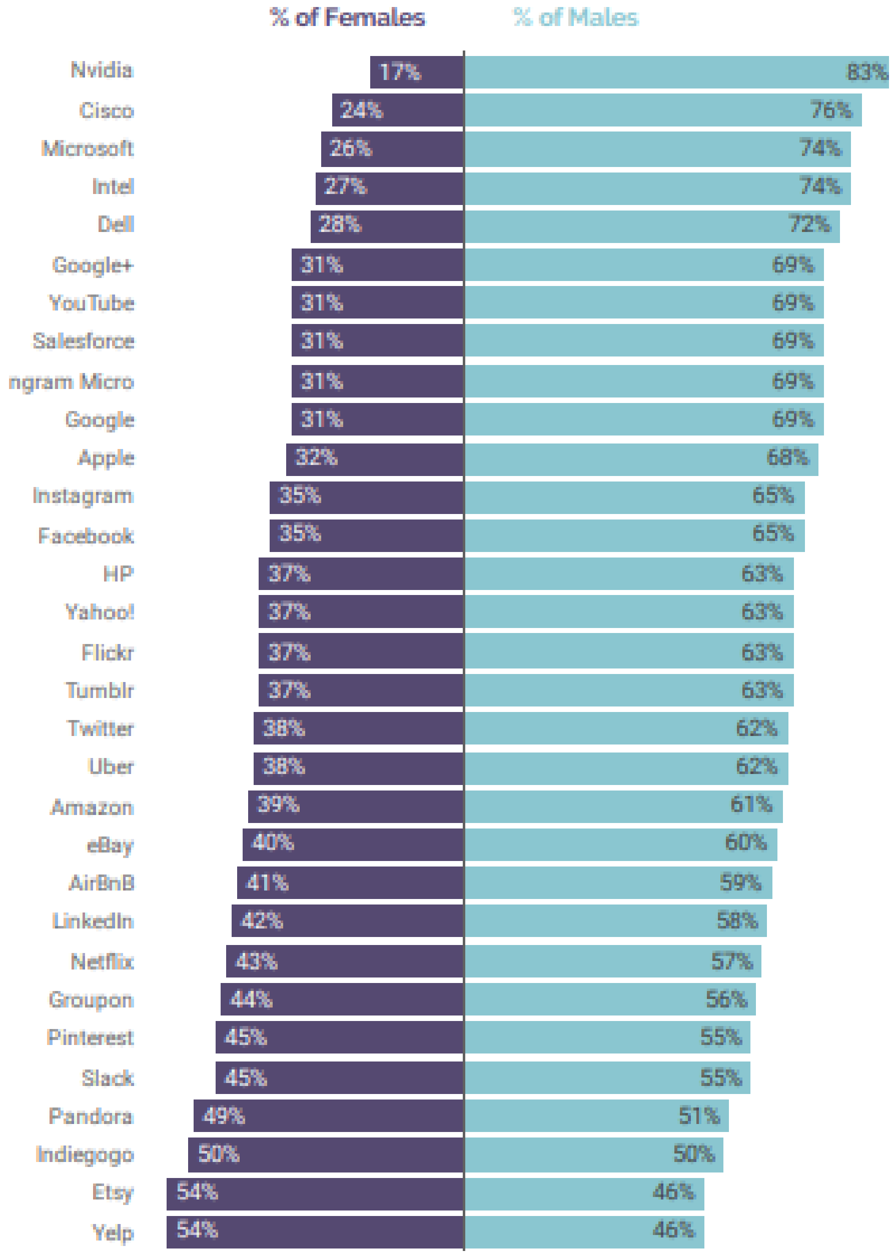
Data Visualisation
Lecture Week 9 – Encodings 2

Dr. Cathy Ennis

Learning Outcomes Week 9

- Create and deploy successful data visualisations using leading software tools.
- Design effective visualizations based on principles from perceptual psychology, cognitive science, graphic design and visual art.
- Analyse and evaluate how mental models aid in the interpretation of complex visual displays.
- Select, formulate and integrate metaphors to suit data-driven tasks

In May 2018, many key technology companies still have an imbalance in their employee gender split.



Source: Company and Press Reports

#MakeoverMonday - 2018 - Week 46

Created by

sianmiller.co.uk

@sianmiller141

Overview

- In today’s lecture we will look at how to use visual encodings appropriately:
    - Colour
    - Size
    - Text
    - Shapes
    - Labels
- In today’s Lab we will use R to create visualisations: Introduction to R to create visualisations

Apply Your Encodings Well

Apply Your Encodings Well

- Colour
  - Leverage Common Colour Associations
  - Colour Theory
  - Cognitive Interference and the Stroop Test
- Size
  - Conveying Size
  - Comparing Sizes
- Text and Typography
  - Use Text Sparingly
  - Fonts and Hierarchies
  - Beware of All Caps
  - Avoid Drop Shadows
- Shape
  - Cultural Connotations
  - Icons
  - Illusions
- Keys versus Direct Labelling of Data Points

Apply Your Encodings Well

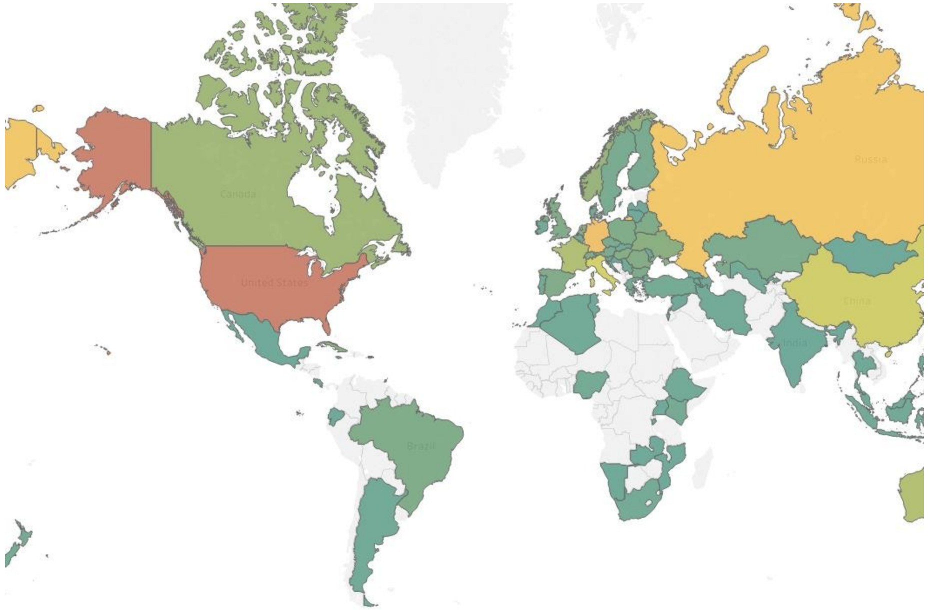
- Colour
  - Leverage Common Colour Associations
  - Colour Theory
  - Cognitive Interference and the Stroop Test
- Size
  - Conveying Size
  - Comparing Sizes
- Text and Typography
  - Use Text Sparingly
  - Fonts and Hierarchies
  - Beware of All Caps
  - Avoid Drop Shadows
- Shape
  - Cultural Connotations
  - Icons
  - Illusions
- Keys versus Direct Labelling of Data Points

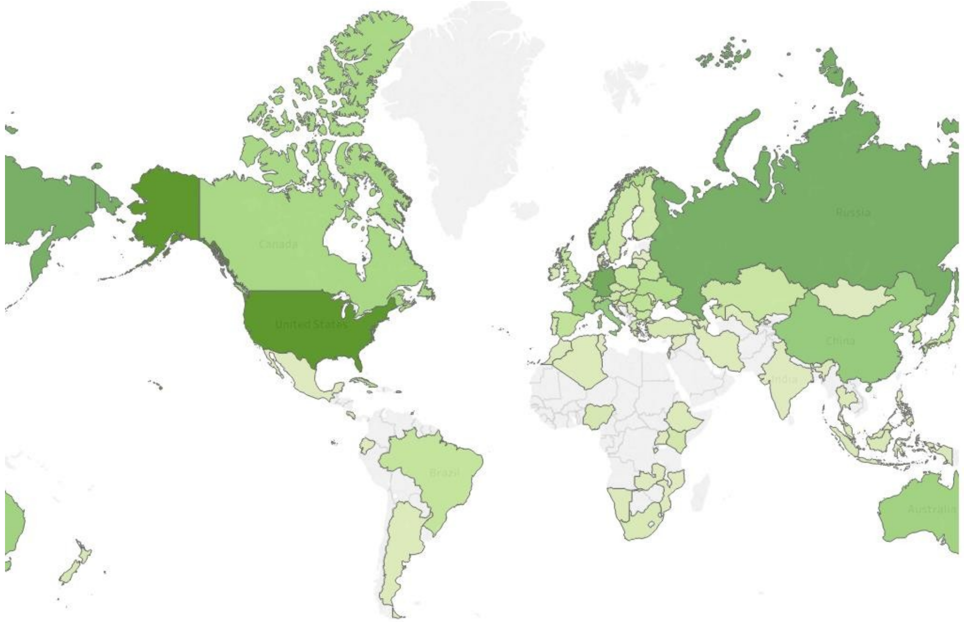
Colour

- It bears repeating here because it is such a common mistake: avoid using colour (hue) for order
  - Vary brightness or saturation

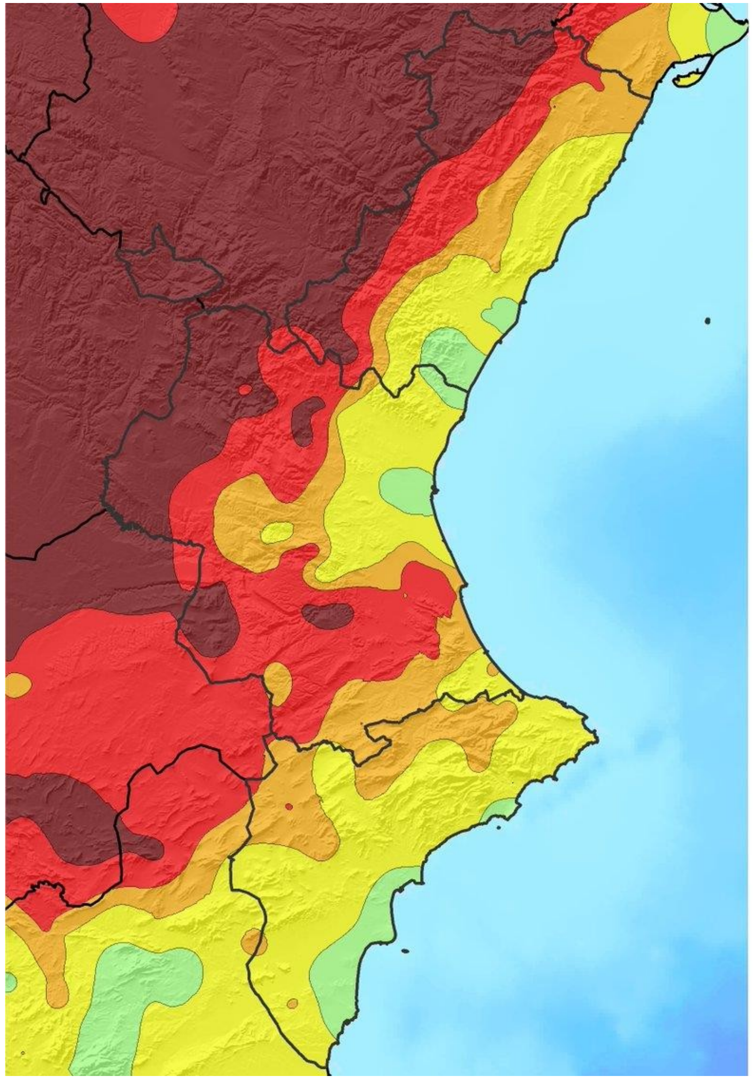
- Colour is an excellent property for labelling categorical data, or non-ordered categories for differentiation purposes
  - e.g. operating system, gender, region, conference track, and genre
  - Just make sure you don’t need too many distinct values

Use of colour



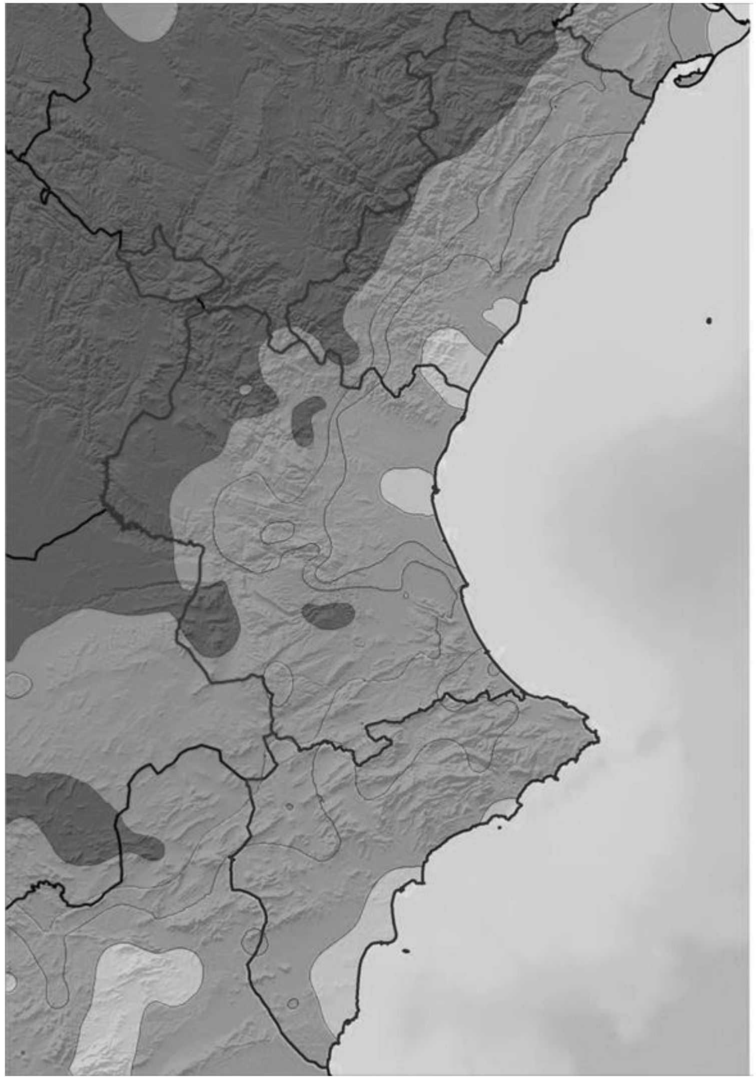


Colour Use



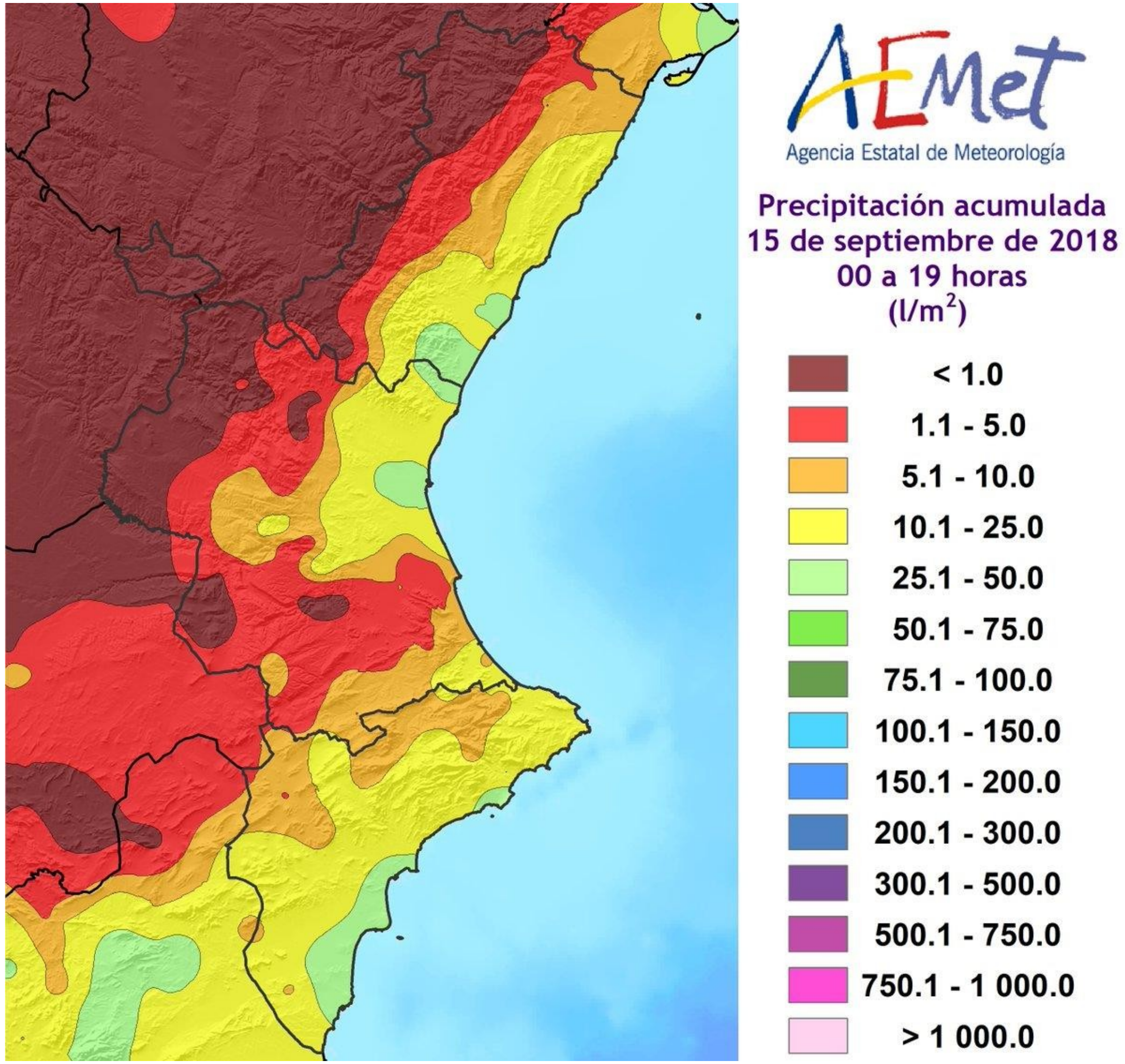
Rainfall in 19 hour period (l/m2)

Colour Use



Grey scale Rainfall in 19 hour period (l/m2

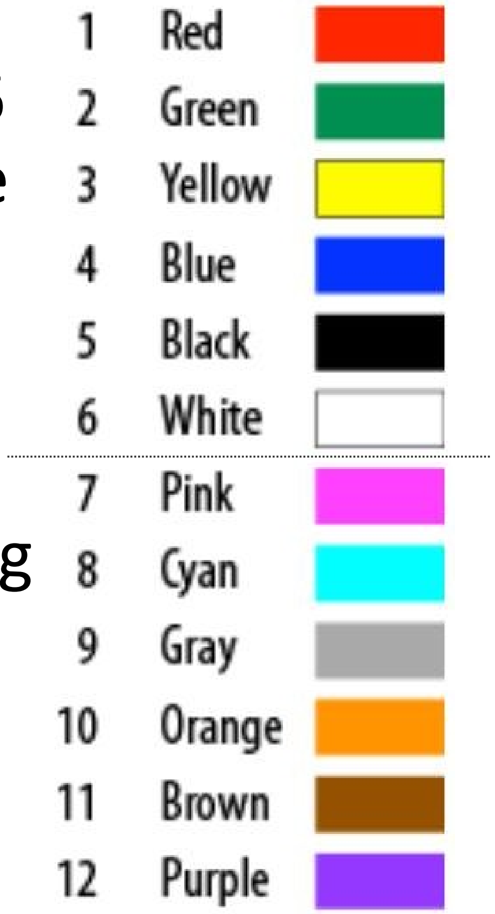
Colour Use



Rainfall in 19 hour period (l/m2

Categorical Data - 12 Colours

- The standard advice for using colour to encode categories is to limit your selection to ideally about 6 - hopefully no more than 12, and absolutely no more than 20 - colours and corresponding categories
- It is recommended to use primary colours first, then secondary colours.
- Select first from the first half of the list before moving on to the second half



1	Red	🟥
2	Green	🟩
3	Yellow	🟨
4	Blue	🟦
5	Black	⬛
6	White	⬜
7	Pink	🟪
8	Cyan	🟦
9	Gray	▨
10	Orange	🟧
11	Brown	🟫
12	Purple	🟪

Leverage Common Colour Associations

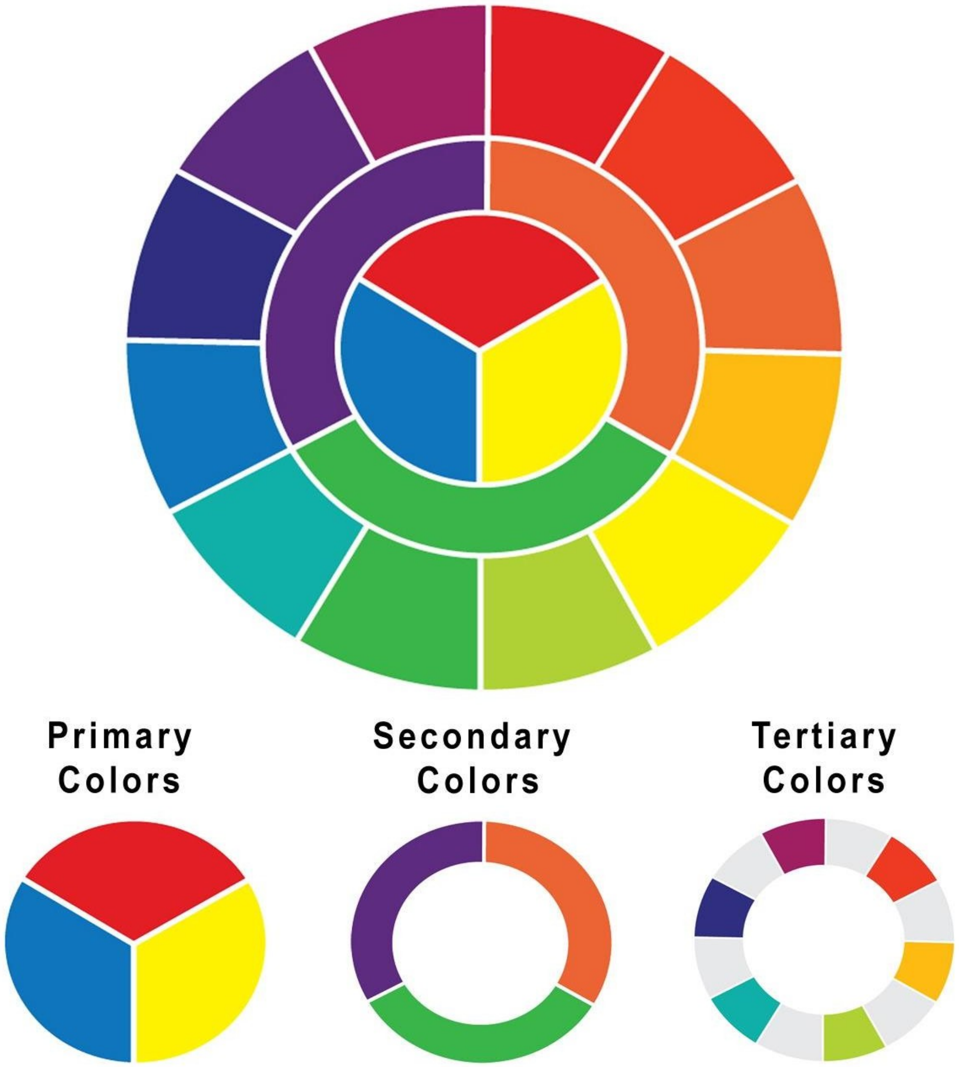
- Colour may not have a natural ordering, but it does carry a lot of cultural conventions, including many common emotional or aesthetic associations

Red	Warning, danger, warfare Passion, love
Green	Nature, earth
Yellow	Sunshine, happiness
Blue	Water, coolness, calm, religion, military
Black	Death, mourning Luxury
White	Weddings, innocence
Pink	Affection, imagination
Grey	Neutrality, conservatism
Orange	Fire, energy
Brown	Dirt, leather, stone
Purple	Royalty, magic

Colours For Labelling

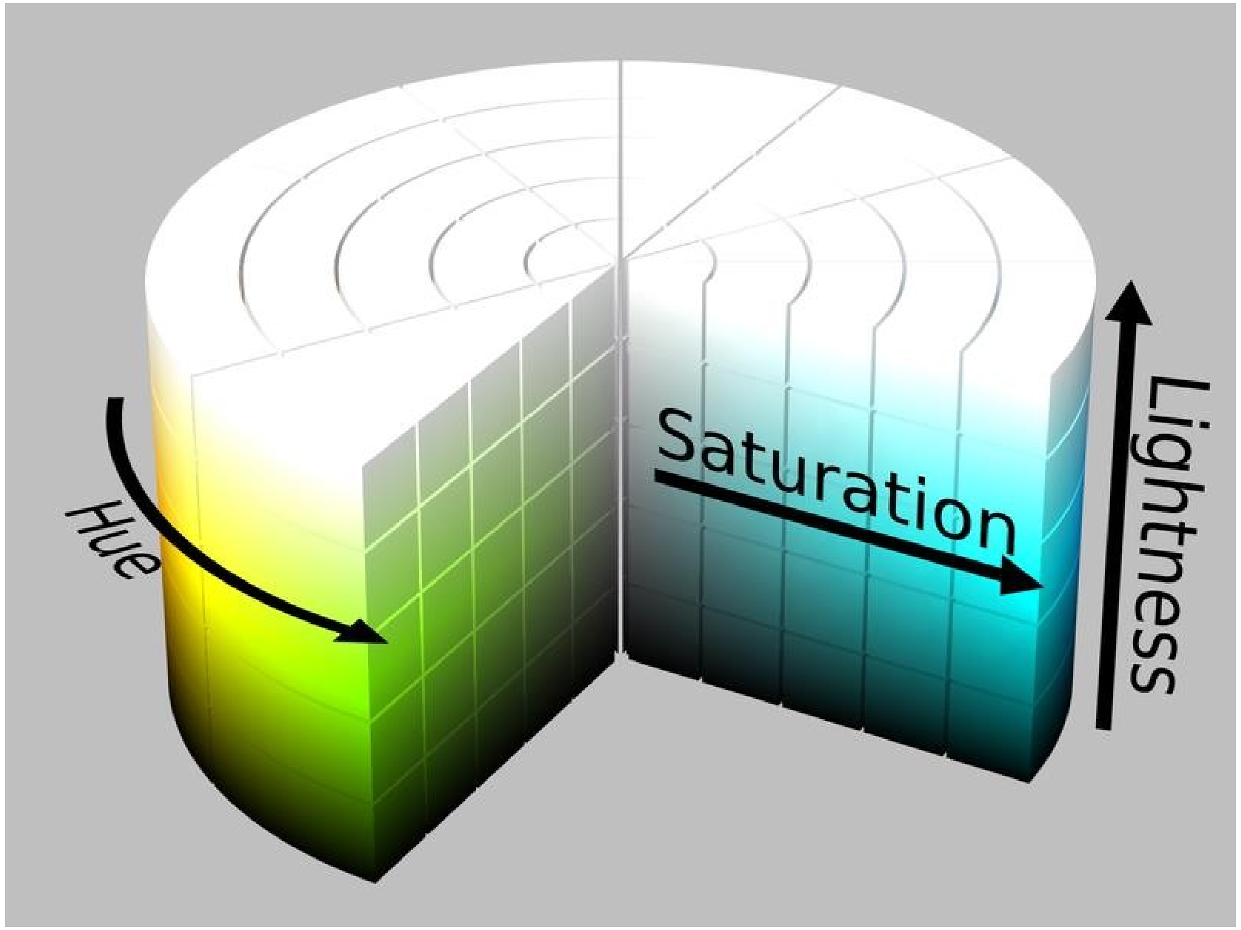
- Things to take into account when selecting your palette:
  - Type of data
  - Distinctness
  - Unique hues
  - Contrast with background
  - Colour blindness
  - Field Size
  - Conventions

Colour Wheel



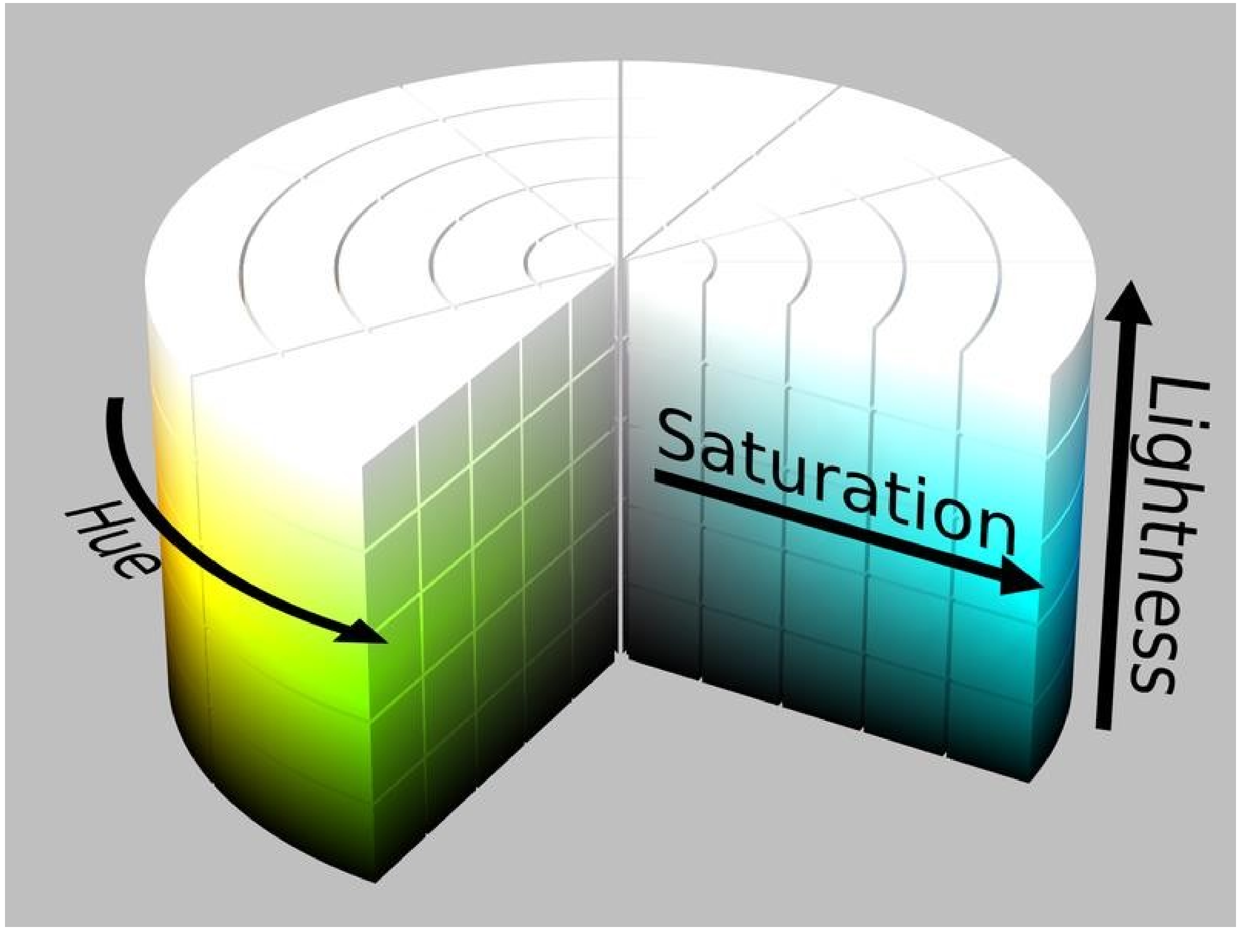
HSL Cylinder - Hue

- Hues refer to the set of "pure" colours within a colour space
- Hues range from 0 to 359 degrees



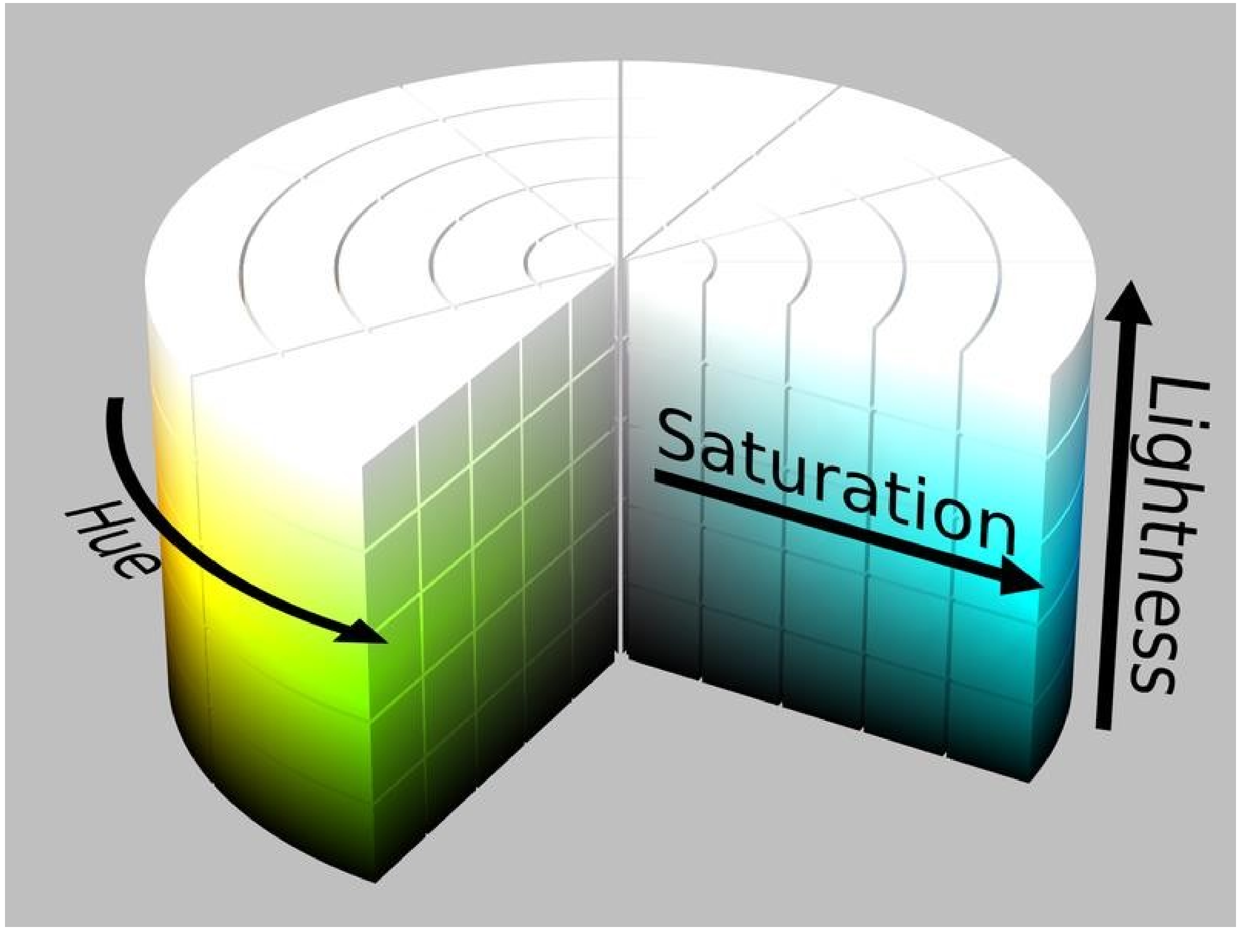
HSL Cylinder - Saturation

- Saturation defines a range from pure colour (100%) to grey (0%) at a constant lightness level. A pure colour is fully saturated
- From a perceptual point of view, saturation influences the grade of purity or vividness of a colour/image
- A desaturated image is said to be dull, less colourful or washed out but can also make the impression of being softer



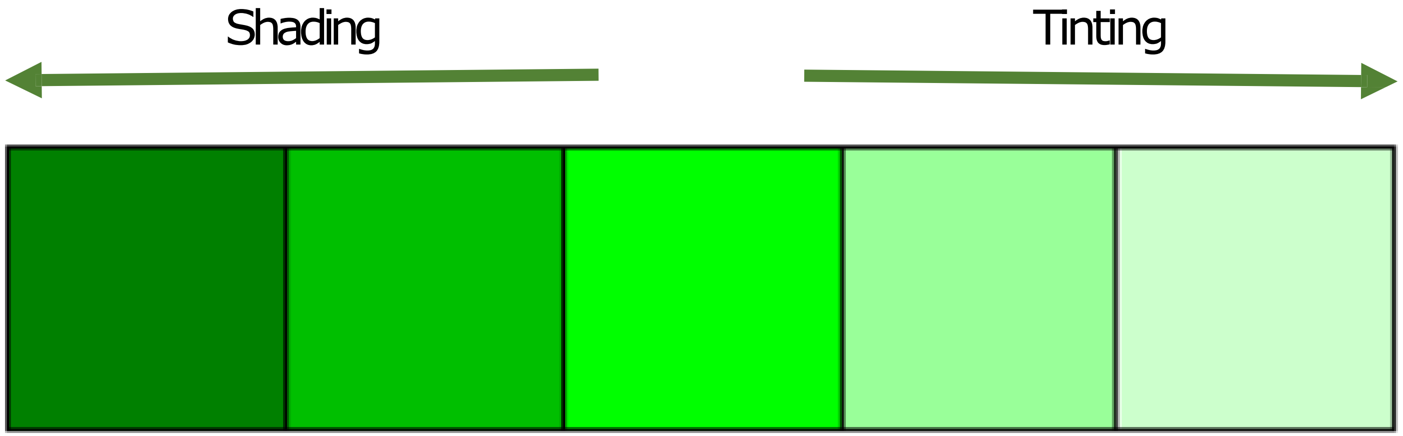
HSL Cylinder - Lightness

- Lightness defines a range from dark (0%) to fully illuminated (100%)
- Any original hue has the average lightness level of 50%
- Lightness is the range from fully shaded to fully tinted



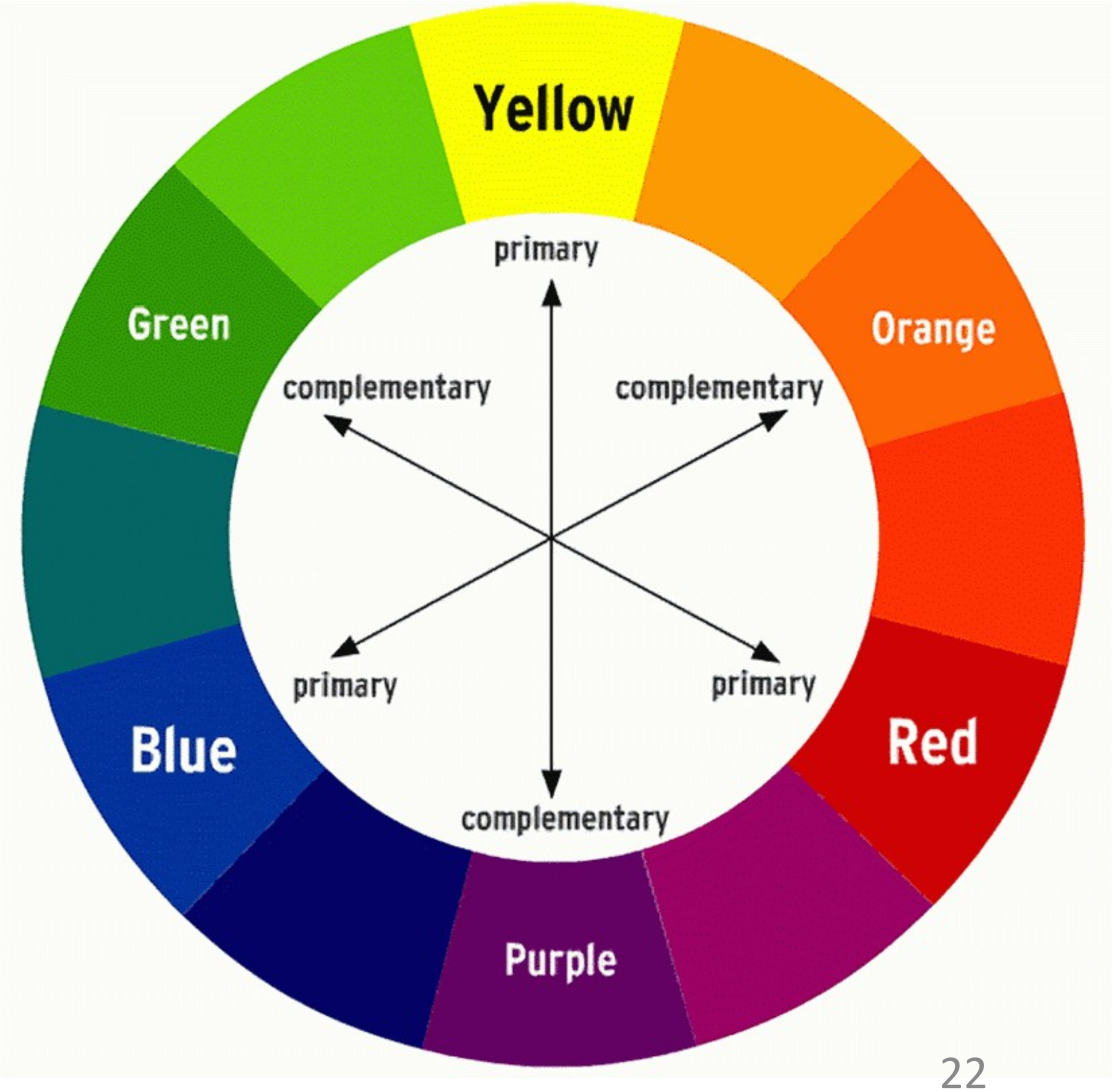
Ordinal data - Monochromatic Palette

- You can create a monochromatic palette by preserving the hue and adding different amounts of either black or white to vary the lightness
  - adding black is called **shading** and adding white is called **tinting**
  - very useful for **heat maps** and encoding relative intensity



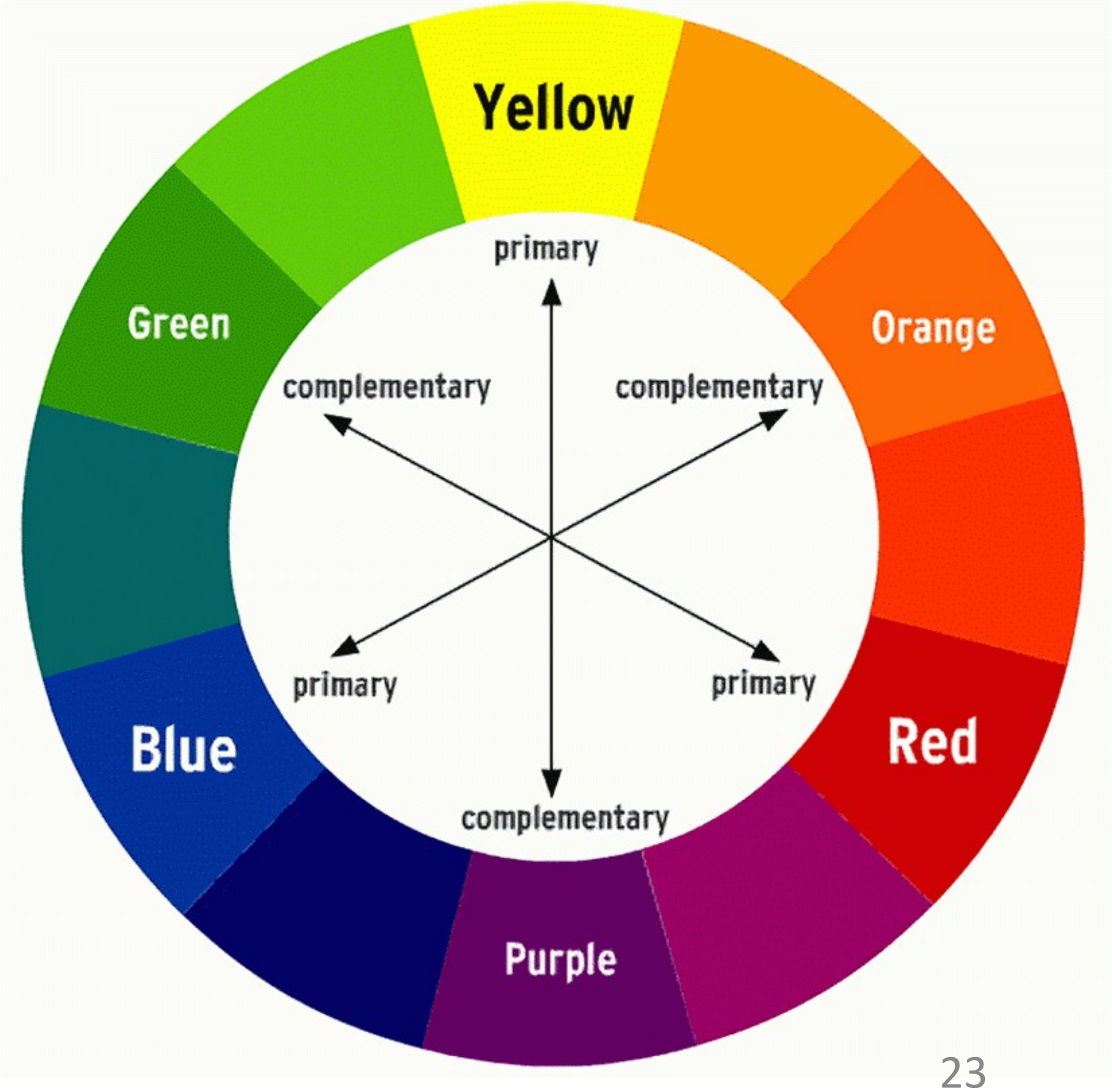
Colour Palettes

- You can create colour palettes from:
  - complimentary colours (colours opposite each other on the colour wheel)
  - analogous colours (neighbouring colours from within the same "pie slice" of the colour wheel)
  - triadic colours (three colours equally spaced around the colour wheel)
- No matter how you select your group of hues, however, if they all have the same saturation or brightness, they will compete with each other for the eye's attention

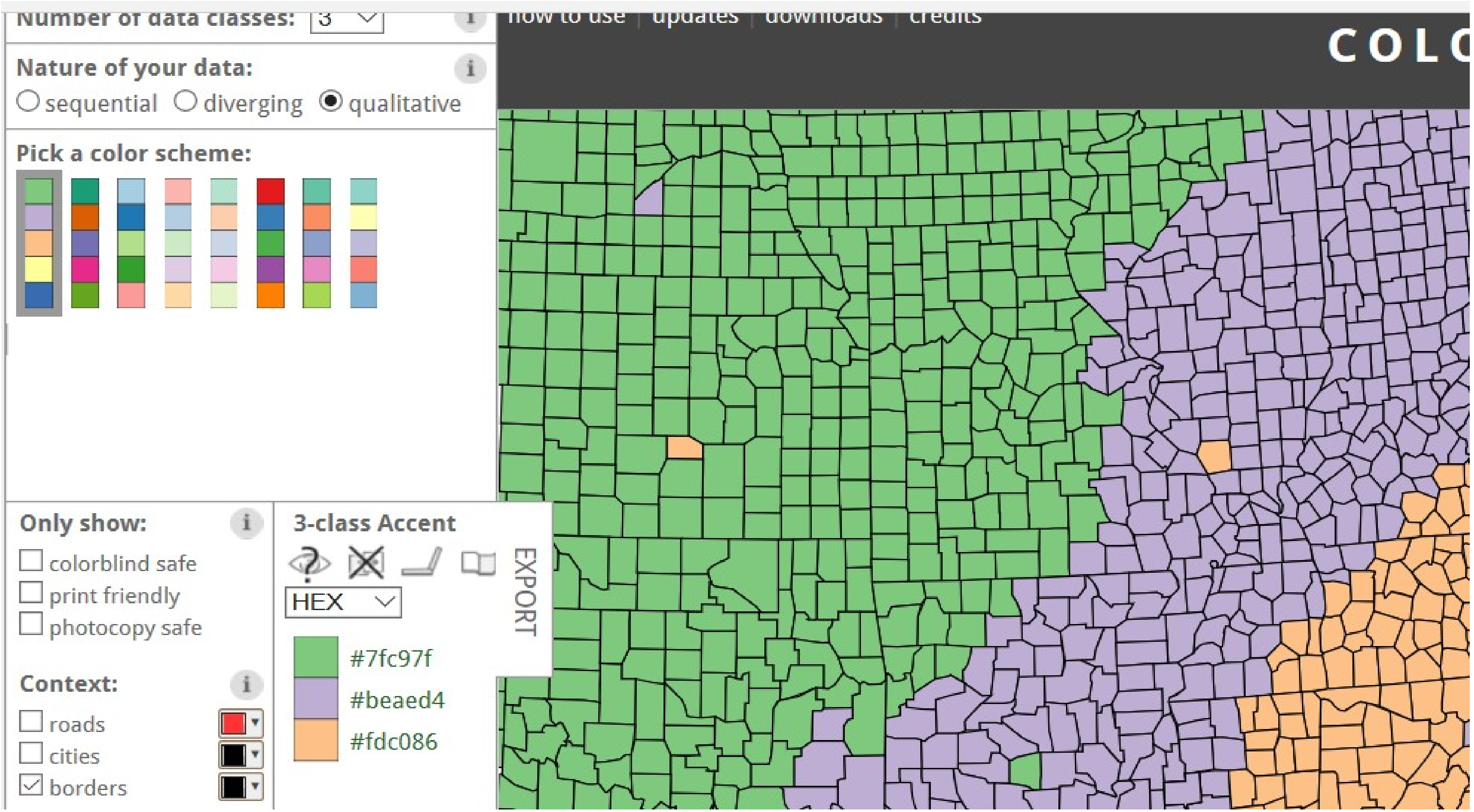


Colour Palettes

- Some colours don’t go together well
  - Colours far apart on the colour wheel but not complimentary or triadic) are perceived as conflicting—they “clash.”
  - When these hues are of similar brightness (such as red and blue) and are placed next to each other, you sometimes get a “shimmering” effect
  - This effect can be startling or jarring, so it can grab attention. But it can also make it difficult for the reader to pay attention to the underlying elements or message



Colorbrewer



Number of data classes: 3

Nature of your data:
○ sequential ○ diverging ● qualitative

Pick a color scheme:

Only show:
□ colorblind safe
□ print friendly
□ photocopy safe

Context:
□ roads
□ cities
☑ borders

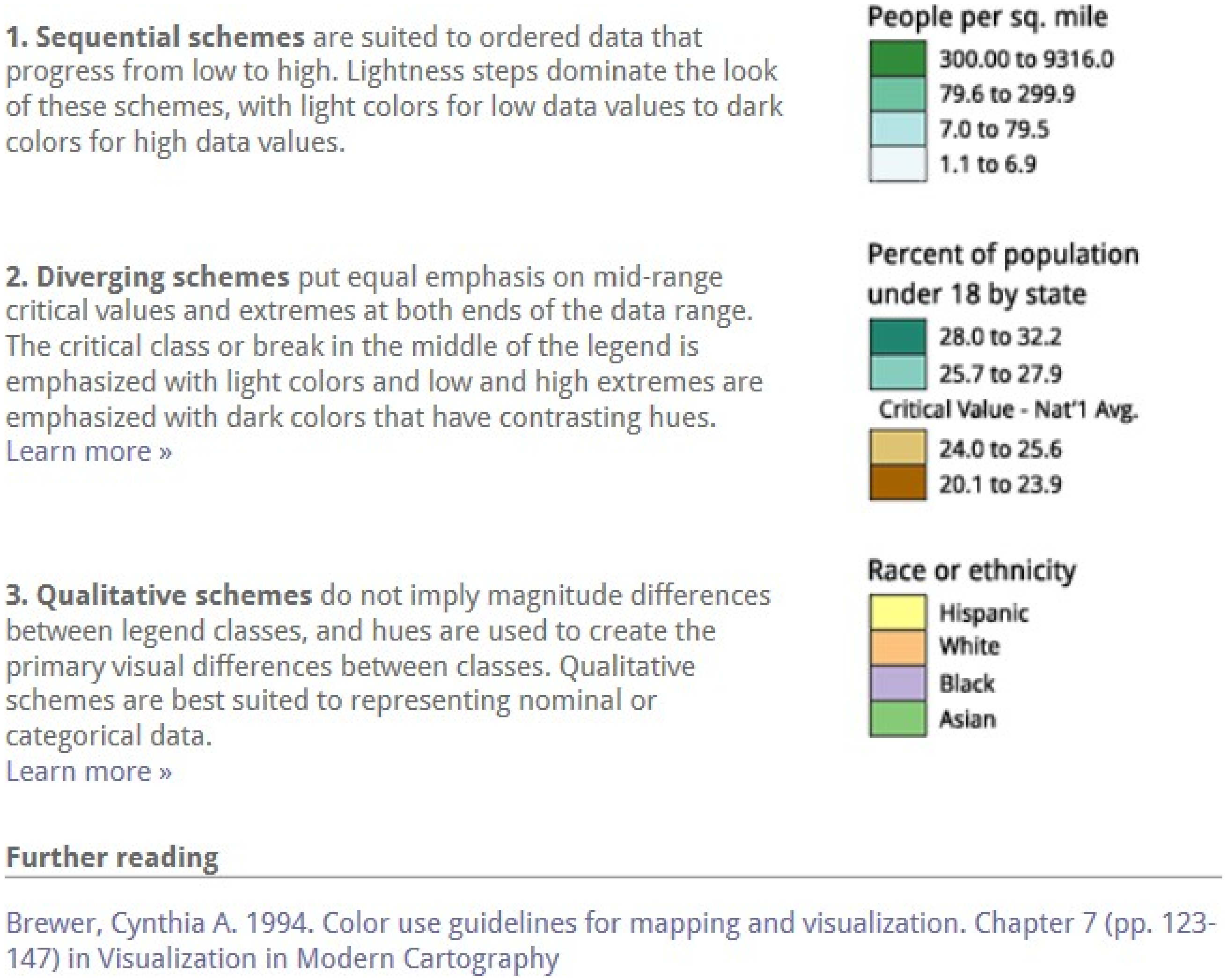
3-class Accent

HEX

🟩	#7fc97f
🟪	#beaed4
🟧	#fdc086

http://colorbrewer2.org/

Colour Palettes Pickers



1. Sequential schemes are suited to ordered data that progress from low to high. Lightness steps dominate the look of these schemes, with light colors for low data values to dark colors for high data values.

People per sq. mile

🟩 300.00 to 9316.0
🟩 79.6 to 299.9
🟦 7.0 to 79.5
⬜ 1.1 to 6.9

2. Diverging schemes put equal emphasis on mid-range critical values and extremes at both ends of the data range. The critical class or break in the middle of the legend is emphasized with light colors and low and high extremes are emphasized with dark colors that have contrasting hues.
Learn more »

Percent of population under 18 by state

🟩 28.0 to 32.2
🟩 25.7 to 27.9
Critical Value - Nat'l Avg.
🟨 24.0 to 25.6
🟫 20.1 to 23.9

3. Qualitative schemes do not imply magnitude differences between legend classes, and hues are used to create the primary visual differences between classes. Qualitative schemes are best suited to representing nominal or categorical data.
Learn more »

Race or ethnicity

🟨 Hispanic
🟧 White
🟪 Black
🟩 Asian

Further reading

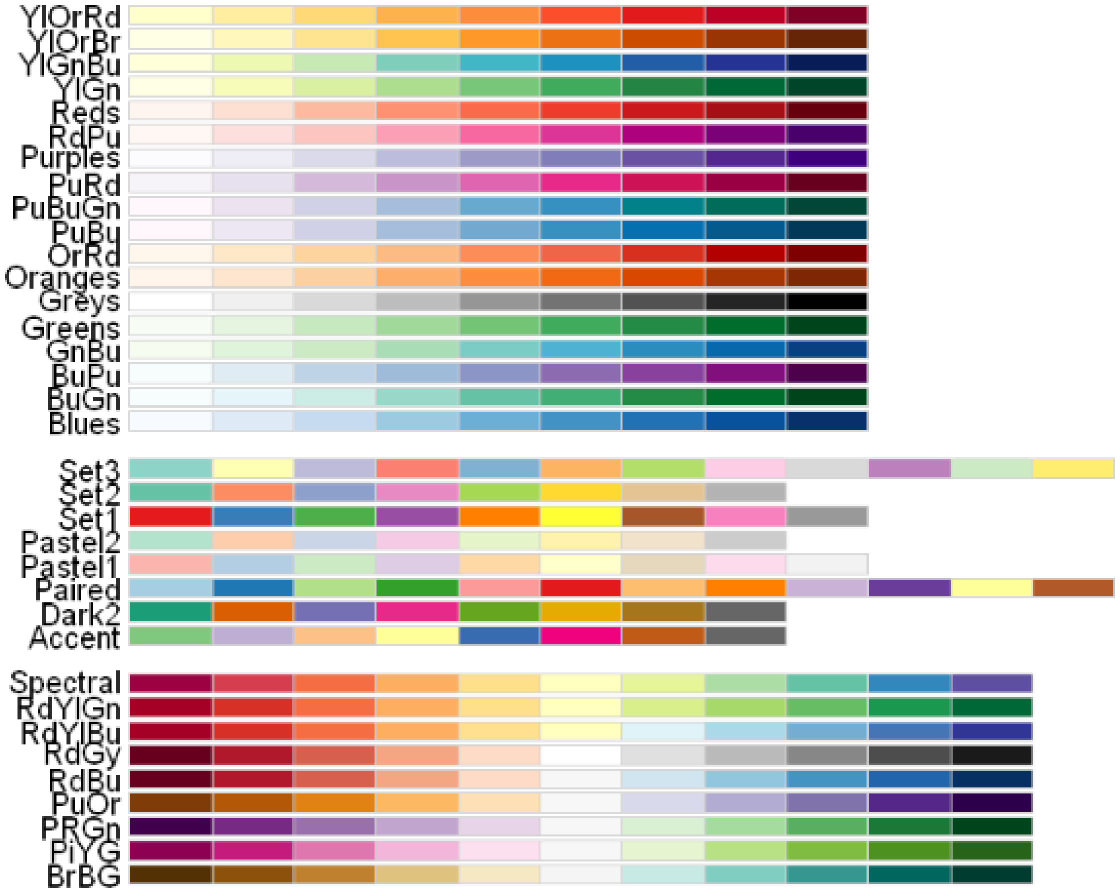
Brewer, Cynthia A. 1994. Color use guidelines for mapping and visualization. Chapter 7 (pp. 123-147) in Visualization in Modern Cartography

http://colorbrewer2.org/

Colour Palettes Pickers

- In R

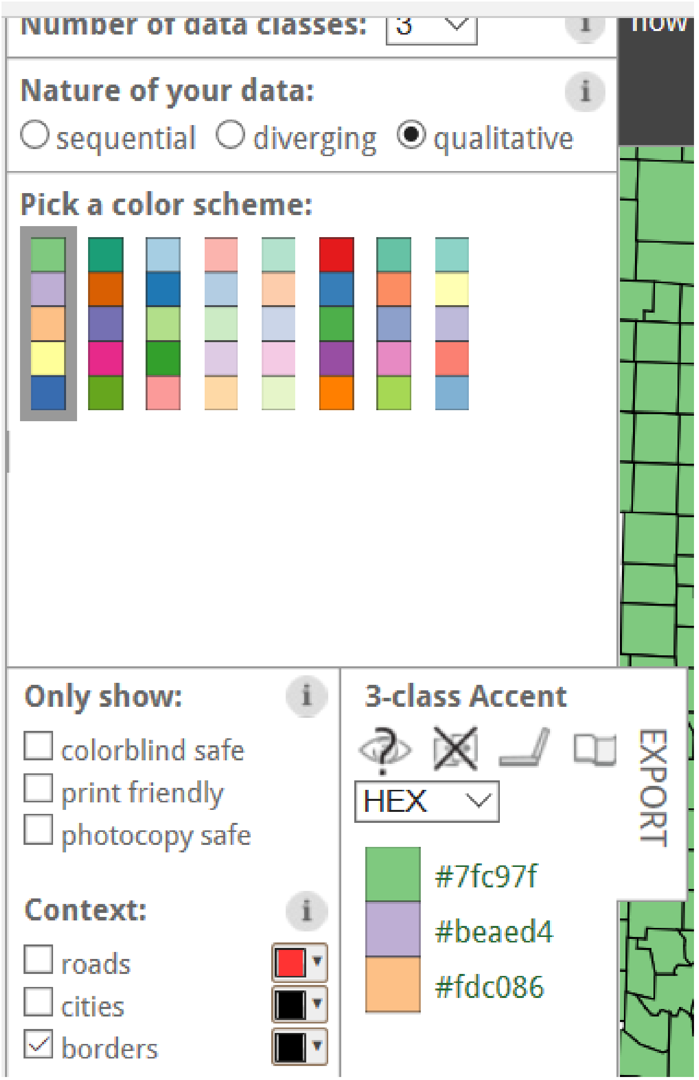
- library(RcolorBrewer)
  display.brewer.all()



http://colorbrewer2.org

Colour Palettes Pickers

- Select the type and nature of your data
- Select the accessibility options
- Use single colours or Palette



Number of data classes: 3

Nature of your data:
○ sequential ○ diverging ● qualitative

Pick a color scheme:

Only show:
□ colorblind safe
□ print friendly
□ photocopy safe

Context:
□ roads
□ cities
☑ borders

3-class Accent

HEX

EXPORT

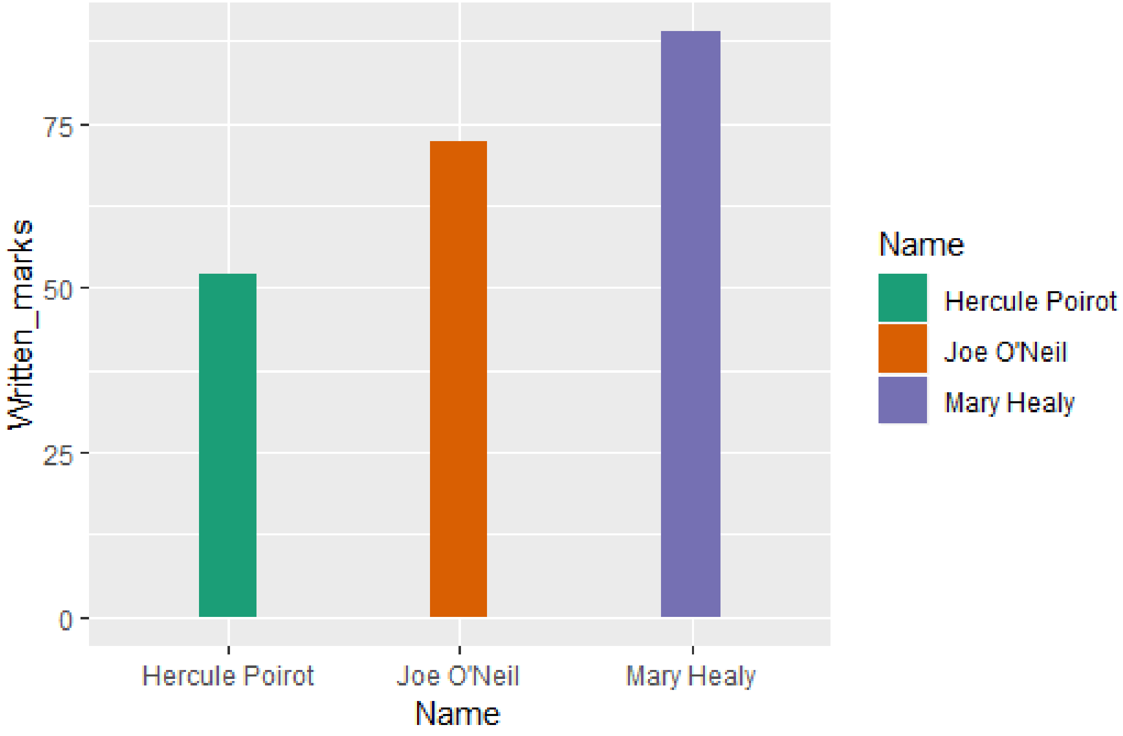
🟩	#7fc97f
🟪	#beaed4
🟧	#fdc086

http://colorbrewer2.org

Colour Palettes Pickers

- ### use colorbrewer to select #specific palettes or colours #type=sequential #scheme=Set2&n=3

- namesAvgmarkplot+ scale_fill_manual(values=c("Hercule Poirot"='#e0f3db', "Joe O'Neil"='#a8ddb5', "Mary Healy"='#43a2ca'))

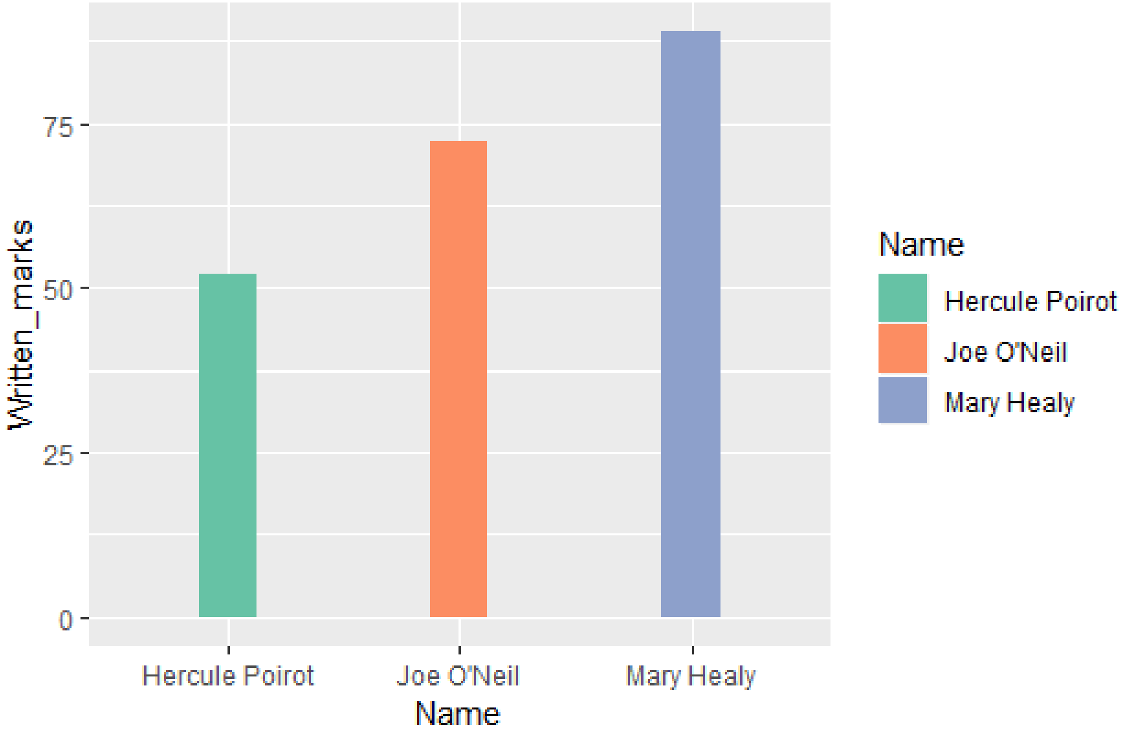


http://colorbrewer2.org

Colour Palettes Pickers

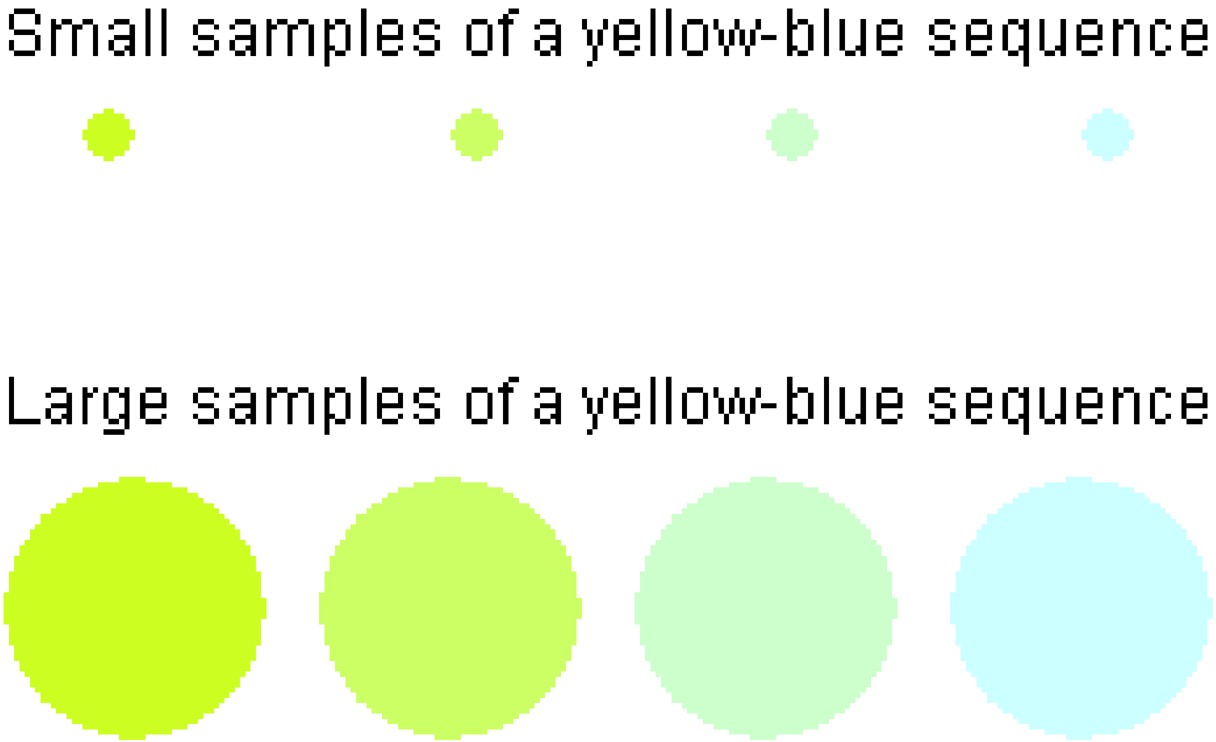
- ### use colorbrewer to select #specific palettes or colours #type=sequential #scheme=Set2&n=3

- namesAvgmarkplot+ scale_fill_brewer(palette=‘Set2')



http://colorbrewer2.org

Small colour Patches More Difficult to Distinguish


Small samples of a yellow-blue sequence
Large samples of a yellow-blue sequence

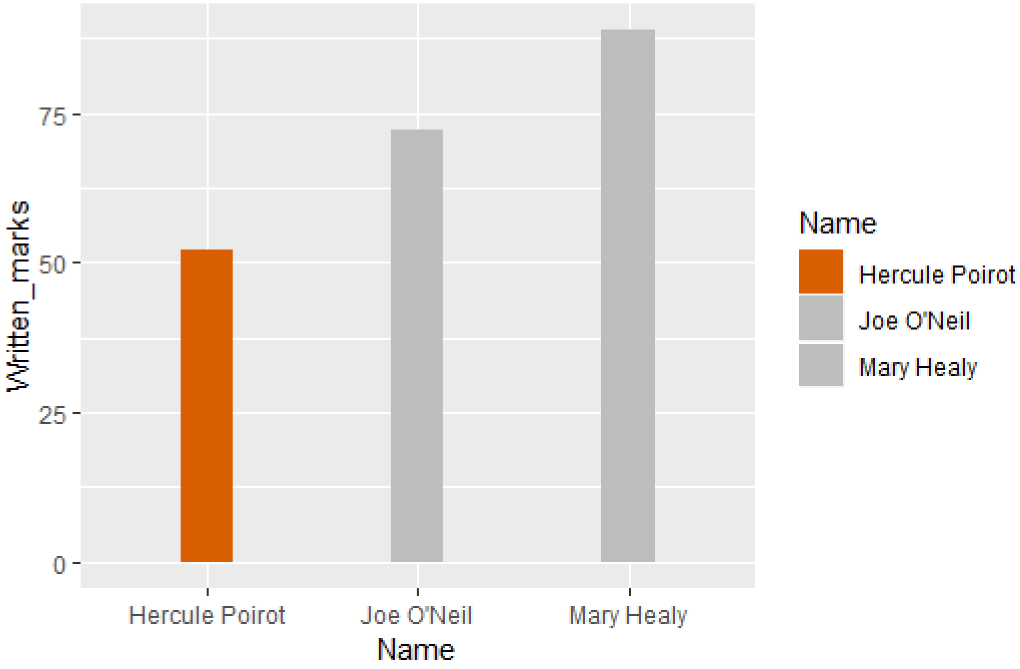
Images from lecture by Terrance Brooke

Spatial Perception of Colour

- Generally speaking, warm colours (reds, oranges, yellows, and browns) will appear to advance to the foreground

- Cool colours (greens, blues, purples, and greys) will appear to recede into the background

- Can be used for focus and editorial salience

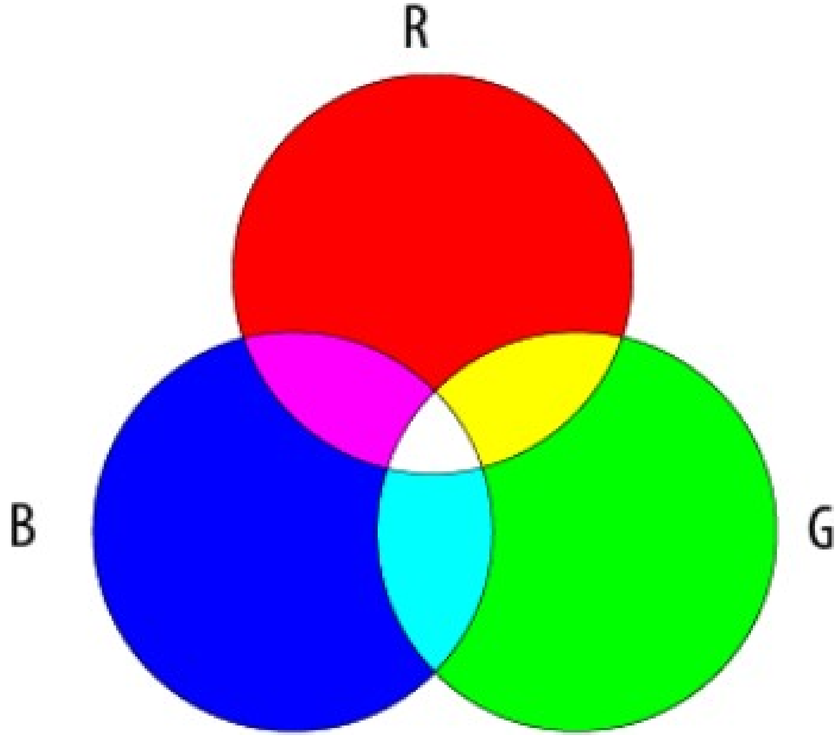


RGB versus CMYK

- Computers and printers use different sets of base colours, or “colour spaces”
  - Computers use RGB (red, green, and blue)
  - Printers use CMYK (cyan, magenta, yellow, and key, or black)
  - Computers and other light-emitting devices use an additive colour model
  - Printers and other ink- or dye-based devices use a subtractive colour model

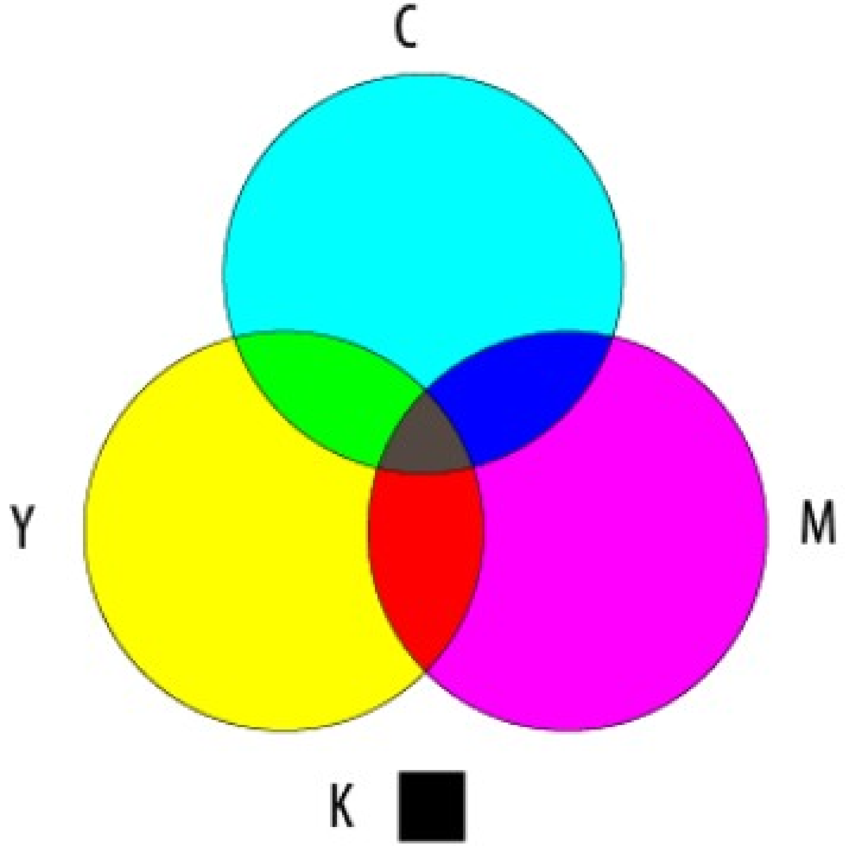
Additive & Subtractive Colour Models

- The additive colour model uses light to project various wavelengths in the spectrum, with red, blue, and green as the primary wavelengths
    - when you combine all three of these, you get white light



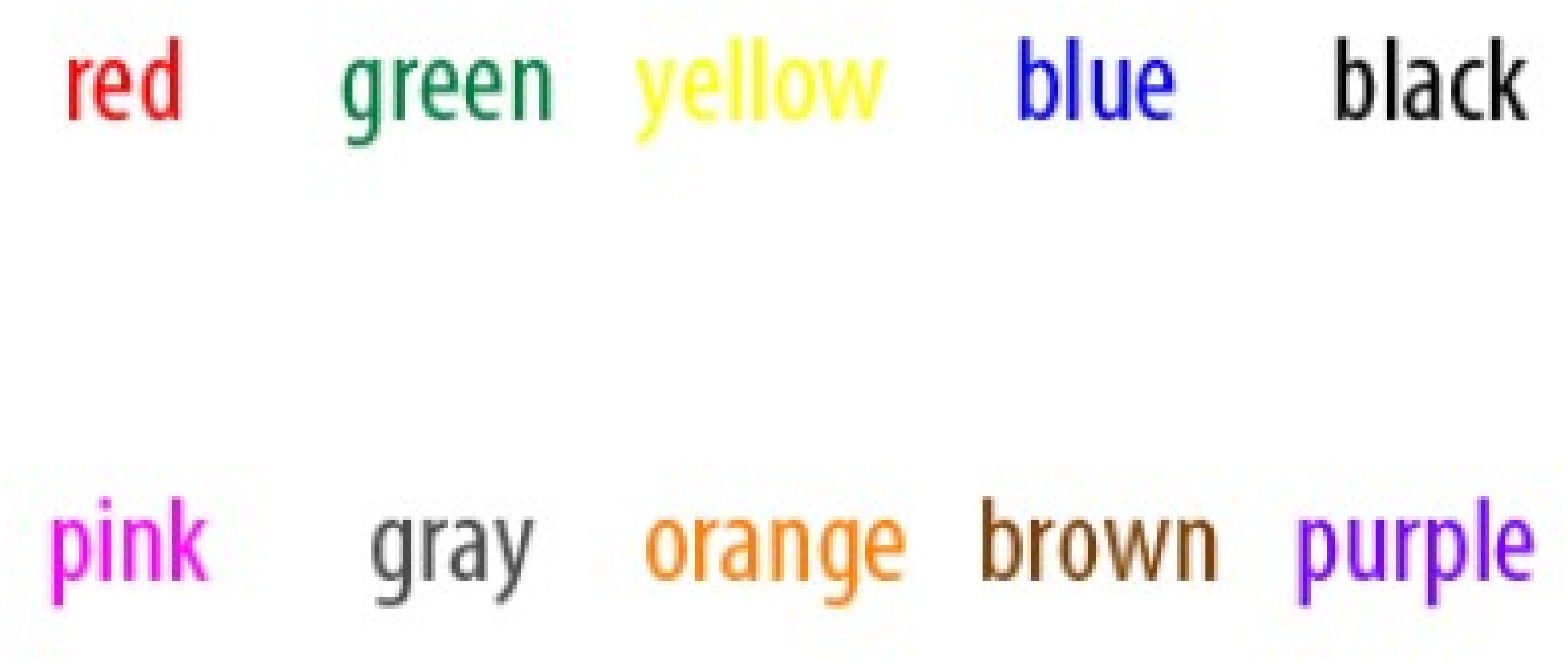
Additive & Subtractive Colour Models

- The subtractive colour model uses pigments to absorb various wavelengths, using cyan, magenta, and yellow as primaries, and black as the pigment that absorbs all wave-lengths



Cognitive Interference & The “Stroop Test”

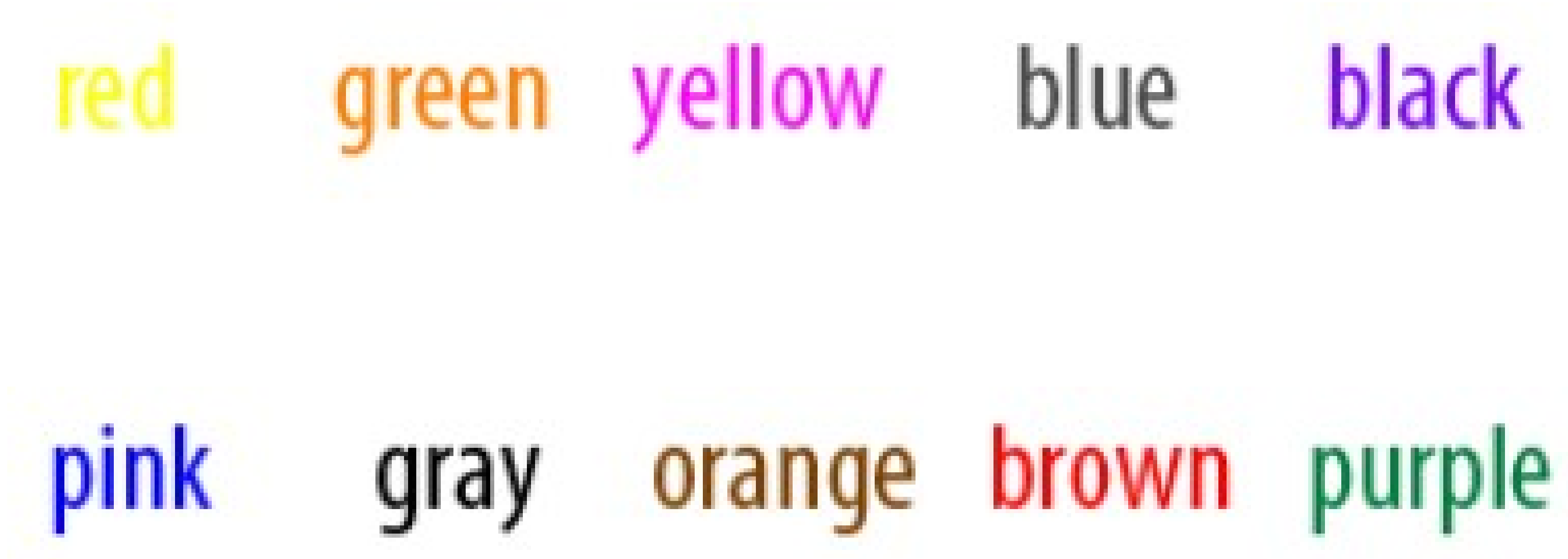
- Sometimes a colour (either the colour itself or the perceived meaning of a colour) can send a message in direct conflict to a message being sent by the element that colour has been applied to (for example, a shape or some text)
- The reader’s brain requires extra time to resolve the disparity - cognitive interference



red green yellow blue black

pink gray orange brown purple

John Ridley Stroop, "Studies of Interference in Serial Verbal Reactions"
Journal of Experimental Psychology, 18, pp 643–662 (1935)



red green yellow blue black

pink gray orange brown purple

John Ridley Stroop, "Studies of Interference in Serial Verbal Reactions"
Journal of Experimental Psychology, 18, pp 643–662 (1935)

Data Visualisation
Lecture Week 9 – Encodings 2

Dr. Cathy Ennis

Apply Your Encodings Well

- Colour
  - Leverage Common Colour Associations
  - Colour Theory
  - Cognitive Interference and the Stroop Test
- Size
  - Conveying Size
  - Comparing Sizes
- Text and Typography
  - Use Text Sparingly
  - Fonts and Hierarchies
  - Beware of All Caps
  - Avoid Drop Shadows
- Shape
  - Cultural Connotations
  - Icons
  - Illusions
- Keys versus Direct Labelling of Data Points

Size

- Size can be used to great advantage to represent the relative importance of entities
- Use size to draw your reader to central, key, or fundamentally important entities

S M L

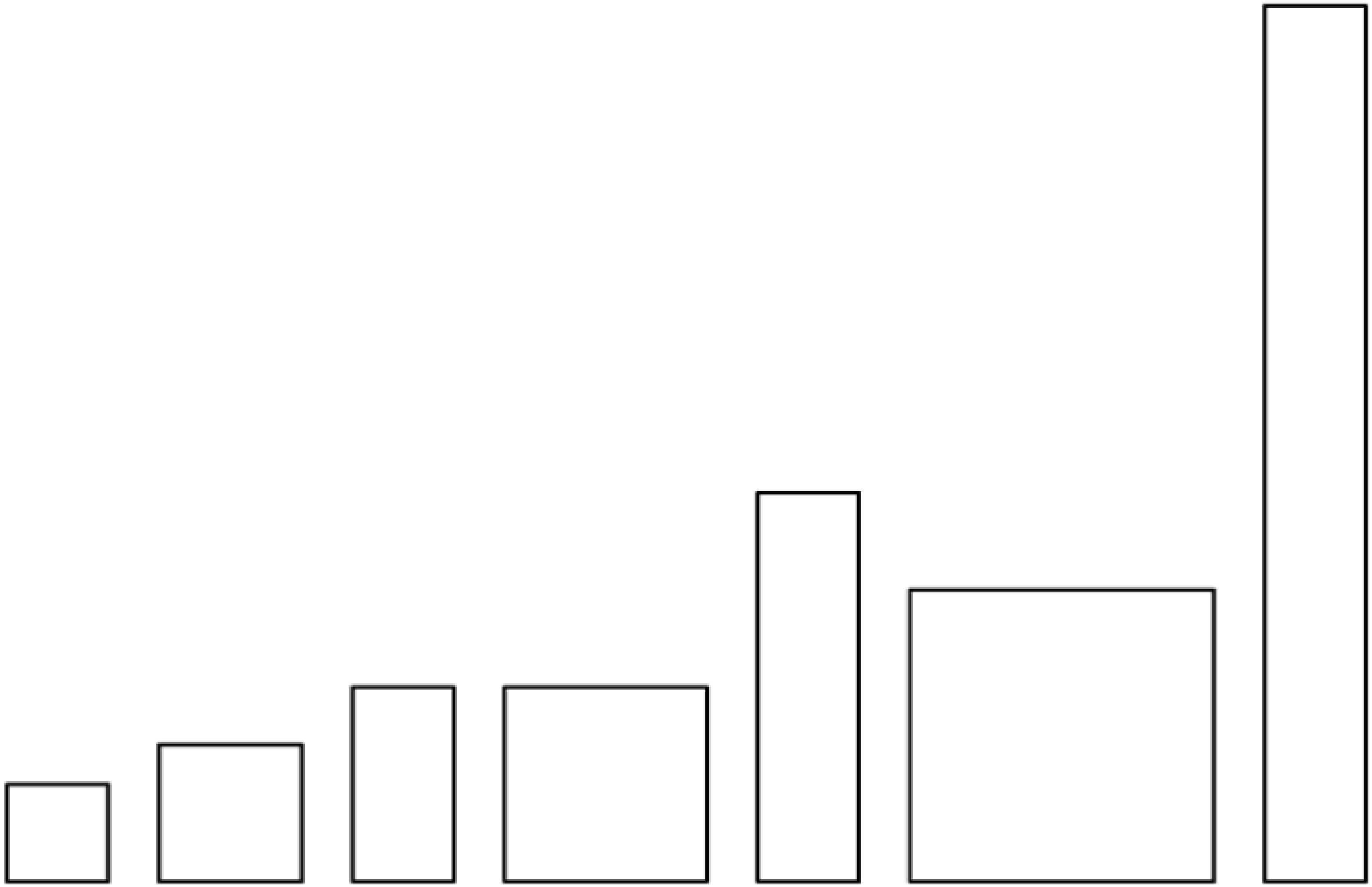
Comparing Sizes

- Humans are quite good at comparing the surface areas of rectangles in situations where the only thing that’s different between them is length

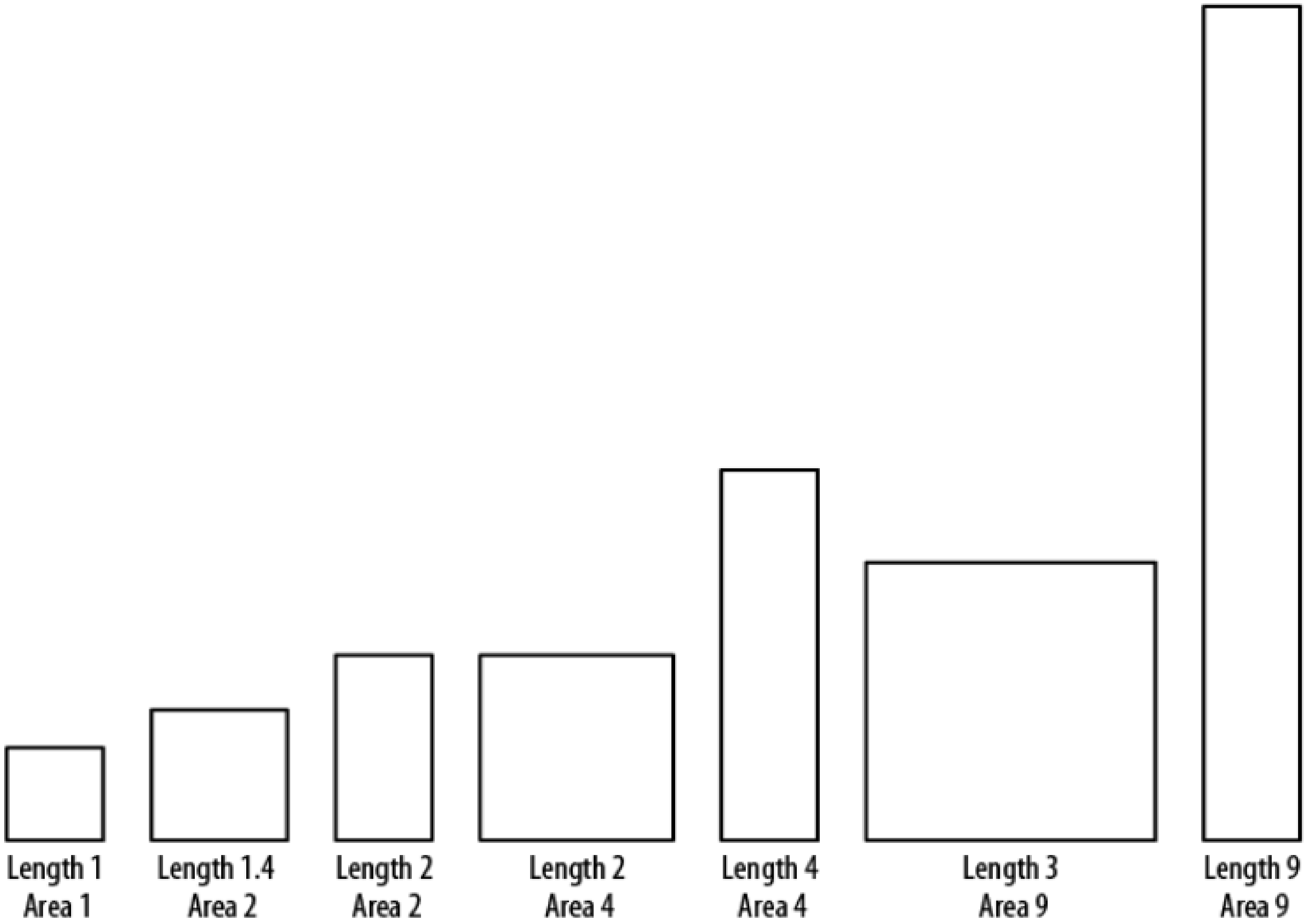
- As long as the width remains constant (think bar graphs)

- When it comes to rectangles where the width and length both change, we don’t judge them as accurately, and we tend to underestimate the differences in size

Comparing Sizes

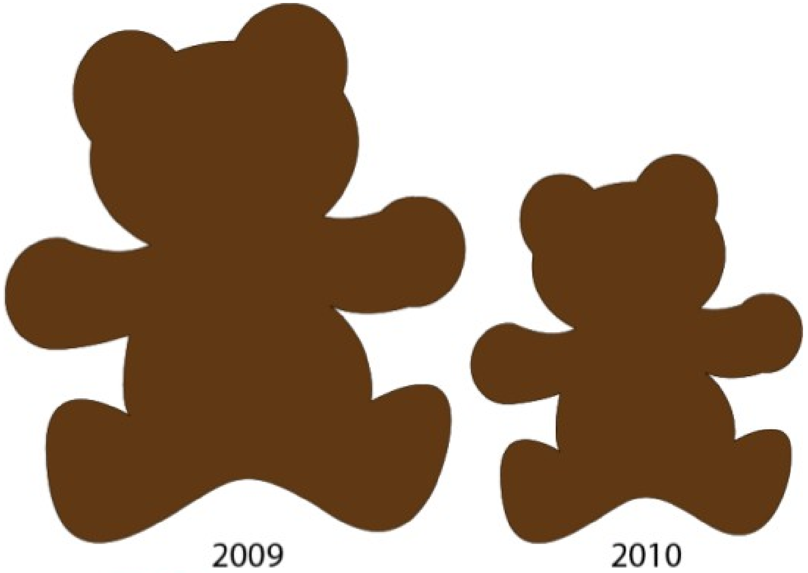


Comparing Sizes

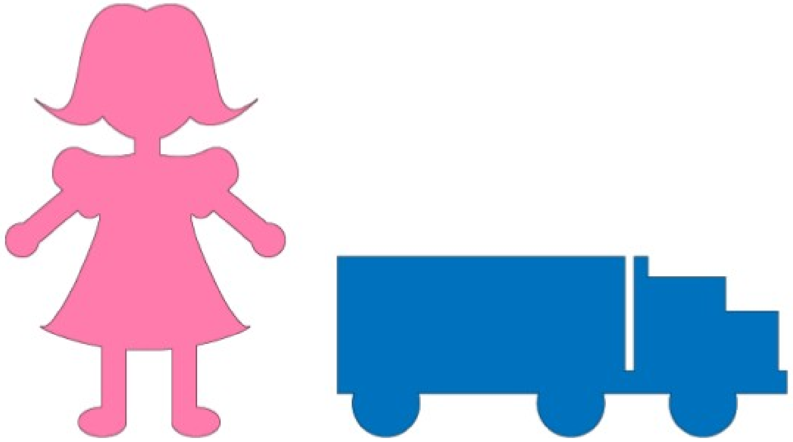


Irregular Surface Areas

- Compare the number of toys donated in two consecutive years



- Compare the number of dolls donated versus the number of trucks donated

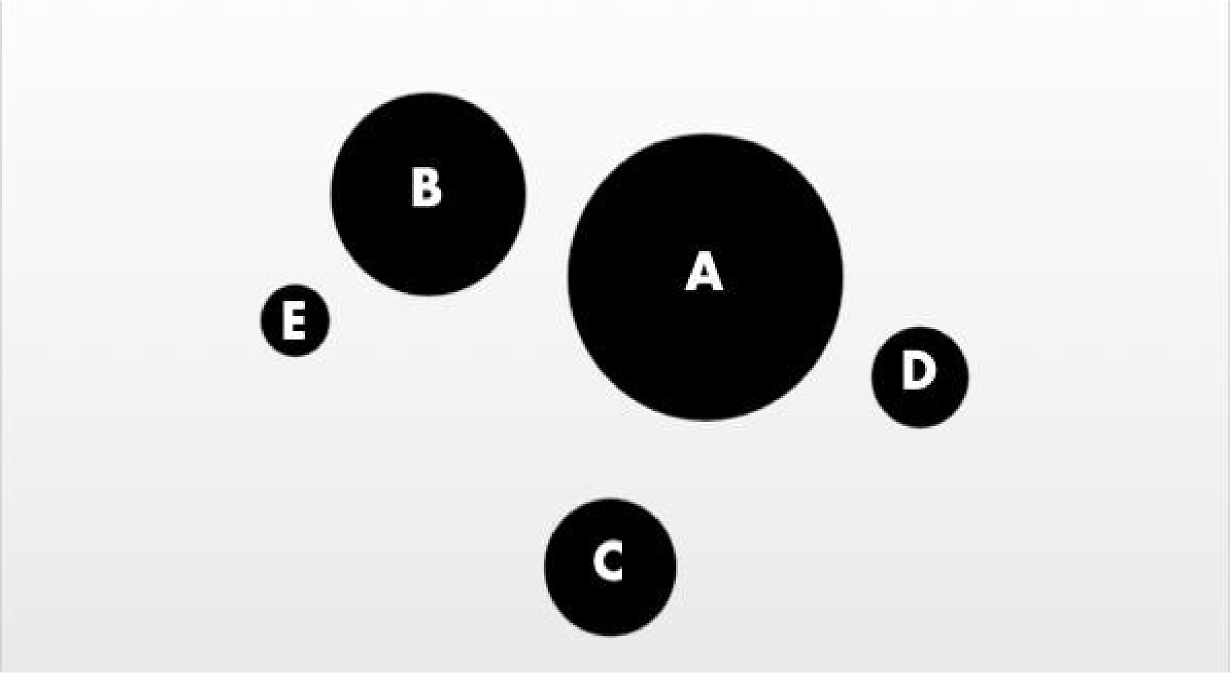


Circles

- Circles are often a poor choice for use in data visualisations
- In particular because our interpretation of their size is often poor
- In particular we run the risk of encoding a data dimension with one dimension of a two-dimensional shape
  - In this example attaching amount to circle diameter, not area

Circles

- If circle area represents company profits would you prefer the profits of company A alone or of companies B, C, D and E combined?



Apply Your Encodings Well

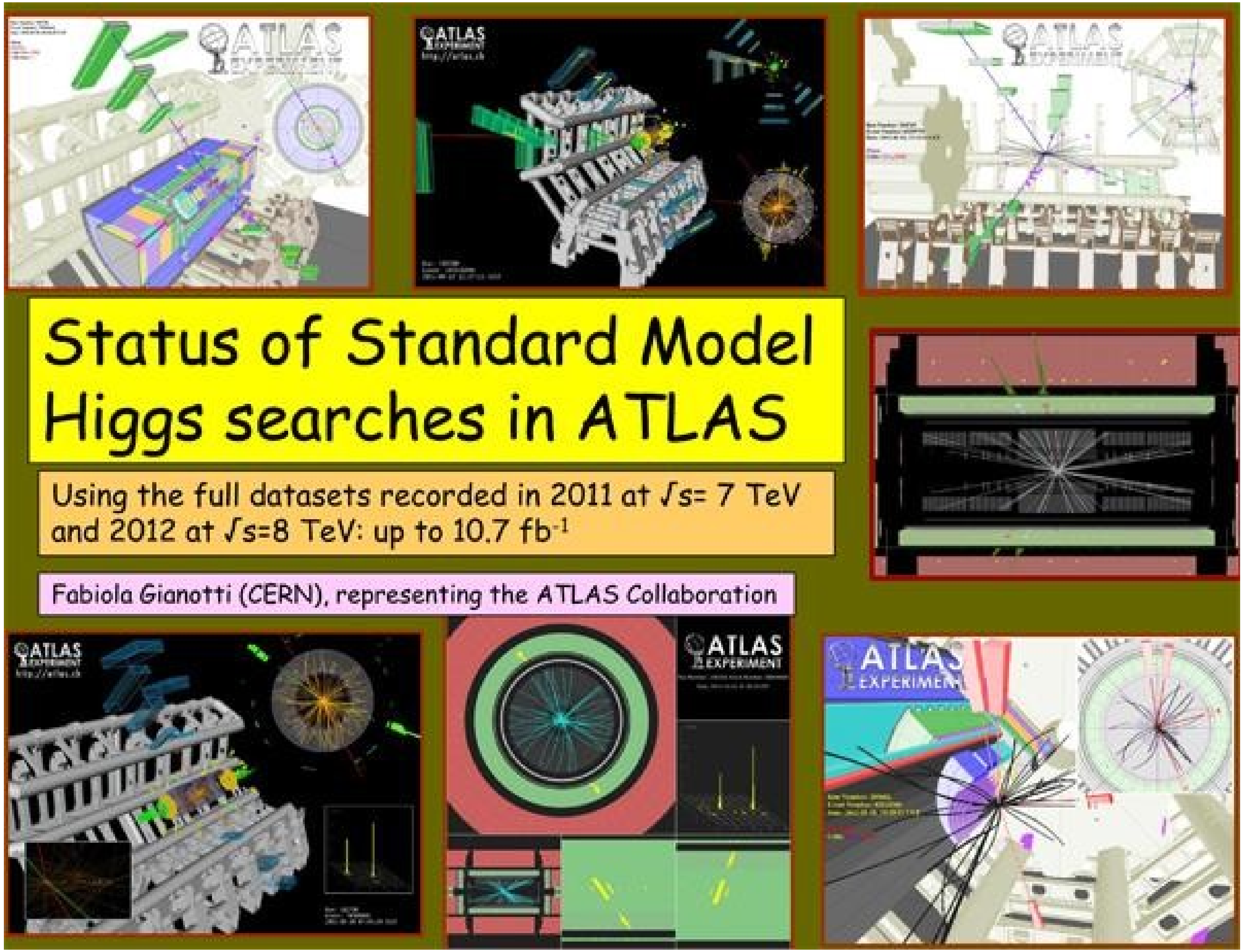
- Colour
  - Leverage Common Colour Associations
  - Colour Theory
  - Cognitive Interference and the Stroop Test
- Size
  - Conveying Size
  - Comparing Sizes
- Text and Typography
  - Use Text Sparingly
  - Fonts and Hierarchies
  - Beware of All Caps
  - Avoid Drop Shadows
- Shape
  - Cultural Connotations
  - Icons
  - Illusions
- Keys versus Direct Labelling of Data Points

Text & Typography

- Visualisations need text for titles, axis labels, or other kinds of labels or indicators
- Many visualizations treat words as second-class citizens
- We need to carefully consider how labels and numbers interact with the visual elements, and what they do or do not say
- Using text sparingly can be effective

Fonts & Hierarchies

- Disagreements over serif versus sans serif fonts are endless
  - beyond the scope of this lecture
- Remember, your goal is to make things clear and easy to navigate for your reader
  - avoid using fancy or trendy fonts just because you can
- Use a consistent font throughout. **Different fonts will call the readers attention and signal meaning**
- Avoid drop shadows, all caps and other silly embellishments



Status of Standard Model Higgs searches in ATLAS

Using the full datasets recorded in 2011 at $\sqrt{s}$= 7 TeV and 2012 at $\sqrt{s}$=8 TeV: up to 10.7 fb^{-1}

Fabiola Gianotti (CERN), representing the ATLAS Collaboration

ATLAS EXPERIMENT
ATLAS EXPERIMENT http://atlas.ch
ATLAS EXPERIMENT
ATLAS EXPERIMENT http://atlas.ch
ATLAS EXPERIMENT
ATLAS EXPERIMENT

https://www.newscientist.com/article/mg21628962-600-tricksy-type-how-fonts-can-mess-with-your-mind/

Apply Your Encodings Well

- Colour
  - Leverage Common Colour Associations
  - Colour Theory
  - Cognitive Interference and the Stroop Test
- Size
  - Conveying Size
  - Comparing Sizes
- Text and Typography
  - Use Text Sparingly
  - Fonts and Hierarchies
  - Beware of All Caps
  - Avoid Drop Shadows
- Shape
  - Cultural Connotations
  - Icons
  - Illusions
- Keys versus Direct Labelling of Data Points

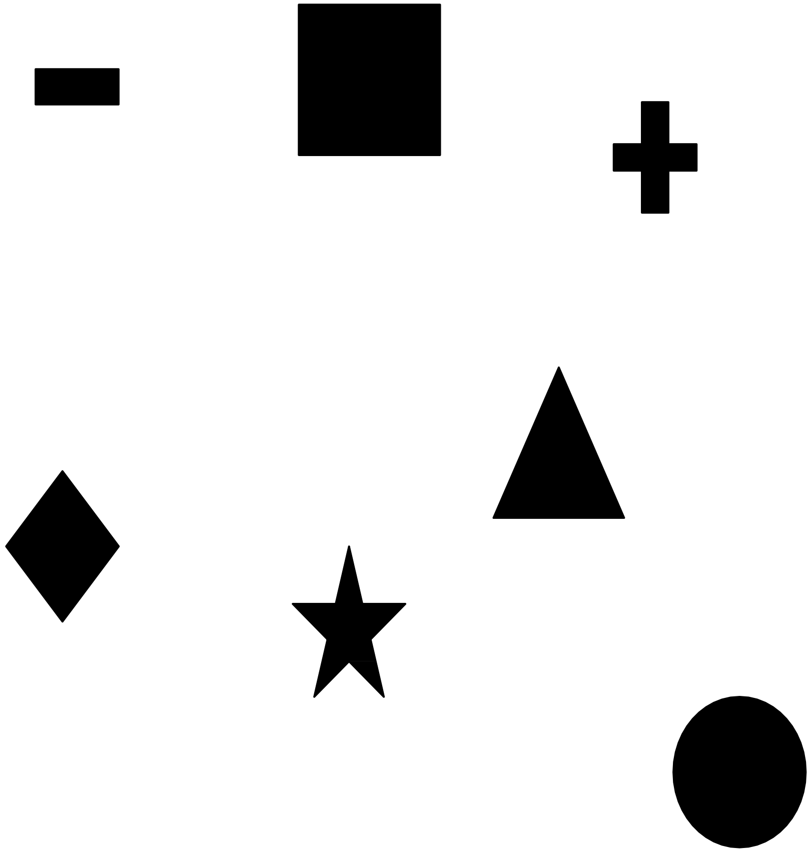
Shape

- Shape is a very useful property for labelling or encoding categories

- Because of the huge variety of shapes available, and the general ease of differentiating them, shape can be much more evocative than some other properties

- The expressive nature of shape has the potential to be both very useful, and very **distracting or misleading**

Simple Shape Sets

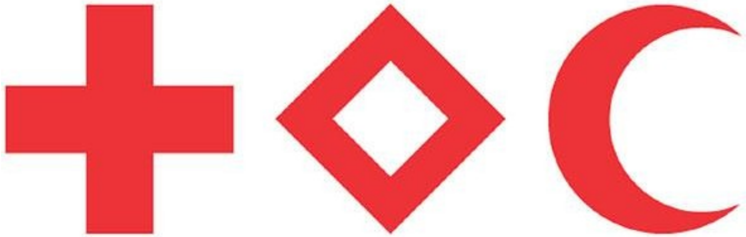


- In R

```
g+geom_point(aes(shape=name))
g+scale_shape(solid=F)
g+scale_shape_manual(values=c(0:3))
```

Cultural Connotations

- Shape can have significant cultural implications:
  - think of the various meanings of crosses, crescents, stars, and shields



- One must be extremely careful to not offend the reader or to convey unintended meaning with shape
- Remember that some readers will not share your assumptions and conventions about shape

Illusions

- Beware of the unintended confluence of shapes forming compound images

- Gestalt Theory
  - Gestalt theory deals with how our mind perceives wholes out of incomplete elements.
  - Things are affected by where they are and by what surrounds them
  - Parts identified individually may have a different characteristics to the whole (gestalt means “organized whole”)

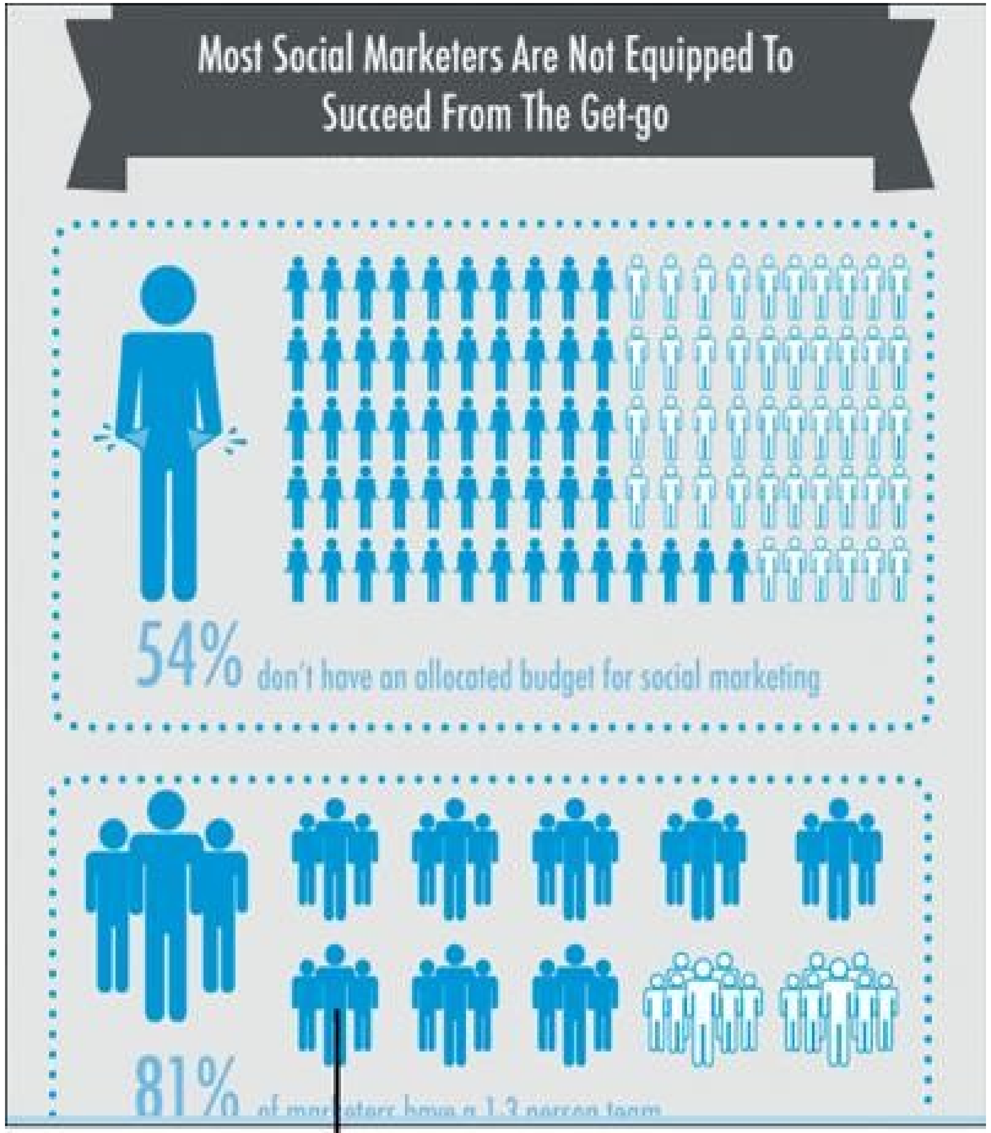
- We look for recognizable shapes in the combination of abstract forms
  - some of these may have meanings conflicting with your intended message

Gestalt

- Gestalt theory is made up of several principles using the concepts of:
  - proximity
  - similarity
  - closure
  - continuation
  - figure/ground
- These concepts describe how the human brain sees visual information
- Designers who understand this theory can develop visuals that communicate information in the most effective ways

Gestalt - Proximity

- When items are placed in close proximity, people assume that they’re in the same group because they’re close to one another and apart from other groups.


Most Social Marketers Are Not Equipped To Succeed From The Get-go
54% don't have an allocated budget for social marketing
81% of marketers have a 1-3 person team

Grouped together

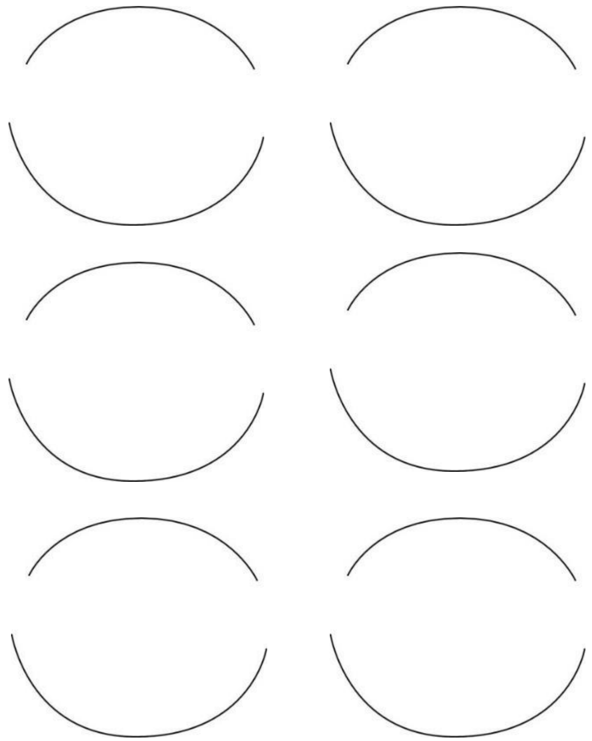
Gestalt - Similarity

- When items look the same, people perceive them to be of the same type. We naturally assume that shapes that look the same are related

- When you create a data visualisation and you keep items together that look the same, you make it easy for someone to understand that those items represent a group

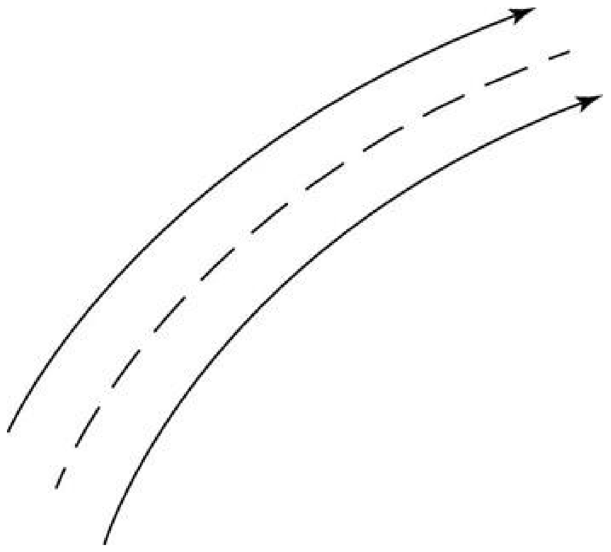
Gestalt - Closure

- Our eyes tend to add any missing pieces of a familiar shape
- If two sections are taken out of a circle, people still perceive the whole circle



Gestalt - Continuation

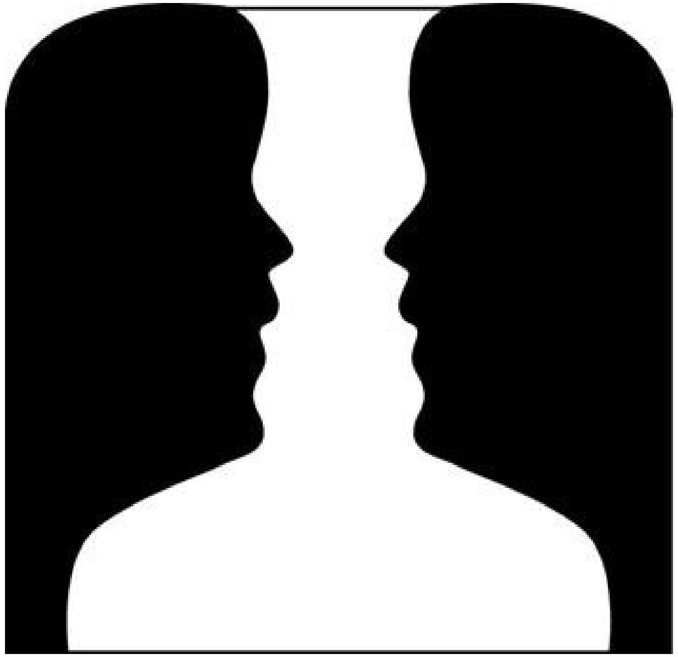
- If people perceive objects as moving in a certain direction, they see them as continuing to move that way



Gestalt – Figure/ground

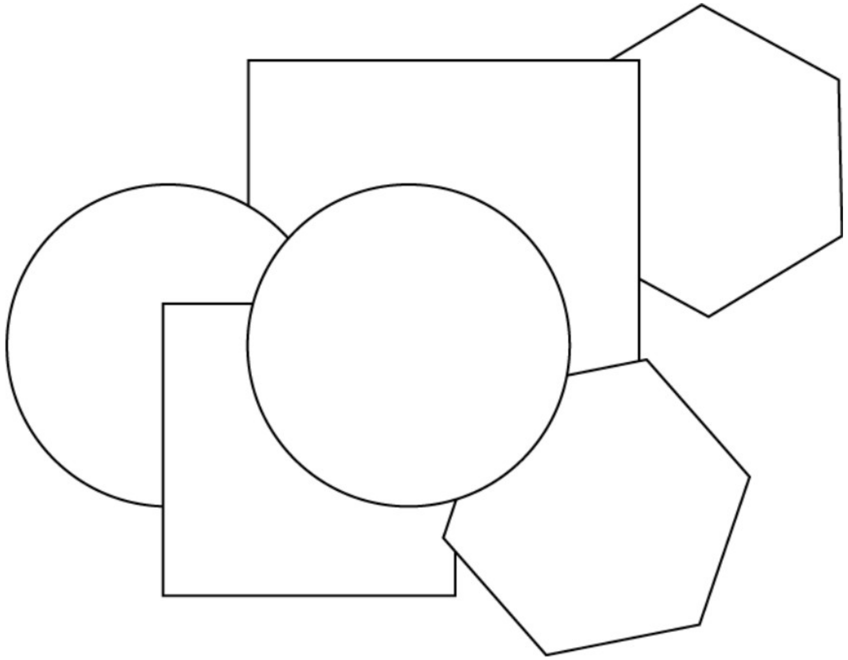
- Depending on how people look at a picture, they see either:
  - the figure (foreground) or
  - the ground (background)

- as standing out



Illusions

- Layering - if your visualization gets so crowded that shapes begin to overlap each other, you may cause the illusion of depth where none is intended



Apply Your Encodings Well

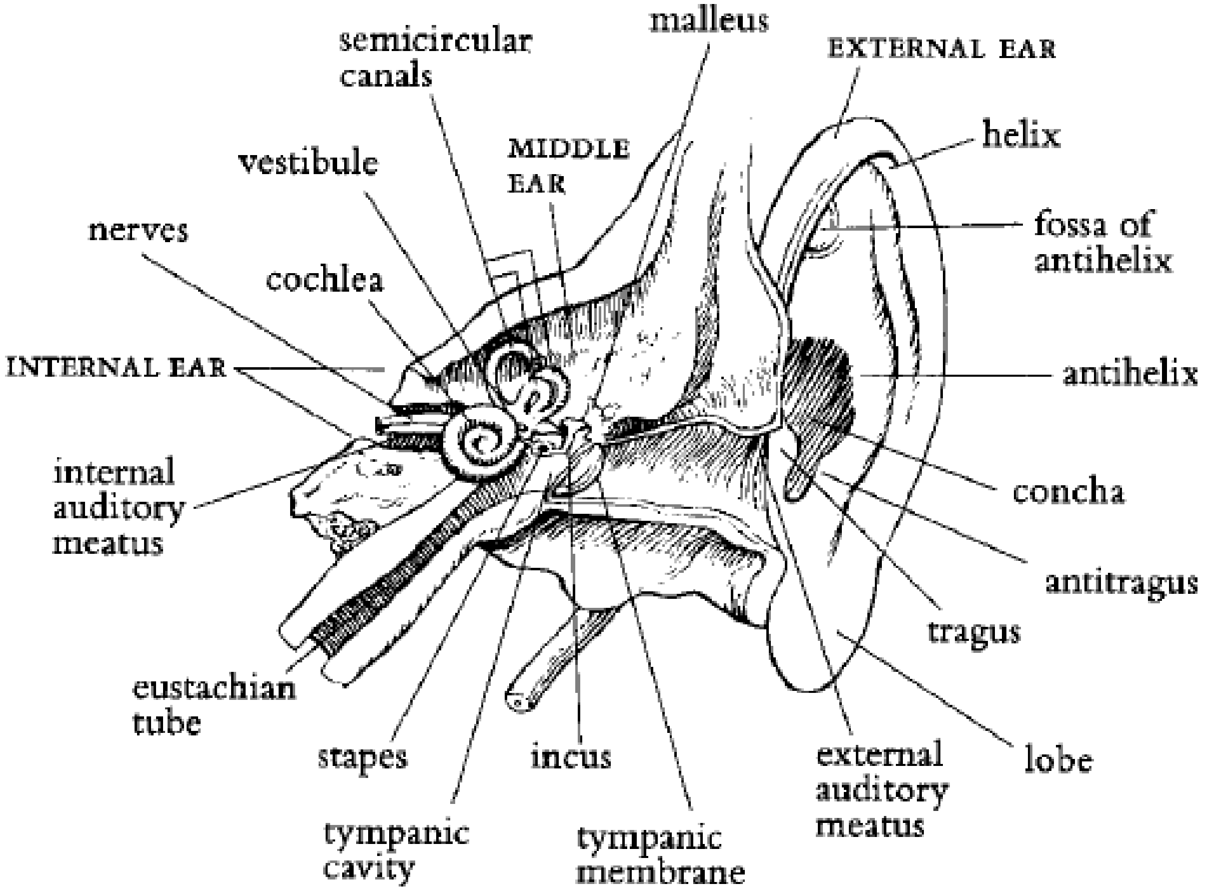
- Colour
  - Leverage Common Colour Associations
  - Colour Theory
  - Cognitive Interference and the Stroop Test
- Size
  - Conveying Size
  - Comparing Sizes
- Text and Typography
  - Use Text Sparingly
  - Fonts and Hierarchies
  - Beware of All Caps
  - Avoid Drop Shadows
- Shape
  - Cultural Connotations
  - Icons
  - Illusions
- Keys versus Direct Labelling of Data Points

Keys Versus Direct Labelling of Data Points

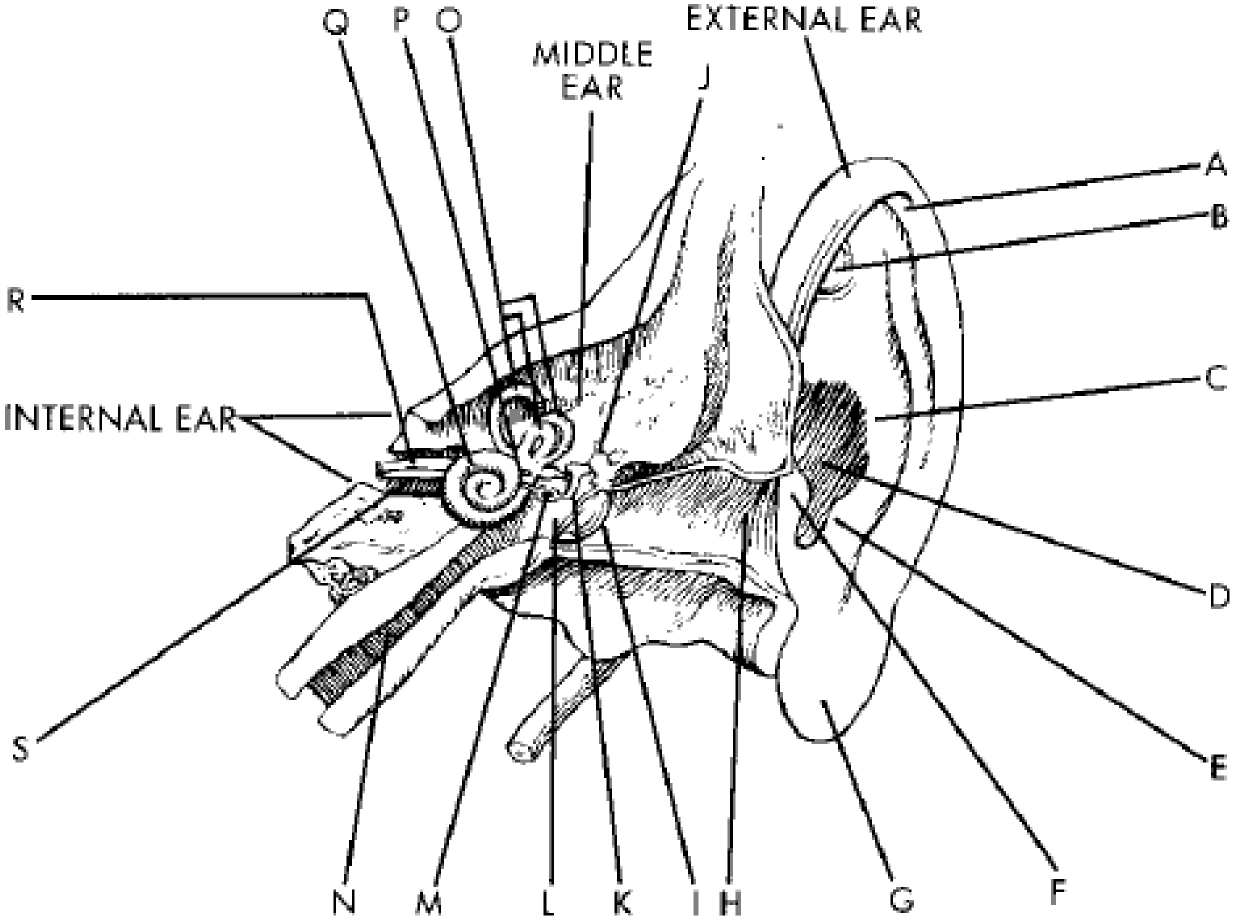
- Legends and labels are an integral part of visualisation design, however:
    - Legends impose extra cognitive burden on your reader – looking back and forth
    - Labels are more accessible, but can obfuscate

- Depends on how many total data points you’re dealing with, and how many possible values exist for those data points
    - Fewer data points -> Direct labelling
    - Fewer possible values -> Key (regardless of the number of data points)
    - Densely located data points and many values -> Labelling only some values

Keys Versus Direct Labelling of Data Points



Keys Versus Direct Labelling of Data Points



Pitfalls

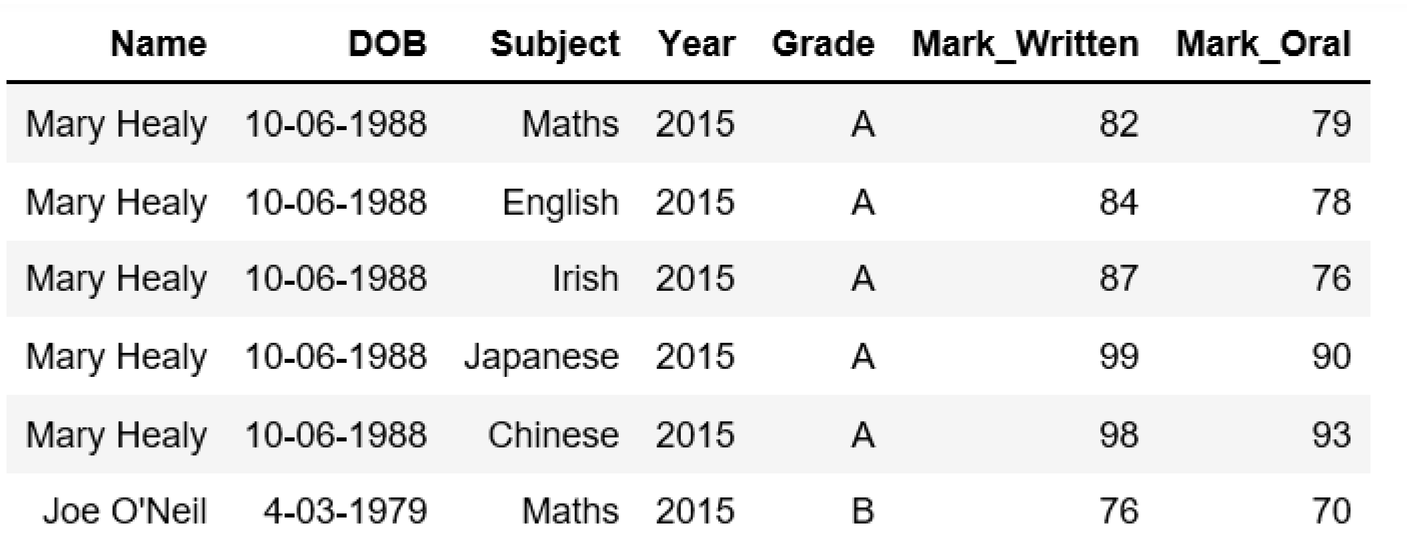
- Primary goal of visualization = communication
- Any element that hinders—rather than helps—the reader, needs to be changed or removed e.g.,
    - labels and tags that are in the way
    - colours that confuse or simply add no value
    - uncomfortable scales or angles
- Efficiency matters - if you're wasting a viewer's time or energy, they're going to move on without receiving your message!

Apply Your Encodings Well

- Colour
  - Leverage Common Colour Associations
  - Colour Theory
  - Cognitive Interference and the Stroop Test
- Size
  - Conveying Size
  - Comparing Sizes
- Text and Typography
  - Use Text Sparingly
  - Fonts and Hierarchies
  - Beware of All Caps
  - Avoid Drop Shadows
- Shape
  - Cultural Connotations
  - Icons
  - Illusions
- Keys versus Direct Labelling of Data Points

Exercise 1 - R

- An analysis of student grades with R
- In this lab, you will use ggplot and sqldf to analyse student marks in a very small file.



Name	DOB	Subject	Year	Grade	Mark_Written	Mark_Oral
Mary Healy	10-06-1988	Maths	2015	A	82	79
Mary Healy	10-06-1988	English	2015	A	84	78
Mary Healy	10-06-1988	Irish	2015	A	87	76
Mary Healy	10-06-1988	Japanese	2015	A	99	90
Mary Healy	10-06-1988	Chinese	2015	A	98	93
Joe O'Neil	4-03-1979	Maths	2015	B	76	70

ggplot structure

```
myplot <- ggplot(data= yourdataset, aes(x=yourx, y =youry))
#this begins your plot by adding the data
```

<u>Examples of extra layers</u>

```
myplot +geom_point()      #this adds a geometry to your plot (scatter plot in the example)

myplot +geom_point(aes(colour=dimension))        #geom layer can be customized

myplot +geom_bar()+scale_fill_brewer(palette=’Reds’)          #customizing the colour
palette

myplot +geom_bar(colour=dimension) +scale_fill_manual(values=c(‘blue’,’red’))

#customize colours manually
```

ggplot structure

- **myplot <- ggplot(data= yourdataset, aes(x=yourx, y =youry))**
- #this begins your plot by adding the data

- <u>Examples of extra layers</u>

- myplot+coord_map(projection=”ortho”, orientation= c(41,-74,0)) #map

projection myplot +theme_classic() #applies a predefined theme to the plot

- myplot +labs(title=”graph title”, x=”xaxistitle”, y=”yaxistitle”) #labels

- myplot +facet_wrap(dimension) #creates small multiples based on dimension

An analysis of student grades with R Examples

- Inspecting data, Dealing with missing values.
- Question 1. What are the average results in written exams across all subjects and all years per student?
- Question 2. What are the average results in oral exams across all subjects and all years per student?
- Question 3. What are the average results in the written exams per student and year?
- Creating custom functions.

An analysis of student grades with R Exercises

1. What are the total marks (oral plus written divided by two) for each student for each subject? (2 marks)

2. What is the relationship between age and mark? (2 marks)

3. Did any students do better on their written compared with their oral (or vice versa)? (2 marks)

4. What subject obtained the best results on average? (2 marks)

5. What are the average results in oral exams across all subjects and all years per student? (2 marks)

Thanks To

- Marisa Llorens-Salvador, John McAuley, Colman McMahon and Brian Mac Namee for an earlier version of these lecture notes