

CS-E4840

Information Visualization

Lecture 4: Visualization

techniques

Tassu Takala <tapio.takala@aalto.fi>

11 March 2021

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 in practice

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

The five principles (on data-ink)

- **Above all show the data**
- The larger the share of data-ink, the better (all other things being equal): **Maximize the data-ink ratio, within reason.**
- Maximizing the data-ink ratio implies minimizing the amount of non-data ink:
 - **Erase non-data-ink, within reason.**
 - **Erase redundant data-ink.**
- **Revise and edit.**

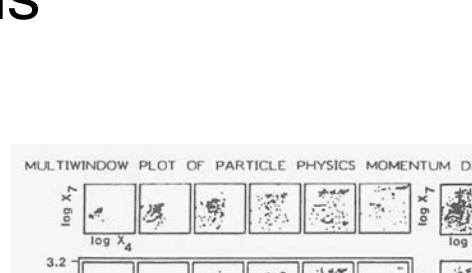
Theory in practice

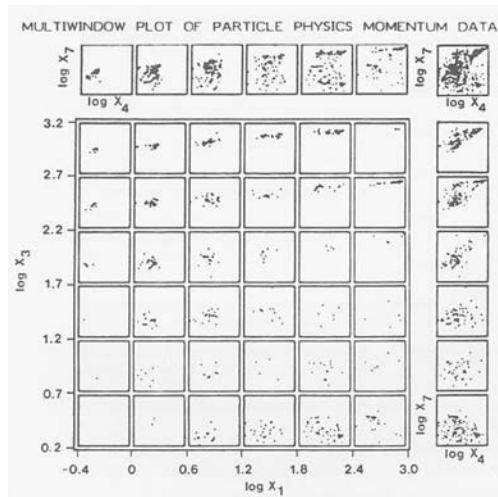
- Data-ink
- **Chartjunk**
- Multifunctioning graphical elements
- Data density and small multiples
- Aesthetics and techniques

Chartjunk

- *Chartjunk* is the interior decoration of graphics that does not tell the viewer anything new
- The purpose of the chartjunk may be to
 - make the graphics appear more scientific and precise (grid lines, excess ticks, redundant representations of simple data etc.)
 - decorate the graphics
 - make the data appear more lively

Types of chartjunk

- Ducks: eye candy and self-promoting graphics)
 - Vibrations
 - Grids

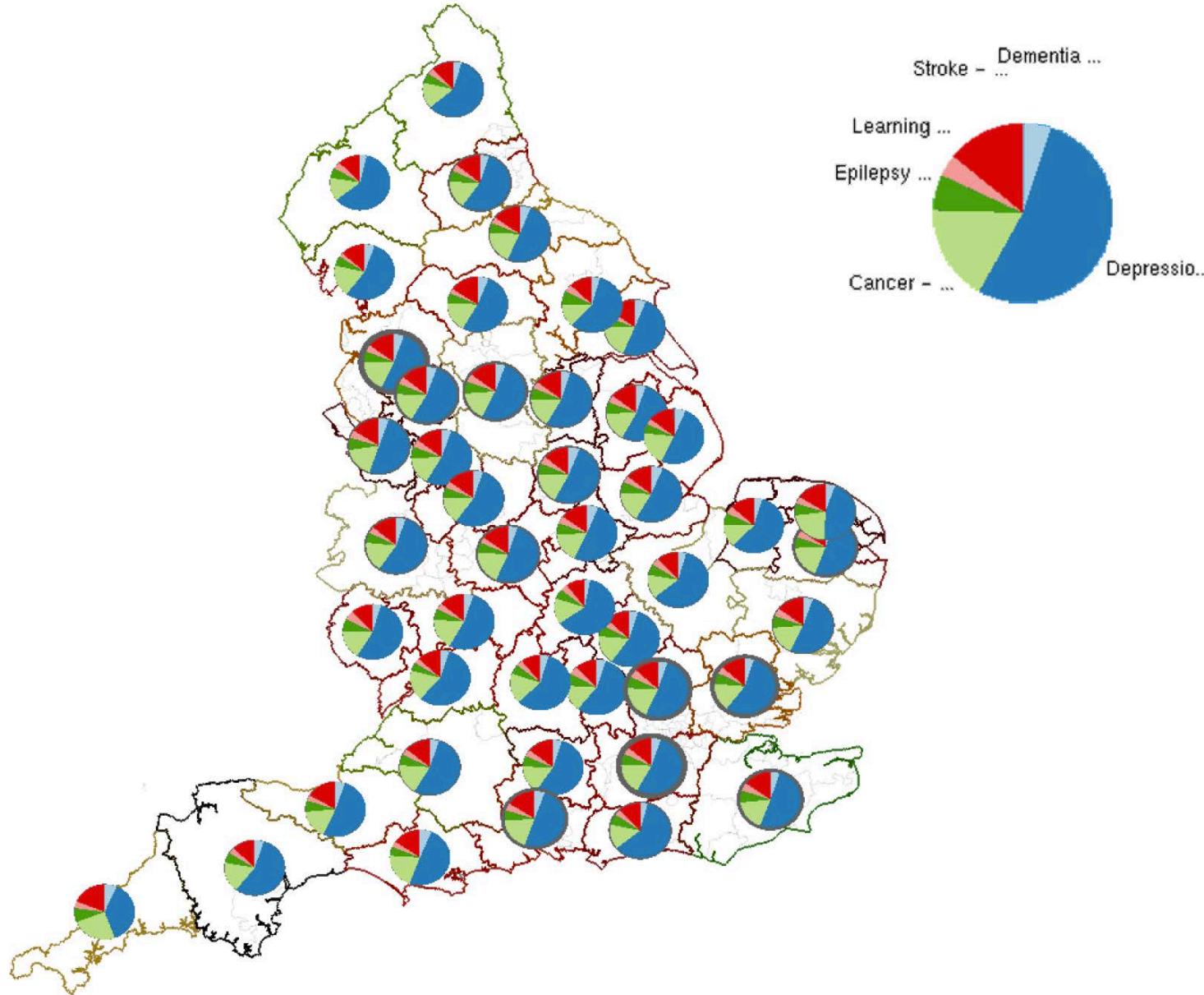


H. Spissler, Reuters, 1999.

Theory in practice

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

Multifunctioning graphical elements: glyphs on a map



Stem-and-leaf display

R Console

```
> attach(faithful)
> stem(eruptions)

The decimal point is 1 digit(s) to the left of the |

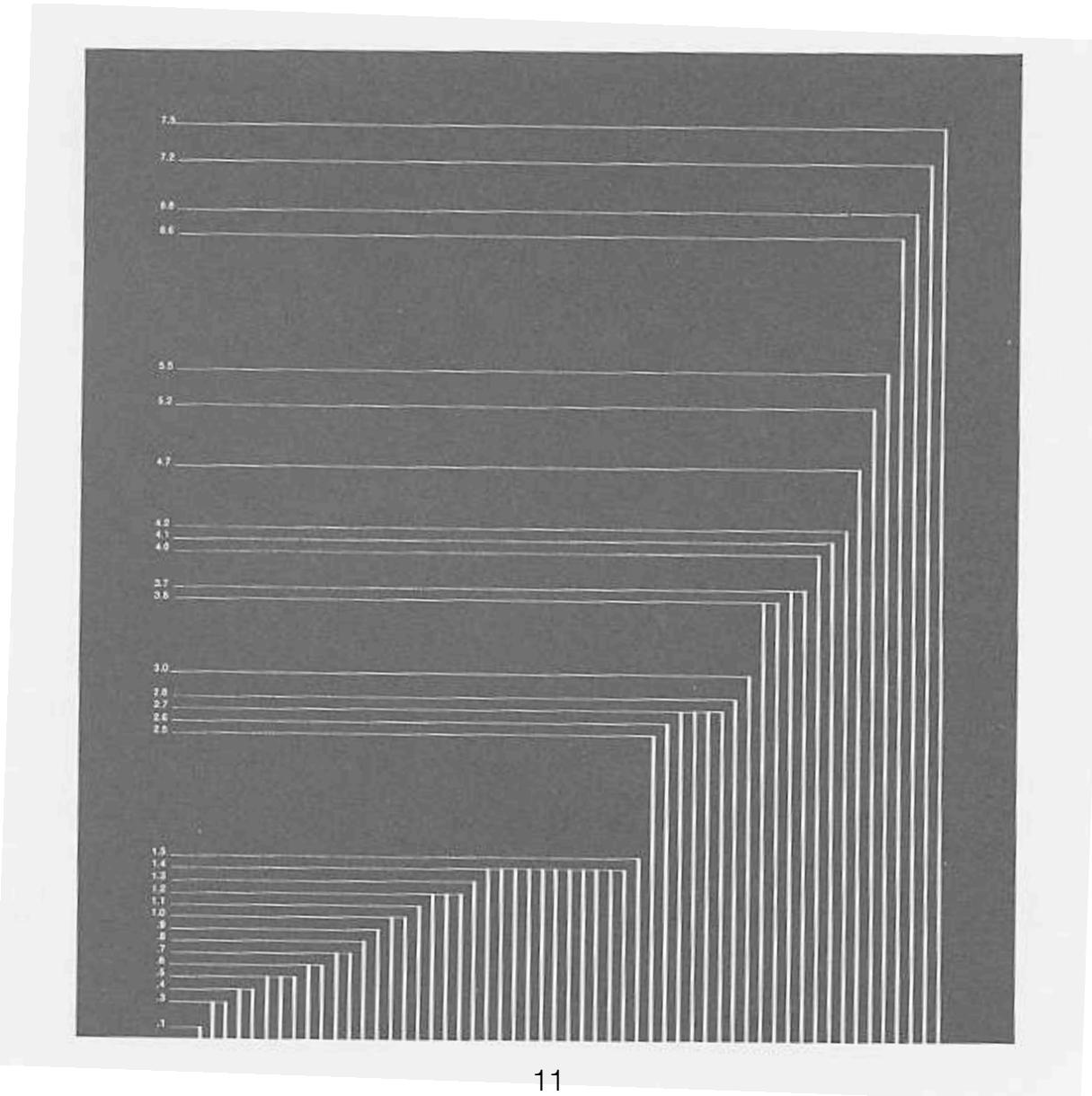
16 | 070355555588
18 | 00002223333333557777777888822335777888
20 | 00002223378800035778
22 | 0002335578023578
24 | 00228
26 | 23
28 | 080
30 | 7
32 | 2337
34 | 250077
36 | 0000823577
38 | 2333335582225577
40 | 0000003357788888002233555577778
42 | 03335555778800233333555577778
44 | 0222233555778000000023333357778888
46 | 0000233357700000023578
48 | 00000022335800333
50 | 0370
```

Old Faithful geysir

The numbers specify exact eruption times (minutes up to one decimal) and form a bar chart.

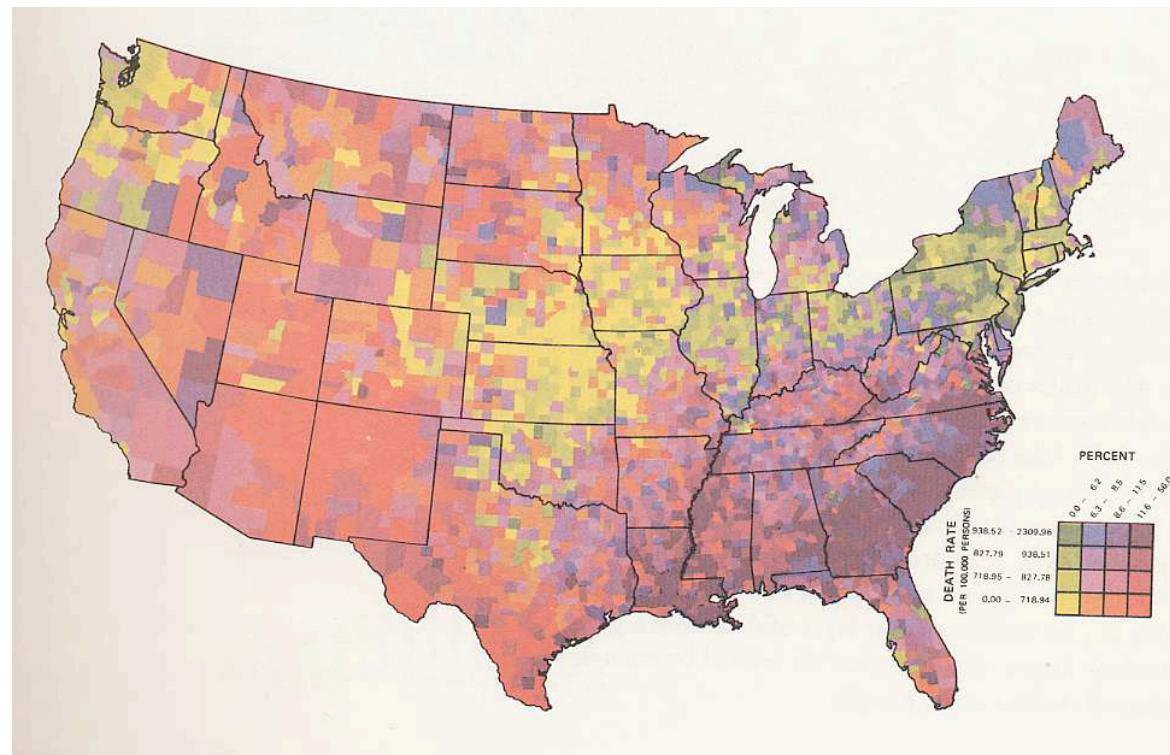


Data-based labels



Graphical puzzles

The complexity of multifunctioning elements may turn the data into visual puzzles.



P. Barabba, Alva L. Finker, 1978 [T 153].

This map must be interpreted through verbal rather than visual process.

Theory in practice

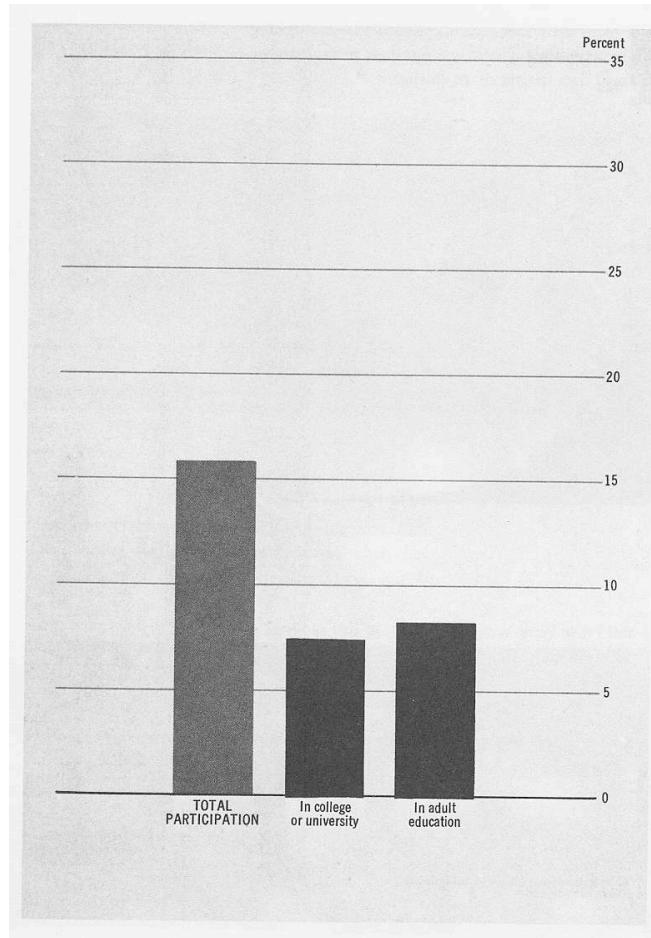
- Data-ink
- Chartjunk
- Multifunctioning graphical elements
- **Data density and small multiples**
- Aesthetics and techniques

Data density and small multiples

- Eye can distinguish patterns of about 10 (or even 60) cycles per degree
- In computer graphics, the resolution may be lower due to limitations in hardware (typical monitor at typical distances has a resolution of about 40 cycles per degree, 150 cycles per degree would be optimal)

$$\text{Data density} = \frac{\text{Number of entries in data matrix}}{\text{Area of graphics}}$$

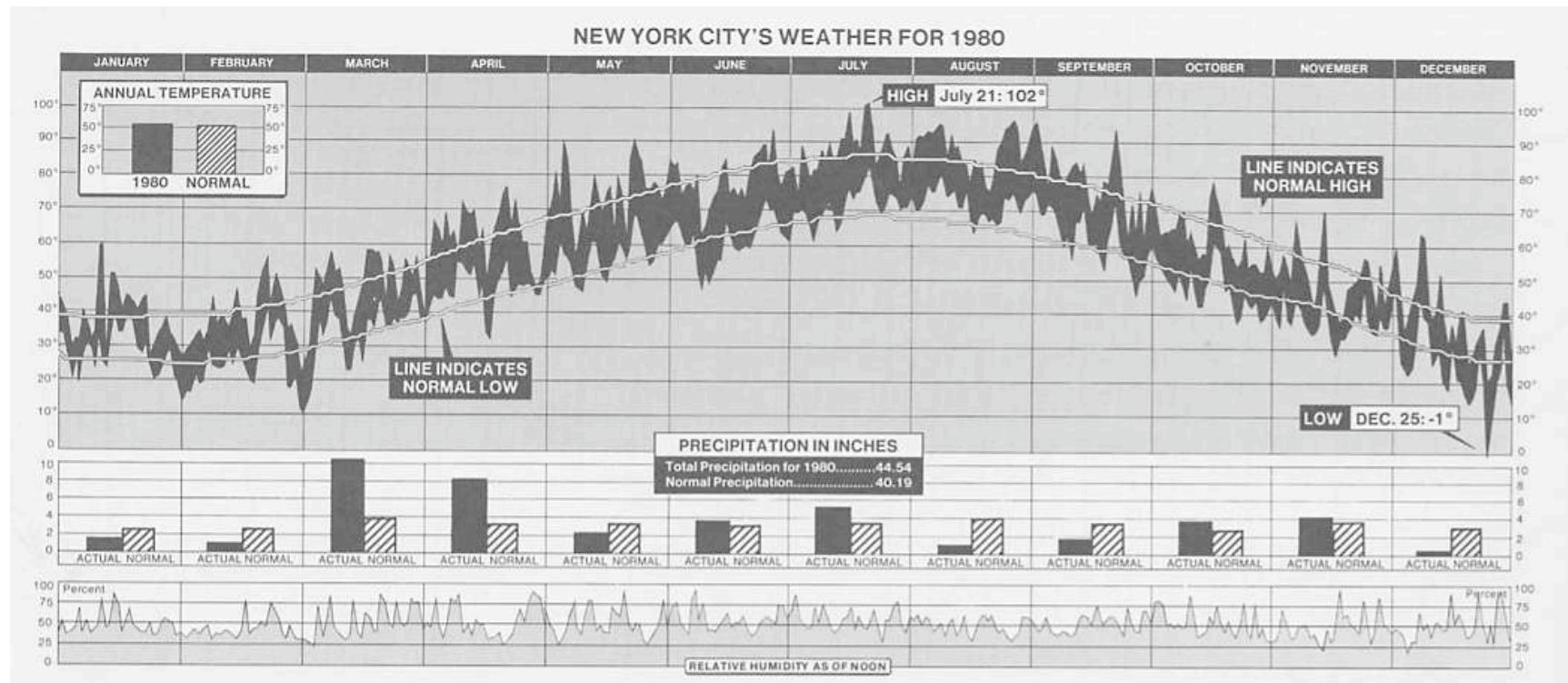
Low data density



Executive Office of the President, 1973 [T 163].

Data density = 0.02 numbers per cm^2 .

High data density

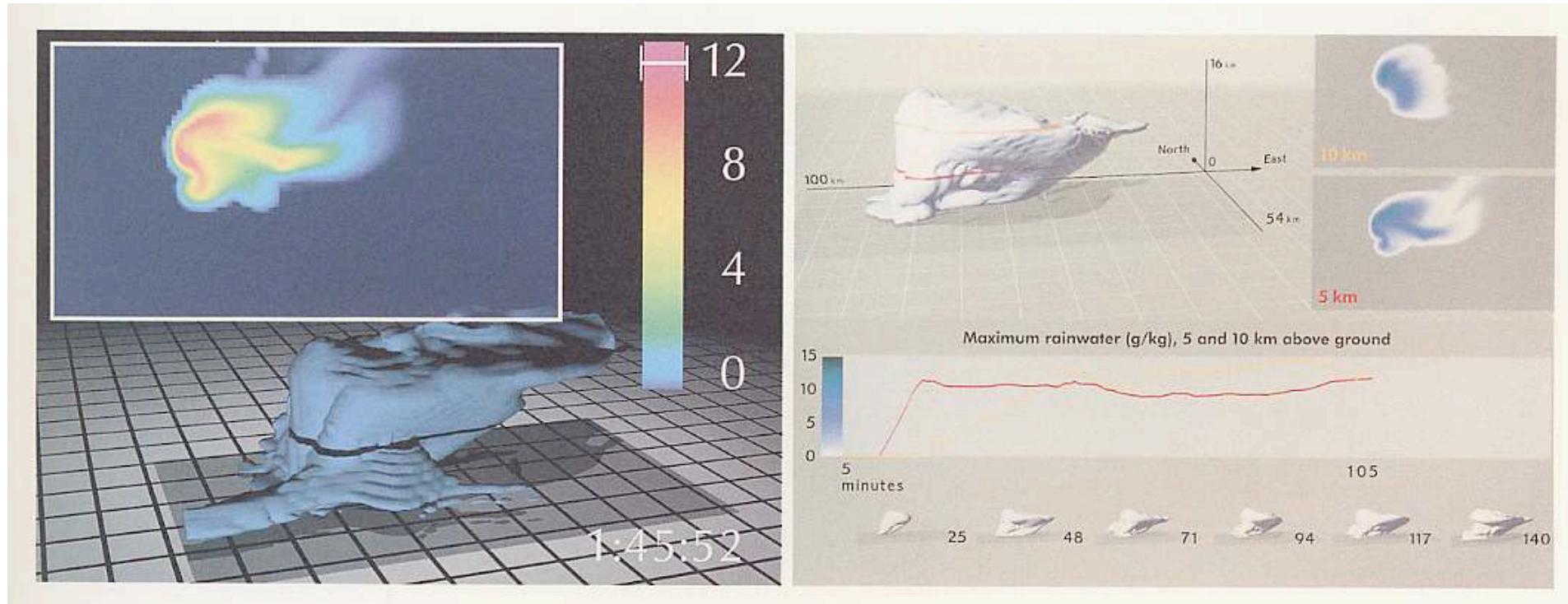


New York Times, 11 January 1981 [T 30].

Data density = 28 numbers per cm^2 .

Use small differences

Make all visual distinctions as subtle as possible, but still clear and effective.

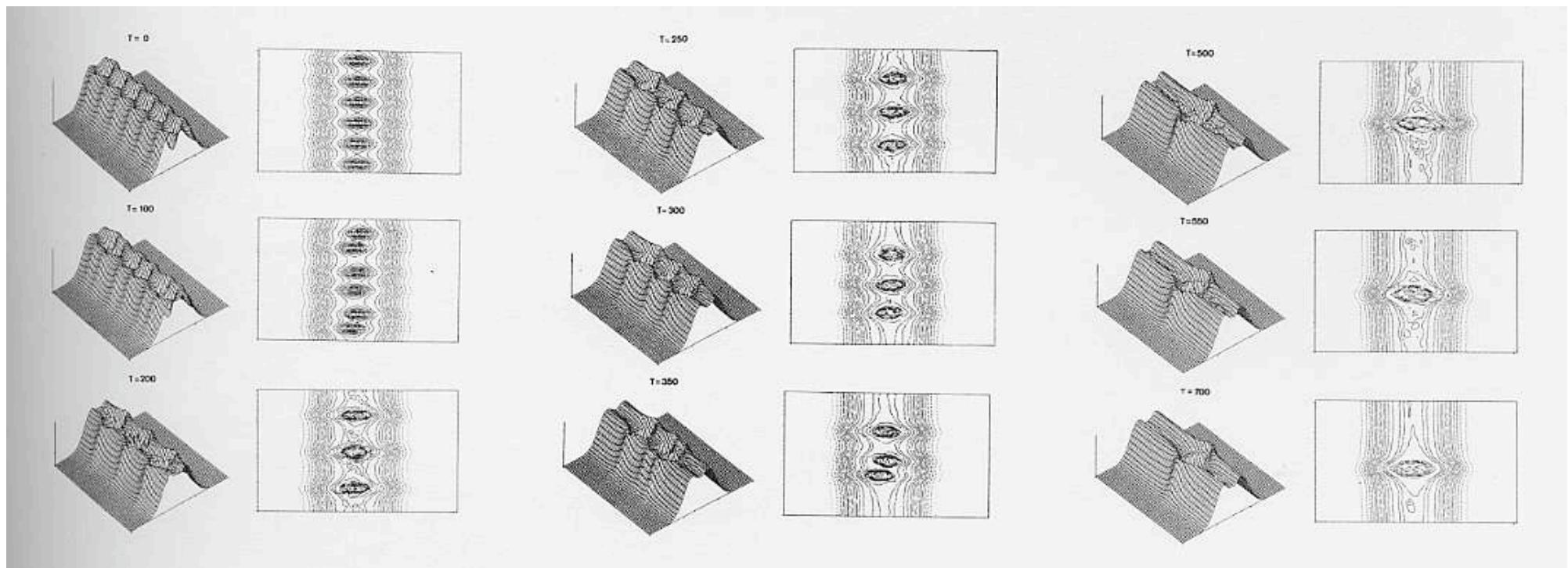


Matthew Arrot et al (original), E. R. Tufte, Polly Baker et al (revised) [VE 75].

Large distinctions generate clutter. Smaller distinctions highlight the data.

Using small multiples to make comparisons

- Comparisons must be positioned within the eye-span for the viewer to make comparisons at glance
- Show changes in data, not in design.



Graphs and tables

I. Cancer Survival Rates



What is the 15 year survival rate for brain and nervous system cancer?

Graphs and tables

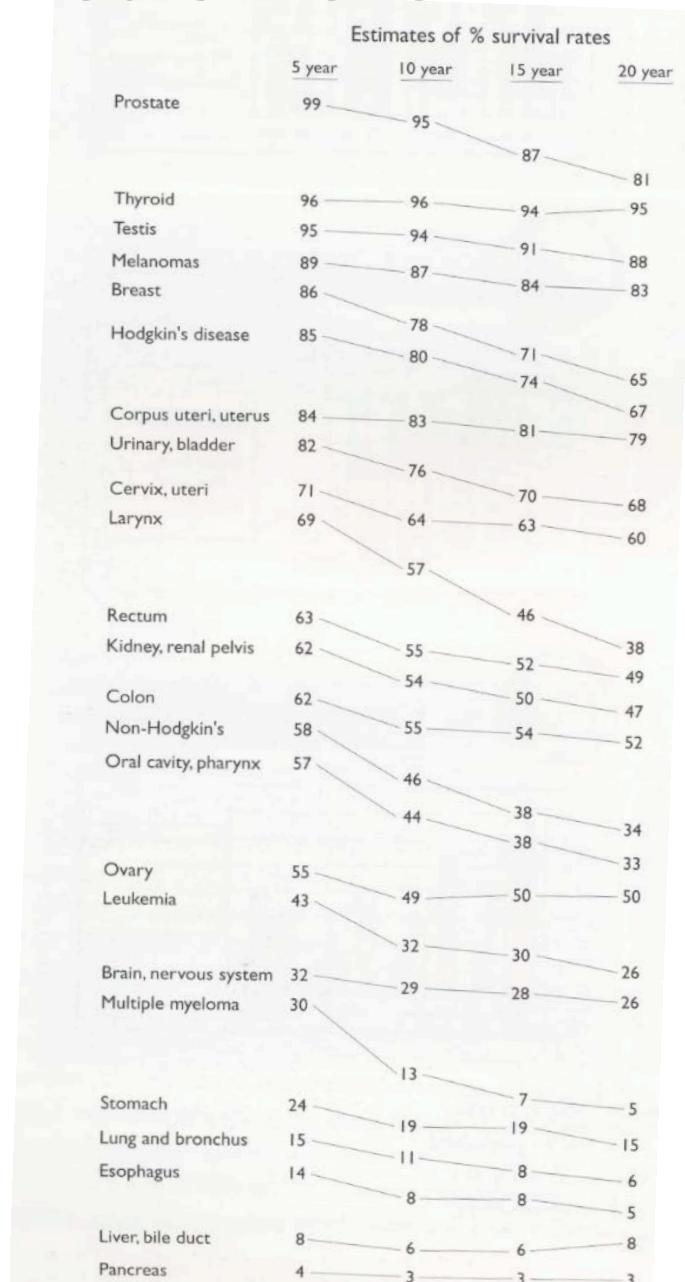
	Cancer survival rates (%)			
	5 year	10 year	15 year	20 year
Leukemia	42.5	32.4	29.7	26.2
Brain, nervous	32.0	29.2	27.6	26.1
Multiple myeloma	29.5	12.7	7.0	4.8
Stomach	23.8	19.4	19.0	14.9

What is the 15 year survival rate for brain and nervous system cancer?

What is the trend over the years?

Graphs and tables

Estimates of relative survival rates, by cancer site ¹²						
	% survival rates and their standard errors					
	5 year	10 year	15 year	20 year		
Prostate	98.8 0.4	95.2 0.9	87.1 1.7	81.1 3.0		
Thyroid	96.0 0.8	95.8 1.2	94.0 1.6	95.4 2.1		
Testis	94.7 1.1	94.0 1.3	91.1 1.8	88.2 2.3		
Melanomas	89.0 0.8	86.7 1.1	83.5 1.5	82.8 1.9		
Breast	86.4 0.4	78.3 0.6	71.3 0.7	65.0 1.0		
Hodgkin's disease	85.1 1.7	79.8 2.0	73.8 2.4	67.1 2.8		
Corpus uteri, uterus	84.3 1.0	83.2 1.3	80.8 1.7	79.2 2.0		
Urinary, bladder	82.1 1.0	76.2 1.4	70.3 1.9	67.9 2.4		
Cervix, uteri	70.5 1.6	64.1 1.8	62.8 2.1	60.0 2.4		
Larynx	68.8 2.1	56.7 2.5	45.8 2.8	37.8 3.1		
Rectum	62.6 1.2	55.2 1.4	51.8 1.8	49.2 2.3		
Kidney, renal pelvis	61.8 1.3	54.4 1.6	49.8 2.0	47.3 2.6		
Colon	61.7 0.8	55.4 1.0	53.9 1.2	52.3 1.6		
Non-Hodgkin's	57.8 1.0	46.3 1.2	38.3 1.4	34.3 1.7		
Oral cavity, pharynx	56.7 1.3	44.2 1.4	37.5 1.6	33.0 1.8		
Ovary	55.0 1.3	49.3 1.6	49.9 1.9	49.6 2.4		
Leukemia	42.5 1.2	32.4 1.3	29.7 1.5	26.2 1.7		
Brain, nervous system	32.0 1.4	29.2 1.5	27.6 1.6	26.1 1.9		
Multiple myeloma	29.5 1.6	12.7 1.5	7.0 1.3	4.8 1.5		
Stomach	23.8 1.3	19.4 1.4	19.0 1.7	14.9 1.9		
Lung and bronchus	15.0 0.4	10.6 0.4	8.1 0.4	6.5 0.4		
Esophagus	14.2 1.4	7.9 1.3	7.7 1.6	5.4 2.0		
Liver, bile duct	7.5 1.1	5.8 1.2	6.3 1.5	7.6 2.0		
Pancreas	4.0 0.5	3.0 1.5	2.7 0.6	2.7 0.8		



What is the 15 year survival rate for brain and nervous system cancer?

Graphs and tables

- The data can be shown in
 - sentences,
 - tables or
 - graphics.
- Table is usually the best choice for (small) collection of numbers

Some Winners and Losers in the Forecasting Game					
Council of Economic Advisors: +4.7%					Chase Econometrics: 7.4%
Data Resources: +4.5%					Wharton Econometric Forecasting: 6.8%
Nat. Assoc. of Business Economists: +4.5%					Conference Board: 6.7%
Wharton Econometric Forecasting: +4.5%					Nat. Assoc. of Business Economists: 6.7%
Congressional Budget Office: +4.4%					I.B.M. Economics Department: 6.6%
Conference Board: +4.2%					Data Resources: 6.5%
I.B.M. Economics Department: +4.1%					Congressional Budget Office: 6.3%
Real G.N.P. Growth: +3.8%					Council of Economic Advisors: 6.3%
Industrial Production Growth: +5.8%					Unemployment Rate: 6%
Change in Consumer Prices: +7.7%					
Corporate Profits Growth: +13.3%					
Chase Econometrics: +2.8%					
Conference Board: +5.5%					
I.B.M. Economics Department: +6.6%					
Data Resources: +10.5%					
Nat. Assoc. of Business Economists: +6.5%					
I.B.M. Economics Department: +10.4%					
Wharton Econometric Forecasting: +4.8%					
Conference Board: +8.2%					
Chase Econometrics: +6.5%					
Chase Econometrics: +1.9%					
Data Resources: +8.2%					
Chase Econometrics: +5.9%					
Council of Economic Advisors: +5.9%					
Wharton Econometric Forecasting: +5.4%					
<i>Forecasters are not listed in categories for which they did not make a prediction.</i>					
<small>*After taxes</small>					

New York Times, 2 January 1979 [T 180].

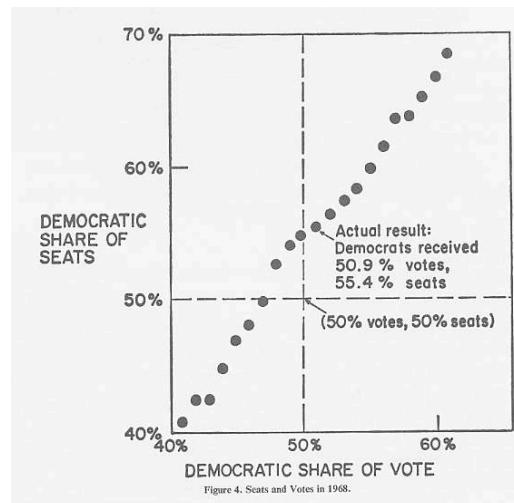
Nearly 53 % of group A did something compared to 46 % of group B and 57 % of C.	Same using a table:	Better(?) order:												
	<table border="1"> <tr> <td>Group A</td><td>53 %</td></tr> <tr> <td>Group B</td><td>46 %</td></tr> <tr> <td>Group C</td><td>57 %</td></tr> </table>	Group A	53 %	Group B	46 %	Group C	57 %	<table border="1"> <tr> <td>Group B</td><td>46 %</td></tr> <tr> <td>Group A</td><td>53 %</td></tr> <tr> <td>Group C</td><td>57 %</td></tr> </table>	Group B	46 %	Group A	53 %	Group C	57 %
Group A	53 %													
Group B	46 %													
Group C	57 %													
Group B	46 %													
Group A	53 %													
Group C	57 %													

Theory in practice

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

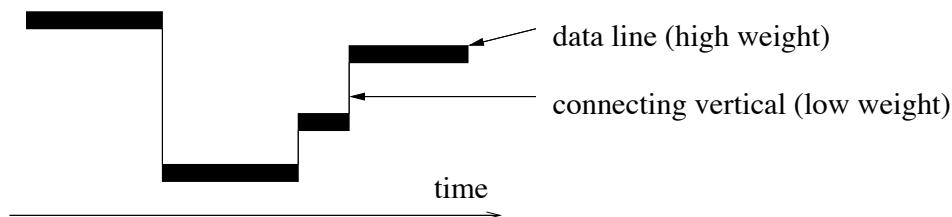
Line weight and lettering

The weights of the letters should be in proportion to the other visual elements:



E. R. Tufte, 1973 [T 184].

The heavier weight should be given to data measures:



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.

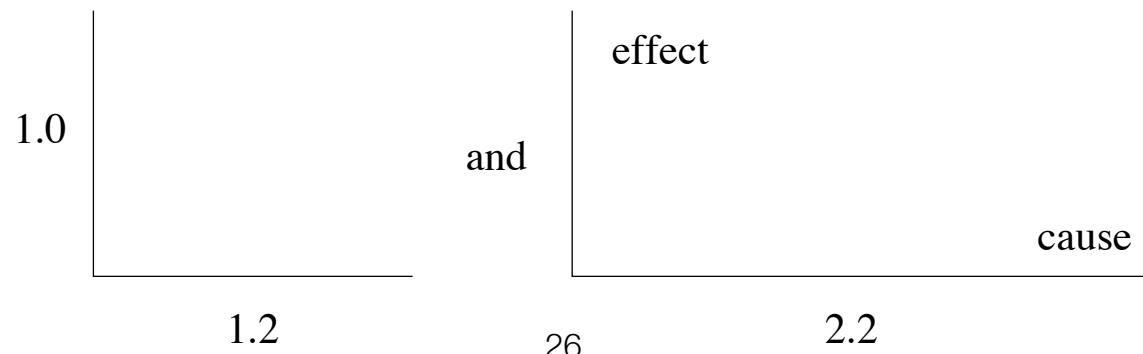
CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS							
CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			X
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			
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			
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

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.

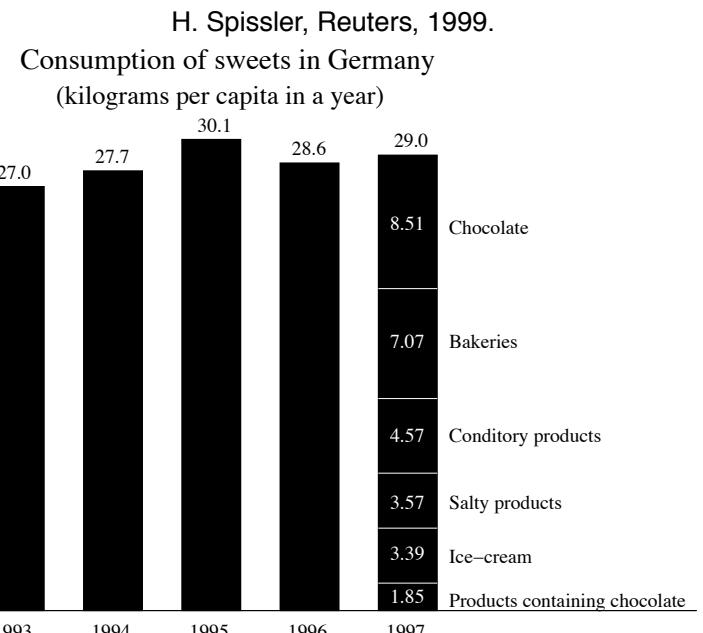


Aesthetics in decision making

- Good aesthetics makes a positive impression
 - Don Norman (2002). *Emotion & Design: Attractive Things Work Better*
- Exaggerating leads to chartjunk
 - e.g. Powerpoint style

Conclusion

- Communicating ideas and information is difficult. Wio's laws:
 - Communication usually fails, except by accident.
 - If a message can be interpreted in several ways, it will be interpreted in a manner that maximizes damages.
- Theory of data graphics: show only the essential in a way that makes the facts obvious. Don't waste space – eliminate all non-essentials and redundancies



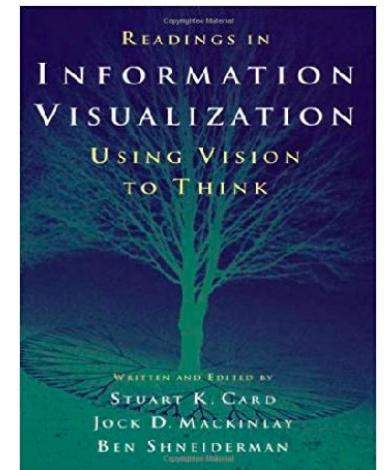
NEW TOPIC IN THIS LECTURE

Presenting statistics

- Techniques
- Problems
- Scenarios

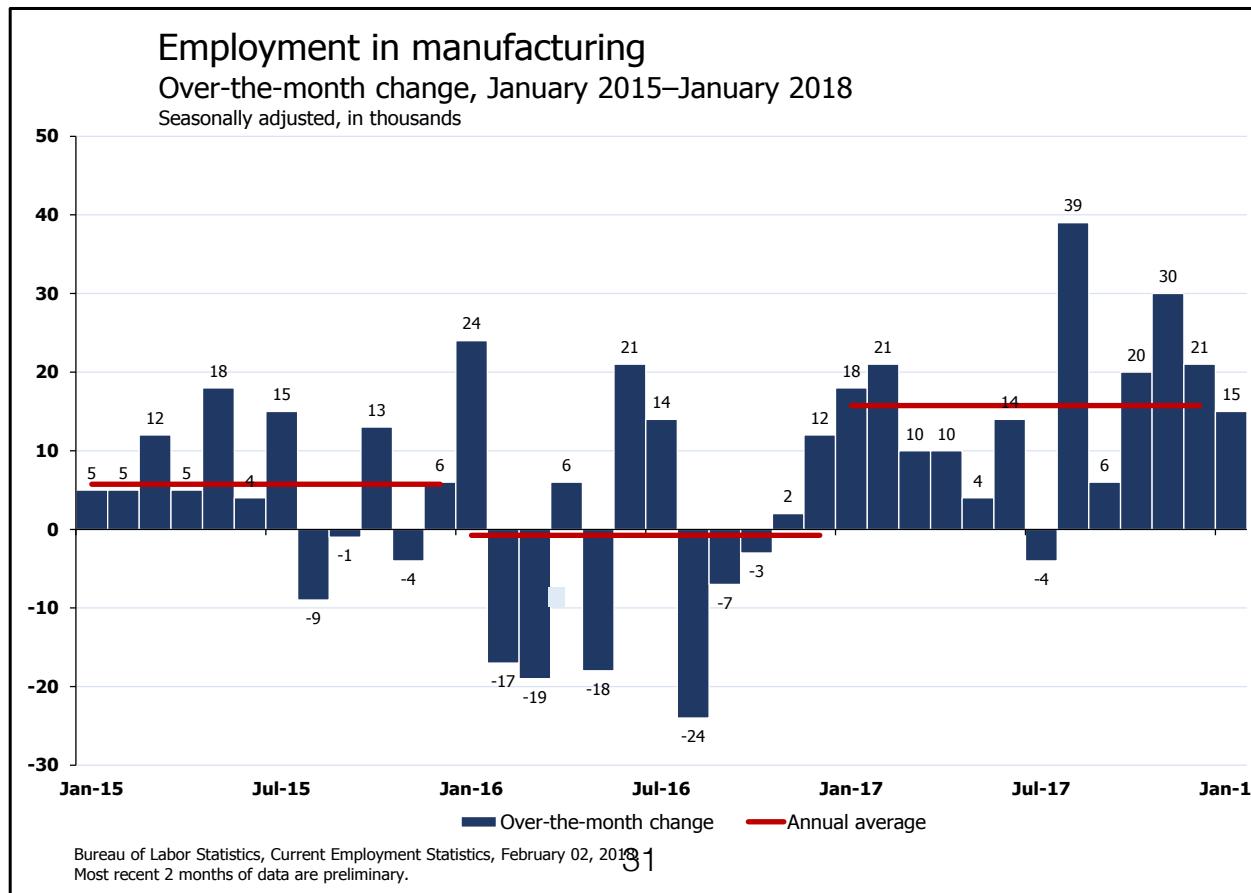
Presenting statistics: Outline

- **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.



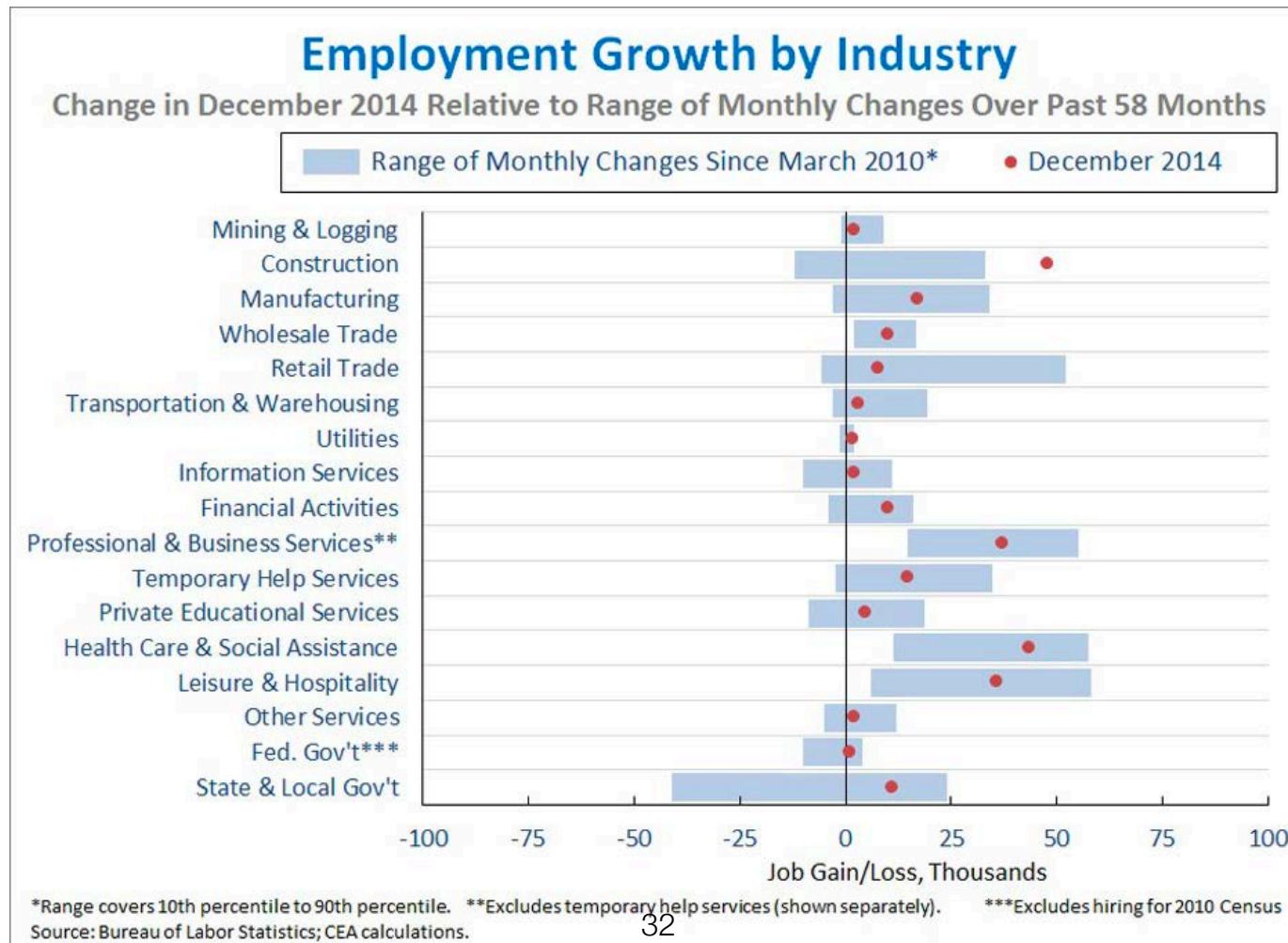
Basic elements: bars

- 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



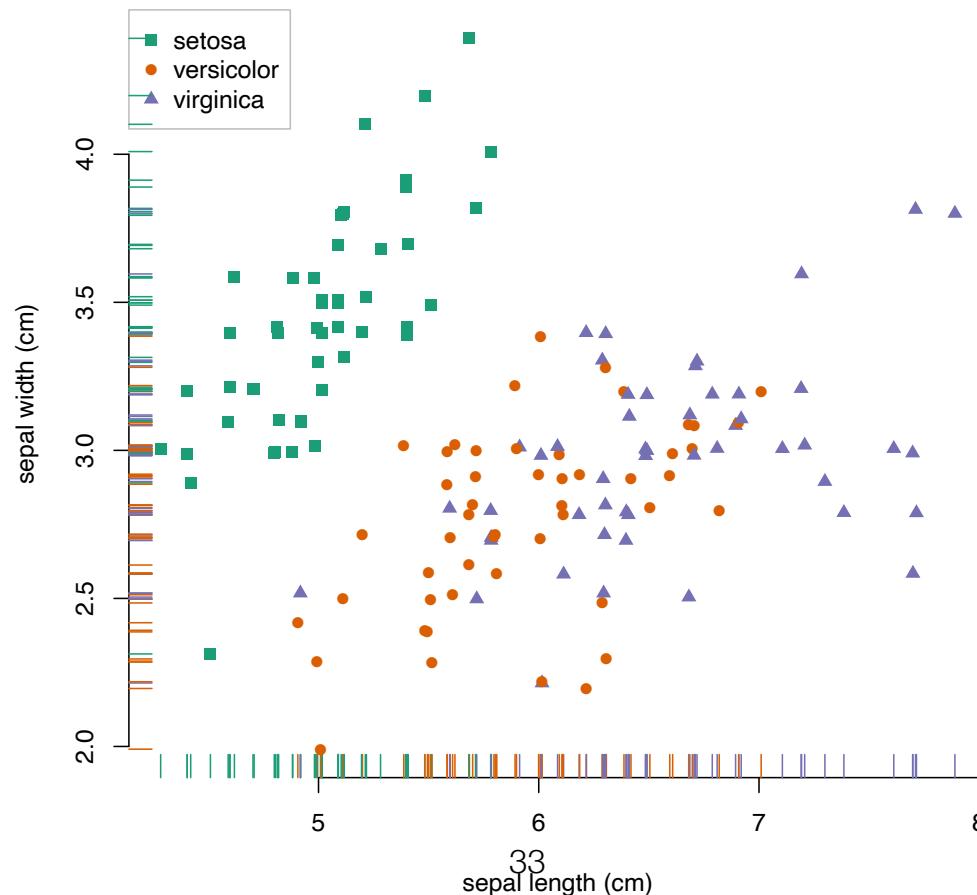
Basic elements: boxes

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



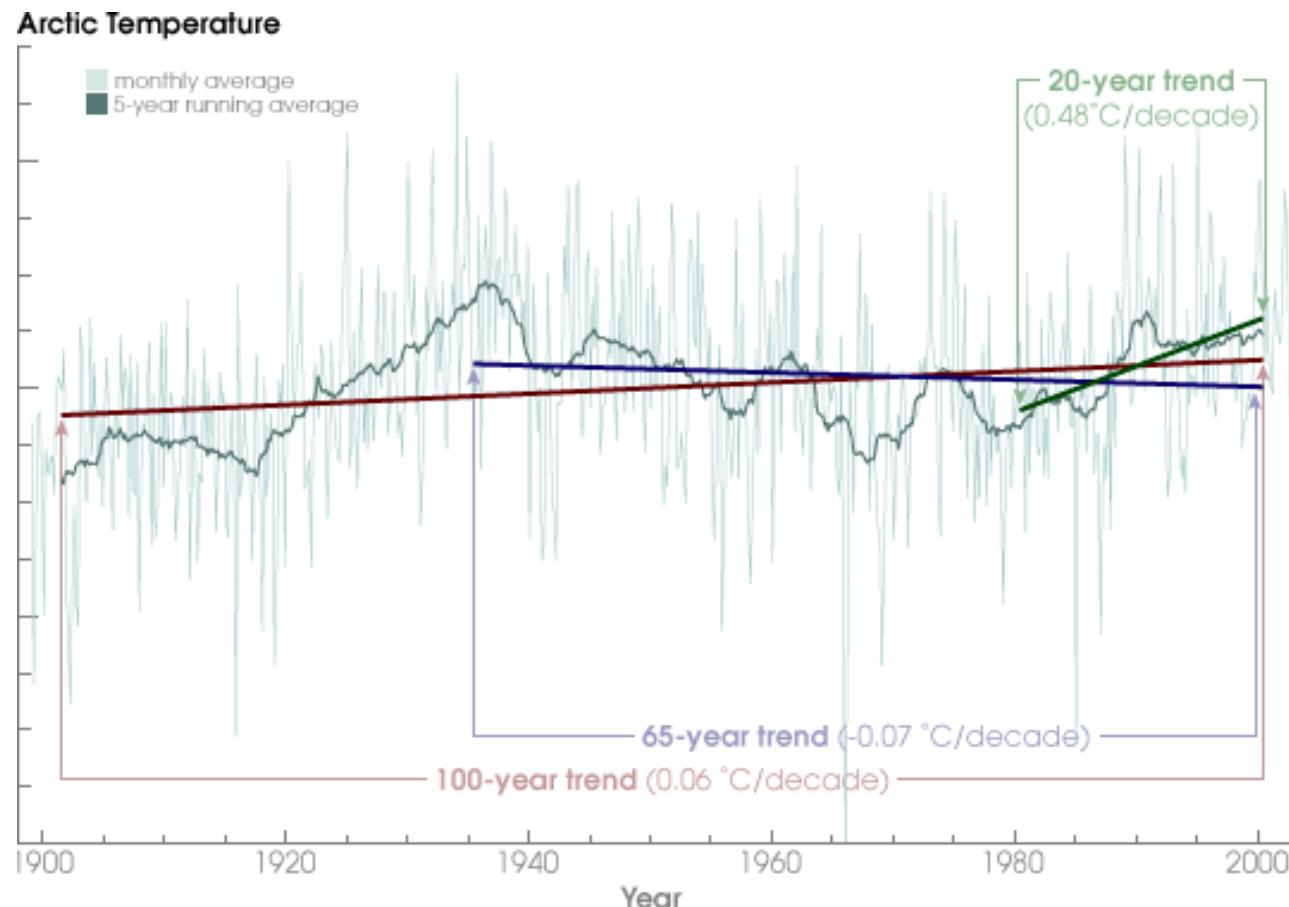
Basic elements: points

- 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 to encode multiple dimensions



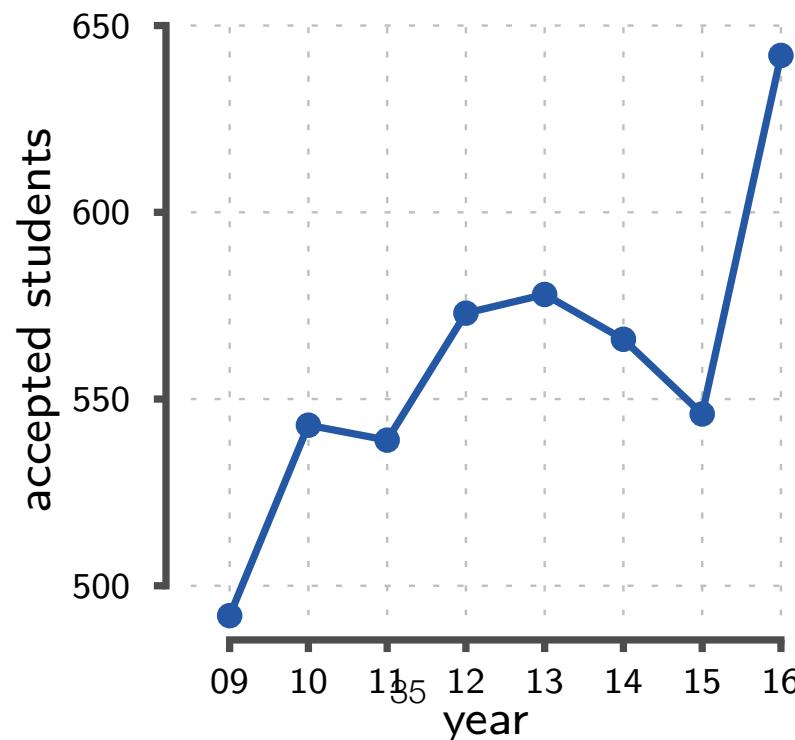
Basic elements: lines

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



When can we use line charts

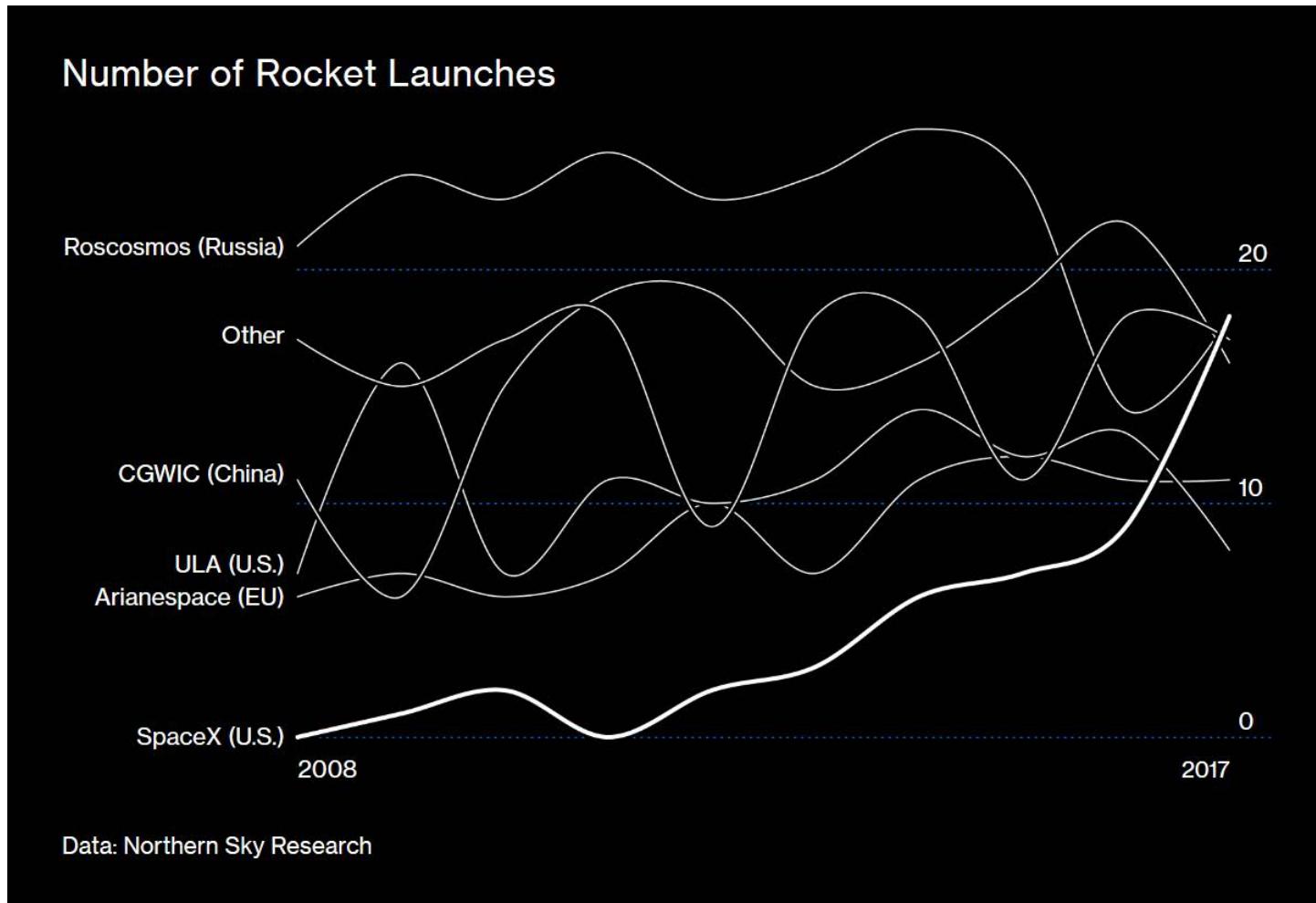
- Line charts are **primarily** meant for **continuous data** on the x-axis but
 - they *can also* be used with discrete data on the x-axis
 - you *can even use* it with categorical data on the x-axis, if it is ordered in a meaningful way



When should we not use line charts

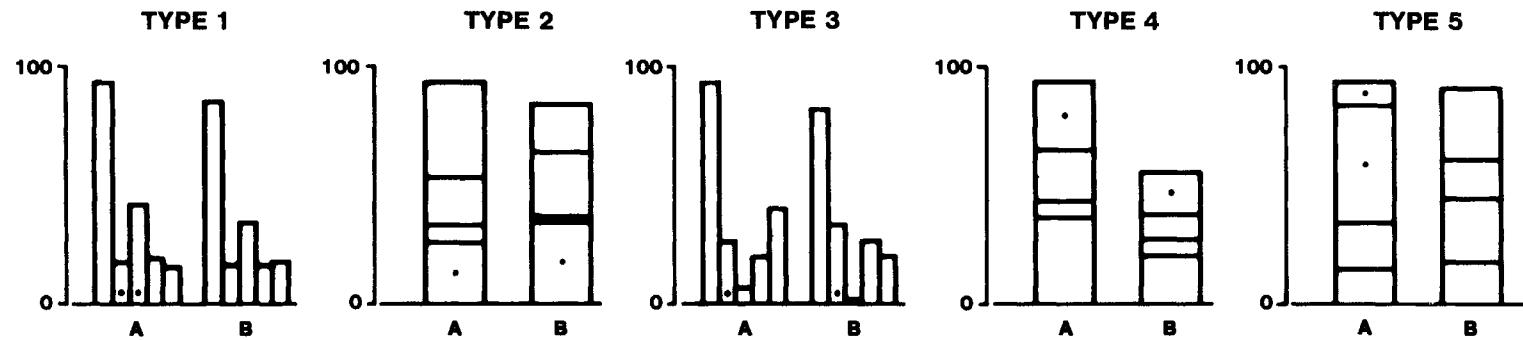
- You should be careful with line charts if
 - you have **gaps** in your data
 - line chart suggests that the **missing data** can be obtained via linear **interpolation** (this may not be true)
 - at minimum, you should indicate the actual data points with markers (e.g., dots)
- line chart doesn't make sense if there is no meaningful order on x-axis

When should we not use line charts



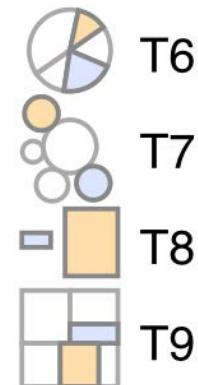
- integer/discrete data should not be interpolated with curves

Which chart type is most effective?

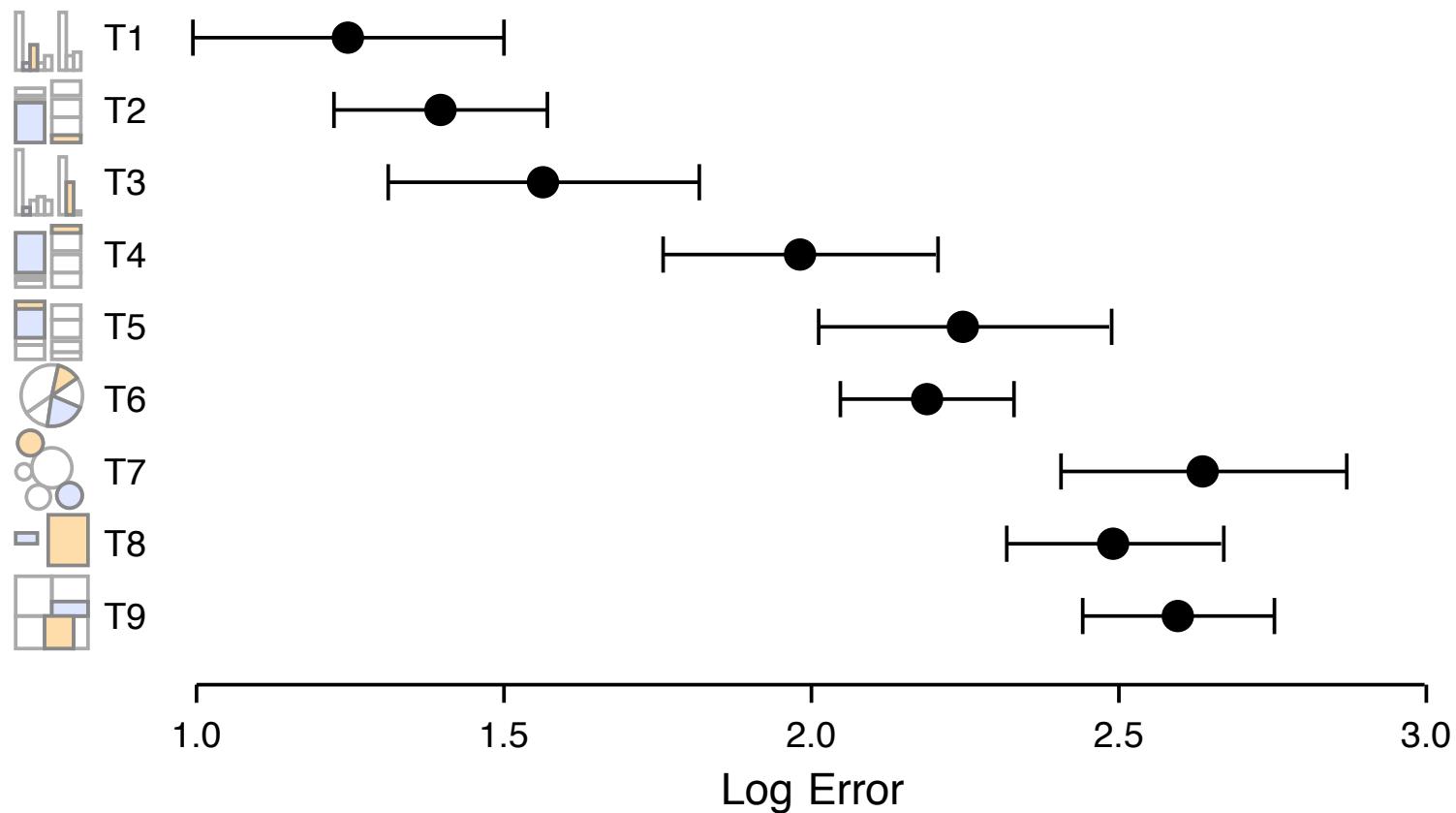


T1-T5

- 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

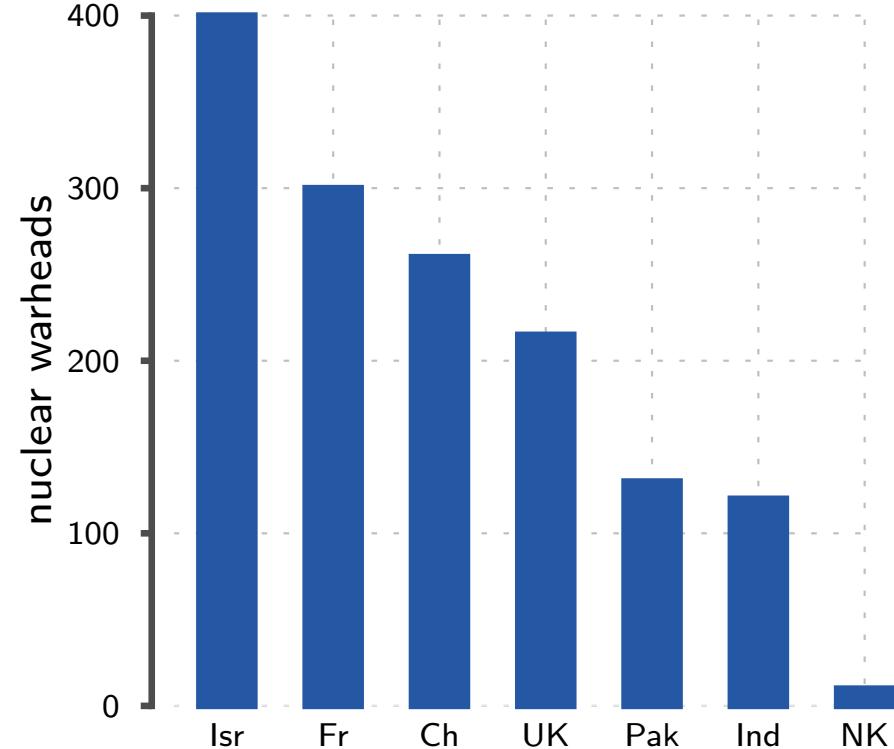
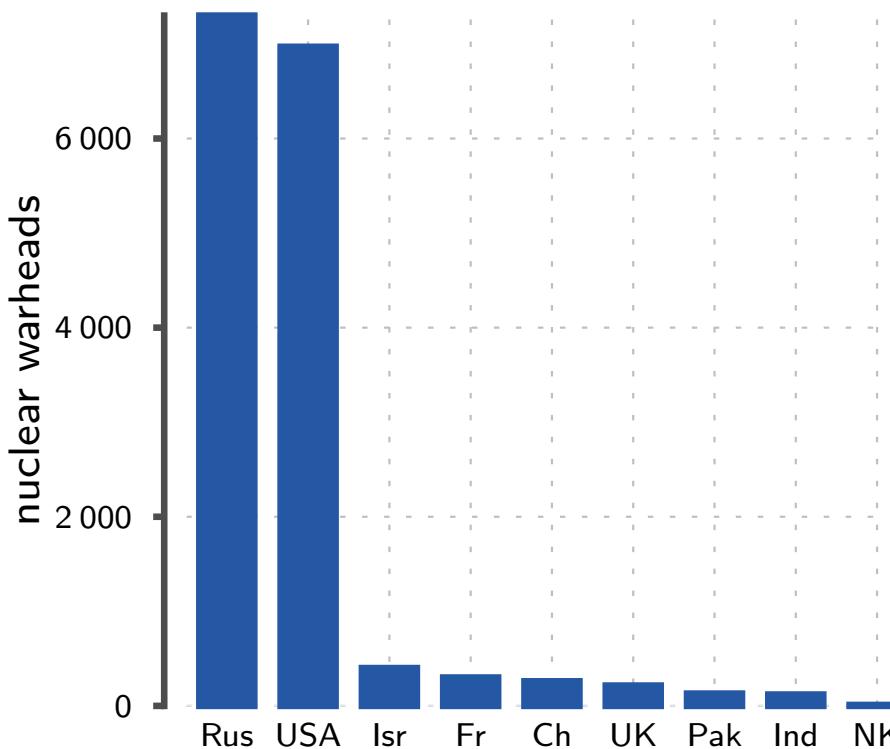
$$\text{log error} = \log_2(|\text{estimated difference} - \text{true difference}| + 1/8)$$

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)

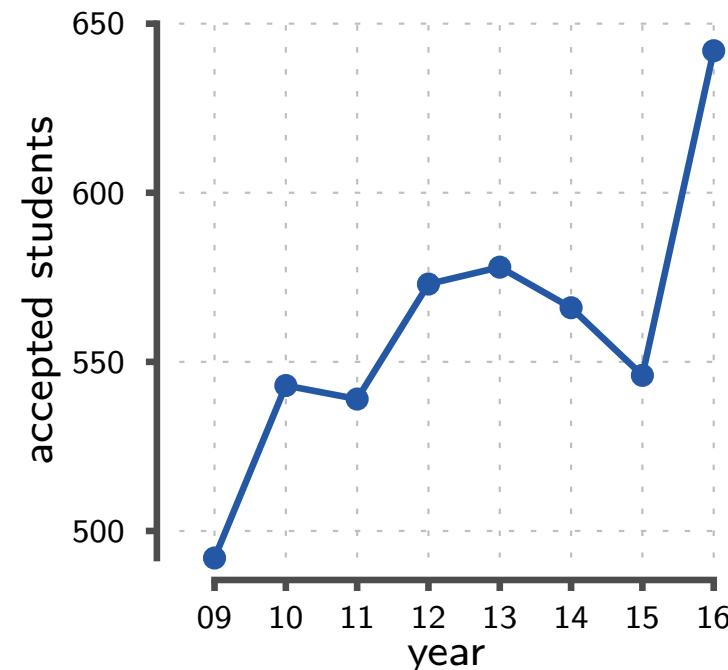
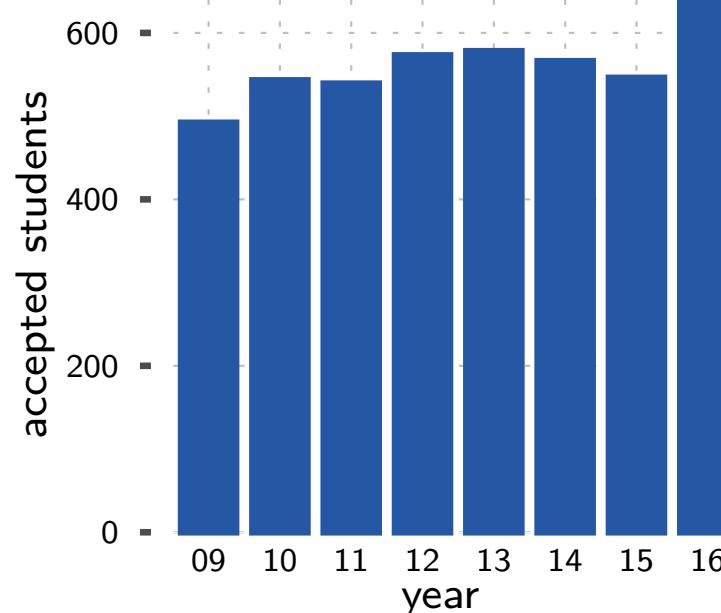
Overview / detail

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



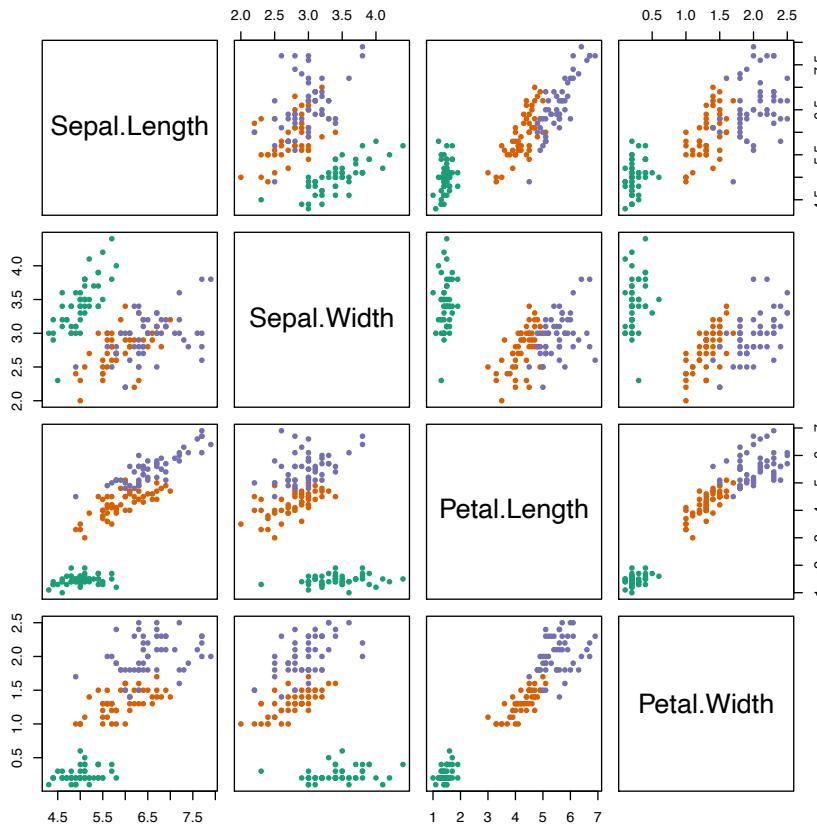
Multiform

- 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)

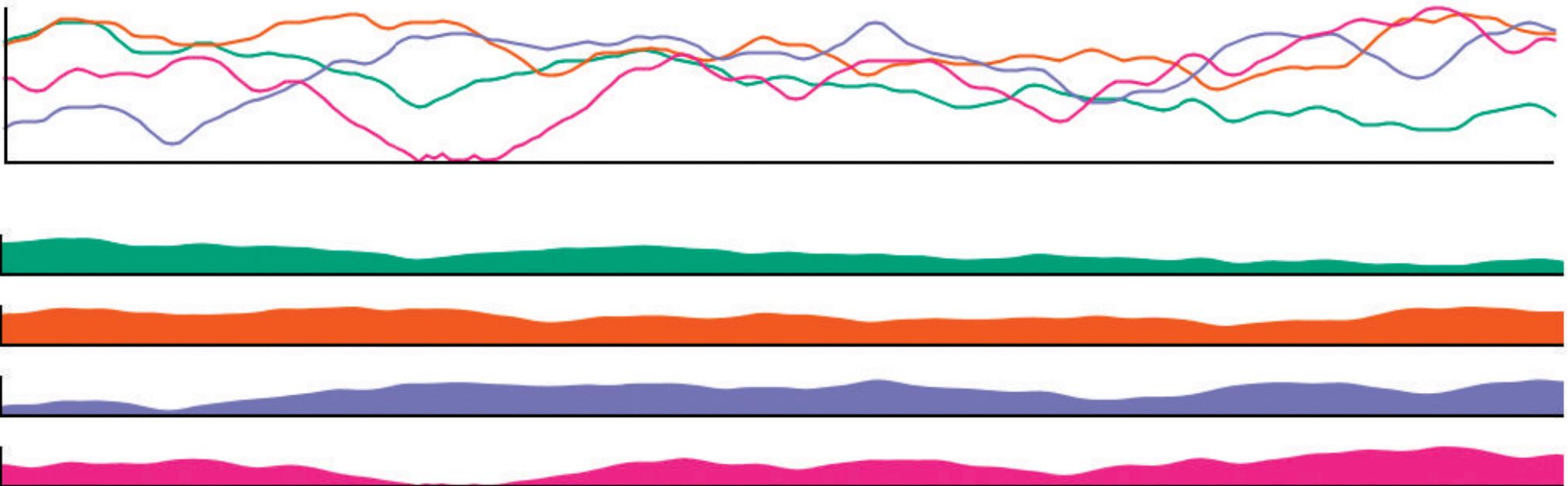
- multiple plots with the same design
- show different parts of the data
- can be arranged into a matrix or other array



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

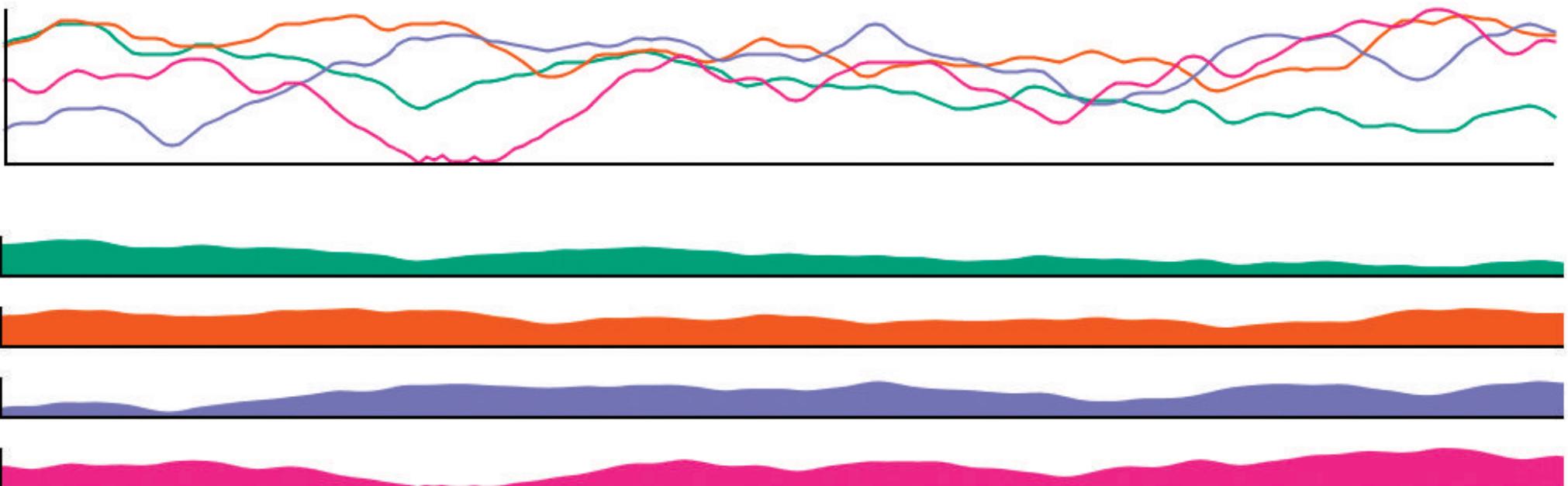
Superimposition vs. small multiples

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



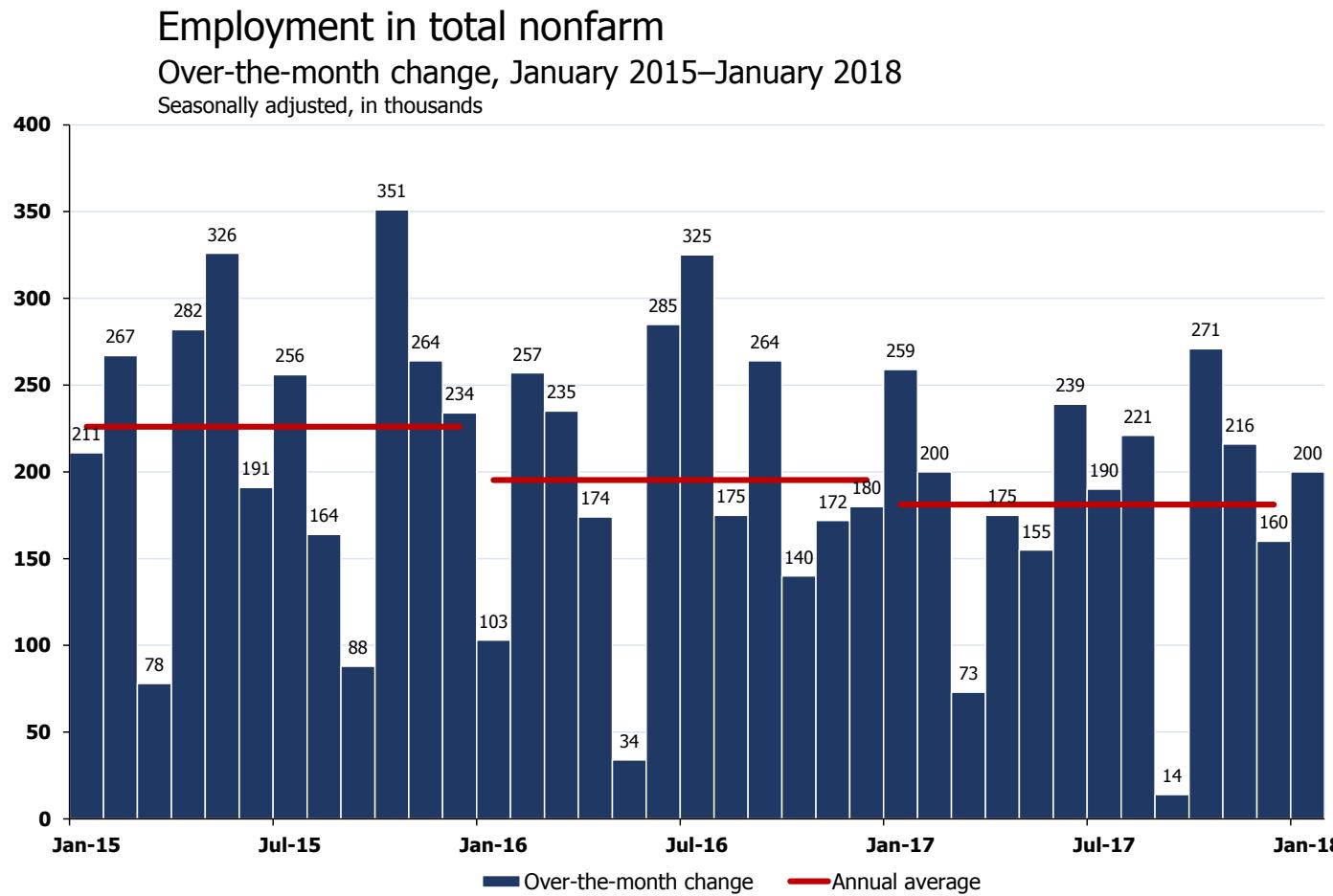
Superimposition vs. small multiples

- (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

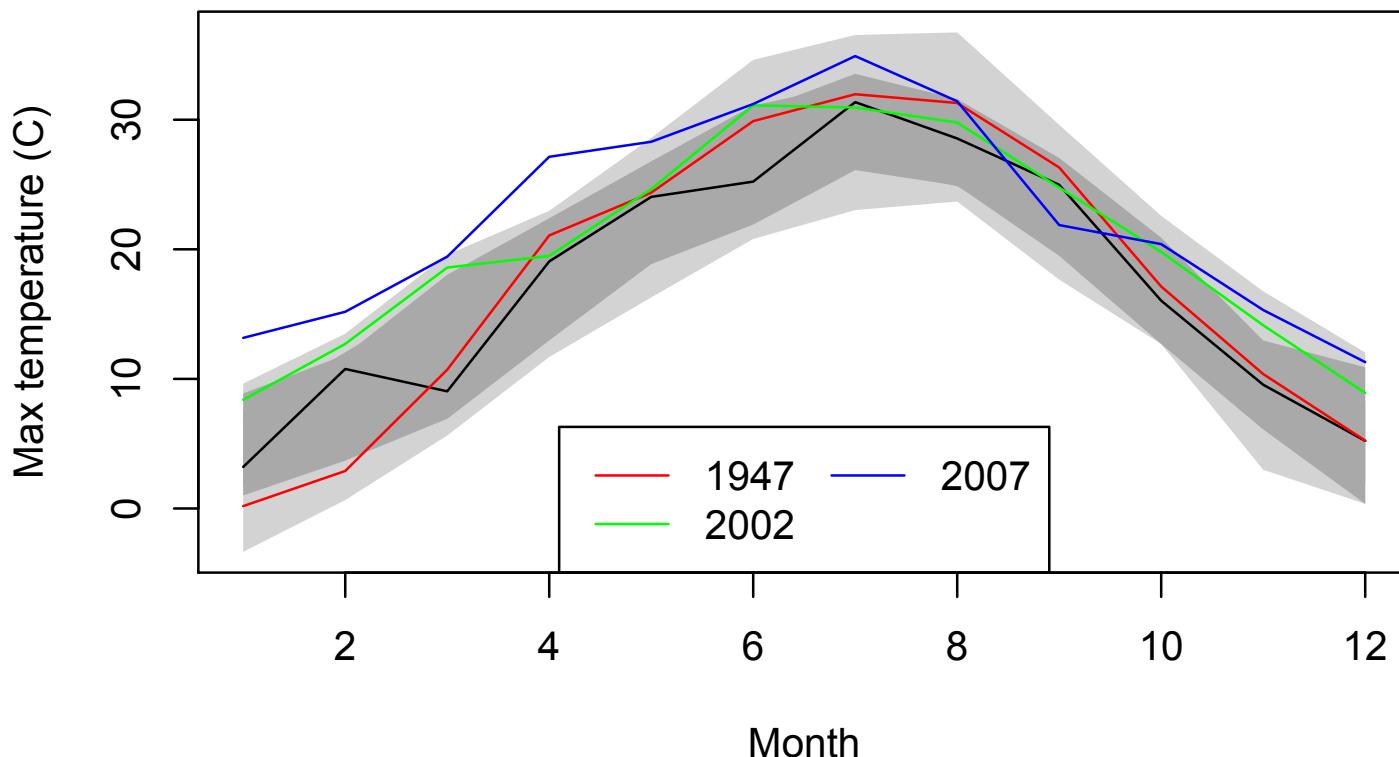


Bureau of Labor Statistics, Current Employment Statistics, February 02, 2018.

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.

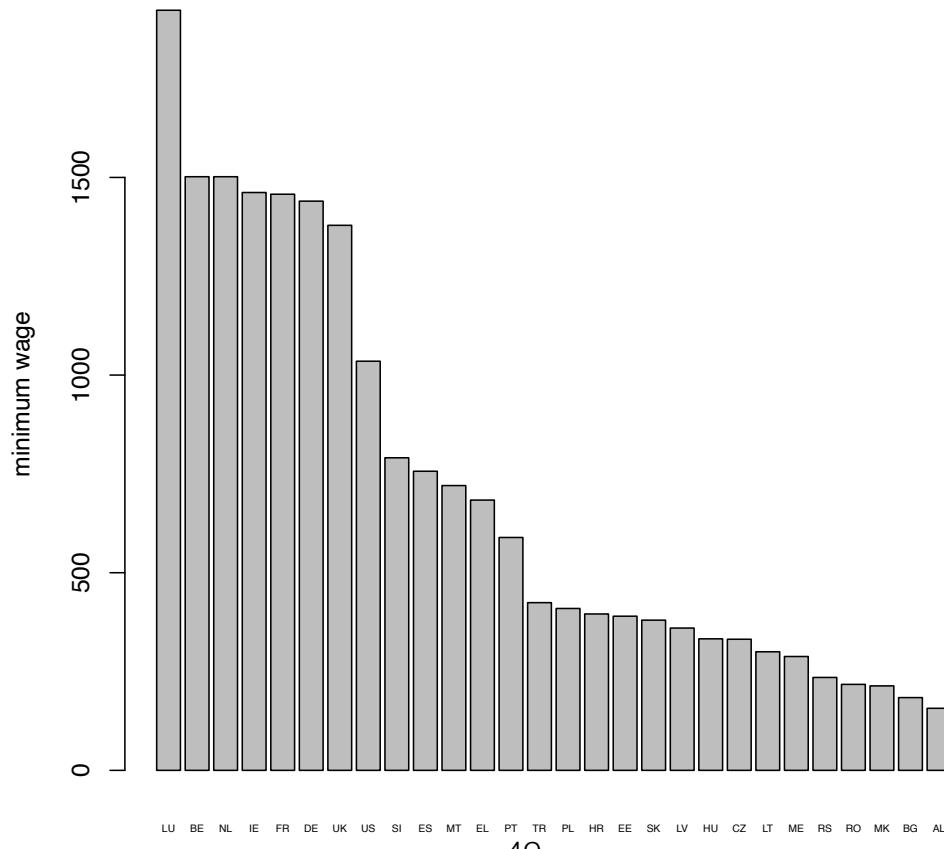


Re-expression

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

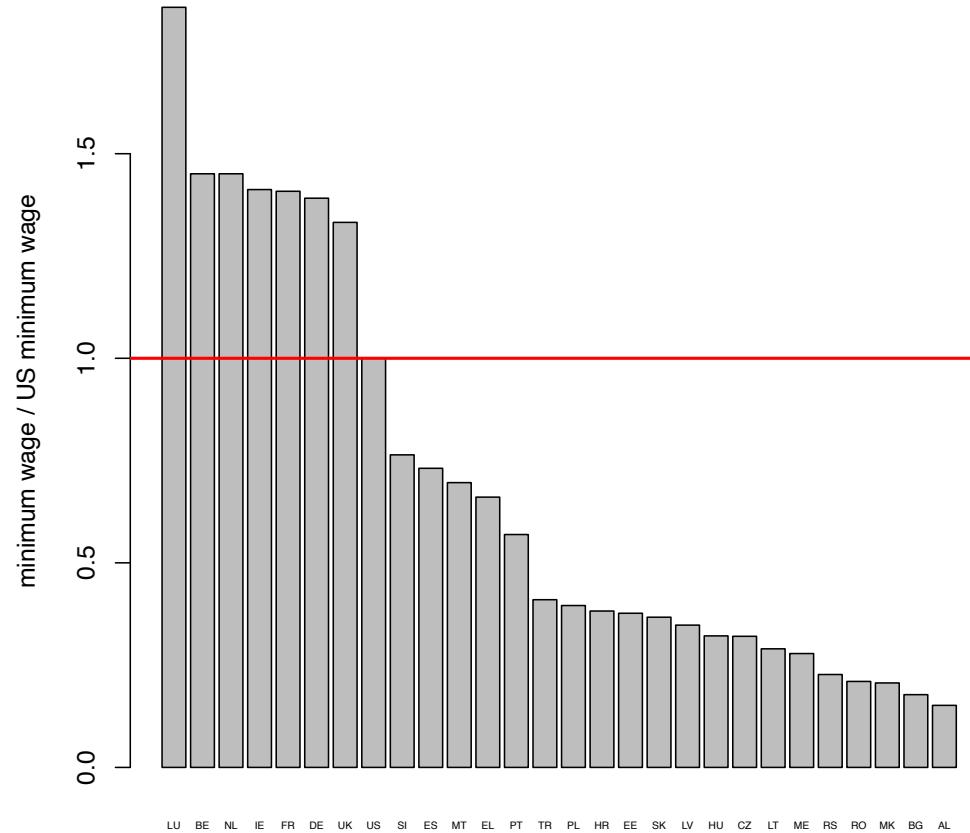
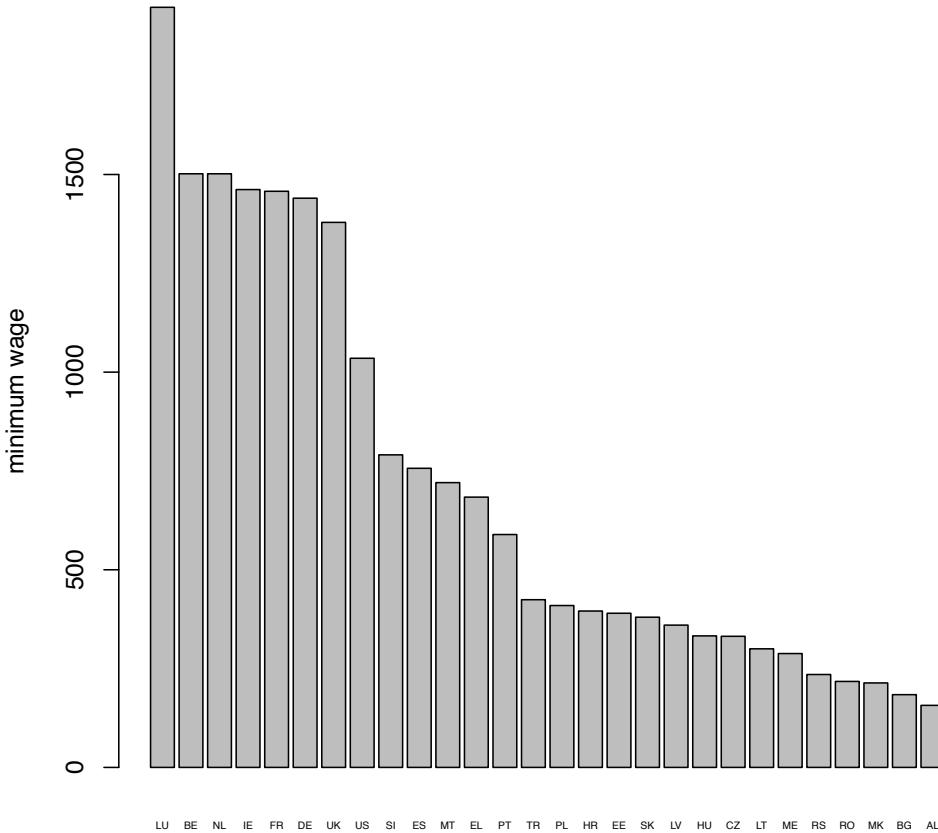
Re-expression

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



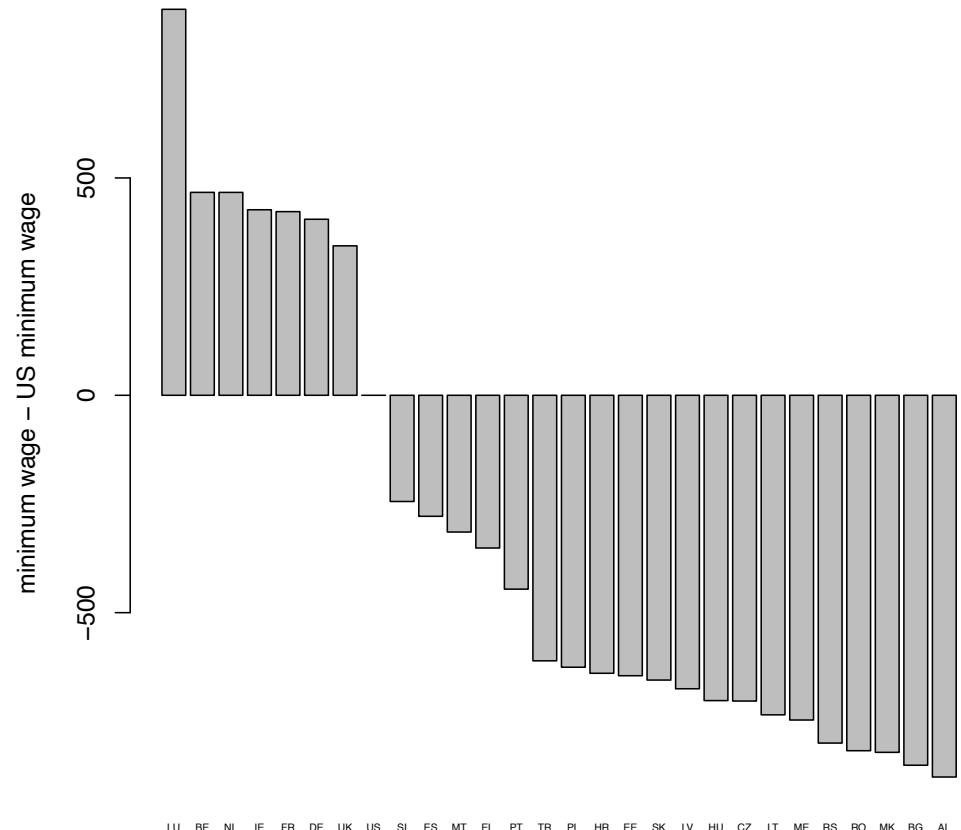
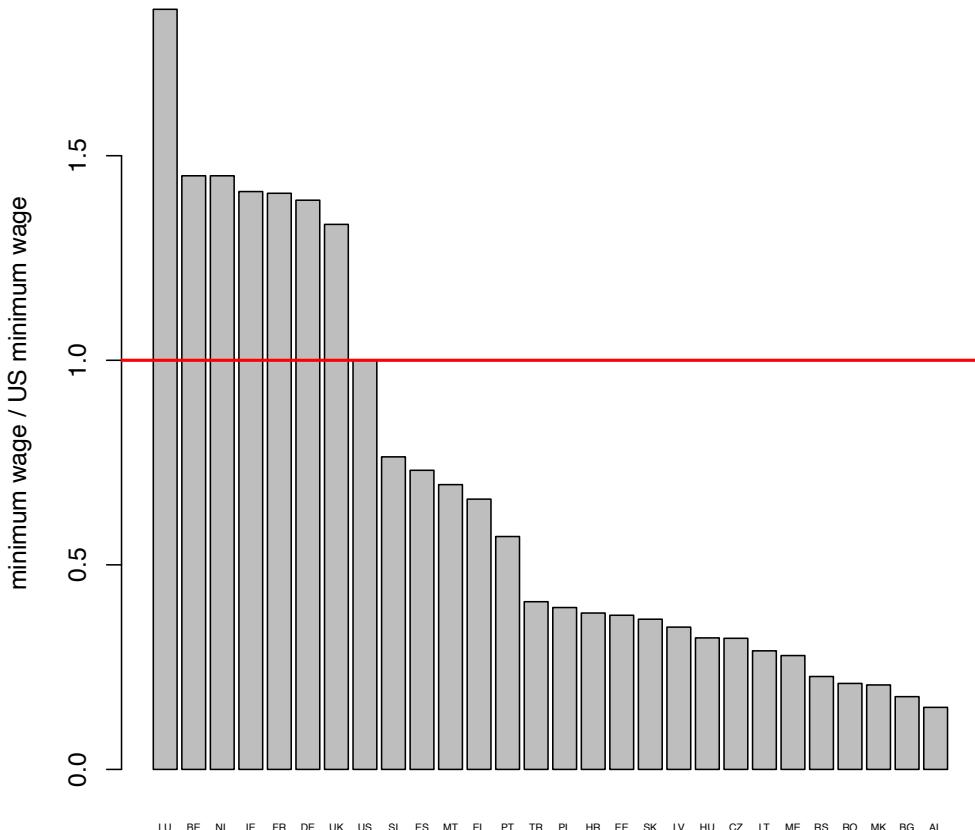
Re-expression

- 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



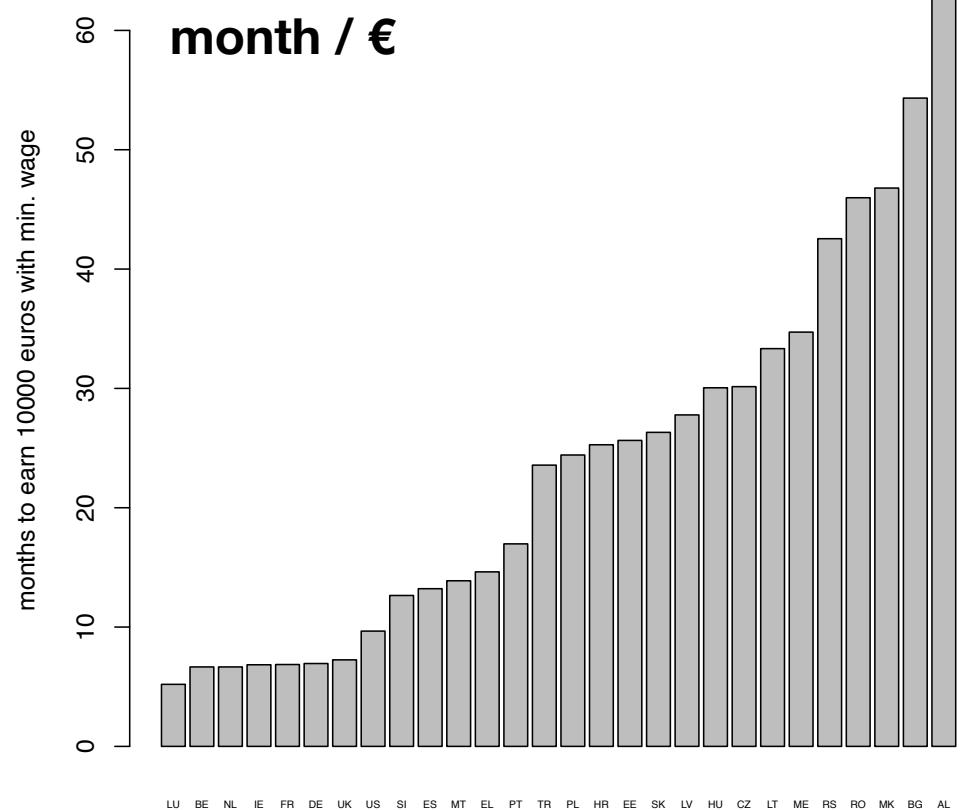
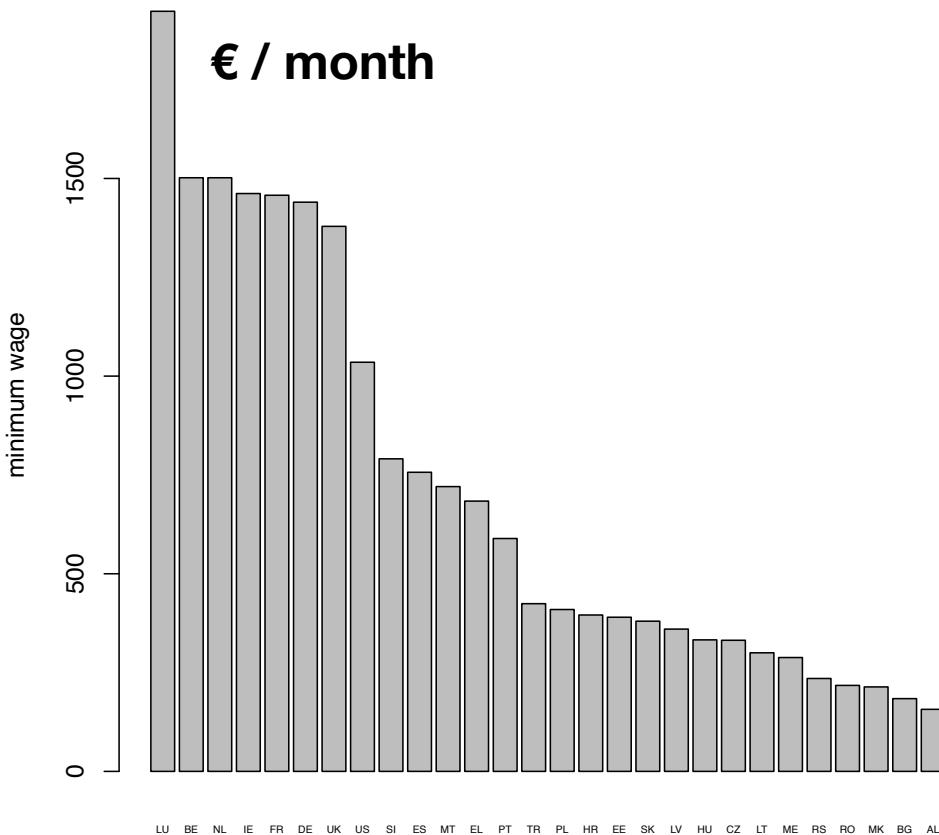
Re-expression

- 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

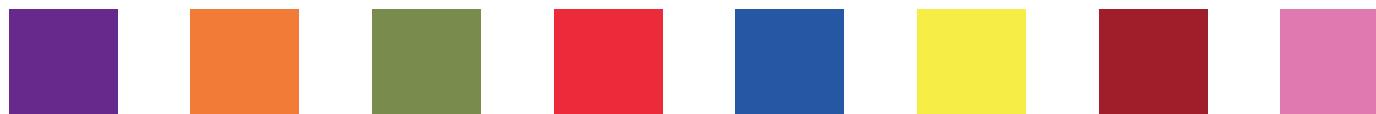
- The most relevant scale?
 - linear scale: shows absolute differences
 - log scale: if relative change (in %) is important
 - inverted scale: plot $1/y$ instead of y works sometimes



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

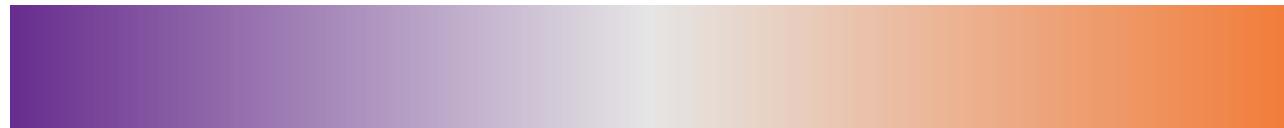


Colors

- General guidelines regarding colors
 - in heat-maps a gray scale is typically not a good idea
 - instead vary from one colour to the 'opponent' color



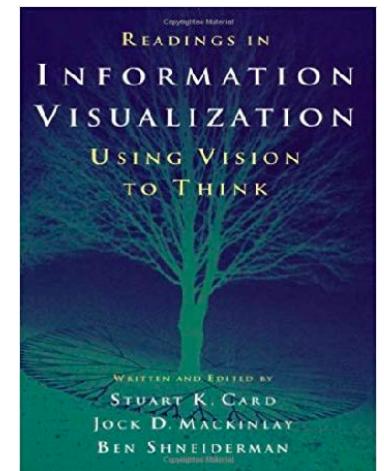
- if heat-map has a neutral value
 - vary from one colour to a neutral gray color and
 - continue to the opponent colour



- See <http://colorbrewer2.org/>

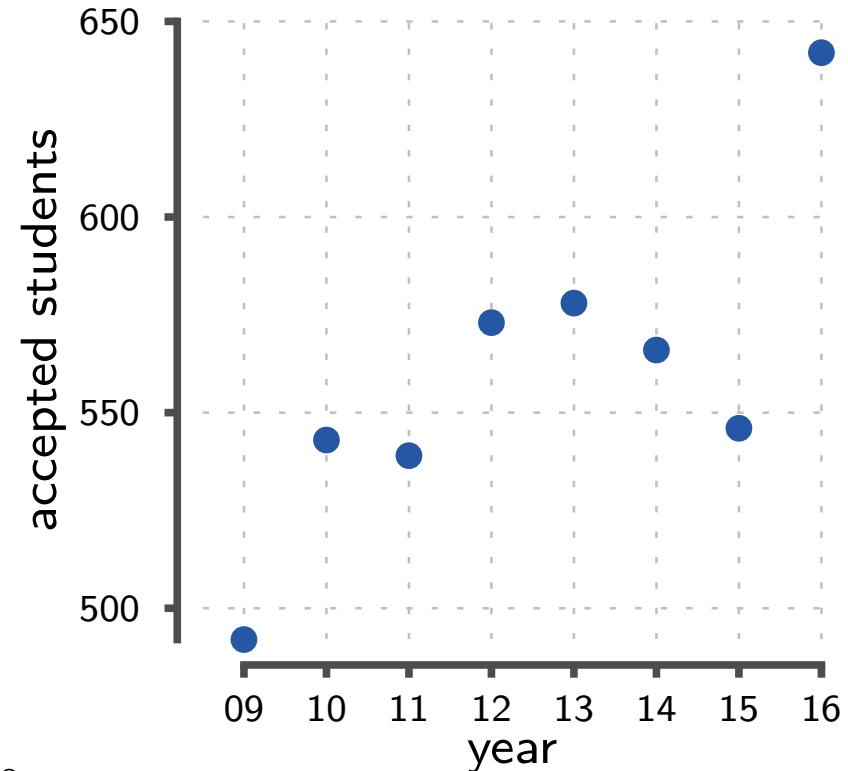
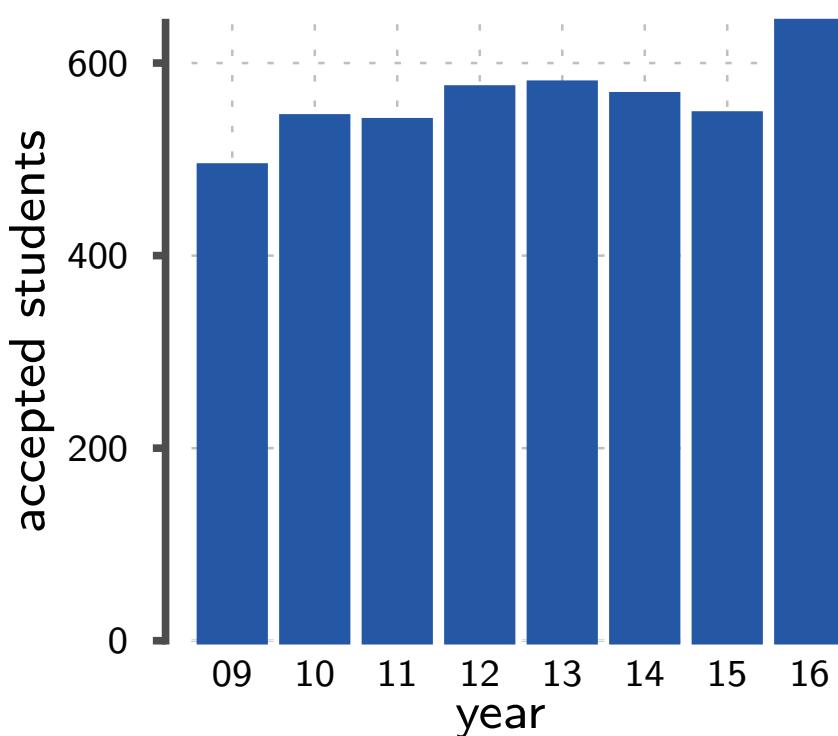
Outline

- **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.



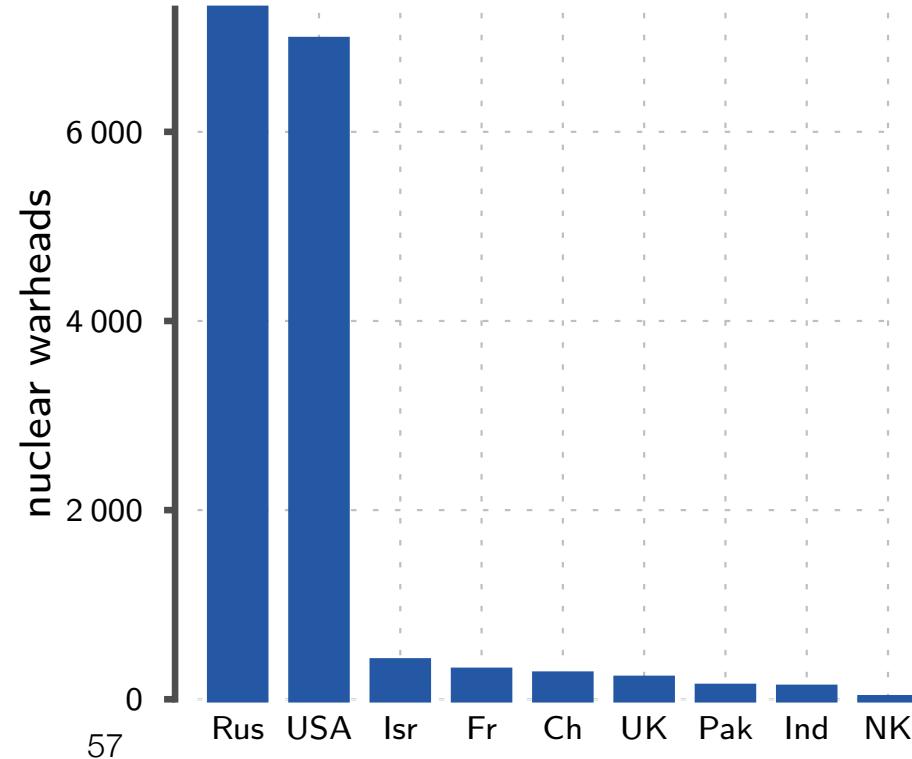
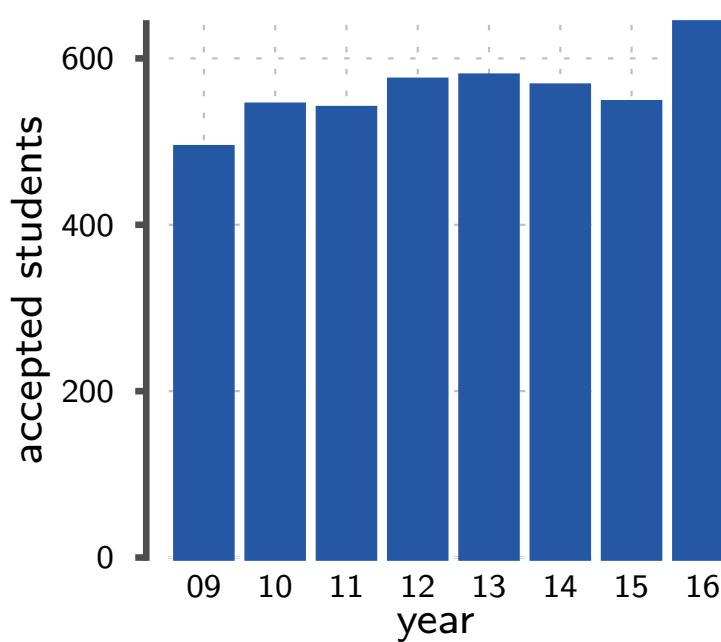
Compressed bar plots

- 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



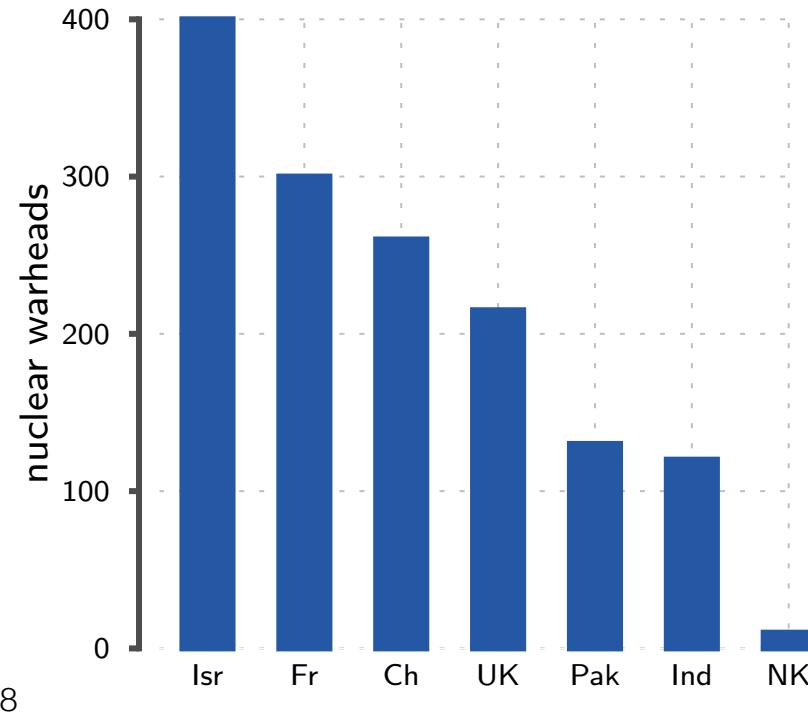
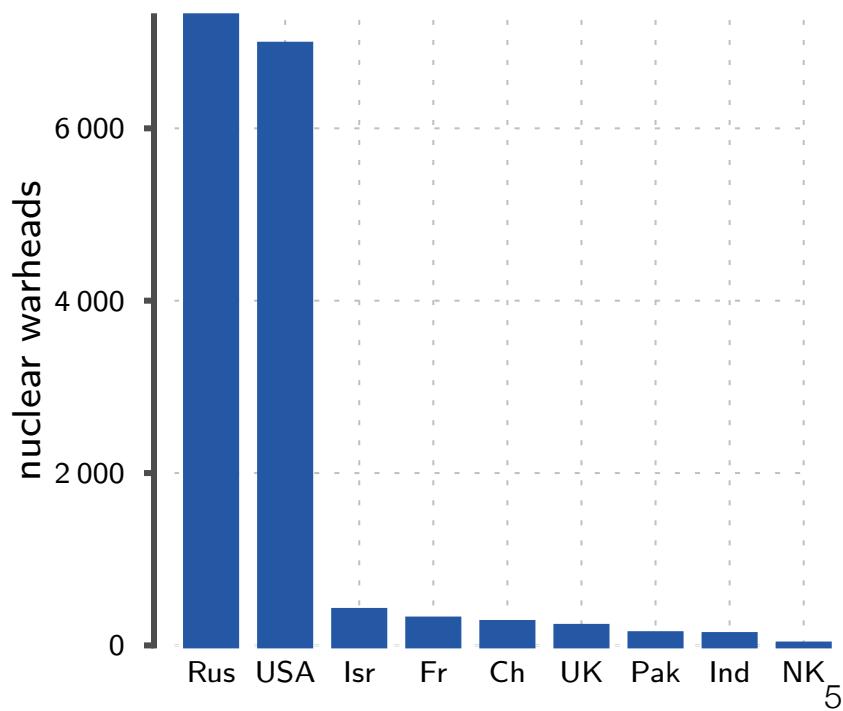
Axis ranges

- 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



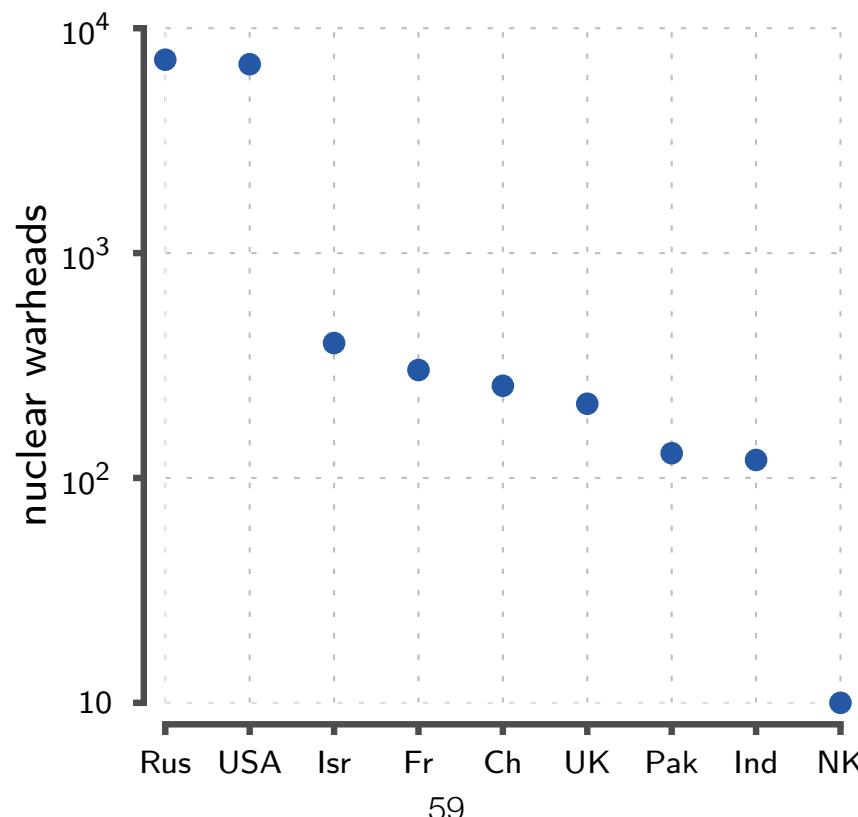
Axis ranges

- 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



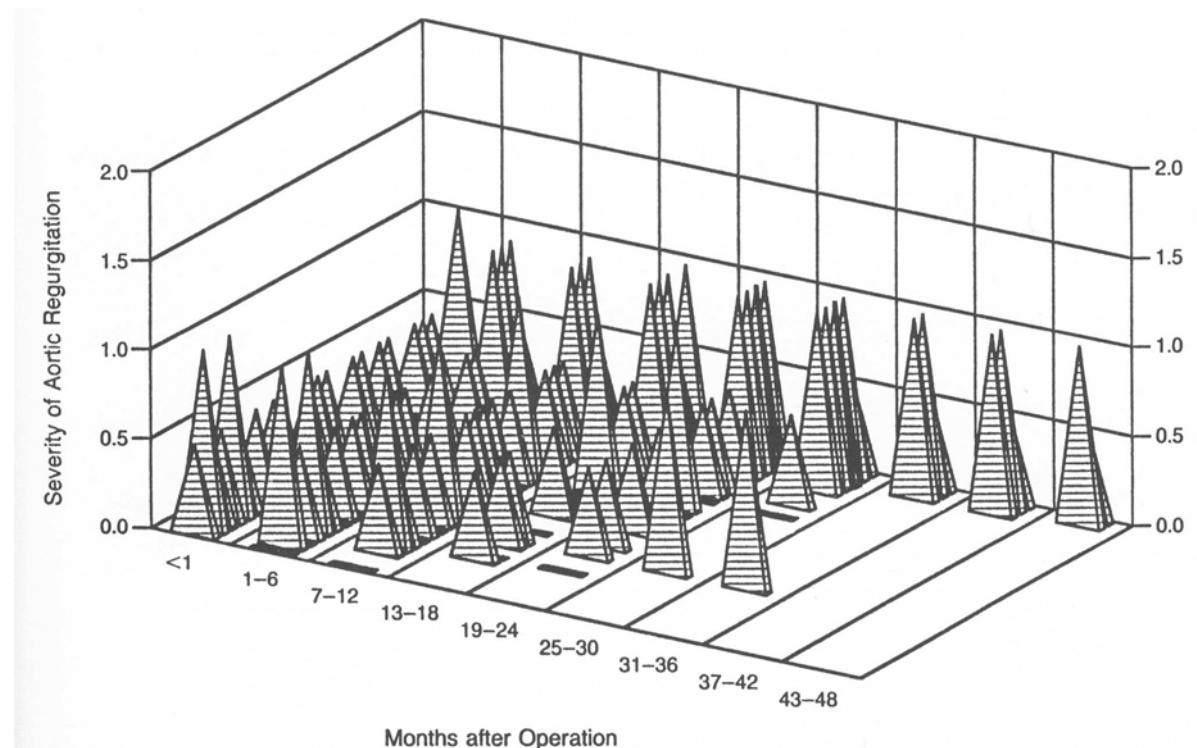
Axis ranges

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

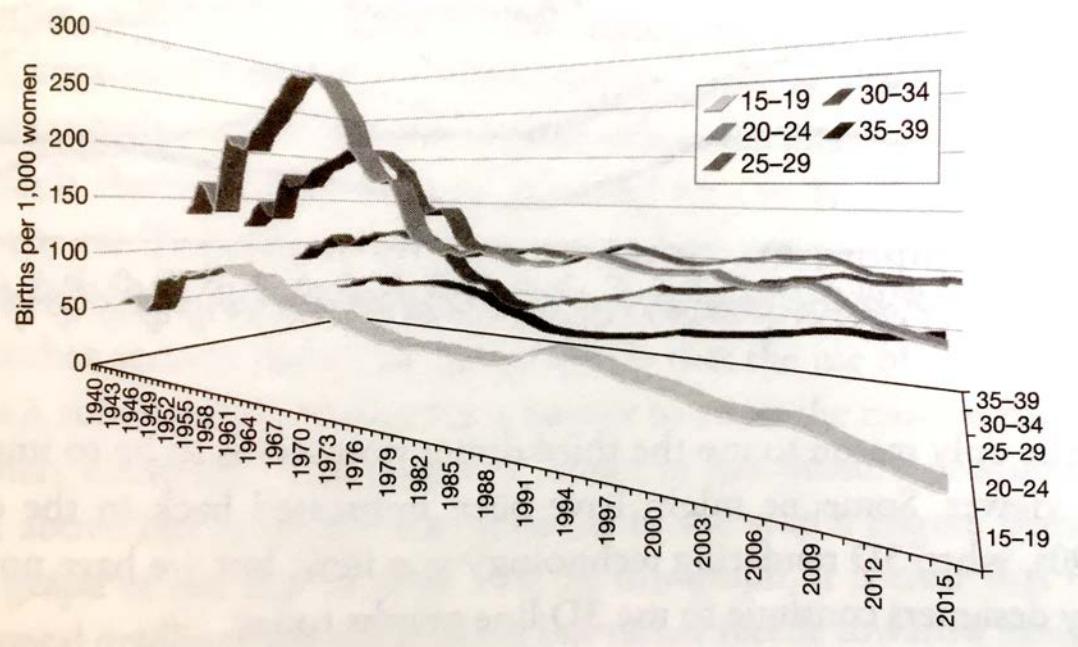


No unjustified 3d

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

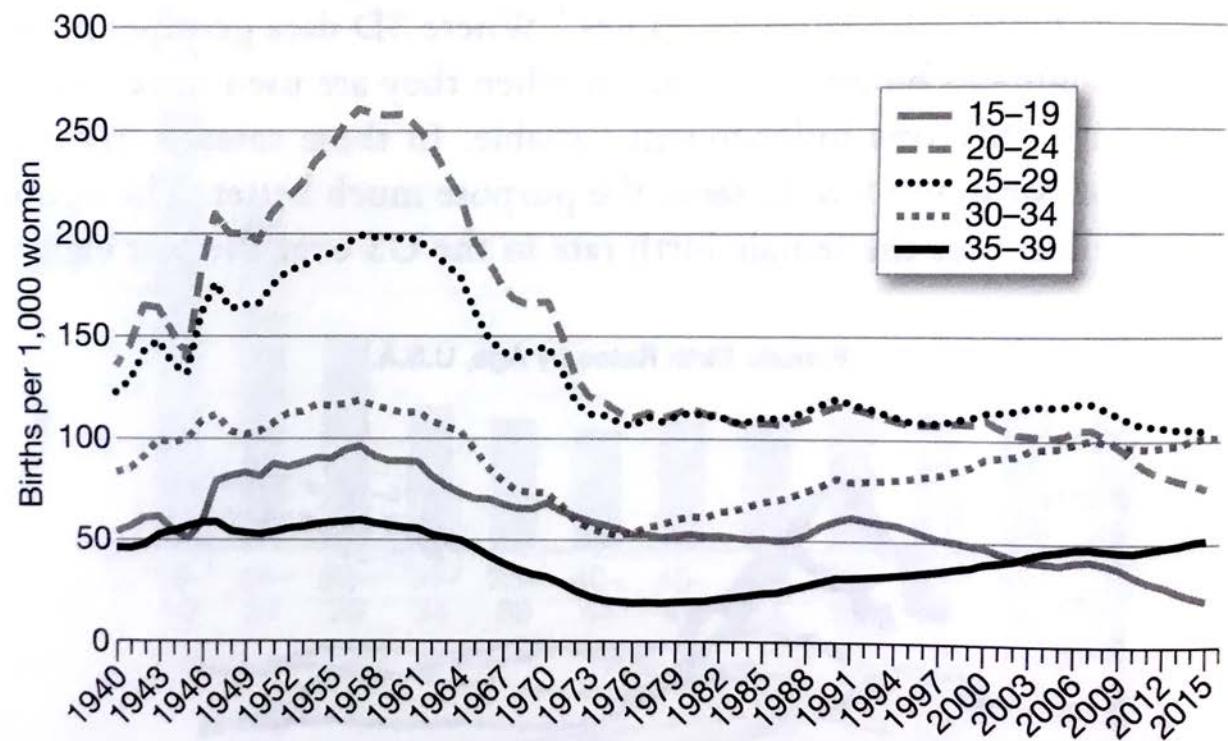


Female Birth Rates by Age, U.S.A.



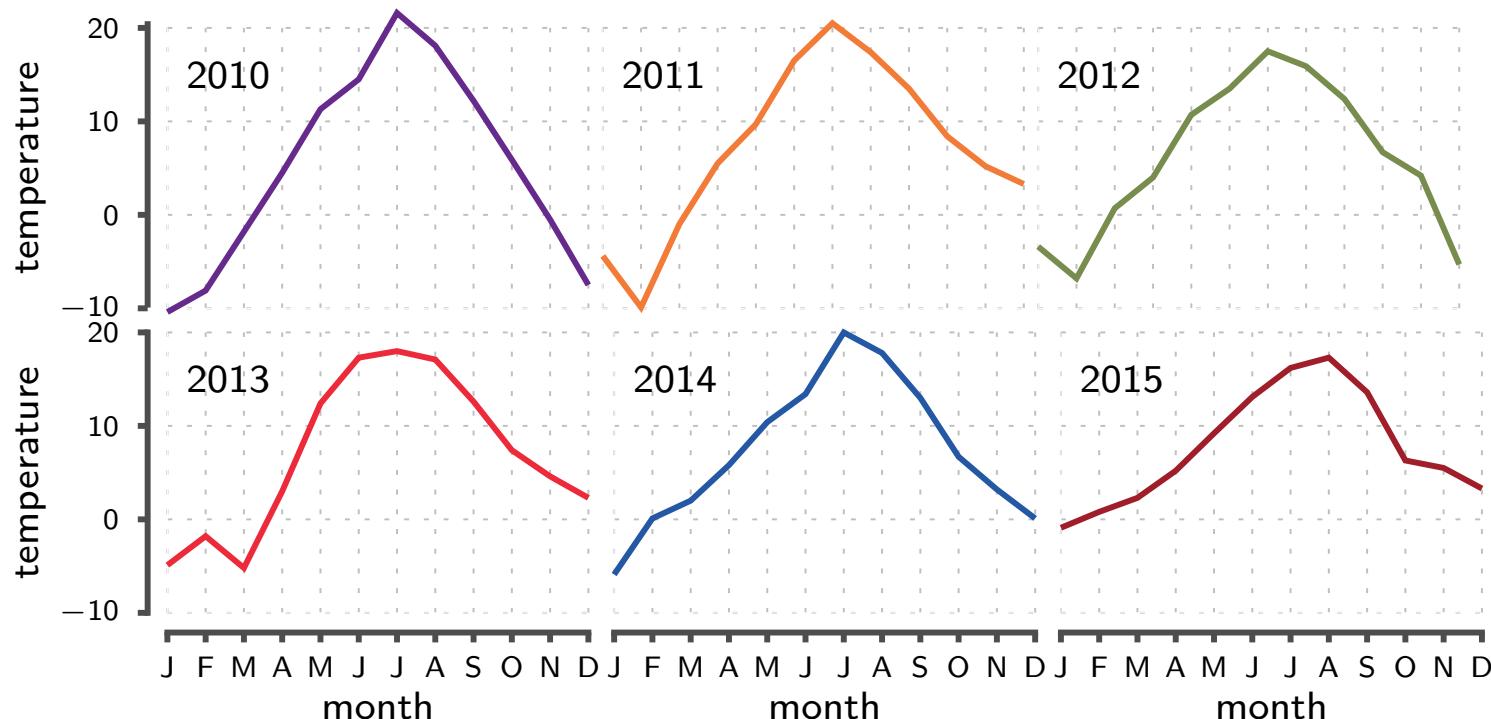
No unjustified 3d

Female Birth Rates by Age, U.S.A.



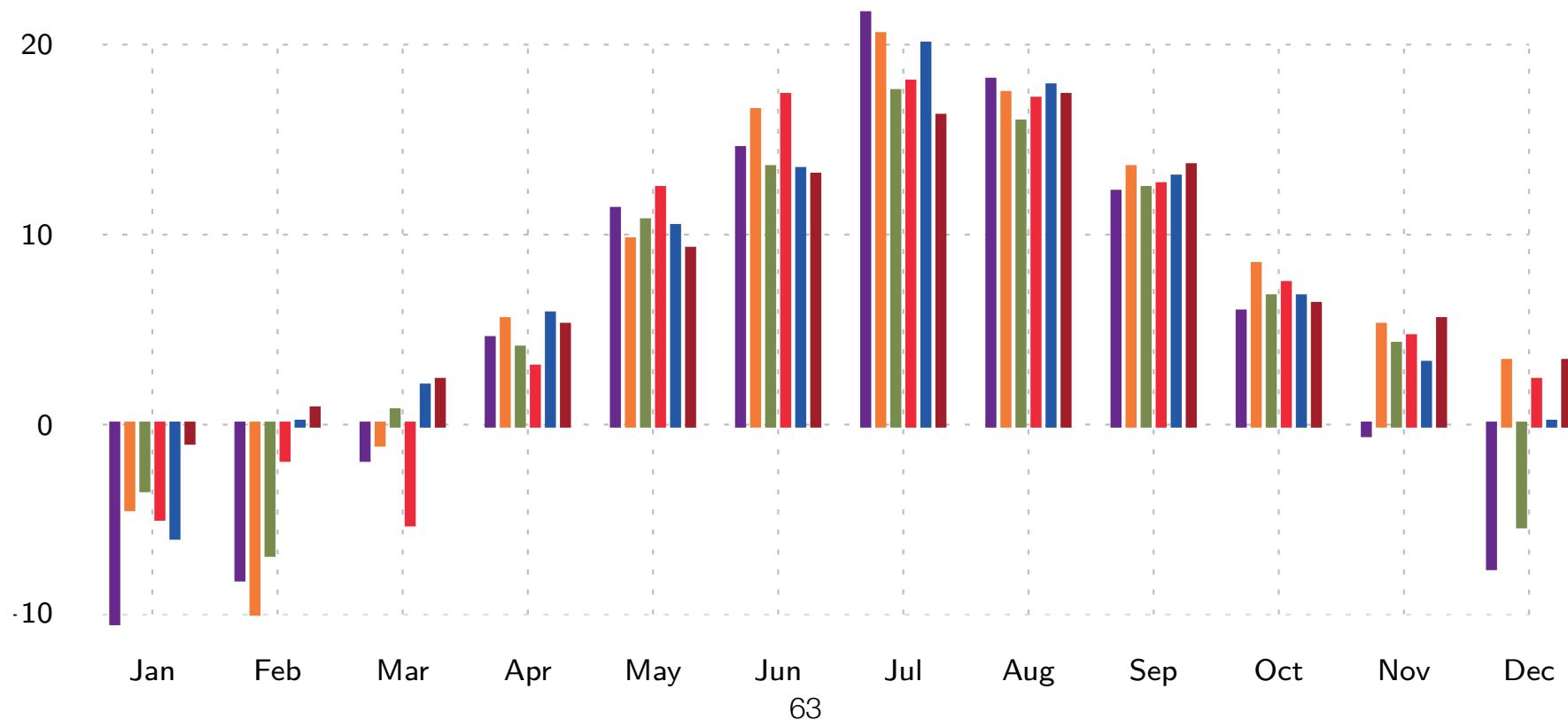
No unjustified 3d

- 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



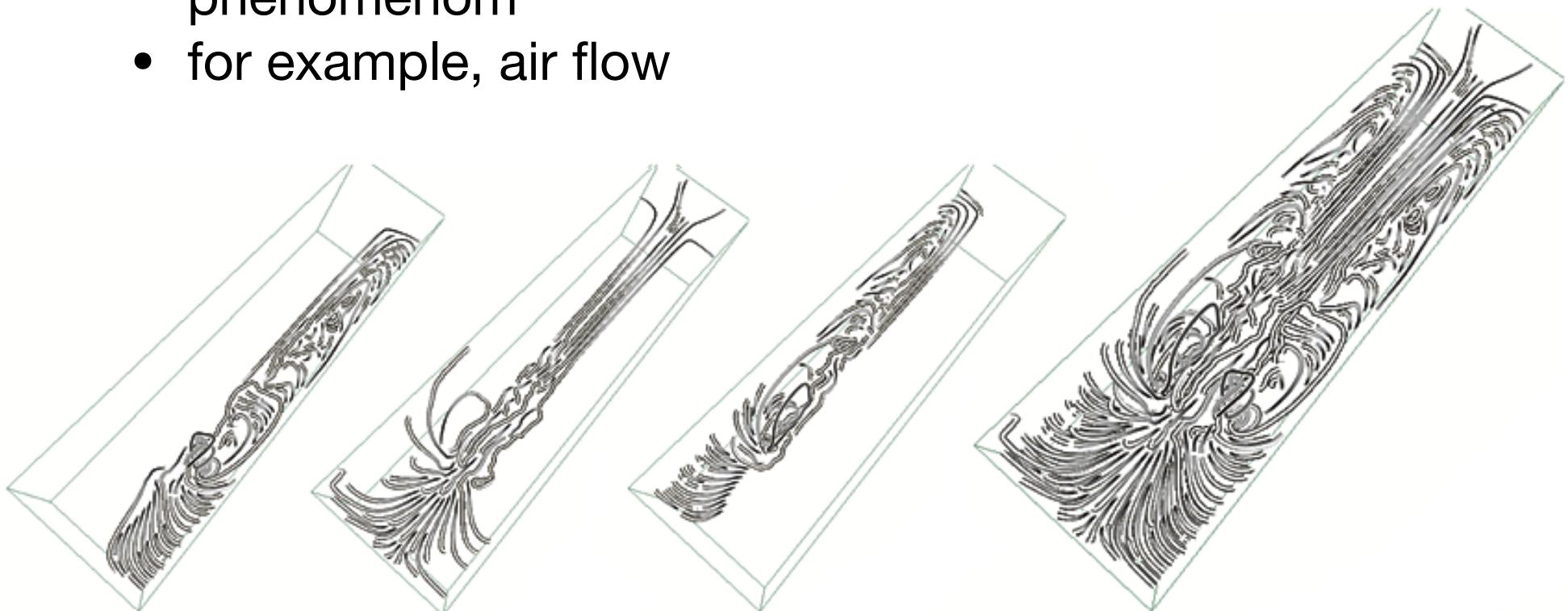
No unjustified 3d

- 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 phenomenon
 - for example, air flow



Graphics by purpose

- numeric table gives exact values
- line graph reveals trends and anomalies
- 2D helps to locate data
- 3D gives an easy overview

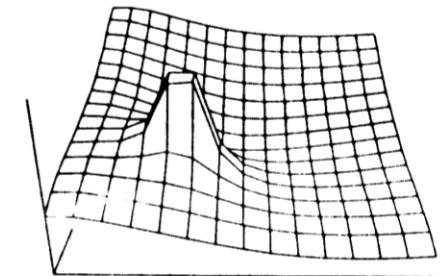
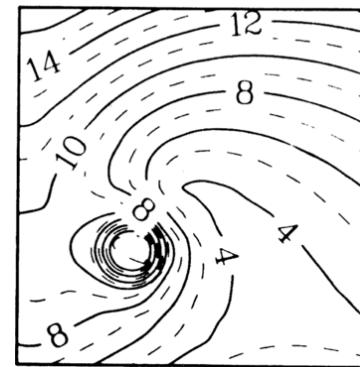
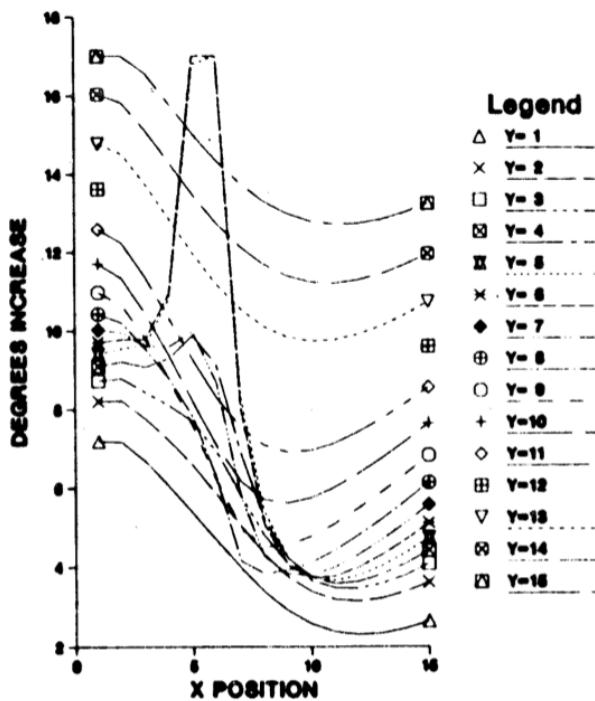
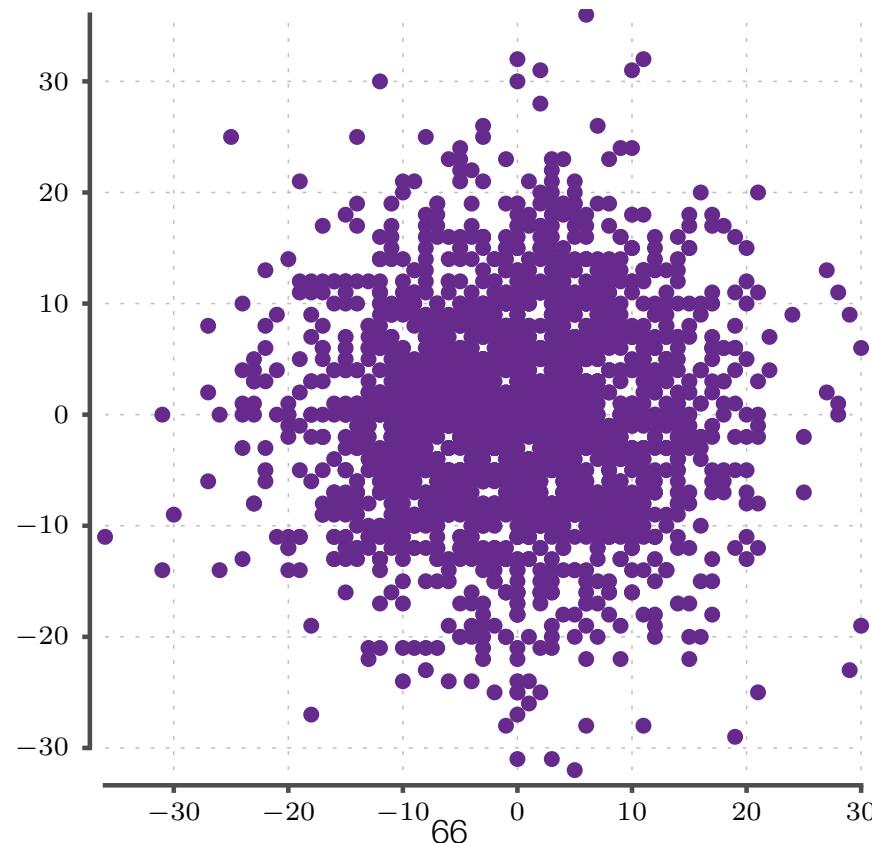


Table 1.
Temperature distribution.

17.0	17.0	16.6	15.8	15.0	14.3	13.8	13.3	13.0	12.8	12.7	12.7	12.8	13.0	13.2
16.0	15.8	15.1	14.3	13.5	12.8	12.2	11.7	11.4	11.2	11.2	11.3	11.4	11.7	12.0
14.8	14.5	13.8	12.9	12.0	11.2	10.6	10.1	9.8	9.7	9.8	9.9	10.1	10.4	10.8
13.6	13.3	12.5	11.5	10.6	9.7	9.0	8.6	8.3	8.3	8.4	8.6	8.9	9.2	9.6
12.6	12.2	11.4	10.3	9.3	8.3	7.6	7.0	6.9	7.0	7.1	7.4	7.8	8.1	8.6
11.7	11.3	10.4	9.4	8.2	7.1	6.2	5.7	5.6	5.8	6.0	6.4	6.8	7.2	7.6
11.0	10.7	9.8	8.8	7.6	6.2	5.0	4.5	4.6	4.8	5.1	5.5	5.9	6.4	6.8
10.4	10.2	9.5	8.8	7.8	6.3	4.2	3.8	3.9	4.1	4.4	4.8	5.2	5.7	6.1
10.0	10.0	9.6	9.6	10.0	8.7	5.8	4.3	3.8	3.8	4.0	4.3	4.7	5.1	5.6
9.7	9.8	9.8	11.0	17.0	17.0	8.1	5.1	4.1	3.7	3.8	4.0	4.3	4.7	5.1
9.4	9.6	9.8	10.9	16.8	17.0	8.4	5.4	4.2	3.8	3.7	3.8	4.0	4.4	4.7
9.1	9.2	9.1	9.3	9.9	9.1	6.9	5.2	4.2	3.8	3.6	3.6	3.8	4.1	4.4
8.7	8.8	8.4	8.1	7.7	6.9	5.8	4.8	4.1	3.7	3.5	3.5	3.6	3.8	4.1
8.2	8.2	7.8	7.2	6.6	5.8	5.1	4.3	3.8	3.4	3.2	3.2	3.2	3.4	3.6
7.2	7.2	6.7	6.0	5.4	4.7	4.0	3.4	2.9	2.6	2.4	2.3	2.3	2.5	2.6

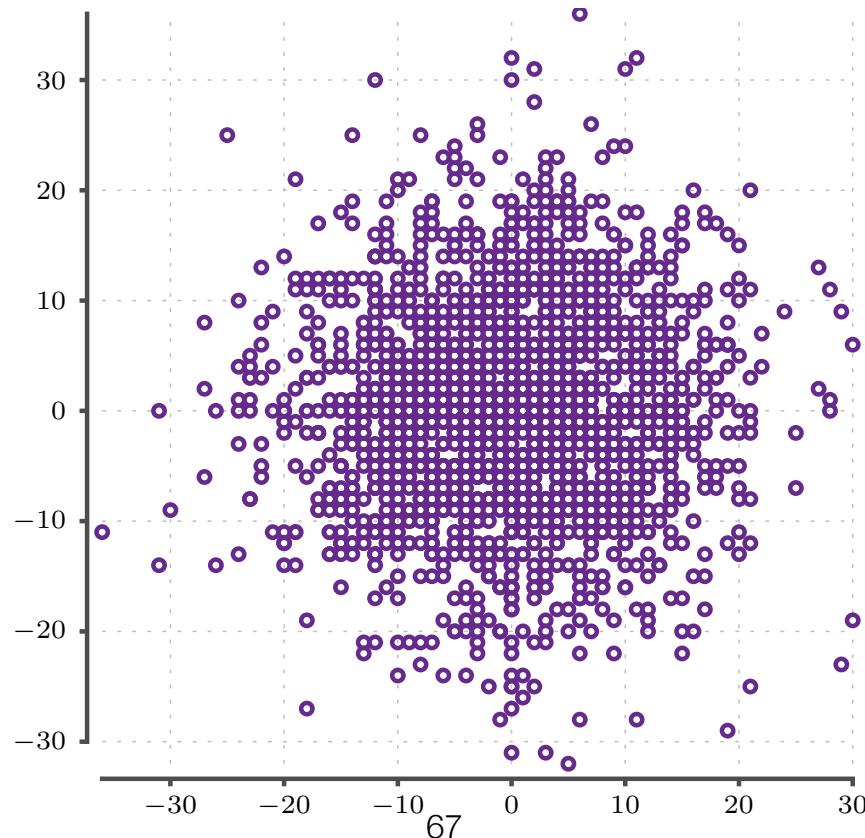
Dealing with overplotting in scatterplot

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



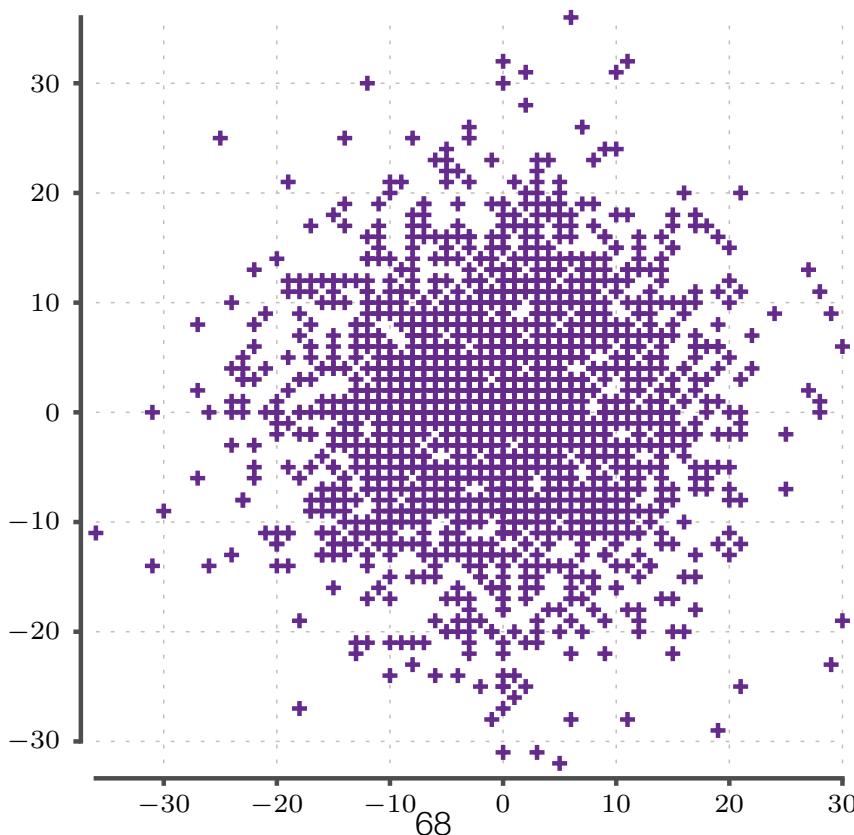
Dealing with overplotting in scatterplot

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



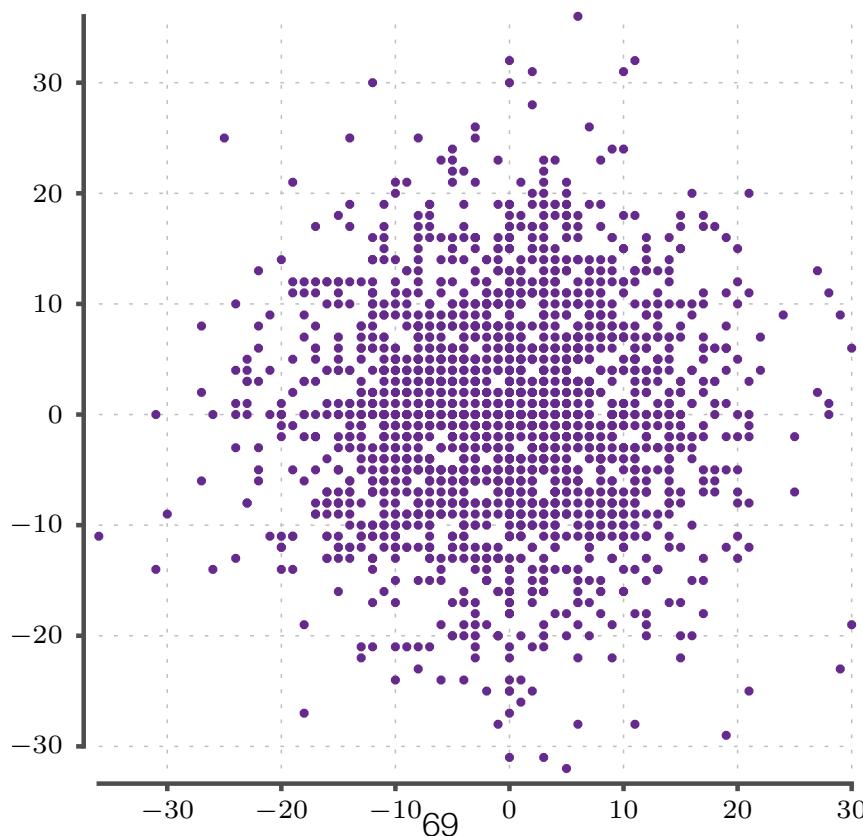
Dealing with overplotting in scatterplot

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



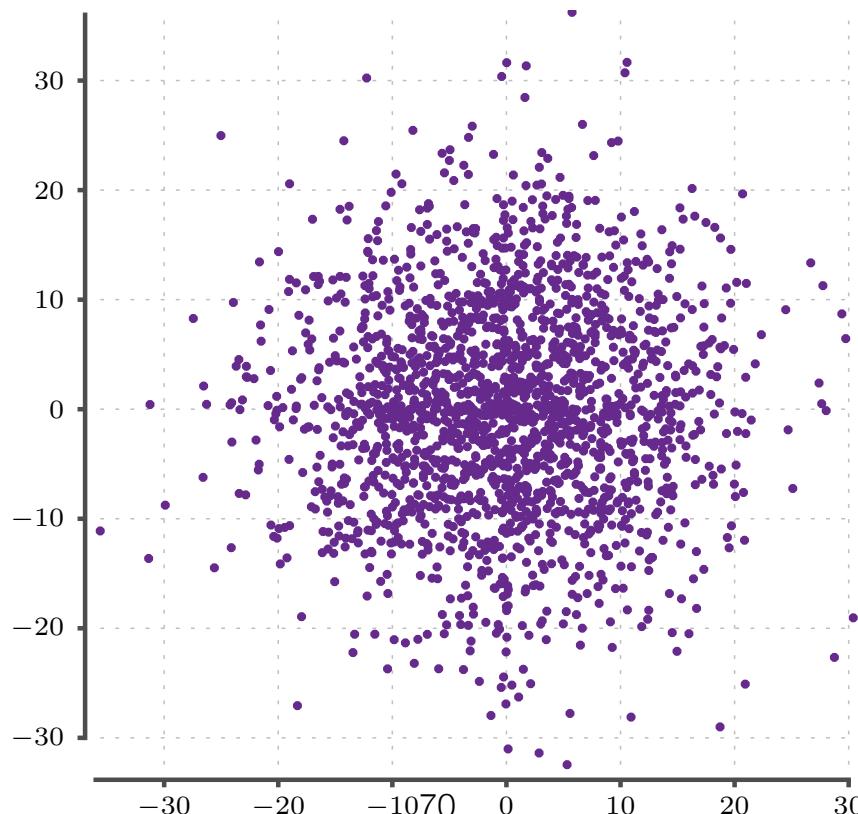
Dealing with overplotting in scatterplot

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



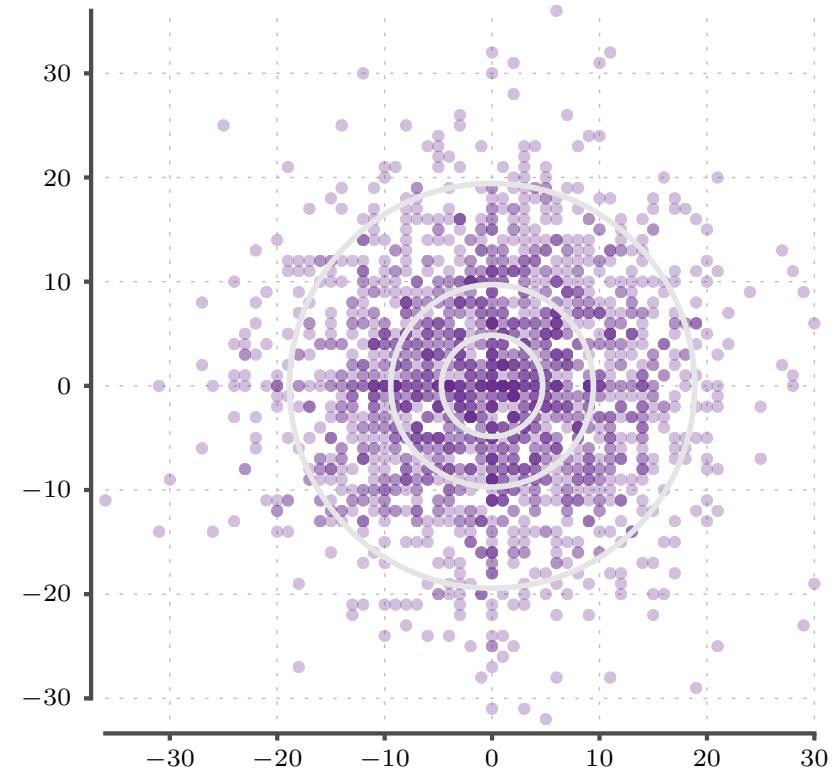
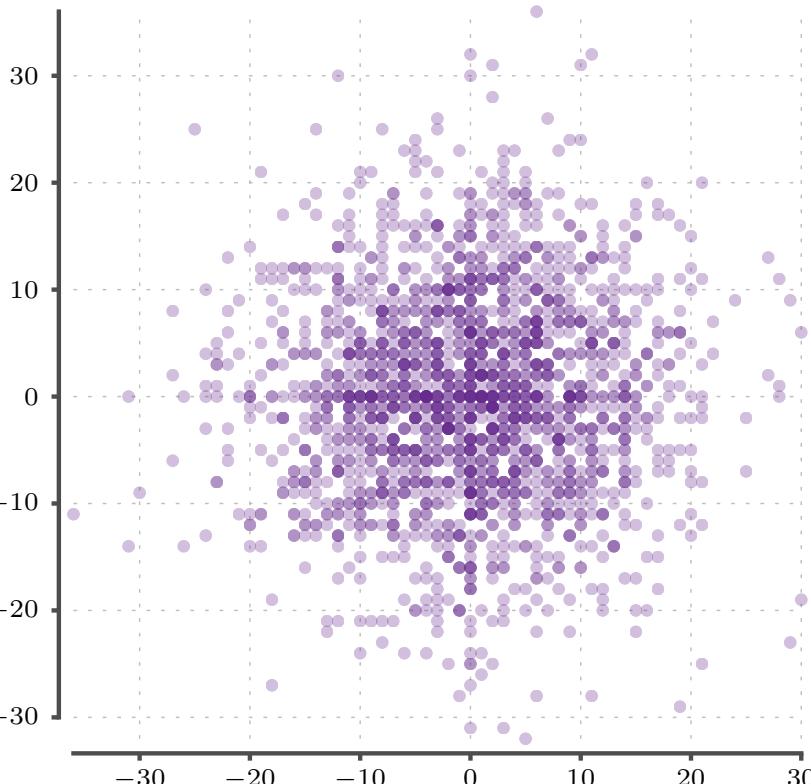
Dealing with overplotting in scatterplot

- 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)



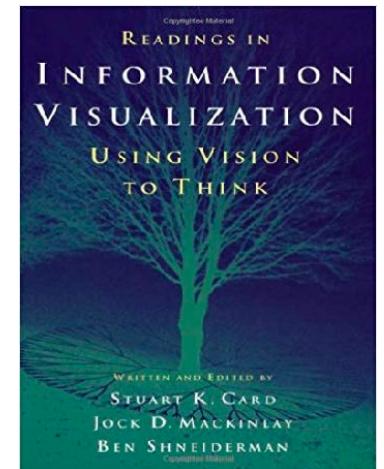
Dealing with overplotting in scatterplot

- 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)



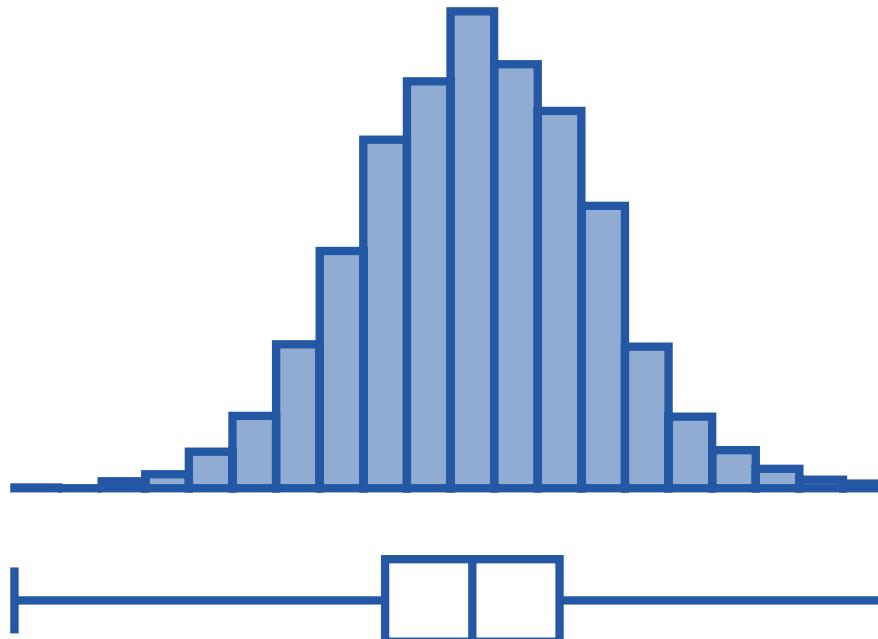
Outline

- **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.



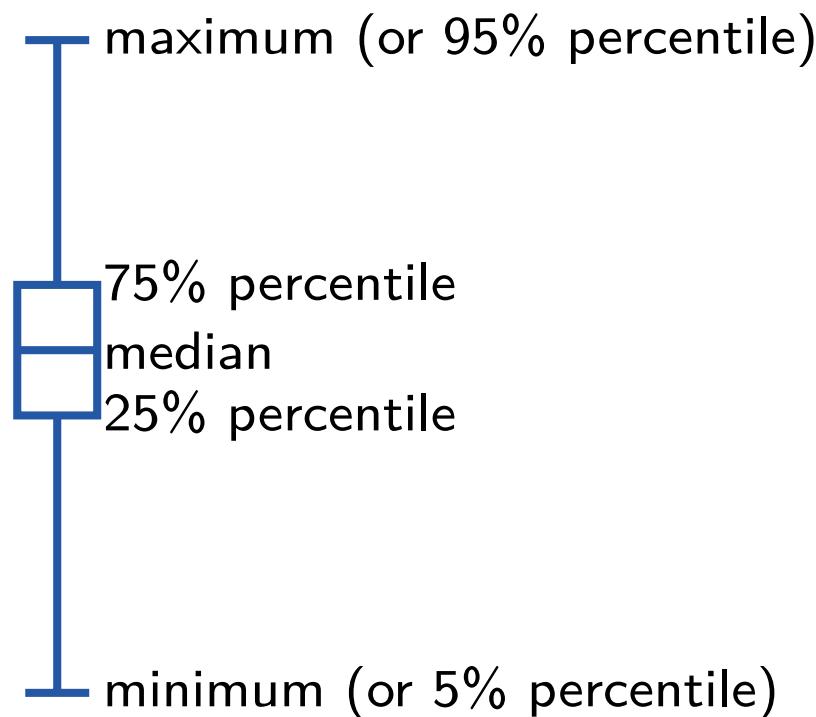
Visualizing distributions

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



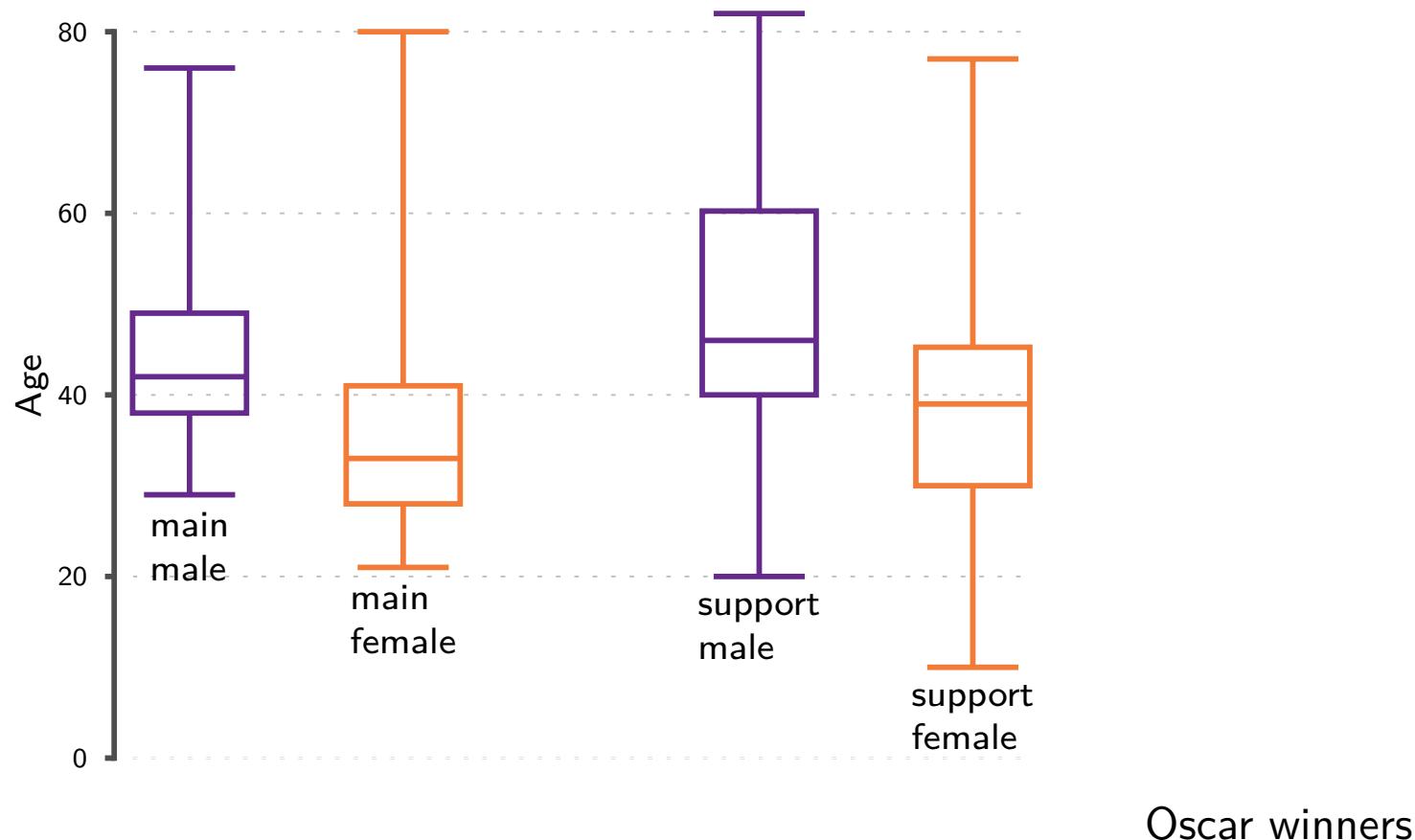
Visualizing distributions

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



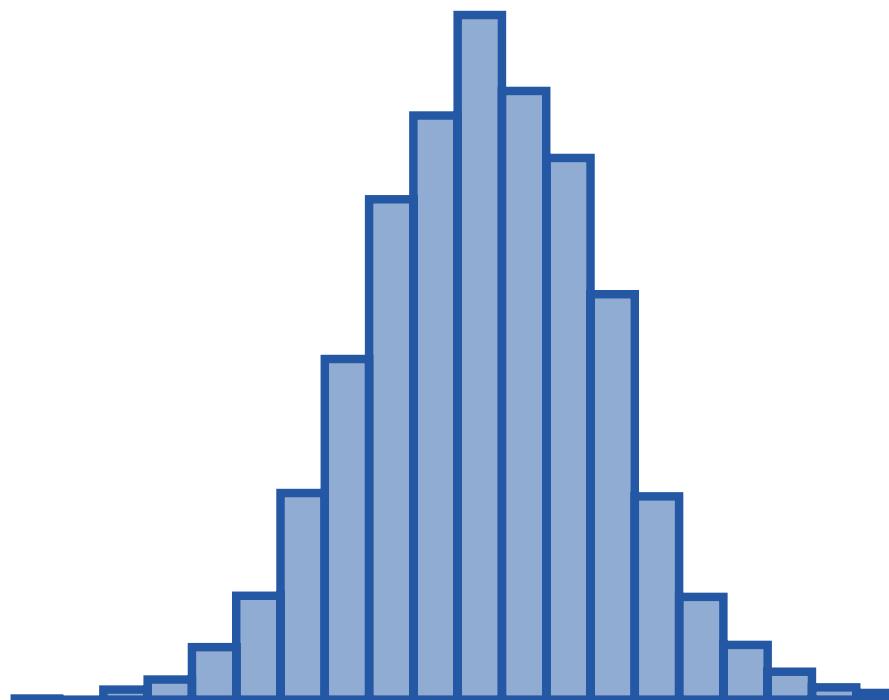
Visualizing distributions

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



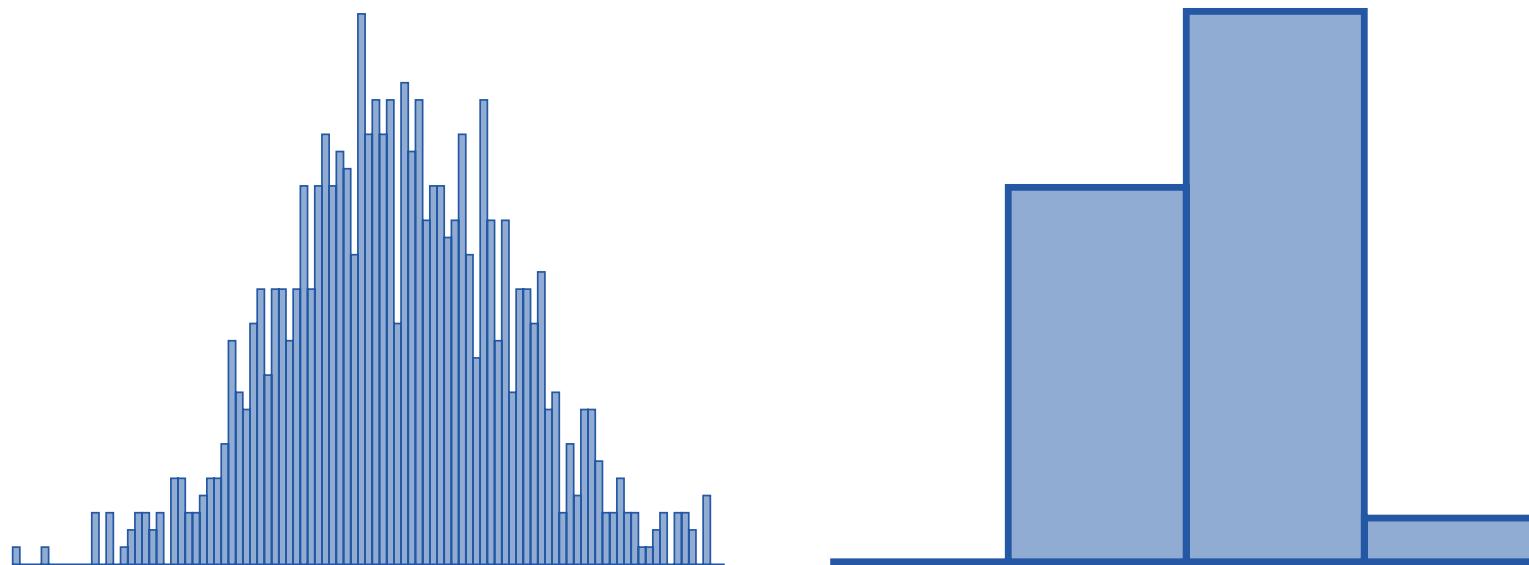
Visualizing distributions

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



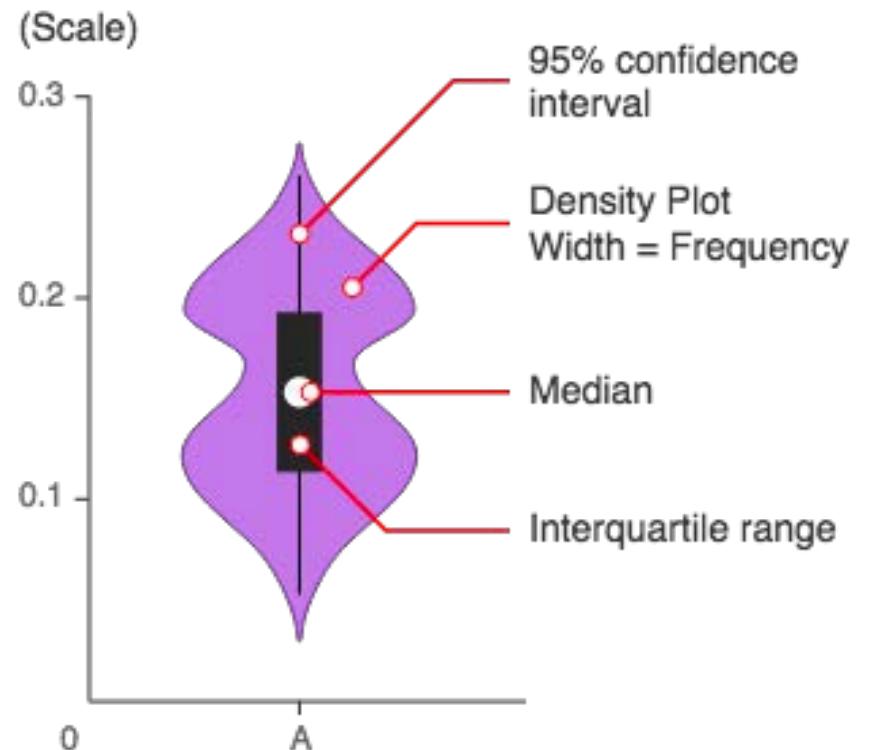
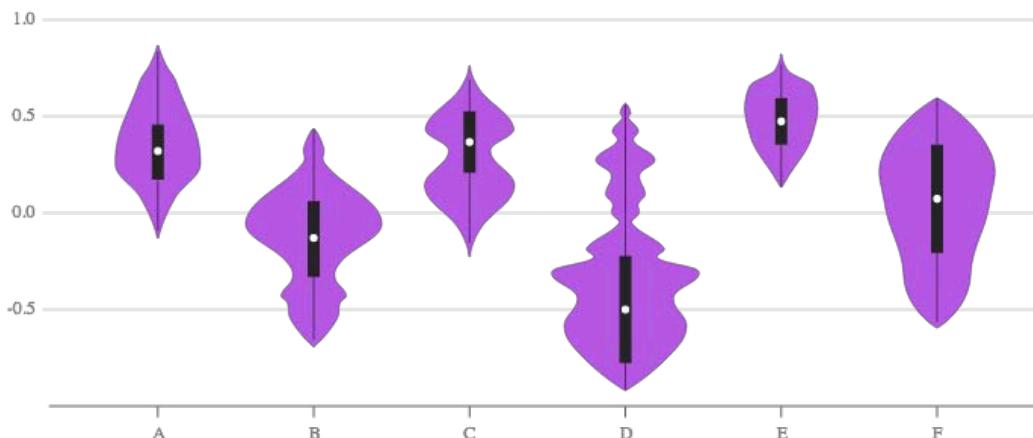
Visualizing distributions

- 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 easier to compare with each other
 - *violin plot* combines both plots together

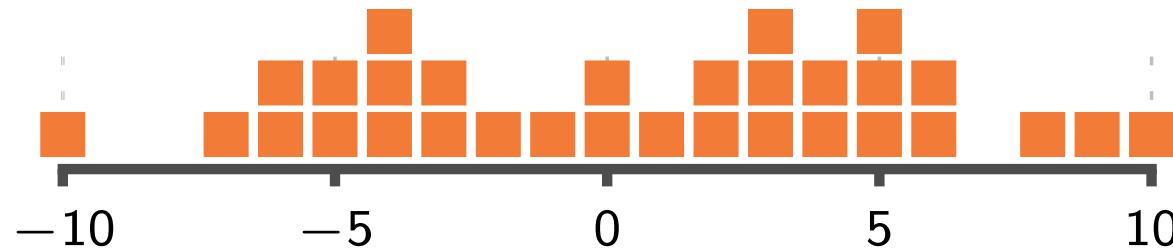


Visualizing distributions

- 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

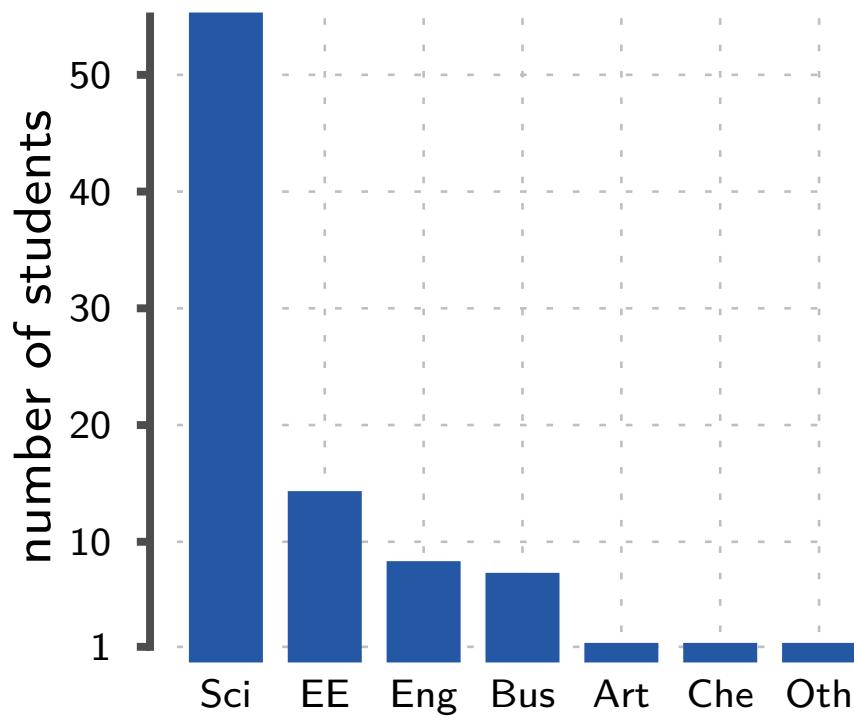
- Stem-and-leaf plot
 - similar to histogram / stacked strip plot
 - each value is split in two parts (stem and leaf)
 - stem is used for y-axis
 - leaf is shown as a number
 - values are sorted and stacked using x-axis
- Data: 1.5, 1.6, 2.1, 2.3, 2.3, 2.6, 2.6, 3.0, 3.2, 4.1
- Stem-and-leaf plot:

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

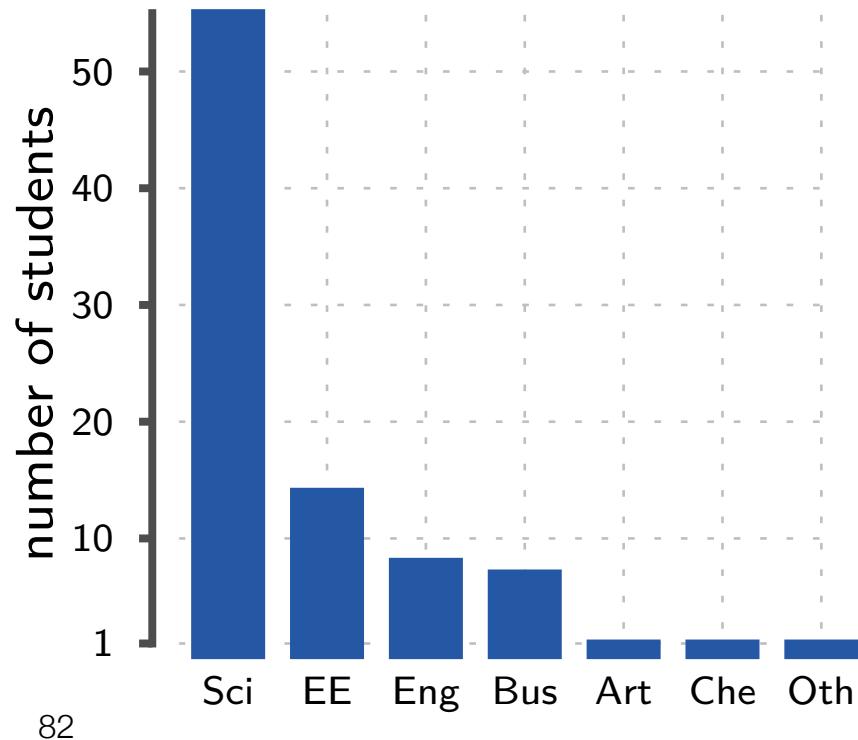
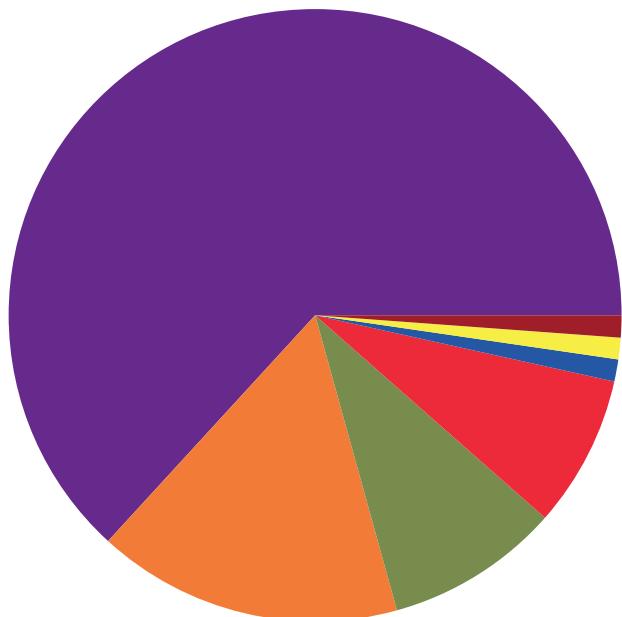
Ranking and part-to-whole analysis

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



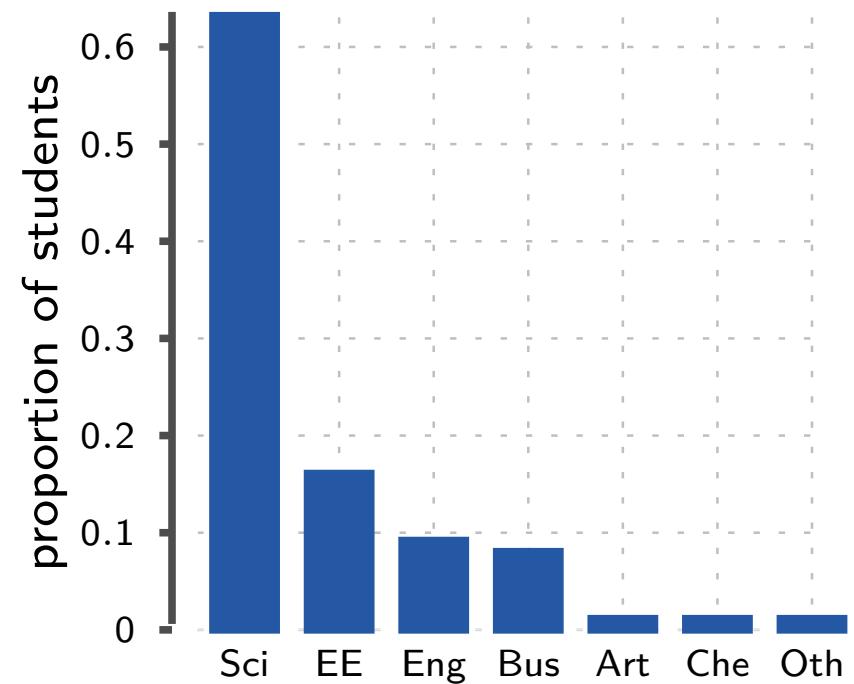
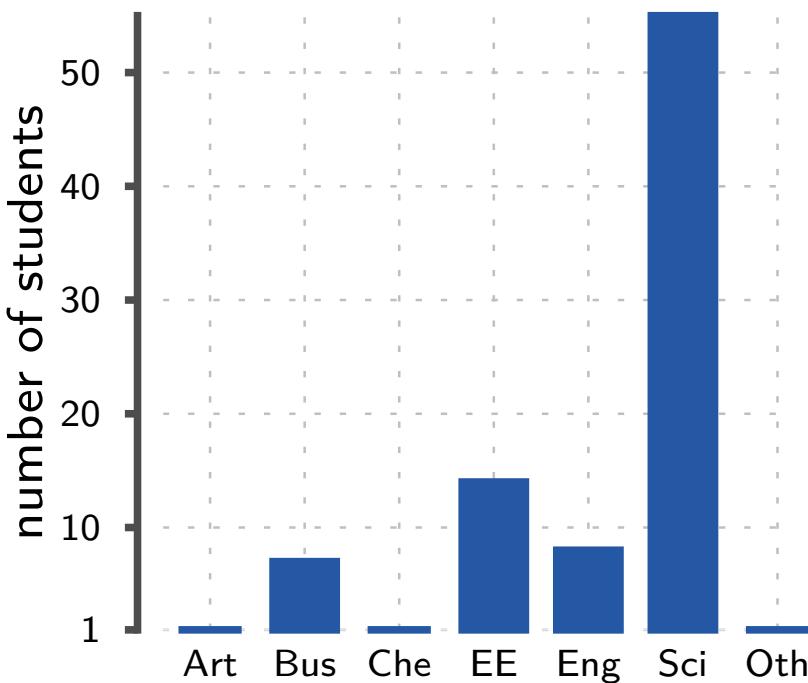
Ranking and part-to-whole analysis

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



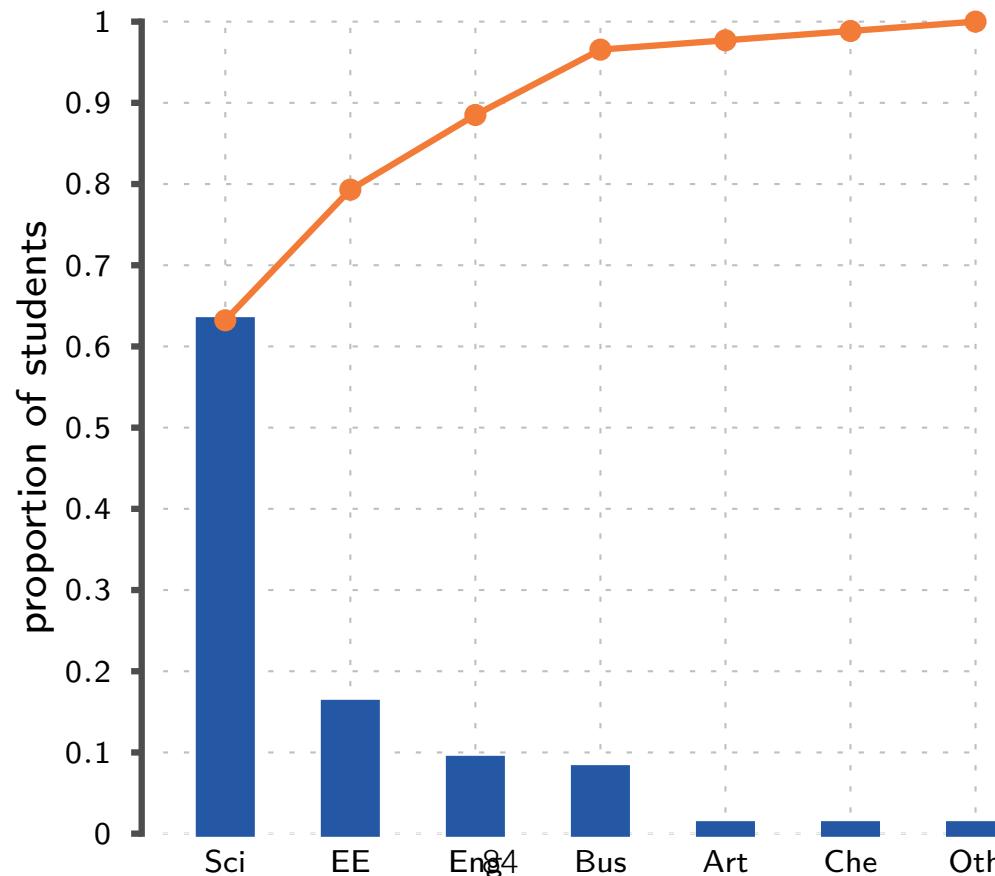
Ranking and part-to-whole analysis

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



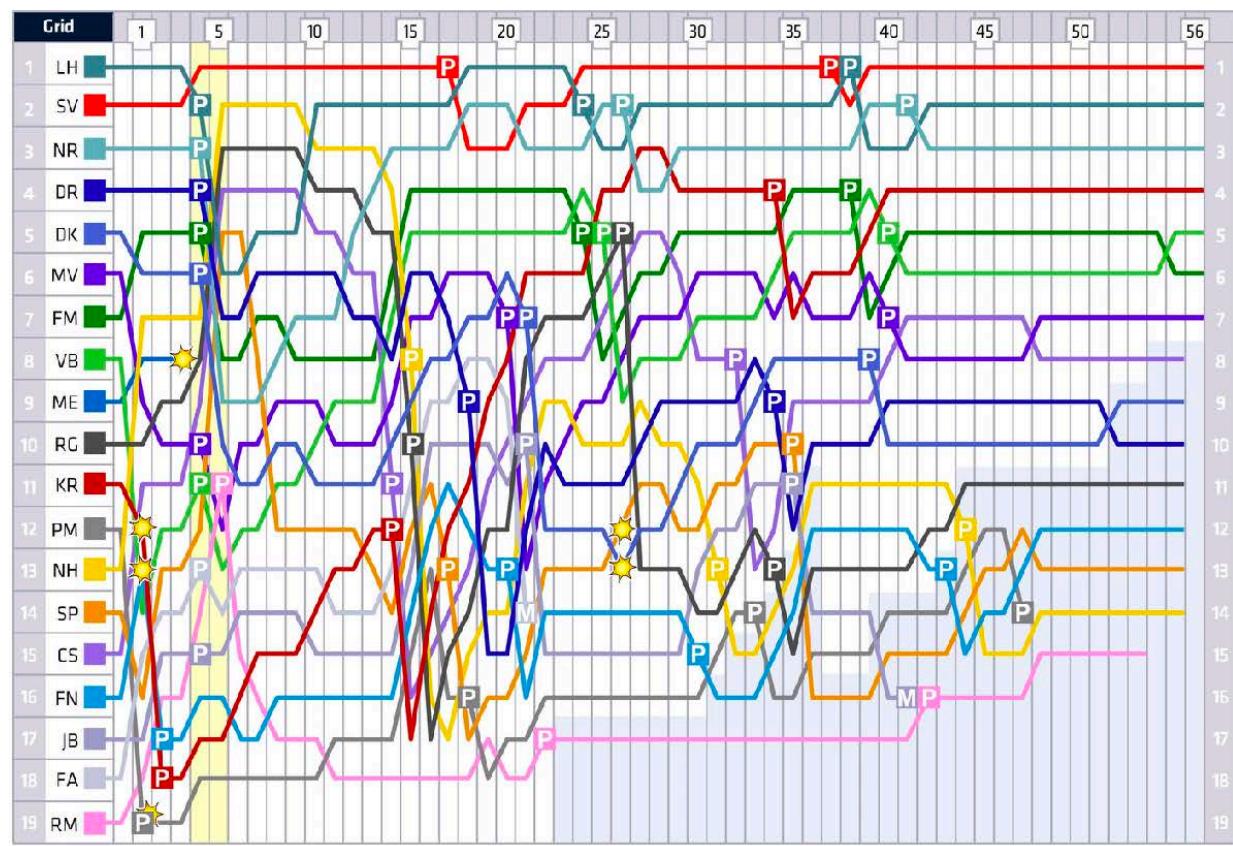
Ranking and part-to-whole analysis

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

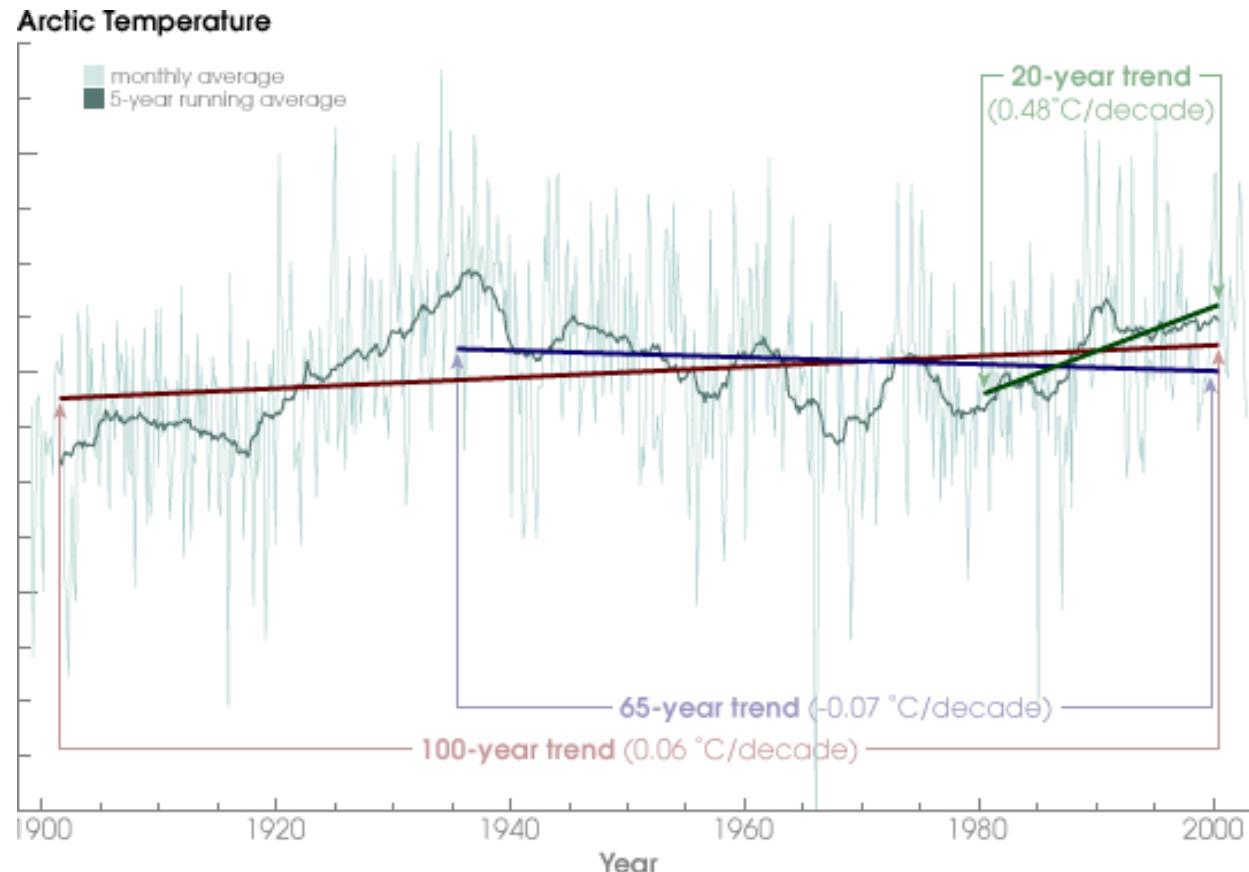


Time-series displays

- 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

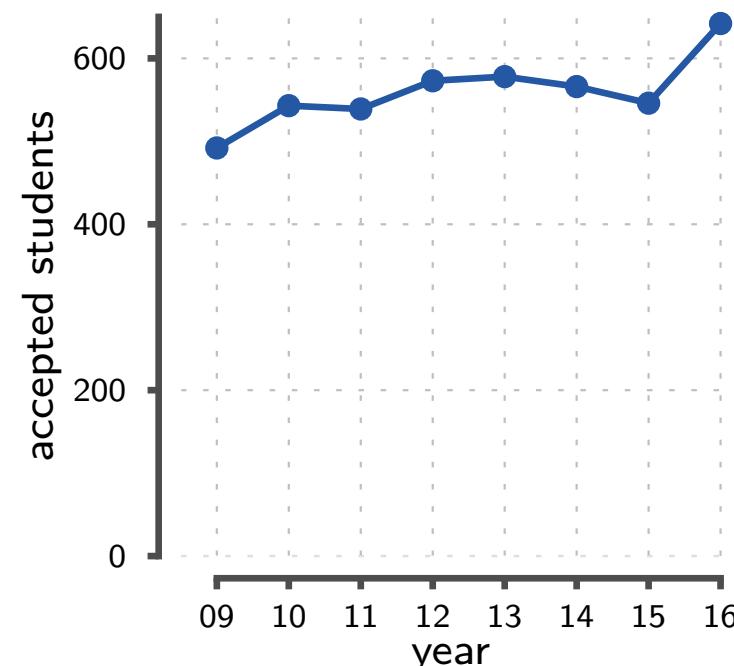
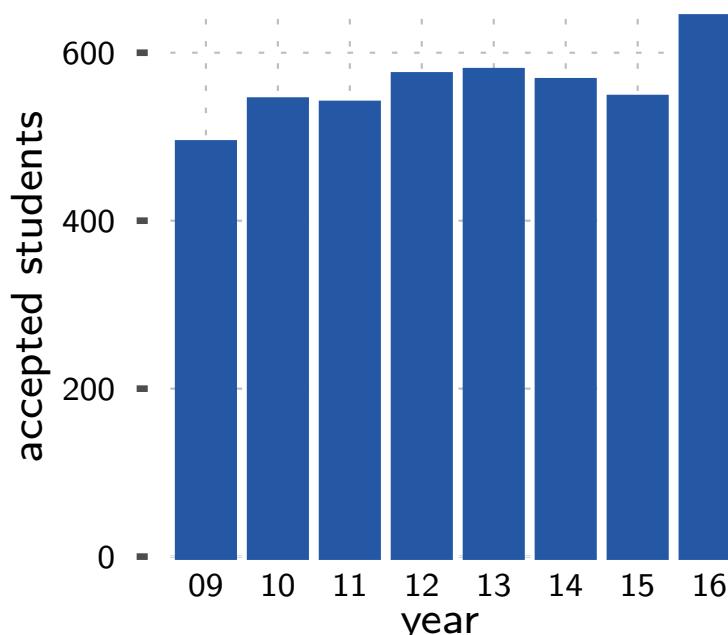
Time-series displays

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



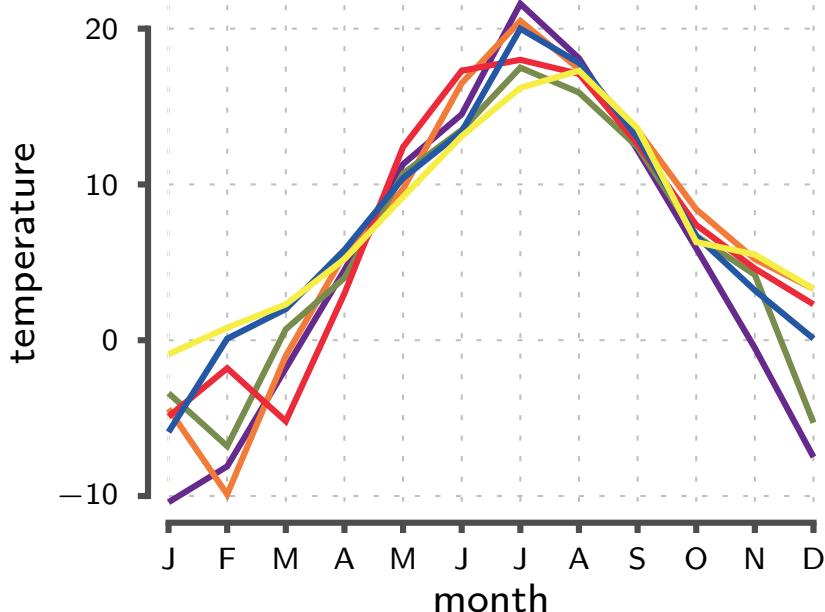
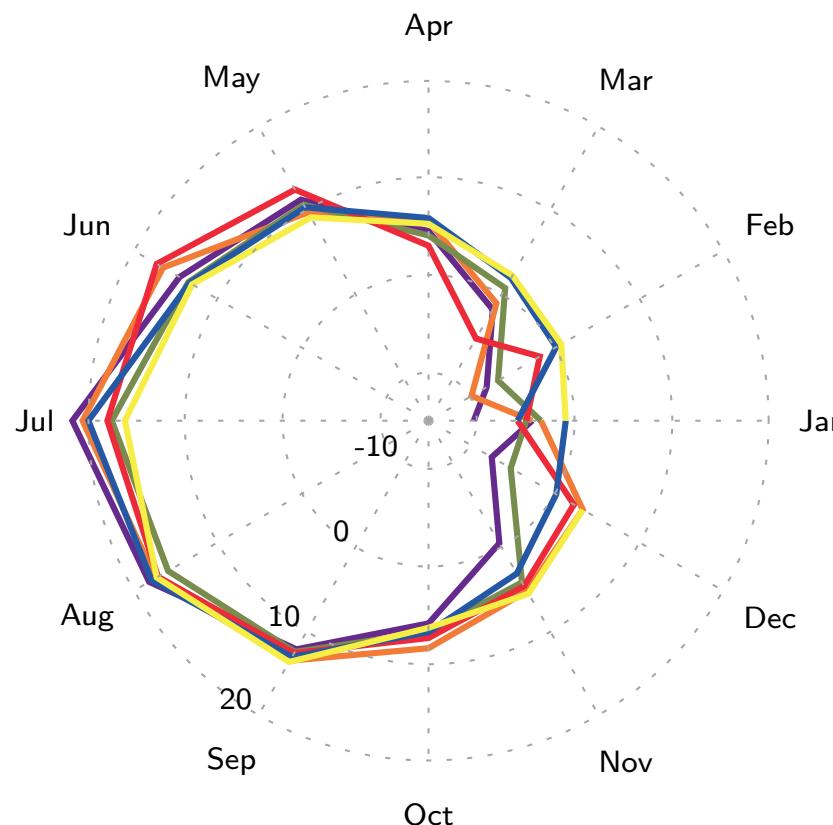
Time-series displays

- 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



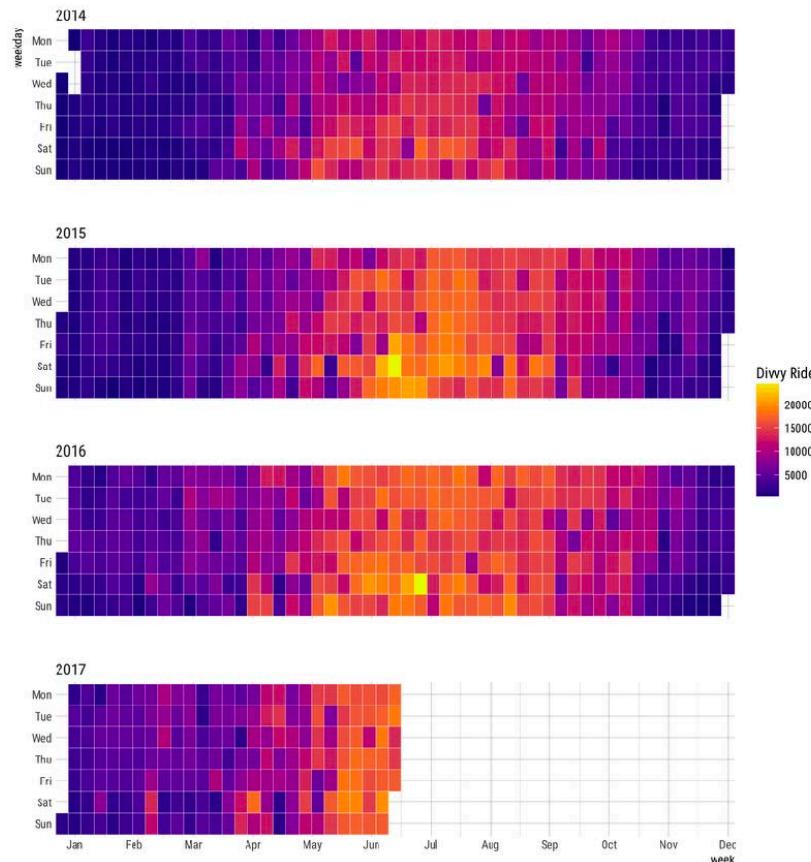
Time-series displays

- 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

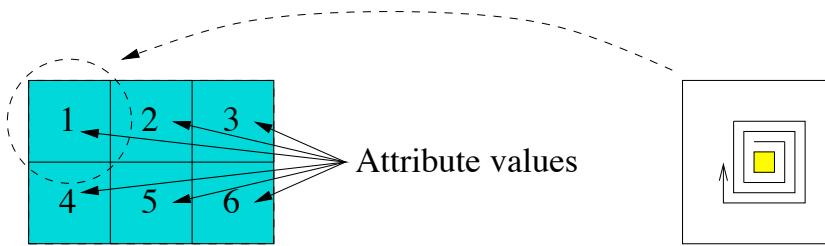


<http://austinwehrwein.com/data-visualization/heatmaps-with-divvy-data/>

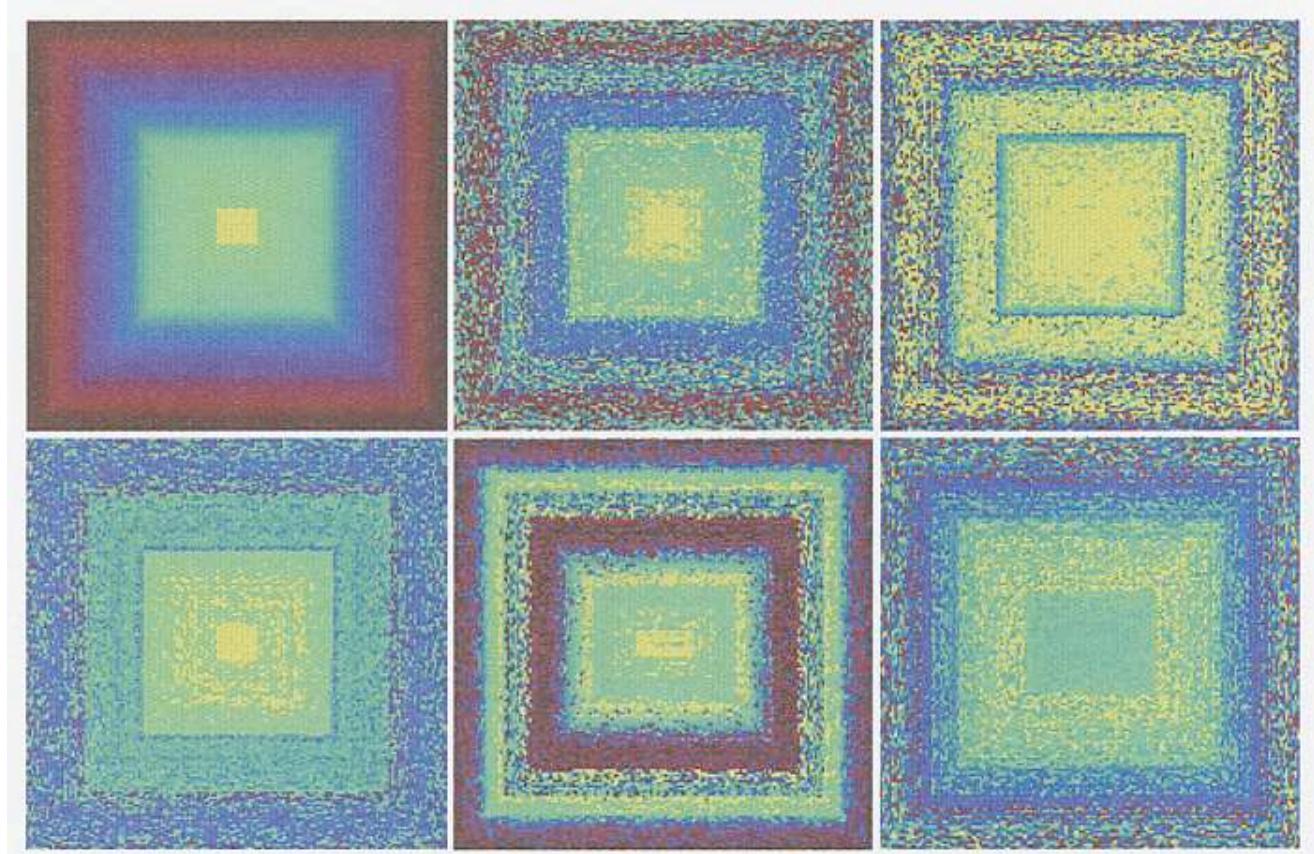
Time-series displays

- **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

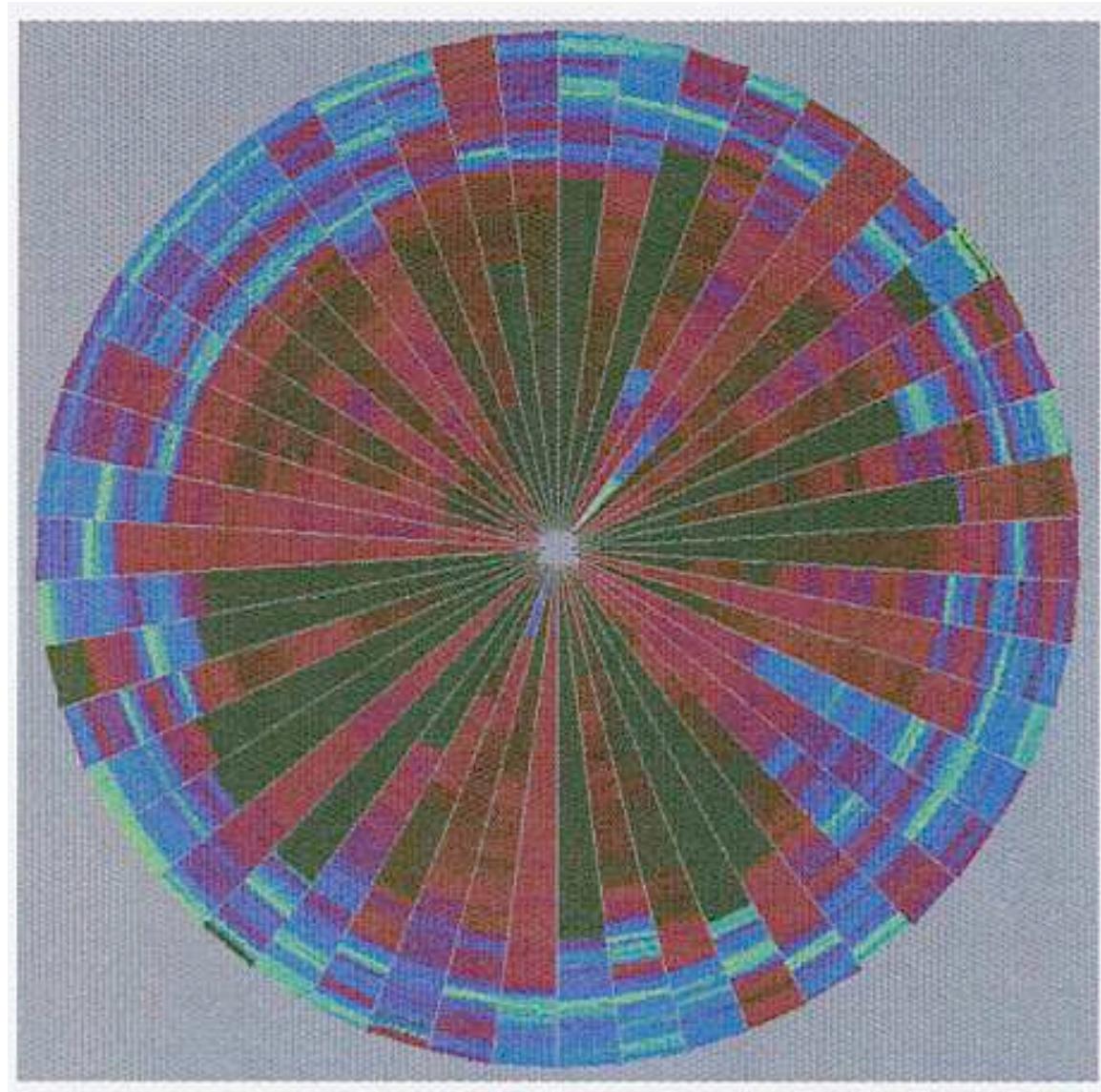


Arrangement in spiral form according to overall distance from the central point



Time-series displays

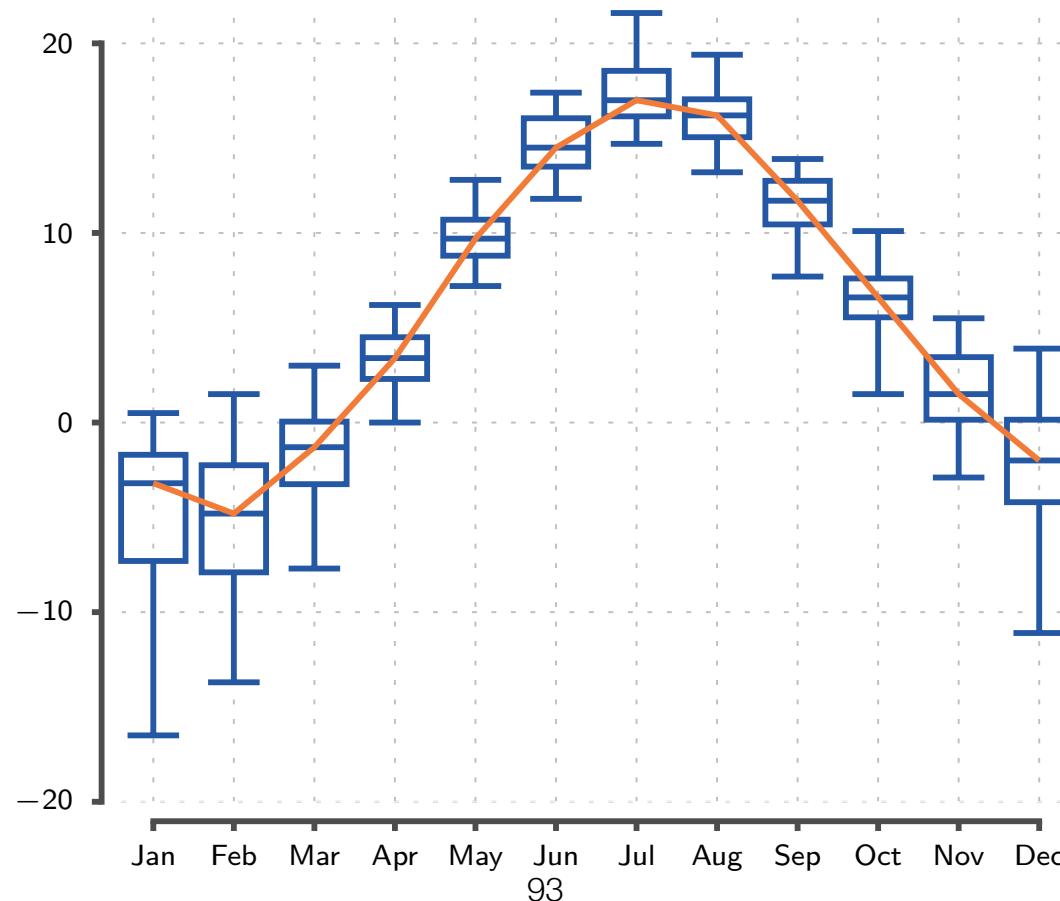
- **Pixel-oriented techniques**
 - Circle segments
(time goes out from center, creating a pseudoperspective)



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

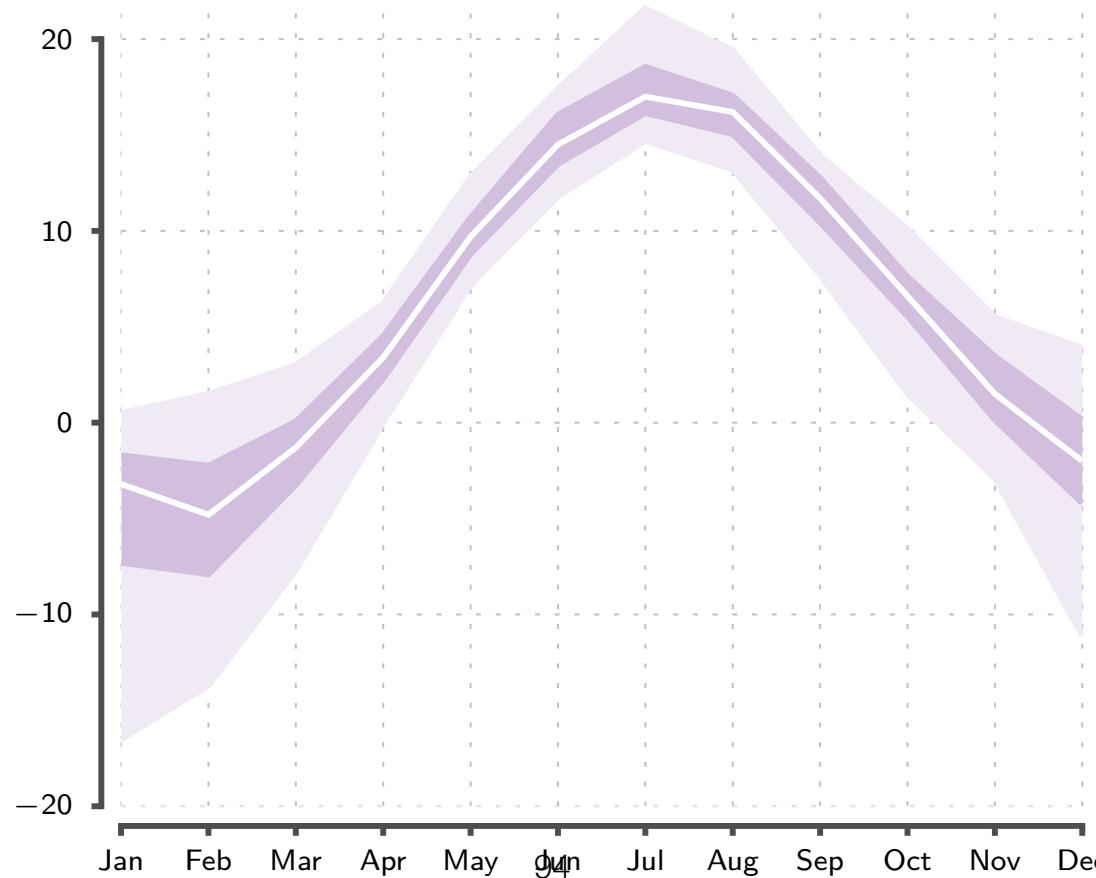
Time-series displays

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



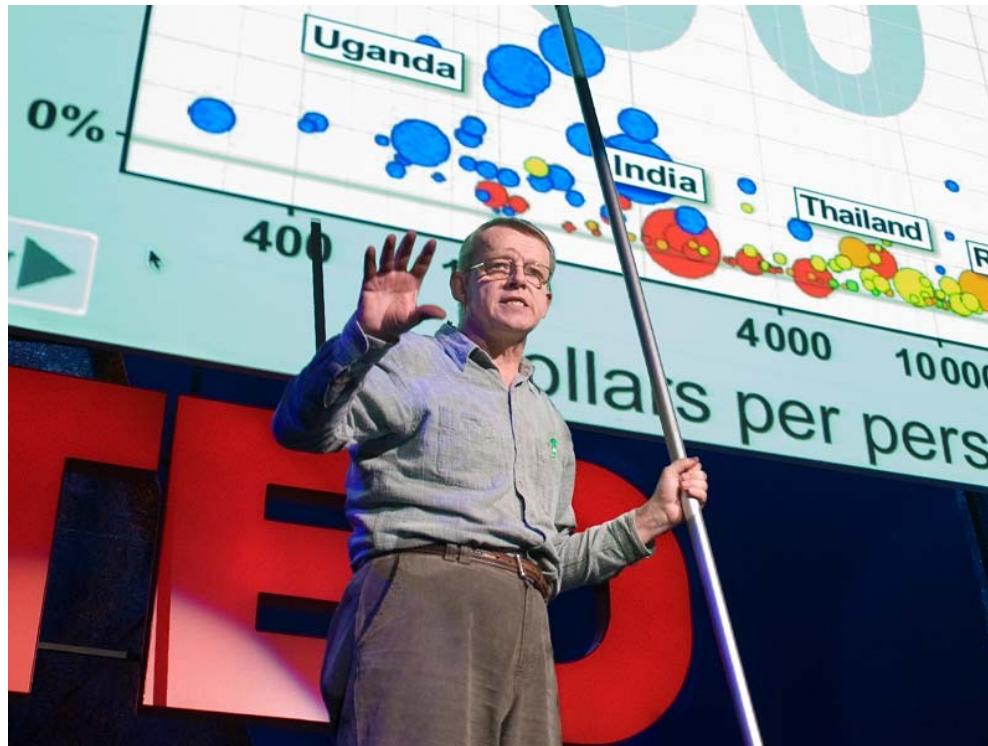
Time-series displays

- 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](https://ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen)

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

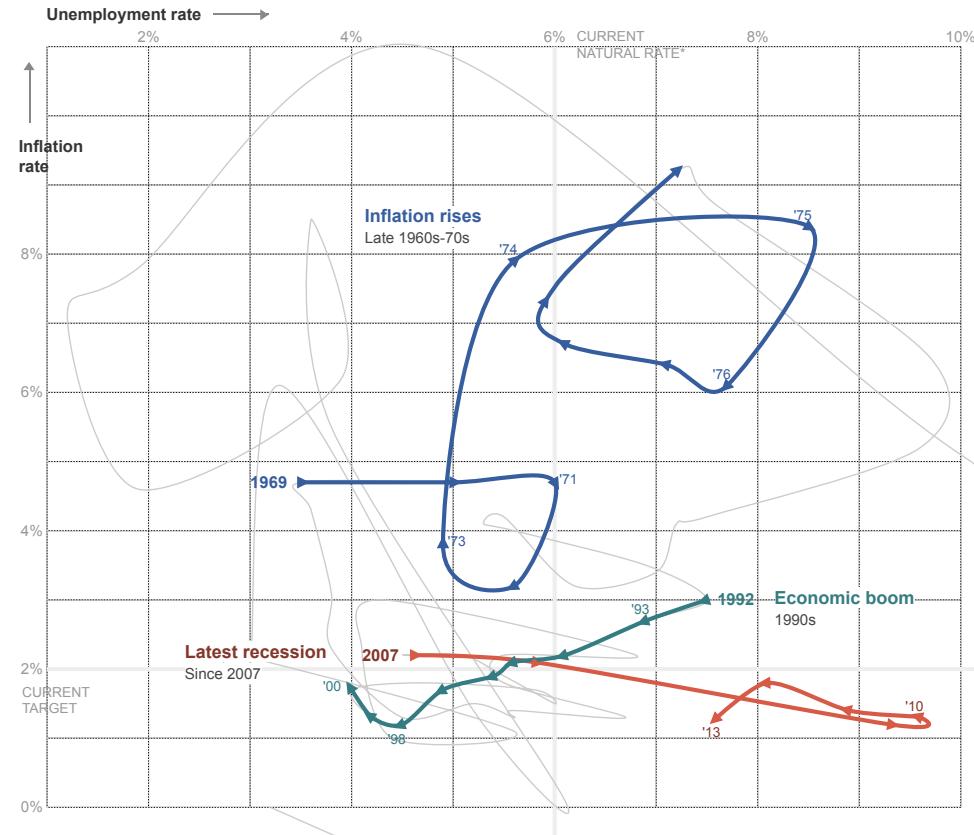
Time-series displays

- 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.

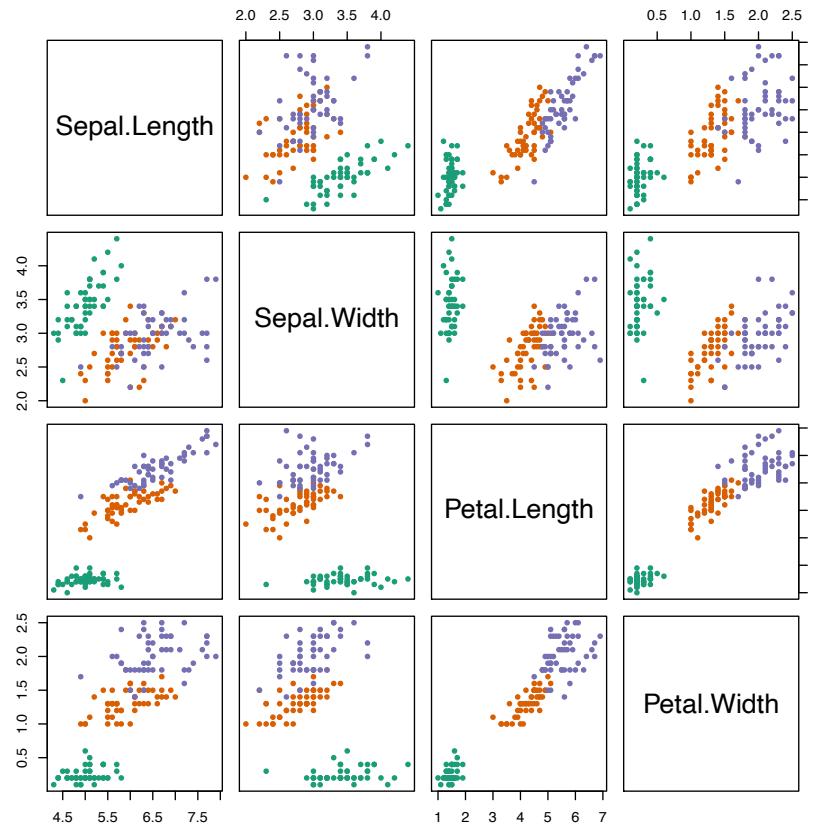


High-dimensional data

- 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

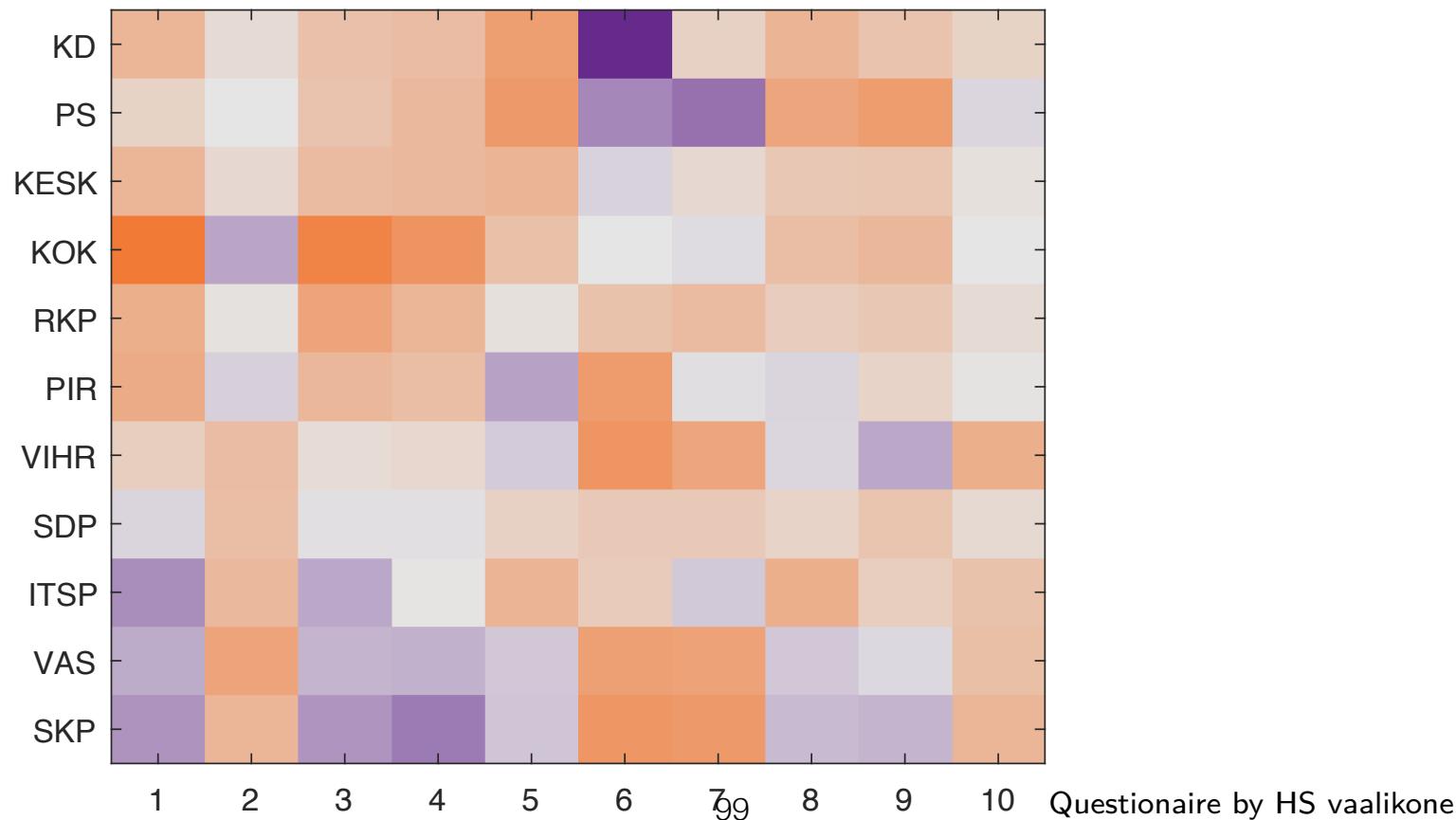
- Small multiples
 - array of a matrix of small plots
 - may result in an overwhelming plot



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

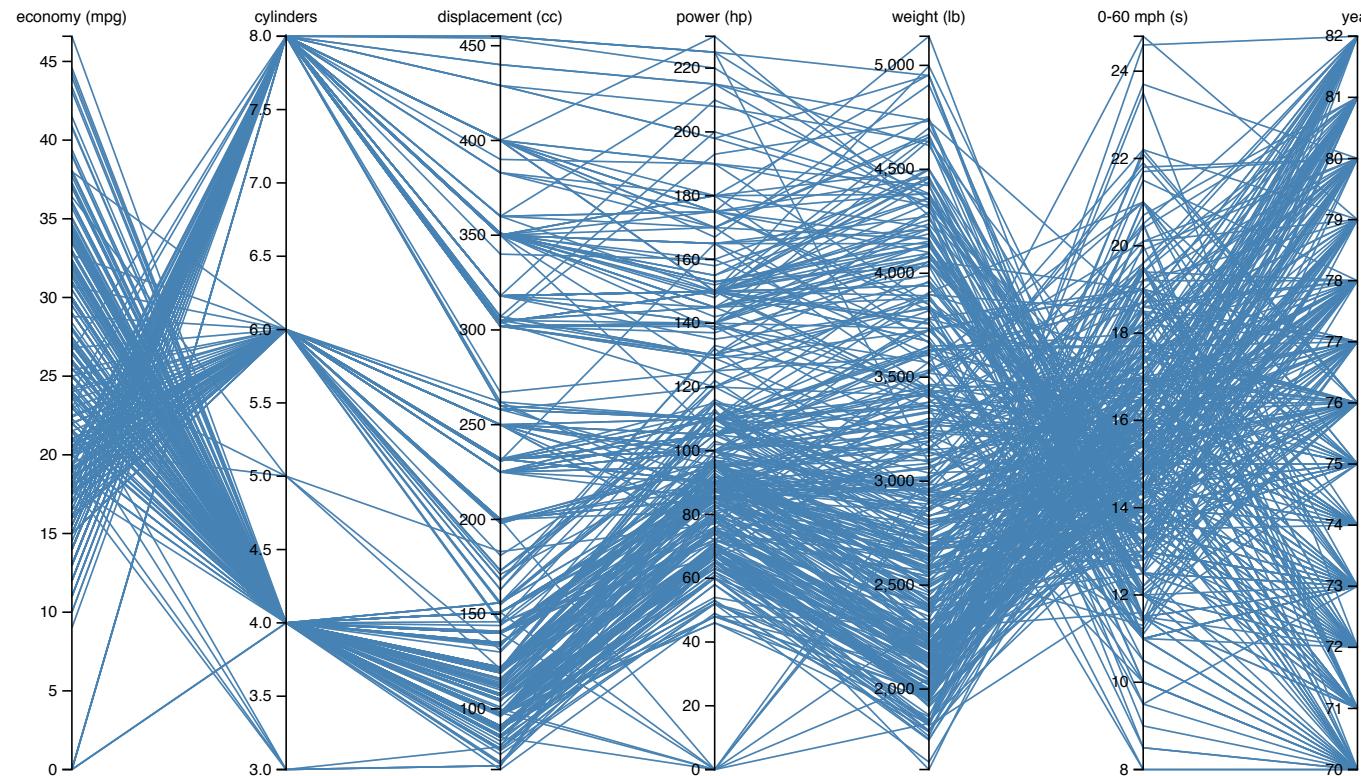
High-dimensional data

- 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



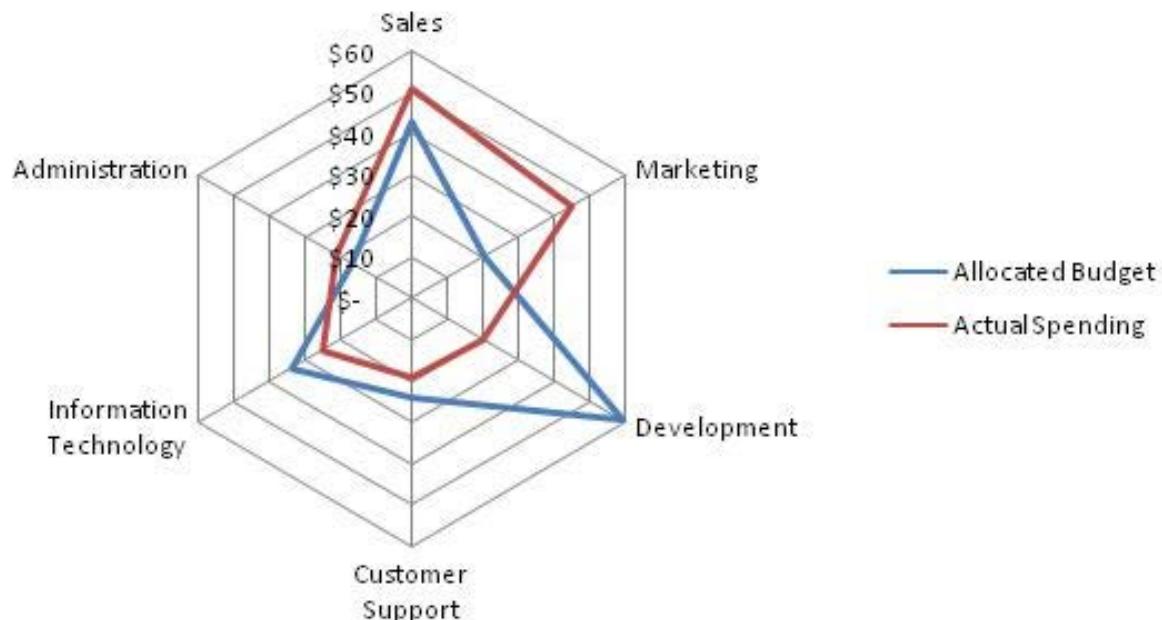
High-dimensional data

- 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



High-dimensional data

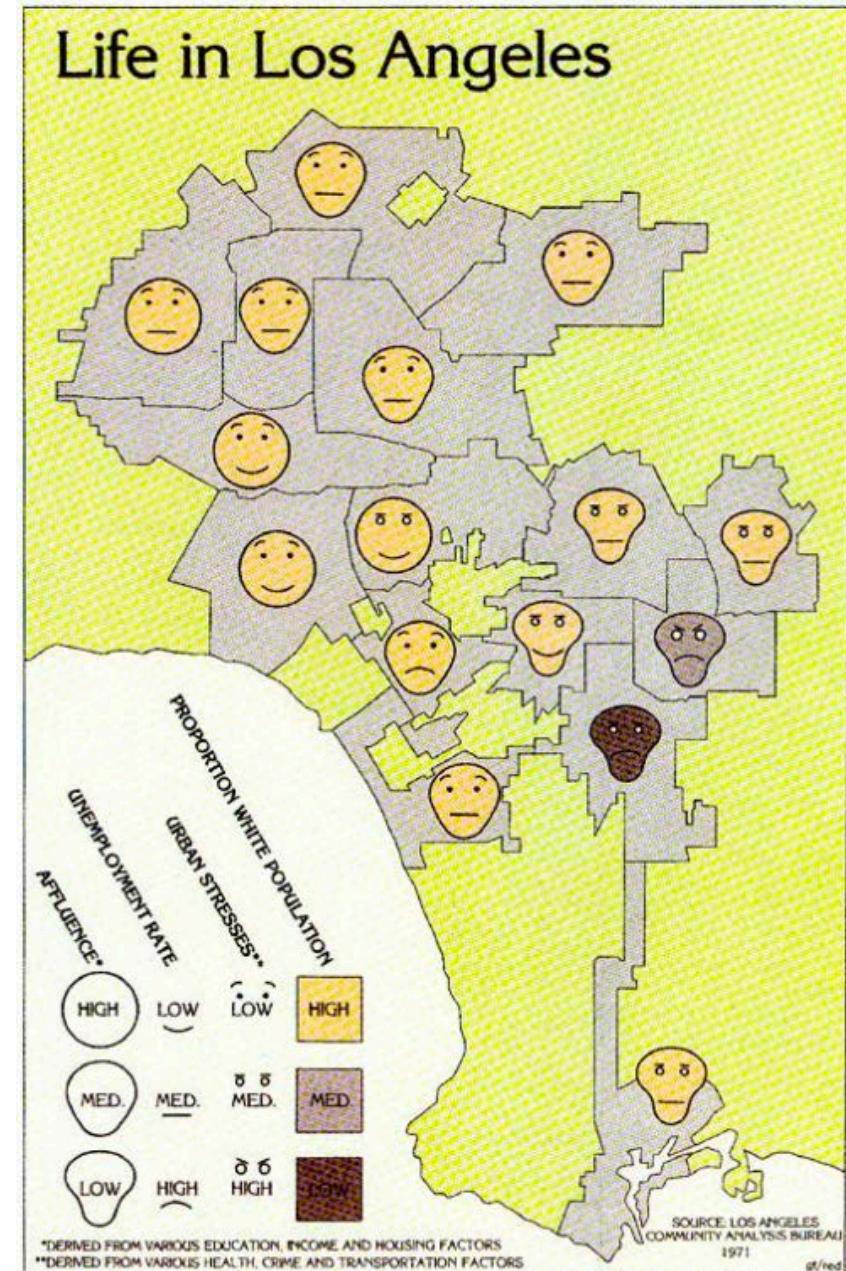
- Radar chart is equivalent to parallel coordinates plot, with the axes arranged radially
- May be useful, if
 - comparing overall similarity
 - consistent ordinal dimensions, e.g. performance scores
 - small/high values at the center/rim mean the same (better/worse) in all dimensions
- Generally not 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

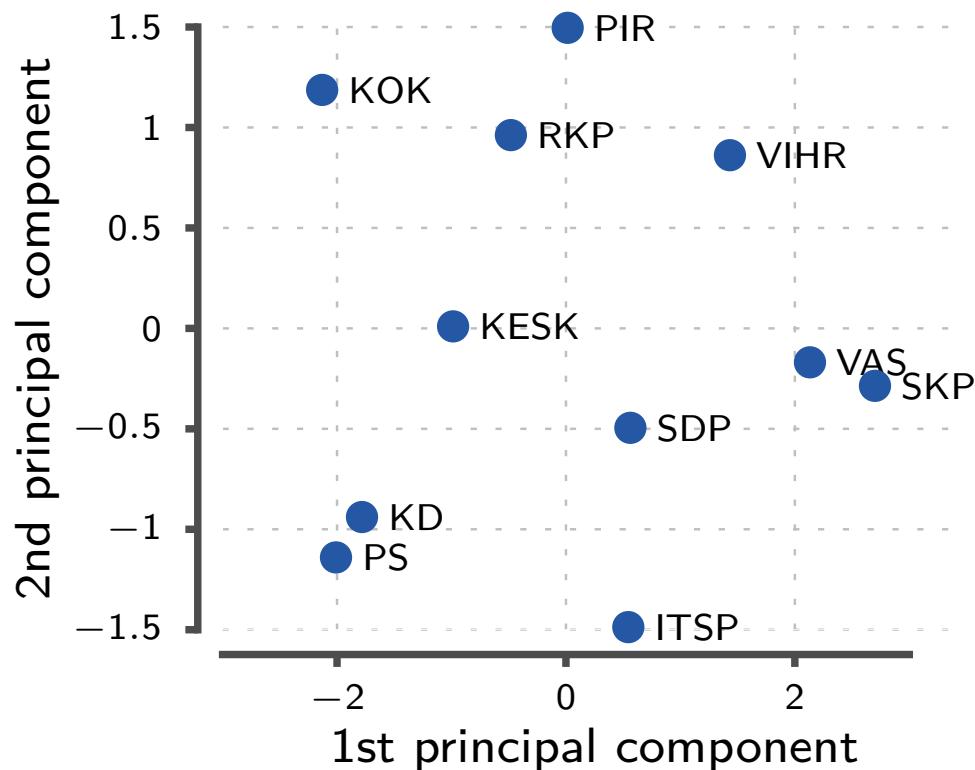


https://en.wikipedia.org/wiki/Chernoff_face

<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
 - rescale/re-express if the relative comparison is important
- use multiple plots if story/data requires it:
 - small multiples, overview/detail, multiform
- muted colors for surfaces, bright colors for objects
- use opponent colors [more during Part III]
- techniques for high-dimensional data [more in Part III]

Next lecture

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