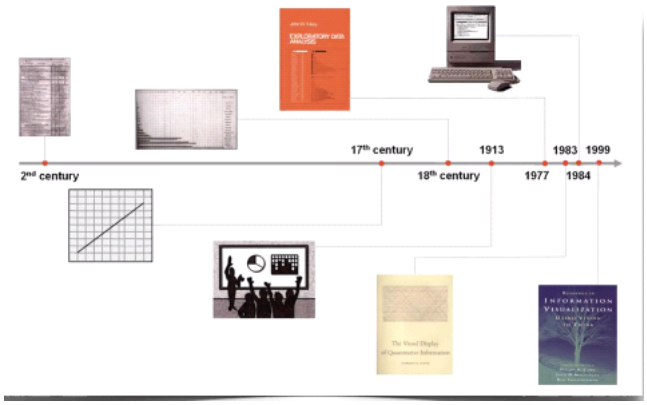
Lesson 3

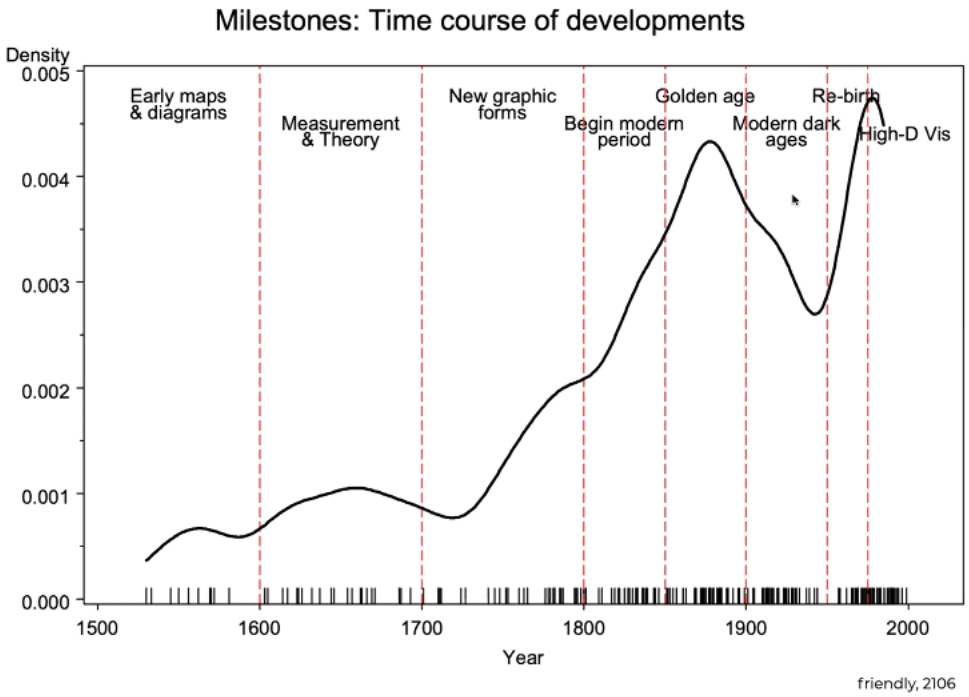
martedì 2 marzo 2021 11:30

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



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

Data visiualization advance is characterised by some important steps.

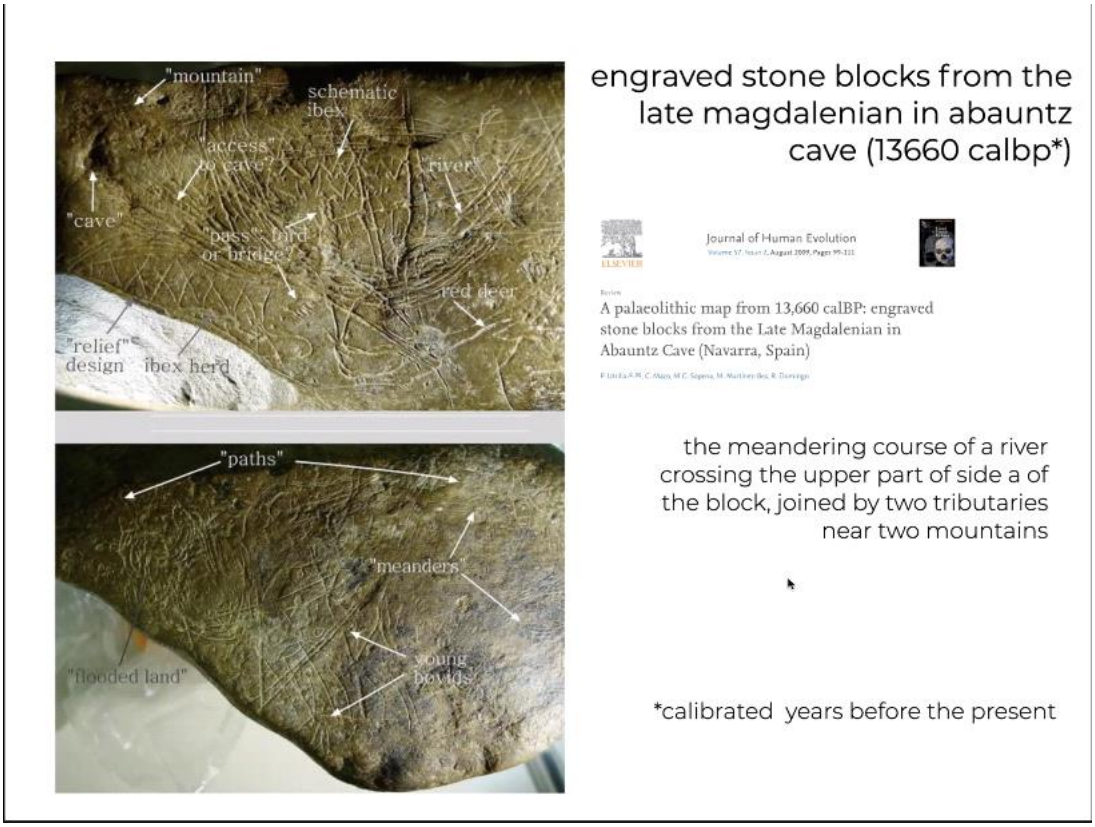


The above image shows different ages and the most important event that characterized the data visualization of that period.

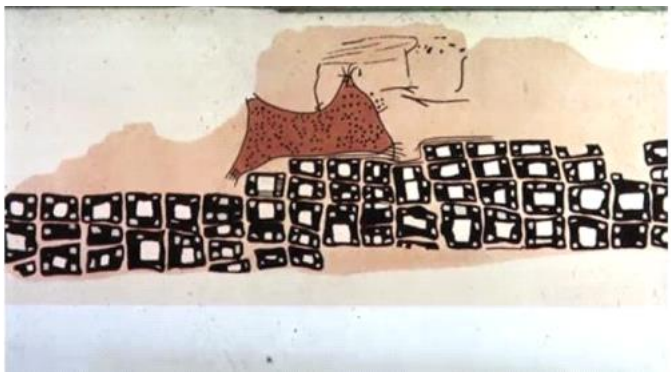
PRE-1600
We treat all this period as a single unit.

- The characterstics of this period are :
- Aboundant 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



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 sourranding territory.
It is a map because the need fro 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 decoration. Although it is a practical tool the graphical aspect is still important.



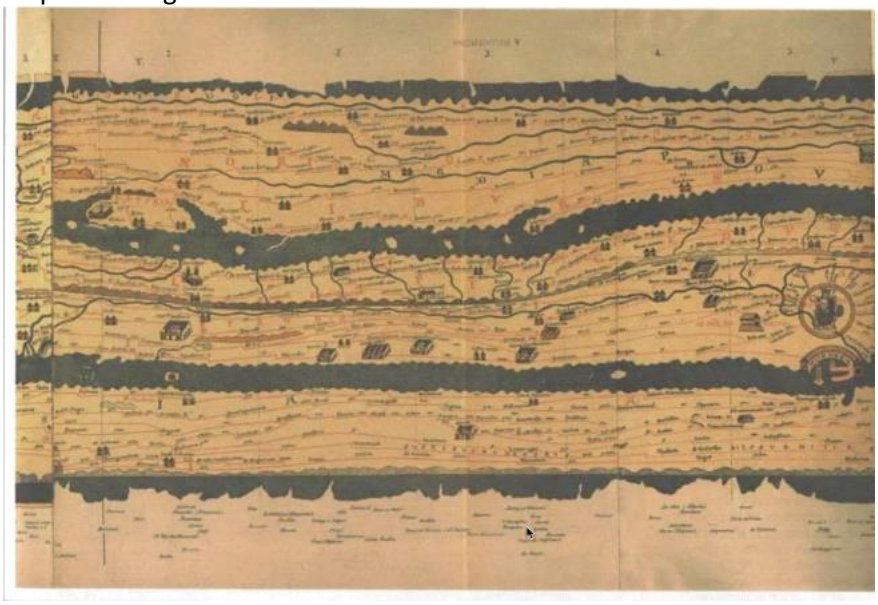
Ritaglio schermata acquisito: 02/03/2021 11:40

This is another map. This shows a town. In the landscape we can also see a vulcano. Archelogist thanks to the rapresentation of the vulcano where 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 that territories.

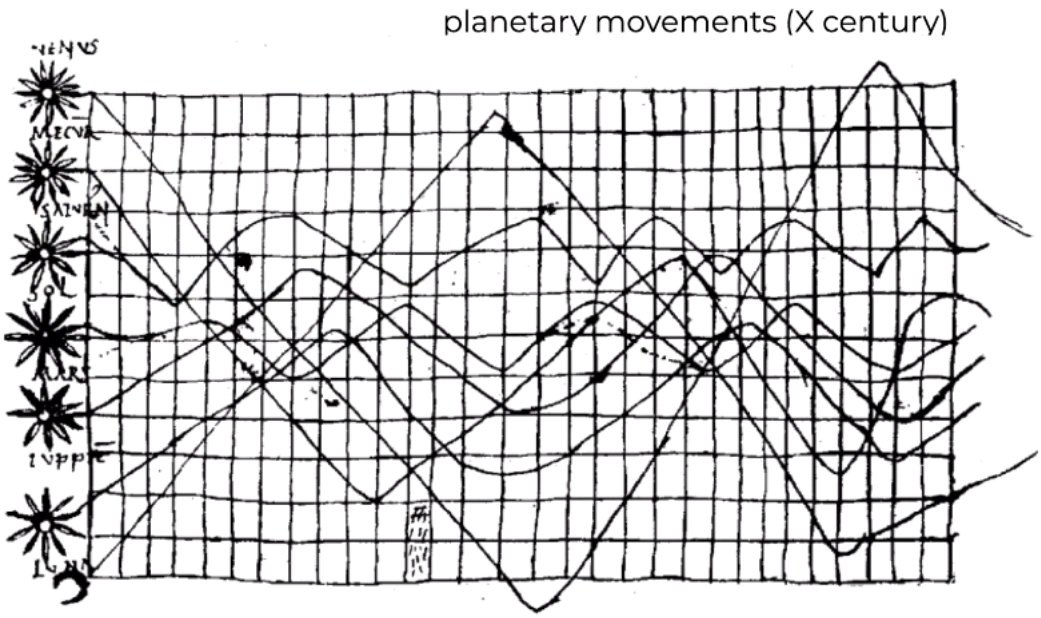


Ritaglio schermata acquisito: 02/03/2021 11:44

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 rapresetnation. 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

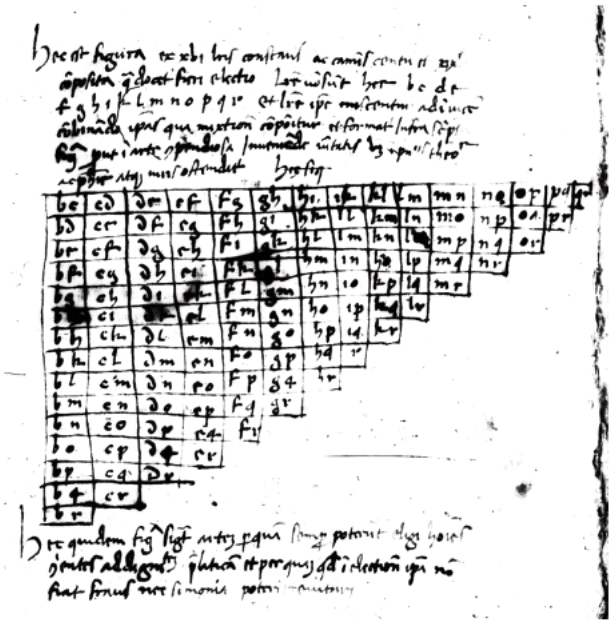
This is a world map. It is a big achievement for the time. The level of details is incredible not only for europe but also for eastern Asia. In this map we can see the first use of longitudinal and latitudinal lines.



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

This is the first example of data rapresentation. 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 bodfies over space and time.
Horizontal scale is indipendent for each planet. We can see also the use of annotations.

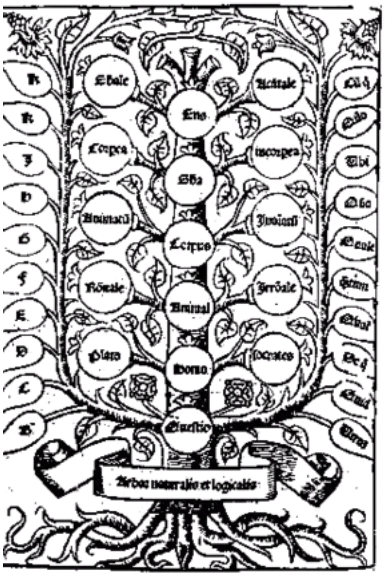
artifitium electionis personarum (r. llull, 1280)



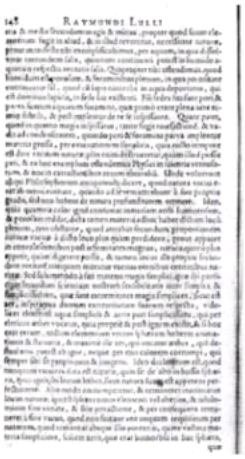
Ritaglio schermata acquisito: 02/03/2021 11:53

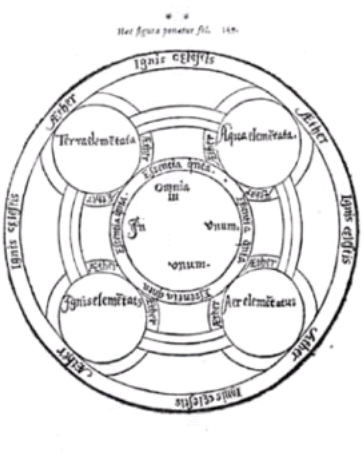
For the first time in a book there is a semi-graphic table. It is a reference table that summarized the results of an election.

mechanical diagrams of knowledge (r. llull, 1305)



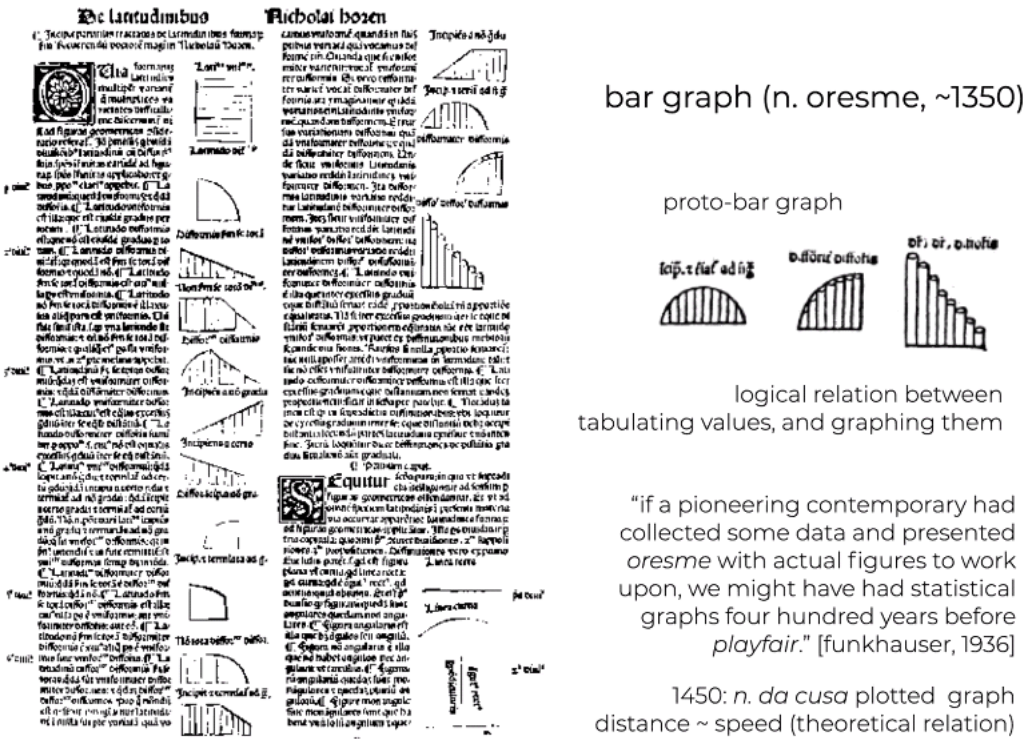
tree of knowledge





mechanical disks

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

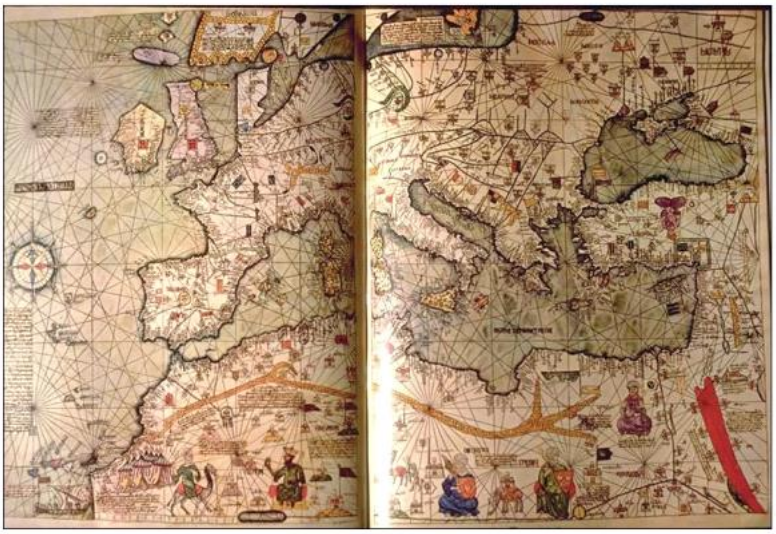


Ritaglio schermata acquisito: 02/03/2021 11:54

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.

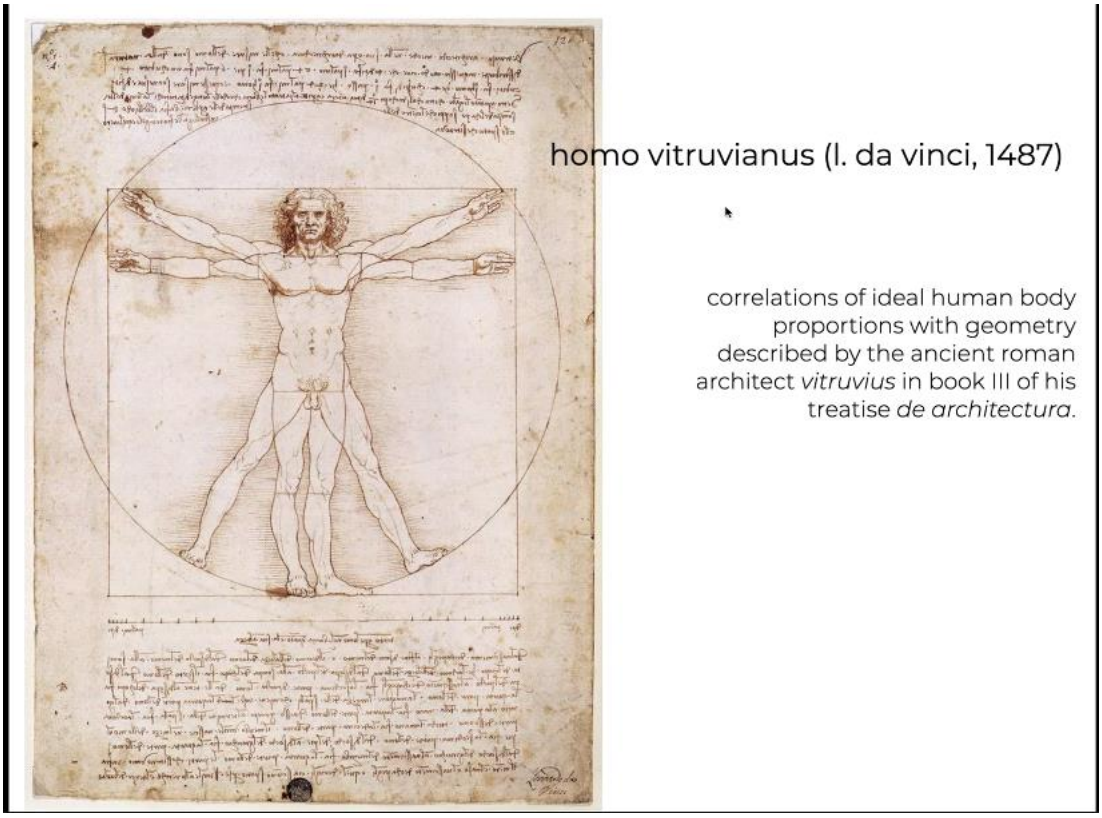
Middle age precursor of modern data visualization.

catalan atlas (a. cresques, 1375)



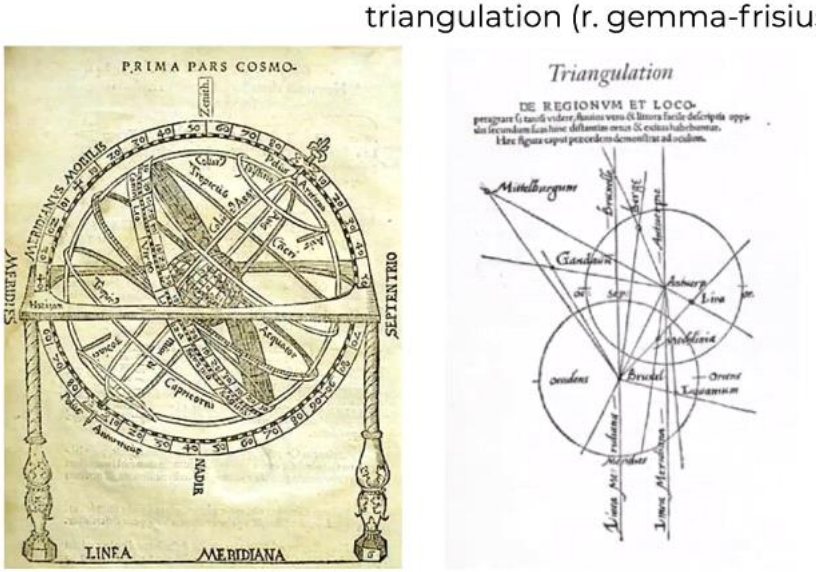
Ritaglio schermata acquisito: 02/03/2021 11:57

Cartography stil went on. There is also a tematic rapresentation. 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.



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

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

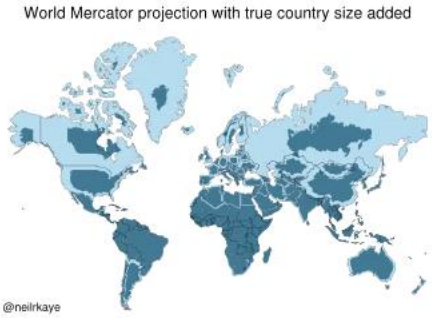


Ritaglio schermata acquisito: 02/03/2021 12:00

1600 modern way of exploring the wolrd. So it is important to understand where you are in a precise moment.
This are two rapresentation which use triangulations (angles and meridians)

cylindrical projection (g. mercator, 1569)

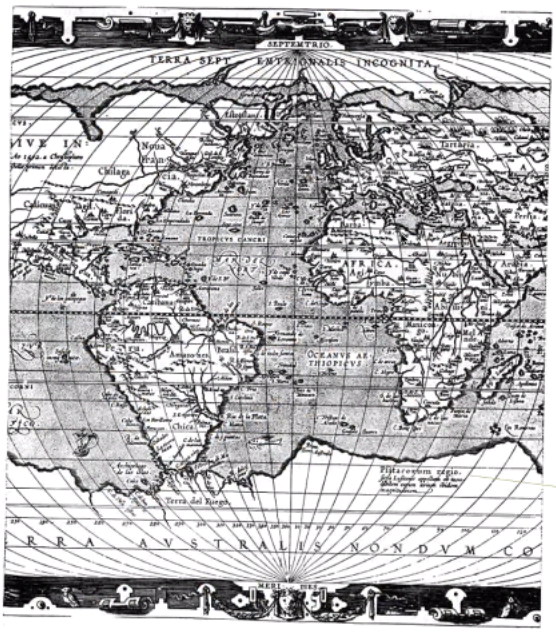
cylindrical projection for portraying the globe on maps, to preserve straightness of rhumb lines

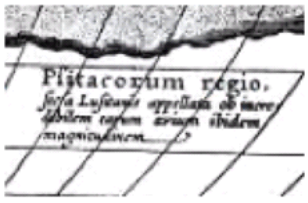


Ritaglio schermata acquisito: 02/03/2021 12:01

A big step occured when geographer understood that you can rapresent the world on a plane. So they developed the concept of projection which is used still today. Mercator projection for norther europe and america distorce the real size.

first modern atlas (a. ortelius, 1570)



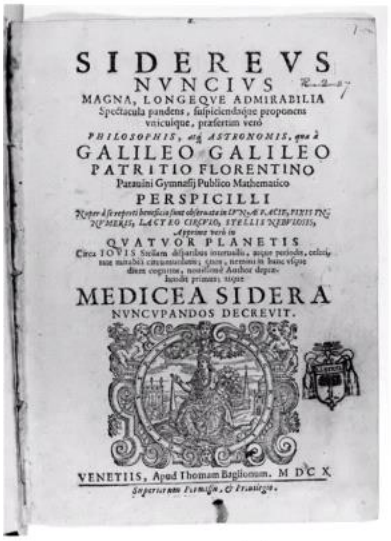


Ritaglio schermata acquisito: 02/03/2021 12:04

Atlanses is a collection of map : global map and extraction of local map from the bigger one.
1600 was characterized by huge differnces:
beginnings of visual thinking : emergence of quantitative tools (distances and measuremnt)
Physical measurement

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

printed astronomical images (g. galilei, 1610)


SIDEREVS
NVNCIVS
MAGNA, LONGEQVE ADMIRABILIA
Spectacula pandens, suspiciendaque proponens
unicuique, praesertim verò
PHILOSOPHIS, atq; ASTRONOMIS, quæ à
GALILEO GALILEO
PATRITIO FLORENTINO
Patauini Gymnasij Publico Mathematico
PERSPICILLI
QVATVOR PLANETIS
MEDICEA SIDERA
NVNCVPANDOS DECREVIT.
VENETIIS, Apud Thomam Baglionum. M D C X

telescope observations

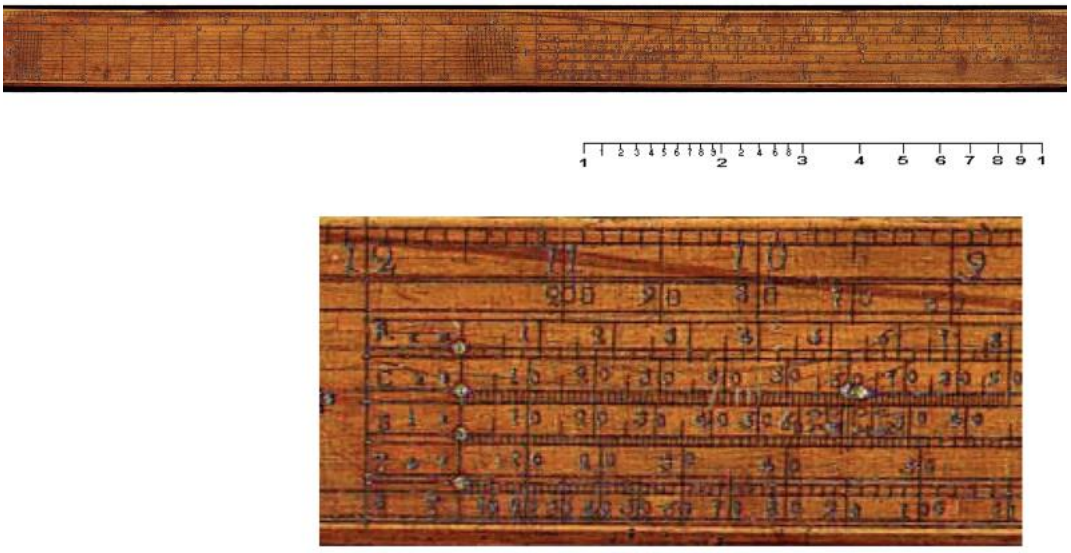

OBSERVAT. SIDEREAE
dum daturam. Depressiores insuper in Luna cernuntur magnæ maculæ, quàm clariores plagæ; in illa enim tam crescente, quam decrescente semper in lucis tenebrarumque confinio, prominente hinc inde circa ipsas magnas maculas contermini partis lucidioris; veluti in describendis figuris obseruauimus; neque depressiores tantummodo sunt dictarum macularum termini, sed æquabiliores, nec rugis, aut asperitatibus interrupti. Lucidior verò pars maximè propè maculas eminet; adeò vt, & ante quadraturam primam, & in ipsa fermè secunda circa maculam quandam, superiorem, boream nempè Lunæ plagam occupantem valdè attollantur tam supra illam, quàm infra ingentes quasdā eminentiæ, velati appositæ præseferunt delineationes.
Hæc

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.

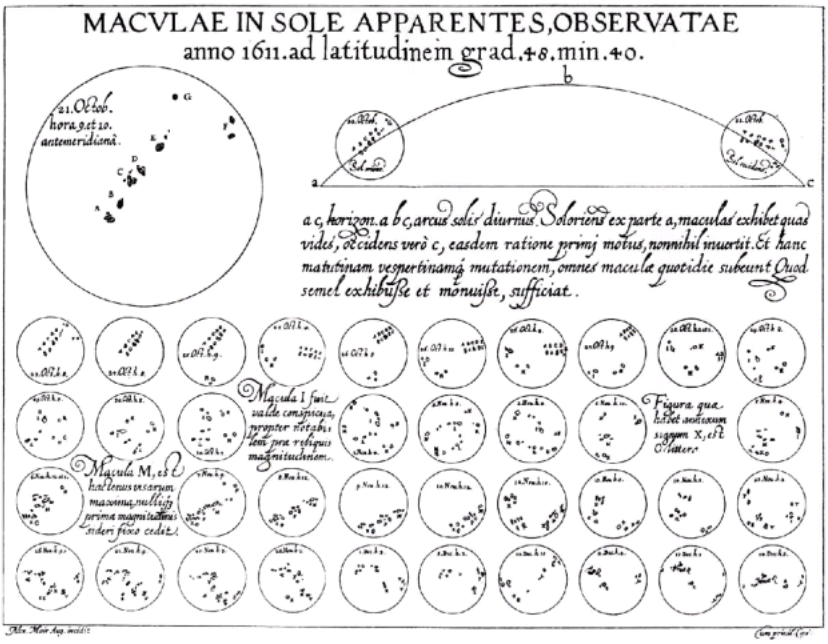
first logarithmic [gunter] scale (e. gunter & w. oughtred, 1620/1628)



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

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

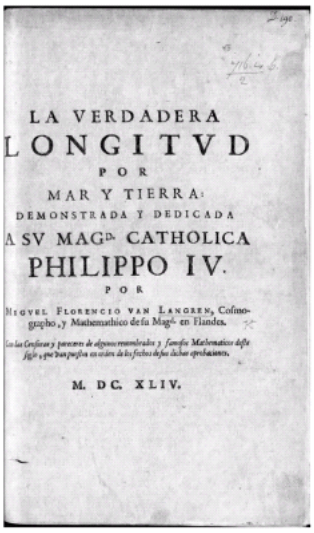
sunspots (c. scheiner, 1626)



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

Here we have a masterpiece.
This rapresentation is devoted to sunspots. This rapresenation 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)


LA VERDADERA
LONGITVD
POR
MAR Y TIERRA:
DEMONSTRADA Y DEDICADA
A SV MAGD. CATHOLICA
PHILIPPO IV.
POR
MIGVEL FLORENCIO VAN LANGREN, Cosmographo, y Mathematico de su Magd. en Flandes.
M. DC. XLIV.

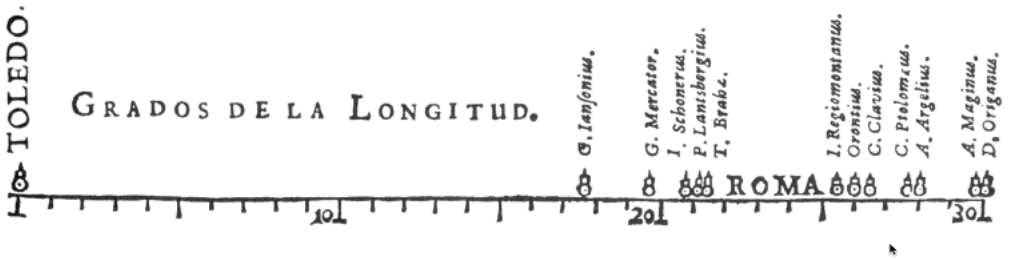
· variations in determination of longitude between *toledo* and *rome*

	Name	Longitude	Year	Longname	City	Country	Latitude
1	G. Jansonius	17.736	1605	Jan Jansson	Amsterdam	Flanders	52.37306
2	G. Mercator	19.872	1567	Gerardus Mercator	Leuven	Flanders	50.88333
3	I. Schonerus	20.638	1536	Johann Schoner	Bamberg	Germany	49.89167
4	P. Lantsbergius	21.106	1530	Phillip van Lansberge	Middelburg	Belgium	51.50000
5	T. Brahe	21.447	1578	Tycho Brahe	Hven	Denmark	55.86667
6	I. Regiomontanus	25.617	1463	Johann Muller	Nuremberg	Germany	49.45000
7	Orontius	26.000	1542	Oronce Fine	Paris	France	48.86667
8	C. Clavius	26.340	1567	Christoph Clavius	Rome	Italy	41.90000
9	C. Ptolomeus	27.787	150	Claudius Ptolemaeus	Alexandria	Egypt	31.19806
10	A. Argelius	28.170	1610	Andrea Argoli	Padua	Italy	45.41667
11	A. Maginus	29.787	1582	Giovanni Antonio Magini	Bologna	Italy	44.48333
12	D. Origanus	30.128	1601	David Origanus	Frankfurt	Germany	52.34710

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

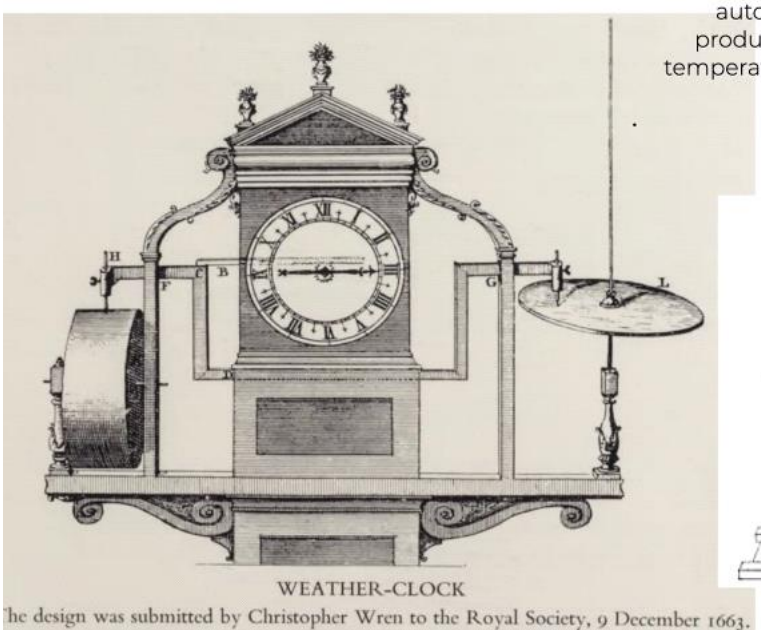
first data graph (m.f. van langren, 1644)

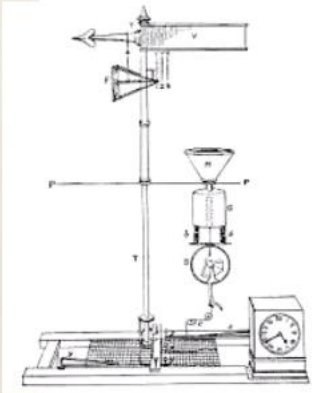
· variations in determination of longitude between *toledo* and *rome*

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

weather clock (c. wren, 1663)

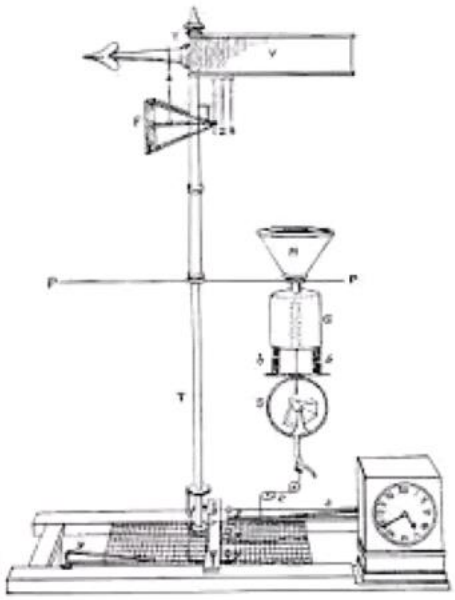
automatic recording device
producing a moving graph of
temperature and wind direction
(in polar coordinates)





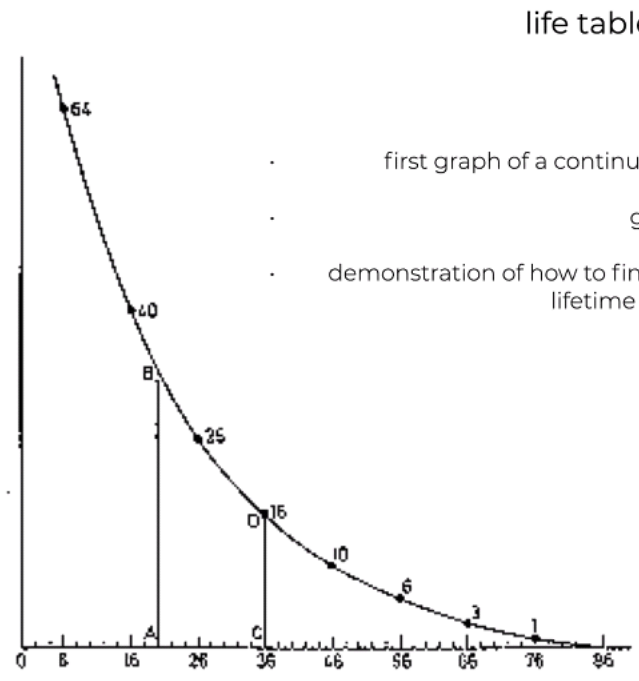
Ritaglio schermata acquisito: 02/03/2021 12:16

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.

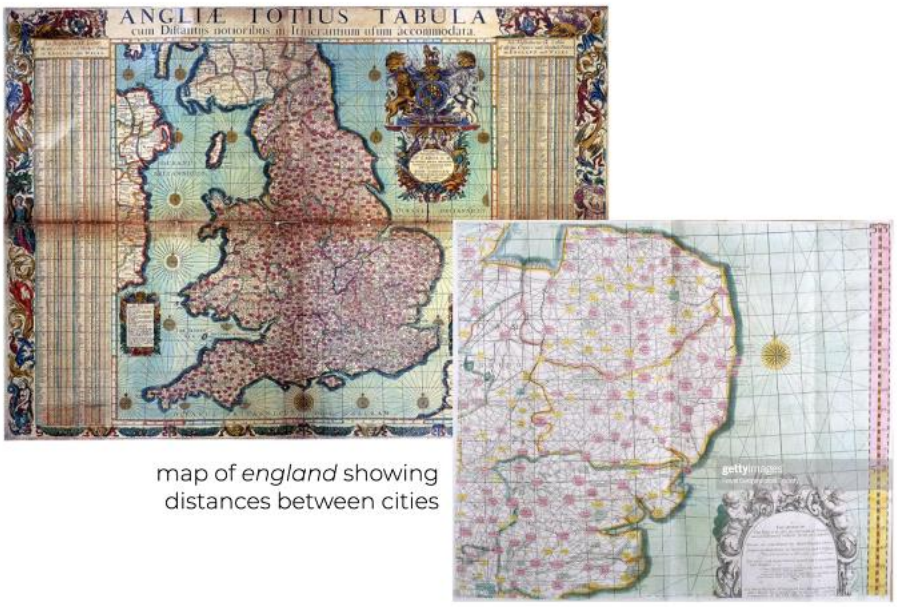
Ritaglio schermata acquisito: 02/03/2021 12:17



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

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

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



map of *england* showing distances between cities

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

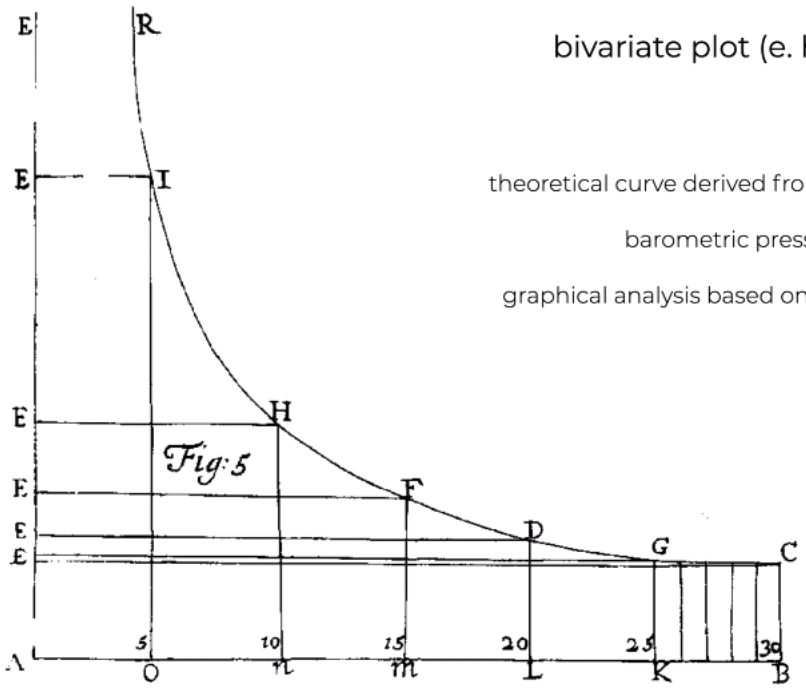
In the same period we have for the first time in cartography, we have an additional layer of information represented as a network. A network of towns connected with lines annotated with numbers (distances). Cartographic information with additional information.

bivariate plot (e. halley, 1686)

theoretical curve derived from observations

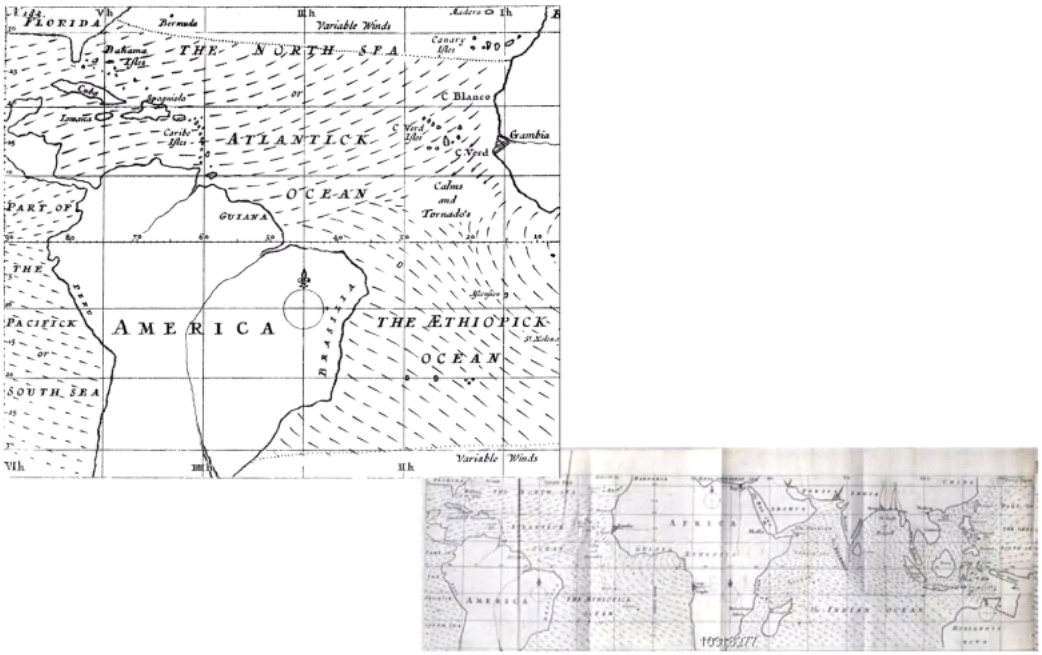
barometric pressure vs. altitude

graphical analysis based on empirical data



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

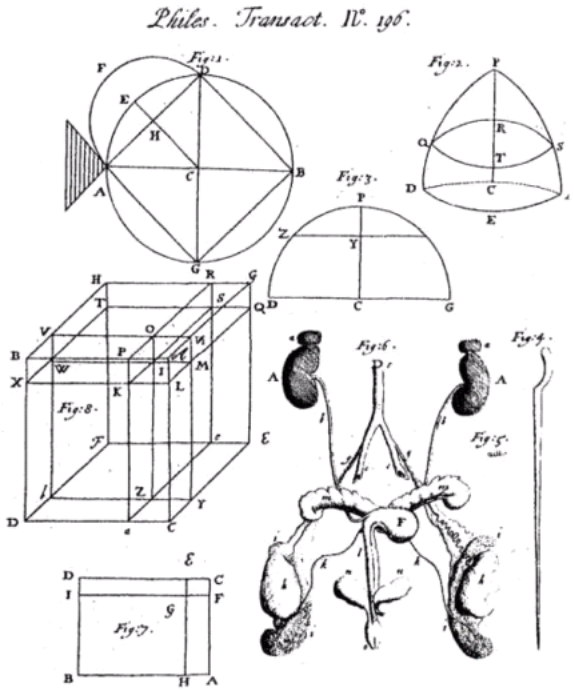
first weather map (e. halley, 1686)



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

Halley had the idea of plotting the corrent 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.

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)

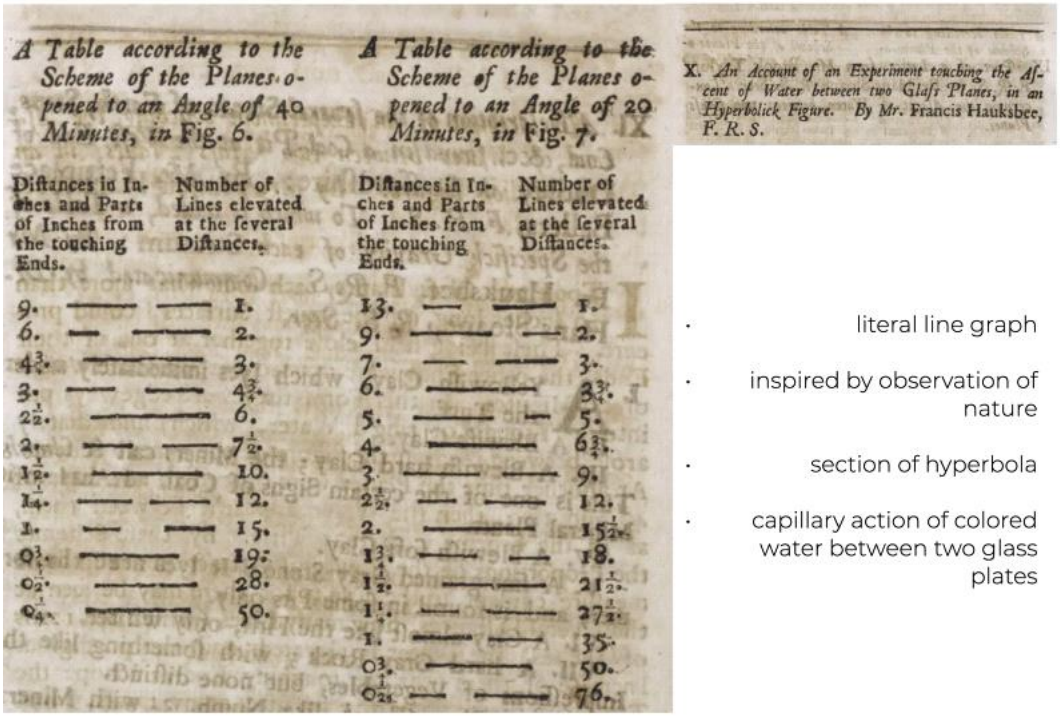
- Thematic mapping of physical quantities, geologic economic and medical data.
- Early beginnings of statistical theory (measurament error)
- Novel visual forms
- Reprodcution of data images (color printing, litography)

first contour map (e. halley, 1701)

- 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

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

literal line graph (f. hauksbee, 1712)



A Table according to the Scheme of the Planes opened to an Angle of 40 Minutes, in Fig. 6.

Diſtances in Inches and Parts of Inches from the touching Ends.	Number of Lines elevated at the ſeveral Diſtances.
9.	1.
6.	2.
4¾.	3.
3.	4¾.
2½.	6.
2.	7½.
1½.	10.
1¼.	12.
1.	15.
0¾.	19.
0½.	28.
0¼.	50.

A Table according to the Scheme of the Planes opened to an Angle of 20 Minutes, in Fig. 7.

Diſtances in Inches and Parts of Inches from the touching Ends.	Number of Lines elevated at the ſeveral Diſtances.
13.	1.
9.	2.
7.	3.
6.	3¾.
5.	5.
4.	6¾.
3.	9.
2½.	12.
2.	15½.
1¾.	18.
1½.	21½.
1¼.	27½.
1.	35.
0¾.	50.
0½.	76.

X. *An Account of an Experiment touching the Aſcent of Water between two Glaſs Planes, 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

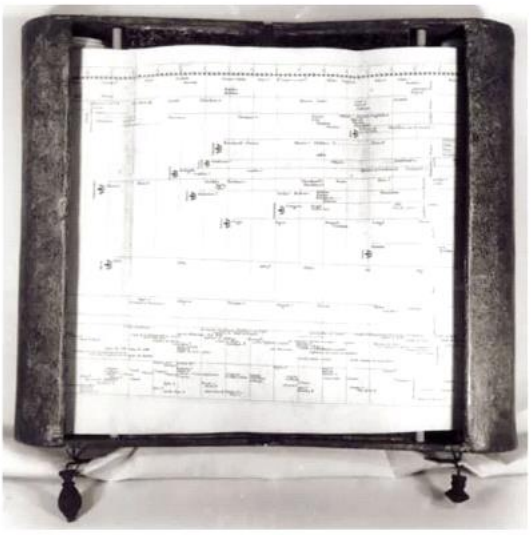
Ritaglio schermata acquisito: 02/03/2021 12:30

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.

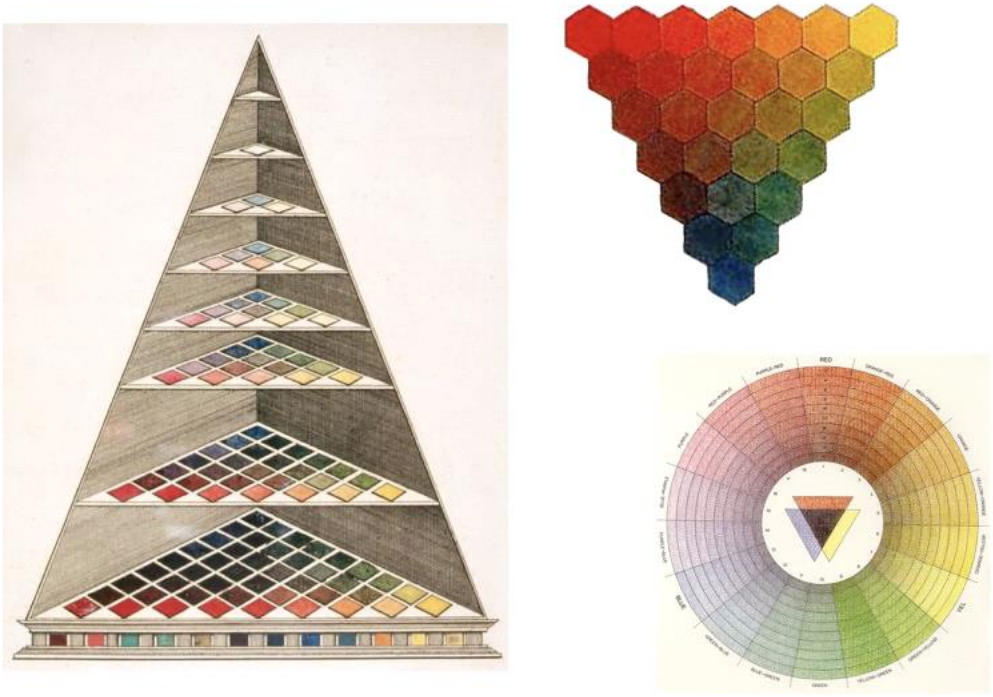
carte chronologique (j. barbeu-dubourg, 1753)



- 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

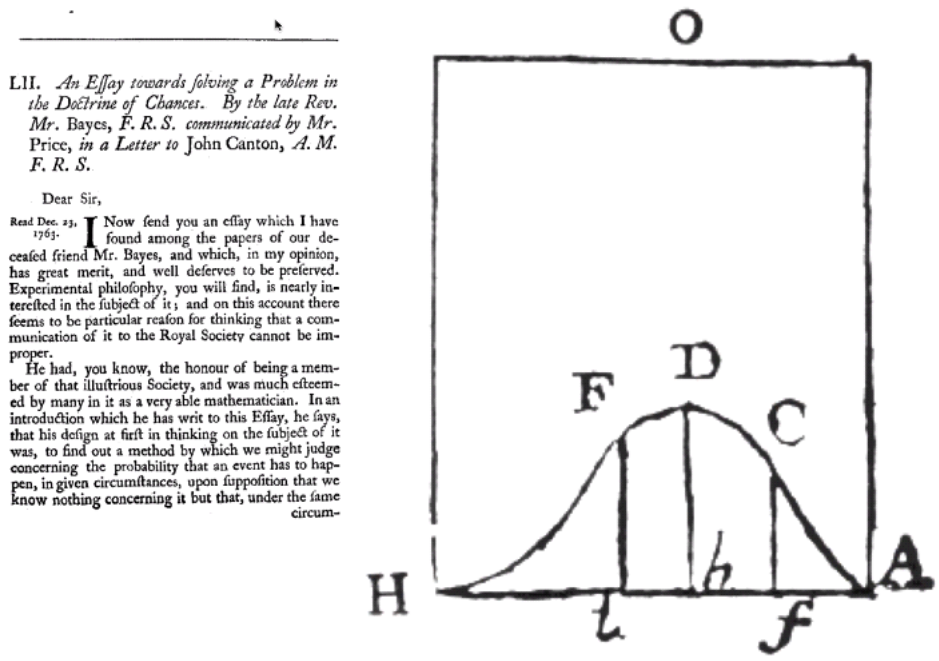
color system (m. harris/j.h. lambert/j.t. mayer, 1758-1772)



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

Systematic theory of colours. Theory of colour bacome rounded. This is a rapresentation 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 rapresentation.

beta density graph (t. bayes, 1763)



LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

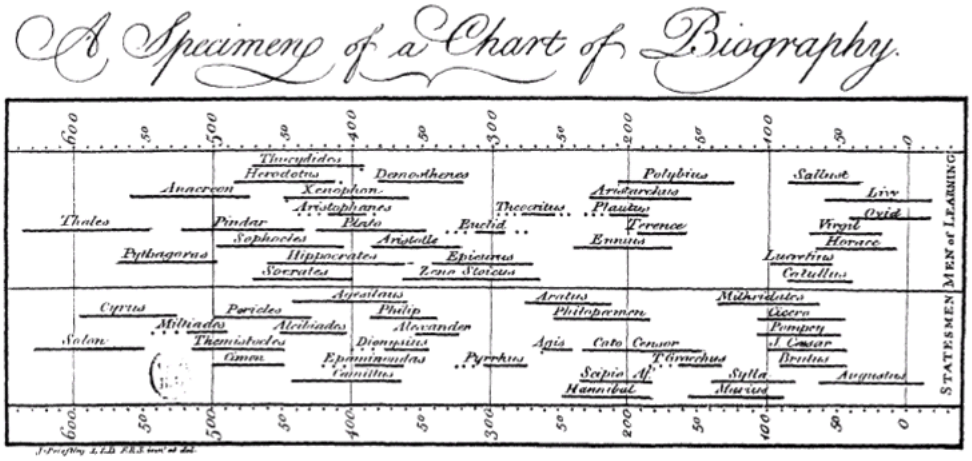
Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

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.

historical timeline (j. priestley, 1765)

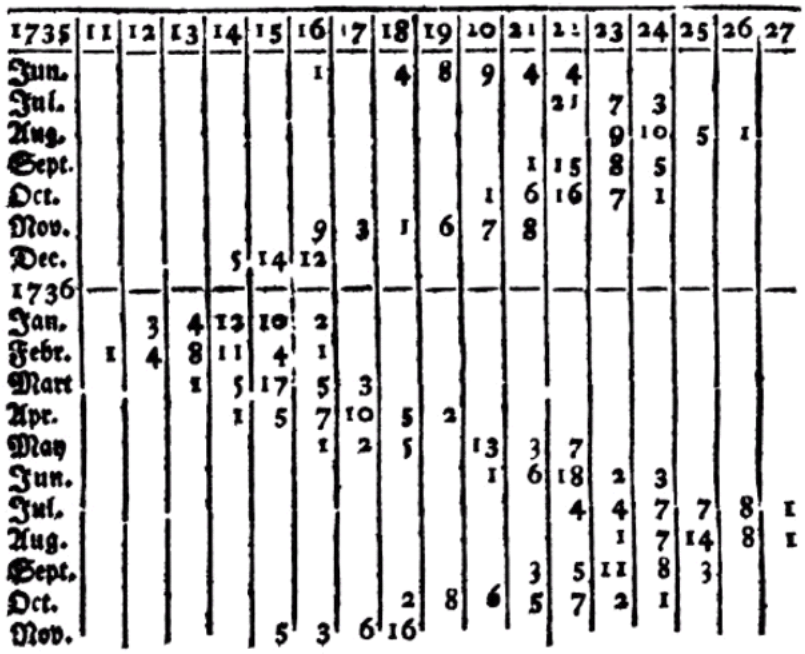


- life spans of 2,000 famous people, 1200 b.c. to 1750 a.d.
- quantitative comparison by means of bars

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.

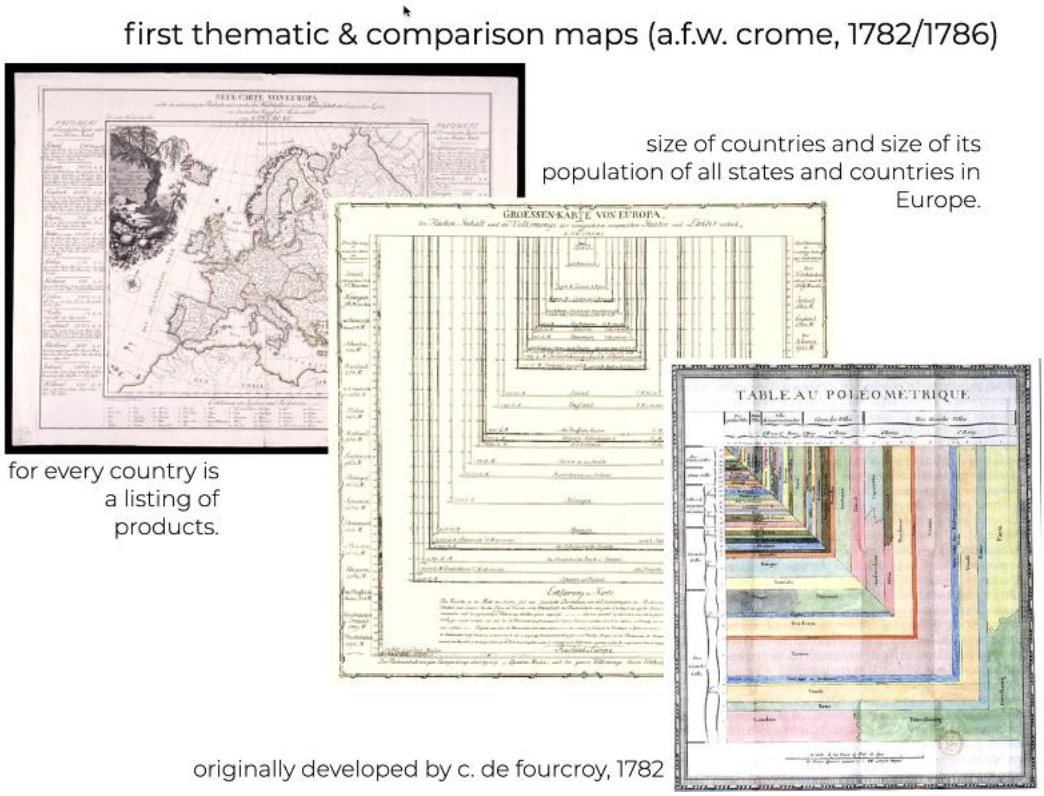
first semi-graphic display (j.h. lambert, 1779)



combining tabular and graphical formats

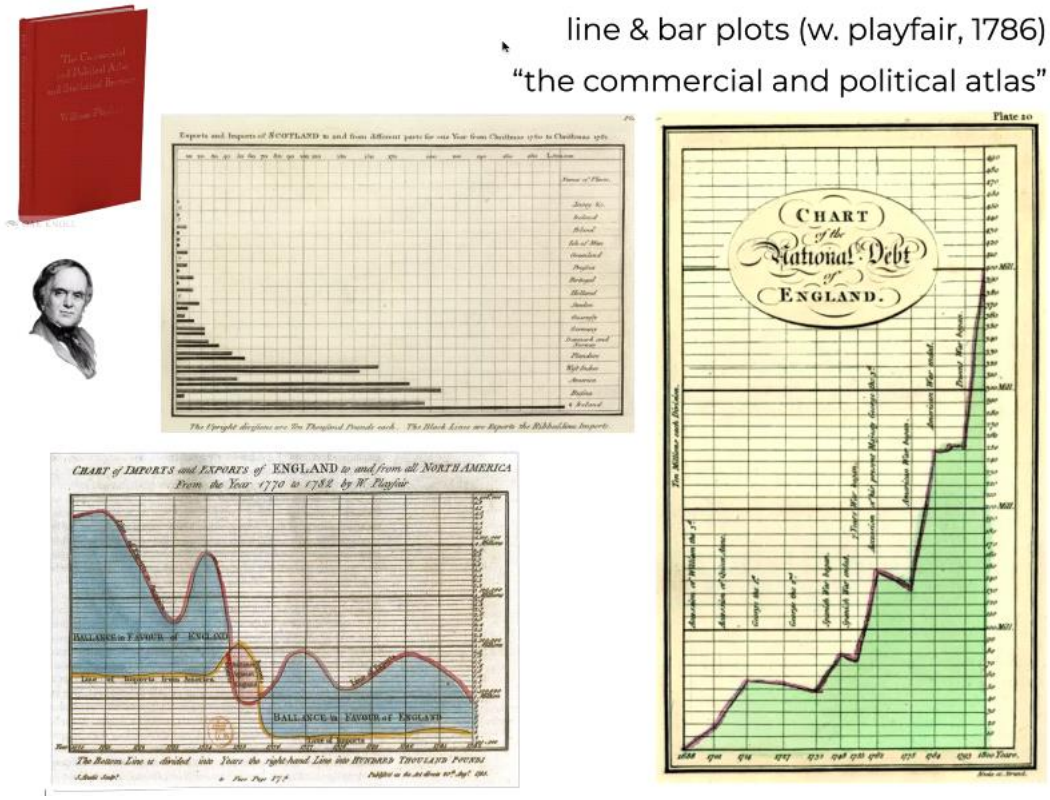
Ritaglio schermata acquisito: 02/03/2021 12:40

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



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

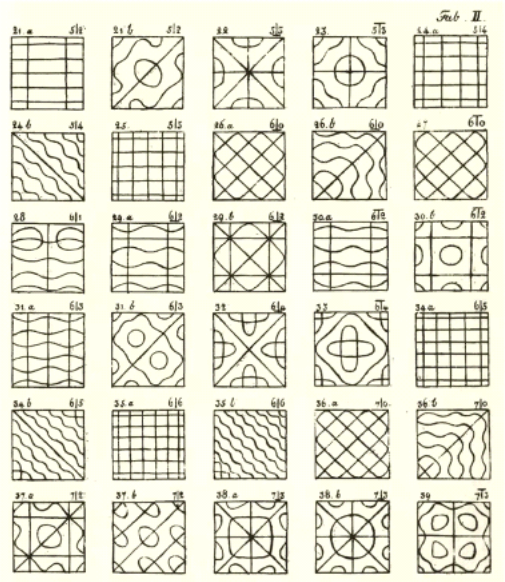
A new figure emerges, the combination of maps with statstics or demography.



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

This is the father of modern data visualization. We have informatipon 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.

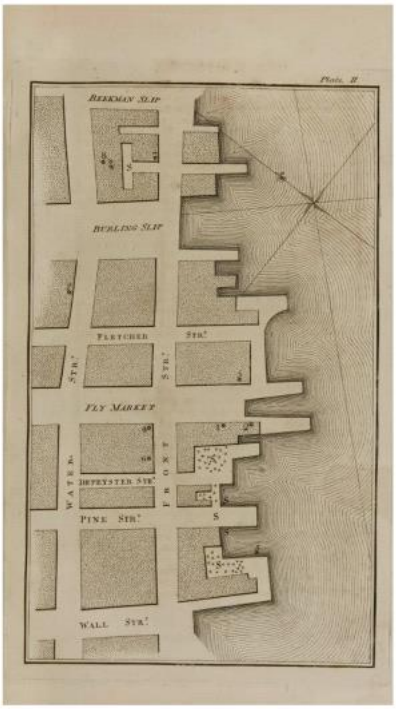
vibrational patterns (e.f.f. chiadni, 1787)



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

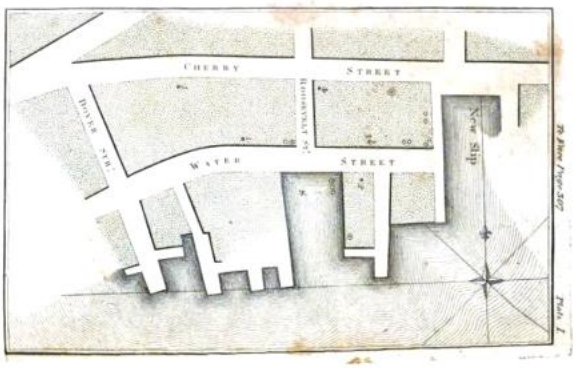
Ritaglio schermata acquisito: 02/03/2021 12:45

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

History of data visualization part 2

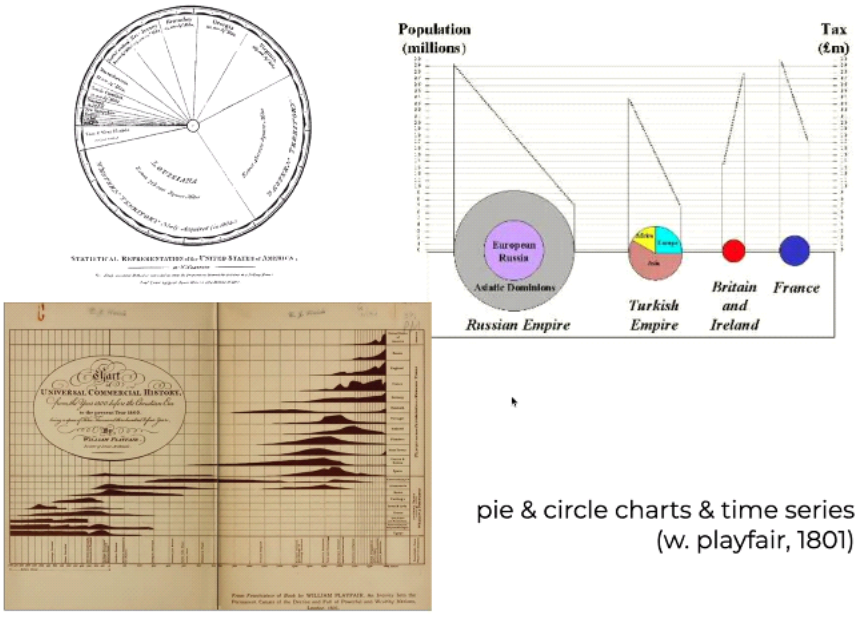
martedì 9 marzo 2021 11:30

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 comprhensive atlases.
Different fields approach to data viz.

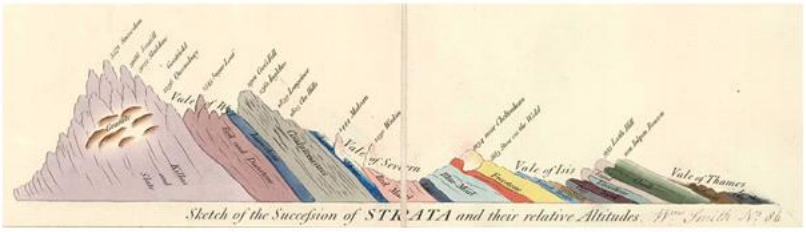


Ritaglio schermata acquisito: 09/03/2021 11:33

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

first uk geological map (w. smith, 1805)

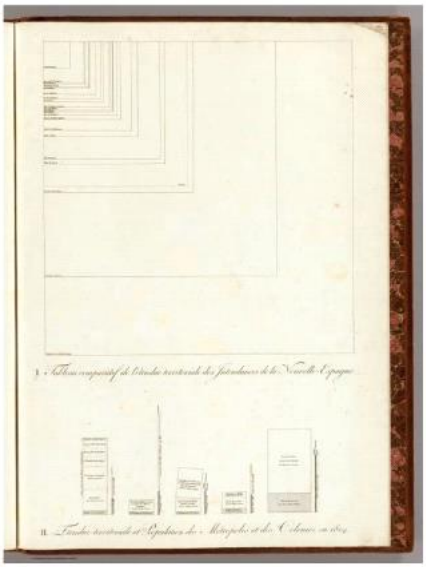
"the map that changed the world" [*winchester*,2001]



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 montain) 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)

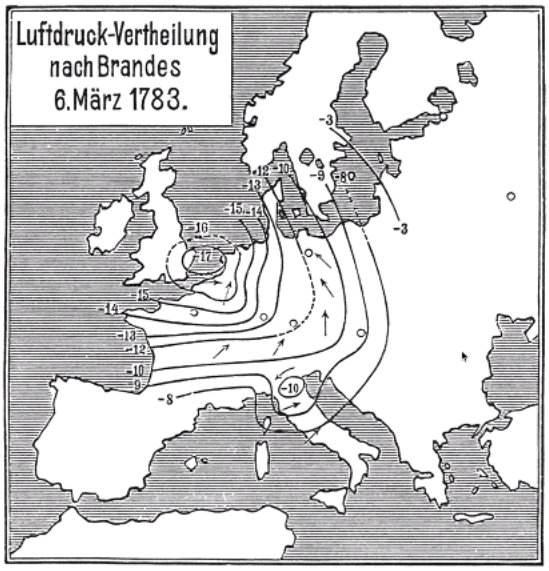


subdivided bar graphs & superimposed squares, showing the relative size of *mexican* territories and populations in the colonies

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

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

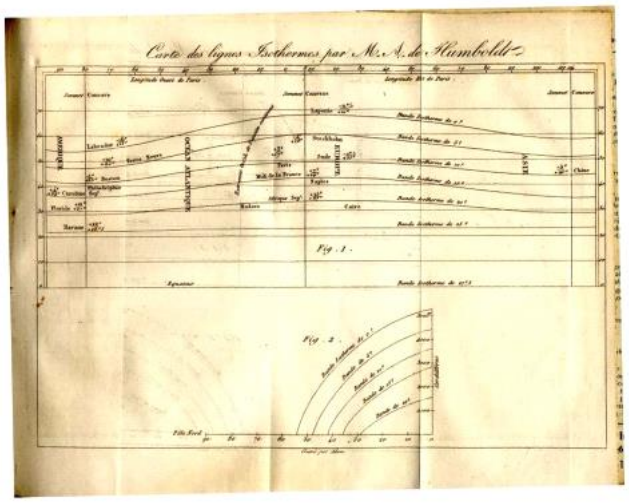
first weather map (h.w. brandes, 1816)



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)

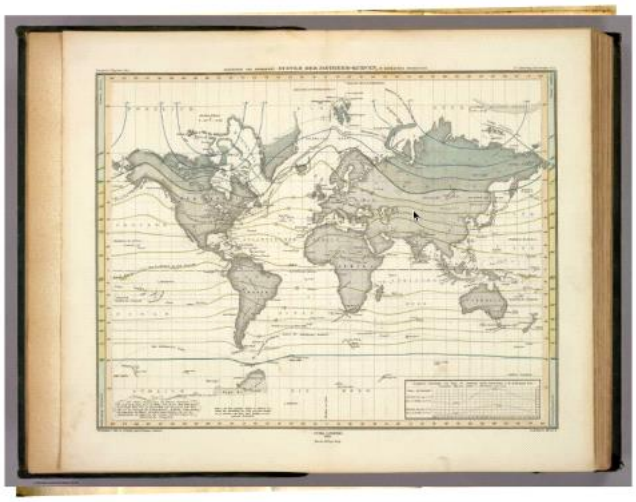


mean temperature around the world by latitude and longitude.

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)



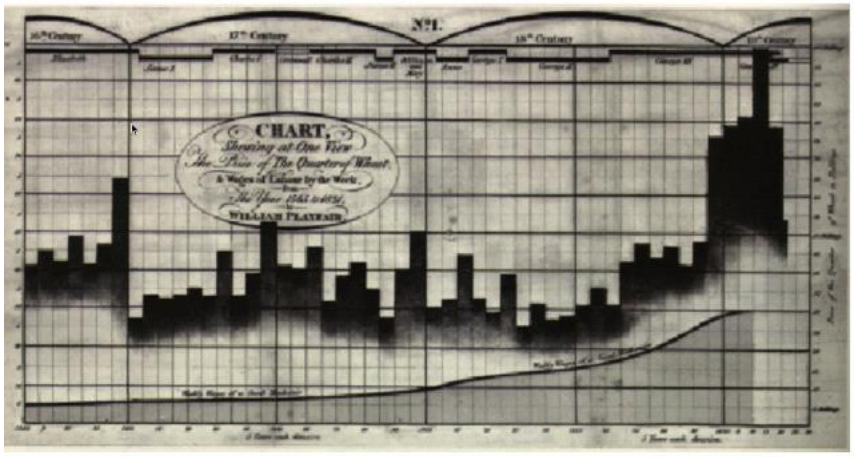
berghaus' 1849 atlas

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)



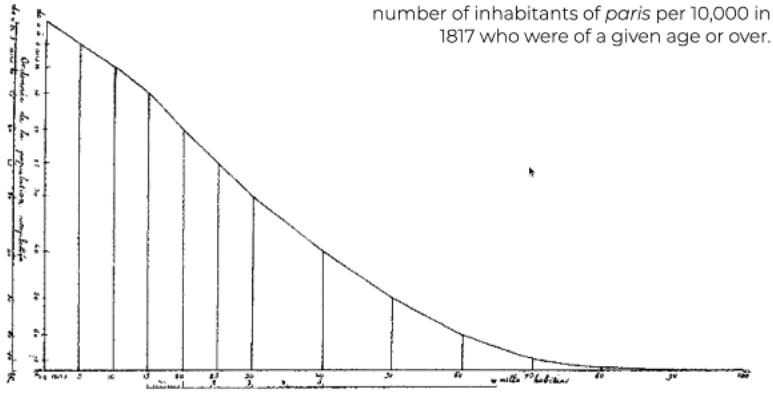
time series graph of prices, wages, and ruling monarch over a 250 year period

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

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

—

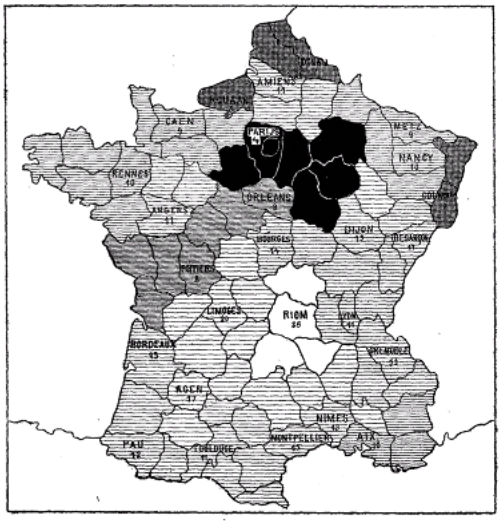
number of inhabitants of *paris* per 10,000 in 1817 who were of a given age or over.



Ritaglio schermata acquisito: 09/03/2021 11:41

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

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

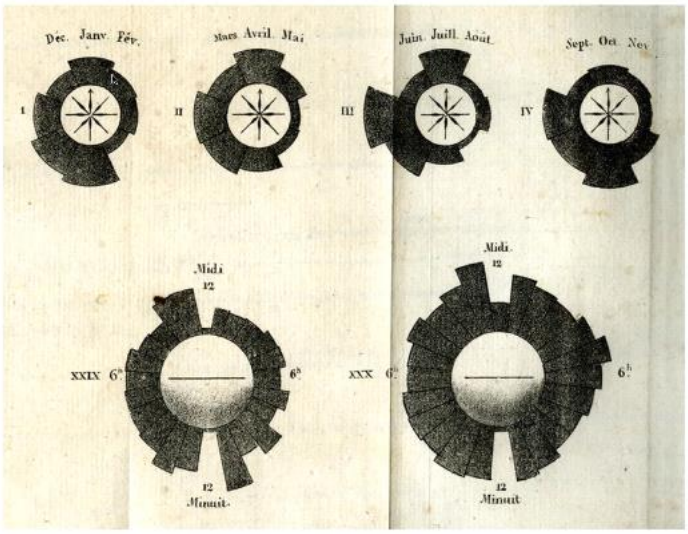


- shadings from black to white
- distribution and intensity of illiteracy in *france*
- first modern statistical map

Ritaglio schermata acquisito: 09/03/2021 11:42

This is a real milestope. 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.

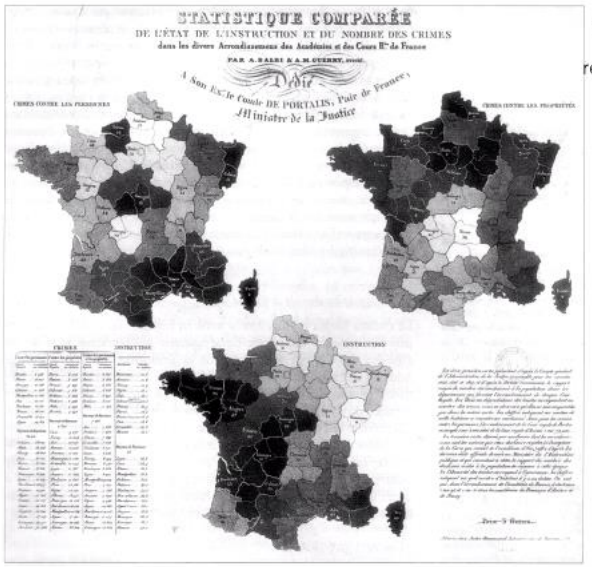
first polar area chart (a.m. guerry, 1829)



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

Ritaglio schermata acquisito: 09/03/2021 11:44

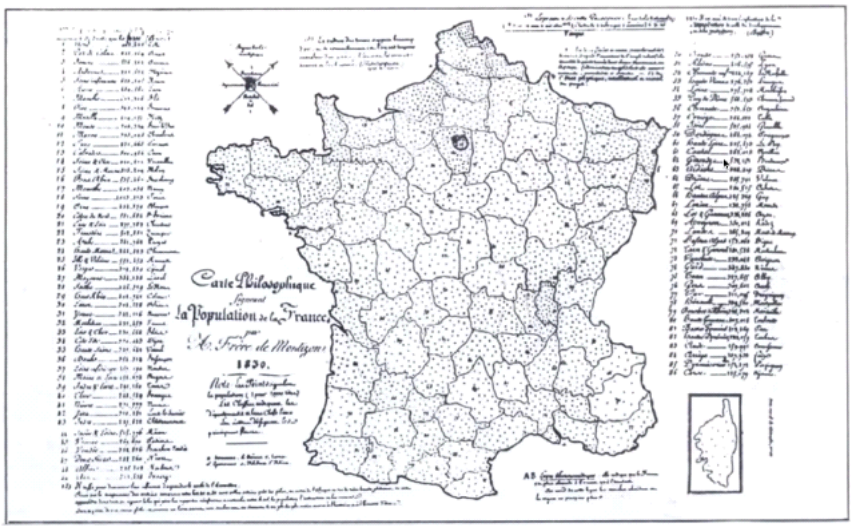
first comparative chloroplet maps (a.m. guerry & a. balbi, 1829)



crimes against persons and crimes against property in relation to level of instruction by departments in *france*

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

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

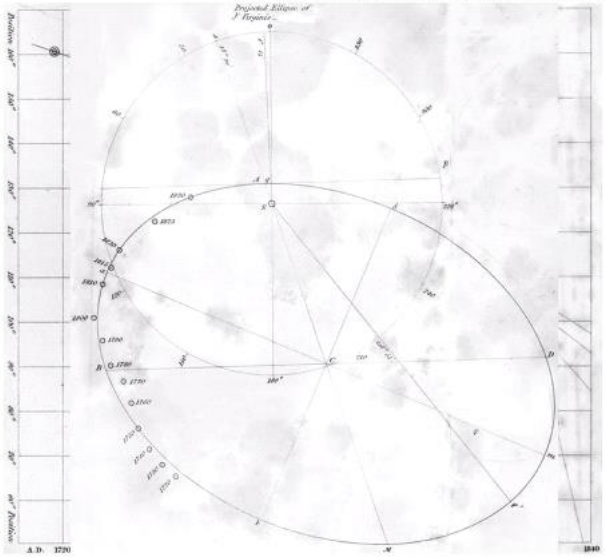


population by department, 1 dot = 10000 people

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.

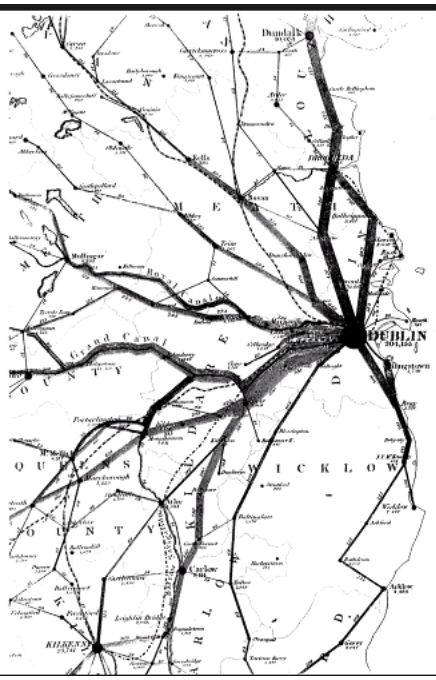
first fitting curve (j.f.w. herschel, 1832)



position vs time of *gamma virginis*

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

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



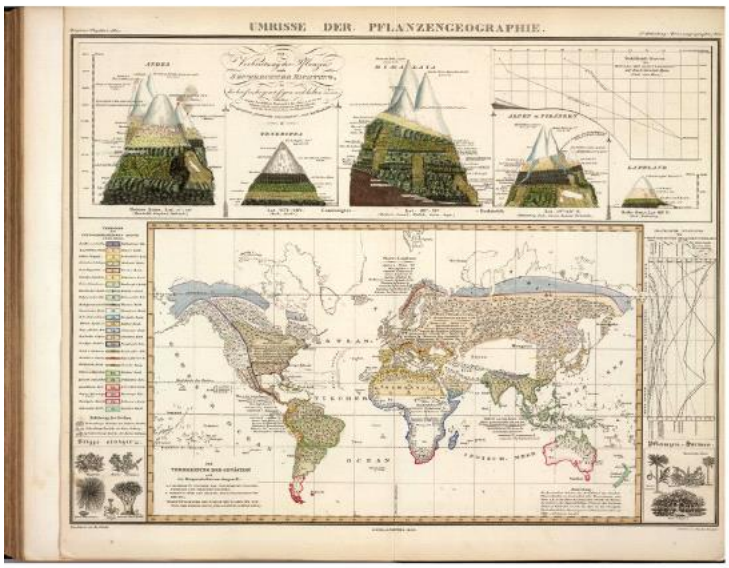
flow map (h.d. harness, 1837)

- passengers in *ireland*
- width proportional to number of passengers

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.

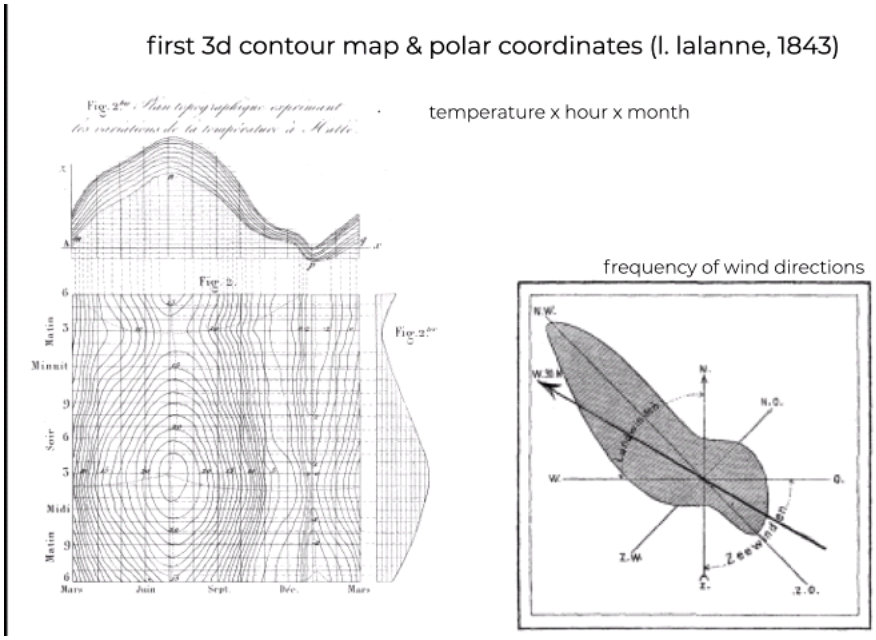
plant atlas (h. berghaus, 183



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

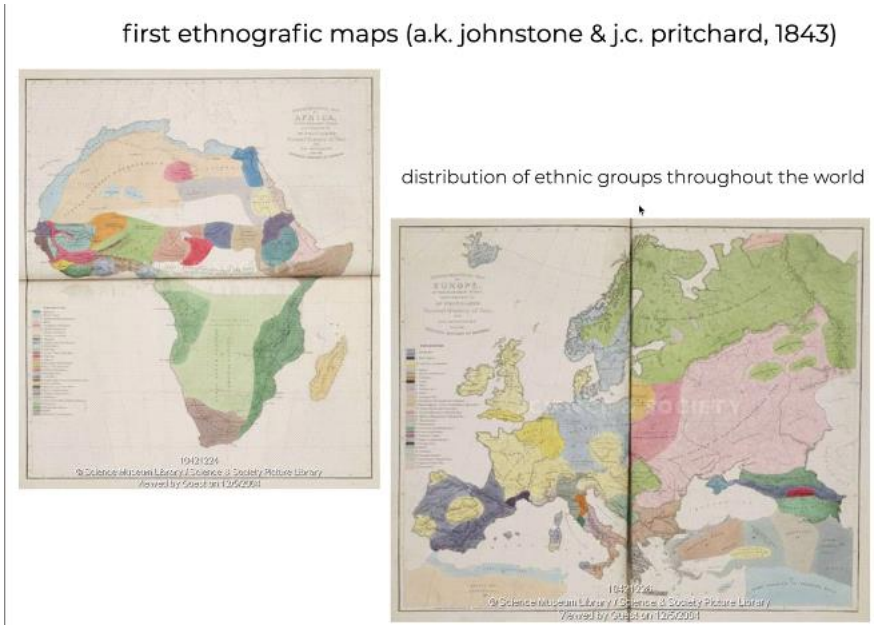
Plants scientists benefits as well from data viz. In this map it is shown the distribution of plants around the world. This is a complex plot representing the phenomenon of distribution of plants around the world.

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



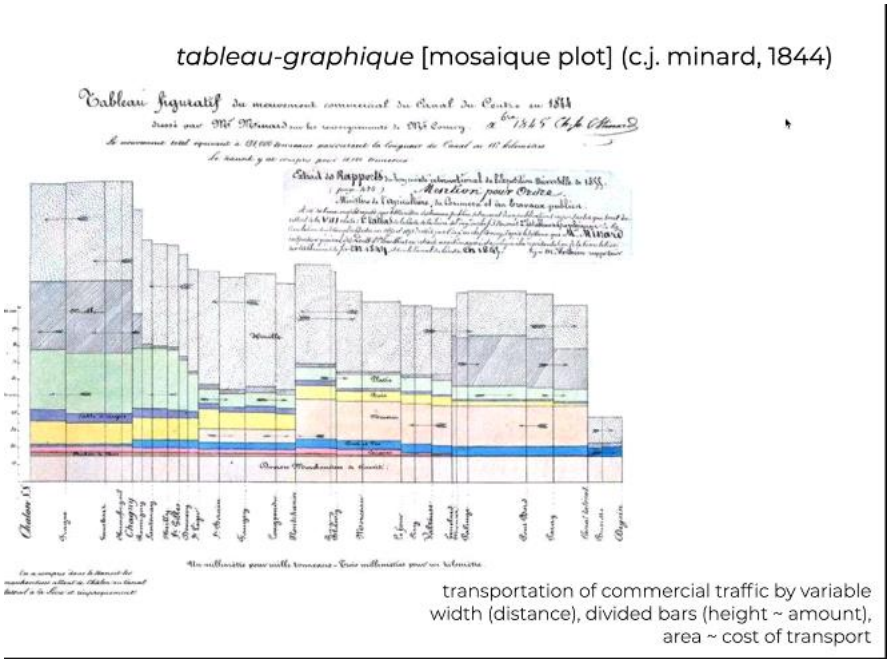
Ritaglio schermata acquisito: 09/03/2021 11:51

First 3-D countour map.



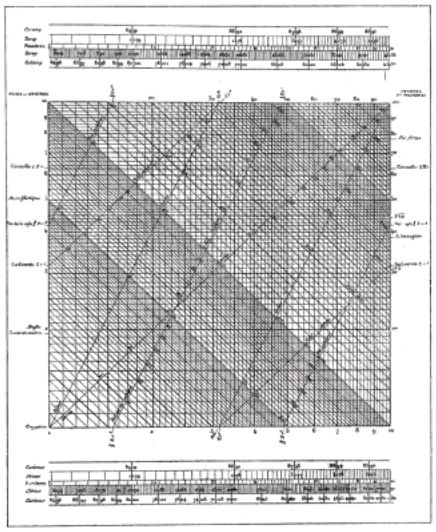
Ritaglio schermata acquisito: 09/03/2021 11:53

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



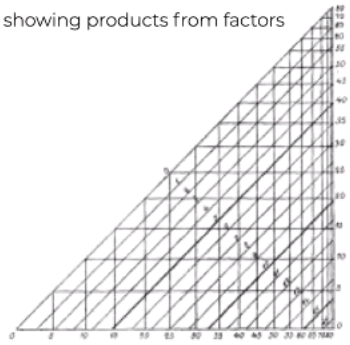
Ritaglio schermata acquisito: 09/03/2021 11:54

first log-log nomogram & universal calculator (l. lalanne, 1846)



logarithms and
trigonometric calculations

showing products from factors

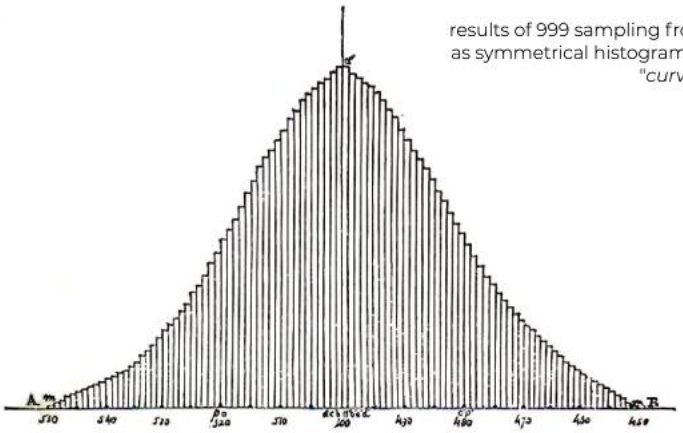


Ritaglio schermata acquisito: 09/03/2021 11:56

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)

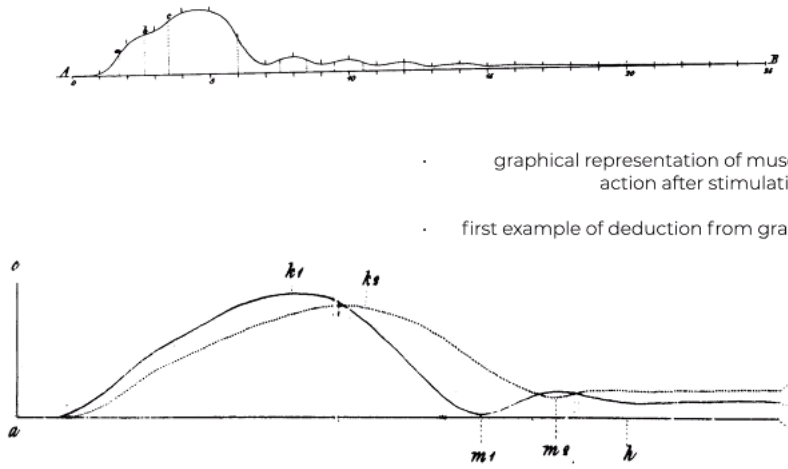
results of 999 sampling from urns shown
as symmetrical histograms, with limiting
"*curve of possibility*"



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)

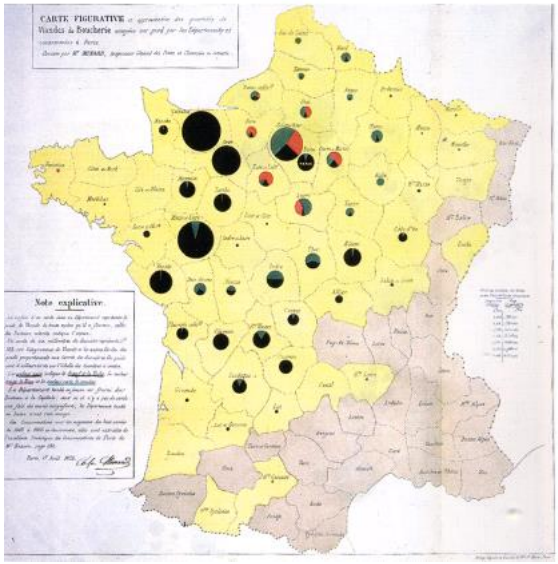


- graphical representation of muscle action after stimulation
- first example of deduction from graph

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 qwas the engine to derive an empircal law.

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

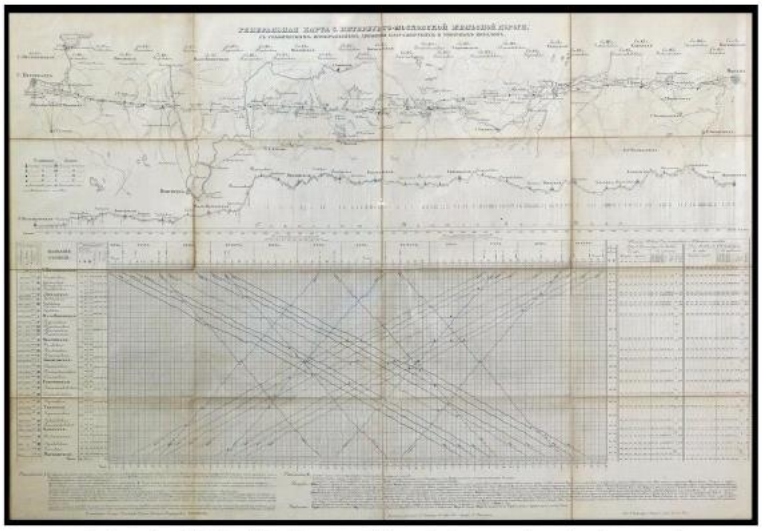


beef consumption

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

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

train schedule (serjev, 1854)

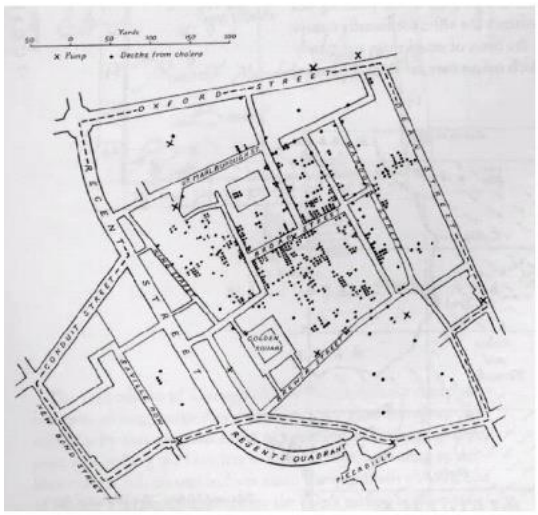


train schedule for 35 railways stations, between *st. petersburg* and *moscow*

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 philosphy. This is a quite complex map but it is thought to be a good support fro travellers.

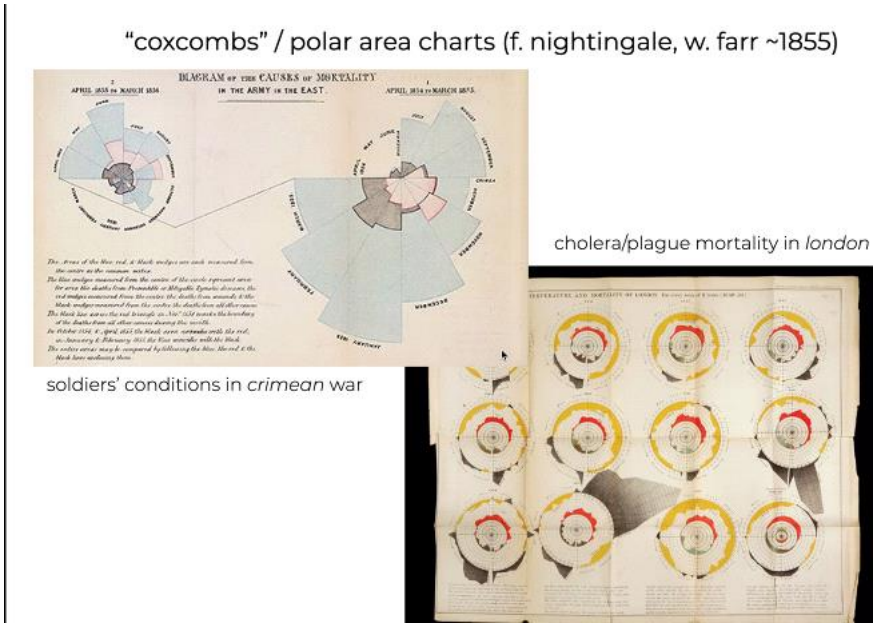
true epidemiological success w/ graphics (j. snow, 1855)



discovering a cholera outbreak in *london* using a map

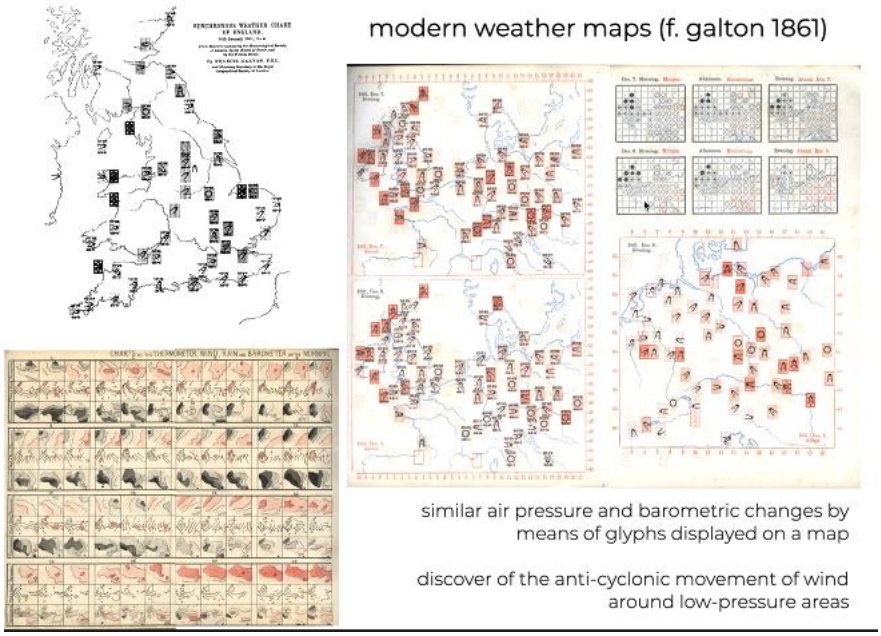
Ritaglio schermata acquisito: 09/03/2021 12:04

Now data viz is not just a representation tool but it is also a foundamental tool for understanding the analized phenomenon. This is the most common example -> the statisticain John Snow who studied the driving reason for the epidemics of colera.
He put a dot which idicates the place where infected people live. He noted that there was a cocnetration in a certain part. He the nundersttod that there was a correlation with the epidemic and the pump.



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.

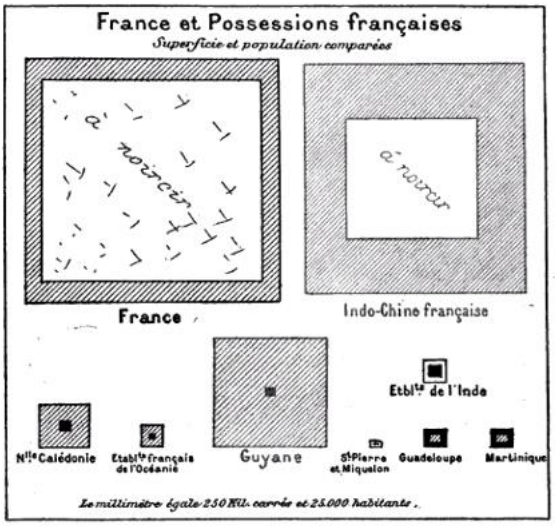


Ritaglio schermata acquisito: 09/03/2021 12:10

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

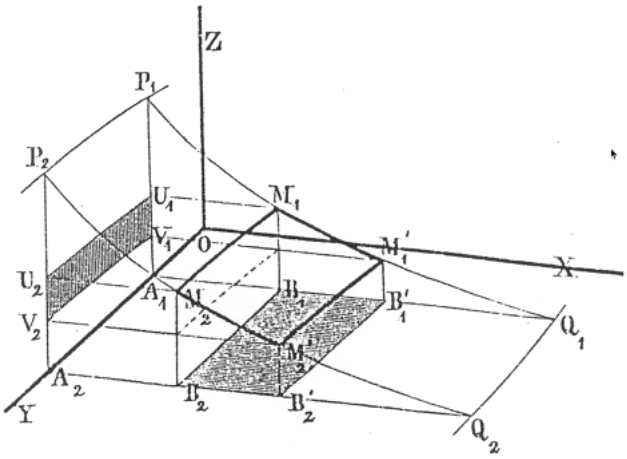
first chart used in textbooks (e. levasseur, 1868)



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

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

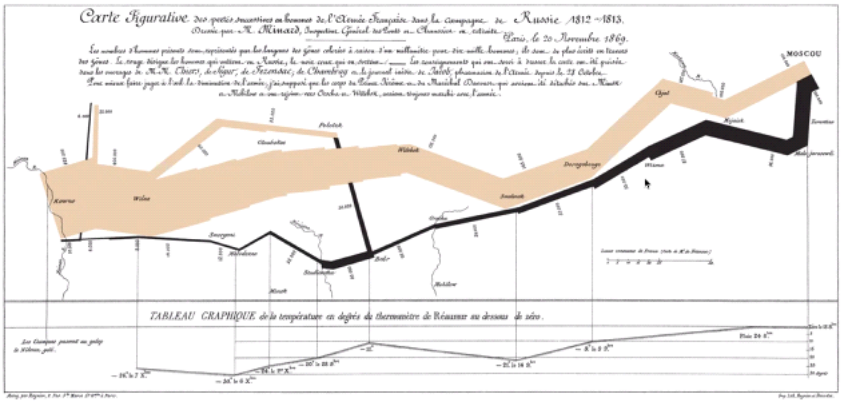
first stereograms (g. zeuner, 1869)



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

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

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

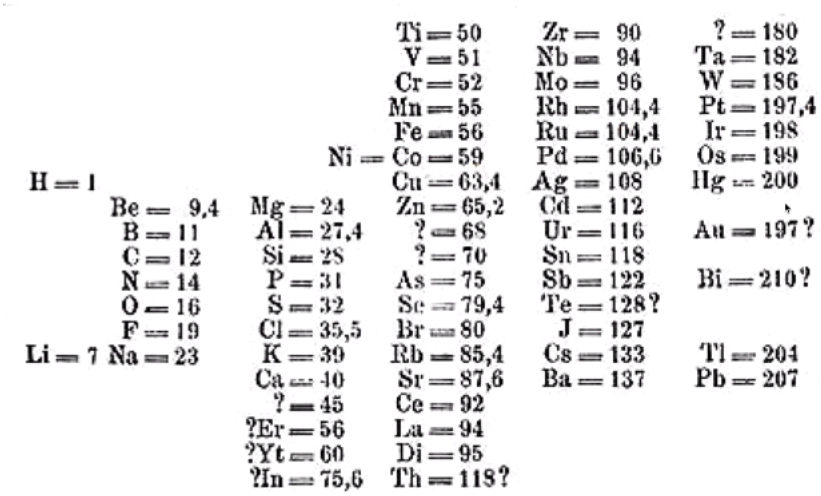


- the size of his army at specific geographic points during their advance/retire
- the distance traveled
- temperature
- slatitude and longitude
- direction of travel
- location relative to specific dates
- later called sankey diagram

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 comphrensive 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)

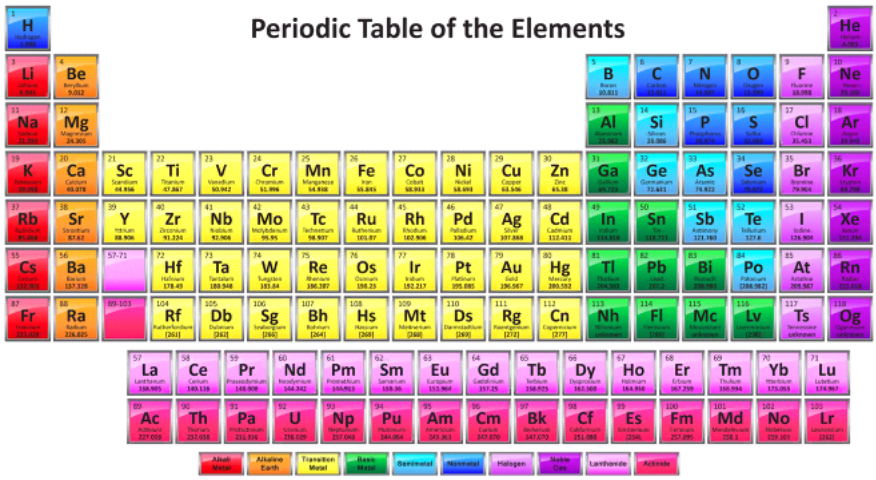


			Ti = 50	Zr = 90	? = 180
			V = 51	Nb = 94	Ta = 182
			Cr = 52	Mo = 96	W = 186
			Mn = 55	Rh = 104,4	Pt = 197,4
			Fe = 56	Ru = 104,4	Ir = 198
			Ni = Co = 59	Pd = 106,6	Os = 199
H = 1			Cu = 63,4	Ag = 108	Hg = 200
	Be = 9,4	Mg = 24	Zn = 65,2	Cd = 112	
	B = 11	Al = 27,4	? = 68	Ur = 116	Au = 197?
	C = 12	Si = 28	? = 70	Sn = 118	
	N = 14	P = 31	As = 75	Sb = 122	Bi = 210?
	O = 16	S = 32	Se = 79,4	Te = 128?	
	F = 19	Cl = 35,5	Br = 80	J = 127	
Li = 7	Na = 23	K = 39	Rb = 85,4	Cs = 133	Tl = 204
		Ca = 40	Sr = 87,6	Ba = 137	Pb = 207
		? = 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.

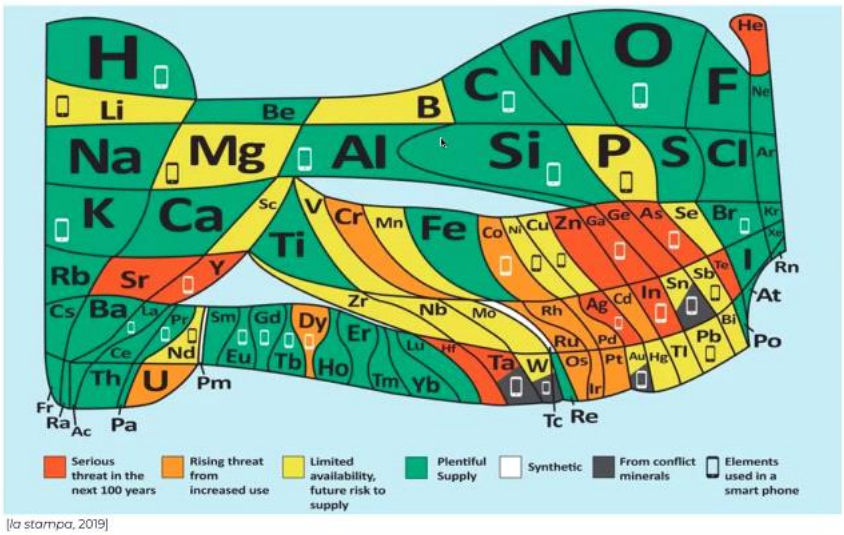
the periodic table (d. mendeleev, 1869)



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

Here we have color, position and annotations.

the periodic table (d. mendeleev, 1869)



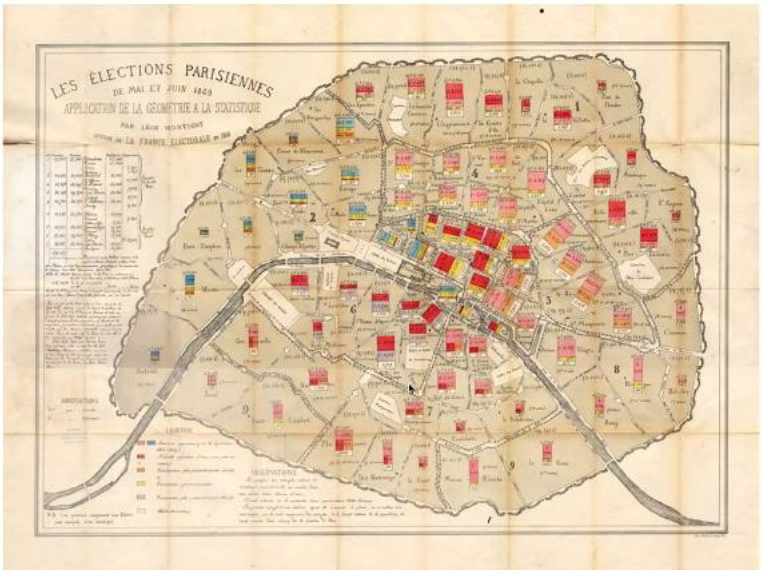
[la stampa, 2019]

even more enhanced - 2019 intl 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 aboundace of this elements on earth.

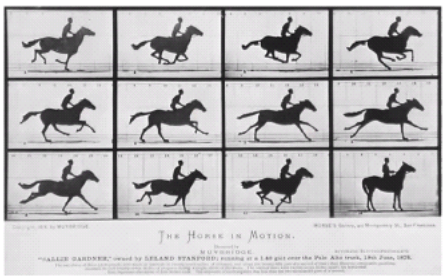
paris election map (l. montigny, 1870)



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

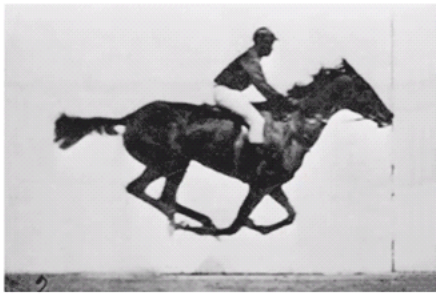
Paris election map -> with all the featrues connected to voting.

galloping horse (e. muybridge, 1872)



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

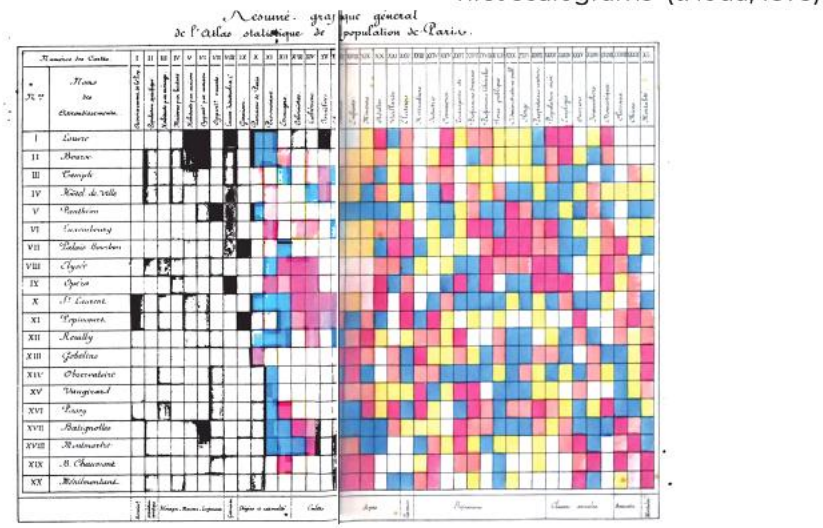
in 1882, *e.-j. marey* carried out several motion-picture camera experiments, recording a series of photographs to study fight of birds, running and walking



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)

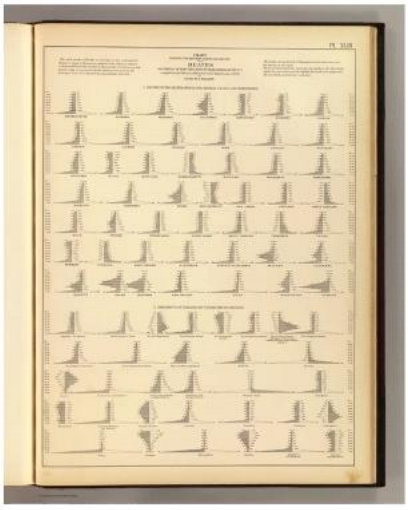


graphic summary of 40 maps of *paris*, each showing some feature of the population by arrondisement.

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)

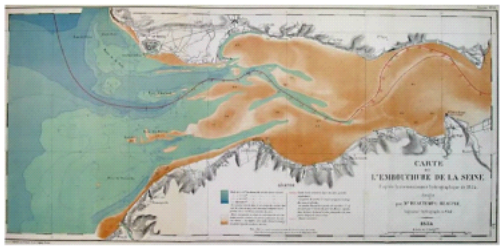


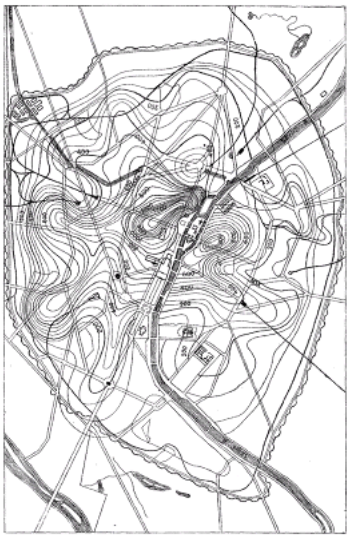
- age pyramid (bilateral histogram)
- bilateral frequency polygon
- use of subdivided squares to show the division of population
- first true u.s. national statistical atlas

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

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

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

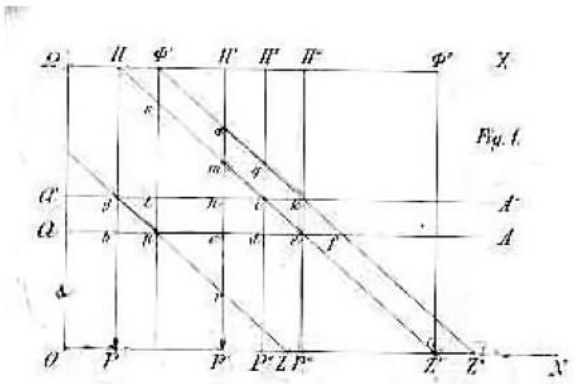




population density shown by contours

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

first lexis diagram (w. lexis, 1875)

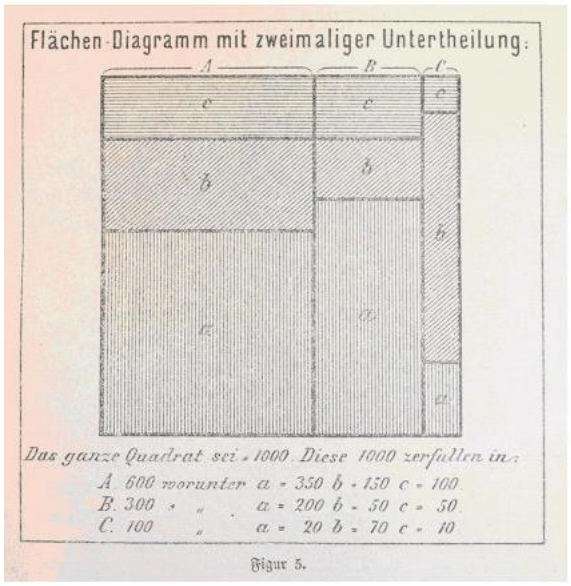


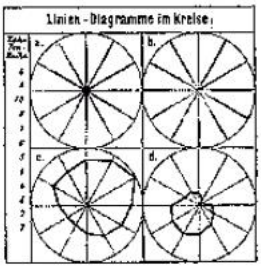
relations among age, calendar time, and life spans of individuals simultaneously

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

This is a quite complex 3D graph. It is diffuclt ot understand this graph.

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

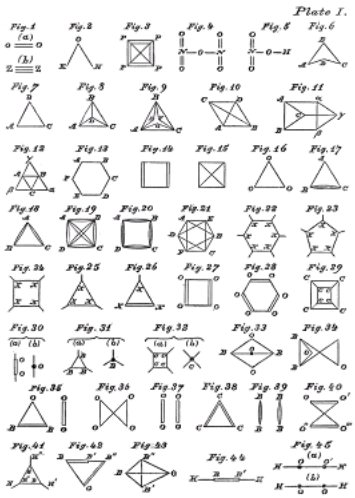




Ritaglio schermata acquisito: 09/03/2021 12:28

More and more types of representations appeared.

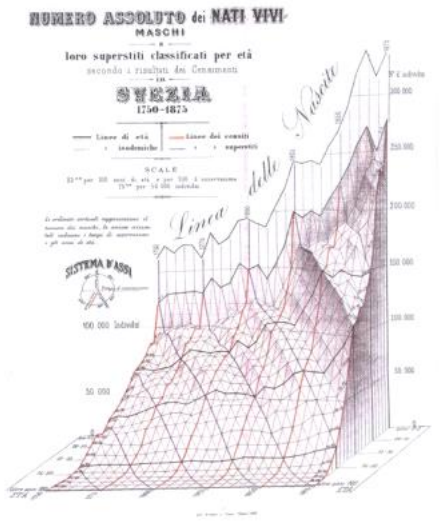
first use of word "graph" (j.j. sylvester, 1878)



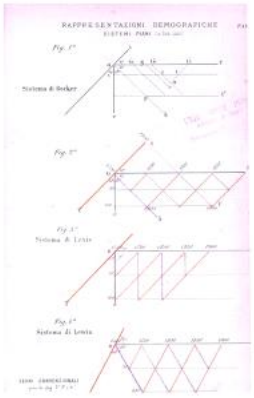
diagrams showing analogies between the chemical bonds in molecules and graphical representations of mathematical invariants

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

more refined stereogram (l. 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

venn diagrams (j. venn, 1880)

THE
LONDON, EDINBURGH, AND DUBLIN
PHILOSOPHICAL MAGAZINE
AND
JOURNAL OF SCIENCE.

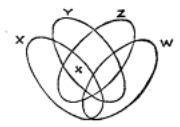
[FIFTH SERIES.]

JULY 1880.

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**.

these simpler cases somewhat carefully. The diagram for two terms, then, is to be thus drawn :—  On the common plan this would represent a *proposition*, and is, indeed, very commonly taken as illustrative of the proposition "Some X is Y."* With us it does not as yet represent a proposition at all, but only the framework into which propositions can be fitted; that is, it represents only the four combinations indicated by the letter-compounds XY, XY, XY, XY. Now conceive that we have to reckon also with the presence, and consequently with the absence, of Z. We just draw a third circle intersecting the two above, thus, 

For instance, the proposition "All X is Y" needs *both* the diagrams,  ; 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,



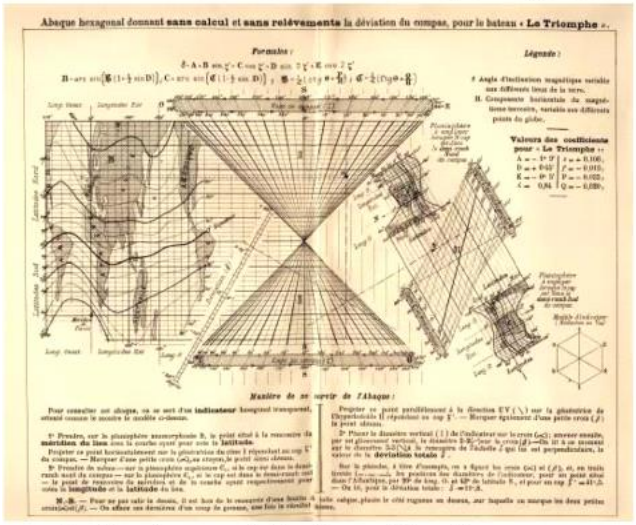
guishing any X that is neither Y nor Z—thus, 

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  . Scratch out, again, the XYZ compart-

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

Venn diagrams appears representing intersections of sets.

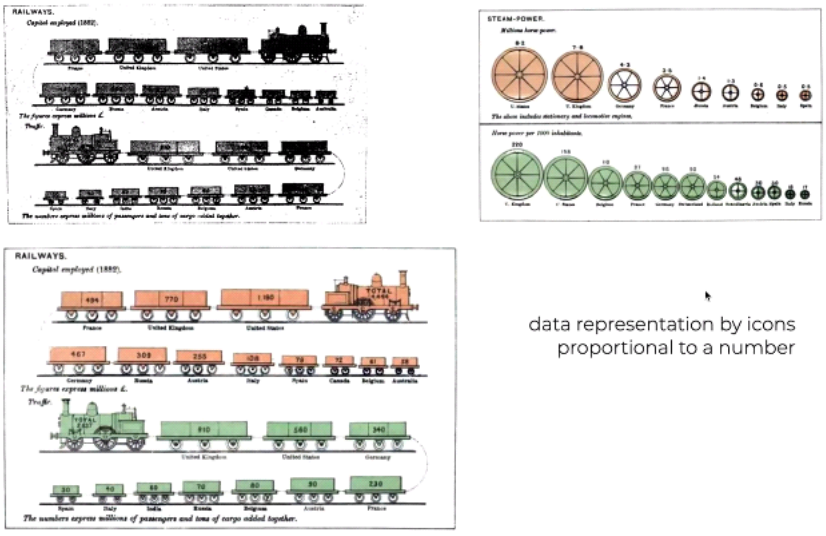
multifunction nomograms (c. lallemand, 1883-1885)



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

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

first pictograms (m.g. mulhall, 1884)

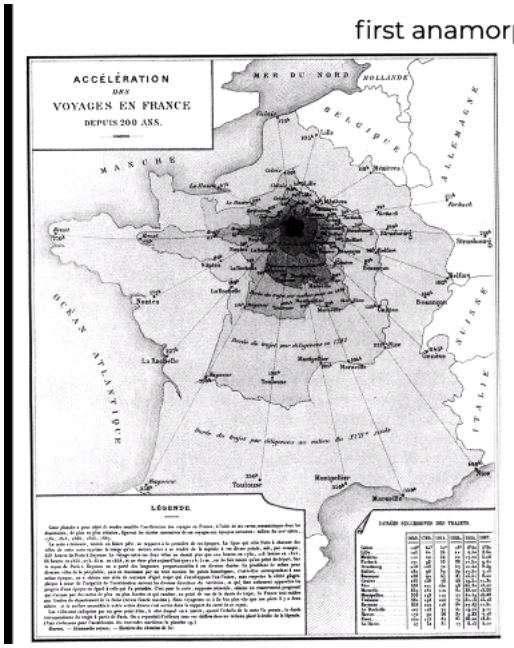


data representation by icons proportional to a number

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

Represting 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 untachable! The isoline of time means that it takes at least that amount of time to travel from paris to the place you want.

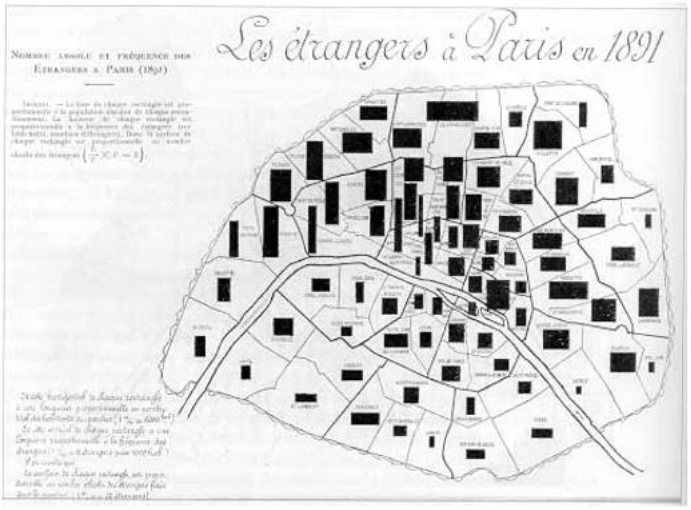
social mapping & color coding (c. booth, 1889)

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)



se 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)

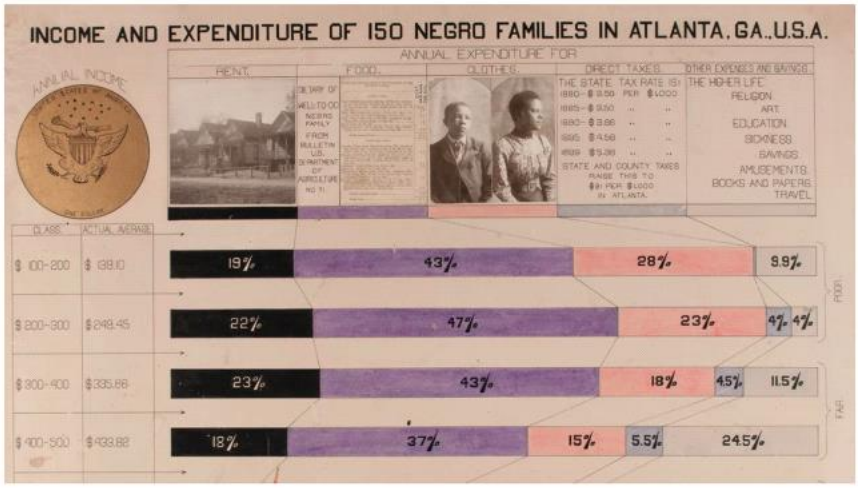
Ritaglio schermata acquisito: 09/03/2021 12:38

1900
We start entering the contepourary 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

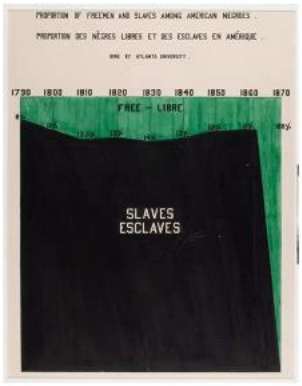
graphs for social sciences (w.e.b. du bois, 1900)

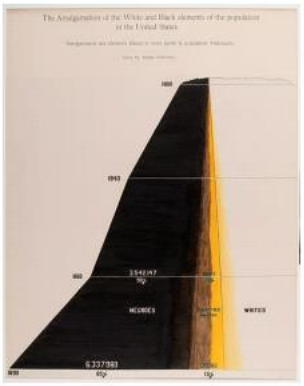
A compendium of new graphs, handmade, for the 1900's world fair

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)

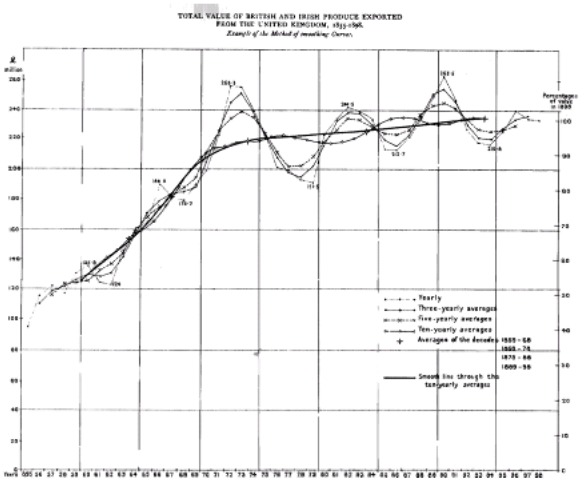




A compendium of new graphs, handmade, for the 1900's world fair

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 thinks 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)

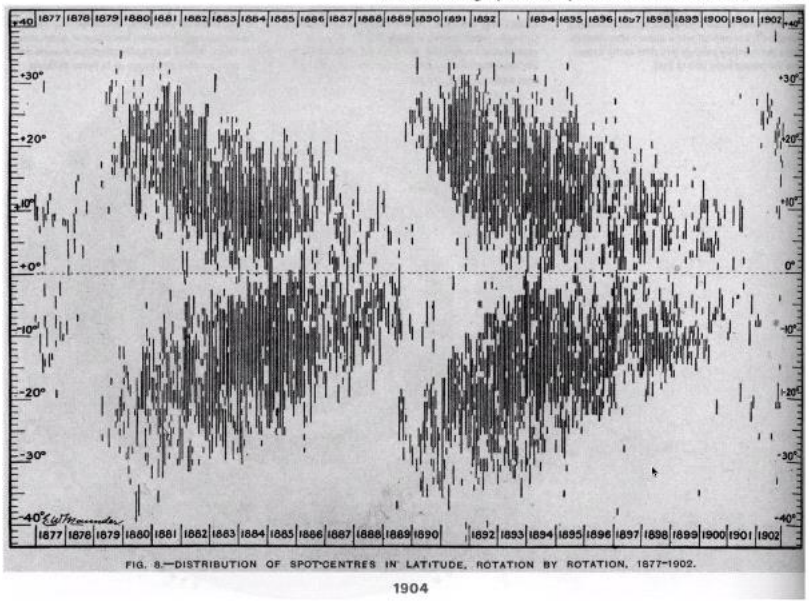


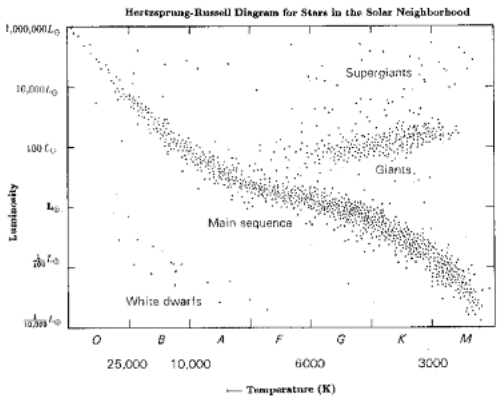
FIG. 8.—DISTRIBUTION OF SPOT-CENTRES IN LATITUDE, ROTATION BY ROTATION, 1877–1902.

1904

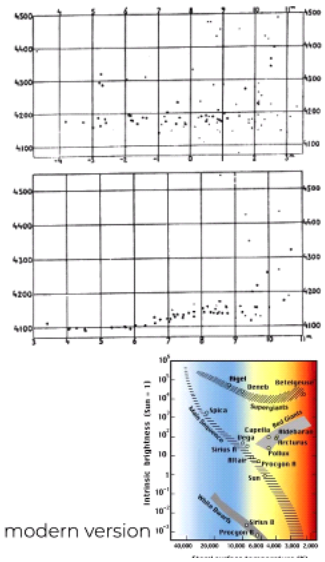
Ritaglio schermata acquisito: 09/03/2021 12:42

An example of unsuccessful graphs.

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



log-log plot of luminosity as a function of temperature for stars, used to explain the changes as a star evolves.



modern version

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

More complex graphs started to appear.

varying pictograms (w.c. brinton, 1914)

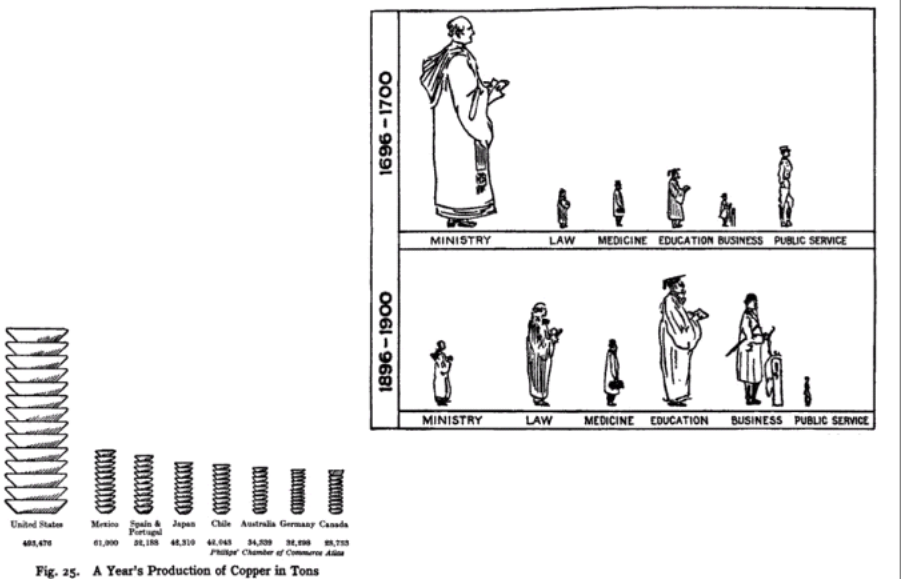
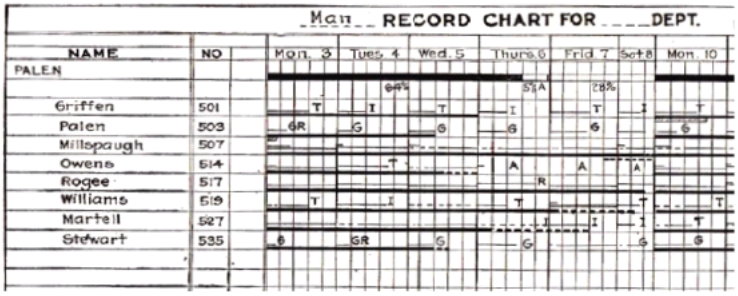
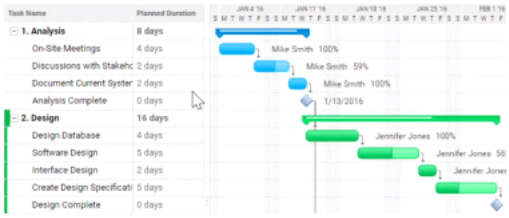


Fig. 25. A Year's Production of Copper in Tons

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

gantt chart (h.l. gantt, 1917)

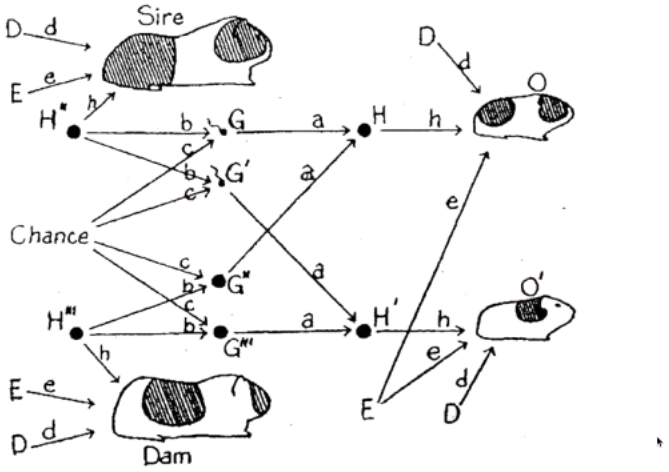




modern gantt chart

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.

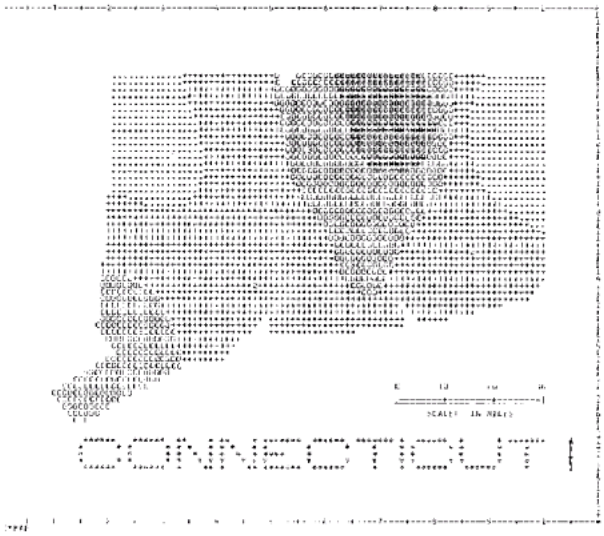
1950 -1975

- 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 emprical it is also crucial for theoretic studies.

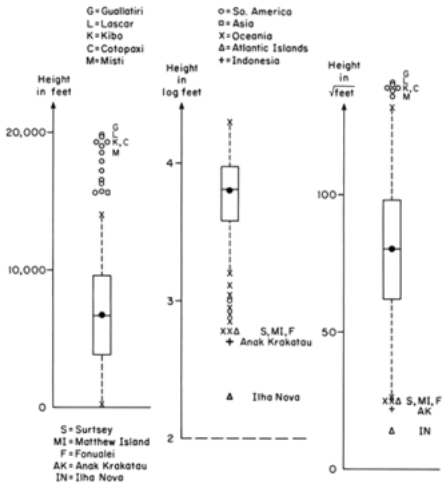
first geographical information systems (gis) (h.t. fisher, 1960)

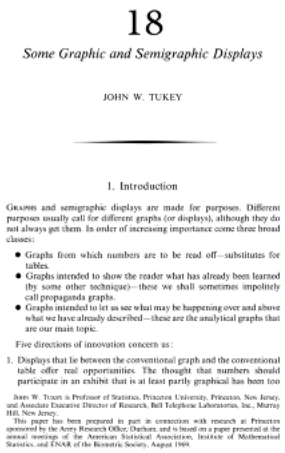


isoline, choropleth and proximal maps on a line printer

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

box-and-whiskers plot (j.w. tukey, 1969)





18

Some Graphic and Semigraphic Displays

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 impolitely 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:

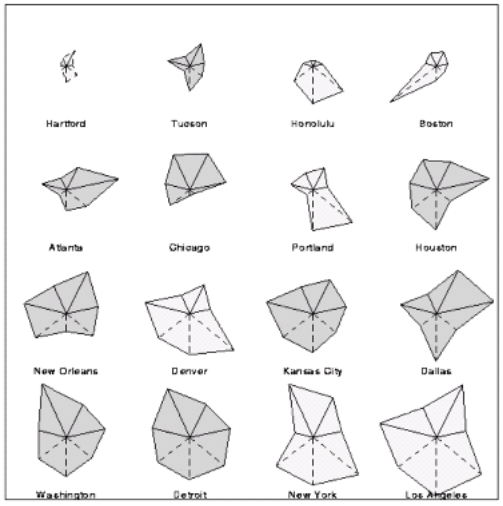
1. Displays that lie between the conventional graph and the conventional table offer real opportunities. The thought that numbers should participate in an exhibit that is at least partly graphical has been too

John W. Tukey is Professor of Statistics, Princeton University, Princeton, New Jersey, and Associate Executive Director of Research, Bell Telephone Laboratories, Inc., 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 is based on a paper presented at the annual meetings of the American Statistical Association, Institute of Mathematical Statistics, and ENAR of the Biometric Society, August 1969.

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)

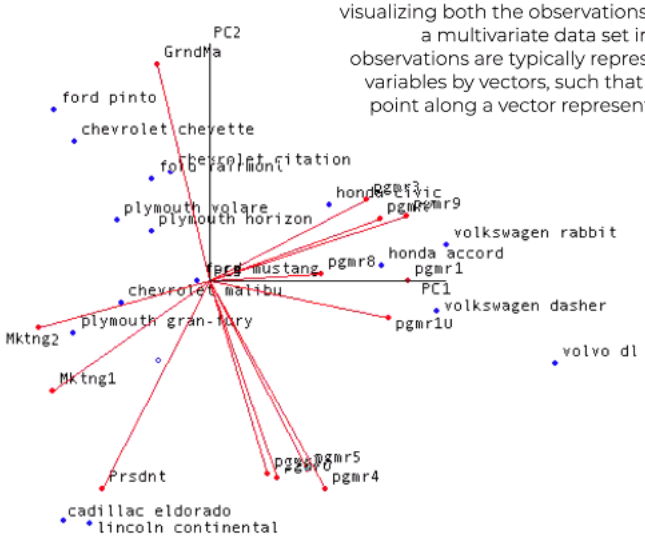


vertices at equally spaced intervals, distance from center proportional to the value of a variable

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

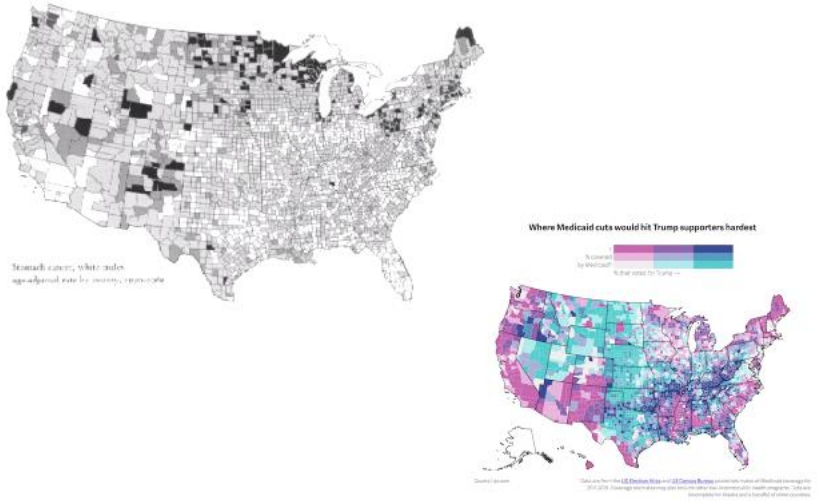
biplot (k.r. gabriel, 1971)

visualizing both the observations and variables in a multivariate data set in a single display
observations are typically represented by points, variables by vectors, such that the position of a point along a vector represents the data value



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

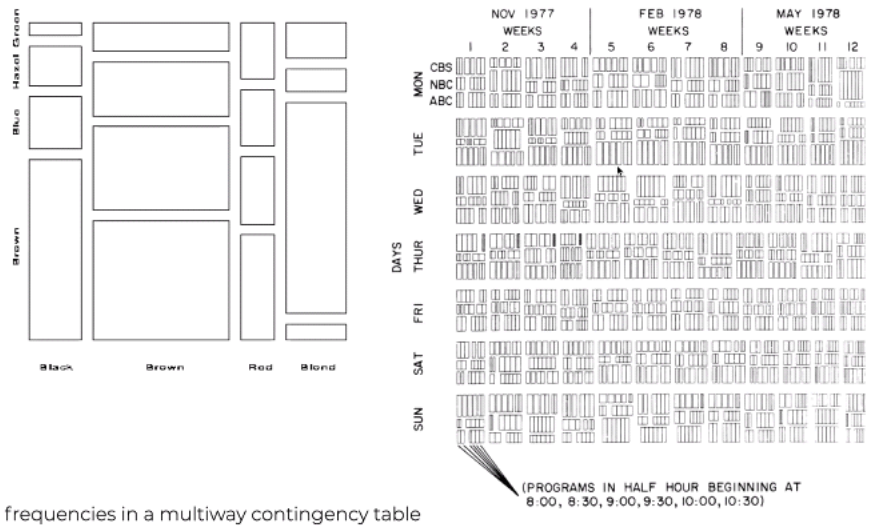
Ritaglio schermata acquisito: 09/03/2021 12:53

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

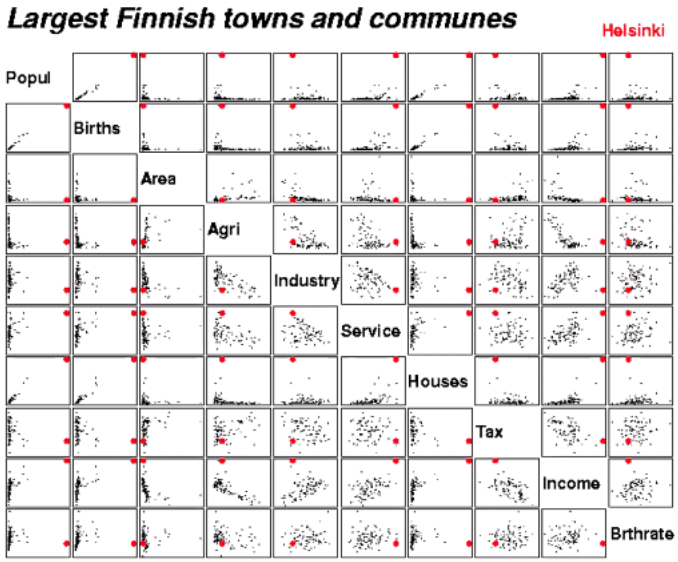
mosaic plot (j. hartigan, b. kleiner 1981)



frequencies in a multiway contingency table

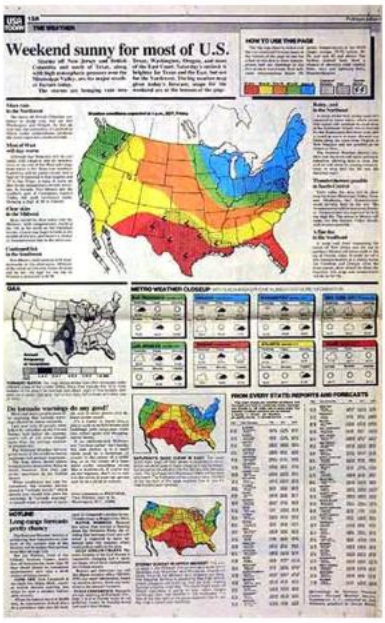
Ritaglio schermata acquisito: 09/03/2021 12:55

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



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

Mutual correlation between variable on the columsn 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 scienifically accurate but full of content and very readable.

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

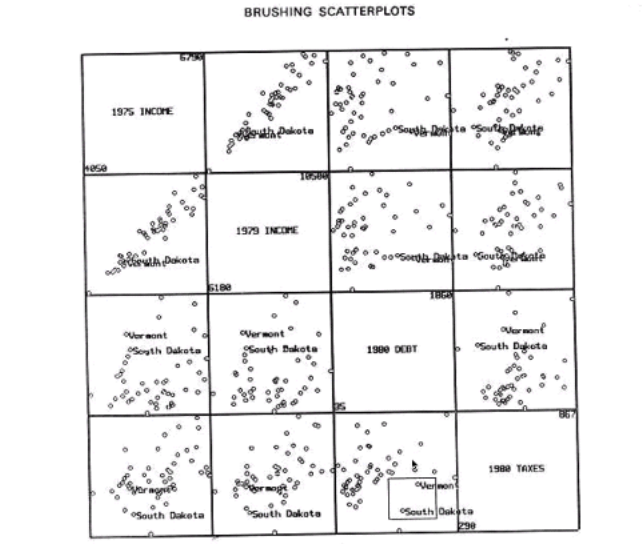
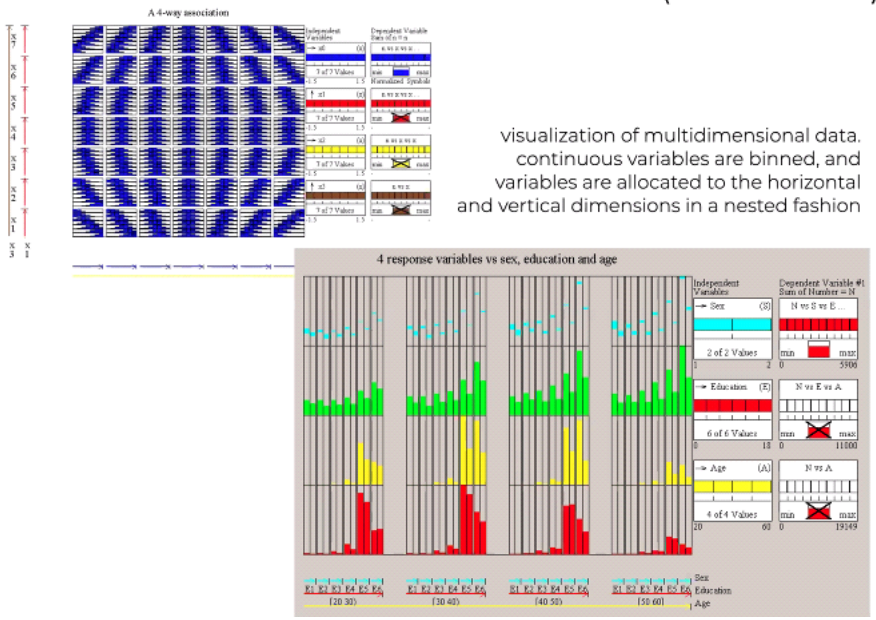


Figure 14. State Data. The label operation is being used. The two points inside the brush on the active panel have their displayed as well as corresponding points on other panels.

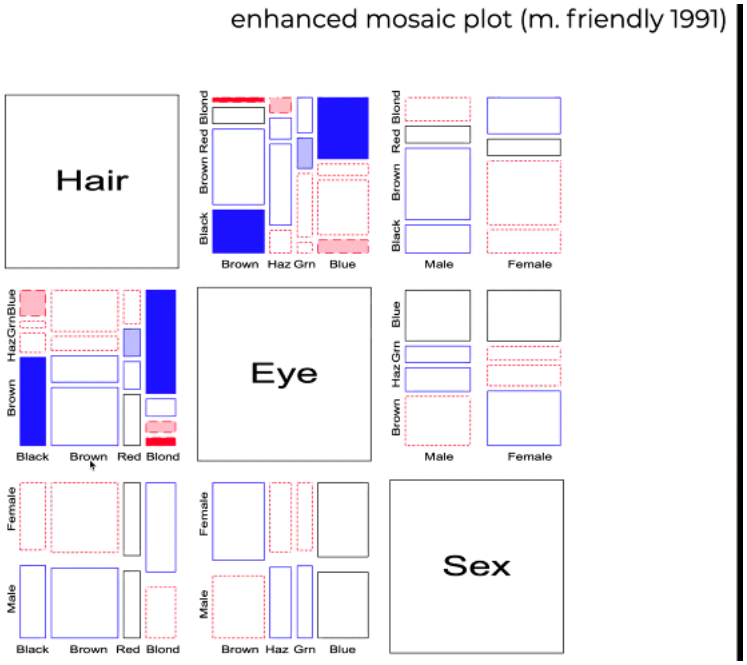
Ritaglio schermata acquisito: 09/03/2021 12:57

nested dimensions (t. mihalisin 1989)

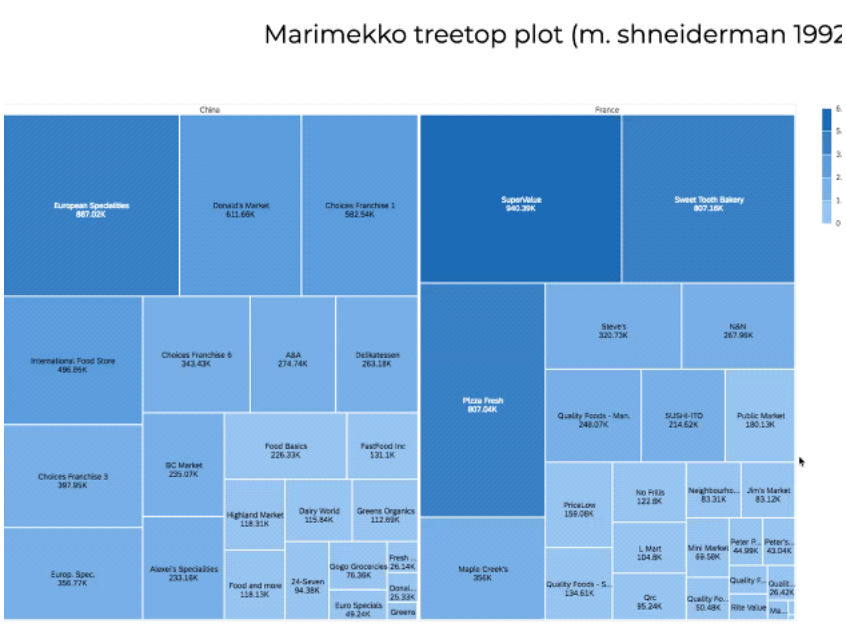


Ritaglio schermata acquisito: 09/03/2021 12:57

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

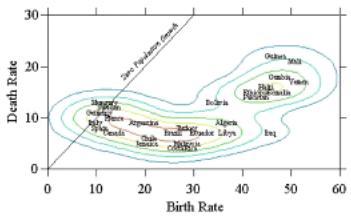


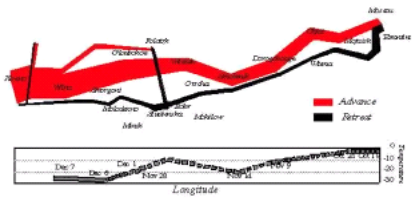
Ritaglio schermata acquisito: 09/03/2021 12:58

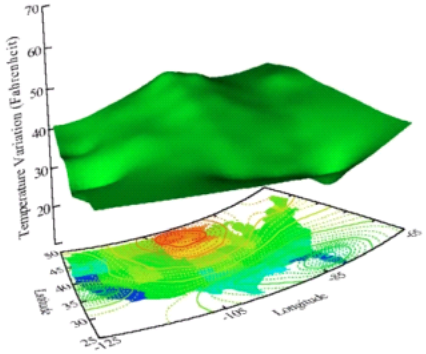


Ritaglio schermata acquisito: 09/03/2021 12:58

grammar of graphics (l. wilkinson 1999)







A comprehensive systematization of grammatical rules for data and graphs and graph algebras within an object-oriented, computational framework

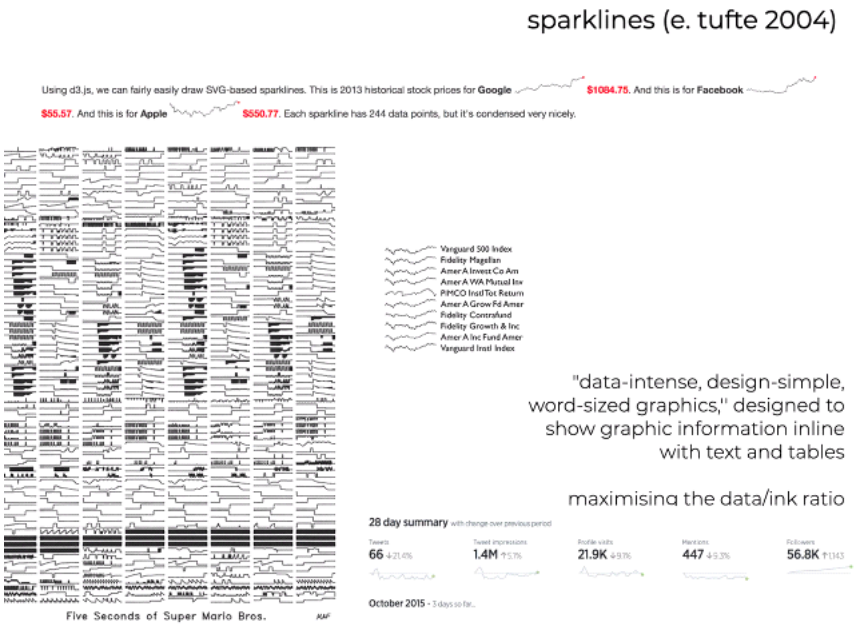
Ritaglio schermata acquisito: 09/03/2021 12:59

tag/word cloud (j. flanagan 2002)

he *search referral zeitgeist*

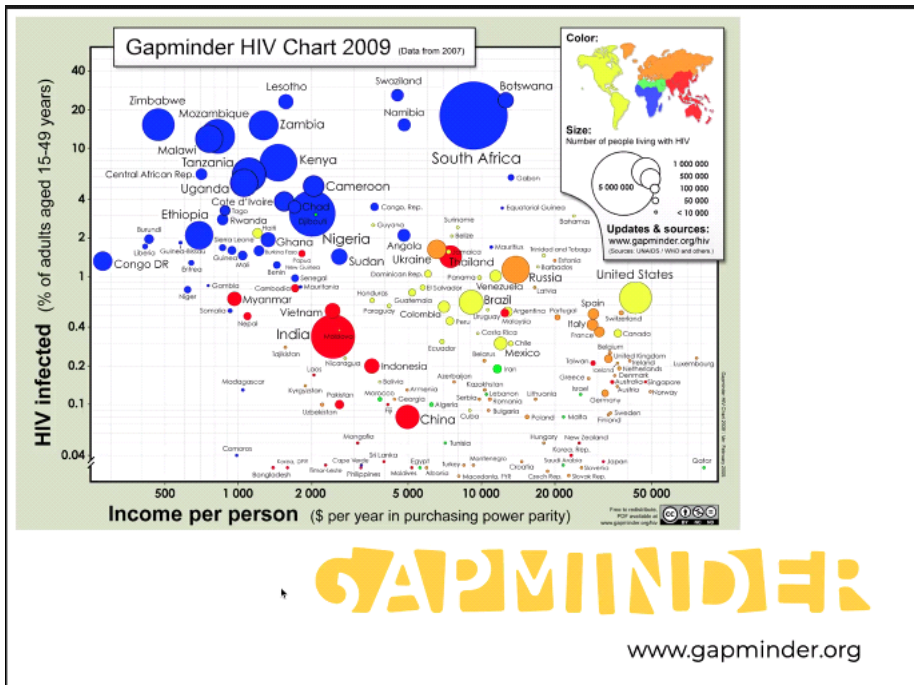
isually display the frequency of
ne most commonly used words
within a document.

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



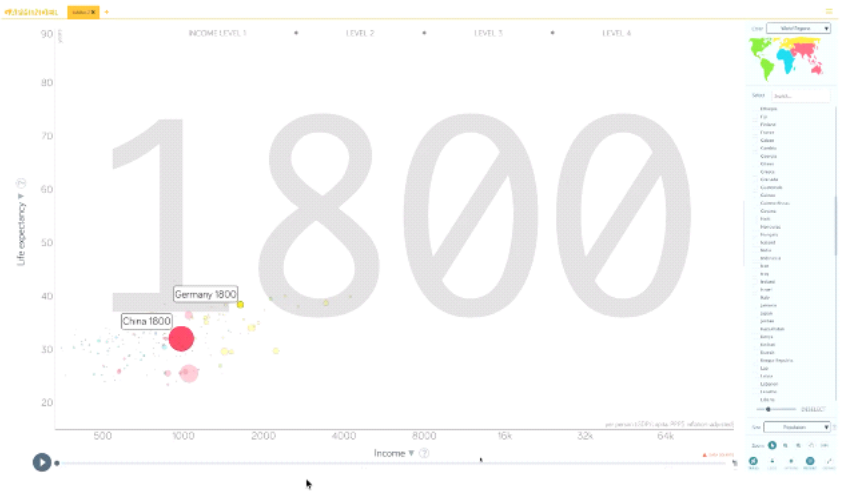
Ritaglio schermata acquisito: 09/03/2021 13:00

Tufte is a key figure for the moder data visualization.
They mazimize the data/ink ratio.
No ink is wasted in thingfs 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.

GAPMINDER

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

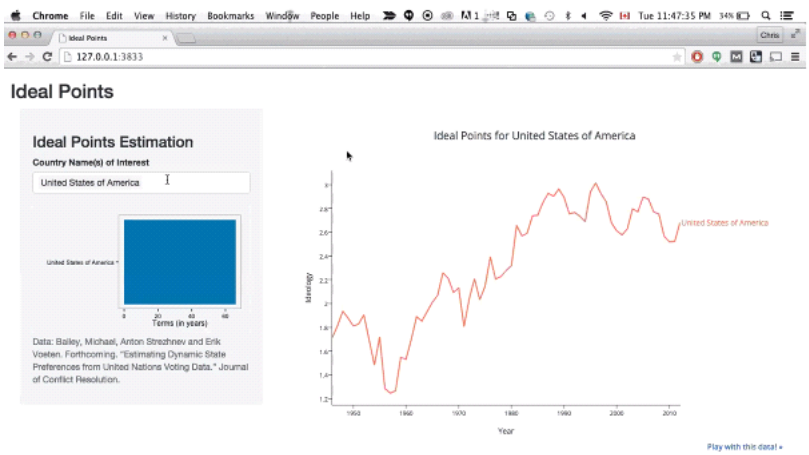
the 5 dimensions of the gapminder graphs are:

- 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

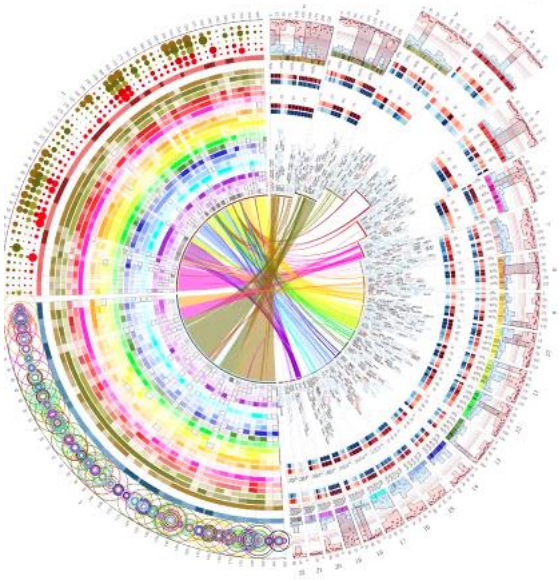
ggplot2 (h. wickham, 2006)



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)

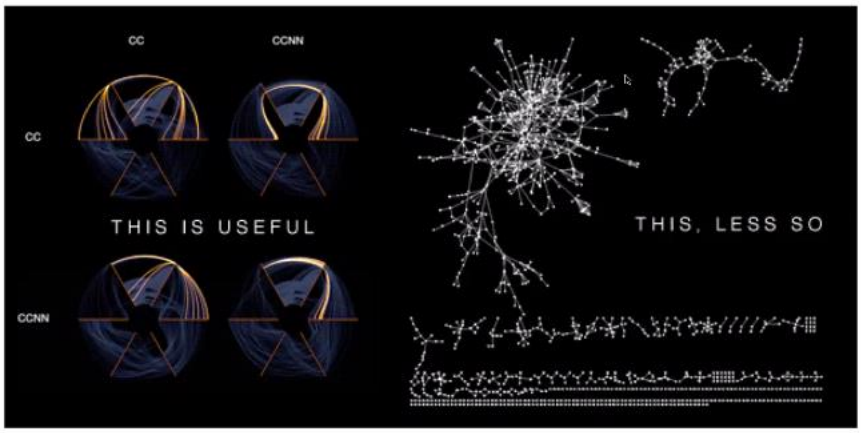


circular diagram designed to facilitate the analysis of relationship among categorical and other variables using chords of a circle with various visual attributes.
the main application is to genomic structure, where the chords can encode various properties of genomic sequences.

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

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

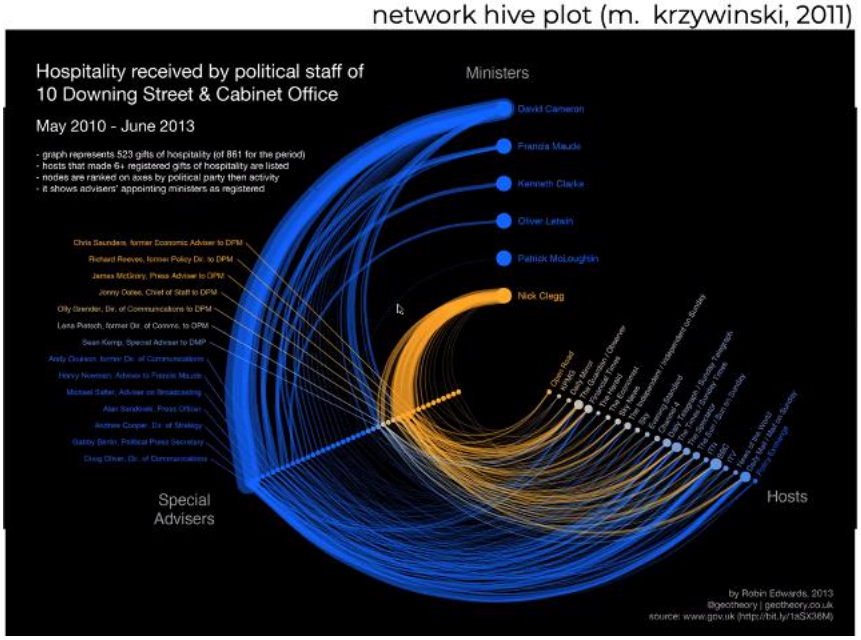
network hive plot (m. krzywinski, 2011)



from the hairball to the hive

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 excpect virtual 3D plots enhancing the interaction with the users.

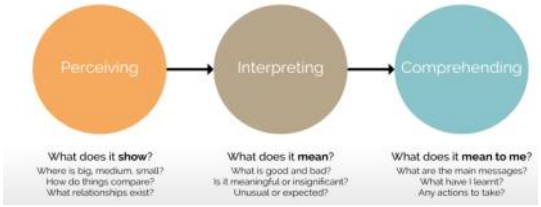
Foundations of data visualization

venerdì 19 marzo 2021 19:09

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

There are 4 pillars:
Representation -> transforming the data in something that can be graphycally exposed and explored
Presentation -> all the surrounding material is used to nicely represent data and information (data optimization)
A good representation enables a better understanding

This is the definition of visualization provided by Kirk



Ritaglio schermata acquisito: 19/03/2021 20:34

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 constrasted effectively.
Data representation -> is the visualization of data through the combination of marks (lines; points ...) and attributes (the tickness of the linesand so on...)
Data presentation -> concerns all other visible desing decision beyond the representation: interactivity, annotations, colours.
Facilitating understanding should be the goal.
Understanding -> according to the latest psycological description, when interpreted, dataviz can be structured in three key moments:
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 visualization? What is good and whatis bad? Graphical representation tells me something useful or something that I completly 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 with the graph that I have seen.
Perceiving it -> the art of simply being able to read a chart efficiently decode the representation of the data (shapes/sizes/colors).

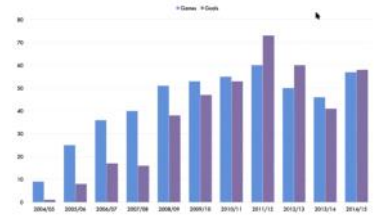
- *What is the relation with the small things and the whole picture?*
- *Where are the largest/middle-sized/smallest values?*
- *What proportion of the total does that values holds?*
- *How do these values compare in ranking terms?*
- *To which other values does this have a connected relationship?*

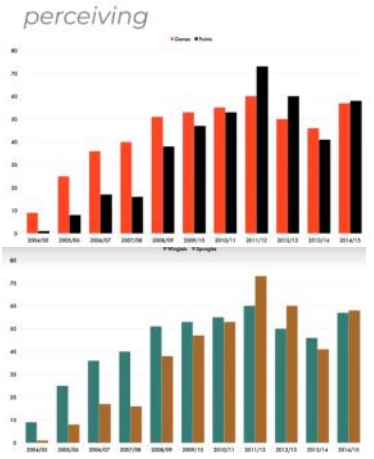
Interpretation-> this requires finding out the real meaning of the plot, good and bad elements.
What is interesting and what is obvious? This are the main milestone in interpreting.
Interpreting is the art of converting the perception into meaning; use the pre-existing knowledge to frame the implication of the viz.

- *Is it good to be big or it is 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 englightened me with new knowledge?*
- *Has this impacted me emotionally or left me indifferent?*
- *Does this understaning force me to take action on the subject?*





This is an example of a barplot(first barplot on the right) which represent lionel messi goals per match.
The perception in this example involves idetifying 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 middle values) we can spota growing trend until 2012 and then there is a small decrease.
Clearly I can also compare the differnt values among the years.
Interpreting-> I put my background. Previous knowledge: scoring 25 is good, scoring 70 is amazing; scoring them in la liga/ucl 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 confirm that messi is you of the best player in the world.
This chart also opens an interesting window: there is a small dicrease in the perfermonace in 2012/13.
What is the reason why?
We know from our expericence that Ronaldinho left barcellona in that year and this could have effected 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 footballso it is difficult to find a proper meaning.
it is not easy to interpet this chart without being an expert of hockey.

Now we consider a third graph. Just changed the tytle -> Winglets and Spungles.
They are two no-sense word. The perception is the same but the other two steps are completly different because I have zero knowledge; no interpretation is possible. This yelds that comprehension is completly senseless.

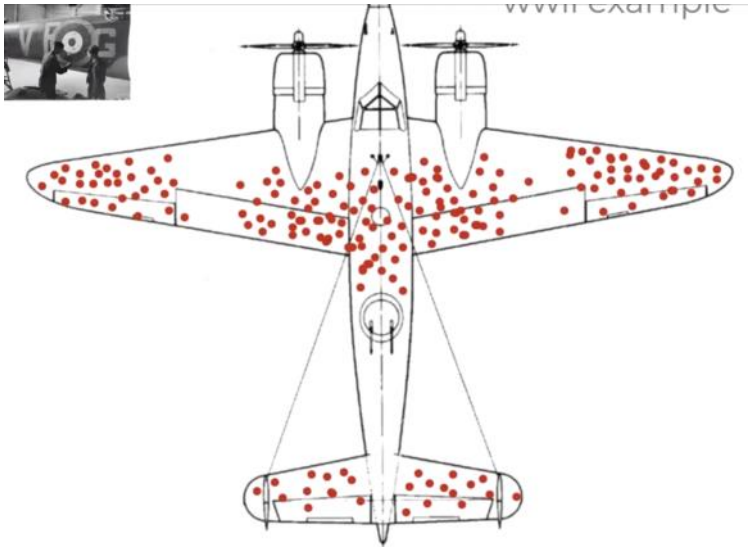
Let's consider now another example realted to the importance of the comprehension of graphs.

abraham wald
1902 1950
ph.d. maths, uni vienna, 1928
member of columbia us statistical research group during wii

how to protect bombers to get shot by german planes?

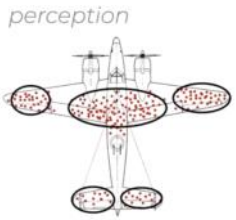
How to optimise position of additional protecting armor?

The idea was to use panel of steel to better protect the aircrafts.
Let's consider this probelm in the field of statistical problems.
So first thing is to collect data regarding the damages caused to the airplanes.
From that they create statstical 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

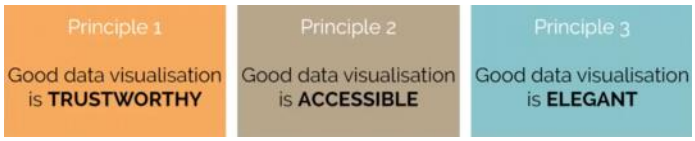
Where should the new armors be put ?



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

First thing -> distribution of the bullets holes. Moving to the interpretation -> statstics tells me that bullet holes should be equally distirbuted.
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 potentail principle of good desing. Transfer concept of good desing into dataviz.
Good design is **innovative**; makes a product **useful**, **aestethetic**, makes a product understandable, is **unobtrusive**, **honest**, **long lasting** , is through 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 exists 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

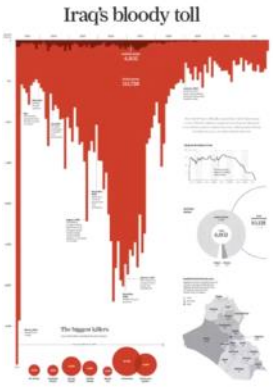
Pure objective is rarely possible. Objective way can be expressed from different point of view and being still true.
Strong need to show that your data/representation is trustable. You may present a true graph and being not trustable or at the contrary being trustable but false.

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

T*hese two graphs are identical but they have a different level of trustworthiness.*
The lefter 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.*

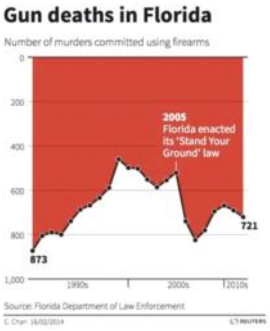
- colors
- fonts
- background
- data description
- data source

The above graph are also influence 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.

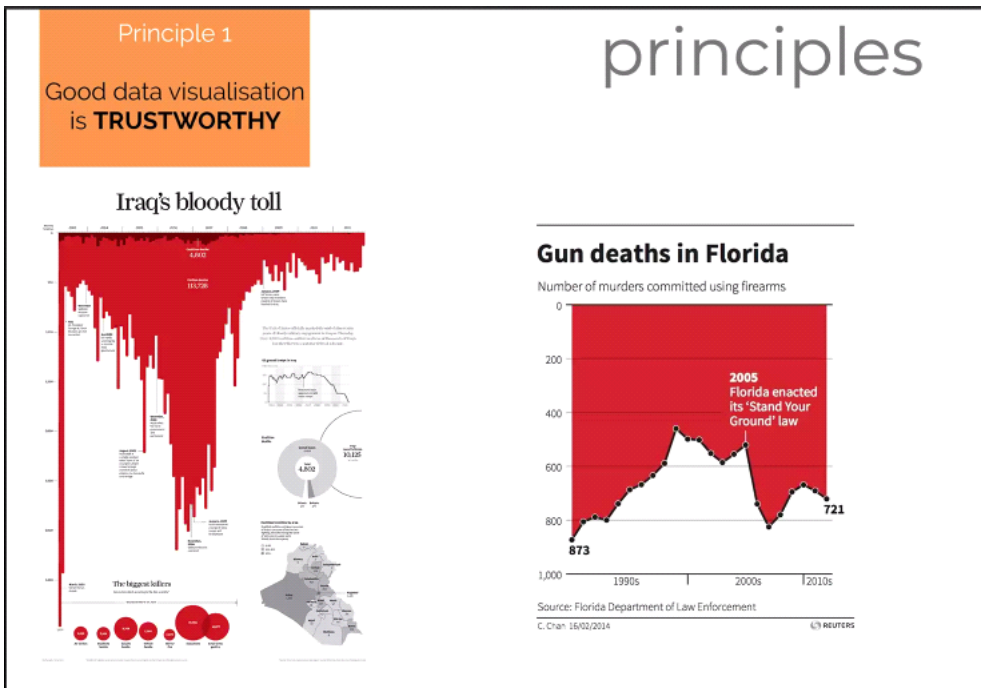


Ritaglio schermata acquisito: 05/04/2021 18:57

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

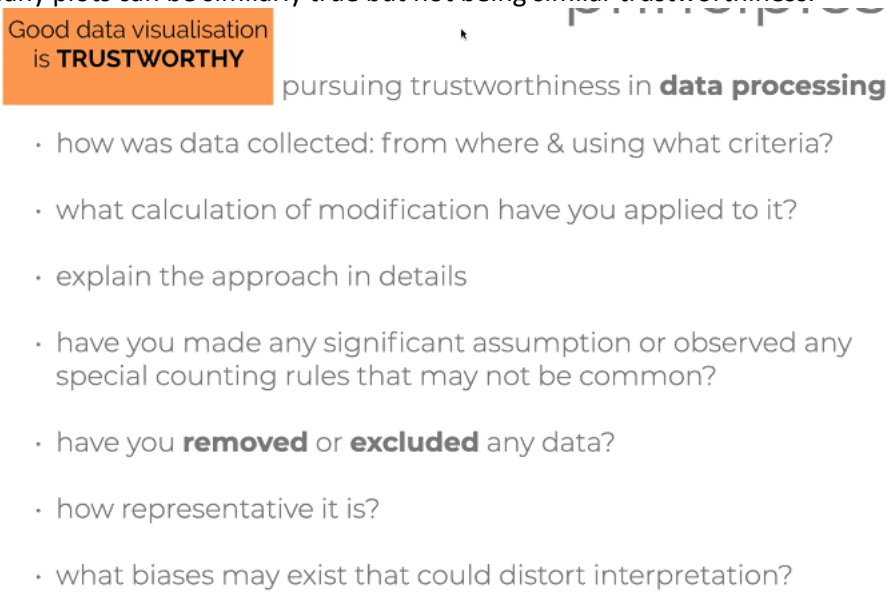
Principles

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.



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

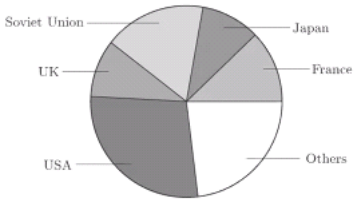
We may have reason to remove some data (**outliers**, **data wrongly collected**, but we have to clearly express this to the reader).

pursuing trustworthiness in **data representation**

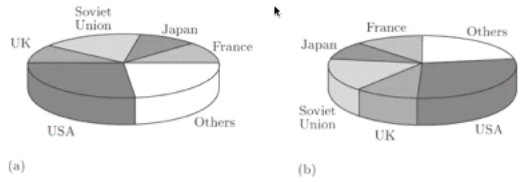
- never **deceive** the receiver
- avoid misunderstandings, inaccuracies, confusions and distortions
  - ⋆ quantitative values represented by areas can be disproportionally perceived

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

Avoid 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

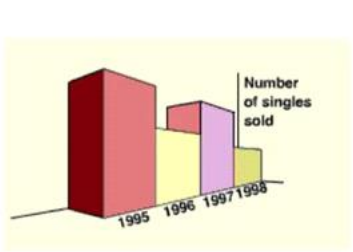
This are two different ways to represent the same pie chart.
The perspective can give you different feelings.
UK seems to be much larger in the a representation compared to the b representation.
With this representation you communicate misleading information due to how human eyes perceive the perspecitve.
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 supsect that the author wants to convince you of something that is not really right.

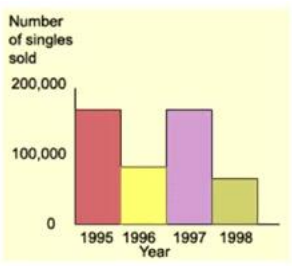
- avoid misunderstandings, inaccuracies, confusions and distortions
  - ⋆ 3d representations are often nothing more than distraction — distortion — decoration; use only if there are 3 dimensions in data and the viewer can change point of view to see different 2d perspectives

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

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

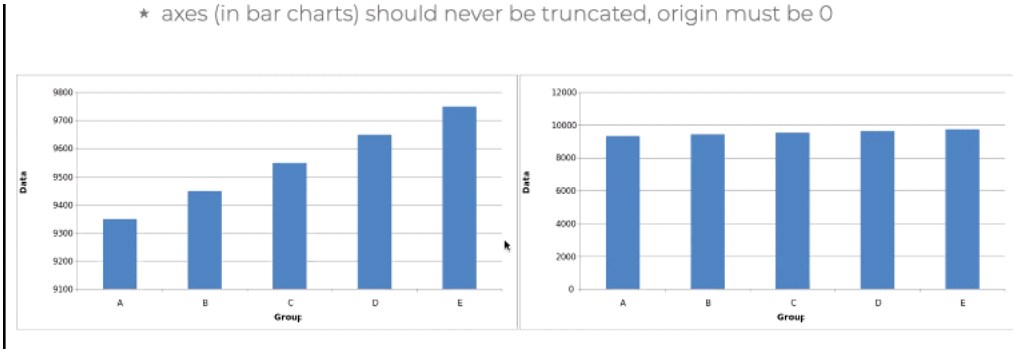
3D change the point of view.





Ritaglio schermata acquisito: 24/03/2021 11:47

Both the graph refers to sales records .
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 !

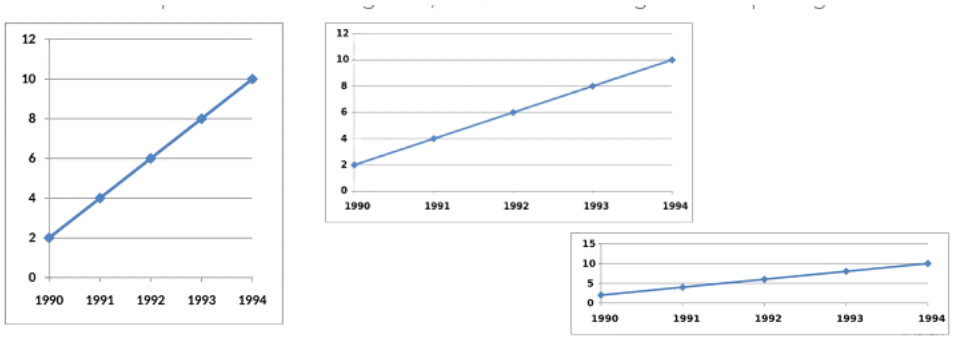


Ritaglio schermata acquisito: 24/03/2021 11:51

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 graps 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 appreaciate a trend.
The left graph is also important because shows different information. According to the different tasks the choice of representation we use is influenced by many factor.
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.

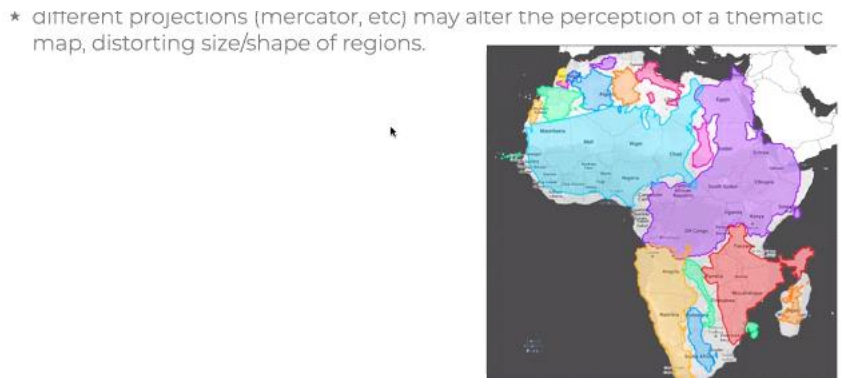
⋆ aspect ratio in line chart is influential, since it modifies the perception of the steepness of connecting lines, either embellishing it or dampening it

Ritaglio schermata acquisito: 24/03/2021 11:59



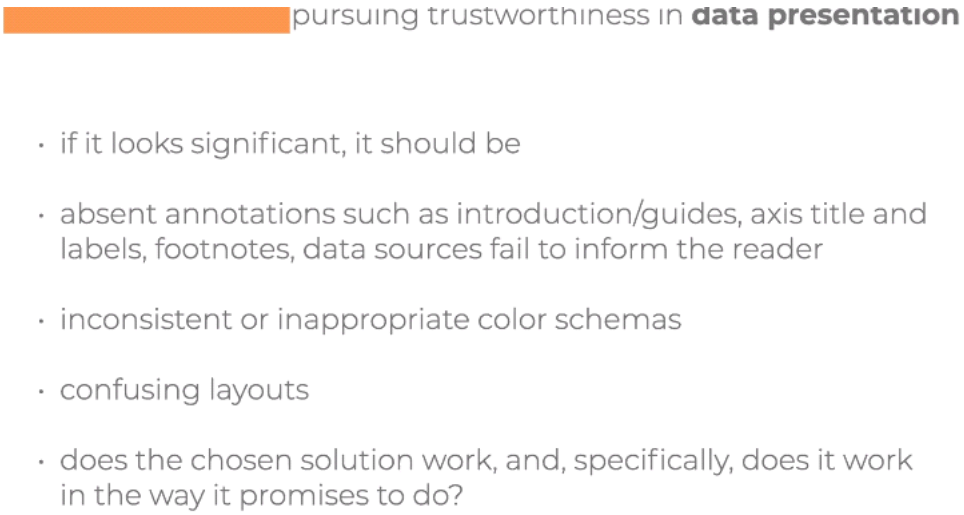
Ritaglio schermata acquisito: 24/03/2021 11:59

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.



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

Mercator projection stretches whatever is near to the equator and distorce what ever is near the poles.
In this case is up to the data visualizer to decide which projection is the more apporpriate.



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 support 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*.

Good data visualisation is **ACCESSIBLE**

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 thing that are not under the control of the author for example the *reader taste*. For example the reader has problem 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 representaiton.

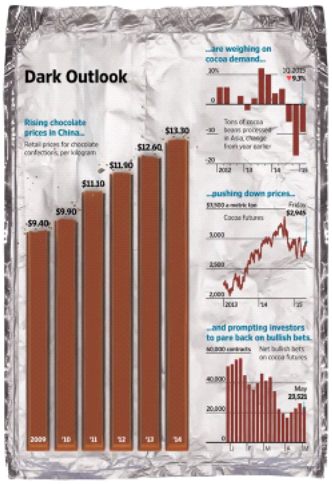
Principle 3

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



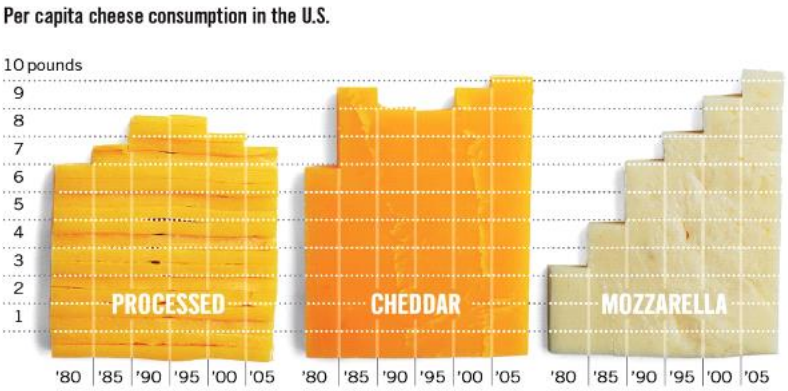
Ritaglio schermata acquisito: 24/03/2021 12:21

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

- Try to *eliminate everything that is arbitrary*. The decoration should stimulate the readder to understand the underlying topic.
- *Not pursuing minimalism at all costs* even if there is a large community fighting to minimazing at all costs but this is not always a good strategy.
- The *representation recalls its masseage*, using chocolate bars for the chart is nice and does not overwhelmed 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 backgroud which recalls the chocolate wrapping -> it is not distrubing and is nice. This is quite a good way to add haestetics. Decorating an infographic without disturbing the audience.

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



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

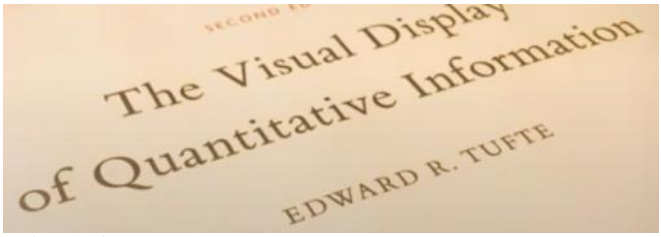
Decoration should be additive, not negative.
For each cheese we have just six points. There is no separion between bars. The first impression is that the plot is more focused on the area rather than the trend of consumption. Your eye is more focused on comapring the three areas rather than the trend.
In many cases the representation is a matter of taste.

"if vienna would be an apartment"

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 desing, plan and implement and deploy a data viz.

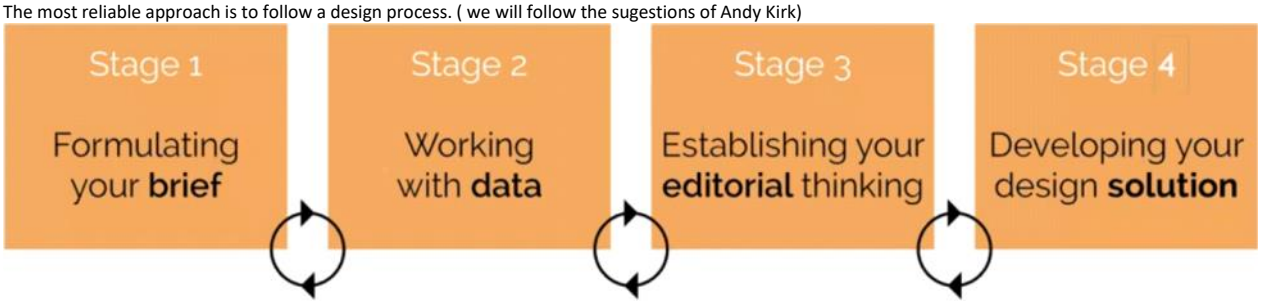
excellence in dataviz consists of complex ideas communicated with clarity, precision and efficiency; graphical displays should

- 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 pourpose of the data viz should be clear.

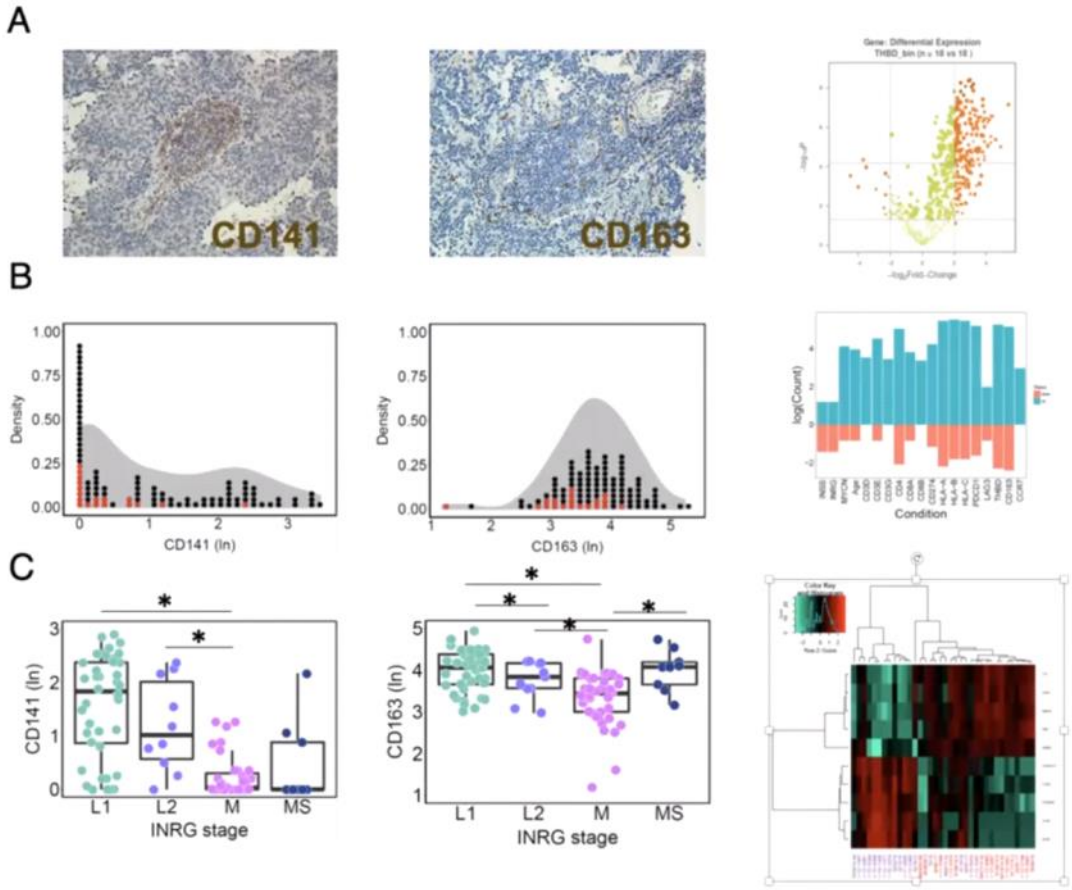
20% of the decision you make influence the 80% of the results.
The most reliable approach is to follow a design process. (we will follow the sugestions of Andy Kirk)



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.
Morevoer, 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 alternaitves.
The desing process needs heauristics 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. Deppen 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.



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 -> 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	vision
defining the origin curiosity	planning a purpose map
identifying circumstances	harnessing ideas
defining the purpose	

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

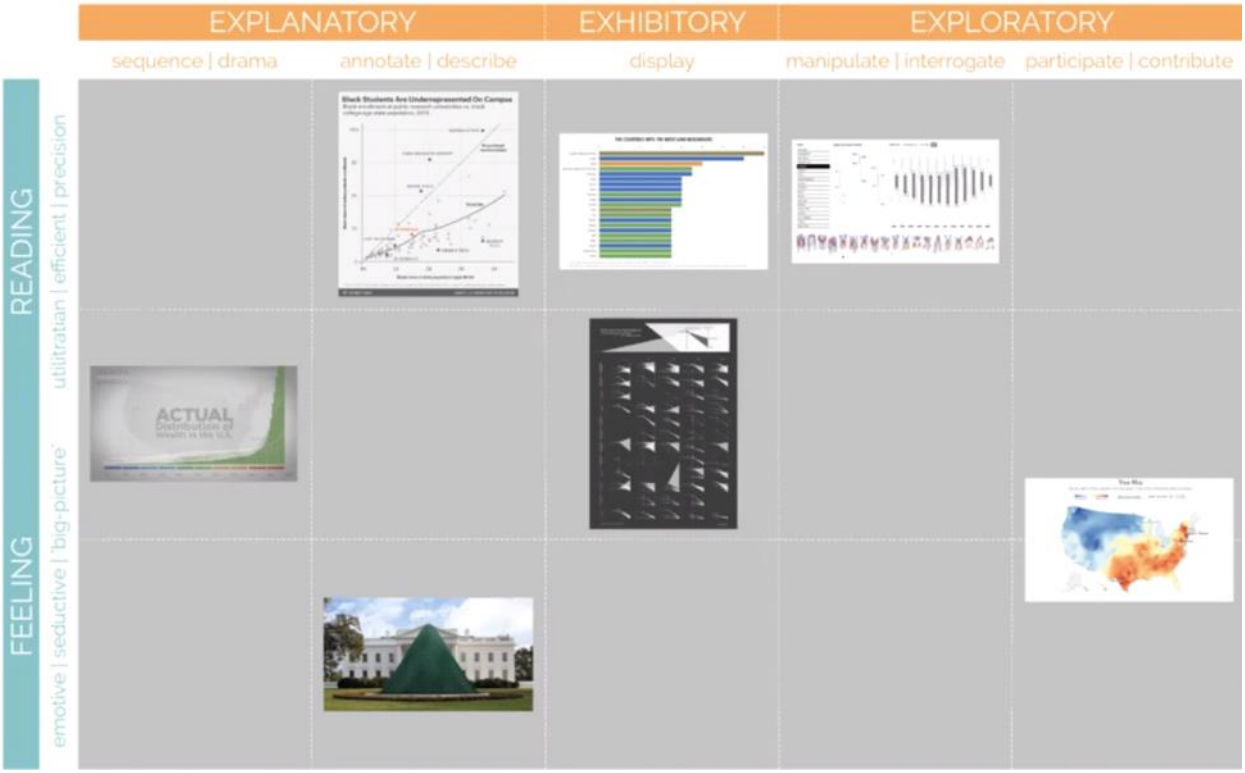
Brief as a set of all the expectation that we have.
Context -> why I want to study this particular topic.
Context -> define your origin curiosity -> it can arise from personal intrigue (specific question you raised); stakeholder intrigue (specific question someone else raised, no anticipation of interest); audience intrigue (combination of knowing what will be needed and anticipating wh at could be needed); anticipate intrigue (audience did not explicitly ask for, but is perceived to be relevant); potential intrigue (opportunity of explorat ion 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?
Constriants -> 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 oftern 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; coffe 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 purpose -> what is it ypu specifically hope to accomplish with your visualization?
Do you want to be impactive? Are you attempting to shock or insipire or persuade? Do you just want 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

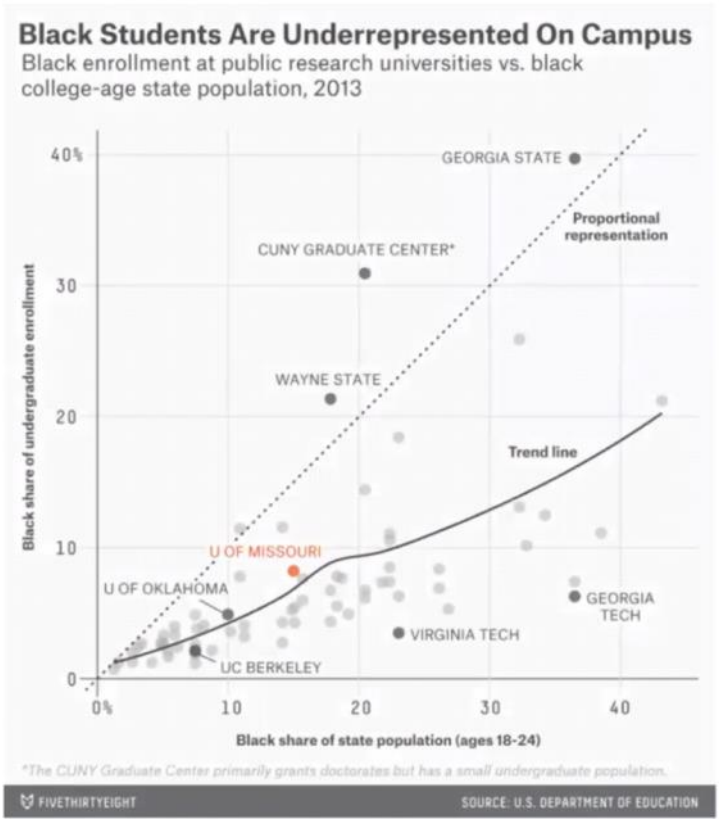
Ritaglio schermata acquisito: 05/04/2021 20:14

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 po ssile, 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 meassages; self-explanatory chart, without the need for in-person explanation.

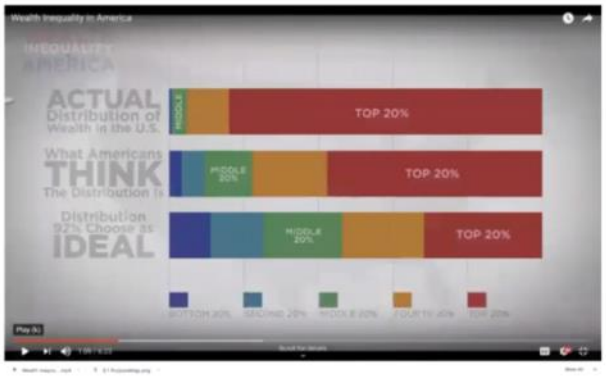


Ritaglio schermata acquisito: 05/04/2021 20:21

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

An alternaitve 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.

Experience Explanatory



wealth inequality in America
[https://www.youtube.com/watch?v=QPKKQnijnsM]

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; over different feautures 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.

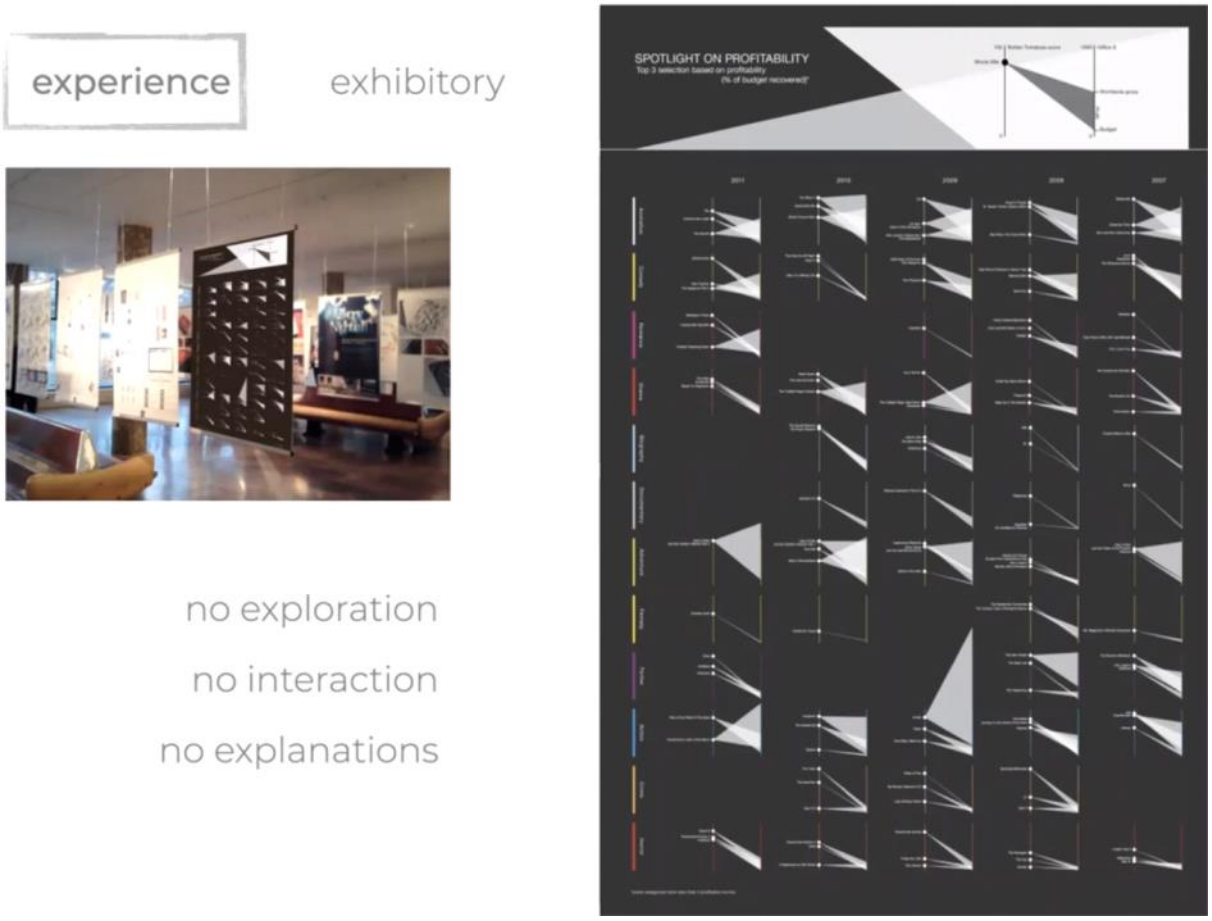
Then we have participate and contribute -> far deeper explanatory experience; greater control and deeper array of features; contributing one's own data; audience drawn to challenges (acitve way)

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

Audience is involved.

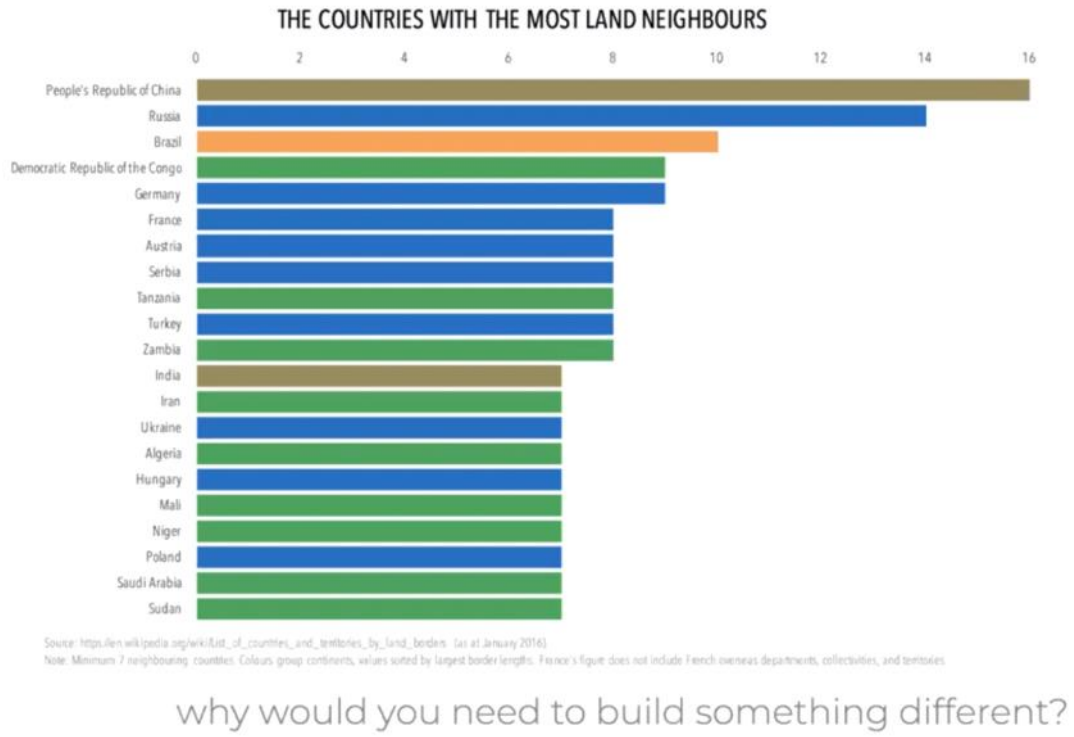
Experience Exhibitory

Exhibitory -> 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*; supporting a written article or report; an inconvenient truth/ gapminder.



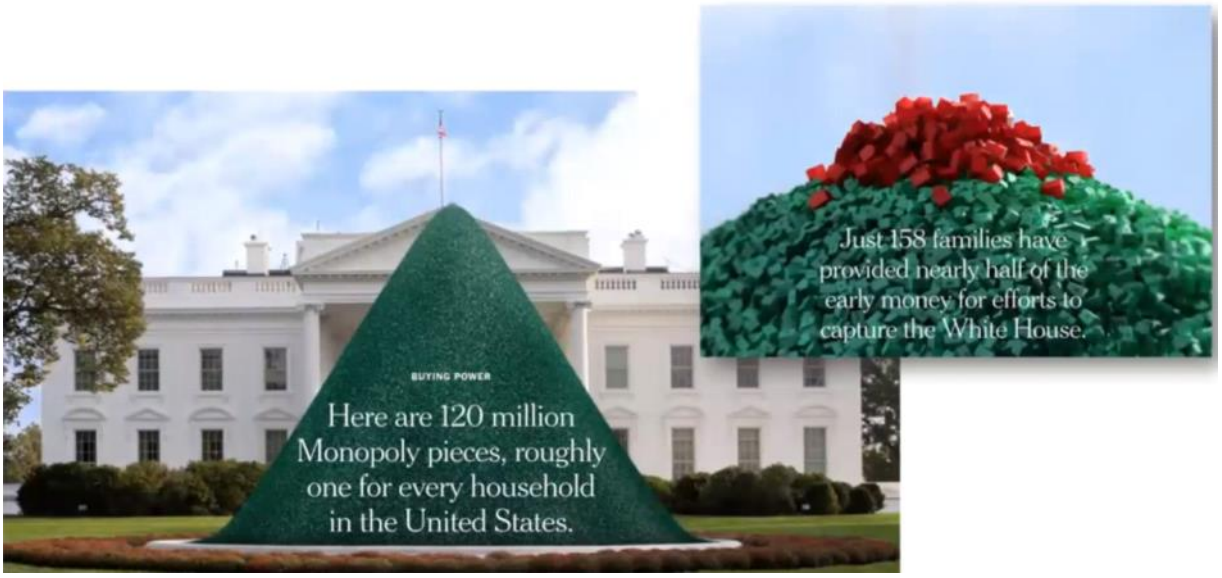
Ritaglio schermata acquisito: 05/04/2021 20:38

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.



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 certaibn degree.



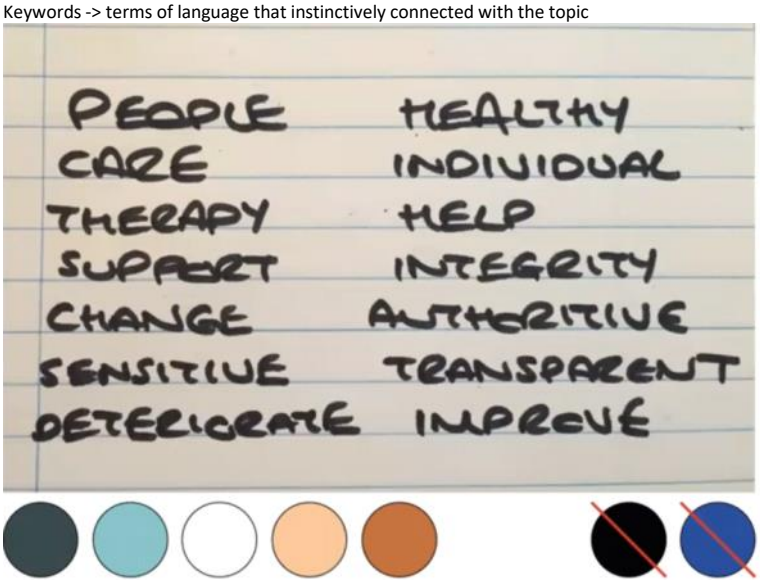
Ritaglio schermata acquisito: 05/04/2021 20:43

Pungulate 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 kahneman0s model of thought

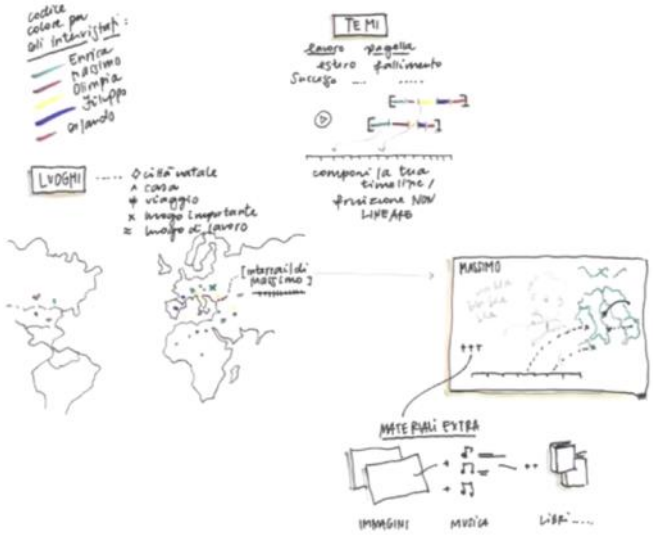
Ritaglio schermata acquisito: 05/04/2021 20:46

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

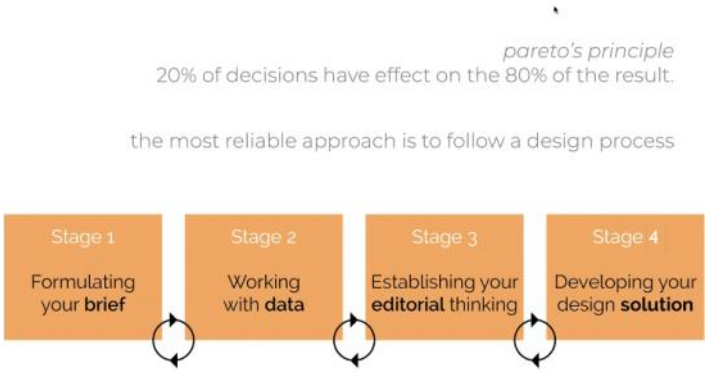
➢ data transformation

➢ data exploration

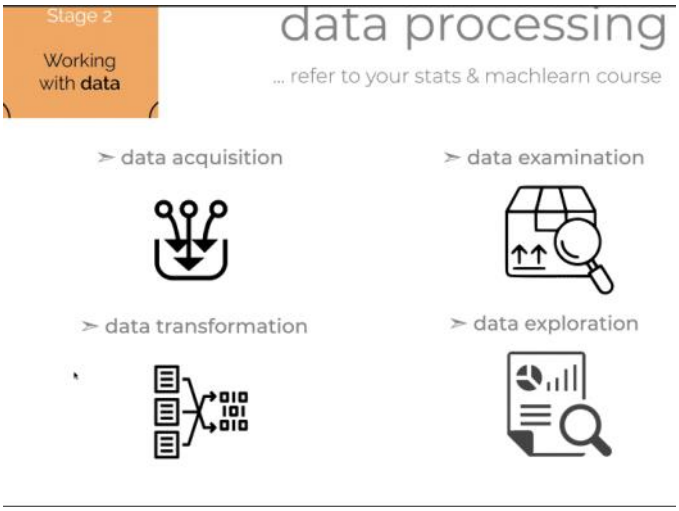
Ritaglio schermata acquisito: 05/04/2021 20:49

Dataviz workflow 2

mercoledì 7 aprile 2021 11:29

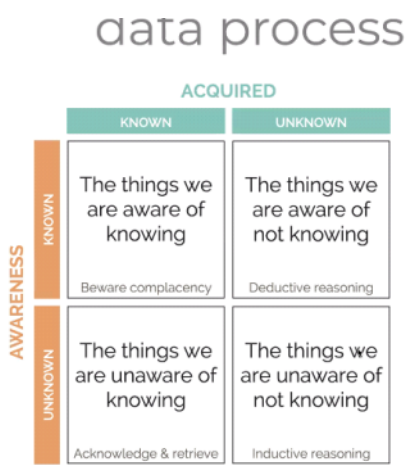


Ritaglio schermata acquisito: 07/04/2021 11:31



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.

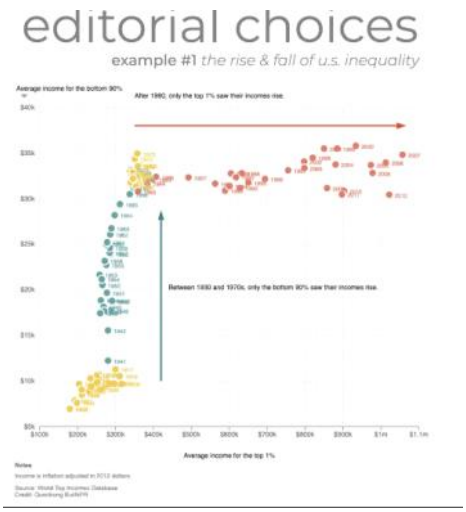


Ritaglio schermata acquisito: 07/04/2021 11:35

EDITORIAL CHOICE
Editorial choice -> many possible 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 subjects (when too complex); what should we ask: are the choosen dimension relevant? Are enough to express what we want to comunicate?; why did we choose this particular pov? Why this pov is worth showing to the audience?
Different audiences can have different perceptions.
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

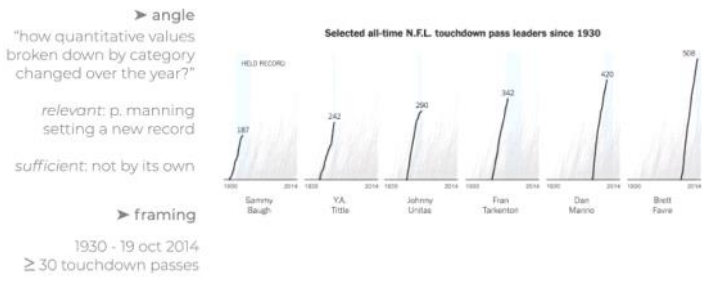
are not a good choice.
The FRAMING -> filtering which data to include or to exclude; remove all the unnecessary clutter; how much content the audience can process- not too small and not too big. It is an hard to balance choice.
The FOCUS -> emphasising what is more important. Provide *visual hierarchy*. Choosing fore/mid/back ground.



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

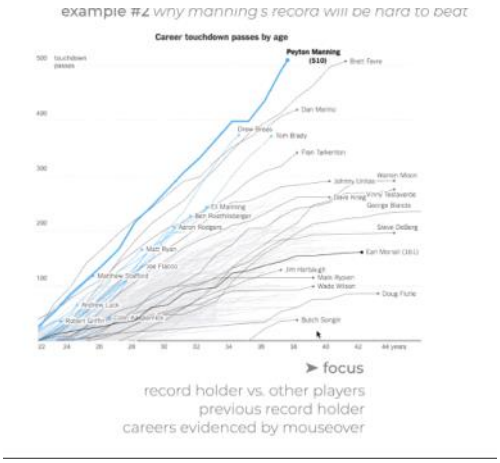
Ritaglio schermata acquisito: 07/04/2021 11:48

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



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 perfromance of single player.

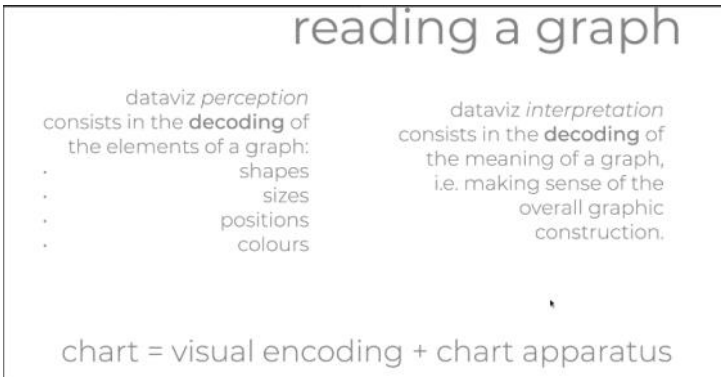


Ritaglio schermata acquisito: 07/04/2021 11:54

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

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

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



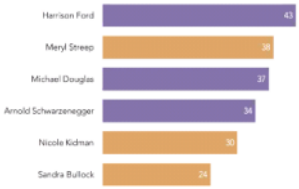
Ritaglio schermata acquisito: 07/04/2021 12:00

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
Meryl Streep	Female	38
Michael Douglas	Male	37
Arnold Schwarzenegger	Male	34
Nicole Kidman	Female	30
Sandra Bullock	Female	24



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

marks

mark	example	description
point		· no spatial variations · quantity as position on a scale
line		· linear spatial dimension · quantity as variation in size
area		· quadratic spatial dimension · quantity as variation in size & position
form		· cubic spatial dimension · quantity as variation in volume

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

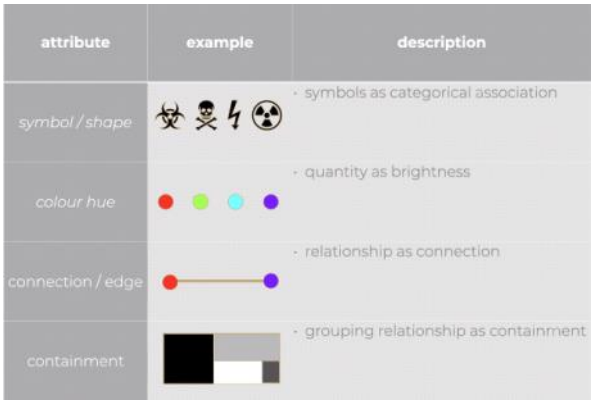
attribute	example	description
position		· quantity as position on a scale
size		· quantity as variation in size
angle / slope		· quantity as variation of angle · quantity as different slope
quantity		· quantity as repeated set of point marks

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		· quantity as saturation
colour lightness		· quantity as brightness
pattern		· quantity as density / shape of pattern
motion		· movement as binary indicator to draw focus or to represent quantitative scale ramp

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



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

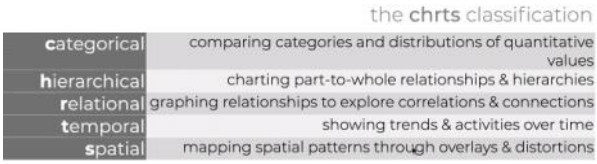
How many choices do we have in this part of the process?
We have different chart types and approaches.

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



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

Stage 1

Formulating your **brief**

the brief

skill & resources + frequency

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?

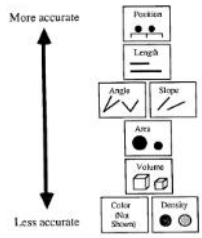
Ritaglio schermata acquisito: 07/04/2021 12:17

The purpose map (different ways of econding data might offer verying 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.



Ritaglio schermata acquisito: 07/04/2021 12:20

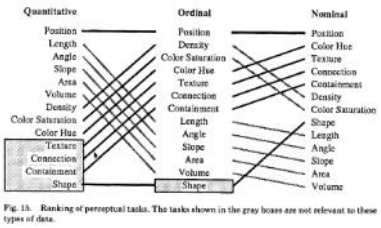


Fig. 13. Ranking of perceptual tasks. The tasks shown in the gray boxes are not relevant to these types of data.

Ritaglio schermata acquisito: 07/04/2021 12:21

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	•		

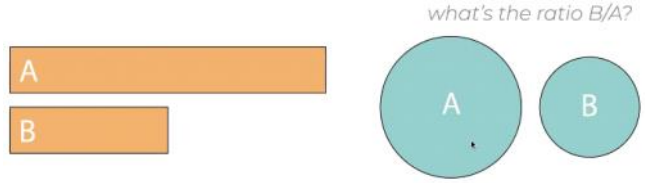
Ritaglio schermata acquisito: 07/04/2021 12:23

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

Ritaglio schermata acquisito: 07/04/2021 12:23

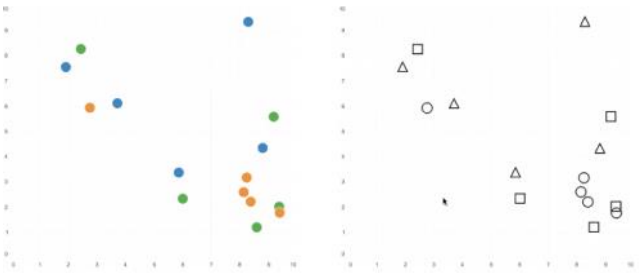
Ritaglio schermata acquisito: 07/04/2021 12:24

Studying measures and shape are important in some instances



Ritaglio schermata acquisito: 07/04/2021 12:25

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



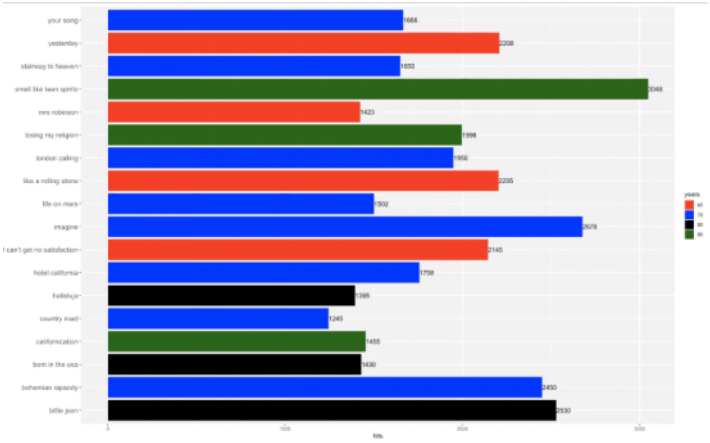
Ritaglio schermata acquisito: 07/04/2021 12:25

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

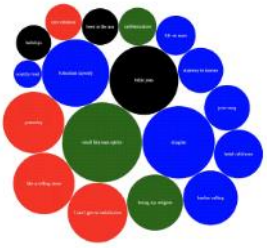
What is the purpose of this representation? Do we need a precise perception? We need to properly separate the data?



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

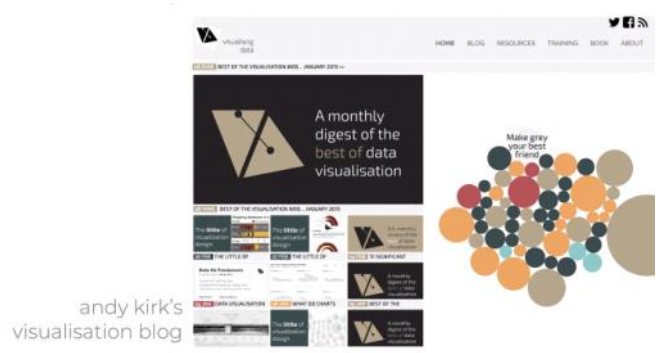
If precise perception is the aim this is the most suitable way of doing it.

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.



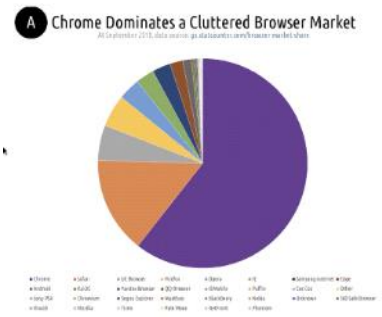
Ritaglio schermata acquisito: 07/04/2021 12:31

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

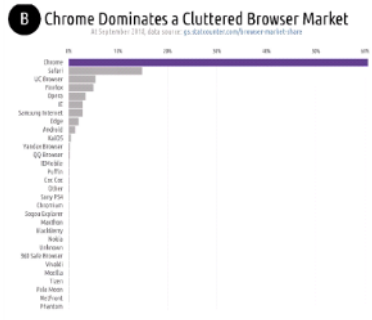
Ritaglio schermata acquisito: 07/04/2021 12:32

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



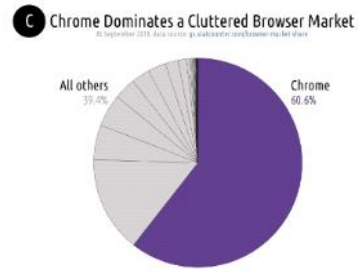
Ritaglio schermata acquisito: 07/04/2021 12:35

This pie chart gives a clear idea of the content.



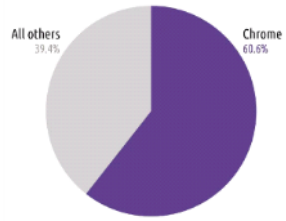
Ritaglio schermata acquisito: 07/04/2021 12:36

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.

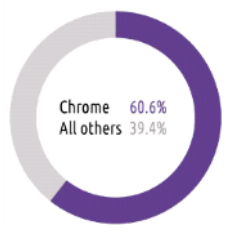


Ritaglio schermata acquisito: 07/04/2021 12:37

We can choose this other type of representation. It is eaiser 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



Ritaglio schermata acquisito: 07/04/2021 12:39

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

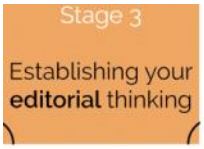
ANGLE

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

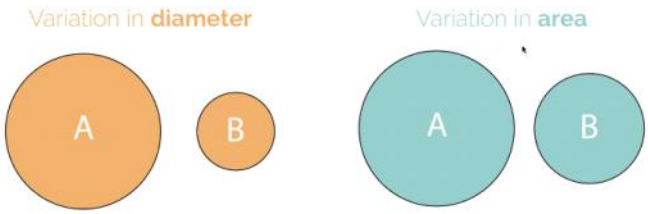
Ritaglio schermata acquisito: 07/04/2021 12:40



trustworthy design

avoiding deception

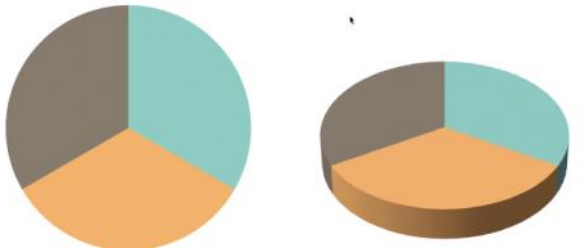
geometric distortions



Ritaglio schermata acquisito: 07/04/2021 12:42

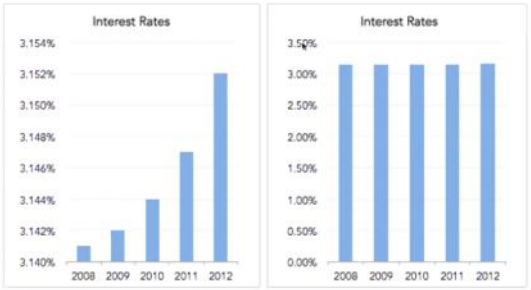
The feeling you are providing to the audience is quite different.

3d decorative distortions



Ritaglio schermata acquisito: 07/04/2021 12:43

truncated axes



Ritaglio schermata acquisito: 07/04/2021 12:44

ACCESSIBLE DESING -> encoded overlays

incorporation of additional details:

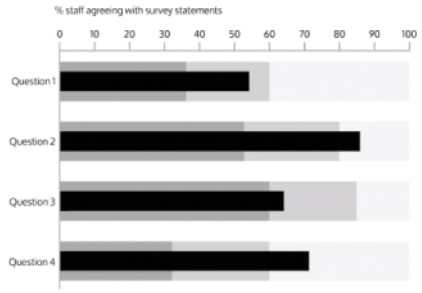
to explain further the context of values

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

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

Ritaglio schermata acquisito: 07/04/2021 12:44

Bandings

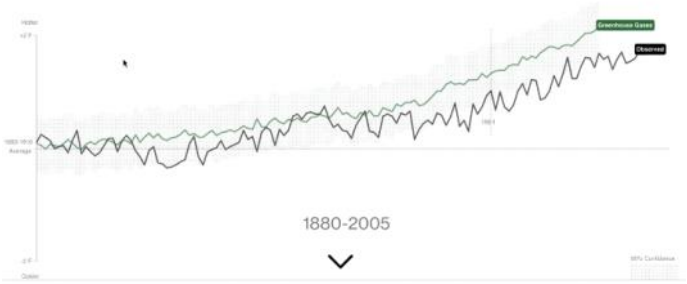


Ritaglio schermata acquisito: 07/04/2021 12:45

Barplot embedded with other barplots.

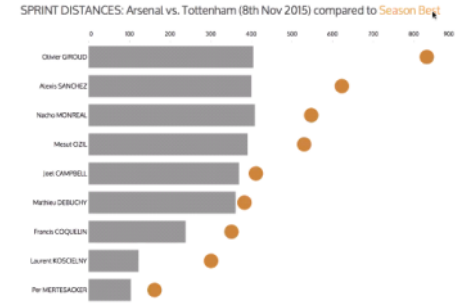
indicate the contrast between the main data value marks and the contextual judgement of historic or expected values

Ritaglio schermata acquisito: 07/04/2021 12:46



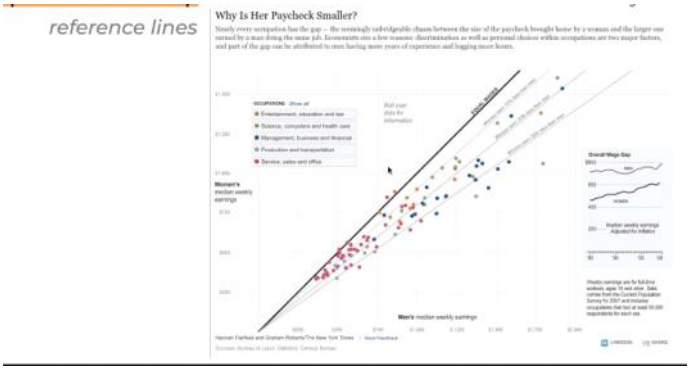
Ritaglio schermata acquisito: 07/04/2021 12:47

Contextual overlay showing 95% confidence interval.



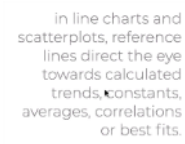
Ritaglio schermata acquisito: 07/04/2021 12:49

Another example of encoded overlays
Additional points comparision versus a miximum values.
Different markers are a very good driver for the reader.



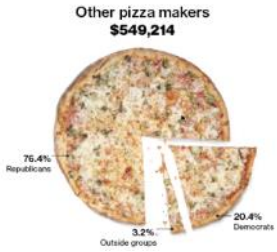
Ritaglio schermata acquisito: 07/04/2021 12:50

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



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

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

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 -> d*efining the shape, size, layout choices for all the components of the chart.*

PROJECT COMPOSITION -> how to layout and size all the visual components?
Although established conventions do exist, this is usually an iterative process towards what feels like an optimal layout.
Hierachy of content can be reached thorugh caregul chouces in relative position and relative variation in size (and variation in color for significance).

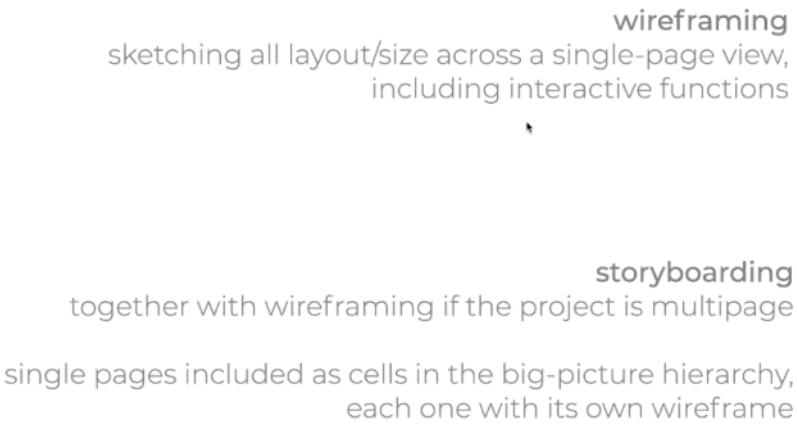
Let's consider this project.
This represent an historical building in hong kong. The panel wants to show each step of the life of the building.

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 sourranded by small componets that help the reader to be orienteted.
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.



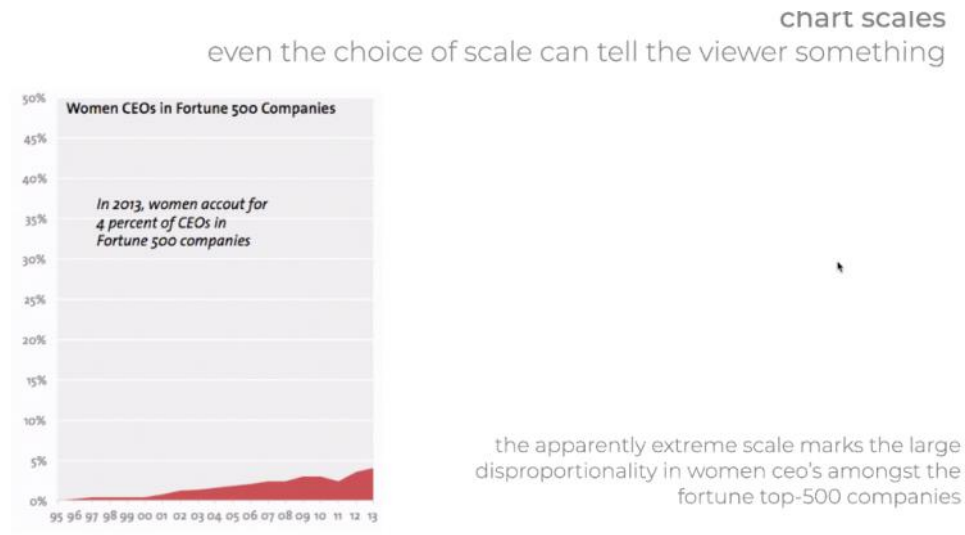
Ritaglio schermata acquisito: 14/04/2021 11:44

The other step in project composition is Storyboarding.

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

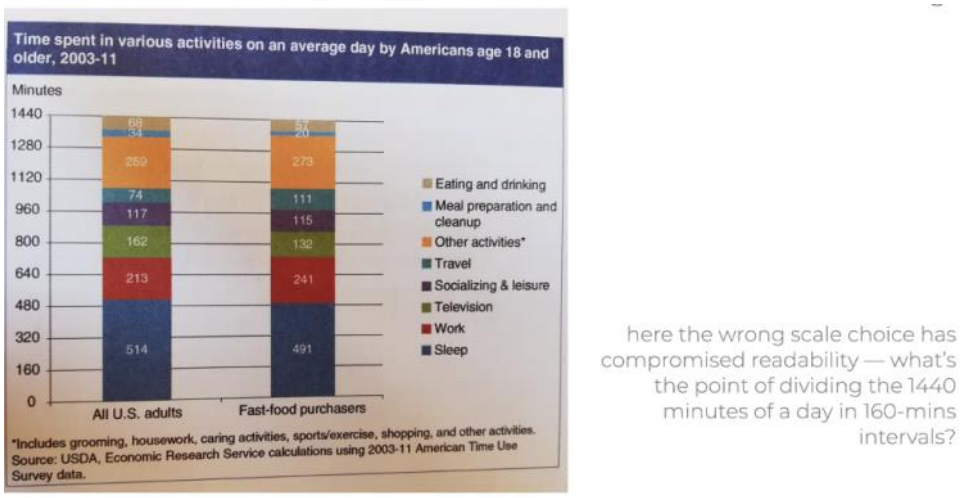
The chart is the central element of data viz. Understand how to desing the layout is a central moment. Understanding the proper size is crucial. Especially if you need to organize multiple information you should not be scared to shrink. Watchout: small writings can be dangerous.

Another component is the scale.



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

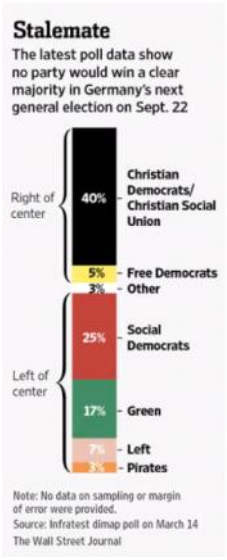
The graphical effect is that the graph is very EMPTY. In ideal world the number of women as CEO should be around 50%. The author choose a vary impacting way of representing it.



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

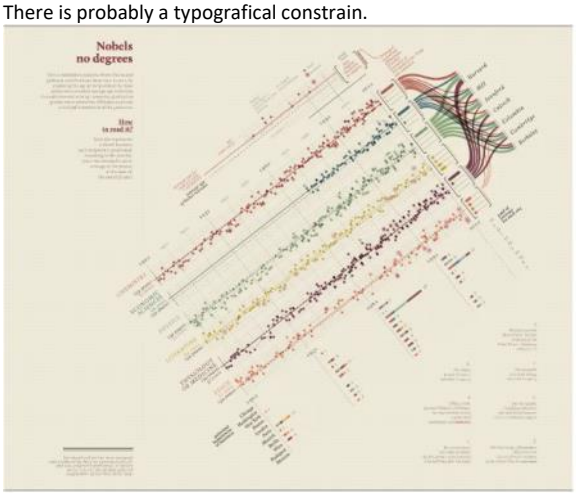
The intervals are really innatural. Subdivision should be done in 60 minutes. It is not a natural way to represent it in that way. Here there is little attention paid to the details.

Chart orientation -> adding an extra degree of readability; avoid label overlap; remeber 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 typografical 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 infromation insiede a single chart.

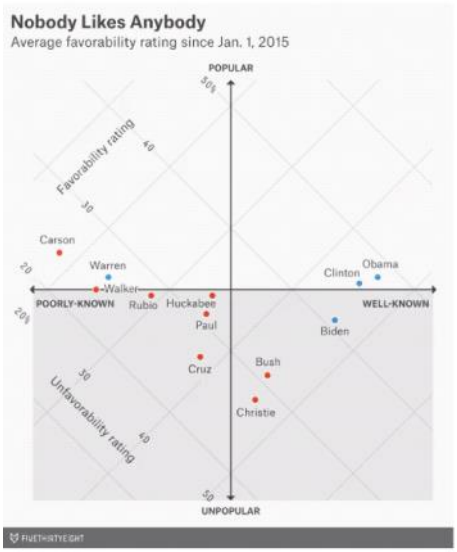


chart orientation
not only 90 degrees

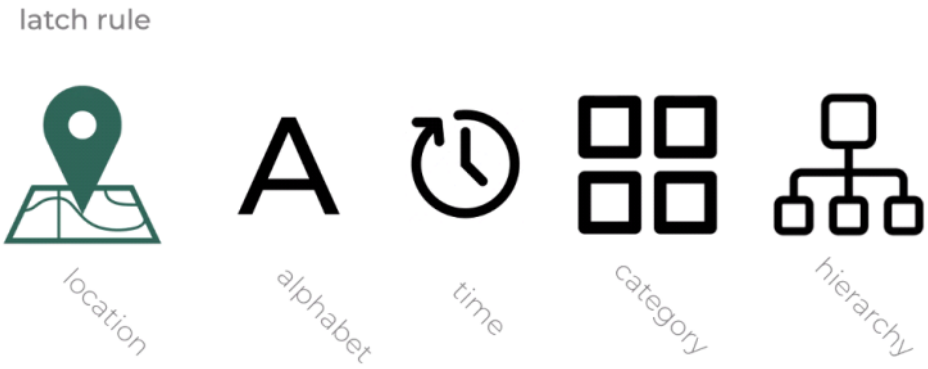
45 degrees rotated scatterplot with 2x2 overplayed grid to make it easier to observe which values are located in each quadrant.

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.

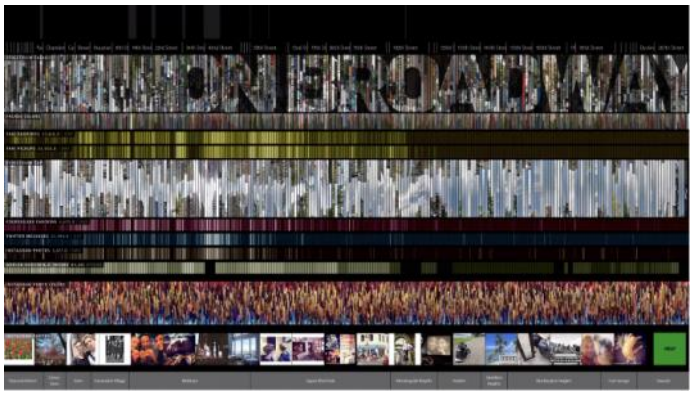
Another important compenent is how you *SORT your value*.
You need to specify an order.
According to which element you sort the values.



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.

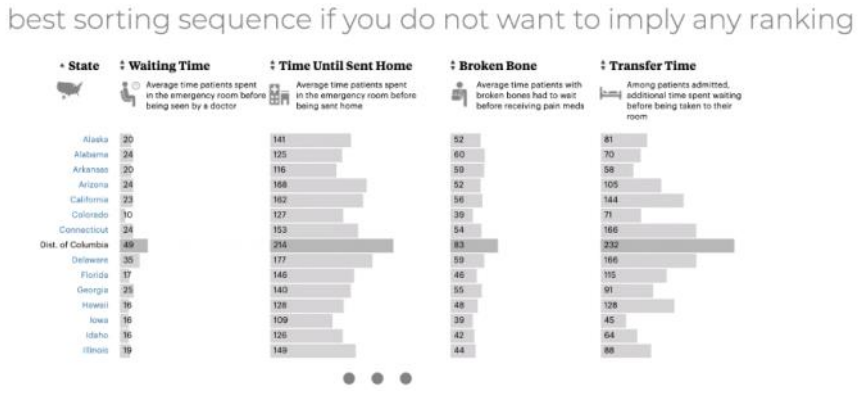
Ritaglio schermata acquisito: 14/04/2021 12:30



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

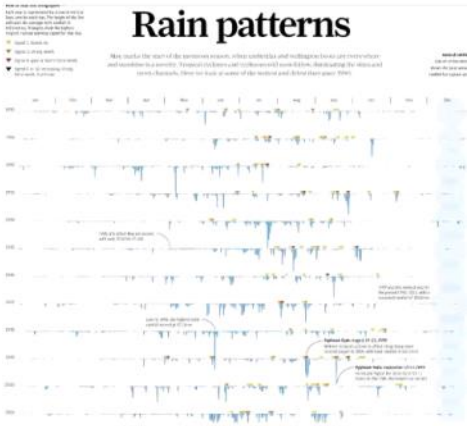
Lot of elements belonging to the street itself.

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



Ritaglio schermata acquisito: 14/04/2021 12:32

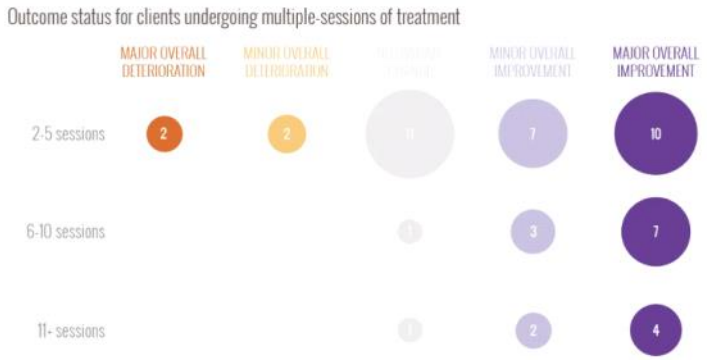
Time -> meaningufl if helps comparing changes over time.



Ritaglio schermata acquisito: 14/04/2021 12:35

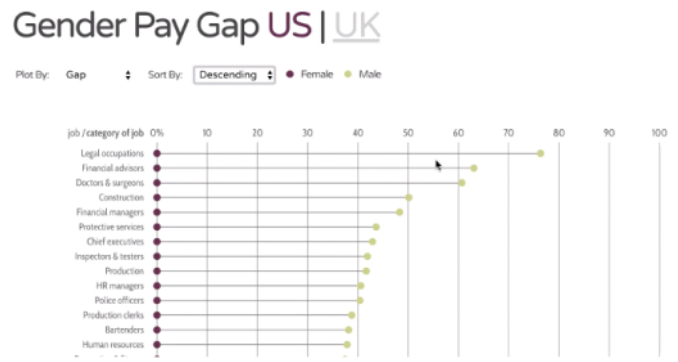
In the example, the interest is in the seasonality of patterns -> chronological sorting is meaningful.

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



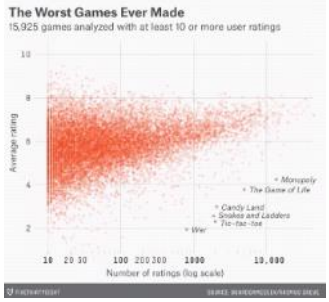
Ritaglio schermata acquisito: 14/04/2021 12:38

Hierachical sorting is defined by increasing/decresasing quantitites efficient perception of size/distribution/raking.



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 distirbution is evident although the different scales.

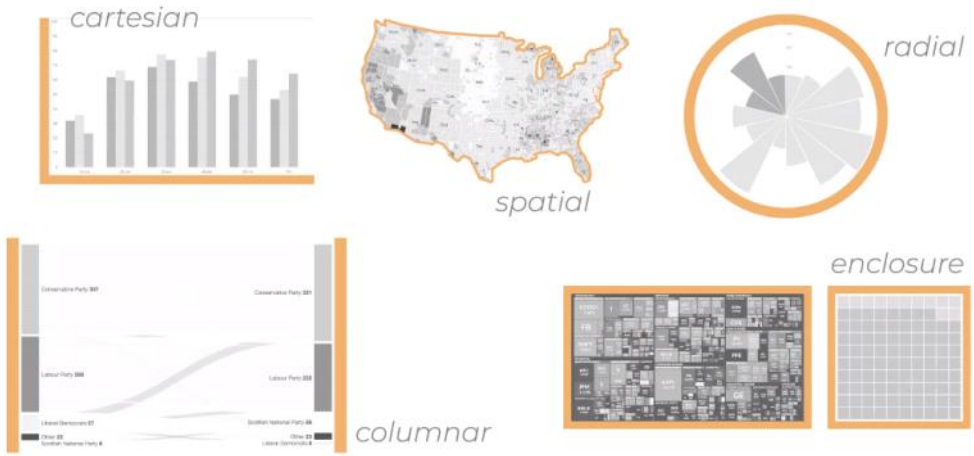
Month on month inflation % 2000–2008

	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	60	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	

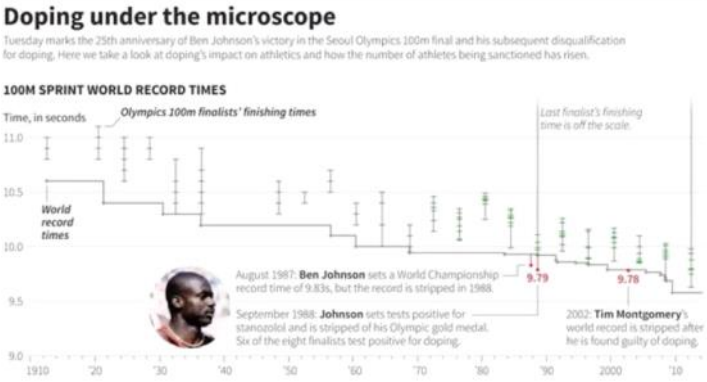
Ritaglio schermata acquisito: 14/04/2021 12:46

Tabular data.
Although they are number -> 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 the double of the space compared to the other. Also the values are out of the range of the previous one.

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

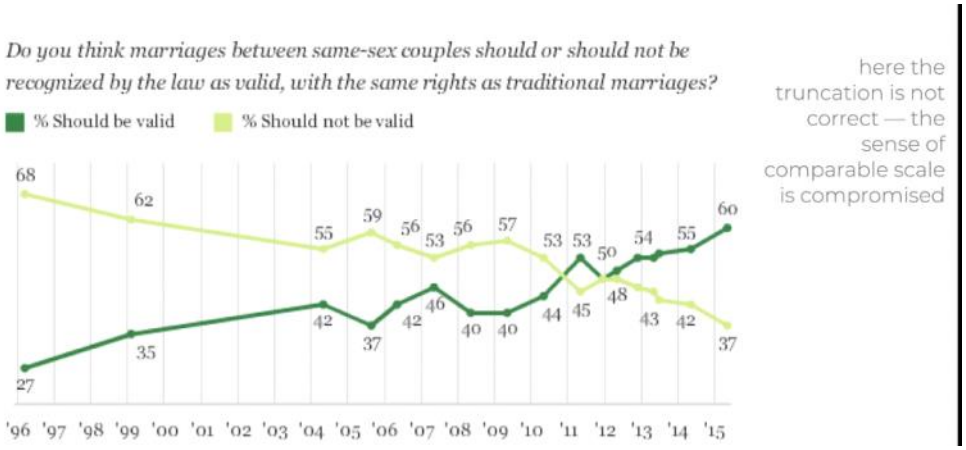


Ritaglio schermata acquisito: 14/04/2021 12:52
Barplot should never have truncated axis but lineplots may have them, since size is not used for encoding



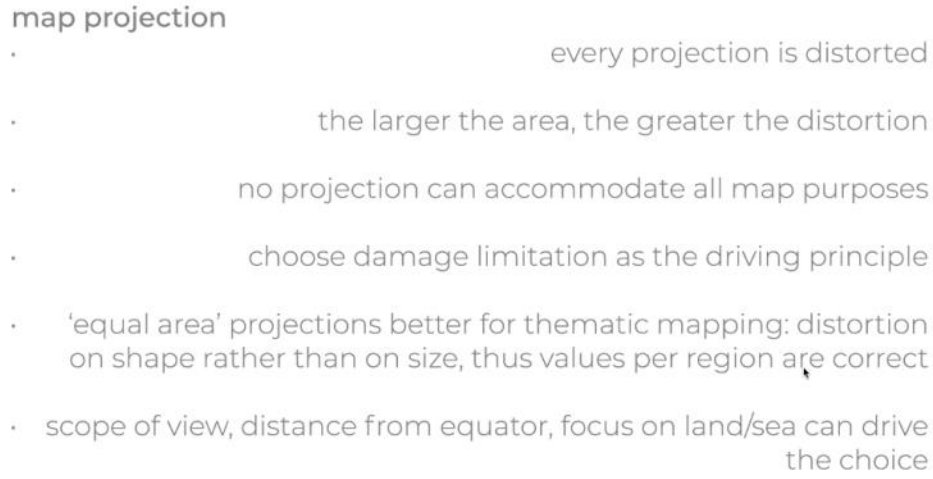
Ritaglio schermata acquisito: 14/04/2021 12:53

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 mantains the trustworthiness of the graph.

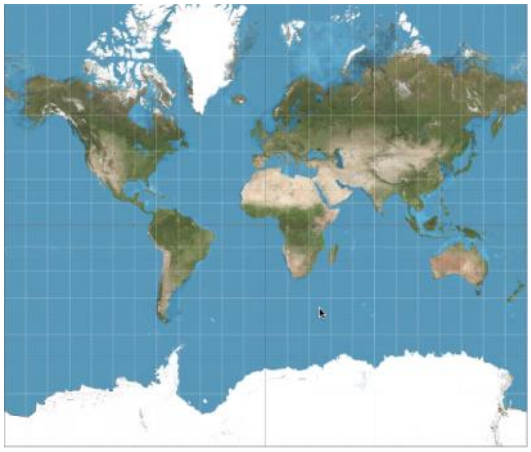


Ritaglio schermata acquisito: 14/04/2021 12:55

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.



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

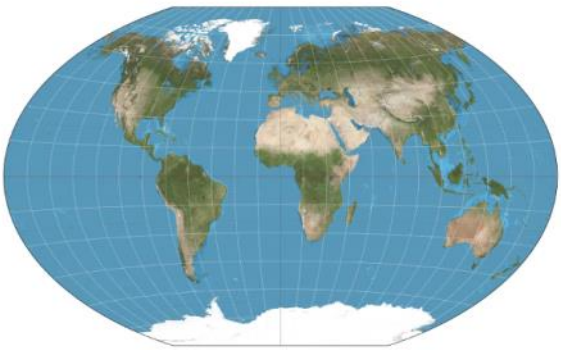
- azimuthal equal-area
- spherical projection
- good for hemisphere or continent-level view
- recommended by european environment agency for any european mapping purposes

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

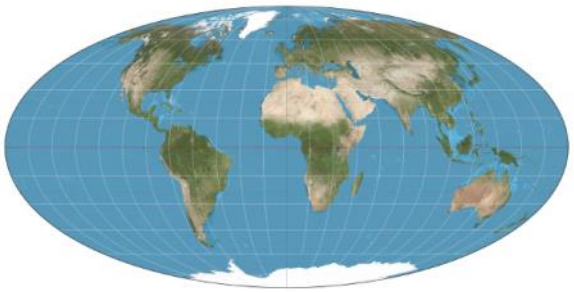
NATIONAL GEOGRAPHIC

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

map projection

quick'n'dirty guidelines

mollweide

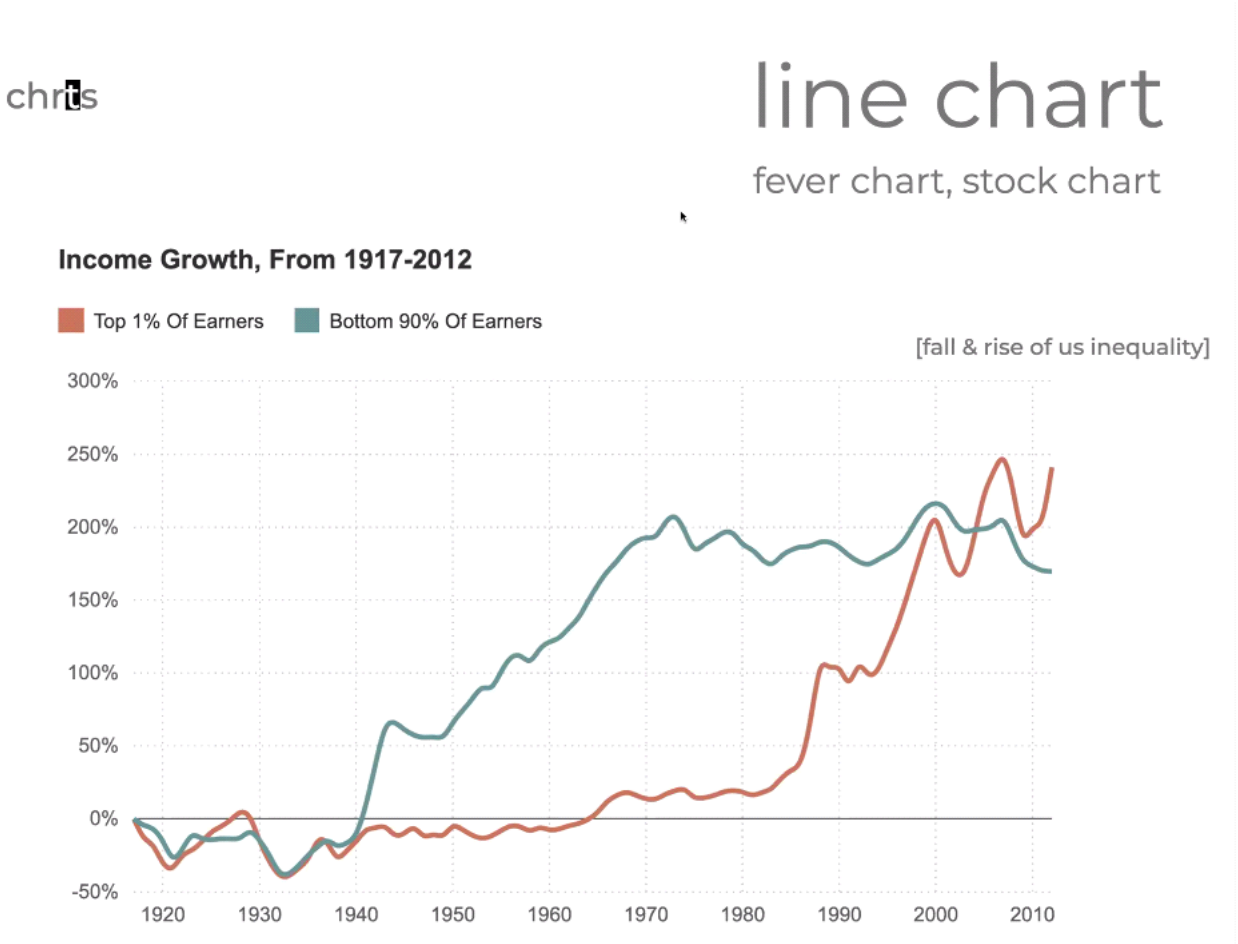


- greater emphasis on the accuracy of ocean areas
- useful for atmospheric mapping and flight paths

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

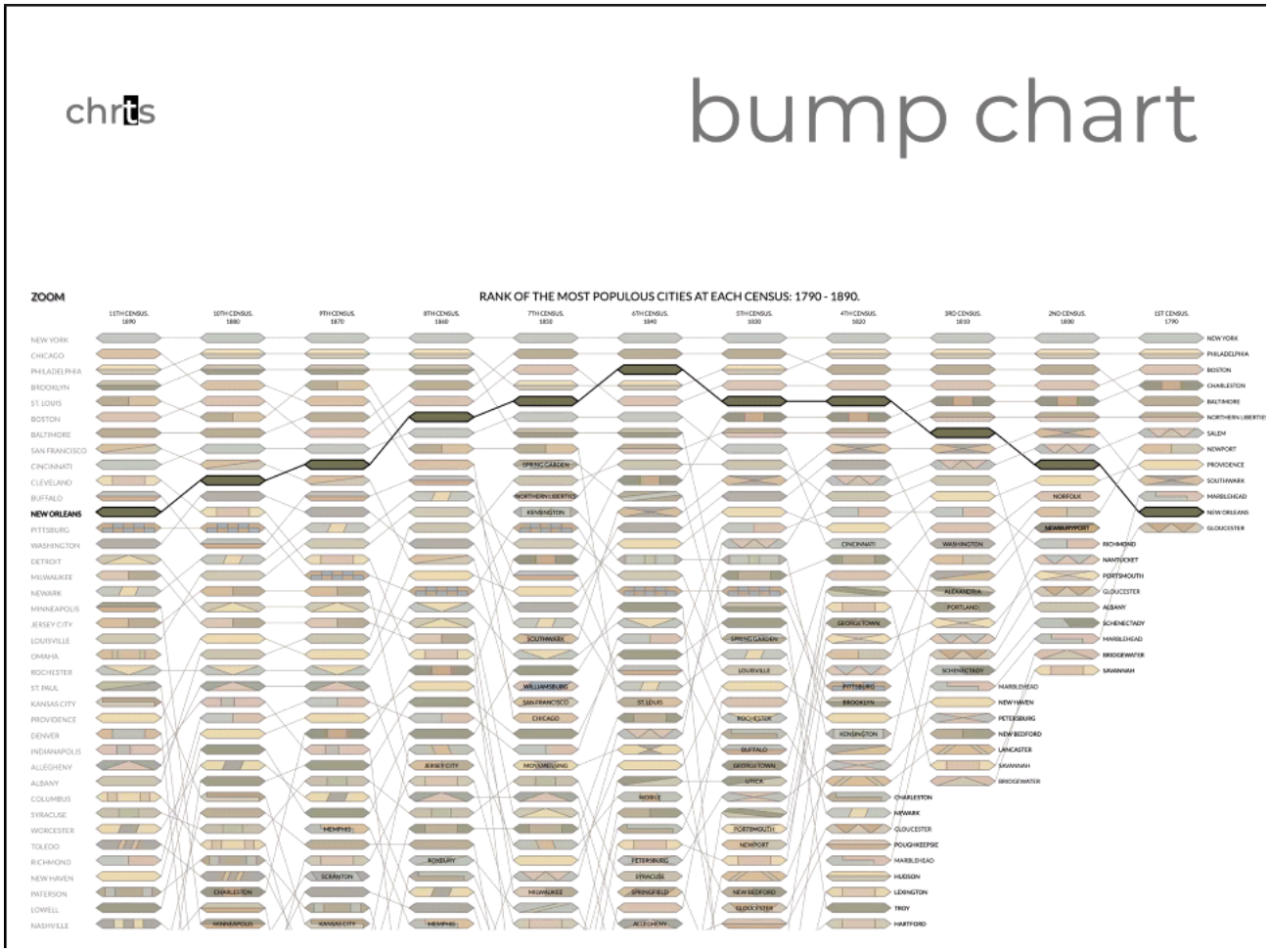
Charts

mercoledì 21 aprile 2021 11:31



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

Line plot with the differecne 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 indepedent 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.

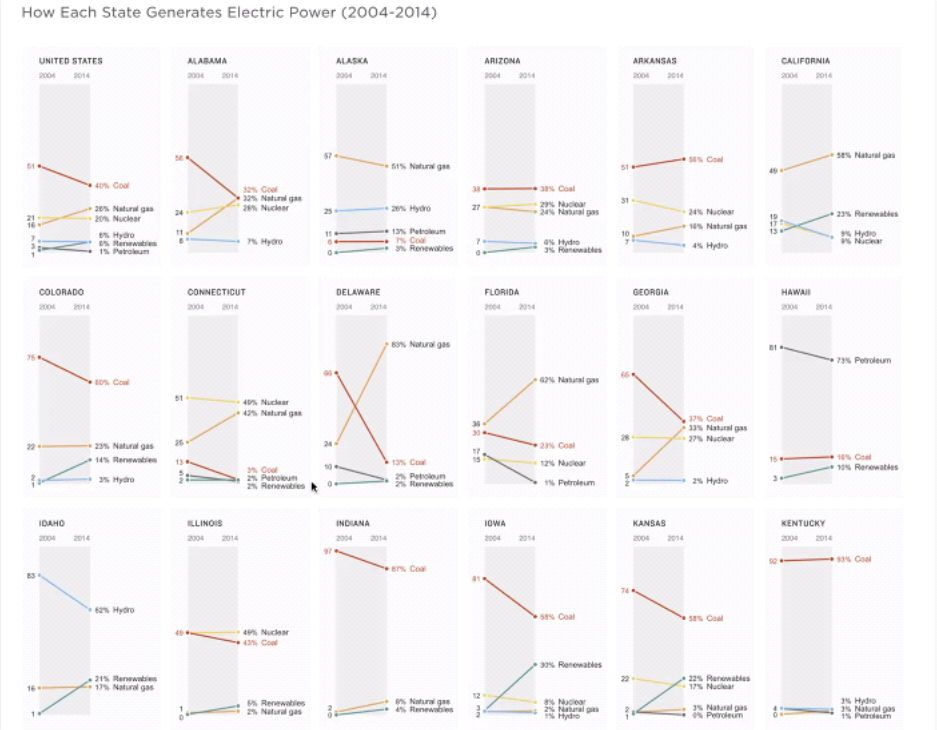


Screen clipping taken: 21/04/2021 11:37

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 characterstic 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.

chrts

slope graph chart



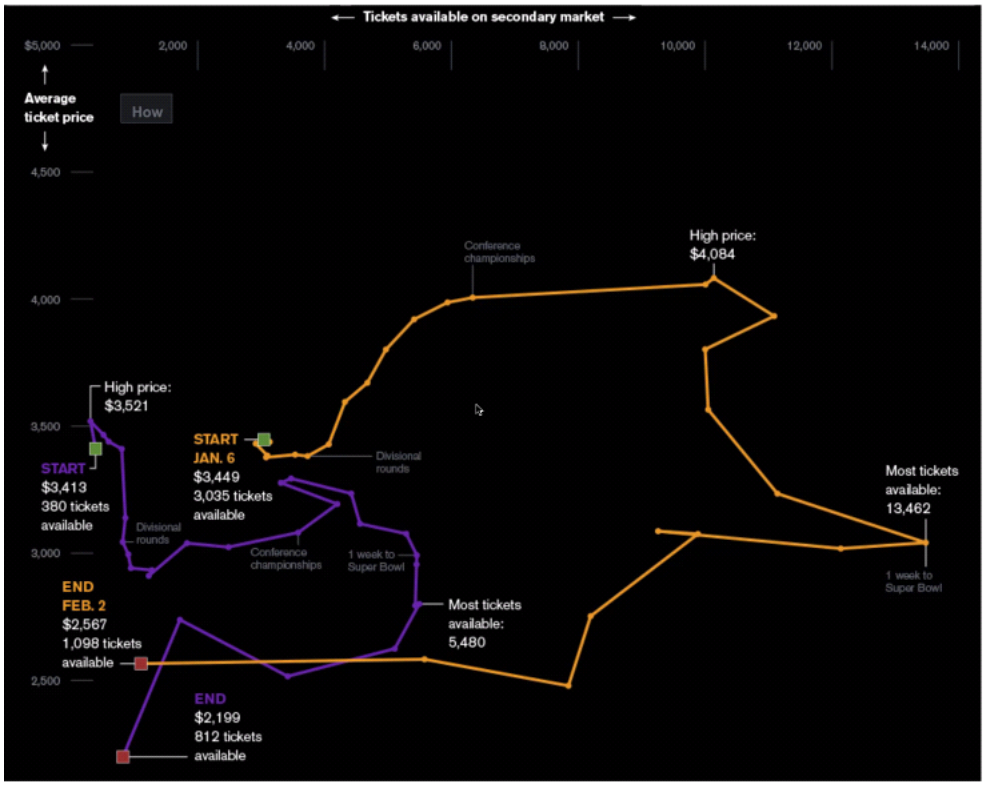
Screen clipping taken: 21/04/2021 11:40

Similarly here we have the slope graph chart. What changes is the presentation. Here we have multiple layouts. Usaully there are two steps in time and several categroeis marked by different colors.
The line is connecting the values for the same category for each sample. The intersting 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 influecne the slope of the connecting lines.

chris

connected scatter plot

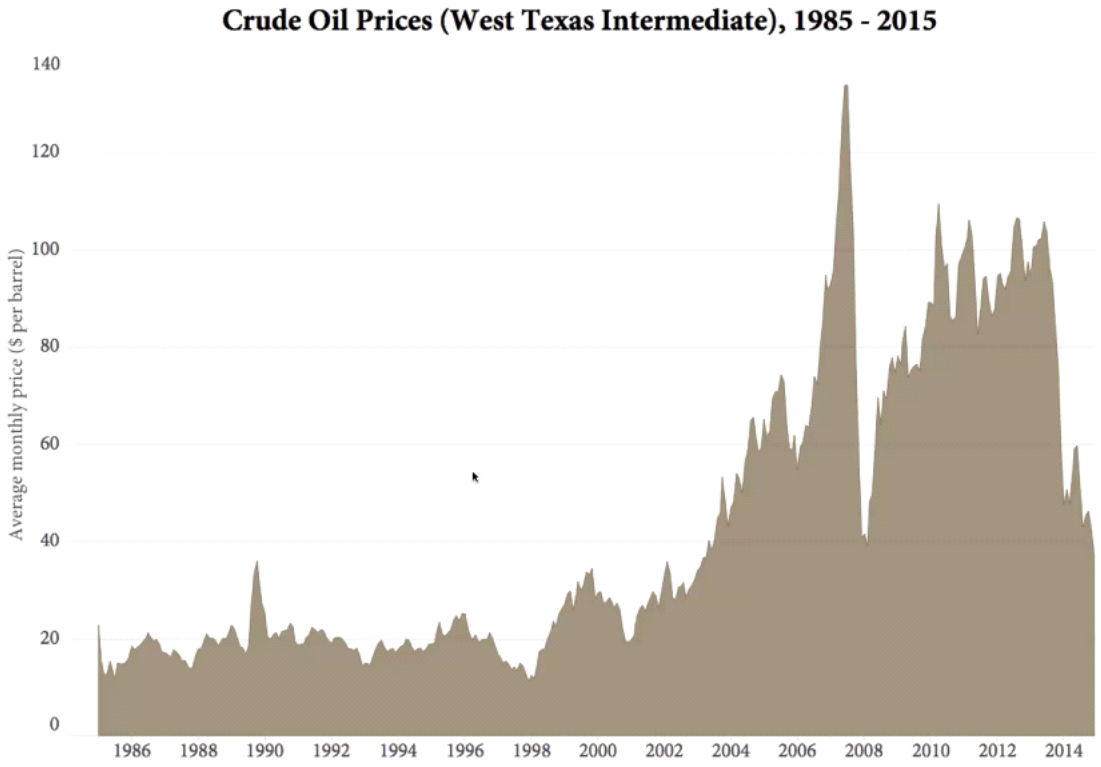
trail chart



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 superbowle 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 alloe to have a very nice narrative.

area chart



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

The area chart where the longitudial 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.

chrts

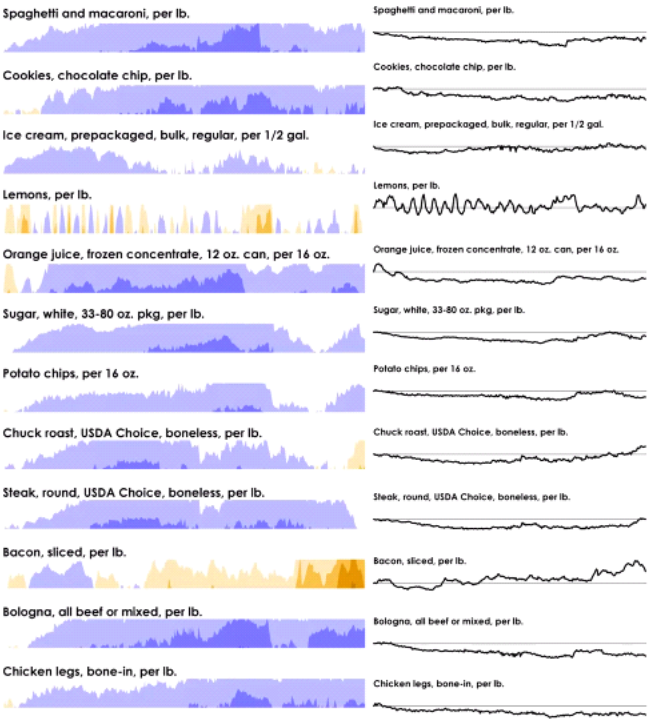
horizon chart

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.

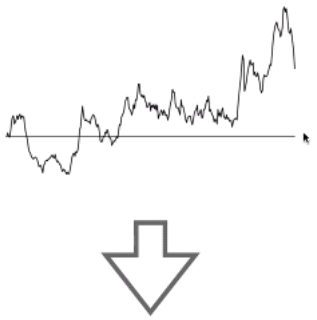
chrts

horizon chart



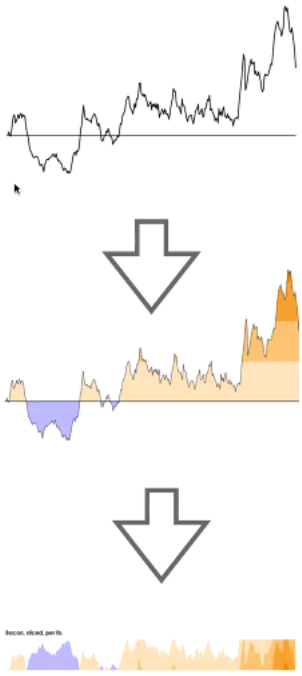
Screen clipping taken: 21/04/2021 11:49

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.



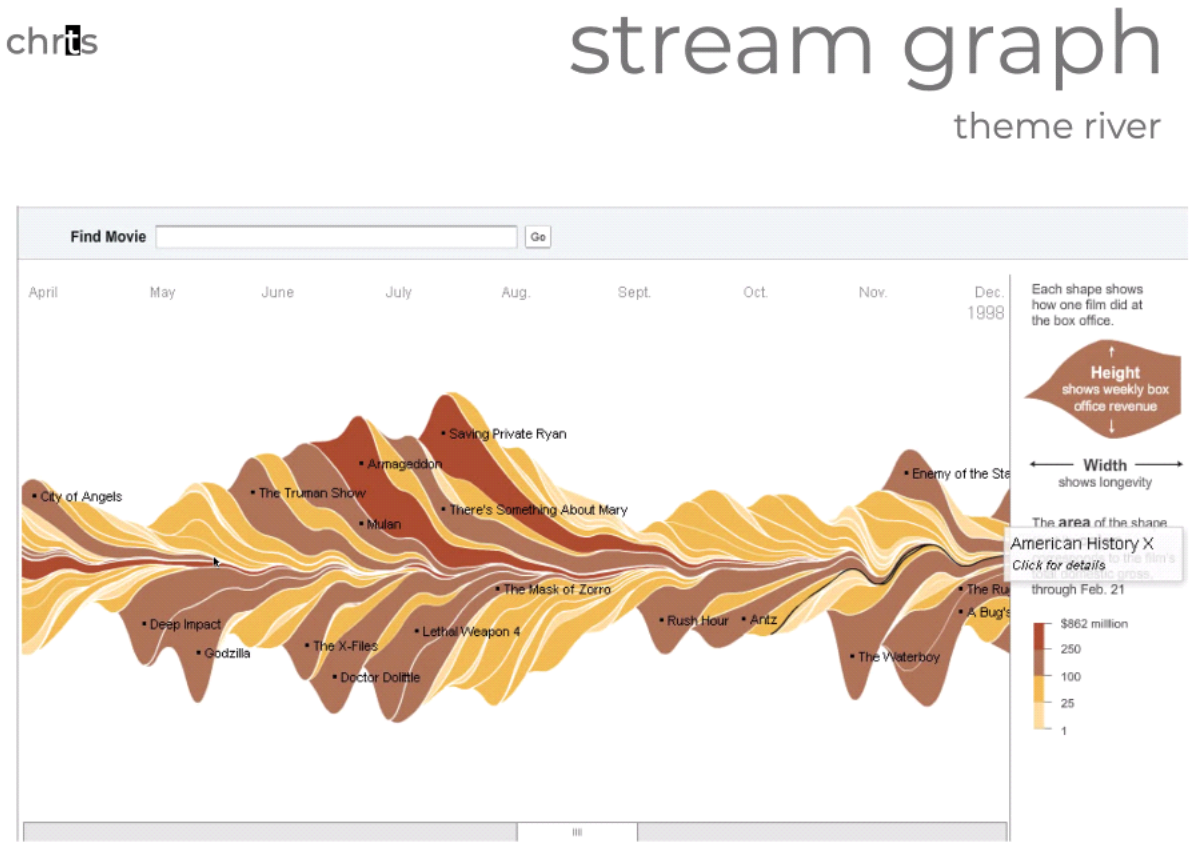
Screen clipping taken: 21/04/2021 11:50

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.



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

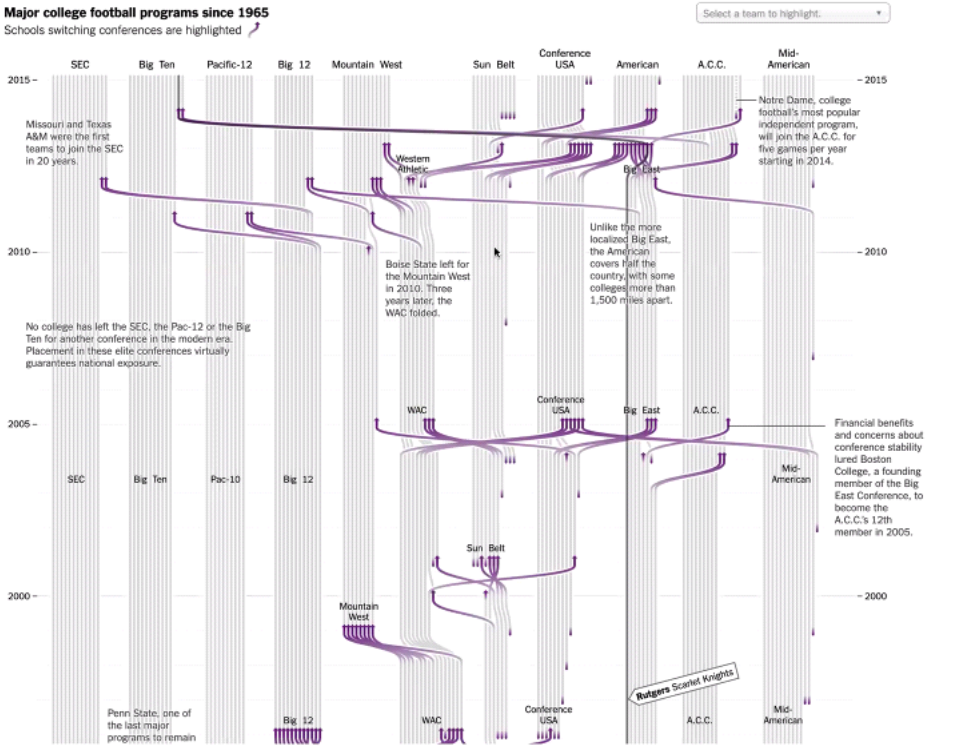
another famous plot is the stream graph.
Here we are showing the box office receipts for some blockbuster.
This is a very visual way of representing data and is not for exact value perception.
It is good for a graphical visual impact and is good for detect seasonal patterns.
The color legend is essential. Using smoothed lines gives a much better impression of the streaming

effect.

chrts

connected timeline

relationship timeline, storyline viz, swim-lane chart



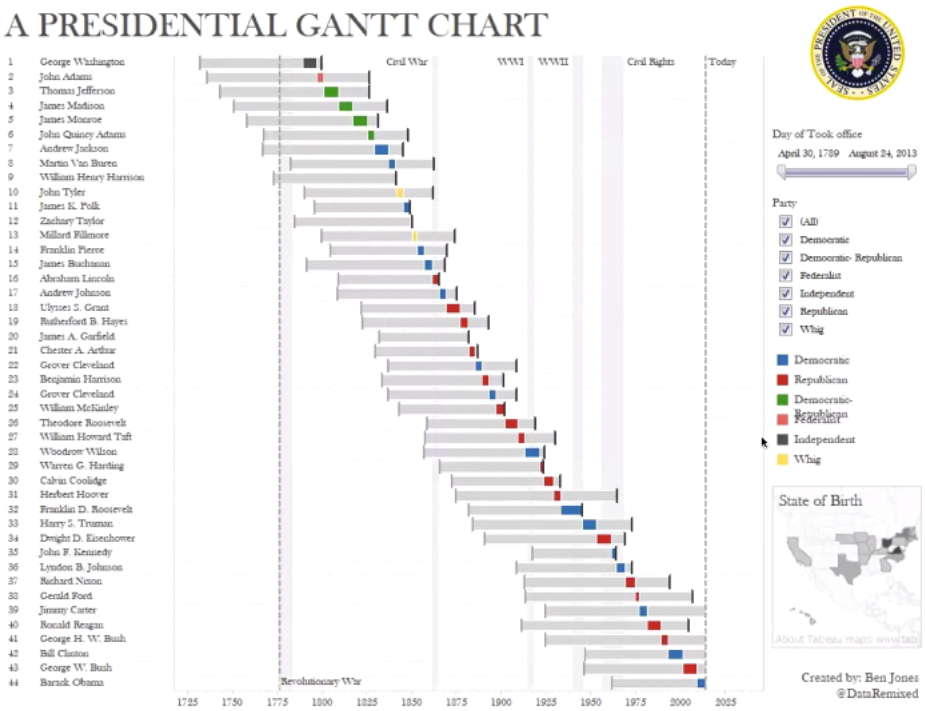
Screen clipping taken: 21/04/2021 11:57

Connected timeline -> usually in this kind of plot everysample 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 unsual kind of data.
This is r*ather complex* layout and somehow innatural. 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.

chris

gannt chart

range chart, floating bar chart



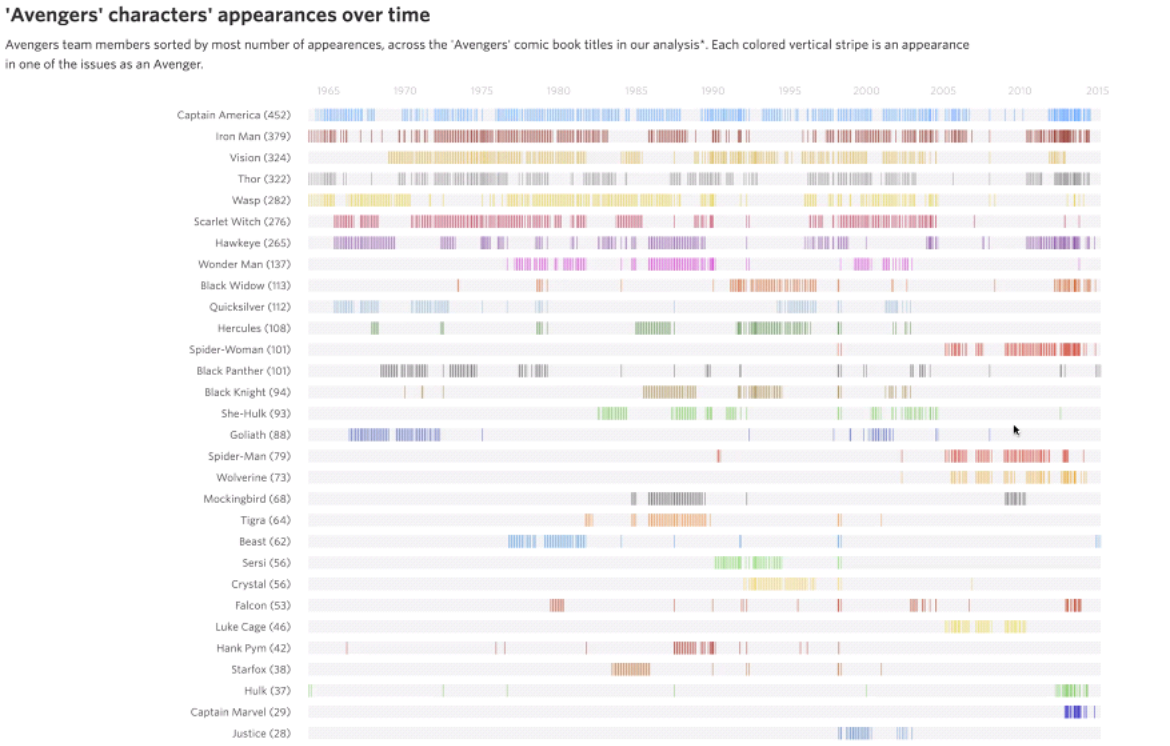
Screen clipping taken: 21/04/2021 12:00

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.

chris

instance chart

milestone map, barcode chart, strip plot



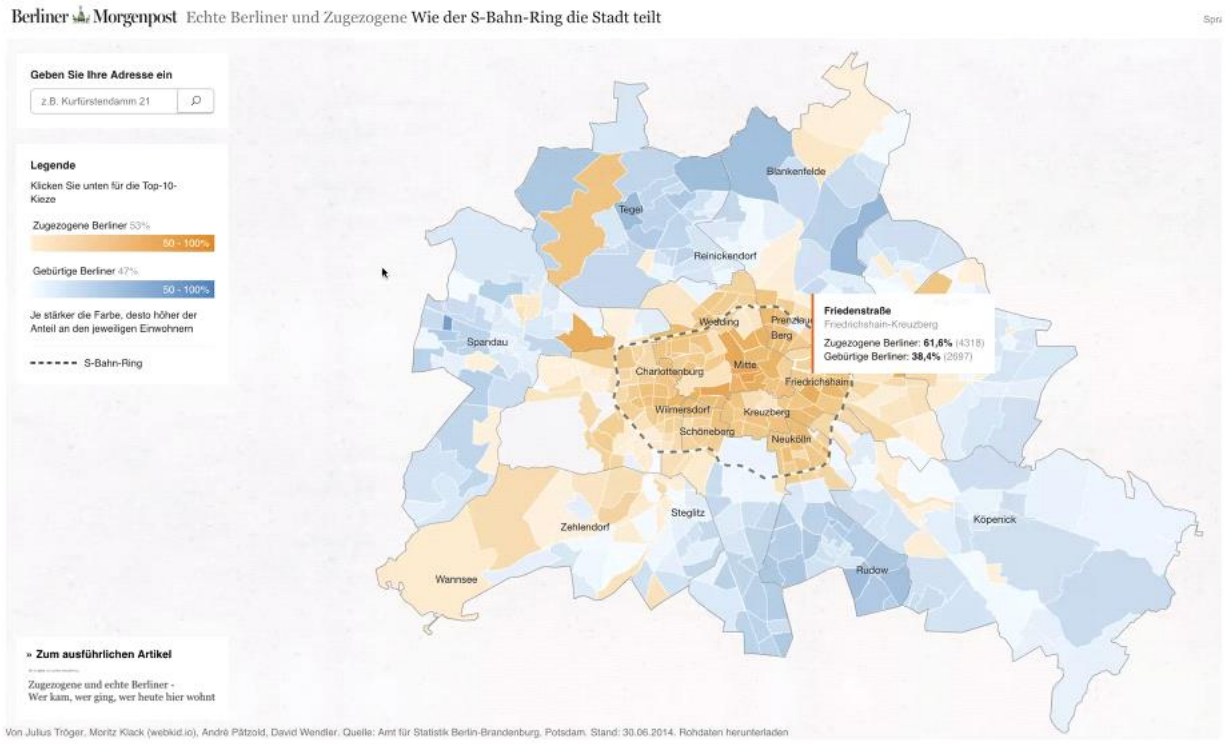
Screen clipping taken: 21/04/2021 12:03

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.

chrts

chloroplet map

heat map



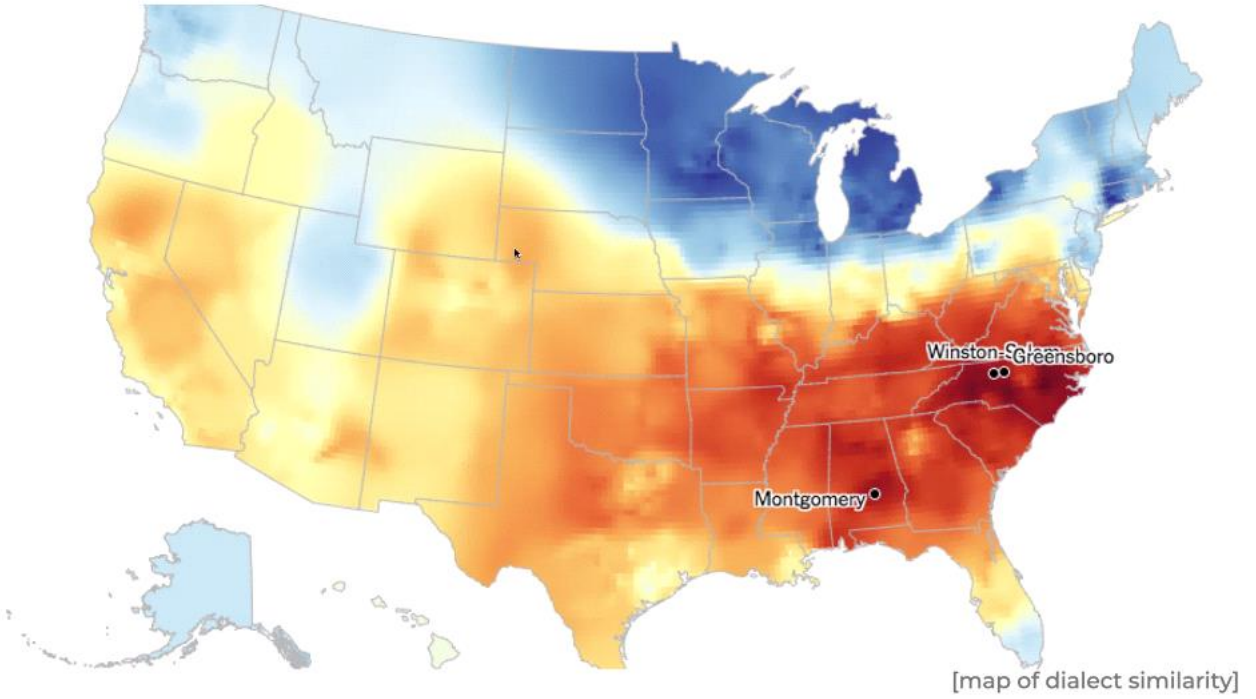
Screen clipping taken: 21/04/2021 12:05

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 your 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.

chrts

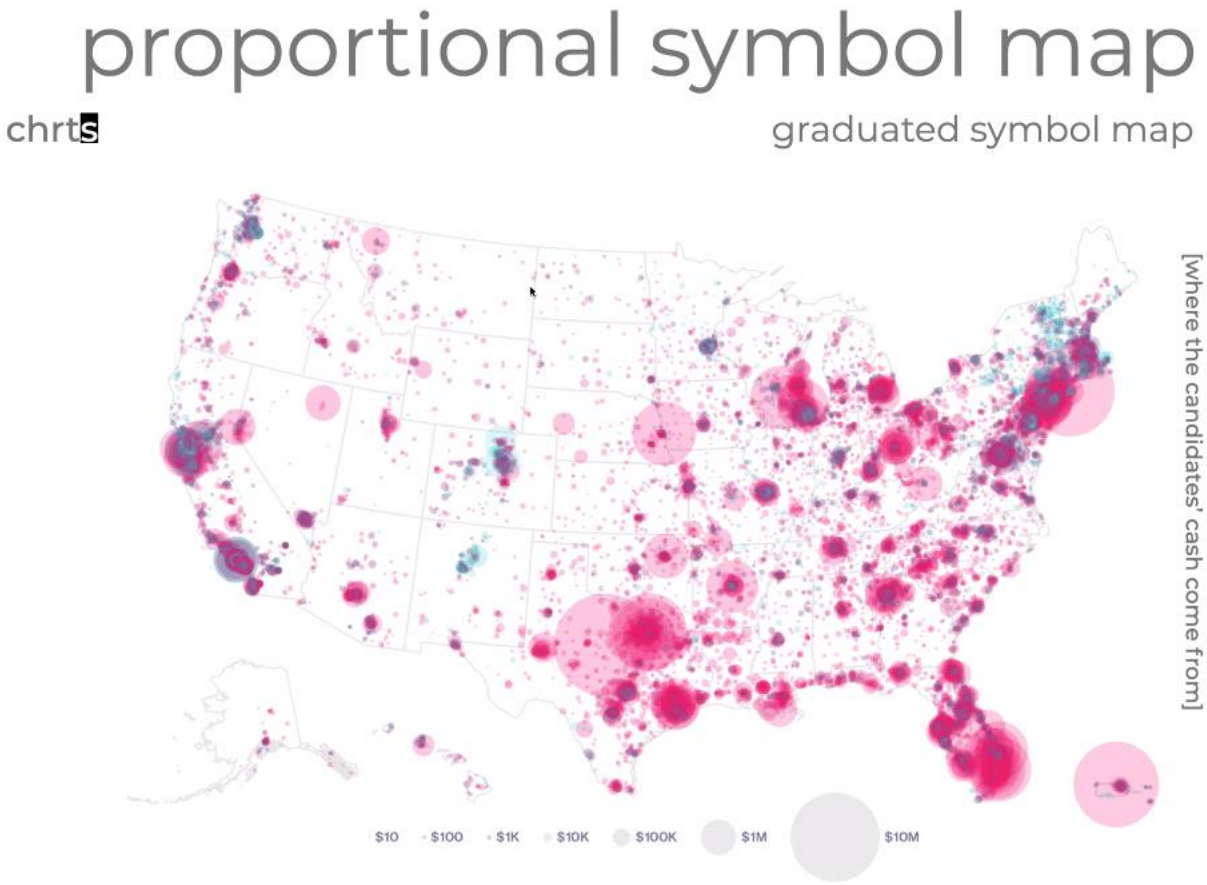
isarithmic map

contour map, isopleth map, isochrone map



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

Difference between isarithmic map and chloroplet.
In countour map we are not associated coloros to different regions subdivsions.
The data do not match with adminsitrative boarders.
So it is better when data and subregional boarders do not match.
Mind interpolation reducing precision of what you are seeing.
To optimize you have to keep the lengend close to the map and color borders not sharp.
If you need to ehnance the gradients of changes you can add isolines.



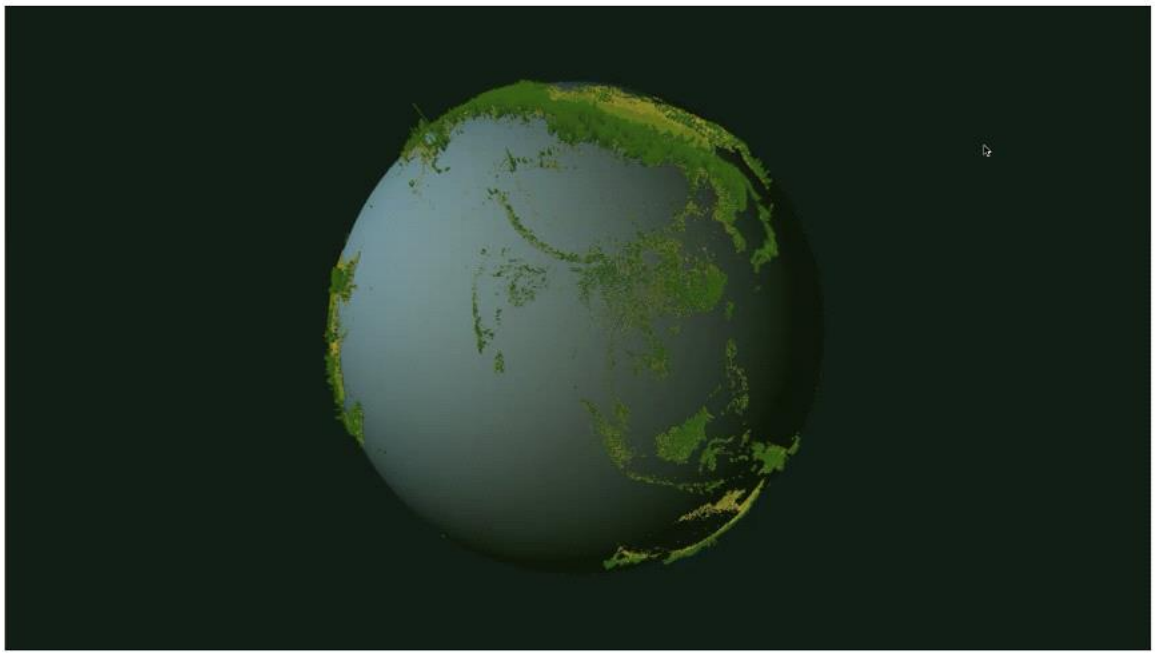
Screen clipping taken: 21/04/2021 12:15

Here the datum is pointled 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.

chrts

prism map

isometric map, spike map, datascape



[trillions of trees]

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.

chrts

dot map

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 coordinations 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 neccesarily applied to maps.

chrts

flow map

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

[arteries of the city]

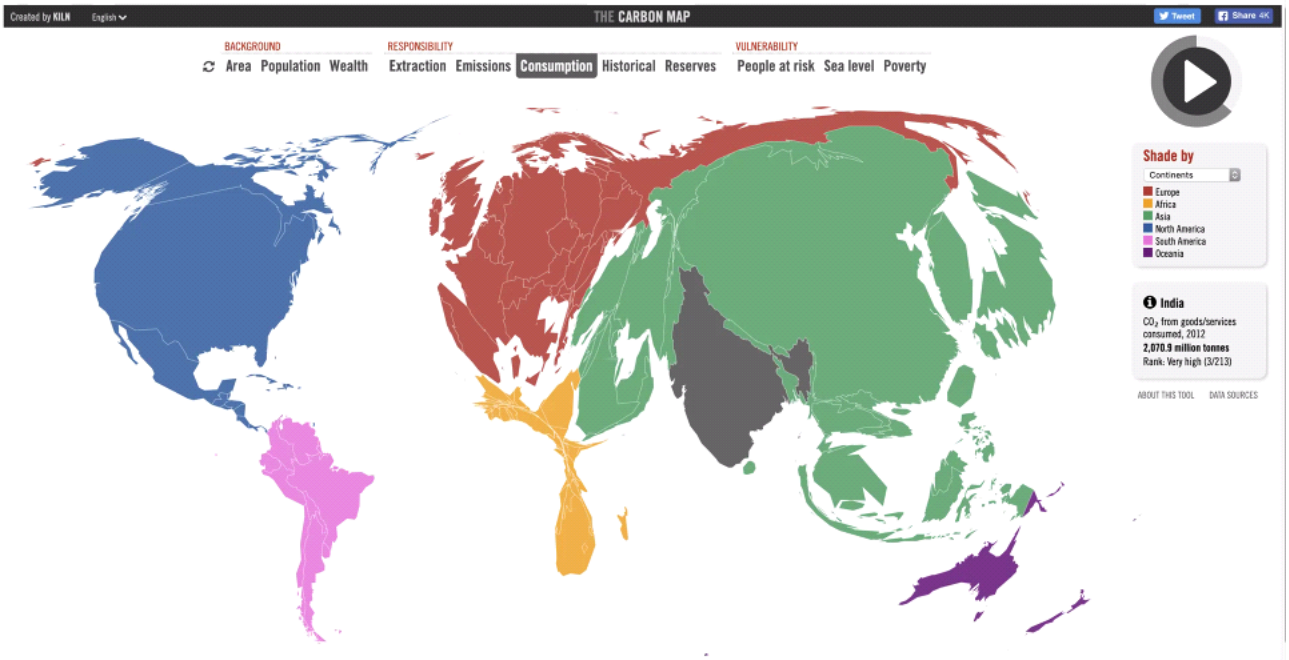
Screen clipping taken: 21/04/2021 12:21

Instead of dots we want trajcetories we use the flow map.
We select some interesting paths. We color such stream with appropriate color and the thickeness 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.

chrts

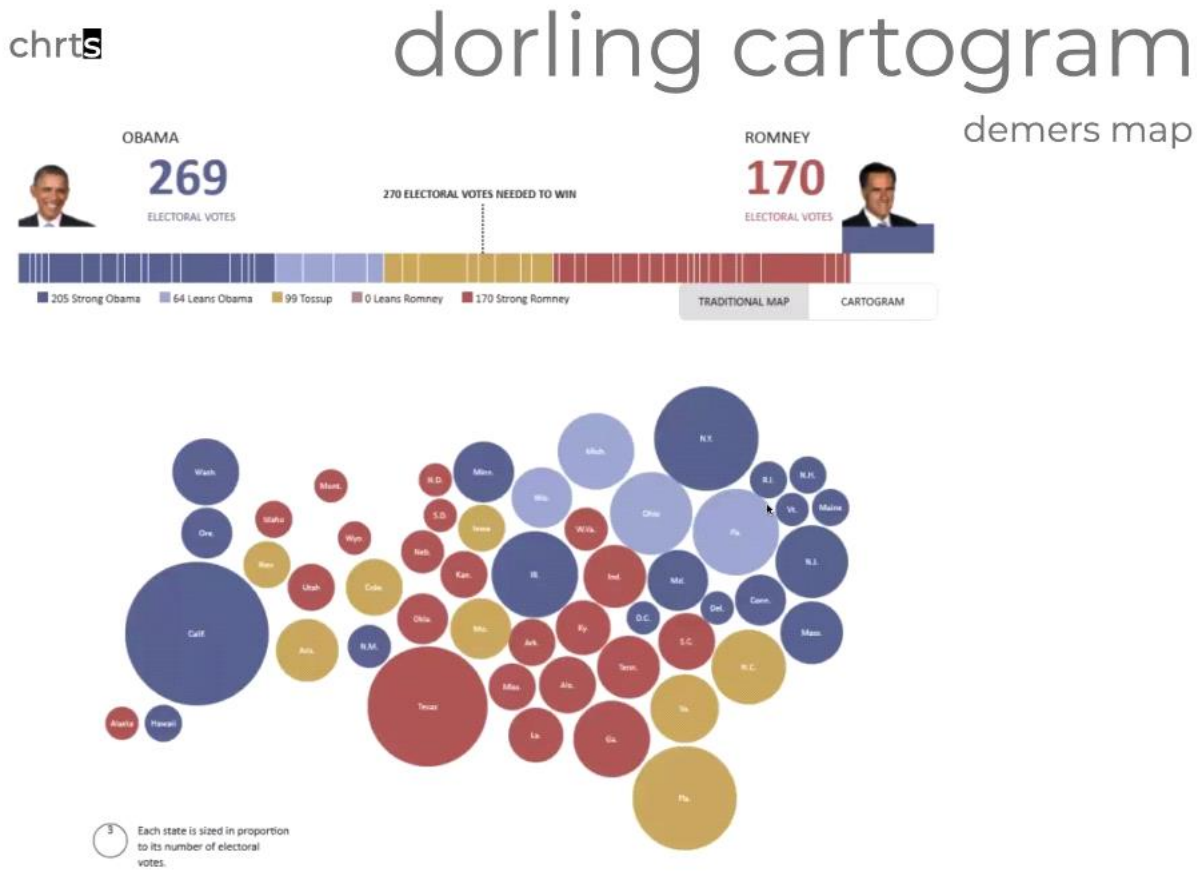
area cartogram

density-equalising map



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

They are very popular. Not many colors but every country is deformated according to their different values associated to colors.
We are looking for carbonium dioxite.
Need the user know the real country shape/size.
Compare rations with real size.
Manual animation helps.



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

Dorling cartograpm 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.

chrts

grid map

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

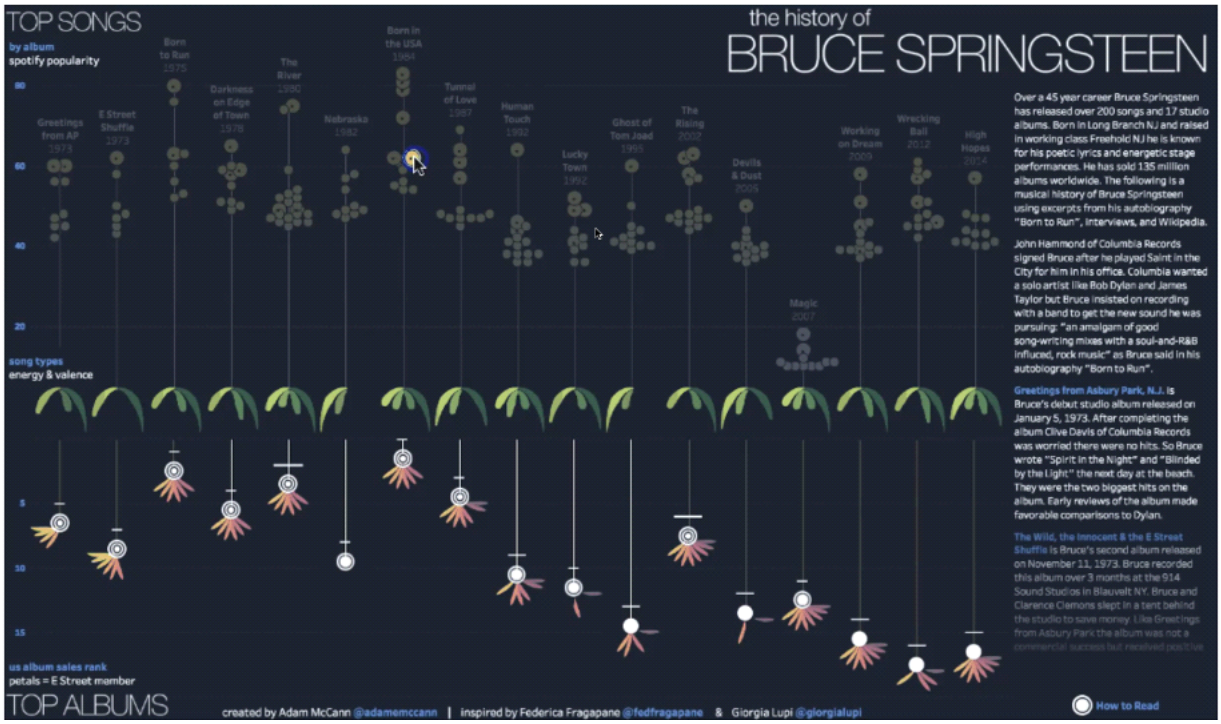
[mapping the spread of obesity]

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.

chrts

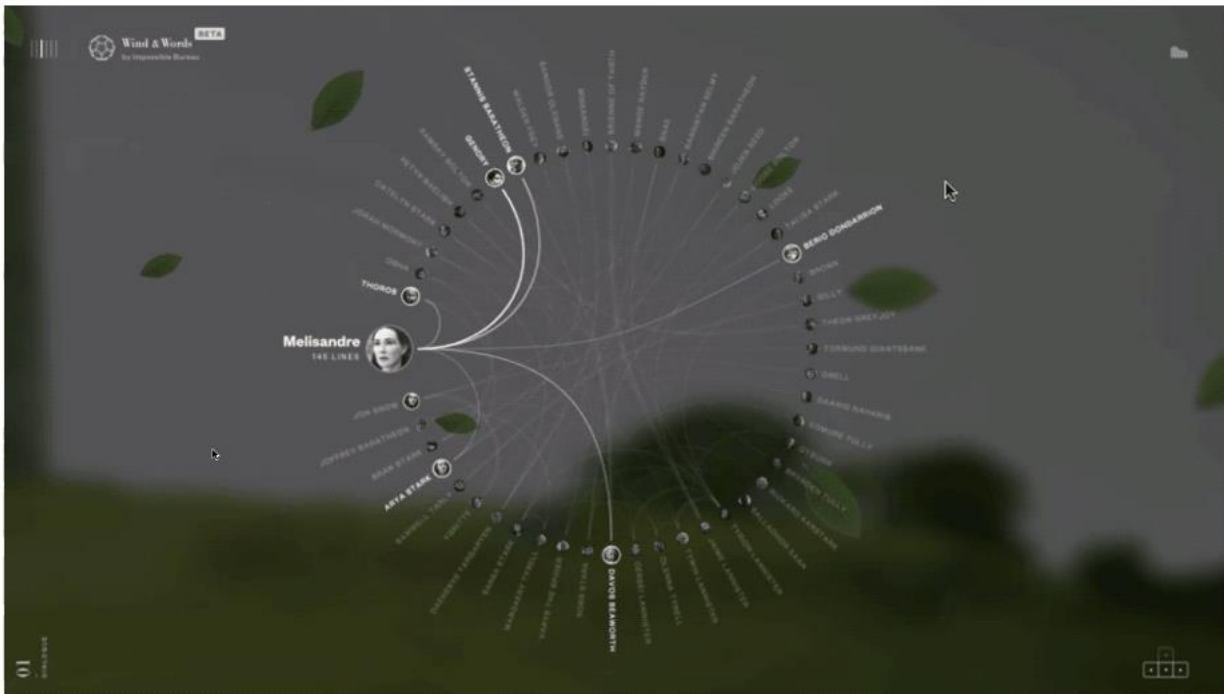
+ many untamed beasts



Screen clipping taken: 21/04/2021 12:32

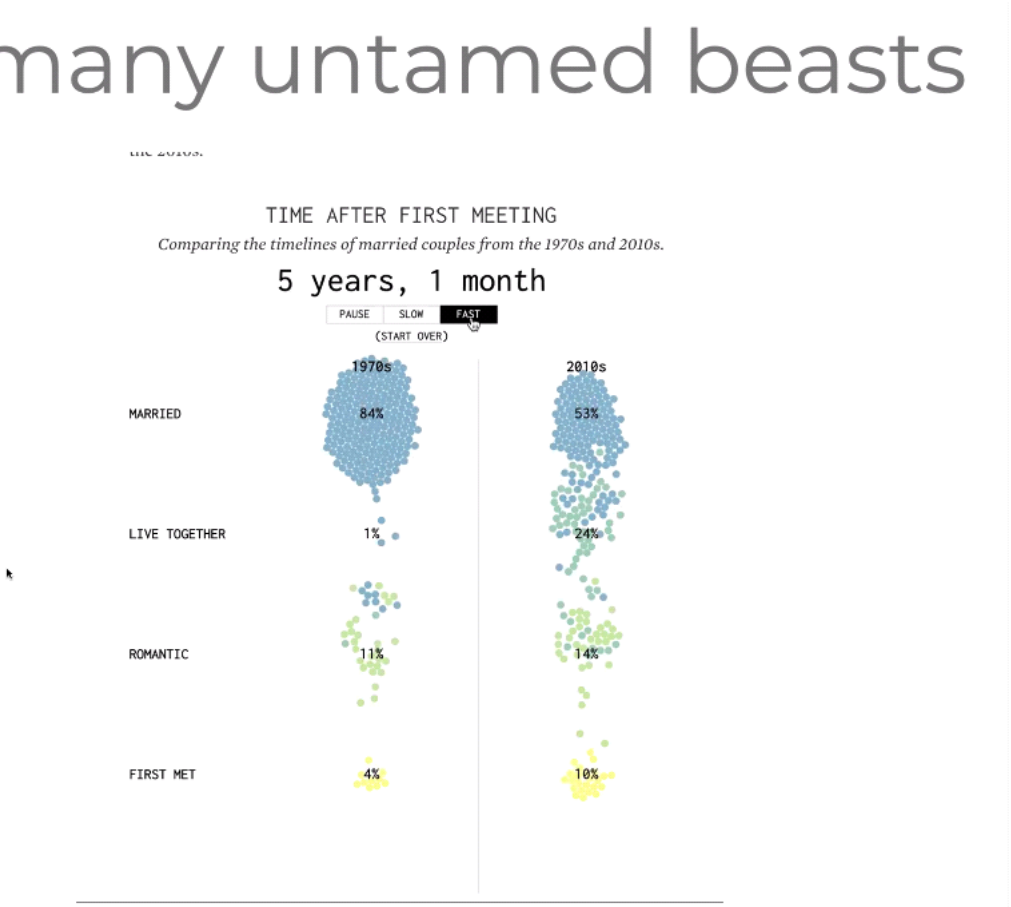
Musical history of bruce springsteen. For each album we have the popularity according to spotify. We have many different categories.

+ many untamed beasts



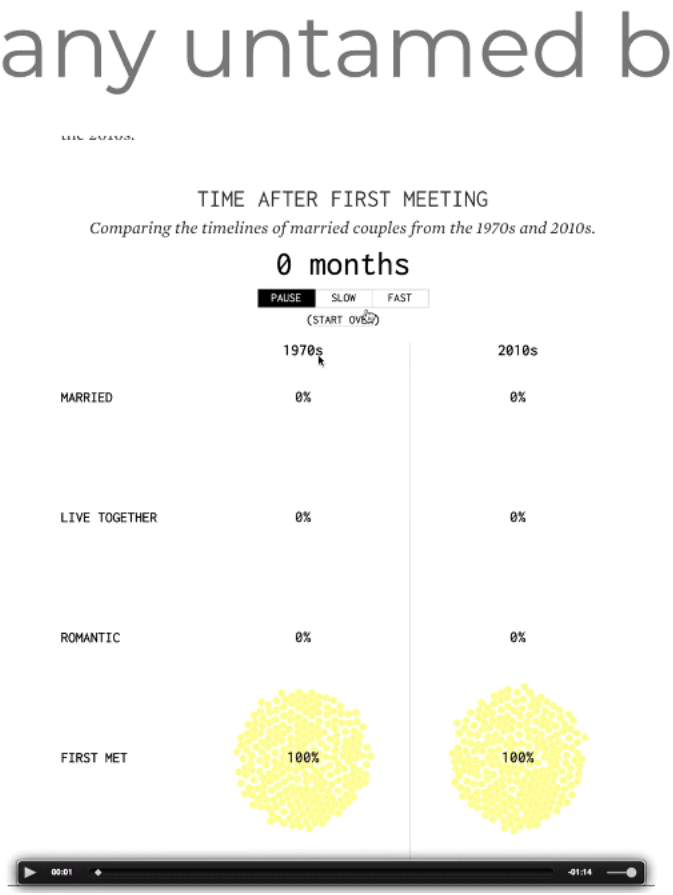
Screen clipping taken: 21/04/2021 12:34

Network for defining relationships inside a tv show.

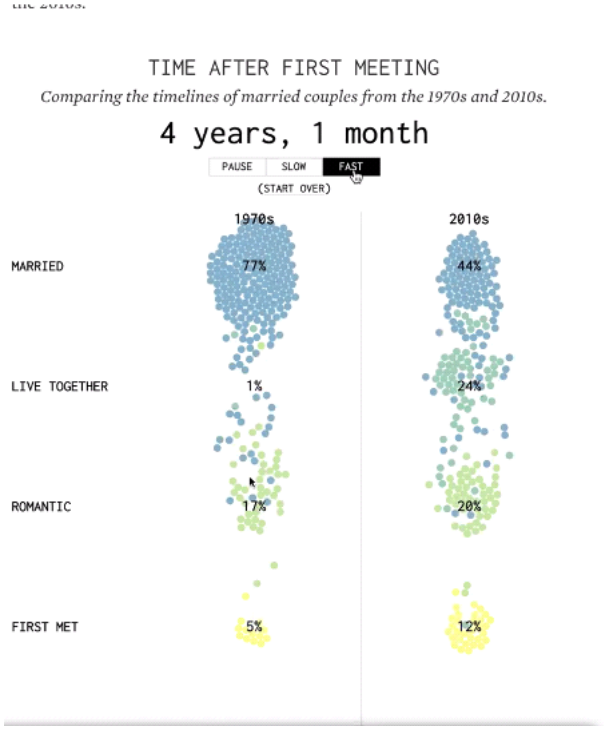


Screen clipping taken: 21/04/2021 12:35

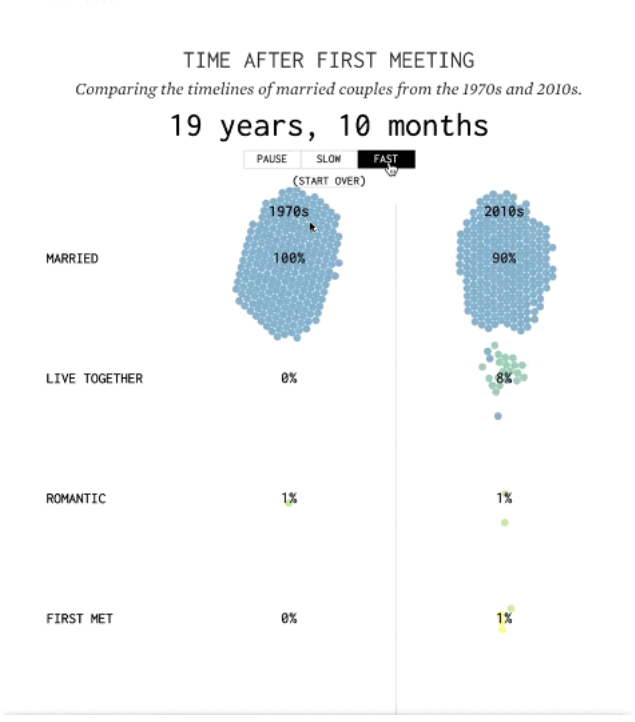
Longitudinal plot of how much married couples are taking from first meeting to getting married.



Screen clipping taken: 21/04/2021 12:36

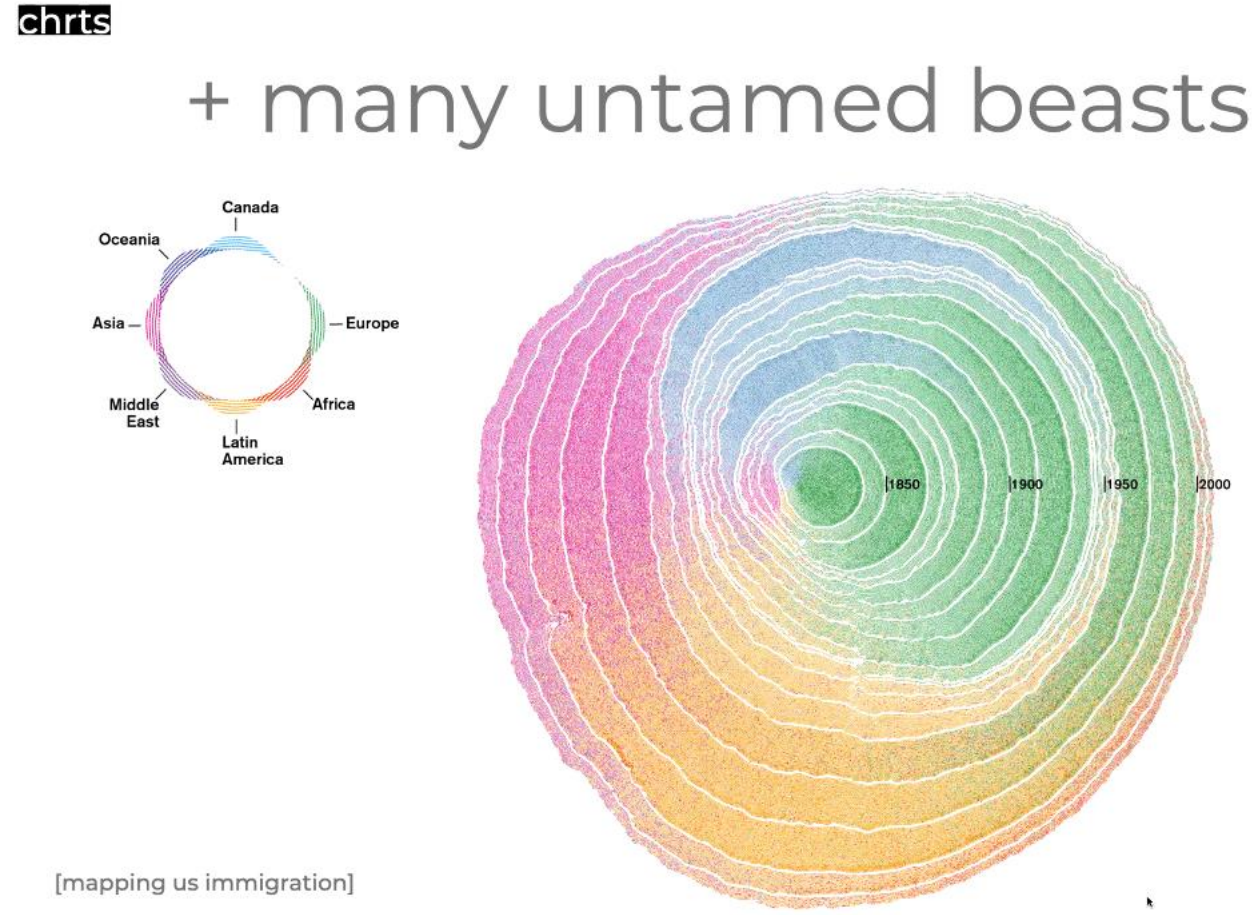


Screen clipping taken: 21/04/2021 12:36



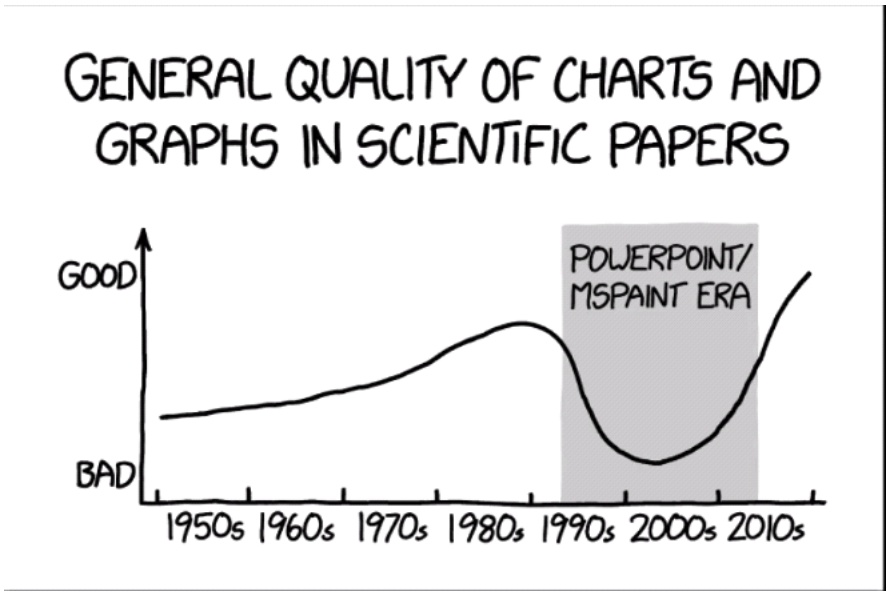
Screen clipping taken: 21/04/2021 12:37

Very famous representation -> very intuitive and gives the reader a very immediate taste of the underlining dynamic.



Screen clipping taken: 21/04/2021 12:39

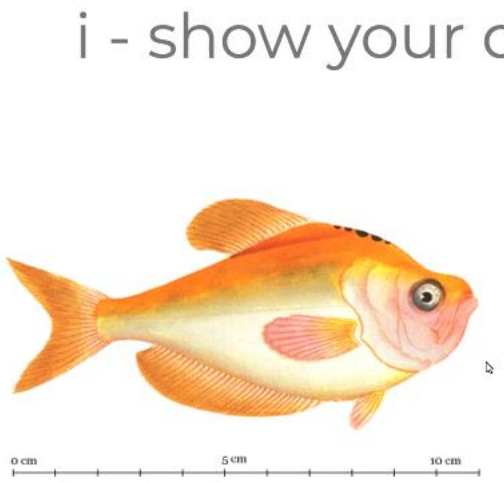
Colors are marking the provinence of the people who go in the US.
This is a tree stump.
Very intuitive and the color is very meaningful.



Screen clipping taken: 21/04/2021 12:41

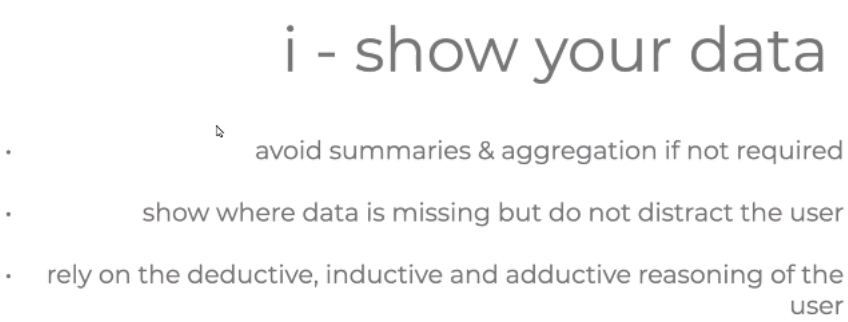
Rules

martedì 27 aprile 2021 11:39

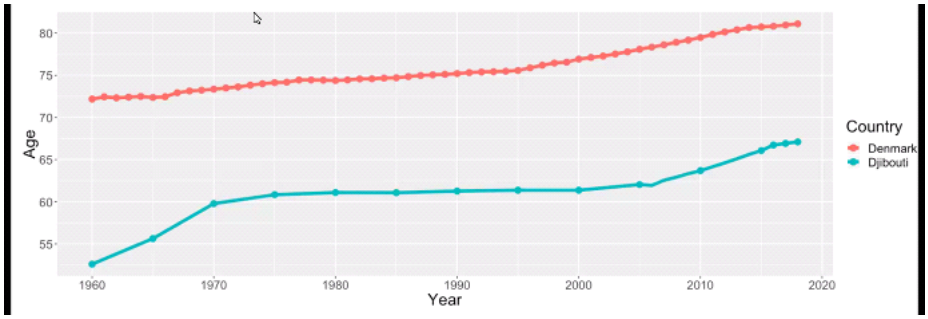


Ritaglio schermata acquisito: 27/04/2021 11:40

Insert all the elements that help to intepret your data.

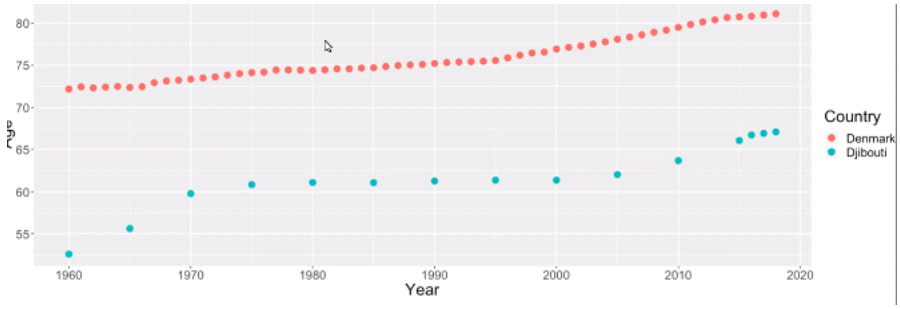


Ritaglio schermata acquisito: 27/04/2021 11:40



Ritaglio schermata acquisito: 27/04/2021 11:41

This is a time series of the life expectancy in two countries : a fisrt world country and a thrid world country.
We see how the difference between these two countries is still the same.
The problem is that if you do not look carefully you see only lines



Ritaglio schermata acquisito: 27/04/2021 11:44

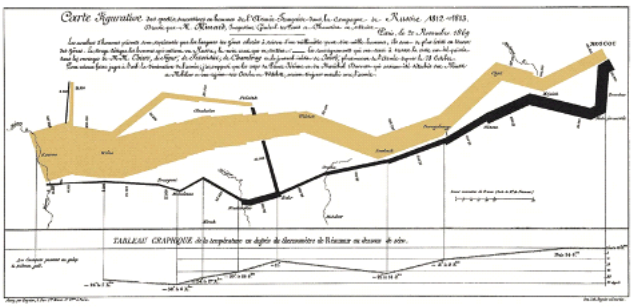
This is a much more honest plot.
Do not connect points in order to hide the fact that you have less observations -> be honest and show the data as they are.

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 strenght of the graph is that they can summarize many textual informtion.

iii - avoid chartjunks

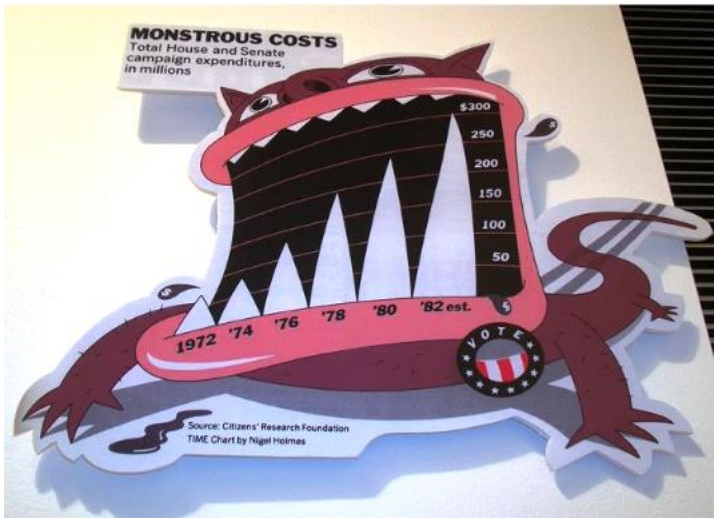
too many decorative elements added, detracting the user from the actual message

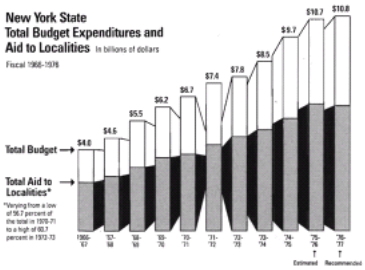
typical example: the unneeded 3d effects

''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 decoratives elements.
If you have little data you cannot compasate this with many decorative elements to fill up your graph.

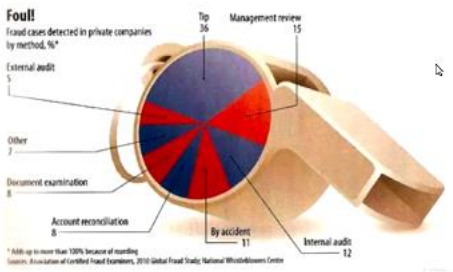


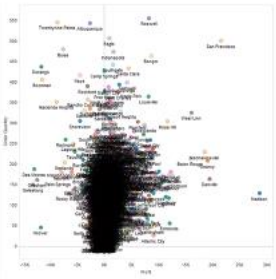


[new york times, 1976]

Ritaglio schermata acquisito: 27/04/2021 11:52

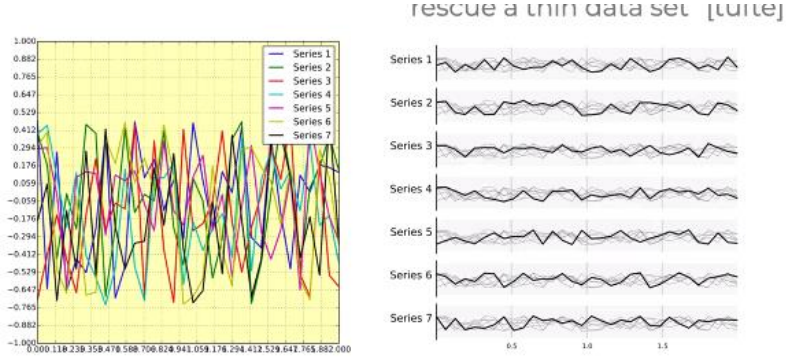
Here the underlying message is that they focus on the 1972 which should be the ideal situation.
What happen before too small, what happen later is too big.
Using shadows in this way should suggest that





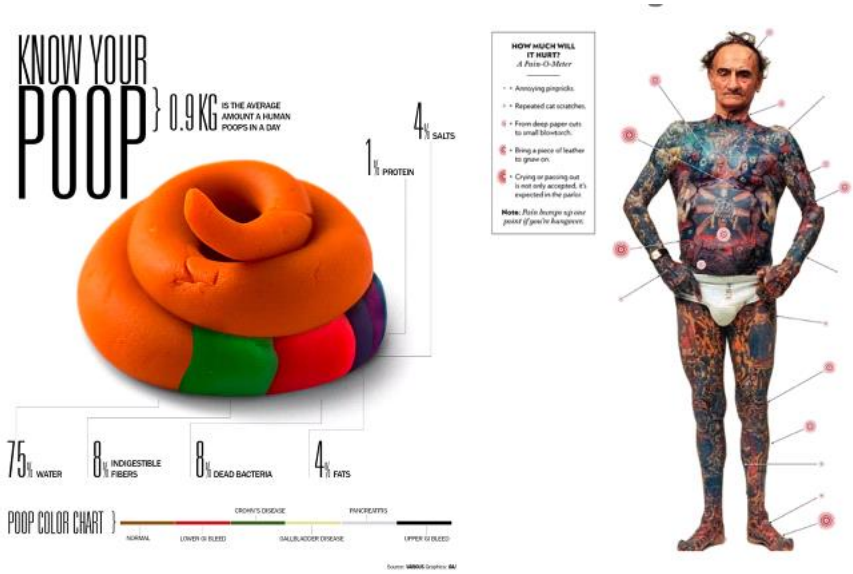
Ritaglio schermata acquisito: 27/04/2021 11:58

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.



Ritaglio schermata acquisito: 27/04/2021 12:11

Other level of chart junks in scientifc 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"

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

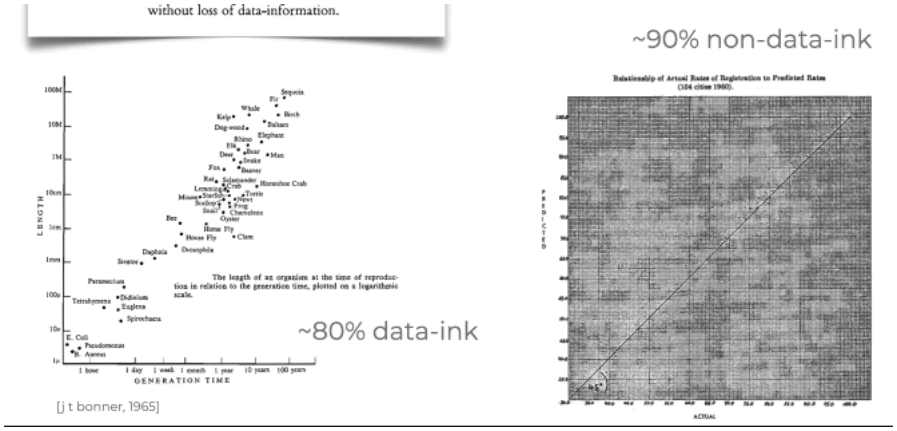
= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1.0 – proportion of a graphic that can be erased without loss of data-information.

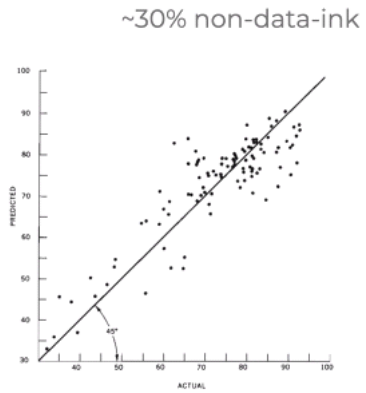
Ritaglio schermata acquisito: 27/04/2021 12:17

Another rule is more quantitative: the data-ink ratio. It can be computed as a formula. Quantity of

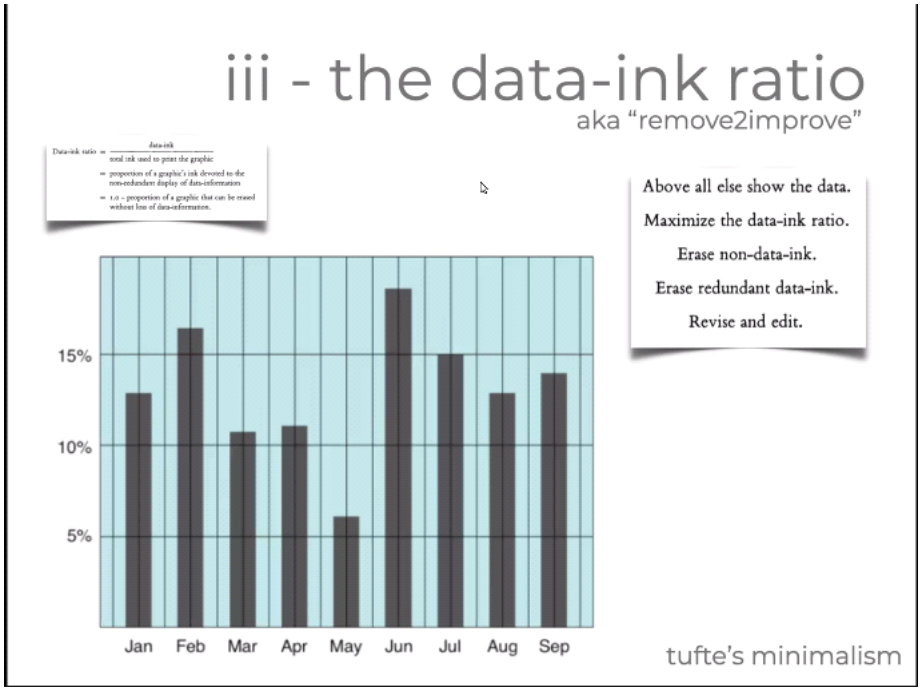
ink dedicated to data representaiton.



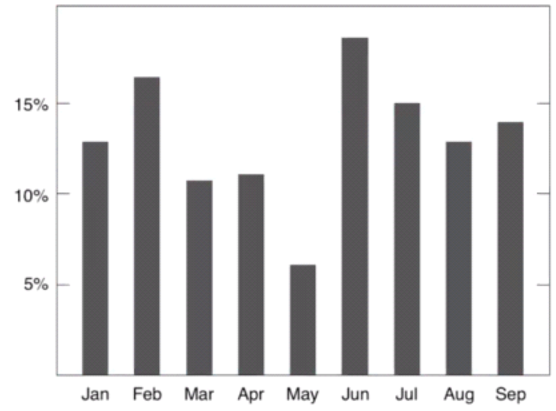
Ritaglio schermata acquisito: 27/04/2021 12:19



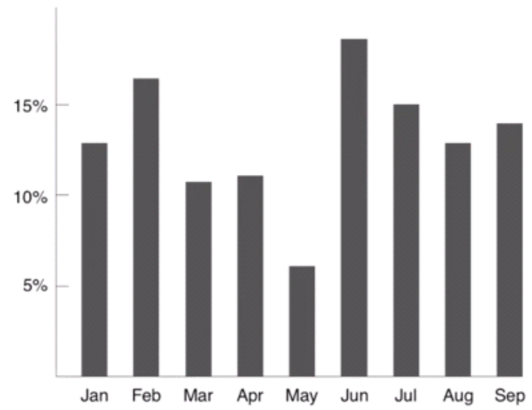
Ritaglio schermata acquisito: 27/04/2021 12:20



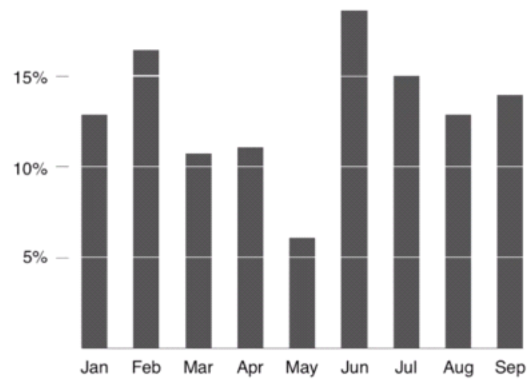
Ritaglio schermata acquisito: 27/04/2021 12:22



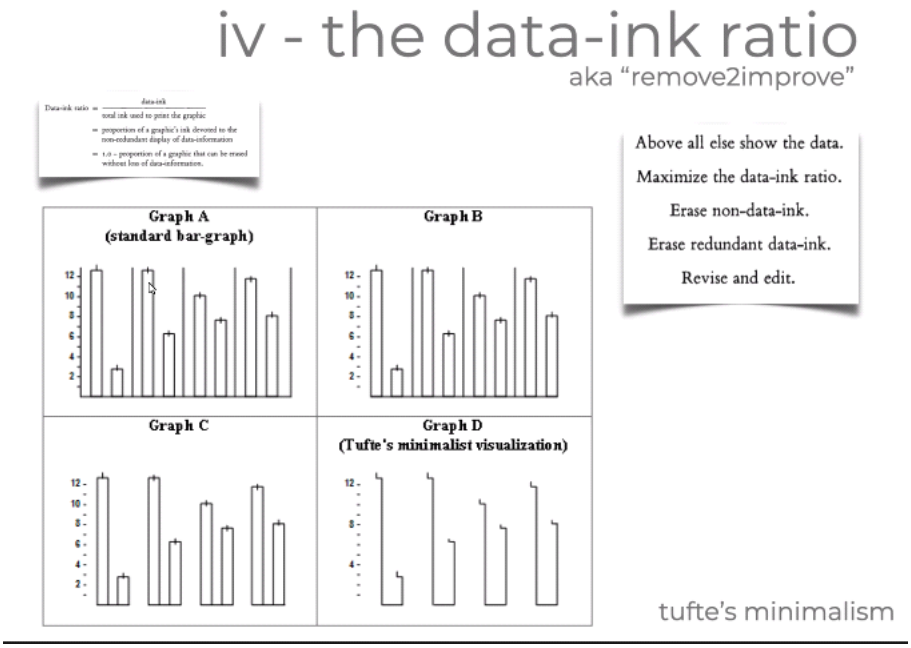
Ritaglio schermata acquisito: 27/04/2021 12:23



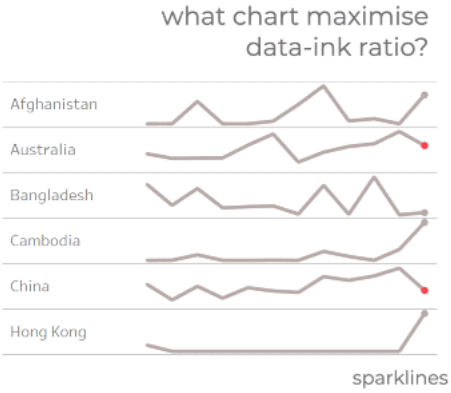
Ritaglio schermata acquisito: 27/04/2021 12:23



Ritaglio schermata acquisito: 27/04/2021 12:23



Ritaglio schermata acquisito: 27/04/2021 12:24



Ritaglio schermata acquisito: 27/04/2021 12:28

Sparklines are built for this purpose.

v - annotation

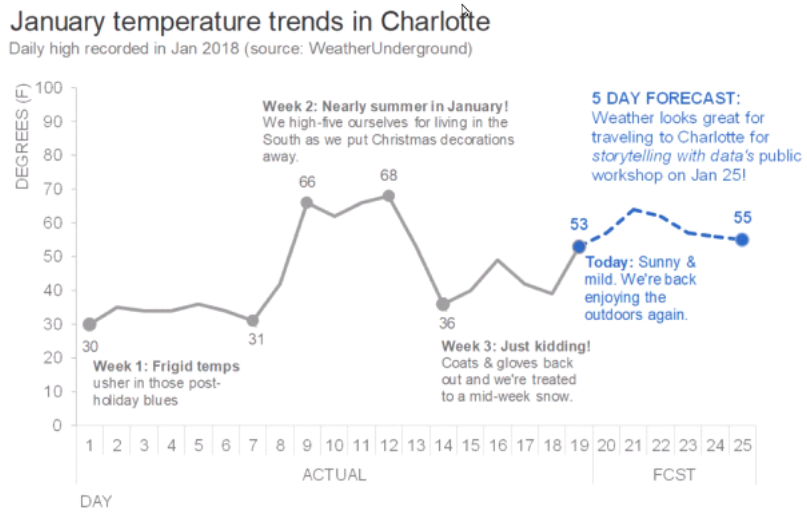
pictures are the focus, but words are important, too

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

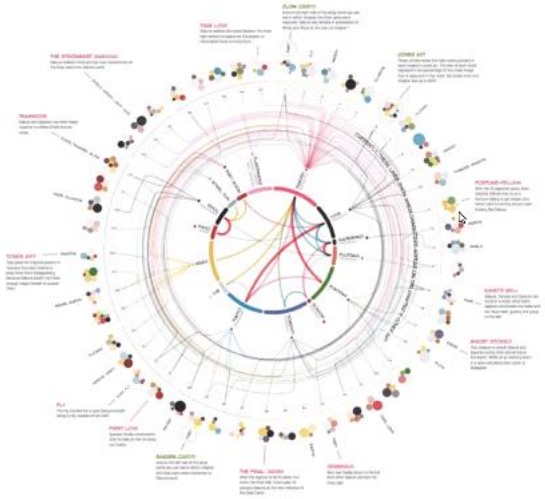
fine micro-level details become textures when viewed at the macro level

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

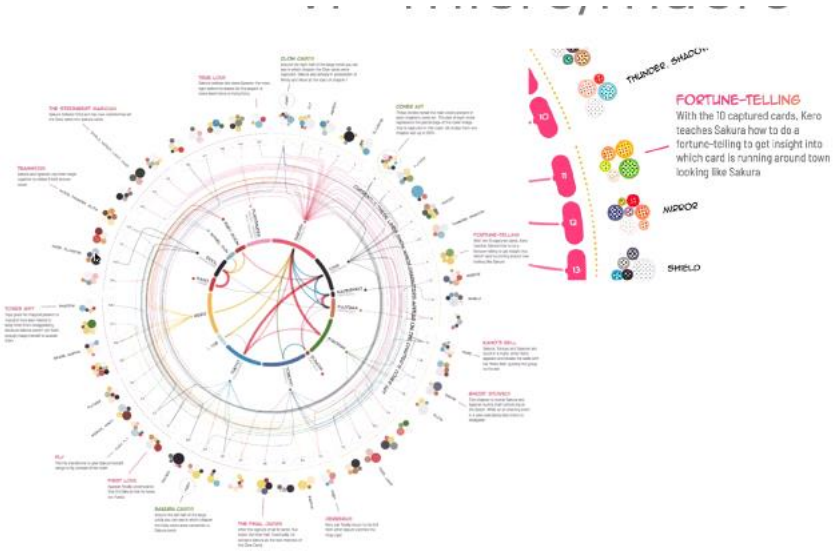
use interactive zooming when possible

Ritaglio schermata acquisito: 27/04/2021 12:35

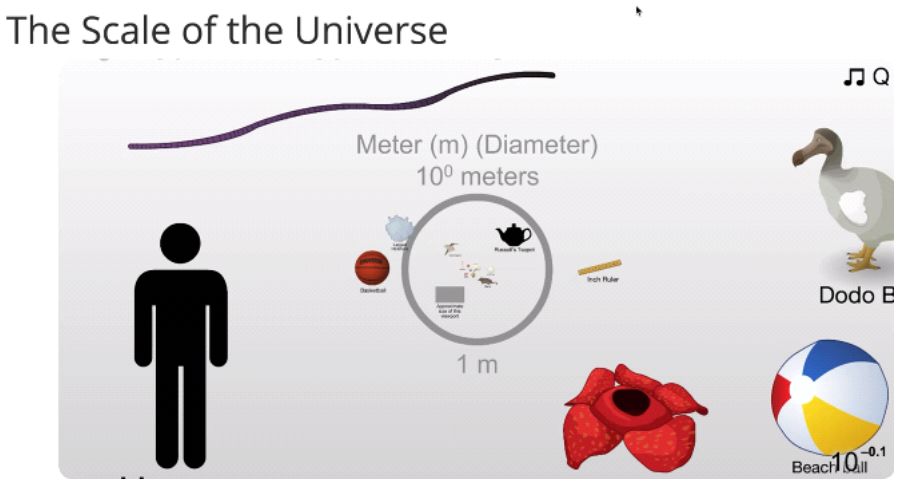


Ritaglio schermata acquisito: 27/04/2021 12:35

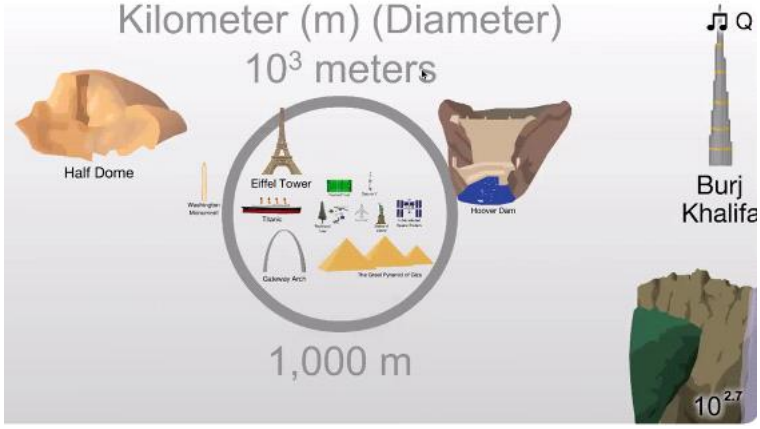
We have some nodes connected by links and many information outside.
We have other then other little circles.



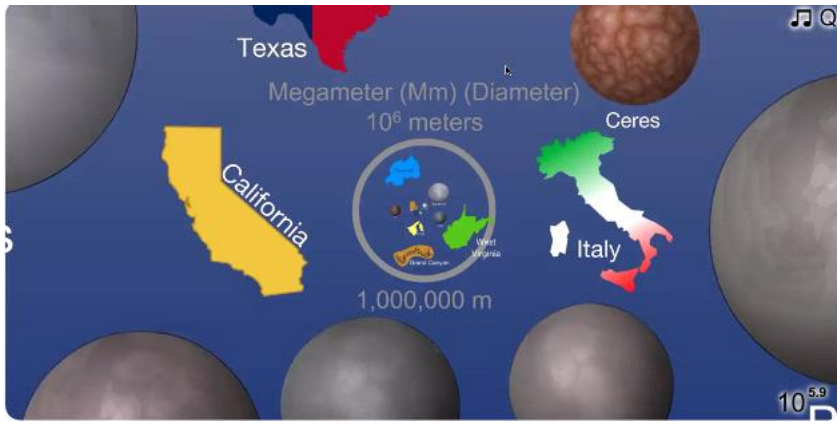
Ritaglio schermata acquisito: 27/04/2021 12:36



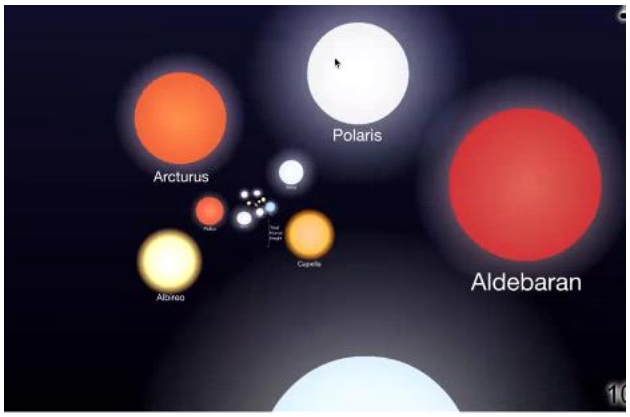
Ritaglio schermata acquisito: 27/04/2021 12:39



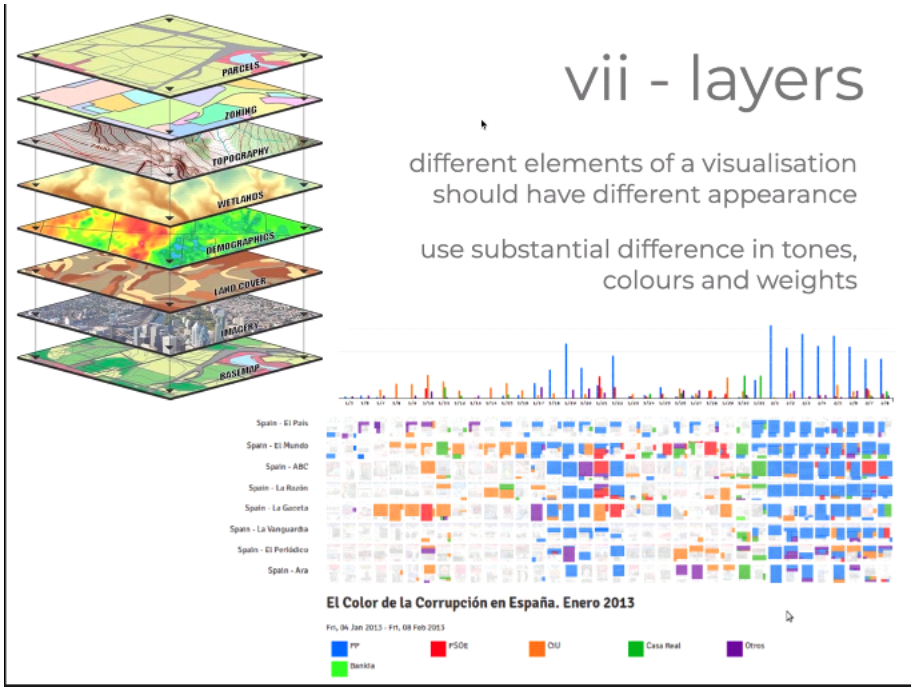
Ritaglio schermata acquisito: 27/04/2021 12:39



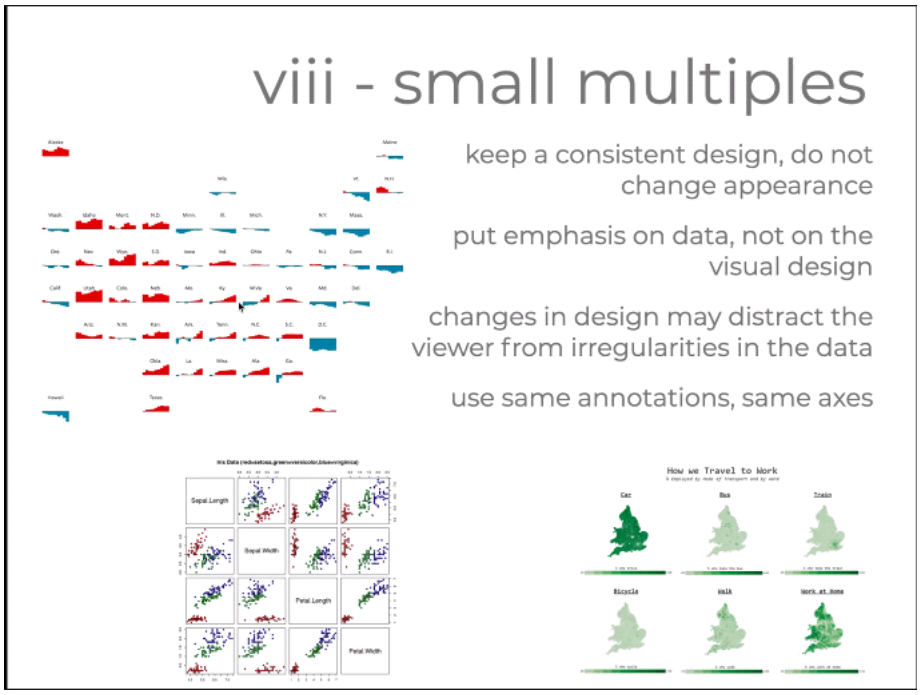
Ritaglio schermata acquisito: 27/04/2021 12:39



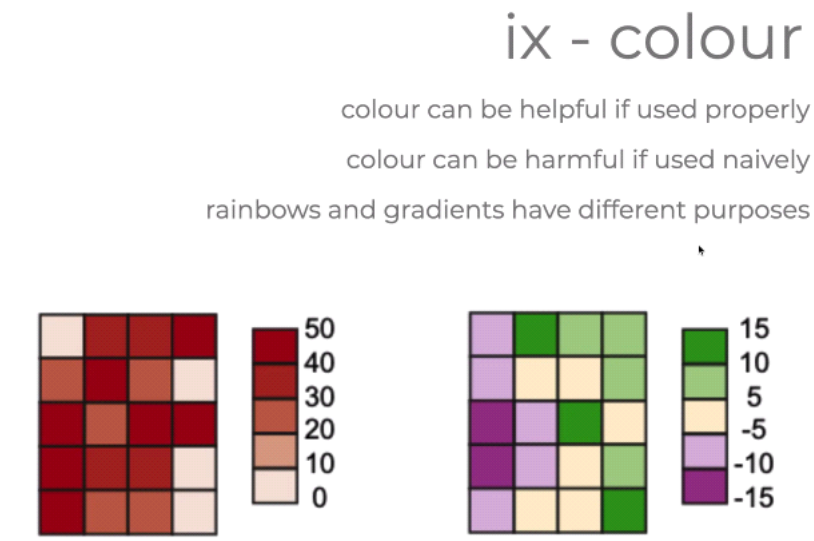
Ritaglio schermata acquisito: 27/04/2021 12:39



Ritaglio schermata acquisito: 27/04/2021 12:42

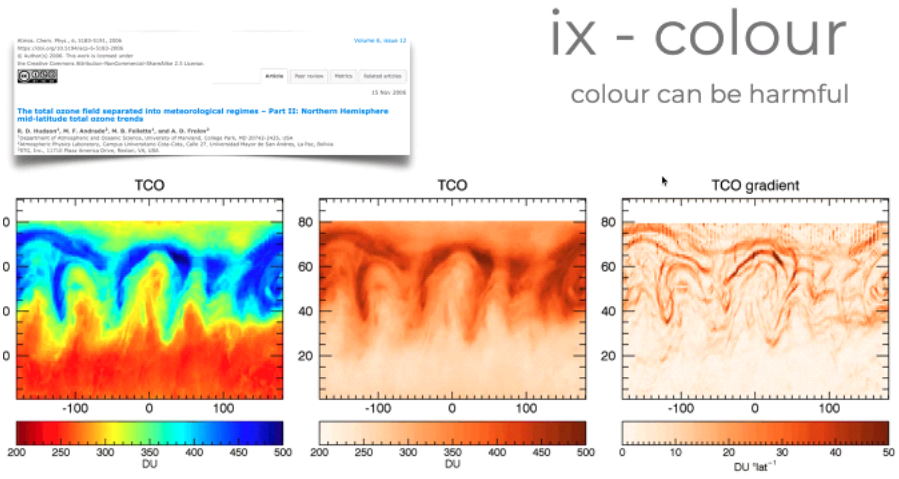


Ritaglio schermata acquisito: 27/04/2021 12:45



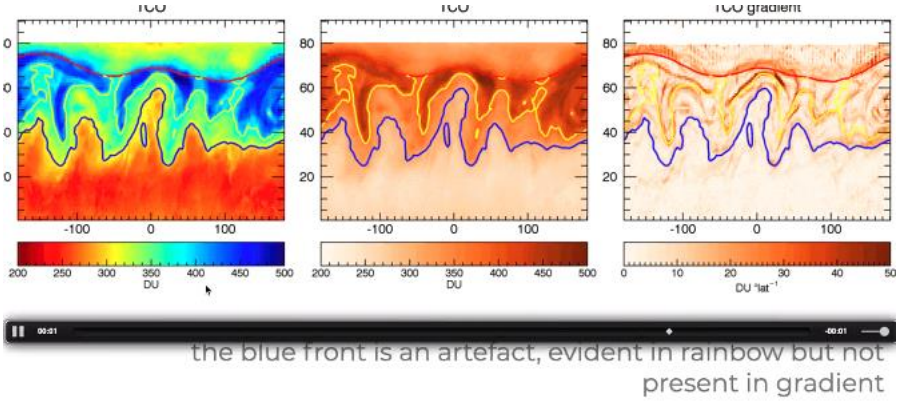
Ritaglio schermata acquisito: 27/04/2021 12:48

Why color can be armful?



Ritaglio schermata acquisito: 27/04/2021 12:49

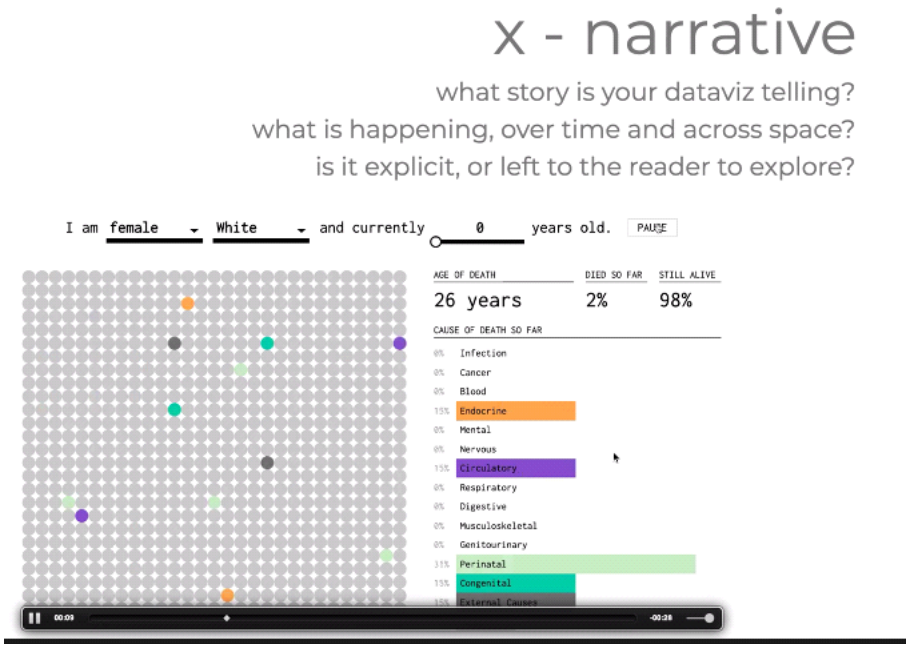
What changes between the three graph is the palettes.



Ritaglio schermata acquisito: 27/04/2021 12:51

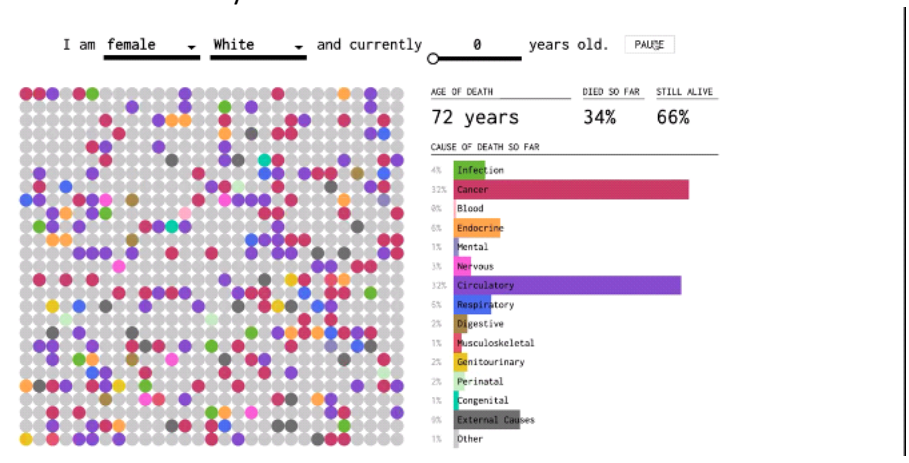
Ritaglio schermata acquisito: 27/04/2021 12:52

The first two map hide the frequency of data at the bottom.



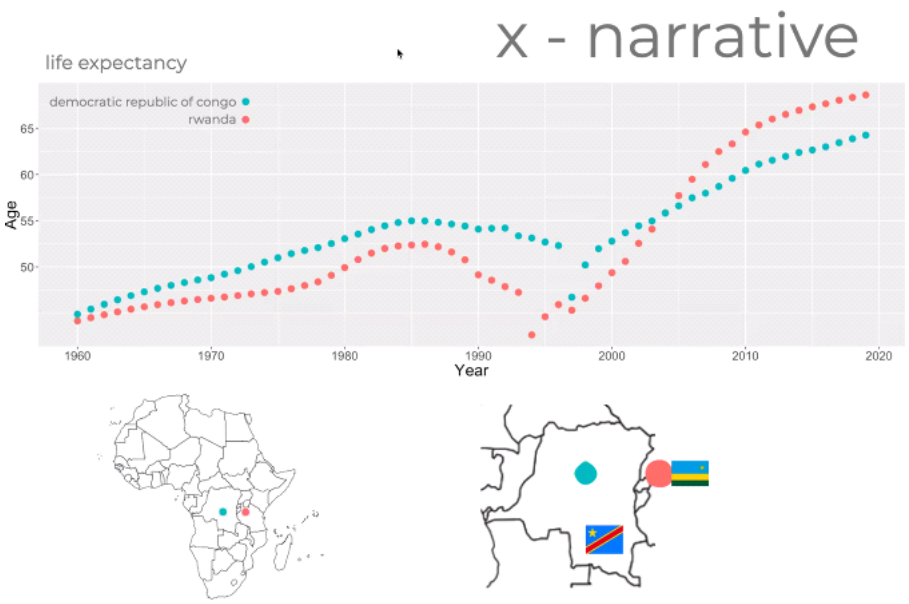
Ritaglio schermata acquisito: 27/04/2021 12:54

Narative is the story connected to the visualization.



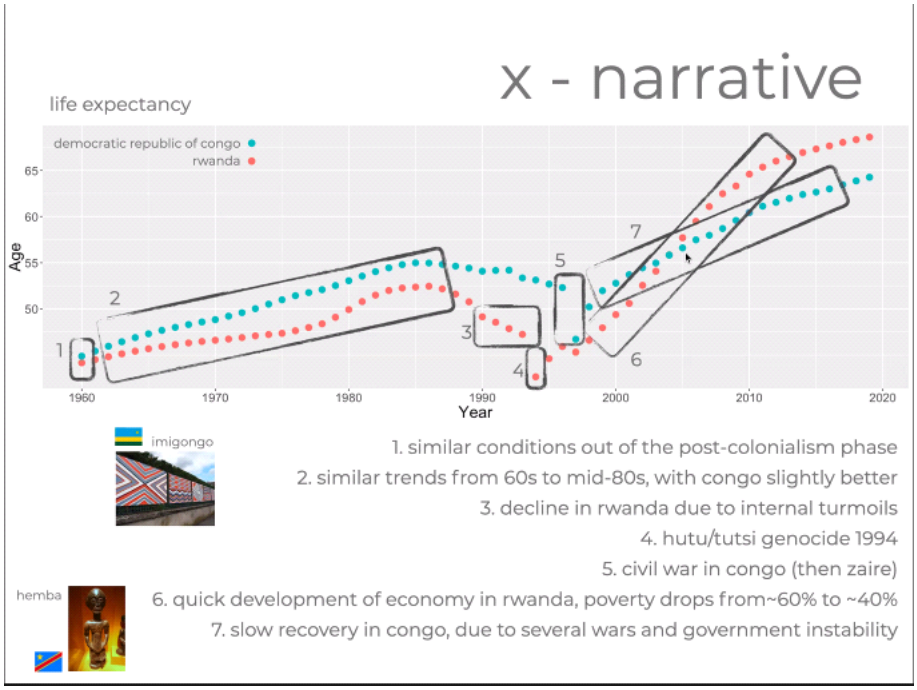
Ritaglio schermata acquisito: 27/04/2021 12:54

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

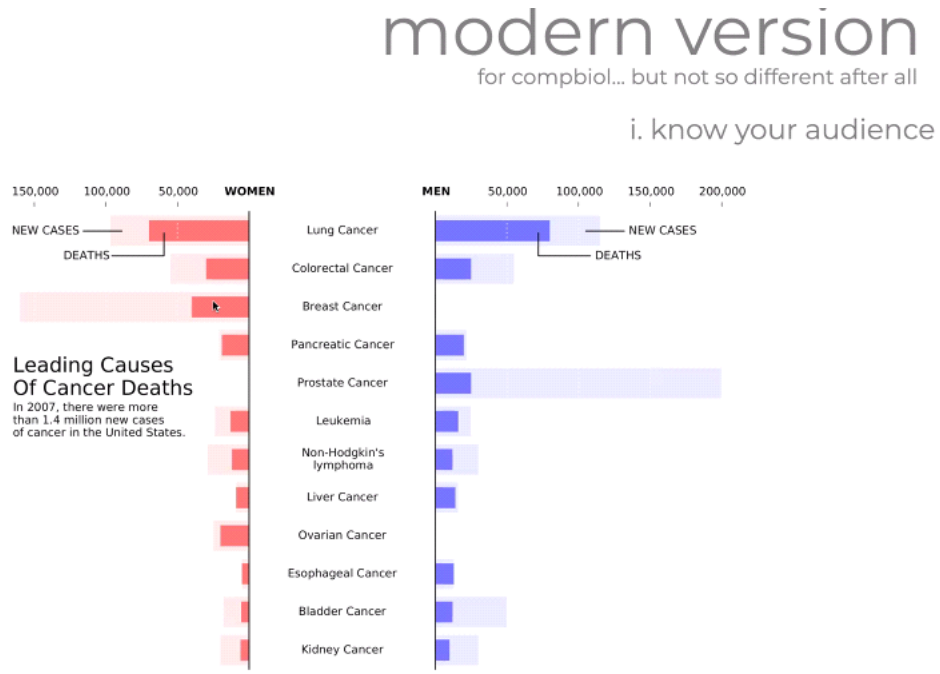
Ten rules for data visualization by Tufte are old but still used.

However, there are some new version of these rules.
There is a paper: 10 simple rules for better figures:

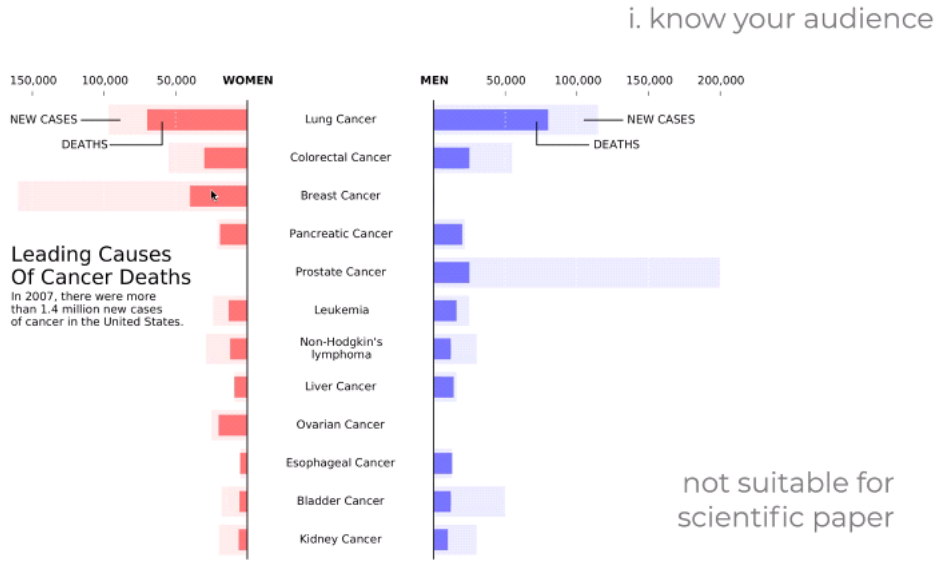
1) **Know your audience**. -> providing data viz is not an absolute task.
2) **Identify** your **message** -> your message can be very simple or very complex depending on the audience.
3) **Adapt the figure** to the support medium. Several journals have both the printed and the paper

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 reade**r. Do not disvtract 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.



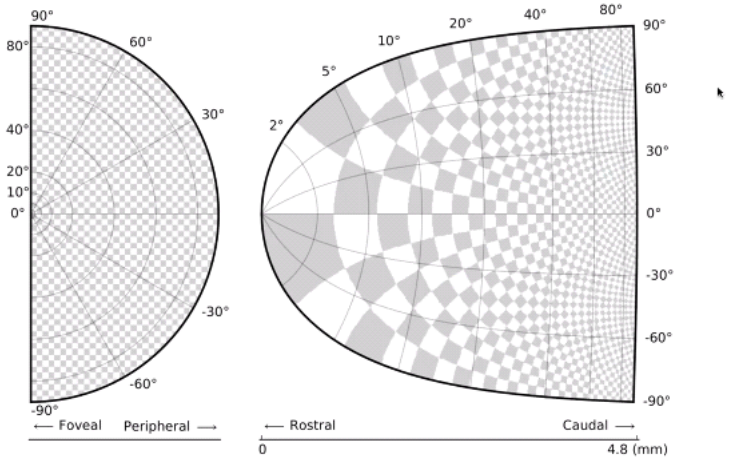
Ritaglio schermata acquisito: 28/04/2021 11:40



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



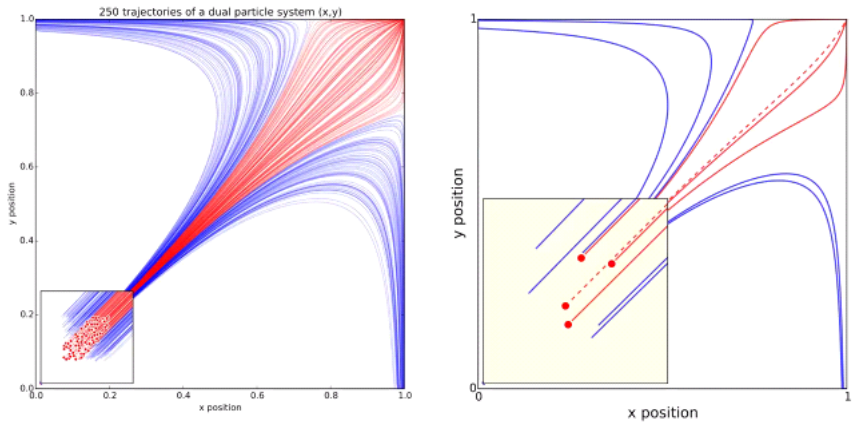
Ritaglio schermata acquisito: 28/04/2021 11:43

Here we have a dataviz, a scientific plot regarding scientiic properties of eyey sight. The interesting thing here is that the author put a very iimportant details -> drives the attention of the reader to the most important concept.
The checkerboard pattern highlights the logarithmic mapping involved in the model.

modern version

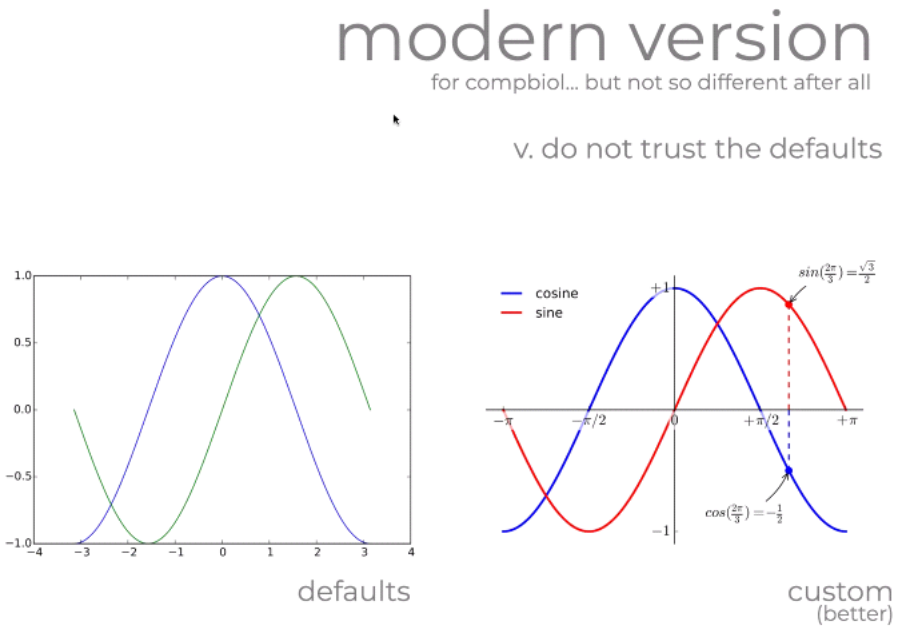
for compbiol... but not so different after all

iii. adapt the figure to the support medium



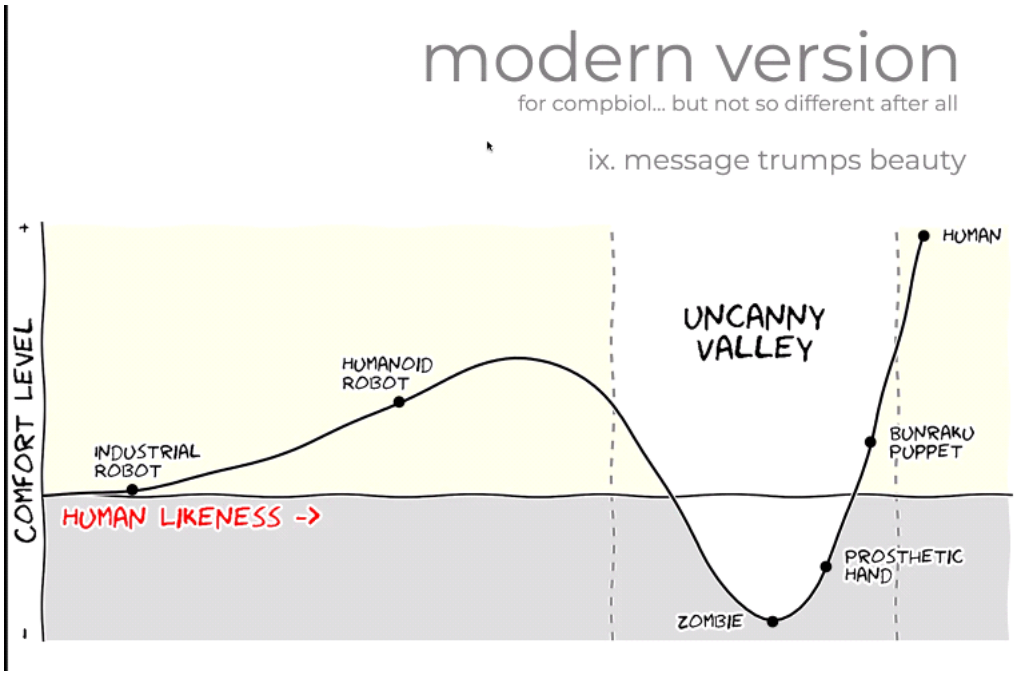
Ritaglio schermata acquisito: 28/04/2021 11:46

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 convay 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 preseting. You have to be impactive, efficient.



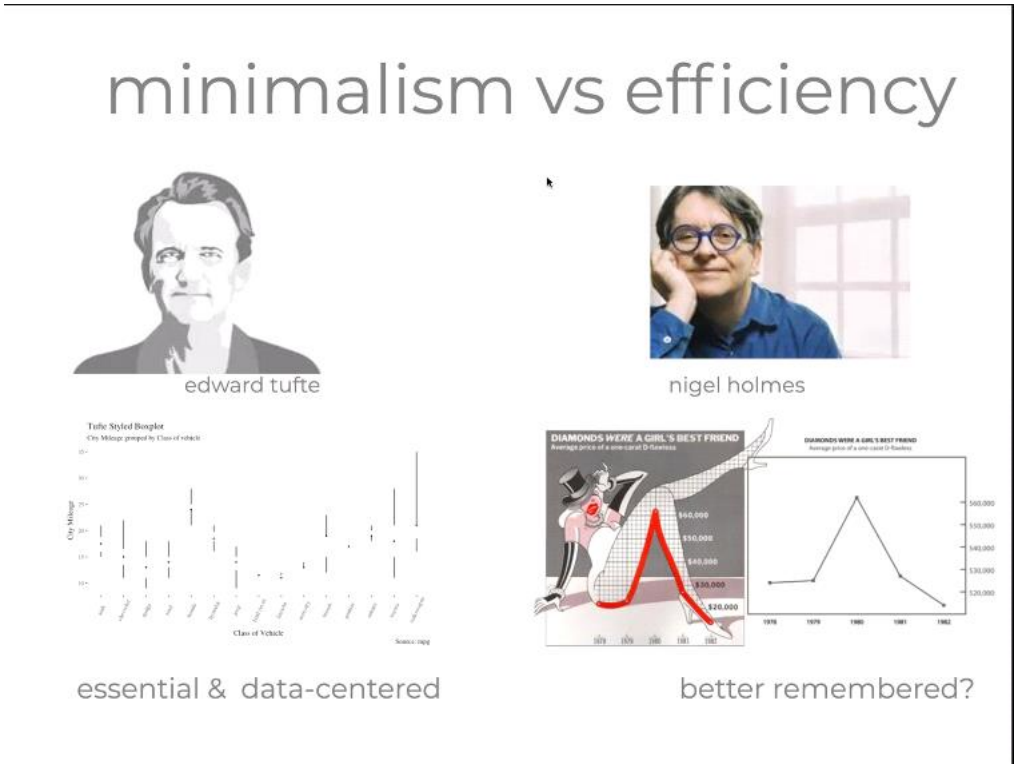
Ritaglio schermata acquisito: 28/04/2021 11:51

Customize all the details. We have a lenged, centered axes with the right labels, a focus on points with the same ascissa.



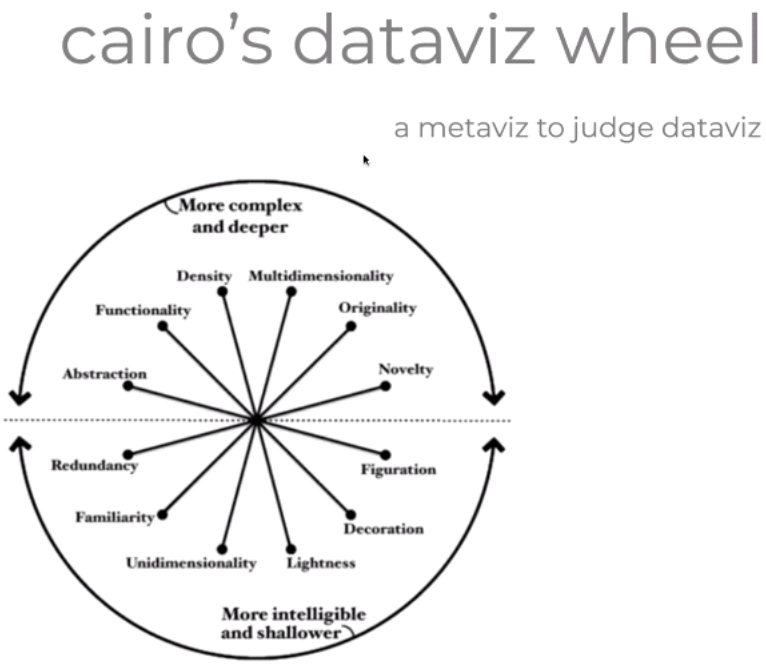
Ritaglio schermata acquisito: 28/04/2021 11:55

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.



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.

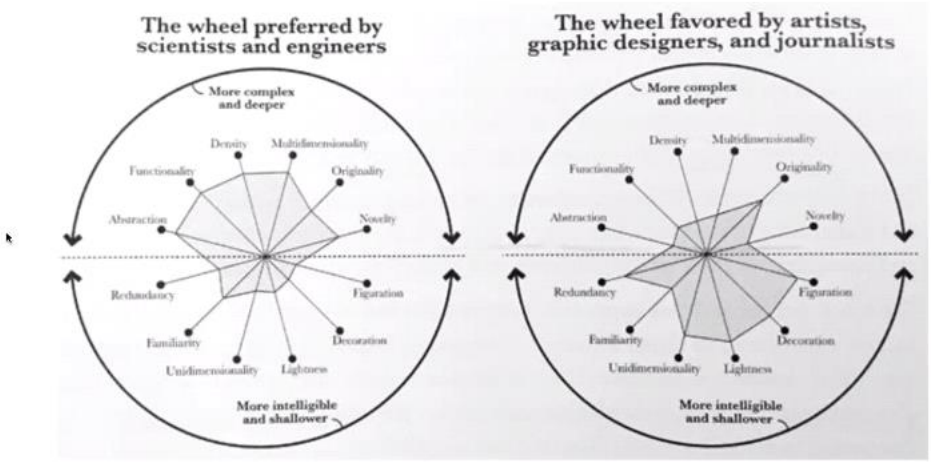


Ritaglio schermata acquisito: 28/04/2021 12:08

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 -> hte 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

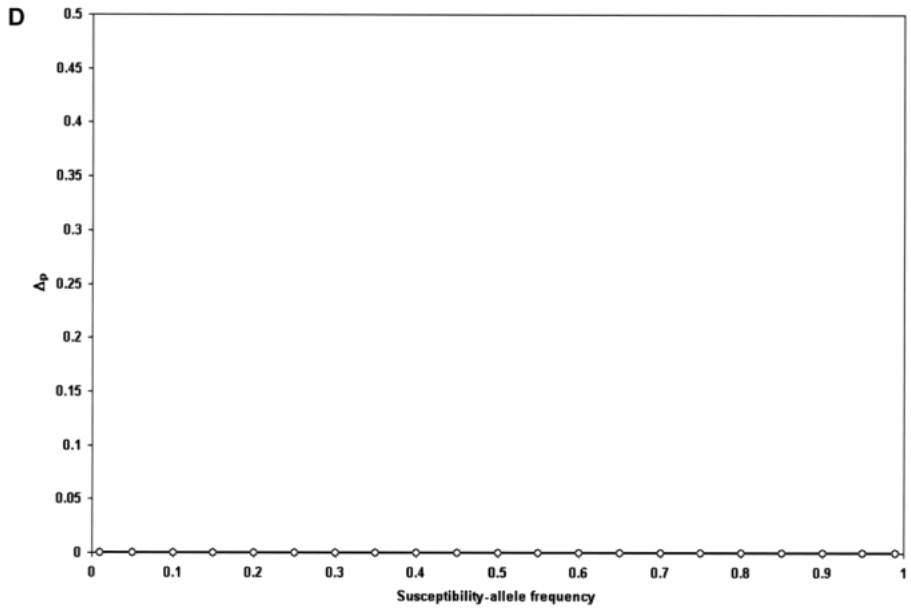
Originality -> new way to shape your data, something that is not seen so far
Novelty -> use the same data viz to highlight different concepts. Not all elements in data viz are novel.
Horror gallery ?
Very bad representation that made their way to pubblication
Rational inferences about departures hardy-weinberg equilibrium. This is an example of paper with terrible data viz.

Figure 1 Δ_γ 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

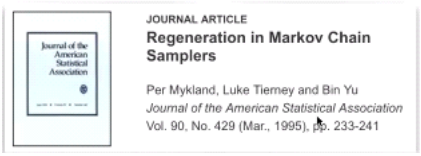
This is the caption. -> we have many gliphs and colors

horror gallery



Ritaglio schermata acquisito: 28/04/2021 12:20

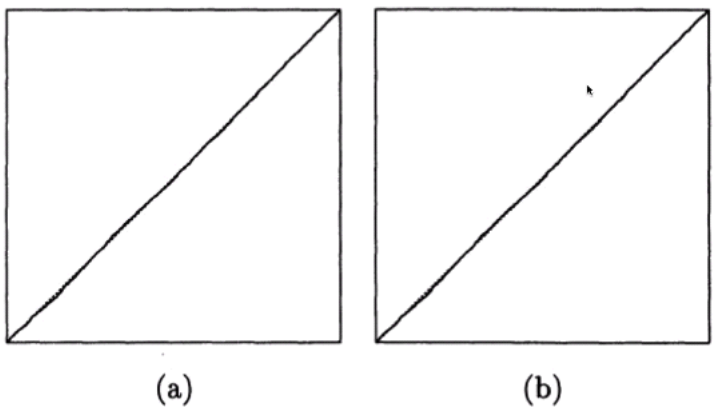
We have one single set of point on the x-axes. This was used to point 10 zeros and was used a whole graph to do it.



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.

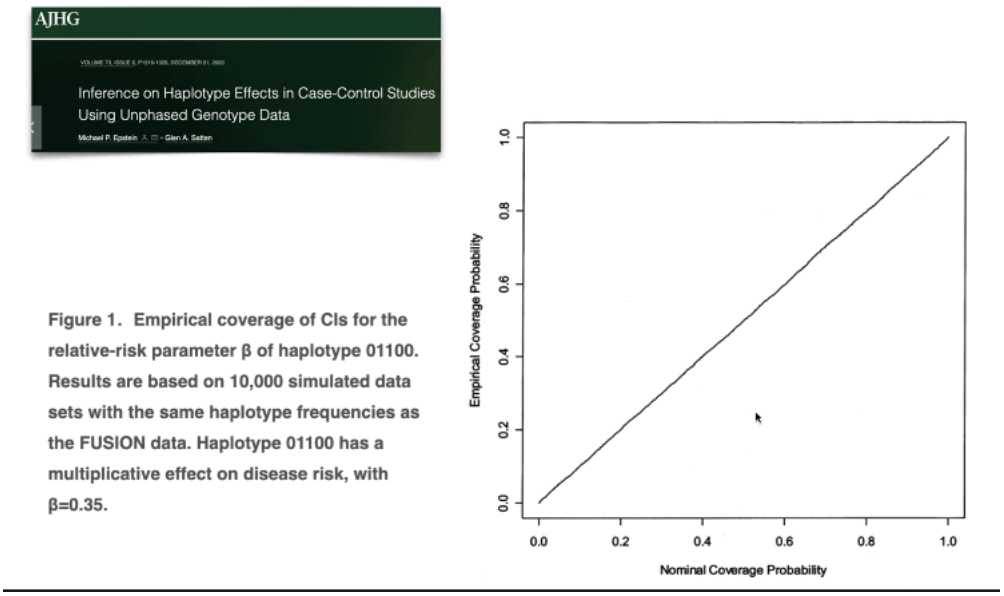
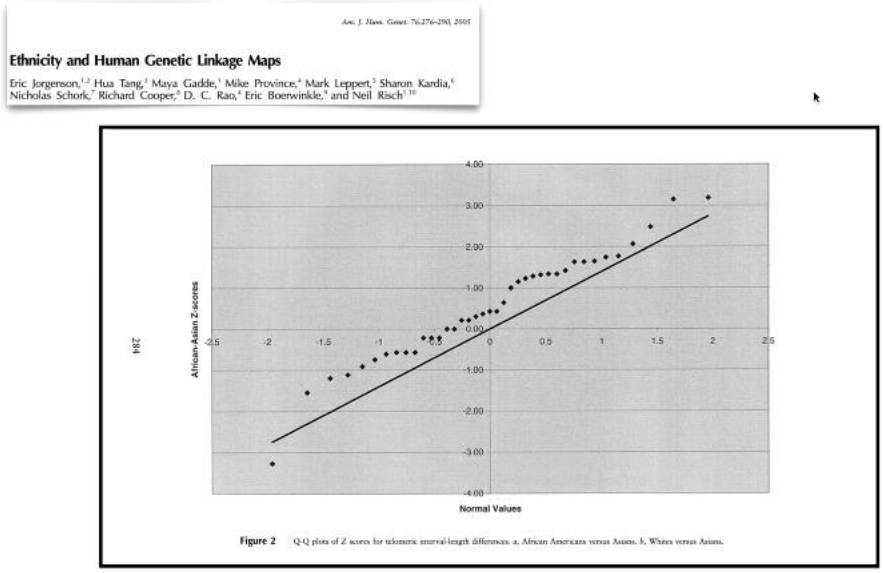


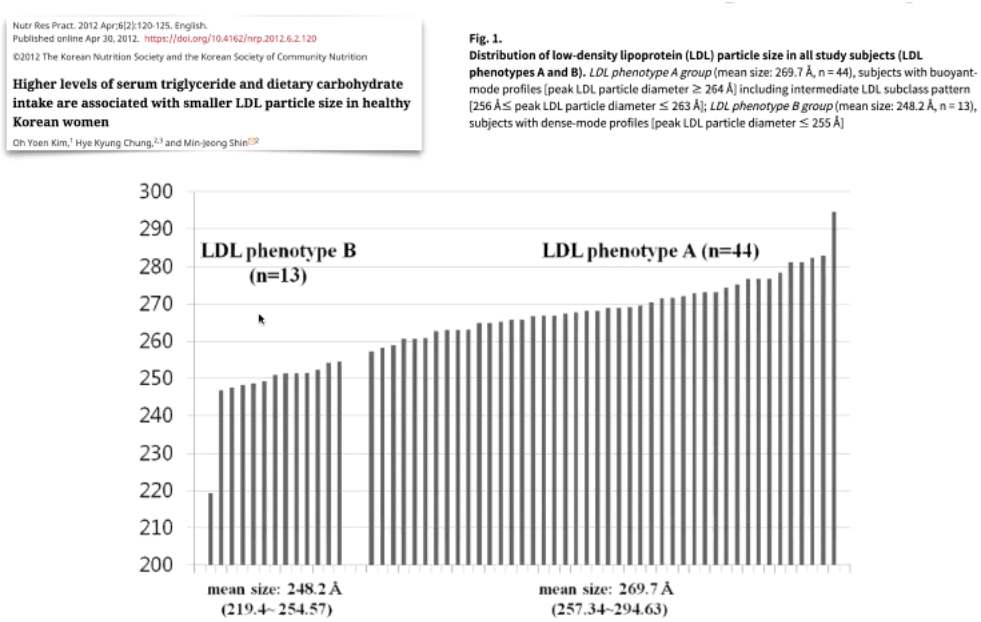
Figure 1. Empirical coverage of CIs for the relative-risk parameter β of haplotype 01100. Results are based on 10,000 simulated data sets with the same haplotype frequencies as the FUSION data. Haplotype 01100 has a multiplicative effect on disease risk, with β=0.35.

Ritaglio schermata acquisito: 28/04/2021 12:26

Probably a confidence interval would have helped to make this representation more meaningful.

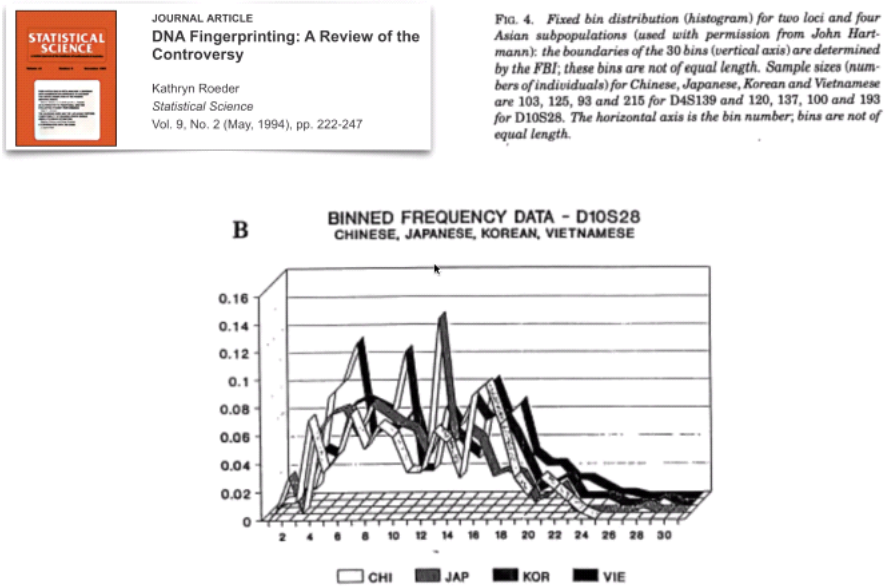


Ritaglio schermata acquisito: 28/04/2021 12:30



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 redability.
We have the real numbers and ranges.



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.

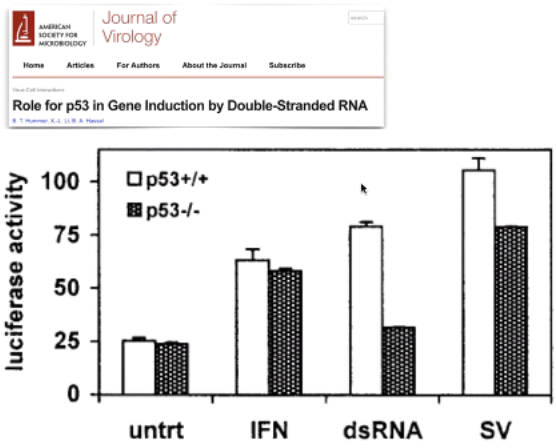


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA 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-α/ml, 50 μg 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 μl 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 ± 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.

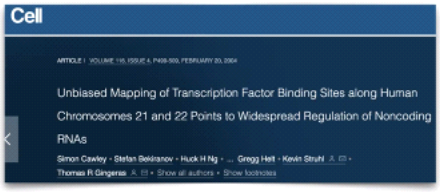
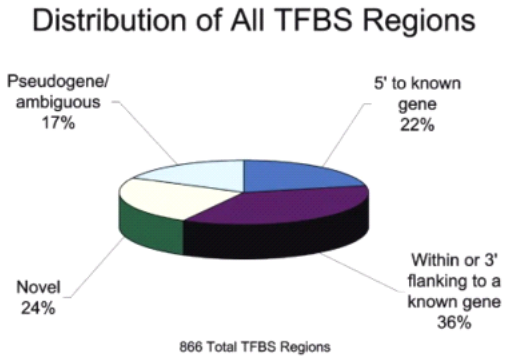


Figure 1. Classification of TFBS Regions

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 c 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).



Ritaglio schermata acquisito: 28/04/2021 12:51

horror gallery

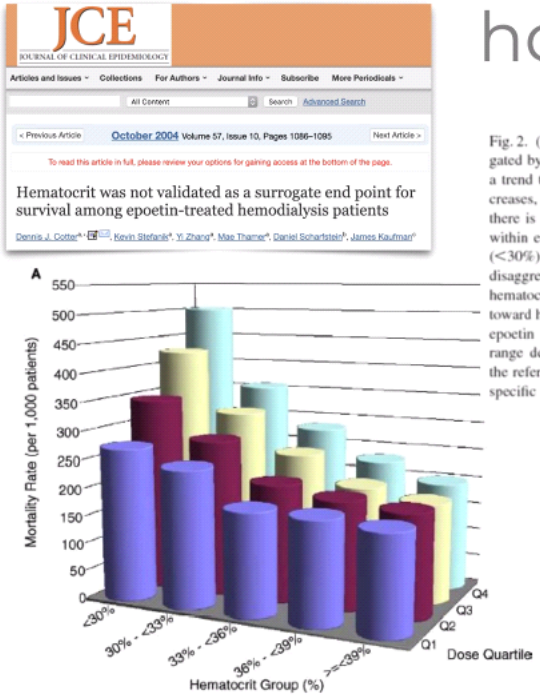


Fig. 2. (A) Unadjusted 1-year mortality rates by hematocrit group disaggregated by epoetin dose quartile. Within each epoetin dose quartile, there is a trend toward increasing mortality as the observed study hematocrit decreases, most notably in the fourth quartile (>21,692 units/wk). Similarly, there is a trend toward increasing mortality as the epoetin dose increases within each observed study hematocrit range, most notably in the lowest (<30%) hematocrit range. (B) Relative risk of death by hematocrit group disaggregated by epoetin dose quartile. For the three lowest observed study hematocrit ranges, compared with the reference group, there is a trend toward higher relative risk of mortality within each hematocrit range as the epoetin dose increases and within each dose quartile as the hematocrit range decreases. For the two highest hematocrit ranges, compared with the reference group, the relative risk of mortality varies, depending on the specific hematocrit range and dose quartile.

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.

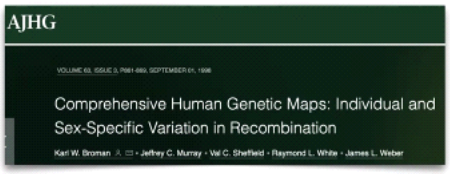
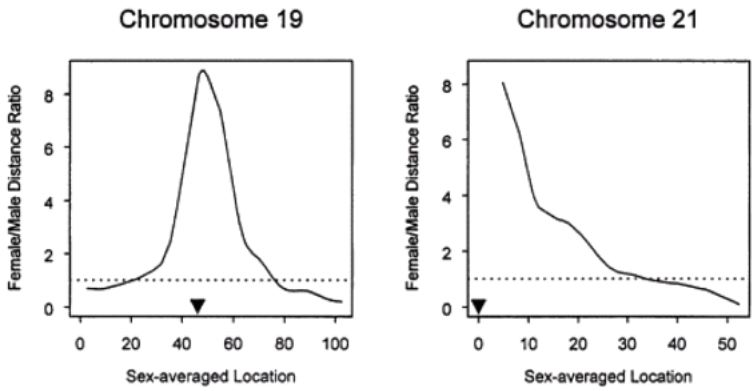
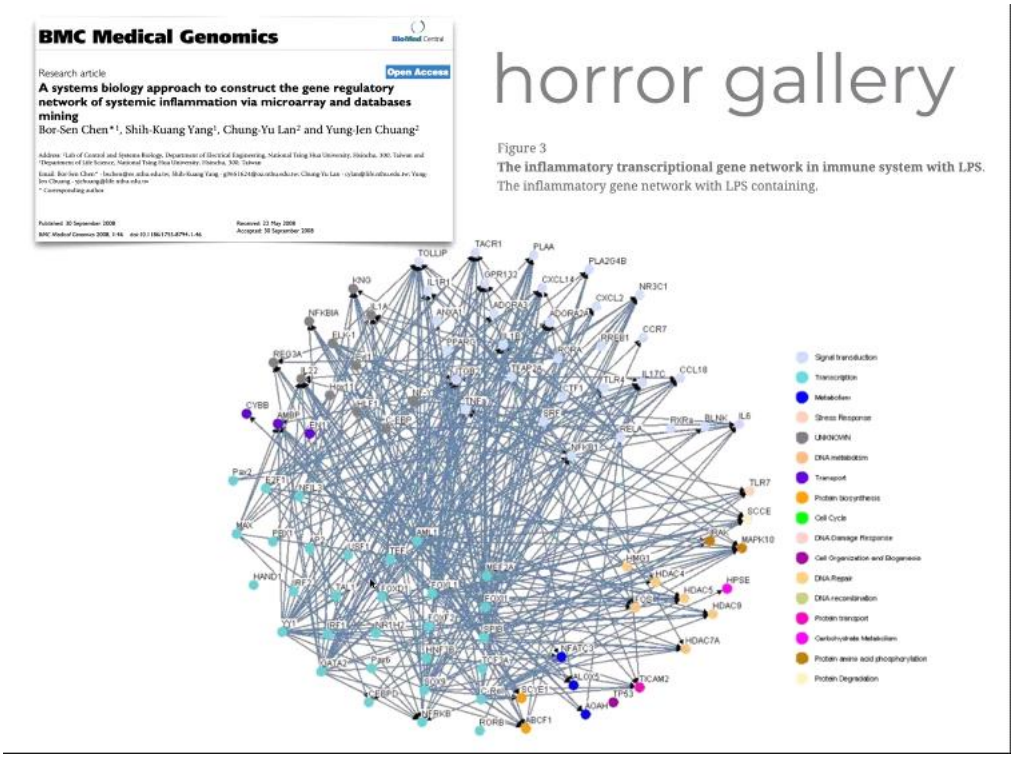


Figure 1. Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

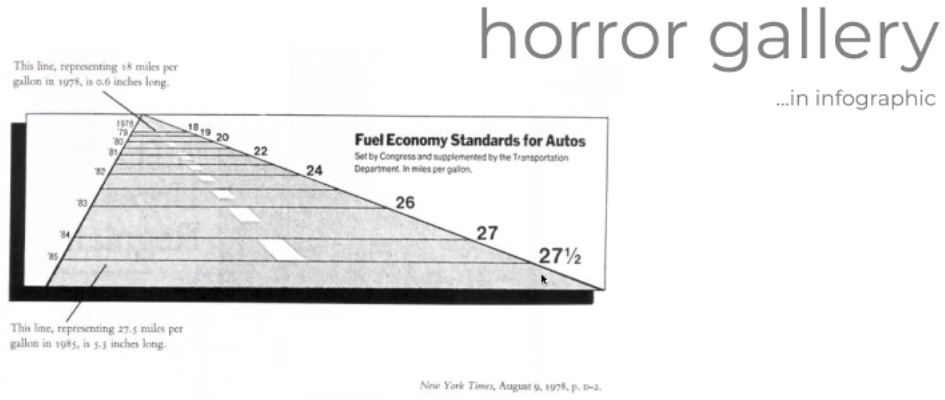


Ritaglio schermata acquisito: 28/04/2021 12:53

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.

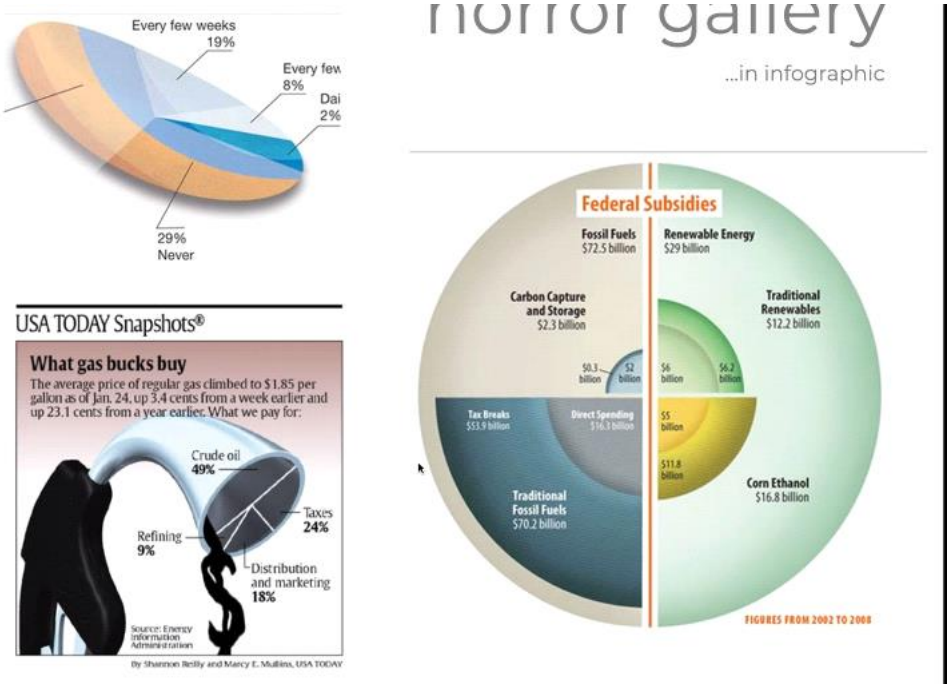


Ritaglio schermata acquisito: 28/04/2021 13:01

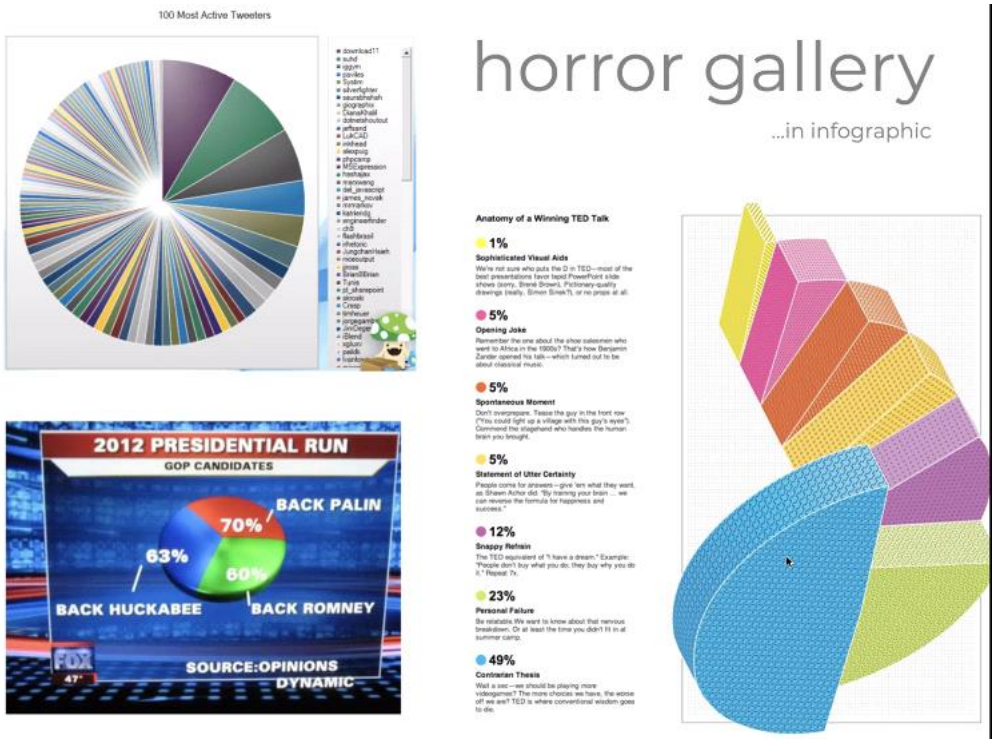


Ritaglio schermata acquisito: 28/04/2021 13:03

Numbers of miles that can be runned with a gallon.
Perspective exagerate the effect



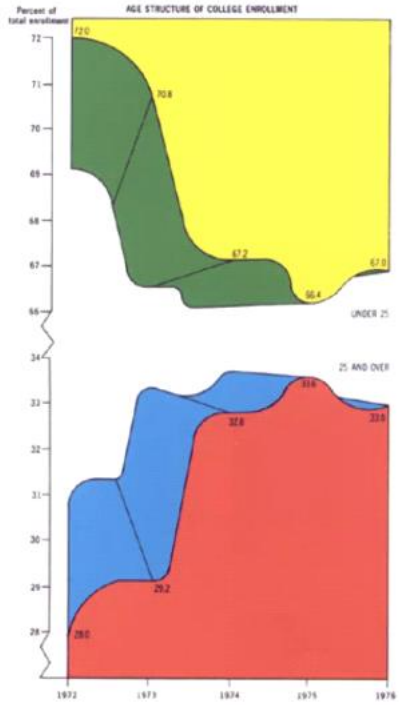
Ritaglio schermata acquisito: 28/04/2021 13:04



Ritaglio schermata acquisito: 28/04/2021 13:04

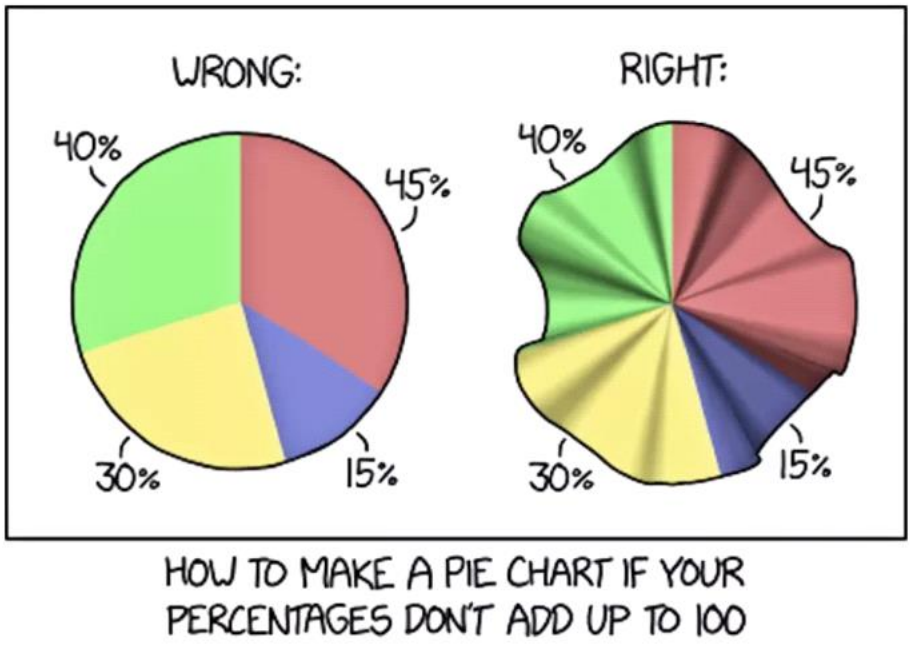
Ritaglio schermata acquisito: 28/04/2021 13:06

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

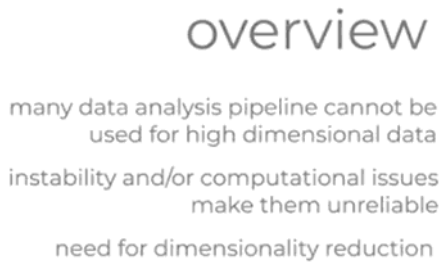
Age structure of collage enrollment. The two groups are complementary. They just show a percentages. They collected just 5 points.
There is also a break in the axes making the plot out of scale.



Ritaglio schermata acquisito: 28/04/2021 13:12

Dimensionality Reduction

mercoledì 5 maggio 2021 16:42



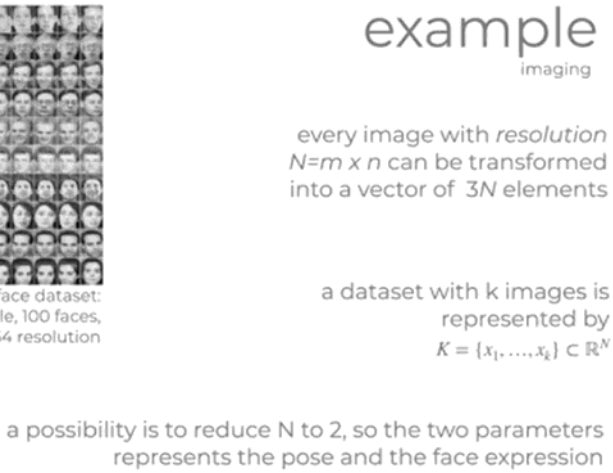
Ritaglio schermata acquisito: 05/05/2021 16:42

We are talking about images, or text processing, hyperspectral imaging, omics and fluid dinamics.

Ritaglio schermata acquisito: 05/05/2021 16:44

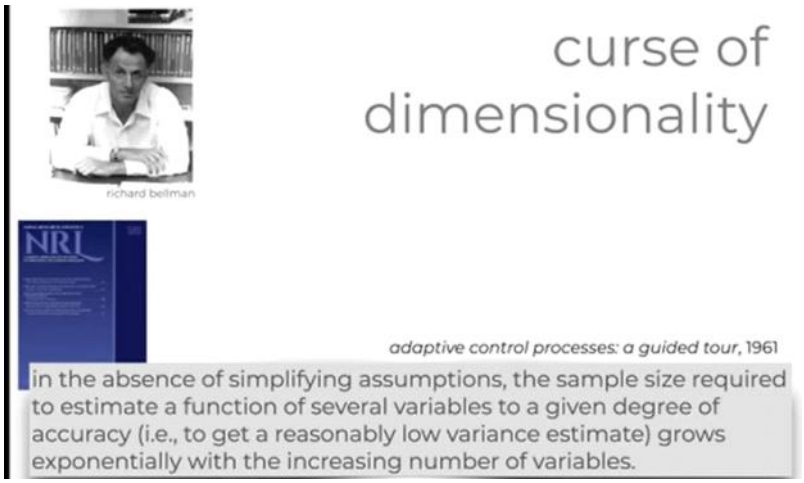
Dr -> dimensionality reduction algorithm.

olivetti face dataset:
10 people, 100 faces,
64x64 resolution

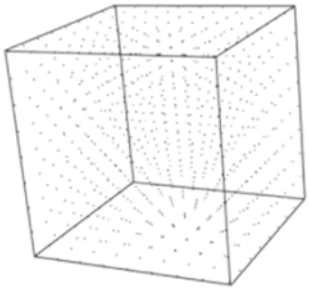


Ritaglio schermata acquisito: 05/05/2021 16:45

Pixel resolution described with RGB vector.



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 dimensionaly has very aspects -> an important one is the **empty space phenomenon** which states that:
High-dim spaces are inherently sparse.

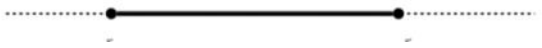
in the unit 9-hypersphere S^9_1 in $\mathbb{R}^{10}, P(d(x,0) \leq 0.9) \sim 0.35$

only 0.02% of the mass of N(0,1) in $\mathbb{R}^{10}$ is in S^9_1

Ritaglio schermata acquisito: 05/05/2021 16:51

(hyper)-cubes & -spheres
of diameter/side 2r

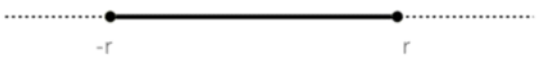
D=1



Ritaglio schermata acquisito: 05/05/2021 16:51

(hyper)-cubes & -spheres
of diameter/side 2r

D=1



$V_S^D = 2r \qquad V_C^D = 2r$

Hyper spheres -> all the points in the space whose distance from the origin is in absolute value less than one.

(hyper)-cubes & -spheres

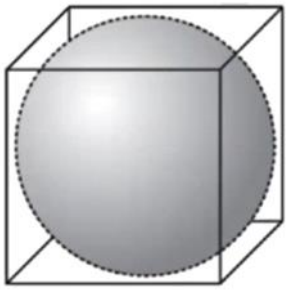
of diameter/side 2r

D=2

Ritaglio schermata acquisito: 05/05/2021 16:52

$$V_S^D = \pi r^2 \qquad V_C^D = 4r^2$$

Ritaglio schermata acquisito: 05/05/2021 16:52



D=3

$$V_S^D = \frac{4}{3}\pi r^3 \qquad V_C^D = 8r^3$$

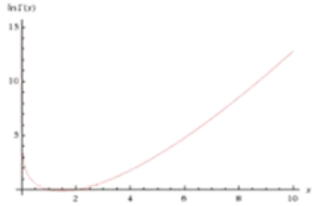
Ritaglio schermata acquisito: 05/05/2021 16:53

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$$

$$\Gamma(n) = (n-1)!$$

$$\Gamma\left(n+\frac{1}{2}\right) = \binom{(n-\frac{1}{2})}{n} n!\sqrt{\pi}$$

D



$$V_s^D = \frac{\sqrt{\pi^D} r^D}{\Gamma\left(\frac{D}{2}+1\right)}$$

$$V_c^D = (2r)^D$$

Ritaglio schermata acquisito: 05/05/2021 16:56

Gamma function -> analytical extension of the factorial

$$\lim_{D\to\infty} \frac{V_s^D}{V_c^D} = \lim_{D\to\infty} \frac{\frac{\sqrt{\pi^D} r^D}{\Gamma\left(\frac{D}{2}+1\right)}}{(2r)^D} = 0$$

Ritaglio schermata acquisito: 05/05/2021 16:56

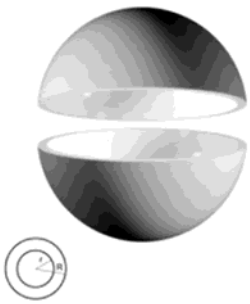
This means that when we have a cube inclosing the sphere the ration between the cube and the sphere is zero -> the vlolume of the spaheres is much smaller than the one of the cube.

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

$$V_{ss}^D = V_{s_R}^D - V_{s_r}^D$$



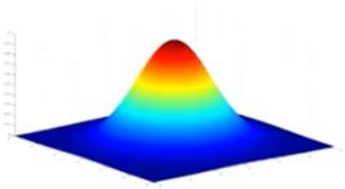
$$\lim_{D\to\infty} \frac{V_{ss}^D}{V_{s_R}^D} = \lim_{D\to\infty} 1 - \frac{V_{s_r}^D}{V_{s_r}^D} = \lim_{D\to\infty} 1 - \left(\frac{r}{R}\right)^D = 1$$

Ritaglio schermata acquisito: 05/05/2021 17:00

As far as the dimension grows all the volume of the sphera 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

$N_D(0,1)$



pdf $G(x) = K(r) = \frac{1}{\sqrt{(2\pi)^D}} e^{-\frac{r^2}{2}}, ||x|| = r$

equiprobable contours

$\partial B_r = \partial\{x \in \mathbb{R}^D, ||x|| \leq r\} = S_r$

$P(x \in B_r) = \frac{\int_0^r V(S_r)K(r)dr}{\int_0^\infty V(S_r)K(r)dr}$

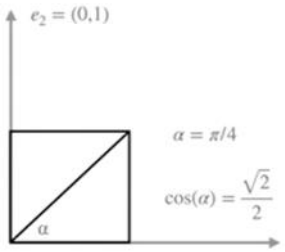
n	1	2	5	10	20	100
Probability	0.04550	0.13534	0.54942	0.94734	0.99995	1.00000

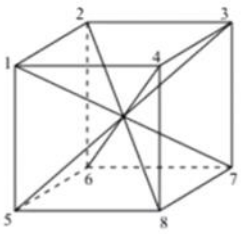
r>2

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

$d = (\pm 1, \pm 1, \ldots, \pm 1)$

$e_i = (\delta_{ij})_{j=1}^D$

$\cos(\alpha) = \left\langle \frac{d}{||d||}, e_i \right\rangle = \frac{\pm 1}{\sqrt{D}} \to 0 \text{ for } D \to \infty$

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 -> D ei are orthogonal. When the dimension grows all the diagonal are nearly orthogonal to all cordinatex axes.

concentration of norms

Theorem 1.1. *Let $\boldsymbol{x} = [x_1, \cdots, x_D]^T \in \mathbb{R}^D$ be a random vector whose components $x_k, 1 \leqslant k \leqslant D$, 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 $\boldsymbol{x}$, and $\rho_j = E\{(x_k - \mu)^j\}$ be their common central j^{th} moment. Then the mean $\mu_{\|\boldsymbol{x}\|}$ and the variance $\sigma^2_{\|\boldsymbol{x}\|}$ of the Euclidean norm of $\boldsymbol{x}$ are*

$$\mu_{\|\boldsymbol{x}\|} = \sqrt{aD - b} + O(D^{-1}),$$
$$\sigma^2_{\|\boldsymbol{x}\|} = b + O(D^{-1/2}), \quad (1.4)$$

respectively, where a and b are the parameters defined by

$$a = \rho^2 + \mu_2,$$
$$b = \frac{4\mu^2\mu_2 - \mu_2^2 + 4\mu\mu_3 + \mu_4}{4(\mu^2 + \mu_2)}$$

[demartines, 1994]

means grow as $O(D^{0.5})$,
variances grow as $O(1)$

⇩

random vectors are nearly normalized

$P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}$ ⇨ $P(|\,||x|| - \mu_{||x||}\,| \geq \varepsilon) \leq \frac{\sigma^2_{||x||}}{\varepsilon^2}$ independent of D

Chebyshev's inequality

for large D $||x|| \sim \mu_{||x||} = r$ ⇨ 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

$$X = \{x_a\}_{a\in A}, x_a \in M,\ \dim(M) = s,\ M \subset \mathbb{R}^D$$

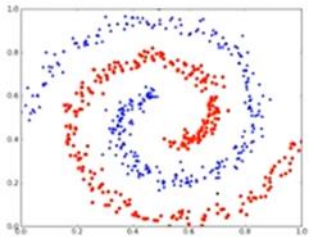
D extrinsic dimension of X, s intrinsic dimension of X

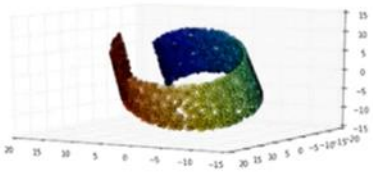
X sample set of random vectors **X** in dimension D

there exist a random vector **Y** in dim s and an invertible analytic function *f* s.t. *f*(**Y**)=**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

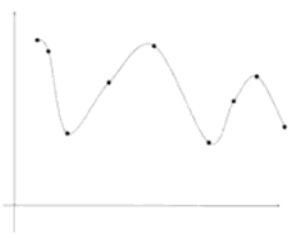
dr is finding a parameterization of the manifold which the points of data reside on

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 $T : \mathbb{R}^s \to S$

$$x = T(y) = Uy + x_0$$

where U is orthogonal (<u,v>=<Uu,Uv>) 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,...,u_s], then {u_1,...,u_s} is an orthonormal basis of S

Ritaglio schermata acquisito: 05/05/2021 17:16

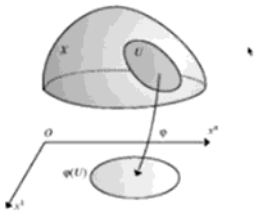
Orthonormal -> each couple of columns is orthogonal and their products is 1.
Distances are preserved and we have an example of linear dimensionality reduction.

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 $\mathbf{R}^D$ has dimension s<D, then the neighbourhood of each point of M is isomorphic to $\mathbf{R}^s$, i.e. there is an invertible differentiable map from M to $\mathbf{R}^s$ whose inverse if 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 space
set V of elements called vectors close under vector sum and scalar multiplication for elements in **R**
basic example: $\mathbf{R}^n$

linear combination
$a_1 v_1 + a_2 v_2 + \ldots + a_n v_n$ for $a_i \in \mathbf{R}$ and $v_i \in V$

linear independent vectors
set $M=\{m_1,\ldots,m_q\}$ of elements of V such that no vector in M can be written as linear combination of the others

generators
set $M=\{m_1,\ldots,m_p\}$ of elements of V such that each element of V can be written as a linear combination of vectors of M

basis
set $B=\{b_1,\ldots,b_n\}$ of linearly independent generators of V

real vector spaces

dimension
cardinality of all the bases of V

example
$V=\mathbf{R}^3$ — basis $E=\{e_1=(1,0,0), e_2=(0,1,0), e_3=(0,0,1)\}$
or basis $G=\{g1=(1,2,3), g2=(-1,3,0), g3=(0,1,-1)\}$

coordinates
each vector in V can be written in a unique ay 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 $T:V\to$ such that $T(u+v)=T(u)+T(v)$ and $T(kv)=k\,T(v)$ for $u,v \in V$ and $k \in \mathbf{R}$

Ritaglio schermata acquisito: 05/05/2021 17:34

matrices

basis change
$B_1=\{v_1,\ldots,v_n\}$, $B_2=\{w_1,\ldots,w_n\}$ bases of V

$$w_j = \sum_{i=1}^{n} a_{ij} v_i \quad \text{if } v = \sum_{i=1}^{n} x_i v_i = \sum_{i=1}^{n} y_i w_i \qquad X = AY \qquad A = (a_{ij})$$

linear transformation
$T:V\to$ with $B_V=\{v_1,\ldots,v_n\}$, $B=\{w_1,\ldots,w_m\}$

$$T(v_j) = \sum_{i=1}^{n} a_{i,j} w_j$$

then T is represented by the matrix $A=(a_{ij})$ in $M_{\mathbf{R}}(m,n)=\mathbf{R}^{mxn}$
different bases correspond to different A

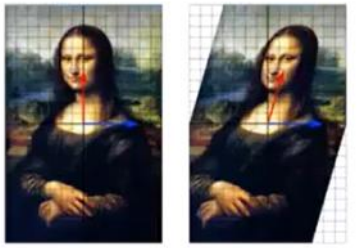
linear transformation
$T:V\to W$
$B_V=\{v_1,\ldots,v_n\}$, $B=\{w_1,\ldots,w_m\}$, $B'_V=\{v'_1,\ldots,v'_n\}$, $B'=\{w'_1,\ldots,w'_m\}$
P matrix $B_V\to B'_V$, Q matrix $B_W\to B_{W'}$
$A' = Q^{-1} A P$

Ritaglio schermata acquisito: 05/05/2021 17:36

eigen*

linear transformation
T:V→V
if T(v)=λv, v is an eigenvector and λ is an eigenvalue;
as matrices, (A-λ**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+αy,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:VxV→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)=($\mathbf{R}^n$, ·)
v · u =0 <=> cos(θ)=0

Ritaglio schermata acquisito: 05/05/2021 17:39

orthogonality

inner product
a bilinear form <,> on a real vector space V that is symmetric
<x,y>=<y,x>
and positive definite <x,x>>0 if x≠0

orthogonal transformation
T:V→V linear on a inner product space with <x,y>=<T(x),T(y)>

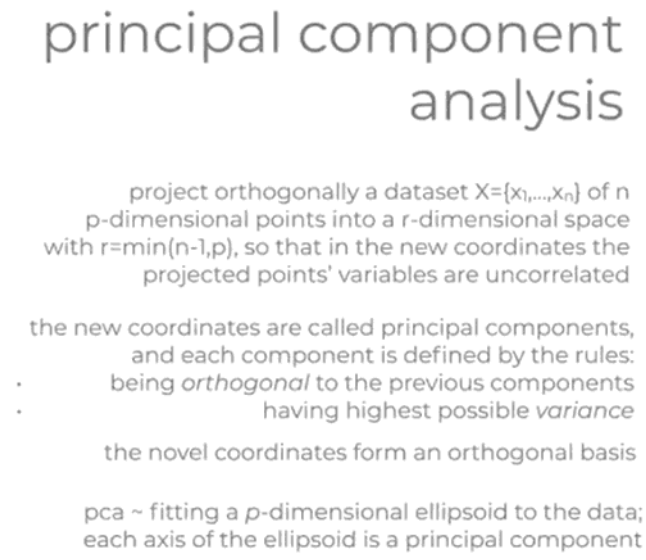
dot product
in euclidean space, <,> is · and then u · v = uv^T

norm
$||w|| = \sqrt{< w, w >}$; for $\cdot, ||w|| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$

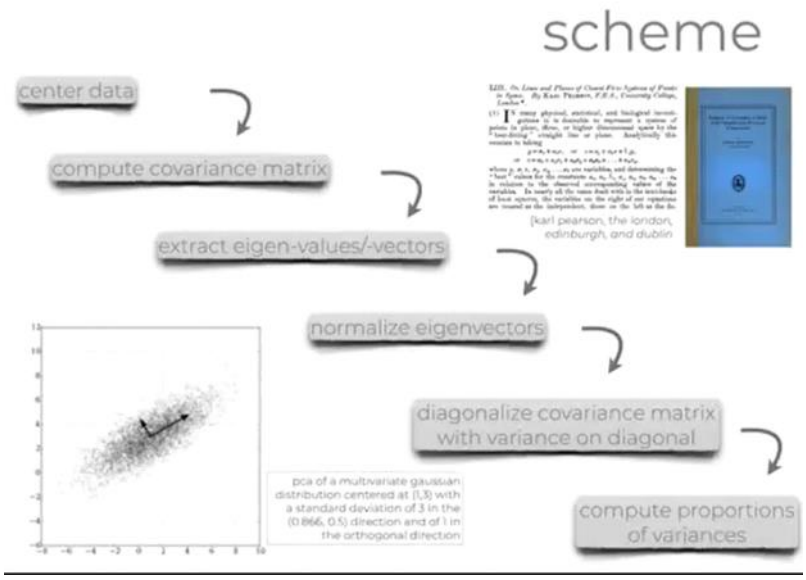
orthogonal basis
a set of orthogonal linearly independent vectors that generate V and are orthogonal;
if ||x||=1 for each element of the basis, they are called *orthonormal*

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.



Ritaglio schermata acquisito: 05/05/2021 17:44



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 $\mathbf{R}^{n\times p}$ — 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 : $t_{ik}=x_{i\cdot}\cdot w_{\cdot k}$

$$\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1l} \\ t_{21} & t_{22} & \cdots & t_{2l} \\ \cdots & & & \\ t_{n1} & t_{n2} & \cdots & t_{nl} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1l} \\ w_{21} & w_{22} & \cdots & w_{2l} \\ \cdots & & & \\ w_{p1} & w_{p2} & \cdots & w_{pl} \end{pmatrix}$$

s.t. $t_1,\dots,t_l$ inherit the maximum possible variance from X and $w_{\cdot k}=(w_{1k},\dots,w_{pk})$ 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

$$\begin{pmatrix} t_{11} \\ t_{21} \\ \cdots \\ t_{n1} \end{pmatrix} = \begin{pmatrix} x_{11} x_{12} \cdots x_{1p} \\ x_{21} x_{22} \cdots x_{2p} \\ \cdots \\ x_{n1} x_{n2} \cdots x_{np} \end{pmatrix} \cdot \begin{pmatrix} w_{11} \\ w_{21} \\ \cdots \\ w_{p1} \end{pmatrix}$$

$$t = Xw$$

$$\mathrm{Var}(t_{\cdot 1}) = \overline{t_{\cdot 1}^2} - (\overline{t_{\cdot 1}})^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} t_{i1}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}w_{j1}\right)^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}w_{j1}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2 - \frac{1}{n^2}\left(\sum_{j=1}^{p} w_{j1}\sum_{i=1}^{n} x_{ij}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} t_{i1}^2$$

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:

$$w_{\cdot 1} = \arg\max_{||w_{\cdot 1}=1||} \sum_{i=1}^{n} t_{i1}^2 = \arg\max_{||w_{\cdot 1}=1||} \sum_{i=1}^{n} (x_{i\cdot}\cdot w_{\cdot 1})^2 = \arg\max_{||w_{\cdot 1}=1||} ||Xw||^2 = \arg\max_{||w_{\cdot 1}=1||} wX^TXw$$

and, since $||w||=1$

$$w_{\cdot 1} = \arg\max \frac{w^TX^TXw}{w^Tw}$$

it is known that the solution correspond to w_1 being the eigenvector associated to the largest eigenvalue

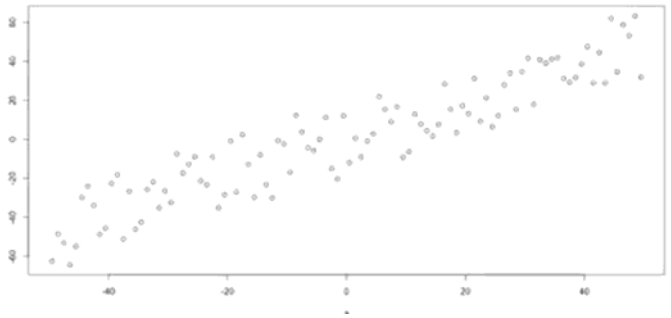
thus the transformed datum is the real number $t_{i1} = x_{i\cdot}\cdot w_1$ in the new coordinates and the vector $t_{i\cdot 1}$ in the old coordinates

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



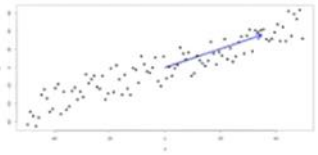
Ritaglio schermata acquisito: 05/05/2021 17:52

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 larget egainvalue of XtX

1st component

example

read samples on the w coordinates and on the x coordinates



original

$t_{[1]} = x_{i,} \cdot w_{,1}$ $t_{[1],}$

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:

$$w_{,k} = \arg\max_{||w_{,1}=1||} ||\hat{X}_k w||^2 = \arg\max \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w}$$

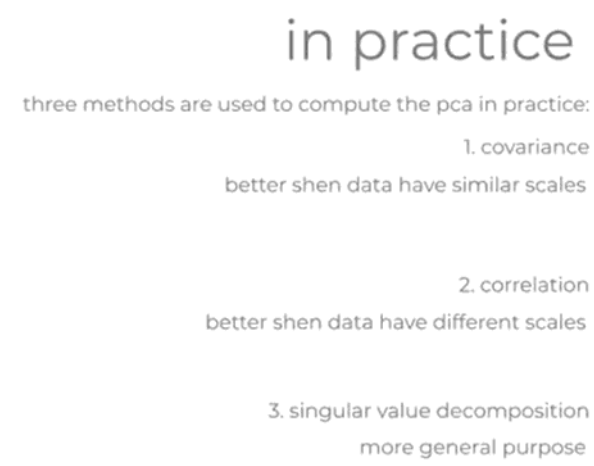
solutions are exactly the remaining eigenvectors of X^TX

the full decomposition is thus given by T=XW with W in $\mathbf{R}^{pxp}$

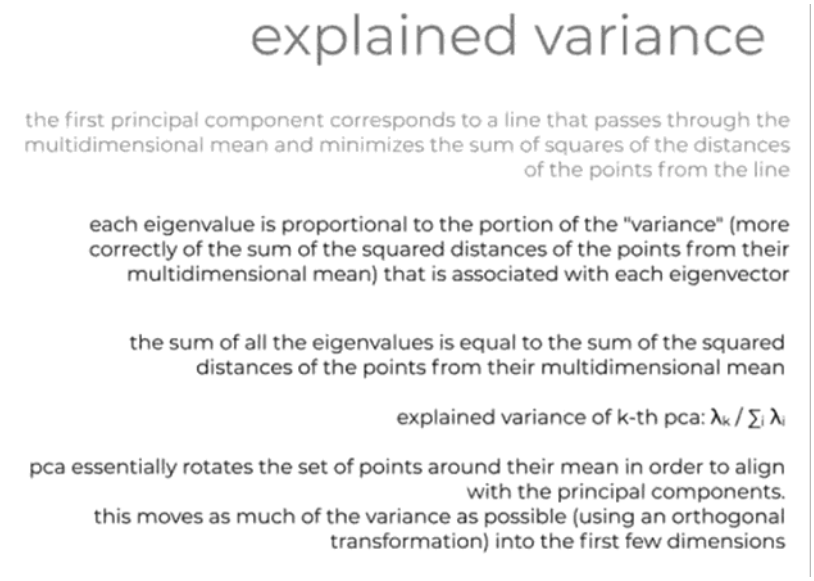
thus the transformed datum is the real number $t_{[1]} = x_{i,} \cdot w_{,1}$ in the new coordinates and the vector $t_{[1],}$ in the old coordinates

Ritaglio schermata acquisito: 05/05/2021 18:07

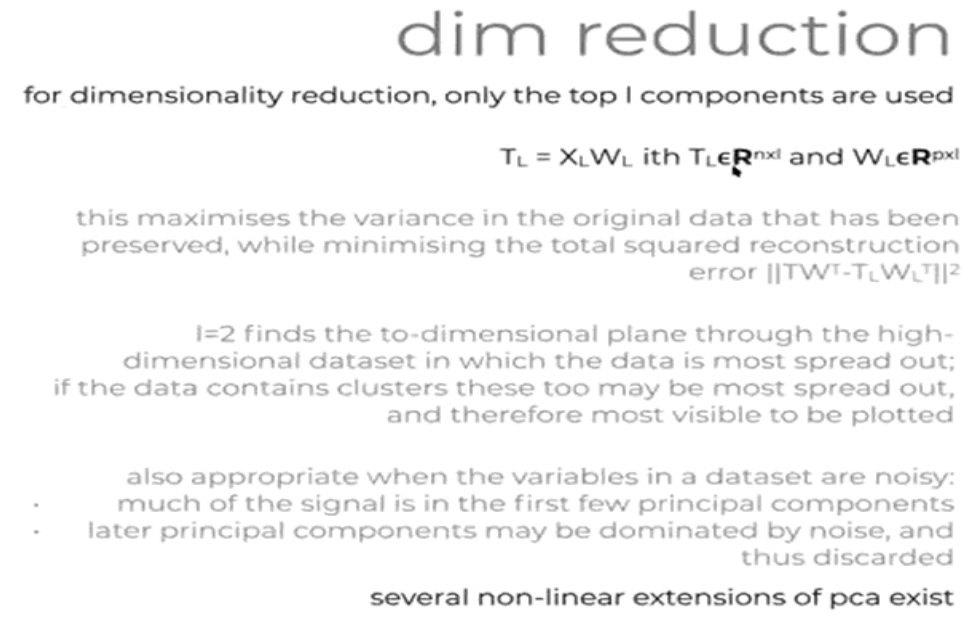
3 methods can be used to compute PCA in practice.



Ritaglio schermata acquisito: 05/05/2021 18:07



Ritaglio schermata acquisito: 05/05/2021 18:08



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 $M \epsilon \mathbf{R}^{d \times r}$, the functional to be maximised becomes

$$\sum_i \sum_j ||x_i M - x_j M||^2$$

that can be rewritten as proportional to

$$tr(M^T X^T X M) - 1 X M M^T X^T 1^T = tr(M^T X^T (I - 1/n \, 1^T 1) X M) = tr(M^T X^T X M)$$

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 $\mathbf{R}^n$ 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

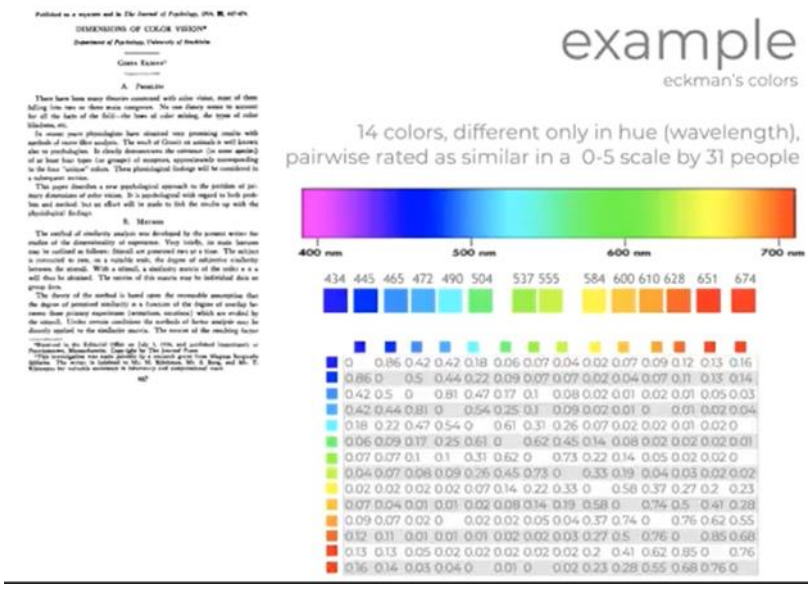
the dimensionality reduction core is the same as pca

if there exists a space $\mathbf{R}^p$ 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

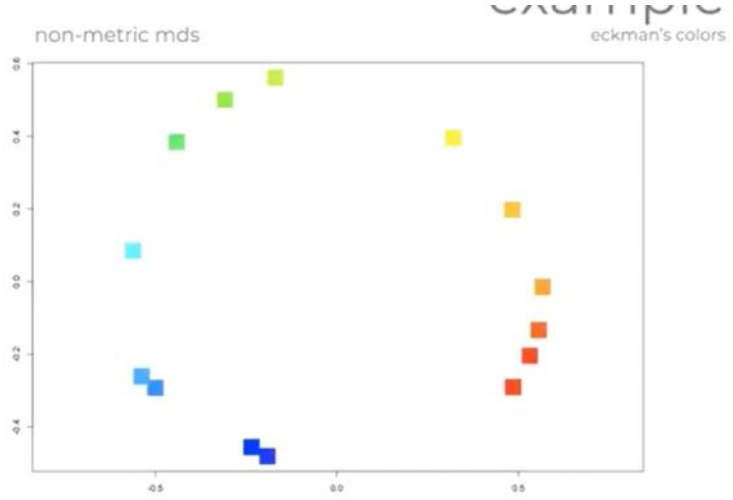
in metric and non-metric mds, the process is given by optimization

Ritaglio schermata acquisito: 05/05/2021 18:24



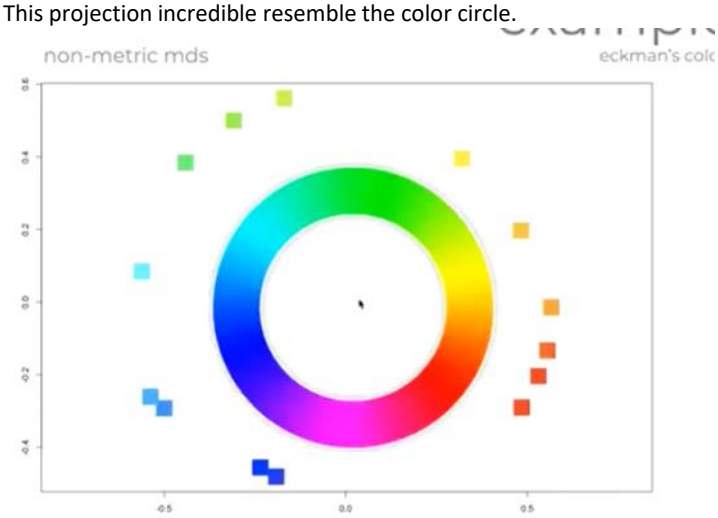
Ritaglio schermata acquisito: 05/05/2021 18:25

The matrix tells how people perceive the similiarities among the colors. What we have is a matrix that summarize and scales t he 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 gr aph.
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 $\mathbf{R}^p$ so that $||x_i-x_j|| \approx d_{ij}$

If d satisfies:

1. $d(x,y)>0$ for $x \neq y$
2. $d(x,y)=0 \iff x=y$
3. $d(x,y)=d(y,x)$ for all x,y
4. $d(x,y) \leq d(x,z)+d(z,y)$ for all x,y,z

the dissimilarity d is called a distance, or a metric

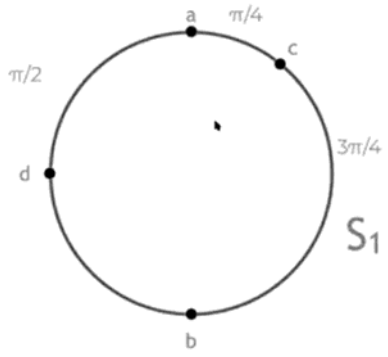
if for some p $||x_i-x_j||=d_{ij}$, d is called an euclidean distance

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 $[0,\pi]$)

	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

not embeddable in any $\mathbf{R}^p$

Ritaglio schermata acquisito: 05/05/2021 18:31

We have this matrix -> this space is not embeddable in euclidean space.

classical mds algorithm

suppose d_{ij} is a (euclidean) distance: find $X=[x_1,\dots,x_n] \in \mathbf{R}^{n \times q}$ so that $||x_i-x_j||=d_{ij}$

X is not unique: for any $c \in \mathbf{R}^q$, 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 $B=XX^T \in \mathbf{R}^{n \times n}$ 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 $d^2_{ij}=||x_i-x_j||^2=\langle x_i-x_j,x_i-x_j \rangle=\langle x_i,x_i \rangle+\langle x_j,x_j \rangle-2\langle x_i,x_j \rangle=b_{ii}+b_{jj}-2b_{ij}$
thus $b_{ij} = -1/2\ (d^2_{ij} - b_{ii} - b_{jj})$ (**2**)

Ritaglio schermata acquisito: 05/05/2021 18:32

classical mds algorithm

call $T=tr(B)=\sum_k b_{kk}$

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**), $\sum_i d^2_{ij}=\sum_i (b_{ii}+b_{jj}-2b_{ij})=\sum_i b_{ii}+\sum_i b_{jj}-2\sum_i b_{ij}$
by (**3**) = T + nb_{jj} - 0 ,
thus $\sum_i d^2_{ij} = T+nb_{jj}$ (**4**)

by symmetry $\sum_j d^2_{ij} = T+nb_{ii}$ (**5**)

by (**2**), $\sum_j\sum_i d^2_{ij}=\sum_j\sum_i (b_{ii}+b_{jj}-2b_{ij})=\sum_j\sum_i b_{ii}+\sum_j\sum_i b_{jj}-2\sum_j\sum_i b_{ij}$
by (**3**) = nT + nT - 0 ,
thus $\sum_j\sum_i d^2_{ij} = 2nT$ (**6**)

Ritaglio schermata acquisito: 05/05/2021 18:34

classical mds algorithm

now plugging (**4**), (**5**) & (**6**) in (**2**):

$$
\begin{aligned}
b_{ij} &= -\frac{1}{2}\left(d_{ij}^2 - b_{ii} - b_{jj}\right) \\
&= -\frac{1}{2}\left(d_{ij}^2 - \frac{\sum_{j=1}^n d_{ij}^2 - T}{n} - \frac{\sum_{i=1}^n d_{ij}^2 - T}{n}\right) \\
&= -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_{j=1}^n d_{ij}^2 - \frac{1}{n}\sum_{i=1}^n d_{ij}^2 + \frac{1}{n}2T\right) \\
&= -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_{j=1}^n d_{ij}^2 - \frac{1}{n}\sum_{i=1}^n d_{ij}^2 + \frac{1}{n}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n d_{ij}^2\right) \qquad (7)
\end{aligned}
$$

Ritaglio schermata acquisito: 05/05/2021 18:36

classical mds algorithm

find now a matrix form for (**7**) defining the matrices
$D_2=(d^2_{ij})$ and $\mathbf{O}$ in $\mathbf{R}^{n\times n}$ and $\mathbf{1}$ in $\mathbf{R}^{n\times 1}$

$$O_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ & & \cdots & \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} \cdot (11\cdots 1) = \mathbf{1}\cdot\mathbf{1}^T$$

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 $C_n=\mathbf{I}_n-(1/n)\mathbf{1}\mathbf{1}^T$, becomes
$B=(-1/2)\, C_nD_2C_n$ (**8**)

C_n is called entering matrix because for any Y, C_nYC_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≠0 be an eigenvector for the eigenvalue λ∈**C**, so Bv=λv

$$\bar{v}^T B v = \bar{v}^T (B v) = \bar{v}^T (\lambda v) = \lambda (\bar{v}^T \cdot v)$$
$$=$$
$$\bar{v}^T B v = (B\bar{v})^T v = (\bar{\lambda}\bar{v})^T v = \bar{\lambda}(\bar{v}^T \cdot v)$$

then λ is real; let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ the n real eigenvalues of B and $z_1,\ldots,z_n$ the corresponding eigenvectors

then B can be *diagonalised*, i.e. written as
B = V ΛV^T
where Λ 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) $C_n D_A C_n$=B=XXT= V ΛV^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 pxp diagonal matrix with the p largest √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

$$D_{(2)} = D = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$B = (1/8) \cdot \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

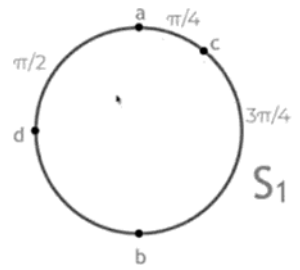
$$\Lambda = (1/2) \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

restricting to $\mathbf{R}^3$

$$X_{(3)} = V_{(3)}\Lambda_{(3),\frac{1}{2}} = \begin{pmatrix} 0.0000000 & 0.6123724 & 0.0000000 \\ -0.1893048 & -0.2041241 & 0.5454329 \\ -0.3777063 & -0.2041241 & -0.4366592 \\ 0.5670111 & -0.2041241 & -0.1087736 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

A,B,C,D are exactly the vertices of regular tetrahedron of edge 1 in $\mathbf{R}^3$

Ritaglio schermata acquisito: 05/05/2021 18:40

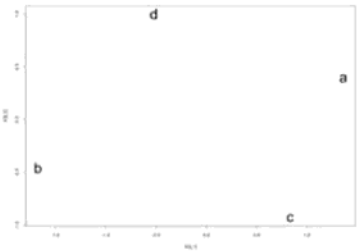


example 2

circular distance

D=π ·

	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



spec(B)={5.6117, 2.2234 ,0,-1.2039)
thus we need to restrict to $X_{(2)}$

dist=π ·

	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

Ritaglio schermata acquisito: 05/05/2021 18:43

alternative mds algorithms

relax the objective $||x_i-x_j|| \approx d_{ij}$ to $||x_i-x_j|| \approx f(d_{ij})$

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)

example 1:

$f(d_{ij})=\alpha d_{ij}+\beta$ and $= (\sum_{i<j} (||x_i-x_j||-(\alpha d_{ij}+\beta))^2 / \sum_{i<j} d^2_{ij})^{(1/2)}$

note that $\alpha=1$ and $\beta=0$ gives the classical case, with a different solution

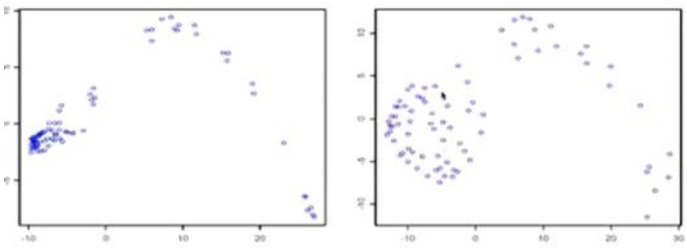
example 2: sammon mapping

$$\mathscr{L} = \frac{1}{\sum_{i<k} d_{ik}} \sum_{i<j} \frac{(||x_i - x_j|| - d_{ij})^2}{d_{ij}}$$

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



classical sammon

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

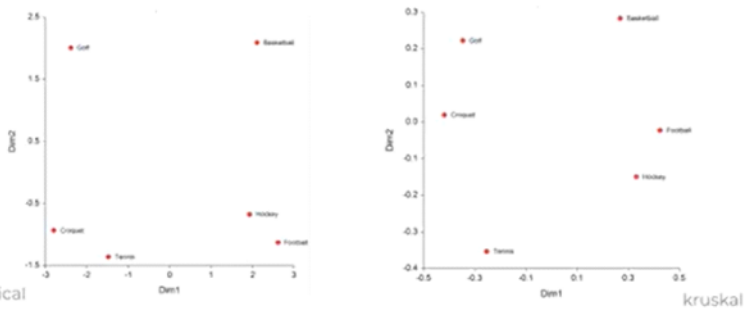
here f is only implicitly defined as a regression curve, and only preserves the order of d, that is $f(d_{ij})<f(d_{kl})$ if $d_{ij}<d_{kl}$
thus only the order of d is needed, not the actual values

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



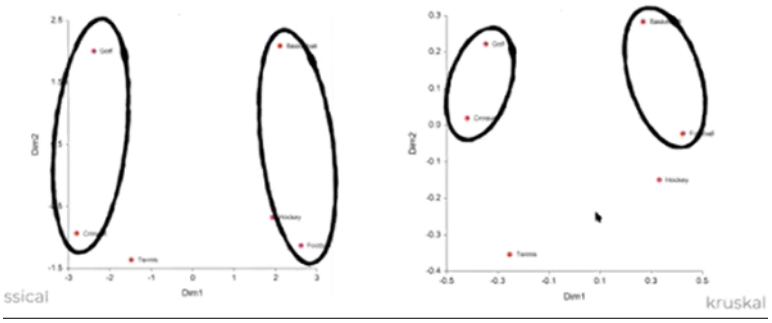
sical kruskal

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	3
Golf					0	2
Croquet						0

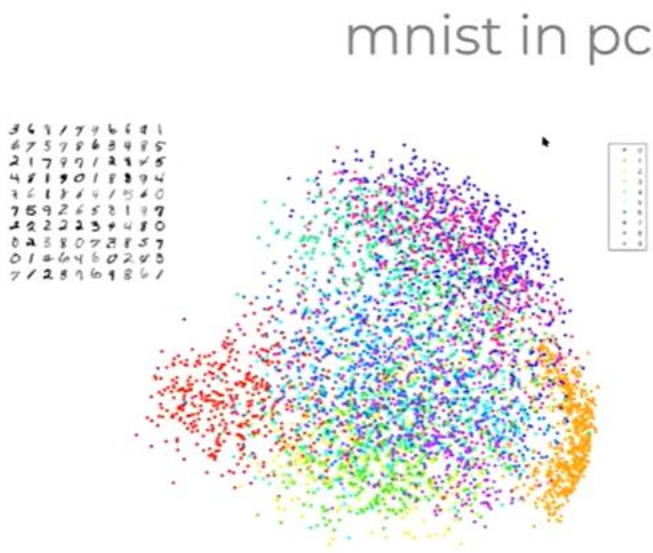


ssical kruskal

Ritaglio schermata acquisito: 05/05/2021 18:49

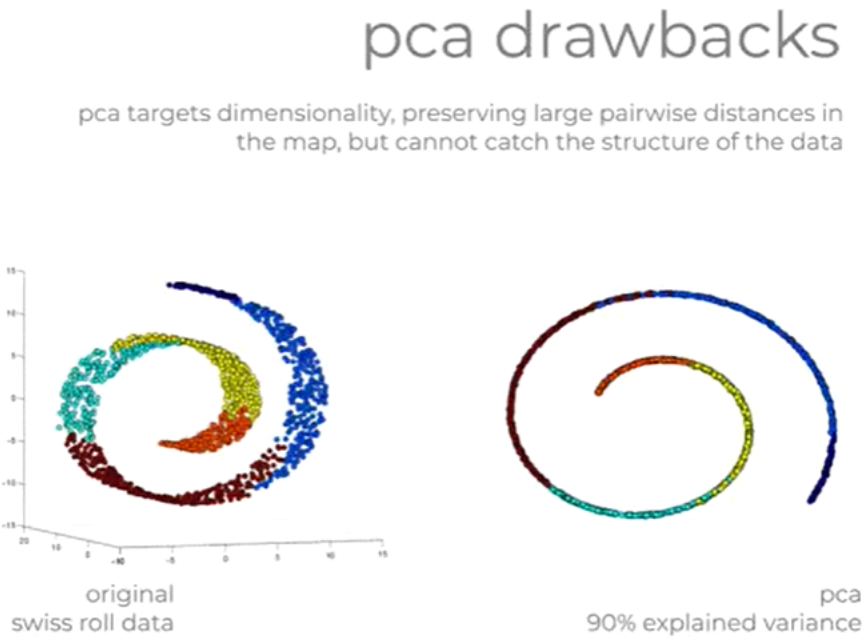
TSNE & UMAP

domenica 16 maggio 2021 16:27



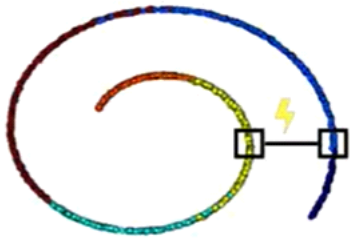
Ritaglio schermata acquisito: 16/05/2021 16:29

Dataset where 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.



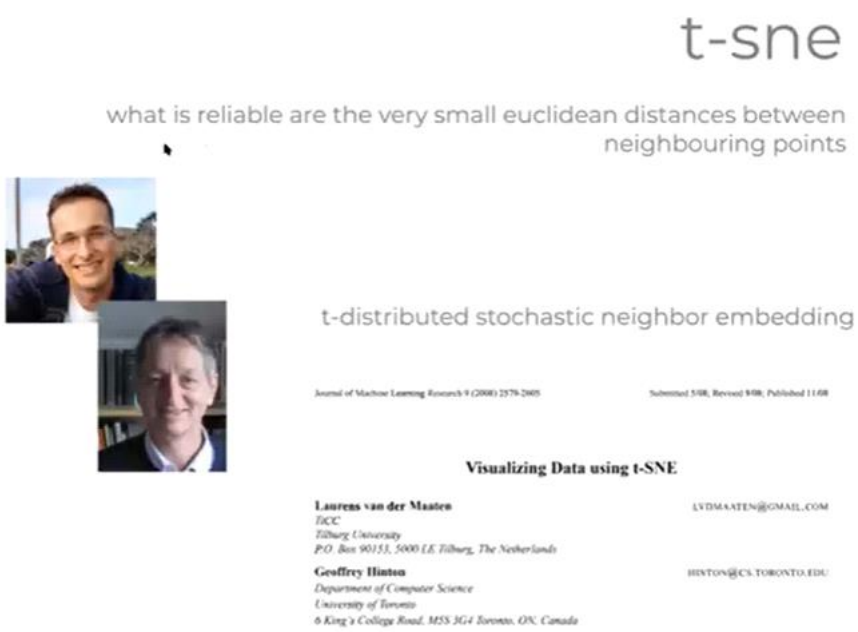
Ritaglio schermata acquisito: 16/05/2021 16:34

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.



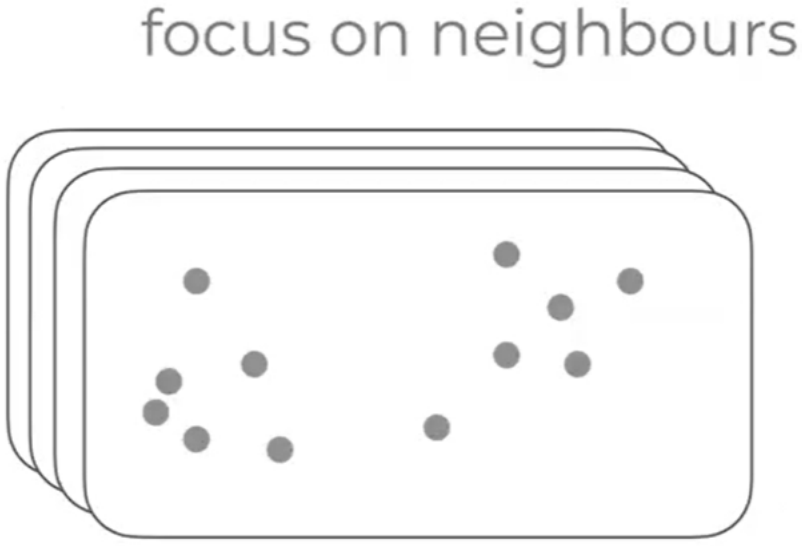
Ritaglio schermata acquisito: 16/05/2021 16:35

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.



Ritaglio schermata acquisito: 16/05/2021 16:36

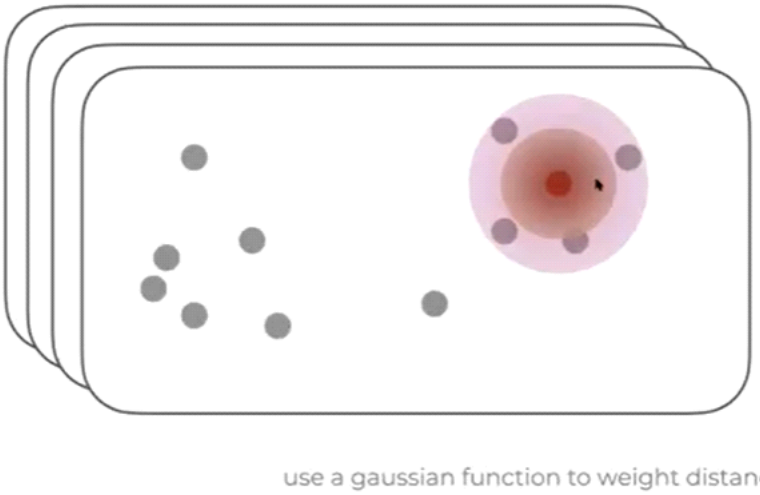
What is reliable in the dimensionality reduction is indeed is the concept of neighborhood.



Ritaglio schermata acquisito: 16/05/2021 16:37

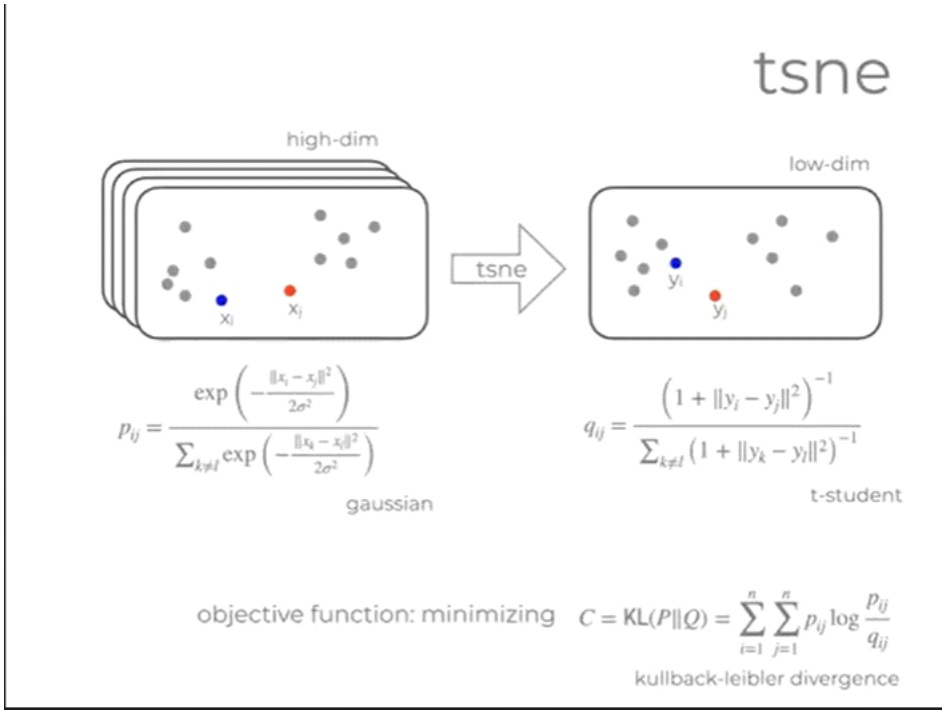
Suppose to have the above data set. Let's consider the red point. Considering the big circle around

the circle the more it is closer to it the more it is similar to the red point.



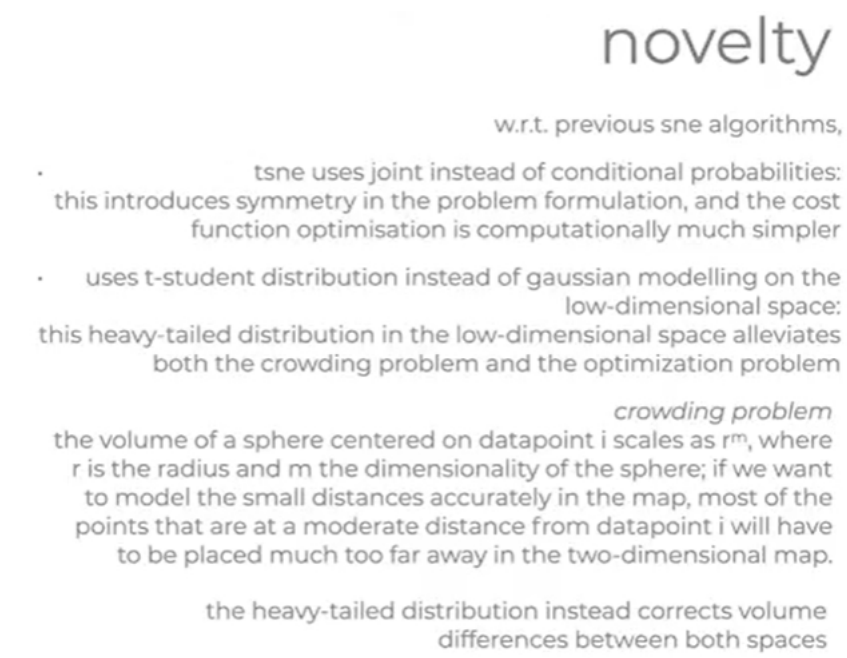
Ritaglio schermata acquisito: 16/05/2021 16:38

We do so to construct the concept of metric neighborhood.



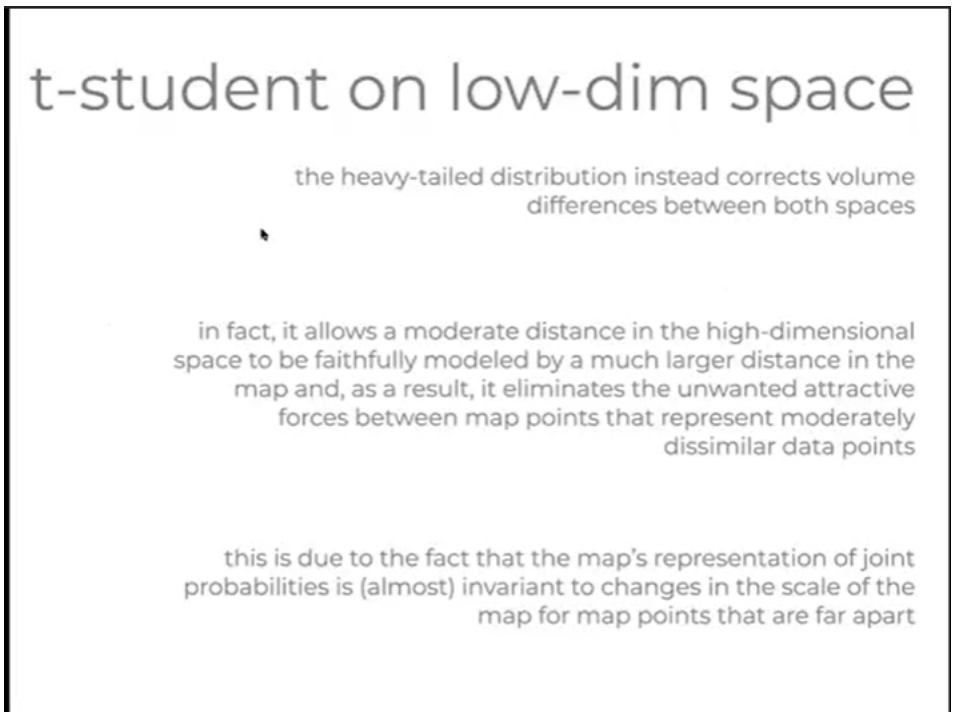
Ritaglio schermata acquisito: 16/05/2021 16:39

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.



Ritaglio schermata acquisito: 16/05/2021 16:41

Join probability instead of conditional one -> this introduce symmetry.
t-student distirbution on the arraiving low dimensional space is helping the solution of the problem.



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

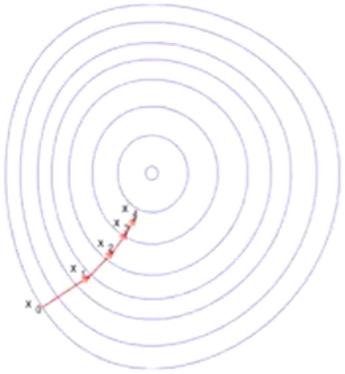
in details, one takes consecutive steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.

Ritaglio schermata acquisito: 16/05/2021 16:45

gradient descent

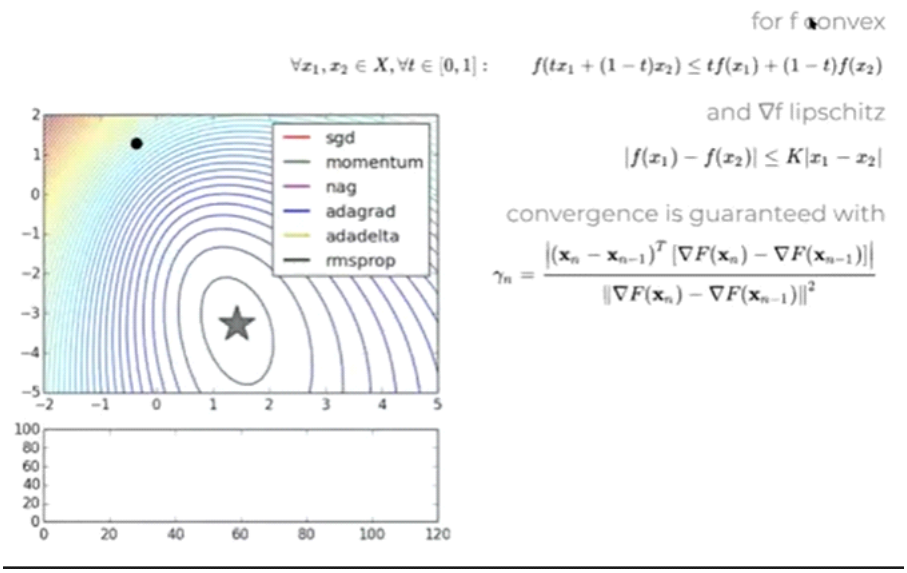
$f:\mathbf{R}^p \to \mathbf{R}$, defined and differentiable in a neighborhood of a point a, decreases *fastest* if one goes from a in the direction of the negative gradient of f at a, $-\nabla f(a)=-(\partial f/\partial x_1,\dots,\partial f/\partial x_p)|_a$

if $a_{n+1}=a_n-\gamma\nabla f(a_n)$ for a small positive γ, then $f(a_{n+1})\leq f(a_n)$



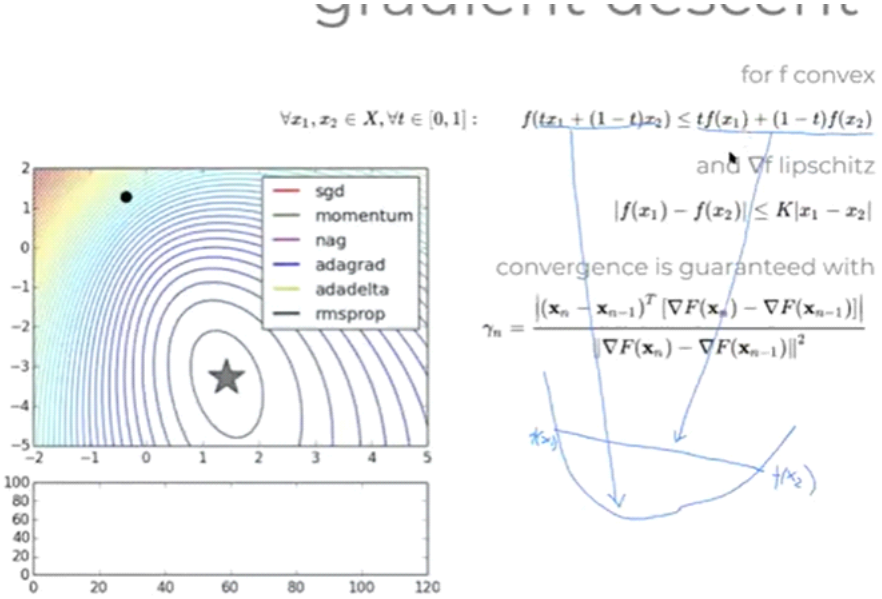
start with a guess x_0 of a local minimum for f and consider the sequence $x_0,x_1,x_2,\dots$ with $x_{n+1}=x_n-\gamma_n\nabla f(x_n)$: then we have a monotonic sequence $f(x_0)\geq f(x_1)\geq f(x_2)\geq\dots$ so that the sequence $x_0,x_1,x_2,\dots$ can converge to the minimum of f

Ritaglio schermata acquisito: 16/05/2021 16:48

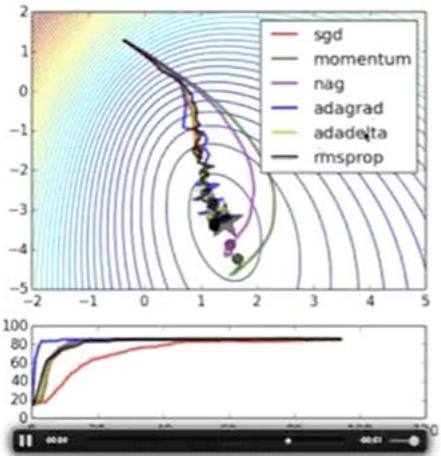


Ritaglio schermata acquisito: 16/05/2021 16:50

Convex -> if we construct the segment from x1 to x2 the image of the segment is smaller than the segment between the two images..



Ritaglio schermata acquisito: 16/05/2021 16:54



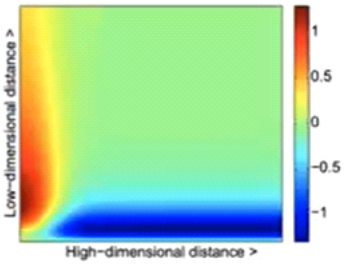
Ritaglio schermata acquisito: 16/05/2021 16:55

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

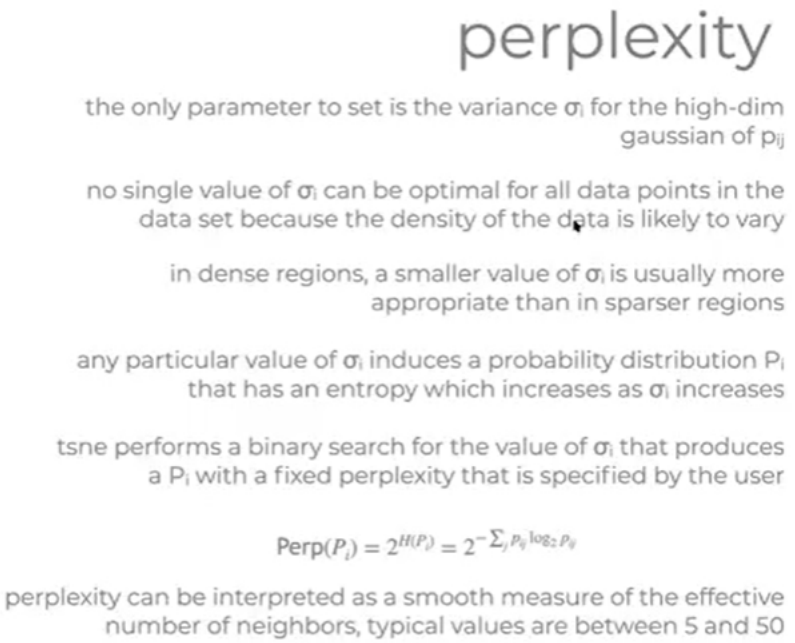


early compression:
force the map points to stay close together at the start of the optimization, implemented by adding an additional l2-penalty to the cost function that is proportional to the sum of square distances of the map points from the origin

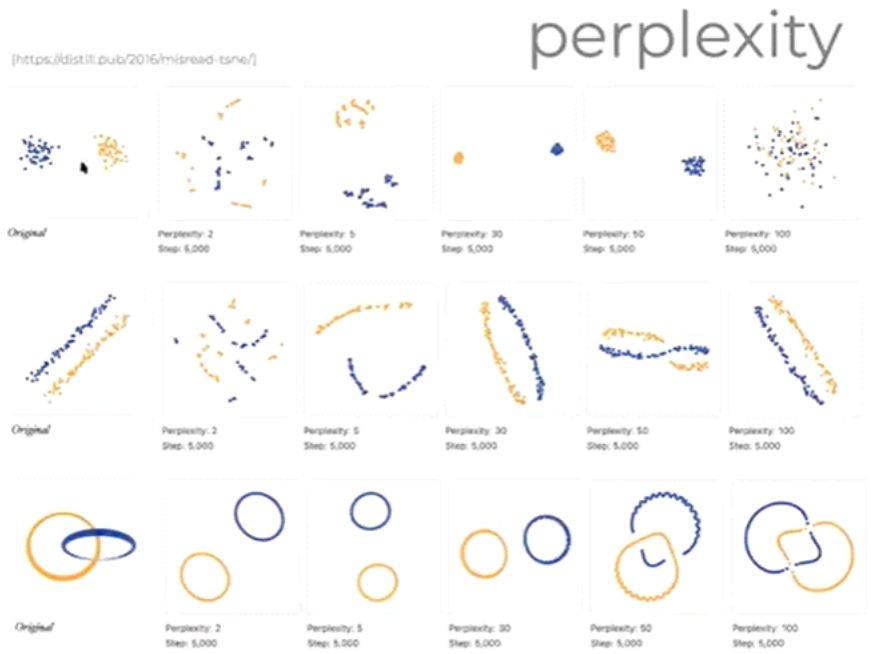
early exaggeration:
multiply all of the p_{ij}'s by, for example, 4, in the initial stages of the optimization, modeling the large p_{ij}'s by fairly large q_{ij}'s; natural clusters in the data tend to form tight widely separated clusters in the map

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.



Ritaglio schermata acquisito: 16/05/2021 16:58

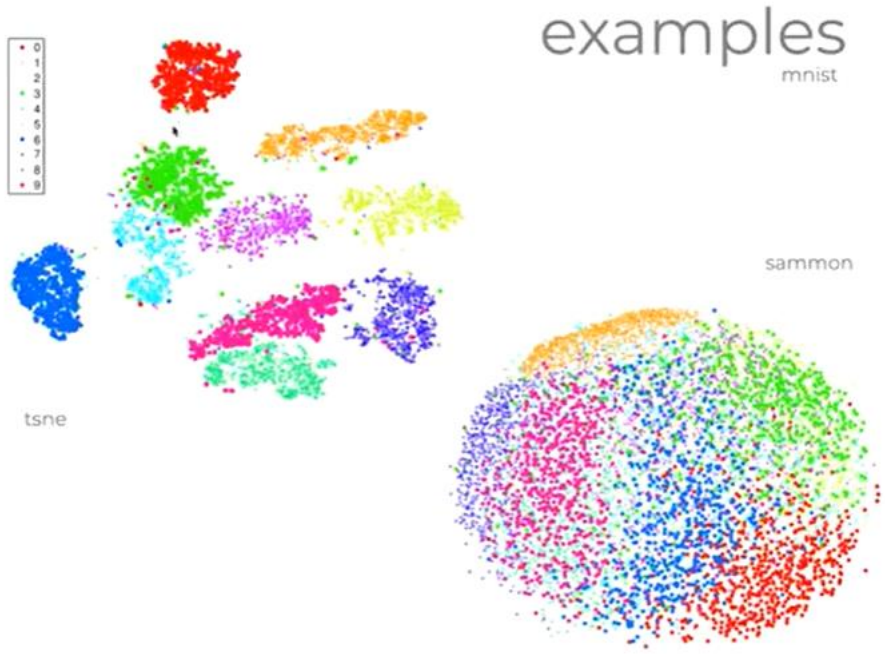


Ritaglio schermata acquisito: 16/05/2021 17:11

The pictures on the left are the orginal 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.
Amunting perplexity -> means using more data points. If we increase before the cluster 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 becasue the two rings overlap.

Hence, there is not a unique and best value for the perplexity.

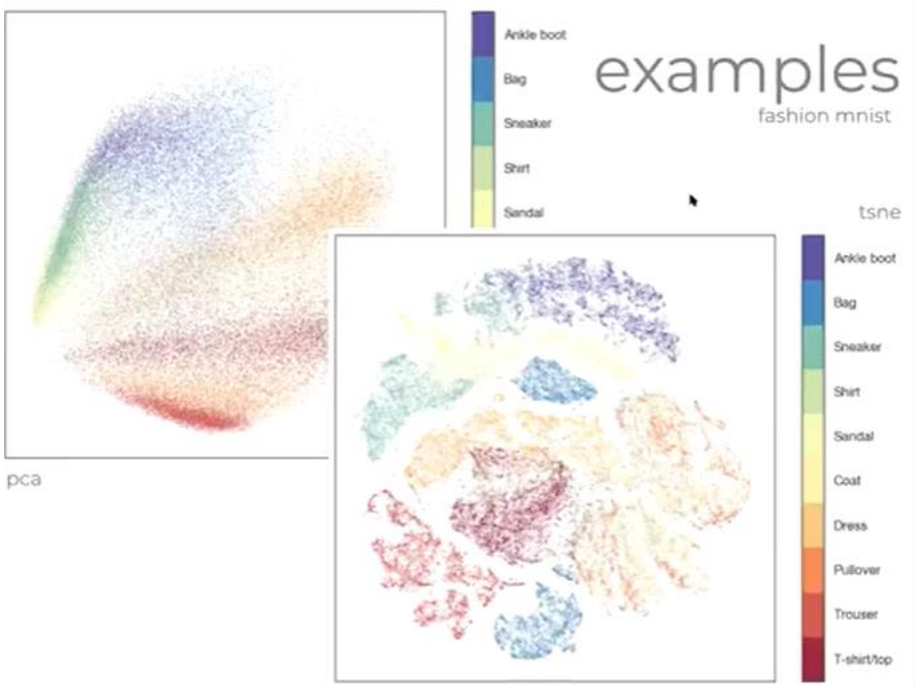


Ritaglio schermata acquisito: 16/05/2021 17:17

Tsne perform quite well

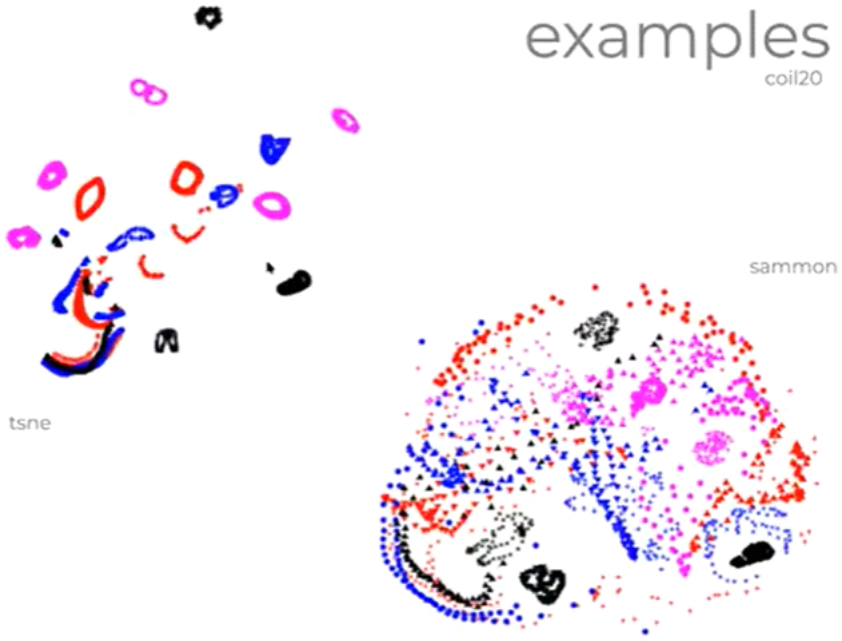
Ritaglio schermata acquisito: 16/05/2021 17:17

Images here are clothing.



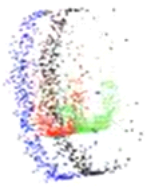
Ritaglio schermata acquisito: 16/05/2021 17:18

Tsne on the right.

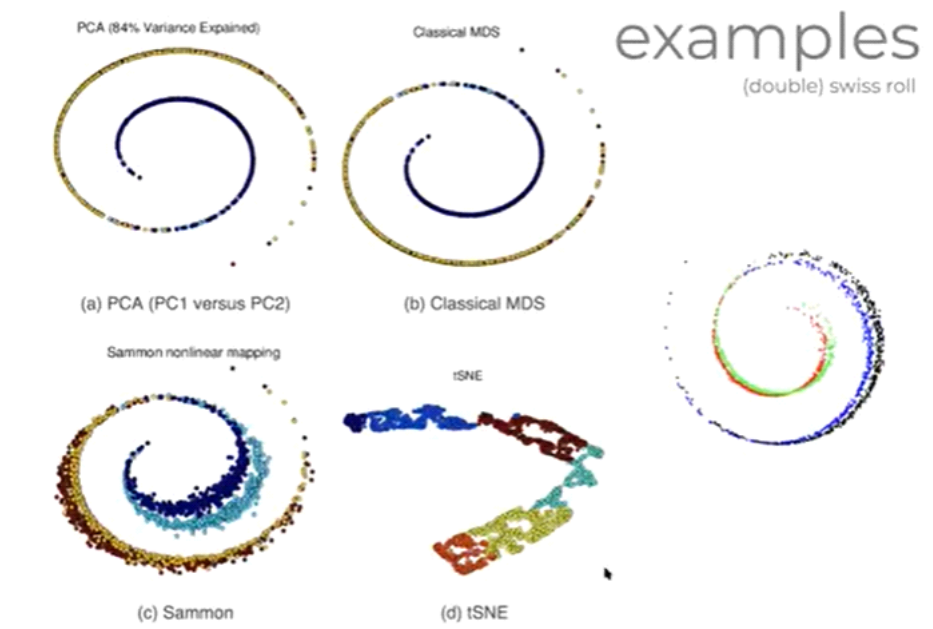


Ritaglio schermata acquisito: 16/05/2021 17:18

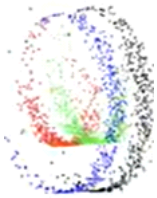
Tsne separed well the classes



Ritaglio schermata acquisito: 16/05/2021 17:19

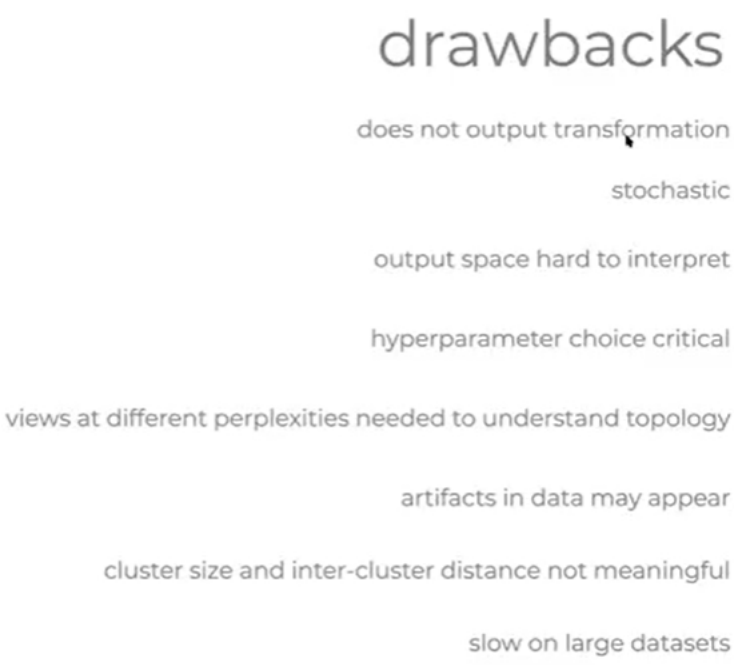


Ritaglio schermata acquisito: 16/05/2021 17:19



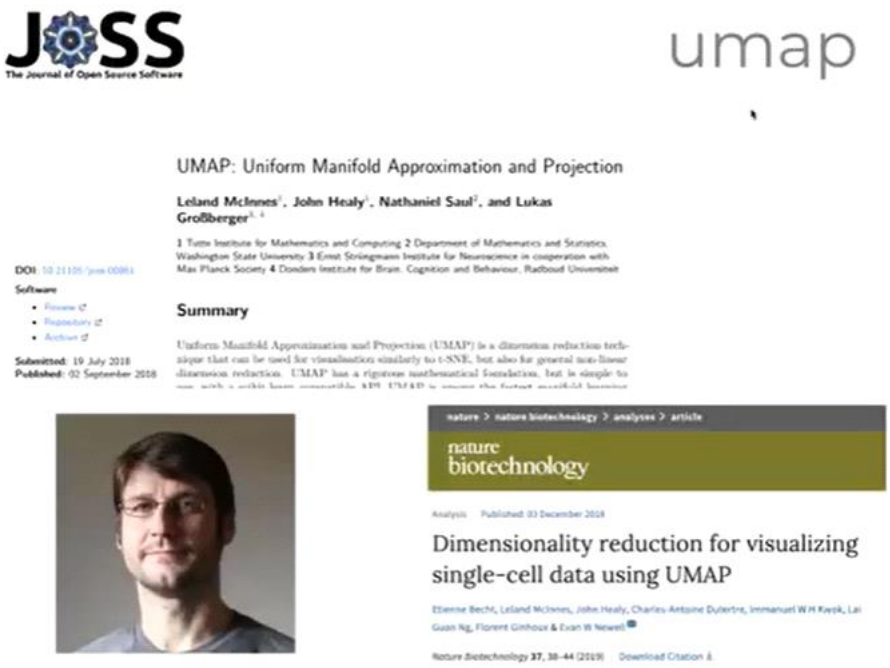
Ritaglio schermata acquisito: 16/05/2021 17:19

Tsne is good also in the case of the double swiss roll which is a really complex data set.



Ritaglio schermata acquisito: 16/05/2021 17:20

Sthocastic -> 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.



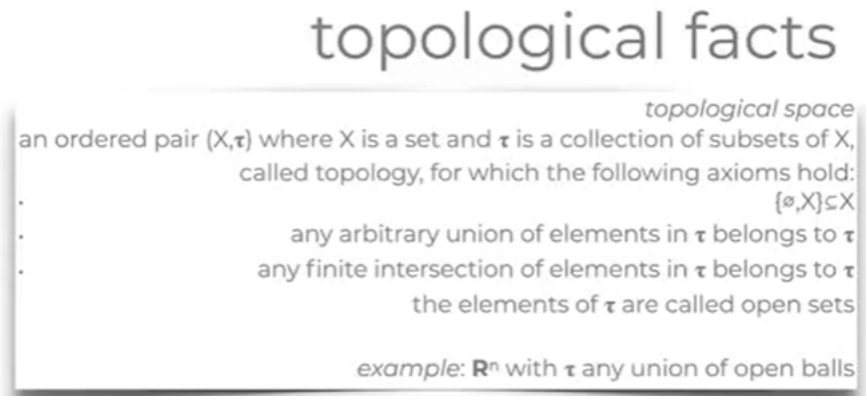
Ritaglio schermata acquisito: 16/05/2021 17:23

As a solution for these problems, has been proposed a new method.

The last dimensionality reduction method is umap, uniform manifold approximation and projection.

UMAP:

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.

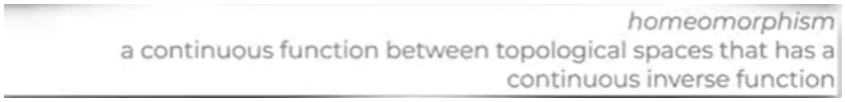


Ritaglio schermata acquisito: 16/05/2021 17:32

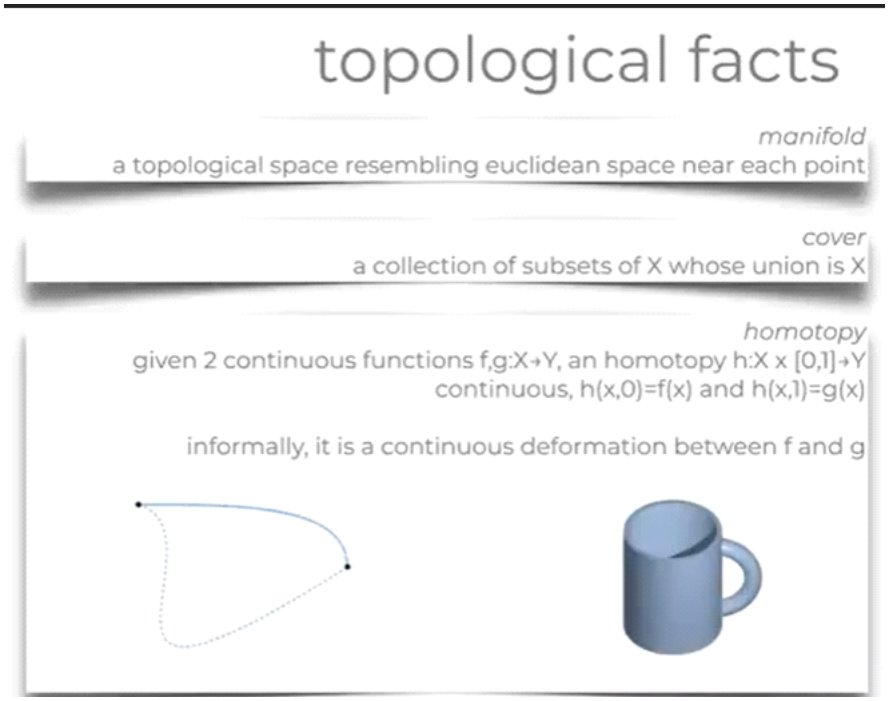
open balls -> is the circle surrounding a given point but without the borders.

Ritaglio schermata acquisito: 16/05/2021 17:34

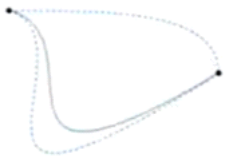
A continous function is a function that goes from a topological space x to a topological space y such that the counter image of an elmenent in tau_y lies in tau_x
The pre image in an open set in a pre image in open set in concise terms.



Ritaglio schermata acquisito: 16/05/2021 17:37

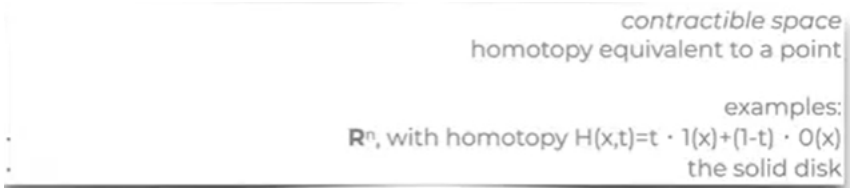


Ritaglio schermata acquisito: 16/05/2021 17:37

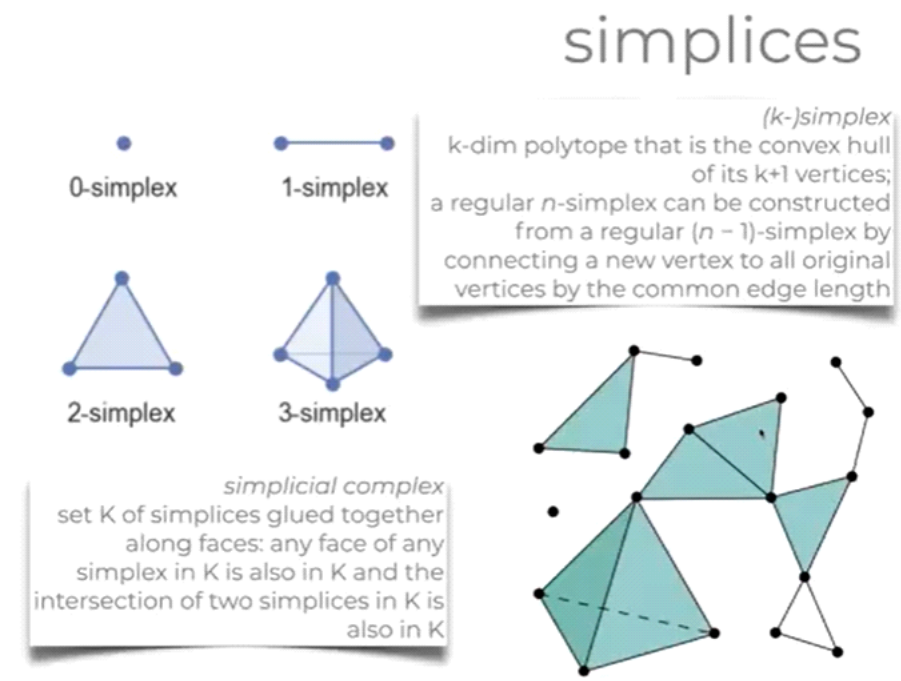


Ritaglio schermata acquisito: 16/05/2021 17:39

Ritaglio schermata acquisito: 16/05/2021 17:40

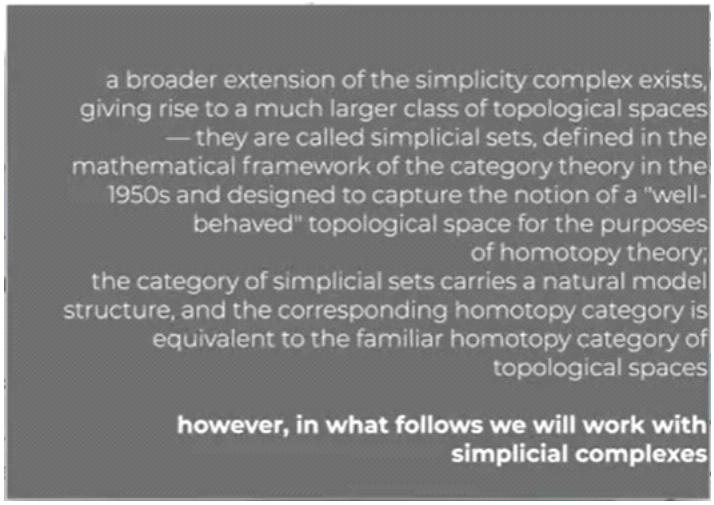


Ritaglio schermata acquisito: 16/05/2021 17:47



Ritaglio schermata acquisito: 16/05/2021 17:48

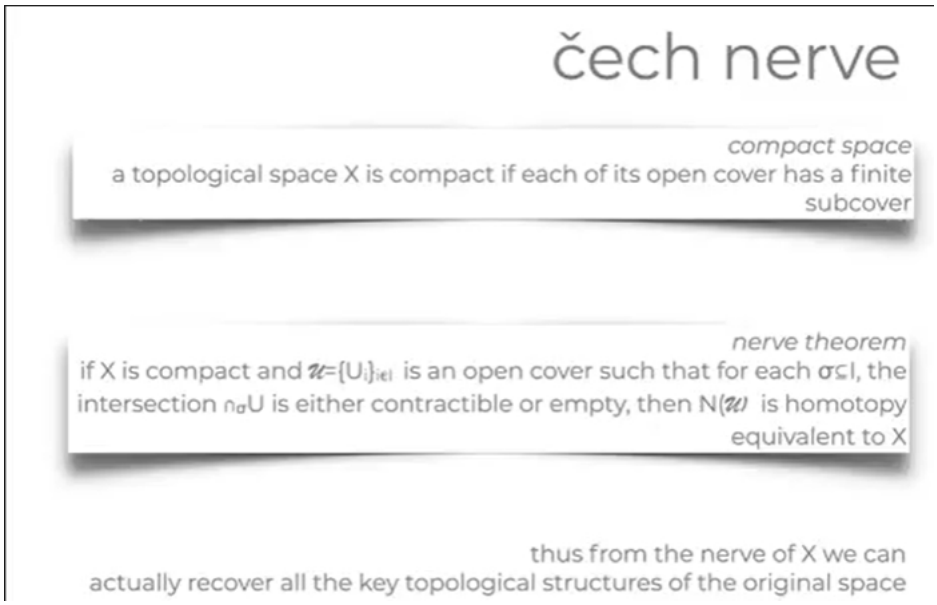
SIMPLICES -> Smallest convex geometrical figures that has k+1 vertices.
In general an n-simplex can be constructed from the previous simplex adding a new point and connecting it to all the other points.
The interesetion of two simplices is still a simplex.



Ritaglio schermata acquisito: 16/05/2021 18:01

Ritaglio schermata acquisito: 16/05/2021 18:02

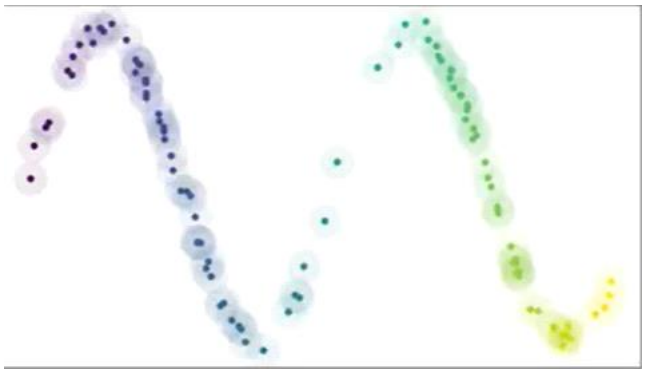
Cech nerve -> all the possible intersection of element in *U*



Ritaglio schermata acquisito: 16/05/2021 18:06

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.

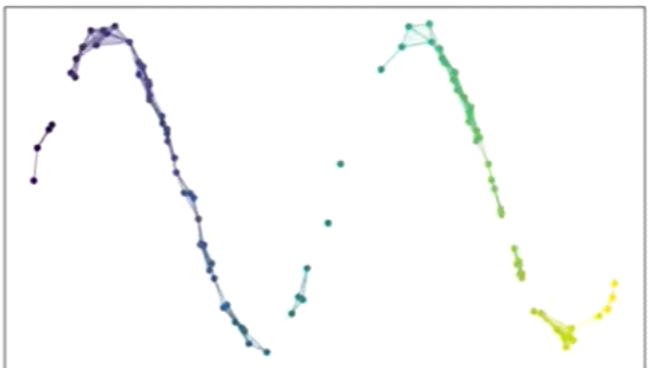


Ritaglio schermata acquisito: 16/05/2021 18:09

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.

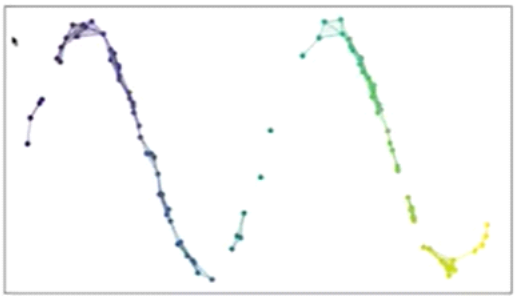
issues

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

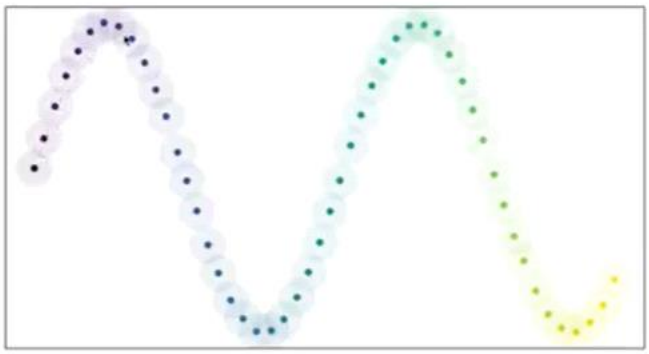
if points are too scattered, the structure cannot be properly captured
if points are too dense, the cover generates a too high-dim nerve



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



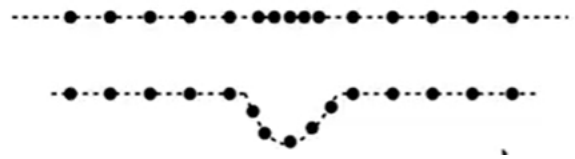
Ritaglio schermata acquisito: 16/05/2021 18:12

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 strecth 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.

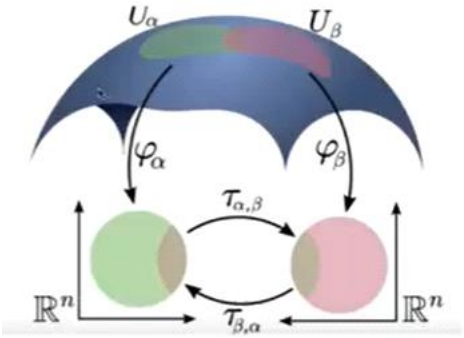


Ritaglio schermata acquisito: 16/05/2021 18:14

Bending the space -> guarantess that all the point are uniformelly distributed as we wanted.

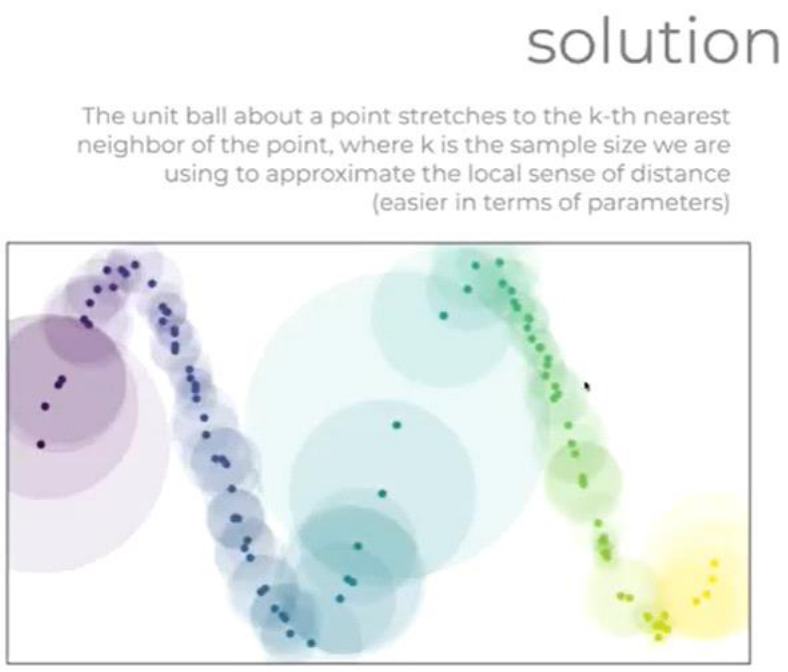
thus give to each point its own unique distance function, and select balls of radius one with respect to that local distance function

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 R to the n where the distance is defined.
There is a map that allows to go from one space to another.

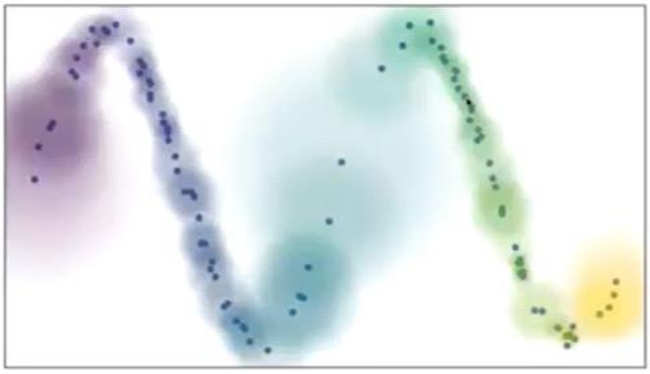


Ritaglio schermata acquisito: 16/05/2021 18:16

fuzziness

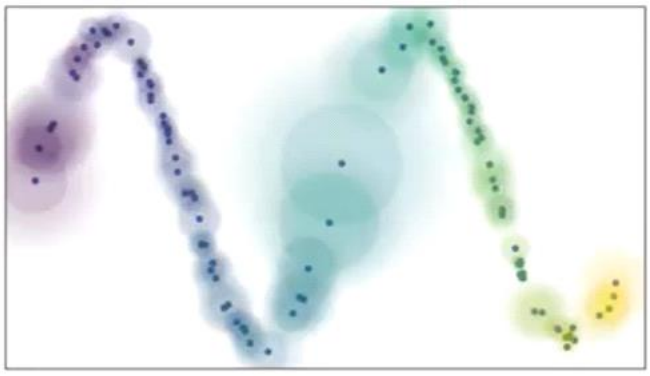
given the local metric, we can weight the edges of the graph according to the edge' vertices distance

mathematically, we move from the simplicial complexes to the simplicial sets (category theory), which corresponds to move from classical balls to fuzzy balls, where ϵ 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

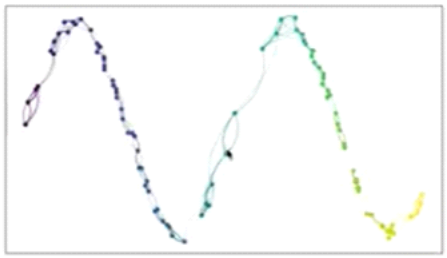


Ritaglio schermata acquisito: 16/05/2021 18:18

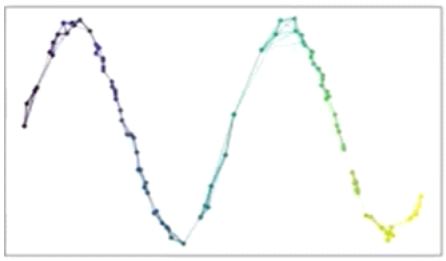
Fuzziness start to take place only behind the nearest neighbors.
The focus is more on the focus on the distance among the neighbor rather than the distance among all the points.

incompatibility

since each point has its own metric, distance from a to b may be different than distance from b to a



Ritaglio schermata acquisito: 16/05/2021 18:19



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)

Ritaglio schermata acquisito: 16/05/2021 18:19

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

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)$$

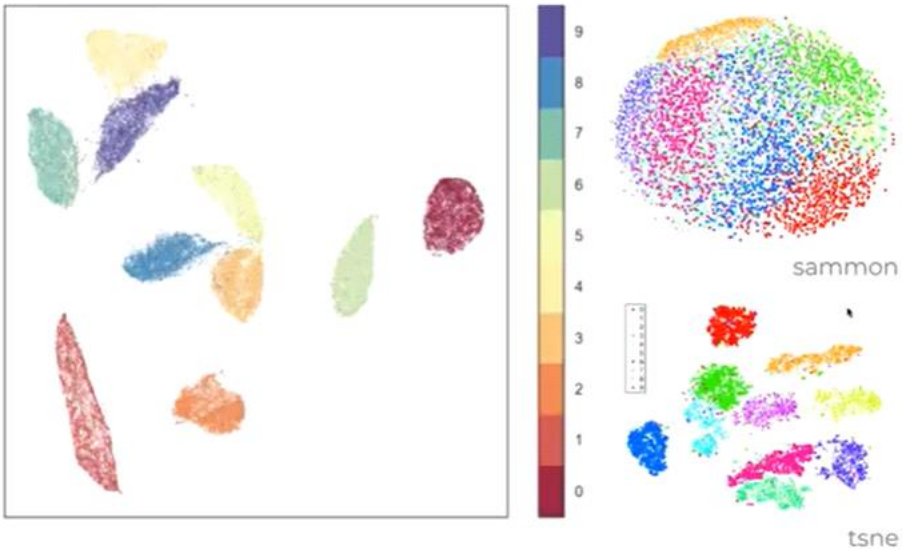
attractive force between points when w is large in high-dimension optimizing clumps

repulsive force between points when w is small in high-dimension: optimizing gaps

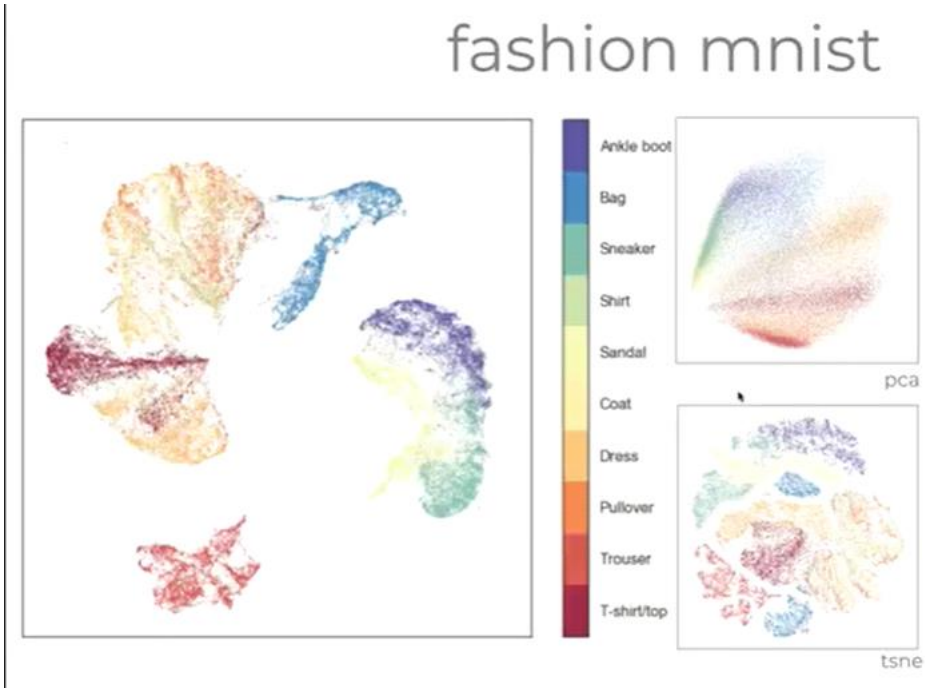
Ritaglio schermata acquisito: 16/05/2021 18:20

Bernoulli variables -> cross-entropy.

mnist



Ritaglio schermata acquisito: 16/05/2021 18:22



Ritaglio schermata acquisito: 16/05/2021 18:23

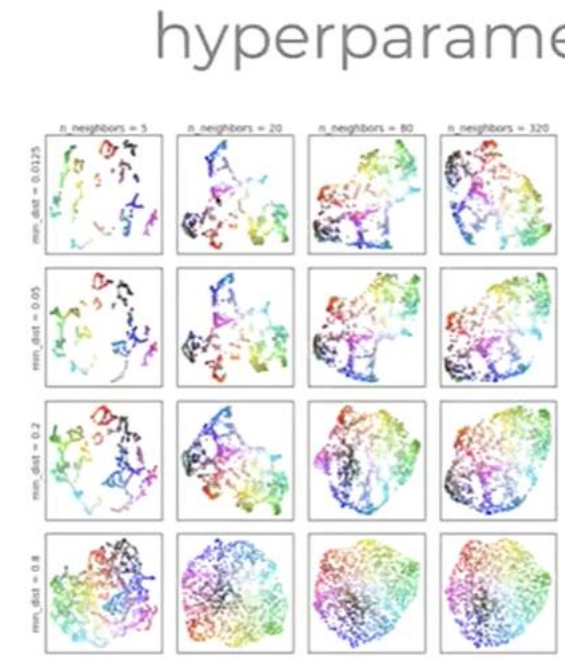
hyperparameters

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.

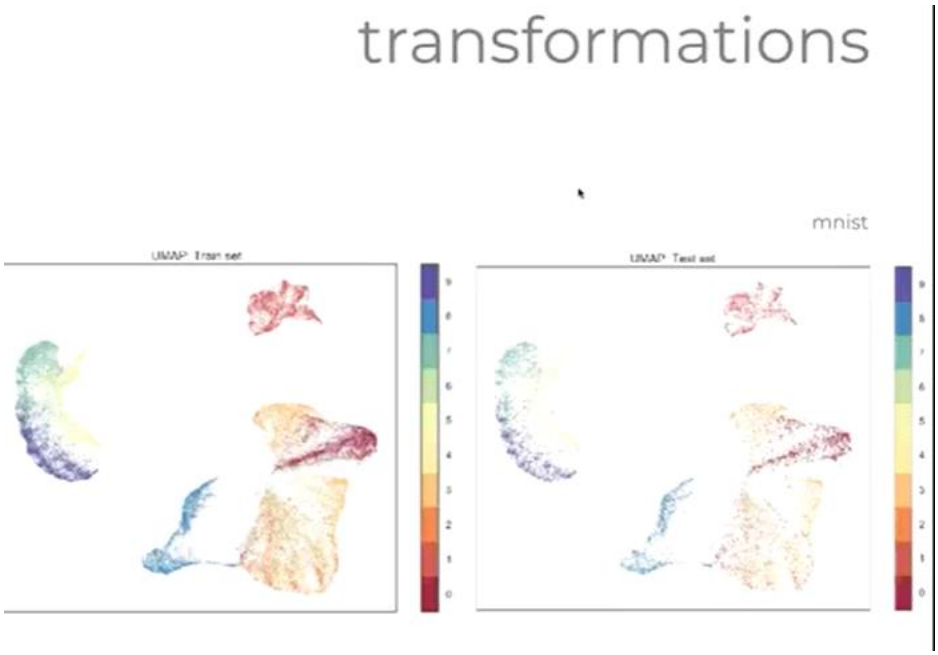
min_dist
controls how tightly the embedding is allowed compress points together
larger values ensure embedded points are more evenly distributed, while smaller values allow the algorithm to optimise more accurately with regard to local structure
sensible values [0.001,0.5], with 0.1 being a reasonable 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

Ritaglio schermata acquisito: 16/05/2021 18:23

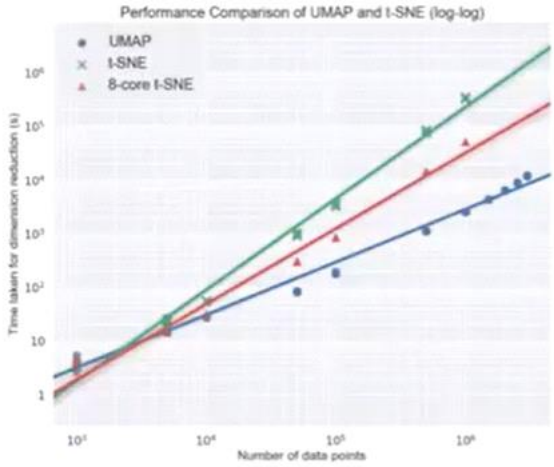


Ritaglio schermata acquisito: 16/05/2021 18:33



Ritaglio schermata acquisito: 16/05/2021 18:34

Umap -> transform the whole space.



cputime

dataset	dataset size	t-SNE	UMAP
COIL20	1440x16384	20s	**7s**
COIL100	72000x49152	683s	**121s**
Shuttle	58000x9	741s	**140s**
MNIST	70000x784	1337s	**98s**
F-MNIST	70000x784	906s	**78s**
GoogleNews	200000x300	16214s	**821s**

Ritaglio schermata acquisito: 16/05/2021 18:35

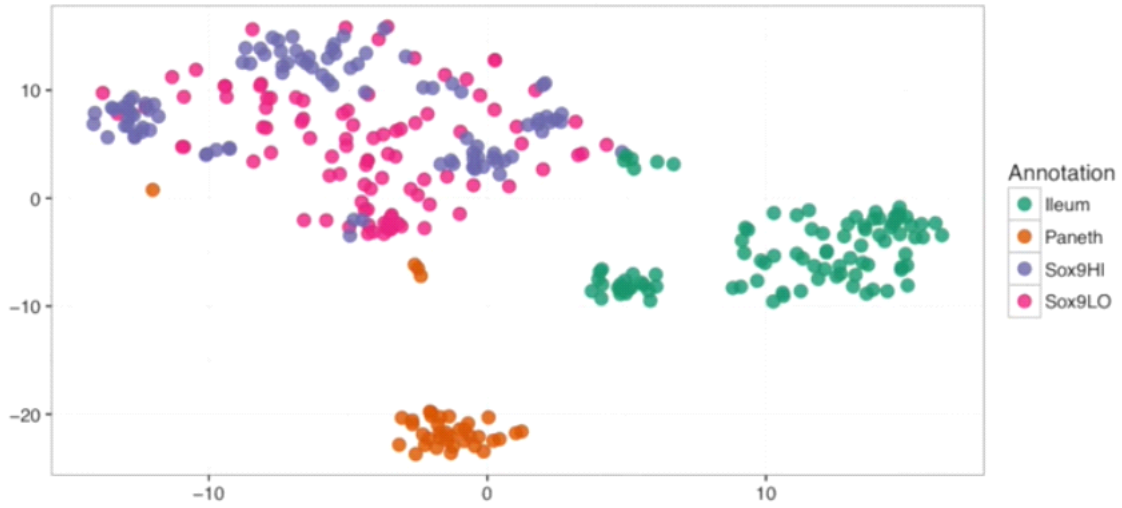
Ritaglio schermata acquisito: 16/05/2021 18:36

T-SNE & UMAP

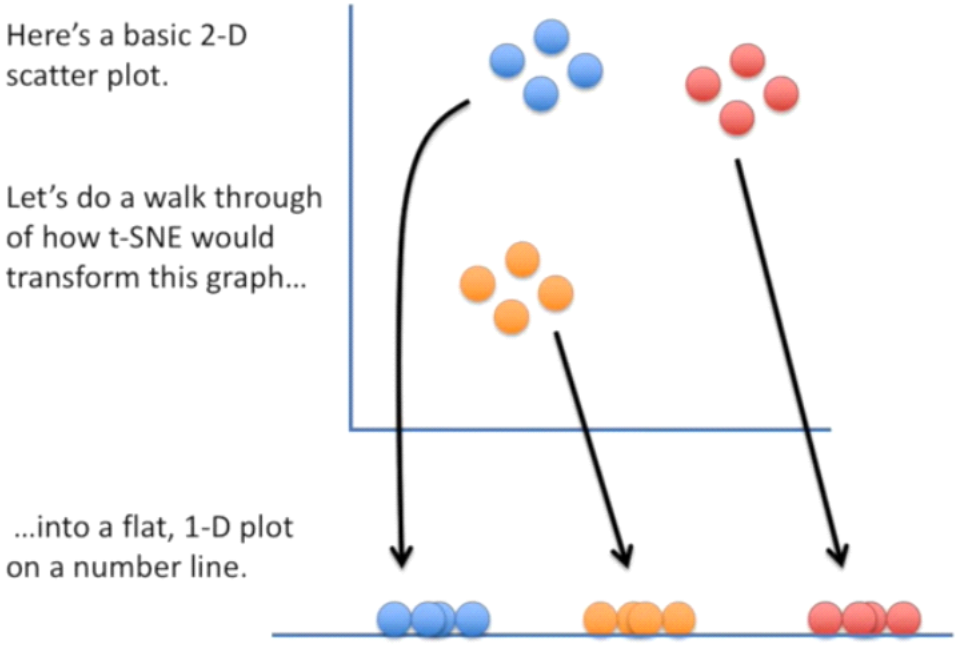
giovedì 26 agosto 2021 10:58

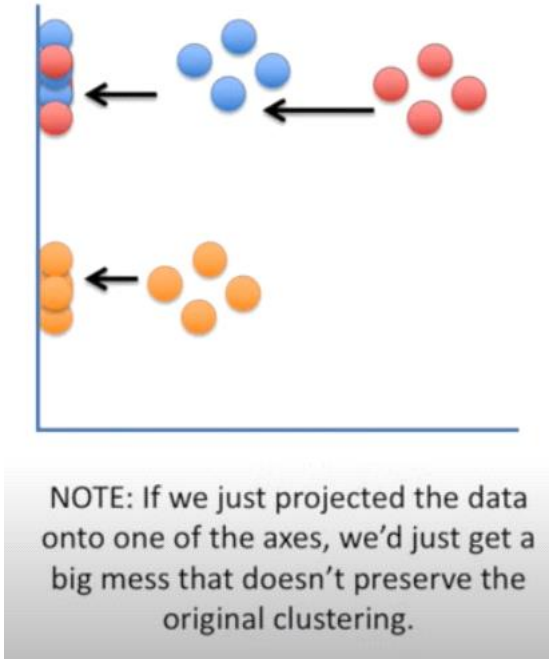
T-SNE

What t-SNE does is taking an high dimensional dataset and reduce it to a low dimensional graph that retains a lot of the original information.

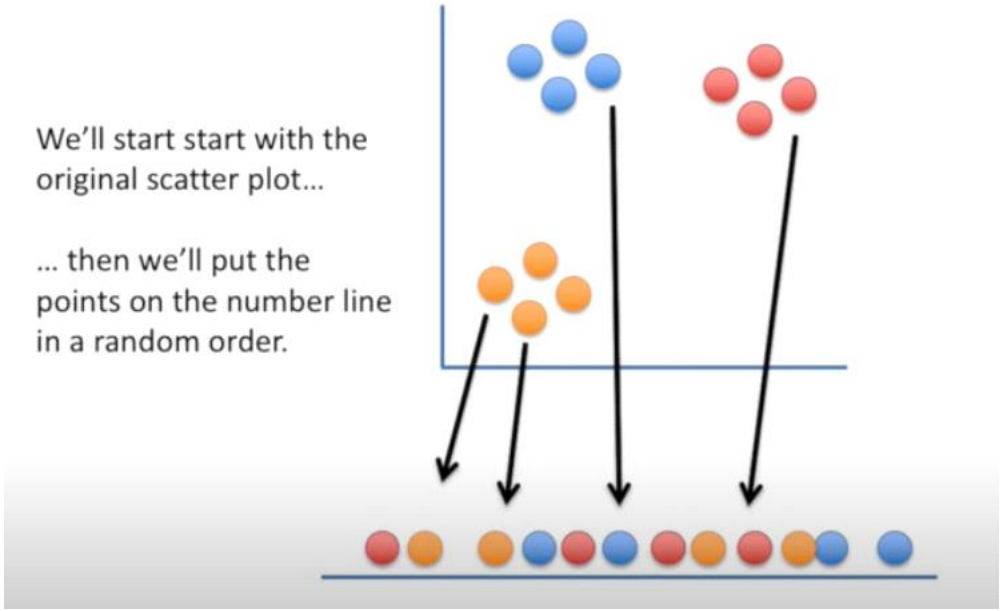


Ritaglio schermata acquisito: 26/08/2021 11:00





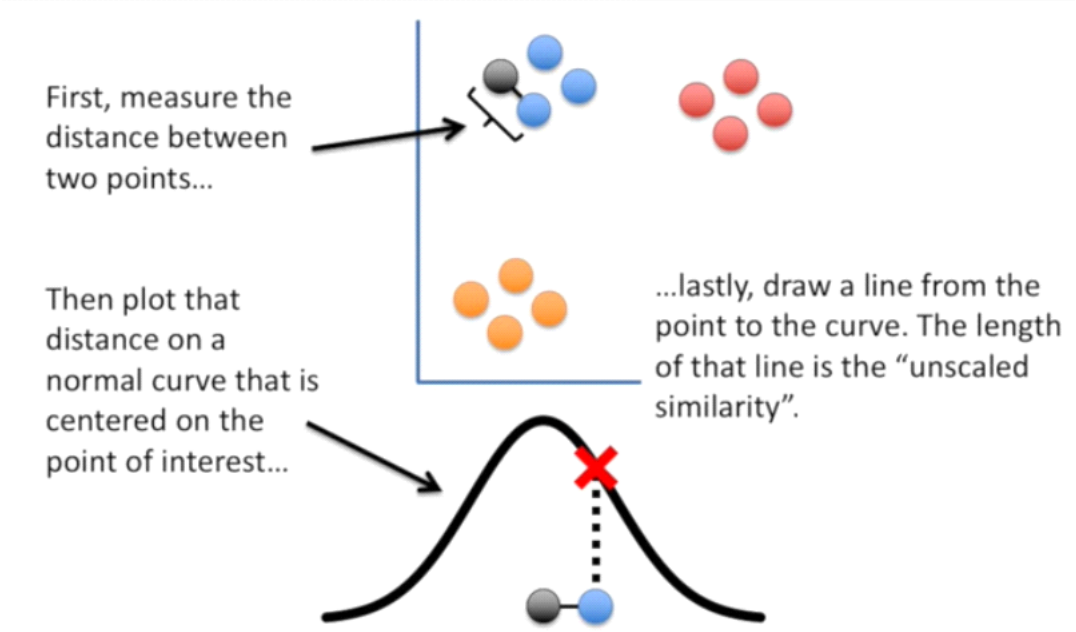
What t-SNE does is find a way to project data inot a low dimensional space (in this case, the 1-D number line) so that the clustering in the high dimensional space (in this case, the 2-D scatter plot) is preserved.



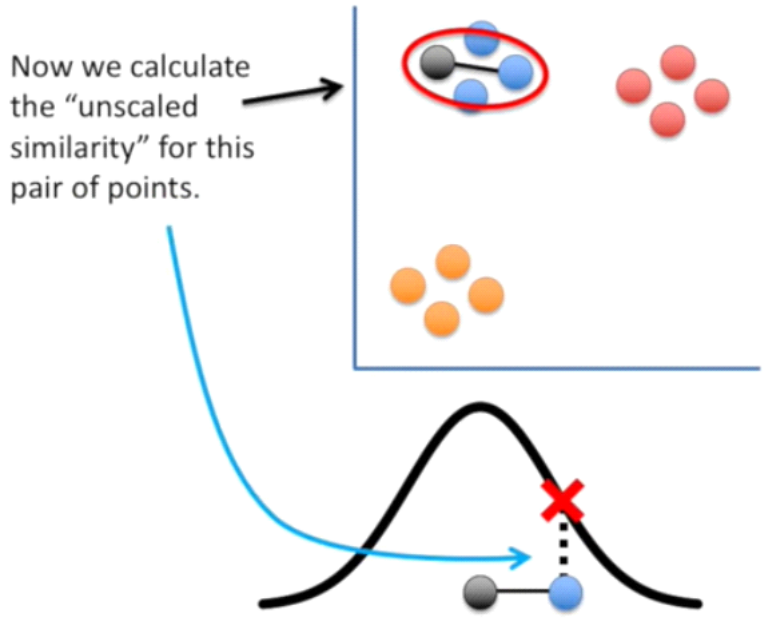
Ritaglio schermata acquisito: 26/08/2021 11:05

From here on out, t-SNE moves these points, a little bit at a time,, until it has clustered them.

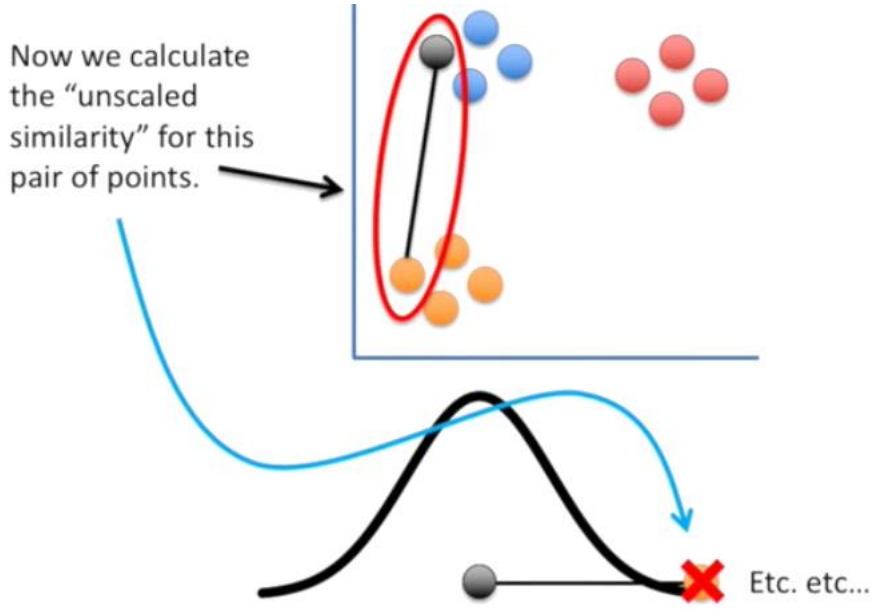
Step 1 -> Determine the 'similarity' of all the points in the scatter plot.



Ritaglio schermata acquisito: 26/08/2021 11:08

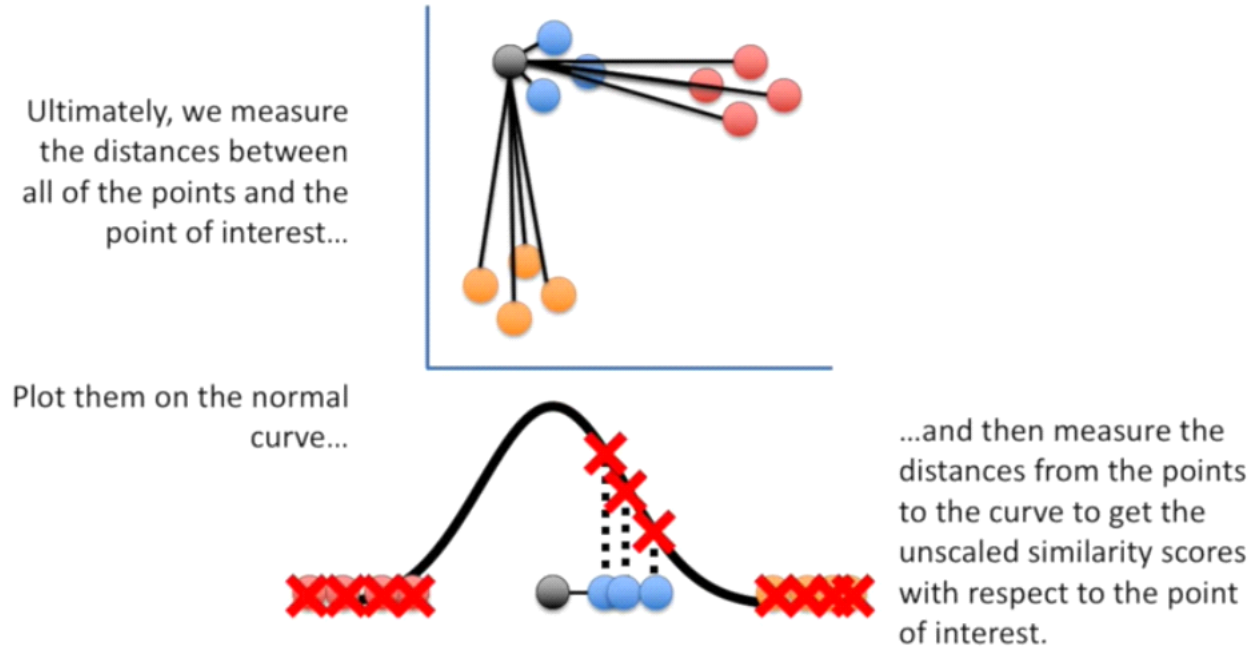


Ritaglio schermata acquisito: 26/08/2021 11:09



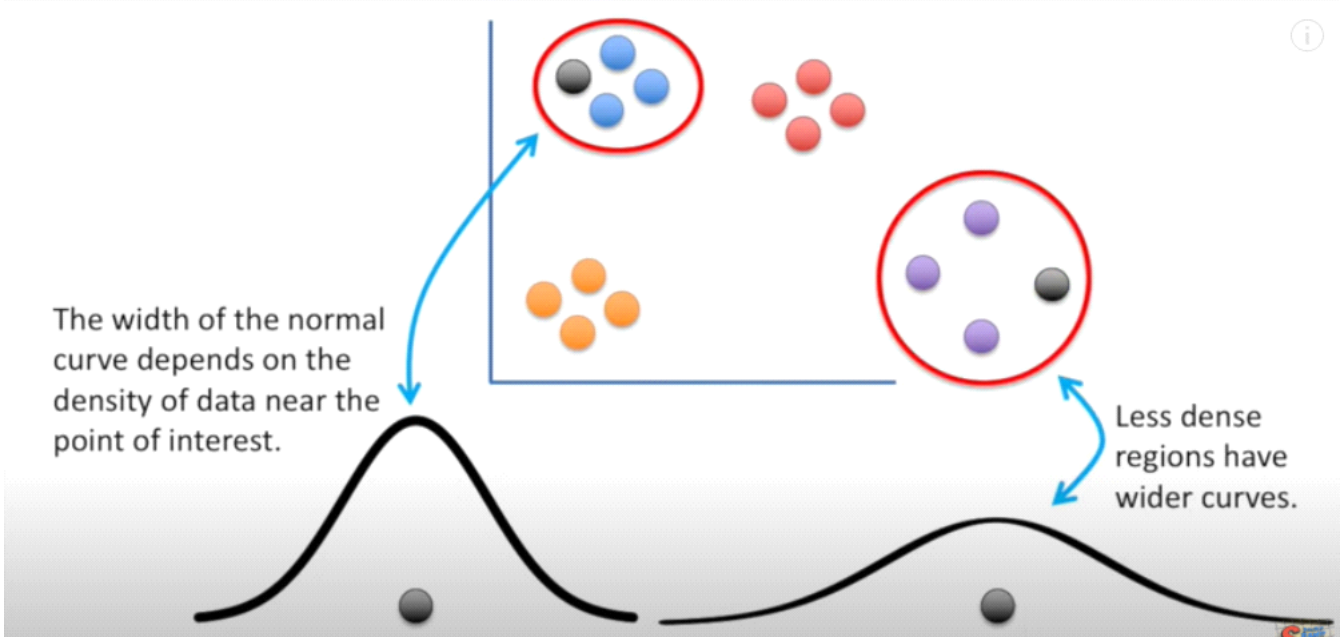
Ritaglio schermata acquisito: 26/08/2021 11:09

Using a normal distribution means that distant points have very low similarity values, and close points have high similarity values.

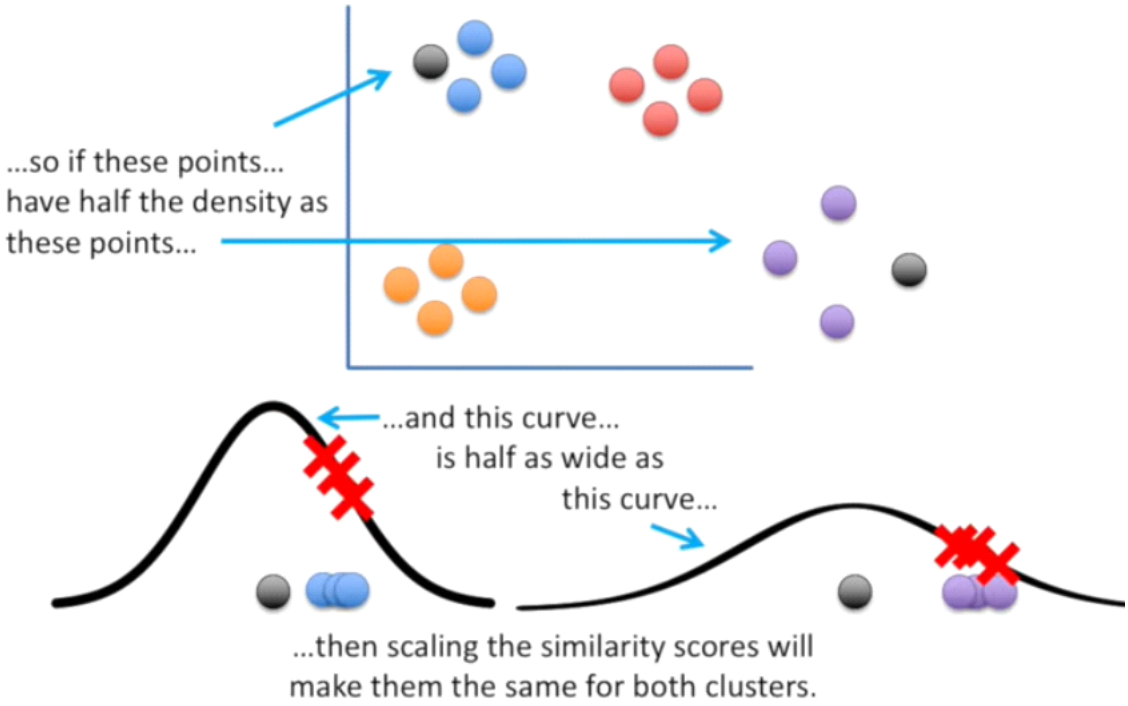


Ritaglio schermata acquisito: 26/08/2021 11:10

Then we need to scale the unscaled similarity.



Ritaglio schermata acquisito: 26/08/2021 11:11



Ritaglio schermata acquisito: 26/08/2021 11:12

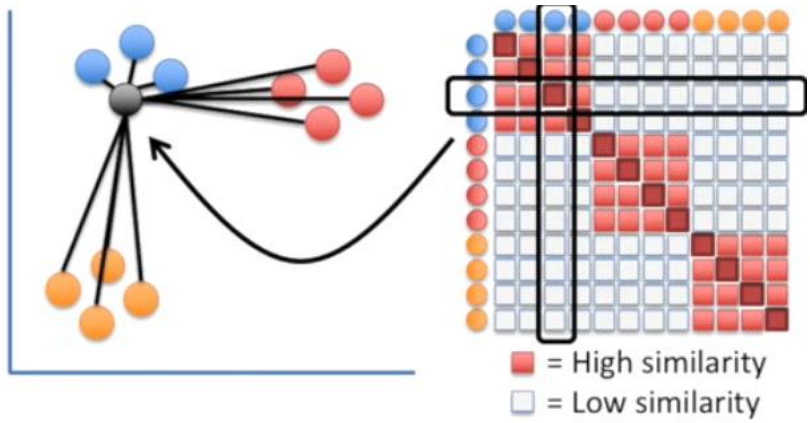
Thus, we need to scale the similarity scores so they sums to 1:
Scaled score = Score/Sum of all scores

However, this is an oversemplification. Indeed, t-SNE has a '**perplexity**' parameter equal to the expected density, and that comes into play, but these clusters are still more 'similar' than you might expect.

Coming back to the original example, we continue to compute the similarity scores for each point. However, it is important to mention that the similarity score for one point in cluster A to one point in

cluster B might be different from the similarity score of the point in cluster B and that point in cluster A due to the width of the distribution (which is based on the density of the surrounding data points).
So t-SNE just averages the two similarity scores from the two directions.

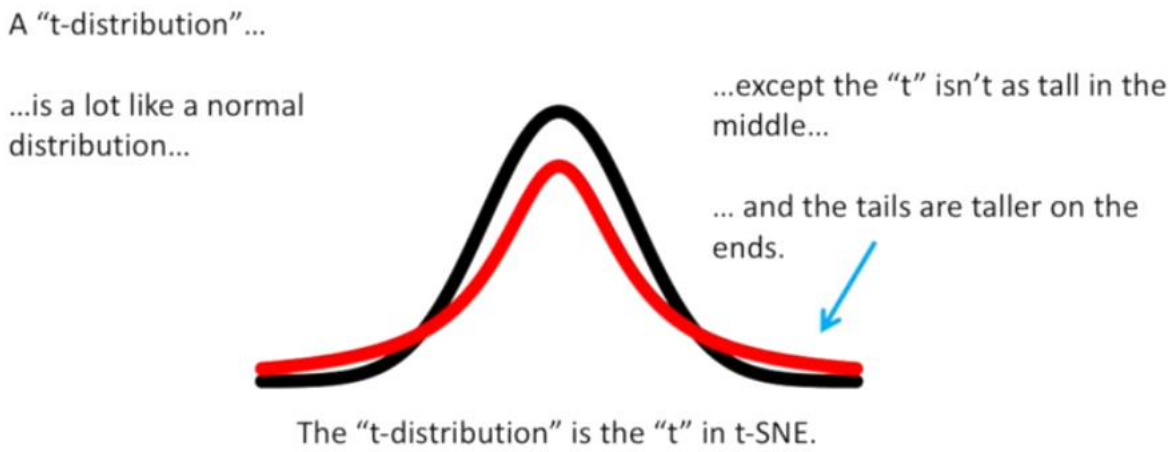
We end up with a matrix of similarity scores:



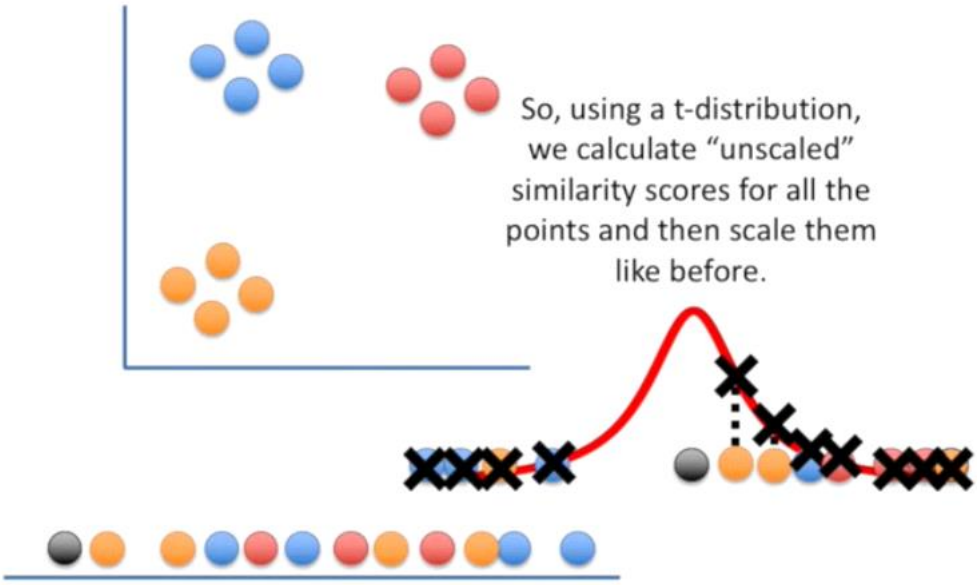
Ritaglio schermata acquisito: 26/08/2021 11:21

The dark red in the above matrix occurs when we compare an observation with itself. However, this notion of being similar to itself does not help the algorithm, thus in t-SNE this value is set to 0.

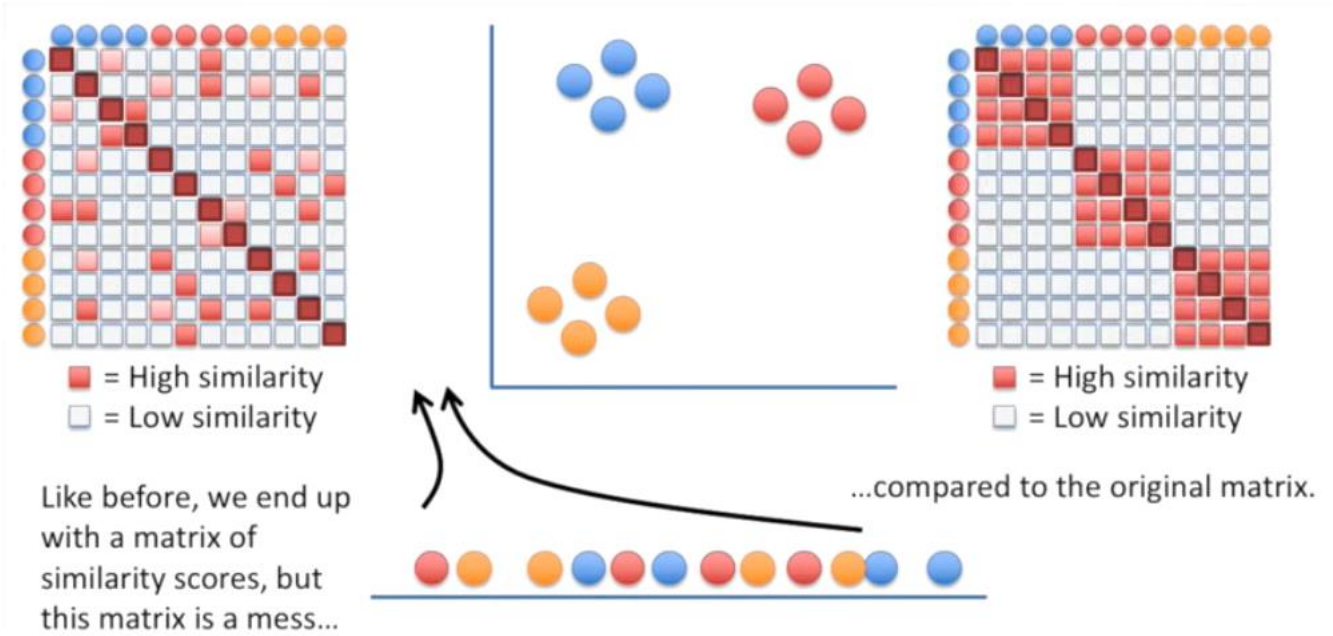
The next step is randomly project the data onto the number line and calculate similarity scores for the points on the number line.
However, now we do not use the normal distribution but the t-distribution.



Ritaglio schermata acquisito: 26/08/2021 11:25



Ritaglio schermata acquisito: 26/08/2021 11:25



Ritaglio schermata acquisito: 26/08/2021 11:25

Thus, we want that the left matrix is as close as possible to the one on the right.
-> t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.

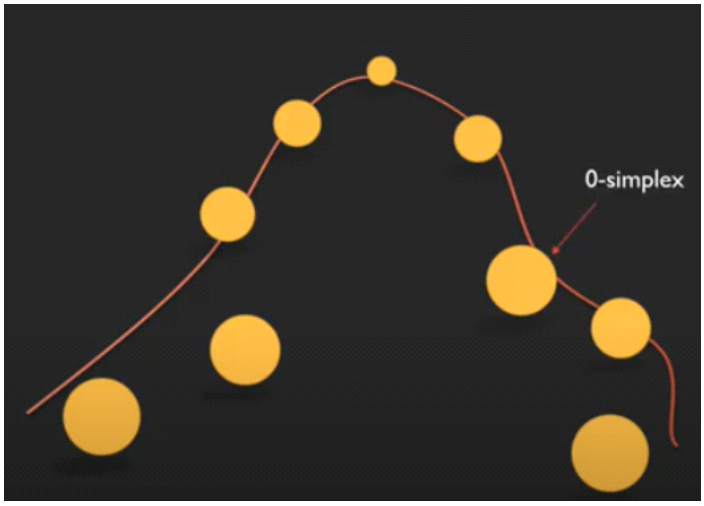
But why we use the t-distribution? It is used because without it the clusters would all clump up in the middle and be harder to see.

UMAP

stands for Uniform Manifold Approximation and Projection.

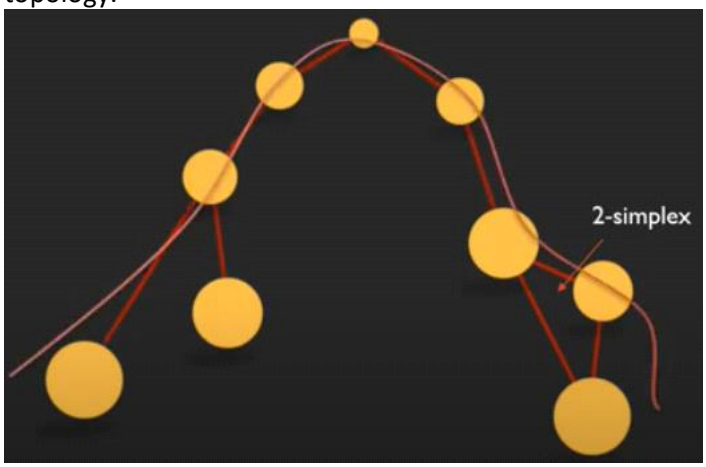
UMAP is basically done in two main steps: the graph construction (high dimension) and the graph projection (low dimension).
The cool part of UMAP is that its steps are mathematically proven to work.

Firstly, we have the data in the high dimensional space and we want to approximate their shape/topology. Each point is called 0-simplex.



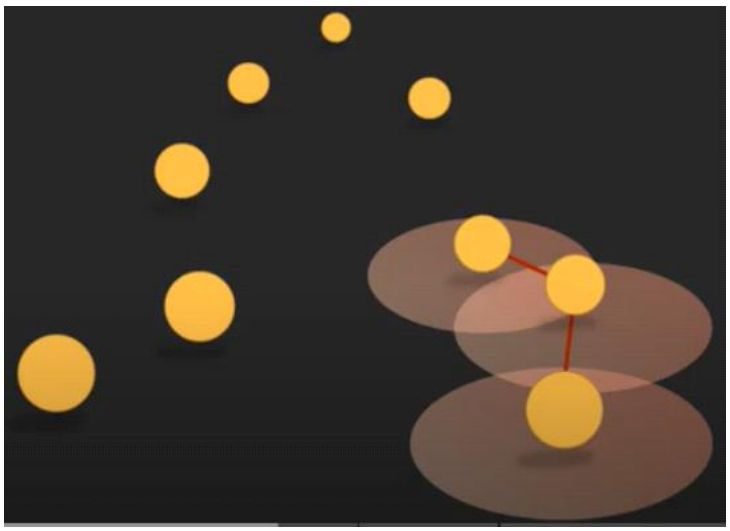
Ritaglio schermata acquisito: 26/08/2021 11:35

Nerve theorem -> shape of data can be approximate when we connect the 0 simpleces with their neighbouring data point forming 1, 2 or higher simpleces. With this we can approximate the topology.



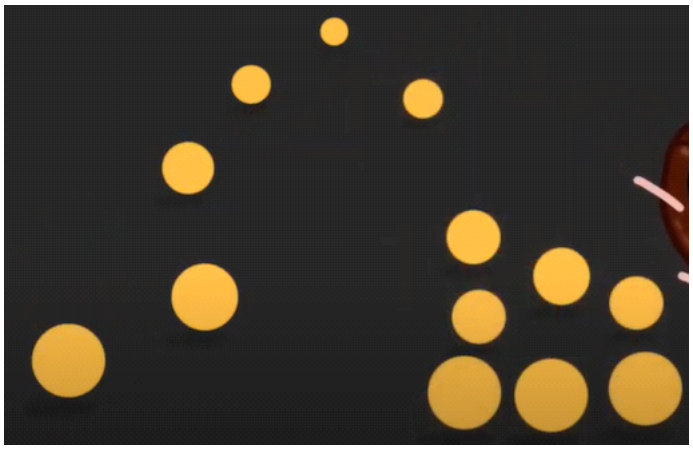
Ritaglio schermata acquisito: 26/08/2021 11:37

So what we need to do is making this connections. For this, the UMAP algorithm extends a radius around each point and makes a connection between each point and its neighbors eith intersecting radii.

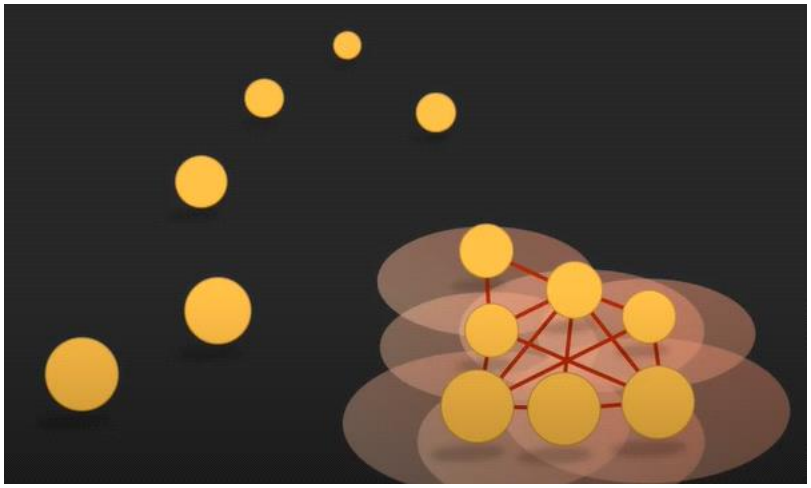


Ritaglio schermata acquisito: 26/08/2021 11:40

So far the radii are equal. But rember, we want to approximate the shape of the data, so we want a connected graph containing all our data points. But this whish brings in two problems:
Firstly, it often happens that in the data there are larger gaps, where there is no next point to connect to in the graph. This usually happens in low density regions.
Secondly, there are oftne high density regions where there are a lot of neighbors in the given radius and everything is way to connected.



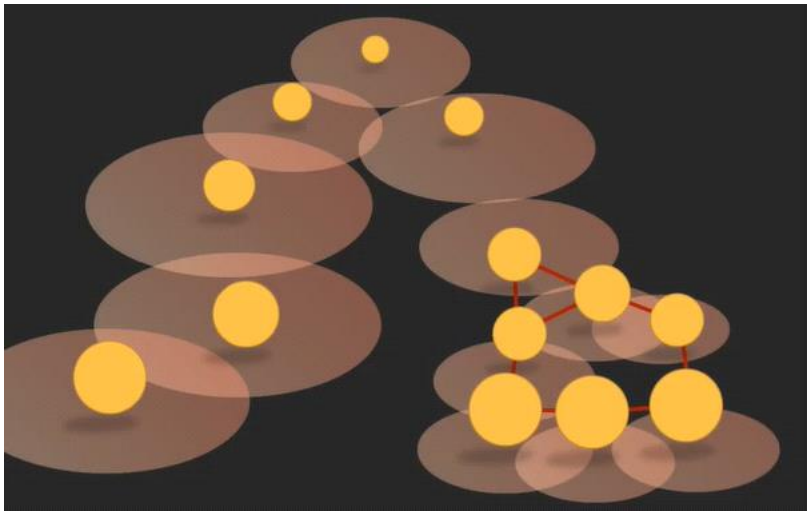
Ritaglio schermata acquisito: 26/08/2021 11:49



Ritaglio schermata acquisito: 26/08/2021 11:51

The second problem becomes even worse with the curse of dimensionality, where in high dimensional spaces the distances between points become more and more similar.

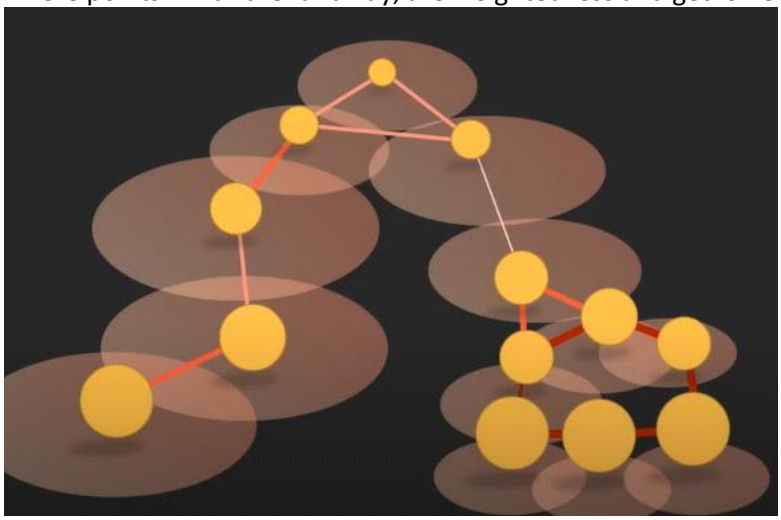
Thus, we can use instead of equal redius, variable radius.



Ritaglio schermata acquisito: 26/08/2021 11:53

This choice is also mathematically supported by the definition of a Riemannian metric on the manifold.
To recap, in small densiyty regions the radius is bigger while in high density regions the radius is smaller.
Moreover, UMAP does not estimate density directly as a number but uses a proxy: the density is estimated to be higher when the kth neighbor is close and lower when the kth nearest neighbor is far away.
As we know, the k is an hyperparameters that needs to be chosen.
K big -> more global structure is preserved
K small -> then the radius decreases and the local structure is more preserved.
So we need to wisely choose the k in order to achieve the perfect balance between the local and the global structure preservation.
Unfortunately, there is not a magic recipes that allows us to find the optimal k automatically -> when need the trial and error procedure.

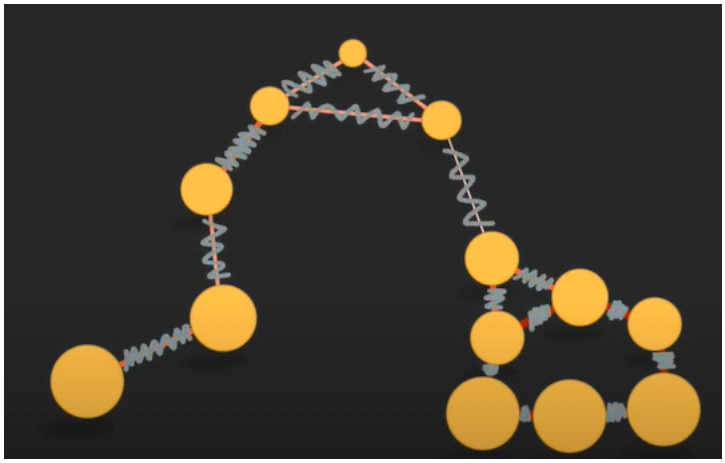
But not all k neigbors are equal, since each have different distances from the point we are looking at. Then the connections between each point and their neigbors get a weight, a connection probability, where points which are far away, are weighted less and get lower connection probability



Ritaglio schermata acquisito: 26/08/2021 12:03

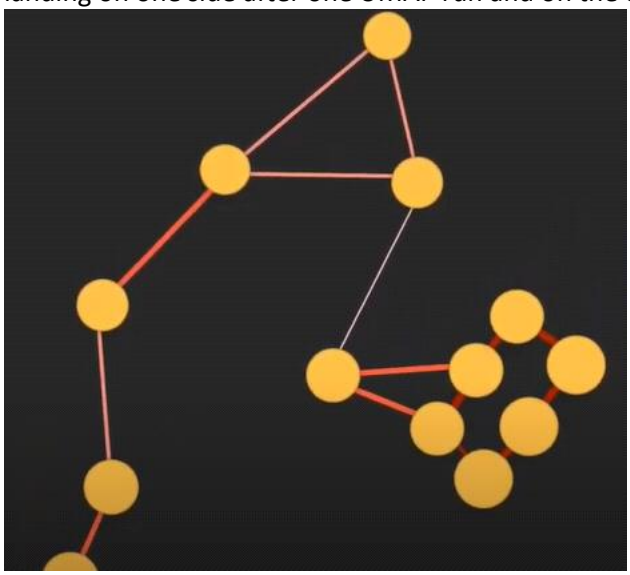
Now our high dimensional graph is constructed and it is ready to be projected to lower dimensions.

We can imagine this projection as taking the high-dimesional graph, with their edges being springs, where each spring is stronger as the edge probability increase.



Ritaglio schermata acquisito: 26/08/2021 12:27

Which means that points connected by high weighted edges are more likely to stay together in the lower dimensional space, because the spring holds these points together. And perhaps interesting to notice is that these spring forces are rotationally rymmetric which leads to clusters sometimes landing on one side after one UMAP run and on the other side after another projection



Ritaglio schermata acquisito: 26/08/2021 12:36

So compared to t-SNE, UMAP:
is faster due to its optimizations and strong mathematical foundations;
Has a better balance between locality and globality in clustering;

UMAP excels at reducing from a lot of dimensions.

Interesing additional reading: https://pair-code.github.io/understanding-umap/