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

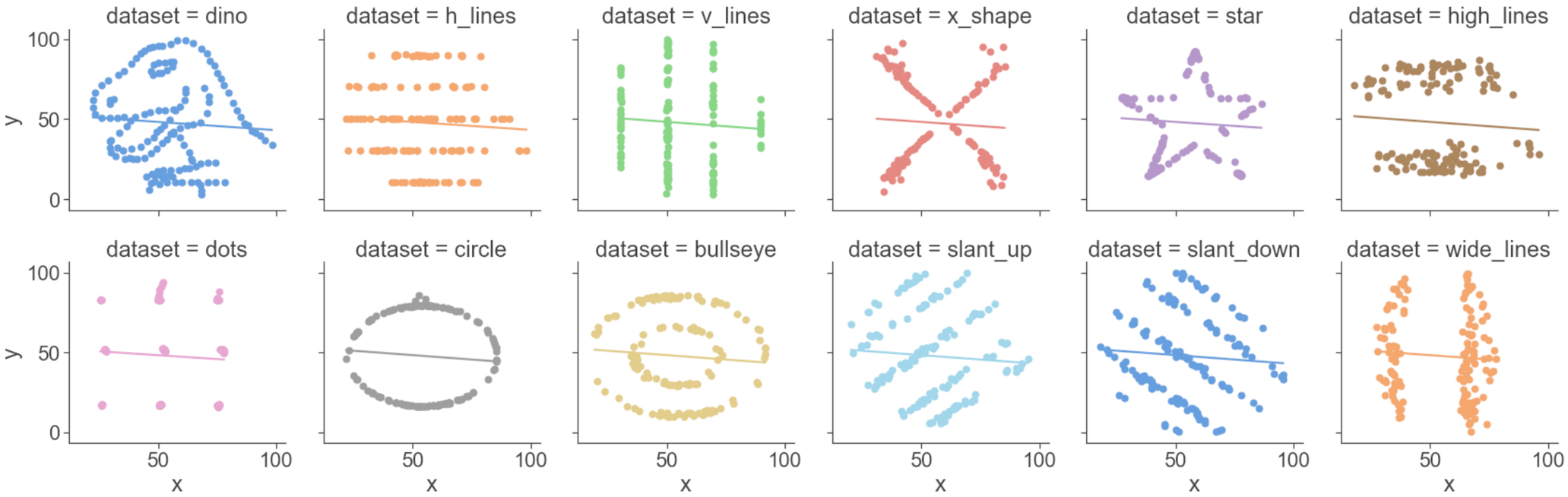
Two conditions:

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

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 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
- Good use of design and storytelling principles –> Effective

understand the **context**

UNDERSTANDING

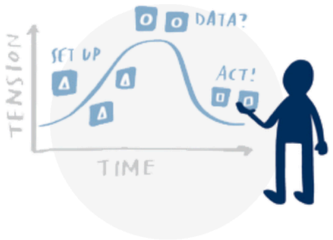
choose an **effective visual**

PERCEPTION

eliminate **clutter**

focus **attention**

DESIGN



tell a **story**

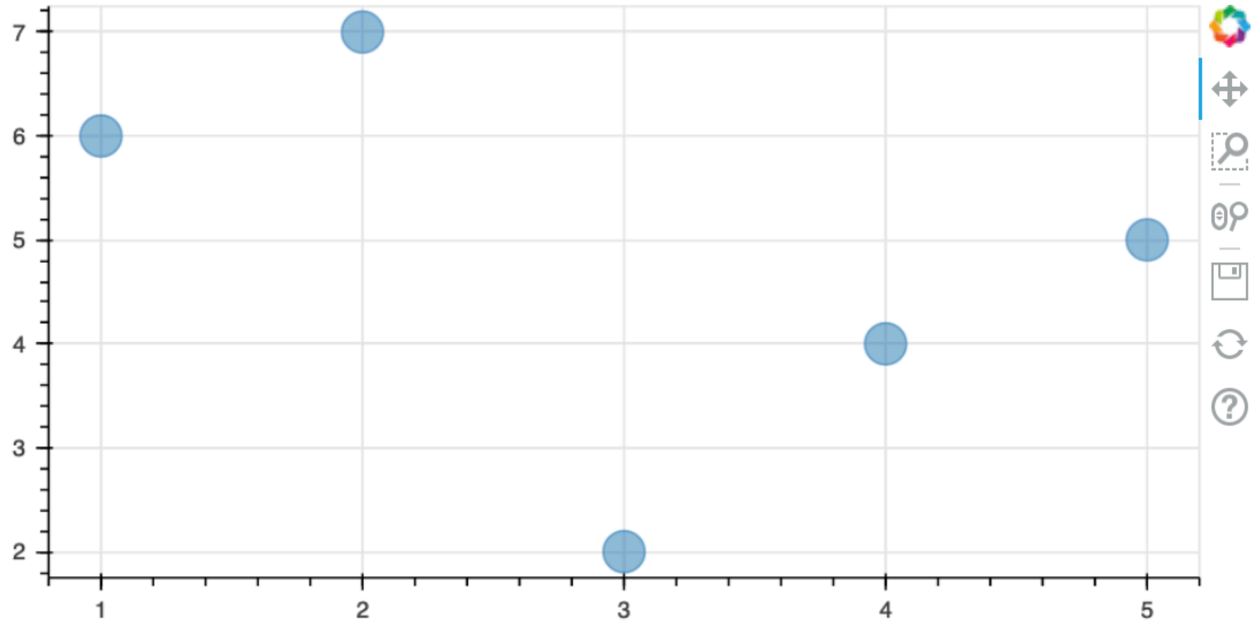
STORYTELLING

Source: Storytelling with data, Cole Nussbaumer Knaflic

PART 0
BOKEH

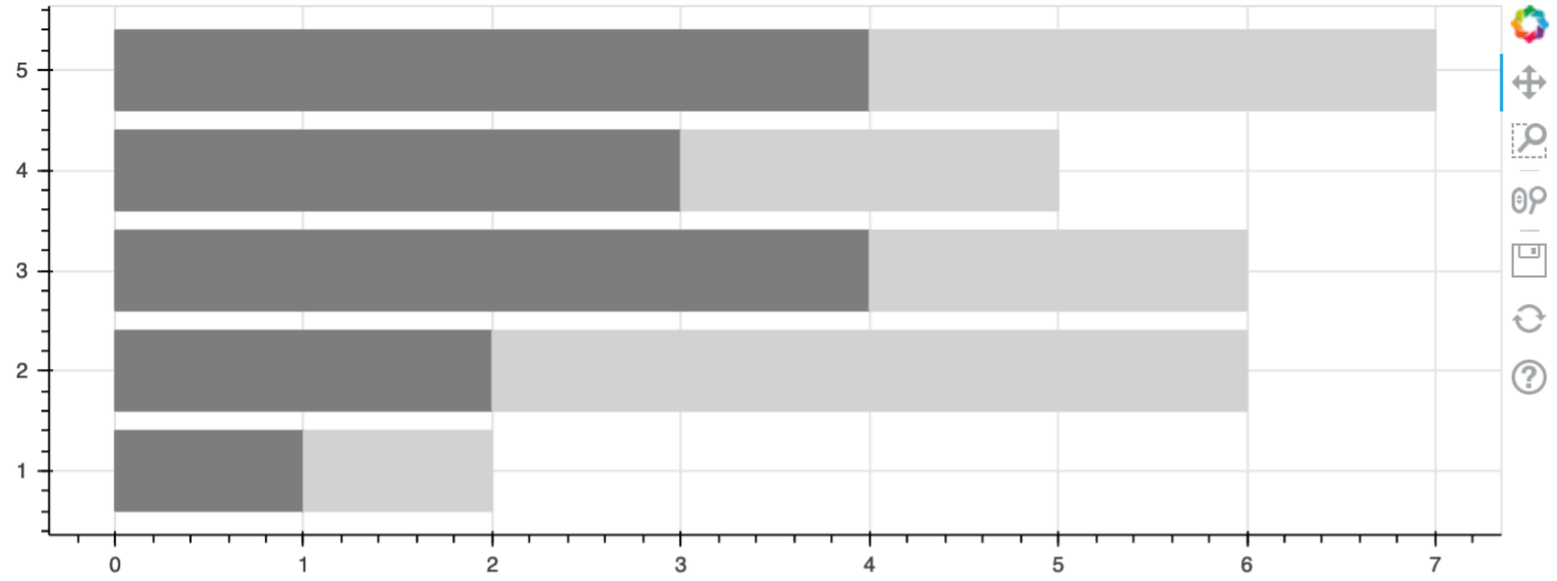
```
from bokeh.plotting import figure, output_file, show

output_notebook() #output_file("output.html")

p = figure(height=300)

# add a circle renderer with a size, color, and alpha
p.circle([1, 2, 3, 4, 5], [6, 7, 2, 4, 5], size=20, alpha=0.5)

# show the results
show(p)
```

BokehJS 2.4.3 successfully loaded.



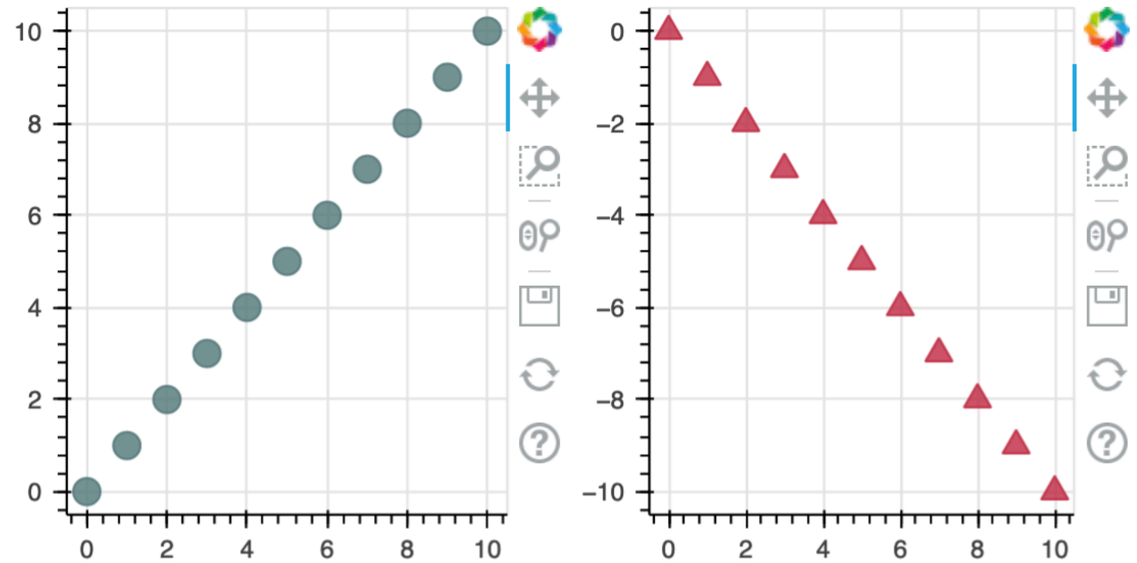
BASIC

```
source = pd.DataFrame(dict(
    y=[1, 2, 3, 4, 5],
    x1=[1, 2, 4, 3, 4],
    x2=[1, 4, 2, 2, 3],
))
p = figure(width=800, height=300)

p.hbar_stack(['x1', 'x2'], y='y', height=0.8, color=("grey", "lightgrey"), source=source)

show(p)
```



BASIC

```
from bokeh.layouts import row

x = np.arange(11)

# create three plots
s1 = figure(width=250, height=250)
s1.circle(x=x, y=x, size=12, color="#53777a", alpha=0.8)

s2 = figure(width=250, height=250)
s2.triangle(x=x, y=-x, size=12, color="#c02942", alpha=0.8)

# put the results in a row and show
show(row(s1, s2))
```

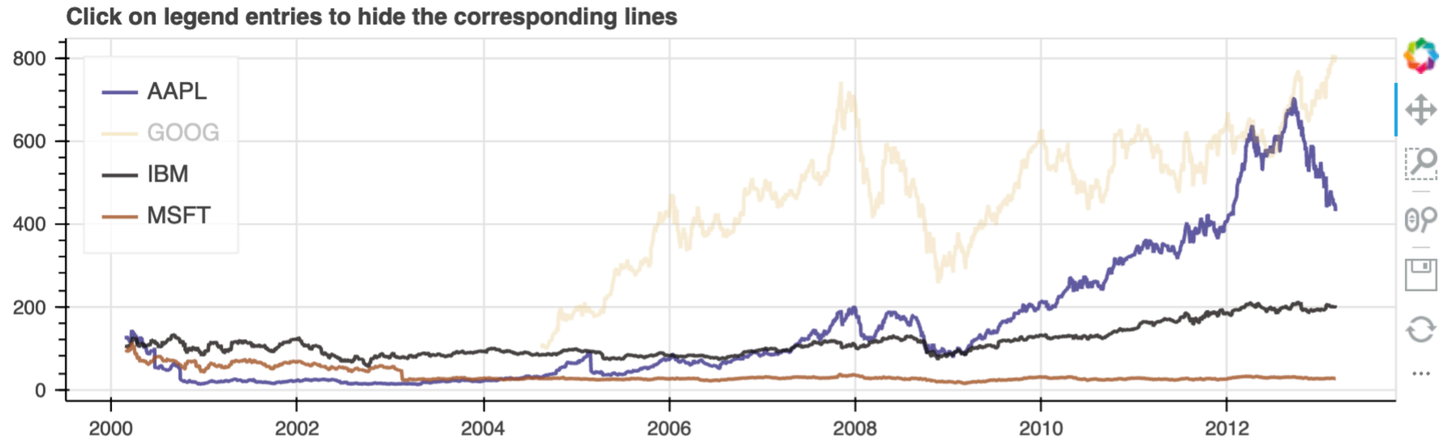


LAYOUT

```
p = figure(width=800, height=250, x_axis_type="datetime")
p.title.text = 'Click on legend entries to hide the corresponding lines'

i = 0
for name, data in df.groupby("stock"):
    p.line(data['date'], data['close'], line_width=2, color=palette[i], alpha=0.8, legend_label=name)
    i += 1

p.legend.location = "top_left"
p.legend.click_policy = "mute"

show(p)
```

date	open	high	low	close	volume	adj_close	stock
2000-03-01	118.56	132.06	118.50	130.31	38478000	31.68	AAPL
2000-03-02	127.00	127.94	120.69	122.00	11136800	29.66	AAPL
2000-03-03	124.87	128.23	120.00	128.00	11565200	31.12	AAPL
2000-03-06	126.00	129.13	125.00	125.69	7520000	30.56	AAPL
2000-03-07	126.44	127.44	121.12	122.87	9767600	29.87	AAPL

INTERACTIVE LEGEND

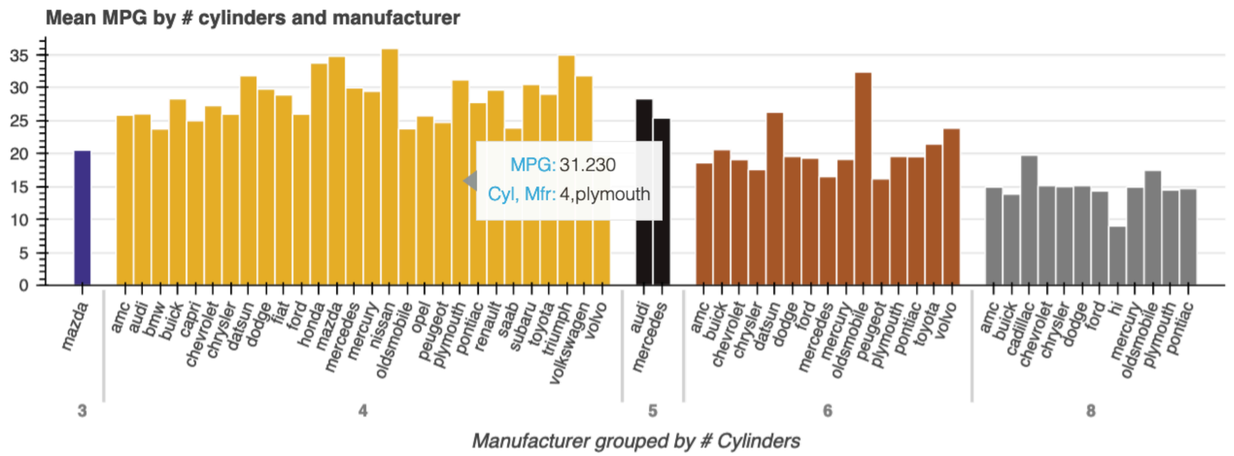


PART 0: BOKEH

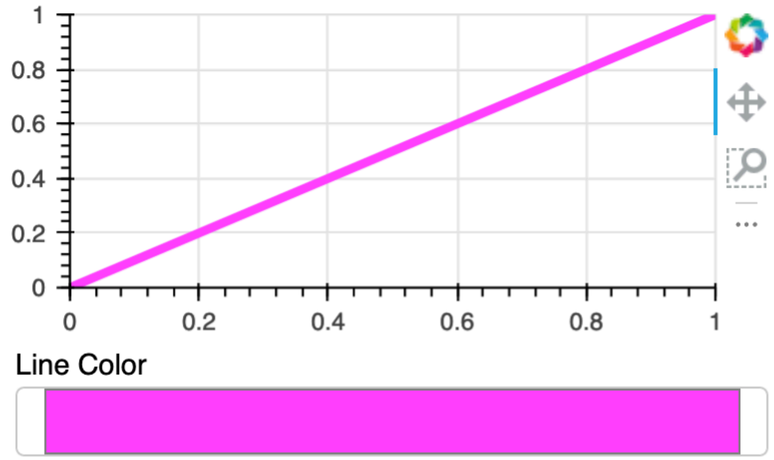
```
from bokeh.transform import factor_cmap
from bokeh.palettes import Viridis5

group = df.groupby(by=['cyl', 'mfr'])

p = figure(width=800,
           height=300,
           title="Mean MPG by # cylinders and manufacturer",
           x_range=group,
           toolbar_location=None,
           tooltips=[("MPG", "@mpg_mean"), ("Cyl, Mfr", "@cyl_mfr")])

p.vbar(x='cyl_mfr',
       top='mpg_mean',
       width=1,
       source=group,
       line_color="white",
       fill_color=factor_cmap('cyl_mfr', palette=palette, factors=sorted(df.cyl.unique()), end=1))

p.y_range.start = 0
p.x_range.range_padding = 0.05
p.xgrid.grid_line_color = None
p.xaxis.axis_label = "Manufacturer grouped by # Cylinders"
p.xaxis.major_label_orientation = 1.2
p.outline_line_color = None

show(p)
```

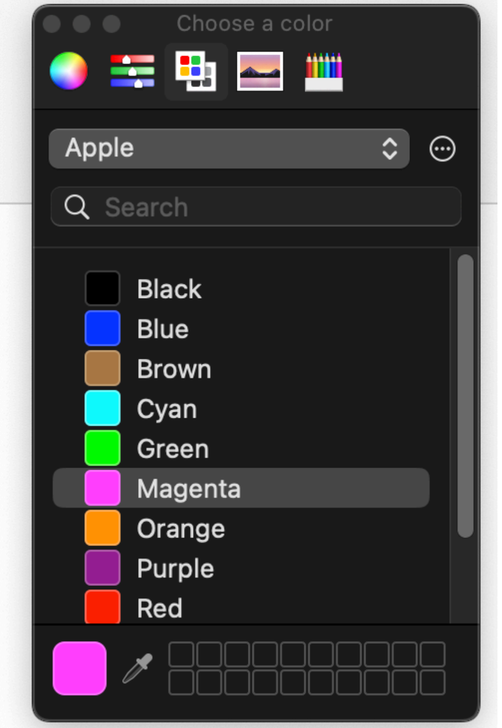
mpg	cyl	displ	hp	weight	accel	yr	origin	name	mfr
18.0	8	307.0	130	3504	12.0	70	North America	chevrolet chevelle malibu	chevrolet
15.0	8	350.0	165	3693	11.5	70	North America	buick skylark 320	buick
18.0	8	318.0	150	3436	11.0	70	North America	plymouth satellite	plymouth
16.0	8	304.0	150	3433	12.0	70	North America	amc rebel sst	amc
17.0	8	302.0	140	3449	10.5	70	North America	ford torino	ford

INTERACTIVE TOOLTIPS

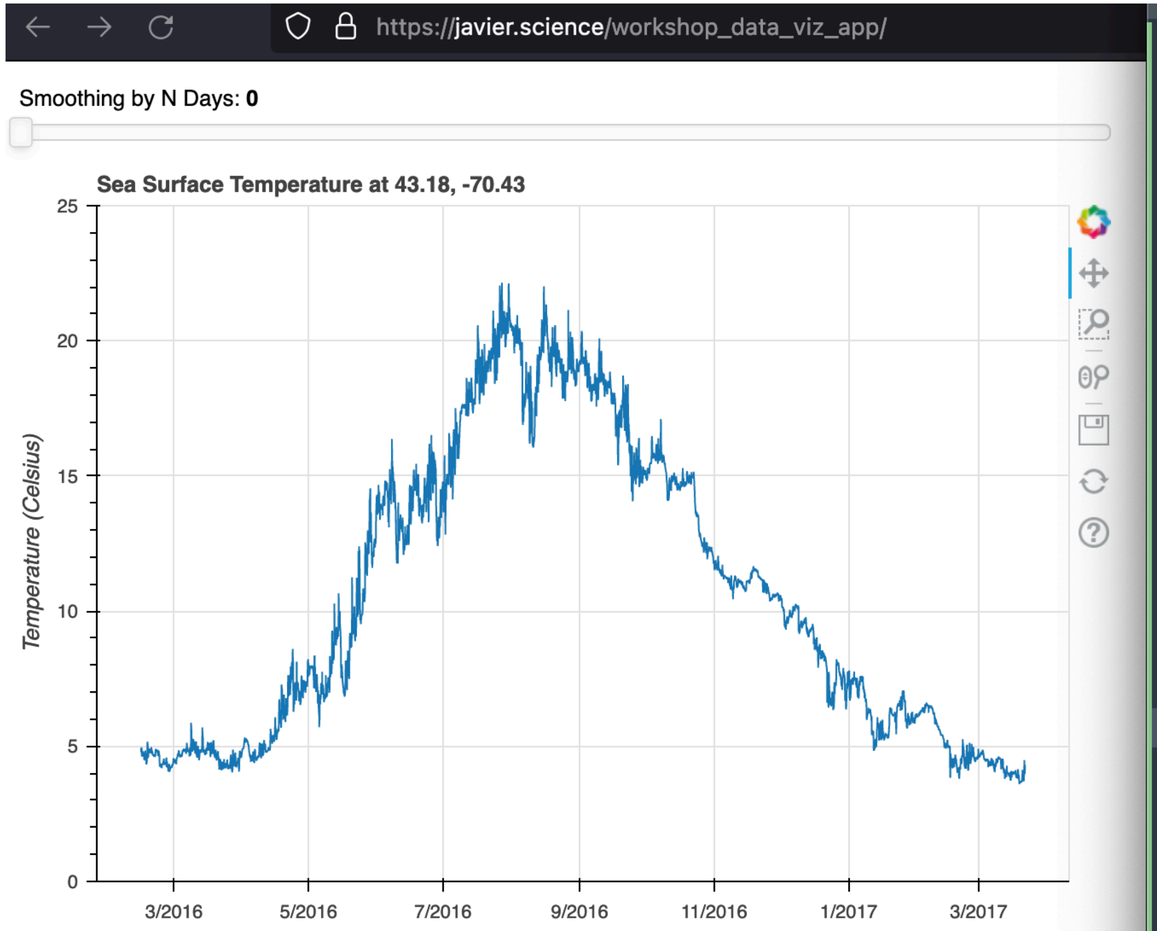


```
from bokeh.models import ColorPicker

plot = figure(x_range=(0, 1), y_range=(0, 1), width=350, height=150)
line = plot.line(x=(0,1), y=(0,1), color="black", line_width=4)

picker = ColorPicker(title="Line Color")
picker.js_link('color', line.glyph, 'line_color')

show(column(plot, picker))
```





ADDING JAVASCRIPT




```python
def bkapp():
    # read data and convert it to bokeh's data structure
    df = sea_surface_temperature.copy()
    source = ColumnDataSource(data=df)

    # make plot
    plot = figure(plot_width=650,
                  plot_height=450,
                  x_axis_type='datetime', y_range=(0, 25),
                  y_axis_label='Temperature (Celsius)',
                  title="Sea Surface Temperature at 43.18, -
    # add line
    plot.line('time', 'temperature', source=source)

    # callback for reactivity
    def callback(attr, old, new):
        if new == 0:
            data = df
        else:
            data = df.rolling('{0}D'.format(new)).mean()
        source.data = ColumnDataSource.from_df(data)

    # somethign to interact with
    slider = Slider(start=0, end=30, value=0, step=1, title=
    slider.on_change('value', callback) # when to use callba

    return column(slider, plot)

# update plot
bokeh_app = pn.pane.Bokeh(bkapp()).servable()
```

CREATING APPS

HOW TO REDUCE COGNITIVE LOAD

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

understand the **context**

UNDERSTANDING

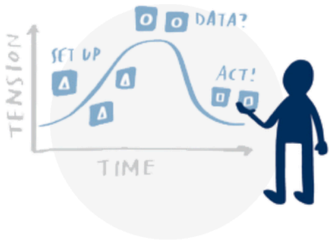
choose an **effective visual**

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?

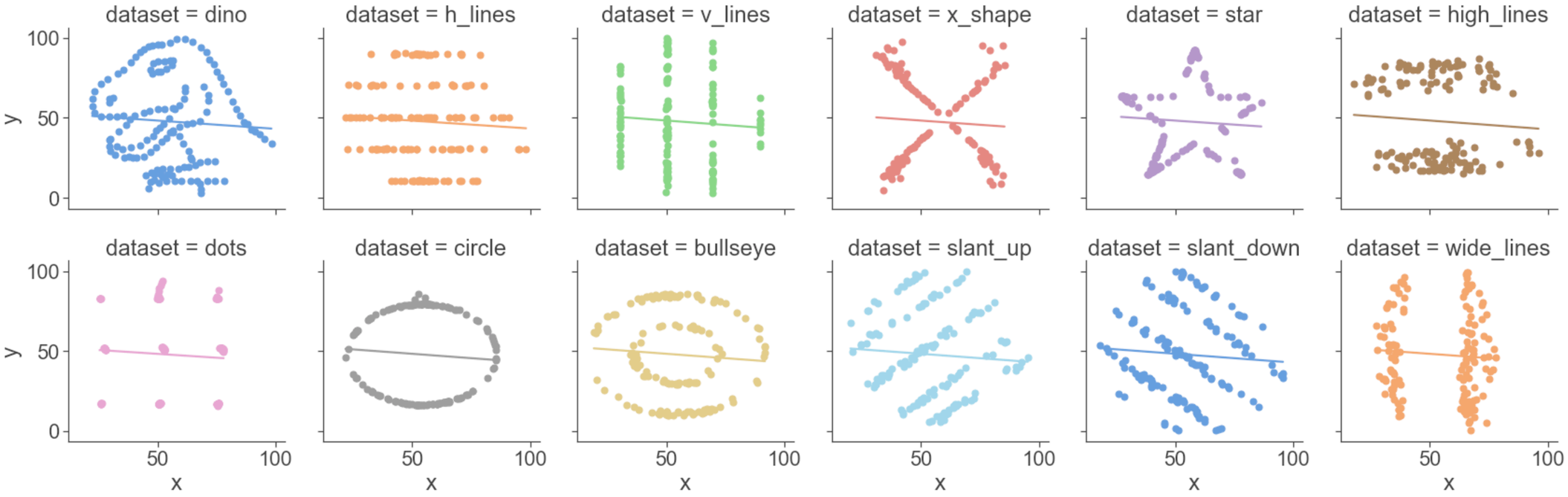
(2) WHAT IS THE MAIN MESSAGE?

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

- Yourself (exploratory) vs **others (explanatory)**

- Academics vs journalists vs lay audience

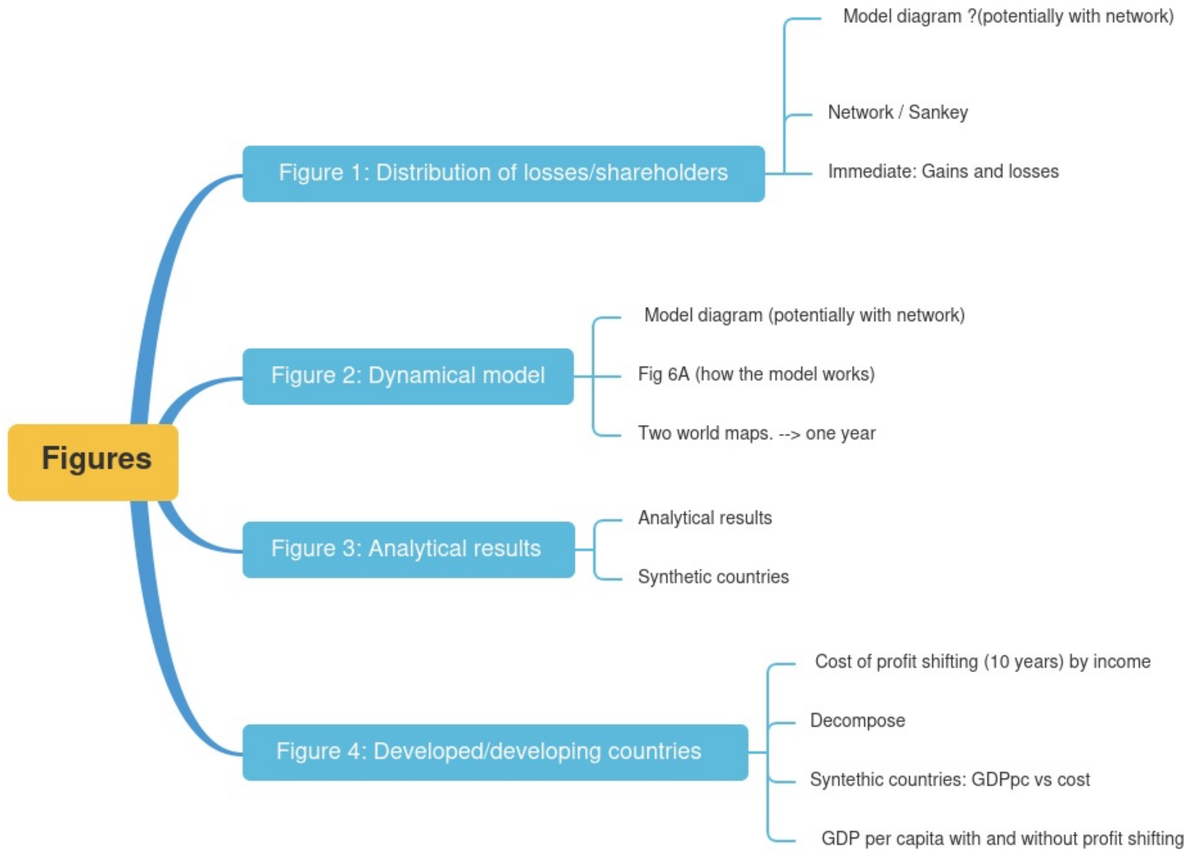
- Online audience (color/bw) vs presentation



(2) WHAT IS THE MAIN MESSAGE?

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

Write it down in one sentence, be specific.



PRACTICE: EXPLORE DATASET

- Derek's:
  - Collection of (microbial DNA) samples of the inside of the dental unit water system
  - Columns: dosing of antimicrobial, amount of legionella, amount of DNA
  - Message idea: X and Y cleaning factors give the safest microbiome in the water system
- School Shooting 1: (ss.tsv)
  - School shooting per day
  - Columns: date, number of victims, state of attack
  - Message idea: the probability of a new attack increases after each attack
- School Shooting 2: (tweets_and_ss.tsv)
  - School shooting per day and number of tweets with the terms "school" and "shooting"
  - Columns: date, number of victims, number of tweets, state of attack
  - Message idea: the probability of a new attack increases with the buzz in social media

- Steps:
  - 15 minutes to load data, explore it and come up with a message/audience

PART 2
EFFECTIVE VISUALS

WHICH FIGURE SHALL I USE?

GOALS:

Two conditions:

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

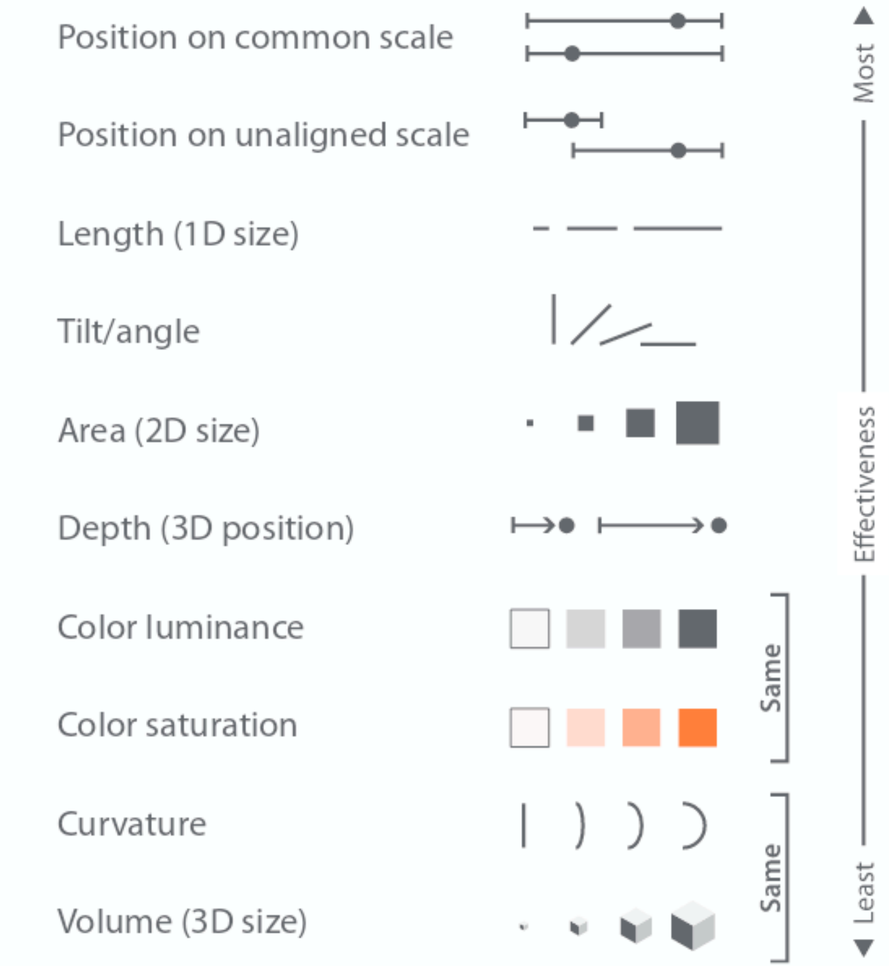
ELEMENTS OF A PLOT:

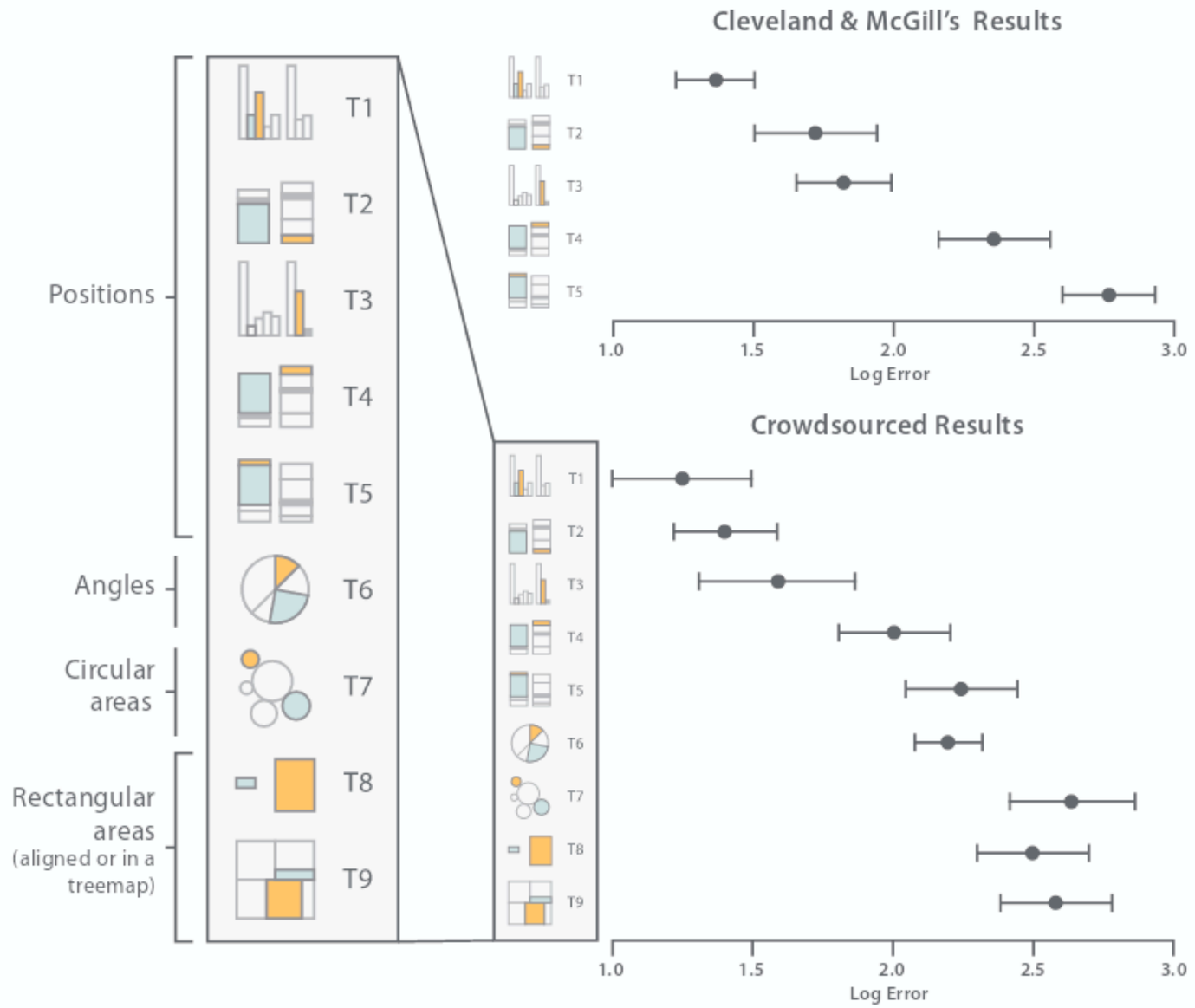
CHANNELS

PART 2: CHOOSE AN EFFECTIVE VISUAL

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes



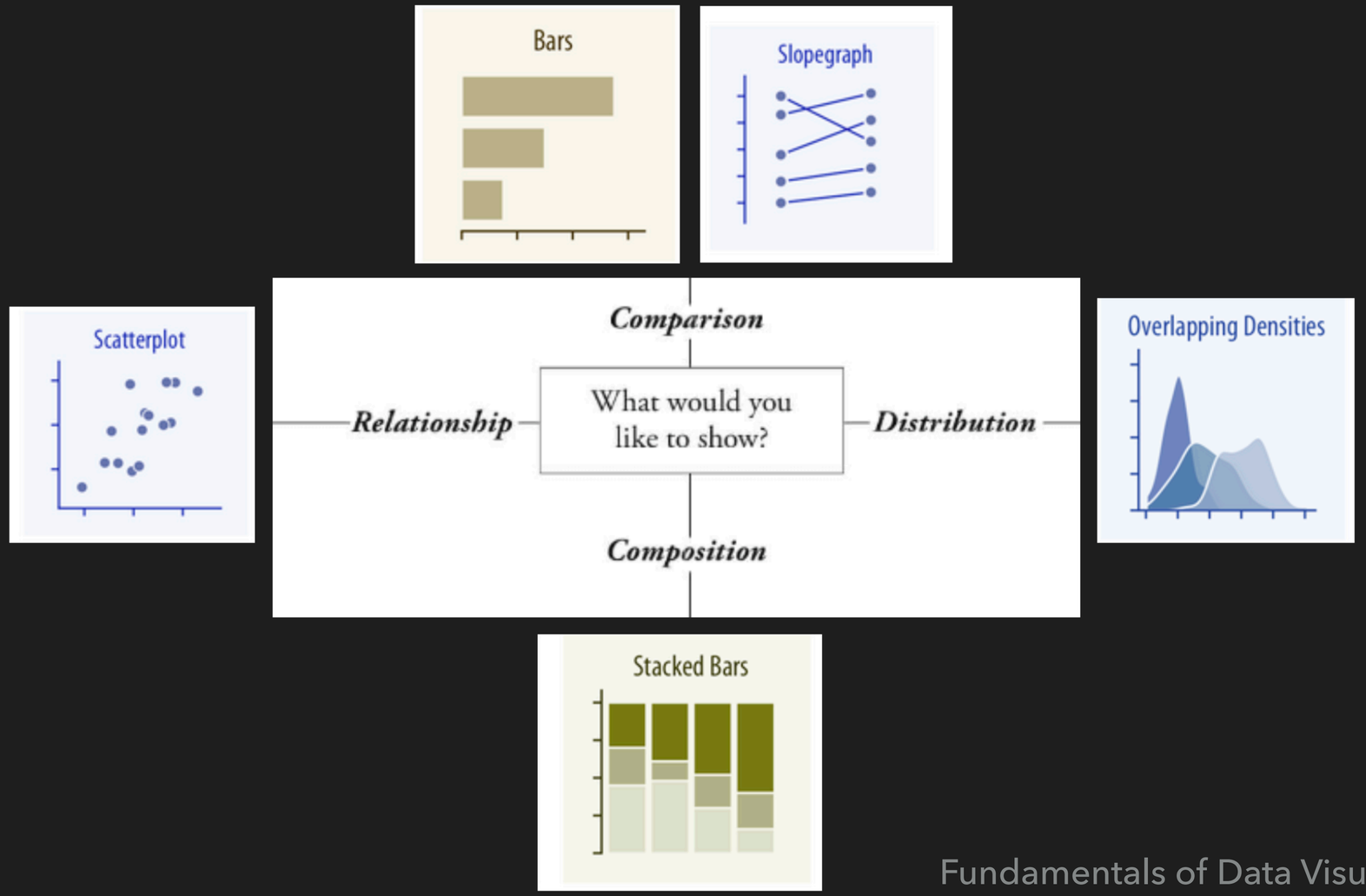


NOT ALL CHANNELS ARE EQUAL

ELEMENTS OF A PLOT:

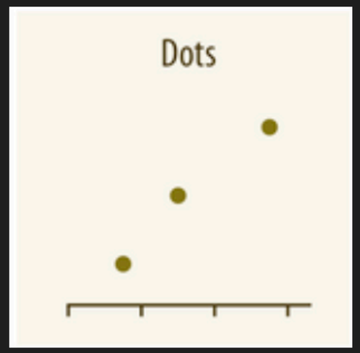
GRAPHICAL OBJECTS

THE TYPE OF GRAPH DEPENDS ON THE GOAL



Fundamentals of Data Visualization by Claus O. Wilke

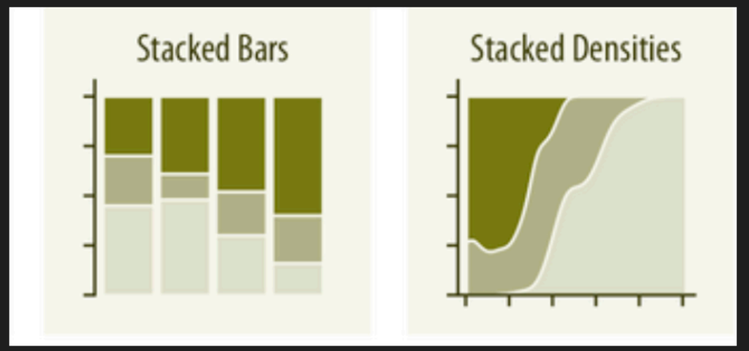
AMOUNTS AND PROPORTIONS



Required with log-scales

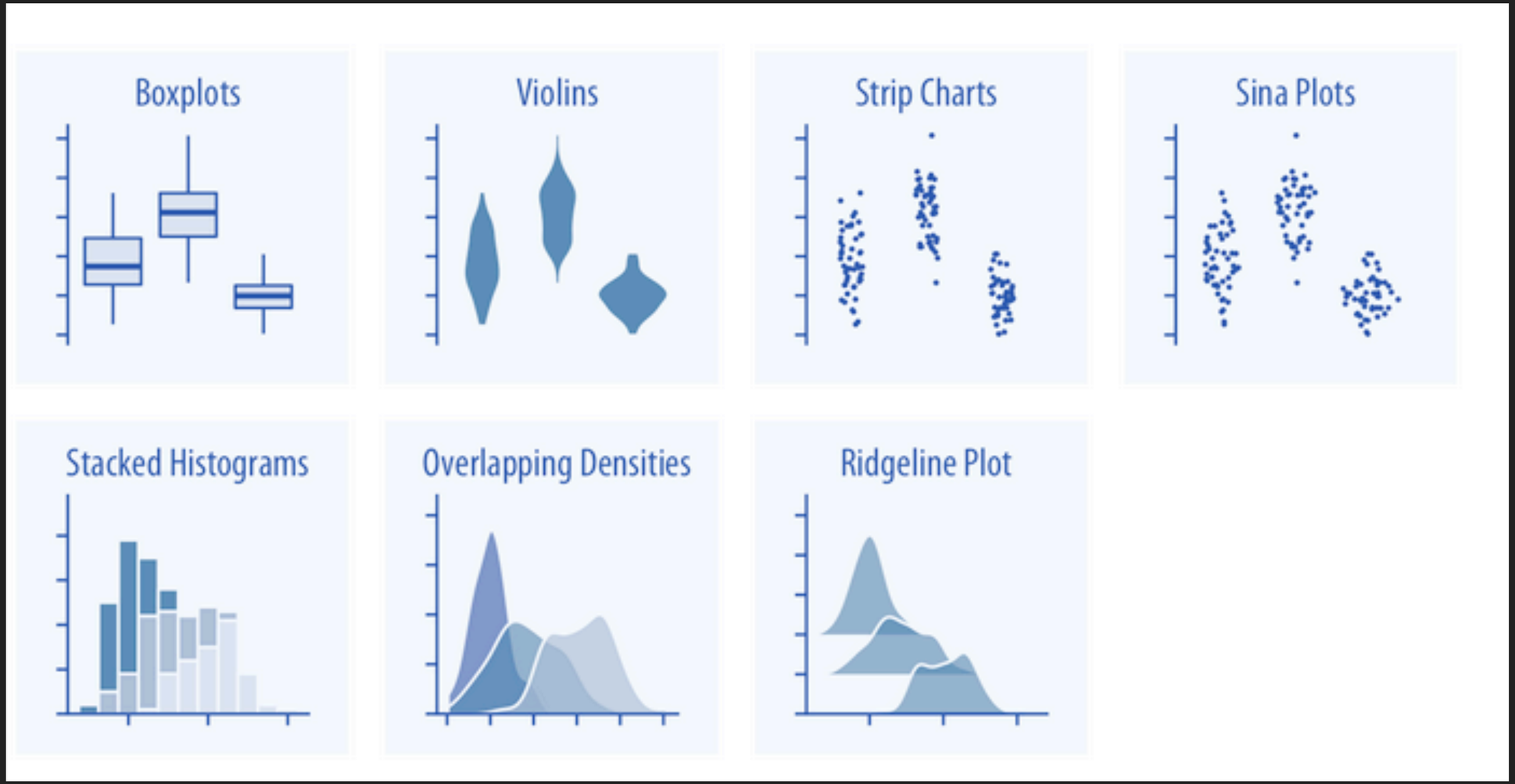
Trends

Differences within row



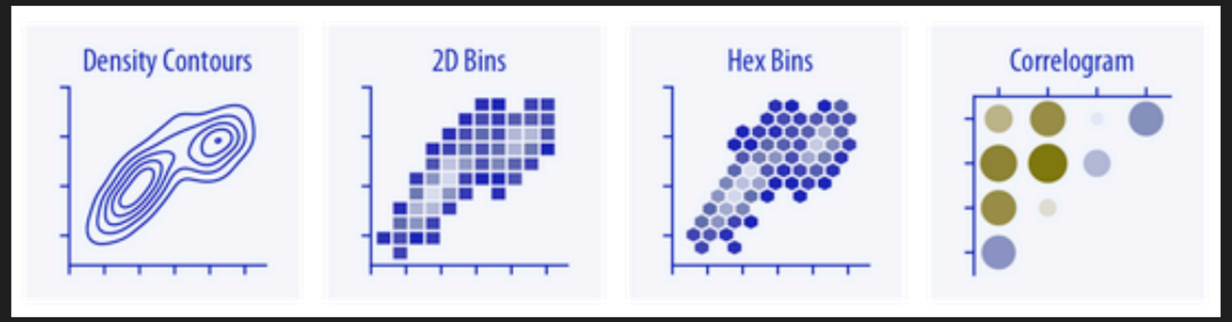
Proportions over x

DISTRIBUTIONS



RELATIONSHIPS

Too many points?

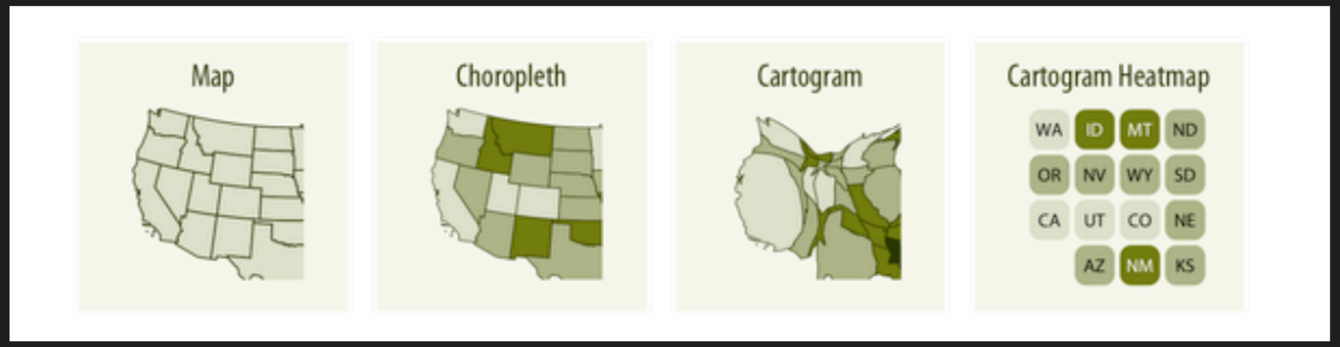


Time series?

GEOGRAPHICAL DATA

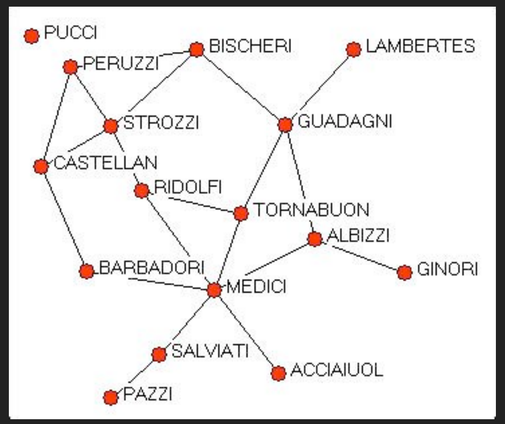
- Color is key (more on this later)
- Combine with a barplot or bubbles if the values are important

Do you *actually* need a map?

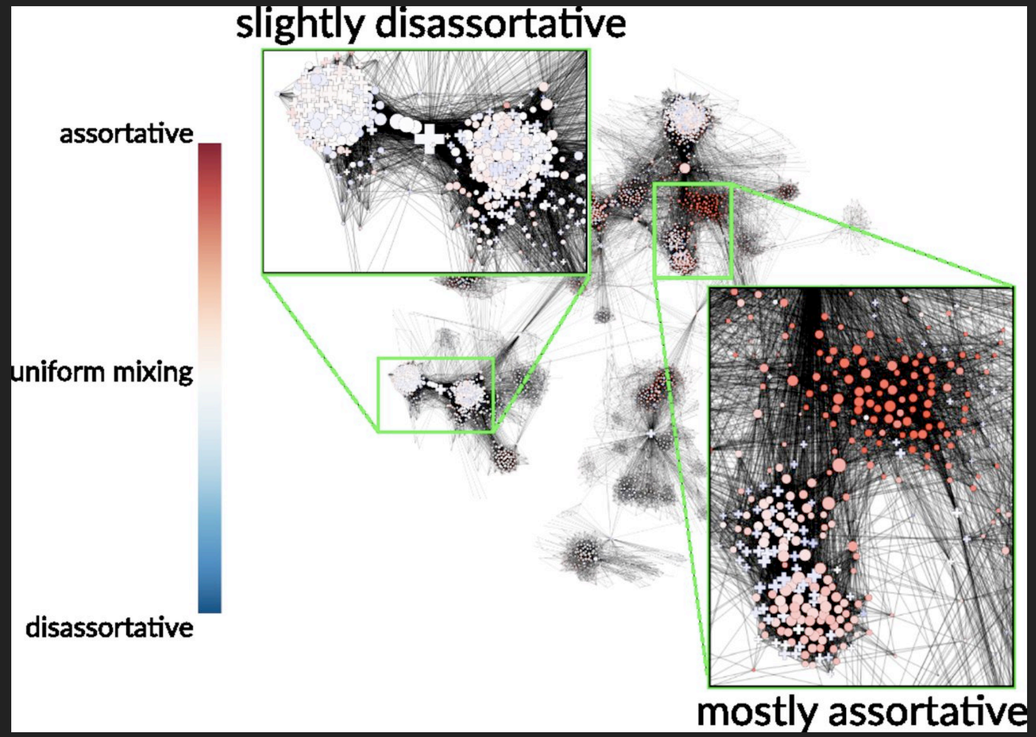


NETWORKS: DO YOU ACTUALLY NEED A NETWORK?

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



Florencian families



Peel et al

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 numbers next to each other is better
- **Transformations** –> Jitter, etc

PRACTICE: MAP TO CHANNELS AND SKETCH VISUALIZATION

- Link your message to the type of figure
  - Type of graph(s)
  - How are you mapping variables to channels?
  - Which transformations (if any)

- Steps:
  - 10 minutes to sketch your idea on paper
  - Mix groups and get feedback from other groups – 10 min

PRACTICE: MAP TO CHANNELS AND SKETCH VISUALIZATION

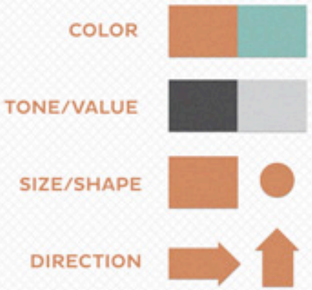
- Link your message to the type of figure
  - Type of graph(s)
  - How are you mapping variables to channels?
  - Which transformations (if any)

- Steps:
  - 20 minutes to code your plot

PART 3
DESIGN

PRINCIPLES OF DESIGN: CRAP

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.

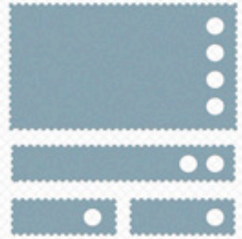
To the left, you can see 4 ways to create contrast in your design.

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.

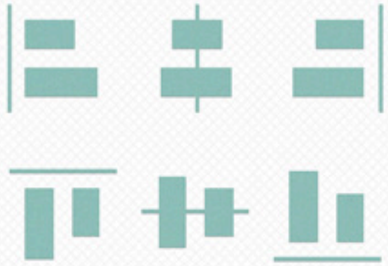


PRINCIPLES of DESIGN

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

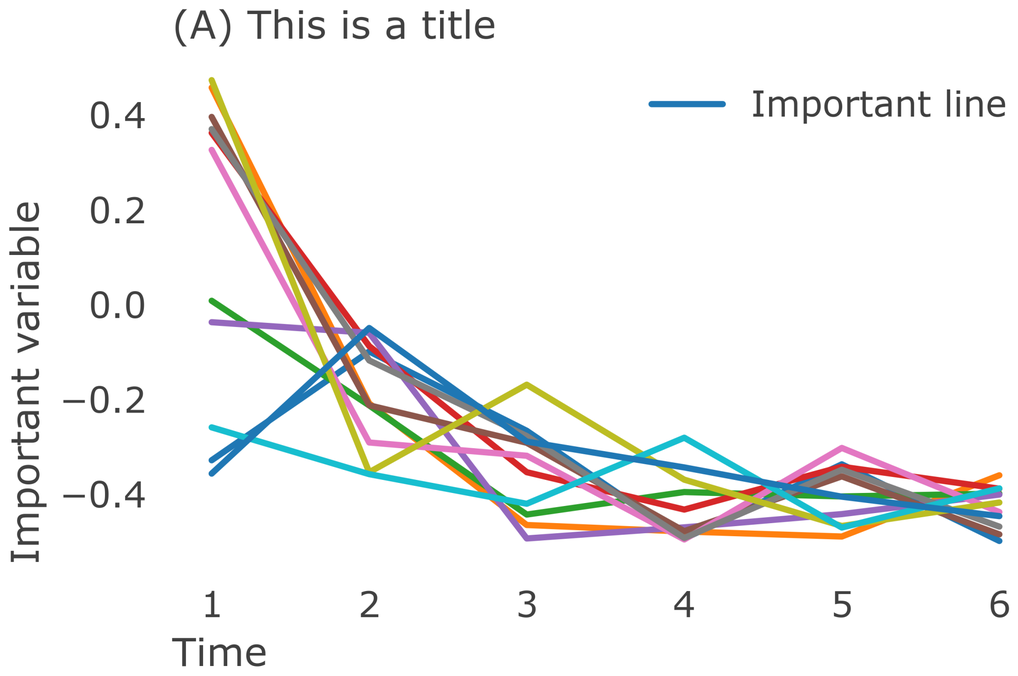
Proximity allows for visual unity in a design. If two elements are related to each other, they should be placed in close proximity to one another. Doing so minimizes visual clutter, emphasizes organization, and increases viewer comprehension.

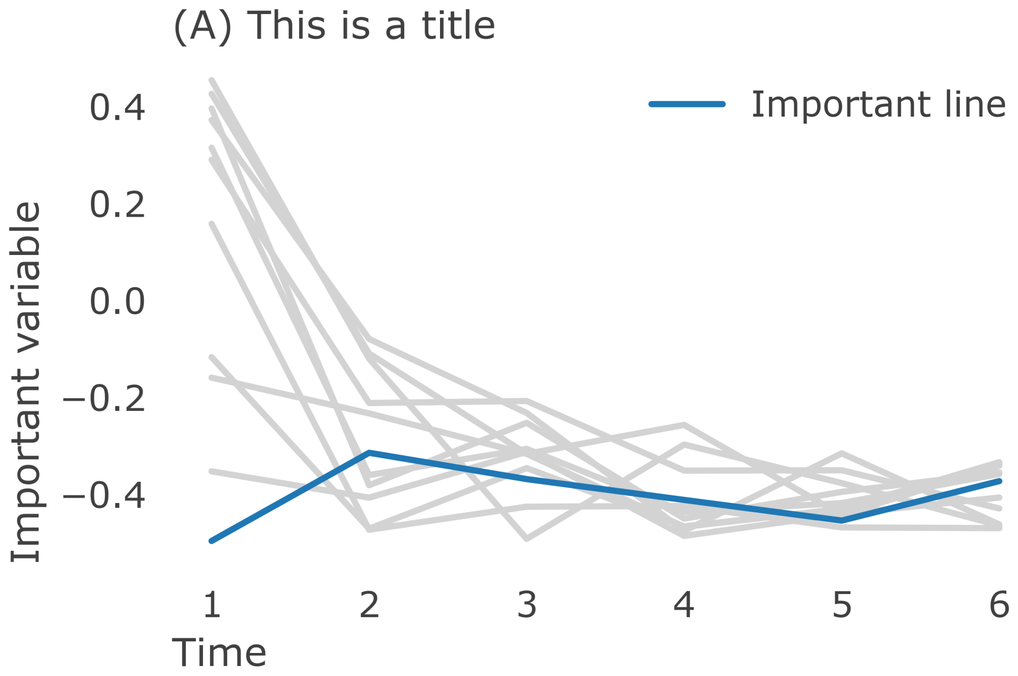
Imagine how ridiculous it would be if the proximity icons on this graphic were located on the other side of this document.

a handy paper leaf resource

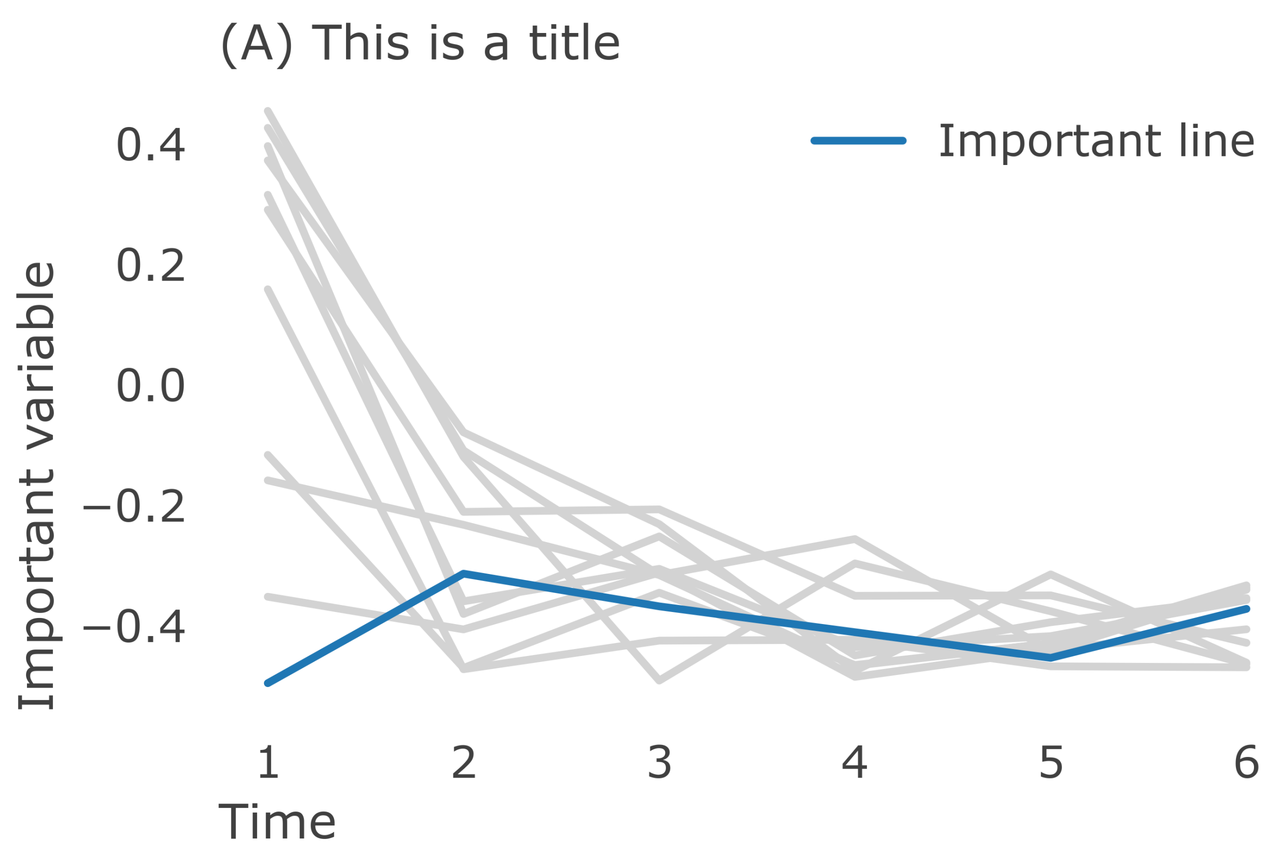
Source: Paper leaf

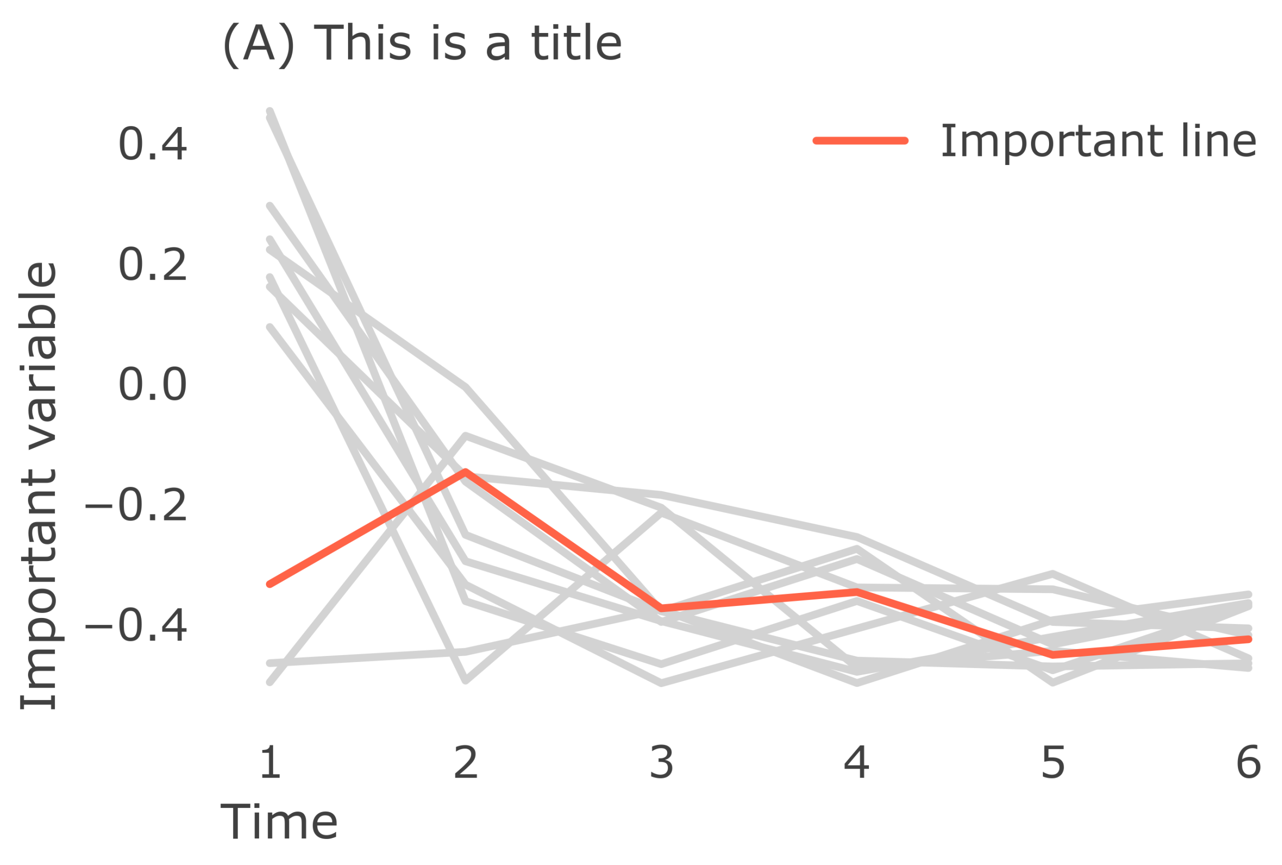
CONTRAST: ELIMINATE CLUTTER

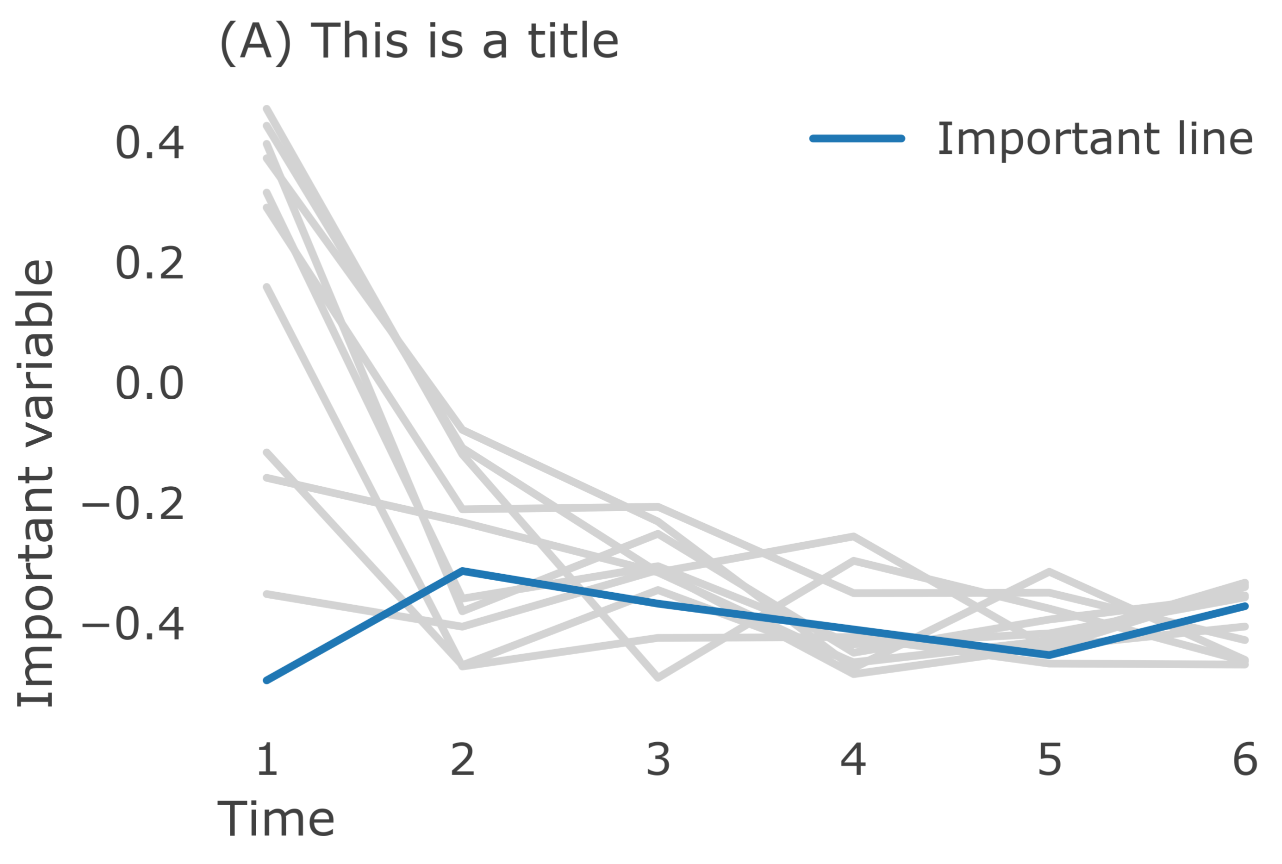


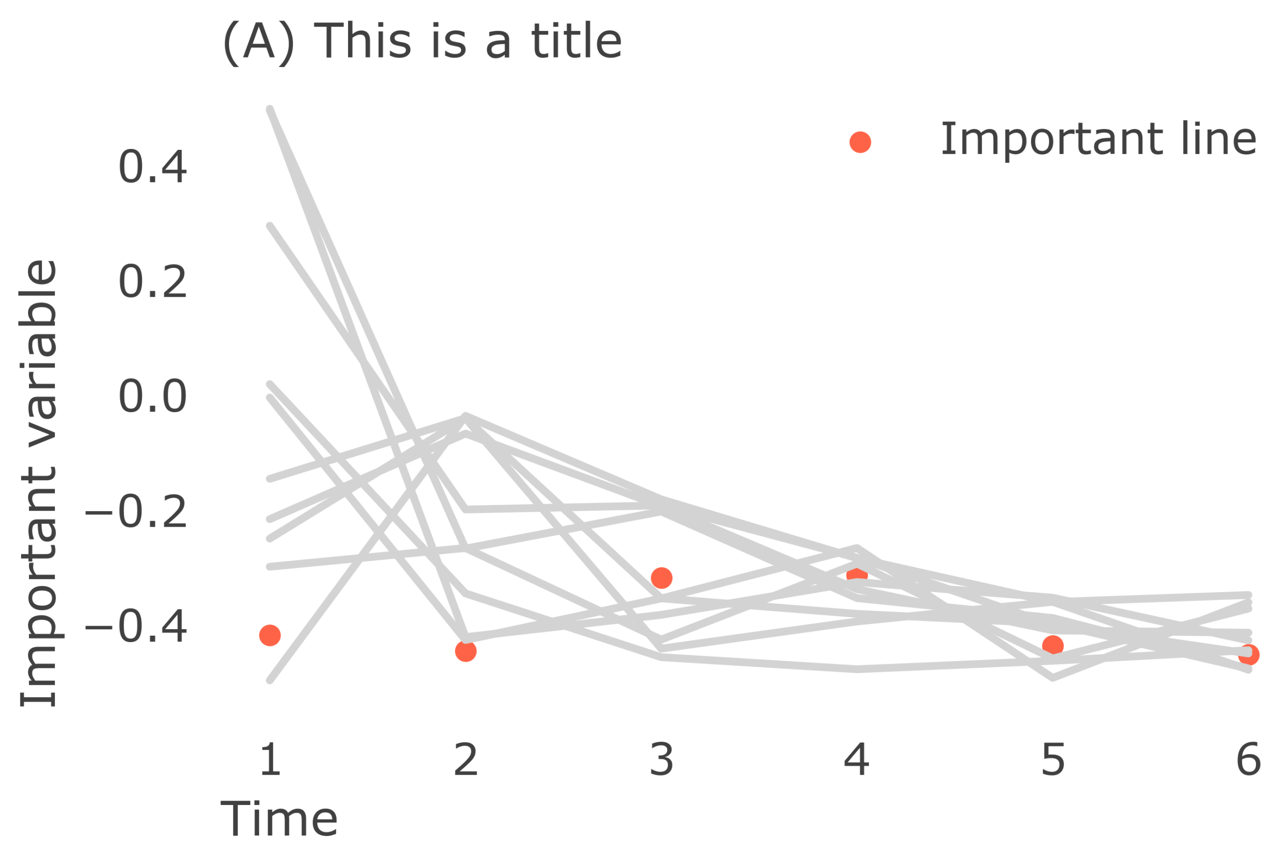


REPETITION


(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

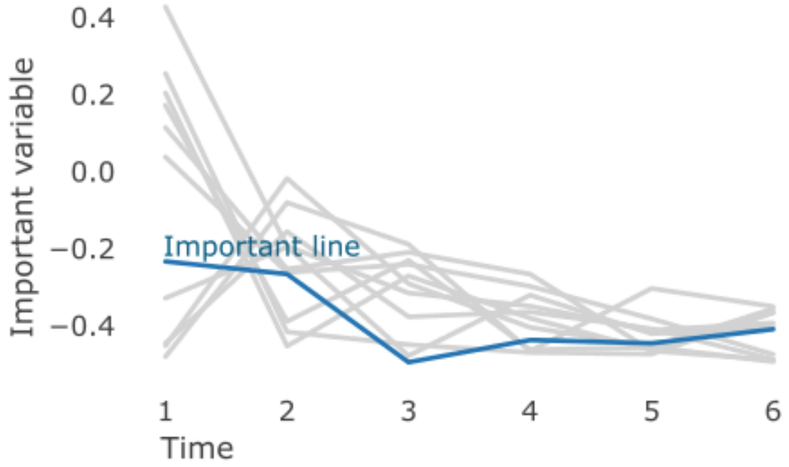

(A) This is a title
Important line
0.4
0.2
0.0
−0.2
−0.4
Important variable
1
2
3
4
5
6
Time

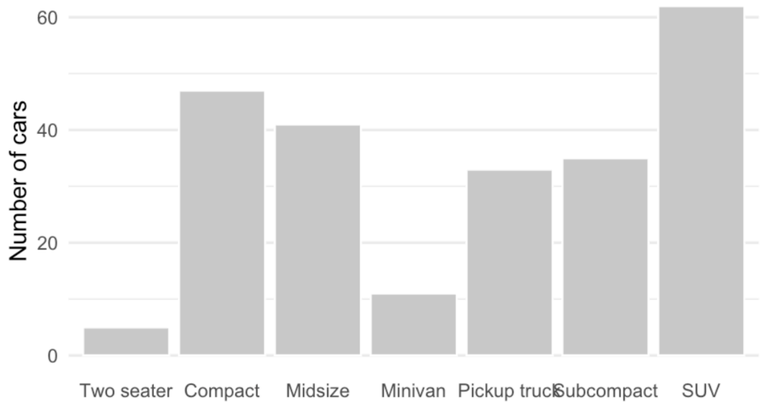

(A) This is a title
Important line
0.4
0.2
0.0
−0.2
−0.4
Important variable
1
2
3
4
5
6
Time

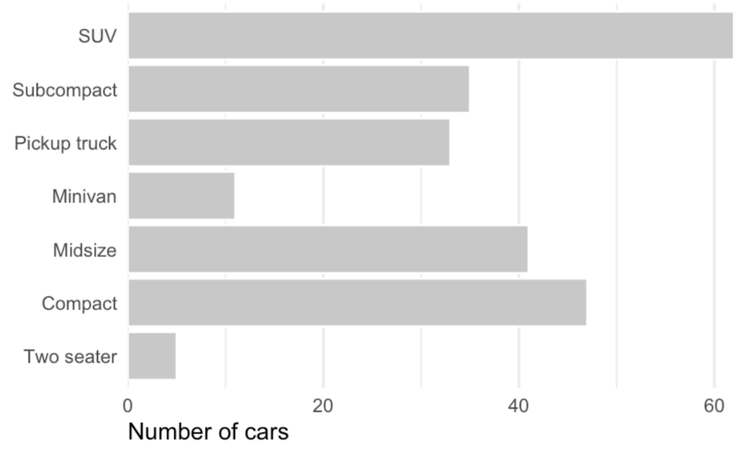
ALIGNMENT

LOOKS GOOD

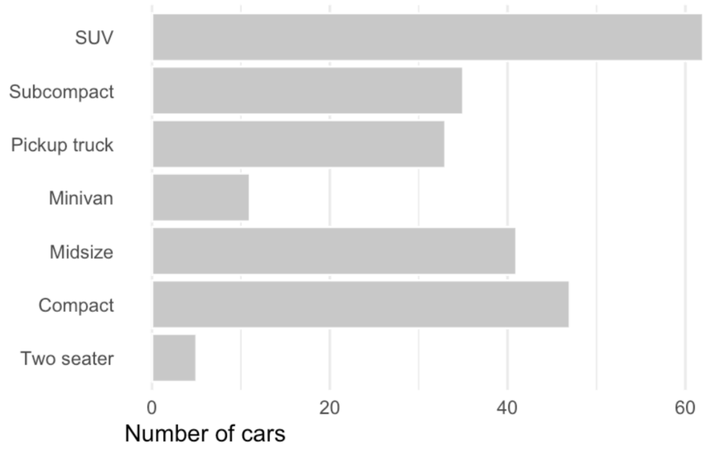
DOES NOT LOOK SO GOOD

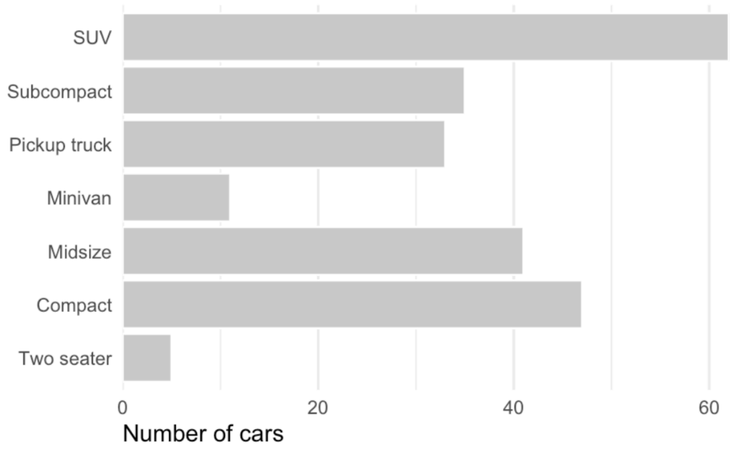


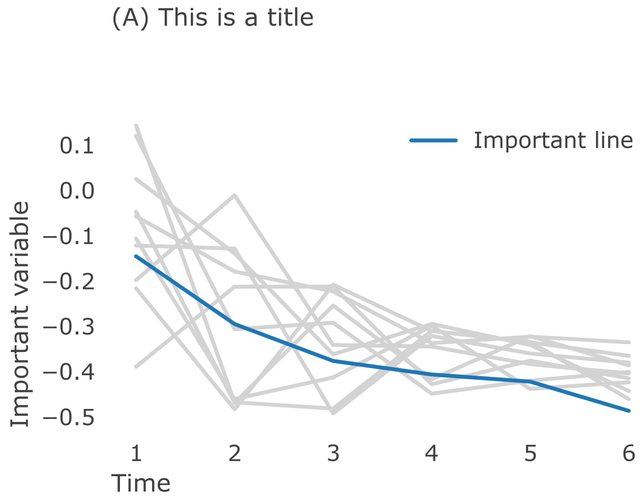


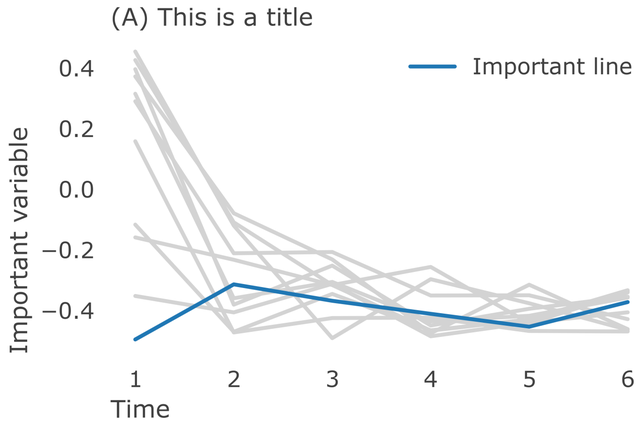


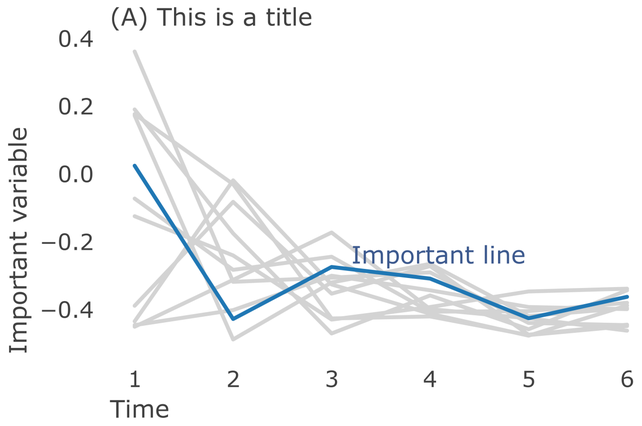
PROXIMITY


SUV
Subcompact
Pickup truck
Minivan
Midsize
Compact
Two seater
0
20
40
60
Number of cars


SUV
Subcompact
Pickup truck
Minivan
Midsize
Compact
Two seater
0
20
40
60
Number of cars


(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

PRINCIPLE 1

Contrast blah

PRINCIPLE 2

Repetition blah

Principle 3

Alignment blah

PRINCIPLE 4

Proximity blah

PRINCIPLE 1

Contrast blah

PRINCIPLE 2

Repetition blah

Principle 3

Alignment blah

PRINCIPLE 4

Proximity blah

PRINCIPLE 1

Contrast blah

PRINCIPLE 2

Repetition blah

PRINCIPLE 3

Alignment blah

PRINCIPLE 4

Proximity blah

PRINCIPLE 1

Contrast blah

PRINCIPLE 2

Repetition blah

PRINCIPLE 3

Alignment blah

PRINCIPLE 4

Proximity blah

PRACTICAL GUIDE

- Reduce cognitive load:
  - Removing unnecessary clutter
  - More professional/aesthetically pleasant
- 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 –> Edit them in AI or Inkscape
- Repetition: Be consistent in different figures
- Alignment: Make sure you align subplots/labels
- Proximity: When possible, label data directly (instead of using legends)

PRINCIPLES OF VISUAL PERCEPTION

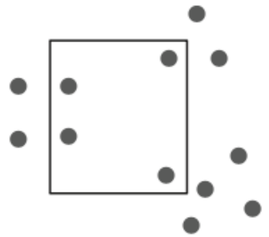
GESTALT PRINCIPLES OF VISUAL PERCEPTION

Principles/laws of human perception that describe how humans group similar elements, recognize 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



ENCLOSURE

Enclosed areas contain related objects (highlight areas)

CLOSURE

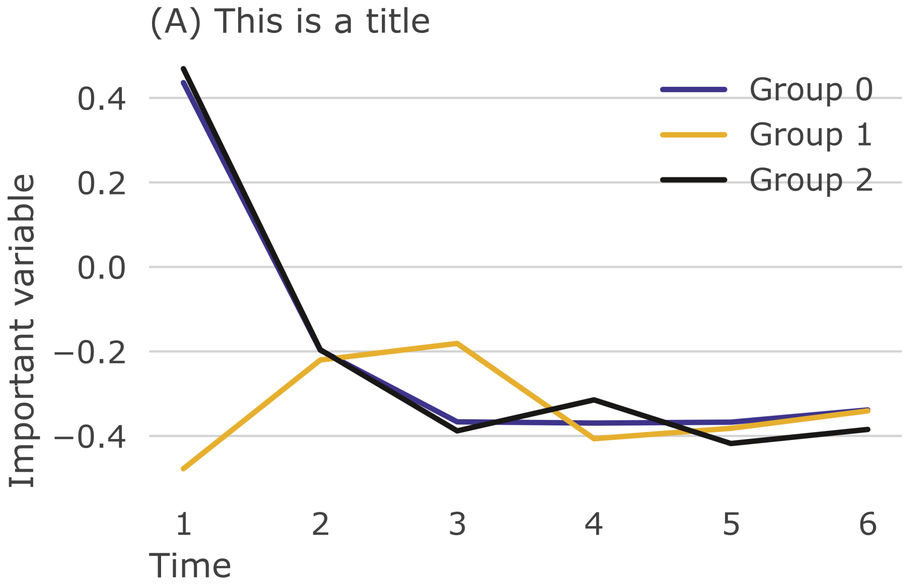
No need to close frames

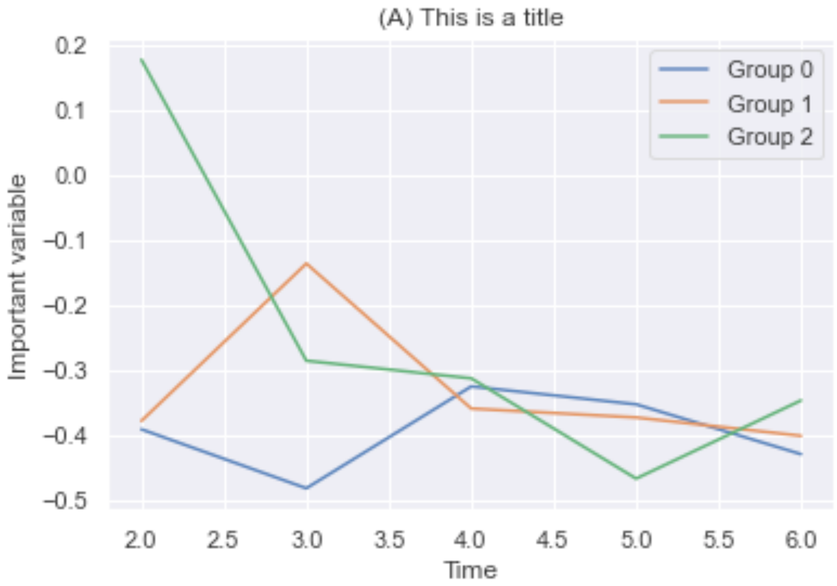
GESTALT PRINCIPLES OF VISUAL PERCEPTION

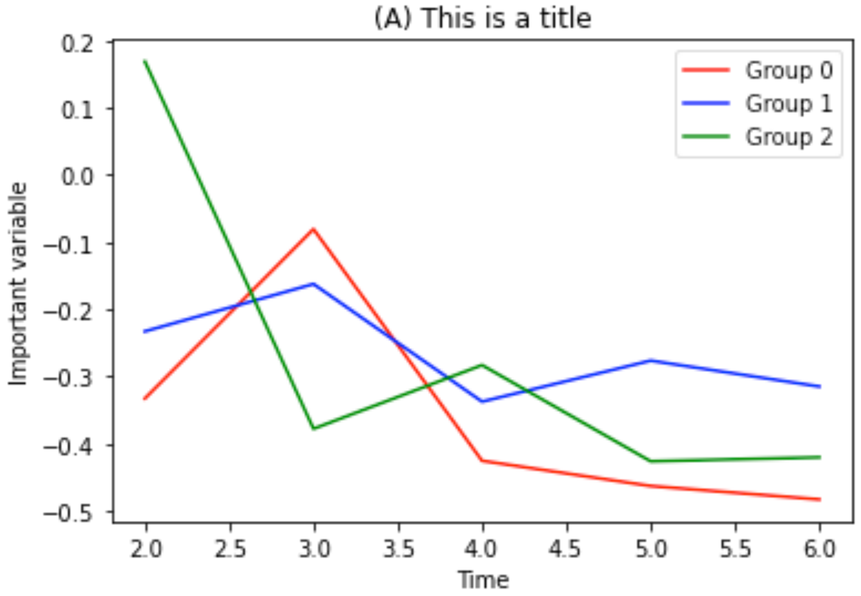
CONTINUITY e.g., No need to have a left line

CONNECTION e.g., Networks

PART 3: DESIGN





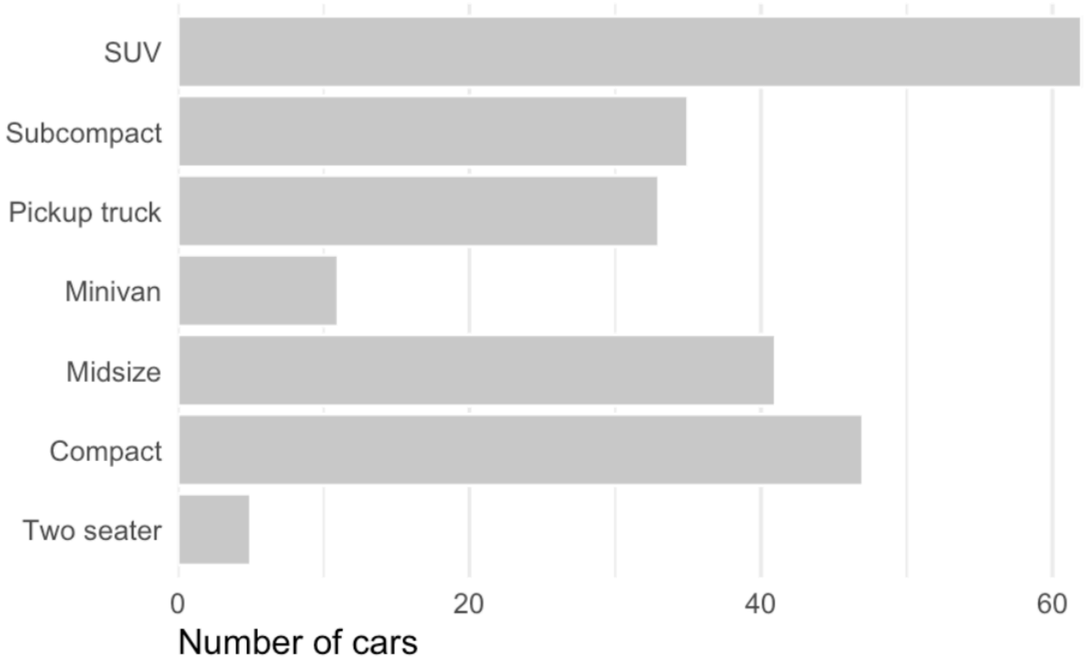


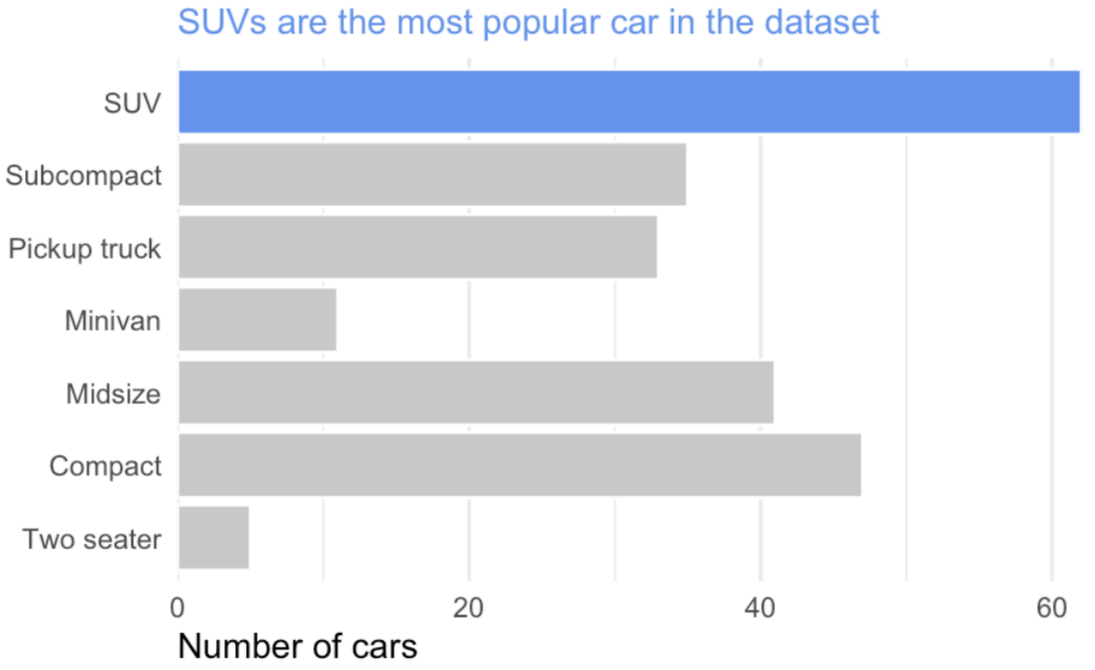
Closure: No need for a frame, we understand that it is one plot
Proximity: We know that the labels belong to the axes because they are nearby
Continuity: 2, 3, 4, 5, 6 are perceived as connected
Connection: All dots within one line
Similarity: The three lines represent similar data

GUIDE THE READER:

PRACTICAL GUIDES

- Compare how you read the figure in the right and in the left





WHY DO WE NEED TO FOCUS ATTENTION?

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

HOW DO WE FOCUS ATTENTION?

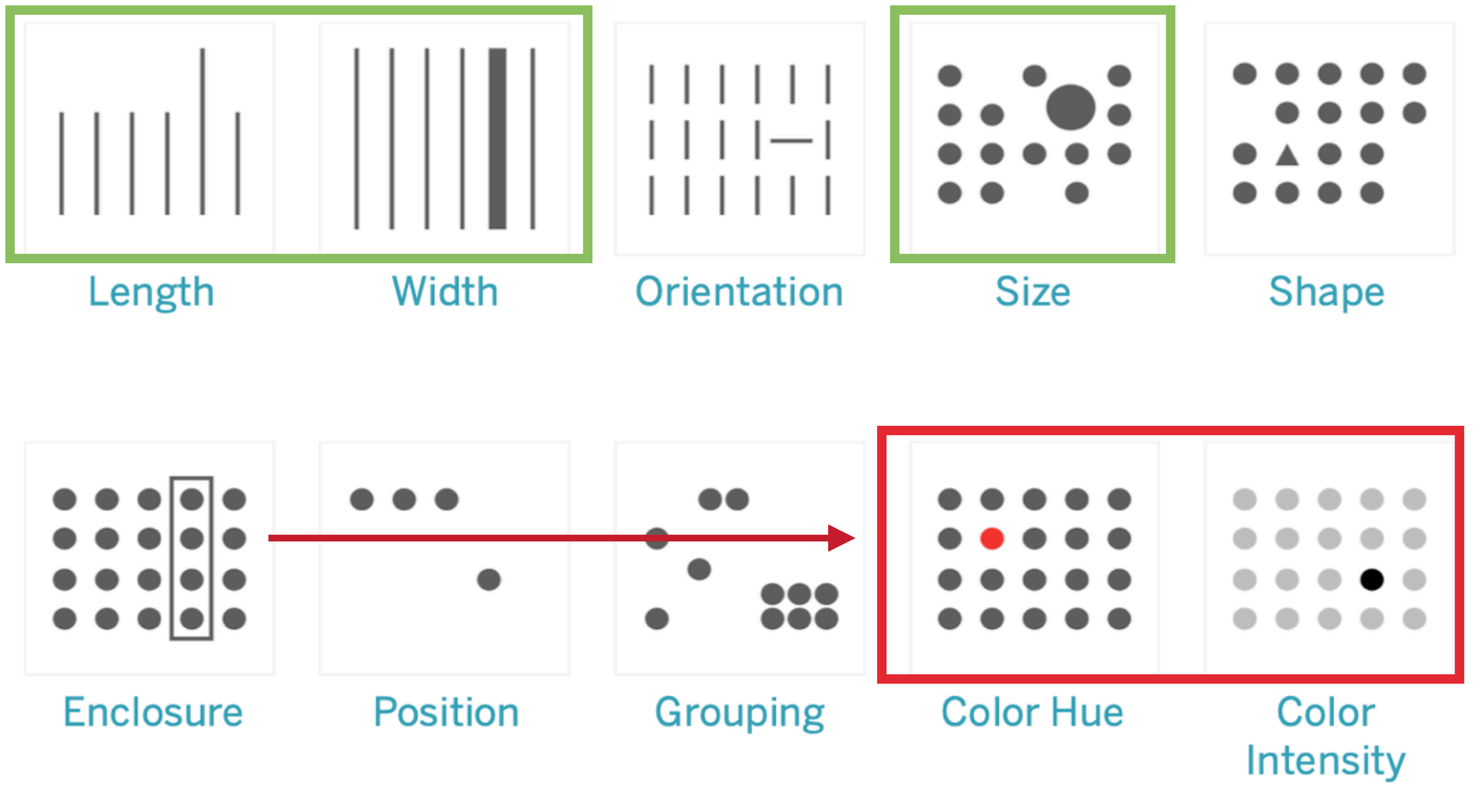
Using pre-attentive attributes

Story telling with data

IMPORTANT PRE ATTENTIVE ATTRIBUTES

What do we focus on:

- Large objects
- Bright objects
- Contrasting objects



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

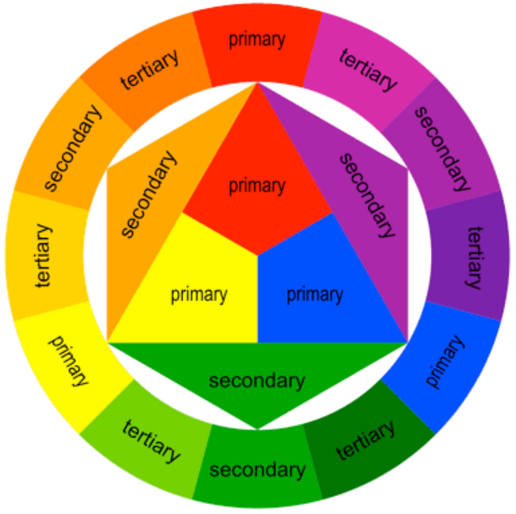
Color affect emotion, culture-dependent. But some responses are nearly universal

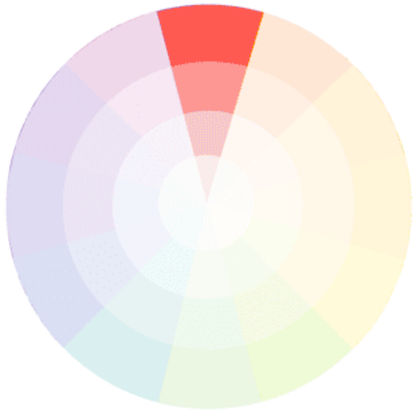
- Warm colors --> alive/alert
- Blue colors --> calming/focus

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

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

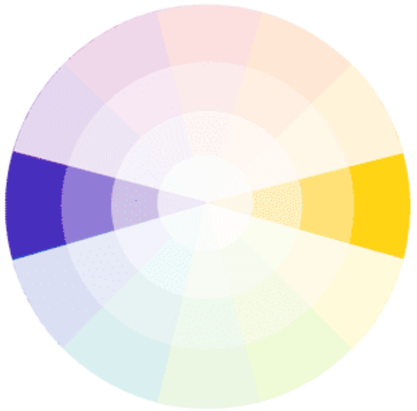
COLORS THAT LOOK GOOD TOGETHER



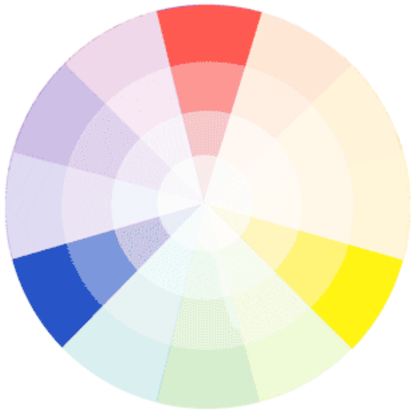


Monochromatic

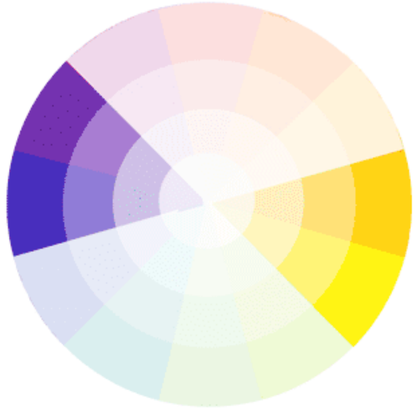
Analogous



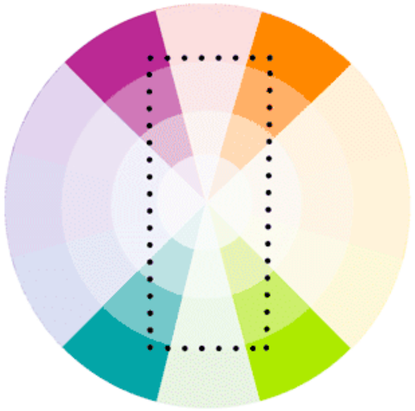
Complementary



Triad



Split-Complementary



Tetradic

COLOR PALETTES

SEQUENTIAL

Minimum is important

DIVERGING

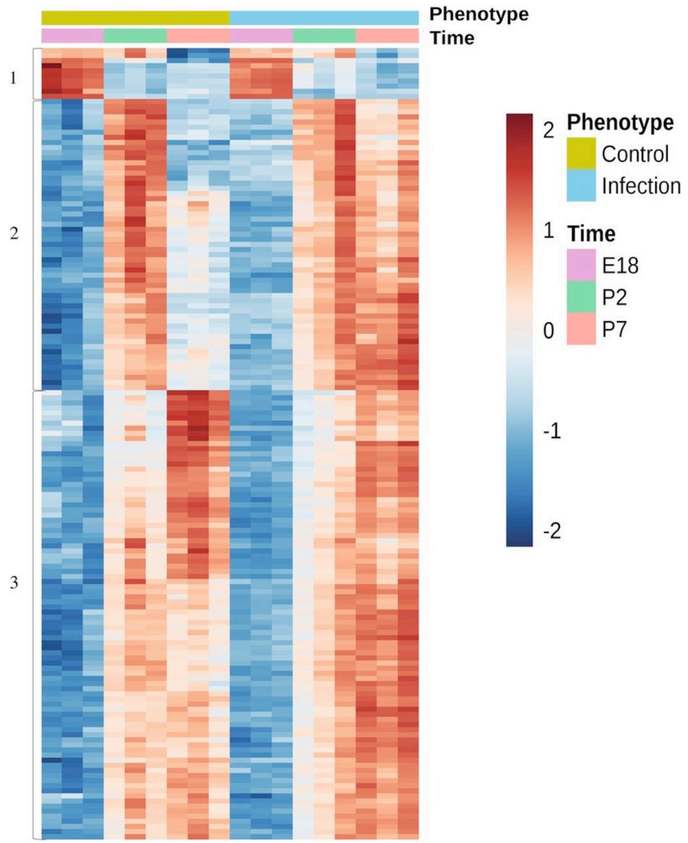
Mean is important

QUALITATIVE

Represent categories

COLOR TO REPRESENT VALUES

ONLY IF THE EXACT VALUES ARE NOT IMPORTANT



Haque, Koski and Scott, 2019

PRACTICE: MAKE YOUR VIZ BEAUTIFUL AND EFFECTIVE

- Think of the CRAP principles. Are you following them?
- Color:
  - Which reason? (guide the reader, represent categories, represent values)
  - Which palette?
- How are you using pre-attentive attributes?

- Steps:
  - 20 minutes to refine your plot

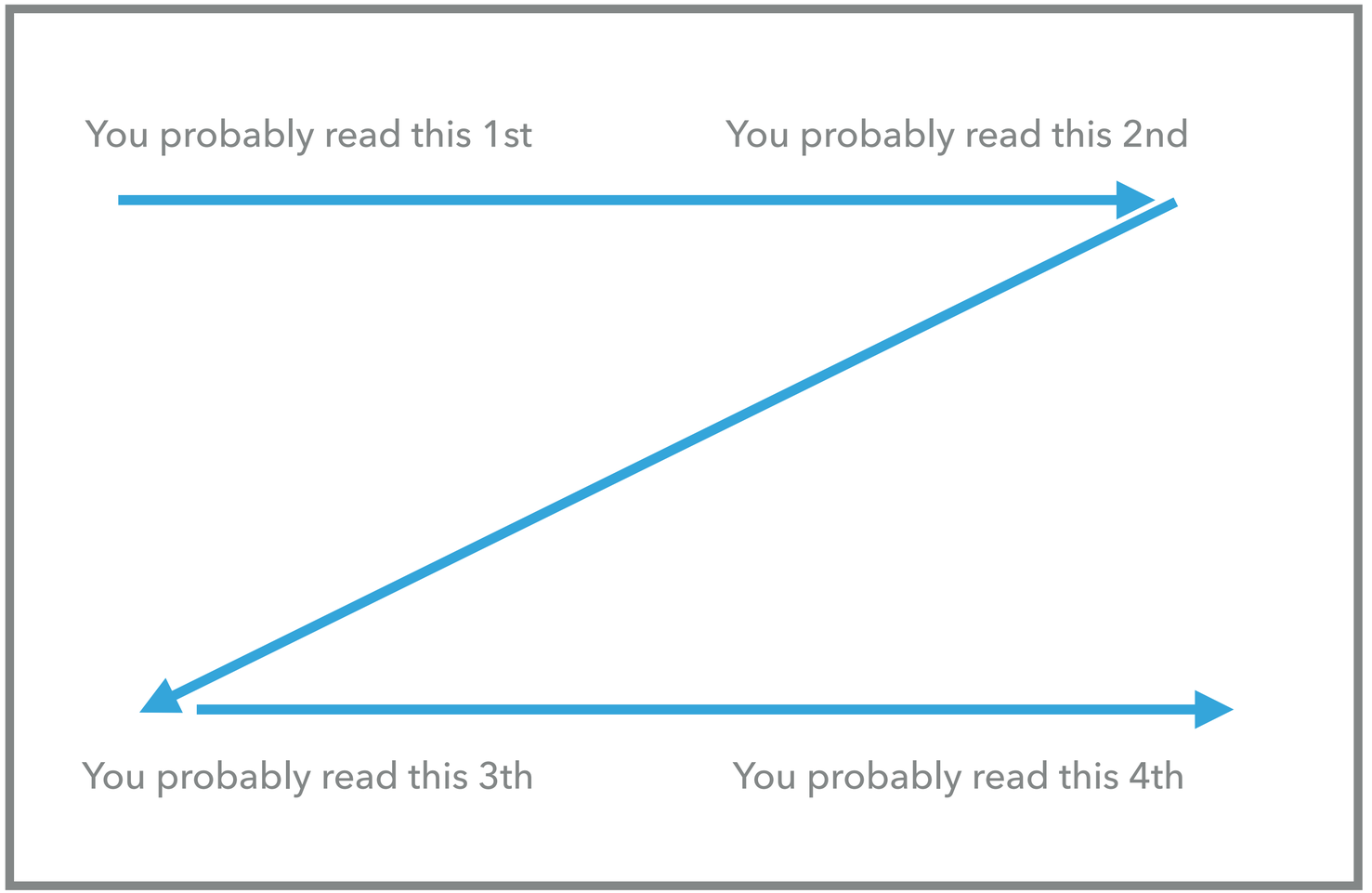
PART 4

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*


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

HOW TO TELL A STORY

Narrative arc

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

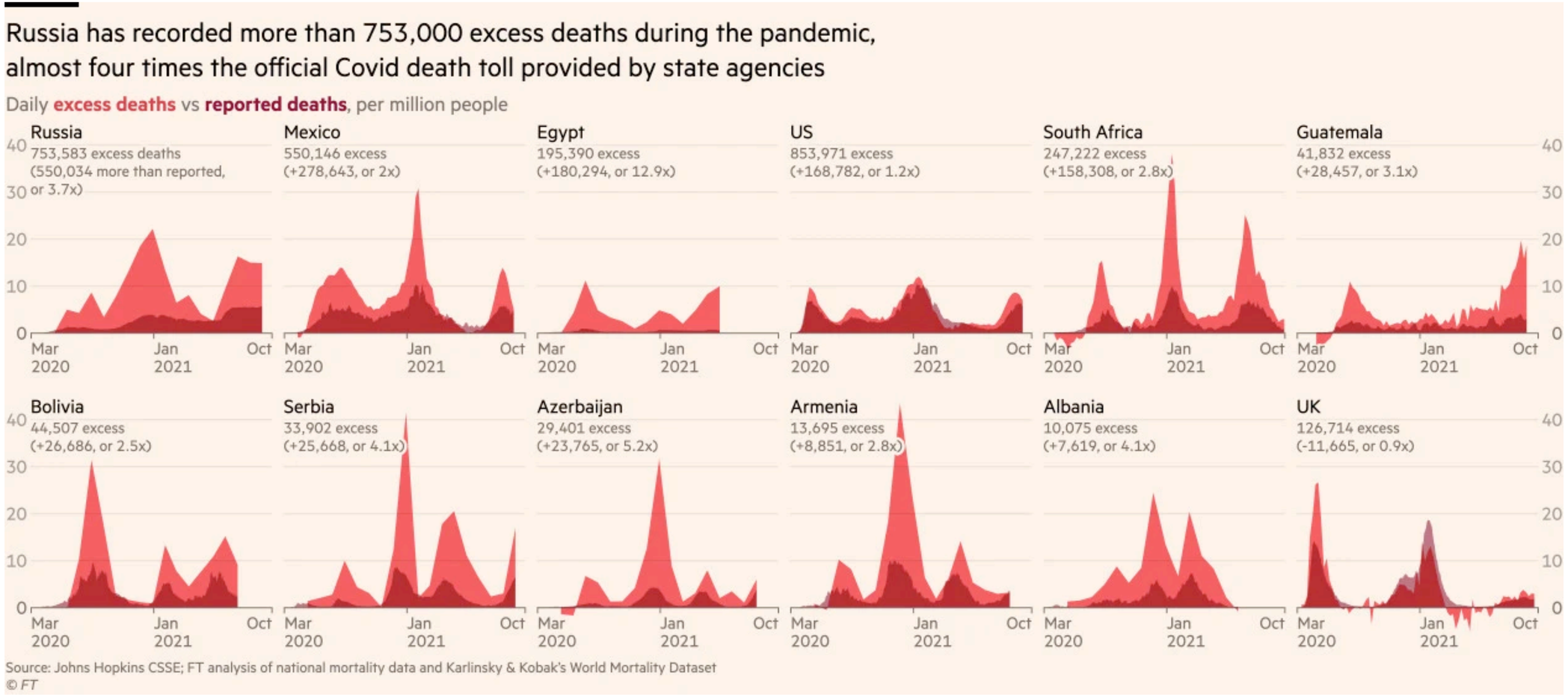
- Leverage how we read plots (Z)
- Call to action (tell the reader what do do)

Middle: What is the point of the plot?

- Use pre attentive attributes: Guide the attention
- Create emotions using color
- Based on conflict/danger

Conclusion:

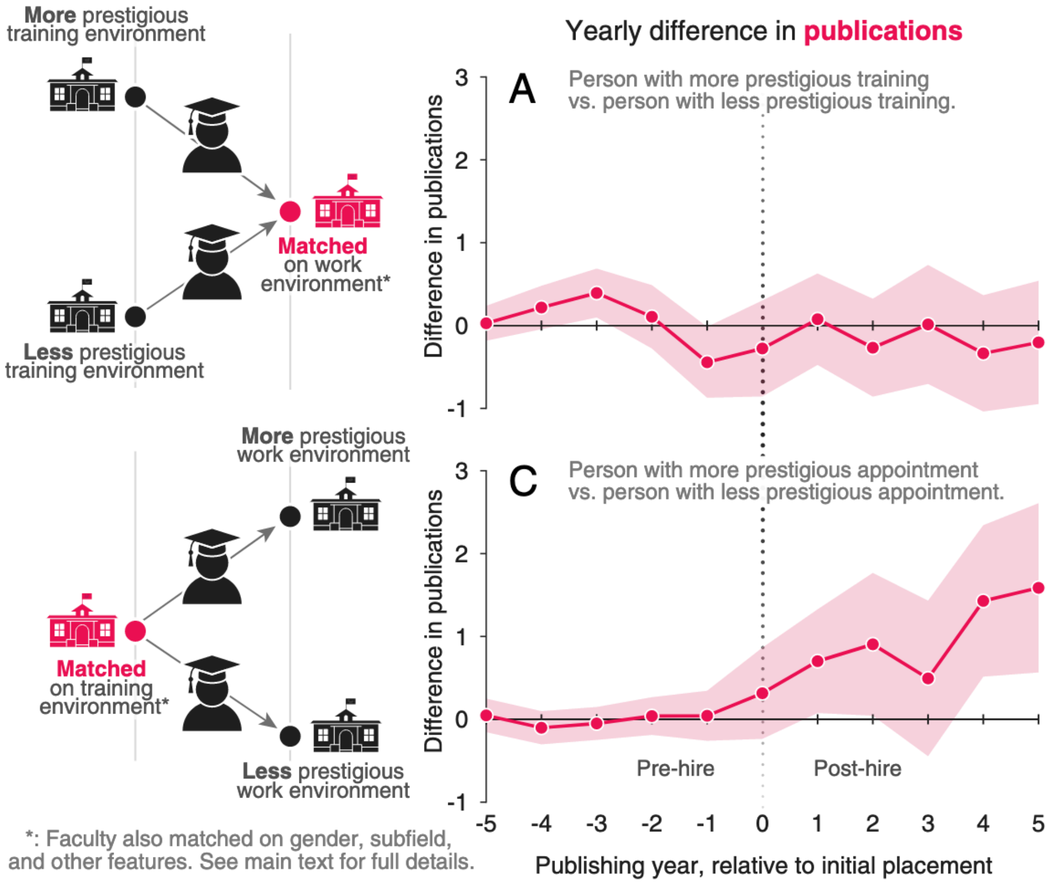
- They should get the message of the plot



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

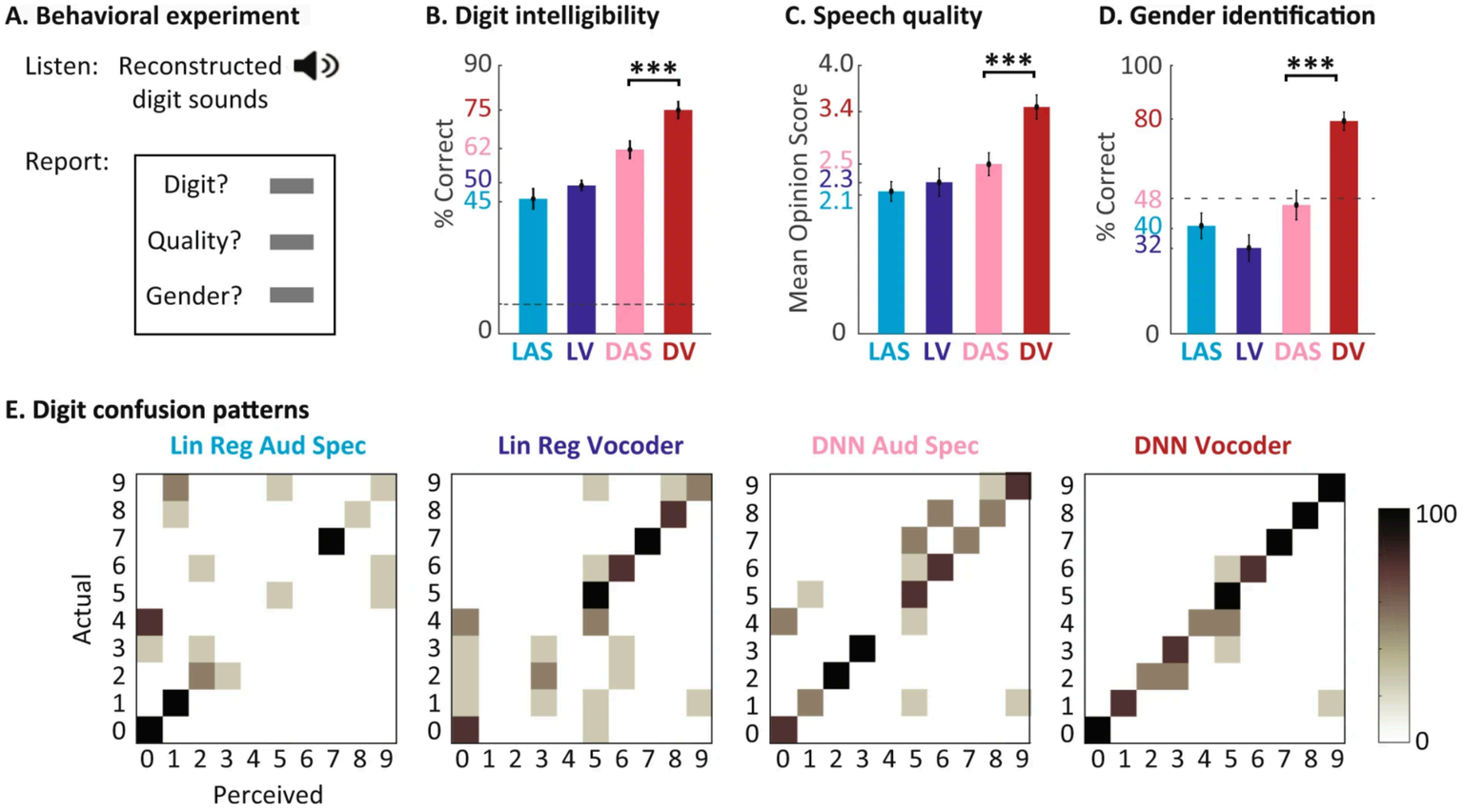


Exposition: Label and drawings.

Conflict: Between both matched persons.

Resolution: Reader has learned that prestige helps get published.

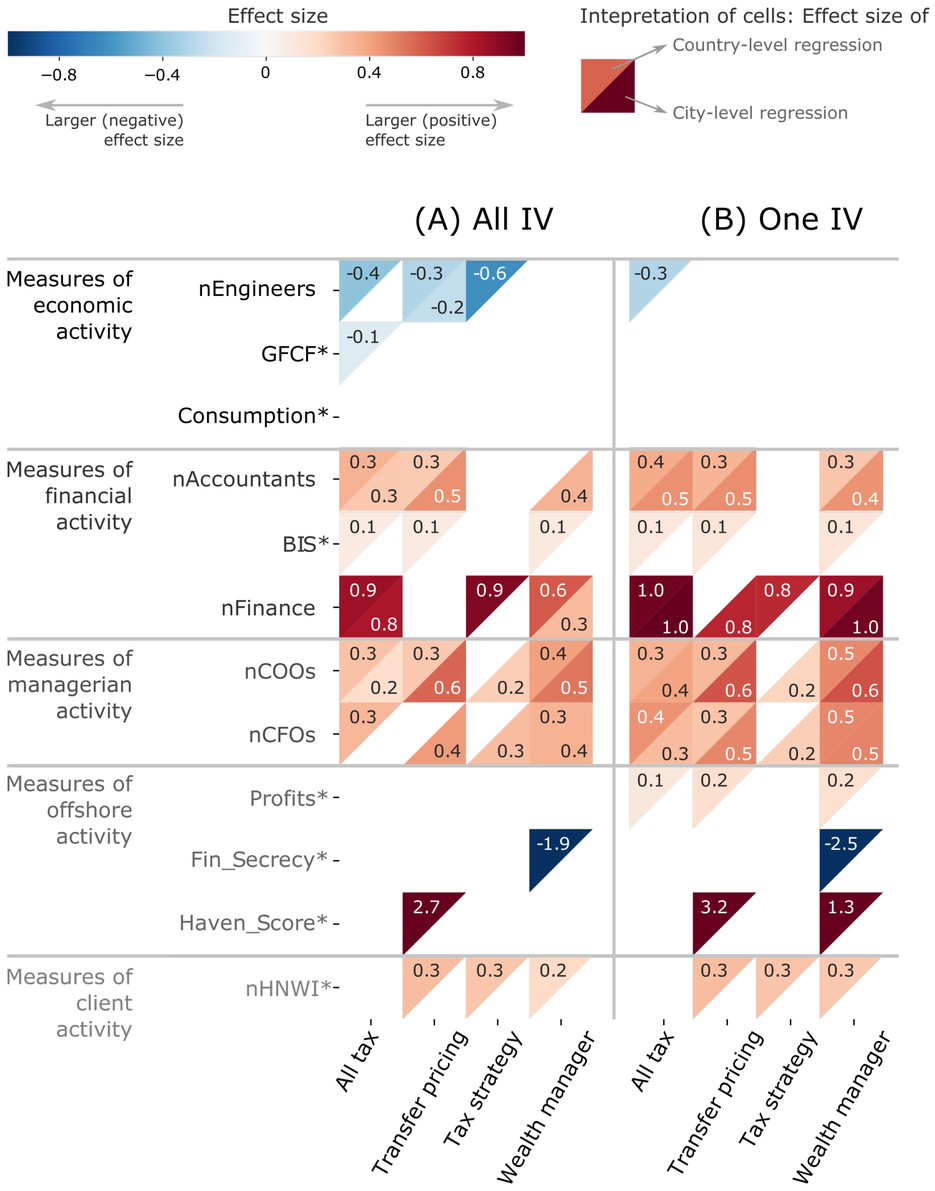
https://www.pnas.org/content/pnas/116/22/10729.full.pdf



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

PRACTICE: TELL A STORY

- (1) What is the exposition, conflict and resolution in your graph?
- (2) How will your graph be perceived by the audience? What moods and emotions might users experience as they engage with your work? (not frustrated!)
- (3) How can you guide them in the interpretation?

- Steps:
  - Do (1) – 5 minutes. Mix groups and do (2) – 10 min. Return to original groups to do (3) – 10 min