INST0065
Data Visualization and GIS

Week 2: Data types and visualization types, and, a history of visualization

Dr. Oliver Duke-Williams

o.duke-williams@ucl.ac.uk
(Please use Moodle forums for messages about this module)

Twitter: @oliver_dw

INFORMATION STUDIES

Celebrating 100 years

Contents

- Types of data
- Types of visualization
- A brief history of data visualization

Types of data, types of story

- This section based on:
  Few, S. (2004). *Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message*.
  http://www.perceptualedge.com/articles/ie/the_right_graph.pdf

Data types

- In order to plan a visualization, we need to recognise different types of data
- The primary division is between quantitative and categorical data
  - Quantitative data are numerical
  - Categorical data have labels (sometimes partially numeric)

Examples

Quantitative	Categorical
Height	Nationality
Weight	Day of the week
Age	Whether vowel/consonant
Temperature	Gender
Price	Parts of speech

Sub-divisions of categorical data

- Nominal
  - Categories have labels but have no inherent order (gender, lemma (linguistics))
- Ordinal
  - Categories have an intrinsic order (day of the week, gold/silver/bronze medals)
- Interval
  - Categories converted from numerical data
  - (age 0-4, 5-9, 10-14...)

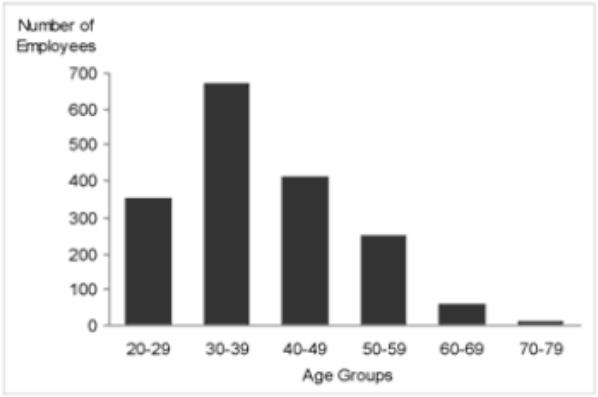
It is not always clear how to define these

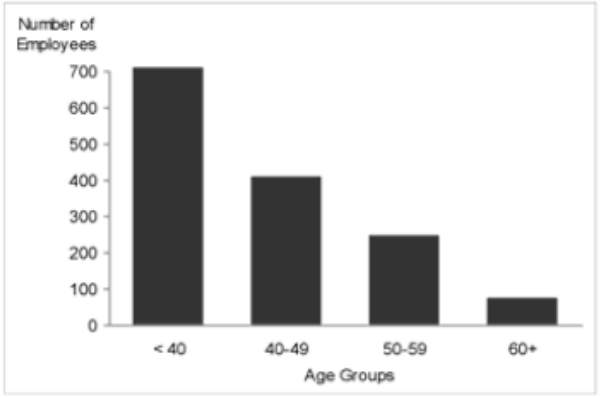
- Colour – usually nominal, but consider wavelength spectra
- Days, months etc – usually ordinal, but need definition: when does the week start?
  - For data display: finite set
  - For data analysis: may be cyclical (Jan, Feb,...,Nov,Dec,Jan)

Dichotomous data

- Dichotomous data are nominal data that can have exactly two possible states e.g. agree/disagree
  - Often framed as: 1 – has characteristic; 0 – does not have characteristic
  - Nominal data with >2 categories might be presented for analysis as a set of dichotomous variables
  - Sometimes called binary data or binomial data; these terms are also used for other things

Intervals should be carefully considered





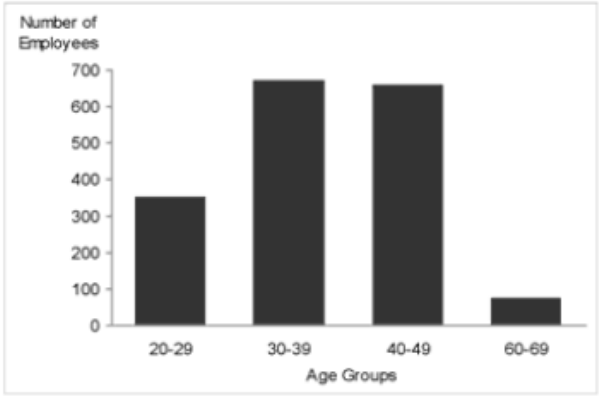


Figure 2: Changing interval scales can alter the data and result in misinformation.

Sub-divisions of quantitative data

- Continuous vs discrete
  - Discrete data have fixed units; e.g. persons
  - Continuous data can be divided into smaller units: e.g. time

Sub-divisions of quantitative data

- Interval v ratio
  - Interval data: differences between units are the same
    - e.g. temperature
  - Ratio data is interval data with a natural zero
    - e.g. time, distance
    - Temperature is not ratio data: 20°C is not twice as hot as 10°C (consider what happens if we quote temperature in °F)

Individual vs aggregate observations

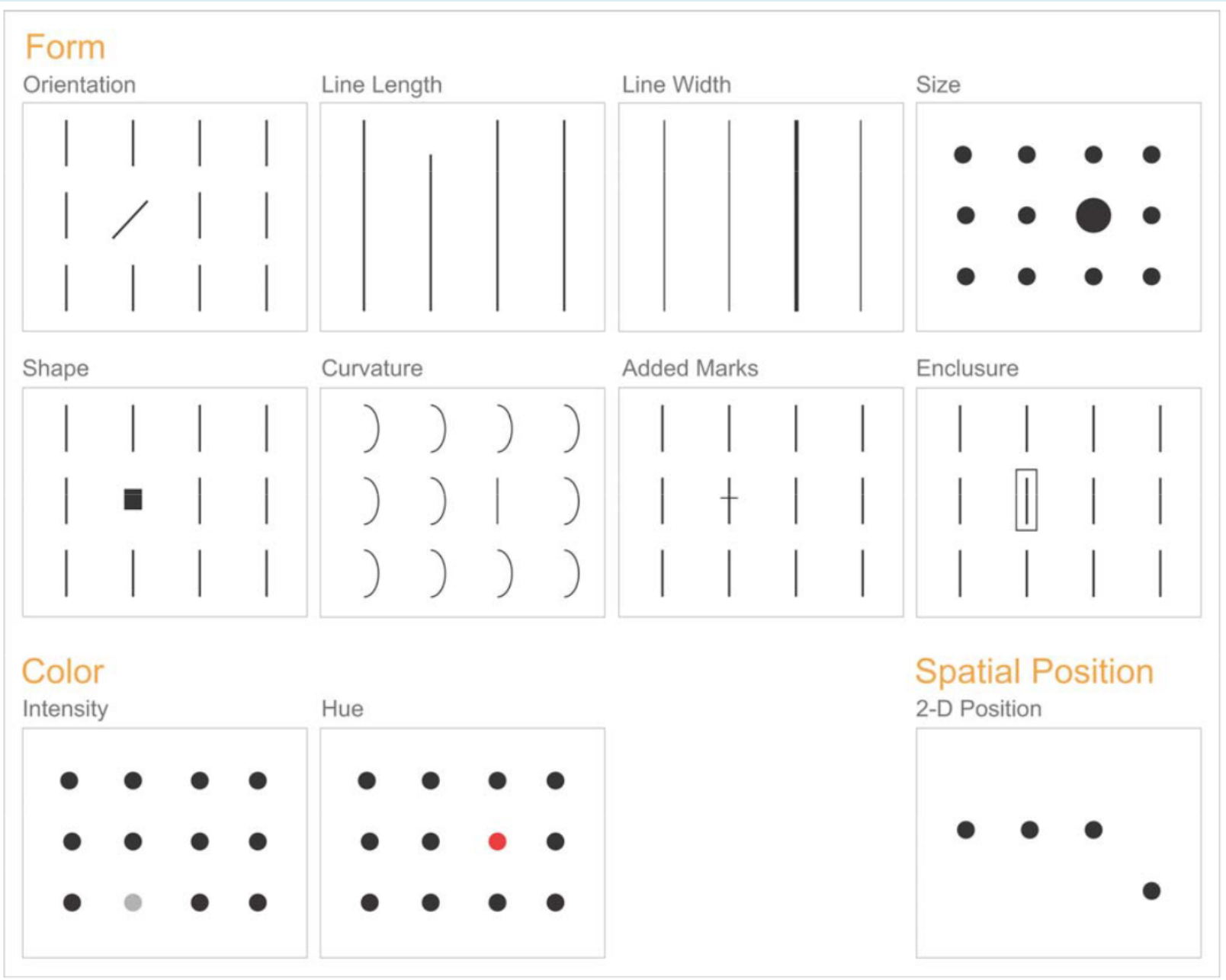
- Individual records will typically have both categorical and quantitative variables
- As we group these, we will introduce the fundamental quantitative variable 'count' or 'number'

Visual attributes

- Few (2004) identifies a set of visual attributes
  - 2-D location
  - Line length
  - Size
  - Shape
  - Orientation
  - Colour

Visual attributes

- A slightly fuller set are described in another paper (Few, 2004b)
  - Form
    - Orientation, line length, line width, size, shape, curvature, added marks, enclosure
  - Colour
    - Intensity, hue
  - Position
    - x,y


Form
Orientation
Line Length
Line Width
Size
Shape
Curvature
Added Marks
Enclosure
Color
Intensity
Hue
Spatial Position
2-D Position

Should we be limited to visual attributes?

- Data 'visualization' tautologically promotes visual presentation
- Data can be presented in other ways
  - Data sonification
  - Haptic data

Visual attributes

- Of these attributes, Few states:

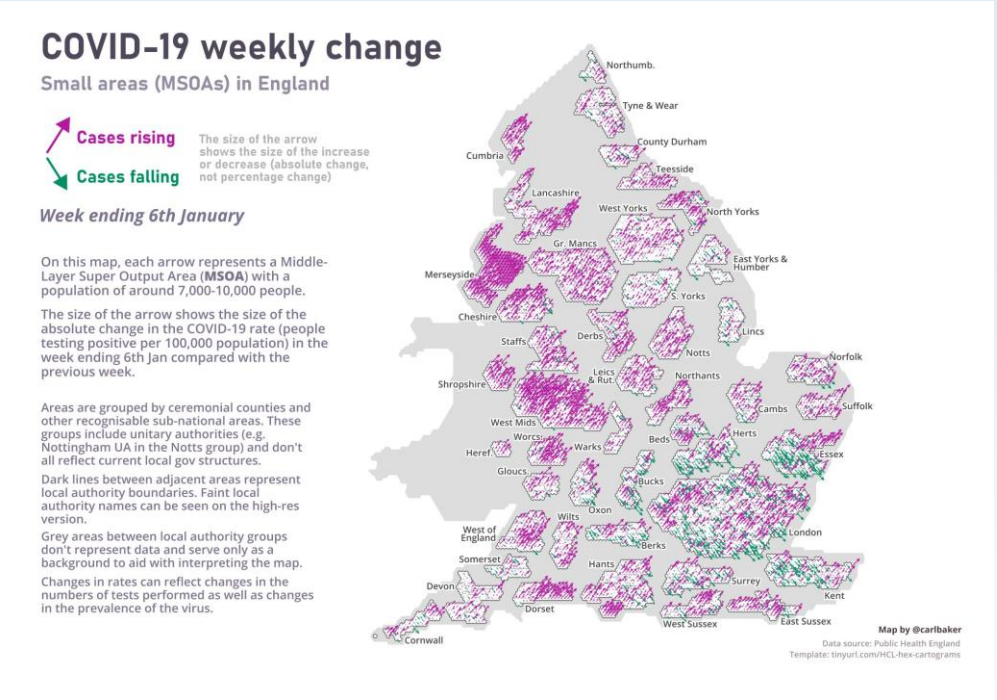
  "Of the full list of these attributes, only two emerge as highly effective means to visually encode quantitative values: 2-D location and line length."

  Do you agree?

Orientation

- Orientation, as one example, might be seen as either:
  - a distinct characteristic
  - a line formed by a pair of 2D locations

- Value is shown by orientation as well as be colour



Source: https://twitter.com/carlbaker/status/1348933992100188162

COVID-19 weekly change

Small areas (MSOAs) in England

Cases rising

Cases falling

The size of the arrow shows the size of the increase or decrease (absolute change, not percentage change)

Week ending 6th January

On this map, each arrow represents a Middle-Layer Super Output Area (**MSOA**) with a population of around 7,000-10,000 people.

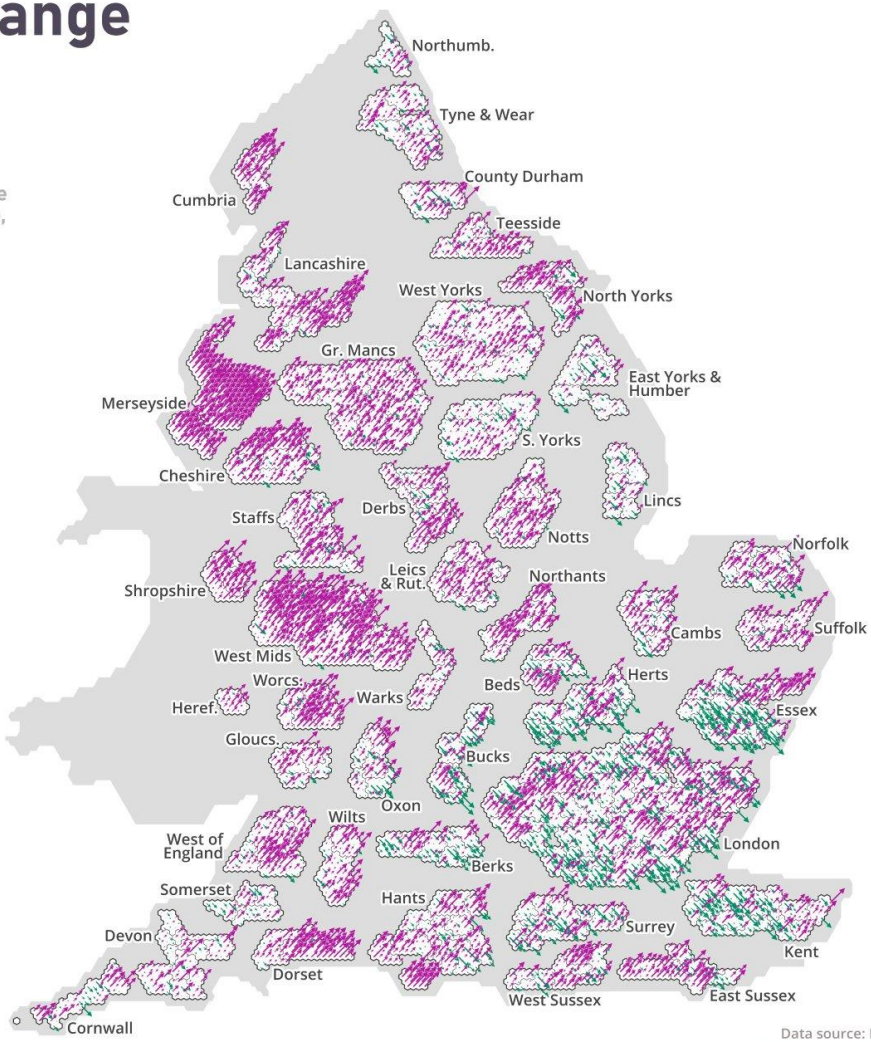
The size of the arrow shows the size of the absolute change in the COVID-19 rate (people testing positive per 100,000 population) in the week ending 6th Jan compared with the previous week.

Areas are grouped by ceremonial counties and other recognisable sub-national areas. These groups include unitary authorities (e.g. Nottingham UA in the Notts group) and don't all reflect current local gov structures.

Dark lines between adjacent areas represent local authority boundaries. Faint local authority names can be seen on the high-res version.

Grey areas between local authority groups don't represent data and serve only as a background to aid with interpreting the map.

Changes in rates can reflect changes in the numbers of tests performed as well as changes in the prevalence of the virus.

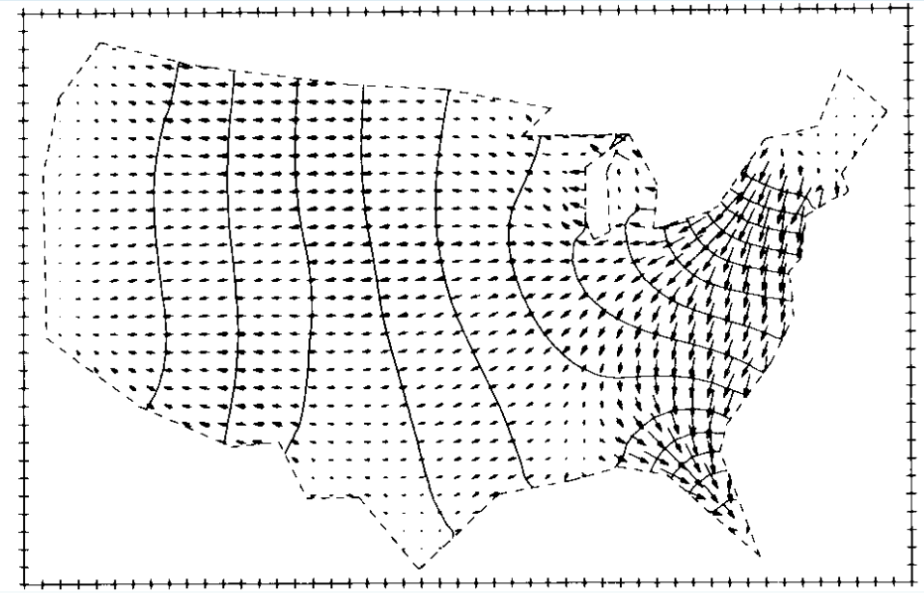


Map by @carlbaker

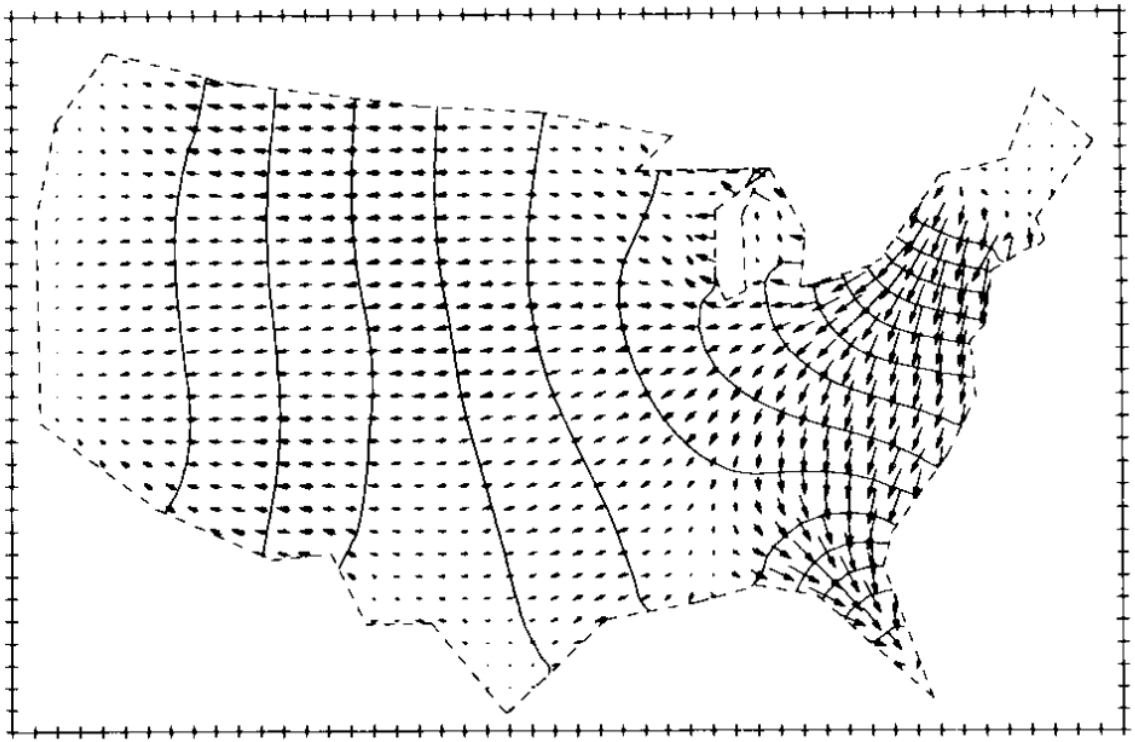
Data source: Public Health England

Template: tinyurl.com/HCL-hex-cartograms

- Value is shown solely by orientation



Source: Tobler (1981)

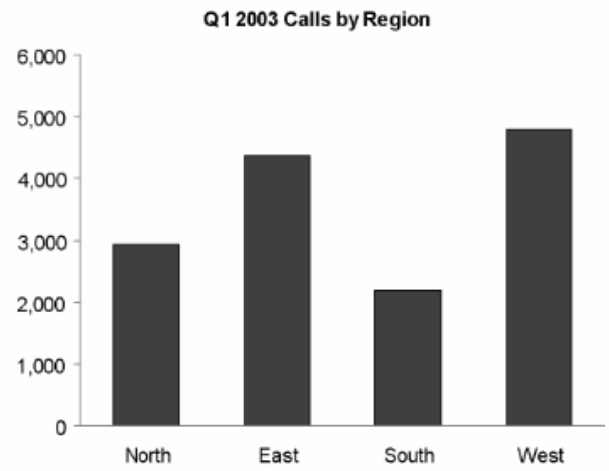


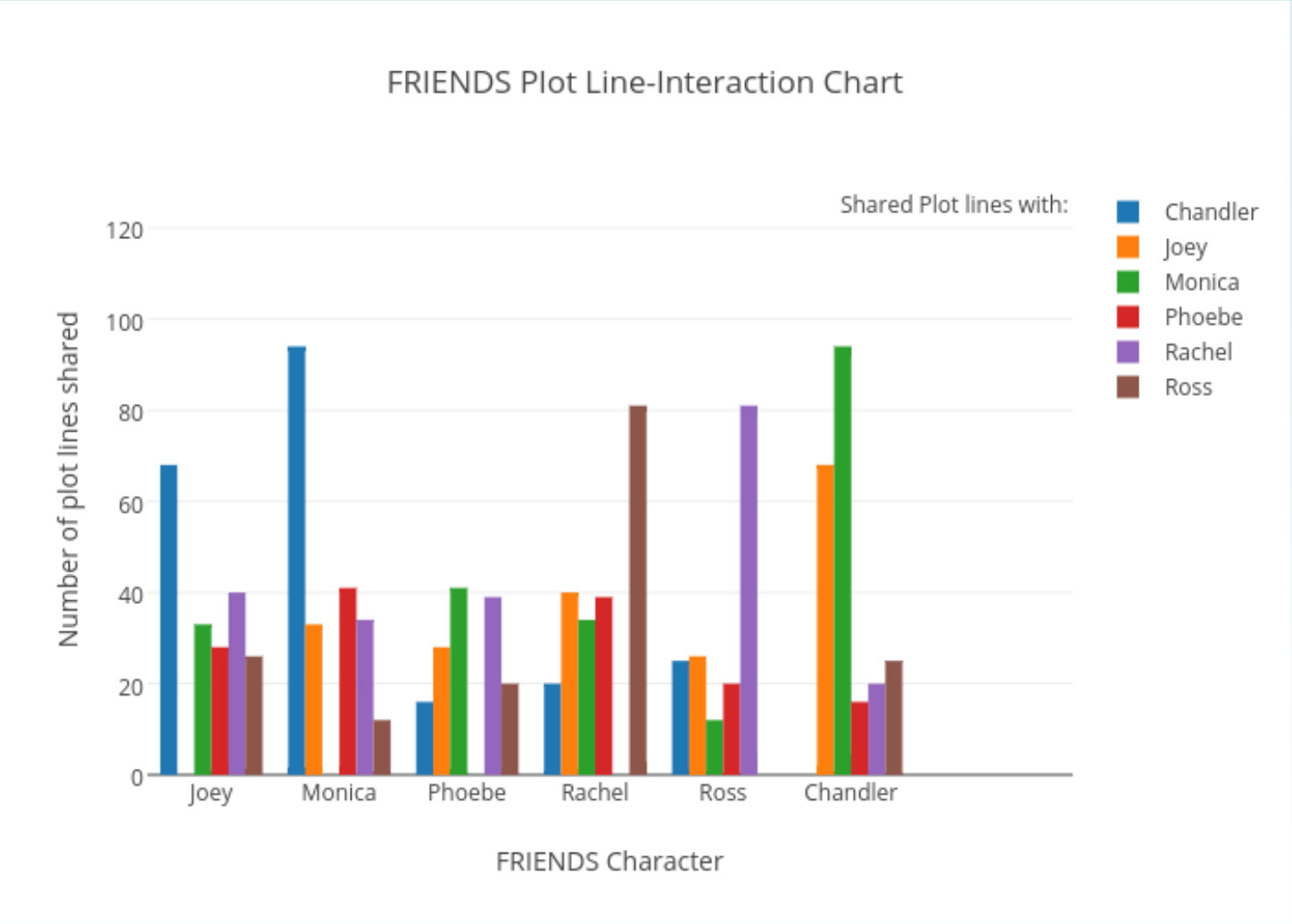
MAP 10. Estimated Field of the 1965/1970 Net Population Flow. Obtained as the solution to Poisson's equation. The potential field is indicated by contour lines and the flow is shown by the gradient vectors.

Types of quantitative messages (Few, 2004)

- Nominal comparison
- Time-series
- Ranking
- Part-to-whole
- Deviation
- Frequency distribution
- Correlation

Nominal comparison

Nominal Comparison A simple comparison of the categorical subdivisions of one or more measures in no particular order	• Bars only (horizontal or vertical)	

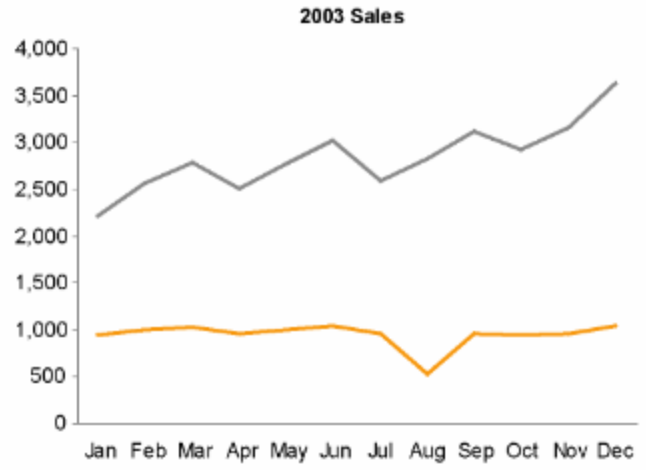


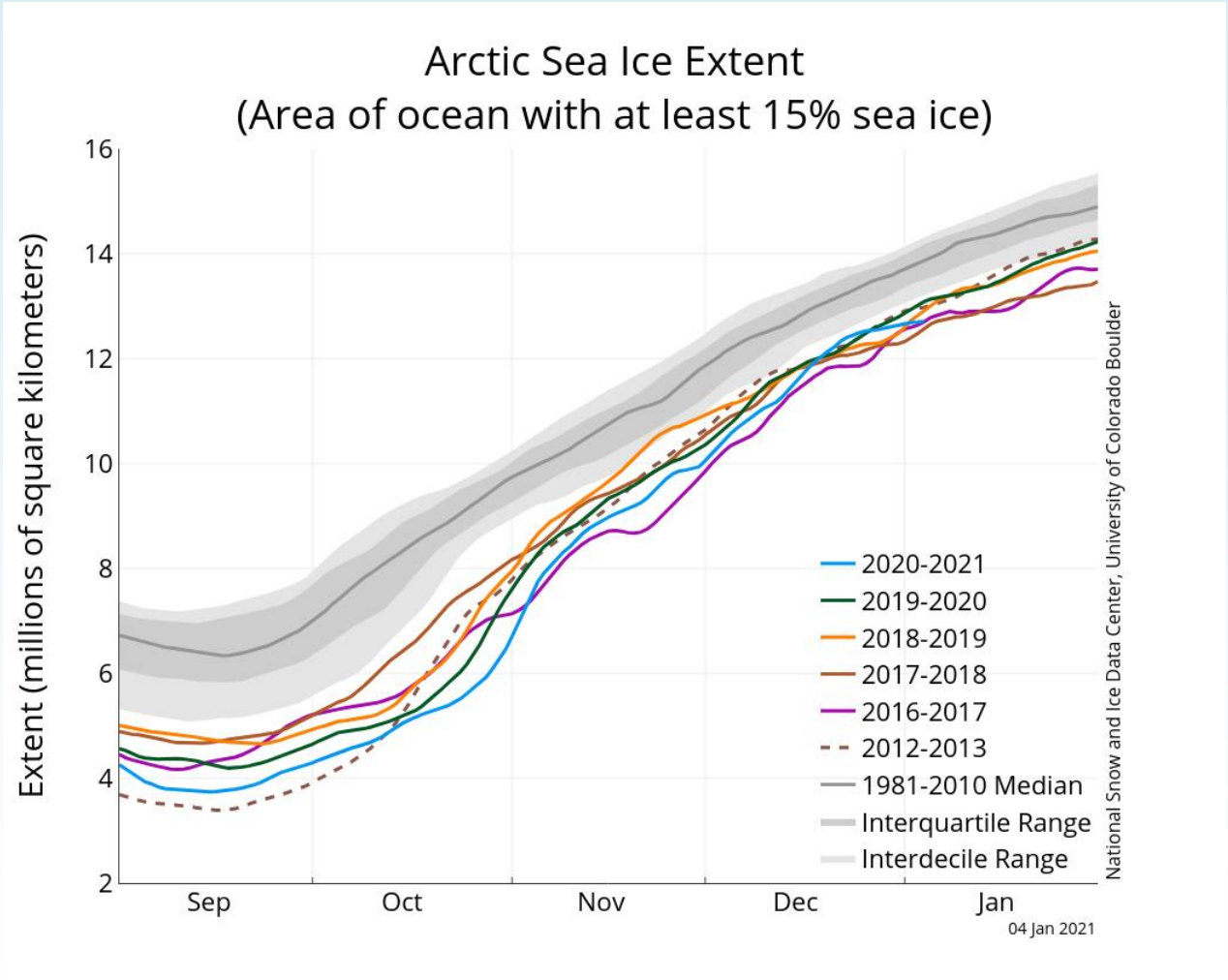
Source: https://chart-studio.plotly.com/~kirktjared/117.embed

Time series

Time Series
Multiple instances of one or more measures taken at equidistant points in time

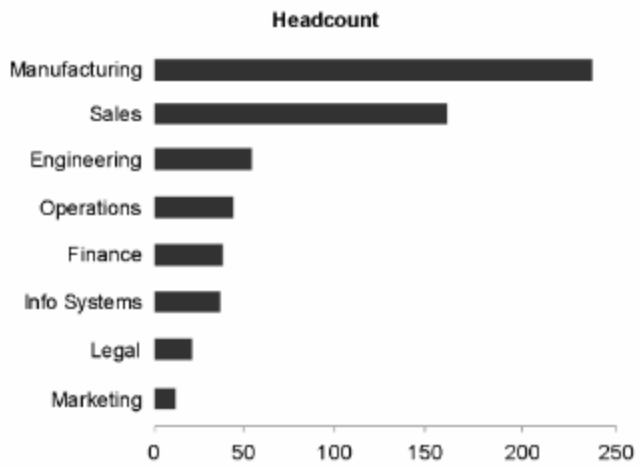
- Lines to emphasize overall pattern
- Bars to emphasize individual values
- Points connected by lines to slightly emphasize individual values while still highlighting the overall pattern
- Always place time on the horizontal axis

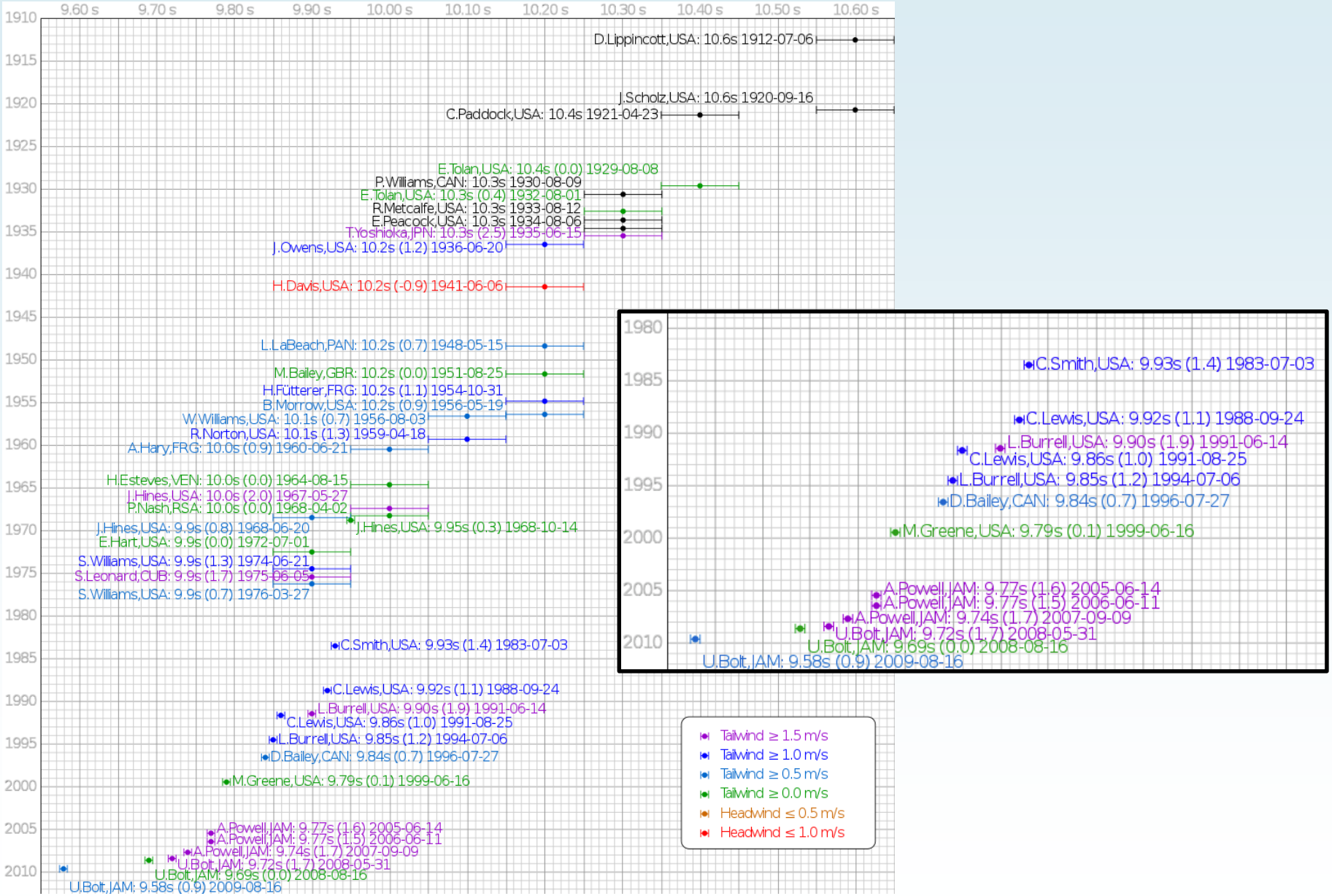




Source: http://nsidc.org/arcticseaicenews/

Ranking

Ranking Categorical subdivisions of a measure ordered by size (either descending or ascending)	• Bars only (horizontal or vertical) • To highlight high values, sort in descending order • To highlight low values, sort in ascending order	

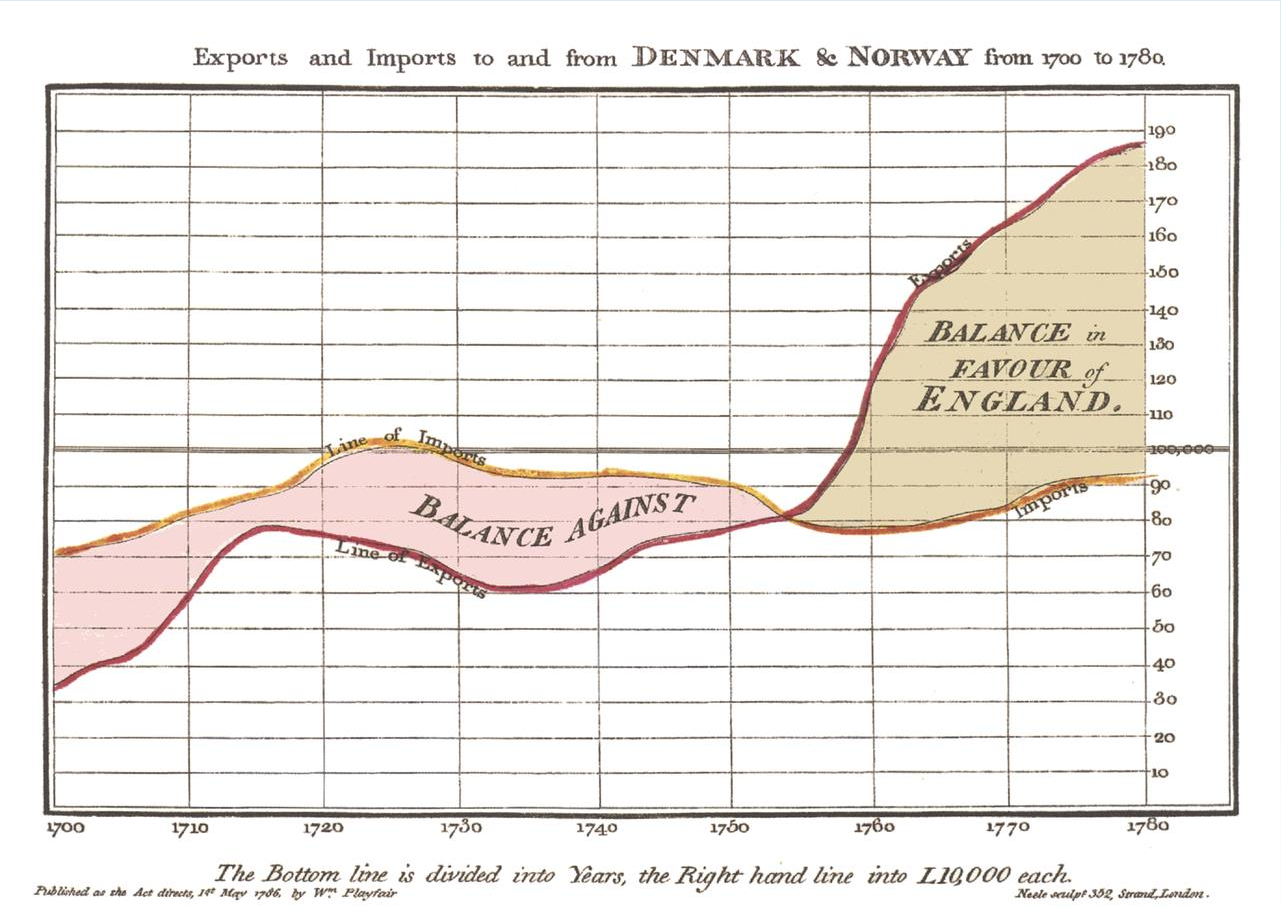


Source: https://en.wikipedia.org/wiki/Men%27s_100_metres_world_record_progression#/media/File:World_record_progression_100m_men.svg

Deviation

Deviation
Categorical subdivisions of a measure compared to a reference measure, expressed as the differences between them

- Lines to emphasize the overall pattern only when displaying deviation and time-series relationships together
- Points connected by lines to slightly emphasize individual data points while also highlighting the overall pattern when displaying deviation and time-series relationships together
- Bars to emphasize individual values, but limit to vertical bars when a time-series relationship is included
- Always include a reference line to compare the measures of deviation against

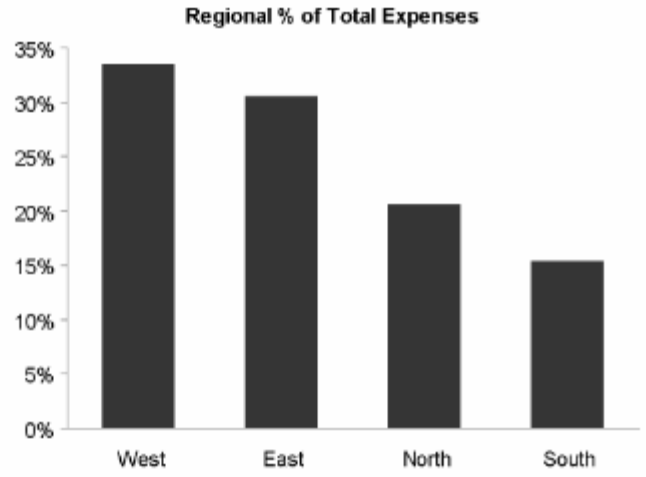


Source (original): Playfair, W (1786)

Part-to-whole

Part-to-Whole
Measures of individual categorical subdivisions as ratios to the whole

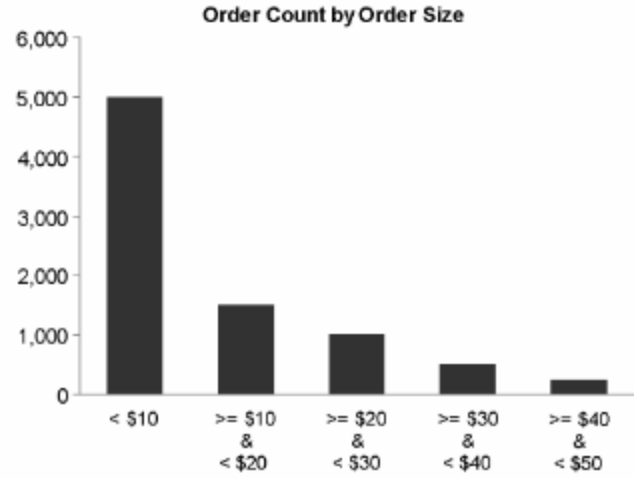
- Bars only (horizontal or vertical)
- Use stacked bars only when you must display measures of the whole as well as the parts



Frequency distribution

Frequency Distribution
Counts of something per categorical subdivisions (intervals) of a quantitative range

- Vertical bars to emphasize individual values (called a *histogram*)
- Lines to emphasize the overall pattern (called a *frequency polygon*)



Variable Description | Representation

Name us5_e_indresp.e_libfreq
Label library frequency

Value	Label	Frequency	% of valid
-9	missing	0	
-8	inapplicable	28,731	
-7	proxy	3,856	
-2	refusal	0	
-1	don't know	3	
1	at least once a week	1,693	13.84%
2	less often than once a week but at least once a month	4,011	32.78%
3	less often than once a month but at least 3 or 4 times a yea	3,786	30.94%
4	twice in the last 12 months	1,847	15.09%
5	once in the last 12 months	899	7.35%

Valid
12236

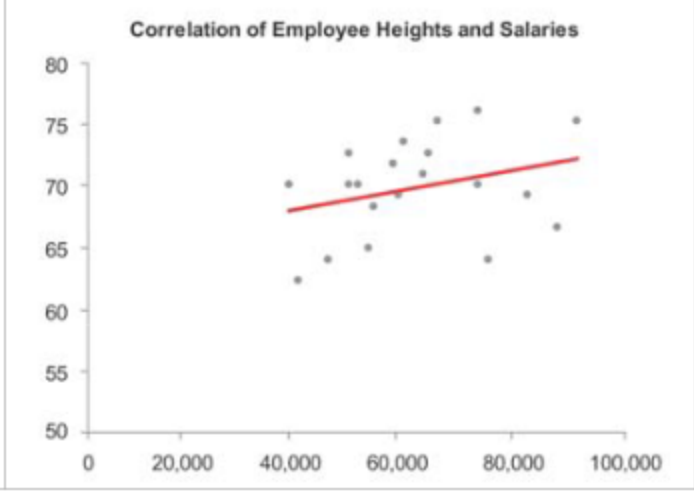
at least once a week	1,693	13.84%
less often than once a week but at least once a month	4,011	32.78%
less often than once a month but at least 3 or 4 times a yea	3,786	30.94%
twice in the last 12 months	1,847	15.09%
once in the last 12 months	899	7.35%

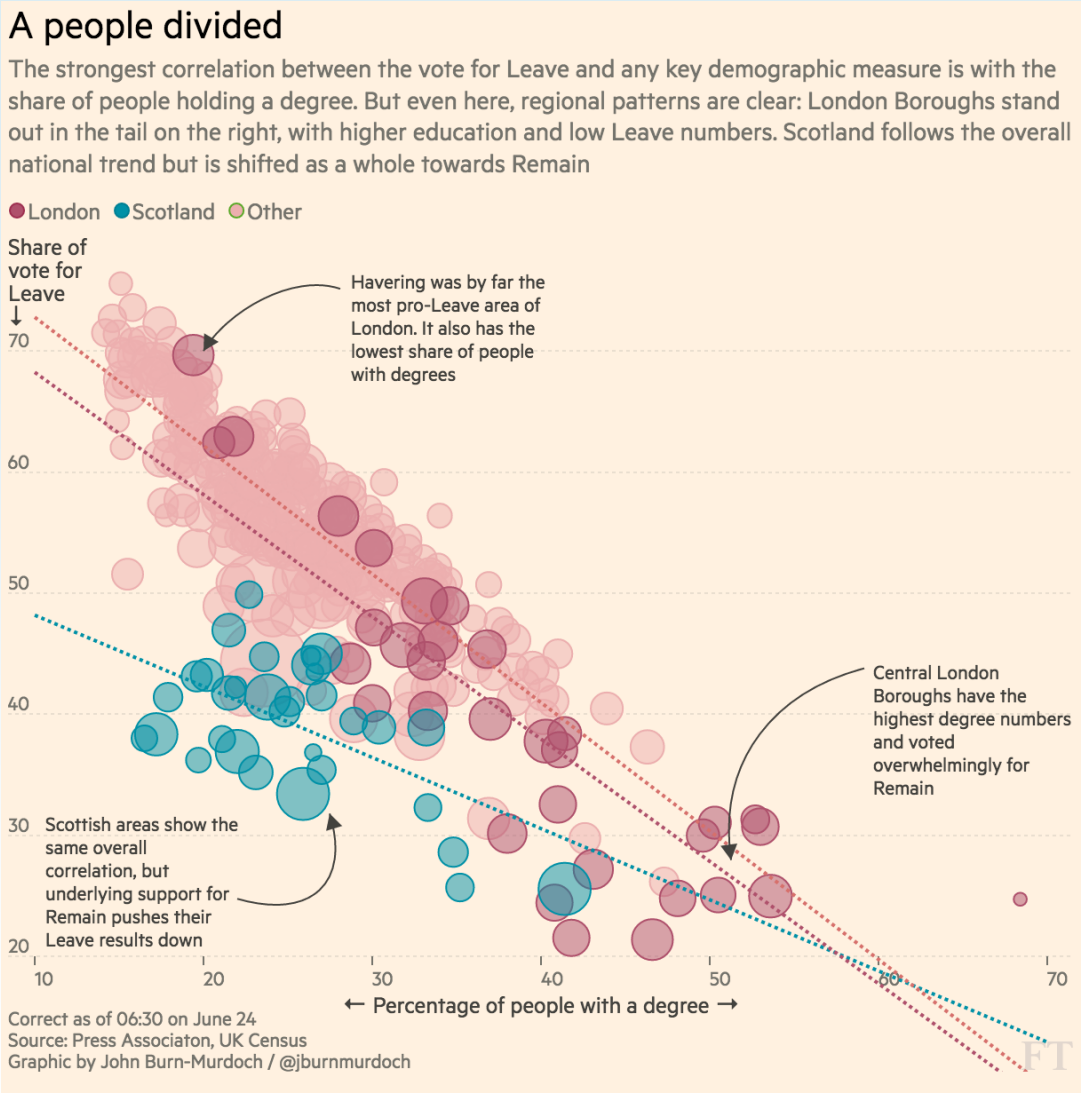
Source: https://discovery.closer.ac.uk/item/uk.iser.ukhls/c88edab1-d82a-4ef3-8634-38cfb95e2343

Correlation

Correlation
Comparisons of two paired sets of measures to determine if as one set goes up the other set goes either up or down in a corresponding manner, and if so, how strongly

- Points and a trend line in the form of a scatter plot
- Bars may be used, arranged as a *paired bar graph* or a *correlation bar graph*, if scatter plots are unfamiliar
- (Note: For descriptions of these graphs, see my book *Show Me the Numbers*.)





Source: Financial Times, via:
https://stats.stackexchange.com/questions/260238/how-should-this-bbc-chart-brexit-correlation-between-education-and-results-hav

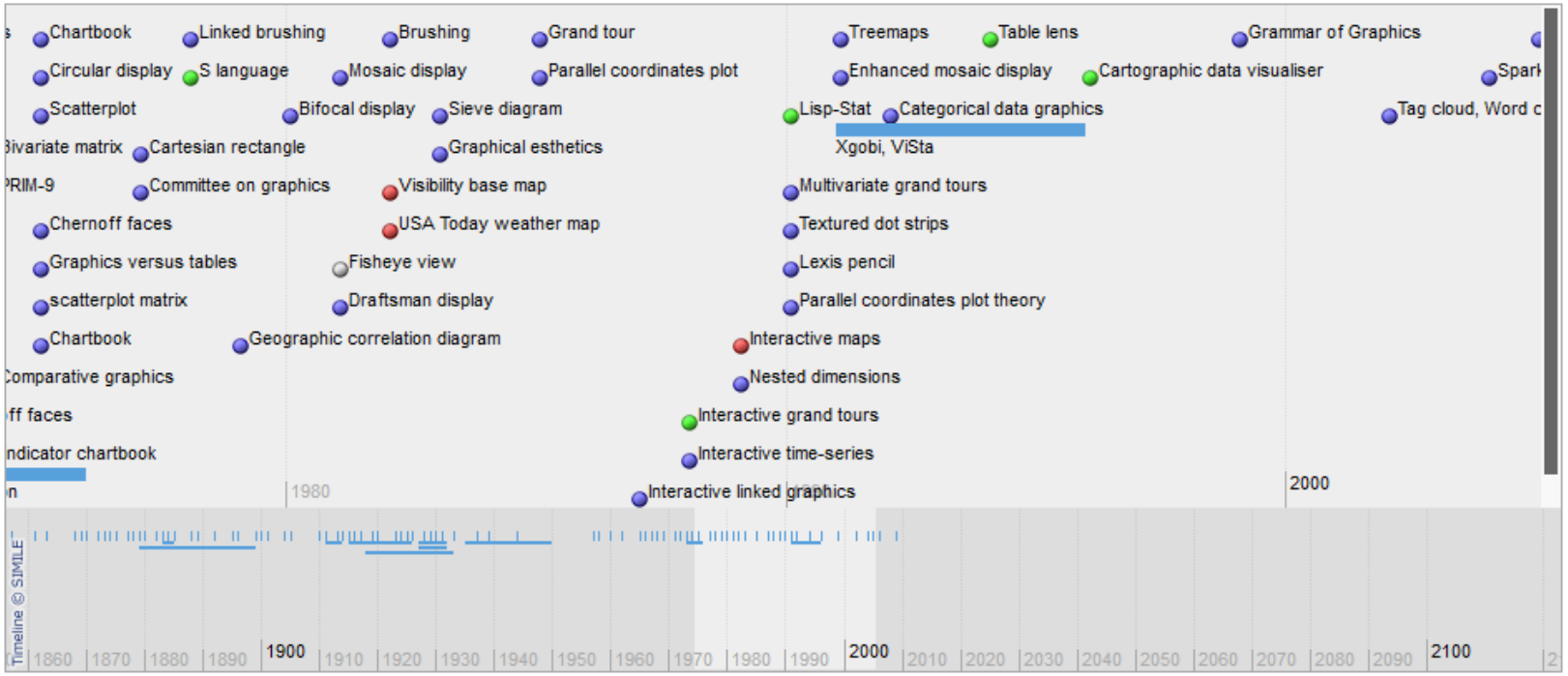
The history of data visualization

- This section identifies some key developments in the history of data visualization
- It is based on:

  Friendly M. (2005) Milestones in the History of Data Visualization: A Case Study in Statistical Historiography. In: Weihs C., Gaul W. (eds) Classification — the Ubiquitous Challenge. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-28084-7_4

The Milestones website

- https://www.datavis.ca/milestones/



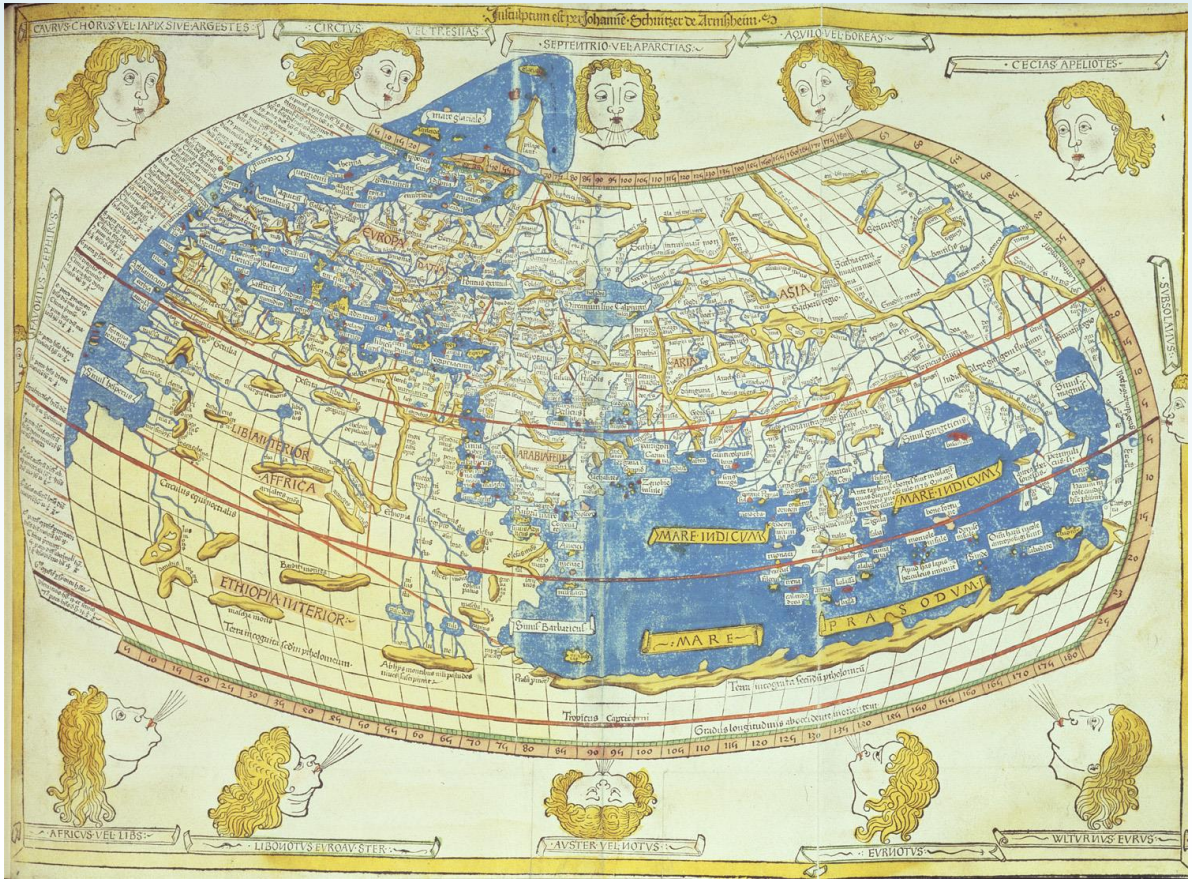
Peutinger map, 13th century

- Based on early maps, possibly dating to c 4th century BC

Source: https://www.datavis.ca/milestones/index.php?group=pre-1600

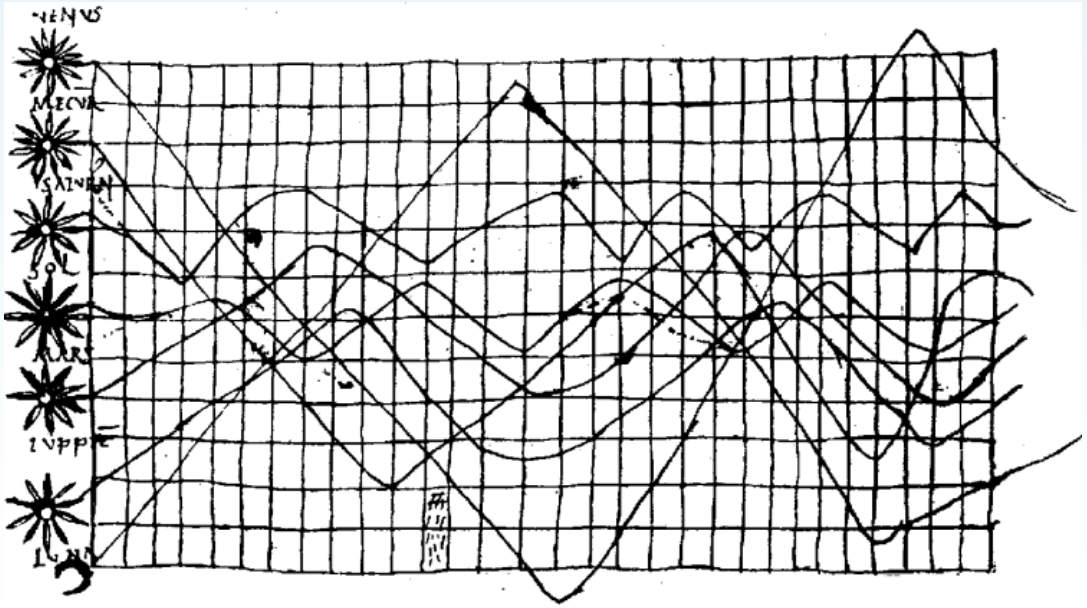
Ptolemy's map, c.150

- Uses latitude and longitude



Unknown, c. 950

- "Earliest known attempt to show changing values graphically (positions of the sun, moon, and planets throughout the year)"



Source: https://www.datavis.ca/milestones/index.php?group=pre-1600#

Langren, 1644

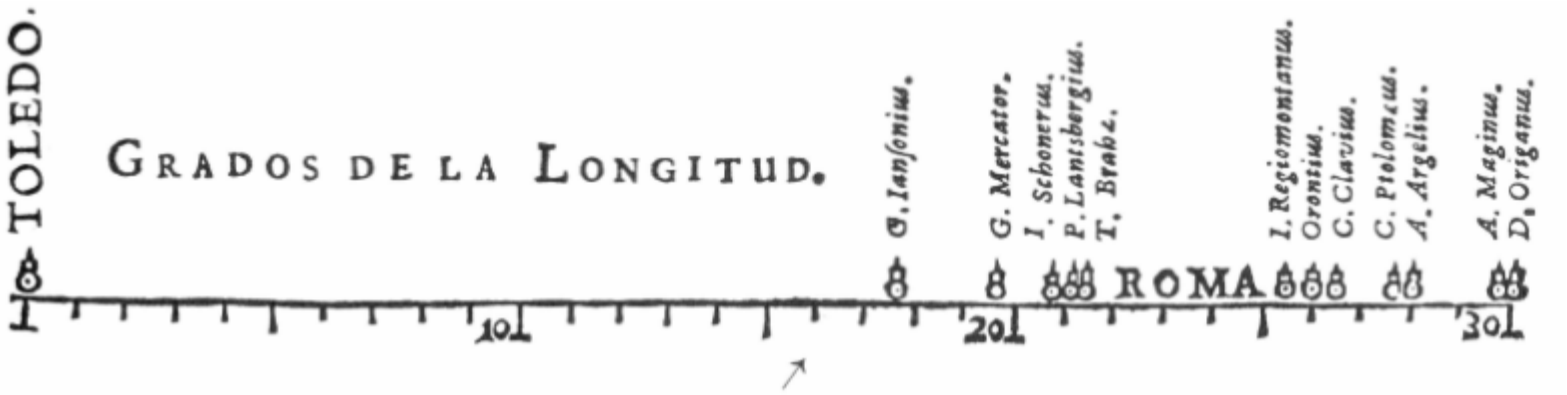
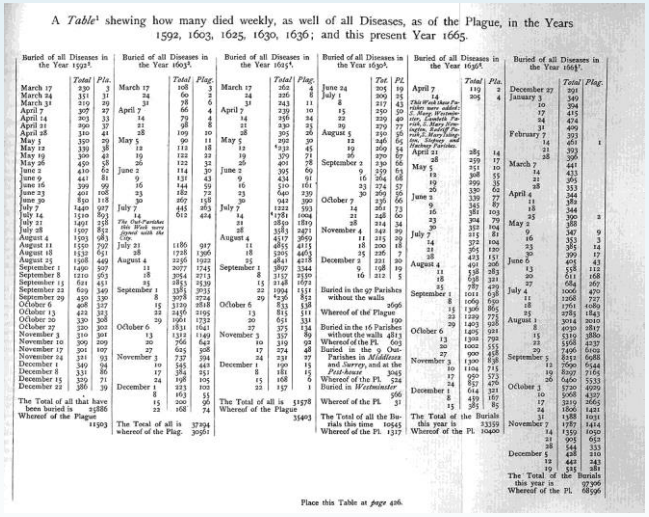


Fig. 1. Langren's 1644 graph of determinations of the distance, in longitude, from Toledo to Rome. The correct distance is 16°30′. Source: Tufte (1997, p. 15)

Life tables – Graunt, 1662
Graphical representation – Huygens, 1669



A *Table*[1] shewing how many died weekly, as well of all Diseases, as of the Plague, in the Years 1592, 1603, 1625, 1630, 1636; and this present Year 1665.

Buried of all Diseases in the Year 1592.

	Total	Plag.
March 17	230	3
March 24	351	31
March 31	219	29
April 7	307	27
April 14	303	33
April 21	290	37
April 28	340	41
May 5	350	29
May 12	339	38
May 19	300	43
May 26	450	58
June 2	410	62
June 9	441	81
June 16	399	99
June 23	401	108
June 30	830	118
July 7	1440	997
July 14	1530	893
July 21	1491	758
July 28	1507	852
August 4	1593	983
August 11	1550	797
August 18	1532	651
August 25	1508	449
September 1	1490	307
September 8	1210	163
September 15	611	431
September 22	603	349
September 29	450	330
October 6	408	337
October 13	422	385
October 20	330	308
October 27	320	302
November 3	310	301
November 10	309	209
November 17	301	107
November 24	321	93
December 1	349	94
December 8	331	86
December 15	359	71
December 22	386	39
The Total of all that have been buried is		25886
Whereof of the Plague		11503

Buried of all Diseases in the Year 1603.

	Total	Plag.
March 17	108	3
24	60	3
31	38	6
April 7	66	4
14	78	4
21	98	8
28	109	10
May 5	90	11
12	112	18
19	122	22
26	122	32
June 2	114	30
9	131	43
16	144	59
23	182	72
30	267	118
July 7	445	263
14	612	424
The Bills Parishes this Week were figured with the Copy.		
July 21	1186	917
28	1728	1366
August 4	2256	1902
11	2077	1745
18	2054	2213
25	2851	2530
September 1	3381	3035
8	3078	2724
15	2170	2818
22	2456	2195
29	1961	1732
October 6	1831	1641
13	1312	1143
20	766	642
27	625	508
November 3	737	594
10	585	442
17	384	251
24	198	105
December 1	223	102
8	163	55
15	200	90
22	168	74
The Total of all is		37294
whereof of the Plag.		30561

Buried of all Diseases in the Year 1625.

	Total	Plag.
March 17	262	4
24	226	8
31	243	11
April 7	239	10
14	256	34
21	230	23
28	305	26
May 5	292	39
12	*332	45
19	379	71
26	401	78
June 2	395	69
9	434	91
16	510	161
23	640	239
30	941	390
July 7	1222	593
14	*1781	1004
21	2850	1809
28	3583	2471
August 4	4517	3169
11	4855	4115
18	5201	4465
25	4841	4218
September 1	3807	3344
8	2157	2550
15	2148	1672
22	1904	1551
29	*836	823
October 6	853	538
13	811	511
20	651	351
27	375	234
November 3	352	89
10	319	92
17	274	48
24	231	27
December 1	190	15
8	181	15
15	168	6
22	137	1
The Total of all is		51578
Whereof of the Plague		35493

Buried of all Diseases in the Year 1630.

	Tot	Pl.
June 24	305	89
July 1	309	85
8	217	43
15	250	50
22	229	40
29	279	77
August 5	230	56
12	246	65
19	269	84
26	270	67
September 2	230	66
9	259	63
16	264	68
23	274	57
30	269	56
October 7	236	66
14	262	73
21	248	60
28	214	34
November 4	242	29
11	215	29
18	200	18
25	226	7
December 2	221	20
9	198	19
16	212	5
Buried in the 97 Parishes without the walls		2596
Whereof of the Plague		190
Buried in the 16 Parishes without the walls		4813
Whereof of the Pl.		603
Buried in the 9 Out-Parishes in *Middlesex* and *Surrey*, and at the *Pest-house*		3045
Whereof of the Pl.		524
Buried in *Westminster*		566
Whereof of the Pl.		31
The Total of all the Burials this time		10545
Whereof of the Pl.		1317

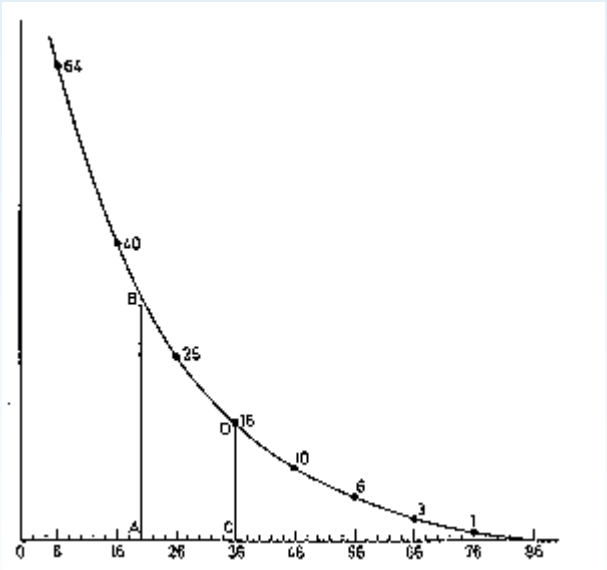
Buried of all Diseases in the Year 1636.

	Total	Plag.
April 7	119	2
14	205	4
This Week these Parishes were added: St. Mary Woolnoth, St. Mary Woolchurch, Lambeth Parish, Christ-Church Parish, Ratcliff Parish, St. Mary Islington, Stepney and Hackney Parishes.		
April 21	285	14
28	259	17
May 5	751	10
12	308	55
19	299	35
26	330	62
June 2	339	77
9	345	87
16	381	103
23	304	79
30	352	104
July 7	315	81
14	372	104
21	365	120
28	402	137
August 4	491	206
11	538	283
18	636	321
25	787	419
September 1	1011	638
8	1069	630
15	1306	865
22	1229	778
29	1403	916
October 6	1405	901
13	1302	798
20	1002	555
27	900	458
November 3	1300	856
10	1104	715
17	950	573
24	857	476
December 1	614	321
8	489	167
15	385	85
The Total of the Burials this year is		13359
Whereof of the Pl.		10400

Buried of all Diseases in the Year 1665.

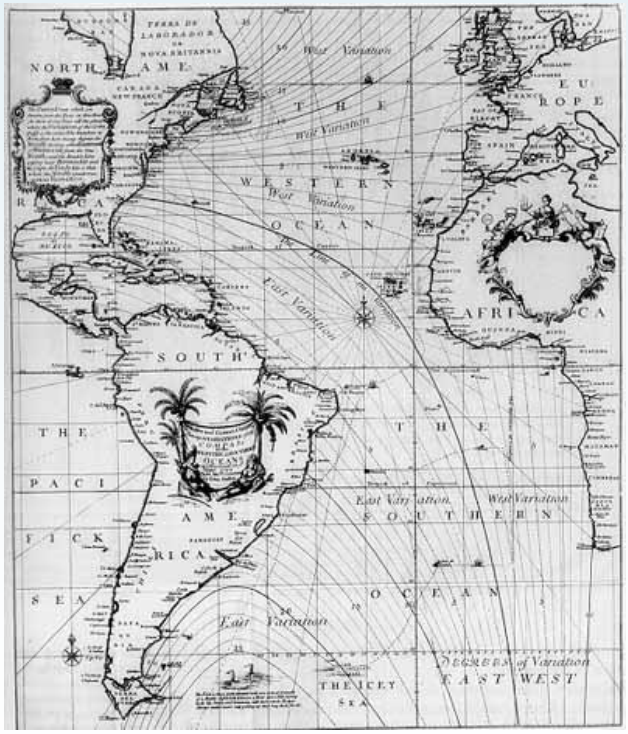
	Total	Plag.
December 27	291	
January 3	349	
10	394	
17	415	
24	474	
31	499	
February 7	393	
14	461	1
21	393	
28	396	
March 7	441	
14	433	
21	365	
28	353	
April 4	344	
11	382	
18	344	
25	390	2
May 2	388	
9	347	9
16	353	3
23	385	14
30	399	17
June 6	405	43
13	538	112
20	611	168
27	684	267
July 4	1006	470
11	1268	727
18	1761	1089
25	2785	1843
August 1	3014	2010
8	4030	2817
15	5319	3880
22	5568	4237
29	7496	6102
September 5	8311	6988
12	7690	6544
19	8297	7165
26	6460	5535
October 3	5720	4929
10	5068	4327
17	3219	2665
24	1806	1421
31	1388	1031
November 7	1187	1414
14	1359	1050
21	905	652
28	544	333
December 5	428	210
12	442	243
19	385	281
The Total of the Burials this year is		97306
Whereof of the Pl.		68596

Place this Table at *page* 416.



Source: https://www.datavis.ca/milestones/index.php?group=1600s

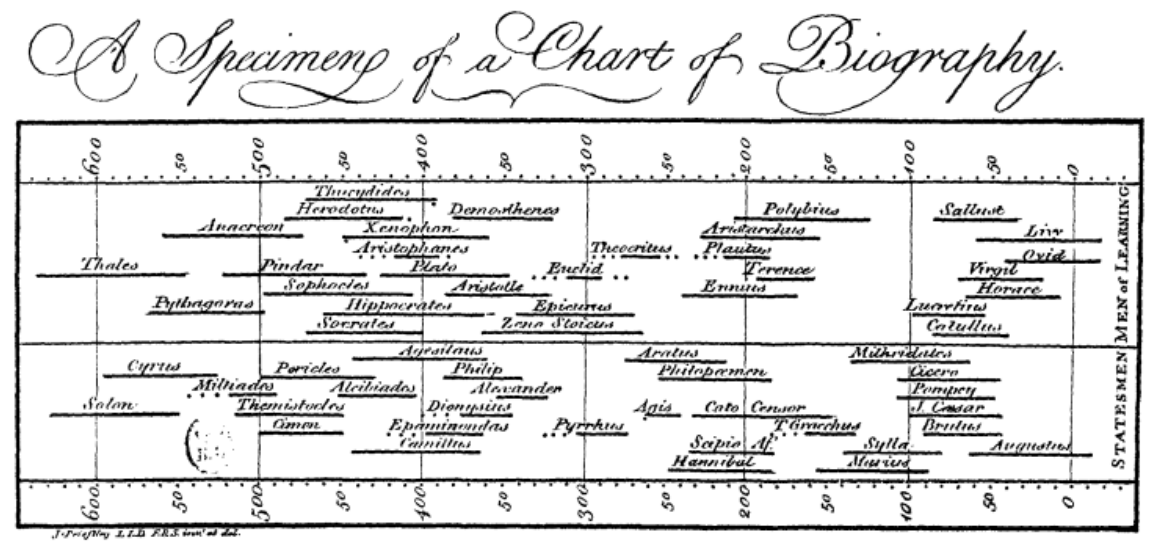
Halley, 1701



- "Contour maps showing curves of equal value (an isogonic map, lines of equal magnetic declination for the world, possibly the first contour map of a data-based variable)"

Priestly, 1765
Historical timeline

- "Historical timeline (life spans of 2,000 famous people, 1200 B.C. to 1750 A.D.), quantitative comparison by means of bars"



William Playfair

"William Playfair (1759–1823) is widely considered the inventor of most of the graphical forms widely used today— first the line graph and bar chart (Playfair,1786), later the pie chart and circle graph (Playfair, 1801). A somewhat later graph(Playfair, 1821), shown in Figure 2, exemplifies the best that Playfair had to offer with these graphic forms. Playfair used three parallel timeseries to show the price of wheat, weekly wages, and reigning monarch over a~250 year span from 1565to 1820, and used this graph to argue that workers had become better off in the most recent years"

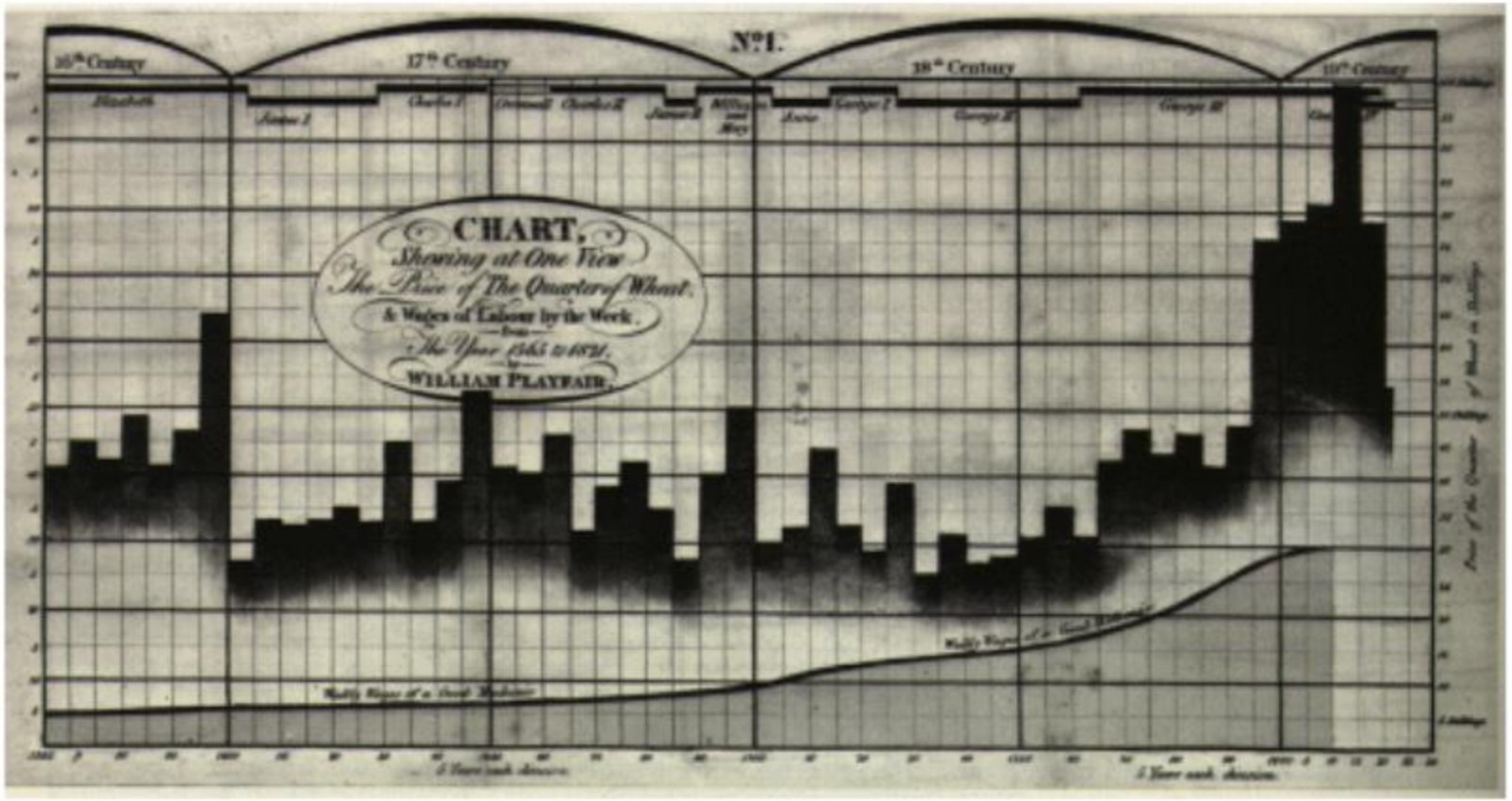
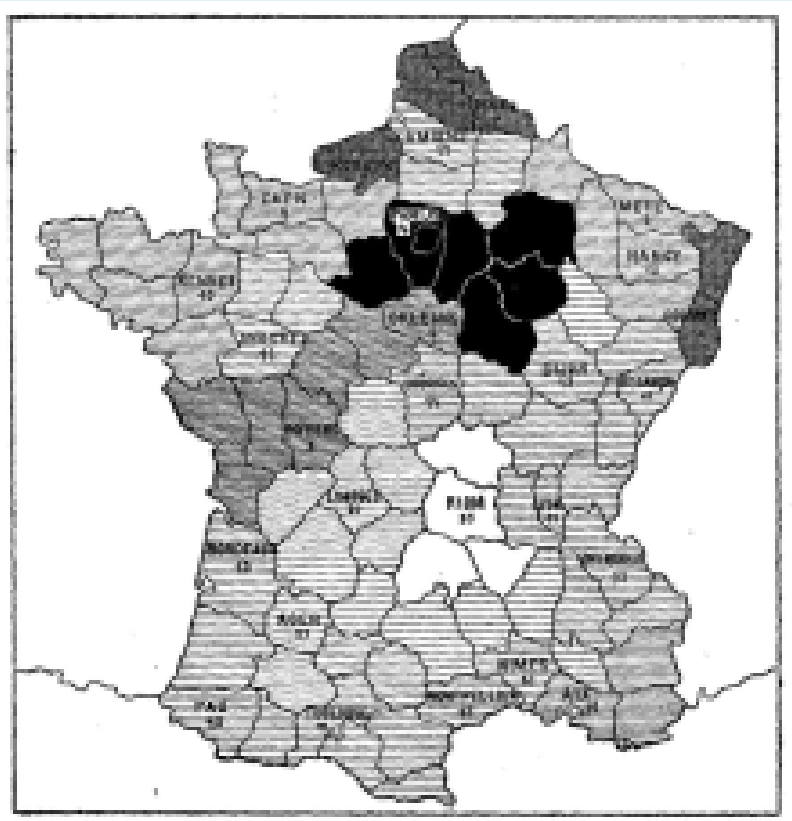


Fig. 2. William Playfair's 1821 time series graph of prices, wages, and ruling monarch over a 250 year period. *Source*: Playfair (1821), image from Tufte (1983, p. 34)

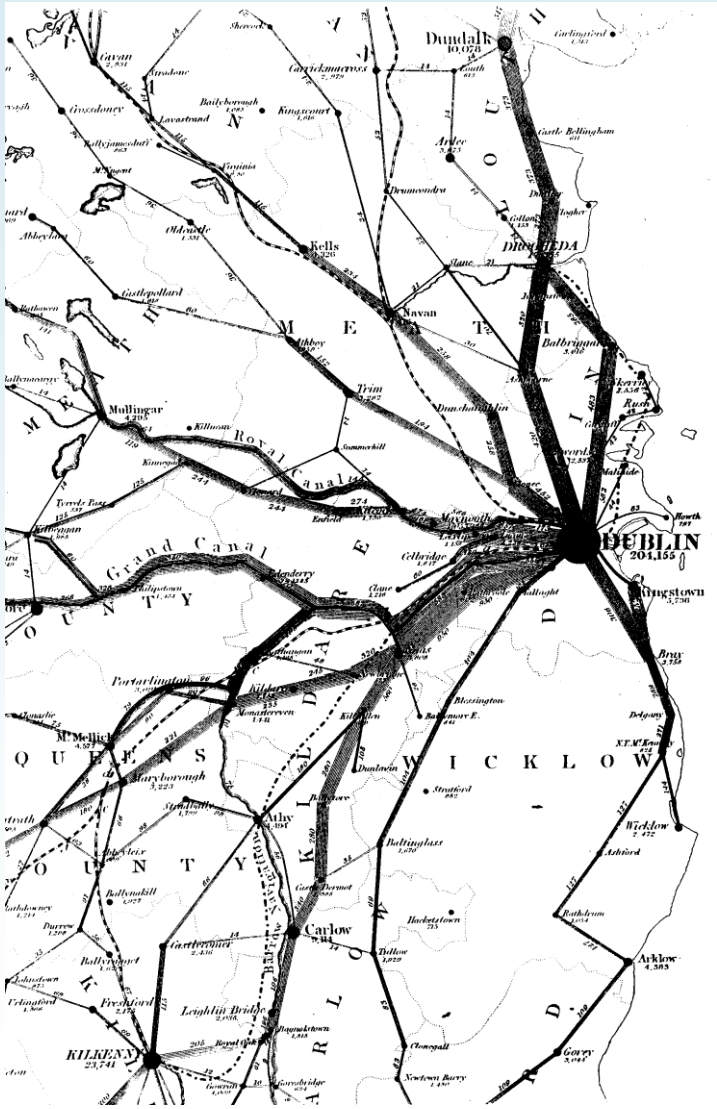
Dupin, 1819

- First choropleth map?



Harness, 1837

- First flow map



Minard, 1844

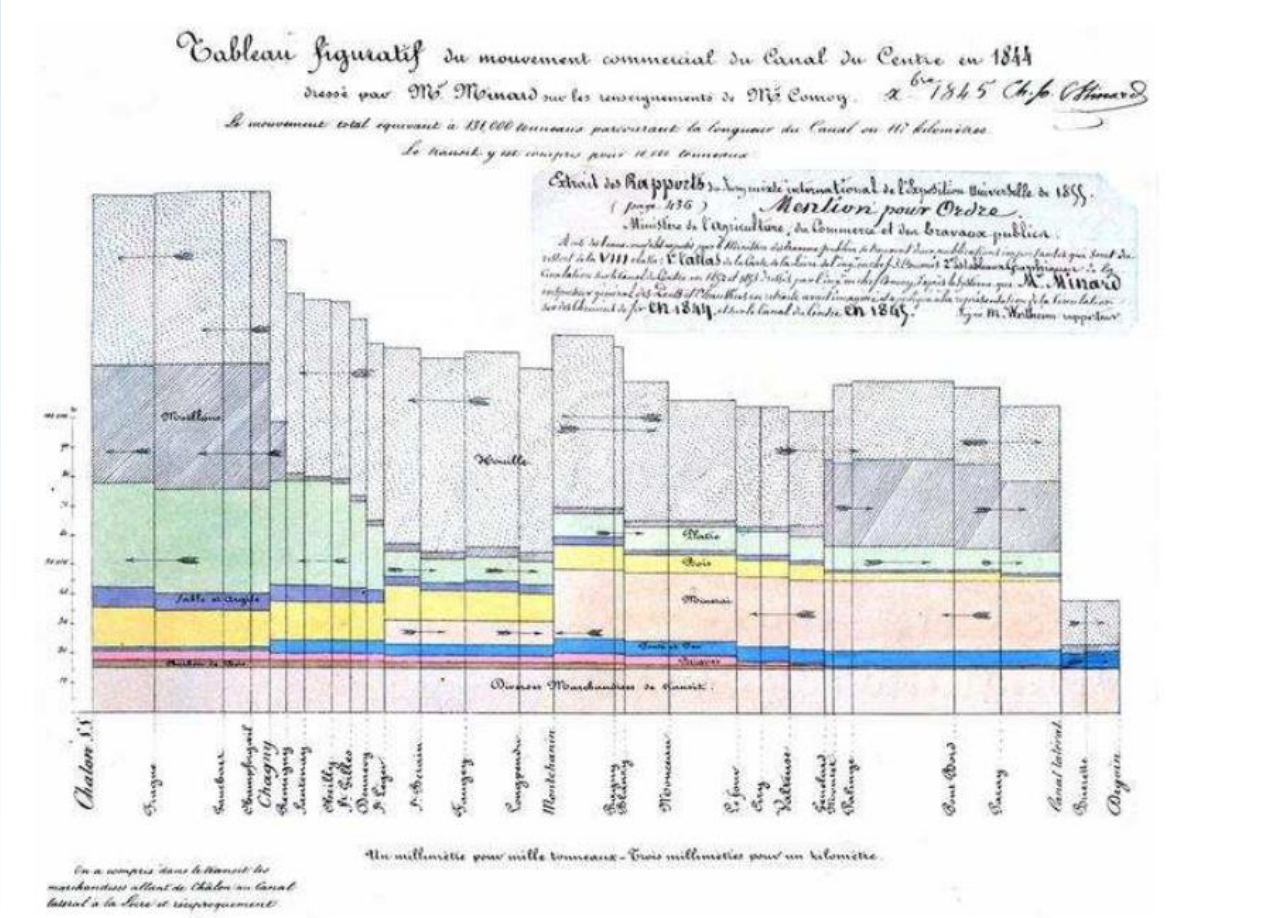


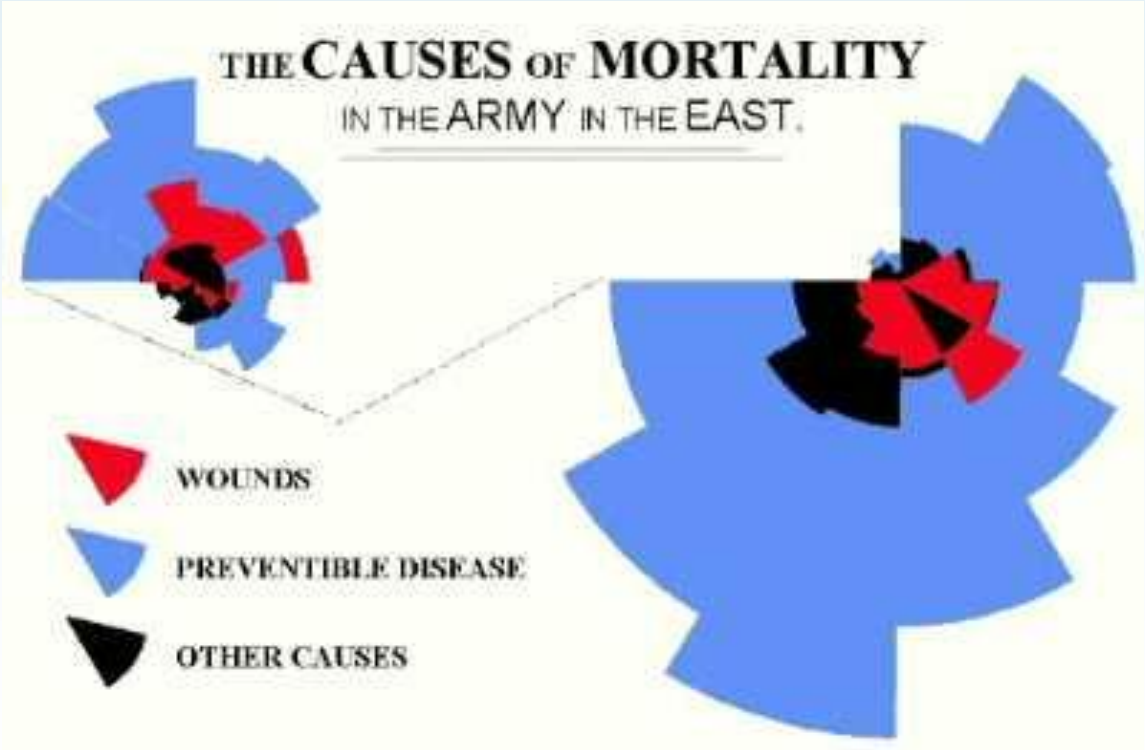
Fig. 3. Minard's Tableau Graphique, showing the transportation of commercial goods along the Canal du Centre (Chalon–Dijon). Intermediate stops are spaced by distance, and each bar is divided by type of goods, so the area of each tile represents the cost of transport. Arrows show the direction of transport. *Source:* ENPC:5860/C351 (Col. et cliché ENPC; used by permission)

Friendly

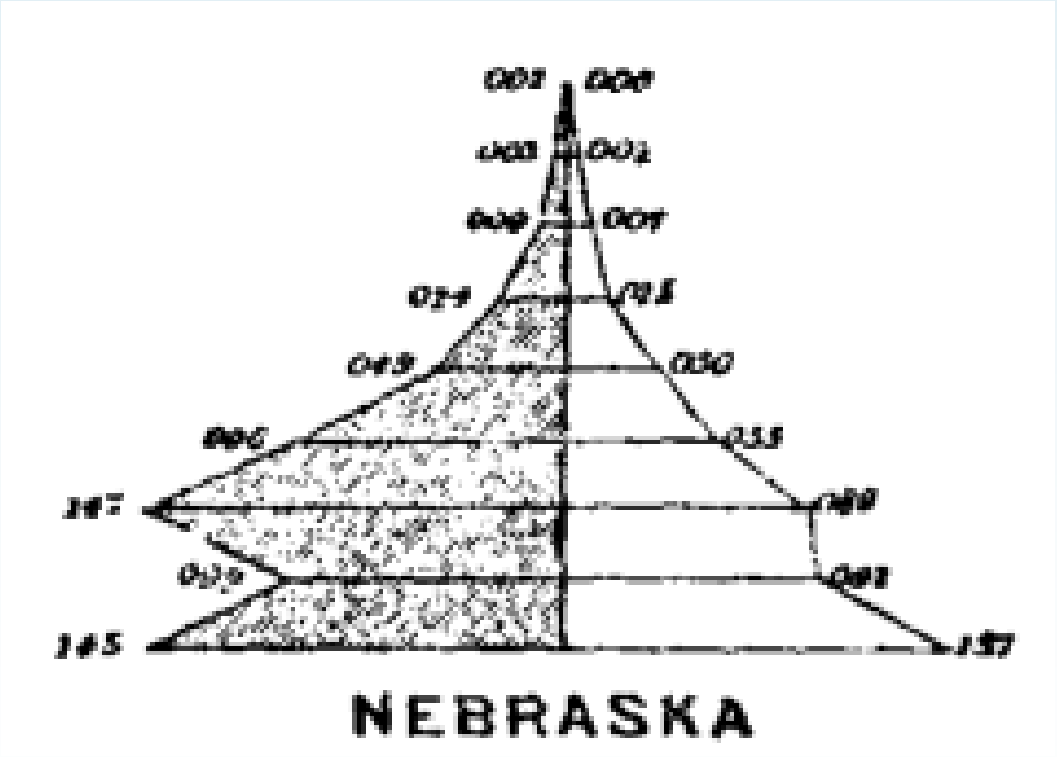
Period	Description
1600-1699	Measurement and theory
1700-1799	New graphic forms
1800-1850	Beginnings of modern graphics
1850-1899	The Golden Age of statistical graphics
1900-1950	The modern dark ages
1950-1975	Re-birth of data visualization

Snow, 1855 Cholera map

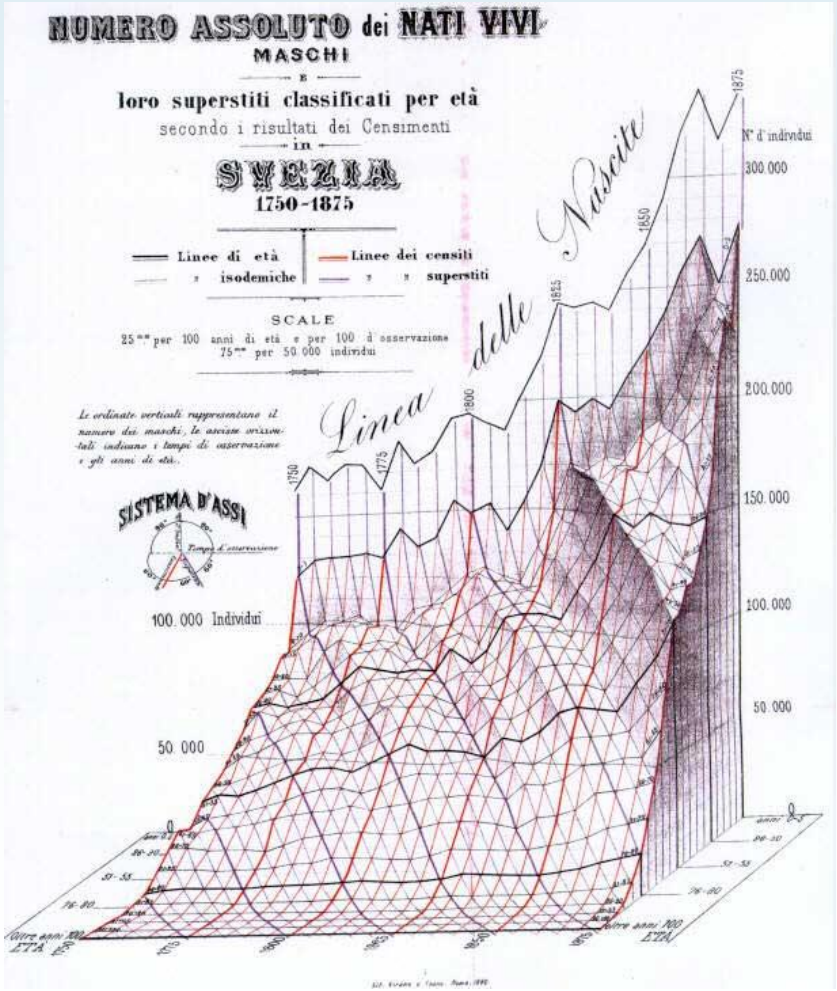
Nightingale, 1857
Causes of mortality



Walker, 1874
Population pyramid

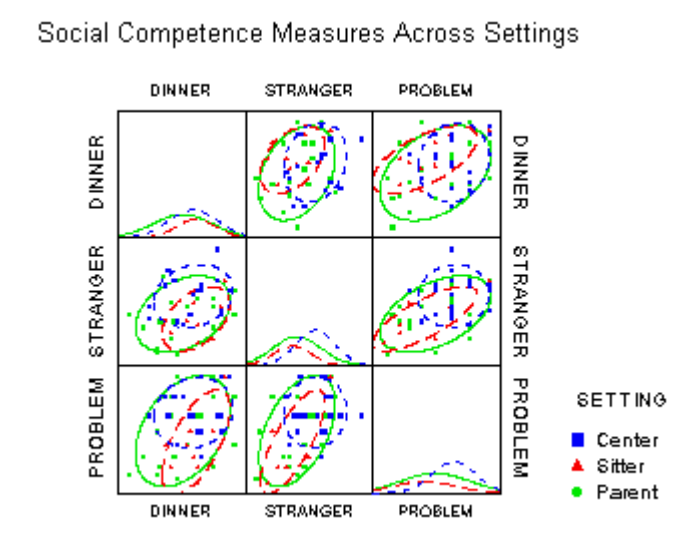


Perozzo, 1879



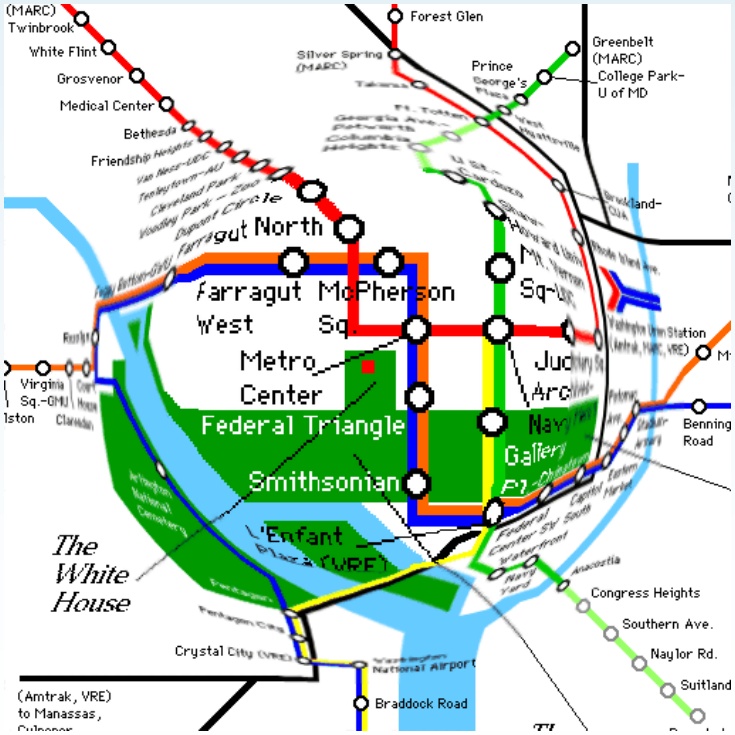
"Stereogram (three-dimensional population pyramid) modeled on actual data (Swedish census, 1750--1875)"

Hartigan, 1975



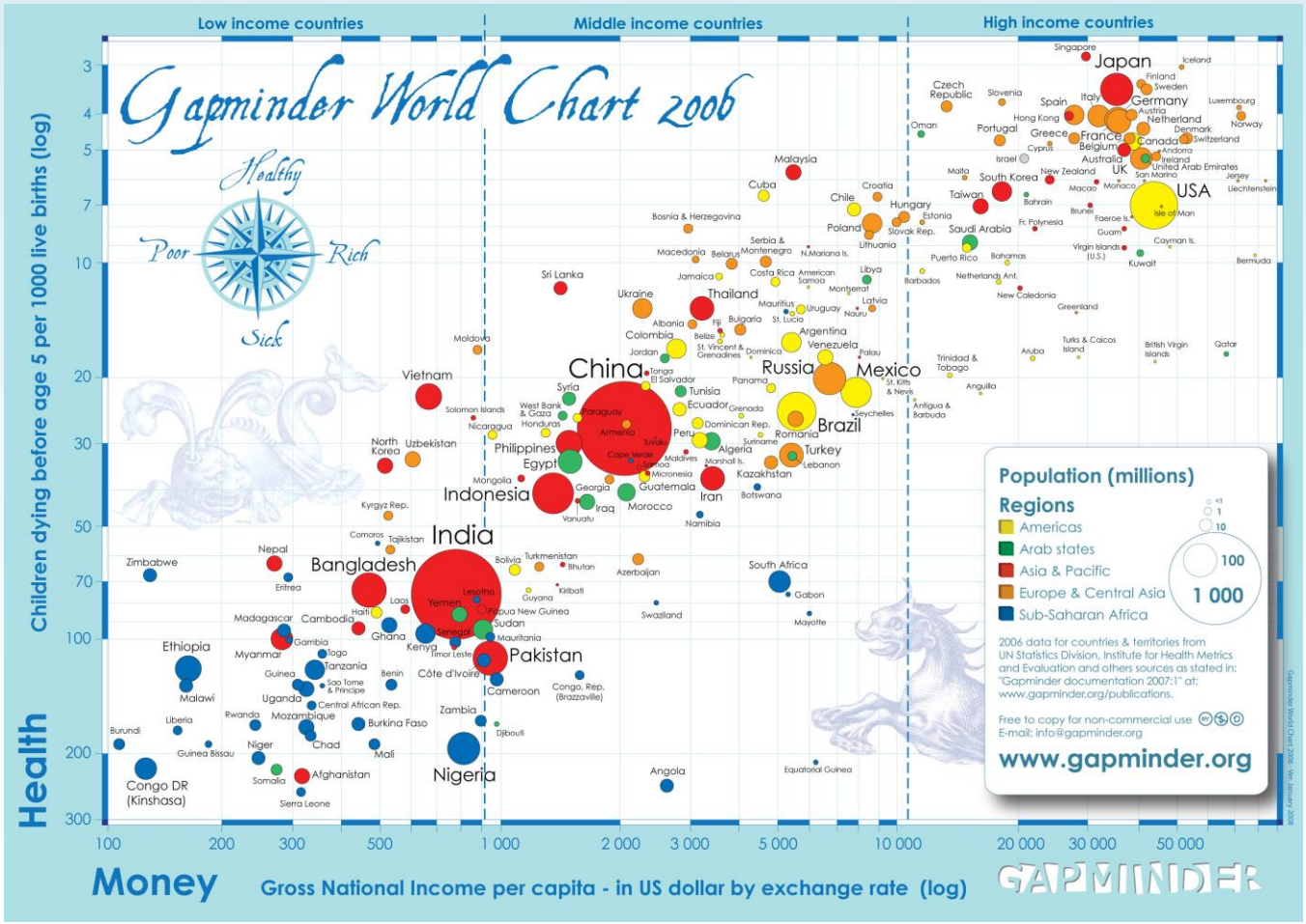
"Scatterplot matrix, the idea of plotting all pairwise scatterplots for *n* variables in a tabular display"

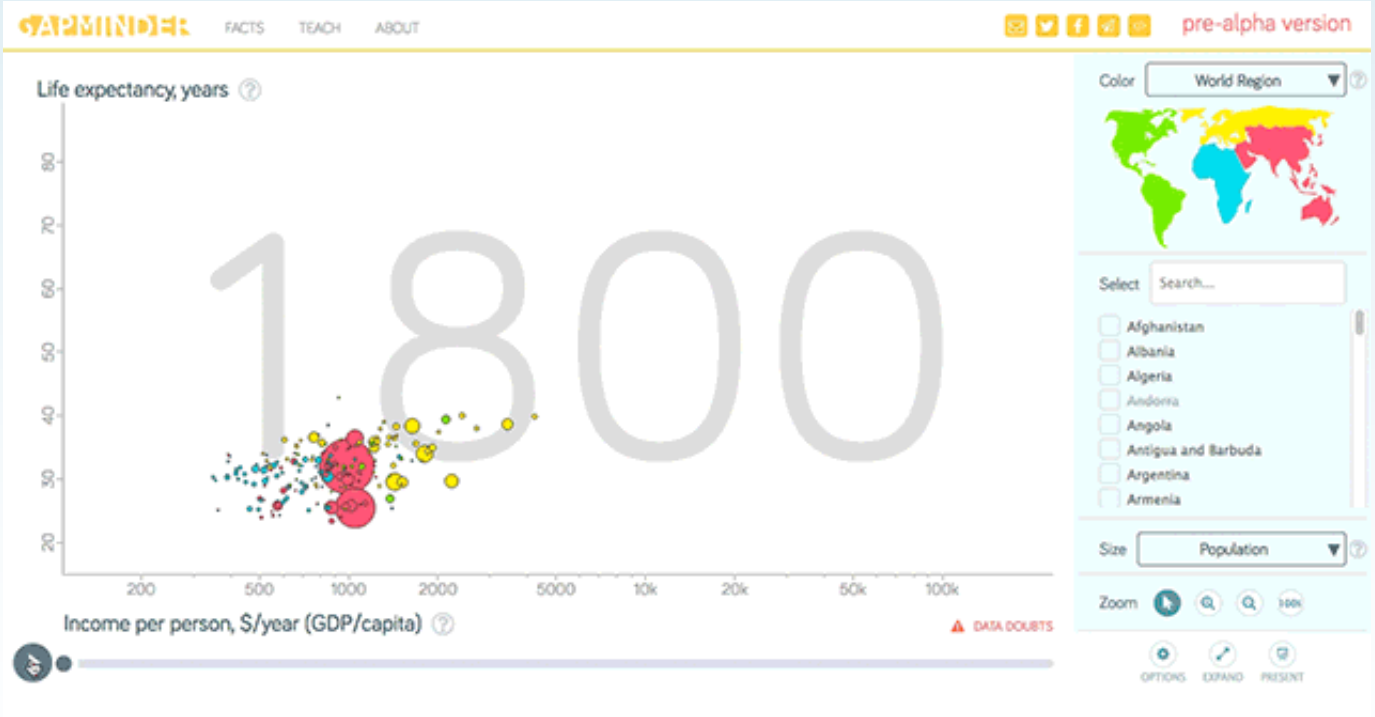
Furnas, 1981



"Fisheye view: an idea to provide focus and greater detail in areas of interest of a large amount of information, while retaining the surrounding context in much less detail"

Rosling, 2006





Source: https://www.storybench.org/role-motion-visualizations/gapminder/