Data Communication

Graphics that tell stories in an engaging way

Cédric Scherer // rstudio::conf // July 2022

Data Visualization

is any graphical representation
of information and data.

Cédric Scherer // rstudio::conf // July 2022

Data Visualization

converts information into visual forms as quantifiable features.

Cédric Scherer // rstudio::conf // July 2022

Data Visualization

helps to amplify cognition, gain insights, discover, explain, and make decisions.

Cédric Scherer // rstudio::conf // July 2022

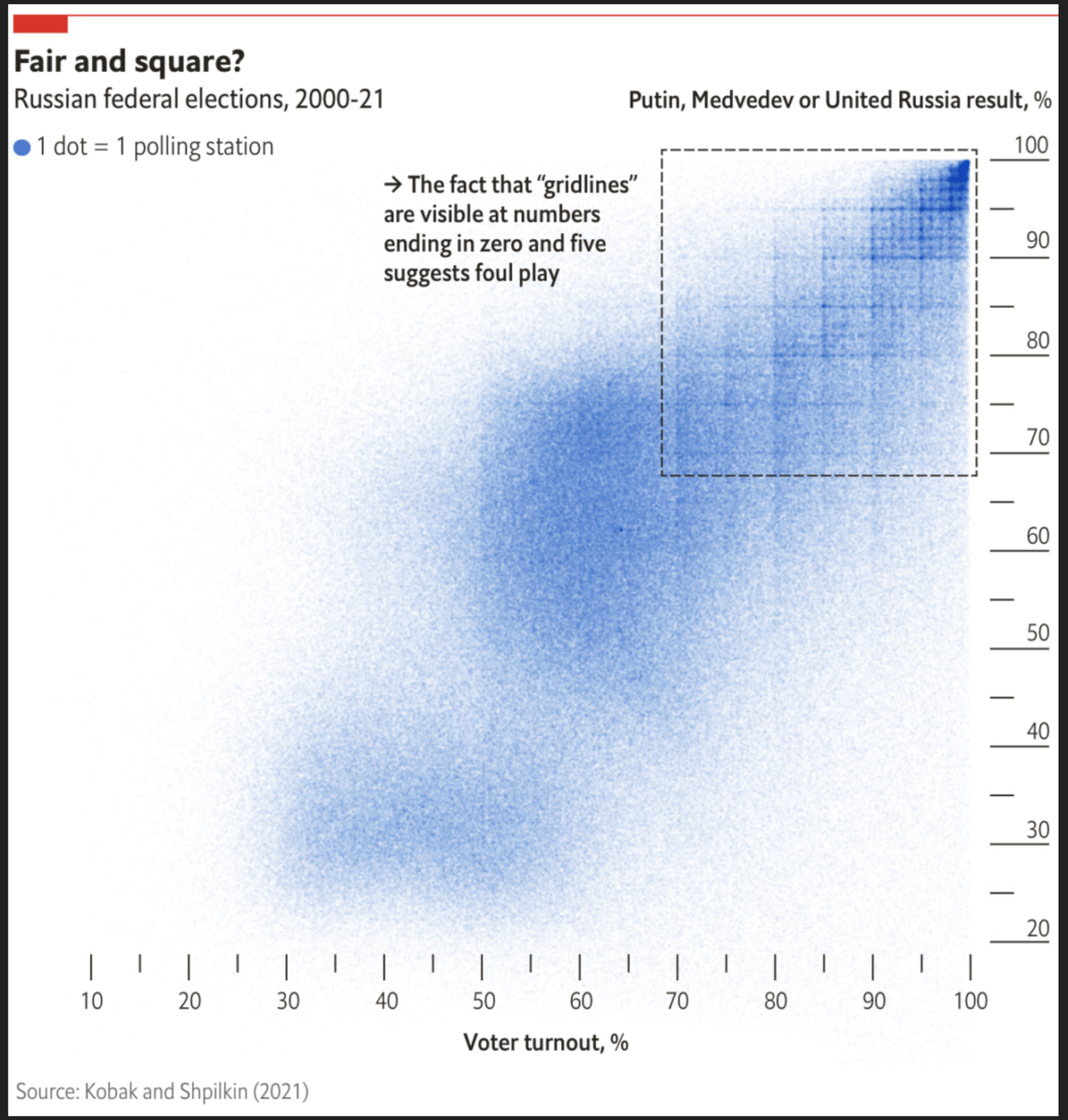
Visualize Your Data

Cédric Scherer // rstudio::conf // July 2022

Visualize Your Data

"When Dmitry Kobak and Sergey Shpilkin [...] analysed the results, they found that **an unusually high number of turnout and vote-share results were multiples of five** (eg, 50%, 55%, 60%), a tell-tale sign of manipulation."

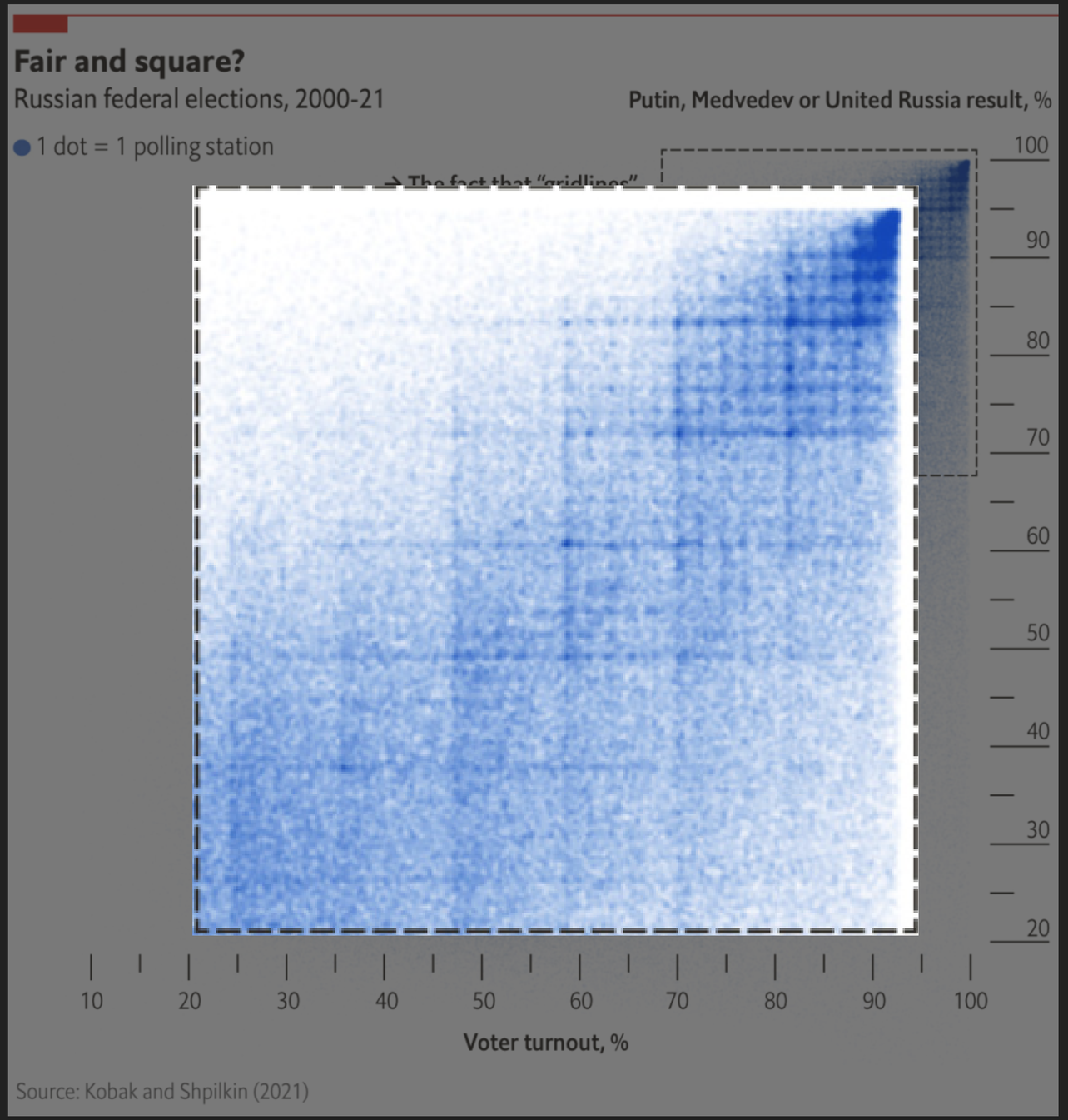
"Russian elections once again had a suspiciously neat result"
by The Economist



Visualize Your Data

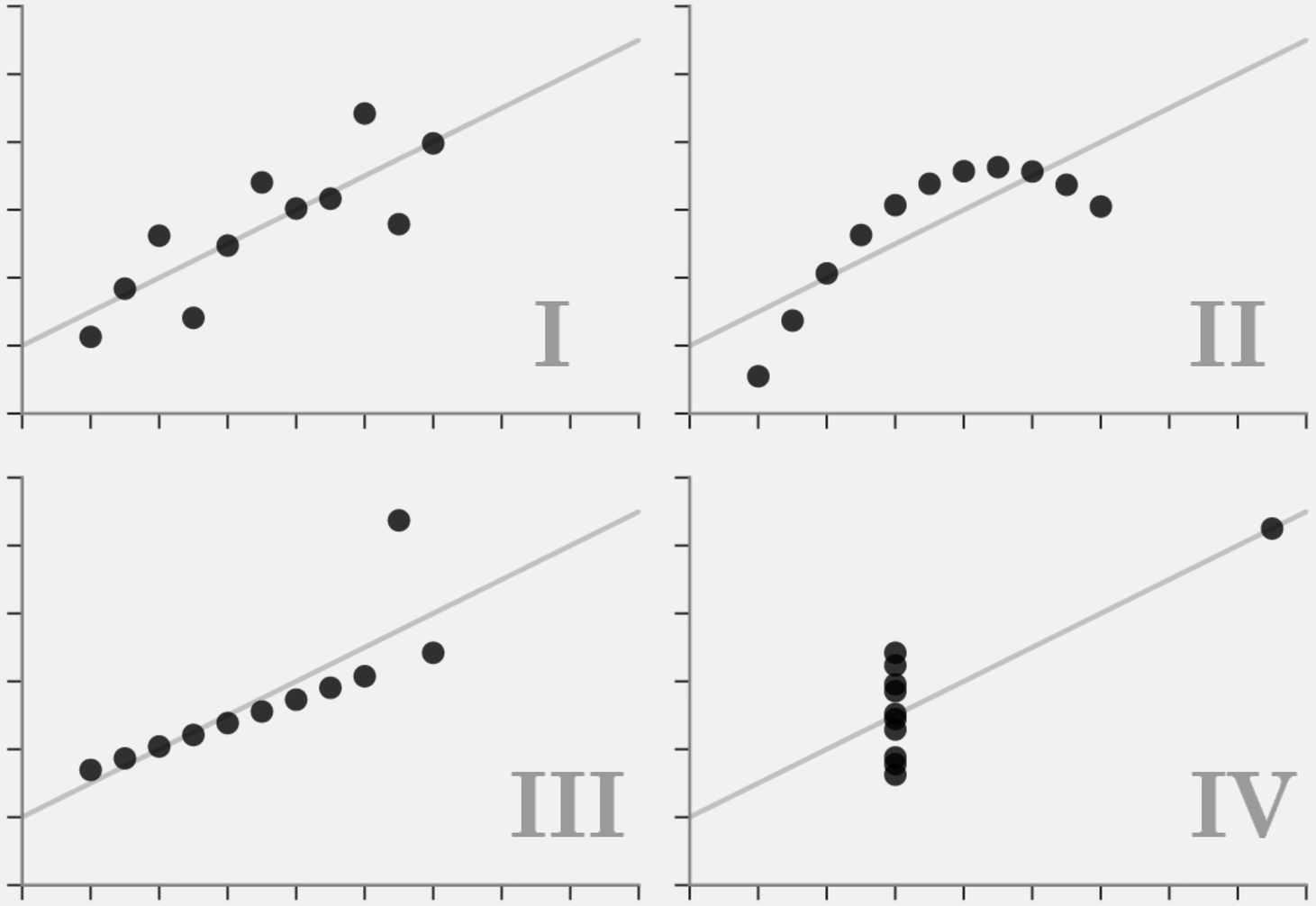
"When Dmitry Kobak and Sergey Shpilkin [...] analysed the results, they found that **an unusually high number of turnout and vote-share results were multiples of five** (eg, 50%, 55%, 60%), a tell-tale sign of manipulation."

"Russian elections once again had a suspiciously neat result"
by The Economist



Anscombe's Quartet

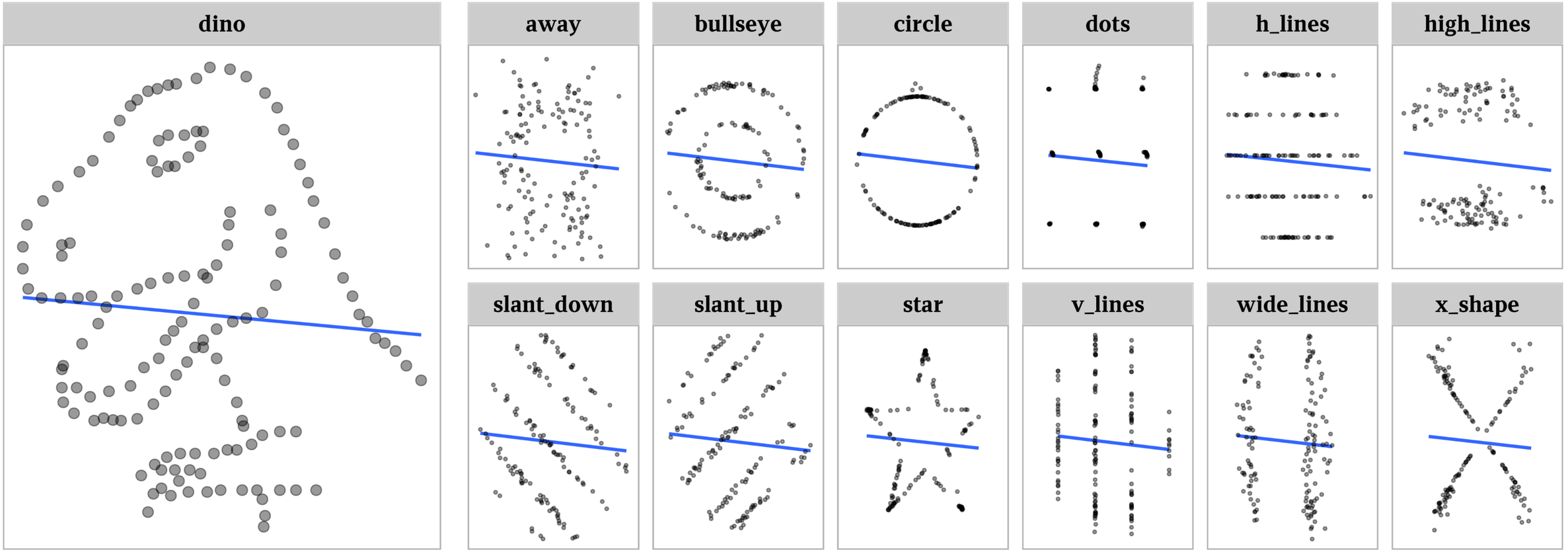
each dataset has the same summary statistics
mean, standard deviation, and correlation
but are visually distinct.



"Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing"
by Justin Matejka & George Fitzmaurice , ACM SIGCHI Conference on Human Factors in Computing Systems 2017

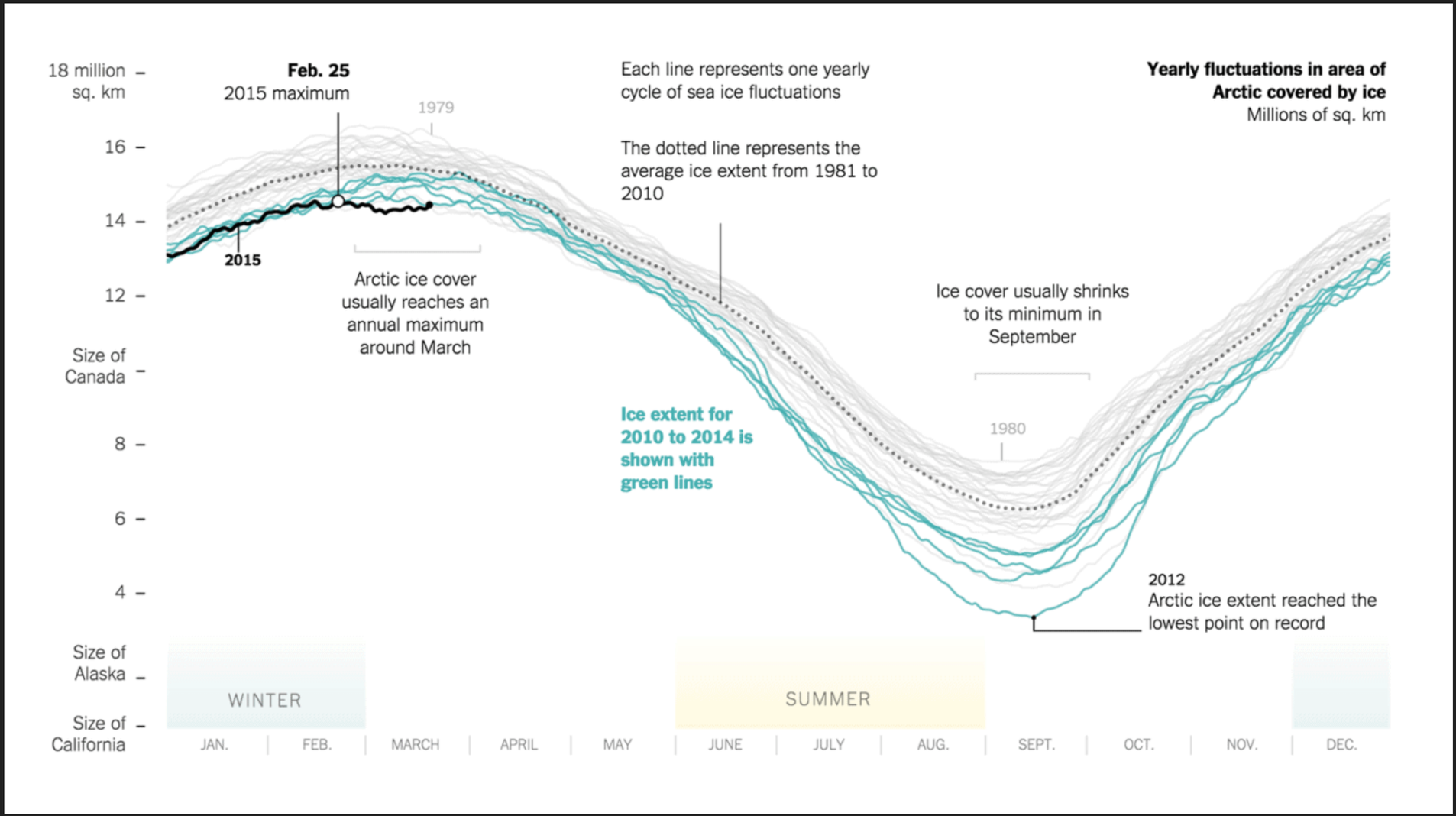
Datasaurus Dozen

The Datasaurus by Alberto Cairo shows us why visualisation is important, not just summary statistics.



"Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing"
by Justin Matejka & George Fitzmaurice , ACM SIGCHI Conference on Human Factors in Computing Systems 2017

What makes it a good data visualization?



"Yearly Fluctuations in Area of Arctic Covered by Ice" by Derek Watkins (New York Times)

What makes it a good data visualization

☛ **INFORMATION** (integrity)

What makes it a good data visualization

☛ **INFORMATION** (integrity)

☛ **STORY** (interestingness)

What makes it a good data visualization

👉 **INFORMATION** (integrity)

👉 **STORY** (interestingness)

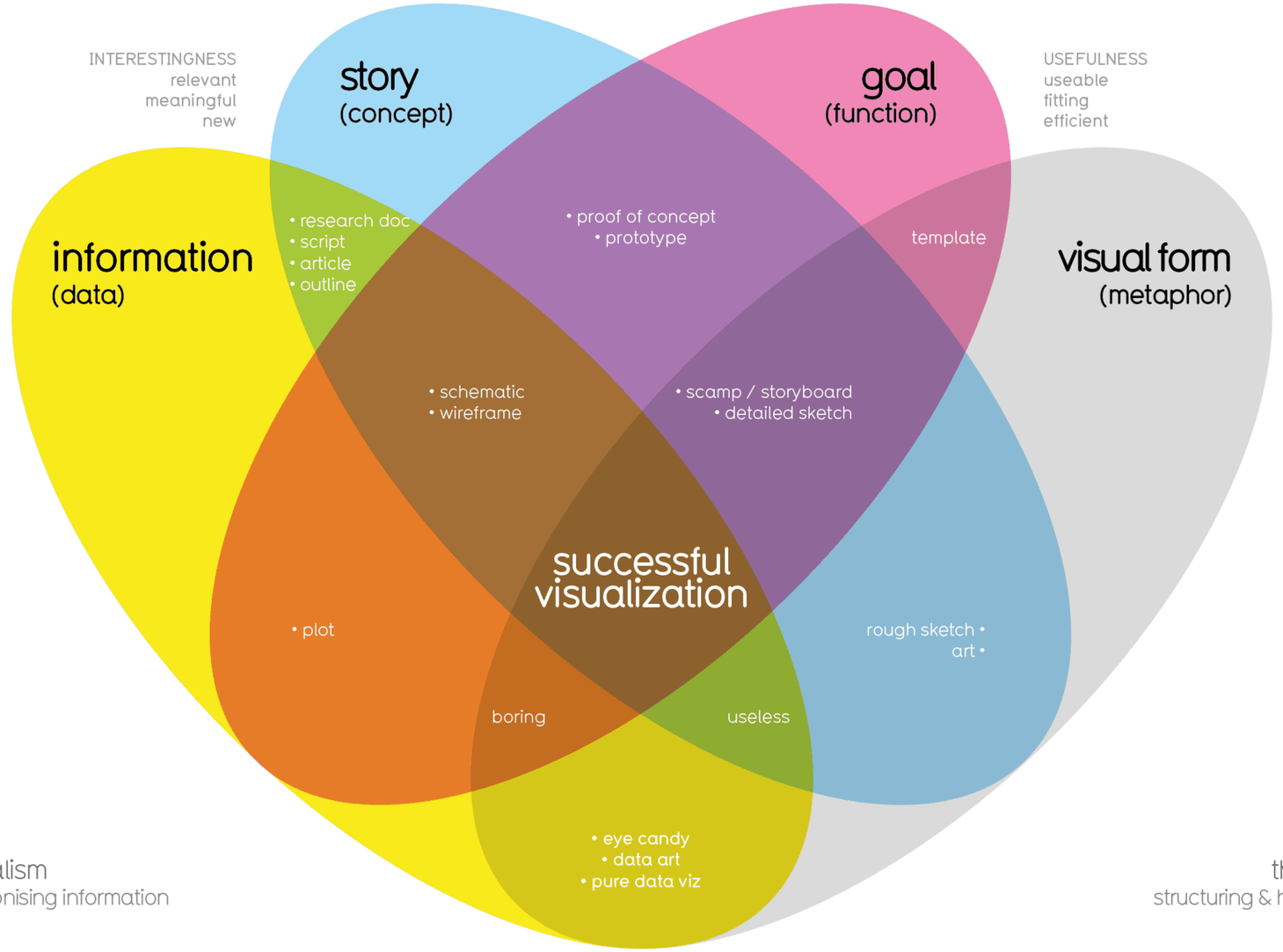
👉 **GOAL** (usefulness)

What makes it a good data visualization

- ☛ **INFORMATION** (integrity)
- ☛ **STORY** (interestingness)
- ☛ **GOAL** (usefulness)
- ☛ **VISUAL FORM** (beauty)

What Makes a Good Visualization?

explicit (implicit)



information
the art of journalism
structuring & harmonising information

visualization
the art of design
structuring & harmonising visuals

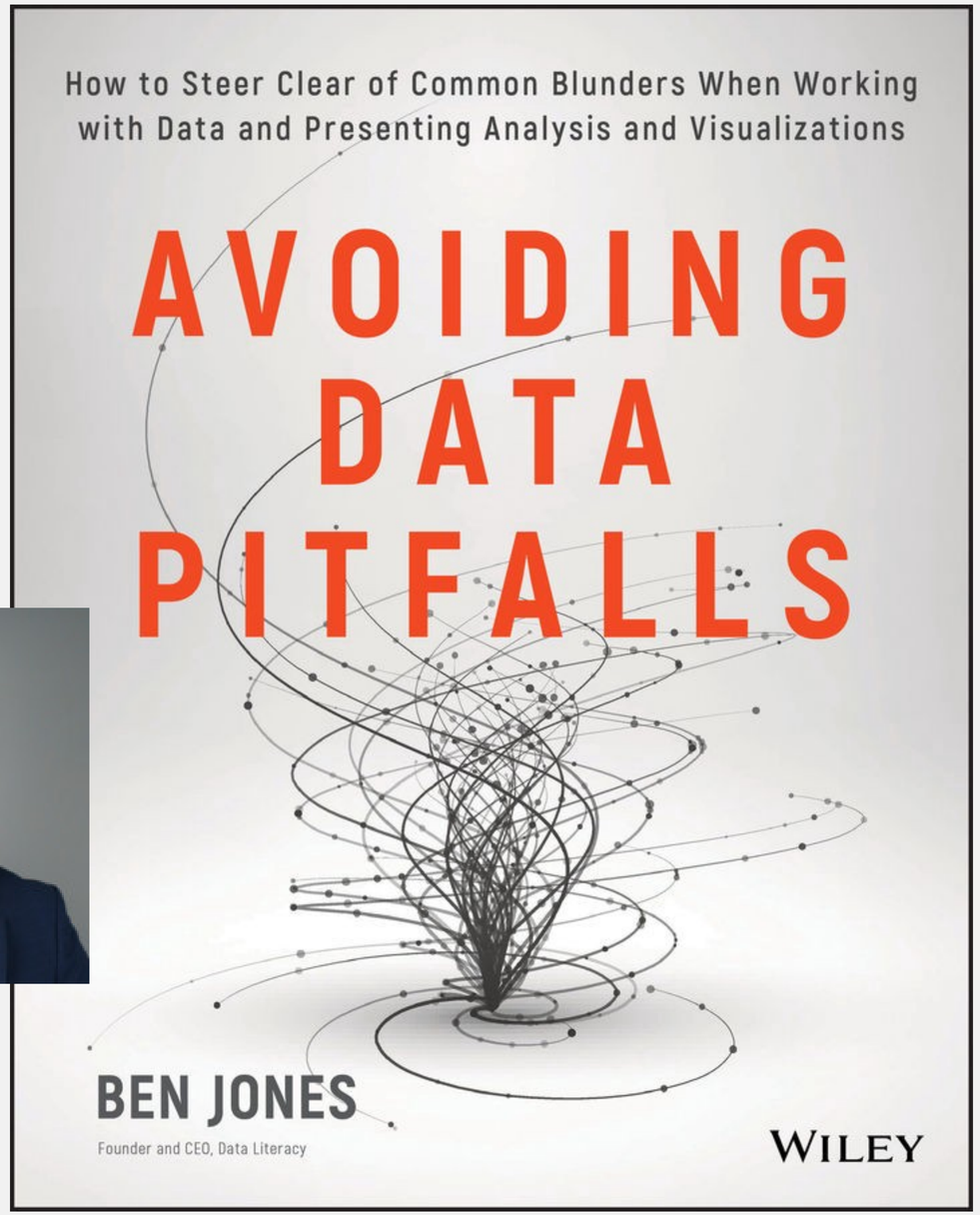
David McCandless
InformationisBeautiful.net

taken from new book
Knowledge is Beautiful

find out more
bit.ly/KIB_Books

Information

Understand your data and be accurate


How to Steer Clear of Common Blunders When Working with Data and Presenting Analysis and Visualizations
AVOIDING DATA PITFALLS
BEN JONES
Founder and CEO, Data Literacy
WILEY

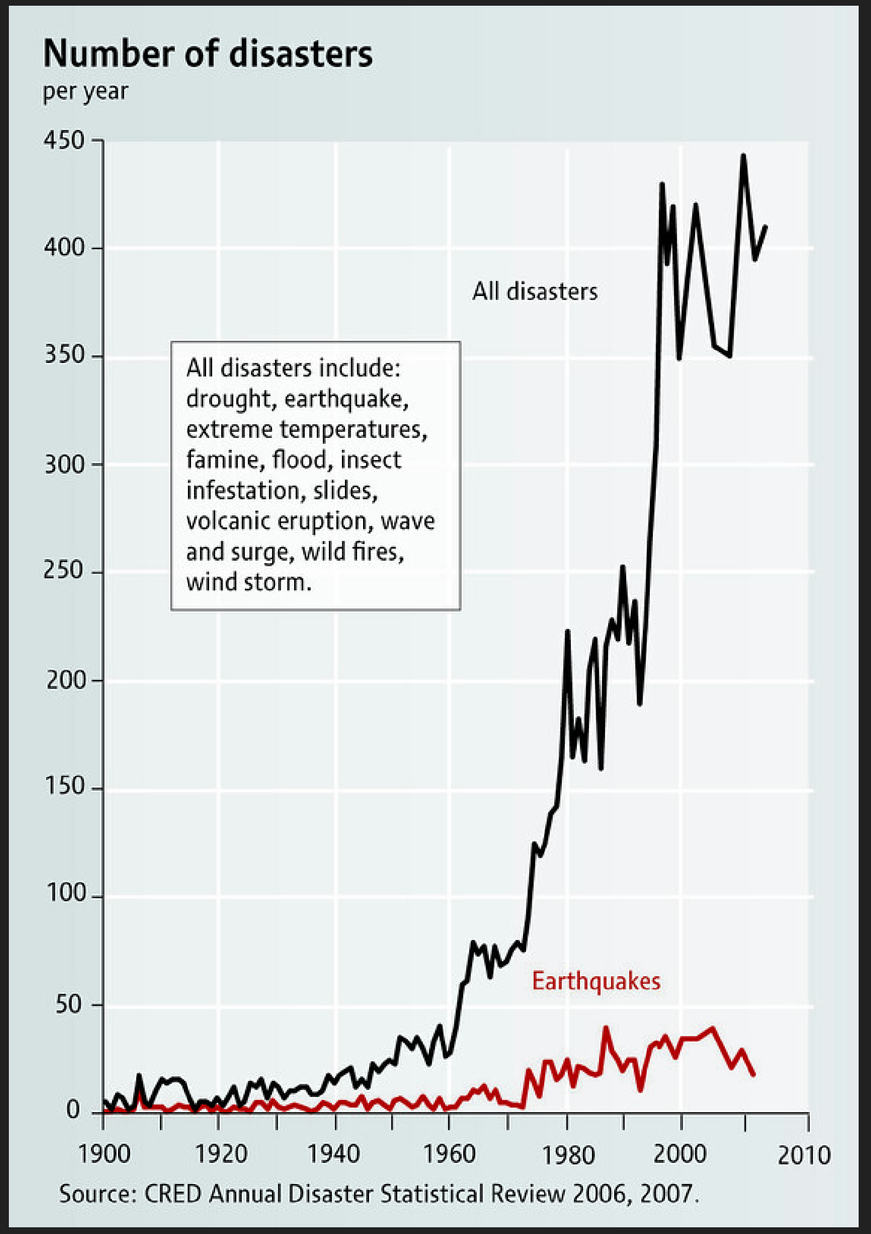
Our data is never a perfect reflection of the real world.

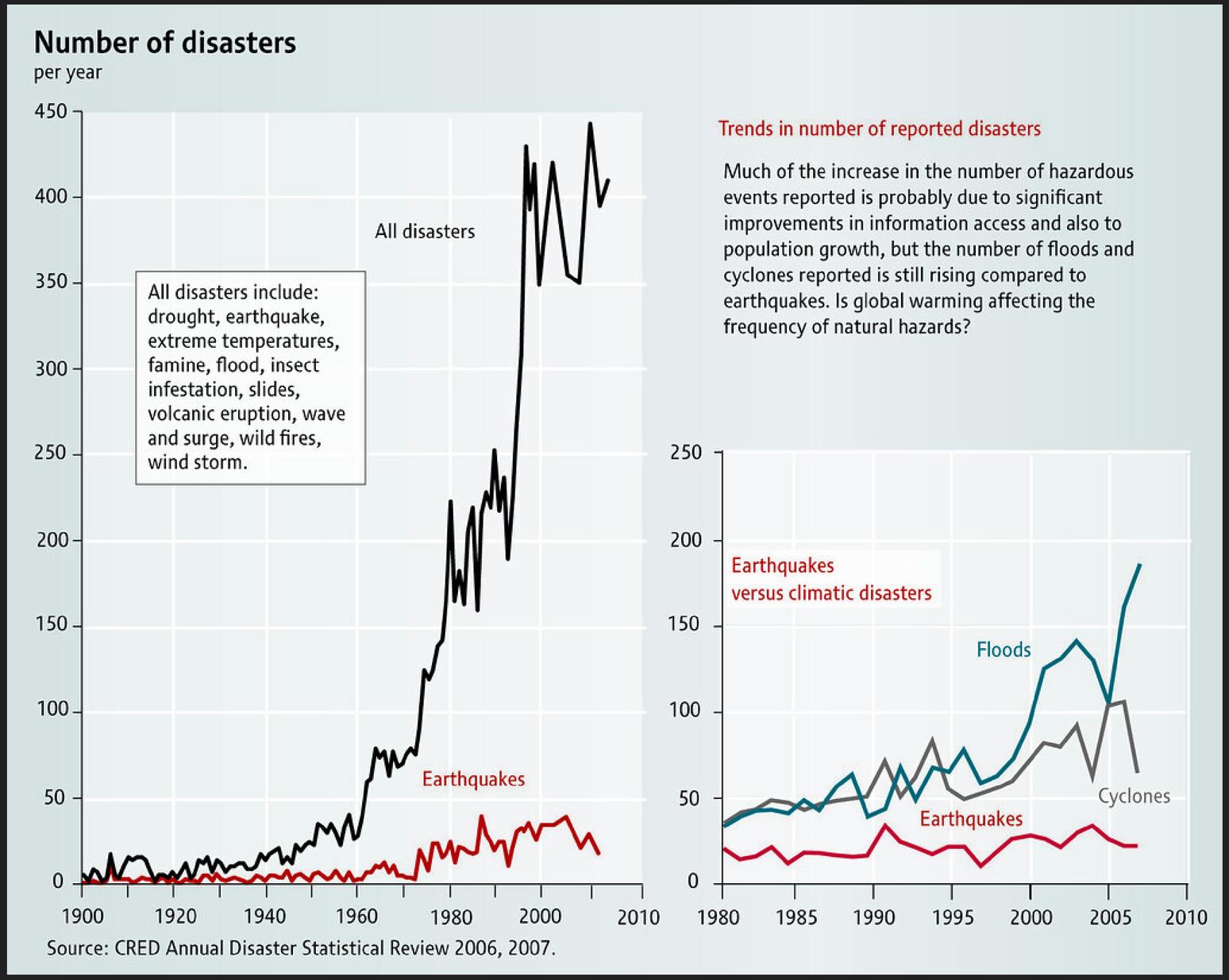
Our data is never a perfect reflection of the real world.

➔ **only a subset:** not crime but reported crime

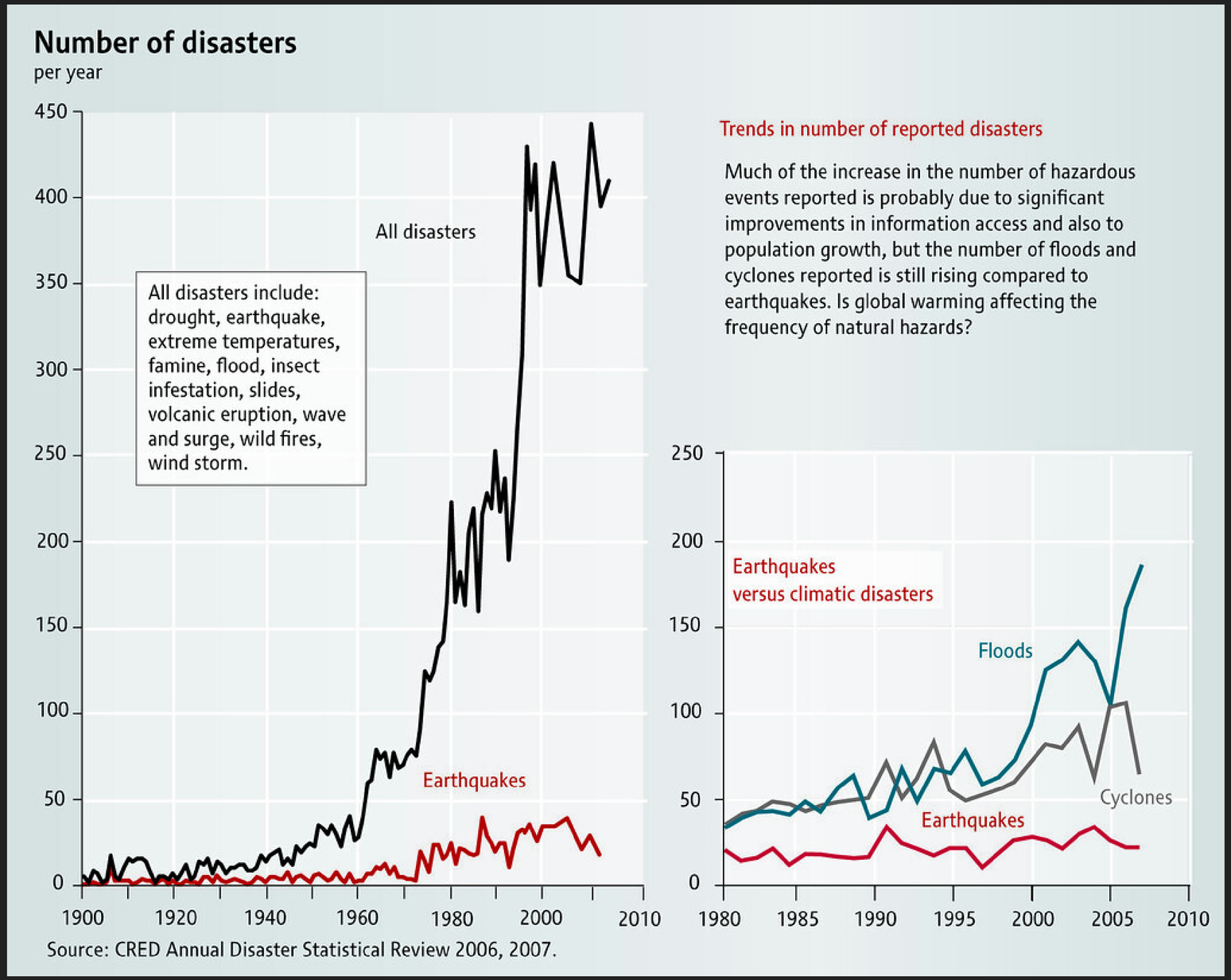
➔ **collected by humans:** guesstimation, precision and errors

➔ **collected by machines:** precisions and errors

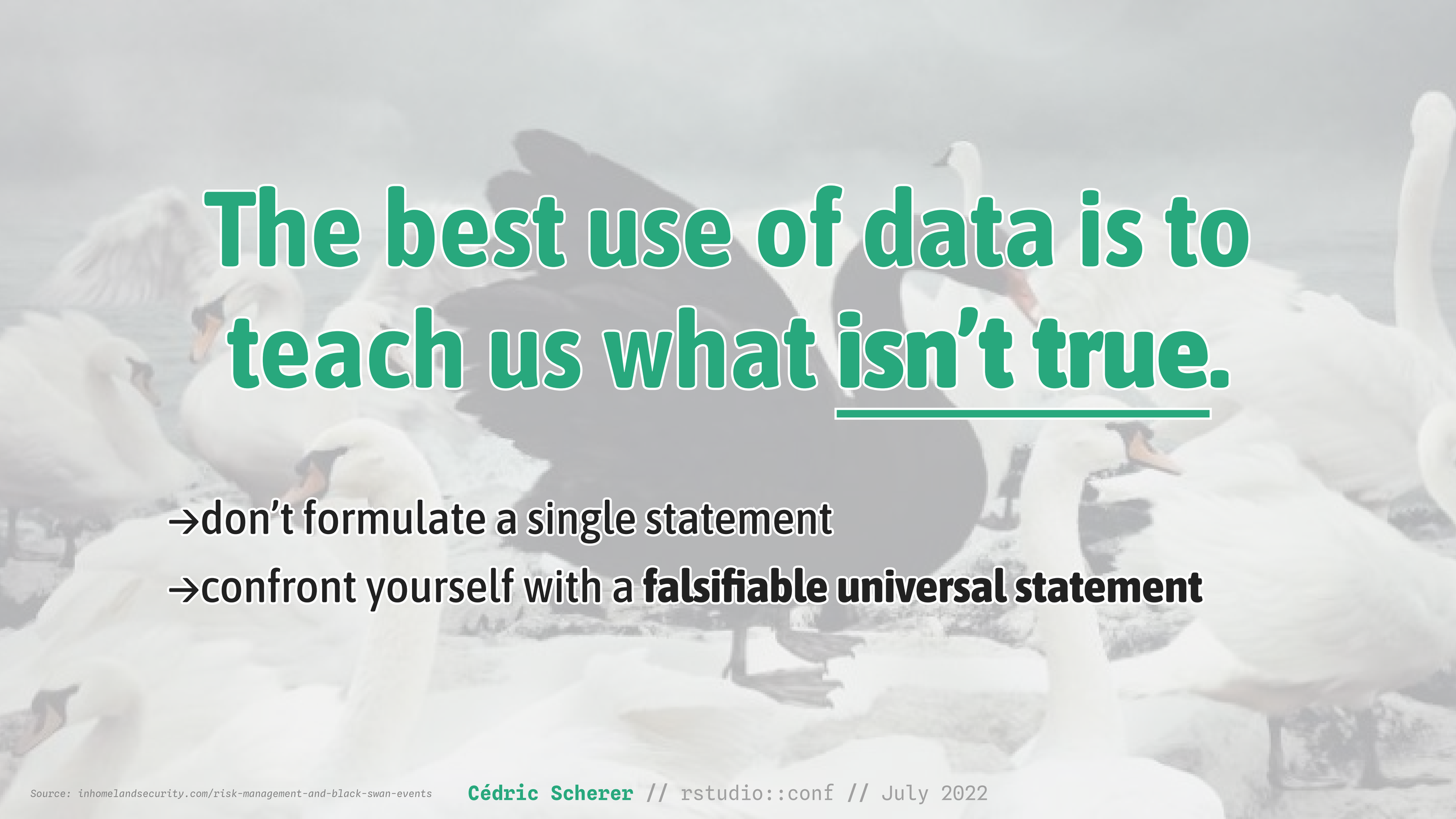

Number of disasters
per year
450
400
350
300
250
200
150
100
50
0
All disasters
All disasters include: drought, earthquake, extreme temperatures, famine, flood, insect infestation, slides, volcanic eruption, wave and surge, wild fires, wind storm.
Earthquakes
1900
1920
1940
1960
1980
2000
2010
Source: CRED Annual Disaster Statistical Review 2006, 2007.


Number of disasters
per year
450
400
350
300
250
200
150
100
50
0
All disasters
All disasters include: drought, earthquake, extreme temperatures, famine, flood, insect infestation, slides, volcanic eruption, wave and surge, wild fires, wind storm.
Earthquakes
1900
1920
1940
1960
1980
2000
2010
Trends in number of reported disasters
Much of the increase in the number of hazardous events reported is probably due to significant improvements in information access and also to population growth, but the number of floods and cyclones reported is still rising compared to earthquakes. Is global warming affecting the frequency of natural hazards?
Earthquakes versus climatic disasters
Floods
Cyclones
Earthquakes
250
200
150
100
50
0
1980
1985
1990
1995
2000
2005
2010
Source: CRED Annual Disaster Statistical Review 2006, 2007.

"Much of the increase of **hazardous events reported** is probably due to significant **improvements in information access**"



The best use of data is to teach us what isn't true.



The best use of data is to teach us what isn't true.

→don't formulate a single statement

→confront yourself with a **falsifiable universal statement**

Source: inhomelandsecurity.com/risk-management-and-black-swan-events

The best use of data is to teach us what isn't true.

Source: inhomelandsecurity.com/risk-management-and-black-swan-events

Story

Be clear about the message of your visualization

Who is my audience?

Who is my audience?

Which story is **interesting** for them?

Who is my audience?

Which story is **interesting** for them?

What are **relevant** details to include?

Who is my audience?

Which story is **interesting** for them?

What are **relevant** details to include?

Which variables are **meaningful** to them?

Who is my audience?

Which story is **interesting** for them?

What are **relevant** details to include?

Which variables are **meaningful** to them?

How will they **encounter** the visualization?

Who is my audience?

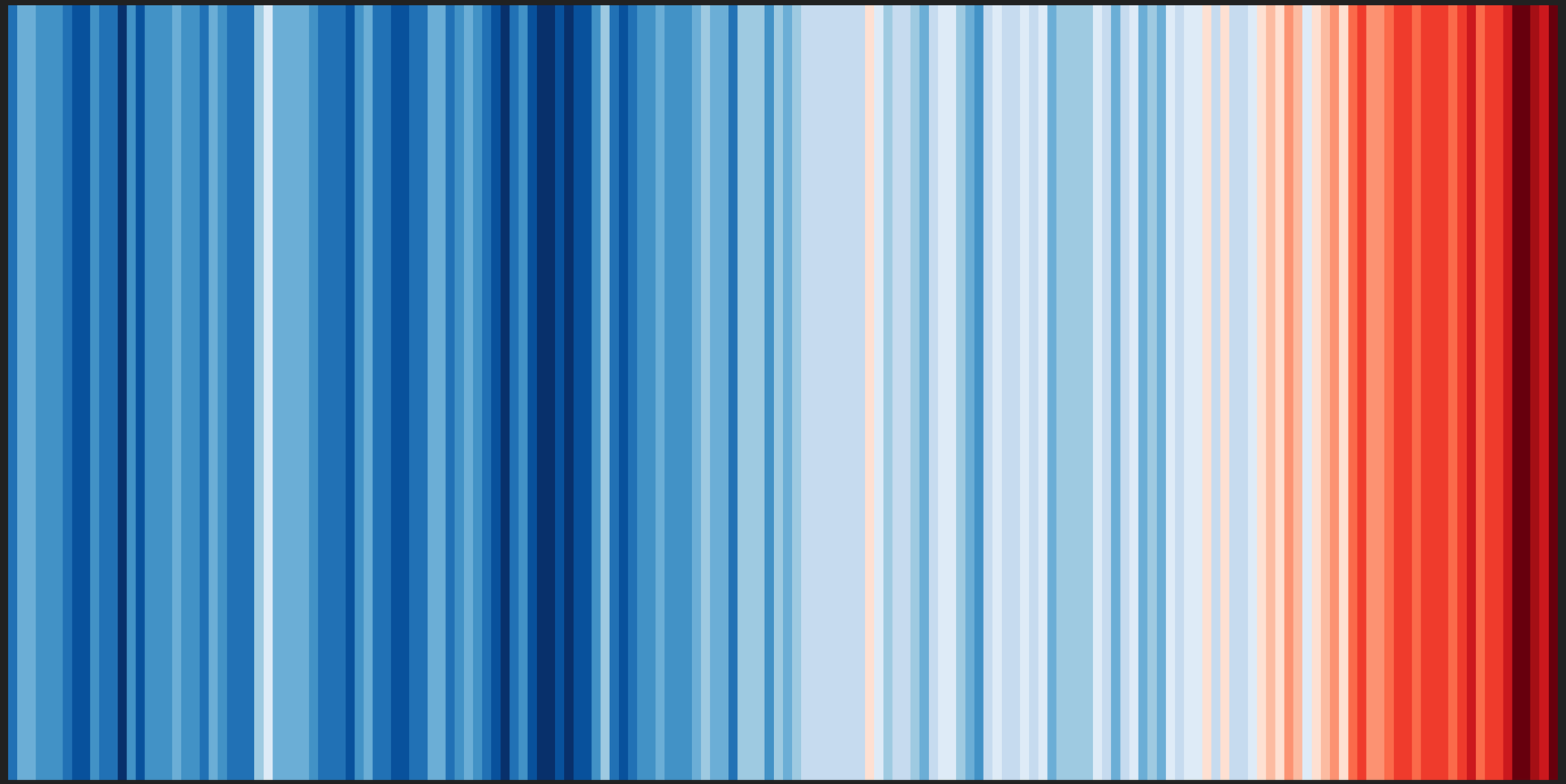
Which story is **interesting** for them?

What are **relevant** details to include?

Which variables are **meaningful** to them?

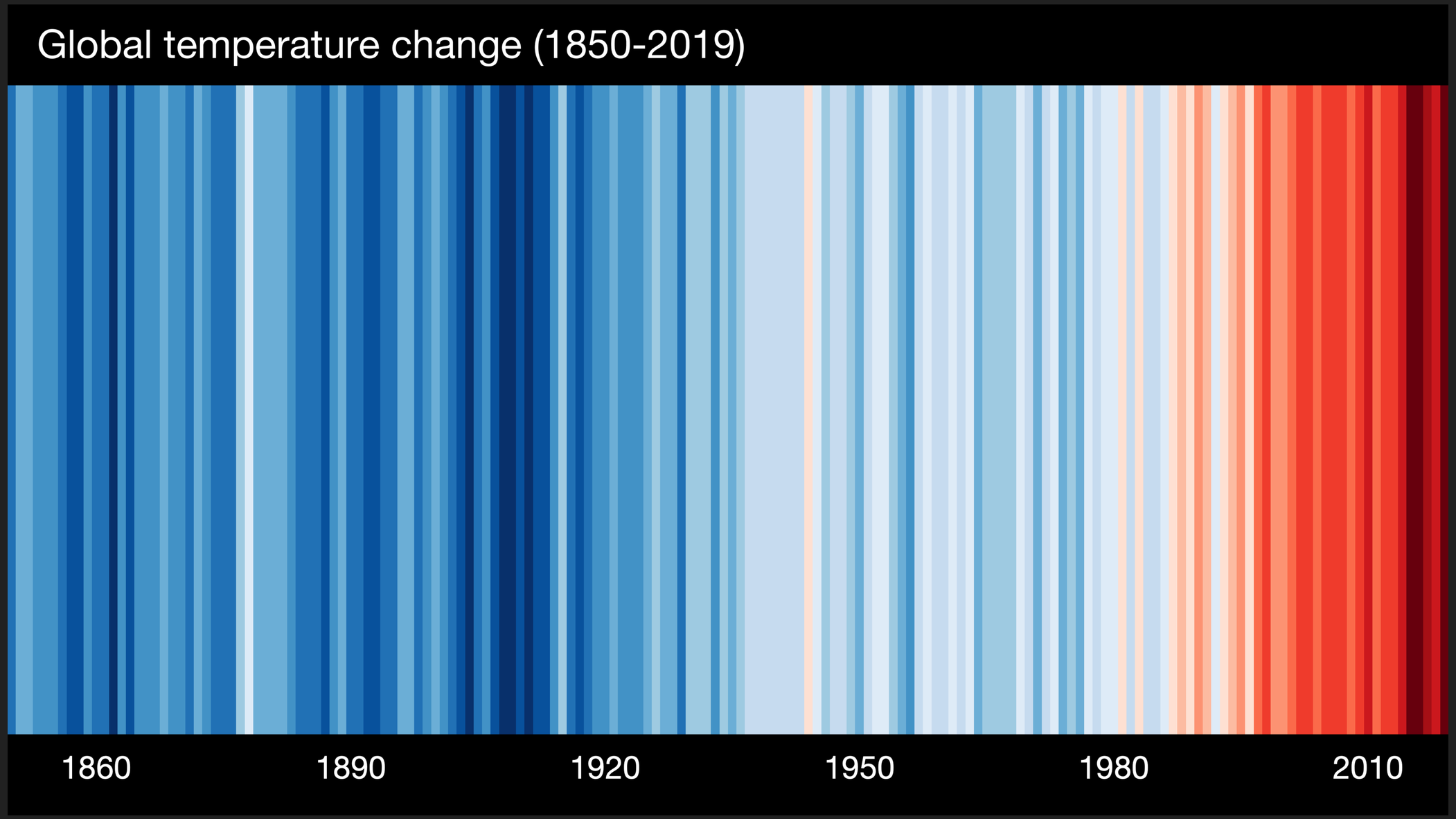
How will they **encounter** the visualization?

Do I need a visualization at all??

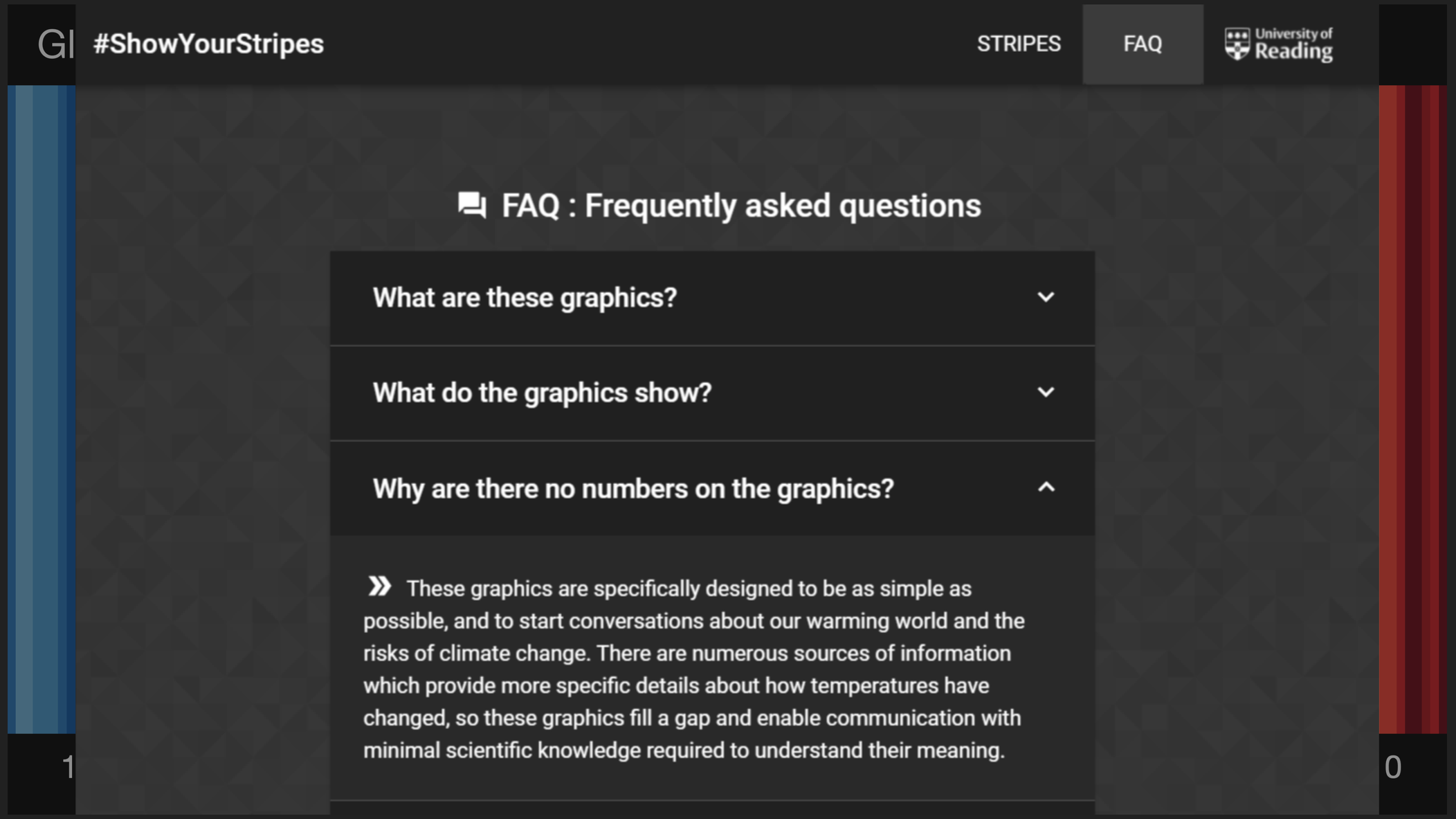


Warming Stripes *by Ed Hawkins*

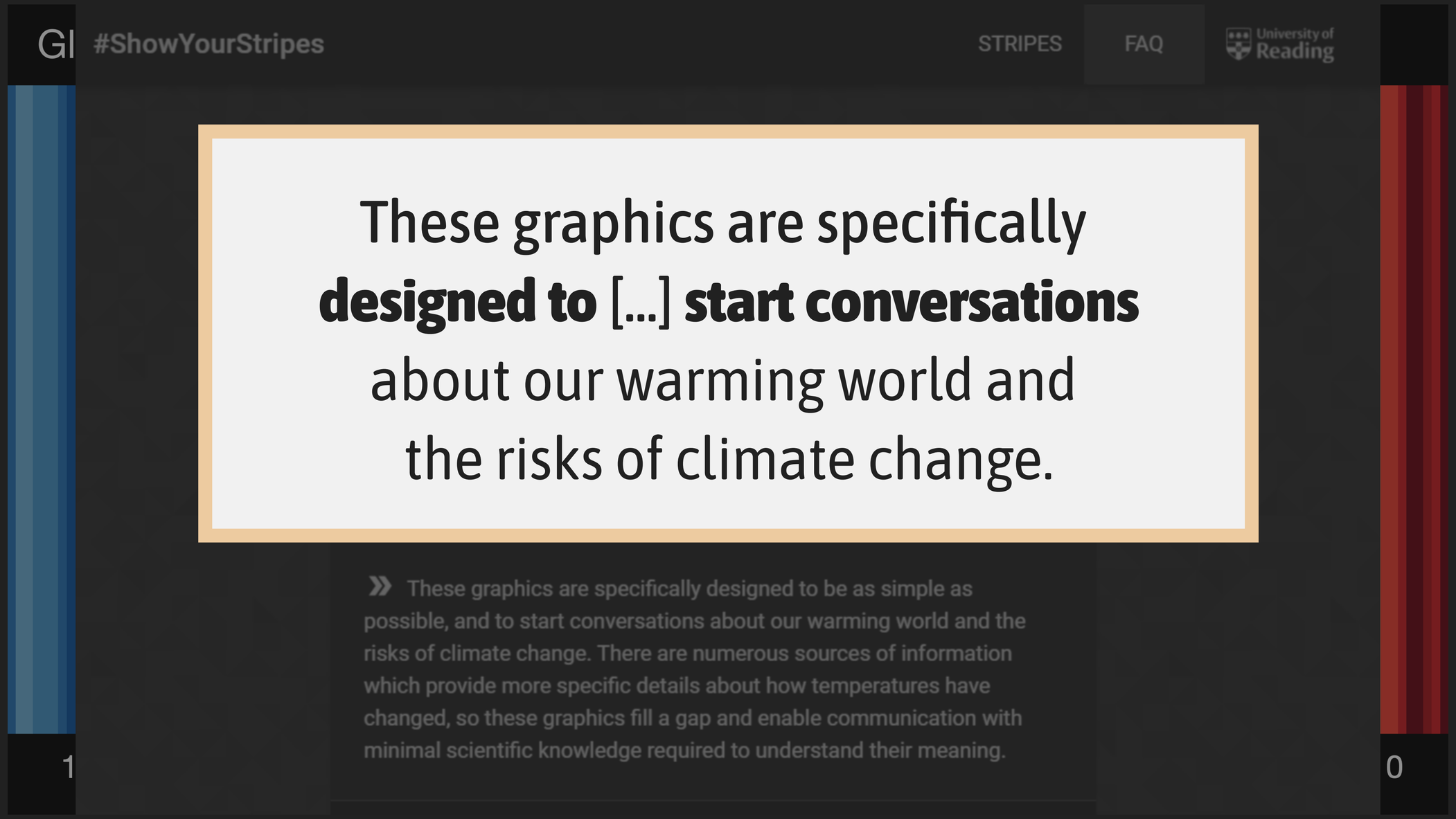
Cédric Scherer // rstudio::conf // July 2022



Warming Stripes *by Ed Hawkins*

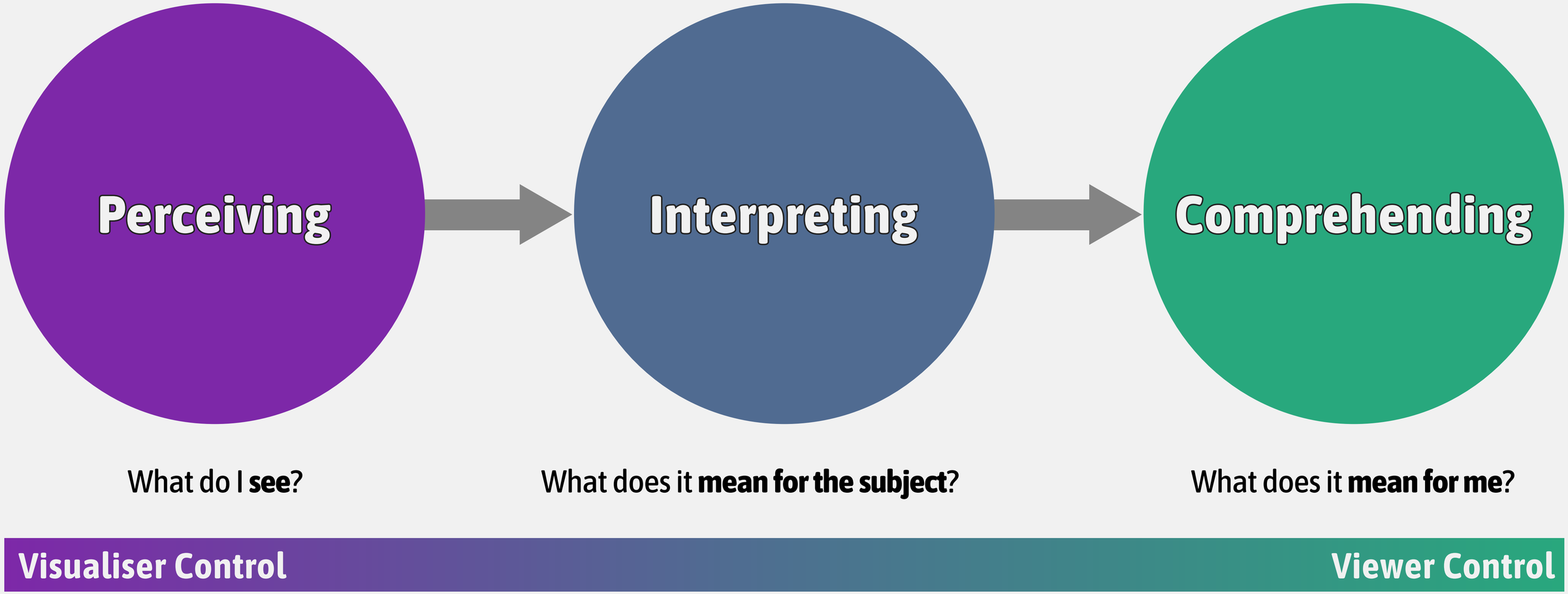


showyourstripes.info/faq



These graphics are specifically
designed to [...] **start conversations**
about our warming world and
the risks of climate change.

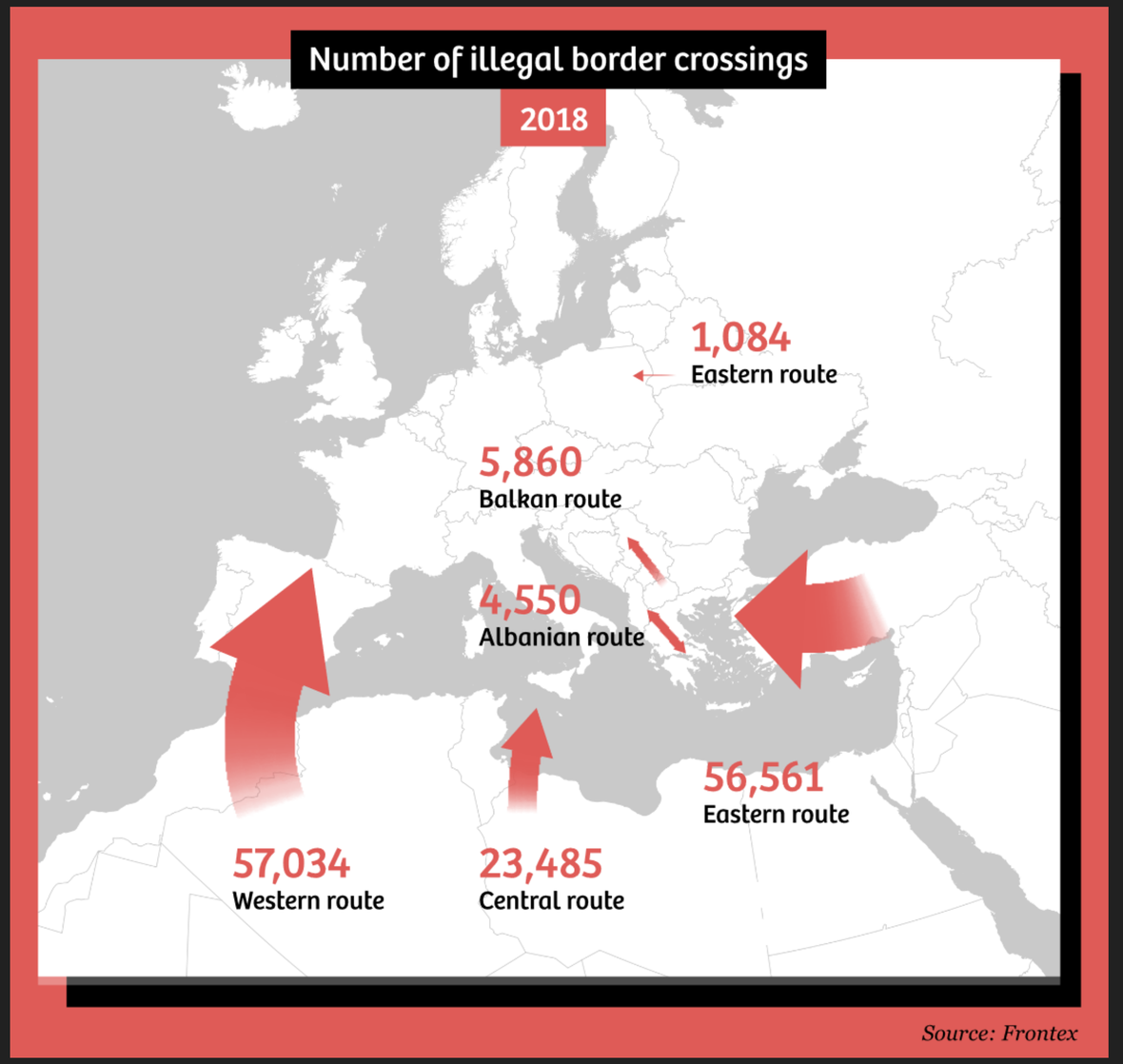
showyourstripes.info/faq



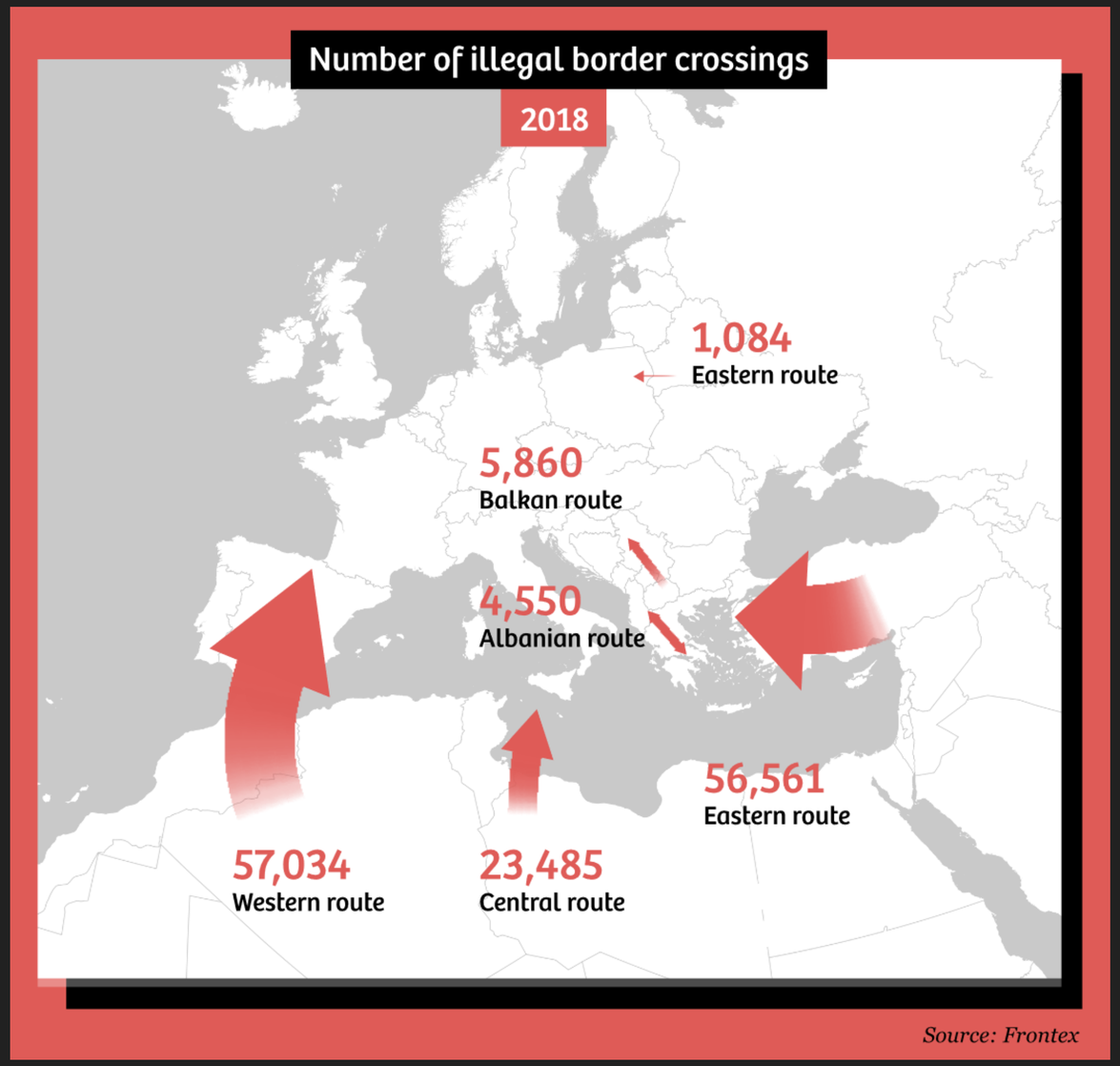
Scheme by Andy Kirk

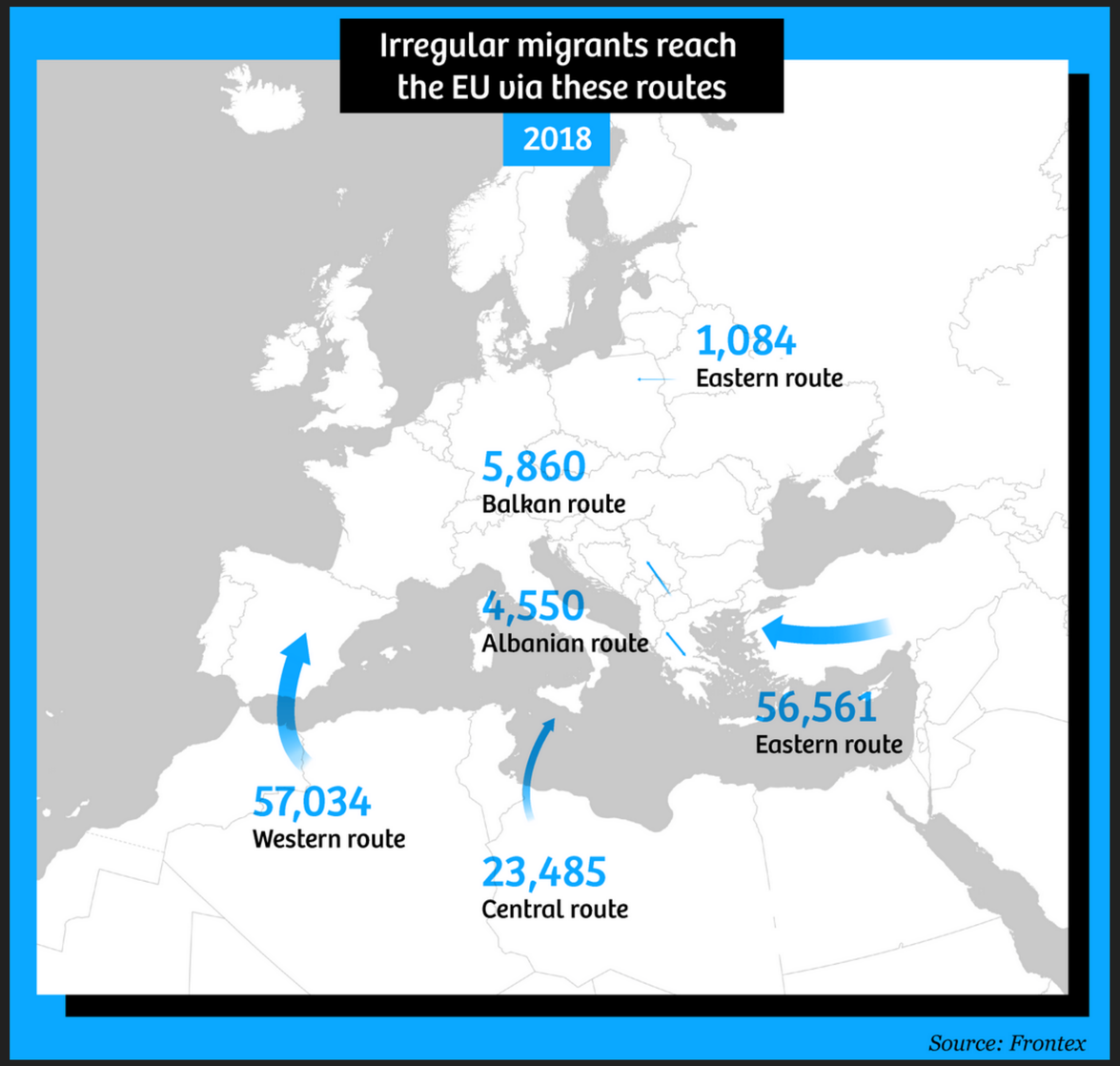
Goal

Select charts that successfully transport your story

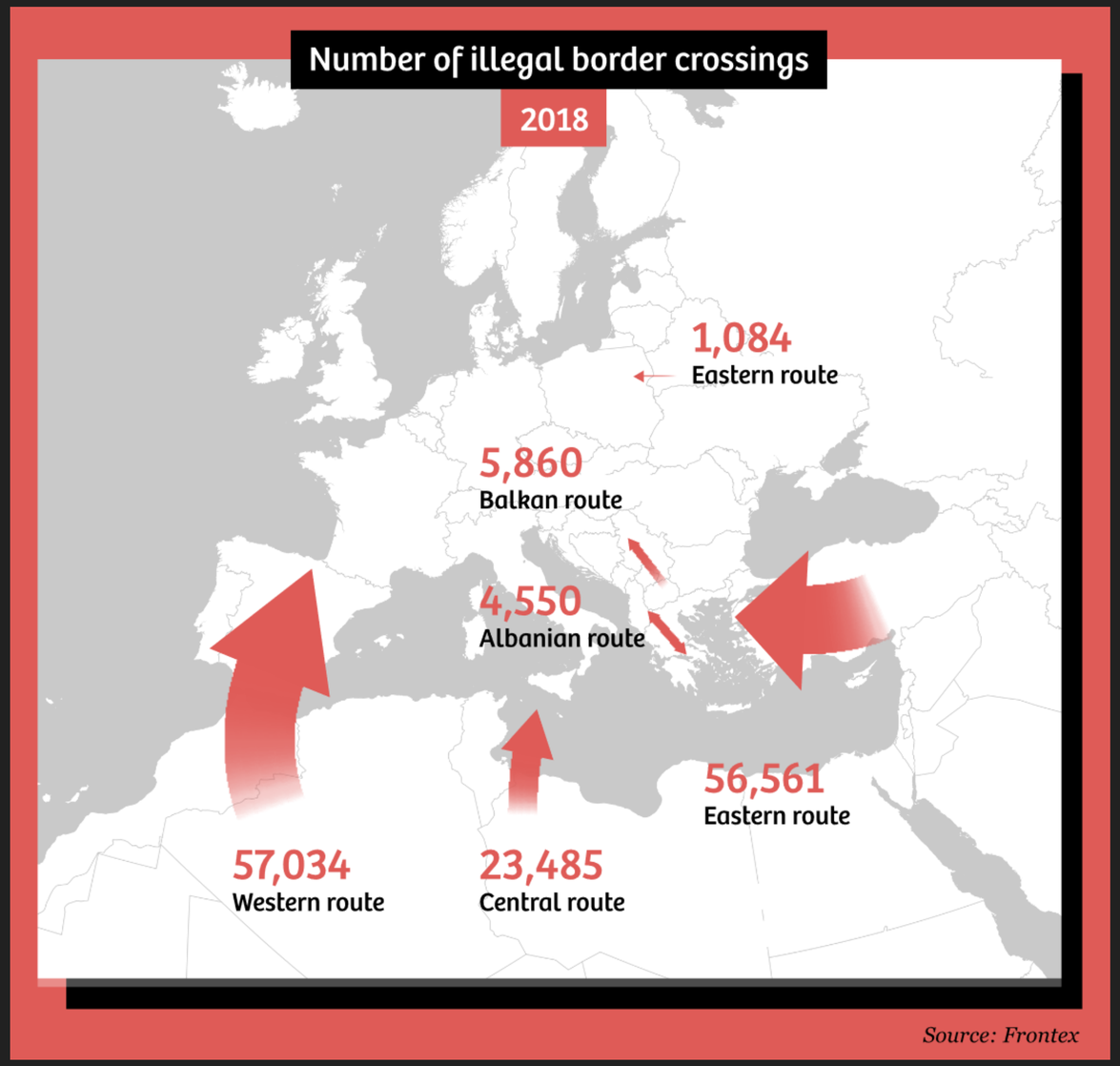


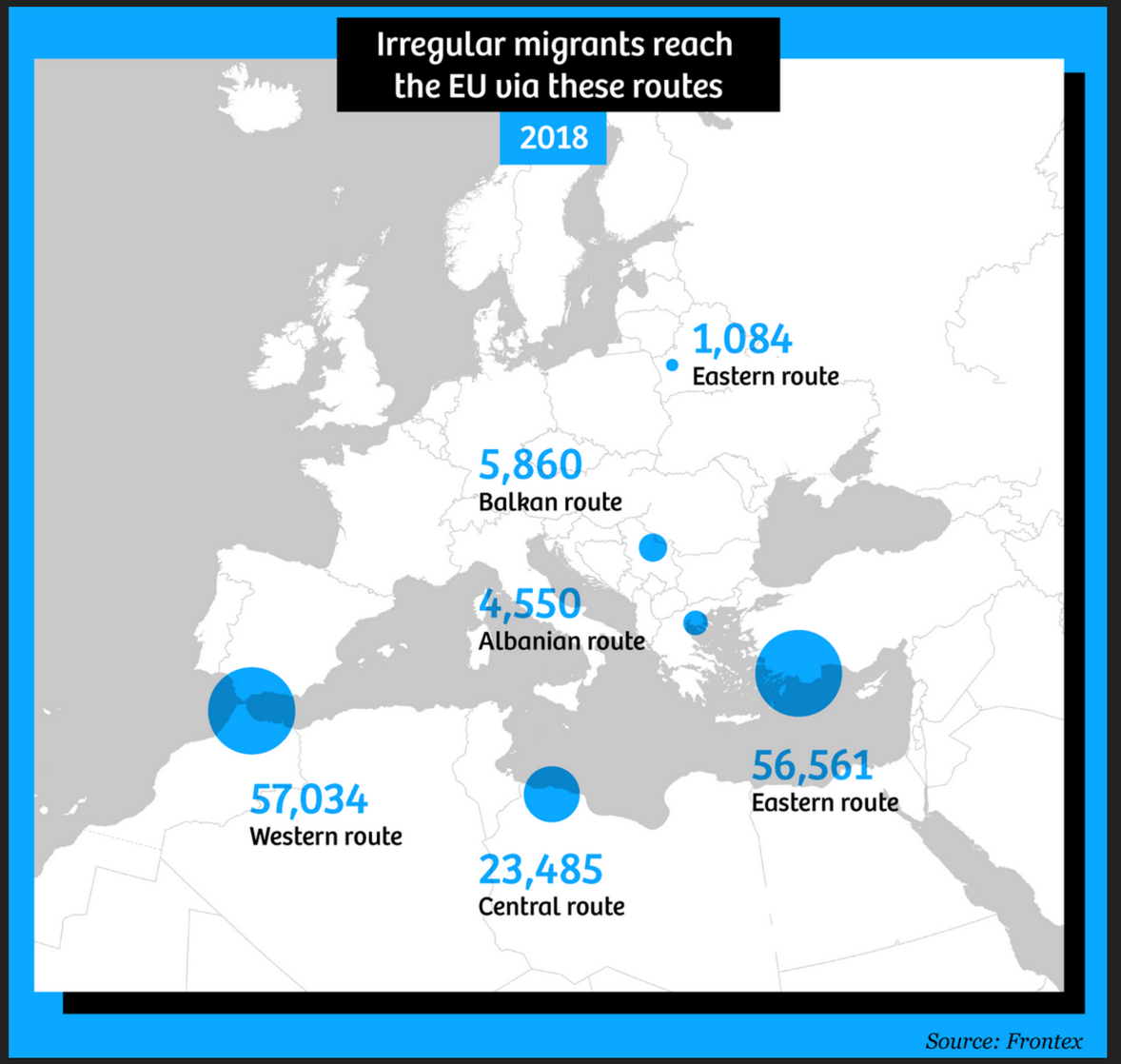
"How maps in the media make us more negative about migrants " by Maite Vermeulen, Leon de Korte & Henk van Houtum



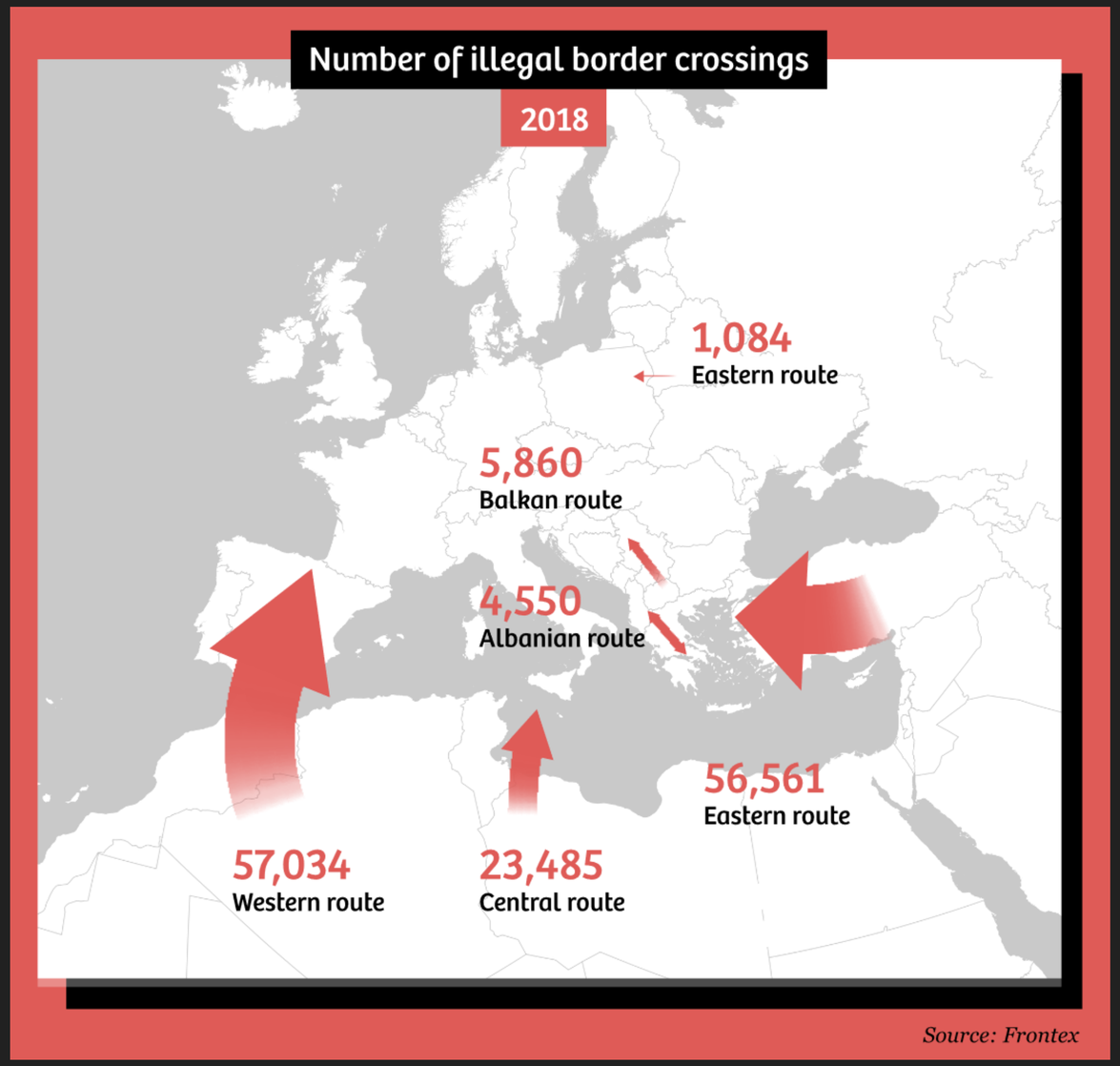


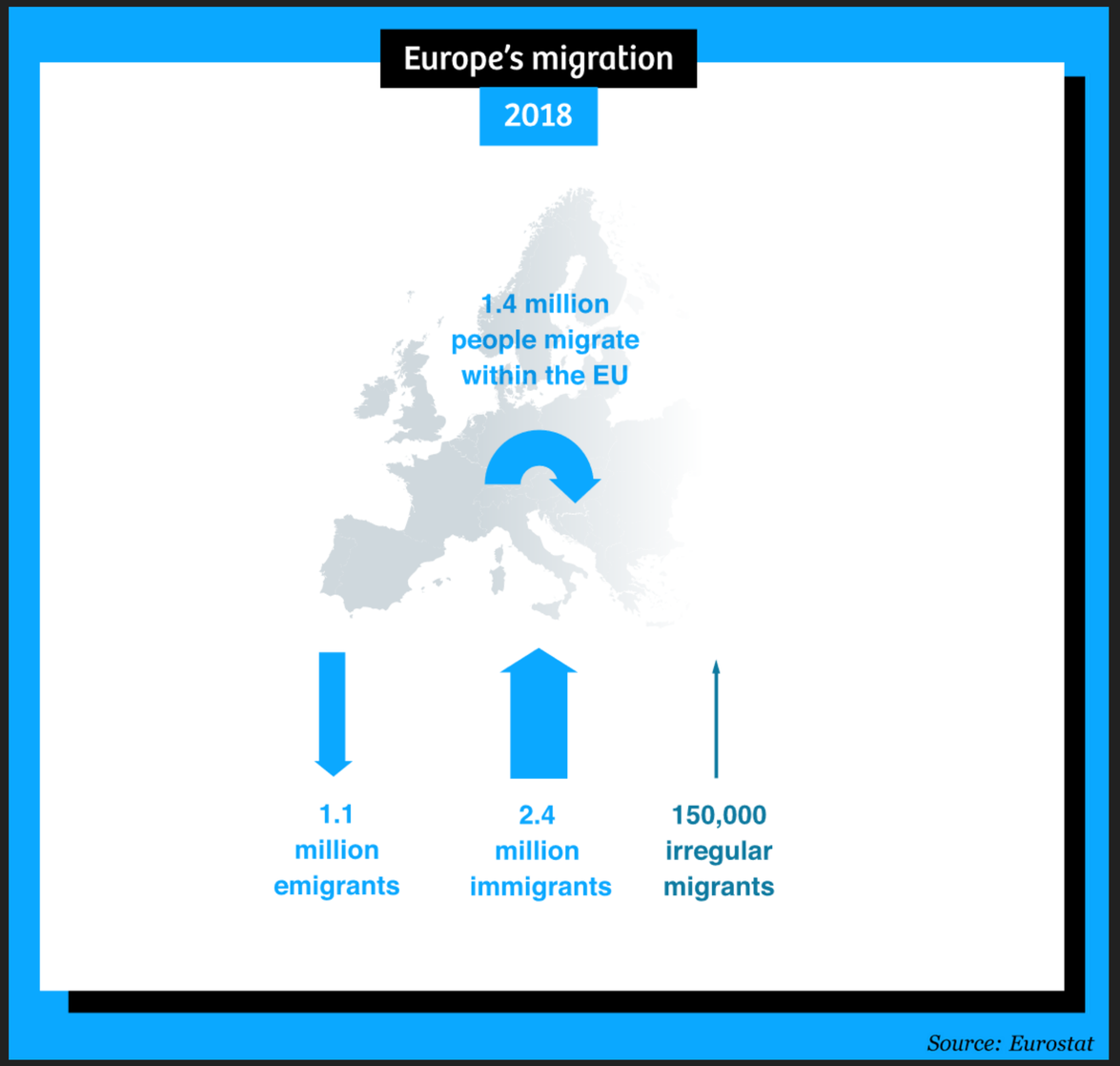
"How maps in the media make us more negative about migrants " by Maite Vermeulen, Leon de Korte & Henk van Houtum



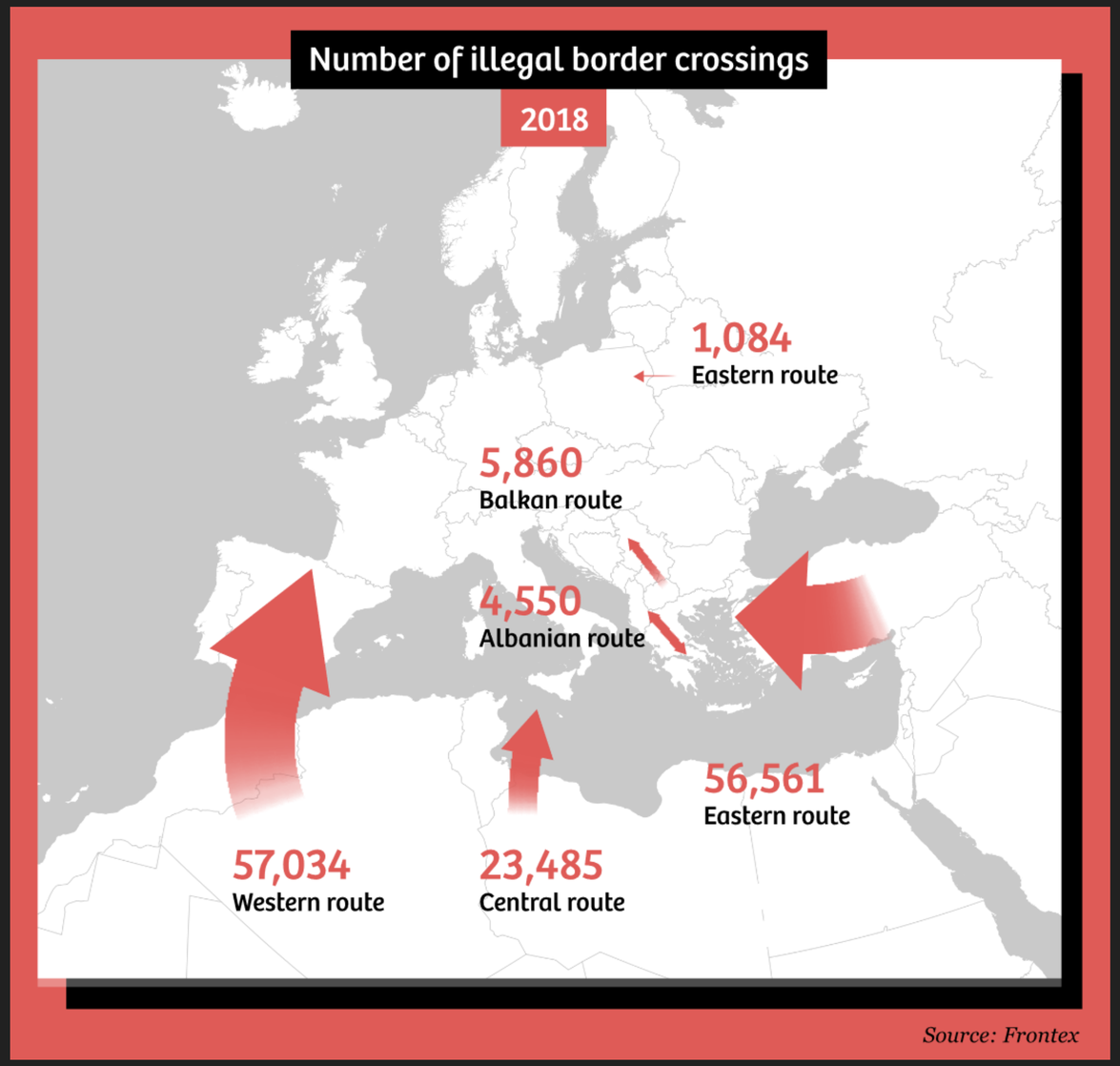


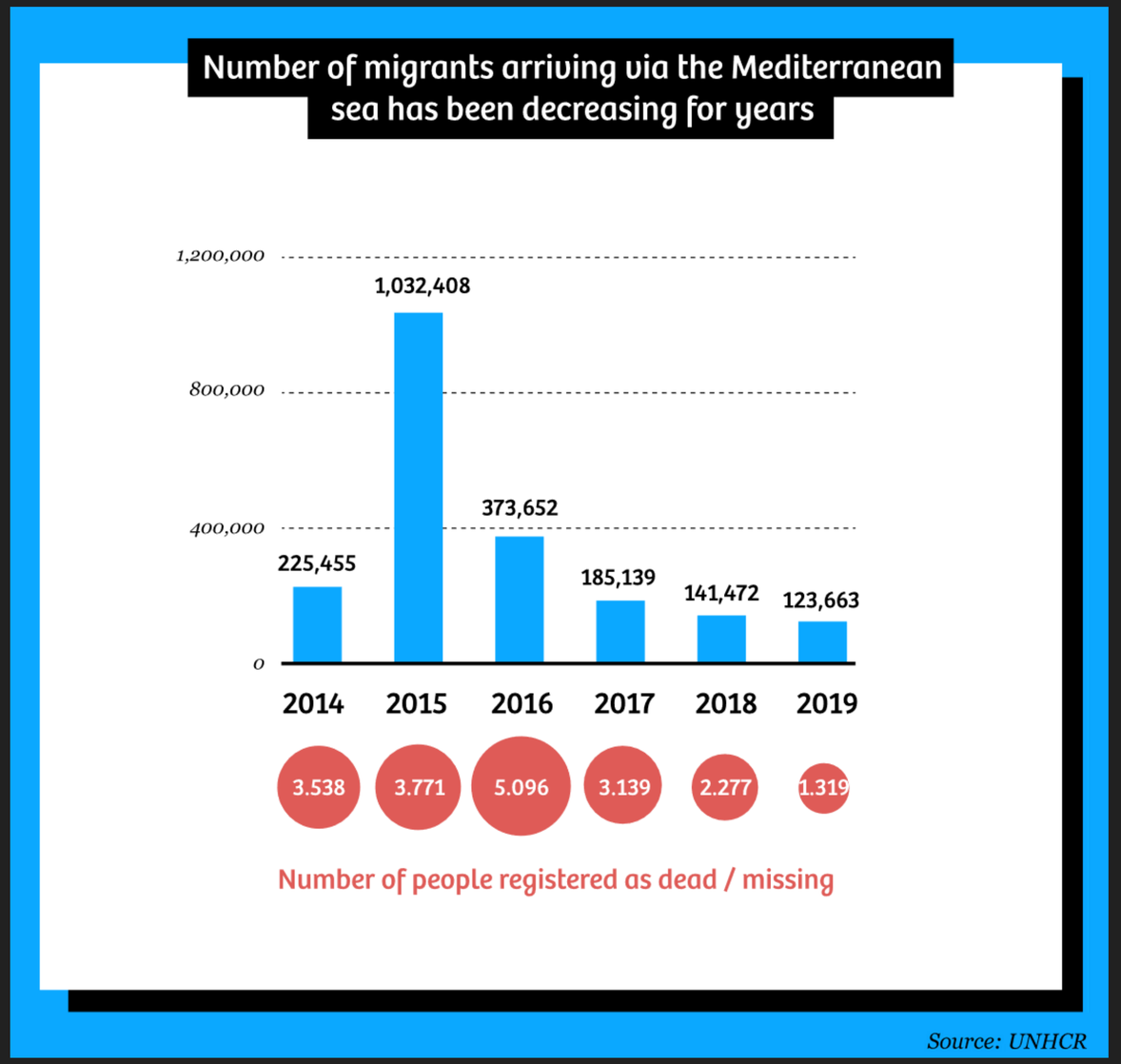
"How maps in the media make us more negative about migrants " by Maite Vermeulen, Leon de Korte & Henk van Houtum





"How maps in the media make us more negative about migrants " by Maite Vermeulen, Leon de Korte & Henk van Houtum





"How maps in the media make us more negative about migrants " by Maite Vermeulen, Leon de Korte & Henk van Houtum

Typology of Information Graphics

by Juuso Koponen & Jonatan Hildén, "Data Visualization Handbook" (2020), p. 25

Is the information conceptual or measurable?

☞ **Type of information**: depict conceptual information <> convert information into visual forms

Typology of Information Graphics

by Juuso Koponen & Jonatan Hildén, "Data Visualization Handbook" (2020), p. 25

Is the information conceptual or measurable?

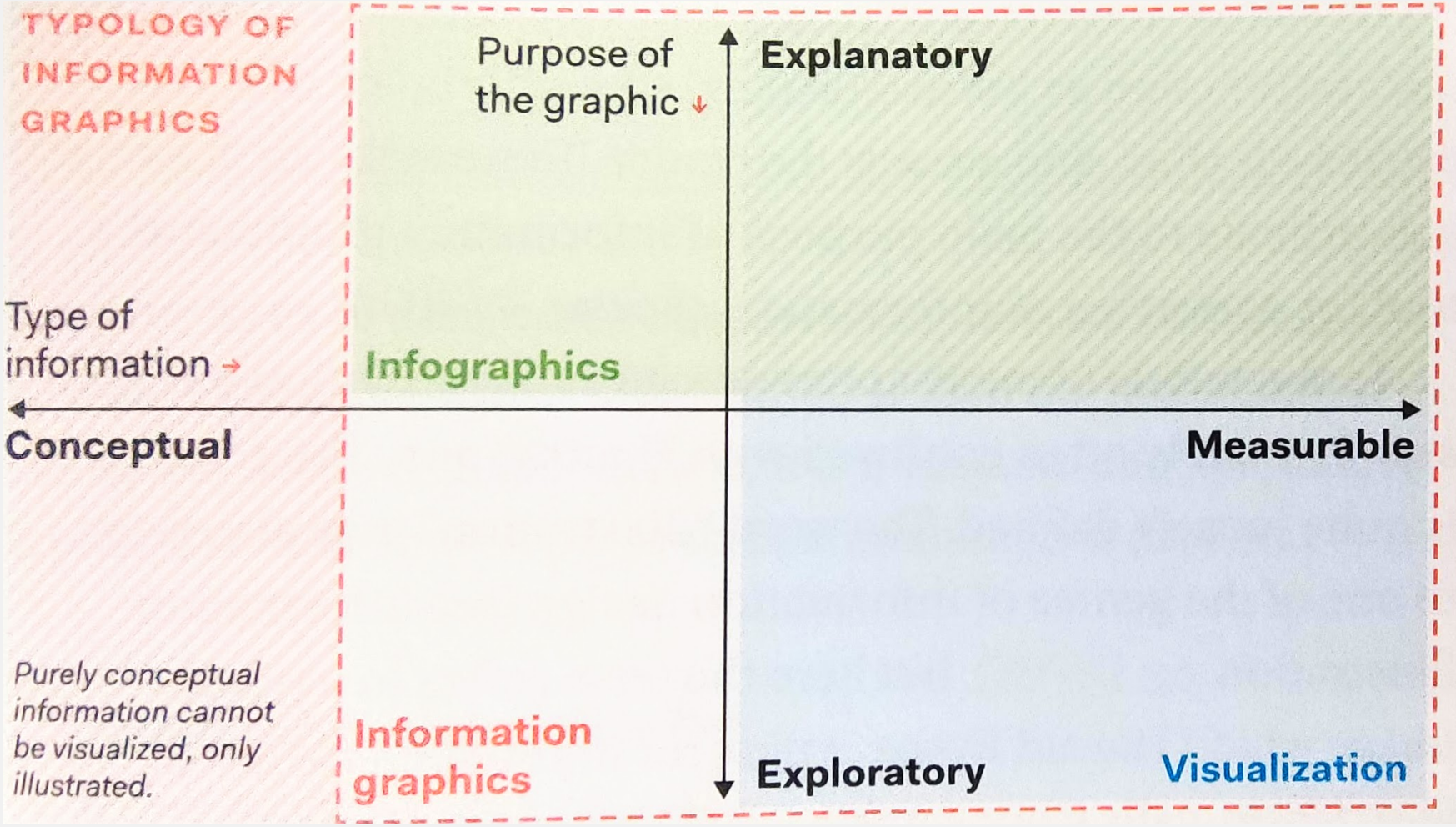
☞ **Type of information**: depict conceptual information <> convert information into visual forms

Is the purpose to explore or to explain the information?

☞ **Purpose of the graphic**: facilite discovery <> communicate information

Typology of Information Graphics

by Juuso Koponen & Jonatan Hildén, "Data Visualization Handbook" (2020), p. 25



“Visualizations can be designed and experienced
in various ways, by people of various backgrounds,
and in various circumstances. That's why
**reflecting on the purpose of a visualization is paramount
before we design it—or before we critique it**.”

Alberto Cairo

Excerpt from the foreword to “Data Sketches” by Nadieh Bremer & Shirley Wu (CRC Press 2021)

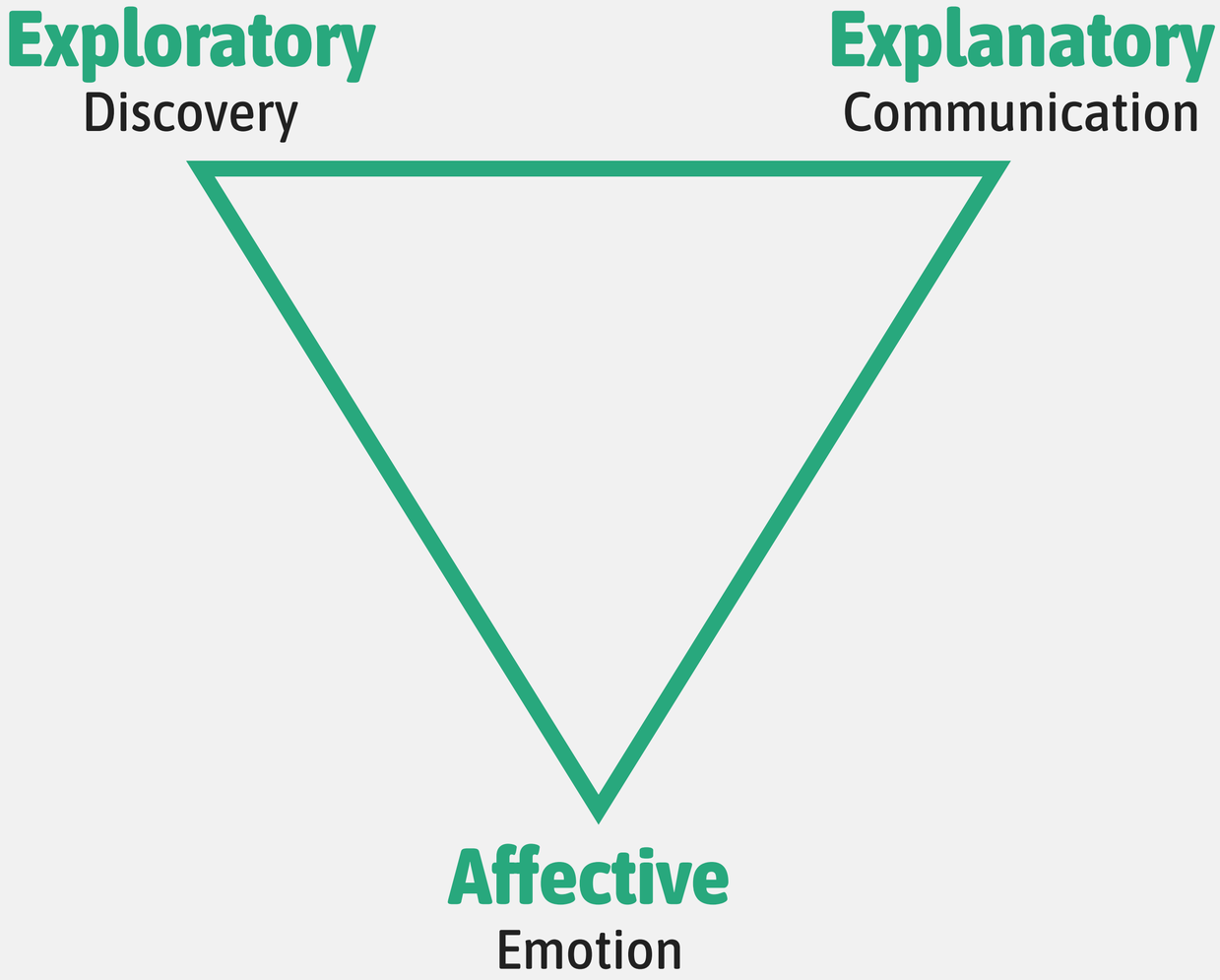
“A common truism about information visualization is that it is primarily about ‘showing the data’. […]

While this might be true for scientific (or financial, or many other) application fields, there are many good uses of visualization that go beyond a precise, “neutral” display of data.“

Moritz Stefaner

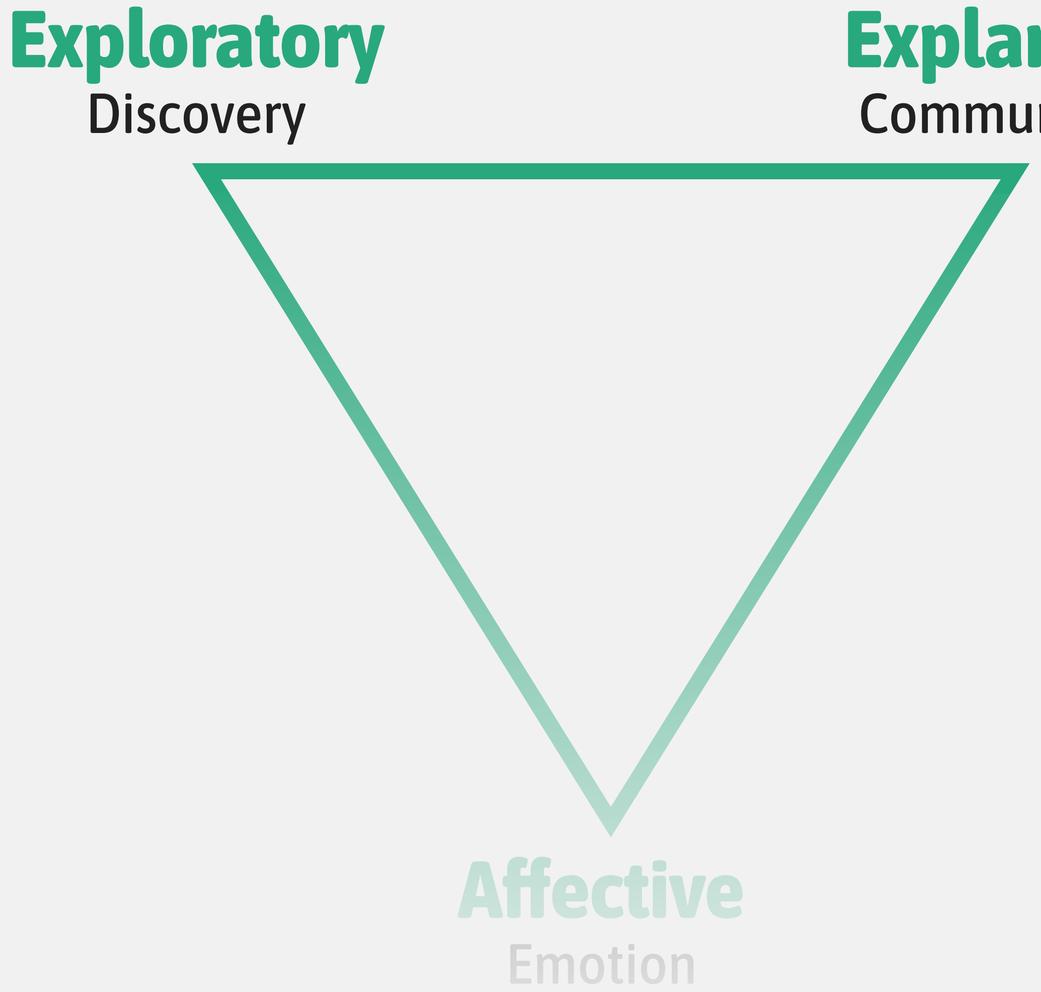
The Vertices of Visualization

by Alberto Cairo, personal communication



The Vertices of Visualization

by Alberto Cairo, personal communication

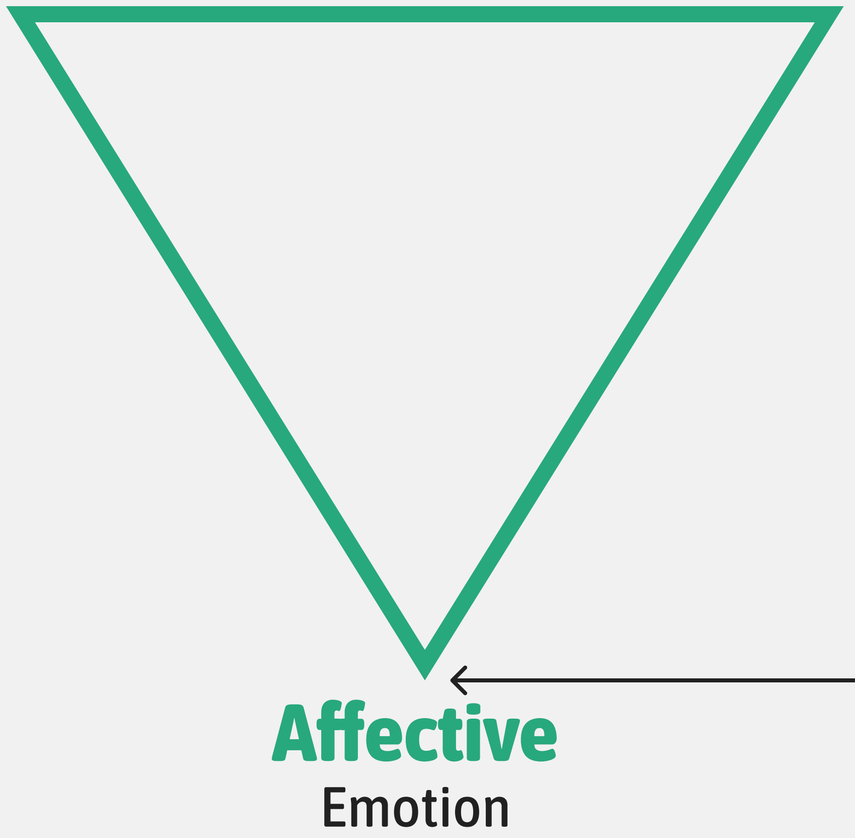


Priority:
efficiency + effectiveness

Goal:
functional

The Vertices of Visualization

by Alberto Cairo, personal communication



Priority:
efficiency + effectiveness

Goal:
functional

Priority:
creativity + novelity

Goal:
emotional

PLOS | BIOLOGY

PERSPECTIVE

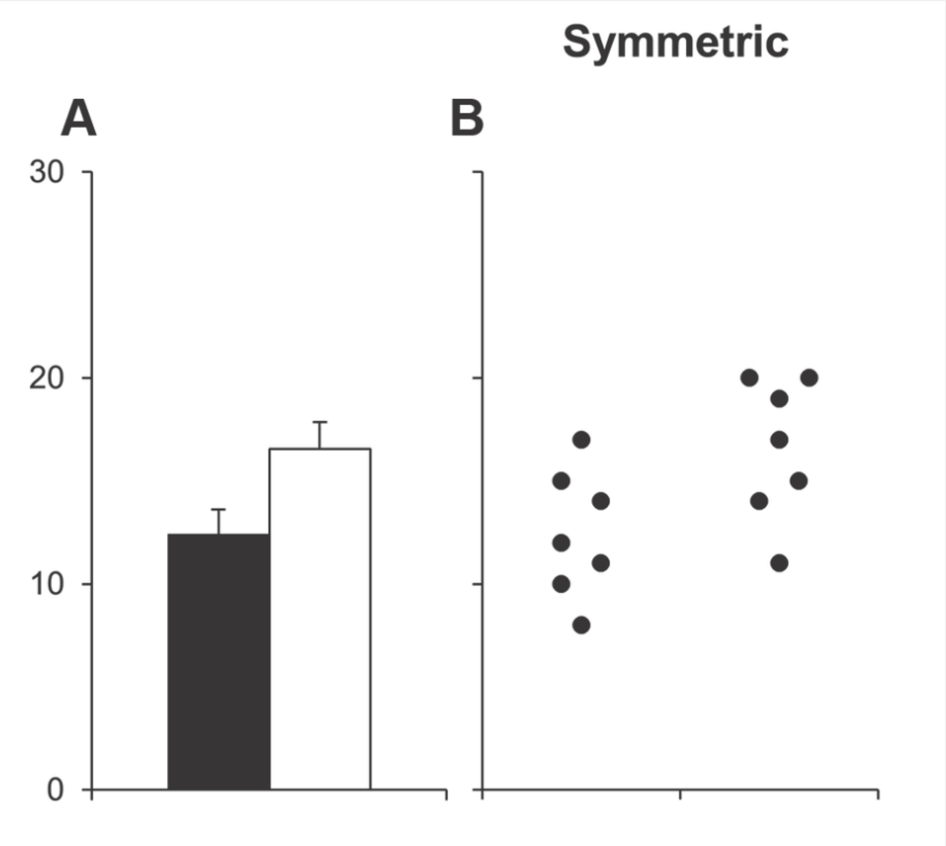
Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm

Tracey L. Weissgerber[1]*, Natasa M. Milic[1,2], Stacey J. Winham[3], Vesna D. Garovic[1]

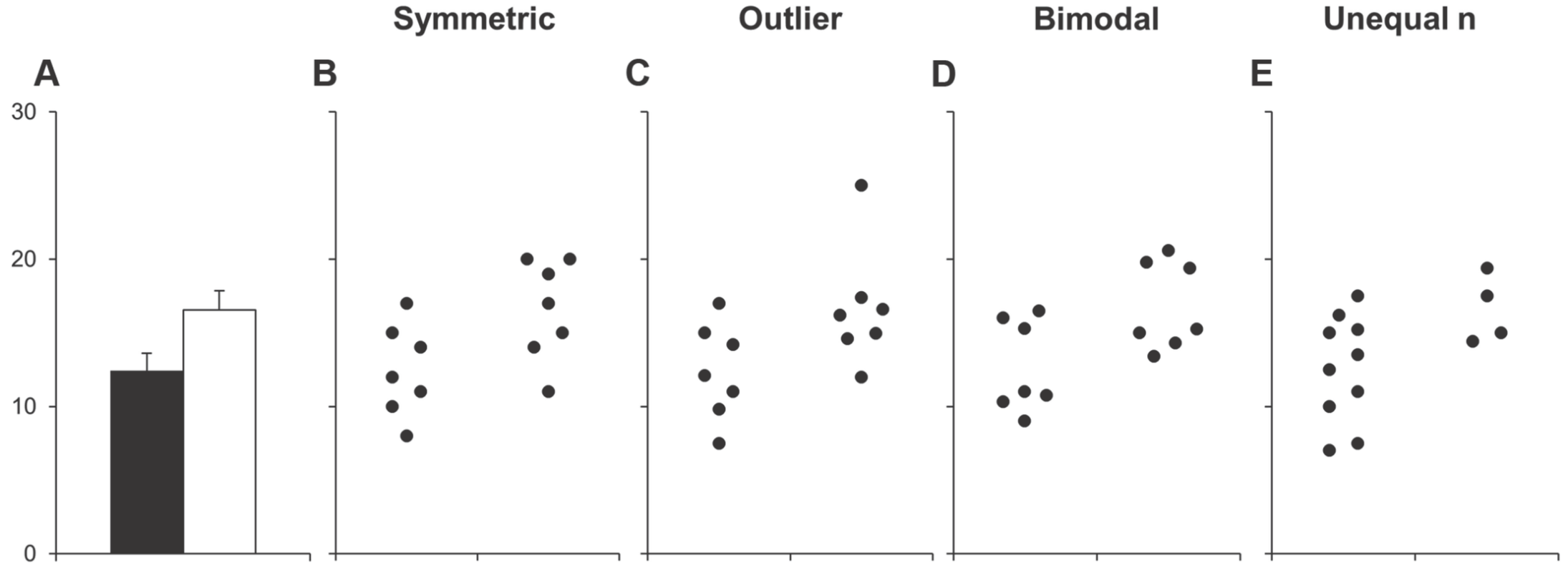
1 Division of Nephrology & Hypertension, Mayo Clinic, Rochester, Minnesota, United States of America, **2** Department of Biostatistics, Medical Faculty, University of Belgrade, Belgrade, Serbia, **3** Division of Biomedical Statistic and Informatics, Mayo Clinic, Rochester, Minnesota, United States of America

* weissgerber.tracey@mayo.edu

Weissgerber et al. (2015) PLoS Biology

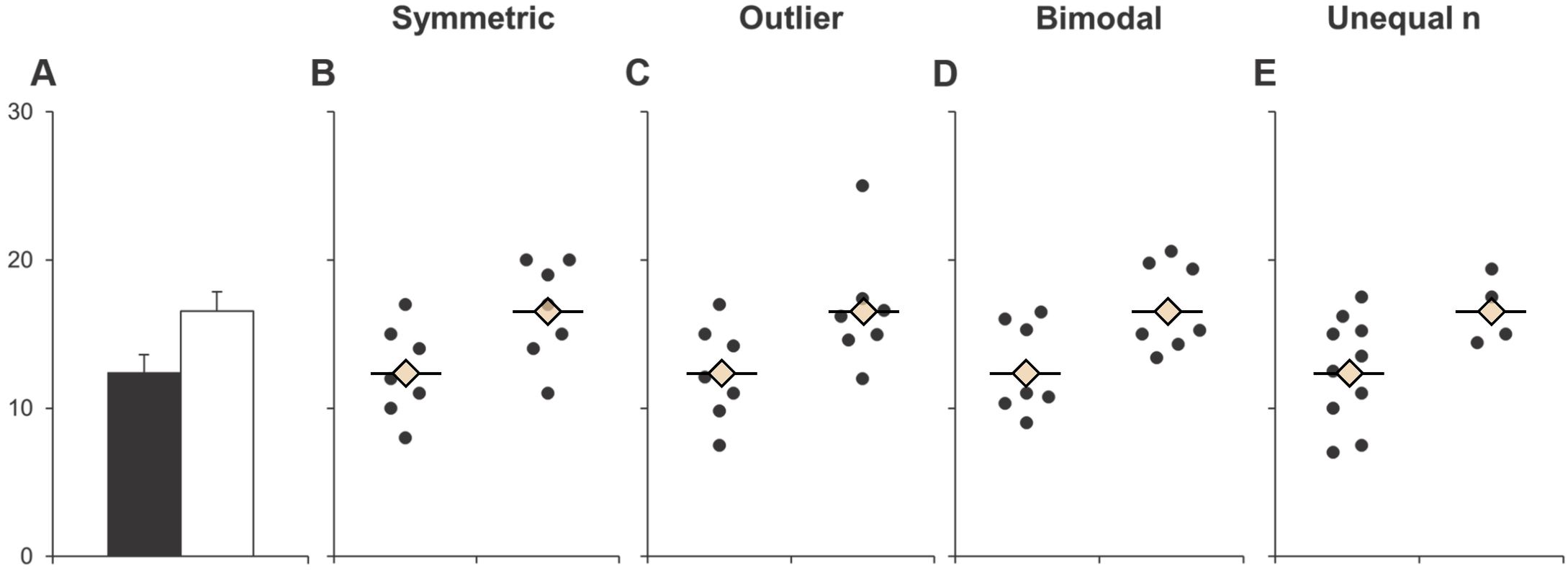


Weissgerber et al. (2015) PLoS Biology



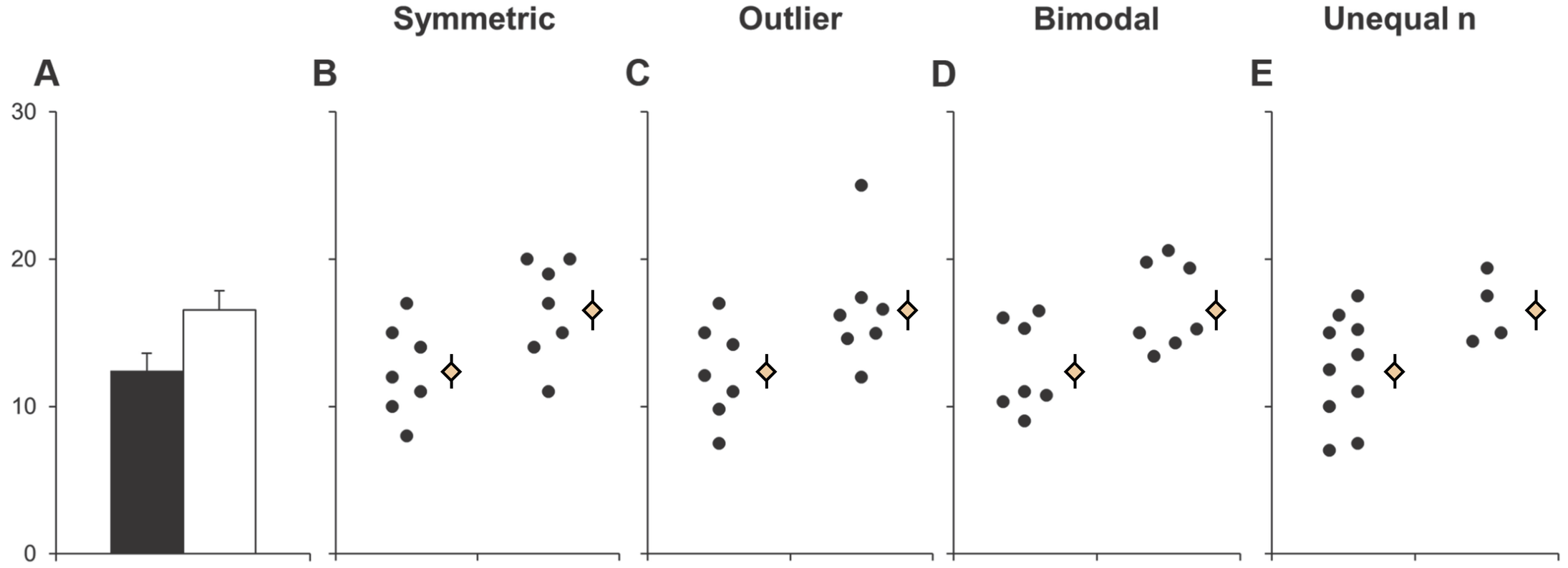
Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Weissgerber et al. (2015) PLoS Biology



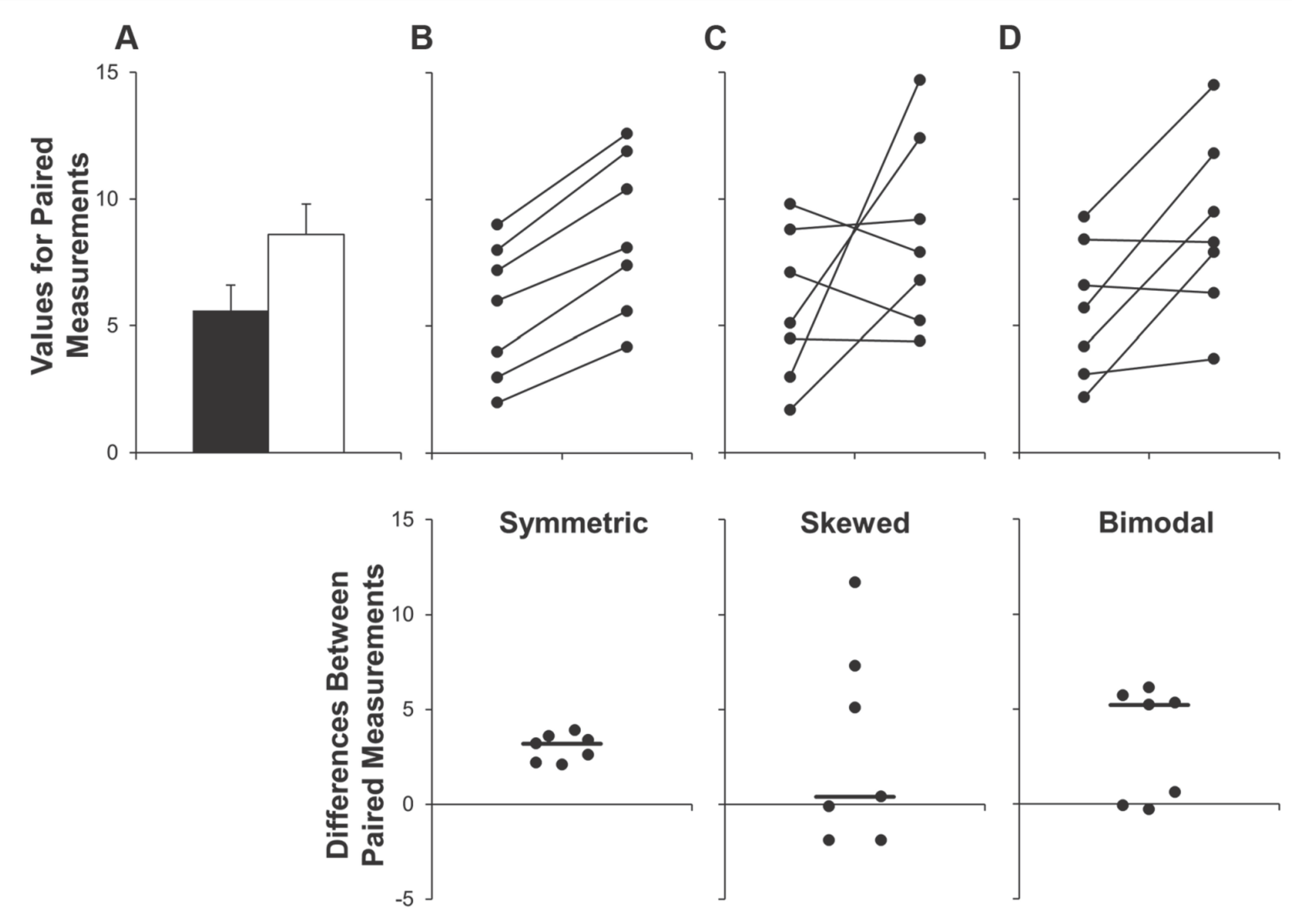
Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Modified from Weissgerber et al. (2015) PLoS Biology

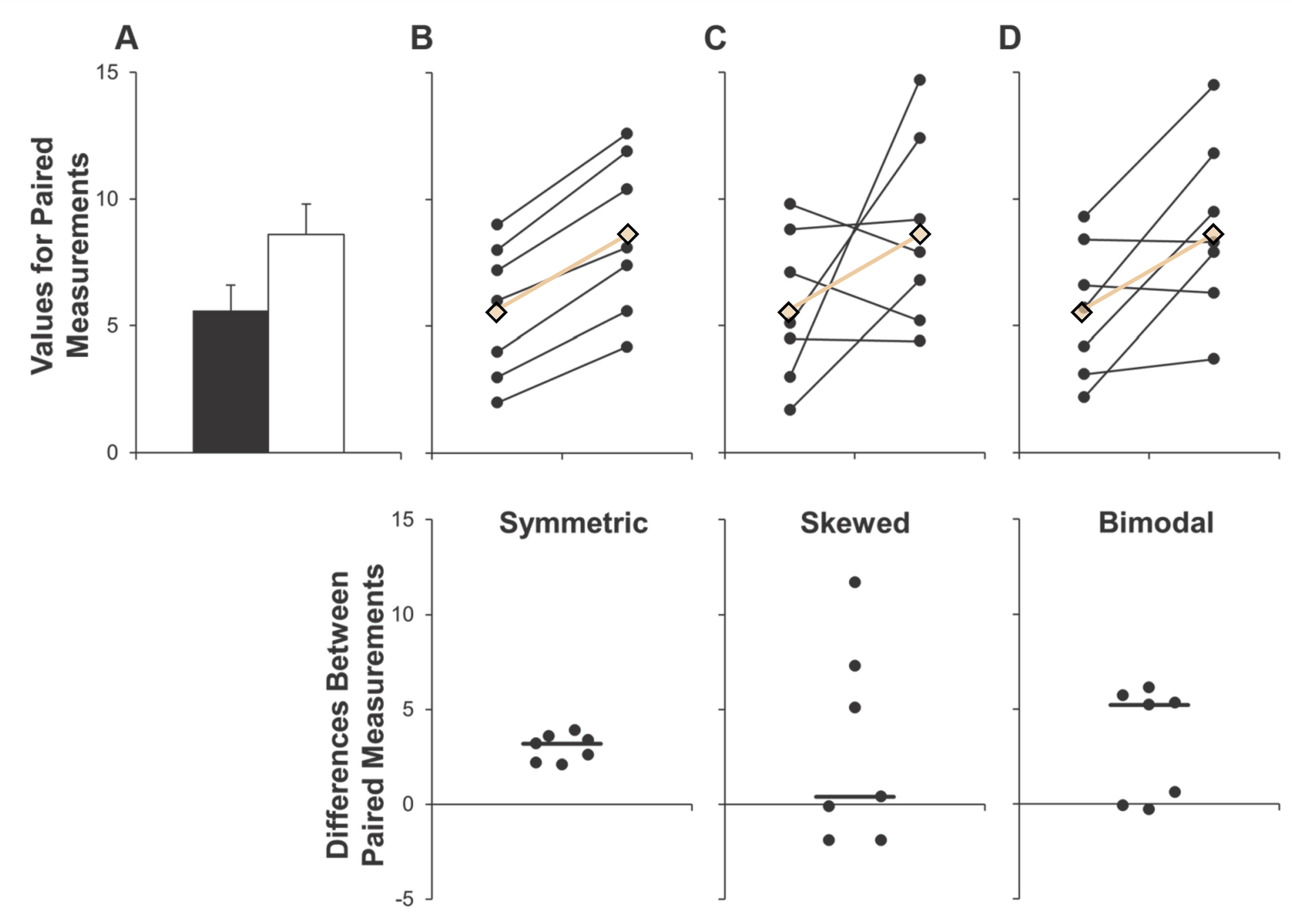


Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

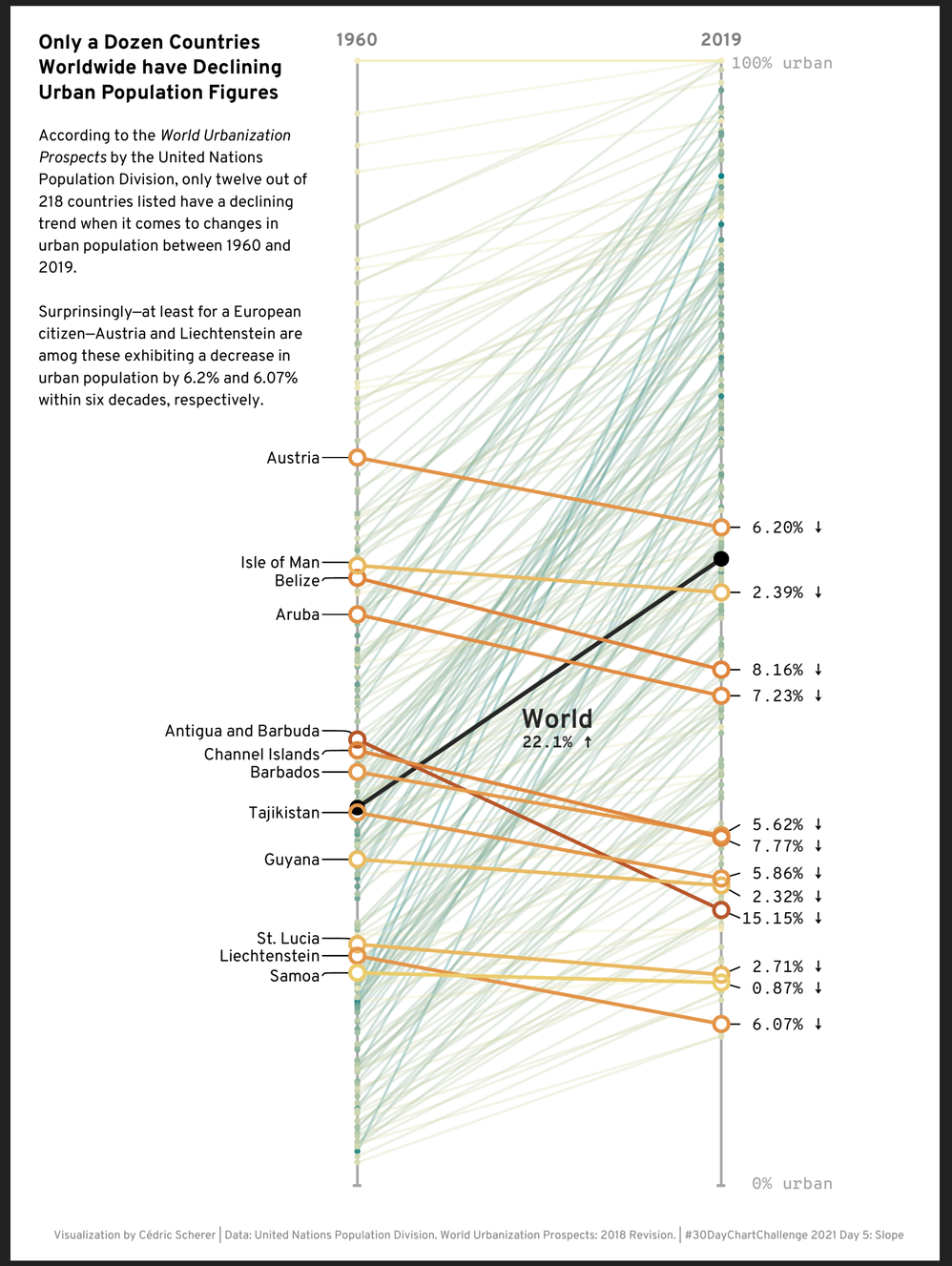
Modified from Weissgerber et al. (2015) PLoS Biology

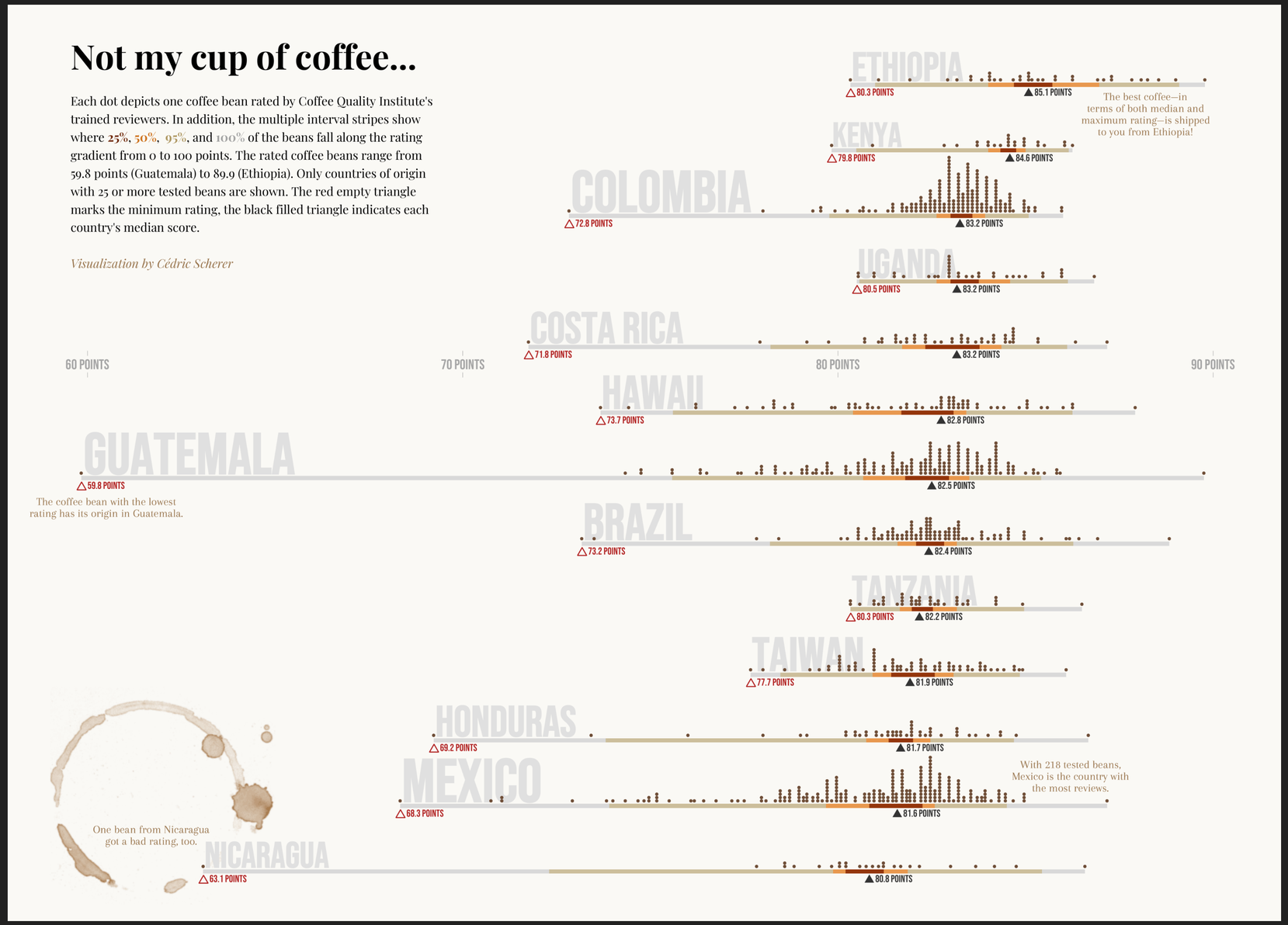


Weissgerber et al. (2015) PLoS Biology



Weissgerber et al. (2015) PLoS Biology


Only a Dozen Countries Worldwide have Declining Urban Population Figures
According to the *World Urbanization Prospects* by the United Nations Population Division, only twelve out of 218 countries listed have a declining trend when it comes to changes in urban population between 1960 and 2019.
Surprinsingly—at least for a European citizen—Austria and Liechtenstein are amog these exhibiting a decrease in urban population by 6.2% and 6.07% within six decades, respectively.
1960
2019
100% urban
0% urban
Austria
Isle of Man
Belize
Aruba
Antigua and Barbuda
Channel Islands
Barbados
Tajikistan
Guyana
St. Lucia
Liechtenstein
Samoa
World
22.1% ↑
6.20% ↓
2.39% ↓
8.16% ↓
7.23% ↓
5.62% ↓
7.77% ↓
5.86% ↓
2.32% ↓
15.15% ↓
2.71% ↓
0.87% ↓
6.07% ↓
Visualization by Cédric Scherer | Data: United Nations Population Division. World Urbanization Prospects: 2018 Revision. | #30DayChartChallenge 2021 Day 5: Slope



“Not my cup of coffee”, *#TidyTuesday Contribution*

Cédric Scherer // rstudio::conf // July 2022

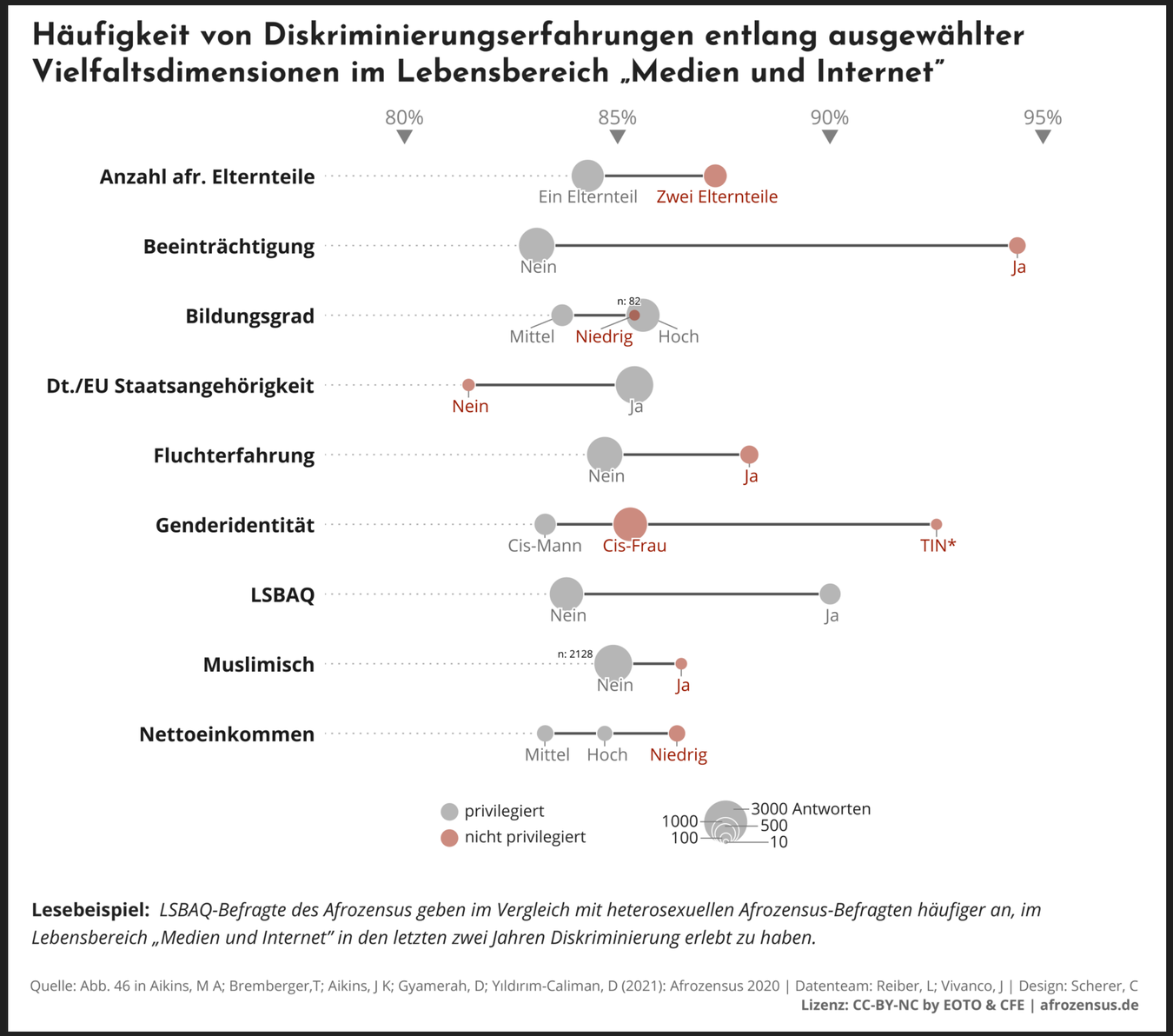


Abb. 46 "Afrozensus 2020" *by Citizens For Europe & EOTO e.V.*

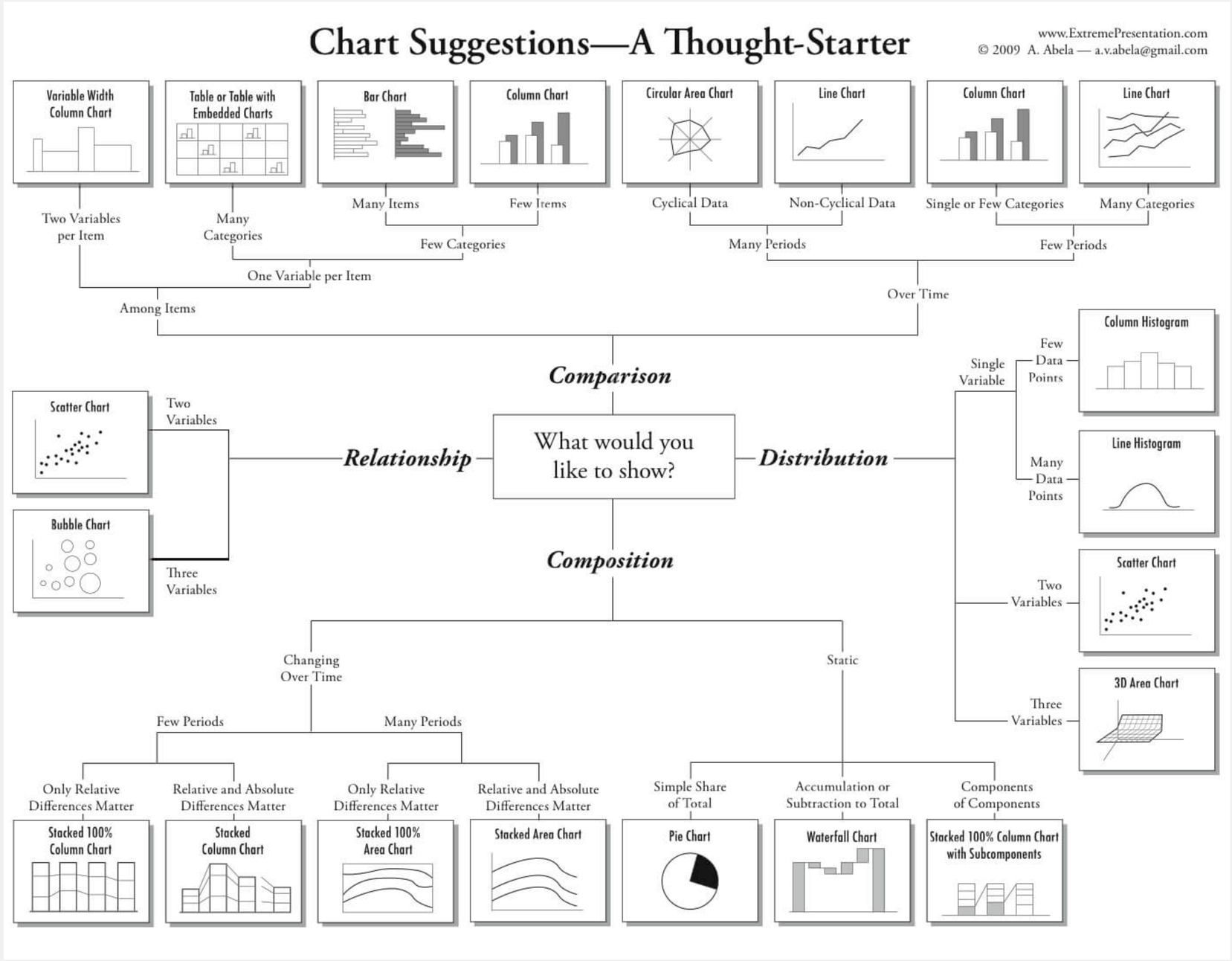
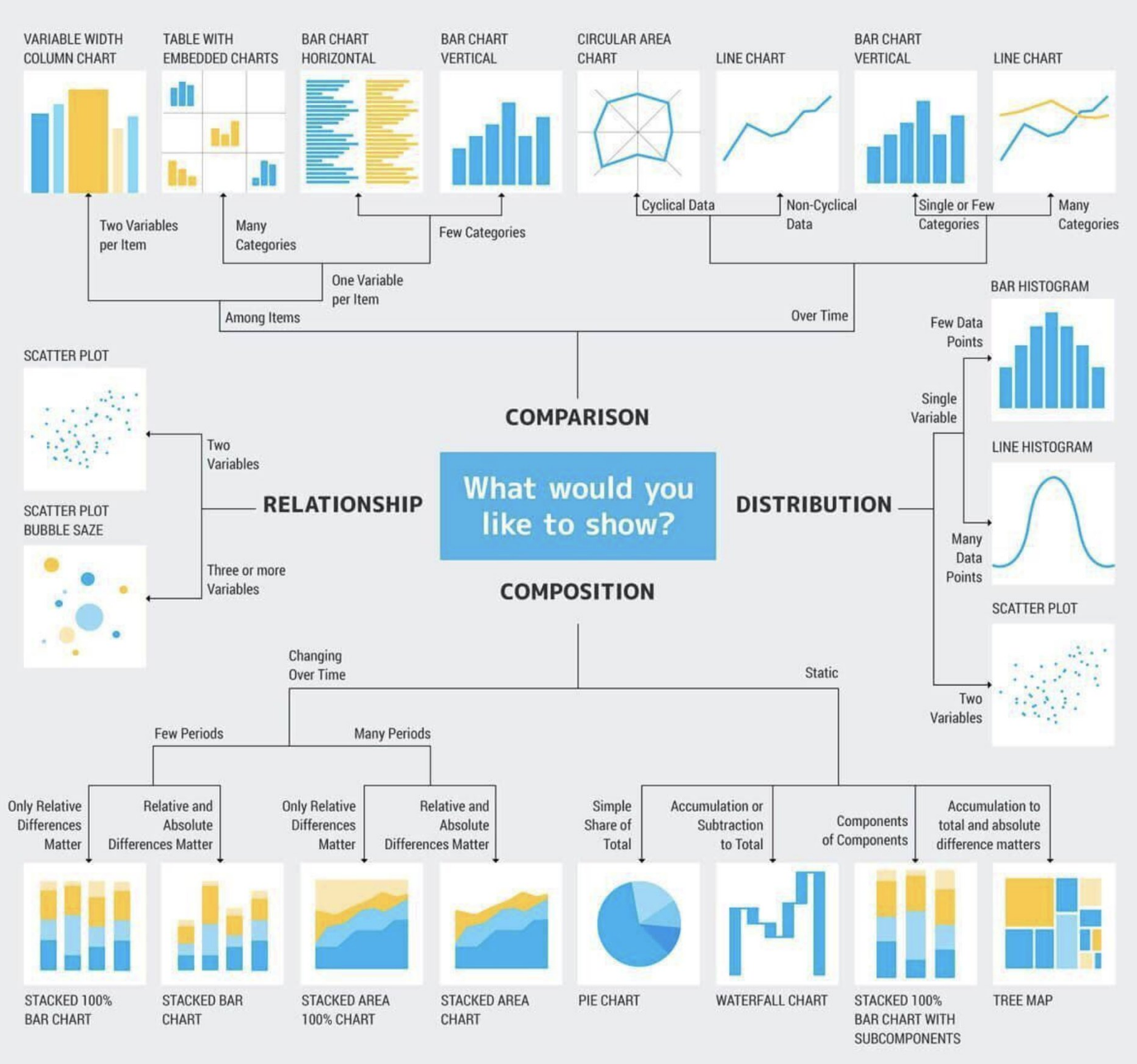
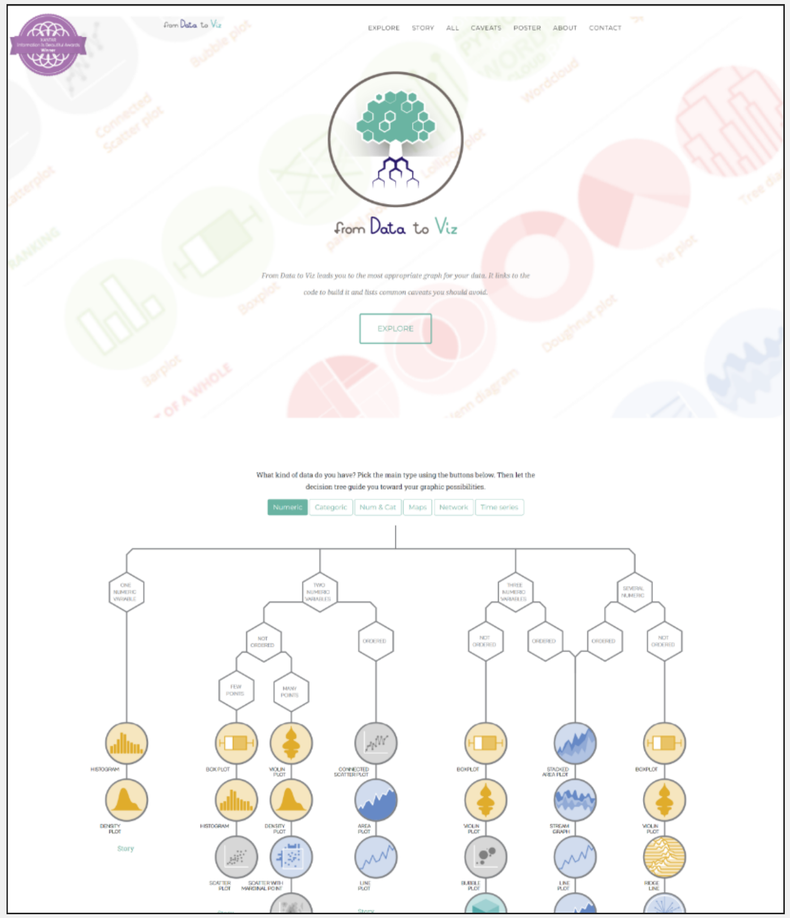


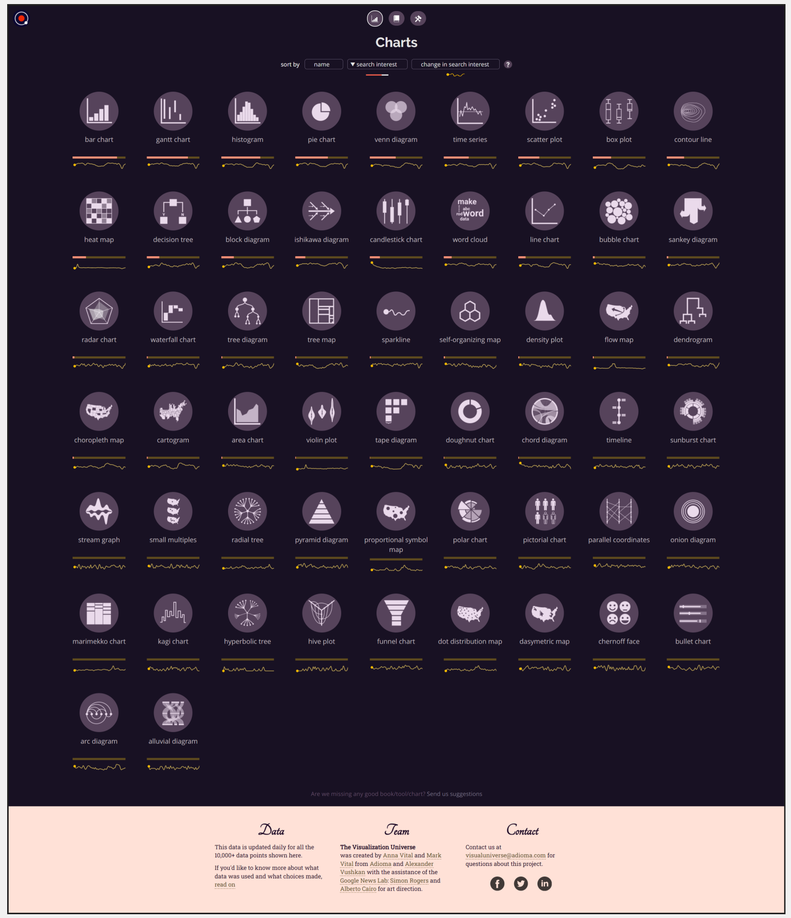
Chart Suggestions—A Thought-Starter
www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com
Variable Width Column Chart
Table or Table with Embedded Charts
Bar Chart
Column Chart
Circular Area Chart
Line Chart
Column Chart
Line Chart
Two Variables per Item
Many Categories
Many Items
Few Items
Cyclical Data
Non-Cyclical Data
Single or Few Categories
Many Categories
Few Categories
Many Periods
Few Periods
One Variable per Item
Over Time
Among Items
Comparison
Scatter Chart
Two Variables
Relationship
What would you like to show?
Distribution
Bubble Chart
Three Variables
Composition
Single Variable
Few Data Points
Column Histogram
Many Data Points
Line Histogram
Two Variables
Scatter Chart
Three Variables
3D Area Chart
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Stacked 100% Column Chart
Stacked Column Chart
Stacked 100% Area Chart
Stacked Area Chart
Pie Chart
Waterfall Chart
Stacked 100% Column Chart with Subcomponents


What would you like to show?
COMPARISON
RELATIONSHIP
DISTRIBUTION
COMPOSITION
VARIABLE WIDTH COLUMN CHART
TABLE WITH EMBEDDED CHARTS
BAR CHART HORIZONTAL
BAR CHART VERTICAL
CIRCULAR AREA CHART
LINE CHART
BAR CHART VERTICAL
LINE CHART
Two Variables per Item
Many Categories
Few Categories
One Variable per Item
Among Items
Cyclical Data
Non-Cyclical Data
Single or Few Categories
Many Categories
Over Time
SCATTER PLOT
Two Variables
SCATTER PLOT BUBBLE SAZE
Three or more Variables
BAR HISTOGRAM
Few Data Points
Single Variable
LINE HISTOGRAM
Many Data Points
SCATTER PLOT
Two Variables
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Accumulation to total and absolute difference matters
STACKED 100% BAR CHART
STACKED BAR CHART
STACKED AREA 100% CHART
STACKED AREA CHART
PIE CHART
WATERFALL CHART
STACKED 100% BAR CHART WITH SUBCOMPONENTS
TREE MAP

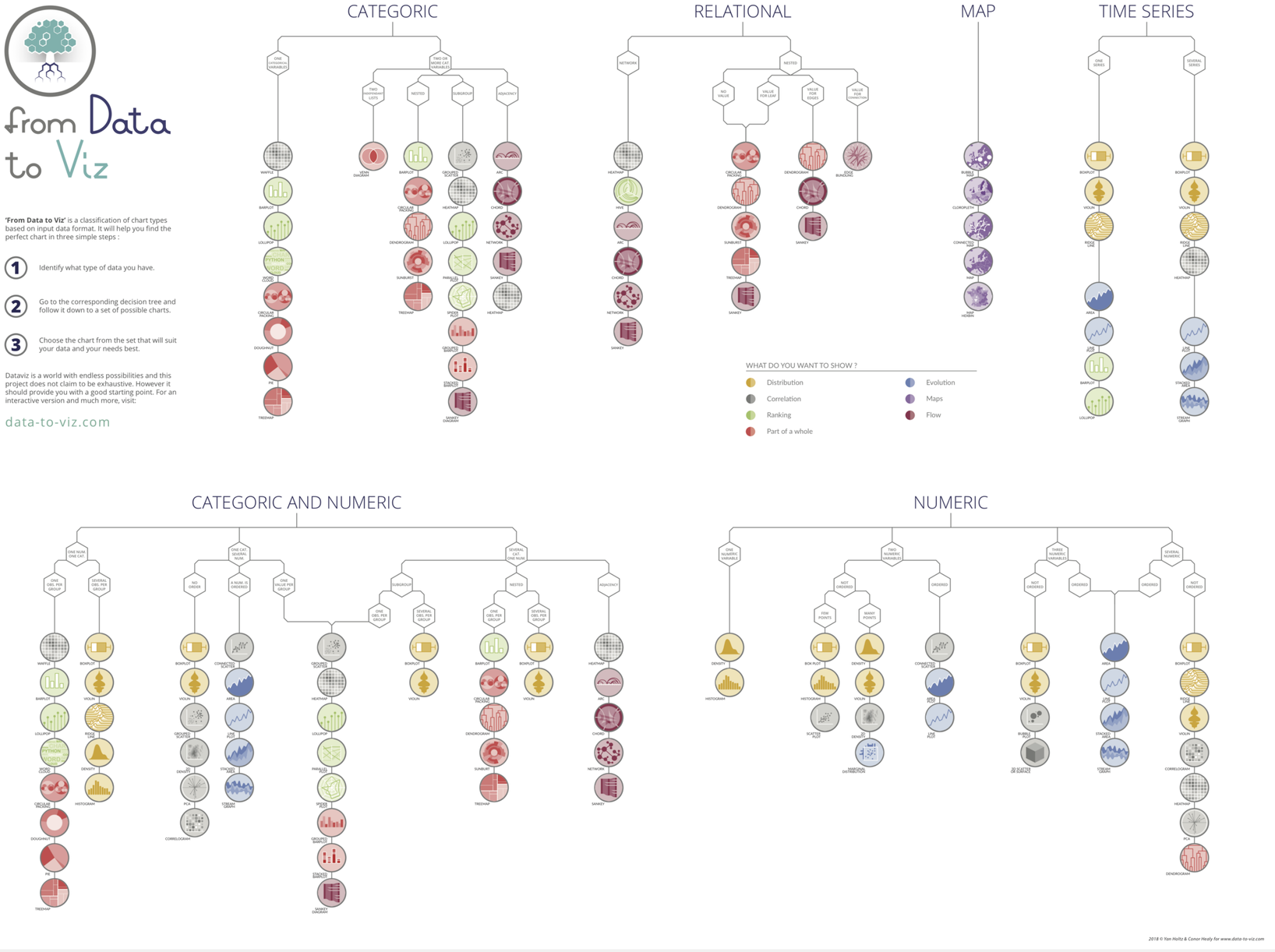


data-to-viz.com

datavizproject.com

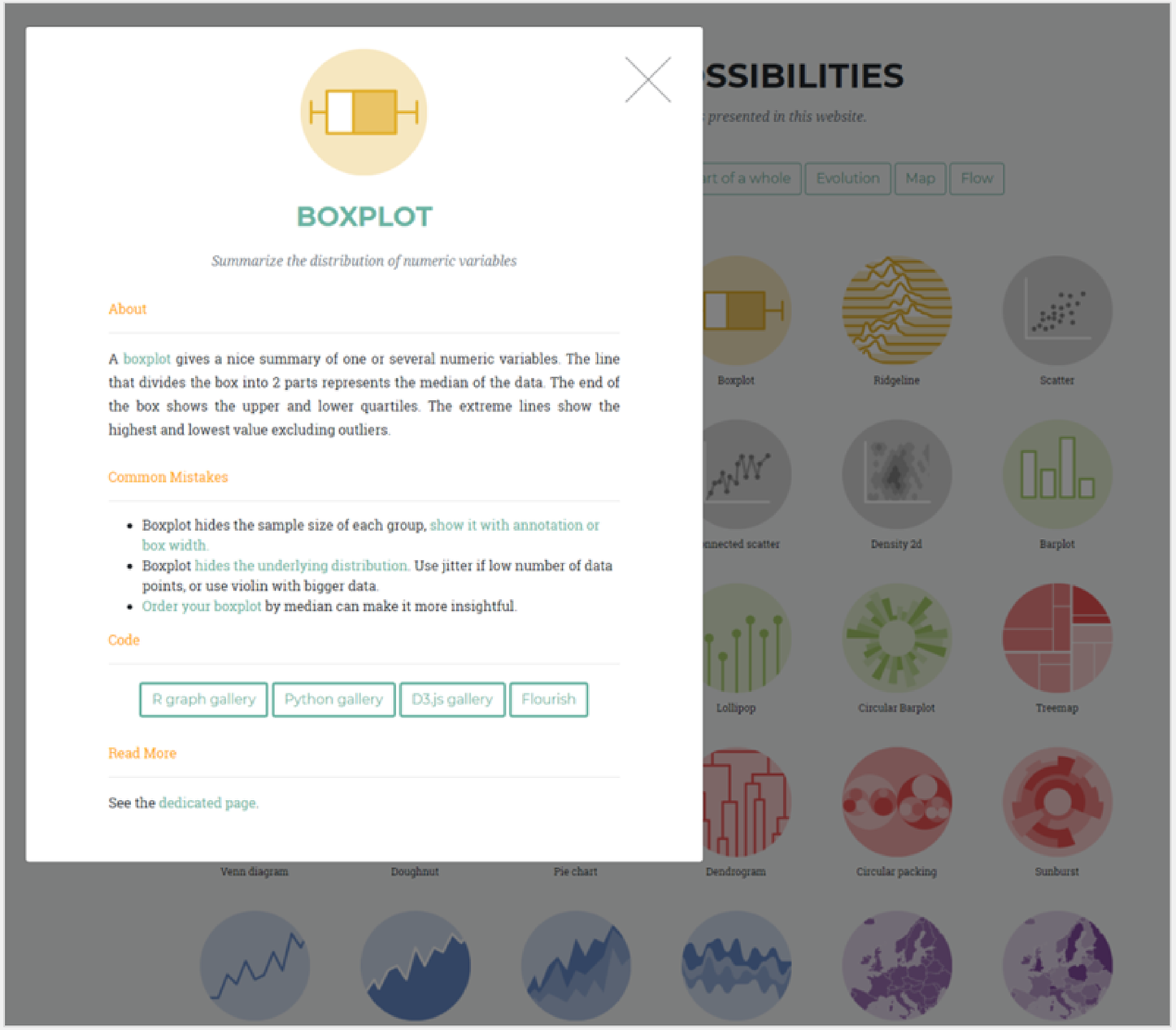


visualizationuniverse.com



data-to-viz.com

Cédric Scherer // rstudio::conf // July 2022

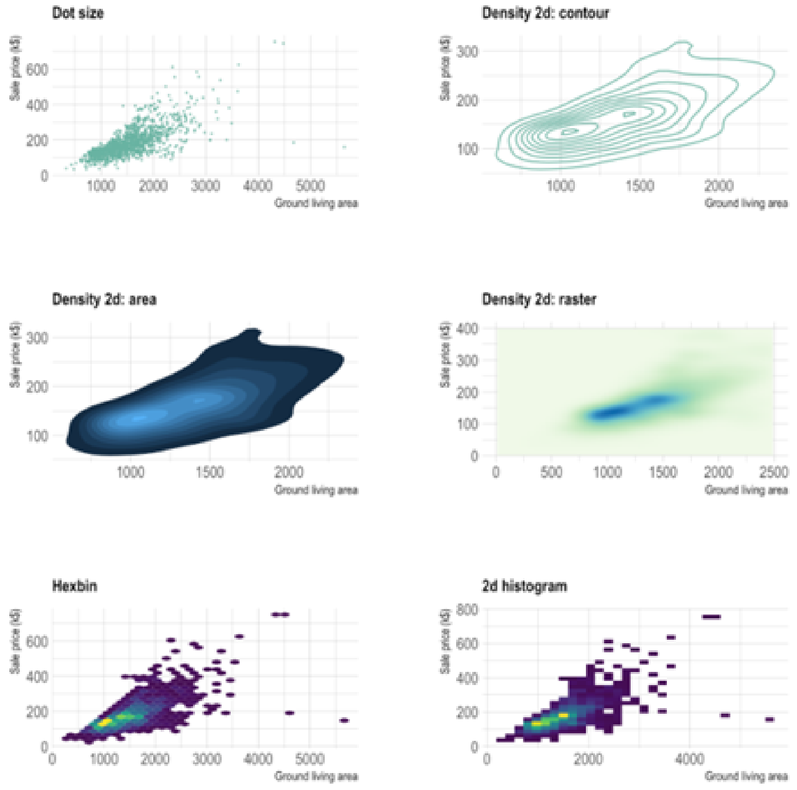


data-to-viz.com

Overplotting

The most common pitfall with scatterplot is overplotting: when the sample size gets big, dots are plotted on top of each other what makes the chart unreadable. There are several work around to avoid this issue as describe in this specific post. Here is a summary of the different offered techniques:

CODE

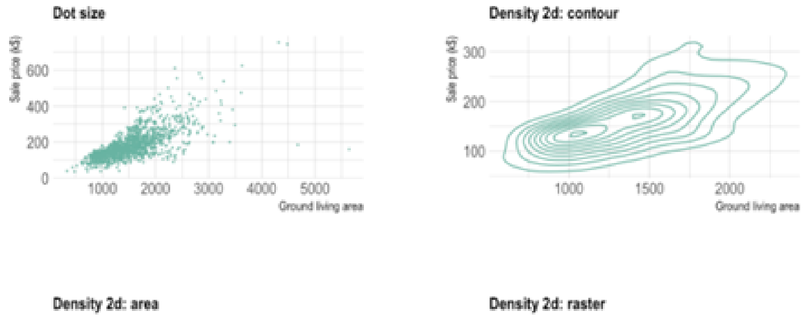


Going further

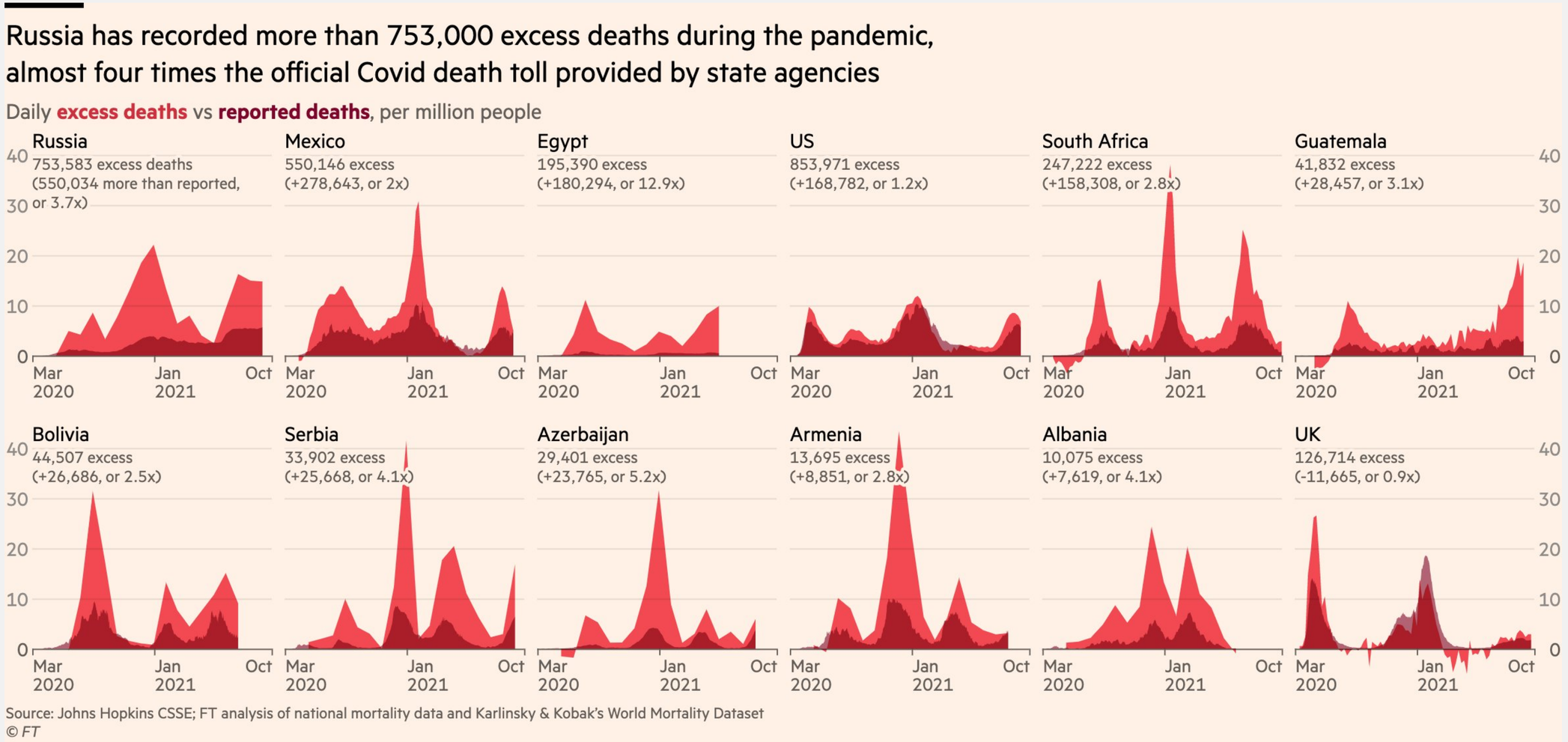
You can learn more about each type of graphic presented in this story in the dedicated

HIDE

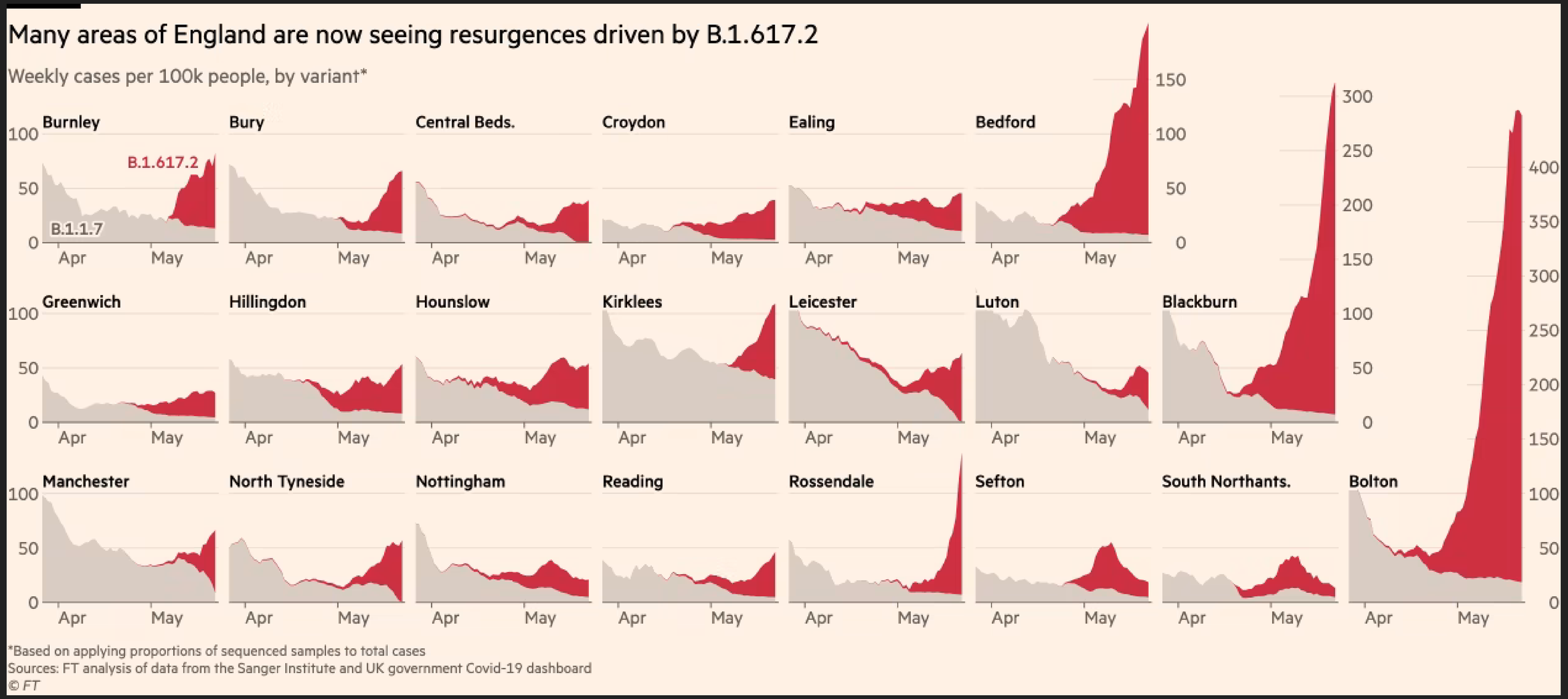
```
# code for all graphics:
p <- data %>%
  ggplot( aes(x=GrLivArea, y=SalePrice/1000)) +
    theme_ipsum() +
    theme(
      plot.title = element_text(size=12)
    ) +
    ylab('Sale price (k$)') +
    xlab('Ground living area')

# Reduce dot size
p1 <- p + geom_point(color="#69b3a2", alpha=0.8, size=0.2) + ggtitle("Dot size")

# Use density estimate
p2 <- p + geom_density2d(color="#69b3a2") + ggtitle("Density 2d: contour")

# Use density estimate (area)
p3 <- p + stat_density_2d(aes(fill = ..level..), geom = "polygon") + ggtitle("Density 2d: area") + theme(legend.positio
n="none")

# With raster
p4 <- p +
  stat_density_2d(aes(fill = ..density..), geom = "raster", contour = FALSE) +
  scale_fill_distiller(palette=4, direction=1) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  theme(
    legend.position='none'
  ) +
  ggtitle("Density 2d: raster") +
  xlim(0,2500) +
  ylim(0,400)

# Hexbin
p5 <- p + geom_hex() +
    scale_fill_viridis() +
    theme(legend.position="none") +
    ggtitle("Hexbin")

# 2d histogram
p6 <- p + geom_bin2d( ) +
    scale_fill_viridis( ) +
    theme(legend.position="none")  +
    ggtitle("2d histogram")

p1 + p2 + p3 + p4 + p5 + p6 + plot_layout(ncol = 2)
```



The Power of Small Multiples



"Russia's excess mortality soars since start of Covid pandemic" by John Burn-Murdoch (Financial Times)



"UK virus cases hit 6-week high but vaccines diminish threat" by John Burn-Murdoch (Financial Times)

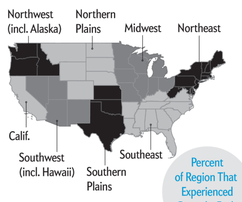
GRAPHIC SCIENCE

Text by Clara Moskowitz | Graphic by Cédric Scherer and Georgios Karamanis

Escalating Drought

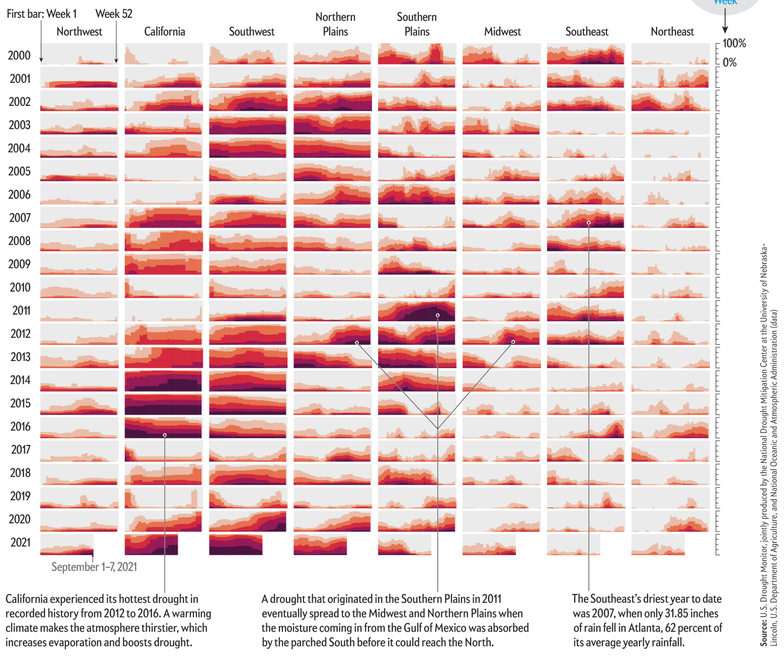
Climate change is intensifying periods of extreme dryness, particularly in the U.S. West

For more than 20 years the National Drought Mitigation Center (NDMC) has been monitoring dozens of indices of drought around the country, including satellite measurements of evaporation and color in vegetation, soil-moisture sensors, rainfall estimates, and river and streamflow levels. Although the agency's weekly assessments have identified periods of exceptional drought before, lately dryness has been ramping up. "The changing climate is definitely contributing to more natural disasters, drought being one of them," says Brian Fuchs, a climatologist who oversees the weekly report at the NDMC. "We're seeing more frequent and high-intensity episodes. This year some of these areas in the West have been in drought more than they have been without drought."



Drought Extent and Intensity by Region over Time

Category: Abnormally Dry | Moderate Drought | Severe Drought | Extreme Drought | Exceptional Drought

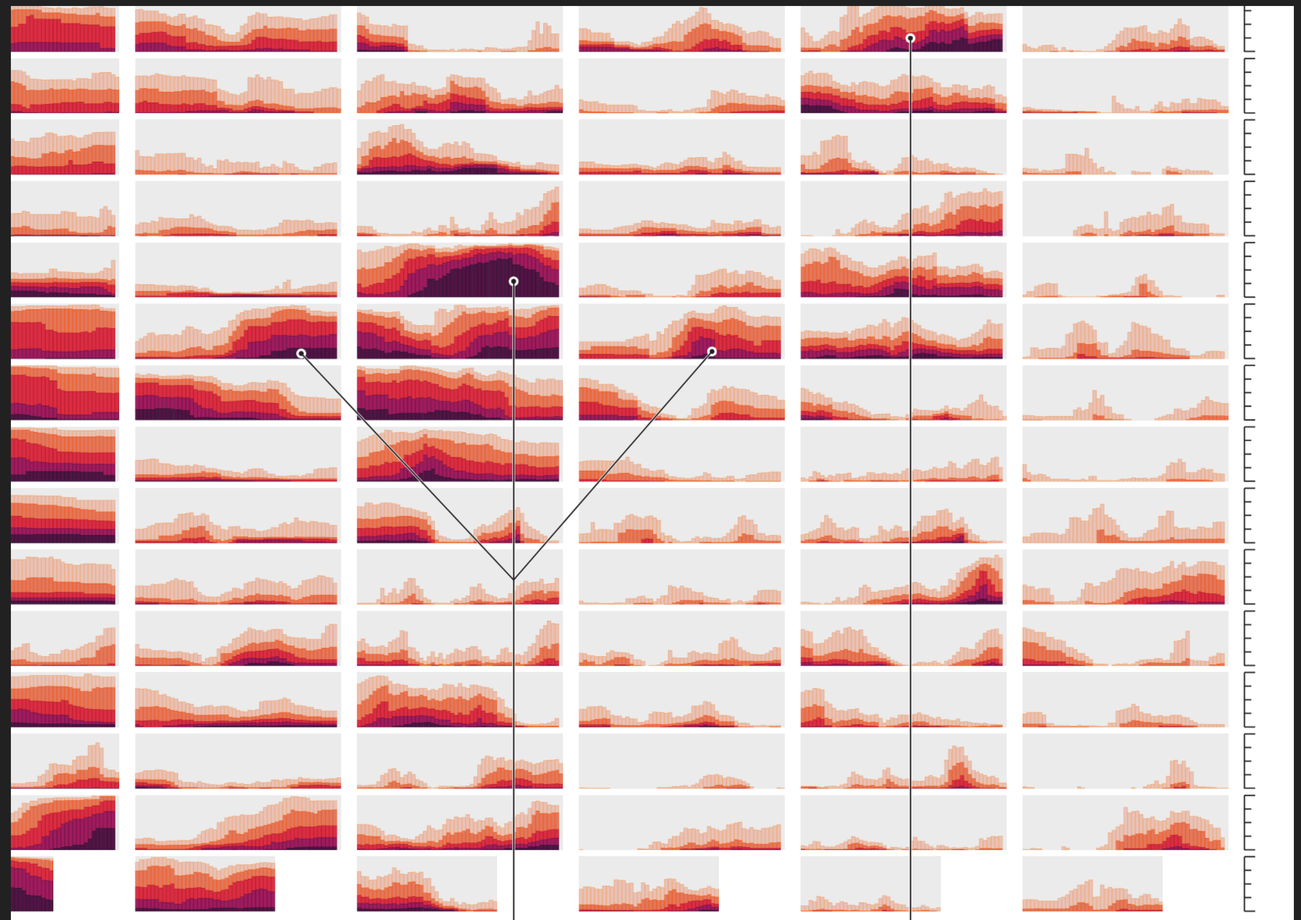


California experienced its hottest drought in recorded history from 2012 to 2016. A warming climate makes the atmosphere thirstier, which increases evaporation and boosts drought.

A drought that originated in the Southern Plains in 2011 eventually spread to the Midwest and Northern Plains when the moisture coming in from the Gulf of Mexico was absorbed by the parched South before it could reach the North.

The Southeast's driest year to date was 2007, when only 31.85 inches of rain fell in Atlanta, 62 percent of its average yearly rainfall.

Source: U.S. Drought Monitor, jointly produced by the National Drought Mitigation Center at the University of Nebraska–Lincoln, U.S. Department of Agriculture, and National Oceanic and Atmospheric Administration (data)



A drought that originated in the Southern Plains in 2011 eventually spread to the Midwest and Northern Plains when the moisture coming in from the Gulf of Mexico was absorbed by the parched South before it could reach the North.

The Southeast's driest year to date was 2007, when only 31.85 inches of rain fell in Atlanta, 62 percent of its average yearly rainfall.

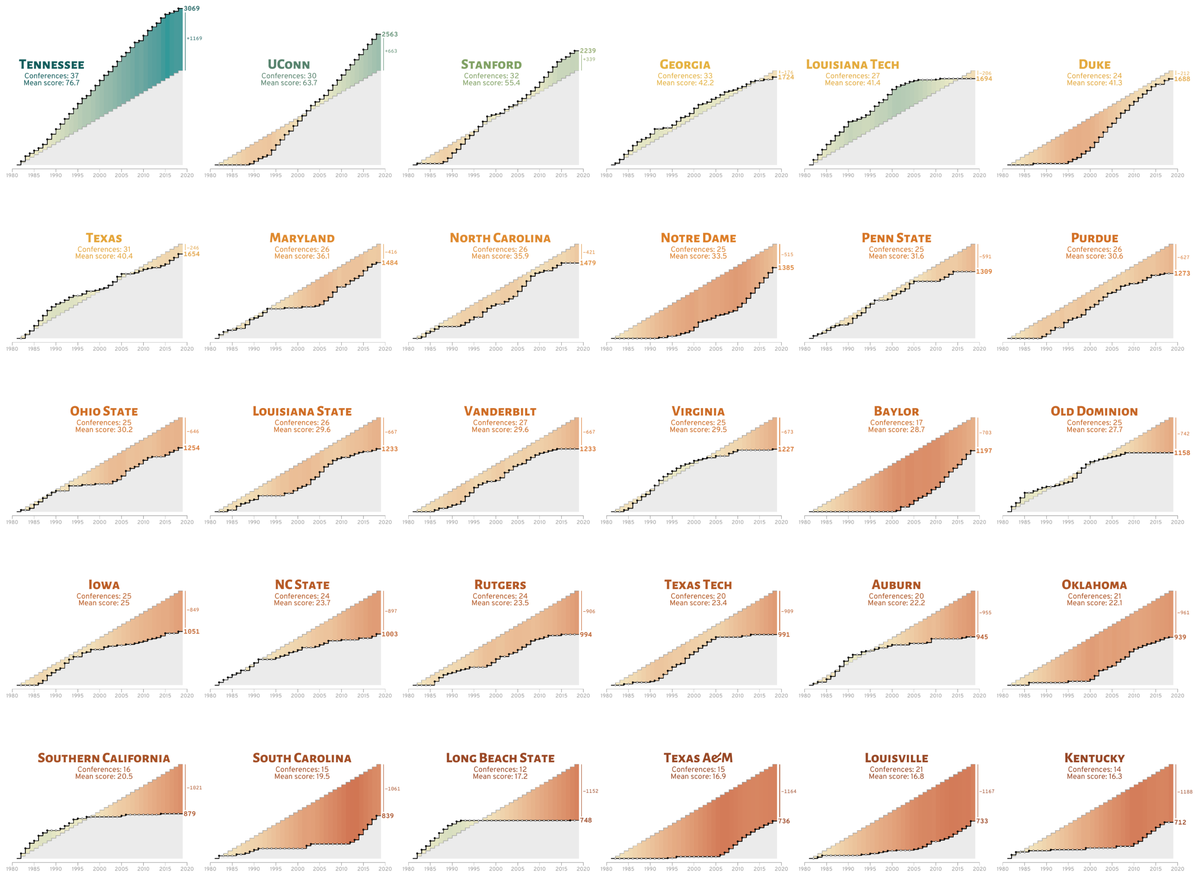
"Escalating Drought", *together with Georgios Karamanis for Scientific American, Issue Nov 2021*

Cédric Scherer // rstudio::conf // July 2022

— The Rise & Fall of Women's College Basketball Dynasties —

A number of teams that were the titans of the early NCAA women's basketball tournament have struggled in recent decades. And in their place, a new ruling class of schools has emerged to become the defining programs of the modern age. FiveThirtyEight estimated the team strength over time based on NCAA Tournament seeds as a proxy in the absence of game-level data. To measure this, FiveThirtyEight awarded "seed points" in proportion to a given seed number's expected wins in the tournament, calibrated to a 100-point scale where the No. 1 seed gets 100 points, No. 2 gets 70 points, and so forth.

The visualization shows the cumulative sum of awarded seed points on a 100-point scale from the very first women's NCAA basketball tournaments in 1982 until 2018 in comparison to a hypothetical team that participated in all of the 37 conferences and gained half of the points each time (grey line). The curves highlight the fall of yesterday's women's basketball powerhouses such as **Louisiana Tech**, **Long Beach State**, **Southern California**, and **Old Dominion** that have been very good throughout the history of the women's tournament but have experienced big drop-offs in seed points over the last years. At the same time, schools such as **UConn**, **Stanford**, **Notre Dame**, **Baylor**, and **Duke** started slow but picked up steam into the present day. Some teams, such as **Tennessee**, have been relatively consistent throughout the NCAA era gathering always more seed points than an average team.
Shown are the top 30 college teams that participated in at least ten conferences between 1982 and 2018, sorted by the cumulative sum of seed points.



Visualization by Cédric Scherer · Data by FiveThirtyEight

"The Rise and Fall of Women's College Basketball Dynasties", *#TidyTuesday Contribution*

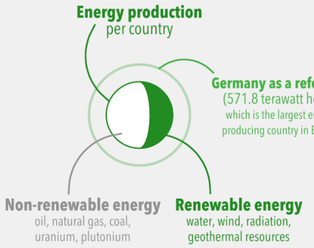
Cédric Scherer // rstudio::conf // July 2022

How European countries generated electricity in 2018

Germany is the largest energy producing country in Europe. It generates the most renewable and conventional thermal energy, representing 31% and 56% of its overall production respectively. **France** is the second largest energy European producer and by far the largest nuclear energy provider: 71% of its production is based on nuclear fission to generate heat.

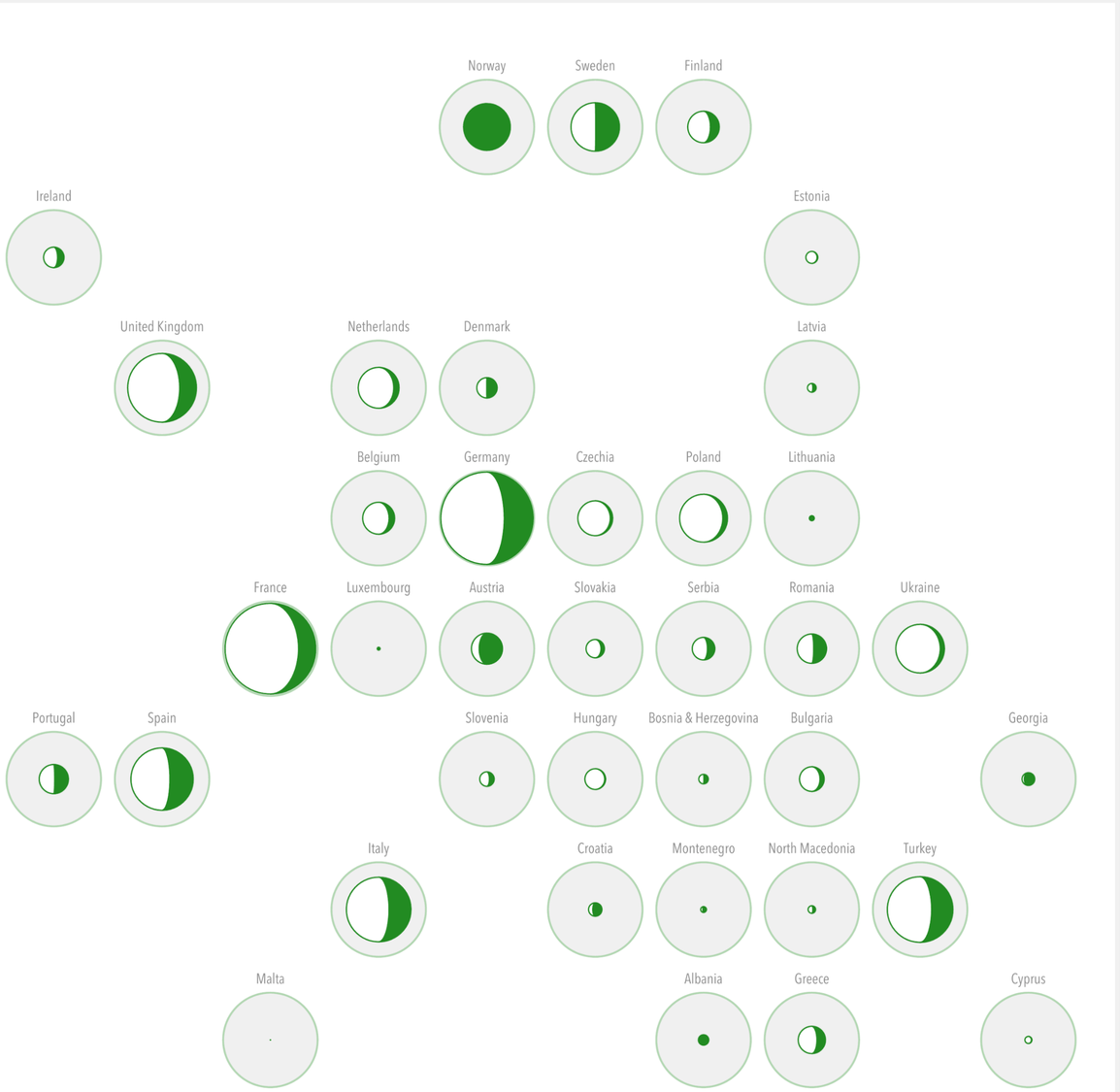
Renewable:	Nuclear:	Conventional thermal:
Germany	**France**	**Germany**
179.1 TWh	393.2 TWh	320.4 TWh
(31% of its production)	(71% of its production)	(56% of its production)

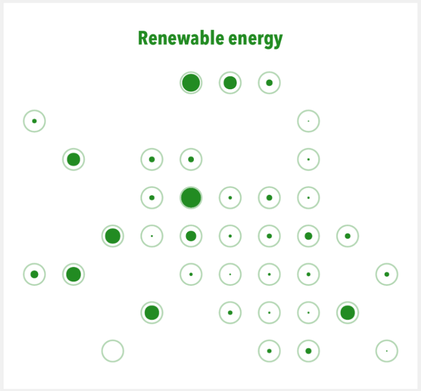
Renewable energy is energy that comes from resources that are naturally replenished such as sunlight, wind, water, and geothermal heat. Unlike fossil fuels, such as oil, natural gas and coal, or nuclear power sources such as uranium and plutonium, renewable energy regenerates naturally in a short period of time.

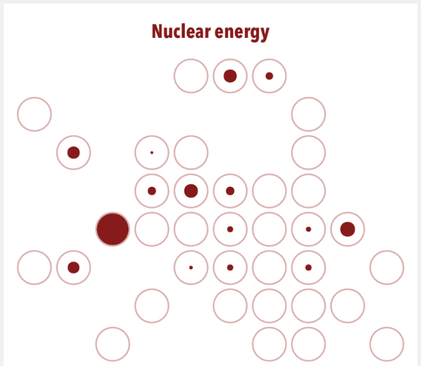


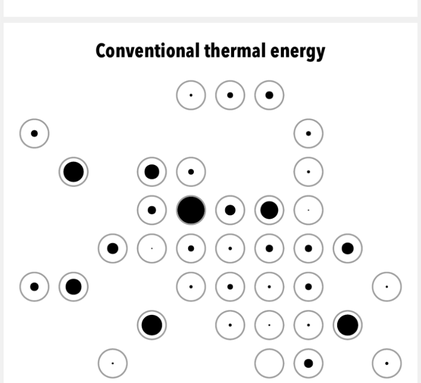
Norway had an electricity production almost entirely made up of renewable energy (98%). This makes Norway the second largest producer of this energy type in Europe. Interestingly, most of the renewable energy is produced by hydro power that take up 95% and only 3% by wind. In contrast, twelve European countries were reported to produce less than 20% of their energy with renewable resources: **Malta** (0%), **Hungary** (5%), **Estonia** (6%), **Czechia** (7%), **Cyprus** (9%), **Ukraine** (9%), **Poland** (10%), **Netherlands** (13%), **Bulgaria** (17%), **Belgium** (18%), **Slovakia** (19%), and **France** (19%).

Note: Energy production is mapped to the area of the circles.
Visualization by Cédric Scherer • Data by Eurostat



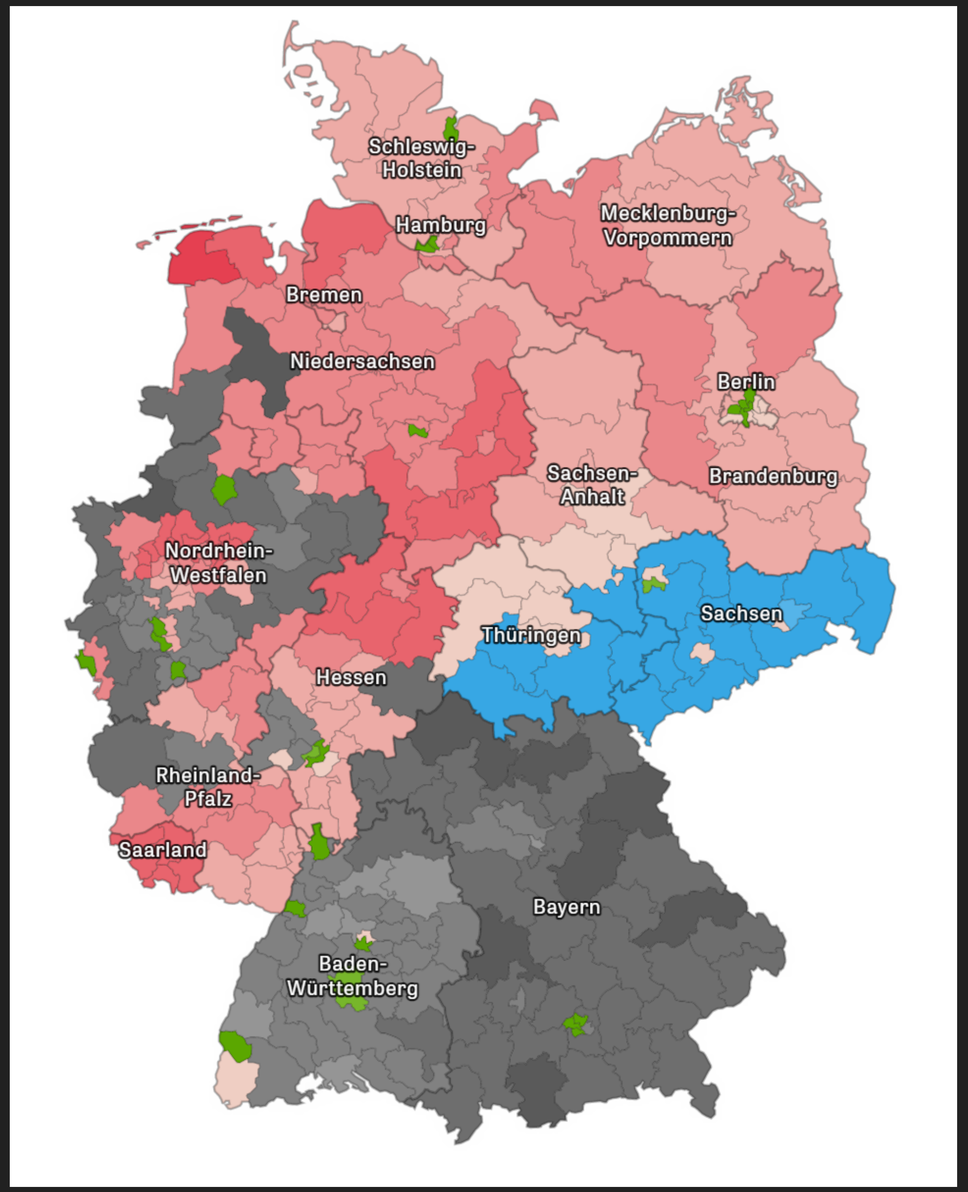


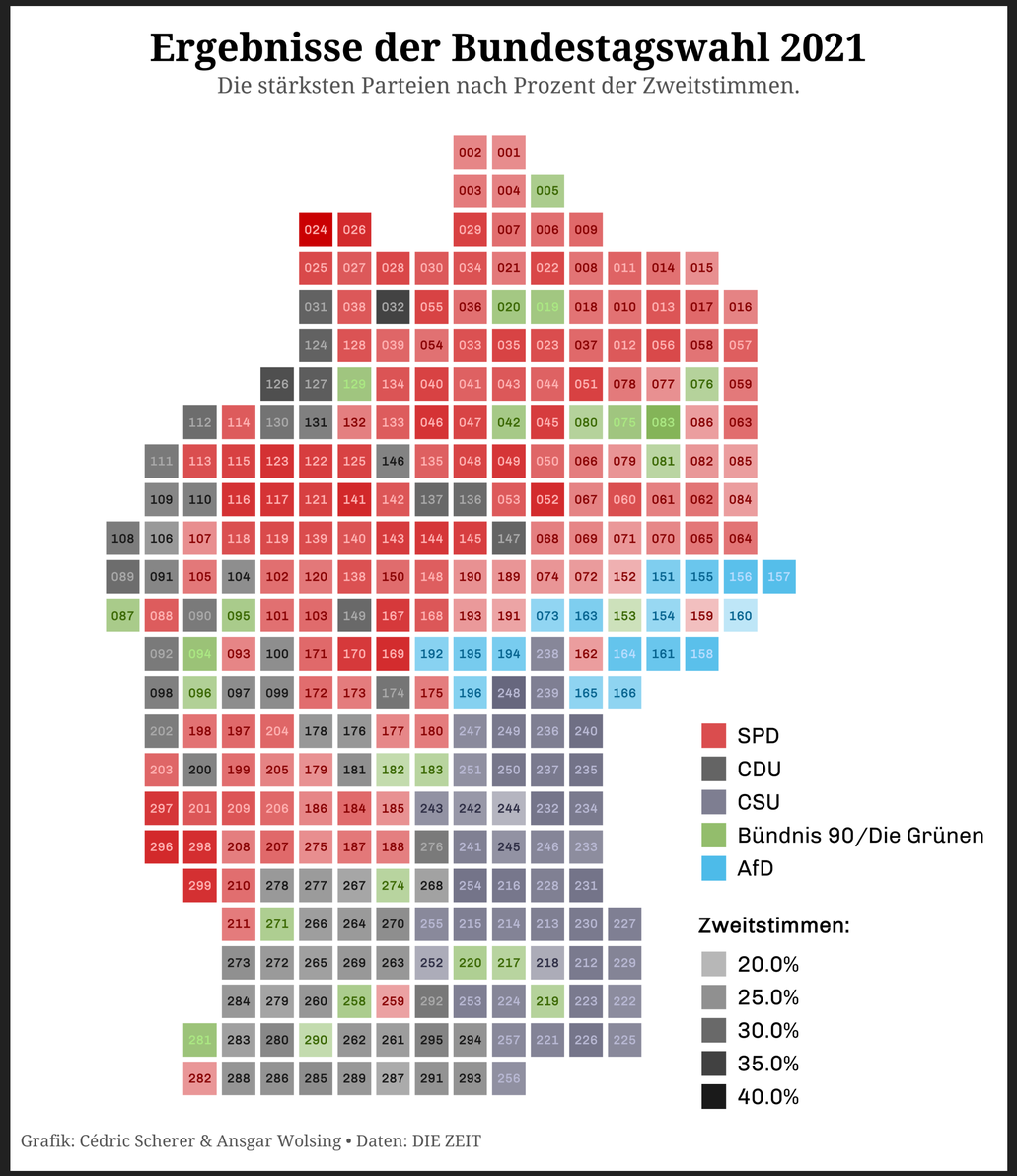




"How European countries generated electricity in 2018", #TidyTuesday Contribution

Cédric Scherer // rstudio::conf // July 2022





Left: Choropleth Map by Die Zeit | Right: Tile Grid Map by Cédric Scherer & Ansgar Wolsing

Visual Form

Follow design rules and data visualization principles

Cédric Scherer // rstudio::conf // July 2022

What is good DataViz design?

Cédric Scherer // rstudio::conf // July 2022

What is good DataViz design?

- Clean layout — “less is more”

What is good DataViz design?

often but not neccesarily!

- Clean layout — “less is more”

What is good DataViz design?

Often but not neccesarily!

- Clean layout — “less is more”
- Use direct annotations to ease readability + interpretability

What is good DataViz design?

Often but not neccesarily!

- Clean layout — “less is more”
- Use direct annotations to ease readability + interpretability
- Make use of hierarchy to guide the reader

What is good DataViz design?

Often but not neccesarily!

- Clean layout — “less is more”
- Use direct annotations to ease readability + interpretability
- Make use of hierarchy to guide the reader
- Consistent use of colors, spacing, typefaces, and weights

What is good DataViz design?

Often but not neccesarily!

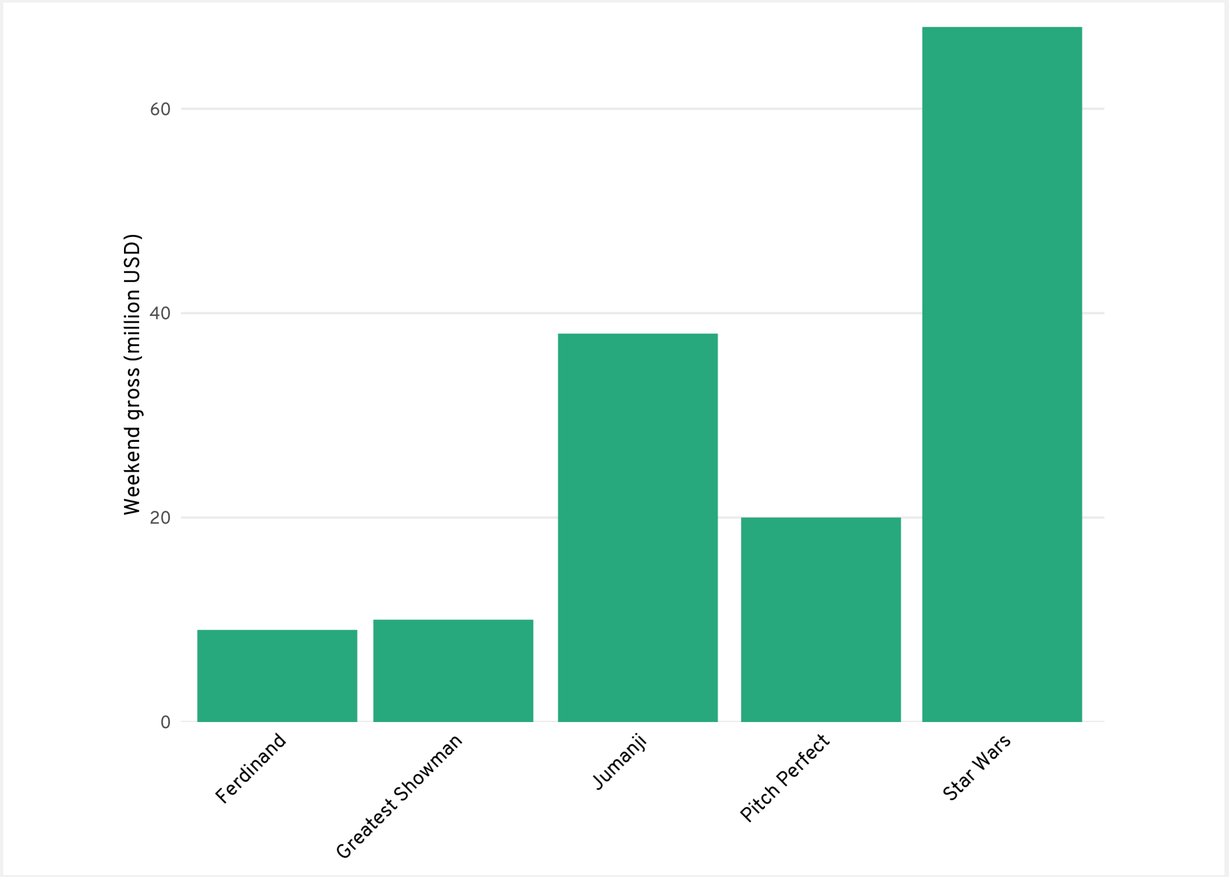
- Clean layout — “less is more”
- Use direct annotations to ease readability + interpretability
- Make use of hierarchy to guide the reader
- Consistent use of colors, spacing, typefaces, and weights
- Use colors wisely and make sure they work for colorblind persons

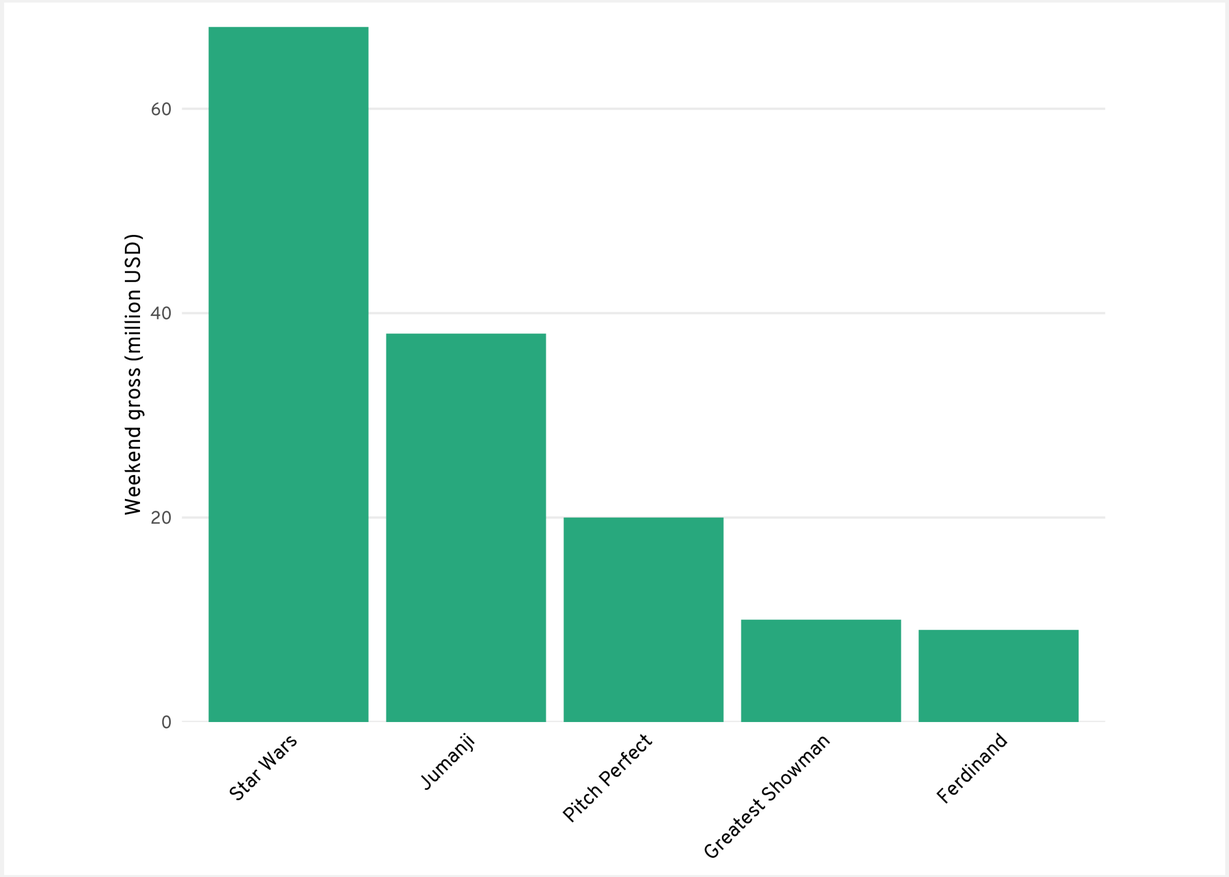
What is good DataViz design?

often but not neccesarily!

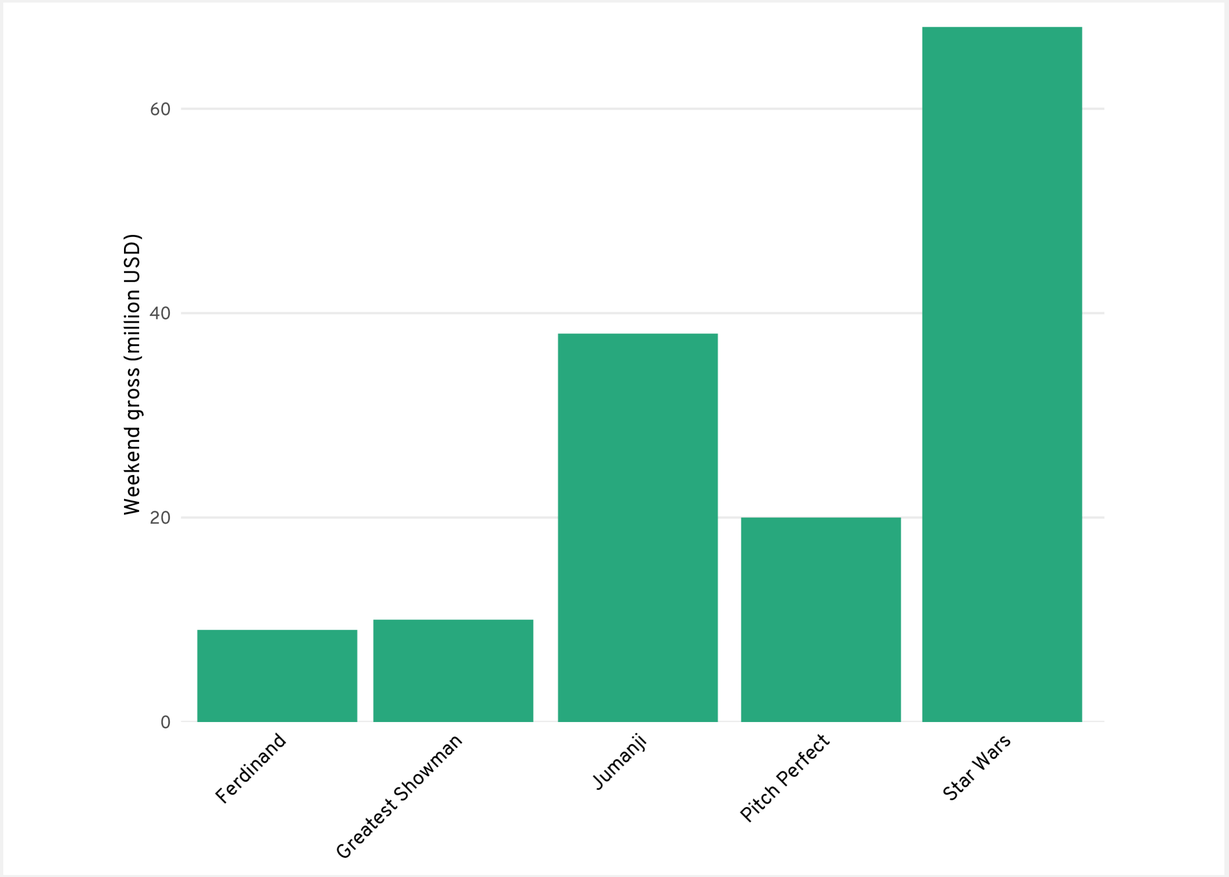
- Clean layout — “less is more”
- Use direct annotations to ease readability + interpretability
- Make use of hierarchy to guide the reader
- Consistent use of colors, spacing, typefaces, and weights
- Use colors wisely and make sure they work for colorblind persons
- Most important information should receive the main attention

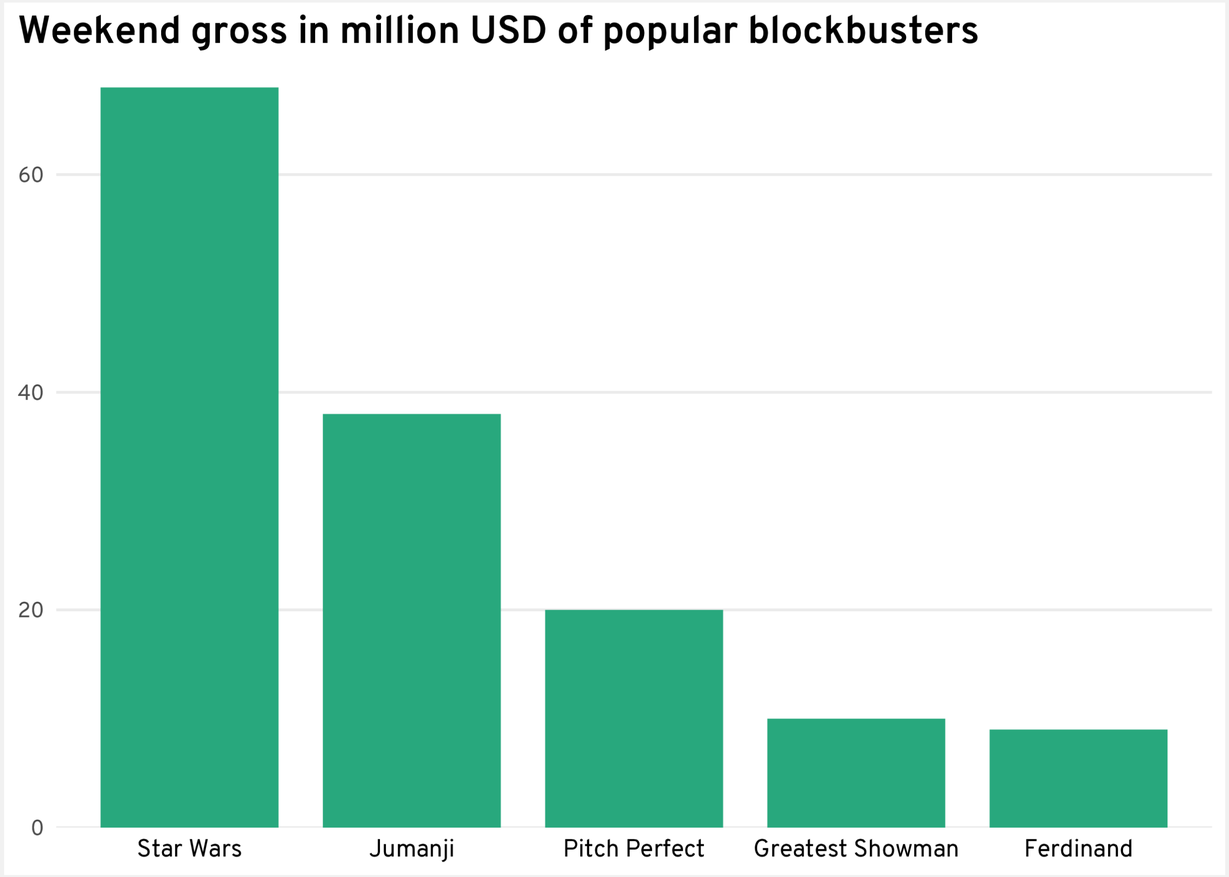
Order your data



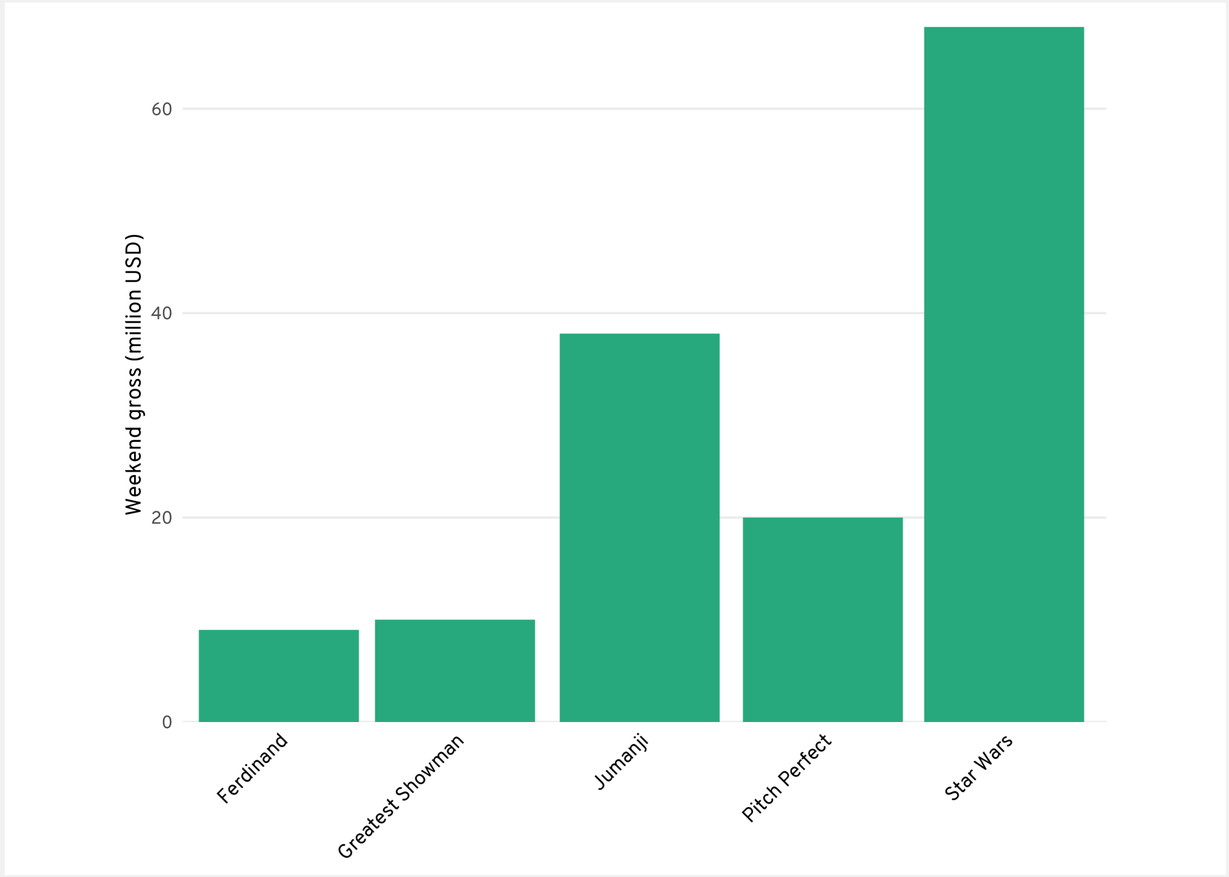


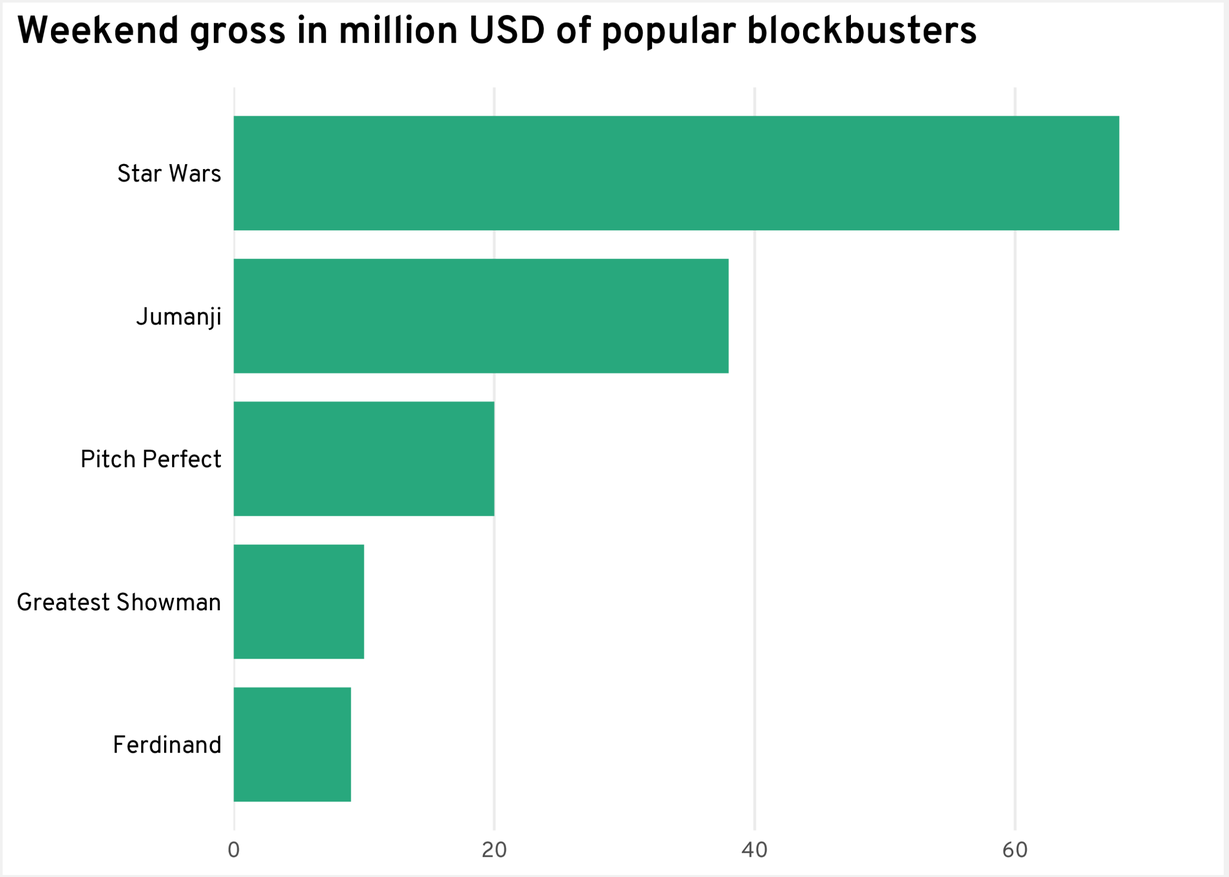
Don’t rotate your text



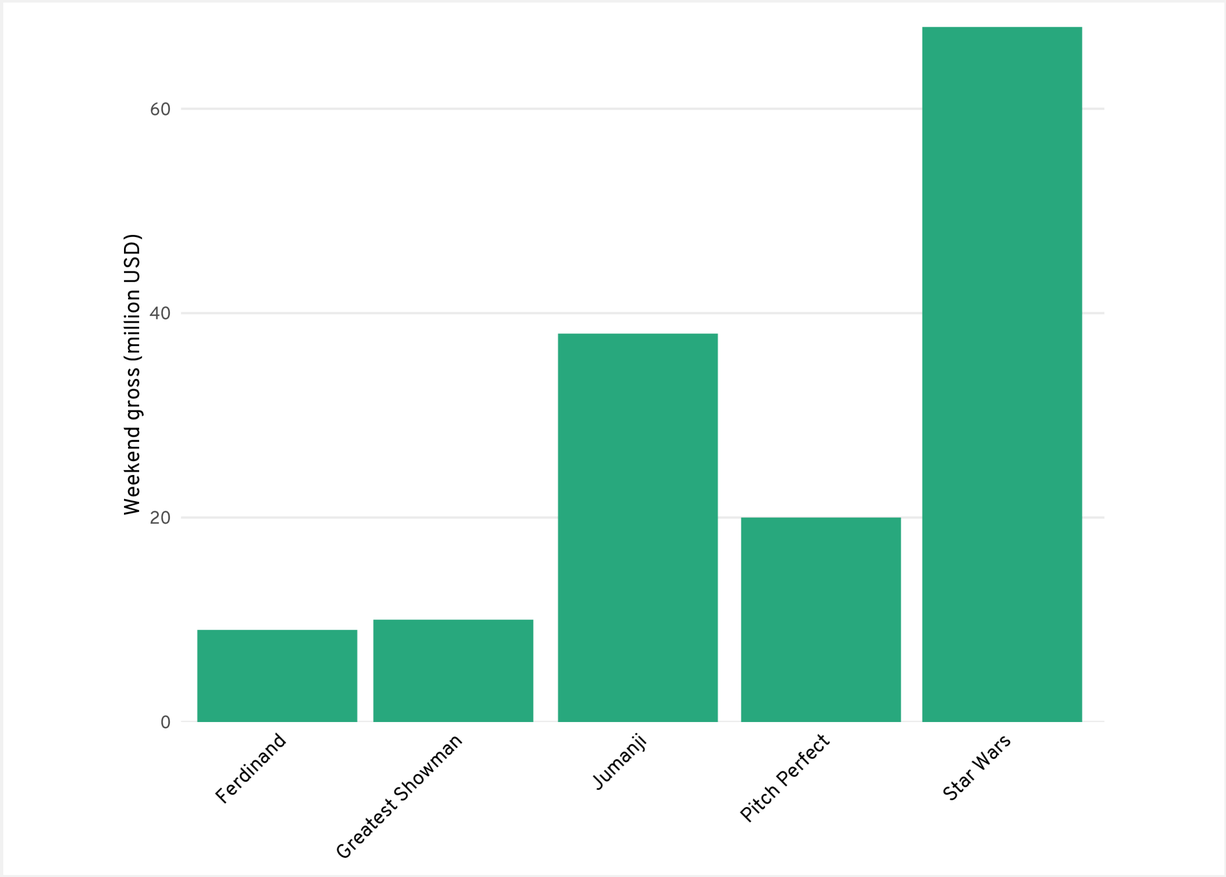


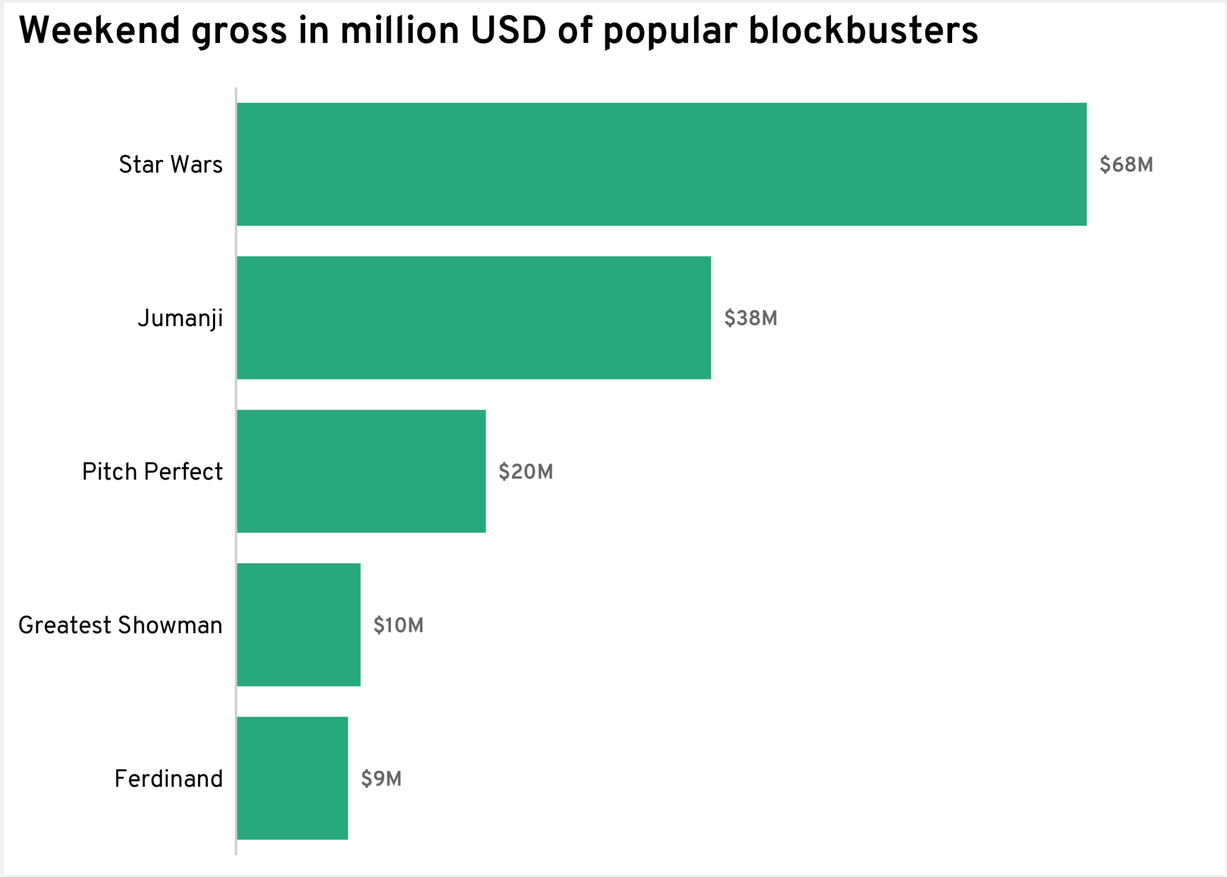
Don’t rotate your text



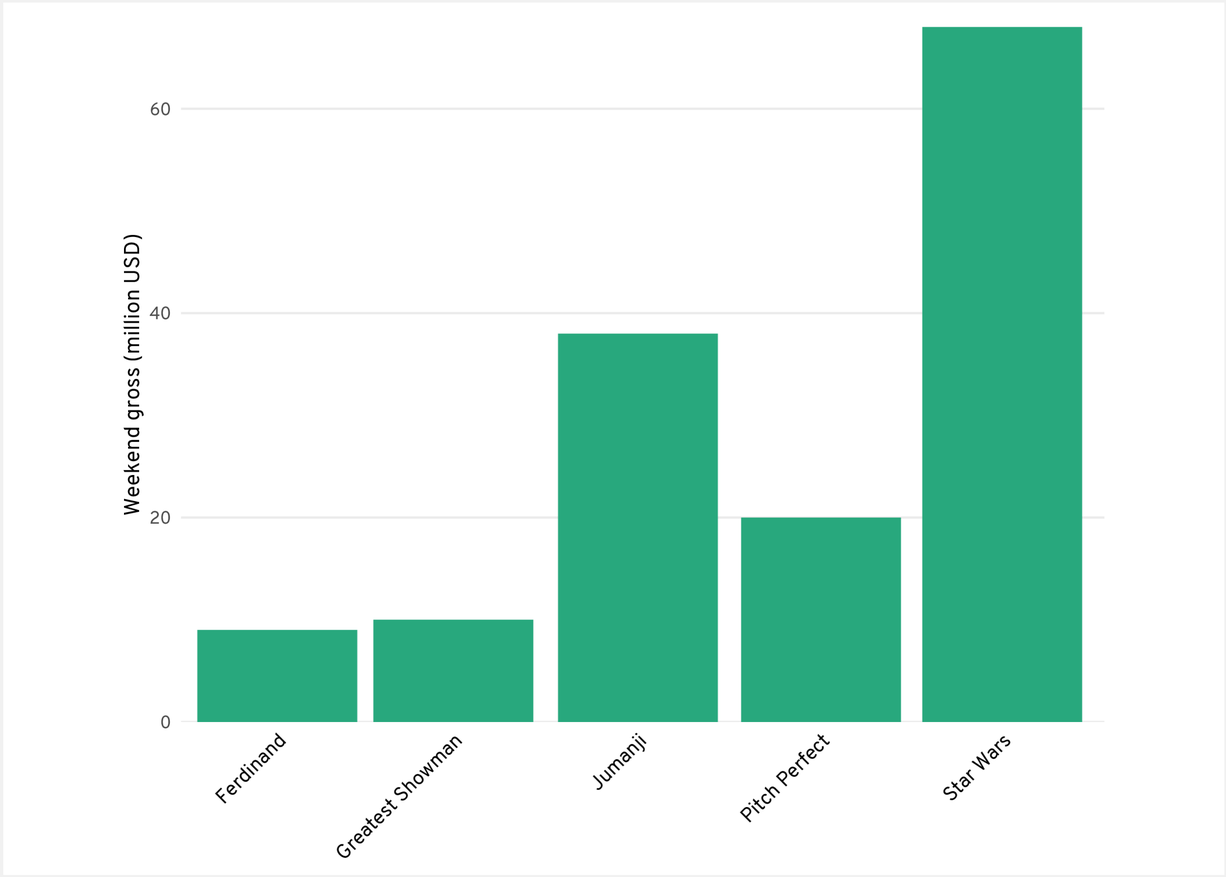


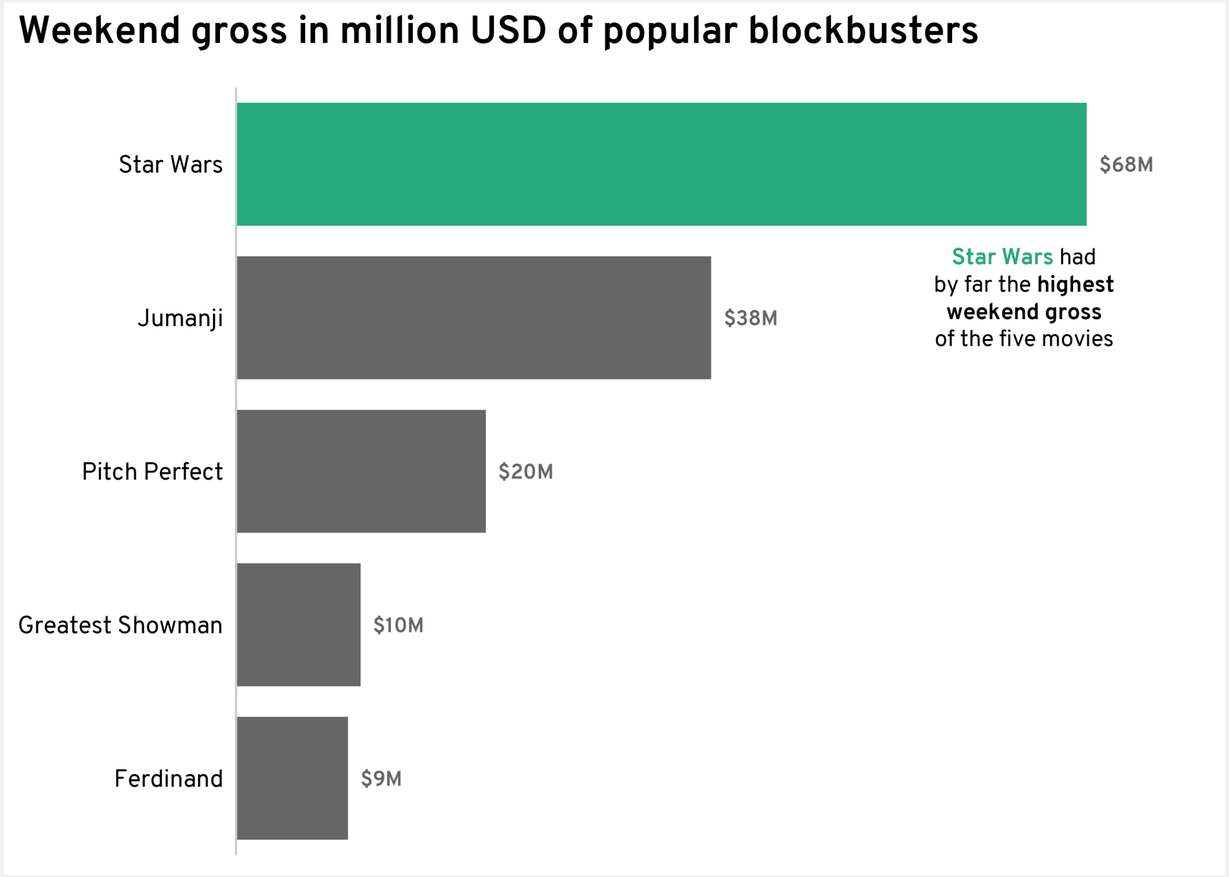
Add direct labels



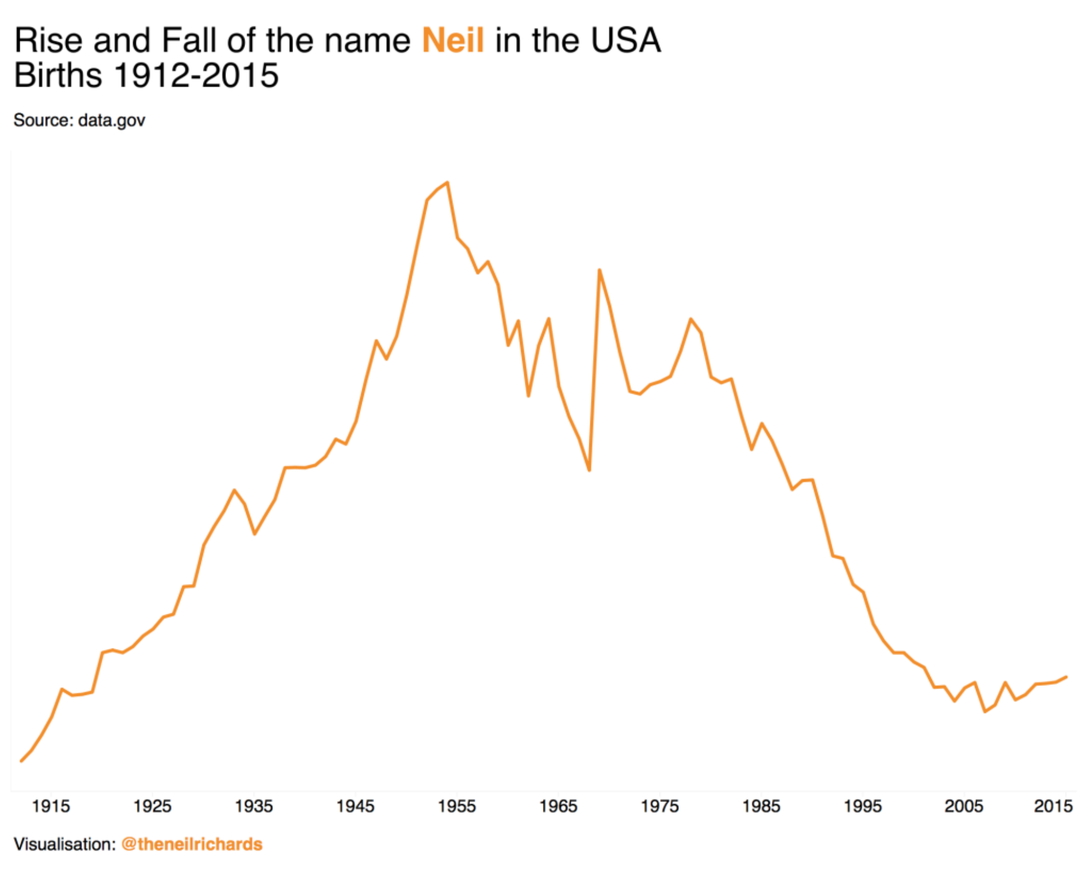


Use colors + annotations wisely



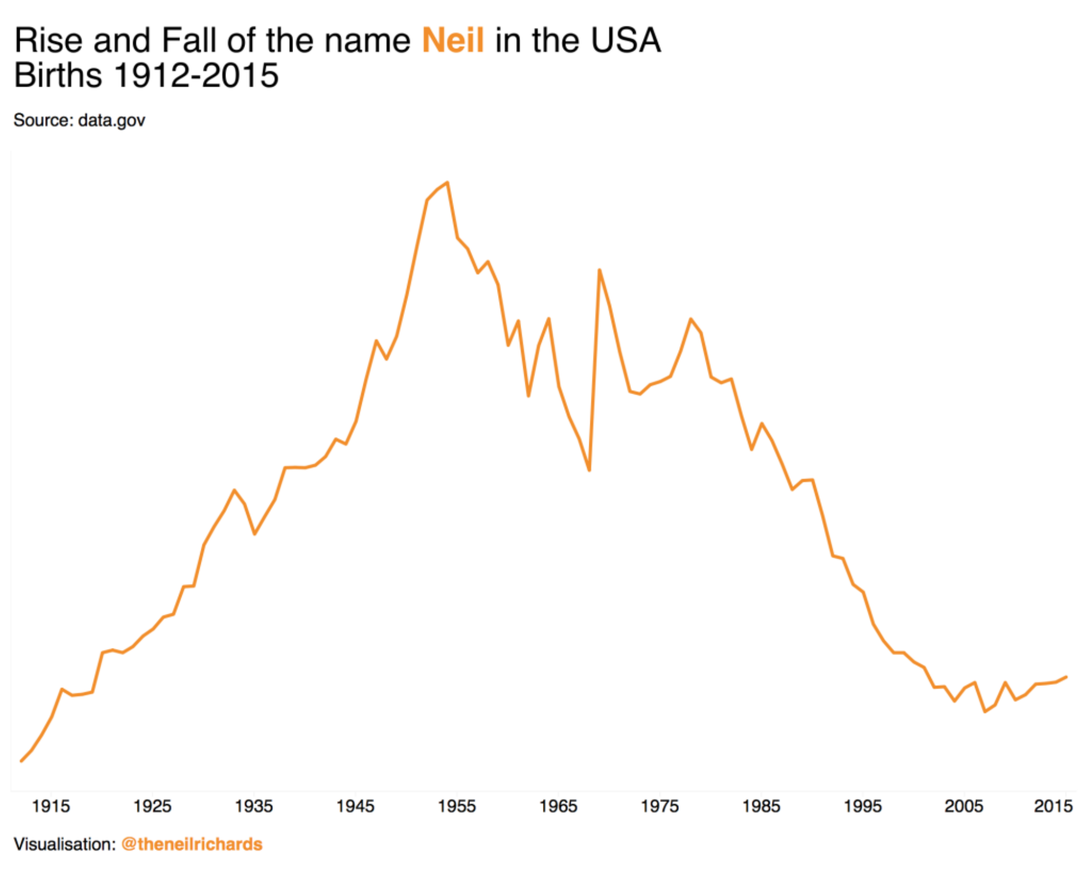


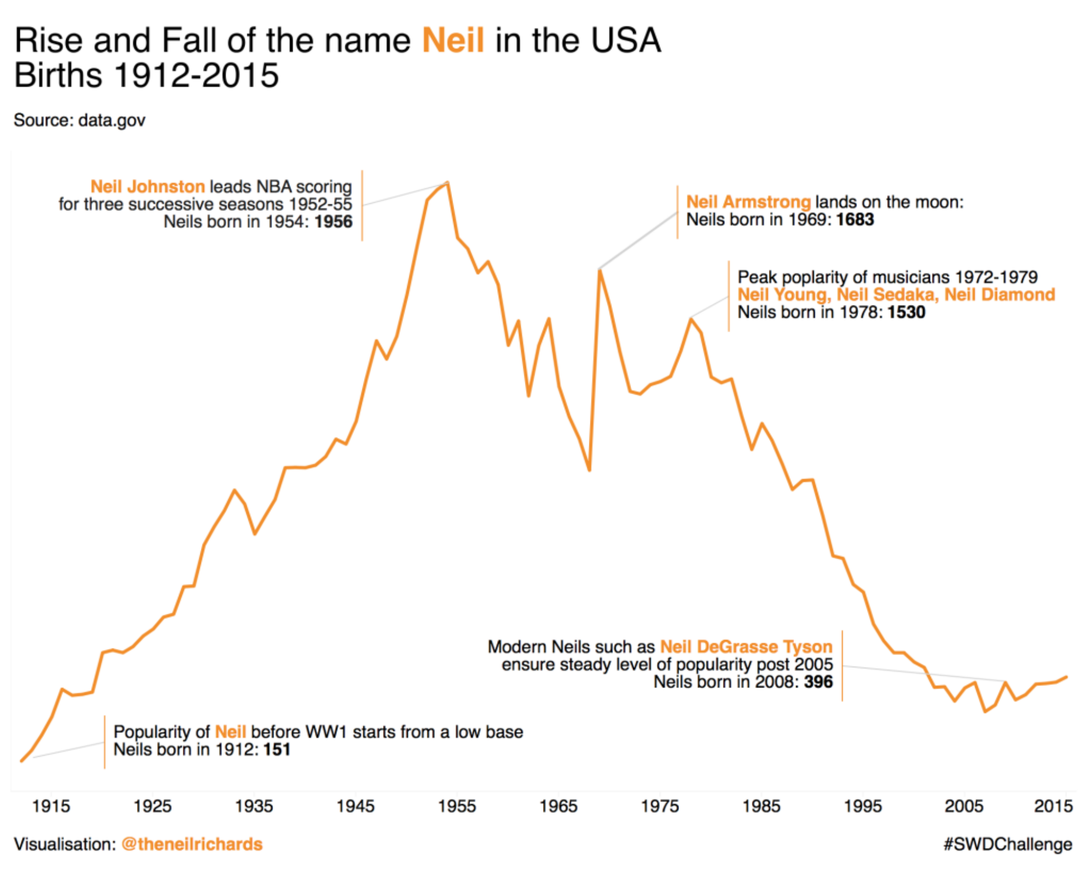
The Power of Annotations



“Is white space always your friend?” by Neil Richards

The Power of Annotations

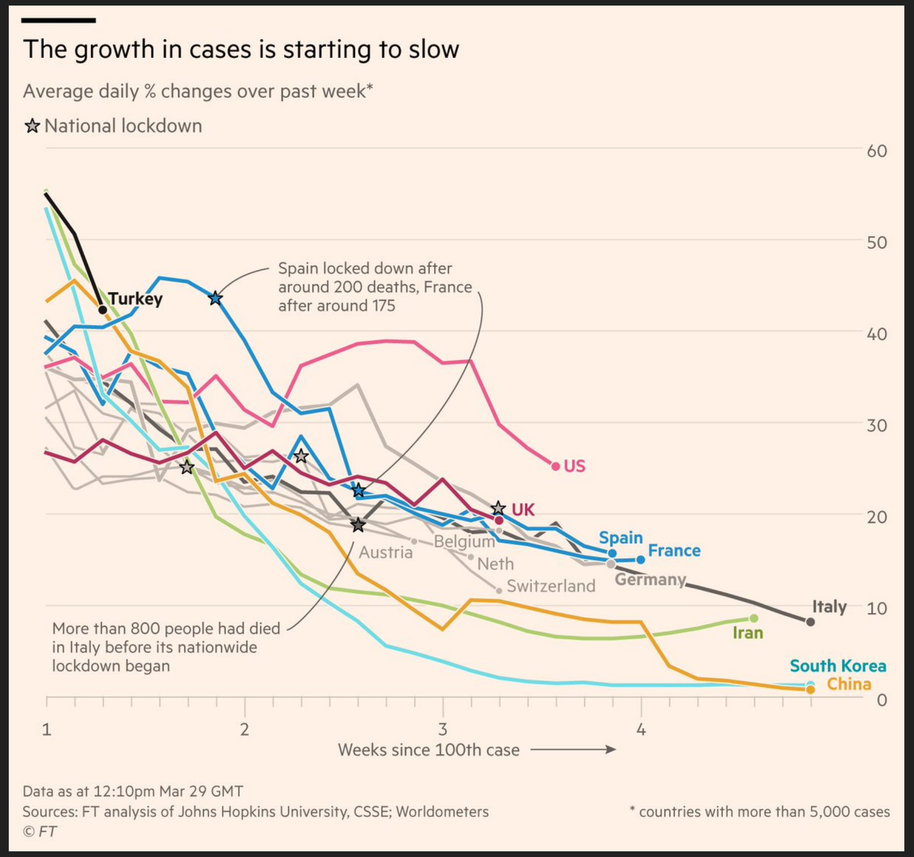


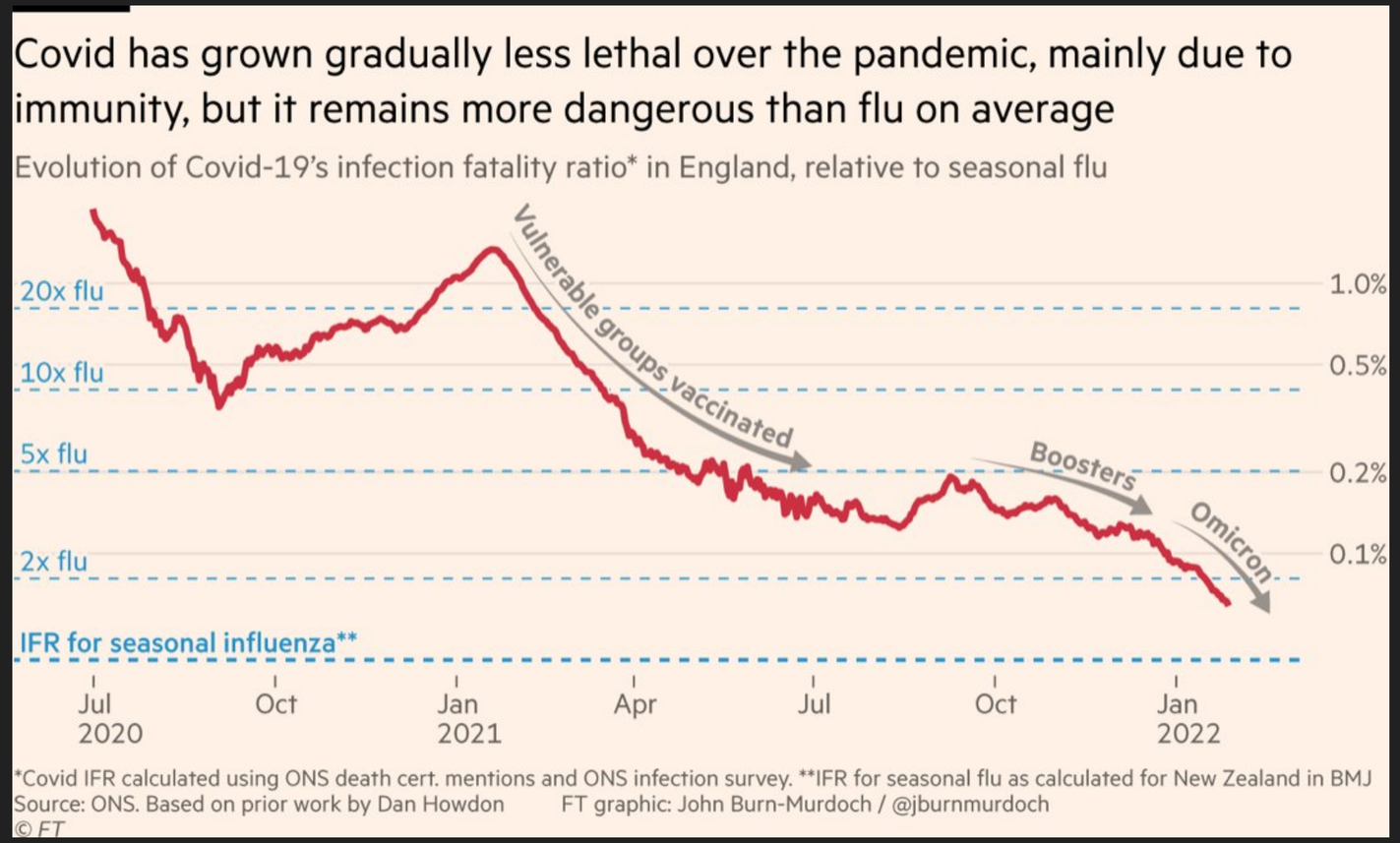


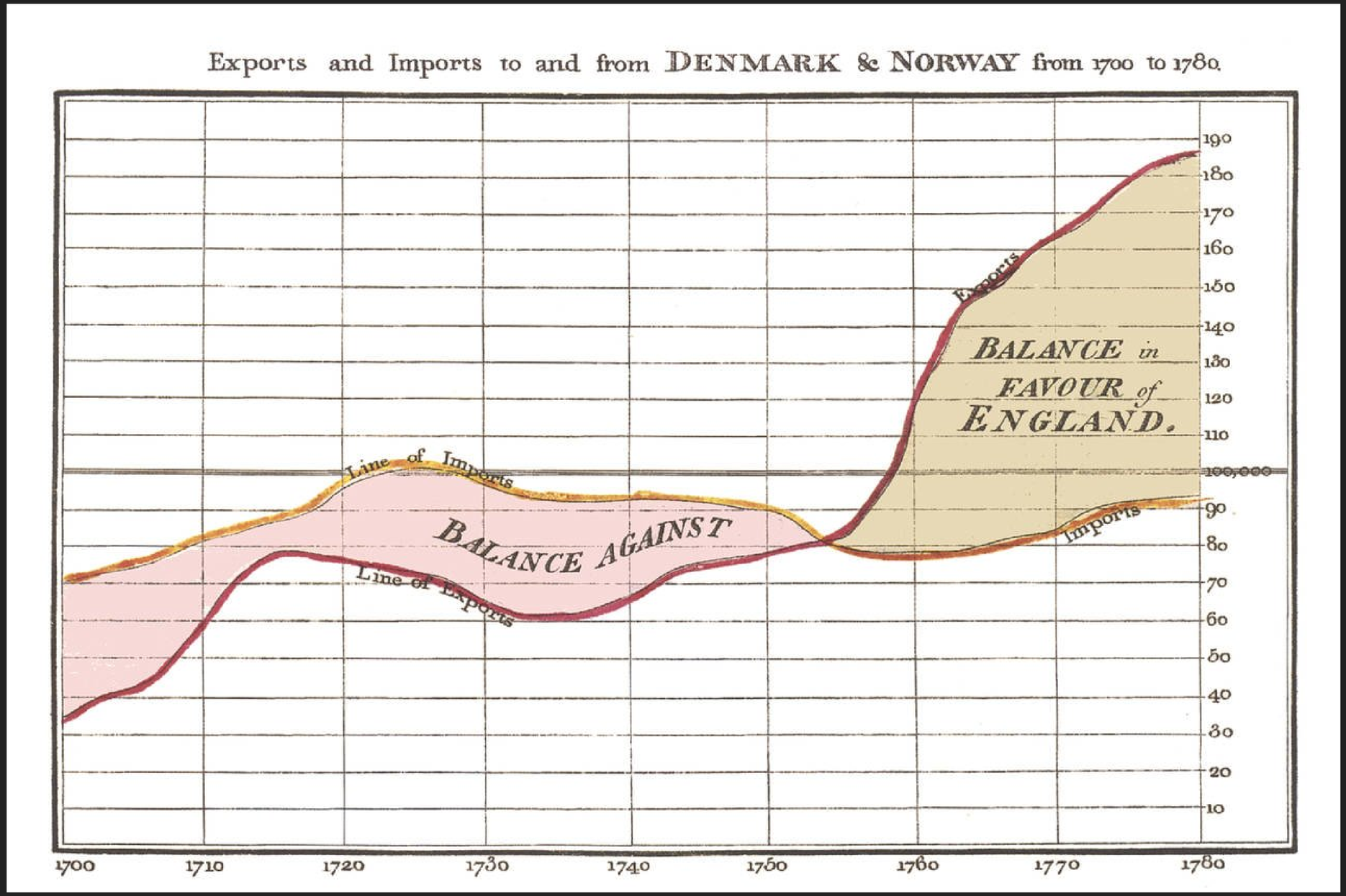
"Is white space always your friend?" by Neil Richards

"The key thing we do is to add a title to the chart,
as an entry point and to explain what is going on.
Text and other annotations add enourmous value for non-chart people."

~ John Burn-Murdoch, Financial Times



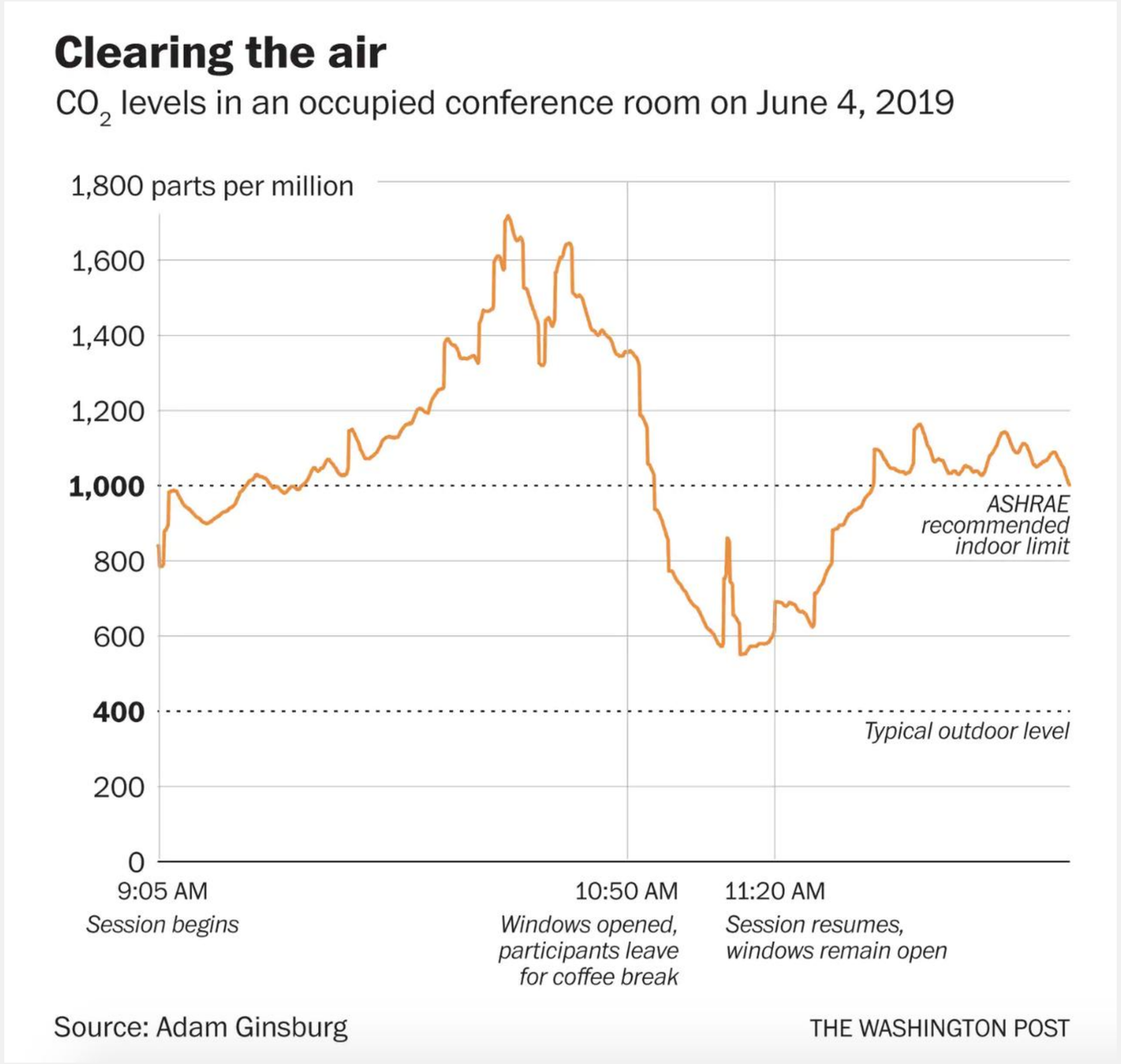




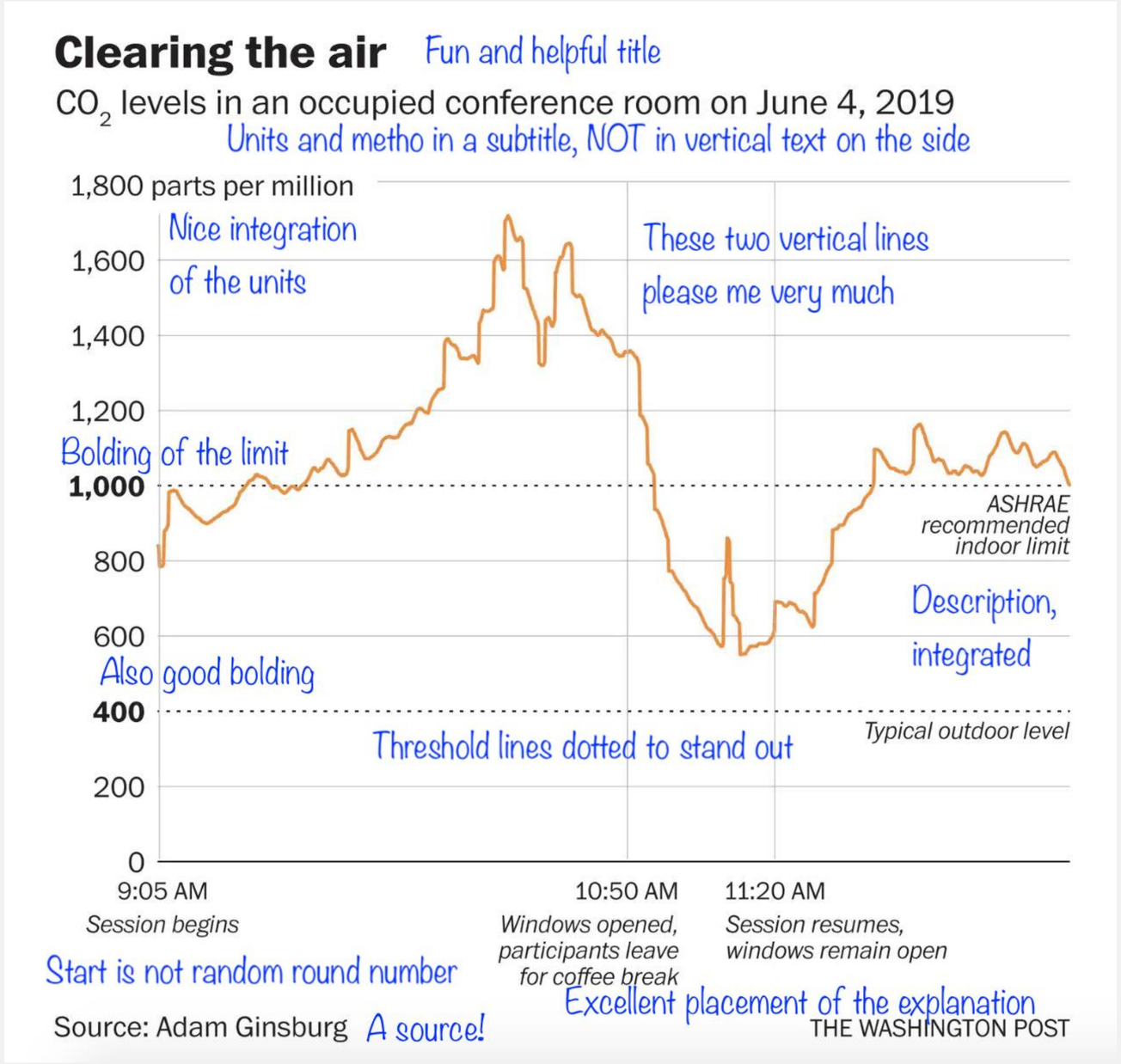
Annotated time-series chart by William Playfair from "The Commercial and Political Atlas and Statistical Breviary" (1786)

Wrap Up

Cédric Scherer // rstudio::conf // July 2022



"Clearing the Air" by Adam Ginsburg (Washington Post)



Notes by Francis Gagnon (Voilà)

Information

Understand your data and be accurate.

Story

Be clear about the message of your visualization.

Goal

Select charts that successfully transport your story.

Visual Form

Follow design rules and data visualization principles.

Your Turn!

We form groups and each group gets a number between 1 and 10.

- Open the image file(s) with the according number in the folder `exercises/4-1-data-communication`

- Discuss the visualization with regard to the 4 levels of dataviz design.
  - Overall, do you think it is a good or a bad visualization?
  - What are details you like?
  - How could one improve the chart?
  - Is there another (potentially better) way to tell the story?
    →Sketch it (and think about how you could build it with ggplot2)