ESLR Workshop

Data Visualisation Pipeline

slides etc: bit.ly/eslr-vis

Dr Chrissy Cuskley (christine.cuskley@ncl.ac.uk) | ccuskley.github.io | @nerdpro

Aims

To build descriptive visual narrative into your analysis pipeline using R.

What we *won't* do

Deal with more complex (and common) analyses, e.g., modelling, other inferential statistics

However...

Visualisation can play an important role in getting inferential analyses right.

See e.g., https://drsimonj.svbtle.com/visualising-residuals

Roadmap

where we'll go

1. Visualising (null) hypotheses

   1.1 Sketching

   1.2 Using real data: Understanding the data; shaping, summarizing, and visualizing

   1.3 More complex measures

   1.4 Comparing expectations and results

Roadmap

how we'll get there

Alternate between slides and exercises; codealong?

All the exercises are in a single RMarkdown (.rmd) file within the project DatavisPipeline

Start at the top and work your way down, reading through the inline instructions **within the .rmd file**

DataVisExercise_complete.html (you can also look into the .rmd file of the same name)

RMarkdown

https://r4ds.had.co.nz/r-markdown.html

"R Markdown provides an unified authoring framework for data science, combining your code, its results, and your prose commentary. R Markdown documents are fully reproducible and support dozens of output formats, like PDFs, Word files, slideshows, and more."

Communication, collaboration, and an environment for doing data science.

1.1 Sketching

You should (obviously) **decide what your measures *are* before you collect data**

You should also think about what your measures will *look like* before data collection.

What does this mean?

Doodle your (null) hypotheses.

I’ll describe an experiment in very basic terms, you’ll sketch your hypotheses.

Cumulative cultural evolution?

(i) a change in behaviour (or product of behaviour, such as an artefact), typically due to asocial learning, followed by

(ii) the transfer via social learning of that novel or modified behaviour to other individuals or groups, where

(iii) the learned behaviour causes an improvement in performance, which is a proxy of genetic and/or cultural fitness, with

(iv) the previous three steps repeated in a manner that generates sequential improvement over time

Mesoudi & Thornton (2018) What is cumulative cultural evolution? *Proc. Royal Soc. B.*, https://doi.org/10.1098/rspb.2018.0712

Cumulative cultural evolution?

(i) a change in behaviour (or product of behaviour, such as an artefact), typically due to asocial learning, followed by

(ii) the transfer via social learning of that novel or modified behaviour to other individuals or groups, where

(iii) the learned behaviour causes an improvement in performance, which is a proxy of genetic and/or cultural fitness, with

(iv) the previous three steps repeated in a manner that generates sequential improvement over time

Mesoudi & Thornton (2018) What is cumulative cultural evolution? *Proc. Royal Soc. B.*, https://doi.org/10.1098/rspb.2018.0712

Cumulative cultural evolution?

(i) a change in behaviour (or product of behaviour, such as an artefact), typically due to asocial learning, followed by

(ii) the transfer via social learning of that novel or modified behaviour to other individuals or groups, where

(iii) the learned behaviour causes an improvement in performance, which is a proxy of genetic and/or cultural fitness, with

(iv) the previous three steps repeated in a manner that generates sequential improvement over time

Mesoudi & Thornton (2018) What is cumulative cultural evolution? *Proc. Royal Soc. B.*, https://doi.org/10.1098/rspb.2018.0712

Research question

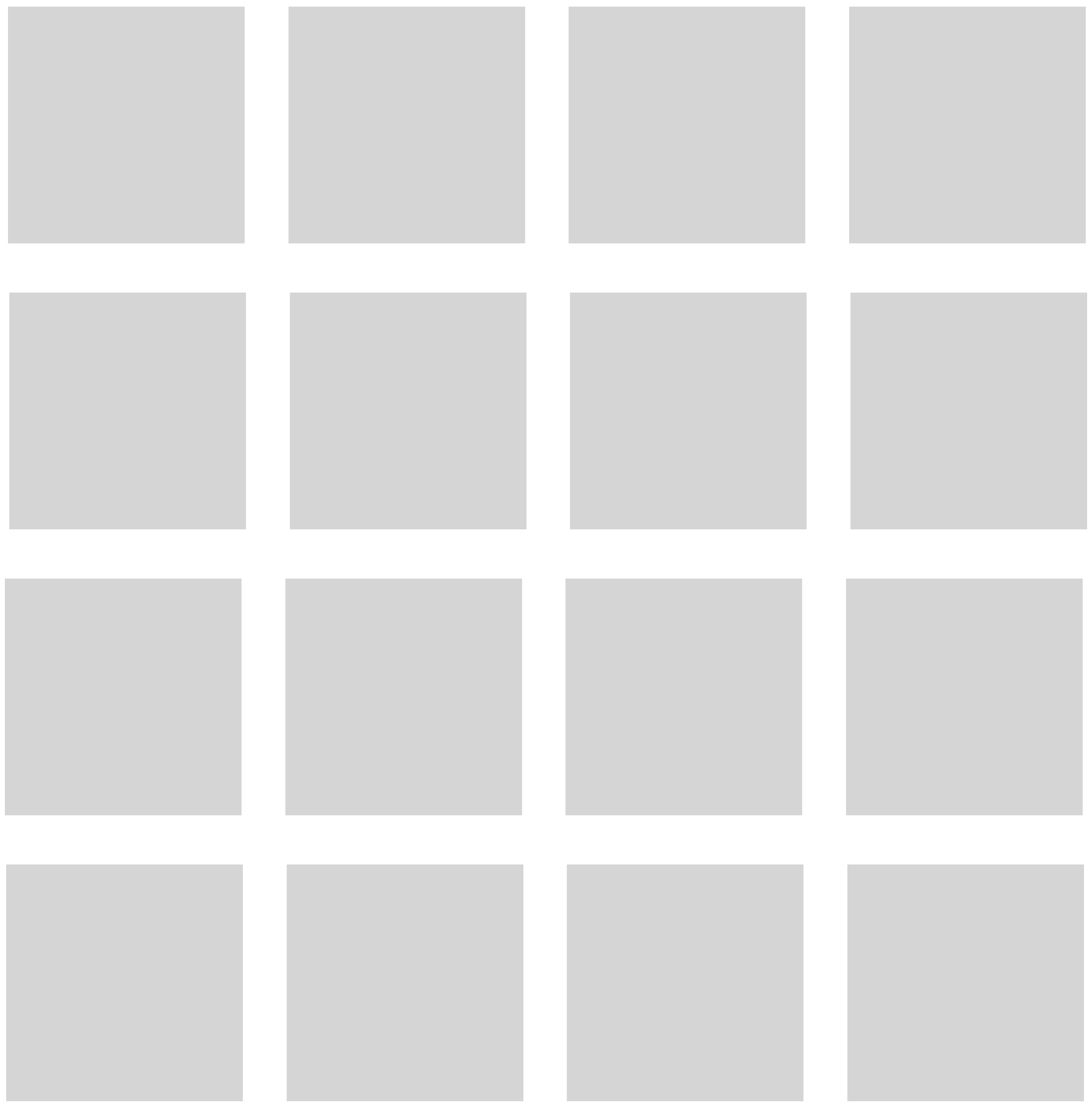
Is copying essential for cumulative cultural evolution?

Saldana et al. (2019). *Proc. R. Soc. B,* 286: 20190729. http://dx.doi.org/10.1098/rspb.2019.0729

Test task

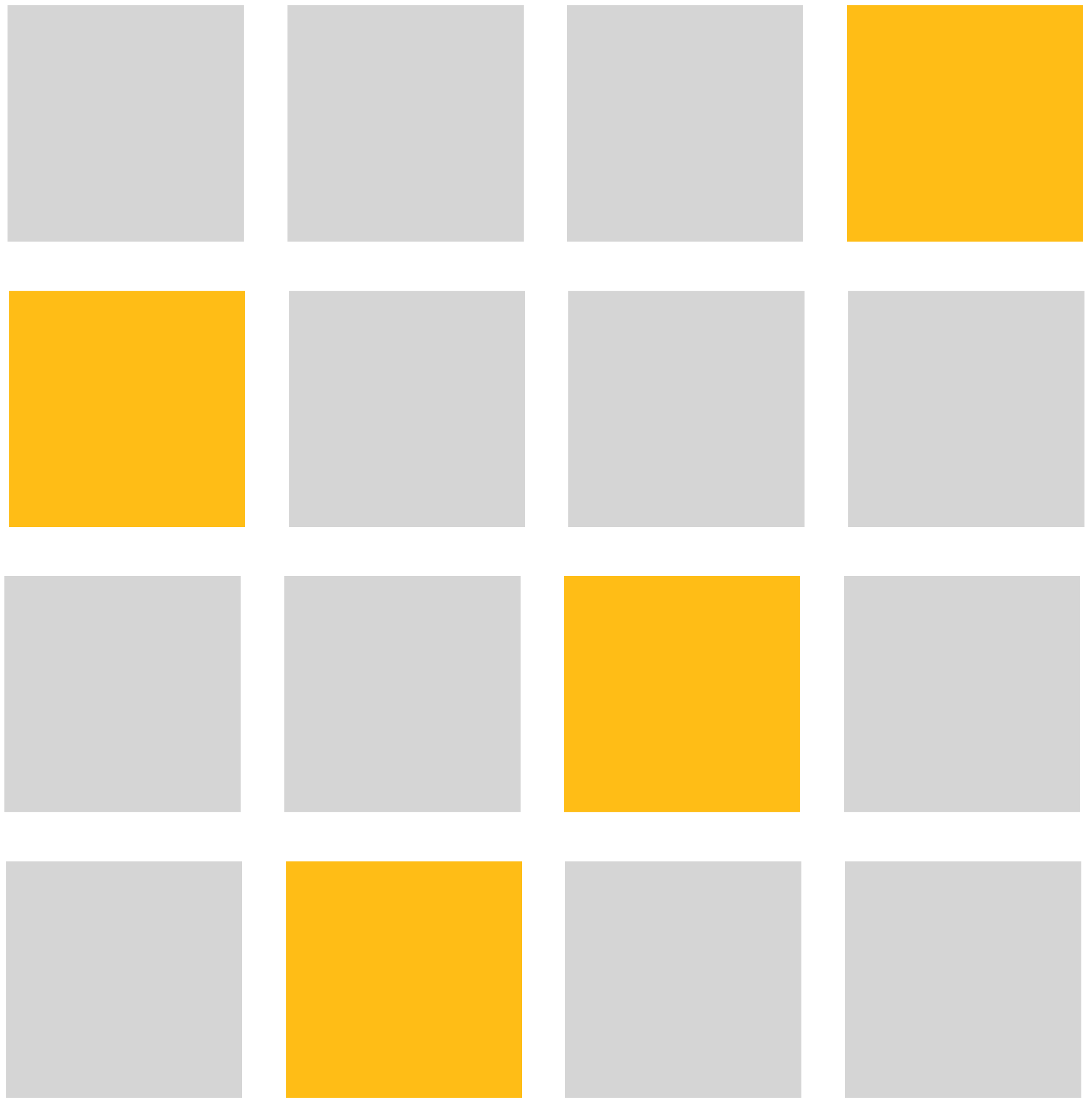
A simple grid task where the remit is to innovate, rather than copy

Cultural transmission & grid copying



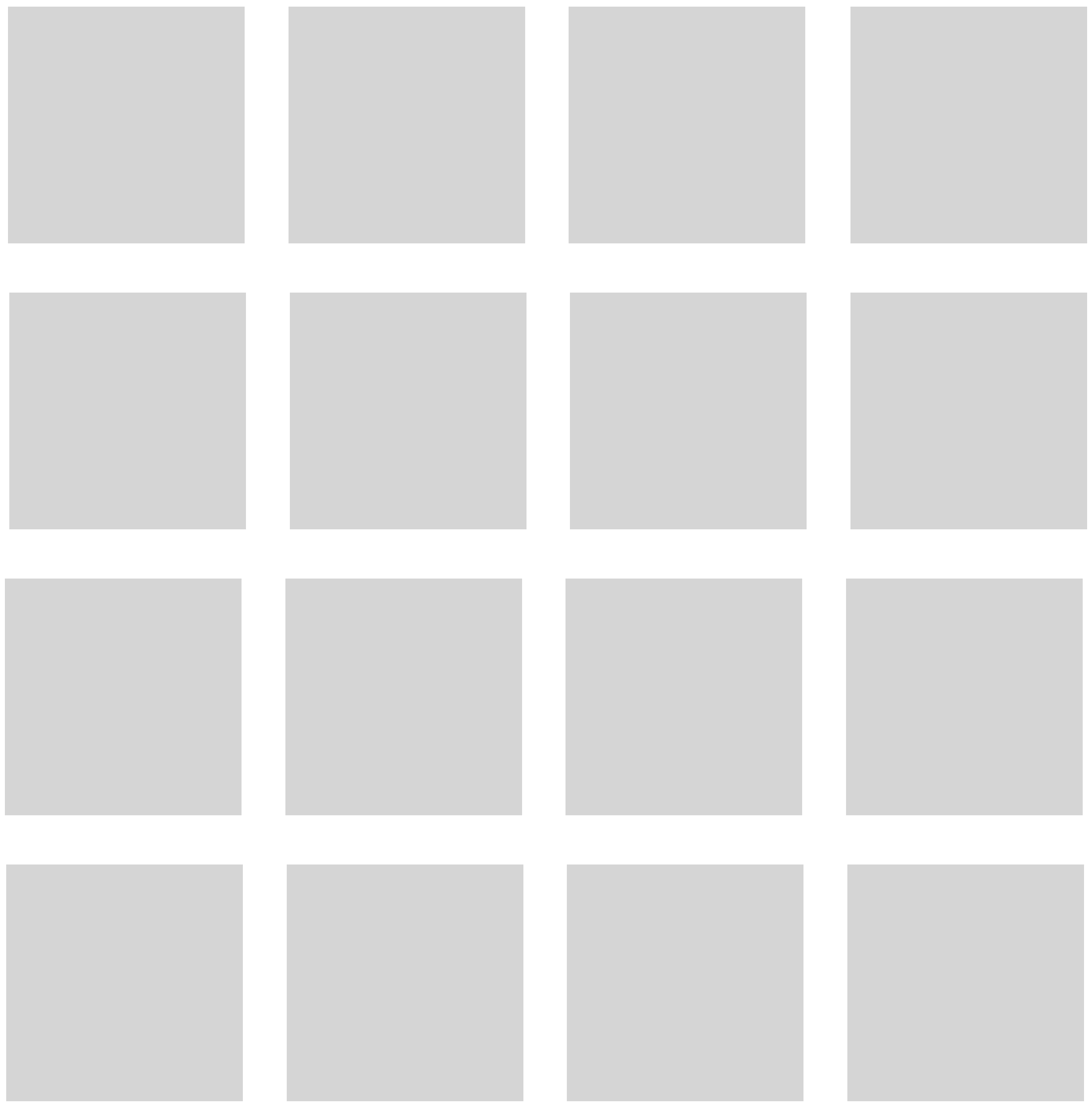
Cultural transmission & grid copying

On each trial, four cells light up.



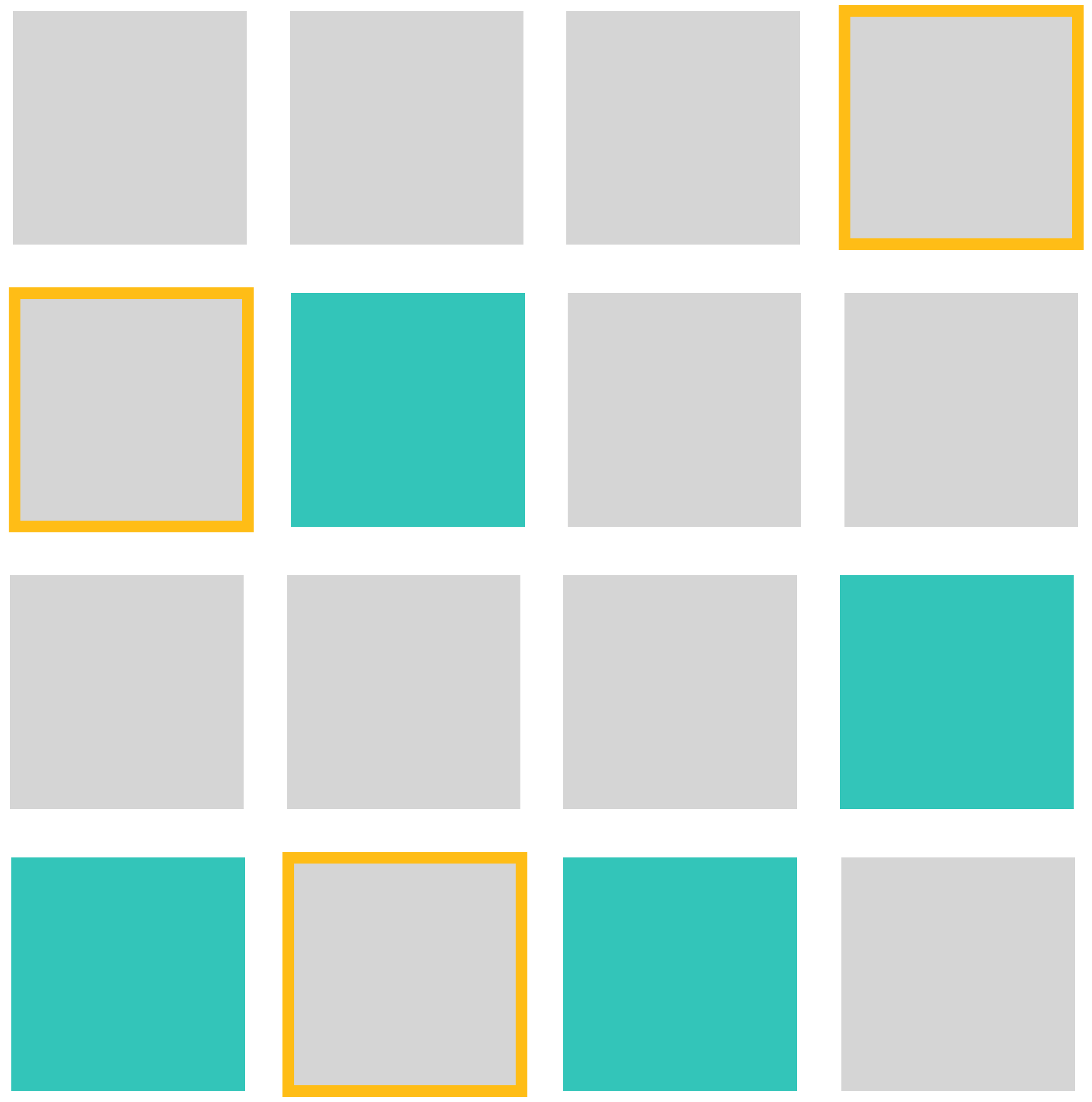
Cultural transmission & grid copying

And then disappear after a short interval.



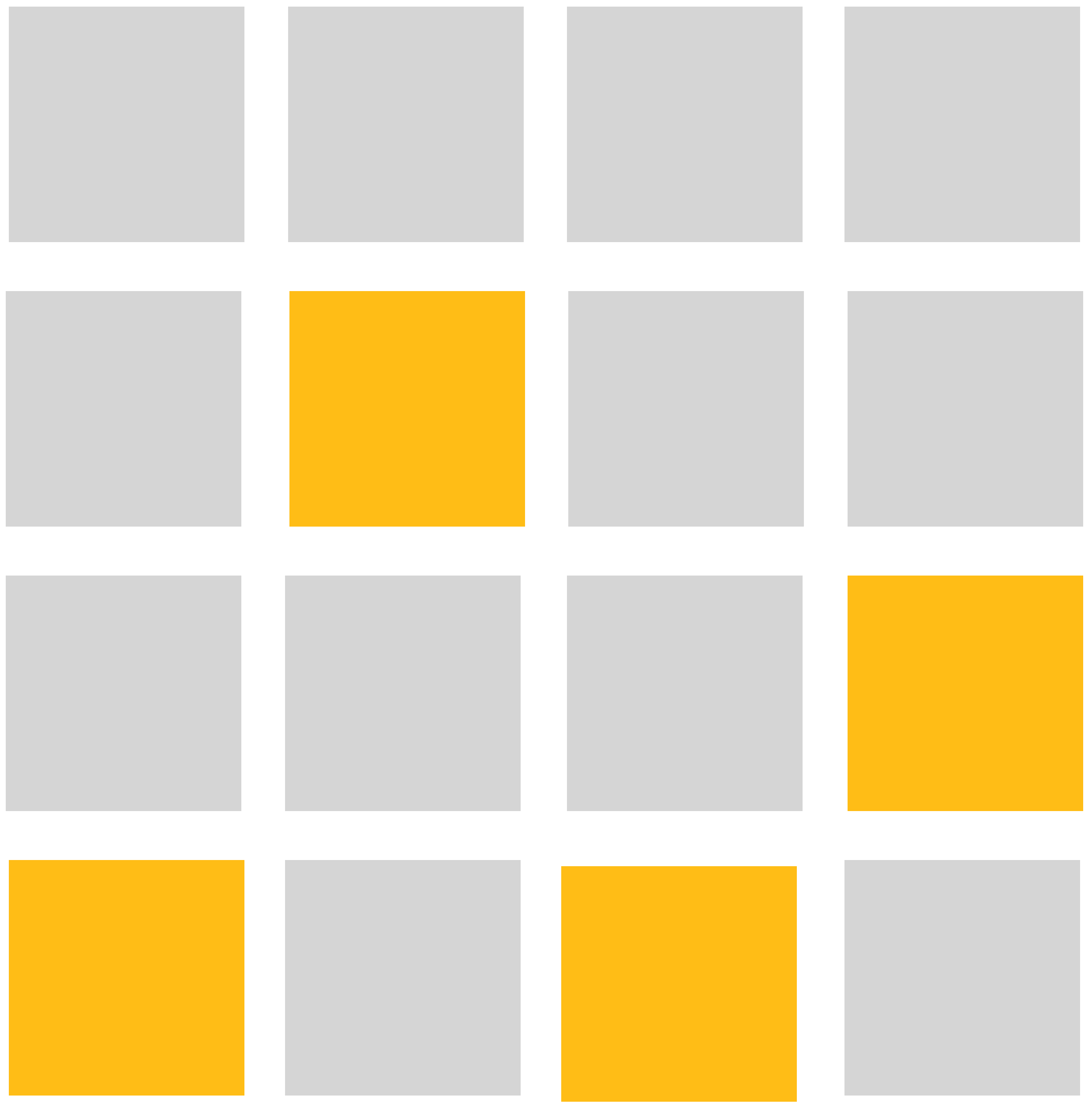
Cultural transmission & grid copying

Choose any four that were **not** lit up in the prompt for a reward.



Cultural transmission & grid copying

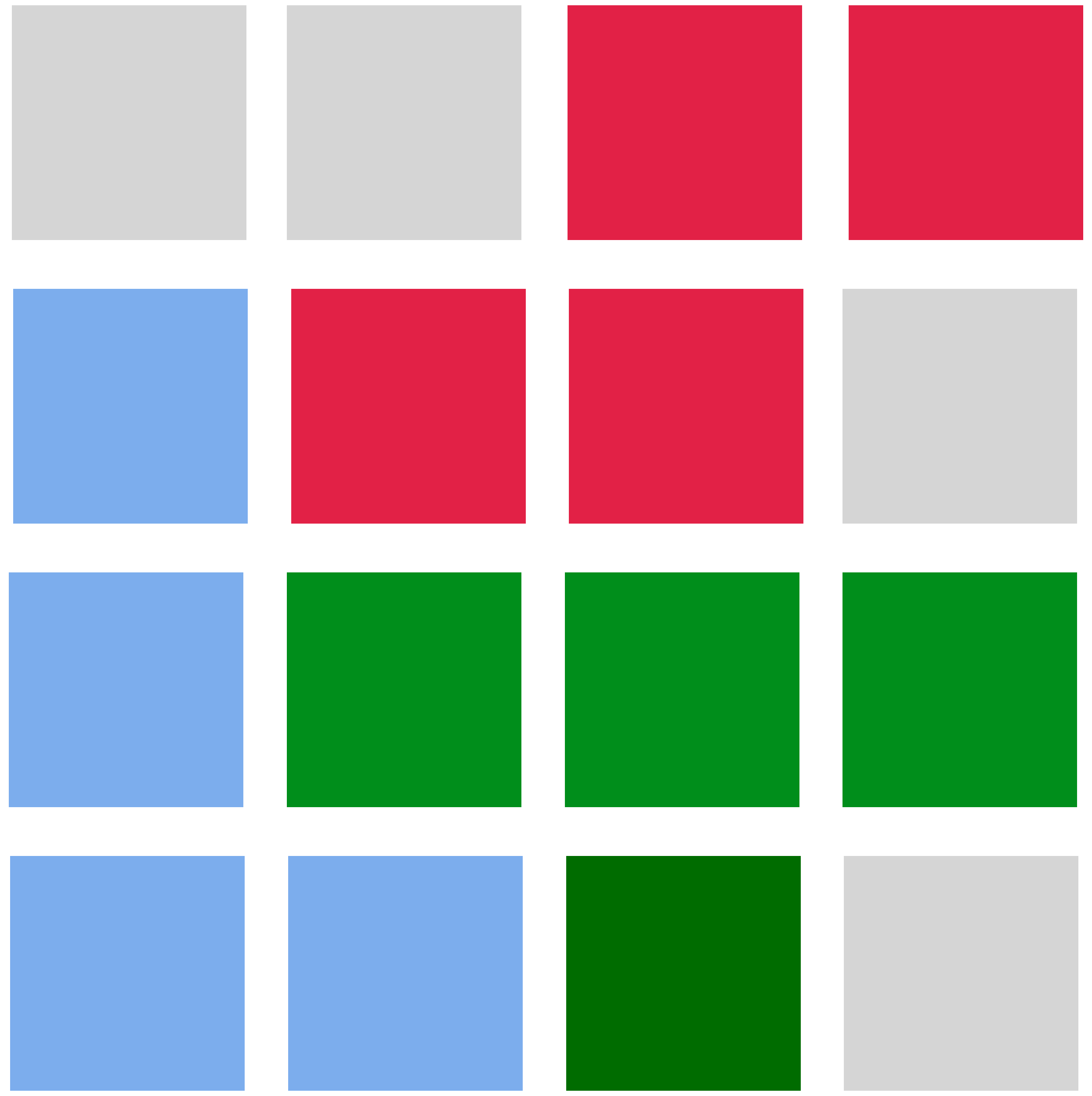
What participant *n* chooses is scored and given to participant *n*+1



Cultural transmission & grid copying

Measuring

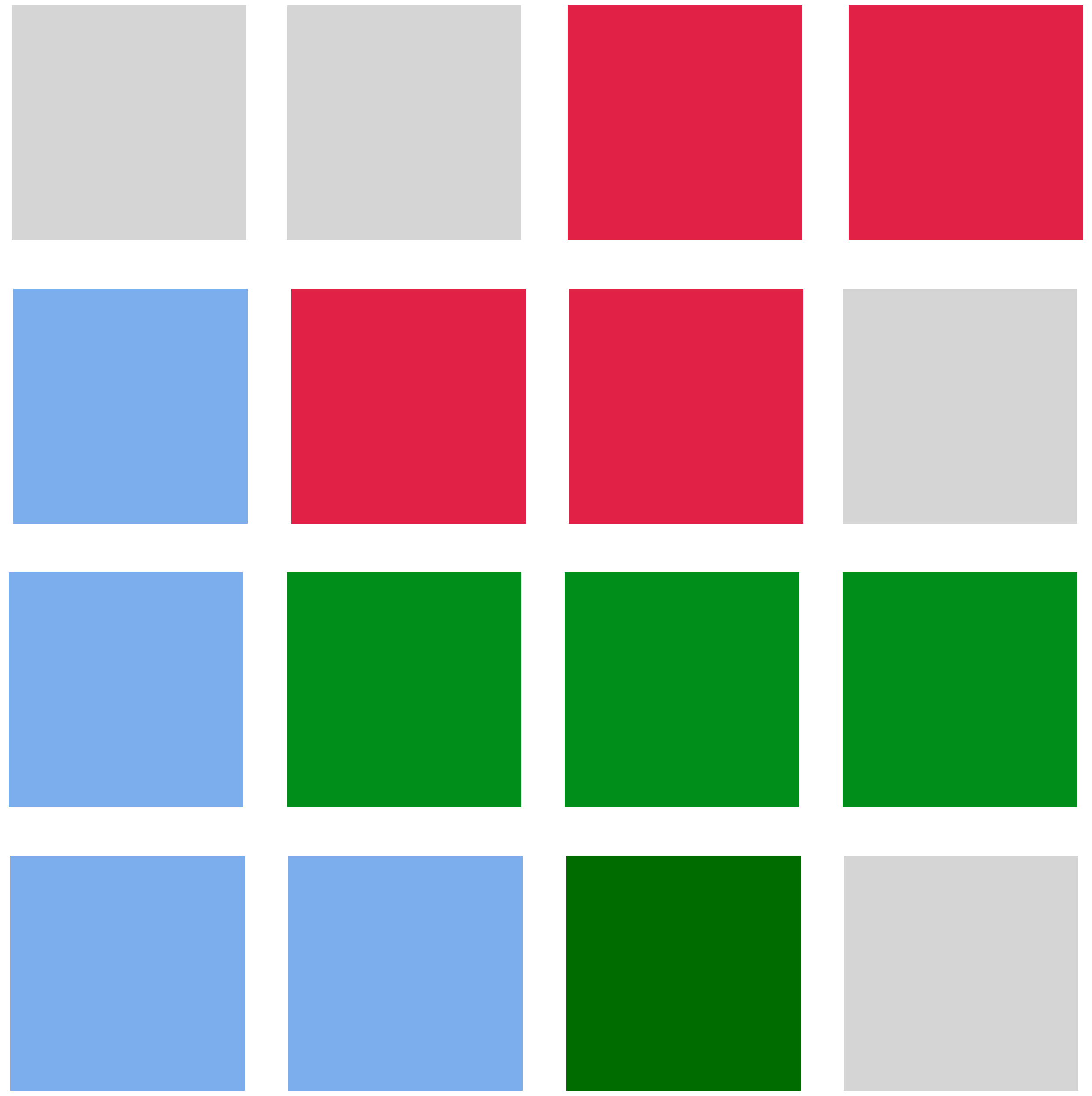
- Performance
- Predictability/Entropy of responses
- Proportion of tetrominoes (Cladière et al., 2014)



Cultural transmission & grid copying

What will results look like over "time" if…

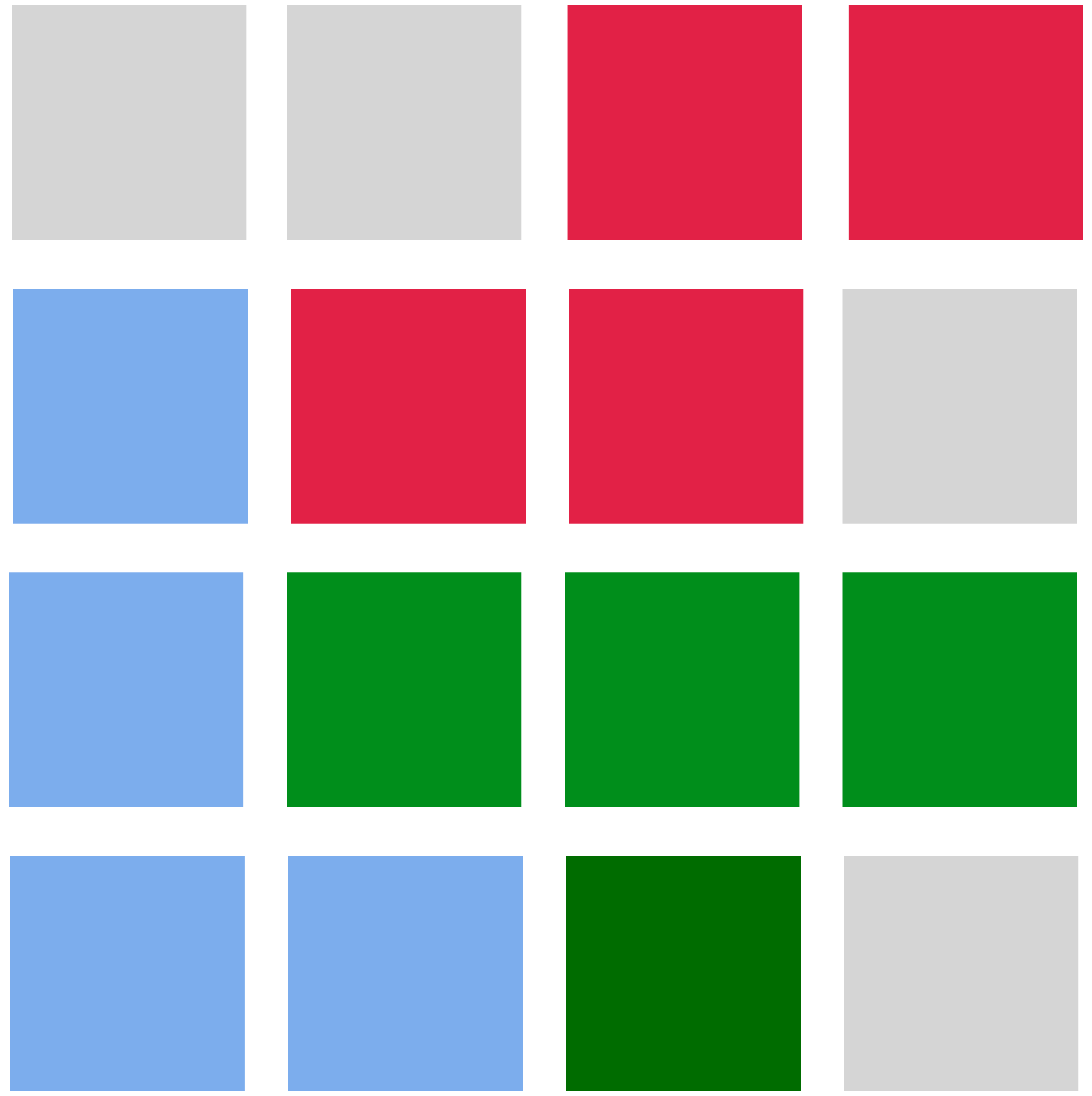
- Performance is random?
- Responses are completely unpredictable?
- No one cares about tetrominoes?



Cultural transmission & grid copying

Versus if…

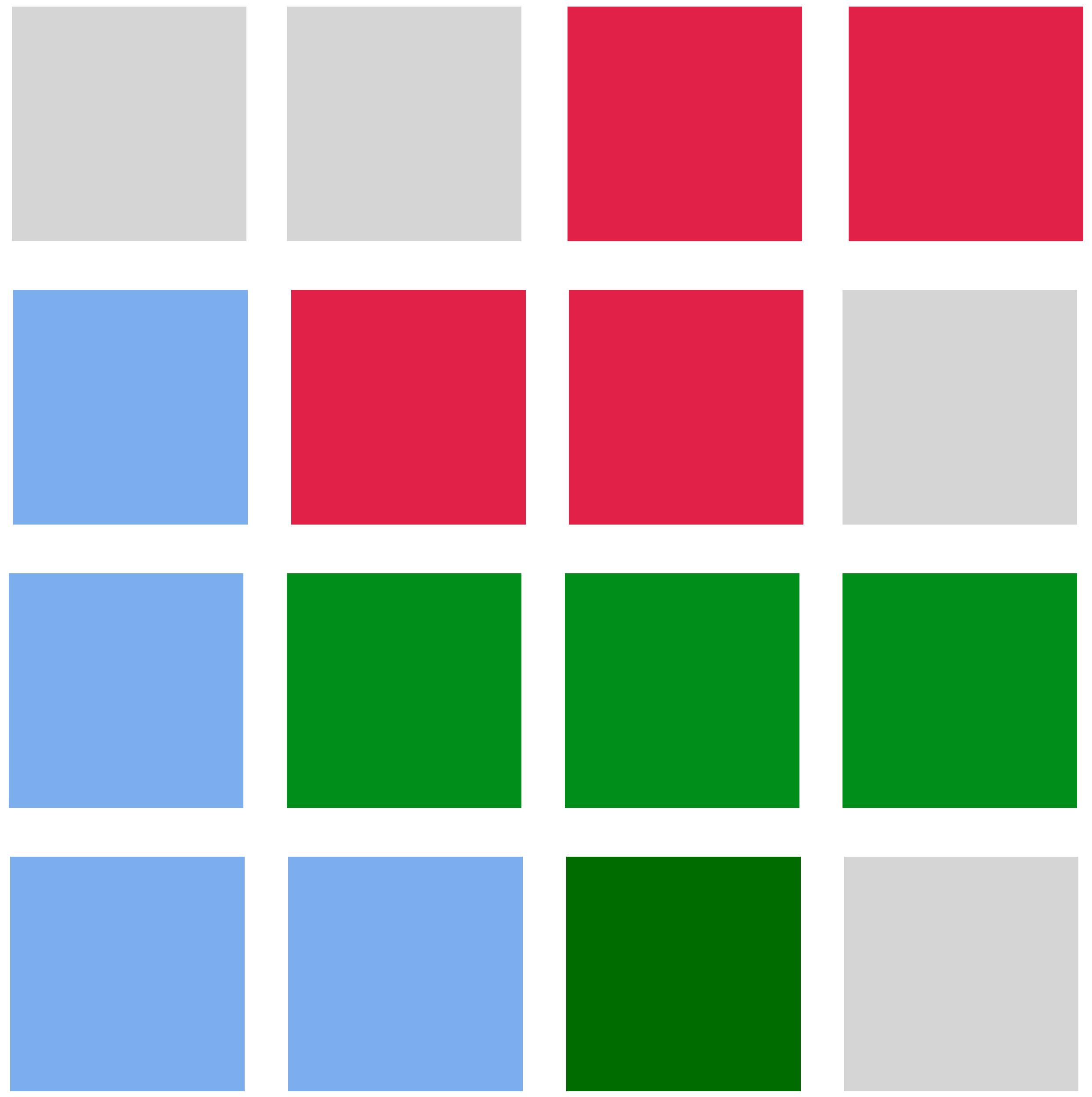
- Performance improves over “time”?
- Responses are highly predictable?
- Tetrominoes are an attractor?



Cultural transmission & grid copying

Would you expect different performance from…

- Children
- Other primates
- Adults?
- Other species?

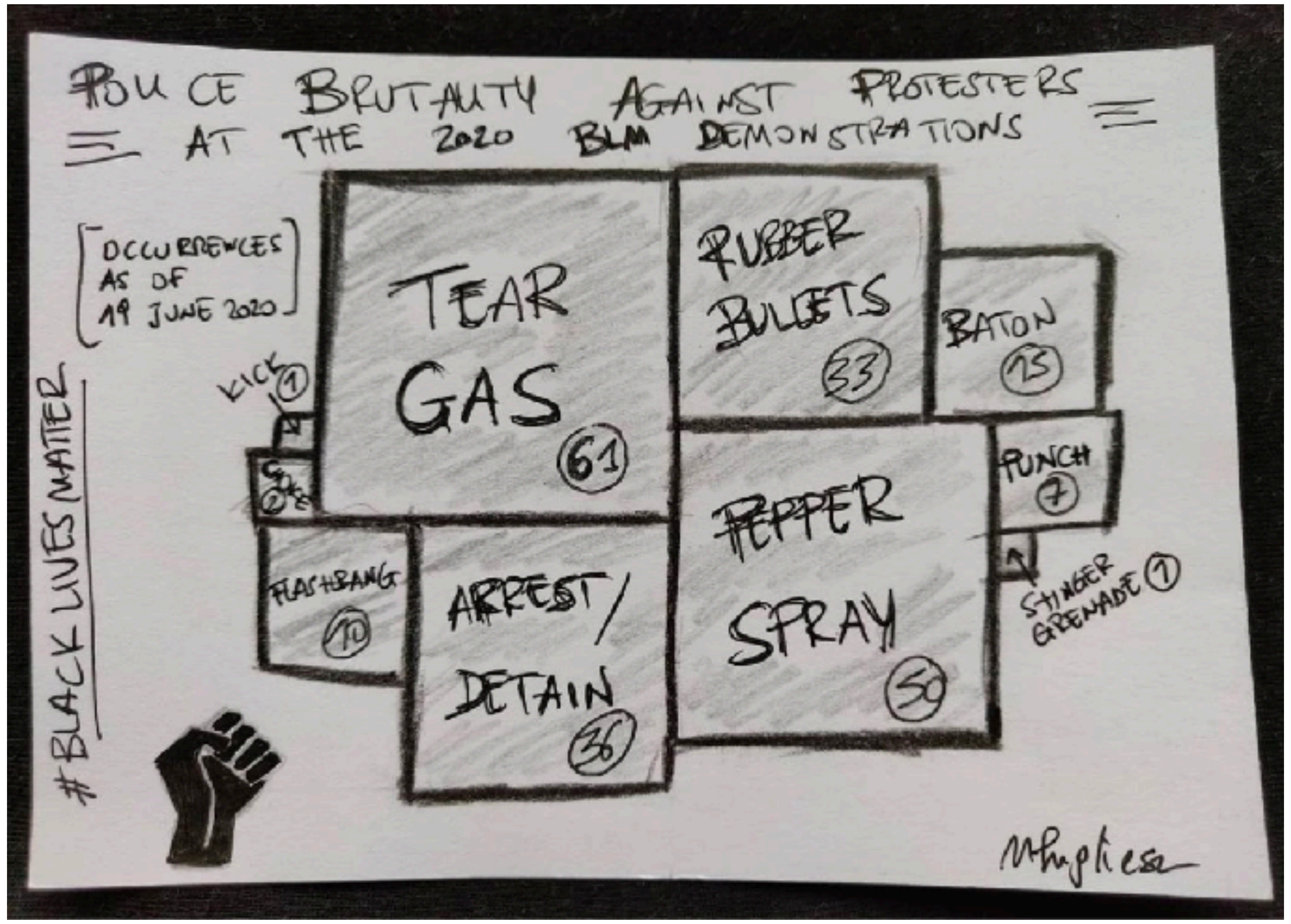


1.1 Exercise in Miro

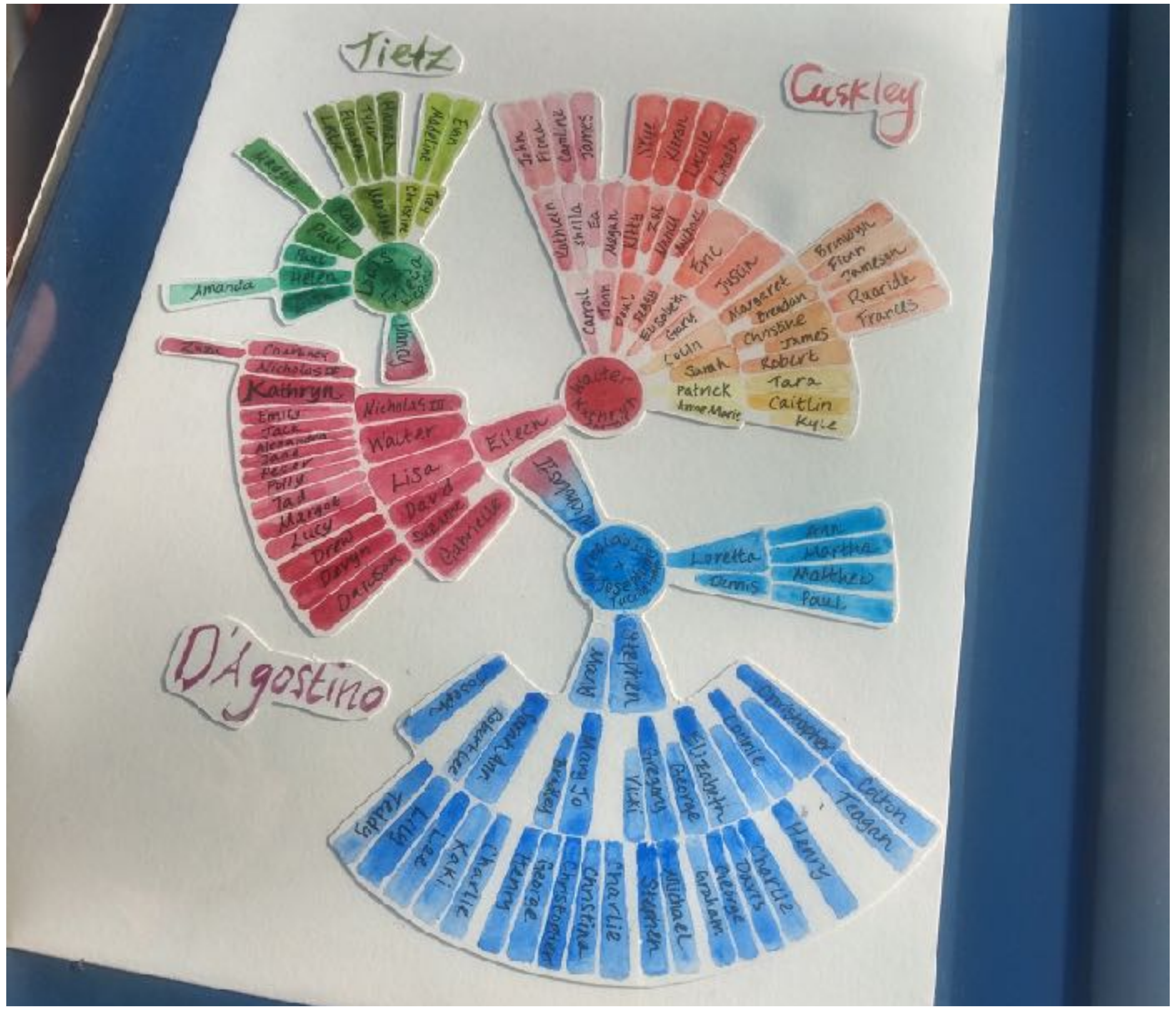
Sketch your hypotheses.

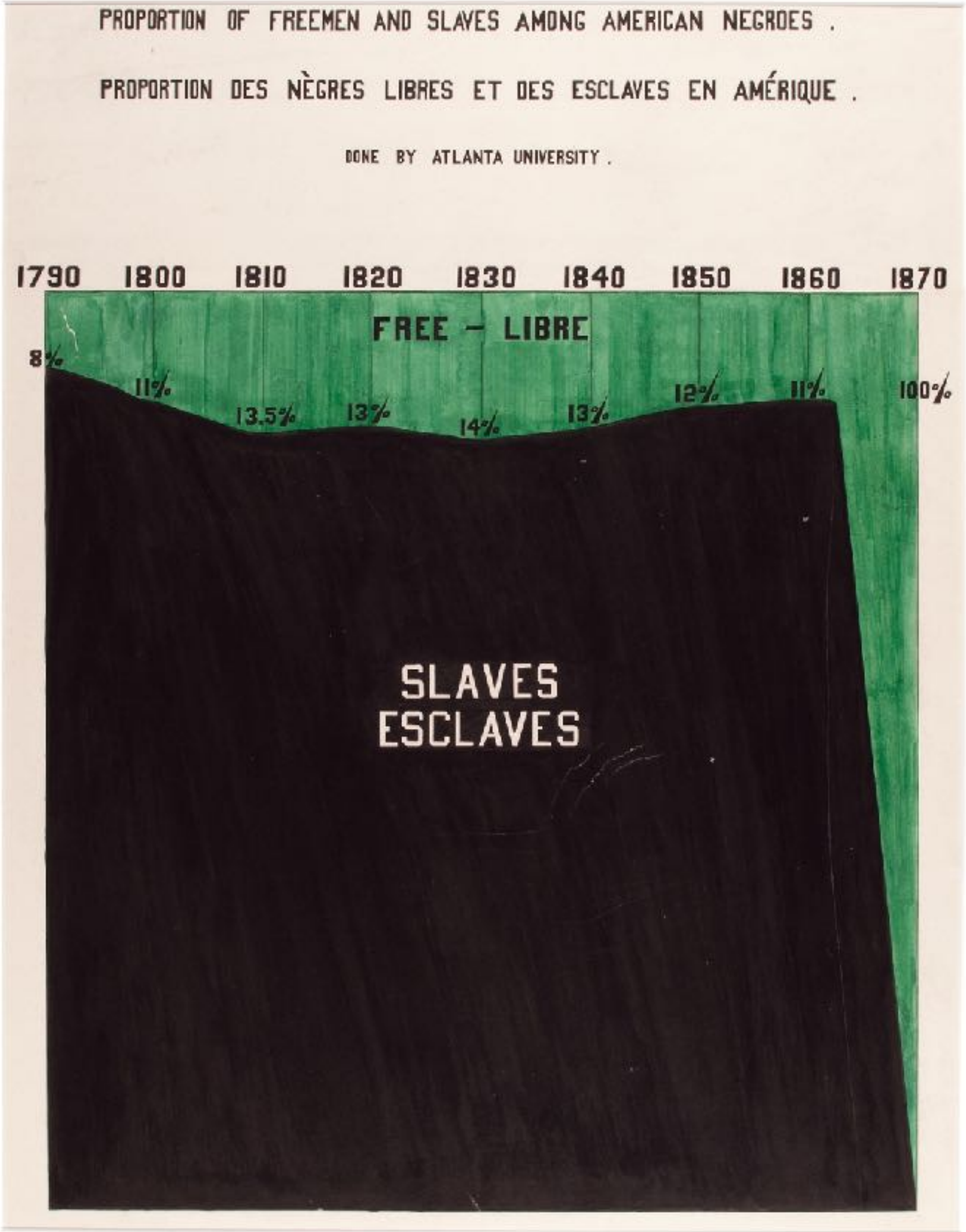
What was the point?

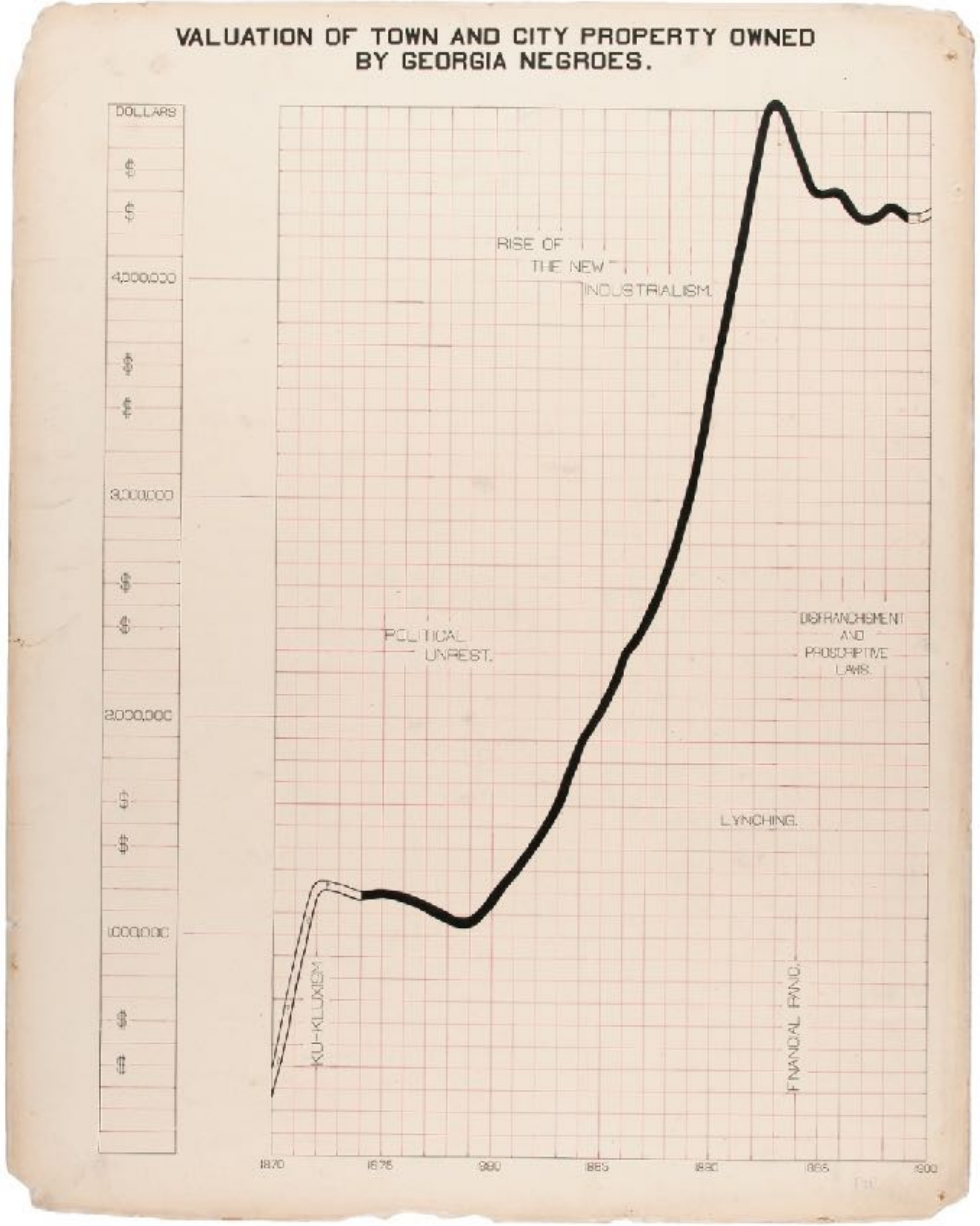
There is a lot of fancy software, but **don’t be afraid to sketch**.



@doodledatcard







https://www.smithsonianmag.com/history/first-time-together-and-color-book-displays-web-du-bois-visionary-

What was the point?

Use sketching to hypothesize, and compare to actual results - get a head start on your results narrative

Learn to have strong **visual** expectations, check whether these line up with actual results and model predictions (though we won't deal with the latter)

1.2 Real Data

The Actual Data

Now we'll look at actual data from Saldana et al., 2019, who performed this experiment with Children and Baboons

Your sketches may have set out the data in different ways; for this exercise, we'll aim to **map each variable to a specific visual dimension**

Shaping the Data

We want to look at two main independent variables

- Generation, which gets at transmission (**X axis**)
- Learner Type, which gets at human uniqueness (or not) (**Colour**)

Since these IVs are our main manipulations, they'll be the same on every graph

Shaping the Data

We want to look at three dependent variables

- Performance/Score
- Proportion of Tetrominoes
- Predictability (diversity) of responses in a set

Each DV will have it's own graph, and occupy the **Y Axis**

Note that this gives consistency to the results across variables.

The tidyverse:

“...an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.”

This will not (cannot) be a comprehensive introduction to the tidyverse.

We'll *use* the tidyverse to deal with familiar problems, which should give you an idea of the general philosophy/approach.

for more, see https://r4ds.had.co.nz/

Raw Data

In general

Raw data can take many forms, but we generally import our data in to R in **long format:**

- Each response is on a single row, with information about the attributes of that response in columns

Starting with long data is a tidyverse tenet, see e.g., Wickham (2014)

Raw Data

Saldana et al 2019

There is a separate file for children, and a separate file for baboons

They mostly have the same columns, with some exceptions.

Let's dive into the data, first, some tidyverse concepts we'll be using

Tidyverse Functions

- read_csv(): Reads a CSV file into a data frame (R’s name for “spreadsheet style” data)
- filter(): allows us to remove rows from a data frame depending on their value
- select(): allows us to take entire columns in/out of a data frame
- add_column(): allows us to add a column to the data frame
- group_by(): allows you to group by 1 or more variables
- summarise(): after group_by, allows you to summarise groups, e.g., getting mean of values in column A based on categorical variable in col B
- ggplot(): the visualisation arm of of the tidyverse

Other functions

- Vanilla R
  - head(): this allows us to see the first 6 rows of a dataframe to check what’s going on
  - unique(): gives the unique values for a variable
  - rbind(): allows us to bind multiple dataframes together, as long as they have the some column names (stands for *row bind*)
  - mean(): calculates the mean of a vector of values
- DescTools: A package for descriptive statistics in R
  - MeanSE(): calculates the standard error of a vector of values
  - Entropy(): calculates the Shannon Entropy from a frequency table

1.2-1.4 Exercises in RStudio/Rmarkdown

Go through the exercises in 1.2-1.4

Ask a peer, raise a hand, or queue up if you have questions.

What was the point?

Get to know the tidyverse a bit - there is much more to this, Google is your friend, and there are a TON of tutorials out there

Understand how you have to manipulate your data to get good visuals, and where you might run into pitfalls.

Take a break!

Roadmap

where we'll go

2. Camera-ready visualisation

  2.1 Principles of good datavis

  2.2 Camera-ready plots with ggplot: tidy geoms and custom theming

  2.3 Network visualisation using ggnet

When data is presented for people to take in, it's more often visual than anything else

This is both for general public consumption *and* scholarly publication.

Even scholarly readers often report skipping ahead to - or *only* looking at - graphs.

Visual Dimensions

Spatial (x,y axes)

Size (e.g., weight of line, size of shape)

Form (shape, e.g. of point)

Text (labels etc)

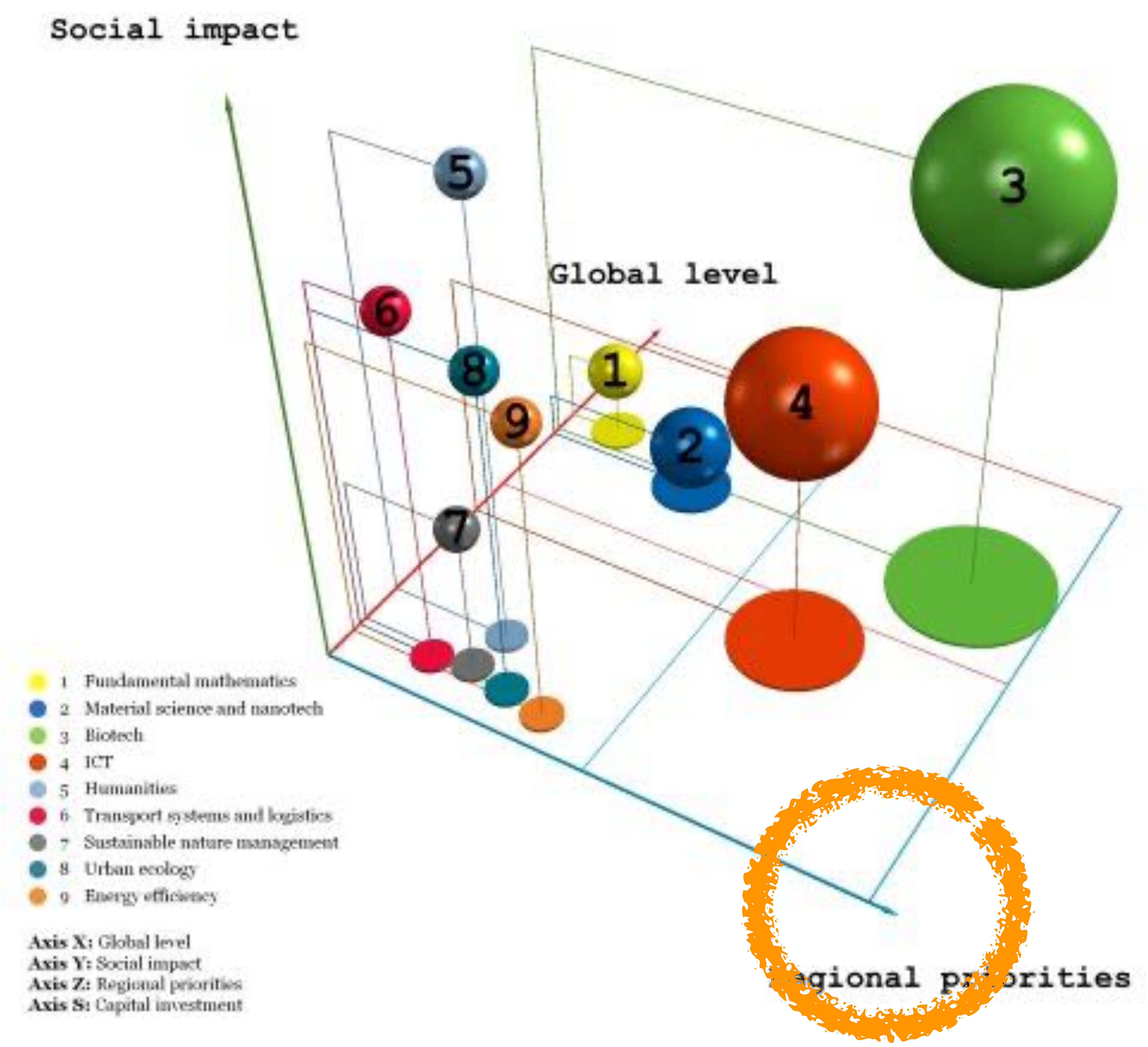
Colour (hue, but also light/dark)

Do's and Don'ts of data visualisation.

Don’ts

Don’t overcomplicate

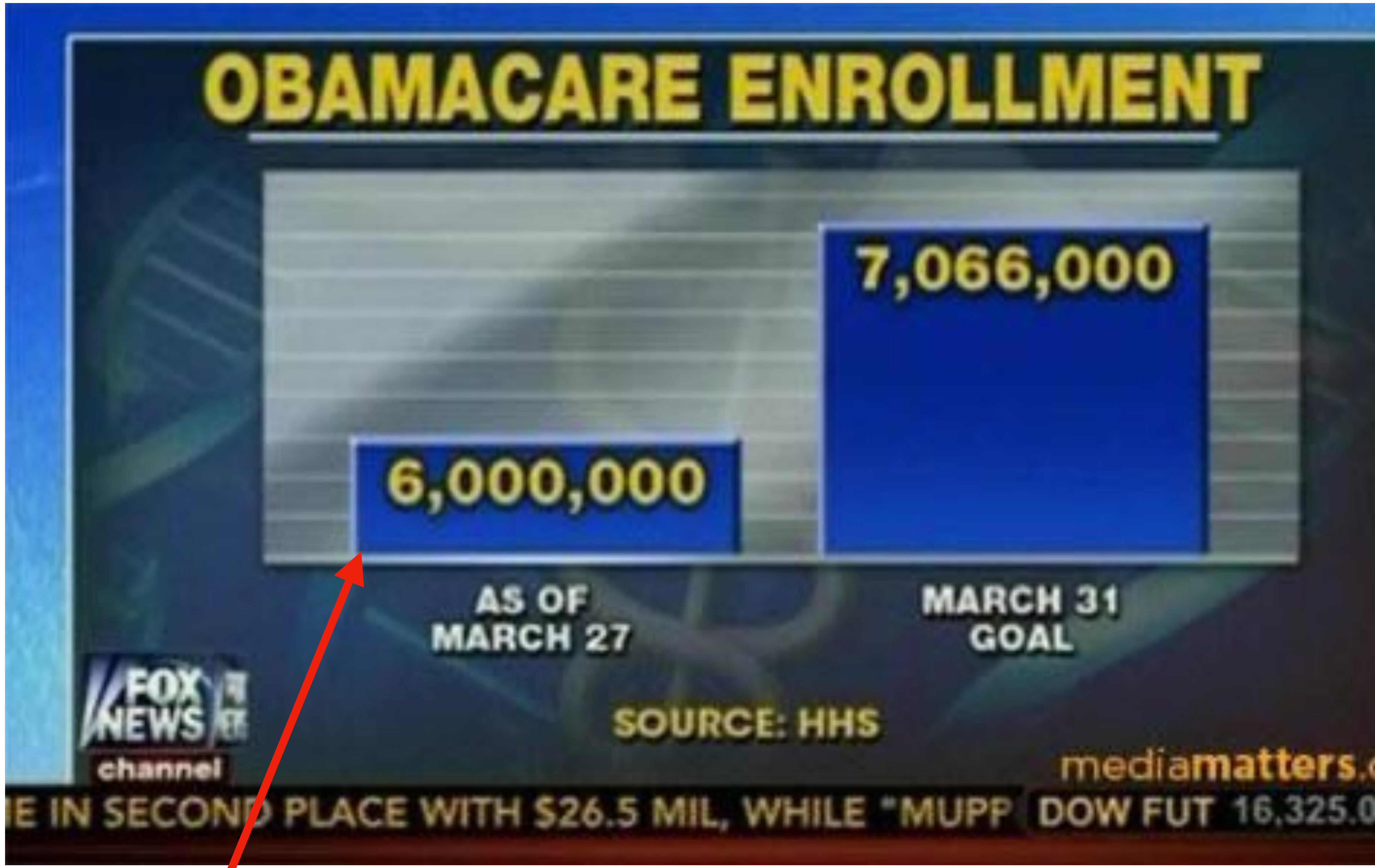
- Highlight relevant information for the viewer
- Don’t add redundant information unless its helpful



Don’ts

Fiddle with your axis ranges

- Truncate if it *really* makes sense
- But don’t do this just to mislead your viewer

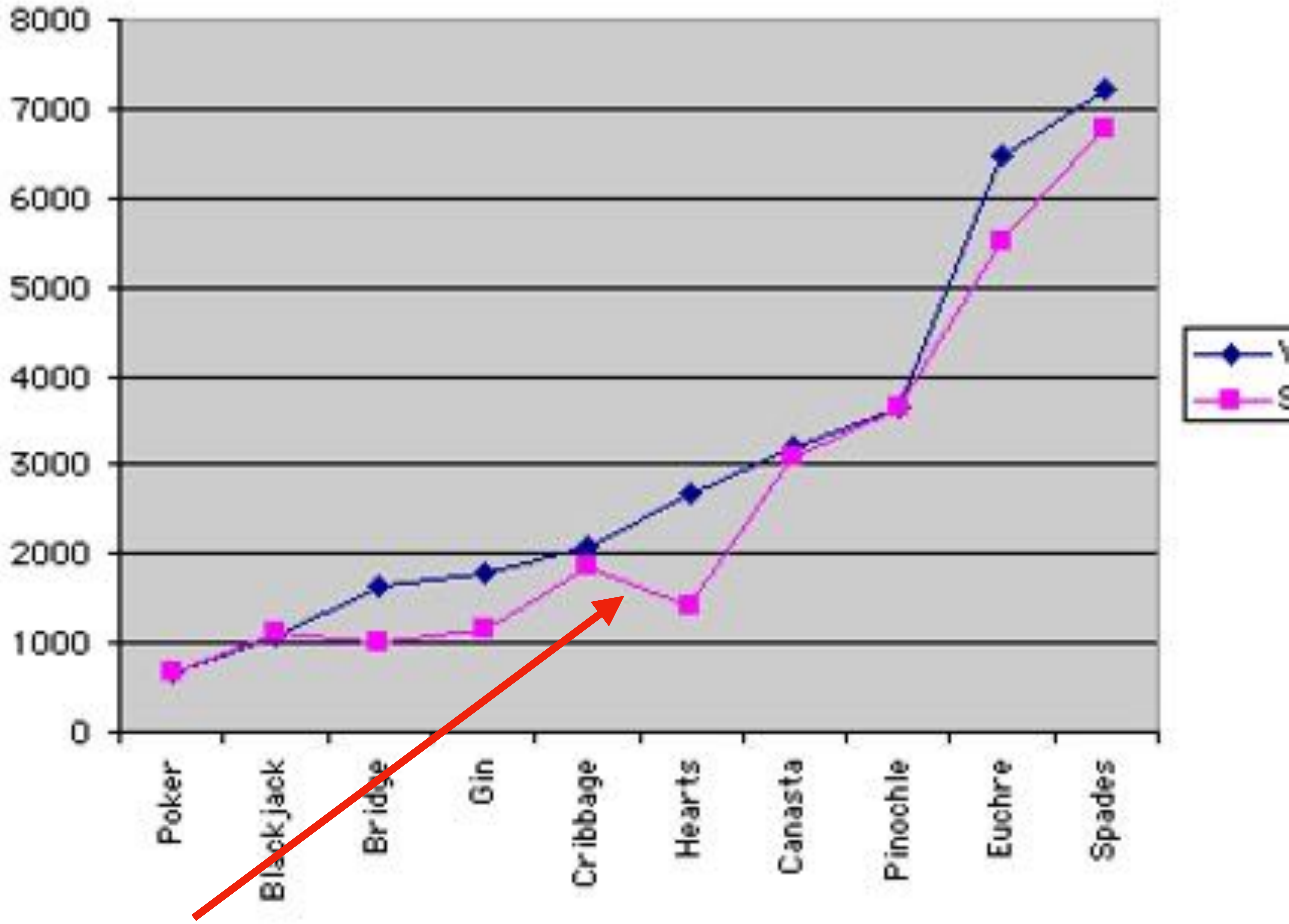


5,400,000

Don'ts

Fiddle with your axis orders

- Don't sort nominal categories to make it look like something is there when it isn't
- Don't connect nominal categories

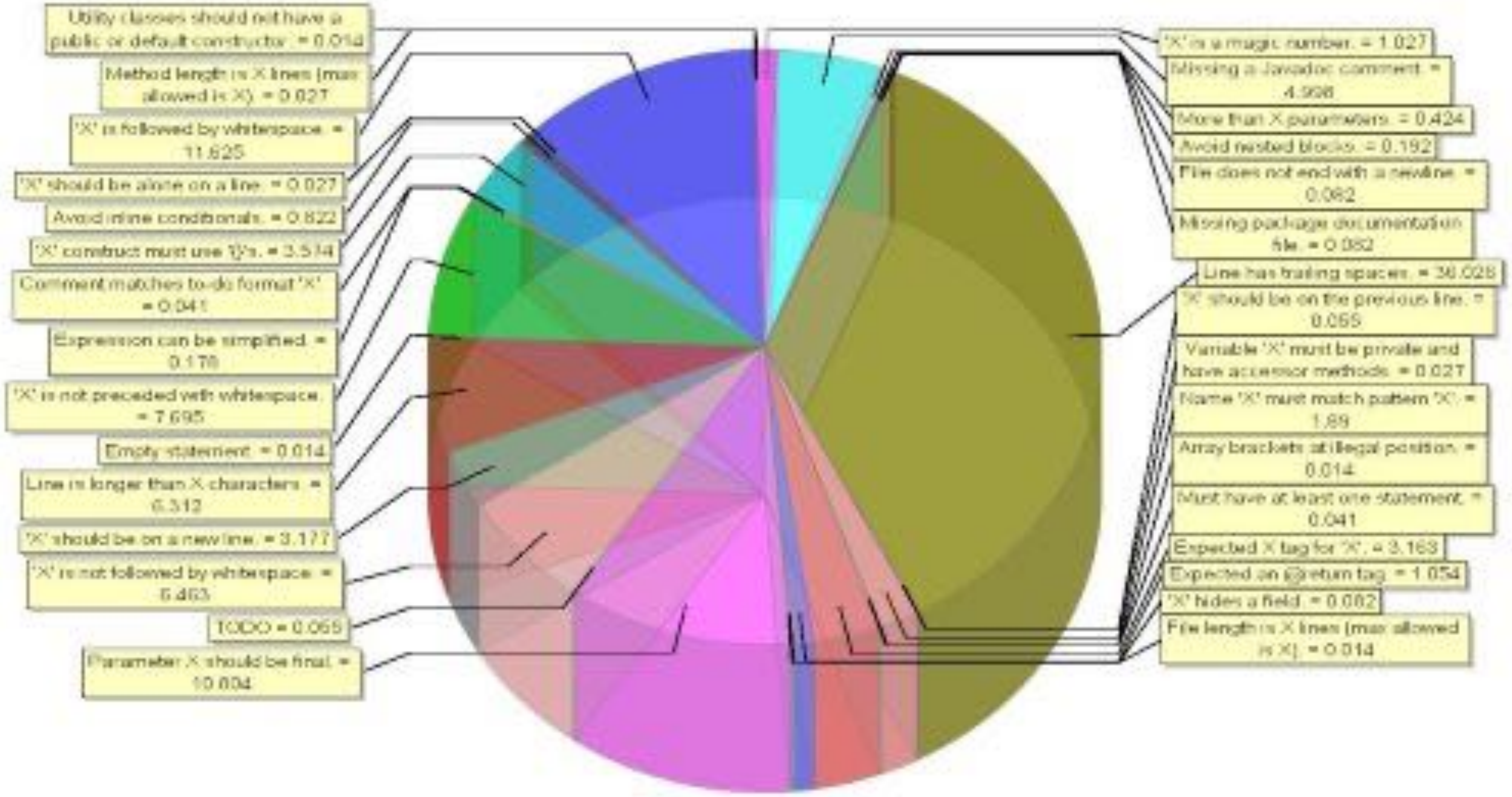


This means nothing

Don’ts

Pie chart

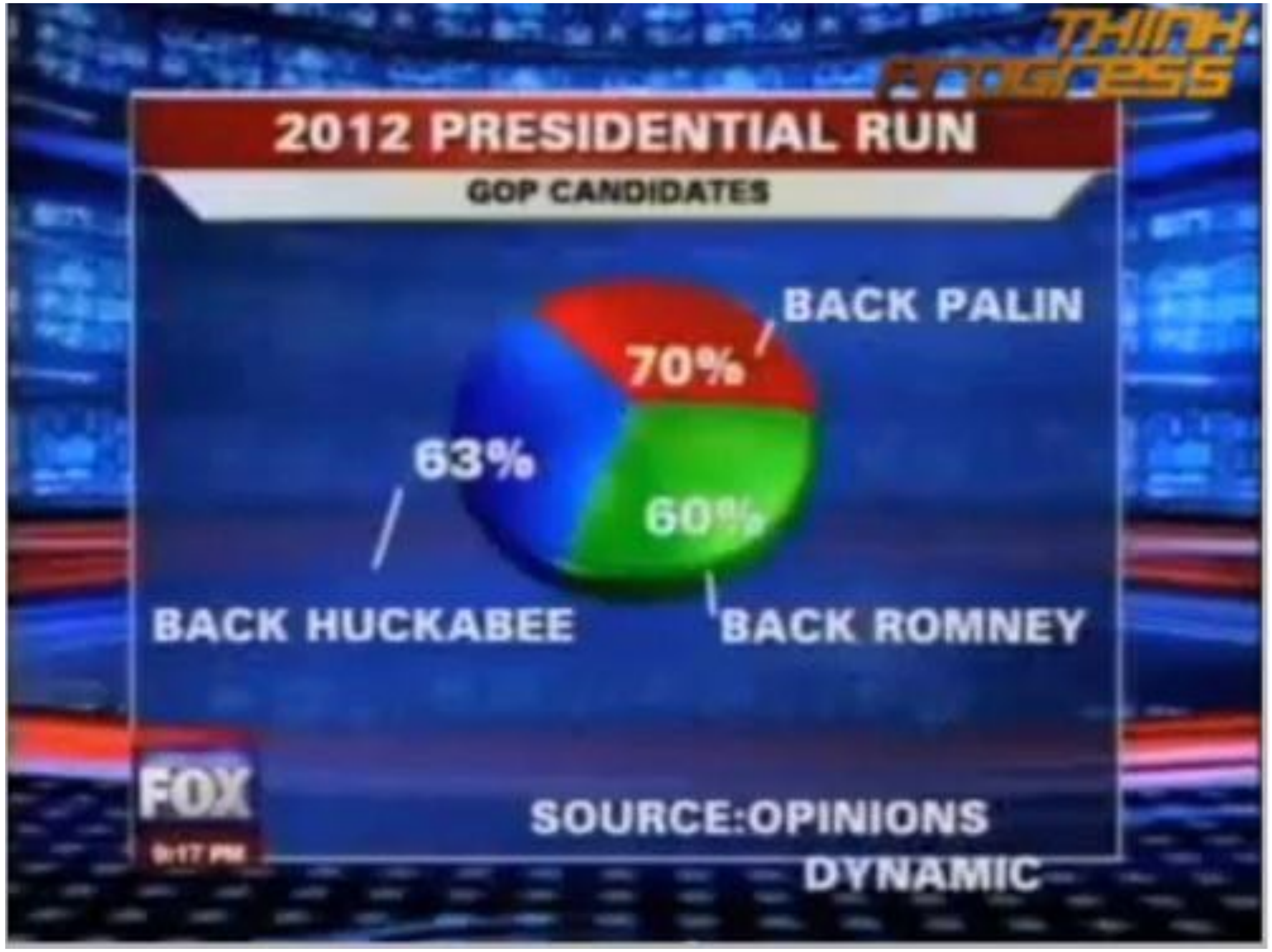
- If you *must*, make sure there are only 3-4 categories



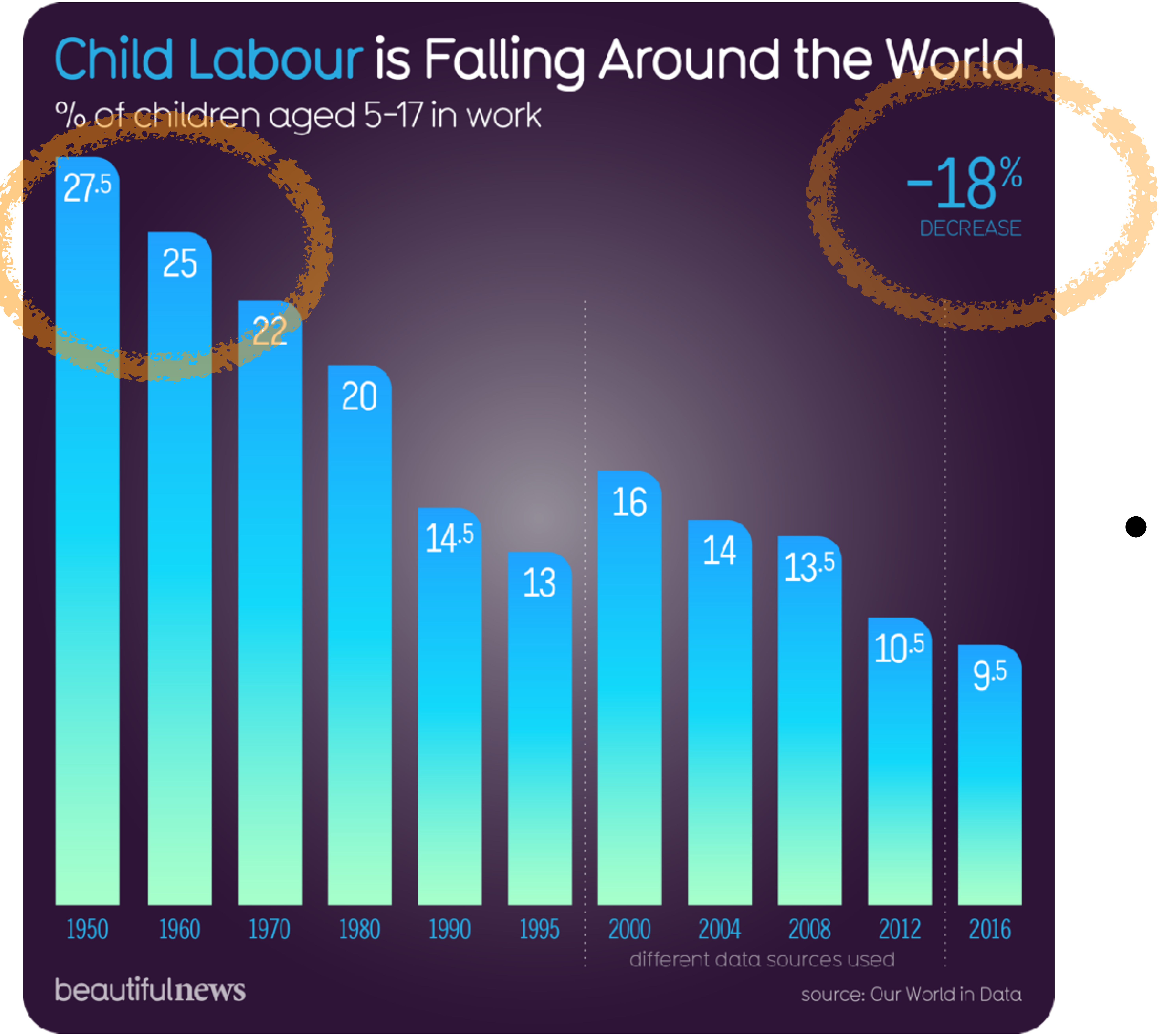
Don’ts

Pie chart

- If you must, make sure there are only 3-4 categories
- But also, just use a bar graph



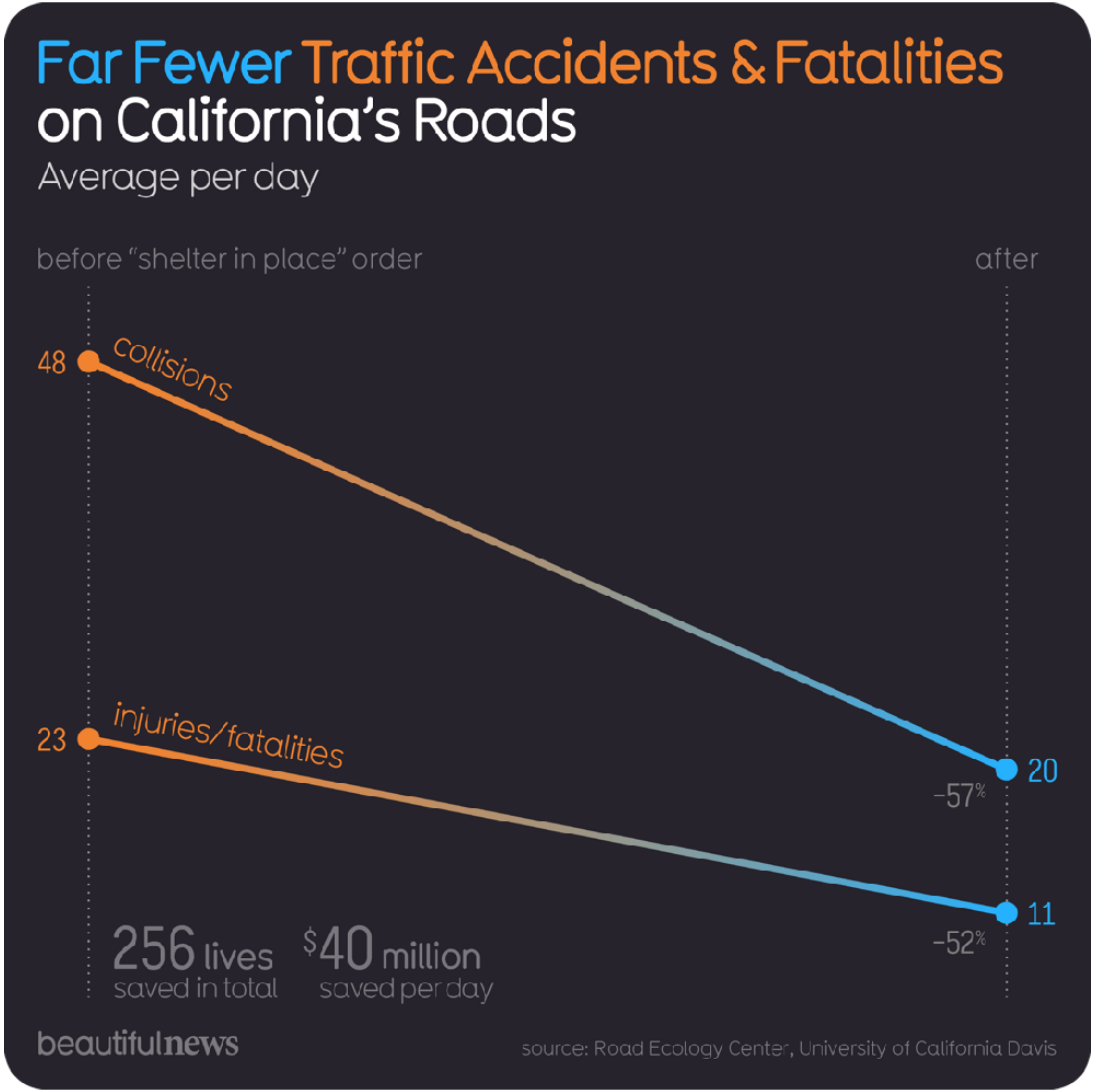
193% BACKED



Do's

Have a clear story

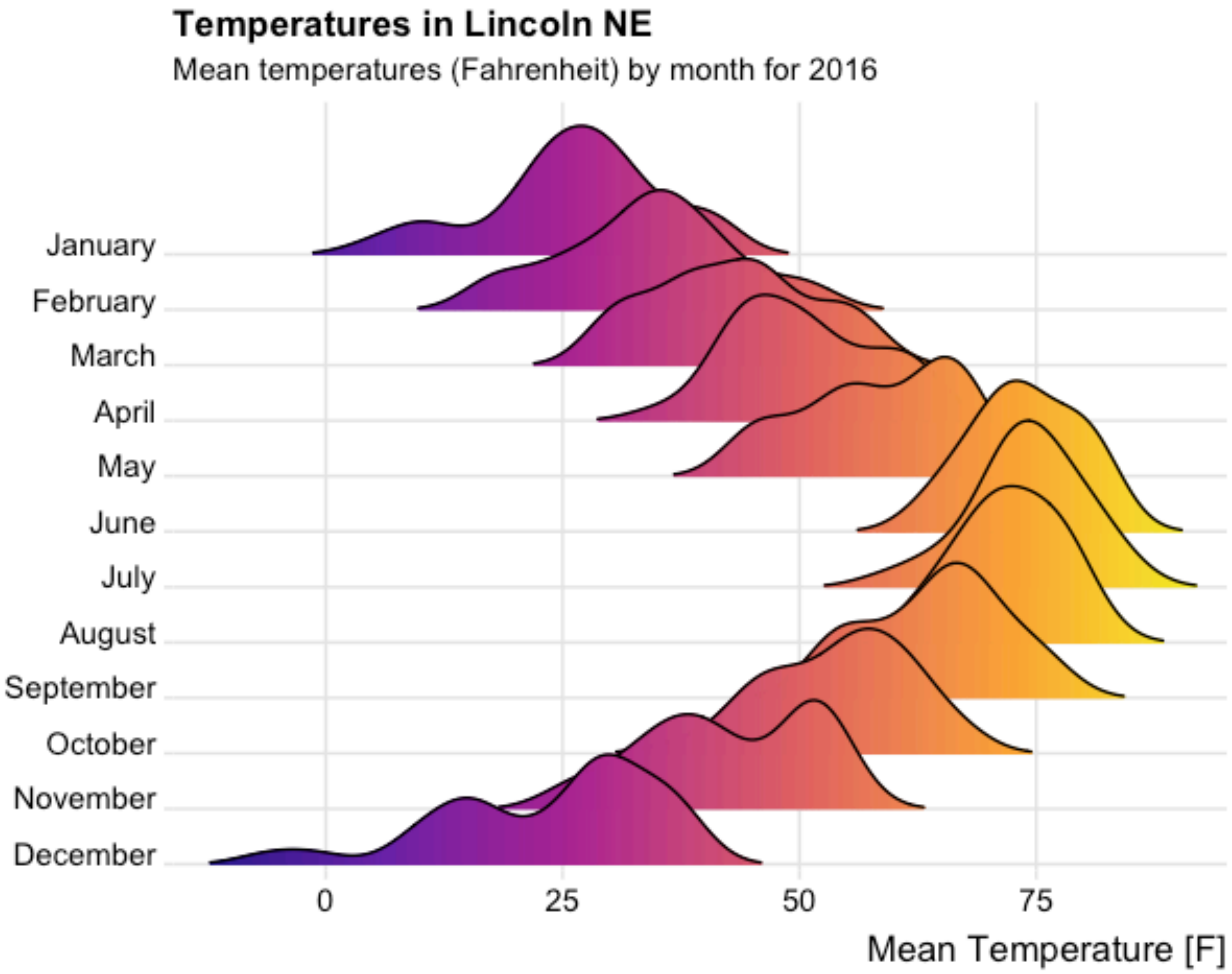
- Don't get misleading or fraudulent with your data, but do higlight salient results for the viewer



Do's

Use visual elements strategically

- Use colour, line, text etc. to highlight information for the viewer



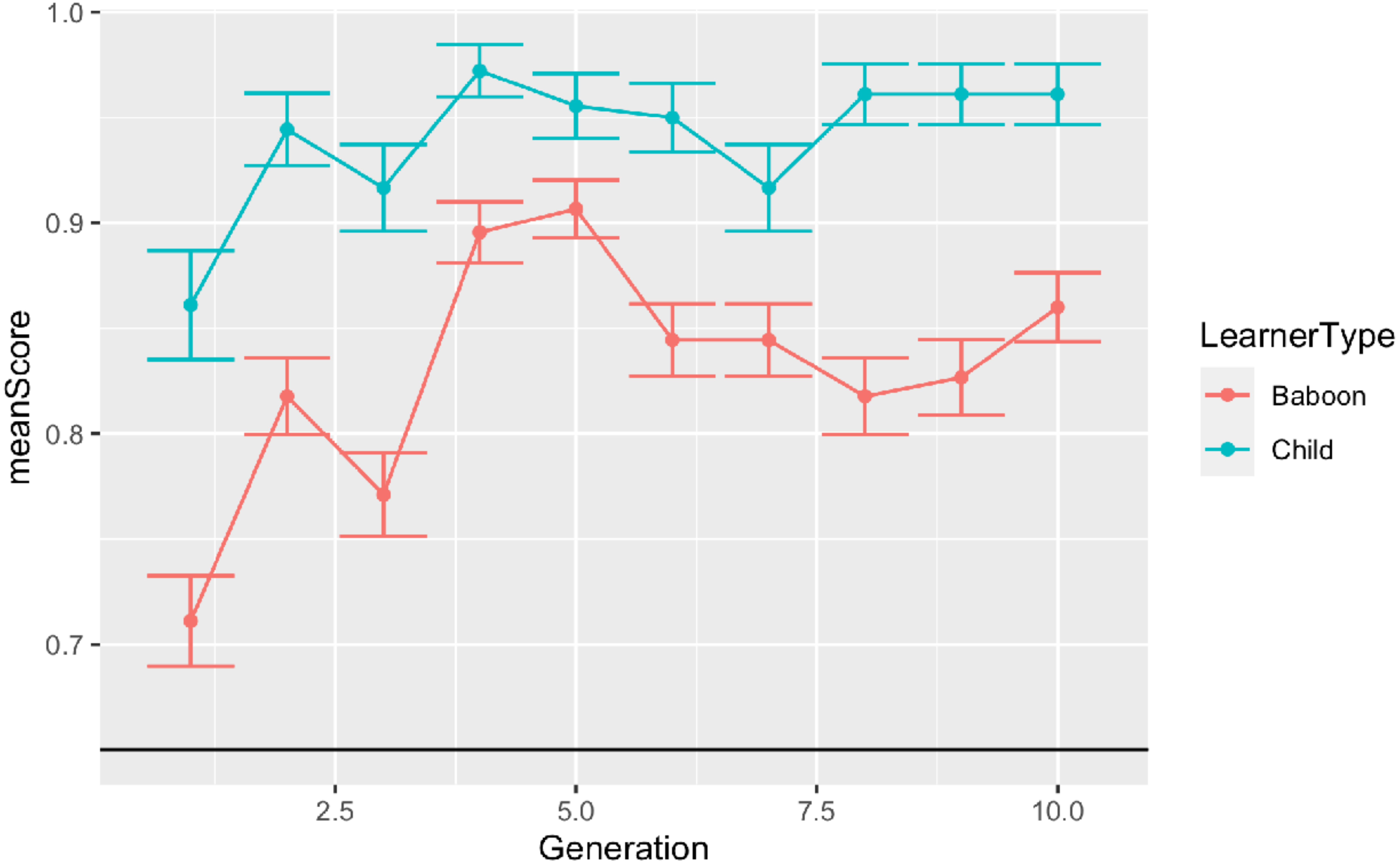
Do's

Use visual elements in ways that make sense

- Think about what viewers *expect* from colour, size, form, etc.

Let's make this look better.

(back to the markdown…)



Links etc:
bit.ly/eslr-links