Foundations of Data Science

Lecture 3: Visualization

Prof. Gilles Louppe
g.louppe@uliege.be

LIÈGE université

Encoding data with visual cues

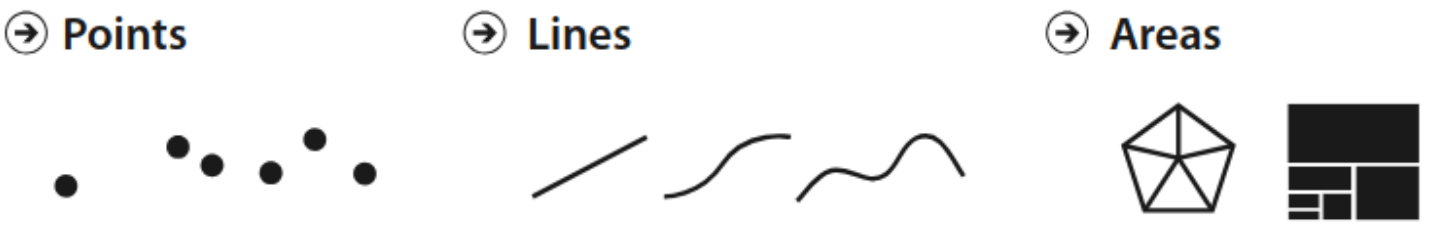
14, 37, 75

In pairs, try to come up with as many representations/encodings of this "data".

Visual cues

Visual cues are elements of a visualization that encode data. They are expressed as marks and channels:

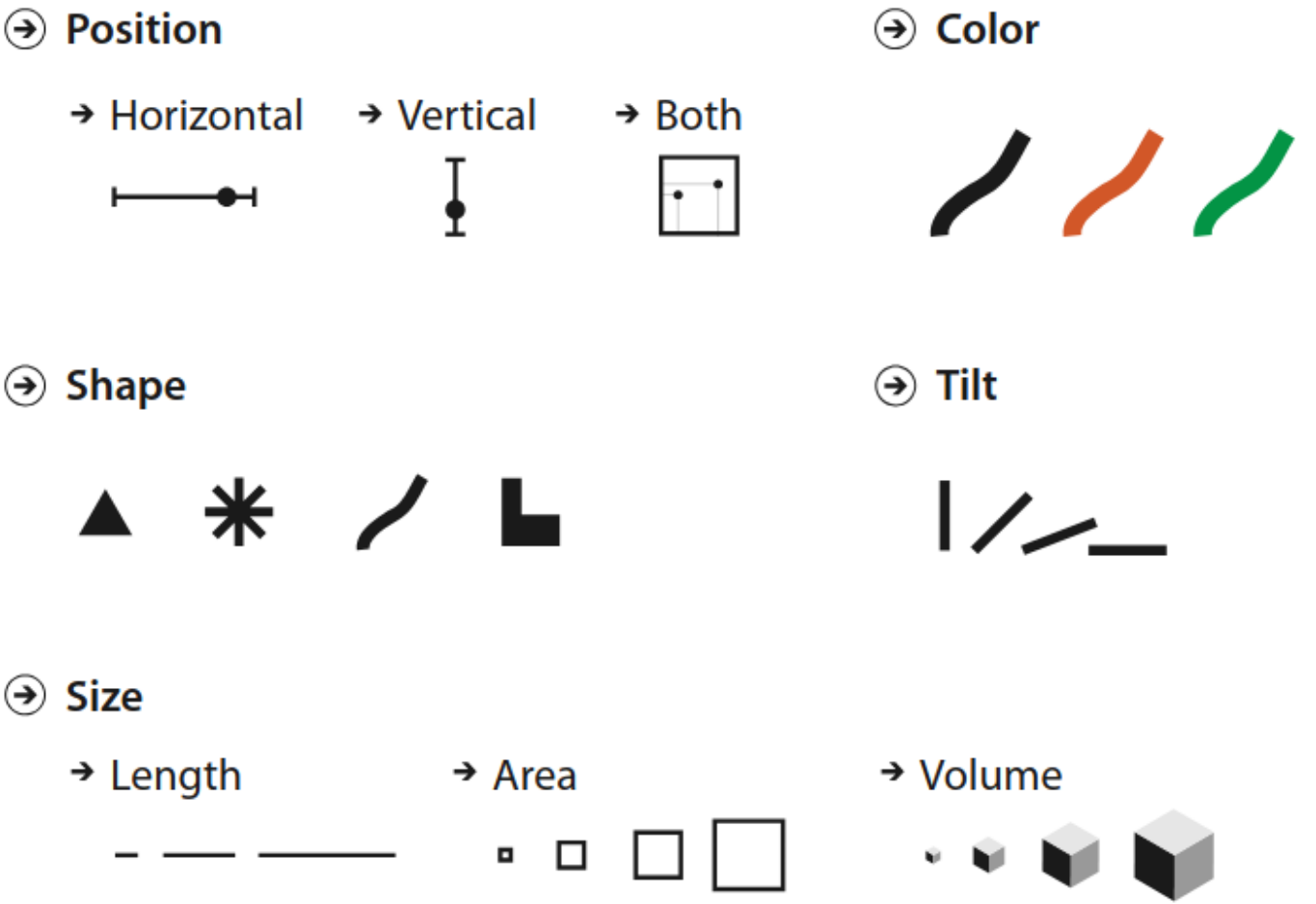
- A **mark** is a geometric primitive such as points, lines, or areas.
- A **channel** is an attribute of a mark that can be used to encode data, such as position, size, shape, or color.



Marks are geometric primitives.

—

Credits: T. Munzner, "Visualization Analysis and Design", 2014.



Visual channels control the appearance of marks.

—

Credits: T. Munzner, "Visualization Analysis and Design", 2014.

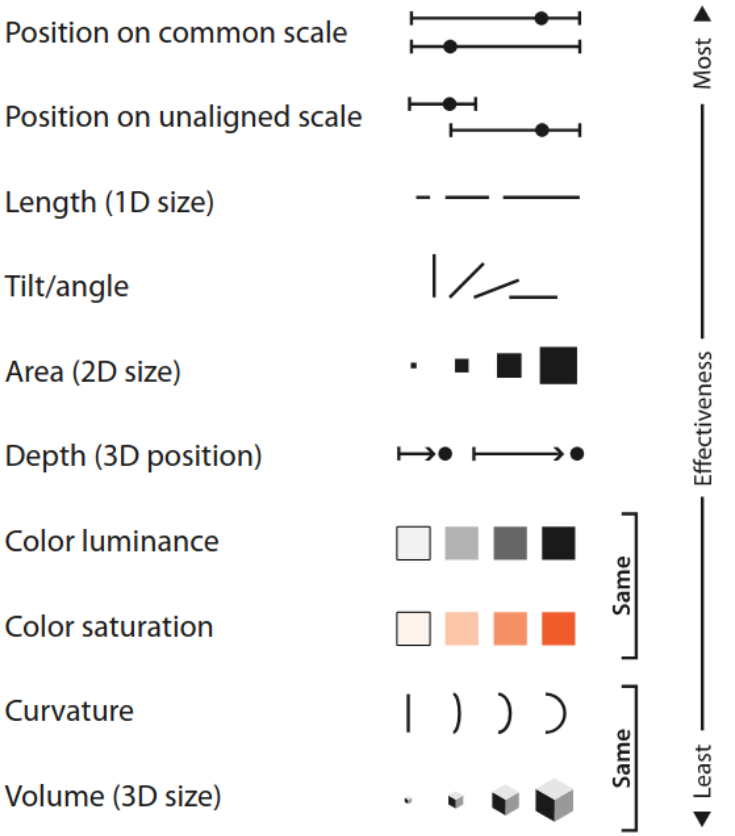
(a) Bar charts encode data using line marks, controlled by a vertical position channel (height) and a horizontal position channel (category). (b) Scatterplots encode data using point marks, controlled by two position channels (x and y). (c) A third variable can be encoded using a color channel. (d) A fourth variable can be encoded using a size channel (area of the point).

—

Credits: T. Munzner, "Visualization Analysis and Design", 2014.

Perceptual hierarchy

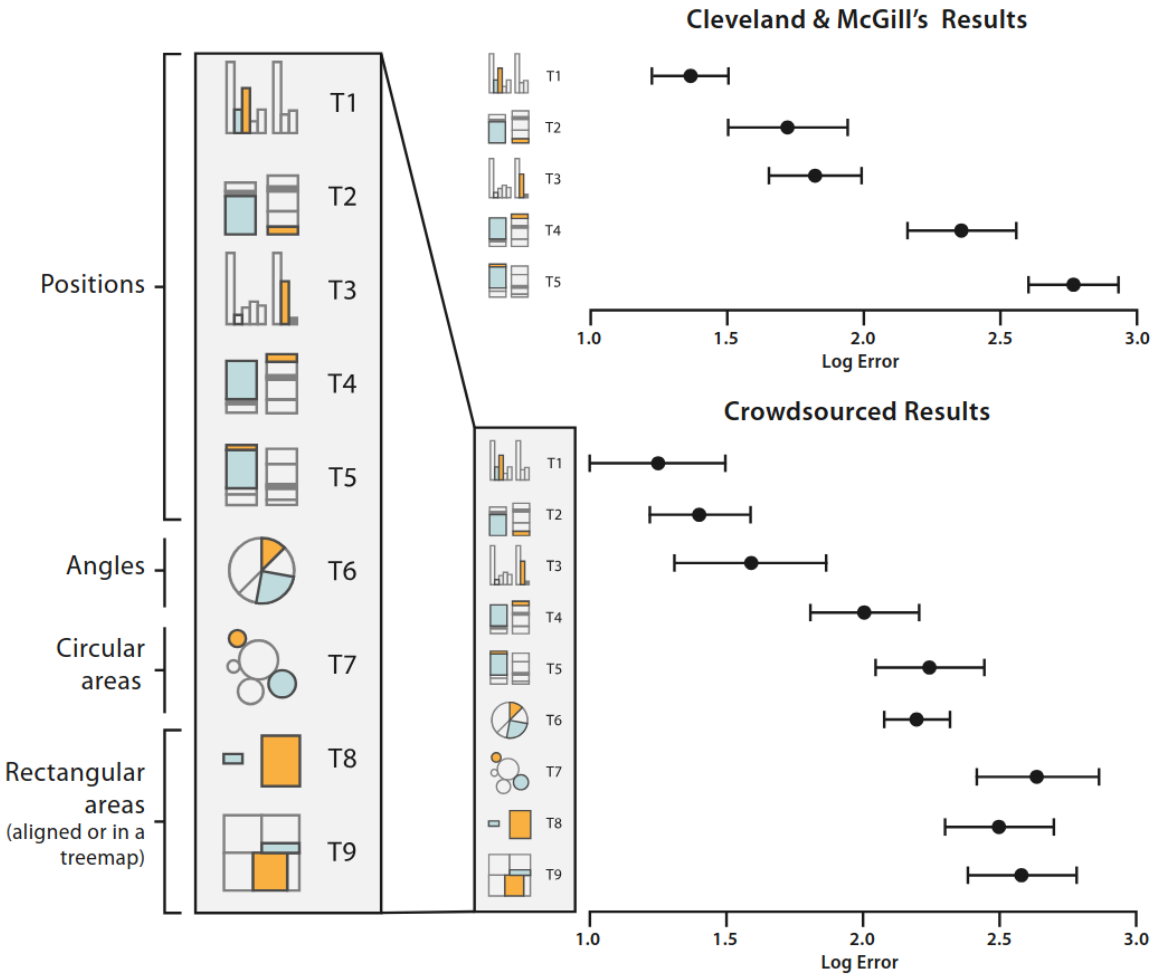
Data can be encoded through a variety of visual channels. However, not all channels are equally effective for conveying information.



—

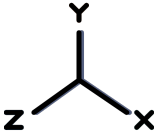
Credits: T. Munzner, "Visualization Analysis and Design", 2014.

Cleveland and McGill (1984) conducted experiments to evaluate the accuracy of visual channels for encoding quantitative data.



—

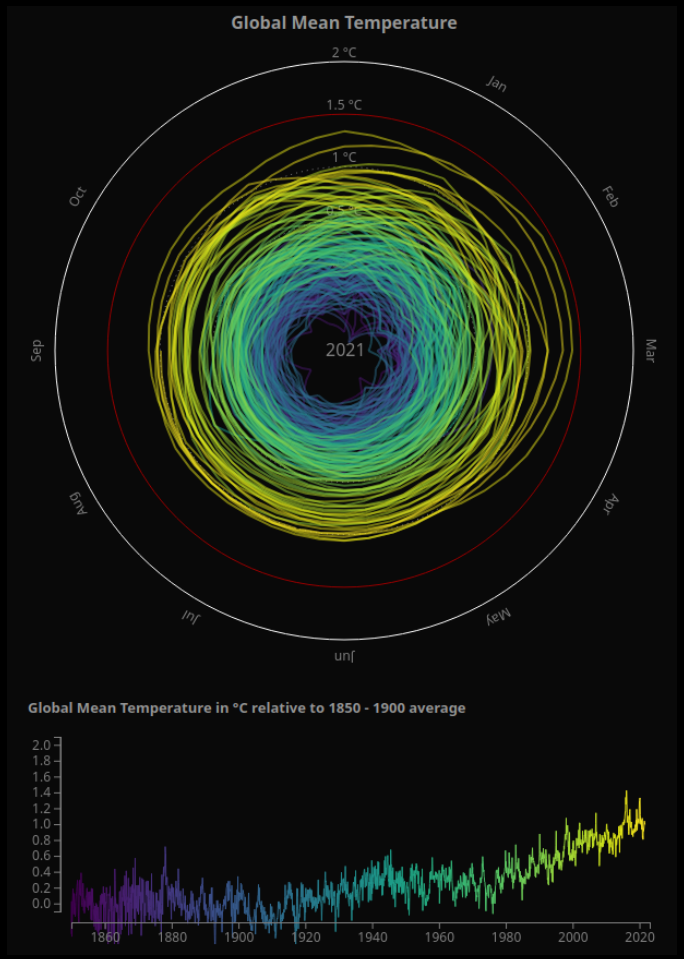
Credits: T. Munzner, "Visualization Analysis and Design", 2014.



Coordinate systems

Before mapping (numerical) data to visual channels, it is important to choose an appropriate coordinate system as it affects the perception and effectiveness of the visualization.

- **Cartesian coordinates**: intuitive and effective for most data types.
- **Polar coordinates**: useful for cyclic data (e.g., time of day, seasons), but can distort perception of lengths.



Climate spiral vs. line chart showing global mean temperature change over time.

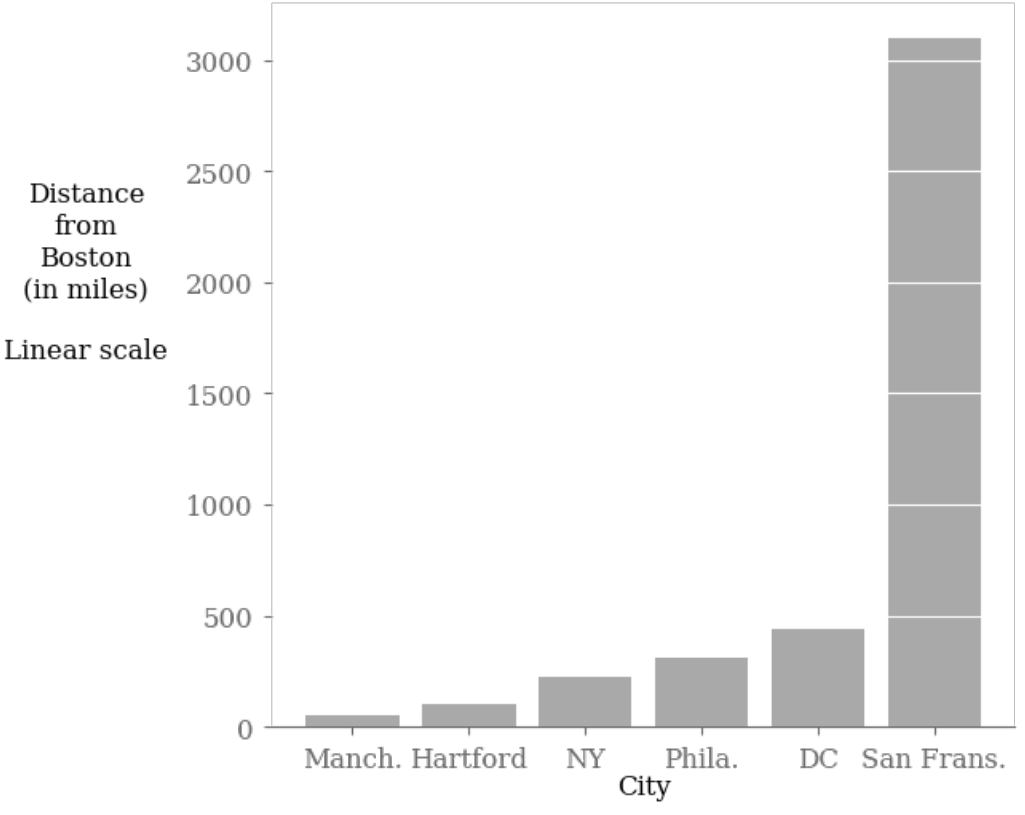
—

Credits: Open climate data, adapted from Ed Hawkins' climate spiral.

Scales and transformations

Sometimes, data spans several orders of magnitude or has a skewed distribution. In such cases, applying a transformation or using a different scale can improve the interpretability of the visualization.

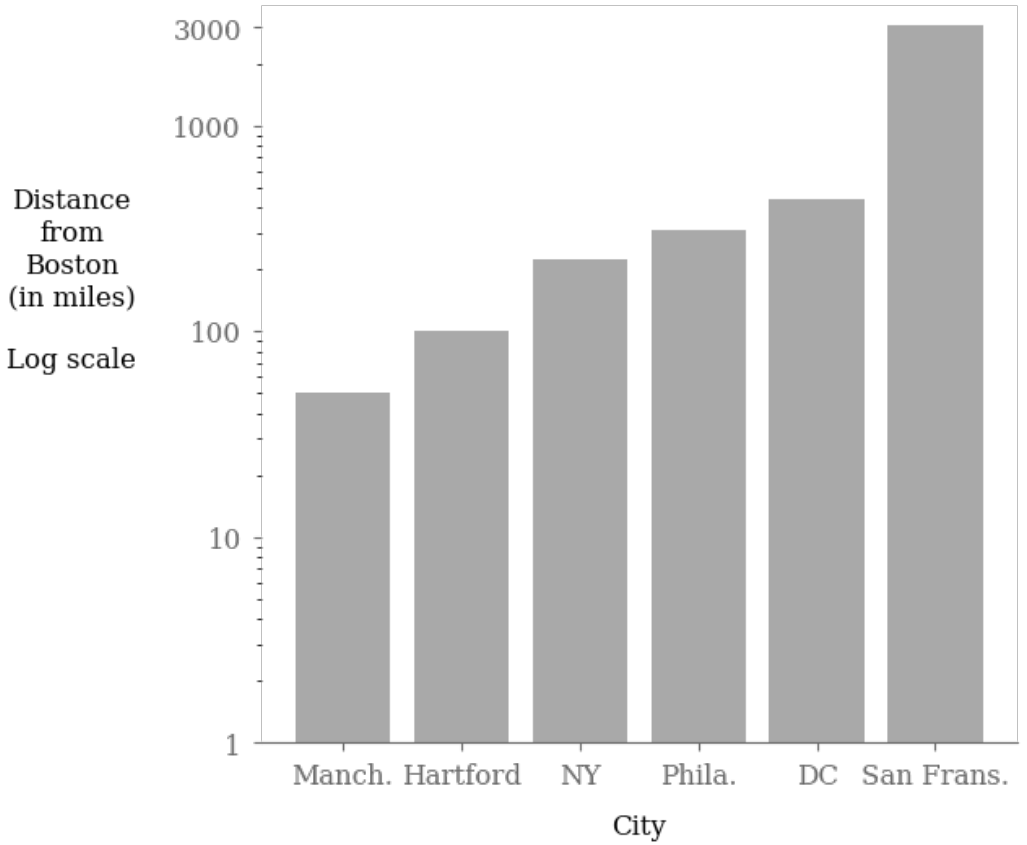
- **Linear scale**: preserves the original data values.
- **Logarithmic scale**: useful for data spanning several orders of magnitude.
- **Quantile scale**: divides data into equal-sized bins, useful for skewed distributions.



Linear scales can be dominated by large values, dwarfing smaller values and making it hard to see small variations.

—

Data source: Effective data visualization, USML, 2021.



Showing data on a logarithmic scale can prevent large values from dominating the visualization and reveal patterns among smaller values.

—

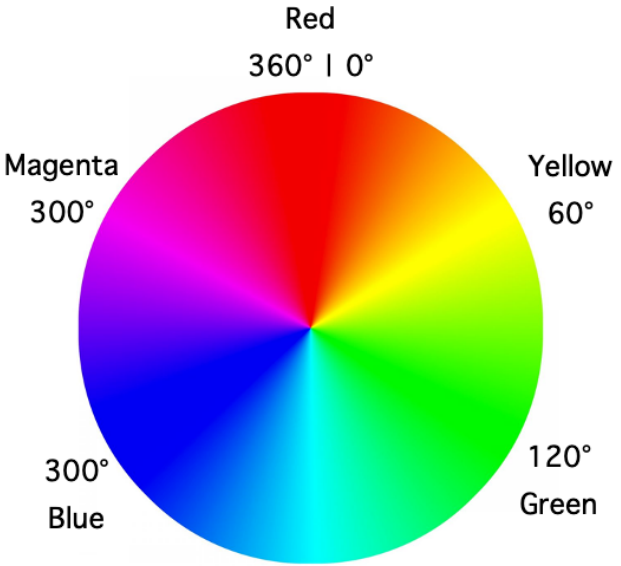
Data source: Effective data visualization, USML, 2021.

Colors

Color is a powerful channel for encoding categorical and quantitative data.

The primary representation system is the Hue, Saturation, Value (HSV) model:

- **Hue**: the type of color (e.g., red, green, blue), numerically represented as an angle on the color wheel (0-360 degrees).
- **Saturation**: the intensity or purity of the color (from gray to full color), represented as a percentage (0-100%).
- **Value**: the brightness of the color (from black to full brightness), represented as a percentage (0-100%).

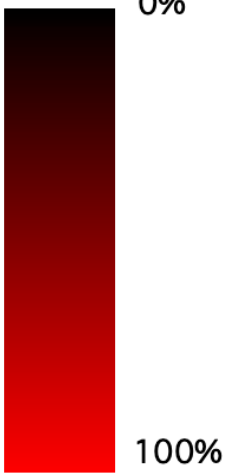


Hue

(red, green, etc.)

Saturation

(essence of the hue)



Lightness

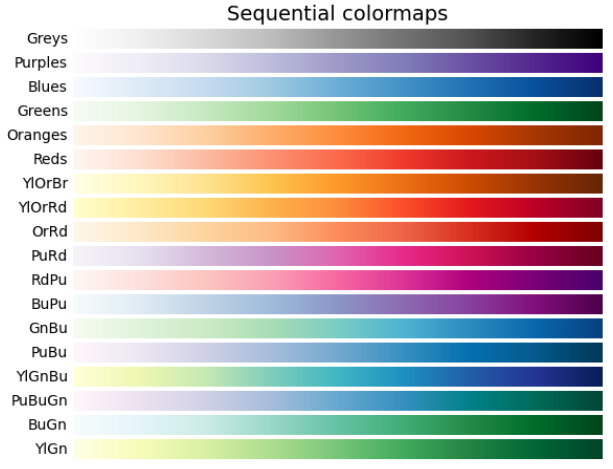
(darkness/brightness)

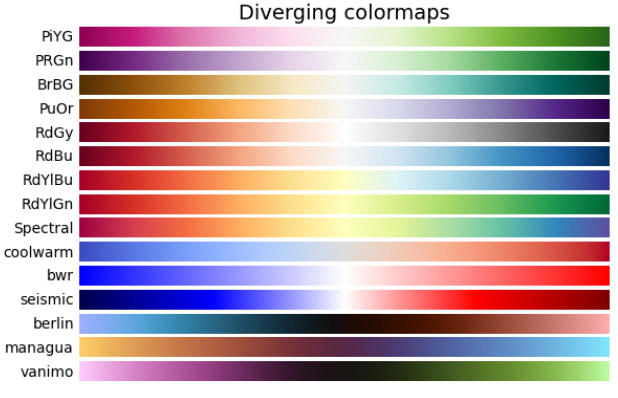
—

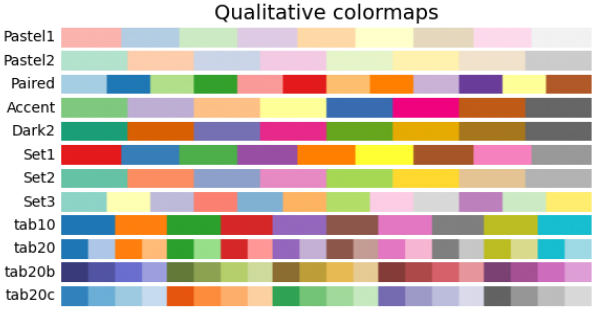
Data source: Effective data visualization, USML, 2021.

A colormap specifies a mapping between data values and colors. There are three main types of colormaps:

- **Sequential colormaps**: vary smoothly from light to dark colors, often using a single hue; should be used for representing ordered data.
- **Diverging colormaps**: vary smoothly between two different hues, with a neutral color in the middle; should be used for representing ordered data with a critical midpoint.
- **Categorical colormaps**: consist of distinct colors; should be used for representing categorical data without inherent ordering.


Sequential colormaps
Greys
Purples
Blues
Greens
Oranges
Reds
YlOrBr
YlOrRd
OrRd
PuRd
RdPu
BuPu
GnBu
PuBu
YlGnBu
PuBuGn
BuGn
YlGn

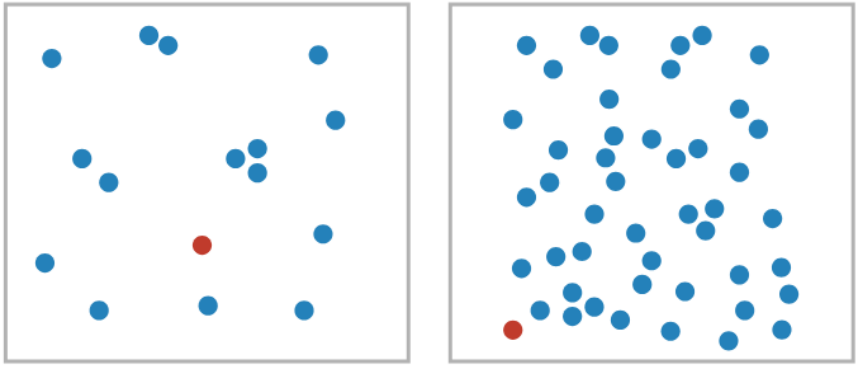

Diverging colormaps
PiYG
PRGn
BrBG
PuOr
RdGy
RdBu
RdYlBu
RdYlGn
Spectral
coolwarm
bwr
seismic
berlin
managua
vanimo


Qualitative colormaps
Pastel1
Pastel2
Paired
Accent
Dark2
Set1
Set2
Set3
tab10
tab20
tab20b
tab20c

Color vision deficiency affects approximately **1 in 12 men** and **1 in 200 women** worldwide. **Color is not a reliable channel for encoding information**.

Perceptually uniform colormaps ensure that equal steps in data are perceived as equal steps in color.

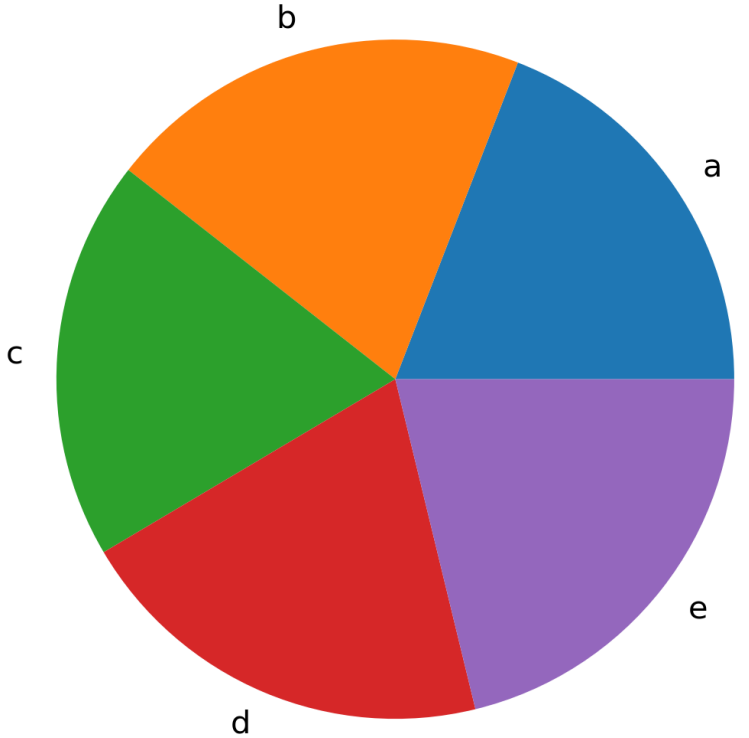
Finally, colors can also be used to **draw attention** to specific elements in a visualization, such as highlighting important data points or trends.



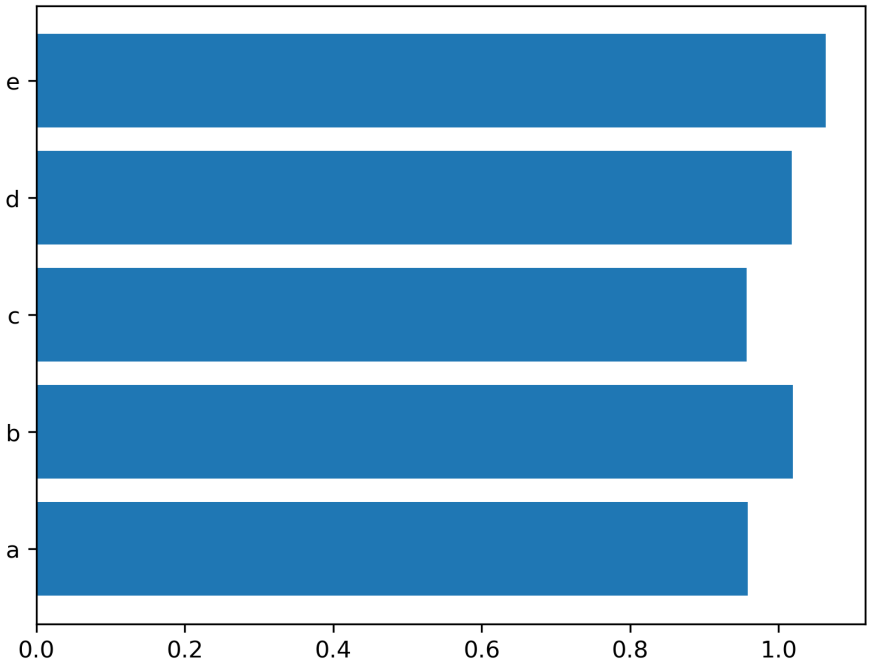
—

Credits: T. Munzner, "Visualization Analysis and Design", 2014.

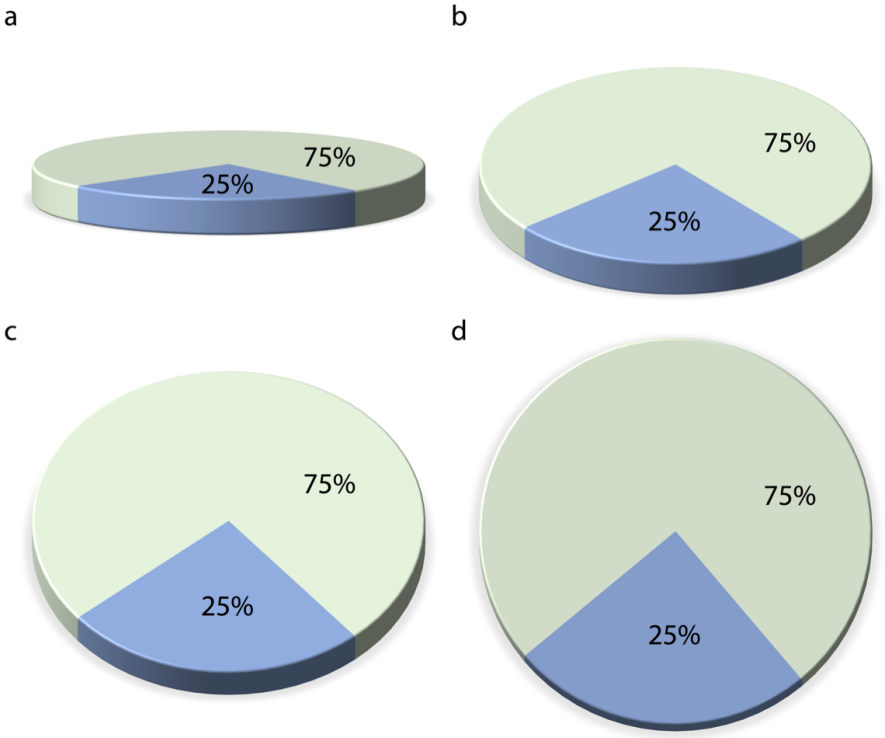
Anti-patterns



What category is the largest?



Same data! **Lengths and positions are easier to compare** than angles and areas.



Do not go for 3D. It distorts perception and adds unnecessary complexity.

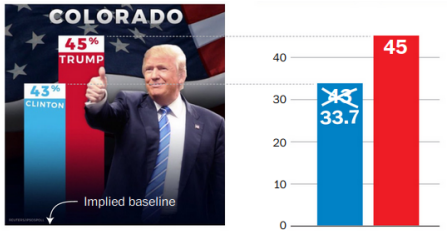
—

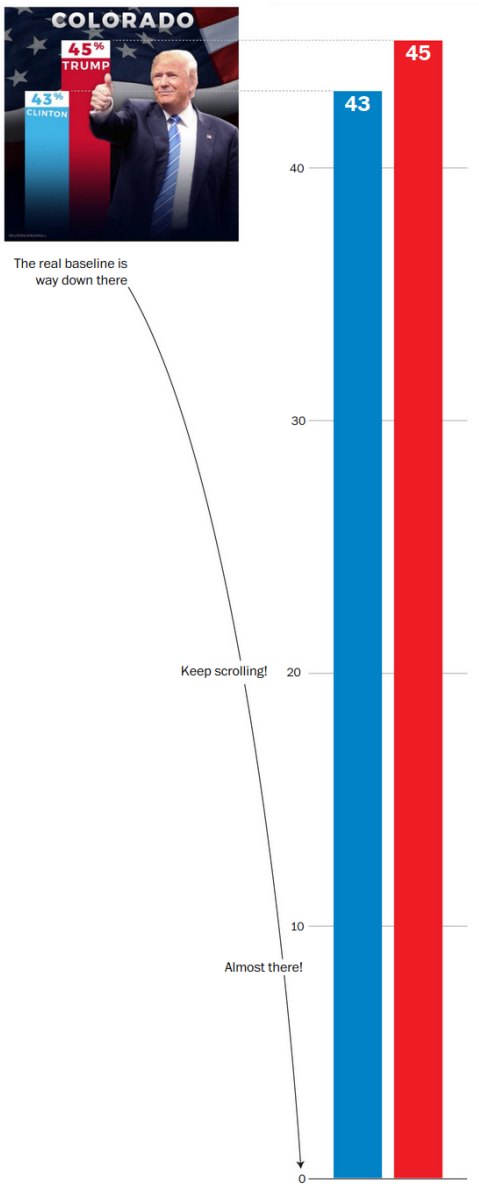
Credits: Claus O. Wilke, Fundamentals of Data Visualization, 2019.

Do not exaggerate reality by truncating axes or using misleading aspect ratios.

—

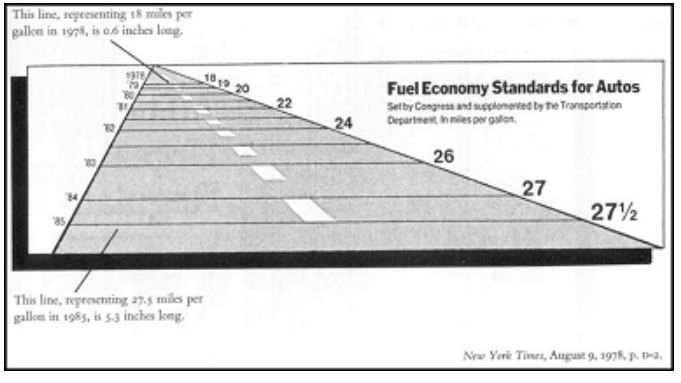
Credits: John Muyskens, Most of Trump's charts skew the data, The Washington Post, 2016.

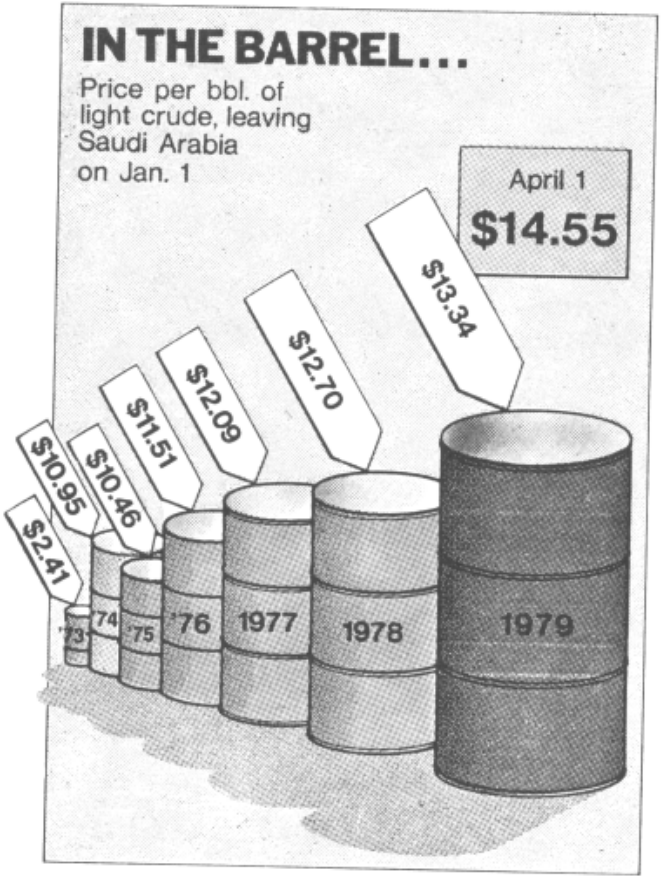




—

Credits: John Muyskens, Most of Trump's charts skew the data, The Washington Post, 2016.



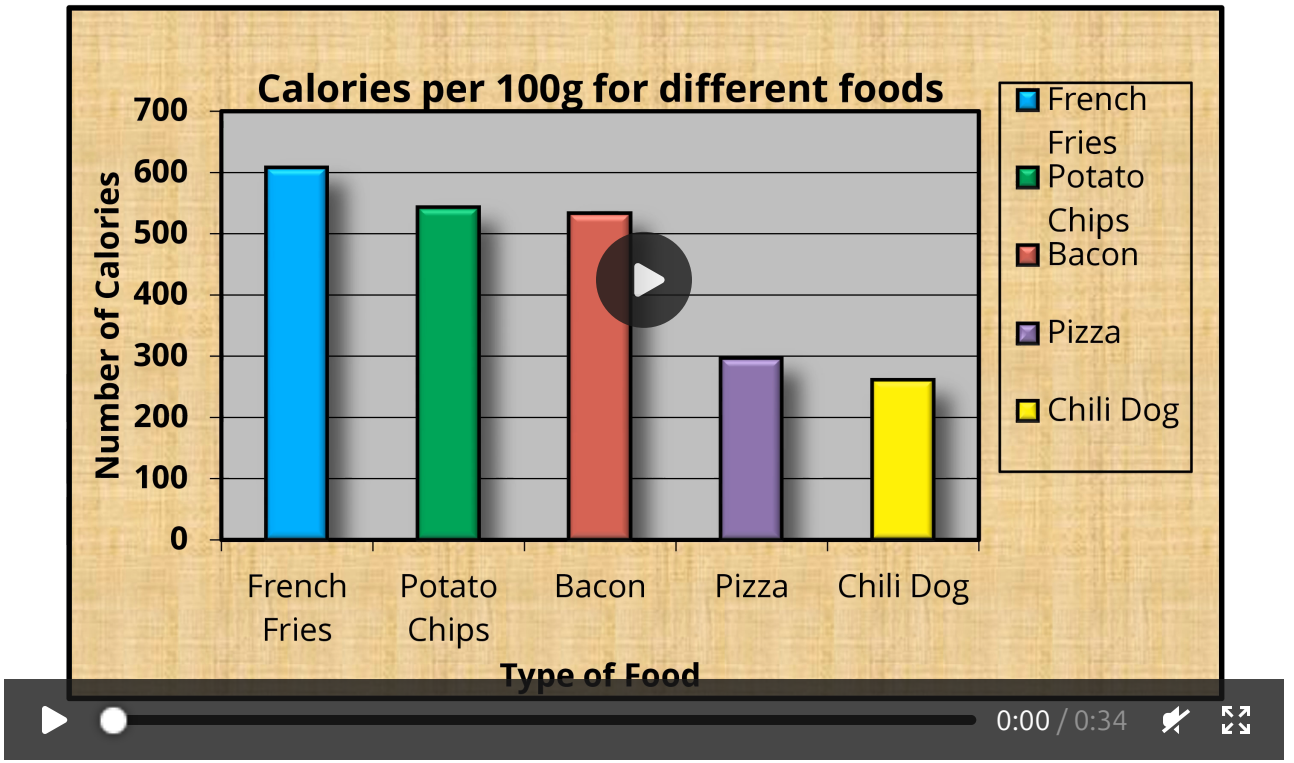


Do not lie. Use visual channels that accurately represent the data.

—

Credits: David Giard, Data visualization.

Remove backgrounds



Maximize the data-ink ratio by removing unnecessary elements.

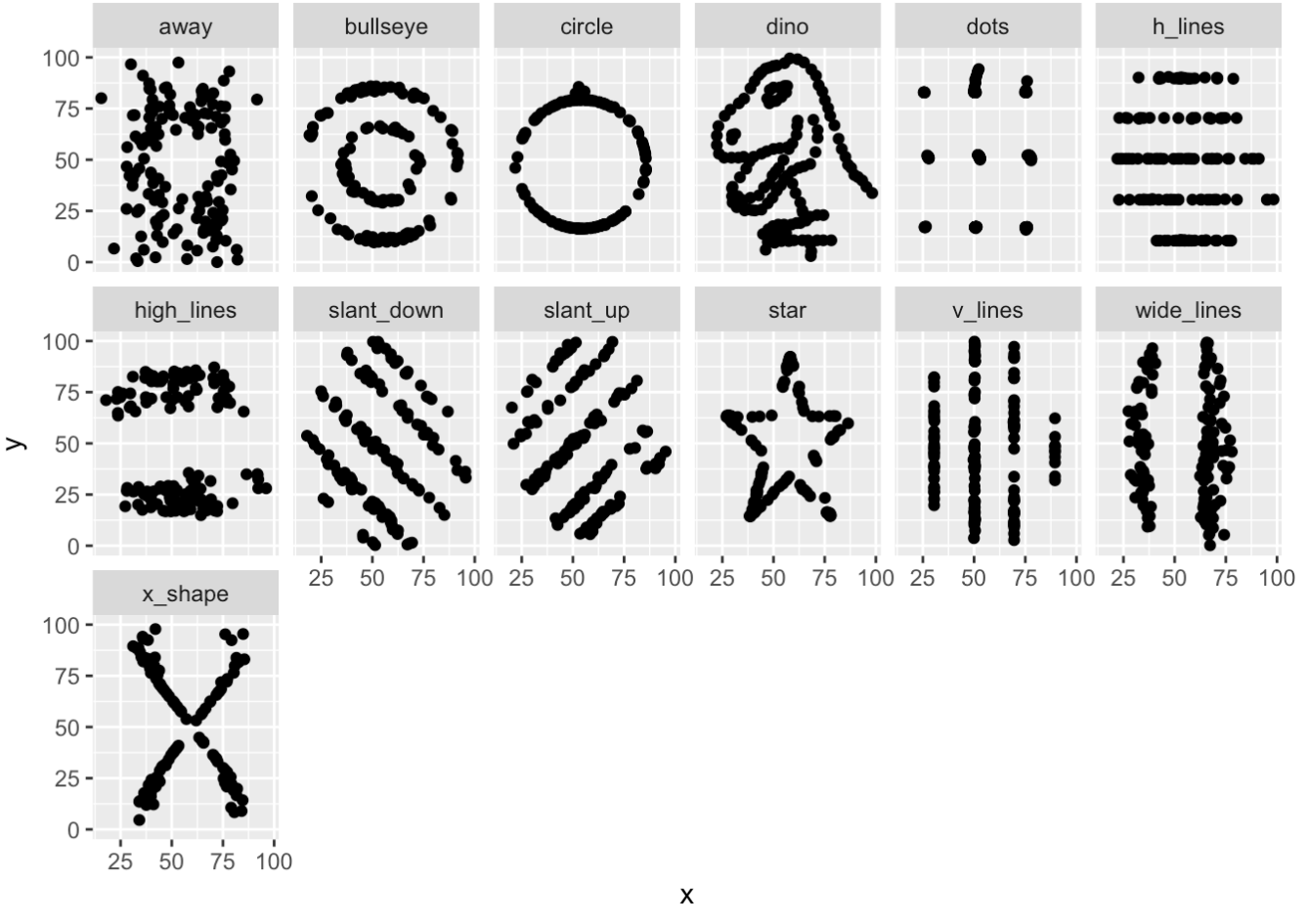
—

Credits: Joey Cherdarchuk, Data looks better naked.

dataset	mean_x	mean_y	sd_x	sd_y	cor
away	54.26610	47.83472	16.76982	26.93974	-0.06412835
bullseye	54.26873	47.83082	16.76924	26.93573	-0.06858639
circle	54.26732	47.83772	16.76001	26.93004	-0.06834336
dino	54.26327	47.83225	16.76514	26.93540	-0.06447185
dots	54.26030	47.83983	16.76774	26.93019	-0.06034144
h_lines	54.26144	47.83025	16.76590	26.93988	-0.06171484
high_lines	54.26881	47.83545	16.76670	26.94000	-0.06850422
slant_down	54.26785	47.83590	16.76676	26.93610	-0.06897974
slant_up	54.26588	47.83150	16.76885	26.93861	-0.06860921
star	54.26734	47.83955	16.76896	26.93027	-0.06296110
v_lines	54.26993	47.83699	16.76996	26.93768	-0.06944557
wide_lines	54.26692	47.83160	16.77000	26.93790	-0.06657523
x_shape	54.26015	47.83972	16.76996	26.93000	-0.06558334

Do not summarize the data without visualizing it.

The Datasaurus dozen: 13 datasets with identical summary statistic.



Summary statistics can be misleading. **Always visualize the rawest data**!

Choosing the right plot

Jean-Luc Doumont, "Choosing the right graph", 2017.

Key takeaways

- Visualize data using the most perceptually effective channels (positions, lengths)
- Use appropriate coordinate systems and scales to enhance interpretability.
- Choose colormaps that accurately represent the data and are accessible to all viewers.
- Avoid common pitfalls such as 3D effects, misleading axes, and unnecessary complexity.

Wrap-up exercise

Let us discuss the following examples. For each of them, identify what is good and what could be improved.

(All plots are taken from MSc theses of previous years. Author names have been removed to protect the innocent.)

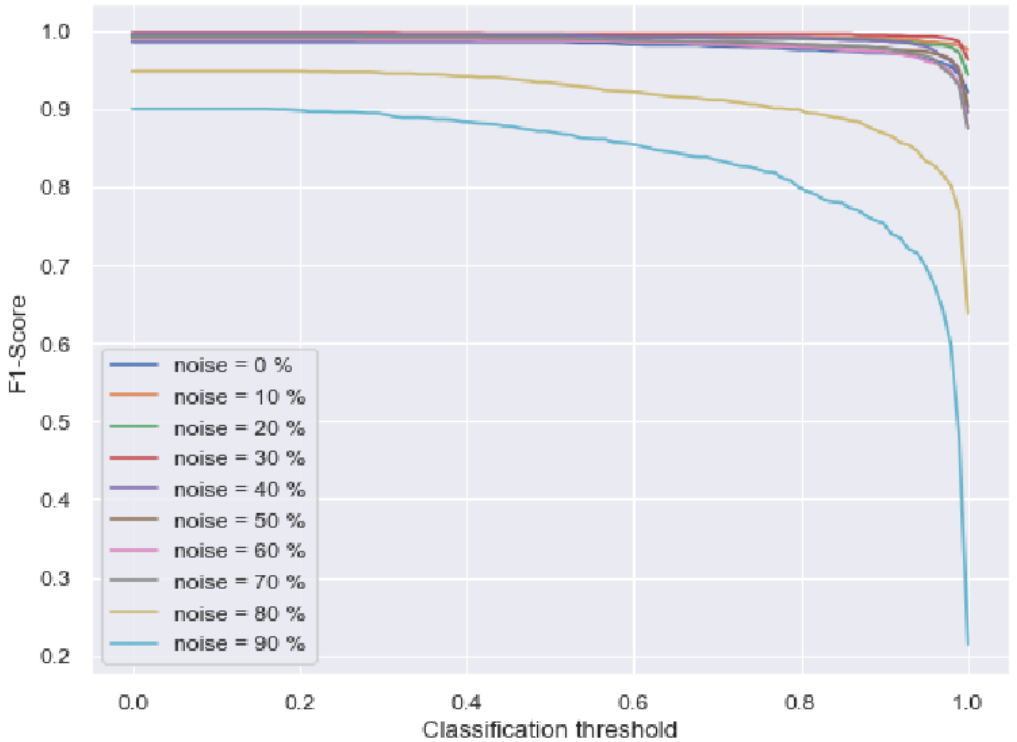


FIGURE 4.21: F1-Score of the best model trained with different quantities of noise going from 0 to 90% and tested on the real conditions dataset according to a classification threshold.

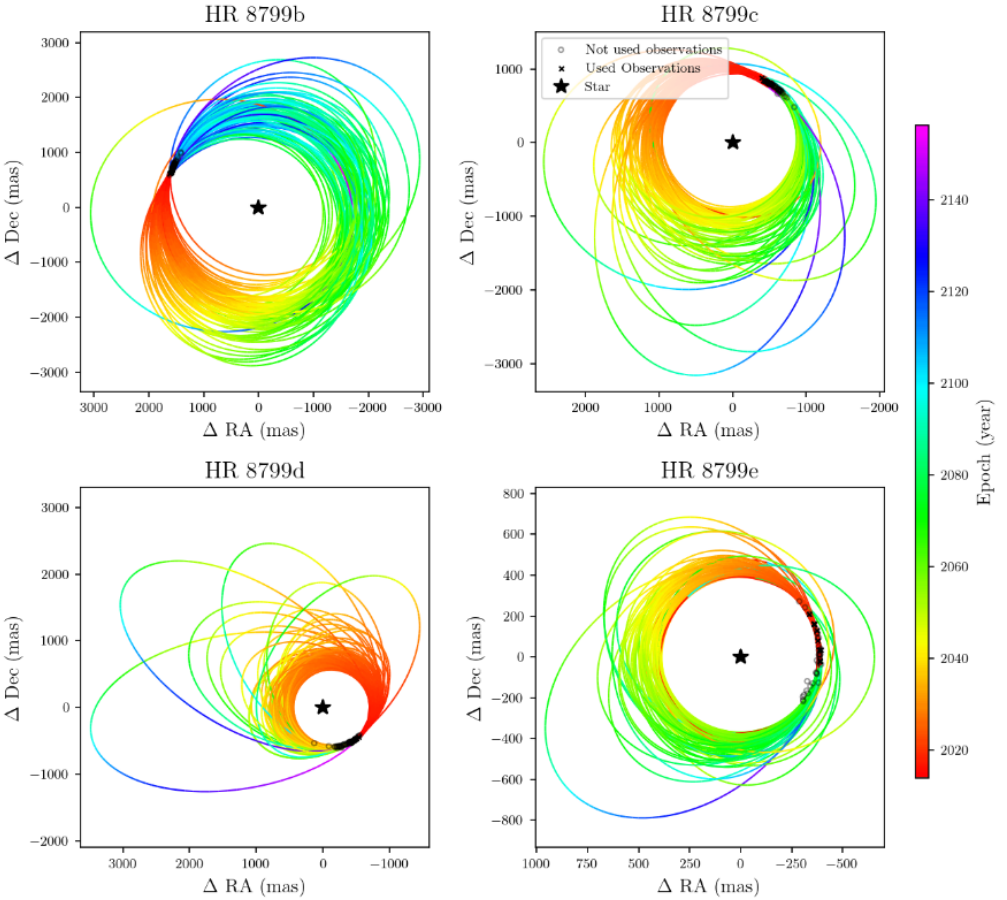
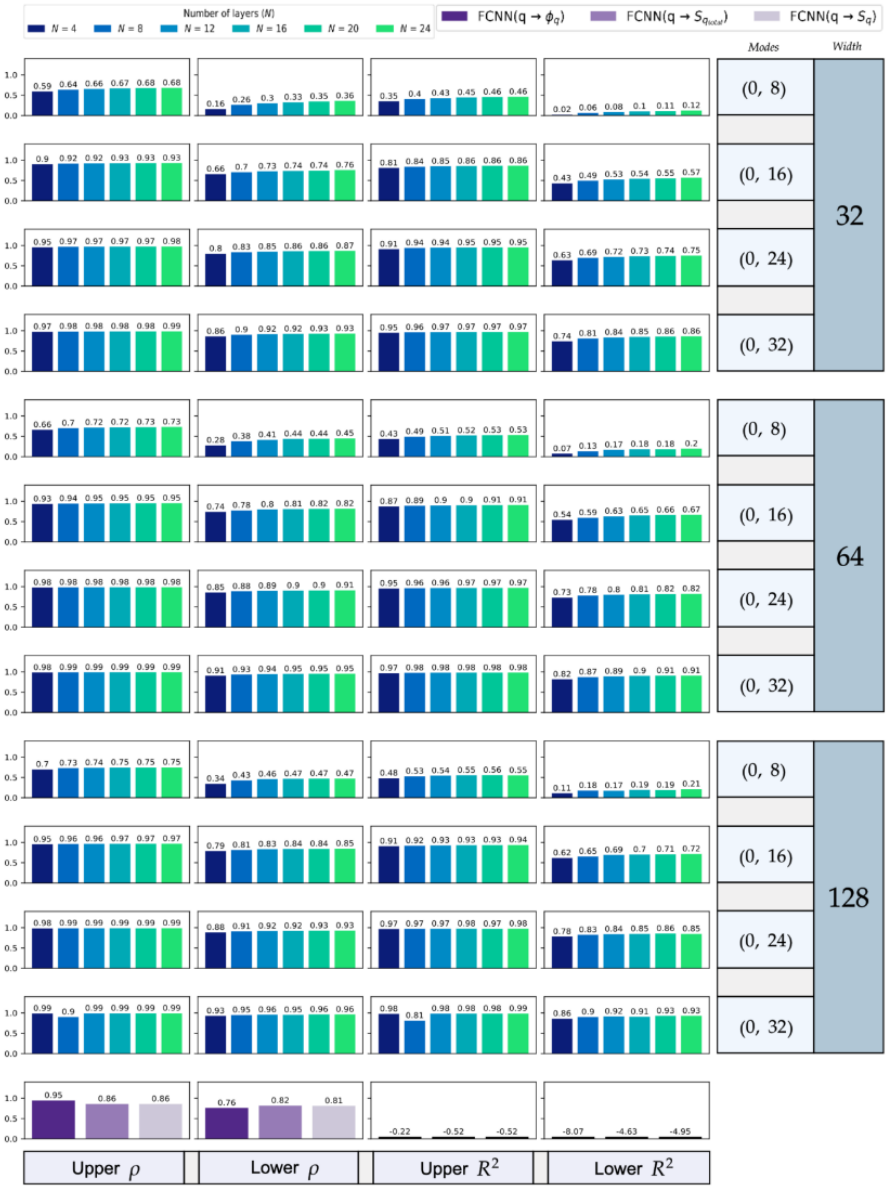


Figure 5.12. Posterior predictive check of the 4 exoplanets of HR 8799, 1000 orbits are generated for each exoplanet. The NPE produces a good approximation of those orbits, however, some impossible orbits are still generated like on HR8799b, HR8799c or HR8799e where there are orbits that are not passing through the observations. The x-axis has been inverted to make the plot comparable with the one in the paper of Sepulveda and Bowler (2022) [40].


Number of layers (N)
N = 4
N = 8
N = 12
N = 16
N = 20
N = 24
FCNN(q → ϕ_q)
FCNN(q → $S_{q_{total}}$)
FCNN(q → S_q)
Modes
Width
(0, 8)
(0, 16)
(0, 24)
(0, 32)
32
64
128
Upper ρ
Lower ρ
Upper R^2
Lower R^2

The bottom line

A good plot is one that effectively communicates the underlying data and insights to the audience. It should be clear, accurate, and visually appealing, while avoiding unnecessary complexity or distortion of the data.

The end.