

Basic statistics and data handling – Day 3

Introduction to Scientific Figure Design

Aiora Zabala

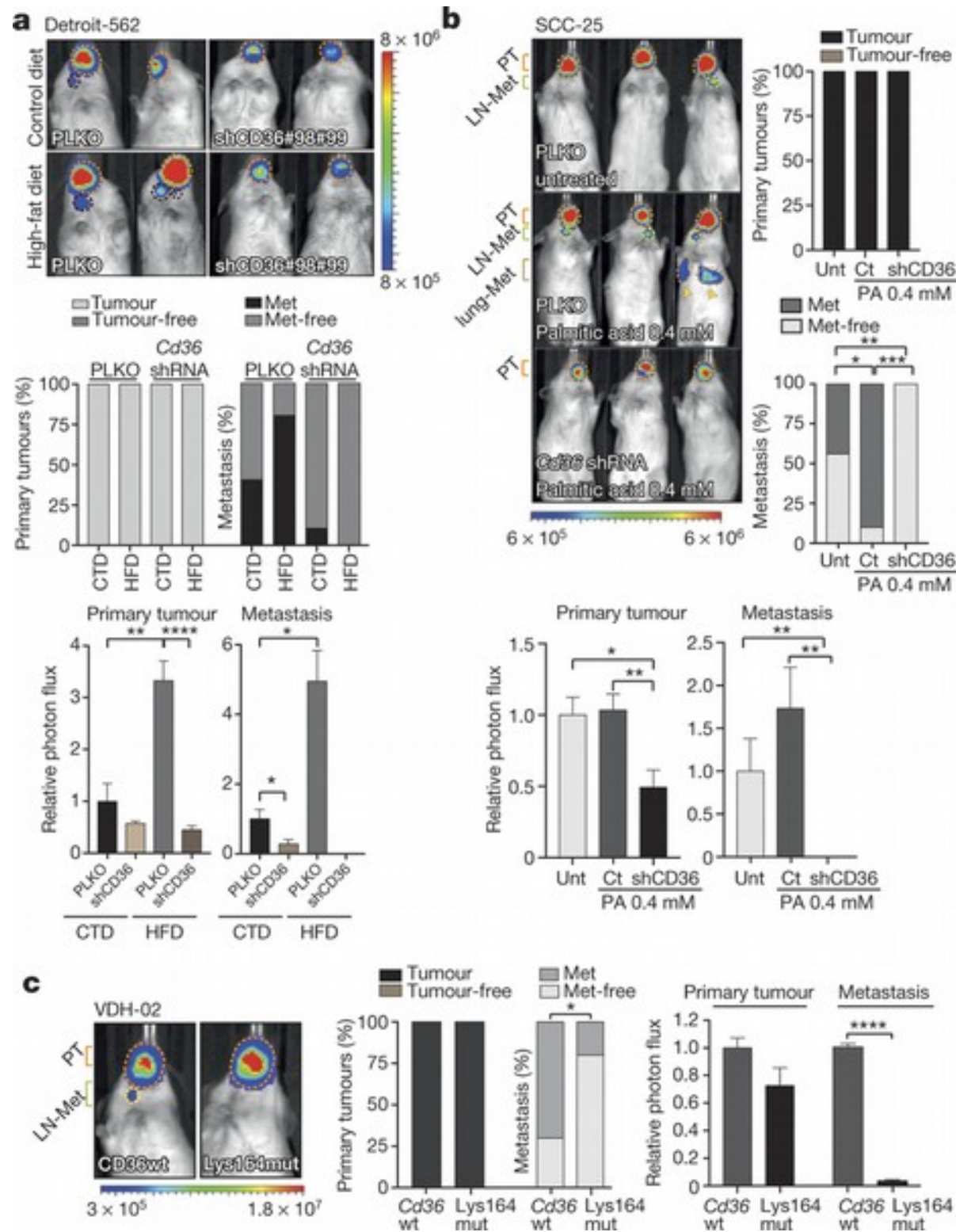
Cancer spread is increased by a high fat diet, ground-breaking evidence shows

Researchers discover new cancer spreading protein

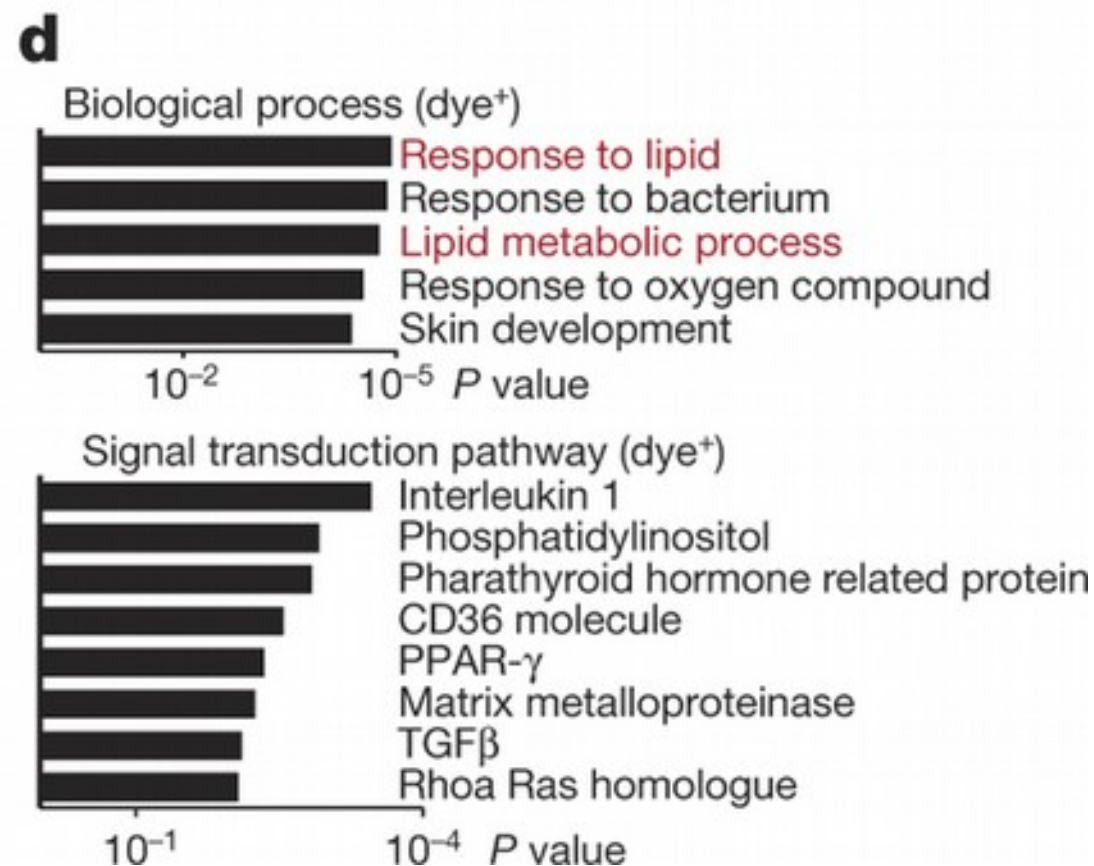
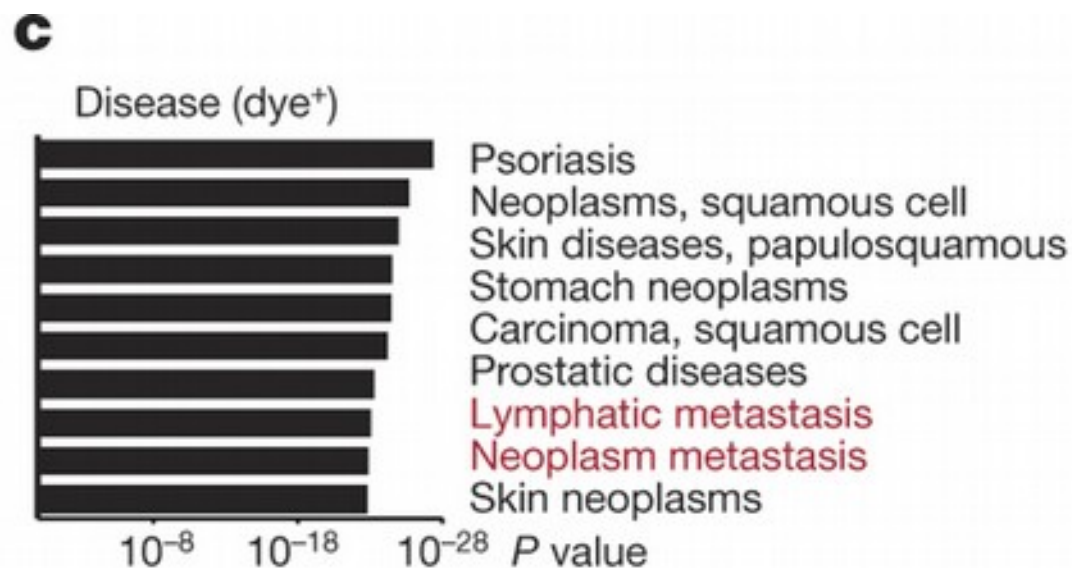
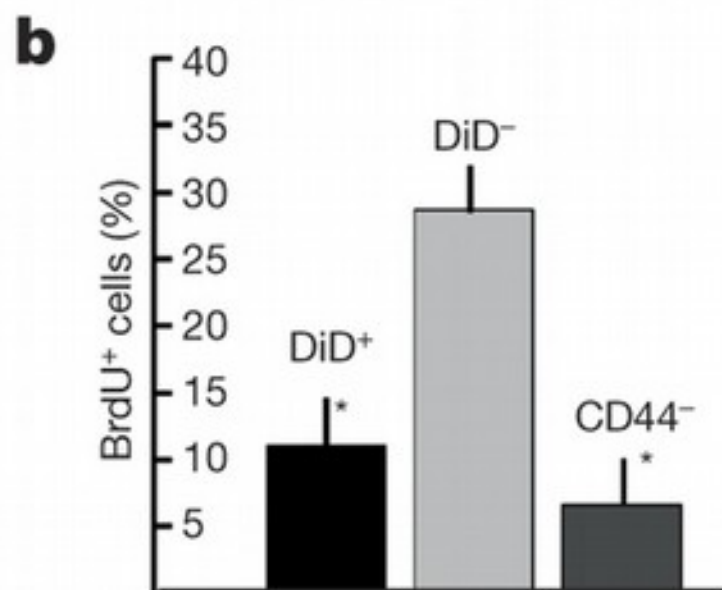
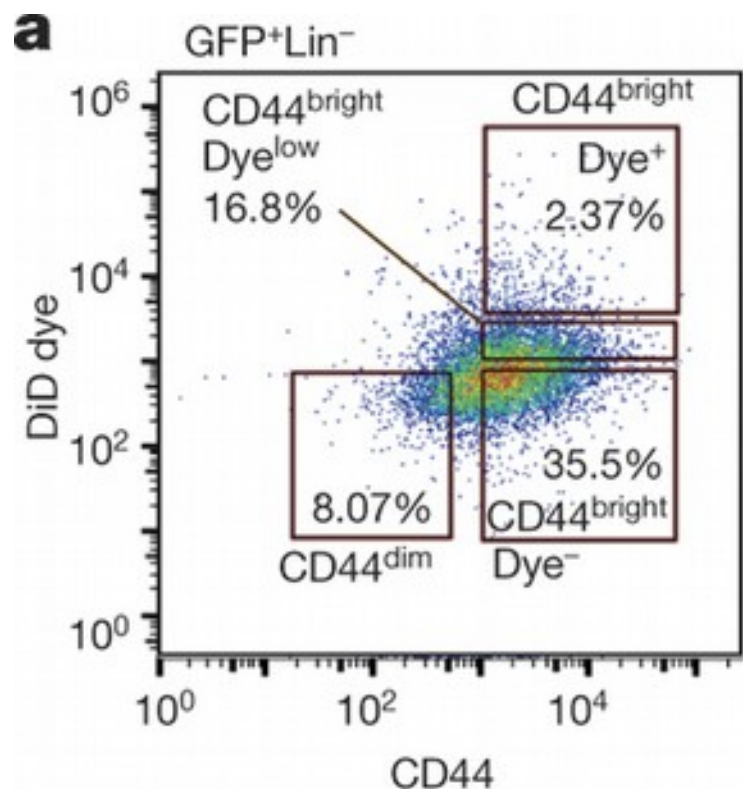
Date: December 7, 2016

Source: Worldwide Cancer Research

Summary: New research shows that the metastatic process (cancer spread) is enhanced by fat intake. Mice given a high fat diet, including palmitic acid (a major component of palm oil which is found in lots of household products) developed the most aggressive cancer spread. The study identifies for the first time a protein called CD36 which has an essential role in cancer spreading.



Pascual et al. Targeting
metastasis-initiating cells
through the fatty acid
receptor CD36.
Nature. 7 December 2016



What is figure design? *Why* design?

‘is not to take bad scientific content and disguise it as great [...] the goal is to **communicate great content in a clear, succinct, and inspiring way** [...] in the best possible light’

‘is not decoration [...] is not adding anything meaningless that lacks information or **purpose**’

Carter

‘Design should never say, ‘Look at me.’
It should always say, ***Look at this.***’

Craib

Goals of this session

Theory

- Explain the key ingredients for sci-figure design
- Discuss what works, what doesn't and what's unethical
- How to choose a type of figure that is appropriate for the data

Practical

- Use Inkscape to apply the theory (free vectorial image editing software)
- Produce a journal-ready figure using R and Inkscape

Structure of this session

Theory

1

Why figure design?
Principles of figure design
Elements of a figure
Dealing with complexity

3

Choosing the right figure
Colour
Typography
Composition & layout

Practical

2

Hands-on Inkscape
Document properties
Create & manipulate
objects

4

Colour
Composition
Import, save & export
for journal submission

What is data visualisation?

Visual representation of data to communicate information clearly and “help people carry out tasks more effectively”

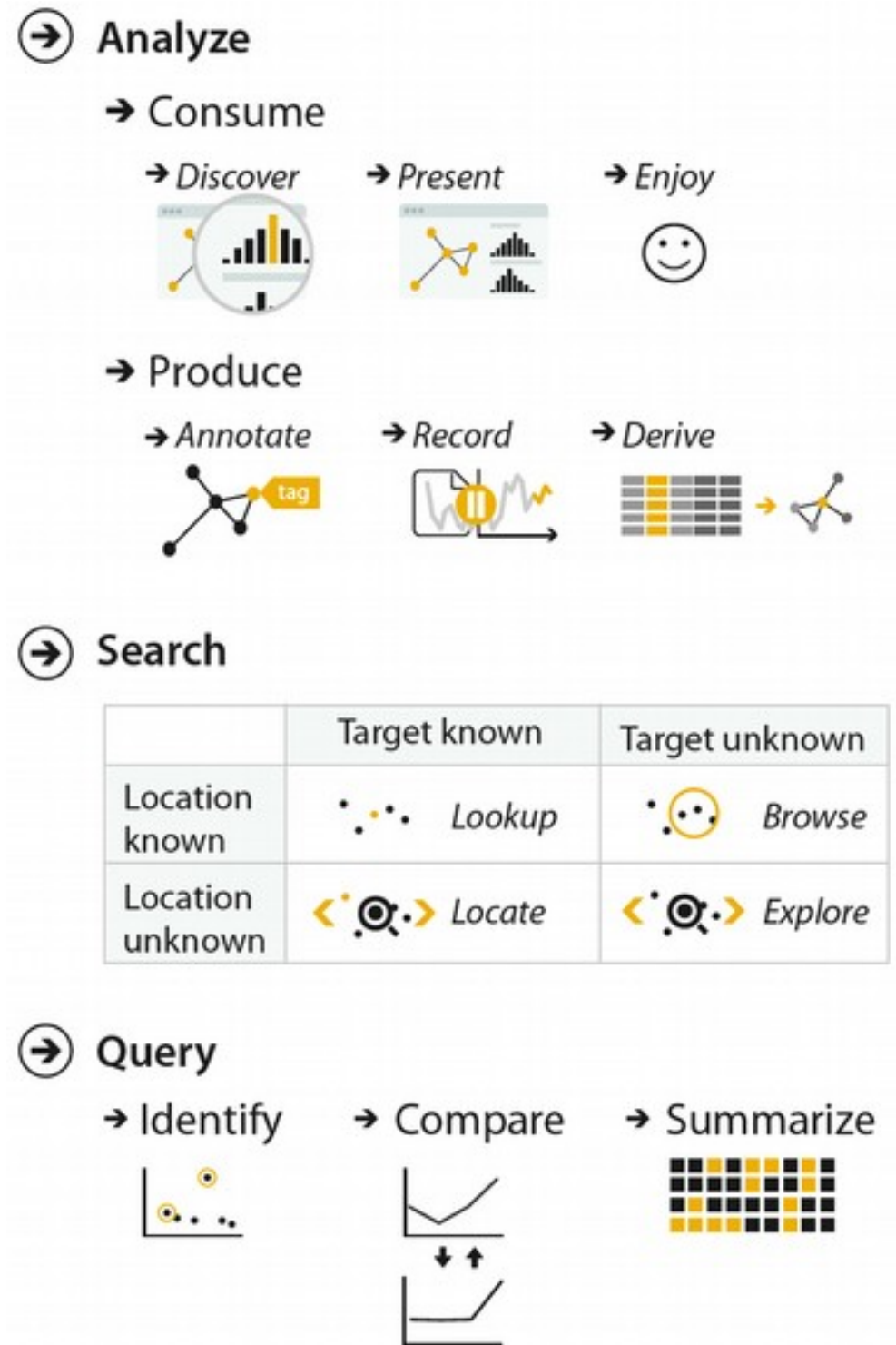
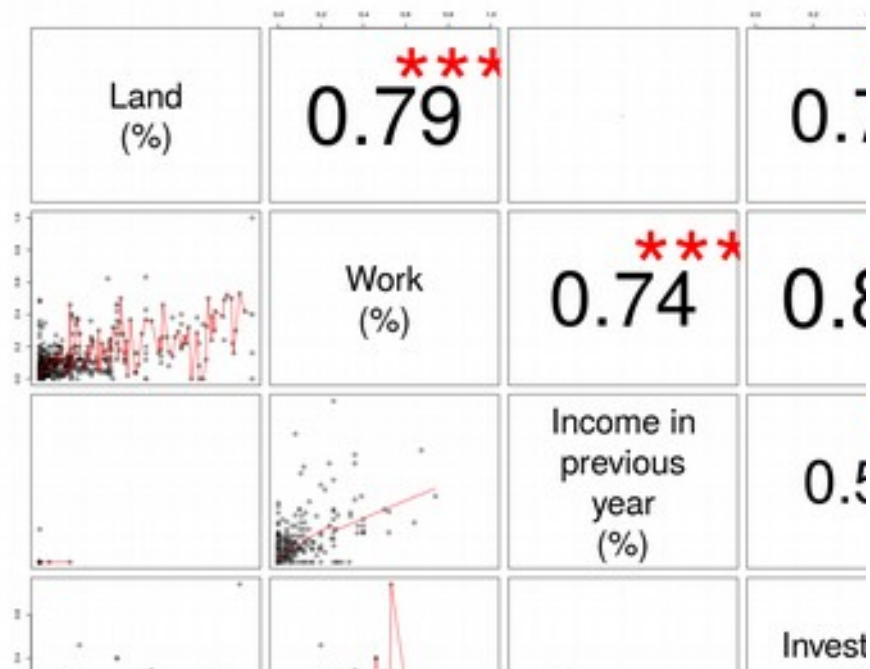


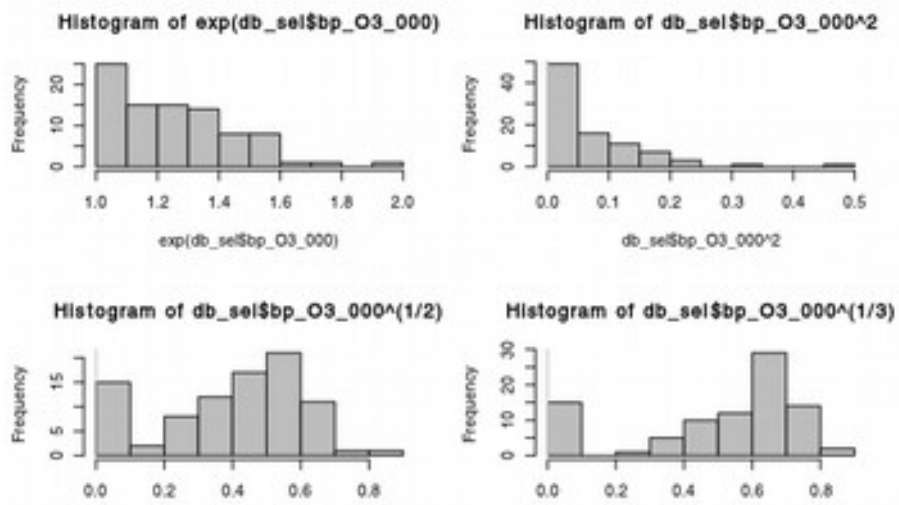
Figure: Tamara Munzner

Exploratory visualisation

- Understand your data
- Multiple ways to present and summarise
- Crude representations
- Interactive
- Not intended for final publication
 - Can be adapted for publication



Correlogram to see bivariate relations



Histograms to see the **distribution** of variables



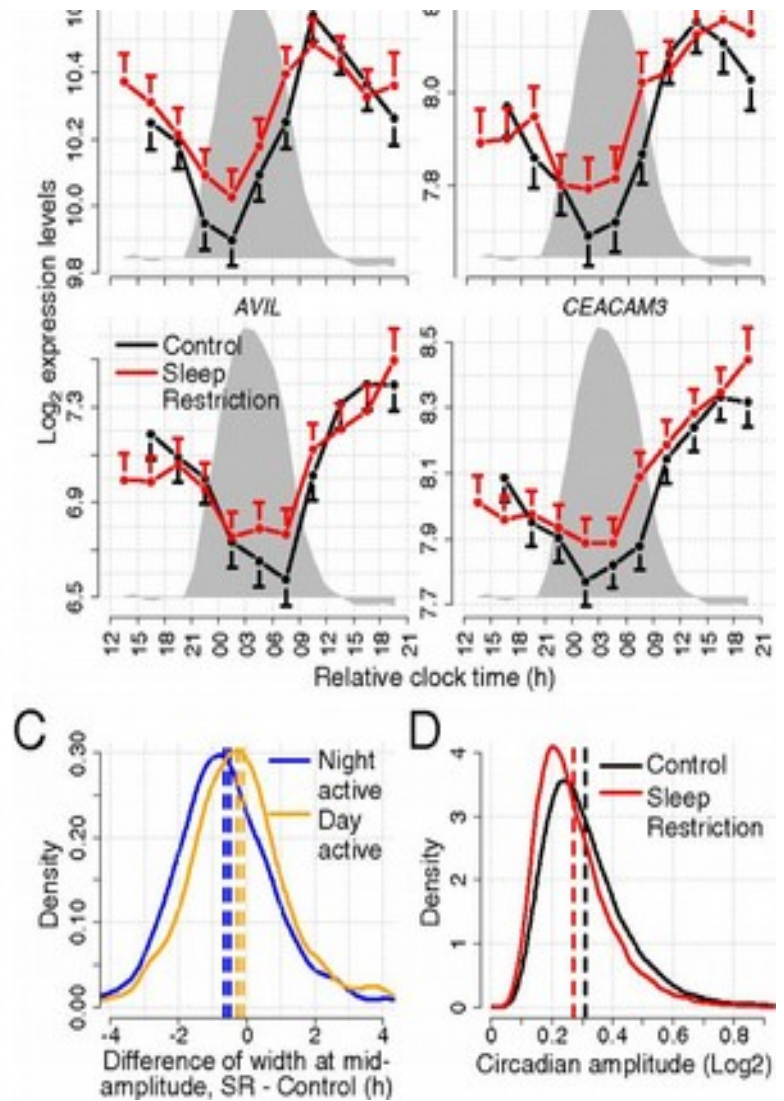
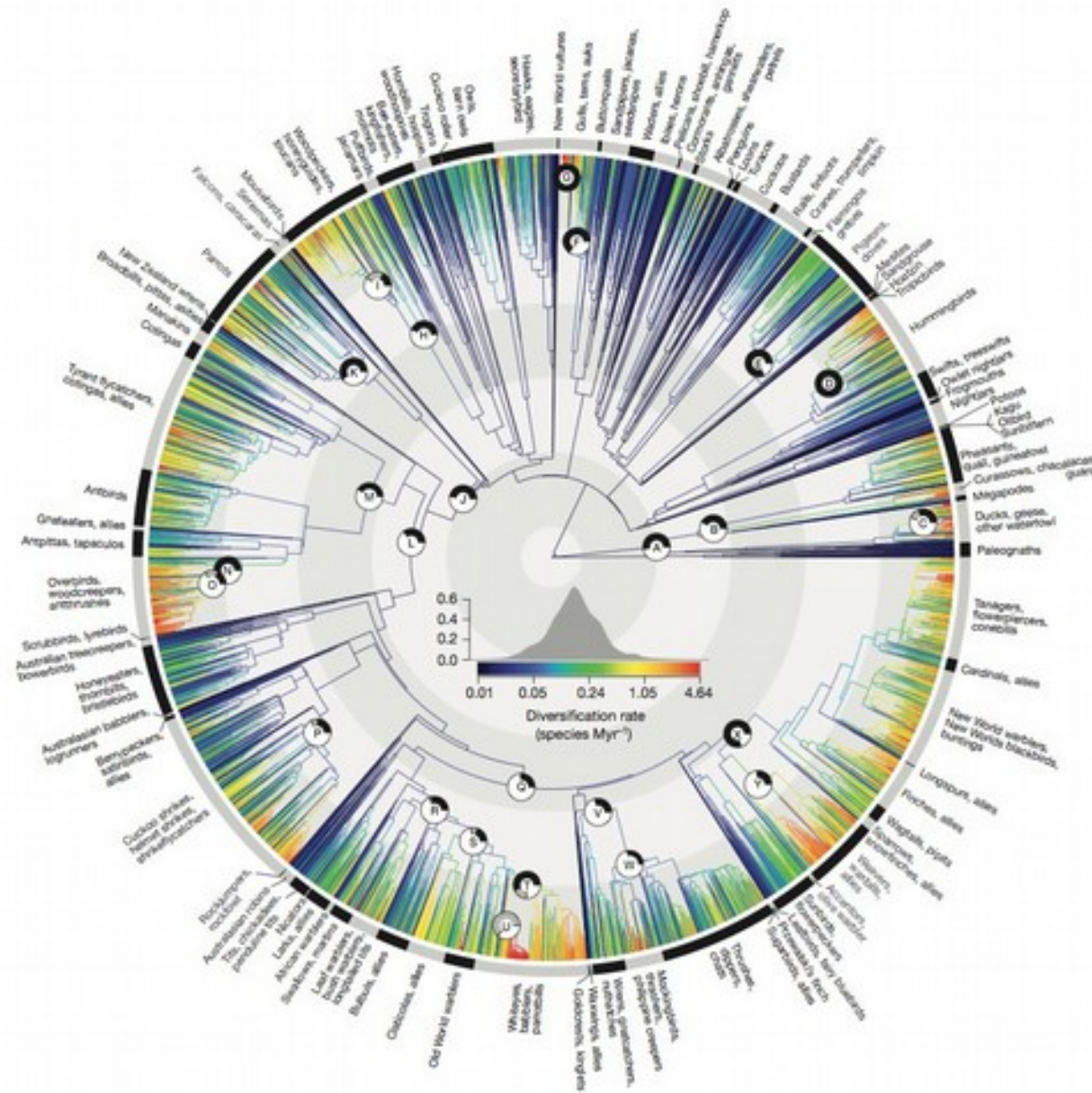
Interactive data exploration with the R package **ggobi**

Reference visualisation

- Using your data as a resource
- Allows users to look up data of interest
- Tabular / Configurable
- Interactive

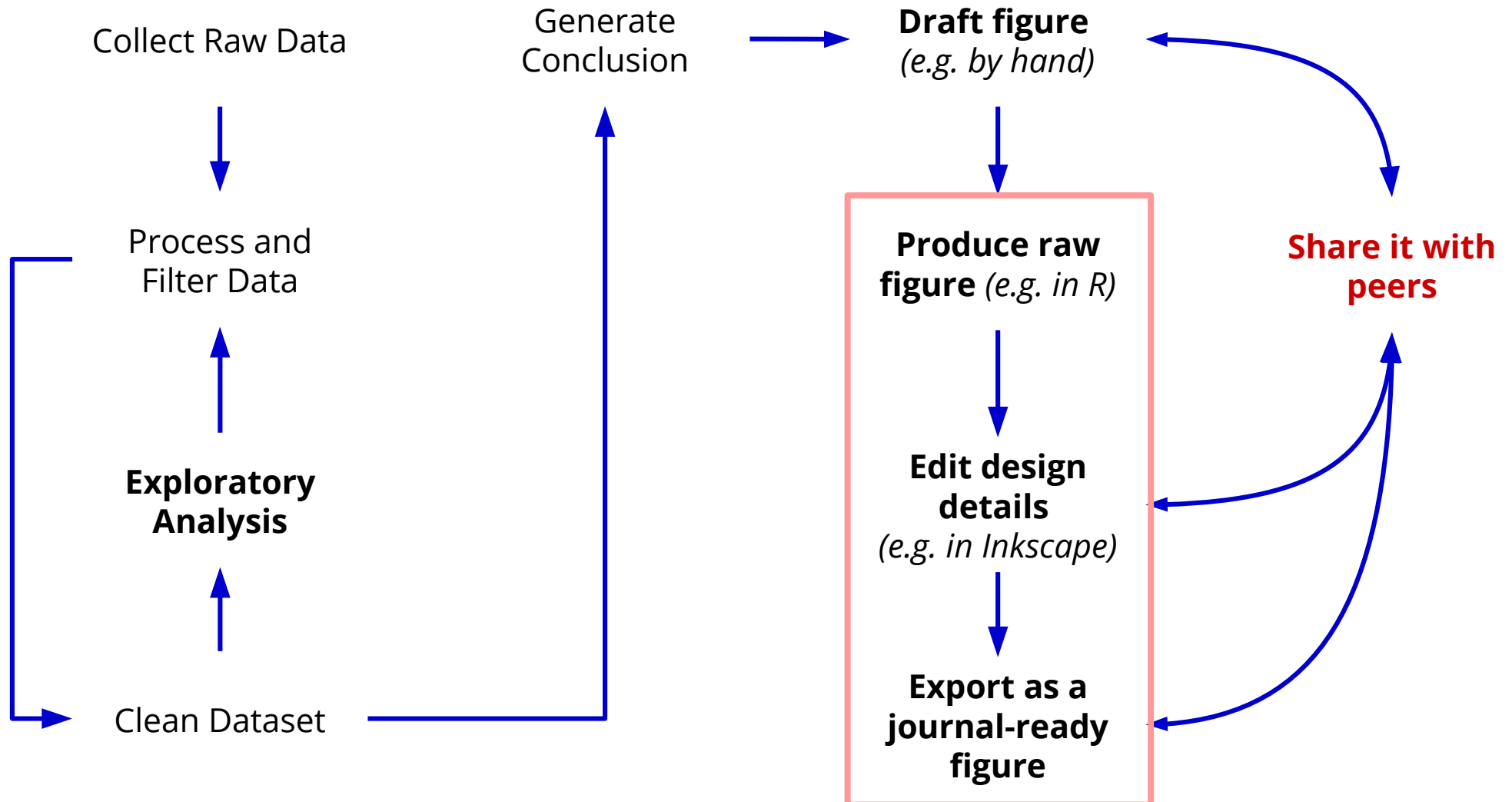
Illustrative visualisation

- Intended to convey a specific point
- Carefully chosen subset of data
- Optimised presentation
- Good design
- Used for figures in papers

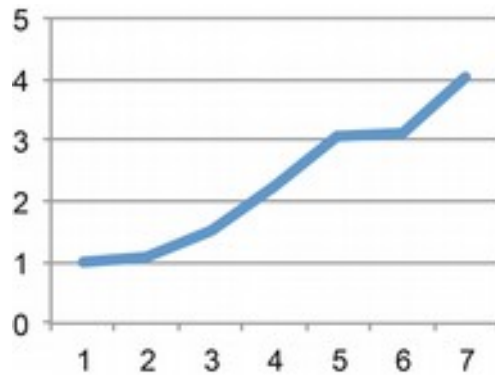


Figures: Avian phylogenetic tree, Jetz et al 2012. Sleep deprivation and genetic expression Möller-Levet et al 2013

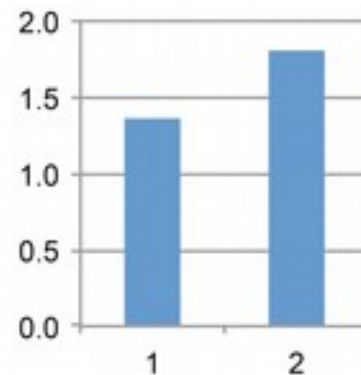
Data Visualisation Process



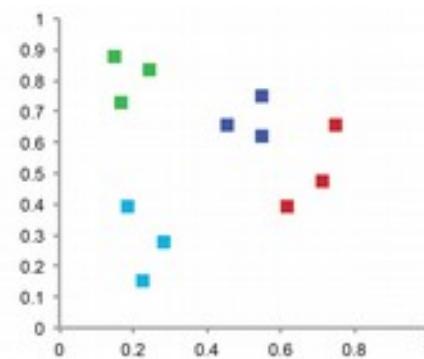
Things you can illustrate



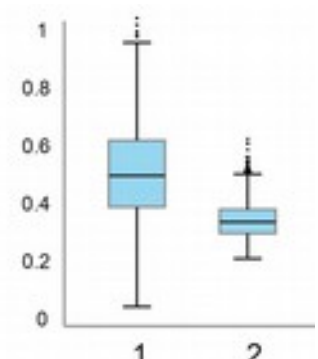
Relationship



Comparison



Composition



Distribution

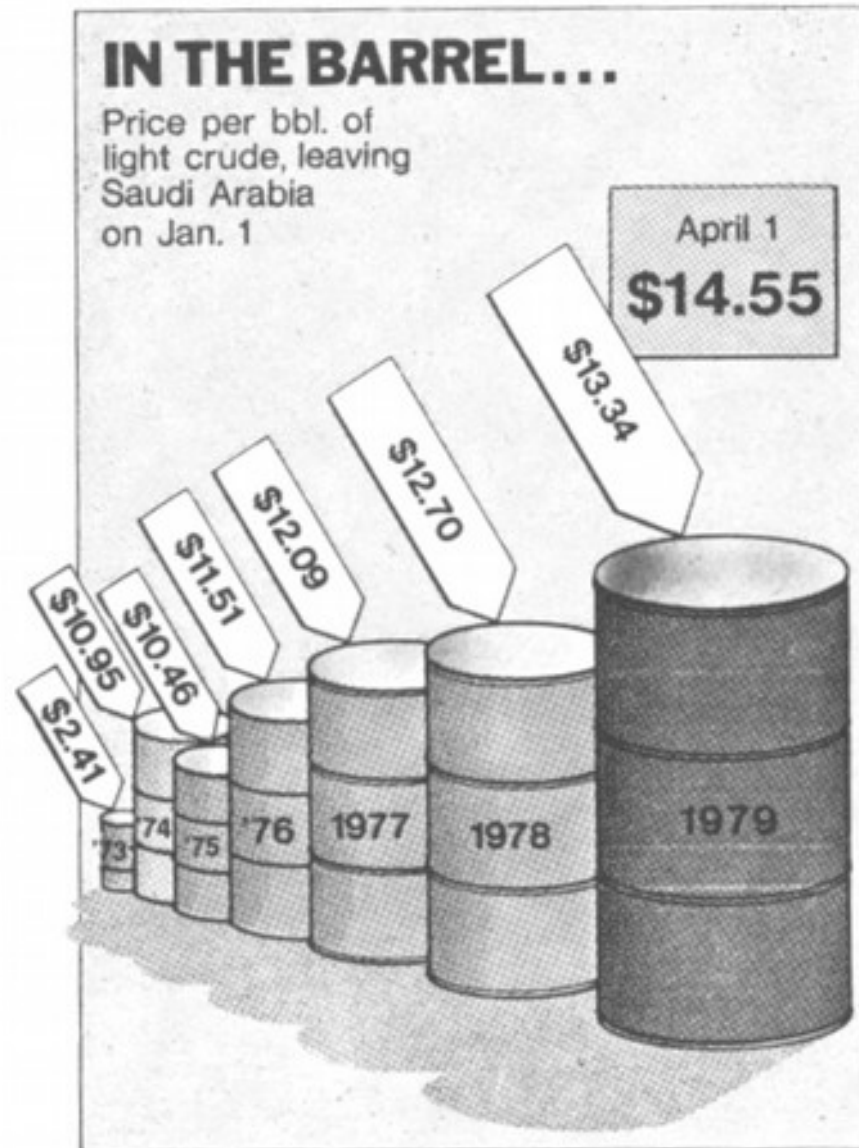
Graphical Representations

- Basic questions:
 - How are you going to **turn the data into a graphical form** (weight becomes length etc.)
 - How are you going to **arrange things in space**
 - How are you going to use **colours, shapes** etc. to clarify the point you want to make

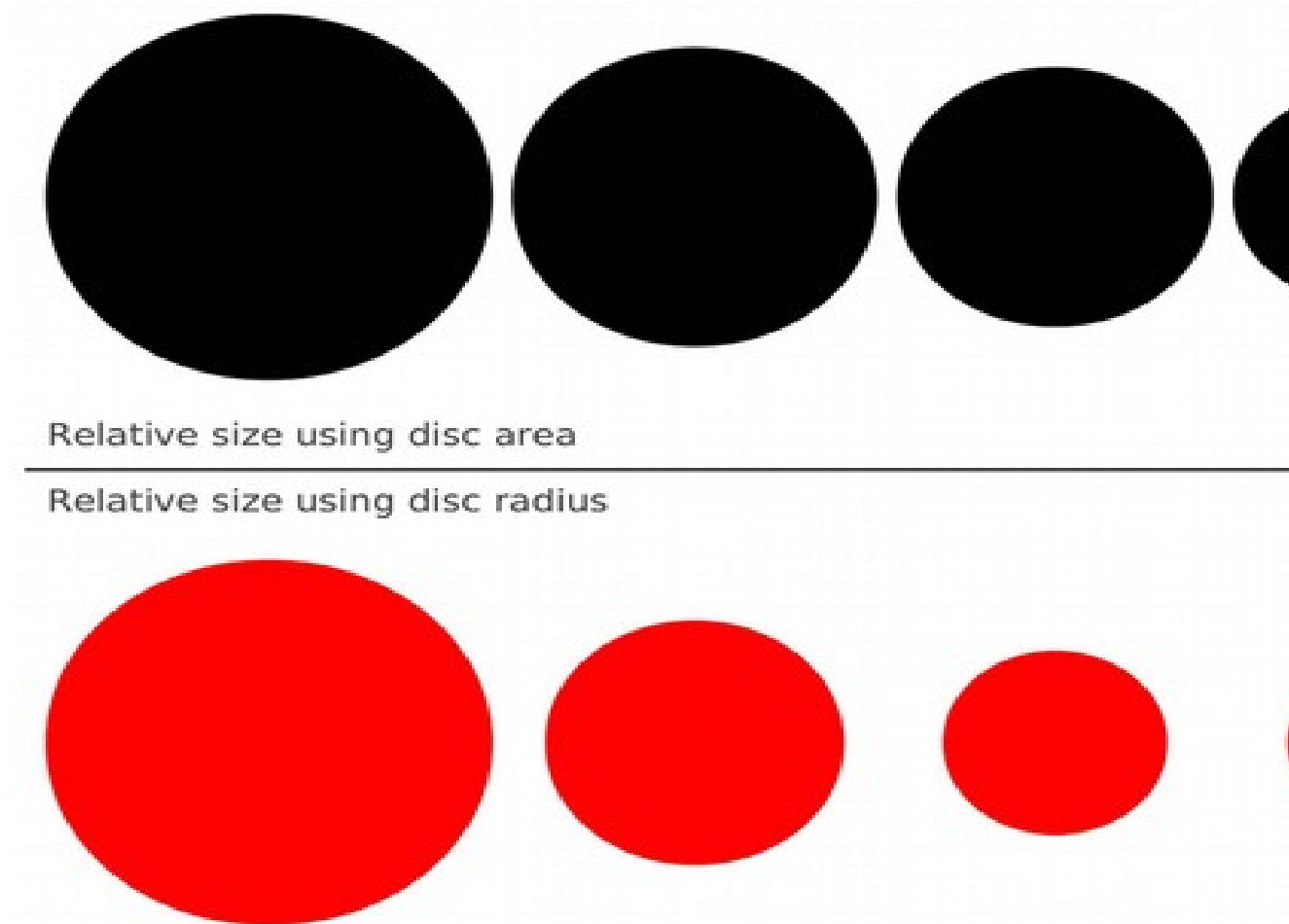
Key ingredients: principles

- **Simplicity**
- **Continuous** process
- *Rules* in graphic design are, as in many other disciplines, rather guidelines: you can break them to allow for creativity and when there is a good reason to break them, but you need to know how to use them.

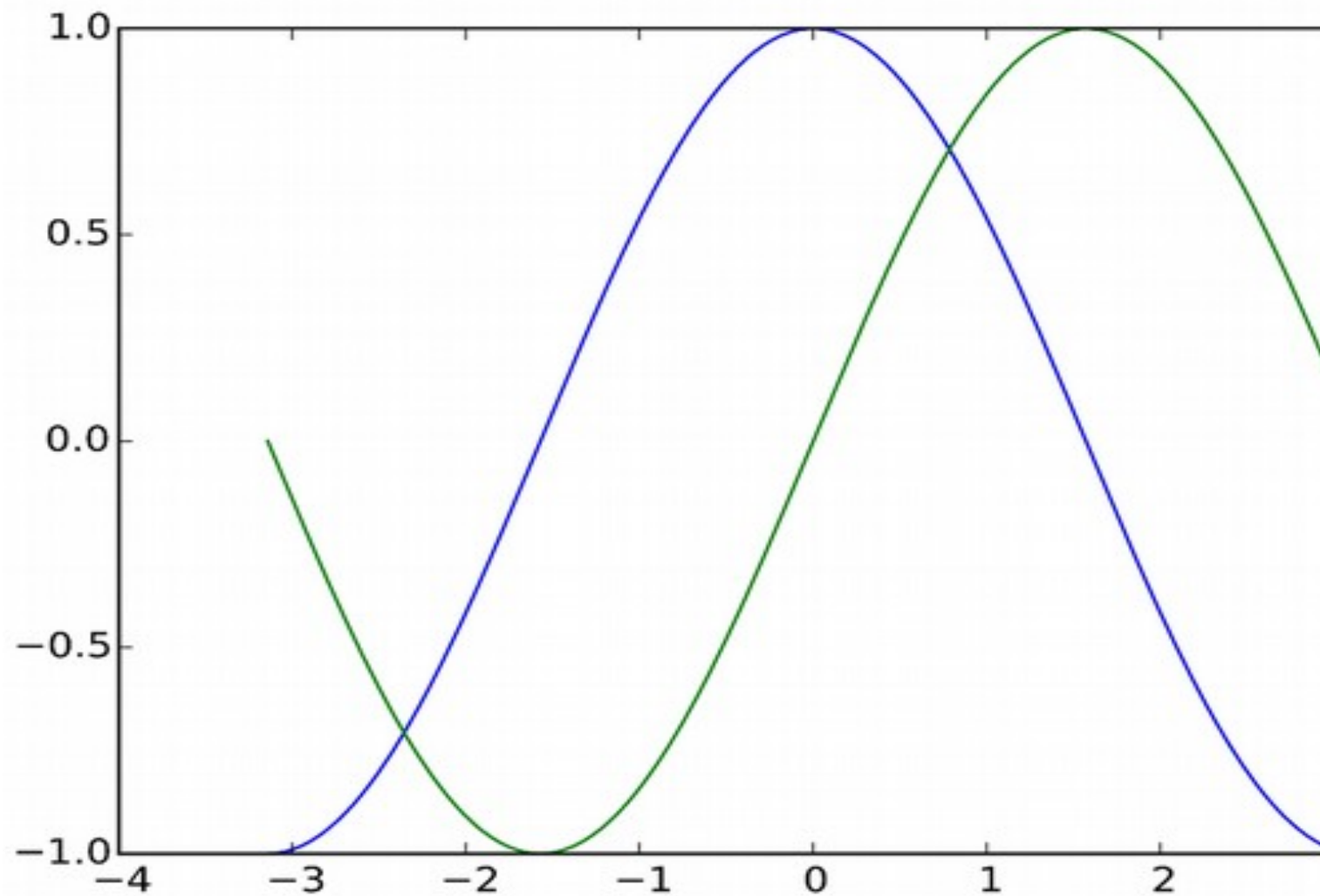
What makes a good figure?



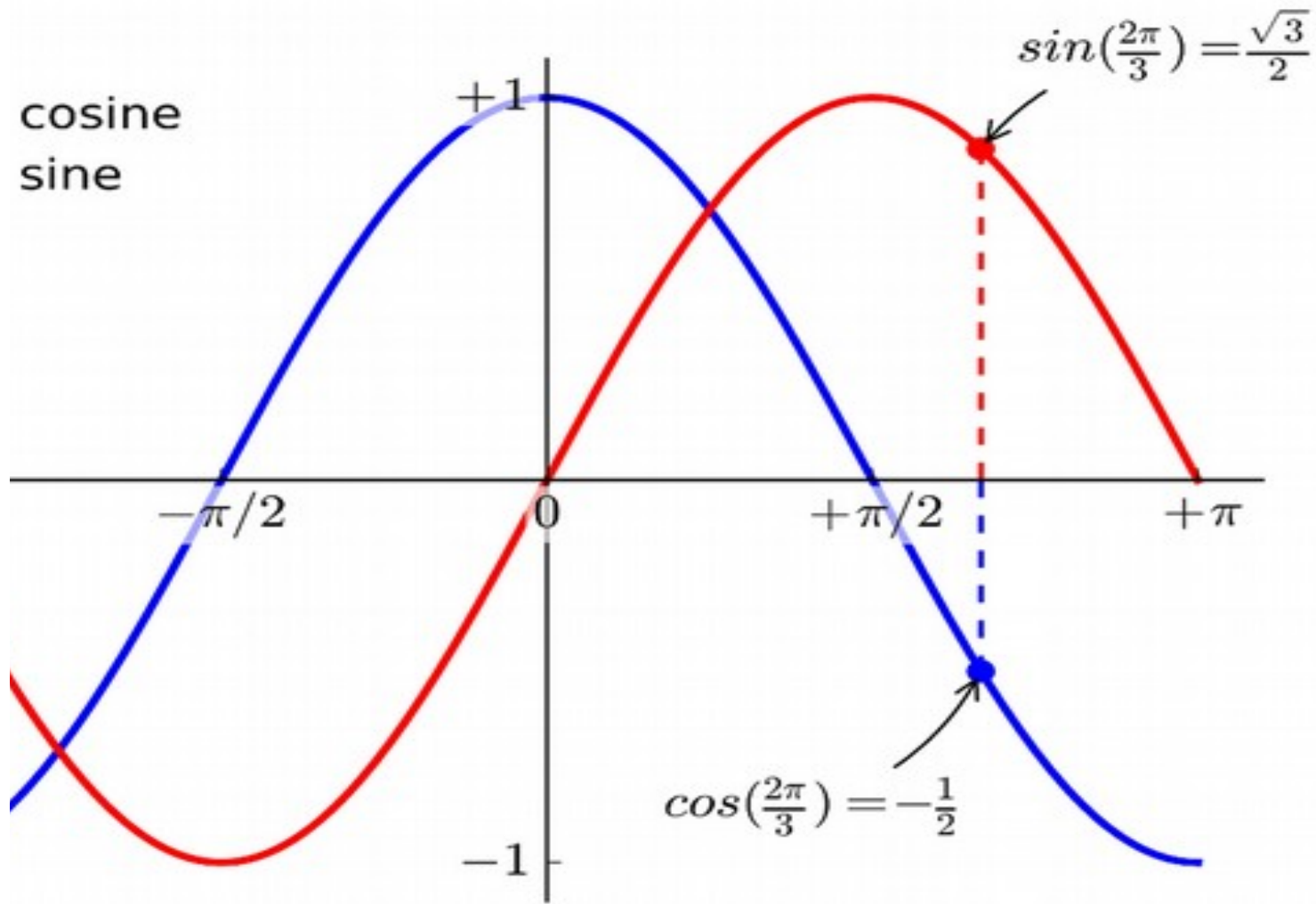
Areas and 3D can be misleading



What makes a good figure?



What makes a good figure?



What makes a good figure?

- Has a clear message
 - Helps to tell a story
 - Adds and relates to the text
- Is focused
 - Don't confuse one message with another
- Is easy to interpret correctly
 - Good data visualisation
 - Good design
- Is a honest and true reflection of the data

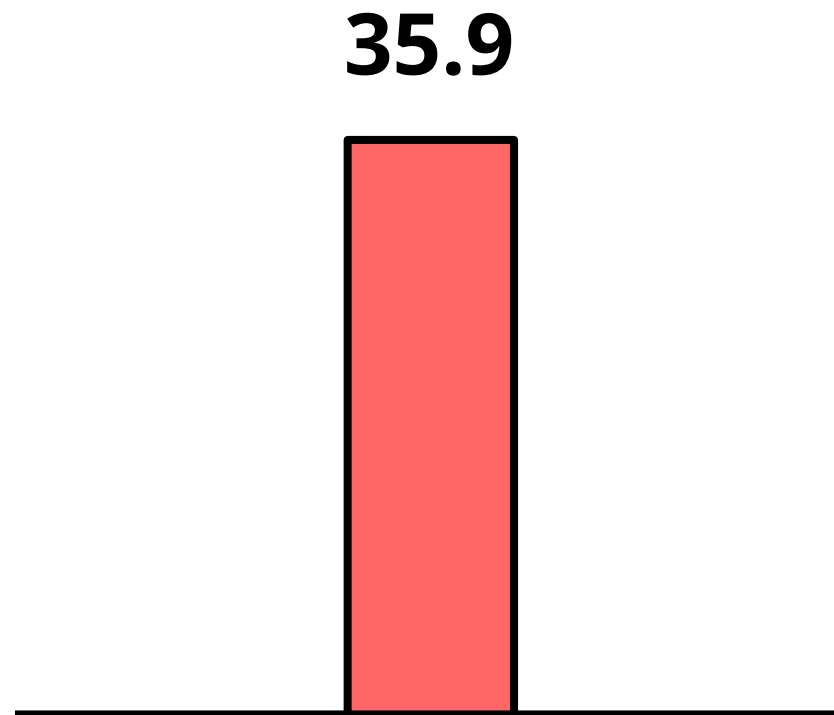
A data visualisation should:

- Show the data
- Link to the accompanying text and statistics
- Summarise to make things clearer
- Serve a clear purpose
- Not distort the data

Simplicity

- Every single element has to be there for a reason, 'distinguish between what is **meaningful** and what is **unnecessary** [...] avoid the latter'
- Simplicity is not boredom, but effectiveness in communicating a message and leaving aside anything unnecessary
- Avoid confounding decorations, e.g. excessive background grids or frames

Simplicity



This figure indicates altitude in **six separate ways**.
Can you find them?

Consistency

- Make the figures uniform to help viewers understand the figure
- Try not to use more than two types of these
 - Font styles and sizes
 - Line weights (thickness)
- When combining more than one chart
 - Use the same colours and shapes for the same groups
 - Use same sizes and scales for comparable charts
 - Position of axis titles and labels
 - Stick to your own rules, e.g. if presented 'Sample A' and then 'Sample B', maintain this throughout

Some useful concepts

- **Data-ink ratio** (and non-data ink)
- **Data density** of a display:
high-info graphics and
the shrink principle



Edward Tufte

Key ingredients: the tools

Elements: marks and channels

- Data
- Points, lines, areas
- Colour
- Typography

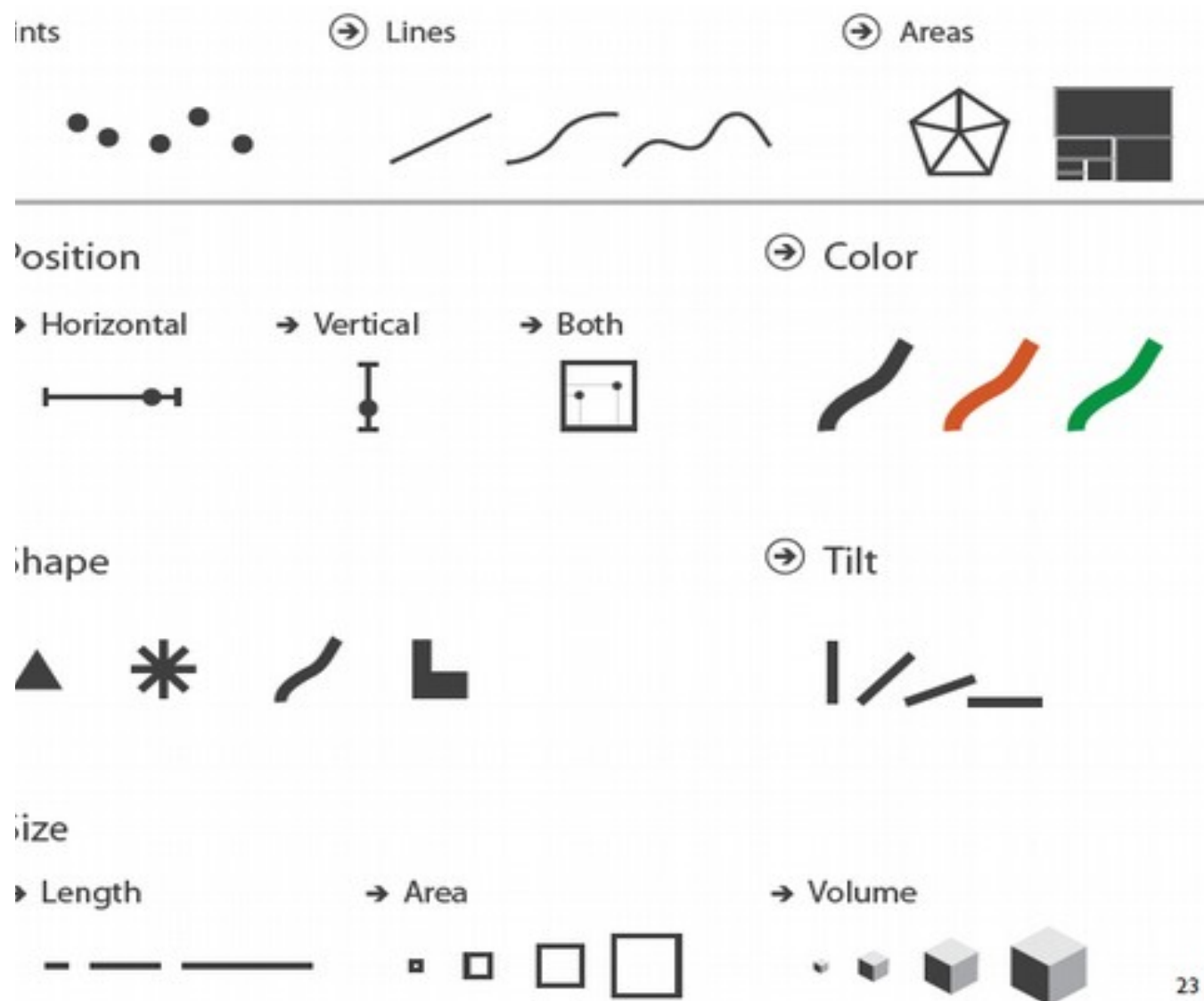
Composition

- Grid and alignments
- Balance
- Hierarchy and focus

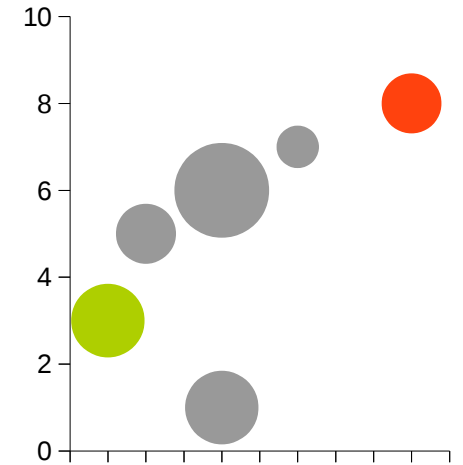
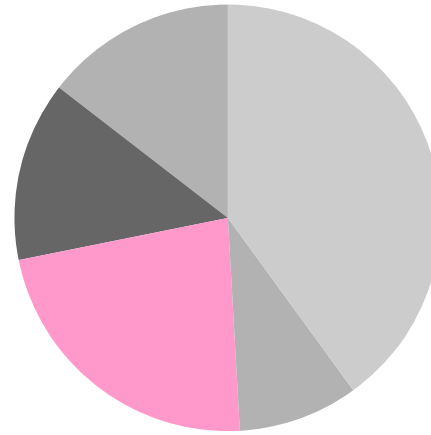
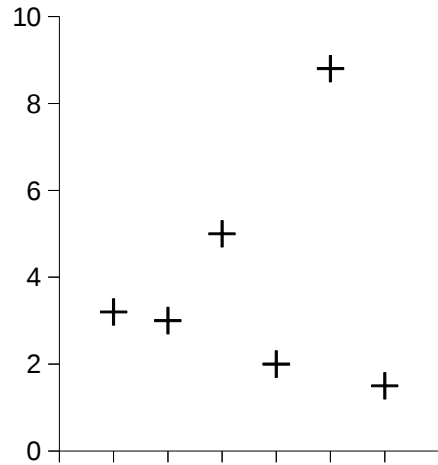
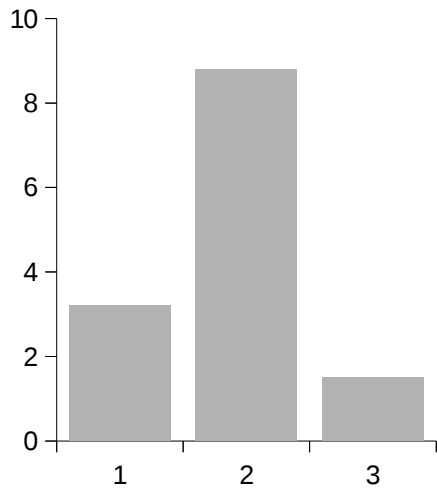
Elements: Marks and channels

Marks (geometric primitives): used to **represent** data

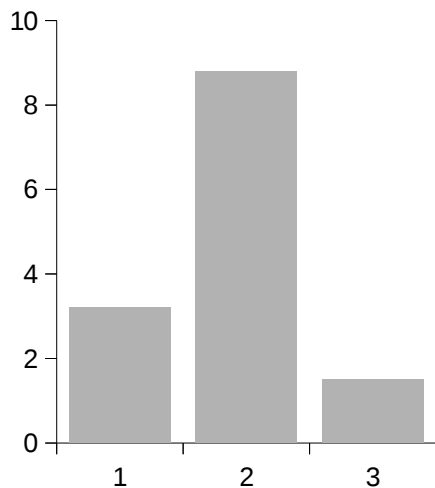
Channels control the graphical appearance of marks: used to **encode** data, can be combined



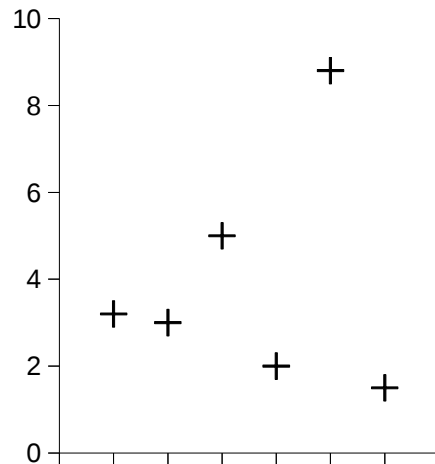
Figures are a combination of marks and channels



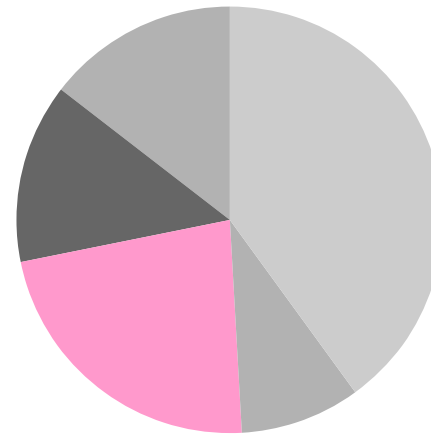
Figures are a combination of marks and channels



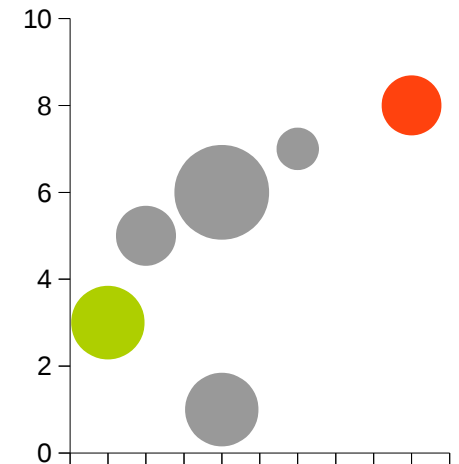
1 Mark =
Rectangle
1 Channel =
Length of
longest side



1 Mark =
Cross shape
2 Channels =
X position
Y position



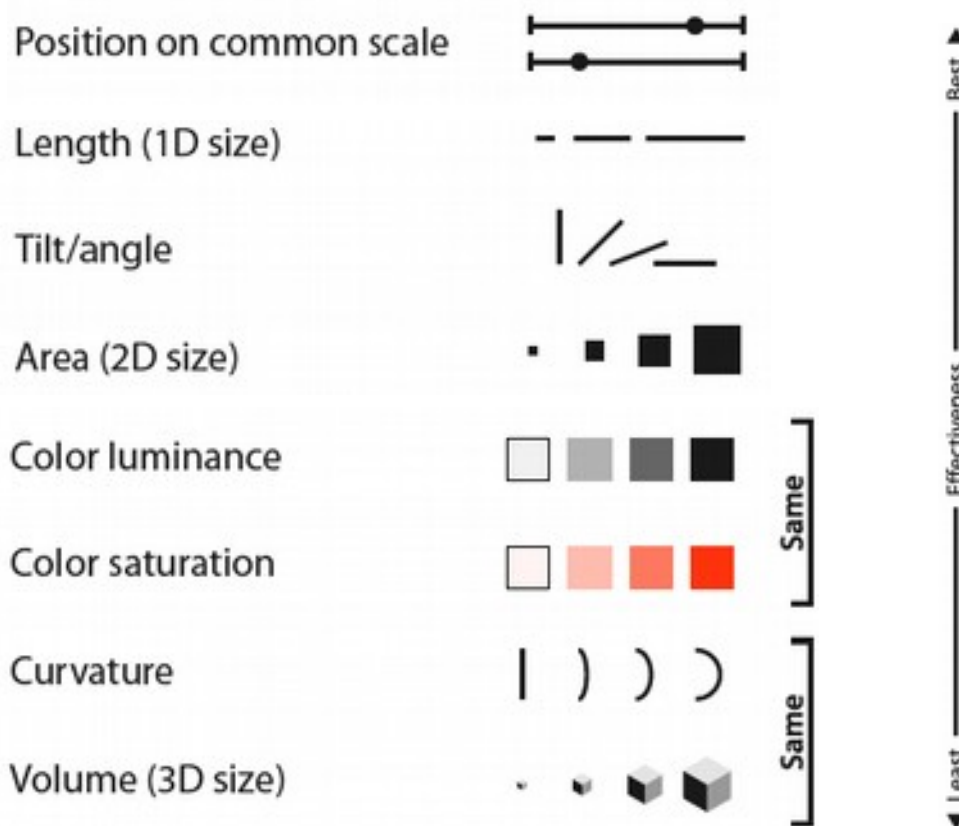
1 Mark =
Circle segment
2 Channel =
Angle
Colour



1 Mark =
Circle
4 Channels:
X position
Y position
Area
Colour

Types of channel

Identity channels: categorical/
qualitative attributes

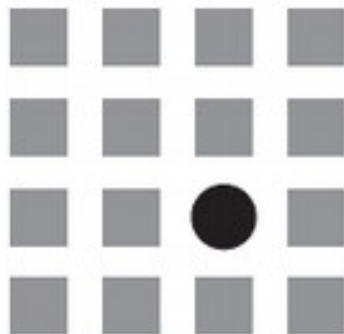


Magnitude channels: ordered/
quantitative attributes



Types of channel (continued)

SHAPE



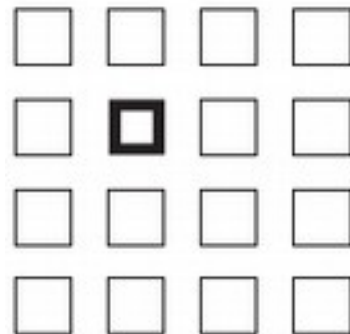
SIZE



ORIENTATION



WEIGHT



POSITION



COLOR



More principles

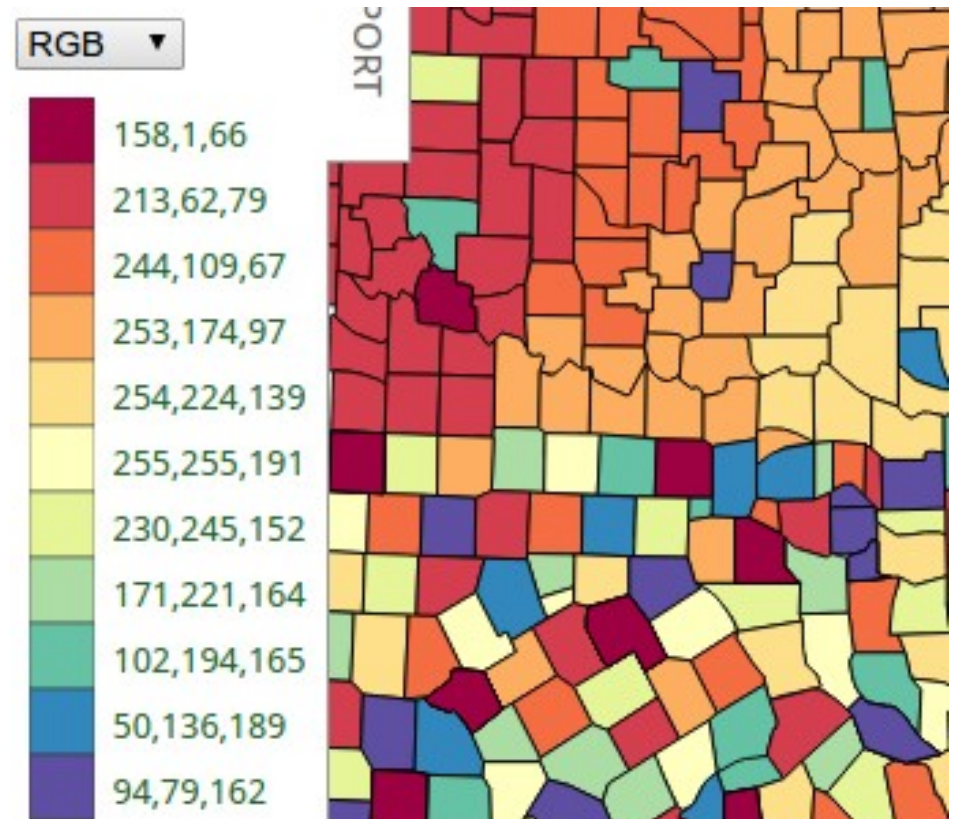
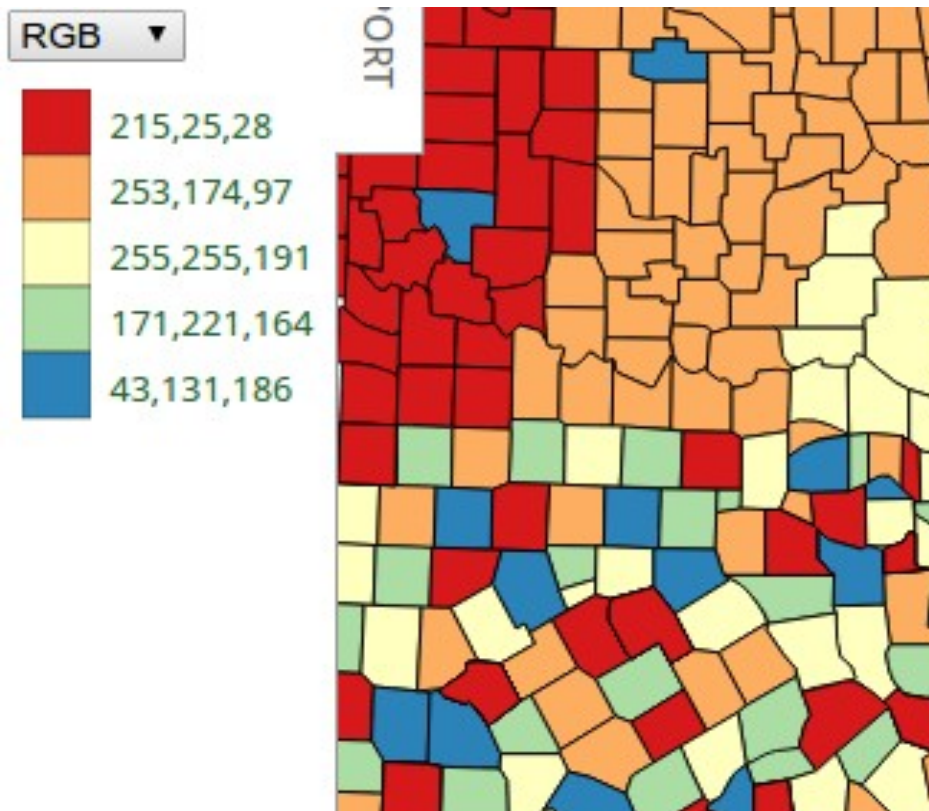
- **Effectiveness:** encode the most important information with the most effective channel
- **Expressiveness:** match the properties of the data and channel
 - i.e. heed whether the data are **quantitative**, **ordered** or **categorical**, and choose accordingly

More principles

Discriminability and separability

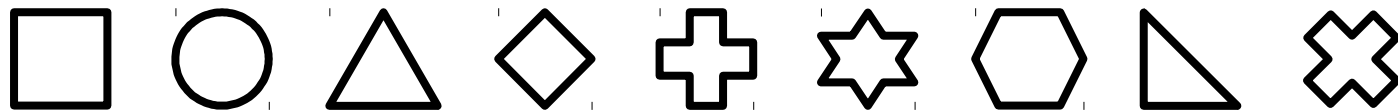
How many different types can you distinguish?

How easy is it to distinguish them?



Qualitative discrimination

Fillable **shapes**: can be combined with colour, but the fillable area needs to be similar,

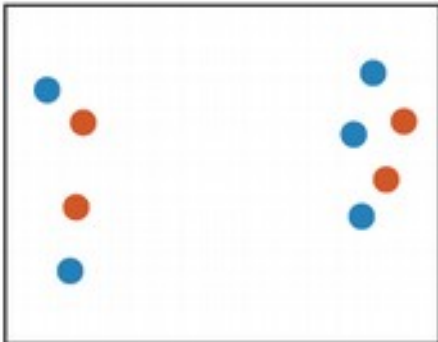


and they have to be distinguishable at small sizes



Separability

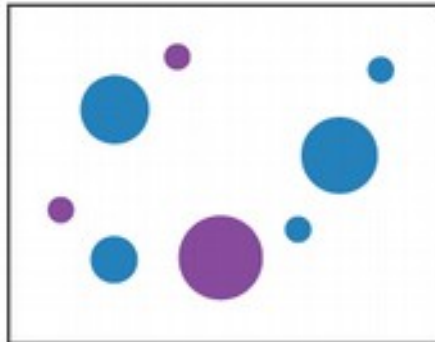
Position
+ Hue (Color)



Fully separable

2 groups each

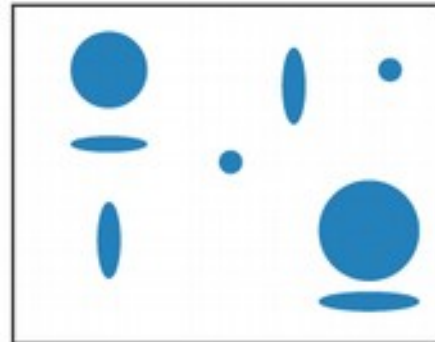
Size
+ Hue (Color)



Some interference

2 groups each

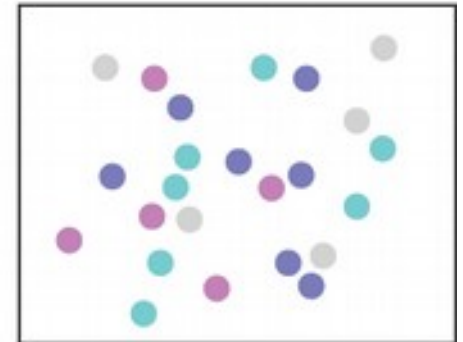
Width
+ Height



Some/significant
interference

3 groups total:
integral area

Red
+ Green



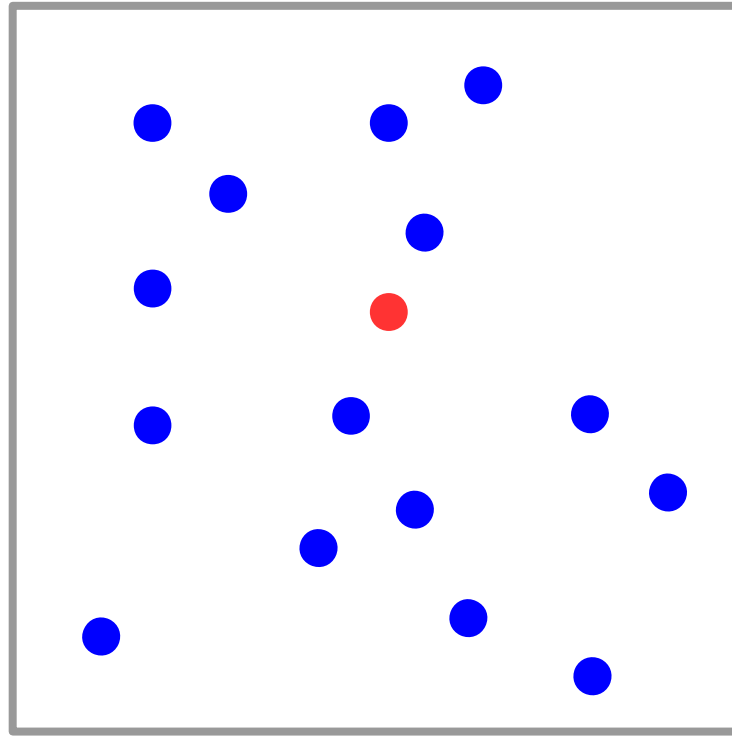
Major interference

4 groups total:
integral hue

Separability

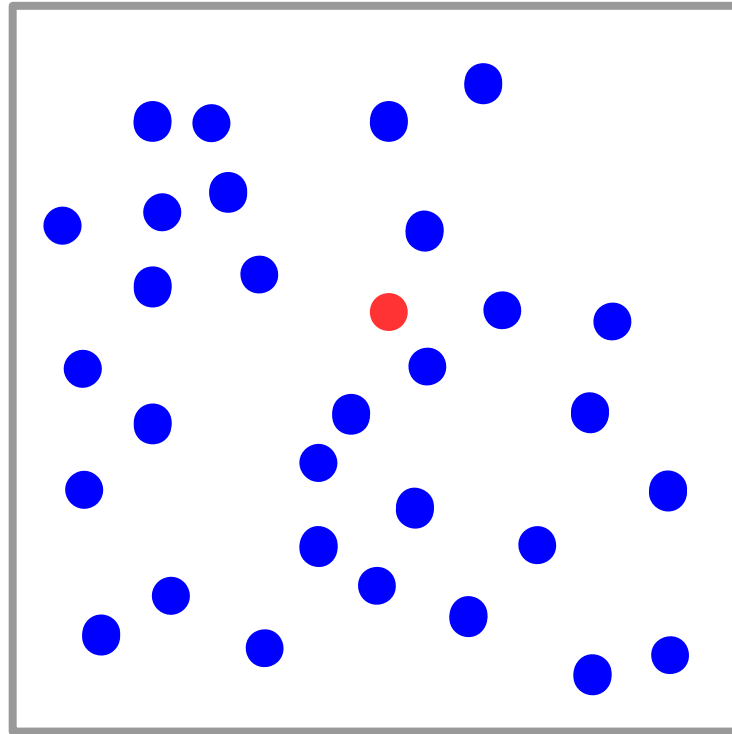
- The effectiveness of a channel does not always survive being combined with a second channel
- There are large variations in how much two different channels interfere with each other
- Trying to put too much information on a figure can erode the impact of the main point you're trying to make

Popout



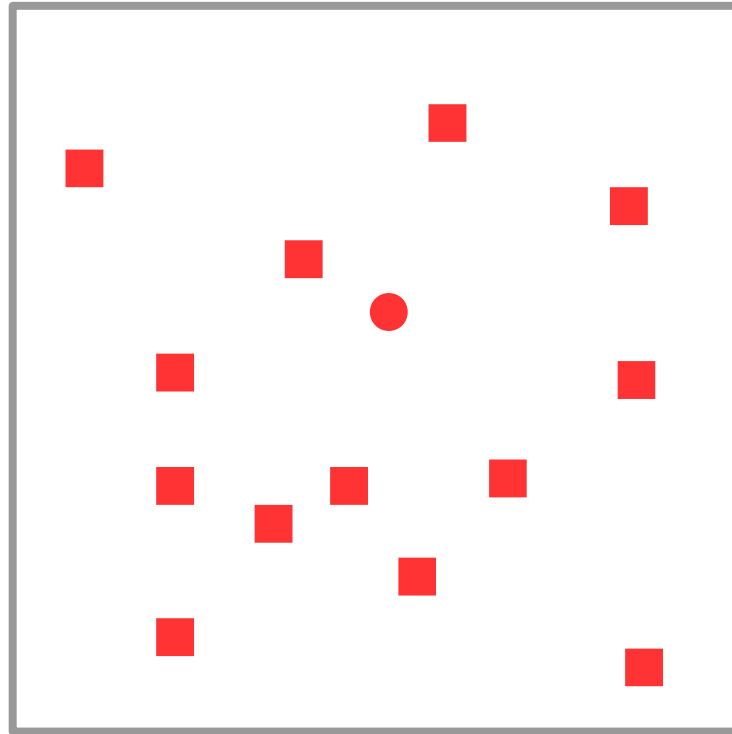
Find the red dot: how long does it take?

Popout



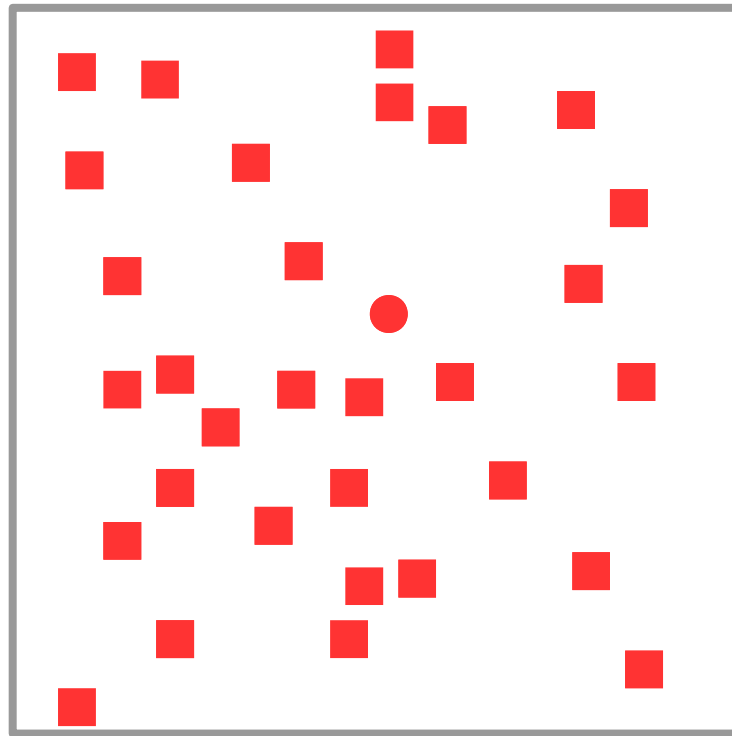
**The speed of identification is independent
of the count of distractors**

Popout



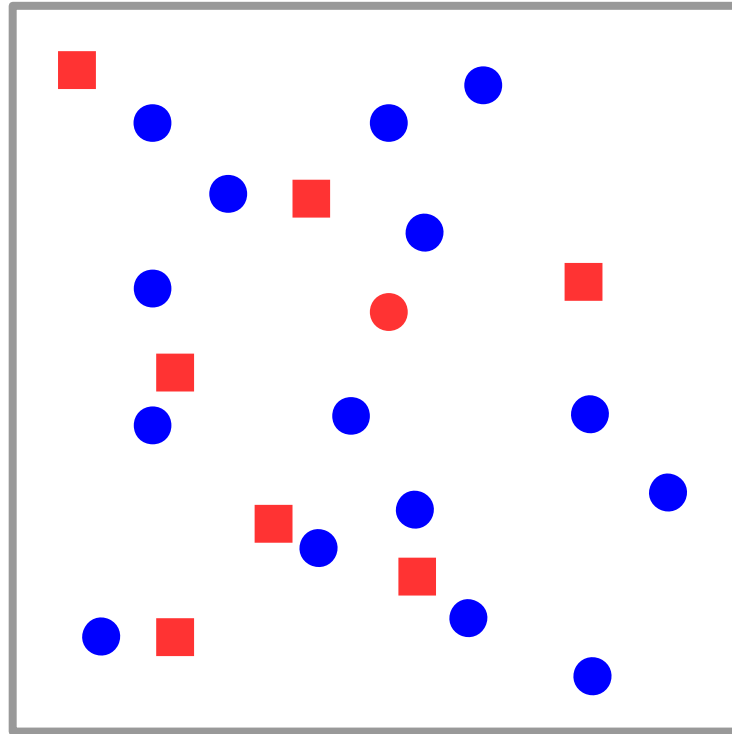
Find the circle

Popout



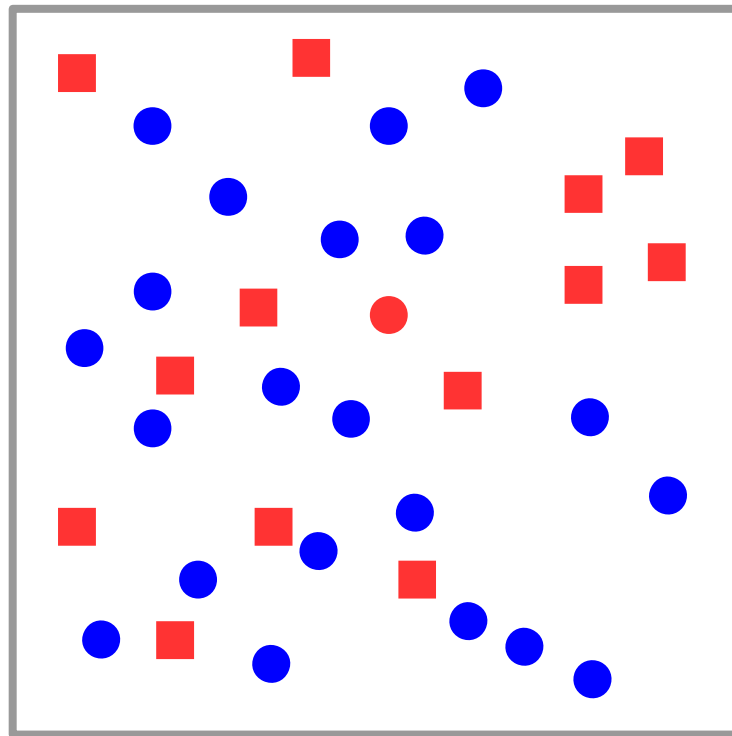
Colour stands out more than shape

Popout



Find the red dot

Popout



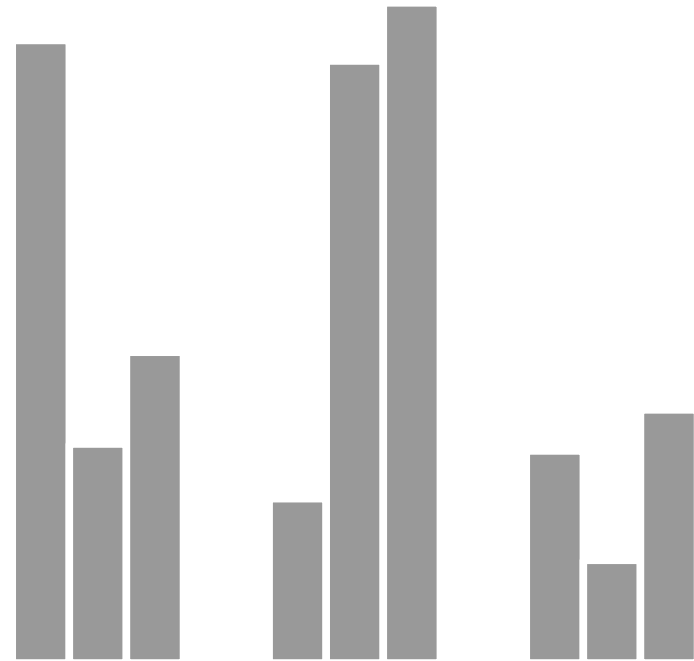
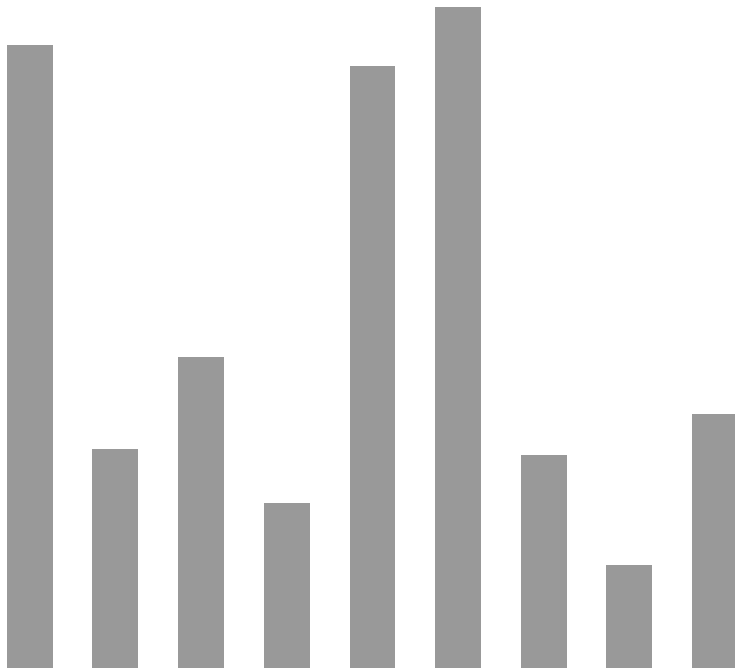
Mixing channels removes the effect

Dealing with complexity

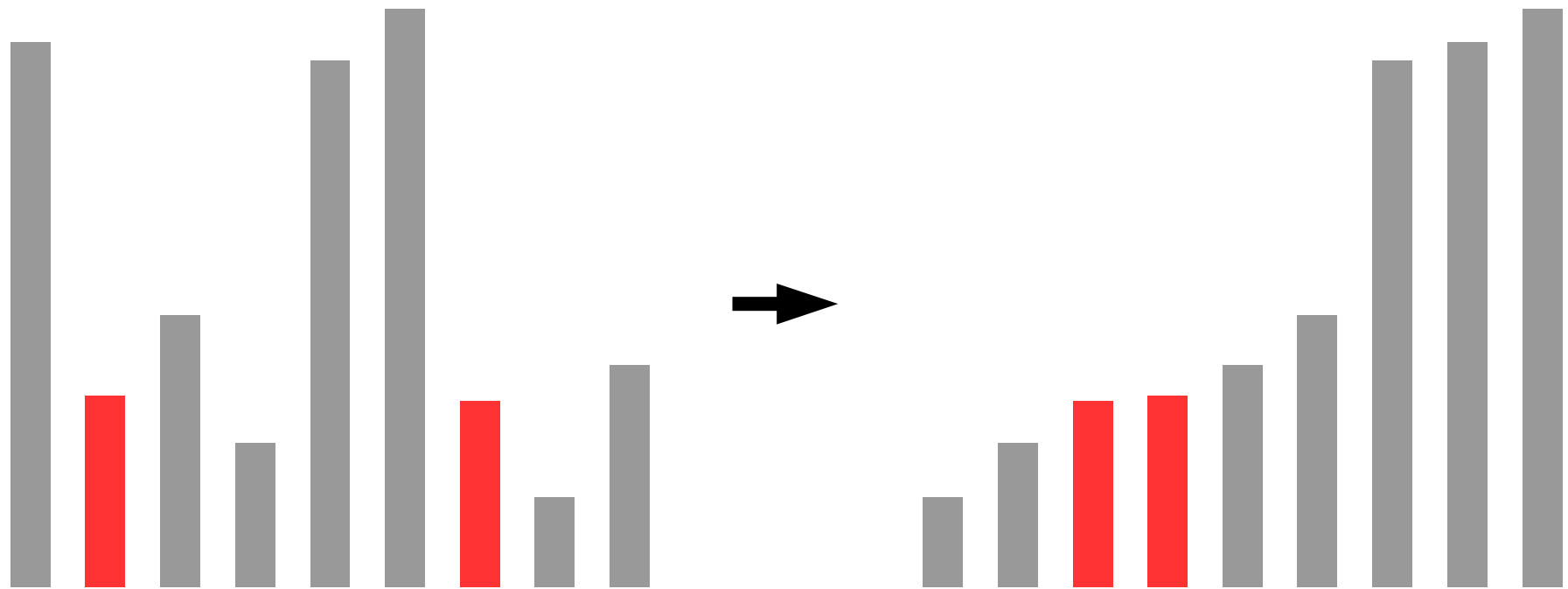
- In order to:
 - Focus the viewer's attention
 - Require less cognitive load for the viewer to understand the message



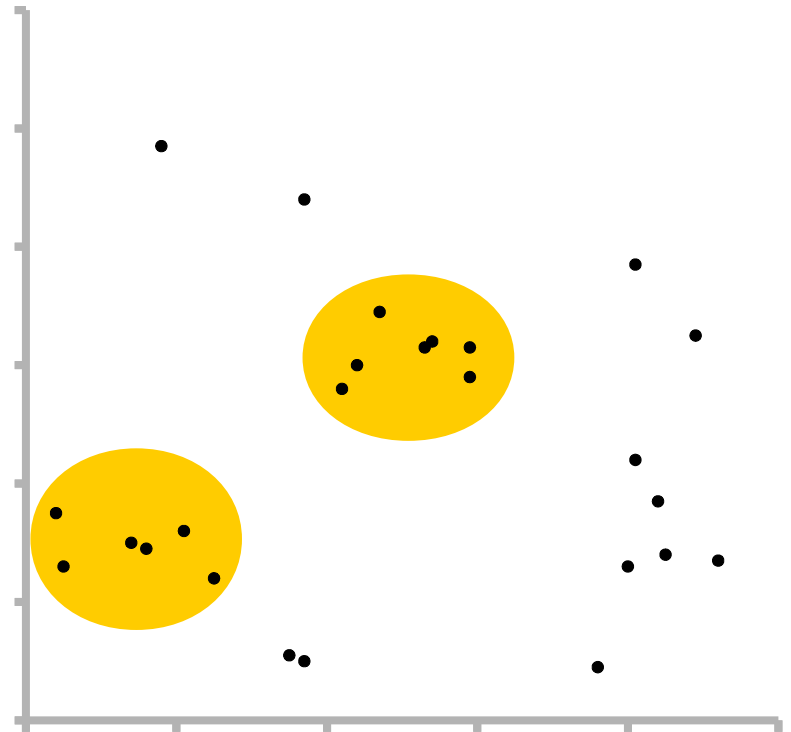
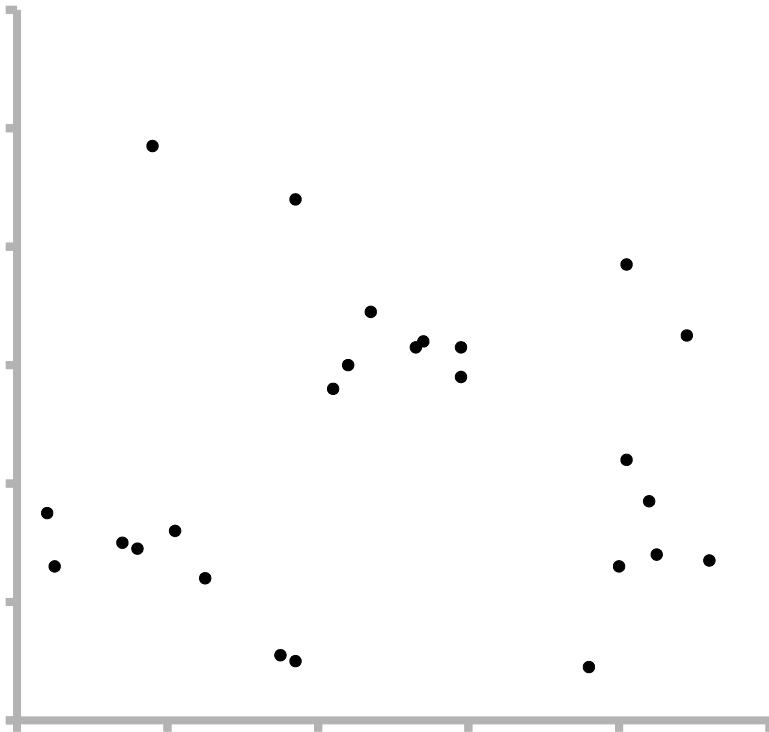
Grouping



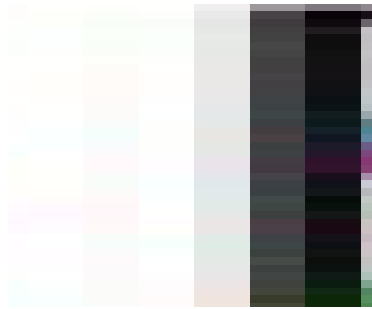
Ordering (only for categories)



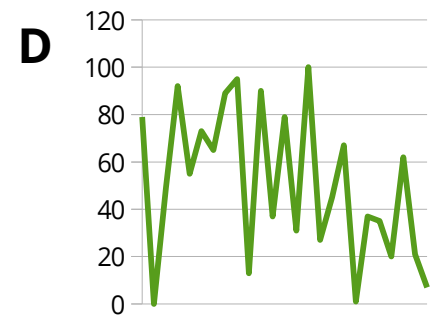
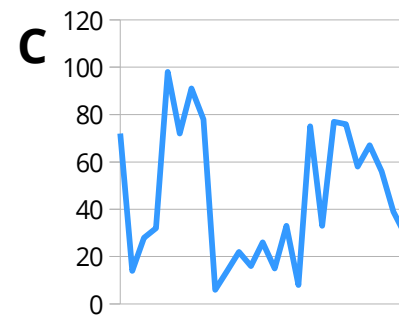
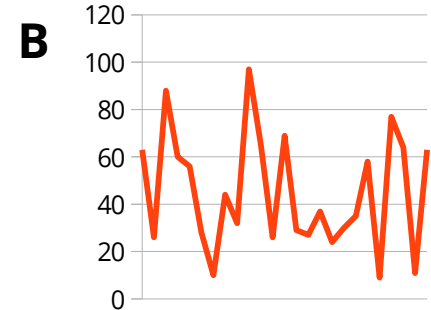
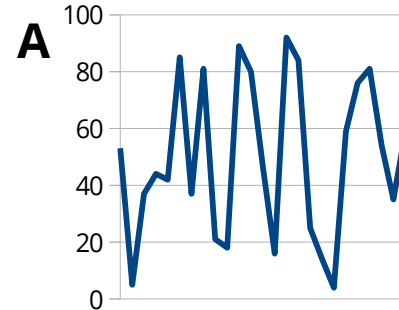
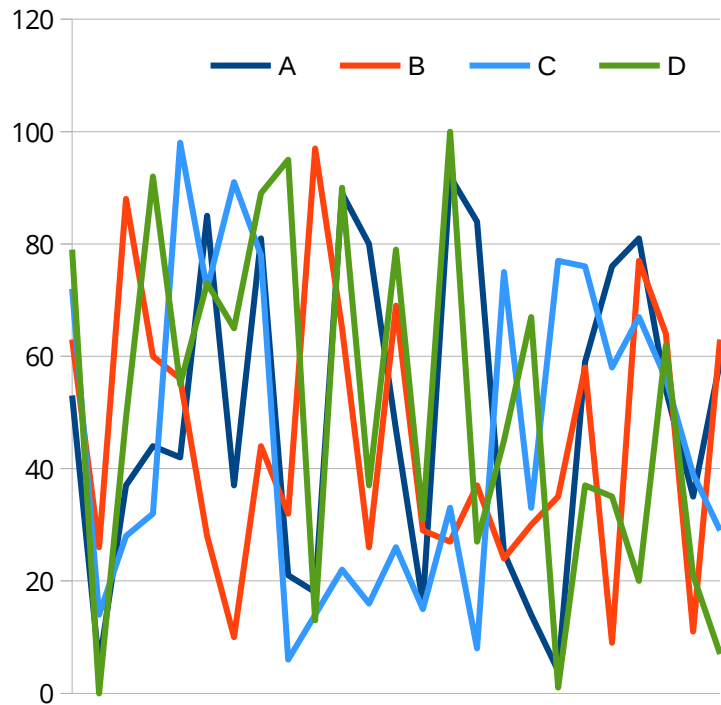
Containment



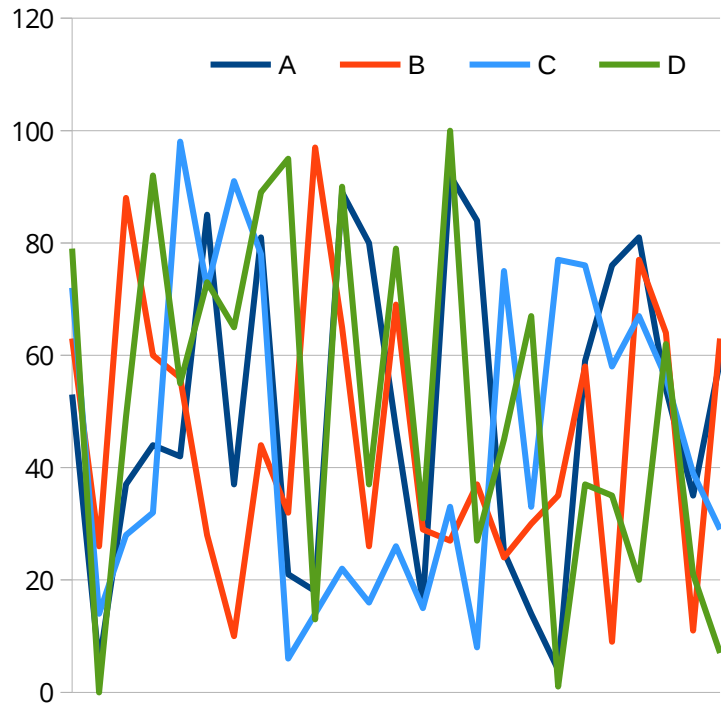
Filter, link, embed



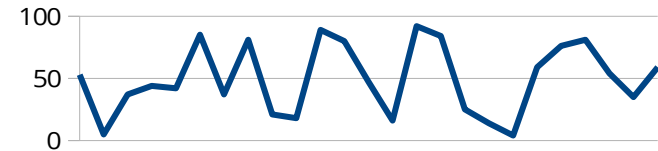
Small multiples



Small multiples



A



C



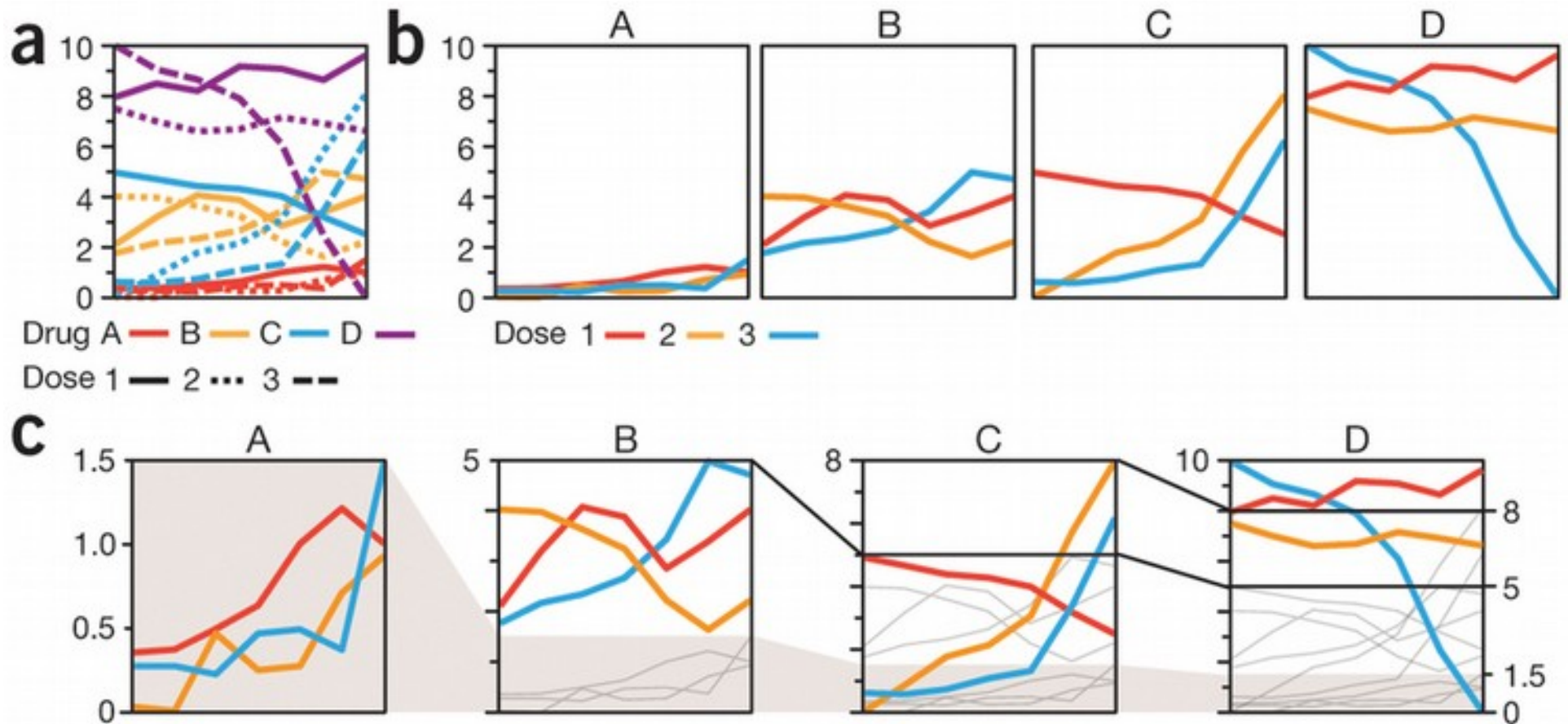
B



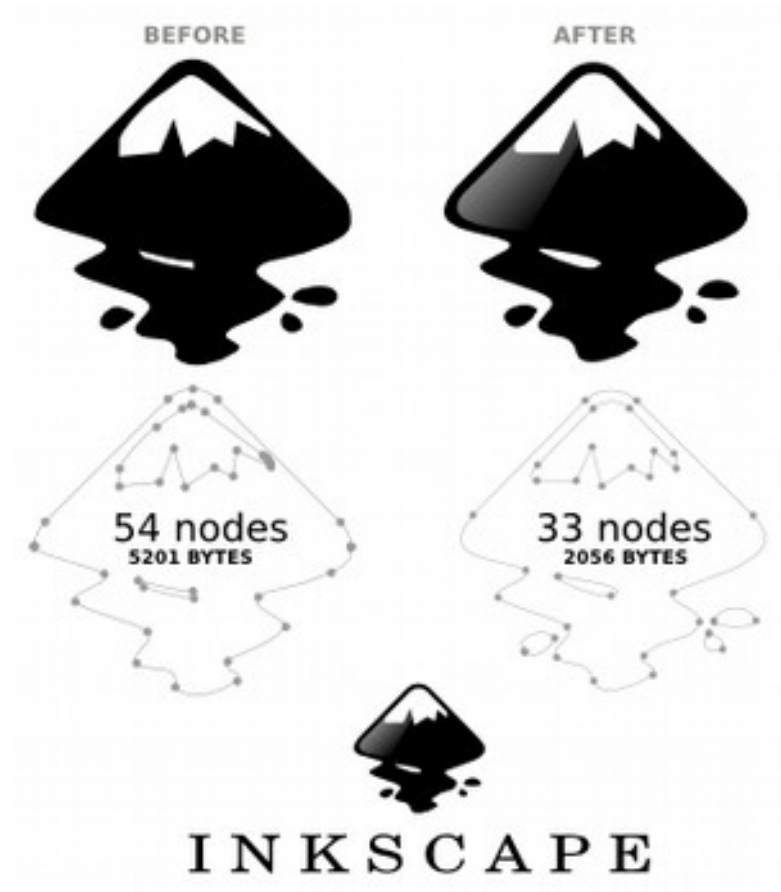
D



Small multiples



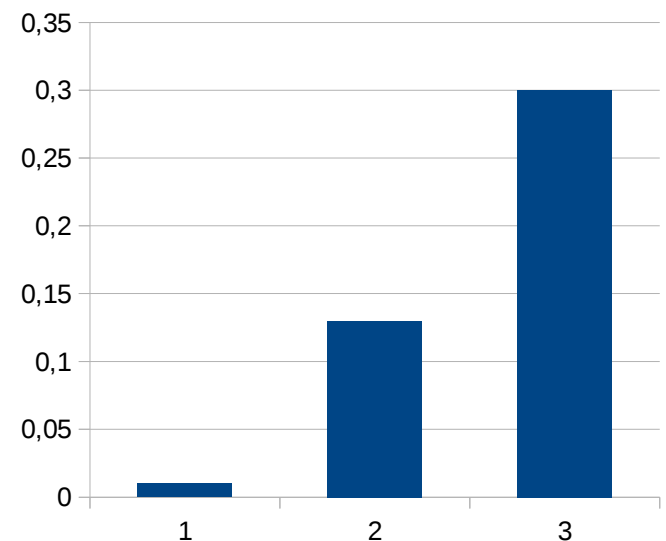
Practical



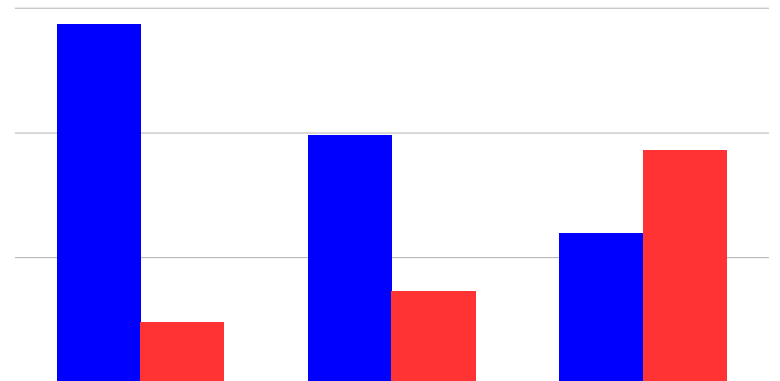
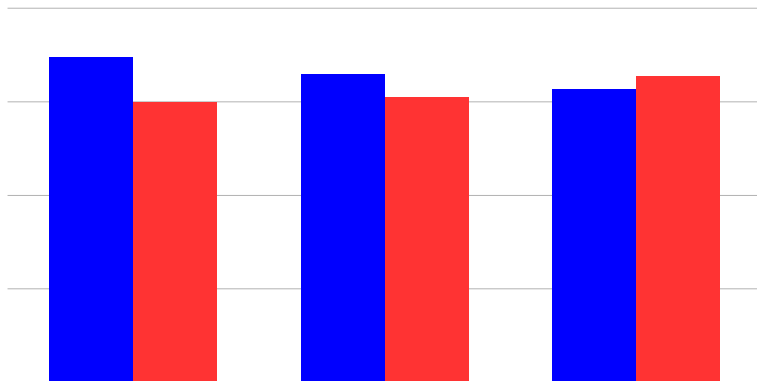
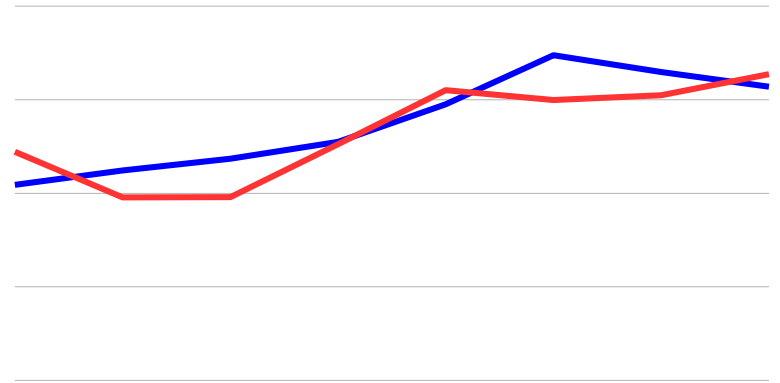
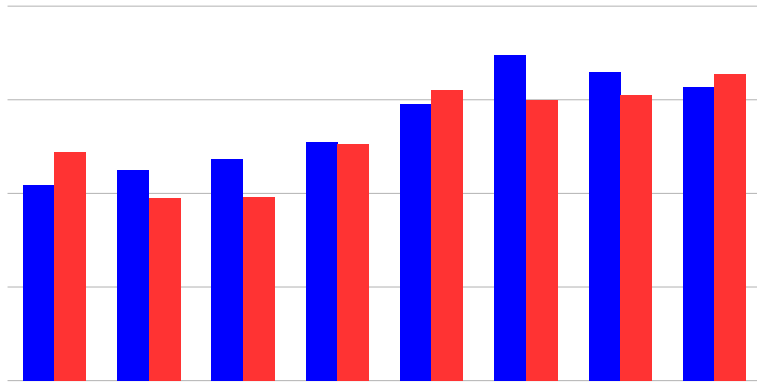
Choosing the type of figure

- Text, table or figure?
 - Text: one or two numbers
 - Table:
 - Exact numerical values
 - Small datasets (a figure may be best avoided if it has low data density)
 - When the data presentation requires many localised comparisons

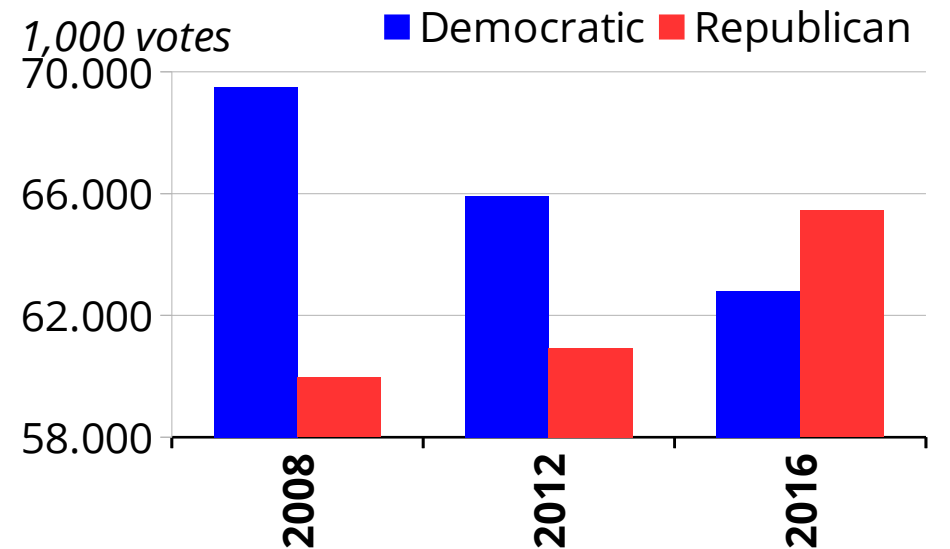
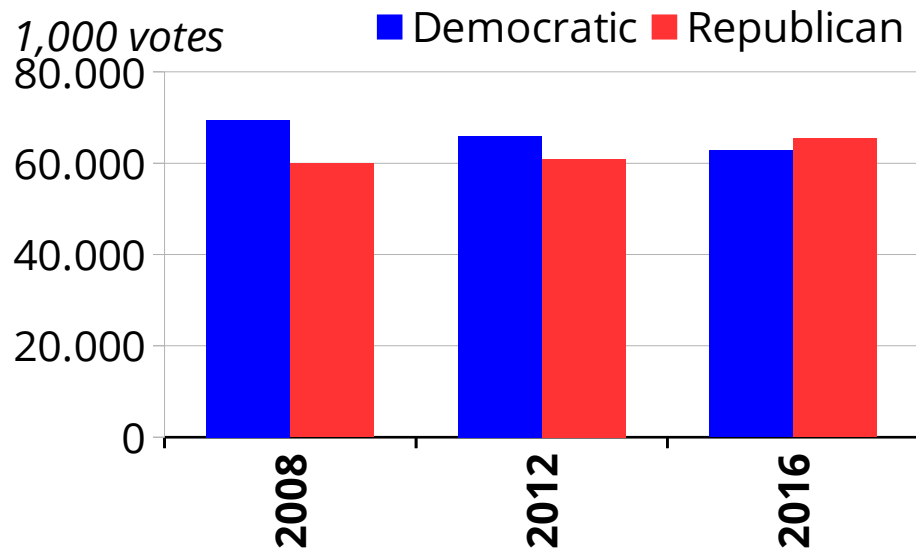
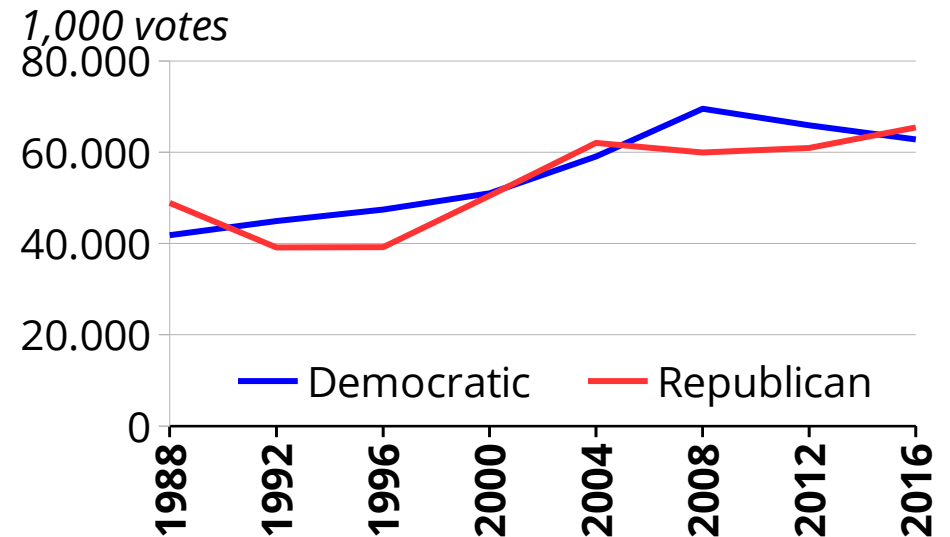
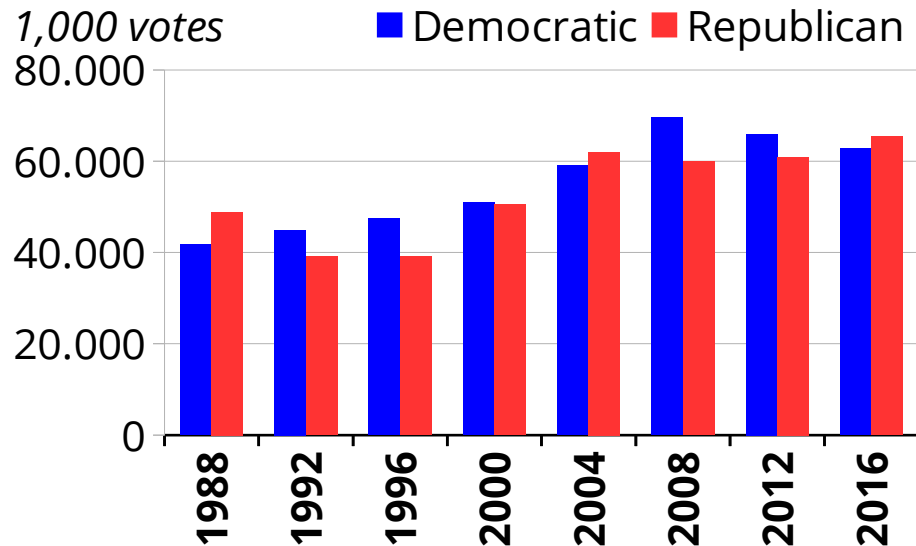
Treatment 1	0.01
Treatment 2	0.13
Treatment 3	0.30



Each figure tells a different story

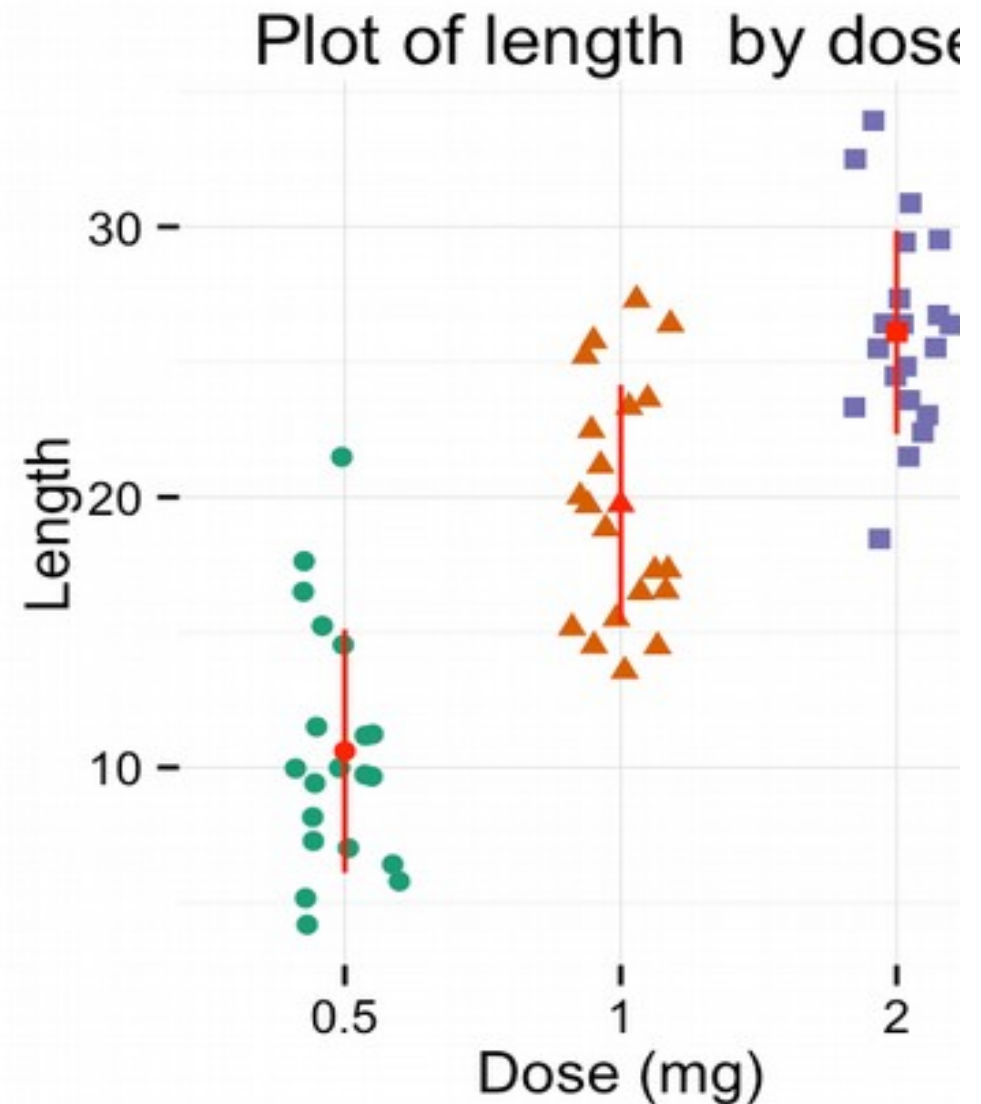


Each figure tells a story differently



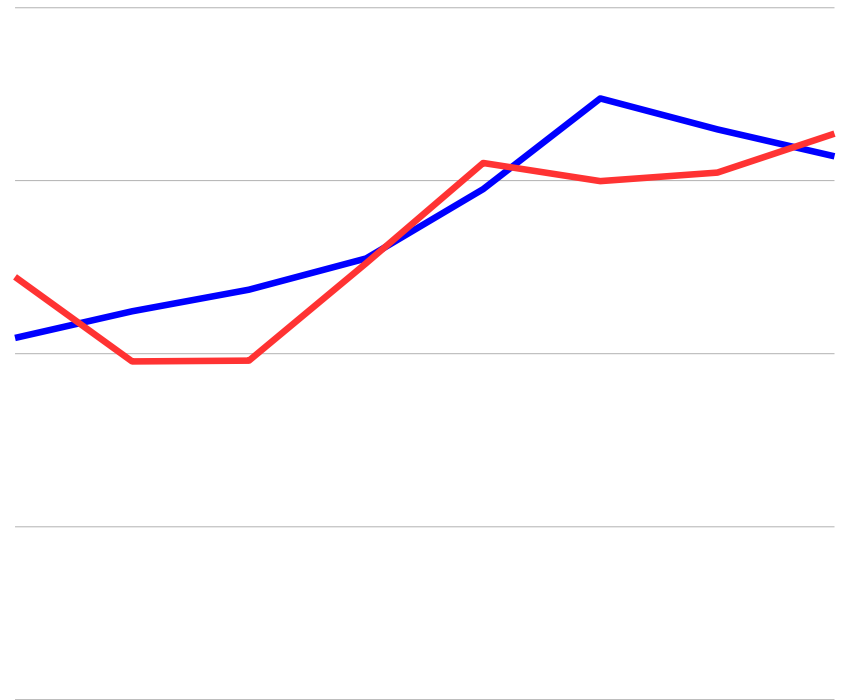
Stripchart – comparison

- Only one of the axis is meaningful
- To explore small datasets
- The most basic plot (rarely in publications)



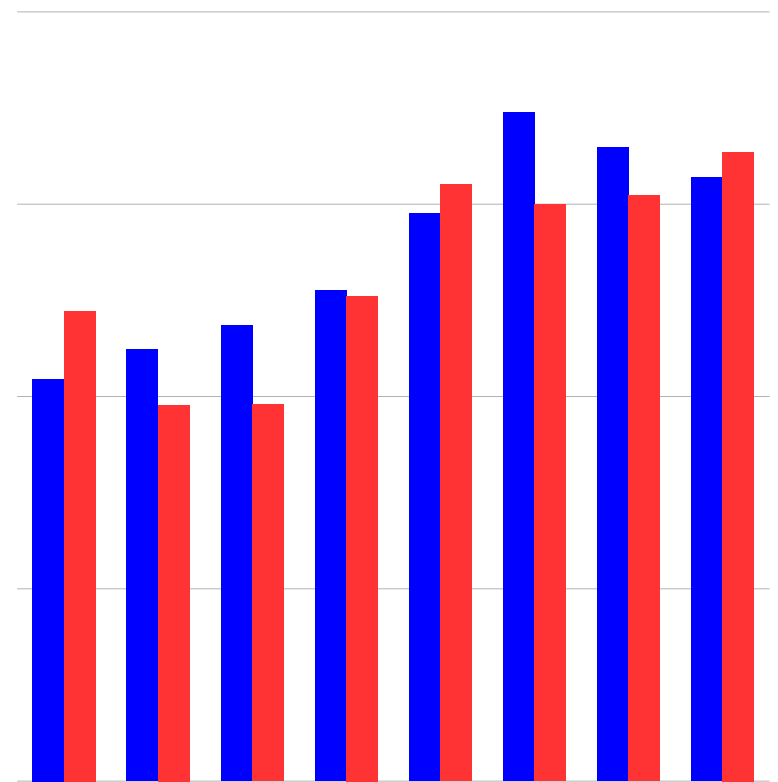
Line chart – relationships

- To show a trend of **continuous** data (usually over time)
- *Story*: how data change, rather than the discrete values of the data



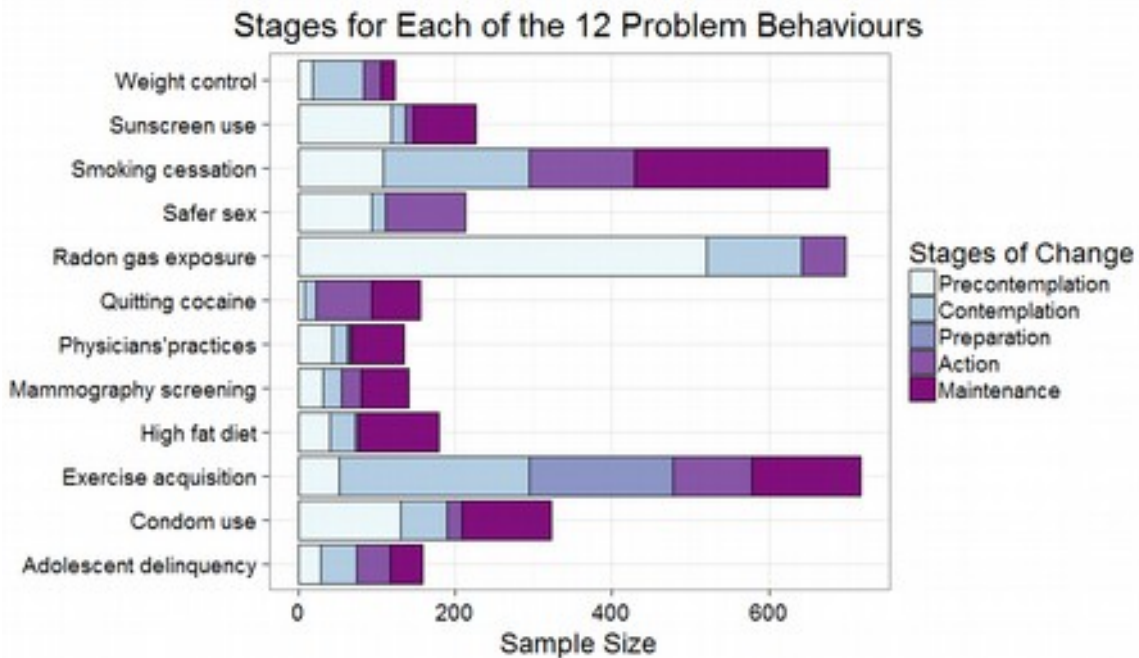
Bar chart – comparison

- To **compare** discrete quantities of **non-continuous** data

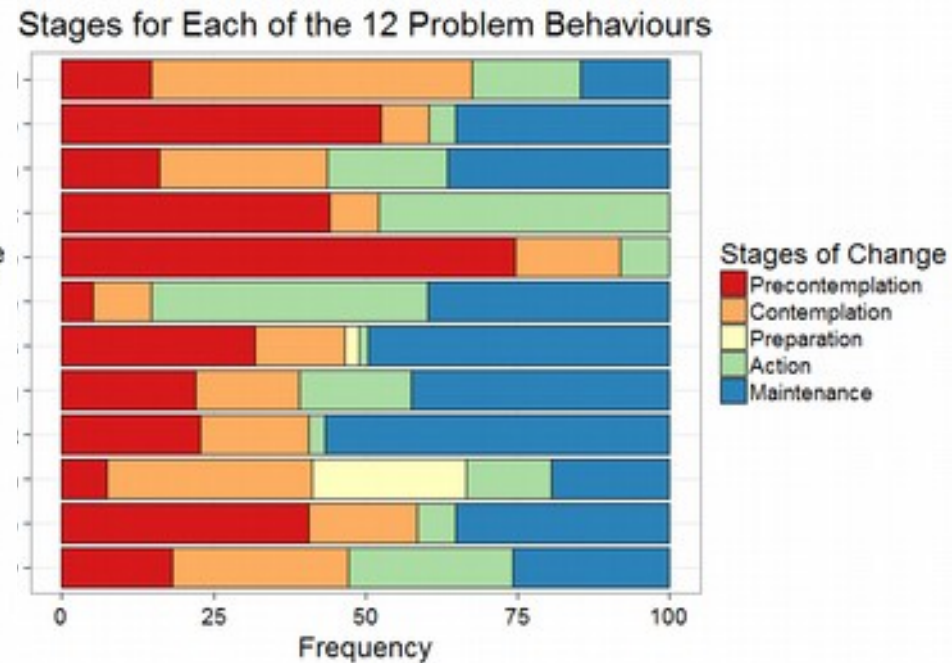


Bar chart variations

Stacked bar chart

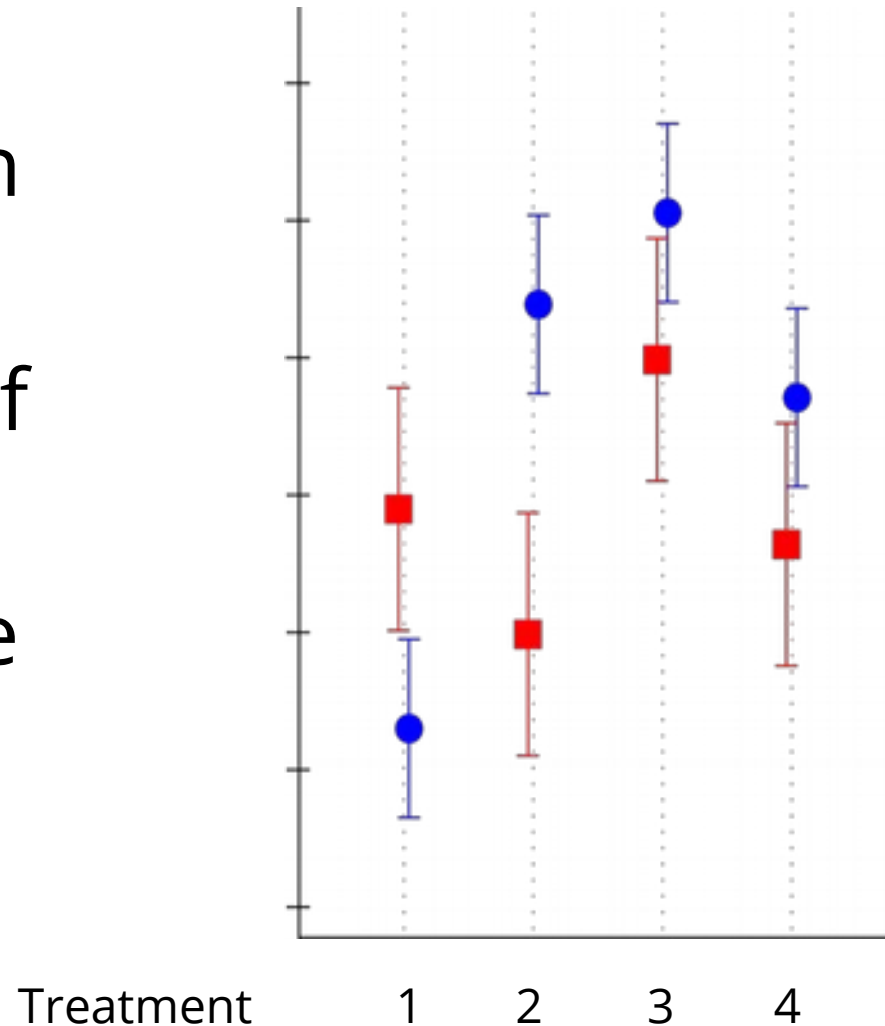


Normalised stacked bar chart



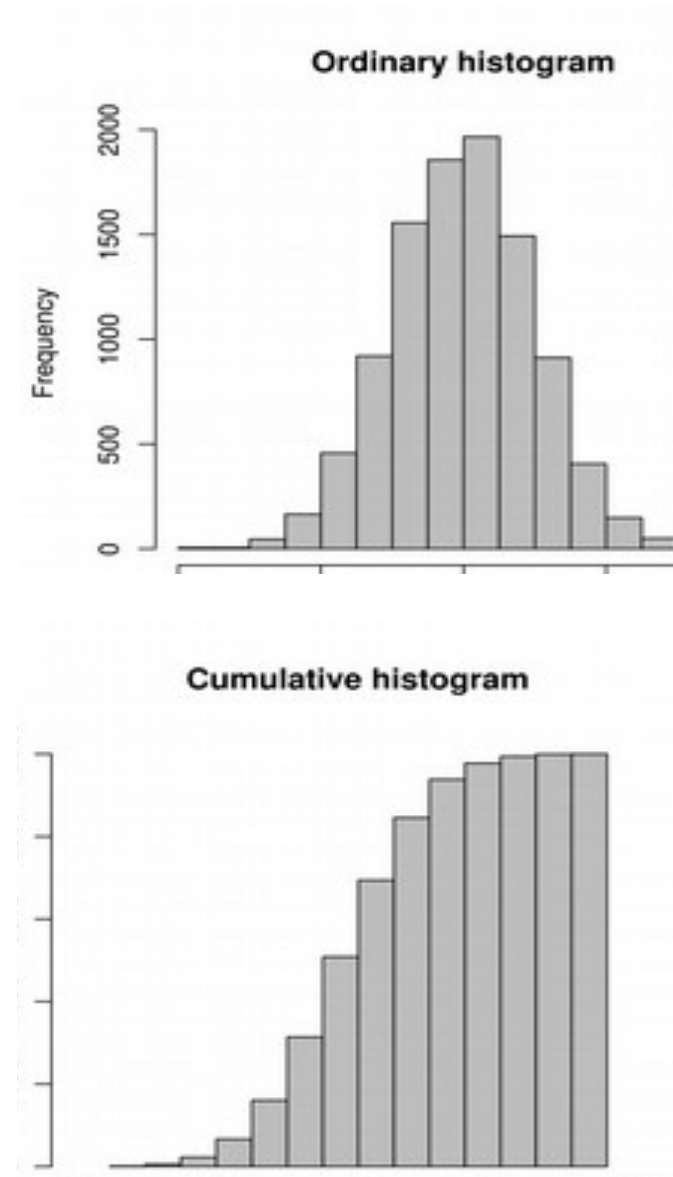
Bar chart alternative for comparisons: **Dotchart with confidence intervals**

- Focuses attention on the relative values and their measure of variability, rather than on the absolute values (height of the bar)



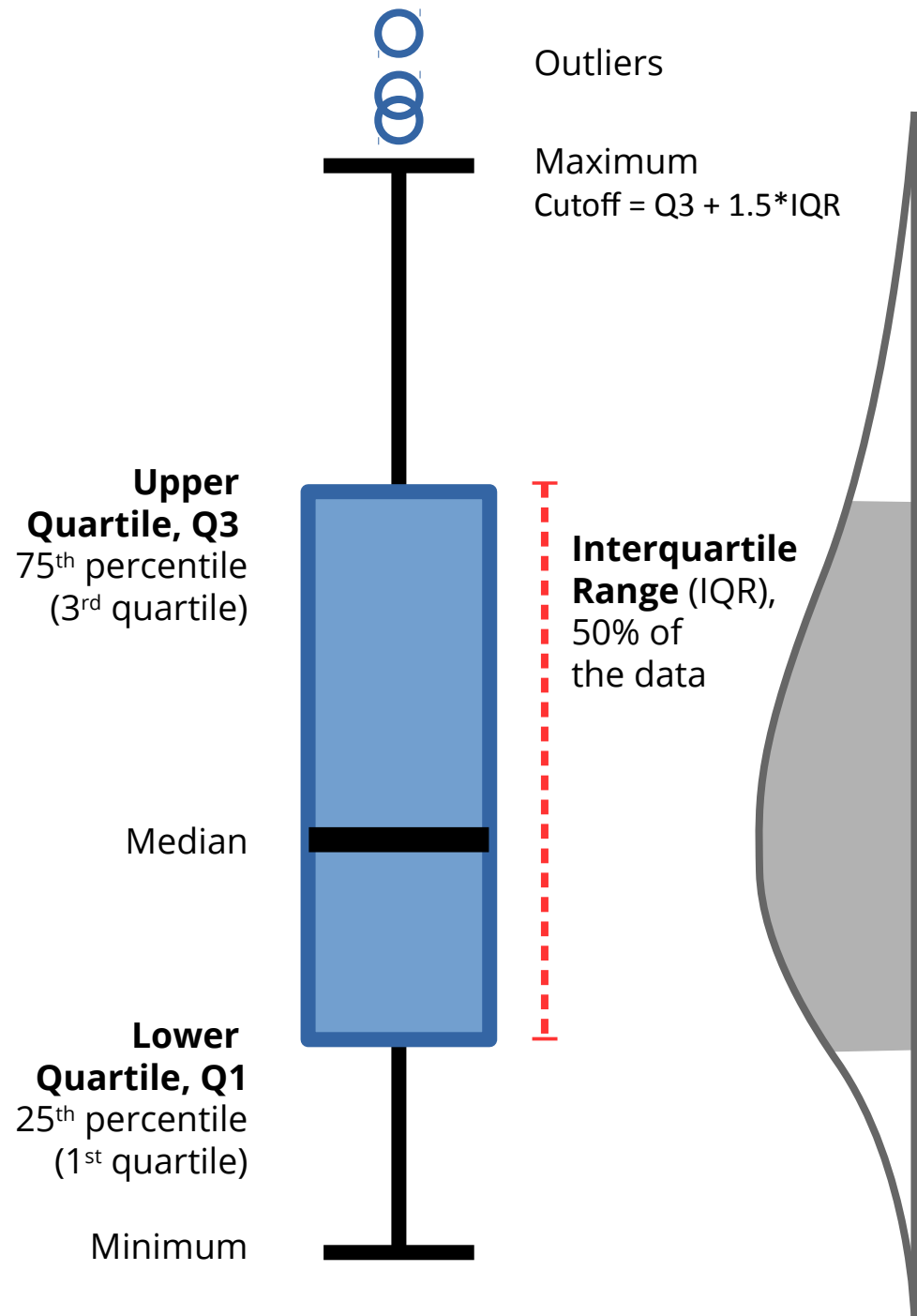
Histogram – distribution

- To show the distribution of a variable and the relative frequency of values
- Estimate of the probability distribution of the variable
- The number of bins (resolution) affects the perceived shape of the distribution
- Rules: Number of intervals $\approx \sqrt{N}$ and Interval width $\approx \text{Range} \div \sqrt{N}$



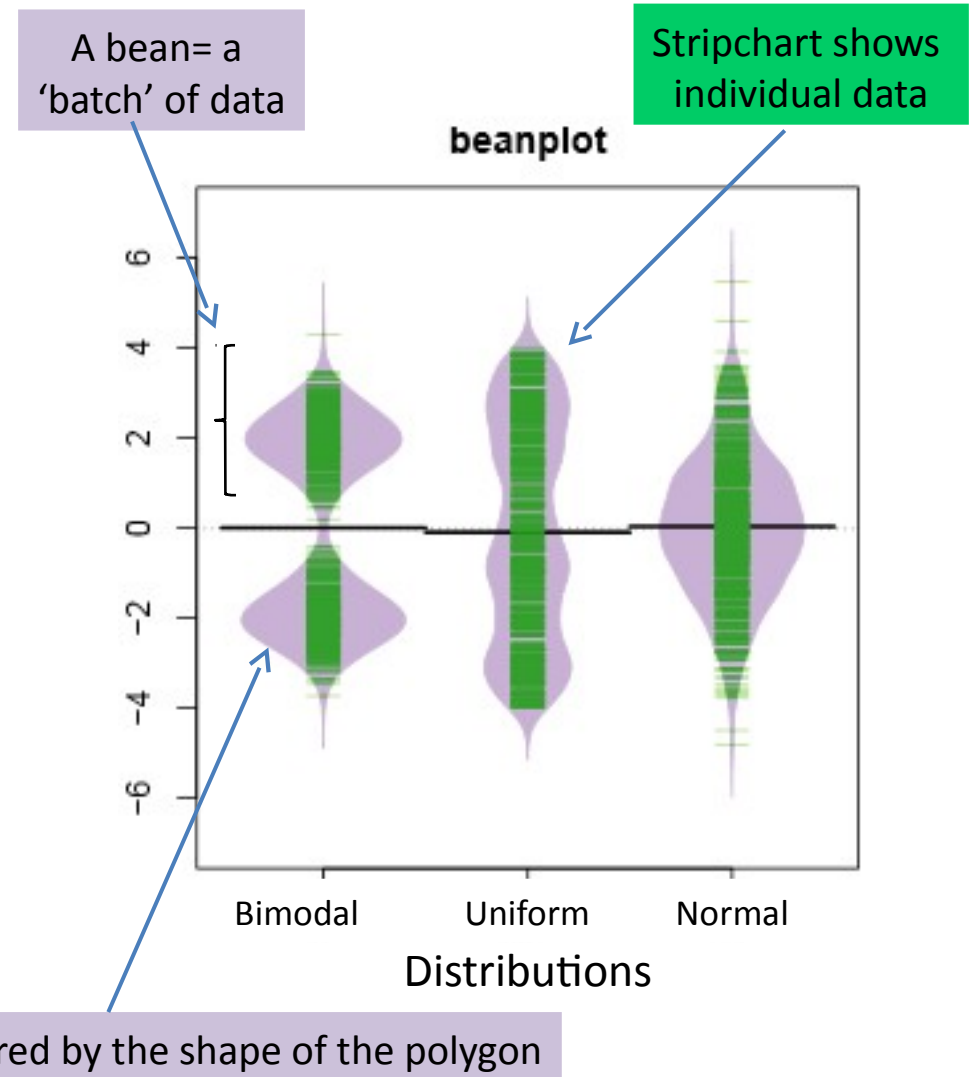
Boxplot – distribution

- Also ***box-and-whisker*** plot
- Shows the central value, the extremes, and the area where 50% of the values are located.
 - Usually median, minimum, maximum, lowest and highest quartiles
- Particularly useful to understand distribution of not-normal data



Boxplot variation: **Violin/ Bean plots**

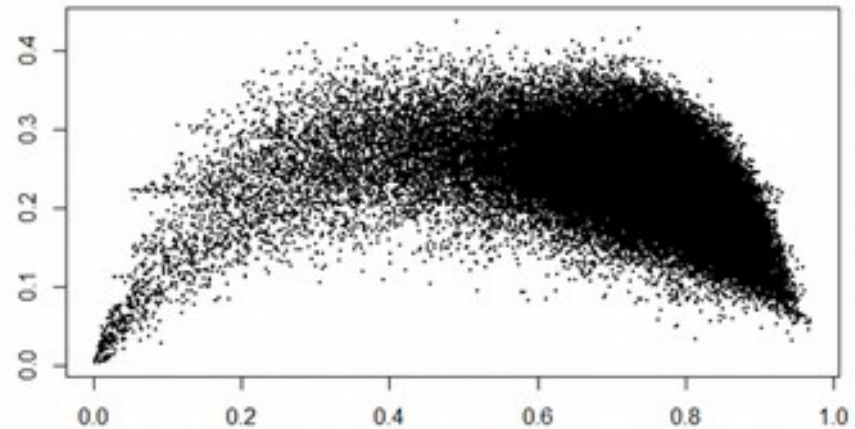
- To the above, it adds a **stripchart** of the actual datapoints and shows the data **density**, to understand the distribution in more detail



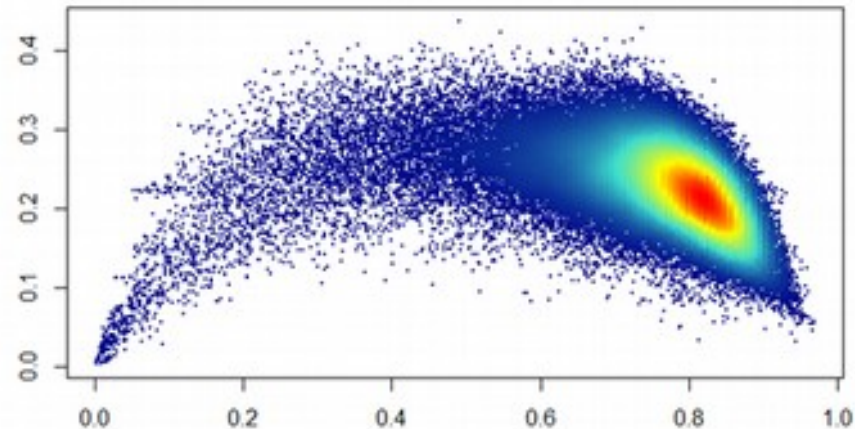
Scatterplot – relationships

- To show the relationship between two continuous variables
- For high-density data: use colours or transparency
- Variation: bubble scatterplot. It adds a 3rd dimension (but only for small datasets)

- Problem: very big dataset

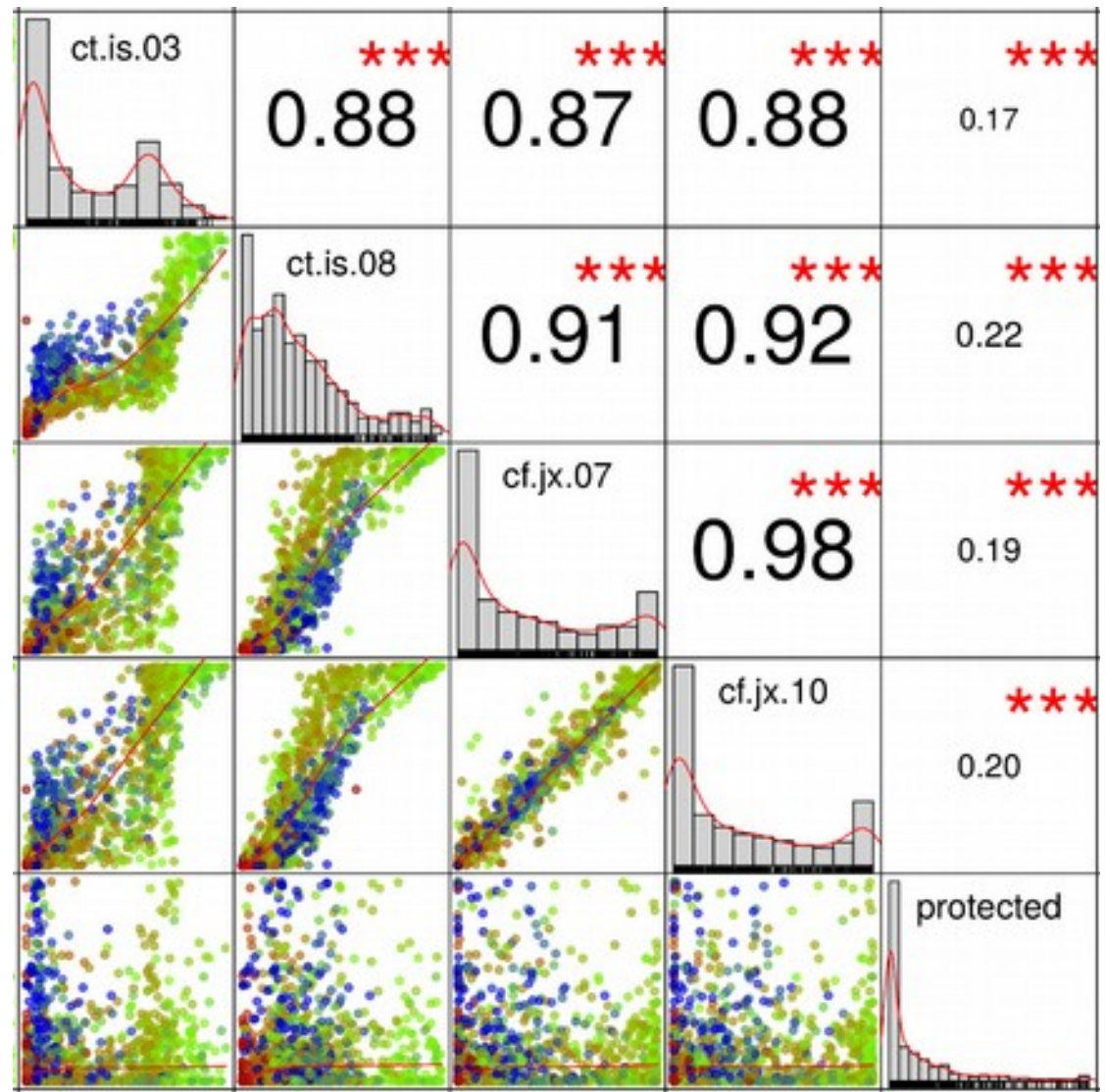


- Solution: smoothed **densities** colour representation



Scatterplot matrix (correlogram)

– relationships

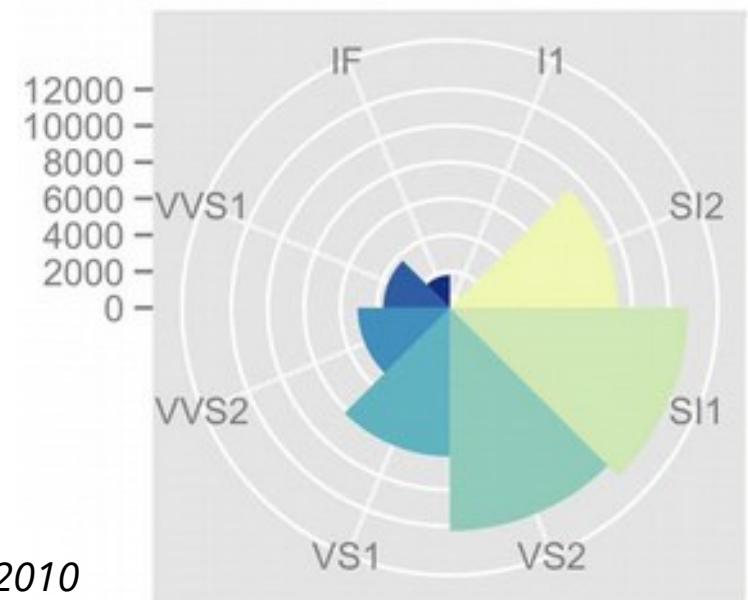
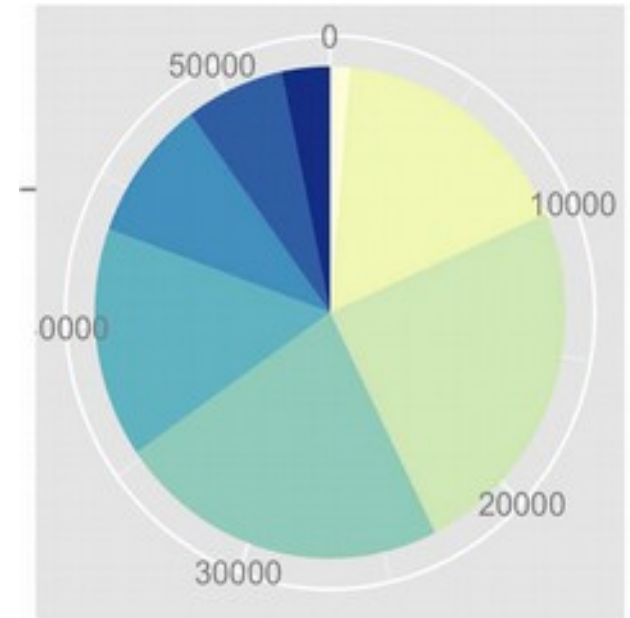


Pie chart – composition/ proportion

- To show relative proportions of a whole
- Not a great idea, 'given their low data-density and failure to order numbers along a visual dimension' (Tufte)

Alternative:

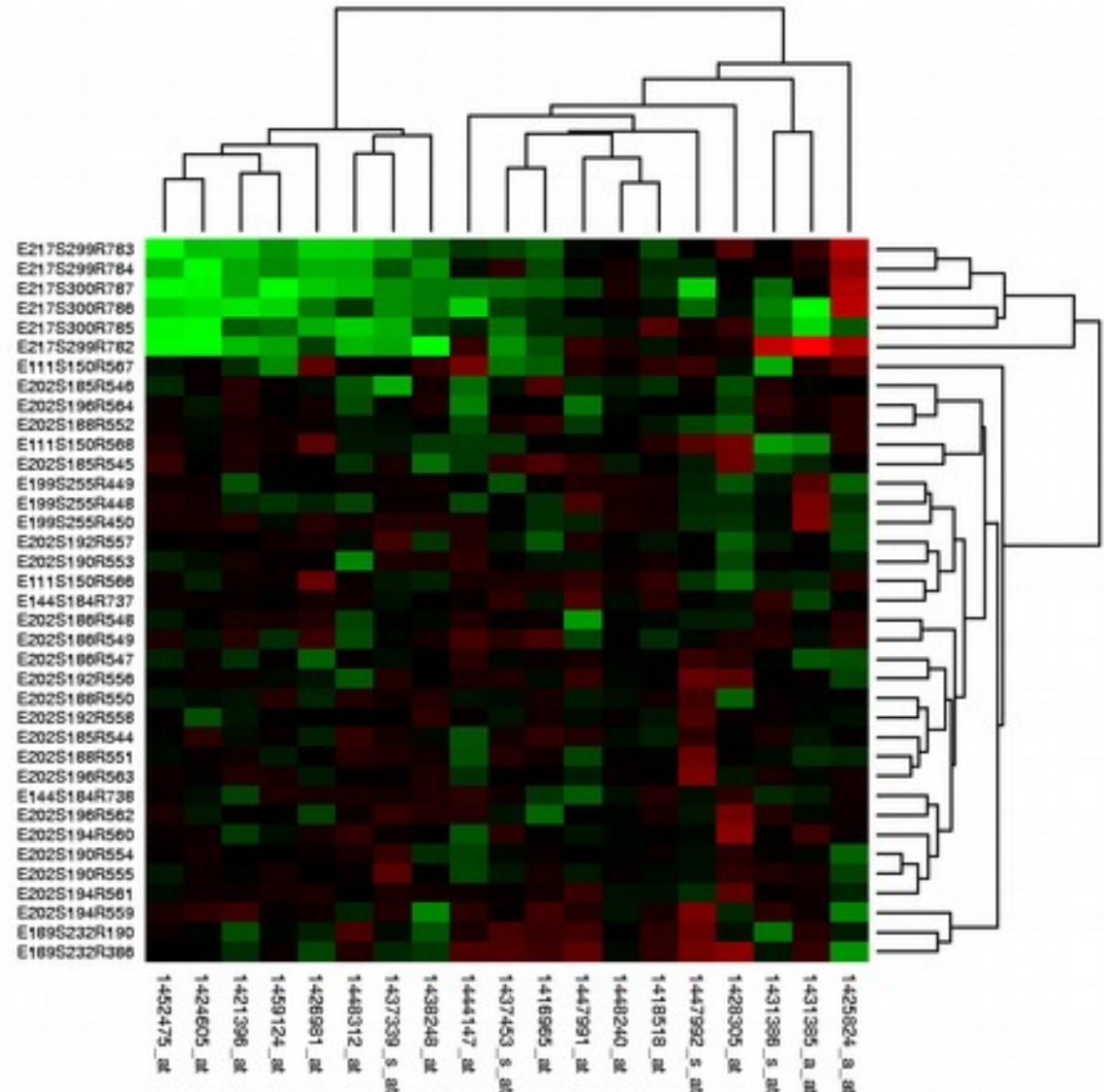
Polar area chart



Wickham, 2010

Heatmap – relationship

- Shows more complex relationships, e.g. many conditions

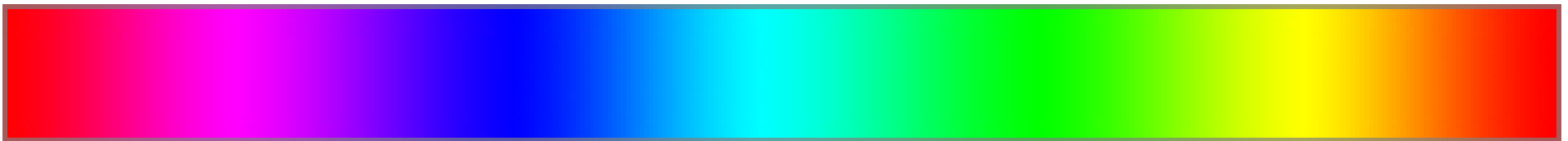


Summary

<i>Plot</i>	<i>Aim</i>	<i>Main R function</i>
Stripchart	distribution	stripchart()
Line chart	relationships	plot(type="l")
Bar chart (stacked, norm. stacked)	comparison (and composition)	barplot()
Dotchart with CI	comparison	dotchart()
Histogram	distribution	hist()
Boxplot (violin/ bean)	distribution	boxplot(), vioplot()
Scatterplot (correlogram)	relationships	plot(x, y), corrgram package
Pie chart	composition	pie()
Heatmap	relationship	heatmap()

Colour

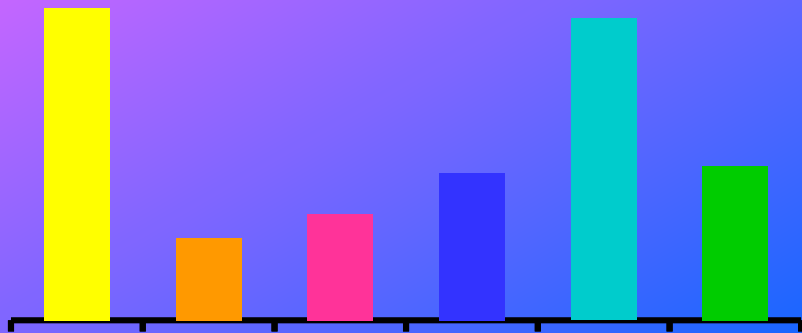
- Colour can be used to:
 - Highlight specific data
 - Group categories of data
 - Encode quantitative values
- Colours: primary, secondary, intermediate
- Our perception of hue is not linear



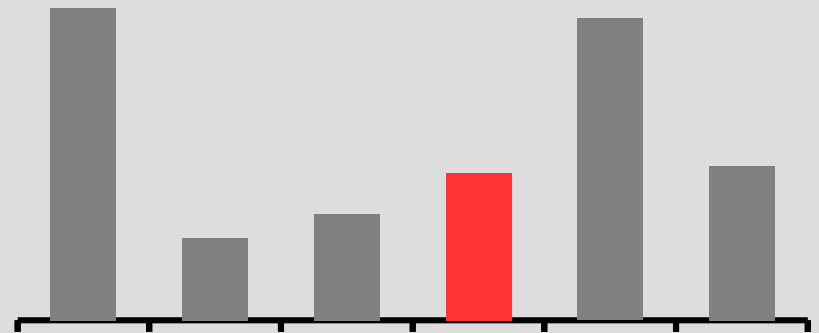
Don't let your
colours
distract from your
message...

Instead, use colour
to **communicate**

Don't let your colours
overwhelm your data

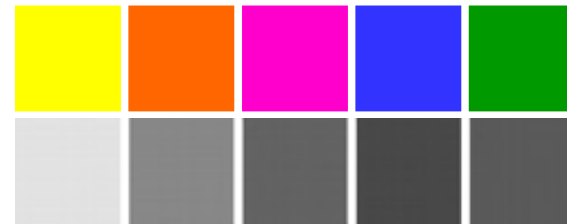


Instead, use colour to
emphasise your data



Characteristics of a colour

- **Hue**: the *actual* colour (qualitative)
- **Saturation**: the intensity of the hue (quantitative)
- **Value**: the lightness/ darkness of a colour (quantitative; useful to know how a colour will behave when transformed to grayscale)
- ***Lightness***
 - ***Shade***: the amount of black
 - ***Tint***: the amount of white



Convert to
grayscale



Colour: How computers identify colours



CMYK: percentage of Cyan + Magenta + Yellow + Black





RGB: intensity of Red + Green + Blue



HSL: Hue + Saturation + Lightness

Hexvalue: 0 to F values of Red, Green, Blue. 0: no intensity. F: maximum, *what colour is this?* #FF 00 00

	<code>#ff00ff</code>	<code>rgb(255, 0, 255)</code>	<code>hsl(300, 100%, 50%)</code>
	<code>#ff0000</code>	<code>rgb(255, 0, 0)</code>	<code>hsl(0, 100%, 50%)</code>

Colour in screen and in print usually differ slightly (especially greens). To match them, the screen has to be calibrated (a cumbersome process!).

Playing with colours: http://www.w3schools.com/colors/colors_picker.asp

Three ways to name colours in R

1. By **name**, see available colours using **colors()**, and the list with the actual colours:

<http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>

2. Using **hexadecimal**, e.g. **"#00FF00"**, or **"#00FF0055"** (the 7th and 8th digits, if any, correspond to opacity)

3. Converting **RGB** to hexadecimal, using the function **rgb**, e.g. **rgb(1, 1, 1)**

Colour tools in R

Colour ramps:

```
# rainbow, heat.colors, topo.colors, terrain.colors,  
cm.colors
```

```
plot(1:5, col=palette(terrain.colors(7)), pch=15, cex=3)
```

RcolorBrewer:

```
# install.packages("RColorBrewer")
```


```
library(RColorBrewer)
```

```
# check out the available palettes
```

```
display.brewer.all(n=NULL, type="all",  
                  select=NULL, exact.n=TRUE)
```

```
brewer.pal(5, "Set1")
```

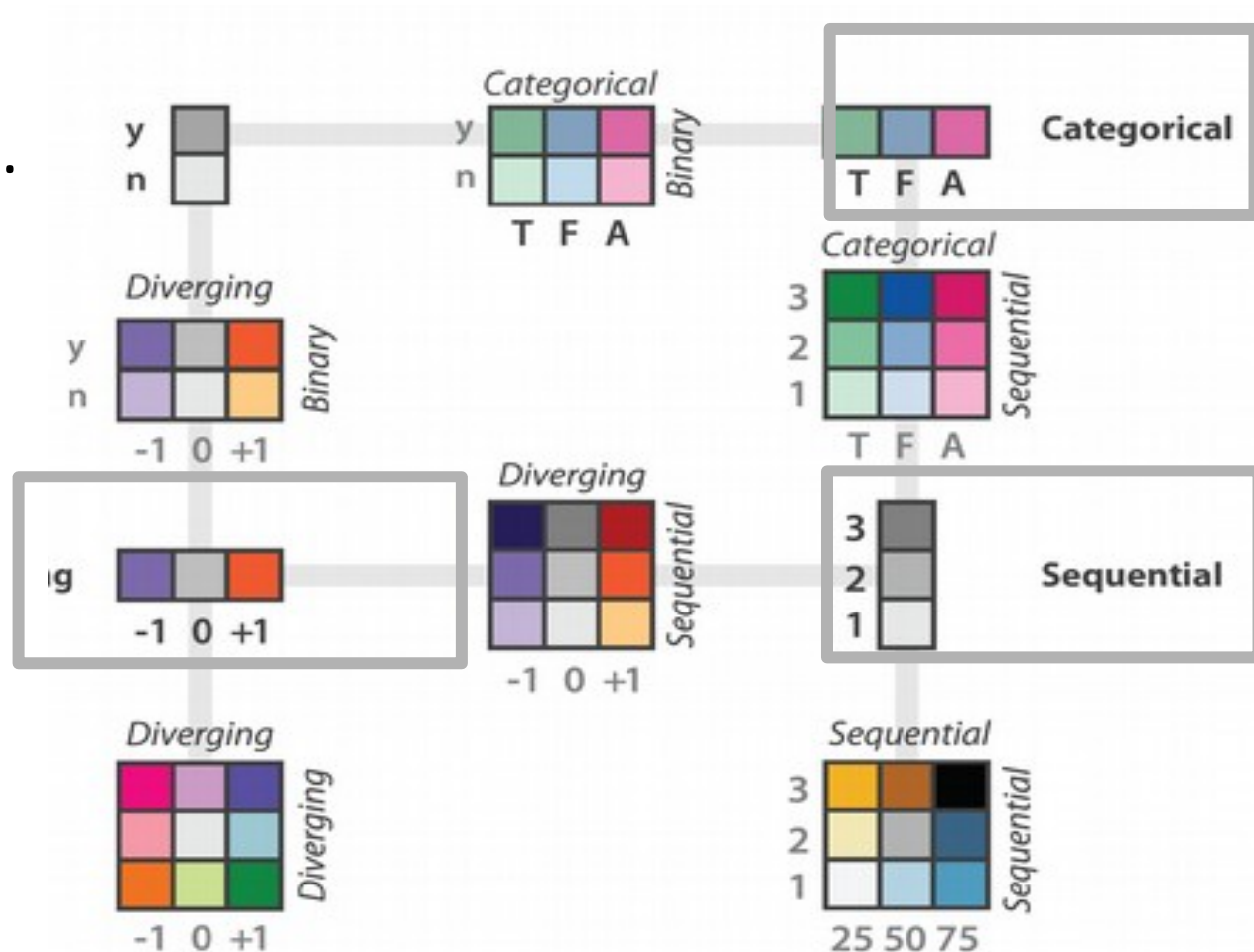
Colour palettes

 **Sequential:** between two values or colours. For quantitative distinctions.

 **Divergent:** colours diverge in opposite directions from a central value.

 Quantitative and qualitative.

 **Categorical:** no order in the colours. Qualitative.

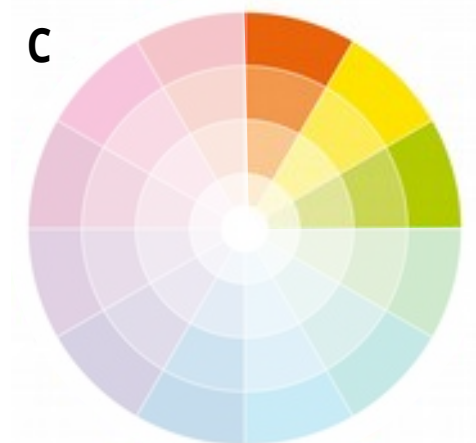
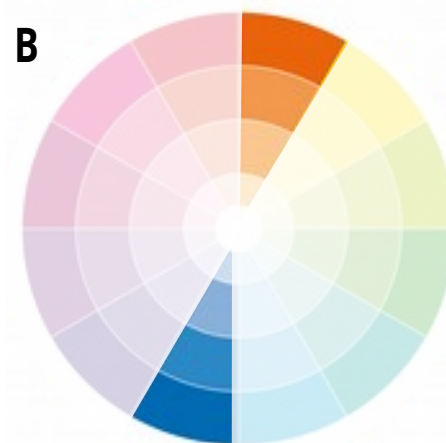
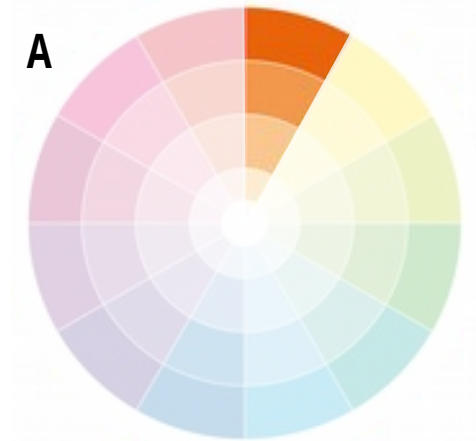


Images from Munzner

after [Color Use Guidelines for Mapping and Visualization. Brewer, 1994.
<http://www.personal.bsu.edu/faculty/c/a/cab38/ColorSch/Schemes.html>]

Colour: Choosing palettes

- The **colour wheel**. Choose combinations that are:
 - A. Monochromatic**: for a uniform look
 - B. Complementary**: to highlight differences between categories
 - C. Analogous**: for both
- Online colour pickers
e.g. <http://colorbrewer2.org/>

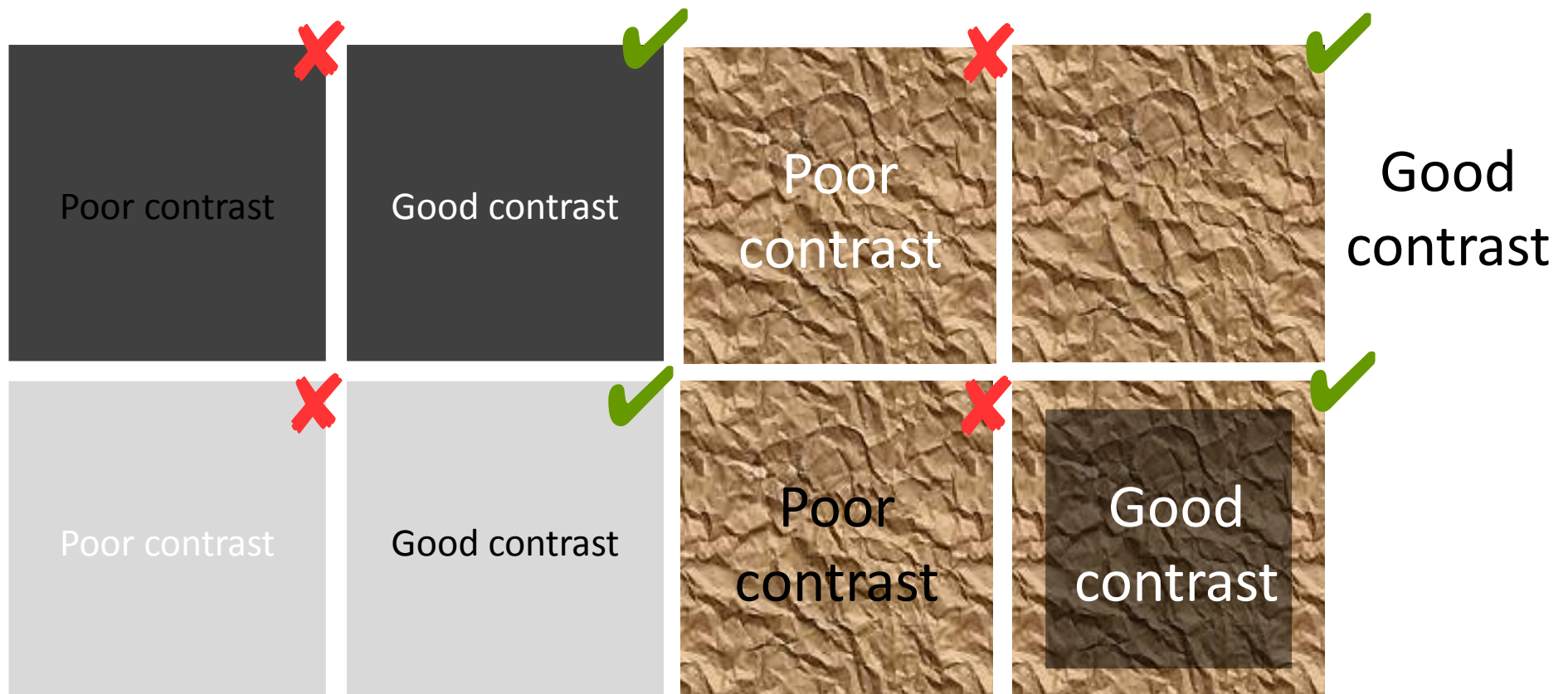


Colour: Choosing palettes

- Principles for choosing colours:
 - **Contrast**
 - Colour **blindness**
 - Black and white/ **grayscale** printing
 - How many **separable** colours in a legend?
- *“Black and white are colours, too”*

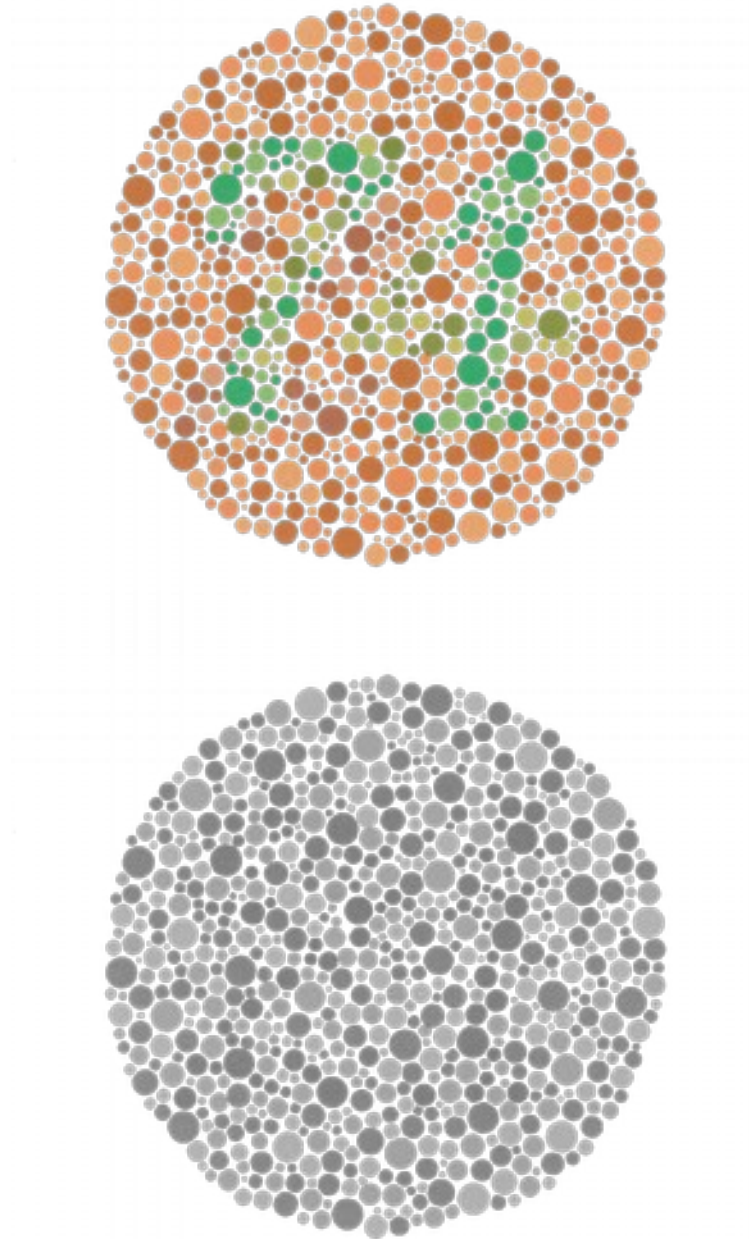
Contrast

Fine detail (e.g. text) needs good contrast to be visible
Beware of patterned backgrounds



Colour Blindness

- Affects 1:12 men and 1:200 women worldwide
- ***“If a submitted manuscript happens to go to three male reviewers of Northern European descent, the chance that at least one will be colour blind is 22 percent.”***



Types of colour blindness

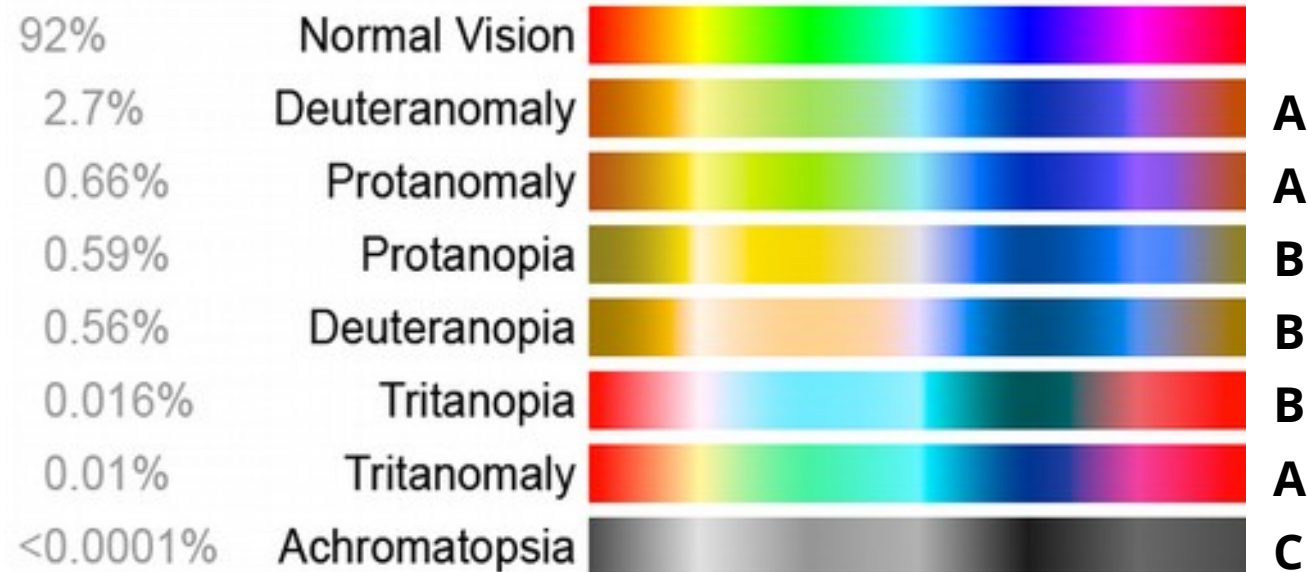
Normal vision: trichromacy (all 3 primary colours)

Colour deficient vision:

A. Anomalous Trichromacy: unusual 3 colour vision

B. Dichromacy: 2 colour vision

C. Monochromacy: black and white vision

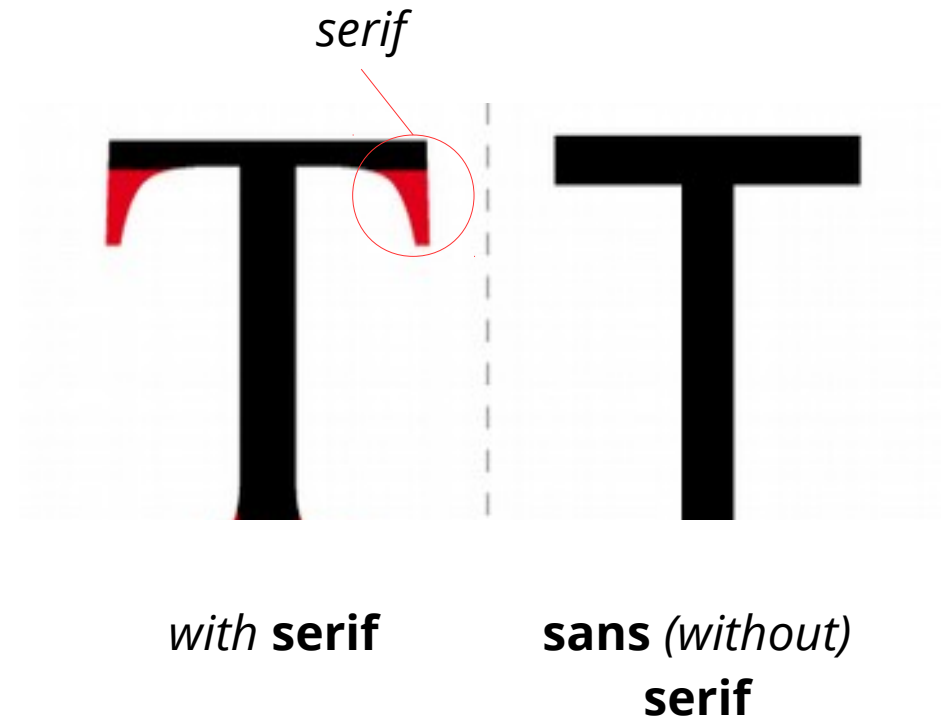


Typography

- All the elements need to be labelled
- The essential criteria for choosing fonts is **readability**:
 - **Scalability** (readable at small sizes)
 - **Contrast** with the background
- Fonts convey a personality or attitude (some more than others)

Labels and fonts

- **Serif** for large blocks of text, **sans serif** for titles and labels
- Monospace (e.g. m vs m and i vs i)
- Sizing, the size of fonts is given in points, and it's the size of an imaginary block of metal that is used in printing. In practice, the only way to know exactly how well your font will be read is to print it.
- Casing: UPPERCASE, lowercase, Sentence case, Title Case.
- Check the journal guidelines for font types



Typography: Typesetting

- Is the arrangement (spacing) of characters in words, lines or paragraphs
 - **Tracking:** space between characters
 - **Leading:** line height
 - **Paragraph alignment:** left, justified, etc.
- Important considerations where figures have many annotations, and in axis and figure titles.

Typography: Guidelines

- Avoid **aspect-ratio distortions**: changing font height or size.
 - The same applies to images and circular objects
 - Scale axes using comparable units
- **Avoid colour** in text (to maximise contrast)
- **Do not tilt** text, always horizontal or vertical
- Check **scalability**: text should be readable after resizing



Typeset in blocks of text that are **solid shapes**



Typeset in blocks of text that are **solid shapes**

Typography:

Heed the numbers in your font

1	l	1	1
2	2	2	2
7	7	7	7
6	6	6	6

- Each font has different styles of numbers
- Make sure that the font you choose distinguishes them well (e.g. l in *Gill Sans*) and is legible at small sizes

Typography:

Think your words carefully

- Avoid wordiness... it's a figure!
- Choose words that “precisely convey what you mean”
- Avoid contractions and spell out whenever possible

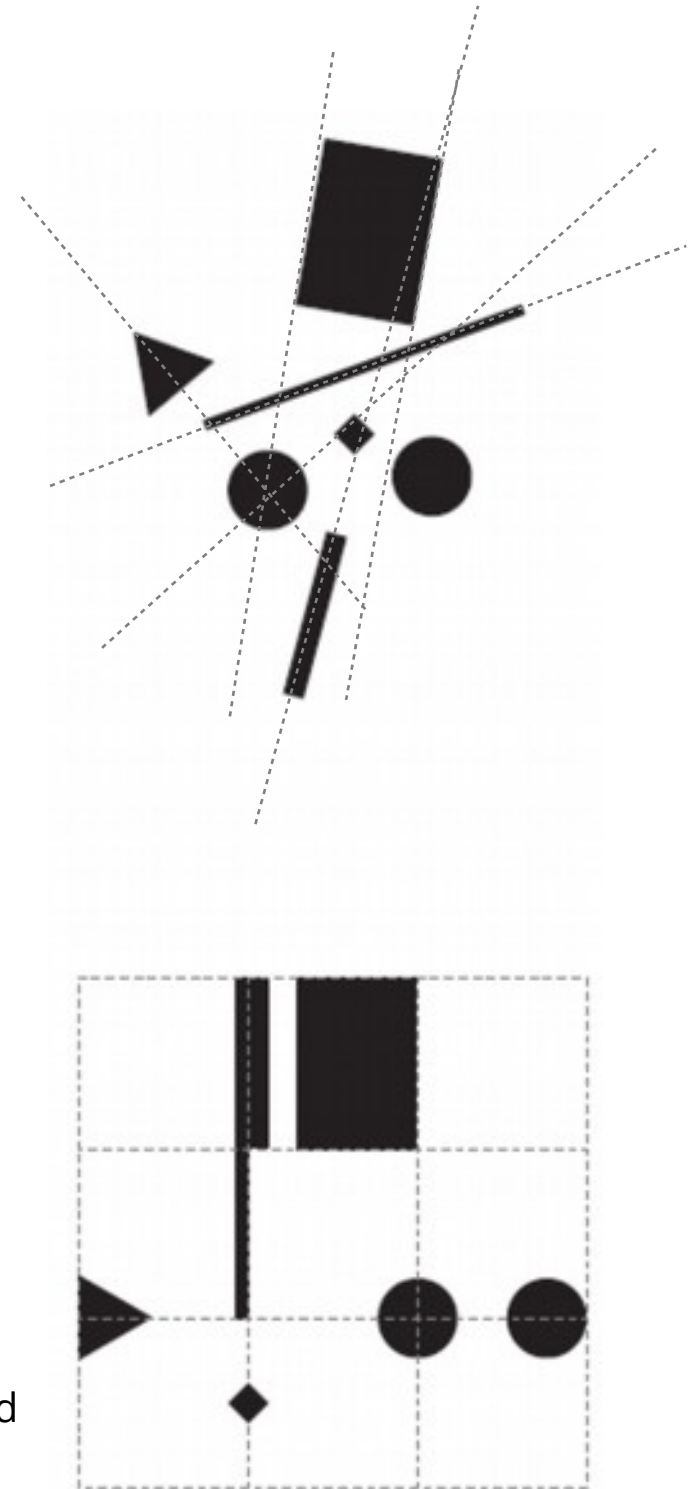
Composition and layout

- Grid and alignments
- Balance and hierarchy

Grids

- Grids are the invisible structure behind a composition that makes it look balanced
- Every alignment (of a box, column, text line and text margin) creates a **visual line** in the grid
- Conversely, a composition where elements are aligned to a grid creates a sense of balance

Grids can help to organize the spaces around and in-between elements. *Rolandi et al 2011*



Alignments

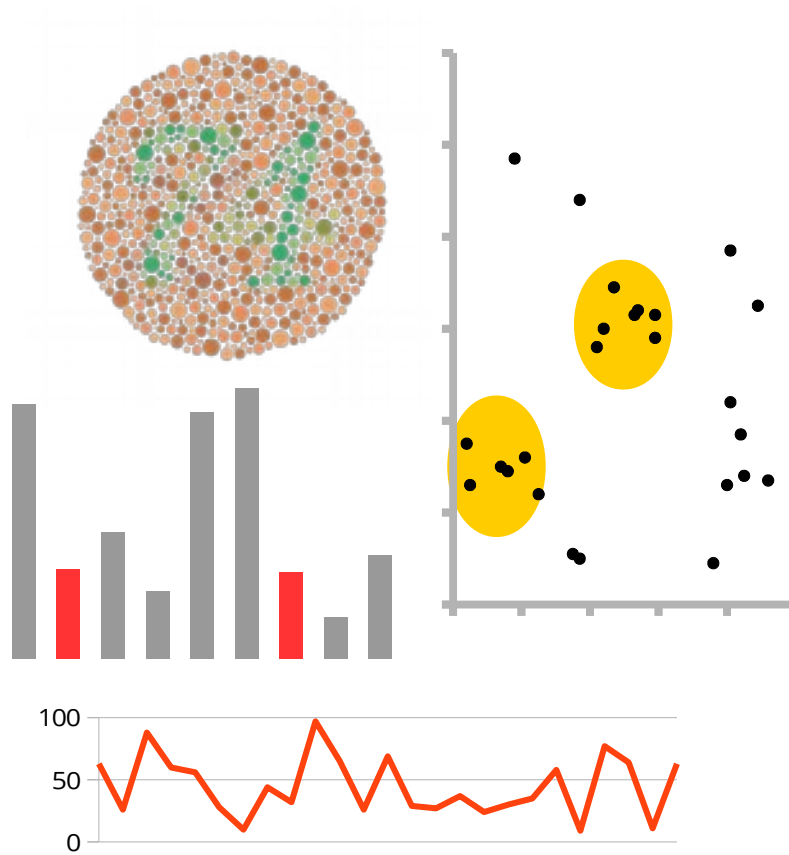


Alignments

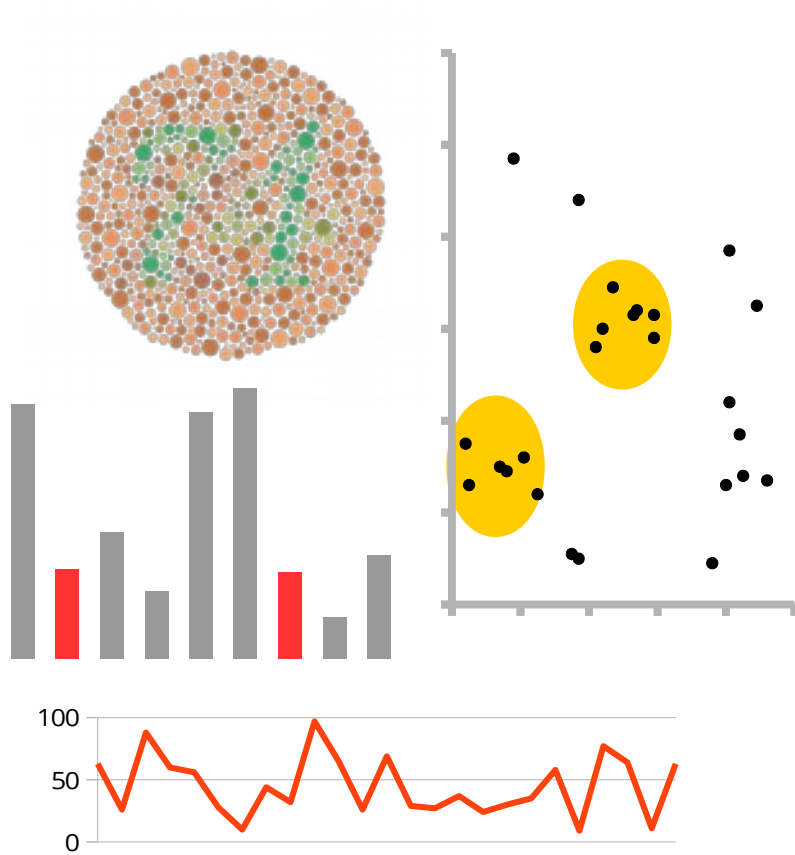


Most programmes have tools for automatic alignment and to distribute objects with equal space, e.g. Libreoffice, Inkscape (doing it by eye is not sufficient!)

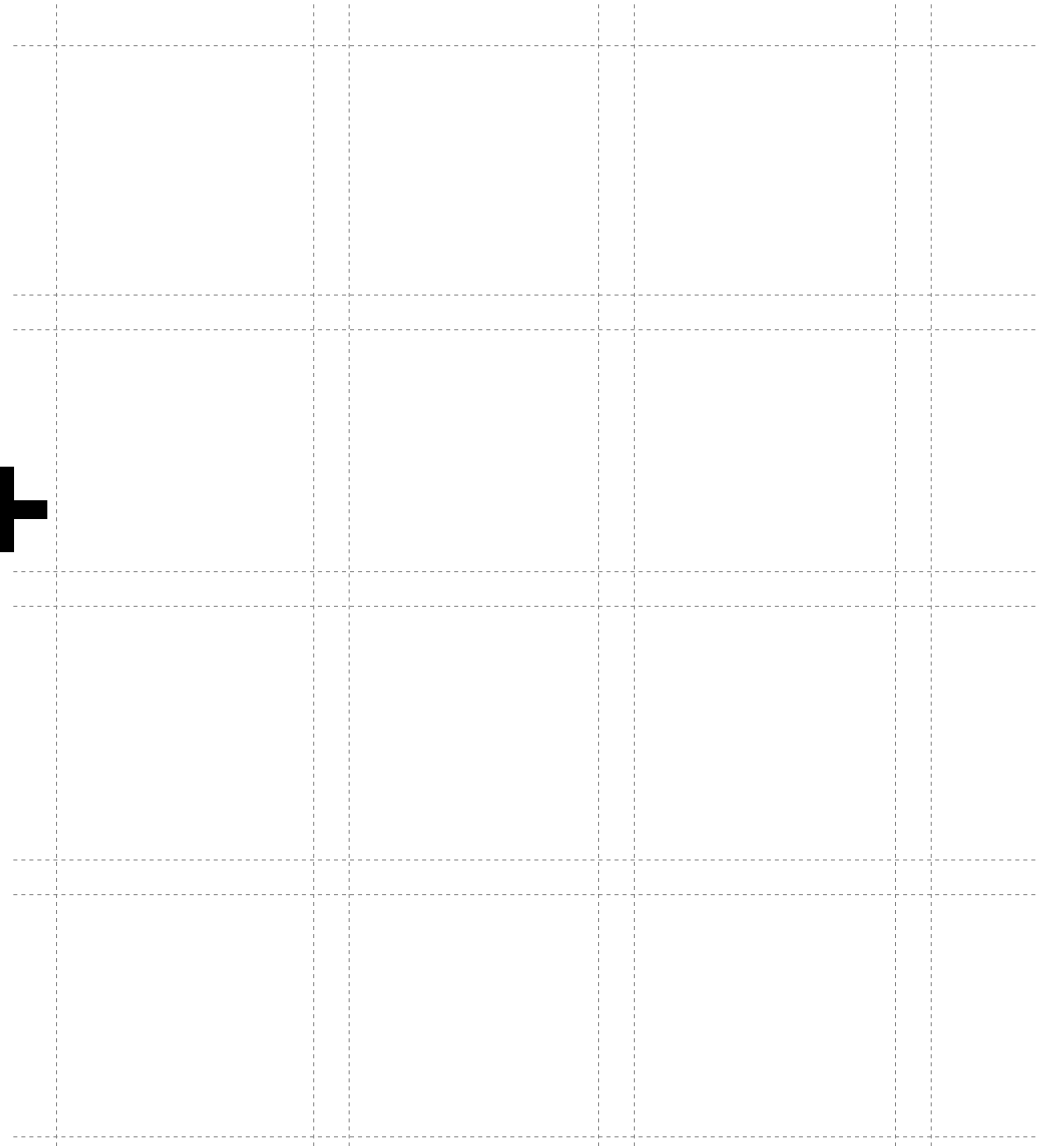
Using grids



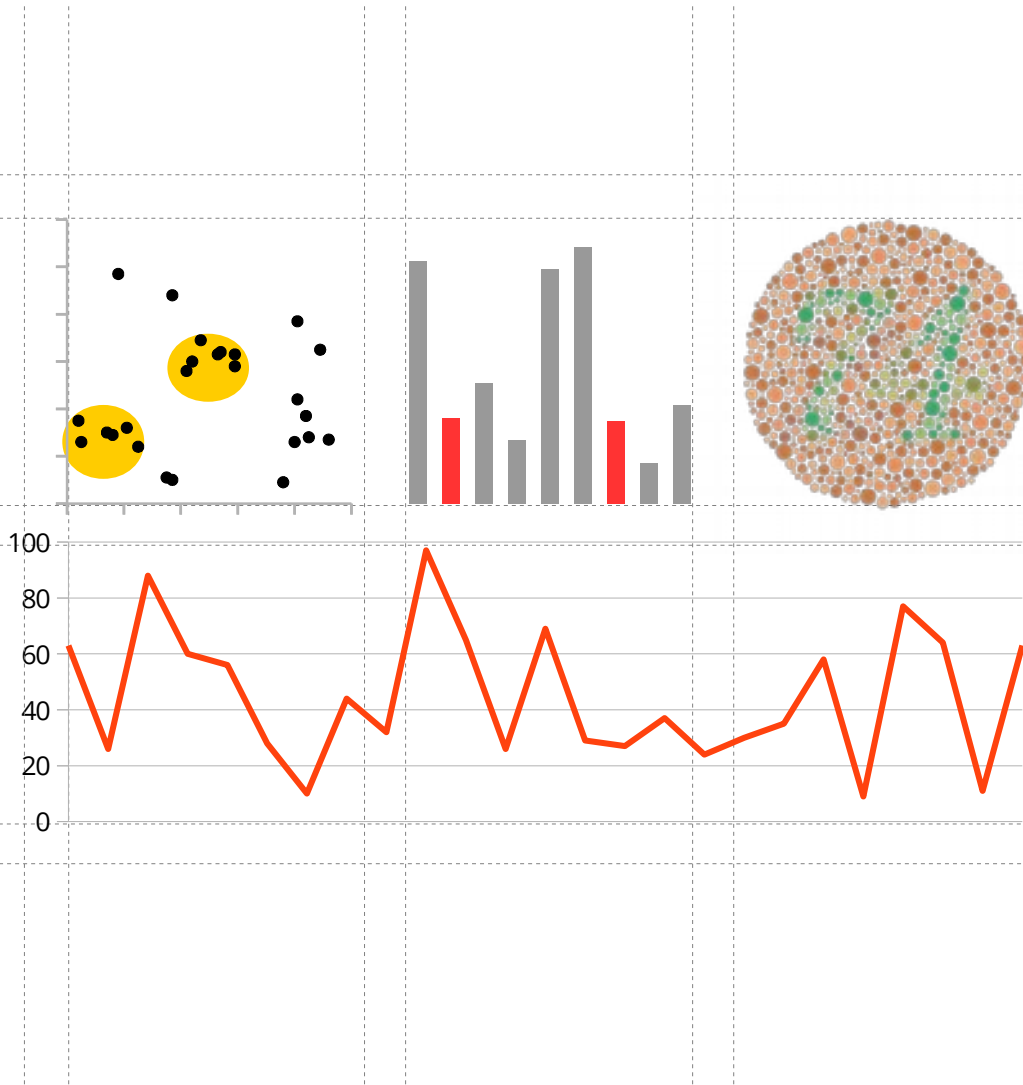
Using grids



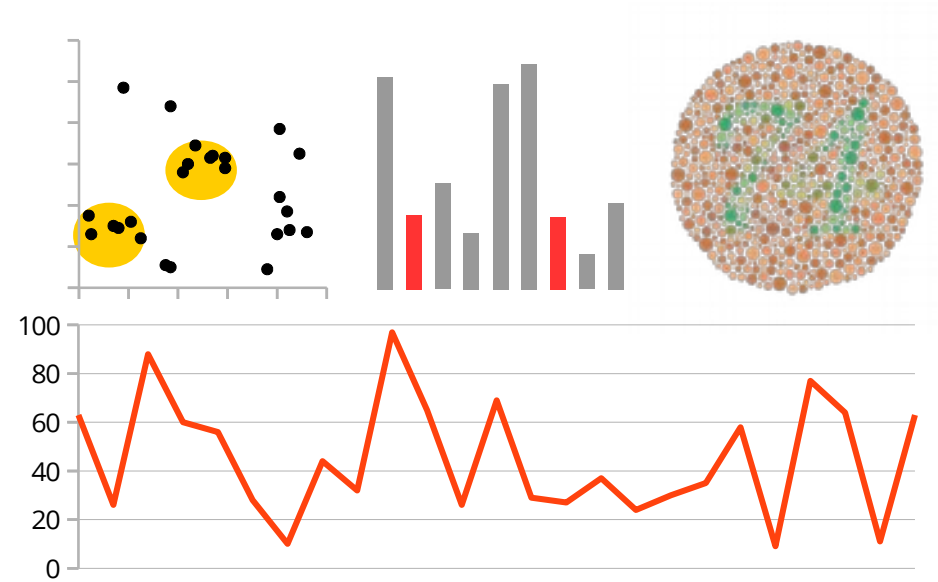
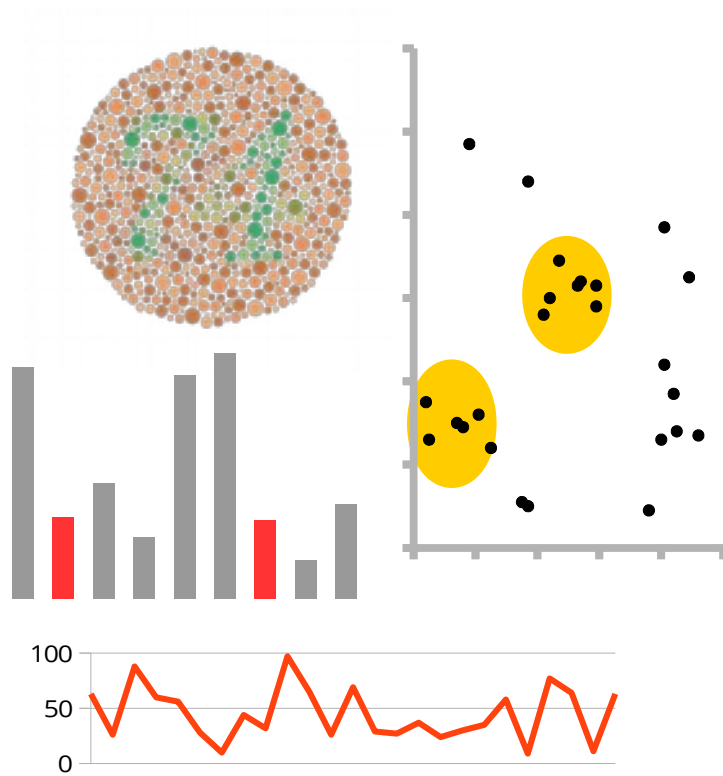
+



Using grids



Using grids



Visual balance and hierarchy

The composition of a graphic object and the **emphasis** on each element will determine what is the **hierarchy between elements**, and how the eye will **flow** and where it will **focus**

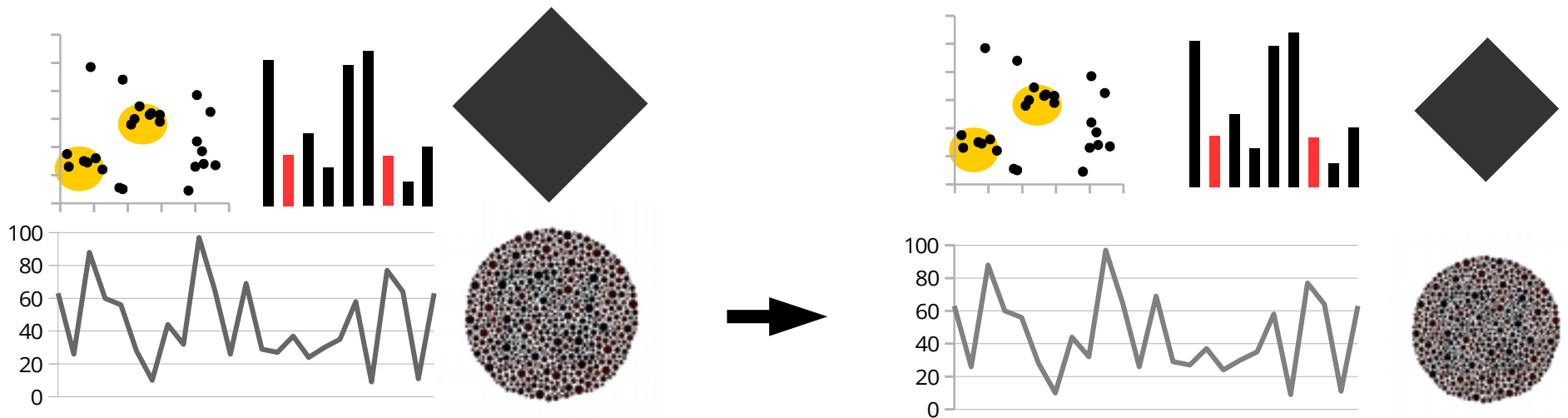
Keep a balance between **white space**, text and figures

Visual weight/ emphasis:

- How much an object on the page attracts and retains the attention of your viewer
- Depends on size, colour, position, etc.
- Should match the relevance of the information

These are some questions you can make to assess visual balance and flow: *Is there a clear (and justified) hierarchy or arrangement between elements? Can adjustments be made to make more relevant connections? Does the place feel cluttered/ scattered?* (Krause, 2004)

Visual balance



In the left figure, the black diamond and, to a lesser extent, the circle stand out (is this our intention?). There is also little separation between the charts, which makes the figure look cluttered.

General tips

Don't-s:

- Don't distort the data
- No unnecessary figures or elements: *do we really need a figure? or a table would suffice?*
- Don't rely absolutely on colour
- No 3D: in most cases it distorts perception

Do-s:

- One point per figure
- Summarise to clarify
- Have a clear purpose/ message
- Link to accompanying text and statistics

Figure ethics

- The figure/graph/image should show **what is actually happening** and **not what you want** to happen
- **Is my plot ethical?**
 - Would a reader come to a different conclusion if they could see the details of the data which were omitted from the plot?

Unethical figures

- Not exploring/getting to know the data well enough
- Choosing the wrong graph to present the data
- Choosing the wrong axis/scale.
 - e.g. logarithmic scale: For cheating, a bar graph using a log axis is a great tool, as it lets you either exaggerate differences between groups or minimize them.
- Simply cheating: choosing the 'most representative' experiment or manipulating images

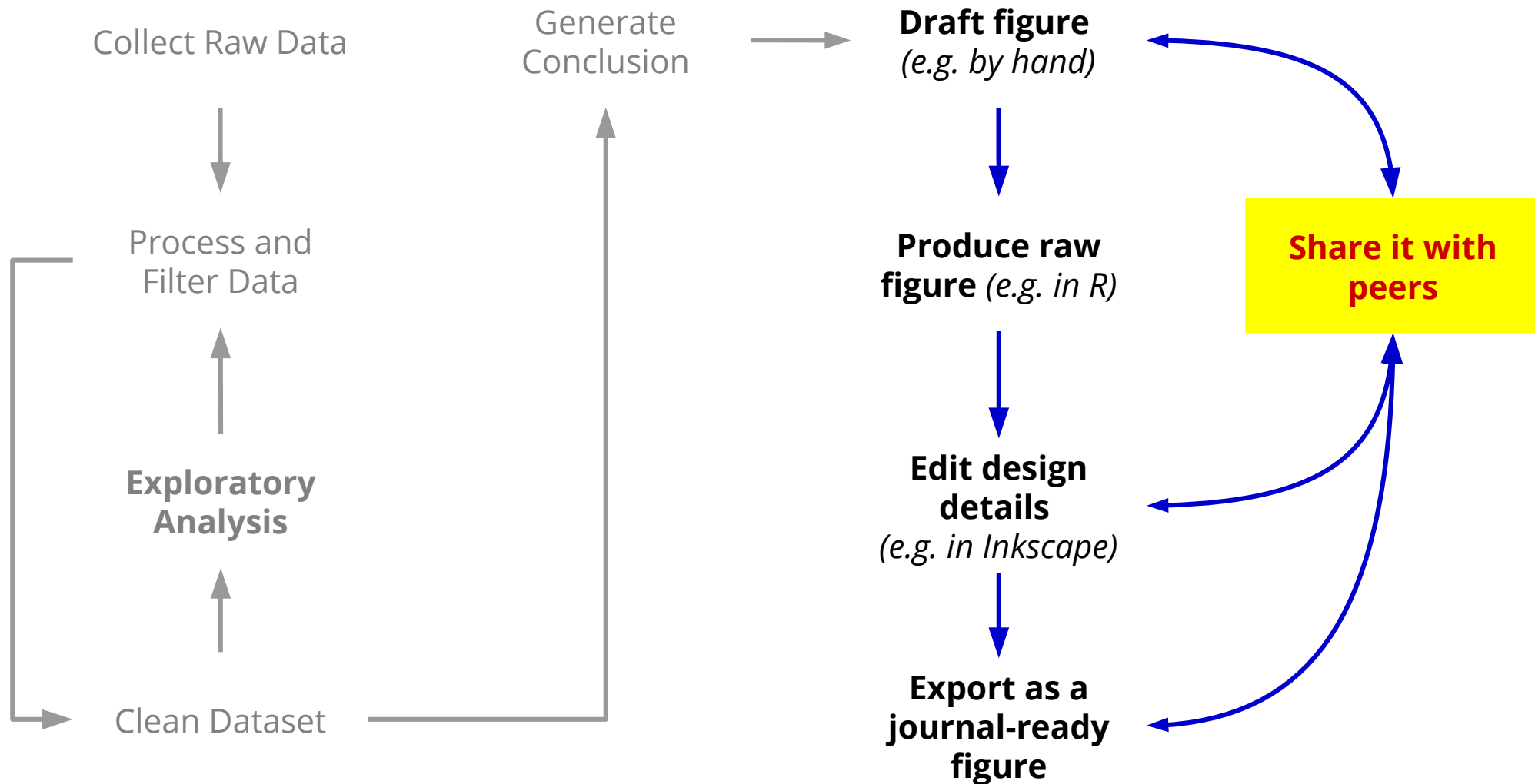
Checklist

Is your figure effective?

- ☐ The figure is **self contained**: understandable without additional information
- ☐ Every element is **labelled** or explained in the caption, including x and y units
- ☐ x and y axis: **scales** show appropriate variation of the data, or are comparable
- ☐ **Readability** and **contrast** are appropriate
- ☐ Every use of **colour** has a reason
- ☐ The figure works in **grayscale** (except for very complex figures)
- ☐ If there are **groupings**, they help understand the message without manipulating
- ☐ There are no channel **inconsistencies** within the figure
- ☐ It is as **simple** as possible: i.e. no decorations, every piece that could be eliminated without losing information has been eliminated
- ☐ Has been **validated** with other people...

Validation

Data Visualisation Process



Validation

- Always try to validate plots you create
- You have seen your data too often to get an unbiased view
- Show the plot to someone not familiar with the data
 - What does this plot tell you?
 - Is this the message you wanted to convey?
 - If they pick multiple points, do they choose the most important one first?

Not covered in this session

Diagrams

- Definition
- Workflow:
 - Clarify the purpose: essential elements to depict and their relation
 - Draft the structure of the diagram by hand and share and discuss it
- Use grids and think carefully about the label choice and position
- Types: Venn diagrams (composition of datasets), flowcharts (for decision making processes), tree diagrams, timelines, networks, pathways, procedural diagrams
- Remember: the key “is not the quality of the diagram or drawing, but the clarity of the information” Carter p128

Photos

- Avoid unethical manipulation (deleting noise, etc.), even if it doesn't change the results
- Crop to emphasize important bits
- Rule of thirds
- Use good quality images (sufficient resolution and colour/ brightness settings)
- Format differences: JPEG, TIFF, GIF, PNG
- Resolution
- Cropping and image composition
- Image size and proportions
- In context: contrast and relation with surrounding content
- Check license for use

Some useful resources

- Short papers:
 - **Rolandi** et al 2011. *A Brief Guide to Designing Effective Figures for the Scientific Paper*. *Advanced Materials* 23
 - **Rougier** et al 2014. *Ten Simple Rules for Better Figures*. *Plos Computational Biology* 10:9
- Design for scientists/ data:
 - **Carter**. 2013. Designing science presentations – *not just figures, very clear*
 - **Munzner**. 2014. Visualization, analysis and design
 - *from a computer-graphics perspective*
 - **Tufte**. 2001. The visual display of quantitative information
 - *from a theory-of-design perspective*
 - **Meirelles**. 2013. Design for information
 - *advanced information visualizations (maps, time-space, flows)*
- Graphic design more generally:
 - **Krause**. 2004. Design basics index – *very concise and to the point*
 - **Samara**. 2014. Design elements: a graphic design manual – *reference book*
- Nature Points of View:
<http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html>

If you need additional help: **az296@cam.ac.uk**

<http://aiora.zabala.net/portfolio>