Cancer Research UK

Designing more effective scientific figures

Aiora Zabala
PhD Environment. VTP Graphic Design
az296, aiora.zabala@gmail.com

ScienceDaily®

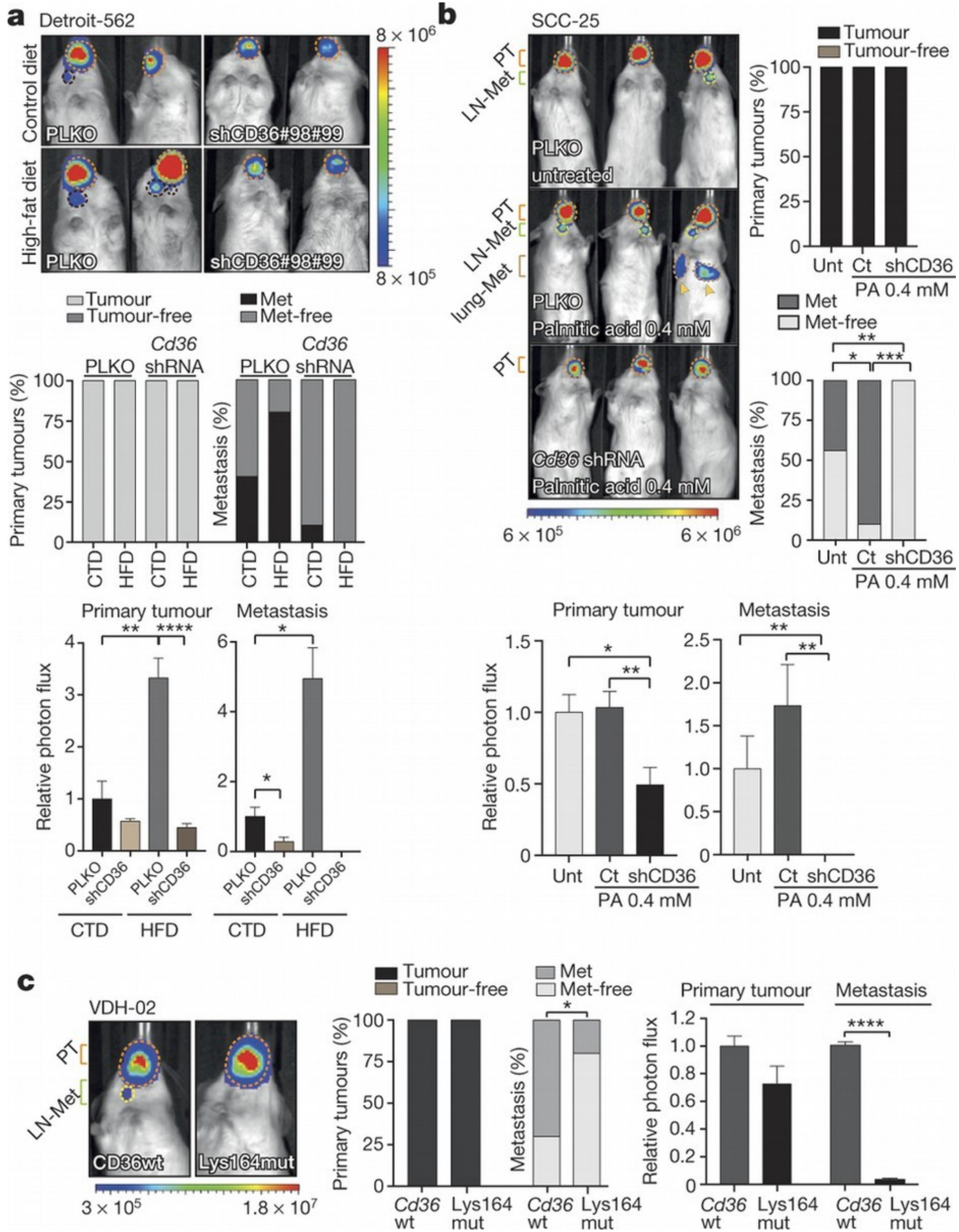
Cancer spread is increased by a high fat diet, ground-breaking evidence shows

Researchers discover new cancer spreading protein

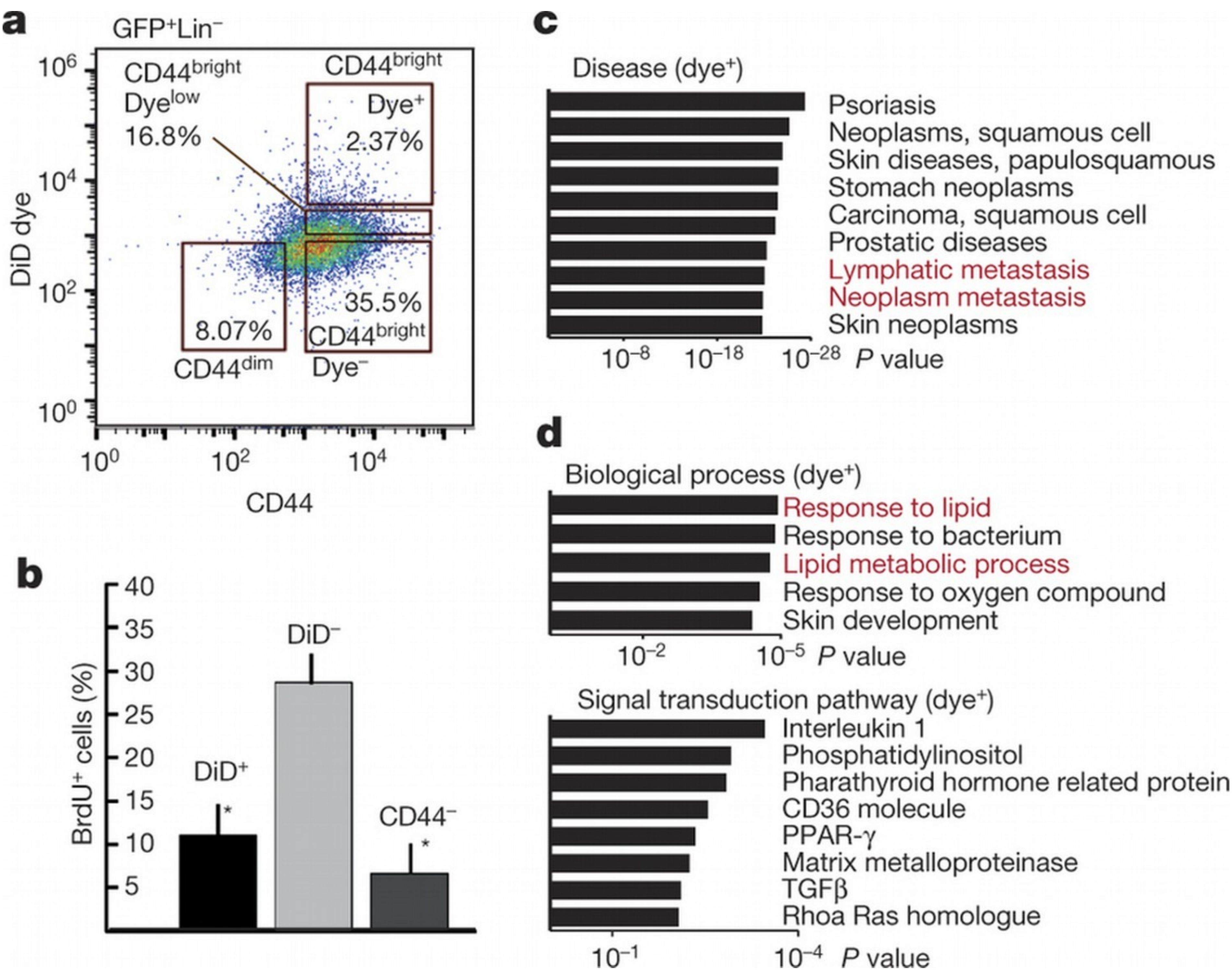
Date: December 7, 2016

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


a
GFP+Lin−
CD44bright Dyelow 16.8%
CD44bright Dye+ 2.37%
35.5% CD44bright Dye−
8.07% CD44dim
DiD dye
CD44
b
BrdU+ cells (%)
DiD+
DiD−
CD44−
c
Disease (dye+)
Psoriasis
Neoplasms, squamous cell
Skin diseases, papulosquamous
Stomach neoplasms
Carcinoma, squamous cell
Prostatic diseases
Lymphatic metastasis
Neoplasm metastasis
Skin neoplasms
P value
d
Biological process (dye+)
Response to lipid
Response to bacterium
Lipid metabolic process
Response to oxygen compound
Skin development
P value
Signal transduction pathway (dye+)
Interleukin 1
Phosphatidylinositol
Pharathyroid hormone related protein
CD36 molecule
PPAR-γ
Matrix metalloproteinase
TGFβ
Rhoa Ras homologue
P value

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 [efficient], 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

- An introduction to two (free) programmes to edit images for professional results:
  - *Gimp*: for bitmap images (equiv. to *Ad. Photoshop)*
  - *Inkscape*: for vectorial images (equiv. to *Adobe Illustrator*)
- Produce a journal-ready figure using standard analysis software and Inkscape

Structure of this course

	THEORY	PRACTICAL
Morning	1 Why figure design? Principles of figure design Elements of a figure Colour & ethics	2 **Gimp** – *bitmap (e.g. jpg)* Setting up a canvas Layers and importing files Editing colour Export formats and qualities
Afternoon	3 Dealing with complexity Choosing the right figure Typography Composition & layout	4 **Inkscape** – *vectorial (e.g. pdf)* Document properties Create & manipulate objects Composition Import & export for publication

What is data visualisation?

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

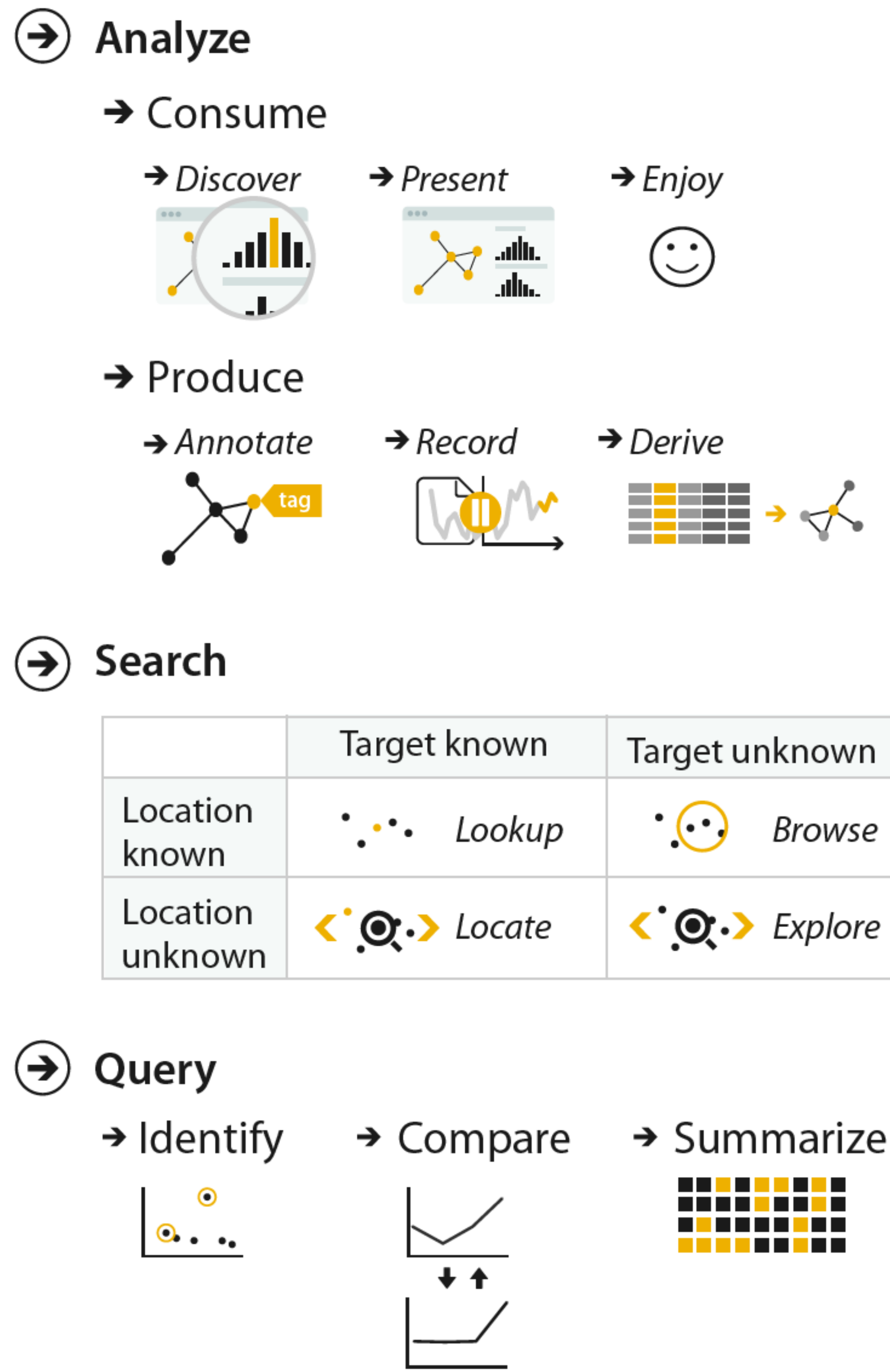


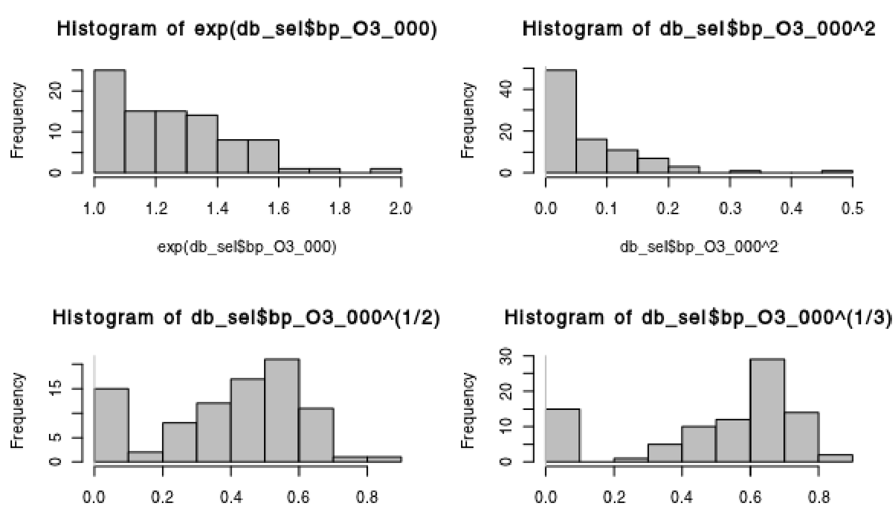
Figure: Tamara Munzner

Exploratory visualisation

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

Slide from Babraham Bioinformatics

Correlogram to see bivariate relations



Histograms to see the **distribution** of variables

Interactive data exploration with **GGobi** (in R via *rggobi* package)

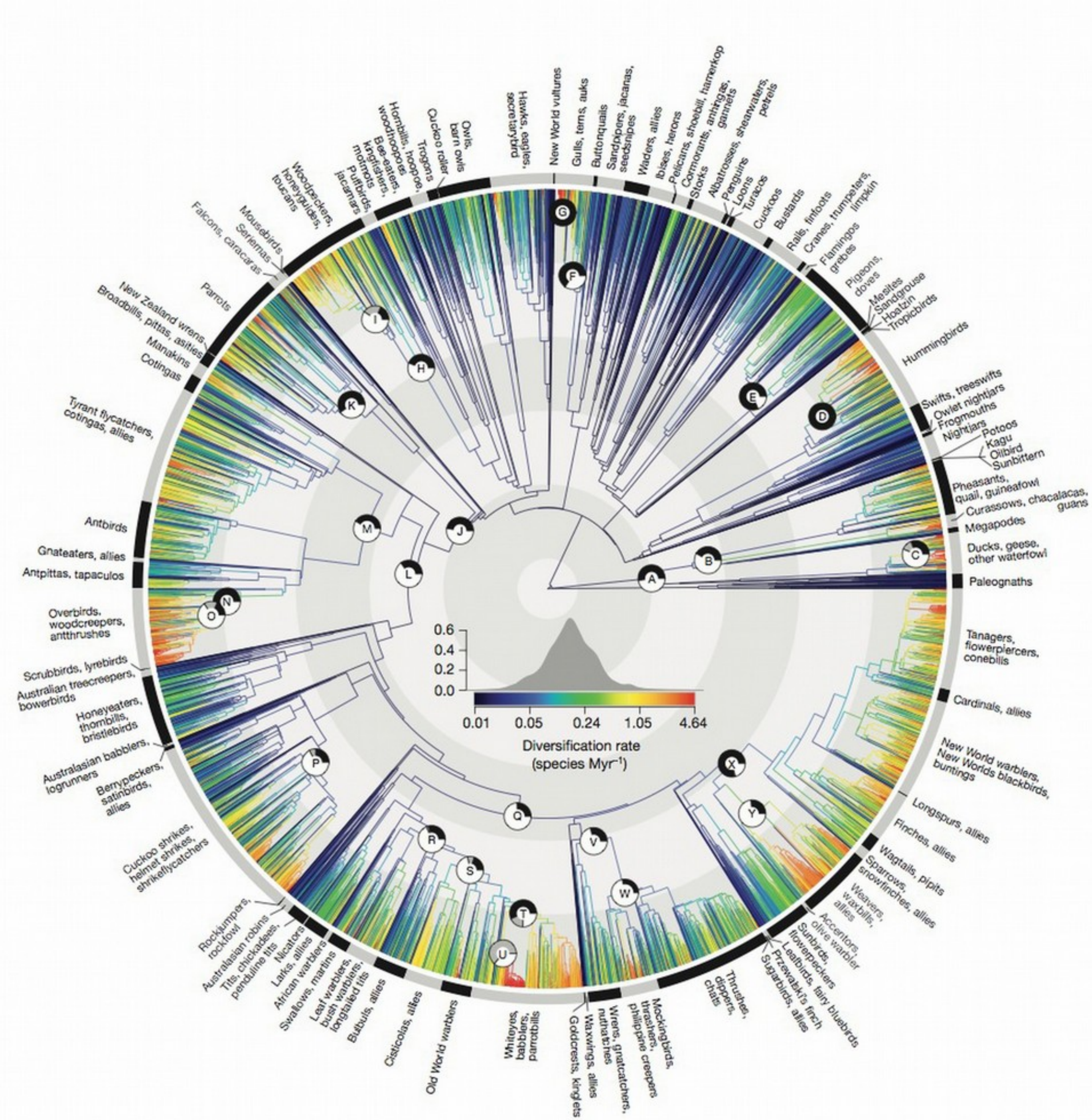
Reference visualisation

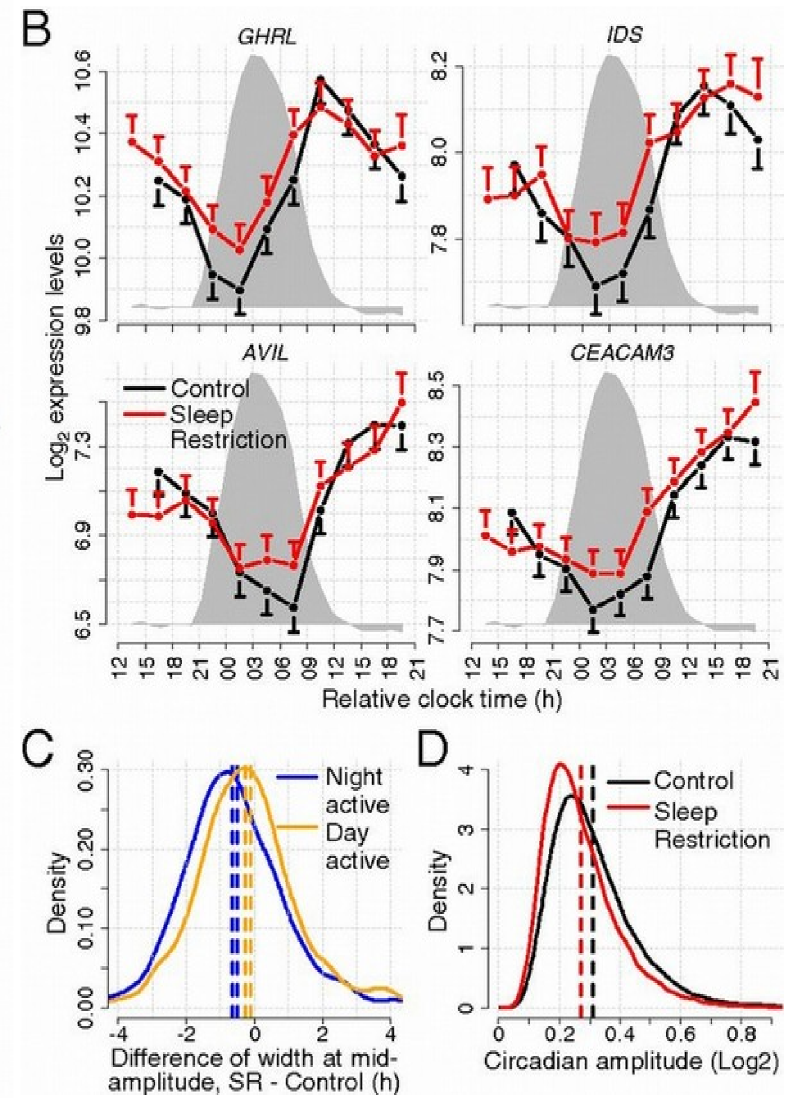
- Using your data as a resource
- Allows users to look up data of interest
- Tabular / Configurable
- Interactive

Illustrative visualisation

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

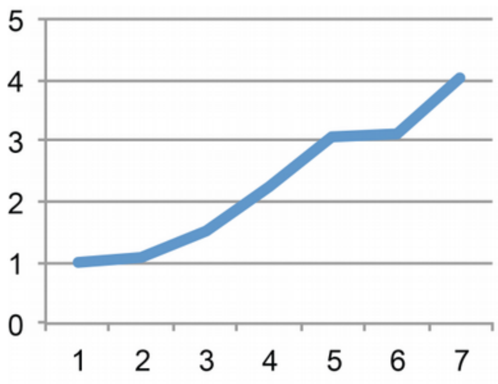
Slide from Babraham Bioinformatics



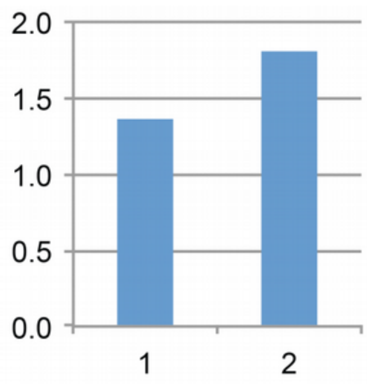


Figures: Avian phylogenetic tree, Jetz et al 2012. Sleep deprivation and genetic expression Möller-Levet et al 2013

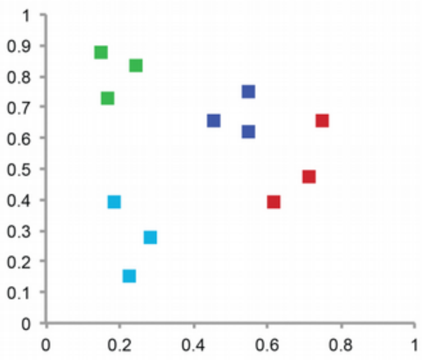
Things you can illustrate



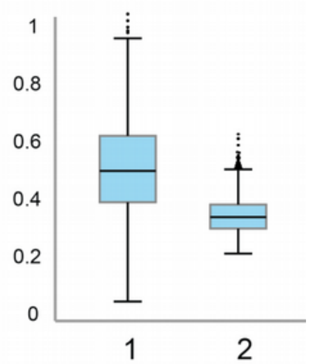
Relationship



Comparison



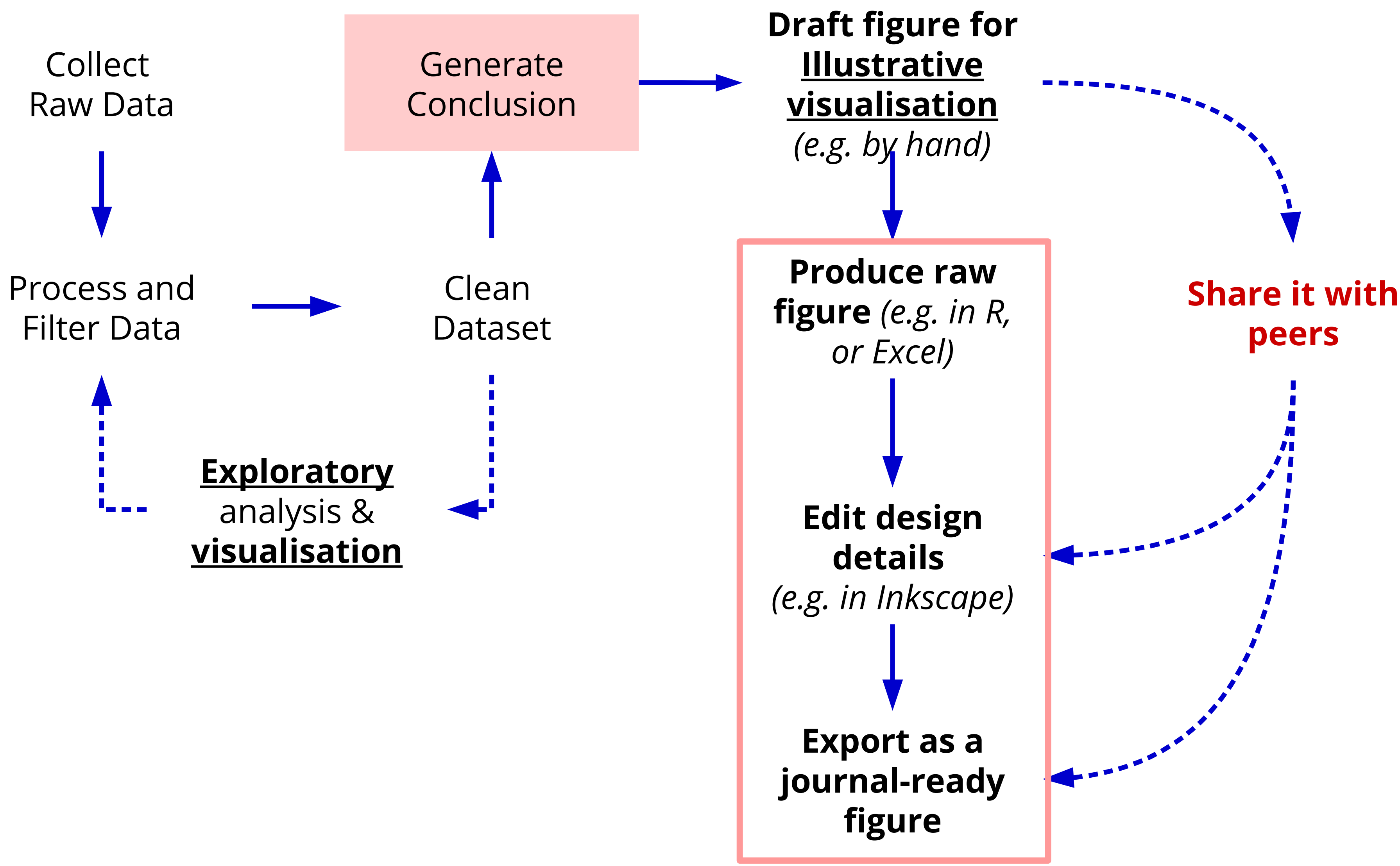
Composition



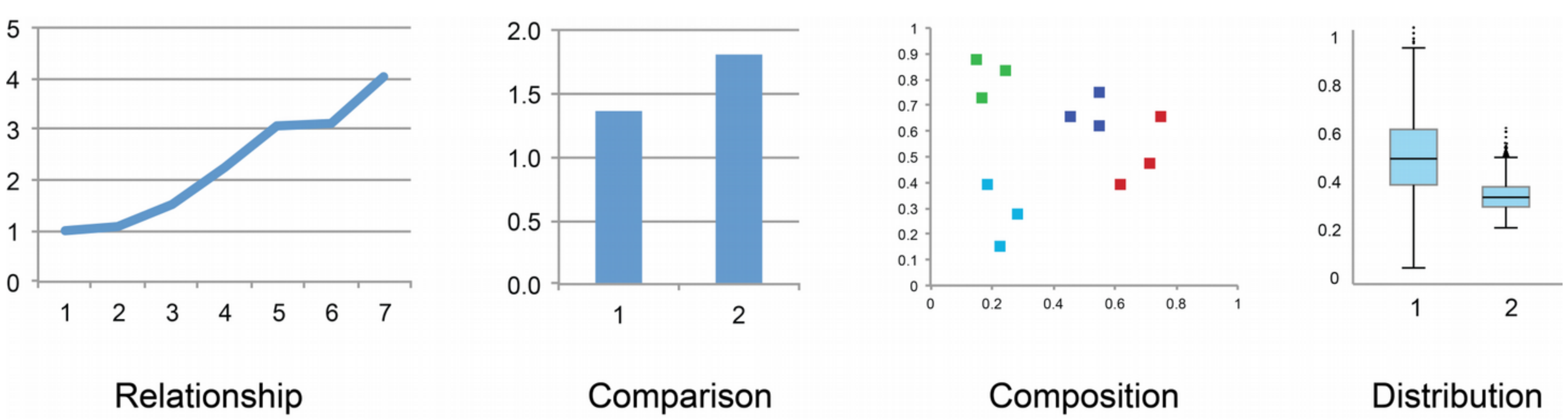
Distribution

Slide from Babraham Bioinformatics

Data Visualisation Process



Things you can illustrate



Slide from Babraham Bioinformatics

Graphical Representations

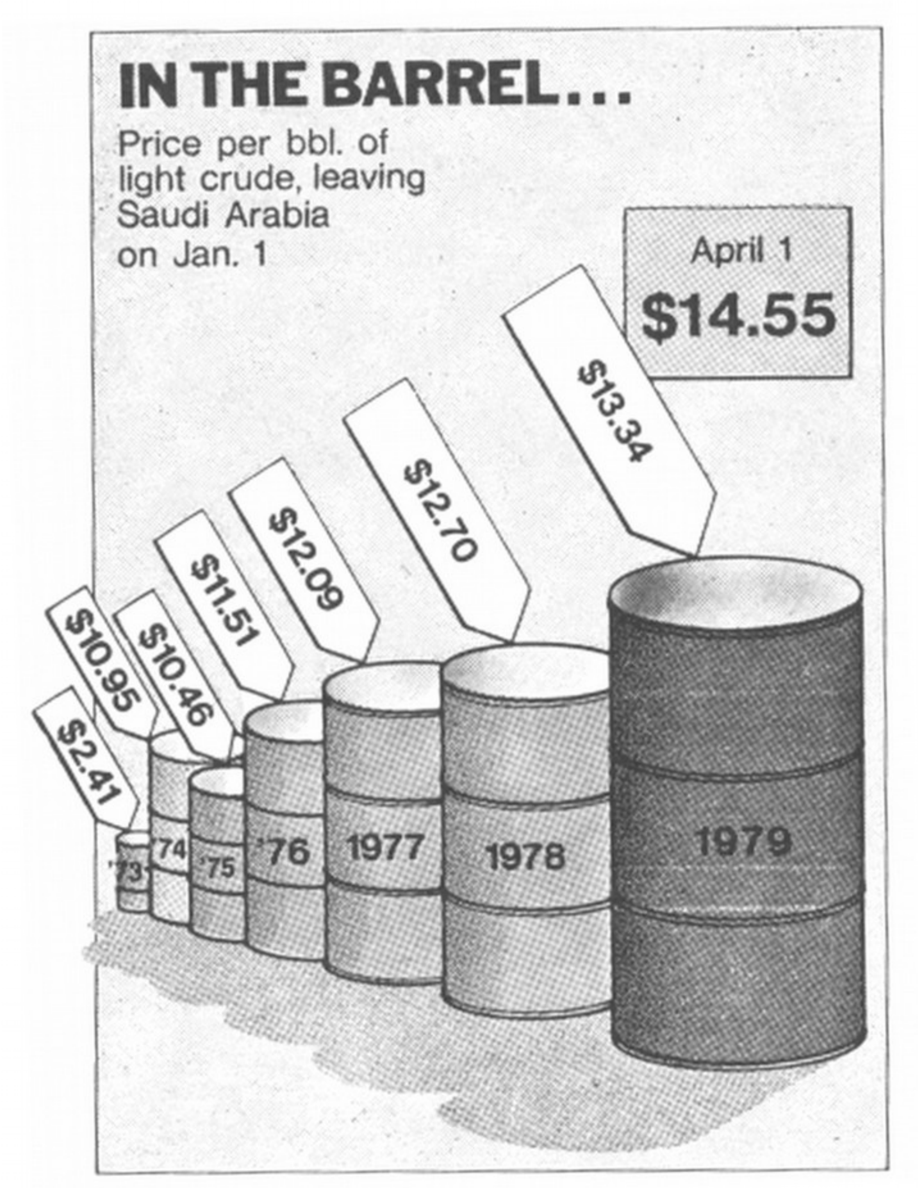
- Basic questions:
  - How are you going to **turn the data into a graphical form** (weight becomes length etc.)
  - How are you going to **arrange things in space**
  - How are you going to use **colours, shapes** etc. to clarify the point you want to make

Slide from Babraham Bioinformatics

Key ingredients: principles

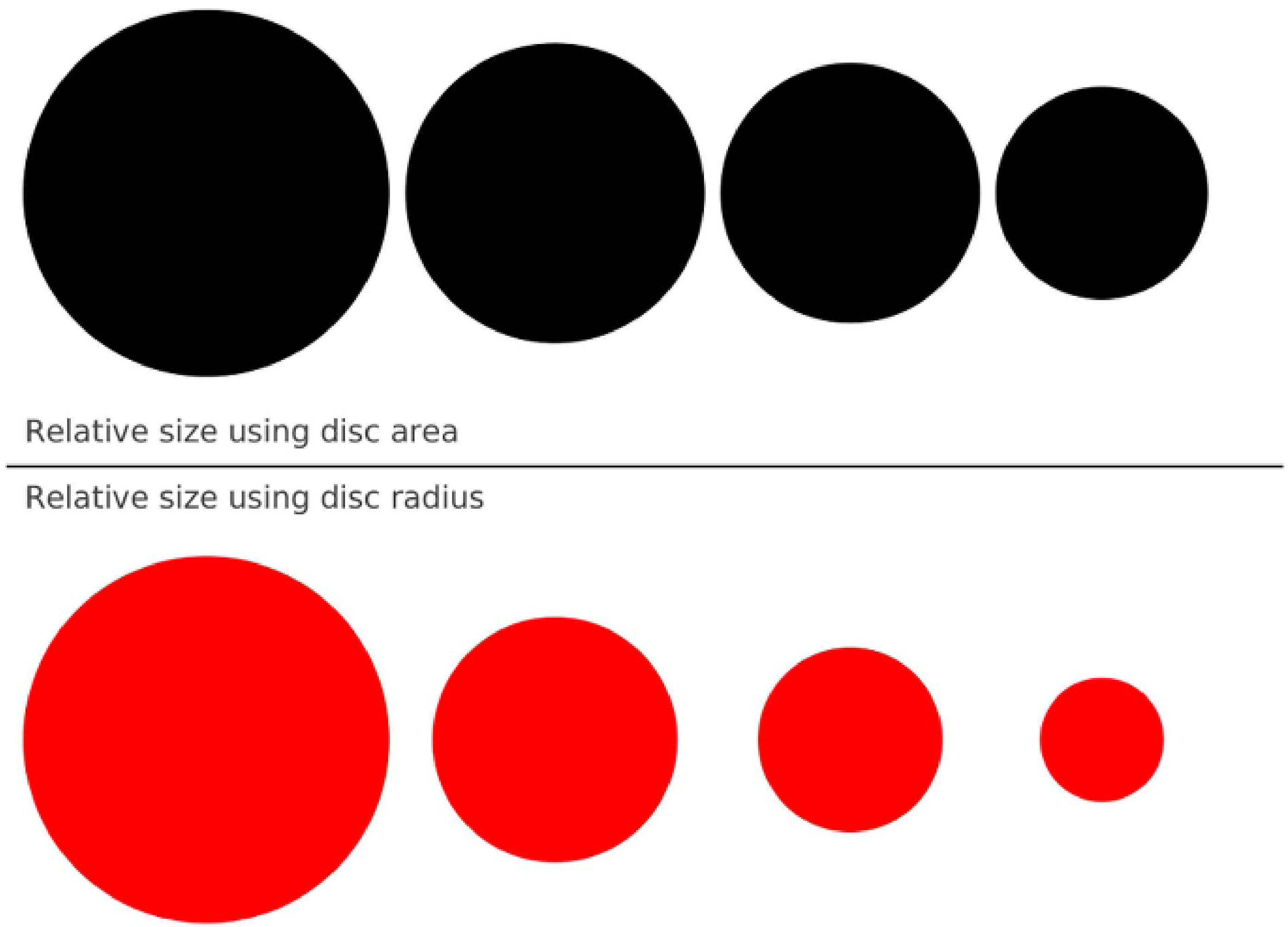
- **Simplicity**
- **Continuous** process
- *Rules* in graphic design are, as in many other disciplines, rather guidelines: you can break them to allow for creativity and when there is a good reason to break them, but you need to know how to use them.

What makes a good figure?



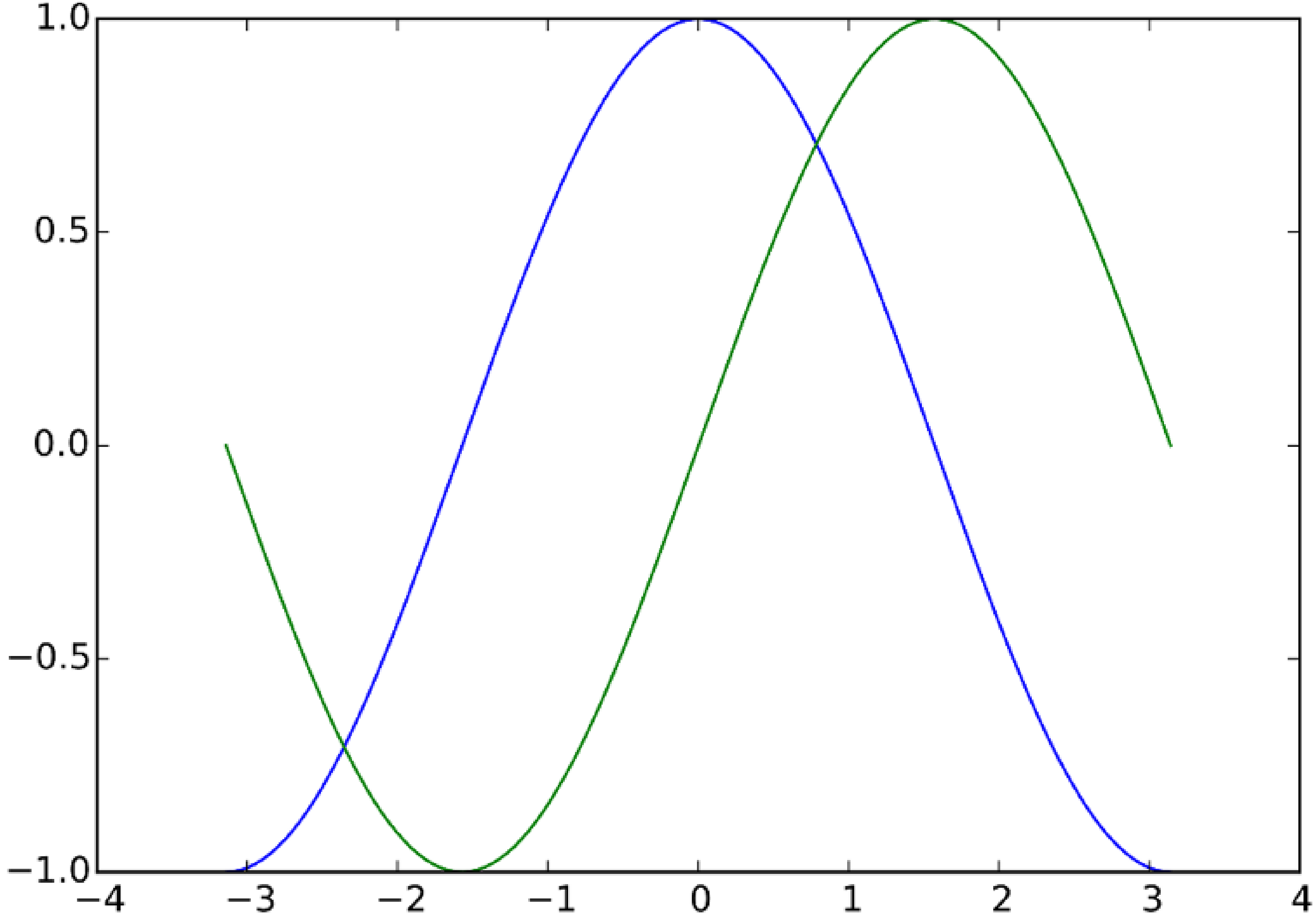
Time c1979 in Tufte

Areas and 3D can be misleading



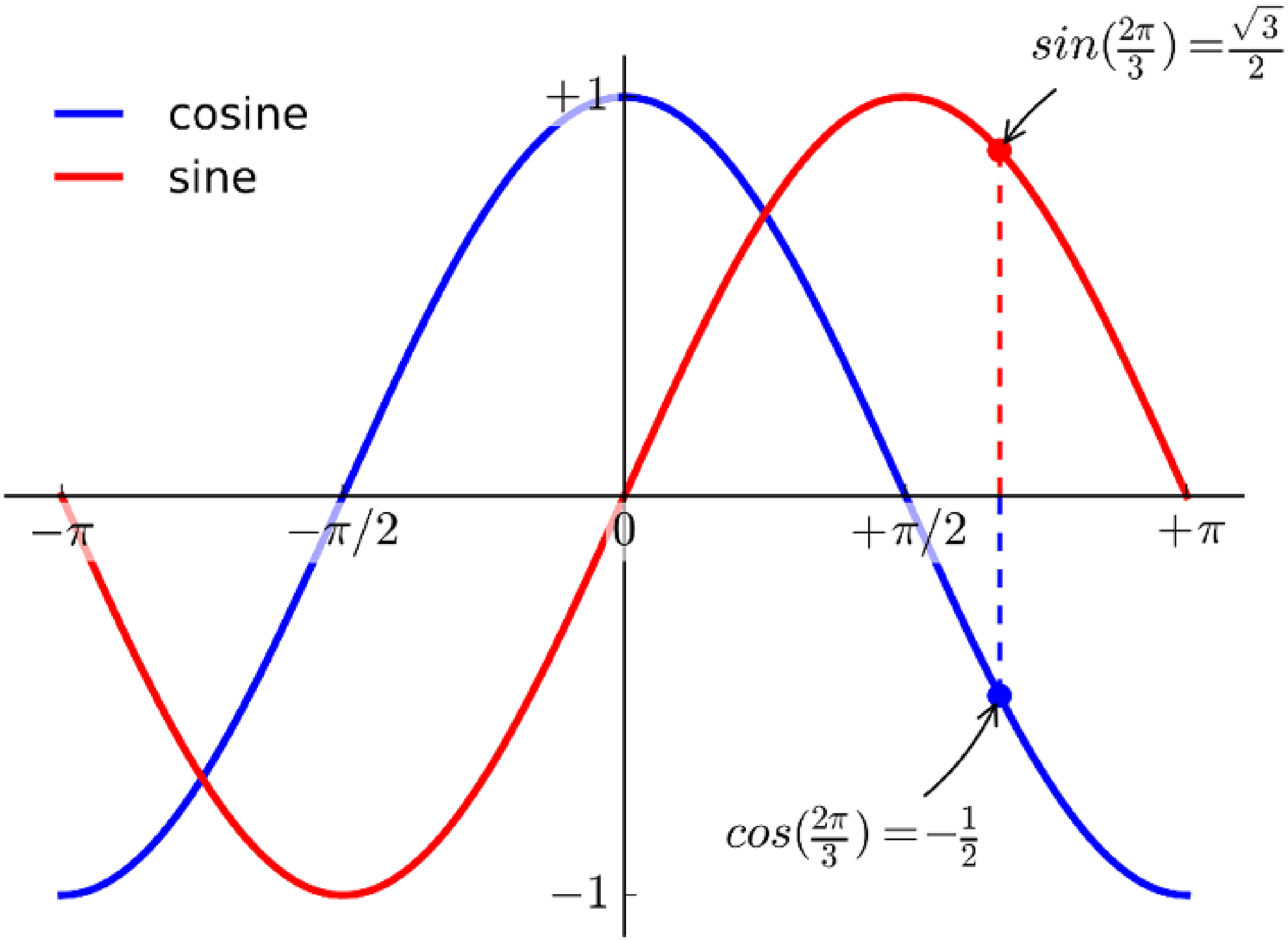
Rougier et al 2014

What makes a good figure?



Rougier et al 2014

What makes a good figure?



Rougier et al 2014

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:

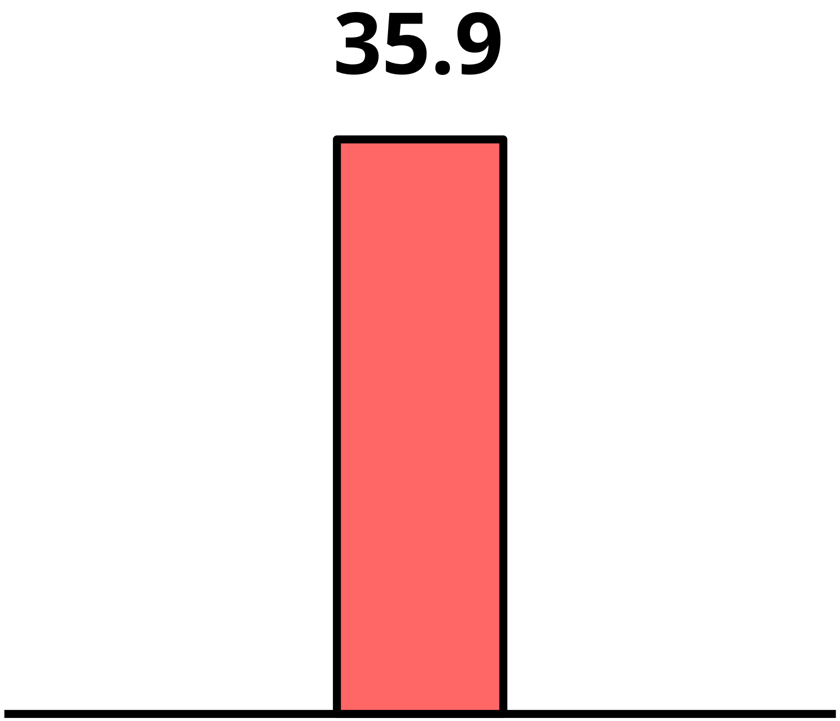
- Show the data
- Link to the accompanying text and statistics
- Summarise to make things clearer
- Serve a clear purpose
- Not distort the data

Slide from Babraham Bioinformatics

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?

Example from Tufte

Consistency

- Make the figures uniform to helps 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

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
- Think about the **audience**:
  - Specialised journal or broader readership?
  - Be aware of disciplinary conventions

Slide from Babraham Bioinformatics

Key ingredients: the tools

Elements: marks and channels

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

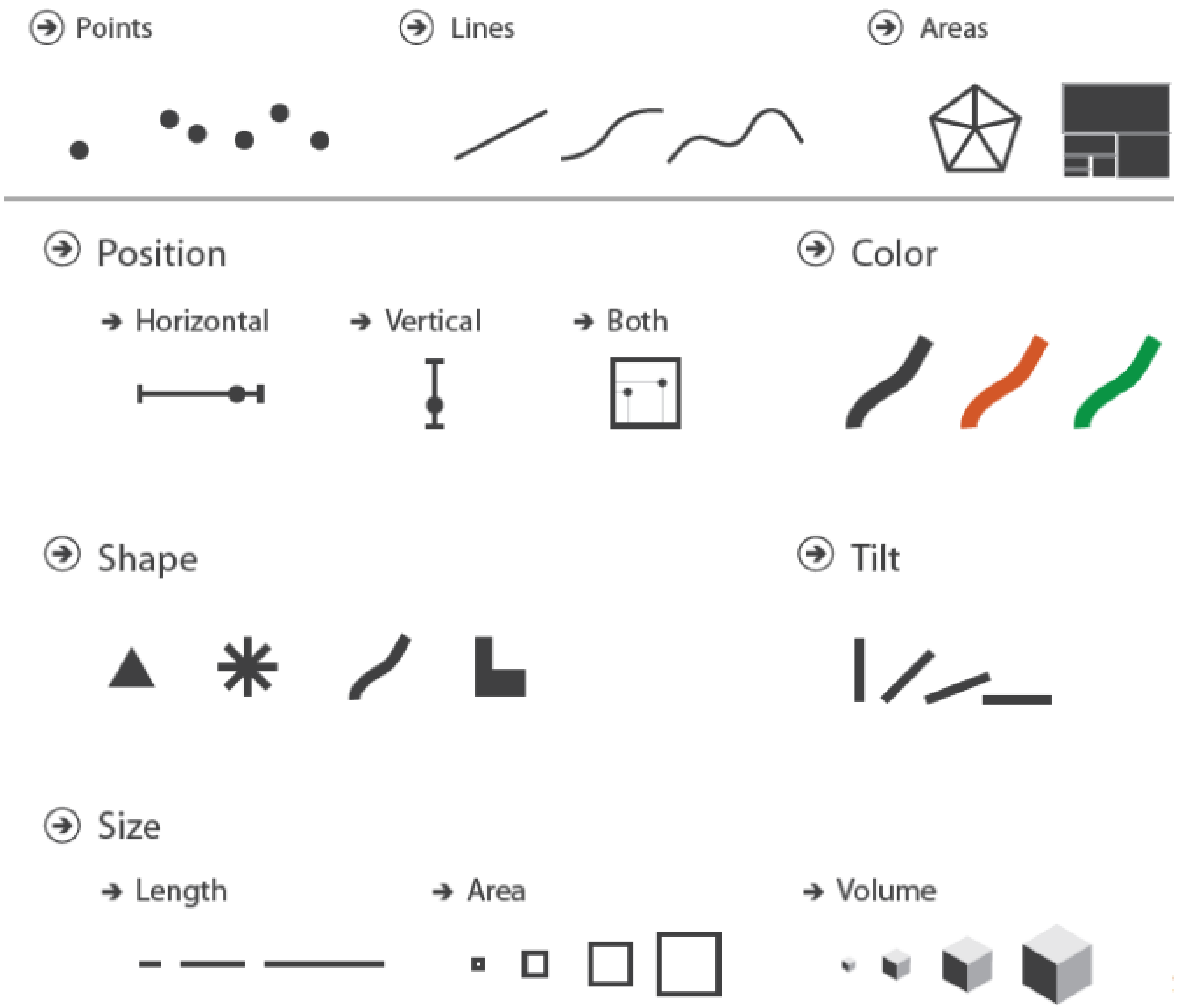
Composition

- Grid and alignments
- Balance
- Hierarchy and focus

Elements: Marks and channels

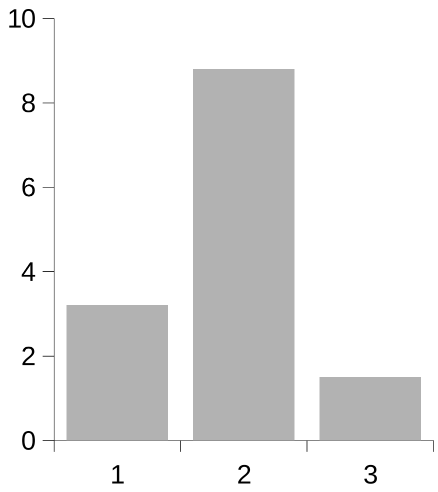
Marks (geometric primitives): used to **represent** data

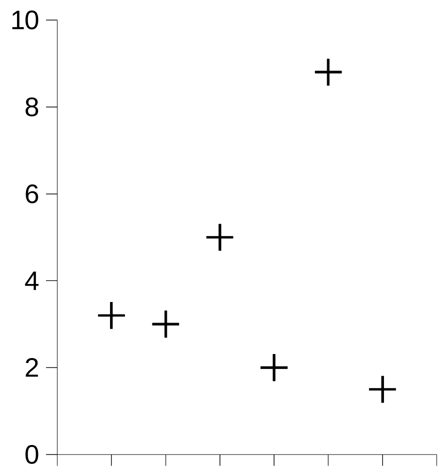
Channels control the graphical appearance of marks: used to **encode** data, can be combined

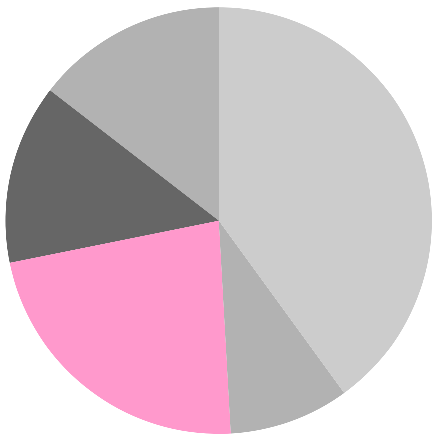


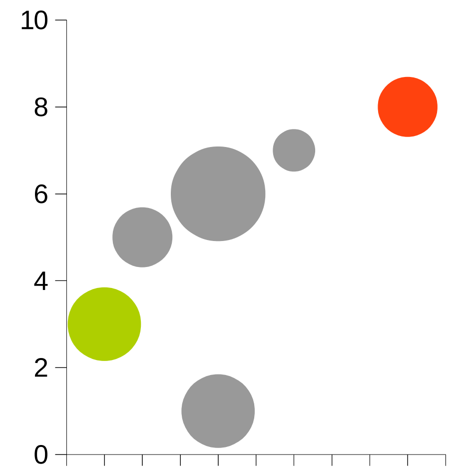
Images from Munzner

Figures are a combination of marks and channels

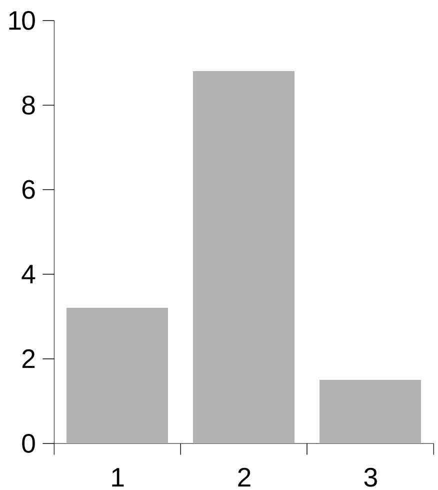




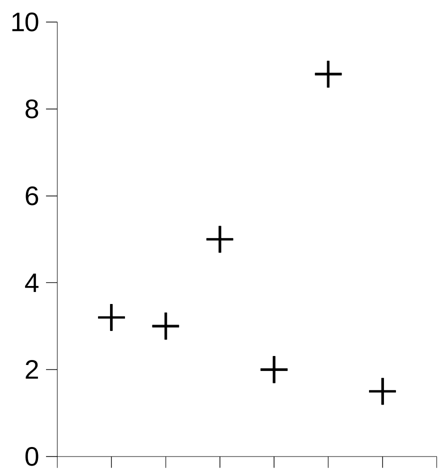




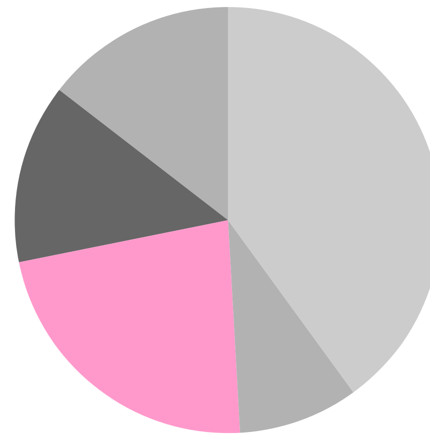
Figures are a combination of marks and channels



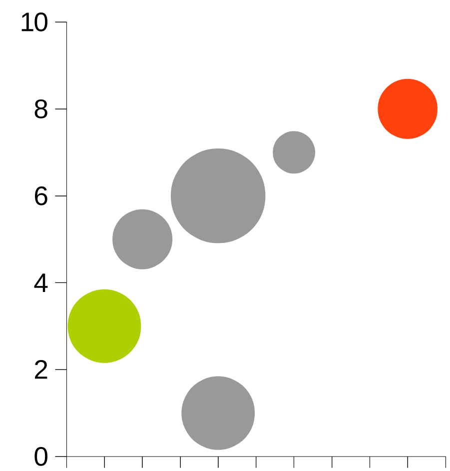
1 Mark =
Rectangle
1 Channel =
Length of
longest side



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



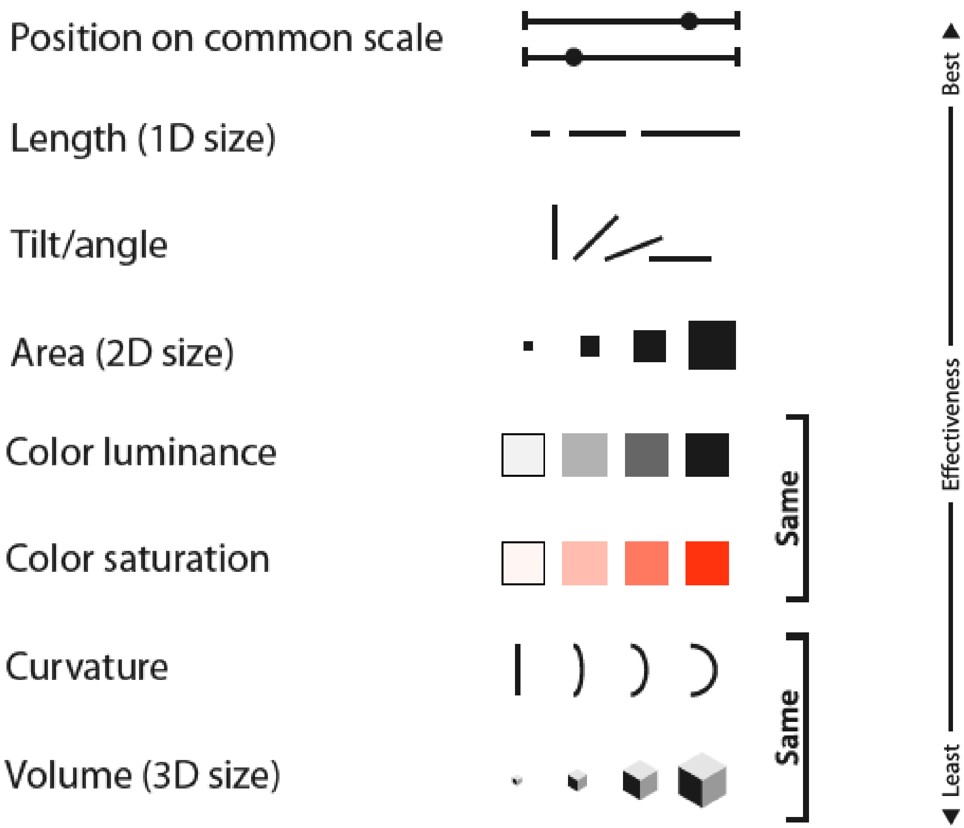
1 Mark =
Circle segment
2 Channel =
Angle
Colour



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

Types of channel

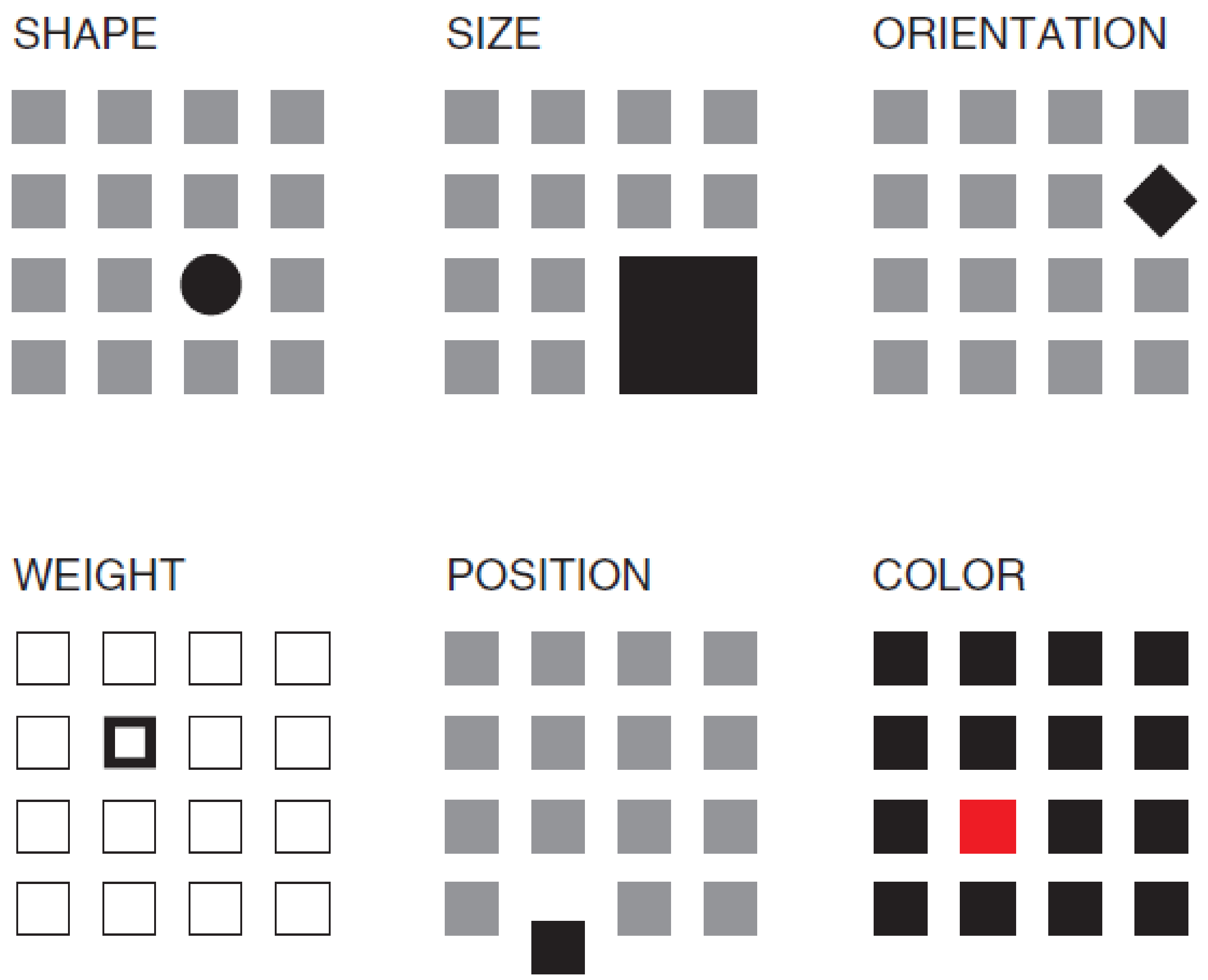
Identity channels: categorical/ **qualitative** attributes



Magnitude channels: ordered/ **quantitative** attributes

Images from Munzner

Types of channel (continued)



Rolandi et al 2011

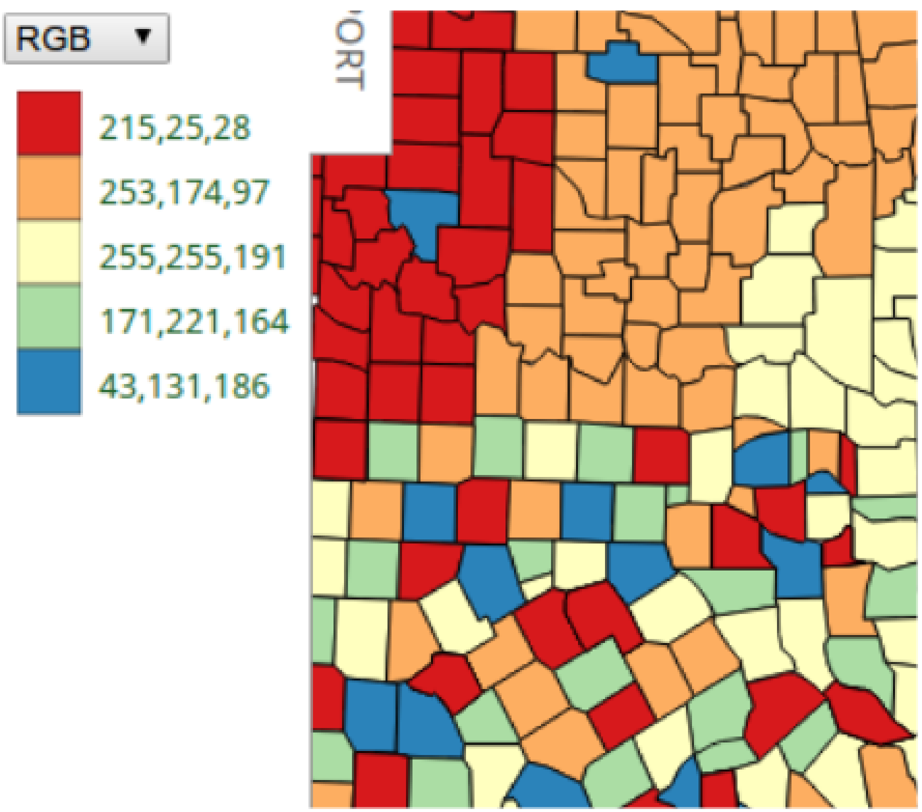
More principles

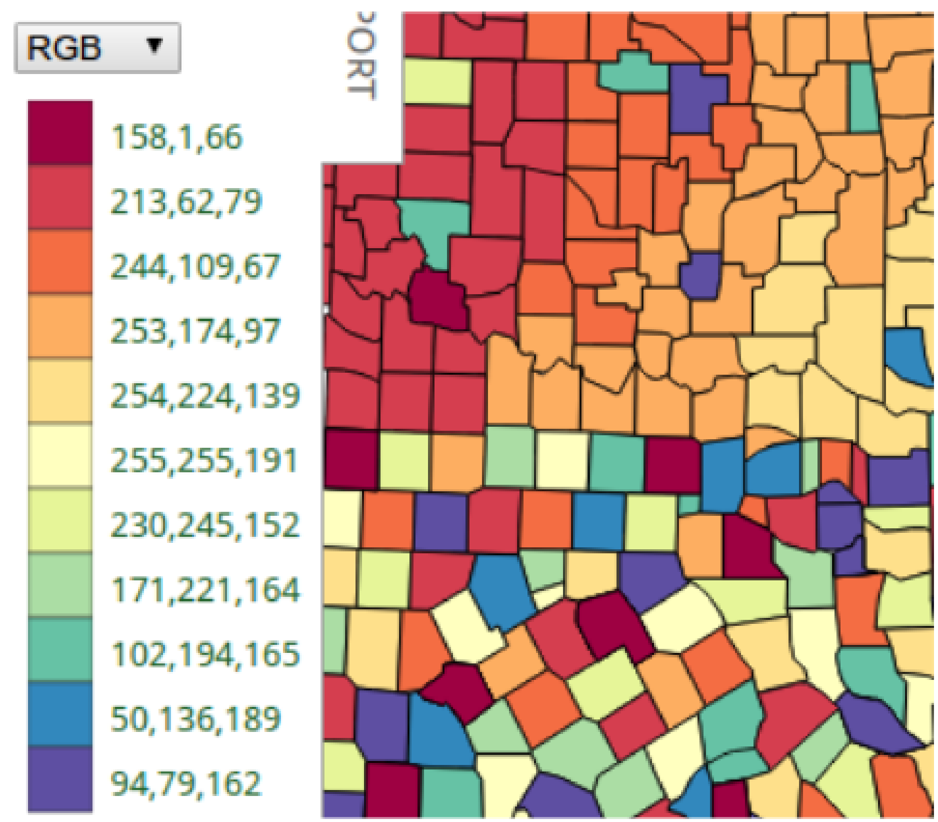
- **Effectiveness**: encode the most important information with the most effective channel
- **Expressiveness**: match the properties of the data and channel
  - i.e. heed whether the data are **quantitative**, **ordered** or **categorical**, and choose accordingly

More principles

Discriminability and **separability**

How many different types can you distinguish?
How easy is it to distinguish them?

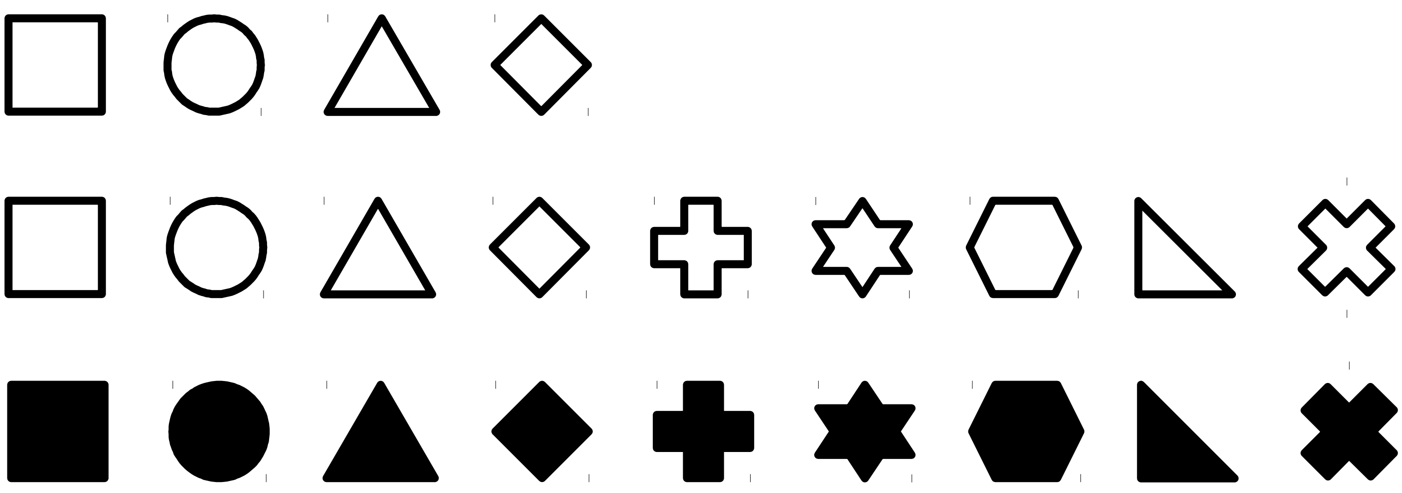




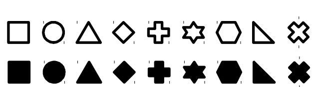
http://colorbrewer2.org/

Qualitative discrimination

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

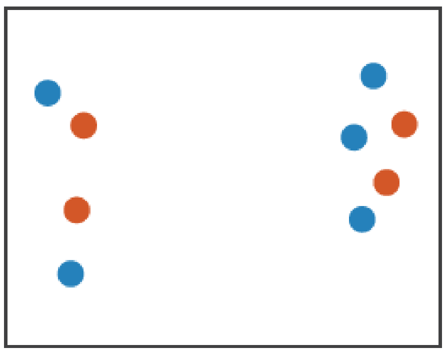


and they have to be distinguishable at small sizes



Separability

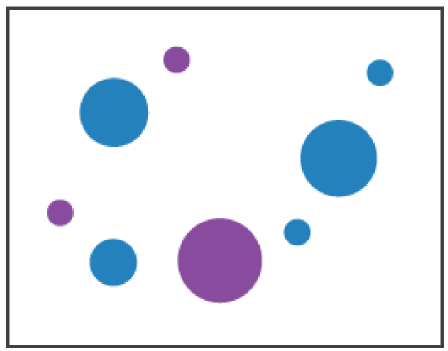
Position
+ Hue (Color)



Fully separable

2 groups each

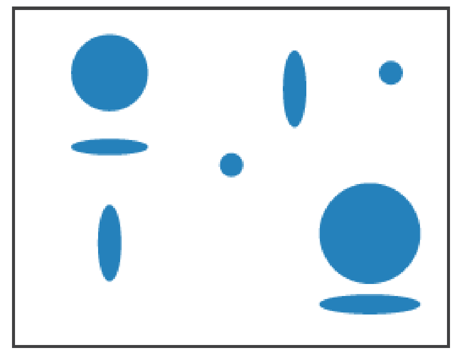
Size
+ Hue (Color)



Some interference

2 groups each

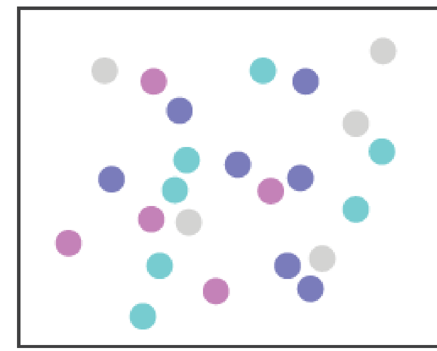
Width
+ Height



Some/significant interference

3 groups total: integral area

Red
+ Green



Major interference

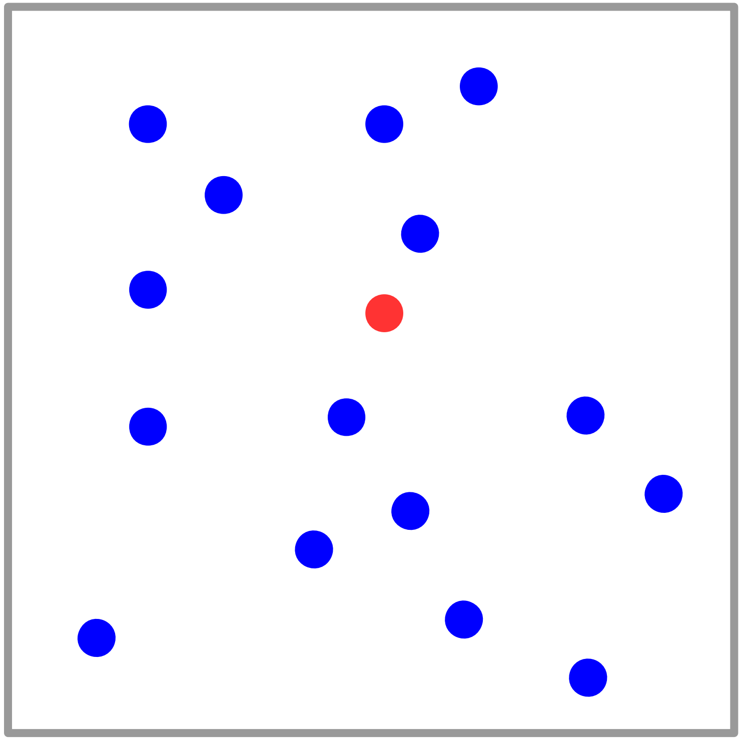
4 groups total: integral hue

Images from Munzner

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

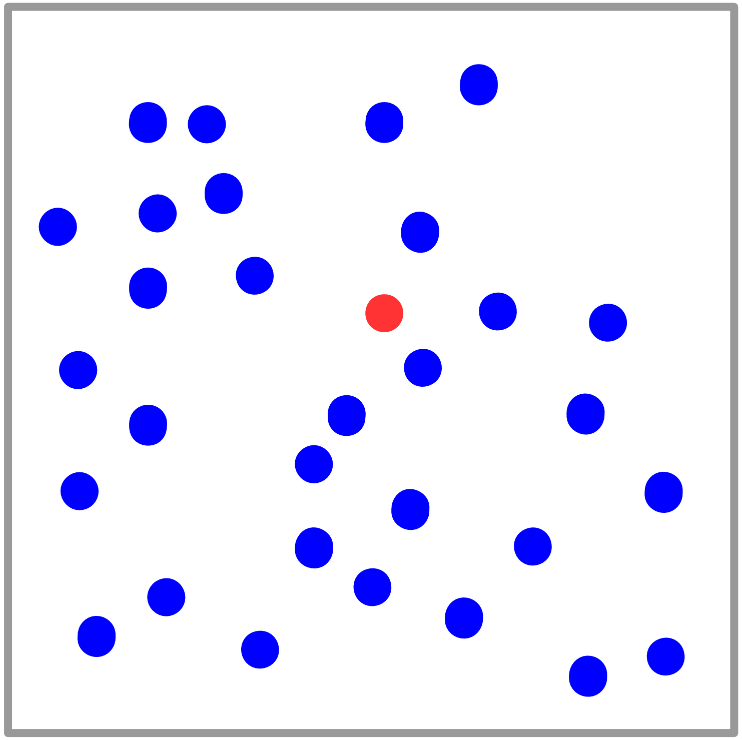
Popout



Find the red dot: how long does it take?

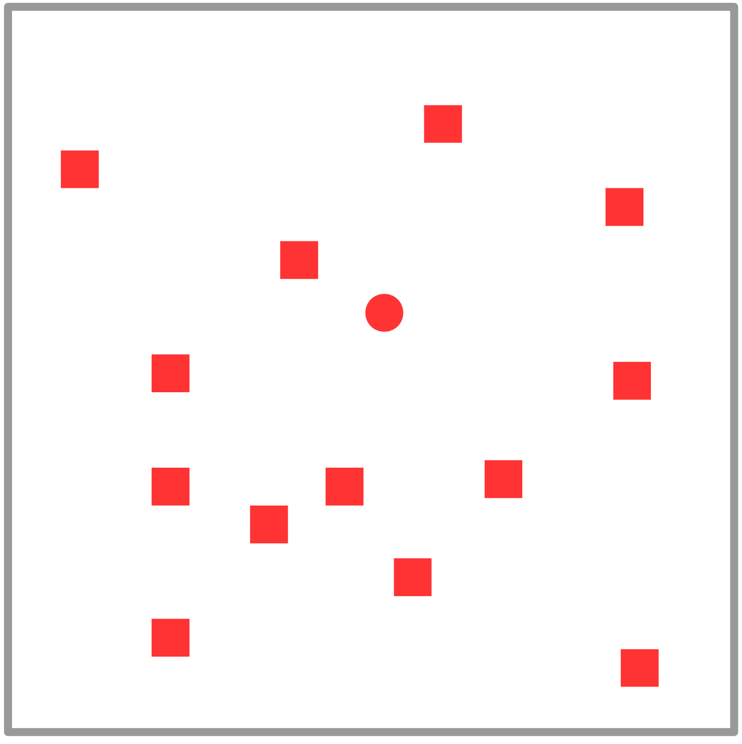
Example based on Munzner
and Babraham Bioinformatics

Popout



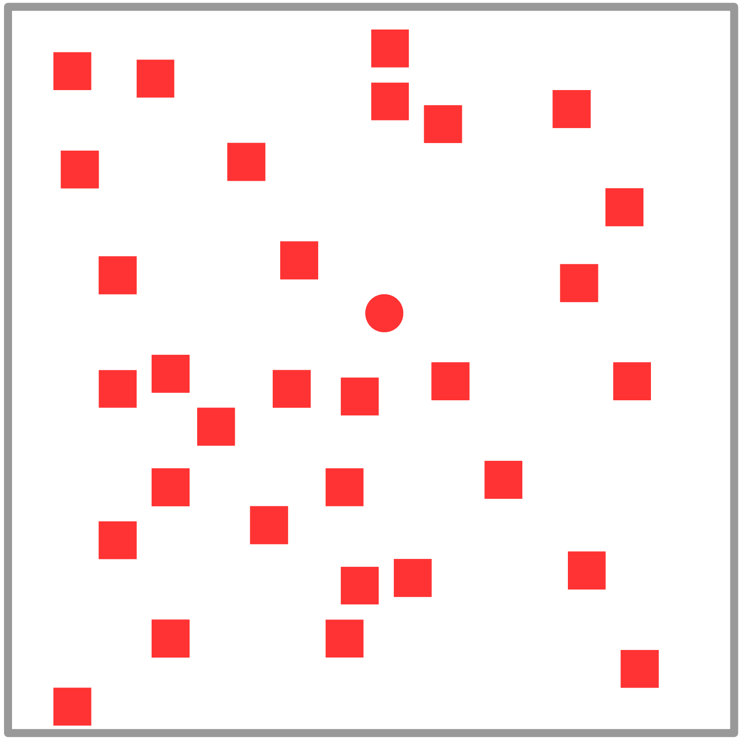
The speed of identification is independent of the count of distractors

Popout



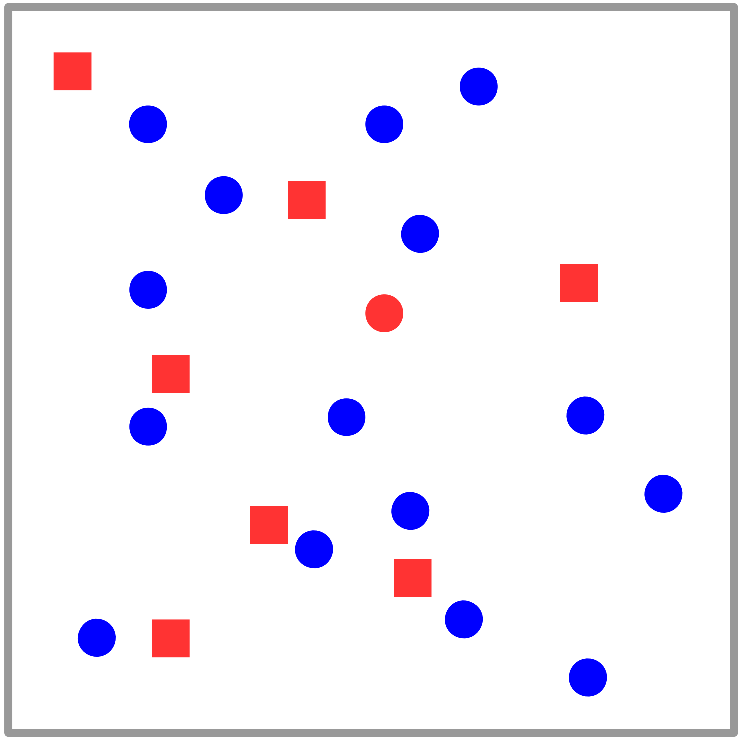
Find the circle

Popout



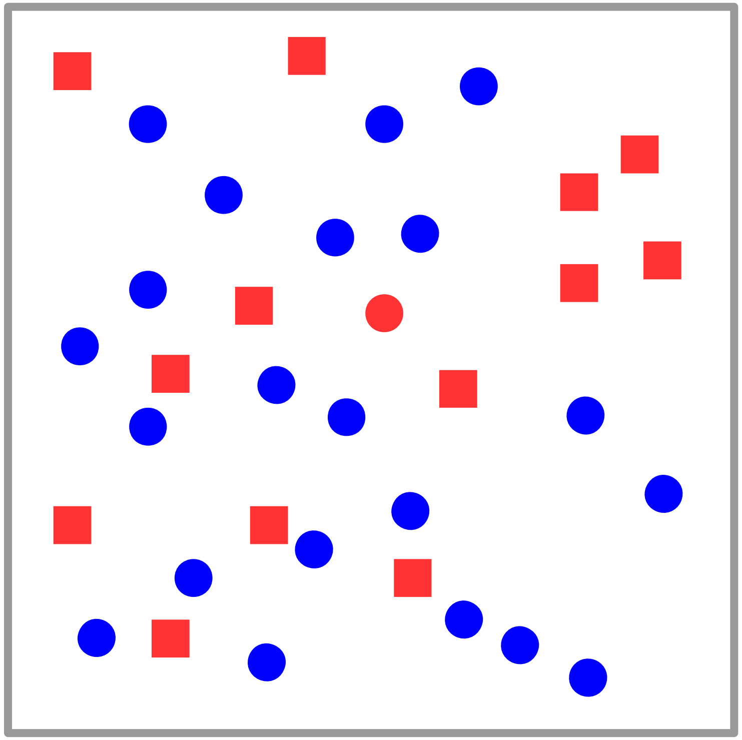
Colour stands out more than shape

Popout



Find the red dot

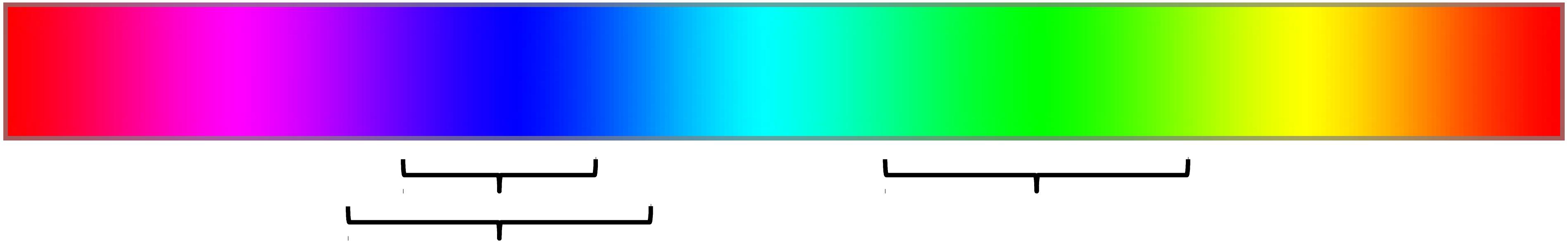
Popout



Mixing channels removes the effect

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

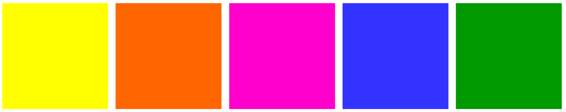


Slide from Babraham Bioinformatics

Example based on Carter

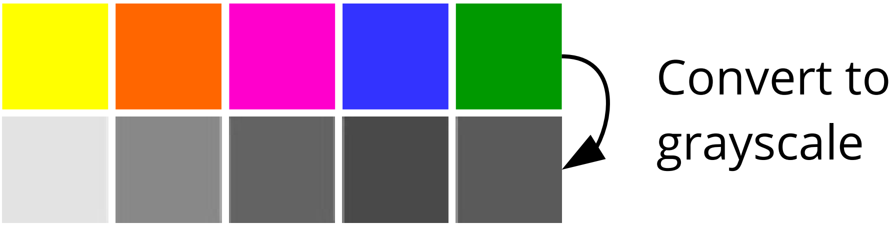
Characteristics of a colour

- **<u>Hue</u>**: the *actual* colour (qualitative)



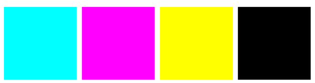
- **<u>Saturation</u>**: the intensity of the hue (quantitative)

- **<u>Value</u>**: the lightness/ darkness of a colour (quantitative; useful to know how a colour will behave when transformed to grayscale)

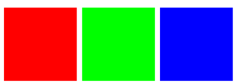


- ***Lightness***
  - ***Shade:*** the amount of black
  - ***Tint:*** the amount of white

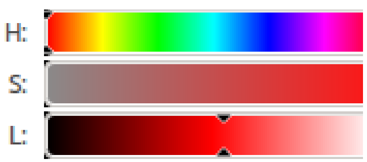
Colour: How computers identify colours



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

	#ff00ff	rgb(255, 0, 255)	hsl(300, 100%, 50%)
	#ff0000	rgb(255, 0, 0)	hsl(0, 100%, 50%)

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

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.

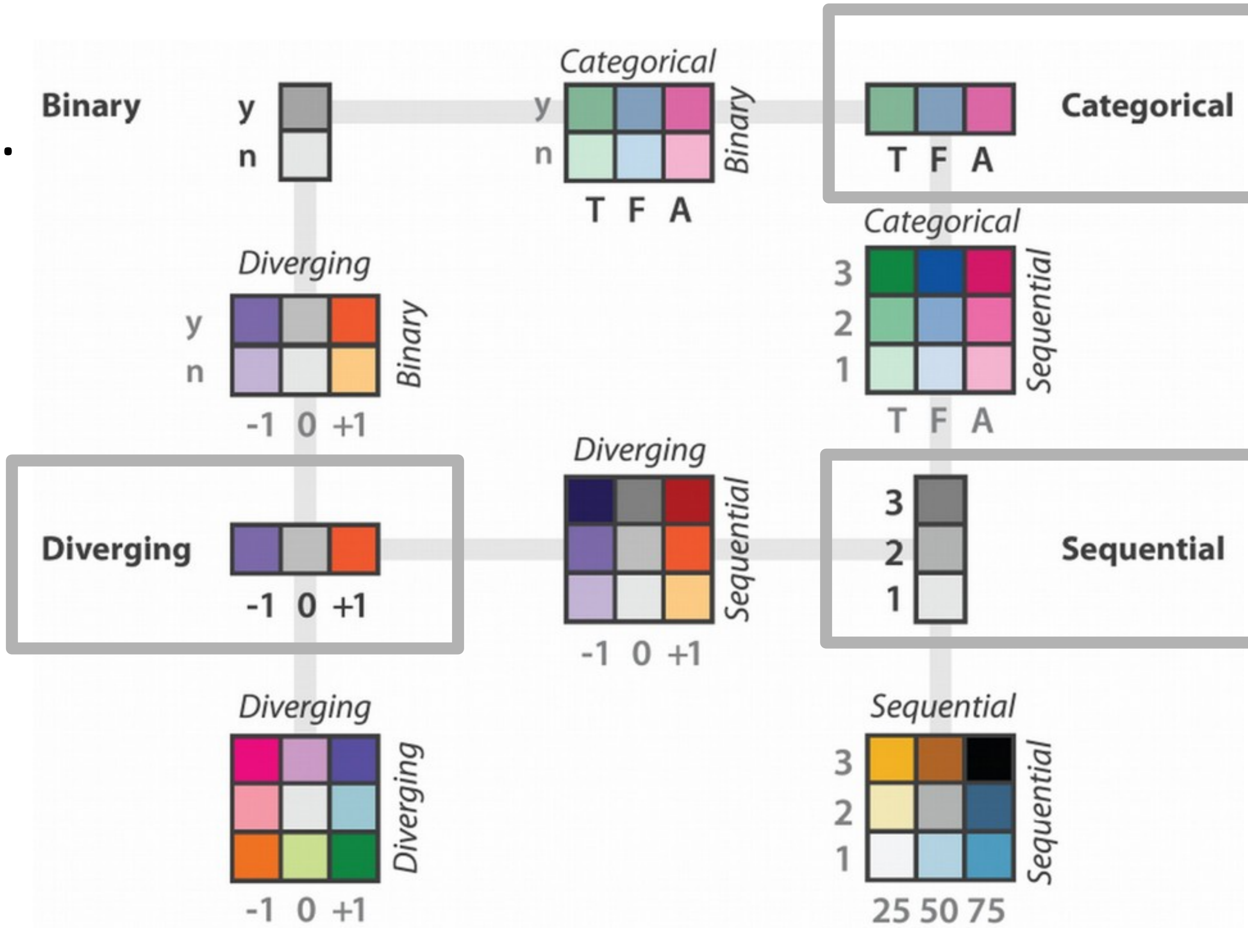
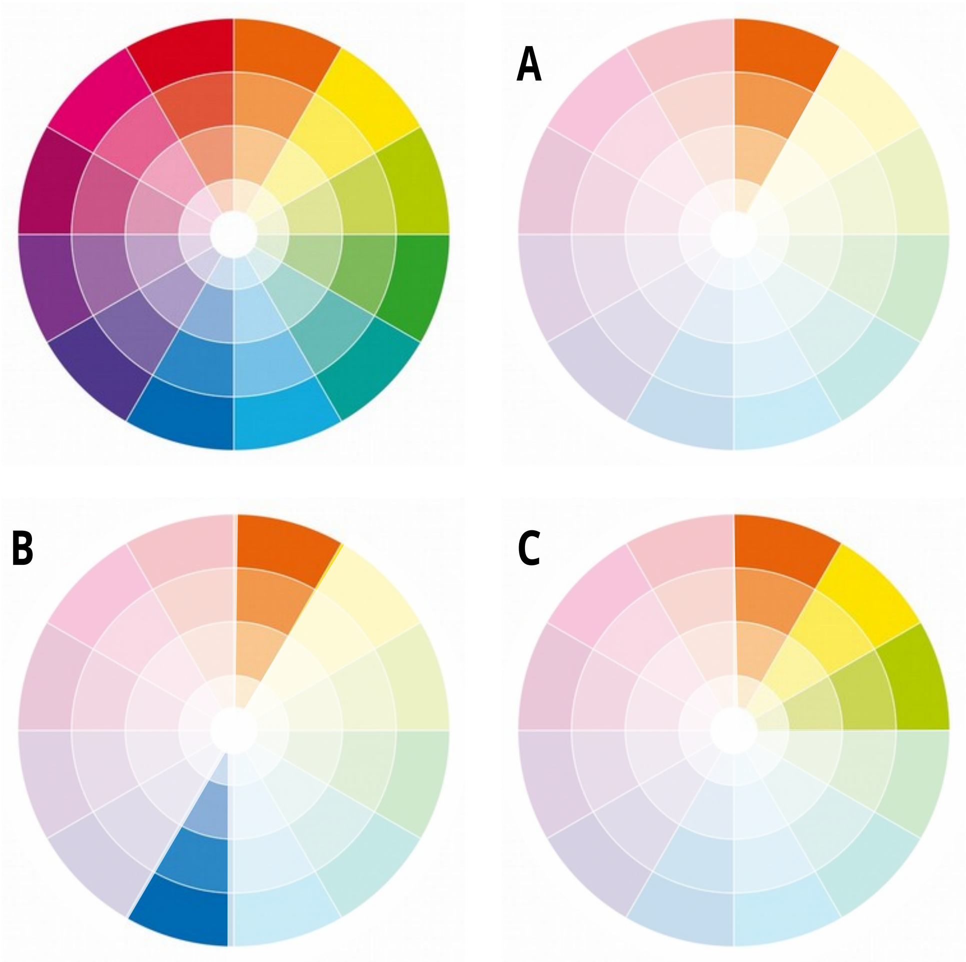


Image from Munzner after "Color use guidelines for mapping and visualization. Brewer 1994"

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/



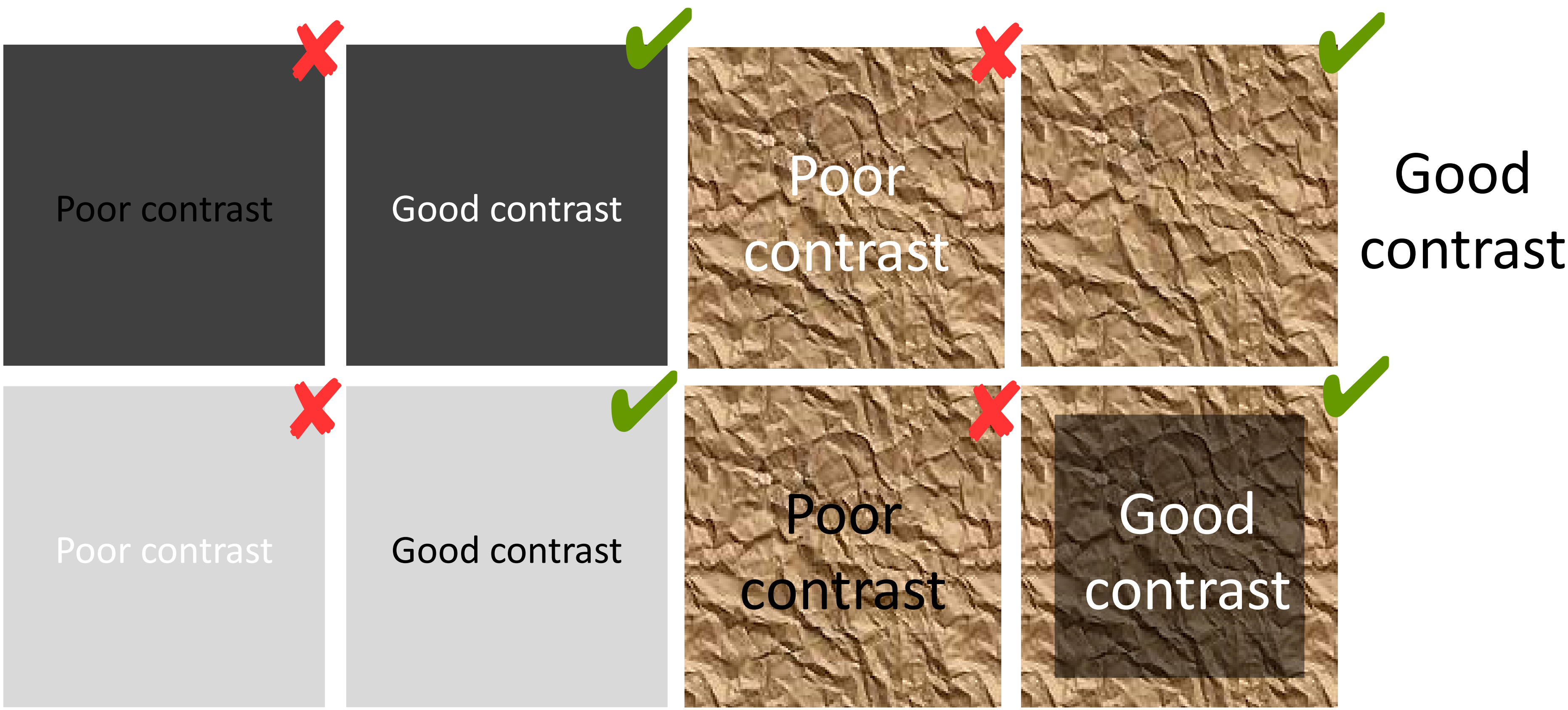
Colour: Choosing palettes

- 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

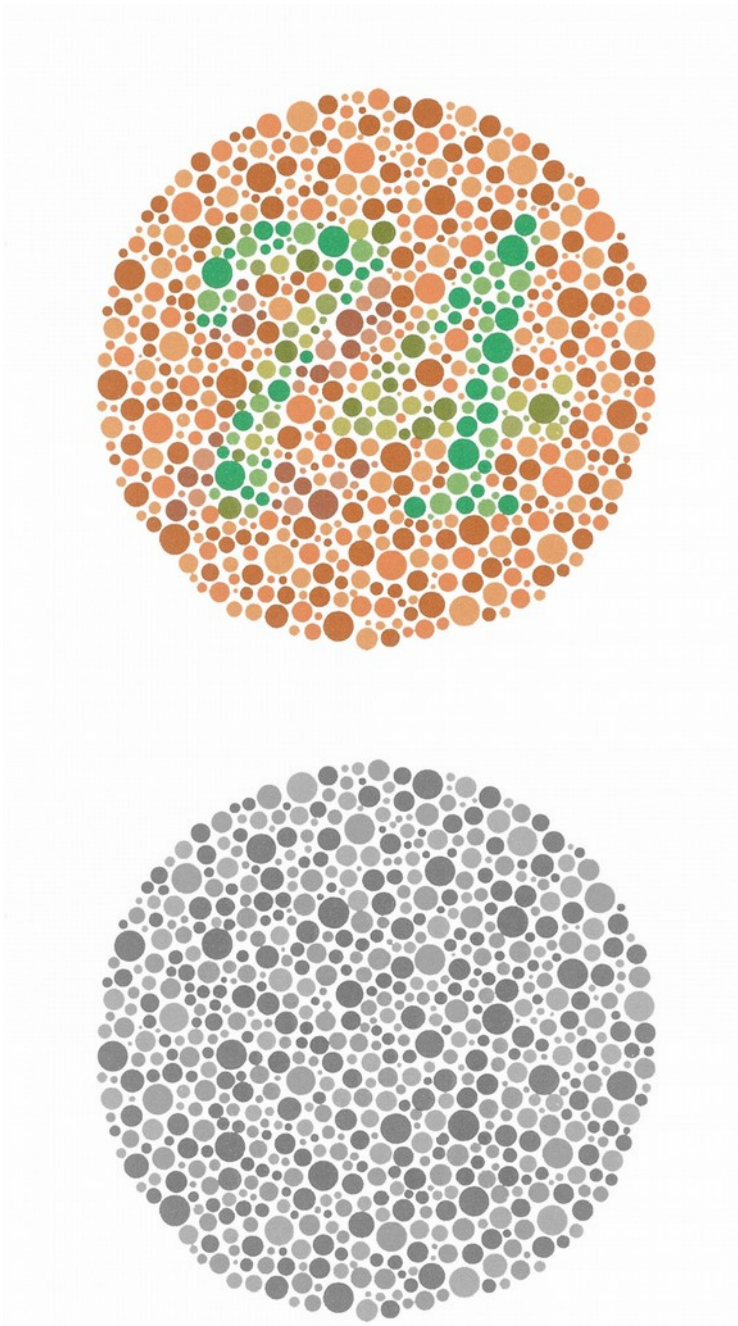
Beware of patterned backgrounds



Slide from Babraham Bioinformatics

Colour Blindness

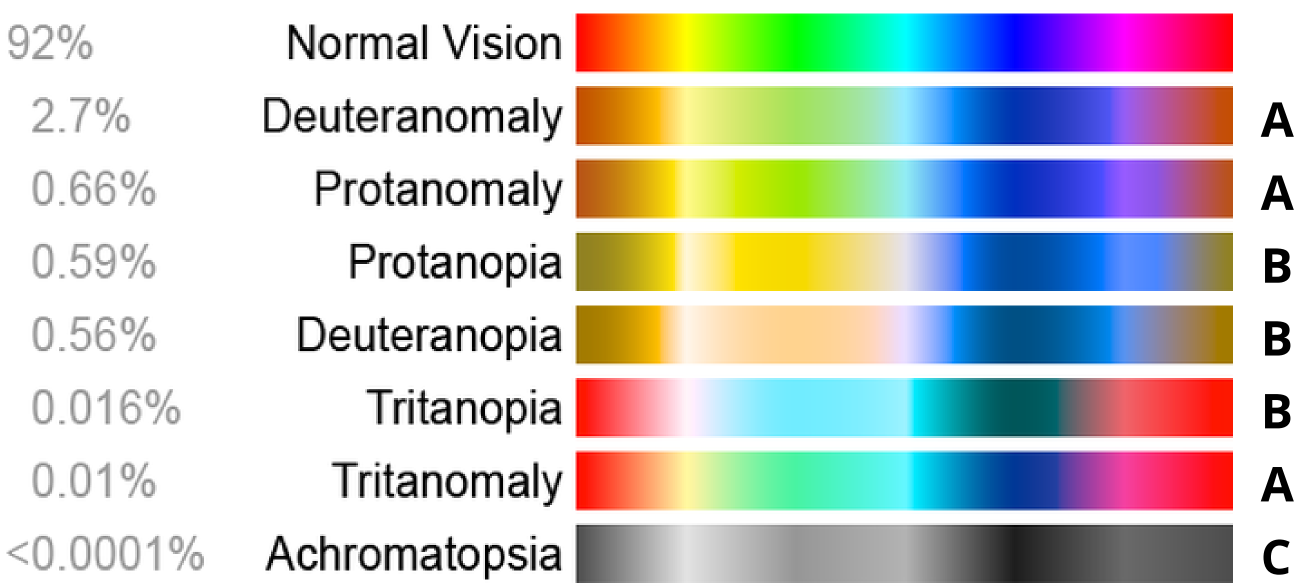
- Affects 1:12 men and 1:200 women worldwide
- ***"If a submitted manuscript happens to go to three male reviewers of Northern European descent, the chance that at least one will be colour blind is 22 percent."***



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



http://colorsscience.weebly.com/color-blindness.html

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:

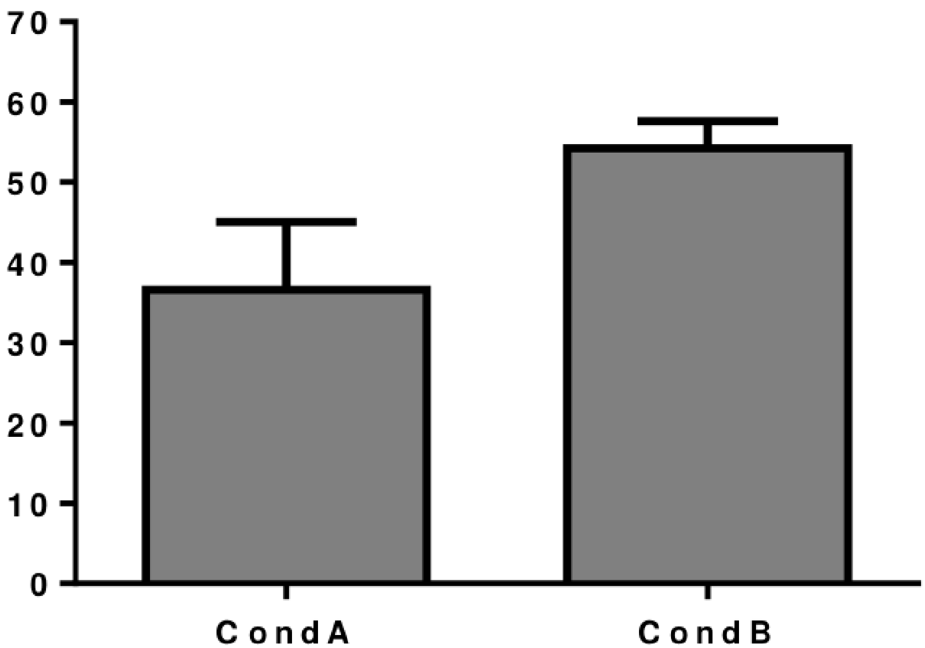
```
# install.packages("RColorBrewer")
library(RColorBrewer)
# check out the available palettes
display.brewer.all(n=NULL, type="all",
                   select=NULL, exact.n=TRUE)
brewer.pal(5, "Set1")
```

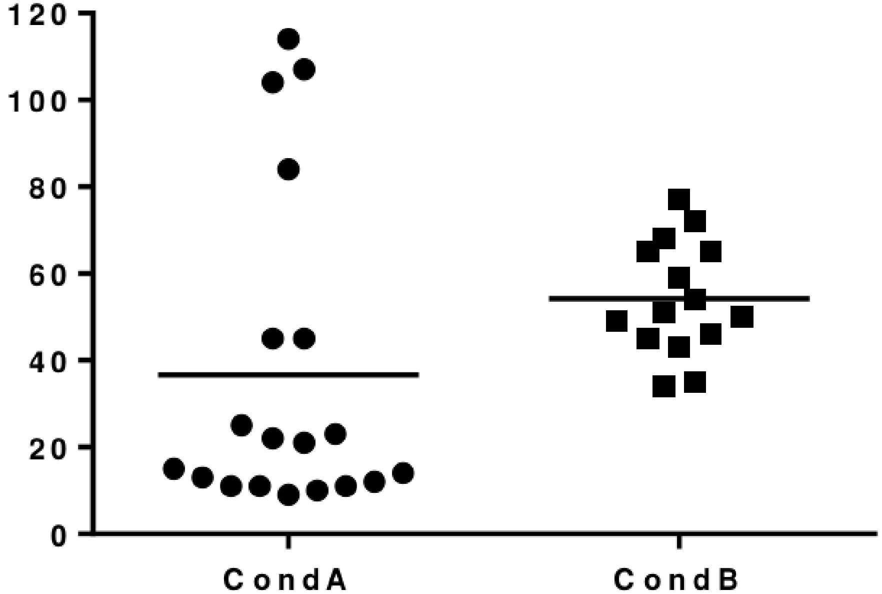
Ethics of data representation

- The figure/graph/image should show **what is actually happening** and **not what you want to happen**
- Different ways of being unethical:
  - Not exploring/getting to know the data well enough
  - Misusing your chosen graphical representation
  - Deliberately showing the data in a misleading manner
  - Choosing the 'most representative' image/experiment

Ethics section adapted from Babraham Bioinformatics

Not exploring/getting to know the data well enough - Example 1

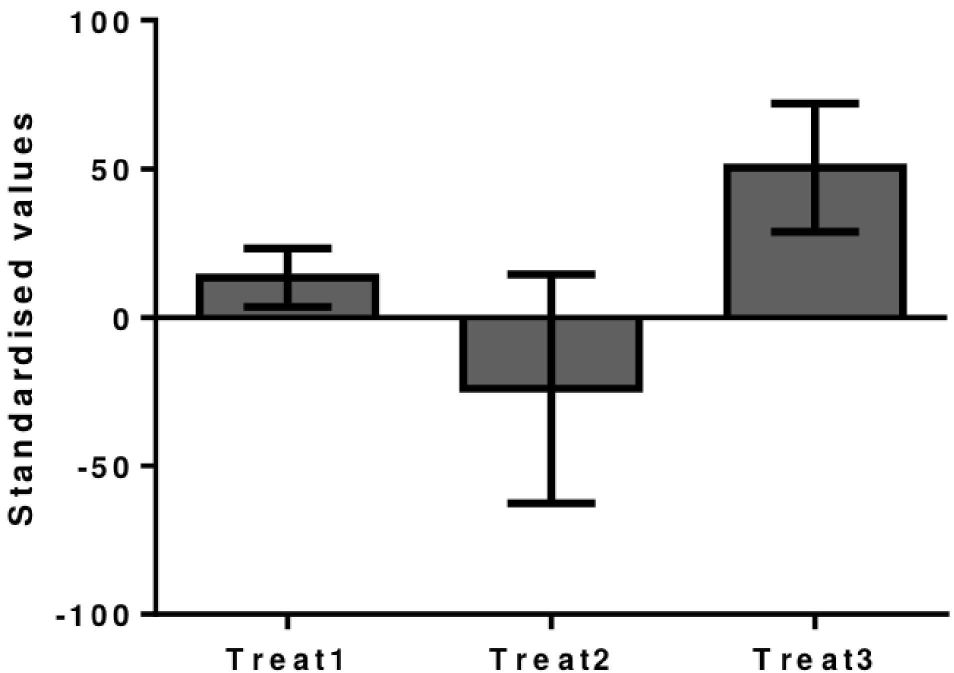


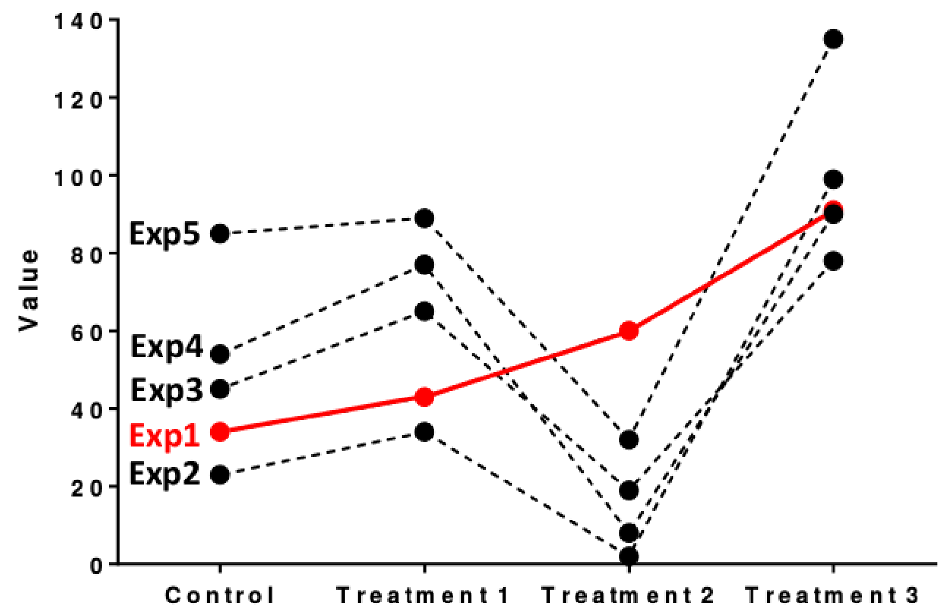


One experiment: change in the variable of interest between CondA to CondB.

✘ Data plotted as a bar chart.

Not exploring/getting to know the data well enough - Example 2

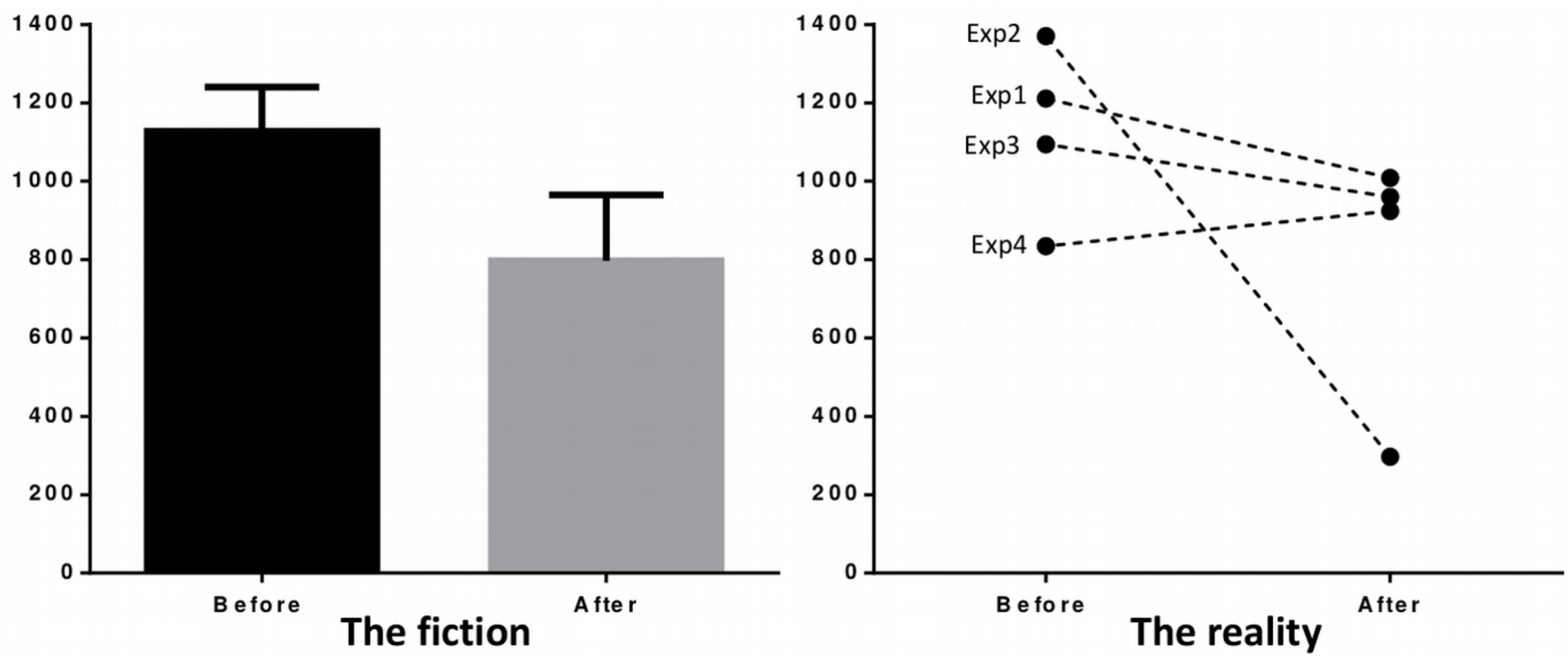




Five experiments: change in the variable of interest between 3 treatments and a control.

✘ Data plotted as a bar chart.

Choosing the wrong graph to present

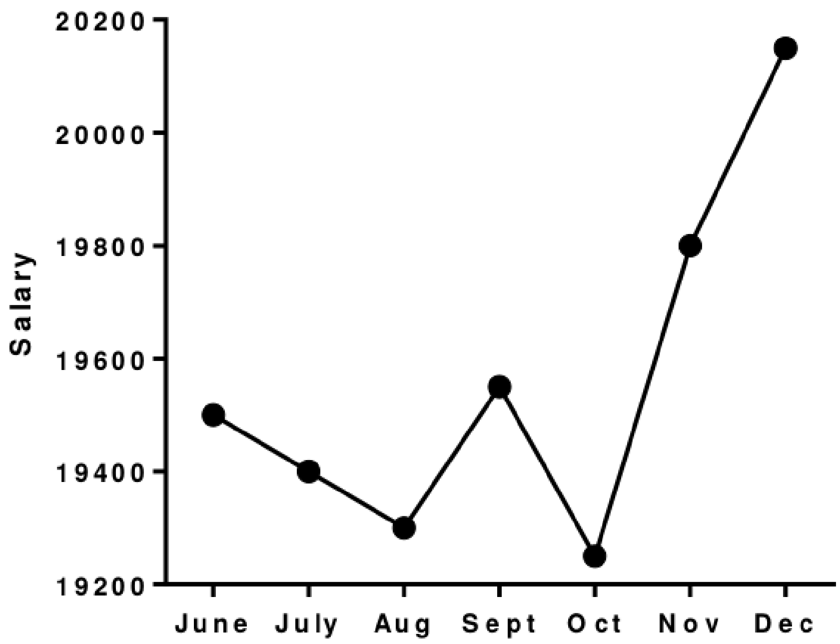


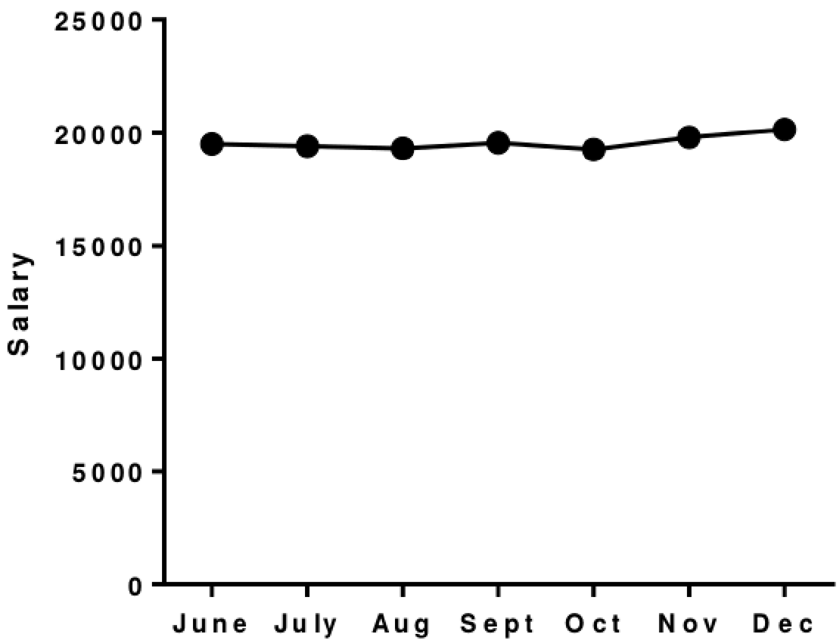
Four experiments: Before-After treatment effect on a variable of interest. **Hypothesis**: Applying a treatment will decrease the levels of the variable of interest.

✘ Data plotted as a bar chart.

Choosing the wrong axis/scale

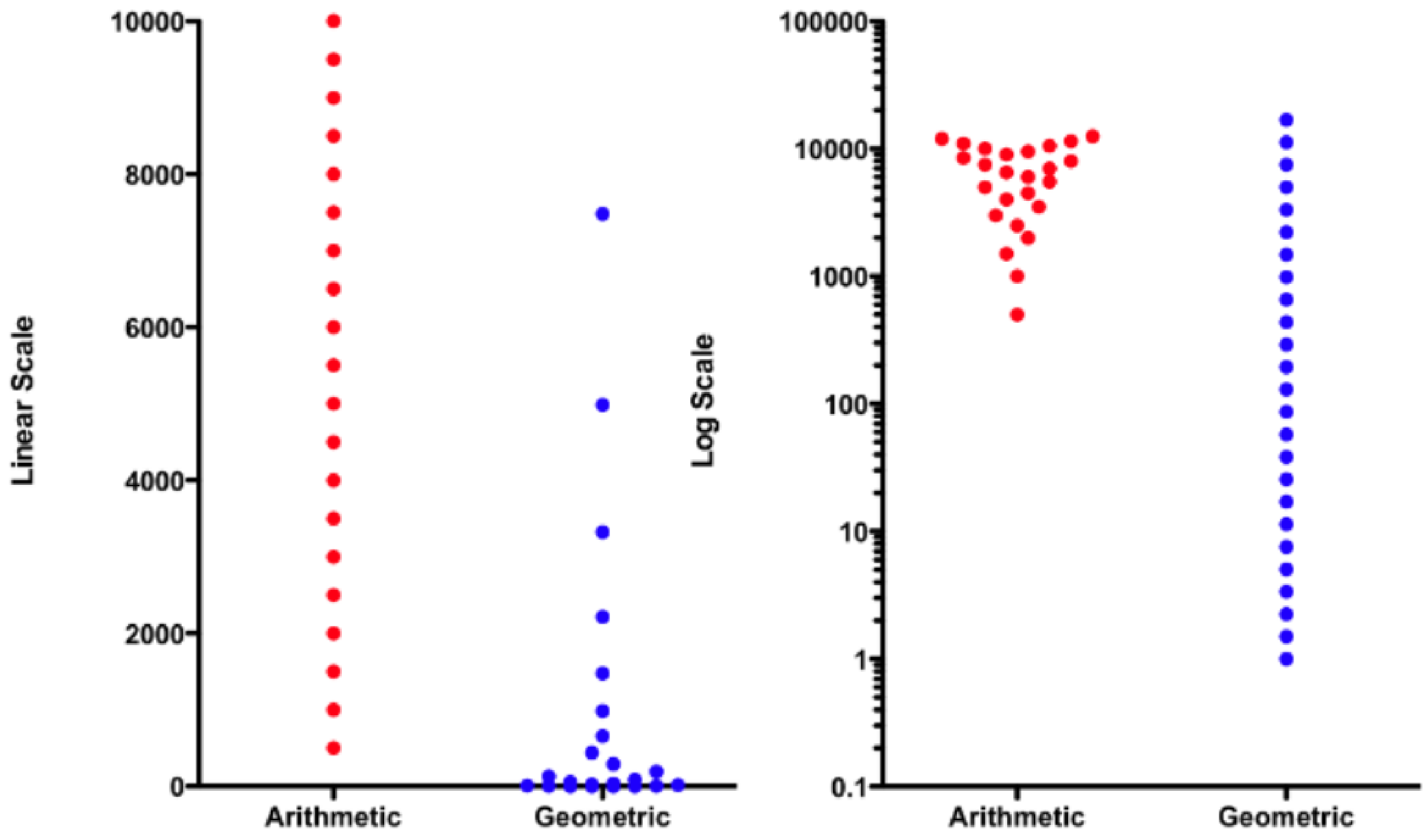
Example: increase in salary in the last term





Choosing the y-axis/scale

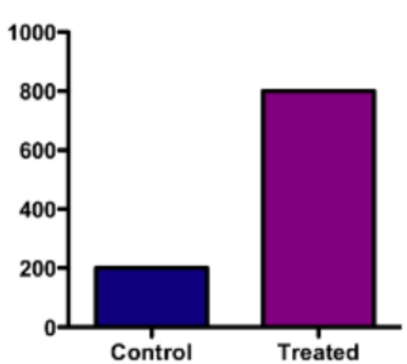
Be careful with Linear vs. logarithmic scale.



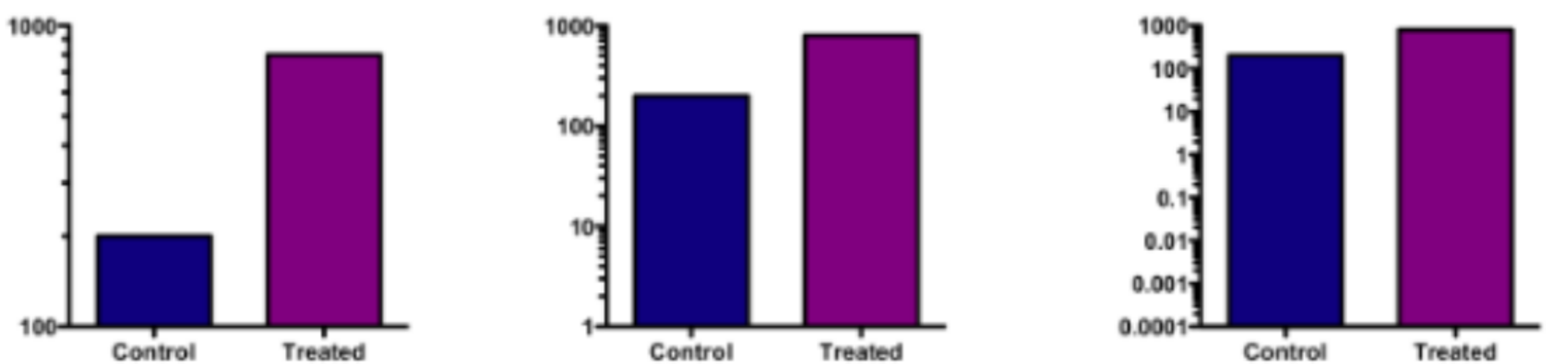
Choosing the y-axis/scale

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.

Linear

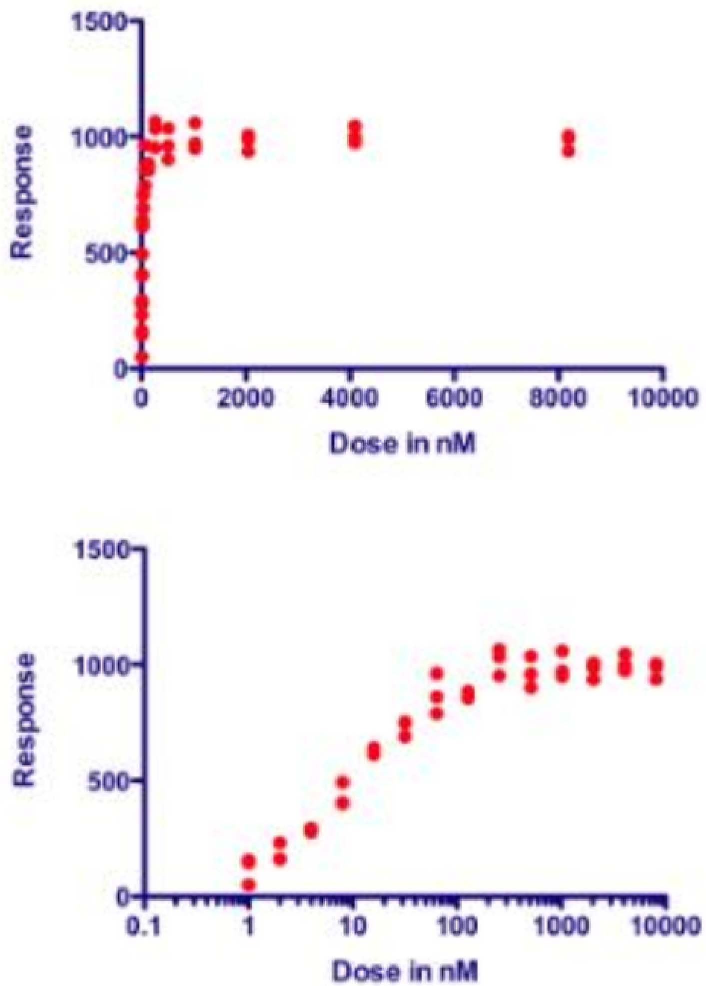


Logarithmic

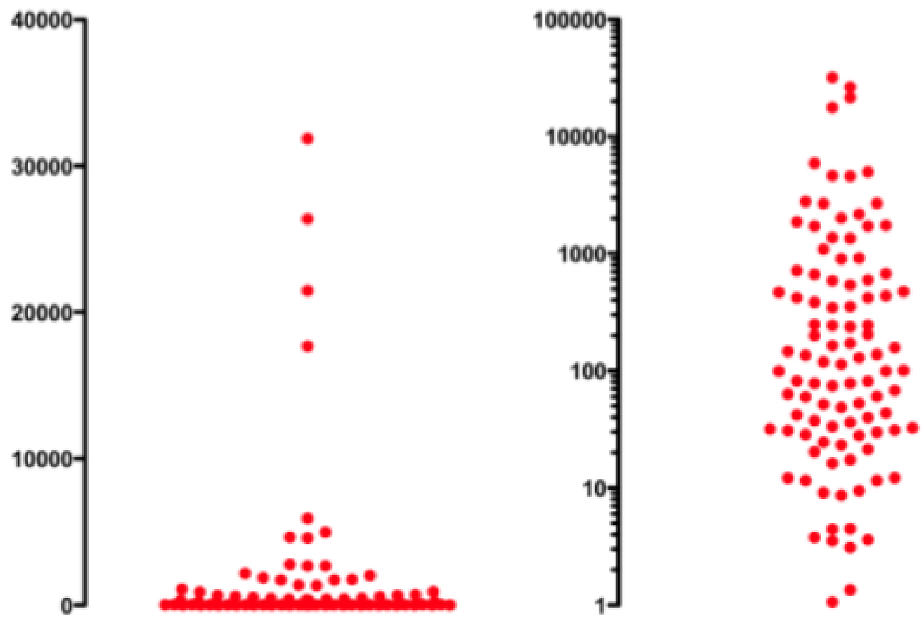


Logarithmic axis should be used for:

Logarithmically spaced values



Lognormal data

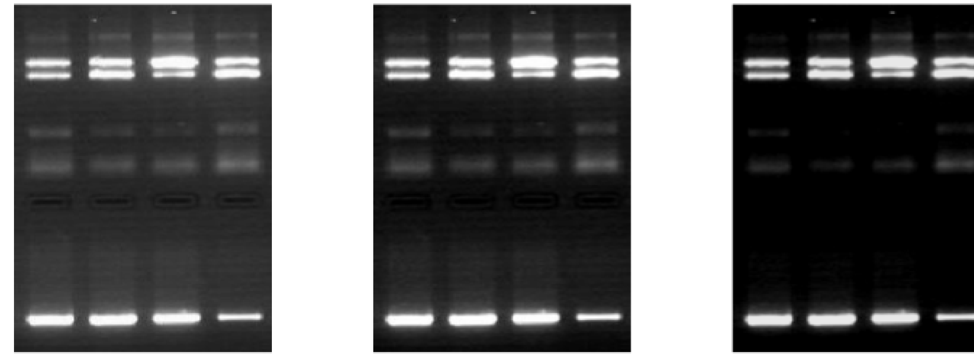


Simply Cheating

Manipulating images: Western blot

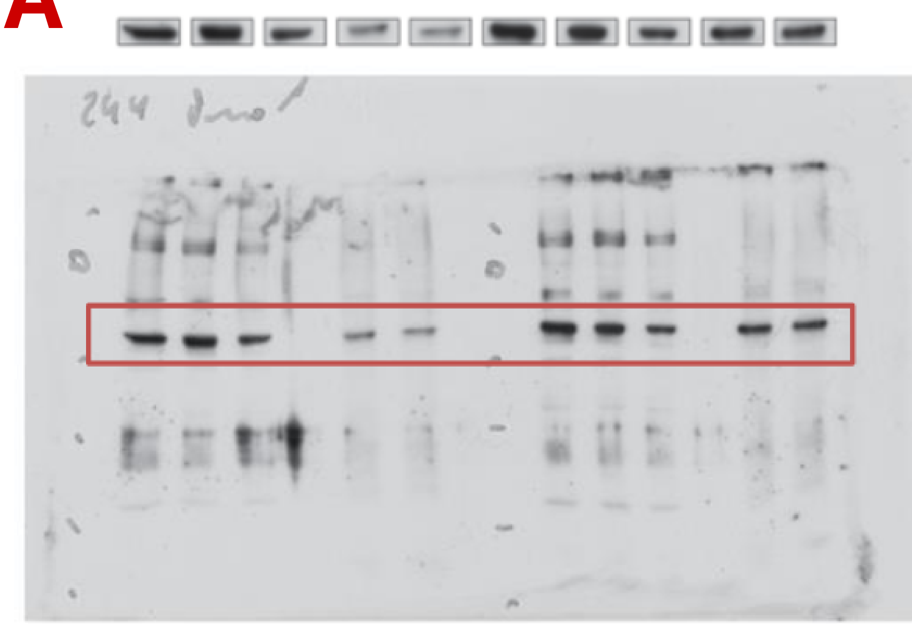
A) Presenting bands out of context
B) 'Playing' too much with contrast
C) 'Rebuilding' a Western blot from several cuts

B

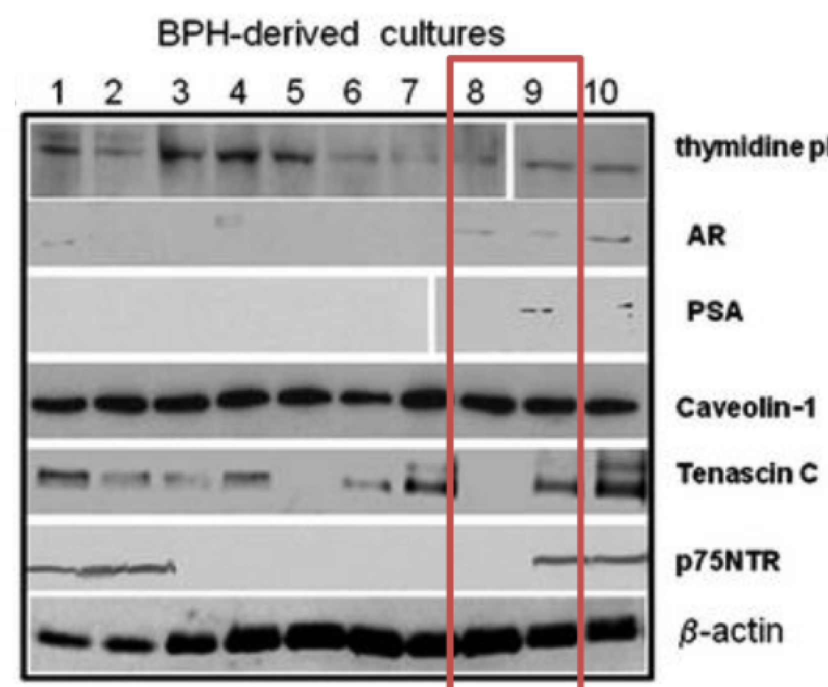


Original | Brightness and Contrast Adjusted | Brightness and Contrast Adjusted Too Much: Oversaturation

A



C



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?

Practical

GIMP

For editing bitmap images (photos)