DATA VISUALIZATION

JAVIER GARCIA-BERNARDO

Department of Methodology & Statistics

Utrecht University

"Those hired into analytical roles typically have quantitative backgrounds that suit them well for the other steps (finding the data, pulling it together, analyzing it, building models), but not necessarily any formal training in design to help them when it comes to the communication of the analysis—which, by the way, is typically the only part of the analytical process that your audience ever sees"

Cole Nussbaumer Knaflic. "Storytelling with Data".

WHAT IS A GOOD VISUALIZATION

Two conditions:

- Fair (e.g. correct data, not hiding important information)
- Effective and efficient: Reduce cognitive load

Cognitive load

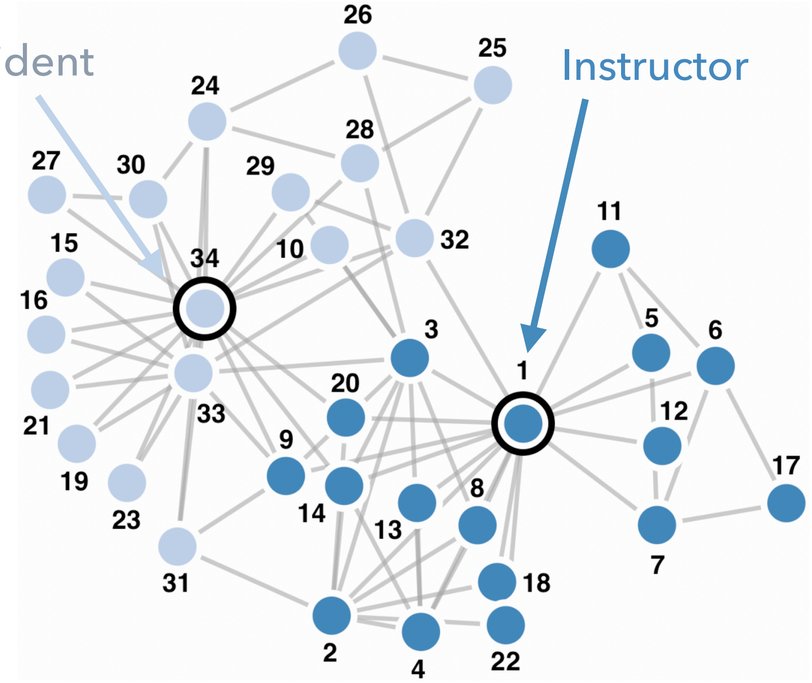
The amount of working memory used to take in information and consolidate it into long-term memory.

How much longer would it take you to understand the information presented in the left?

Individual Number

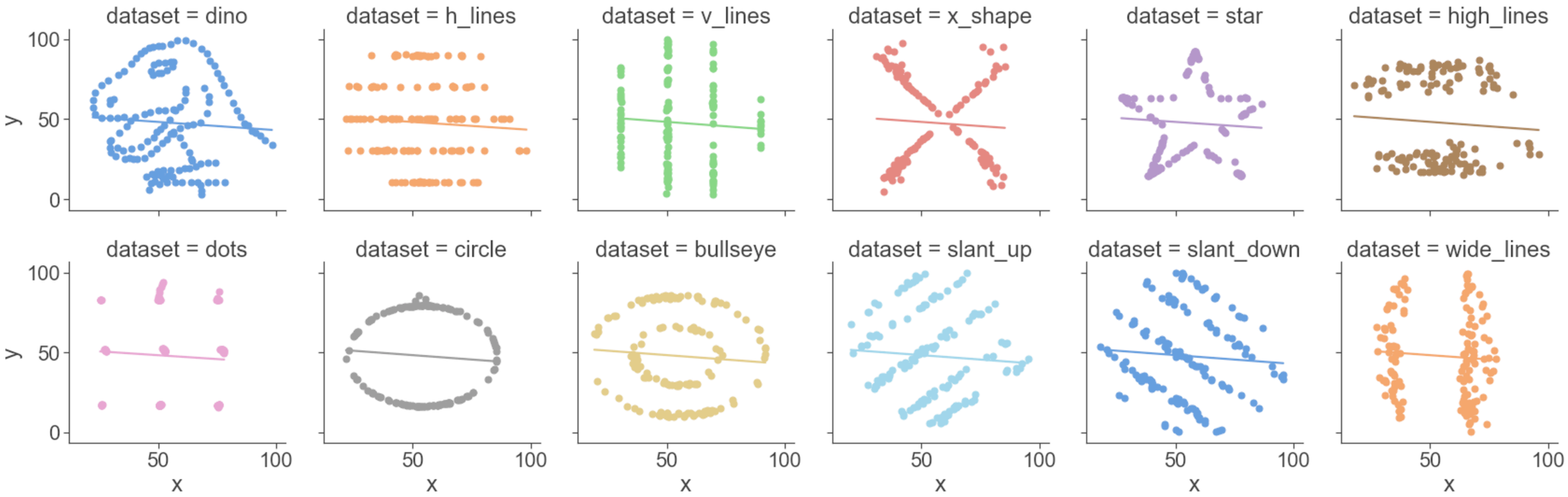
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
3	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0
4	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
17	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1
31	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1	1
33	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	1	1	1	0	1
34	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	0

Zachary's karate club network



Adapted from: https://towardsdatascience.com/preventing-and-tackling-outbreaks-f790f2fca5d2

How much longer would it take you to understand that the datasets are different?



WHY DO WE WANT TO REDUCE COGNITIVE LOAD

- More willing to **read** your paper
- More likely to **understand** the data/results
- More willing to **accept** the results
- More likely to **remember** them

HOW TO REDUCE COGNITIVE LOAD

- Good use of perception principles –> Ethical and efficient information
- Good use of design and storytelling principles –> Effective

understand the **context**

INTRO

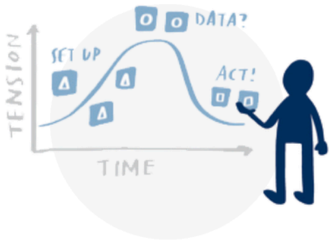
choose an **effective visual**

PERCEPTION

eliminate **clutter**

focus **attention**

DESIGN



tell a **story**

STORYTELLING

Source: Storytelling with data, Cole Nussbaumer Knaflic

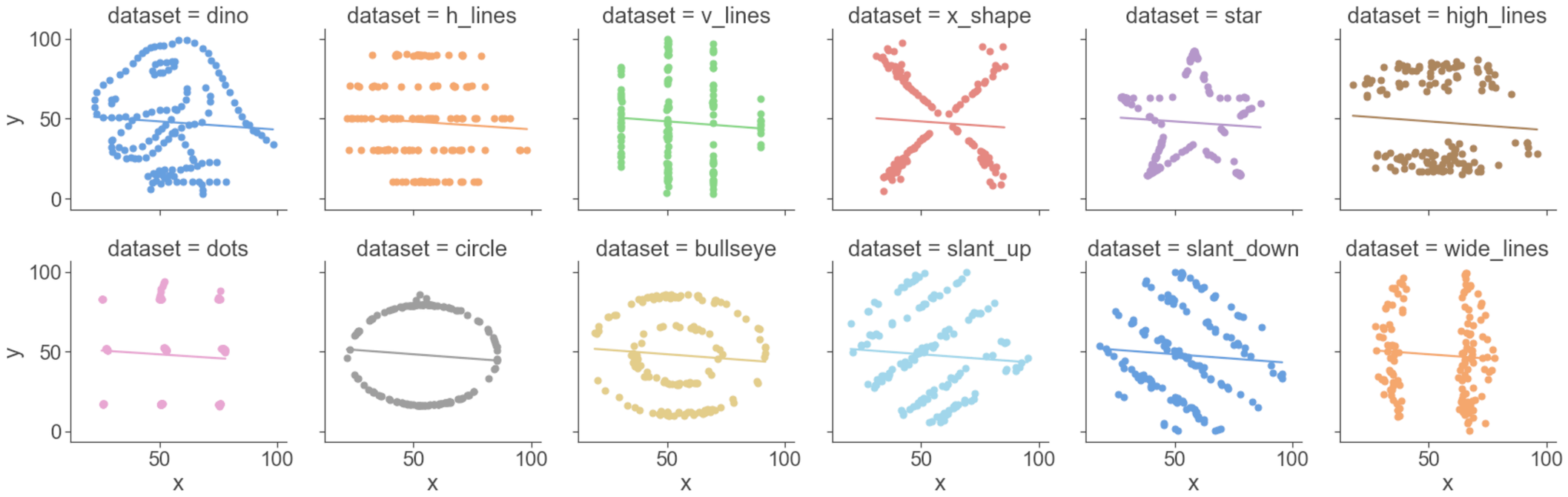
PART 1
UNDERSTAND

TWO MAIN QUESTIONS

(1) WHO IS IT FOR?

(2) WHAT IS THE MAIN MESSAGE?

(1) WHO IS IT FOR? — EXPLORATORY VS EXPLANATORY VISUALIZATION



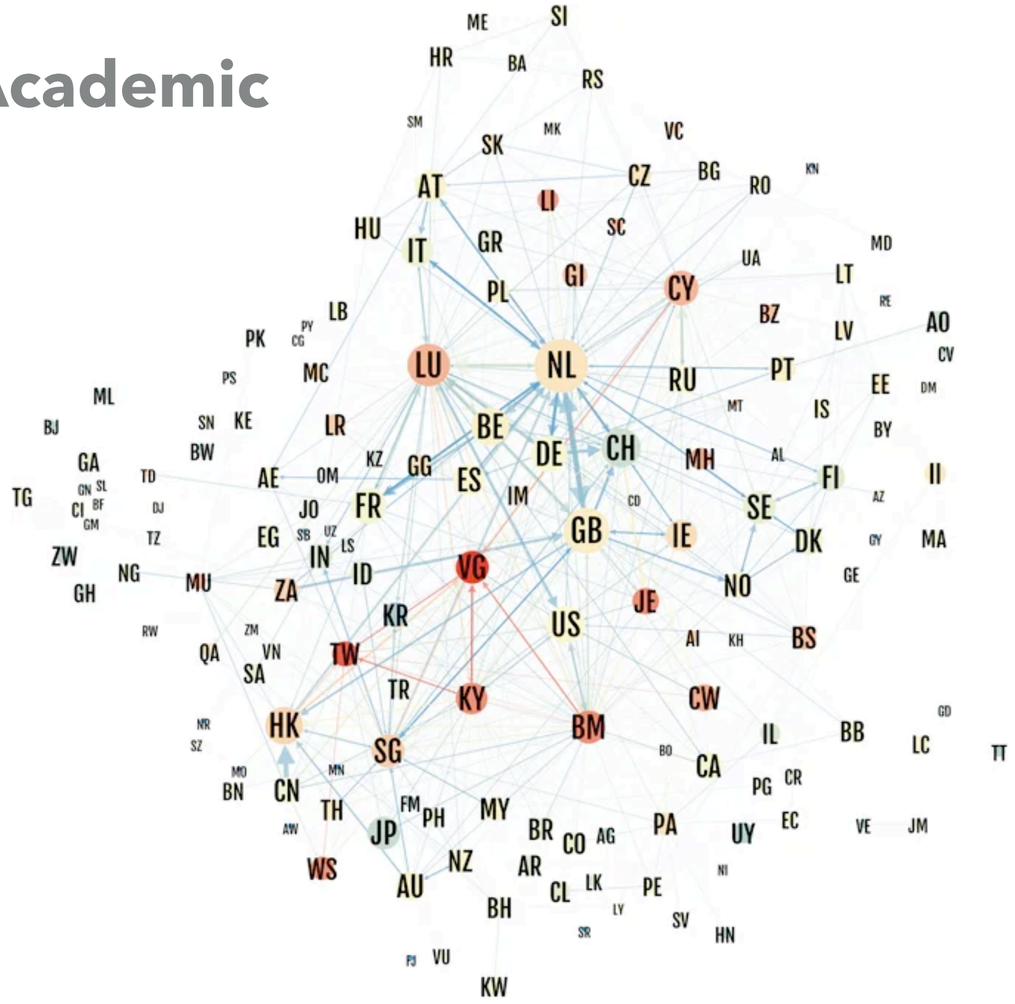
(1) WHO IS IT FOR? — AUDIENCE

	Academic journal	Presentation	Journalistic
Goal:	Convince	Inform/convince	Inform
Time devoted:	High	High	Low
Advantages:	Engaged reader	Interactive	Focused
Challenges:	Convince reader	Keep attention	Keep attention
Type:	Printed	Digital	Digital/Printed

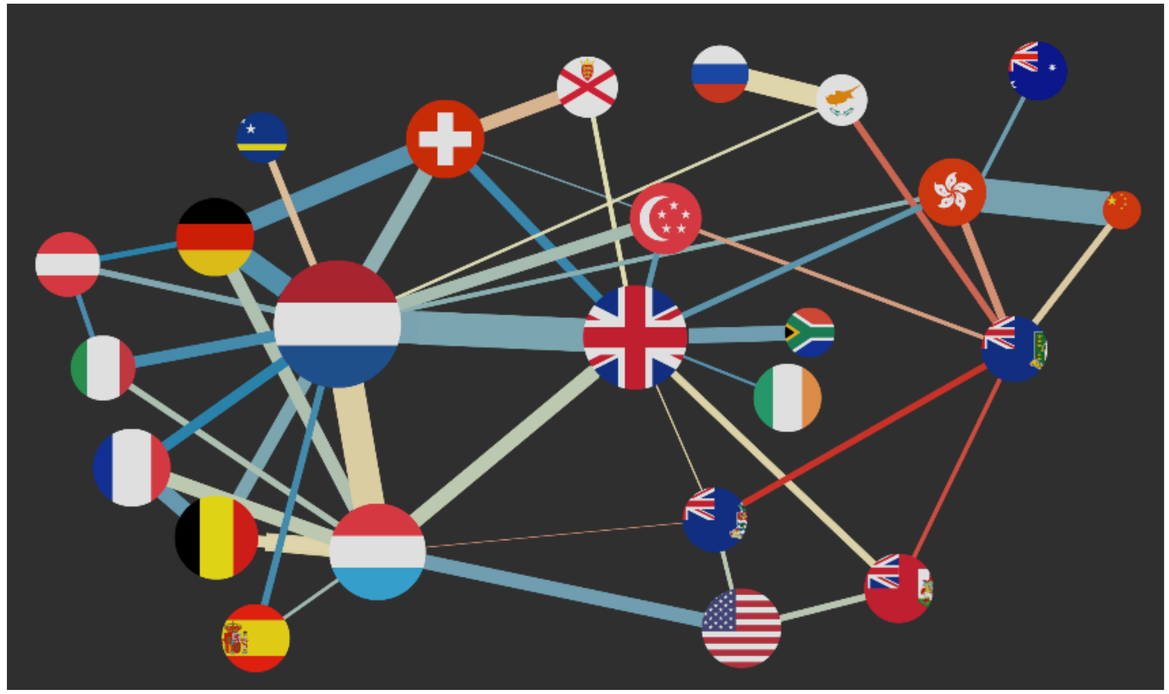
General struggle: Draw attention and keep it

(1) WHO IS IT FOR? — AUDIENCE

Academic



Journalism



(1) WHO IS IT FOR? — MEDIUM

Digital

- Can be interactive
- Higher range of colours
- Careful with contrast in presentations

Printed

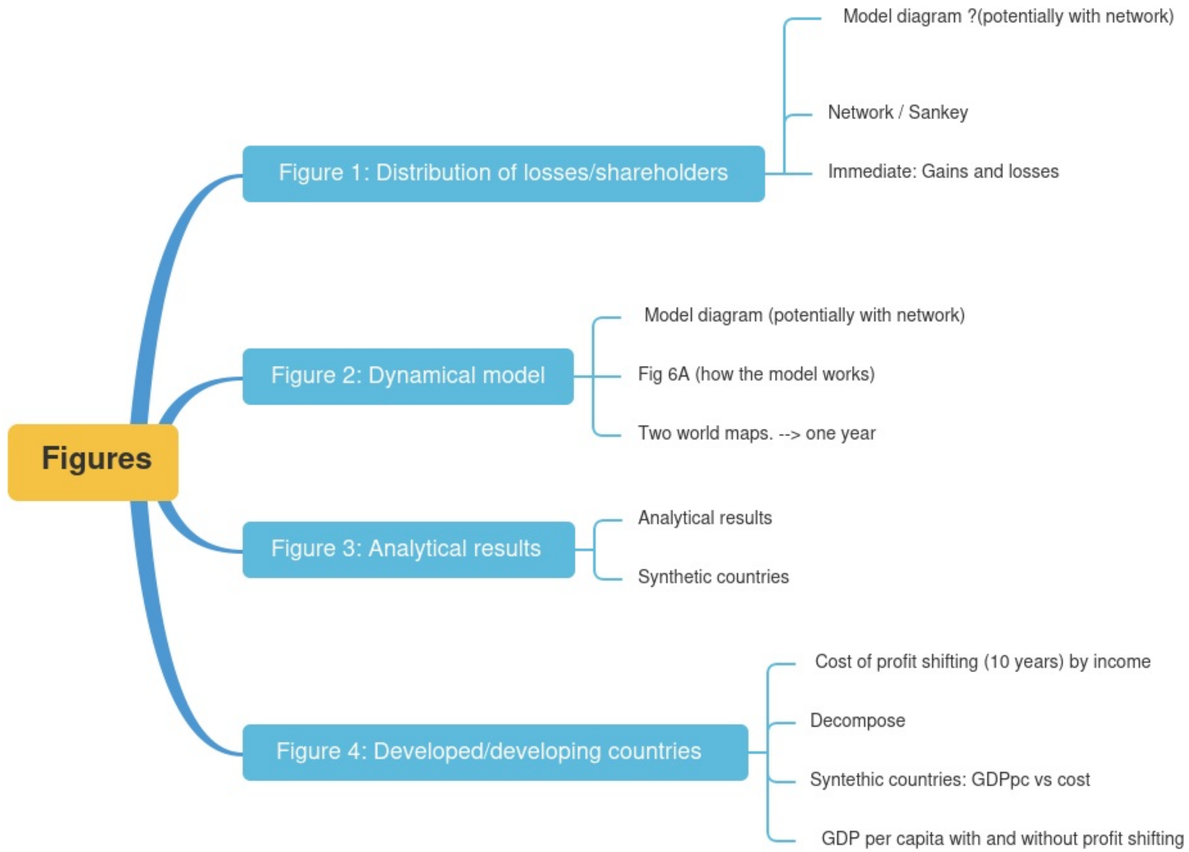
- Static
- Grayscale is important

Use **vector graphics!**

(2) WHAT IS THE MAIN MESSAGE?

Each figure should have **one** (and only one) main message

Write it down in one sentence, be specific.



EXERCISE

- Dataset: https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country
- Think about the main message
- Write it down

Testing policies and the number of tests performed per 100 000 persons, vary markedly across the EU/EEA. More extensive testing will inevitably lead to more cases being detected.

The daily reported COVID-19 cases and deaths number should be used in combination with other factors including testing policies, number of tests performed, test positivity, excess mortality and rates of hospital and Intensive Care Unit (ICU) admissions, when analysing the epidemiological situation in a country. Most of these indicators are presented for EU/EEA Member States in the Country Overview report.

Even when using several indicators in combination, comparisons between countries should be done with caution and relevant epidemiological expertise.

Variable	Definition	Code
dateRep	Date of reporting “dd/mm/yyyy”	string
day		unit8
month		unit8
year		unit16
cases	Number of newly reported cases	int64
deaths	Number of newly reported deaths	int64
countriesAndterritories	Name of the country or territory	string
geoId	2-letter code	string
countriesAndterritoryCode	3-letter ISO code	string
popData2020	Eurostat 2020 data	int64
continentExp	Name of the continent reporting	string

PART 2
EFFECTIVE VISUALS

tinyurl.com/uva-dataviz

WHICH FIGURE SHALL I USE?

GOALS:

Two conditions:

- Fair (e.g. correct data, not hiding important information)
- Effective and efficient: Reduce cognitive load

ELEMENTS OF A GRAPH (GRAMMAR OF GRAPHICS, WICKHAM 2010)

- **Mapping data to channels** (position, shape, color, …)
- **Geometric objects** (points, lines, bars, …)
- Scales (continuous, discrete, …)
- Facets (small multiples)

Additionally:

- Statistical transformation (identity, binning, unique, jitter, …)
- Coordinate system (cartesian, polar, parallel, …)

ELEMENTS OF A PLOT:

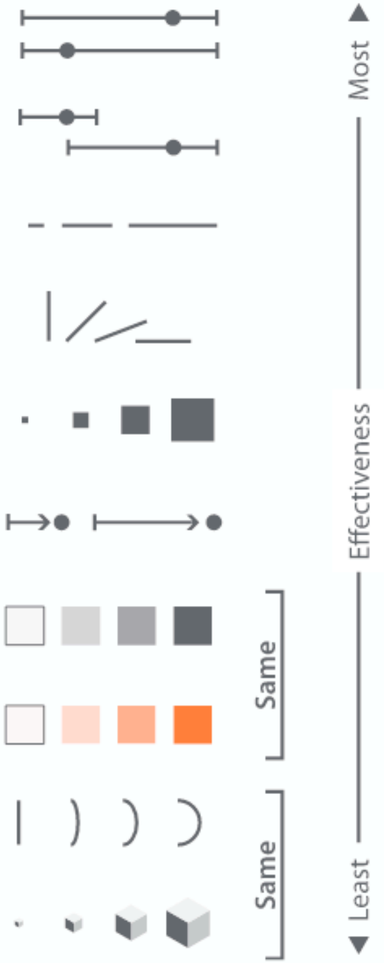
CHANNELS

PART 2: CHOOSE AN EFFECTIVE VISUAL

Channels: Expressiveness Types and Effectiveness Ranks

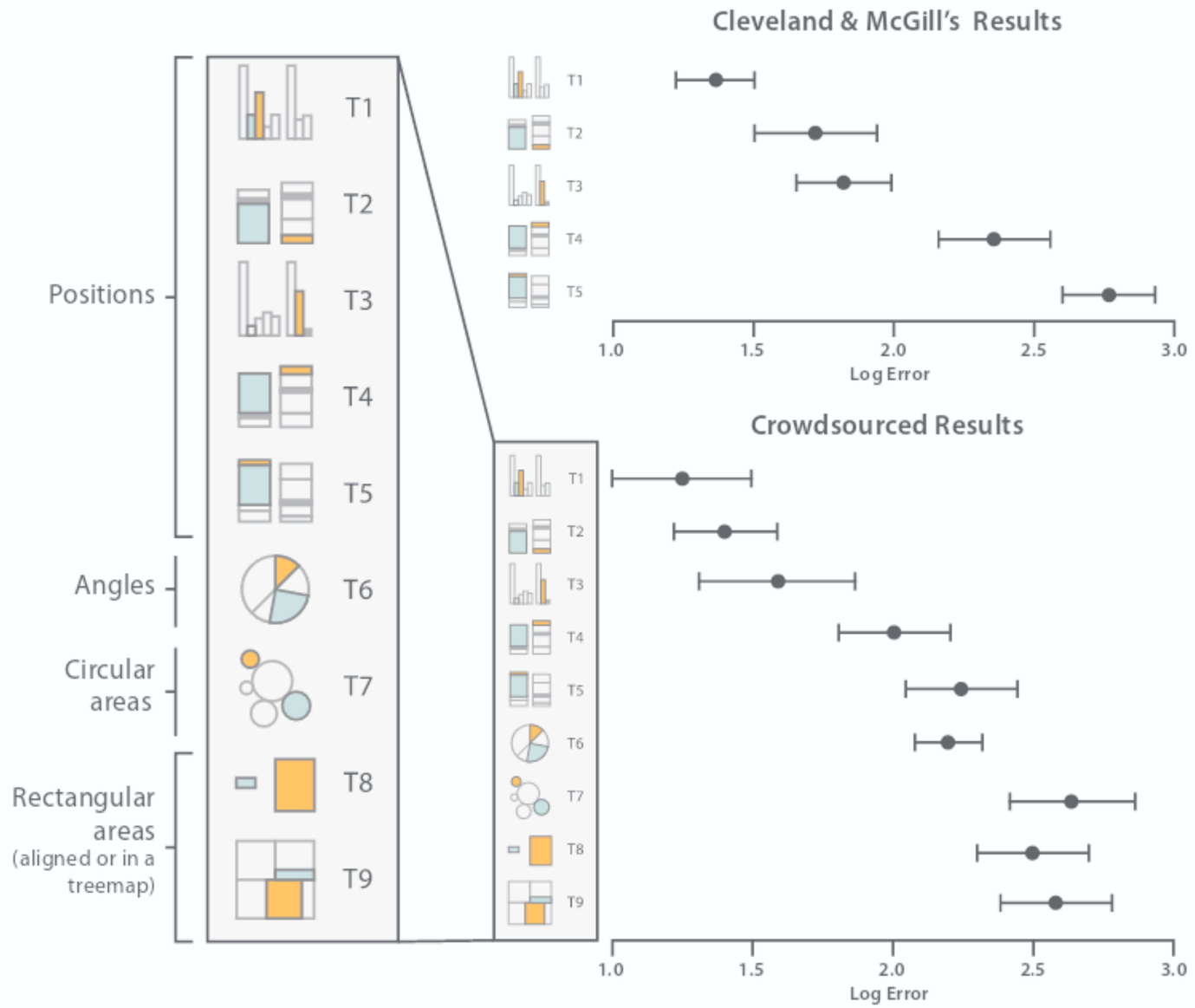
Magnitude Channels: Ordered Attributes

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)



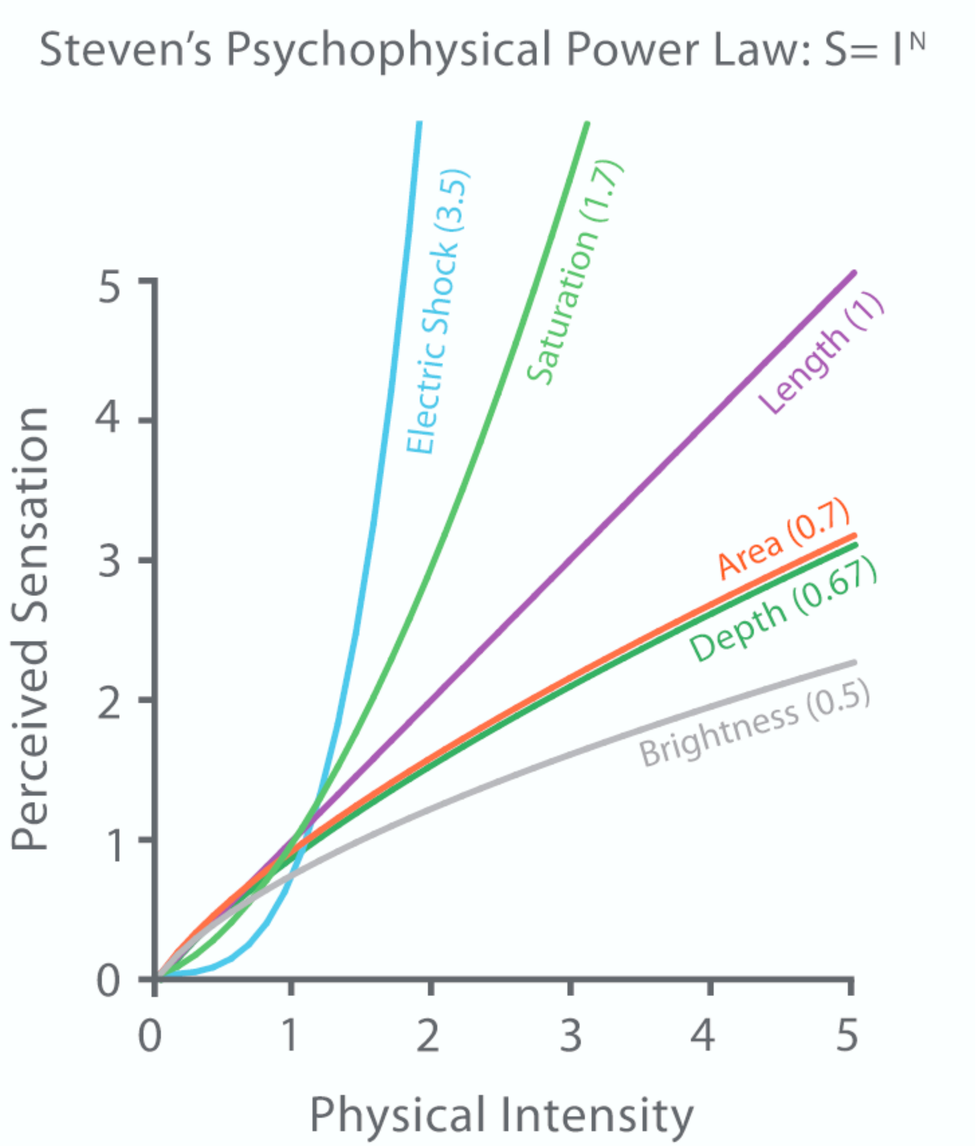
Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape



NOT ALL CHANNELS ARE EQUAL

HUMANS ARE BIASED



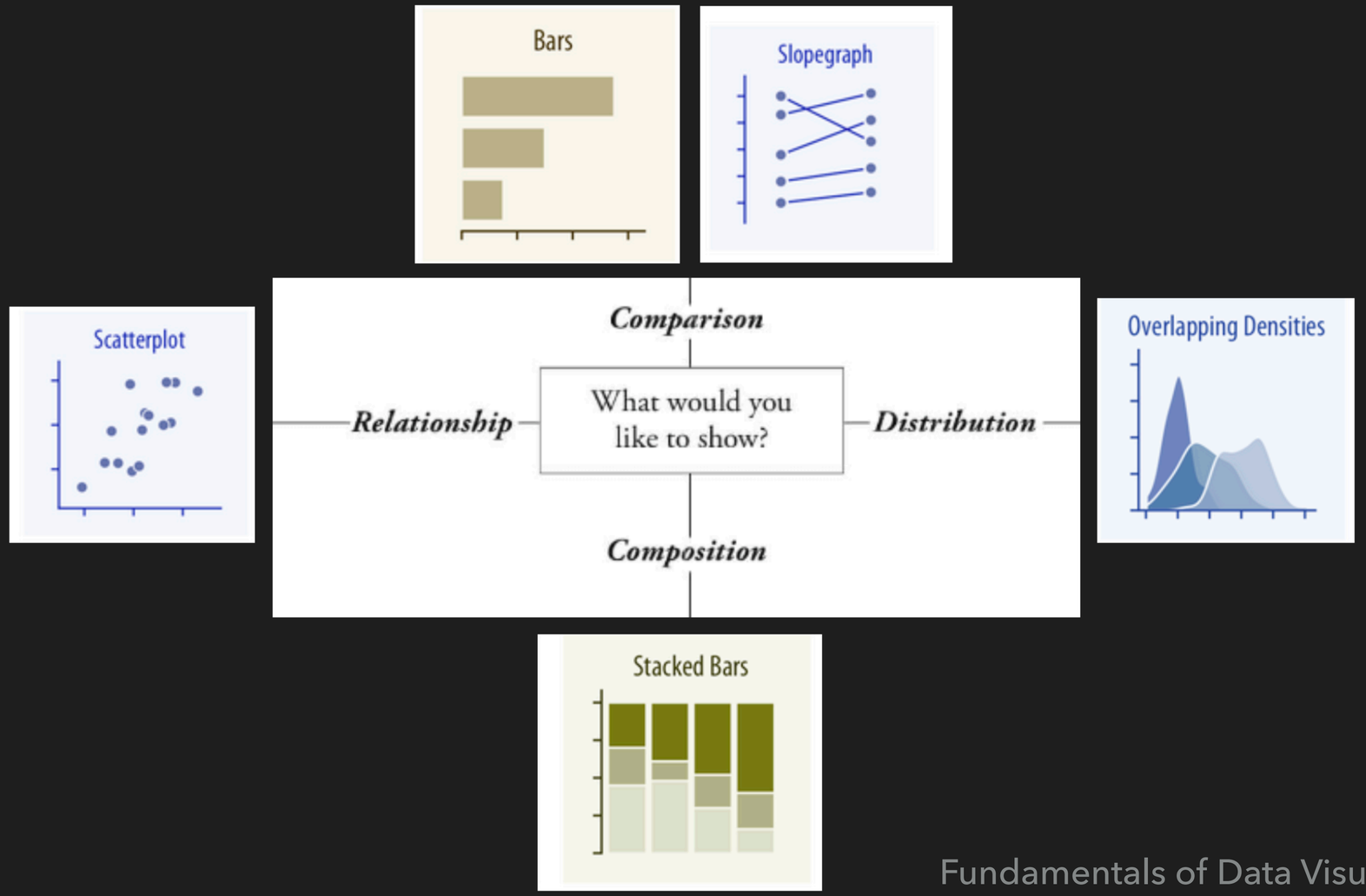
HOW TO USE CHANNELS

- Length is unbiased
- Comparing length in the same scale is best
- Use color for categorical variables

ELEMENTS OF A PLOT:

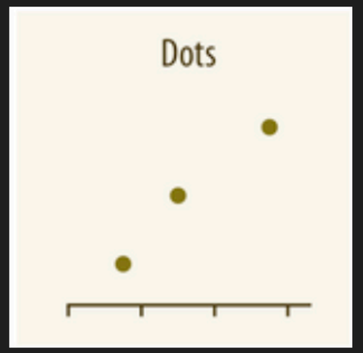
GRAPHICAL OBJECTS

THE TYPE OF GRAPH DEPENDS ON THE GOAL



Fundamentals of Data Visualization by Claus O. Wilke

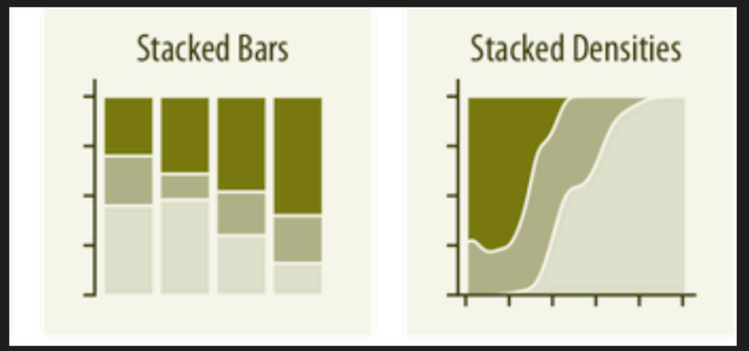
AMOUNTS AND PROPORTIONS



Required with log-scales

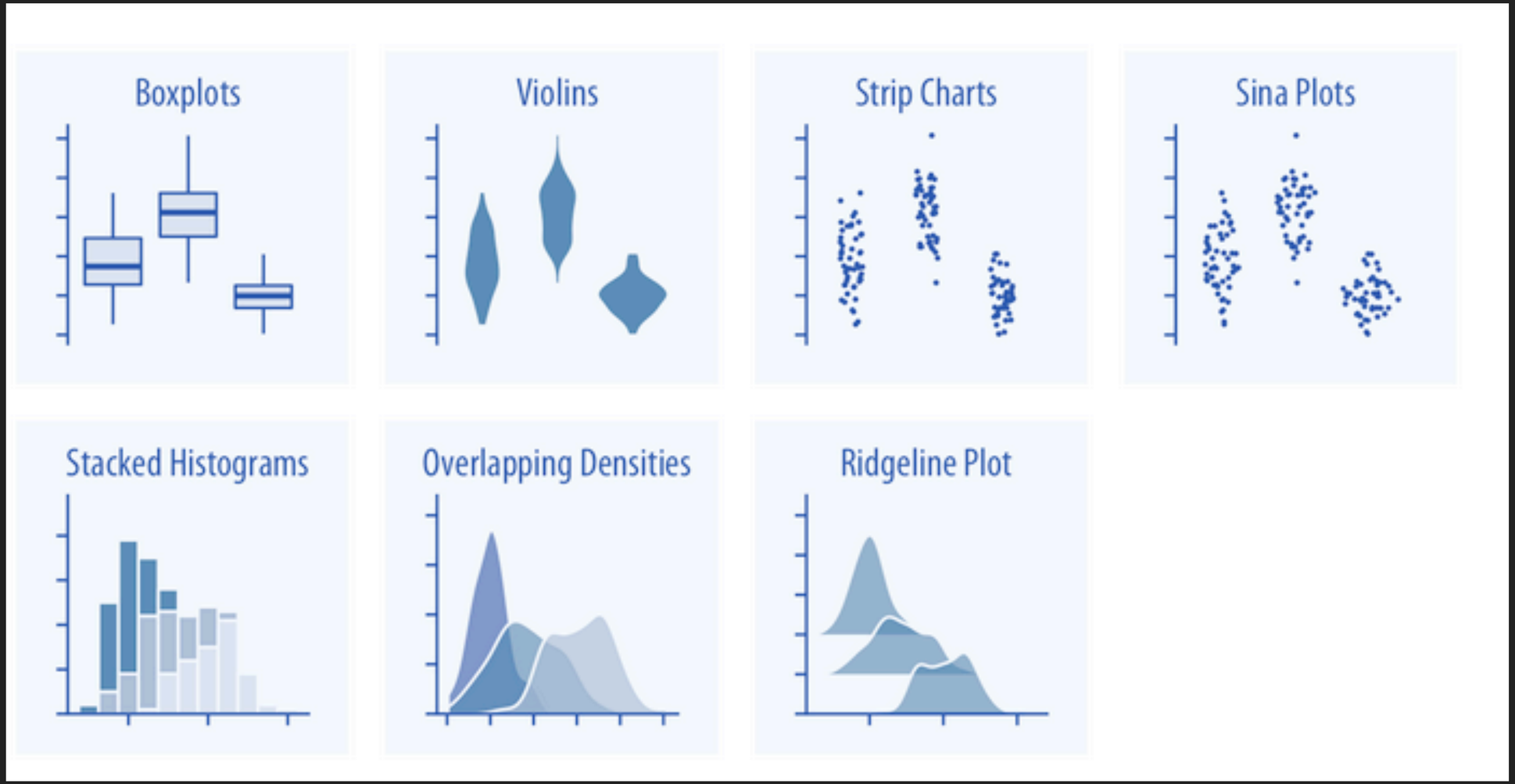
Correlation plots

Differences within row

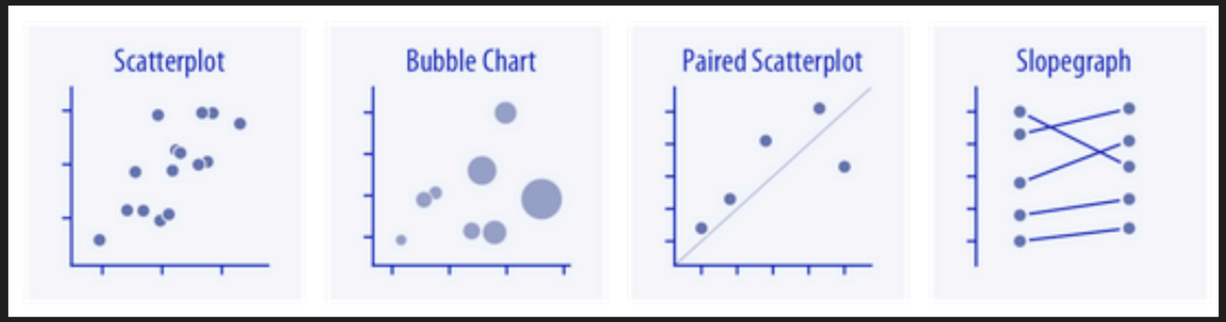


Proportions over x

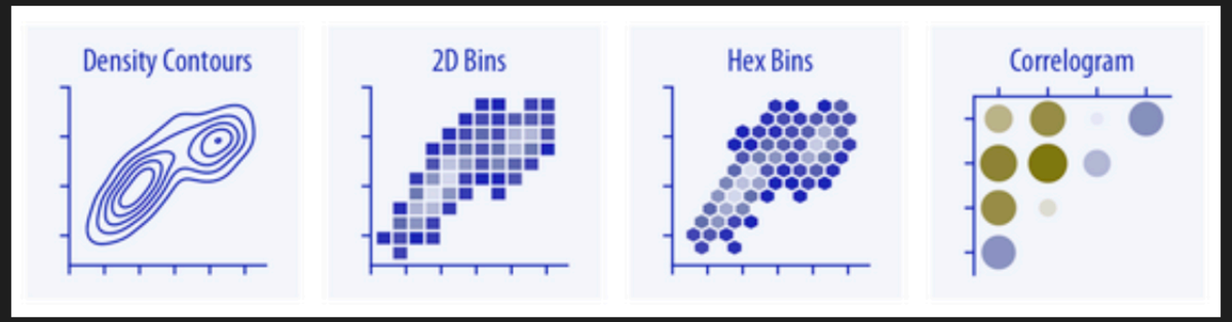
DISTRIBUTIONS



RELATIONSHIPS



Too many points?

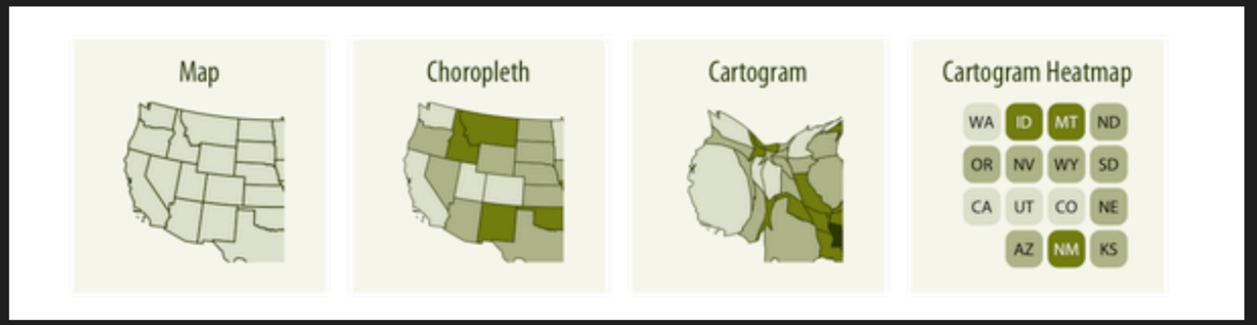


Time series?

GEOGRAPHICAL DATA

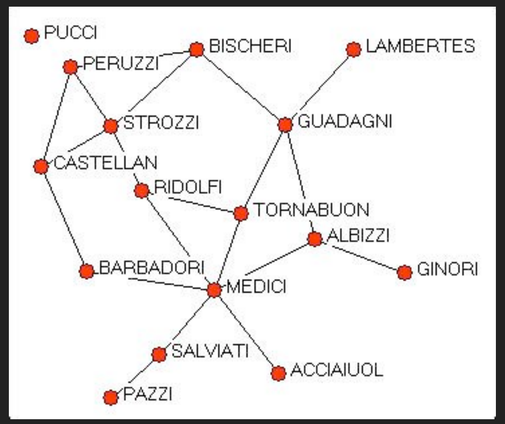
- Color is key (more on this later)
  - No rainbows!
  - Scale
    - Diverging scales: when the middle point
    - Continuous scales: start
- Combine with a barplot or bubbles if the values are important

Do you *actually* need a map?

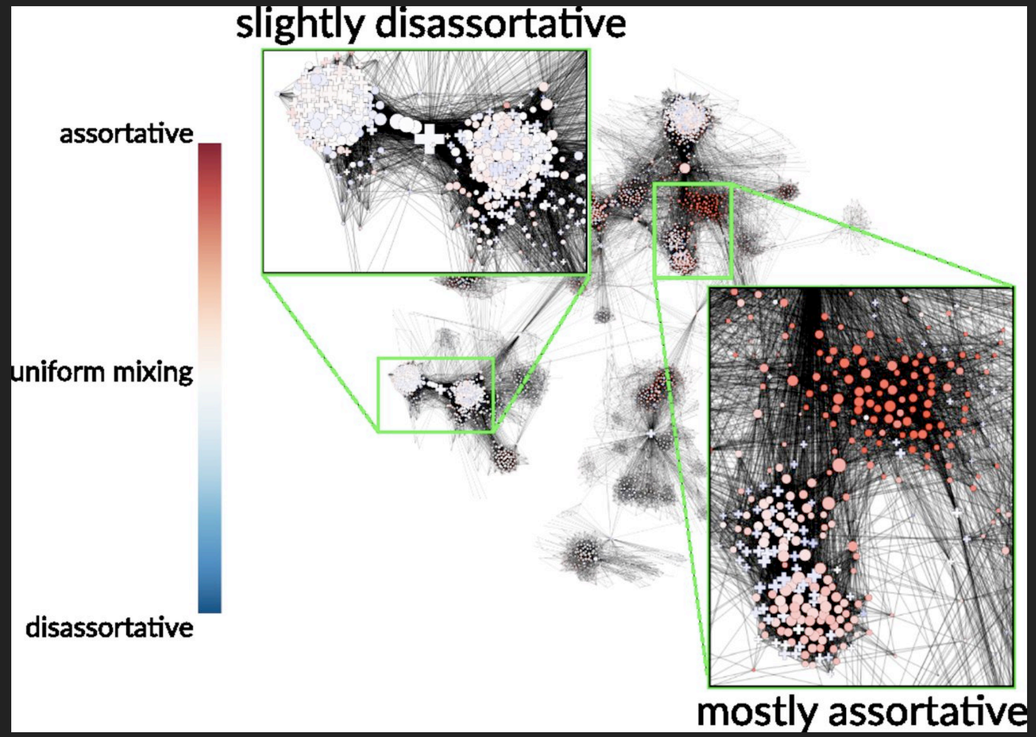


DO YOU ACTUALLY NEED A NETWORK?

- Nobody wants to see your hairball
- Show if:
  - Small networks
  - Show a macro-pattern



Florencian families



Peel et al

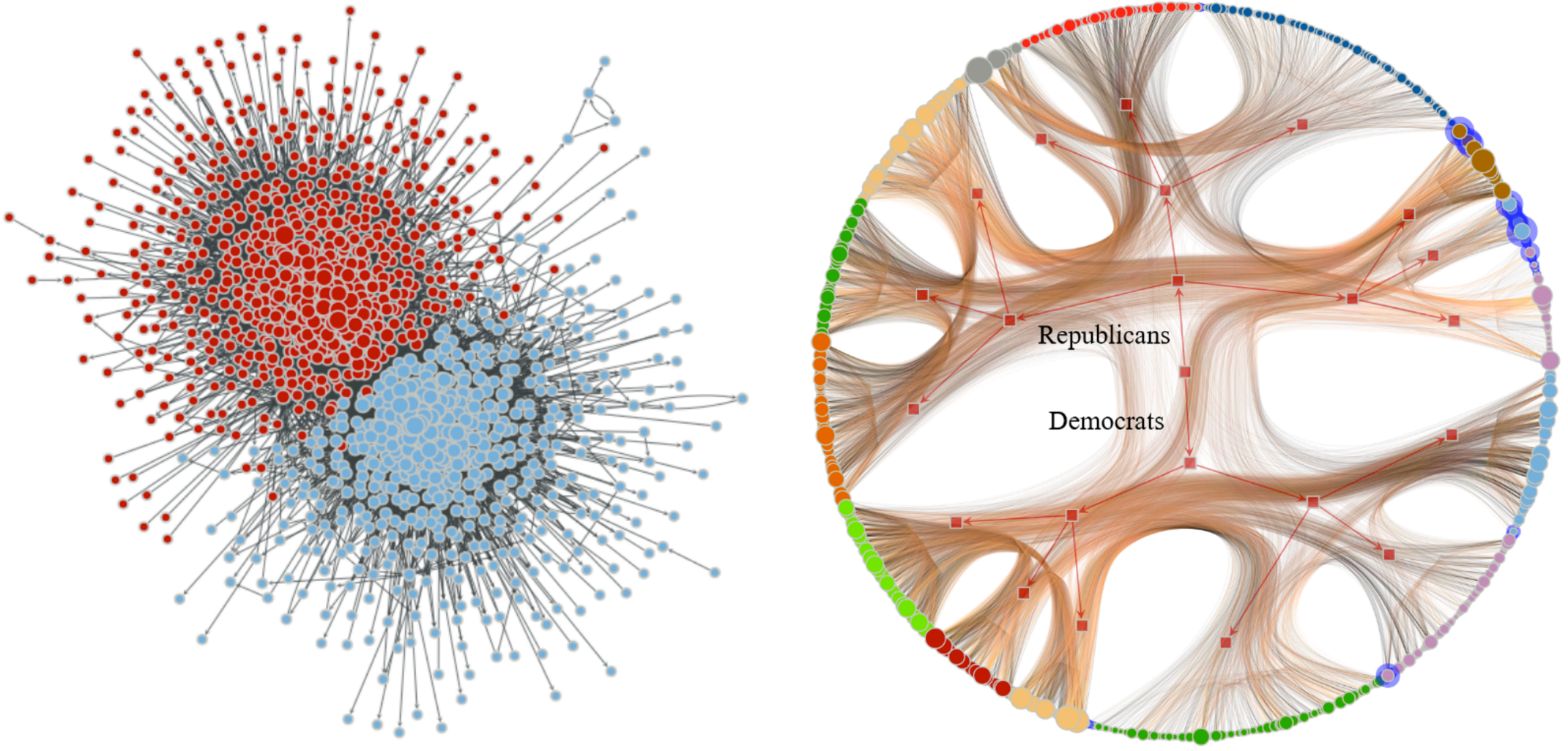


Figure 5. The political blog network of Adamic and Glance [67]. Left: Topmost partition of the hierarchy inferred with

HOW TO USE GRAPHICAL OBJECTS

- We know that length is perceived biased
- Type of plots:
  - One numerical variable: **barplot** or histogram
  - Two numerical variables: **scatter** plot
  - Three numerical variables: scatter plot + **bubble size**
  - Use color for categorical variables (more on this later)

ELEMENTS OF A PLOT:

SCALES, FACETS, TRANSFORMATIONS

SCALES, FACETS, TRANSFORMATIONS

- **Scales:** When to use logarithmic scale
  - Represent ratios or percentages (1:2 and 2:1 are equidistant from 1:1 in a log-scale)
  - Increase visibility (too many values with small values)
  - Show that our distribution follows a exponential (lin-log scale), lognormal (log-lin scale) or power-law (log-log scale) distribution
- **Facets** –> Comparing same scale is better
- **Transformations** –> Jitter, etc

PRINCIPLES OF VISUAL PERCEPTION

GESTALT PRINCIPLES OF VISUAL PERCEPTION

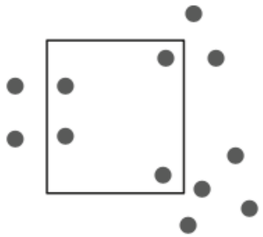
Early 1900s, how people perceive order. Principles/laws of human perception that describe how humans group similar elements, reco
patterns and simplify complex images when we perceive objects

PROXIMITY e.g. barplot: Bars next to each other are perceived as related

SIMILARITY e.g. color across plots are perceived as related (be consistent)

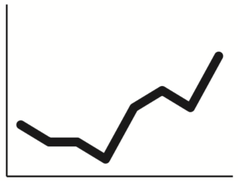
GESTALT PRINCIPLES OF VISUAL PERCEPTION

Principles of human perception that describe how humans group similar elements, recognize patterns and simplify complex images when we perceive objects



ENCLOSURE

Enclosed areas contain related objects (highlight areas)



CLOSURE

No need to close frames

GESTALT PRINCIPLES OF VISUAL PERCEPTION

Early 1900s, how people perceive order. Principles/laws of human perception that describe how humans group similar elements, reco
patterns and simplify complex images when we perceive objects

CONTINUITY e.g., No need to have a left line

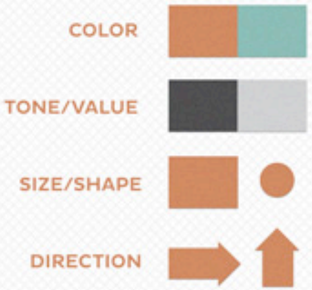
CONNECTION e.g., Networks

EXERCISE

- Using the covid data
- Think of how to show the one message
  - Which variables connect to which channels (please don’t choose a map!)
  - Which type of plot
  - Which transformations
- Create the visualization

PART 3
DESIGN

CONTRAST



Unique elements in a design should stand apart from one another. One way to do this is to use contrast. Good contrast in a design – which can be achieved using elements like color, tone, size, and more – allows the viewer's eye to flow naturally.

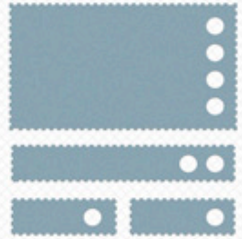
To the left, you can see 4 ways to create contrast in your design.

REPETITION

Repetition breeds cohesiveness in a design. Once a design pattern has been established – for example, a dotted border or a specific typographic styling – repeat this pattern to establish consistency.

The short version?

Establish a style for each element in a design and use it on similar elements.

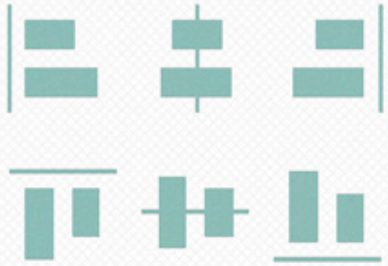


PRINCIPLES of DESIGN

quick reference poster

ALIGNMENT

Proper alignment in a design means that every element in it is visually connected to another element. Alignment allows for cohesiveness; nothing feels out of place or disconnected when alignment has been handled well.



PROXIMITY

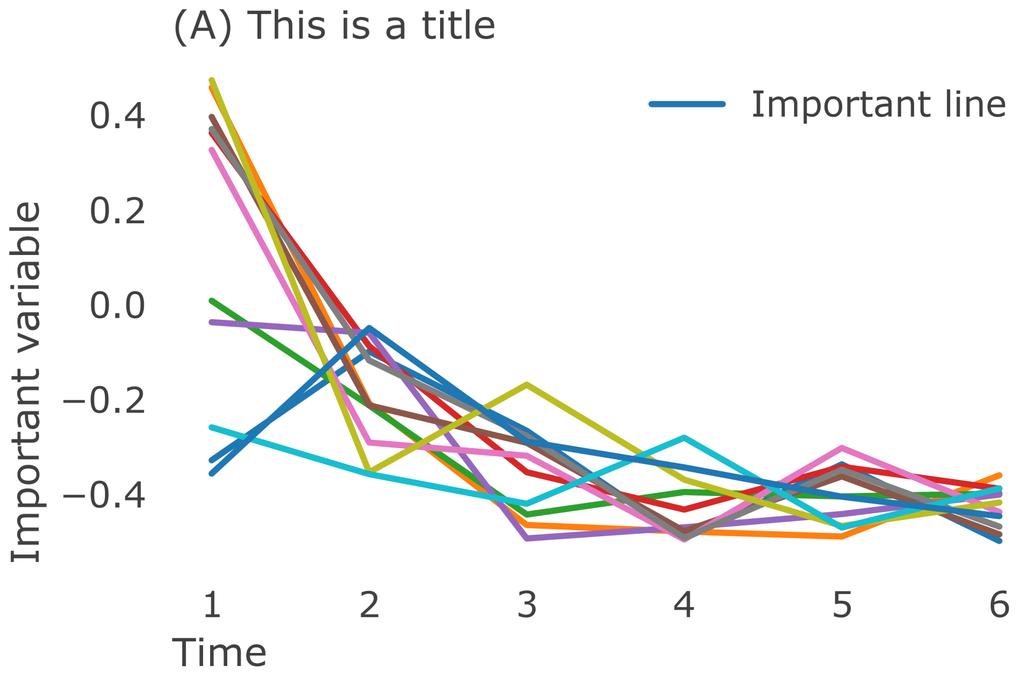
Proximity allows for visual unity in a design. If two elements are related to each other, they should be placed in close proximity to one another. Doing so minimizes visual clutter, emphasizes organization, and increases viewer comprehension.

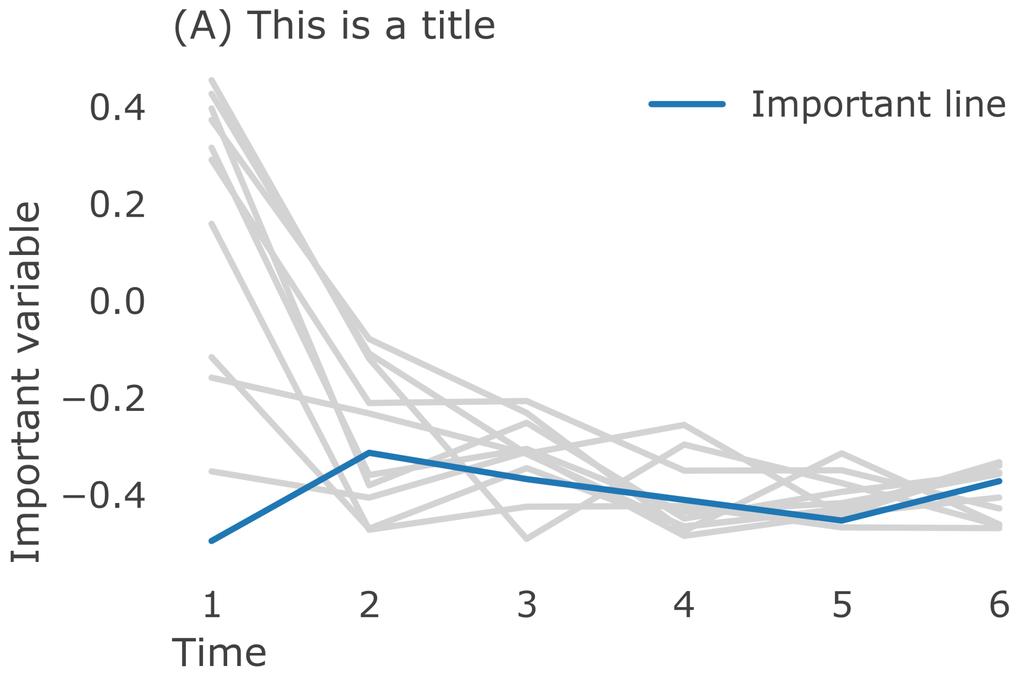
Imagine how ridiculous it would be if the proximity icons on this graphic were located on the other side of this document.

a handy paper leaf resource

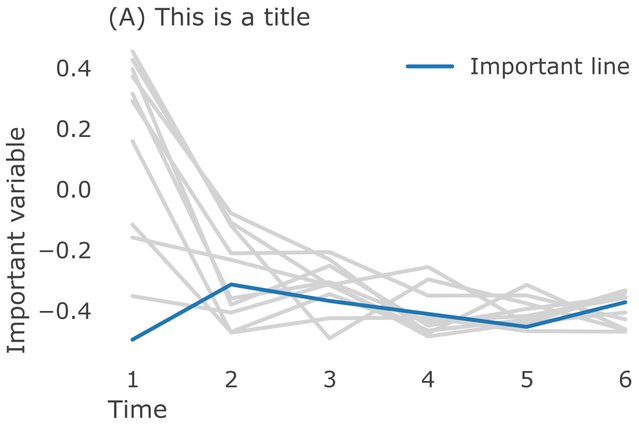
Source: Paper leaf

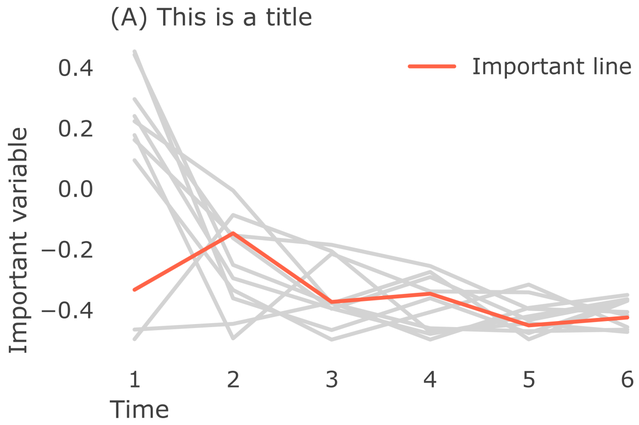
CONTRAST

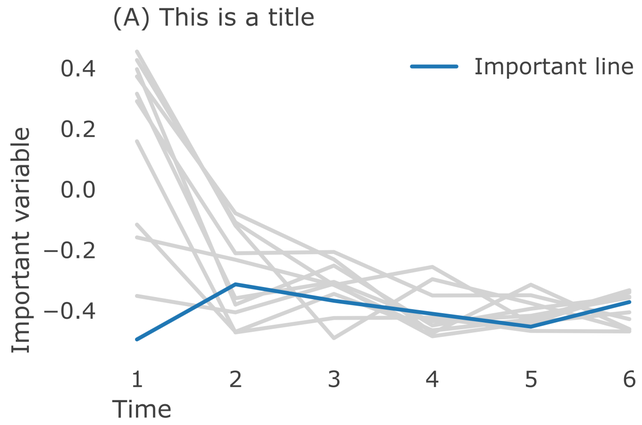

(A) This is a title
Important line
0.4
0.2
0.0
−0.2
−0.4
Important variable
1
2
3
4
5
6
Time

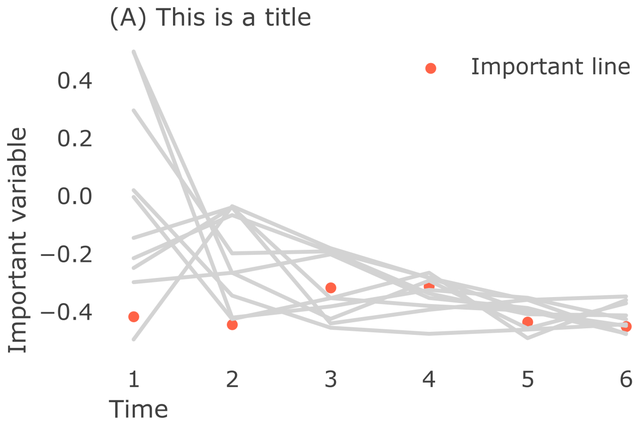

(A) This is a title
Important line
0.4
0.2
0.0
−0.2
−0.4
Important variable
1
2
3
4
5
6
Time

REPETITION

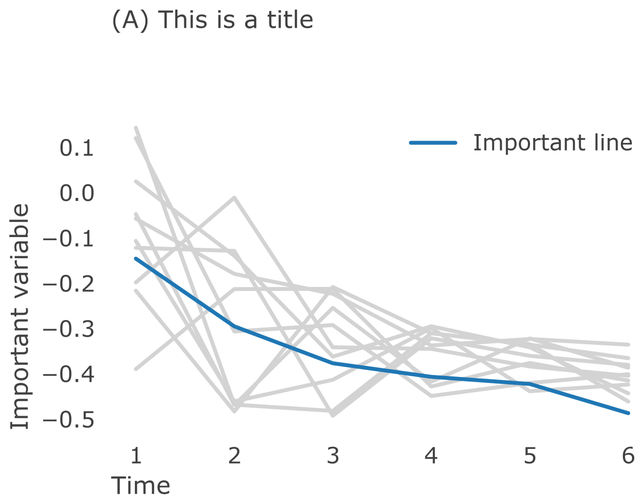

(A) This is a title
Important line
Important variable
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
6
Time

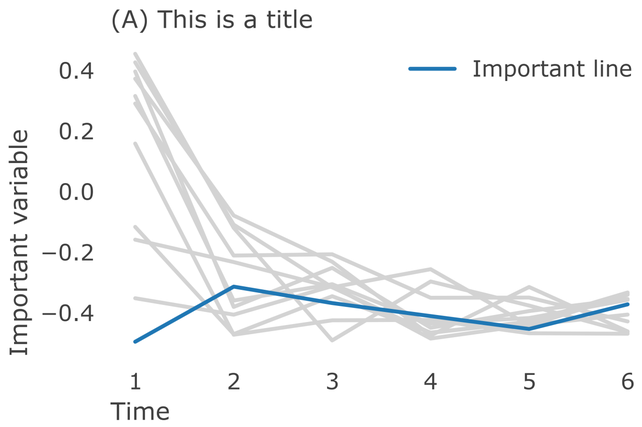

(A) This is a title
Important line
Important variable
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
6
Time

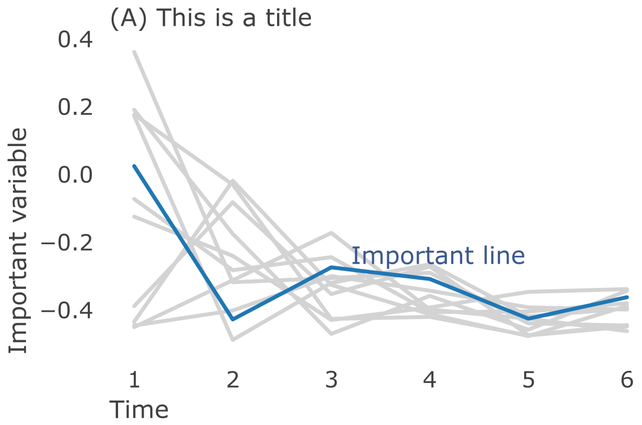

(A) This is a title
Important line
Important variable
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
6
Time


(A) This is a title
Important line
Important variable
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
6
Time

PROXIMITY


(A) This is a title
Important line
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
Important variable
1
2
3
4
5
6
Time

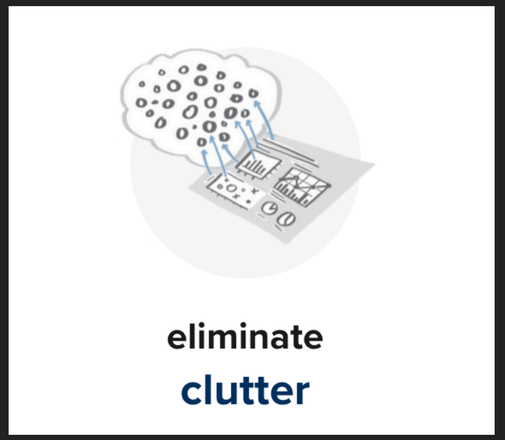

(A) This is a title
Important line
0.4
0.2
0.0
−0.2
−0.4
Important variable
1
2
3
4
5
6
Time


(A) This is a title
0.4
0.2
0.0
−0.2
−0.4
Important variable
Important line
1
2
3
4
5
6
Time

ALIGNMENT

LOOKS GOOD

NOT SO GOOD



ELIMINATE CLUTTER:

PRACTICAL GUIDES

ELIMINATE CLUTTER = MAXIMIZE INK-TO-DATA RATIO

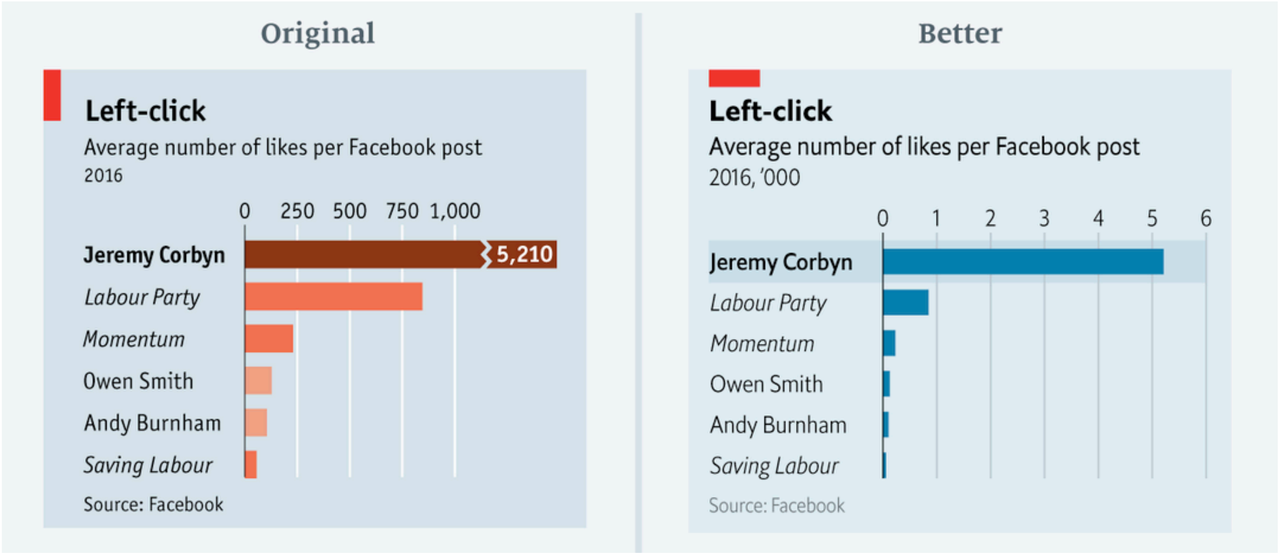
- Reduce cognitive load
  - Removing unnecessary clutter
  - Looking more professional/aesthetically pleasant

ELIMINATE CLUTTER = MAXIMIZE INK-TO-DATA RATIO

- Contrast:
  - Eliminate unnecessary lines (all frames, use gray grid lines, etc)
  - Don’t use a gray background
  - White space is your friend (allows for “breathing”)
  - Enlarge the labels
  - Use vector graphics (svg/pdf/eps) to avoid blurry figures
- Repetition: Be consistent in different figures
- Alignment: Make sure you align subplots/labels
- Proximity: When possible, label data directly (instead of using legends)

ELIMINATE CLUTTER DOESN'T MEAN ELIMINATE CONTEXT

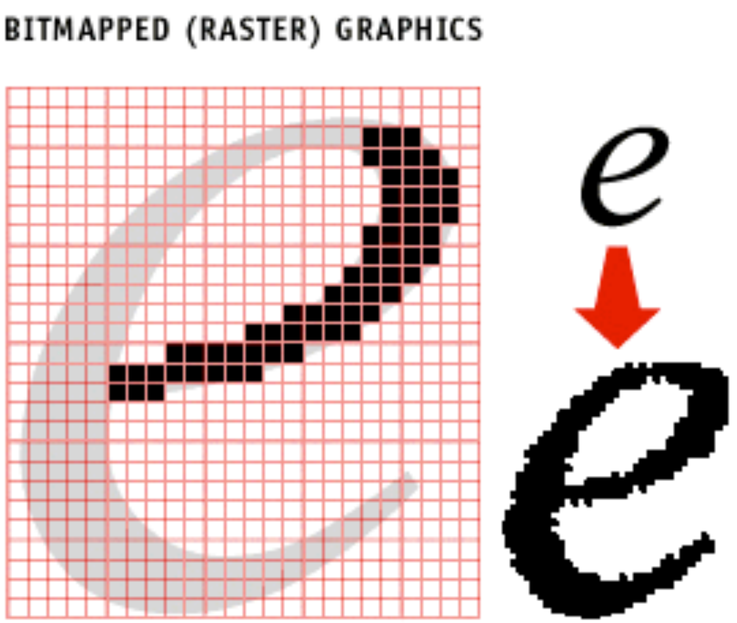
- Show the data: The axis should reflect the range of the data (no need to start at zero!)
- Show the context: Variability in the data (error bars).
- Don't use double scales

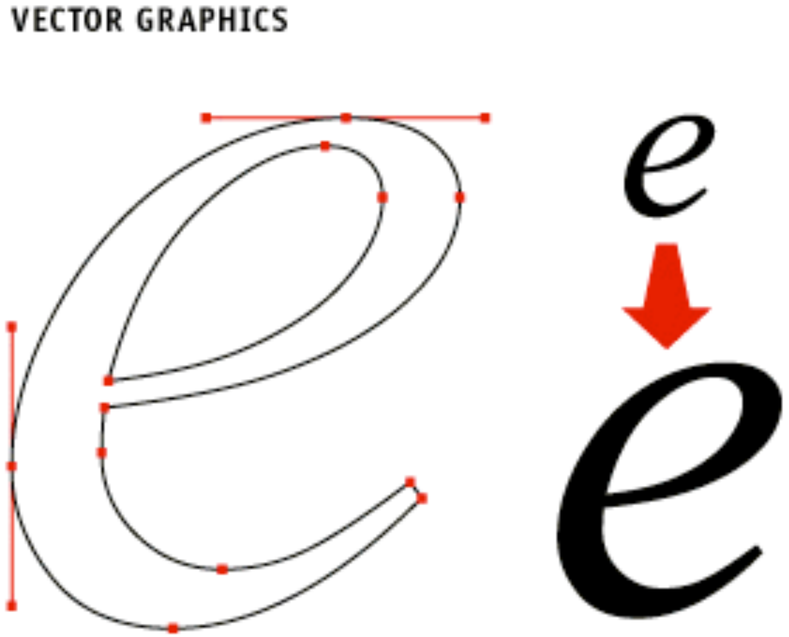


ELIMINATE BLURRY FIGURES: USE VECTOR GRAPHICS

Raster: GIF, JPEG, PNG

Vector: SVG, PDF, EPS



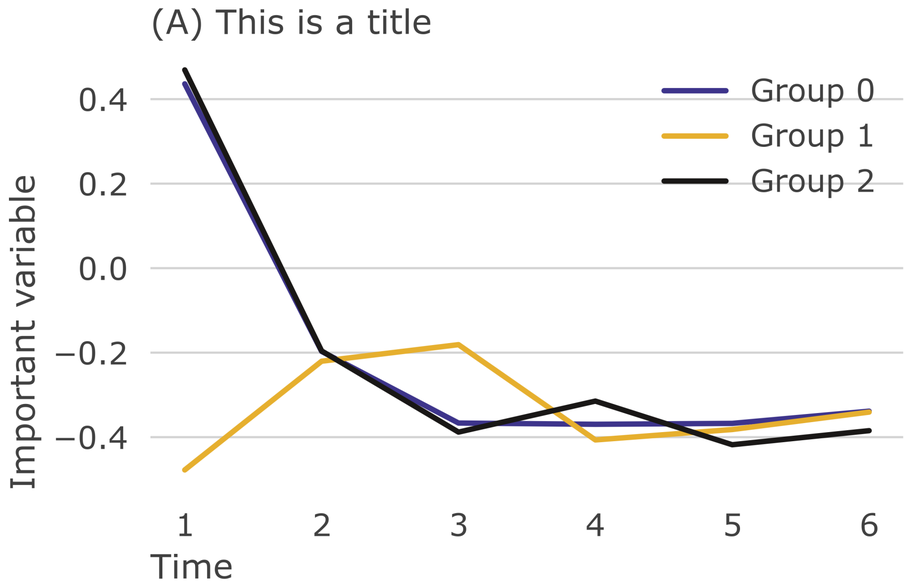


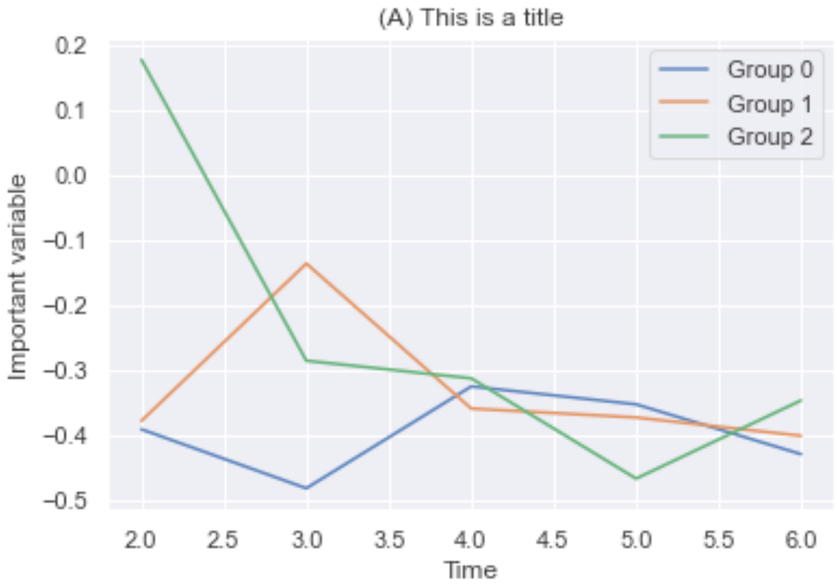
Source: https://signalizenj.wordpress.com/2015/01/29/vector-vs-raster/

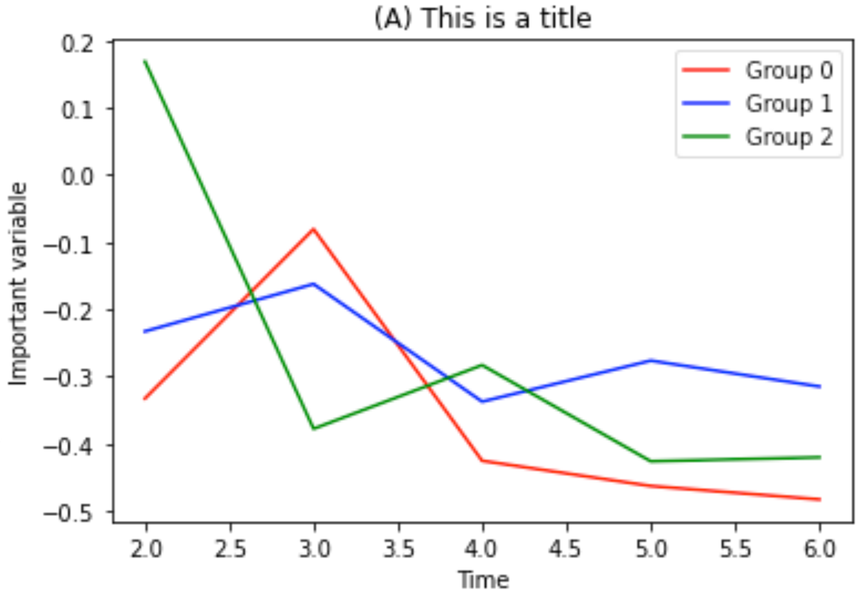
PYTHON (MATPLOTLIB/SEABORN)

```
custom_params = {"axes.spines.right": False, "axes.spines.top": False, "axes.spines.left": False, "axes.spines.bottom": False,"lines.linewidth": 2, "grid.color": "lightgray", "legend.frameon": False, "xtick.labelcolor": "#484848", "ytick.labelcolor": "#484848", "xtick.color": "#484848", "ytick.color": "#484848","text.color": "#484848",  "axes.labelcolor": "#484848", "axes.titlecolor":"#484848","figure.figsize": [5,3], "axes.titlelocation":"left","xaxis.labellocation":"left","yaxis.labellocation":"bottom"}

palette = ["#3d348b","#e6af2e","#191716","#e0e2db"] #use your favourite colours

sns.set_theme(context='paper', style='white', palette=palette, font='Verdana', font_scale=1.3, color_codes=True, rc=custom_params)
```

TEXT





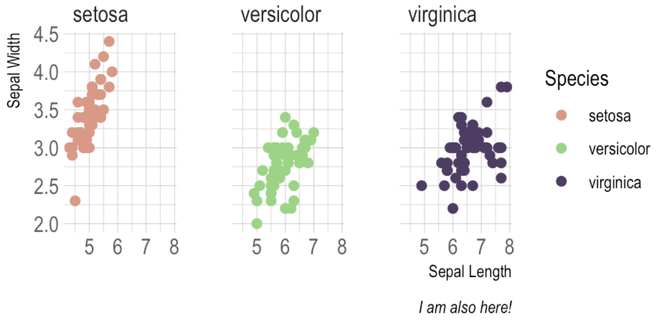


Closure: No need for a frame, we understand that it is one plot
Proximity: We know that the labels belong to the axes because they are nearby
Continuity: 2, 3, 4, 5, 6 are perceived as connected
Connection: All dots within one line
Similarity: The three lines represent similar data

GGPLOT2 —> HBRTHEMES::THEME_IPSUM

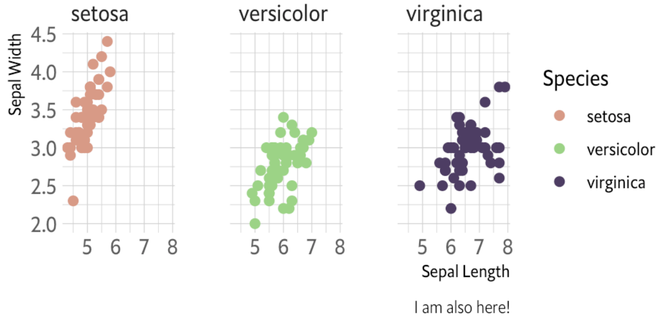
theme_ipsum

Arial Narrow



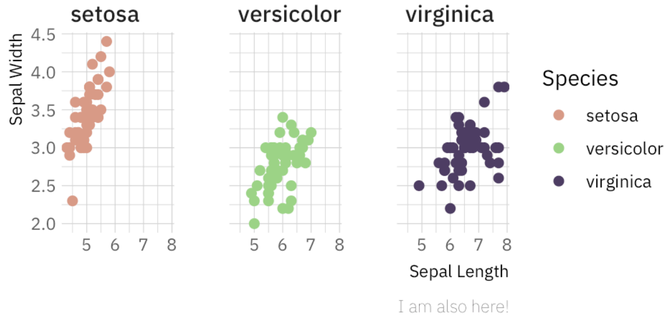
theme_ipsum_es

Econ Sans



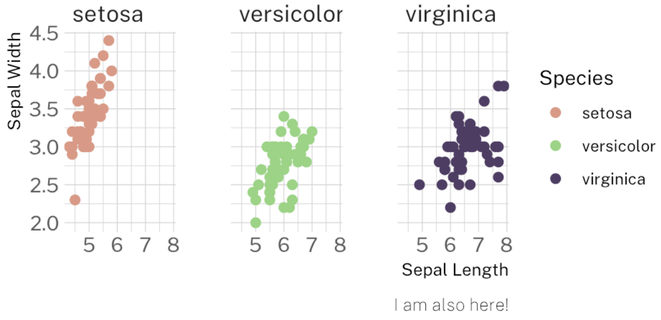
theme_ipsum_ps

IBM Plex Sans



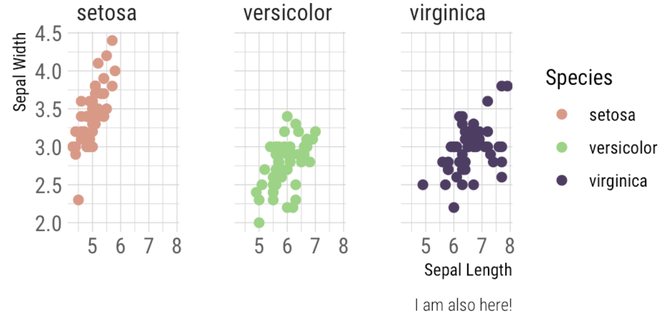
theme_ipsum_pub

Public Sans



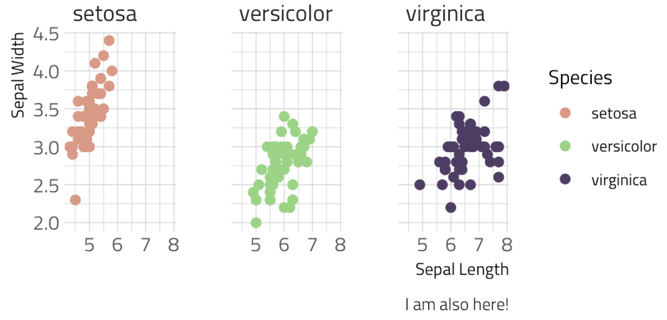
theme_ipsum_rc

Roboto Condensed



theme_ipsum_tw

Titillium Web



FOCUS ATTENTION:

PRACTICAL GUIDES

WHY DO WE NEED TO FOCUS ATTENTION?

Help the person interpret the plot (reduce cognitive load, make it enjoyable)

HOW DO WE FOCUS ATTENTION?

Using pre-attentive attributes

PRE ATTENTIVE ATTRIBUTES

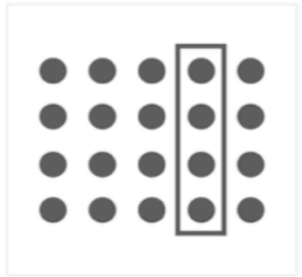
Length

Width

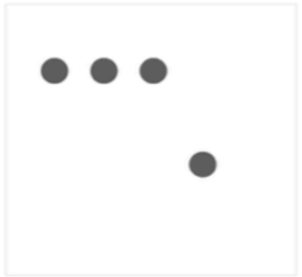
Orientation

Size

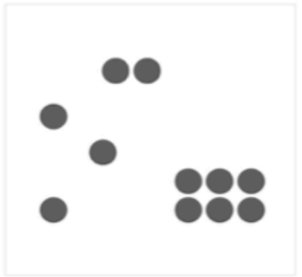
Shape



Enclosure

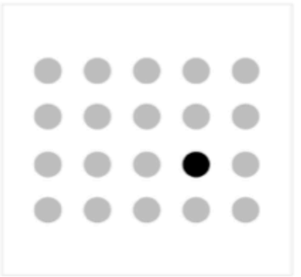


Position



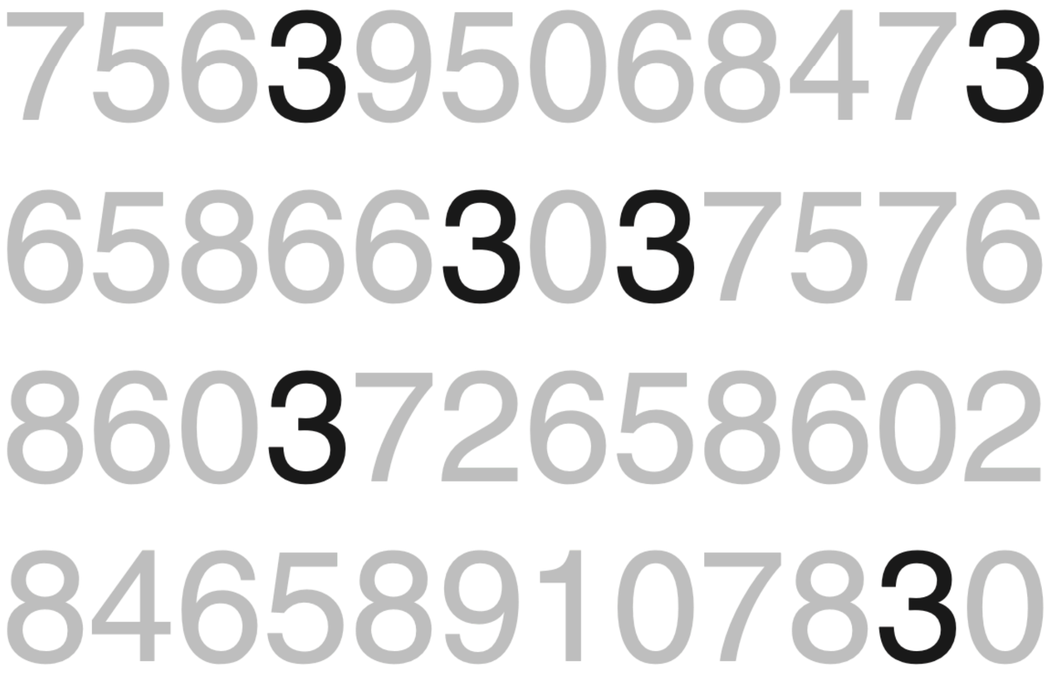
Grouping

Color Hue



Color Intensity

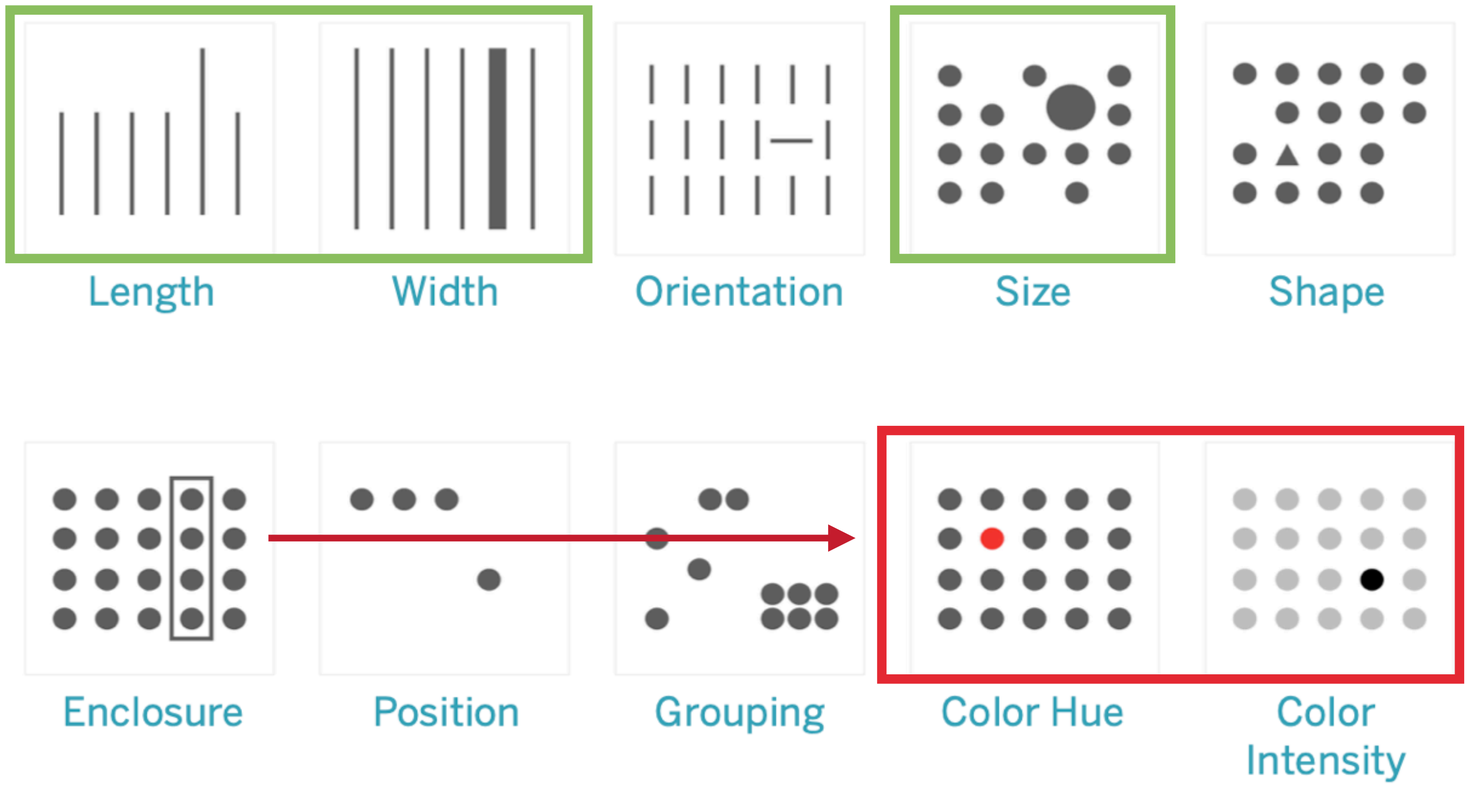

756395068473
658663037576
860372658602
846589107830

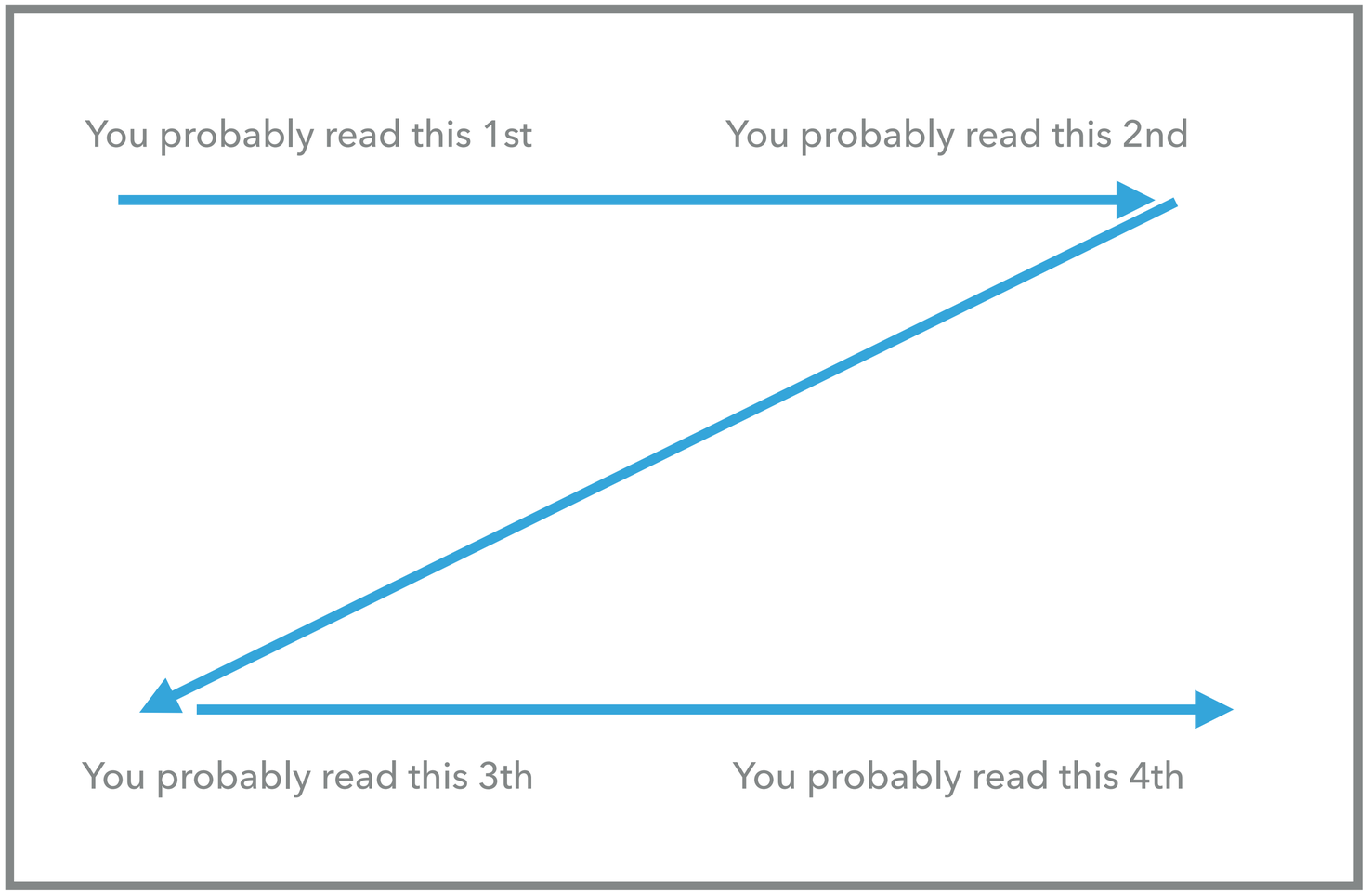

756395068473
658663037576
860372658602
846589107830

IMPORTANT PRE ATTENTIVE ATTRIBUTES

What do we focus on:

- Large objects
- Bright objects
- Contrasting objects




You probably read this 1st
You probably read this 2nd
You probably read this 3th
You probably read this 4th

Color makes ice cream taste sweeter,
veggies taste fresher,
and coffee taste richer

Ellen Lupton

COLOR

The most useful pre-attentive attribute

- Increases contrast
- Allows for consistency

Color affect emotion, culture-dependent. But some responses are nearly universal

- Warm colors --> alive/alert
- Blue colors --> calming/focus

Resources:

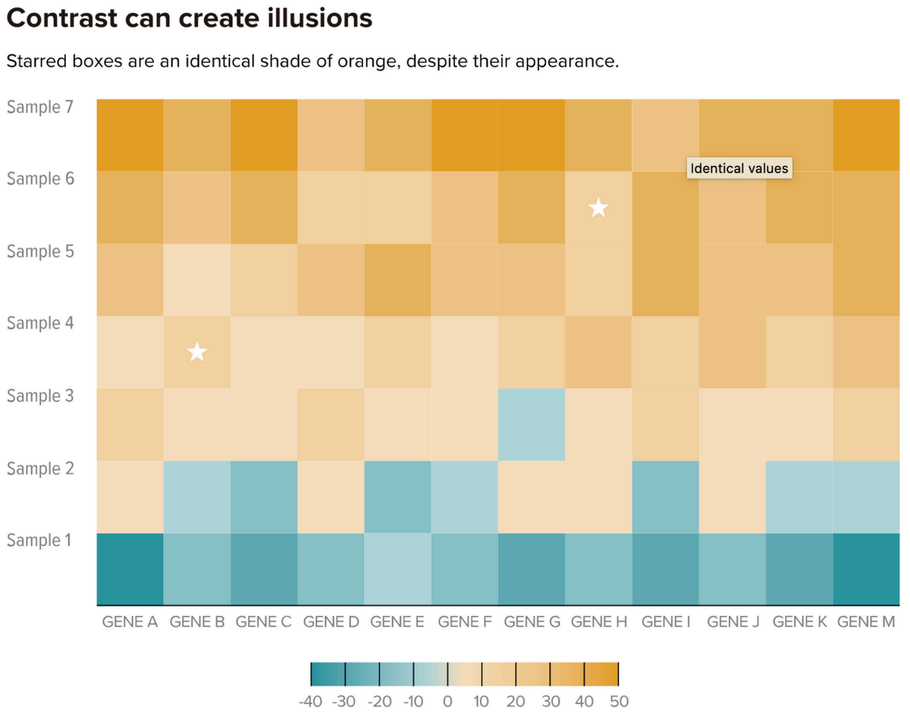
https://blog.datawrapper.de/which-color-scale-to-use-in-data-vis/

https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40

COLOR

In addition of highlighting, colours can be used to:

- Represent categories (not more than 4 colors)
- Represent values: Only if strictly needed



COLOR PALETTES

SEQUENTIAL

Minimum is important

DIVERGING

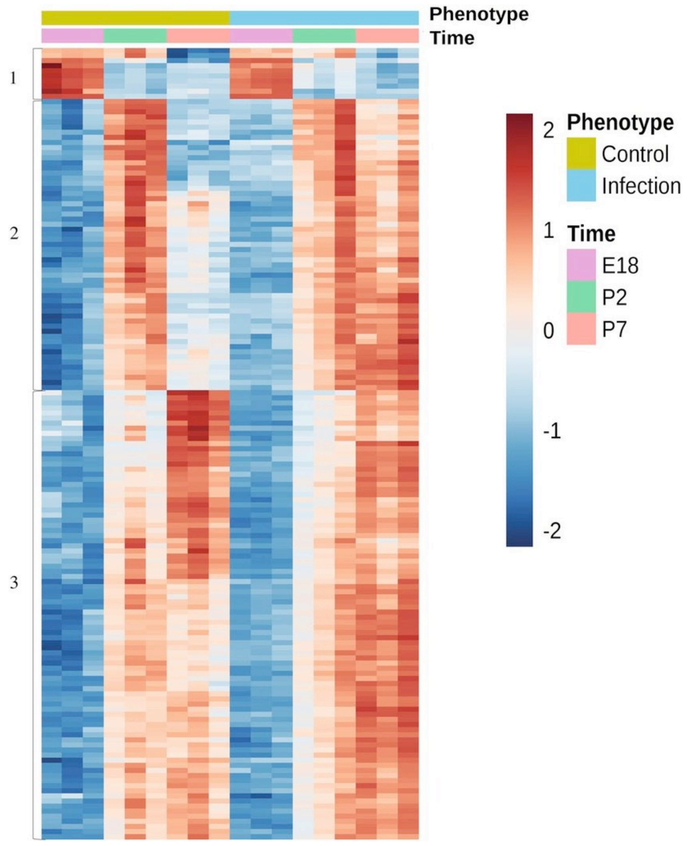
Mean is important

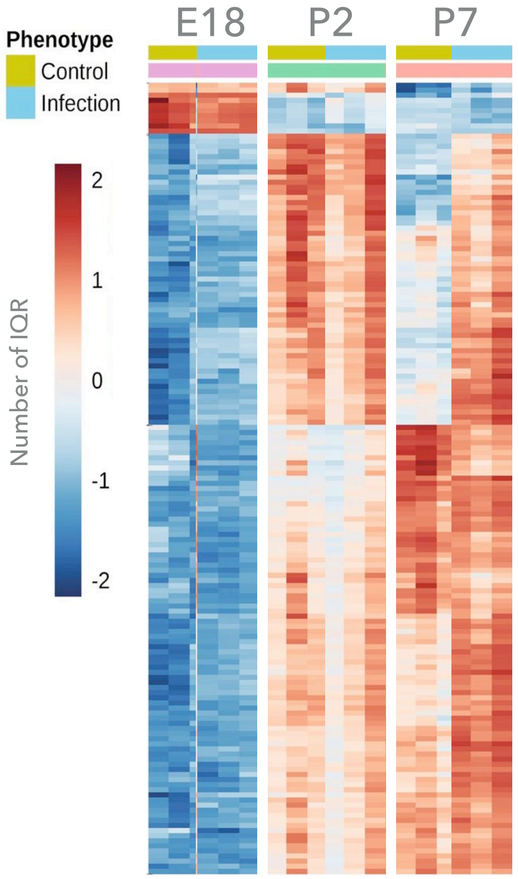
QUALITATIVE

Represent categories

COLOR TO REPRESENT VALUES

ONLY IF THE EXACT VALUES ARE NOT IMPORTANT





Haque, Koski and Scott, 2019

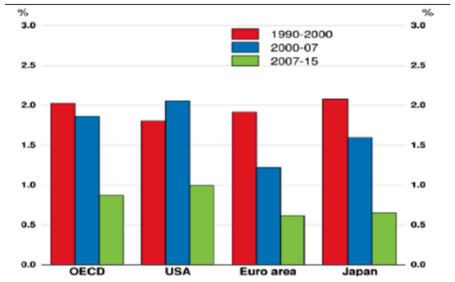
COLOR: PRACTICAL ADVISE

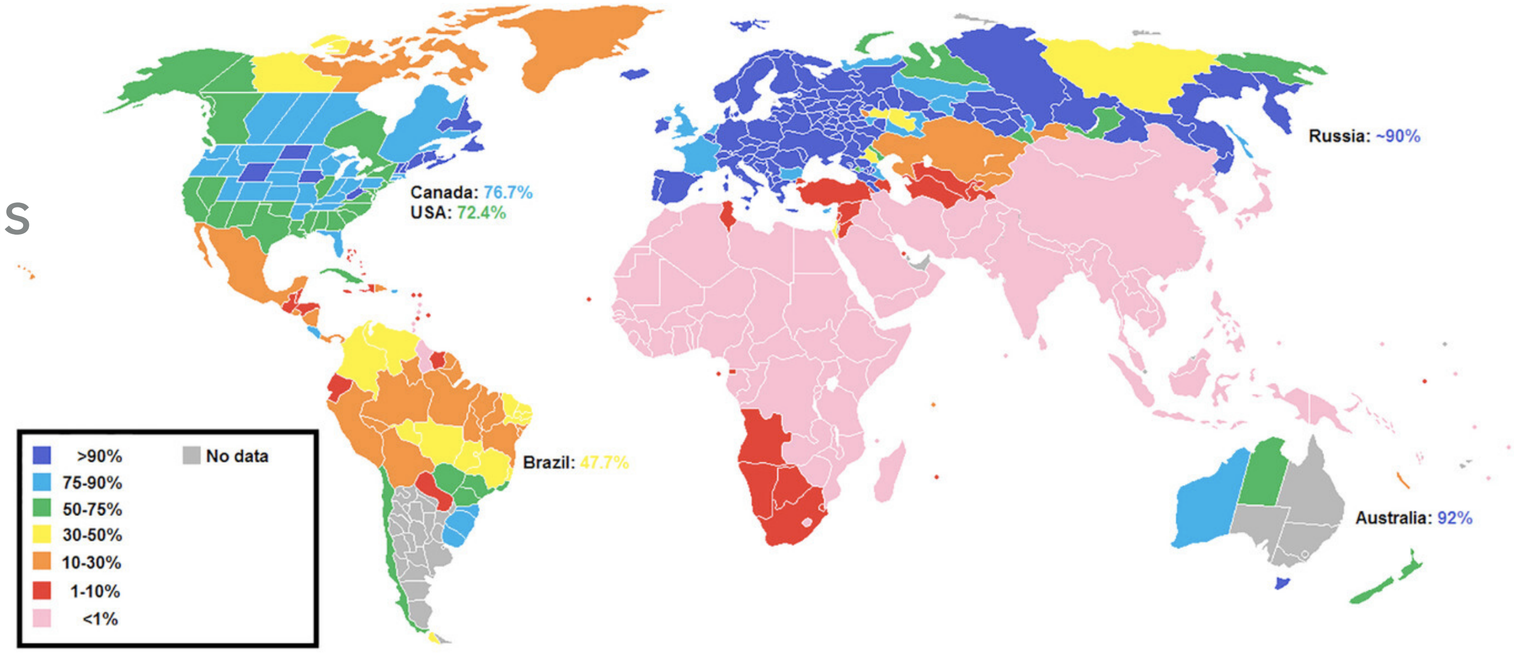
Try to be color-blind friendly (e.g. install Color Oracle to check it)

Leverage color associations

Don’t use default colours

Keep the rainbow out of graphs





COLOR: PRACTICAL ADVISE

Colors have names

–> lightgray

–> cornflowerblue

–> tomato

EXERCISE

- Declutter and emphasise the main message

PART 4

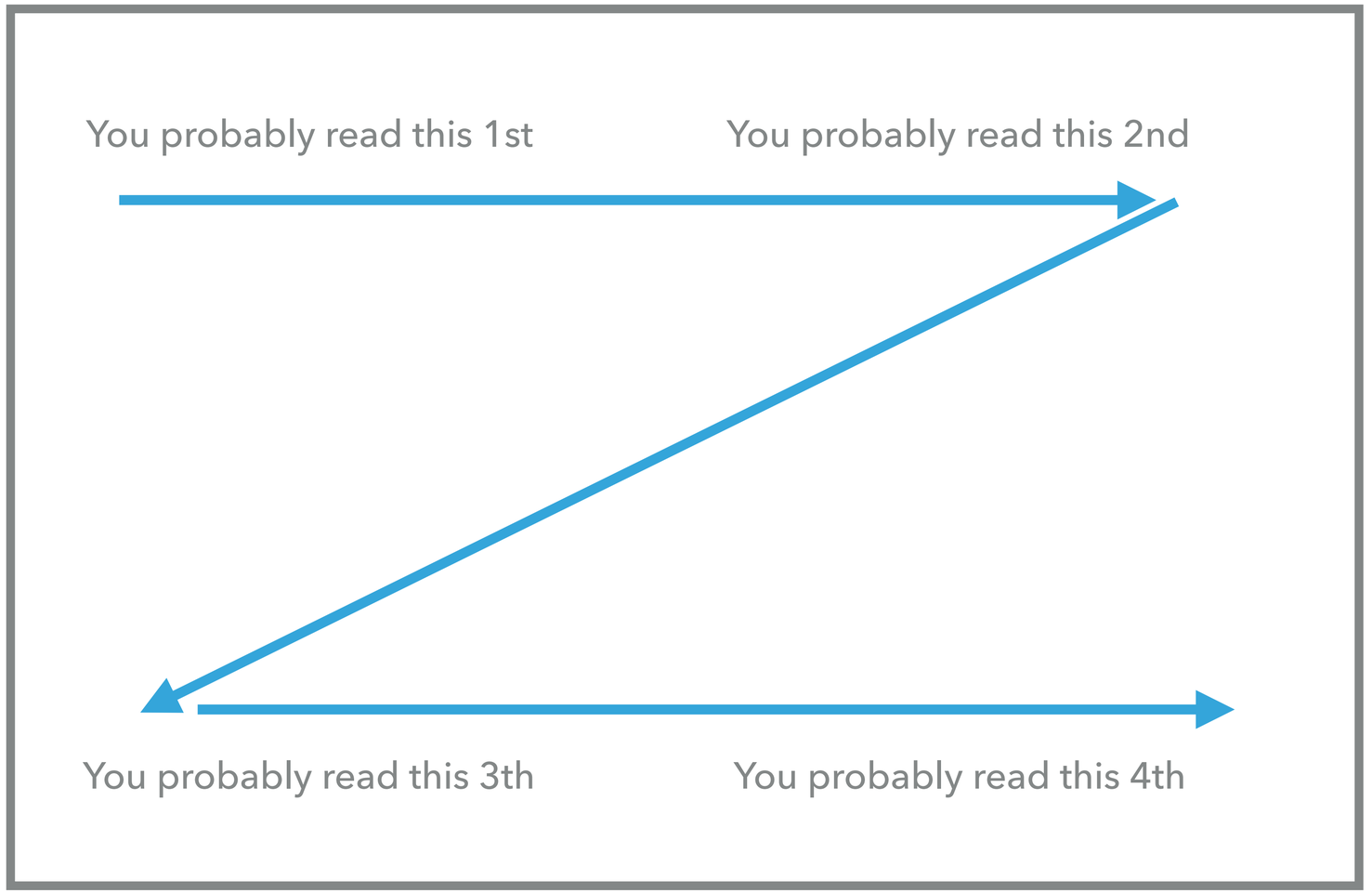
STORYTELLING

STORYTELLING

Storytelling is the **most effective tool** to make audiences enjoy a presentation, make them patient and curious to accept an idea, help them better understand an instruction, and keep them away in lectures. People love cute stuff — *Mengyan Li*

A great story does more than represent emotion from a distance. It makes us feel an emotional charge — *Ellen lupton*

HOW DO WE READ PLOTS



HOW TO TELL A STORY

Narrative arc

Exposition: What does the reader need to know to understand the plot?

- Leverage how we read plots (Z)
- Use pre attentive attributes
- Call to action (tell the reader what do do)

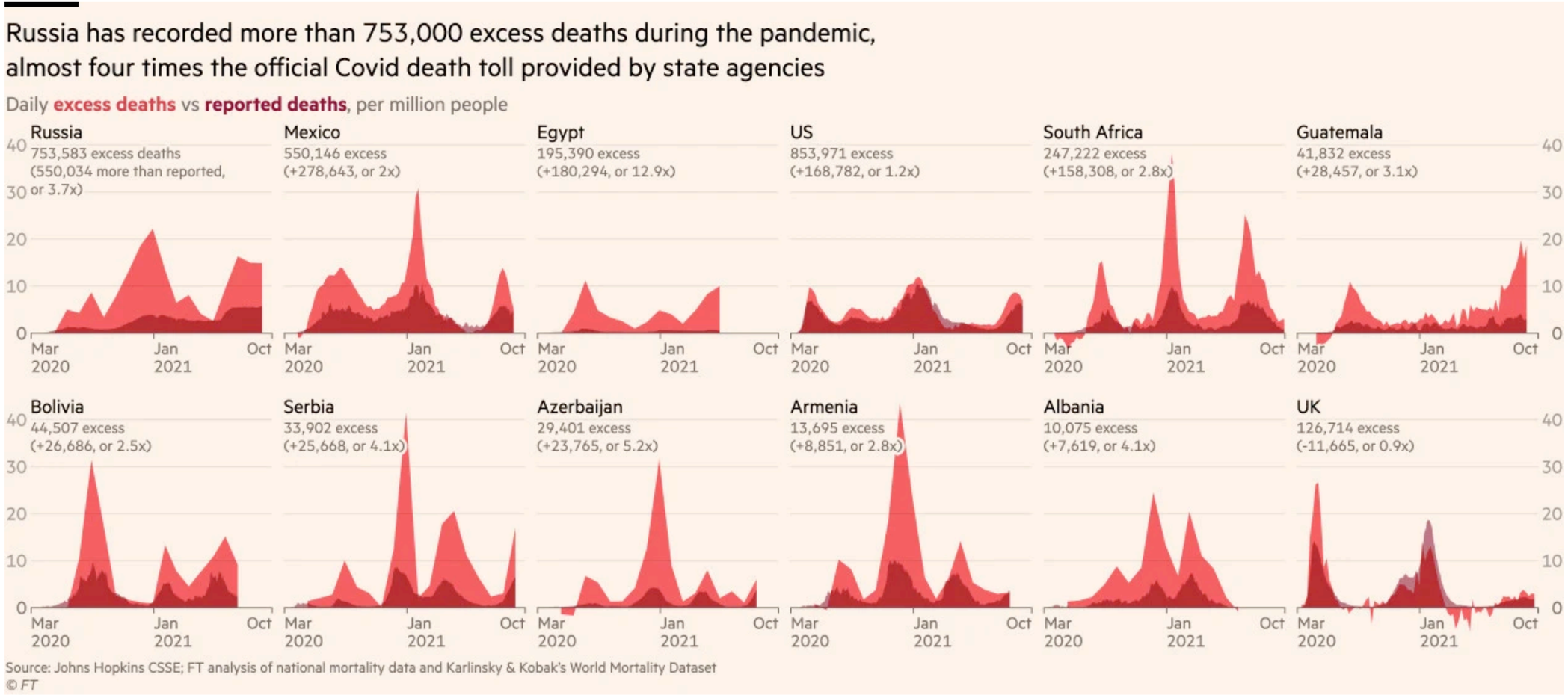
Middle: What is the point of the plot?

- Guide the attention and create emotions using color
- Based on conflict

Conclusion:

- Make sure the

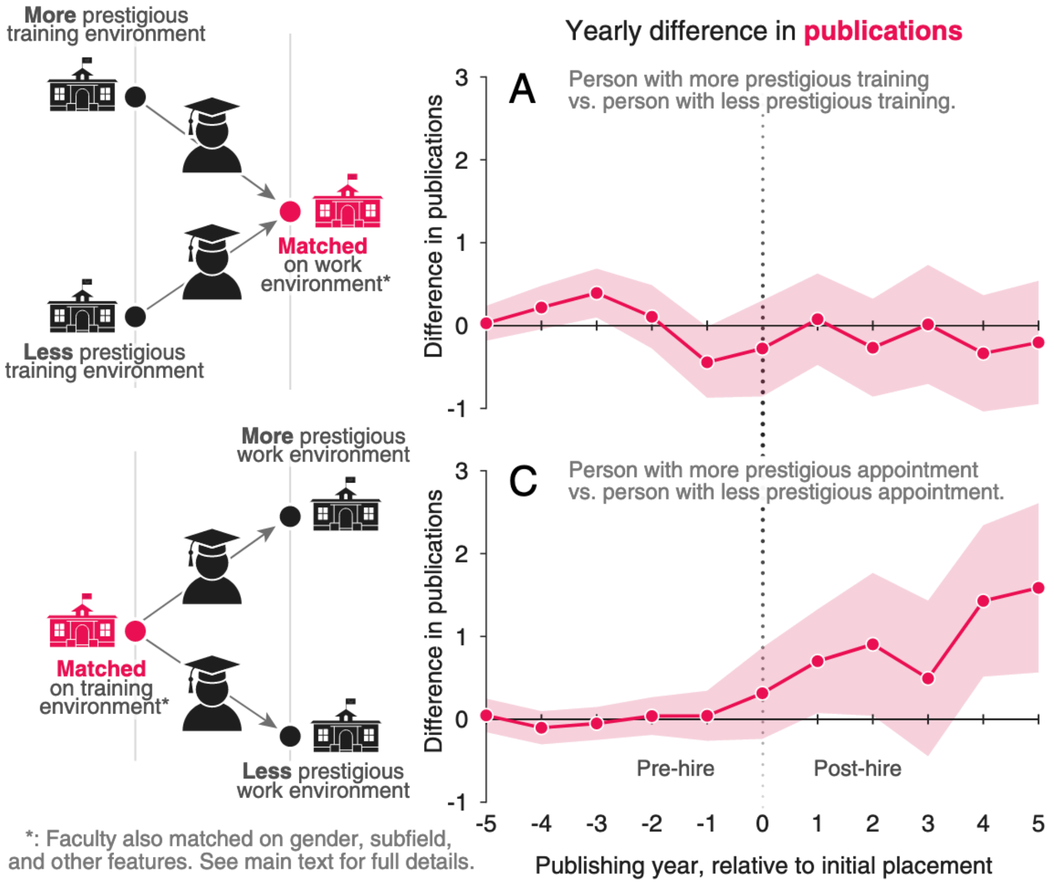
- Prototype and ask for feedback



Exposition: Label on top-left corner, tells the reader what to do

Conflict: Excess and reported deaths are very different

Resolution: The reader has understood how it looks in different countries

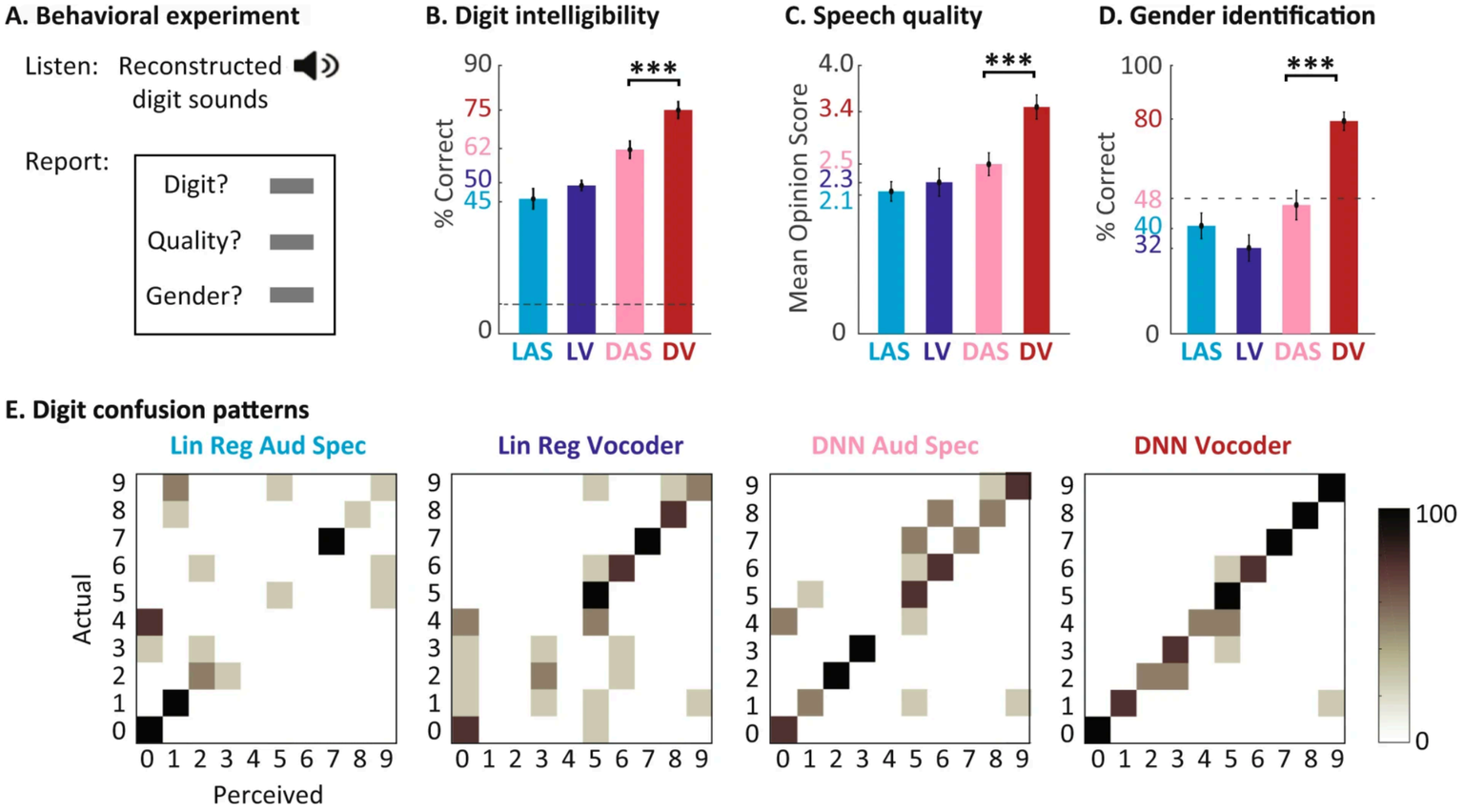


Exposition: Label and drawings.

Conflict: Between both matched persons.

Resolution: Reader has learned that prestige helps get published.

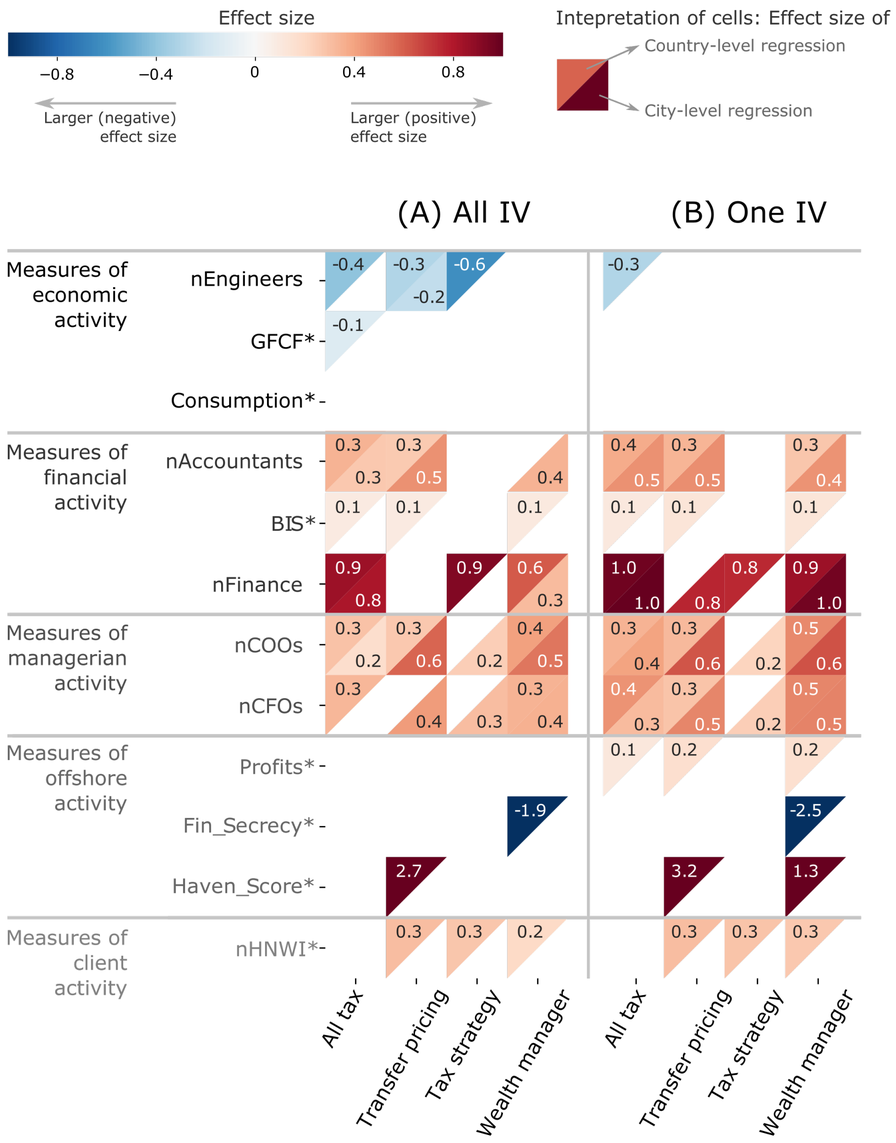
https://www.pnas.org/content/pnas/116/22/10729.full.pdf



Exposition: (A) tells you how to read the rest of the plot

Conflict: Between the four methods

Resolution: DNN Vocoder much better



Exposition: Top-left tells you how to read the rest of the plot

Conflict: Between economic, financial, managerial and offshore

Resolution: Financial and managerial activities overrepresented

EXERCISE

Think about the personas who may read the plot: Archetypal users with specific needs, e.g. researcher in neuroscience, casual reader vs reviewer.

How will your graph be perceived by them?What moods and emotions might users experience as they engage with your work? (not frustrated!)

How can you guide them in the interpretation?

EXERCISE

- Using the covid data
- Create a full figure & story, think where to put the legend, plots, how to use color, which color, etc.

SUMMARY

Think about the *who* and *what* questions

Write down the *main message* of the plot

Map your data to the possible *channels*

Pay attention to the *CRAP* (contrast / repetition / alignment / proximity)

Guide the user:

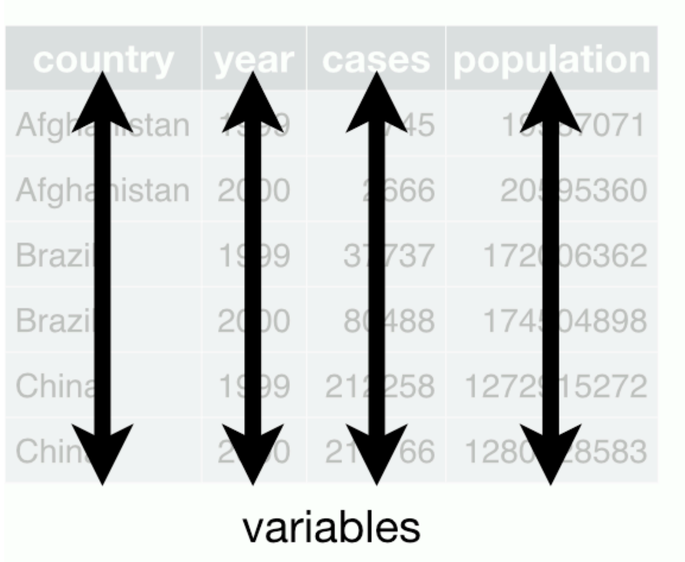
- Preattentive attributes to create structure
- Storytelling

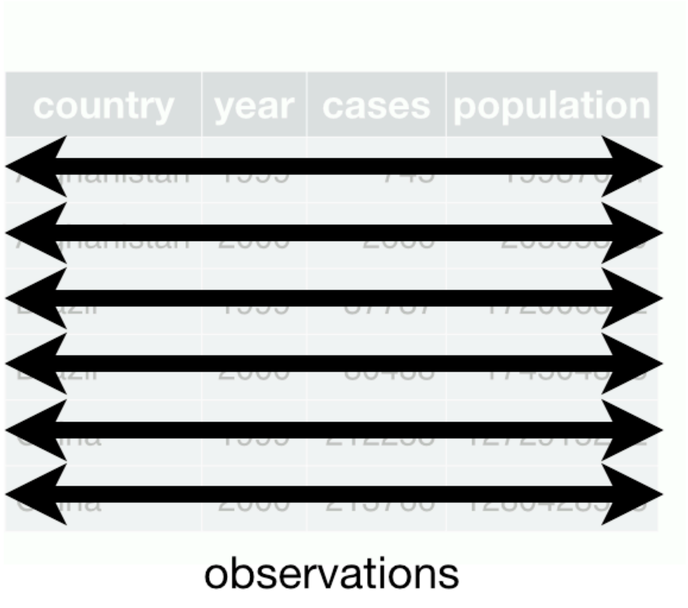
Iterate and ask for feedback

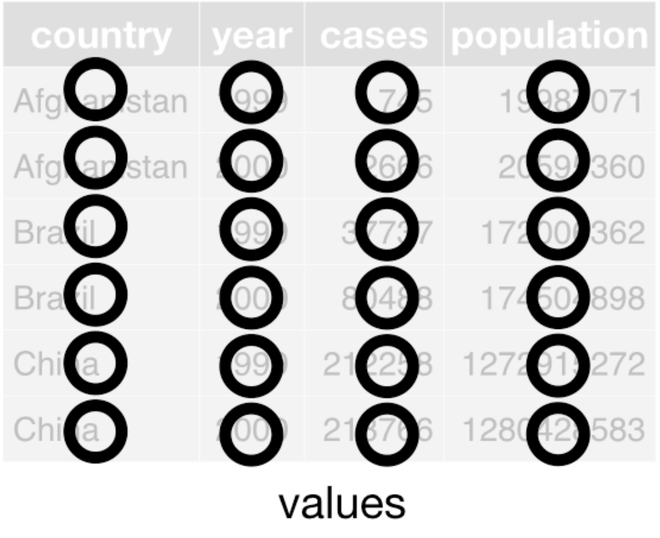
PART 5
PROGRAMMING

TIDY DATA

Use tidy data from the first day







MATPLOTLIB

Control what goes on top → *plt.plot([1,2,3], [1,4,2], zorder=0) –> background (9 = top)*

Label data directly → *plt.text(x_coord, y_coord, "Text"). Combines with the adjustText library*

Remove gridlines → *ax.grid(axis=None) or ax.grid(axis="x")*

Add $ or thousands separator to the axis: →

```
import matplotlib as mpl
ax.yaxis.set_major_formatter(mpl.ticker.StrMethodFormatter('{x:,.0f}'))
```

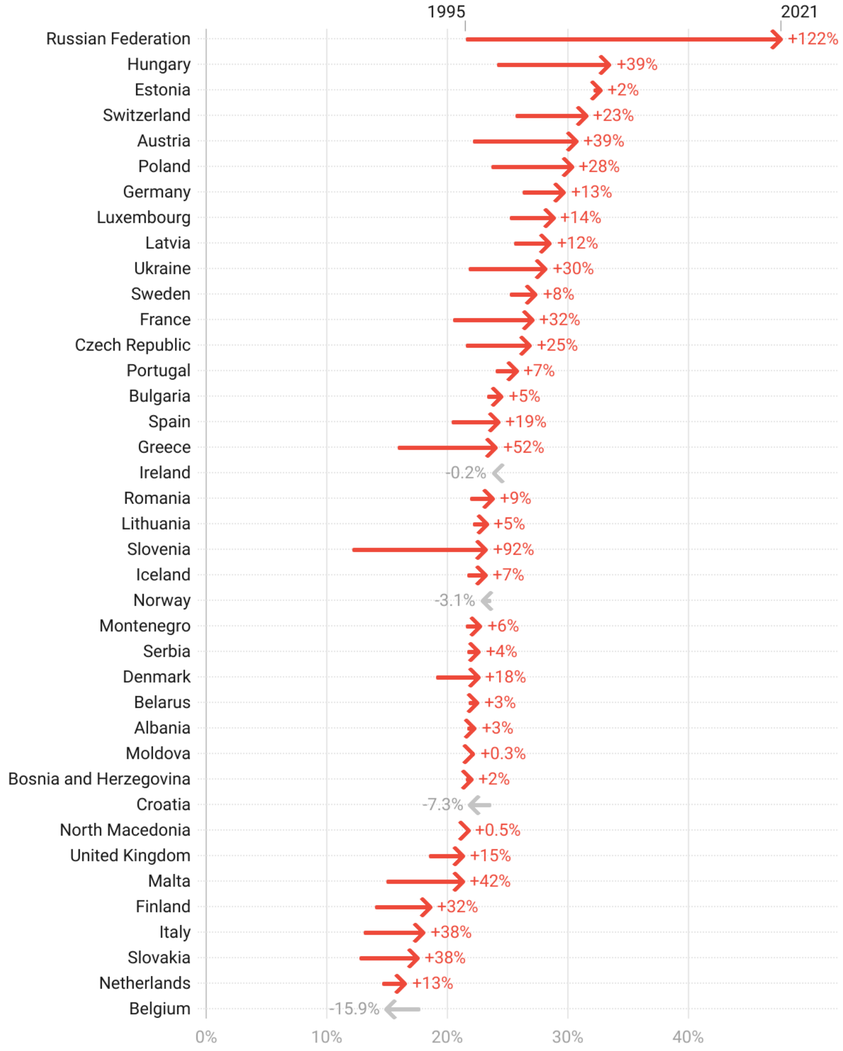
Consistent color →

```
colors_hue = {"Netherlands": "orange", "Spain": "tomato" …}
sns.relplot(x="time", y="value", hue="country', palette=colors_hue)
```

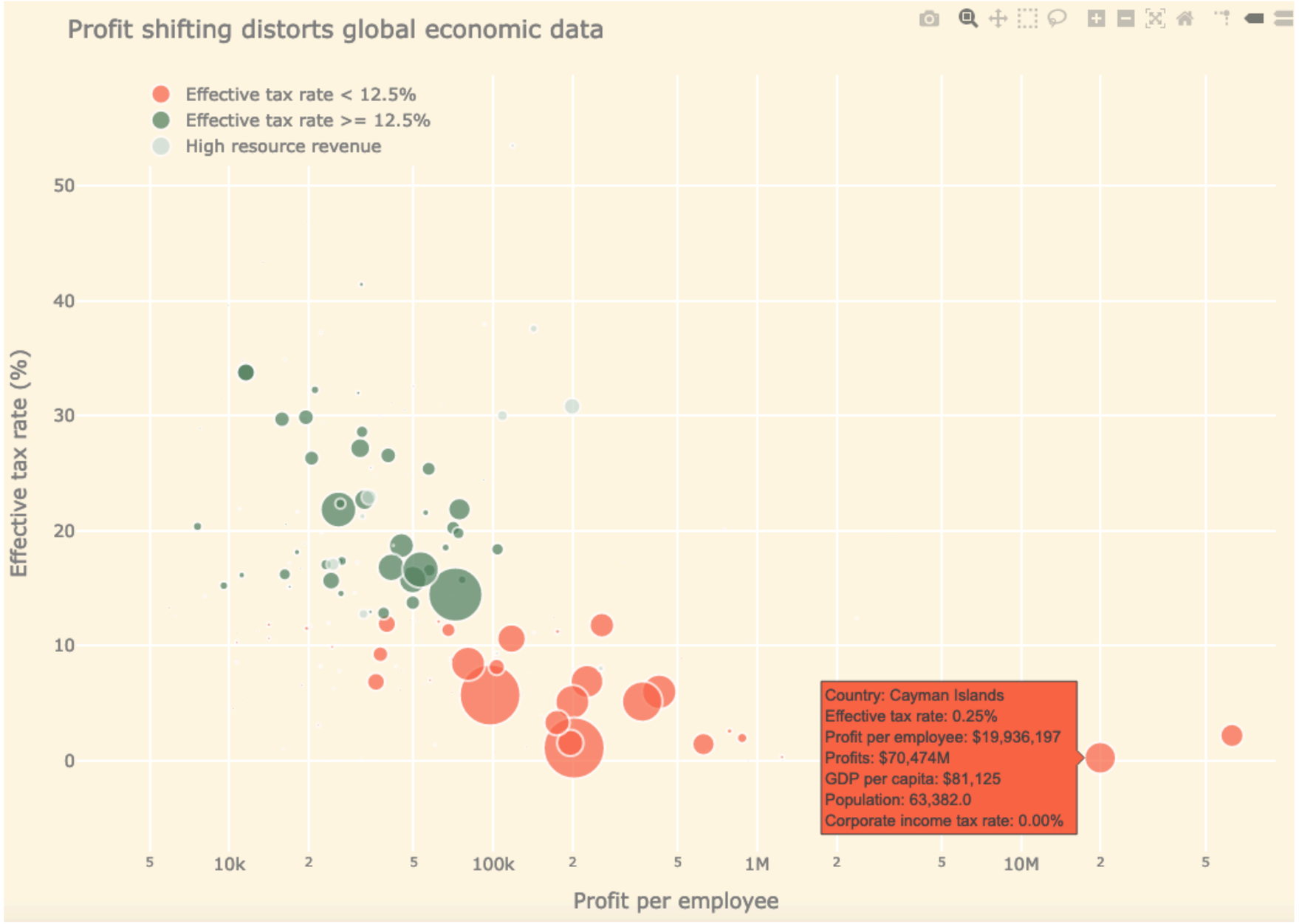
Do as much as possible programming (faster in the long run)

DATAWRAPPER

Increase in wealth inequality since 1995



PLOTLY


Profit shifting distorts global economic data
Effective tax rate < 12.5%
Effective tax rate >= 12.5%
High resource revenue
Effective tax rate (%)
50
40
30
20
10
0
5
10k
2
5
100k
2
5
1M
2
5
10M
2
5
Profit per employee
Country: Cayman Islands
Effective tax rate: 0.25%
Profit per employee: $19,936,197
Profits: $70,474M
GDP per capita: $81,125
Population: 63,382.0
Corporate income tax rate: 0.00%

INKSCAPE / ILLUSTRATOR