CS-E4840
Information Visualization
Lecture 4: Visualization techniques

Tassu Takala <tapio.takala@aalto.fi>
7 March 2019

Theory of data graphics

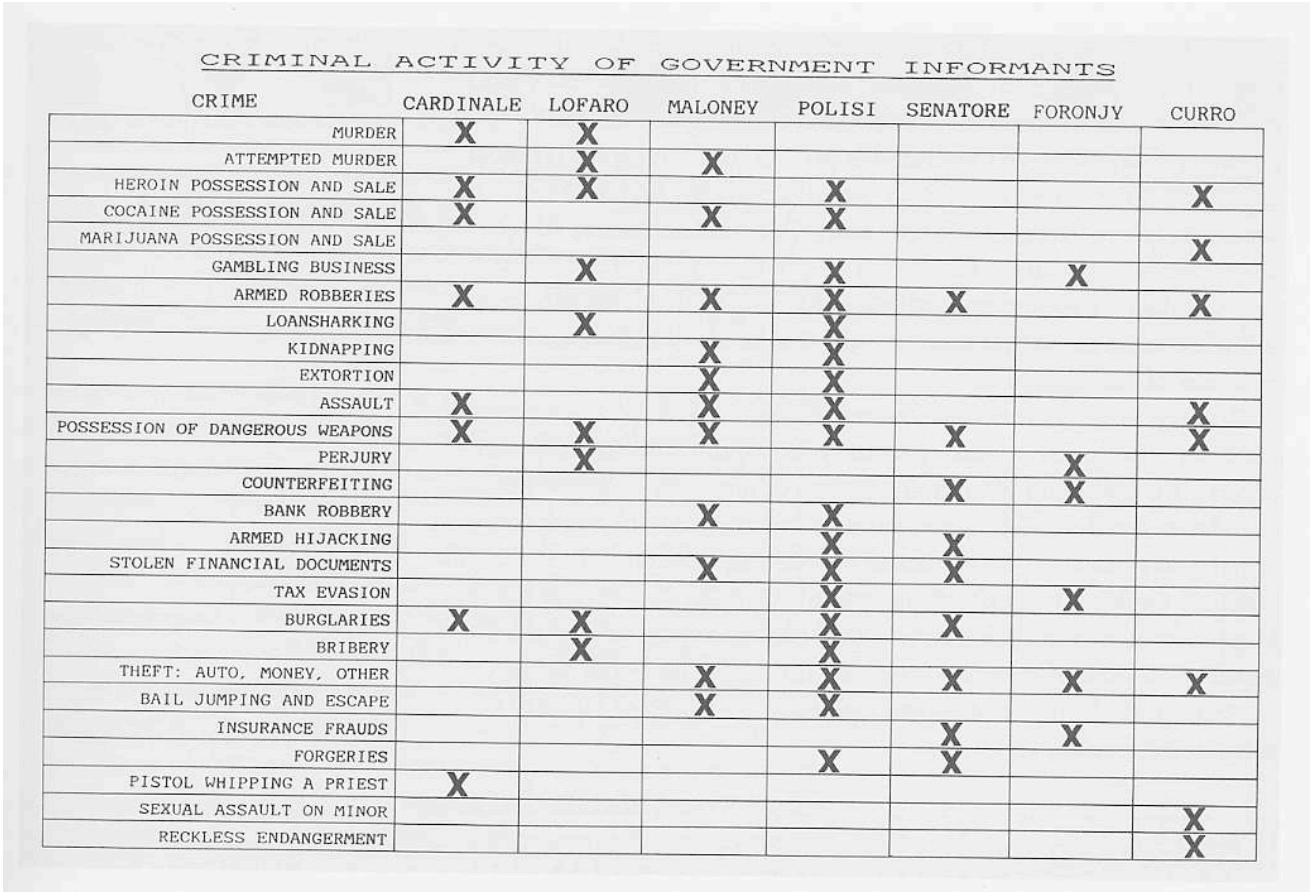
- The idea:
  - Give the viewer the greatest number of ideas in the shortest time
  - Use the least amount of ink
  - Don't waste space
  - Eliminate non-essentials and redundancies
- Or:
  - Make the graphics as easy to read and as simple as possible, while displaying the data fully.

Theory of data graphics

- Data-ink
- Chartjunk
- Multifunctioning graphical elements
- Data density and small multiples
- Aesthetics and techniques

Line weight and lettering

An excellent summary of crimes committed by state's witnesses in a Mafia trial. Notice the thick glyphs and how the most horrid crimes are listed first and last.



CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS

CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			X
POSSESSION OF DANGEROUS WEAPONS	X	X	X	X	X		X
PERJURY		X				X	
COUNTERFEITING					X	X	
BANK ROBBERY			X	X			
ARMED HIJACKING				X	X		
STOLEN FINANCIAL DOCUMENTS			X	X	X		
TAX EVASION				X		X	
BURGLARIES	X	X		X	X		
BRIBERY		X		X			
THEFT: AUTO, MONEY, OTHER			X	X	X	X	X
BAIL JUMPING AND ESCAPE			X	X			
INSURANCE FRAUDS					X	X	
FORGERIES				X	X		
PISTOL WHIPPING A PRIEST	X						
SEXUAL ASSAULT ON MINOR							X
RECKLESS ENDANGERMENT							X

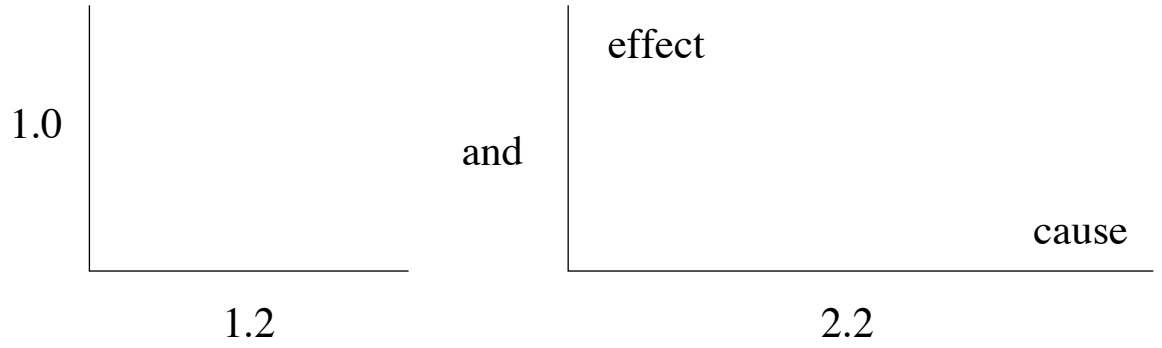
United States vs. Gotti et al, 1987 [EI 31].

Proportion of graphics

Graphics should usually have greater length than height:

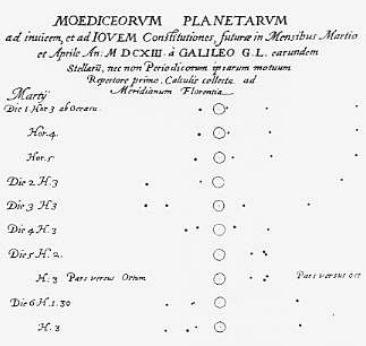
- Our eye is practiced in detecting deviations from the horizon. Thus e.g. horizontal time-series are easier to read.
- It is easier to write words and labels horizontally.
- Longer horizontal helps to emphasize the causal variable

Preferred height/length ratios vary depending on the circumstances; the golden ratio 1:1.618 is a good rule of thumb.



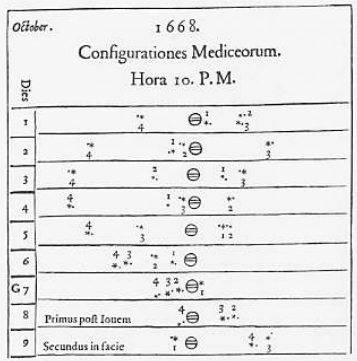
Jovian moons

- On 10 January 1610 Galileo Galilei was able to separate the motion of the Jovian satellites from that of the planet.
- It took 300 years to move from dots to continuous curves, with muted horizontal lines, that report every position of the moons.



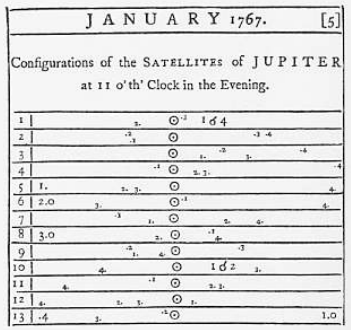
MOEDICEORVM PLANETARVM

Galileo Galilei, *Istoria e dimostrazioni intorno alle macchie solari* . . . [Welser sunspot letters], (Rome, 1613), illustration of satellites (called by Galileo "Medicean stars" in honor of his patron) following p. 150.



October. 1668. Configurationes Mediceorum. Hora 10. P.M.

Jean Domenique Cassini, *Ephemerides Bononienses Mediceorum syderum ex hypothesibus, et tabulis Io,* (Bologne, 1668), p. 34.

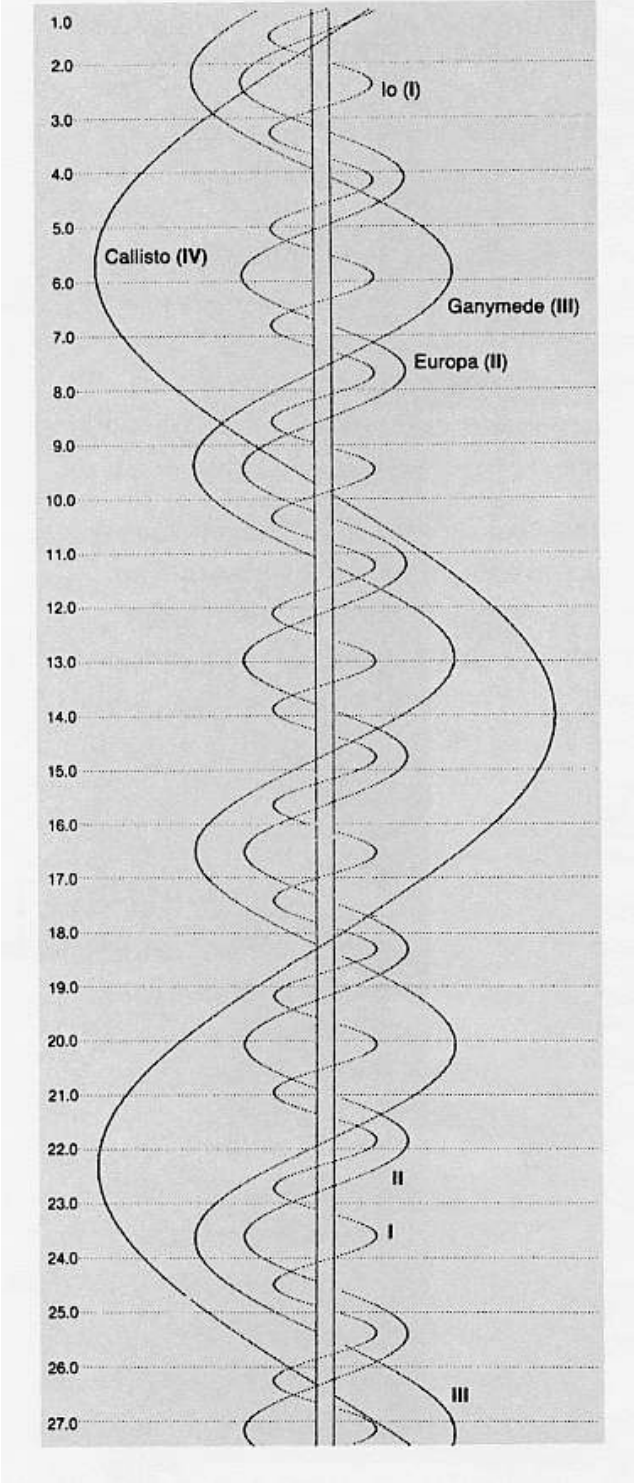


JANUARY 1767. [5]

Configurations of the SATELLITES of JUPITER at 11 o'th' Clock in the Evening.

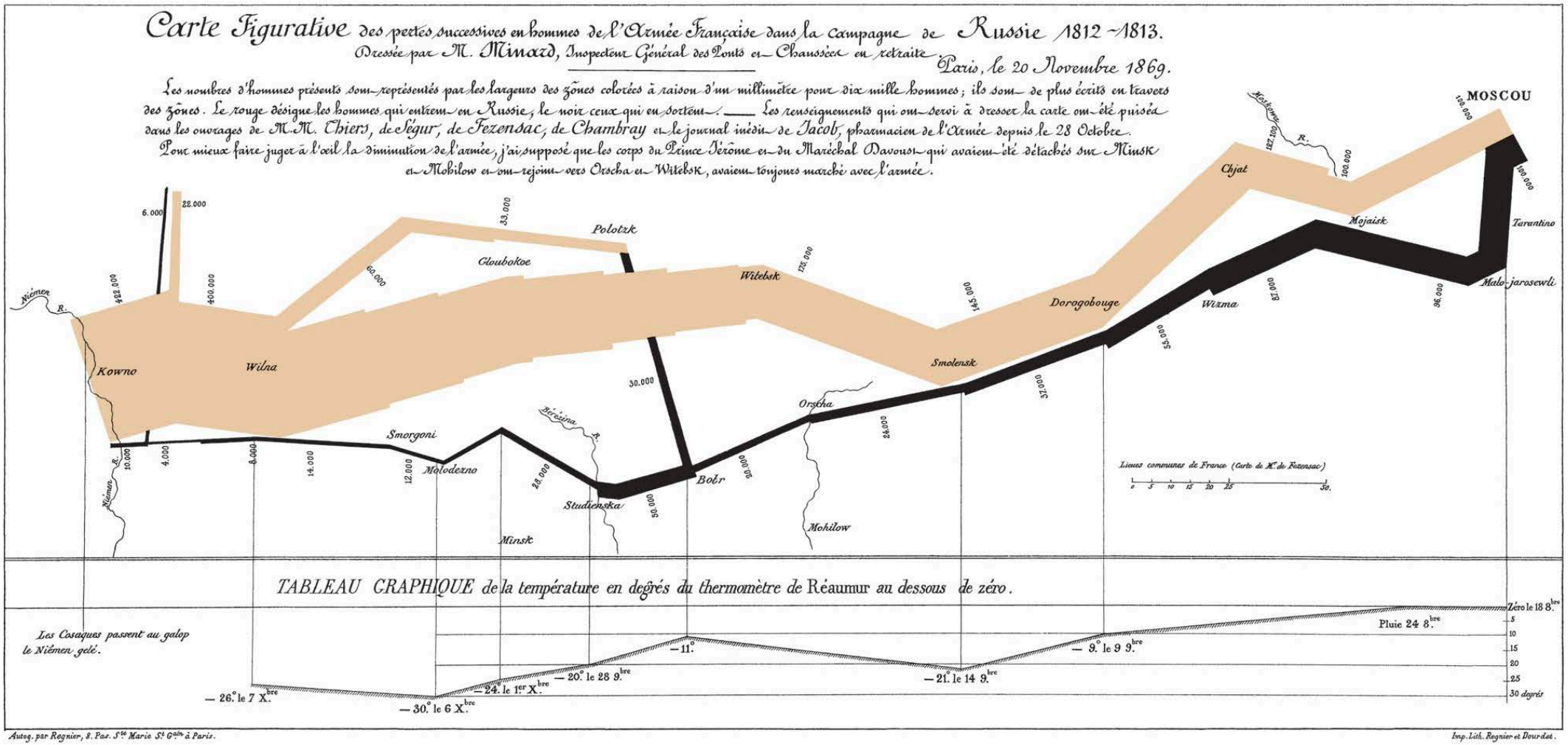
Bureau des Longitudes, *Connaissance des Temps* (Paris, 1766), p. 5.

[2] Translation of *The Starry Messenger* by Stillman Drake, in his *Telescopes, Tides, and Tactics* (Chicago, 1983), pp. 59–63.



Sky & Telescope, 1988 [El 100].

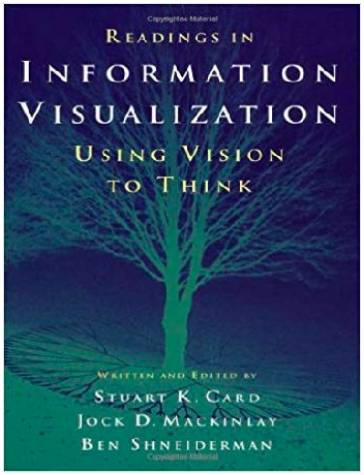
Carte Figurative by Minard



Outline

- **Techniques:**
  - Bars, boxes, lines, dots
  - multiple plots
  - reference lines and regions
  - rescaling /normalising / re-expressing
  - colours
- **Problems:**
  - axis ranges
  - use of 3D
  - overplotting
- **Scenarios:**
  - distribution analysis
  - ranking and part-of-whole analysis
  - time-series
  - high-dimensional data

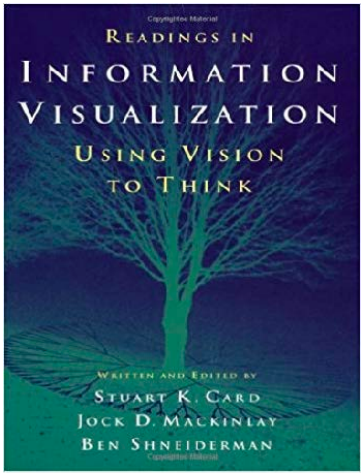
- **Related reading:** *Few. Now you see it. Analytic Press, 2009.*
- **Older but relevant:** *Card et al. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, 1999.*



Outline

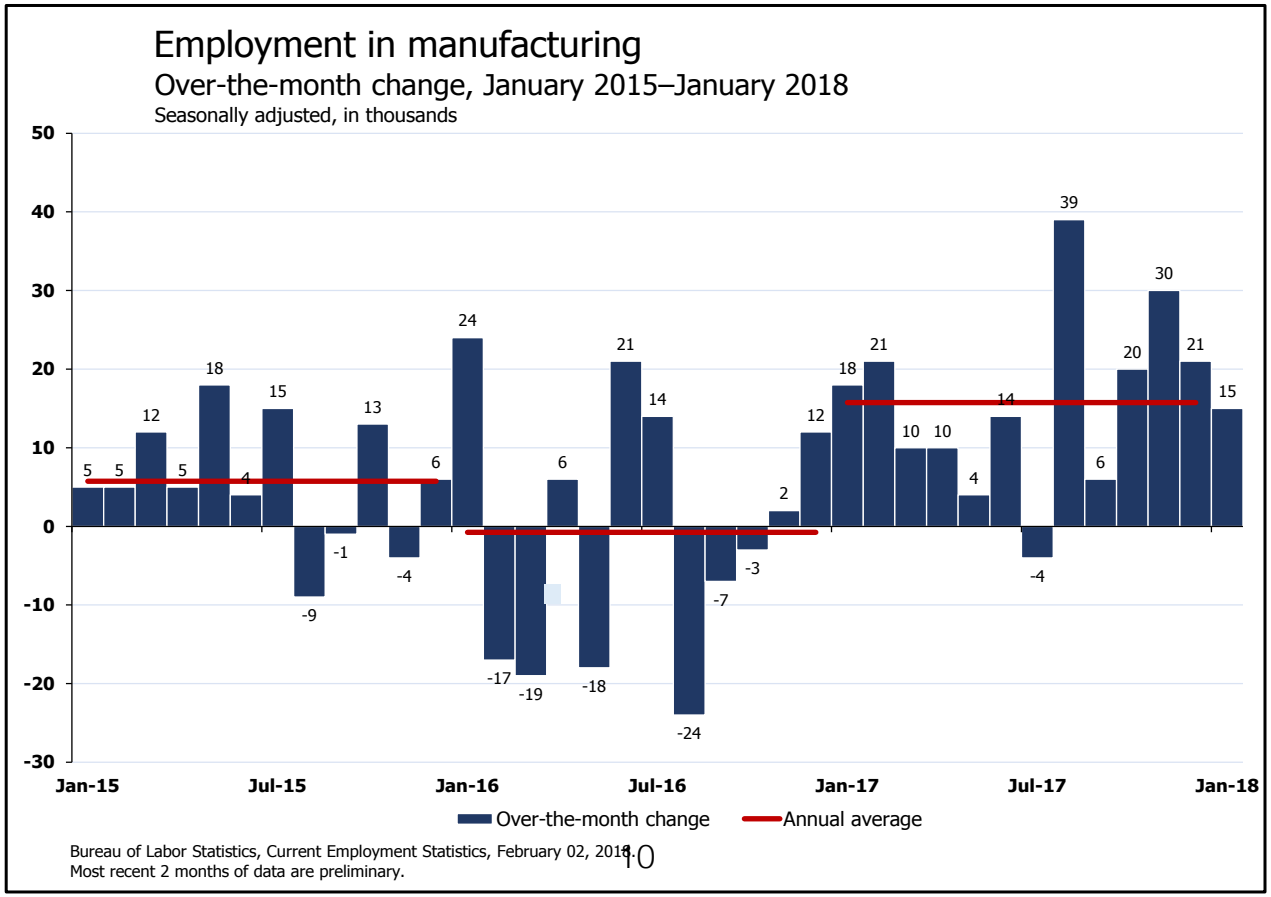
- **Techniques:**
  - Bars, boxes, lines, dots
  - multiple plots
  - reference lines and regions
  - rescaling /normalising / re-expressing
  - colours
- **Problems:**
  - axis ranges
  - use of 3D
  - overplotting
- **Scenarios:**
  - distribution analysis
  - ranking and part-of-whole analysis
  - time-series
  - high-dimensional data

- **Related reading:** *Few. Now you see it. Analytic Press, 2009.*
- **Older but relevant:** *Card et al. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, 1999.*



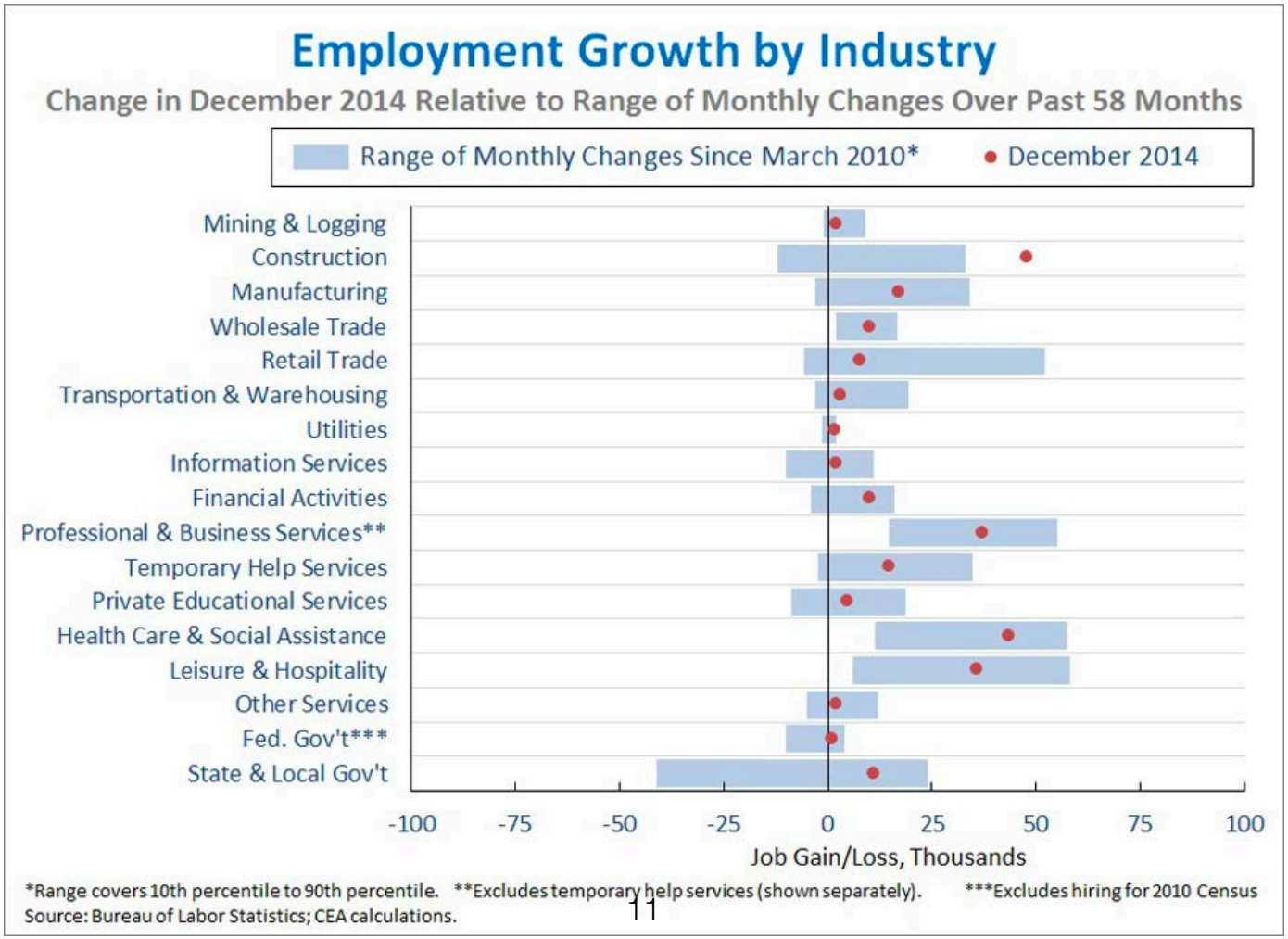
Basic elements: bars

- vertical (or horizontal) height to encode a value
- bars must have a 0 baseline
- but they can be negative
- bars can be displayed as groups
- best option for comparing individual values



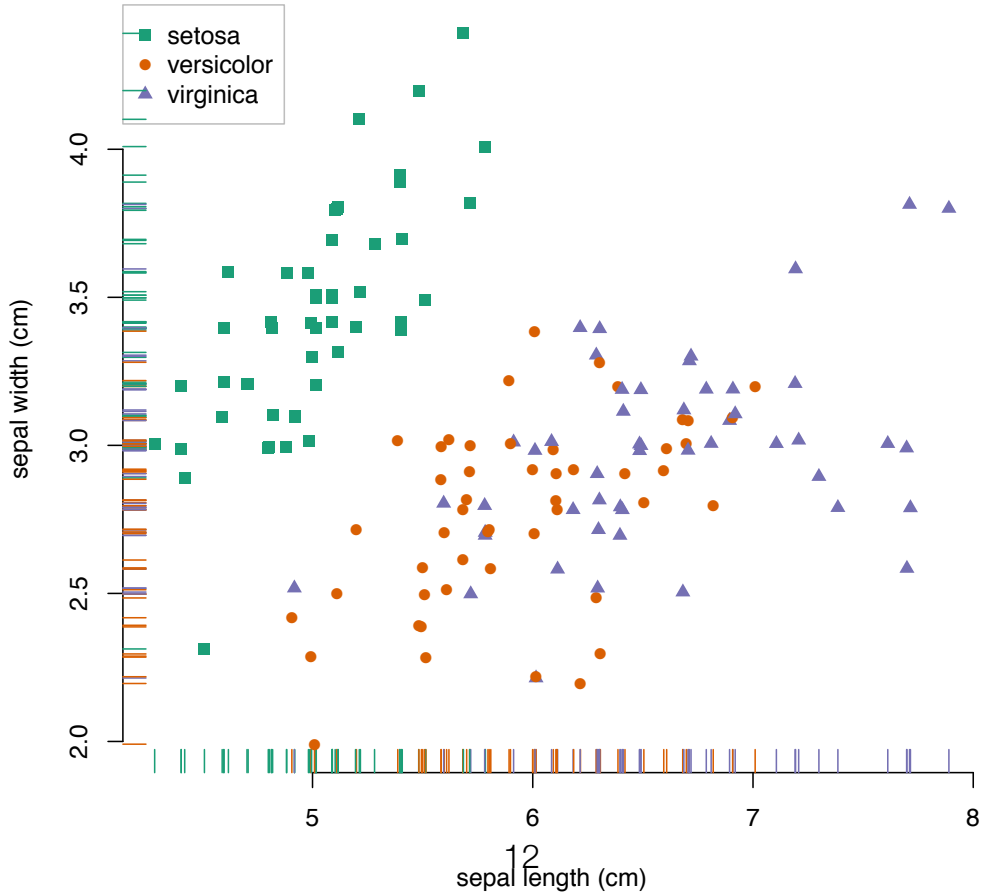
Basic elements: boxes

- boxes are used to encode ranges, for example
  - error bars or
  - percentiles of data



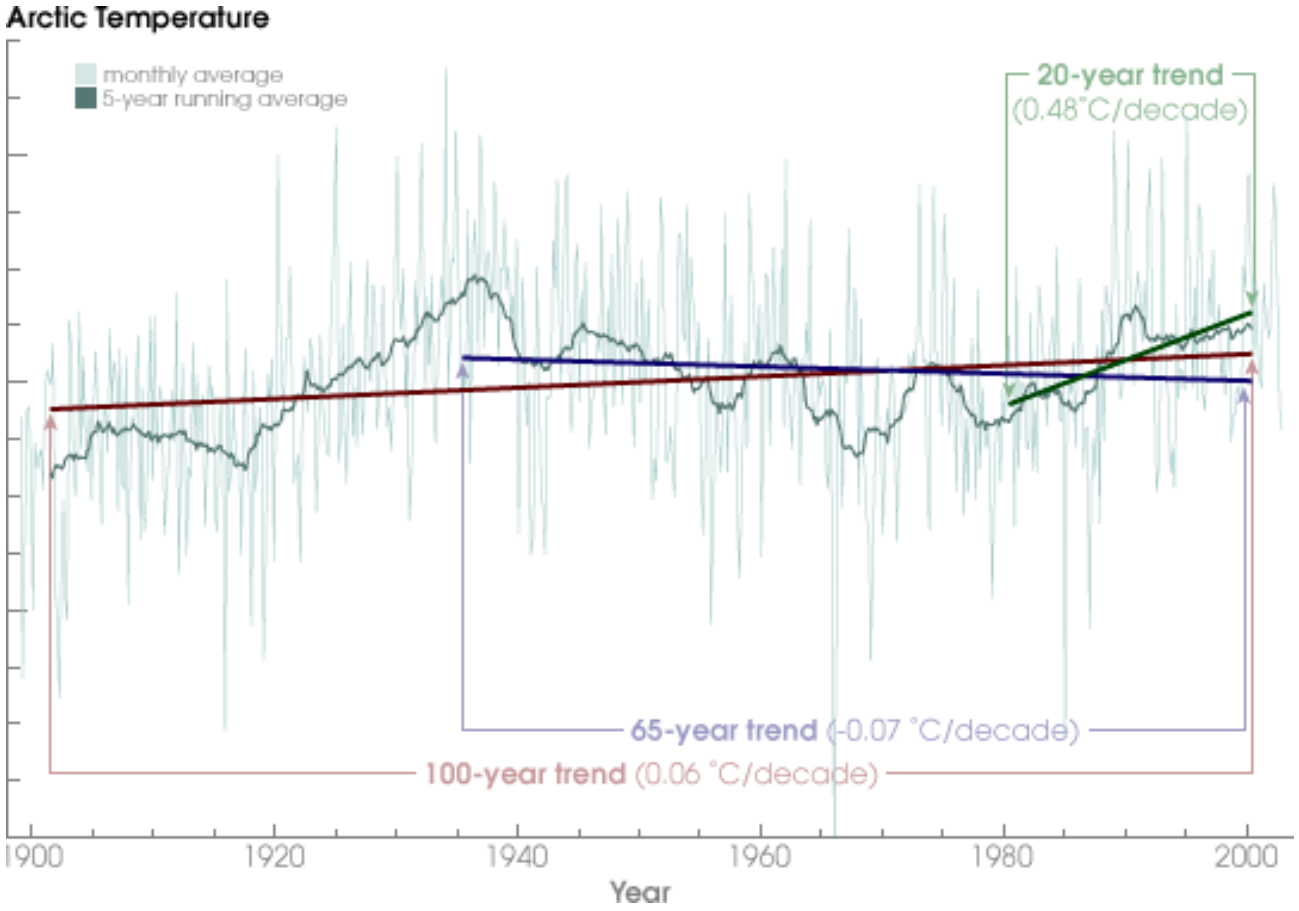
Basic elements: points

- use location on a plane to encode data
- points can have different colors
- ...and different shapes to encode additional value
- more complex shapes (glyphs) can be used used to encode multiple dimensions



Basic elements: lines

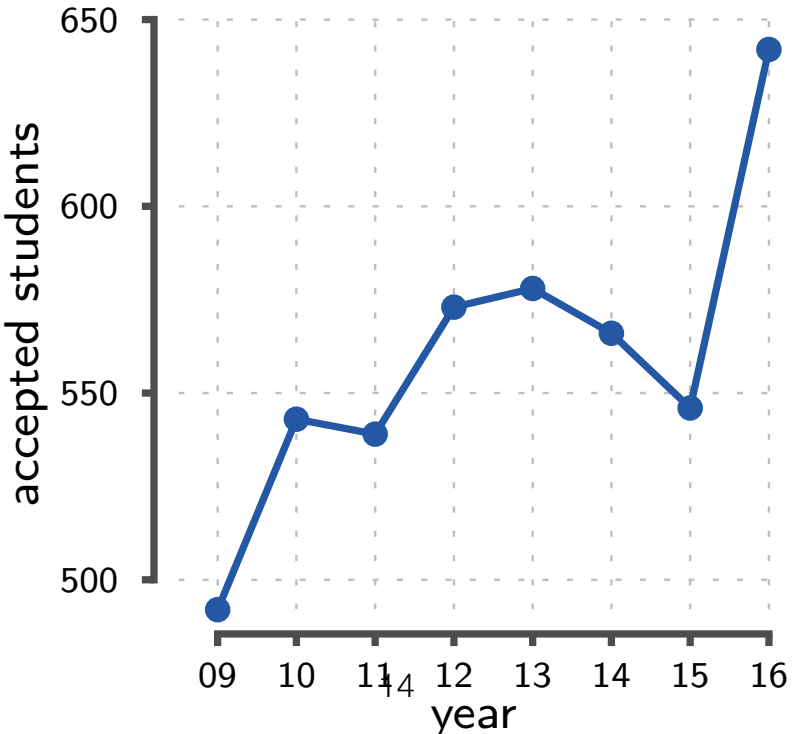
- combines successive data points in a 2D plane
- useful for revealing trends / variation / outliers



https://earthobservatory.nasa.gov/Features/ArcticIce/arctic_ice3.php

When can we use line charts

- Line charts are **primarily** meant for **continuous data** on the x-axis but
  - they *can also* be used with discrete data on the x-axis
  - you *can even use* it with categorical data on the x-axis, if it is ordered in a meaningful way

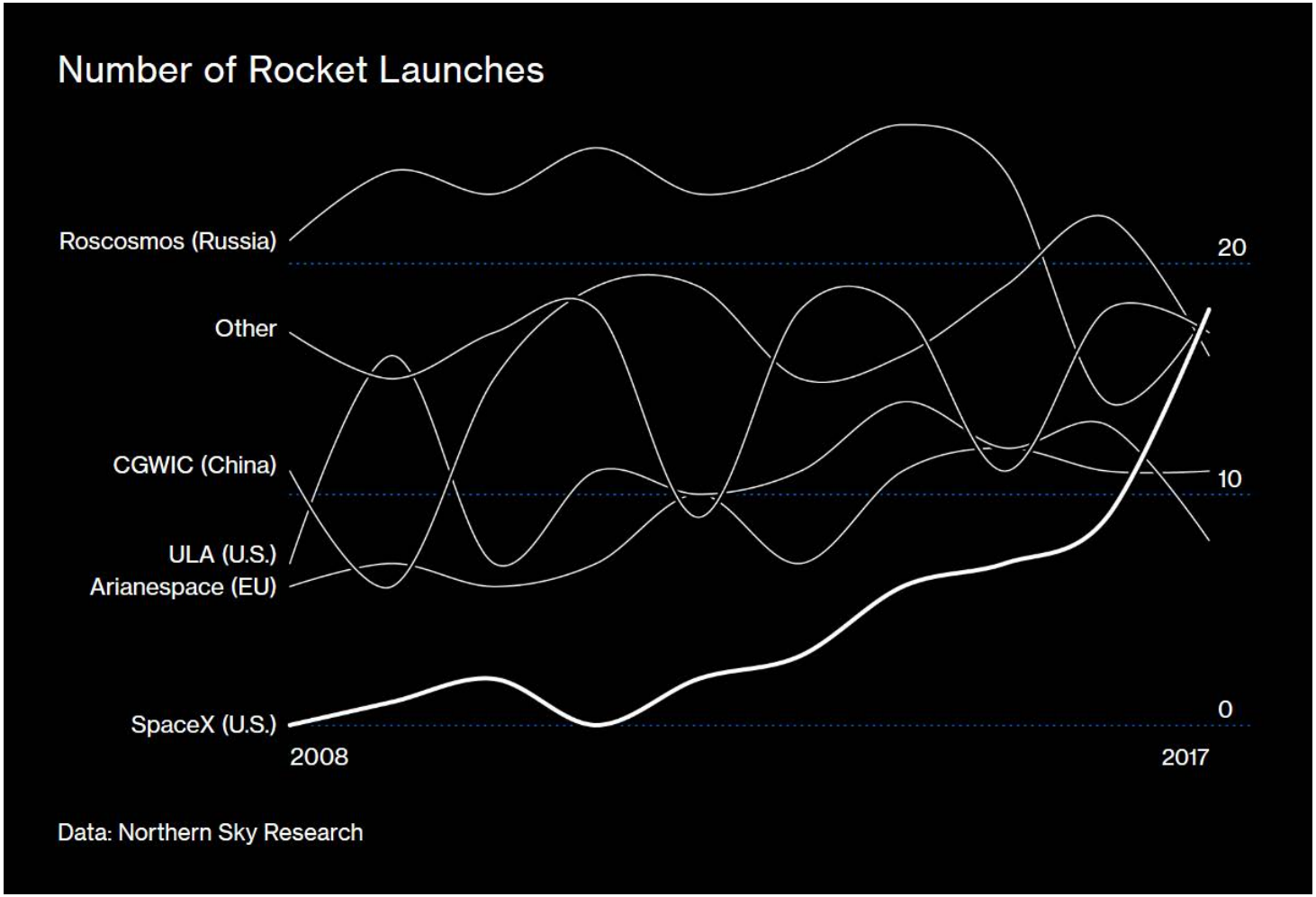


When should we **not** use line charts

- You should be careful with line charts if
  - you have **gaps** in your data
  - line chart suggests that the **missing data** can be obtained via linear **interpolation** (this may not be true)
  - at minimum, you should indicate the actual data points with markers (e.g., dots)

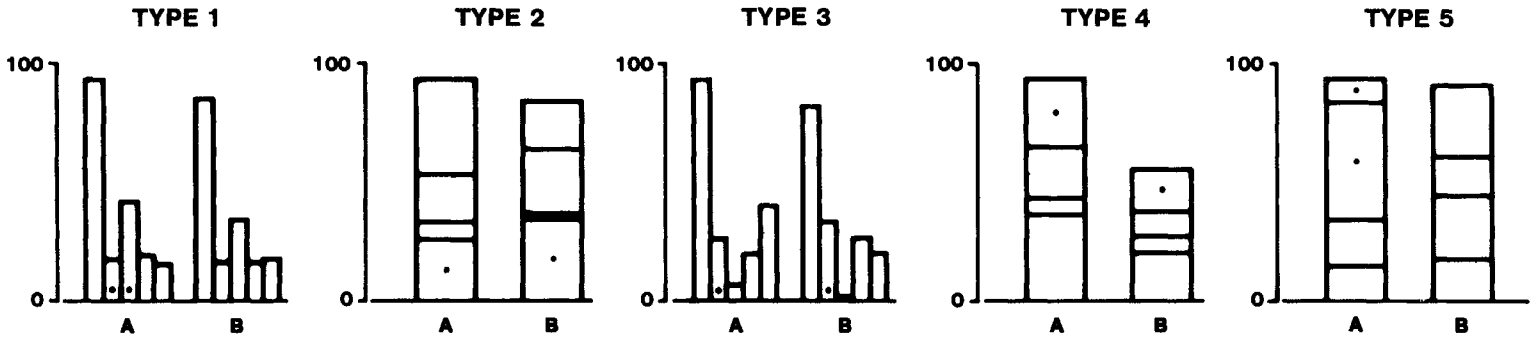
- line chart doesn't make sense if there is no meaningful order on x-axis

When should we not use line charts



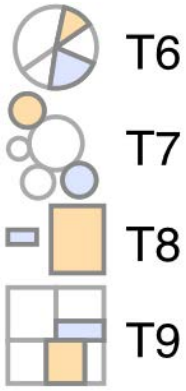
- strictly integer data should not be interpolated with curves

Which chart type is most effective?

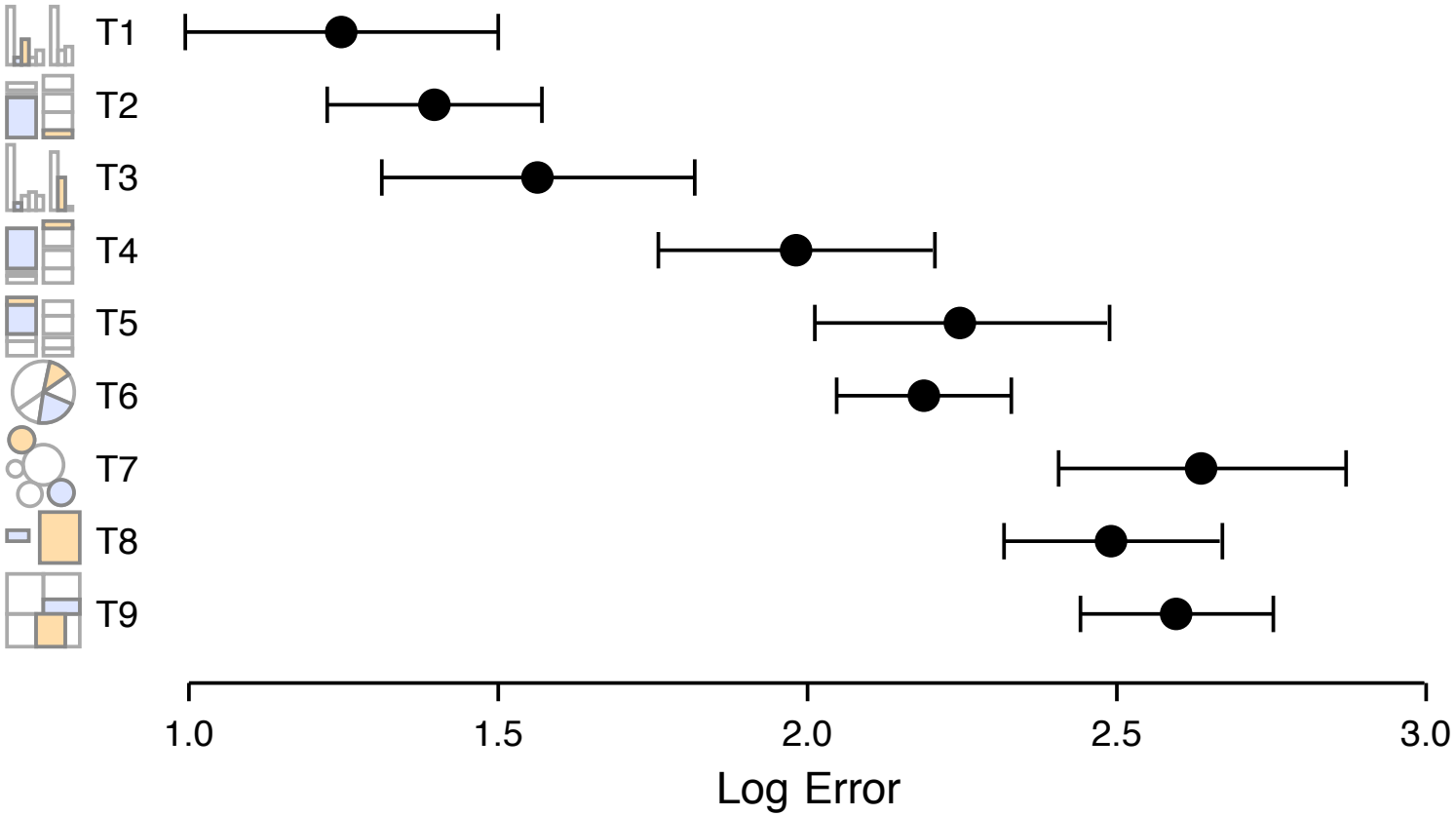


- Seminal paper by Cleveland & McGill (1984)
  - 55 subjects, 5 types of tasks
  - 10 copies of each task with different proportions
  - asked proportional difference in percentage

- Replicated and extended by Heer and Bostock (2010)
  - T6: angle (pie chart)
  - T7: area (bubble chart)
  - T8: area of vertically centered rectangles
  - T9: area of rectangles
  - crowdsourced with Amazon's Mechanical Turk



Effectiveness of different graphical designs



Heer and Bostock, 2010

$$\text{log error} = \log_2(|\text{estimated difference} - \text{true difference}| + 1/8)$$

Multiple plots

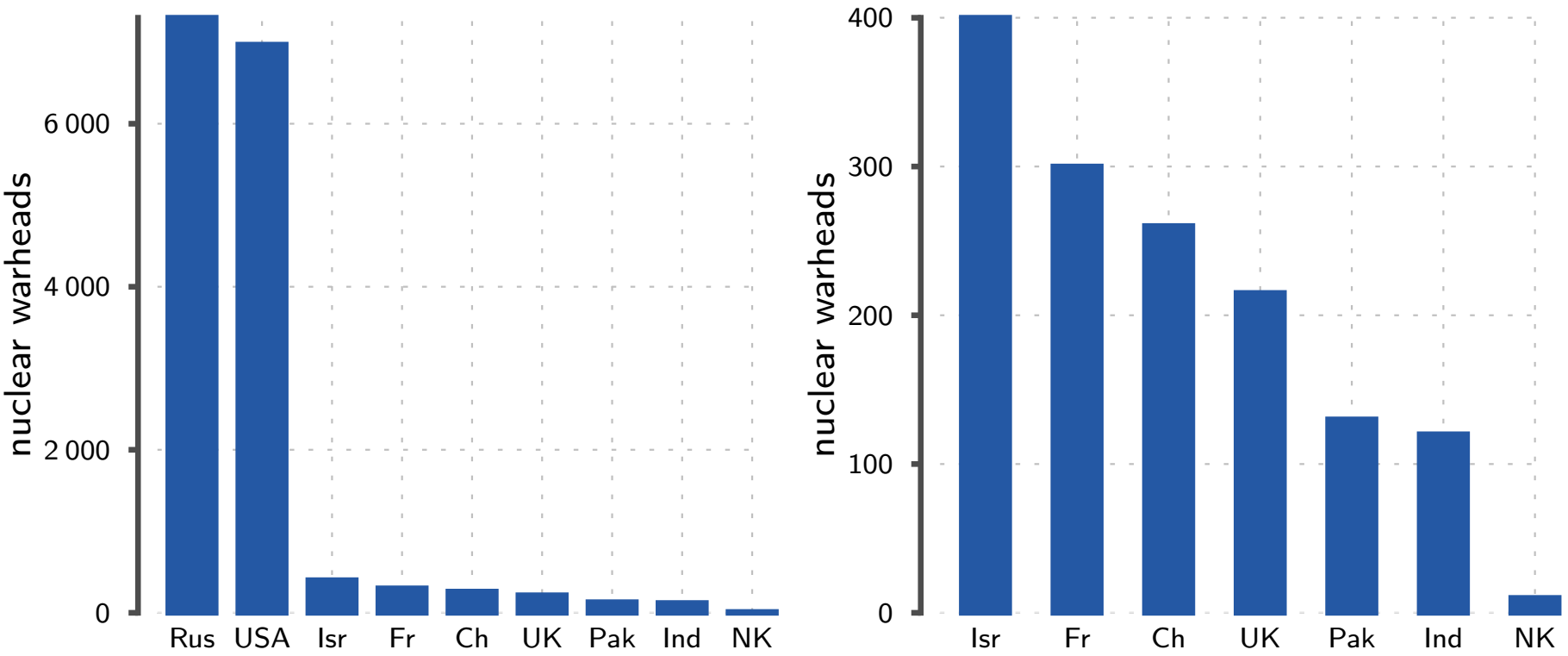
- Complex data/story is often difficult to present in one figure
  - use several figures that are somehow linked to each other to tell a story
    - pros: gives flexibility beyond a single plot
    - cons: may take a lot of space
    - cons: may create a graphical puzzle

- Common techniques for combining multiple plots
  - small multiples = trellis displays
  - overview / detail
  - multiform = multiple concurrent views

  (naming varies from author to author)

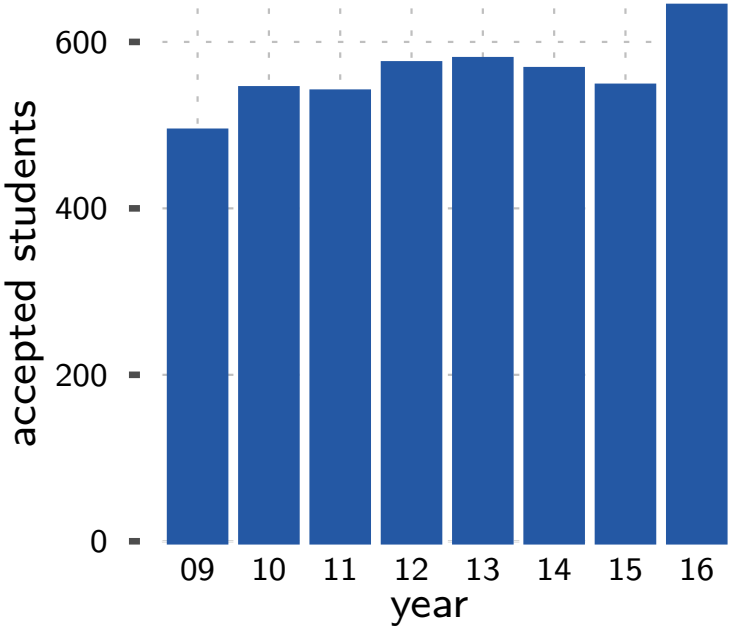
Overview / detail

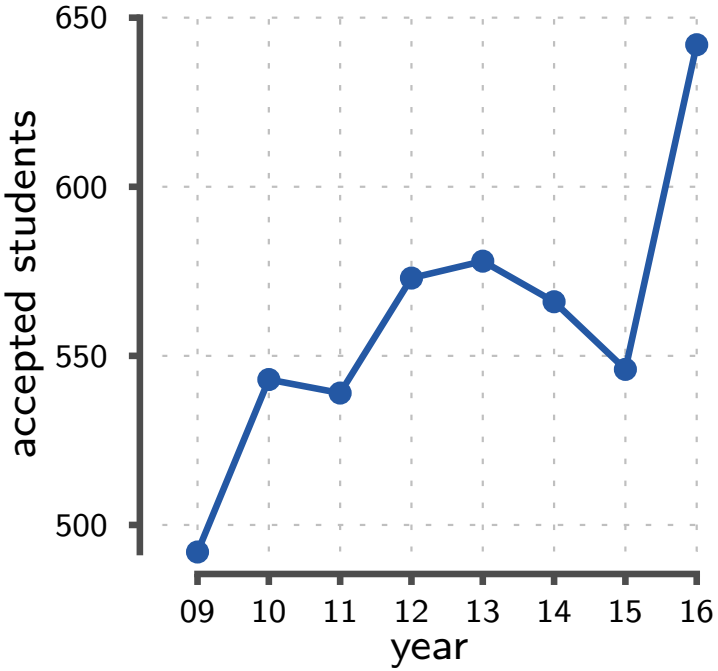
- Show several graphics side-by-side
  - one graphic shows the overview of the whole data
  - other graphics show details / zoom-ins



Multiform

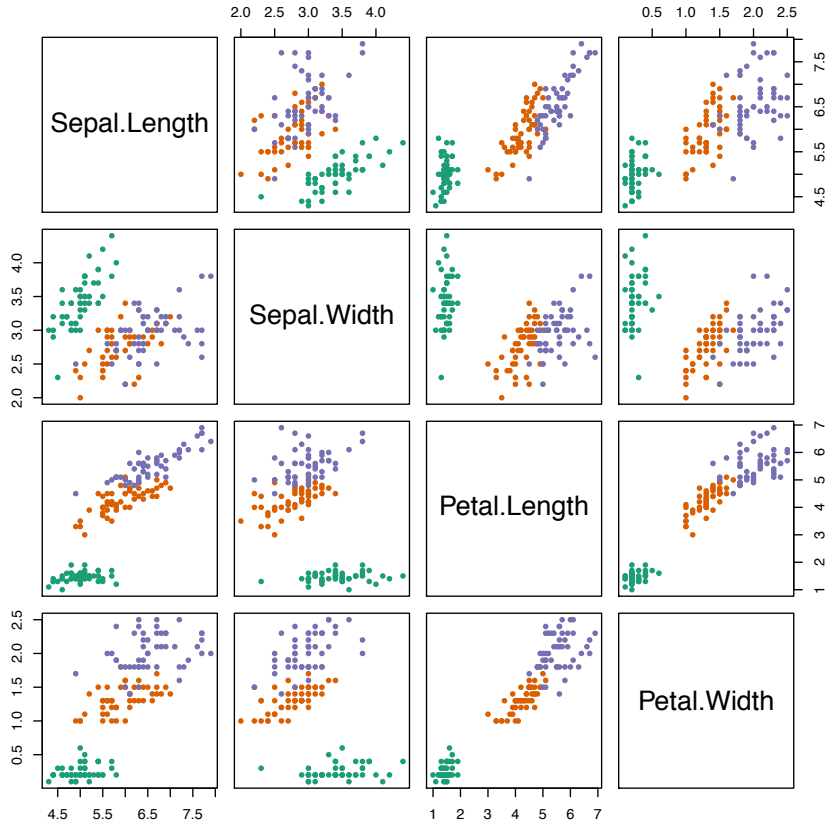
- Show the same data with different designs
  - different designs are more helpful to tell different parts of the story
  - for example, bar charts help you to compare individual values while line charts reveal trends
  - introduces redundant data-ink so needs to be justified





Small multiples (trellis)

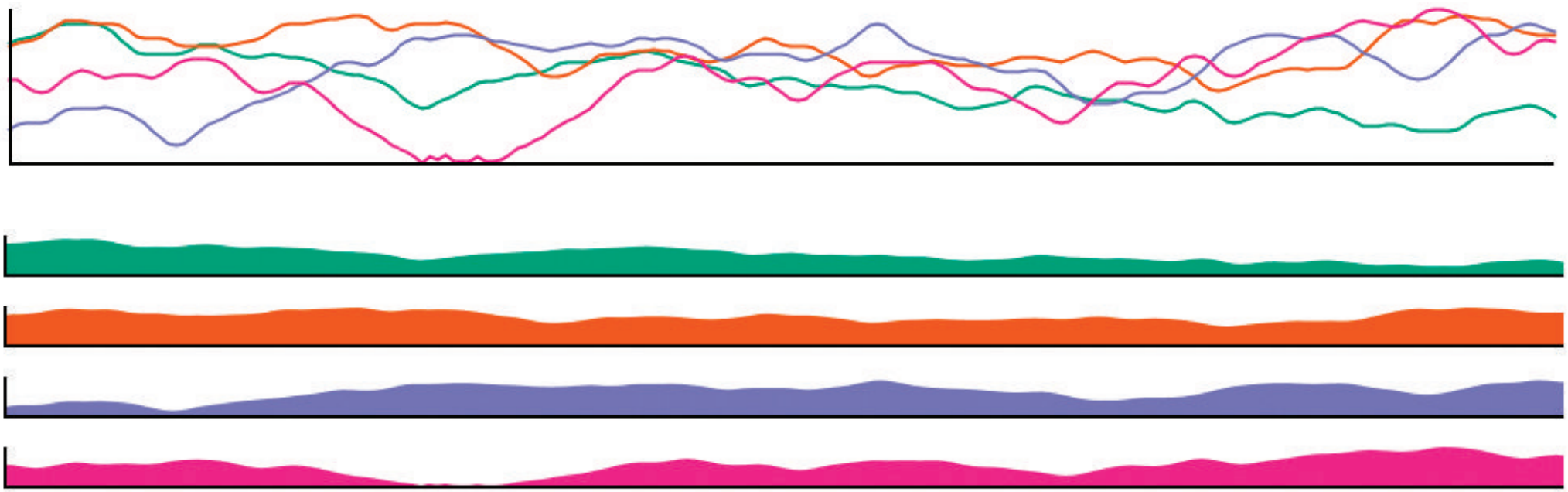
- multiple plots with the same design
- show different parts of the data
- can be arranged into a matrix or other array




```
pairs(iris[,1:4],col=brewer.pal(3,"Dark2")[iris[,5]],pch=20)
```

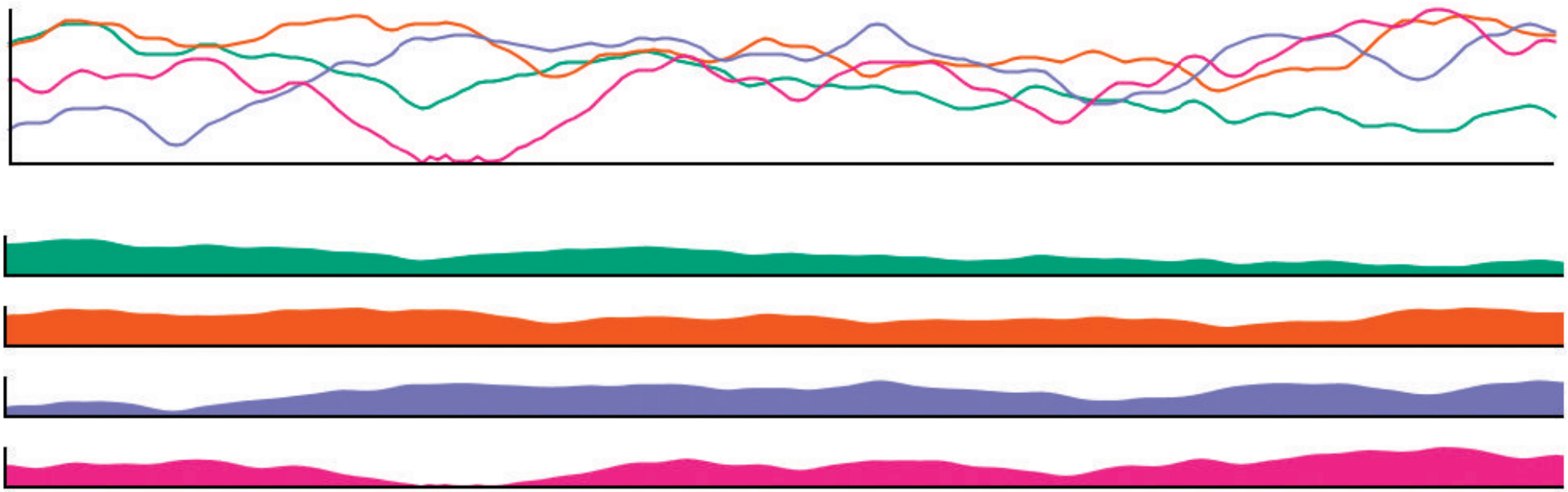
Superimposition vs. small multiples

- superimposition is not possible for large amount of series
  - too cluttered / run out of colors
- small multiples require more space
  - squished y-axis



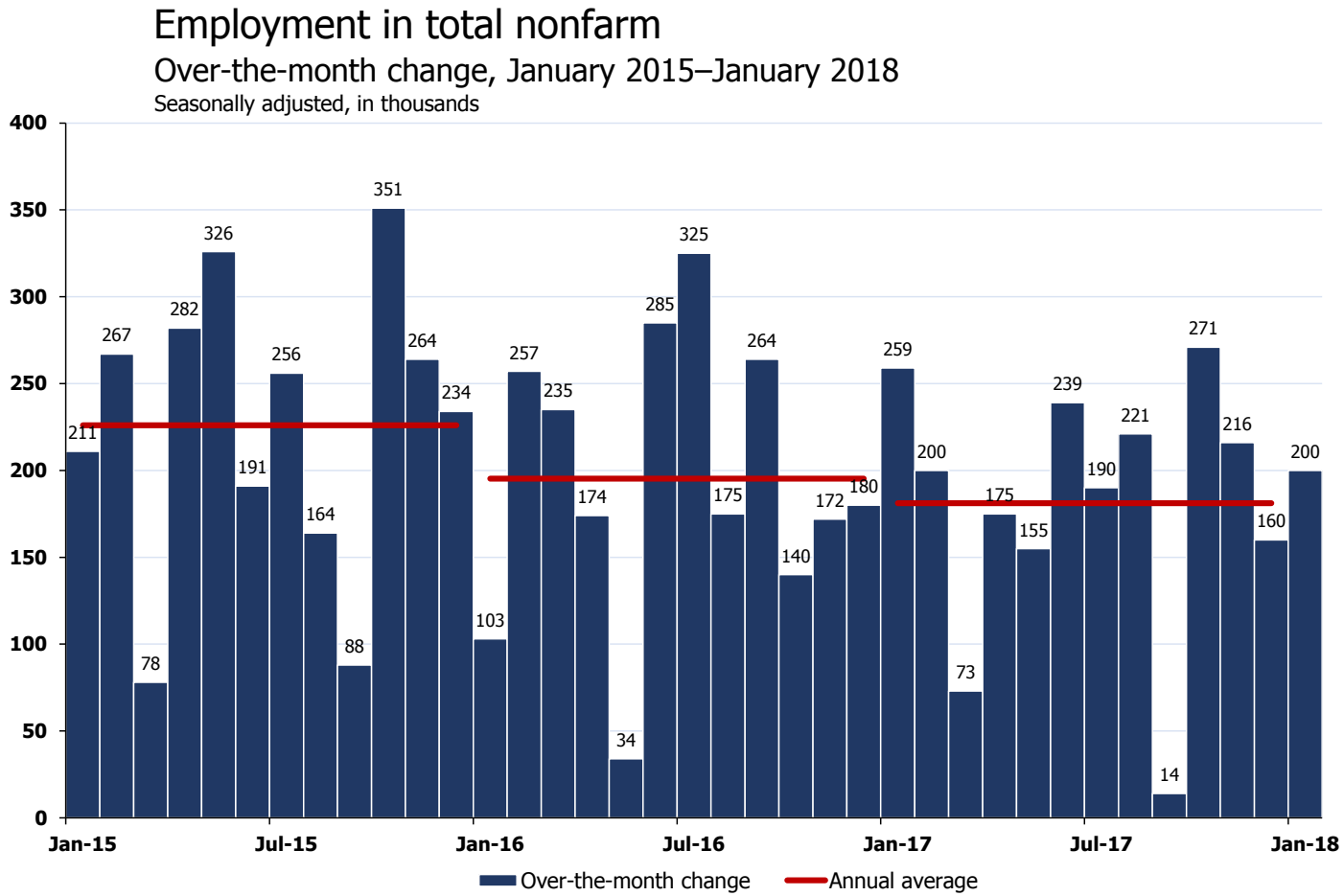
Superimposition vs. small multiples

- (local) find time series with the highest point at one time point
- (global) which of the time series has the highest slope?
- Javed et al. (2010) showed that superimposition is better for local, small multiples is better for global.



Reference lines

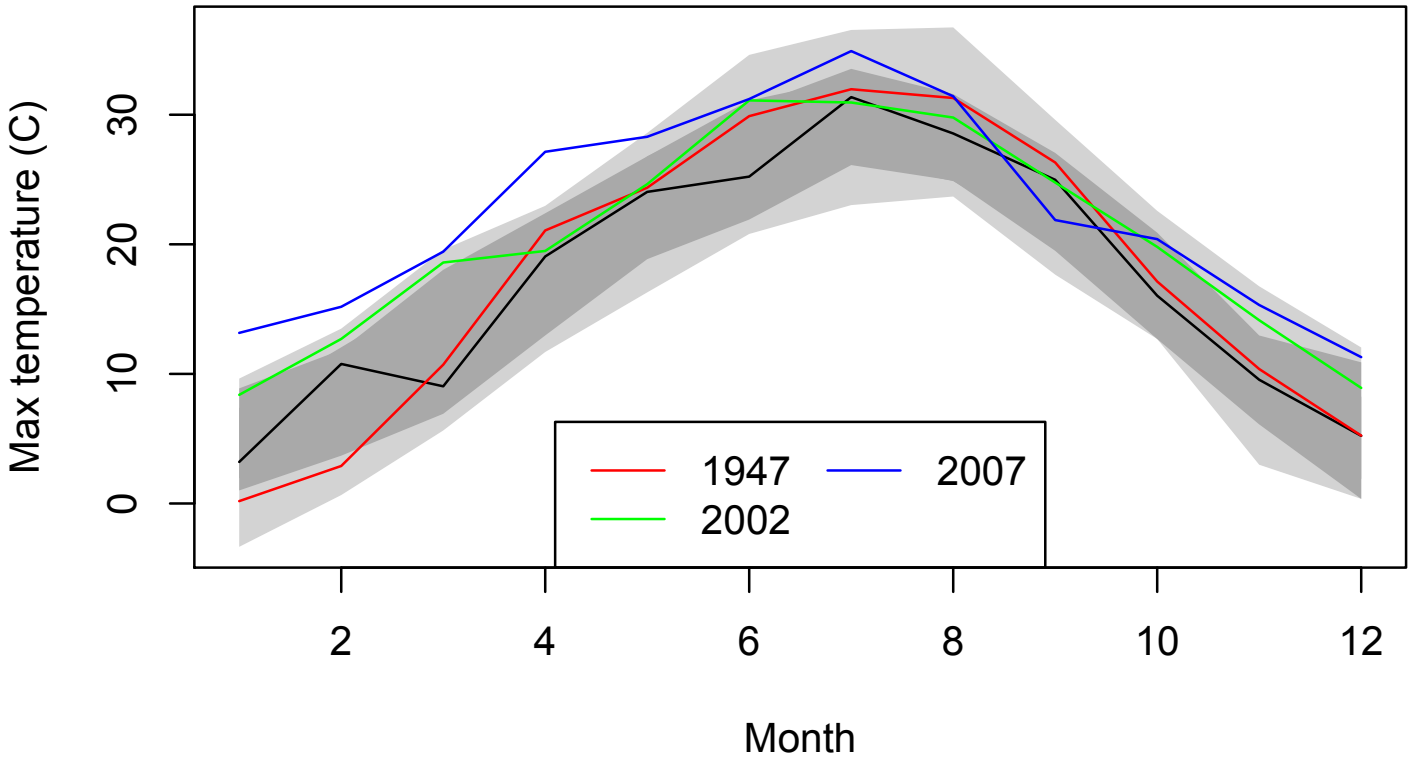
- Adding a reference line can be extremely helpful for telling the story



Bureau of Labor Statistics, Current Employment Statistics, February 02, 2018.
Most recent 2 months of data are preliminary.

Reference region

- Similarly, reference regions can be helpful, for example,
  - showing the acceptable values, and
  - highlighting abnormal behaviour.



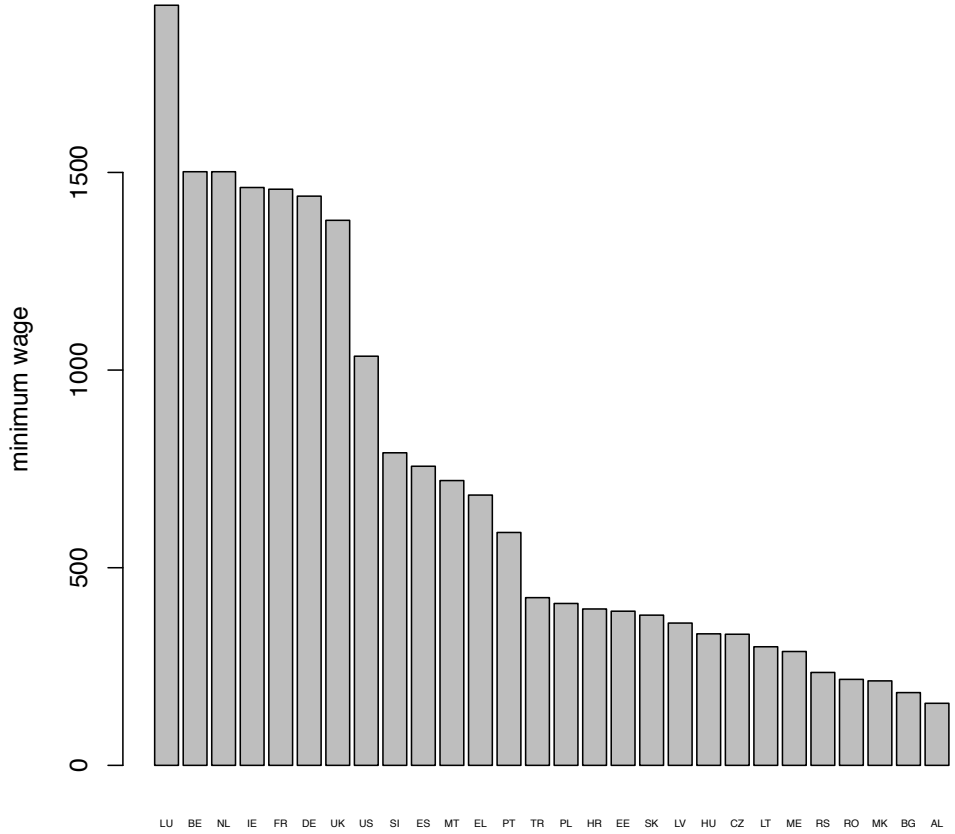
J. Korpela (2018), http://urn.fi/URN:ISBN:978-952-60-7855-7

Re-expression

- What is important:
  - relative performance to a baseline?
  - difference to a baseline?
  - the most relevant scale?

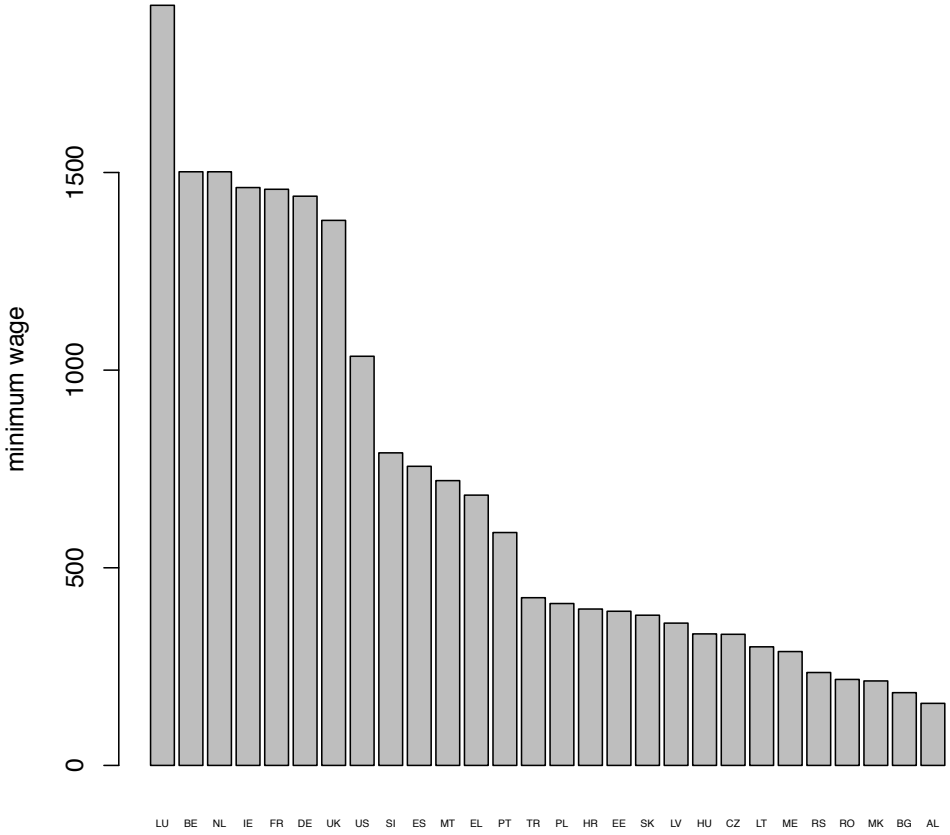
Re-expression

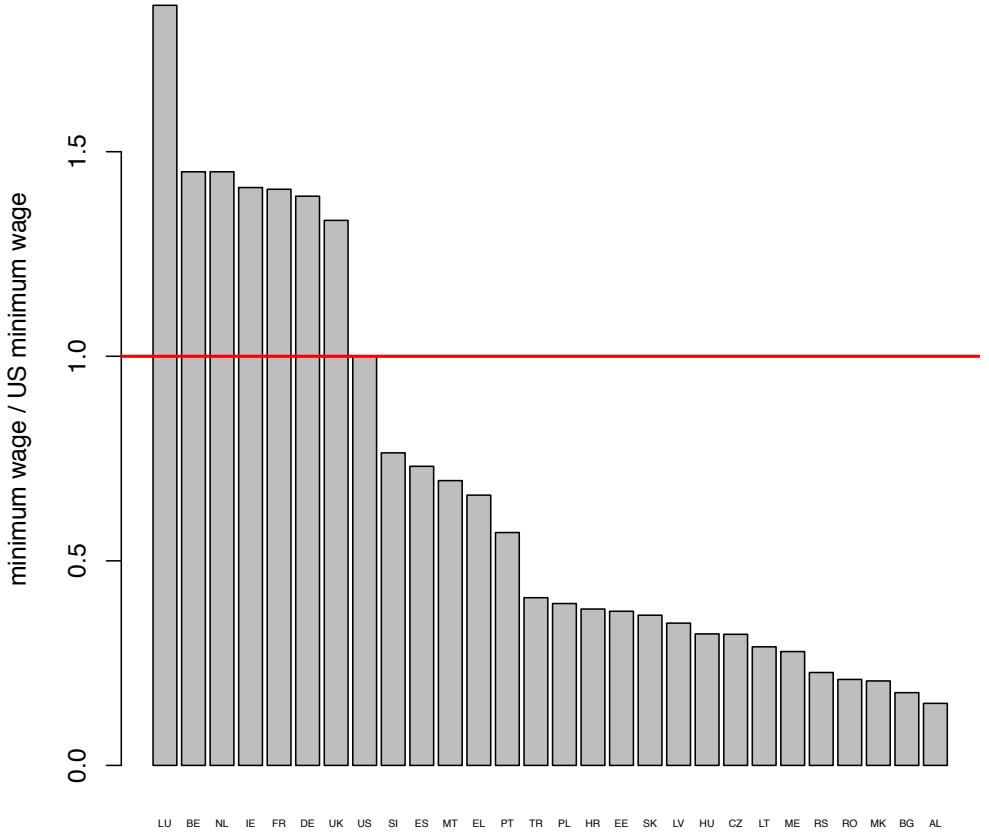
- minimum wages in euros in 2015
- http://ec.europa.eu/eurostat/web/labour-market/earnings/main-tables



Re-expression

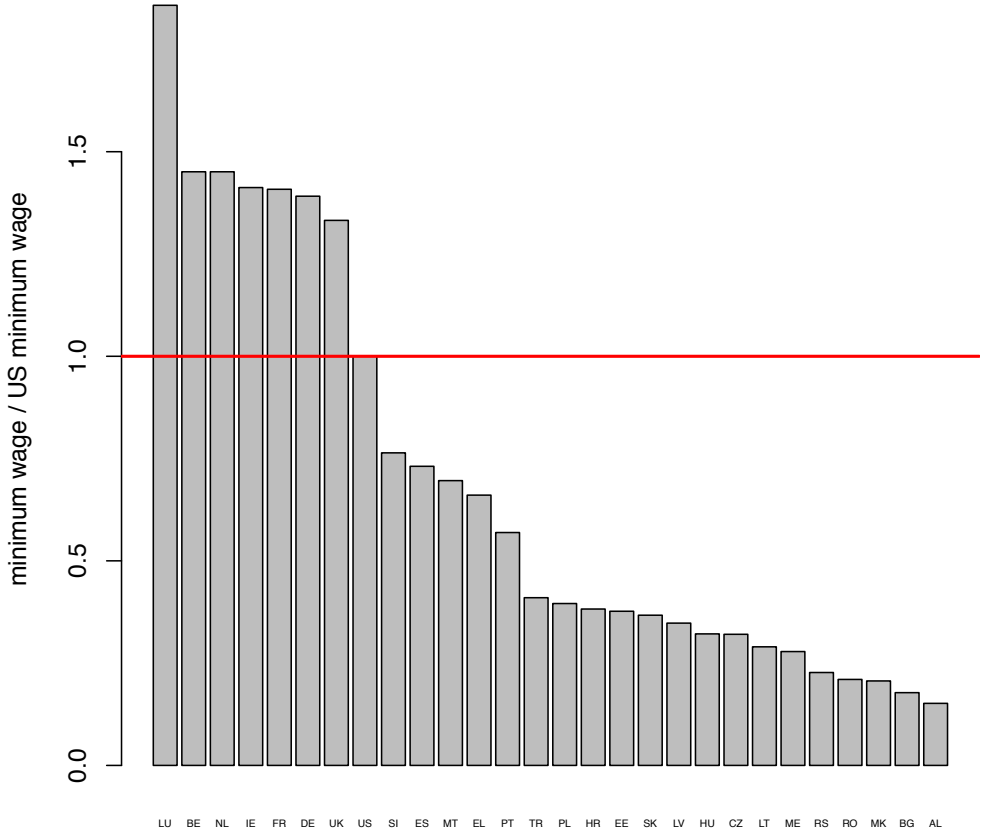
- relative performance to a baseline?
  - if performance **relative** (proportional) to a baseline b is important, then scale the y-axis by dividing with b.
  - indicate the scaling in labels and in caption

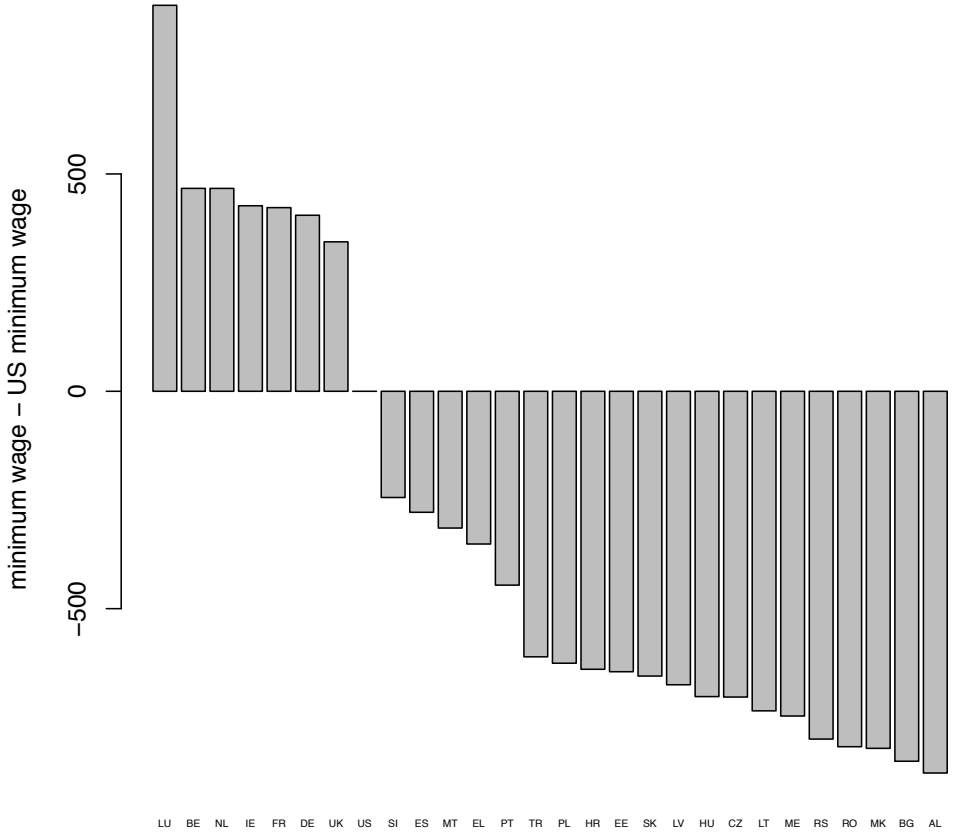




Re-expression

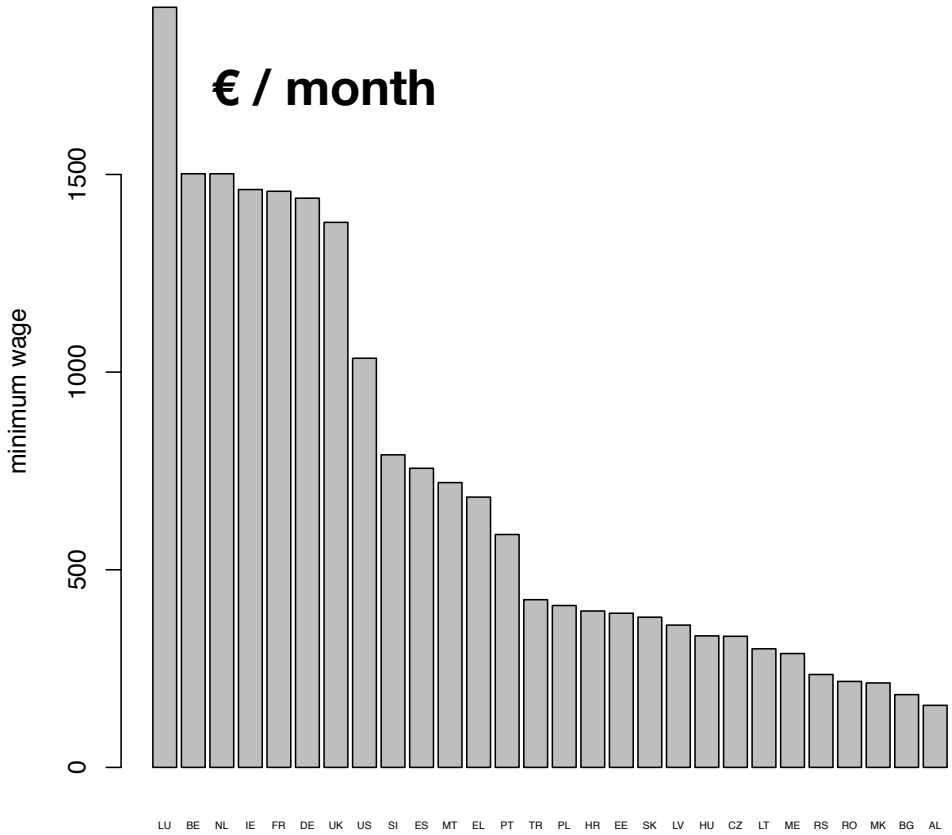
- difference to a baseline?
  - if the (absolute) **difference** to a baseline b is important, then scale the y-axis by subtracting b.
  - indicate the scaling in labels and in caption

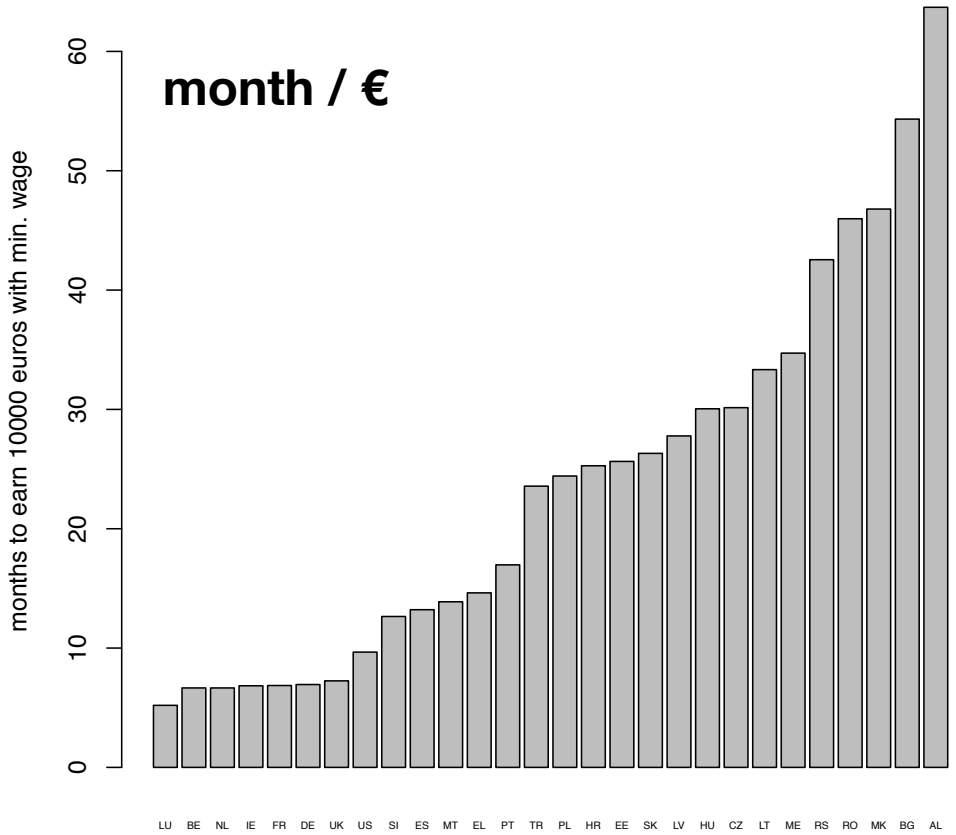




Re-expression

- The most relavant scale?
  - log scale: if relative change (in %) is important
  - inverted scale: plot 1/y instead of y works sometimes

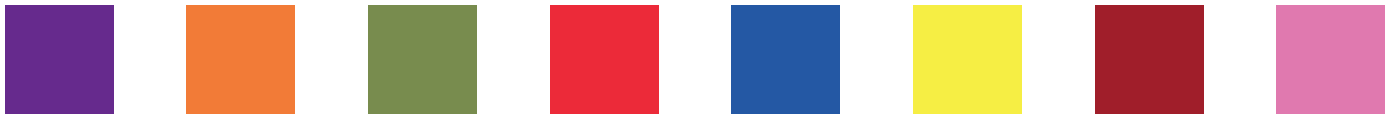




Colors

[more on colors in Part II of the course]

- General guidelines regarding colors
  - use muted colors when applying to surfaces
  - use bright colors for small objects (points)
  - most common colour blindness is red-and-green
  - sometimes it is a good idea to vary luminance as well as colors



Colors

- General guidelines regarding colors
  - in heat-maps using a gray scale is typically not a good idea

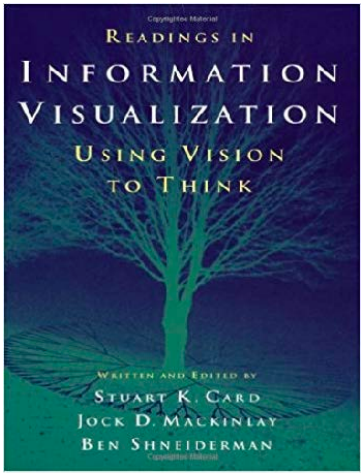
- instead vary from one colour to the ’opponent’ color
- if heat-map has a neutral value
  - vary from one colour to a neutral gray color and
  - continue to the opponent colour

- See http://colorbrewer2.org/

Outline

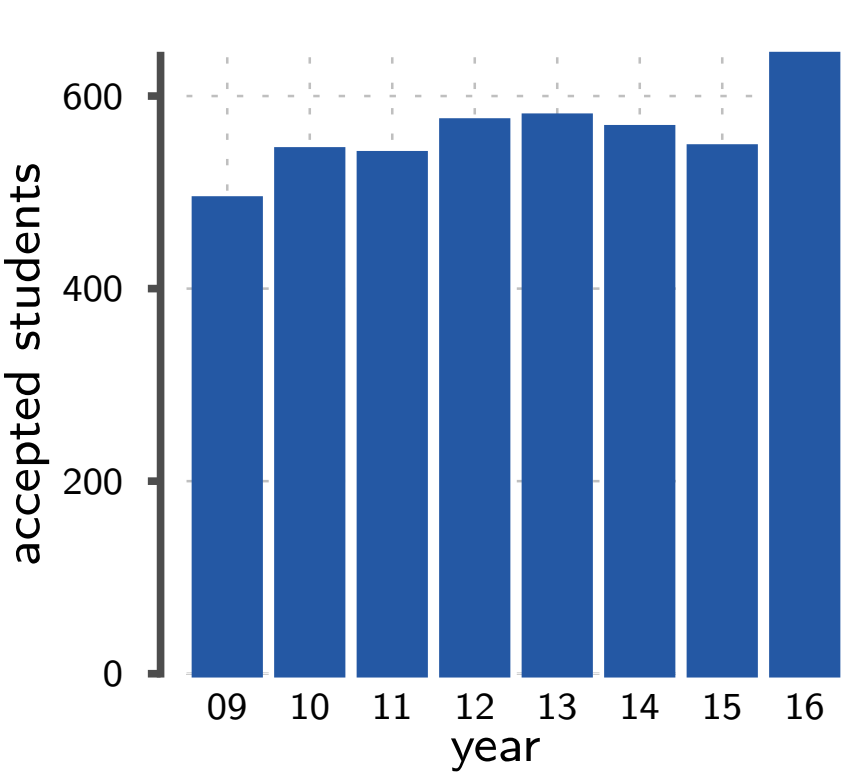
- **Techniques:**
  - Bars, boxes, lines, dots
  - multiple plots
  - reference lines and regions
  - rescaling /normalising / re-expressing
  - colours
- **Problems:**
  - axis ranges
  - use of 3D
  - overplotting
- **Scenarios:**
  - distribution analysis
  - ranking and part-of-whole analysis
  - time-series
  - high-dimensional data

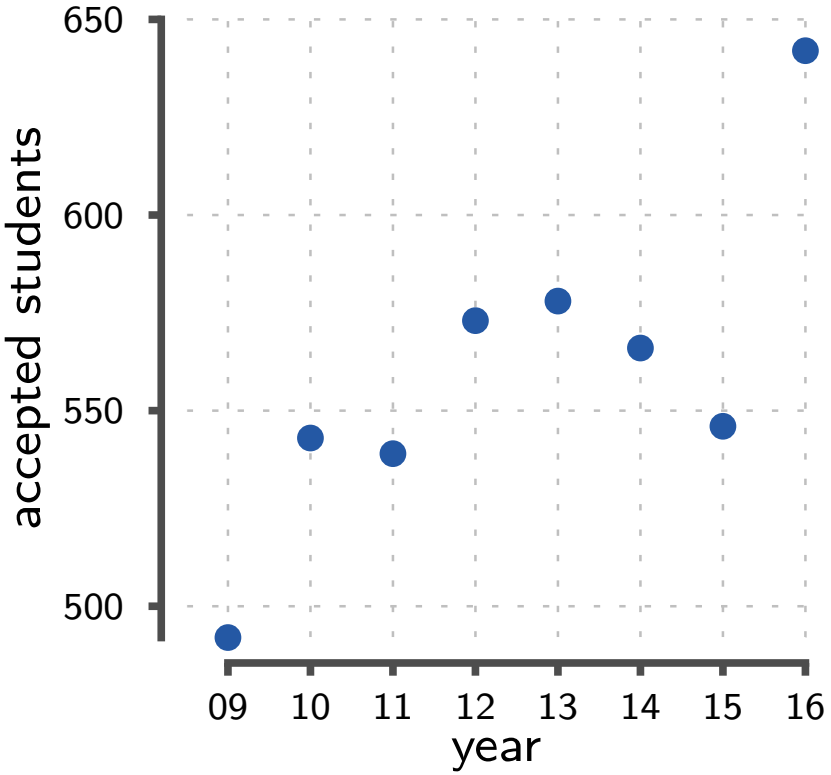
- **Related reading:** *Few. Now you see it. Analytic Press, 2009.*
- **Older but relevant:** *Card et al. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, 1999.*



Compressed bar plots

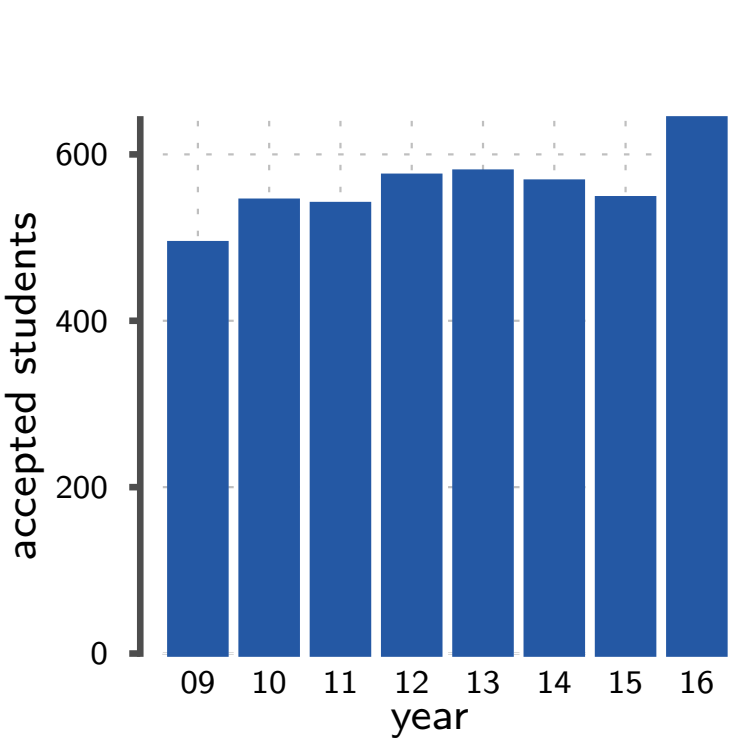
- You should probably not change the baseline for bar plots
  - change the format: use a dot plot
  - consider also a line plot if it makes sense
    - this allows to rescale the y-axis

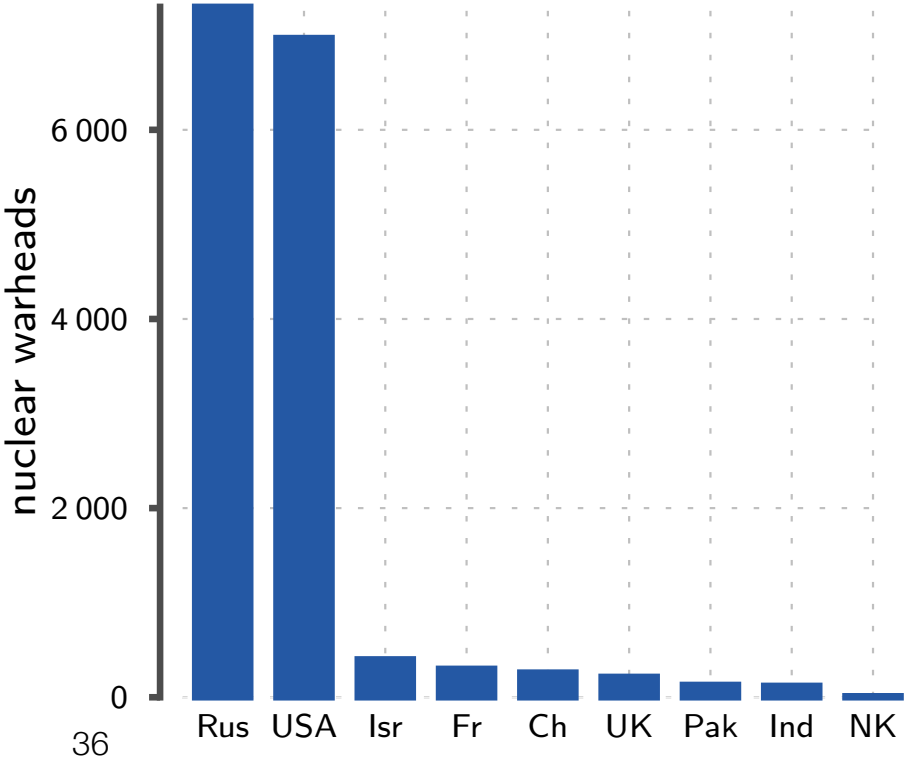




Axis ranges

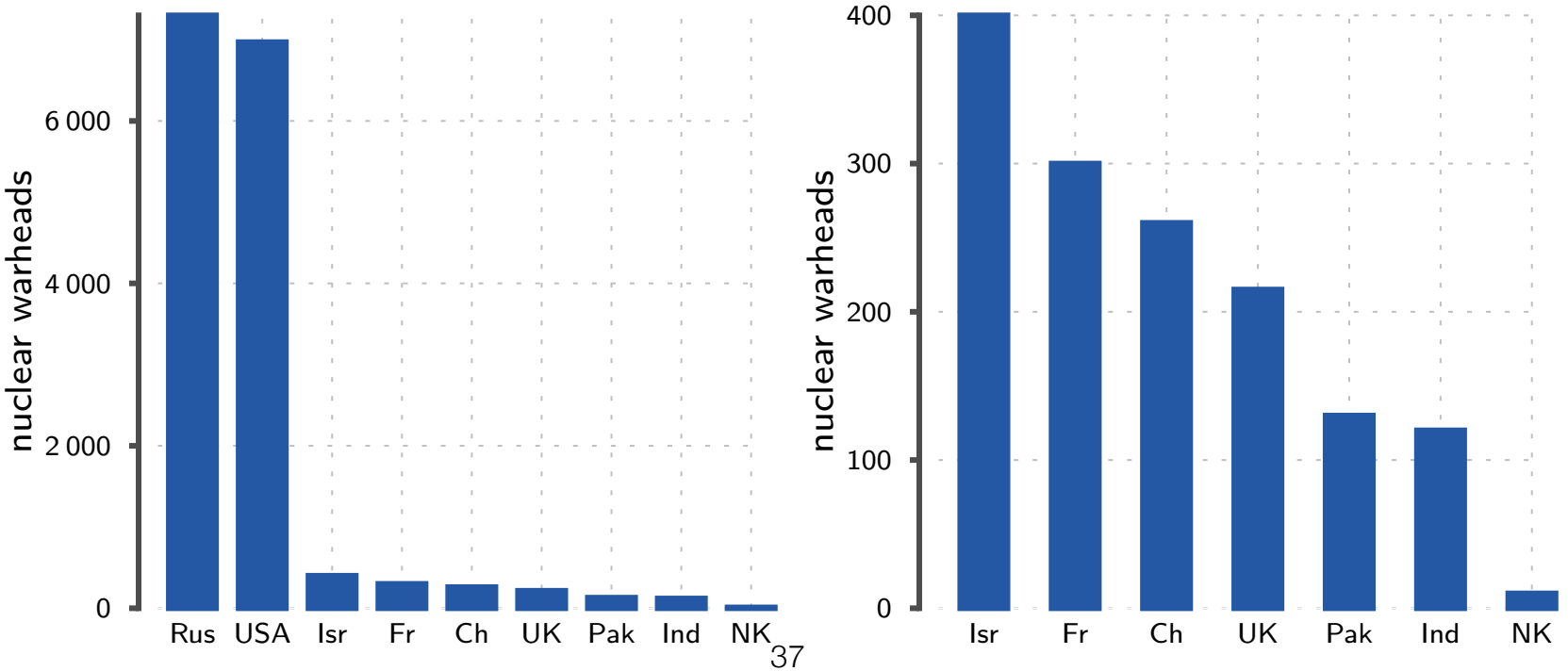
- Common problem with plots are compressed axis ranges:
  - zero baseline with bar plots causes compressed variation
  - large variation in data causes small values to be unreadable





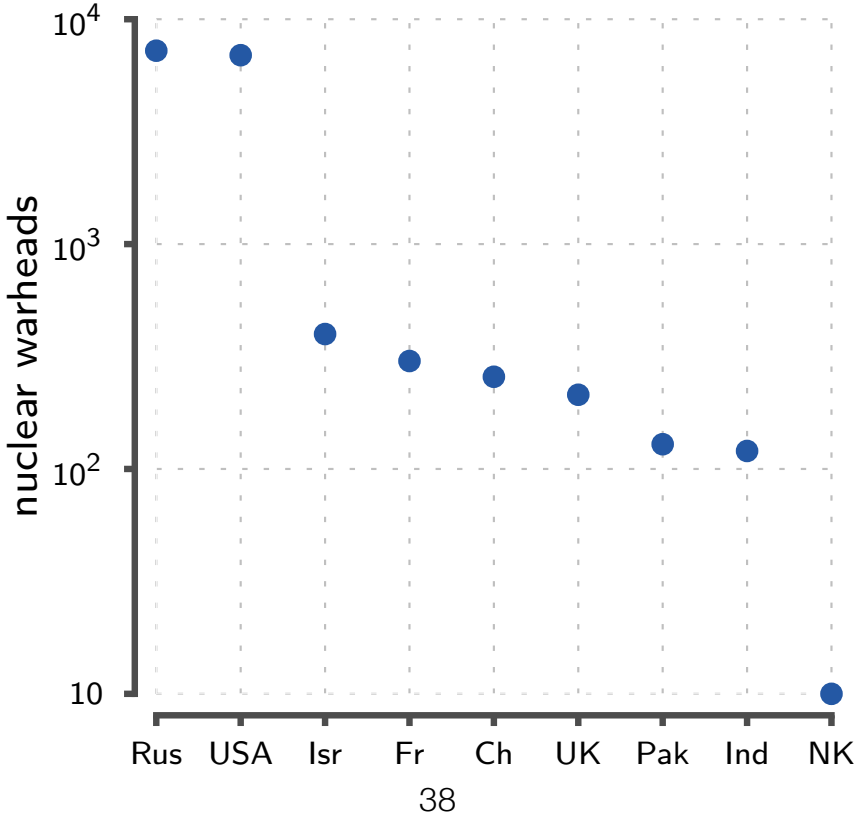
Axis ranges

- Solving problems associated with large variation
  - create two plots: first showing the overview, second showing the small values.
  - if possible, you can also group related variables in the overview



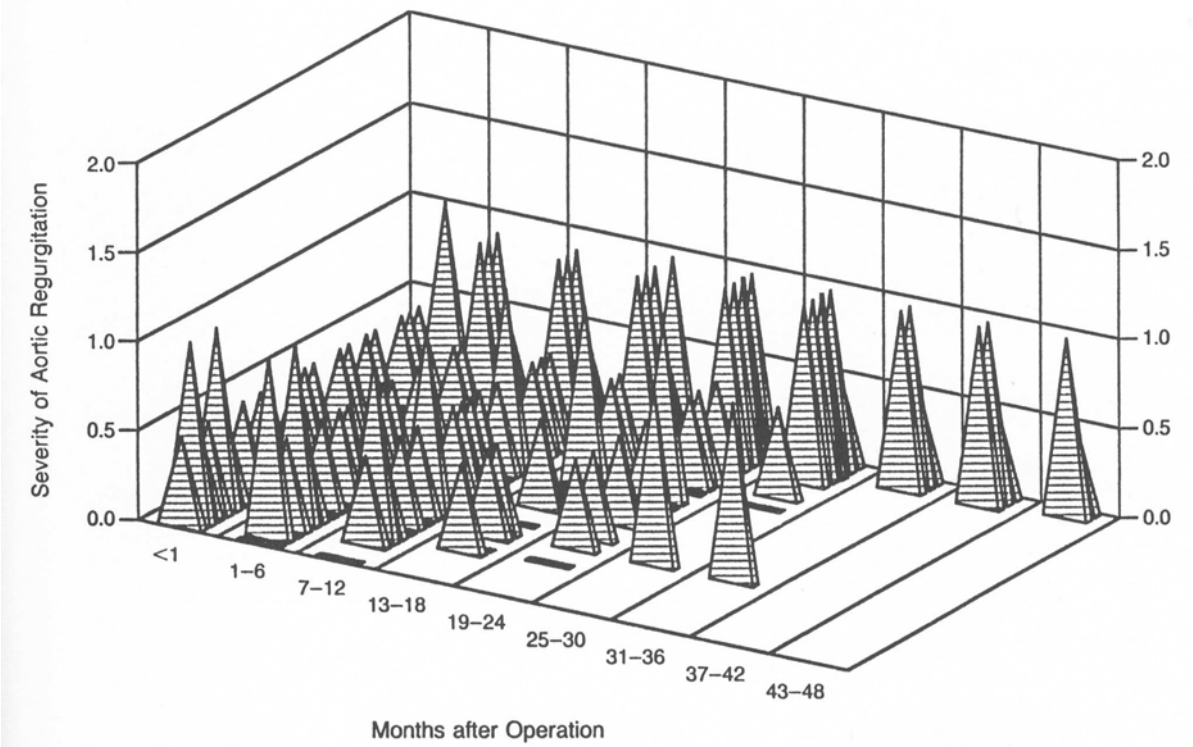
Axis ranges

- Solving problems with large variation
  - the other approach is to re-express data:
    - switch to log-scale



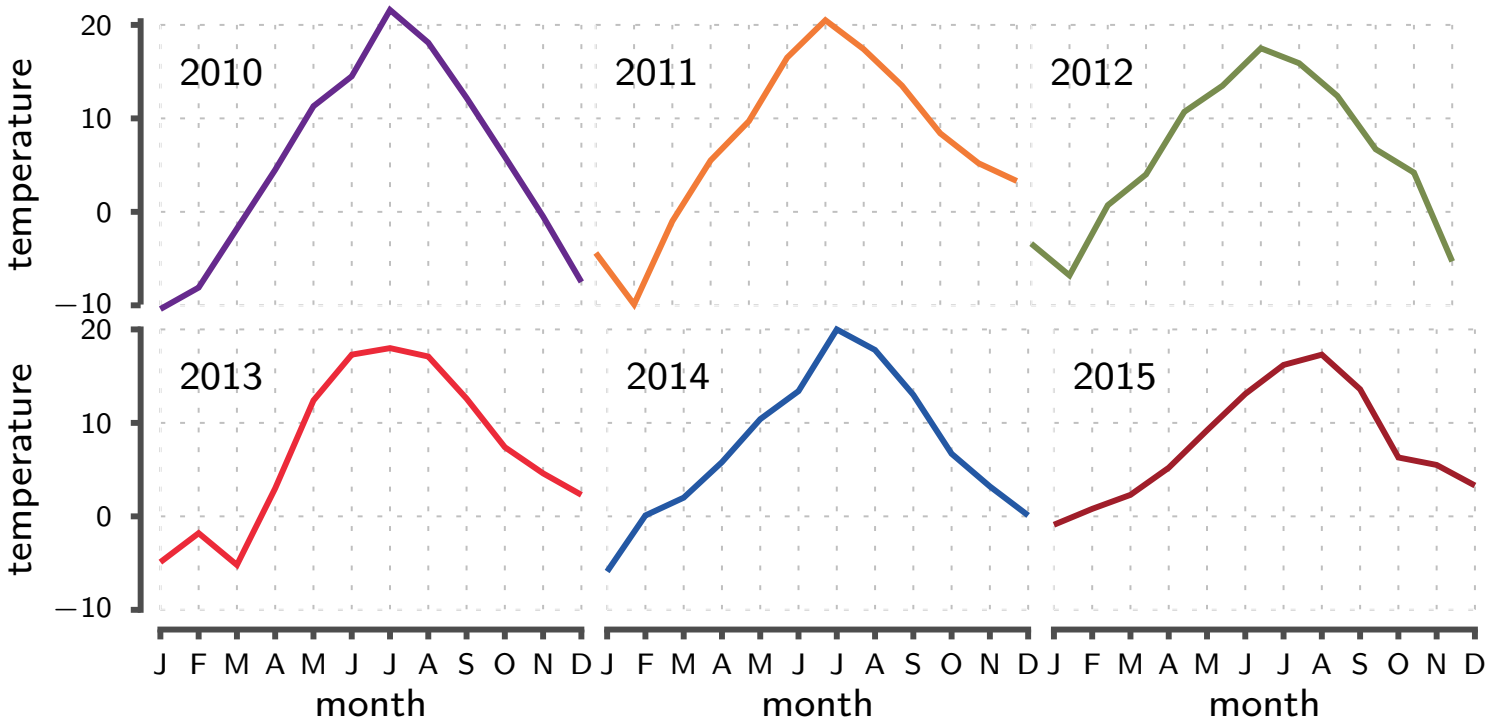
No unjustified 3d

- 3D graphics are highly problematic
  - hard to read
  - occlusion
  - perspective error
  - necker illusion
  - would often need interaction to work



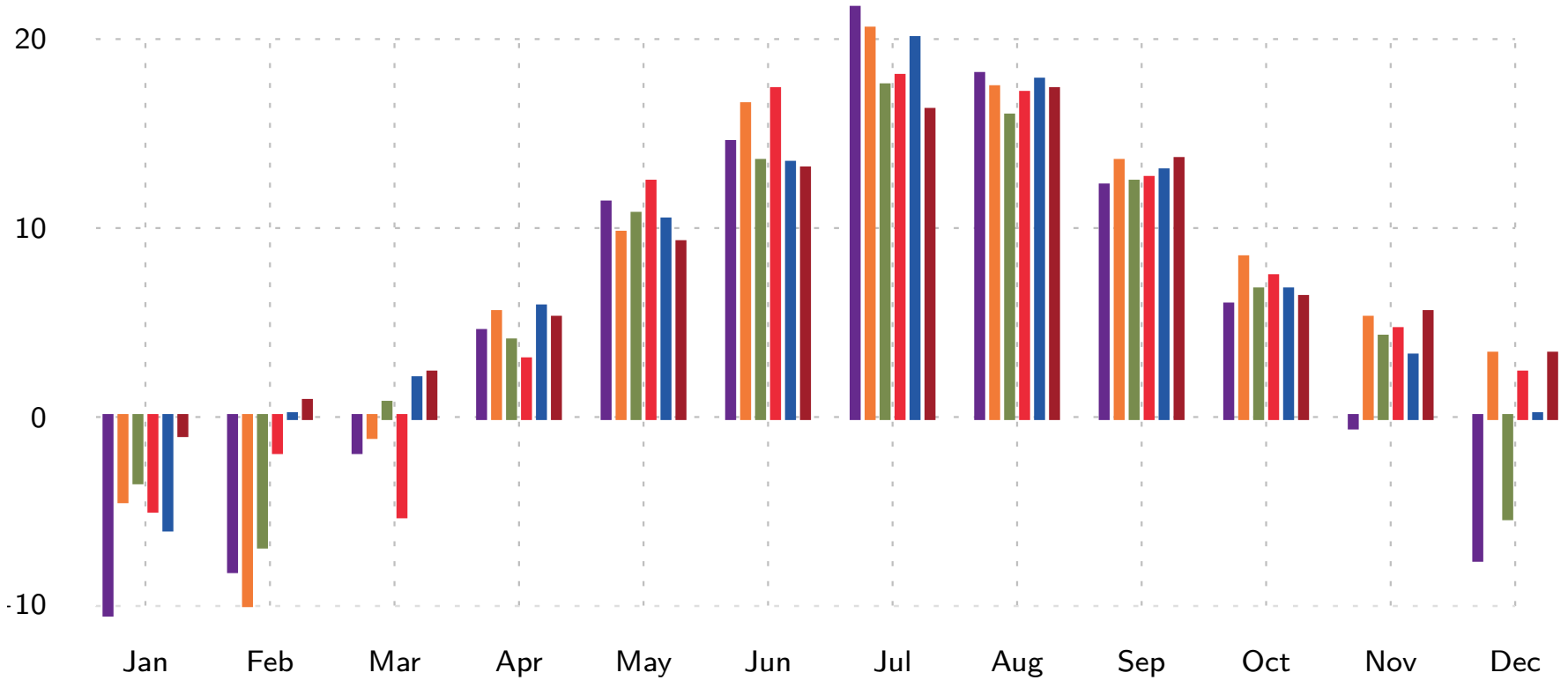
No unjustified 3d

- Ways to go around
  - if one axis is integer or categorical, then you can split the figure in small multiples
    - takes more space but is significantly more readable



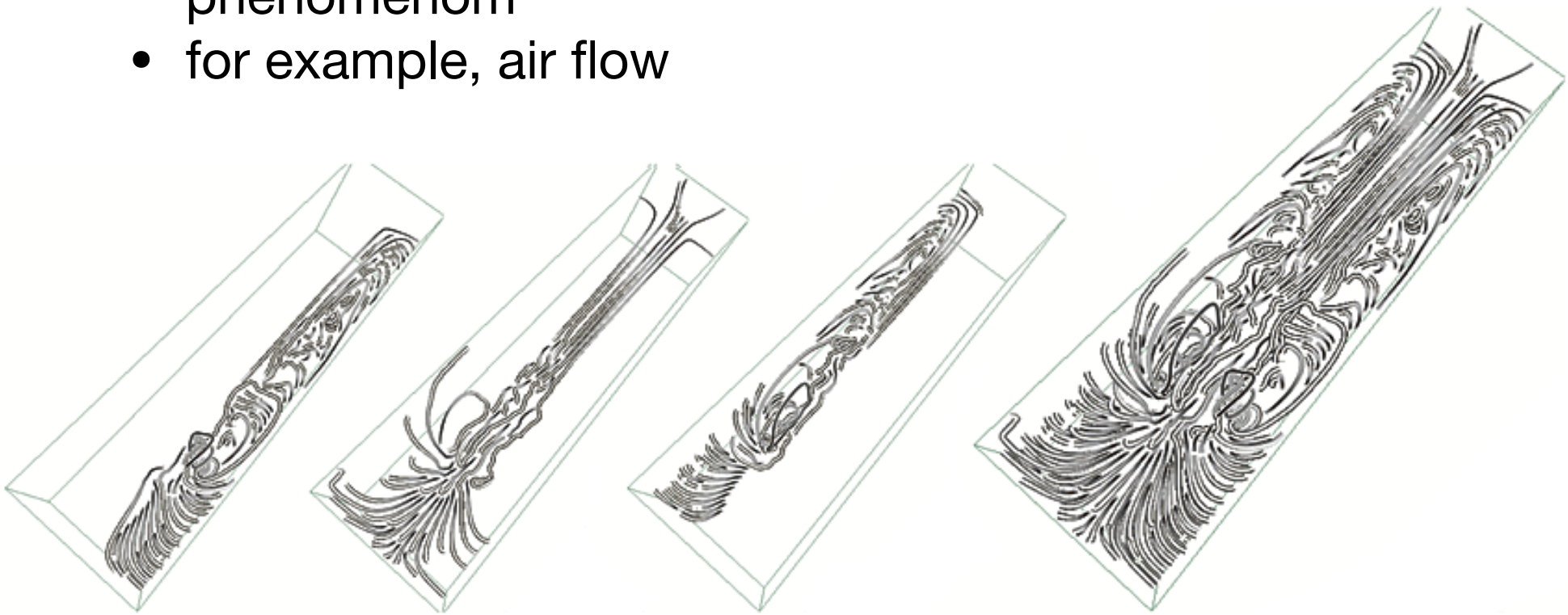
No unjustified 3d

- Ways to go around
  - if one axis is integer or categorical, you can also try superimposition (may result in clutter) or
  - grouped bar chart / dot plot



No unjustified 3d

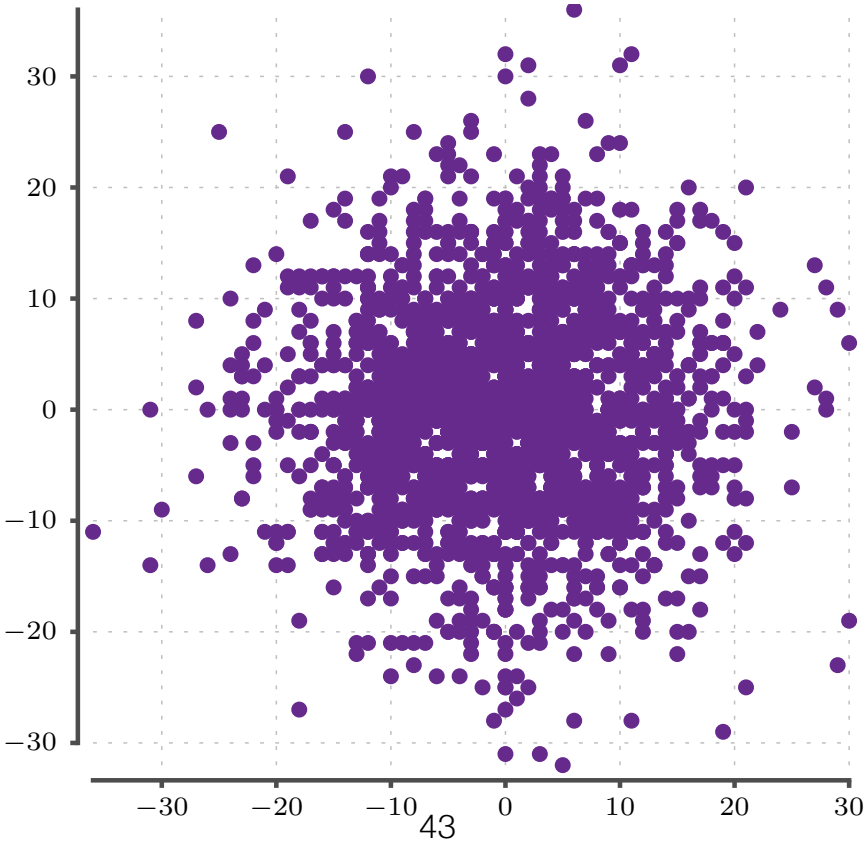
- Sometimes 3d works
  - if you are studying the shape of a physical 3d phenomenom
  - for example, air flow



Li & Shen (2007), https://doi.org/10.1109/TVCG.2007.1009

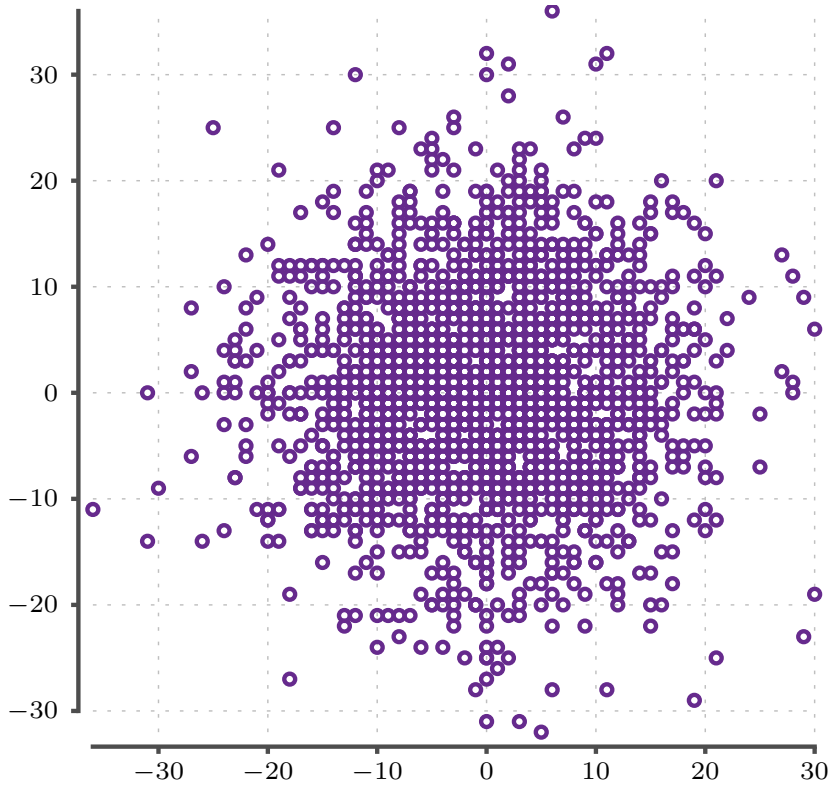
Dealing with overplotting in scatterplot

- Large datasets create cluttered scatterplots
  - individual points are hidden
  - density is not seen



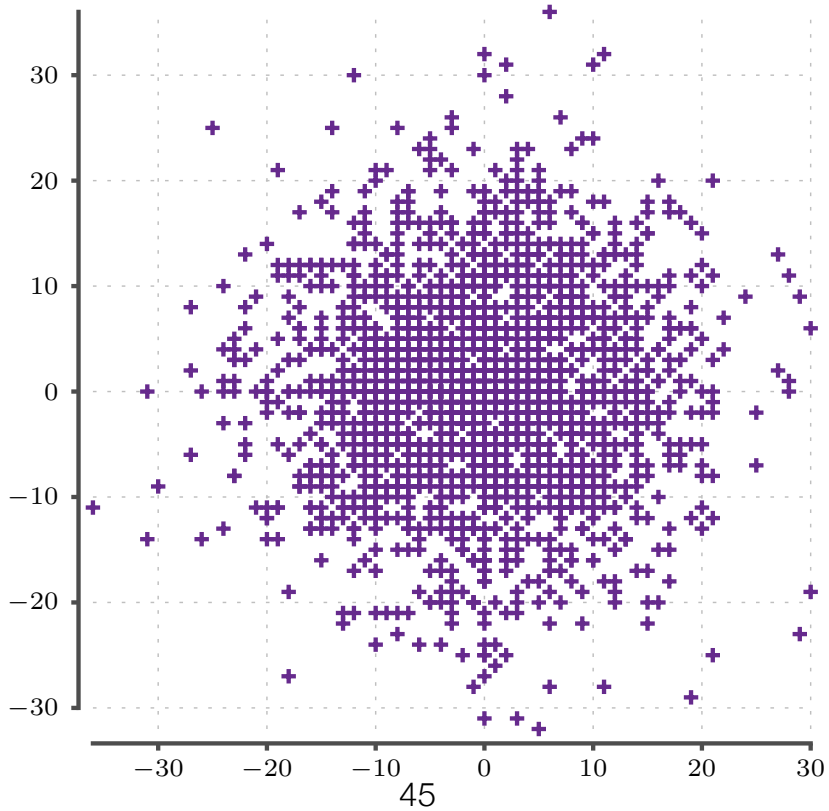
Dealing with overplotting in scatterplot

- Tricks to reduce clutter:
  - remove the **fill** of the markers



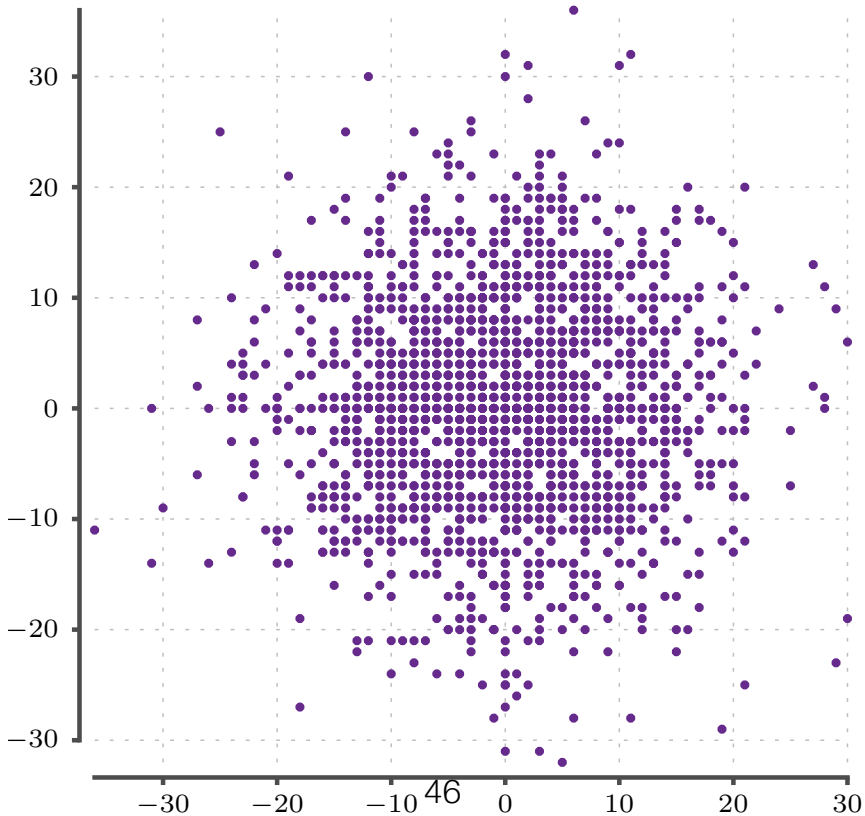
Dealing with overplotting in scatterplot

- Tricks to reduce clutter:
  - change the **shape** of the markers



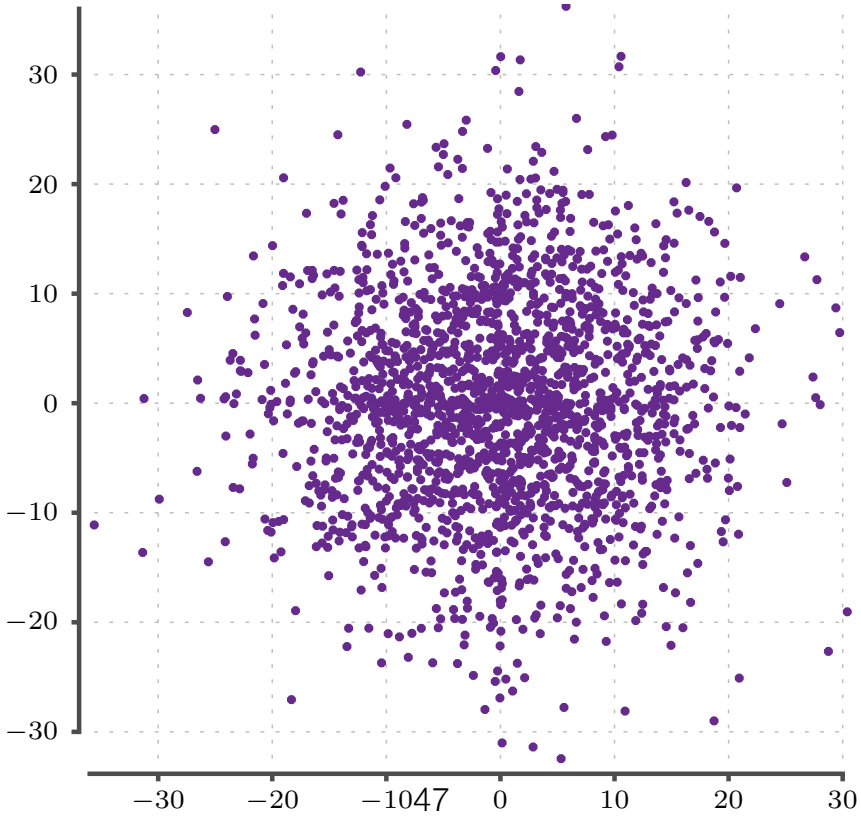
Dealing with overplotting in scatterplot

- Tricks to reduce clutter:
  - reduce the **size** of the markers



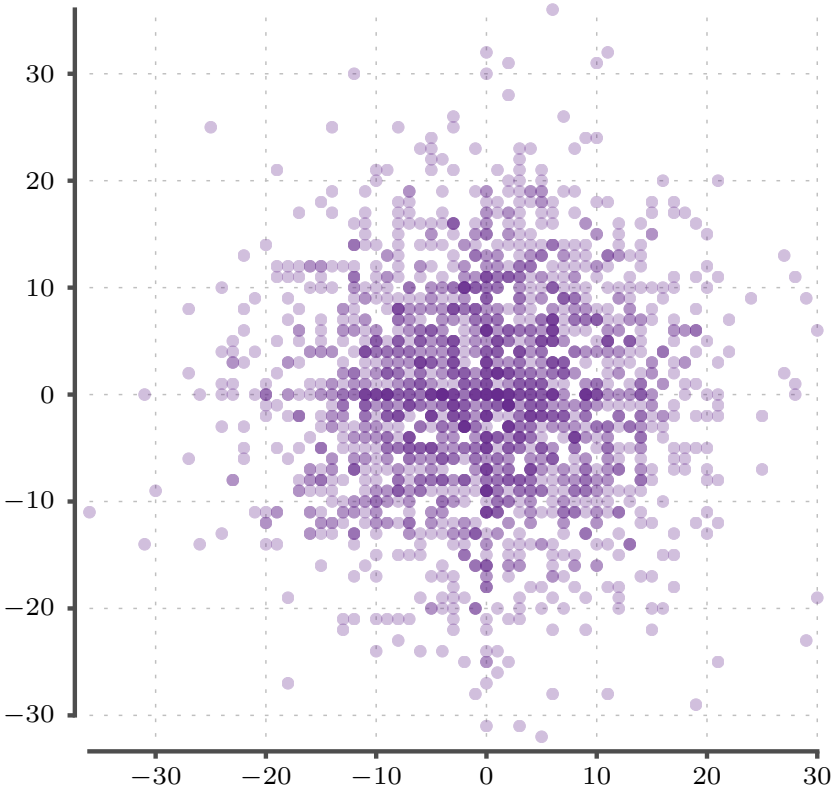
Dealing with overplotting in scatterplot

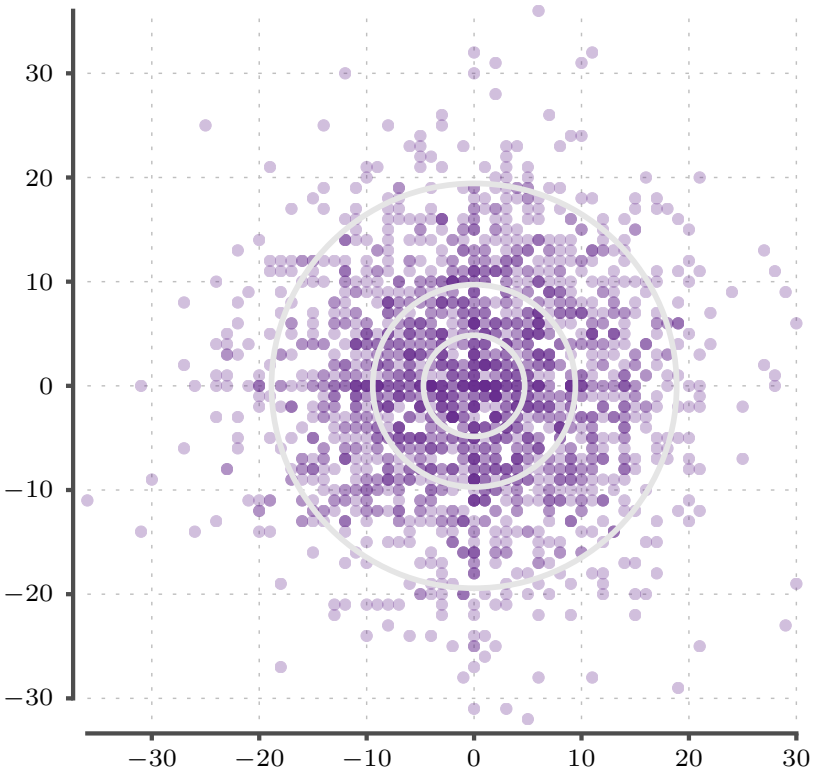
- Tricks to reduce clutter:
  - problem occurs if many points share the same value
  - **jitter** the points around with random noise
    (... and explain in the caption)



Dealing with overplotting in scatterplot

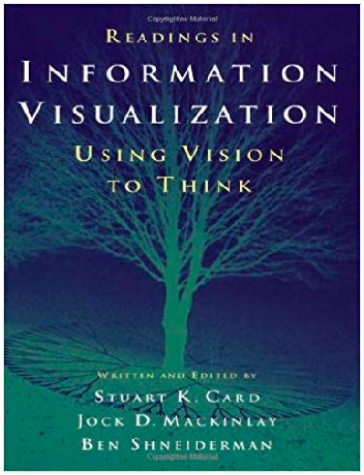
- Tricks to reduce clutter:
  - make your markers **transparent**
  - consider **contour lines** or switching, e.g., to heat map
  - **downsample** data (it probably makes no sense to show million points anyway)





Outline

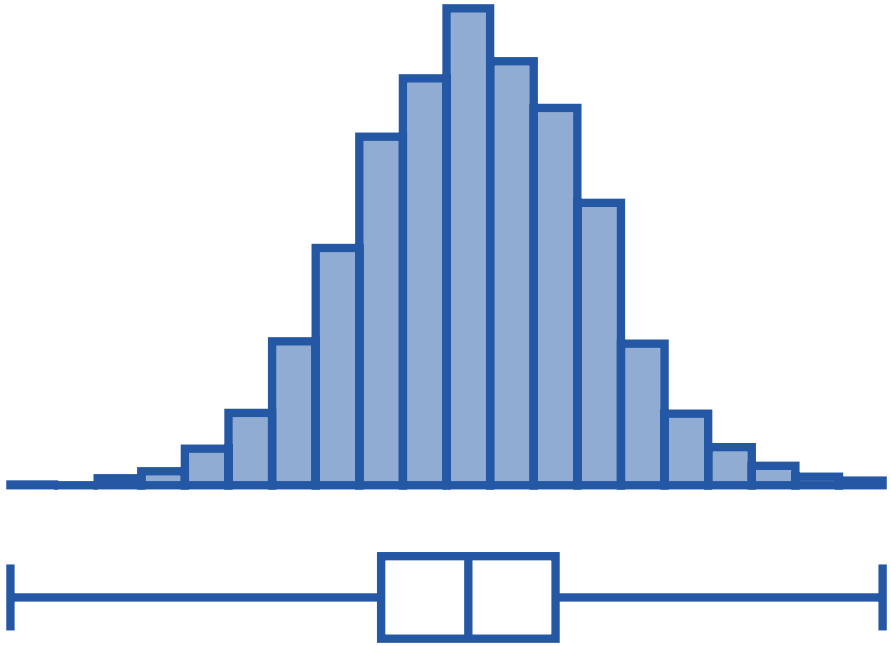
- **Techniques:**
  - Bars, boxes, lines, dots
  - multiple plots
  - reference lines and regions
  - rescaling /normalising / re-expressing
  - colours
- **Problems:**
  - axis ranges
  - use of 3D
  - overplotting
- **Scenarios:**
  - distribution analysis
  - ranking and part-of-whole analysis
  - time-series
  - high-dimensional data



- **Related reading:** *Few. Now you see it. Analytic Press, 2009.*
- **Older but relevant:** *Card et al. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, 1999.*

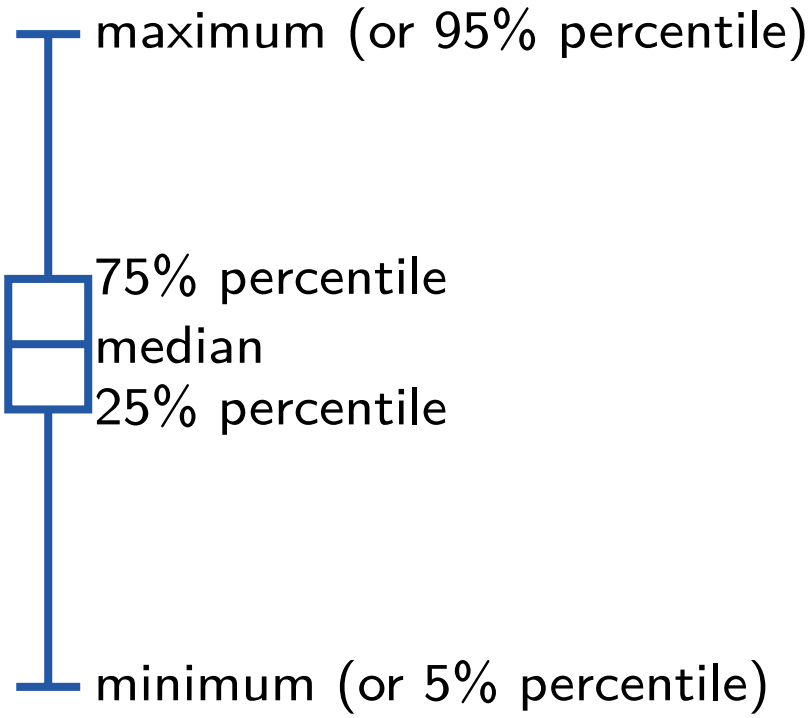
Visualizing distributions

- Two most common techniques to show distributions
  - box plots
  - histograms



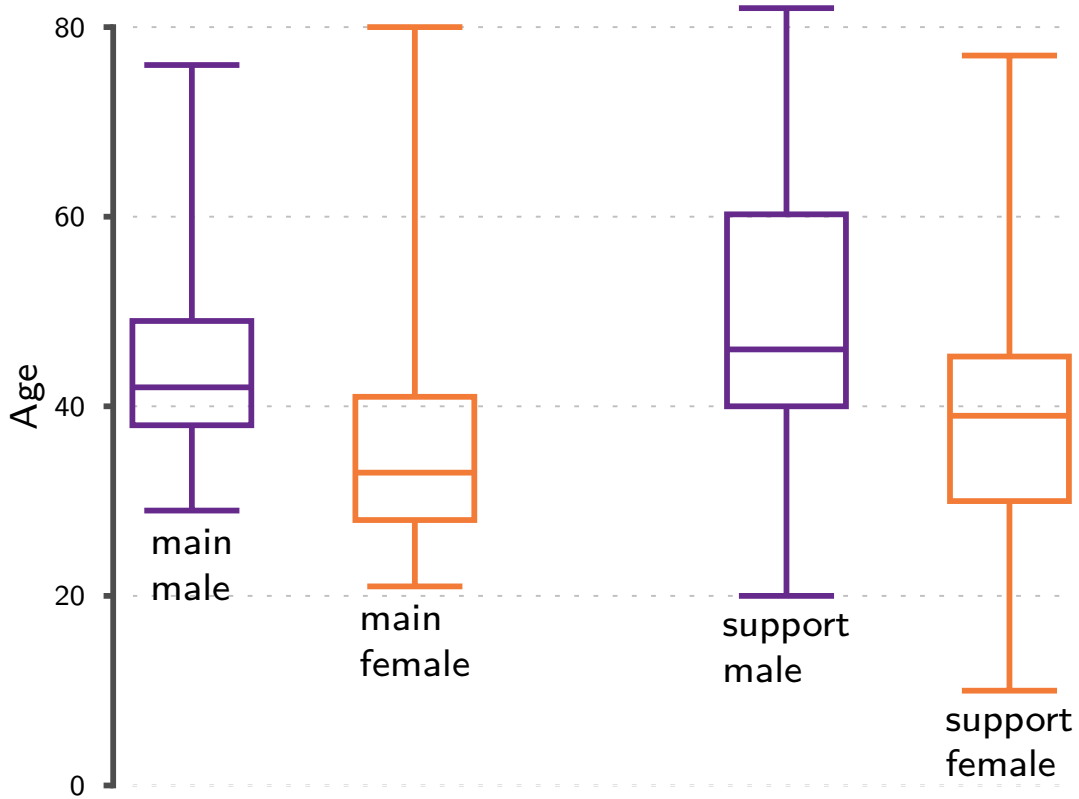
Visualizing distributions

- Box plot summarizes data (= set of numbers) with 5 numbers
  - median, 25%, 75% percentiles
  - minimum and maximum
- additionally it can show outliers



Visualizing distributions

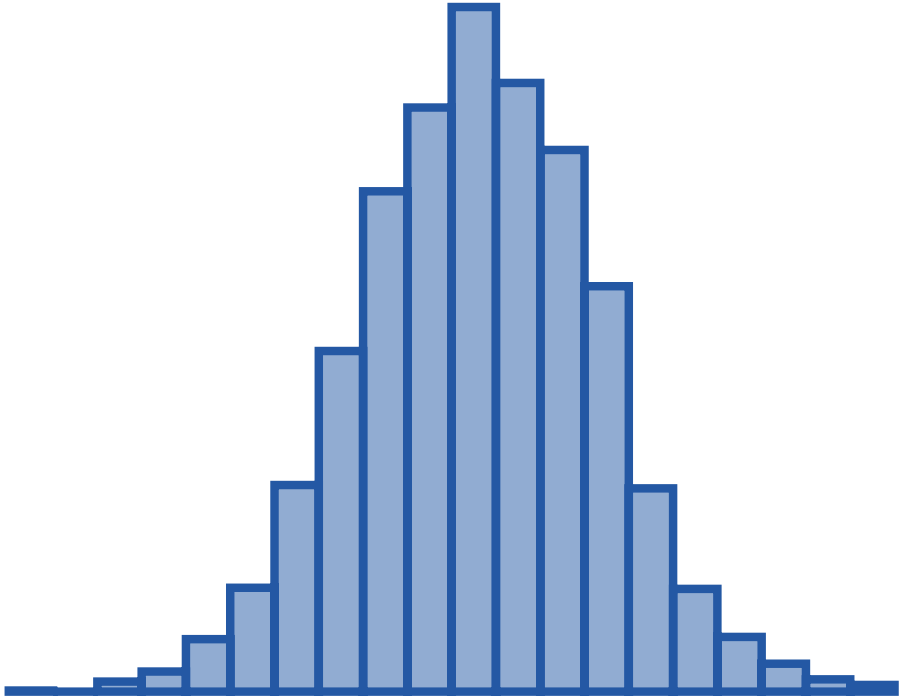
- Box plots are at best when multiple plots are shown at the same time



Oscar winners

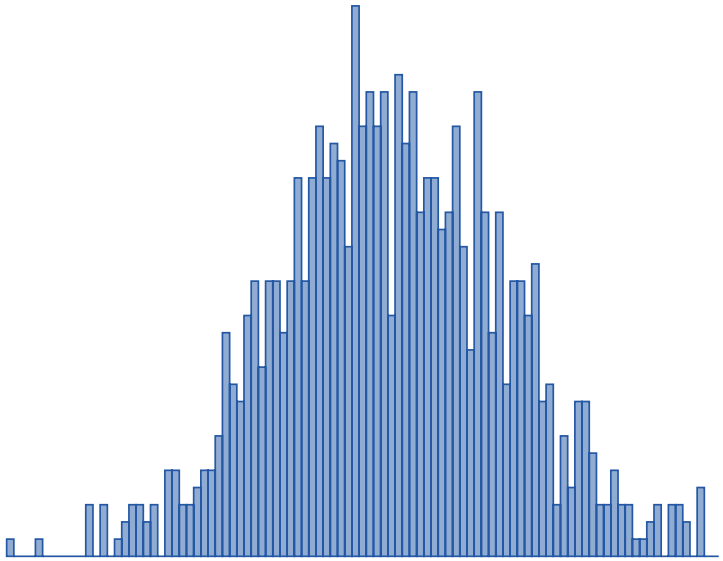
Visualizing distributions

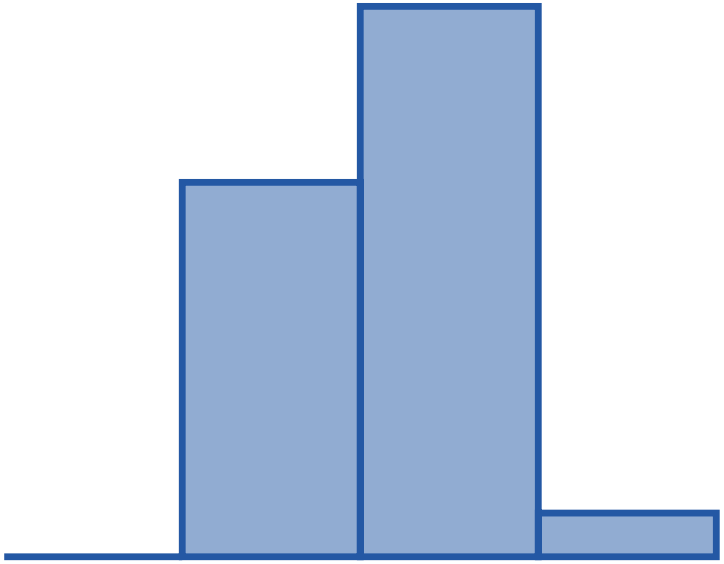
- Histograms show the shape of the distribution
  - x-axis is divided in bins
  - display the numbers of data points that are within a bin



Visualizing distributions

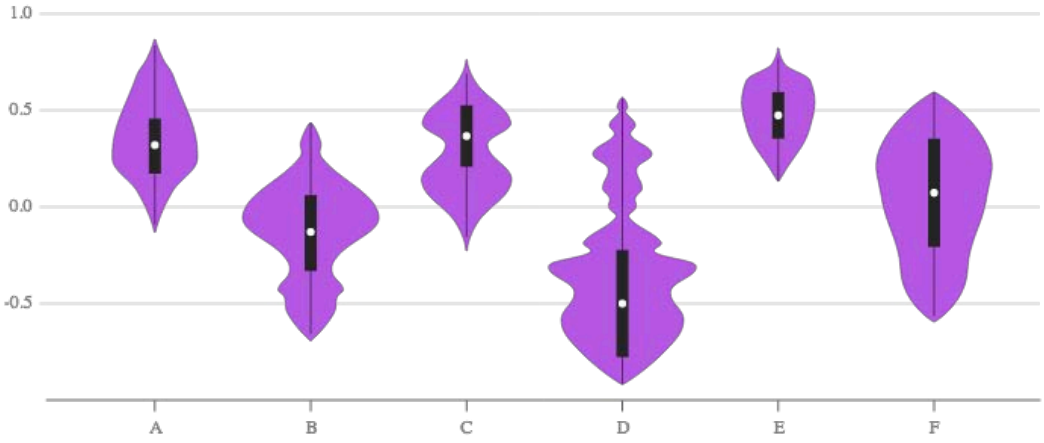
- Selecting bins
  - selecting the width of a bin is not straightforward
  - too small bin leads to a messy plot
  - too large bin leads to loss of information
  - bins can vary in length but one needs to normalise the counts by the width of the bins
  - there exist techniques that can automatically select appropriate bins

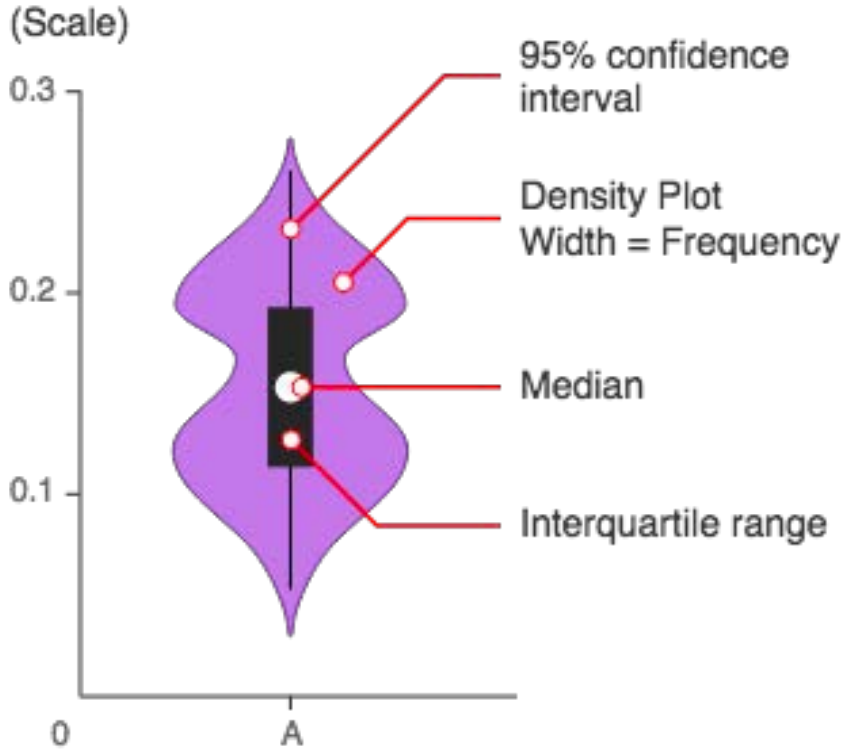




Visualizing distributions

- Box plots vs. histograms
  - histograms reveal more information about distribution
  - especially if the distribution is multimodal (several peaks)
  - box plots are easer to compare with each other
  - violin plot combines both plots together

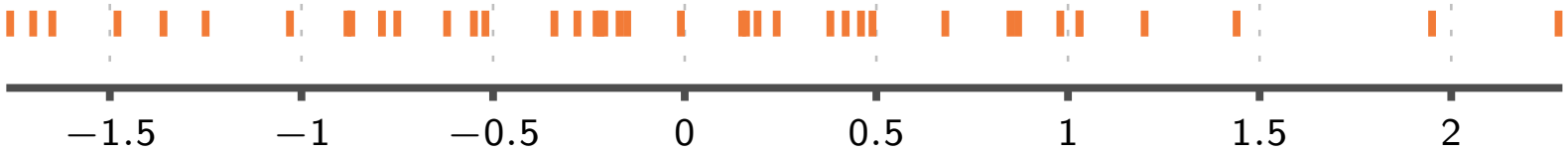




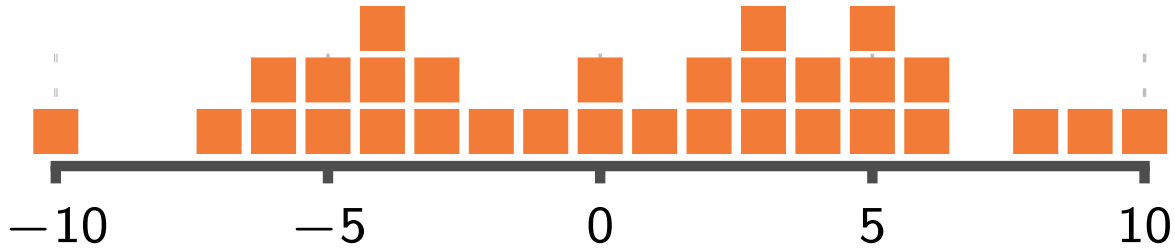
https://datavizcatalogue.com/methods/violin_plot.html

Visualizing distributions

- Strip (rug) plot
  - show individual values at the x-axis
  - apply techniques to reduce overplotting



- stack individual points using y-axis
- results in a histogram-like picture



Visualizing distributions

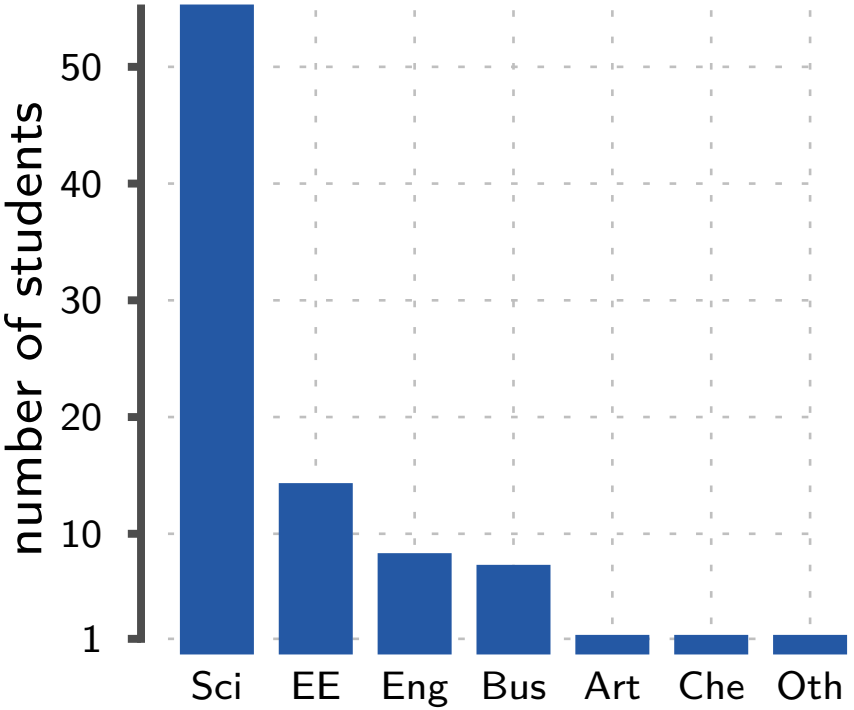
- Stem-and-leaf plot
  - similar to histogram / stacked strip plot
  - each value is split in two parts (stem and leaf)
  - stem is used for y-axis
  - leaf is shown as a number
  - values are sorted and stacked using x-axis
- Data: 1.5, 1.6, 2.1, 2.3, 2.3, 2.6, 2.6, 3.0, 3.2, 4.1
- Stem-and-leaf plot:

```
1 | 56
2 | 13366
3 | 02
4 | 1
```

- leaf value can be truncated
  the cutting point needs to be consistent, for example, integral part is stem and fractional part is leaf

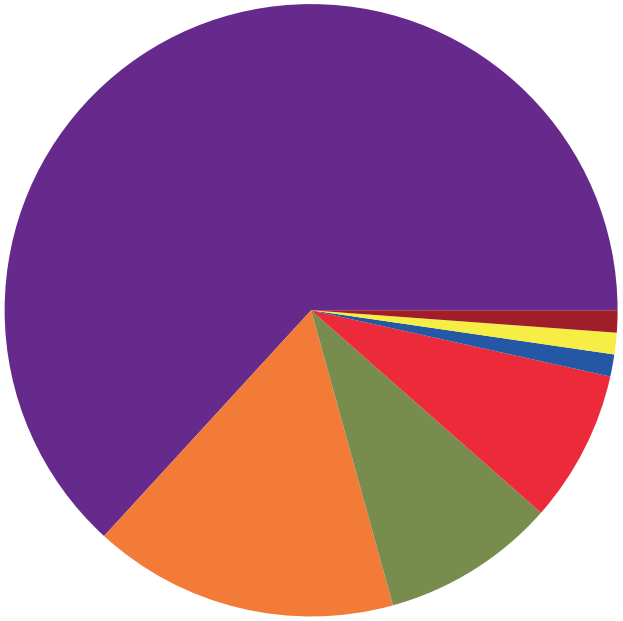
Ranking and part-to-whole analysis

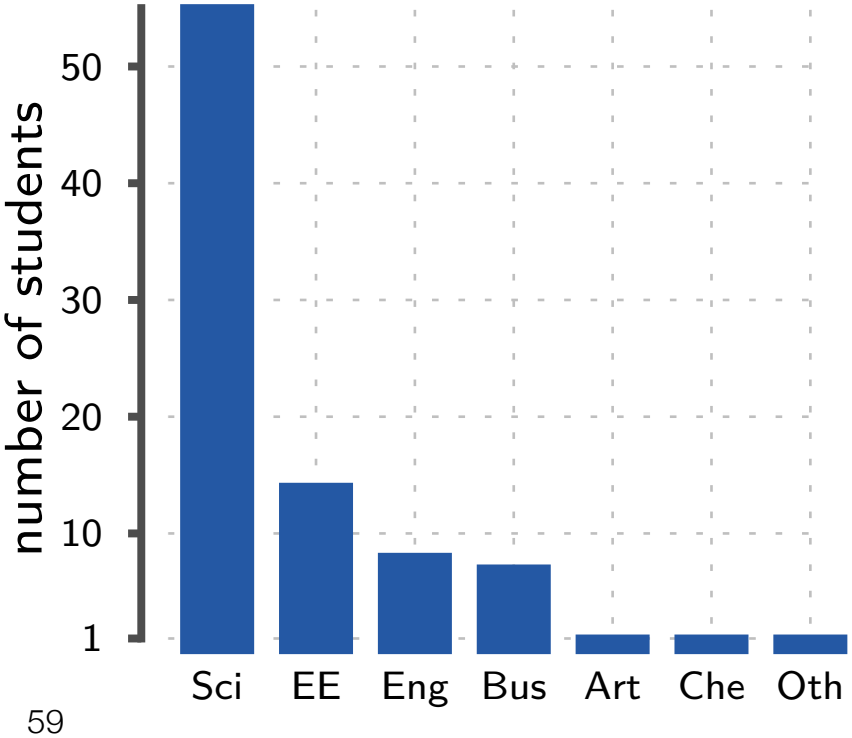
- The goal is to compare
  - how well individual components compare to other
  - what is their ranking



Ranking and part-to-whole analysis

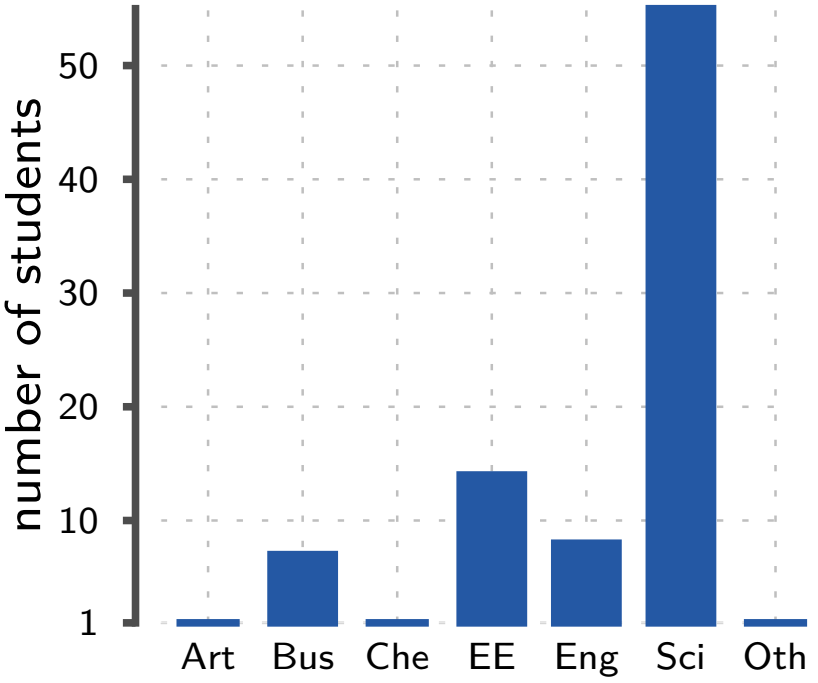
- a common way of displaying such data is a pie chart
- bar graphs are much more effective: easier to compare positions than angles

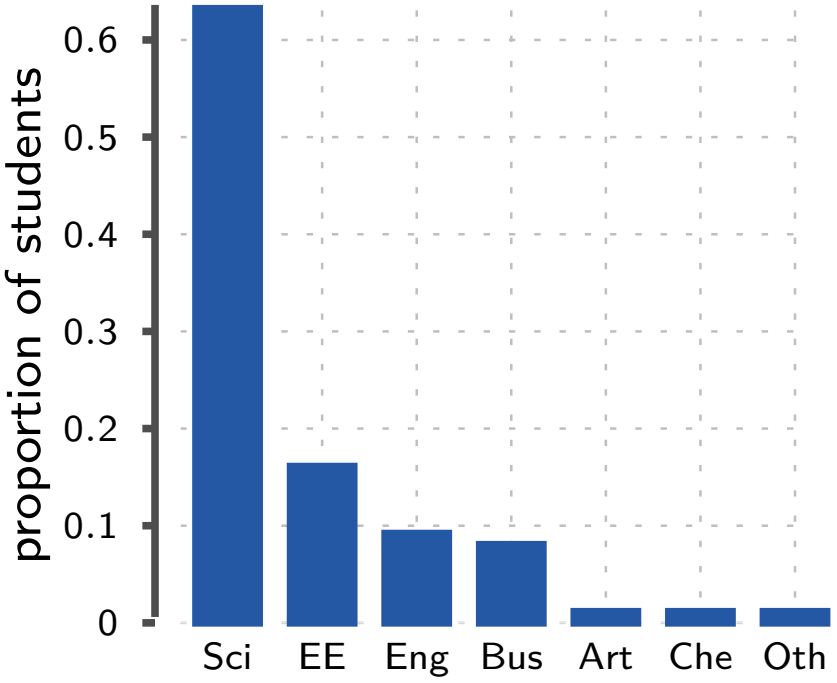




Ranking and part-to-whole analysis

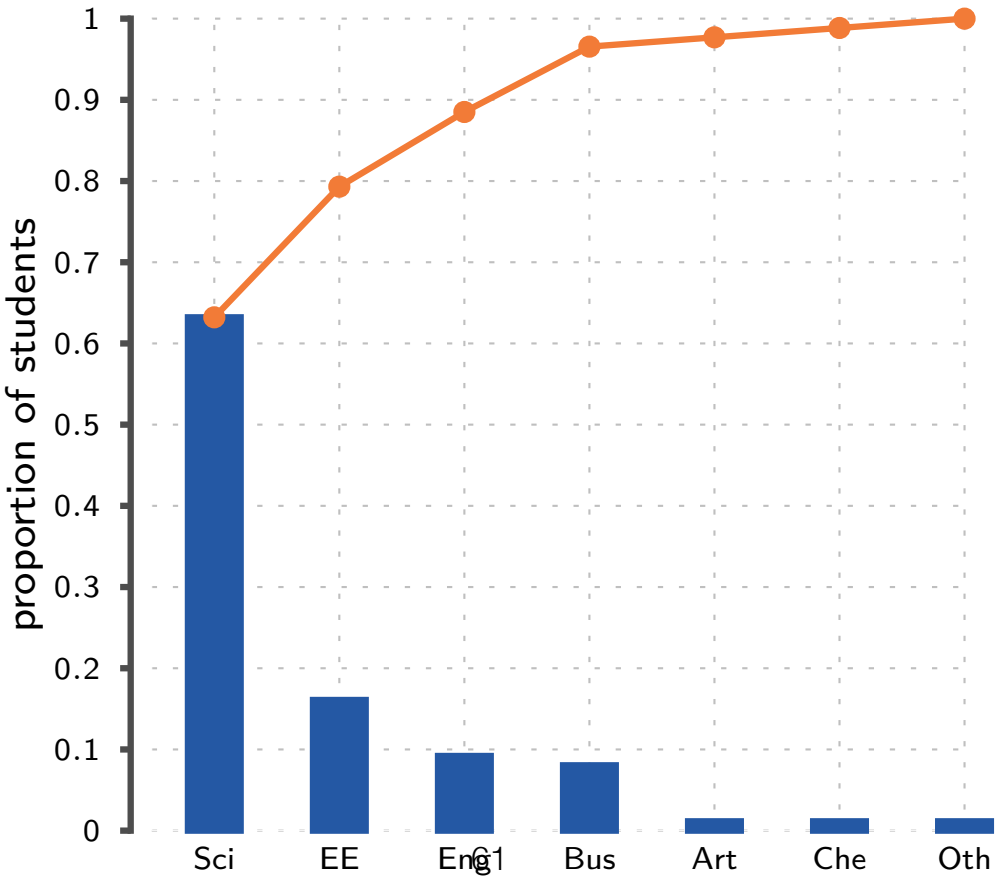
- Two techniques that can greatly improve readability
  - **sort** individual values
  - either normalise the y-axis, or indicate the percentages with labels





Ranking and part-to-whole analysis

- Pareto chart (bar chart + line of cumulative numbers) can be very useful in analysing compositions
  - 80% of students come from two schools



Ranking and part-to-whole analysis

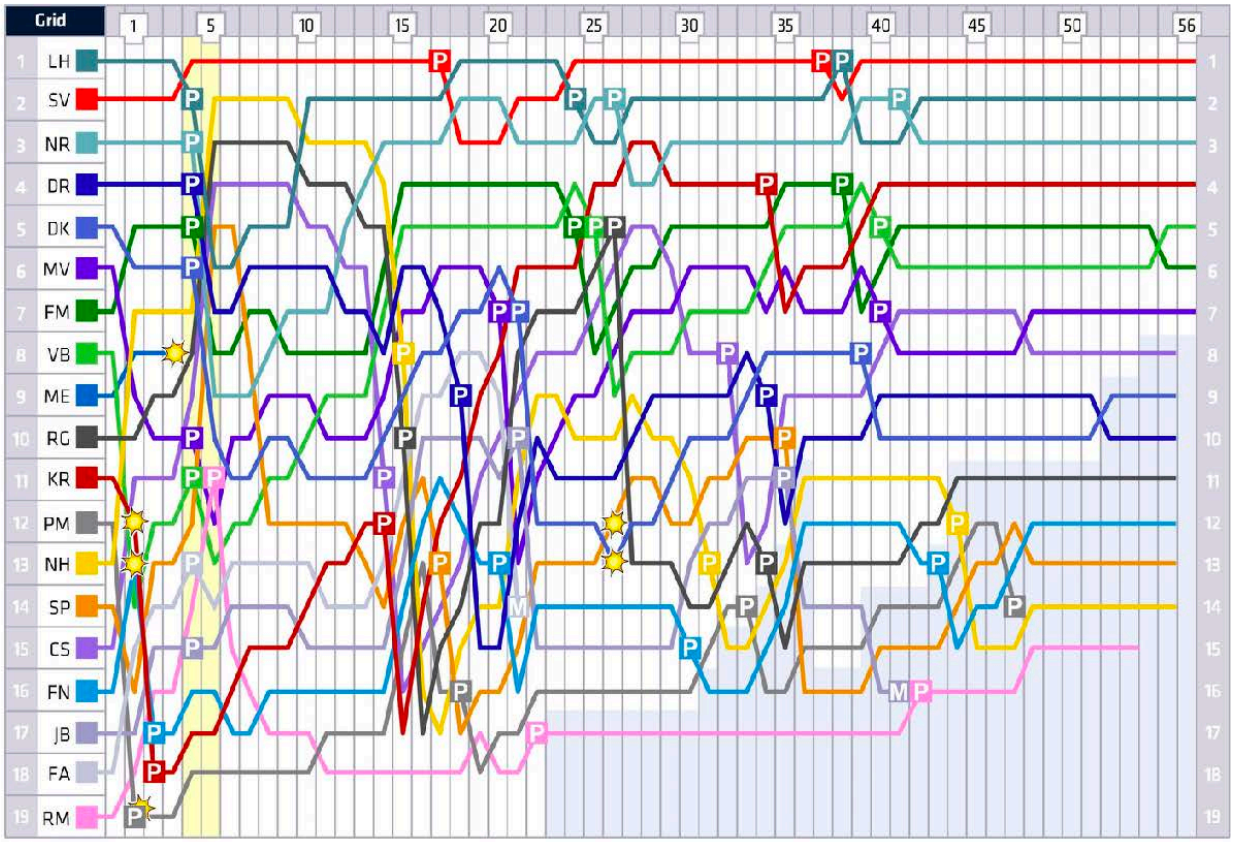
- Ranking over time can be expressed with line graphs

ROUND 02	MALAYSIAN GRAND PRIX
RACE DATE:	29 MAR 2015
CIRCUIT NAME:	SEPANG INTERNATIONAL CIRCUIT
NUMBER OF LAPS:	56
START TIME	15:00 Local - 07:00 GMT
CIRCUIT LENGTH:	5.543KM
RACE DISTANCE:	310.408KM
LAP RECORD:	1:34.223 - J P Montoya [2004]

KEY

Accident
M Mechanical failure
P Pit stop
E Excluded
Black Flagged
Red Flagged
Safety Car
Lapped

LH: L Hamilton
NR: N Rosberg
DR: D Ricciardo
DK: D Kvyat
FM: F Massa
VB: V Bottas
SV: S Vettel
KR: K Raikkonen
FA: F Alonso
JB: J Button
SP: S Perez
NH: N Hulkenberg
MV: M Verstappen
CS: C Sainz
RG: R Grosjean
PM: P Maldonaldo
ME: M Ericsson
FN: F Nasr
RM: R Merhi



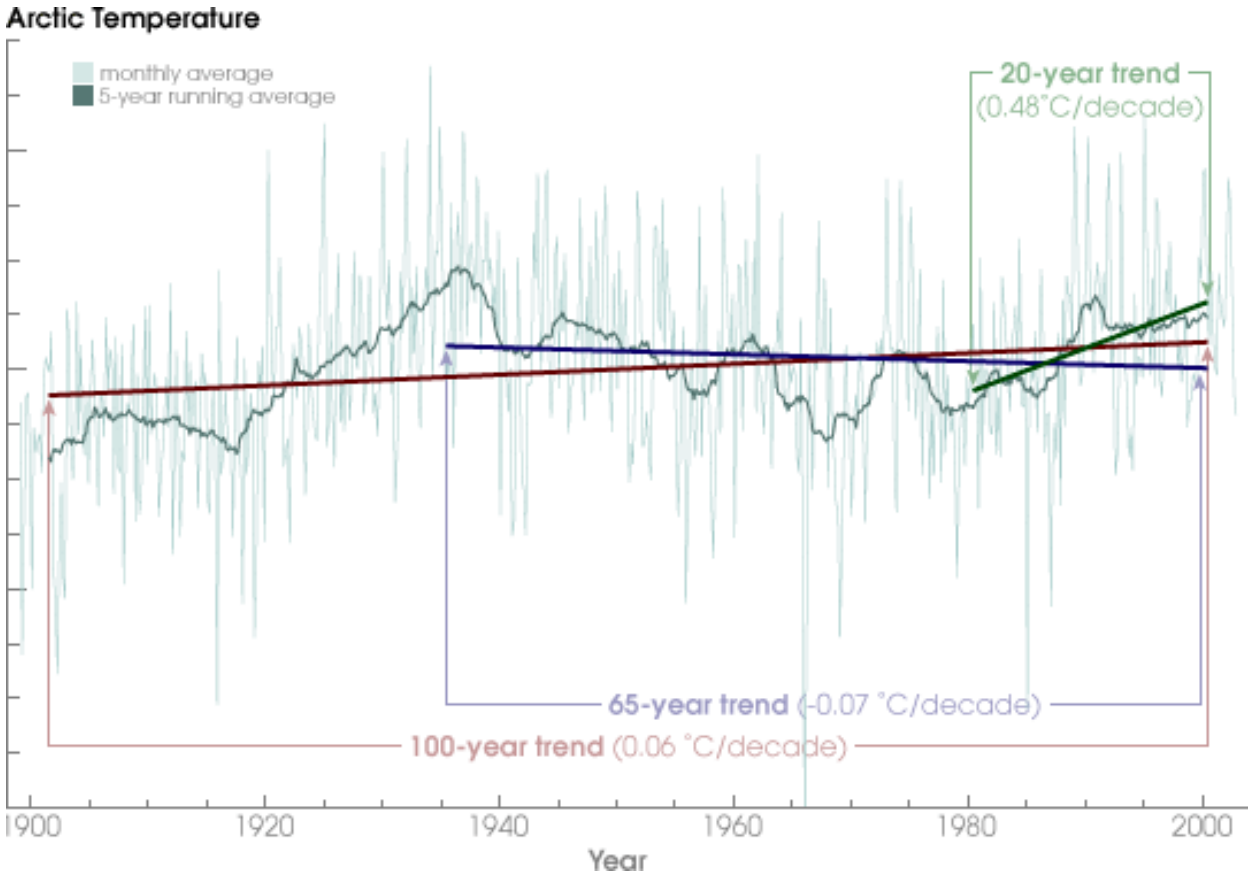
https://www.fia.com/events/fia-formula-one-world-championship/season-2015/lap-chart-0

Time-series displays

- 7 types of graphics are useful for examining time-series:
  - Line graphs
  - Bar graphs
  - Dot plots
  - Radar graphs
  - Heat-maps
  - Box plots (and similar) for analyzing distribution over time
  - (Animated) scatter plots

Time-series displays

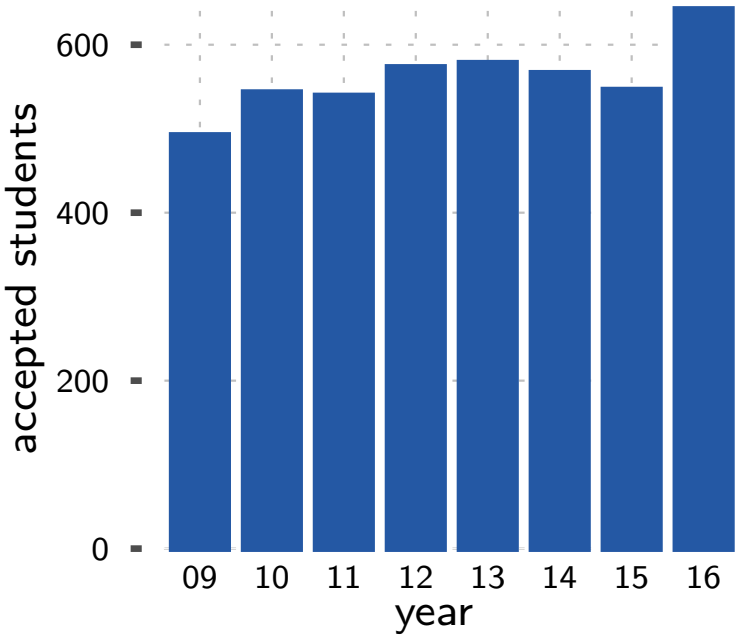
- Line graph is typically the best choice
- use it to show patterns over time:
  - increasing / decreasing trend
  - variation / volatility over time
- exceptions / outlier behaviour

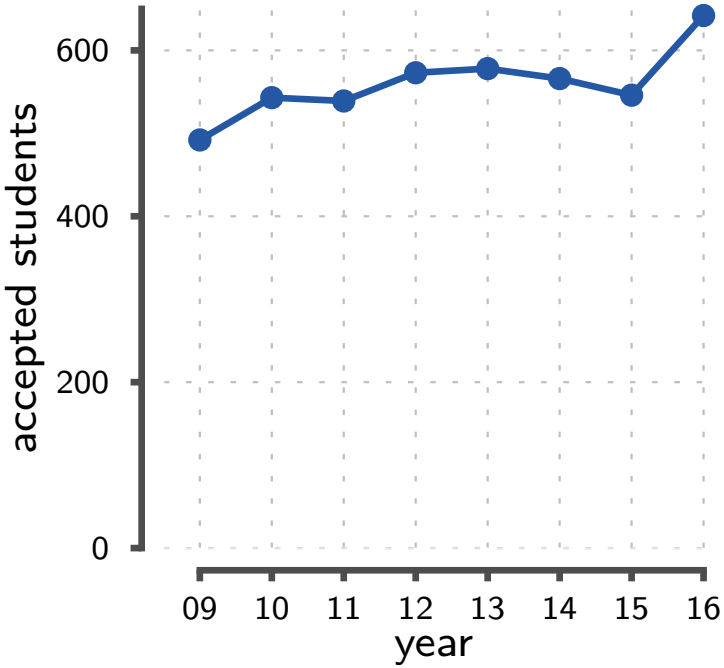


https://earthobservatory.nasa.gov/Features/ArcticIce/arctic_ice3.php

Time-series displays

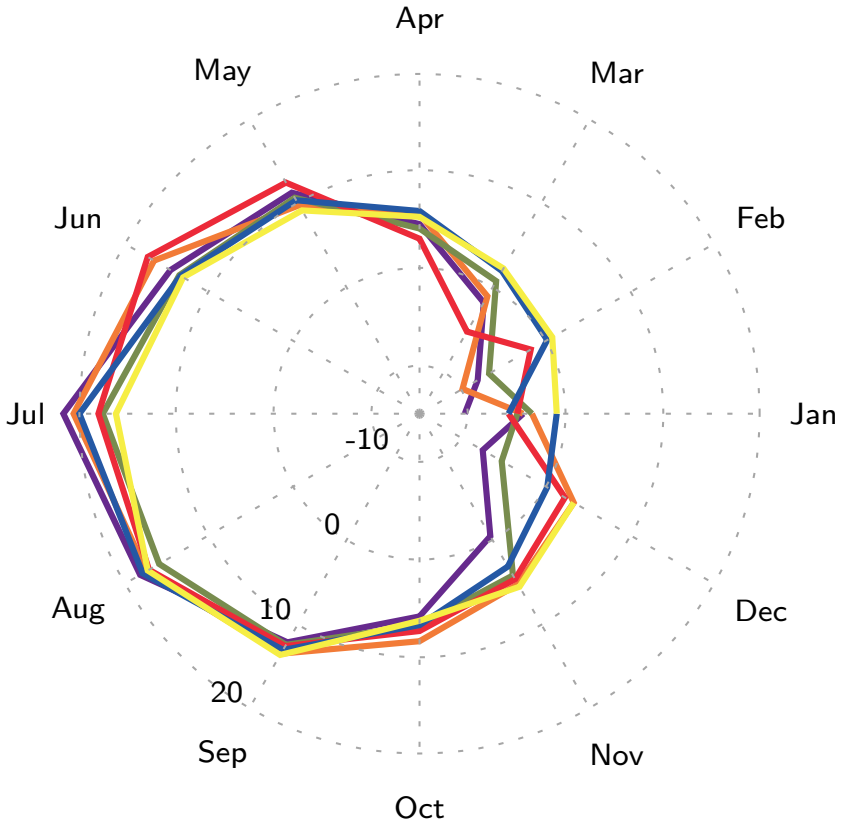
- Bar graph works better if
  - you want to compare individual values
- You should use dot plot if you cannot use line chart, e.g.when
  - measurements at irregular time intervals
  - you cannot guarantee that the intermediate values are close to linear interpolation

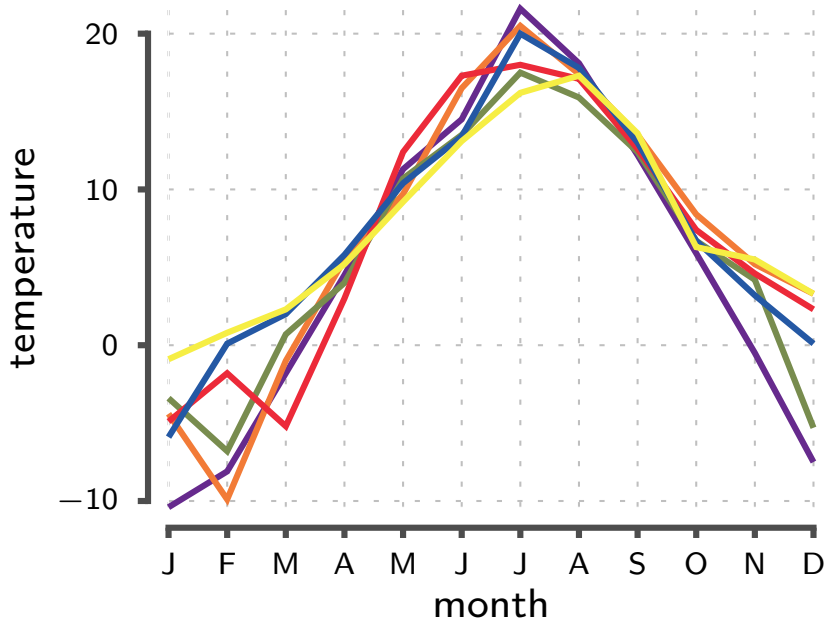




Time-series displays

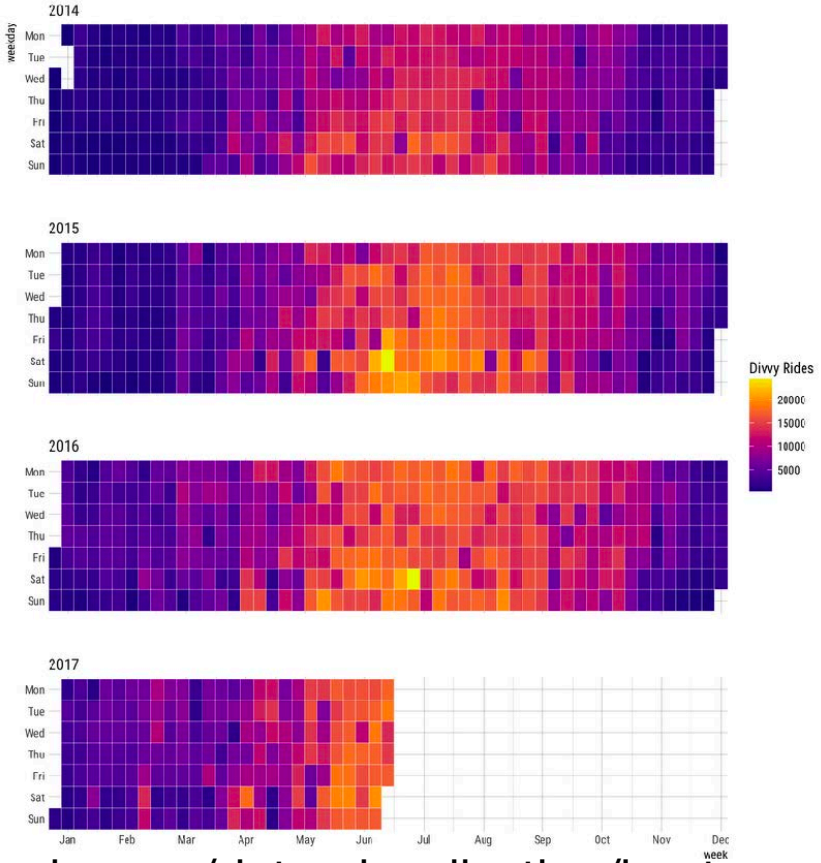
- Radar plots
  - can be used to show cycles over time
  - a line chart with superimposed lines is probably better





Time-series displays

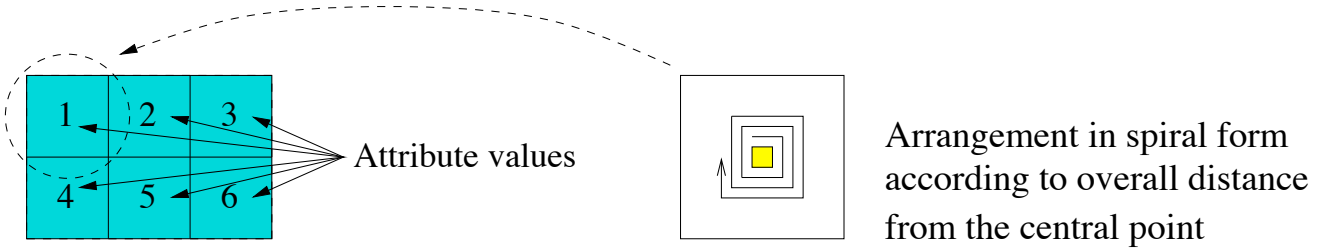
- A **heatmap** shows many time-series, if superimposition creates too cluttered picture
- y-axis is individual time series, x-axis is time, colour shows the value
- similar arrangement (but different encoding) than with small multiples

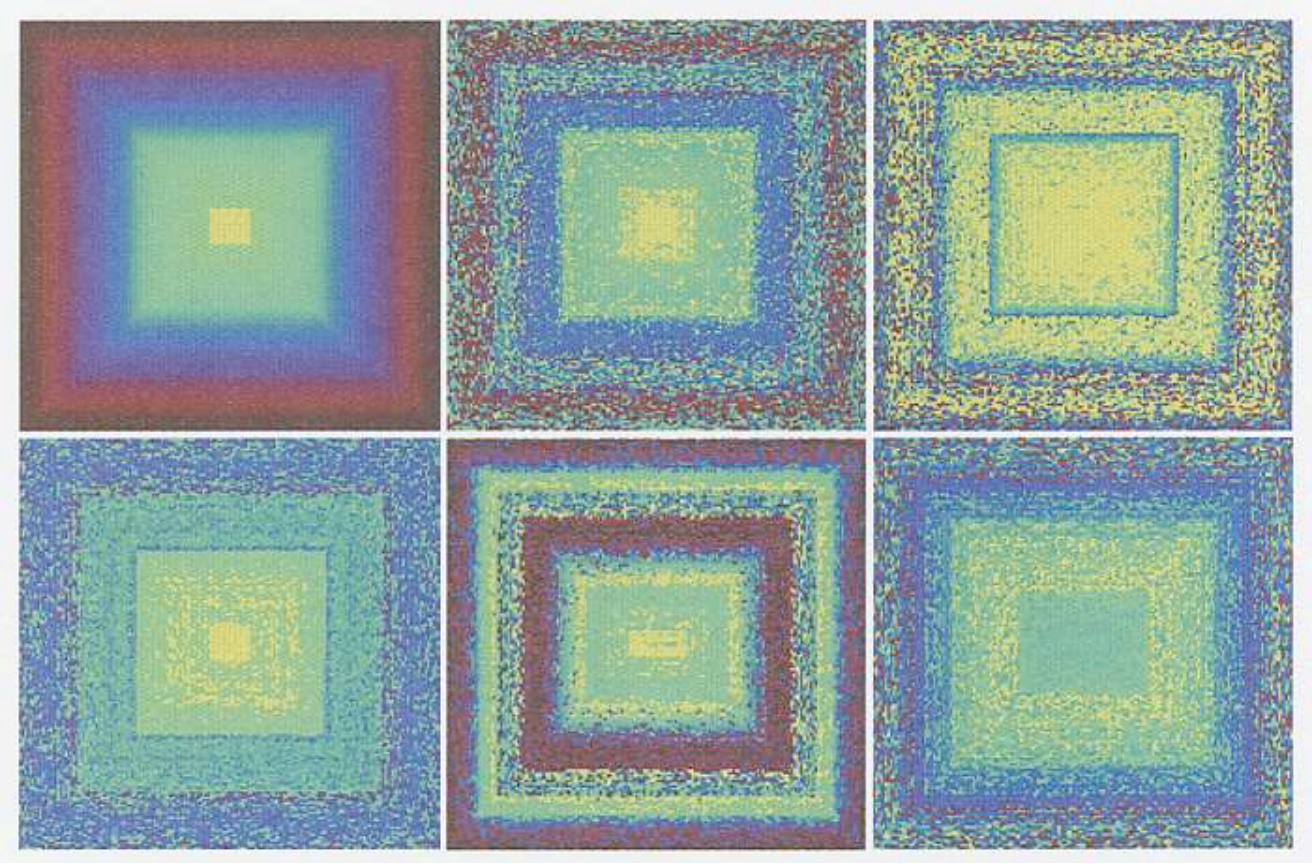


http://austinwehrwein.com/data-visualization/heatmaps-with-divvy-data/

Time-series displays

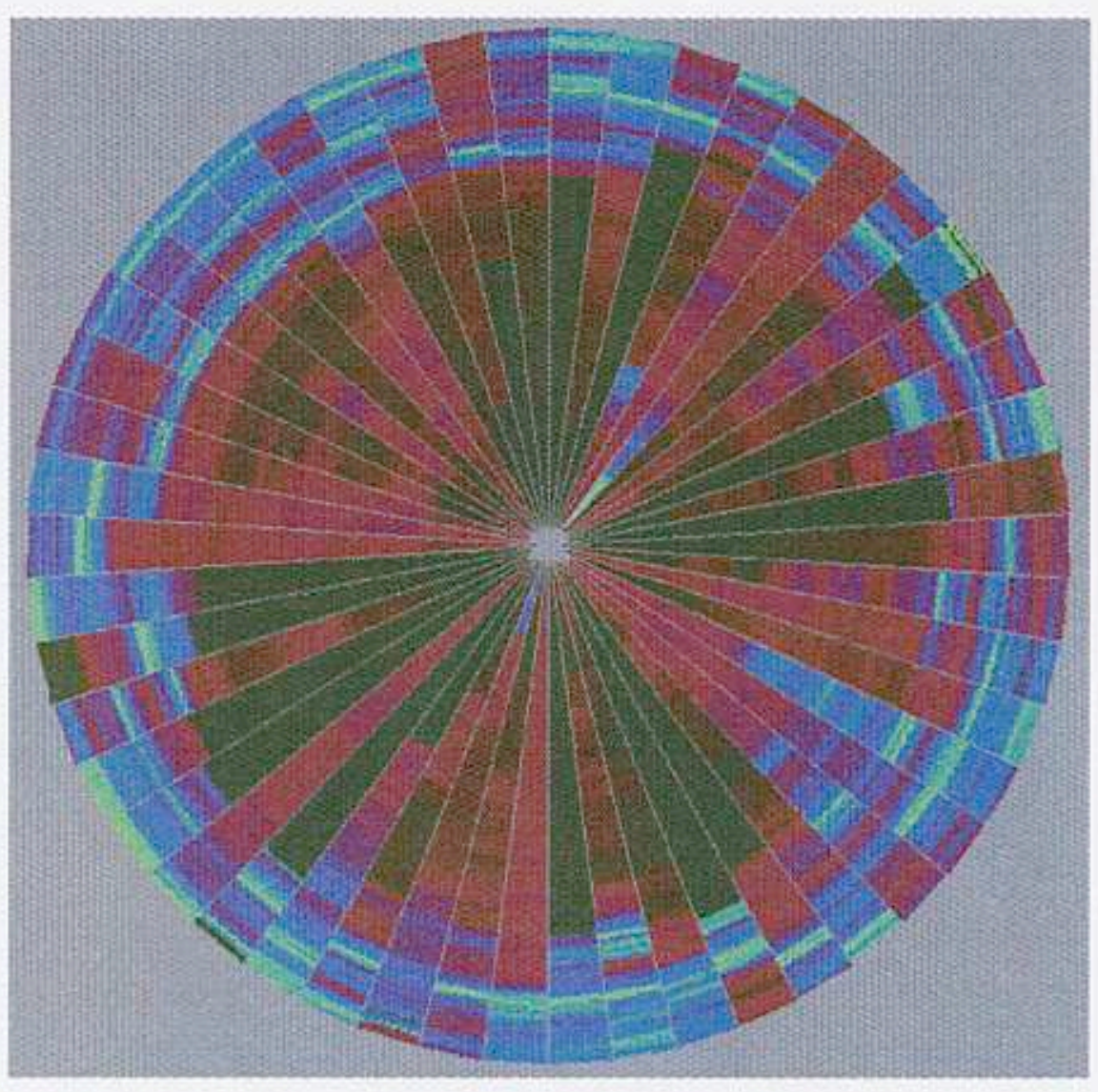
- **Pixel-oriented techniques**
  - Each attribute value is represented by one pixel (the value ranges are mapped to a fixed colour-map)
  - The attribute values for each attribute are represented in separate sub-windows





Time-series displays

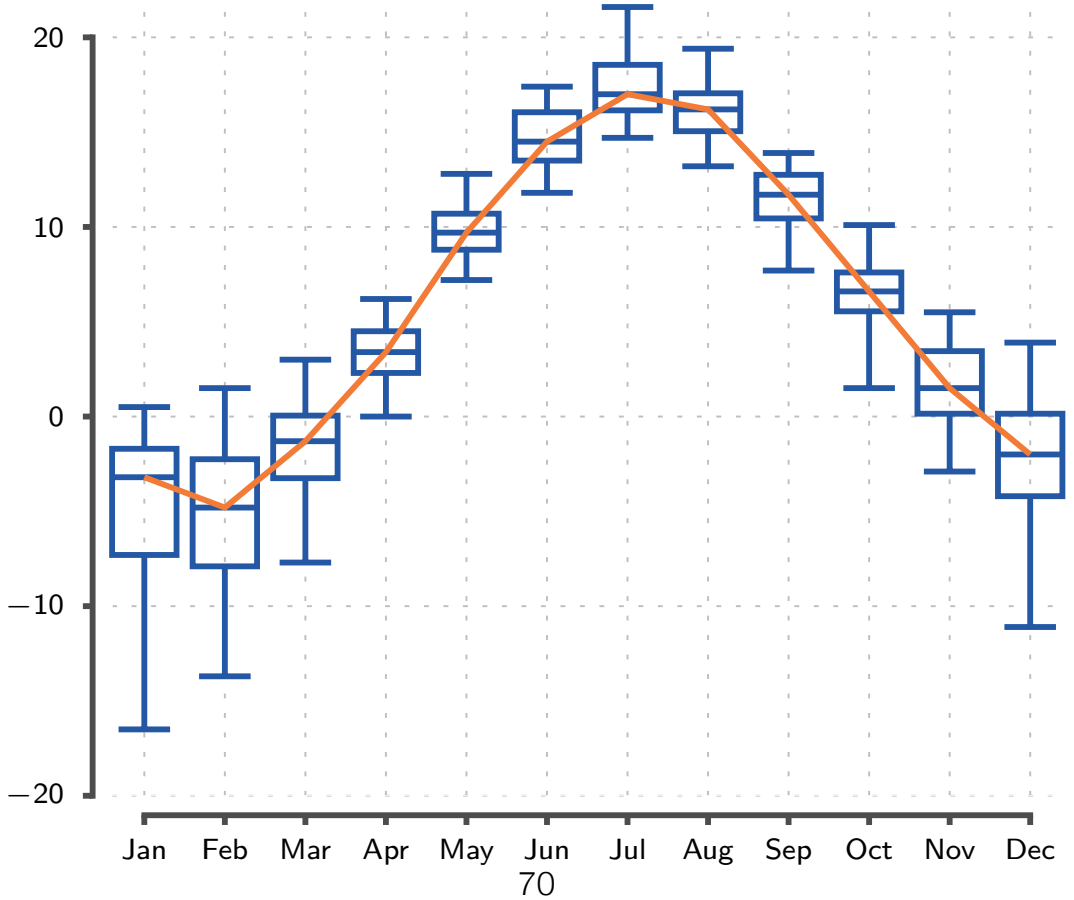
- **Pixel-oriented techniques**
  - Circle segments (time goes out from center, creating a pseudoperspective)



Time series of 50 stocks of the Frankfurt Stock Index [K 50].

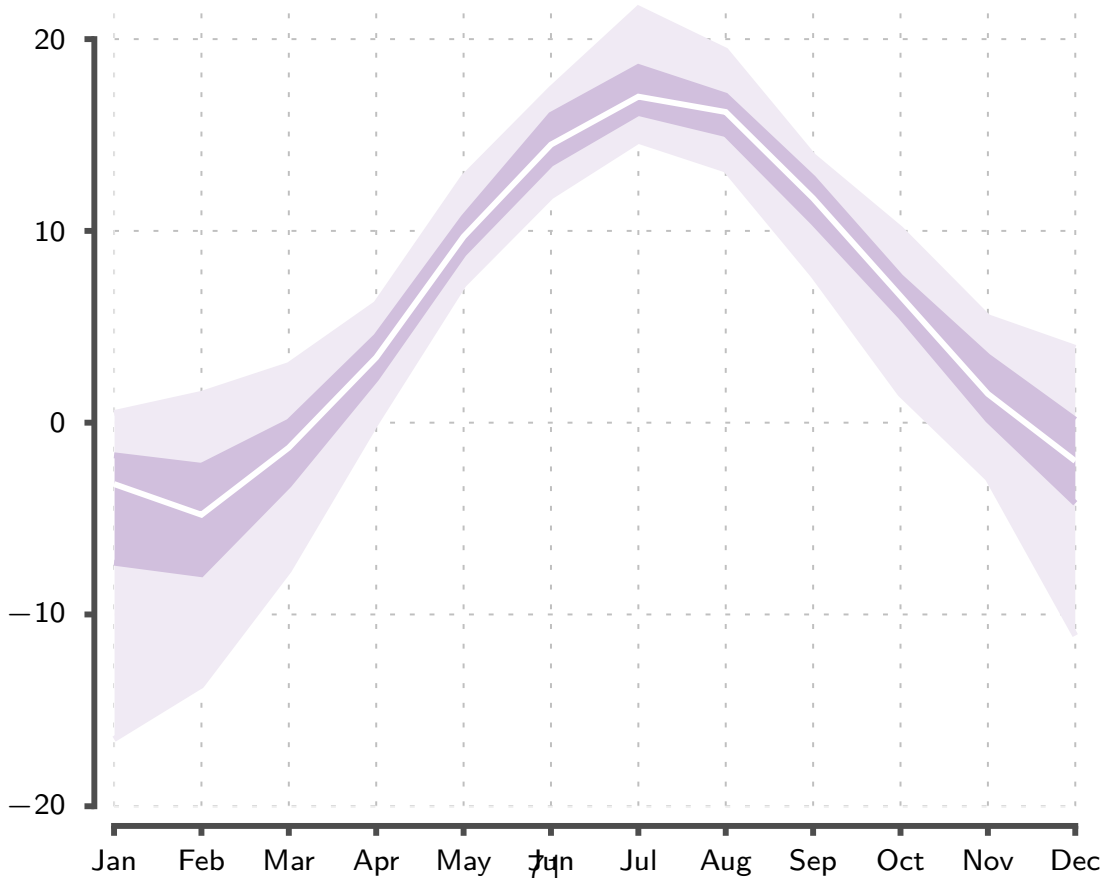
Time-series displays

- Box plots are useful
  - if you have significant number of time-series
  - you are only interested in how the distribution changes over time



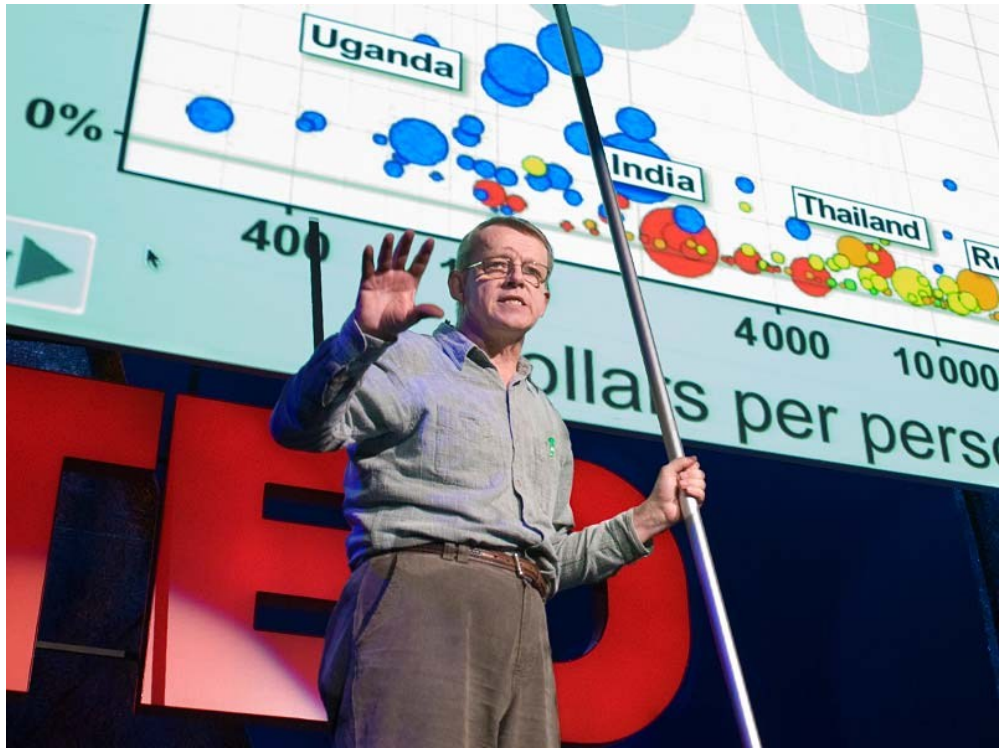
Time-series displays

- You can also plot the statistics as lines or areas
  - if you have significant number of time-series
  - you are only interested in how the distribution changes over time



Time-series displays

- Animation and scatter plots can be used
  - if you wish to show how two quantities change over time
  - more complex plots are possible (for example bubble maps)



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

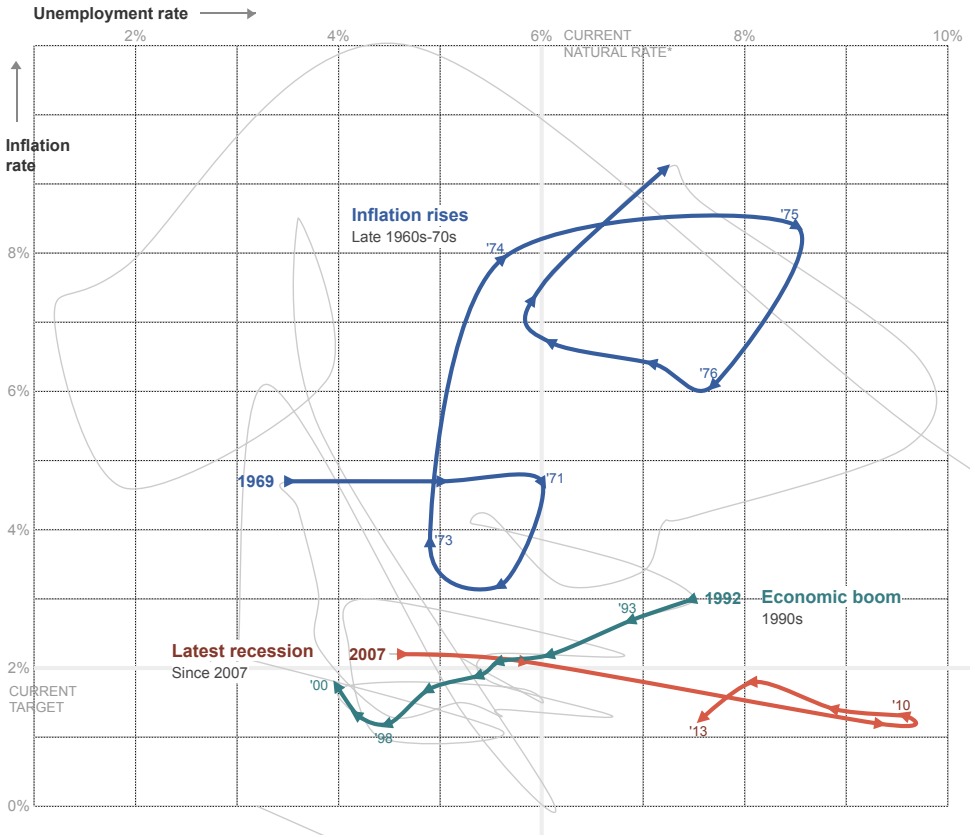
try https://www.gapminder.org/tools/

Time-series displays

- These animations can be simulated on paper by showing trails
  - similar to space-time narratives but with abstract quantities

Inflation and unemployment
The Federal Reserve is said to have a "dual mandate": keeping inflation in check and the unemployment rate low. These measures, which tend to change cyclically and in concert with each other, are charted for every year since the Great Depression.

In speeches and in meetings, Ms. Yellen, the nominee for the next Fed leader, has commented on the Fed's actions during significant periods, providing a window into her views and priorities.



http://www.nytimes.com/interactive/2013/10/09/us/yellen-fed-chart.html

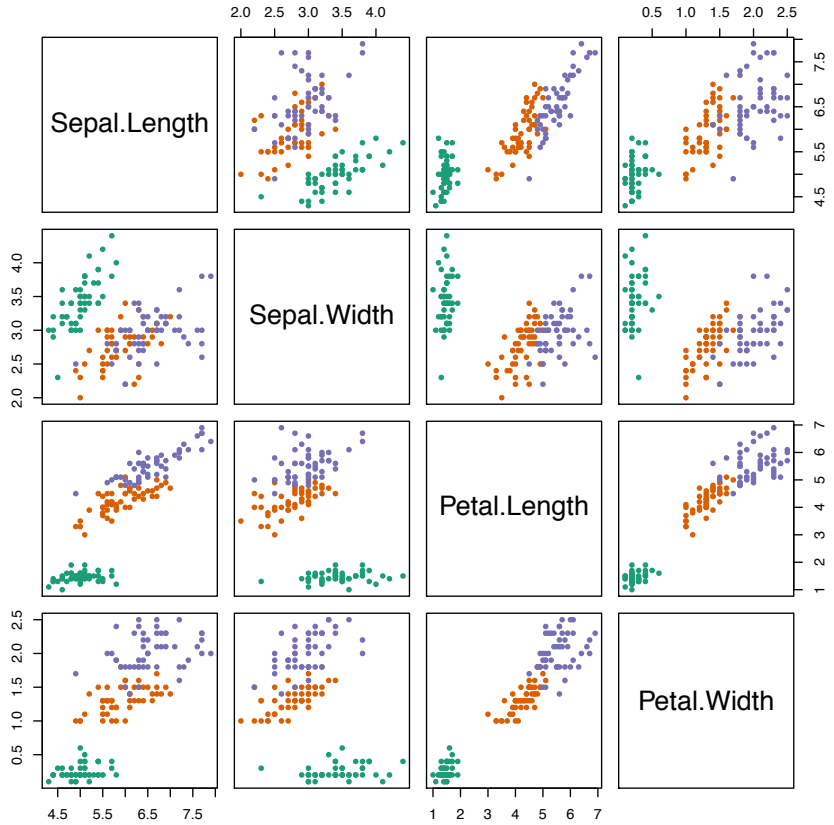
High-dimensional data

- Options for visualising high-dimensional multivariate data
  - small multiples with simple plots
  - heat-maps
  - parallel coordinates
  - glyphs
  - dimension reduction techniques

- All these techniques have problems
  - ask yourself what is the story that you are trying to tell and
  - visualise accordingly or
  - apply data mining tools to extract new knowledge from the data

High-dimensional data

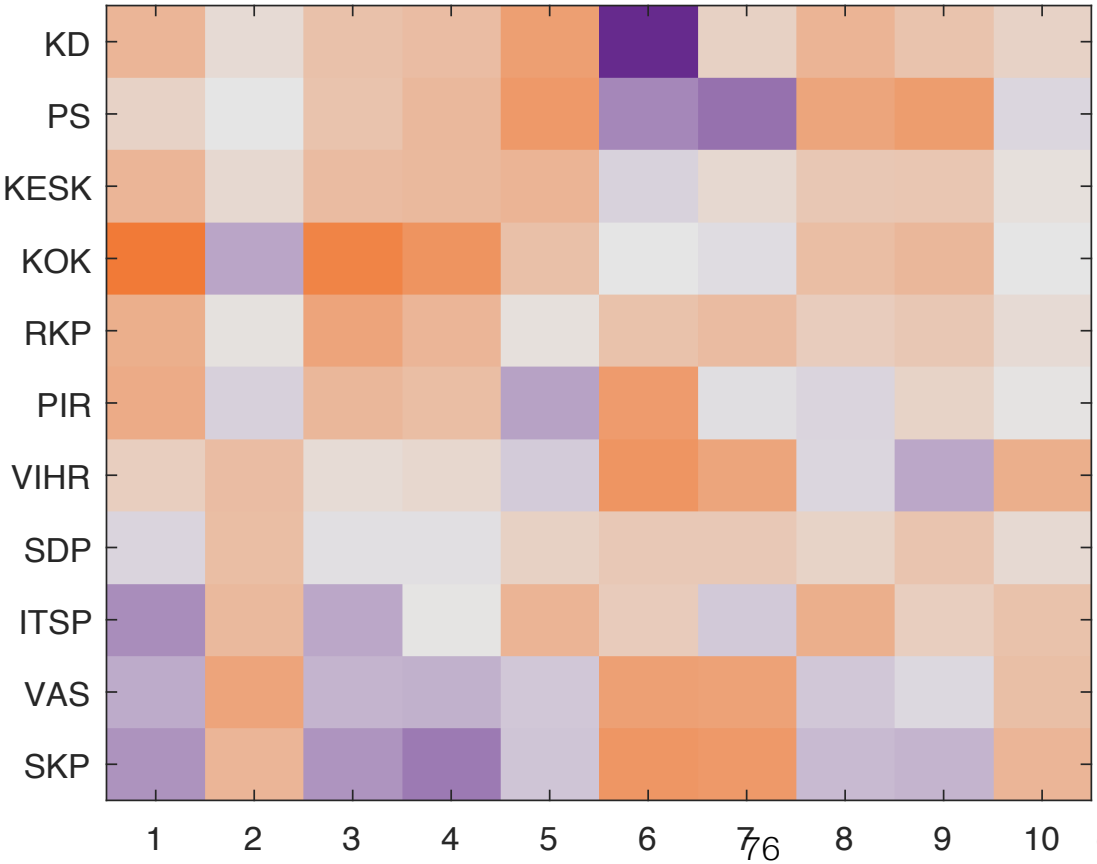
- Small multiples
  - array of a matrix of small plots
  - may result in an overwhelming plot




```
pairs(iris[,1:4],col=brewer.pal(3,"Dark2")[iris[,5]],pch=20)
```

High-dimensional data

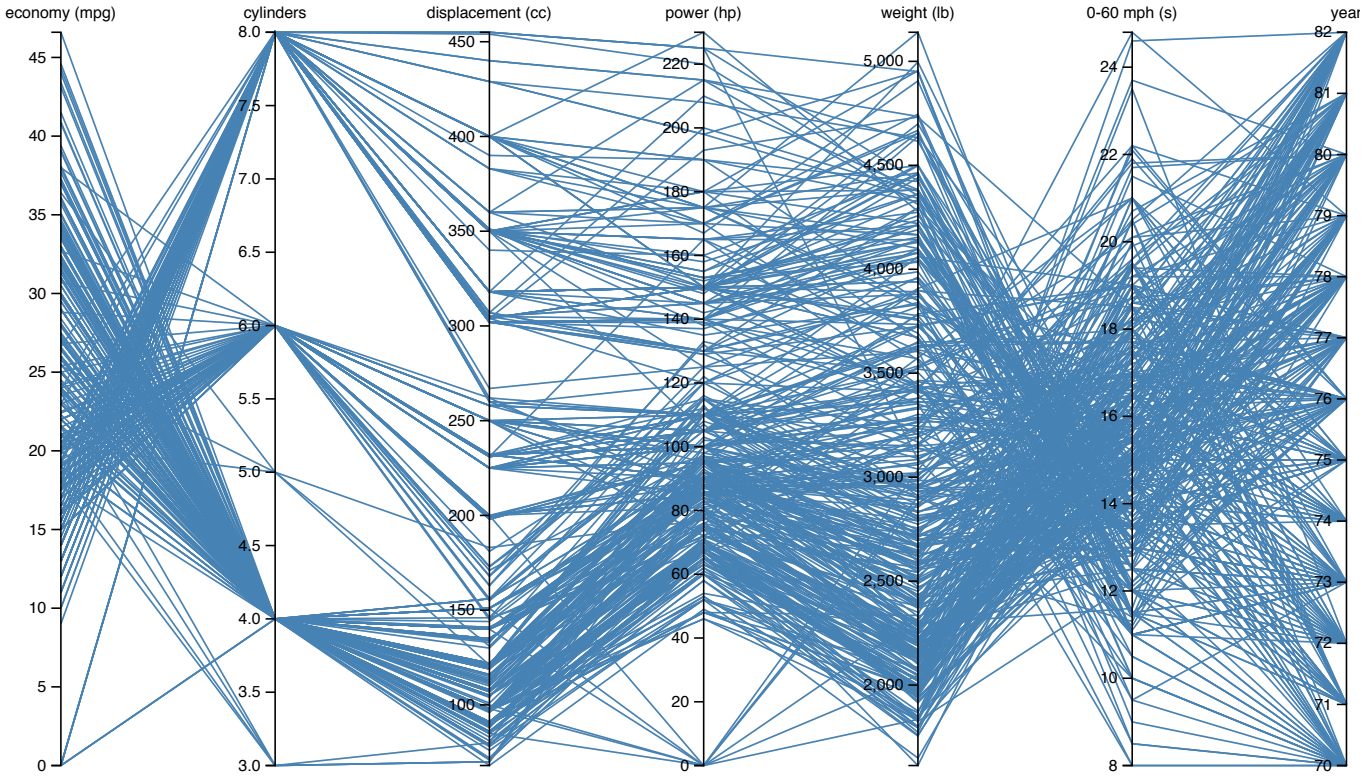
- Heat-map
  - dimensions are on x-axis
  - data points are on y-axis
  - colour represent the value
  - may be difficult to extract any information
  - ordering dimensions and data points is crucial



Questionaire by HS vaalikone

High-dimensional data

- Parallel coordinates
  - line graphs: x-axis are individual dimensions
  - each data point is a line
  - somewhat counter-intuitive and may result in cluttered picture
  - order of dimensions matter but may reveal information that is not visible in other designs
  - highlights clusters
  - may work better as an interactive tool

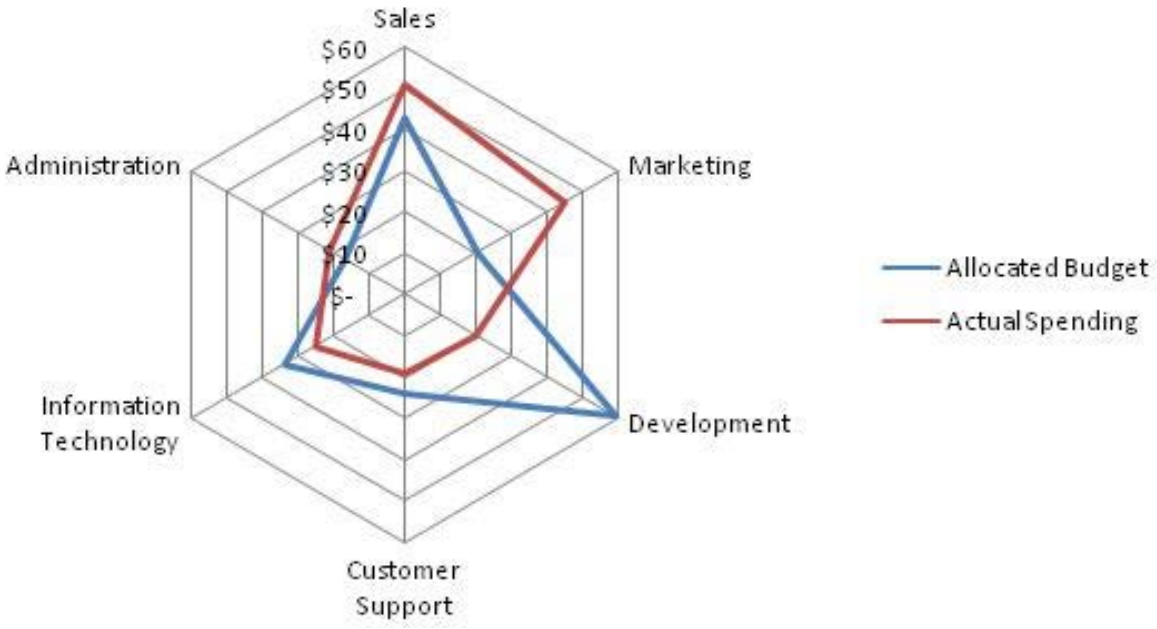


https://bl.ocks.org/jasondavies/1341281

High-dimensional data

- Radar chart is equivalent to parallel coordinates plot, with the axes arranged radially
- May be useful
  - when used to compare overall similarity
  - with consistent ordinal dimensions, e.g. performance scores
    - small values near the center mean better in each dimension
- Generally not recommended because
  - may impose artificial cyclic structure; areas may be misleading

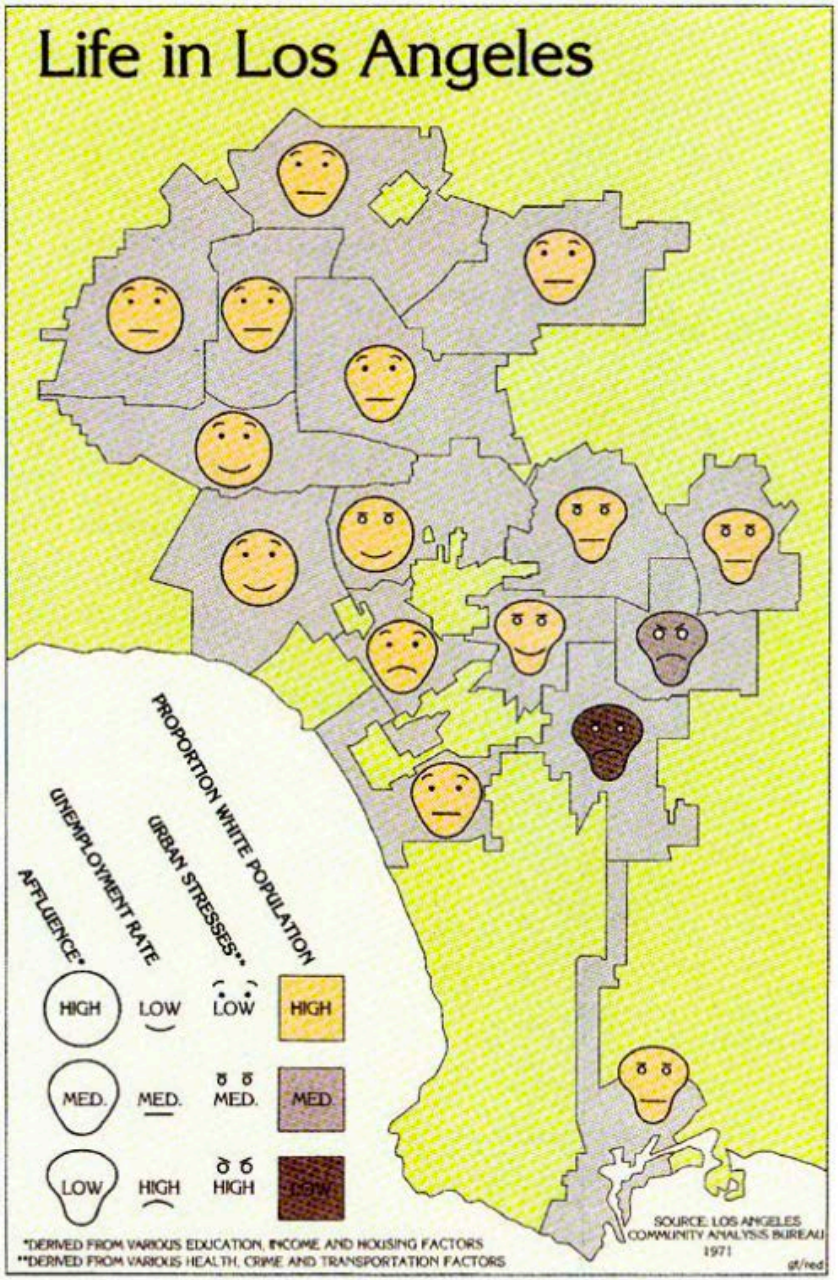
(do not fill the polygon with color!)



High-dimensional data

[more on glyphs in Part II]

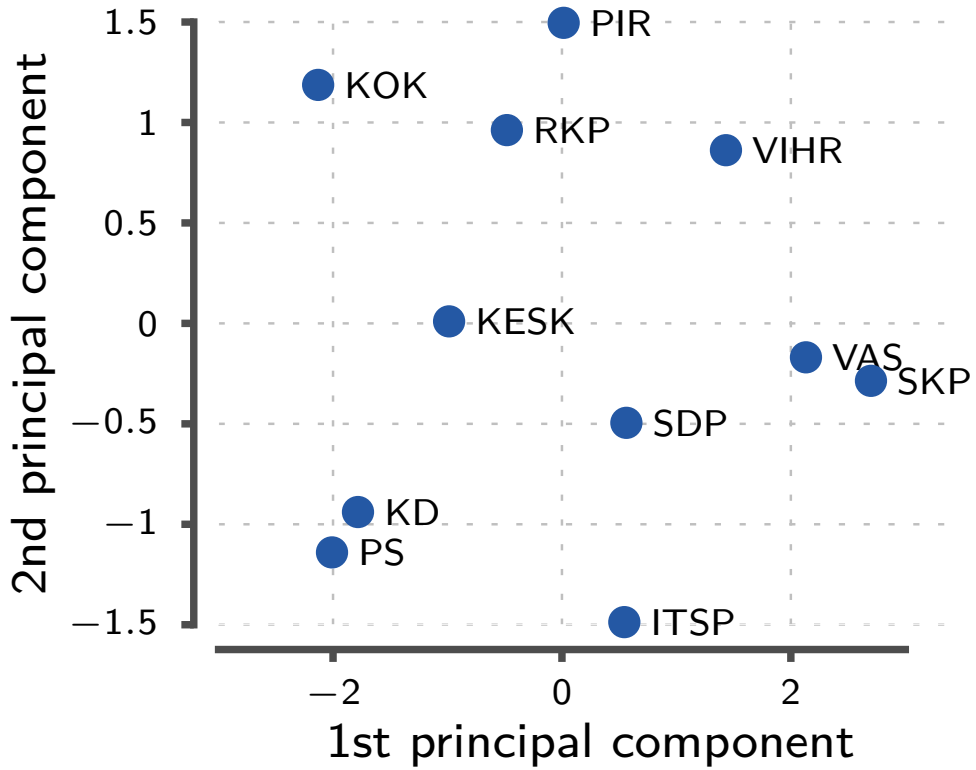
- Glyphs
  - shortened for hieroglyphs
  - small 'subplots' that can be placed in a scatter plot
  - at simplest glyphs are coloured dots
  - more complicated glyphs are possible to indicate more dimensions (for example Chernoff faces)
  - can only carry limited information before the plot gets too cluttered



http://mapdesign.icaci.org/2014/12/mapcarte-353365-life-in-los-angeles-by-eugene-turner-1977/

High-dimensional data

- Reducing dimension
  - different techniques that try to plot high-dimensional data as a scatter plot
  - points that are 'close'/'far' in the data are also close/far in the plot
  - powerful to reveal new knowledge hidden within the data
  - but almost always introduce distortion



[more during Part III]

Recap

- four basic graphic elements:
  - points, lines, bars, and boxes
- use lines to show trends, variability
- use bars to compare individual numbers
- use reference lines/regions for comparison
- use multiple plots if story/data requires it:
  - small multiples, overview/detail, multiform
- rescale/re-express if the relative comparison is important
- muted colors for surfaces, bright colors for objects
- use opponent colors [more during Part II]

Next lecture

- Part II: Human perception (following Ware's book)