

The theory of data visualisation

V2017-10

Simon Andrews, Phil Ewels

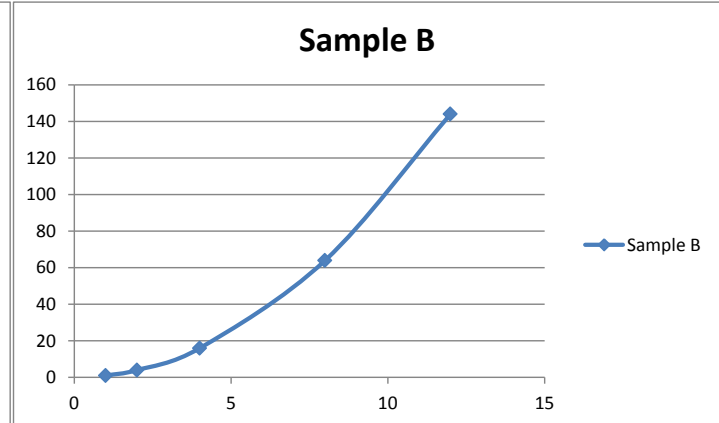
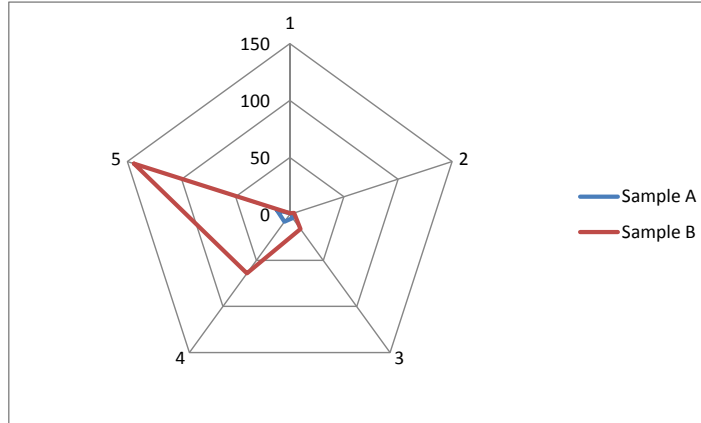
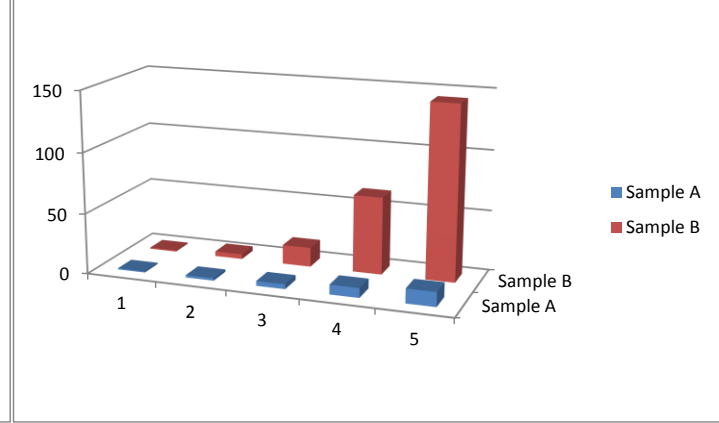
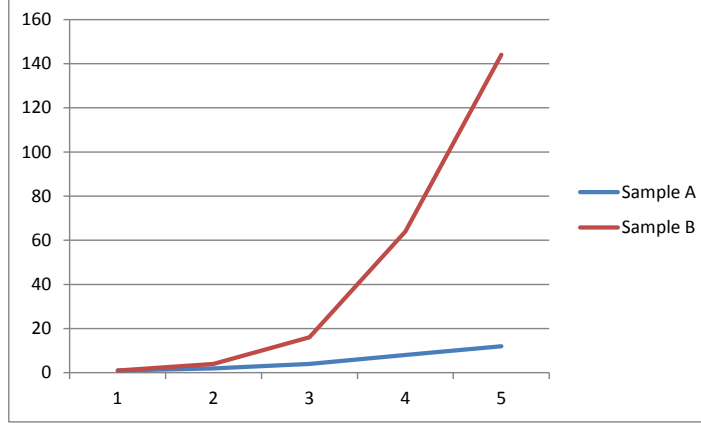
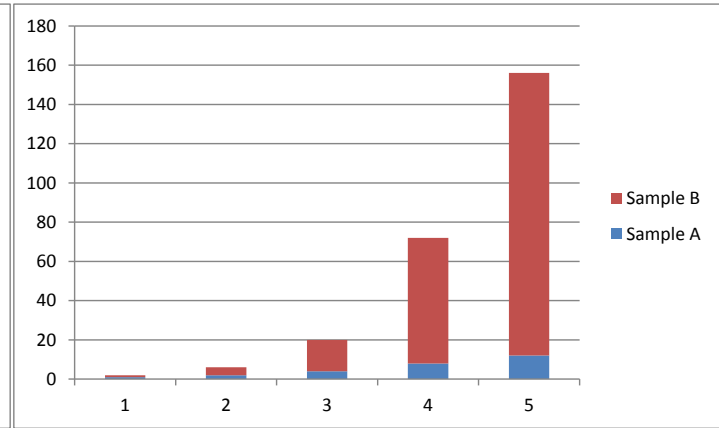
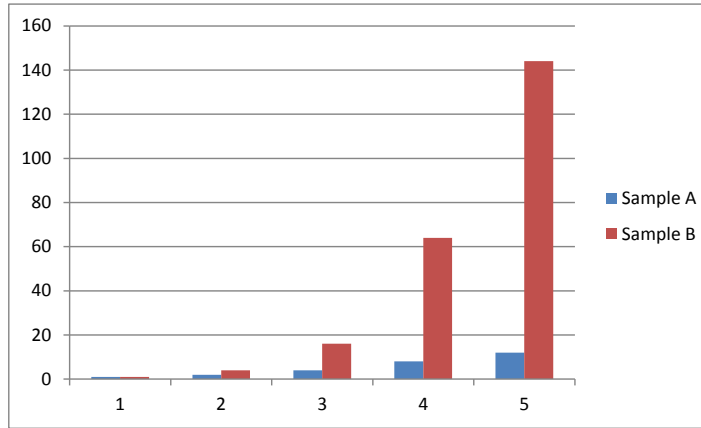
simon.andrews@babraham.ac.uk

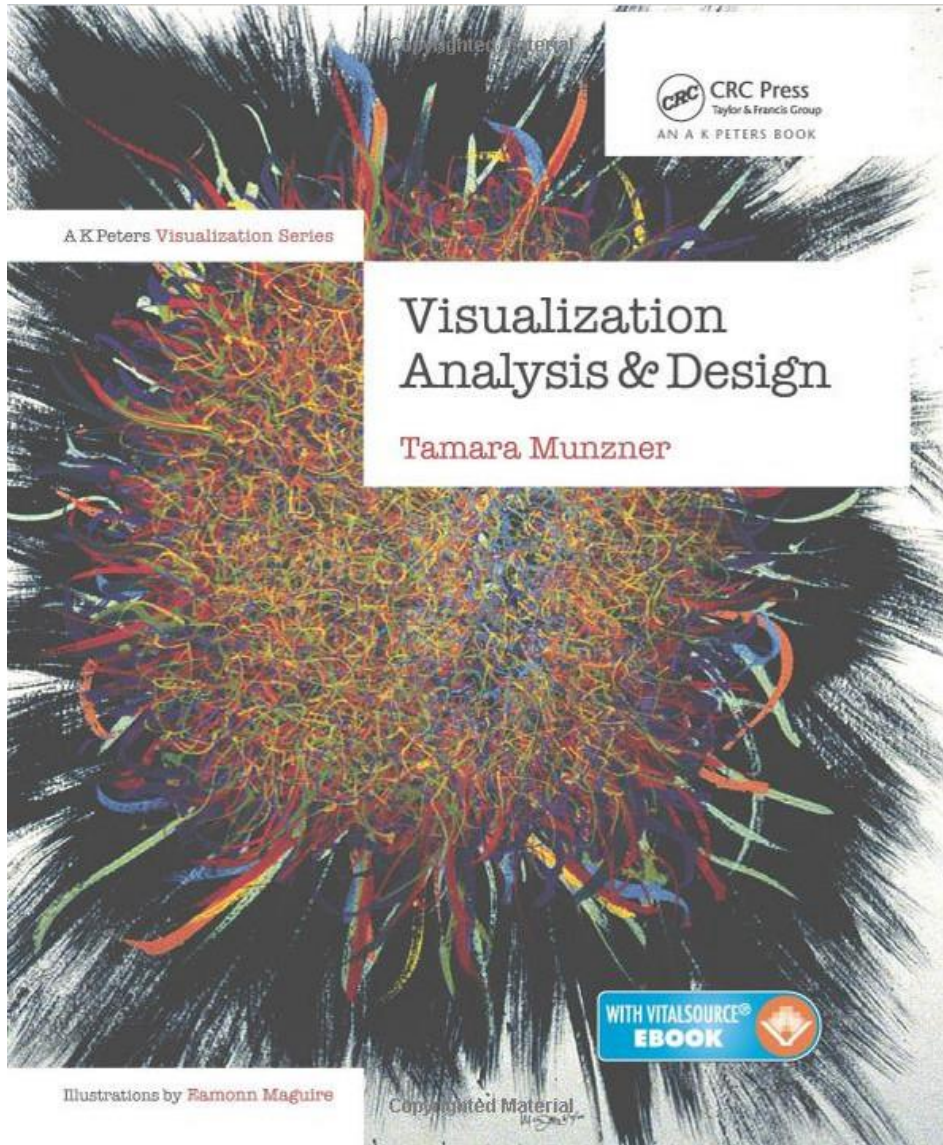
phil.ewels@scilifelab.se

Data Visualisation

- A scientific discipline involving the creation and study of the visual representation of data whose goal is to communicate information clearly and efficiently to users.
- Data Visualisation is both an art and a science.

Sample A	Sample B
1	1
2	4
4	16
8	64
12	144

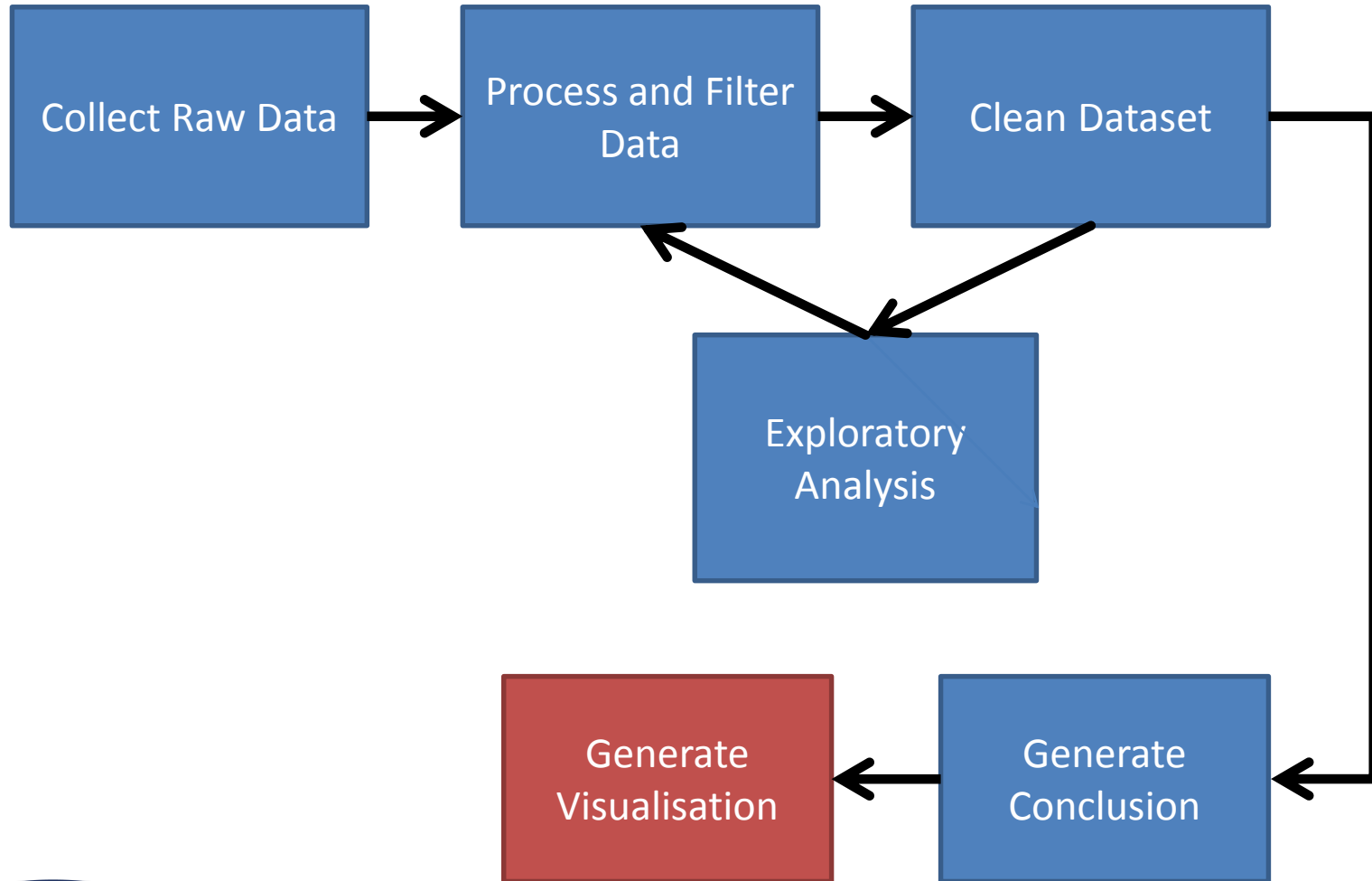




ISBN-10: 1466508914

<http://www.cs.ubc.ca/~tmm/talks.html>

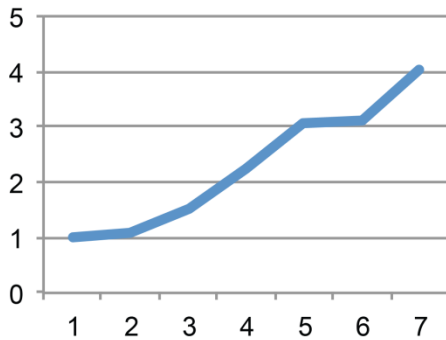
Data Viz Process



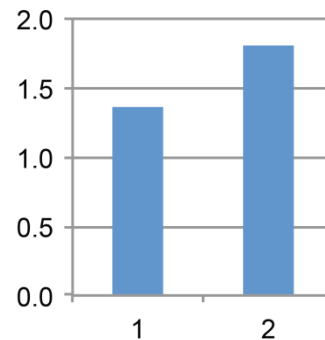
A data visualisation should...

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

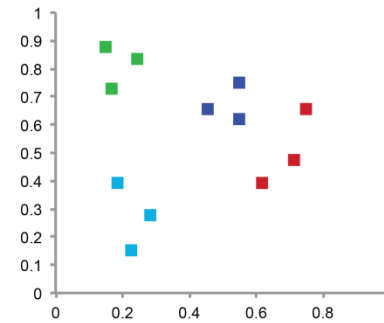
Things you can illustrate



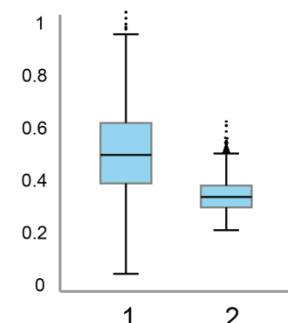
Relationship



Comparison



Composition



Distribution

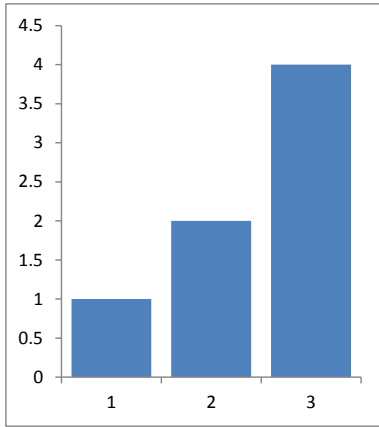
Graphical Representations

- 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

Marks and Channels

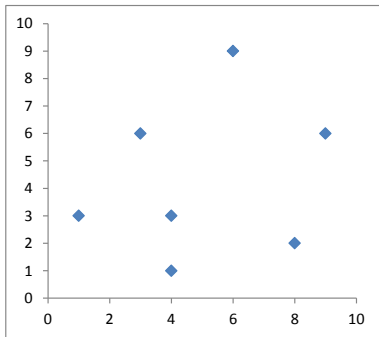
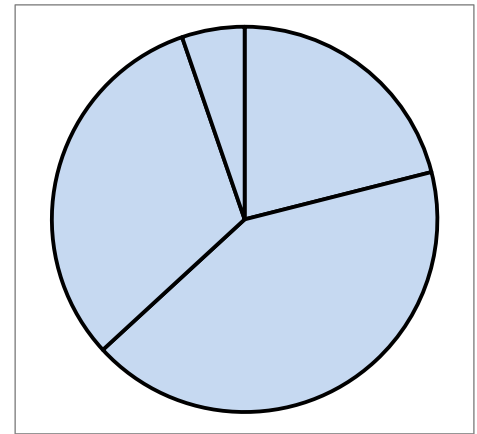
- Marks
 - Geometric primitives
 - Lines
 - Points
 - Areas
 - Used to represent data sets
- Channels
 - Graphical appearance of a mark
 - Colour
 - Length
 - Position
 - Angle
 - Used to encode data

Figures are a combination of marks and channels



1 Mark = Rectangle
1 Channel = Length of longest side

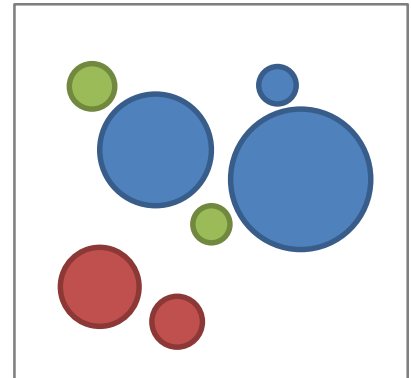
1 Mark = Circle segment
1 Channel = Angle



1 Mark = Diamond shape
2 Channels = X position, Y position

1 Mark = Circle
4 Channels:

X position
Y position
Area
Colour



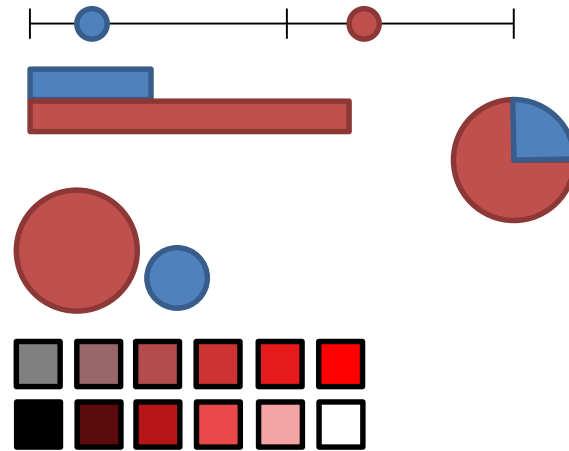
Golden Rules

- Effectiveness
 - Encode the most important information with the most effective channel
- Expressiveness
 - Match the properties of the data and channel

Types of channel

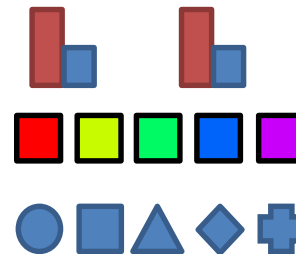
- Quantitative

- Position on scale
- Length
- Angle
- Area
- Colour (saturation)
- Colour (lightness)



- Qualitative

- Spatial Grouping
- Colour (hue)
- Shape



Colour

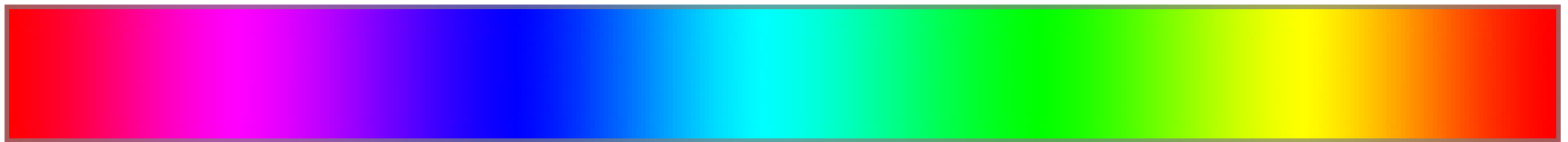
- Technical representations of colour
 - Red + Green + Blue (RGB)
 - Cyan + Magenta + Yellow + Black (CMYK)
- Perceptual representation of colour
 - Hue + Saturation + Lightness (HSL)

HSL Representation

- Hue = Shade of colour = Qualitative
- Saturation = Amount of colour = Quantitative
- Lightness = Amount of white = Quantitative

- Humans have no innate quantitative perception of hue but we have learned some (cold – hot, rainbow etc.)

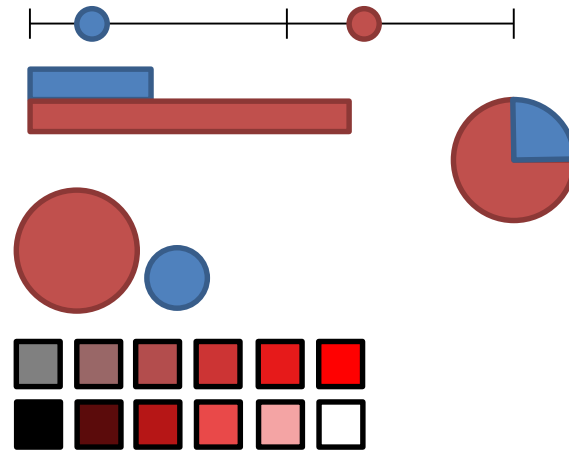
- Our perception of hue is not linear



Types of channel

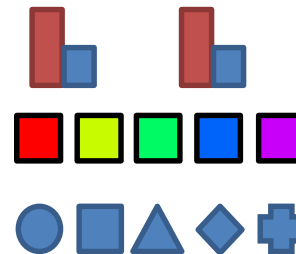
- Quantitative

- Position on scale
- Length
- Angle
- Area
- Colour (saturation)
- Colour (lightness)



- Qualitative

- Spatial Grouping
- Colour (hue)
- Shape



Data Types

- Quantitative
 - Height, Length, Weight, Expression etc.
- Ordered
 - Small, Medium, Large
 - January, February, March
- Categorical
 - WT, Mutant1, Mutant2
 - GeneA, GeneB, GeneC

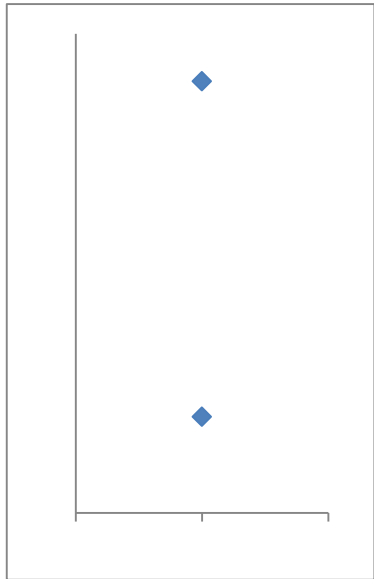
Golden Rules

- Effectiveness
 - Encode the most important information with the most effective channel
- Expressiveness
 - Match the properties of the data and channel

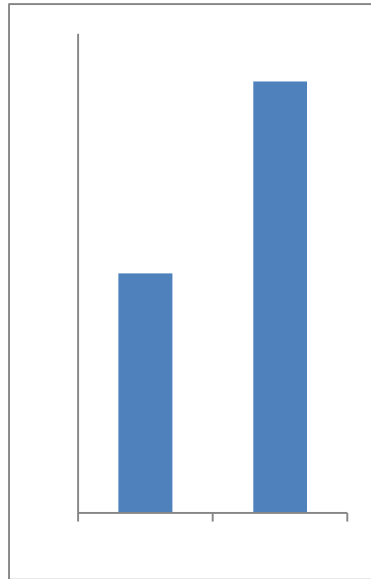
Golden Rules

- Effectiveness
 - Encode the most important information with the most effective channel
- Expressiveness
 - Match the properties of the data and channel

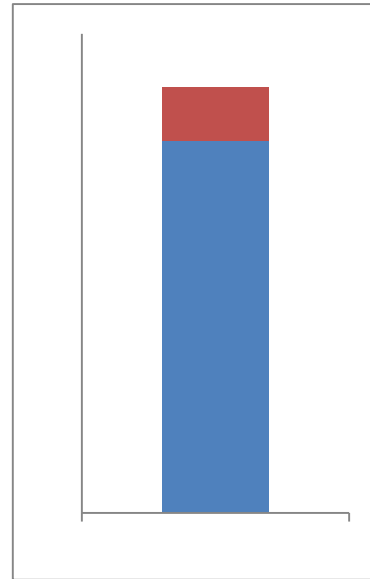
Effectiveness of quantitation



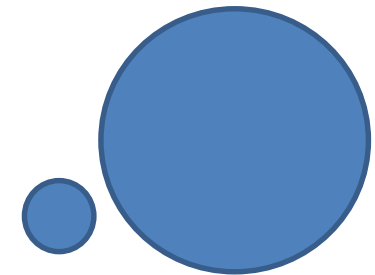
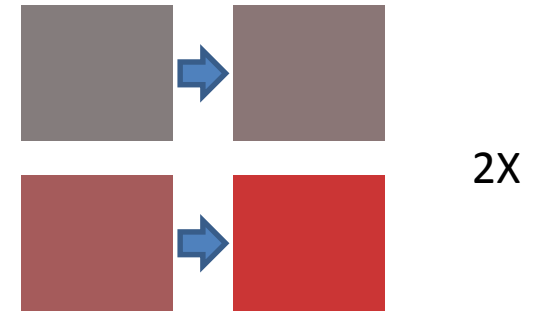
4.5X



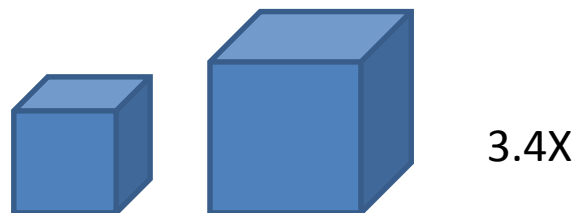
1.8X



7X



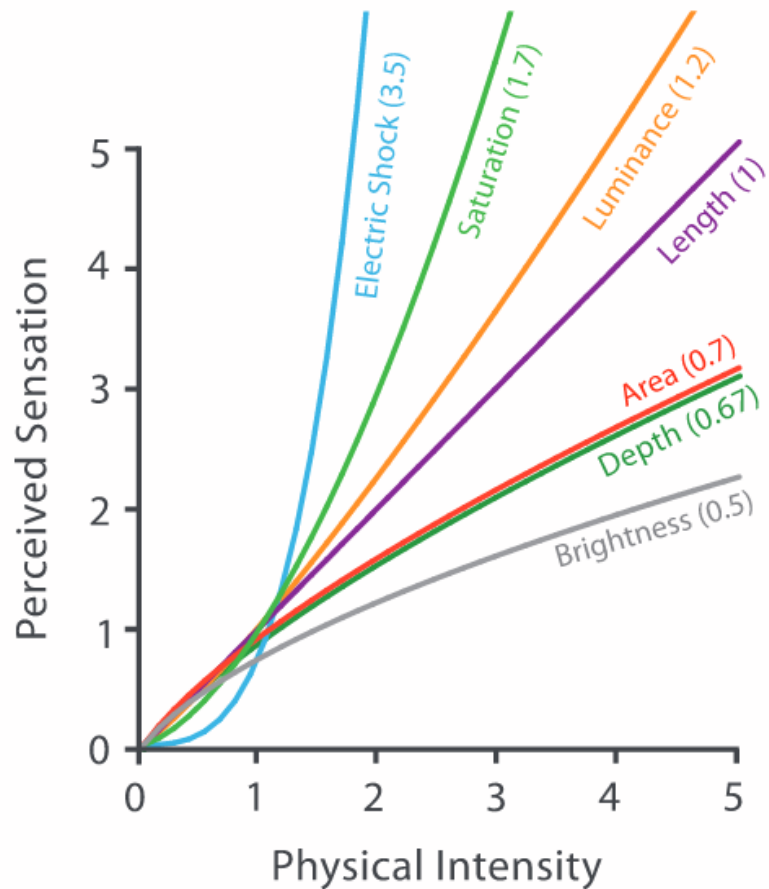
16X



3.4X

Quantitation Perception

Steven's Psychophysical Power Law: $S = I^N$

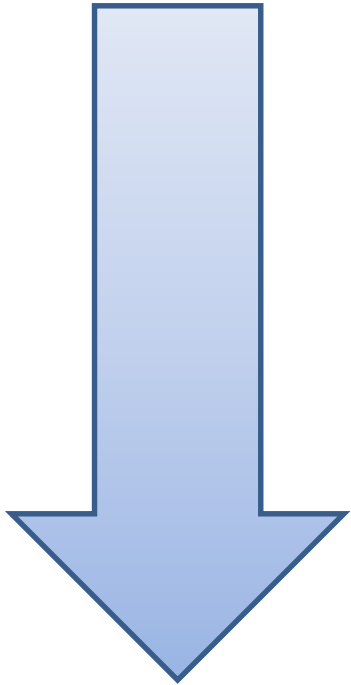


Golden Rules

- Effectiveness
 - Encode the most important information with the most effective channel
- Expressiveness
 - Match the properties of the data and channel

Most Quantitative Representations

Good quantitation



- Bar chart
- Stacked bar chart with common start
- Stacked bar chart with different starts
- Pie charts
- Bubble plots (circular area)
- Rectangular area
- Colour (luminance)
- Colour (saturation)

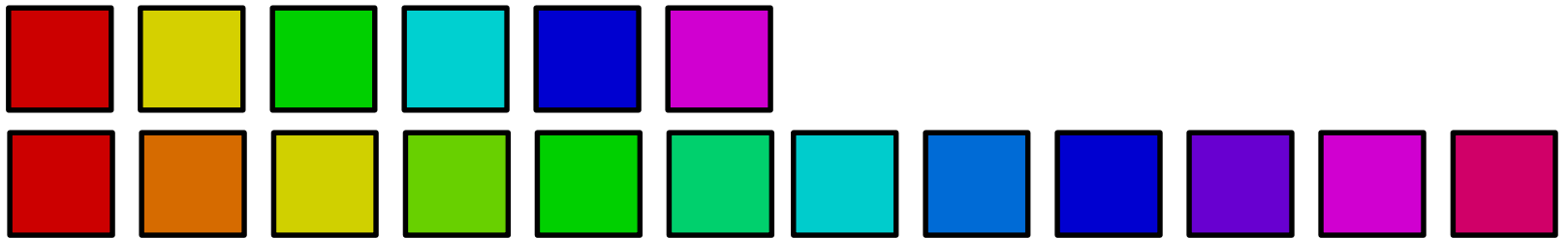
Poor quantitation

Discriminability

- If you encode categorical data are the differences between categories easy for the user to perceive correctly?

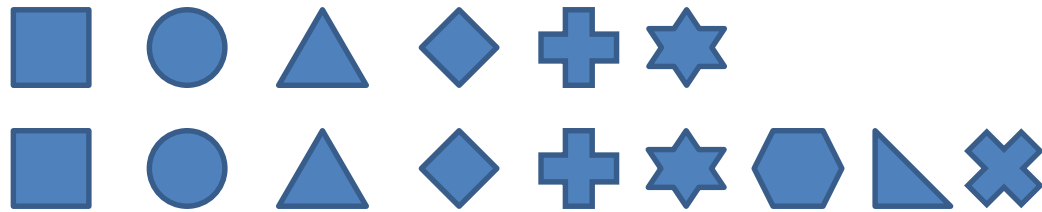
Qualitative Discrimination

- How many colours can you discriminate?



Qualitative Discrimination

- How many (fillable) shapes can you discriminate?



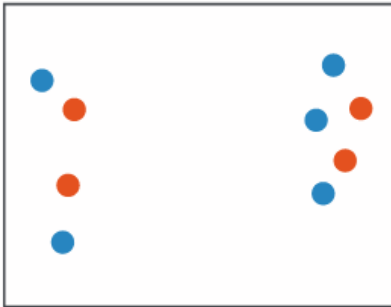
- Can combine with colour, but need to maintain similar fillable areas

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

Separability

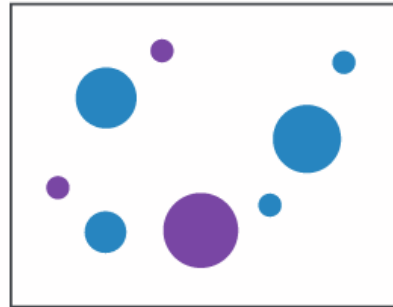
Position
+ Hue (Color)



Fully separable

There is no confusion between the two channels

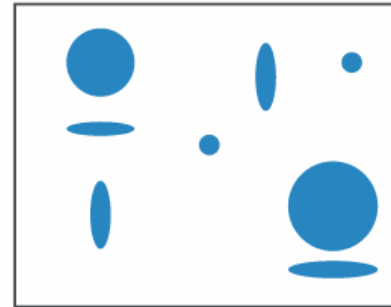
Size
+ Hue (Color)



Some interference

Larger points are easier to discriminate than smaller ones

Width
+ Height



Some/significant interference

We tend to focus on the area of the shape rather than the height/width separately

Red
+ Green



Major interference

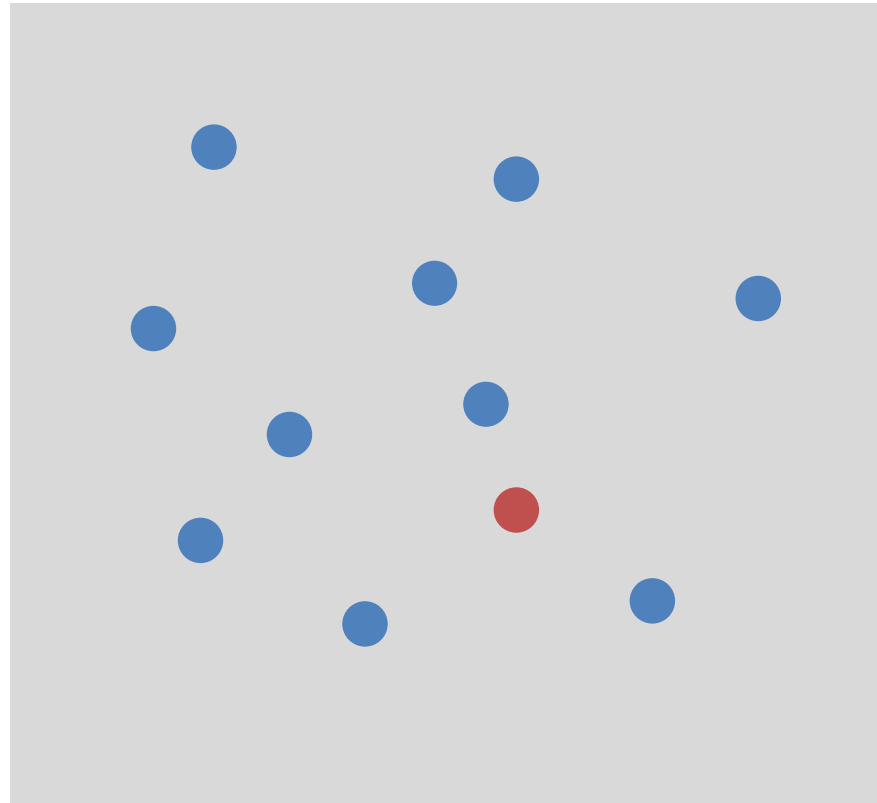
Humans are very bad at separating combined colours

Popout

- A distinct item immediately stands out from the others
- Triggered by our low level visual system
- You don't need to actively look at every point (slow!) to see it

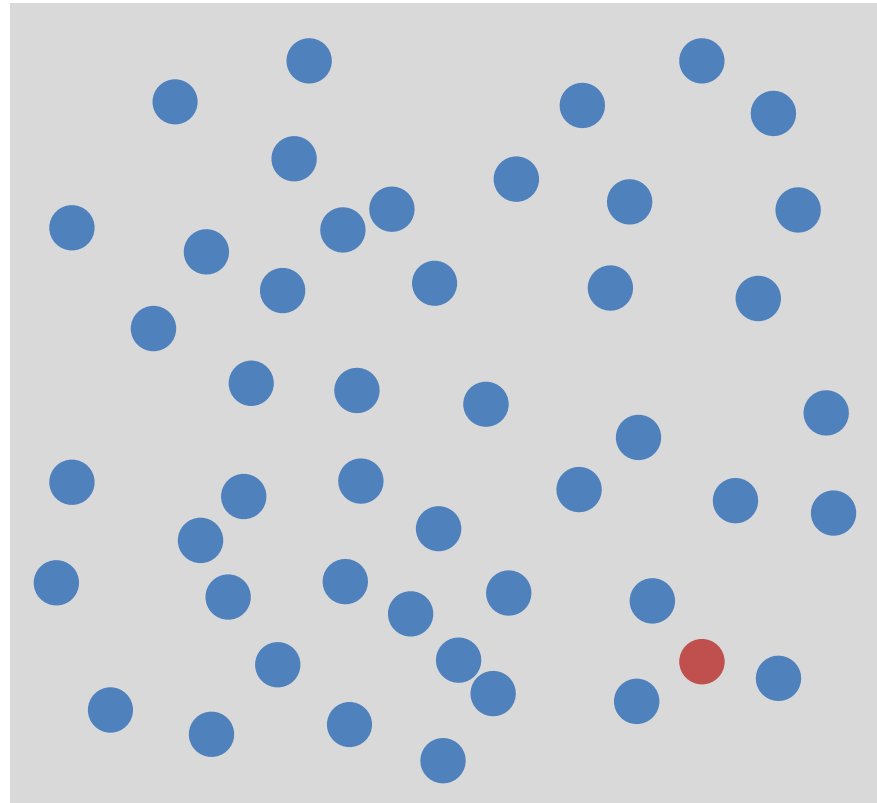
Popout

(find the red circle)



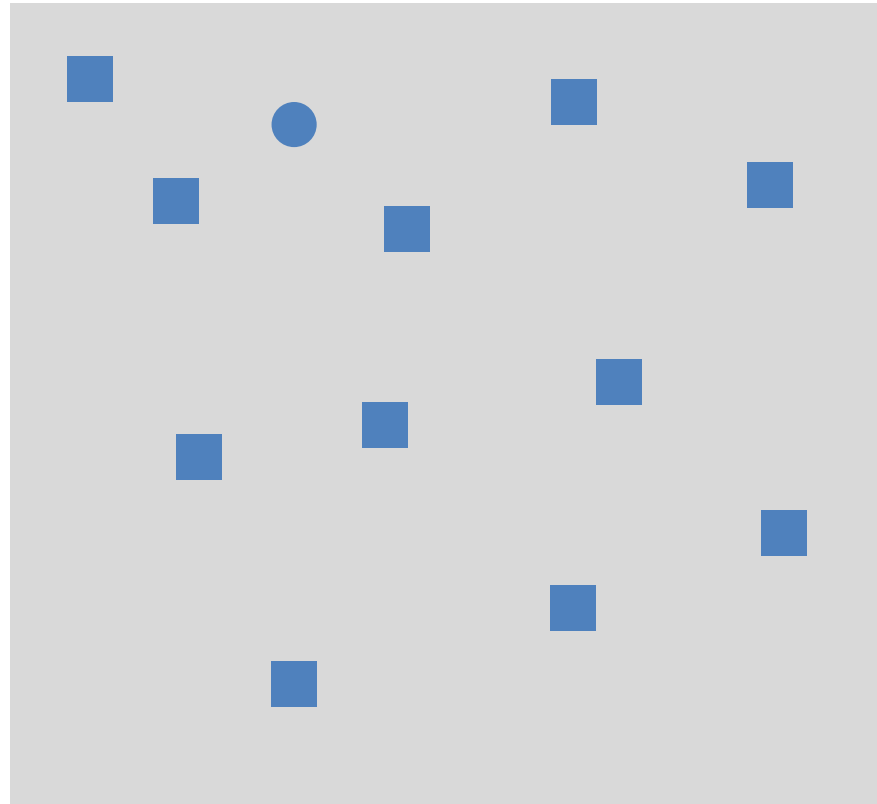
Popout

Speed of identification is independent of the number of distracting points



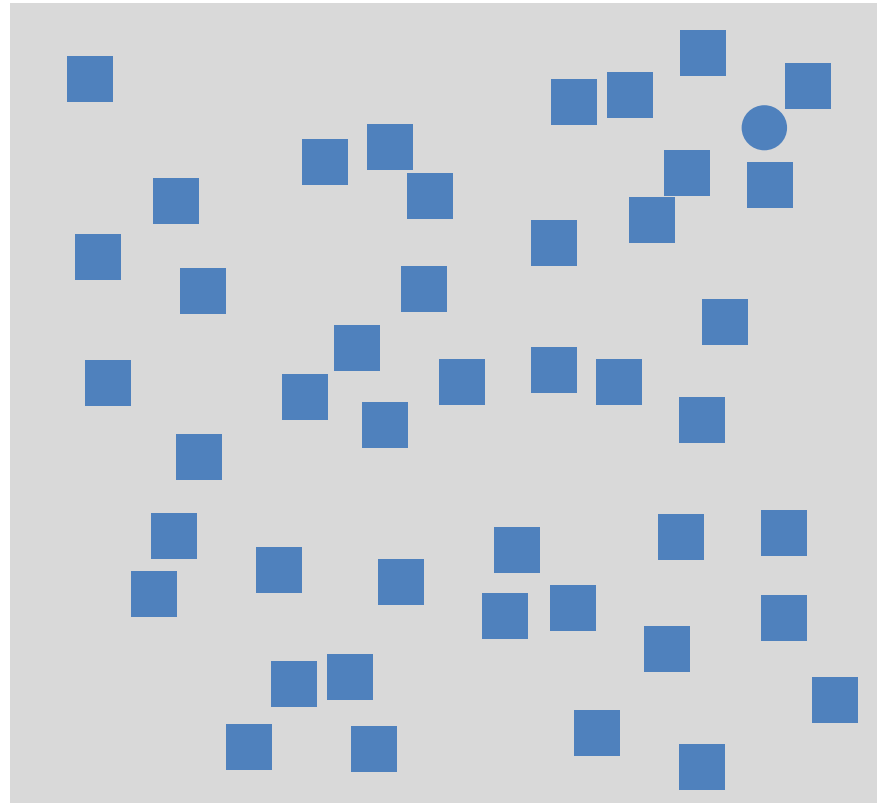
Popout

(Find the circle)



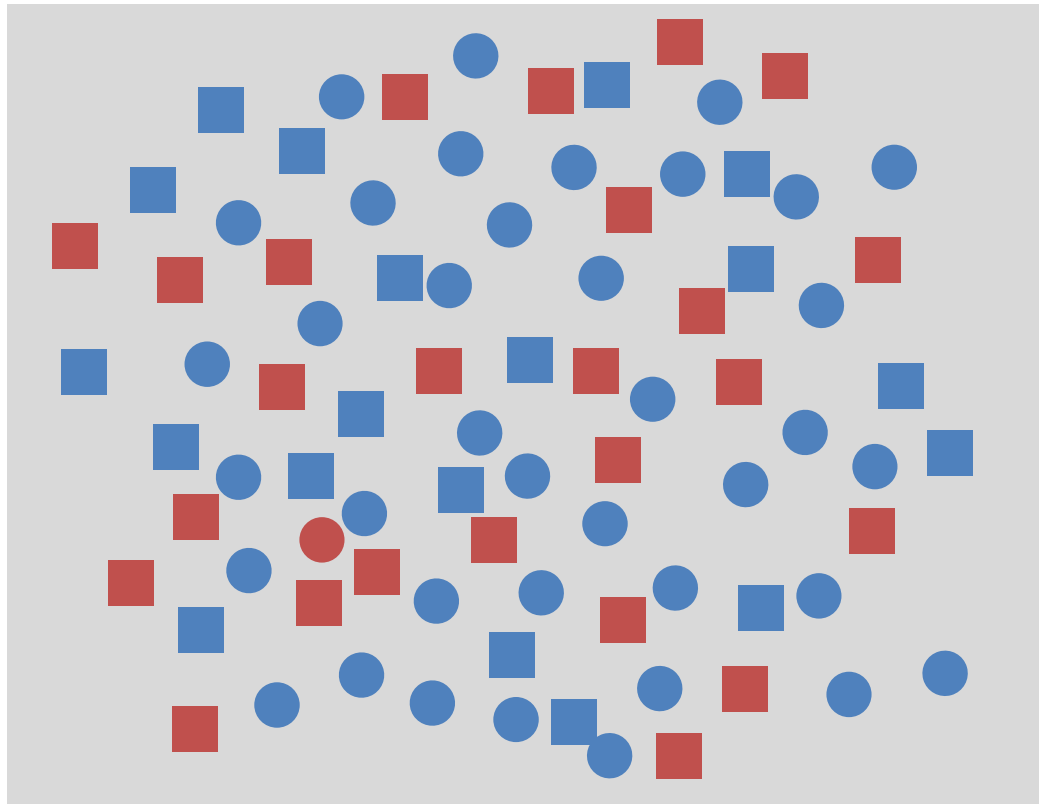
Popout

Colour pops out more than shape



Popout

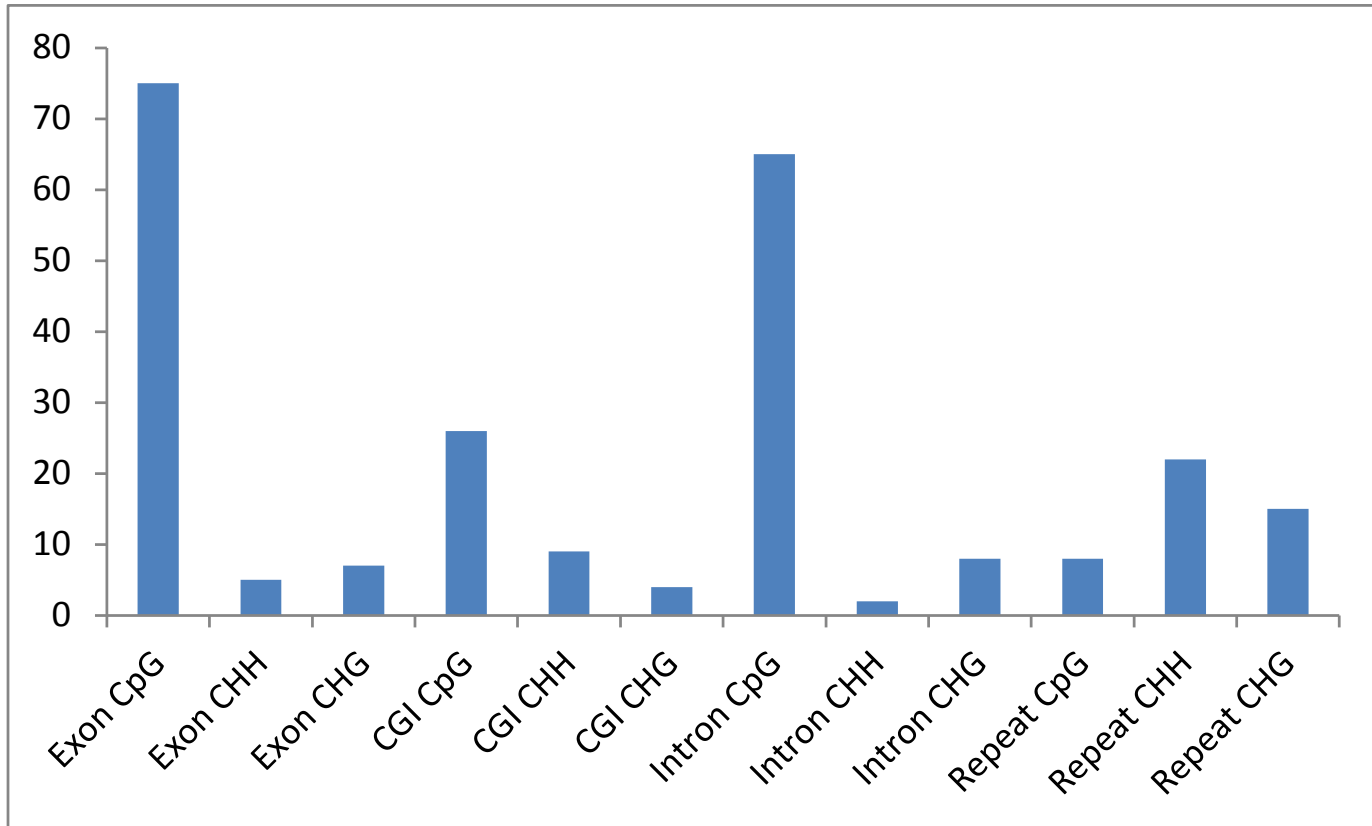
Mixing channels removes the effect
(Find the red circle)



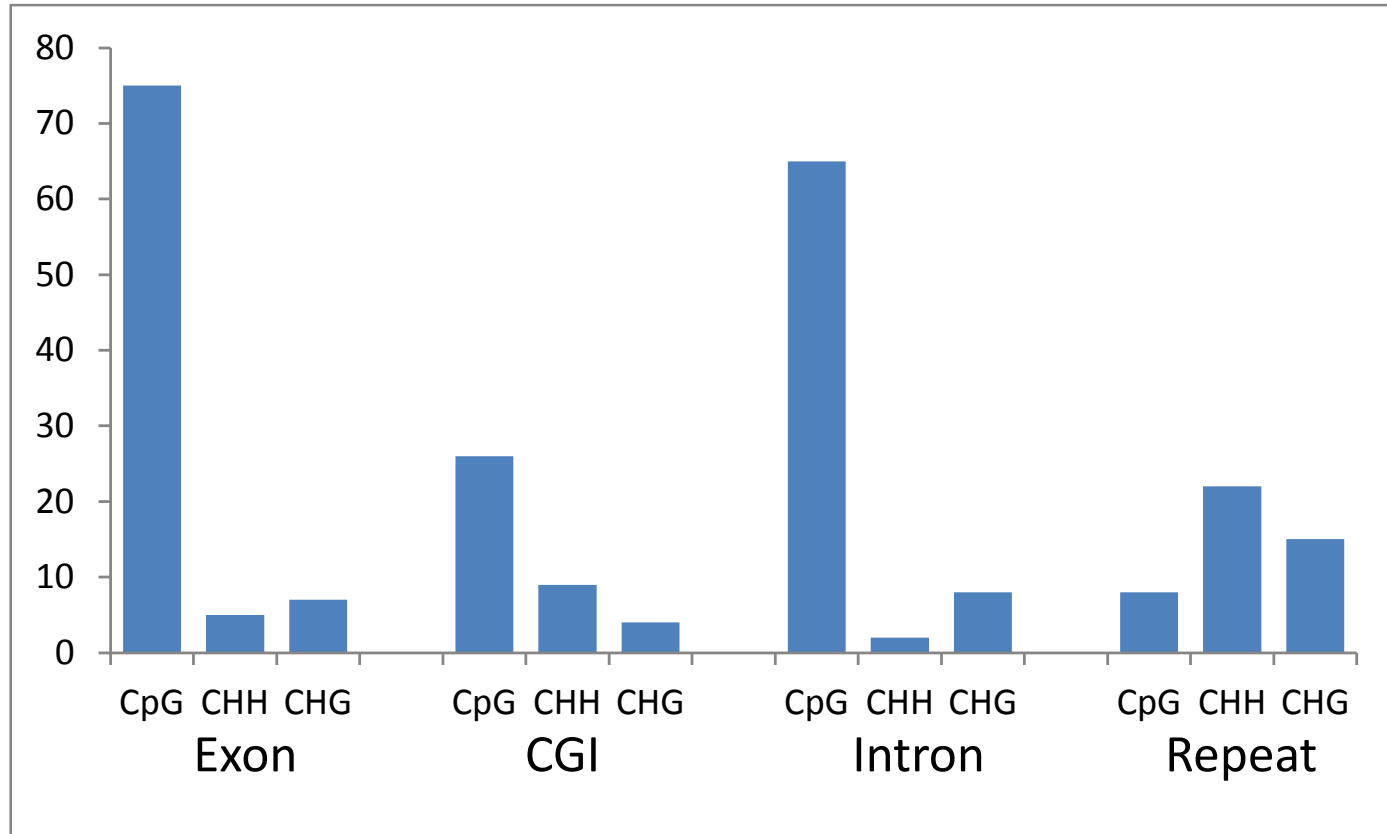
Use of space

- Where you want a viewer to focus on specific subsets of data you can help their perception by using the layout or highlighting of data to draw their attention to the point you're making

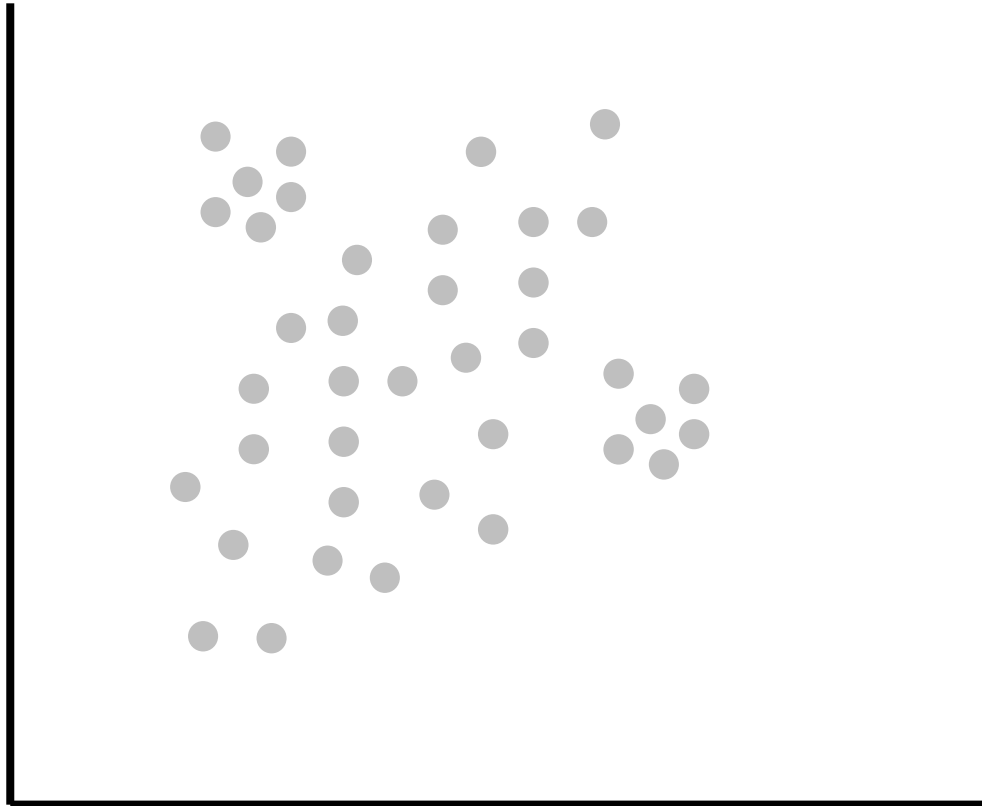
Grouping



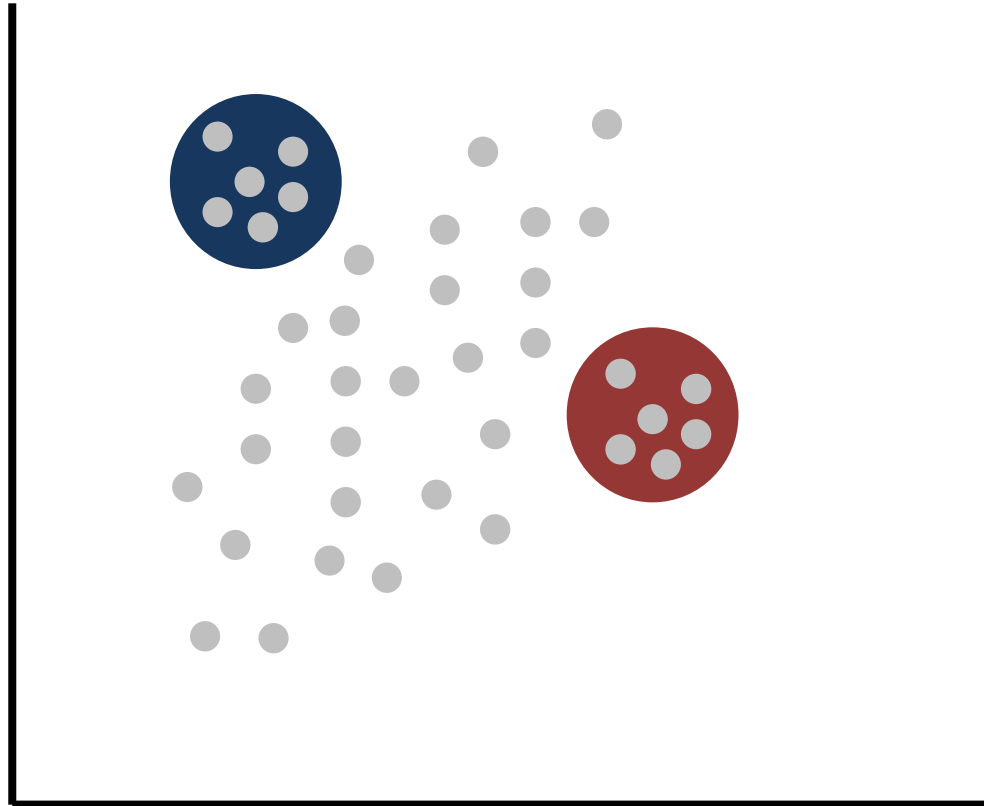
Grouping



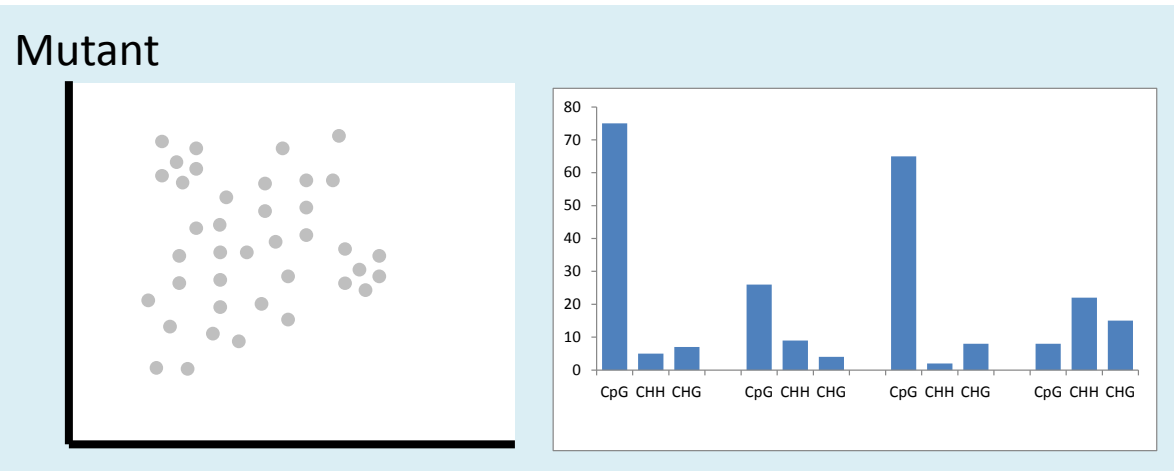
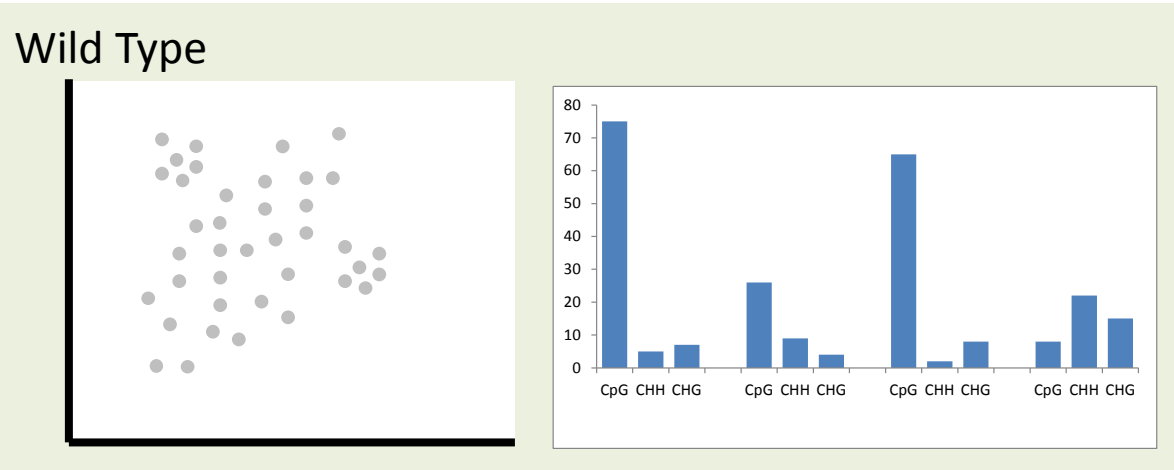
Containment



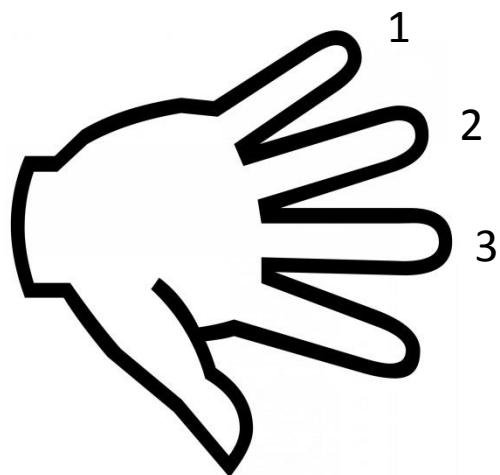
Containment



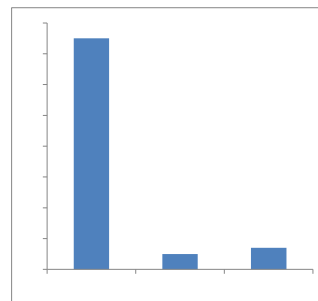
Containment



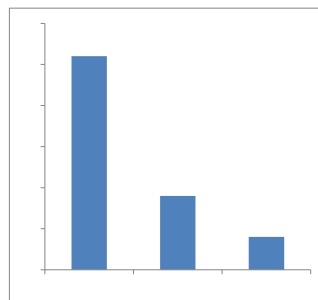
Linking



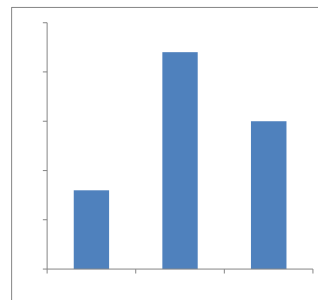
1



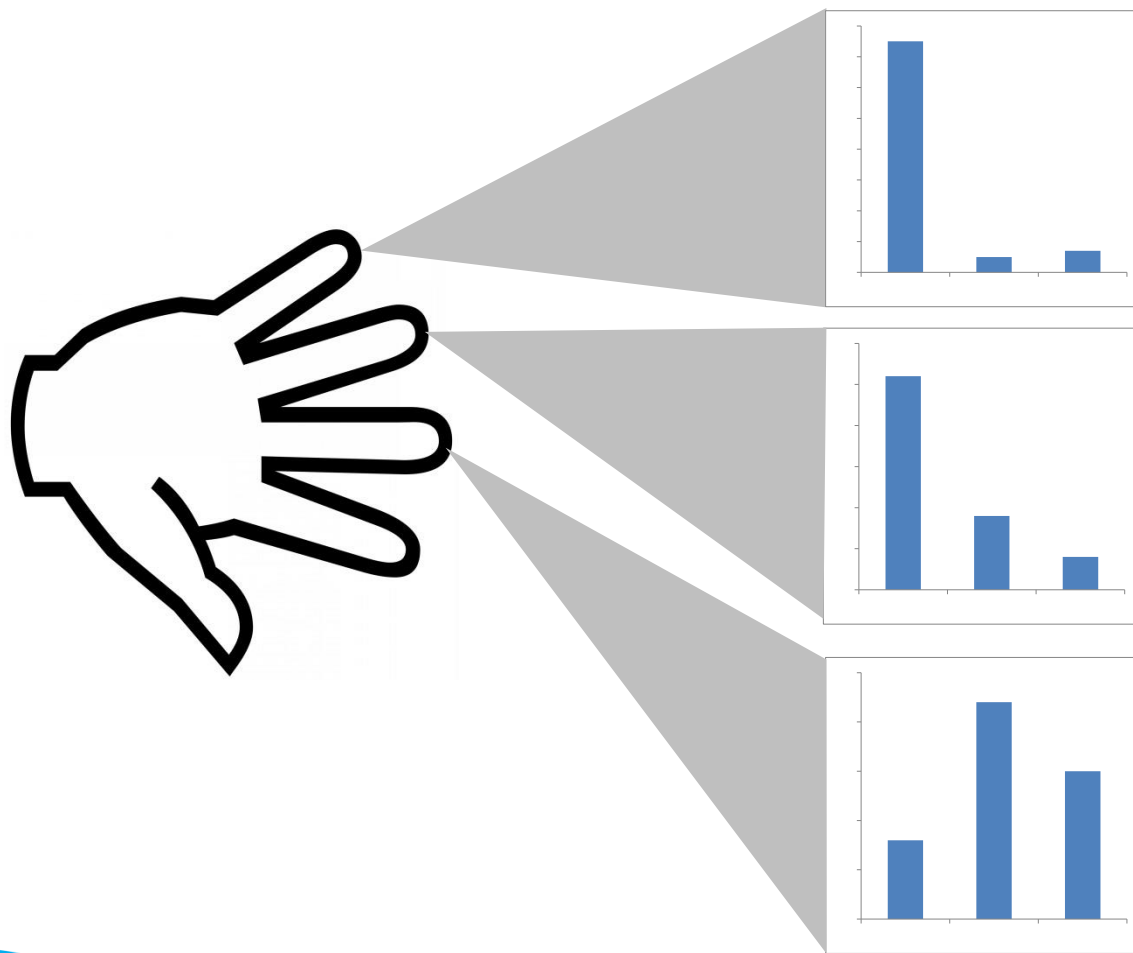
2



3

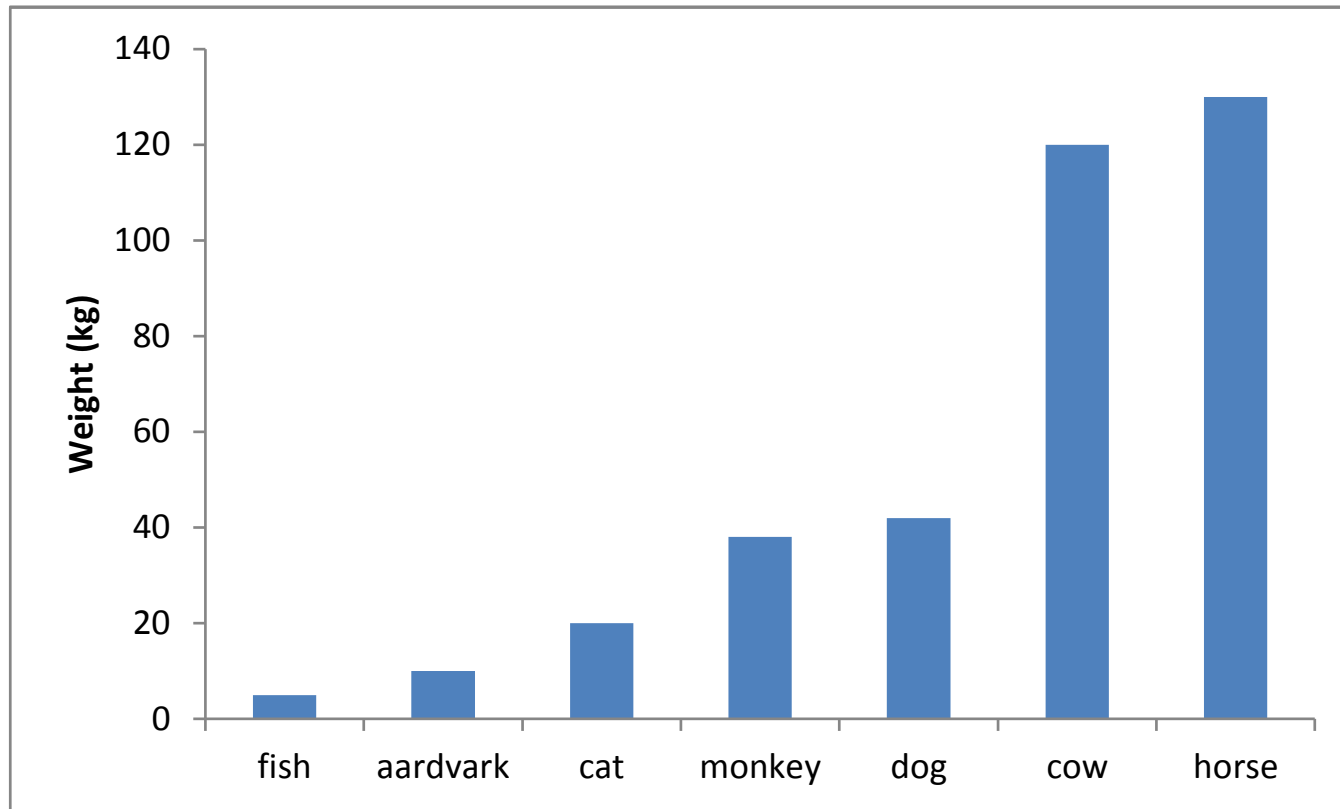


Linking



Ordering

- Is a monkey heavier than a dog?



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?

General Rules

- No unnecessary figures
 - Does a graphical representation make things clearer?
 - Would a table be better?
- One point per figure
 - Design each figure to illustrate a single point
 - Adding complexity compromises the effectiveness of the main point
- No absolute reliance on colour
 - Figures should ideally still work in black and white
 - Colour should help perception
- No 3D
 - 3D is hardly ever justified and makes things less clear
- Figures should be self-contained
 - Must be understandable without additional information