a short

DATA VISUALIZATION HANDBOOK

by IRAKLIS VRETZAKIS

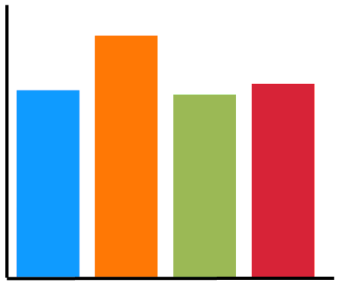
Outline

Data visualization is an essential component of scientific research and communication. Graphs and charts are employed to visually represent numerical information in the physical space, both for exploring your dataset during the analysis phase, but also for communicating your results in the publication and outreach phase. Picking up the right graphical format for visualizing your dataset might prove a more cumbersome task than predicted, especially for an inexperienced eye. Multiple configurations might be appropriate to plot your dataset, and thus there is no such thing as the perfect recipe. Yet, there are certain suggestions that will aid you bring your scientific message forward and create more impactful visualizations. Those suggestions are briefly contextualized herein the above chapters.

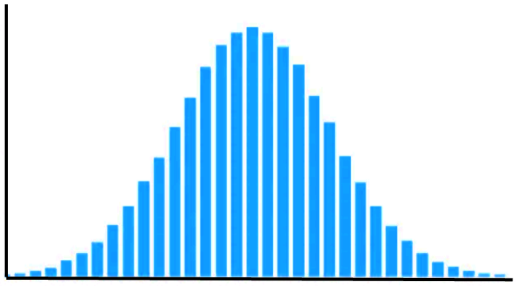
.

Graphical forms of data charts

There is an abundant variety of graphs that you might select from to plot your datasets. Either simple or more complex, this selection suggestively depends on:
a) the **nature** of your **variable(s)**, for example:



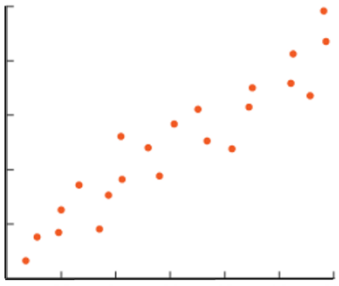
bar chart
for categorical variables



histogram
for continuous variables

b) the **number** of your **variables**, for example [1]:

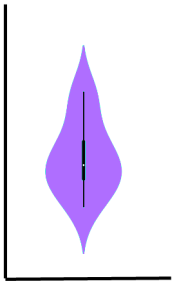
pie chart
for univariate plotting



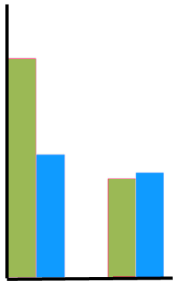
scatterplot
for bivariate plotting

Chernoff faces
for multivariate plotting

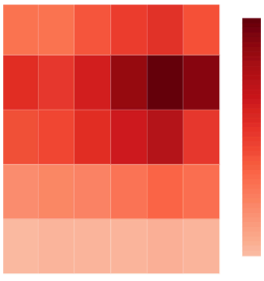
c) the **type of relationship** that you wish to illustrate among your data, for example:



violin chart
to illustrate
distribution



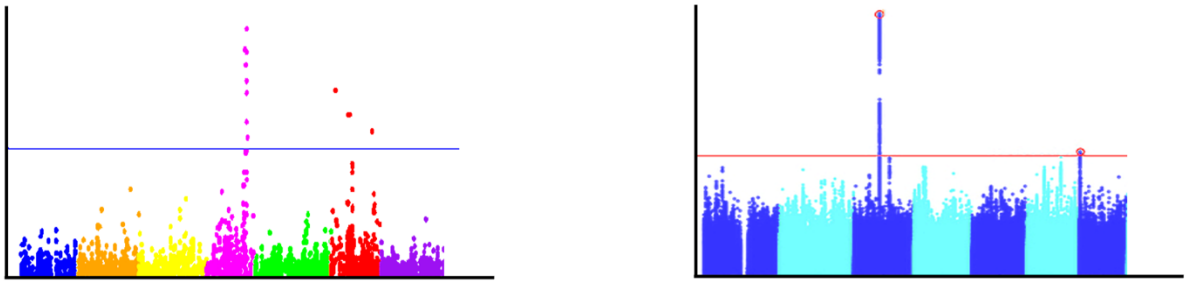
bar chart
to illustrate
comparison



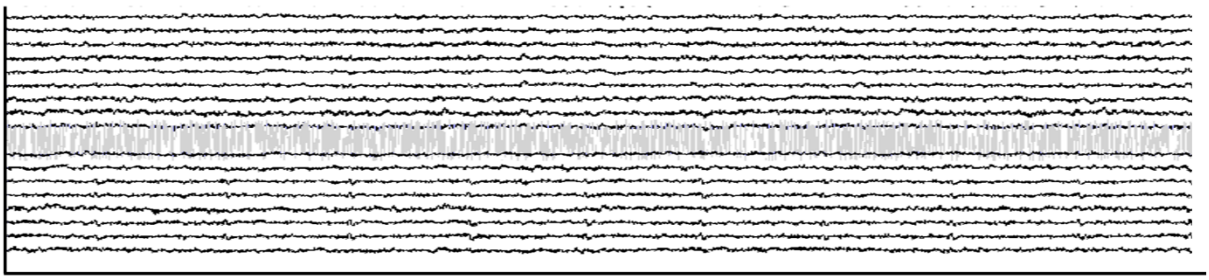
heatmap
to illustrate
correlation

sunburst diagram
to illustrate
parts-to-a-whole

Furthermore – depending on your **research (sub)discipline** – you will get acquainted with more complex graphical formats, which in essence can be reduced to more simple ones. For example:



Manhattan plots in genome-wide association studies are very dense scatterplots. The x axis displays the chromosomal position of a given SNP, while the y axis the $\log_{10}$(P-value) of the association statistic [2].



MEG magnometric graphs in neuroimaging studies are multilayered line graphs. Each waveform displays the magnetic field generated by neuronal activity, as measured in individual channels [3].

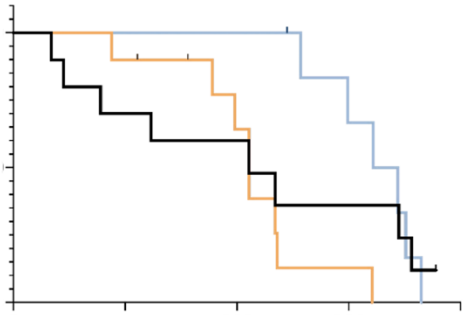
It is hence apparent that picking up the right graphical form requires in-depth comprehension of your dataset and the conceptual relationships that can be inferred among your data. This selection should always be in line with the story you wish to illustrate on a given audience, as well as the established standards of your research community. Having that done, it is now time to proceed with additional graphical elements that are essential for accurately visualizing scientific data.

Navigational cues

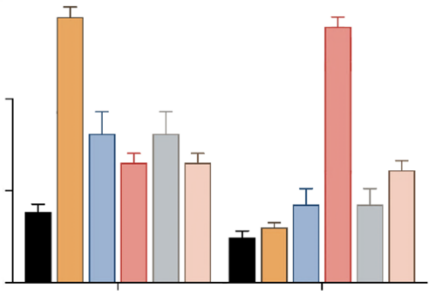
Axes, units and gridlines are the primary navigational cues of most charts and plots. **Axes** constitute the visual representation of scales and thus, they are essential for bringing your dataset into the physical space [4]. A typical axis with ticks resembles a ladder, providing a reference to guide the reader's eyes throughout your graphical display. **Grids** are used to establish clear comparison of proportions or relative positions among the different data at display [5]. Those together ultimately determine how a reader perceives the size, shape and location of your data points or classes.

At a first glance, designing the appropriate axes and gridlines seem to be simply a matter of selecting the range within which the numbers/intervals of your datasets fall. However, depending on the nature of the variables and the range of the observed values, there are certain additional things that you may consider:

- *Opt for axes that capture the entire range of your data values*, i.e. extend the scales beyond the observed minimum and maximum of the variables. Data points that fall right on the boundaries will be obscured by the axes.

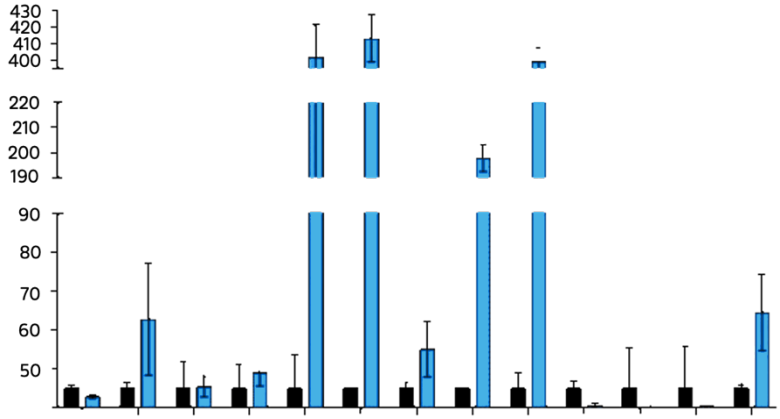


a good example

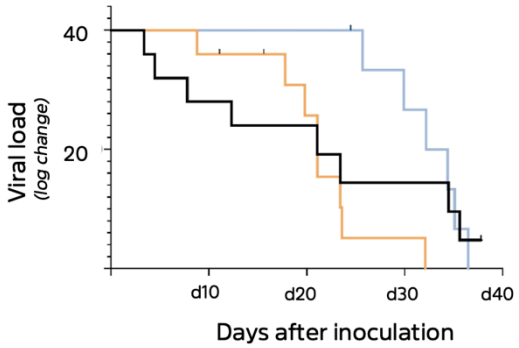


not that good of an example

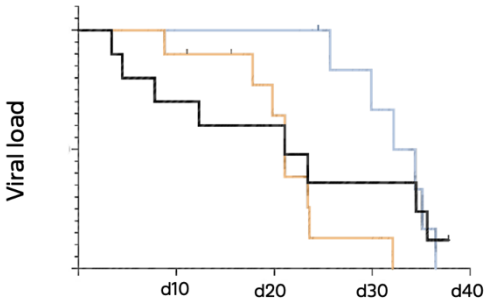
- In general, *tick intervals should be proportionally scaled*. But when your data range includes extreme values like below, consider using **interval breaks** (or transforming to log scale depending on the data). Make sure that *the breaks extend all the way and include the axes* – not only in the data you plot.



- The axes and ticks should be *appropriately labelled*; *units of measurement* should be present to avoid ambiguity.



a good example

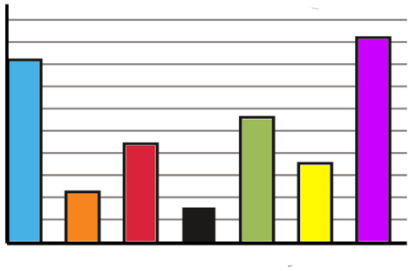


not that good of an example

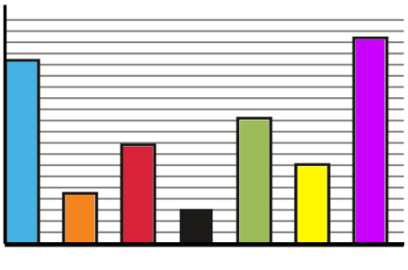
- Make sure your *tick numbers always read horizontally*, (while the y-axis title/label can be read vertically); to facilitate readability, label the axes' ticks using round numbers when applicable.

- Consider the **origin** (i.e. where the two axes meet) based on conventions or expectations [6]. In other words, not all axes should start with a zero (e.g. the origin of a double log scale is 1).

- When including gridlines, *the density of the grid is suggestive of the scale at which variation is observed*. However, dense grids might create redundant visual clutter and factually make it harder for the reader to trace them back to the axes ticks.



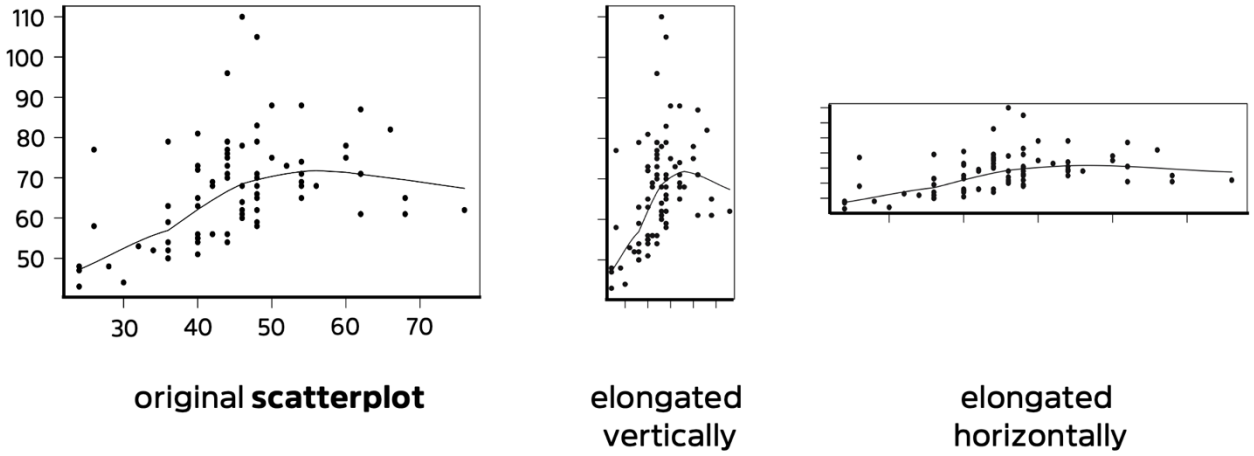
a good example



not that good of an example

- In multi-panel figures, *use the same axes range across graphs* when applicable, to help the reader compare them faster. That way you may avoid unnecessary repetition of axes labels/units.

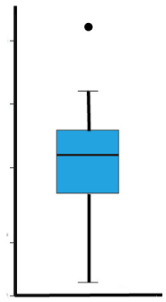
- In multi-panel figures, be also mindful of *the aspect ratio of your graphs*. **Aspect ratio** (i.e. the ratio of the width to the height of a given display) does affect the way we perceive the relationship illustrated in a graph. An elongated x axis with a shrunk y axis will highlight gradual change (e.g. over time), while the reverse will suggest more dramatic change [7]. For example:



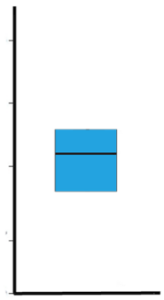
Graphical overlays of statistical information

Statistics are a necessary evil when visualizing research data. When submitting graphs in scientific journals for example, those most often require elaborate statistical information (e.g. measures of variance, regression coefficients, p-values and many more). This is also the case – even though less strictly – when presenting in poster sessions and research symposia. Below are some examples of how statistics are commonly visualized or annotated on basic graphical shapes. Regardless of your analyses, it is absolutely suggested that you include this information in your visualizations.

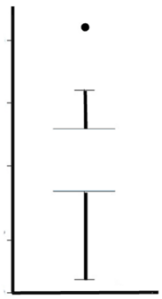
By design, certain charts already contain statistical information graphically. In a box-whisker chart for example, the box represents the interquartile range (Q1-Q3) with the median value as the horizontal line. The whiskers represent the minimum to the lower quartile and then from the upper quartile (the end of the box) to the maximum, with a dot that represents an outlier [8].



box-whisker chart

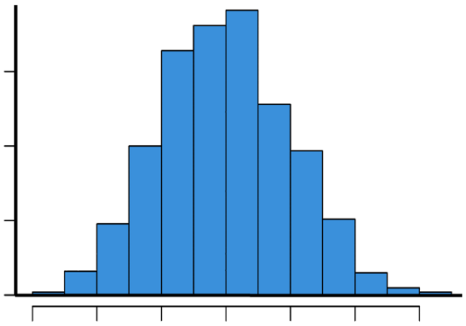


interquartile range + median

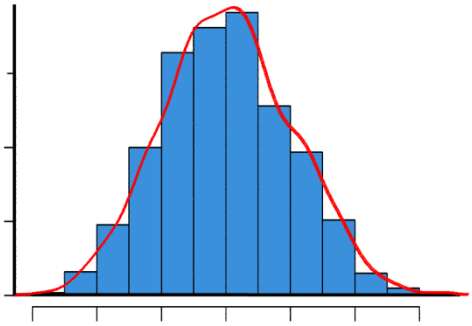


upper and lower limit + outlier

In other cases, additional shapes need to be visualized on top of a chart to show the hypothesized overall structure or model of your dataset [7]. Those options are part of most statistical software scripts. For example, fitting a kernel density plot over a histogram to visualize the shape/structure of a distribution:

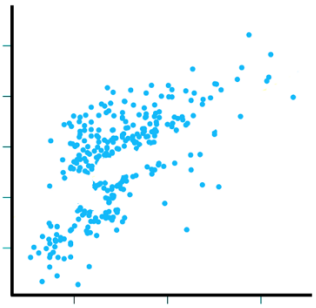


original **histogram**

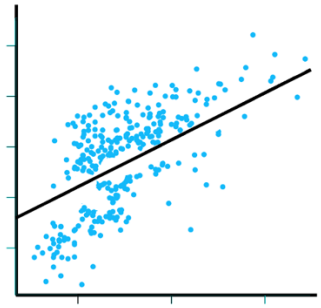


histogram + density plot

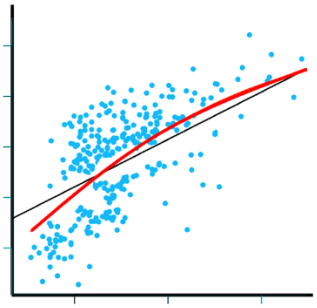
Or adding regression lines to show the comparison of fitness for different models:



original
scatterplot

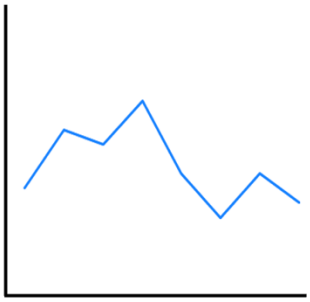


linear regression
fit

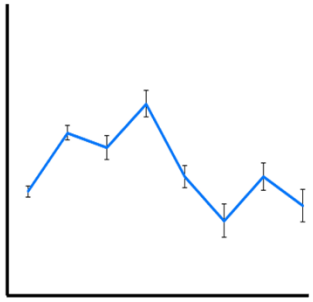


polynomial regression
fit over linear

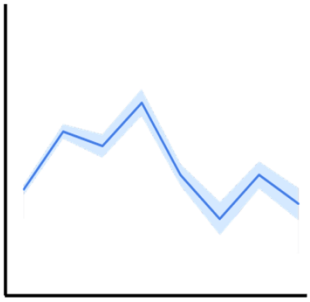
Or overlaying confidence intervals (or confidence bands) to visualize uncertainty of an estimate:



original
line graph

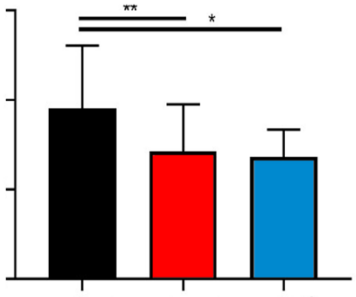


line graph with
confidence intervals

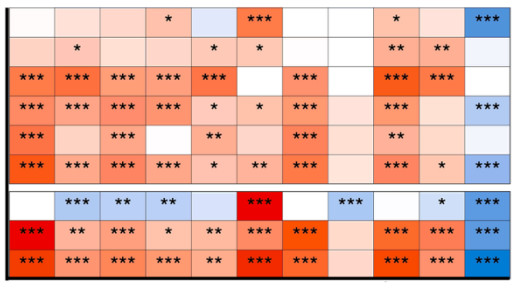


line graph with
confidence bands

Nonetheless, certain output parameters may not be visualized graphically and they need to be annotated on top of your graphs (either by using symbols, numbers or equations). For example, the most common convention is using asterisks to point out statistical significance. In APA style for example, one asterisk (*) refers to $P \leq 0.05$, two asterisks (**) to $P \leq 0.01$ and three (***) to $P \leq 0.001$:



annotating p-value
over a **bar chart**

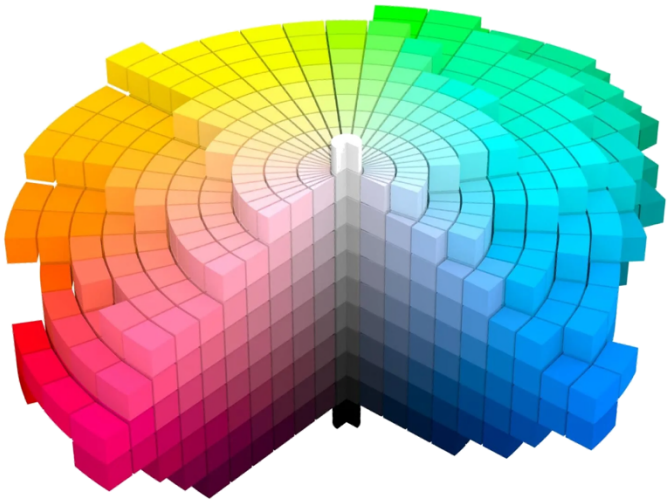


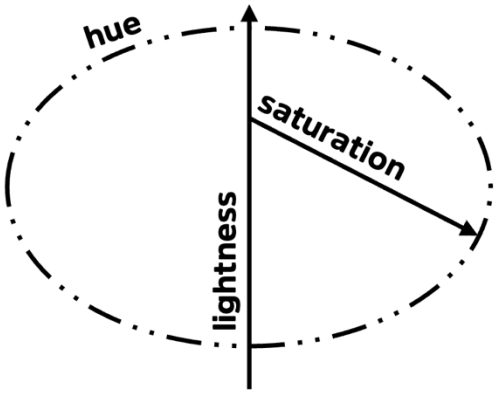
annotating p-value
over a **heatmap**

Color theory in graphs

Although often neglected, color is an equally important feature of scientific visualizations and their perception. When appropriately manipulated, color might draw the reader's attention to specific values or locations in your graphs and thus better facilitate legibility. Too many colors on the other hand may create a busy display with a lot of visual weight, rendering your output overwhelming and in fact difficult to read. It is therefore suggested that you *set up a minimal color palette, with limited and meaningful color choices to visualize your results*. There are several online tools that may help you with this selection, such as ColorBrewer, Coolors and Adobe – yet how should your selection be optimal?

To make color an ally when designing graphs, one should first consider its properties that are typically discussed in photography and visual arts. Depending on which editing software you use for your design manipulation, there are several models for color coding (e.g. RGB, CMYK, HSL and HCV). For example, the **HSL system** breaks down color into:



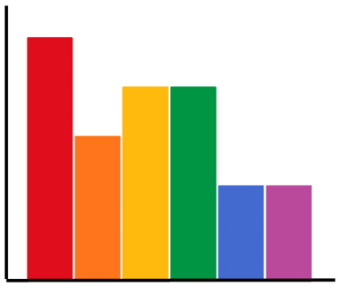


1) **Hue** (i.e. what we typically perceive as the color),
2) **Saturation** (i.e. the colorfulness, the vividness of the color) and
3) **Lightness** (i.e. how bright is the color shade).

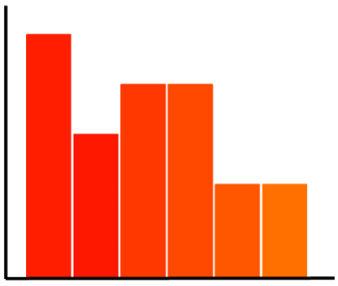
In this 3D space, every color is composed of three coordinates, with each one corresponding to one color property (H, S, L). For example, considering a vivid red as baseline and manipulating one property per color:

manipulating **hue**	manipulating **saturation**	manipulating **lightness**
(**0**, 100%, 50%)	(0, **100%**, 50%)	(0, 100%, **50%**)
(**235**, 100%, 50%)	(0, **50%**, 50%)	(0, 100%, **30%**)
(**280**, 100%, 50%)	(0, **30%**, 50%)	(0, 100%, **10%**)

These properties may appear theoretical, yet they define the way we perceive the graphical formations at sight. When it comes to plotting scientific results, *effective color manipulation will help highlight the relationship that you wish to convey about your data via your charts*. For example, if you wish to plot different groups on a bar chart, then picking a palette with different hues (rather than proximal hues) will create a more noticeable effect:

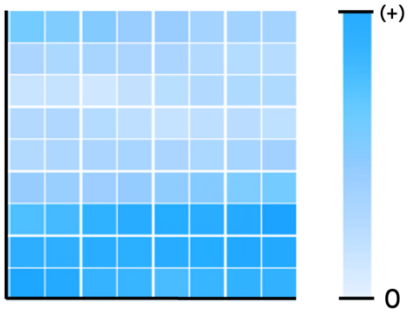


bar chart
with different hues

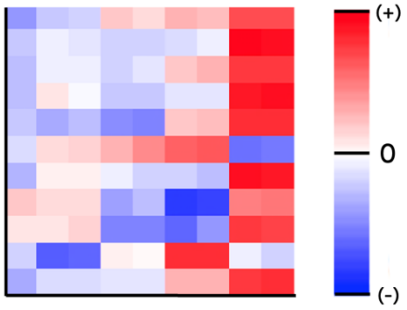


bar chart
with proximal hues

On the contrary, when you wish to highlight *gradual change*, then a palette with *succeeding proximal hues or lightness* would be a better choice. Heatmaps are the most representative example of such a case. For example, when effect is unidirectional (i.e. only absolute values matter), the effect size may be visualized with a simple gradient of lightness. But when effect is bidirectional (i.e. when negative values matter), using a lightness gradient anchored in two different hues is a more effective visualization:

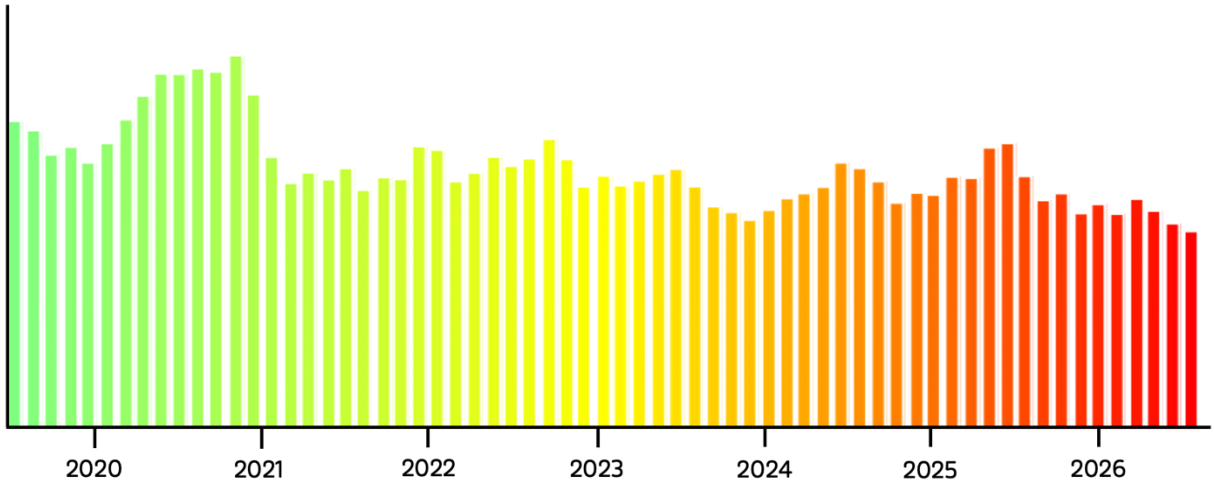


heatmap with
a lightness gradient



heatmap with
hue + lightness gradient

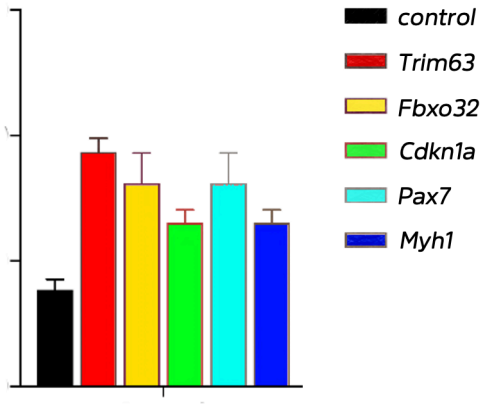
The use of such a palette is not limited to heatmaps; it may be generally used to suggest a *continuum*. For example, using a hue gradient on a dense bar chart might suggest time flow:



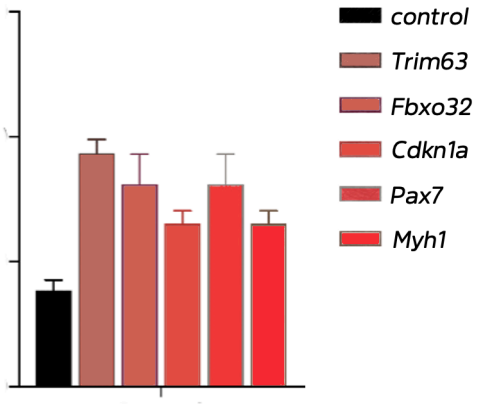
Your color palette is thus necessarily entangled with the relationship that you wish to illustrate about your dataset. This is especially true in populated datasets, where visualizations turn out even more complex. However, to ensure that all information are appropriately conveyed, there are certain things to take into account:

- All colors should be *annotated on the graph* (via a color-key) or *described in the legend* if a color-key is not applicable.

- The color palette should create *noticeable differences* that the reader can observe on the color-key.

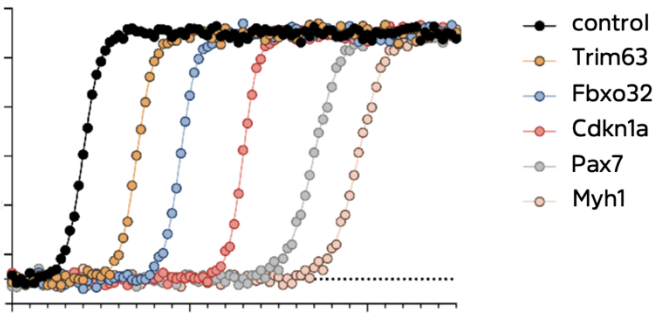


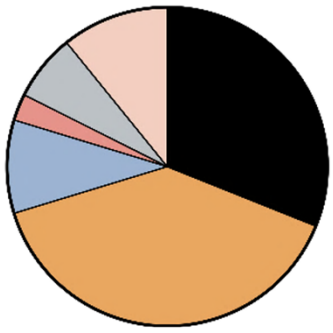
a good example



not that good of an example

- In multi-panel figures, if a given color signifies a specific concept (e.g. red for treatment group), then *this color should be consistent* across the different graphs.





- When using *saturation* of the same lightness to differentiate data, any *variations will go unobserved in the grayscale*. It is hence suggested that you refrain from this manipulation, in case a reader wants to print your graph(s) in black and white.

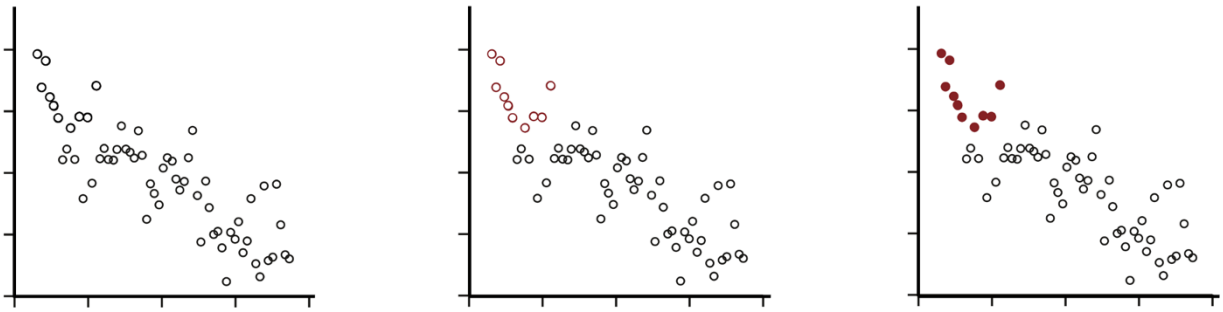
- *Color blindness* affects as many as 8% of men and 0.4% of women of Northern European ancestry, with vision deficiency in green and red being the most prominent. Online tools like Viz Palette may help you create a color palette that is inclusive of such an audience.

Additional design guidelines

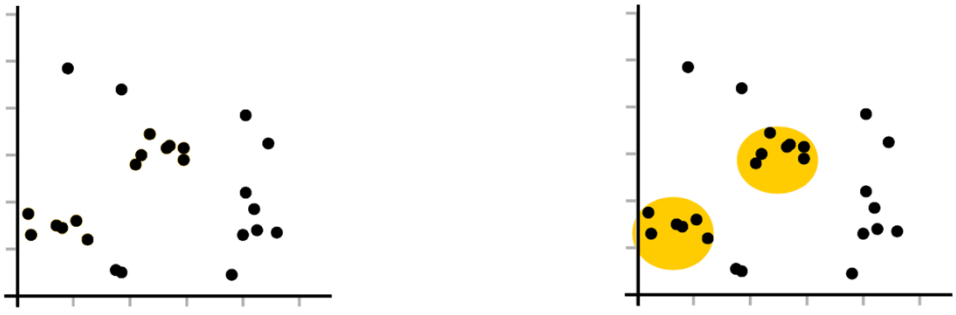
Thus far, the suggestions provided are mostly relevant for preparing graphs for scientific reports and manuscripts, somehow close to the final export of your software. It would be a remiss not to mention that there are *certain additional design guidelines* that will help you render your graphical message stronger. Those are suggested when designing panels for intentions of lectures, lab presentations, science communication pieces, poster presentations etc. To check if those apply to manuscript submissions, consult the guidelines of your journal of choice.

1. Salience

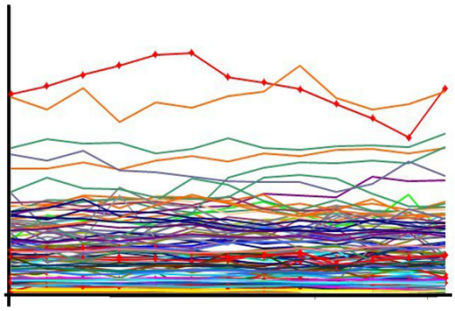
In design terms, **salience** refers to the physical property of an object that makes it stand out from its surroundings [9]. In other words, how an object may be more visually prominent from its close counterparts. Graphical elements like color, shape, size, and location are now tools that you can manipulate to attract the reader's attention on specific parts and minimize their cognitive load while interpreting your display. For example, *adding a color filling* in a specific set of values on a scatterplot *will increase the salience*:

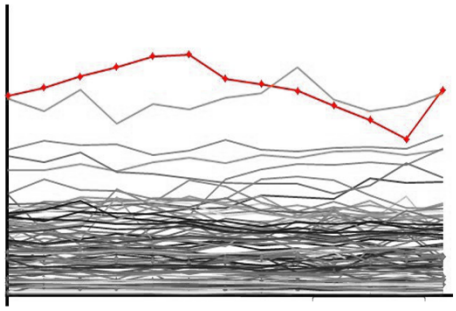


Or for example, adding a background color suggestive of grouping:

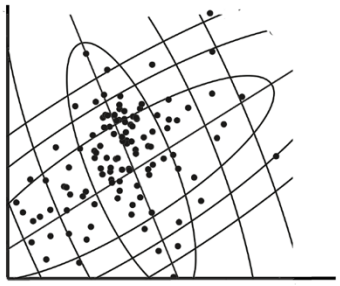


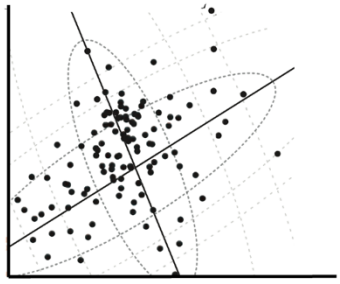
In extremely populated datasets, the reverse is also true. Instead of adding, *removing colors (i.e. desaturating)* will increase the salience of the colored parts. For example:



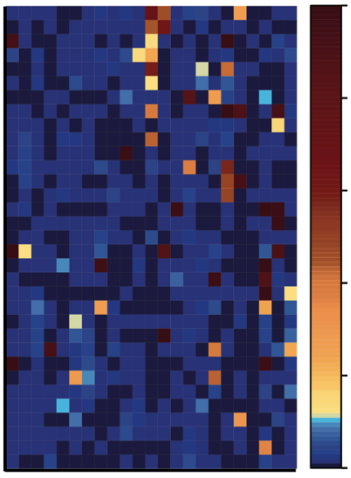


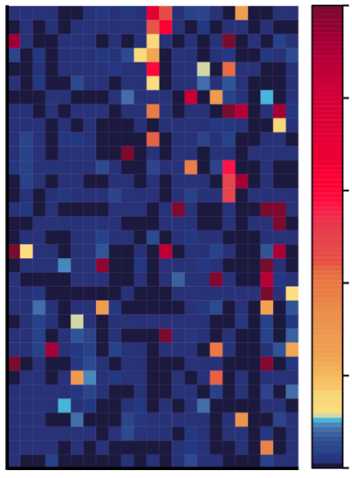
Similarly, this might also apply to *line thickness* or other graphical properties. For example, rendering navigational cues less salient, yet present on a scatterplot:





Bear in mind that such *manipulations should be done parsimoniously and should serve a purpose*. Irresolute or unintentional assignment of salient features may have the reverse effect and actually make it harder for the reader to focus on your data visualizations. For example, in the heatmap below lower values are more salient than higher ones because the color coding of deep red of the scale is hardly visible against the deep blue background of the lowest values. Changing the color scale with a brighter red in the upper scale will make the effect differences more observable:

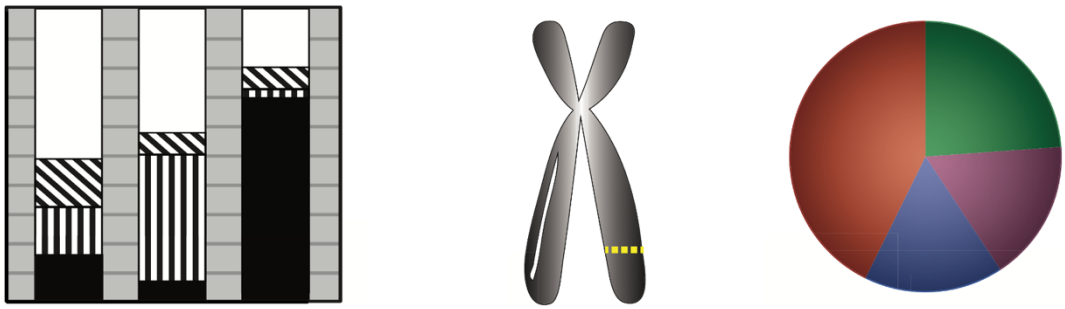




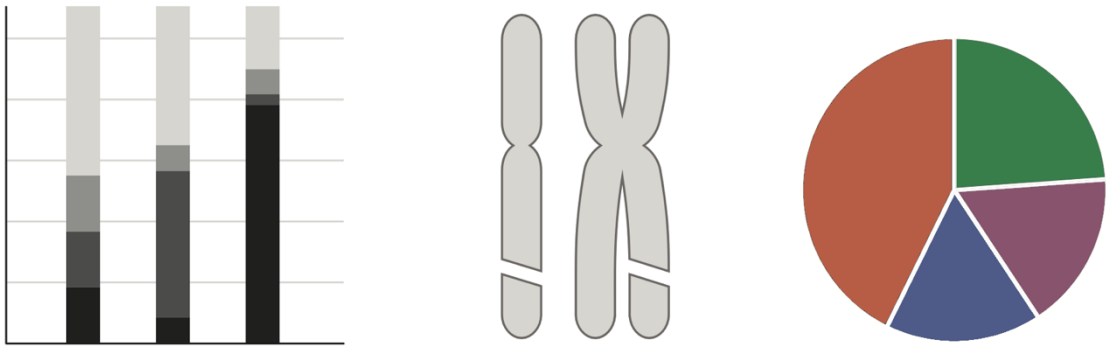
2. Clarity

On the same train of thought, overly detailed features with complex configurations create the impression of a *visually unclear display*. On the opposite tip of the balance, the concept of (optical) **clarity** instructs you to *avoid rich and ornated graphical forms* to ensure that your graphical message comes across as clear and simple as possible. After all, […] the key "is not the quality of the diagram or drawing, but the clarity of the information" [10].

For example, a multi-panel figure with *redundant visual elements* such as shades, patterns, and thick lines *results in an inconsistent design* across graphs:



Removing those redundant visual elements will result in a *more minimalistic and consistent display*, while retaining the main message that you wish to illustrate:

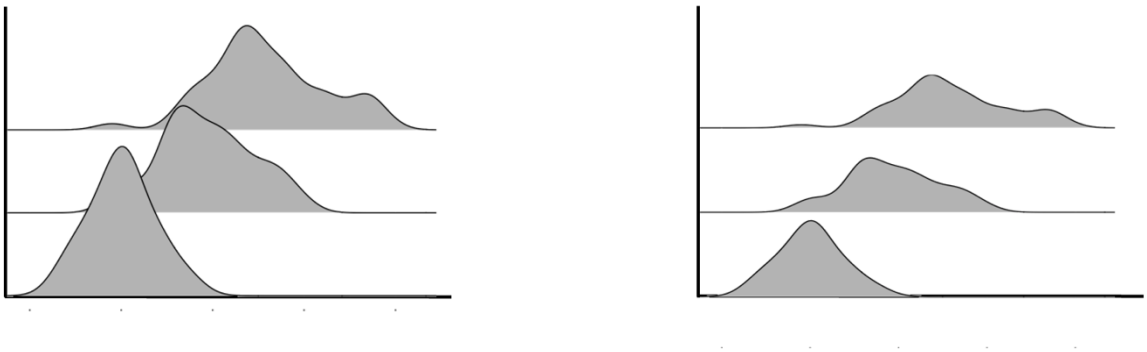


A good strategy for achieving clarity includes a) reducing your charts down to *basic shapes*, b) adding *secondary graphical information* (such as grids, statistical annotation, and patterns) only if they serve a purpose in the interpretation and c) finishing up with a *minimal and meaningful color palette*.

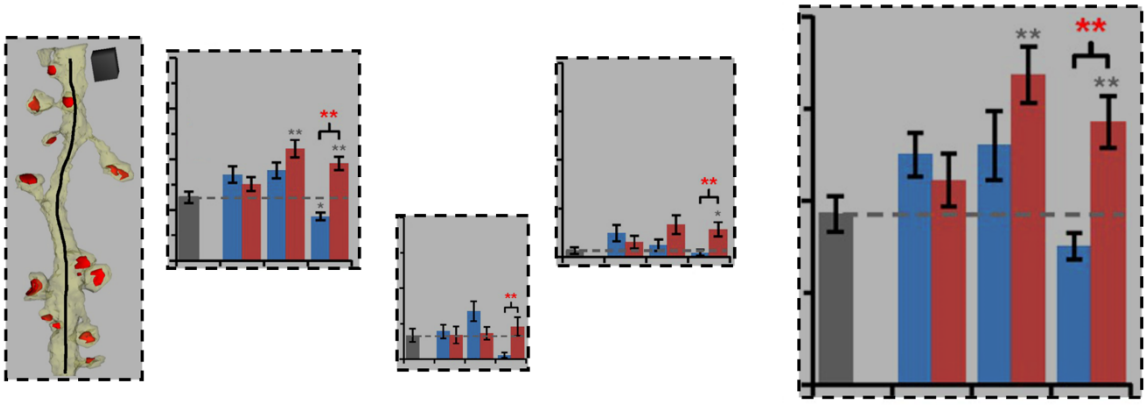
3. *White space and alignment*

White (or negative) **space** refers to the areas of a display that do not entail any graphical or verbal information, i.e. the in-between. Even though it lacks information, it is important to highlight that white space is also an essential part of your graphical composition that you should account for to avoid *visual clutter*. Appropriate spacing can provide structure and *sectioning* in your display, rendering it easier for a reader to navigate through the information presented.

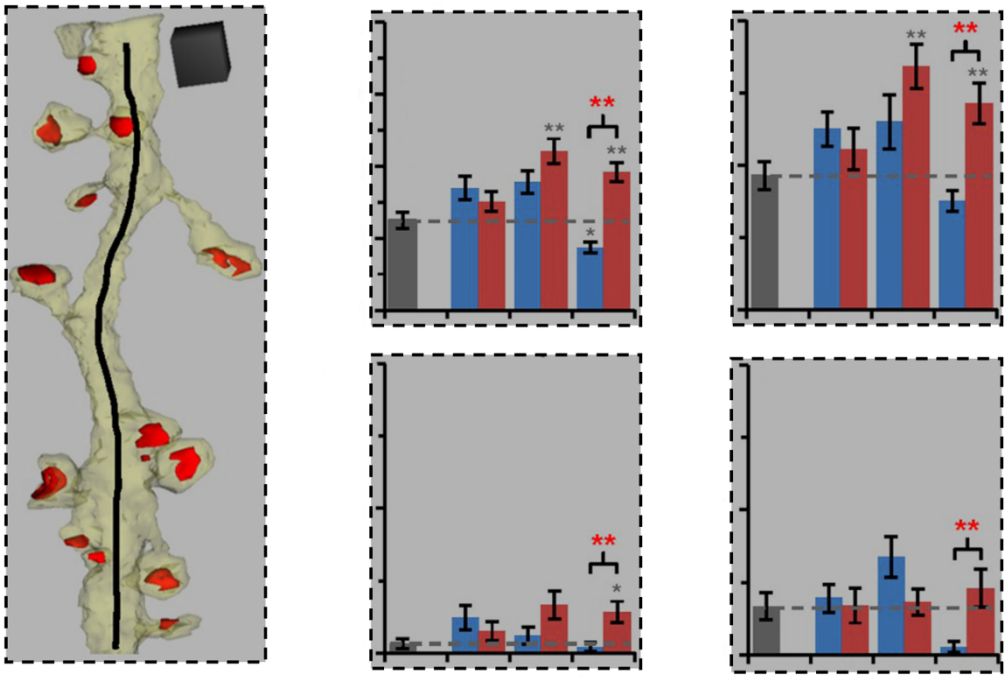
For single graphs, effective white space ensures that the graph's resolution and scaling allow the reader to fully perceive the data range. For example, reconfiguring the scale parameter of a ridgeline plot so that the tallest density curve just touches the baseline of the next higher one:



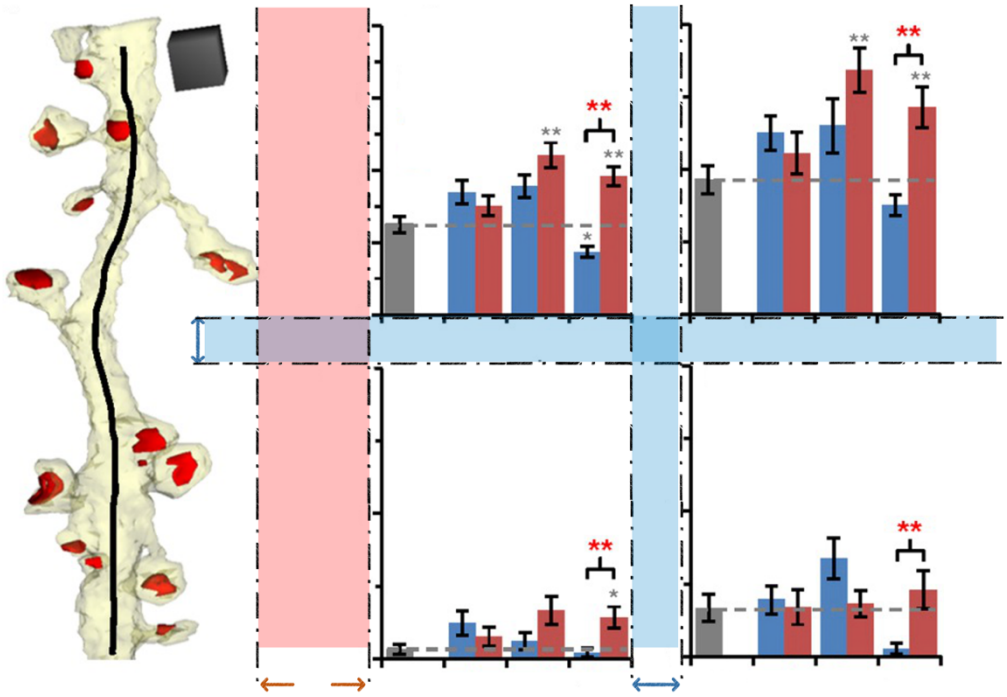
When it comes to designing multi-panel figures, posters and PowerPoint presentations, **alignment** should be also accounted for. In combination with appropriate white space manipulation, alignment will create *visually balanced* outcomes with no cluttered graphical elements. A useful approach for designing multiple graphs starts by *containing graphical* (i.e. charts, plots and figures) and *textual elements* into *virtual boxes*. For example, imaging having to form a figure from four different bar charts along and an illustration figure:



This provides an estimation of the positive (i.e. information filled) and the negative white space distribution in your display. By manipulating the size and aspect ratio of the individual graphs, you should *align those blocks both vertically and horizontally* to provide a more structured layout. For example:



This undoubtedly provides a somehow neater visualization than before. To make sectioning even clearer, *graphical objects that belong together should be proximal in space to highlight similarity*. This can be applied by reducing the white space among them (blue) and respectively increasing it between objects that are less similar (red). For example:



Data Chart(s) Checklist ☑

- ☐ **[Graphical form]** Is the graph appropriate for the given dataset?
  - *Does the graphical form correspond appropriately to dataset-specific parameters such as number and nature of the variables?*
  - *Does the graphical form illustrate the desired relationship among the data?*

- ☐ **[Navigational cues]** Are axes, ticks and gridlines well-designed?
  - *Do the axes cover the entire range of data plotted?*
  - *Are axes and ticks labelled, and units of measurement present?*
  - *If gridlines are present, do they have the optimal density?*

- ☐ **[Statistical information]** Are statistical measures (such as confidence intervals, model parameters, p-value) visualized or annotated on the graph?

- ☐ **[Color]** Are the color choices minimal and informative? Does each color correspond to a different concept, with no redundant colors used?
  - *Does the color palette highlight the data relationship you wish to convey?*
  - *Are the colors consistent across different panels?*
  - *Is the color palette inclusive of a color-blind audience?*

- ☐ **[Salience]** Are the most important elements salient (prominent/highlighted) enough (e.g. by manipulating color and/or size)? Are the less important elements less salient?

- ☐ **[Clarity]** Are the charts clearly designed, with no redundant elements (e.g. text is present only where necessary, the schematics are clearly designed)?
  - *Are the elements presented self explicable as much as possible (even for interdisciplinary audience)?*

- ☐ **[White space and alignment]** Is white space managed effectively?
  - *Are elements that belong together close (in proximity) and elements that do not belong together more distal?*
  - *Is there enough space between the sections?*
  - *In multi-panel figures, are the panels sections aligned clearly (horizontally, vertically, circularly)?*

Further reading

[1] R. Raciborski, "Graphical representation of multivariate data using Chernoff faces," 2009. [Online]. Available: http://www.math.yorku.ca/SCS/sasmac/faces.html

[2] S. D. Turner, "qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots", doi: 10.1101/005165.

[3] A. Gramfort *et al.*, "MEG and EEG data analysis with MNE-Python," *Front Neurosci*, no. 7 DEC, 2013, doi: 10.3389/fnins.2013.00267.

[4] L. Wilkinson, *The Grammar of Graphics*. Chicago, 2005.

[5] M. Krzywinski, "Points of view: Axes, ticks and grids," *Nature Methods*, vol. 10, no. 3. p. 183, Mar. 2013. doi: 10.1038/nmeth.2337.

[6] C. Kelleher and T. Wagener, "Ten guidelines for effective data visualization in scientific publications," *Environmental Modelling and Software*, vol. 26, no. 6, pp. 822–827, Jun. 2011, doi: 10.1016/j.envsoft.2010.12.006.

[7] W. Härdle, A. Unwin, and C. Chun-houh, *Handbook of Data Visualization*. 2008. doi: 10.1007/978-3-540-33037-0.

[8] J. H. Park *et al.*, "The principles of presenting statistical results using figures," *Korean J Anesthesiol*, vol. 75, no. 2, pp. 139–150, Apr. 2022, doi: 10.4097/kja.21508.

[9] B. Wong, "Salience to relevance.," *Nat Methods*, vol. 8, no. 11, p. 889, Nov. 2011, doi: 10.1038/nmeth.1762.

[10] A. Zabala, "Designing more effective scientific figures." Accessed: Mar. 27, 2023. [Online]. Available: https://bioinformatics-core-shared-training.github.io/effective-figure-design/DesigningEffectiveScientificFigures_Zabala_afternoon_v00.pdf