Lecture 10
Presentation of results, time series, texts

Data visualization · 1-DAV-105

Lecture by Broňa Brejová

Acknowledgement: materials inspired by lectures from Martina Bátorová in 2021

Data analysis project phases

Recall from L08: Data analysis project phases

- Obtaining data
- Data preprocessing, **checking**, cleaning
- **Exploratory** analysis
- Formation of **hypotheses**
- Testing hypotheses
- **Explanatory** visualizations for the final report / presentation

Details: obtaining data

- Obtaining data
    - This course: we download whole datasets in a tabular form
    - But often also web scraping, manual collection of data, measurements, surveys,...
    - Requires careful planning

- Data preprocessing, **checking**, cleaning
- **Exploratory** analysis
- Formation of **hypotheses**
- Testing hypotheses
- **Explanatory** visualizations for the final report / presentation

Details: preprocessing data

- Obtaining data
- Data preprocessing, **checking**, cleaning
  - Try to understand how the data was obtained and processed
  - Convert them to a convenient format
  - Check for missing values and suspicious outliers
  - Very important phase: "Garbage in, garbage out"

- **Exploratory** analysis
- Formation of **hypotheses**
- Testing hypotheses
- **Explanatory** visualizations for the final report / presentation

Details: exploratory analysis

- Obtaining data
- Data preprocessing, **checking**, cleaning
- **Exploratory** analysis
  - Try many analyses
  - This course: visualizations and simple statistics
  - Later you learn advanced statistical and machine learning models
  - Even less successful attempts may suggest new directions

- Formation of **hypotheses**
- Testing hypotheses
- **Explanatory** visualizations for the final report / presentation

Details: Formation of hypotheses

- Obtaining data
- Data preprocessing, **checking**, cleaning
- **Exploratory** analysis
- Formation of **hypotheses**
  - Select visualizations showing interesting trends / exceptions in the data
  - Formulate possible relationships
    (but remember, correlation does not imply causation)

- Testing hypotheses
- **Explanatory** visualizations for the final report / presentation

Details: Testing hypotheses

- Obtaining data
- Data preprocessing, **checking**, cleaning
- **Exploratory** analysis
- Formation of **hypotheses**
- Testing hypotheses
    - Recheck your code and data, try other related analyses
    - Try to find other relevant data or existing analyses by other people
    - If important decisions will be based on your result,
      test it particularly thoroughly
      (what would happen if our plot was all wrong?)
- **Explanatory** visualizations for the final report / presentation

Details: Explanatory visualizations

- Obtaining data
- Data preprocessing, **checking**, cleaning
- **Exploratory** analysis
- Formation of **hypotheses**
- Testing hypotheses
- **Explanatory** visualizations for the final report / presentation
  - Formulate your conclusions
  - Support them with your analysis and visualizations
  - Do not include all exploratory analyses
    (but do not hide data contradicting your conclusion)
  - Polish visualizations that you selected

Presentation of results

Presentation of results

- Understand **context**, audience, goals (more later)
- Tell a **story** (more later)
- Choose appropriate **visuals** (text / table / chart, appropriate type of graph, pre-attentive attributes, hierarchy of graph elements)
- Eliminate clutter, **focus attention** on the main points (pre-attentive attributes, such as color, size, spacing)
- Pay attention to **design** (accessibility due to font size and colors, aesthetics...)
- Get **feedback** and a lot of **practice**

(see Cole Nussbaumer Knaflic: Storytelling with data)

Understand the context of your presentation

Before writing any text or preparing any presentation try to find out:

- Who is your expected audience?
- What do they know and what do you need to explain?
- What might be interesting / new for them?
- What is the medium (live presentation, video, printed text, website)?
- What is the appropriate length (time, number of pages)?
- What do you want to achieve by the presentation?
  (inform / entertain / inspire to take a specific action)

Examples

Try to list some examples of situations where data visualization might be presented: who are speakers and audiences, what are goals

Situations where data visualizations are presented

- A **company** presents to potential **consumers**, persuades them to **buy** their products
- A **charity** presents to general **public**, persuades them to **donate**
- A **nonprofit / government** present to general **public**, persuades them to take **action** (live healthily, protect environment, ...)
- An **employee** presents to **colleagues**, persuades them to **change** processes
- A **politician** presents to general **public**, persuades them to **vote** for something
- A **journalist** writes for general **public**, informs them about important **issues**
- A speaker talks to general **public**, **entertains / informs** about interesting topics
- A **teacher** presents to **students**, teaches them a given **topic**
- A **student** presents to a **teacher**, **demonstrates** his / her achievements and skills
- A speaker talks to **experts**, informs about **new discoveries**, technologies etc.

Presentation of results

- Understand context
- **Tell a story**
- Choose appropriate visuals
- Eliminate clutter, focus attention on the main points
- Get feedback and a lot of practice

Storytelling

- We are easily captivated by a good story (book, movie, play)
  - We do not want to interrupt reading / watching
  - We can recall the plot afterwards
  - We want to achieve similar effects by your presentation
- Traditional stories structured as basic plot - twists - ending
- This roughly corresponds to introduction, actual content, conclusion
- Repetition useful in stories as well as in presentation

(see Cole Nussbaumer Knaflic: Storytelling with data)

Storytelling: structuring presentation

- One option is to describe your process of discovery roughly **chronologically** (omitting some dead ends): identifying question, getting data, analyzing data, coming to conclusion, recommending action
- Another option is to **lead with the ending**: starting with a call to action, backing it up with data

(see Cole Nussbaumer Knaflic: Storytelling with data)

Cognitive biases

Cognitive bias (kognitívne skreslenie)

- Cognitive bias is a systematic deviation from rational judgement
- A brain mechanism to create shortcuts, allow fast reasoning
- Term introduced by Amos Tversky and Daniel Kahneman in 1972

Very long list of biases discovered by researchers:

https://commons.wikimedia.org/wiki/File:The_Cognitive_Bias_Codex_-_180%2B_biases,_designed_by_John_Manoogian_III_(jm3).png

Three cognitive biases

- **Patternicity bias:** See non-existent patterns in data, even in random noise (related to seeing faces in the clouds)
- **Storytelling bias:** Invent "stories", explanations, cause-effect relationships for these patterns
- **Confirmation bias:** It is hard to discard our beliefs. We search for evidence that back our theories and interpret contradicting evidence the opposite way.

See Alberto Cairo: The Truthful Art

Cognitive biases in analysis and presentation

- Beware of biases in yourselves during analysis and in your audience during presentation
- *"The first principle is that you must not fool yourself---and you are the easiest person to fool"* Richard Feynman

Do not oversimplify

Story from Alberto Cairo: The Truthful Art

"*Study finds that more than a quarter journalism grads wish they'd chosen a different career*" Poynter Institute, 2013

Storytelling bias suggests:

- A change from printed to online media leads to worse job market for journalists
- Cairo as a journalism professor starts to worry about his future

Journalism grads (cont.)

"*Study finds that more than a quarter journalism grads wish they'd chosen a different career*" Poynter Institute, 2013

Actual value is 28%, as found by a survey

- This value by itself is presumably correct
- However it is not put into perspective, compared with other values

Results of Cairo's investigation

- The dissatisfaction among journalism students did not change much over the years
- Decreases in the number of news reporters and their low salaries
- Survey results imply sampling error which should be considered
- (Ideally compare to grads from other fields)

He suggests reformulating the message of the story:

"*Even if jobs prospects for journalists have worsened substantially and they may worsen even further in the future, the percentage of grads who wish they'd chosen a different career hasn't changed at all in more than a decade.*"

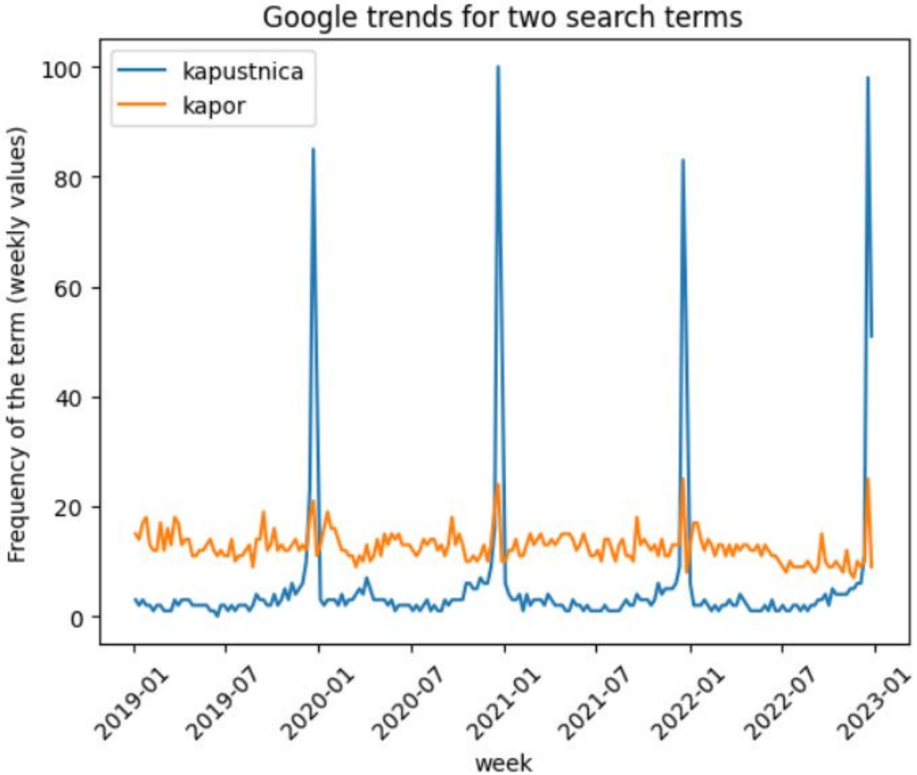
Properties of good visualization

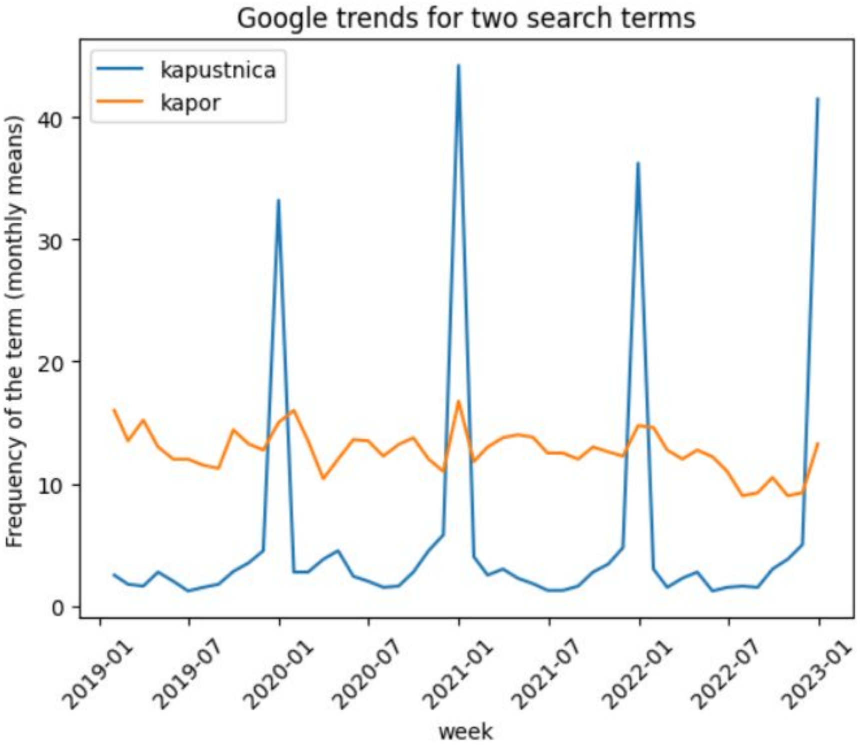
- **Truthful** (based on thorough and honest research, high quality data, appropriate analysis, correct math, no bugs in code)
- **Functional** (constitutes an accurate depiction of the data, allows meaningful comparisons)
- **Beautiful** (attractive, intriguing, aesthetically pleasing for target audience)
- **Insightful** (reveals evidence hard to see otherwise)
- **Enlightening** (changes our minds for the better)

Alberto Cairo: The Truthful Art (journalist's perspective)

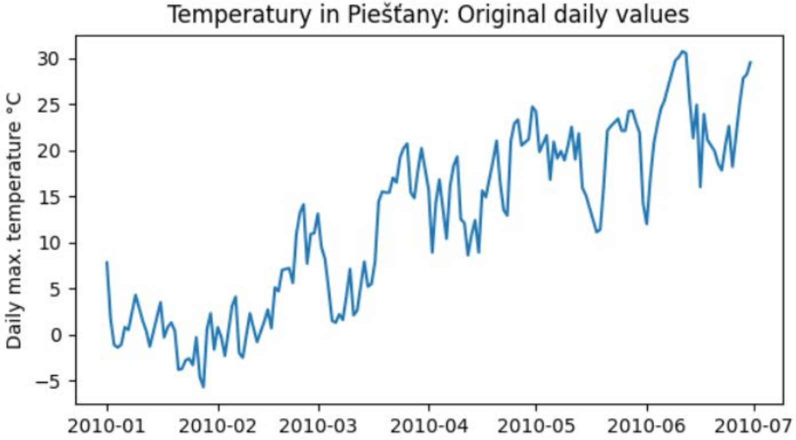
Visualizing time series (continued from L06)

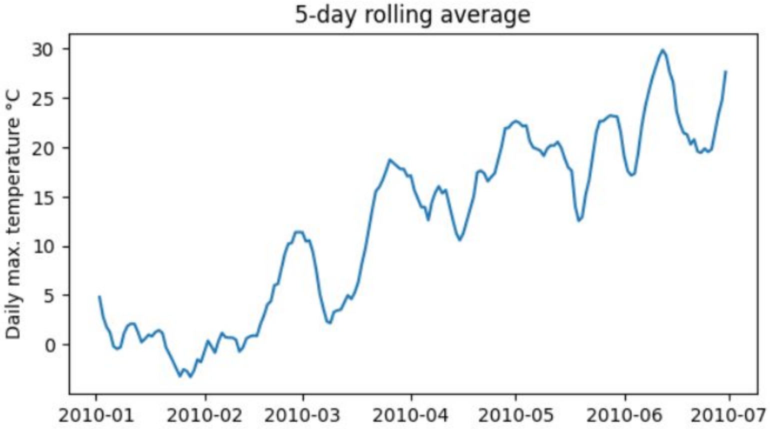
Recall: smoothing by monthly aggregation



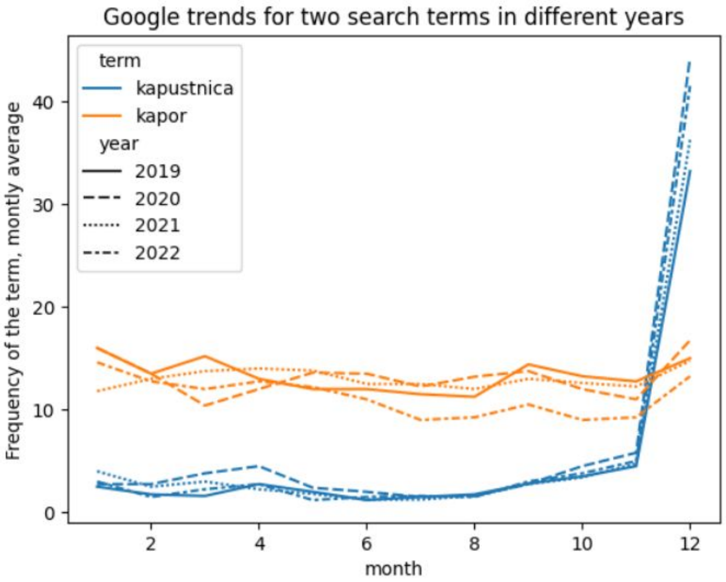


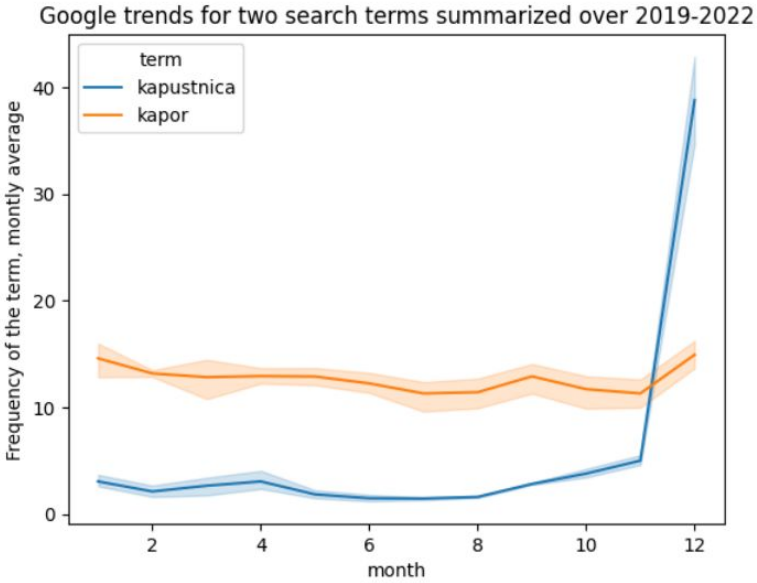
Recall: smoothing by moving window



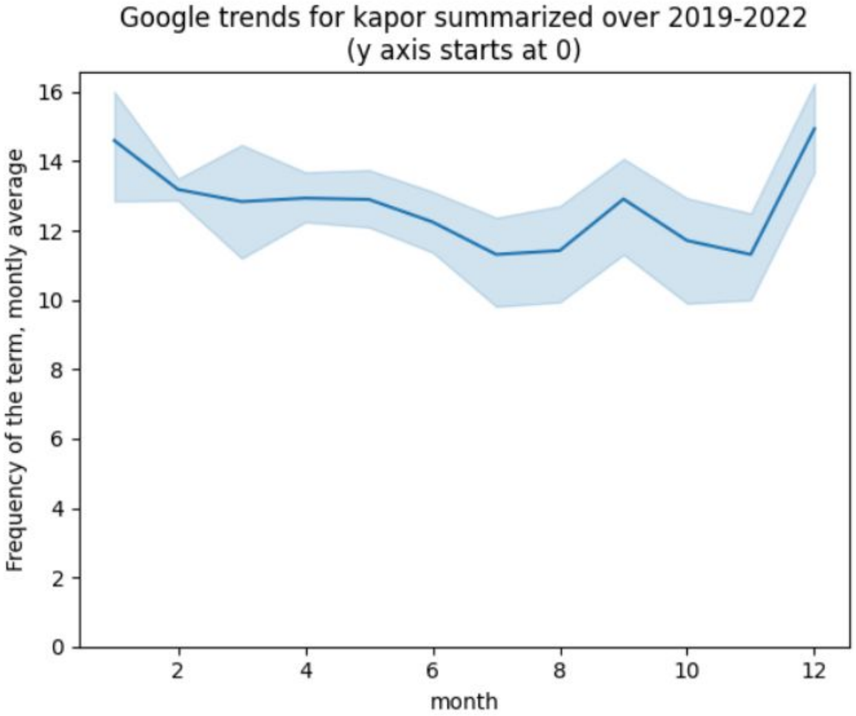


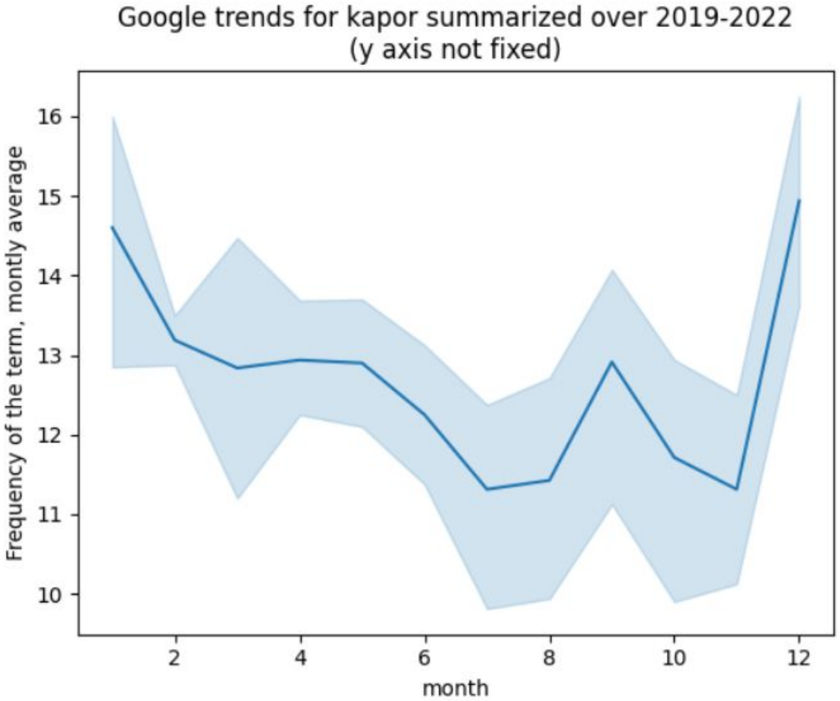
Overlapping timescales to display seasonality



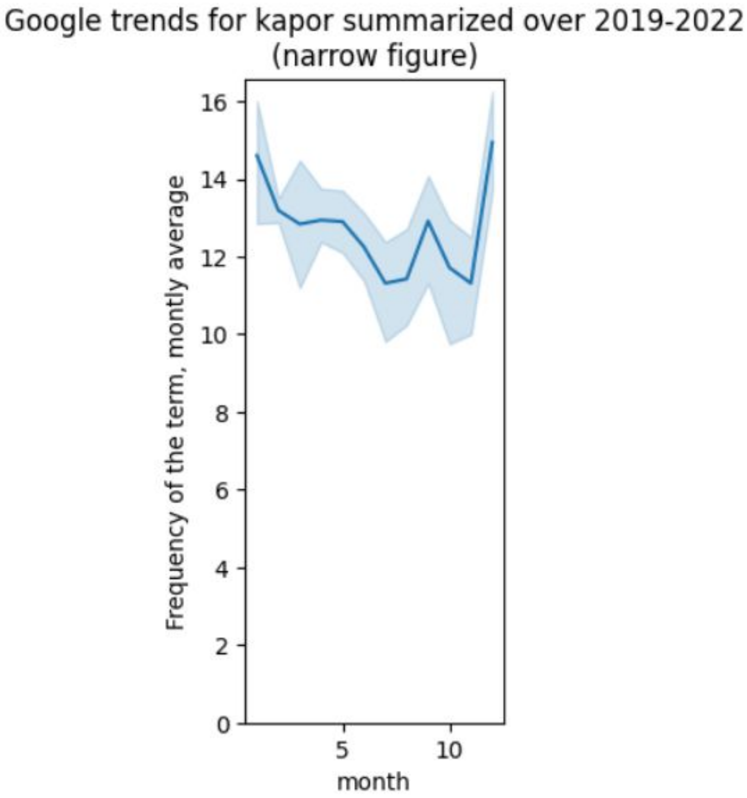


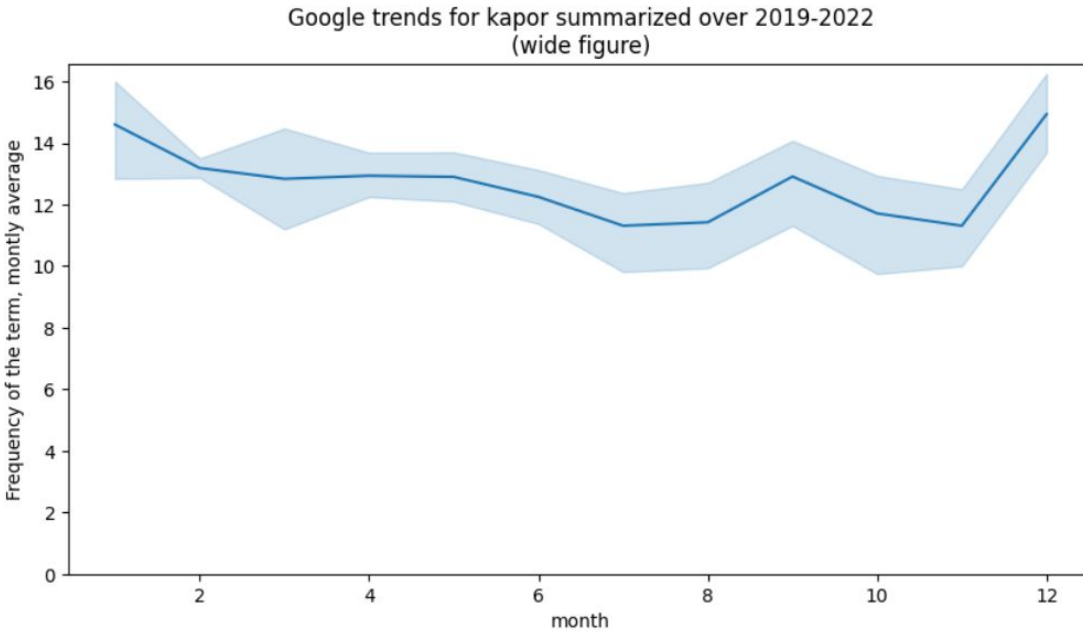
Importance of scales



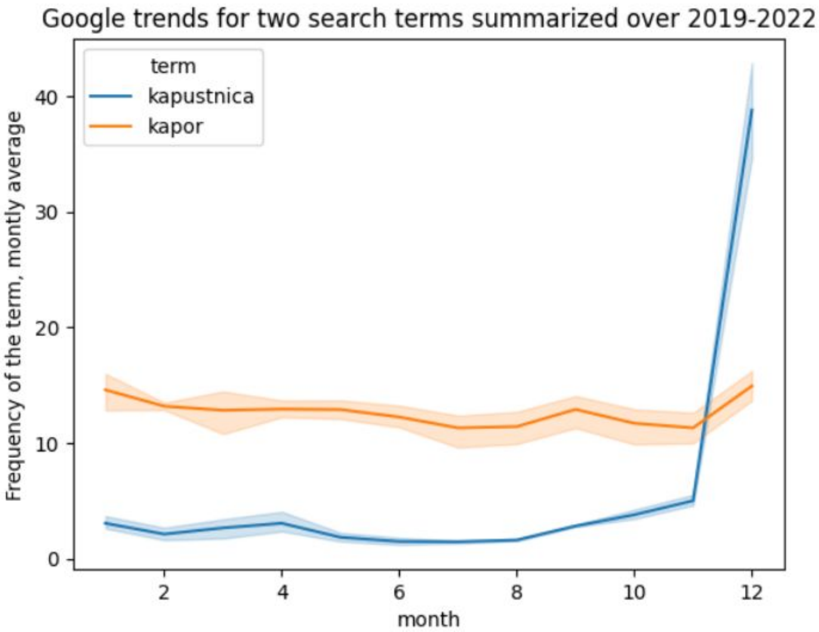


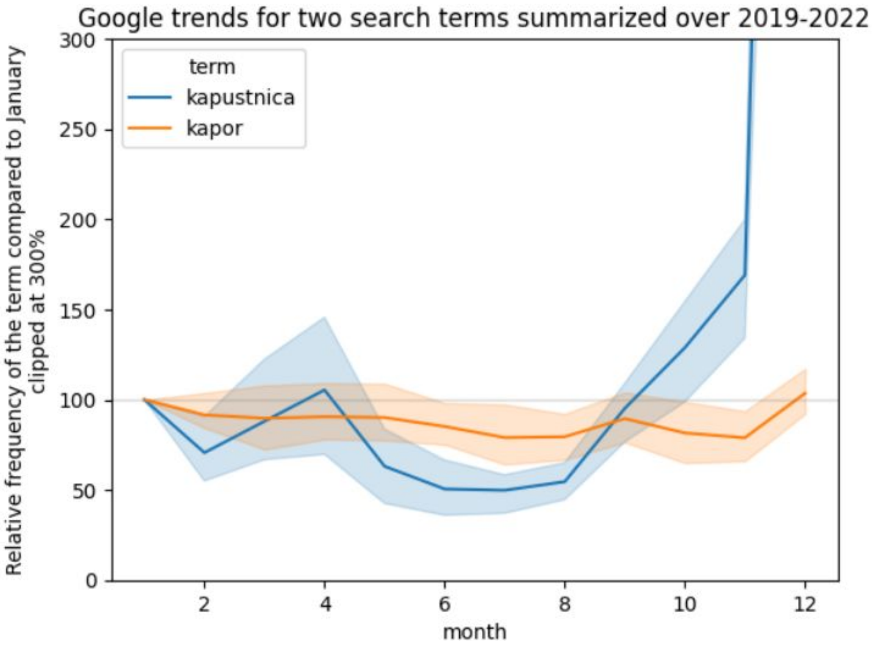
Importance of scales



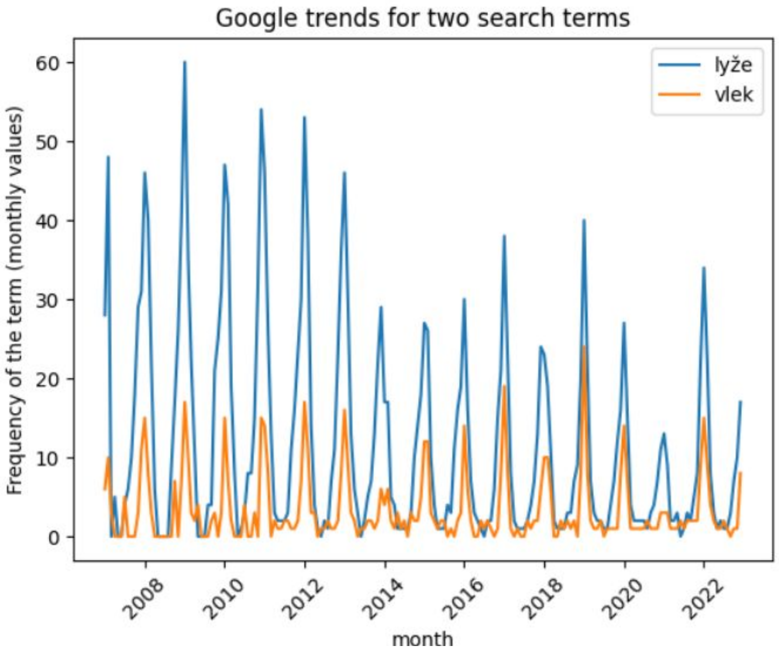


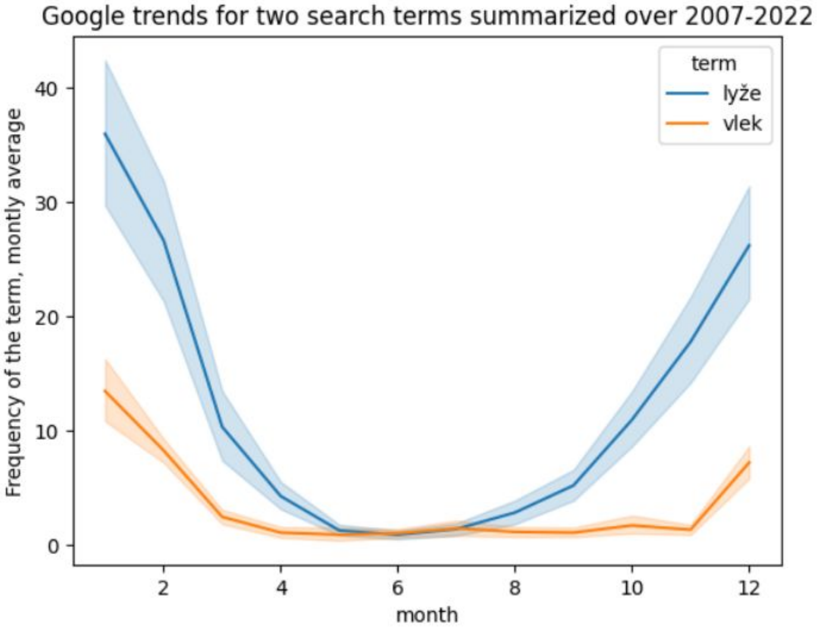
Relative scales



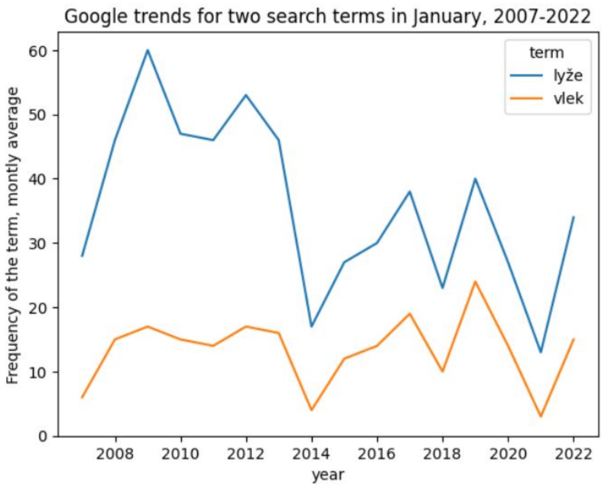


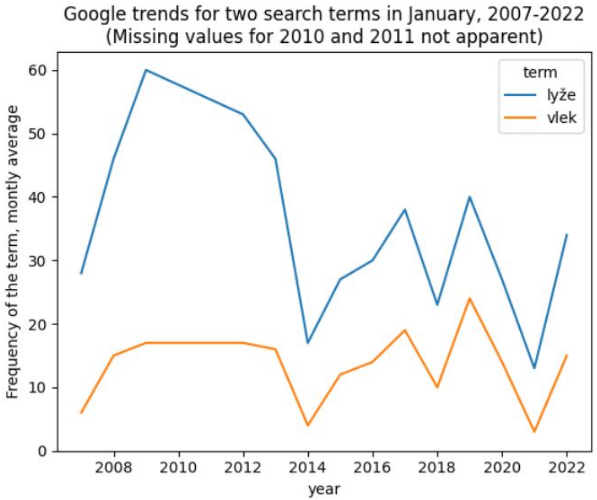
One more pair of Google trend lines

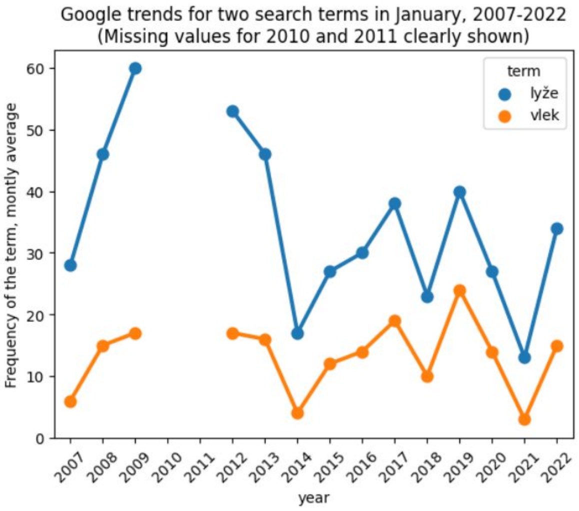




Acknowledging missing values







Recall: frequent goals of time series analysis

Observe and study"

- overall trend (increasing / decreasing / flat; rate of change),
- seasonality (daily / weekly / yearly cycles),
- noise (general variability / outliers).

Visualizing text data

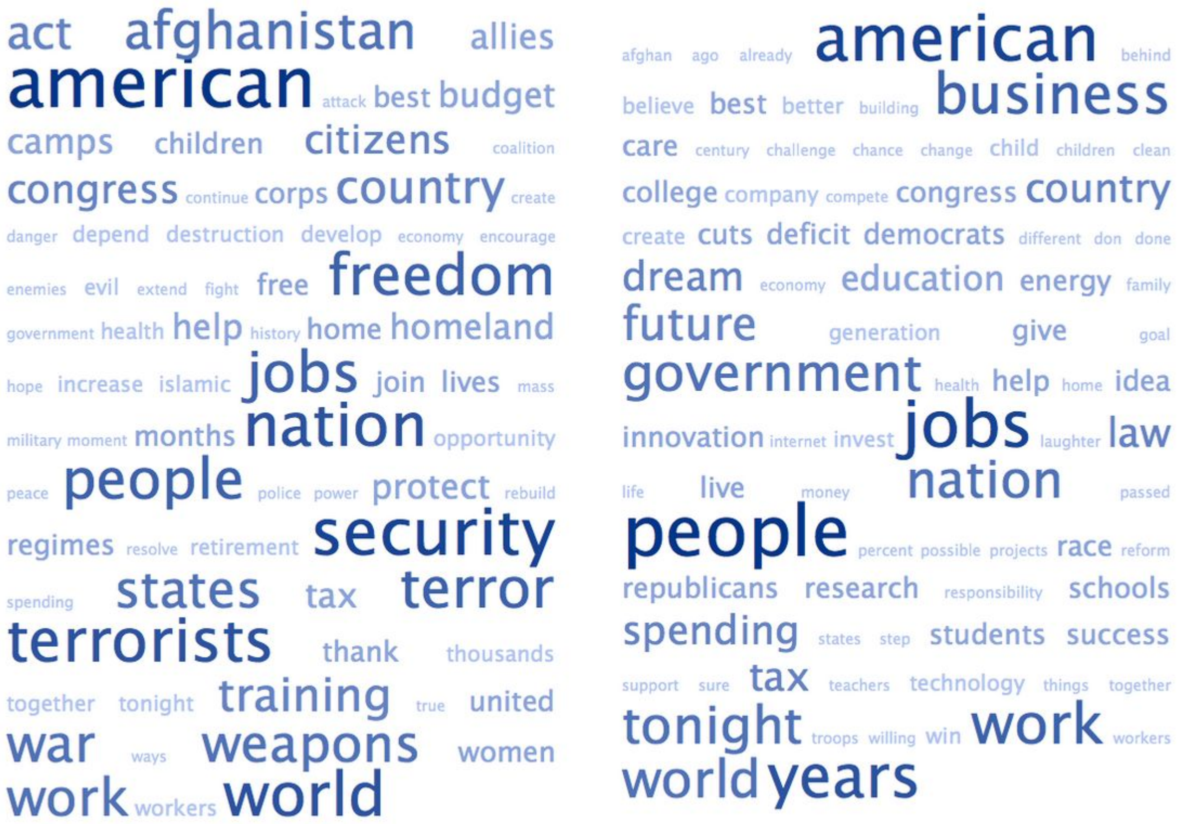
Visualizing text data

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")

Word clouds

State of the Union Address, 2002 vs. 2011



President Bush, January 29, 2002

President Obama, January 25, 2011

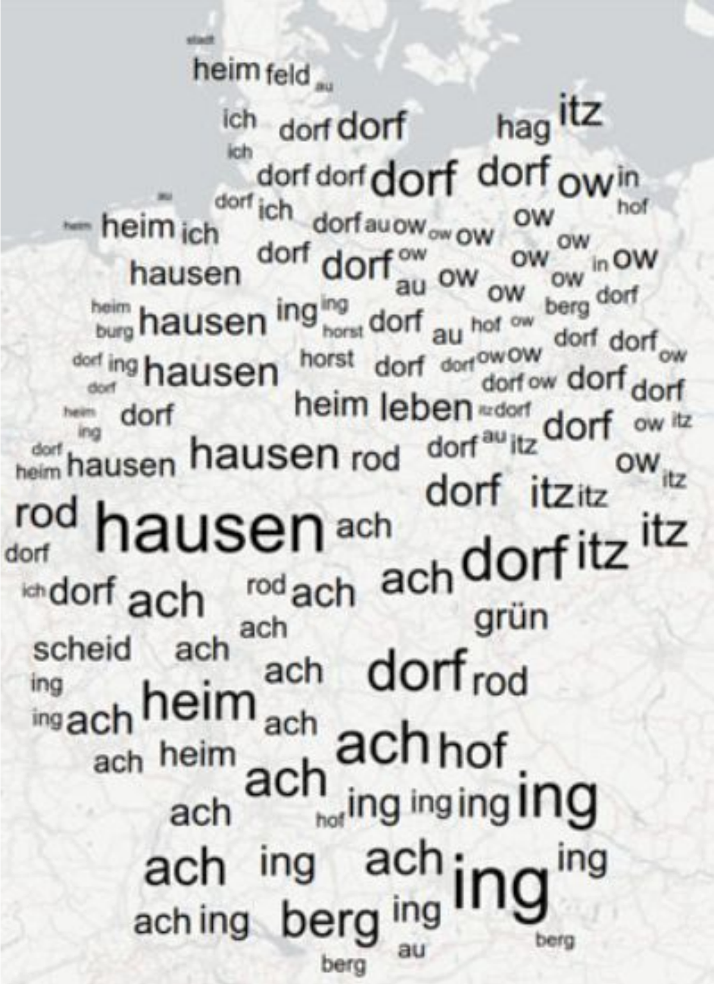
https://commons.wikimedia.org/wiki/File:State_of_the_union_word_clouds.png

Word clouds

- Display the most common words from a text
- Size of words grows with frequency
- Arranged to be visually pleasing
- Not the best option for understanding/comparing word frequencies
- You can also display word frequencies using **bar graphs** and other plot types

Tag cloud

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

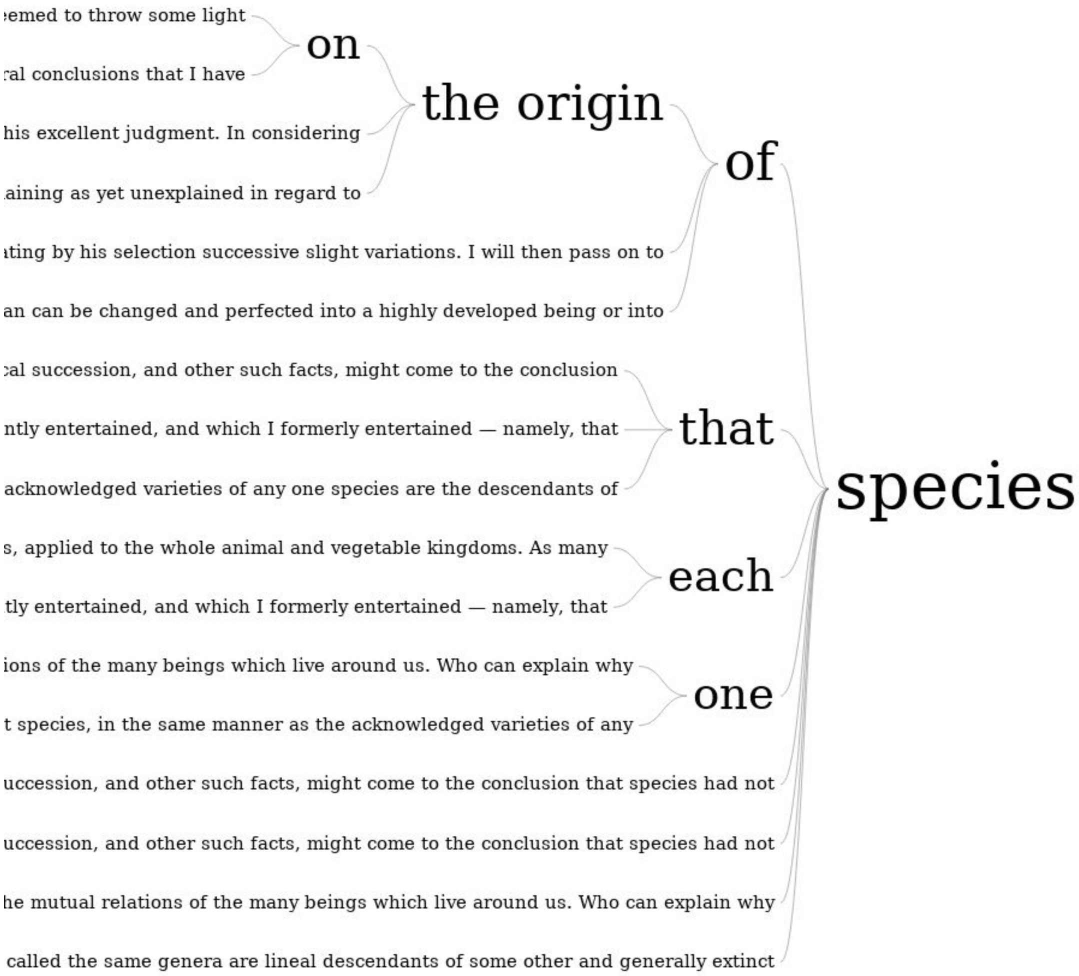

stadt
heim feld au
ich dorf dorf hag itz
ich
dorf dorf dorf dorf ow in
au dorf ich hof
heim heim ich dorf au ow ow ow ow
hausen dorf dorf ow ow ow in ow
au ow ow ow berg dorf
heim hausen ing ing horst dorf au hof ow dorf dorf ow
burg
dorf ing hausen horst dorf dorf ow ow dorf ow dorf dorf
dorf
heim dorf heim leben itz dorf dorf ow itz
ing
dorf heim hausen hausen rod dorf au itz ow itz
rod hausen ach dorf itz itz itz
dorf
ich dorf ach rod ach ach dorf itz
grün
scheid ach ach
ing ach dorf rod
ing ach heim ach
ach heim ach ach hof
ach hof ing ing ing ing
ach ing ach ing ing
ach ing berg ing
au berg
berg

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

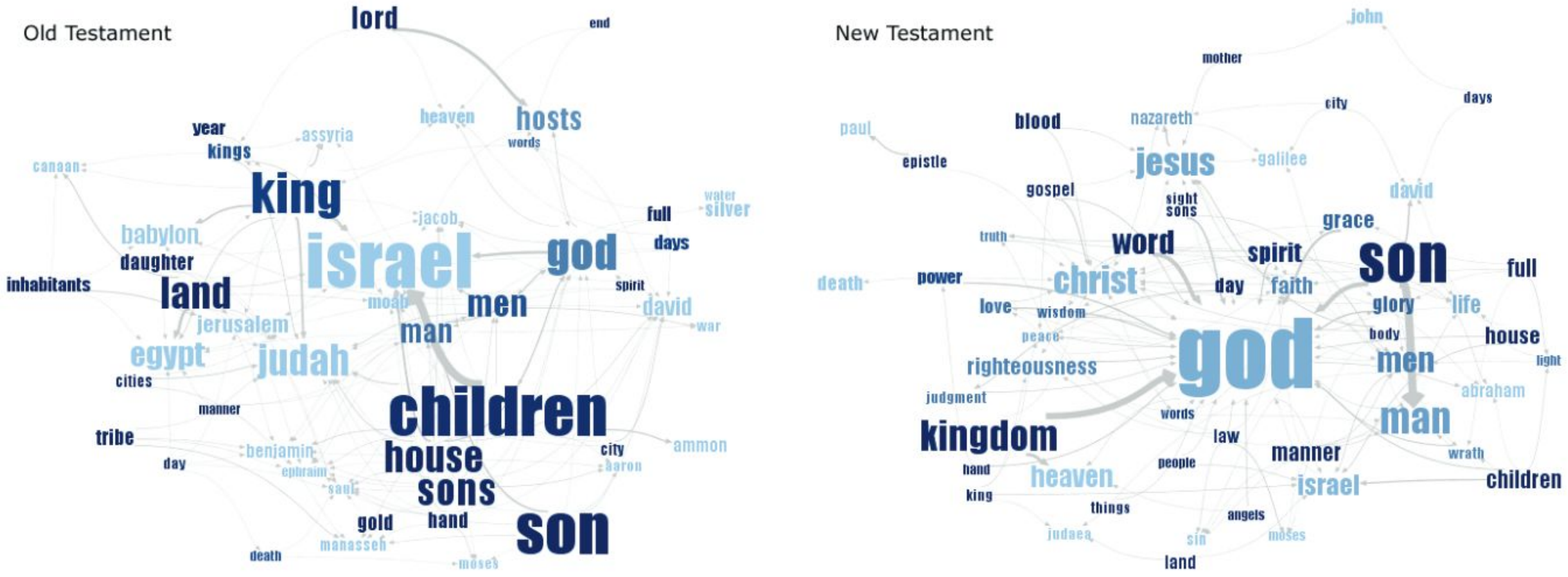
Figure source



eemed to throw some light
ral conclusions that I have
on
his excellent judgment. In considering
aining as yet unexplained in regard to
the origin
ating by his selection successive slight variations. I will then pass on to
an can be changed and perfected into a highly developed being or into
of
cal succession, and other such facts, might come to the conclusion
ntly entertained, and which I formerly entertained — namely, that
acknowledged varieties of any one species are the descendants of
that
s, applied to the whole animal and vegetable kingdoms. As many
tly entertained, and which I formerly entertained — namely, that
each
ions of the many beings which live around us. Who can explain why
t species, in the same manner as the acknowledged varieties of any
one
uccession, and other such facts, might come to the conclusion that species had not
uccession, and other such facts, might come to the conclusion that species had not
he mutual relations of the many beings which live around us. Who can explain why
called the same genera are lineal descendants of some other and generally extinct
species

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 thoughts on good visualization

Last lecture

Pre-attentive attributes are quickly recognized by our brain (size, color, position,...)

Hierarchy of graph elements: not all attributes are good for accurate quantitative reasoning

Gestalt principles: how brain connects elements into larger patterns (proximity, similarity, connection, enclosure, closure, continuity,)

Errors in visual processing lead to **illusions**

This informs our chart type choice (bars vs pies) and elimination of chart junk

Additional aspects of good plot choice

Basic setup: Selecting variables, choosing type of plot, assigning variables to x, y, color...

Data transformations: filtering (e.g. select data from one region), aggregating (e.g. summary per region) to avoid overplotting

Additional settings: sorting (e.g. bar graph columns), rescaling (log axis), re-expressing (e.g. absolute value vs relative change), zooming

Focus and explanation: highlighting, annotating (adding notes to plot)

Inspired by Stephen Few: Now you see it

Speed is not always everything

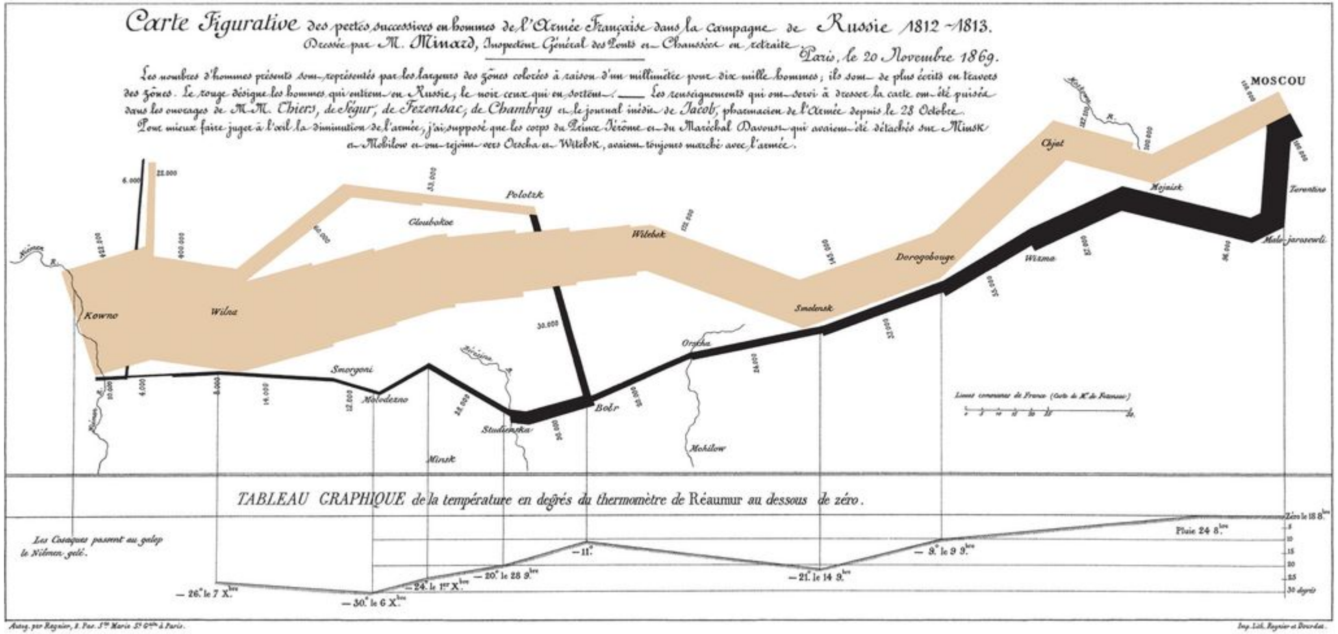
While there is a place for rapidly-understood graphs, it is too limiting to make speed a requirement in science and technology, where the use of graphs ranges from detailed in-depth data analysis to quick presentation. […]
The important criterion for a graph is not simply how fast we can see a result; rather it is whether through the use of the graph we can see something that would have been harder to see otherwise or that could not have been seen at all.

William Cleveland, The Elements of Graphing Data, Chapter 2

Recall: exploratory vs. explanatory analysis, sometimes audience can explore too

Recall Minard's plot of French army losses

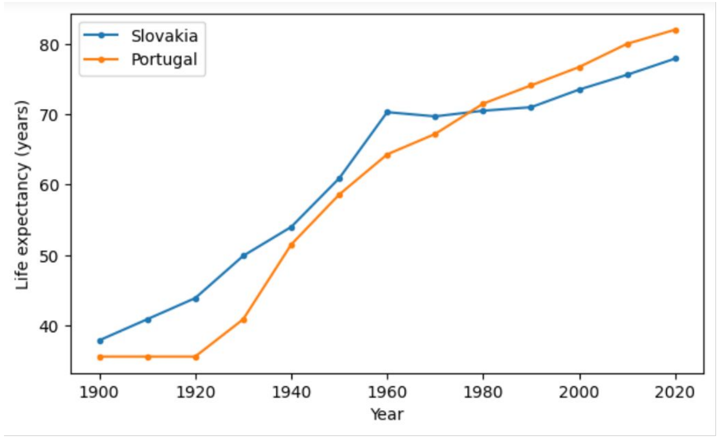
Easy to see big picture but also many minute details



https://commons.wikimedia.org/wiki/File:Minard.png

Tables vs. graphs

When is it good to include a table instead of / in addition to a graph?



	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010	202
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	77.
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**	**82.**

Tables vs. graphs

Advantages of tables:

- **Very few numbers** typically better given directly than in a graph
- In a long table, each reader can **find items** of personal interest (e.g. results of a sport competition, statistics for all countries)
- A table gives **exact values**
- Readers can **re-analyze** the same data (table preferably machine-readable)
- Numbers at very **different scales** are sometimes difficult to display even with log axes

See also
https://www.storytellingwithdata.com/blog/2011/11/visual-battle-table-vs-graph

Examples of bad graphs and their improvements

- http://www.perceptualedge.com/examples.php
- https://eagereyes.org/pie-charts
- https://viz.wtf/