# Lecture 9
# Text, Pre-attentive Attributes, Gestalt, Illusions

Data Visualization · 1-DAV-105

Lecture by Broňa Brejová

Acknowledgement: materials inspired by lectures from Martina Bátorová in 2021

# Visualizing text data

# Visualizing text data

Working with natural text is difficult

- Complex grammar, ambiguous meaning, synonyms, etc.
- Lot of machine learning research
- Nonetheless sometimes simple statistics on frequencies of words or groups of words can be useful
- Usually we remove *stop words* (frequent words such as "and", "is"...) and apply *lemmatization* (convert inflected words to canonical form, such as "seen" -> "see")
- In homework you will see the use on nltk library

# Word clouds

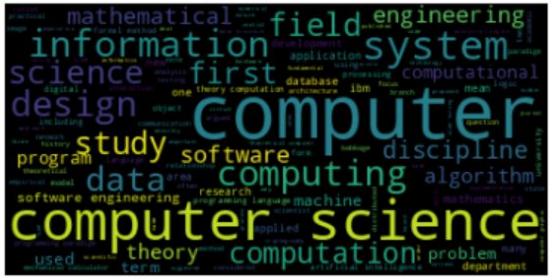
- Display the most common words / phrases from a text
- Size of words grows with frequency
- Arranged to be visually pleasing

Here wikipedia articles on Data science, Computer science and Bioinformatics

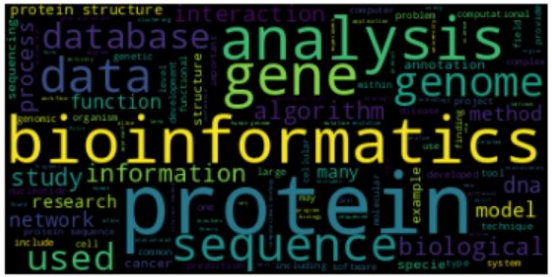
Aside: Data science article was shortest

Data science


Computer science




Bioinformatics



# Word clouds

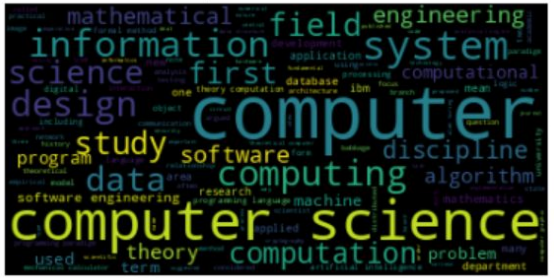
- Display the most common words / phrases from a text
- Size of words grows with frequency
- Arranged to be visually pleasing

Not the best option for understanding/comparing word frequencies.

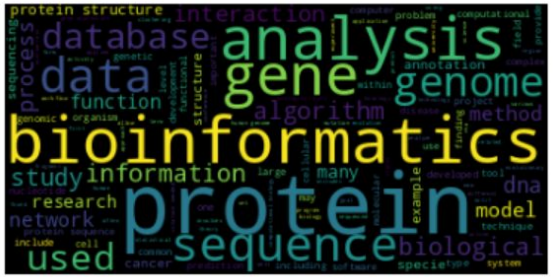
**Why?**

Data science


Computer science

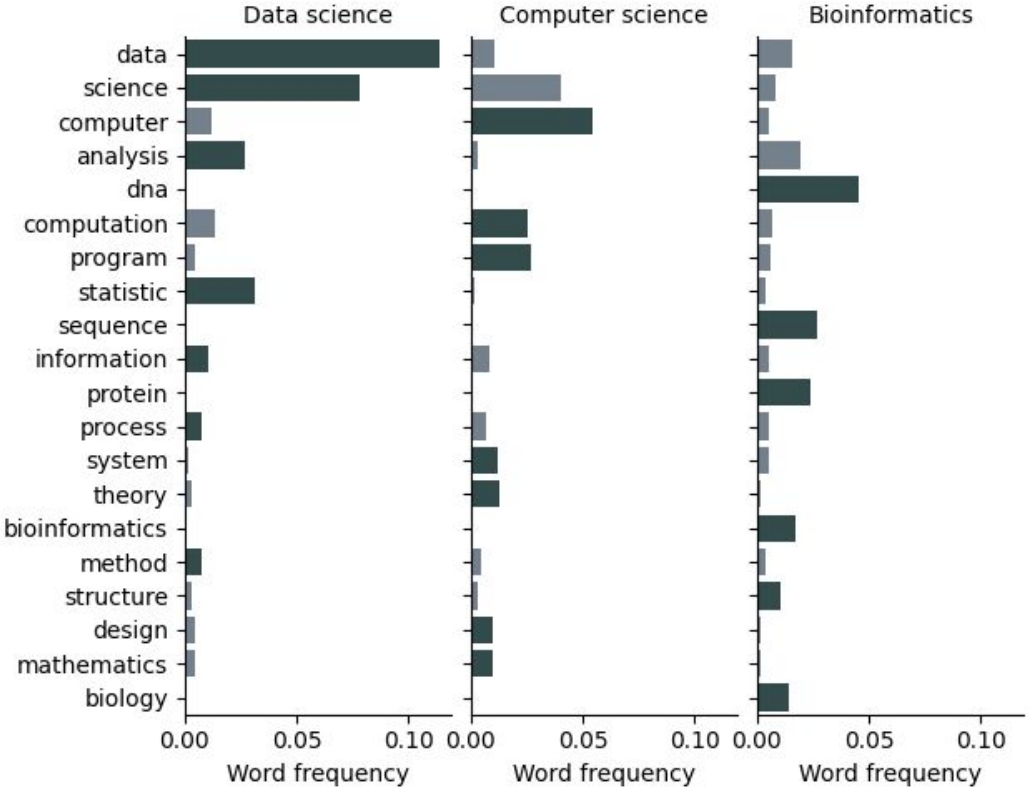



Bioinformatics



# Word clouds

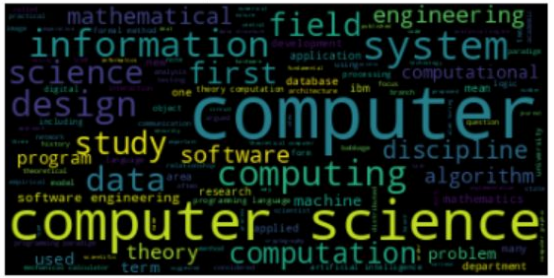
- Display the most common words / phrases from a text
- Size of words grows with frequency
- Arranged to be visually pleasing
- Not the best option for understanding/comparing word frequencies
  - Longer words take greater area making them more prominent
  - Different word positions it difficult to locate the same word and also to compare their size
- You can display word frequencies using **bar graphs** and other plot types for more precise comparison (examples)



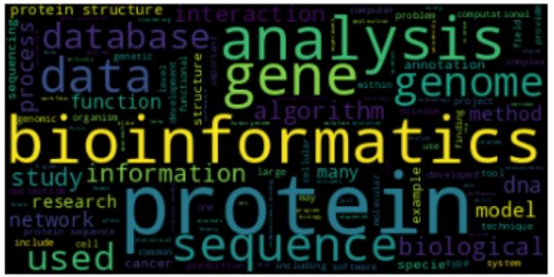

Frequency of words and synonyms;
field with highest frequency highlighted

Data science


Computer science




Bioinformatics



# Tag cloud

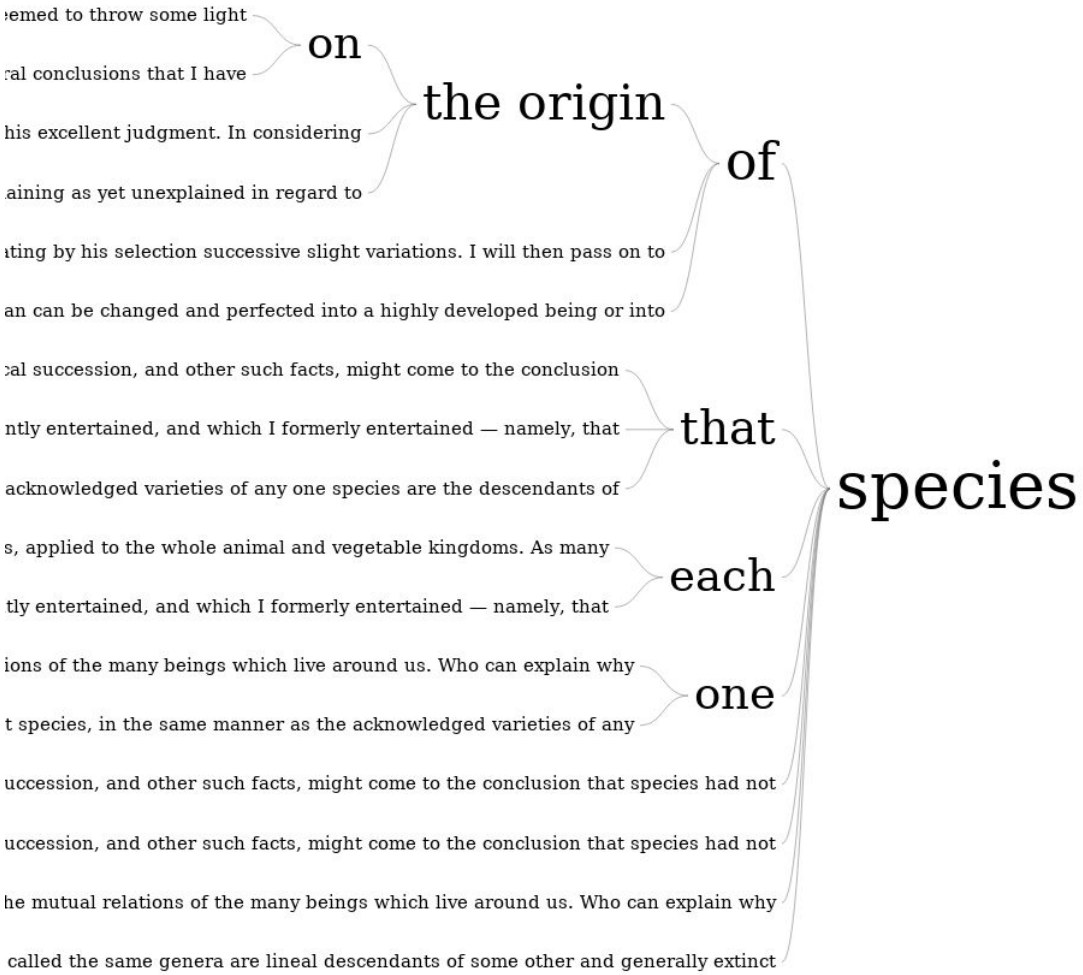
- Endings of German city names typical for individual regions
- Combination of a word cloud and map
- Figure from [Reckziegel et al 2018]

# Word tree

Shows with words most often follow or precede a given word using a hierarchy

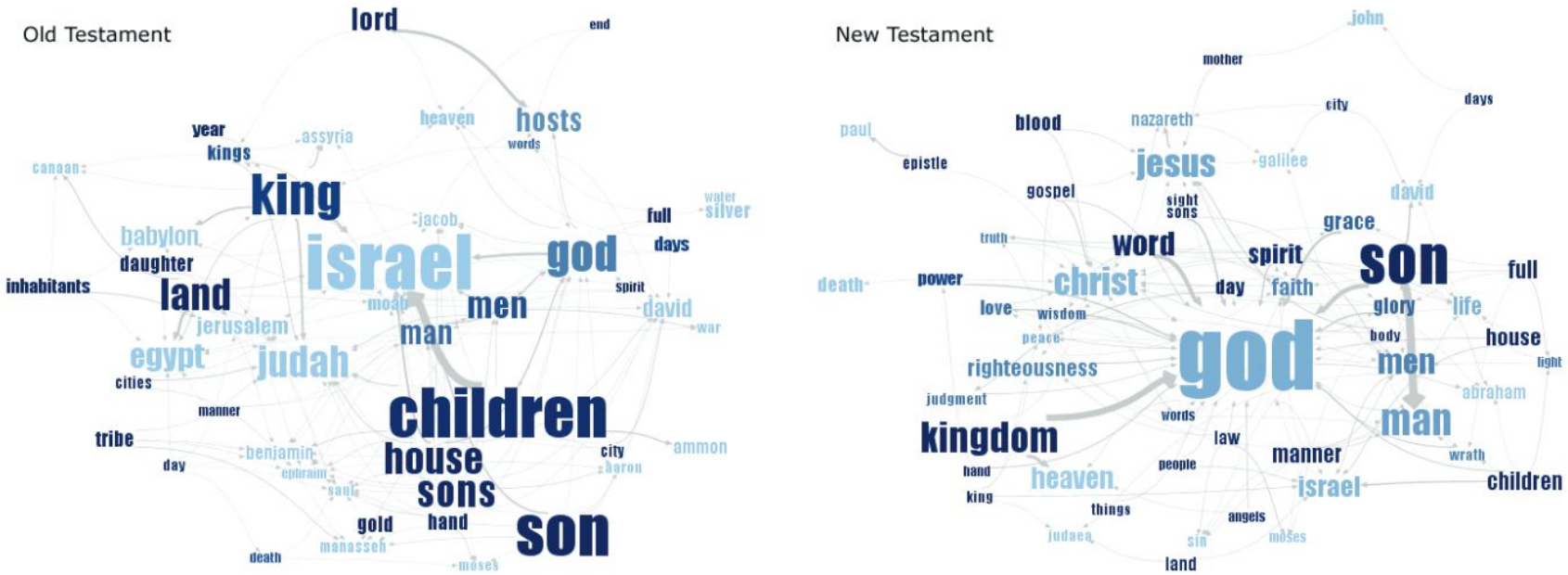
Text: Introduction to The Origin of Species by Charles Darwin, 1859, 1872

Figure source



# Phrase nets

Phrases of type "X of Y", X connected to Y in a graph; source [van Ham et al 2009]



# Text visualization: additional sources

- Courses Data Management (2L), Principles of Data Science (3Z)
- Text visualization browser https://textvis.lnu.se/
- Lecture from Univ. of Washington
- Drawing Elena Ferrante's Profile: Finding out who is Elena Ferrante, bestselling Italian author (My Brilliant Friend) by comparing word frequencies etc. (see e.g. page 100)

# Back to visual perception

# Human visual perception

What happens when we look at the figure?

- The **light** from the screen / projector hits the retinas of our **eyes**
- Photoreceptor cells **transduce** (convert) this signal into nerve impulses
- In the brain:
  - detection of **basic features**
  - recognition of **patterns**
  - interpretation, assignment of **meaning**

# Human visual perception

What happens when we look at the figure?

- The **light** from the screen / projector hits the retinas of our **eyes**
- Photoreceptor cells **transduce** (convert) this signal into nerve impulses
- In the brain:
  - detection of **basic features**
  - recognition of **patterns**
  - interpretation, assignment of **meaning**

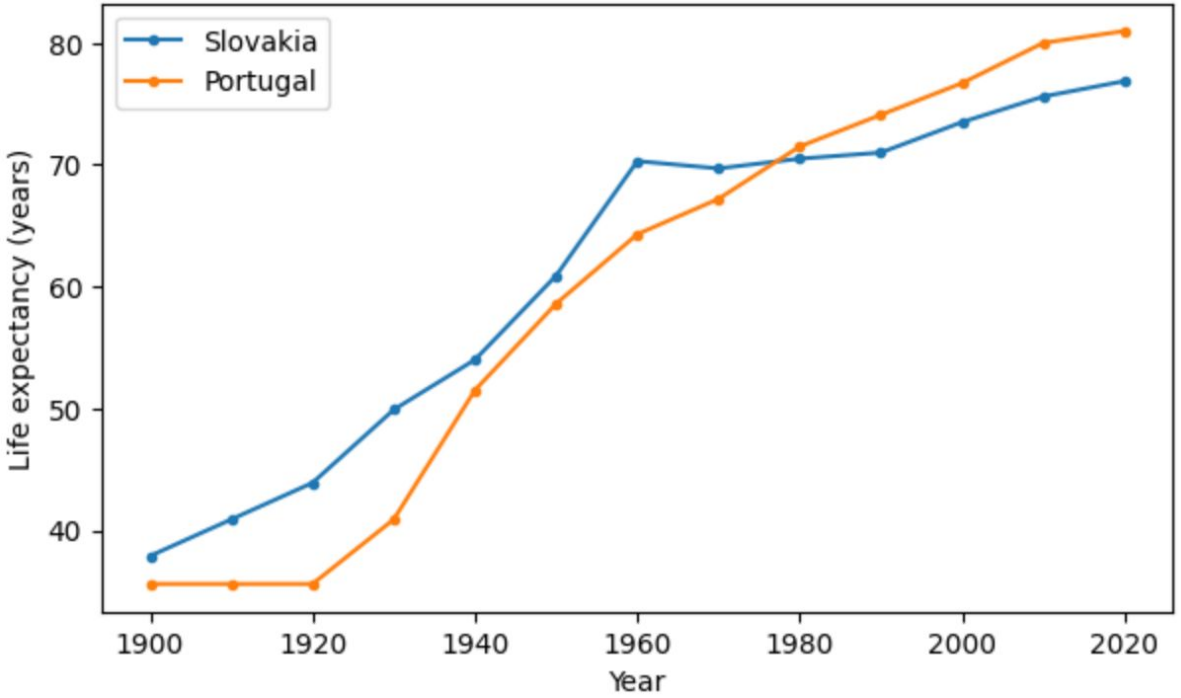
**Today**: Detection of features and patterns, use for visualization

**Note:** Human visual perception is very good for detecting **motion** (danger/prey). This is relevant for animated visualization, but not covered today.

# In which period of time was life expectancy higher in Slovakia than in Portugal?

| Country | 1900 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010 | 2020 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Slovak Republic** | 37.9 | 40.9 | 43.9 | 49.9 | 54.0 | 60.9 | 70.3 | 69.7 | 70.5 | 71.0 | 73.5 | 75.6 | 76.9 |
| **Portugal** | 35.6 | 35.6 | 35.6 | 40.9 | 51.5 | 58.6 | 64.3 | 67.2 | 71.5 | 74.1 | 76.7 | 80.0 | 81.0 |

# In which period of time was life expectancy higher in Slovakia than in Portugal?



How many copies of digit six do you see?

1014508

2530653

6821550

3702967

8622988

What about now?

1014508

2530**6**53

**6**821550

3702**6**7

8**6**22988

# What about Slovakia vs Portugal in this table?

| | 1900 | 1910 | 1920 | 1930 | 1940 | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010 | 2020 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Slovakia | **37.9** | **40.9** | **43.9** | **49.9** | **54.0** | **60.9** | **70.3** | **69.7** | 70.5 | 71.0 | 73.5 | 75.6 | 76.9 |
| Portugal | 35.6 | 35.6 | 35.6 | 40.9 | 51.5 | 58.6 | 64.3 | 67.2 | **71.5** | **74.1** | **76.7** | **80.0** | **81.0** |

# Pre-attentive attributes

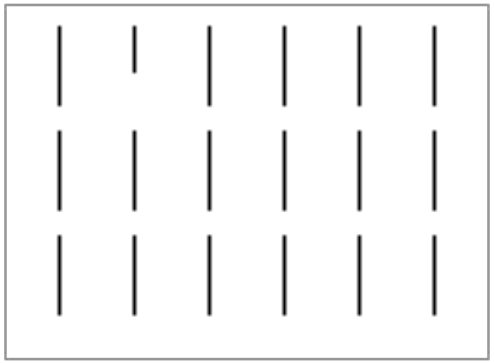
- Features of the seen objects detected by our brain very **fast**
- Prior to and **without** the need of conscious **awareness**
- Brain uses them to **guide attention** / gaze to interesting parts of the scene
- Their correct use allows fast and effortless understanding of our visualizations

**Next:**
Examples of important pre-attentive attributes (form, color, position)
following Few 2009

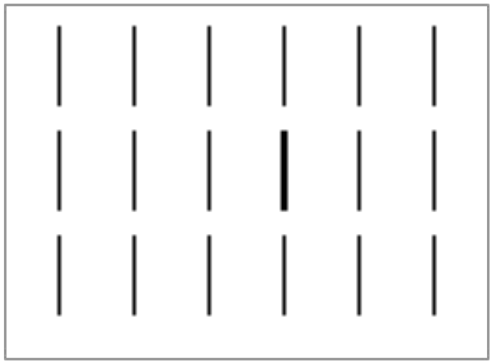
See also https://www.csc2.ncsu.edu/faculty/healey/PP/

# Pre-attentive attributes: form

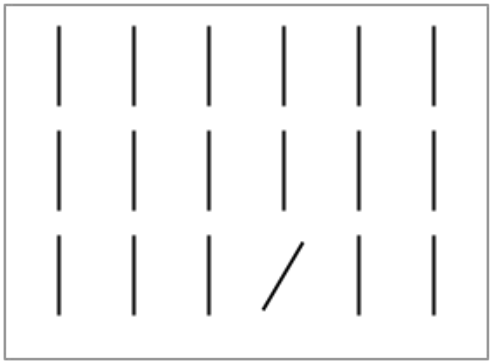
We can quickly distinguish one object that differs from the rest



Length



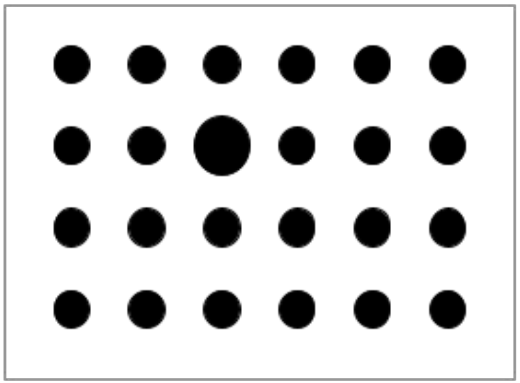
Width



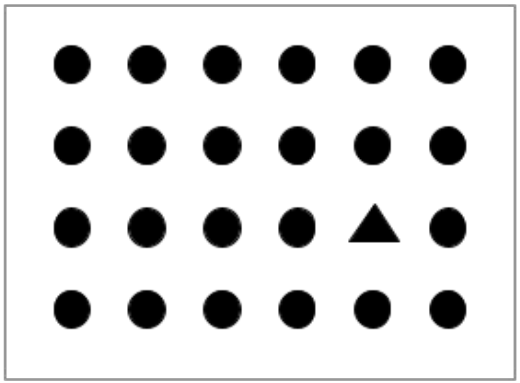
Orientation

# Pre-attentive attributes: form

We can quickly distinguish one object that differs from the rest



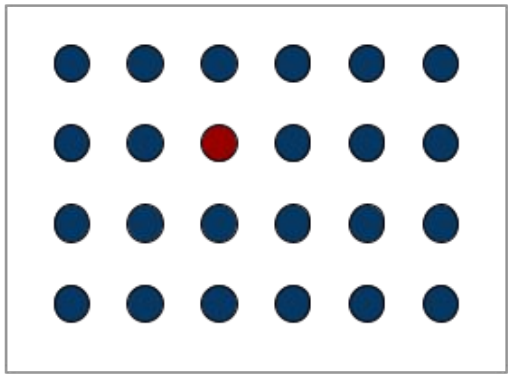
Size



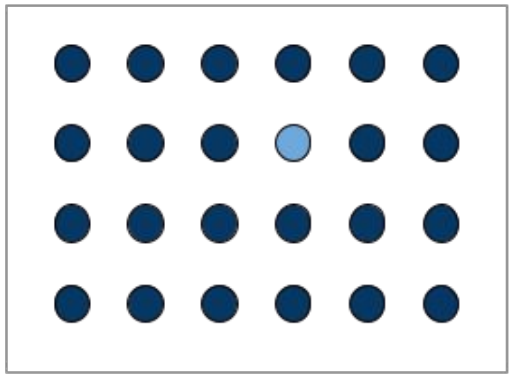
Shape

# Pre-attentive attributes: color

We can quickly distinguish one object that differs from the rest

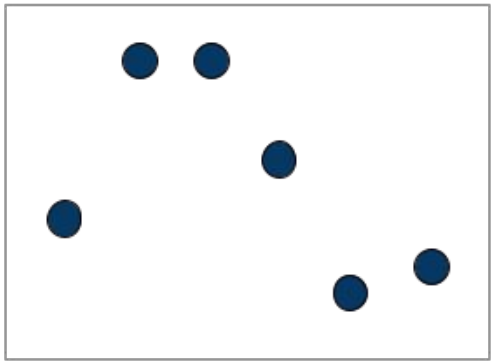


Hue

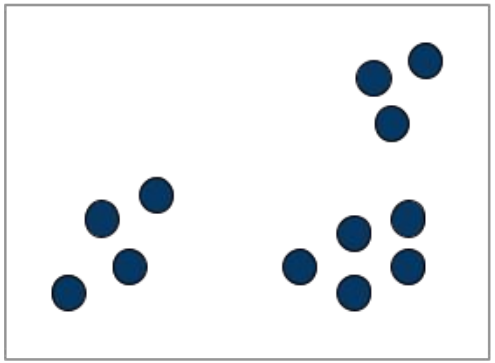


Lightness

# Pre-attentive attributes: position



2D



Groups

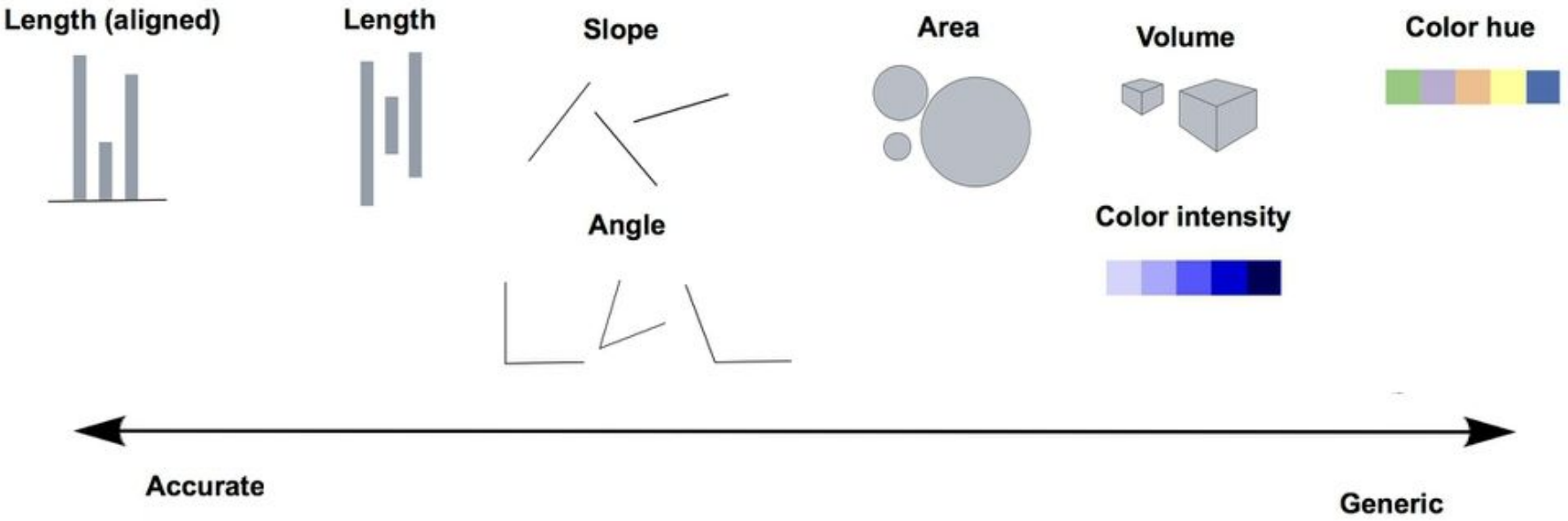
# Hierarchy of graph elements

Cleveland, McGill 1985

Experiments with volunteers of how well they **judge ratios** between values graphically encoded in different ways.

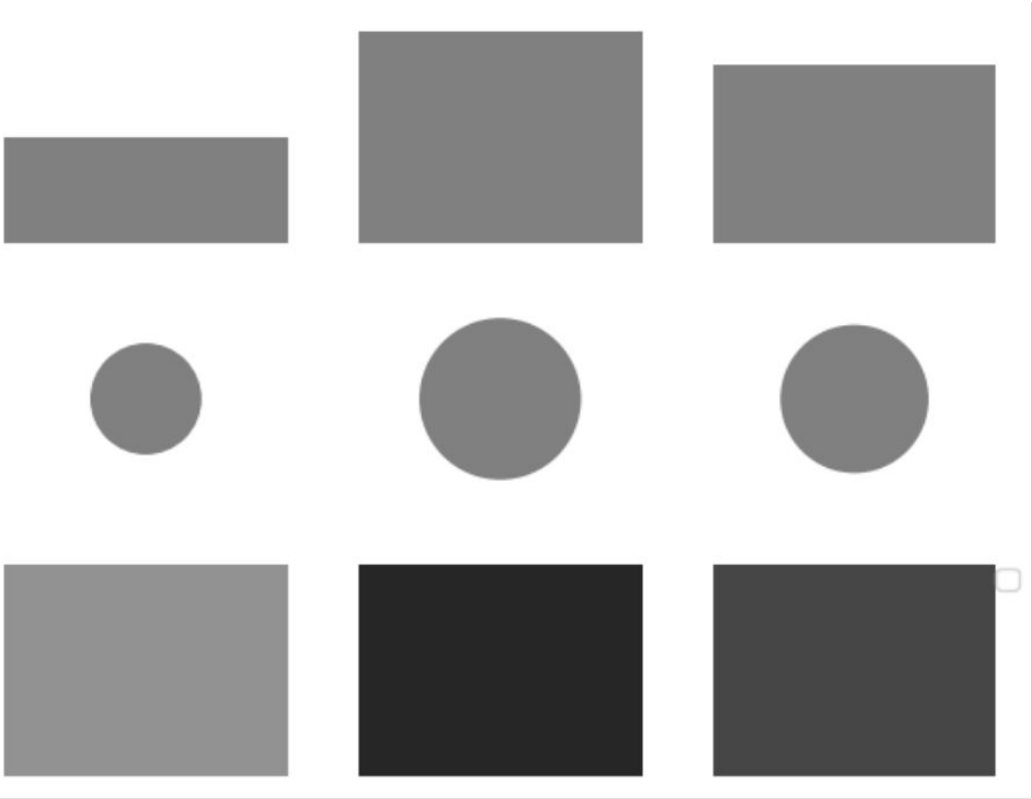
Not all pre-attentive attributes are equally good for **quantitative reasoning**.

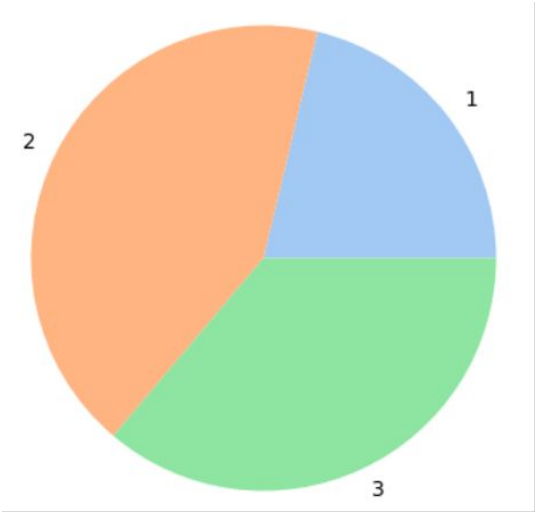
# Prefer elements on the left side for accuracy




Based on https://paldhous.github.io/ucb/2015/dataviz/week2.htm

# The same data with length / area / color / angle






10,20,17

# Chart selection tools

In lecture 3 and later, we have seen many types of graphs

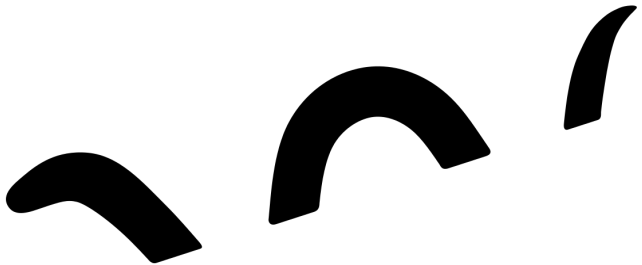
Some websites list them based on variable type and purpose for easier selection:

- https://www.data-to-viz.com/
- https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Let us look at some the suggestions from the first website in terms of the hierarchy of graph elements

# From parts to the whole: gestalt

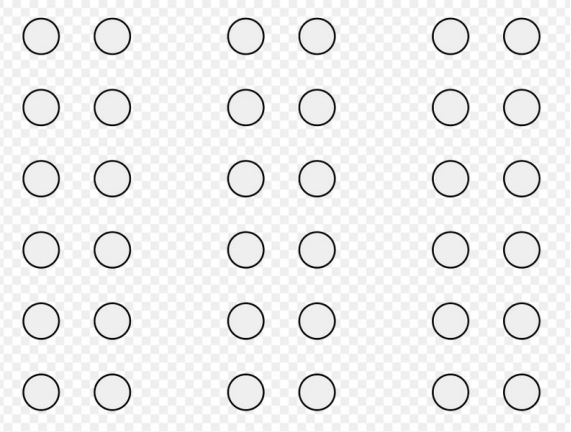
- Gestalt psychology (early 20th century, Austria and Germany)
- **Gestalt** means **pattern**
- Our brains group individual shapes into larger patterns
- The brain favors speed to precision (illusions, errors)
- It also favors symmetry and simplicity
- Several gestalt principles are relevant in data visualization

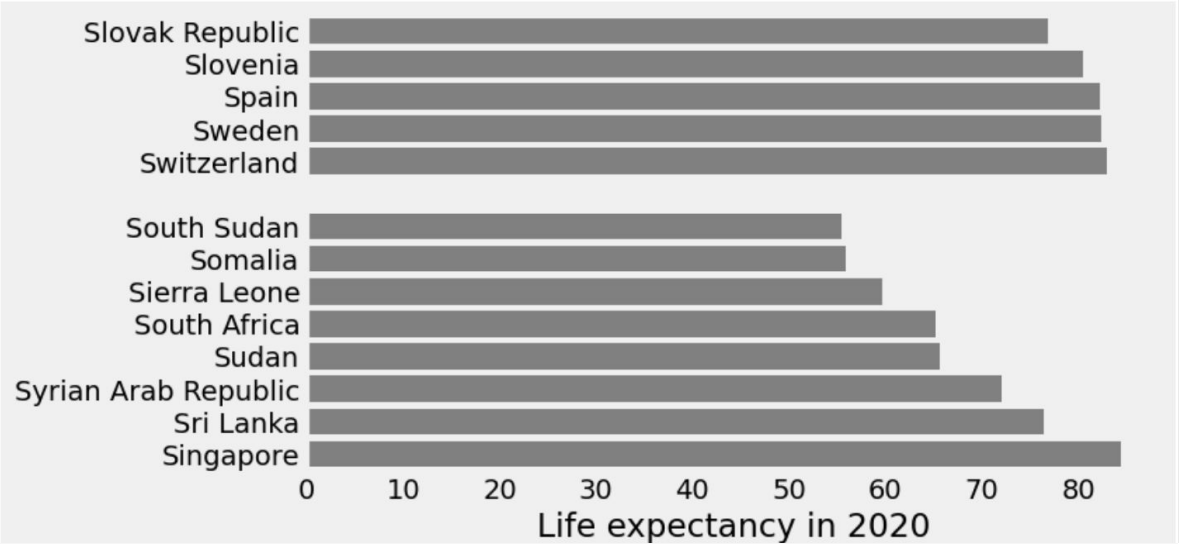


https://commons.wikimedia.org/wiki/File:Reification.svg

# Principle of proximity

- Objects located close to each other are perceived as a group
- Good use of space in plots / tables / presentations can improve readability

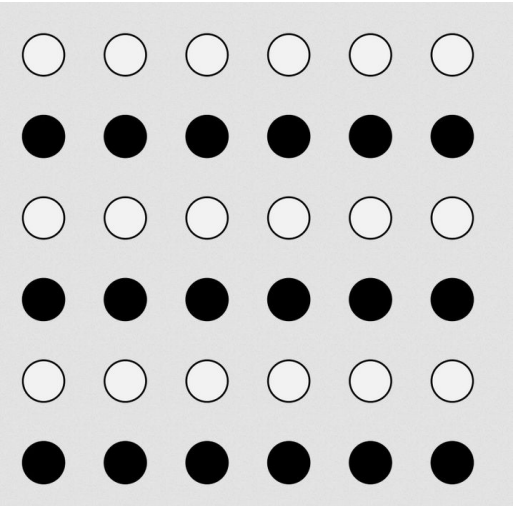


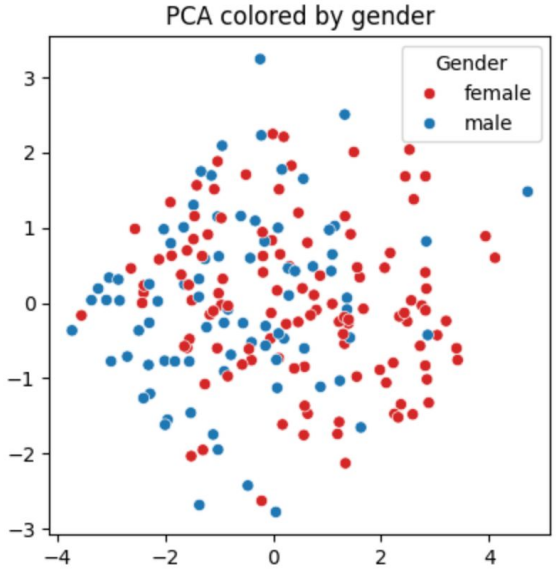



https://commons.wikimedia.org/wiki/File:Gestalt_proximity.svg

# Principle of similarity

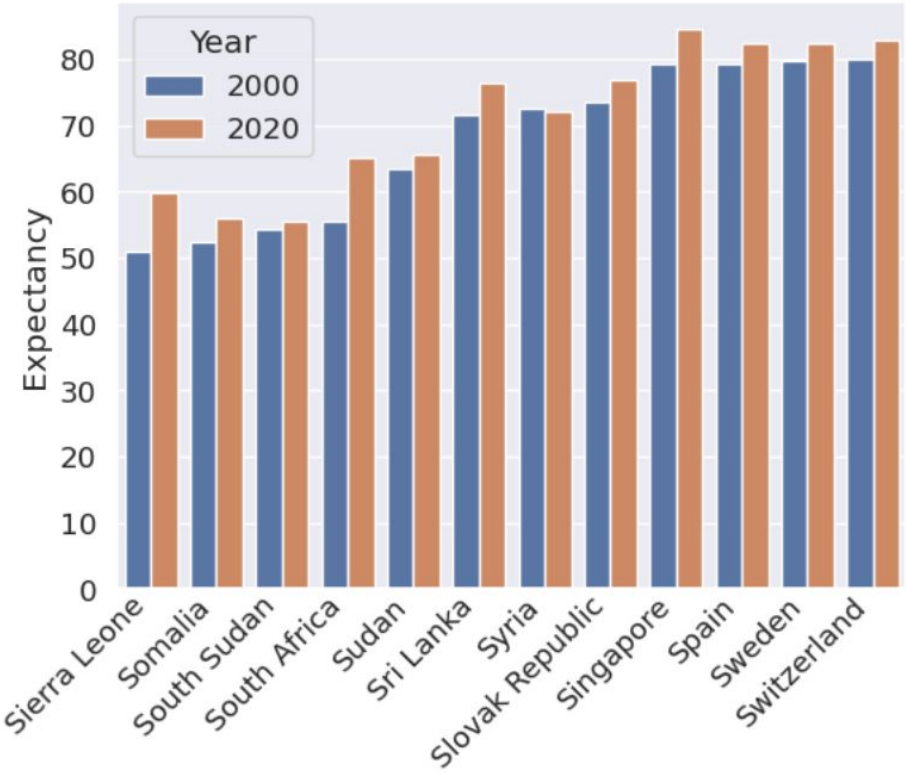
- Similar objects are perceived as a group even of not close by
- Various plots use color / shape to distinguish groups






https://commons.wikimedia.org/wiki/File:Gestalt_similarity.svg

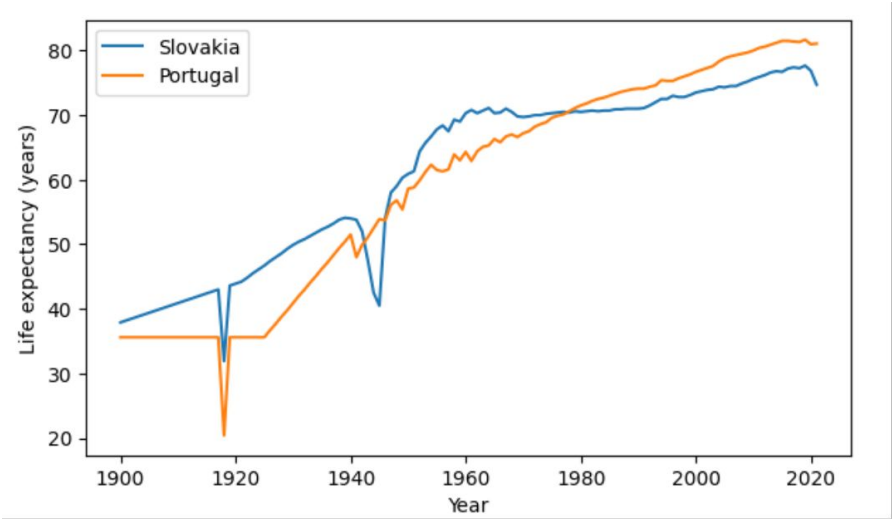
# How are both principles used here?

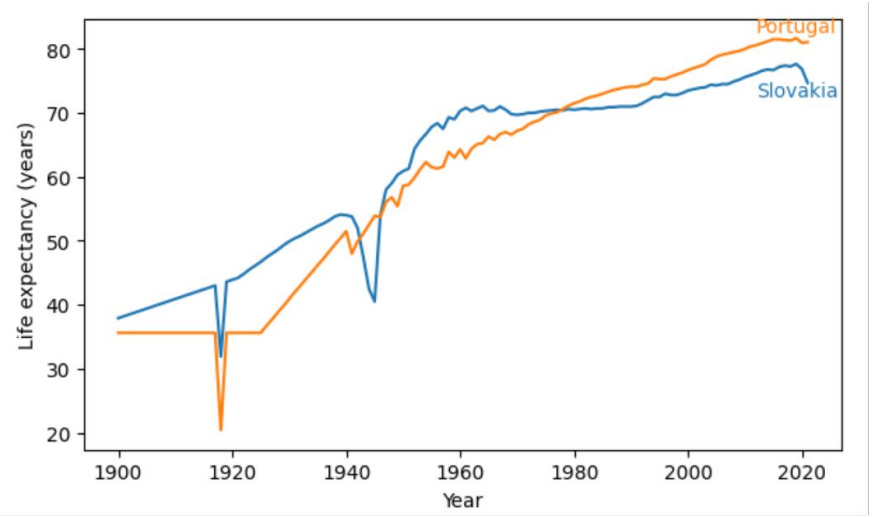


# Example

separate legend vs marking lines with text in the same color

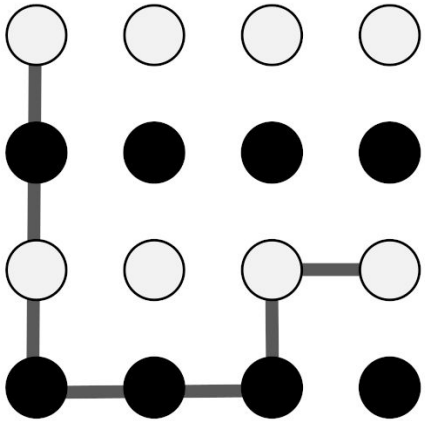
- using principles of proximity and similarity

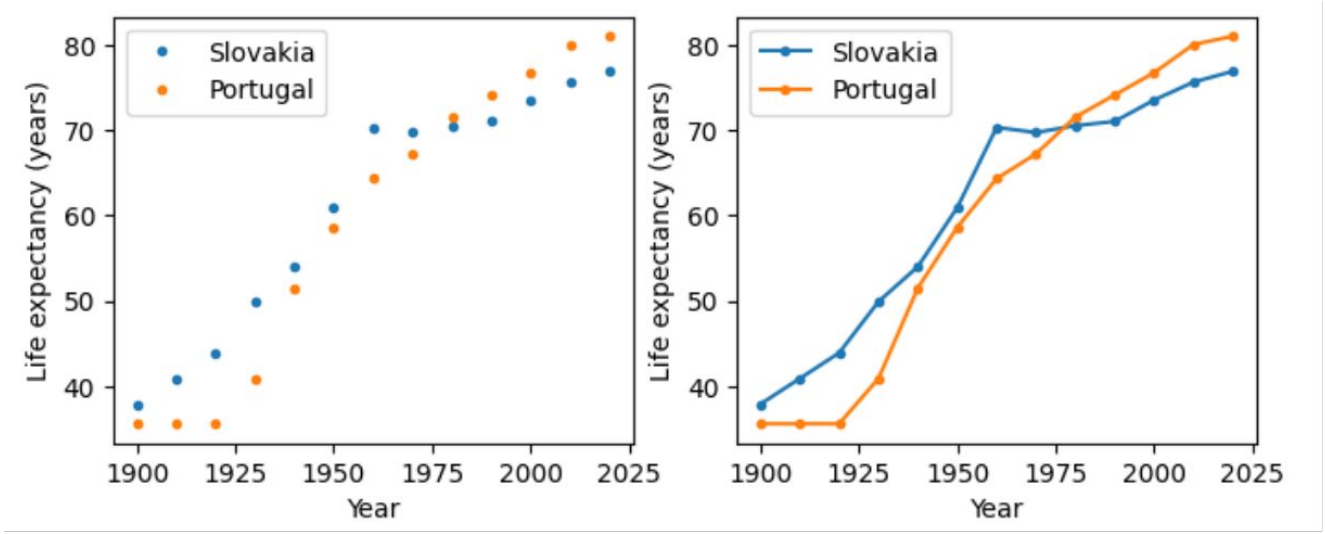


# Principle of connection

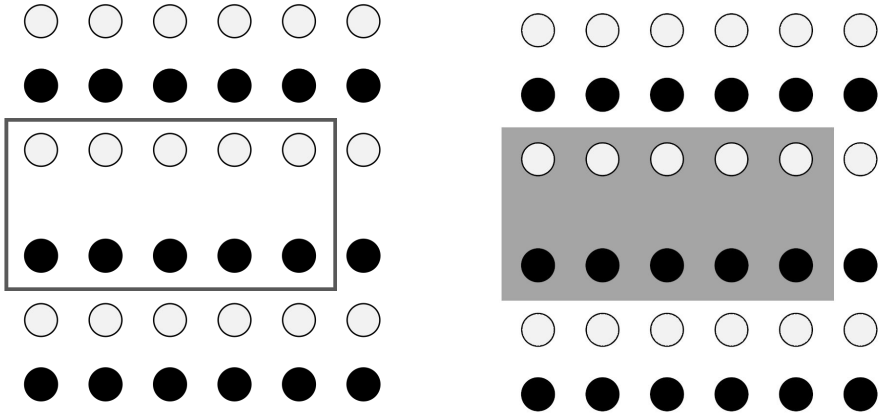
- Connected objects are perceived to form a group
- Stronger than proximity and similarity
- Consider carefully when to use line graph vs. scatter plot

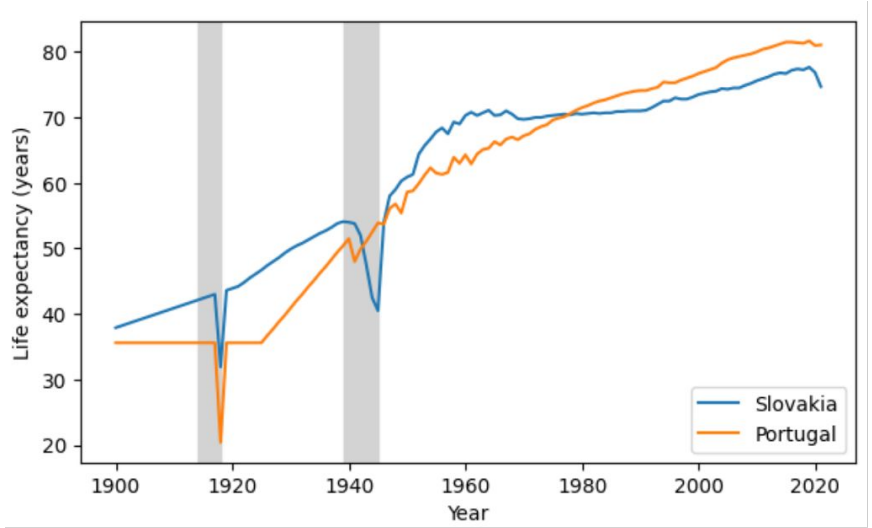




# Principle of enclosure

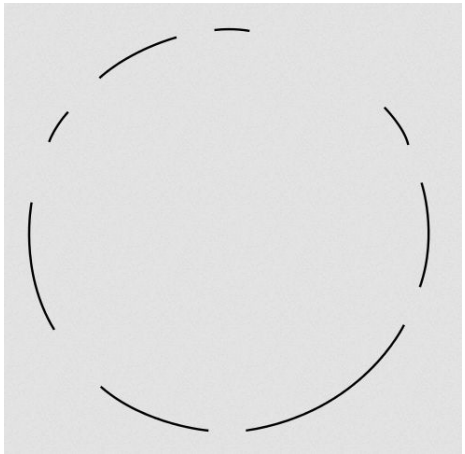
- Enclosed objects are perceived as a member of the group
- Stronger than proximity and similarity
- Useful for highlighting in plots; little is enough (e.g. light background)

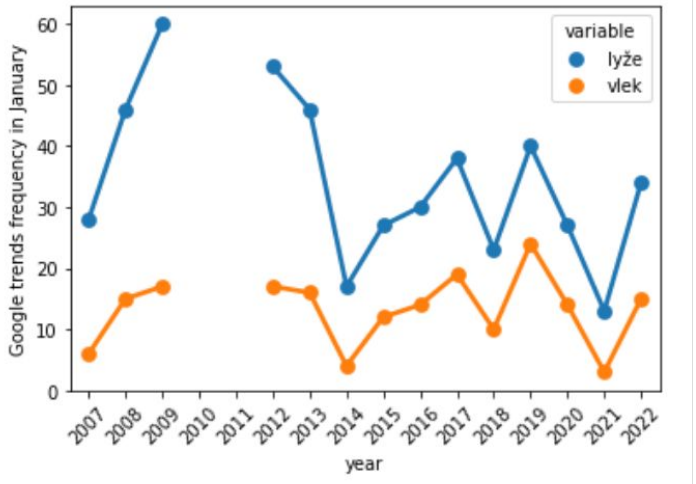




# Principle of closure

- Our brain fills gaps in figures, connects interrupted lines
- Useful / dangerous when interruptions by design

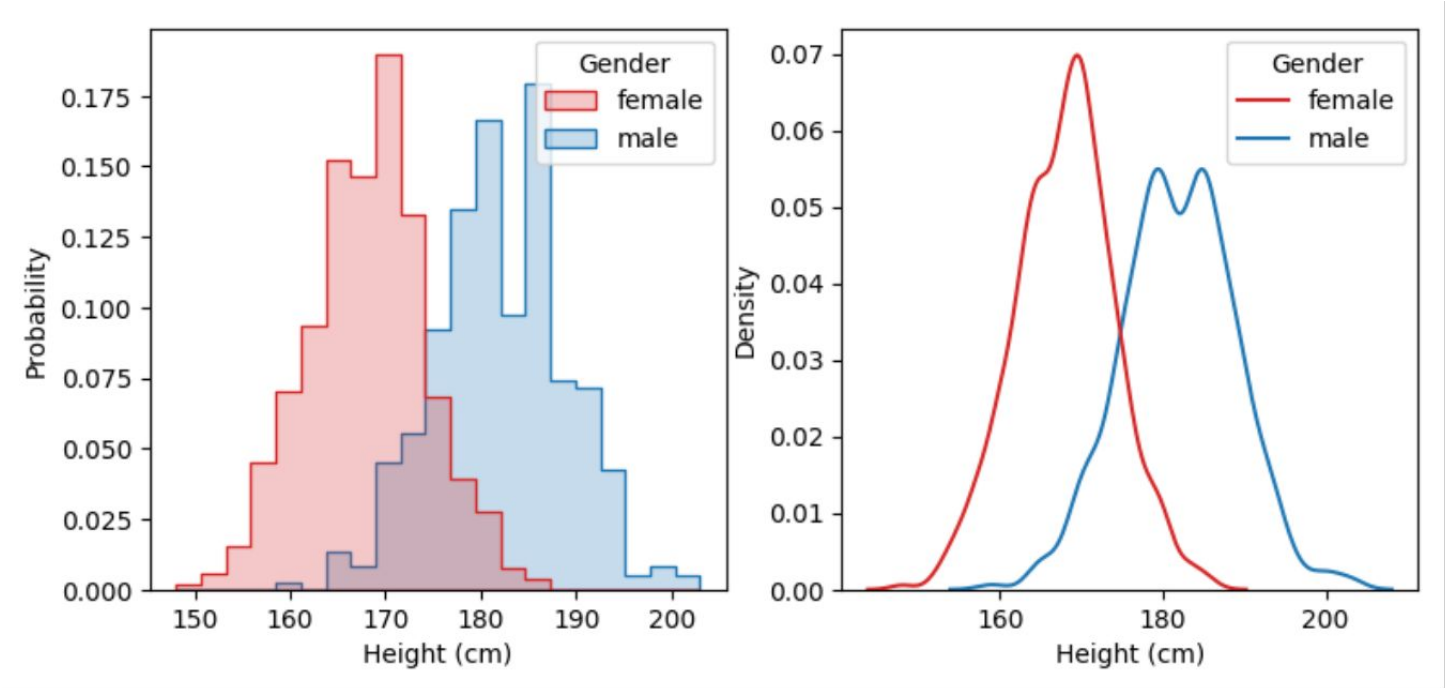





https://commons.wikimedia.org/wiki/File:Gestalt_closure.svg

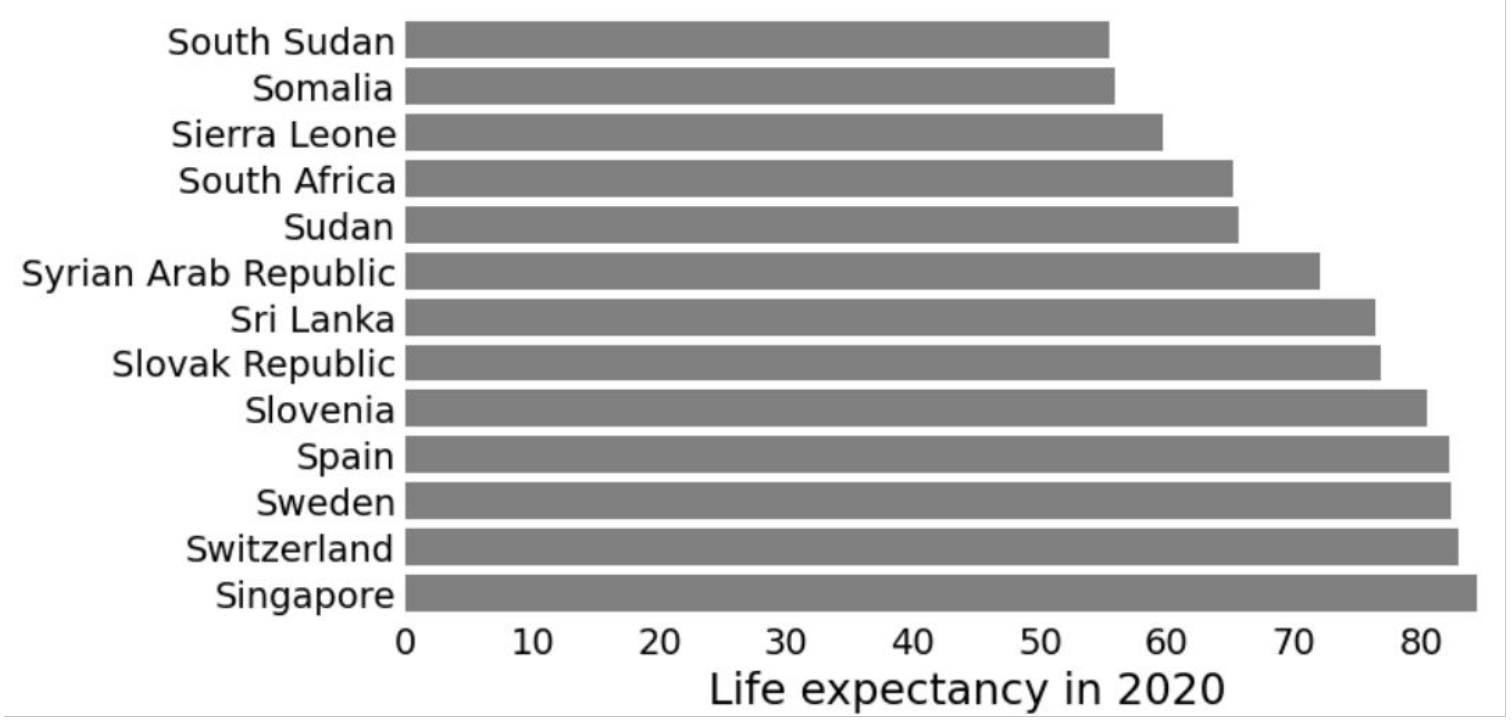
# Principle of continuity

Smooth lines are easier to follow than angular and sharp



# Frames not necessary, gestalt principles fills them in

(principle of closure and proximity)

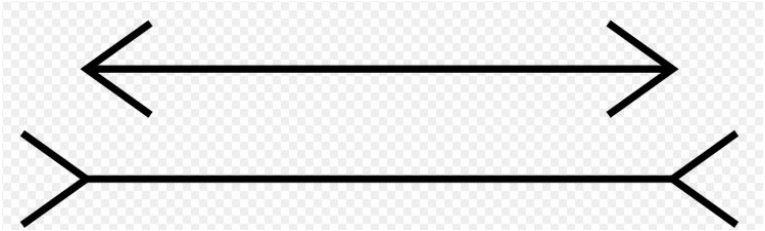


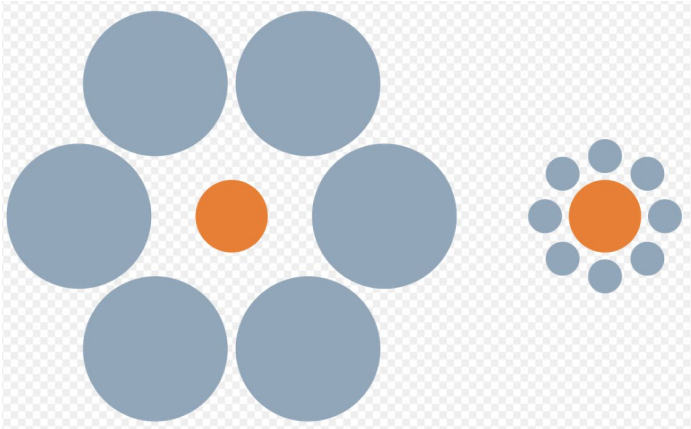
# Illusions

- Fast visual processing leads to errors
- These are demonstrated by many optical illusions
- Avoid creating illusions in your plots

# Illusions: length and size

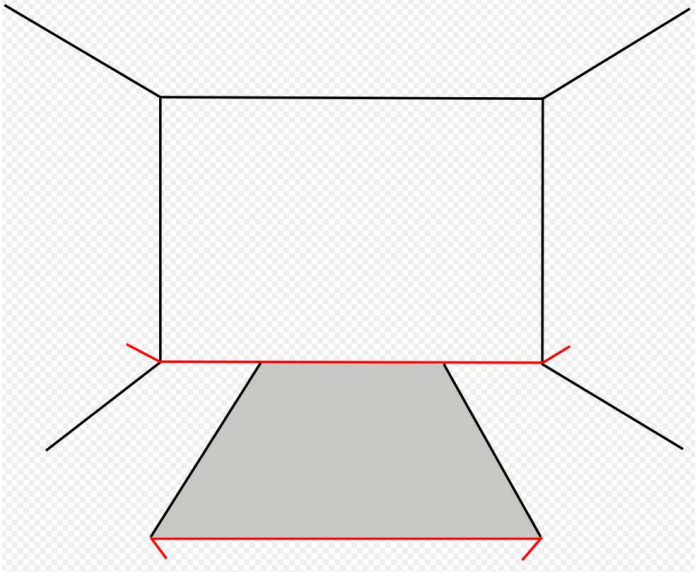
Müller-Lyer and Ebbinghaus illusions

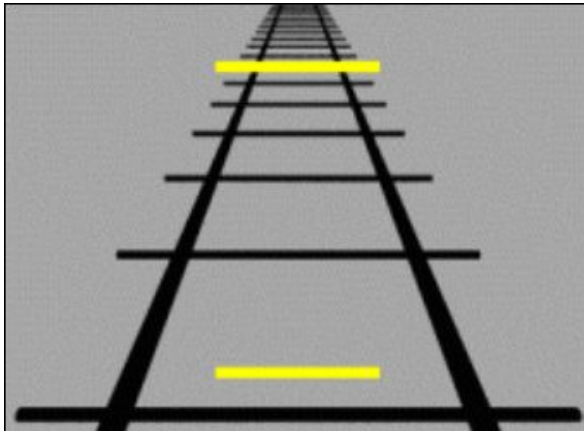




https://commons.wikimedia.org/wiki/File:M%C3%BCller-Lyer_illusion.svg
https://commons.wikimedia.org/wiki/File:Mond-vergleich.svg

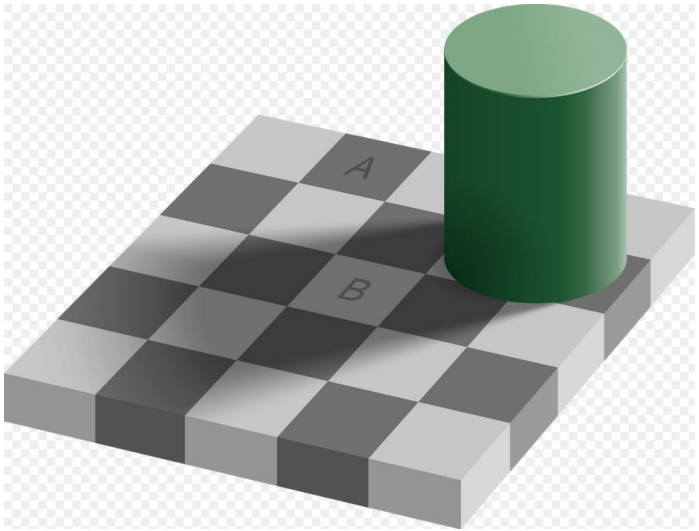
# Illusions: length, perspective, spatial compensation

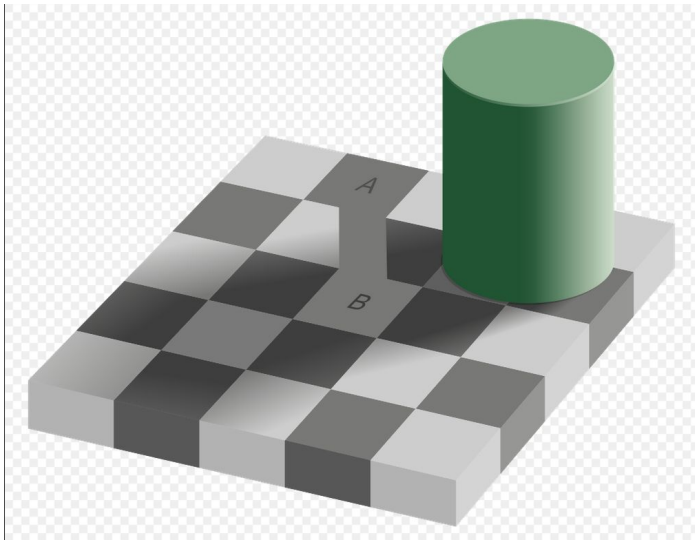


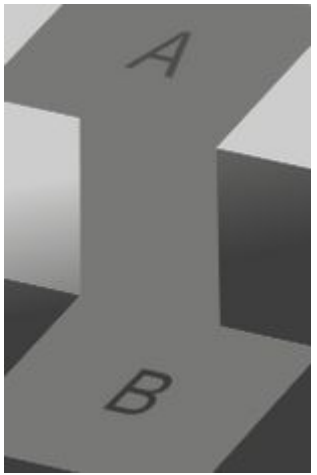


https://commons.wikimedia.org/wiki/File:Mueller_lyer.svg
https://commons.wikimedia.org/wiki/File:Ponzo_illusion.gif

# Illusions: color



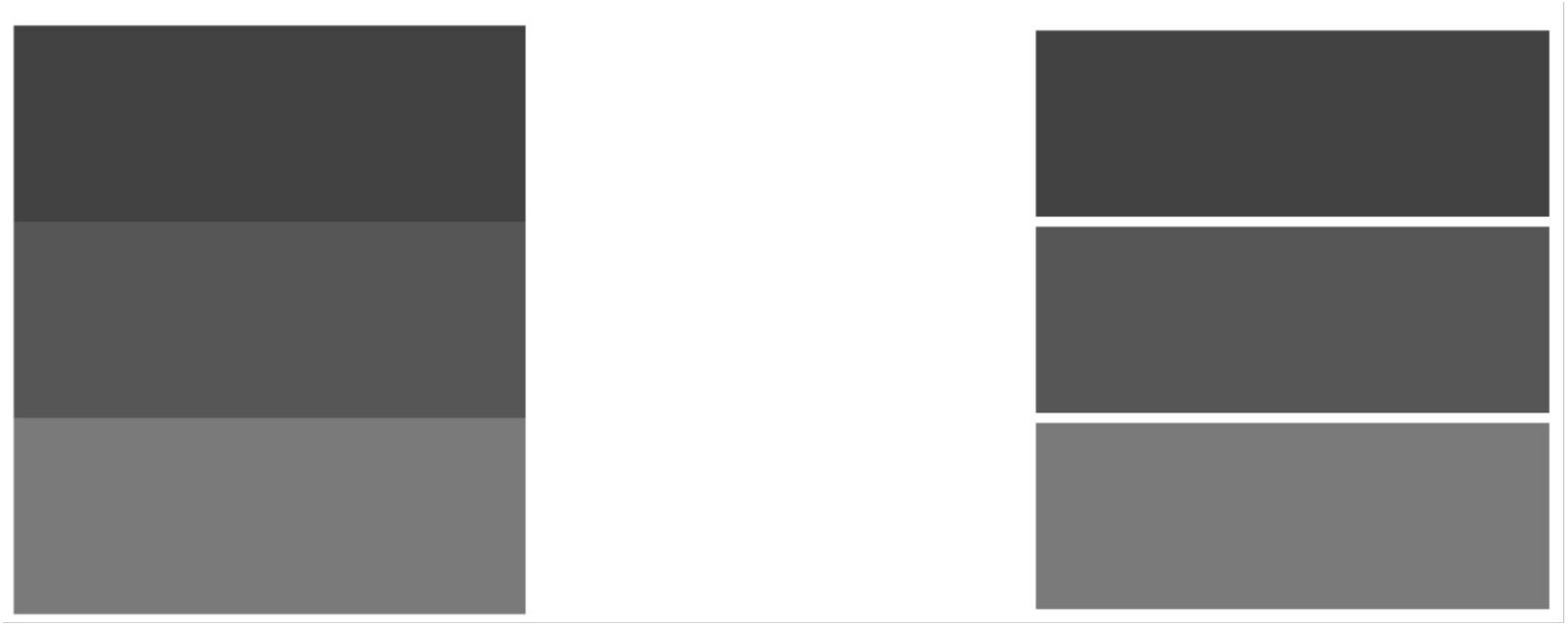




https://en.wikipedia.org/wiki/File:Checker_shadow_illusion.svg
https://commons.wikimedia.org/wiki/File:Grey_square_optical_illusion_proof2.svg

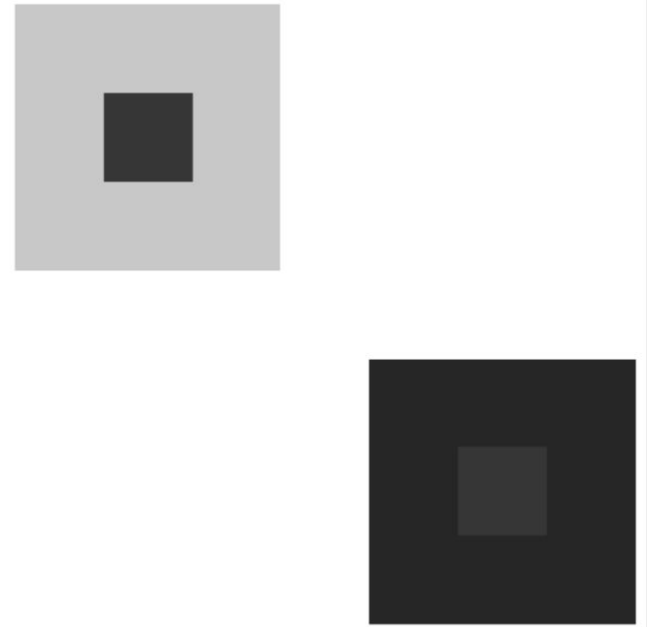
# Illusions: color

Mach bands: when bands touch, the edge effect exaggerates their difference



# Illusions: color

In heatmap the perception of colors is influenced by their surroundings

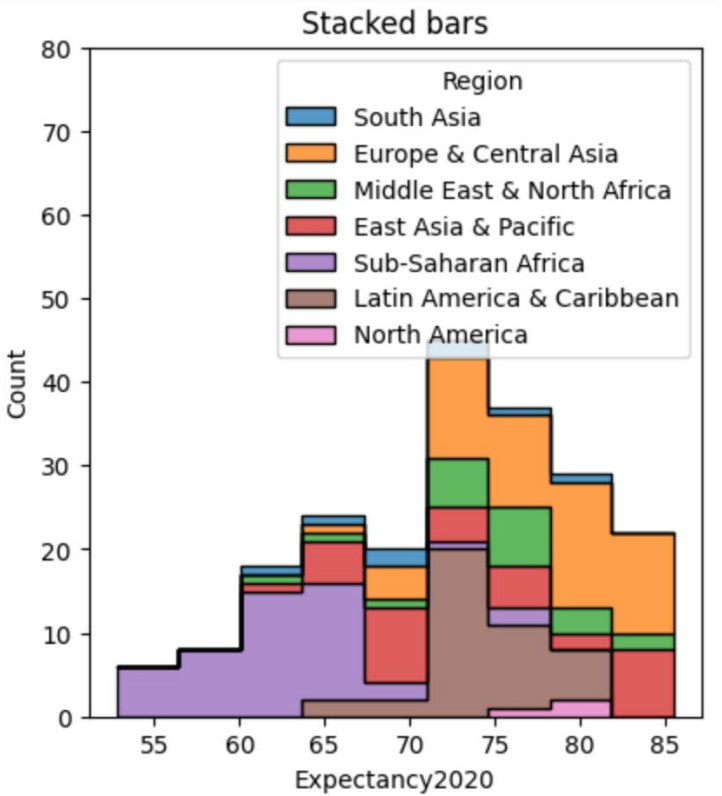


```
a = np.array([
    [10, 10, 10,  0,  0,  0,  0],
    [10, 37, 10,  0,  0,  0,  0],
    [10, 10, 10,  0,  0,  0,  0],
    [ 0,  0,  0,  0,  0,  0,  0],
    [ 0,  0,  0,  0, 40, 40, 40],
    [ 0,  0,  0,  0, 40, 37, 40],
    [ 0,  0,  0,  0, 40, 40, 40],
    ])
pal = sns.color_palette("dark:white_r", as_cmap=True)
axes = sns.heatmap(data=a, square=True,
                   cmap=pal, cbar=False)
axes.axis('off')
```

# Working memory

- **Iconic memory:** extremely short-term (<1s), simple pre-attentive processing
- **Visual short-term memory:** many seconds, but very small capacity (only 3-5 items)
- **Long-term memory:** store and recall selected information

Since working memory is small, looking at a plot with many colors requires back-and-forth between legend and plot



# Chart and table junk

- **Chart junk**: elements of plots not necessary to convey information
- They unhelpfully catch our attention through pre-attentive attributes
- Most visualization can be improved by simplification
- Some redundancy can be useful

Nice visualizations of the simplification process:

- https://www.darkhorseanalytics.com/blog/data-looks-better-naked
- Also tables, maps and the unpopular pie charts

# Summary

- **Pre-attentive attributes** are processed by our brains very fast
- Choosing the right attributes from the **hierarchy** allows accurate quantification
- Principles of **gestalt** describe how the brain connects part to the whole
- The brain can also make errors in visual processing as seen in **illusions**
- Removing **chart junk** concentrates our attention to the important elements

# Additional sources

- Utilizing Gestalt Principles to Improve Your Data Visualization Design
- http://daydreamingnumbers.com/blog/gestalt-laws-data-visualization/
- Albert Cairo: The Functional Art
- C.N. Knaflic: Storytelling with Data
- Stephen Few: Now You See it