

TOPIC: Data Visualization – some theory

SPEAKER: Jan Ewald, Laura Žigutytė

If not stated otherwise, shown images are licensed under CC-BY

<https://creativecommons.org/licenses/by/4.0/>

GEFÖRDERT VOM

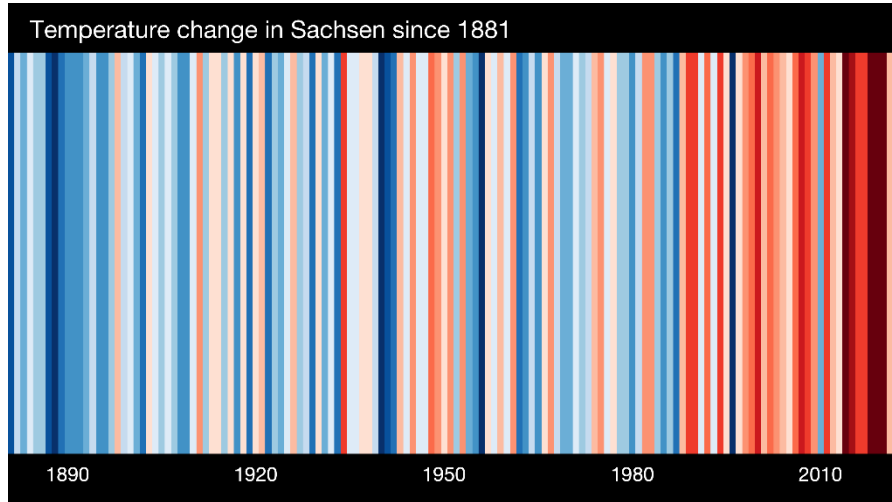


Bundesministerium
für Bildung
und Forschung



SACHSEN Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

Motivation – Example “warming stripes”



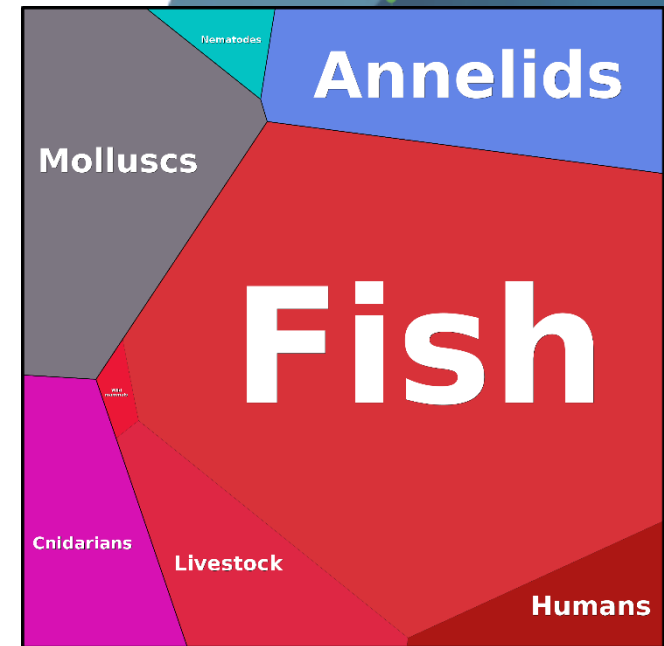
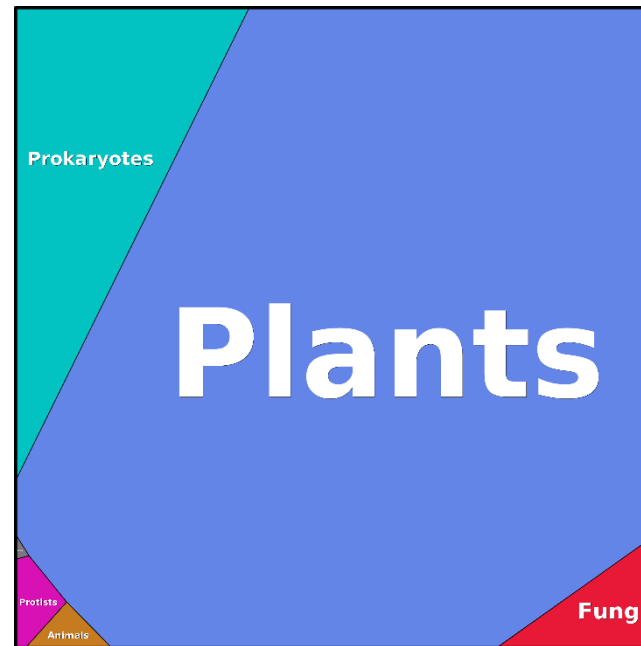
Ed Hawkins, University of Reading
<https://showyourstripes.info/l/europe/germany/sachsen>



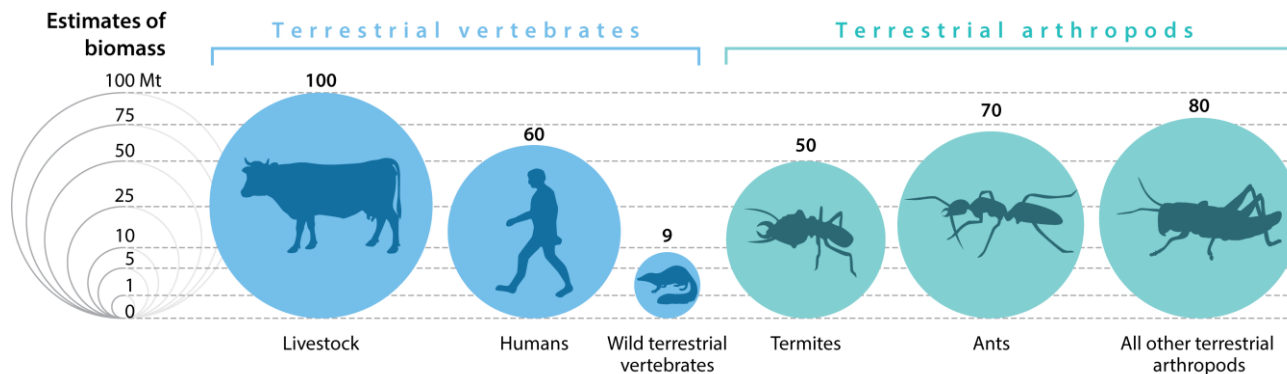
Steffenster, licensed under CC-BY-SA https://commons.wikimedia.org/wiki/File:Warming_Stripes_on_the_Sachsenbr%C3%BCcke_Leipzig.jpg

Motivation – biomass (1)

- Irregular shapes vs. quantification of areas
 - Voronoi-maps
 - Circles
- Color choices for species:
 - Plants are green
 - Fish -> ocean -> blue
 - Animal -> blood -> red



Self-created with <http://bionic-vis.biologie.uni-greifswald.de/> based on:
Bar-On, Yinon M., Rob Phillips, and Ron Milo. "The biomass distribution on Earth." PNAS 115.25 (2018): 6506-6511.



Eggleton P. 2020.
Annu. Rev. Environ. Resour. 45:61–82

Motivation – biomass (2)

- Rectangles: quantification of areas
- Icons for species and annotation

Life on Earth: the distribution of all global biomass

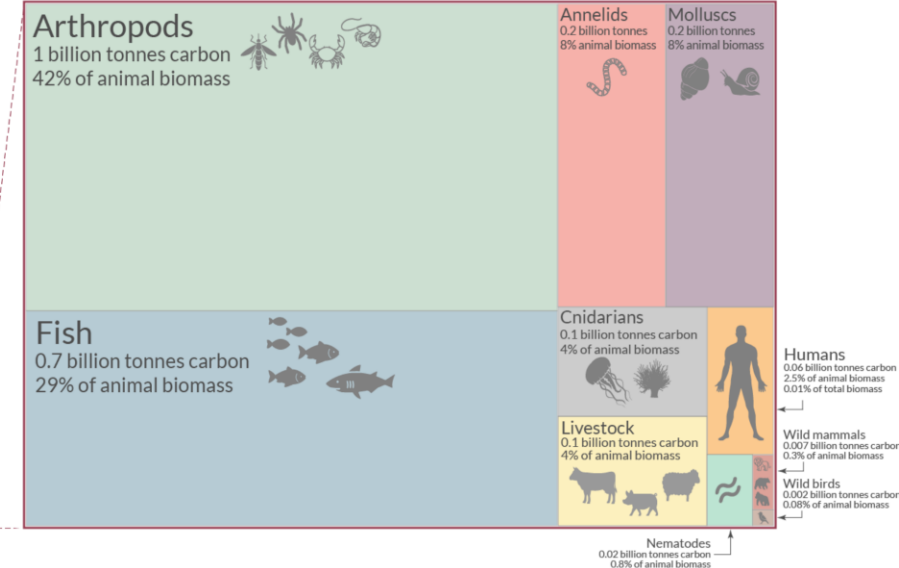
Biomass is measured in tonnes of carbon. The global distribution of Earth's biomass is shown by group of organism (taxa).

Our World
in Data

Global biomass: 546 billion tonnes of carbon



Animal biomass: 2 billion tonnes of carbon (0.4% of total biomass)



Data source: Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*. Icons from Noun Project.
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

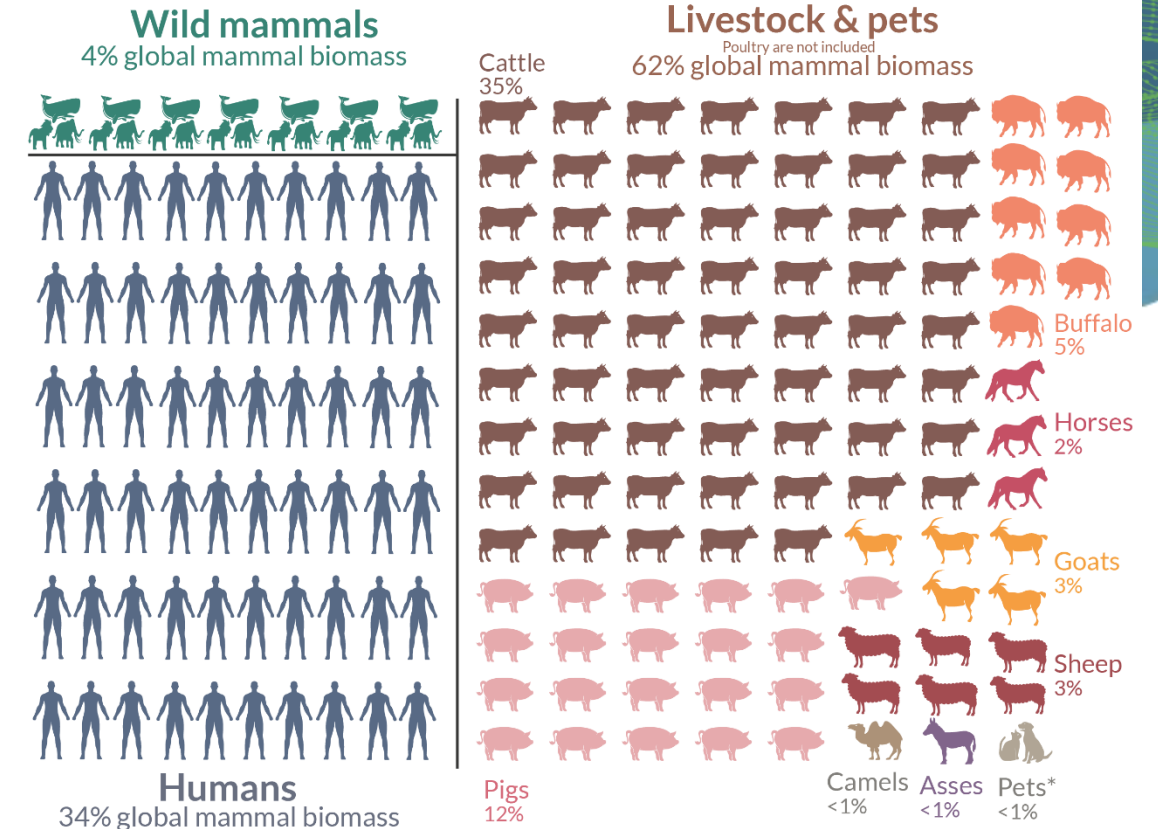
Motivation – biomass (3)

- Fixed size of icons: make data countable!
- Intuitive icons and partially colors

Distribution of mammals on Earth

Mammal biomass is shown for the year 2015.  or  or  = 1 million tonnes carbon (C)

Our World
in Data

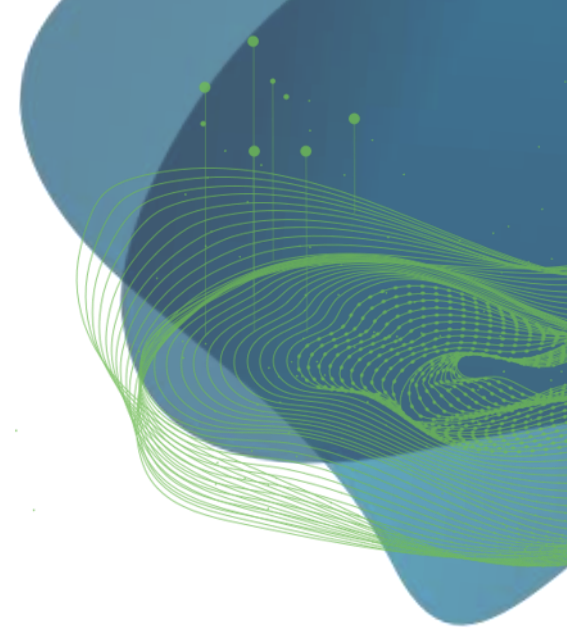


*Bar-On et al. (2018) provide estimates of livestock only, without estimates of mammalian pets (e.g. cats and dogs).
Pets have been added as an additional category based on calculations from estimates of the number of pets globally and average biomass.
Data source: Bar-On et al. (2018). The biomass distribution on Earth. Images sourced from the Noun Project.
OurWorldInData.org Research and data to make progress against the world's largest problems. Licensed under CC-BY by the author Hannah Ritchie.



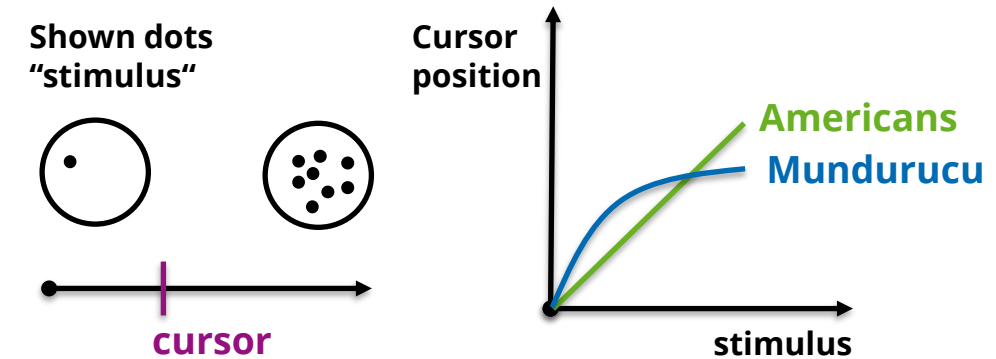
Goals of data visualization

- Summarizing, reduction, different views on data
- Exploration, discover patterns, generation of hypotheses
- Reasoning, justification
- Planning, scheduling, resource distribution
- Easier to transport and memorize content
- Stimulation and creativity
- ...



Why are there “good” / “bad” visualizations?

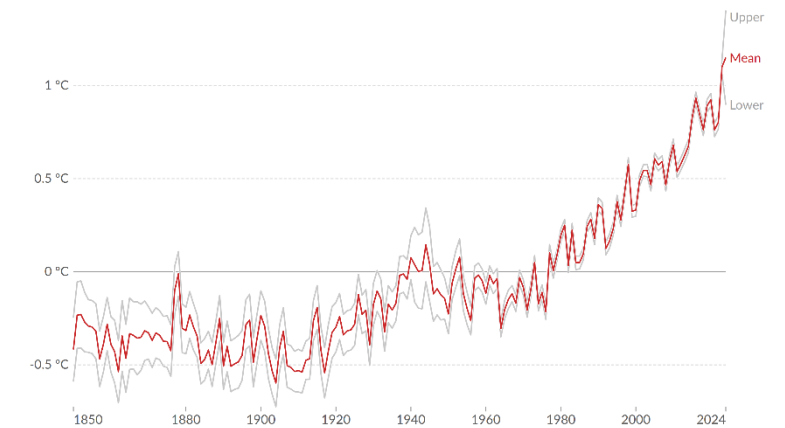
- [Apparent content errors and deceptions]
- biological and physical limitations of human eyes & brain
 - Max. number of distinctive shades of grey
 - Bad quantification of areas and angles
 - Color blindness
 - ...
- Socio-cultural habits and practices
 - Reading directions (left-to-right, top-to-bottom)
 - Different number systems or scales
 - Associations of colors with traits and characteristics (red: warm/attention/signal)
 - ...
- Educational and research field habits and practices
 - Familiarity and interpretation of (complex) plot types (box-plot, heatmaps ...)
 - Trade-off between information accuracy and reduction
 - Visualization and comprehension of errors and uncertainty
 - ...



Dehaene, Stanislas, et al. "Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures." *Science* (2008)

Average temperature anomaly, Global

Global average land-sea temperature anomaly relative to the 1961-1990 average temperature.



Data source: Met Office Hadley Centre (2023)

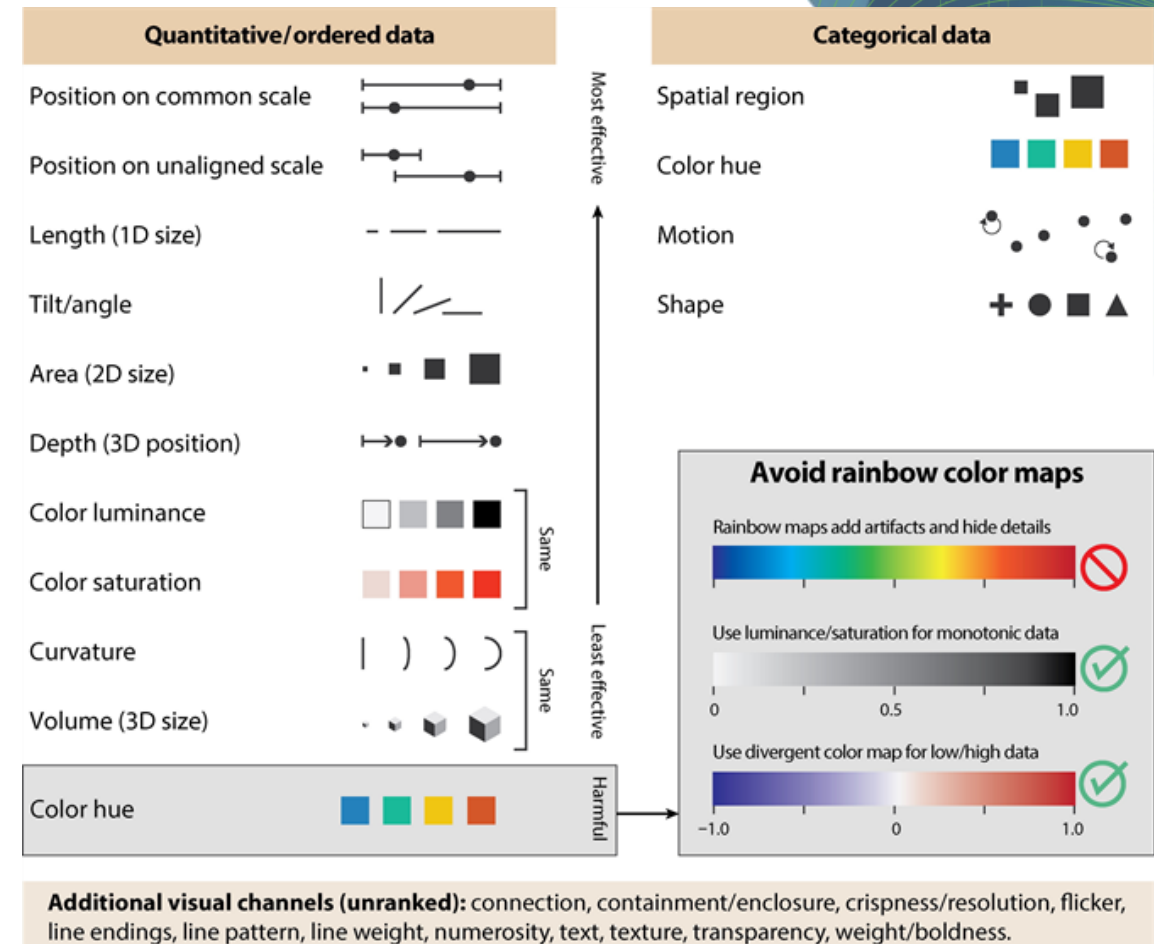
OurWorldInData.org/co2-and-greenhouse-gas-emissions | CC BY

Note: The gray lines represent the upper and lower bounds of the 95% confidence intervals.

Basic elements of visualization (1)

Visual variables

- Position
- Size, length, area
- Shape
- Color: hue, brightness, saturation
- Orientation
- Texture, grain
- ...

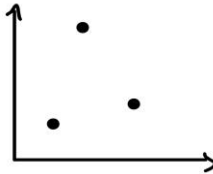
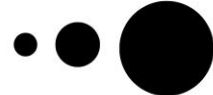

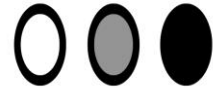





Ikuomenisan, G. and Morgan, Y. (2022) Systematic Review of Graphical Visual Methods in Honeypot Attack Data Analysis. Journal of Information Security, 13, 210-243. <https://doi.org/10.4236/jis.2022.134012>

Basic elements of visualization (1)

Important traits of visual variables

- Selective (group → 1)
- associative (n → cluster)
- Quantitative
- Order (* > * > * > *)
- Distinctive

Visualization (1)		Characteristics					
Visual Variables		Selective	Associative	Quantitative	Order	Length	
	Position		yes	yes	yes	yes	infinite
	Size		yes	no	partially	yes	Selection: ~ 5 Distinction: ~ 20
	Shape		no	mostly	no	no	Infinite
	Value		yes	no	no	yes	Selection: < 7 Distinction: ~ 10
	Color		yes	yes	no	no	Selection: < 7 Distinction: ~ 10
	Orientation		yes	yes	no	no	~5 (Infinite)
	Texture		yes	yes	no	multi	infinite

Theoretical aspects – data characteristics

Data characteristics strongly influence their suitable visualization

- Data type

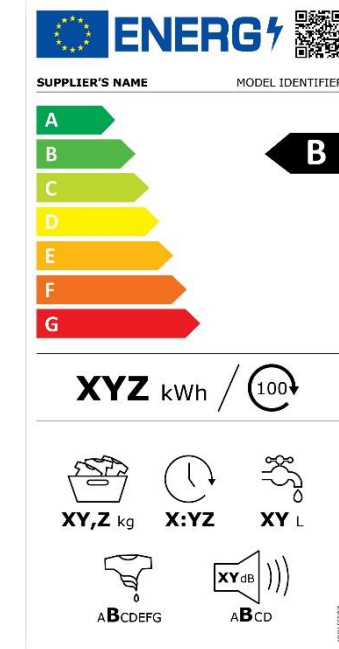
- Nominal, categorical



- Ordinal

- Quantitative:

- scale,
 - range,
 - relation,
 - unit

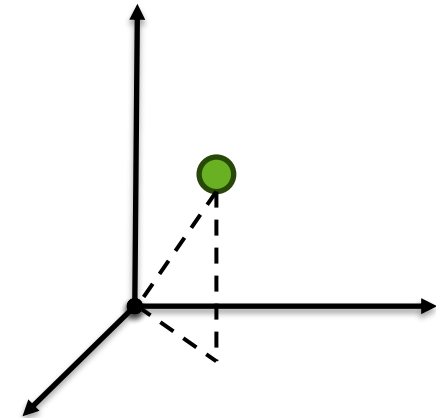
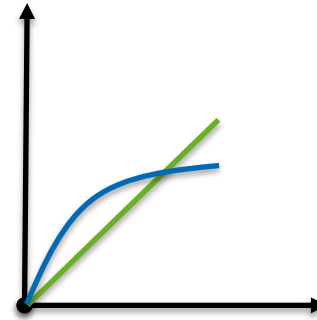


© European Union, <https://eur-lex.europa.eu/>, 1998-2024.
This item is from a European Union agency or department which as its official copyright policy cites the European Union Commission Decision of 12 December 2011, allowing free use for purposes both commercial and non-commercial as long as attribution is given.

Theoretical aspects – data characteristics

Data characteristics strongly influence their suitable visualization

- Dimensionality
 - 1D, series
 - 2D, 3D
 - N-dimensions

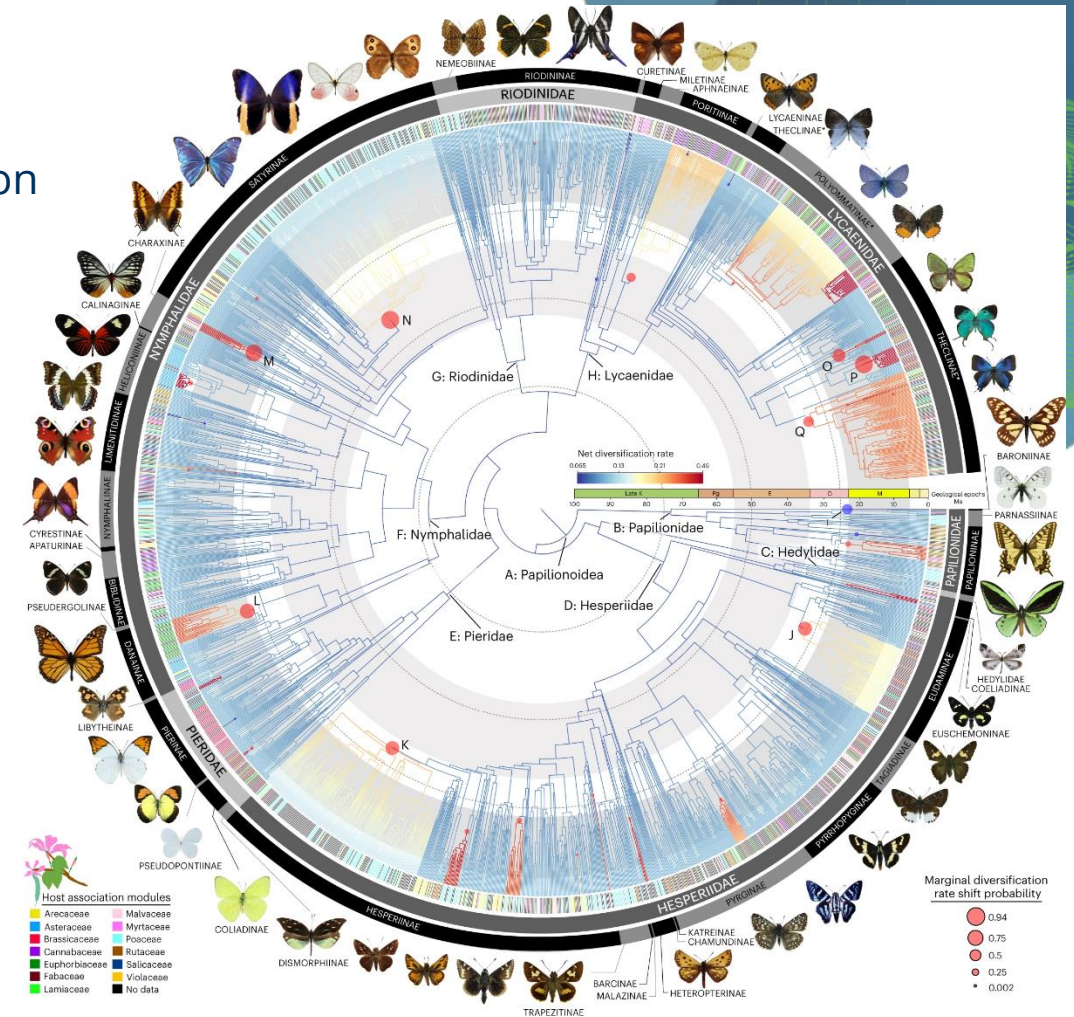
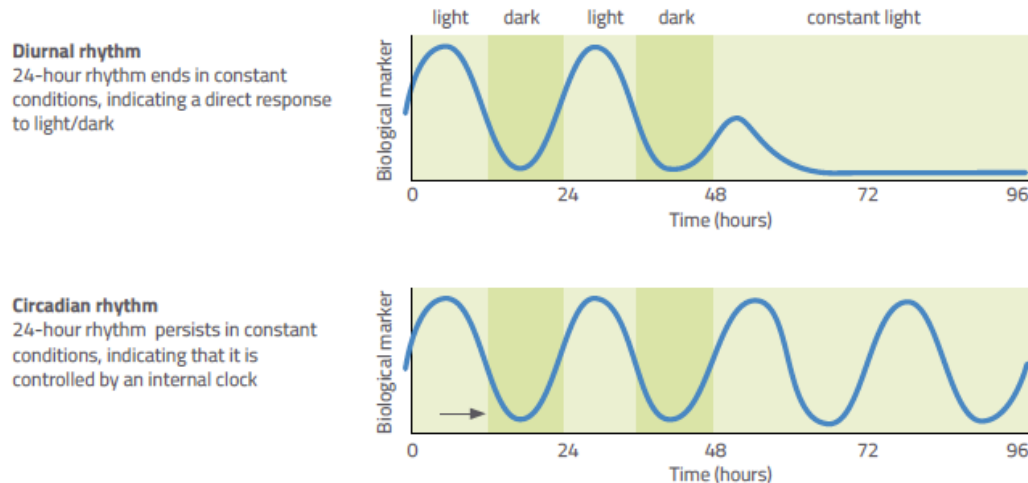




Theoretical aspects – data characteristics

Data characteristics strongly influence their suitable visualization

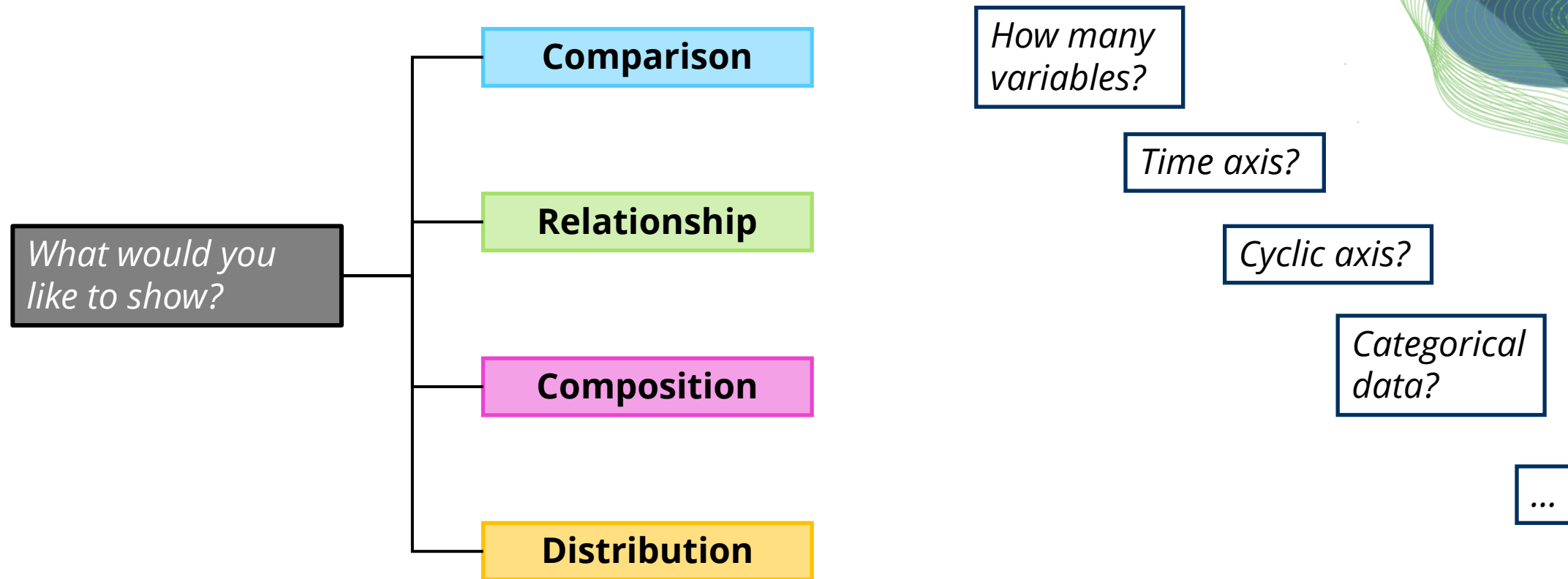
- Other aspects
 - Graphs and networks (+trees)
 - Temporal data and time series
 - Cyclic data and axis



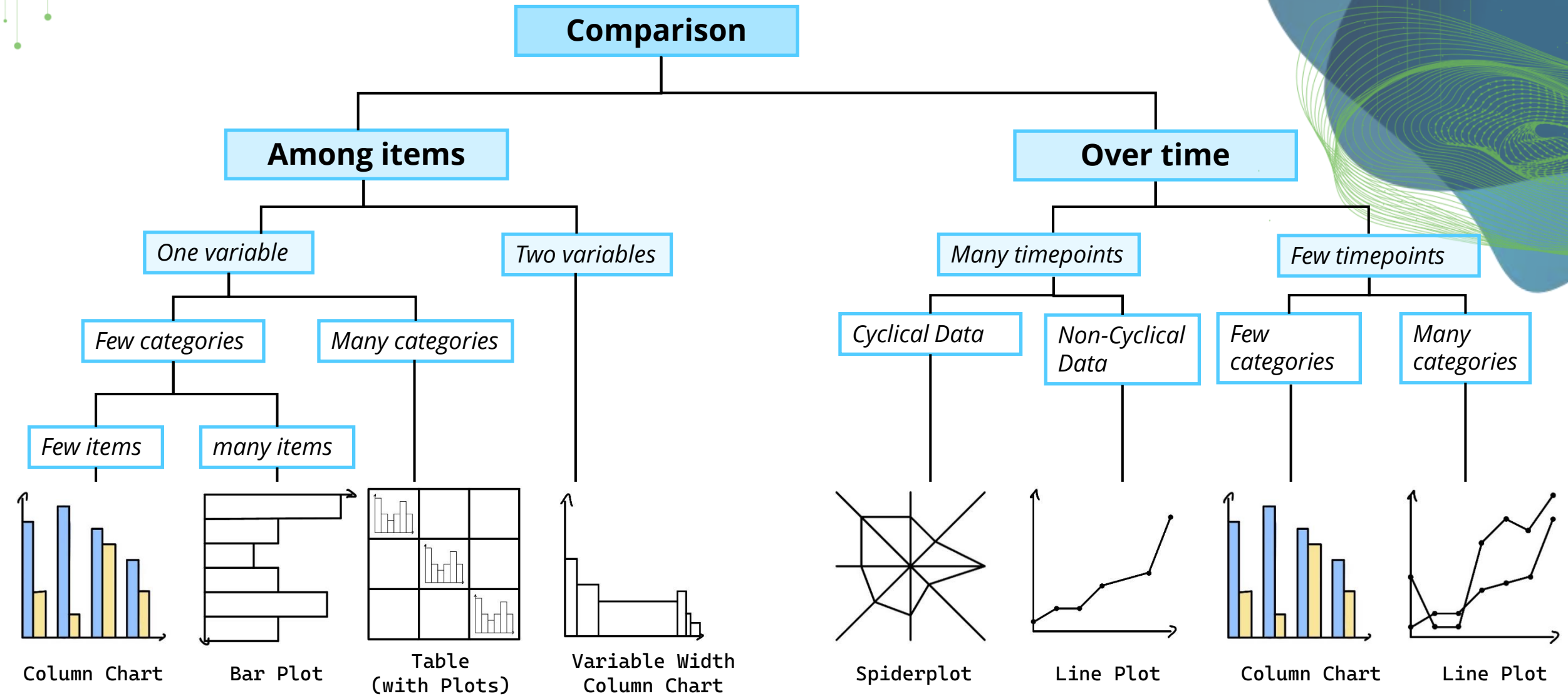
Katharine Hubbard | www.scienceinschool.org | Science in School | Issue 48 : Autumn 2019 | 11

Kawahara, A.Y., Storer, C., Carvalho, A.P.S. et al. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nat Ecol Evol* 7, 903–913 (2023). <https://doi.org/10.1038/s41559-023-02041-9>

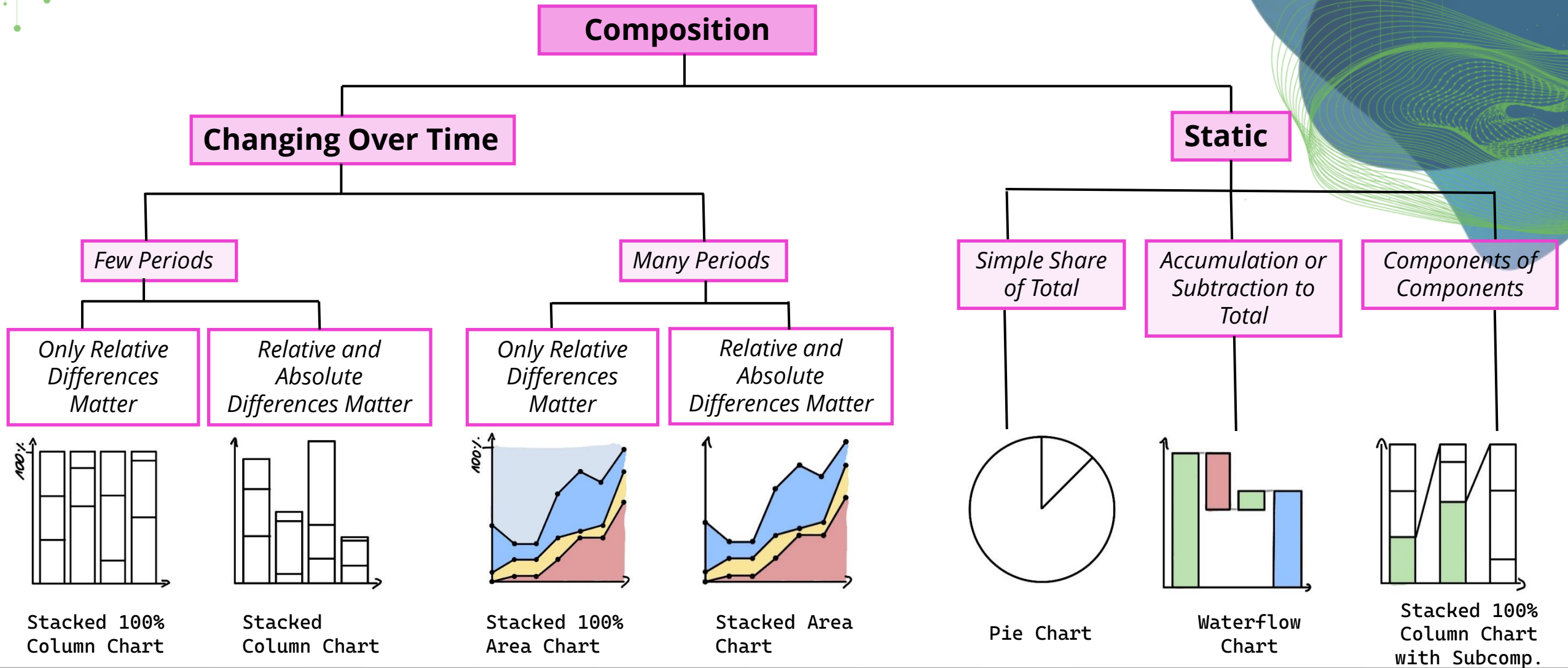
Theoretical aspects – How to decide on plot types



Theoretical aspects – How to decide on plot types (2)



Theoretical aspects – How to decide on plot types (3)



Theoretical aspects – How to decide on plot types (4)

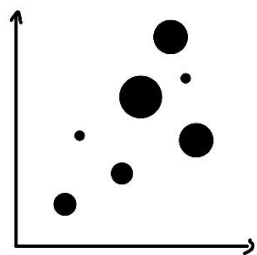
Relationship

Two Variables



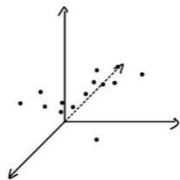
2D Scatterplot

3 Variables



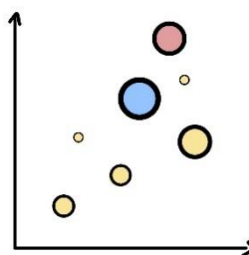
Scatterplot with 1 additional attribute (e.g. Bubble Plot)

Interactive?

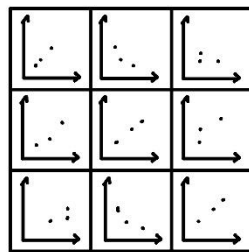


3D Scatterplot

More Variables



Scatterplot with additional attributes

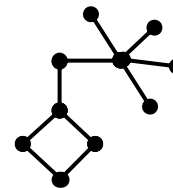


Scatterplot Matrix

Hierarchy

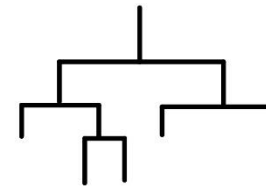
+Composition

other



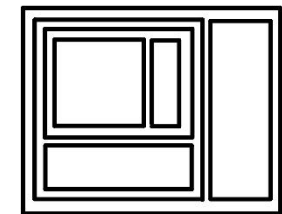
Networks

Tree-hierarchy



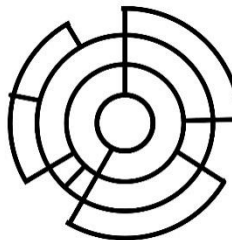
Rooted or unrooted Trees

Few categories



TreeMap

Many categories



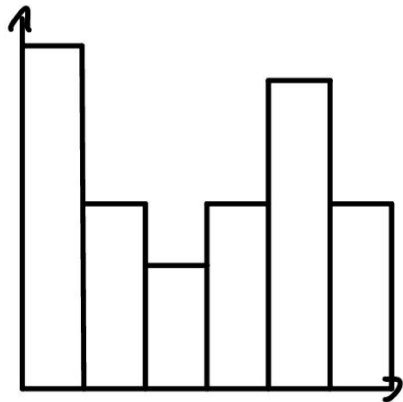
Sunburst

Theoretical aspects – How to decide on plot types (5)

Distribution

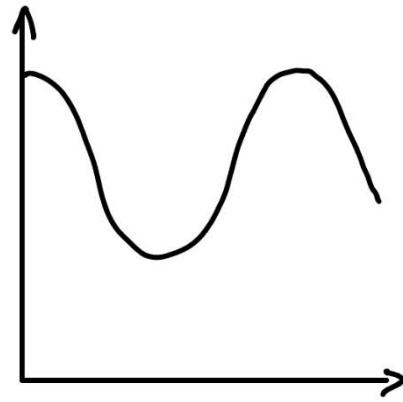
Single Variable

Few Data Points



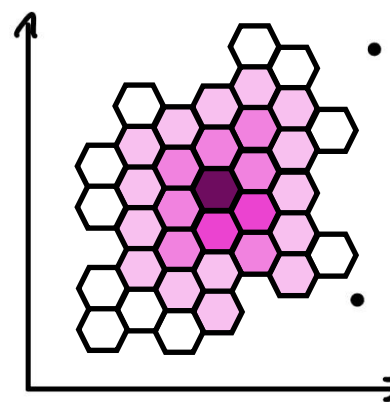
Histogram

Many Data Points



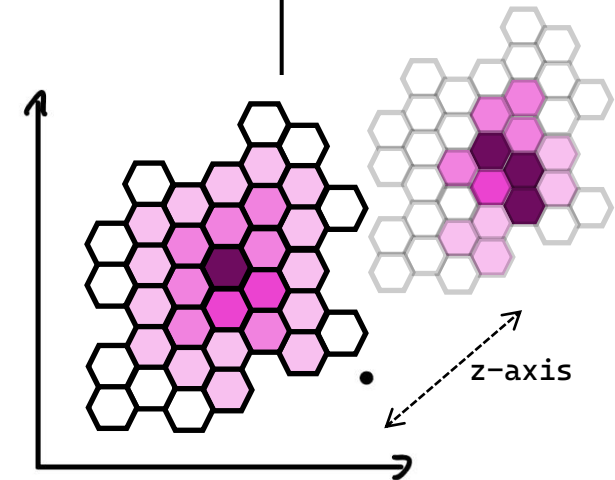
Density Plot
(line)

Two Variables



2D Histogram

Three Variables

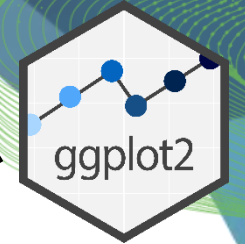


Interactive 2D Histogram

Comparison of plotting interfaces and their philosophies



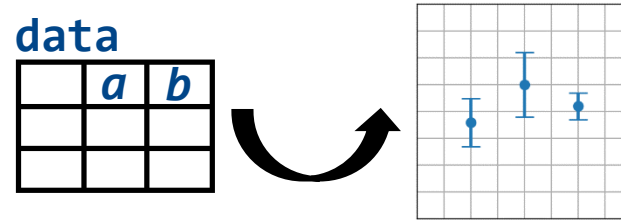
seaborn.objects



The Grammar of Graphics,
Leland Wilkinson, 2005
<https://doi.org/10.1007/0-387-28695-0>

Plot type driven ←————→ Data and geometry driven

Comparison of plotting interfaces and their philosophies (2)



seaborn

seaborn.objects

```
x=data['a'].drop_duplicates()
y=data['b'].groupby(['a']).mean()
yerr=data['b'].groupby(['a']).sd()

errorbar(x,y,yerr)
```

```
pointplot(
    data,
    x='a', y='b',
    errorbar='sd'
)
```

```
so.Plot(data,x='a', y='b')
.add(    so.Dot(),
        so.Agg())
.add(    so.Range(),
        so.Est(errorbar='sd'))
```

- 1) Choose plot type
- 2) Extract or calculate variables
- 3) Stuff into plot API

- 1) Choose basic type
- 2) Define data and variables

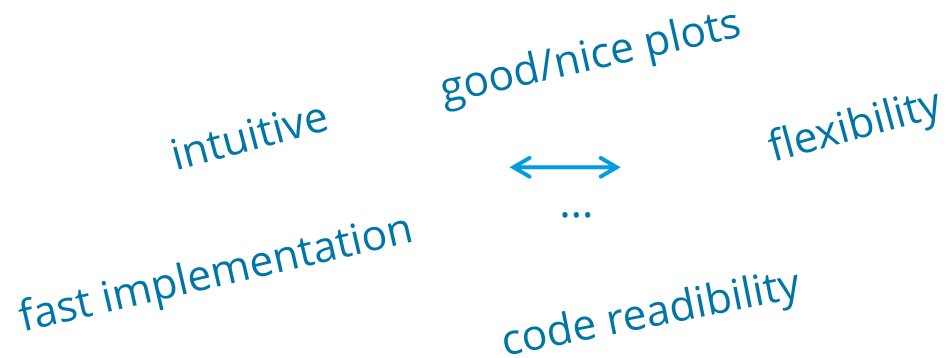
- 1) Define data and variables
- 2) Choose plot geometries
- 3) Define statistics

Plot type driven ← → Data and geometry driven

Comparison of plotting interfaces and their philosophies (3)



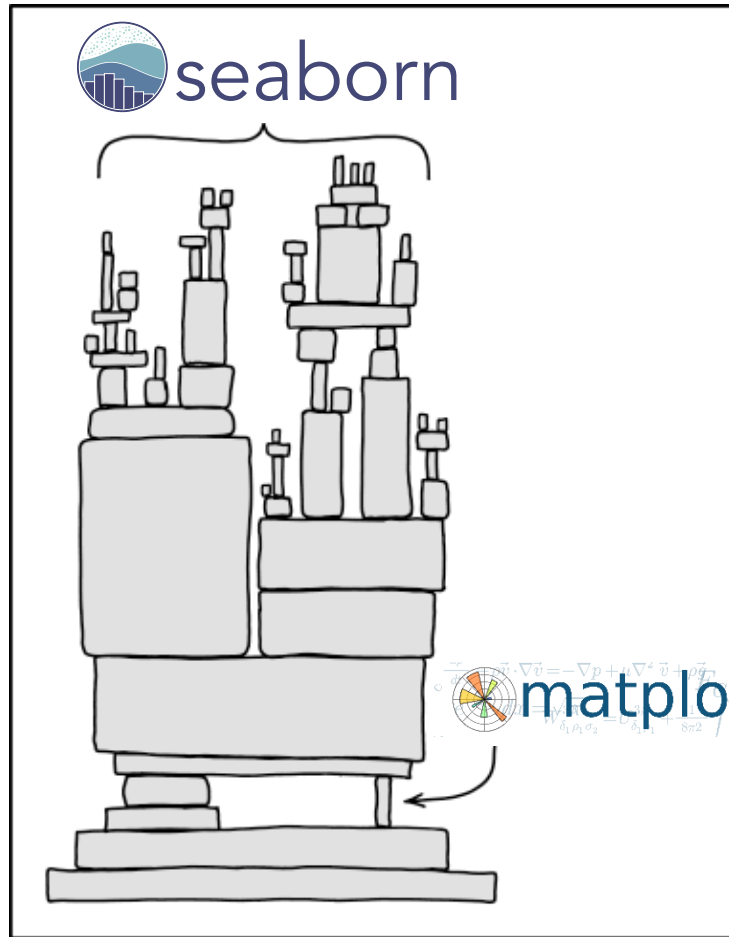
seaborn.objects



**What do you see
as pro's and con's?**

Plot type driven ← —————→ **Data and geometry driven**

History and problem of plotting libraries in python



MATLAB
plot interface

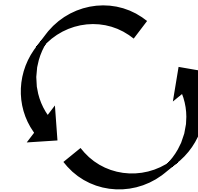
Adapted from <https://xkcd.com/2347>, licensed under CC-BY-NC <https://creativecommons.org/licenses/by-nc/2.5/>

Data format: keep it tidy, save time later

- Mostly long-format is preferable for plotting
- Pandas and other libraries have functions to convert formats
- Keep and make data tidy before plotting saves a lot of time and work

long format

wide format

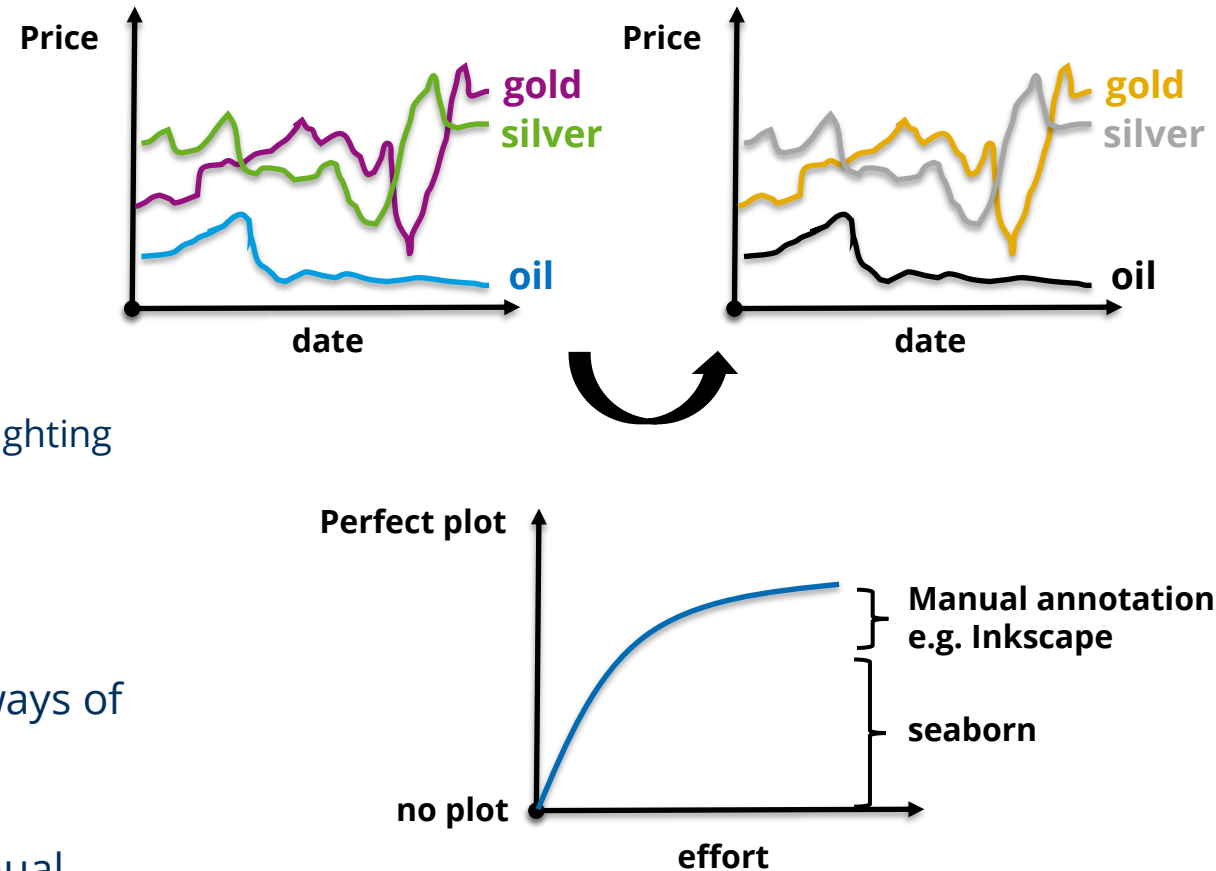


Athlets	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Points
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7	8217
CLAY	10.76	7.4	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5	8122
KARPOV	11.02	7.3	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2	8099
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1	8067

Athlets	Discipline	Value
SEBRLE	100m	11.04
SEBRLE	Long.jump	7.58
SEBRLE	Shot.put	14.83
SEBRLE	High.jump	2.07
SEBRLE	400m	49.81
SEBRLE	110m.hurdle	14.69
SEBRLE	Discus	43.75
SEBRLE	Pole.vault	5.02
SEBRLE	Javeline	63.19
SEBRLE	1500m	291.7
SEBRLE	Points	8217
CLAY	100m	10.76
CLAY	Long.jump	7.4
CLAY	Shot.put	14.26
CLAY	High.jump	1.86
CLAY	400m	49.37
CLAY	110m.hurdle	14.05
CLAY	Discus	50.72
CLAY	Pole.vault	4.92
CLAY	Javeline	60.15
CLAY	1500m	301.5
CLAY	Points	8122

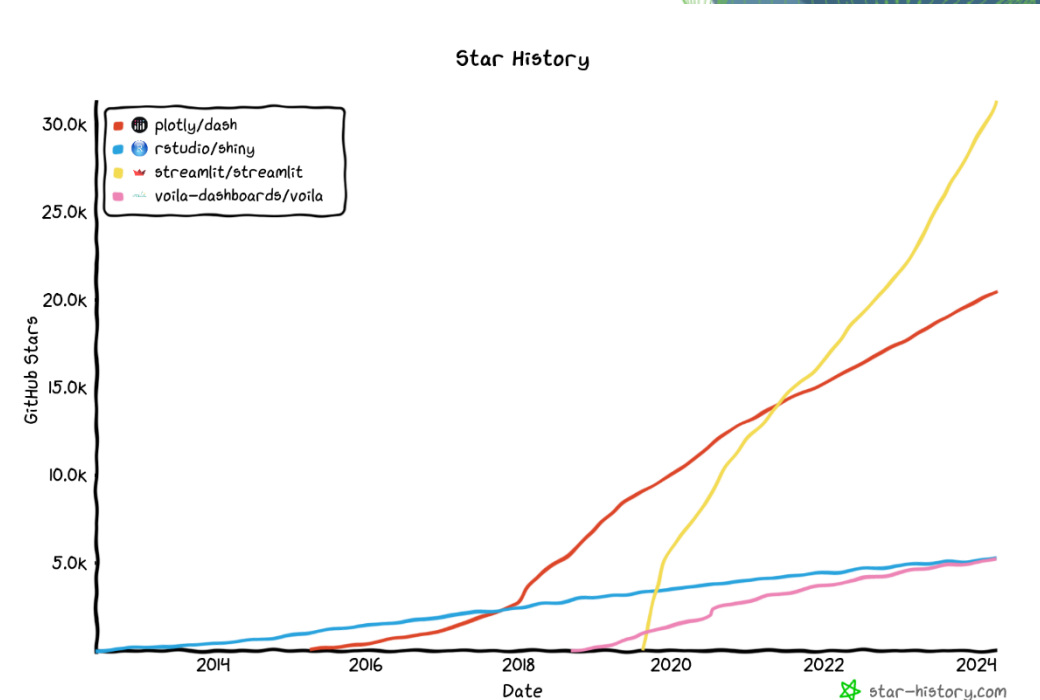
Dos and Don'ts

- Usage of colors, shapes and textures
 - **Consistency!**
 - Intuitive choice of colors etc. (plants → green)
- Purpose and goal determine the visualization
 - Data exploration: high complexity and many details
 - For a publication: **1-2 key messages per plot**, use highlighting and annotations
- Think about the **visual habits** of the target audience
- There is **no single best** visualization, but **many bad** ways of visualization
- Find a good trade-off between programming and manual adjustments or annotations



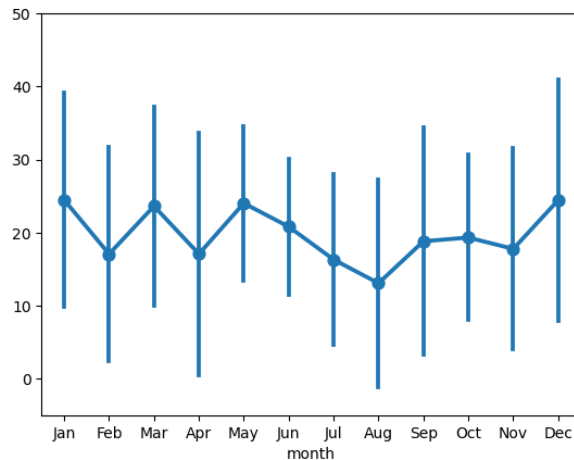
Dos and Don'ts

- Output format
 - Prefer loss-less zoomable formats: svg, pdf
 - Raster vs. vector files
- Be careful with 3D plots
 - Think about if really necessary?
 - Recommended if interactive and not as static plot
- **Interactive plots**, and dashboards
 - Highly recommended, very empowering, more and more common
 - Problem: not supported in static publications like PDF and most journals
 - Examples:
 - <https://plotly.com/examples/>
 - <https://shiny.posit.co/r/gallery/>
 - ...

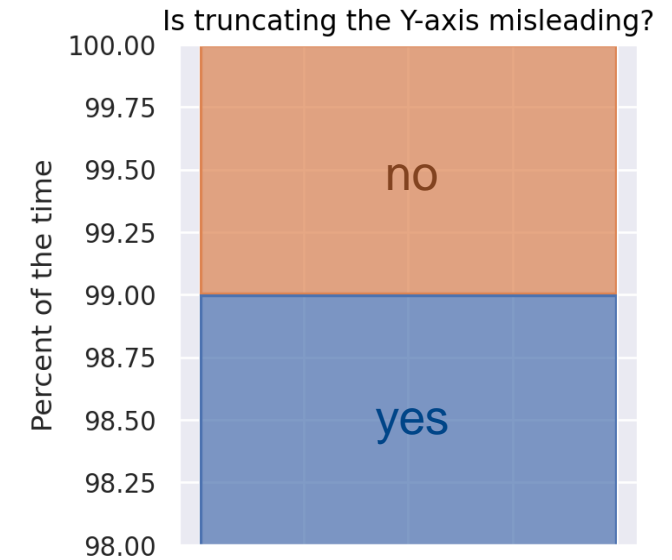
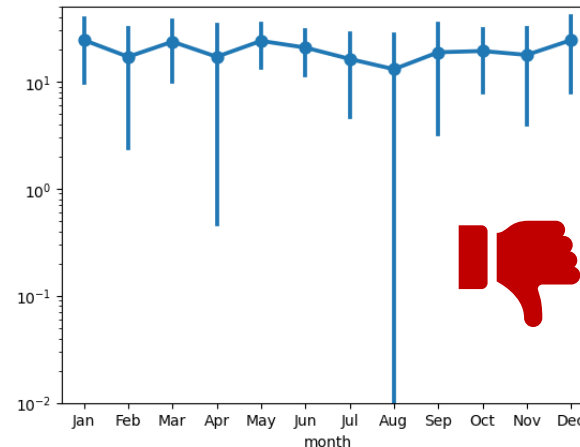


Dos and Don'ts: "Chart-crimes"

- Cheating with Y-axis
 - Multiple Y-axis
 - Free Y-axis, cutted Y-axis
 - No axis at all ...
- Problems with log-scales
 - Zero values
 - Not suitable for error-bars, box-plots



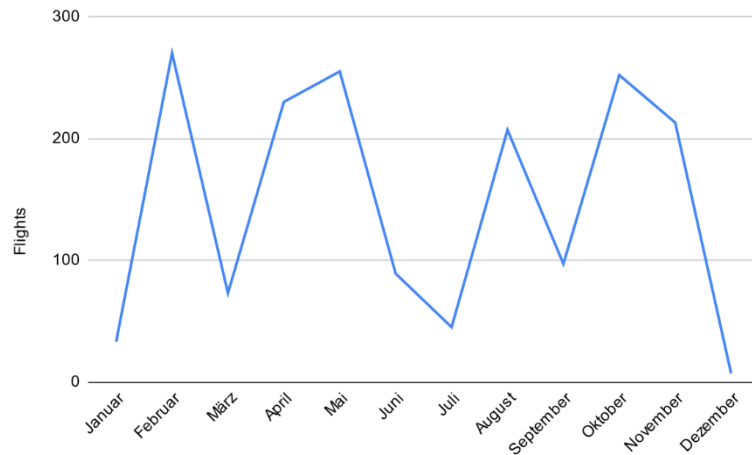
Y-axis
log-scale



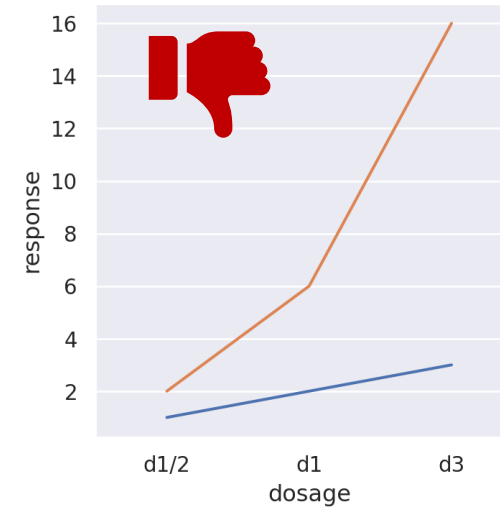
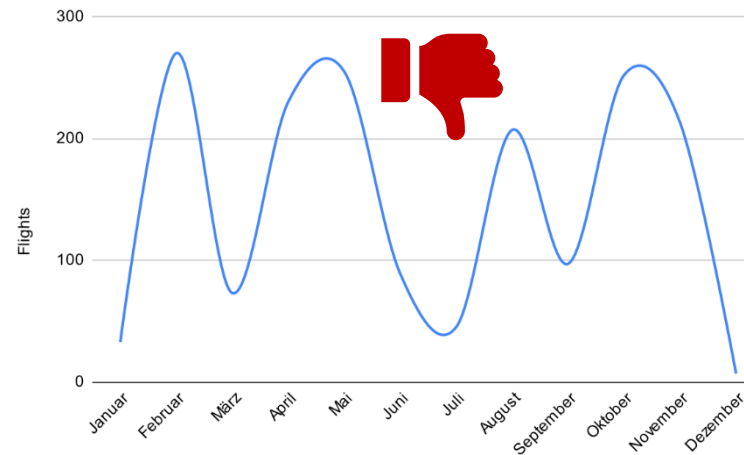
Original meme probably from WyoWeeds
<http://imgur.com/HZe4vKy>

Dos and Don'ts: "Chart-crimes"

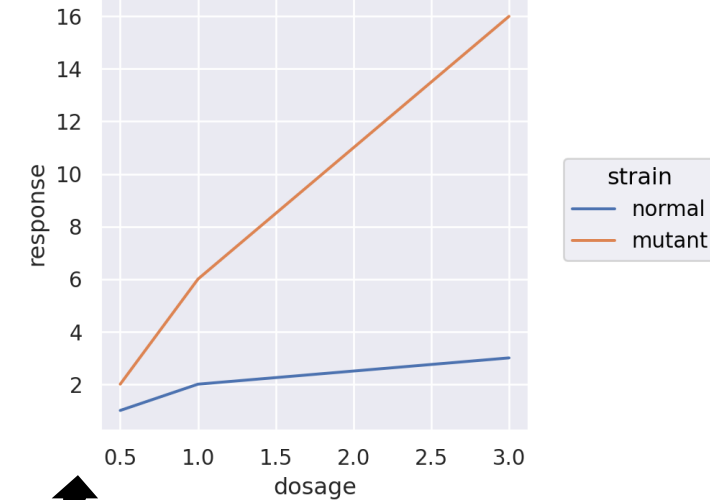
- Line-plots
 - Line-plots not suitable if X-axis is categorical or non-quantitative
 - Spaghetti line-plots
 - In-appropriate use of splines (smoothing)



faking resolution

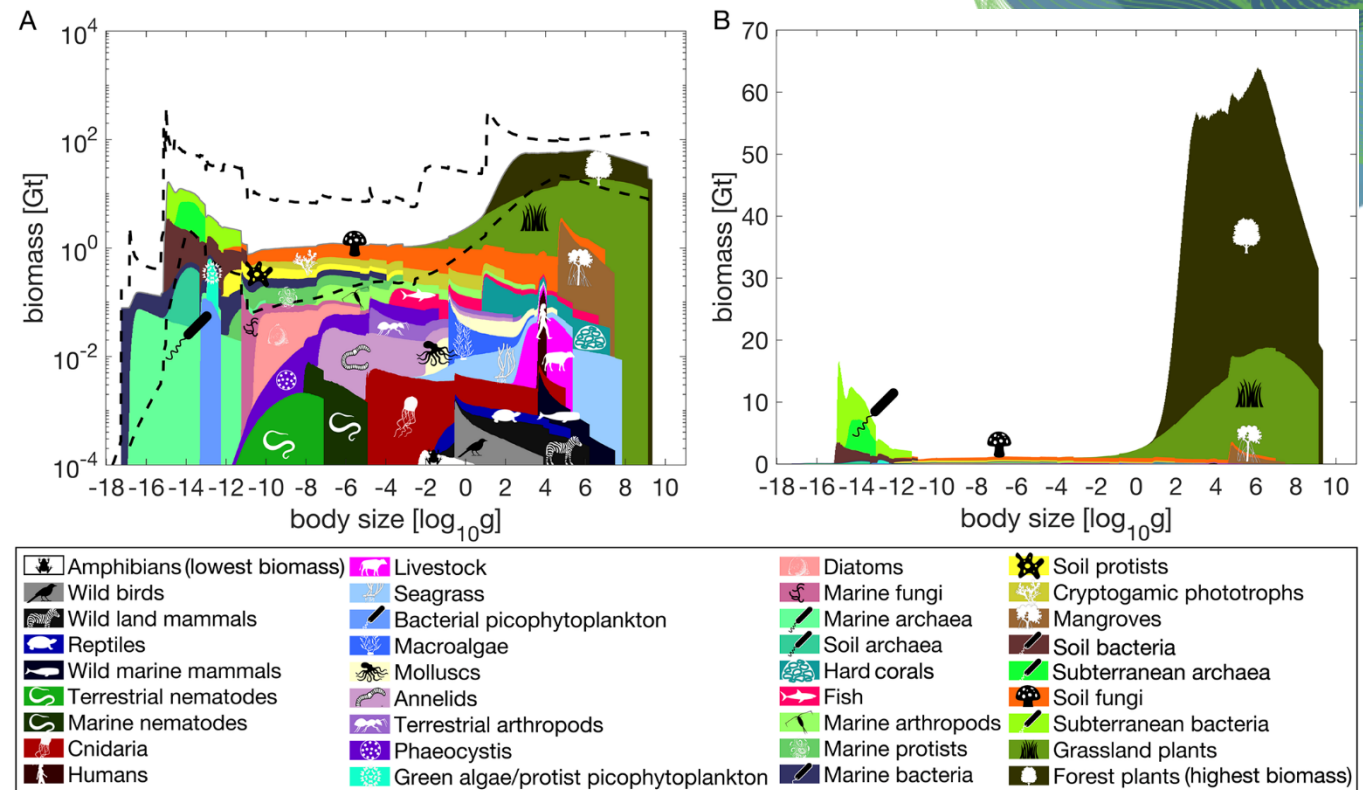


x-axis is quantitative



Dos and Don'ts: "Chart-crimes"

- Not proportional visualization of data
- Missing axis descriptions, legends or plain wrong content representation
- Plots with areas and angles
 - Pie-charts, especially 3D
 - Are areas proportional to data?
- High complexity,
 - ask yourself "Do I need more than 3":
 - Colors
 - Categories
 - Lines
 - Annotations
 - ...



Tekwa EW, Catalano KA, Bazzicalupo AL, O'Connor MI, Pinsky ML (2023) The sizes of life. PLoS ONE 18(3): e0283020.

<https://doi.org/10.1371/journal.pone.0283020>

Resources and further reading

- Prof. Sheelagh Carpendale
 - <https://www.cs.sfu.ca/~sheelagh/>
 - Online lecture <https://www.youtube.com/watch?v=geQcMZV8LZs>
- Books
 - <https://www.storytellingwithdata.com/books>
 - The Grammar of Graphics <https://books.google.de/books?id=YGgUswEACAAJ>
- Online resources
 - <https://seaborn.pydata.org/>
 - <https://matplotlib.org/>
 - <https://www.data-to-viz.com/>
 - <https://r-graph-gallery.com/>
 - Checklist <https://ly.uxlib.net/assets/subject/data-viz/datacated-visual-best-practices-checklist.pdf>
- Dashboard
 - <https://shiny.rstudio.com/tutorial/>
 - <https://plotly.com/examples/>

Hands on sessions



seaborn.objects

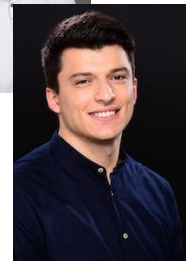
Day 3.3a „Basic Plotting“
13:30–15:00 Room: „Zwenkauer See“ (?)

Day 3.3b „Advanced Plotting“
13:30–15:00 Room: „Markkleeberger See“ (?)



Step-by-step introduction to
plot with standard seaborn
API

Introduction to
seaborn.objects API and
high-level visualization
tips & tricks



Plot type driven ← ————— → Data and geometry driven