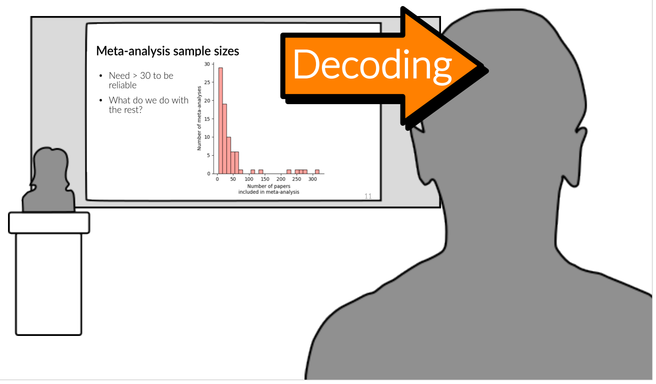
McGill ORIGAMI Lab Fonds de recherche Santé Québec NSERC CRSNG

Introduction to Data Visualization

Part 1: Decoding



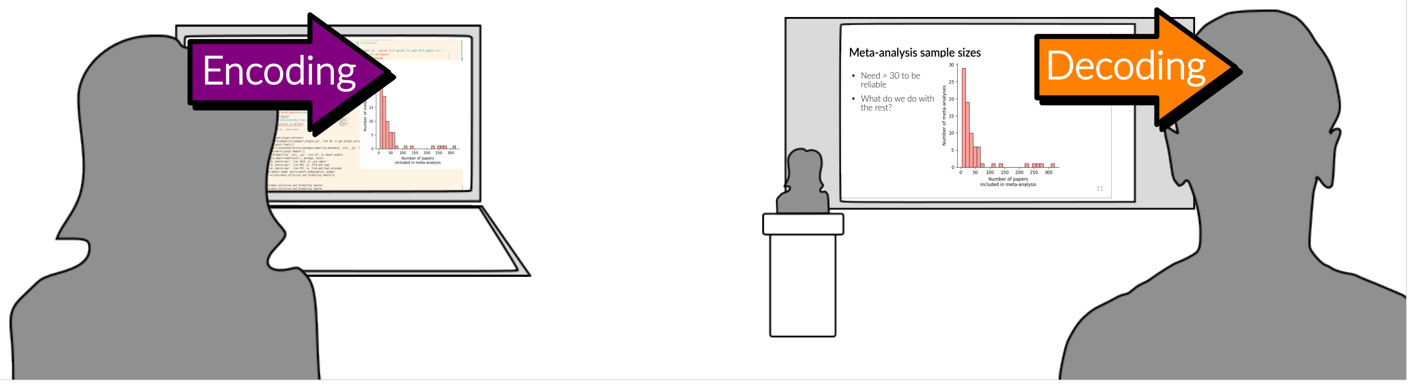
Kendra Oudyk

1

Welcome to intro to Data visualization
I'm Kendra Oudyk, I'm a 5th year PhD student in the ORIGAMI Lab, JB Poline's lab at McGill University

Goal

Use principles of visual encoding and decoding
to efficiently create visualizations
that are effective and reproducible



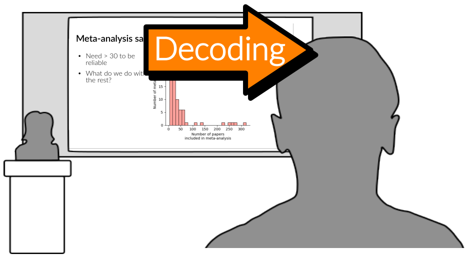
The important division here is encoding and decoding.

Decoding happens when someone sees your figure and understands something about your data.

Encoding happens when you write code to create the figure according to your plan.

In these lectures, I hope you will learn to
* use principles of visual encoding and decoding
To efficiently create visualizations
That are effective and reproducible.

Understanding decoding will help you make effective figures
Understanding encoding will help you make figures more reproducibly and efficiently



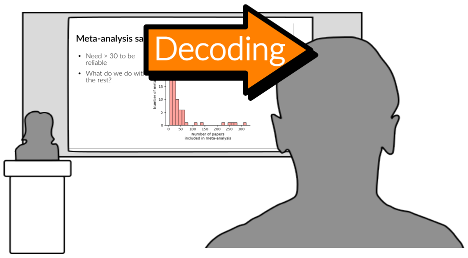
To plan an **effective visualization**, we need to think about

- Message
  - What we want to communicate
- Perception
  - How best to communicate it
- Conventions
  - How it's usually communicated
- Context
  - Where it will be seen

In this lecture, we'll talk about decoding.

This is our outline.

In order to plan an....



To plan an **effective visualization**, we need to think about

- Message
  - **What** we want to communicate
- Perception
  - **How best** to communicate it
- Conventions
  - How it's **usually** communicated
- Context
  - **Where** it will be seen

First, we'll talk about what we want to communicate.

Message

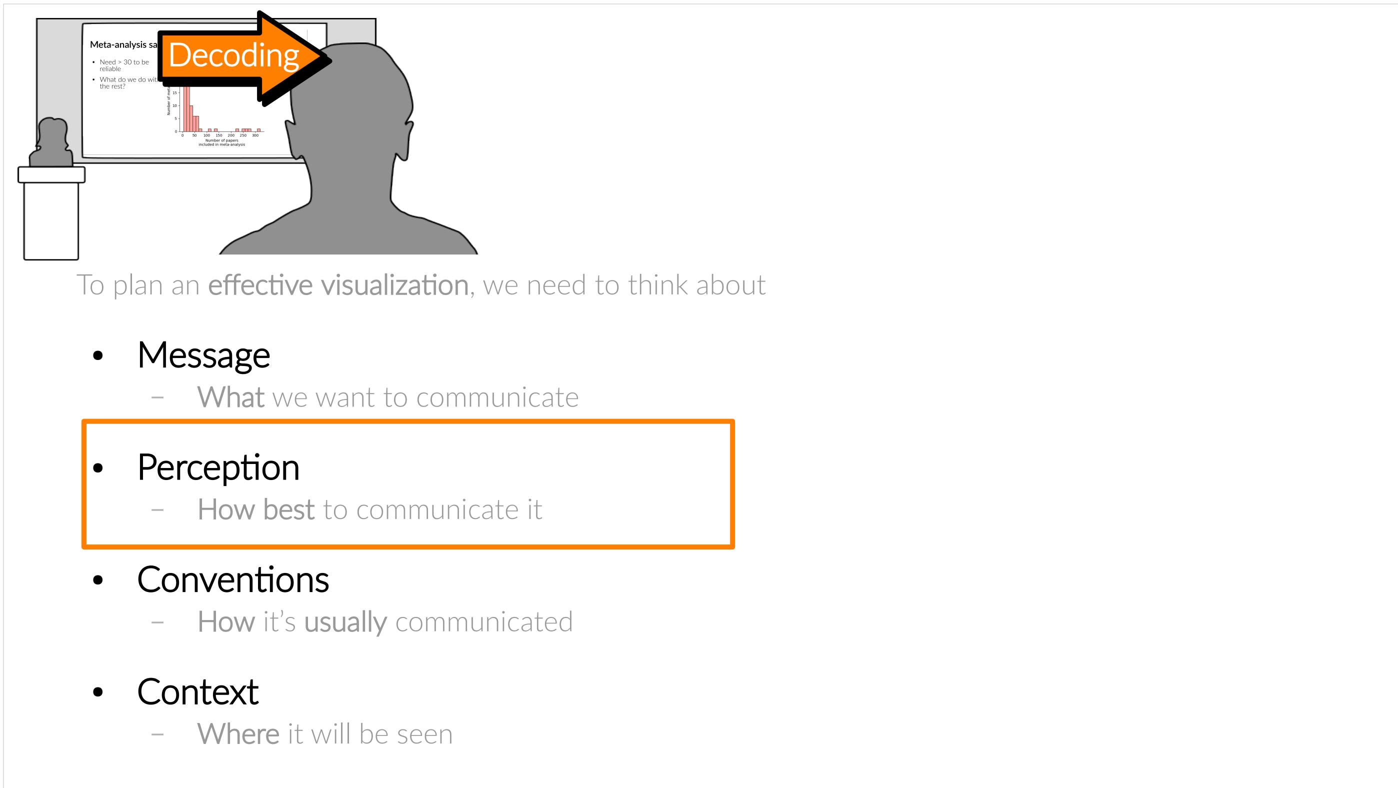
- Raw data has no message
- Abstract it
  - "There are more males than females in science"
    --> A difference between magnitudes
  - "These brain areas activate together"
    --> A grouping / pattern

*It's important to remember that while raw data has no inherent message, a visualization does.
Your design choices display, hide, and emphasize different parts of the data.

Once we know our particular message, it is helpful to abstract it.
* say we want to show that there are more males than females in science.
Essentially, we want to show a difference between magnitudes

*alternatively, say we want to show that certain brain areas activate together.
Essentially, we want to show a grouping or a pattern, and maybe the exact values are not important

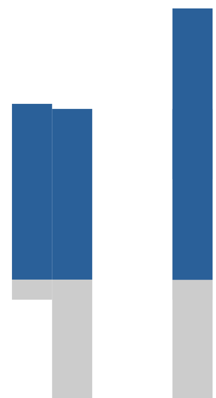
Explicitly stating our message will inform how we show our data.



Now we'll talk about how best to communicate the message, based on psychophysical evidence.

In case you're not familiar with psychophysical experiments, let's do a little demo

Which blue bar is bigger?

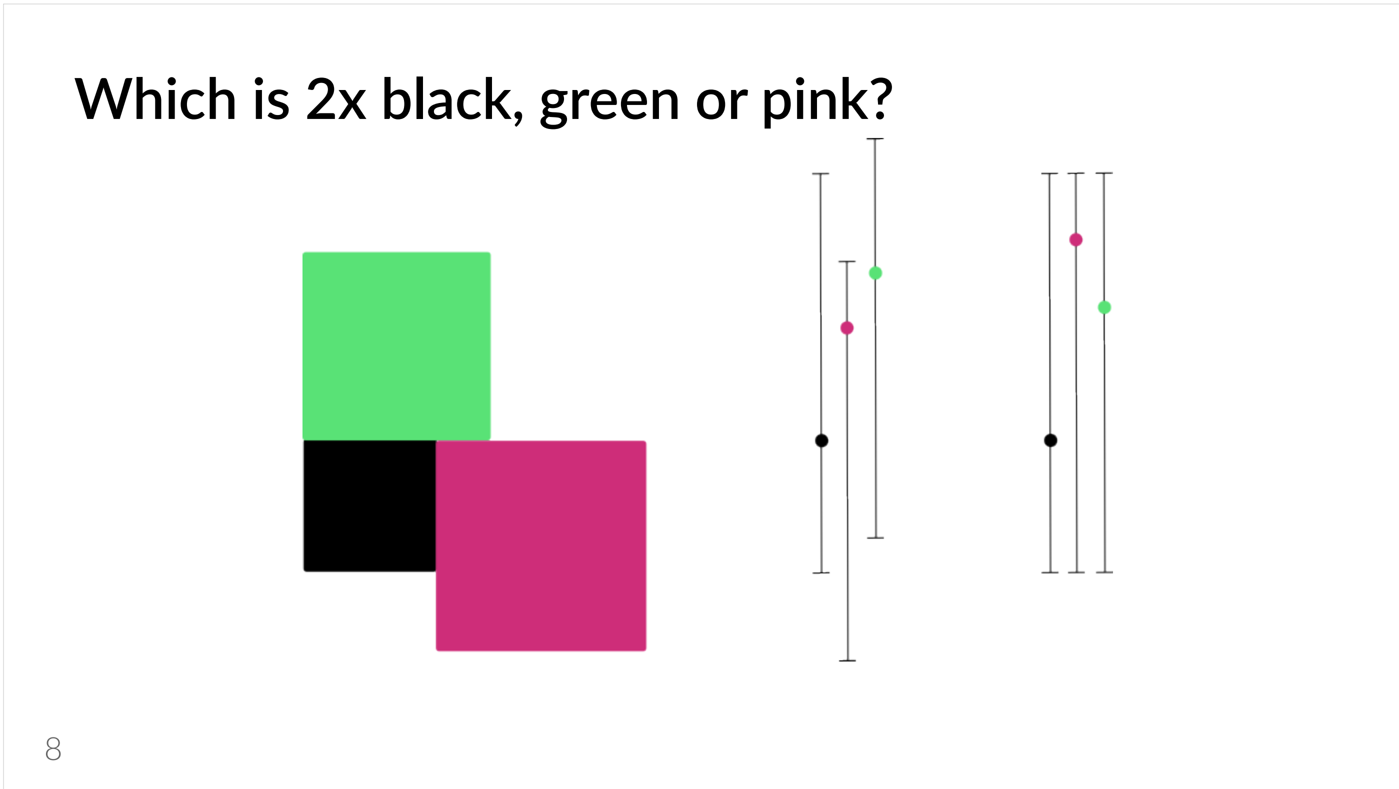


Which blue bar do you think is taller?

* what about now?

* and what about now?

You can see that alignment and proximity are important for judging two magnitudes in this way.



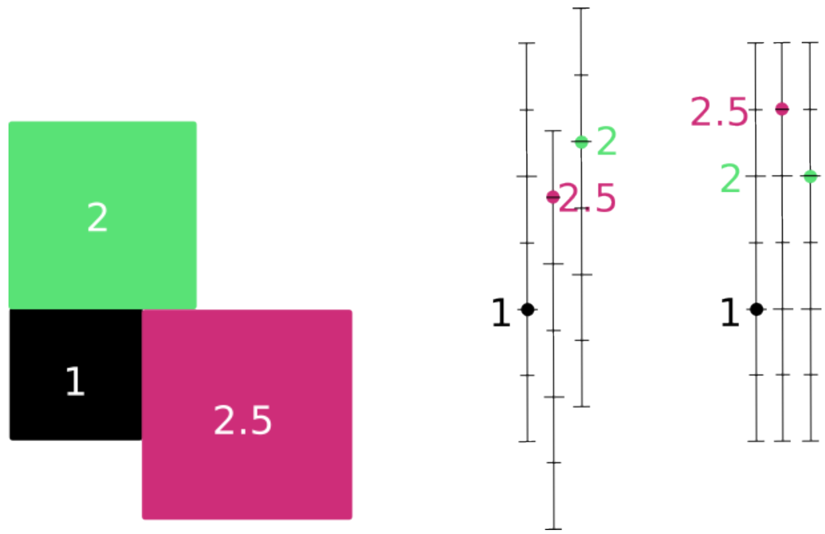
Let's try another.

Which square is 2 times as large as the black square, the pink or green one?

*which dot is 2 times as high as the black one?

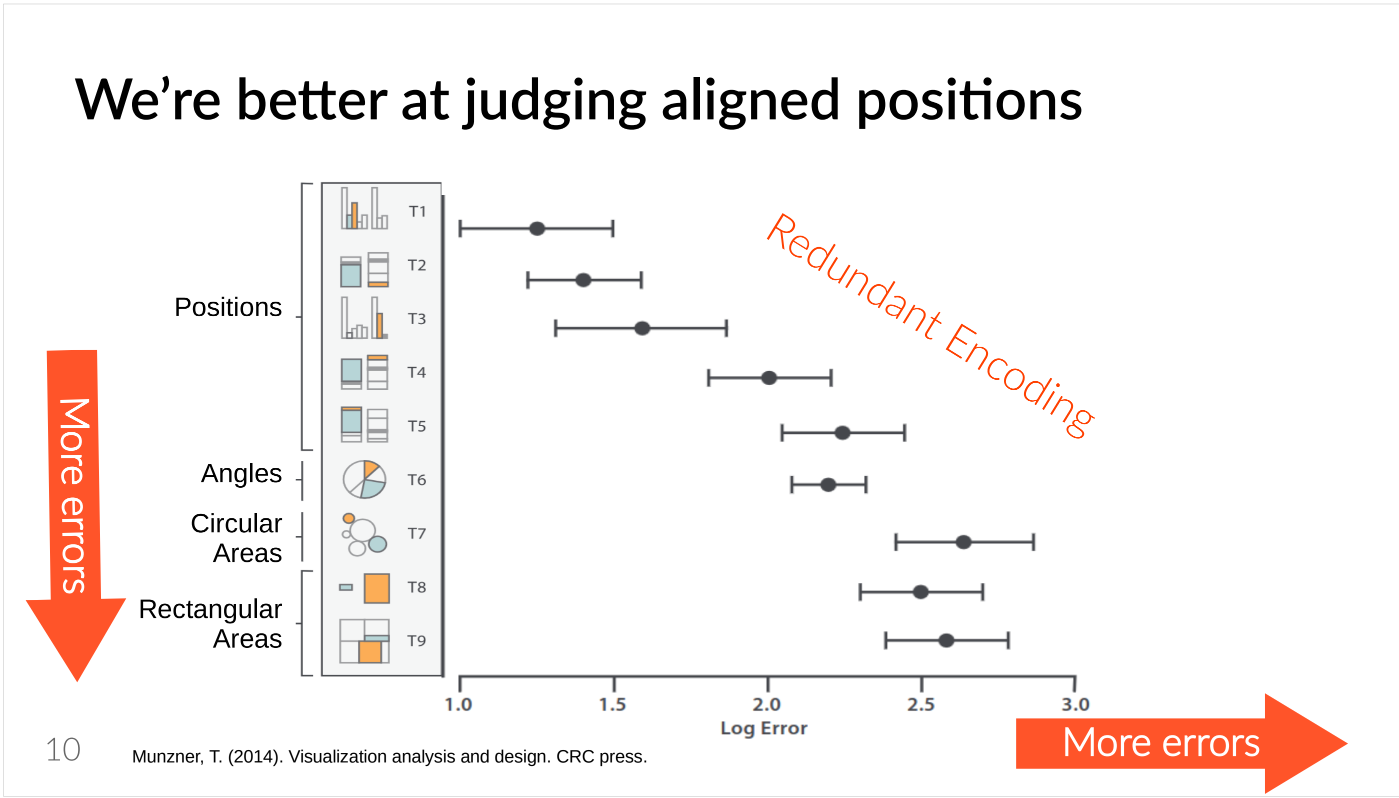
*what about now?

Which is 2x black, green or pink?



With these annotations, you can see that in each case, the green element is 2 x the black one, and the pink is 2.5 times the black one.

But it’s much easier to see this on the aligned scales.



Here's a figure showing the results of an experiment where they had people do something like what you did.

* on the x-axis, we have the error rates; there are more errors made as the data move to the right.

* they've also redundantly encoded the error rates by ordering the chart types by their error rates.

This makes it easier to see the message, that is, to see which charts are more accurately decoded.

You can see that charts that rely on positions elicit fewer errors than those that rely on angles or areas.

Decoding values vs. patterns

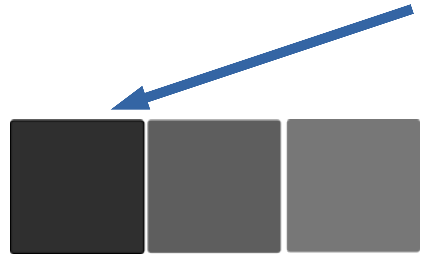
- "There are more males than females in science"
  --> A difference between **magnitudes**
- "These brain areas activate together"
  --> A grouping / **pattern**

Remember this example?

The previous slides showed that certain visualizations are better for decoding magnitudes.

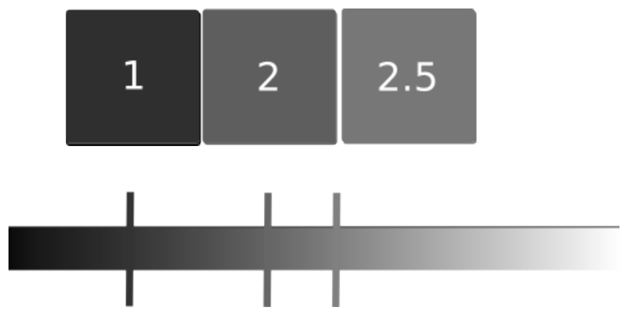
Now we're going to talk about color, which is important for visualizing things like patterns, shapes, and contours.

Which is 2x lighter than the leftmost box?



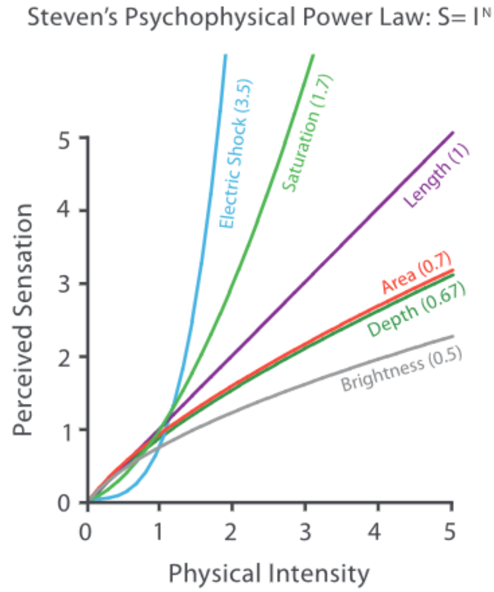
Let’s do this again, which box is two times lighter than the left box?

Which is 2x lighter than the leftmost box?



Even though these values map to positions on a single scale, it’s hard to decode exact values or comparisons.

Physical intensity vs perceived intensity



14

Munzner, T. (2014). Visualization analysis and design. CRC press.

Here we have a plot showing a model of what wejust experienced.

It plots the physical intensity of a stimulus and the corresponding physical sensation.

You can see that length is a strait line, a linear relationship. We can directly sense changes in physical length.
If you make something 2x as long, we're pretty good at decoding the change.

For something like brightness though, the relationship is nonlinear. We need to square the physical intensity in order to get a matching sensation.

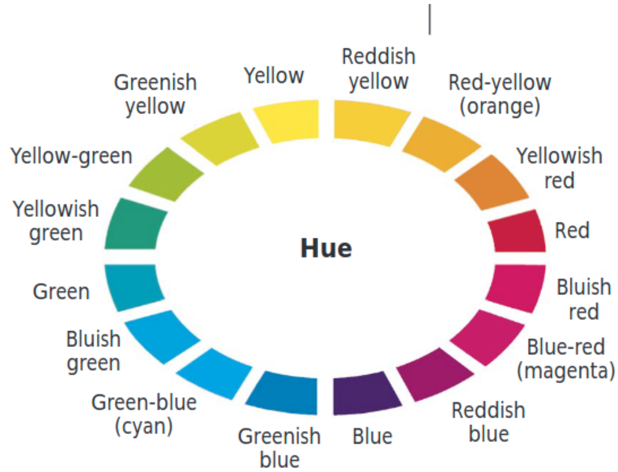
Color: salient but complicated

15

Now let's talk about color.

It is a very salient cue, very eye-catching.
But it's complicated.

Hue: how we talk about color



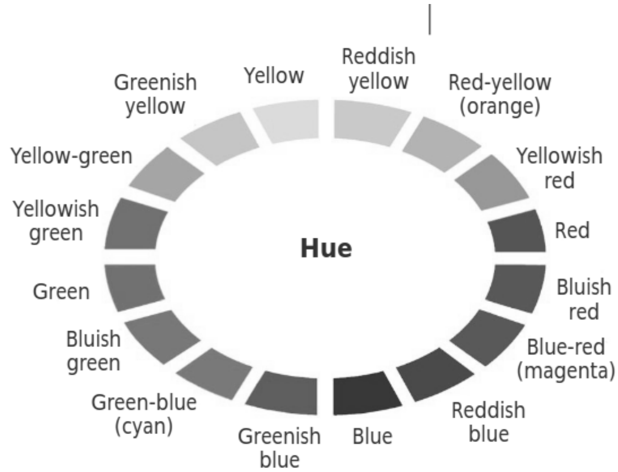
https://www.engr.colostate.edu/ECE666/Handouts/WritingPapers/UsingColorEffectively.pdf

But when most people talk about color, they use color names

I even used them when I described the previous slides.

It is most obvious that each part of this circle is a different color.
This property of color is called “hue”

Luminance: an important subconscious cue



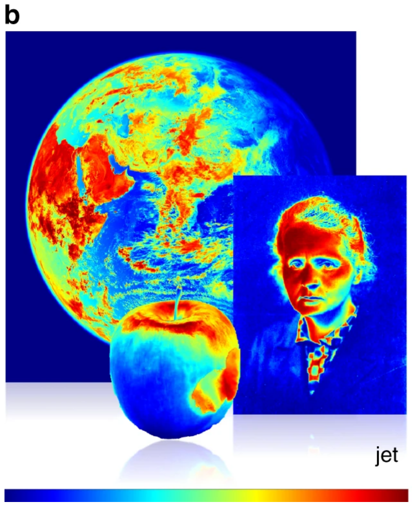
https://www.engr.colostate.edu/ECE666/Handouts/WritingPapers/UsingColorEffectively.pdf

But there are also differences in luminance, or lightness.

This is a less conscious cue, but it is very important subconsciously for letting us decode edges and shapes.

Let’s explore how this is relevant to data visualization.

Named colors don't work well for ordered values

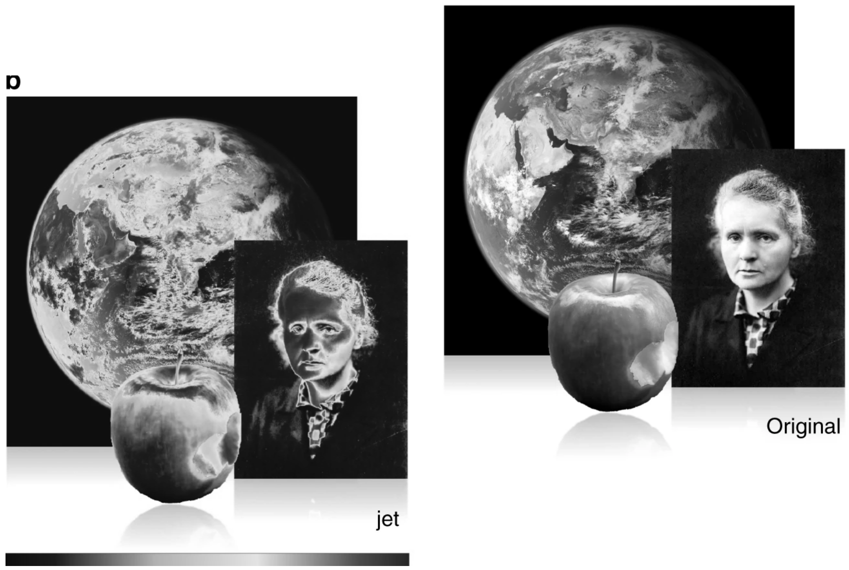


 Crameri, F., Shephard, G.E. & Heron, P.J. The misuse of colour in science communication. Nat Commun 11, 5444 (2020). https://doi.org/10.1038/s41467-020-19160-7

Named colors don't work well for representing ordered data.

These images look funny, don't they.
Let me explain why

Named colors don’t work well for ordered values



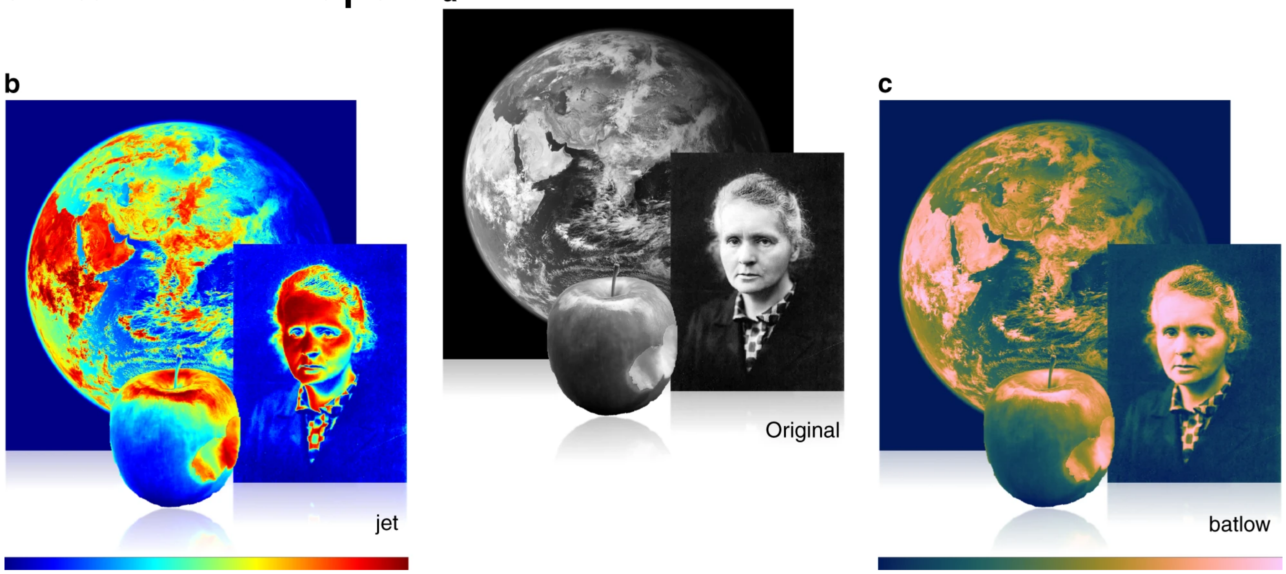
19 Crameri, F., Shephard, G.E. & Heron, P.J. The misuse of colour in science communication. Nat Commun 11, 5444 (2020). https://doi.org/10.1038/s41467-020-19160-7

To get an intuition for this, let’s look at the the rainbow-colored images in grayscale alongside the original image in grayscale.

You can see that, in the grayscaled-jet map, theres dark blotches on the lady’s face where it should be almost white.

The problem is with the luminance, or lightness, of the jet colormap

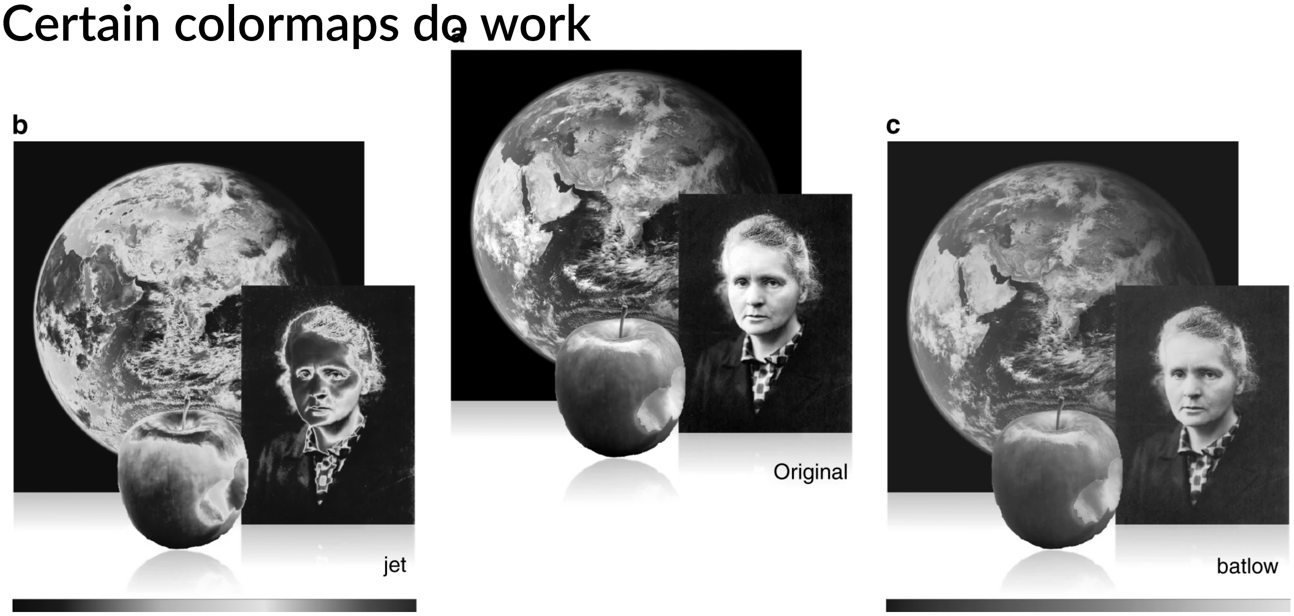
Certain colormaps do work



20 Crameri, F., Shephard, G.E. & Heron, P.J. The misuse of colour in science communication. Nat Commun 11, 5444 (2020). https://doi.org/10.1038/s41467-020-19160-7

But colors aren’t all bad, certain colormaps do work, for example, this one called batlow.

Certain colormaps do work



21 Crameri, F., Shephard, G.E. & Heron, P.J. The misuse of colour in science communication. Nat Commun 11, 5444 (2020). https://doi.org/10.1038/s41467-020-19160-7

When you look at it in grayscale, the contours still make sense.

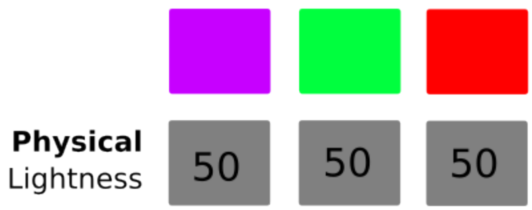
This is called a perceptually uniform colormap.

Why is this?

Which is lightest?

To get an intuition for this we need to understand the difference between physical lightness and spectral lightness.

Which square is lightest?



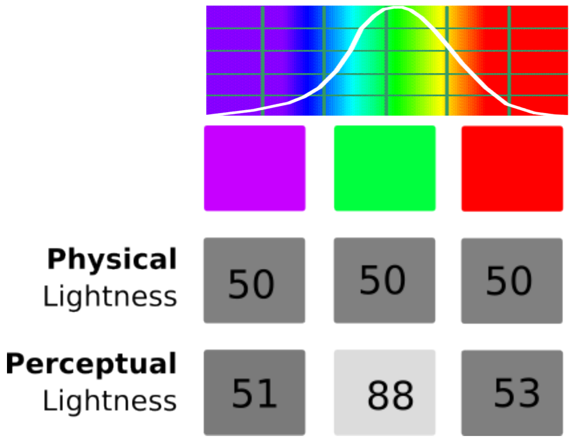
You might think that the green one is lighter.

But they all have the same physical amount of lightness.

What you’re relying on is the perceptual lightness,

For green, there’s a big difference between the physical and perceptual lightness

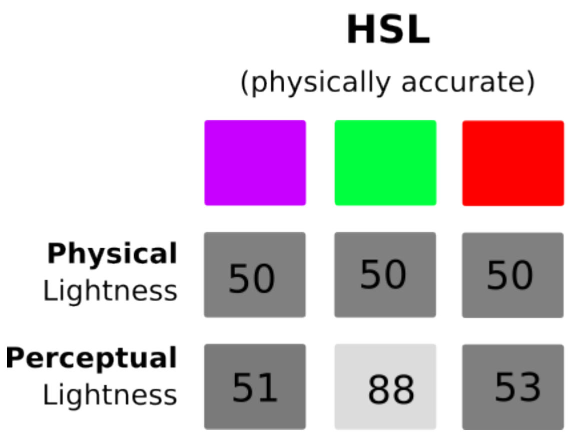
Spectral sensitivity to luminance (lightness)



 (upper figure) https://www.yorku.ca/eye/photopik.htm

This is because we're particularly sensitive to luminance in the green area of the spectrum.

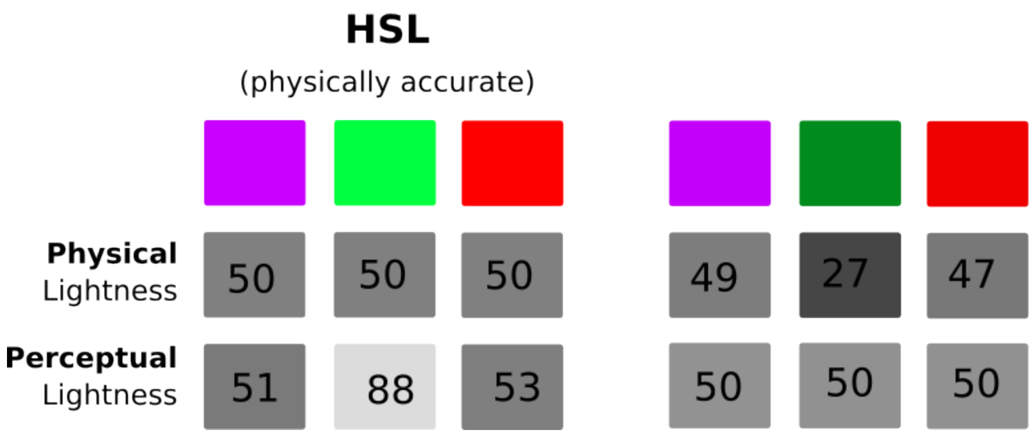
Spectral sensitivity to luminance (lightness)



These values are from the HSL way of defining color.

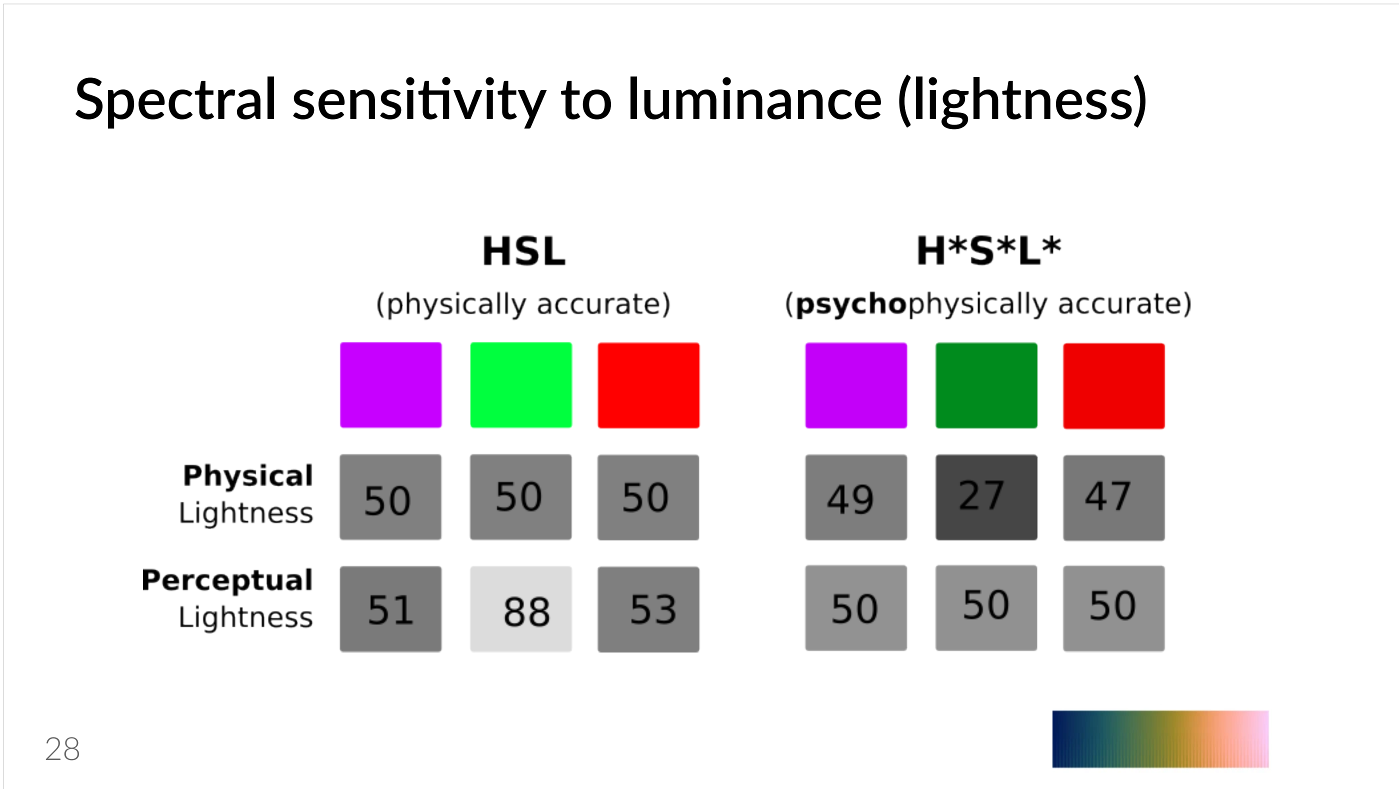
It is physically accurate, but is not perceptually accurate.

Spectral sensitivity to luminance (lightness)



Alternatively, we can use a pallette that has been controlled for spectral sensitivity to luminance.

Here, the colors have the same perceptual lightness but different physical lightness.



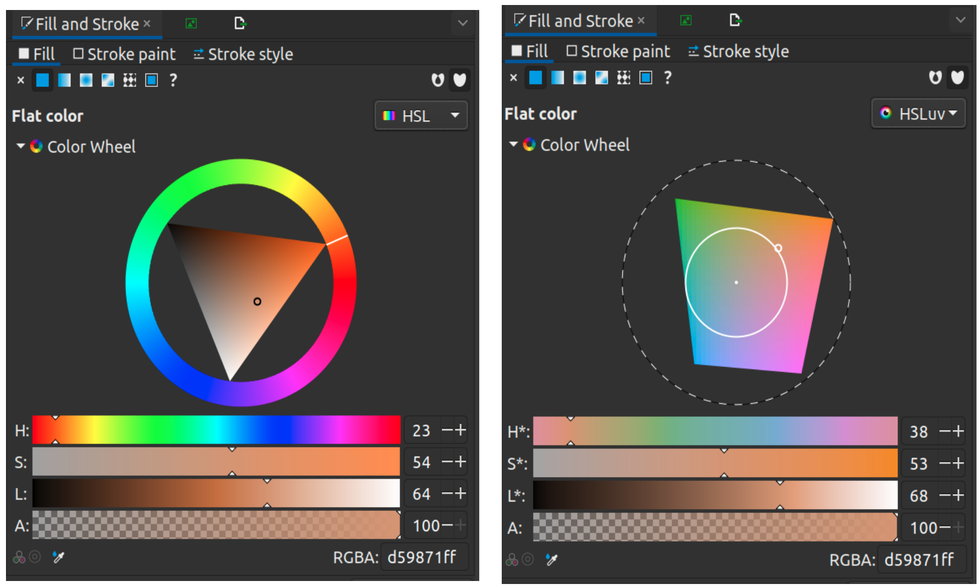
This is pallette is called HSLuv or, HSL with asterisks.

Once a pallette has been adjusted according to perception, it is called perceptually uniform.

* here's a perceptually uniform colormap, derived from this kind of palette. It looks duller because some colors need to be darkened in order to match the perception of other colors

HSL vs H*S*L* (or HSLuv)

Hue
Saturation
Lightness / Luminance



To understand this problem, we need to think about color in terms of Hue, Saturation, and Luminance (or lightness).

Go to Inkscape

I'm just going into Inkscape now to demonstrate these scales. Inkscape is an open-source program for making vector graphics.

Here we have HUE, which range of named colors, or the different physical wavelengths of light.

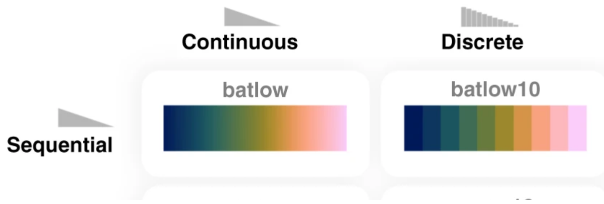
Saturation, here, is the range between grayscale and full color. When I make an image grayscale, I keep the lighness and hue the same and set the saturation to 0.

Luminance or Lightness is the amount of black and white in the color

This is HSL, the physically accurate model of hue, saturation and luminance.

This, alternatively, is HSLuv, which is pyschophysically accurate.

Types of colormaps



30 Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. Nature communications, 11(1), 5444.

It’s not enough to know you need ot use a perceptually-uniform colormap.

You also need to match it to the kind of data you have.

If you have ordered data, you should use a sequential map in order.

The map increases monotonically in luminance.

The choice between a continous map or a discrete map will depend on whether small differences in values are important.

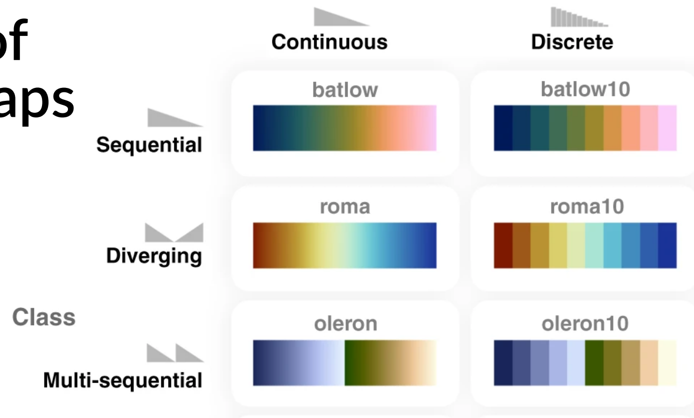
Types of colormaps

Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. Nature communications, 11(1), 5444.

You should use a diverging colormap if you have data that has a logical middle point.

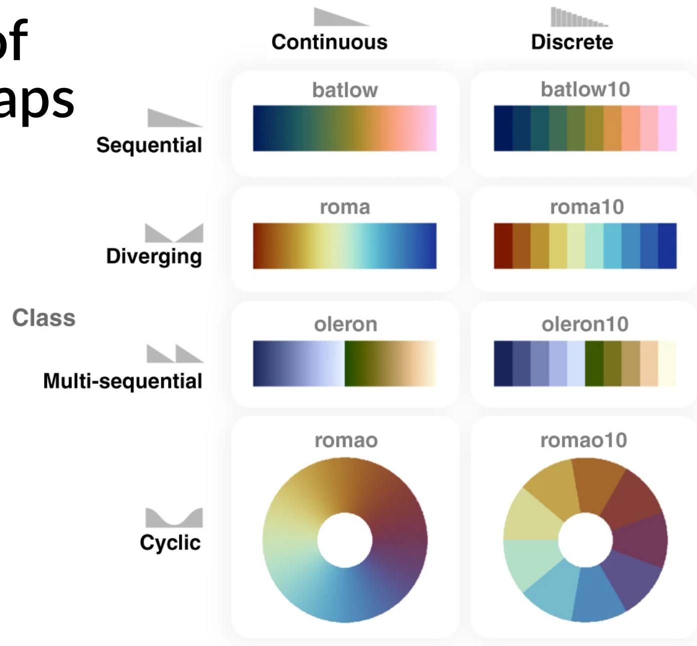
For example, it is useful for plotting correlation coefficients, which can be positive or negative.

Types of colormaps



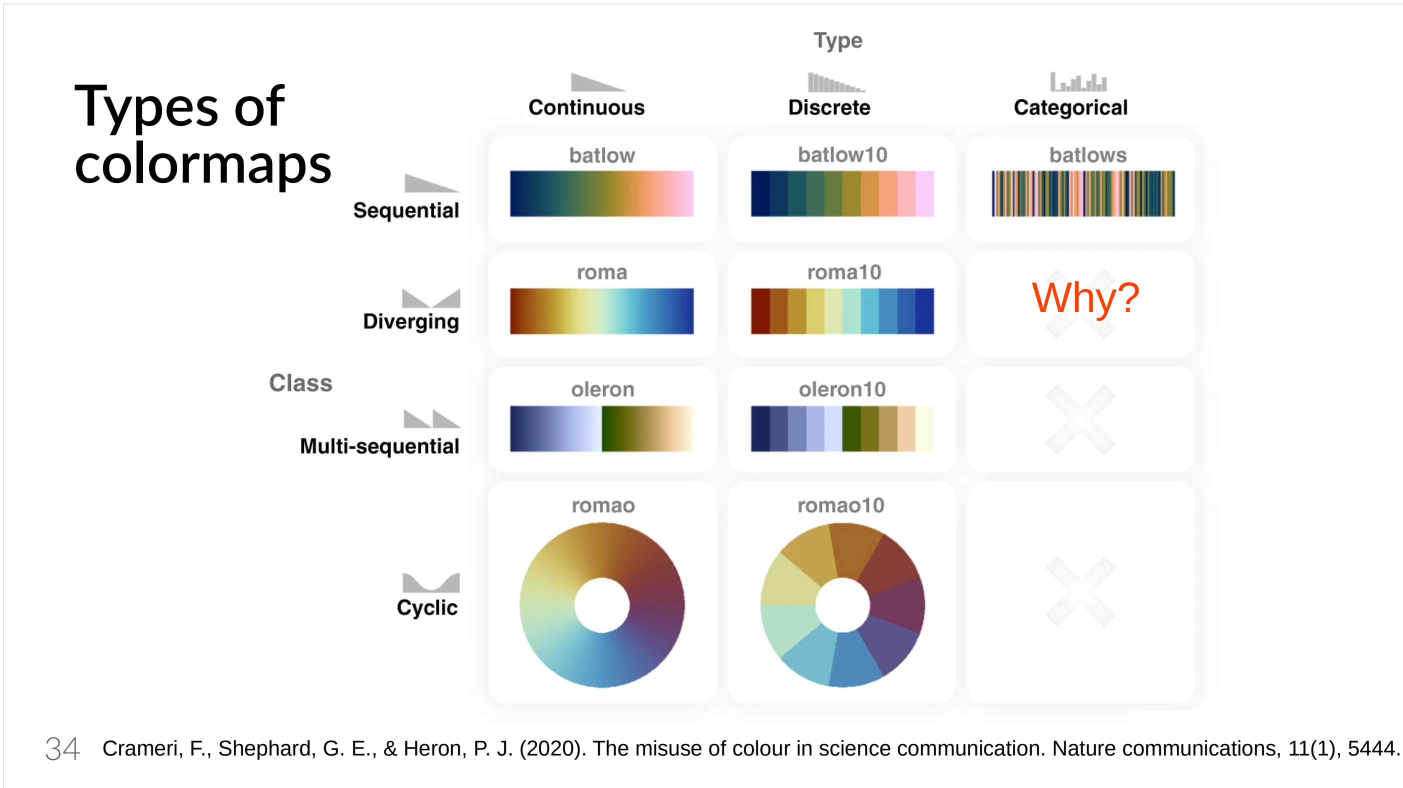
 Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. Nature communications, 11(1), 5444.

Types of colormaps



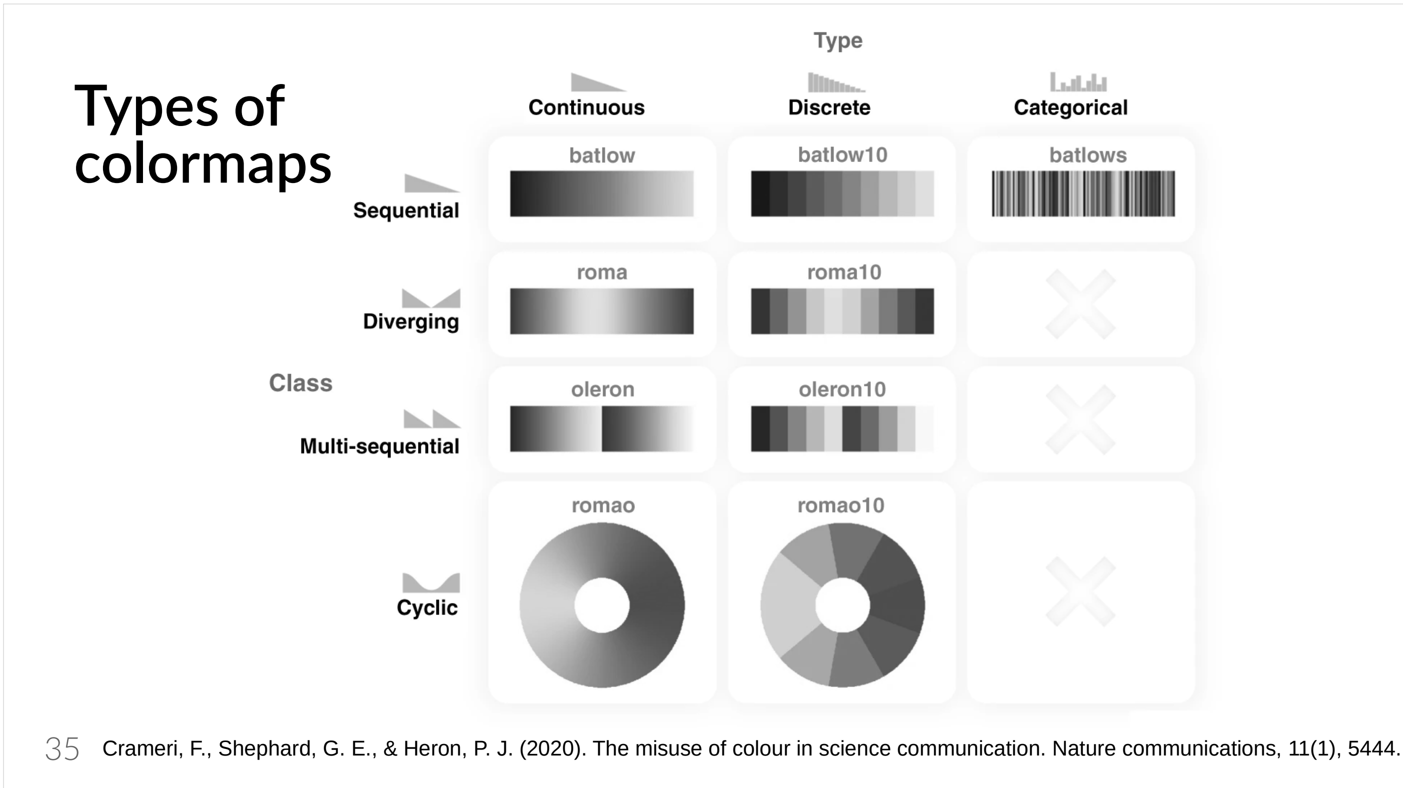
33 Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. Nature communications, 11(1), 5444.

Lastly, if your data is ordered cyclicly, you might want to use a cyclic colormap.



For categorical data, you should use colors from a sequential colormap.

Can any of you think of why?



It becomes more obvious when you look at the maps in grayscale.

If each color maps onto only one level of luminance, then the chart should be understandable in grayscale.

How do I remember all that??

- Test yourself, your lab mate, your friend
  “Which bar is higher?”

- Look at the image in grayscale

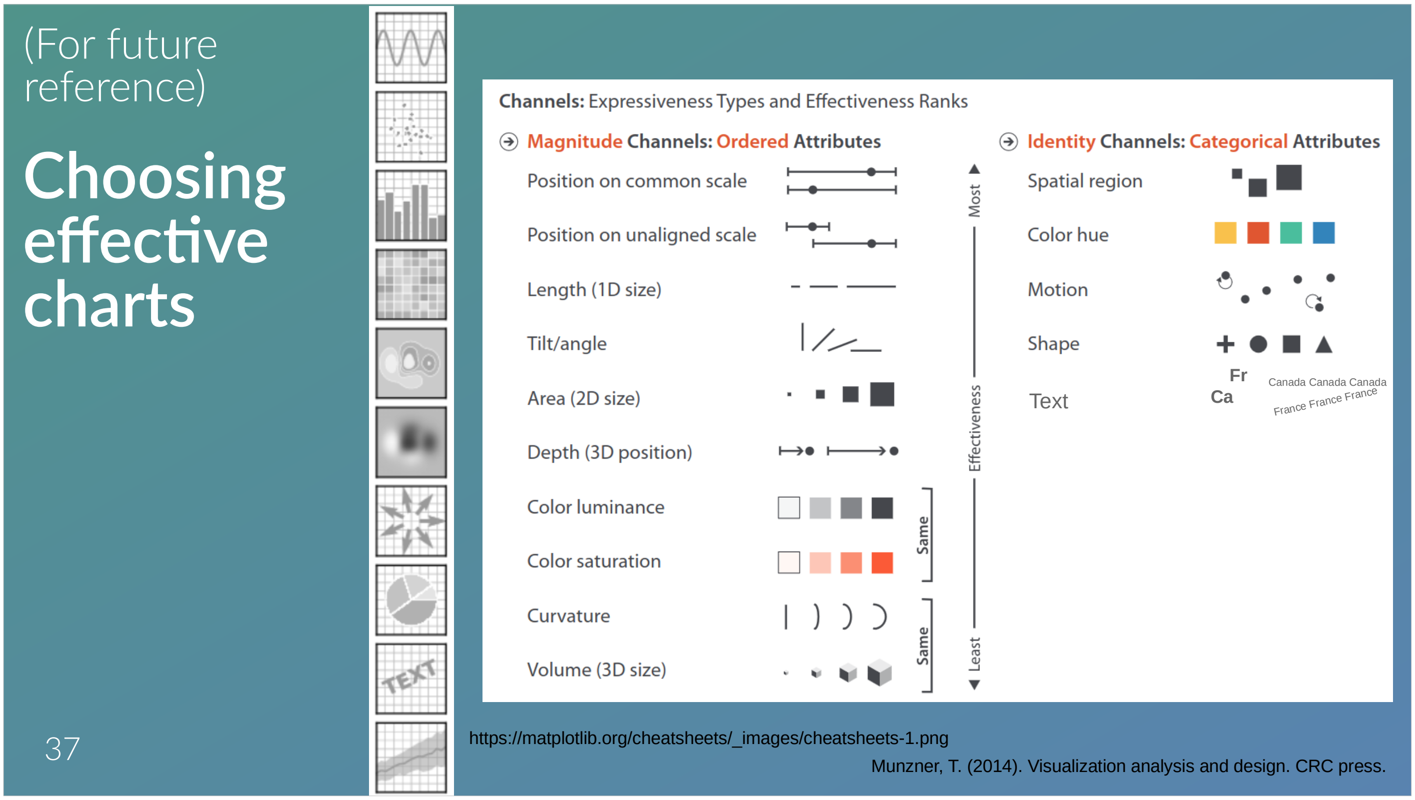
Now, that was a lot of information.

I’ve created some ‘cheatsheets’ for different topics so you can easily refer back.

But as a rule of thumb, you can just test yourself or someone else as to what different elements in your visualization mean.

For example, which bar is higher?
What does blue mean?
If they say “I don’t know”, you might need to change it.

Finally, I hope you’ve realized it’s useful to look at things in grayscale to check if the colors are working.



Here’s the cheatsheet I made for choosing effective charts.

It’s a busy slide because I wanted you to be able to see and compare all of the info.

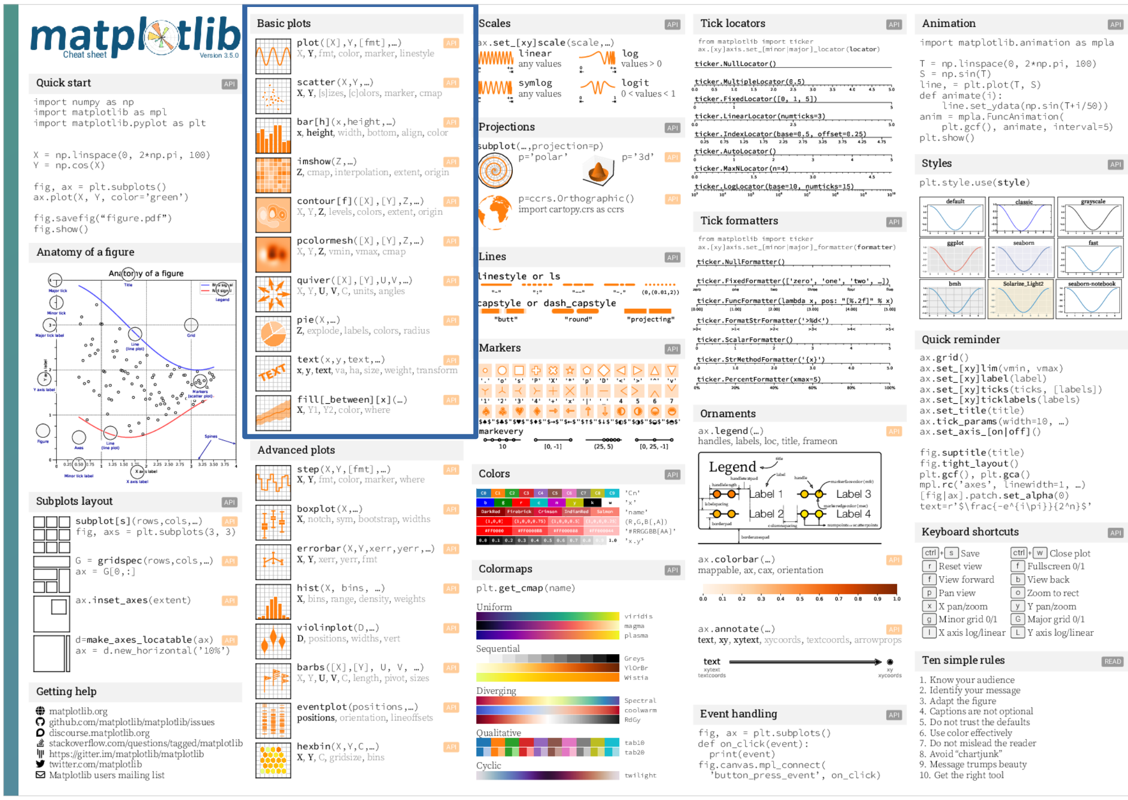
On the right, we see a summary table showing which channels are most effective for expressing magnitides and identities.

For example, it’s easier to tell that one value is twice another if they are represented as dots on the a common scale, as opposed to having different color saturation.

Notice that this chart on the right doesn’t outline which channels are best at expressing higher-level messages, like patterns.

Luminance may not be effective for decoding exact values, but it can be effective for decoding patterns and shapes.

On the left we have different kinds of charts.



https://matplotlib.org/cheatsheets/_images/cheatsheets-1.png

Just an FYI, this list of charts comes from a matplotlib cheatsheet that we'll use in the next lecture.

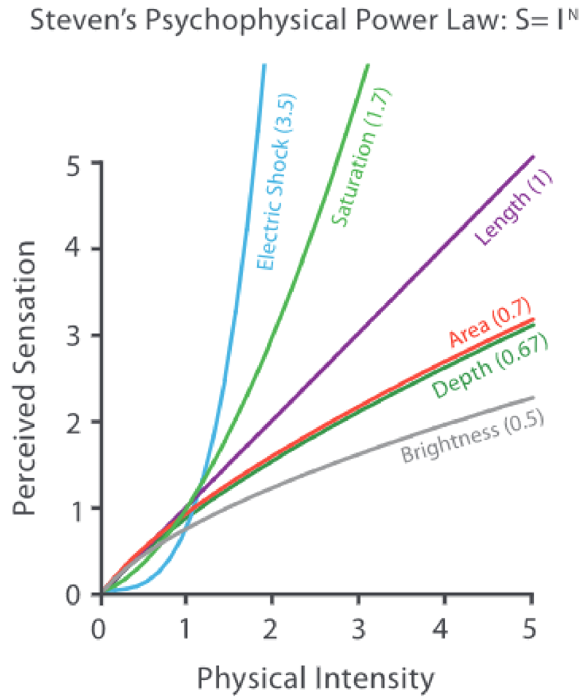
So you'll be able to easily match the figure you want to the code you need.

Now we'll go through some examples of chart types to see what visual elements they use to show data

Lineplot

Ordered

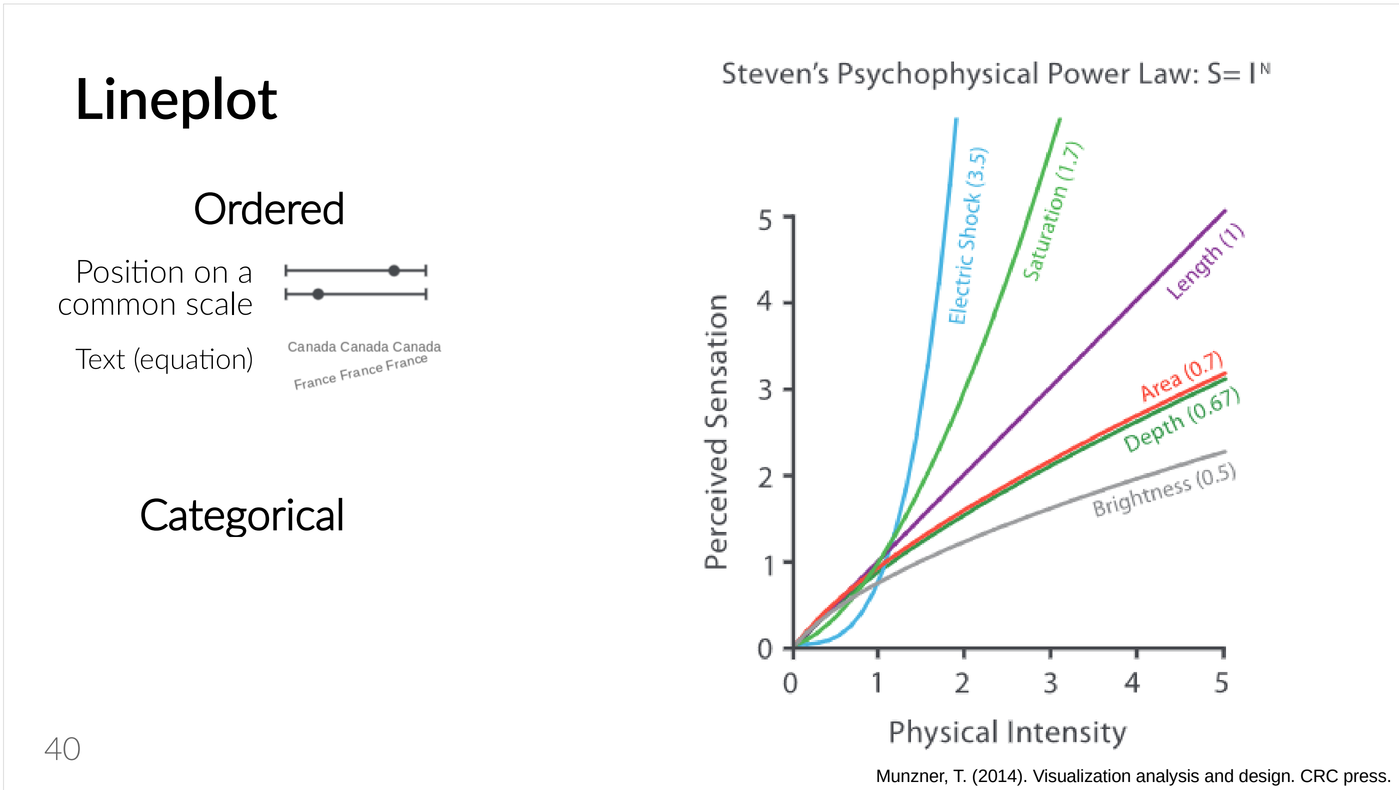
Categorical



Munzner, T. (2014). Visualization analysis and design. CRC press.

First, we have the line plot.
We've seen this plot before.

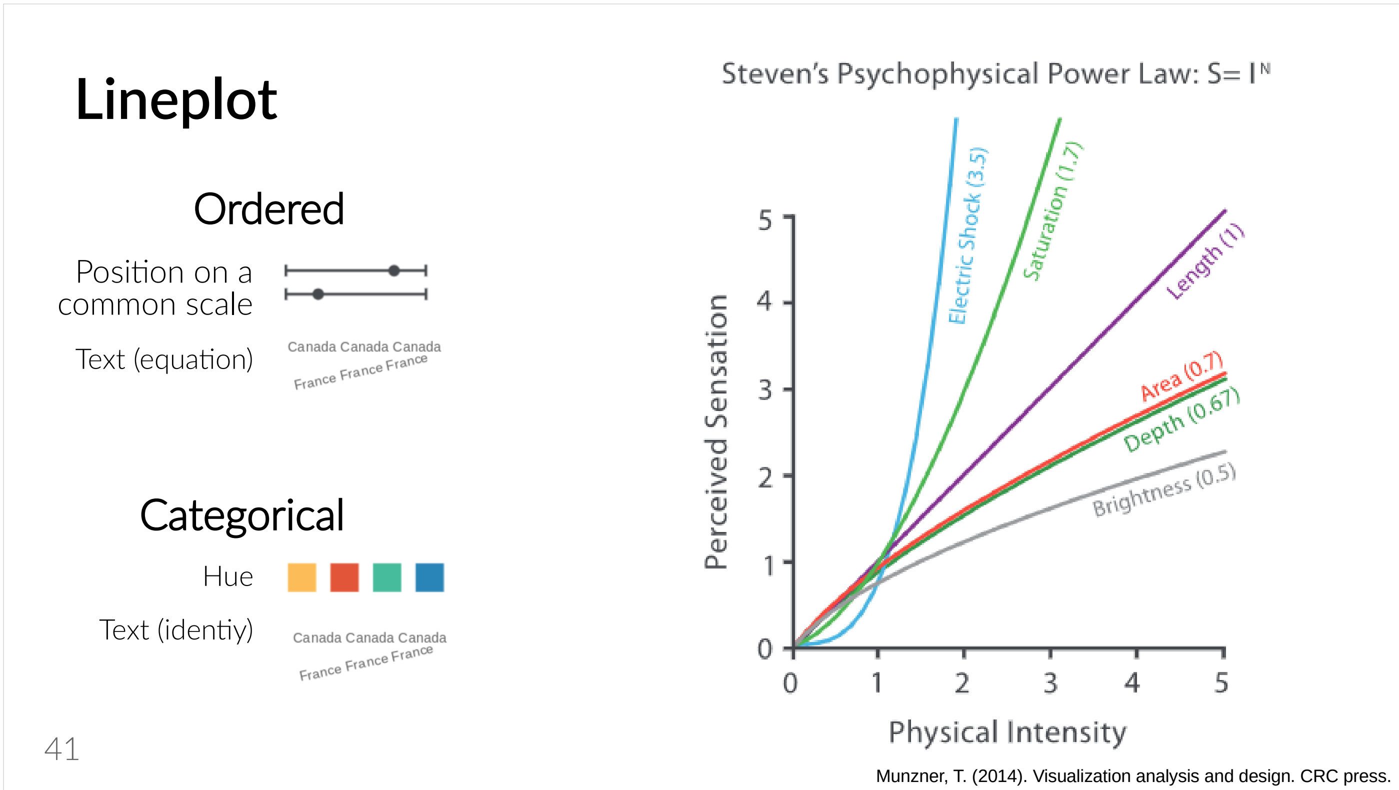
Now let's unpack how it communicates



Ordered values are encoded by their position on common scales, which are the axes.

In a way, this is redundantly encoded with the text along each line because it shows the number needed to understand the equation for that line.

For example, for brighness the number is 0.5. The equation would be the physical intensity to the power of 0.5. So we know that we could get the perceived sensation by squaring the physical intensity.

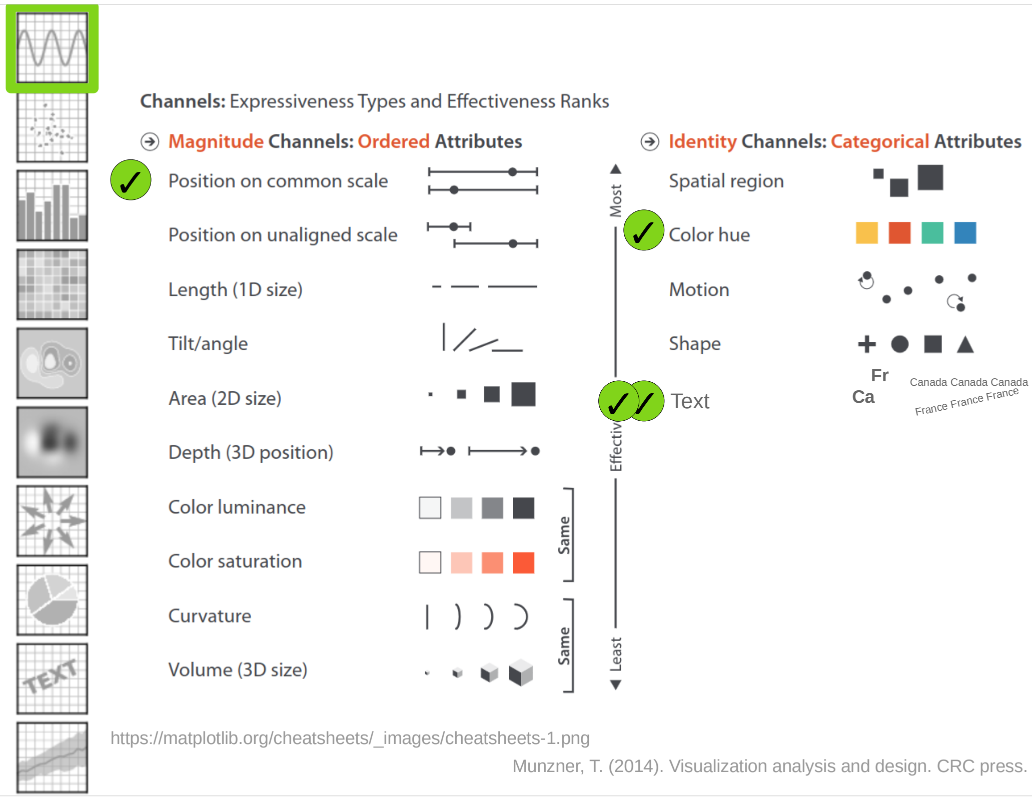


The categorical values are encoded with color and text.

Putting the text right beside the lines is very effective here.

If they were in a legend, it would require more cognitive effort to remember the color and look up its meaning.

Lineplot



https://matplotlib.org/cheatsheets/_images/cheatsheets-1.png

Munzner, T. (2014). Visualization analysis and design. CRC press.

Here we see the cheatsheet

I’ve put a checkmark beside the elements that our example used.

We may know intuitively that it was a good visualization.

But now we can see why.

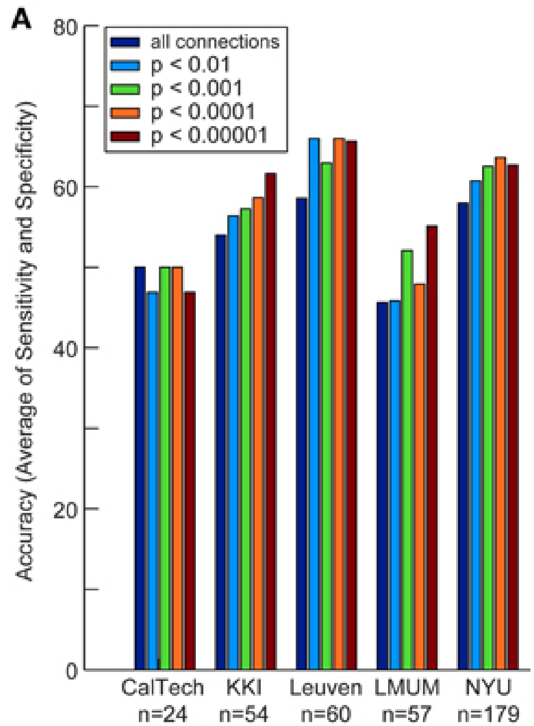
It used elements that are high on the effectiveness scale.

And it redundantly encoded both the categories and magnitudes with effectively-positioned text.

Barplot

Ordered

Categorical



Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., ... & Anderson, J. S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in human neuroscience, 7, 599.

Here we have a bar chart.

It is from the original paper for the ABIDE dataset. They classified participants with autism using brain functional connectivity.

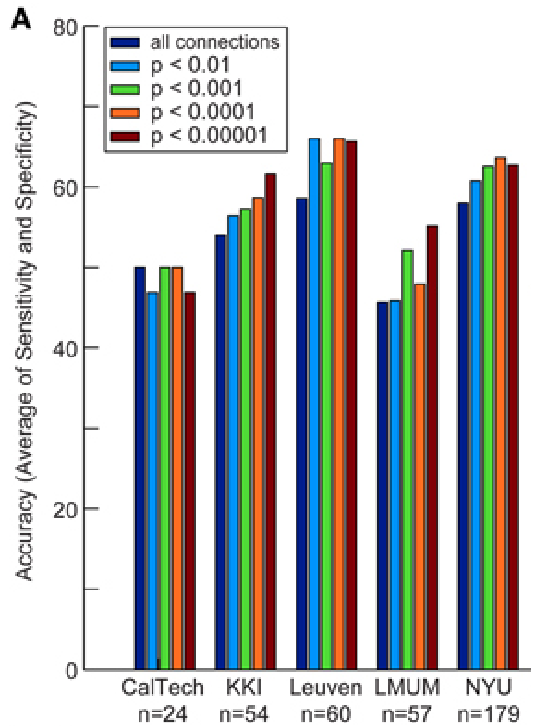
Here we see the classification accuracy for different scales of connectomes, and for different sites where they collected the data.

Barplot

Ordered

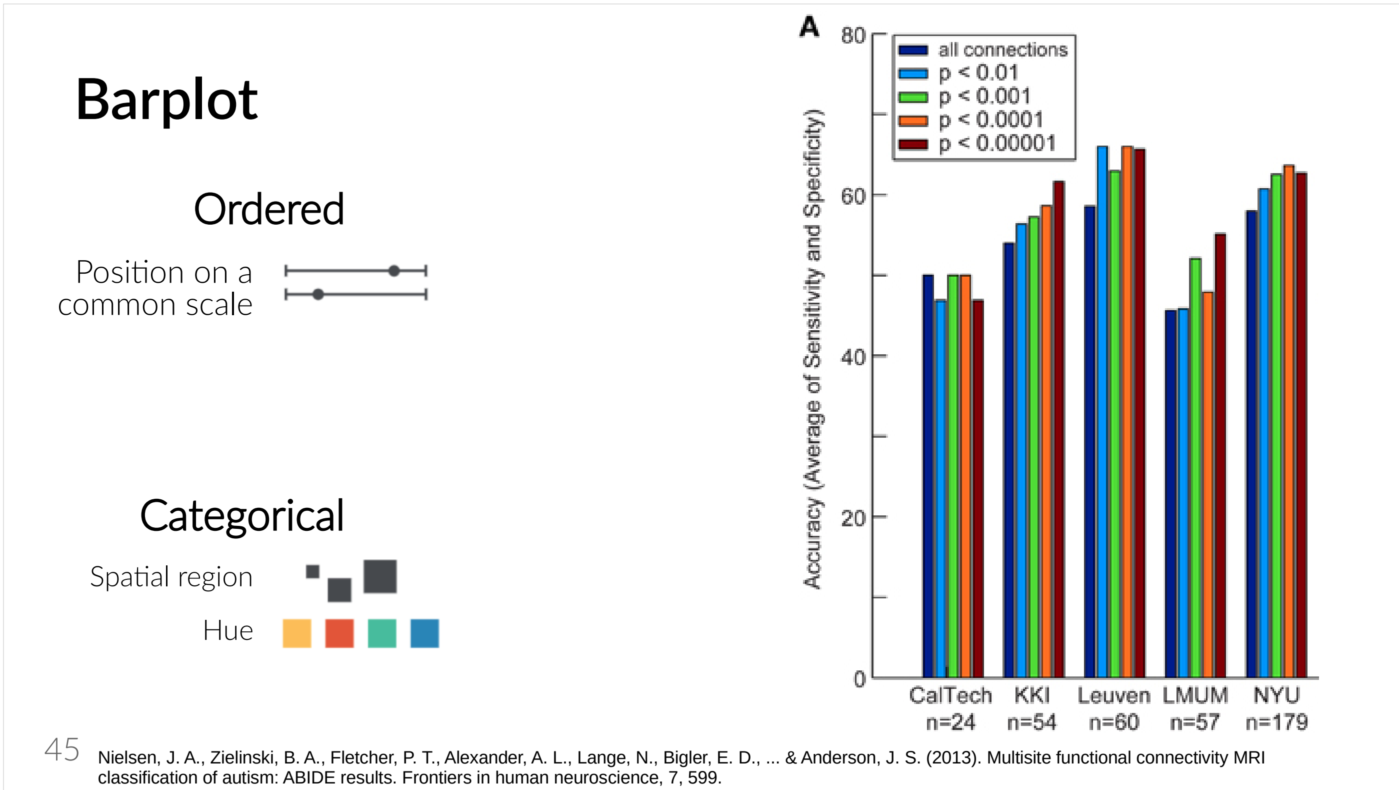
Position on a common scale

Categorical



Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., ... & Anderson, J. S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in human neuroscience, 7, 599.

For the ordered variable of accuracy, they used position on a common scale



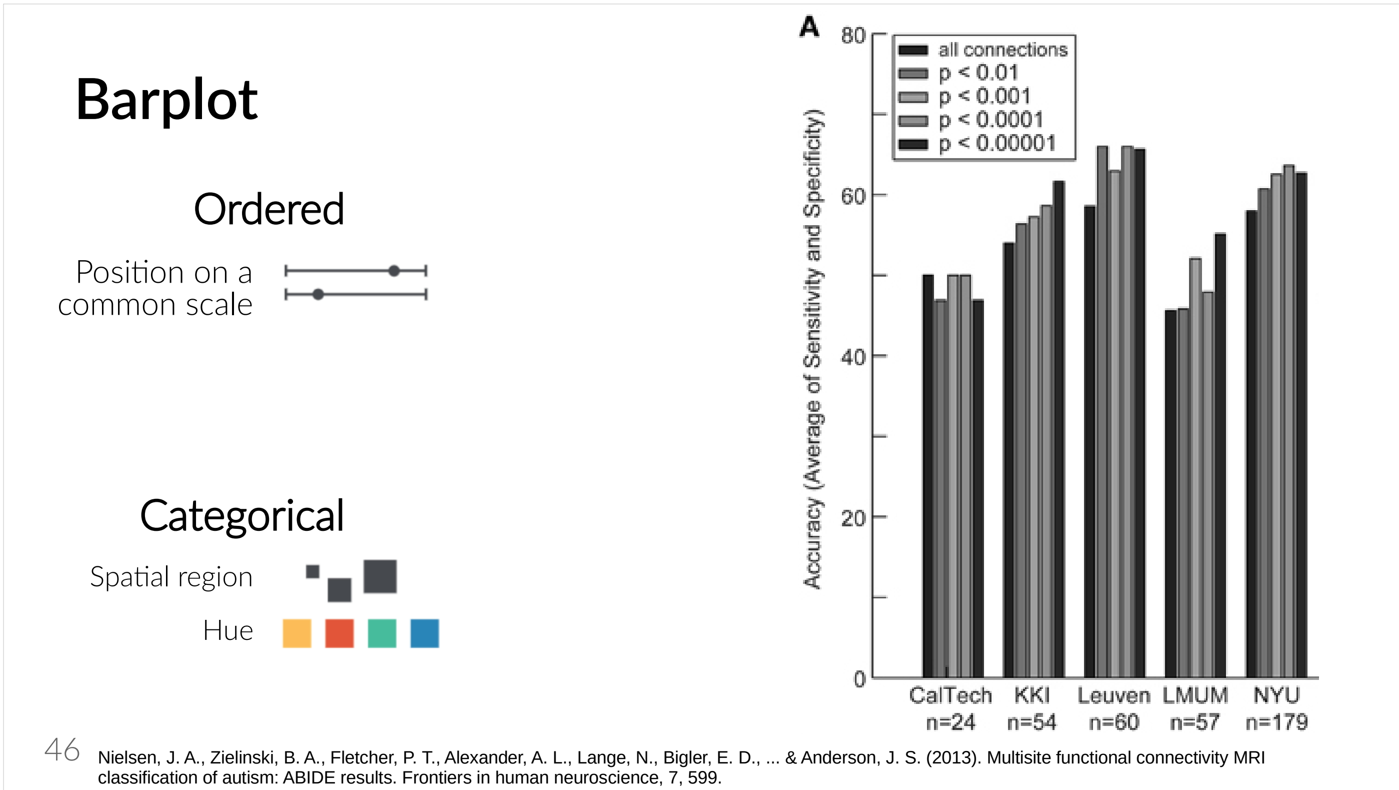
For the categorical variable of site, they used spatial region.

For the variable in the legend, they used hue.

But let's think about that variable.
It's the threshold that determined which connections were included in the connectome that was used to classify diagnosis.

As we decrease the significance threshold, we decrease the number of connections included.

So essentially, this is an ordered variable, not really categorical.



Again, it becomes obvious that their colors aren't effective when we see them in grayscale.

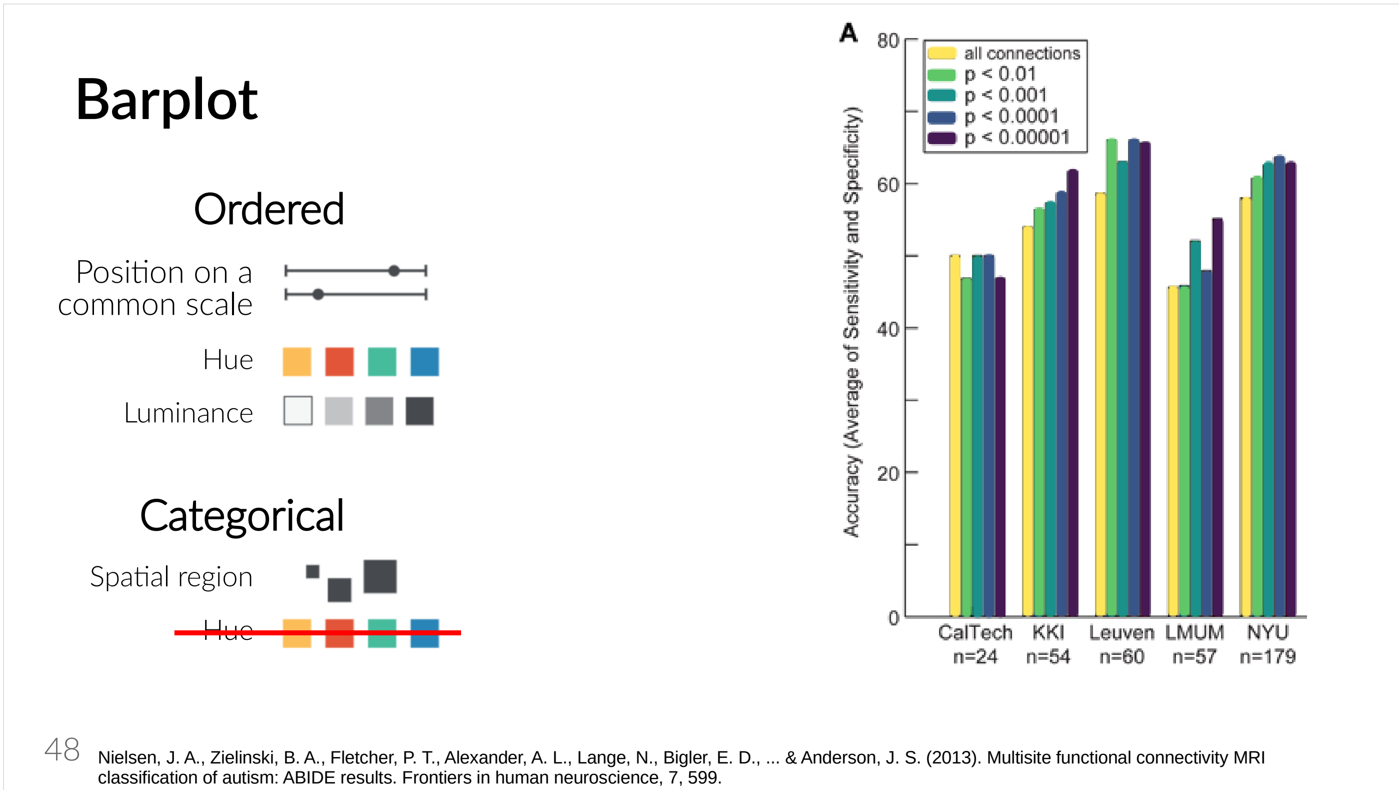
They essentially used a diverging colormap, but the values don't have a logical center value.

Barplot

https://matplotlib.org/cheatsheets/_images/cheatsheets-1.png

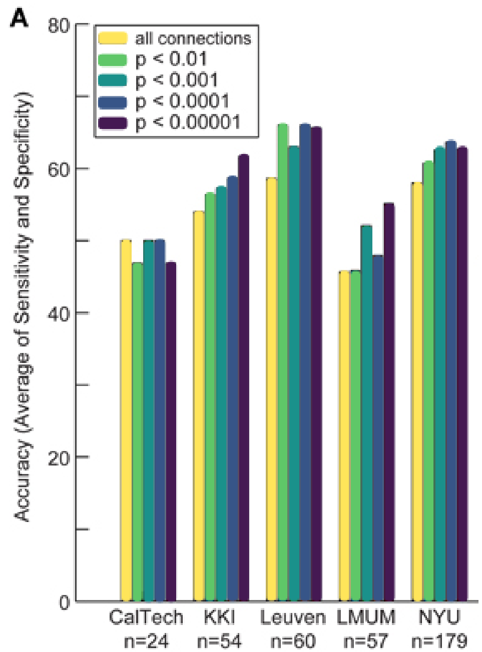
Munzner, T. (2014). Visualization analysis and design. CRC press.

Here is the cheatsheet, and I've marked which elements they used correctly and which could be misleading



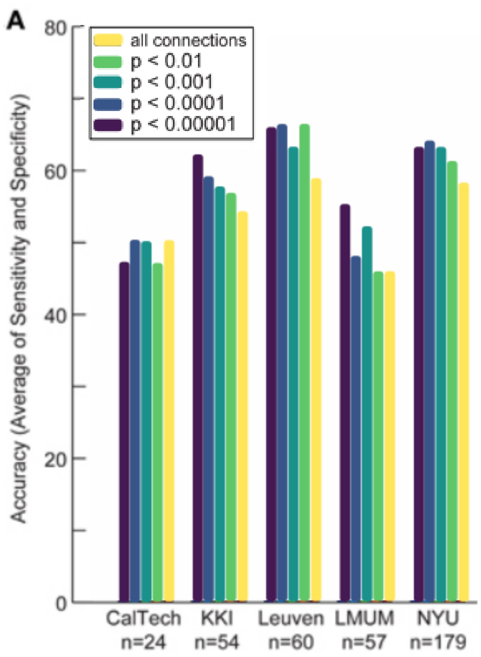
We can improve this by using luminance to encode this variable as an ordered variable.

Let’s focus in on that chart and see what else we can improve.



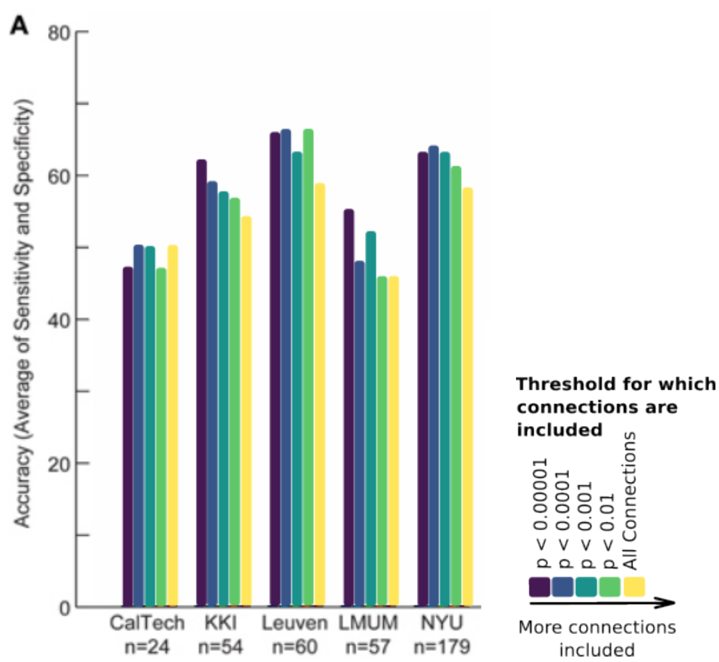
In our current layout, the bars decrease in the number of connections as we move right on the x-axis.

This is opposite of how an axis usually works,
So let’s switch the order



50

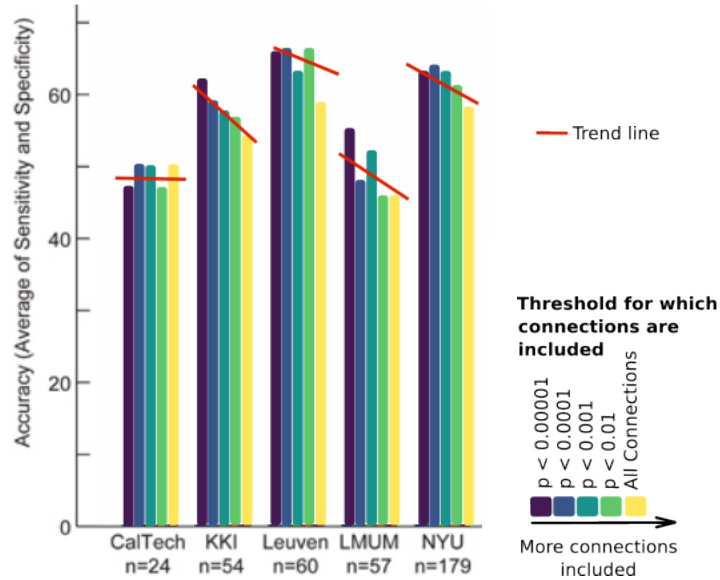
The colors and values are the same, but I've re-ordered the bars so that the number of connections included increases towards the right.



We can make this more obvious by rotating the legend and including more explanatory text.

With things aligned in this way, there's a trend that becomes more obvious: the accuracy tends to decrease as we increase the number of connections.

Across locations, the diagnostic accuracy is fairly consistent around 55%.
The diagnostic accuracy tends to decrease as we include more connections in the connectomes.

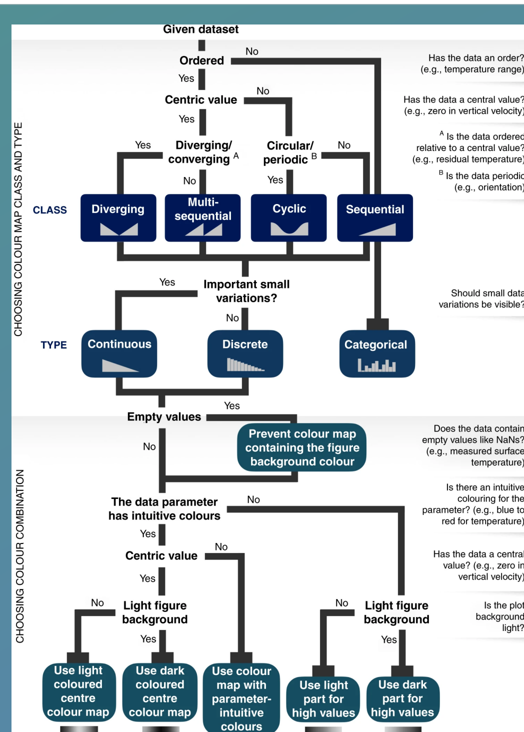


We could make this even more obvious by including trend lines.

Further, we could use the figure title to explicitly state our message.

(For future reference)

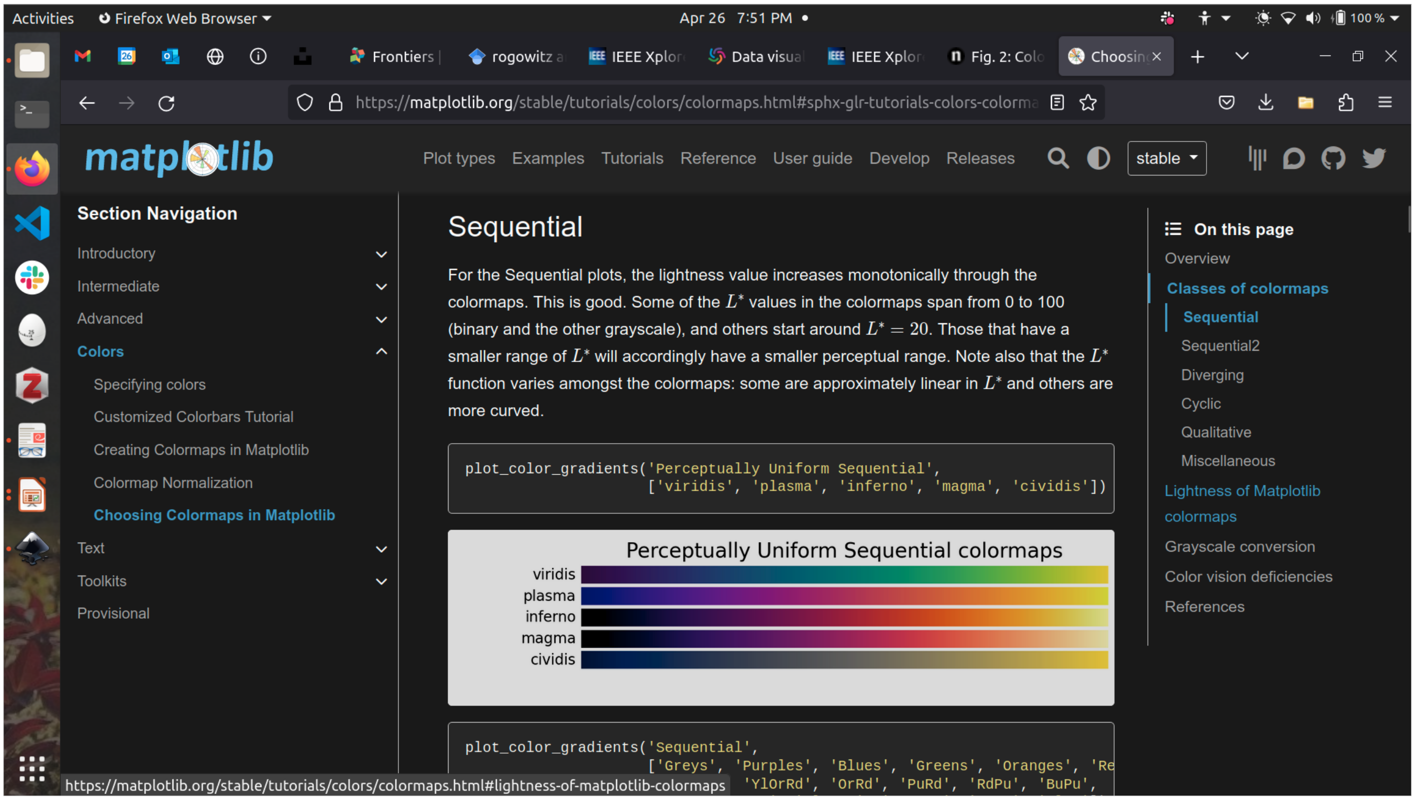
Choosing effective colormaps

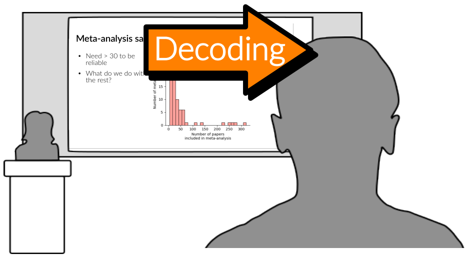


 Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. Nature communications, 11(1), 5444.

Next, I have a cheatsheet for choosing colormaps.

It’s pretty easy to follow the flowchart, so I’ll let you try that when you need to.


Activities
Firefox Web Browser
Apr 26 7:51 PM
100 %
Frontiers |
rogowitz a
IEEE Xplor
Data visua
IEEE Xplor
Fig. 2: Colo
Choosin
https://matplotlib.org/stable/tutorials/colors/colormaps.html#sphx-glr-tutorials-colors-colorma
matplotlib
Plot types
Examples
Tutorials
Reference
User guide
Develop
Releases
stable
Section Navigation
Introductory
Intermediate
Advanced
Colors
Specifying colors
Customized Colorbars Tutorial
Creating Colormaps in Matplotlib
Colormap Normalization
Choosing Colormaps in Matplotlib
Text
Toolkits
Provisional
Sequential
For the Sequential plots, the lightness value increases monotonically through the colormaps. This is good. Some of the L^* values in the colormaps span from 0 to 100 (binary and the other grayscale), and others start around $L^* = 20$. Those that have a smaller range of L^* will accordingly have a smaller perceptual range. Note also that the L^* function varies amongst the colormaps: some are approximately linear in L^* and others are more curved.
plot_color_gradients('Perceptually Uniform Sequential',
['viridis', 'plasma', 'inferno', 'magma', 'cividis'])
Perceptually Uniform Sequential colormaps
viridis
plasma
inferno
magma
cividis
plot_color_gradients('Sequential',
['Greys', 'Purples', 'Blues', 'Greens', 'Oranges', 'Re
'YlOrRd', 'OrRd', 'PuRd', 'RdPu', 'BuPu',
https://matplotlib.org/stable/tutorials/colors/colormaps.html#lightness-of-matplotlib-colormaps
On this page
Overview
Classes of colormaps
Sequential
Sequential2
Diverging
Cyclic
Qualitative
Miscellaneous
Lightness of Matplotlib colormaps
Grayscale conversion
Color vision deficiencies
References



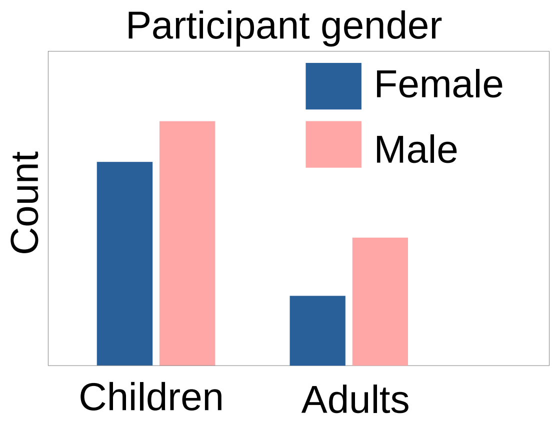
To plan an **effective visualization**, we need to think about

- Message
  - What we want to communicate
- Perception
  - How best to communicate it
- Conventions
  - How it's usually communicated
- Context
  - Where it will be seen

We've just seen that some ways of visualizing data are more effective, based on perceptual evidence.

But we also need to consider conventions that could influence our design choices.

Culture: Color associations



58

As a really basic example, consider cultural color associations.

* in western culture, we associate pink and blue with male and female respectively.

If we make a visualization where this is backwards, it might take more cognitive load for the viewer to understand it.

Science: Rainbow colormaps are disliked

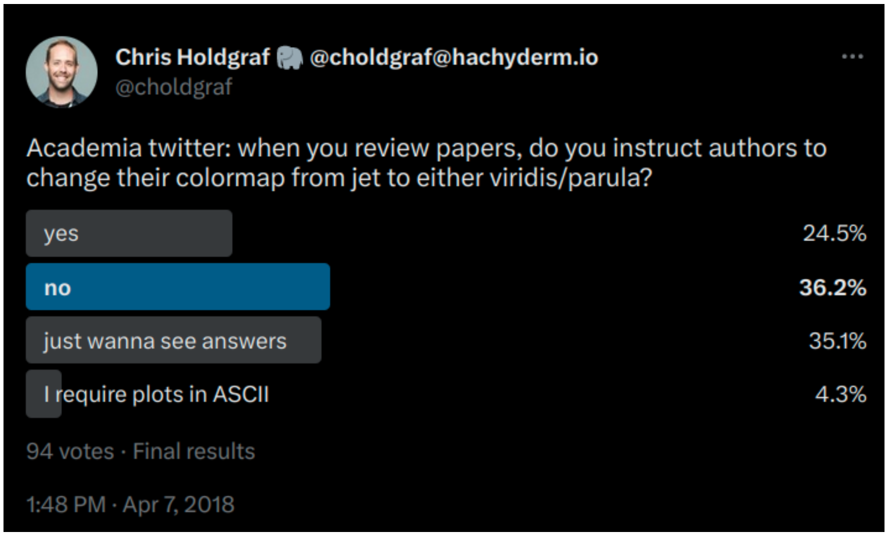
59

We've already talked about rainbow colormaps.

But I just want to emphasise that people have strong opinions about it.

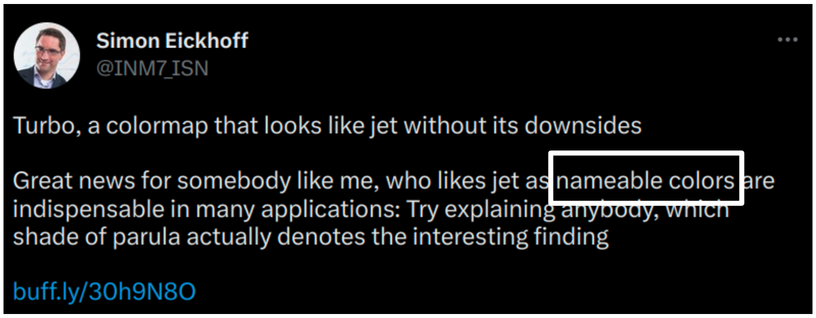
This person says,

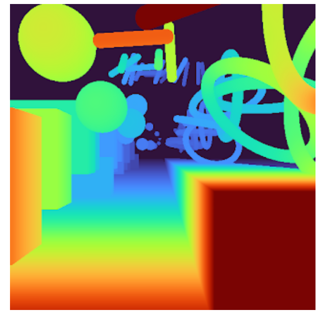
Science: Rainbow colormaps are disliked



About a quarter of these respondants indicated that, if they were reviewing a paper, they would instruct authors to change their colormap from jet to a perceptually-uniform colormap

Science: Rainbow colormaps are disliked?





61

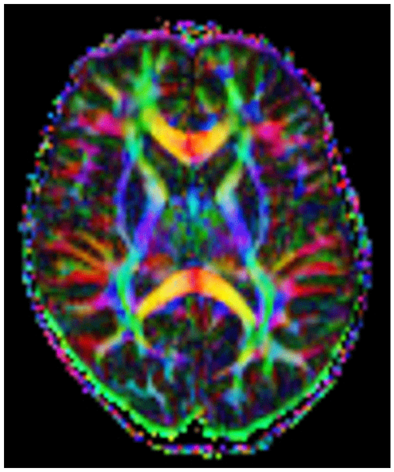
On the other hand, there is a rainbow map that is a perceptually uniform diverging colormap! It’s called turbo.

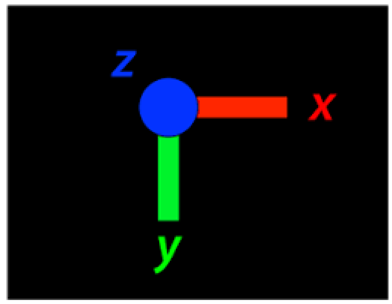
Like this person says, it’s great news for someone who likes having nameable colors.

But I’d be wary because it still looks a lot like jet, and people have such strong opinions about jet.

Neuroscience: Colors = 3D direction

Principal Diffusion Direction





62

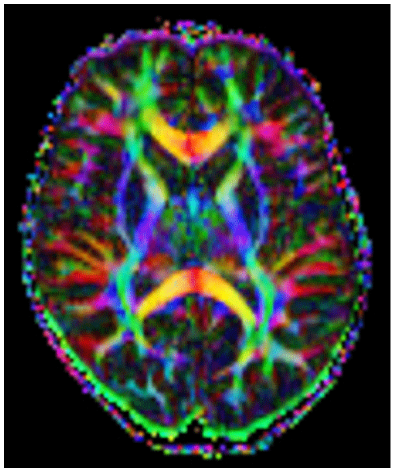
https://mriquestions.com/dti-tensor-imaging.html

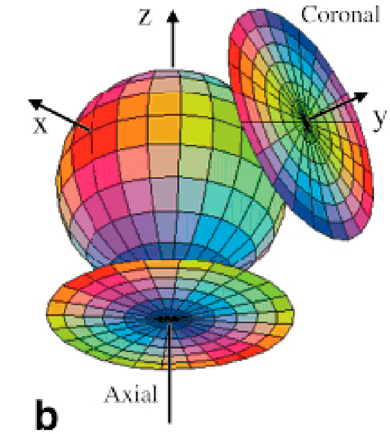
In Neuroscience, we have this convention that colors can represent 3D directions.

This is a map of the principal diffusion direction of water in the brain, which roughly shows white matter tracts.

Neuroscience: Colors = 3D direction

Principal Diffusion Direction





https://mriquestions.com/dti-tensor-imaging.html

https://mipav.cit.nih.gov/pubwiki/index.php/DTI_Color_Display

The mapping of color to direction is not intuitive, and the color choices are arbitrary.

But if everyone visualizes diffusion directions in this way, then people might get good at understanding these maps without too much cognitive effort.

So you might want to use these kinds of visualizations anyways, if you work with this kind of data.

(For future reference)

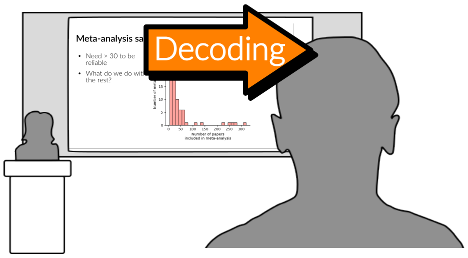
Examples of visualization conventions

- Culture
  - Pink is female
  - Time goes left to right
- Science
  - Rainbow colormaps are disliked
- Neuroscience
  - MRI in grayscale
  - DWI colors = directions

64

Here’s another summary slide for future reference.

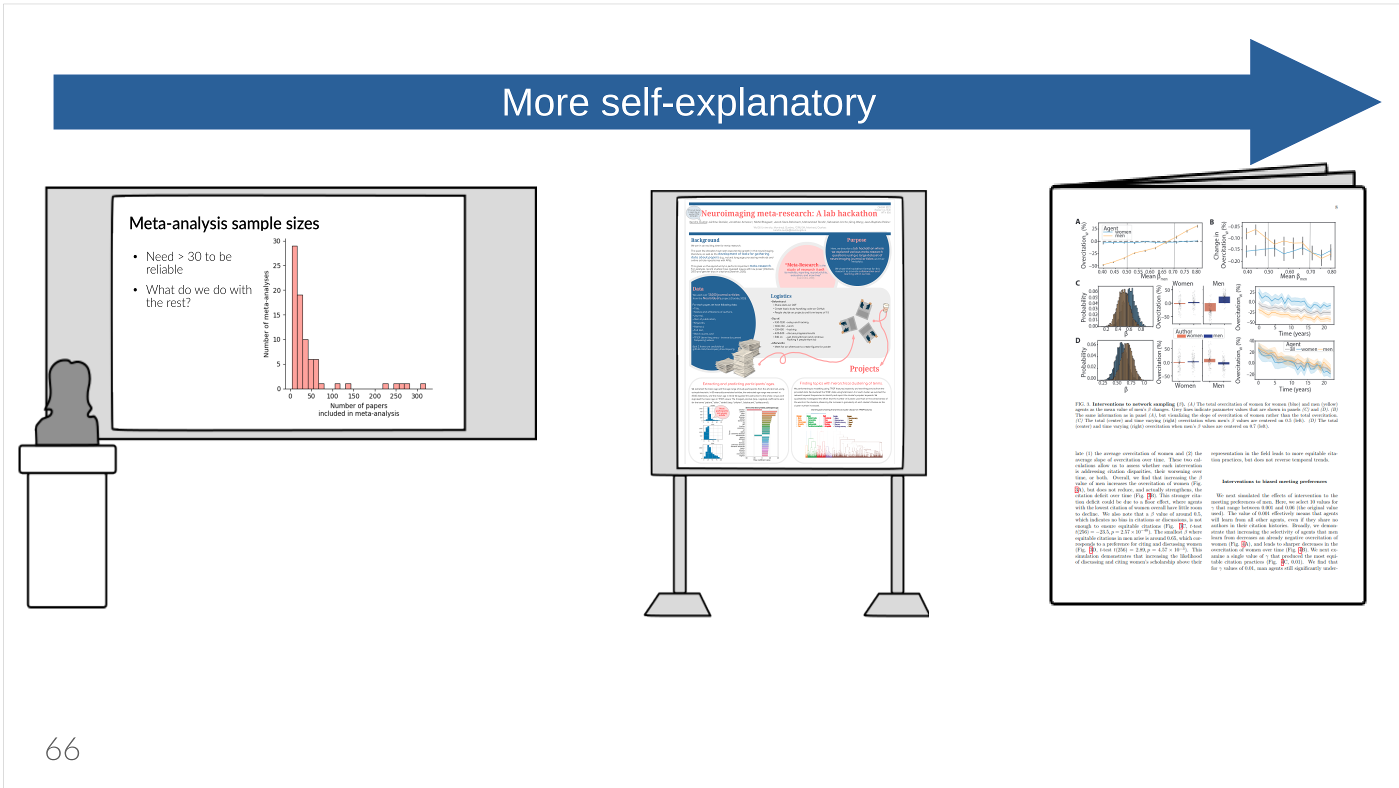
I think this stuff is easy to understand, so we’ll move on.



To plan an **effective visualization**, we need to think about

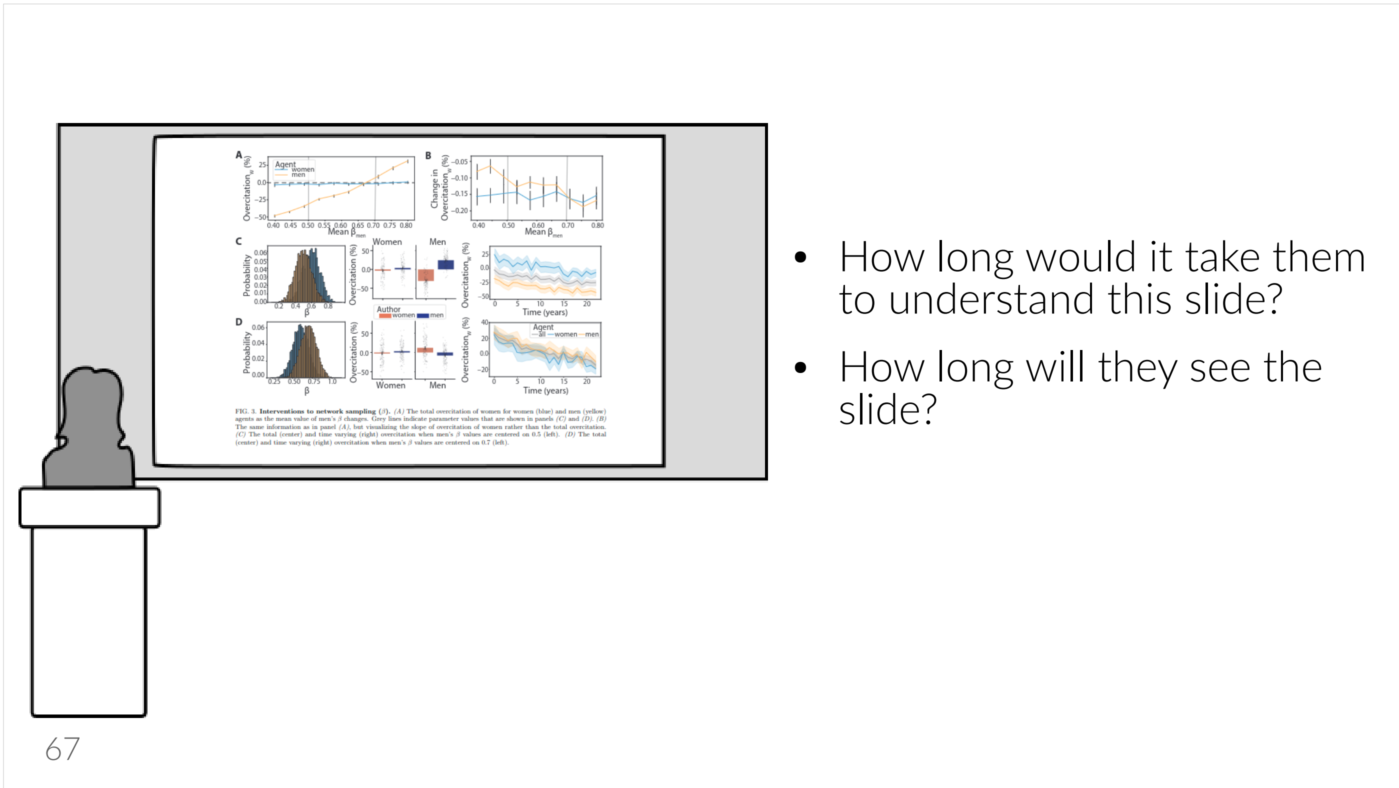
- Message
  - **What** we want to communicate
- Perception
  - **How best** to communicate it
- Conventions
  - How it's **usually** communicated
- Context
  - **Where** it will be seen

Finally, we'll touch on how to adjust visualizations according to where people will see it.



There are basically three places we see data visualizations: in presentations, on posters, and in papers

* I'd argue that these have a rough order of how self-explanatory they need to be.



For example, a presentation slide does not need to be self-explanatory.

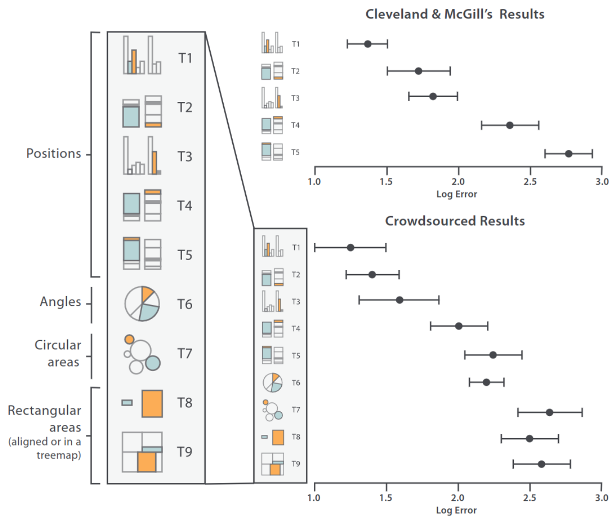
In fact, it should be very simple.

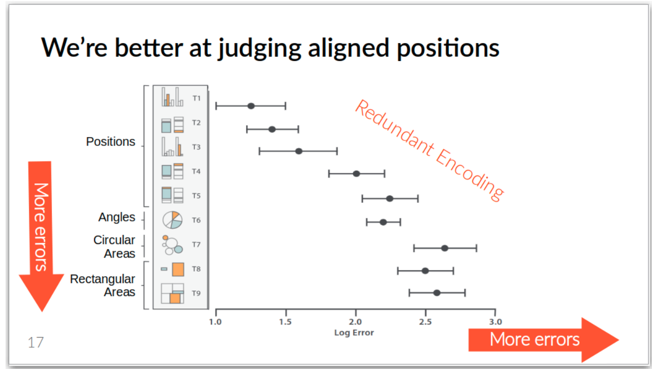
You should ask yourself.... and ...

And if they mismatch, you should simplify the visualization.

E.g., Figure in textbook

My slide



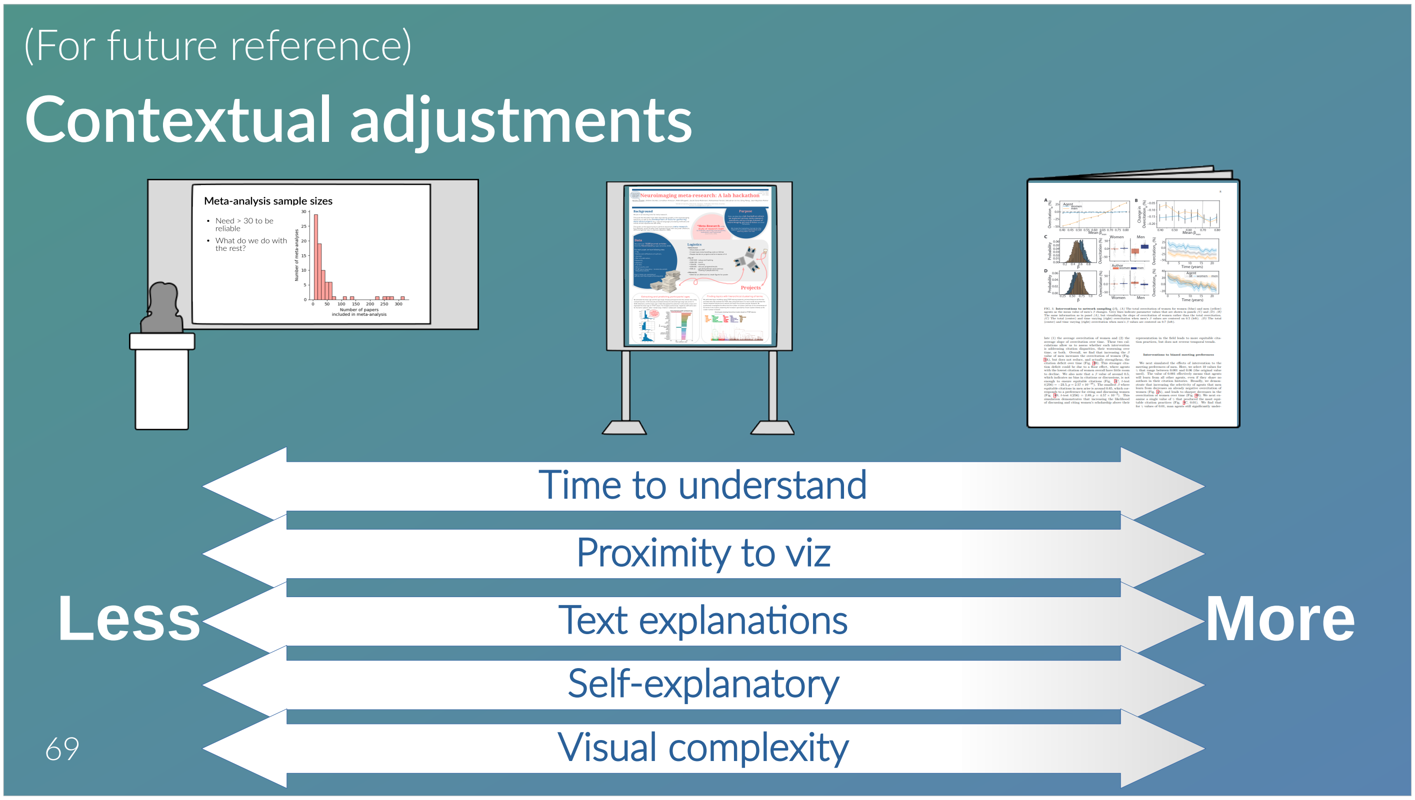


68

Munzner, T. (2014). Visualization analysis and design. CRC press.

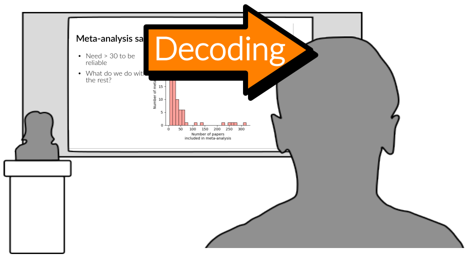
For example, here's a figure from a textbook,
And here's the way I used it in a slide.

I used only part of it to make it simpler,
I increased the font
And I added annotations to help people understand it quickly.



Basically, you should adjust figures according to how much time people will have to understand it, and how close they'll be to the figure.

This will inform how much text explanations there should be, how self-explanatory the entire thing should be, and how visually complex it should be.



To plan an **effective visualization**, we need to think about

- Message
  - **What** we want to communicate
- Perception
  - **How best** to communicate it
- Conventions
  - How it's **usually** communicated
- Context
  - **Where** it will be seen

Now that we've discussed all these principles, we'll plan a simple figure to drive these points home.

Example 1 – Simple data and figure

- Original paper on the ABIDE dataset
  - 964 subjects
    - 396 male
    - 51 female

Lets visualize this

71 Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., ... & Anderson, J. S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in human neuroscience, 7, 599.

We'll use a tiny bit of data from the ABIDE dataset, which I mentioned before.

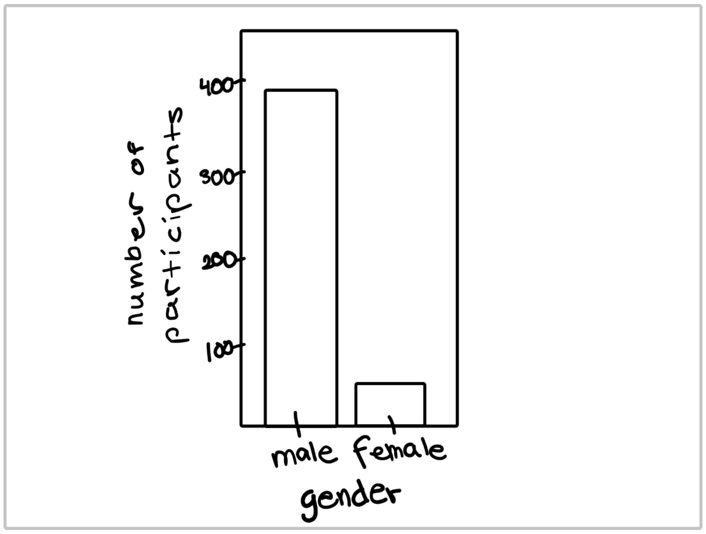
We'll simply use the number of male and female participants

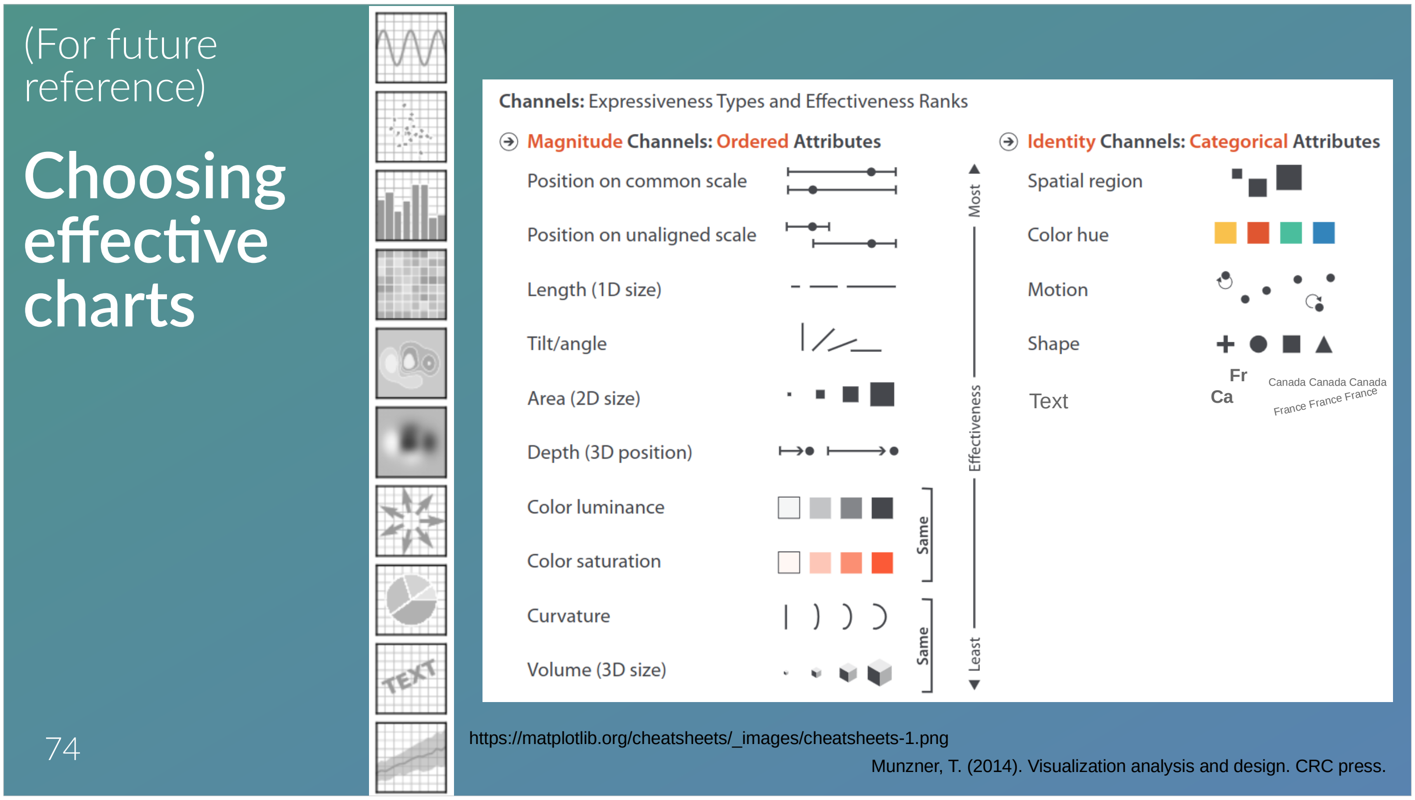
What's our message?

- More male than female participants
  - 2 categories, 2 values

How should we communicate this message?

- Clearly show 2 categories and 2 different values





Here's the cheatsheet I made for choosing effective charts.

It's a busy slide because I wanted you to be able to see and compare all of the info.

On the right, we see a summary table showing which channels are most effective for expressing magnitides and identities.

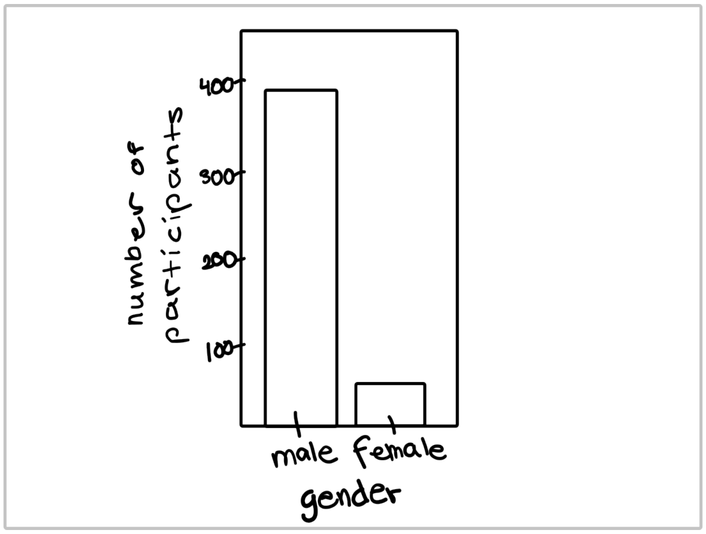
For example, it's easier to tell that one value is twice another if they are represented as dots on the a common scale, as opposed to having different color saturation.

On the left we have different kinds of charts.

Now we'll go through some of them to demo how to use this cheat sheet.

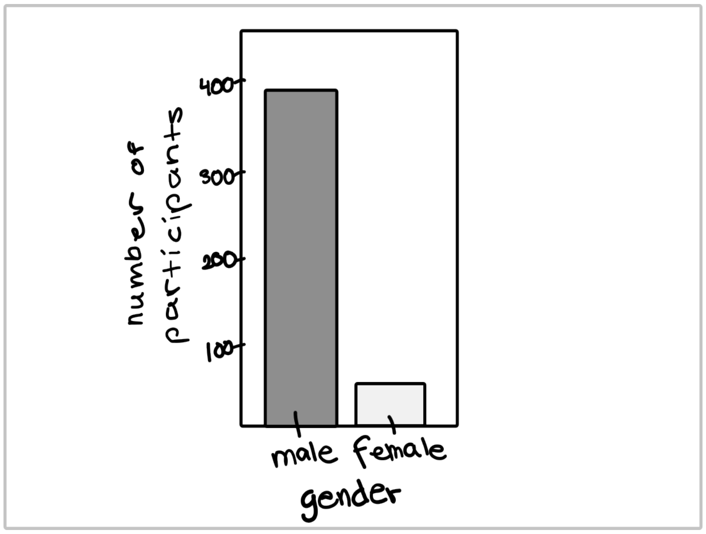
How should we communicate this message?

- Clearly show 2 categories and 2 different values



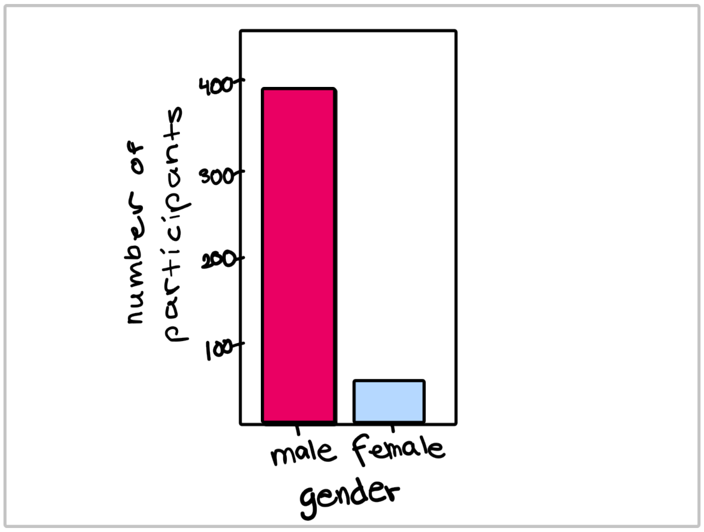
How should we communicate this message?

- Clearly show 2 categories and 2 different **values**



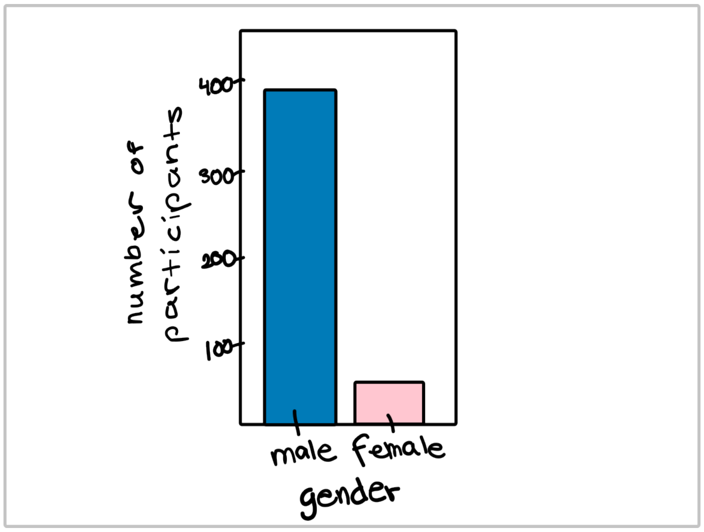
How should we communicate this message?

- Clearly show 2 **categories** and 2 different values



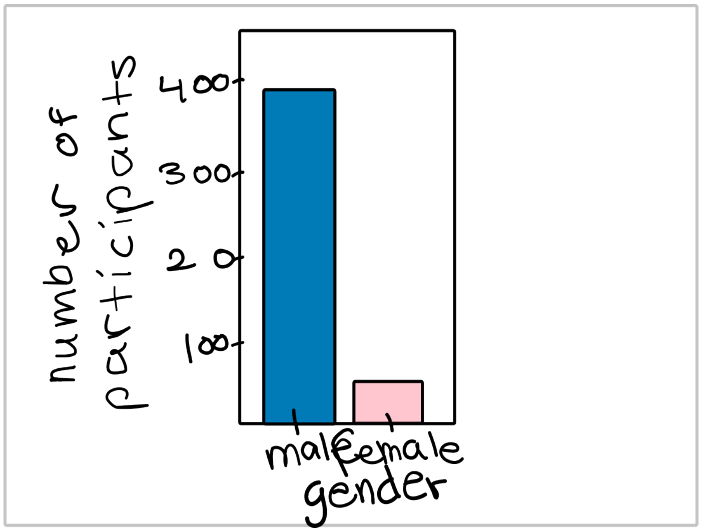
Are there any conventions we should think about?

- Gender & color



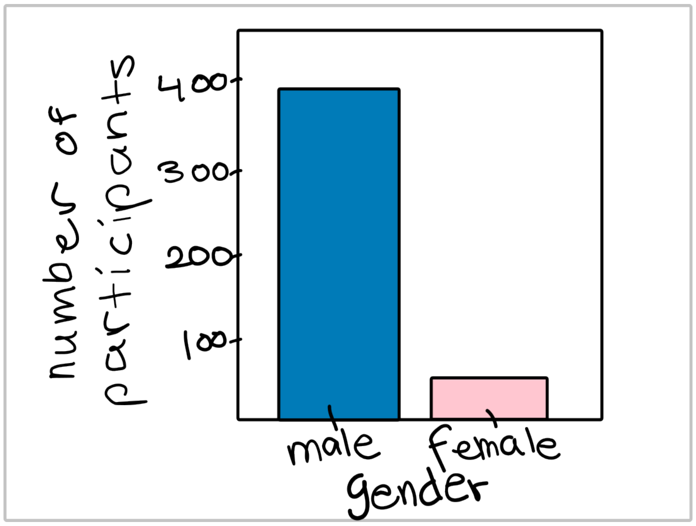
Do we need to adapt it to the context?

- Academic presentation



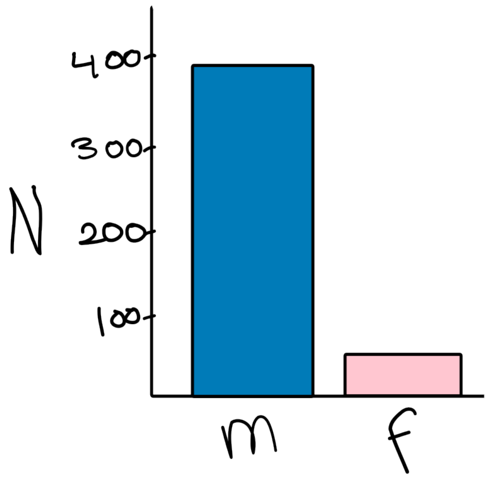
Do we need to adapt it to the context?

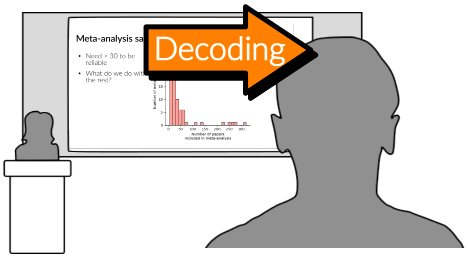
- Academic presentation



Do we need to adapt it to the context?

- Academic presentation





To plan an **effective visualization**, we need to think about

- Message
  - What we want to communicate
- Perception
  - How best to communicate it
- Conventions
  - How it's usually communicated
- Context
  - Where it will be seen

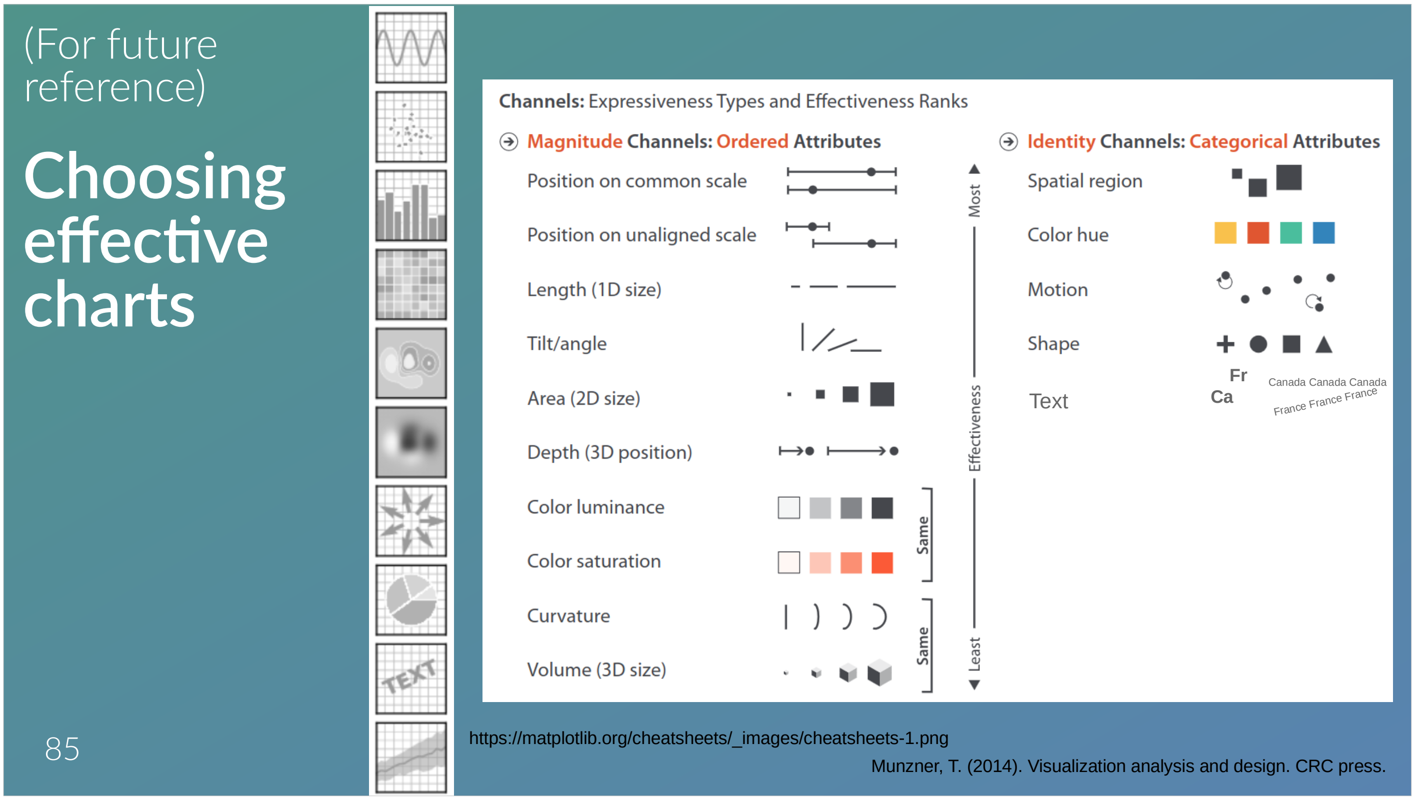
Now we've finished part 1 of intro to data visualization, where what we need to consider about how people decode visualizations.

We need to know what we want to communicate,
How best to communicate it,
How it's usually communicated,
And the context in which it will be received.

In the next lecture, we'll discuss how to encode figures that follow these principles

The end of part 1

Reference slides...



Here's the cheatsheet I made for choosing effective charts.

It's a busy slide because I wanted you to be able to see and compare all of the info.

On the right, we see a summary table showing which channels are most effective for expressing magnitides and identities.

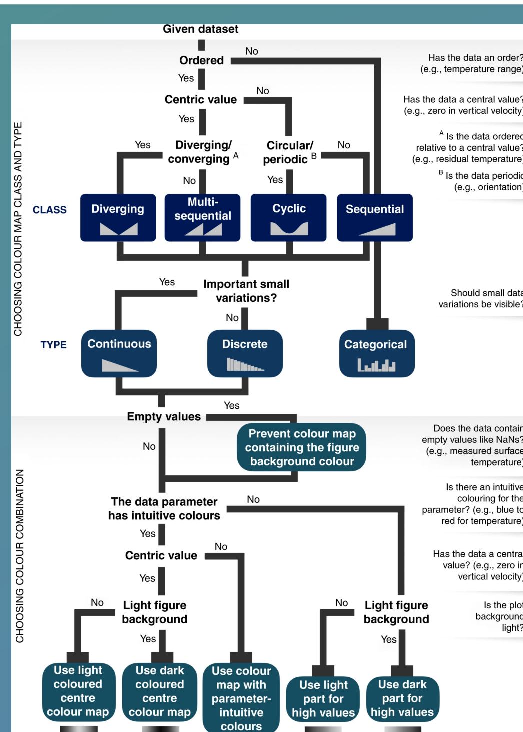
For example, it's easier to tell that one value is twice another if they are represented as dots on the a common scale, as opposed to having different color saturation.

On the left we have different kinds of charts.

Now we'll go through some of them to demo how to use this cheat sheet.

(For future reference)

Choosing effective colormaps



Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. Nature communications, 11(1), 5444.

(For future reference)

Examples of visualization conventions

- Culture
  - Pink is female
  - Time goes left to right

- Science
  - Rainbow colormaps are disliked

- Neuroscience
  - MRI in grayscale
  - DWI colors = directions

(For future reference)

Contextual adjustments