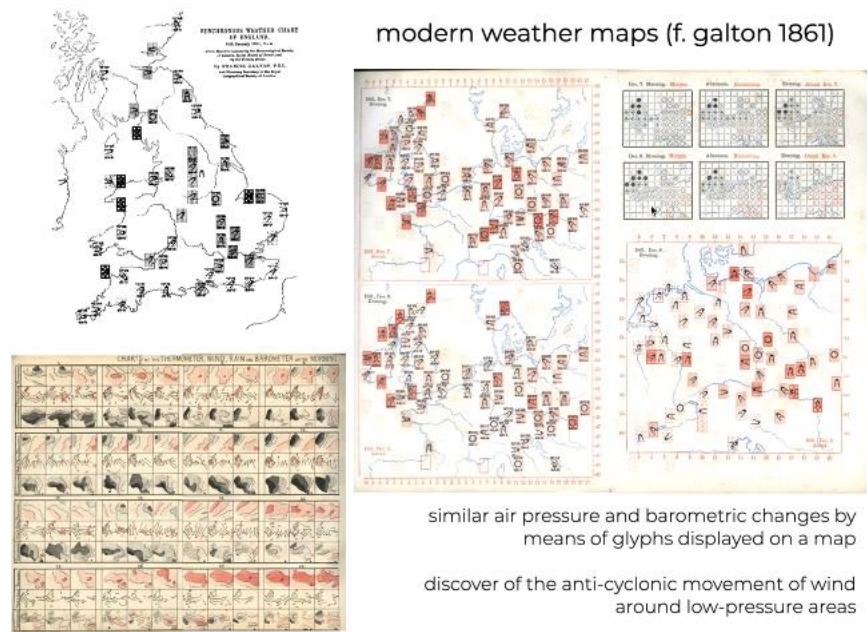
"coxcombs" / polar area charts (f. nightingale, w. farr ~1855)



Ritaglio schermata acquisito: 09/03/2021 12:08

How data viz changed people life. Through a radar chart they represent the status of the soldier in crimian war. This was useful for nursery to take action.

Visualizing is no more something at posteriori.

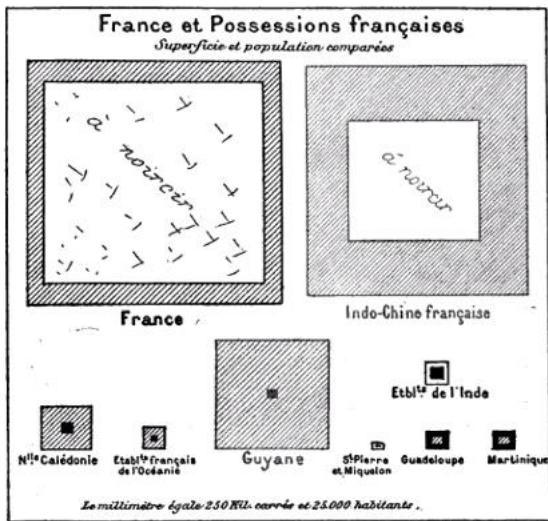


Ritaglio schermata acquisito: 09/03/2021 12:10

Wheather maps started to be similar to the modern ones. This helped to discover anticyclonic movement. This conquer was possible only thanks to data viz.

Data visualizer also started to get attention also for the characterstic of graphs itself e.g. semilog chart. What it is important is the scale

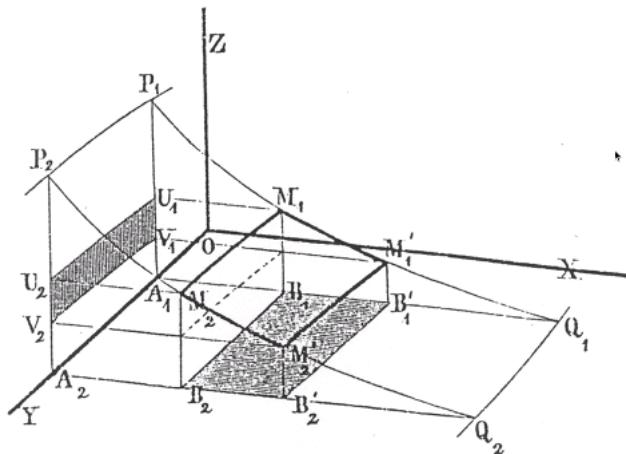
first chart used in textbooks (e. levasseur, 1868)



Ritaglio schermata acquisito: 09/03/2021 12:13

Data viz has now a larger audience. It is used for educational purpose

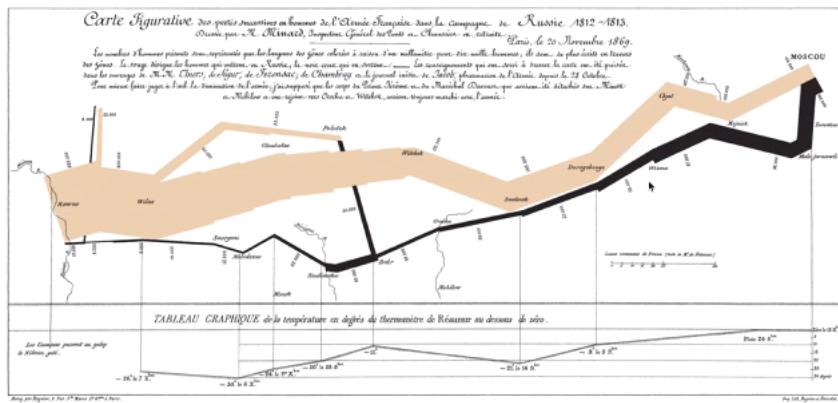
first stereograms (g. zeuner, 1869)



Ritaglio schermata acquisito: 09/03/2021 12:14

This changed the framework of data viz.  
This is a 3-d representation of euclidean space.

## napoleon russian campaign (c. j. minard, 1869)



- the size of his army at specific geographic points during their advance/retire
- the distance traveled
- temperature
- slatitude and longitude
- direction of travel
- location relative to specific dates
- later called sankey diagram

Ritaglio schermata acquisito: 09/03/2021 12:15

Universally acknowledge as the best graph drawn so far.

The important thing here is the number of information in a graphical comprehensive representation and the nicely way through which is represented.

It is represented a specific geographical area, the amount of the army, the distance travelled, the temperature, the long and latitude,... we have 5 levels of information which are very readable. **Sankey diagram**.

## the periodic table (d. mendeleev, 1869)

Ti = 50	Zr = 90	? = 180
V = 51	Nb = 94	Ta = 182
Cr = 52	Mo = 96	W = 186
Mn = 55	Rh = 104,4	Pt = 197,4
Fe = 56	Ru = 104,4	Ir = 198
Ni = 59	Co = 59	Pd = 106,6
Cu = 63,4	Ag = 108	Os = 199
H = 1		Hg = 200
Be = 9,4	Mg = 24	Zn = 65,2
B = 11	Al = 27,4	? = 68
C = 12	Si = 28	? = 70
N = 14	P = 31	As = 75
O = 16	S = 32	Se = 79,4
F = 19	Cl = 35,5	Br = 80
Li = 7 Na = 23	K = 39	Rb = 85,4
	Ca = 40	Sr = 87,6
	? = 45	Ce = 92
	?Er = 56	La = 94
	?Yt = 60	Di = 95
	?In = 75,6	Th = 118?

Ritaglio schermata acquisito: 09/03/2021 12:18

Chemestry and data viz -> periodic table -> table with specific locations for the elements -> semi graphical construction.

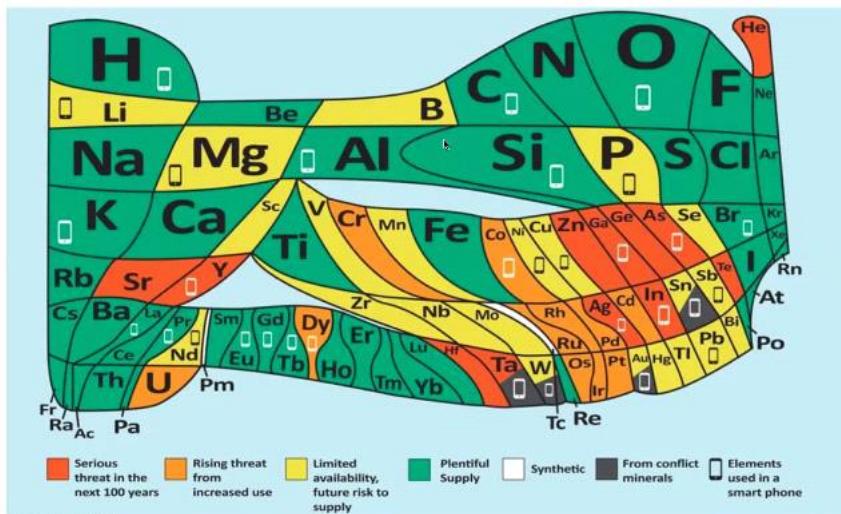
the periodic table (d. mendeleev, 1869)

Periodic Table of the Elements																	
1 H	2 He	3 Li	4 Be	5 B	6 C	7 N	8 O	9 F	10 Ne	11 Na	12 Mg	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Ag	47 Cd	48 In	49 Sn	50 Sb	51 Te	52 I	53 Xe	54 Cs
55 Cs	56 Ba	57-71	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89-103	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
57 La 58 Ce 59 Pr 60 Nd 61 Pm 62 Sm 63 Eu 64 Gd 65 Tb 66 Dy 67 Ho 68 Er 69 Tm 70 Yb 71 Lu  89 Ac 90 Th 91 Pa 92 U 93 Np 94 Pu 95 Am 96 Cm 97 Bk 98 Cf 99 Es 100 Fm 101 Md 102 No 103 Lr																	
Acid Water  Alkaline Earth  Transition Metal  Rare Metal  Divalent Metal  Halogen  Noble Gas  Lanthanide  Actinide																	

Ritaglio schermata acquisito: 09/03/2021 12:20

Here we have color, position and annotations.

the periodic table (d. mendeleev, 1869)

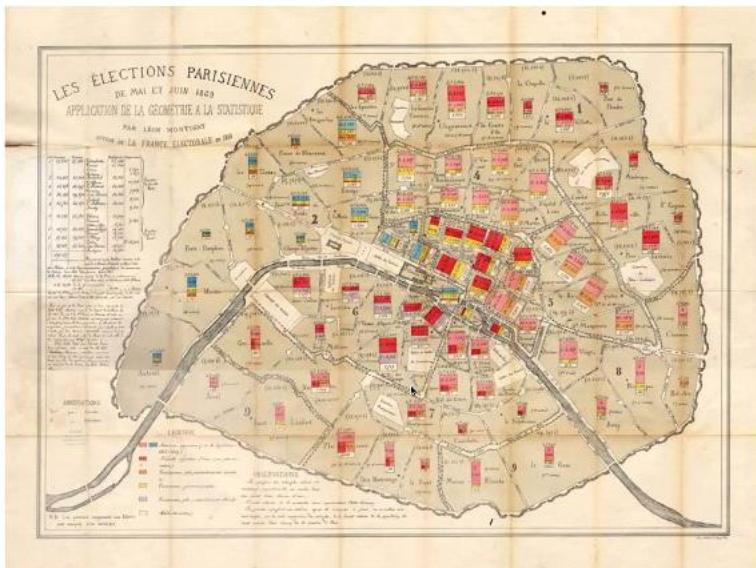


even more enhanced - 2019 int'l year of the periodic table

Ritaglio schermata acquisito: 09/03/2021 12:20

Variation of these theme are very large. Here we have a strange periodic table where elements have different shape and their size indicates the abundance of these elements on earth.

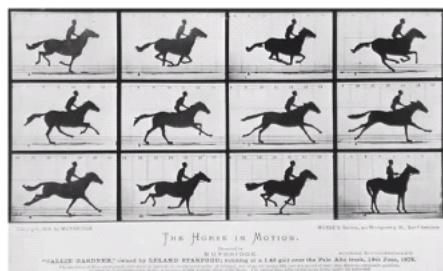
paris election map (l. montigny, 1870)



Ritaglio schermata acquisito: 09/03/2021 12:21

Paris election map -> with all the features connected to voting.

galloping horse (e. muybridge, 1872)



recording of motion (of a running horse)  
by means of a set of glass-plate cameras,  
triggered by strings

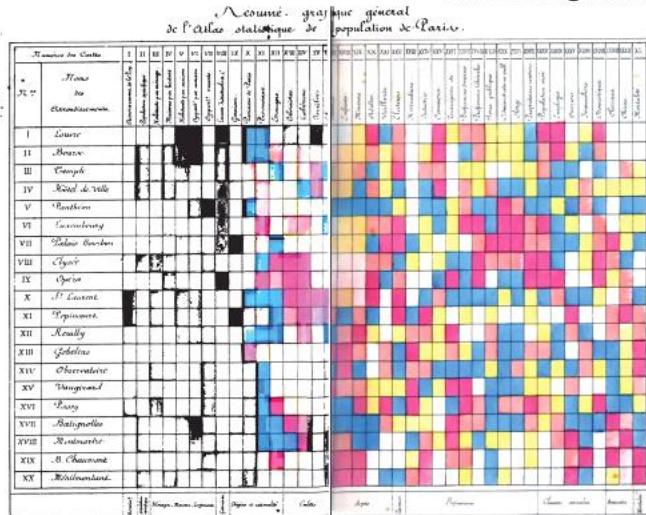


In 1882, e.-j. marray carried out several motion-picture camera experiments, recording a series of photographs to study flight of birds, running and walking

Ritaglio schermata acquisito: 09/03/2021 12:23

People started wondering if representation could also be non-static  
Here we have animation, dynamics.  
For the first time we have the first record of a moving horse.

### first scalograms (t. loua, 1873)

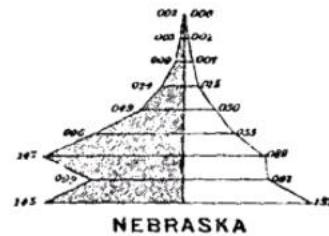
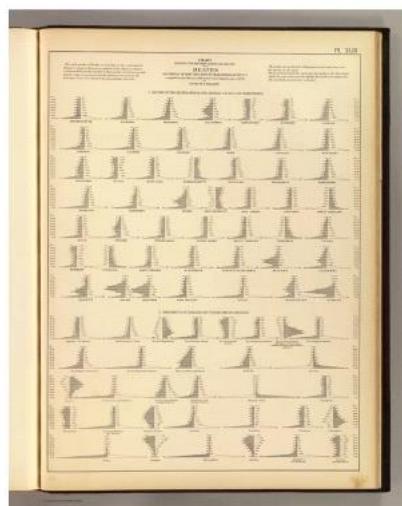


graphic summary of 40 maps of paris, each showing some feature of the population by arrondissement.

Ritaglio schermata acquisito: 09/03/2021 12:24

New type of plotting appears. We have a two scale not sorted, quite independent and a colored element marking the relation between two particoular sqaure.

### first age pyramid (f.a. walker, 1874)



- age pyramid (bilateral histogram)
- bilateral frequency polygon
- use of subdivided squares to show the division of population
- first true u.s. national statistical atlas

Ritaglio schermata acquisito: 09/03/2021 12:25

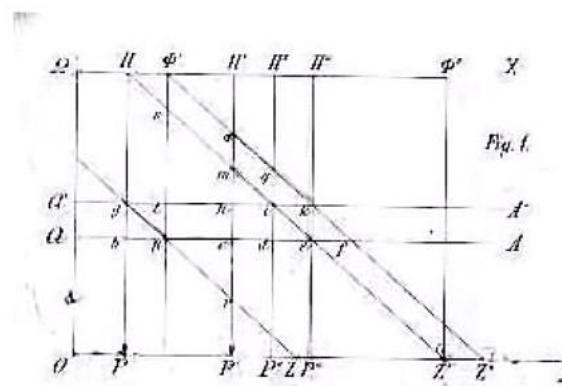
For the first time we met pyramids -> different layers.

first statistical contour map (I.-I. vautier, 1874)



Ritaglio schermata acquisito: 09/03/2021 12:27

first lexis diagram (w. lexis, 1875)

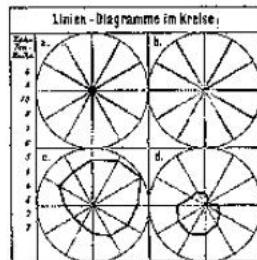
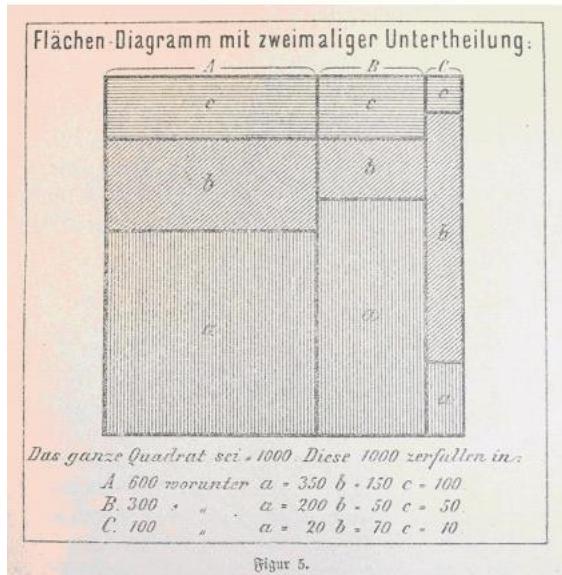


relations among age, calendar time, and life spans of individuals simultaneously

Ritaglio schermata acquisito: 09/03/2021 12:27

This is a quite complex 3D graph. It is difficult to understand this graph.

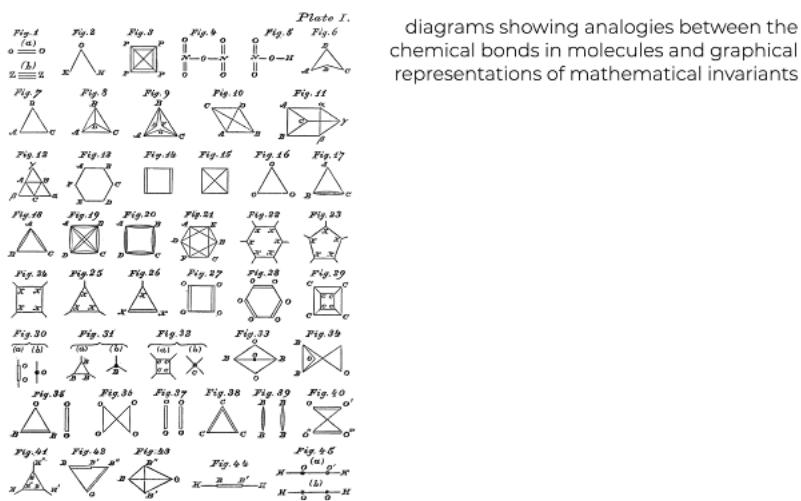
### first mosaic and star plot (g. von mayr, 1877)



Ritaglio schermata acquisito: 09/03/2021 12:28

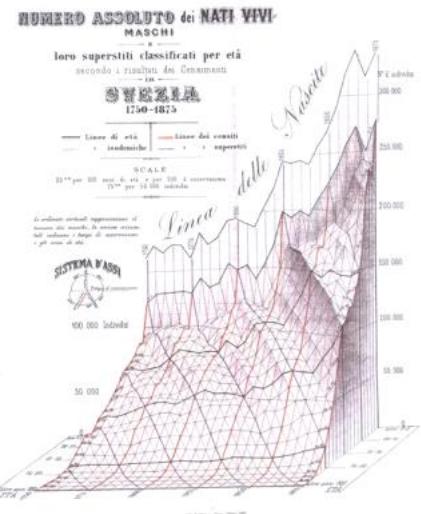
More and more types of representations appeared.

### first use of word "graph" (j.j. sylvester, 1878)

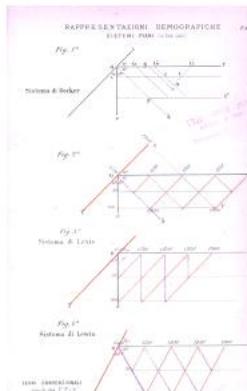


Ritaglio schermata acquisito: 09/03/2021 12:29

## more refined stereogram (I. perozzo, 1879)

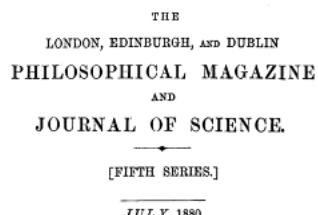


3d representations of data showing the age group of the swedish population between the 18th and 19th centuries.



representation instructions

Ritaglio schermata acquisito: 09/03/2021 12:29

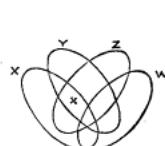


I. On the Diagrammatic and Mechanical Representation of Propositions and Reasonings. By J. VENN, M.A., Fellow and Lecturer in Moral Science, Caius College, Cambridge\*.

## venn diagrams (j. venn, 1880)

For instance, the proposition "All X is Y" needs both the diagrams,  $\bigcap_{x \in X} Y$  ; for we cannot tell, from the mere

verbal statement, whether there are any Y's which are not X. Similarly the proposition "Some X is not Z" needs three other diagrams,



giving any X that is neither Y nor Z—thus,

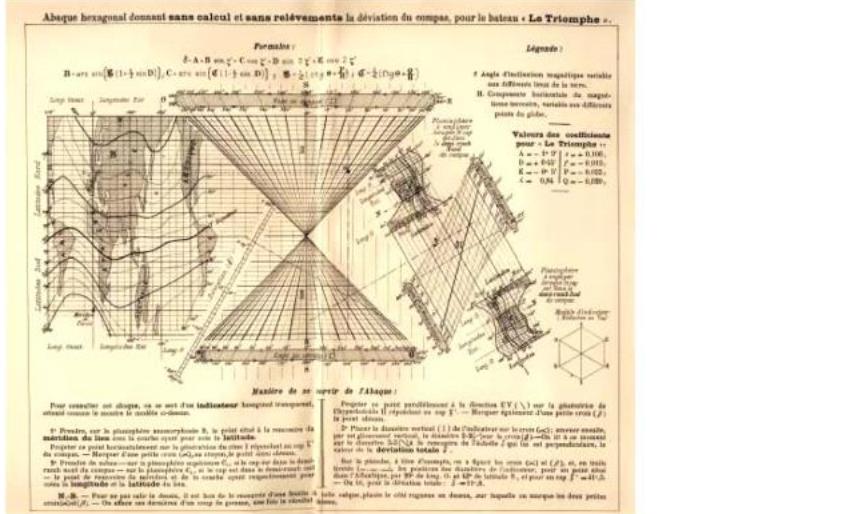


If to this were added the statement that "none but the X's are either Y or Z," we should then abolish the XY and the XZ, and have  $\bigcap_{x \in X} Y \cup \bigcap_{x \in X} Z$ . Scratch out, again, the XYZ compart-

Ritaglio schermata acquisito: 09/03/2021 12:31

Venn diagrams appears representing intersections of sets.

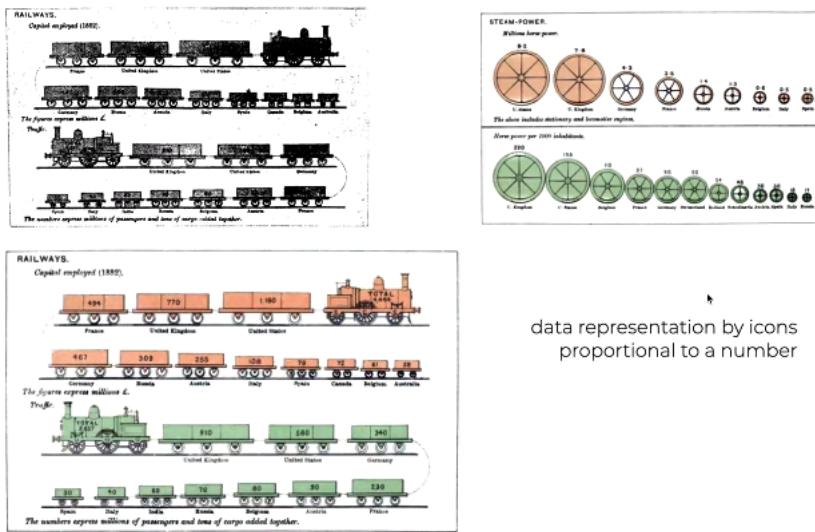
## multifunction nomograms (c. lallemand, 1883-1885)



combination of many variables into multi-function nomograms, using 3d, juxtaposition of maps, parallel coordinate and hexagonal grids (l'abaque triomph)

Ritaglio schermata acquisito: 09/03/2021 12:32

## first pictograms (m.g. mulhall, 1884)

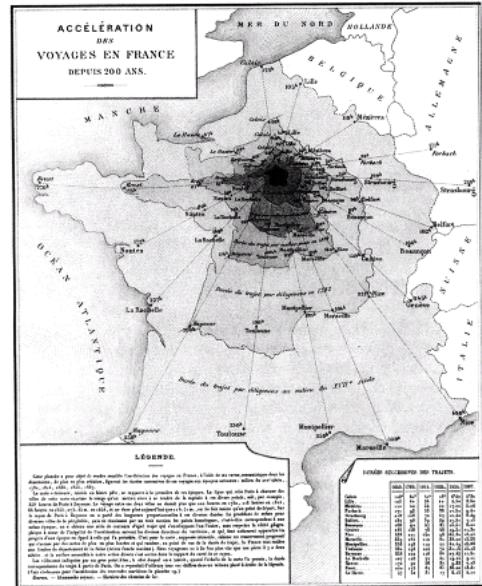


data representation by icons proportional to a number

Ritaglio schermata acquisito: 09/03/2021 12:33

Representing trains with trains -> pictograms.

first anamorphic map (e. cheysson, 1888)



deformation of spatial size to show the decrease in time to travel from paris to various places in france over 200 years

Ritaglio schermata acquisito: 09/03/2021 12:36

Geographical reference is not untouchable! The isoline of time means that it takes at least that amount of time to travel from paris to the place you want.

social mapping & color coding (c. booth, 1889)

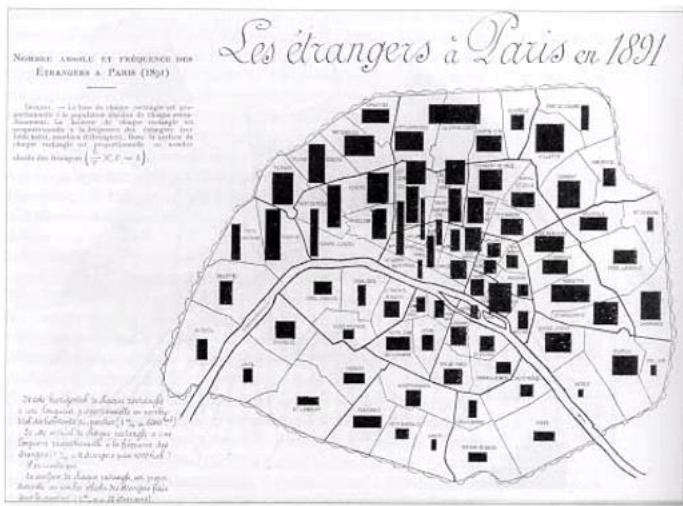


street maps of london, showing poverty and wealth by color coding

Ritaglio schermata acquisito: 09/03/2021 12:37

Printing machines become more complex and so they also produce better representation. Here we can appreciate different tone of the colors. Colors here play a crucial role.

area rectangles (j. bertillon, 1896)



use of area rectangles on a map to display two variables and their product (population of arrondisements in paris, percent foreigners; area = absolute number of foreigners)

Ritaglio schermata acquisito: 09/03/2021 12:38

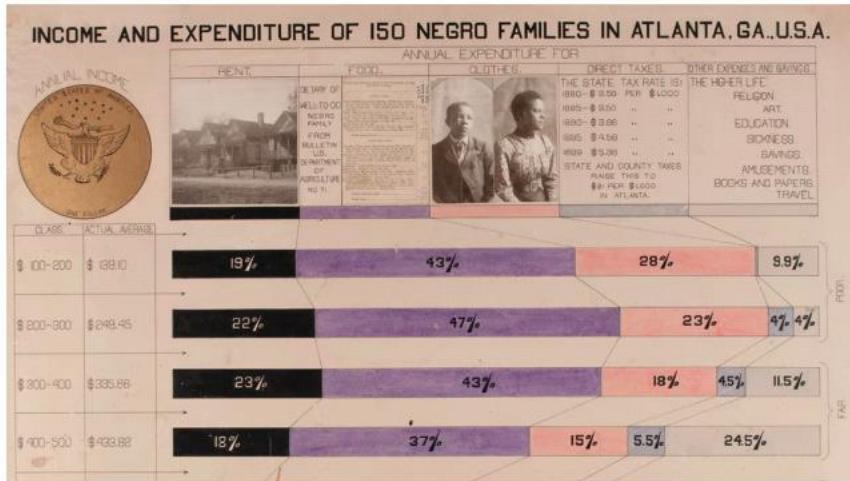
1900

We start entering the contemporary era.

- the “dark age of data visualization”
- few graphical innovations
- time of necessary dormancy, application, and popularization, rather than one of innovation
- experimental comparisons of the efficacy of various graphics forms were begun
- new ideas and methods for multi-dimensional data in statistics and psychology

Ritaglio schermata acquisito: 09/03/2021 12:38

graphs for social sciences (w.e.b. du bois, 1900)



A compendium of new graphs, handmade, for the 1900's world fair

graphs for social sciences (w.e.b. du bois, 1900)

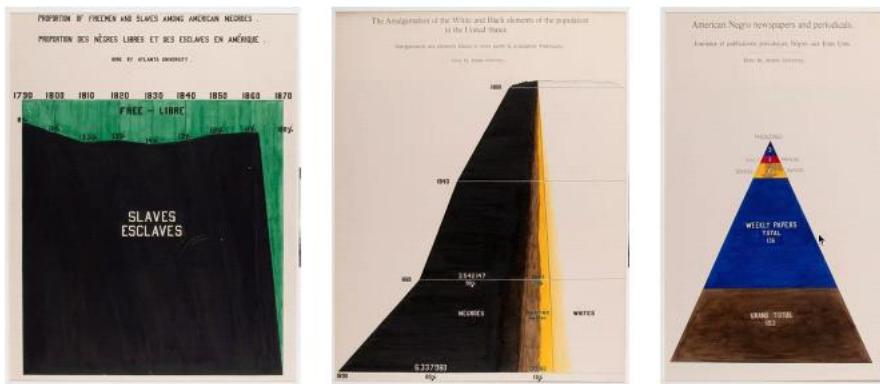


A compendium of new graphs, handmade, for the 1900's world fair

Ritaglio schermata acquisito: 09/03/2021 12:40

Large use of colors and geometrical forms for representations.

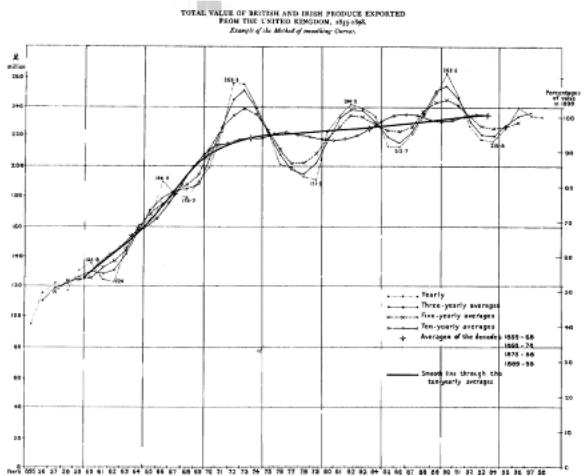
## graphs for social sciences (w.e.b. du bois, 1900)



A compendium of new graphs, handmade, for the 1900's world fair

Ritaglio schermata acquisito: 09/03/2021 12:41

## smoothing time series (a.l. bowley, 1901)

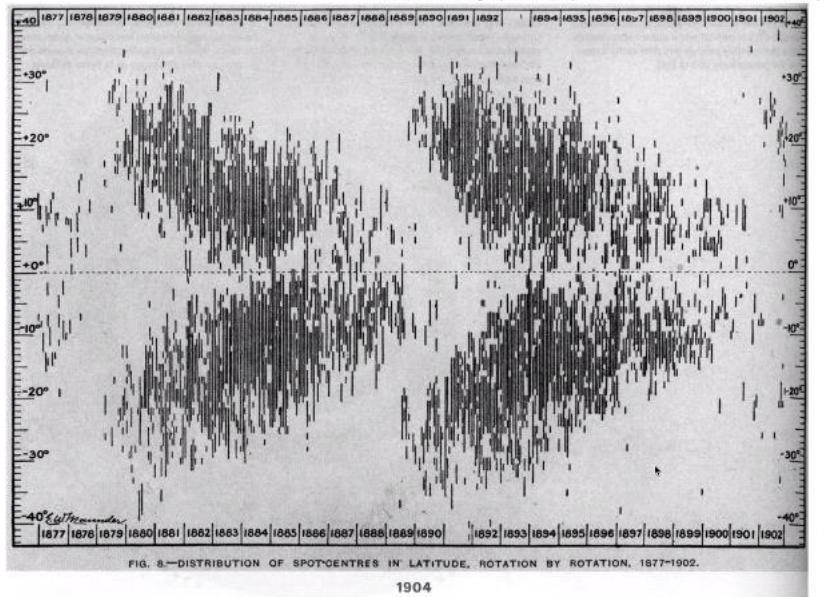


*british and irish exports from 1855-1899:* line graph of the time-series data, supplemented by overlaid line graphs of 3-, 5- and 10-year moving averages

Ritaglio schermata acquisito: 09/03/2021 12:41

People started to think that data can be modified to obtain a more regular representation. Doing it only on numbers would have been much more difficult (in terms of interpretation).

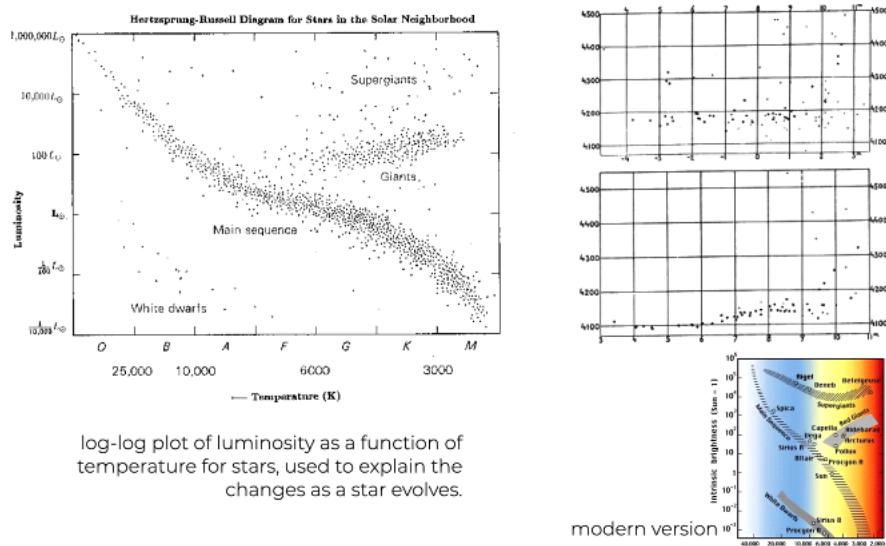
butterfly plots (e.w. maunder, 1904)



Ritaglio schermata acquisito: 09/03/2021 12:42

An example of unsuccessful graphs.

hertzsprung-russell diagram (e. hertzsprung & h.n. russell, 1911-1913)



Ritaglio schermata acquisito: 09/03/2021 12:43

More complex graphs started to appear.

### varying pictograms (w.c. brinton, 1914)

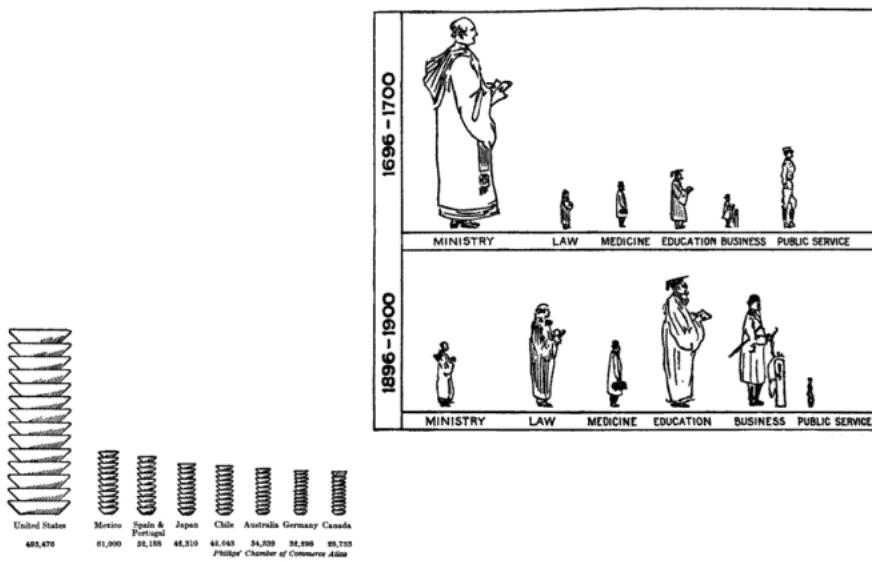
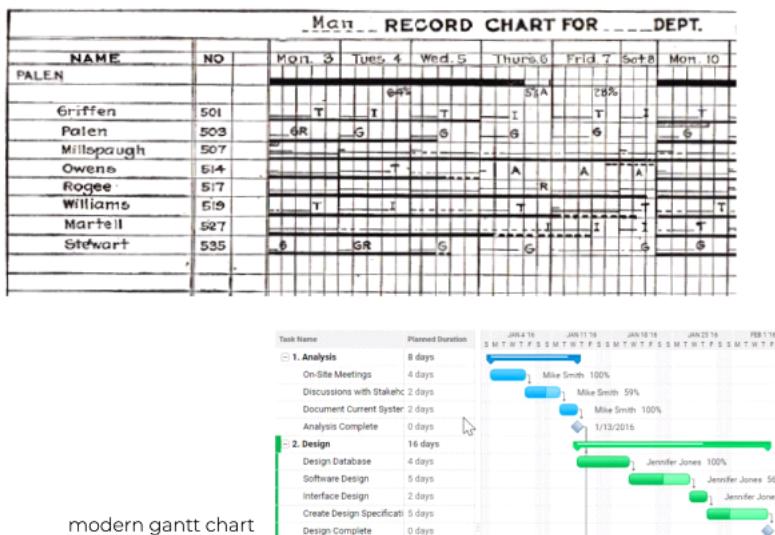


Fig. 25. A Year's Production of Copper in Tons

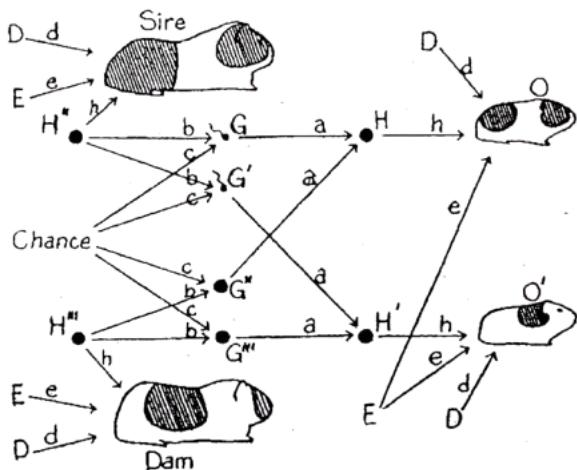
Ritaglio schermata acquisito: 09/03/2021 12:44

### gantt chart (h.l. gantt, 1917)



Ritaglio schermata acquisito: 09/03/2021 12:44

path diagram (s. wright, 1920)



relations among a network of endogenous and exogenous variables forming a system of structural equations

Ritaglio schermata acquisito: 09/03/2021 12:45

Path diagram-> relation between diagrams.

london underground map (h.c. beck, 1933)



inspired to electrical circuit board, w/ only vertical, horizontal and 45 degree angled lines  
stations located according to available space  
geographically inaccurate, but easier to use to determine how to get from point a to b.

Ritaglio schermata acquisito: 09/03/2021 12:46

This is one of the most famous reference map. Organized according to the available space and not according to the real distances. This helps a lot the users. Time travels and distances are not well described, but are less useful information.

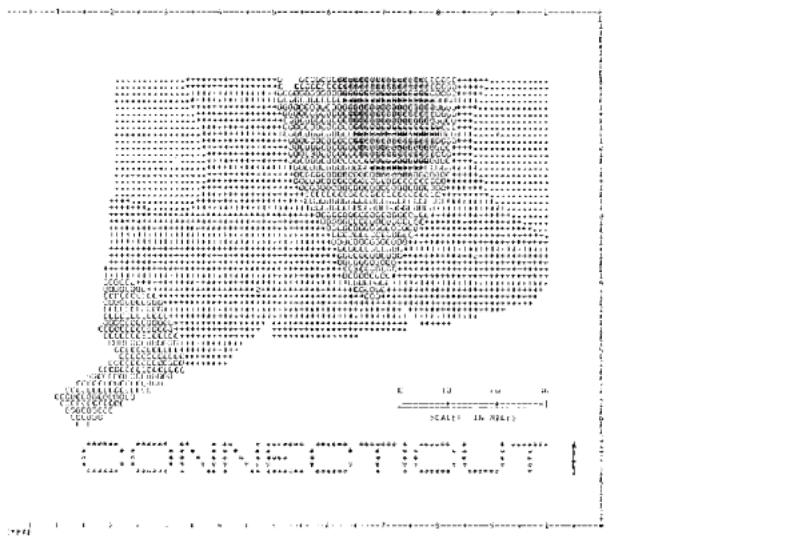
1950 -1975

- the “re-birth of data visualization”
- wide variety of new, simple, effective graphics (*j.w. tukey*)
- organization of the visual and perceptual elements (*j. bertin*)
- computer processing of data begins
- true high-resolution graphics are developed
- new paradigms, languages and software packages for expressing and implementing statistical and data graphics
- visual representations of multivariate data
- animations of a statistical process
- perceptually-based theory

Ritaglio schermata acquisito: 09/03/2021 12:48

Data viz is no more only empirical it is also crucial for theoretic studies.

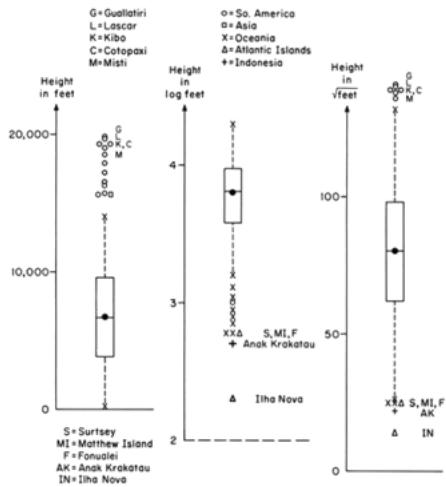
first geographical information systems (gis) (*h.t. fisher*, 1960)



isoline, choropleth and proximal maps on a line printer

Ritaglio schermata acquisito: 09/03/2021 12:49

JOHN W. TUKEY



## 1. Introduction

GRAPHS and semigraphic displays are made for purposes. Different purposes usually call for different graphs (or displays), although they do not always get them. In order of increasing importance come three broad classes:

- Graphs from which numbers are to be read off—substitutes for tables.
- Graphs intended to show the reader what has already been learned (by some other technique)—these we shall sometimes implicitly call propaganda graphs.
- Graphs intended to let us see what may be happening over and above what we have already described—these are the analytical graphs that are our main topic.

Five directions of innovation concern us:

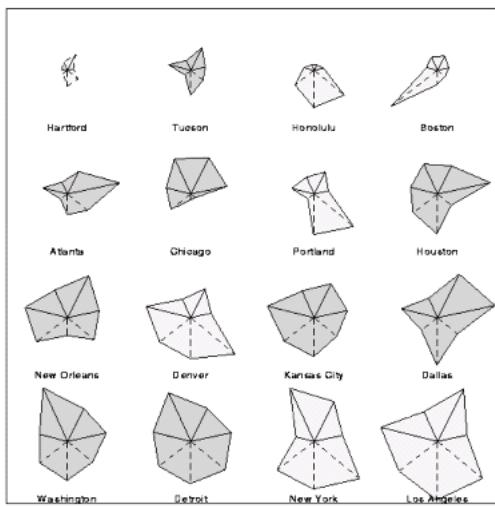
1. Displays that lie between the conventional graph and the conventional table offer real opportunities. The thought that numbers should participate in an exhibit that is at least partly graphical has been too long suppressed.
- John W. Tukey, *Data and Design*, Princeton University Press, Princeton, New Jersey, and Associate Executive Director, Research Bell Telephone Laboratories, Murray Hill, New Jersey.
- This paper has been prepared in part in connection with research at Princeton sponsored by the Army Research Office, Durham, and in based on a paper presented at the International Conference on Data Analysis, held at the Institute of Mathematical Statistics, and INSTAT of the Royal Society, August 1969.

293

Ritaglio schermata acquisito: 09/03/2021 12:50

The box and whiskers plots was invented by Tukey and this was an incredible innovation.

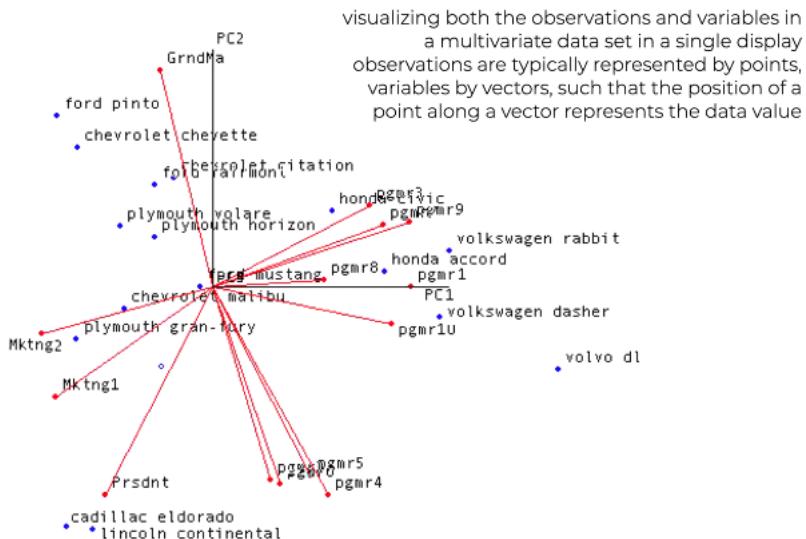
star plot (h.p. friedman, r.m. goldwyn, j.h. siegel, 1971)



vertices at equally spaced intervals, distance from center proportional to the value of a variable

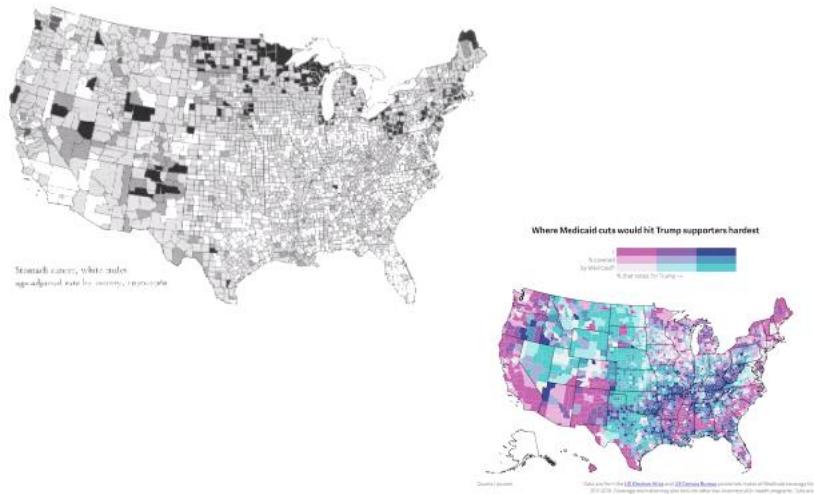
Ritaglio schermata acquisito: 09/03/2021 12:51

### biplot (k.r. gabriel, 1971)



Ritaglio schermata acquisito: 09/03/2021 12:51

### bivariate matrix (u.s. bureau of census, 1974)



color-coded bivariate matrix to represent two intervally measured variables in a single map

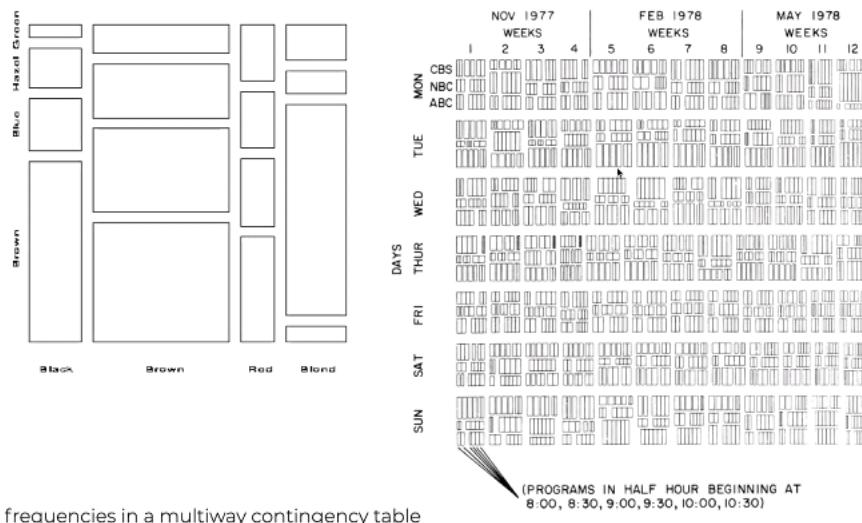
Ritaglio schermata acquisito: 09/03/2021 12:53

1975-Now

- highly interactive computer systems
  - new paradigms of direct manipulation for visual data analysis (linking, brushing, selection, focusing, etc.)
  - new methods for visualizing high-dimensional data (grand tour, scatterplot matrix, parallel coordinates plot, etc.)
  - new graphical techniques for discrete and categorical data (fourfold display, sieve diagram, mosaic plot, etc.)
  - extensions of older ones (diagnostic plots for generalized linear models, mosaic matrices, etc.)
  - application of visualization methods to an ever-expanding array of substantive problems and data structures
  - increased computer processing capacity, allowing computationally intensive methods + big data problems
- 

Ritaglio schermata acquisito: 09/03/2021 12:53

mosaic plot (j. hartigan, b. kleiner 1981)

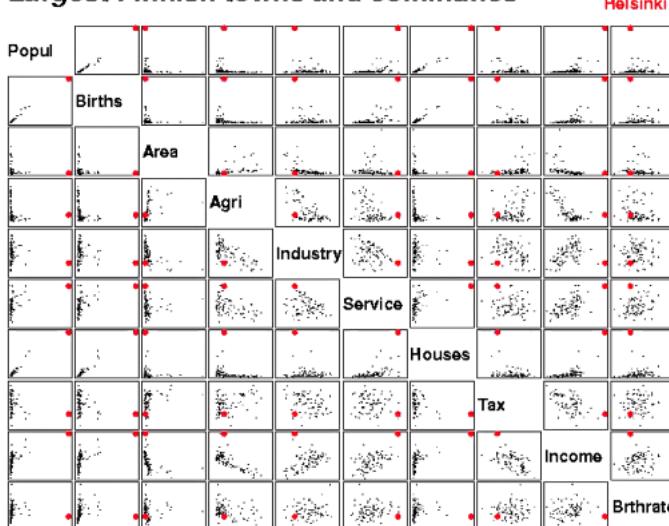


frequencies in a multiway contingency table

Ritaglio schermata acquisito: 09/03/2021 12:55

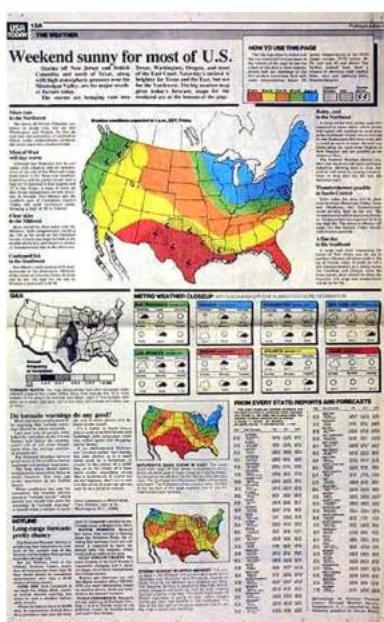
draftsman display (j.w. tukey, p.a. tukey 1981)

### Largest Finnish towns and communes



Ritaglio schermata acquisito: 09/03/2021 12:55

Mutual correlation between variable on the columns and variable on the rows.



usa today weather maps (g. rorick 1982)

the usa today color weather map begins an era of color information graphics in newspapers.

shortly, colorful visual graphics become widespread: the infographics

Ritaglio schermata acquisito: 09/03/2021 12:56

The rise of iconographic -> not very scientifically accurate but full of content and very readable.

## interactive graphics (r.a. becker, w.s. cleveland 1987)

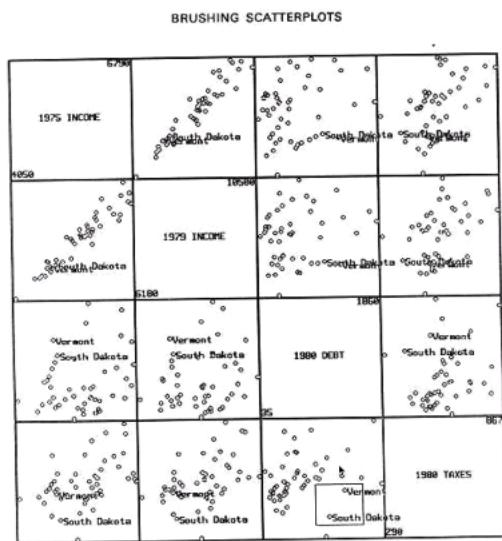
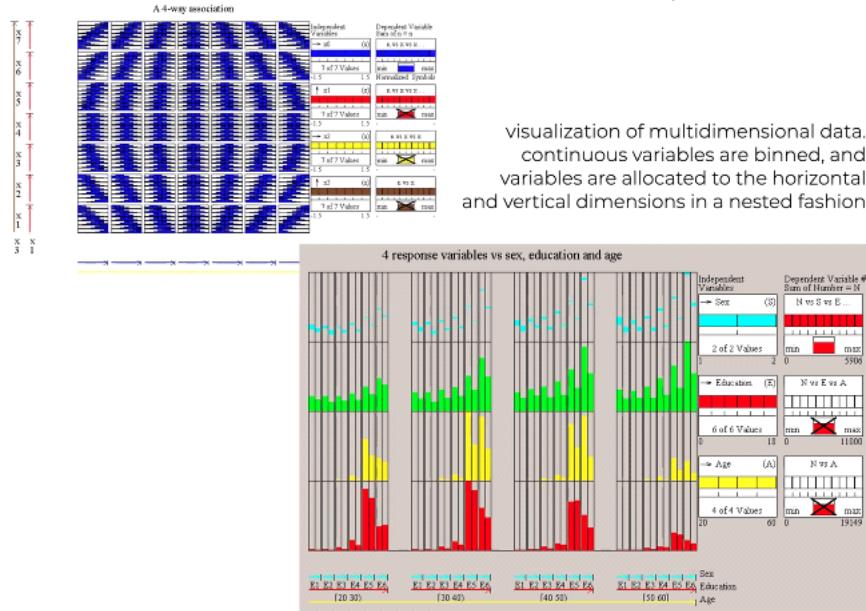


Figure 14. State Data. The label operation is being used. The two points inside the brush on the active panel have their displayed as well as corresponding points on other panels.

Ritaglio schermata acquisito: 09/03/2021 12:57

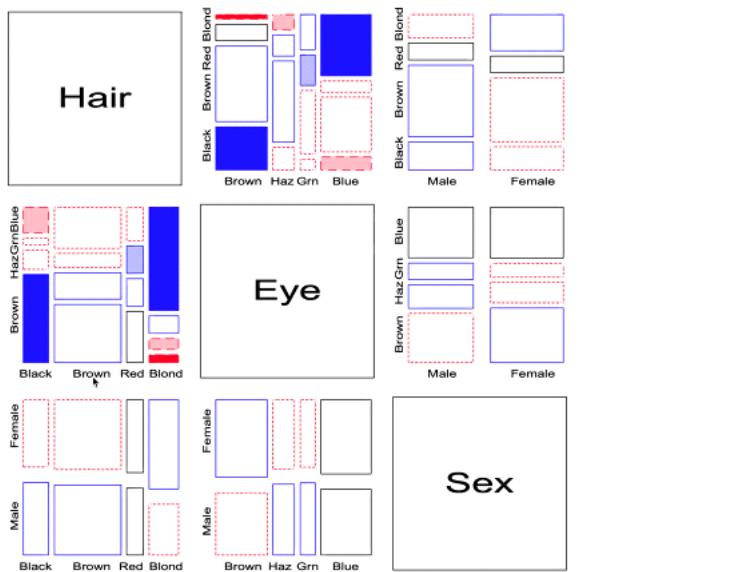
## nested dimensions (t. mihalisin 1989)



Ritaglio schermata acquisito: 09/03/2021 12:57

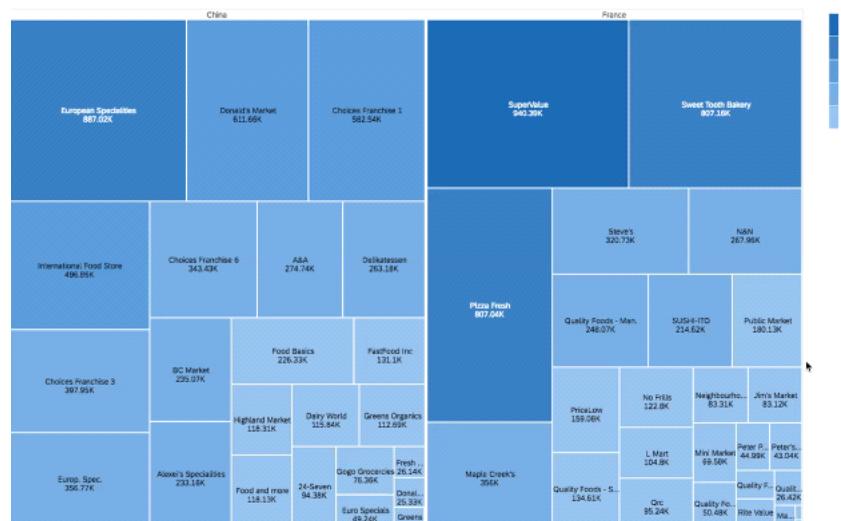
Difficult to read but summarize many information in a unique plot.

## enhanced mosaic plot (m. friendly 1991)



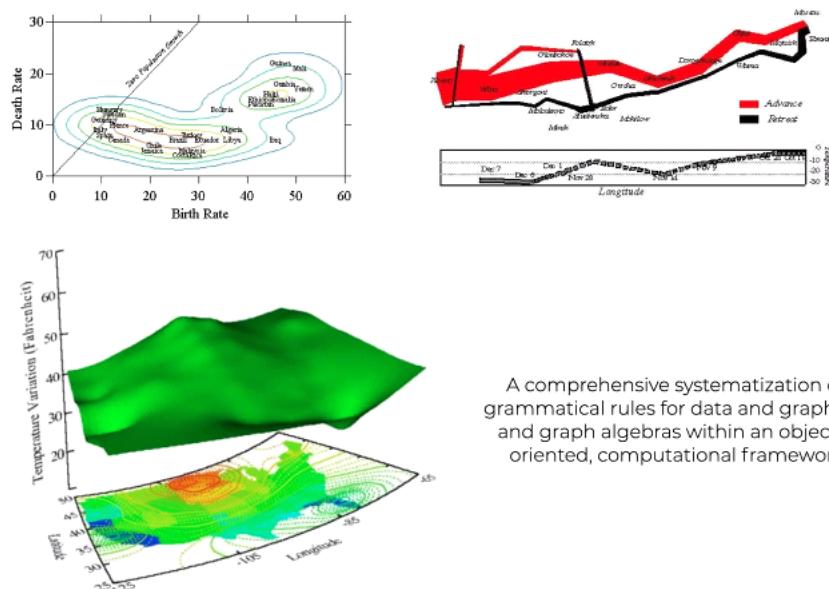
Ritaglio schermata acquisito: 09/03/2021 12:58

## Marimekko treemap plot (m. shneiderman 1992)



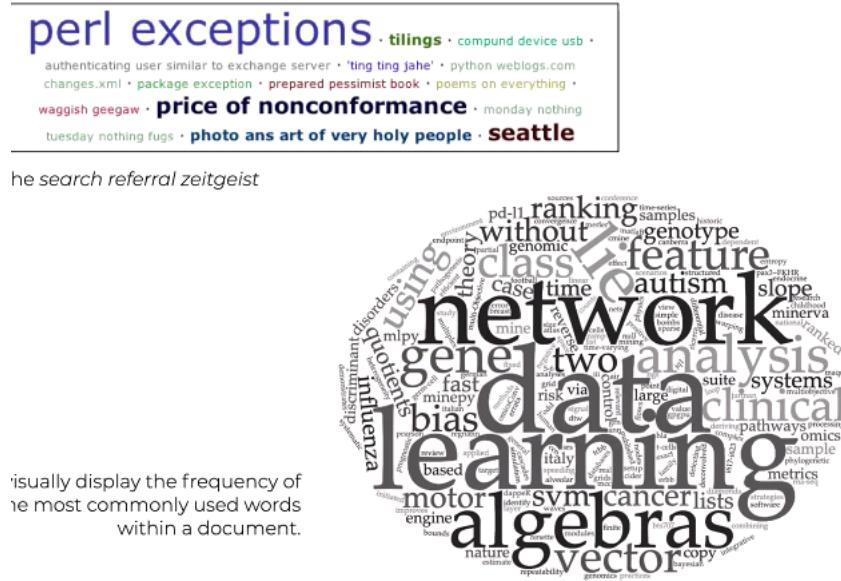
Ritaglio schermata acquisito: 09/03/2021 12:58

### grammar of graphics (l. wilkinson 1999)



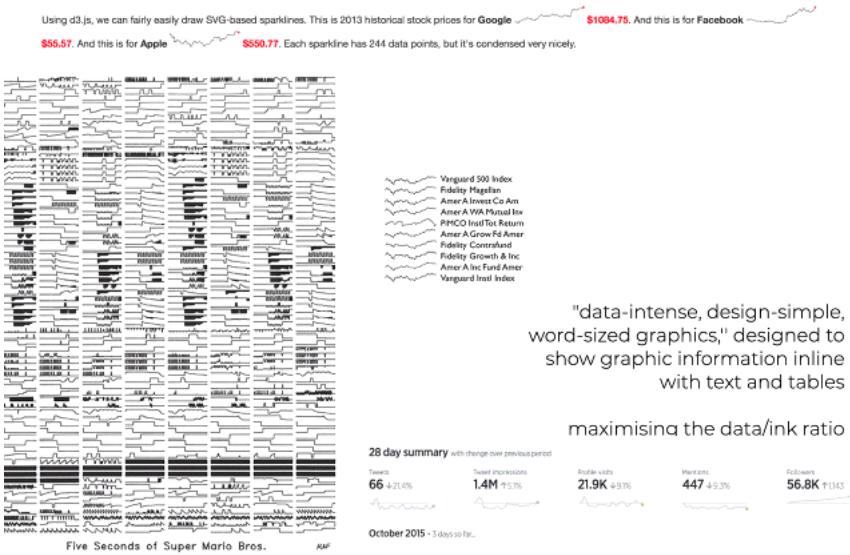
Ritaglio schermata acquisito: 09/03/2021 12:59

### tag/word cloud (j. flanagan 2002)



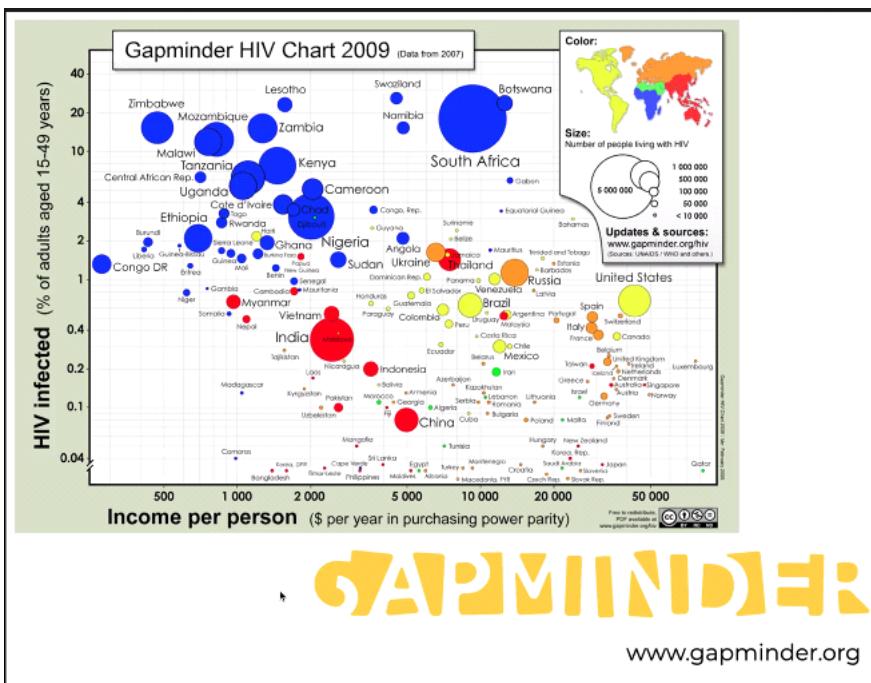
Ritaglio schermata acquisito: 09/03/2021 13:00

## sparklines (e. tufte 2004)



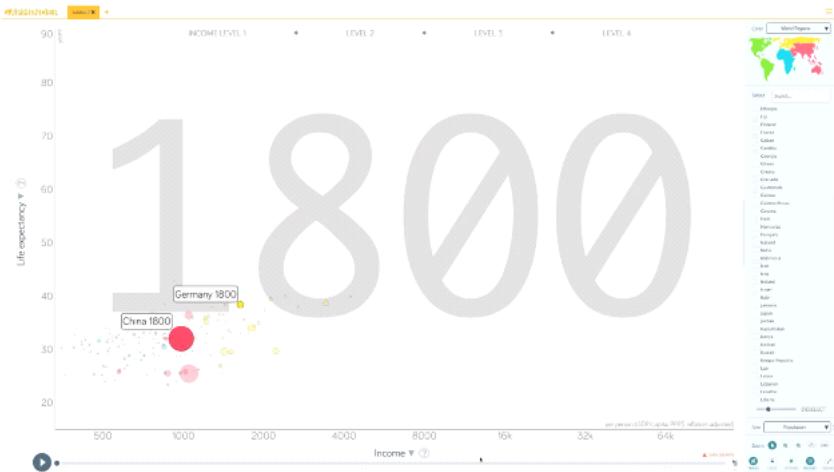
Ritaglio schermata acquisito: 09/03/2021 13:00

Tufte is a key figure for the modern data visualization.  
They maximize the data/ink ratio.  
No ink is wasted in things that are not directly related to data.



Ritaglio schermata acquisito: 09/03/2021 13:02

Gapminder is more than a plot it is a website, an initiative to spread the concepts of data visualization to a larger audience. ( Video on the slides).



<http://www.gapminder.org>

Ritaglio schermata acquisito: 09/03/2021 13:03

Compare socio-economical behavioir for different categories through the time. We have a comparison between china and germany according to income and life expectancy.



non-profit foundation founded in 2005 with a goal of "... increase use and understanding of statistics and other information about social, economic and environmental development at local, national and global levels."

[Ie, 2013] "empower instructors in designing dynamic presentation of real life data, energizing students"

seizes the power of statistics by expressing 5-dimensional data into one place

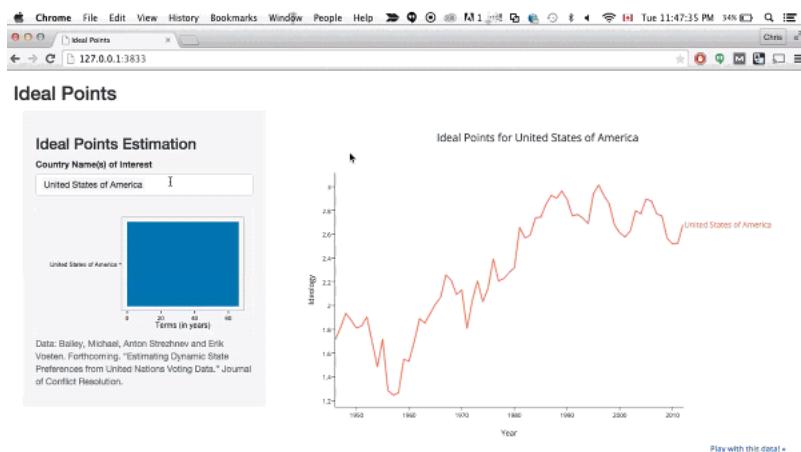
the 5 dimensions of the gapminder graphs are:

- variable on the horizontal axis,
- variable on the vertical axis,
- time,
- geography (color of the dot),
- population (size of the dot).

[smaranda, 2016]

Ritaglio schermata acquisito: 09/03/2021 13:05

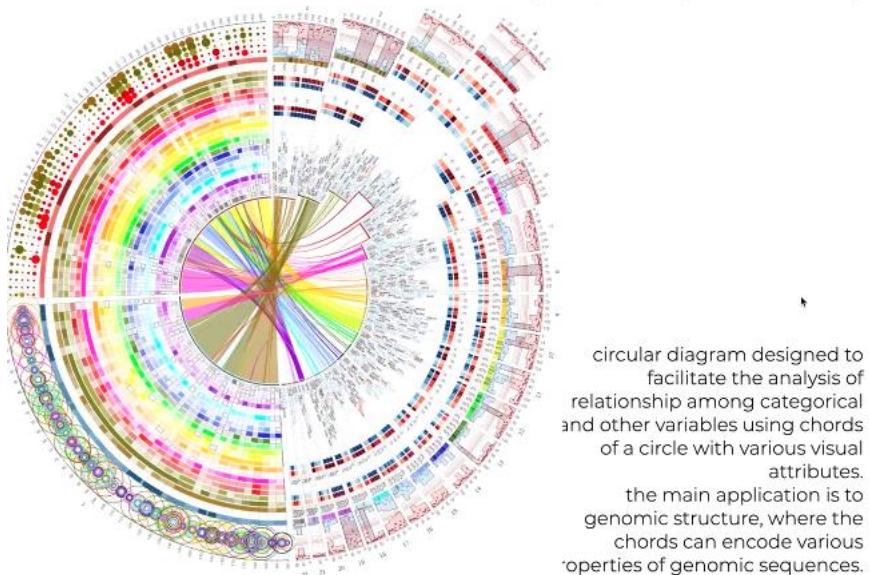
## ggplot2 (h. wickham, 2006)



an influential, open source implementation of the grammar of graphics in r, together with other computational tools to make it easier to produce beautiful and interactive statistical diagrams

Ritaglio schermata acquisito: 09/03/2021 13:06

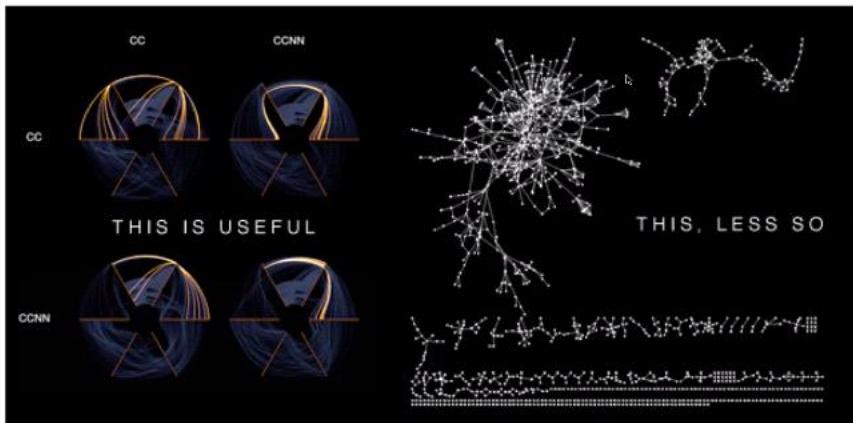
## circos plot (m. krzywinski, 2009)



Ritaglio schermata acquisito: 09/03/2021 13:08

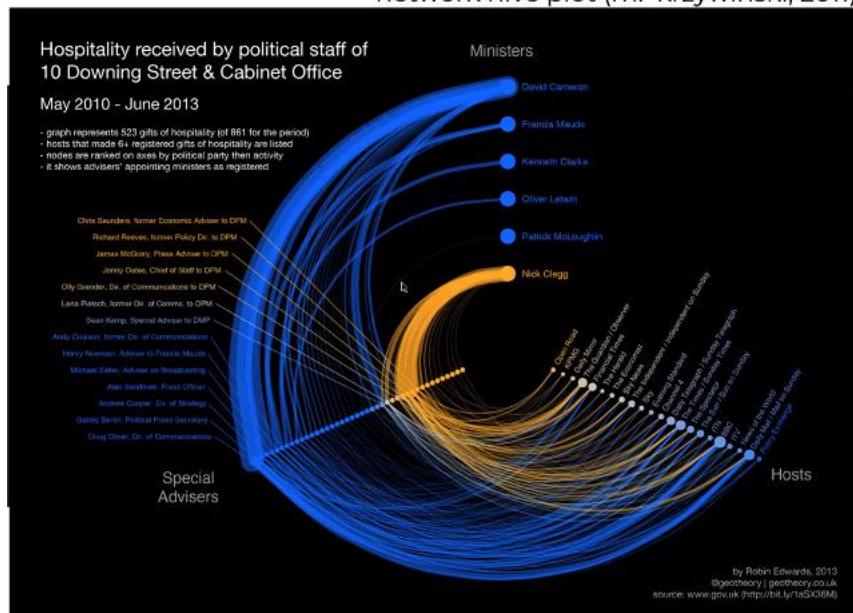
This is the standard for representign several levels of knowledge fro genetic code.

## network hive plot (m. krzywinski, 2011)



Ritaglio schermata acquisito: 09/03/2021 13:10

## network hive plot (m. krzywinski, 2011)



Ritaglio schermata acquisito: 09/03/2021 13:11

For the future we expect virtual 3D plots enhancing the interaction with the users.

## Foundations of data visualization

venerdì 19 marzo 2021 19:09

We will explore the theory regarding how to interpret and design a good data visualization according to the principle recently developed.  
Data visualization is constantly evolving. There is not a set of precise rule. Data Viz is always changing.  
Data visualization

There are 4 pillars:  
Representation -> transforming the data in sth that can be graphically exposed and explored  
Presentation -> all the surrounding material is used to nicely represent data and information ( data optimization)  
Facility understanding



Ingredients:

Data is the key actor in the dataviz process

Tabular form is what we consider the raw form of data

Raw data (table form) cannot be compared and contrasted effectively.

Data representation -> is the visualization of data through the combination of marks (lines; points...) a Data presentation -> concerns all other visible design decisions beyond the representation: interactivity Facilitating understanding should be the goal.

Understanding -> according to the latest psychological description when interpreted for dataviz can be st

Perception -> perceiving an image; the first feeling when we see a graph

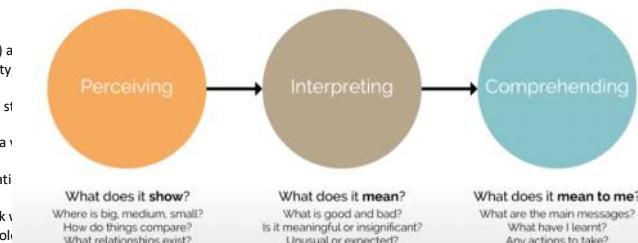
Interpreting -> now that I have understood the element of the graph, which is the meaning of the data

Graphical representation tells me something useful or something that I completely expect?

Comprehension -> now I interpret the plot, but for my purpose which is the meaning of the interpretation

Perception -> where are the big elements the small elements and the chunks. Is there any relevant link or simply being able to read a chart efficiently decode the representation of the data (shapes/sizes/colors)

- What is the relation with the small things and the whole picture?
- Where are the largest/middle-sized/smallest values?
- What proportion of the total does that values holds?
- How do these values compare in ranking terms?
- To which other values does this have a connected relationship?



Interpretation -> this requires finding out the real meaning of the plot, good and bad elements.

What is interesting and what is obvious? This are the main milestones in interpreting.

Interpreting is the art of converting the perception into meaning; use the pre-existing knowledge to frame the implication of the viz.

- Is it good to be big or is it better to be small?
- What does it mean to go up or to go down?
- Is that relationship meaningful or insignificant?
- Is the decline/increase of that category especially surprising?

Comprehension -> turning the meaning to my aim. Comprehension is the art of reasoning about the consequences of interpretation what is the novelty carried by this dataviz about the subject?

- Why is this relevant? To whom?
- Has it confirmed what I suspected or enlightened me with new knowledge?
- Has this impacted me emotionally or left me indifferent?
- Does this understanding force me to take action on the subject?

This is an example of a barplot which represents Lionel Messi's goals per match.

The perception in this example involves identifying the type of chart, axes and labels (year and goals); identifying big, small and medium values -> there are some small elements (left part of the graph); big elements in the middle part of the graph and then smaller values -> we can spot a growing trend until 2012 and then there is a small decrease somehow that puts back the situation to a medium size.

Clearly I can also compare the different values among the years.

Interpreting -> I put my background. Previous knowledge: scoring 25 is good, scoring 70 is amazing; scoring them in La Liga/UCI is even more amazing; a ratio of almost 1 goal/game is very rare in football; even more remarkable if coupled with Messi's age.

Comprehension -> imagine we are doing a report on Messi -> what we understand from the plot is a confirmation that Messi is one of the best players in the world.

This chart also opens an interesting window: there is a small decrease in performance in 2012/13.

What is the reason why?

We now from our experience that Ronaldinho left Barcelona in that year and this could have affected the performance of Messi; another hypothesis could consider injuries.

The next graph is the same -> picking up the same graph we change the color and the labels. This shows the number of points for an hockey player. (this data are invented). We want to explore the differences between this graph and the previous. What can I say if I am not an expert of hockey. What change is that the way points are not given is not the same for football so it is difficult to find a proper meaning.

It is not easy to interpret this chart without being an expert of hockey.

Now we consider a third graph. Just changed the title -> Winglets and Spangles.

They are two nonsense words. The perception is the same but the other two steps are completely different because I have zero knowledge; no interpretation is possible. This yields that comprehension is completely senseless.



abraham wald  
1902 1950  
ph.d. maths, uni vienna, 1928

member of columbia us statistical research group during wwi

how to protect bombers to get shot by german planes?

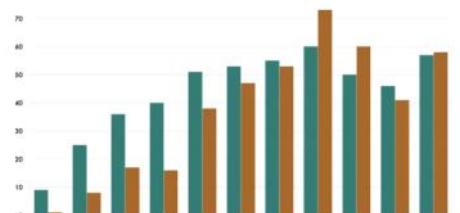
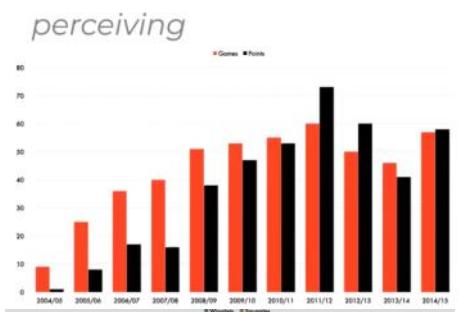
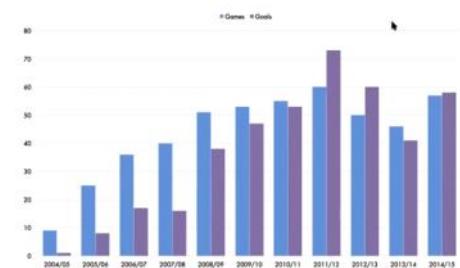


How to optimise position of additional protecting armor?

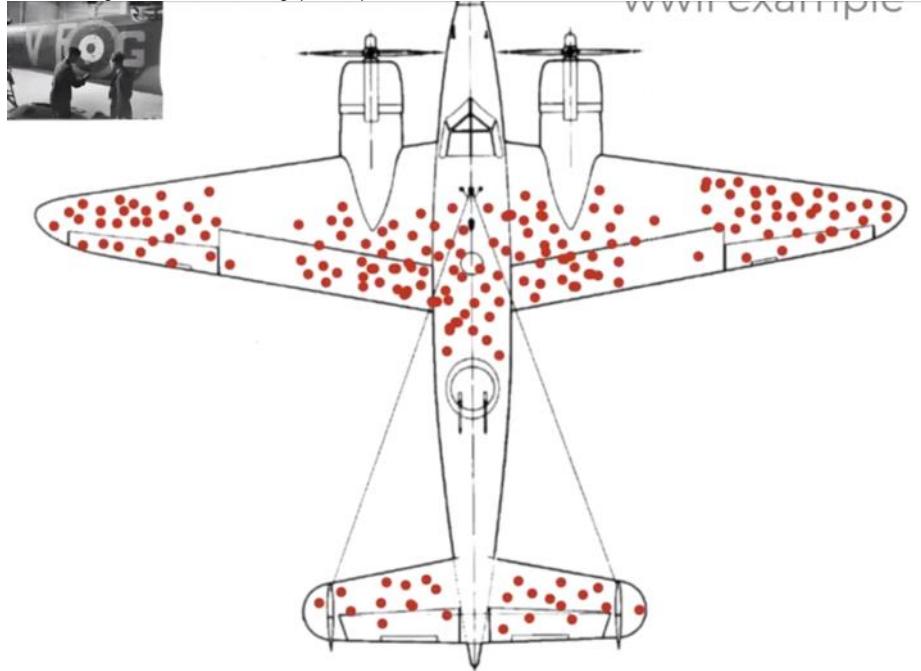
The idea was to use panels of steel to better protect the aircrafts.

Let's consider this problem in the field of statistical problems.

So first thing is to collect data regarding the damages caused to the airplanes.

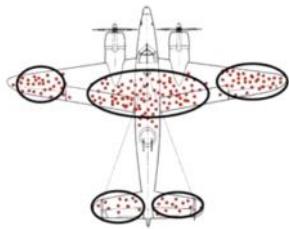


From that they create statistical models and from that they found solutions.  
Instead of using the model we move to the graph so the plot of the aircraft.



Ritaglio schermata acquisito: 05/04/2021 15:54  
Where should the new armors be put?

*perception*



Ritaglio schermata acquisito: 05/04/2021 15:56

First thing -> distribution of the bullet holes. Moving to the interpretation -> statistics tells me that bullet holes should be equally distributed.  
The second observation is the winning one -> data only from returning aircrafts !  
Comprehension -> **Survivor bias** -> when you have partial information about the data. When you have partial view of the data.  
Principle for a good data visualization  
Exploring potential principle of good design. Transfer concept of good design into dataviz.  
Good design is innovative; makes a product useful, aesthetic, makes a product understandable, is unobtrusive, honest, long lasting , is thorough down to the last detail, is a little design as possible.



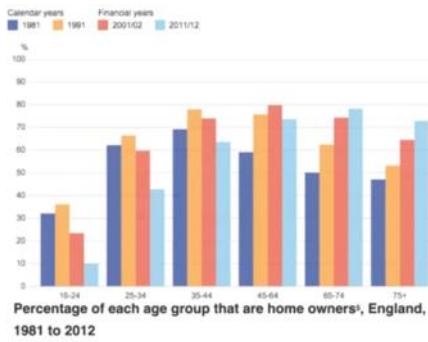
Ritaglio schermata acquisito: 05/04/2021 16:09

integrity + accuracy + legitimacy

- **truth** is an obligation
- different legitimate versions of truth can exist in dataviz, being the outcome of different pathways
- pure objectivity is rarely possible in dataviz
- there is a need to show that the shown truth is **trustable**
- even a true graph may not be viewed as trustworthy

Ritaglio schermata acquisito: 05/04/2021 18:37

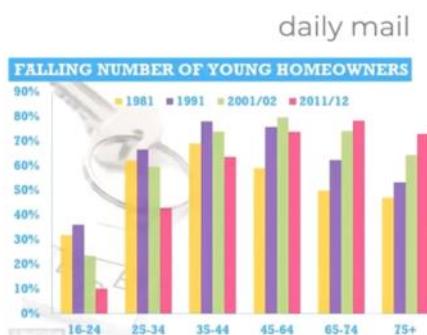
Pure objective is rarely possible. Objective way can be expressed from different point of view and being still true.  
Strong need to show that your data/representation is trustable. You may present a true graph and be not trustable or at the contrary being trustable but false.



Source: English Housing Survey (EHS) 2012 to 2013, Table FC2101, OCLG - EHS 2001/02, Table S106, OCLG

uk office for national statistics

same true data,  
different trustworthiness

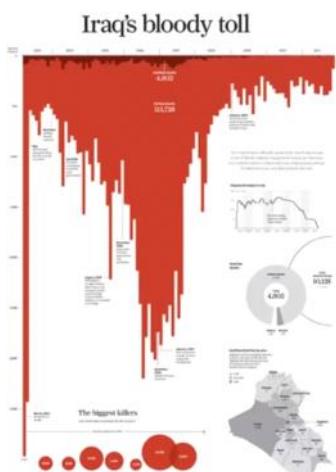


Ritaglio schermata acquisito: 05/04/2021 18:41

These two graphs are identical but they have a different level of trustworthiness.  
The left graph looks more trustworthy for example in the left part the source is shown, the left graph is neutral while the right one influence the readers with its title.

- colors
- fonts
- background
- data description
- data source

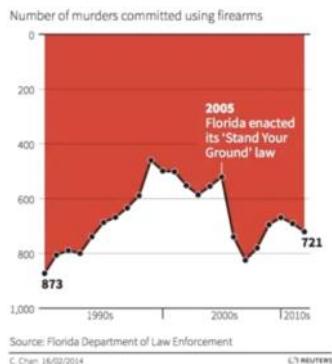
The above graph are also influenced by the above elements.



Ritaglio schermata acquisito: 05/04/2021 18:47

This is a very famous infographic concerning us casualties in the golf war. This is a reverse bar plot that want to recall blood.

## Gun deaths in Florida

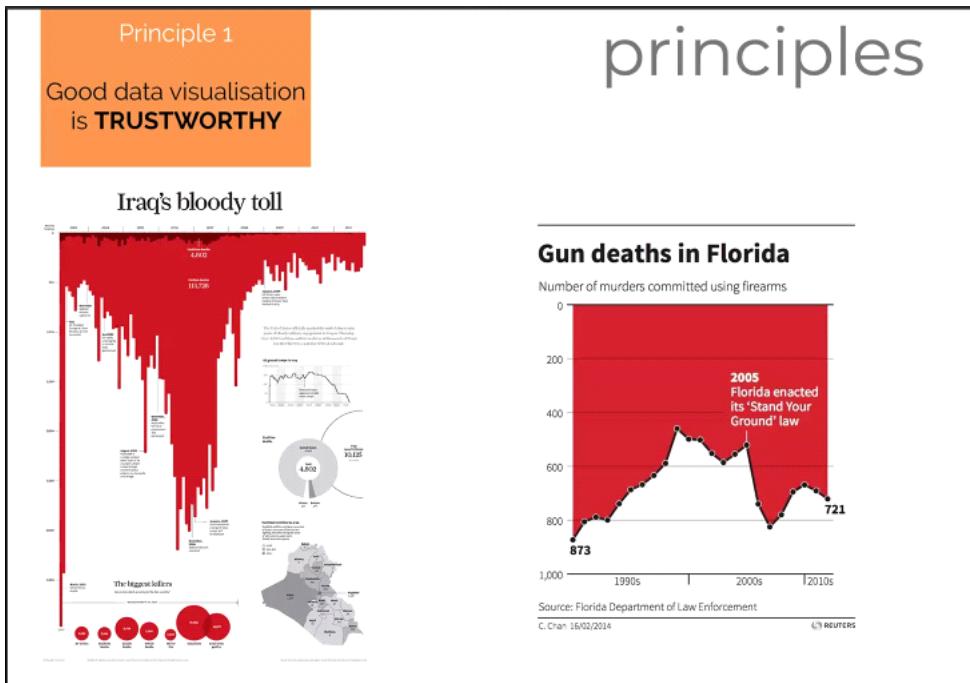


Ritaglio schermata acquisito: 05/04/2021 18:57

This second graph compared to the second one is less effective. We completely lose the blood effect. This graph is perceived differently from the reader. Is not obvious what are the data and what is the background.

# Principles

martedì 23 marzo 2021 11:16



Ritaglio schermata acquisito: 24/03/2021 11:31

Many plots can be similarly true but not being similar trustworthiness.

Good data visualisation is **TRUSTWORTHY**

pursuing trustworthiness in **data processing**

- how was data collected: from where & using what criteria?
- what calculation or modification have you applied to it?
- explain the approach in details
- have you made any significant assumption or observed any special counting rules that may not be common?
- have you **removed** or **excluded** any data?
- how representative it is?
- what biases may exist that could distort interpretation?

Ritaglio schermata acquisito: 24/03/2021 11:32

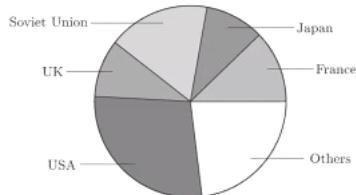
We may have reason to remove some data ( outliers, wrongly collected but we have to clearly express this to the reader).

## pursuing trustworthiness in **data representation**

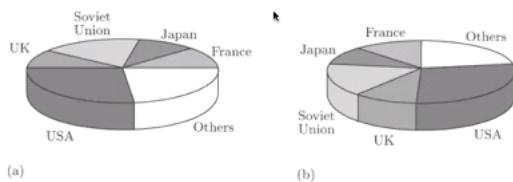
- never **deceive** the receiver
- avoid misunderstandings, inaccuracies, confusions and distortions
  - \* quantitative values represented by areas can be disproportionately perceived

Ritaglio schermata acquisito: 24/03/2021 11:38

Avoiding misunderstanding and distortion. We see many different examples of that. Fooling the receiver in an unwanted way .



Ritaglio schermata acquisito: 24/03/2021 11:39



Ritaglio schermata acquisito: 24/03/2021 11:40

This are two different ways to represent the same pie chart.

The perspective can give you different feelings.

UK seems to be much larger in the a representation compared to the b representation.

With this representation you communicate misleading information due to how human eyes perceive the perspective.

3-d Pie chart -> whatever is close to the reader seems far larger. This distortion is inherited in the way you present data. Tricks like this can make the representation misleading and tell to the reader different things. So you may suspect that the author wants to convince you of something that is not really right.

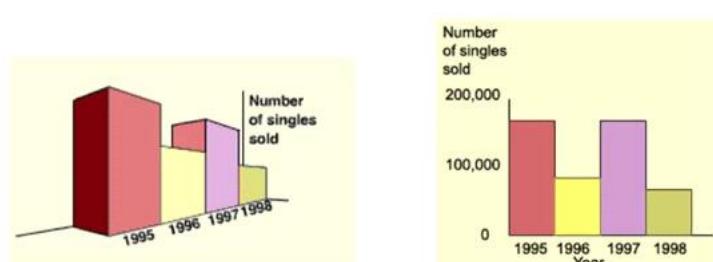
- avoid misunderstandings, inaccuracies, confusions and distortions

\* 3d representations are often nothing more than distraction — distortion — decoration; use only if there are 3 dimensions in data and the viewer can change point of view to see different 2d perspectives

Ritaglio schermata acquisito: 24/03/2021 11:46

3D shuld be used only if data is itself in 3D.

3D change the point of view.

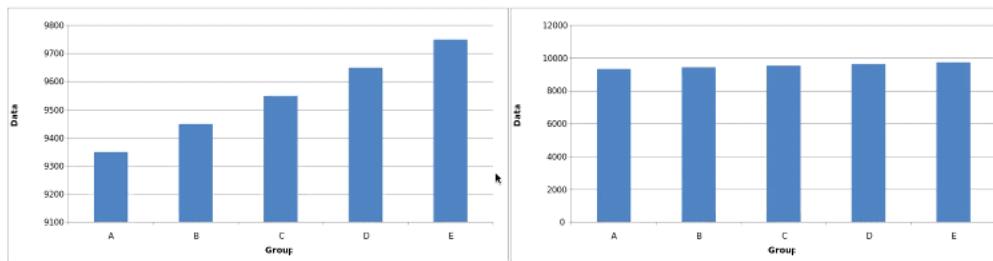


Both the graph refers to records sales.

The number of singles sold in 1995 and 1997 are the same but with the 3D this is not really perceived. This is an example of representation that wants to fool the audience.

Another common effect regards the axes

- \* axes (in bar charts) should never be truncated, origin must be 0



The most trustworthy representation is the right one because consider the origin and has bigger interval so the difference is minimal.

It is also difficult to grasp the increase in the right graph. At a first look you would say that the bars on the right have the same height.

While on the left we have a zoom and we can appreciate a trend.

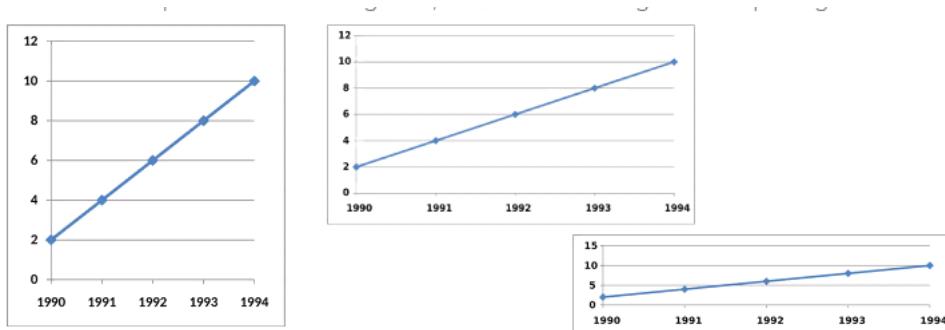
The left graph is also important because shows different information. According to the different task the choice of representation we use is influenced by many factors.

If the left hand is important, the trend growth is important, you should show and tell to the reader that what you are representing is just a zoom.

The important thing is to be clearly explicit.

The aspect ratio is the quotient of the unit of measure of y divided by the one of the x.

- \* aspect ratio in line chart is influential, since it modifies the perception of the steepness of connecting lines, either embellishing it or dampening it



These three graphs are three different versions of the same graph.

The perception of the slope is different but the use of the different representation can be useful for specific purpose.

- \* different projections (mercator, etc) may alter the perception of a thematic map, distorting size/shape of regions.



Ritaglio schermata acquisito: 24/03/2021 12:06

**Mercator projection stretches whatever is near to the equator and distorts whatever is near the poles.**

In this case it is up to the data visualizer to decide which projection is the more appropriate.

### pursuing trustworthiness in **data presentation**

- if it looks significant, it should be
  - absent annotations such as introduction/guides, axis title and labels, footnotes, data sources fail to inform the reader
  - inconsistent or inappropriate color schemas
  - confusing layouts
  - does the chosen solution work, and, specifically, does it work in the way it promises to do?

Ritaglio schermata acquisito: 24/03/2021 12:08

**Pursuing trustworthiness:**

- Adopting inconsistent color schema can be very dangerous when presenting data viz.
- The simplest layout that supports your idea without confusing the reader.
- The representation respects what I wanted to do

**Principle two -> accessibility**

Accessible -> means that the viewer should make the smallest effort to understand what it is represented.

the viewer should experience minimum friction between the act of understanding (**effort**) and the achieving of understanding (**reward**)

#### reader's side:

- subject/matter appeal
- dynamic of need
- subject/matter knowledge
- what they need to know
- issue of unfamiliar representation
- time
- format
- personal taste
- attitude & emotion

#### author's side

- lack of focus
- not deep enough
- oversimplified representation
- unfit for the setting
- visually inaccessible
- misjudged format, details get lost
- too many interactive options
- too complex representation
- absent annotations

Ritaglio schermata acquisito: 24/03/2021 12:13

There are certain things that are not under the control of the author for example the reader taste. For example the reader has problems with a specific kind of representation and so on.

On the other hand there are many things under the control of the author so he should focus on these key points to get the best out of its representation.

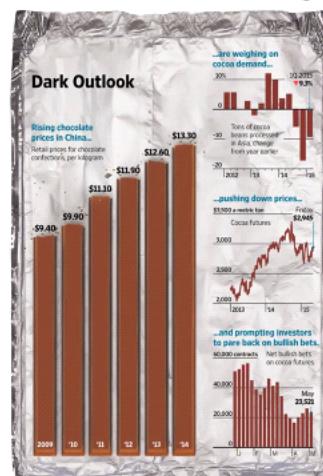
#### Principle 3

Good data visualisation  
is **ELEGANT**

## principles

achieving a visual quality to attract your audience, without letting the style overcome the substance

- eliminate the arbitrary: "remove to improve"
- thoroughness
- develop a style
- decoration should be additive, not negative
- offer elegant & appealing presentation congruent with the subject
- do not pursue minimalism at all costs



Ritaglio schermata acquisito: 24/03/2021 12:21

Being **elegant** -> **attract the audience** -> good visual quality -> do not let the style overcome the message.

Try to eliminate everything that is arbitrary. The decoration should stimulate the reader to understand the underlying topic.

Not pursuing minimalism at all costs even if there is a large community fighting to minimize at all costs but this is not always a good strategy.

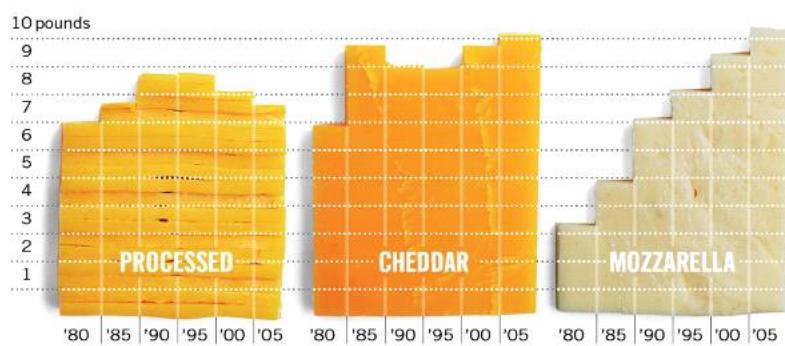
The representation recalls its message, using chocolate bars for the chart is nice and does not overwhelm the message. The chocolate bars give a good impression to the reader who is led to think about chocolate. This is probably the main driver. Another element is the aluminium in the

background which recalls the chocolate wrapping -> it is not disturbing and is nice. This is quite a good way to add hestetics. Decorating an infographic without disturbing the audience.



Too many portrait details compared to the inforgraphic.  
The main plot is about three integers number.

Per capita cheese consumption in the U.S.



Ritaglio schermata acquisito: 24/03/2021 12:40

Decoration should be additive, not negative.

For each cheese we have just six points. There is no separation between bars. The first impression is that the plot is more focused on the area rather than the trend of consumption. Your eye is more focused on comparing the three areas rather than the trend.

In many cases the representation is a matter of taste.



Ritaglio schermata acquisito: 24/03/2021 12:46

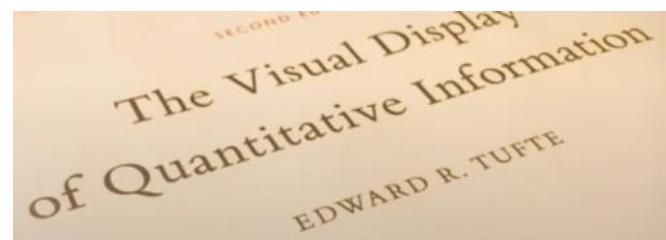
Here the author was intelligent because he knew that for human comparing areas is quite hard. It is not easy to look only at the areas but the addition of the areas representing the squared meters

helps to interpret and having a good idea of the content.

## Data visualization workflow

lunedì 5 aprile 2021 19:07

How to set up a correct pipeline to design, plan and implement and deploy a data viz.



Ritaglio schermata acquisito: 05/04/2021 19:10

excellence in dataviz consists of complex ideas communicated with clarity, precision and efficiency; graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, technology of production
- avoid distorting what data have to say
- present many numbers in a small space
- make large dataset coherent
- encourage the comparison of different pieces of data
- reveal the data at several levels of detail
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal description of the dataset

Ritaglio schermata acquisito: 05/04/2021 19:11

The data has to be the ACTOR. Data is the driver of any plot. Whatever is in the data can be shown. You should not allow that technicalities overcome the message. You have to plan some graphical representation that allow to show many information/data/numeric values in a small space. The purpose of the data viz should be clear.

20% of the decision you make influence the 80% of the results.

The most reliable approach is to follow a design process. (we will follow the suggestions of Andy Kirk)



Ritaglio schermata acquisito: 05/04/2021 19:15

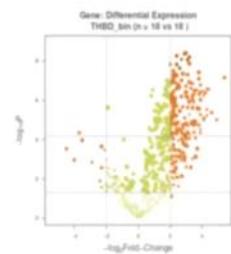
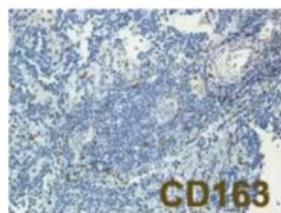
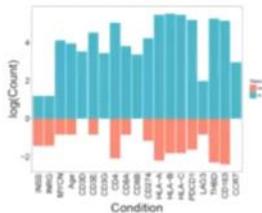
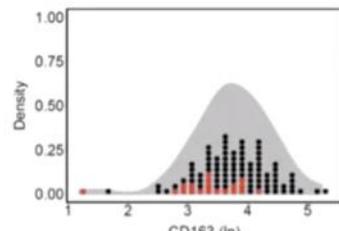
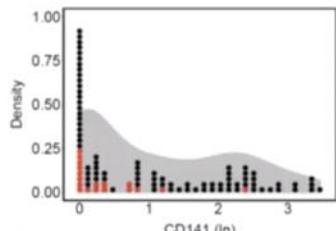
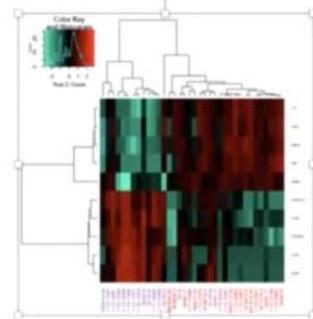
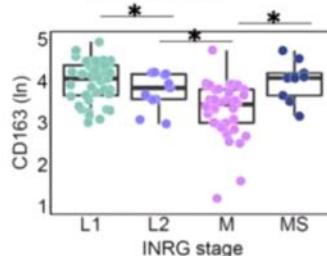
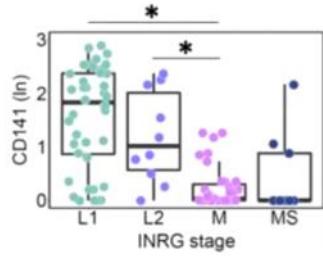
Why a design process? A design process allows to be more pragmatic and less dogmatic. There are rules but flexibility is more important. Moreover, it helps to reduce randomness: break down the goal into a connected system of thinking.

A design process leaves room for experiment; facilitate adaptability and iteration. A design process allows a continuous interaction with the subject; you have always the same process. Lastly, the organization helps to partition your mindset; thinking/doing/making.

How the design process is organized/planned? You need to manage the progress and resources. You need to leave room to think, the moment where you examine all the alternatives.

The design process needs heuristics to support the decisions you want to take. Pen and paper sketch can be useful as well as taking notes. Communicating and exchange ideas is also a very useful step as well as doing research. Paying attention to the details. As a maker of data visualization you are the first reader. Lastly, learning what is happening is crucial. Deepen the knowledge about the topic you are dealing with.

Data viz can be very different but this modular organization should work for every data viz. The modular organization should work for every data viz.

**A****B****C**

Ritaglio schermata acquisito: 05/04/2021 19:32



Ritaglio schermata acquisito: 05/04/2021 19:32

Data viz has very different forms but a plan accomodate very different aspects-

Stage 1 -&gt; formulate your brief

identify the context in which your work will be undertaken and then define its aims: who/what/why/where/when/how

*brief* represents the set of expectations and capture all the relevant information about the project

in dataviz, this means establishing the *context* and the *vision*

### context

defining the origin curiosity  
identifying circumstances  
defining the purpose

### vision

- planning a purpose map
- harnessing ideas

Ritaglio schermata acquisito: 05/04/2021 19:34

Brief as a set of all the expectation that we have.

Context -> why I want to study this particular topic.

Context -> define your origin curiosity -> it can arise from personal intrigue ( specific question you raised); stakeholder intrigue ( specific question someone else raised, no anticipation of interest); audience intrigue ( combination of knowing what will be needed and anticipating what could be needed); anticipate intrigue ( audience did not explicitly ask for, but is perceived to be relevant); potential intrigue (opportunity of exploration without exactly knowing where to go).

Identify your project's circumstances: pointing out all the requirements and restrictions that are inherited by you, imposed on you or determined by you.

People -> stakeholder: what impact will they have on the work? Audience: who are your viewers?

Constraints -> pressure: how much time do you have and what milestones along the realisation path? Rules: layout/size restrictions, style guidelines, functional restrictions.

Deliverables-> quantity: how many things should be made? Are they similar or very different?

Format: digital, print, physical? Poster, website, app?

Consumption-> frequency: how often this project will be repeated? What is the trade-off between effort and lifetime? Is it worth trying to automatise part of the process? Setting: how the work will be consumed? Four types of setting: boardroom - limited time/patience/tolerance, immediate insights; coffee shop - more relaxed setting, time to familiarise; cockpit- instrumentation nature of dataviz, or reference map, many levels, operational; prop - dataviz as a supporting visual device for understanding facilitation.

Resources-> skills: what competences are available among those who will work at the project= technologies: what tools/apps/programming options will you use?

Define your project's purpose -> what is it you specifically hope to accomplish with your visualization?

Do you want to be impactive? Are you attempting to shock or inspire or persuade? Do you want just to inform or actively seek to make a difference?

No single type of data viz will be capable of delivering an experience whereby all flavours of understanding are facilitated?

Do not define your purpose before establishing your trigger curiosity, or you will force data to do your talking.

### PURPOSE MAP



Ritaglio schermata acquisito: 05/04/2021 20:14

Three main level of experience and two different tones. Each category as itself subcategories



Ritaglio schermata acquisito: 05/04/2021 20:15

Experience means how will the data viz practically operate as a means of communication?

Through what functional experience will understanding be achieved by the viewer?

Explanatory first type of experience -> providing the viewer with a visual portrayal of the subject's data; taking responsibility to bring key insights at the surface, rather than leaving it to the viewer; attempting to assist with the viewer's progress of understanding as much as possible, drawing out the meaning of the data.

Explanatory -> annotate and describe -> mildest form of explanatory experience; including simple annotation devices assisting in the interpretation; use color to separate different features; use captions to outline key messages; self-explanatory chart, without the need for in-person explanation.

**experience**

explanatory      annotate & describe

### Black Students Are Underrepresented On Campus

Black enrollment at public research universities vs. black college-age state population, 2013



\*The CUNY Graduate Center primarily grants doctorates but has a small undergraduate population.

FIVETHIRTYEIGHT

SOURCE: U.S. DEPARTMENT OF EDUCATION

Ritaglio schermata acquisito: 05/04/2021 20:21

The main target have different colors; we have a driver, a trend line; the title is totally self explain.

An alternative can be sequence and drama -> a more intense form of explanatory experience; use of a narrative structured around a sequence of information; extensive use of animation; storytelling.



Ritaglio schermata acquisito: 05/04/2021 20:23

(WATCH THE VIDEO)

The third way is manipulate and interrogate -> highlighting/filtering categories of interests; change data parameters; switch between different views; hover over different features to reveal detailed annotations; suitable for audience with foundation knowledge of the subject



Ritaglio schermata acquisito: 05/04/2021 20:31

No caption; no indication of significance/insignificance.

No good/bad values

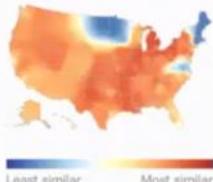
Viewer can decide his own window of analysis.

Then we have participate and contribute -> far deeper explanatory experience; greater control and deeper array of features; contributing one's own data; audience drawn to challenges (active way)

# How Y'all, Youse and You Guys Talk

What does the way you speak say about where you're from?  
Answer all the questions below to see your personal dialect map.

YOUR LAST ANSWER  
How do you pronounce  
aunt?  
to sound like ant



Least similar      Most similar

QUESTION 13 OF 25

How do you pronounce crayon?

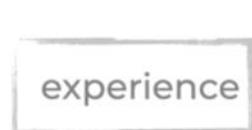
- with one syllable—rhymes with man
- with two syllables—sounds like cray-ahn
- with two syllables, where the second syllable rhymes with dawn
- sounds like crown
- other

Next >

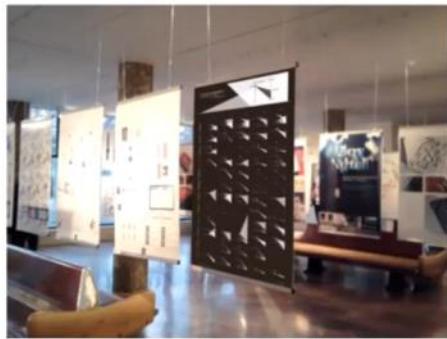
Ritaglio schermata acquisito: 05/04/2021 20:34

Audience is involved.

Exhibitionary -> the viewer has to do the work to interpret the meaning; no explanatory qualities/no scope for interrogatory explorations; the viewer needs to know the content and the context; a supporting a written article or report; an inconvenient truth/gapminder.



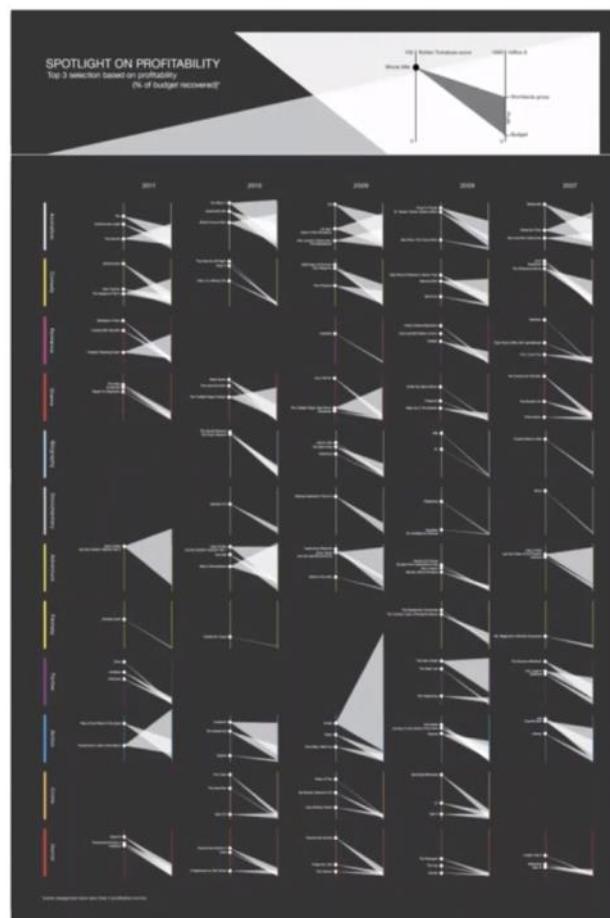
exhibitory



no exploration

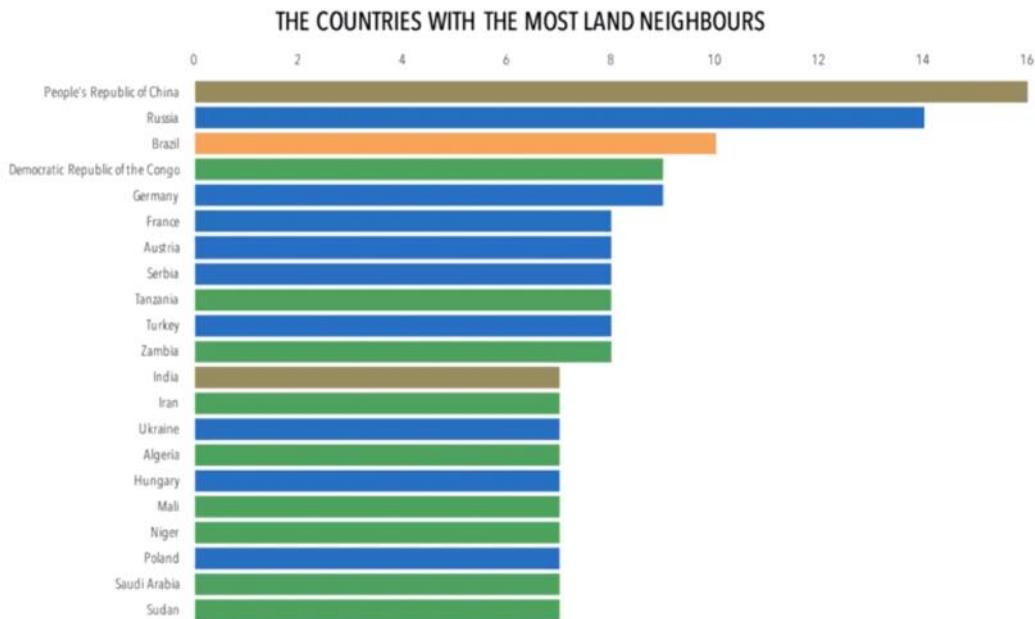
no interaction

no explanations



Ritaglio schermata acquisito: 05/04/2021 20:38

The TONE -> a continuum with subtle and very subjective variation between the two choices of reading and feeling.  
Tone -> reading -> very neutral -> optimising the ease of the viewer's estimation; efficacy of perceiving the data; facilitate understanding at high degree of precision and detail; no need to seduce the audience; no need of visual simulation; analytical, pragmatical and no frills.



Source: [https://en.wikipedia.org/wiki/List\\_of\\_countries\\_and\\_territories\\_by\\_land\\_borders](https://en.wikipedia.org/wiki/List_of_countries_and_territories_by_land_borders) (as at January 2016)

Note: Minimum 7 neighbouring countries. Colours group continents, values sorted by largest border lengths. France's figure does not include French overseas departments, collectivities, and territories.

why would you need to build something different?

Ritaglio schermata acquisito: 05/04/2021 20:41

Tone -> feeling: visual look that attracts and also informs; audience need to be engaged; subject stirring strong emotions; encapsulating emotions in the display; it is a manipulation of certain degree.



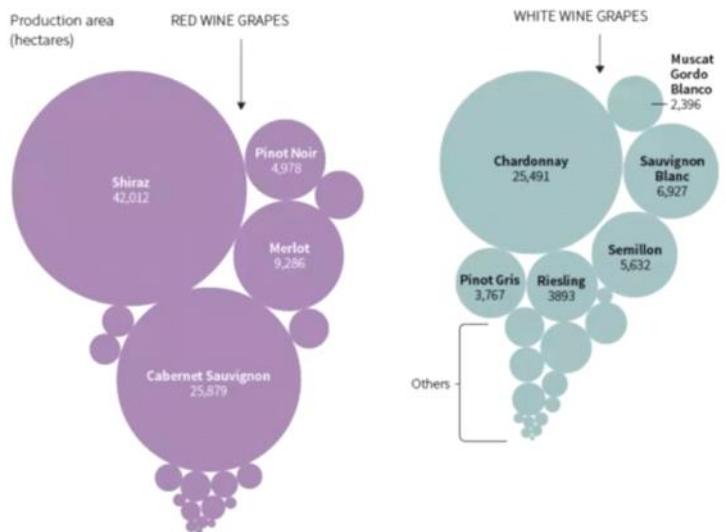
Ritaglio schermata acquisito: 05/04/2021 20:43

Punctuate the audience to understand how unfair is the distribution.

Tone -> harnessing ideas -> building from the earliest seeds of any idea about the solution -> mental visualization; keywords; sketching; research and inspiration; limitation of author's idea; limitation of others' ideas.

Mental visualization -> system one kahnemanOs model of thought

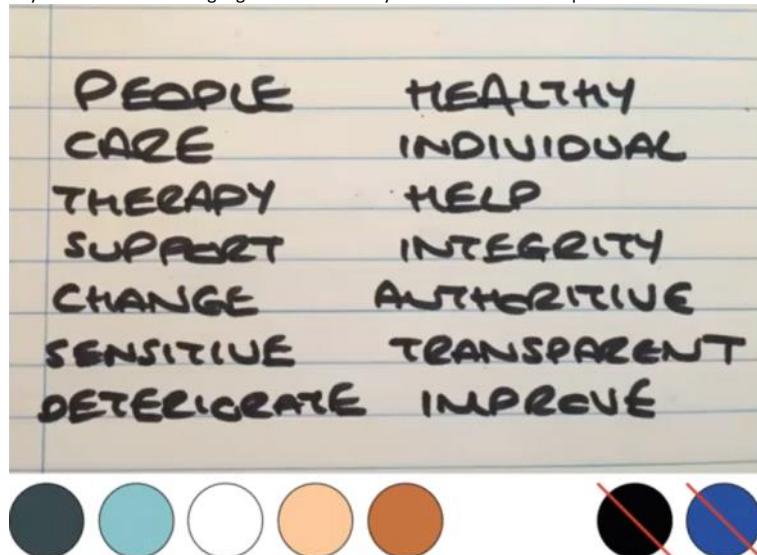
### Top grape varieties grown



the mental impression  
forming in the mind  
when proposed with the  
challenge

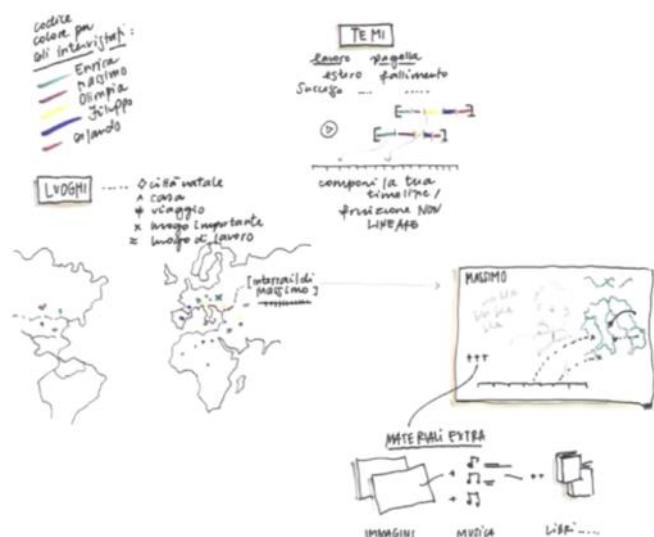
Ritaglio schermata acquisito: 05/04/2021 20:46

Keywords -> terms of language that instinctively connected with the topic



Ritaglio schermata acquisito: 05/04/2021 20:47

Sketching -> freedom and speed when extracting ideas from the mind



Ritaglio schermata acquisito: 05/04/2021 20:48  
Harnessing ideas

► research & inspiration  
consider different sources of imagery:  
colours/patterns/shapes/metaphors

► limitation of author's ideas  
plagiarism, copying and stealing uncredited ideas

► limitation of others' ideas  
author's responsibility to lead on the creative process

➢ data acquisition



➢ data examination



➢ data transformation



➢ data exploration



Ritaglio schermata acquisito: 05/04/2021 20:49

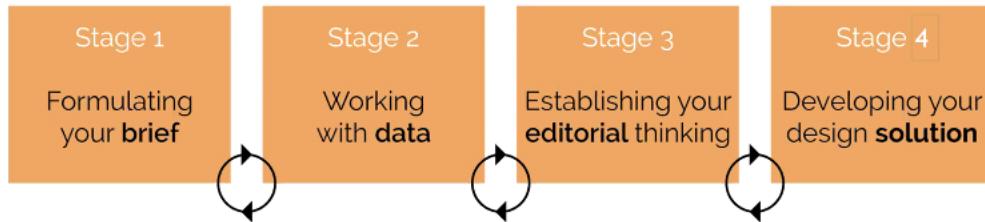
# Dataviz workflow 2

mercoledì 7 aprile 2021 11:29

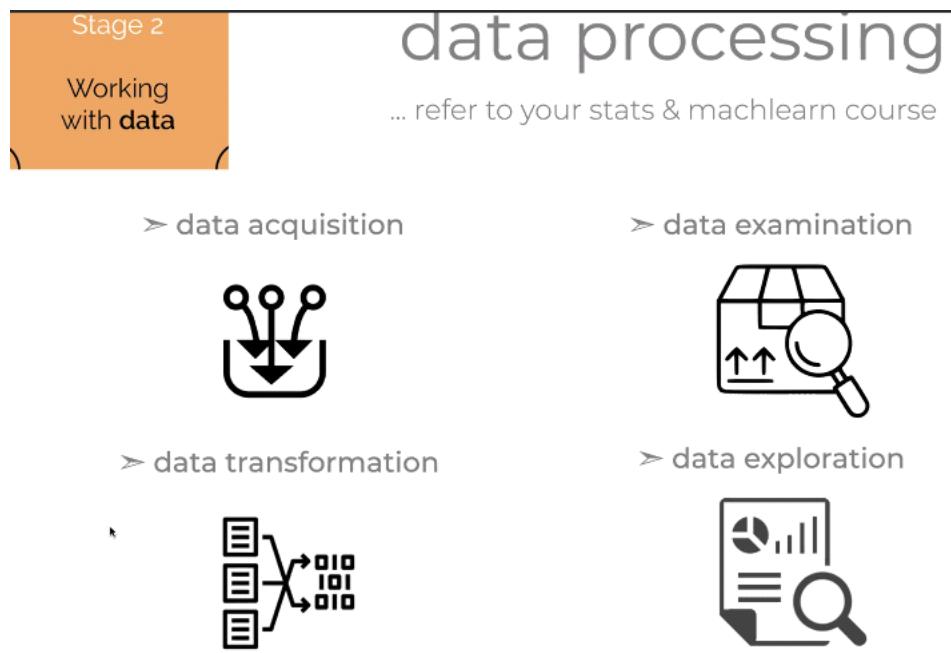
*pareto's principle*

20% of decisions have effect on the 80% of the result.

the most reliable approach is to follow a design process



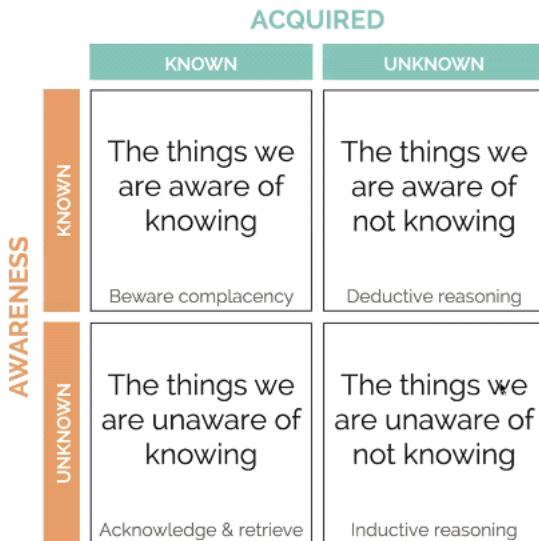
Ritaglio schermata acquisito: 07/04/2021 11:31



Ritaglio schermata acquisito: 07/04/2021 11:34

Render the data in a way that is useful for your purpose.  
We extract what the data have to say according to the research topic.

# data processing



Ritaglio schermata acquisito: 07/04/2021 11:35

## EDITORIAL CHOICE

Editorial choice -> many viable perspectives offered by data you will focus on.

This can be parallelized to what we do when we try to take a picture.

What are the important characteristic -> the angle; the framing ( what are we leaving out); the focus, the particular aspect we want to highlight.

The ANGLE -> choose viewpoints in the analysis, choose the dimension to break down the subject ( when too complex); what should we ask: are the chosen dimension relevant? Are enough to express what we want to communicate?; why did we choose this particular pov? Why this pov is worth showing to the audience?

Different audience can have different perception.

You are the one who choose the view point you have to explain why it is relevant.

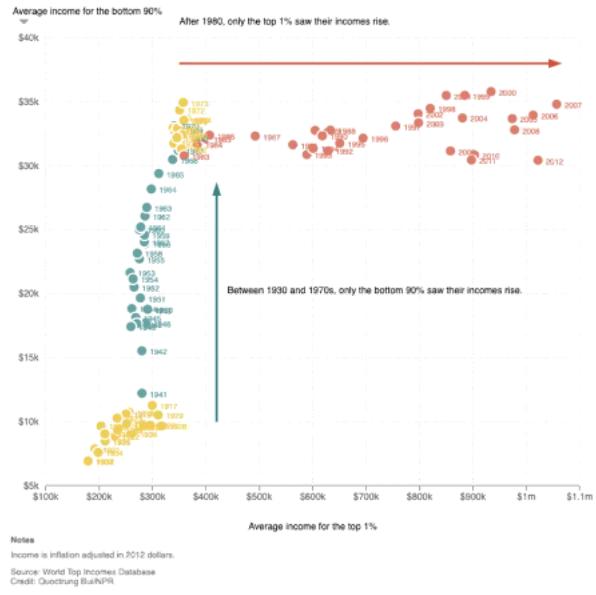
Being enough -> to fully describe a phenomenon is enough one pov? Are they representative? We should consider different angles across space and time; too many pov are not a good choice.

The FRAMING -> filtering which data to include or to exclude; remove all the unnecessary clutter; how much content the audience can process- not too small and not too big. It is hard to balance choice.

The FOCUS -> emphasising what is more important. Provide visual hierarchy. Choosing fore/mid/back ground.

# editorial choices

example #1 the rise & fall of u.s. inequality



Ritaglio schermata acquisito: 07/04/2021 11:46

## ► angle

relation between 2 measures

*relevant*: key indicator of wealth distribution

*sufficient*: to support the article

## ► framing

space: u.s. / time: 1917-2012

## ► focus

highlighted by colours

Ritaglio schermata acquisito: 07/04/2021 11:48

### Stage 3

#### Establishing your editorial thinking

► angle  
how quantitative values broken down by category changed over the year?"

relevant: p. manning setting a new record

sufficient: not by its own

► framing  
1930 - 19 oct 2014  
≥ 30 touchdown passes

# editorial choices

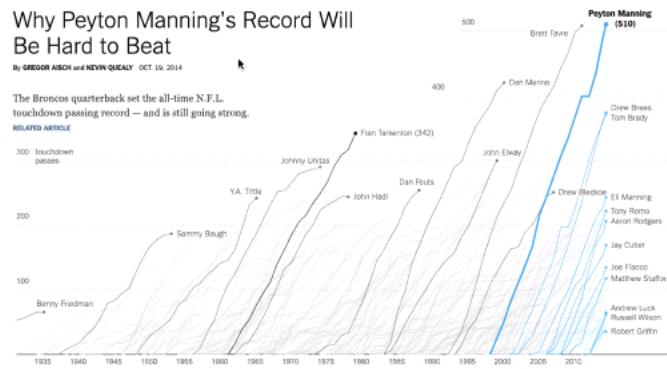
example #2 why manning's record will be hard to beat

### Why Peyton Manning's Record Will Be Hard to Beat

By GREGOR AISCH and KEVIN QUIGLEY OCT. 19, 2014

The Broncos quarterback set the all-time N.F.L. touchdown passing record — and is still going strong.

RELATED ARTICLE



### ► focus

record holder vs. other players  
previous record holder  
careers evidenced by mouseover

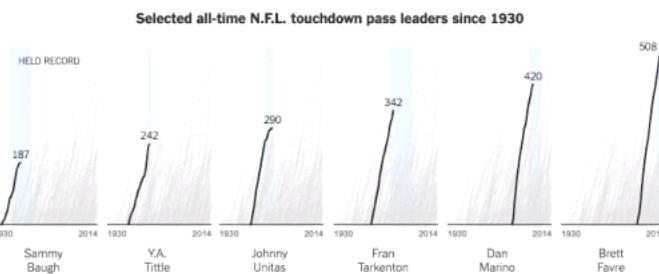
Ritaglio schermata acquisito: 07/04/2021 11:49

► angle  
"how quantitative values broken down by category changed over the year?"

relevant: p. manning setting a new record

sufficient: not by its own

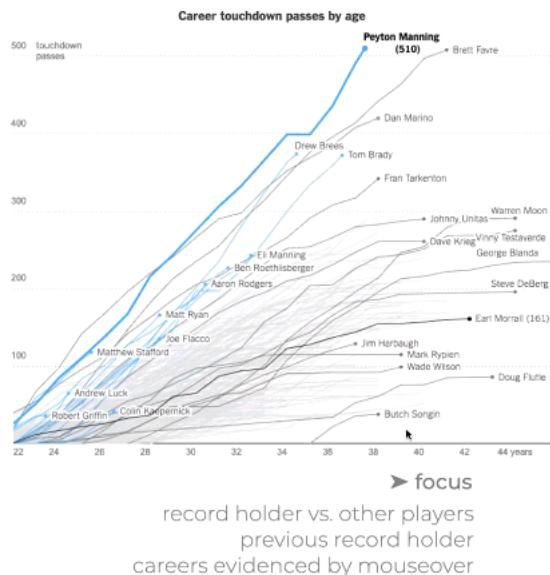
► framing  
1930 - 19 oct 2014  
≥ 30 touchdown passes



Ritaglio schermata acquisito: 07/04/2021 11:53

This is an additional plot that should help the reader to confront data.  
We have also a background band that highlight the performance of single player.

example #2 why manning's record will be hard to beat



Ritaglio schermata acquisito: 07/04/2021 11:54

Third graph but with different x-axes -> the age of the player.

## ➤ data representation

+annotations  
+colors  
+interactivity  
+composition

Ritaglio schermata acquisito: 07/04/2021 11:56

Decoding elements of the graph in the perception moment.

The interpretation moment is still decoding but we are interested in the meaning of the graph.

## reading a graph

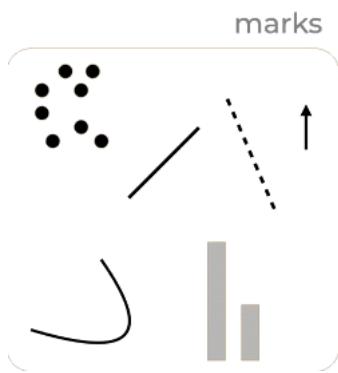
*dataviz perception*  
consists in the **decoding** of  
the elements of a graph:  
• shapes  
• sizes  
• positions  
• colours

*dataviz interpretation*  
consists in the **decoding** of  
the meaning of a graph,  
i.e. making sense of the  
overall graphic  
construction.

chart = visual encoding + chart apparatus

Ritaglio schermata acquisito: 07/04/2021 12:00

Visual encoding -> mostly characterized by marks; glyphs and similar



Ritaglio schermata acquisito: 07/04/2021 12:00

Annotations are -> quantity, size, position, colour.

ACTOR	GENDER	YEARS SINCE FIRST MOVIE		
Harrison Ford	Male	43	Harrison Ford	43
Meryl Streep	Female	38	Meryl Streep	38
Michael Douglas	Male	37	Michael Douglas	37
Arnold Schwarzenegger	Male	34	Arnold Schwarzenegger	34
Nicole Kidman	Female	30	Nicole Kidman	30
Sandra Bullock	Female	24	Sandra Bullock	24

Ritaglio schermata acquisito: 07/04/2021 12:01

## marks

mark	example	description
point		<ul style="list-style-type: none"> <li>no spatial variations</li> <li>quantity as position on a scale</li> </ul>
line		<ul style="list-style-type: none"> <li>linear spatial dimension</li> <li>quantity as variation in size</li> </ul>
area		<ul style="list-style-type: none"> <li>quadratic spatial dimension</li> <li>quantity as variation in size &amp; position</li> </ul>
form		<ul style="list-style-type: none"> <li>cubic spatial dimension</li> <li>quantity as variation in volume</li> </ul>

Ritaglio schermata acquisito: 07/04/2021 12:02

attribute	example	description
position		<ul style="list-style-type: none"> <li>· quantity as position on a scale</li> </ul>
size		<ul style="list-style-type: none"> <li>· quantity as variation in size</li> </ul>
angle / slope		<ul style="list-style-type: none"> <li>· quantity as variation of angle</li> <li>· quantity as different slope</li> </ul>
quantity		<ul style="list-style-type: none"> <li>· quantity as repeated set of point marks</li> </ul>

Ritaglio schermata acquisito: 07/04/2021 12:04

Position -> setting a point in a graduate scale tells you sth about the quantity you are representing.  
Variation is size can be easily relate to variation in the size.

attribute	example	description
colour saturation		<ul style="list-style-type: none"> <li>· quantity as saturation</li> </ul>
colour lightness		<ul style="list-style-type: none"> <li>· quantity as brightness</li> </ul>
pattern		<ul style="list-style-type: none"> <li>· quantity as density / shape of pattern</li> </ul>
motion		<ul style="list-style-type: none"> <li>· movement as binary indicator to draw focus or to represent quantitative scale ramp</li> </ul>

Ritaglio schermata acquisito: 07/04/2021 12:06

attribute	example	description
symbol / shape		<ul style="list-style-type: none"> <li>· symbols as categorical association</li> </ul>
colour hue		<ul style="list-style-type: none"> <li>· quantity as brightness</li> </ul>
connection / edge		<ul style="list-style-type: none"> <li>· relationship as connection</li> </ul>
containment		<ul style="list-style-type: none"> <li>· grouping relationship as containment</li> </ul>

Ritaglio schermata acquisito: 07/04/2021 12:07

How many choices do we have in this part of the process?  
We have different chart types and approaches.

**bottom-up:** from visual encoding to chart type

**top-down:** choosing a chart type first

a whole world other than the classic 3: bar/pie/line

Ritaglio schermata acquisito: 07/04/2021 12:10

### the charts classification

<b>categorical</b>	comparing categories and distributions of quantitative values
<b>hierarchical</b>	charting part-to-whole relationships & hierarchies
<b>relational</b>	graphing relationships to explore correlations & connections
<b>temporal</b>	showing trends & activities over time
<b>spatial</b>	mapping spatial patterns through overlays & distortions

Ritaglio schermata acquisito: 07/04/2021 12:11

*caveat: small multiples not included - they are not a different chart on their own, rather an editorial thinking solution*

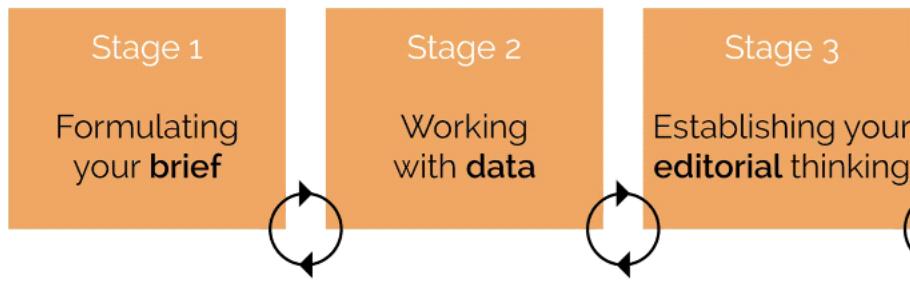
Ritaglio schermata acquisito: 07/04/2021 12:12



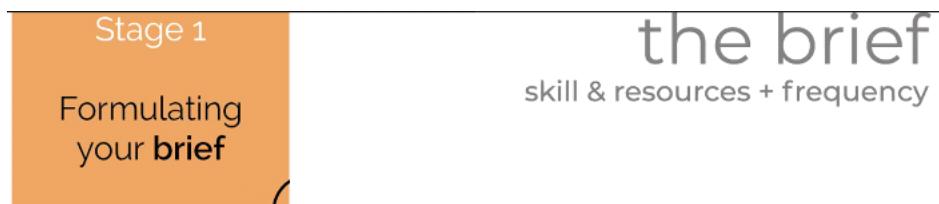
Ritaglio schermata acquisito: 07/04/2021 12:13

Paying attention to all the types of chart and not only the basics ones.

INFLUENCING FACTORS (for design solution)



Ritaglio schermata acquisito: 07/04/2021 12:14



which graph can you actually make and how efficiently?

expressiveness

- maximum expressiveness — can create **any** combination of mark & attribute encoding
- limited expressiveness — limited scope, need for workaround, but quick&simple charting

Ritaglio schermata acquisito: 07/04/2021 12:14

**Influencing factors can also concern the purpose of my brief.**

should you represent your data in a chart form?

will it add any value, new insights, greater perceptual efficiency  
w.r.t to non-visualised form?

will portraying your data in an elegant table actually offer a  
more suitable solution?

maybe an information-based solution (imagery, text, video,  
photo) would work better?

Ritaglio schermata acquisito: 07/04/2021 12:15

**The purpose map -> the tone**

*reading of the data or feeling of the data?*

precise and accurate perception of values  
or sense-making of the gist of values

trade-off between emotional qualities and  
perceptual efficiency?

The purpose map ( different ways of encoding data might offer varying degrees of effectiveness in perception -> subjective factors)

General ranking explaining which attributes facilitate the highest degree of perceptual accuracy.

#### Automating the Design of Graphical Presentations of Relational Information

JOCK MACKINLAY  
Stanford University

ACM Transactions on Graphics, Vol. 5, No. 2, April 1986, Pages 110-141.

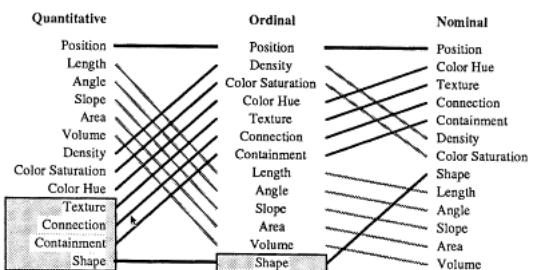
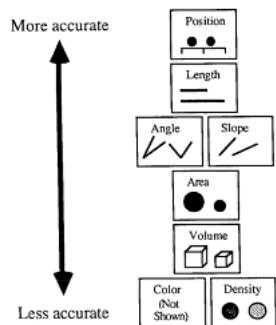


Fig. 15. Ranking of perceptual tasks. The tasks shown in the gray boxes are not relevant to these types of data.

The ranking may change -> in the plot we show the relationship between one level and the other on the same quantitative indicator.

For categorical elements -> the texture is very important.

	Nominal	Ordinal	Quantitative
Size	-	•	•
Saturation	-	•	•
Texture	•	•	
Color	•	•	
Orientation	•		
Shape	•		

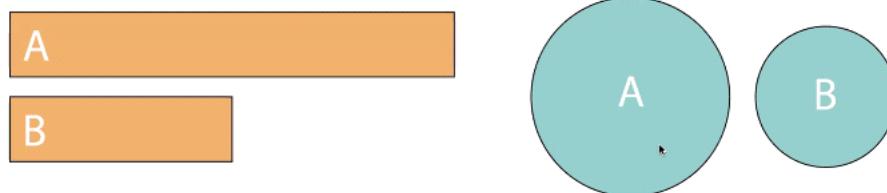
In this table we see what is common and what is not.



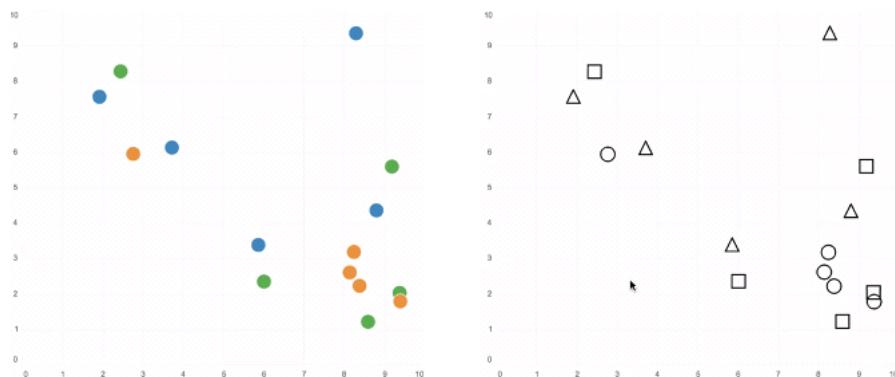
summary: some attributes make it easier  
and others make it harder to judge  
accurately the values being portrayed

judging variations in lines is far more precise than in areas, even worse in circles: this is due to the 2D measure and the shape (and in the previous table, *length* is ranked higher than *area*)

Studying measures and shape are important in some instances  
*what's the ratio B/A?*



We perceive differently the ration between these two graphs. It is much easier when we compare the bar chart than the area of the two circle.



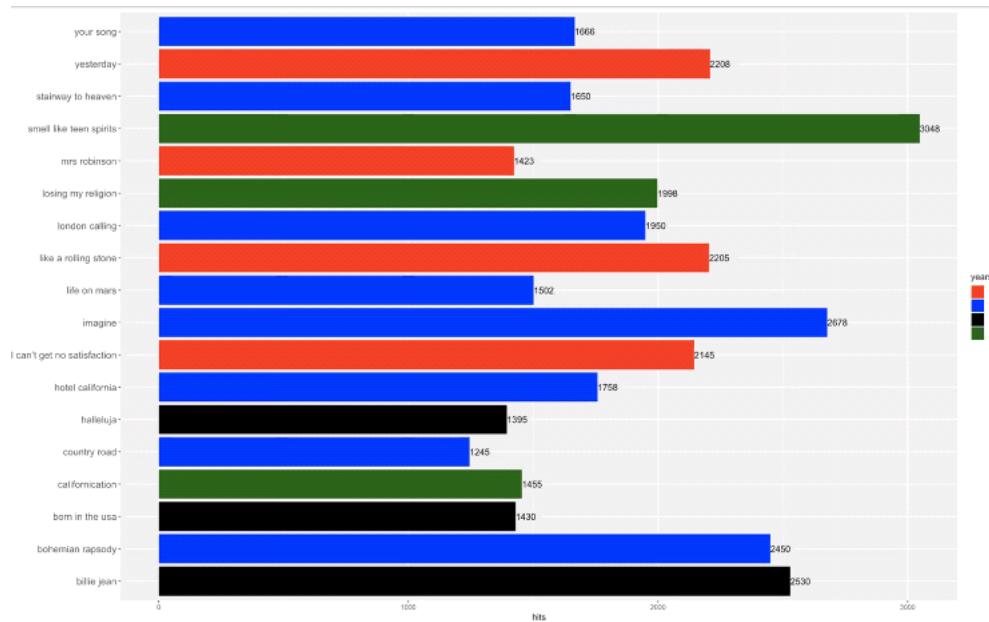
Here we have nominal data. On the left we have different hue(color) and the other with shapes. The preference is subjective.

Let's consider a music website poll

song	yr	hits
smell like teen spirits	90	3048
imagine	70	2678
billie jean	80	2530
bohemian rapsody	70	2450
yesterday	60	2208
like a rolling stone	60	2205
I can't get no satisfaction	60	2145
losing my religion	90	1998
london calling	70	1950
hotel california	70	1758
your song	70	1666
stairway to heaven	70	1650
life on mars	70	1502
californication	90	1455
born in the usa	80	1430
mrs robinson	60	1423
halleluja	80	1395
country road	70	1245

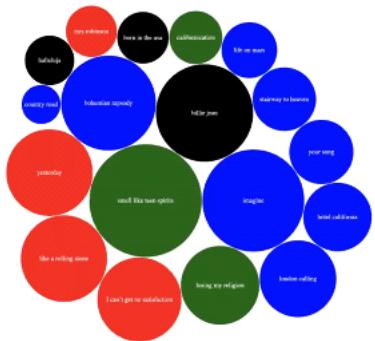
Ritaglio schermata acquisito: 07/04/2021 12:27

What is the purpose of this representation? Do we need a precise perception? We need to properly separate the data?



Ritaglio schermata acquisito: 07/04/2021 12:29

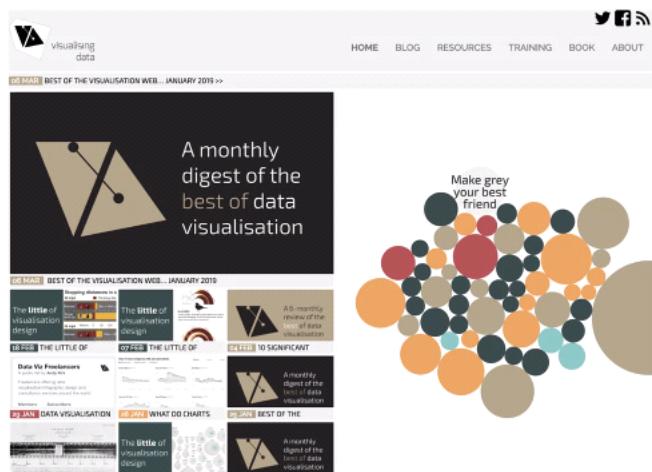
If precise perception is the aim this is the most suitable way of doing it.  
or aesthetically give the sense of data?



Ritaglio schermata acquisito: 07/04/2021 12:29

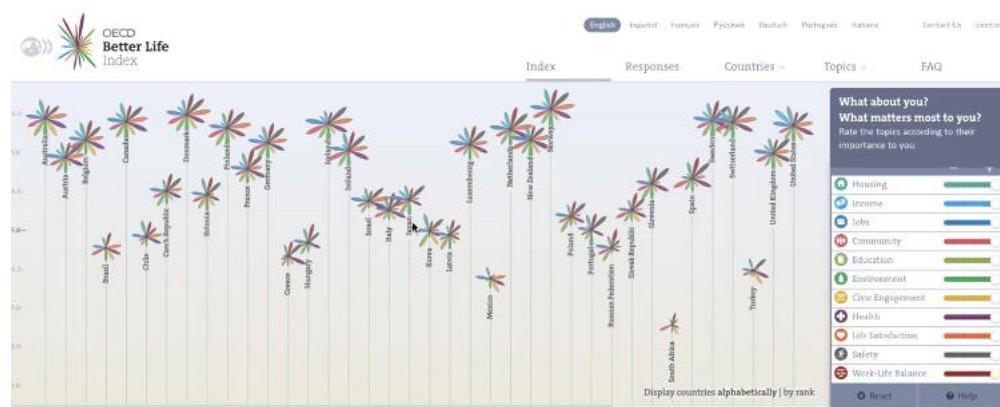
If the purpose is give the sense of the data and it is more aesthetically this second option is preferred.

We have no legend, no figure, bad proportion, trasformed data.



Ritaglio schermata acquisito: 07/04/2021 12:31

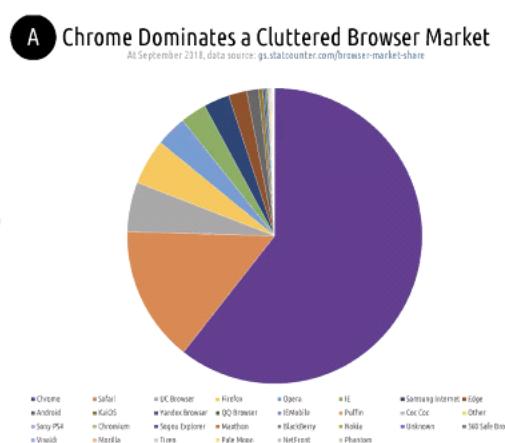
Data processing -> only certain types of data can fit into certain types, and vice versa.



*inherent meaning: flower & blossoming metaphor conveys idea of better life*

Ritaglio schermata acquisito: 07/04/2021 12:32

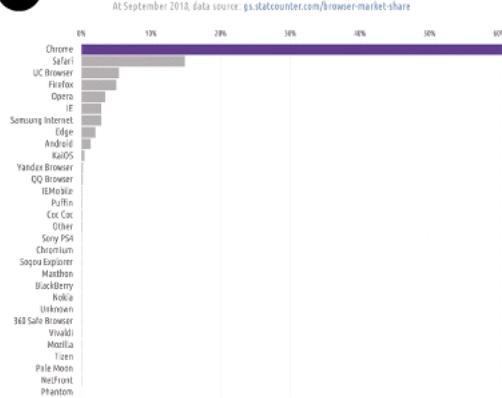
This plot produced by OECD, represent an index of well being called better life.



Ritaglio schermata acquisito: 07/04/2021 12:35

This pie chart gives a clear idea of the content.

## B Chrome Dominates a Cluttered Browser Market

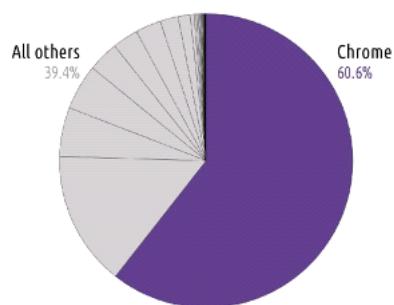


Ritaglio schermata acquisito: 07/04/2021 12:36

Comparing the lenght is much easier than comparing the area. So preferring a bar plot could be a good solution in order to help the reader.

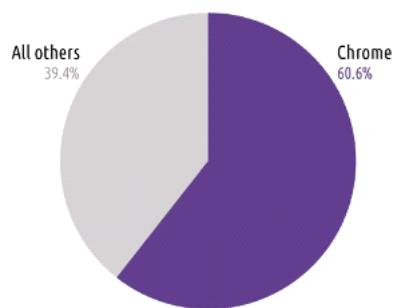
In this way you also need just two color.

## C Chrome Dominates a Cluttered Browser Market

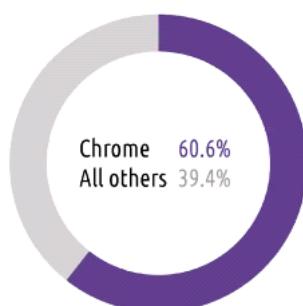


Ritaglio schermata acquisito: 07/04/2021 12:37

We can choose this other type of representation. It is easier to interpret we have only two dimension and we put number as notation. Segments are not needed.



Ritaglio schermata acquisito: 07/04/2021 12:38



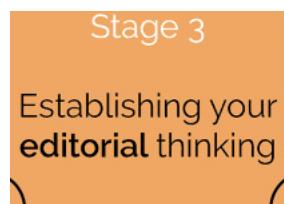
This is an even better solution due to its self contained characteristic.

## ANGLE

choosing the angles of analysis dictates which chart type might be most relevant following the *charts* taxonomy

treat every representation challenge on its own merits: having spatial data does not mean you must use a map

if regional/spatial information are not essential, the map composition may hinder the analysis rather than helping it

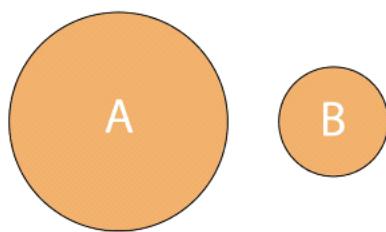


## trustworthy design

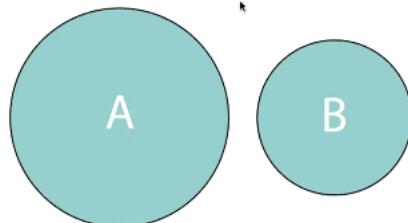
avoiding deception

geometric distortions

Variation in **diameter**

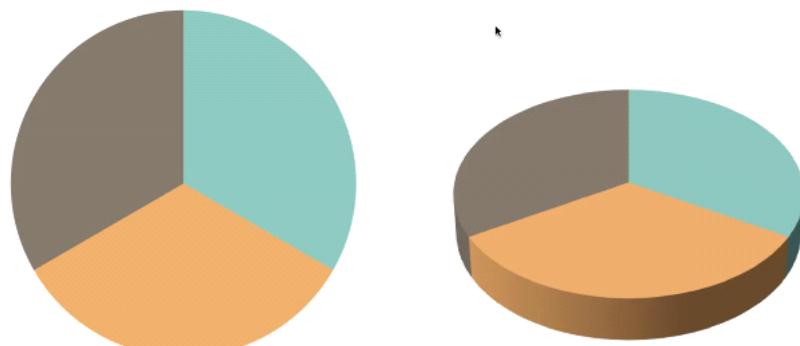


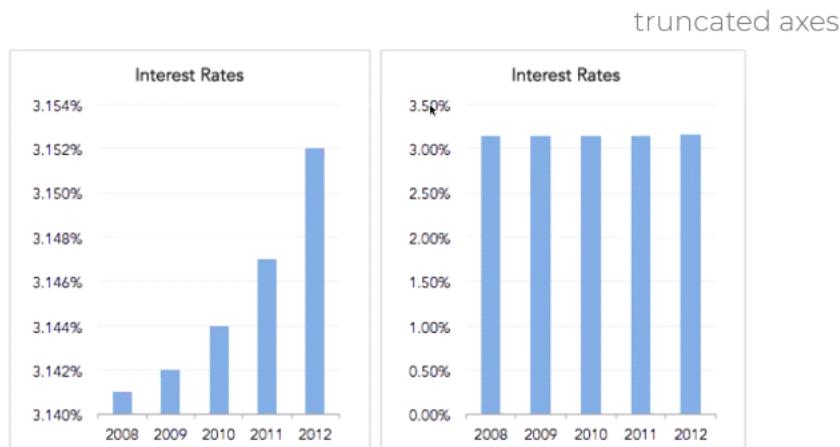
Variation in **area**



The feeling you are providing to the audience is quite different.

3d decorative distortions





### ACCESSIBLE DESIGN -> encoded overlays

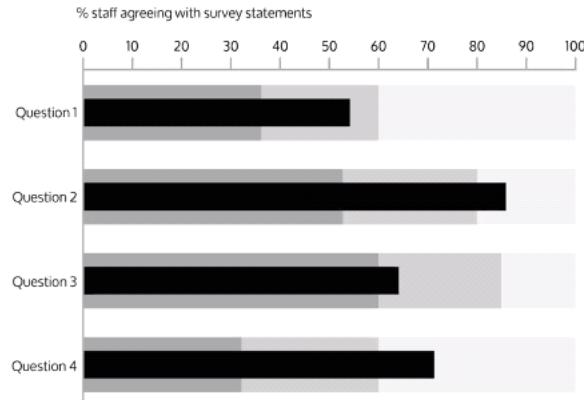
incorporation of additional details:

to explain further the context of values

to amplify the interpretation of good/bad, normal/exceptional

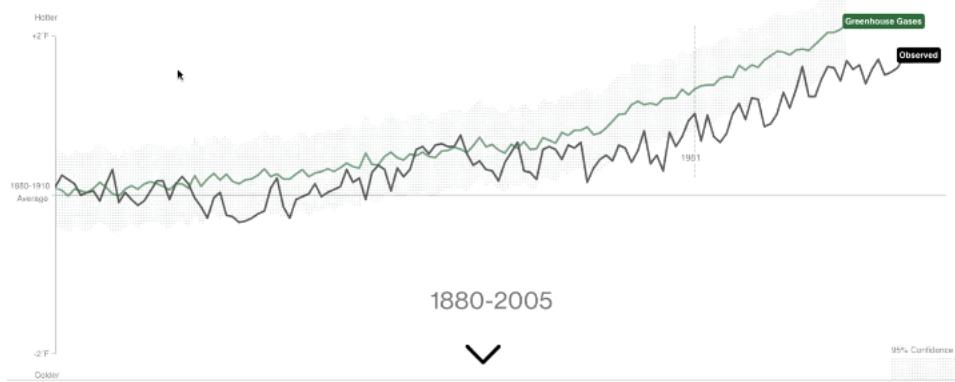
they are not just annotations: they represent data values and require encoding choices

### Bandings



### Barplot embedded with other barplots.

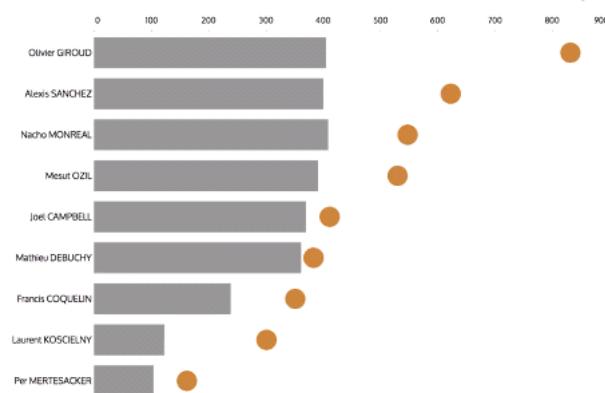
indicate the contrast  
between the main data  
value marks and the  
contextual judgement of  
historic or expected values



Ritaglio schermata acquisito: 07/04/2021 12:47

Contextual overlay showing 95% confidence interval.

SPRINT DISTANCES: Arsenal vs. Tottenham (8th Nov 2015) compared to Season Best

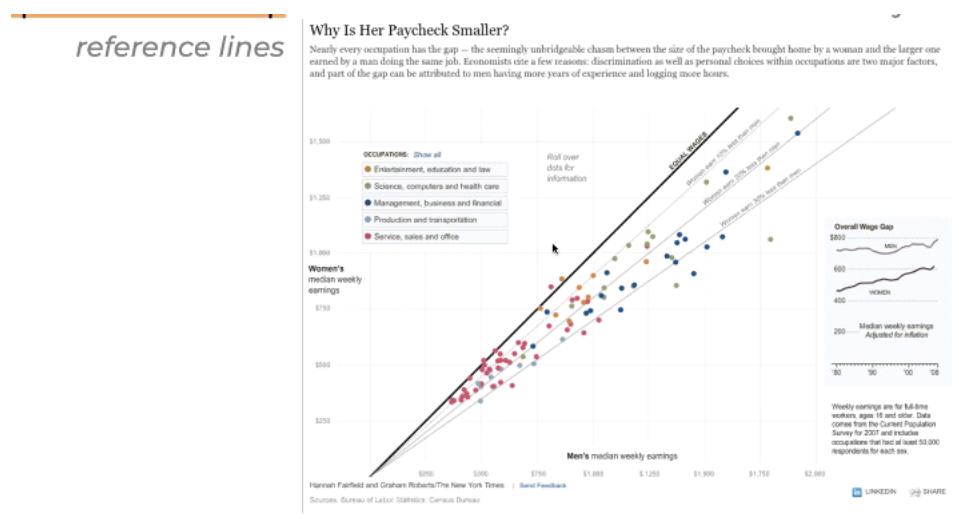


Ritaglio schermata acquisito: 07/04/2021 12:49

Another example of encoded overlays

Additional points comparison versus a maximum values.

Different markers are a very good driver for the reader.



Ritaglio schermata acquisito: 07/04/2021 12:50

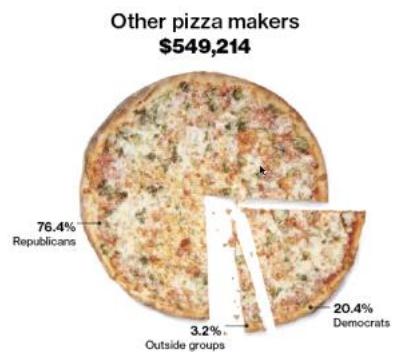
Lines + scatterplot. The encoded overlays are the reference lines that also mark some milestone.

in line charts and scatterplots, reference lines direct the eye towards calculated trends, constants, averages, correlations or best fits.

Ritaglio schermata acquisito: 07/04/2021 12:52

### ELEGANT DESIGN -> Visual appeal

Sometimes there might be scope in squeezing out an extra sense of stylistic association between visual impact and the context



Ritaglio schermata acquisito: 07/04/2021 12:53

Pizza pie -> could be useful to make people remember the data.  
We can use irregular shape and then adding percentage to make it clearer.

# Composition

mercoledì 14 aprile 2021 11:34

Composition as to set up the elements of data viz to maximize the effect of your representation.

final layer of the design anatomy

making careful decisions about the physical attributes of an relationships between every visual property to ensure the optimum readability and meaning of the overall project

Ritaglio schermata acquisito: 14/04/2021 11:35

There are two moment in which the composition step is organized.

The first is the project composition -> defining the layout and hierarchy of the project

2nd -> Chart Composition -> defining the shape, size, layout choices for all the components of the chart

PROJECT COMPOSITION -> how to lay you ans size all the visual components?

Although established conventions do exist, this is usually an iterative process towards what feels like an optimal layout.

Hierarchy of content can be reached thorough careful choices in relative position and relative variation in size ( and variation in color for significance).



Ritaglio schermata acquisito: 14/04/2021 11:39

Let's consider this project.

This represent an historical building in hong kong. The panel wants to show each step of the life of the building.

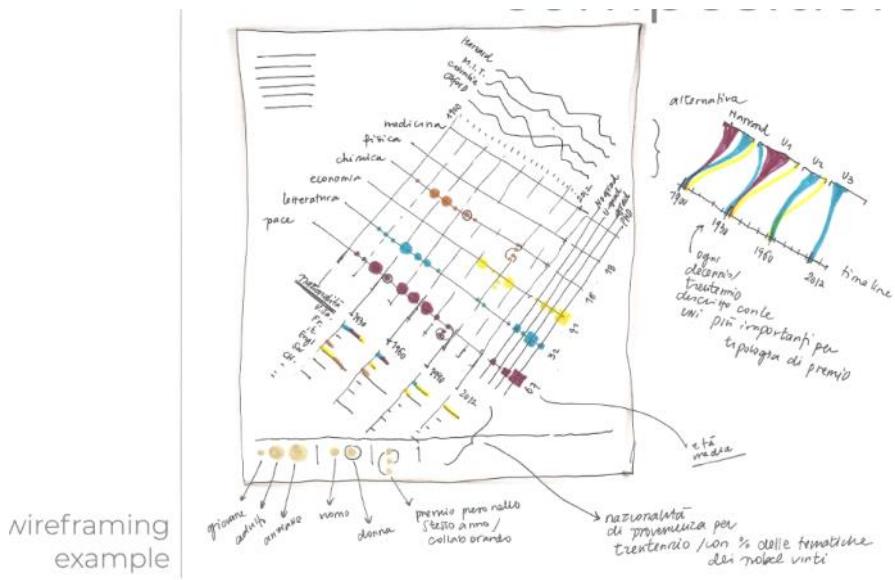
From the pov of project composition we see the organisation based on a primary local point ( the section of the construction). This main focus is surrounded by small components that help the reader to be orientated.

At the bottom there are small supplementary representation.

The final result is something that can be considered a good example of dataviz.

The steps of component composition:

Wireframing



Ritaglio schermata acquisito: 14/04/2021 11:43

We can see the main point of the data viz, possible alternatives to show the same message.

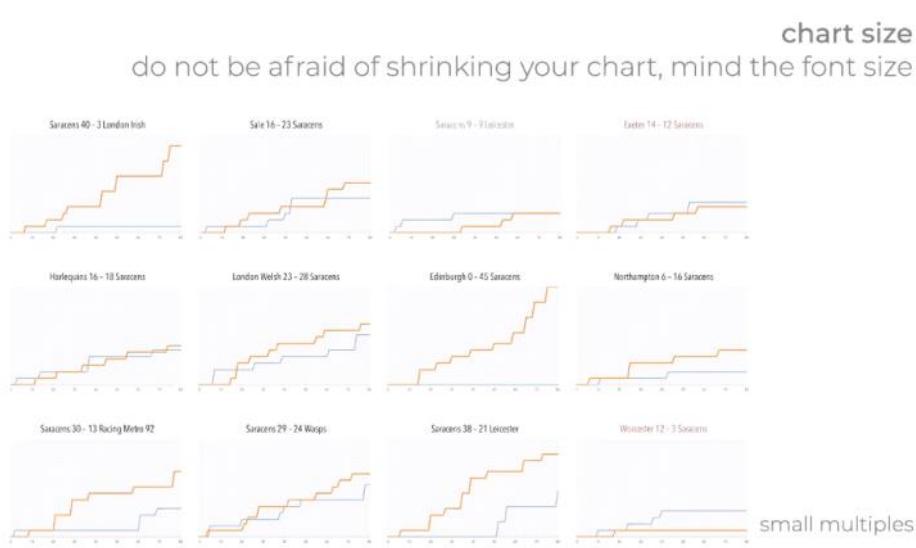
wireframing  
sketching all layout/size across a single-page view,  
including interactive functions

storyboarding  
together with wireframing if the project is multipage

single pages included as cells in the big-picture hierarchy,  
each one with its own wireframe

Ritaglio schermata acquisito: 14/04/2021 11:44

The other step in project composition is Storyboarding.



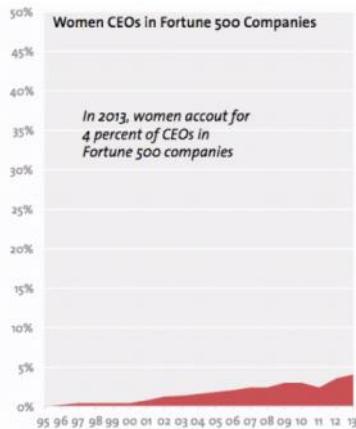
Ritaglio schermata acquisito: 14/04/2021 11:46

The chart is the central element of data viz. Understand how to design the layout is a central moment.  
Understanding the proper size is crucial. Especially if you need to organize multiple information you should not be

scared to shrink. Watchout: small writings can be dangerous.

Another component is the scale.

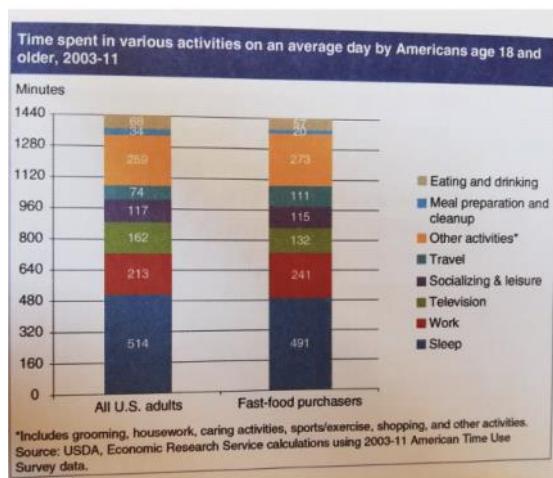
### chart scales even the choice of scale can tell the viewer something



the apparently extreme scale marks the large disproportionality in women ceo's amongst the fortune top-500 companies

Ritaglio schermata acquisito: 14/04/2021 11:48

The graphical effect is that the graph is very EMPTY. In ideal world the number of women as CEO should be around 50%. The author choose a vary impacting way of representing it.

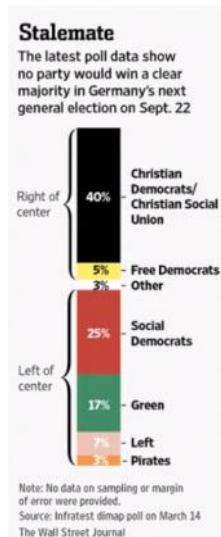


here the wrong scale choice has compromised readability — what's the point of dividing the 1440 minutes of a day in 160-mins intervals?

Ritaglio schermata acquisito: 14/04/2021 11:54

The intervals are really unnatural. Subdivision should be done in 60 minutes. It is not a natural way to represent it in that way. Here there is little attention paid to the details.

Chart orientation -> adding an extra degree of readability; avoid label overlap; remebere the Iraq's bloody toll.

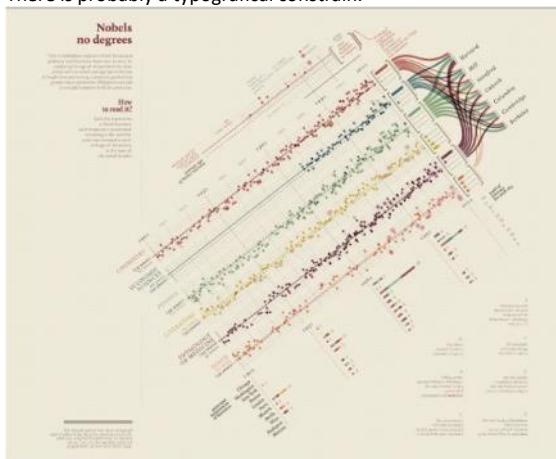


Ritaglio schermata acquisito: 14/04/2021 12:03

Missed opportunity of using the left/right duality.

Representing the bars in horizontal way would have been more effective.

There is probably a typographical constrain.



final version of previous wireframe,  
offering greater room in the page

Ritaglio schermata acquisito: 14/04/2021 12:09

Optimal example of how to optimize the information inside a single chart.

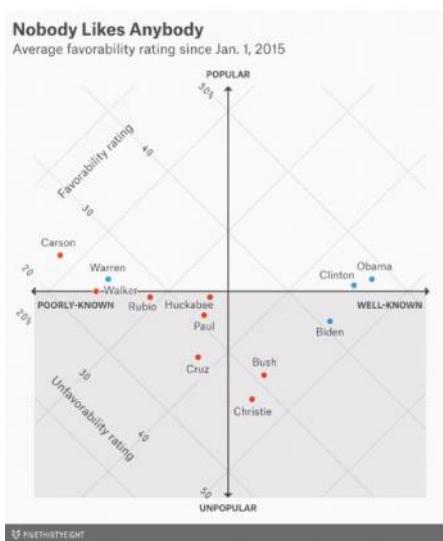


chart orientation  
not only 90 degrees

45 degrees rotated scatterplot with 2x2 overplayed grid to make it easier to observe which values are located in each quadrant.

also emphasising the distinction between location at the top and bottom halves of the chart along the popularity axis, which is the primary focus of analysis

Ritaglio schermata acquisito: 14/04/2021 12:14

There are two set of axes.-> popular/well known; favorability and not

The point is that favorability is quantitative;

While the other two are qualitative.

This choice could be caused by the fact that ordinary people are more likely to understand the qualitative axes.

Another important component is how you SORT your value.

You need to specify an order.

According to which element you sort the values.

latch rule



location



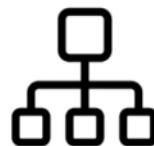
alphabet



time



category



hierarchy

on broadway installation



Ritaglio schermata acquisito: 14/04/2021 12:28

Geographical ordering -> sequencing content according to spatial order -> only if offers the most logical sequence in content readability.

Ritaglio schermata acquisito: 14/04/2021 12:30



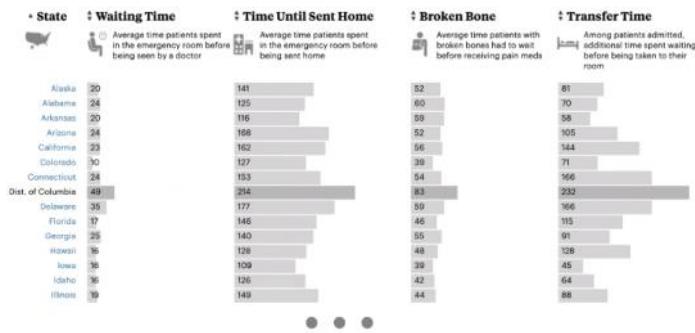
Geographical ordering -> sequencing content according to spatial order -> only if offers the most logical sequence in content readability.



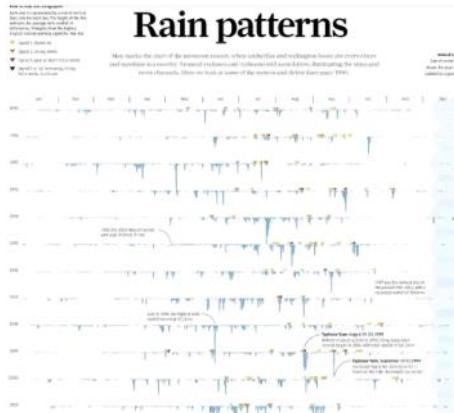
Lot of elements belonging to the street itself.

Alphabetical orientation choice everytime when it helps an efficient lookup and reference.  
Most natural choice for the reader to look for information.

best sorting sequence if you do not want to imply any ranking



Time -> meaningful if helps comparing changes over time.



In the example, the interest is in the seasonality of patterns -> chronological sorting is meaningful.

Categorical -> data organized in categories; by a ranking inherited by their values or unique to the subject.

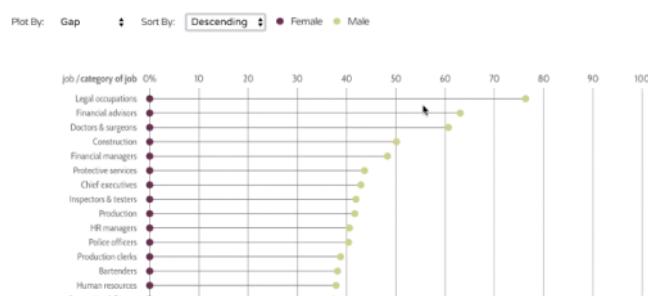
### Outcome status for clients undergoing multiple-sessions of treatment



Ritaglio schermata acquisito: 14/04/2021 12:38

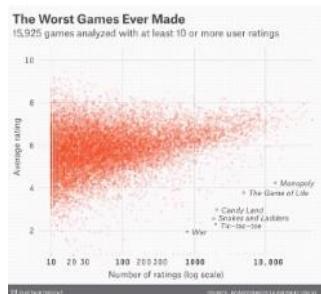
Hierarchical sorting is defined by increasing/decreasing quantities efficient perception of size/distribution/raking.

### Gender Pay Gap US | UK



Ritaglio schermata acquisito: 14/04/2021 12:41

Format -> what is the shape/size of the primary format? How transferable is the solution across different platform?  
Data -> how legitimately fit data into the given canvas?



Ritaglio schermata acquisito: 14/04/2021 12:43

Change scale to fit a square.

Logarithmic scale on the x-axes.

The idea that there is a triangular distribution is evident although the different scales.

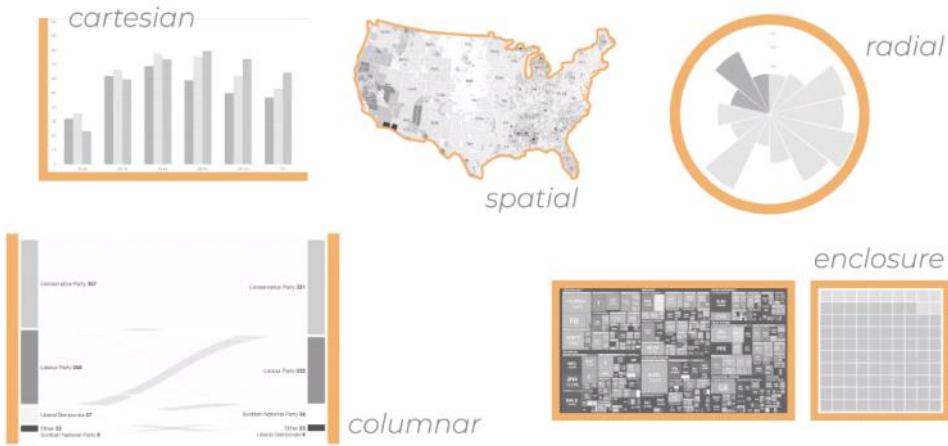
	Month on month inflation % 2000–2008								
	2000	2001	2002	2003	2004	2005	2006	2007	2008
Jan	55	57	116	208	628	133	613	1593	100,580
Feb	48	57	116	220	602	127	782	1729	165,000
Mar	50	55	113	228	583	123	913	2200	355,000
Apr	53	56	114	269	505	129	1092	3714	736,604
May	58	55	122	300	448	144	1193	4530	1,800,000
Jun	59	64	114	364	394	164	1184	7251	
Jul	53	70	123	399	362	254	993	7634	220,000,000
Aug	53	76	135	426	314	265	1204	6592	231,000,000
Sept	62	86	139	455	251	359	1023	7892	
Oct	66	97	144	525	209	411	1070	14840	
Nov	56	103	175	619	149	502	1098	26470	
Dec	55	112	198	598	132	585	1281	66000	

Ritaglio schermata acquisito: 14/04/2021 12:46

Tabular data.

Although they are numbers -> there is also a visual effect due to the last column. All the other columns are in the same range, the values look at the same dimension, they are somehow comparable. The last column uses almost double of the space compared to the others. Also the values are out of the range of the previous one.

The chart type -> what are the spatial consequences of the chosen chart?



Ritaglio schermata acquisito: 14/04/2021 12:52

Barplot should never have truncated axis but lineplots may have them, since size is not used for encoding

### Doping under the microscope

Tuesday marks the 25th anniversary of Ben Johnson's victory in the Seoul Olympics 100m final and his subsequent disqualification for doping. Here we take a look at doping's impact on athletics and how the number of athletes being sanctioned has risen.

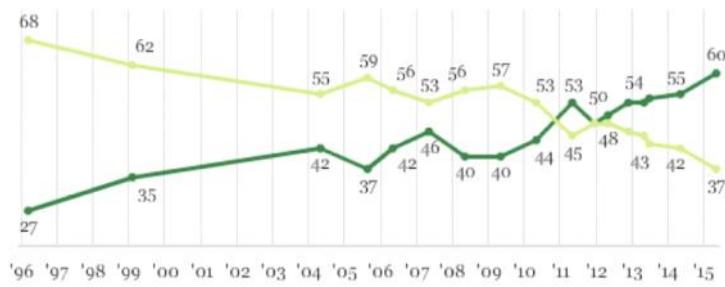


Ritaglio schermata acquisito: 14/04/2021 12:53

Since we are focusing on the dynamic of the record and how it changed over time. So focusing to 9 to 11 seconds is a perfect choice. And maintains the trustworthiness of the graph.

*Do you think marriages between same-sex couples should or should not be recognized by the law as valid, with the same rights as traditional marriages?*

■ % Should be valid ■ % Should not be valid



here the truncation is not correct — the sense of comparable scale is compromised

Ritaglio schermata acquisito: 14/04/2021 12:55

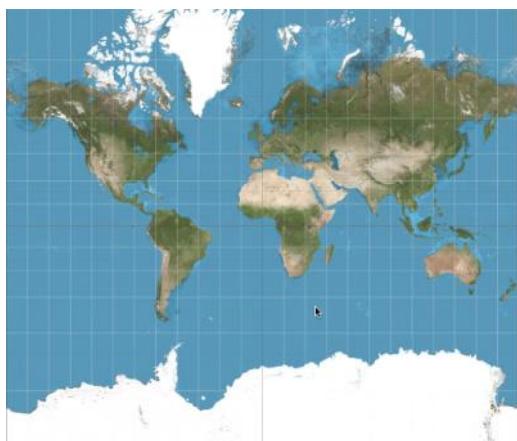
The scale is not reported but we may guess that the scale goes from 20 to 70.

Since you want to compare the two lines focusing on the central values does not help to consider the values of the two lines. This provide a misconception.

## map projection

- every projection is distorted
- the larger the area, the greater the distortion
- no projection can accommodate all map purposes
- choose damage limitation as the driving principle
- 'equal area' projections better for thematic mapping: distortion on shape rather than on size, thus values per region are correct
- scope of view, distance from equator, focus on land/sea can drive the choice

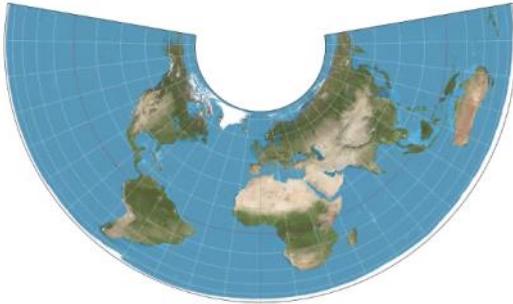
Ritaglio schermata acquisito: 14/04/2021 12:57



Ritaglio schermata acquisito: 14/04/2021 12:59

## map projection

### quick'n'dirty guidelines



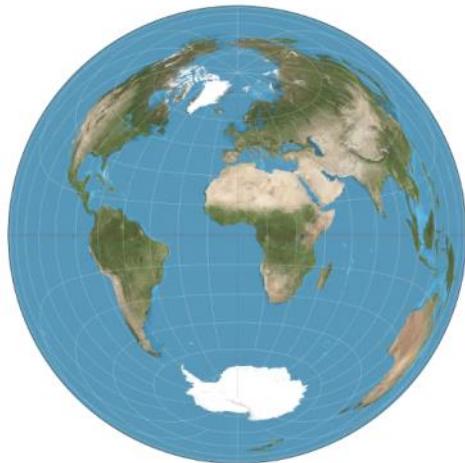
*albers*

equal-area conic

good at country level

Ritaglio schermata acquisito: 14/04/2021 12:59

## map projection



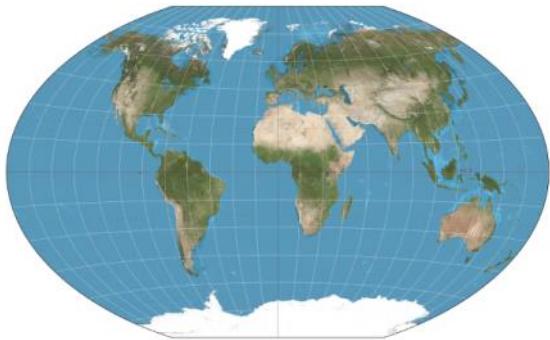
## quick'n'dirty guidelines

### *lambert*

- azimuthal equal-area
- spherical projection
- good for hemisphere or continent-level view
- recommended by european environment agency for any european mapping purposes

Ritaglio schermata acquisito: 14/04/2021 13:00

## map projection



## quick'n'dirty guidelines

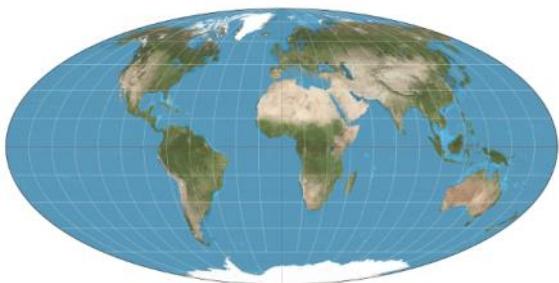
### *winkel-tripel*

- best choice for viewing the whole world
- modern standard world map adopted by



Ritaglio schermata acquisito: 14/04/2021 13:00

## map projection



## quick'n'dirty guidelines

### *mollweide*

- greater emphasis on the accuracy of ocean areas
- useful for atmospheric mapping and flight paths

Ritaglio schermata acquisito: 14/04/2021 13:01

# Charts

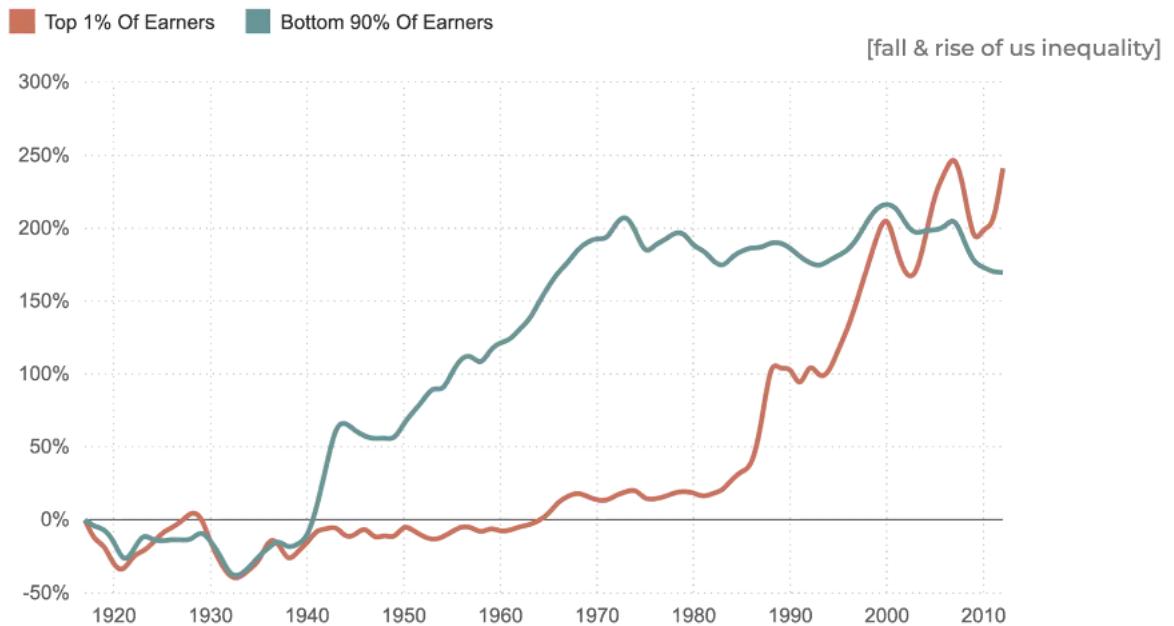
mercoledì 21 aprile 2021 11:31

charts

## line chart

fever chart, stock chart

Income Growth, From 1917-2012



Screen clipping taken: 21/04/2021 11:33

Line plot with the difference that we have a time series as a source of data.

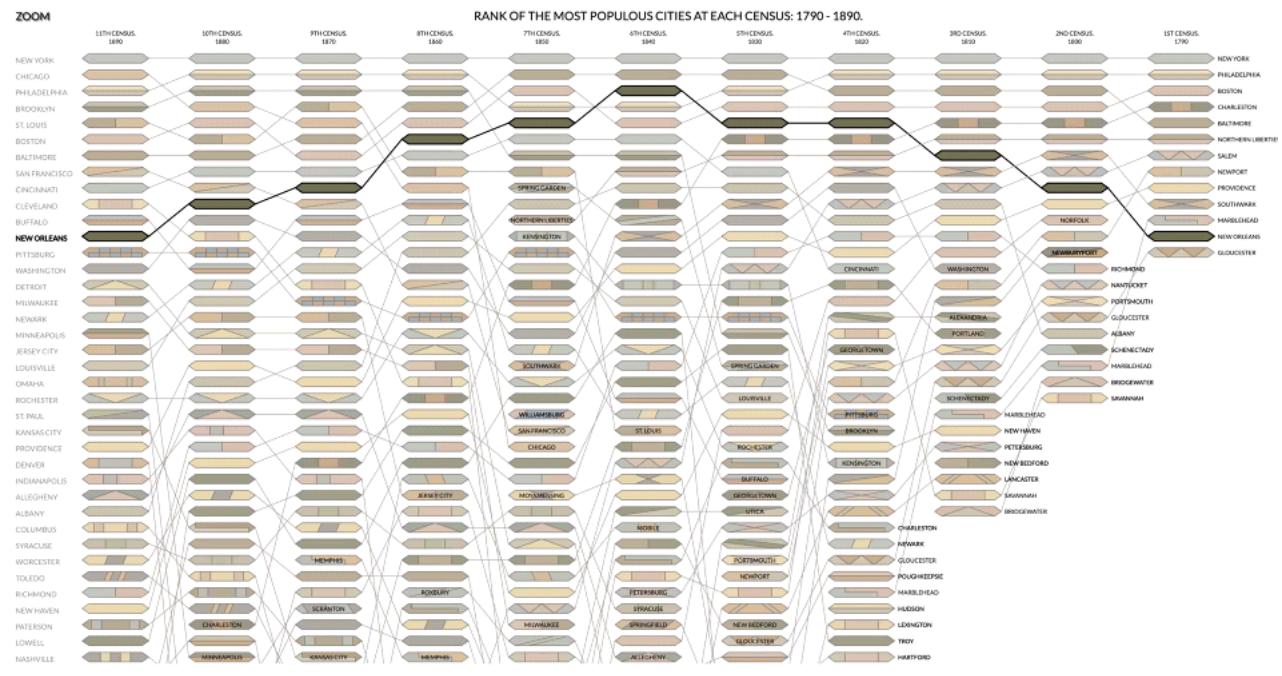
This type of data may or not having the connecting points.

We need to look out for the axes especially if we have more than one.

We look for spikes and trends -> points outside the expected flow. We have to be very careful in the use of aspect ratio. Usually on the y-axes is independent on time. The author has to choose the right view for the message of the lines.

Sparkline -> line chart with minimal decoration about ticks and marks is shown as an inline plot within the text.

# bump chart

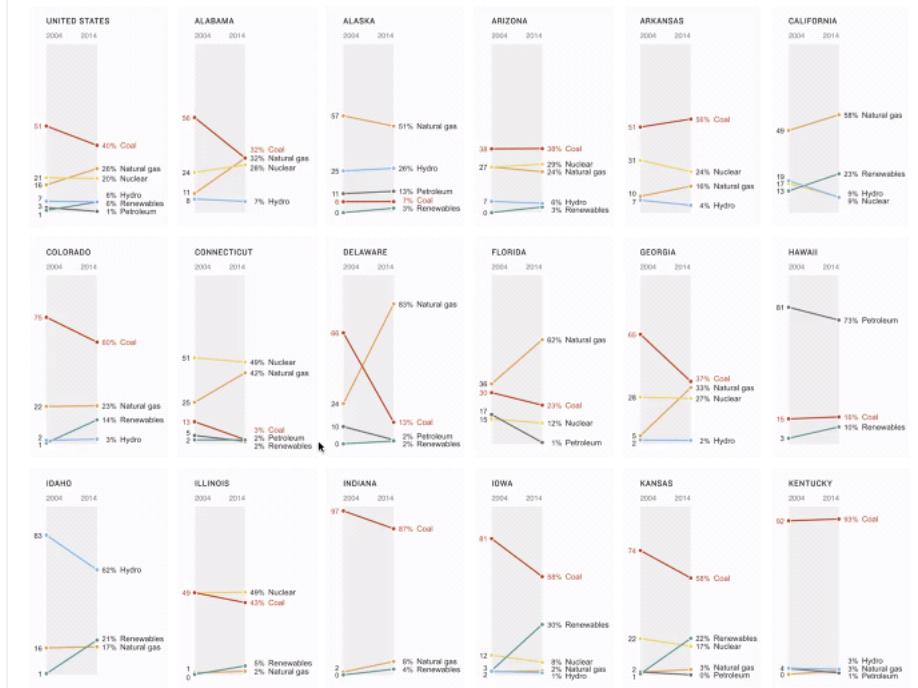


Screen clipping taken: 21/04/2021 11:37

Bump chart -> presenting a set of samples on the columns and all the other columns present a set of features. -< we follow the line connecting all the characteristic for each sample.  
 Here there is the need to carefully analyse the axes. Looking the connecting lines.  
 Interesting find lines with large slope -> means that suddenly there is a huge change.  
 Also here providing the user with interaction tools -> to highlight sample is essential, otherwise all the sample make the plot quite messy.

# slope graph chart

How Each State Generates Electric Power (2004-2014)



Screen clipping taken: 21/04/2021 11:40

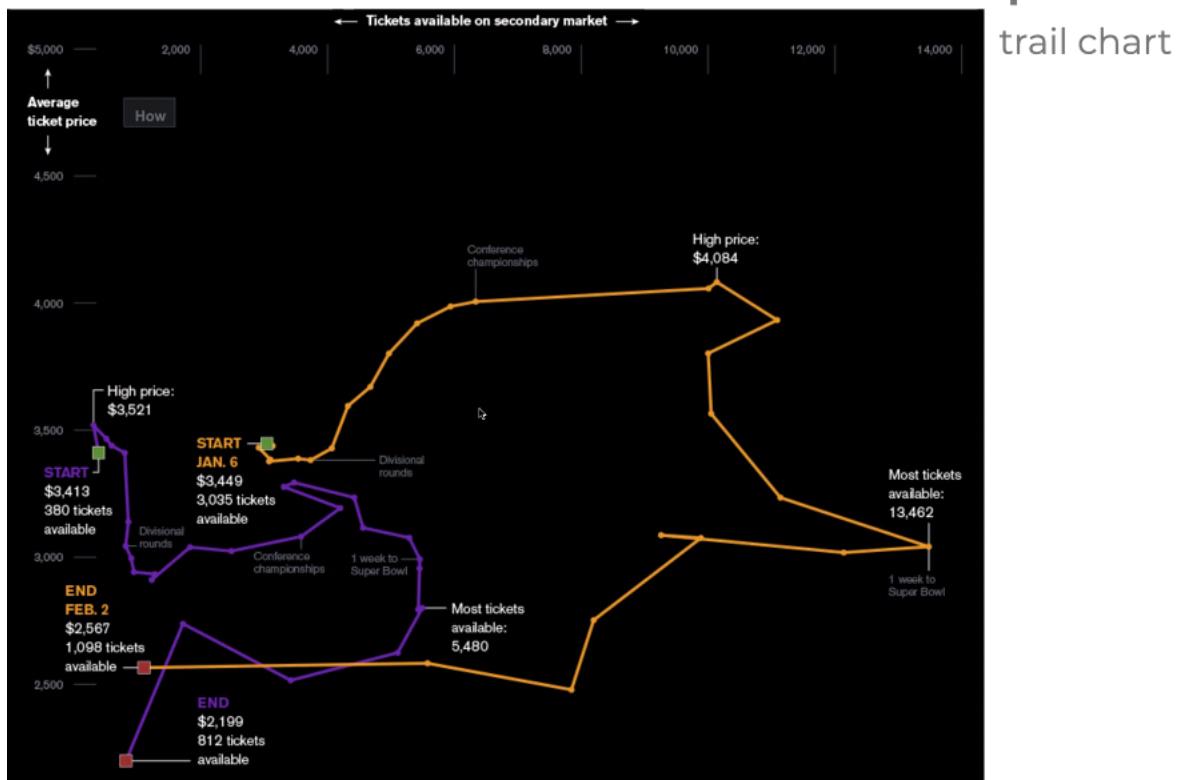
Similarly here we have the slope graph chart. What changes is the presentation. Here we have multiple layouts. Usually there are two steps in time and several categories marked by different colors.

The line is connecting the values for the same category for each sample. The interesting thing is that not all lines are going in same direction otherwise, we would not have any interest in this kind of plot.

It is interesting to try to group samples according to their trends.

Pay attention to the aspect ratio -> it can influence the slope of the connecting lines.

# connected scatter plot



Screen clipping taken: 21/04/2021 11:43

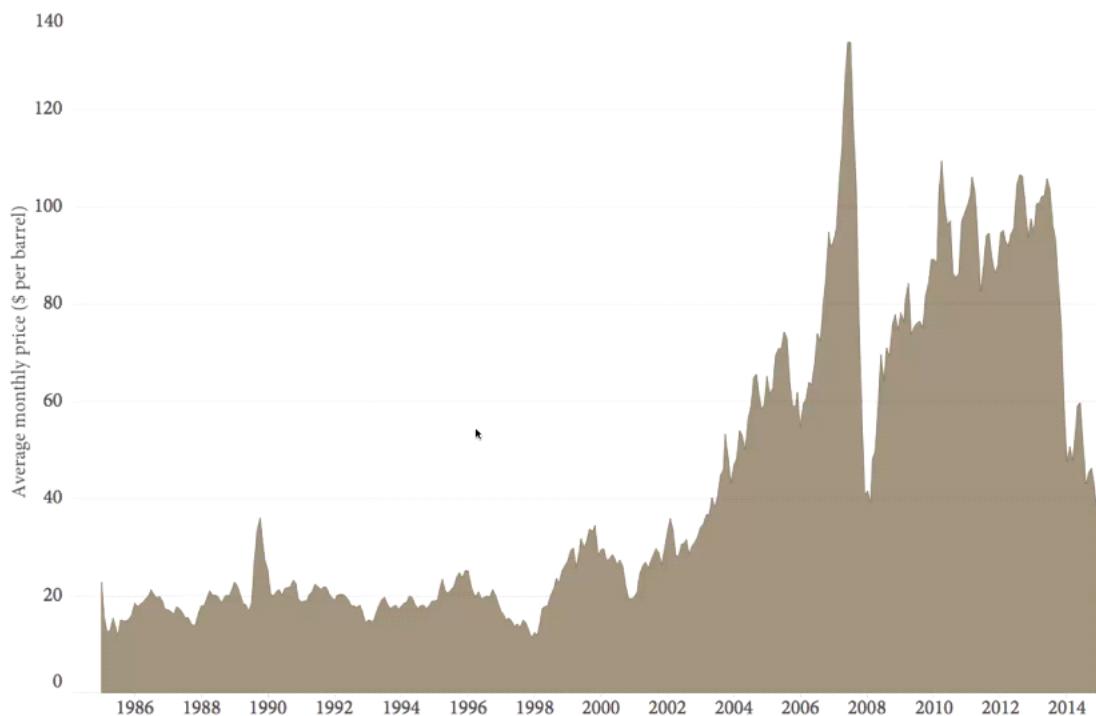
Connected scatter plot, the time, the longitudinal dimension is not represent is inherited in the dynamic of the plot.

Price of superbowl ticket. -> ticket price and ticket availability are the two dimension. Time is given by the connecting lines. The connecting lines give the flow of time until the last day.

We have independent variables -> this allows to have a very nice narrative.

# area chart

**Crude Oil Prices (West Texas Intermediate), 1985 - 2015**



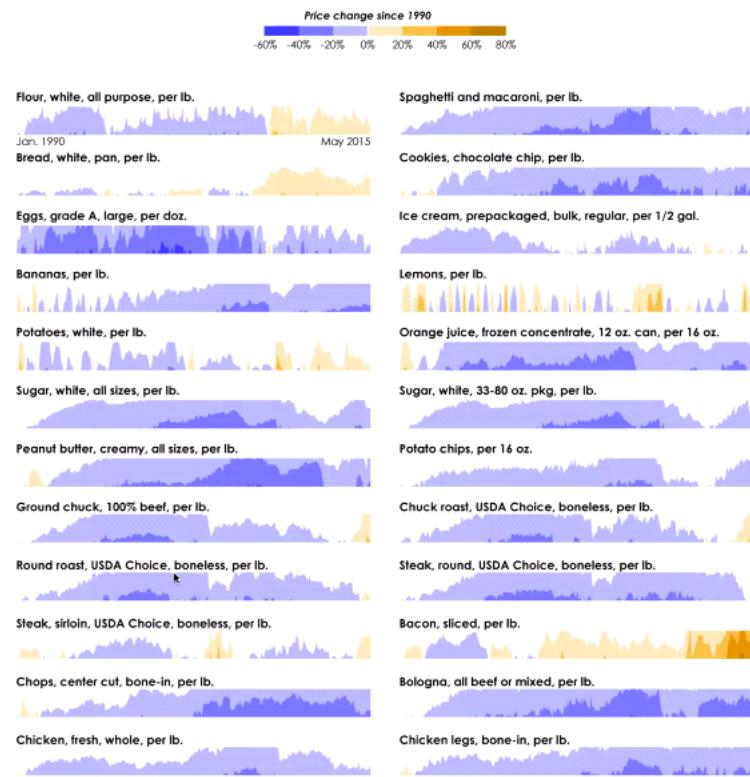
Screen clipping taken: 21/04/2021 11:45

The area chart where the longitudinal dimension represented on the x-axes, we do not have line points but lines and the area between the lines and the reference point is highlighted to better represent the dimension.

The use of colors for stacked area.

If you have many categories we can use small multiples.

# horizon chart

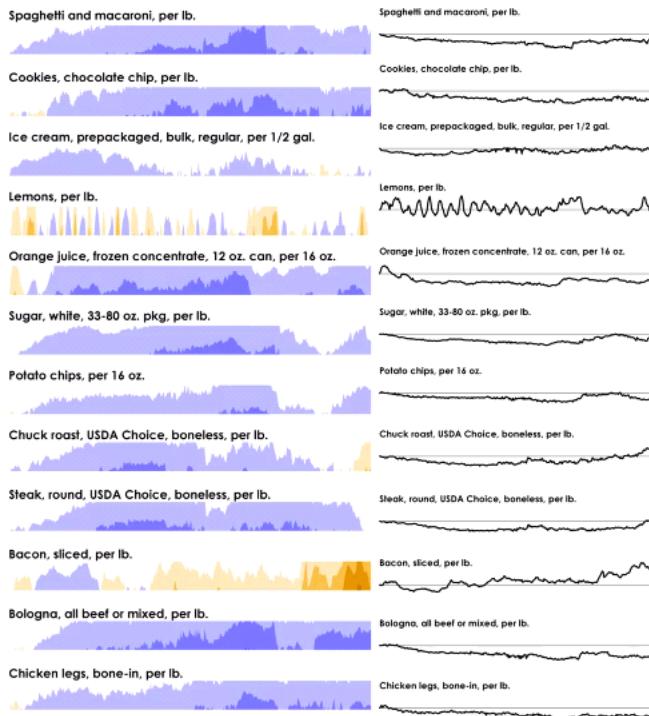


Screen clipping taken: 21/04/2021 11:47

Small multiples, any sample is represented by a kind of area chart but, the difference is that on the y axes we have colored things.

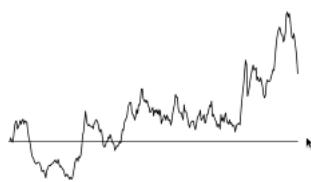
We have a gradient color chart that represent intervals. The area chart is colored according to this particular reference gradient. Here is necessary having a meaningful palette. You have to look out for common trends.

# horizon chart



Screen clipping taken: 21/04/2021 11:49

An alternative is having a line chart expressing the difference as a function respect to the axes. We can start from the line chart to create the horizon chart.

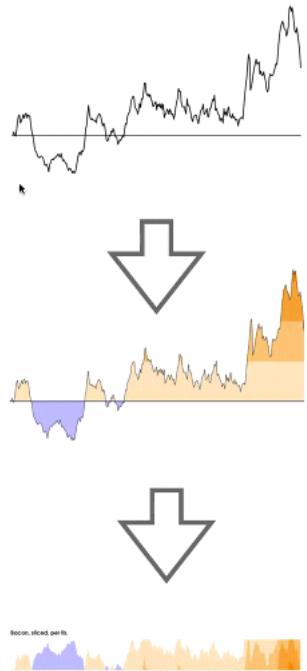


Screen clipping taken: 21/04/2021 11:50

But why change from this graph to the horizon chart ? Because the height of the chart going in the negative part is unpredictable. We want to have a definite way.

So if i want small multiples i prefer to have them of all the same height.

Using htis solution of moving everything udner the reference line to the top and using the color is a smart way to condense such number of different graphs in something more uniform



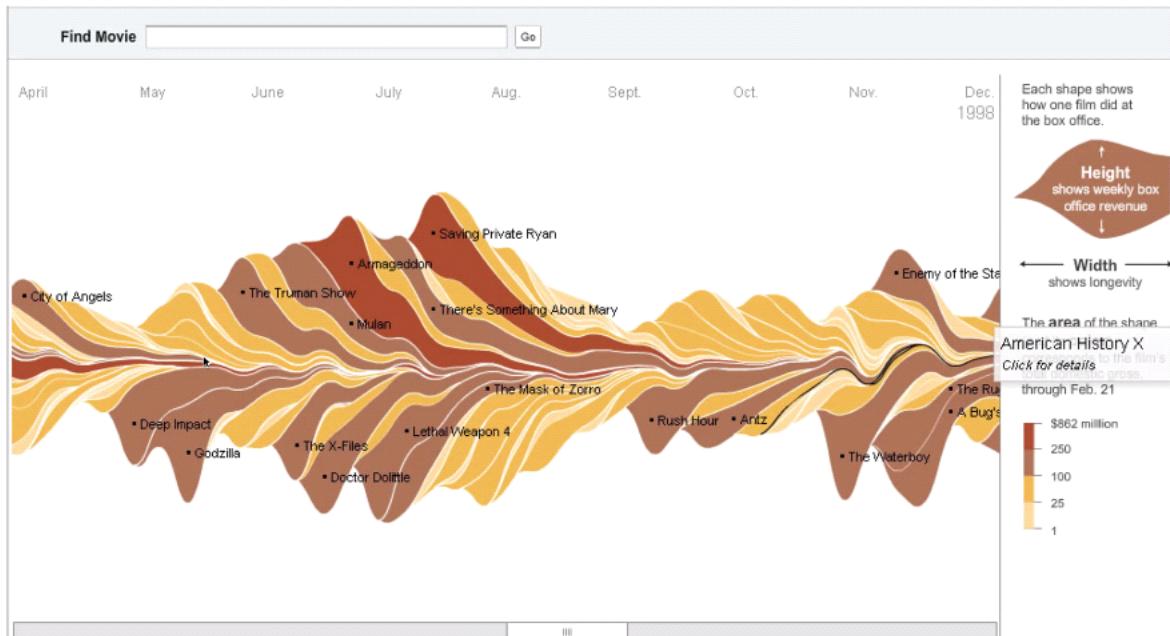
Screen clipping taken: 21/04/2021 11:50

First step -> area chart and then everything is symmetrically rotated on the up part of the graph.

## chrts

# stream graph

theme river



Screen clipping taken: 21/04/2021 11:52

another famous plot is the stream graph.

Here we are showing the box office receipts for some blockbuster.

This is a very visual way of representing data and is not for exact value perception.

It is good for a graphical visual impact and is good for detect seasonal patterns.

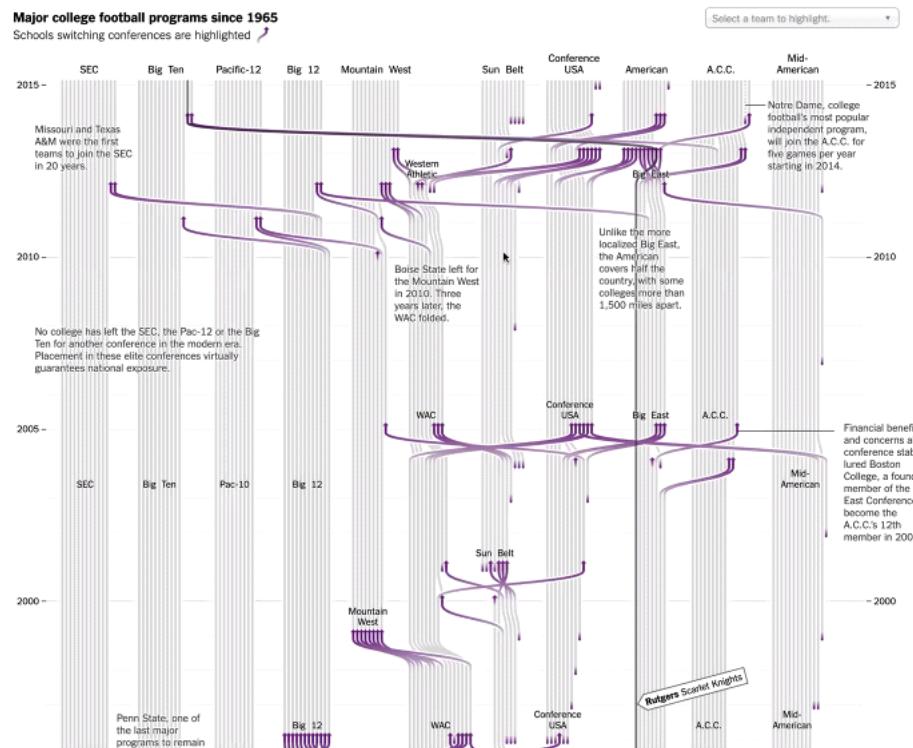
The color legend is essential. Using smoothed lines gives a much better impression of the streaming

effect.

# chrts

# connected timeline

## relationship timeline, storyline viz, swim-lane chart



Screen clipping taken: 21/04/2021 11:57

Connected timeline -> usually in this kind of plot every sample is associated to different evolution between categories. The longitudinal axes is plotted on the y and usually going up. Everyline is a sample. This kind of plot is used for unusual kind of data.

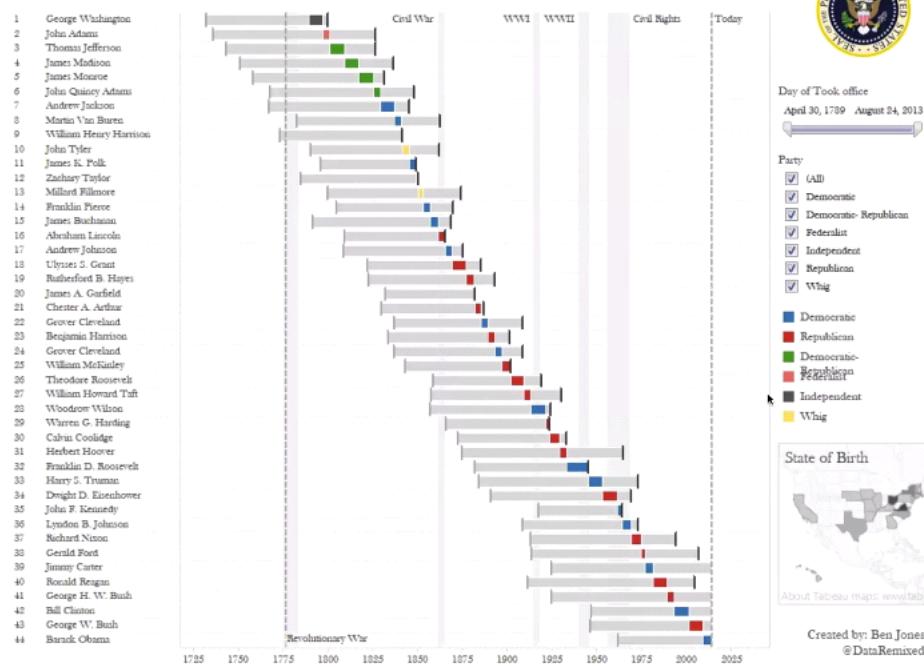
This is rather complex layout and somehow unnatural. Hard to interpret.

Here is essential to properly label the critical events to construct a meaningful storytelling. Using color is very useful to differentiate various trajectories.

# gannt chart

## range chart, floating bar chart

### A PRESIDENTIAL GANTT CHART



Screen clipping taken: 21/04/2021 12:00

Gannt chart -> this is represented for each sample as a row, time line on the x-axes and to each sample is associated a grey bar marked with further annotation defined by colors.  
Here can appreciate in a single graph who were the president with longer governemnt, their political inclinations.

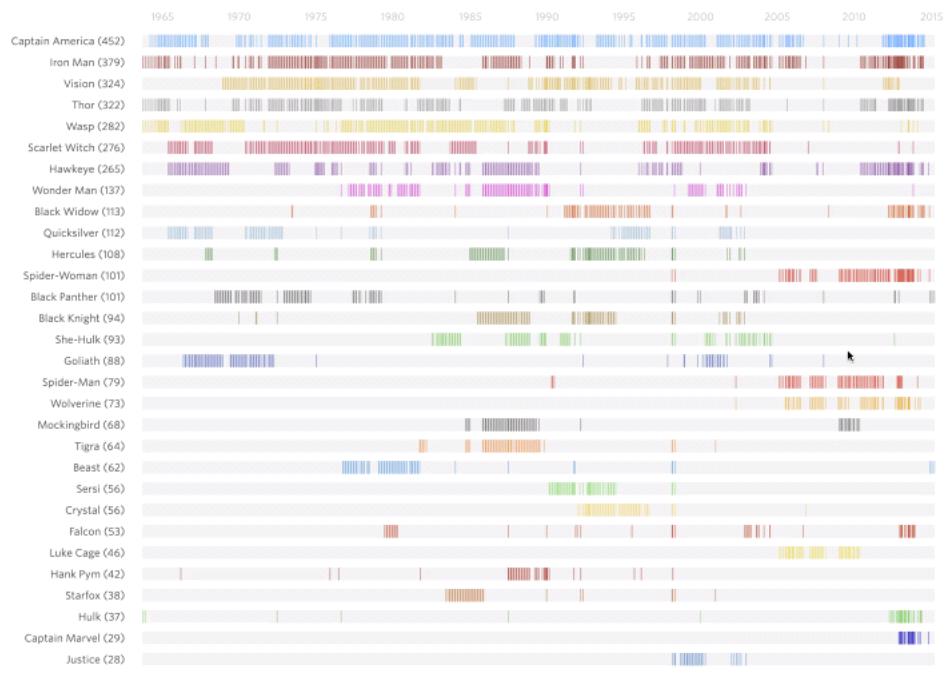
So colors mark particular categories. This compoartive plot help the reader.

# instance chart

milestone map, barcode chart, strip plot

## 'Avengers' characters' appearances over time

Avengers team members sorted by most number of appearances, across the 'Avengers' comic book titles in our analysis\*. Each colored vertical stripe is an appearance in one of the issues as an Avenger.



Screen clipping taken: 21/04/2021 12:03

The time dimension is on the x. We have samples for each rows. And lines represent particular events for particular studies.

This is quite simple but can be arbitrary complex using smart combination of shapes and colors. Here we look for clusters and dispersal and empty zones.

Properly sequence categories.

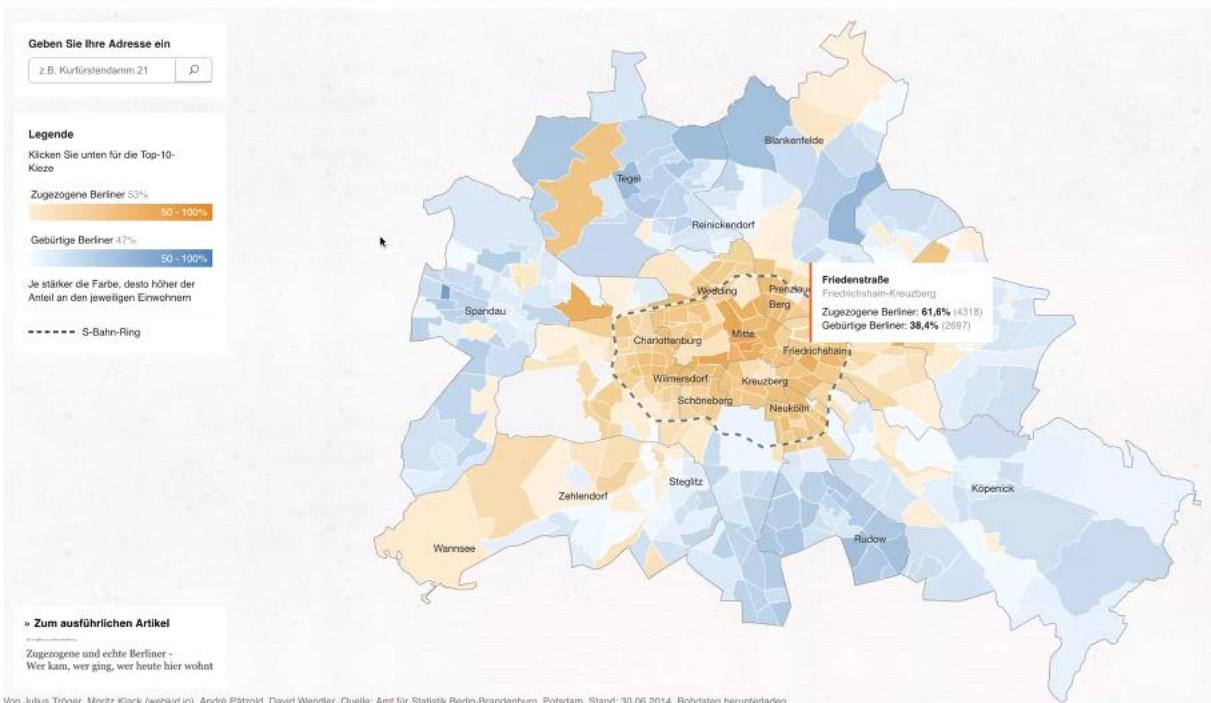
Differnt glyphs and size for different events.

# chloroplet map

## heat map

Berliner Morgenpost Echte Berliner und Zugezogene Wie der S-Bahn-Ring die Stadt teilt

Spir



Screen clipping taken: 21/04/2021 12:05

Maps -> the first kind of spatial data is the chloroplet map. A map, the entity of the map are marked with sub-entities and every administrative division is colored according to a particular function of yourv statistical data georeferenced.

Polygons associated to subregion.

This kind of plot are meaningful if they have association for a particular datum.

Data may be area-normalised -> the dimension of the map is not essential for the understanding.

In general look for outliers and do not fooled by area size.

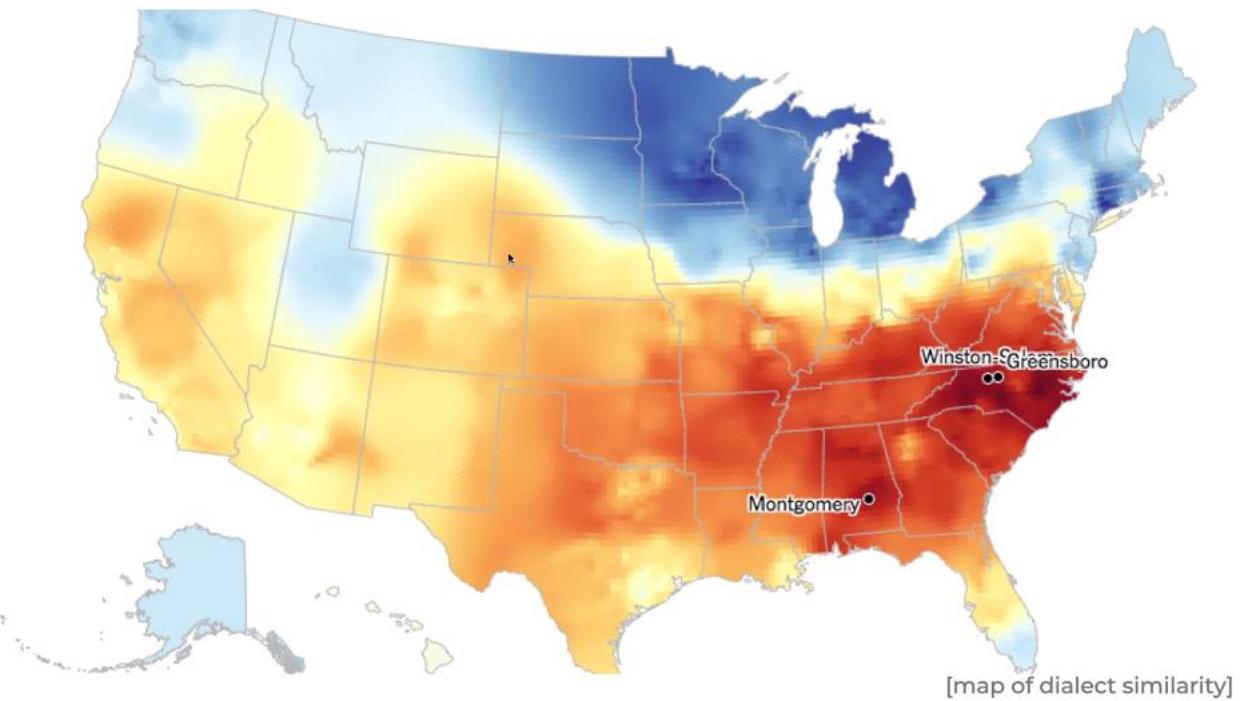
Detect larger common areas.

map almost transpart, no geo details, they are there only to mark reasonability or optimization of my datum on the plot itself.

Morevoer, you may want to add a furhter level of information and you can achive this by interposing texture to have even deeper inside for the data set you are studying.

# isarithmic map

contour map, isopleth map, isochrone map



Screen clipping taken: 21/04/2021 12:11

Different between isarithmic map and chloroplet.

In countour map we are not associated coloros to different regions subdivisions.

The data do not match with adminsitrative boarders.

So it is better whn data and subregional boarders do not match.

Mind interpolation reducing precision of what you are seeing.

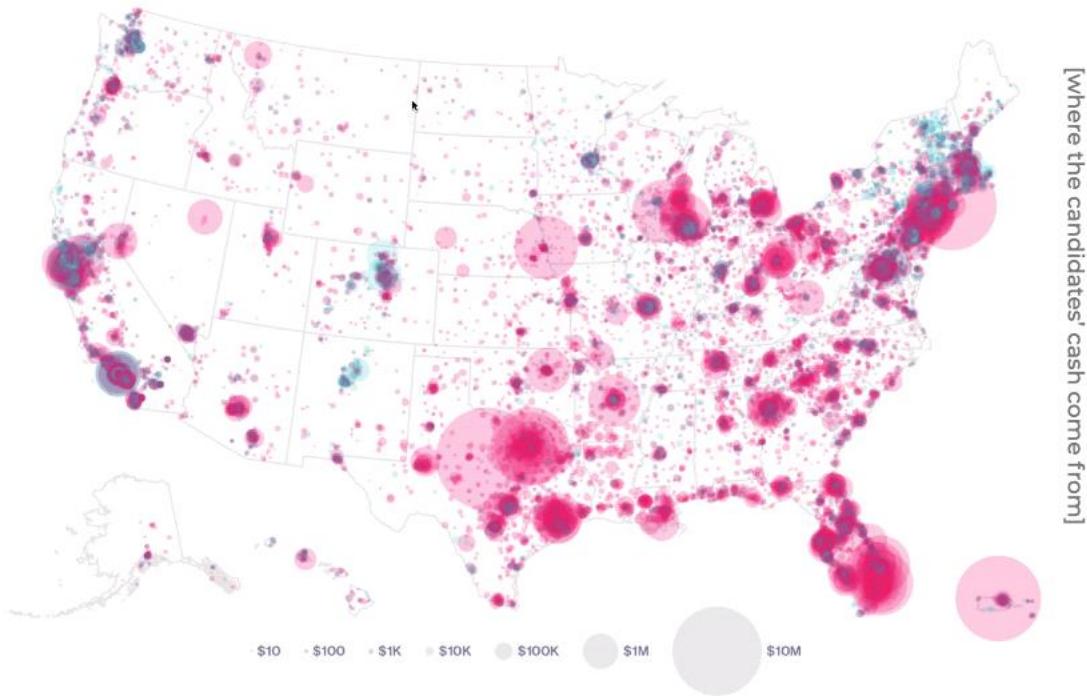
To optimize you have to keep the lengedn close to the map and color borders not sharp.

If you need to inance the gradients of changes you can add isolines.

# proportional symbol map

charts

graduated symbol map

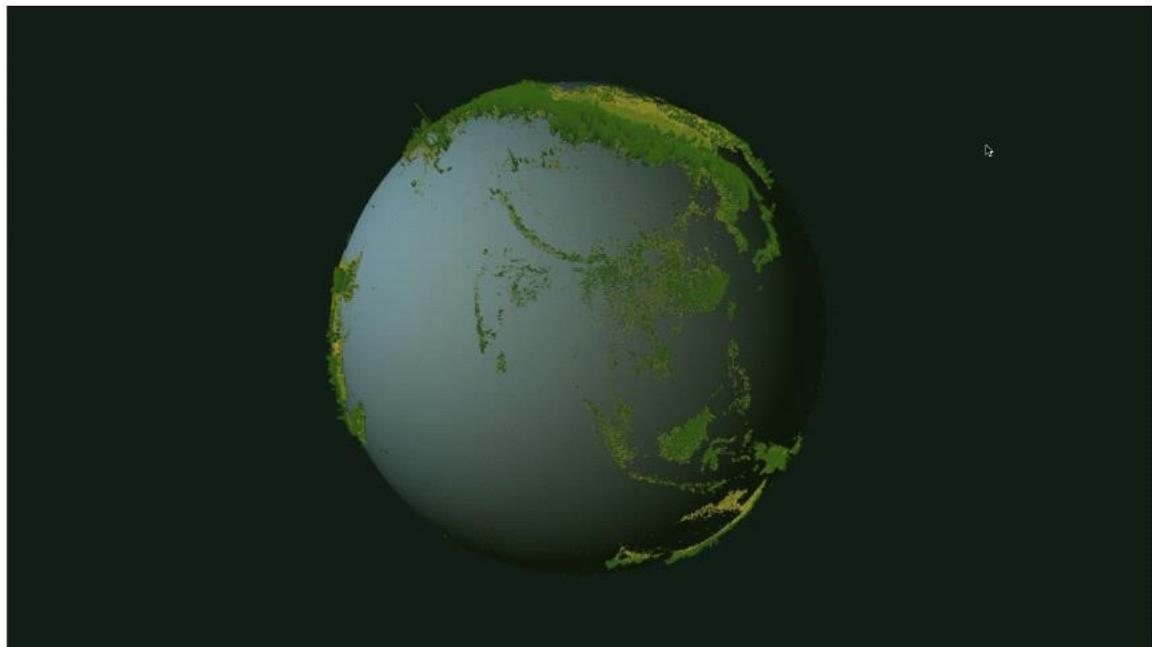


Screen clipping taken: 21/04/2021 12:15

Here the datum is pointlike associated to a set of coordinates. Punctual information.  
Area size and color encoding data.  
Mind overlap of different colors -> hard to understand how many points.  
Adding piecharts not a good idea because this would make the plot very messy.

# prism map

isometric map, spike map, datascape



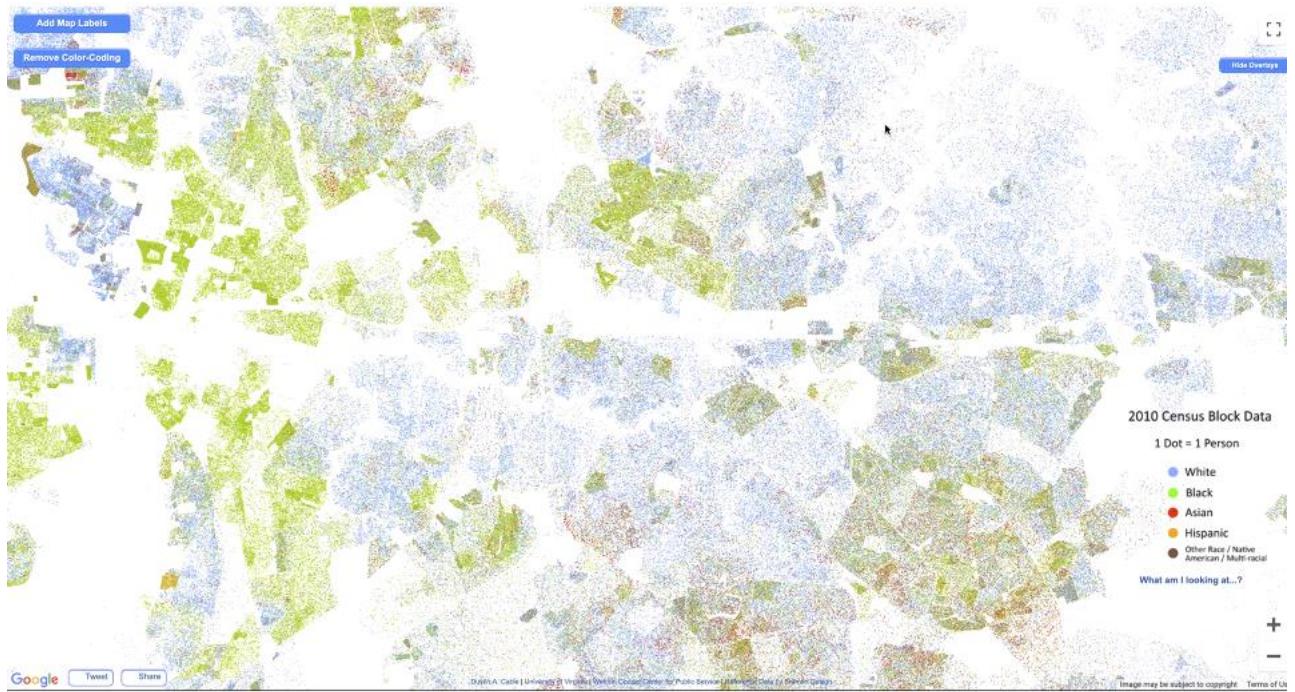
Screen clipping taken: 21/04/2021 12:17

Dinamic maps representing in an esagerated scale a particular kind of data on earth surface.  
Average density of trees.

This is only for gist od data -3d difficult to interpret.  
Here interaction is essential.

dot distribution map, pointillist map, location map

[racial dot map]



Screen clipping taken: 21/04/2021 12:19

We have a special reference system, for each couple of coordinates we can plot the points with different colors

This is kind of a demographic plot.

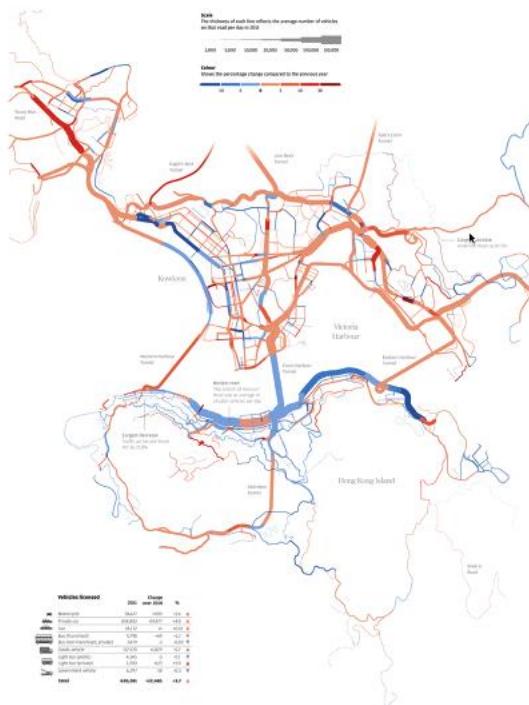
Look for density variations at different levels.

Use very different colors

Interactive zooming is essential.

Not necessarily applied to maps.

connection map, route map, stream map, particle flow map



[arteries of the city]

Screen clipping taken: 21/04/2021 12:21

Instead of dots we want trajectories we use the flow map.

We select some interesting paths. We color such stream with appropriate color and the thickness of the line is proportional to the quantity that the datum is encoding. If you do for many kind of rivers you will have many attributes so that if we can have it animated sequences help greatly.

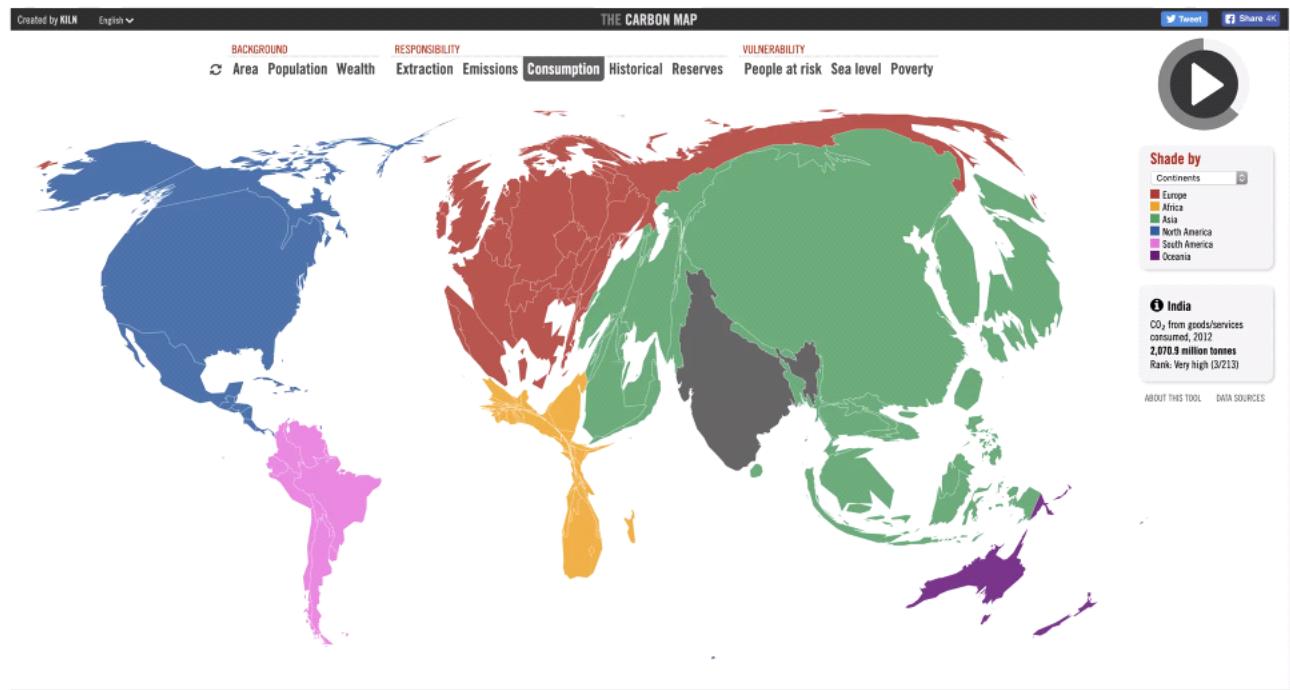
Reconstruct dynamics.

Annotations here is essential.

Use geo distortion if helps readability.

# area cartogram

## density-equalising map



Screen clipping taken: 21/04/2021 12:24

They are very popular. Not only many colors but every country is deformed according to their different values associated to colors.

We are looking for carbon dioxide.

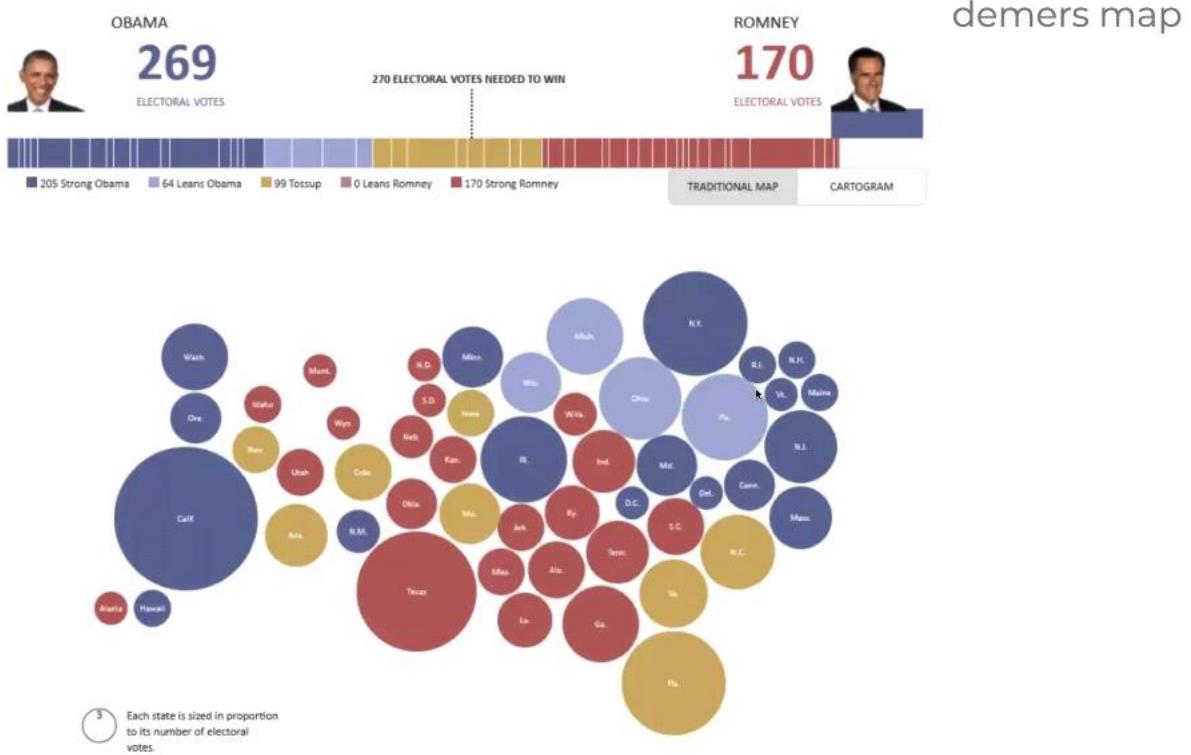
Need the user know the real country shape/size.

Compare ratios with real size.

Manual animation helps.

# dorling cartogram

demers map



Screen clipping taken: 21/04/2021 12:27

Dorling cartogram is a map plot.

Circles are encoded for adjustancy sample in the world.

Here a map of the US each ball is centered in the capital and the size is proportional to the electoral vote.

here we need to mark the correspondence between the datum and the geography.

Preserve adjacency of neighbouring regions.

cartogram, bin map, equal-area cartogram, hexagon bin map



[mapping the spread of obesity]

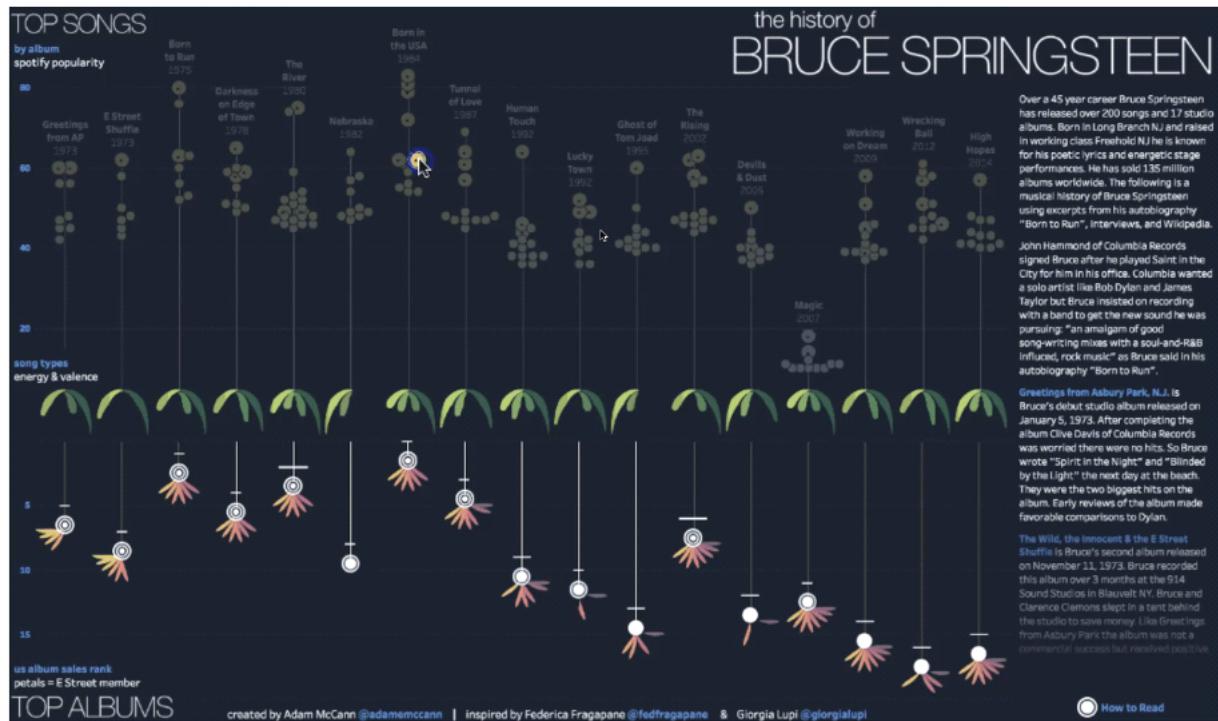
Screen clipping taken: 21/04/2021 12:29

Small multiple panel. Each panel is a geographical entity. Glyphs are marking for administrative entities.

Although it is very schematic is quite easy to interpret. Position is crucial for little glyphs.  
Using tiles fro regions, no area attribute.

Optimise tile-region relationship -> the arrangement of squares should be optimized.

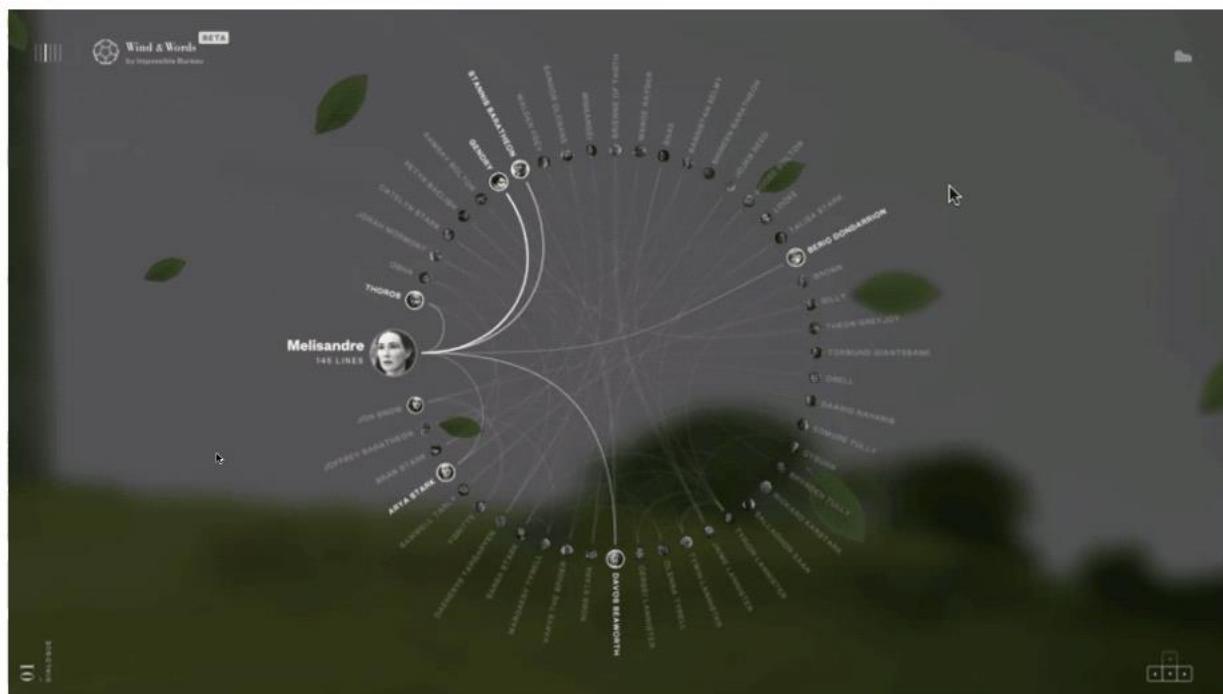
+ many untamed beasts



Screen clipping taken: 21/04/2021 12:32

Musical history of bruce springsteen. For each album we have the popularity according to spotify. We have many different categories.

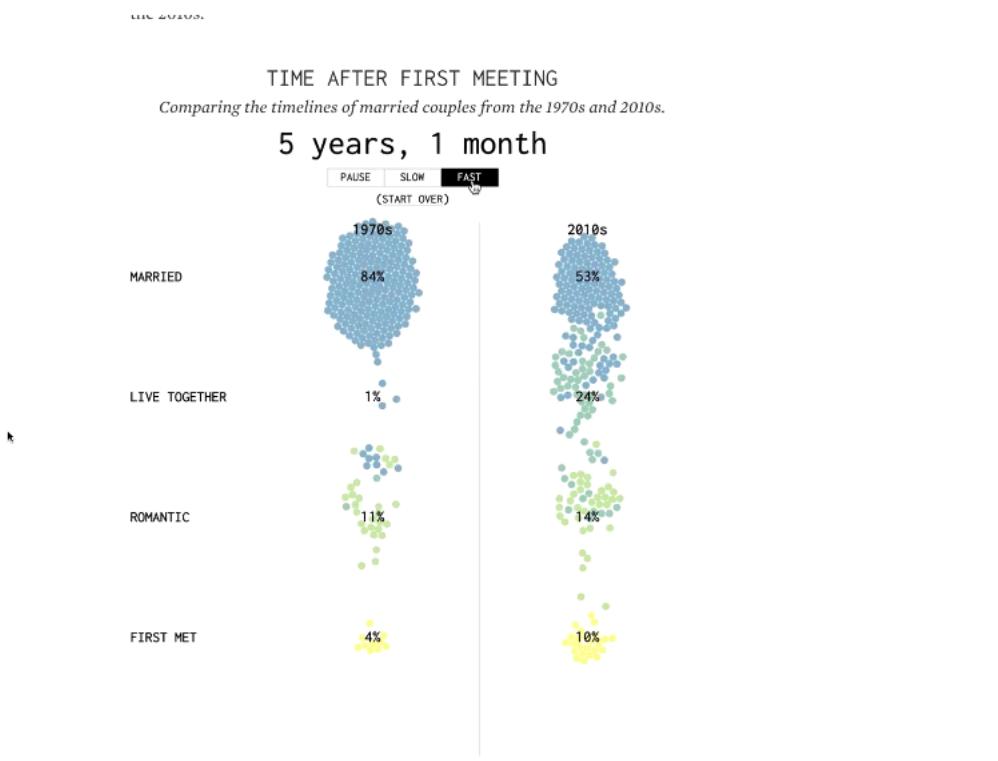
+ many untamed beasts



Screen clipping taken: 21/04/2021 12:34

Network for defining relationships inside a tv show.

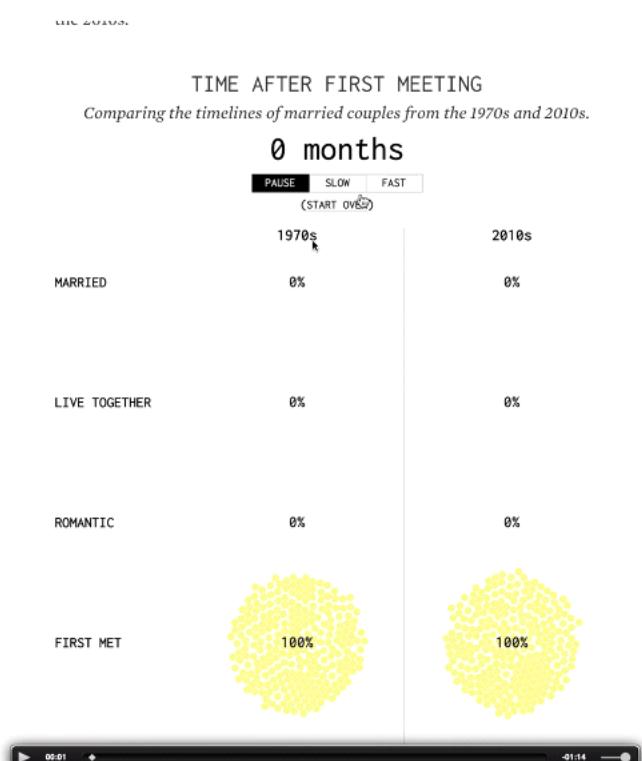
# + many untamed beasts



Screen clipping taken: 21/04/2021 12:35

Longitudinal plot of how much married couples are taking from first meeting to getting married.

# + many untamed beasts

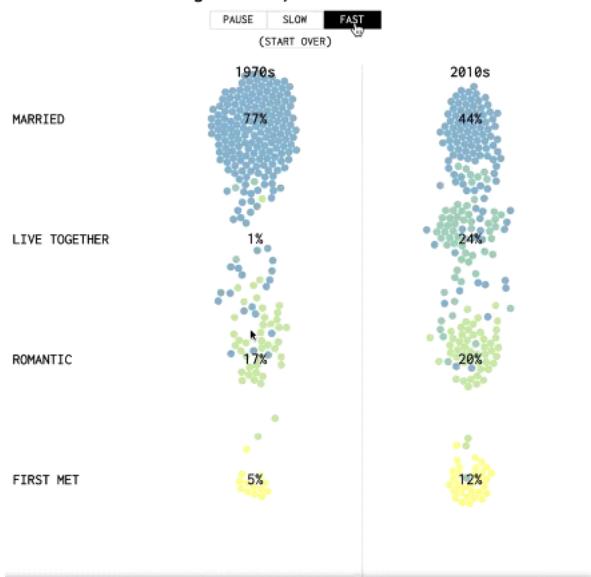


Screen clipping taken: 21/04/2021 12:36

TIME AFTER FIRST MEETING

Comparing the timelines of married couples from the 1970s and 2010s.

4 years, 1 month

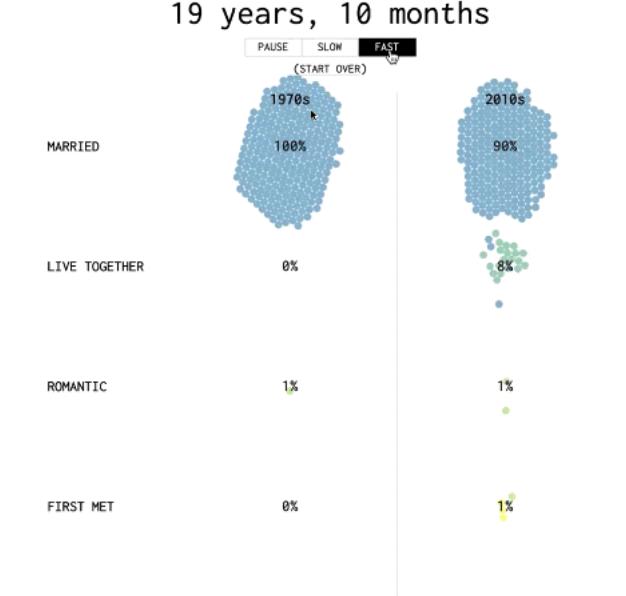


Screen clipping taken: 21/04/2021 12:36

TIME AFTER FIRST MEETING

Comparing the timelines of married couples from the 1970s and 2010s.

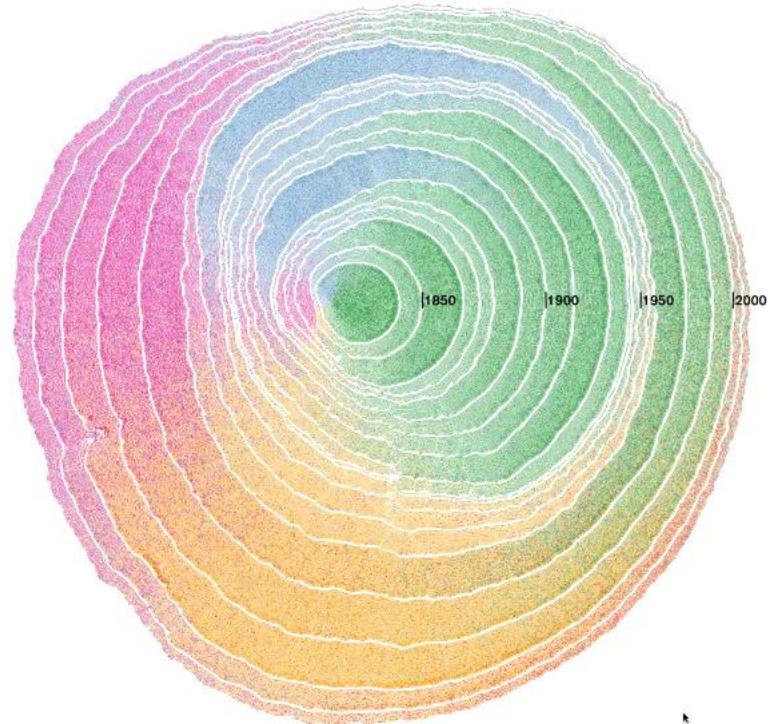
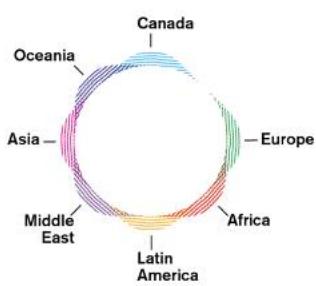
19 years, 10 months



Screen clipping taken: 21/04/2021 12:37

Very famous representation -> very intuitive and gives the reader a very immediate taste of the underlining dynamic.

# + many untamed beasts



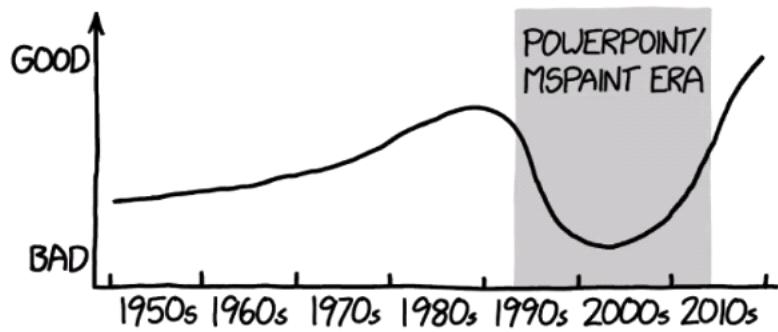
[mapping us immigration]

h

Screen clipping taken: 21/04/2021 12:39

Colors are marking the provenance of the people who go in the US.  
This is a tree stump.  
Very intuitive and the color is very meaningful.

## GENERAL QUALITY OF CHARTS AND GRAPHS IN SCIENTIFIC PAPERS

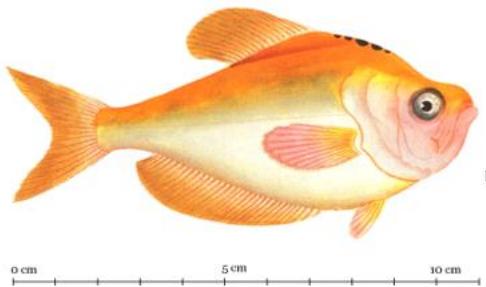


Screen clipping taken: 21/04/2021 12:41

# Rules

martedì 27 aprile 2021 11:39

## i - show your data



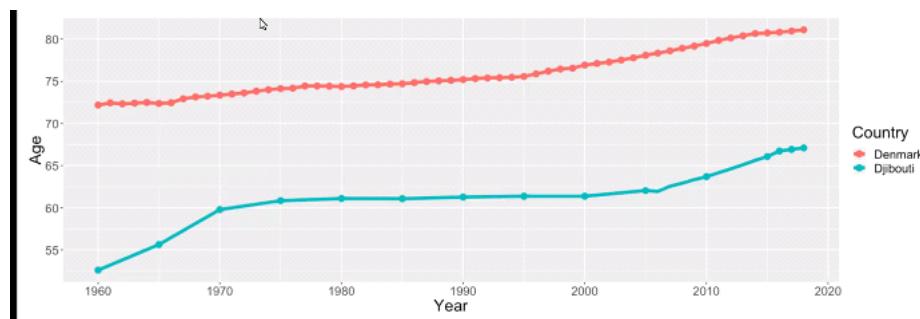
Ritaglio schermata acquisito: 27/04/2021 11:40

Insert all the elements that help to interpret your data.

## i - show your data

- ↗ avoid summaries & aggregation if not required
- show where data is missing but do not distract the user
- rely on the deductive, inductive and adductive reasoning of the user

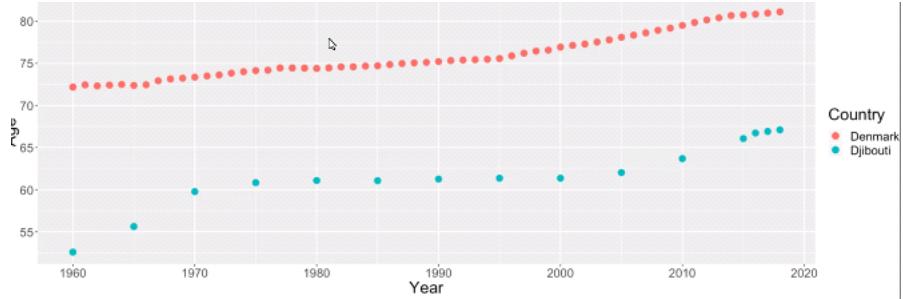
Ritaglio schermata acquisito: 27/04/2021 11:40



Ritaglio schermata acquisito: 27/04/2021 11:41

This is a time series of the life expectancy in two countries : a first world country and a third world country.

We see how the difference between these two countries is still the same.  
The problem is that if you do not see the line carefully you see only lines



Ritaglio schermata acquisito: 27/04/2021 11:44

This is a much more honest plot.

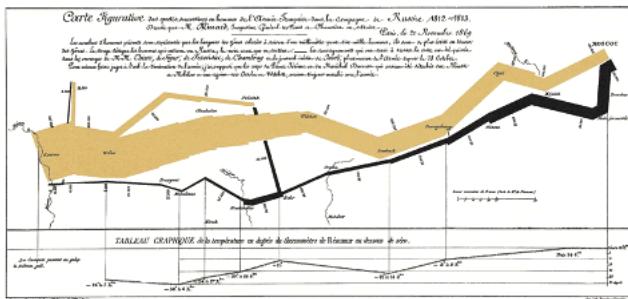
Do not connect points hifing the fact that you have less observation -> be honest and show the data as they are.

## ii - use graphics

a picture is worth a thousand words

show with a diagram or a picture things that it would take  
a lot of text to describe

"only a picture can carry such volume of data in such a small space" [tufte]



Ritaglio schermata acquisito: 27/04/2021 11:46

Graphics is not always the best choice but it is quite frequently. The strength of the graph is that they can summarize many textual information.

### iii - avoid chartjunks

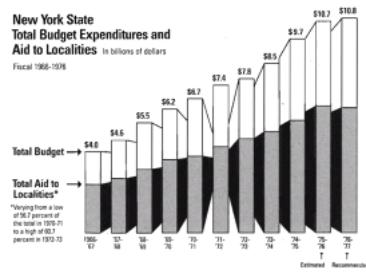
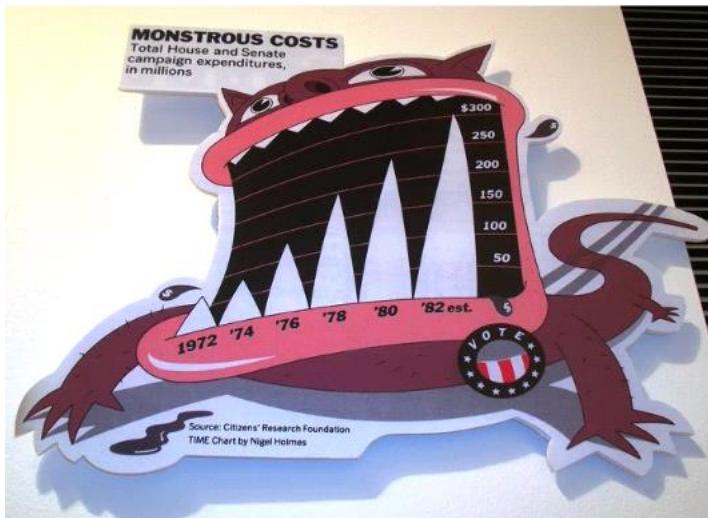
too many decorative elements added, detracting the user from the actual message

typical example: the unneeded 3d effects

"chartjunk can turn bores into disasters, but it can never rescue a thin data set" [tufte]

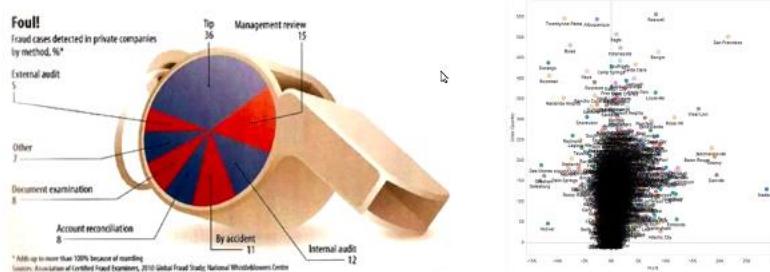
Ritaglio schermata acquisito: 27/04/2021 11:47

Junks included in the chart -> you should carefully avoid too many decorative elements.  
If you have little data you cannot compensate this with many decorative elements to fill up your graph.



Ritaglio schermata acquisito: 27/04/2021 11:52

Here the underlying message is that they focus on the 1972 which should be the ideal situation.  
What happen before too small, what happen later is too big.  
Using shadows in this way should suggest that



Ritaglio schermata acquisito: 27/04/2021 11:58

In the first example we do not see the pie chart frontally -> it is distort and this can also bias the perception of the graph itself.

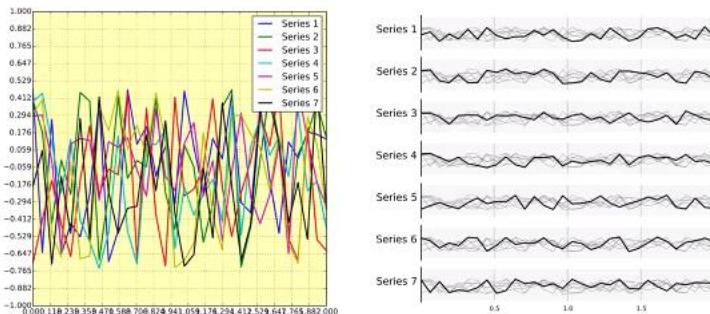
In particular: management review looks very important and seems more than 1 half of tip.  
In the right graph we see a huge concentration of cities.

The author wanted to highlight one particular thing: a huge concentration apart from some small cities and San Francisco .

So the message could have been highlight that San Francisco stands alone.

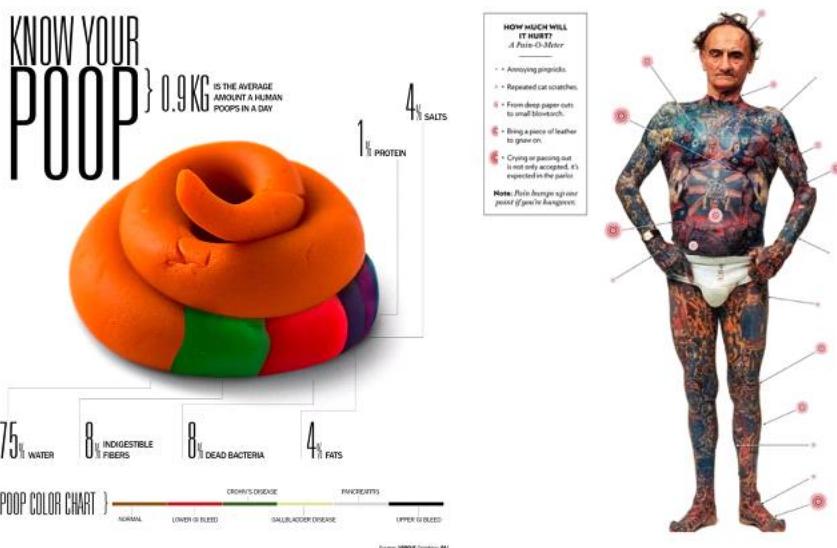
The real answer to why the author create something like that is that the only thing that he really wanted is that San Francisco is an outlier and he did not care about all the other city.

### rescue a thin data set [turte]



Ritaglio schermata acquisito: 27/04/2021 12:11

Other level of chart junks in scientific plot -> slightly change the layout of your plot. Dedicate a little more time in trying to explore solutions that help the reader.



Ritaglio schermata acquisito: 27/04/2021 12:14

Not only mix bad infographic but also mixed with bad taste.

The graph on the left is still a bit informative.

The aim of these kind of graphs is that they catch the attention of the reader.

## iv - the data-ink ratio aka "remove2improve"

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

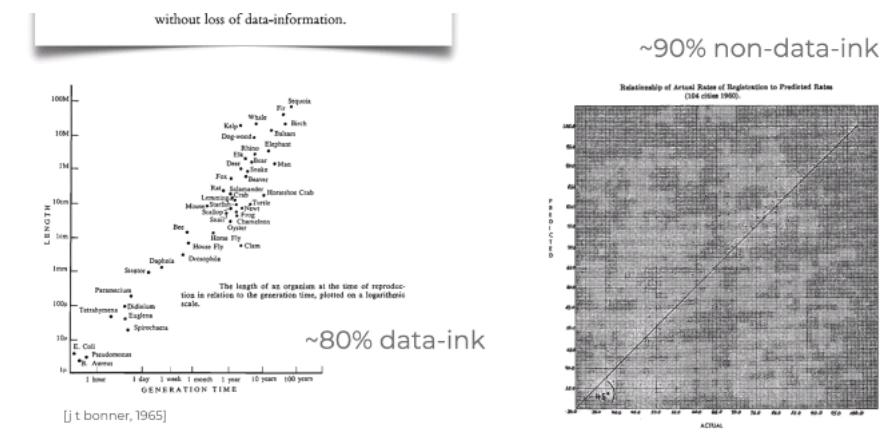
= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1.0 – proportion of a graphic that can be erased without loss of data-information.

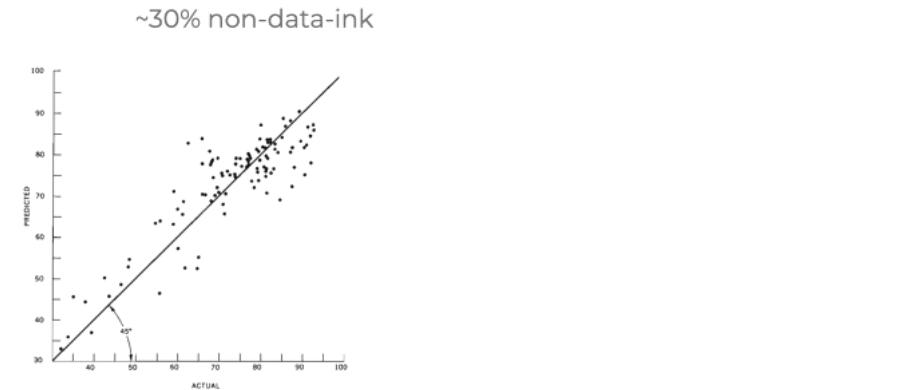
Ritaglio schermata acquisito: 27/04/2021 12:17

Another rule is more quantitative: the data-ink ratio. It can be computed as a formula. Quantity of

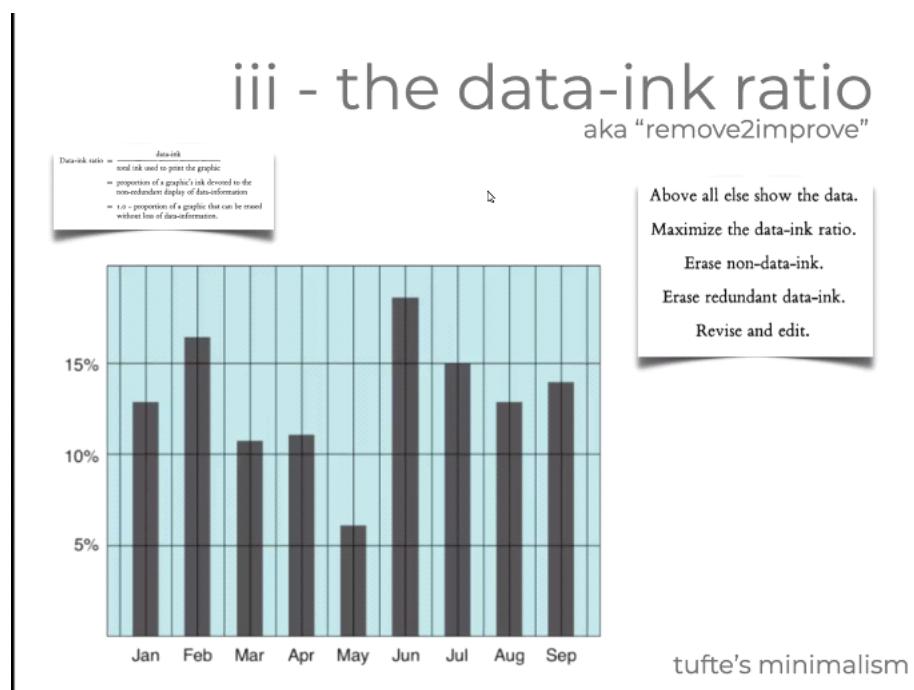
ink dedicated to data representaiton.



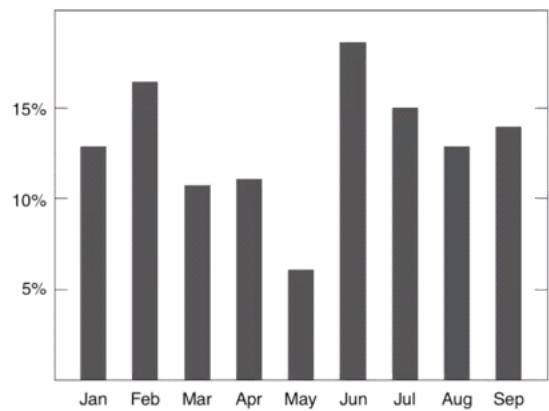
Ritaglio schermata acquisito: 27/04/2021 12:19



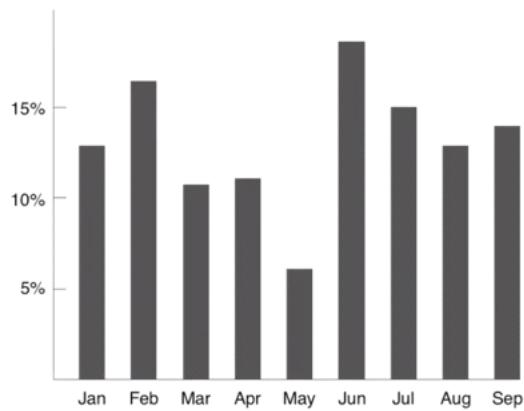
Ritaglio schermata acquisito: 27/04/2021 12:20



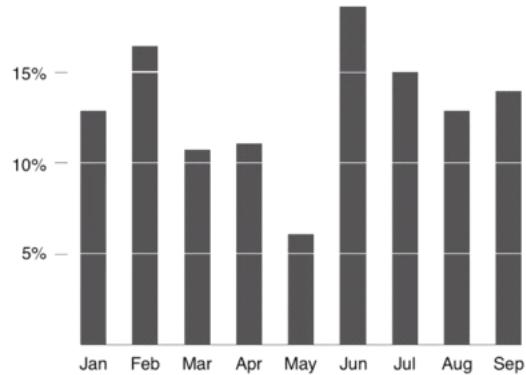
Ritaglio schermata acquisito: 27/04/2021 12:22



Ritaglio schermata acquisito: 27/04/2021 12:23



Ritaglio schermata acquisito: 27/04/2021 12:23

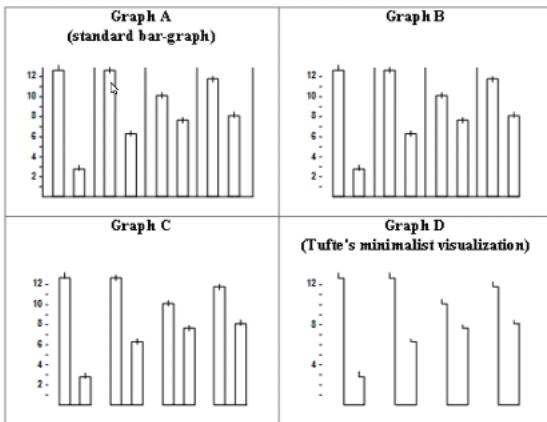


Ritaglio schermata acquisito: 27/04/2021 12:23

# iv - the data-ink ratio

aka "remove2improve"

Data-ink ratio =  $\frac{\text{data-ink}}{\text{total ink used to print the graphic}}$   
= proportion of a graphic's ink devoted to the non-redundant display of data-information  
= 1.0 = proportion of a graphic that can be erased without loss of data-information.



Above all else show the data.

Maximize the data-ink ratio.

Erase non-data-ink.

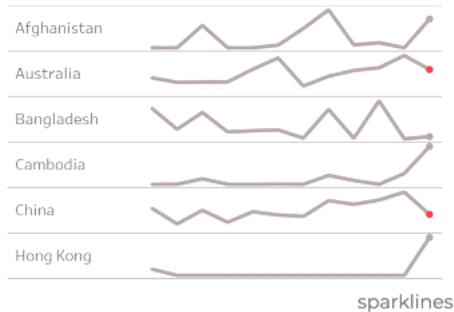
Erase redundant data-ink.

Revise and edit.

tufte's minimalism

Ritaglio schermata acquisito: 27/04/2021 12:24

what chart maximise  
data-ink ratio?



Ritaglio schermata acquisito: 27/04/2021 12:28

Sparklines are built for this purpose.

## v - annotation

↳

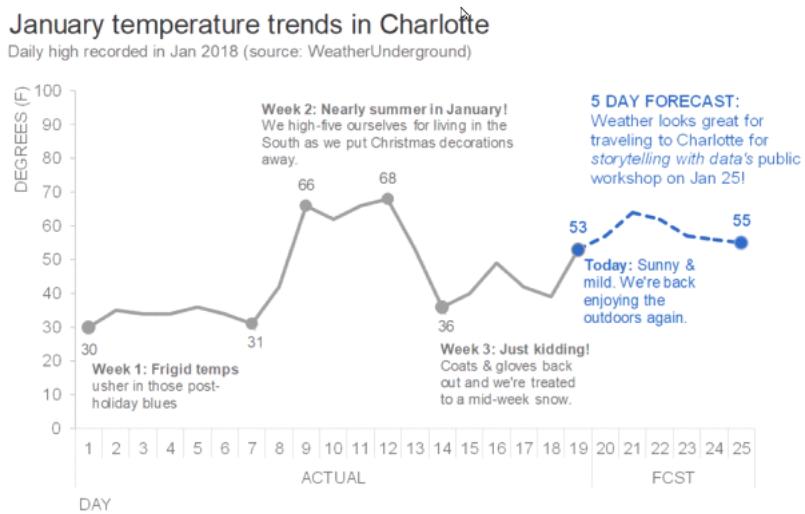
pictures are the focus, but words are important, too

labels are informative guides and they are essential in data design

when labelling, focus on clarity, readability and differentiation

labels should stand out from data

Ritaglio schermata acquisito: 27/04/2021 12:30



Ritaglio schermata acquisito: 27/04/2021 12:31

## vi - micro/macro

fine micro-level details become textures when viewed at the macro level

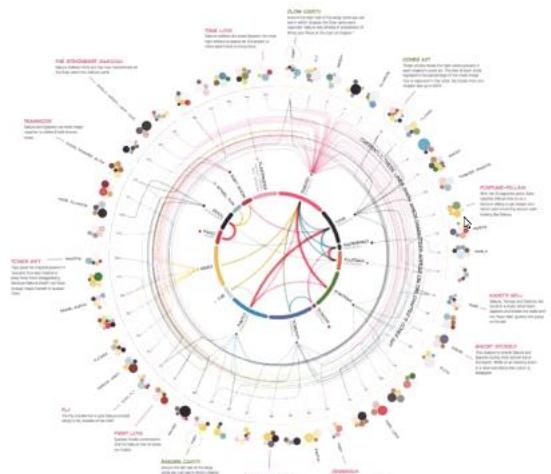
the clarity of the macro is determined by the quality and the quantity of the micro

↳

ben schneiderman's mantra: overview first, then details on demand

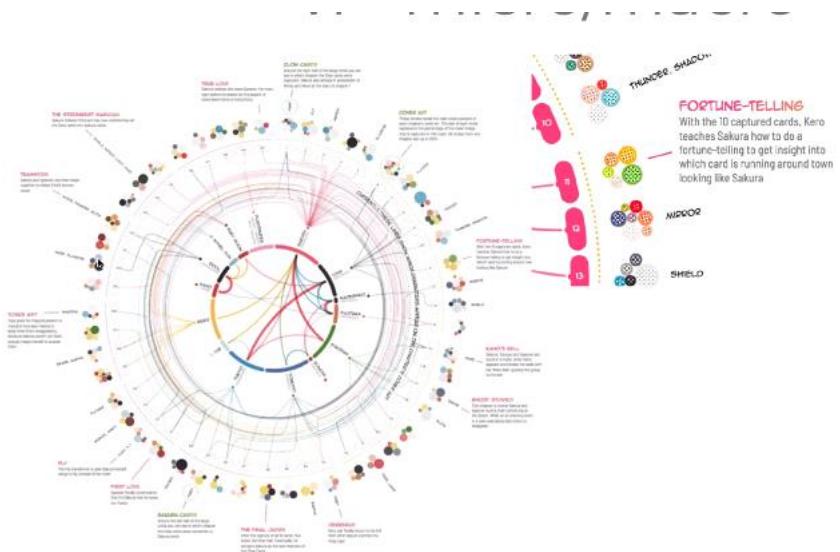
use interactive zooming when possible

Ritaglio schermata acquisito: 27/04/2021 12:35



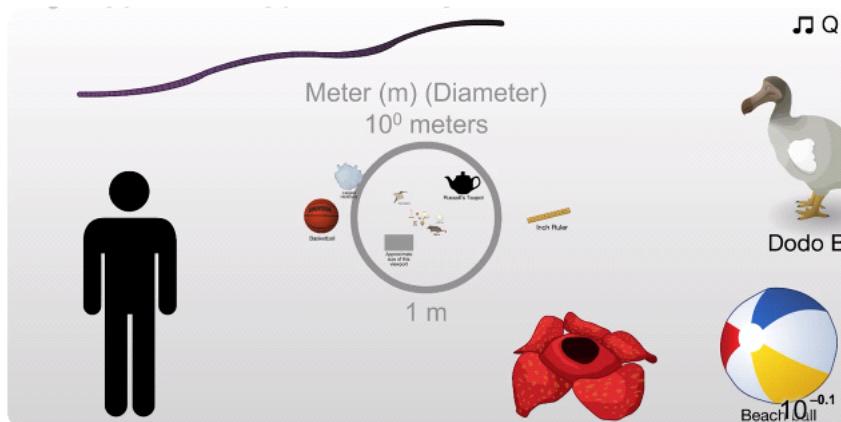
Ritaglio schermata acquisito: 27/04/2021 12:35

We have some nodes connected by links and many information outside.  
We have other then other little circles.

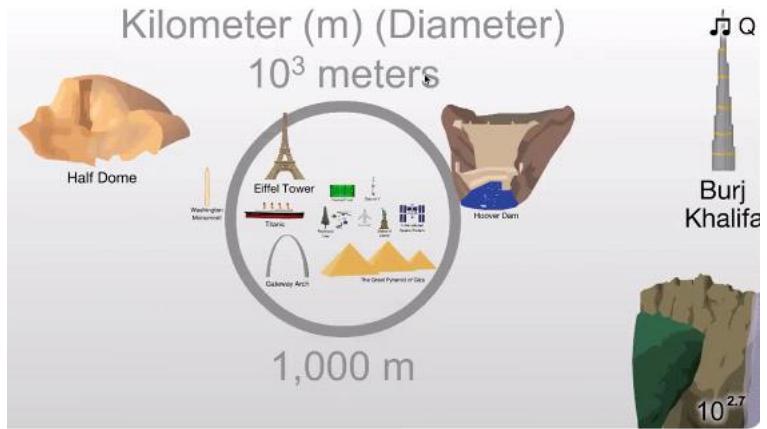


Ritaglio schermata acquisito: 27/04/2021 12:36

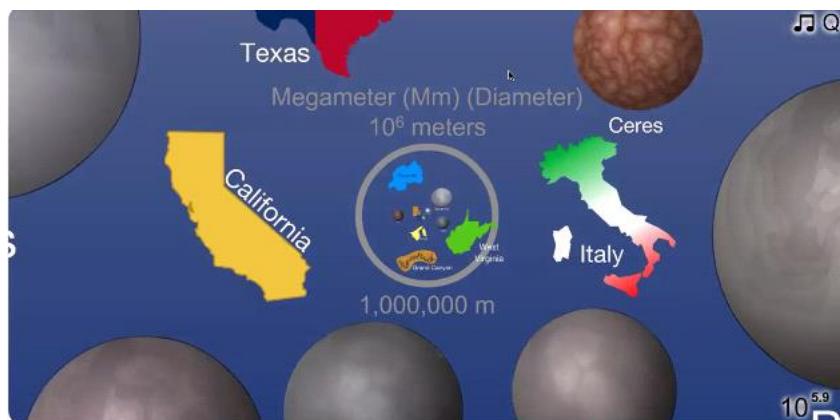
## The Scale of the Universe



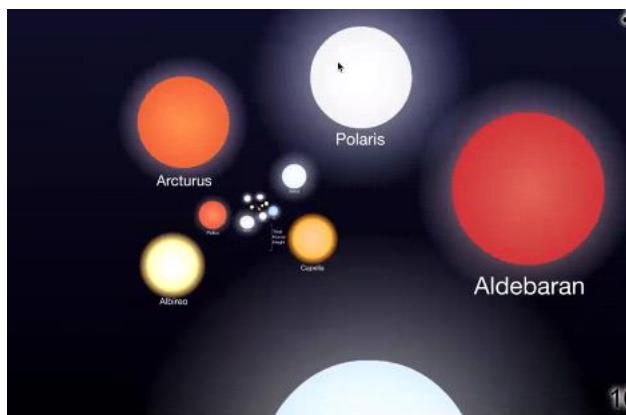
Ritaglio schermata acquisito: 27/04/2021 12:39



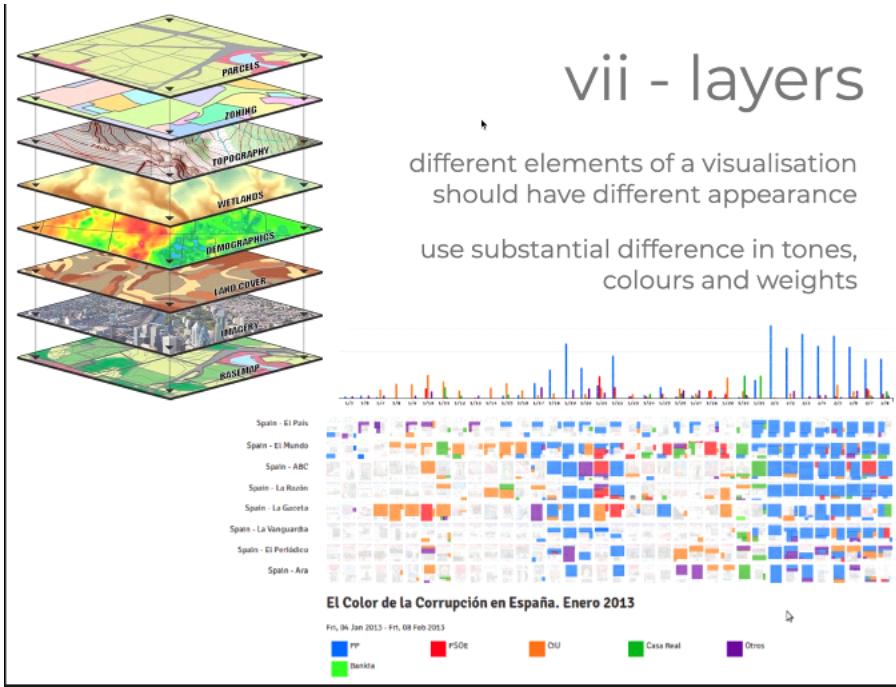
Ritaglio schermata acquisito: 27/04/2021 12:39



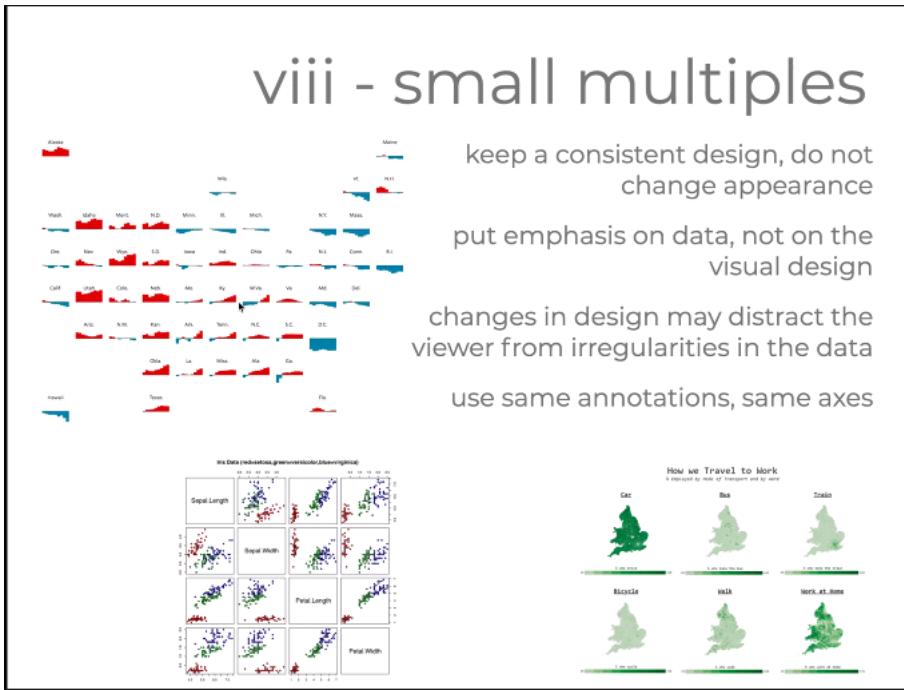
Ritaglio schermata acquisito: 27/04/2021 12:39



Ritaglio schermata acquisito: 27/04/2021 12:39



Ritaglio schermata acquisito: 27/04/2021 12:42



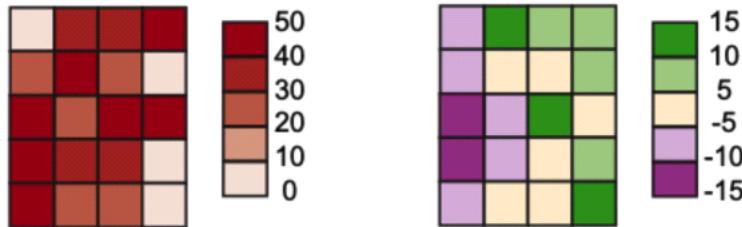
Ritaglio schermata acquisito: 27/04/2021 12:45

# ix - colour

colour can be helpful if used properly

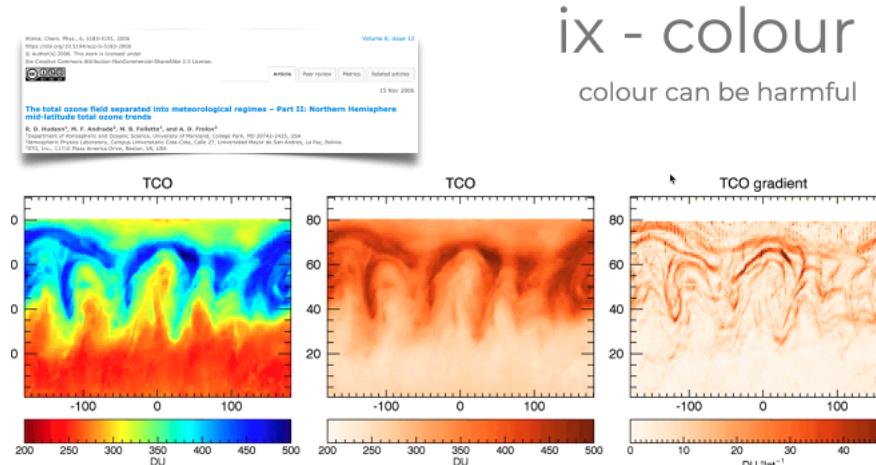
colour can be harmful if used naively

rainbows and gradients have different purposes



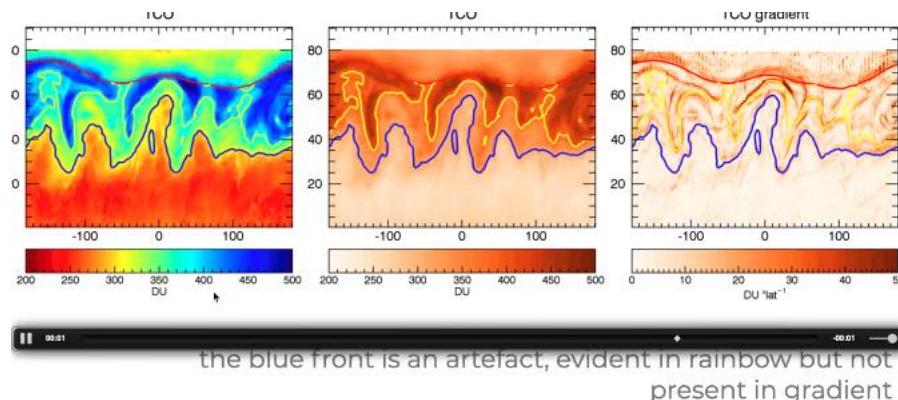
Ritaglio schermata acquisito: 27/04/2021 12:48

Why color can be armful?



Ritaglio schermata acquisito: 27/04/2021 12:49

What changes between the three graph is the palettes.

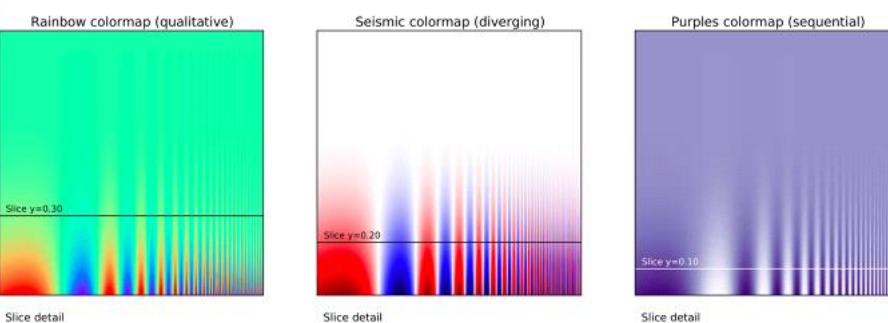


Ritaglio schermata acquisito: 27/04/2021 12:51

IEEE Comput Graph Appl. 2007 Mar-Apr;27(2):14-7.  
**Rainbow color map (still) considered harmful.**  
 Borland D<sup>1</sup>, Taylor MR 2nd.

## ix - colour

colour can be harmful



rainbow & seismic hide details in the bottom-right part

Ritaglio schermata acquisito: 27/04/2021 12:52

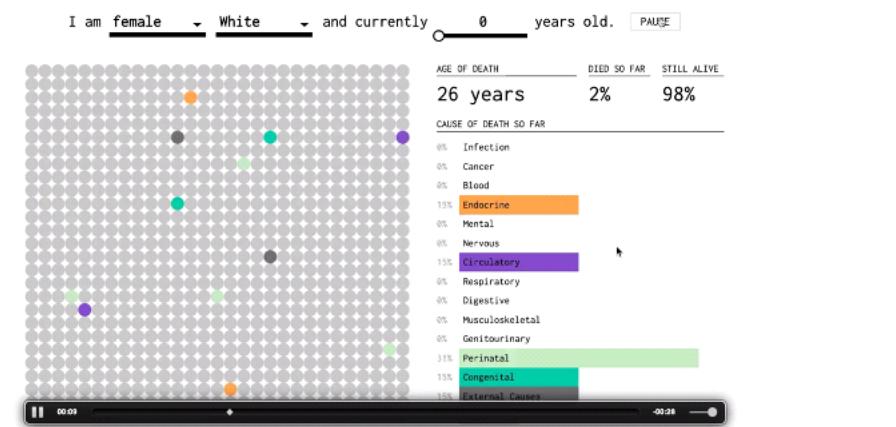
The first two map hide the frequency of data at the bottom.

## X - narrative

what story is your dataviz telling?

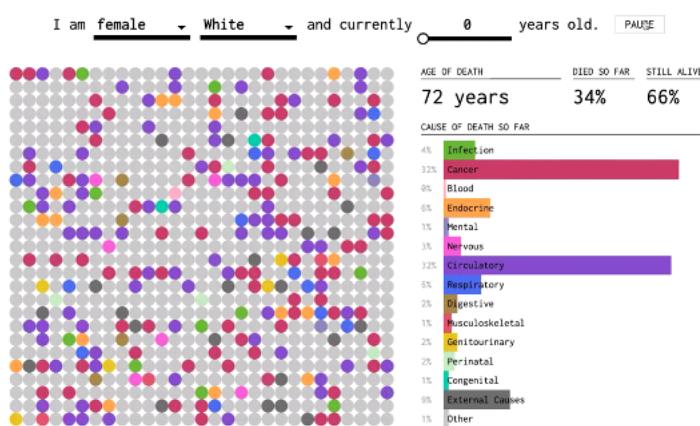
what is happening, over time and across space?

is it explicit, or left to the reader to explore?



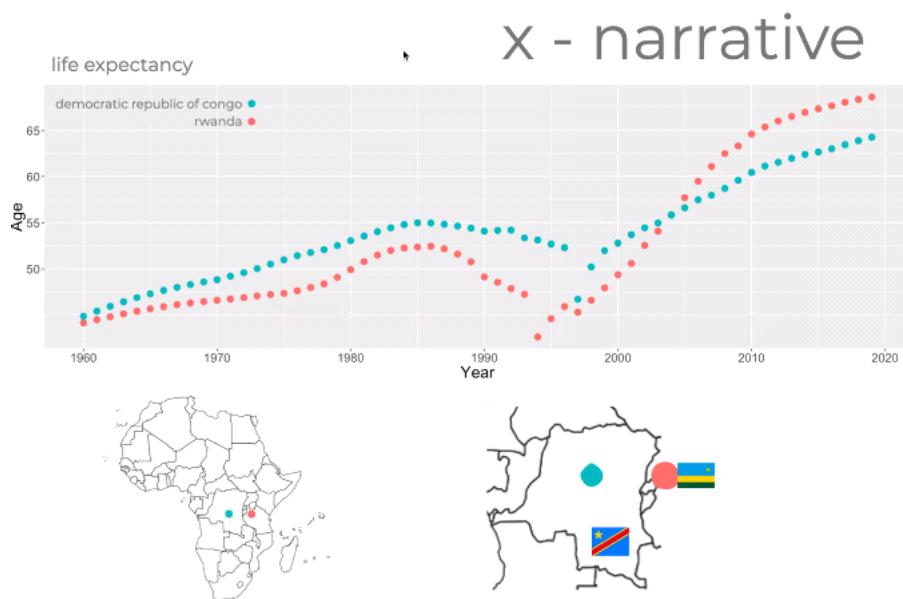
Ritaglio schermata acquisito: 27/04/2021 12:54

Narrative is the story connected to the visualization.



Ritaglio schermata acquisito: 27/04/2021 12:54

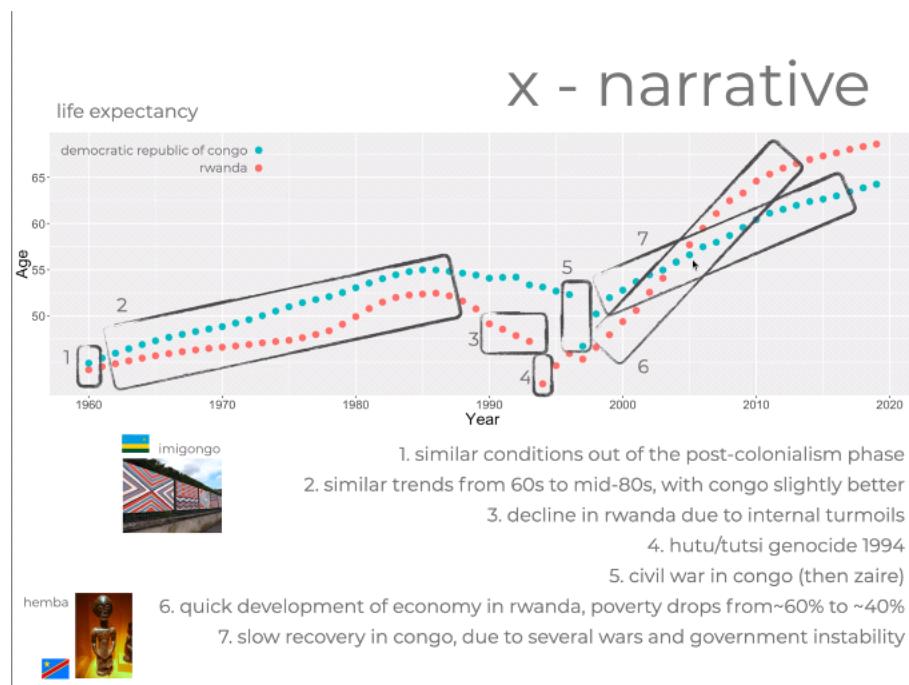
There are many equally valid choices.



Ritaglio schermata acquisito: 27/04/2021 12:57

There is nothing here that helps the reader.

Have to build the narrative considering historic events.



Ritaglio schermata acquisito: 27/04/2021 13:03

Ten rules for data visualization by Tufte are old but still used.

However, there are some new version of these rules.

There is a paper: 10 simple rules for better figures:

- 1) Know your audience. -> providing data viz is not an absolute task.
- 2) Identify your message -> your message can be very simple or very complex depending on the audience.
- 3) Adapt the figure to the support medium. Several journals have both the printed and the paper

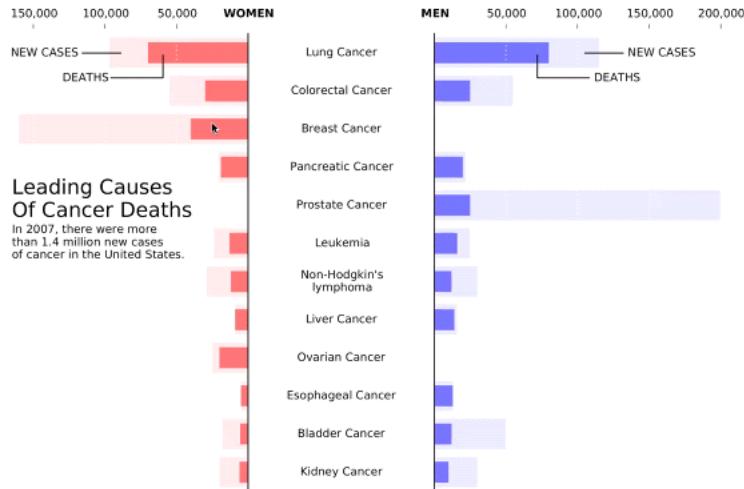
version -> different aspects for data viz.

- 4) Captions are not optional -> they need to be very detailed and self contained especially for scientific publications.
- 5) Do not trust the defaults. Pay attention to what they provide.
- 6) Use colour effectively. You cannot forget about the importance of how you choose your palette and the effect that it has.
- 7) Do not mislead the reader. Do not distract the reader with unnecessary material.
- 8) Avoid chart junk
- 9) Message trumps beauty -> aesthetically pleasant but it does not have to affect the chart
- 10) Get the right tool. -> different tools -> different results.

## modern version

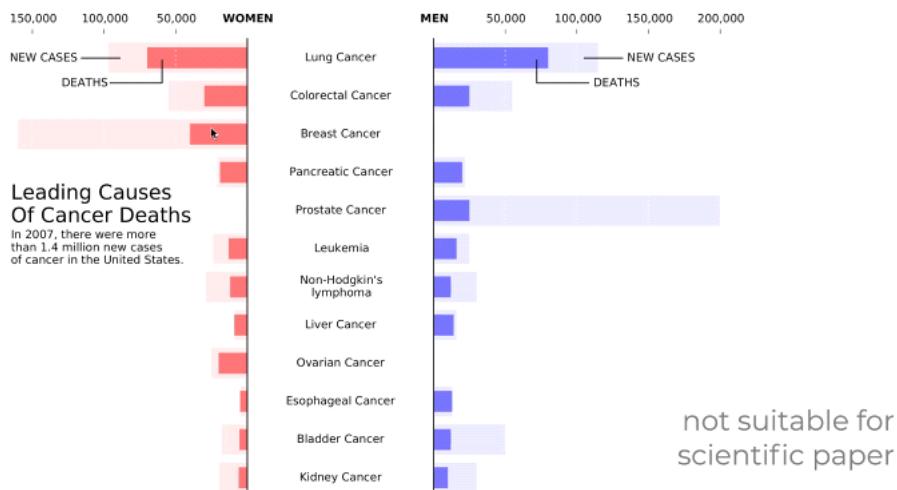
for compbiol... but not so different after all

### i. know your audience



Ritaglio schermata acquisito: 28/04/2021 11:40

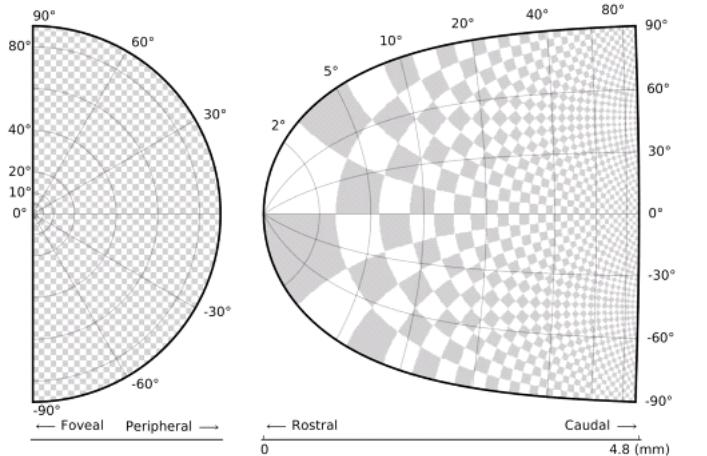
### i. know your audience



Ritaglio schermata acquisito: 28/04/2021 11:41

Why not suitable for scientific paper? It is informative, it tells the story but not the details. In scientific papers you have to be really precise.

## ii. identify your message



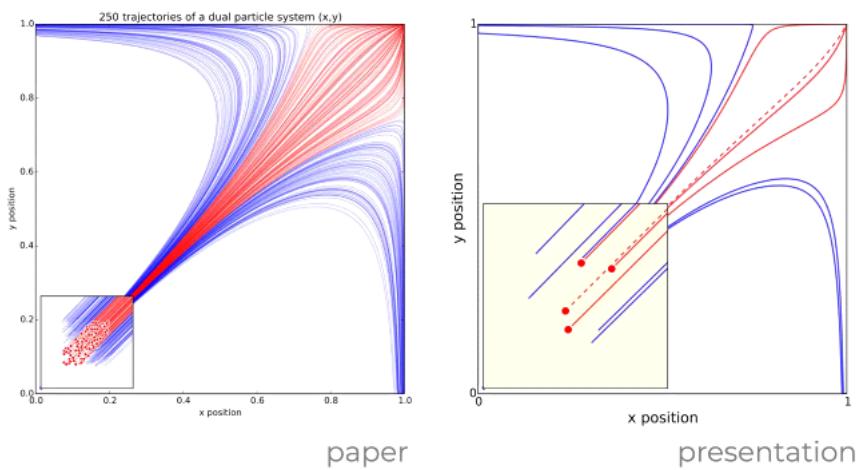
Ritaglio schermata acquisito: 28/04/2021 11:43

Here we have a dataviz, a scientific plot regarding scientific properties of eye sight. The interesting thing here is that the author put a very important detail -> drives the attention of the reader to the most important concept.

The checkerboard pattern highlights the logarithmic mapping involved in the model.

**modern version**  
for compbiol... but not so different after all

## iii. adapt the figure to the support medium



Ritaglio schermata acquisito: 28/04/2021 11:46

These two plots portait the same thing. But one has been put in the paper the other in the presentation.

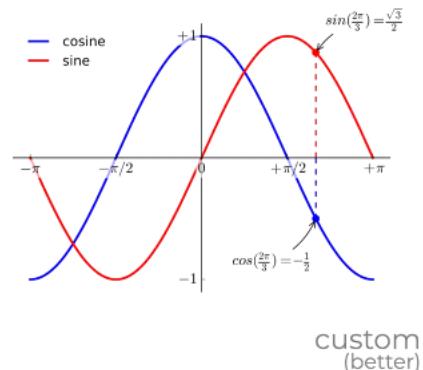
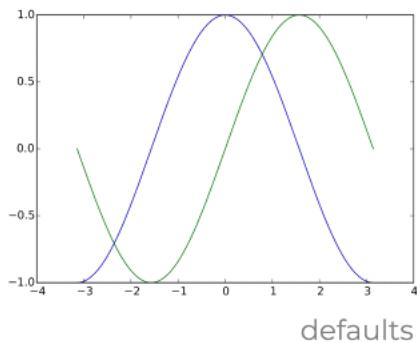
In the paper you can insert all the information the reader can spend more time analyzing and understanding it.

when you are putting a presentation -> you have to convey the same message in few seconds. The audience have to understand and interact with the figure. You need to be very selective and present the main overall effect of what you are presenting. You have to be impactive, efficient.

# modern version

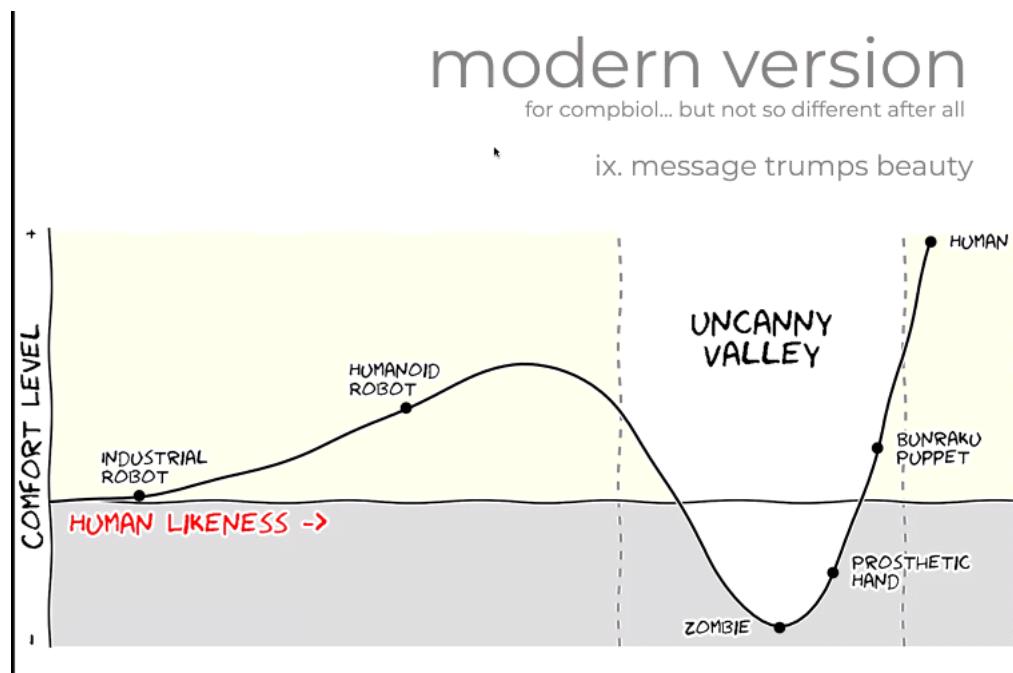
for compbiol... but not so different after all

v. do not trust the defaults



Ritaglio schermata acquisito: 28/04/2021 11:51

Customize all the details. We have a lenged, centered axes with the right labels, a focus on points with the same ascissa.



Ritaglio schermata acquisito: 28/04/2021 11:55

Rough description of the evolution between the human likeness of some robots with the confort level in dealing with them.

This comes from a scientific comic.

But this is not very high levels -> it looks almost human-made. For example the unprecise way the lines is plotted, the absence of numerical values on the axes.

But the message is really well transmitted and it does better than a plot with numerical values.

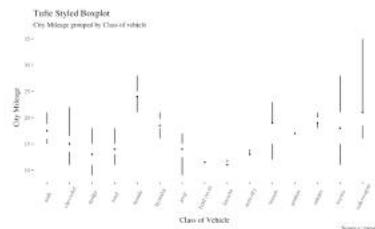
Here the beauty of the sketching part of the plot is transmitted very well.

Questionable aesthetic buy clear message.

# minimalism vs efficiency



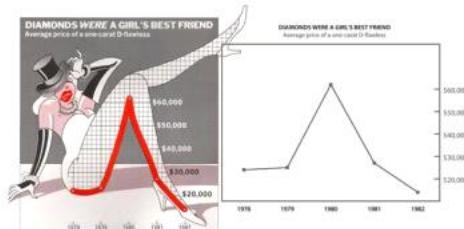
edward tufte



essential & data-centered



nigel holmes



better remembered?

Ritaglio schermata acquisito: 28/04/2021 12:00

Tufte -> minimalism but this is not the only point of view.

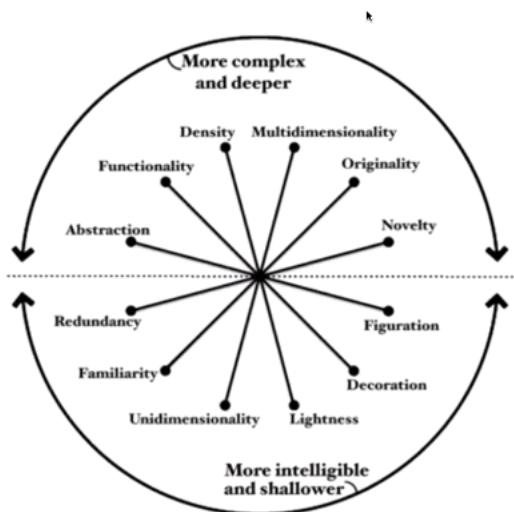
There is indeed, the one of nigel holmes.

For tufte the infographic on the right is a chartjunks -> terrible ink-data ratio.

Nonetheless, the picture on the right of the left part is the rational representation. But which is the graph that you better remember.

## cairo's dataviz wheel

a metaviz to judge dataviz



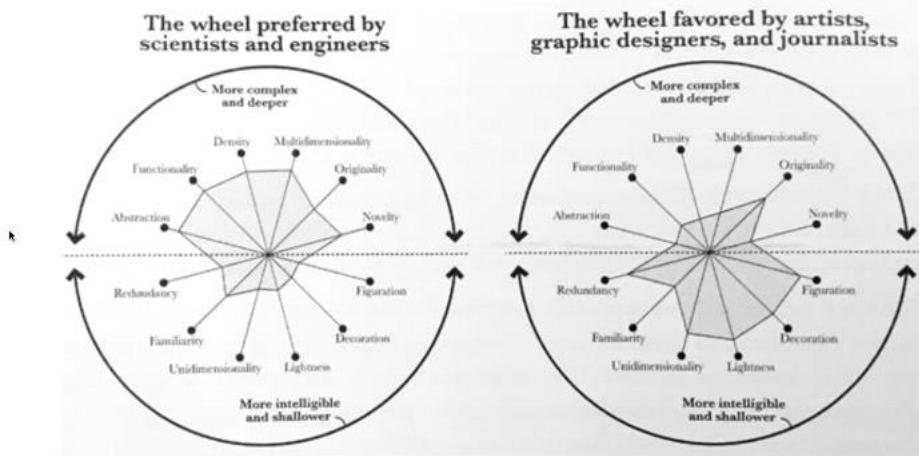
Ritaglio schermata acquisito: 28/04/2021 12:08

Cairo produced many data viz that are cornerstones for dataviz.

We have characteristics to which is associated their opposite e.g. novelty and redundancy. We have these six sticks -> the upper ones are more complex and deeper while the other are more intelligible and shallower representation.

# cairo's dataviz wheel

a metaviz to judge dataviz



Ritaglio schermata acquisito: 28/04/2021 12:11

Originality -> new way to shape your data, something that is not seen so far

Novelty -> use the same data viz to highlight different concepts. Not all elements in data viz are novel.

Horror gallery ?

Very bad representation that made their way to pubblication

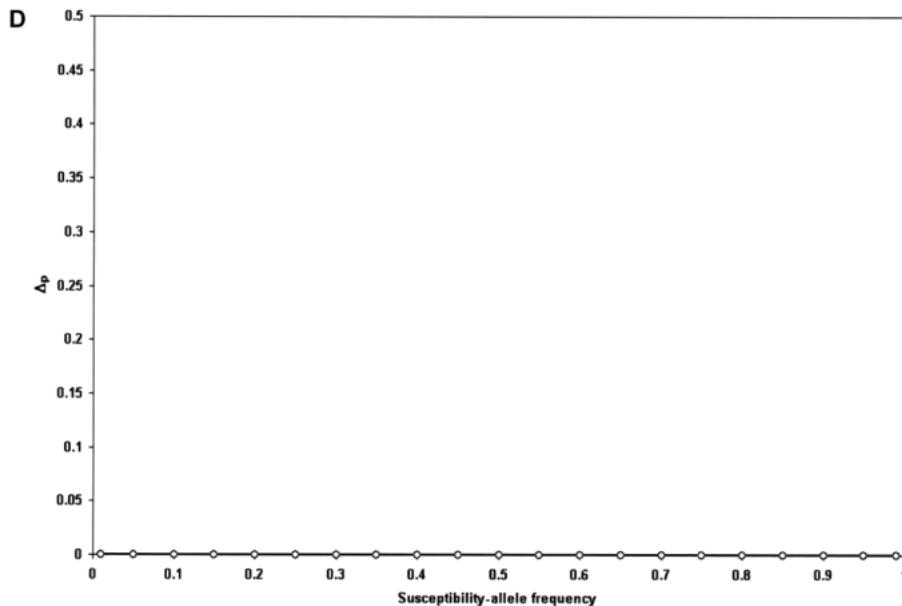
Rational inferences about departures hardy-weinberg equilibrium. This is an example of paper with terrible data viz.

**Figure 1**  $\Delta_p$  plotted versus the susceptibility-allele frequency for patients. A, B, and D, Data points are as follows:  $\gamma = 1.1$  (blackened diamonds),  $\gamma = 1.3$  (unblackened triangles),  $\gamma = 1.5$  (blackened triangles),  $\gamma = 2$  (unblackened diamonds),  $\gamma = 5$  (blackened squares), and  $\gamma = 10$  (unblackened circles). A, Dominant model. B, Recessive model. C, Additive model. Since  $\gamma < 2$  would not satisfy our definition of an additive model as  $\gamma = 2\beta$  and  $\beta > 1$ , the data points in C are as follows:  $\gamma = 2.2$  ( $\beta = 1.1$ ) (blackened diamonds),  $\gamma = 2.6$  ( $\beta = 1.3$ ) (unblackened triangles),  $\gamma = 3$  ( $\beta = 1.5$ ) (blackened triangles),  $\gamma = 5$  (blackened squares),  $\gamma = 2$  (unblackened diamonds). D, Multiplicative model.

Ritaglio schermata acquisito: 28/04/2021 12:18

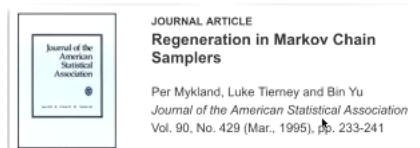
This is the caption. -> we have many glyphs and colors

# horror gallery



Ritaglio schermata acquisito: 28/04/2021 12:20

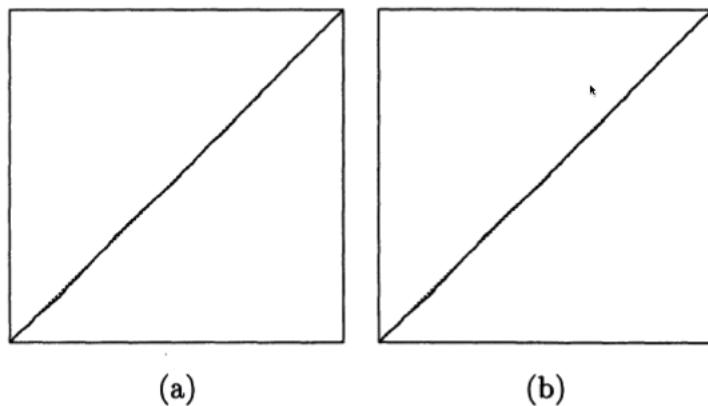
We have one single set of point on the x-axes. This was used to point 10 zeros and was used a whole graph to do it.



Ritaglio schermata acquisito: 28/04/2021 12:22

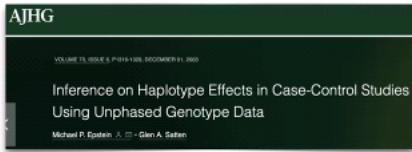
Figure 1. SRQ Plots of  $T_i/T_n$  (Vertical Axes) Against  $i/n$  (Horizontal Axes) for the Gibbs Sampler (a) and an Alternating Gibbs/Independence Sampler (b) for the Pump Failure Data Based on Runs of Length 5,000. Lines through the origin with unit slope are shown dashed; axis ranges are from 0 to 1 for all axes.

Ritaglio schermata acquisito: 28/04/2021 12:22

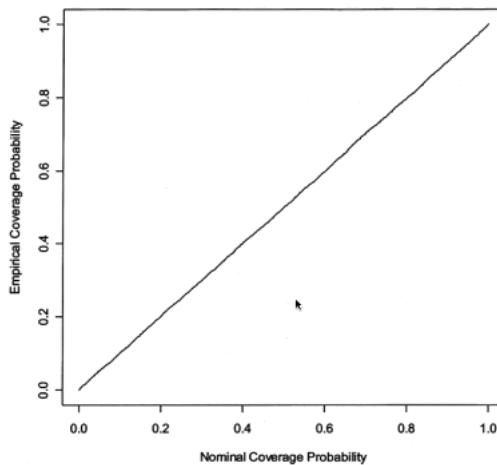


Ritaglio schermata acquisito: 28/04/2021 12:23

These plots are really usefull although a well described caption.

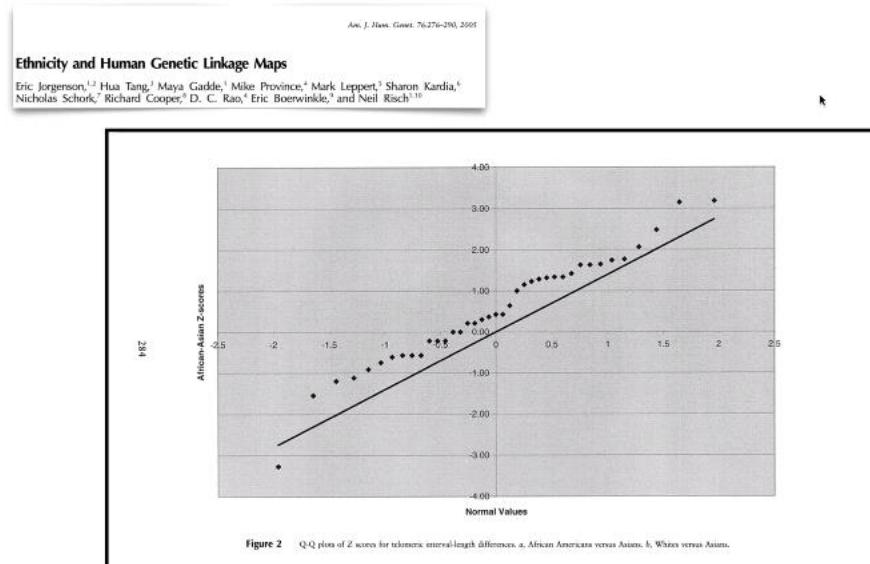


**Figure 1.** Empirical coverage of CIs for the relative-risk parameter  $\beta$  of haplotype 01100. Results are based on 10,000 simulated data sets with the same haplotype frequencies as the FUSION data. Haplotype 01100 has a multiplicative effect on disease risk, with  $\beta=0.35$ .



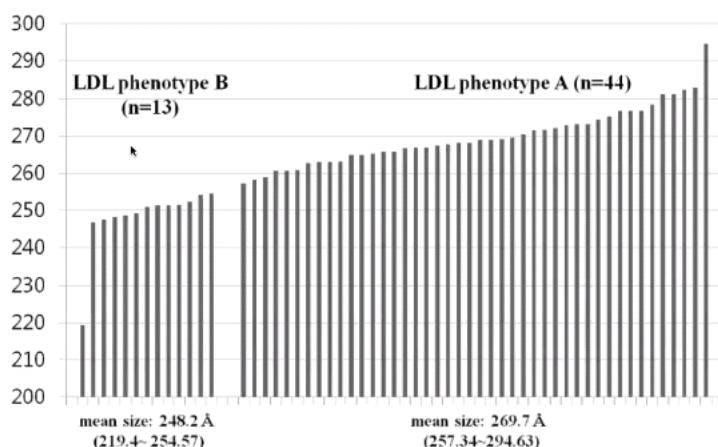
Ritaglio schermata acquisito: 28/04/2021 12:26

Probably a confidence interval would have helped to make this representation more meaningful.



Ritaglio schermata acquisito: 28/04/2021 12:30

**Fig. 1.**  
**Distribution of low-density lipoprotein (LDL) particle size in all study subjects (LDL phenotypes A and B).** *LDL phenotype A group (mean size: 269.7 Å, n = 44), subjects with buoyant-mode profiles [peak LDL particle diameter ≥ 264 Å] including intermediate LDL subclass pattern [256 Å ≤ peak LDL particle diameter ≤ 263 Å]; LDL phenotype B group (mean size: 248.2 Å, n = 13), subjects with dense-mode profiles [peak LDL particle diameter ≤ 255 Å]*

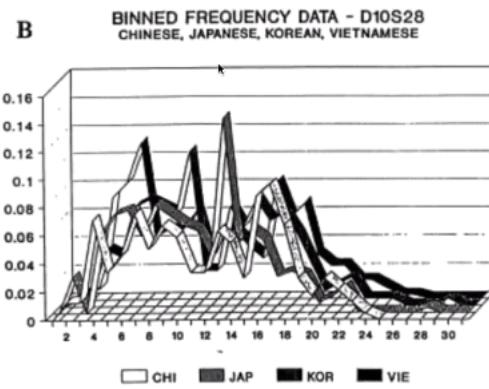


Ritaglio schermata acquisito: 28/04/2021 12:37

Wrong reference point it starts from 200 instead of 0. This accentuate the perceived difference between the two categories.  
 However, putting the actual values at the bottom helps to the readability.  
 We have the real numbers and ranges.



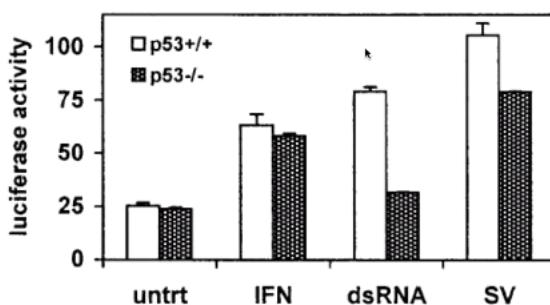
**FIG. 4.** Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are 103, 125, 93 and 215 for D4S139 and 120, 137, 100 and 193 for D10S28. The horizontal axis is the bin number; bins are not of equal length.



Ritaglio schermata acquisito: 28/04/2021 12:42

We have a 3d effect.

This representation does not help the reader to have a better understanding of the data. This effect skews the real slope of the lines. Lastly, it makes things difficult to read.



**FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNAs, and virus.** Cells ( $6 \times 10^3$  HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN- $\alpha$ /ml, 50  $\mu$ g of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20  $\mu$ l of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples  $\pm$  the standard error; the data presented are representative of four independent experiments.

Ritaglio schermata acquisito: 28/04/2021 12:45

Read the last line of the caption.

They use a bar plot to plot three points instead of plotting directly the three points.  
The bar plot hide the real data.

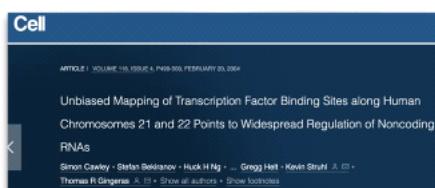
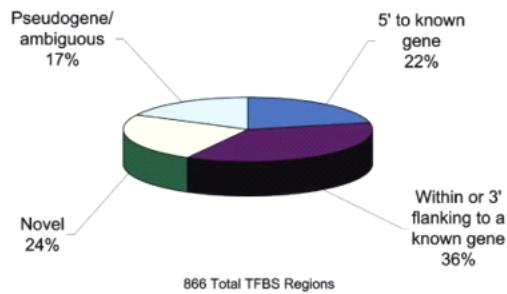


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5' most exon of a gene, within 5 kb of the 3' terminal exon, or within a gene, novel or outside any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

### Distribution of All TFBS Regions

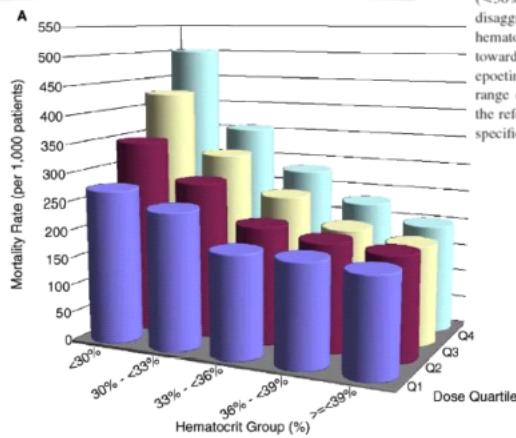


Ritaglio schermata acquisito: 28/04/2021 12:51

The screenshot shows the JCE website with a search bar and navigation links. Below the header, a specific article is highlighted:

Hematocrit was not validated as a surrogate end point for survival among epoetin-treated hemodialysis patients

Dennis J. Cotter\*, Kevin Stefanski\*, Yi Zhang\*, Mai Thamer\*, Daniel Schachter\*, James Kaufman\*



Ritaglio schermata acquisito: 28/04/2021 12:52

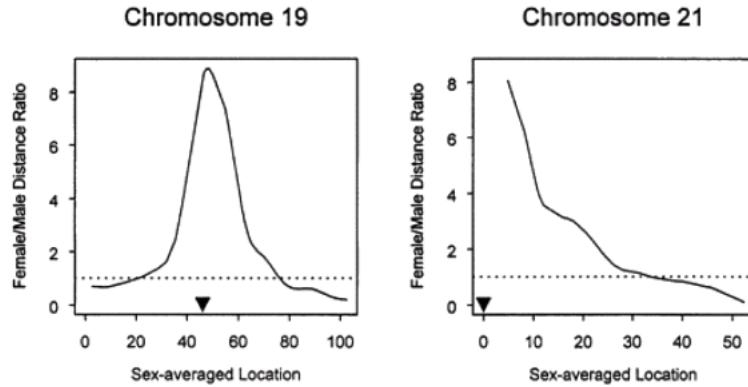
Unwnated 3d effect, perspective, the real trend cannot be see by human eye. This kind of fanciness is hard to be understood.

The screenshot shows the AJHG website with a green header. Below it, a figure is displayed:

Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination

Karl W. Broman, Jeffrey C. Murray, Val C. Sheffield, Raymond L. White, James L. Weber

Figure 1. Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.



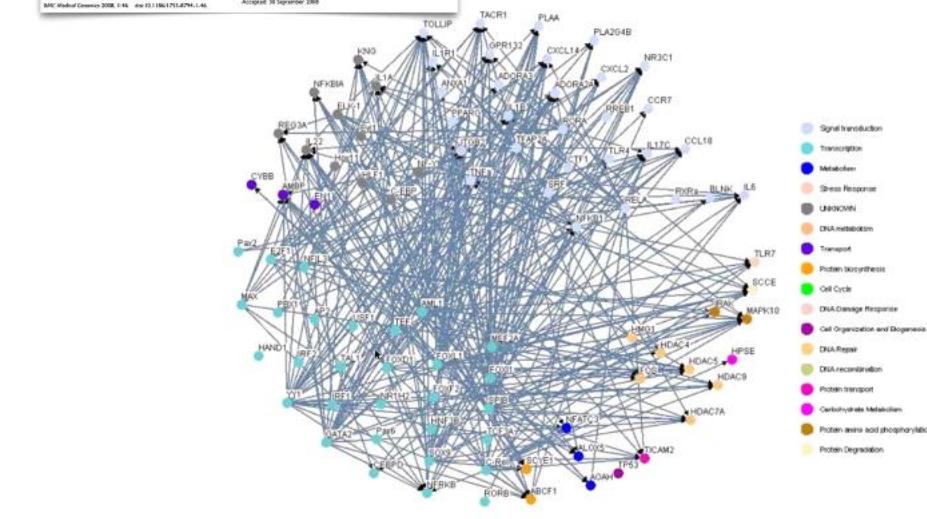
Ritaglio schermata acquisito: 28/04/2021 12:53

They want ot represent the ratio between two numbers.

Matematically, we are representing a ratio. e.g famale values are 8 times the value of the male. But when the we have male values bigger than femals it is very difficult to understand the difference. e.g. the 1/8 is 0.125 and is less interpretable because is very very small.

# horror gallery

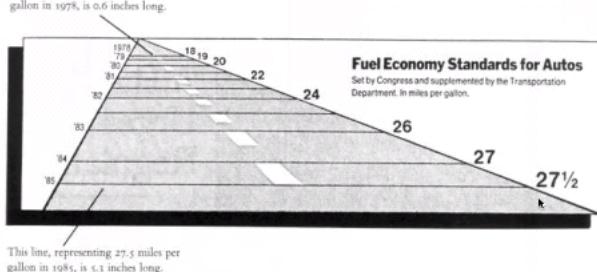
**Figure 3**  
**The inflammatory transcriptional gene network in immune system with LPS.**  
The inflammatory gene network with LPS containing.



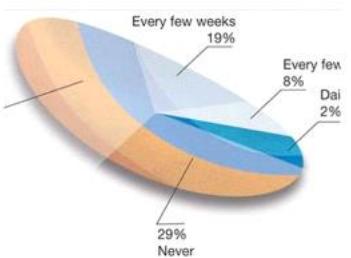
Ritaglio schermata acquisito: 28/04/2021 13:01

# horror gallery

...in infographic

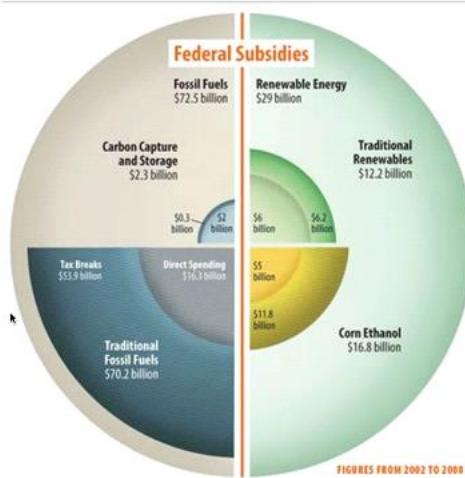
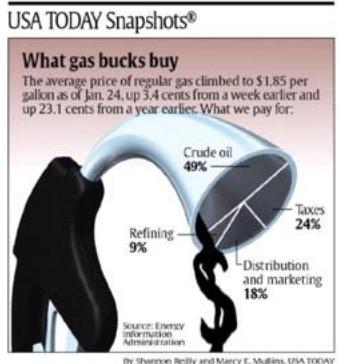


Ritaglio schermata acquisito: 28/04/2021 13:03

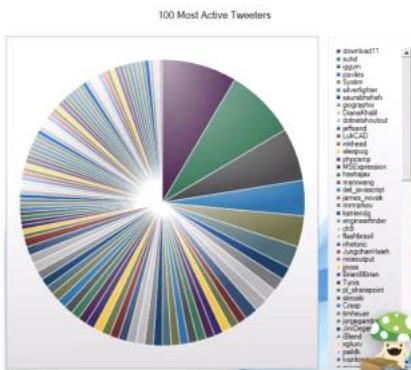


# Horror gallery

...in infographic

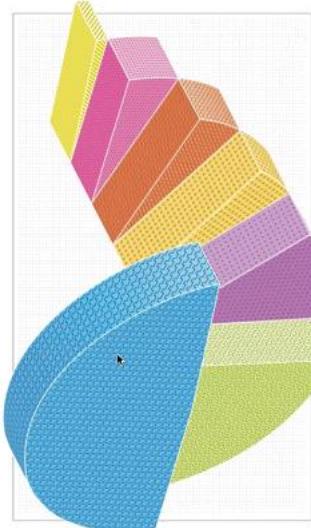
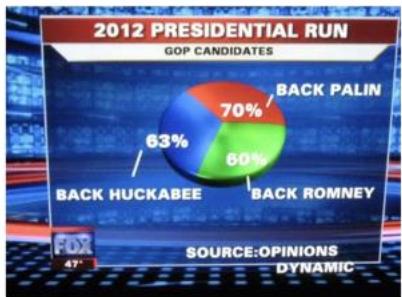


Ritaglio schermata acquisito: 28/04/2021 13:04



# horror gallery

...in infographic



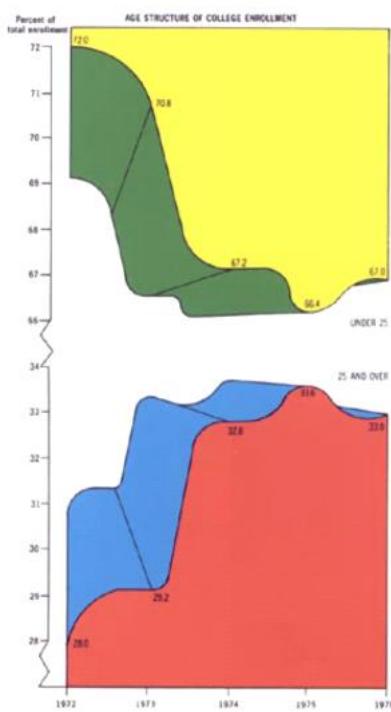
Ritaglio schermata acquisito: 28/04/2021 13:04

#### ...in infographic



Ritaglio schermata acquisito: 28/04/2021 13:06

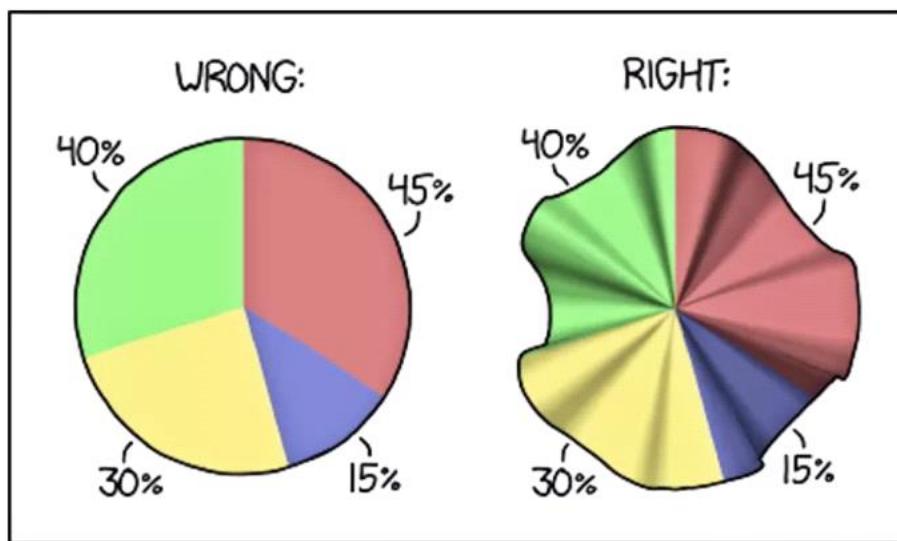
These 4 points are not lying on a straight line. They distorted the y axes to plot these points on the same line. Cheating the reader to convince that all these points define a trend.



Ritaglio schermata acquisito: 28/04/2021 13:08

Age structure of collage enrollment. The two groups are complementary. They just show a percentages. They collected just 5 points.

There is also a break in the axes making the plot out of scale.



HOW TO MAKE A PIE CHART IF YOUR PERCENTAGES DON'T ADD UP TO 100

Ritaglio schermata acquisito: 28/04/2021 13:12



# Dimensionality Reduction

mercoledì 5 maggio 2021 16:42

## overview

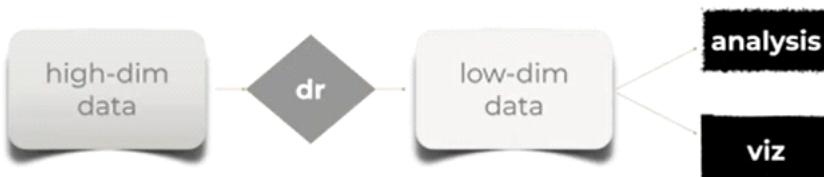
many data analysis pipeline cannot be used for high dimensional data

instability and/or computational issues make them unreliable

need for dimensionality reduction

Ritaglio schermata acquisito: 05/05/2021 16:42

We are talking about images, or text processing, hyperspectral imaging, omics and fluid dynamics.



Ritaglio schermata acquisito: 05/05/2021 16:44

Dr -> dimensionality reduction algorithm.



olivetti face dataset:  
10 people, 100 faces,  
64x64 resolution

## example

imaging

every image with *resolution*  $N=m \times n$  can be transformed into a vector of  $3N$  elements

a dataset with  $k$  images is represented by  
 $K = \{x_1, \dots, x_k\} \subset \mathbb{R}^N$

a possibility is to reduce  $N$  to 2, so the two parameters represents the pose and the face expression

Ritaglio schermata acquisito: 05/05/2021 16:45

Pixel resolution described with RGB vector.



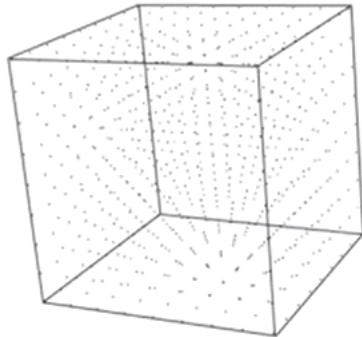
# curse of dimensionality



*adaptive control processes: a guided tour, 1961*

in the absence of simplifying assumptions, the sample size required to estimate a function of several variables to a given degree of accuracy (i.e., to get a reasonably low variance estimate) grows exponentially with the increasing number of variables.

Ritaglio schermata acquisito: 05/05/2021 16:47



grid of spacing  $1/10$  the  $[0,1]^D$  cube  
 $\mathbb{R}^D$  in has  $10^D$  points

Ritaglio schermata acquisito: 05/05/2021 16:47

Multidimensional cube -> d dimensional.

Curse of dimensionality has very aspects -> an important one is the empty space phenomenon which states that:

High-dim spaces are inherently sparse.

in the unit 9-hypersphere  $S^9$  in  $\mathbb{R}^{10}$ ,  $P(d(x,0) \leq 0.9) \sim 0.35$

only 0.02% of the mass of  $N(0,1)$  in  $\mathbb{R}^{10}$  is in  $S^9$

Ritaglio schermata acquisito: 05/05/2021 16:51

# (hyper)-cubes & -spheres

of diameter/side  $2r$

D=1



Ritaglio schermata acquisito: 05/05/2021 16:51

# (hyper)-cubes & -spheres

of diameter/side  $2r$

D=1

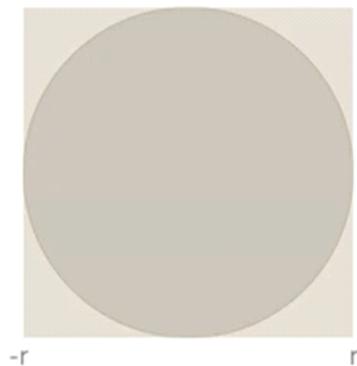


$$V_s^D = 2r \quad V_c^D = 2r$$

Hyper spheres -> all the points in the space whose distance from the origin is in absolute value less than one.

# (hyper)-cubes & -spheres

of diameter/side  $2r$

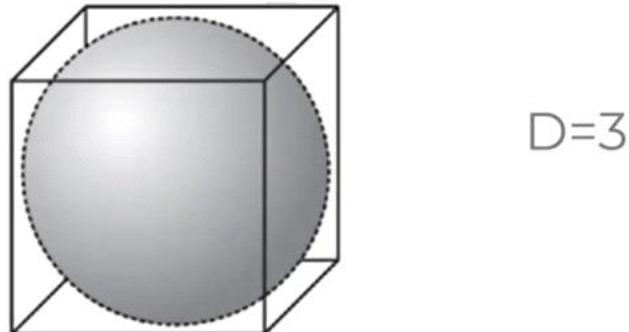


D=2

Ritaglio schermata acquisito: 05/05/2021 16:52

$$V_{\mathbf{S}}^D = \pi r^2 \quad V_{\mathbf{C}}^D = 4r^2$$

Ritaglio schermata acquisito: 05/05/2021 16:52



$$V_{\mathbf{S}}^D = \frac{4}{3}\pi r^3 \quad V_{\mathbf{C}}^D = 8r^3$$

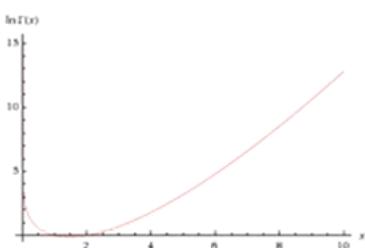
Ritaglio schermata acquisito: 05/05/2021 16:53

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

$$\Gamma(n) = (n-1)!$$

$$\Gamma\left(n + \frac{1}{2}\right) = \binom{(n - \frac{1}{2})}{n} n! \sqrt{\pi}$$

D



$$V_{\mathbf{S}}^D = \frac{\sqrt{\pi^D} r^D}{\Gamma\left(\frac{D}{2} + 1\right)}$$

$$V_{\mathbf{C}}^D = (2r)^D$$

Ritaglio schermata acquisito: 05/05/2021 16:56

Gamma function -> analytical extension of the factorial

$$\lim_{D \rightarrow \infty} \frac{V_S^D}{V_C^D} = \lim_{D \rightarrow \infty} \frac{\frac{\sqrt{\pi^D} r^D}{\Gamma(\frac{D}{2} + 1)}}{(2r)^D} = 0$$

Ritaglio schermata acquisito: 05/05/2021 16:56

This means that when we have a cube inclosing the sphere the ration between the cube and the sphere is zero -> the volume of the spheres is much smaller than the one of the cube.  
For increasing dimensions, the volume of the cube concentrates more in its corners and less in the inscribed sphere.  
Dealing with high dimensio introduce high fluctuation.

## (hyper)-spherical shells



$$V_{ss}^D = V_{S_R}^D - V_{S_r}^D$$

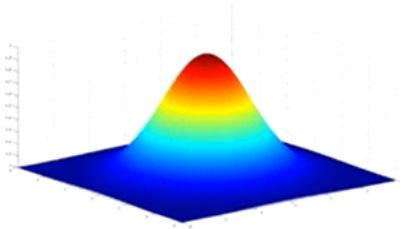
$$\lim_{D \rightarrow \infty} \frac{V_{ss}^D}{V_{S_R}^D} = \lim_{D \rightarrow \infty} 1 - \frac{V_{S_r}^D}{V_{S_R}^D} = \lim_{D \rightarrow \infty} 1 - \left(\frac{r}{R}\right)^D = 1$$

Ritaglio schermata acquisito: 05/05/2021 17:00

As far as the dimension grows all the volume of the sphere gets concentrate on the external shell.  
Virtually all the content of a D dimensional sphere concetrates on its surface, which is only a D-1 dimensional manifold.

# normal distributions

$$N_D(0,1)$$



$$\text{pdf } G(x) = K(r) = \frac{1}{\sqrt{(2\pi)^D}} e^{-\frac{r^2}{2}}, ||x|| = r$$

equiprobable contours  
 $\partial B_r = \partial\{x \in \mathbb{R}^D, ||x|| \leq r\} = S_r$

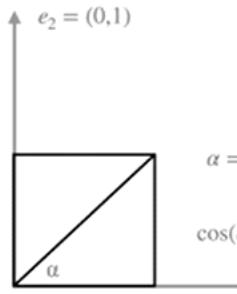
$$P(x \in B_r) = \frac{\int_0^r V(S_r)K(r)dr}{\int_0^\infty V(S_r)K(r)dr}$$

$n$	1	2	5	10	20	100
Probability	0.04550	0.13534	0.54942	0.94734	0.99995	1.00000

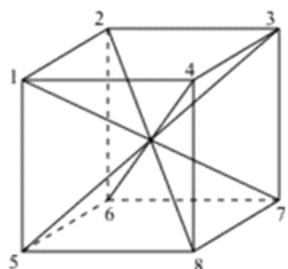
$r > 2$

the probability mass of a multivariate gaussian rapidly migrates into the tails, as the dimension increases

Ritaglio schermata acquisito: 05/05/2021 17:04



diagonals



D-dim hypercube

$$d = (\pm 1, \pm 1, \dots, \pm 1)$$

$$e_i = (\delta_{ij})_{j=1}^D$$

$$\cos(\alpha) = \left\langle \frac{d}{||d||}, e_i \right\rangle = \frac{\pm 1}{\sqrt{D}} \rightarrow 0 \text{ for } D \rightarrow \infty$$

the diagonals are nearly orthogonal to all coordinate axes for large D.

Ritaglio schermata acquisito: 05/05/2021 17:06

Cosine of this angle is 0 means that the angle is a right angle  $\rightarrow$  D ei are orthogonal. When the dimension grows all the diagonal are nearly orthogonal to all coordinate axes.

# concentration of norms

**Theorem 1.1.** Let  $\mathbf{x} = [x_1, \dots, x_D]^T \in \mathbb{R}^D$  be a random vector whose components  $x_k, 1 \leq k \leq D$ , are independent and identically distributed (i.i.d.), with a finite eighth order moment. Let  $\mu = E[x_k]$  be the common mean of all components  $x_k$  of  $\mathbf{x}$ , and  $\mu_j = E[(x_k - \mu)^j]$  be their common central  $j^{\text{th}}$  moment. Then the mean  $\mu_{\|\mathbf{x}\|}$  and the variance  $\sigma_{\|\mathbf{x}\|}^2$  of the Euclidean norm of  $\mathbf{x}$  are

$$\begin{aligned}\mu_{\|\mathbf{x}\|} &= \sqrt{aD - b} + O(D^{-1}), \\ \sigma_{\|\mathbf{x}\|}^2 &= b + O(D^{-1/2}),\end{aligned}\quad (1.4)$$

respectively, where  $a$  and  $b$  are the parameters defined by

$$\begin{aligned}a &= \mu^2 + \mu_2, \\ b &= \frac{4\mu^2\mu_2 - \mu_2^2 + 4\mu\mu_3 + \mu_4}{4(\mu^2 + \mu_2)}.\end{aligned}$$

[demartines, 1994]

means grow as  $O(D^{0.5})$ ,  
variances grow as  $O(1)$



random vectors are  
nearly normalized

$$P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \Rightarrow \quad P(||x|| - \mu_{\|x\|} \geq \varepsilon) \leq \frac{\sigma_{\|x\|}^2}{\varepsilon^2} \quad \text{independent of } D$$

Chebyshev's inequality

$$\text{for large } D \quad ||x|| \sim \mu_{\|x\|} = r \quad \Rightarrow$$

all random vectors are close to  
the surface of a sphere of a  
sphere of radius  $r$

euclidean distance between any two vectors  
is approximately constant.

Ritaglio schermata acquisito: 05/05/2021 17:09

## intrinsic/extrinsic dim

the points of high-dimensional data usually reside  
on a much low-dimensional manifold

$$X = \{x_\alpha\}_{\alpha \in A}, x_\alpha \in M, \dim(M) = s, M \subset \mathbb{R}^D$$

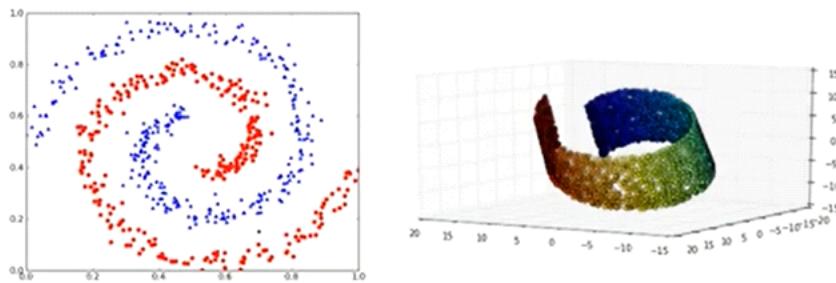
D extrinsic dimension of X, s intrinsic dimension of X

X sample set of random vectors  $\mathbf{X}$  in dimension D

there exist a random vector  $\mathbf{Y}$  in dim s and an  
invertible analytic function  $f$  s.t.  $f(\mathbf{Y}) = \mathbf{X}$

Ritaglio schermata acquisito: 05/05/2021 17:11

# intrinsic/extrinsic dim



due to the low intrinsic dimension of data, we can reduce the (extrinsic) dimension without losing much information for many types of real-life high-dimensional data, avoiding many of the curses of dimensionality

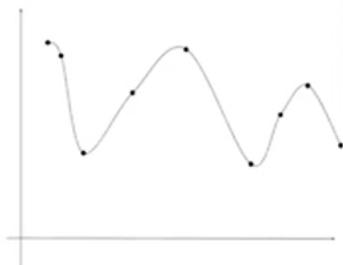
dr is finding a parameterization of the manifold which the points of data reside on

Ritaglio schermata acquisito: 05/05/2021 17:12

Estimate the intrinsic dimension.

# intrinsic dim extimate

a finite data set can be embedded in many manifolds of various dimensions, depending on the geometric structures defined on the data



Ritaglio schermata acquisito: 05/05/2021 17:13

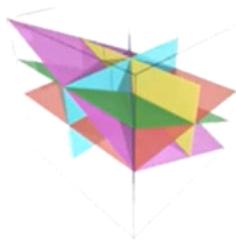
By unisolvence theorem, the dataset here on the plane can be embedded in a regular curve.

Two cases: linear case and non linear case.

# linear case

simplest manifold: linear subspace — hyperplane

among all hyperplanes  $S$  including the dataset  $X$ ,  
choose the one with lowest dimension  $s$



there exists a linear transformation  $T: \mathbb{R}^s \rightarrow S$

$$x = T(y) = Uy + x_0$$

where  $U$  is orthogonal ( $\langle u, v \rangle = \langle Uu, Uv \rangle$ ) and  $x_0$  is  
the nearest point on  $S$  to the origin

the inverse of  $T$  provides a  $s$ -dim parameterisation of  $X$

$$Y = \{y \in \mathbb{R}^s : y = T^{-1}(x), x \in X\}$$

wlog  $x_0=0$  and if  $U=[u_1, \dots, u_s]$ , then  $\{u_1, \dots, u_s\}$  is an orthonormal basis of  $S$

Ritaglio schermata acquisito: 05/05/2021 17:16

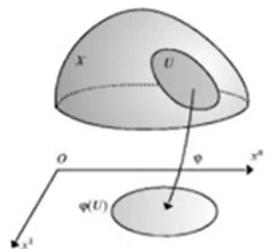
Orthonormal  $\rightarrow$  each couple of columns is orthogonal and their products is 1.  
Distances are preserved and we have an example of linear dimensionality reduction.

# nonlinear case

if the underlying geometry is non linear, the  
previous approach does not give a true dimension

we then use the topological manifold theory

if a manifold  $M$  in  $\mathbb{R}^D$  has dimension  $s < D$ , then the  
neighbourhood of each point of  $M$  is isomorphic to  $\mathbb{R}^s$ ,  
i.e. there is an invertible differentiable map from  $M$  to  
 $\mathbb{R}^s$  whose inverse is differentiable.



Ritaglio schermata acquisito: 05/05/2021 17:18

# nonlinear case

a set  $U$  is open if each point of  $U$  has a neighbourhood included in  $U$  — e.g. open intervals in  $\mathbb{R}$ , or disks without boundary of  $\mathbb{R}^2$

given a space  $S$ , the open covering of  $X$  in  $S$  is a collection  $O$  of open sets in  $S$  whose union contains  $X$

a refinement of  $O$  is another covering  $O'$  such that each set in  $O'$  is a subset of some set in  $O$

an  $s$ -dimensional set  $X$  can be covered by open spheres such that each point belongs to at most  $s + 1$  open spheres.

a subset  $X$  of a topological space  $S$  is said to have topological dimension  $s$  (also called Lebesgue covering dimension) if every covering  $O$  of  $X$  has a refinement  $O'$  such that every point of  $X$  is covered by at most  $s + 1$  open sets in  $O'$ , and  $s$  is the smallest among such integers.

Ritaglio schermata acquisito: 05/05/2021 17:19

# dr methods

in most dimensionality reduction problems, the output data is not required to have the intrinsic dimension, say,  $s$ , but a target dimension  $d$  that is lower than the intrinsic one,  $d < s$ .

## **linear**

output data are a projection of the original data  
*principal component analysis*  
*multidimensional scaling*

## **non-linear**

output data are the manifold coordinate representation of the original data  
*isomap*  
*t-sne, umap*

Ritaglio schermata acquisito: 05/05/2021 17:20



Ritaglio schermata acquisito: 05/05/2021 17:21

## PCA AND MDS

### REAL VECTOR SPACES

# real vector spaces

#### *real vector space*

set  $V$  of elements called vectors close under vector sum  
and scalar multiplication for elements in  $\mathbb{R}$   
basic example:  $\mathbb{R}^n$

#### *linear combination*

$a_1 v_1 + a_2 v_2 + \dots + a_n v_n$  for  $a_i \in \mathbb{R}$  and  $v_i \in V$

#### *linear independent vectors*

set  $M = \{m_1, \dots, m_q\}$  of elements of  $V$  such that no vector in  $M$   
can be written as linear combination of the others

#### *generators*

set  $M = \{m_1, \dots, m_p\}$  of elements of  $V$  such that each element  
of  $V$  can be written as a linear combination of vectors of  $M$

#### *basis*

set  $B = \{b_1, \dots, b_n\}$  of linearly independent generators of  $V$

# real vector spaces

*dimension*  
cardinality of all the bases of V

*example*  
 $V = \mathbb{R}^3$  — basis  $E = \{e_1 = (1,0,0), e_2 = (0,1,0), e_3 = (0,0,1)\}$   
or basis  $G = \{g_1 = (1,2,3), g_2 = (-1,3,0), g_3 = (0,1,-1)\}$

*coordinates*  
each vector in V can be written in a unique way as a linear combination of the basis element; the coefficients are called coordinates:

$$v = (7, -4, 5)_E = 7e_1 + (-4)e_2 + 5(e_3) = \\ (11/4)g_1 + (-17/4)g_2 + (13/4)g_3 = (11/4, -17/4, 13/4)_G$$

*linear transformation*  
map  $T: V \rightarrow W$  such that  $T(u+v) = T(u) + T(v)$  and  
 $T(kv) = kT(v)$  for  $u, v \in V$  and  $k \in \mathbb{R}$

Ritaglio schermata acquisito: 05/05/2021 17:34

## matrices

*basis change*  
 $B_1 = \{v_1, \dots, v_n\}, B_2 = \{w_1, \dots, w_n\}$  bases of V

$$w_j = \sum_{i=1}^n a_{ij} v_i \quad \text{if } v = \sum_{i=1}^n x_i v_i = \sum_{i=1}^n y_i w_i \quad X = AY \quad A = (a_{ij})$$

*linear transformation*  
 $T: V \rightarrow W$  with  $B_v = \{v_1, \dots, v_n\}, B_w = \{w_1, \dots, w_m\}$   
 $T(v_j) = \sum_{i=1}^n a_{ij} w_i$   
then T is represented by the matrix  $A = (a_{ij})$  in  
 $M_{\mathbb{R}}(m, n) = \mathbb{R}^{mn}$   
different bases correspond to different A

*linear transformation*  
 $T: V \rightarrow W$   
 $B_v = \{v_1, \dots, v_n\}, B_w = \{w_1, \dots, w_m\}, B'_v = \{v'_1, \dots, v'_n\}, B'_w = \{w'_1, \dots, w'_m\}$   
P matrix  $B_v \rightarrow B'_v$ , Q matrix  $B_w \rightarrow B'_w$   
 $A' = Q^{-1} A P$

Ritaglio schermata acquisito: 05/05/2021 17:36

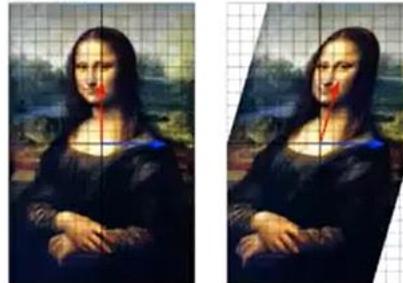
# eigen\*

linear transformation

$T:V \rightarrow V$

if  $T(v) = \lambda v$ ,  $v$  is an eigenvector and  $\lambda$  is an eigenvalue;  
as matrices,  $(A - \lambda I)v = 0$

*geometric interpretation*  
an eigenvector, corresponding to a real nonzero eigenvalue, points in a direction that is stretched by the transformation and the eigenvalue is the factor by which it is stretched; if the eigenvalue is negative, the direction is reversed.



shear mapping  $T((x,y)) = (x+ay, y)$

Ritaglio schermata acquisito: 05/05/2021 17:38

## orthogonality

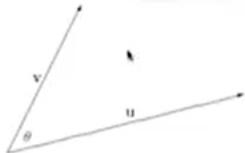
the generalisation of the notion of *perpendicularity*

bilinear forms

$B$  is a bilinear form on a vector space  $V$  if  $B:V \times V \rightarrow K$   
if it is linear in each component

orthogonality

to vectors  $v$  and  $w$  are orthogonal w.r.t.  
a bilinear form  $B$  if  $B(v,w) = 0$



example

euclidean space  $(V, B) = (\mathbb{R}^n, \cdot \cdot)$

$v \cdot u = 0 \iff \cos(\theta) = 0$

Ritaglio schermata acquisito: 05/05/2021 17:39

# orthogonality

a bilinear form  $\langle \cdot, \cdot \rangle$  on a real vector space  $V$  that is symmetric  
 $\langle xy \rangle = \langle yx \rangle$   
and positive definite  $\langle x, x \rangle > 0$  if  $x \neq 0$

*inner product*  
 $T: V \rightarrow V$  linear on a inner product space with  $\langle x, y \rangle = \langle T(x), T(y) \rangle$

*dot product*  
in euclidean space,  $\langle \cdot, \cdot \rangle$  is  $\cdot$  and then  $u \cdot v = uv^T$

$$\|w\| = \sqrt{\langle w, w \rangle}; \text{ for } \cdot, \|w\| = \sqrt{w_1^2 + w_2^2 + \dots + w_n^2} \quad \text{norm}$$

*orthogonal basis*  
a set of orthogonal linearly independent vectors that  
generate  $V$  and are orthogonal;  
if  $\|x\|=1$  for each element of the basis, they are called *orthonormal*

Ritaglio schermata acquisito: 05/05/2021 17:41

PCA algorithm -> reduce dimensionality and deletes correlation among the variables.  
The new coordinates -> are the principle components. -> typically the 1st component is the one that  
catches the most of the variability.

## principal component analysis

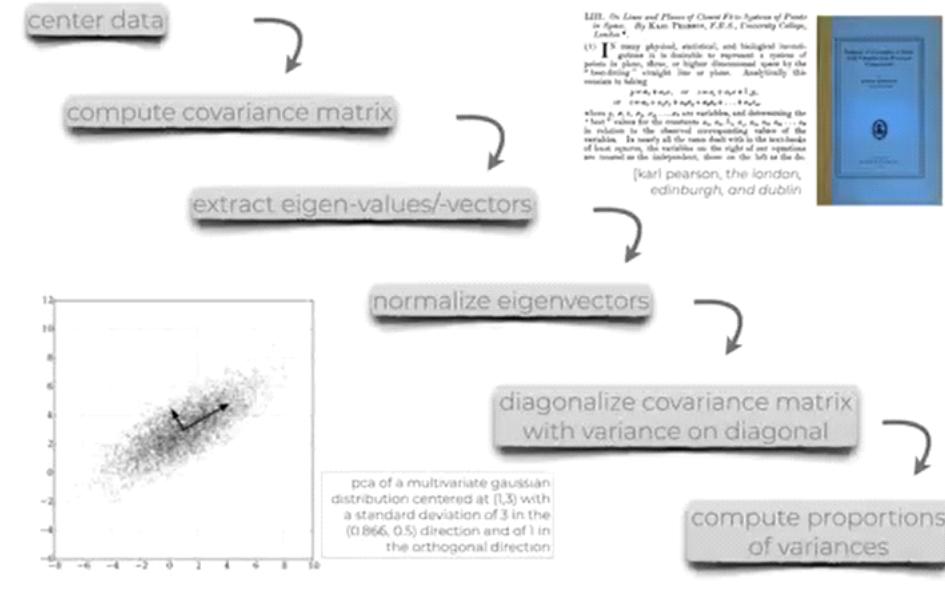
project orthogonally a dataset  $X = \{x_1, \dots, x_n\}$  of  $n$   
 $p$ -dimensional points into a  $r$ -dimensional space  
with  $r = \min(n-1, p)$ , so that in the new coordinates the  
projected points' variables are uncorrelated

the new coordinates are called principal components,  
and each component is defined by the rules:  
• being *orthogonal* to the previous components  
• having highest possible *variance*  
the novel coordinates form an orthogonal basis

pca ~ fitting a  $p$ -dimensional ellipsoid to the data;  
each axis of the ellipsoid is a principal component

Ritaglio schermata acquisito: 05/05/2021 17:44

# scheme



Ritaglio schermata acquisito: 05/05/2021 17:45

## setup

$D = \{z_1, \dots, z_n\}$  dataset of  $n$  samples in  $p$  variables:  $z_i = (z_{i1}, \dots, z_{ip})$

center variables:  $x_{ij} = z_{ij} - (1/n) \sum_i z_{ij}$

obtain  $X$  in  $\mathbb{R}^{n \times p}$  — data matrix columnwise zero centered

goal: transform  $X = \{x_1, \dots, x_n\}$  into a new dataset  $T = \{t_1, \dots, t_n\}$  in  $l$  variables s.t. each sample (row)  $x_i$  is mapped into the sample (row)  $t_i$  by the matrix:  $t_{ik} = x_{ik} \cdot w_{ik}$

$$\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1l} \\ t_{21} & t_{22} & \cdots & t_{2l} \\ \cdots & & & \\ t_{n1} & t_{n2} & \cdots & t_{nl} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1l} \\ w_{21} & w_{22} & \cdots & w_{2l} \\ \cdots & & & \\ w_{p1} & w_{p2} & \cdots & w_{pl} \end{pmatrix}$$

s.t.  $t_{11}, \dots, t_{1l}$  inherit the maximum possible variance from  $X$  and  $w_{ik} = (w_{1k}, \dots, w_{pk})$  has norm one for each  $k=1, \dots, l$

caveat: pca depends on the chosen scaling

Ritaglio schermata acquisito: 05/05/2021 17:47

# 1st component

$$\begin{pmatrix} t_{11} \\ t_{21} \\ \dots \\ t_{n1} \end{pmatrix} = \begin{pmatrix} x_{11}x_{12}\dots x_{1p} \\ x_{21}x_{22}\dots x_{2p} \\ \dots \\ x_{n1}x_{n2}\dots x_{np} \end{pmatrix} \cdot \begin{pmatrix} w_{11} \\ w_{21} \\ \dots \\ w_{p1} \end{pmatrix}$$
$$t = Xw$$

$$\text{Var}(t_{.1}) = \overline{t_{.1}^2} - (\overline{t_{.1}})^2 = \frac{1}{n} \sum_{i=1}^n t_{i1}^2 - \frac{1}{n^2} \left( \sum_{i=1}^n t_{i1} \right)^2 = \frac{1}{n} \sum_{i=1}^n t_{i1}^2 - \frac{1}{n^2} \left( \sum_{i=1}^n \sum_{j=1}^p x_{ij} w_{j1} \right)^2 =$$
$$\frac{1}{n} \sum_{i=1}^n t_{i1}^2 - \frac{1}{n^2} \left( \sum_{j=1}^p \sum_{i=1}^n x_{ij} w_{j1} \right)^2 = \frac{1}{n} \sum_{i=1}^n t_{i1}^2 - \frac{1}{n^2} \left( \sum_{j=1}^p w_{j1} \sum_{i=1}^n x_{ij} \right)^2 = \frac{1}{n} \sum_{i=1}^n t_{i1}^2$$

thus maximising variance means maximising  $\sum_{i=1}^n t_{i1}^2$

Ritaglio schermata acquisito: 05/05/2021 17:50

# 1st component

equation to be solved is thus:

$$w_{.1} = \arg \max_{\|w_{.1}\|=1} \sum_{i=1}^n t_{i1}^2 = \arg \max_{\|w_{.1}\|=1} \sum_{i=1}^n (x_i \cdot w_{.1})^2 = \arg \max_{\|w_{.1}\|=1} \|Xw_{.1}\|^2 = \arg \max_{\|w_{.1}\|=1} w^T X^T X w_{.1}$$

and, since  $\|w\|=1$

$$w_{.1} = \arg \max_{w^T w=1} \frac{w^T X^T X w_{.1}}{w^T w}$$

it is known that the solution correspond to  $w_{.1}$  being the eigenvector associated to the largest eigenvalue

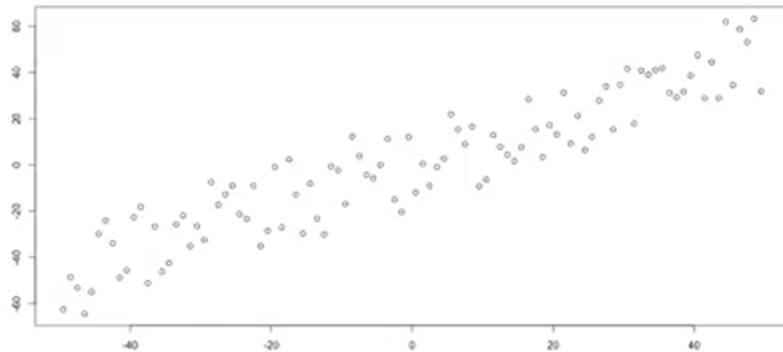
thus the transformed datum is the real number  $t_{i1} = x_i \cdot w_{.1}$  in the new coordinates and the vector  $t_{i1}$  in the old coordinates

Ritaglio schermata acquisito: 05/05/2021 17:51

# 1st component

example

generate a dataset Z with n=100 samples in p=2 dim  
distributed with little noise on y=x



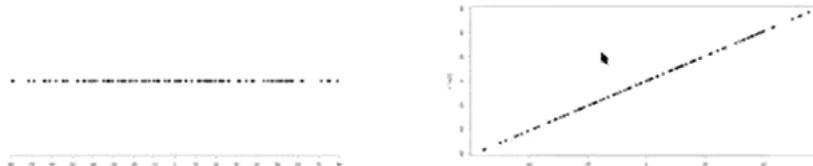
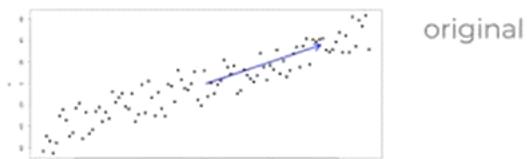
Ritaglio schermata acquisito: 05/05/2021 17:52

The first thing we do is to center both variables to mean zero obtaining new dataset X.  
Then we compute first component s the eigen vector corresponding to largest eigenvalue of  $X^T X$

# 1st component

example

read samples on the w coordinates and on the x coordinates



$$t_{ij} = x_{ij} \cdot w_{j1}$$

$$t_{ij,1}$$

Ritaglio schermata acquisito: 05/05/2021 18:06

# 2nd to last components

subtract all previous components and proceed as in the previous case:

$$\hat{X}_k = X - \sum_{s=1}^{k-1} X w_s w_s^T$$

finding the vector with maximal variance:

$$w_k = \arg \max_{\|w_k\|=1} \| \hat{X}_k w_k \|^2 = \arg \max \frac{w_k^T \hat{X}_k^T \hat{X}_k w_k}{w_k^T w_k}$$

solutions are exactly the remaining eigenvectors of  $X^T X$

the full decomposition is thus given by  $T=XW$  with  $W$  in  $\mathbb{R}^{p \times p}$

thus the transformed datum is the real number  $t_{ij} = x_{ij} \cdot w_j$  in the new coordinates and the vector  $t_{i,1..p}$  in the old coordinates

Ritaglio schermata acquisito: 05/05/2021 18:07

3 methods can be used to compute PCA in practice.

## in practice

three methods are used to compute the pca in practice:

1. covariance

better shen data have similar scales

2. correlation

better shen data have different scales

3. singular value decomposition

more general purpose

Ritaglio schermata acquisito: 05/05/2021 18:07

# explained variance

the first principal component corresponds to a line that passes through the multidimensional mean and minimizes the sum of squares of the distances of the points from the line

each eigenvalue is proportional to the portion of the "variance" (more correctly of the sum of the squared distances of the points from their multidimensional mean) that is associated with each eigenvector

the sum of all the eigenvalues is equal to the sum of the squared distances of the points from their multidimensional mean

explained variance of k-th pca:  $\lambda_k / \sum_i \lambda_i$

pca essentially rotates the set of points around their mean in order to align with the principal components.  
this moves as much of the variance as possible (using an orthogonal transformation) into the first few dimensions

Ritaglio schermata acquisito: 05/05/2021 18:08

# dim reduction

for dimensionality reduction, only the top l components are used

$$T_L = X_L W_L \text{ with } T_L \in \mathbb{R}^{n \times l} \text{ and } W_L \in \mathbb{R}^{p \times l}$$

this maximises the variance in the original data that has been preserved, while minimising the total squared reconstruction error  $\|TW^T - T_LW_L^T\|^2$

$l=2$  finds the 2-dimensional plane through the high-dimensional dataset in which the data is most spread out;  
if the data contains clusters these too may be most spread out,  
and therefore most visible to be plotted

- also appropriate when the variables in a dataset are noisy:
- much of the signal is in the first few principal components
- later principal components may be dominated by noise, and thus discarded

several non-linear extensions of pca exist

Ritaglio schermata acquisito: 05/05/2021 18:09

# a different obj function

pca minimizes low-dimensional reconstruction error

alternative: maximize the scatter of the projection, so to obtain the most informative projection; this is known as classical scaling

given a map  $M \in \mathbb{R}^{d \times r}$ , the functional to be maximised becomes

$$\sum_i \sum_j \|x_i M - x_j M\|^2$$

that can be rewritten as proportional to

$$\text{tr}(M^T X^T X M) - 1 X M M^T X^T = \text{tr}(M^T X^T (I - 1/n 1^T 1) X M) = \text{tr}(M^T X^T X M)$$

because  $X$  has mean zero

thus, scaling is equivalent to pca in the trivial case.

Ritaglio schermata acquisito: 05/05/2021 18:11

## general case

multidimensional scaling is a family of algorithms visualizing the level of similarity of individual cases of a dataset

in practice, mds finds an embedding of  $n$  objects into a  $r$ -dimensional euclidean space  $\mathbb{R}^r$  so to preserve as well as possible (a function of) the distances between original objects

mds can be distinguished in 3 kinds, depending on the objective function:

classical mds - the obj fun is called strain and involves directly the original distances between objects

metric mds - the obj fun is called stress and involves a function of the original distances

non-metric mds - the original distances are dissimilarities, so the stress function finds a non-parametric monotonic relationship between the dissimilarities in the item-item matrix and the Euclidean distances between items, and defines the location of each item

Ritaglio schermata acquisito: 05/05/2021 18:13

# general case

in the classical mds case, the solution is deterministic

the dimensionality reduction core is the same as pca

if there exists a space  $\mathbb{R}^p$  where all the original distances between objects are preserved, the distance  $d$  is called euclidean

for an euclidean distance, the classical mds solution is unique up to isometries

in metric and non-metric mds, the process is given by optimization

Ritaglio schermata acquisito: 05/05/2021 18:24

Published on a website and in *The Journal of Physiology*, 196, 35, 457-474.

DIMENSIONS OF COLOR VISION\*

Department of Psychology, University of Stockholm

GÖTE ECKMAN<sup>a</sup>

A. PEARLINE<sup>b</sup>

There have been many theories concerned with color vision, most of them divide into two or three main categories. No one theory seems to account for all the facts of the field—the laws of color mixing, the types of color blindness, etc.

In recent years physiologists have obtained a promising result with respect to color blindness. The work of Ganz at Salzburg has well shown who are color blind. It clearly shows the existence (in the average) of at least four types (or groups) of receptors, approximately corresponding to the four "natural" colors. These physiological findings will be considered in this paper.

This paper attempts a new psychological approach to the problem of many dimensions of color vision. It is psychological with regard to both problem and method, but no effort will be made to link the results up with the physiological findings.

R. MARSH

The method of similarity analysis was developed by the present writer for studies of the dimensionality of experience. Very briefly, an n-set between can be outlined as follows: stimuli are presented two at a time. The subject is asked to rate, on a scale from zero to one, the degree of similarity between the two stimuli. With a stimuli, a similarity matrix of the order  $n \times n$  will thus be obtained. The entries of this matrix may be individual data or group data.

The measure of the method is based upon the reasonable assumption that the degree of perceived similarity is a function of the degree of overlap between those primary experiences (perceptions, sensations) which are evoked by the stimuli. Under certain conditions the methods of factor analysis may be directly applied to the stimulus matrix. The essence of the resulting factor

\*Received in the Editorial Office on July 1950, and published immediately at the suggestion of the Editor-in-Chief. The author wishes to thank Dr. H. W. Fairhurst for his help in the preparation of the manuscript. This investigation was made possible by a research grant from the National Institute of Mental Health.

## example

eckman's colors

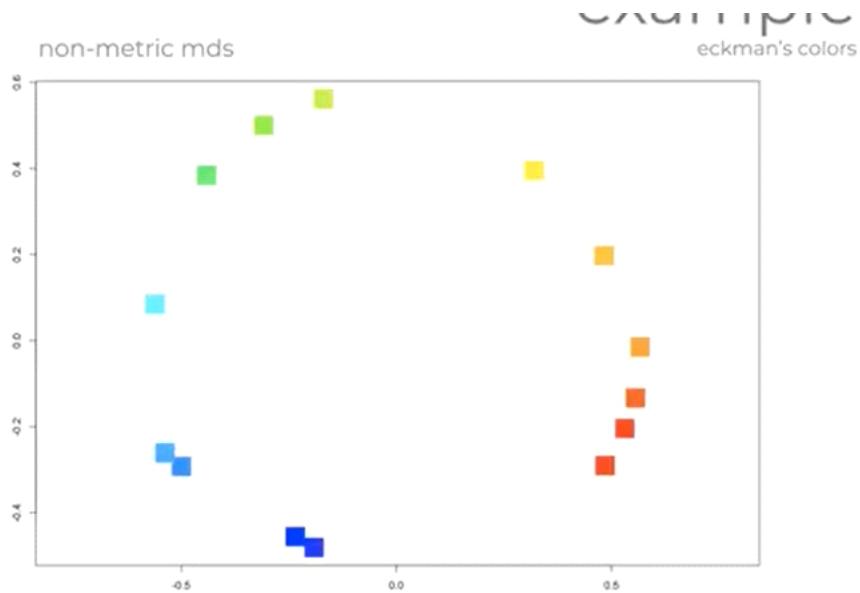
14 colors, different only in hue (wavelength), pairwise rated as similar in a 0-5 scale by 31 people



	434	445	465	472	490	504	537	555	584	600	610	628	651	674
0	0	0.86	0.42	0.42	0.18	0.06	0.07	0.04	0.02	0.07	0.09	0.12	0.13	0.16
0.86	0	0.5	0.44	0.22	0.09	0.07	0.07	0.02	0.04	0.07	0.11	0.13	0.14	
0.42	0.5	0	0.81	0.47	0.17	0.1	0.08	0.02	0.01	0.02	0.01	0.05	0.03	
0.42	0.44	0.81	0	0.54	0.25	0.01	0.09	0.02	0.01	0	0.01	0.02	0.04	
0.18	0.22	0.47	0.54	0	0.61	0.31	0.26	0.07	0.02	0.02	0.01	0.02	0.20	
0.06	0.09	0.17	0.25	0.61	0	0.62	0.45	0.14	0.08	0.02	0.02	0.02	0.01	
0.07	0.07	0.1	0.1	0.31	0.62	0	0.73	0.22	0.16	0.05	0.02	0.02	0.0	
0.04	0.07	0.08	0.09	0.26	0.45	0.73	0	0.33	0.19	0.04	0.03	0.02	0.02	
0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	0	0.58	0.37	0.27	0.2	0.23	
0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	0	0.74	0.5	0.41	0.28	
0.09	0.07	0.02	0	0.02	0.02	0.05	0.04	0.37	0.74	0	0.76	0.62	0.55	
0.12	0.11	0.01	0.01	0.01	0.02	0.03	0.27	0.5	0.76	0	0.85	0.68		
0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.2	0.41	0.62	0.85	0	0.76	
0.16	0.14	0.03	0.04	0	0.01	0	0.02	0.23	0.28	0.55	0.68	0.76	0	

Ritaglio schermata acquisito: 05/05/2021 18:25

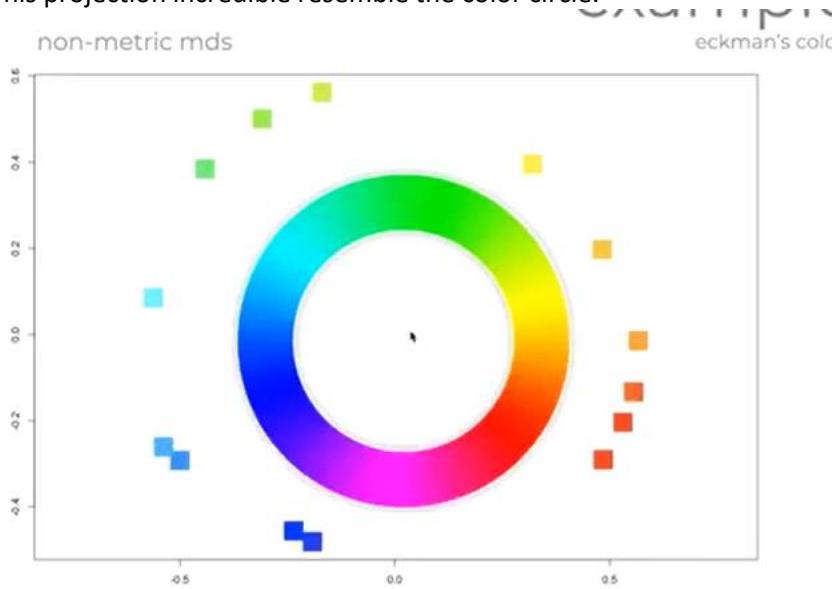
The matrix tells how people perceive the similarities among the colors. What we have is a matrix that summarizes and scales the results into a correlation matrix.



Ritaglio schermata acquisito: 05/05/2021 18:28

Identify on two dimension the projections of these point in a way that the distance shows in the matrix is reported in the graph.

This projection incredible resemble the color circle.



Ritaglio schermata acquisito: 05/05/2021 18:29

# distances

let  $O = \{o_1, \dots, o_n\}$  be a dataset of  $n$  objects with a dissimilarity measure  $d_{ij}$

mds aims at finding  $n$  points  $x_1, \dots, x_n$  in  $\mathbb{R}^p$  so that  $\|x_i - x_j\| = d_{ij}$

- 1.
- 2.
- 3.
- 4.

If  $d$  satisfies:  
 $d(x,y) > 0$  for  $x \neq y$   
 $d(x,y) = 0 \iff x = y$   
 $d(xy) = d(y,x)$  for all  $x,y$   
 $d(xy) \leq d(x,z) + d(z,y)$  for all  $x,y,z$   
the dissimilarity  $d$  is called a distance, or a metric

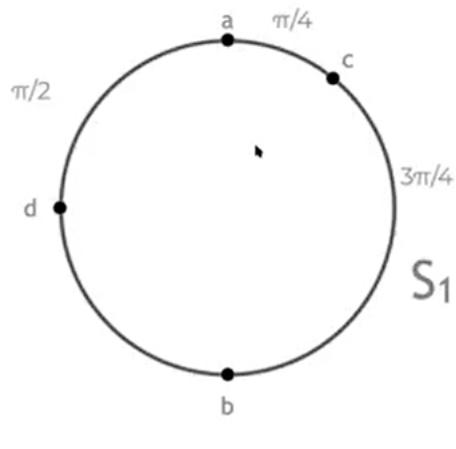
if for some  $p$   $\|x_i - x_j\| = d_{ij}$ ,  $d$  is called an euclidean distance

if no such  $p$  exists but  $d$  is a metric,  $d$  is called a non-euclidean distance

Ritaglio schermata acquisito: 05/05/2021 18:29

# distances

a non-euclidean distance



$d(x,y)$  is the length of the shortest arc connecting  $x$  and  $y$   
(or the absolute value of the smaller arc measure between  $x$  and  $y$  in  $[0,\pi]$ )

not embeddable in any  $\mathbb{R}^p$

	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>
<b>a</b>	0	1	1/4	1/2
<b>b</b>	1	0	3/4	1/2
<b>c</b>	1/4	3/4	0	3/4
<b>d</b>	1/2	1/2	3/4	0

Ritaglio schermata acquisito: 05/05/2021 18:31

We have this matrix  $\rightarrow$  this space is not embeddable in euclidean space.

# classical mds algorithm

suppose  $d_{ij}$  is a (euclidean) distance: find  $X = [x_1, \dots, x_n] \in \mathbb{R}^{n \times q}$  so that  $\|x_i - x_j\| = d_{ij}$

$X$  is not unique: for any  $c \in \mathbb{R}^q$ ,  $X + c$  (rowwise) is also a solution, since  $\|(x_i + c) - (x_j + c)\| = \|x_i - x_j\| = d_{ij}$

usually the centered configuration constraint is added:  
 $\sum_i x_{ik} = 0$  for all columns  $k$  (1)

let  $B = XX^T \in \mathbb{R}^{n \times n}$  be the Gram matrix  
 $b_{ij} = x_{i1}x_{j1} + x_{i2}x_{j2} + \dots + x_{in}x_{jn} = \langle x_i, x_j \rangle$

now  $d_{ij}^2 = \|x_i - x_j\|^2 = \langle x_i - x_j, x_i - x_j \rangle = \langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2\langle x_i, x_j \rangle = b_{ii} + b_{jj} - 2b_{ij}$   
thus  $b_{ij} = -1/2 (d_{ij}^2 - b_{ii} - b_{jj})$  (2)

Ritaglio schermata acquisito: 05/05/2021 18:32

# classical mds algorithm

call  $T = \text{tr}(B) = \sum_k b_{kk}$

because of (1),  $\sum_i b_{ij} = \sum_i \sum_k x_{ik} x_{jk} = \sum_k x_{jk} \sum_i x_{ik} = 0$  (3)

by (2),  $\sum_i d_{ij}^2 = \sum_i (b_{ii} + b_{jj} - 2b_{ij}) = \sum_i b_{ii} + \sum_i b_{jj} - 2 \sum_i b_{ij}$   
by (3) =  $T + nb_{jj} - 0$ ,  
thus  $\sum_i d_{ij}^2 = T + nb_{jj}$  (4)

by symmetry  $\sum_j d_{ij}^2 = T + nb_{ii}$  (5)

by (2),  $\sum_j \sum_i d_{ij}^2 = \sum_j \sum_i (b_{ii} + b_{jj} - 2b_{ij}) = \sum_j \sum_i b_{ii} + \sum_j \sum_i b_{jj} - 2 \sum_j \sum_i b_{ij}$   
by (3) =  $nT + nT - 0$ ,  
thus  $\sum_j \sum_i d_{ij}^2 = 2nT$  (6)

Ritaglio schermata acquisito: 05/05/2021 18:34

# classical mds algorithm

now plugging (4), (5) & (6) in (2):

$$\begin{aligned} b_{ij} &= -\frac{1}{2} \left( d_{ij}^2 - b_{ii} - b_{jj} \right) \\ &= -\frac{1}{2} \left( d_{ij}^2 - \frac{\sum_{j=1}^n d_{ij}^2 - T}{n} - \frac{\sum_{i=1}^n d_{ij}^2 - T}{n} \right) \\ &= -\frac{1}{2} \left( d_{ij}^2 - \frac{1}{n} \sum_{j=1}^n d_{ij}^2 - \frac{1}{n} \sum_{i=1}^n d_{ij}^2 + \frac{1}{n} 2T \right) \\ &= -\frac{1}{2} \left( d_{ij}^2 - \frac{1}{n} \sum_{j=1}^n d_{ij}^2 - \frac{1}{n} \sum_{i=1}^n d_{ij}^2 + \frac{1}{n} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 \right) \quad (7) \end{aligned}$$

Ritaglio schermata acquisito: 05/05/2021 18:36

# classical mds algorithm

find now a matrix form for (7) defining the matrices  
 $D_2 = \{d_{ij}\}$  and  $\mathbf{O}$  in  $\mathbb{R}^{n \times n}$  and  $\mathbf{1}$  in  $\mathbb{R}^{n \times 1}$

$$\mathbf{O}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \cdot (1 \cdots 1) = \mathbf{1} \cdot \mathbf{1}^T$$

then (7) reads:  
 $B = (-1/2)(D_2 - (1/n)\mathbf{O}D_2 - (1/n)D_2\mathbf{O} + (1/n)\mathbf{O}D_2(1/n)\mathbf{O})$

that, introducing the centering matrix  $C_n = I_n - (1/n)\mathbf{1}\mathbf{1}^T$ , becomes  
 $B = (-1/2) C_n D_2 C_n \quad (8)$

$C_n$  is called centering matrix because for any  $Y$ ,  $C_n Y C_n$  has rowwise and columnwise sum equal to zero

# classical mds algorithm

but  $B = XX^T$ , so it is real symmetric:  $B = B^T$ ,  
let  $v \neq 0$  be an eigenvector for the eigenvalue  $\lambda \in \mathbb{C}$ , so  $Bv = \lambda v$

$$\begin{aligned} \bar{v}^T B v &= \bar{v}^T (Bv) = \bar{v}^T (\lambda v) = \lambda (\bar{v}^T \cdot v) \\ &= \\ \bar{v}^T B v &= (B\bar{v})^T v = (\bar{\lambda} \bar{v})^T v = \bar{\lambda} (\bar{v}^T \cdot v) \end{aligned}$$

then  $\lambda$  is real; let  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$  the  $n$  real eigenvalues of  $B$  and  $z_1, \dots, z_n$  the corresponding eigenvectors

then  $B$  can be *diagonalised*, i.e. written as  
 $B = V \Lambda V^T$   
where  $\Lambda$  is the diagonal matrix of the eigenvalues,  
and  $V$  is the orthogonal matrix ( $V^{-1} = V^T$ ) with the corresponding eigenvalues as columns.

# classical mds & dr

thus we have  
 $(-1/2) C_n D_{(2)} C_n = B = X X^T = V \Lambda V^T$

which yields  
 $X = V \Lambda^{1/2}$

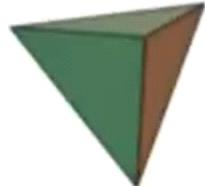
the solution is then unique, and the above algorithm is constructive  
 for non-euclidean distances, only approximation if possible

for dimensionality reduction, since the first  $p < q$  components best  
 preserve the distances among all other  $p$ -dim reductions, one can use

$$X_{(p)} = V_{(p)} \Lambda_{(p)1/2}$$

where  $\Lambda_{(p)1/2}$  is the  $p \times p$  diagonal matrix with the  $p$  largest  $\sqrt{\text{eigenvalues}}$   
 of  $B$ , and  $V_{(p)}$  is collected through the first  $p$  columns of  $V$

Ritaglio schermata acquisito: 05/05/2021 18:39



## example 1

unitary tetrahedron

distance matrix of the four vertices

$$B = (1/8) \cdot \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

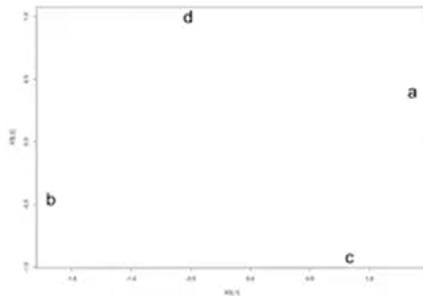
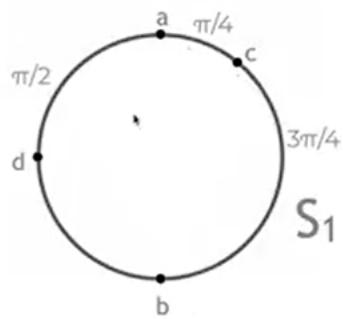
$$\Lambda = (1/2) \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

restricting to  $\mathbb{R}^3$

$$X_{(3)} = V_{(3)} \Lambda_{(3), \frac{1}{2}} = \begin{pmatrix} 0.0000000 & 0.6123724 & 0.0000000 \\ -0.1893048 & -0.2041241 & 0.5454329 \\ -0.3777063 & -0.2041241 & -0.4366592 \\ 0.5670111 & -0.2041241 & -0.1087736 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

A,B,C,D are exactly the vertices of regular tetrahedron of edge 1 in  $\mathbb{R}^3$

Ritaglio schermata acquisito: 05/05/2021 18:40



## example 2

circular distance

	a	b	c	d
a	0	1	1/4	1/2
b	1	0	3/4	1/2
c	1/4	3/4	0	3/4
d	1/2	1/2	3/4	0

D=π

spec(B)={5.6117, 2.2234, 0, -1.2039)  
thus we need to restrict to X\_{(2)}

	a	b	c	d
a	0	1.00	0.46	0.63
b	1.00	0	0.81	0.59
c	0.46	0.81	0	0.75
d	0.63	0.59	0.75	0

dist=π

Ritaglio schermata acquisito: 05/05/2021 18:43

## alternative mds algorithms

relax the objective  $\|x_i - x_j\| = d_{ij}$  to  $\|x_i - x_j\| = f(d_{ij})$

metric/non-metric mds depending on d being quantitative or not (e.g. ordinal)

becomes an optimisation process aimed at minimising a stress function, solved by iterative algorithms

Ritaglio schermata acquisito: 05/05/2021 18:44

# metric mds

given a dimension  $p$ , and a monotone function  $f$ , metric mds aims at finding  $X = \{x_1, \dots, x_n\}$  such that  $\|x_i - x_j\| \approx f(d_{ij})$  as close as possible, i.e. minimising a chosen loss function (stress)  $\mathcal{L}$

example 1:

$$f(d_{ij}) = \alpha d_{ij} + \beta \text{ and } \mathcal{L} = (\sum_{i < j} (\|x_i - x_j\| - (\alpha d_{ij} + \beta))^2 / \sum_{i < j} d_{ij}^2)^{1/2}$$

note that  $\alpha=1$  and  $\beta=0$  gives the classical case, with a different solution

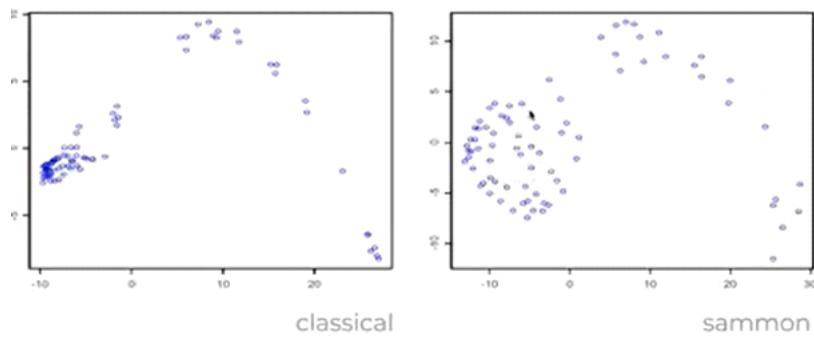
example 2: sammon mapping

$$\mathcal{L} = \frac{1}{\sum_{l < k} d_{lk}} \sum_{i < j} \frac{(\|x_i - x_j\| - d_{ij})^2}{d_{ij}}$$

sammon mapping preserves the small  $d_{ij}$ , giving them a greater degree of importance in the fitting procedure than for larger values of  $d_{ij}$

Ritaglio schermata acquisito: 05/05/2021 18:44

## sammon mapping



sammon mapping better preserves inter-distances for smaller dissimilarities, while proportionally squeezes the inter-distances for larger dissimilarities.

Ritaglio schermata acquisito: 05/05/2021 18:46

# non-metric mds

when dissimilarities are known only by their rank order, and the spacing between successively ranked dissimilarities is of no interest or is unavailable

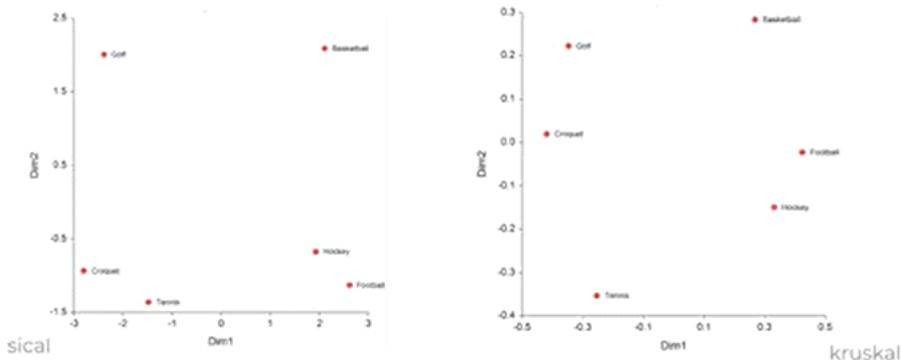
here  $f$  is only implicitly defined as a regression curve, and only preserves the order of  $d$ , that is  $f(d_{ij}) < f(d_{kl})$  if  $d_{ij} < d_{kl}$  thus only the order of  $d$  is needed, not the actual values

most common algorithm is the Kruskal mds

Ritaglio schermata acquisito: 05/05/2021 18:46

dissimilarity rating between sports

Sport	Hockey	Football	Basketball	Tennis	Golf	Croquet
Hockey	0	2	3	4	5	5
Football		0	3	5	6	5
Basketball			0	5	4	6
Tennis				0	4	3
Golf					0	2
Croquet						0

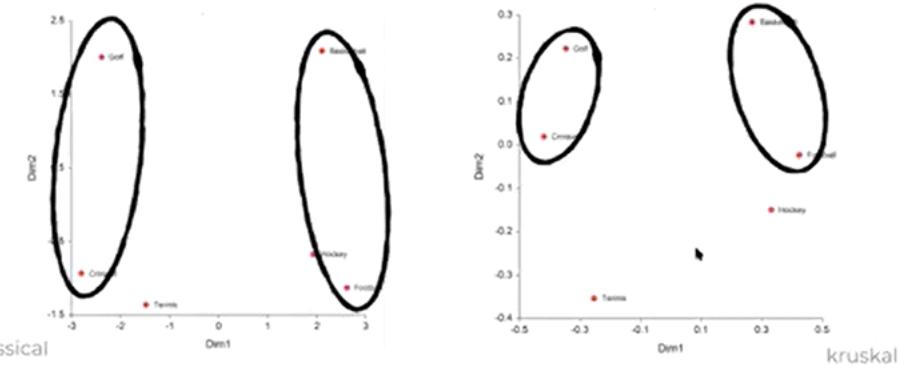


Ritaglio schermata acquisito: 05/05/2021 18:47

# kruskal mds

dissimilarity rating between sports

Sport	Hockey	Football	Basketball	Tennis	Golf	Croquet
Hockey	0	2	3	4	5	5
Football		0	3	5	6	5
Basketball			0	5	4	6
Tennis				0	4	
Golf					0	2
Croquet						0

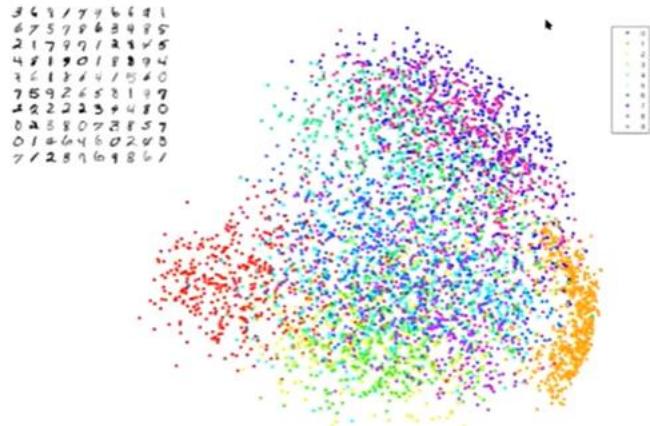


Ritaglio schermata acquisito: 05/05/2021 18:49

# TSNE & UMAP

domenica 16 maggio 2021 16:27

## mnist in pca

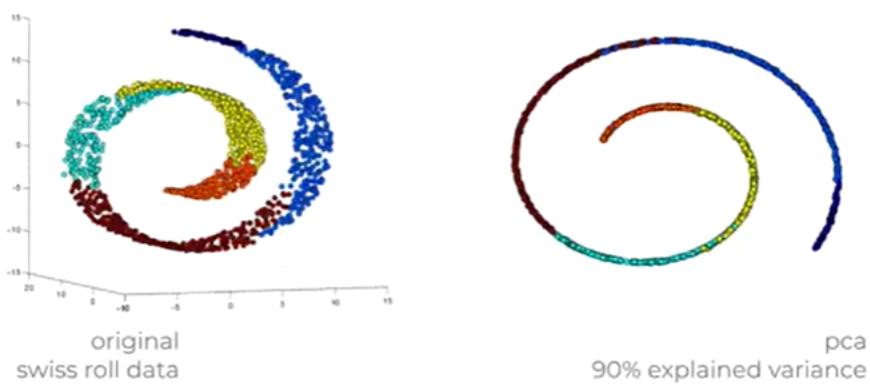


Ritaglio schermata acquisito: 16/05/2021 16:29

Dataset we have many images and each image is an hand written digit.  
Each number is an image, which can be expressed as a matrix of pixels.  
We feed this data in a PCA -> we can represent this data in a two dimensional space  
In the picture we can spot some clusters (orange area on the right, or the red one in the bottom left part of the plot) however, we notice that the PCA is not able to clearly detect the important features that distinguish the digit.

## pca drawbacks

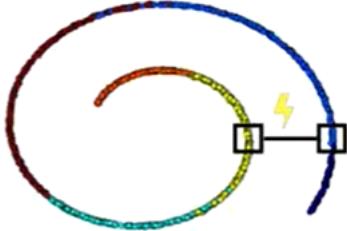
pca targets dimensionality, preserving large pairwise distances in the map, but cannot catch the structure of the data



Ritaglio schermata acquisito: 16/05/2021 16:34

Pca cannot capture the small variation in the data, it is only able to detect big changes in the variance.

Now we consider this new data set, here the pca is good to represent the data.



Ritaglio schermata acquisito: 16/05/2021 16:35

Let's now consider the distance between the blue and the yellow data. They are closer than actually are. The pca shows a relation that does not exist.

## t-sne

what is reliable are the very small euclidean distances between neighbouring points



t-distributed stochastic neighbor embedding

Journal of Machine Learning Research 9 (2008) 2579-2605

Submitted 5/08; Revised 9/08; Published 11/08

### Visualizing Data using t-SNE

**Laurens van der Maaten**  
Tilburg University  
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Lvdmaaten@gmail.com

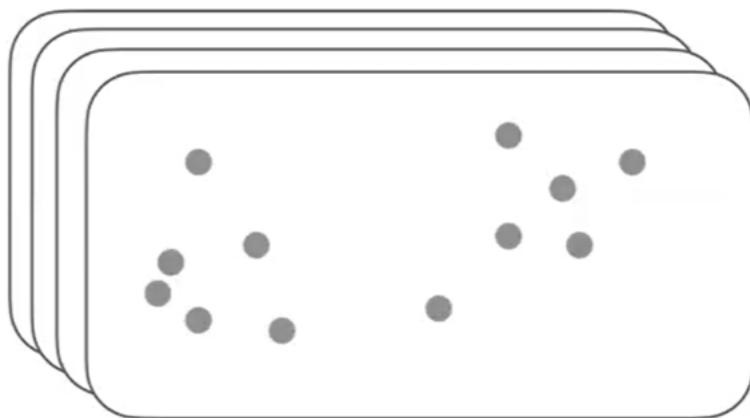
**Geoffrey Hinton**  
Department of Computer Science  
University of Toronto  
6 King's College Road, M5S 3G4 Toronto, ON, Canada

hinton@cs.toronto.edu

Ritaglio schermata acquisito: 16/05/2021 16:36

What is reliable in the dimensionality reduction is indeed the concept of neighborhood.

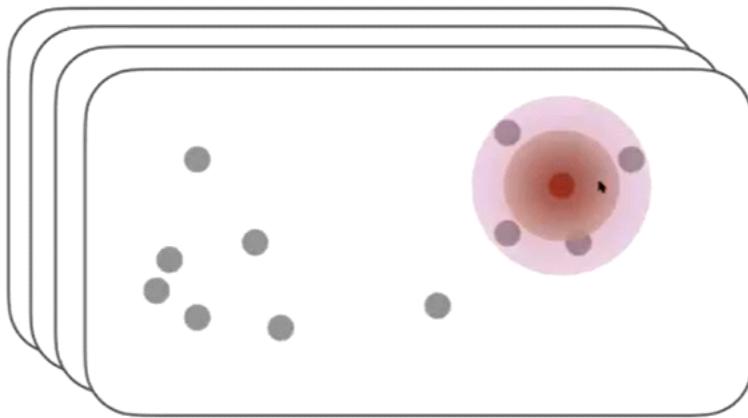
## focus on neighbours



Ritaglio schermata acquisito: 16/05/2021 16:37

Suppose to have the above data set. Let's consider the red point. Considering the big circle around

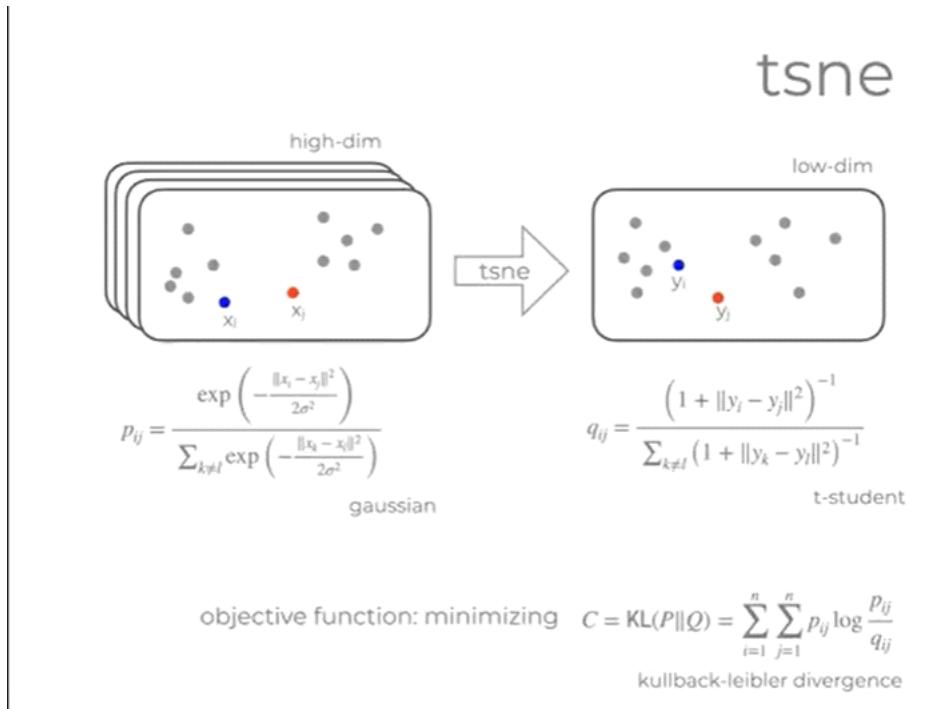
the circle the more it is closer to it the more it is similar to the red point.



use a gaussian function to weight distances

Ritaglio schermata acquisito: 16/05/2021 16:38

We do so to construct the concept of metric neighborhood.



Ritaglio schermata acquisito: 16/05/2021 16:39

The aim of tsne is to project data into a low dimensional space. The idea is to define properly the distance among the points.

The core of tsne algorithm is to find the minimum of the kullback-leiber function.

# novelty

w.r.t. previous sne algorithms,

tsne uses joint instead of conditional probabilities:  
this introduces symmetry in the problem formulation, and the cost  
function optimisation is computationally much simpler

- uses t-student distribution instead of gaussian modelling on the  
low-dimensional space:  
this heavy-tailed distribution in the low-dimensional space alleviates  
both the crowding problem and the optimization problem

*crowding problem*

the volume of a sphere centered on datapoint i scales as  $r^m$ , where  
r is the radius and m the dimensionality of the sphere; if we want  
to model the small distances accurately in the map, most of the  
points that are at a moderate distance from datapoint i will have  
to be placed much too far away in the two-dimensional map.

the heavy-tailed distribution instead corrects volume  
differences between both spaces

Ritaglio schermata acquisito: 16/05/2021 16:41

Join probability instead of conditional one -> this introduce symmetry.  
t-student distribution on the arriving low dimensional space is helping the solution of the problem.

## t-student on low-dim space

the heavy-tailed distribution instead corrects volume  
differences between both spaces

in fact, it allows a moderate distance in the high-dimensional  
space to be faithfully modeled by a much larger distance in the  
map and, as a result, it eliminates the unwanted attractive  
forces between map points that represent moderately  
dissimilar data points

this is due to the fact that the map's representation of joint  
probabilities is (almost) invariant to changes in the scale of the  
map for map points that are far apart

Ritaglio schermata acquisito: 16/05/2021 16:44

# the optimisation step

since the kullback-leibler divergence is not symmetric, different types of error in the pairwise distances in the low-dimensional map are not weighted equally

there is a large cost for using widely separated map points to represent nearby data points, but there is only a small cost for using nearby map points to represent widely separated data points

solution is found by using gradient descent, a first-order iterative optimization algorithm for finding the minimum of a function

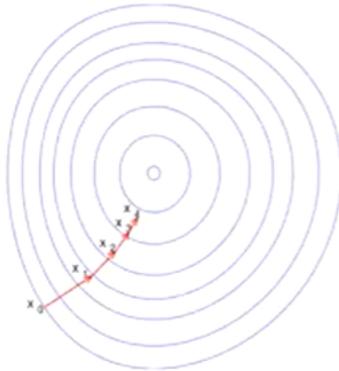
in details, one takes consecutive steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.

Ritaglio schermata acquisito: 16/05/2021 16:45

## gradient descent

$f: \mathbb{R}^p \rightarrow \mathbb{R}$ , defined and differentiable in a neighborhood of a point  $a$ , decreases fastest if one goes from  $a$  in the direction of the negative gradient of  $f$  at  $a$ ,  $-\nabla f(a) = -(\partial f / \partial x_1, \dots, \partial f / \partial x_p)|_a$

if  $a_{n+1} = a_n - \gamma \nabla f(a_n)$  for a small positive  $\gamma$ , then  $f(a_{n+1}) \leq f(a_n)$



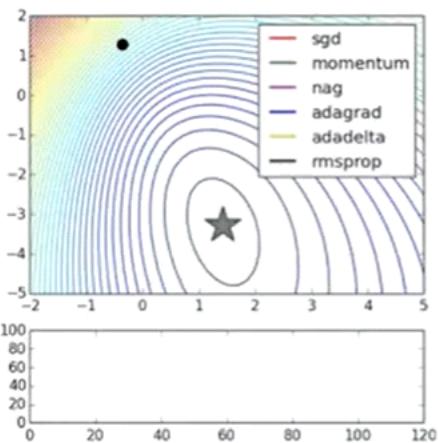
start with a guess  $x_0$  of a local minimum for  $f$  and consider the sequence  $x_0, x_1, x_2, \dots$  with  $x_{n+1} = x_n - \gamma_n \nabla f(x_n)$ : then we have a monotonic sequence  $f(x_0) \geq f(x_1) \geq f(x_2) \geq \dots$  so that the sequence  $x_0, x_1, x_2, \dots$  can converge to the minimum of  $f$

Ritaglio schermata acquisito: 16/05/2021 16:48

# gradient descent

for  $f$  nonconvex

$$\forall x_1, x_2 \in X, \forall t \in [0, 1] : f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$



and  $\nabla f$  lipschitz

$$|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$$

convergence is guaranteed with

$$\gamma_n = \frac{|(\mathbf{x}_n - \mathbf{x}_{n-1})^T [\nabla F(\mathbf{x}_n) - \nabla F(\mathbf{x}_{n-1})]|}{\|\nabla F(\mathbf{x}_n) - \nabla F(\mathbf{x}_{n-1})\|^2}$$

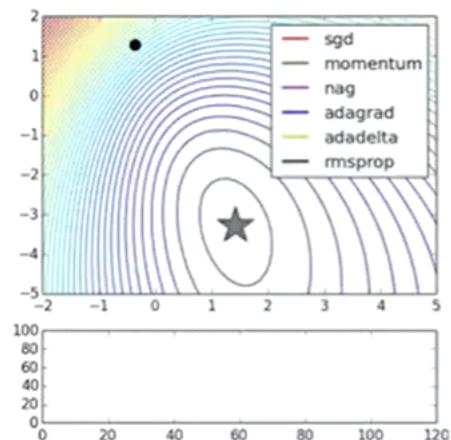
Ritaglio schermata acquisito: 16/05/2021 16:50

Convex -> if we construct the segment from  $x_1$  to  $x_2$  the image of the segment is smaller than the segment between the two images..



for  $f$  convex

$$\forall x_1, x_2 \in X, \forall t \in [0, 1] : f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$



and  $\nabla f$  lipschitz

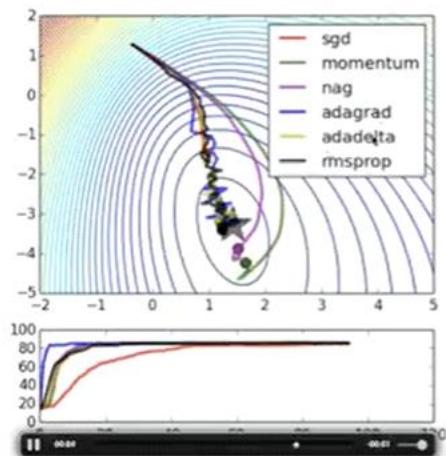
$$|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$$

convergence is guaranteed with

$$\gamma_n = \frac{|(\mathbf{x}_n - \mathbf{x}_{n-1})^T [\nabla F(\mathbf{x}_n) - \nabla F(\mathbf{x}_{n-1})]|}{\|\nabla F(\mathbf{x}_n) - \nabla F(\mathbf{x}_{n-1})\|^2}$$



Ritaglio schermata acquisito: 16/05/2021 16:54



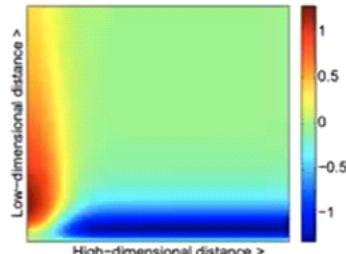
Ritaglio schermata acquisito: 16/05/2021 16:55

## tsne optimization

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \left(1 + \|y_i - y_j\|^2\right)^{-1}$$

n bodies elastic system

- strongly repels dissimilar data points that are modeled by a small pairwise distance in the low-dimensional representation
- these repulsions do not go to infinity



early compression:  
force the map points to stay close together at the start of the optimization, implemented by adding an additional L2-penalty to the cost function that is proportional to the sum of square distances of the map points from the origin

early exaggeration:  
multiply all of the  $p_{ij}$ 's by, for example, 4, in the initial stages of the optimization, modeling the large  $p_{ij}$ 's by fairly large  $q_{ij}$ 's; natural clusters in the data tend to form tight widely separated clusters in the map

Ritaglio schermata acquisito: 16/05/2021 16:56

How tsne does the optimization? We have the above formula ( n bodies elastic system). We can use two methods to overcome the fact that these repulsions do not go to infinity.

# perplexity

the only parameter to set is the variance  $\sigma_i$  for the high-dim gaussian of  $p_{ij}$

no single value of  $\sigma_i$  can be optimal for all data points in the data set because the density of the data is likely to vary

in dense regions, a smaller value of  $\sigma_i$  is usually more appropriate than in sparser regions

any particular value of  $\sigma_i$  induces a probability distribution  $P_i$  that has an entropy which increases as  $\sigma_i$  increases

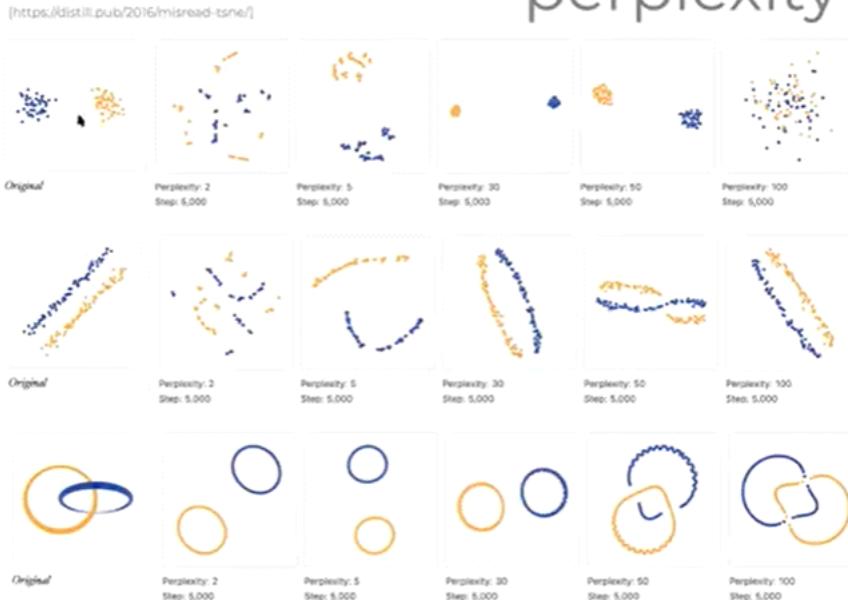
tSNE performs a binary search for the value of  $\sigma_i$  that produces a  $P_i$  with a fixed perplexity that is specified by the user

$$\text{Perp}(P_i) = 2^{-H(P_i)} = 2^{-\sum_j P_{ij} \log_2 P_{ij}}$$

perplexity can be interpreted as a smooth measure of the effective number of neighbors, typical values are between 5 and 50

Ritaglio schermata acquisito: 16/05/2021 16:58

# perplexity



Ritaglio schermata acquisito: 16/05/2021 17:11

The pictures on the left are the original one (we have two clusters well separated), the other on the right are tSNE projections of the original dataset with different values of perplexity.

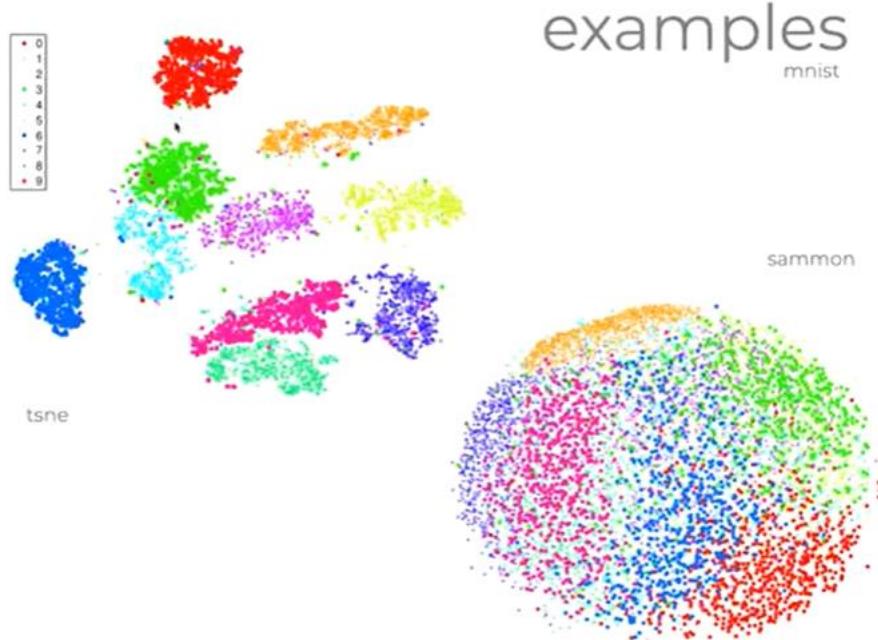
Too low perplexity -> not satisfactory.

Increasing perplexity -> means using more data points. If we increase before the clusters get more dense and then become more relaxed

Too large perplexity -> too many datapoints -> clusters get lost

Two rings -> with perplexity from 2 to 30 we have two separated circles. Then the situation gets more complicated because the two rings overlap.

Hence, there is not a unique and best value for the perplexity.



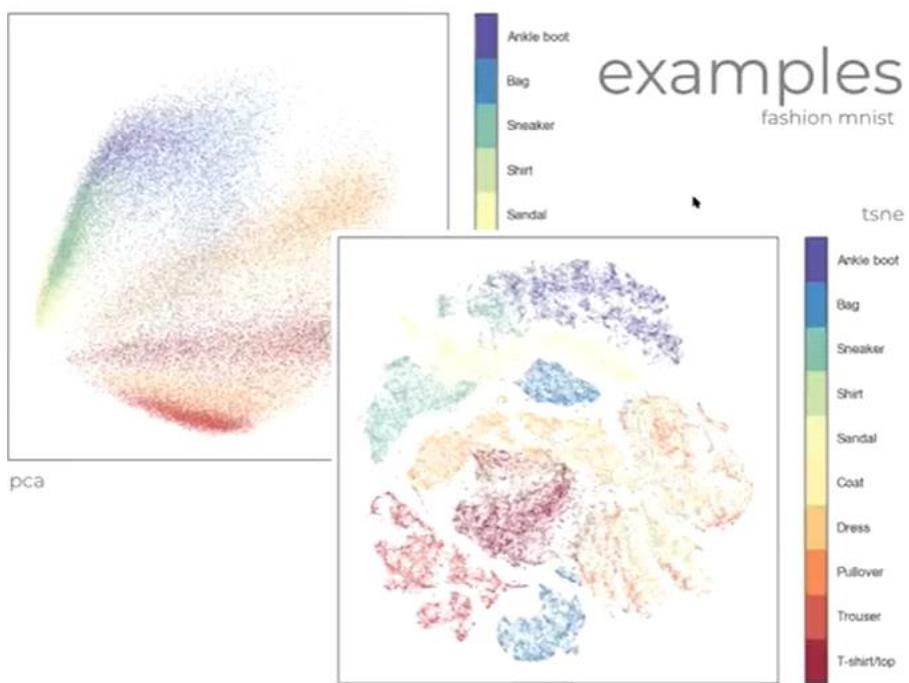
Ritaglio schermata acquisito: 16/05/2021 17:17

Tsne perform quite well



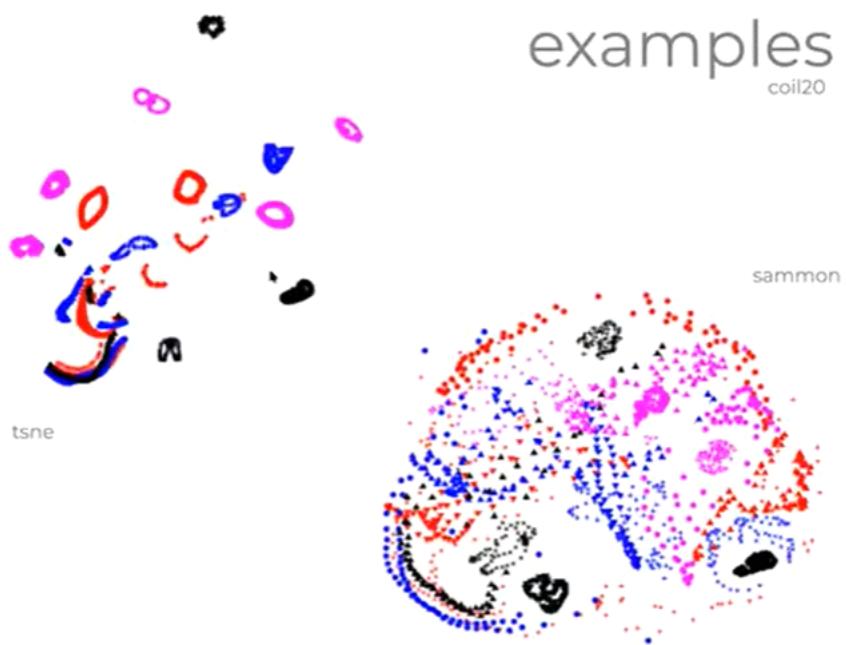
Ritaglio schermata acquisito: 16/05/2021 17:17

Images here are clothing.



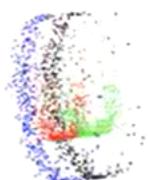
Ritaglio schermata acquisito: 16/05/2021 17:18

Tsne on the right.

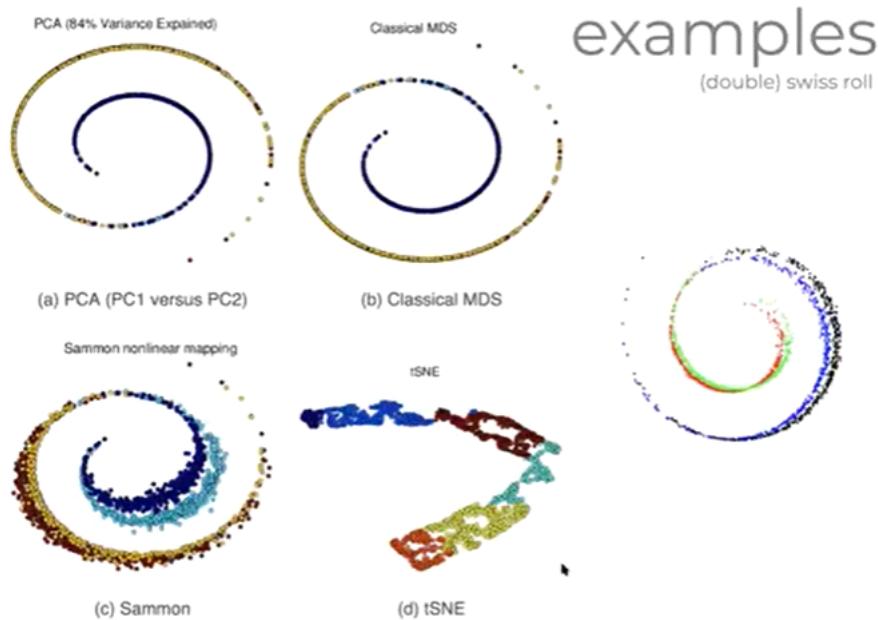


Ritaglio schermata acquisito: 16/05/2021 17:18

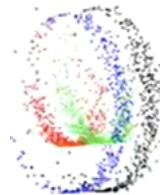
Tsne separated well the classes



Ritaglio schermata acquisito: 16/05/2021 17:19



Ritaglio schermata acquisito: 16/05/2021 17:19



Ritaglio schermata acquisito: 16/05/2021 17:19

Tsne is good also in the case of the double swiss roll which is a really complex data set.

## drawbacks

does not output transformation

stochastic

output space hard to interpret

hyperparameter choice critical

views at different perplexities needed to understand topology

artifacts in data may appear

cluster size and inter-cluster distance not meaningful

slow on large datasets

Ritaglio schermata acquisito: 16/05/2021 17:20

Sthochastic -> statistically based.

Hard to interpret in term of geometry. Choice of the perplexity is critical.

Although it does a good job in clustering, the size of the cluster and the distance between clusters is not meaningful.



umap

UMAP: Uniform Manifold Approximation and Projection

Leland McInnes<sup>1</sup>, John Healy<sup>1</sup>, Nathaniel Saul<sup>2</sup>, and Lukas Großberger<sup>3, 4</sup>

1 Tufts Institute for Mathematics and Computing 2 Department of Mathematics and Statistics, Washington State University 3 Ernst Strüngmann Institute for Neuroscience in cooperation with Max Planck Society 4 Donders Institute for Brain, Cognition and Behaviour, Radboud Universiteit

DOI: 10.21105/joss.0081

Software

- [Source](#)
- [Repository](#)
- [Archive](#)

Submitted: 19 July 2018  
Published: 02 September 2018

**Summary**

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualizations similarly to t-SNE, but also for general non-linear dimension reduction. UMAP has a rigorous mathematical foundation, but is simple to use with a width range controllable API. UMAP is released under the Apache 2.0 license.



nature > nature biotechnology > analyses > article

**nature biotechnology**

Analysis Published: 02 December 2018

Dimensionality reduction for visualizing single-cell data using UMAP

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutour, Immanuel WH Keyk, Lal Guan Ng, Florent Ginhoux & Evan W Newell

Nature Biotechnology 37, 38–44 (2019) Download Citation

Ritaglio schermata acquisito: 16/05/2021 17:23

As a solution for these problems, has been proposed a new method.

The last dimensionality reduction method is umap, uniform manifold approximation and projection. The underlying idea is to detect the manyfold the geometrical structure where the points are leaving and try to approximate it and project it in a two (three) dimensional space.

## topological facts

*topological space*

an ordered pair  $(X, \tau)$  where  $X$  is a set and  $\tau$  is a collection of subsets of  $X$ , called topology, for which the following axioms hold:

- $\{\emptyset, X\} \subseteq \tau$
- any arbitrary union of elements in  $\tau$  belongs to  $\tau$
- any finite intersection of elements in  $\tau$  belongs to  $\tau$
- the elements of  $\tau$  are called open sets

*example:  $\mathbb{R}^n$  with  $\tau$  any union of open balls*

Ritaglio schermata acquisito: 16/05/2021 17:32

open balls -> is the circle surrounding a given point but without the borders.

*continuous function*

$f: (X, \tau_X) \rightarrow (Y, \tau_Y)$  s.t.  $f^{-1}(V) \in \tau_X$  for each  $V \in \tau_Y$

Ritaglio schermata acquisito: 16/05/2021 17:34

A continuous function is a function that goes from a topological space  $x$  to a topological space  $y$  such that the counter image of an element in  $\tau_y$  lies in  $\tau_x$ .  
The pre image in an open set in a pre image in open set in concise terms.

**homeomorphism**  
a continuous function between topological spaces that has a continuous inverse function

Ritaglio schermata acquisito: 16/05/2021 17:37

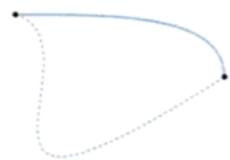
## topological facts

**manifold**  
a topological space resembling euclidean space near each point

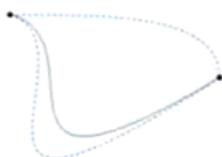
**cover**  
a collection of subsets of  $X$  whose union is  $X$

**homotopy**  
given 2 continuous functions  $f,g:X \rightarrow Y$ , an homotopy  $h:X \times [0,1] \rightarrow Y$  continuous,  $h(x,0)=f(x)$  and  $h(x,1)=g(x)$

informally, it is a continuous deformation between  $f$  and  $g$



Ritaglio schermata acquisito: 16/05/2021 17:37



Ritaglio schermata acquisito: 16/05/2021 17:39

## topological facts

**homotopy equivalence**  
two topological spaces  $X,Y$  are homotopy equivalent if  $\exists f:X \rightarrow Y$  and  $g:Y \rightarrow X$  continuous s.t.  $f \cdot g$  is homotopic to  $1_X$  and  $g \cdot f$  is homotopic to  $1_Y$

every homeomorphism is an homotopy equivalence, but not the converse; (informally) two spaces  $X$  and  $Y$  are homotopy equivalent if they can be transformed into one another by bending, shrinking and expanding operations

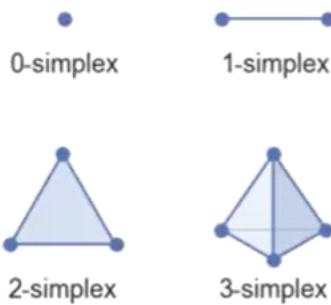
Ritaglio schermata acquisito: 16/05/2021 17:40

*contractible space*  
homotopy equivalent to a point

examples:  
 $\mathbb{R}^n$ , with homotopy  $H(x,t) = t \cdot 1(x) + (1-t) \cdot 0(x)$   
the solid disk

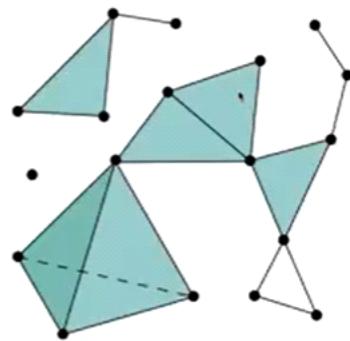
Ritaglio schermata acquisito: 16/05/2021 17:47

## simplices



*(k-)simplex*  
k-dim polytope that is the convex hull  
of its  $k+1$  vertices;  
a regular  $n$ -simplex can be constructed  
from a regular  $(n-1)$ -simplex by  
connecting a new vertex to all original  
vertices by the common edge length

*simplicial complex*  
set  $K$  of simplices glued together  
along faces: any face of any  
simplex in  $K$  is also in  $K$  and the  
intersection of two simplices in  $K$  is  
also in  $K$



Ritaglio schermata acquisito: 16/05/2021 17:48

SIMPLICES -> Smallest convex geometrical figures that has  $k+1$  vertices.

In general an  $n$ -simplex can be constructed from the previous simplex adding a new point and connecting it to all the other points.

the intersection of two simplices is still a simplex.

a broader extension of the simplicial complex exists,  
giving rise to a much larger class of topological spaces  
— they are called simplicial sets, defined in the  
mathematical framework of the category theory in the  
1950s and designed to capture the notion of a "well-  
behaved" topological space for the purposes  
of homotopy theory;  
the category of simplicial sets carries a natural model  
structure, and the corresponding homotopy category is  
equivalent to the familiar homotopy category of  
topological spaces

**however, in what follows we will work with  
simplicial complexes**

Ritaglio schermata acquisito: 16/05/2021 18:01

# čech nerve

given an open cover  $\mathcal{U} = \{U_i\}_{i \in I}$  of a topological space  $X$ , consider the fibre product  $U_{ij} = U_i \times_X U_j = U_i \cap U_j$ ; generalizing to arbitrary intersection we obtain the simplicial object  $N(\mathcal{U}) = \coprod_{x \in X} U_x$ , called the čech nerve (complex) of  $X$

*in practice*

let each set in the cover be a 0-simplex  
create a 1-simplex between two such sets if they have a non-empty intersection  
create a 2-simplex between three such sets if the triple intersection of all three is non-empty  
and so on

*does this simple process produce something that represents the topological space  $X$  itself in a meaningful way?*

Ritaglio schermata acquisito: 16/05/2021 18:02

Cech nerve -> all the possible intersection of element in  $U$

## čech nerve

*compact space*  
a topological space  $X$  is compact if each of its open cover has a finite subcover

*nerve theorem*  
if  $X$  is compact and  $\mathcal{U} = \{U_i\}_{i \in I}$  is an open cover such that for each  $\sigma \subseteq I$ , the intersection  $\cap_{i \in \sigma} U_i$  is either contractible or empty, then  $N(\mathcal{U})$  is homotopy equivalent to  $X$

thus from the nerve of  $X$  we can actually recover all the key topological structures of the original space

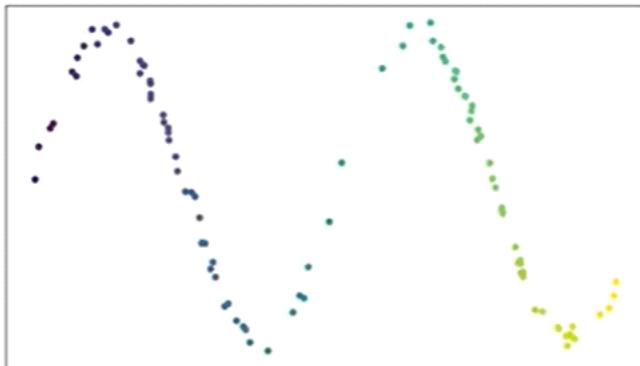
Ritaglio schermata acquisito: 16/05/2021 18:06

---

# from spaces to data

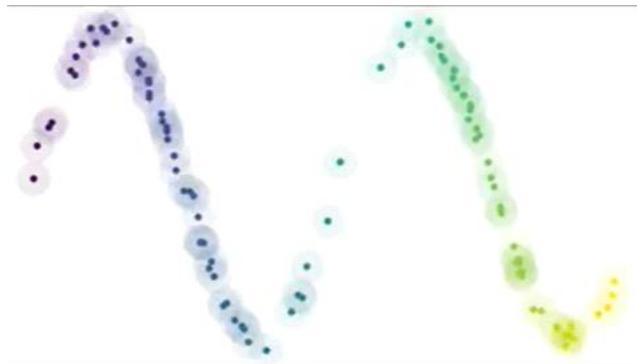
to apply the above construction to a dataset, a open cover is needed

if data lie in a metric space (i.e. there exists a distance), the open cover can be obtained by the balls centered in each data point; thus there is a 0-simplex for each data point



Ritaglio schermata acquisito: 16/05/2021 18:07

We have a data set. The first thing we need is a open cover. In particular if the dataset lies in a metric space -> the open cover can be obtained from all the open balls of the data points.

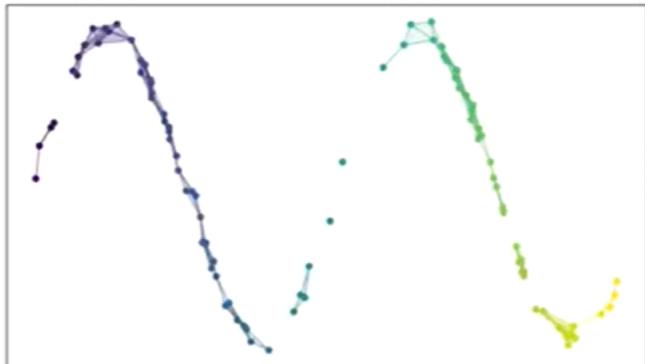


Ritaglio schermata acquisito: 16/05/2021 18:09

Cover -> the union include the whole starting data set.  
Then I can start construct my nerve.

note that most of the job can be done just considering 0- and 1-simplices: mathematically, this is equivalent to consider the victories-rips complex instead of the čech complex

this yields that a dimensional reduction can be obtained by mean of a graph representation



Ritaglio schermata acquisito: 16/05/2021 18:09

Points and lines are nodes and links in a network.

## issues

### *problem 1*

what's the optimal radius for the open cover?

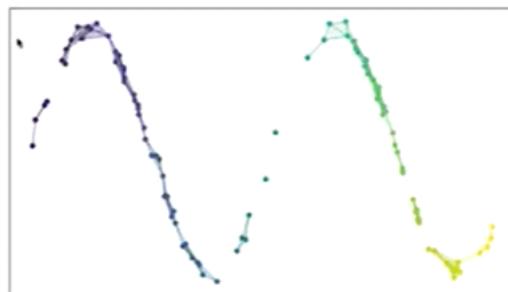
too small: too many connected components

too large: few very high dimensional simplices, and no structure

### *problem 2*

if points are too scattered, the structure cannot be properly captured

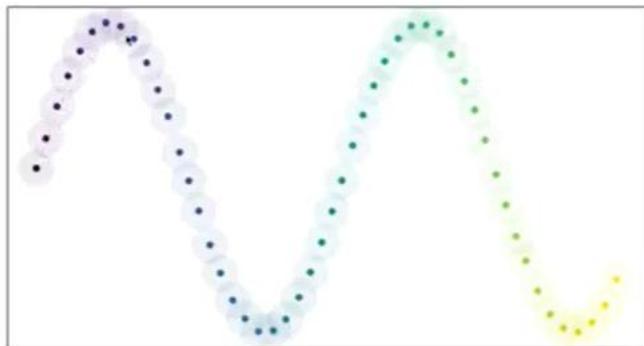
if points are too dense, the cover generates a too high-dim nerve



Ritaglio schermata acquisito: 16/05/2021 18:11

# best of all possible worlds

having data points uniformly distributed across the manifold  
radius = half the average distance between points  
no gaps & no clumps in the cover



Ritaglio schermata acquisito: 16/05/2021 18:12

But this does not happen with real world data.

SOLUTION:

adapt the notion of distance on the manifold (stretch) so that all the points seem to be uniformly distributed

Ritaglio schermata acquisito: 16/05/2021 18:13

I need to stretch the space -> this is why we introduce the topology.

We simplify the situation with a line. There is a part of the line where there are many points -> the space is not uniform.

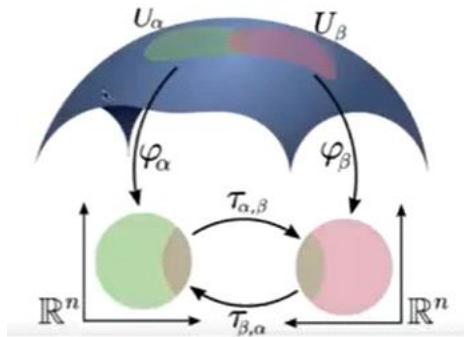


Ritaglio schermata acquisito: 16/05/2021 18:14

Bending the space -> guarantees that all the points are uniformly distributed as we wanted.

thus give to each point its own unique distance function, and select balls of radius one with respect to that local distance function

Ritaglio schermata acquisito: 16/05/2021 18:14



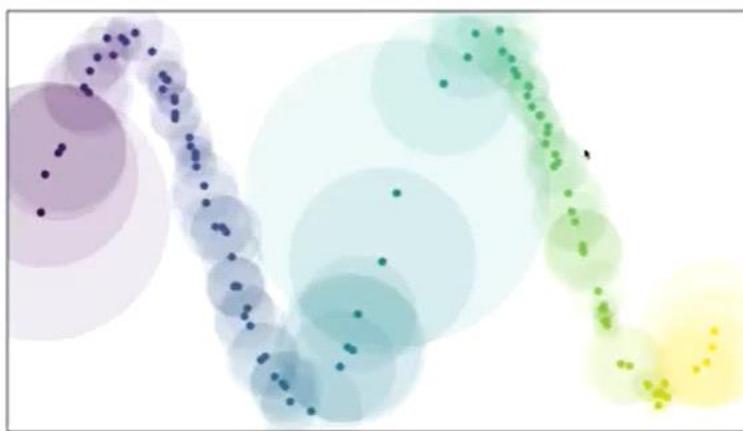
Ritaglio schermata acquisito: 16/05/2021 18:14

We have the manifold, we pick up a point and its neighbor, project the neighbour to a particular subset of  $\mathbb{R}^n$  where the distance is defined.

There is a map that allows to go from one space to another.

## solution

The unit ball about a point stretches to the k-th nearest neighbor of the point, where k is the sample size we are using to approximate the local sense of distance (easier in terms of parameters)

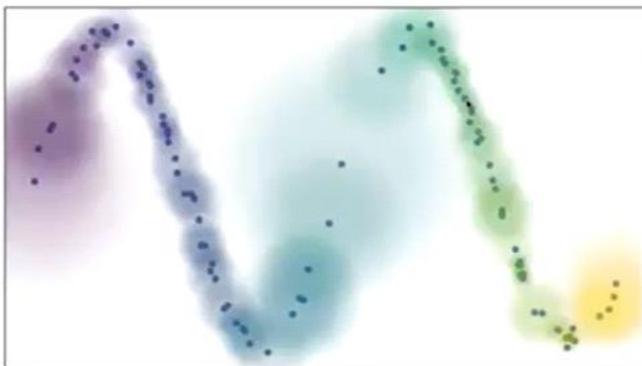


Ritaglio schermata acquisito: 16/05/2021 18:16

# fuzziness

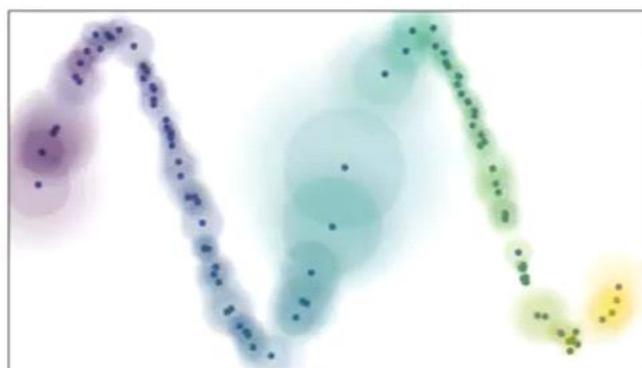
given the local metric, we can weight the edges of the graph according to the edge' vertices distance

mathematically, we move from the simplicial complexes to the simplicial sets (category theory), which corresponds to move from classical balls to fuzzy balls, where  $\epsilon$  becomes a function in  $[0,1]$  decreasing further away from the center  $(1+ax^{2b})^{-1}$



Ritaglio schermata acquisito: 16/05/2021 18:16

- points may end up being separated by the rest of the manifold
- local connectivity is introduced: the fuzziness decays only beyond the nearest neighbour
- the focus is on the difference in distances among nearest neighbors rather than the absolute distance, avoid the curse of dimensionality



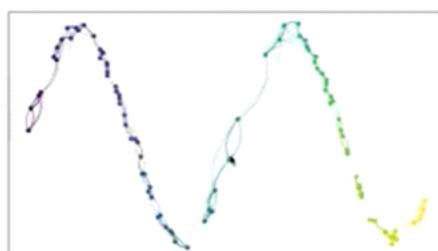
Ritaglio schermata acquisito: 16/05/2021 18:18

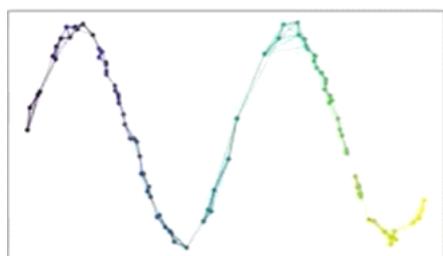
Fuzziness start to take place only behind the nearest neighbors.

The focus is more on the focus on the distance among the neighbor rather than the distance among all the points.

# incompatibility

since each point has its own metric,  
distance from a to b may be  
different than distance from b to a





the theoretically grounded solution is correctly defining the fuzzy union of simplicial sets as the probability that at least one of the edge exists, thus ending with a single fuzzy simplicial complex (weighted graph)

## low-dim representation

we need now to faithfully embed the graph into a low-dim euclidean space so to preserve the original manifold structure

this means deciding which  $f(w_h, w_l)$  to optimize, where  $w_h, w_l$  are the graph edges' weights in high and low dimension

since  $w_h, w_l$  are bernoulli variables ( $w$  exists with prob.  $p$  and does not exist with prob.  $1-p$ ), the correct function is the cross-entropy

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$



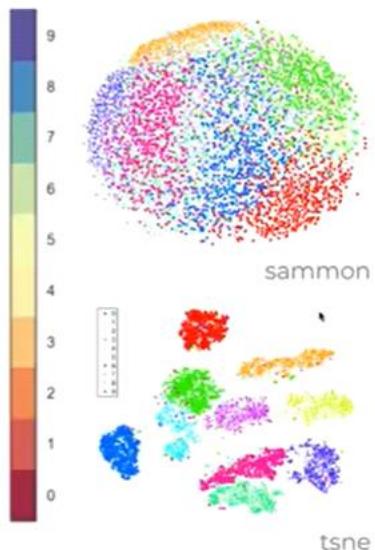
attractive force between points when  $w$  is large in high-dimension optimizing clumps



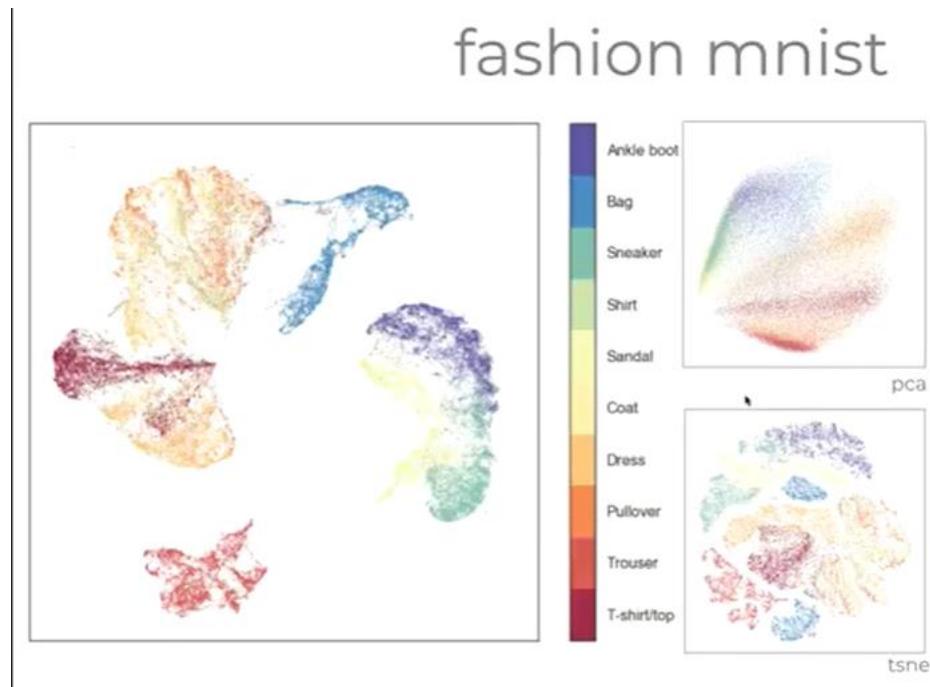
repulsive force between points when  $w$  is small in high-dimension: optimizing gaps

Bernoulli variables -> cross-entropy.

## mnist



tsne



## hyperparameters

*n\_neighbors*

determines the number of neighboring points used in local approximations of manifold structure  
larger values will result in more global structure being preserved at the loss of detailed local structure.  
should often be in the range 5 to 50, with a choice of 10 to 15 being a sensible default.

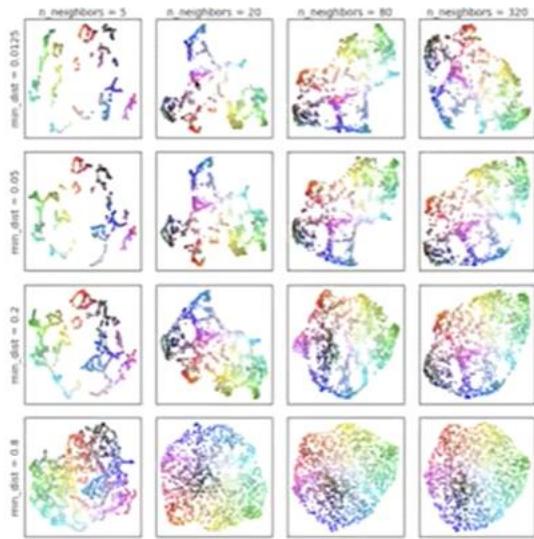
*min\_dist*

controls how tightly the embedding is allowed compress points together  
larger values ensure embedded points are more evenly distributed, while smaller values allow the algorithm to optimise more accurately with regard to local structure  
sensible values [0.001,0.5], with 0.1 being a reasonable default

*metric*

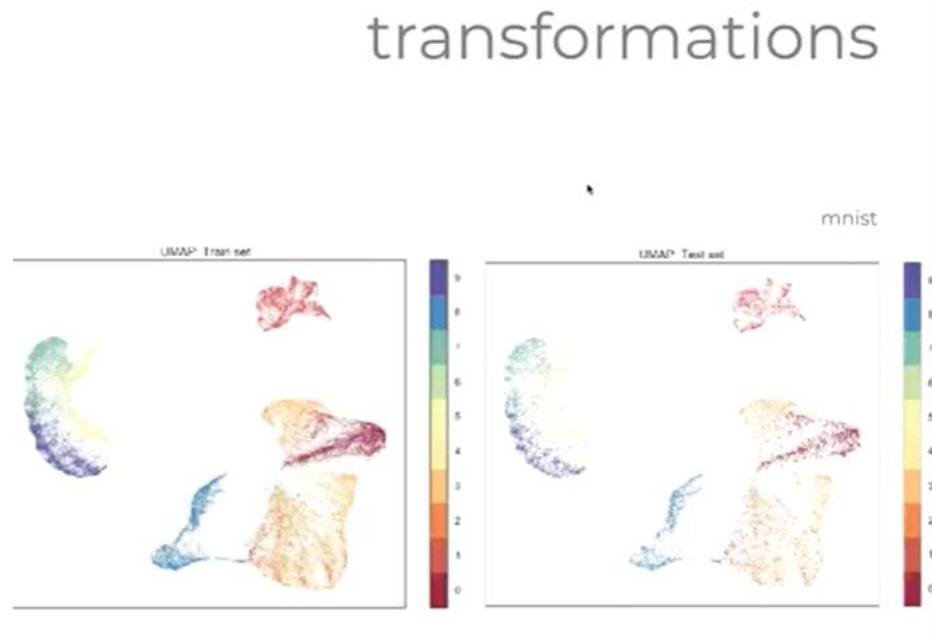
determines the choice of metric used to measure distance in the input space  
wide variety of metrics are already coded  
a user defined function can be passed

# hyperparameters



Ritaglio schermata acquisito: 16/05/2021 18:33

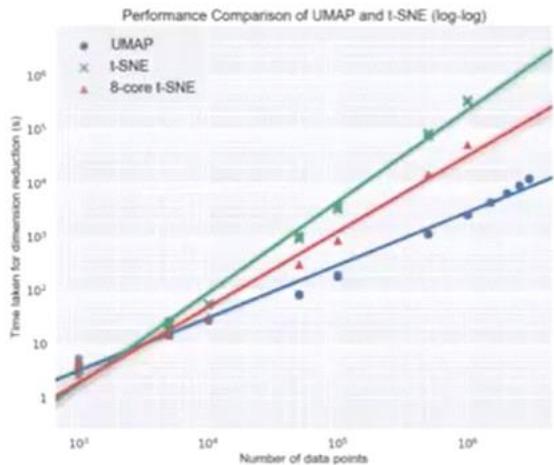
# transformations



Ritaglio schermata acquisito: 16/05/2021 18:34

Umap -> transform the whole space.

# cputime



dataset	dataset size	t-SNE	UMAP
COIL20	1440x16384	20s	<b>7s</b>
COIL100	72000x49152	683s	<b>121s</b>
Shuttle	58000x9	741s	<b>140s</b>
MNIST	70000x784	1337s	<b>98s</b>
F-MNIST	70000x784	906s	<b>78s</b>
GoogleNews	200000x300	16214s	<b>821s</b>

Ritaglio schermata acquisito: 16/05/2021 18:35



Ritaglio schermata acquisito: 16/05/2021 18:36