

ROMEO AND JULIET
Number of characters **41 | 37%** Network density

Varieties of information visualization

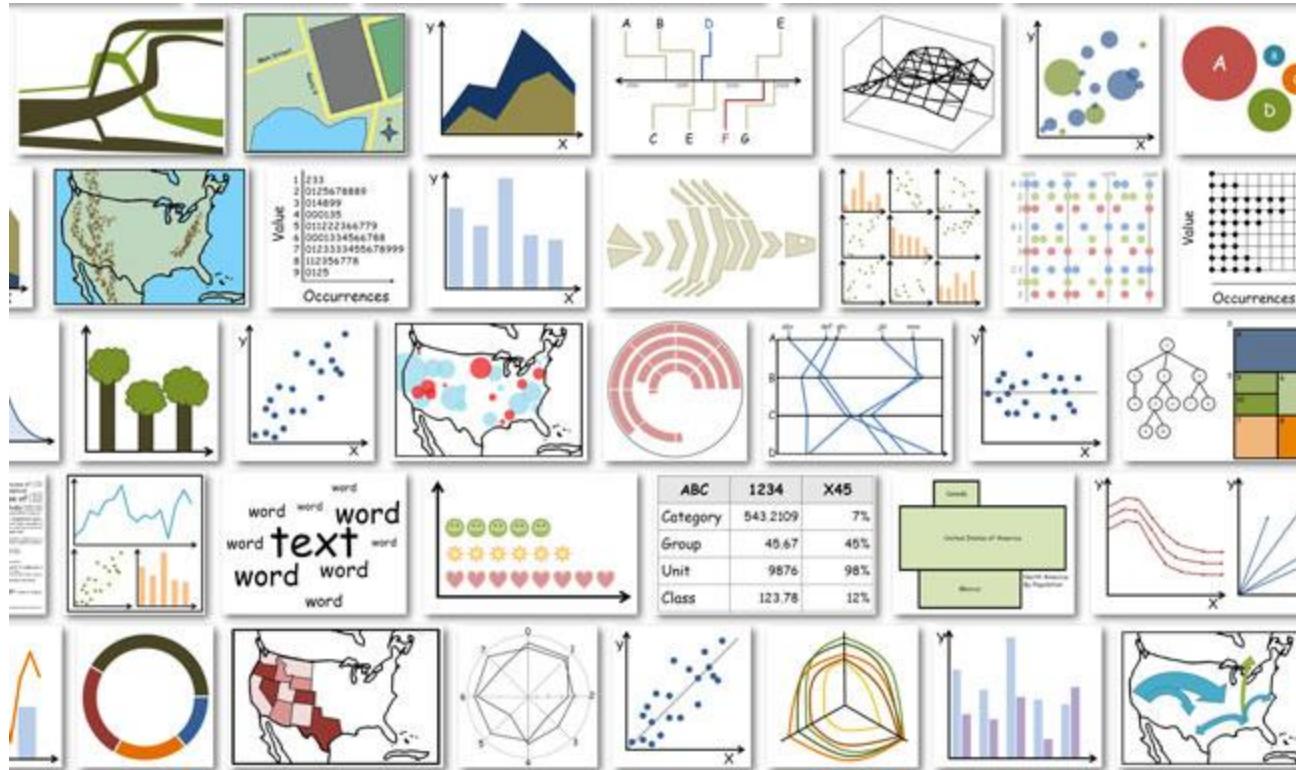
Michael Friendly
Psych 6135

So many types

There are so many kinds of charts, diagrams, graphs, maps

What are their features?

What tasks are they good for? – Accuracy or speed of judgment? Memorability?



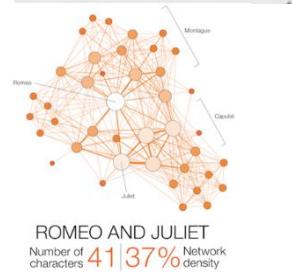
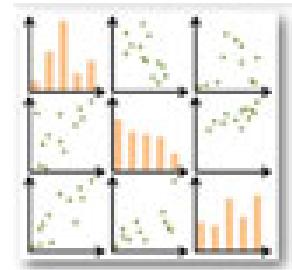
Classify by: ???

For purposes of “What kind of graph should I use?” usually most useful to think:
“What do I want to show?”

Deviation	Correlation	Ranking	Distribution	Change over Time	Magnitude
Example FT uses: Trade surplus/deficit, climate change	Example FT uses: Inflation & unemployment, income & life expectancy	Example FT uses: Worst, departing, league tables, constituency election results	Example FT uses: Income distribution, population (geographic) distribution	Example FT uses: Share, oil movements, economic time series	Example FT uses: Commodity production, market capitalisation
Diverging bar A simple standard bar chart that can handle both negative and positive magnitude values.	Scatterplot The standard way to show the relationship between two continuous variables, each of which has its own axis.	Ordered bar Standard bar charts display the ranks of values much more easily when sorted into order.	Histogram The standard way to show a unimodal distribution - keep the gaps between columns small to highlight the shape of the data.	Line The standard way to show a changing time series. If data are irregular, consider markers to represent data points.	Column The standard way to compare parts of things. Must always start at 0 on the axis.
Diverging stacked bar Perfect for presenting survey results which involve subtracting (eg. disagree/neutral/agree).	Line + Column A good way of showing the relationship between an amount (column) and a rate (line).	Ordered columns See above.	Boxplot Summarise multiple distributions by showing the median, quartiles and range of the data.	Column Columns work well for showing change over time - but usually best with only one series of data at a time.	Bar See above. Good when data are not fine series and labels have long category names.
Spike chart Splits a single issue into two contrasting components (eg. market/bonds).	Connected scatterplot Usually used to show how the relationship between 2 variables has changed over time.	Ordered proportional symbol Used when there are big variations between values and seeing the relative size of the data between data is not so important.	Violin plot Similar to a box plot but more effective with complex distributions. Gives a clear sense of the data with simple messages.	Line + column A good way of showing the relationship over time between something (columns) and a rate (line).	Paired column As per standard column but allows for multiple series. Can make it easier to read with these than 2 series.
Balanced/deficit filled line The standard way of filling a chart to show a balance to be shown either against a baseline or between two series.	Bubble Like a scatterplot but with additional detail by sizing the circles according to a third variable.	Dot strip plot Data placed in order on a grid as a space-efficient method of laying out ratios across multiple categories.	Population pyramid A standard way for showing the age and sex breakdown of a population often illustrating effectively back to back histograms.	Stacked price Usually focused on day-to-day activity. These charts show opening/closing and high/low points of each day.	Paired bar See above.
XY heatmap A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts.	Slope Perfect for showing how ratios have changed over time or vary between categories.	Dot strip plot Good for showing individual values in a distribution, can be a problem when too many dots have the same value.	Step chart Good for showing changing data as long as the data can be simplified into 1 or 3 points without missing a lot of data.	Proportional stacked bar A good way of showing the size and proportion of data at the same time - as long as the data are not too complex.	

Topics, by graph type

- Statistical data graphs
 - 1D: dotplot, boxplot, violin plot
 - 1.5D: time-series plot, density plot, bar chart, pie chart
 - 2D: scatterplot, ridgeline plot
 - 3D: contour plot, 3D scatterplot, surface plot
- Thematic maps
 - Choropleth map
 - Anamorphic map
 - Flow maps
- Network & tree visualization
- Animation & interactive graphics



What are dimensions

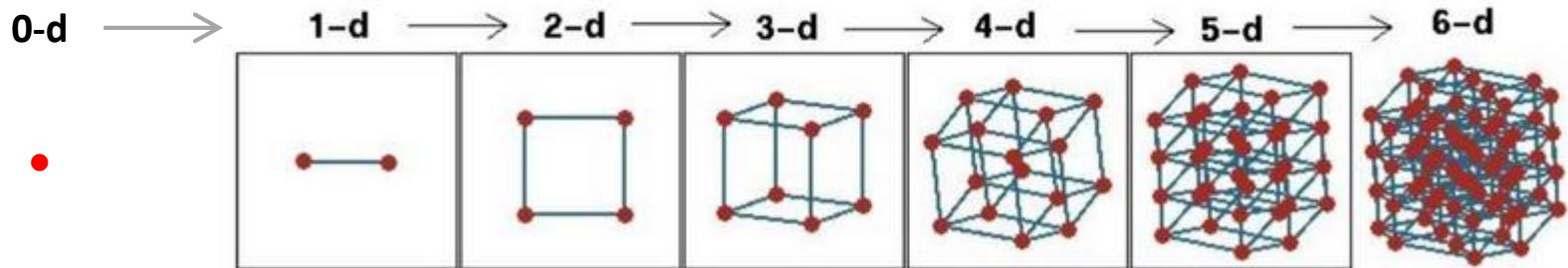


Fig credit: Di Cook [@visnut](#)

1 D
1.5 D
2 D
3 D
 n D ?

Data graphs can be classified by the number of variables, dimensions shown in a given graph

Data graphs

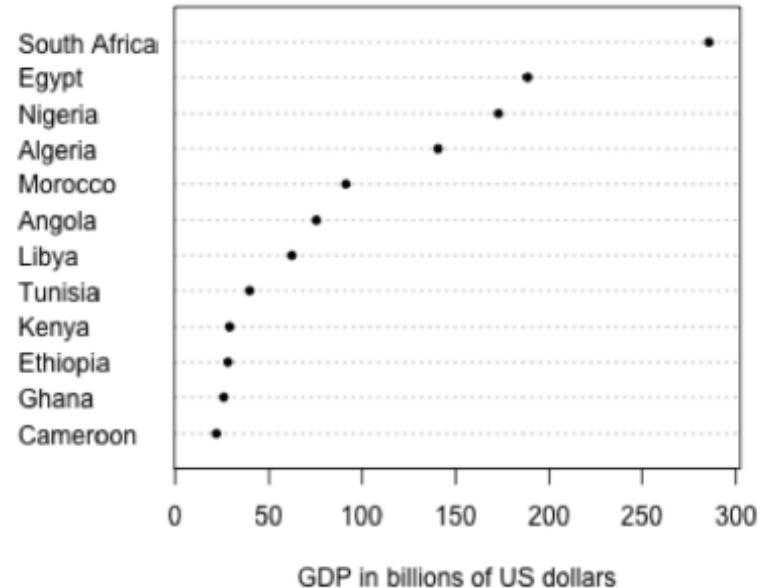
1D: Infographic vs. Data graphic

The same data can be shown in different forms, for different purposes

African Countries by GDP



African Countries by GDP



One might argue that this infographic has greater impact in showing the relative size of GDP

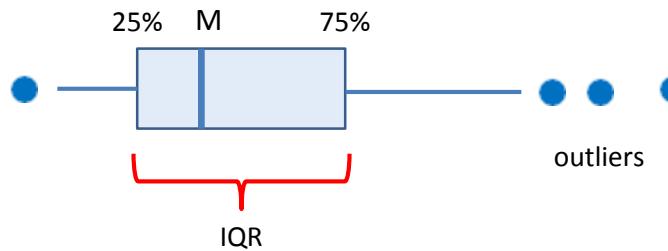
One might argue that this statistical graph makes comparisons easier

1.5D: Dotplots & boxplots

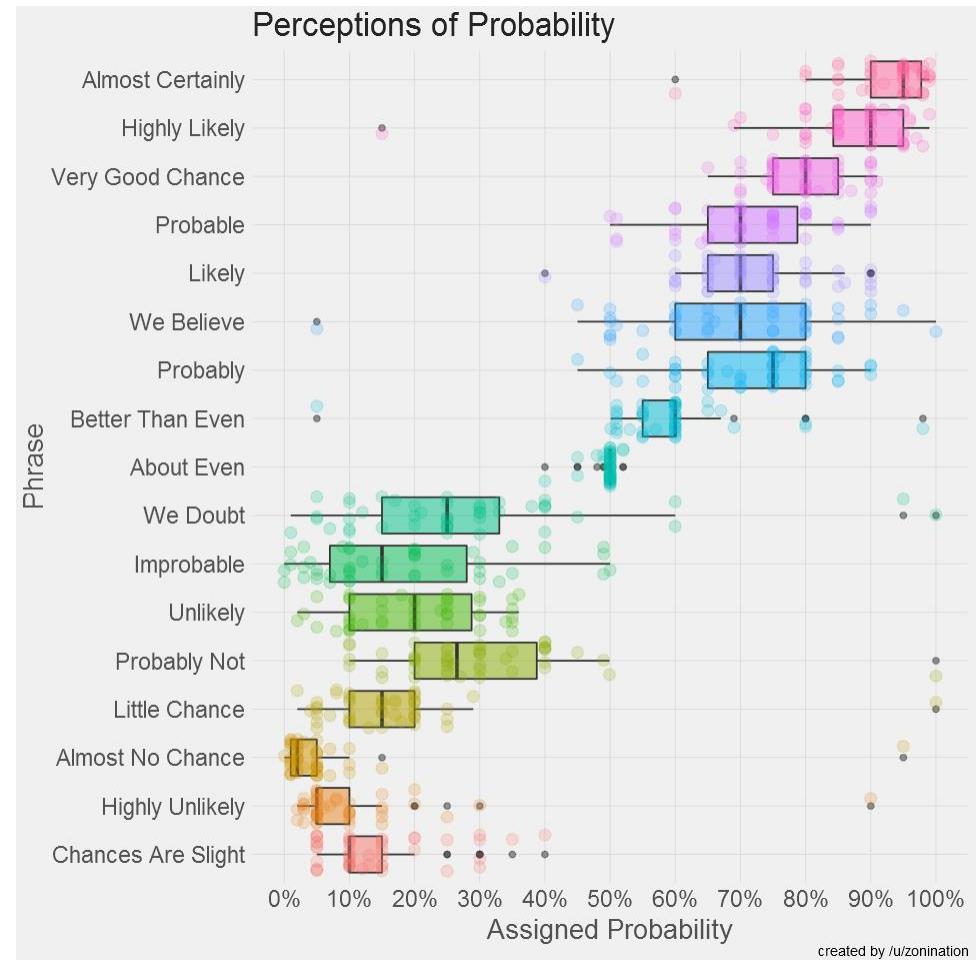
What number do you give to a probability phrase?

Boxplots summarize the important characteristics of a univariate data distribution:

- center (median)
- spread (IQR)
- shape (symmetric? skewed?)
- outliers?



This example overlays the boxplot with a jittered dotplot, so we can also see the individual observations



This visualization made the longlist for the 2015 Kantar Information is beautiful award. Data & R code:
<https://github.com/zonation/perceptions>

1.5D: Density ridgeline plots

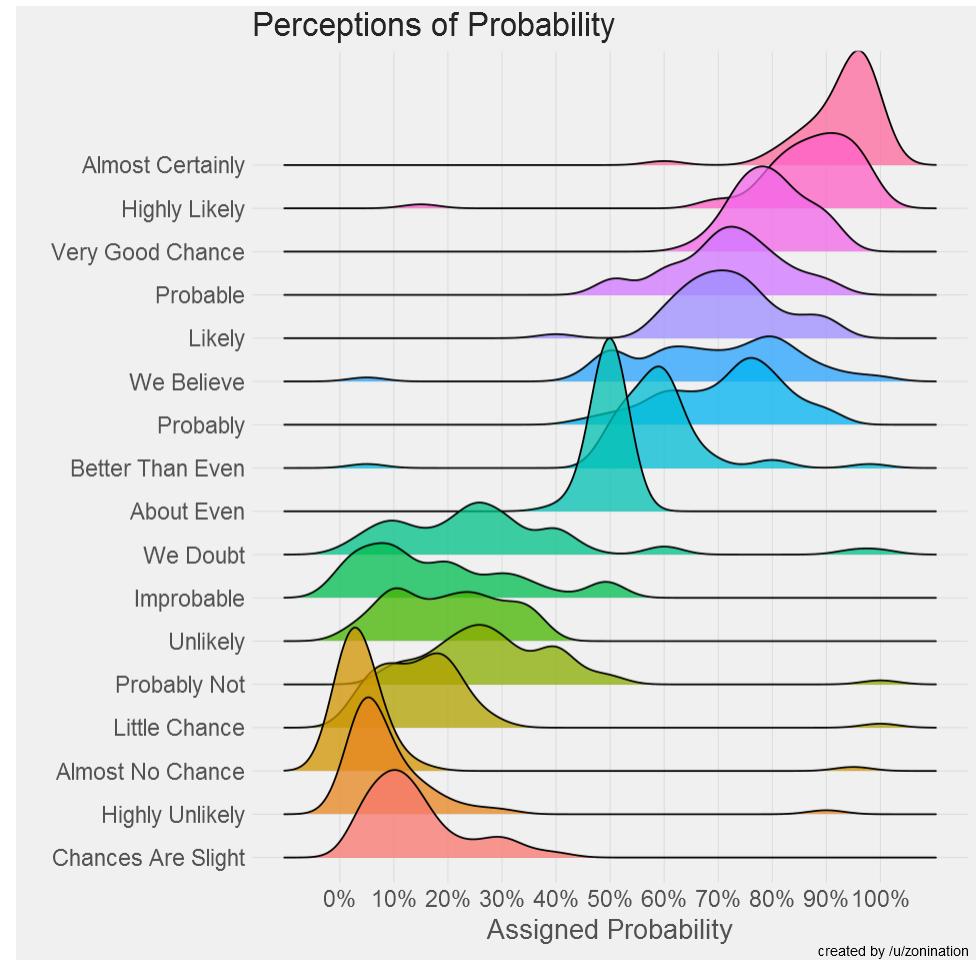
Another possible 1D display is a **density estimate**— a statistically smoothed histogram.

For comparing a set of them, a **ridgeline** plot stacks them vertically to create the impression of a mountain range.

As in the boxplot version, this uses:

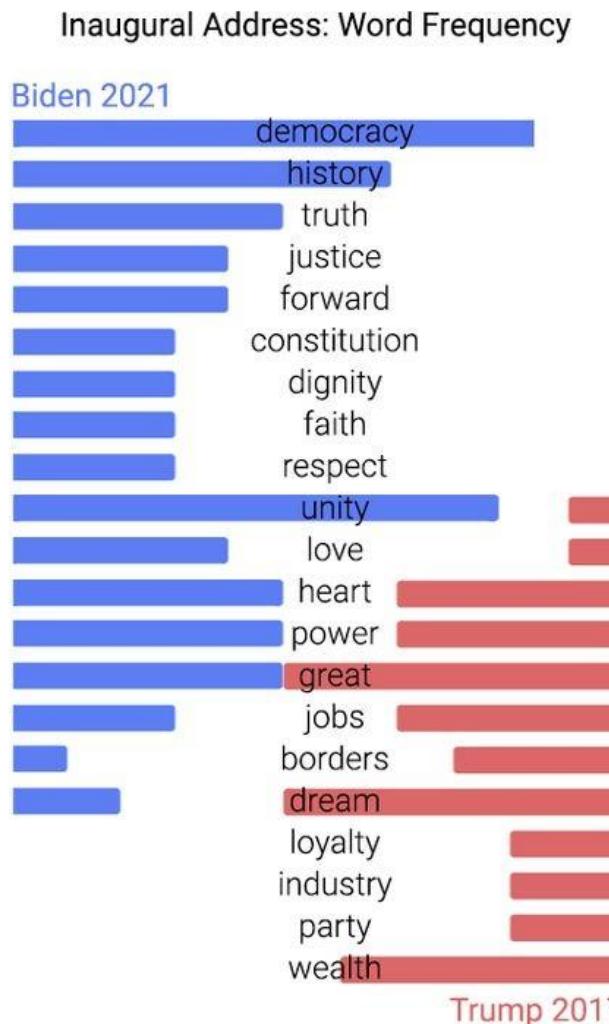
- a progressive scale of colors
- transparent colors to handle overlap

Q: What features stand out here?



Software note: These figures are drawn with R, using ggplot2 and the ggridges package. See:
<https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html>

1.5D: Text bar charts



- Text can be analyzed as data also, most often in frequency counts.
- This chart uses a novel design to compare the most frequent words by Biden (2021) & Trump 2017) in their inaugural addresses.
- The contrast is striking!
 - **democracy, unity** vs. **great, dream**

From:

https://www.reddit.com/r/dataisbeautiful/comments/l7k0f0/us_inauguration_address_word_frequency_biden_vs/

1.5D: Time series line graphs

William Playfair (1786), *The Commercial and Political Atlas*, invented the time series line graph as a way to show data on England's trade with other countries

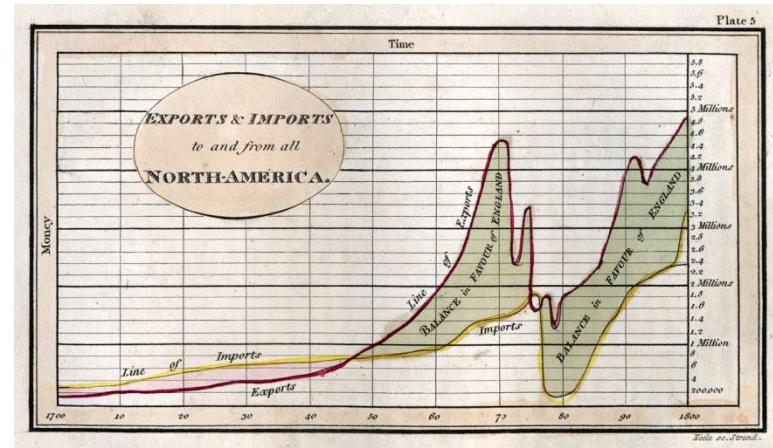
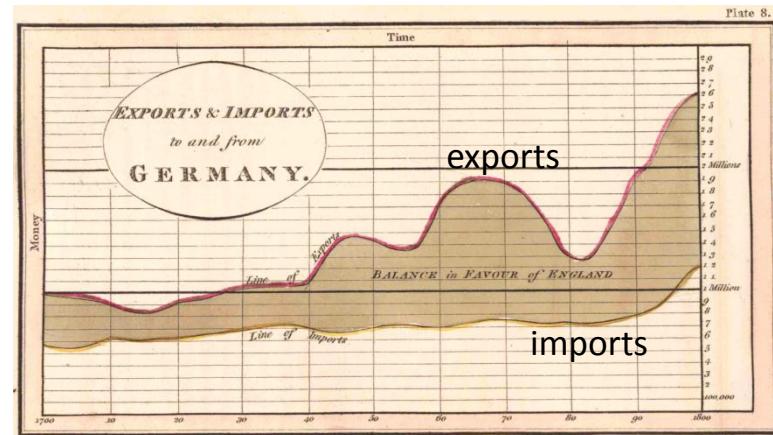
One curve for imports, one for exports

The **balance of trade** could be seen as the difference between the curves

Trade with Germany was consistently in favor of England

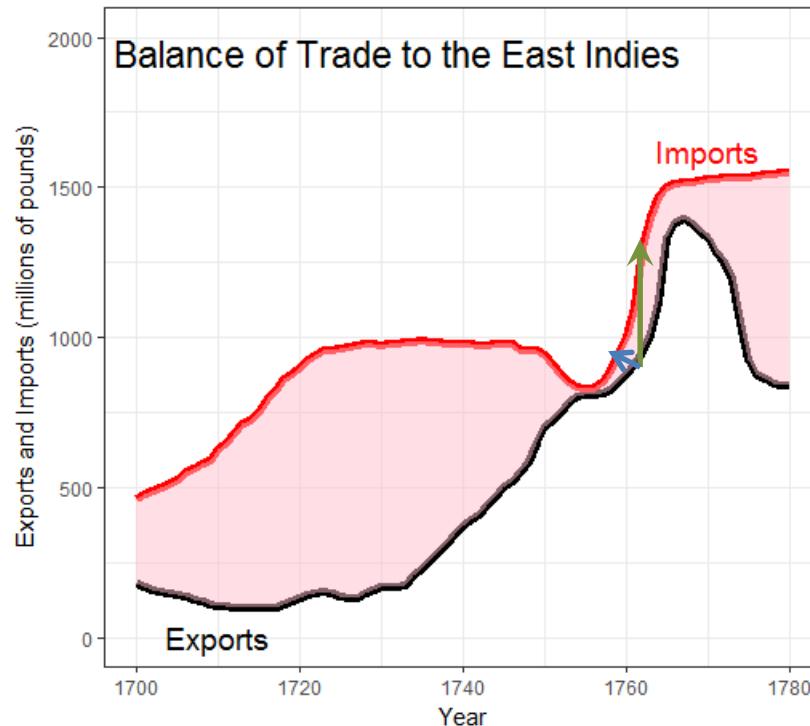
With North America, the balance changed back and forth over time

Economic 'history' could now be visualized and explained

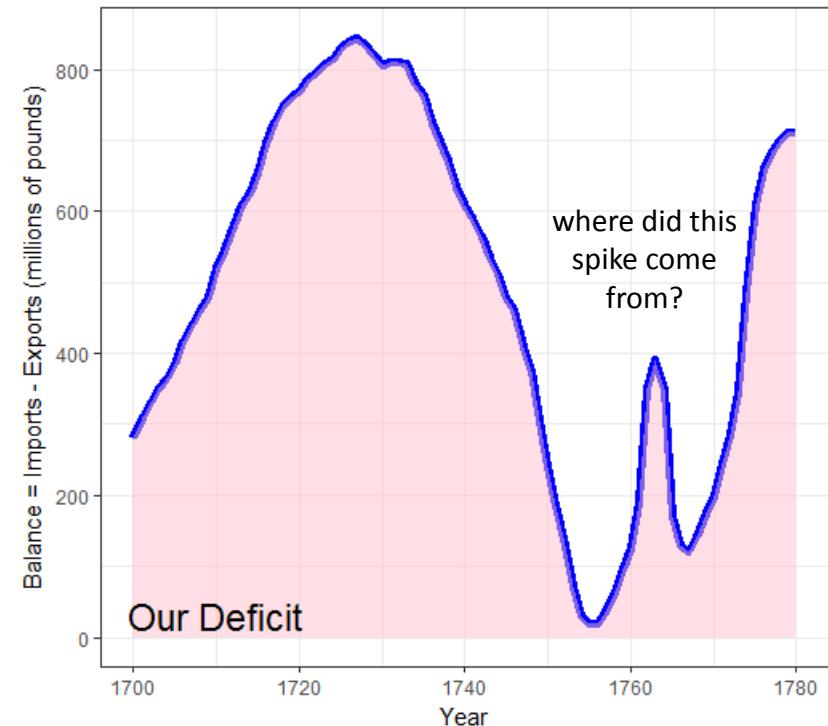


Psychology: Distances between curves

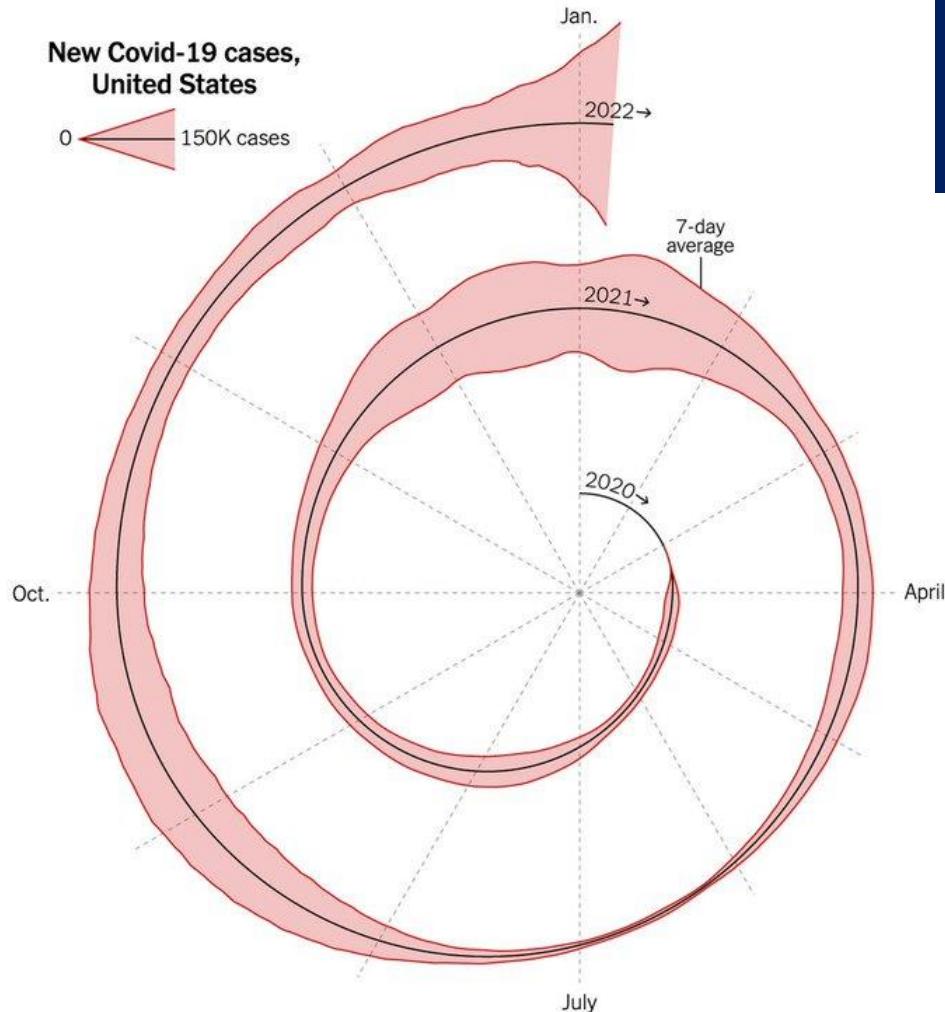
What Playfair didn't know is that judgments of **distance** between curves are **biased**
We tend to see the **perpendicular** distance rather than the **vertical** distance



Plotting balance of trade directly



Jan. 6, 2022



1.5 D COVID spiral graph

This graph shows a 1.5 D time series of COVID-19 cases over the entire span of the pandemic

It is wrapped into a spiral to compare months over years.

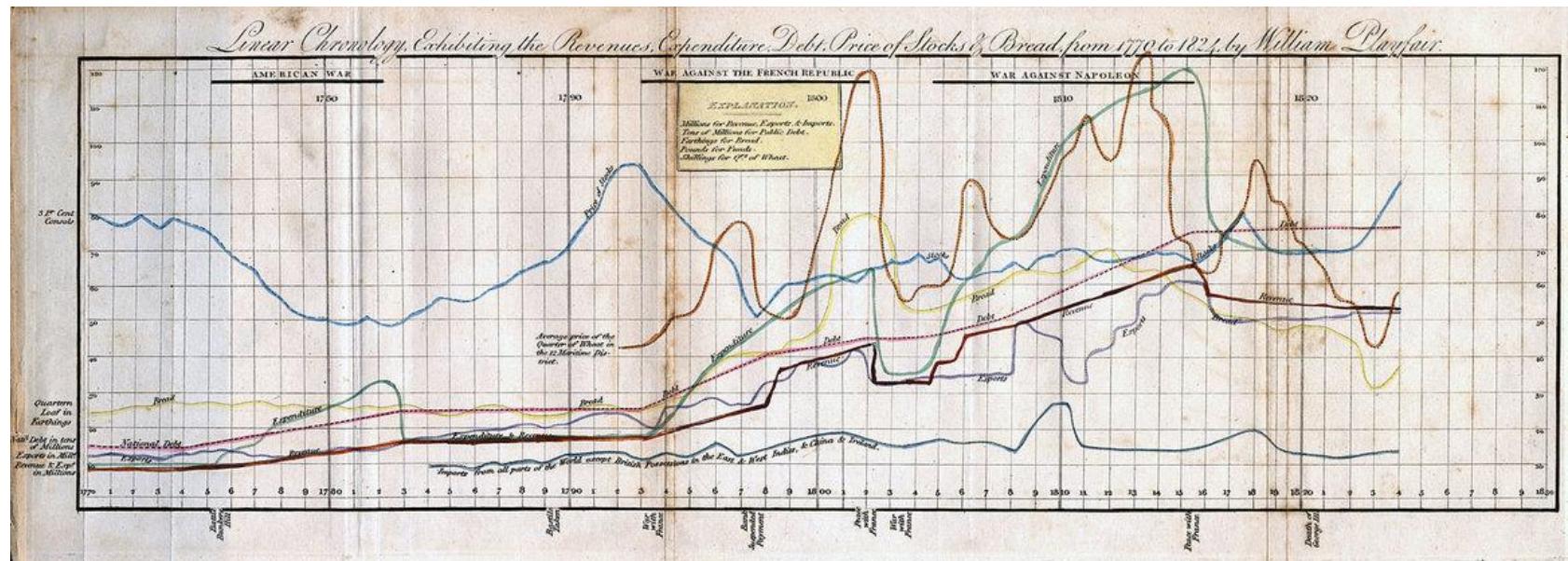
Does this work for you, or is it too weird?
Is it the “tapeworm of doom”?

See commentary: <https://www.youtube.com/watch?v=YWowF9Vi4Mw>

Multiple time series graphs

Things get messy when there are many series to be compared

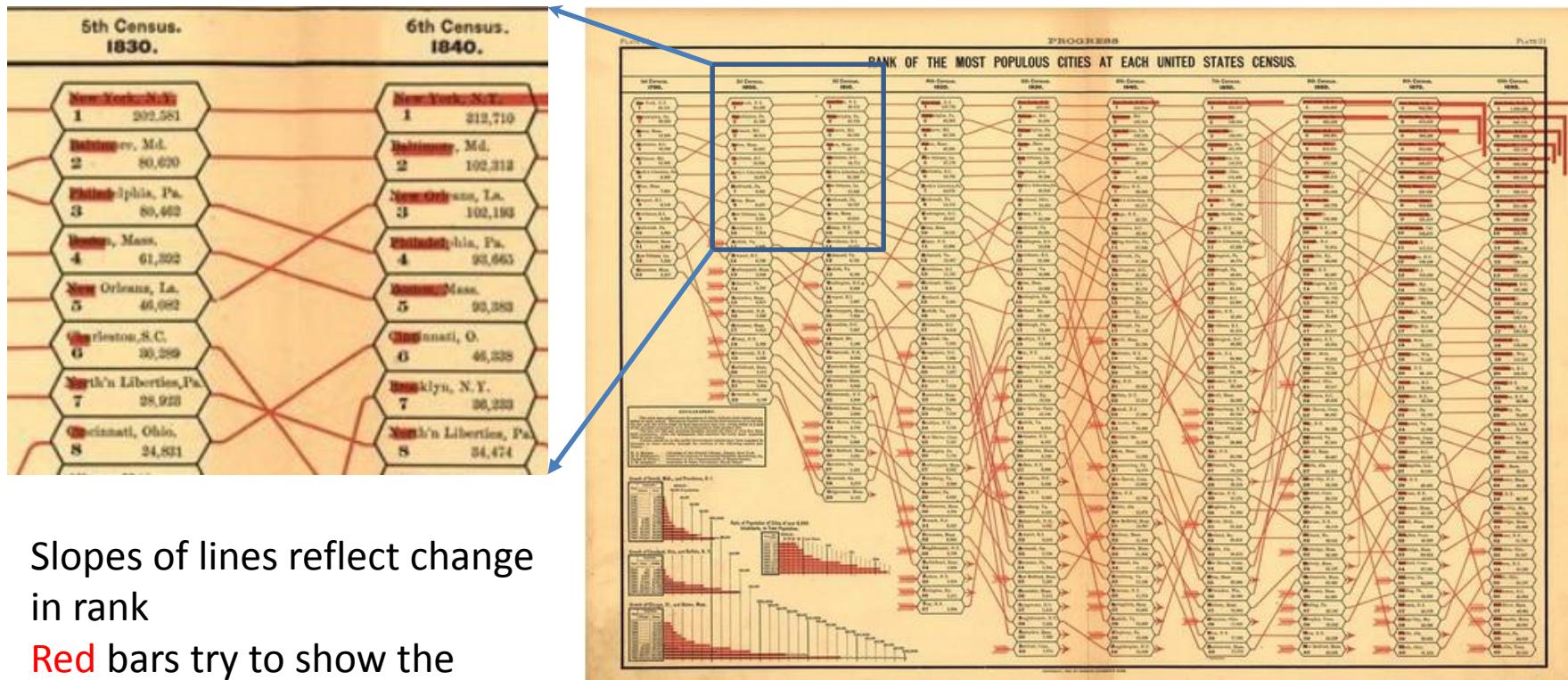
- To be fair, this was designed as **timeline of history** – a visual story of economics.
- History shown as a **strip-chart recording** (e.g., EKG)
- It was Playfair's last graph.



Playfair, W. (1824) *Chronology of Public Events and Remarkable Occurrences*.

Parallel ranked list charts

Another solution for multiple time series is to chart the **ranks** of observations and connect them with lines to show changes in relative position.



Slopes of lines reflect change
in rank
Red bars try to show the
numbers

Ranks of populations of US cities across census years
From: *Statistical Atlas of the United States* (1880)

Propensity of crimes across age

André-Michel Guerry (1833), rank order of crimes at different ages

- sexual assault **against adults** decreases with age
- sexual assault **against children** increases with age
- **parricide** relatively low until 60-70

INFLUENCE DE L'AGE.

DISTRIBUTION DES CRIMES AUX DIFFÉRENS ÂGES, PAR PÉRIODE DE DIX ANNÉES.

IV.

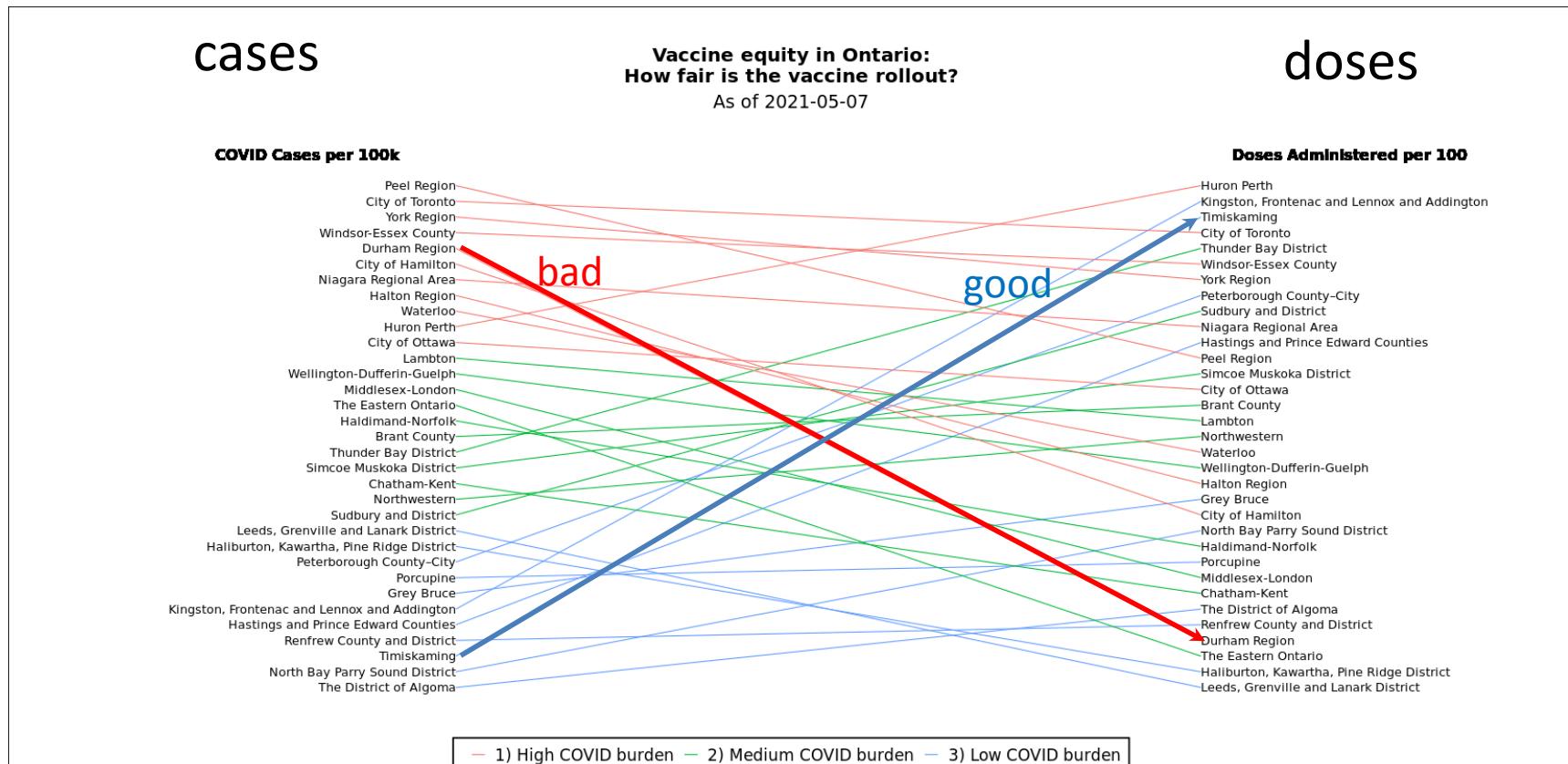
CRIMES CONTRE LES PERSONNES.												AU-DESSUS DE 70 ANS.		
AU-DESSOUS DE 21 ANS.			DE 21 À 30.		DE 30 À 40.		DE 40 À 50.		DE 50 À 60.		DE 60 À 70.		AU-DESSUS DE 70 ANS.	
	NATURE DES CRIMES.	Sur 1,000	NATURE DES CRIMES.	Sur 1,000										
1	Blessures et coups.	184	Blessures et coups.	216	Blessures et coups.	179	Assassinat	194	Meurtre	185	Meurtre	173	Viol sur des enfants	318
2	Viol sur des adultes.	169	Viol sur des adultes.	157	Viol sur des adultes.	157	Assassinat	161	Assassinat	182	Viol sur des enfants	165	Blessures et coups	137
3	Meurtre	147	Assassinat	140	Meurtre	152	Meurtre	155	Blessures et coups	175	Assassinat	159	Meurtre	125
4	Viol sur des enfants	147	Rebellion	141	Rebellion	110	Rebellion	100	Rebellion	98	Blessures et coups	135	Assassinat	102
5	Assassinat	101	Viol sur des adultes	105	Viol sur des adultes	73	Viol sur des enfants	94	Viol sur des enfants	68	Faux témoignage	98	Faux témoignage	102
6	Rebellion	79	Infanticide	83	Infanticide	63	Faux témoignage	69	Faux témoignage	76	Rebellion	78	Rebellion	91
7	Infanticide	48	Viol sur des enfants	58	Viol sur des enfants	59	Viol sur des adultes	61	Viol sur des adultes	52	Infanticide	43	Empoisonnement	25
8	Bless. env. ascend.	47	Bless. env. ascend.	50	Bless. env. ascend.	59	Bless. env. ascend.	49	Infanticide	24	Empoisonnement	35	Infanticide	25
9	Associat. de malfait.	52	Faux témoignage	33	Faux témoignage	41	Infanticide	41	Empoisonnement	20	Parricide	21	Viol sur des adultes	23
10	Faux témoignage	29	Empoisonnement	18	Empoisonnement	25	Empoisonnement	23	Parricide	18	Associat. de malfait.	11	Associat. de malfait.	11
11	Empoisonnement	11	Crim. env. des enf.	10	Crim. env. des enf.	16	Associat. de malfait.	19	Avortement	19	Avortement	14	Voies de fait, etc.	11
12	Crim. env. des enf.	6	Associat. de malfait.	10	Associat. de malfait.	12	Bigamie	13	Avortement	15	Crim. env. des enf.	11	Bless. env. ascend.	11
13	Voies de fait, etc.	6	Parricide	8	Parricide	9	Bigamie	13	Bigamie	15	Bless. env. ascend.	2	Parricide	11
14	Mend. av. violence.	6	Mend. av. violence.	6	Mend. av. violence.	8	Mendicité	8	Mendicité	10	Associat. de malfait.	7	Crim. env. les enf.	11
15	Crim. env. des enf.	5	Autres crimes	5	Autres crimes	8	Parricide	7	Parricide	10	Bigamie	7	Avortement	11
16	Parricide	5	Avortement	5	Avortement	7	Avortement	7	Avortement	10	Bigamie	7	Bigamie	11
17	Avortement	5	Bigamie	2	Bigamie	6	Parricide	4	Crim. env. des enf.	6	Voies de fait, etc.	7	Voies de fait, etc.	11
18	Bigamie	1	Voies de fait, etc.	6	Voies de fait, etc.	6	Voies de fait, etc.	4	Voies de fait, etc.	6	Mend. av. violence.	7	Mend. av. violence.	11
19	Autres crimes	7	Autres crimes	12	Autres crimes	12	Autres crimes	12	Autres crimes	26	Autres crimes	11	Autres crimes	34
	TOTAL	1,000		1,000		1,000		1,000		1,000		1,000		1,000

Friendly, (2007). "A.-M. Guerry's Moral Statistics of France: Challenges for Multivariable Spatial Analysis," *Statistical Science*, vol. 22, no. 3. <https://www.datavis.ca/papers/guerry-STS241.pdf>

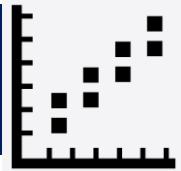
Slope graph of COVID: Cases vs. Doses

Vaccine equity → all lines should be ≈ flat

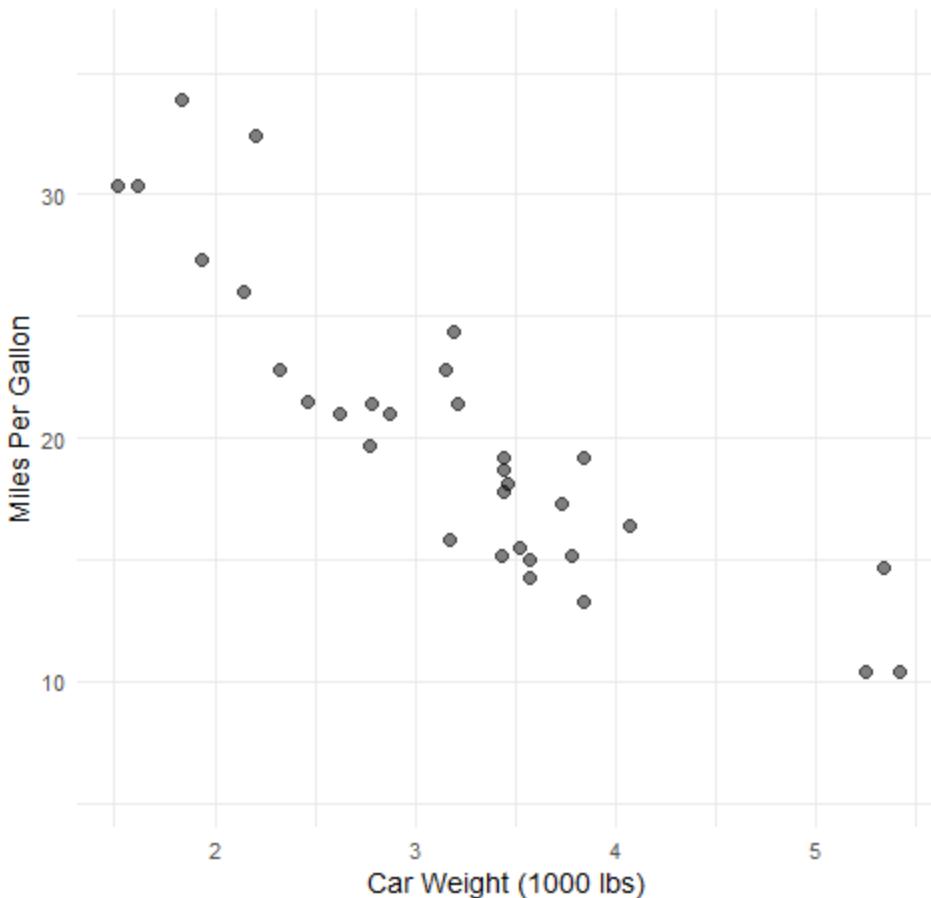
Which health regions stand out?
How could this graph be better?



2D: Scatterplots



1974 Motor Trend Cars: Gas Mileage vs. Weight [Frame 1]



Goals: Want to see--

- Overall pattern:
 - direction
 - form
 - strength
- Deviations from pattern:
 - residuals
 - outliers

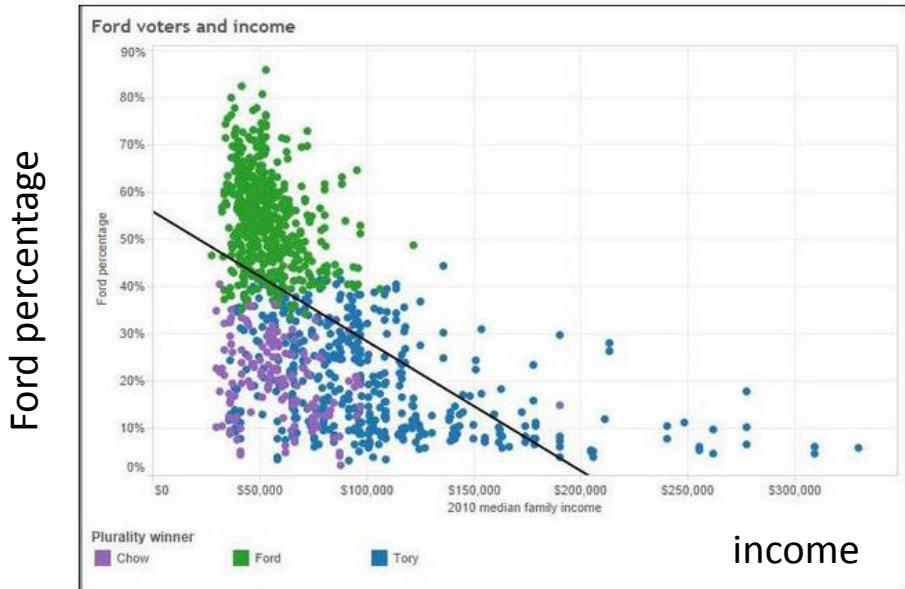
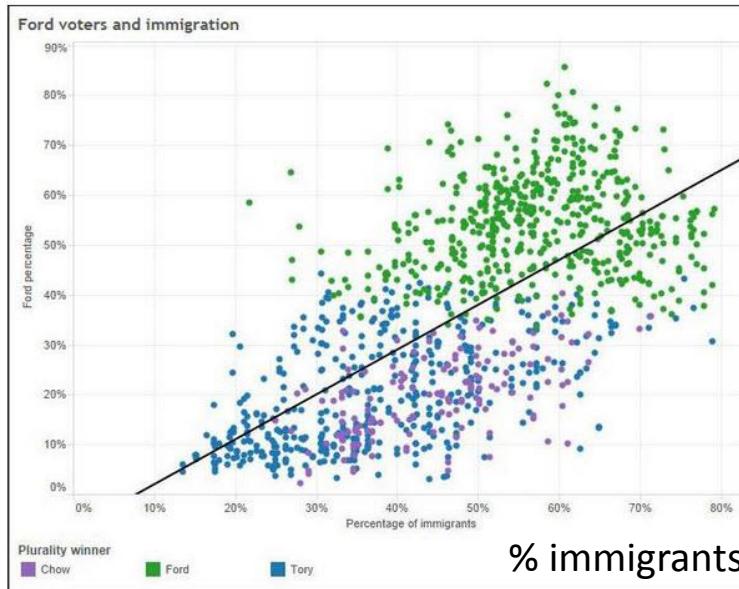
Annotations can help:

- regression line
- confidence band
- smoothed curve (nonlinear?)
- residuals

2D: Scatterplots: Ford Nation



Who voted for Rob Ford in the 2014 Toronto mayoral election?



These simple scatterplots by data journalist Patrick Cain use simple enhancements:

- Color, for candidate (Chow, Ford, Tory)
- Overall regression line

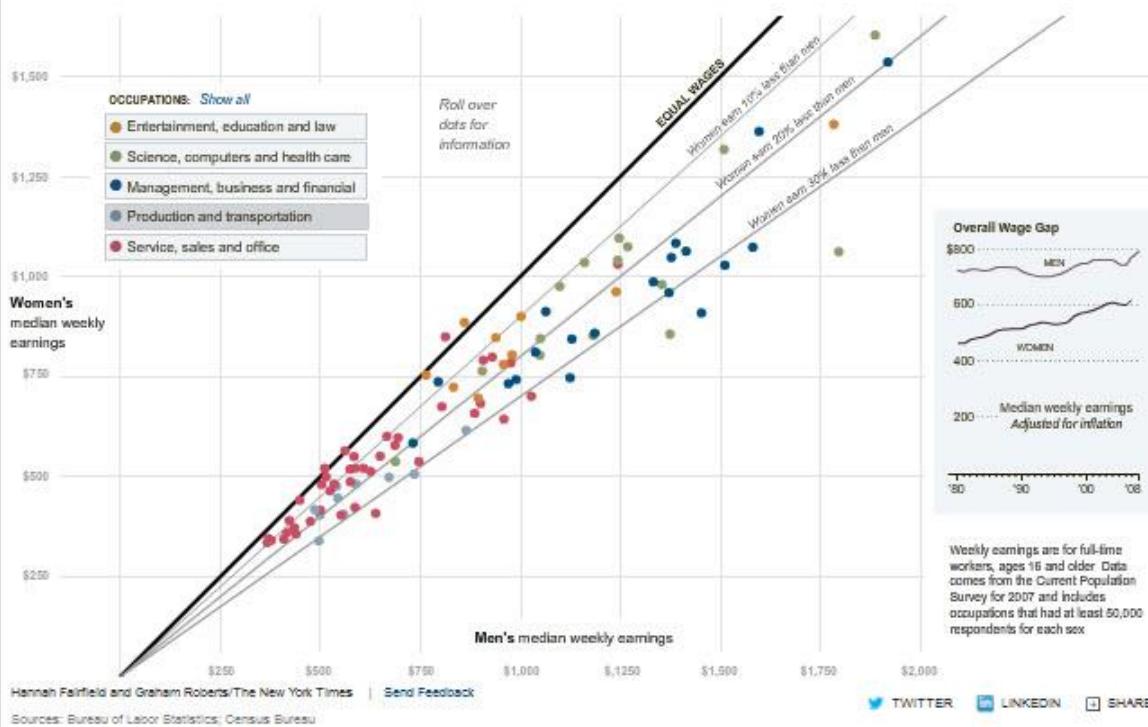
Source: <https://globalnews.ca/news/1652571/ford-nation-2014-15-things-demographics-tell-us-about-toronto-voters/>

Scatterplots: Wage gap

Published: May 18, 2010

Why Is Her Paycheck Smaller?

Nearly every occupation has the gap — the seemingly unbridgeable chasm between the size of the paycheck brought home by a woman and the larger one earned by a man doing the same job. Economists cite a few reasons: discrimination as well as personal choices within occupations are two major factors, and part of the gap can be attributed to men having more years of experience and logging more hours.



How to compare salaries of men & women in different occupations?

The NYT chose to plot median salaries for women against those for men, in different occupational groups

The 45° line represents wage parity
Other lines show 10, 20, 30% less for women

How else to show this?

Alberto Cairo, *The Truthful Art*, Fig 9.19, from:

http://www.nytimes.com/interactive/2009/03/01/business/20090301_WageGap.html

Scatterplots: InfoVis

This graph, from fivethirtyeight.com was designed to show how some presidential candidates had shifted positions before the 2016 election.

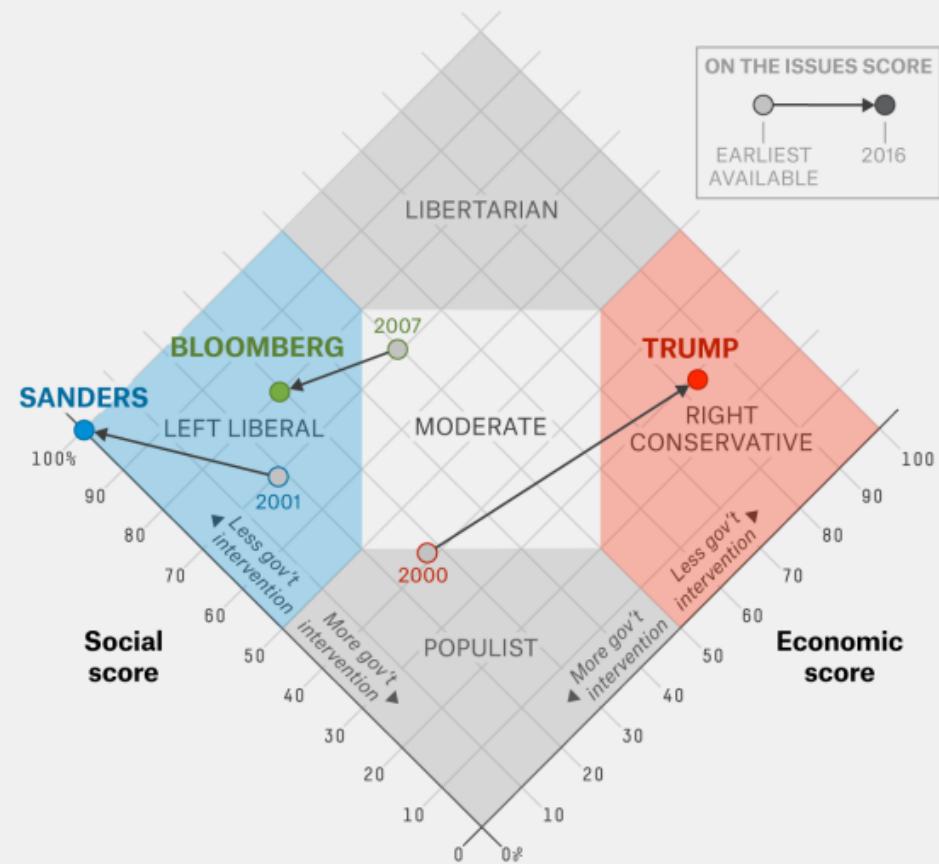
The axes are a score on **social** and **economic** policy, but they rotate the axes by 45° to create zones related to political thought.

This info graphic is **eye-catching** and **self-explanatory**:

- colored/labeled zones
- interpretive labels on axes
- arrows showing movement to extremes

Candidates abandoning the middle ground

Earliest available and current OnTheIssues score for Bernie Sanders, Donald Trump and Michael Bloomberg



Scatterplots: Annotations enhance perception

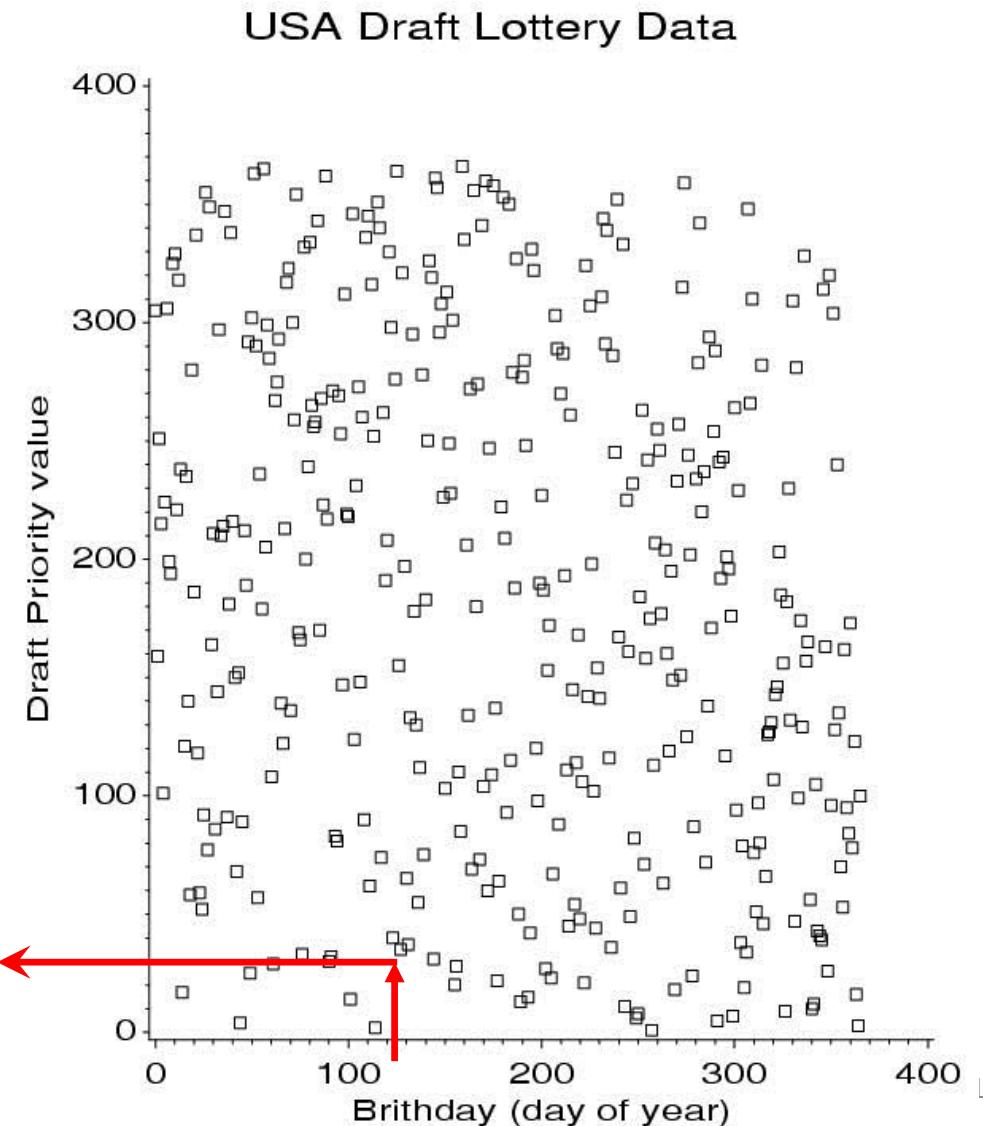
Data from the US draft lottery,
1970

- Birth dates were drawn at random to assign a “draft priority value” (1=bad)
- Can you see any pattern or trend?

This is an example of data with a weak signal and a lot of noise



Me (May 7):
127 → priority = 35

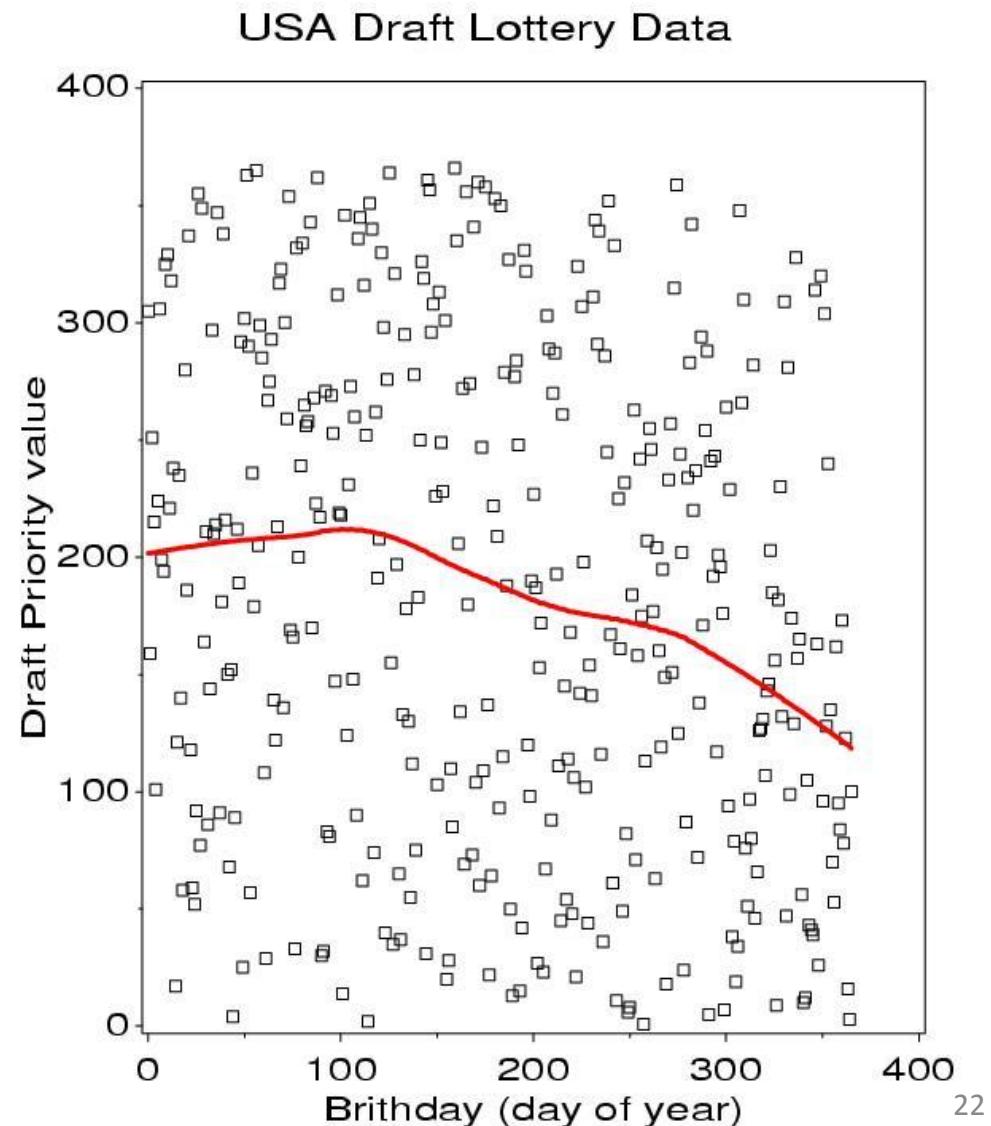


Scatterplots: Smoothing enhances perception

Drawing a smooth curve shows a systematic decrease toward the end of the year.

- The smooth curve is fit by **loess**, a form of non-parametric regression.

Visual explanation:

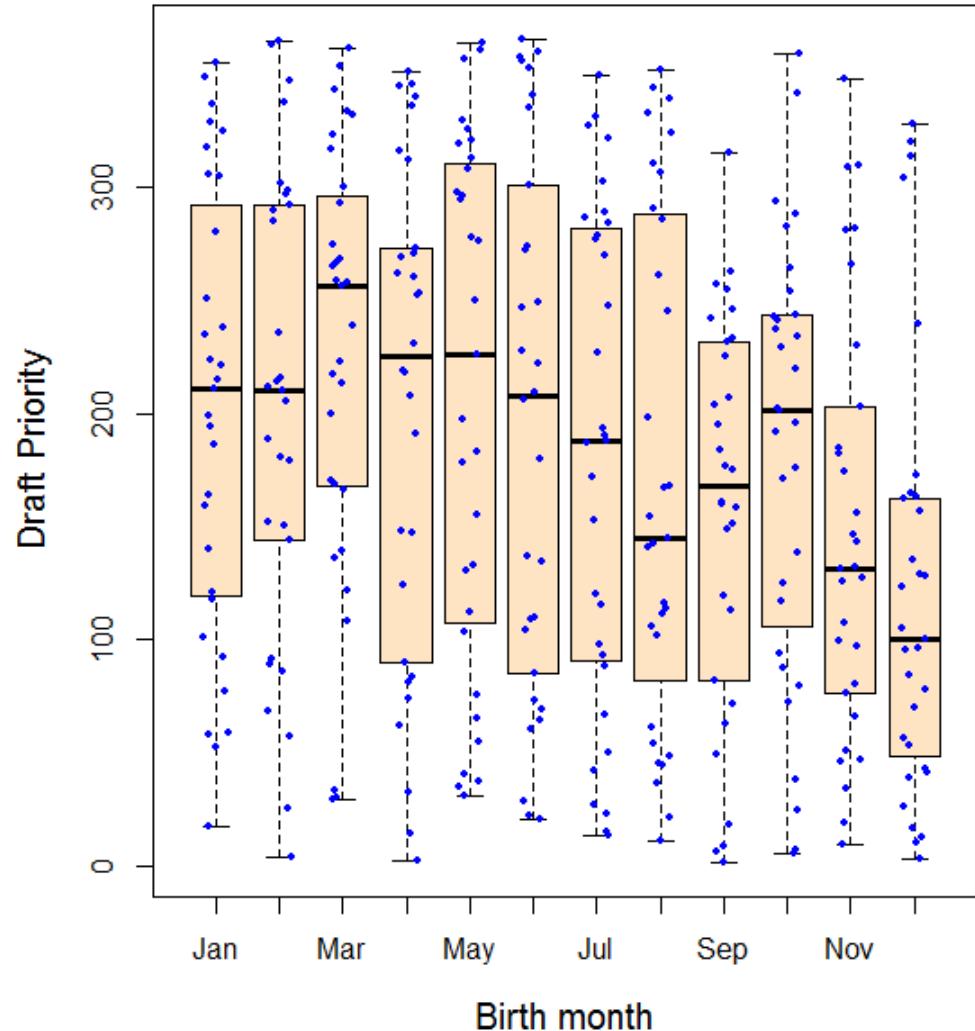


Smoothing by grouping and summarization

Another form of smoothing is to make one variable discrete & show a graphical summary – here a boxplot

The decrease in later months becomes apparent

Perception: the boxplots form the foreground; the jittered points show the data



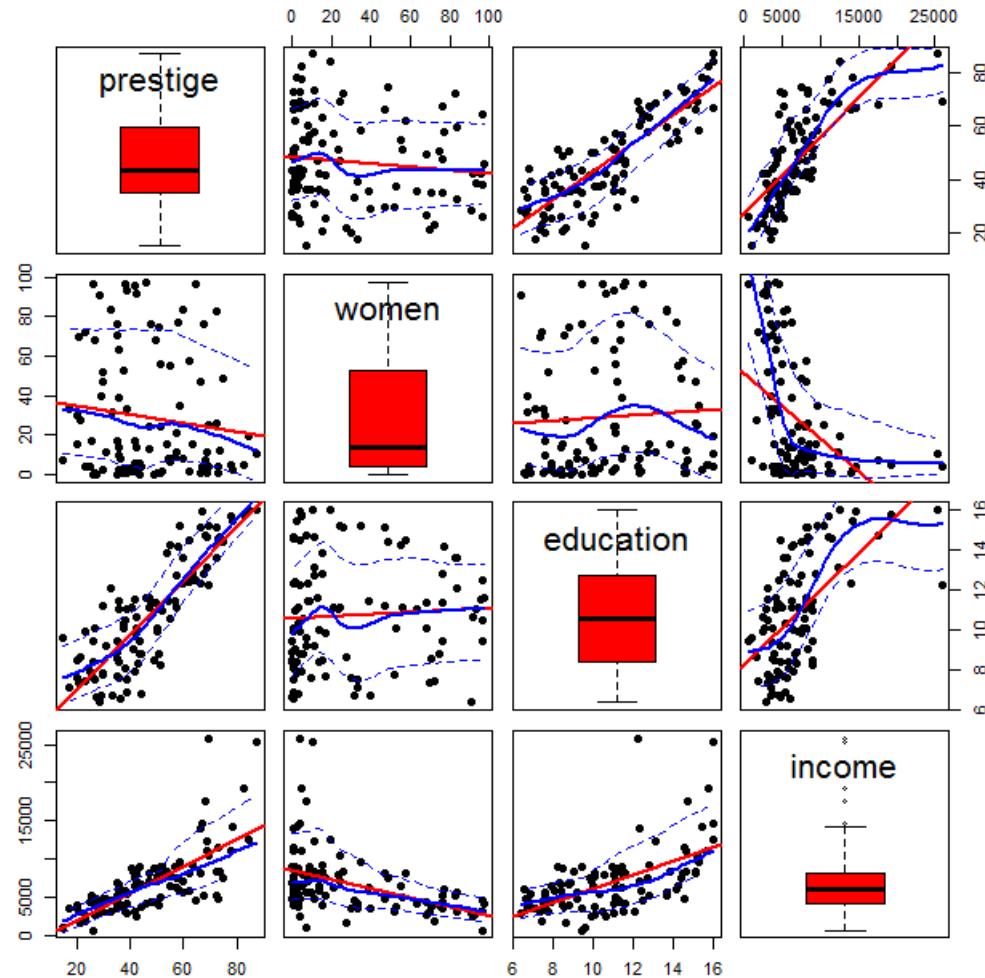
Scatterplot matrices

A scatterplot matrix shows the bivariate relation between all **pairs** of variables. Seeing these all together is more useful than a collection of separate plots.

How does occupational prestige depend on %women, education and income?

The individual plots are enhanced with linear regression lines and non-parametric smooths to show non-linearity

This figure uses `scatterplotMatrix()` in the [car](#) package. There are many options.



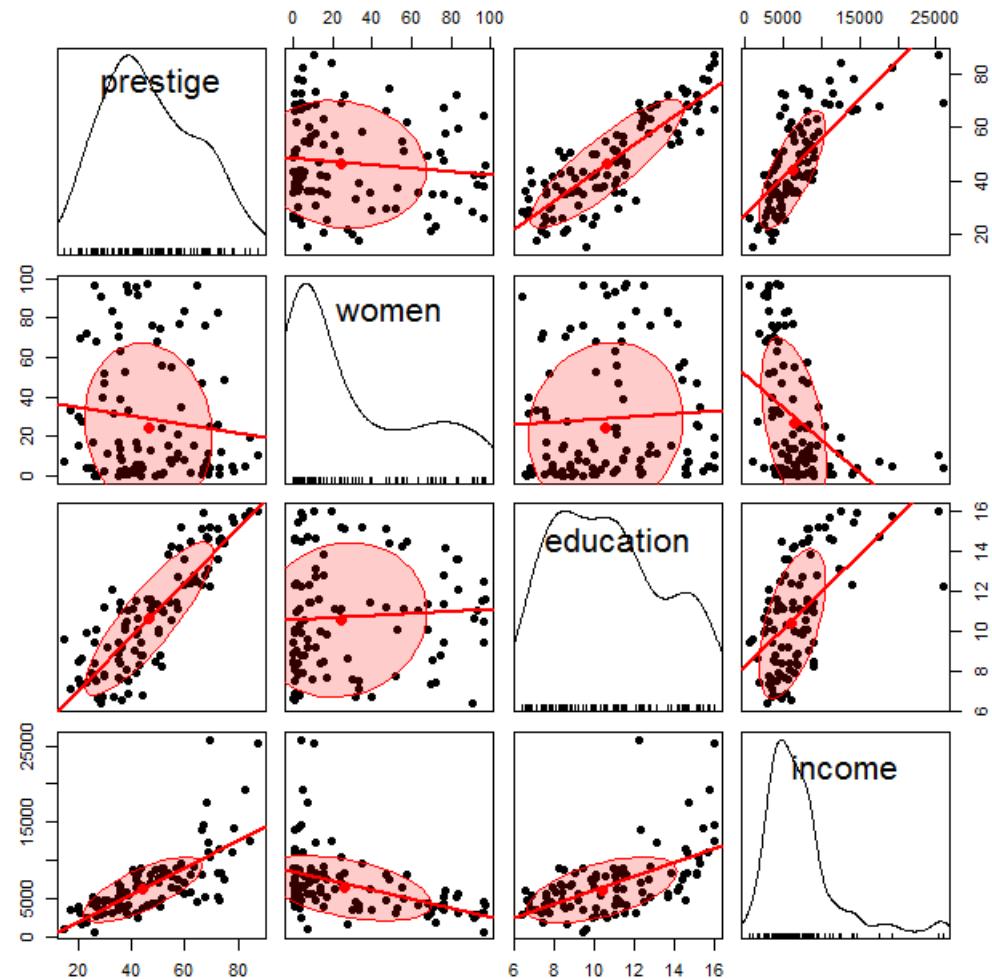
Scatterplot matrices

Density plots are often more useful for showing the shapes of distributions

- women: bimodal
- income: highly skewed

A data ellipse gives a visual summary of the direction and strength of the relationship

Again, graphical annotation provides aids for interpretation.



Larger data sets

Scatterplot matrices hold up reasonably well with a larger number of variables

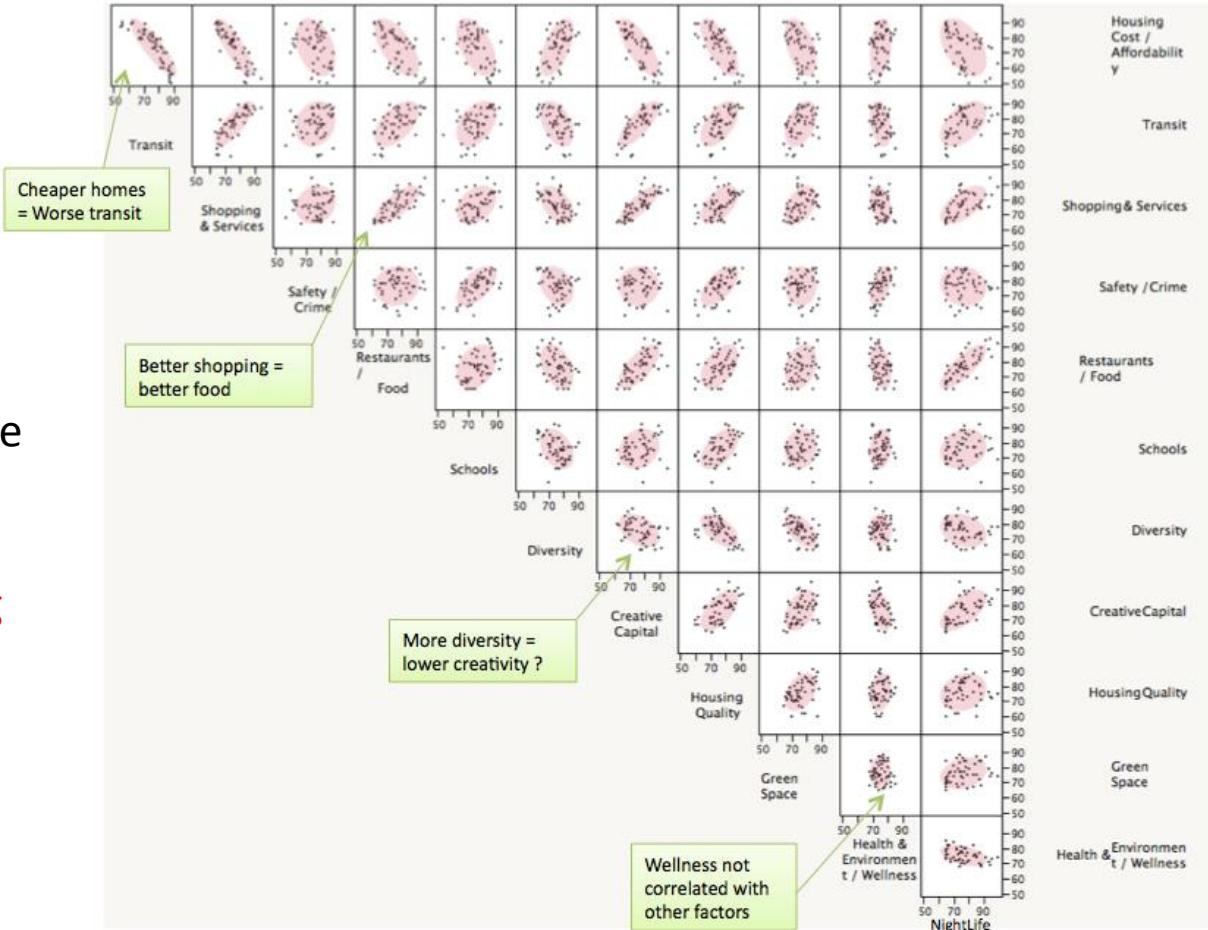
Where to live in NYC?

This SPM shows 12 variables on ~ 60 neighborhoods

The data **ellipses** provide a visual summary

I call this **visual thinning** – reducing details in a larger picture

In an interactive display we can **zoom** in/out



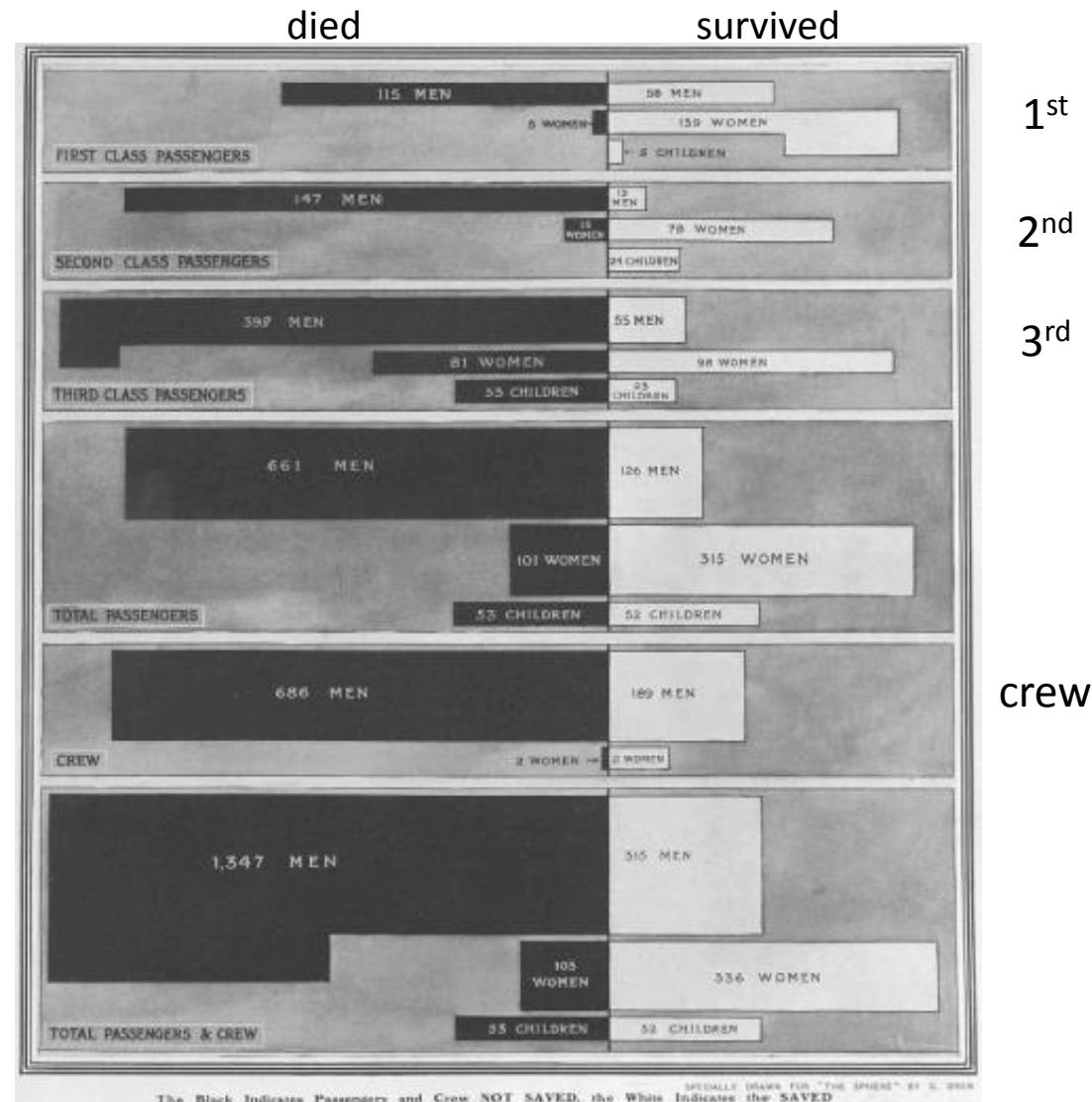
Categorical data

This remarkable chart shows survival on the *Titanic*, by Class for passengers and Gender and Age.

It was drawn by G. Bron, a graphic artist, and published in *The Sphere*, one month after the *Titanic* sank.

It uses back-to-back bar charts, with area ~ frequency

See our web page:
<http://datavis.ca/papers/titanic/>



Categorical data: Mosaic plots

Similar to a grouped bar chart

Shows a frequency table with tiles,
area \sim frequency

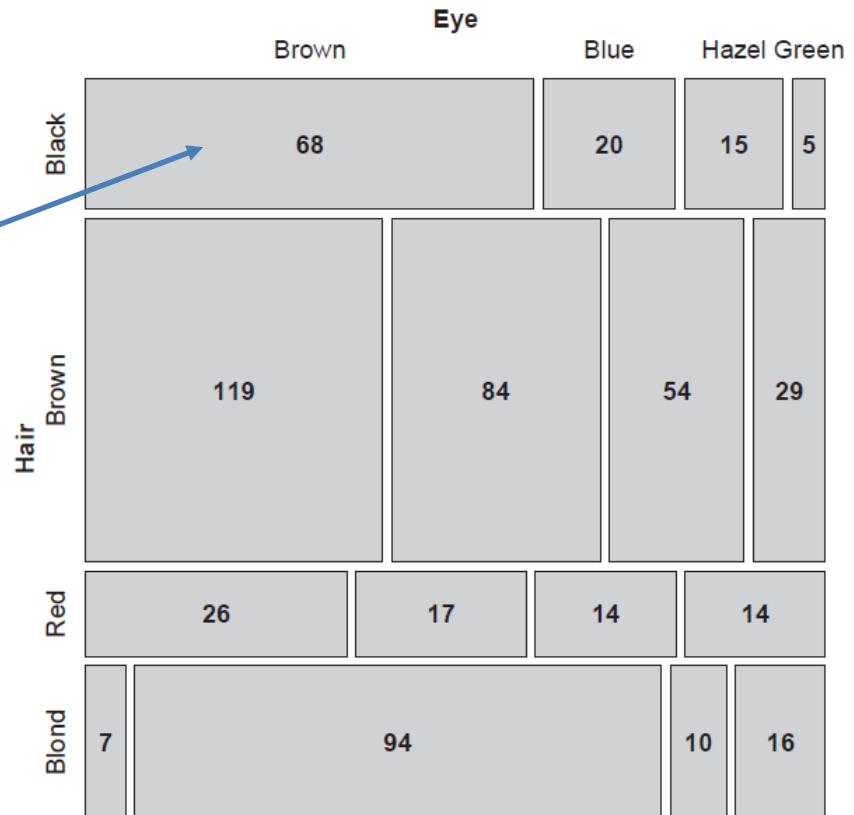
```
> data(HairEyeColor)
> HEC <- margin.table(HairEyeColor, 1:2)
> HEC
```

		Eye			
Hair		Brown	Blue	Hazel	Green
Black		68	20	15	5
Brown		119	84	54	29
Red		26	17	14	14
Blond		7	94	10	16

```
> chisq.test(HEC)
```

Pearson's Chi-squared test

```
data: HEC
X-squared = 140, df = 9, p-value <2e-16
```



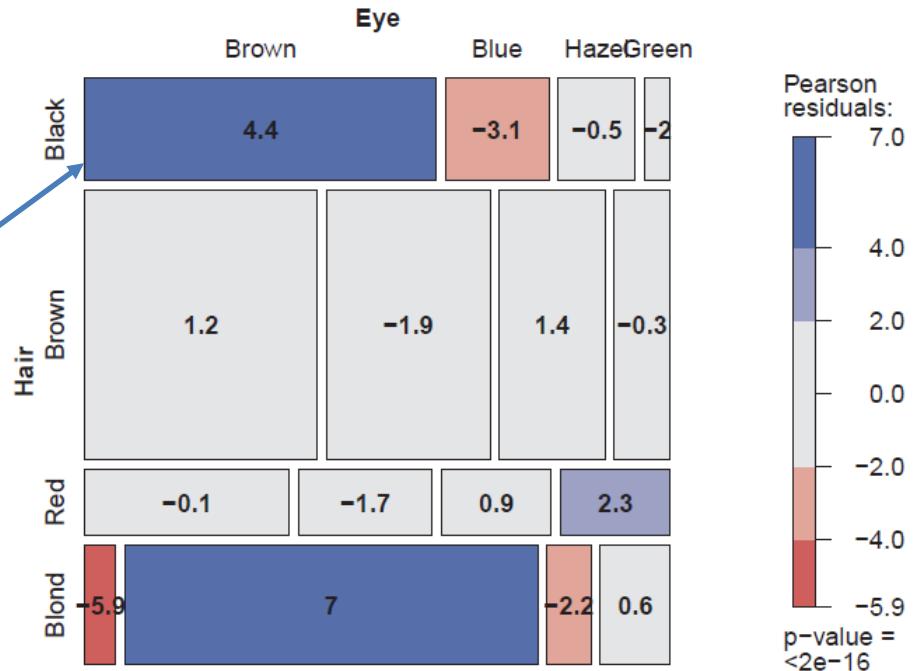
How to understand the association
between hair color and eye color?

Mosaic plots

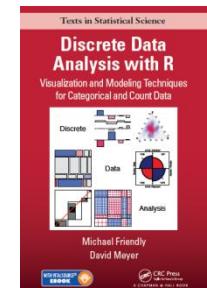
Shade each tile in relation to the contribution to the Pearson χ^2 statistic

$$\chi^2 = \sum r_{ij}^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

```
> round(residuals(chisq.test(HEC)), 2)
   Eye
Hair  Brown Blue Hazel Green
Black  4.40 -3.07 -0.48 -1.95
Brown  1.23 -1.95  1.35 -0.35
Red    -0.07 -1.73  0.85  2.28
Blond -5.85  7.05 -2.23  0.61
```

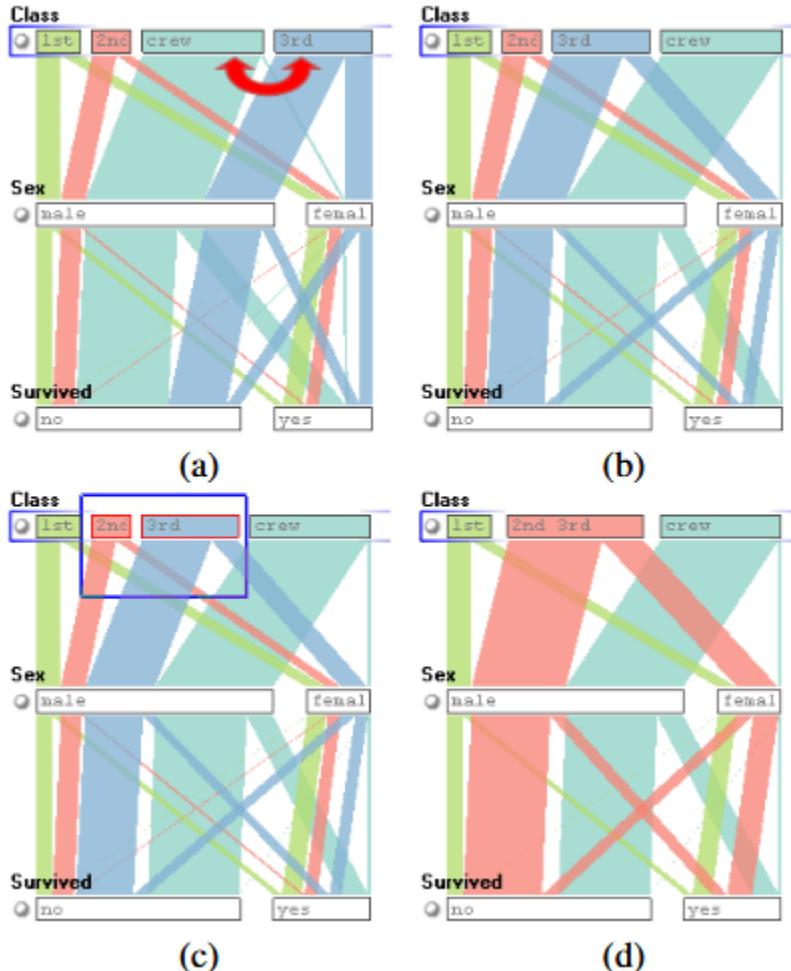


Mosaic plots extend readily to 3-way + tables
 They are intimately connected with loglinear models
 See: Friendly & Meyer (2016), Discrete Data Analysis with R, <http://ddar.datavis.ca/>



Parallel Sets

Titanic data: Who survived?



Parallel sets use **parallel coordinate** axes to show the relations among categorical variables.

The frequencies of one variable (Class) are sub-divided according to the joint frequencies in the next (Sex) and shown by the width of the connecting line.

The ParSets application is interactive:

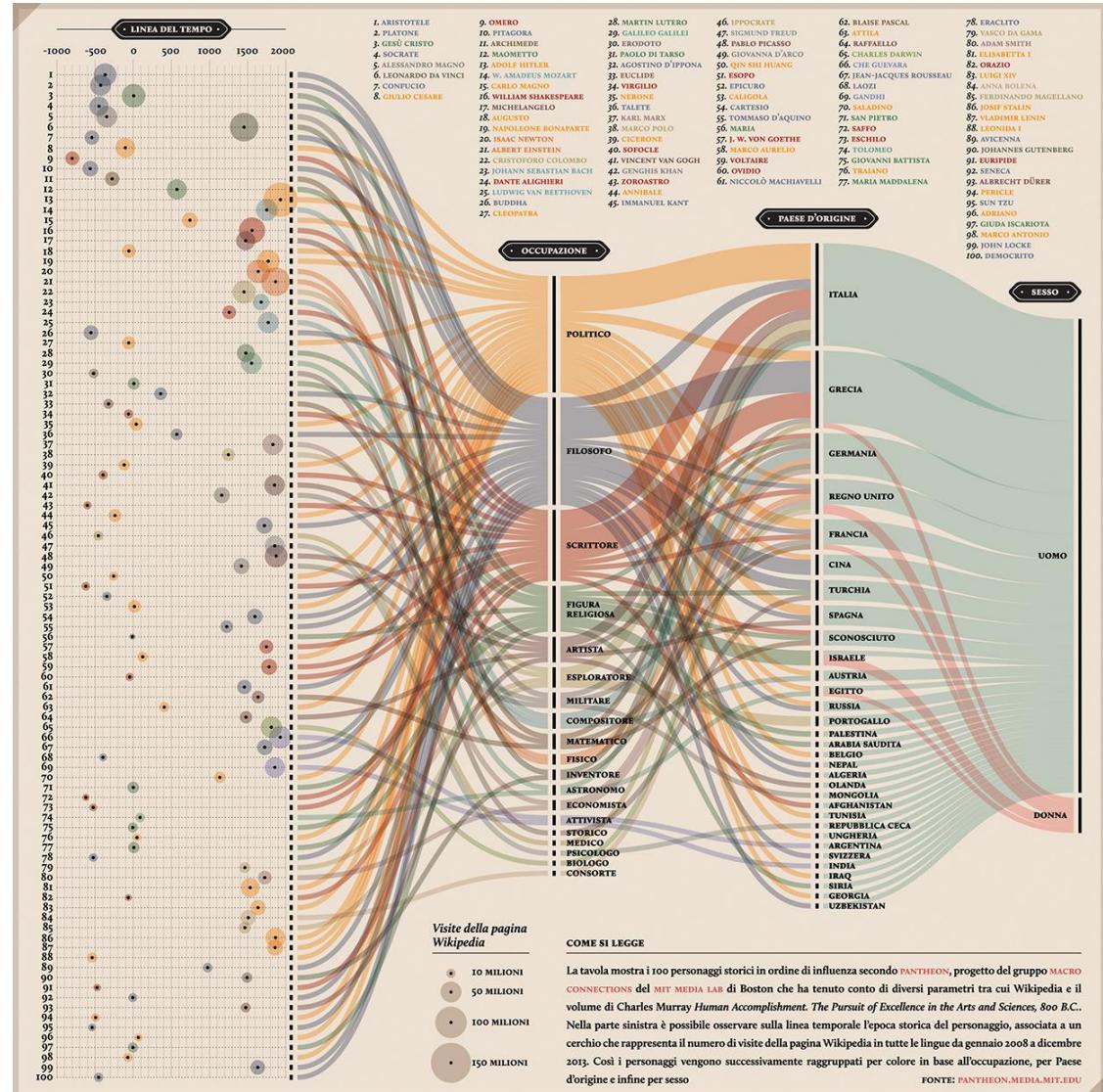
- categories can be reordered (a, b)
- categories can be grouped (c, d)

Sankey diagram

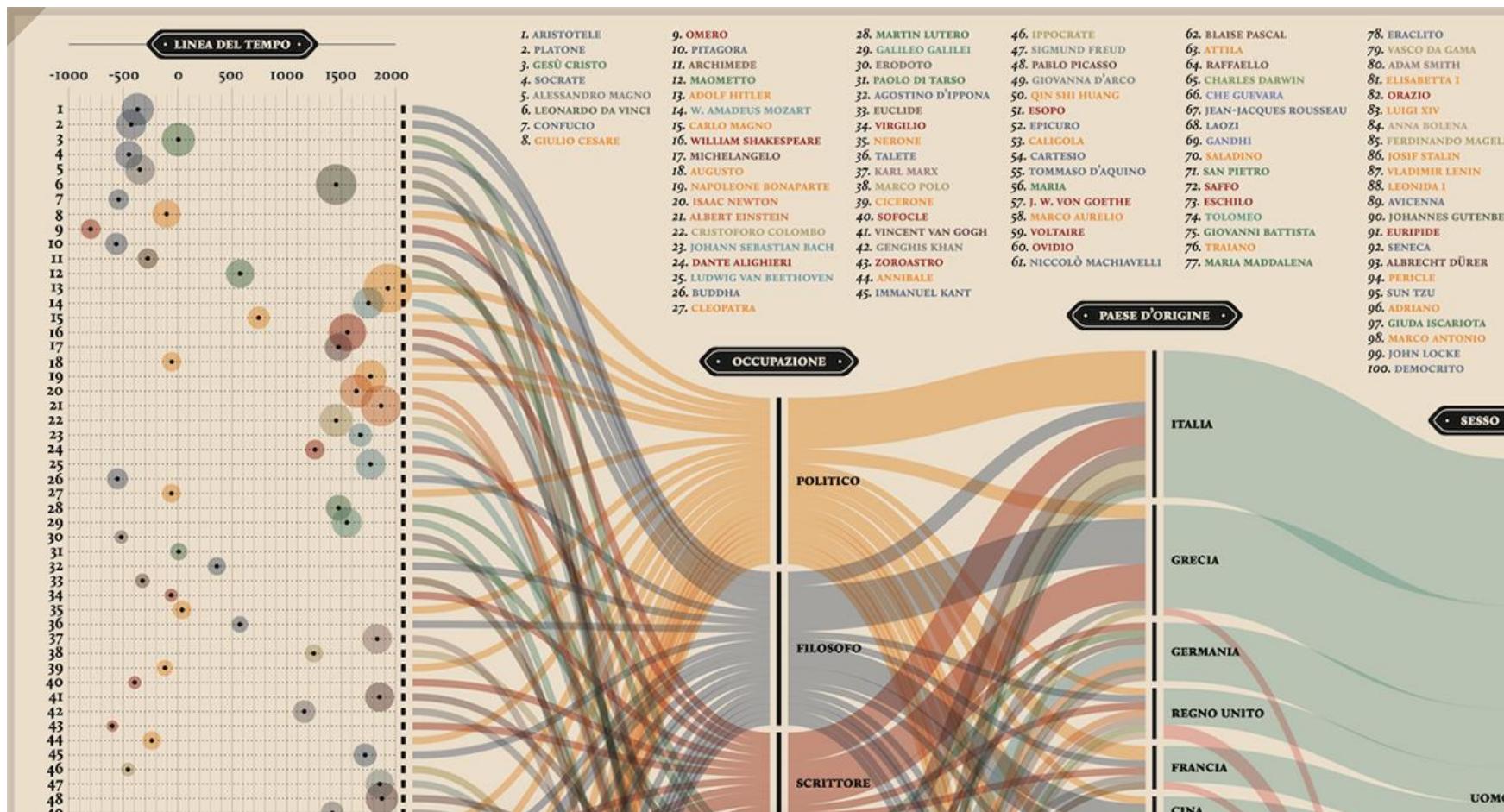
Pantheon, by Valerio Pellegrini
 Visualizing the 100 most
 influential figures in History
 (Wikipedia visits)

Columns show **occupation**,
country of origin and
gender

Flow lines link individuals to
 the column variables, width
 ~ influence



Sankey diagram



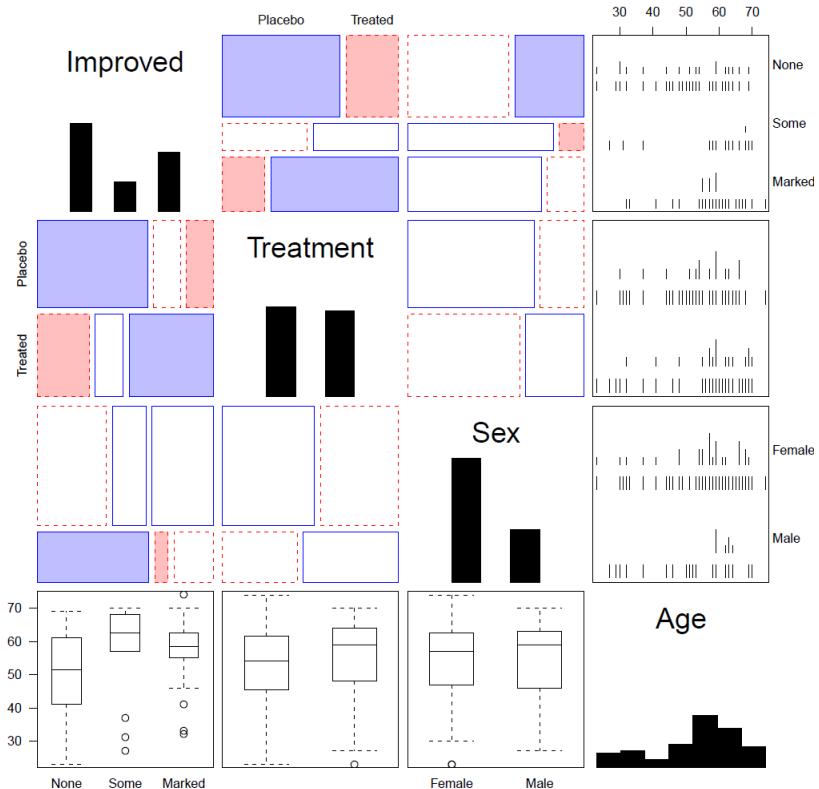
Multiple dimensions of the most influential people in history

From: <http://visualoop.com/blog/83382/pantheon-by-valerio-pellegrini>

Generalized pairs plots

Generalized pairs plots from the [gpairs](#) package handle both categorical (**C**) and quantitative (**Q**) variables in sensible ways

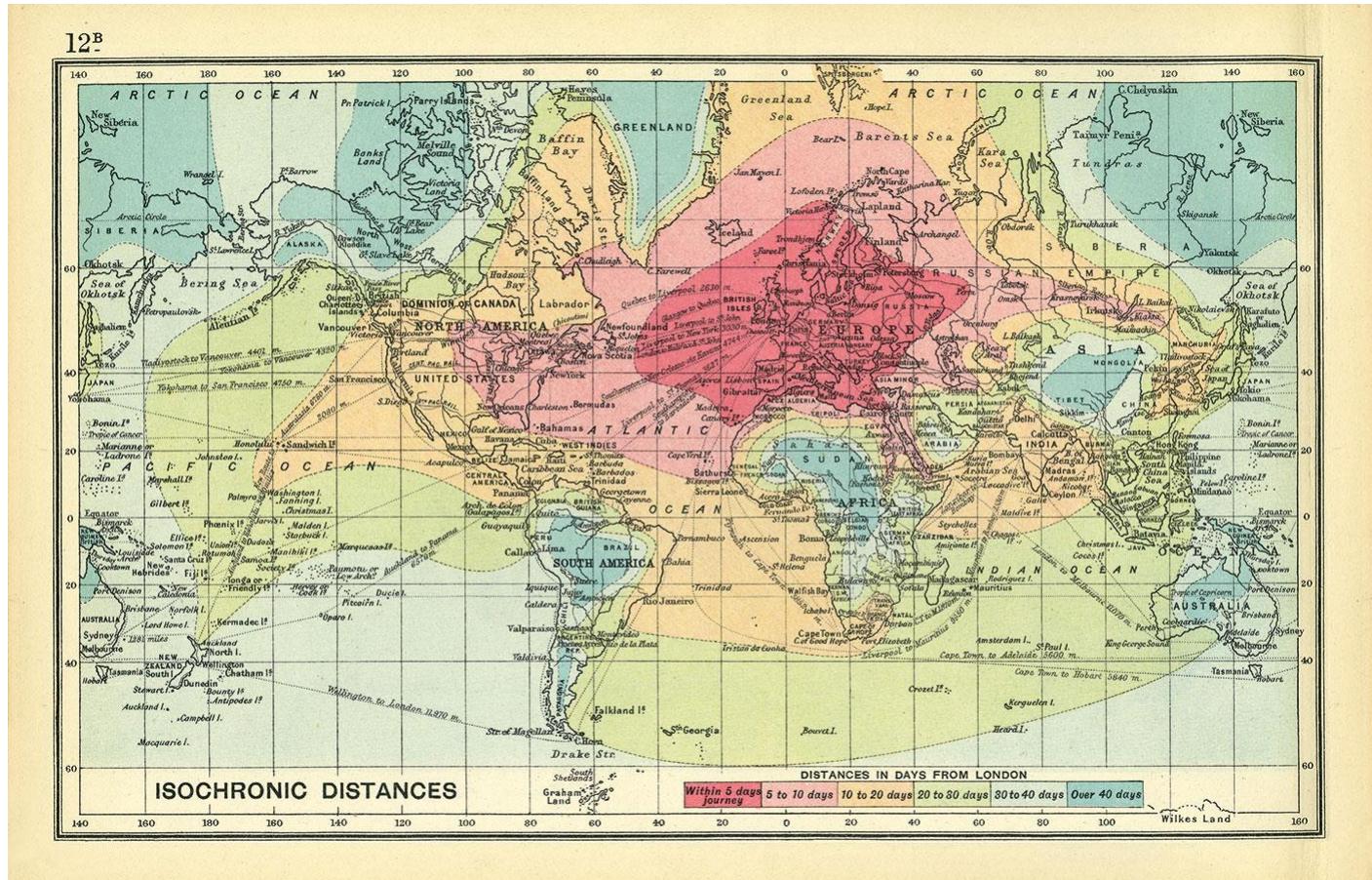
x	y	plot
Q	Q	scatterplot
C	Q	boxplot
Q	C	barcode
C	C	mosaic



```
library(gpairs)
data(Arthritis)
gpairs(Arthritis[, c(5, 2:5)], ...)
```

3D: Iso-contour maps

Early attempts to show 3D data used **contours of equal value** on a map
The data was actually very thin; the contours the result of imaginative smoothing



Francis Galton, *Isochronic chart of travel time, 1881*

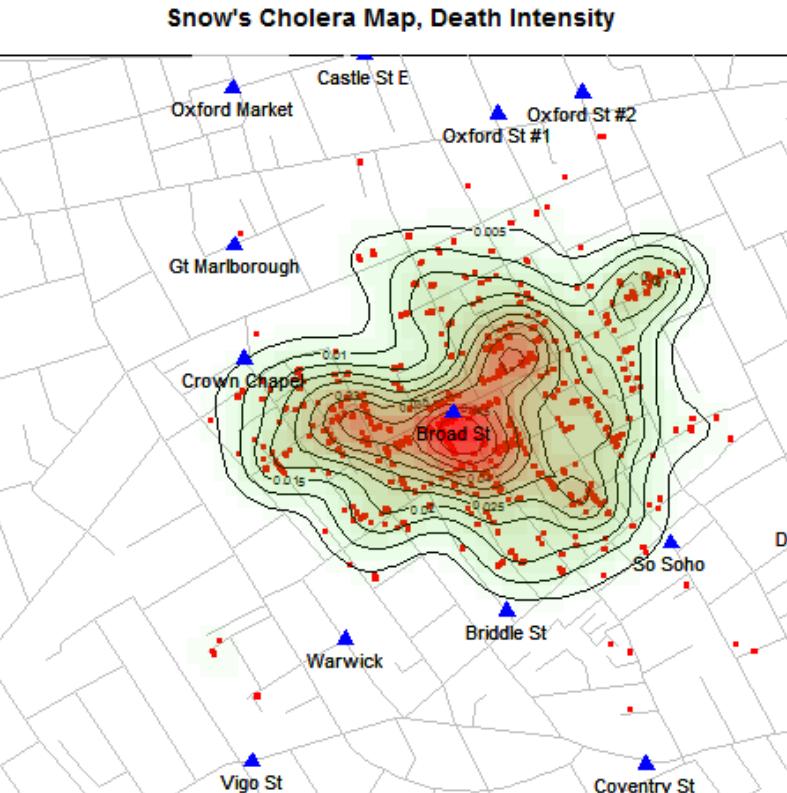
3D: Bivariate density estimation

John Snow's map of cholera deaths in London, 1854



Broad St. pump

Modern statistical techniques can compute contours of constant density



Data: HistData package for R



3D: population pyramid

Italian demographer Luigi Perozzo (1880) develops the first true 3D diagram showing the population of Sweden over years and age groups as a 3D surface

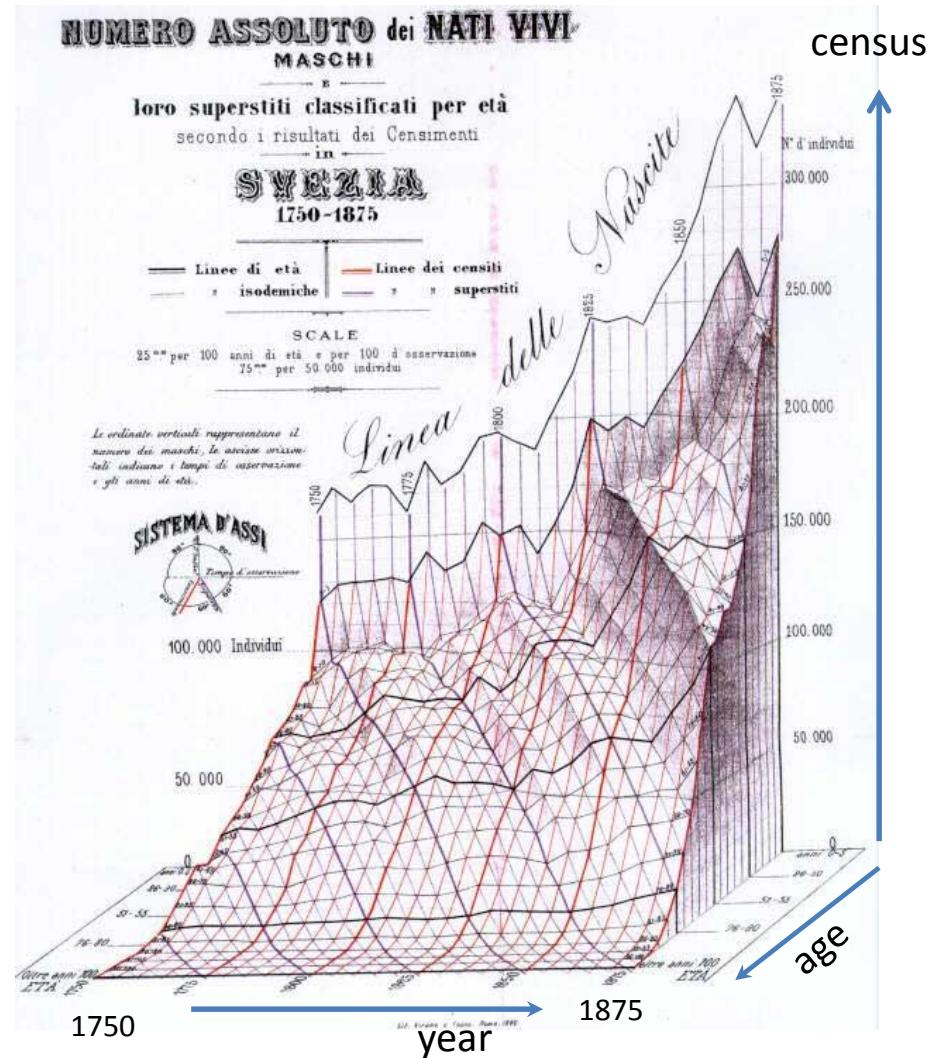
Census counts for a given **year** are shown by the red lines

Survival of a given **age** are shown by black lines

Cohorts are shown by lines down & to the right

These 3 variables are primary in demography.

A mystery here: what caused the decline at the upper right?

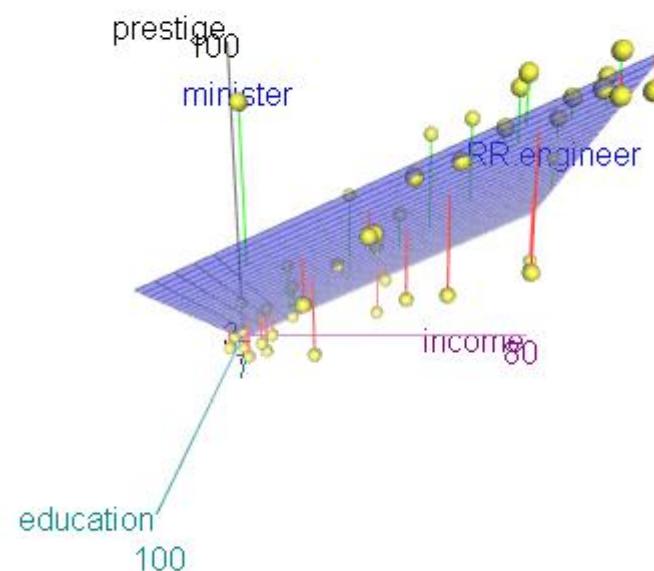


3D: scatterplot & regression surface

How does occupational prestige depend on income & education?

This plot shows the data and a fitted multiple regression surface, connecting the points to the regression plane

It is hard to see in a static view, but easier when the plot is rotated dynamically



This plot is produced in R, using the [car](#) and [rgl](#) packages

```
data("Duncan", package="car")
scatter3d(prestige ~ income + education, data=Duncan, id.n=2)
movie3d(spin3d(c(0,1,0), rpm=6), duration=6, movie="duncan-reg3d")
```

Thematic maps & Spatial visualization

Thematic maps use a wide variety of techniques to display quantitative or qualitative variables on the geographic framework of a map

Once the domain of cartographers, these ideas are now being developed as an area of geospatial visualization and geospatial statistical methods

	Point	Linear	Areal	2½-D	True 3-D
Spacing					
Size					
Perspective Height					None Possible
Orientation					
Shape					
Arrangement					
Lightness					

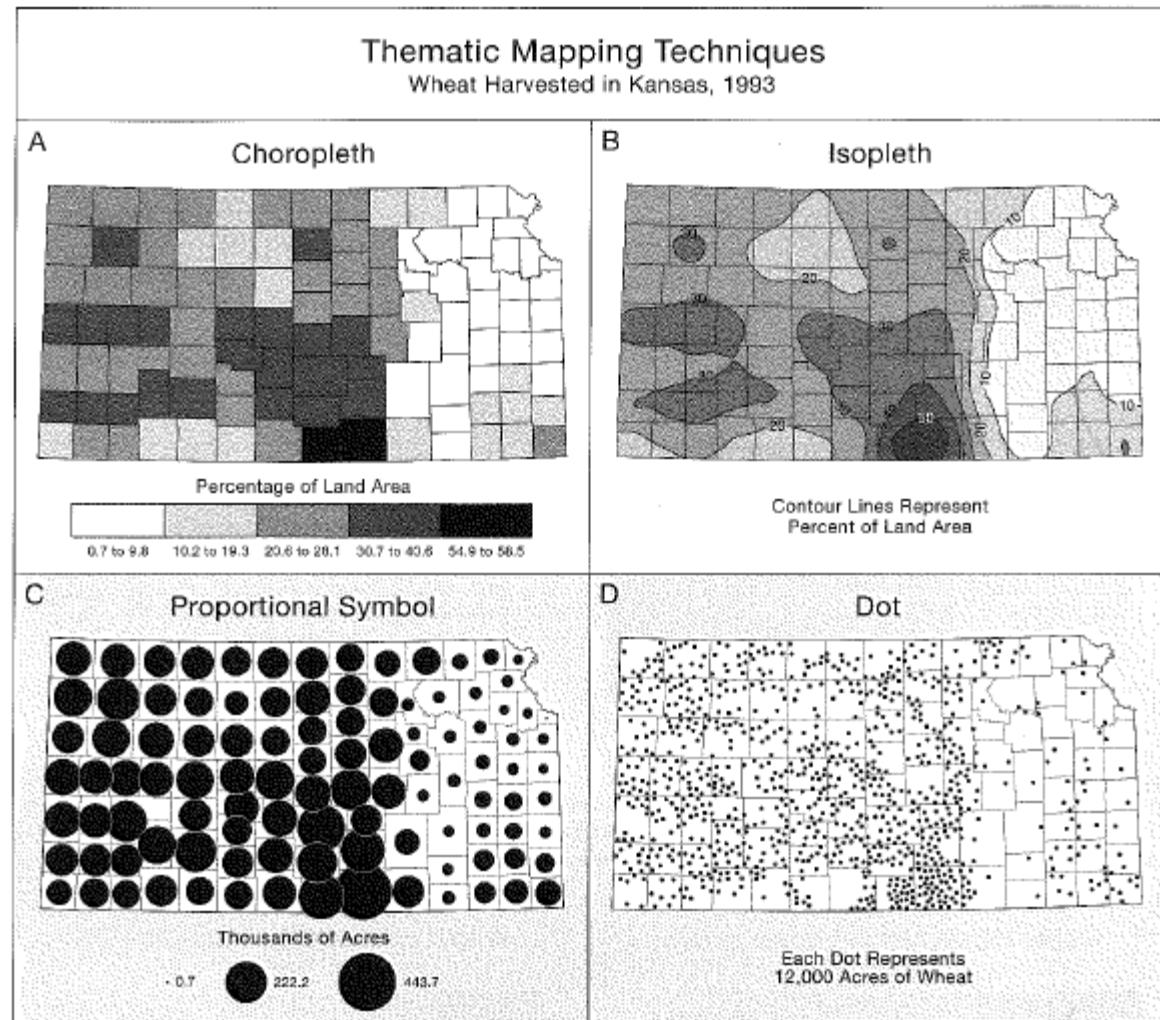
From: Slocum et al., *Thematic cartography and geographical visualization*, Fig 4.3

Thematic maps: Types

Basic types of thematic maps

Most are direct mappings of numbers to visual variables

Isopleth maps combine some analysis (contours) with display



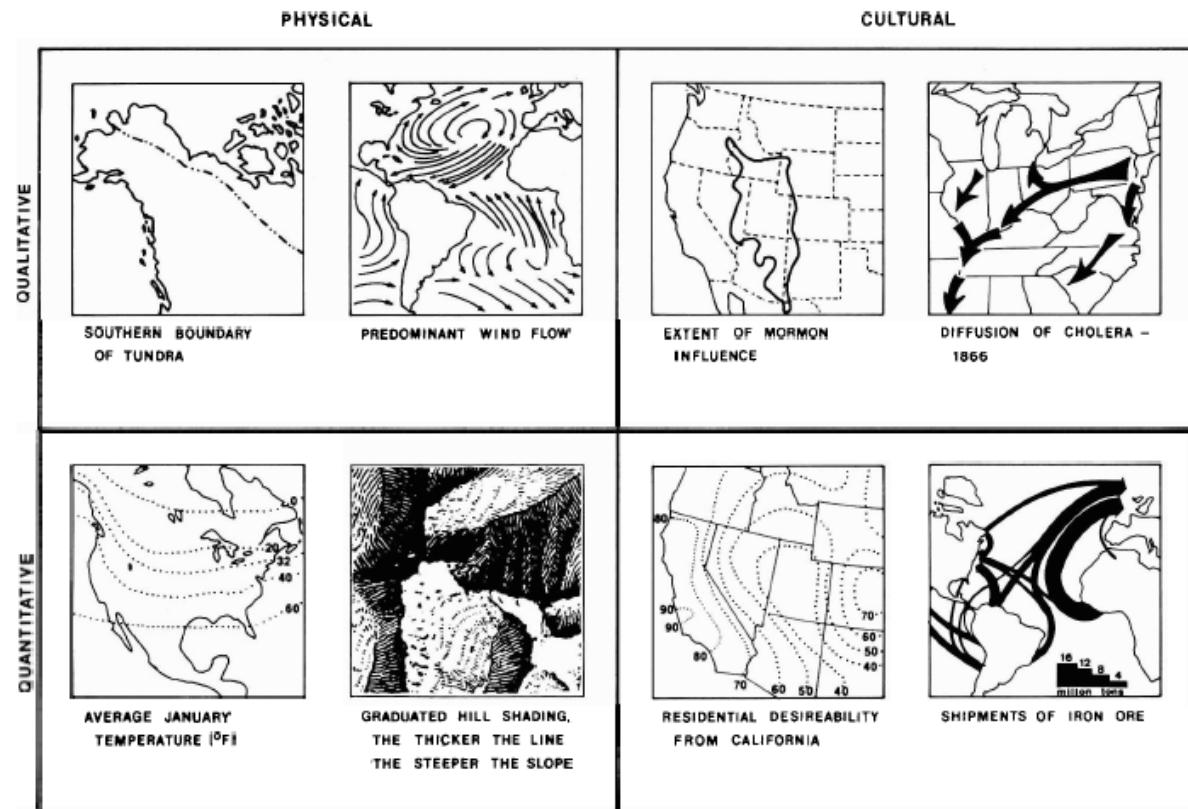
From: Slocum et al., *Thematic cartography and geographical visualization*, Fig 4.9

Thematic maps: Theory

Alan MacEachern (1979) classifies point, line and area symbols on thematic maps according to whether they depict **quantitative** or **qualitative** phenomena, in the **physical** or **cultural** domain.

This is a coarse classification.

Theories, ideas, and methods have advanced considerably since this time.

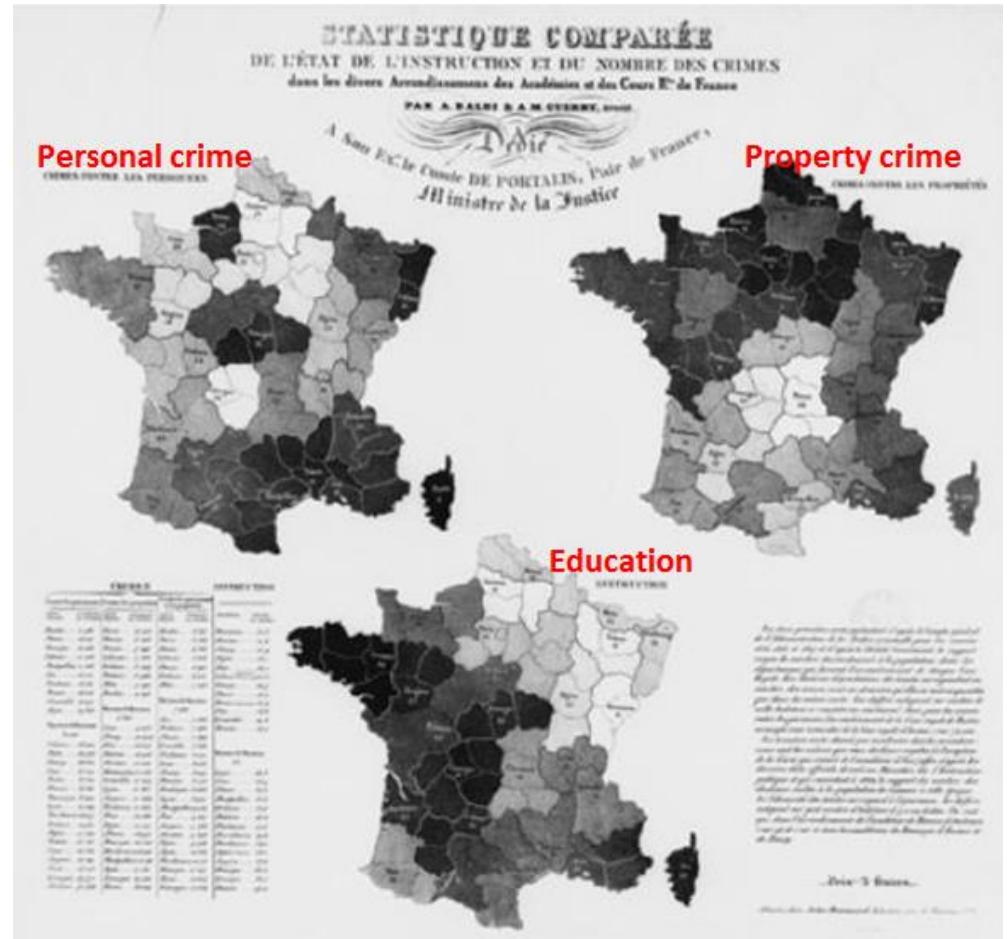


MacEachern, A. (1979). The Evolution Of Thematic Cartography / A Research Methodology and Historical Review, *The Canadian Cartographer* 16(1) June 1979, p. 17-33

Choropleth maps

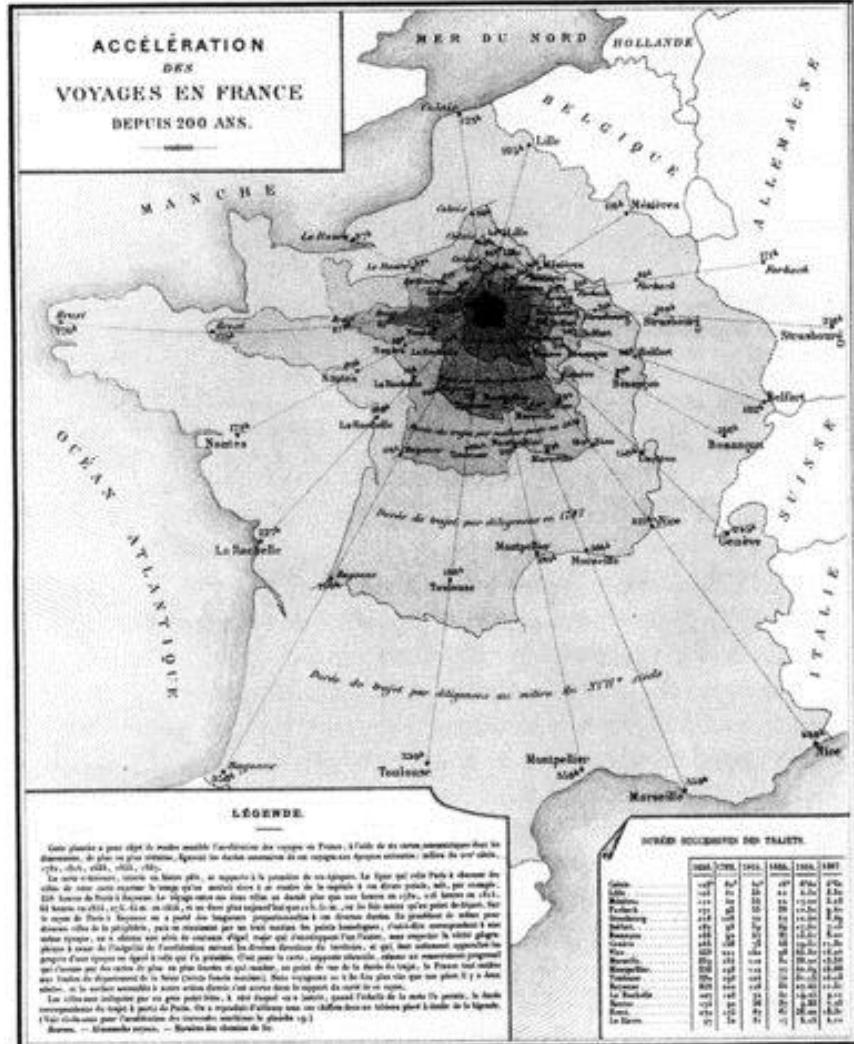
Balbi & Guerry (1829)

- First thematic maps of crime data
- First comparative maps (“small multiples”)
- Crime against persons inversely related to crime against property
- Education: *France obscure* & *France éclairée*
- N. of France highest in education & also property crime



Anamorphic maps

- *Anamorph*: Deforming a spatial size or shape to show a quantitative variable
- Émile Cheysson used this to show the decrease in travel time from Paris to anywhere in France over 200 years



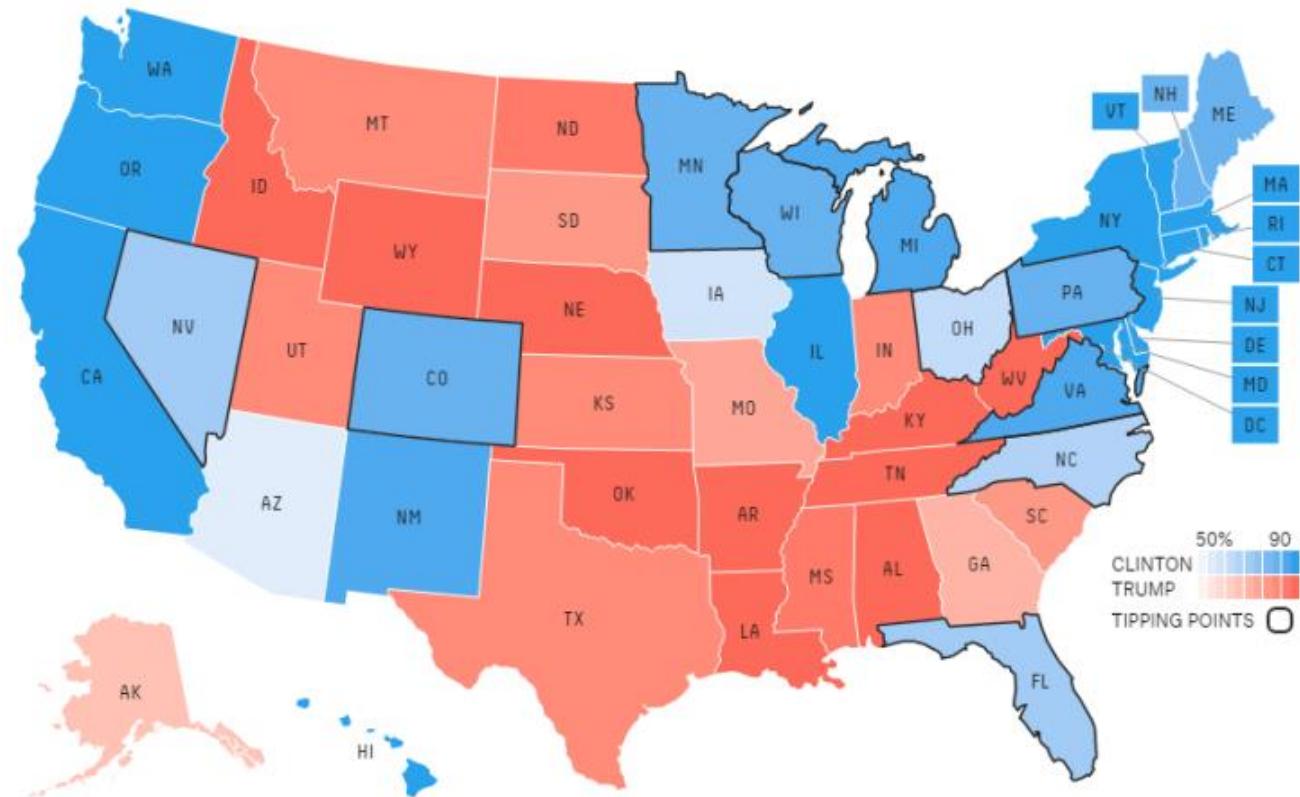
Album de Statistique Graphique, 1888, plate 8

What's wrong with choropleth maps?

Choropleth maps are misleading because size (area) of units dominates perception. This is particularly true for maps of the US & Canada. Not so for France (why?)

Montana looks
bigger than
Washington

Note use of labels
for small NE states



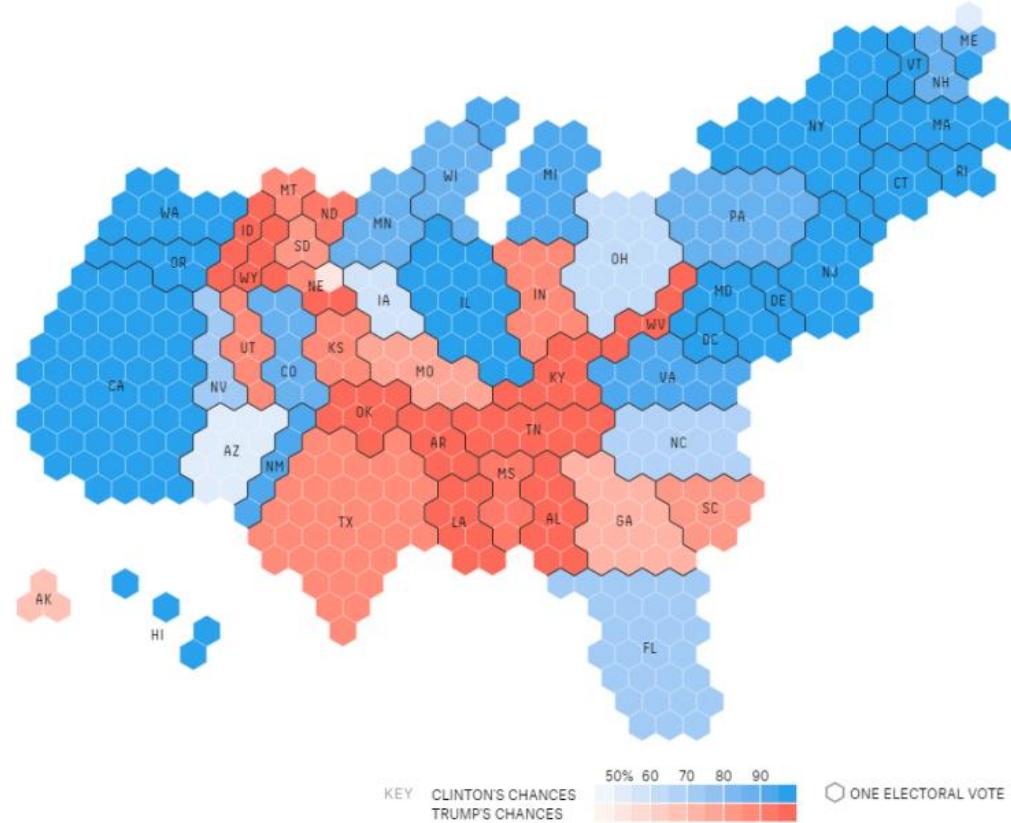
Cartogram (tilegrams)

A **tilegram** uses hexagonal tiles to make area proportional to a given variable

Here, the size of each state is made \sim number of electoral college votes

Now, it is easy to see the impact of states

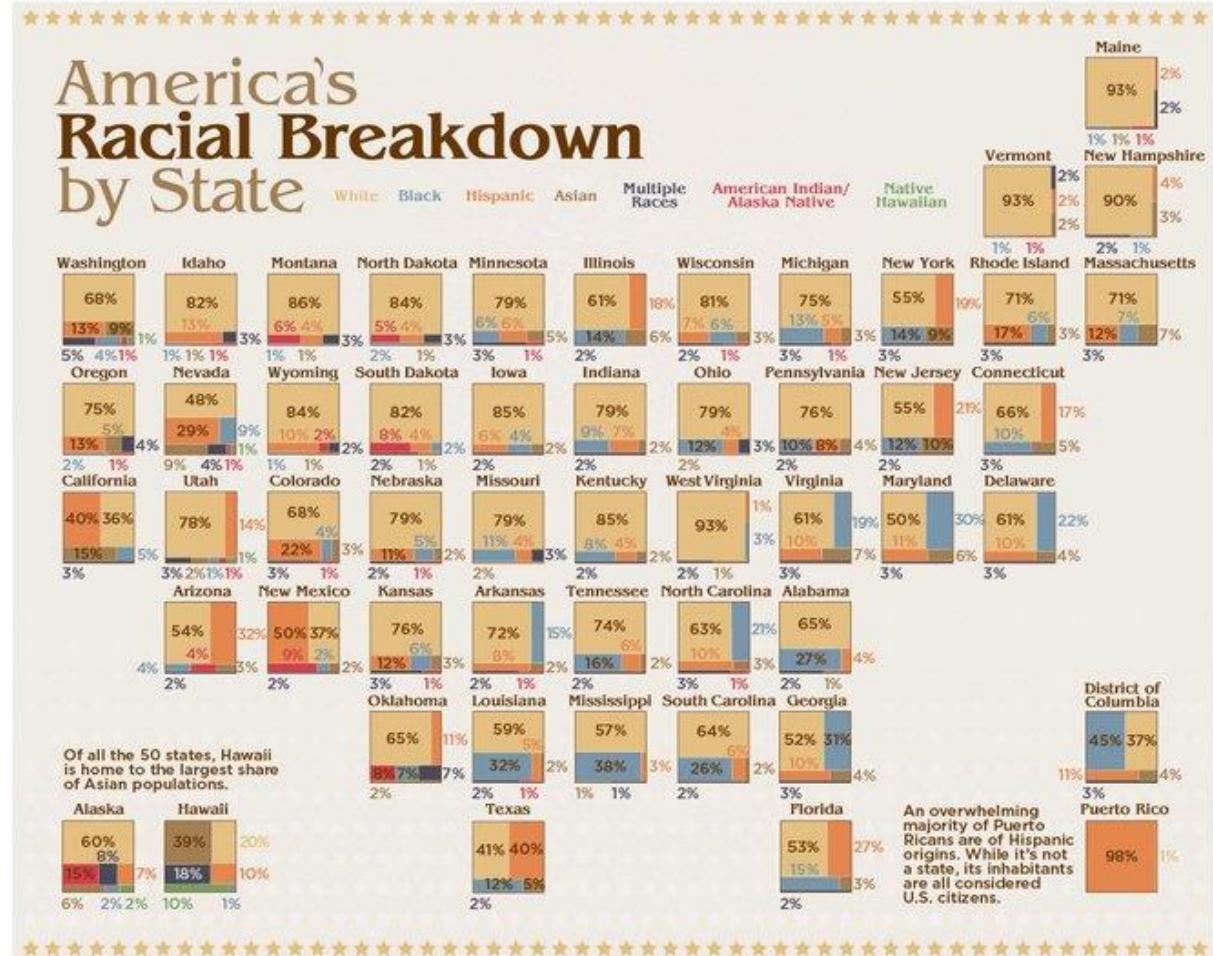
Take-away: Area doesn't vote; People do!



Mosaic cartograms

US map provides a spatial framework for showing the distribution of categorical data

Each tile is positioned as in a schematic US map



Sources: Kaiser Family Foundation, U.S. Census Bureau

[/visualcapitalist](#)

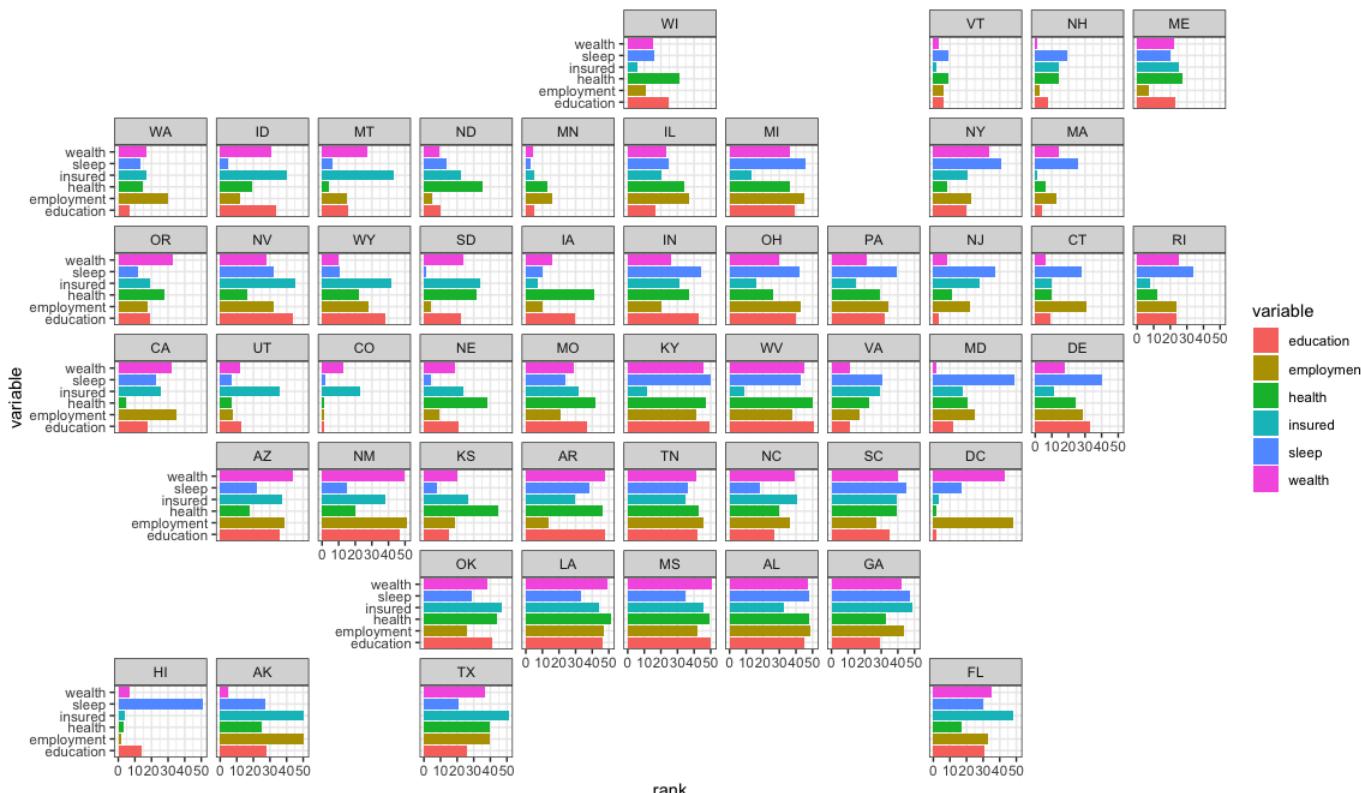
[@visualcap](#)

[visualcapitalist.com](#)

Geo-faceting

More general idea:

- A grid of (x, y) locations defines geographic units
- Each grid cell can be composed of any kind of plot conceivable with ggplot2



From: <https://cran.r-project.org/web/packages/geofacet/vignettes/geofacet.html>

Worldmapper: The world in cartograms

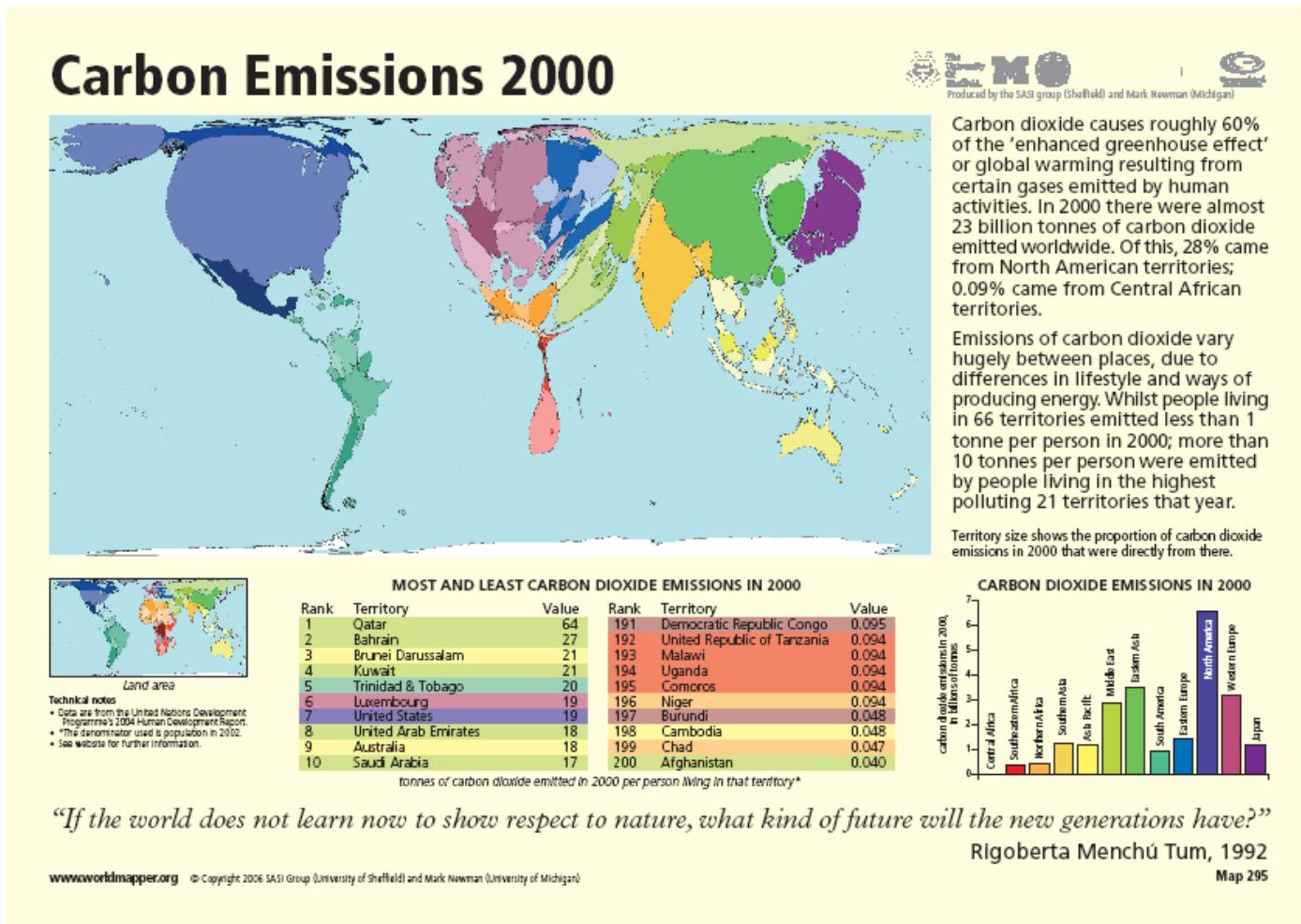
How to visualize social, economic, disease, ... data for geographic units?

worldmapper.com : **cartograms: area ~ variable of interest** (700+ maps)

The screenshot shows the homepage of Worldmapper. At the top left is the logo "WORLD MAPPER" with a globe icon. To its right is the tagline "The world as you've never seen it before". On the far right is a search bar with the placeholder "Search for a map:" and a "Go" button. Below the header is a horizontal menu with links: Home, Map Categories, Thumbnail Index, A-Z Map Index, About Worldmapper, and Help. A descriptive text block below the menu states: "Worldmapper is a collection of world maps, where territories are re-sized on each map according to the subject of interest. There are 366 maps, also available as PDF posters. Use the menu above or click on a thumbnail image below to view a map." Below this text are two sections of maps. The left section, titled "Reference maps ...", contains four thumbnails: "Total Population", "Land Area", "Labelled Map", and "Appendix A (Areas included)". The right section, titled "Newest maps ...", contains four thumbnails: "Often Preventable Deaths", "Morphing animation", "Deaths from Non-Communicable Illnesses", and "All Injury Deaths".

Worldmapper: Carbon emissions

These pages are well-designed according to data vis. Ideas: high impact graph + interpretive details & explanation



Worldmapper: Cholera deaths

Deaths from cholera in 2004. Territory size ~ proportion of worldwide deaths

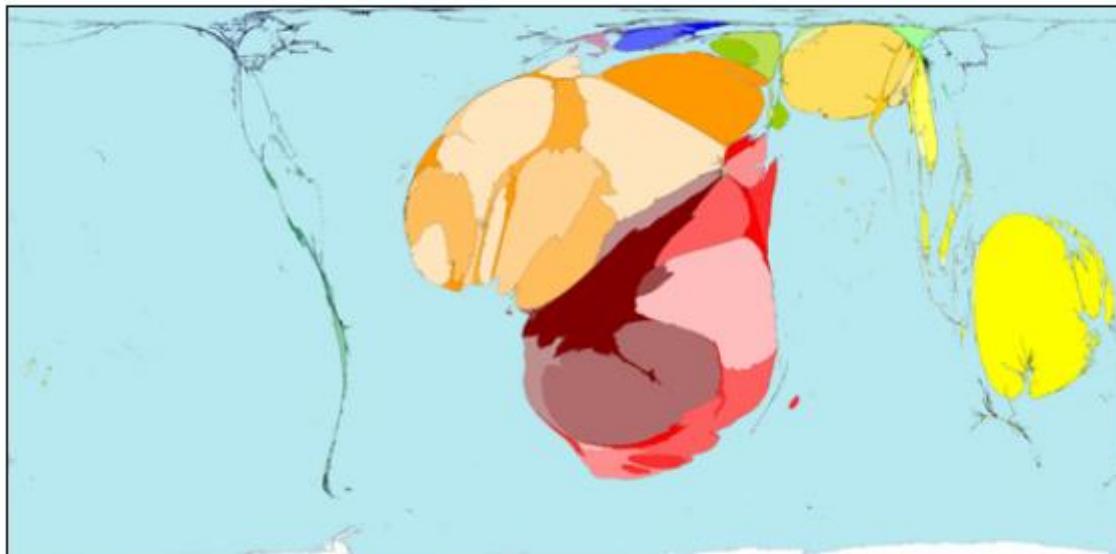
[< Previous Map](#)

Cholera Deaths

Map No. 232

[Open PDF poster](#)

[Next Map >](#)



'The cholera outbreak has continued ... water provided by the tankers is not enough and they try to boost their supply from the wells, which are not covered. The rain washes faeces and other pollutants into the wells ...' Pierre Kahozzi, 2004

Cholera deaths result from severe dehydration caused by diarrhoea. This is treatable: in 2004 the number of cholera deaths was only 2.5% of the number of cholera cases that year. Distributions of cholera cases and deaths differ due to differing availability of treatments.

In 1962, in Papua New Guinea, 36% of cholera cases, which was 464 people, died. In 2004, in the Central African Republic, 15% of cholera cases, which was 48 people, died.

In contrast, there were 73 territories where nobody died from cholera, because of good sanitation, clean water and available treatment. These territories have no area on this map.

Territory size shows the proportion of worldwide deaths from cholera that occurred there in 2004 or most recent year available.

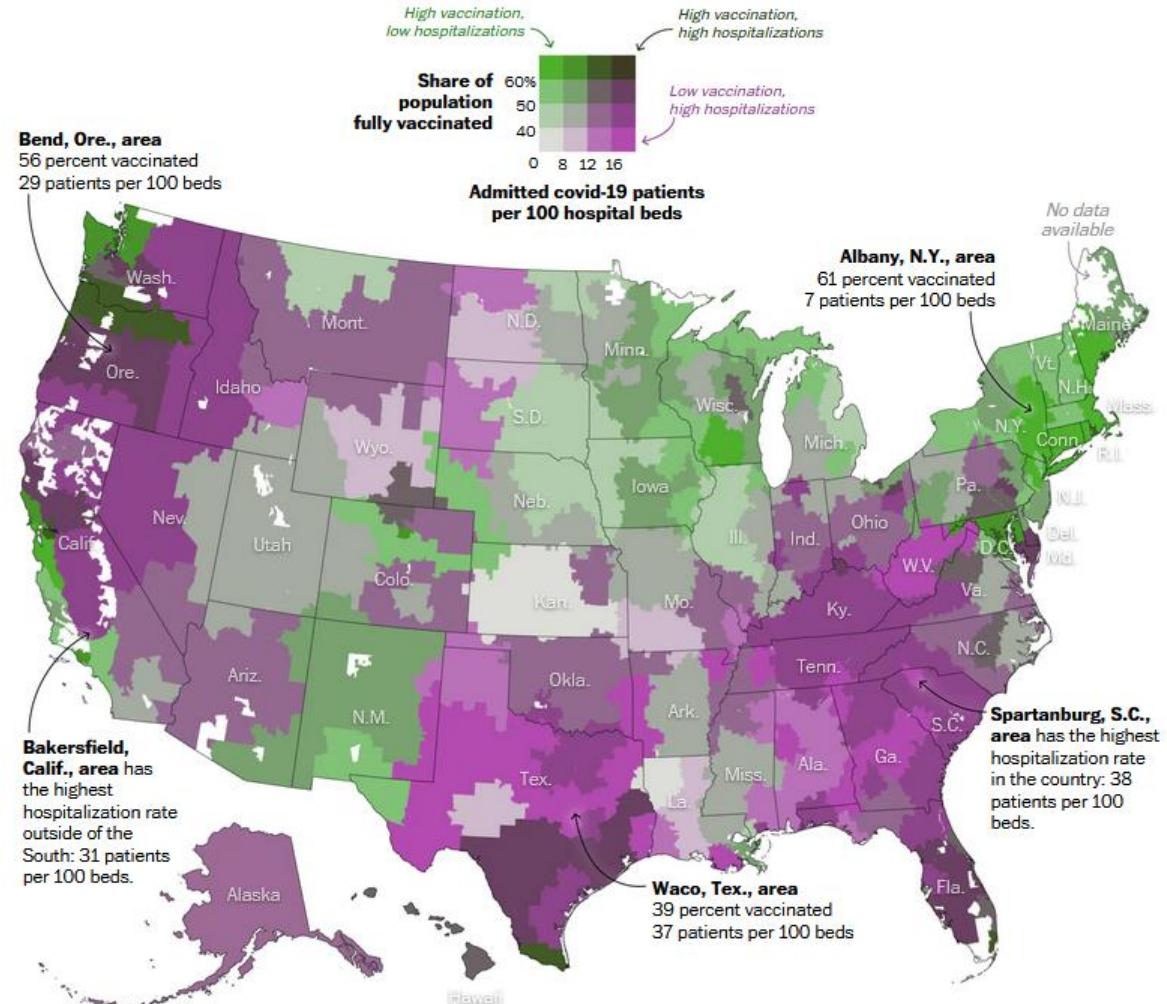
Bivariate maps

How to show relation between two variables on a map?

Bivariate maps combine two colors with degrees of saturation

Claim:
Regions with more vaccinations have fewer hospitalizations

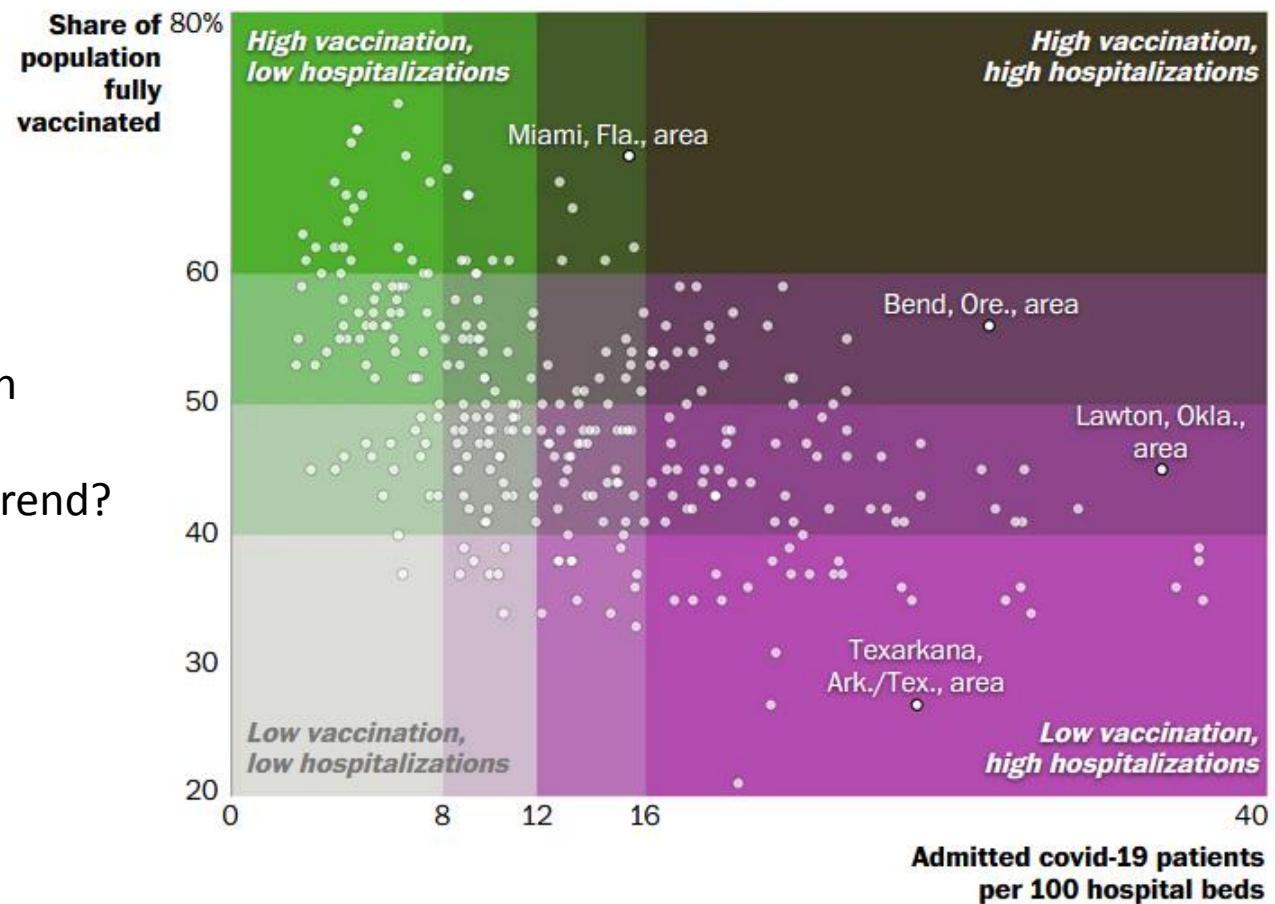
Vaccination rates and COVID hospitalizations



Vs. a scatterplot

Distribution of vaccination and hospitalization rates

Most hospital regions with high vaccination rates have lower hospitalization rates – and places with lower vaccination rates tend to see higher rates of admitted patients.

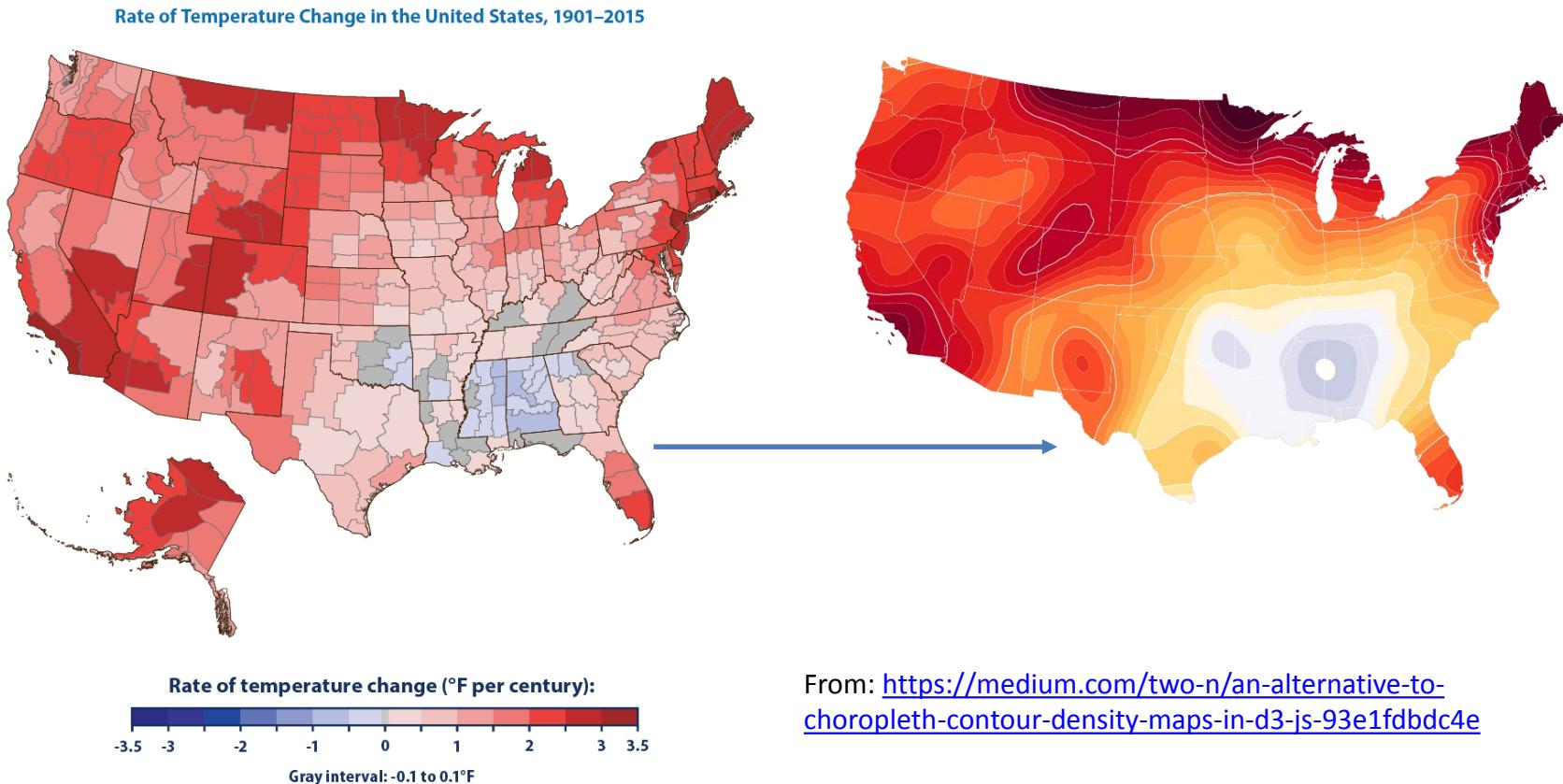


This lovely scatterplot shows:

- Yes, inverse relation
- Classing of color
- Whoa: non-linear trend?

Contour maps

Contour maps ignore region boundaries and estimate constant contours of a phenomenon over geographical space. This is a form of **geo-smoothing**.



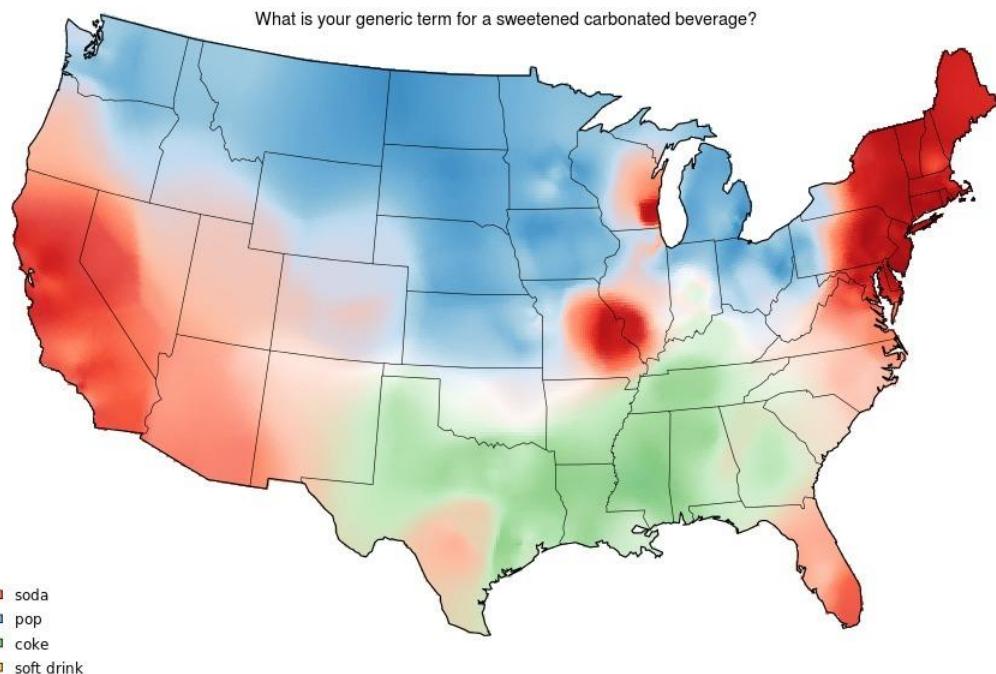
Spatial visualization: Analysis + maps

Linguistics: Food dialect maps— visualizing how people speak

In the *Cambridge Online Survey of World Englishes*, Bert Vaux and Marius L. Jøhndal surveyed 11,500 people to study the ways people use English words.

NC State Univ. student Joshua Katz turned the US data into shaded **kernel density maps**.

soda vs. pop?



Take the survey: http://www.tekstlab.uio.no/cambridge_survey

Programming in R: <http://blog.revolutionanalytics.com/2013/06/r-and-language.html>

Spatial visualization: Analysis + maps

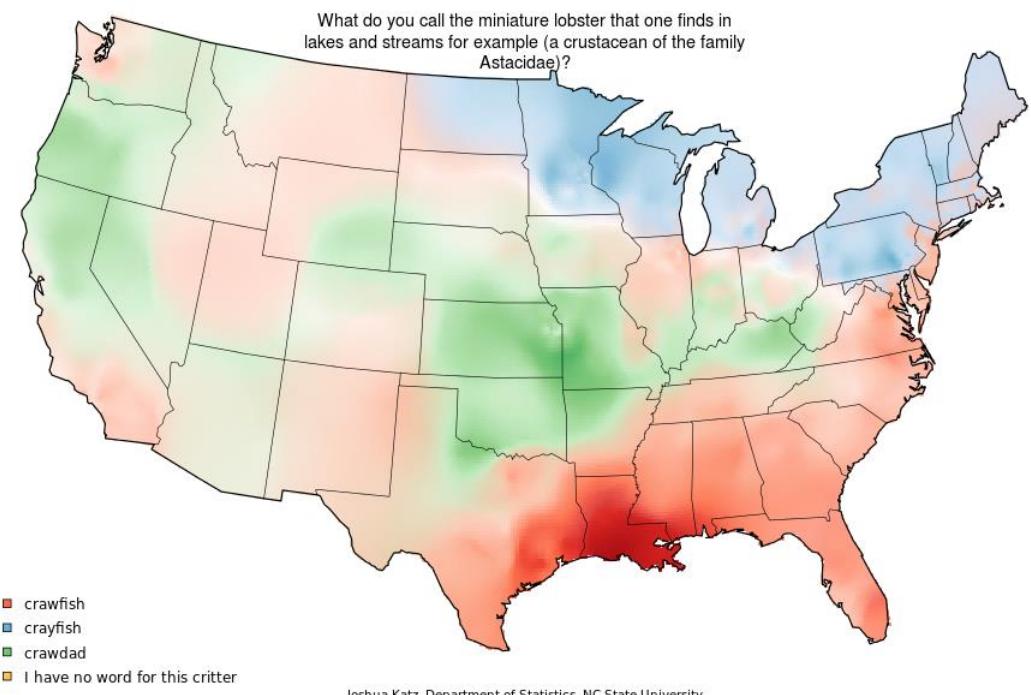
Linguistics: Food dialect maps— visualizing how people speak

crawfish, crawfish, crawdad?

A k -nearest neighbor **kernel density estimate** over (x,y) locations gives a smoothed & interpretable display of the choice probabilities.

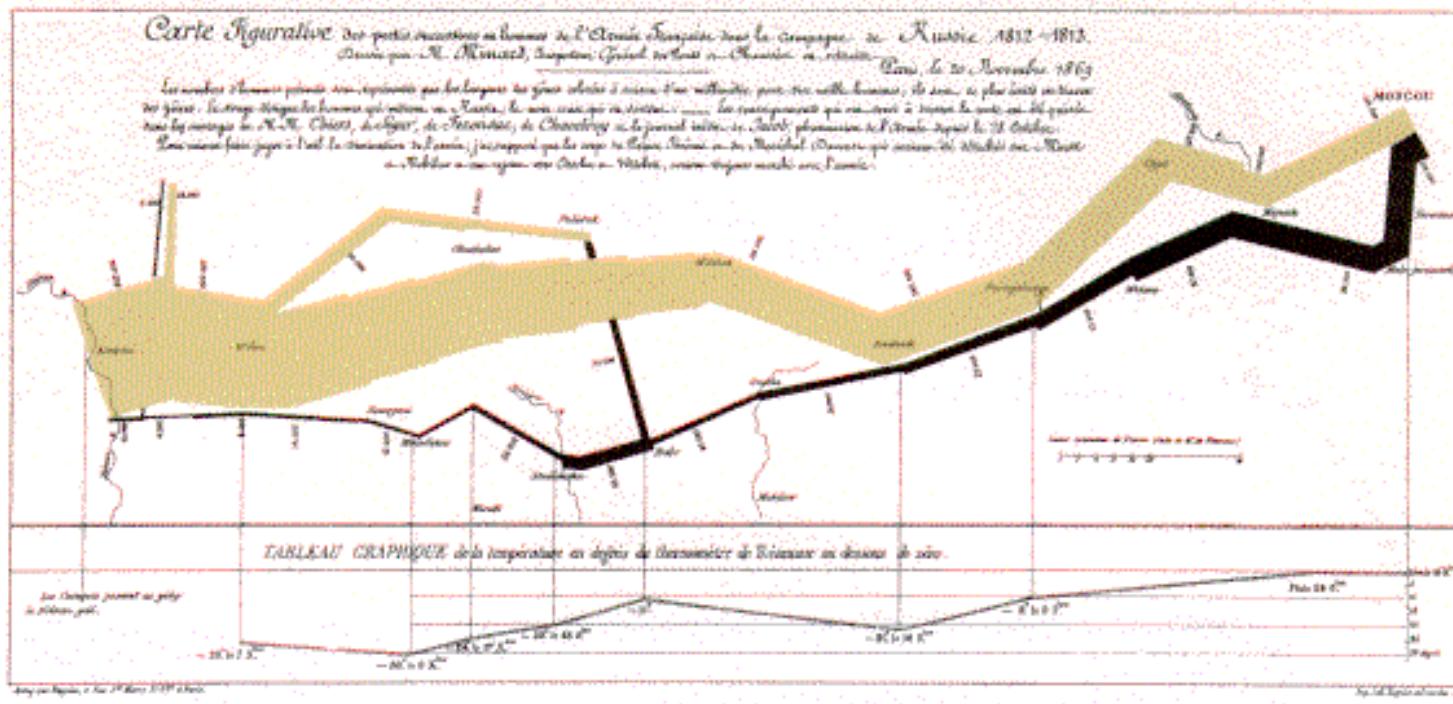
Regional differences are quite apparent.

The use of **color** combines discrete categories with intensity to give a meaningful display



Flow maps

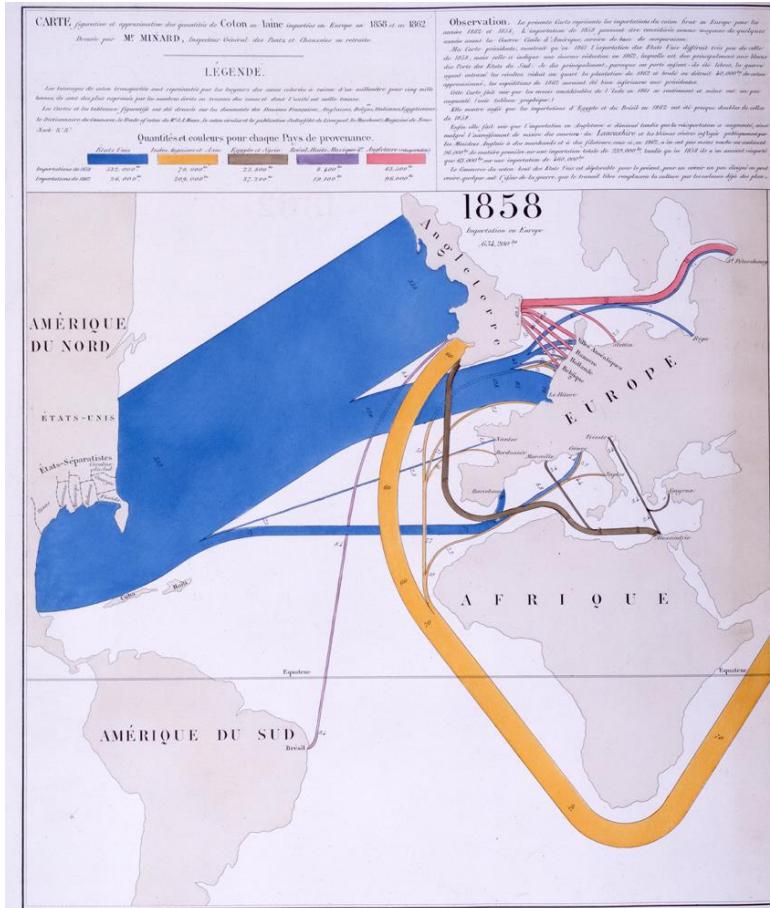
Flow maps show **movement** or **change** in a geographic framework
The master work is this image by Charles-Joseph Minard (1869)



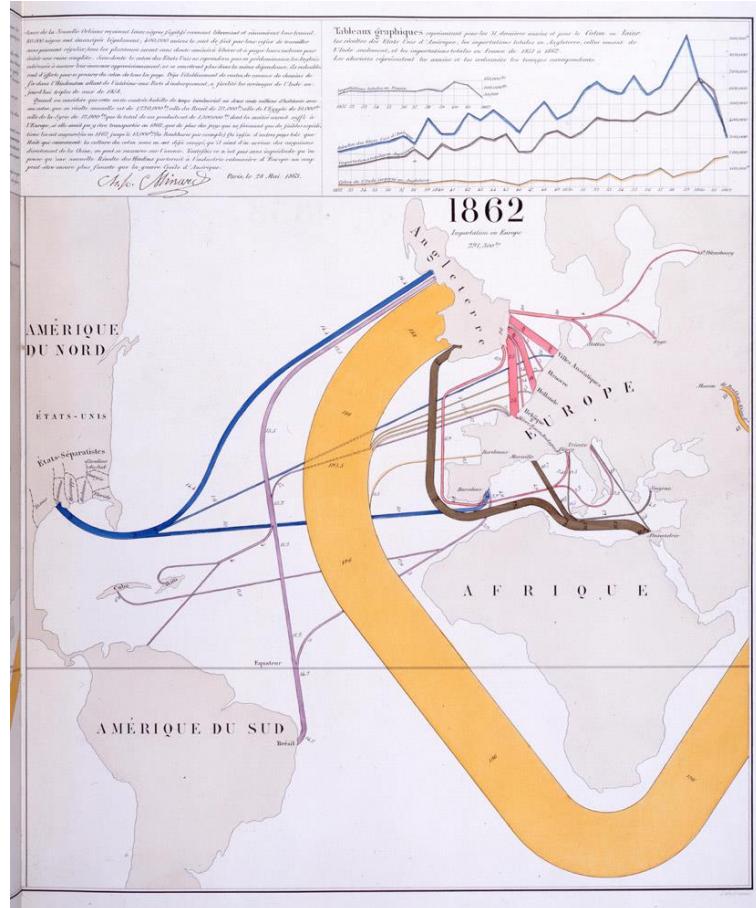
- Marey (1878): “defies the pen of the historian in its brutal eloquence”
- Tufte (1983): “the best statistical graphic ever produced”

Effect of US civil war on cotton trade

Before



After



Note the deformation of the map to accommodate the data

The Great Migration

In a graphic tribute to C.-J. Minard and W. E. B. Du Bois, Raymond Andrews & Howard Wainer tell the story of the migration of blacks from the southern US after freedom from slavery.

Figurative Maps showing the flows of Nonwhite Migrants in America 1880 ~ 1940. design inspired by M. Charles Joseph Minard ~ and ~ W. E. B. Du Bois.

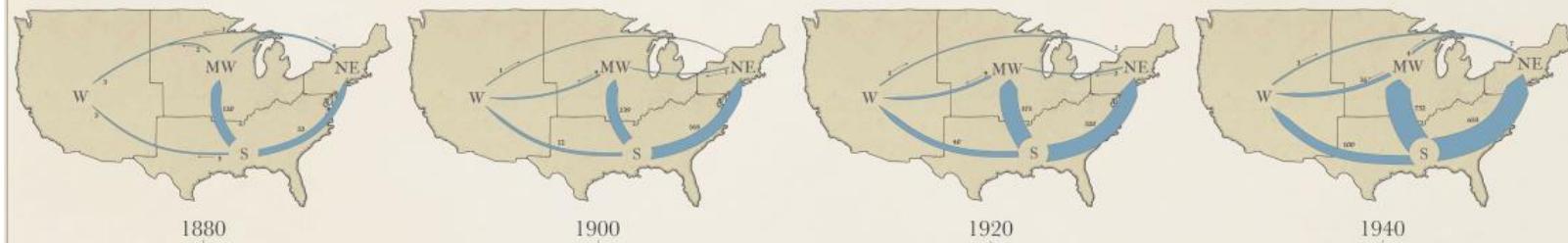
Legend.

The number of net migrants is labeled in thousands and indicated by the width of each flow - proportional to the square root of the raw number.

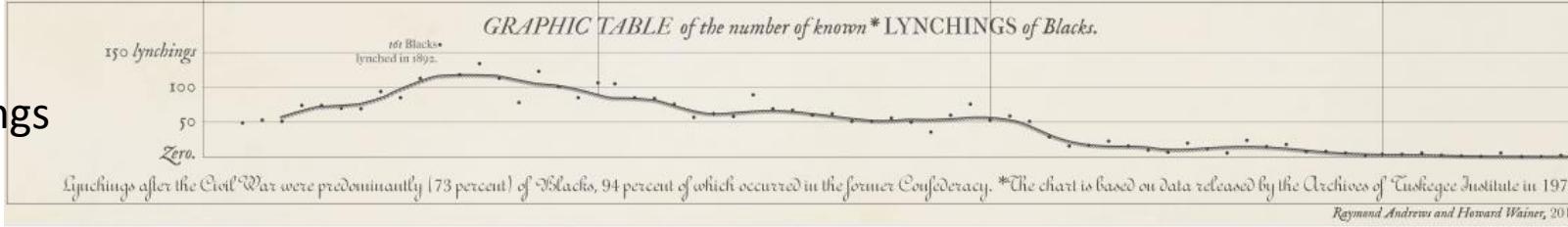
October, 2017.

Explanation & Observations: Internal migrants are displayed on a journey from their region of birth to their place of residence as recorded in the decennial U.S. Census. Migrant flows are shown between major regions which include the Northeast (NE), the South (S), the Midwest (MW), and the West (W) - geographic areas constructed by the authors from the original nine Census regions. The Great Migration is seen as Blacks leaving the South, fleeing the former Confederacy for economic opportunities in the Northeast, Midwest, and (later) the West. Maps are paired with known lynchings of Blacks in order to add a causal dimension, using it as a proxy measure of the broader social environment.

Migration



Lynchings



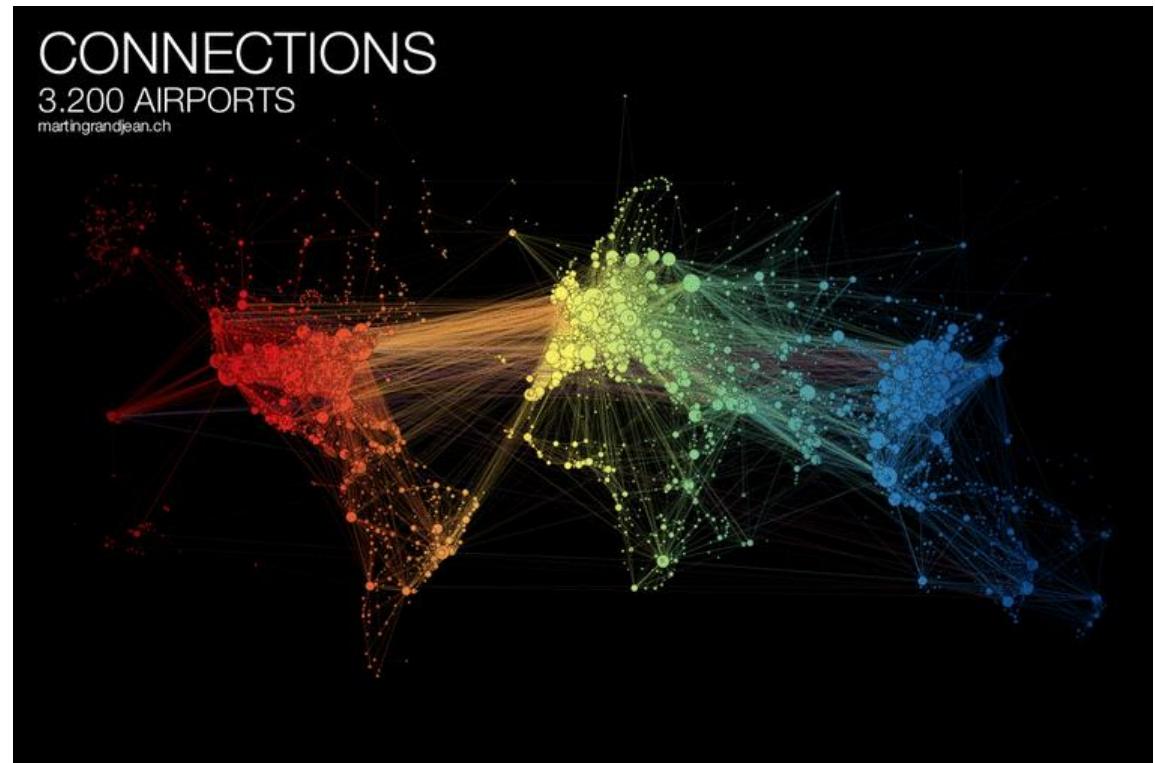
Network visualization



Once the domain of mathematicians & computer scientists, graph theory and network visualization turn out to have surprising & interesting applications.

Animated demo by Martin Granjean showing transport of passengers from/to world airports.

It illustrates the difference between geography & **force-directed layout** to focus on volume & connections



From: <http://www.martingrandjean.ch/connected-world-air-traffic-network/>
See more: <https://flowingdata.com/2016/05/31/air-transportation-network/>

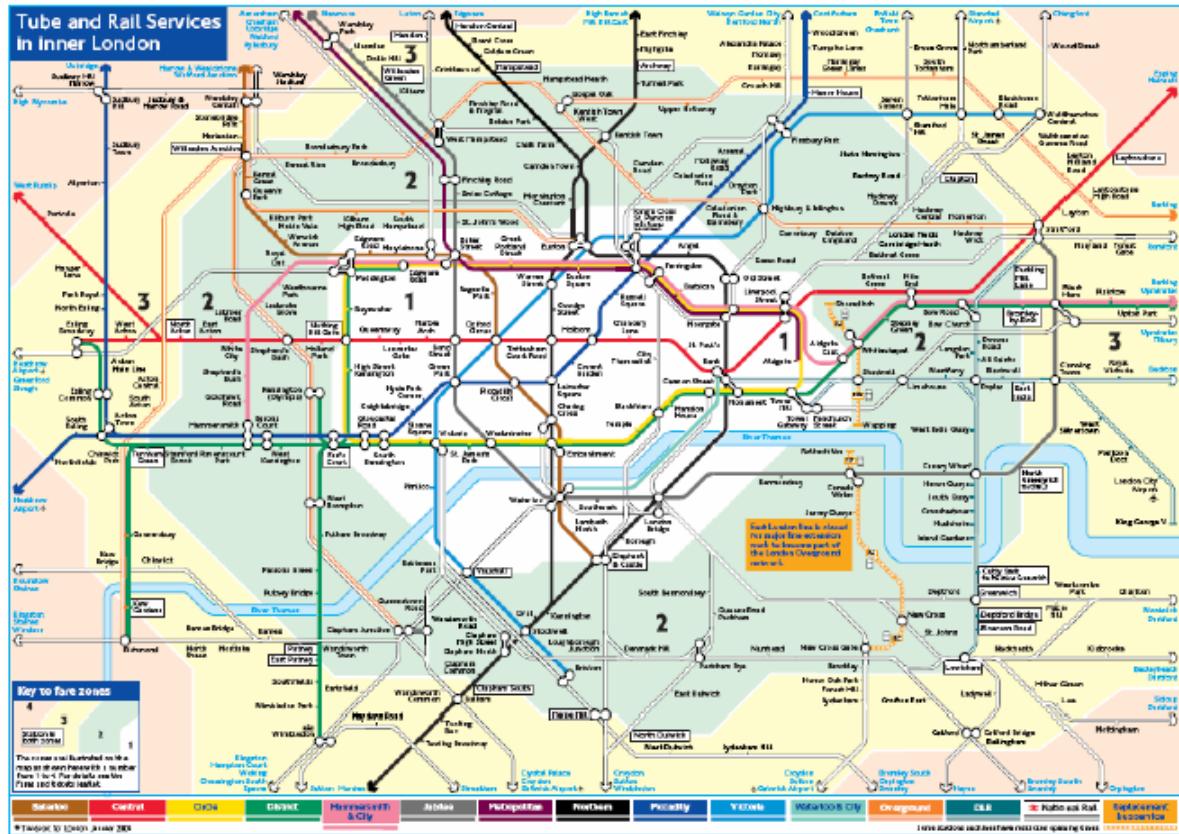
Network visualization: Transport maps

How do I get from Chigwell to Charing Cross?
How much will it cost?

This route map shows the connections and fare zones

The first one was designed by Henry Beck in 1931.

The modern version is zoomable and available on your phone.



See: <https://tfl.gov.uk/maps/track>

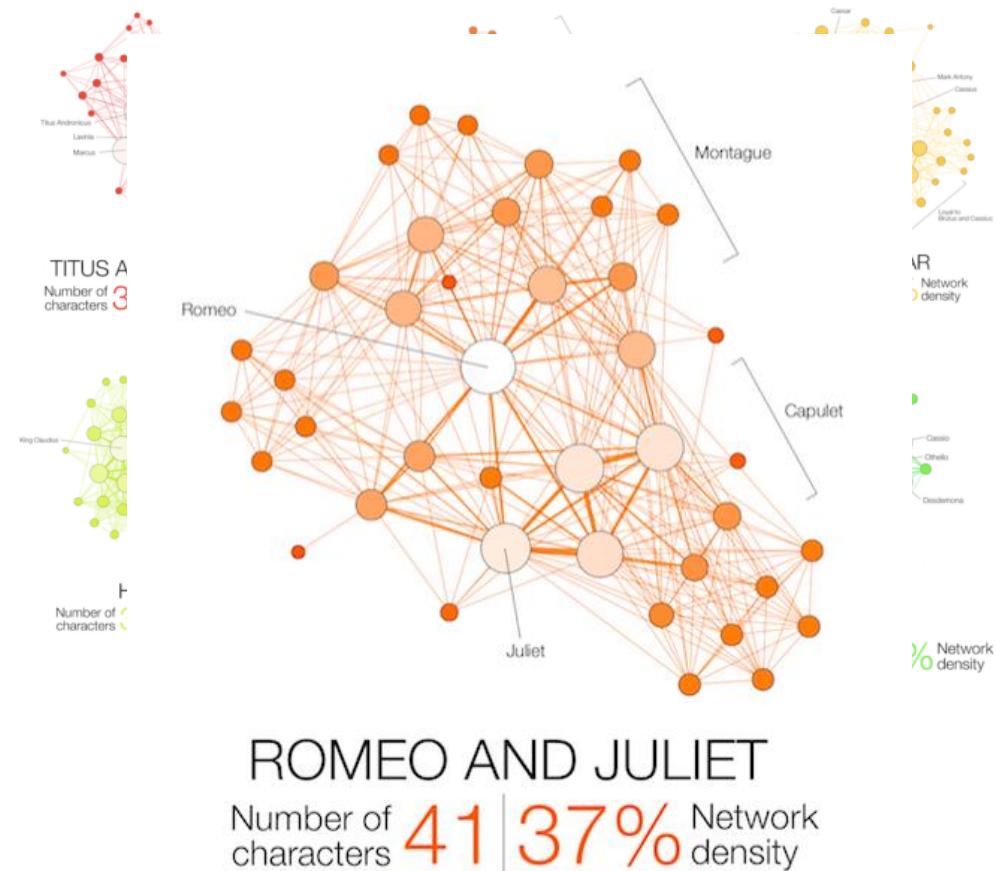
Network visualization: Shakespeare tragedies

A new form of literary criticism?

Martin Grandjean looked at the structure of Shakespeare tragedies through character interactions.

Each circle (node) represents a character, and an edge represents two characters who appeared in the same scene.

The structural characteristics of the graphs have meaningful interpretations.



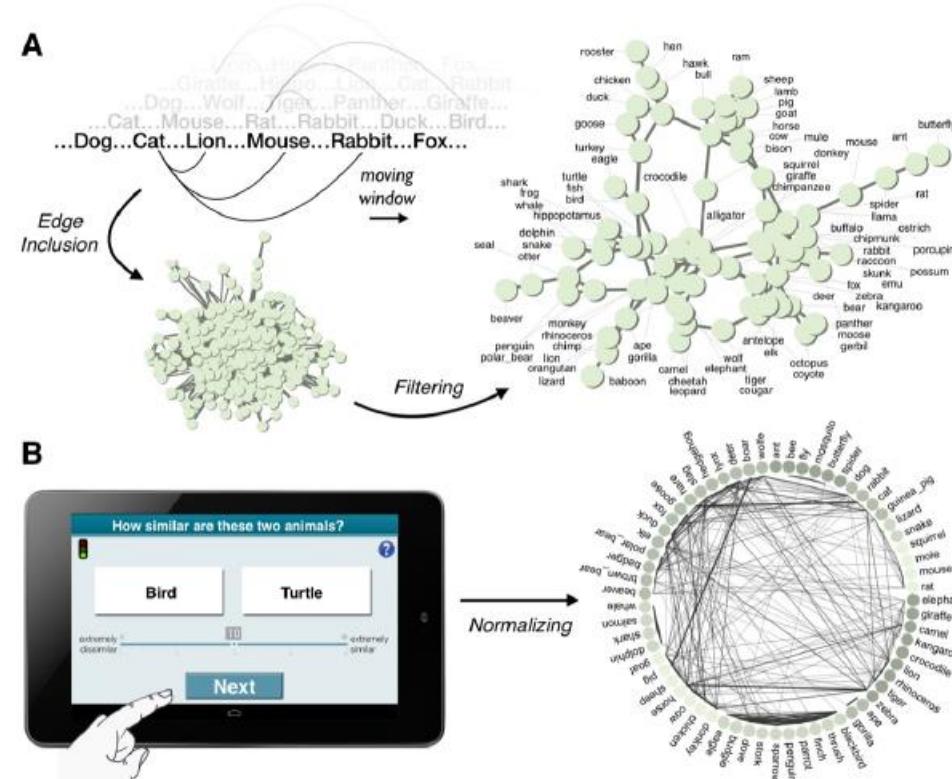
From: <https://flowingdata.com/2015/12/30/shakespeare-tragedies-as-network-graphs/>

Semantic memory: Cognitive structure

Various tasks can be used to assess the relations among words/concepts in our semantic memory

The data can be used to calculate measures of **similarity**, and be shown in network or other diagrams

Verbal fluency task: Say/write all the names of [animals, countries, ...] you can in 1 minute.



Similarity ratings: For each pair, indicate how similar they are

From: Wulff et al. (2018), Structural differences in the semantic networks of younger and older adults

Semantic memory: Cognitive structure

Do younger and older adults differ on measures calculated from their network diagrams?

$\langle k \rangle$: Average “degree” # of connections

C : average local clustering

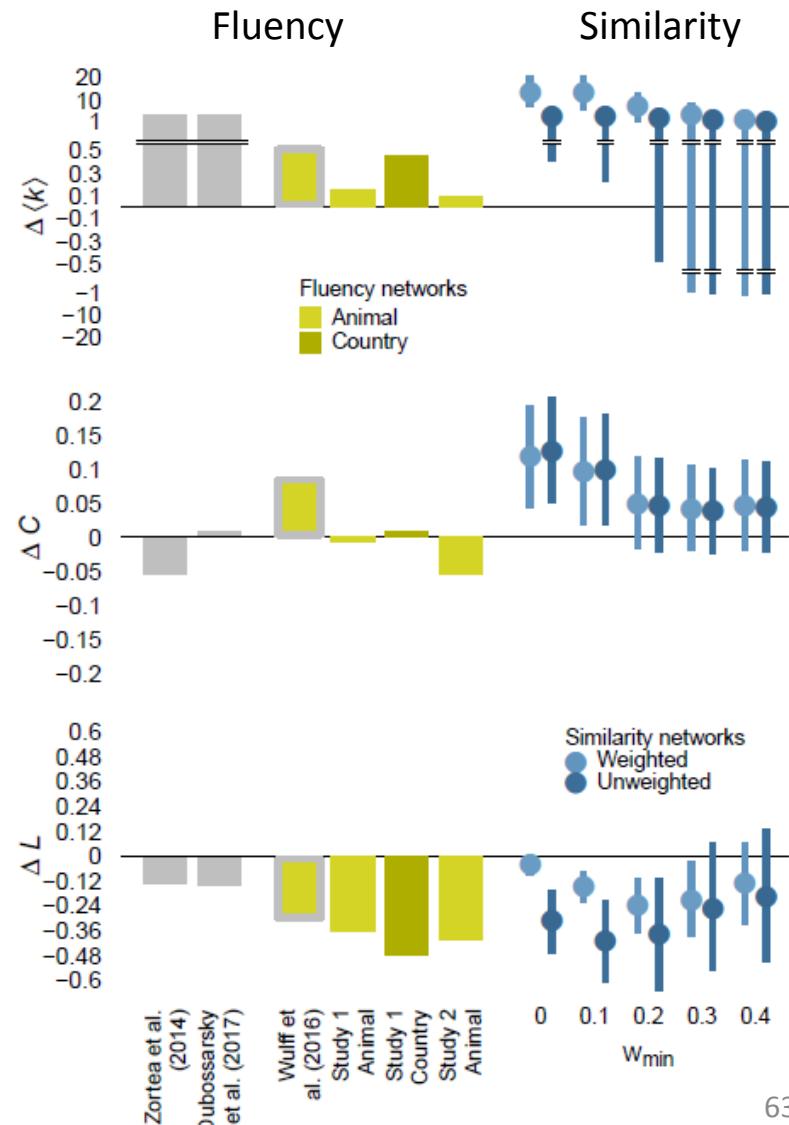
L : average path length in network

$\Delta()$: young – old difference

IMHO, this graph tries to do too much.

The fluency data is most important to their argument.

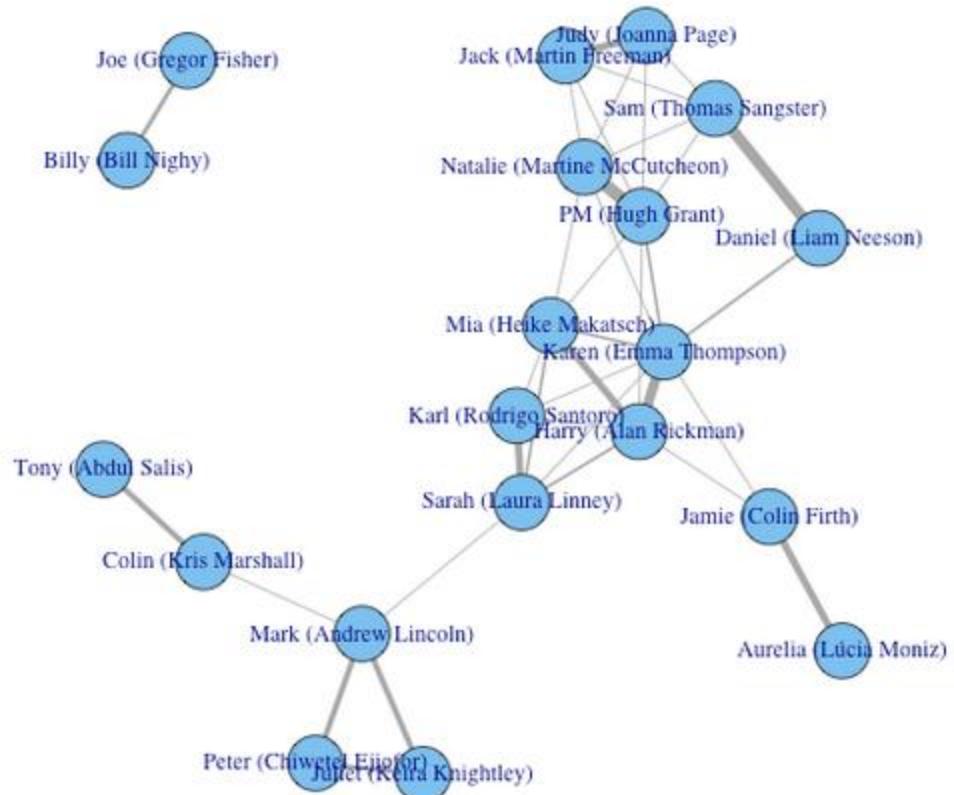
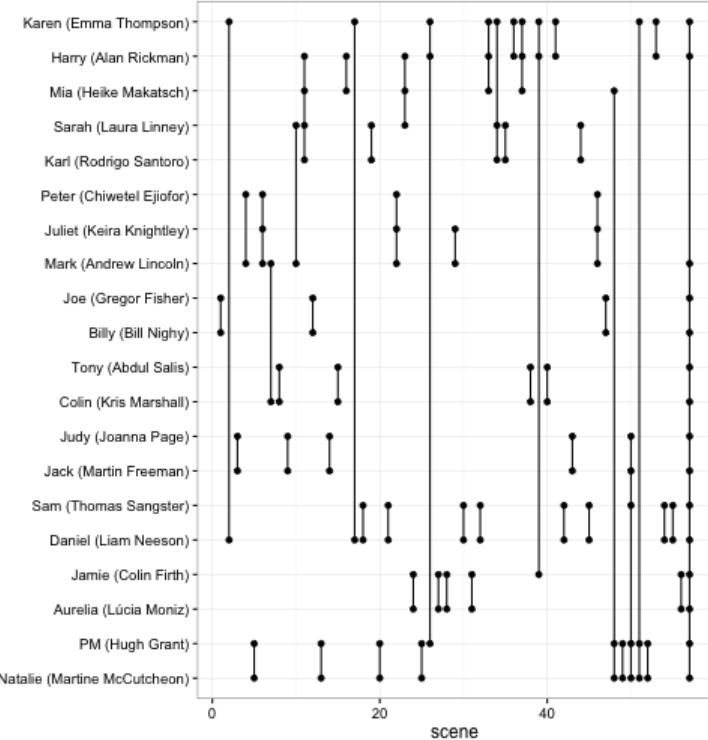
ΔL & $\Delta \langle k \rangle$ show consistent differences between young & old



Love, Actually: Interactive app

Interactions among characters in *Love, Actually*

Data:



Interactive Shiny app: <https://dgrtwo.shinyapps.io/love-actually-network/>

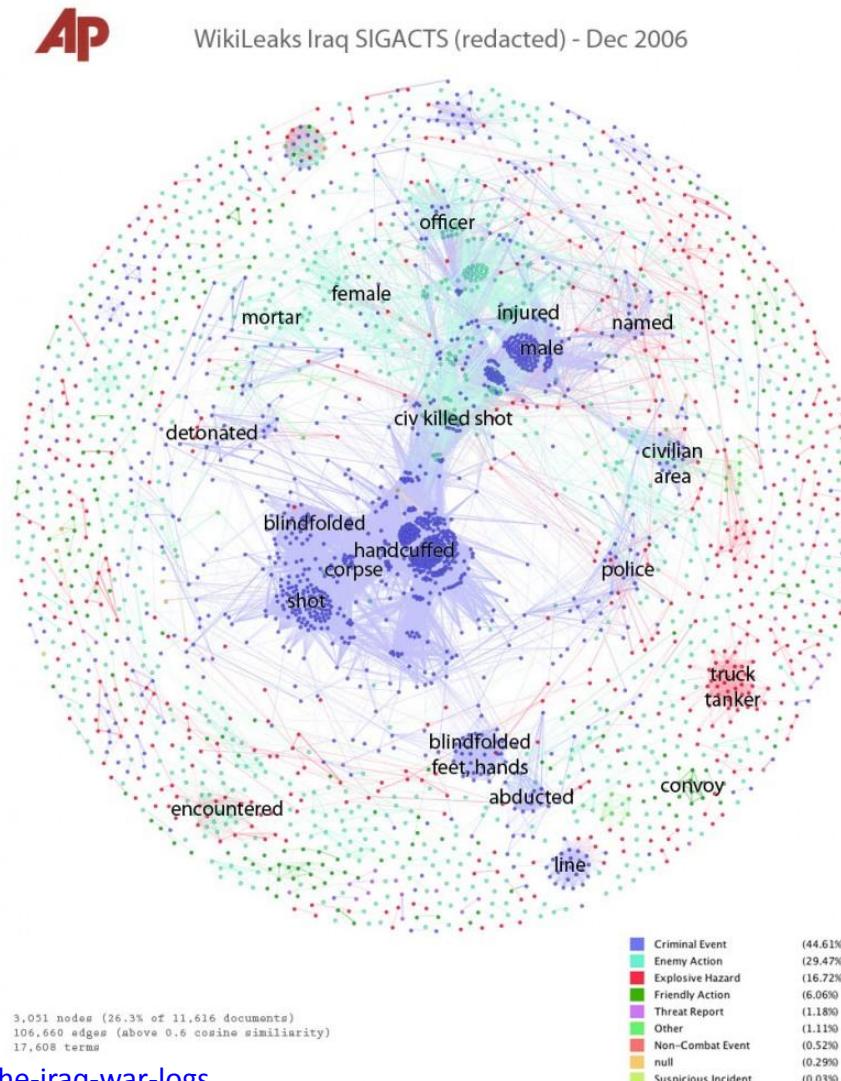
WikiLeaks Iraq war logs

Johnathan Stray & Julian Burgess analyzed > 11,000 documents for SIGACT (“significant action”) reports from the 2006 Iraqi civil war made available by WikiLeaks.

Each report is a dot. Each dot is labelled by the three most “characteristic” words in that report.

Documents that are “similar” have edges drawn between them, width ~ similarity

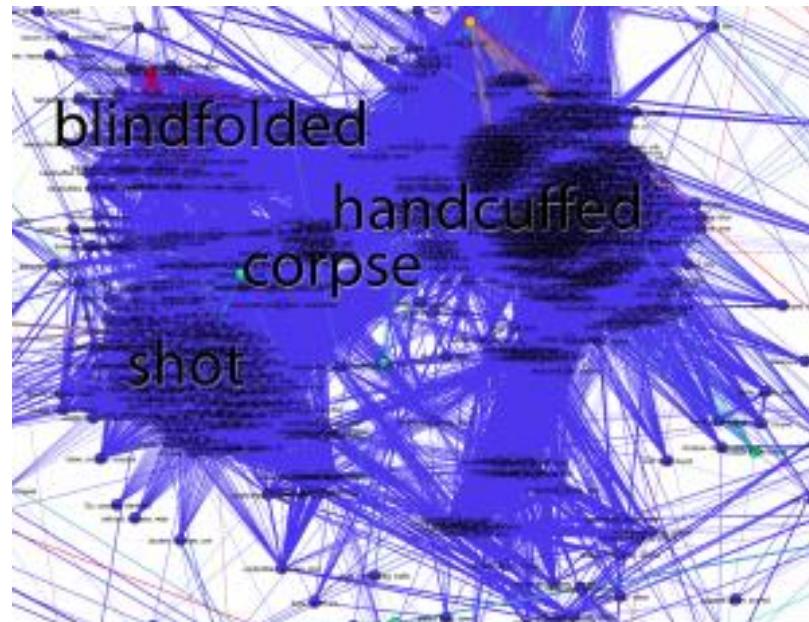
The graph-drawing algorithm placed similar nodes together



WikiLeaks Iraq war logs

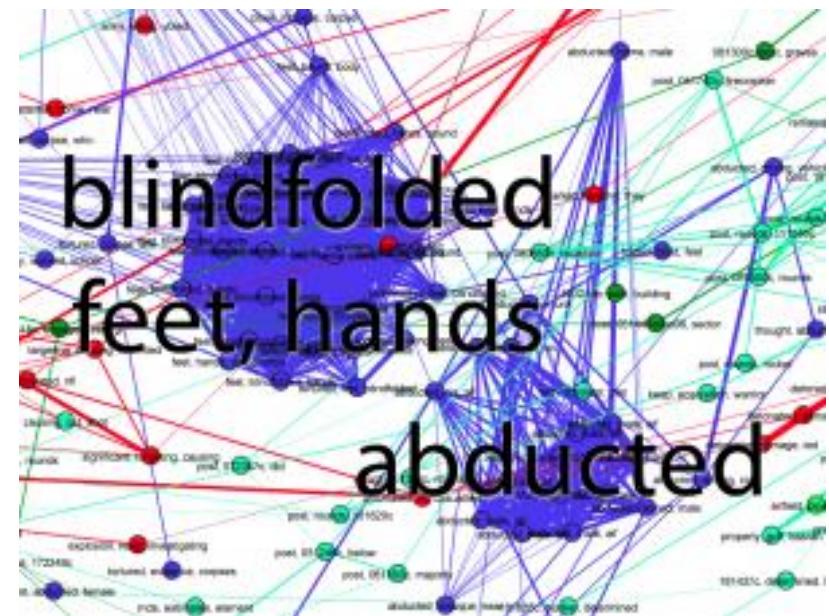
Certain themes became clear, and could be studied in rich detail
The underlying methods use “term frequency–inverse document frequency”
measures of **text-mining**.

Murder cluster. All contain the word
“corpse”



<http://jonathanstray.com/wp-content/uploads/2010/12/Murders.png>

Torture-abduction cluster



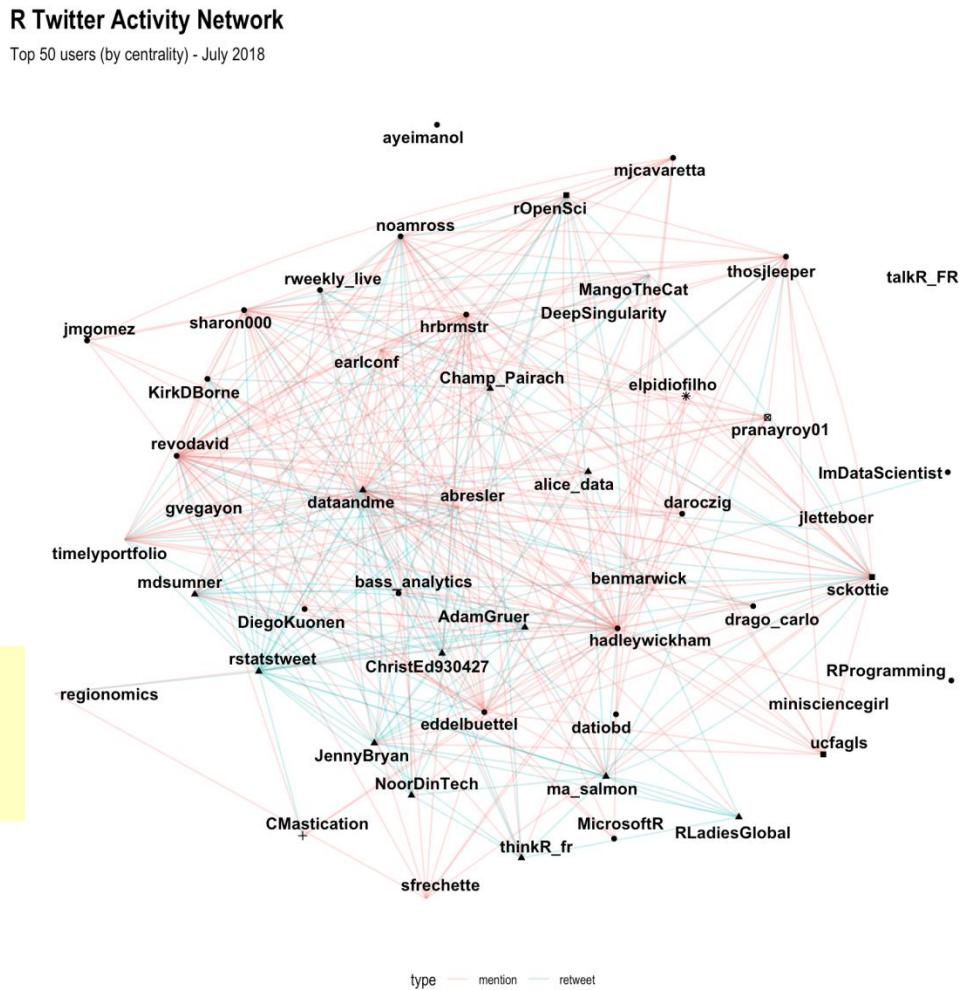
<http://jonathanstray.com/wp-content/uploads/2010/12/Torture-abduction.png>

Twitter network of R users

Perry Stephenson explores the connections among the top 50 R users on Twitter

The rtweet package provides access to Twitter info

```
library(rtweet)
followers <-
get_followers("datavisFriendly")
```



From: <https://perrystephenson.me/2018/09/29/the-r-twitter-network/>

Twitter circles

Who do I most often interact with?

Three rings to show my twitter world

One ring to rule them all:
@datavisFriendly

Other rings: #datavis,
#maps, #rstats, #psy6135

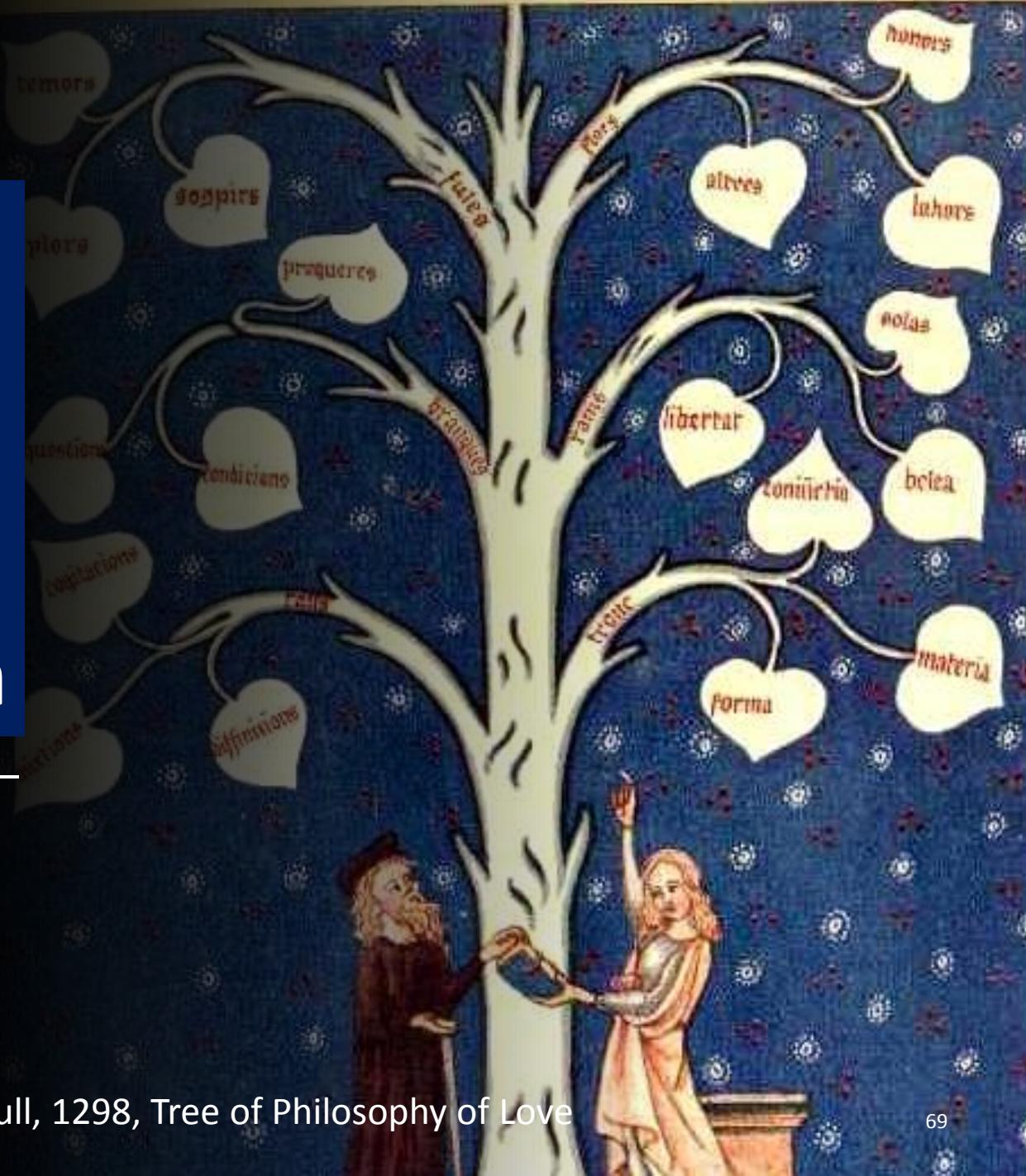


Tree-based Visualization

Branching patterns

History as a tree

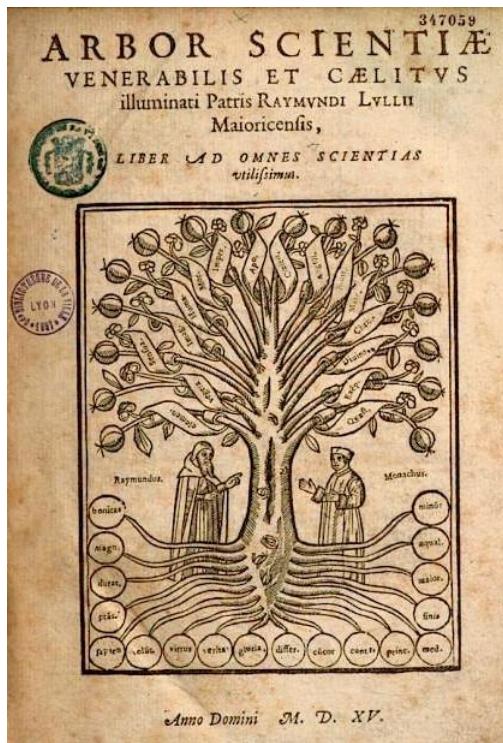
Treemaps



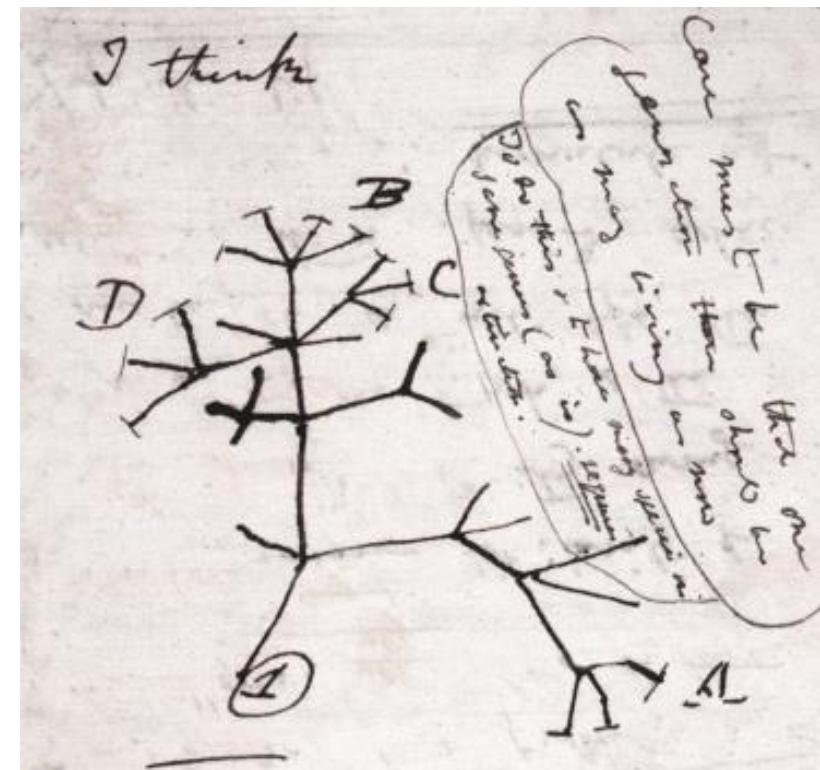
Ramon Llull, 1298, Tree of Philosophy of Love

Tree diagrams

Trees are natural, organic visual metaphors for branching processes and space-filling designs.



Ramon Llull's tree of science, showing roots and branches of knowledge



Charles Darwin's first visual sketch of the evolution of species

History as a Tree: *Geschichtsbaum* Europa (2003)

- The entire history of Europe in one diagram
- space-filling design: resolution \sim time²
- natural metaphors for roots, branches

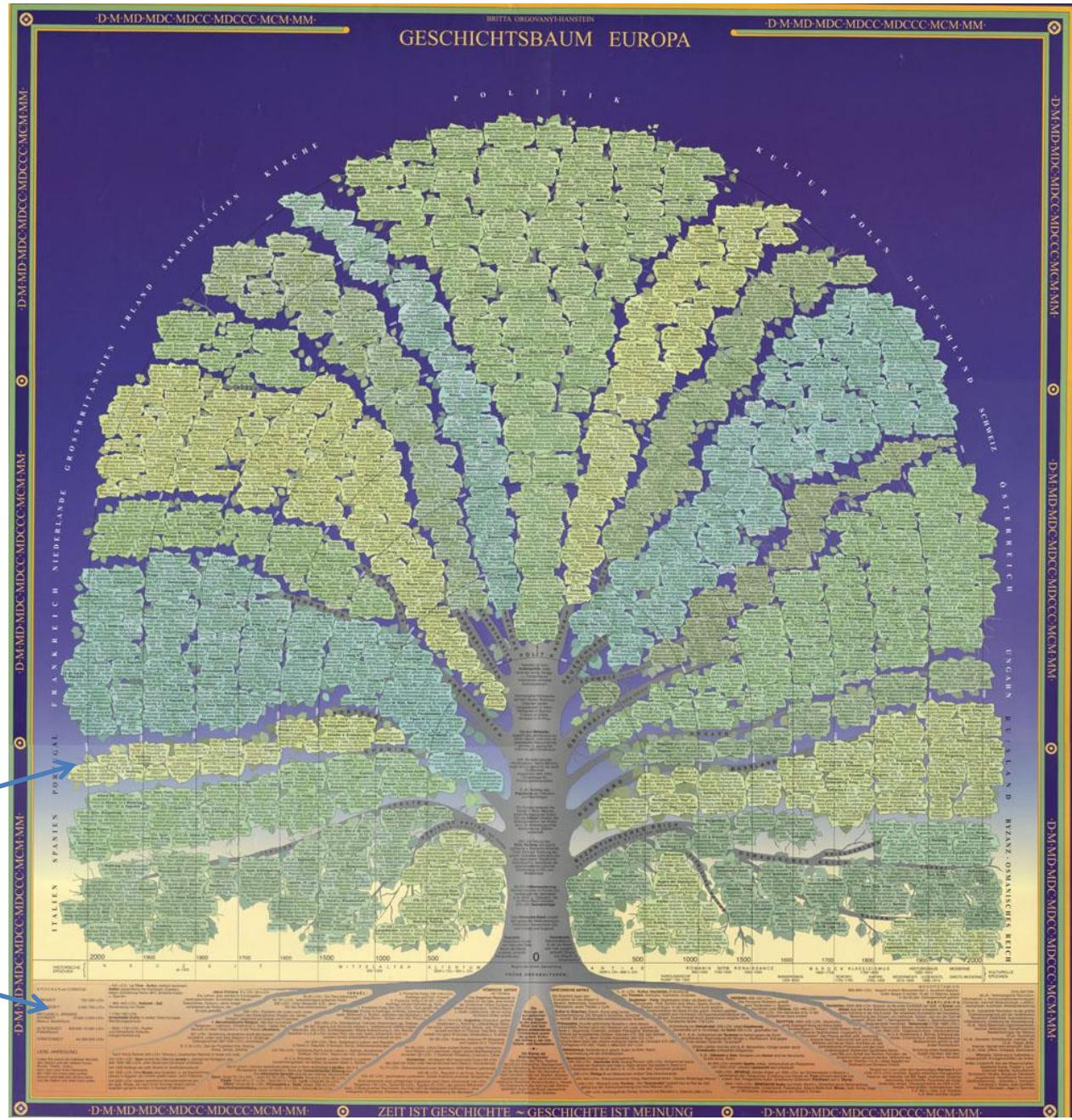
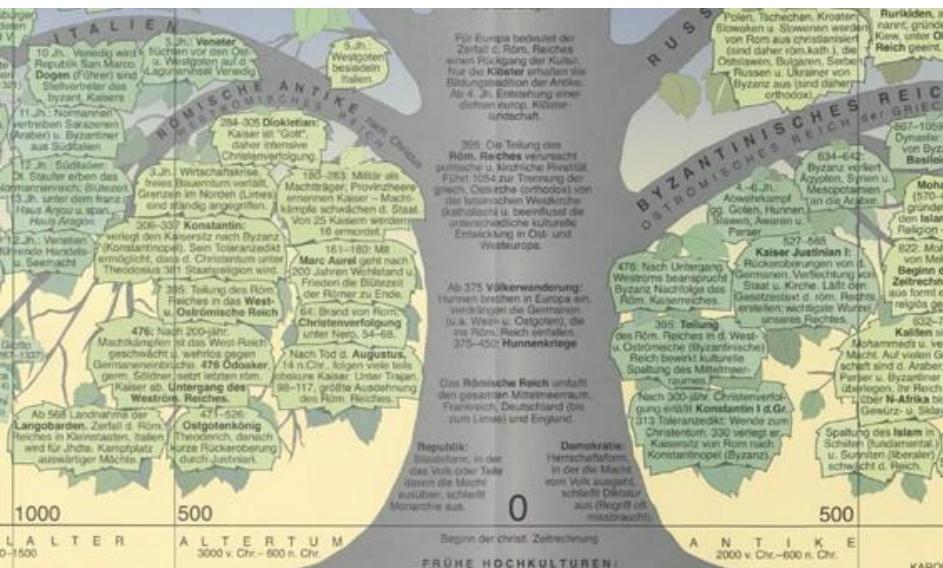
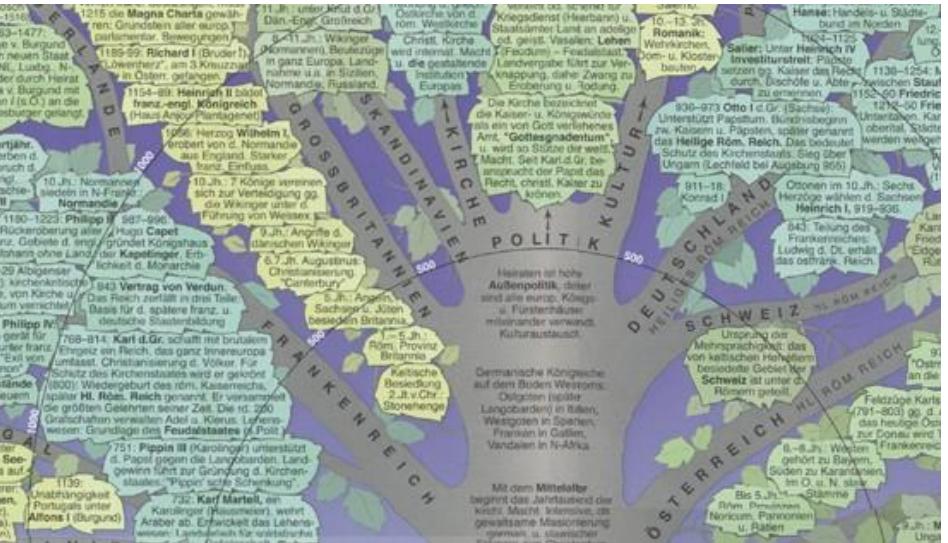


Image: <http://euclid.psych.yorku.ca/SCS/Gallery/images/timelines/geschicht1000.jpg>

History as a Tree

- Branches for countries & domains of thought
- Leaves for all the details



- linear horizontal scale → area ~ time²

Treemaps

Treemaps display hierarchical data as a set of nested rectangles.
Each node (leaf) has an area \sim size (CO_2)

Who emits the most CO_2 ?

Global carbon dioxide (CO_2) emissions were 36.2 billion tonnes in 2017.

Asia

19 billion tonnes CO_2
53% global emissions

China

9.8 billion tonnes CO_2
27% global emissions

Japan
1.2 billion tonnes
3.3%

Saudi Arabia
635 million tonnes
1.8%

Thailand
931M tonnes
0.9%

UAE
232M tonnes
0.6%

Pakistan
199M tonnes
0.5%

Vietnam
192M tonnes
0.55%

Canada
573M tonnes
1.6%

Mexico
490M tonnes
1.4%

Russia
1.7 billion tonnes
4.7%

Turkey
443M tonnes
1.2%

Ukraine
212M tonnes
0.6%

Belarus
(61M t)

Georgia
(22M t)

Nicaragua
(12M t)

South Korea
616 million tonnes
1.7%

Taiwan
272M tonnes
0.8%

Philippines
128M tonnes
0.35%

Qatar
130M tonnes
0.4%

Bangladesh
80M tonnes
0.2%

South Africa
456M tonnes
1.3%

Nigeria
101M tonnes
0.3%

Brazil
476M tonnes
1.3%

Australia
414M t
1.1%

International aviation
& shipping
1.15 billion tonnes
3.2%

Argentina
204M tonnes (0.6%)

Morocco
100M tonnes
(0.3%)

Colombia
80M tonnes
(0.2%)

Iran
672 million tonnes
1.9%

Indonesia
489 million tonnes
1.4%

Malaysia
255M tonnes
0.7%

Kuwait
89M tonnes
0.2%

Egypt
80M tonnes
0.6%

Algeria
65M tonnes (0.4%)

Chile
50M tonnes (0.2%)

Venezuela
100M tonnes
(0.4%)

Colombia
80M tonnes
(0.2%)

Angola
40M tonnes
(0.1%)

Peru
30M tonnes
(0.1%)

Uruguay
20M tonnes
(0.1%)

Africa
1.3 billion tonnes CO_2
3.7% global emissions

South America
1.1 billion tonnes CO_2
3.2% global emissions

Oceania
0.5 billion tonnes CO_2
1.3% global emissions

Our World
in Data

The construction
makes efficient use of
space

Nesting shows relative
size at multiple levels

No limit to the depth
of the branches

Treemaps: Google Newsmap

They turn out to be useful in a wide range of applications

Google NewsMap shows top news stories with

- Size ~ popularity
- Color: domain— **world news, sports, national, ...**
- Shades: recency

Interactivity: Hover, click to show details



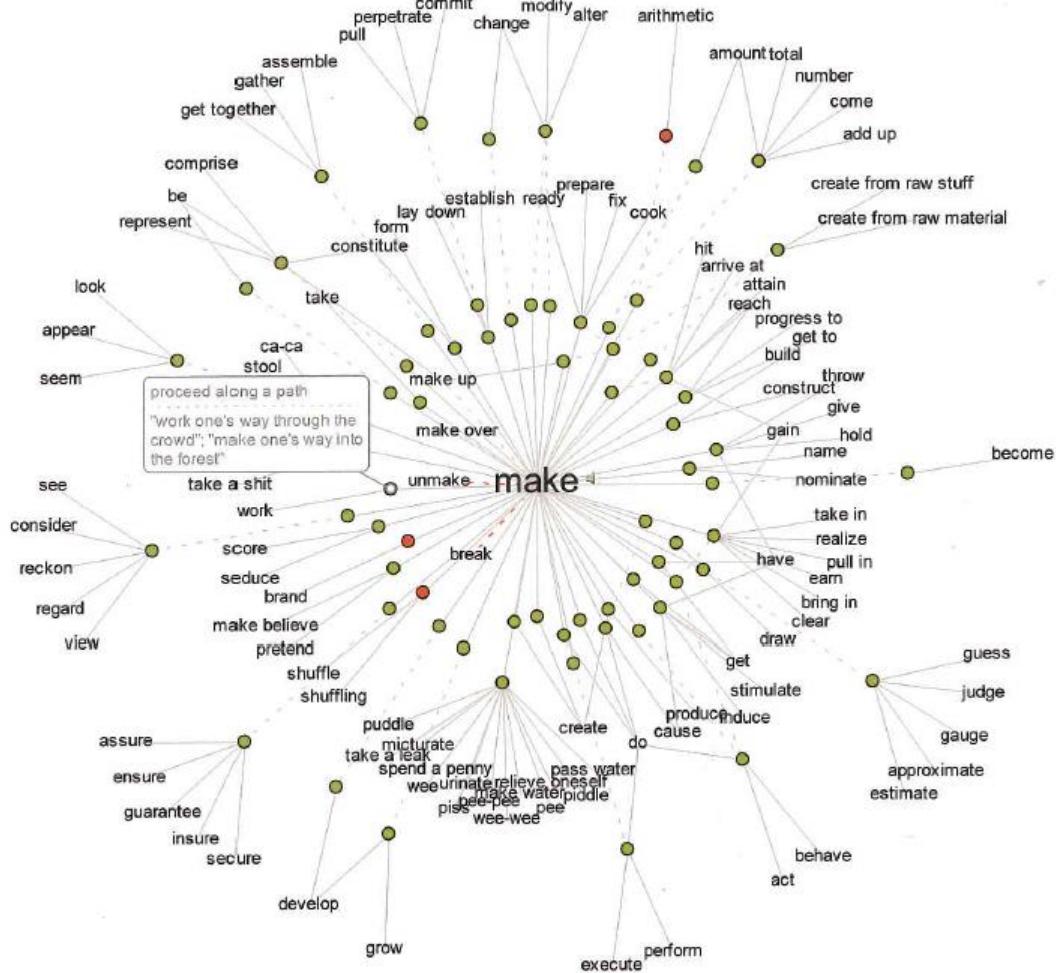
Radial trees: *Visual Thesaurus*

The *Visual Thesaurus*, from Thinkmap was the first application to make word meanings **visual** and **interactive**.

They used a radial layout to show the various related senses of given focus word.

This application was incisive in promoting ideas of interaction with tree-based data: query, zoom, tool-tips,

...



This fig from Manuel Lima, *The Book of Trees*, p. 127

Animation & Interactive Graphics

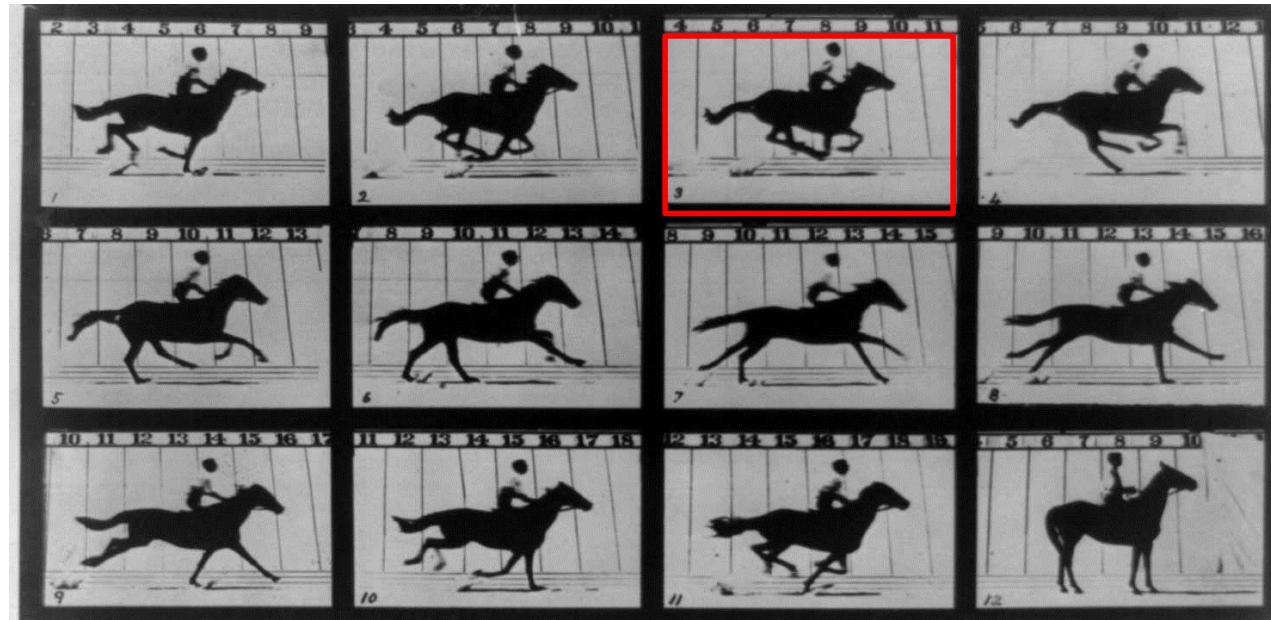
Origins: Visualizing motion
Animated graphics
Dynamically updated
graphics
Linking views
Interactive application
development frameworks



A wager about a horse in motion

In the late 1800s, a popular quasi-scientific question was: Does a horse, in a trot, cantor or gallop ever have all four feet off the ground?
This came to be called the **Hypothesis of Unsupported Transit**

Eadweard Muybridge solved the problem by automating multiple photographs



THE HORSE IN MOTION.

Illustrated by
MUYBRIDGE.

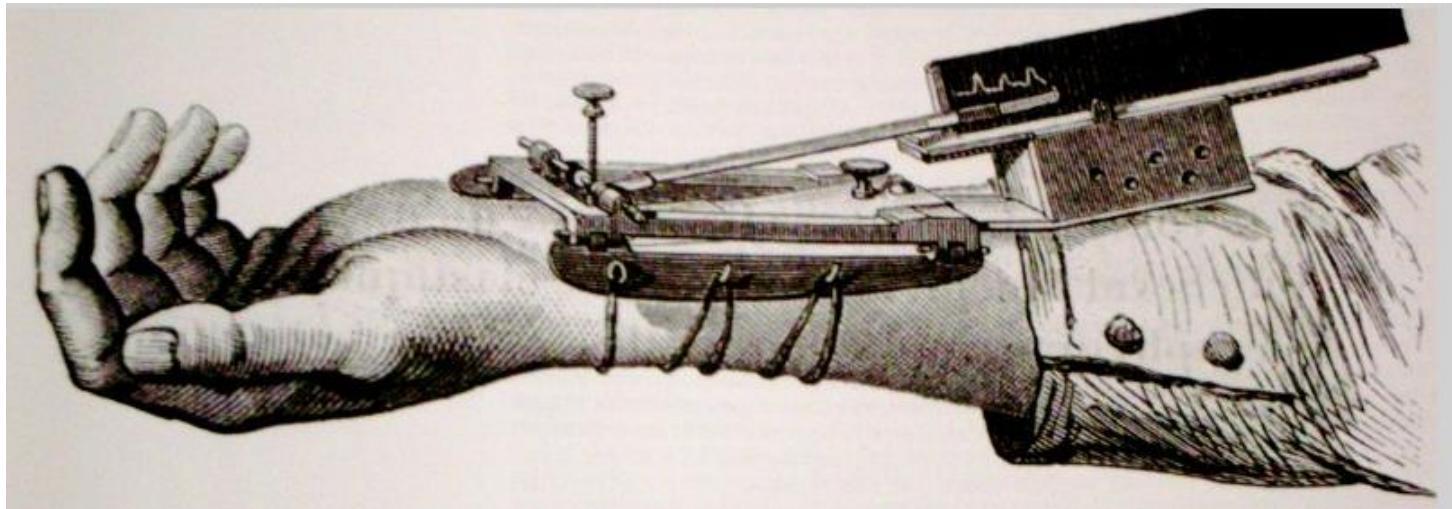
AUTOMATIC ELECTRO-PHOTOGRAPH.

"SALLIE GARDNER," owned by LELAND STANFORD; ridden by G. DOMM, running at a 1.40 gait over the Palo Alto track, 19th June, 1878.
The negatives of these photographs were made at intervals of twenty-seven inches of distance, and about the twenty-fifth part of a second of time; they illustrate consecutive positions assumed during a single stride of the mare. The vertical lines were twenty-seven inches apart; the horizontal lines represent elevations of four inches each.
The negatives were each exposed during the two-thousandth part of a second, and are absolutely "untouched."

É.-J. Marey: A science of visualizing motion

- Physiology: How to make internal physiological processes subject to visual analysis?
 - Invented many graphic recording devices (heart rate, blood pressure, muscle contraction, etc.)
 - “Every kind of observation can be expressed by graphs”

Marey's sphygmograph, recording a visual trace of arterial blood pressure



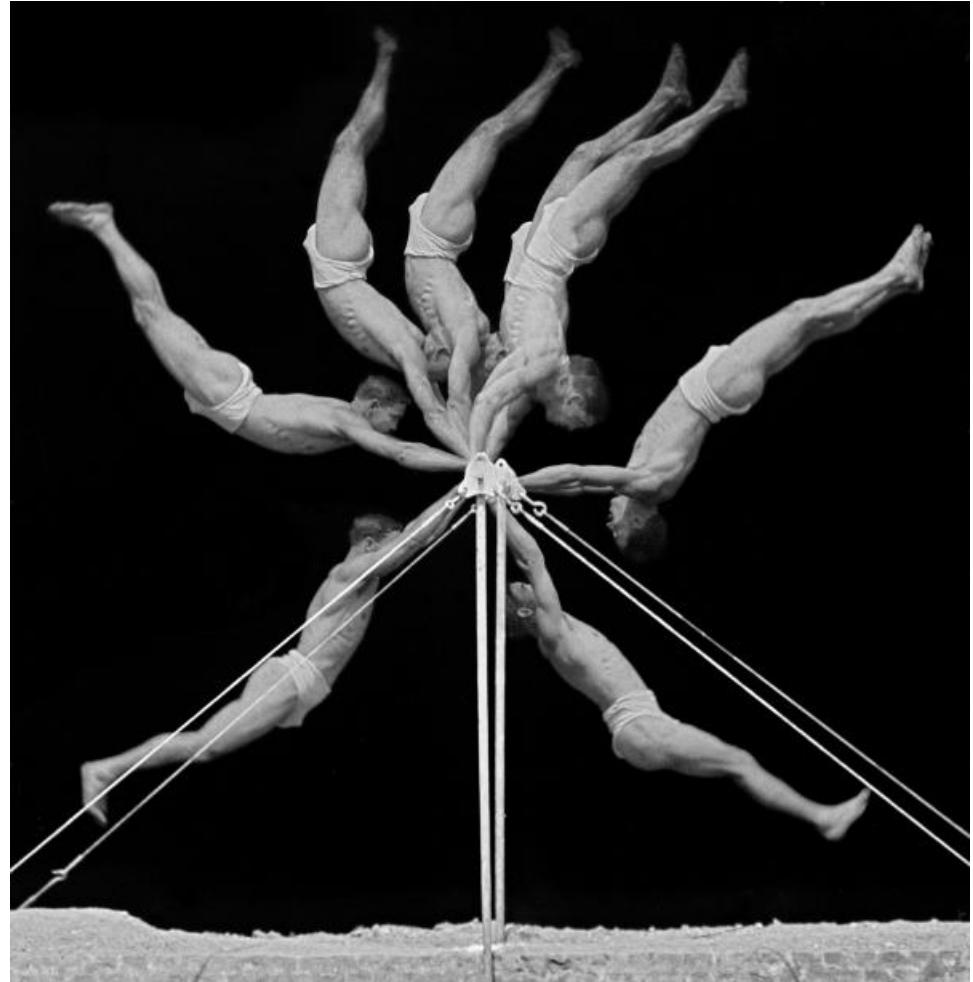
Animation: Chronophotography

Marey pioneered the study of human and animal motion photographically



Fig. 6. Mode d'emploi du fusil photographique.

The photographic gun, allowing recording of 12 frames/sec. at intervals of 1/720 of a second

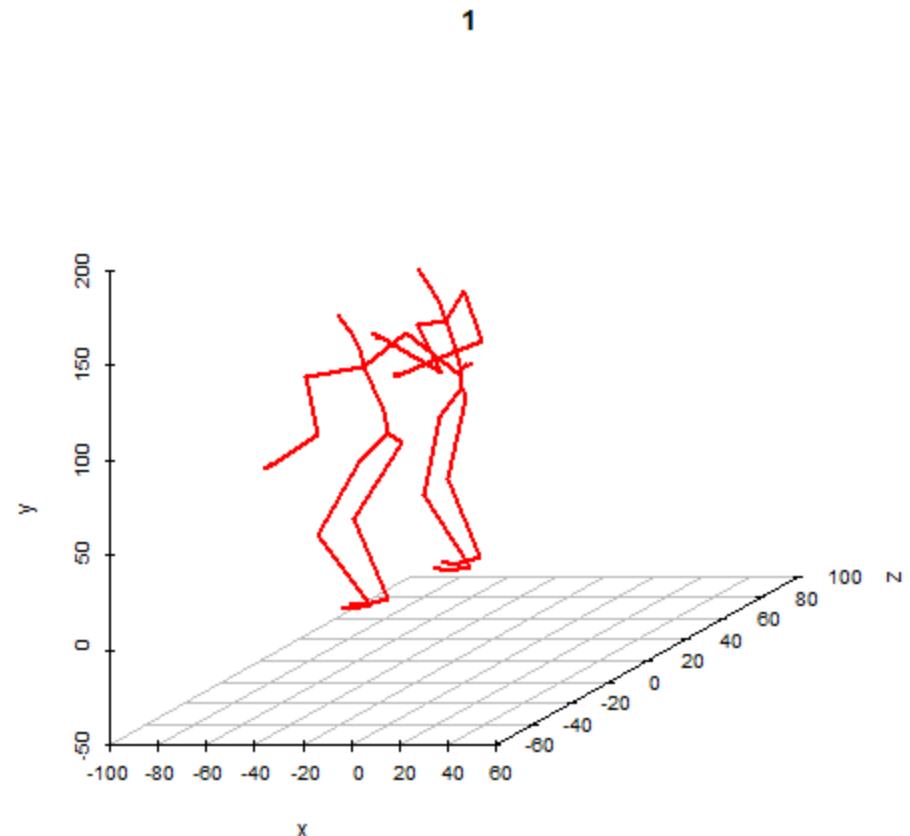


Animated graphics

Animated graphics, like movies
are just a series of frames strung
together in a sequence

The data for this animation come
from human figures in motion-
capture suits dancing the Charleston.

The Carnegie-Mellon Graphics Lab
maintains a Motion Capture
Database, <http://mocap.cs.cmu.edu/>

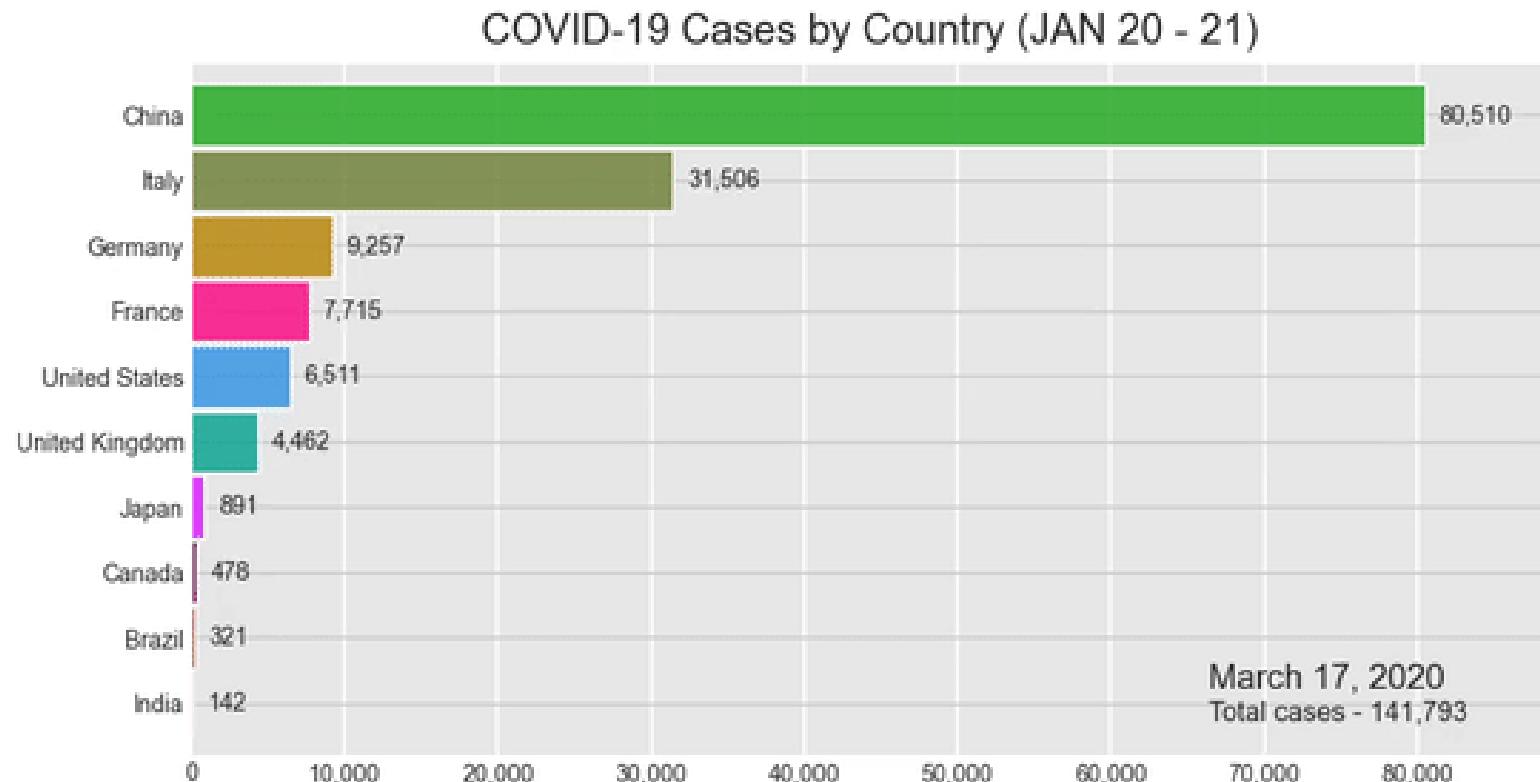


From: <http://blog.revolutionanalytics.com/2017/08/3-d-animations-with-r.html>

Bar chart races

Data that changes over time can often be shown in a simpler animated graphic

This example of a **bar chart race** shows the strengths & weaknesses of this approach.



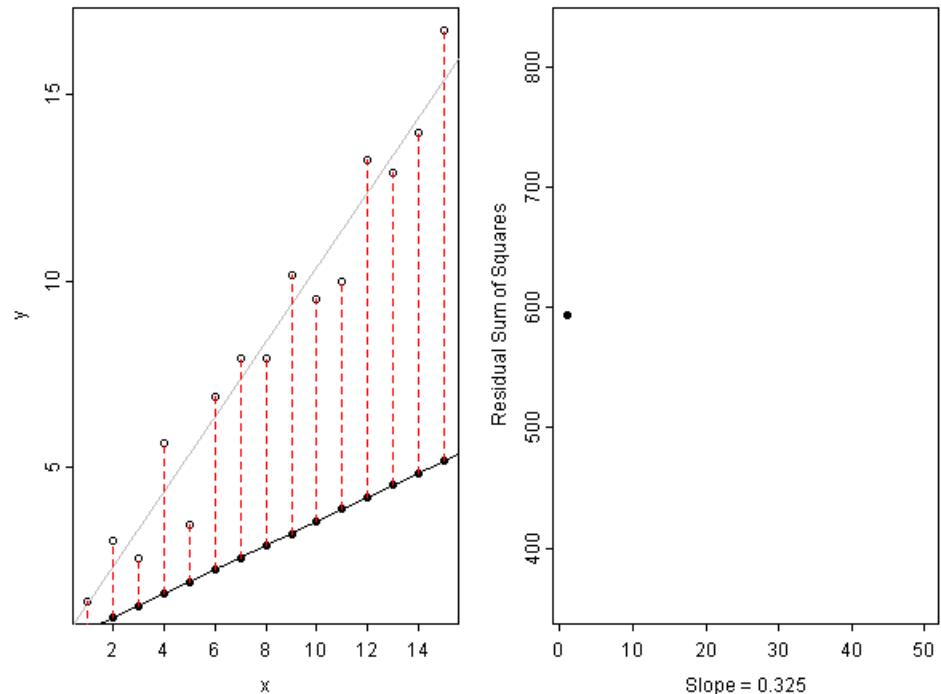
Statistical animations

Statistical concepts can often be illustrated in a dynamic plot of some process.

This example illustrates the idea of least squares fitting of a regression line.

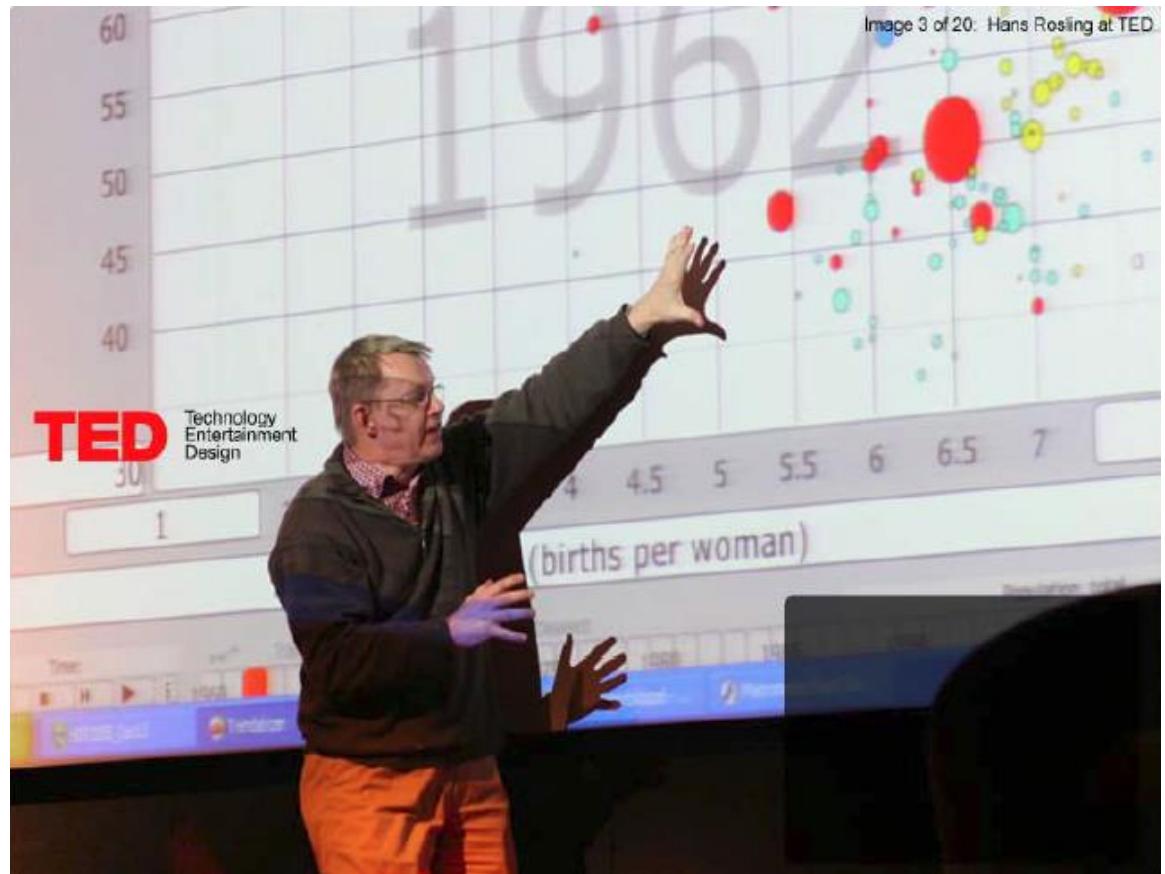
As the slope of the line is varied, the right panel shows the residual sum of squares.

This plot was done using the [animate](#) package in R.



Animated graphics

Hans Rosling captivated audiences with dynamic graphics showing changes over time in world health data



Video: Hans Rosling, “The best stats you’ve ever seen,”

https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

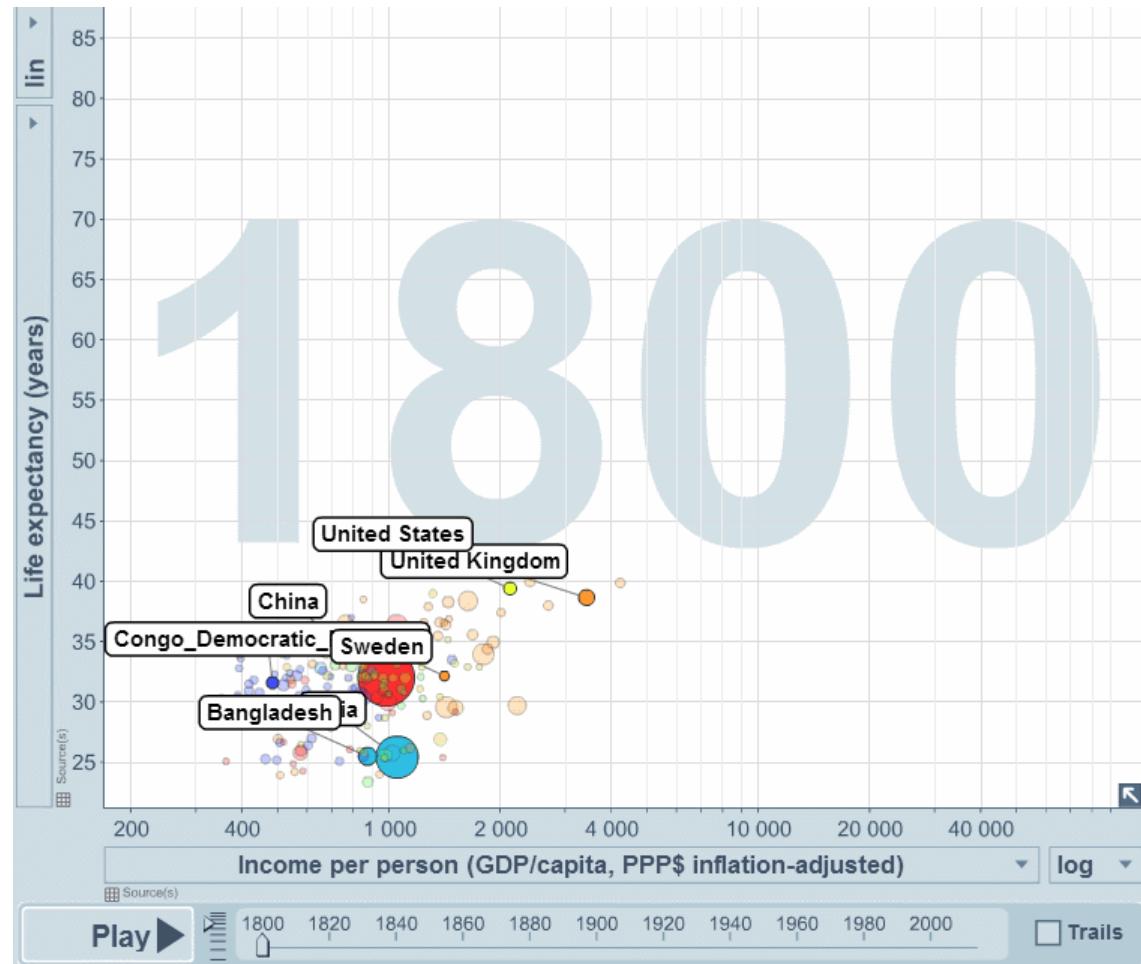
Animation & Interactivity

The Gapminder “moving bubble chart” was the vehicle.

- Choose (x, y) variables
- Choose bubble size variable
- Animate this over time

Liberating the X axis from time opened new vistas for data exploration

Software made this available as a general tool

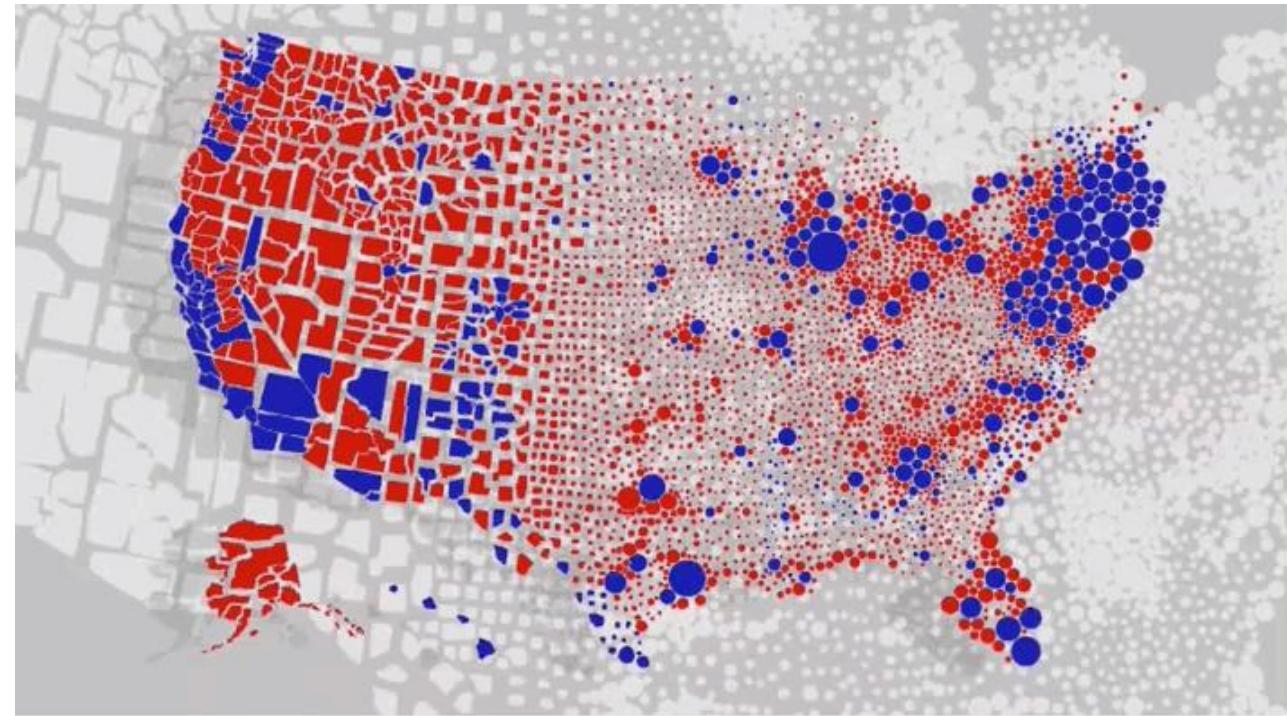


Animation: Interpolated views

Animation can also be used to show the difference between two views, using interpolated transitions: $\text{Current} = \alpha \text{ view}_1 + (1 - \alpha) \text{ view}_2$

This image showed Rep vs. Dem votes in the 2016 US election, contrasting shading by area vs. shading by population.

Land doesn't vote;
people do



[Image: Karim Douieb/Jetpack.ai]

Linking animated views

This example links a **dendrogram** to a **grand tour** and **map** of the USArrests data to visualize a classification in 5 dimensions

The grand tour animates a series of 2D projections of the 5D data

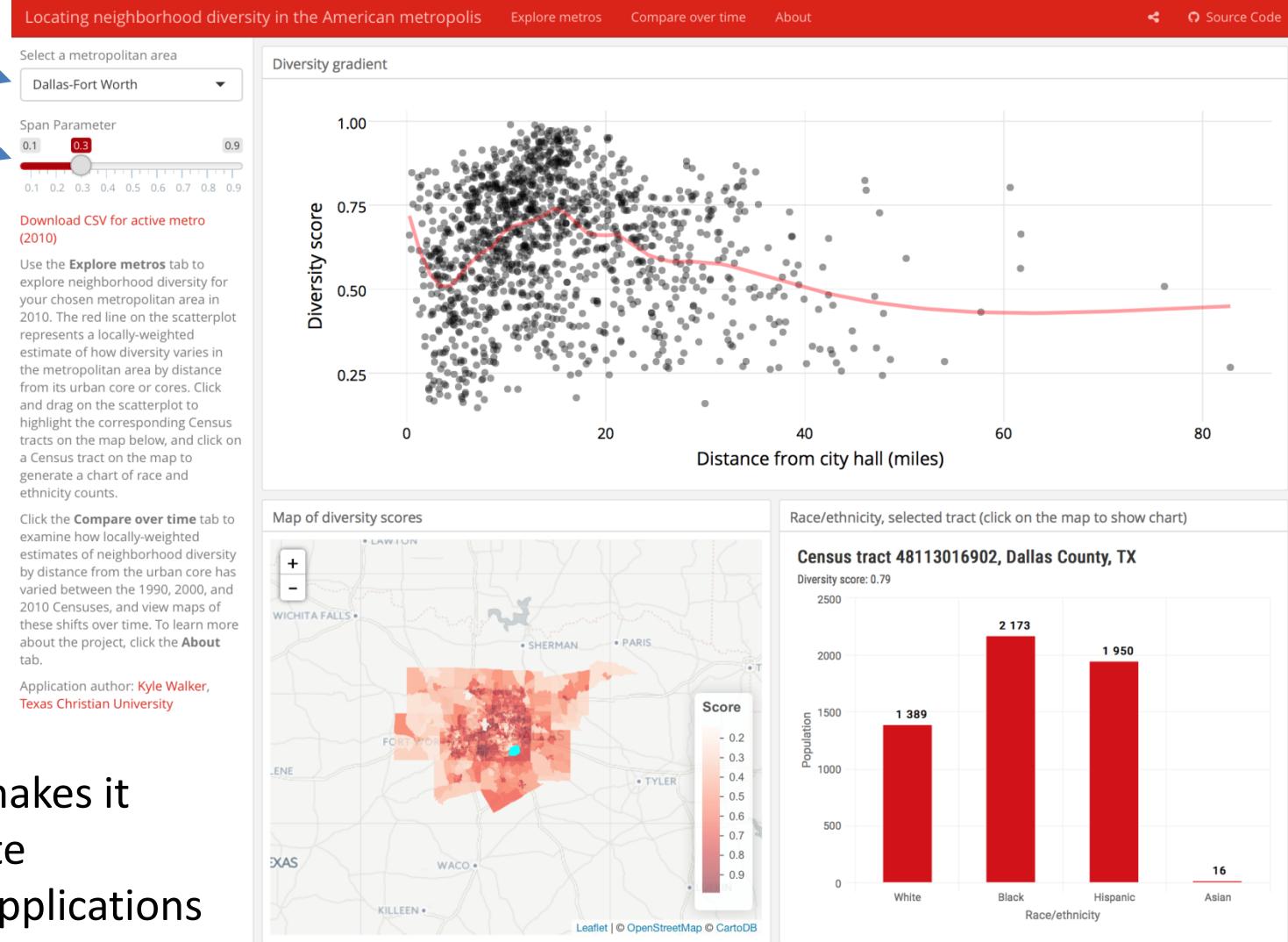
The image is recorded as a GIF



Interactive application frameworks

selectors

inputs



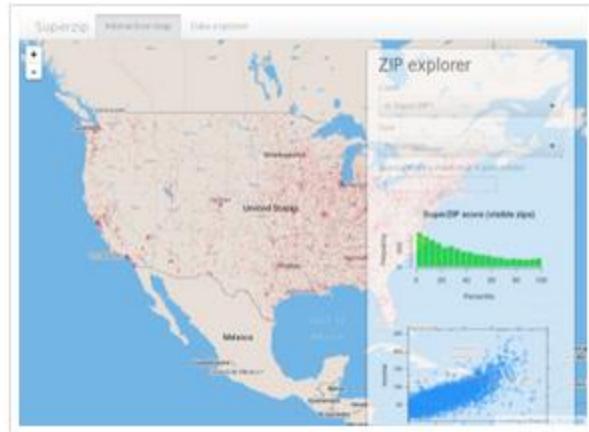
shiny for R makes it
easy to create
interactive applications

shiny gallery

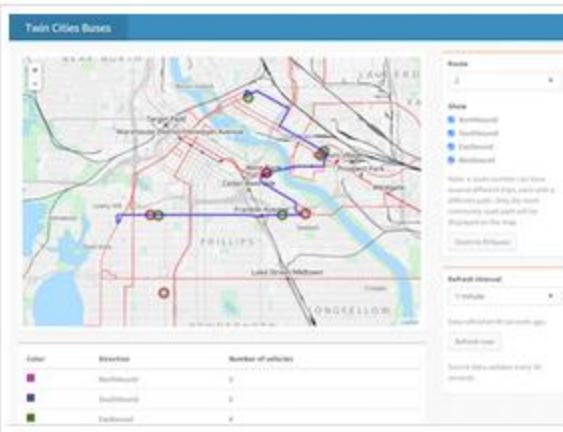
There is now a large collection of shiny applications, <https://shiny.rstudio.com/gallery/>
These integrate other interactive web software: d3, Leaflet, Google Charts, ...

Interactive visualizations

Shiny is designed for fully interactive visualization, using JavaScript libraries like d3, Leaflet, and Google Charts.



SuperZip example



Bus dashboard



Movie explorer

Summary

- The topics here were largely about data graphs, for analysis & presentation. Mainly not Info-graphics
 - Quantitative data: different forms for 1D, 1.5D, 2D, 3+D data
 - Categorical data: often best shown as areas ~ frequency (bar plots, mosaic plots)
- Thematic maps: visualizing spatially varying data
 - Raw data with different visual encodings
 - Spatial statistical models provide some smoothings
- Networks/trees: visualizing connections
- Animation: show changes over time or space
- Interaction: allow the viewer to explore the data