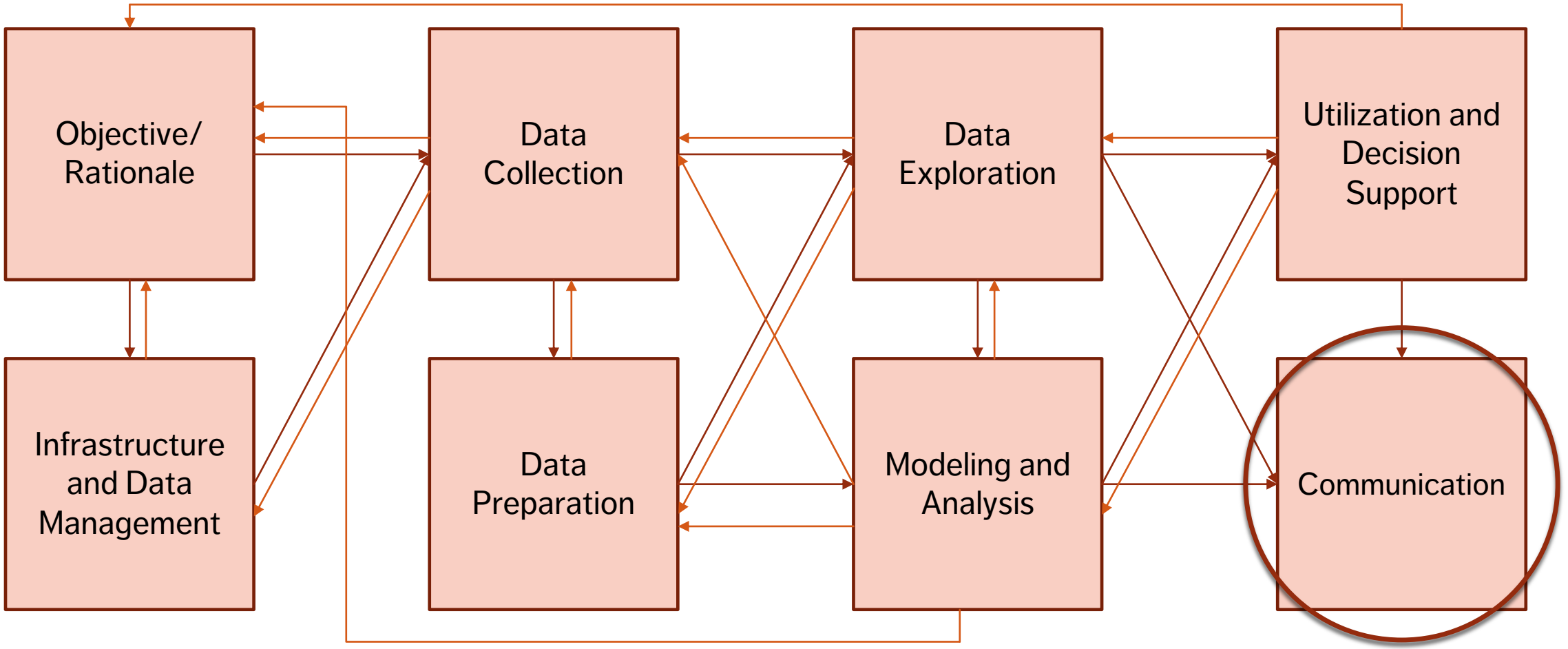
STORYTELLING WITH DATA

THE (MESSY) ANALYSIS PROCESS



WHO IS THE AUDIENCE?

Avoid general audiences: address **Lines of Business** (finance, engineering, HR, etc.)

Identify **decision-makers** and the various audience **roles**

Ask the following questions:

- what relationship do you have with them?
- how do they perceive you?
- how do you establish trust and credibility?

WHAT IS NEEDED FOR THEM TO KNOW OR DO?

Ask for **action:**

- what decisions are people going to make from the analysis?
- how often are they going to be looking at the data?
- how often do they expect the data to be refreshed?

HOW DO WE COMMUNICATE EFFECTIVELY WITH THE AUDIENCE?

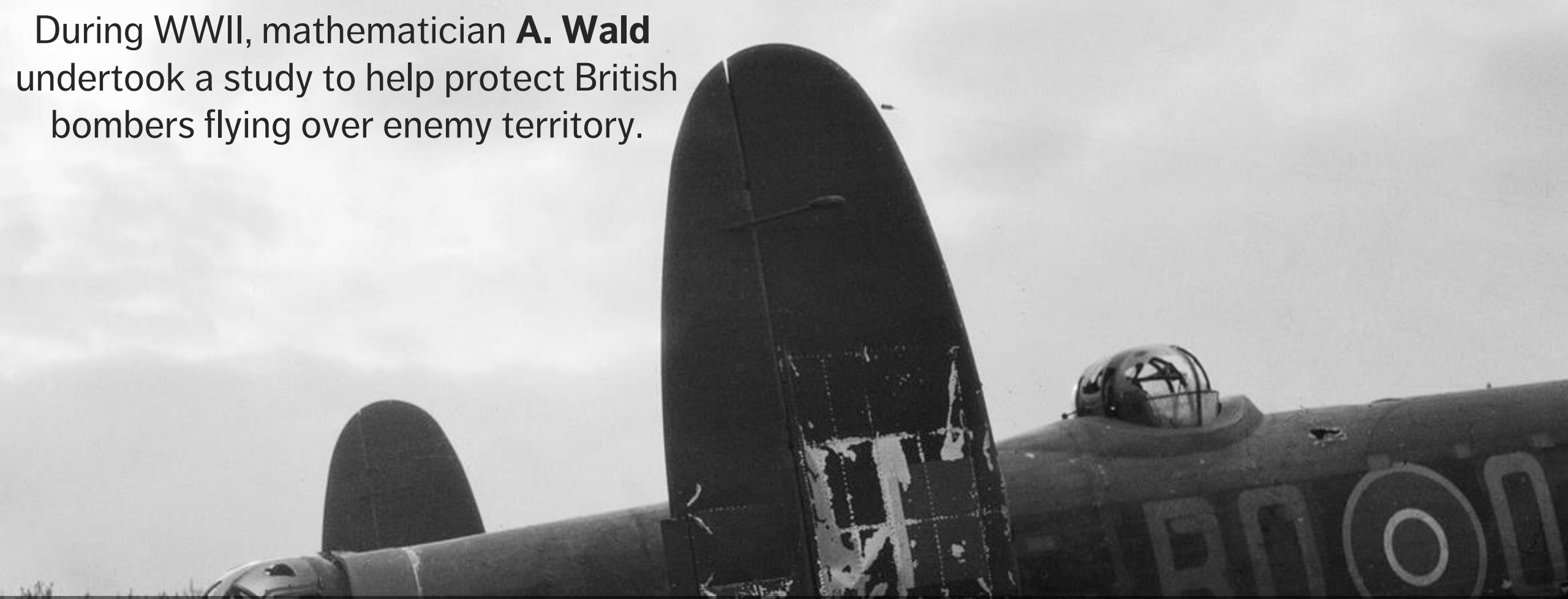
What data is **actually available**?

- is the data clean?
- can it be accessed?
- is it being “massaged”, used to paint a rosy picture?

How much will the audience need/want to **interact** with the charts?

- are they passive?
- can they run limited filtering?
- what data can they download (if any)?

During WWII, mathematician **A. Wald** undertook a study to help protect British bombers flying over enemy territory.



Data included: the **number** and **location** of **bullet holes** on returning aircraft, and the goal was to use this information to determine where to add armor to best protect the plane's structure.

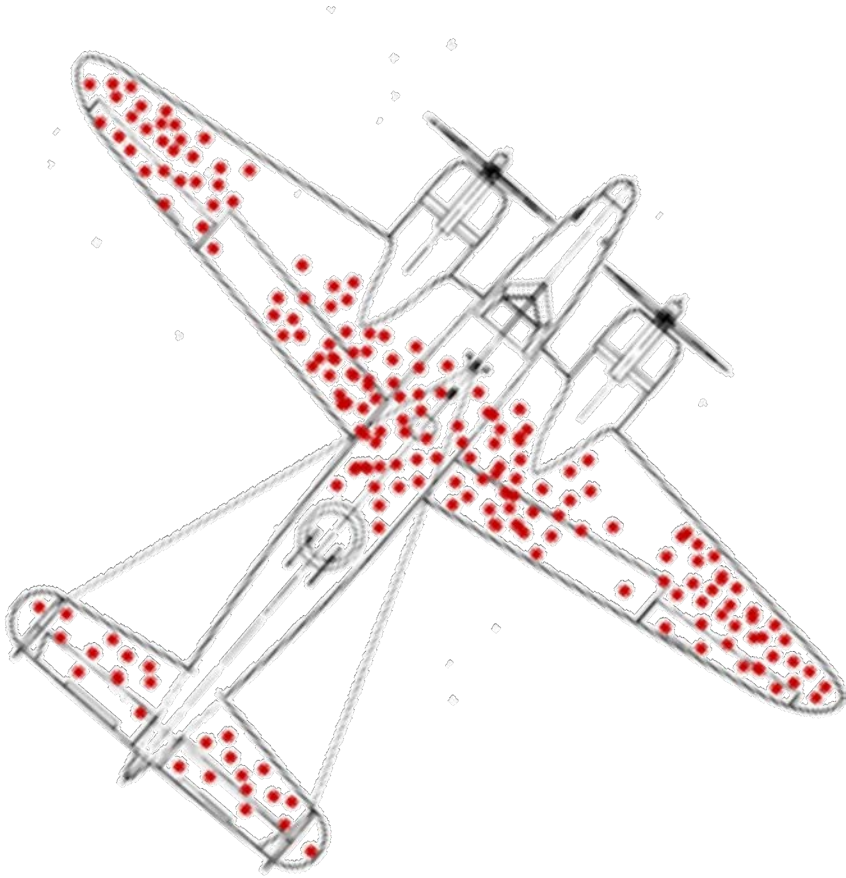
A chart was created to show where the maximum number of bullet holes were located on **returning aircraft**. This chart showed greatest damage on the **aircraft extremities**, not on the main wing and tail spars, engines, and core fuselage areas.

WALD'S STORY

As such, the Air Ministry wanted to add armor to the **extremities**. Wald suggested they were **dead wrong**.

To avoid **"survivorship bias",** armor should be added to the areas with the **fewest holes**: if no returning planes had holes in their wing spars and engines, then even a few holes in those locations were **deadly**.

Take-Away: the data that is missing may be as important to story than the data that is there.



CREATING A NARRATIVE

There are a number of ways of constructing a **narrative**, including:

- chronological
- most important first, or least important first
- begin with the end
- success first, bad news last, or bad new first, success last

Advice: tell the story of the data in a number of different ways

Some dashboards are temporary but some will be a constant reference: this has an impact on how the data should be presented.

MAINTAINING A CLEAR NARRATIVE

Horizontal logic:

- if your visualizations span many pages then the title of each page should tell you the story
- reinforce with an executive summary dashboard or report at the beginning

Vertical logic:

- one page or many, the content should reinforce the title and *vice versa* (self-reinforcement)
- there should be a logical link between all the elements, tags and visual aids on the page

TYPES OF MEMORIES

Telling stories engages our memory:

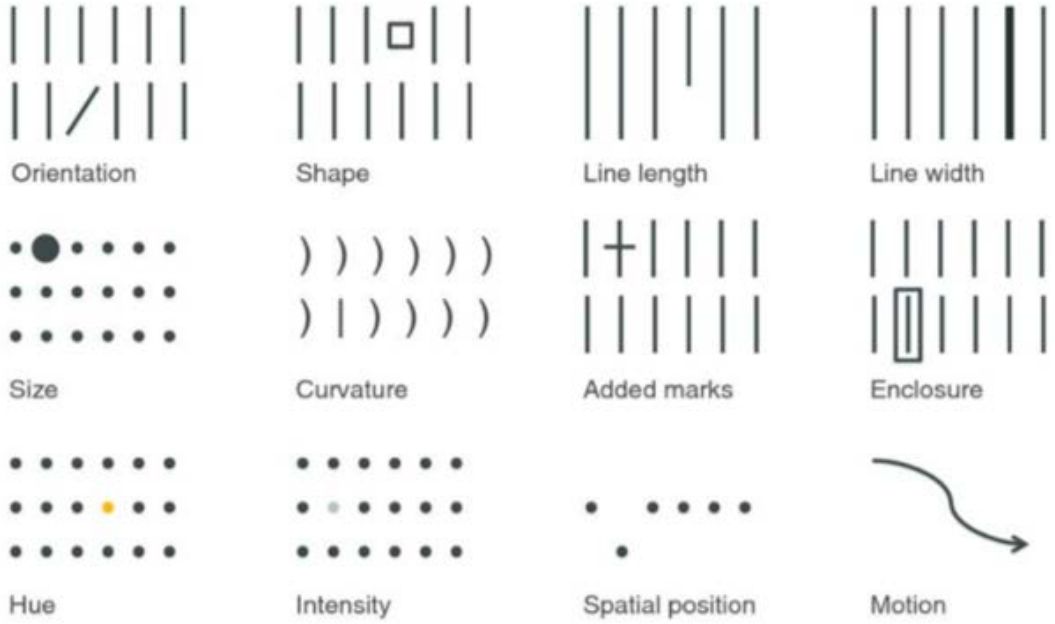
1. iconic memory
2. short-term memory
3. long-term memory

ICONIC MEMORY

Iconic memory is the **visual sensory memory** (SM) register relating to the visual domain and a fast-decaying, high-capacity store of visual information.

Iconic memory is **very brief** (< 1000 ms) and provides a **coherent representation of our entire visual perception**.

Tuned to **pre-attentive attributes** (subconscious accumulation of information from the environment).



SHORT-TERM MEMORY

We can hold ~4 chunks of visual information in **short-term memory** at a given time.

When presented with more chunks (such as data points on a graph), chunks need to be **processed in and out of memory**.

Generally, we try to form **bigger**, **focused** hierarchies of chunks (Gestalt principles).

LONG-TERM MEMORY

Long-term memory is built up over a lifetime and is the basis for pattern recognition and general cognitive processing.

It is an aggregate of **visual** memory and **verbal** memory.

Images help us recall long-term memory, making the story **“stick”**.

Context-providing text also makes a difference:

You have currently selected 28,711 ATIP requests totaling 6,597,612 pages of information

VS

28,711
requests

6,597,612
pages

STORYBOARDING

Storyboarding is a way to summarize the flow of information into a **coherent whole**.

It helps us determine **how many pages/elements per page** we might need.

This is NOT the same as designing the layout of a dashboard.

Storyboarding is used to **define the story** and the dashboard's **content**.

STORYBOARDING – EXAMPLE

1. State intended hiring goal for the year

2. Describe what is driving the hiring (Fed Gov't Init)

3. Show how close/far the goal is as of today

4. Show which departments have the highest requirements

5. Demonstrate which groups are impacted the most

6. Ask/tell the reader how they can help

EXERCISE

Individually or in teams, prepare a storyboard for the dashboard you designed in the previous section.

FUNDAMENTAL PRINCIPLES OF DATA VIZ

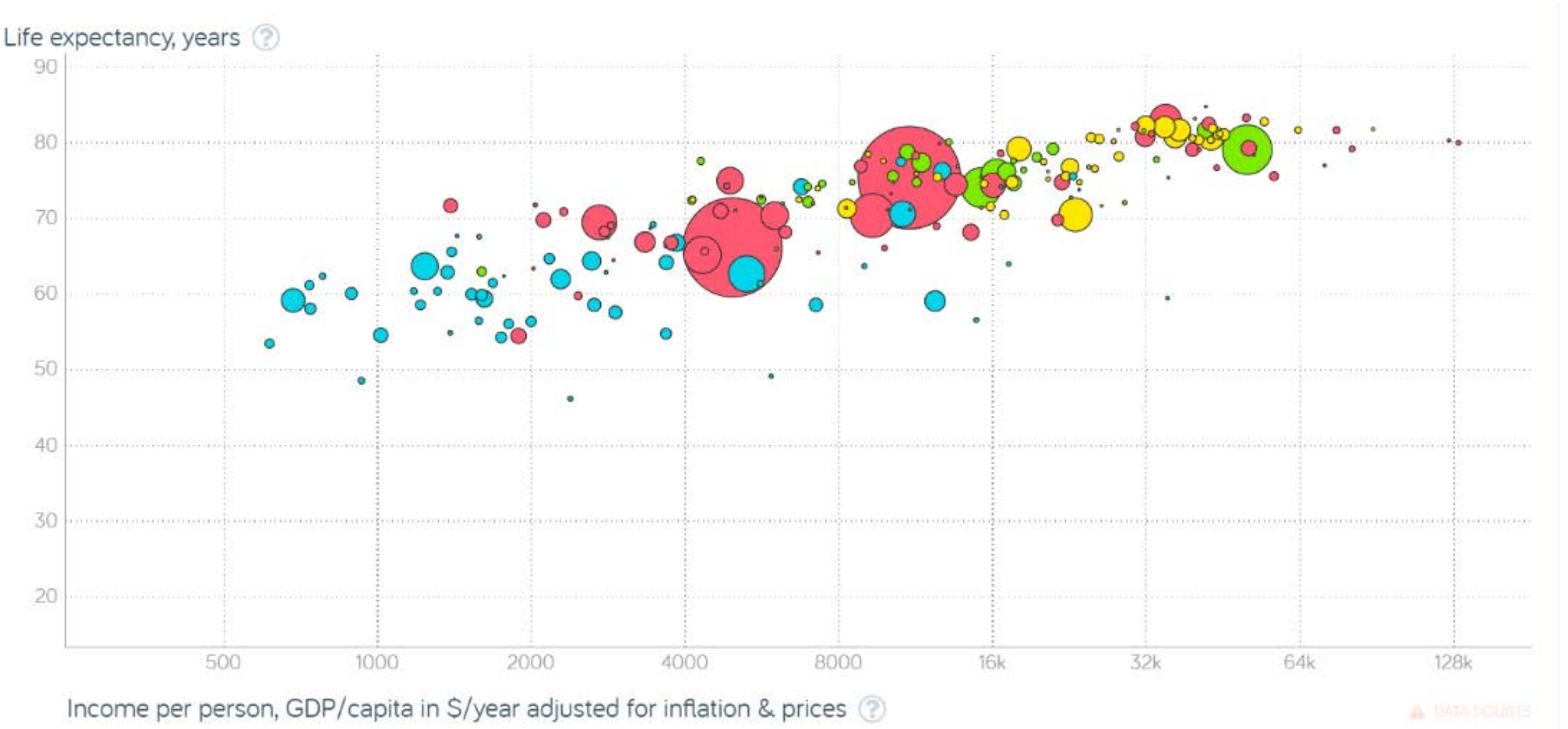
[E. Tufte, Beautiful Evidence]

FUNDAMENTAL PRINCIPLES OF ANALYTICAL DESIGN

Symmetry to visual displays of evidence: consumers should be seeking exactly what producers should be providing, namely

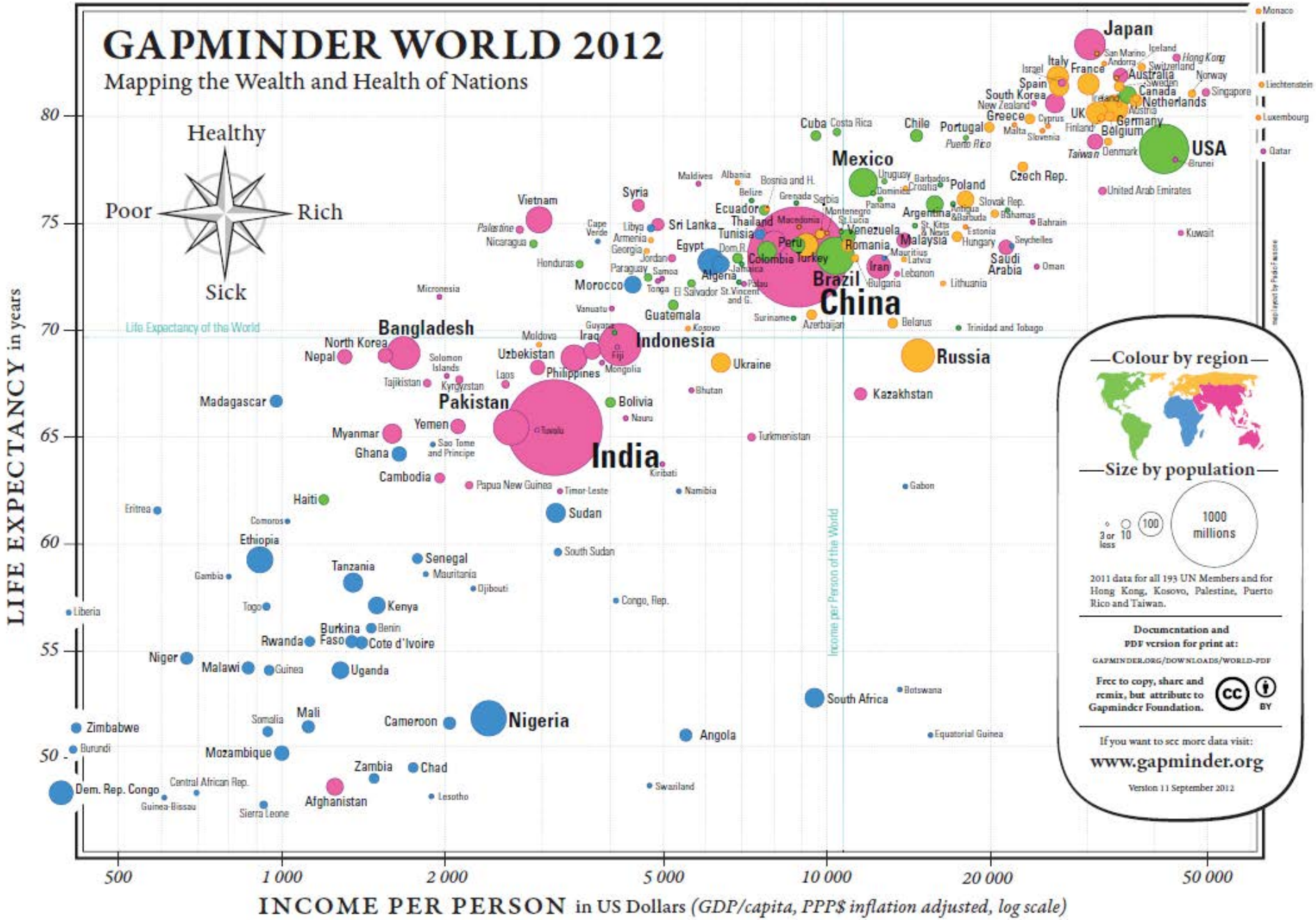
- meaningful comparisons
- causal networks and underlying structure
- multivariate links
- integrated and relevant data
- honest documentation
- primary focus on content

[Gapminder Foundation]

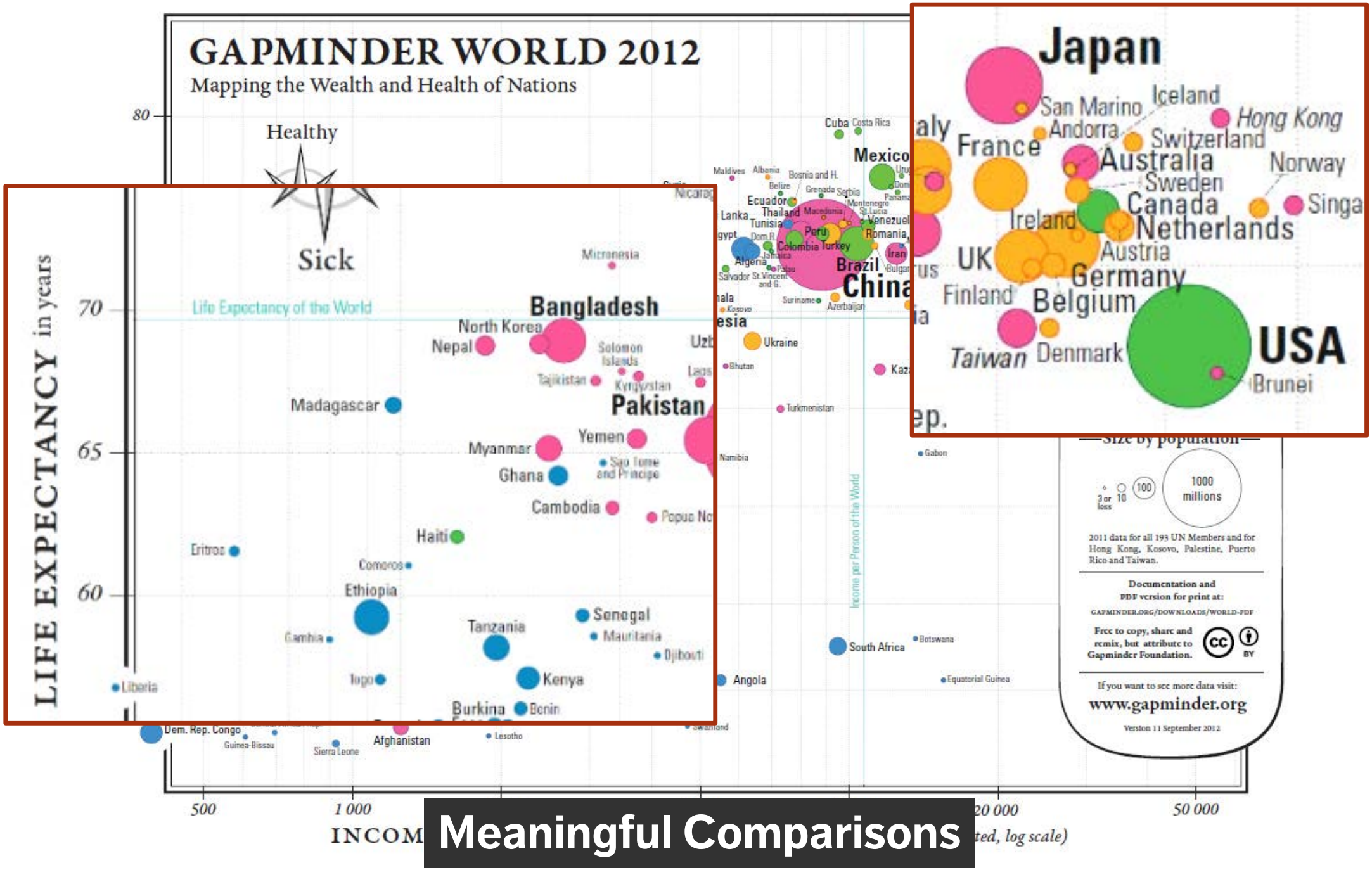


Non-Integrated Data

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]

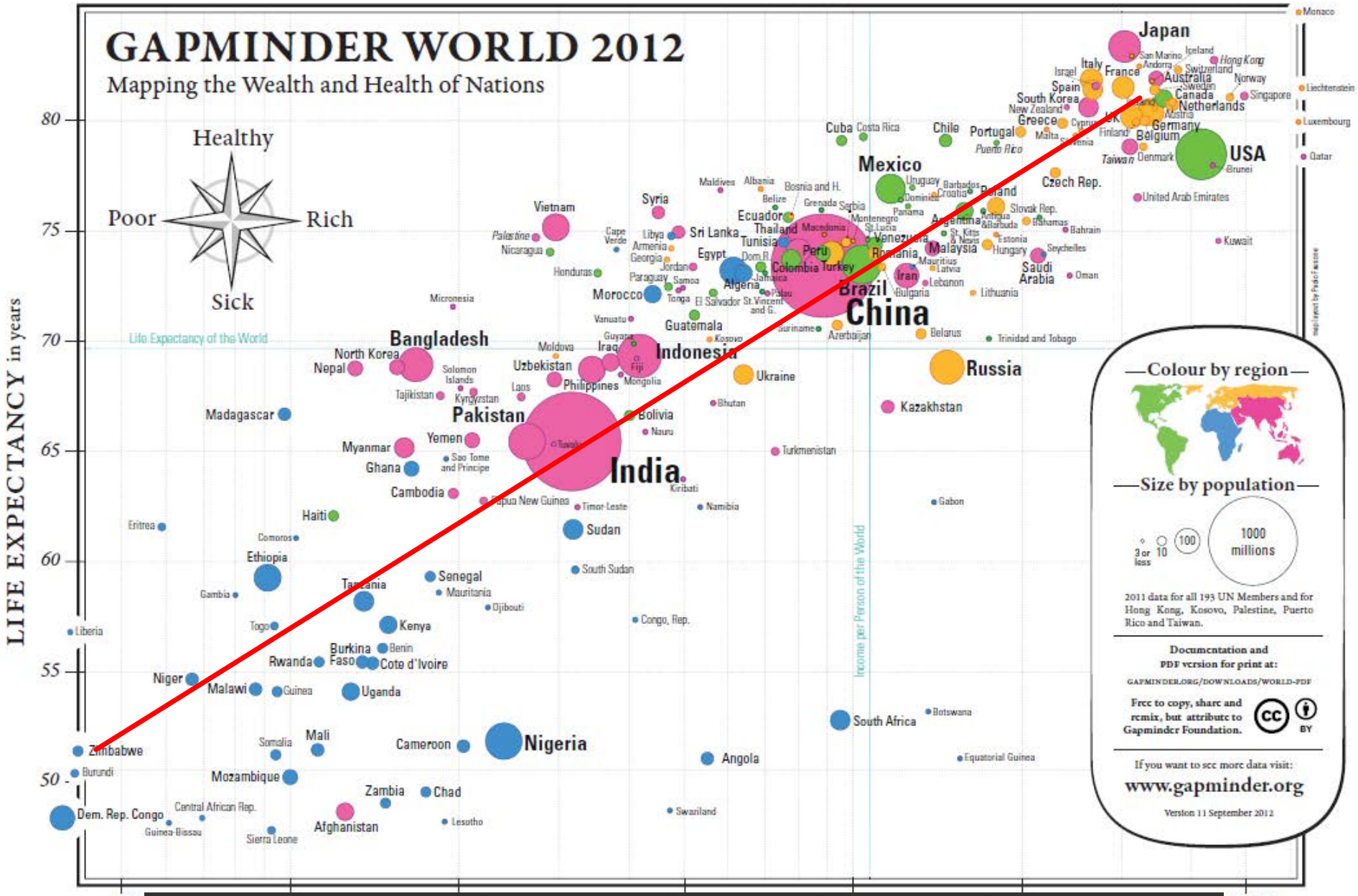


[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



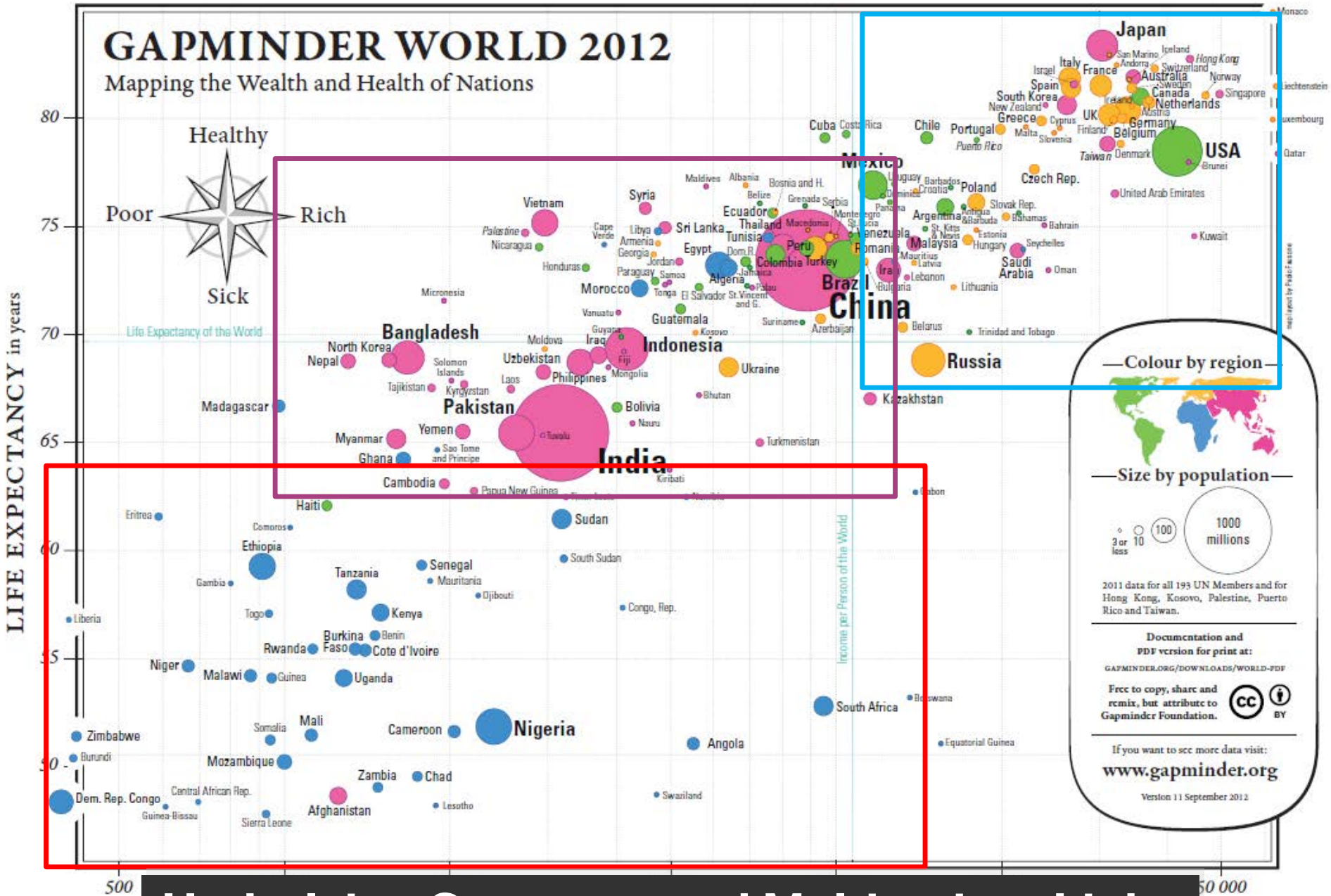
Meaningful Comparisons

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



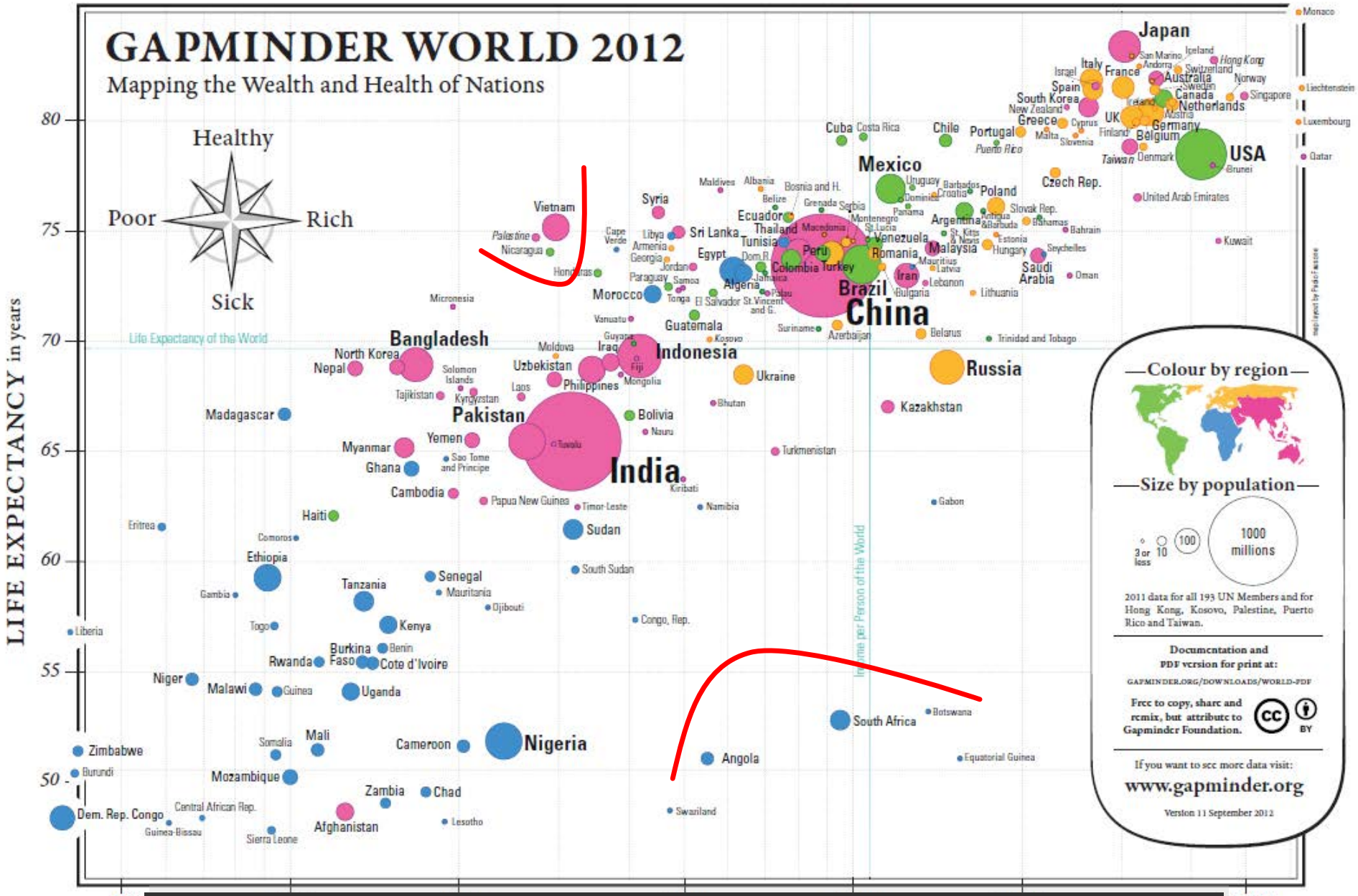
Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



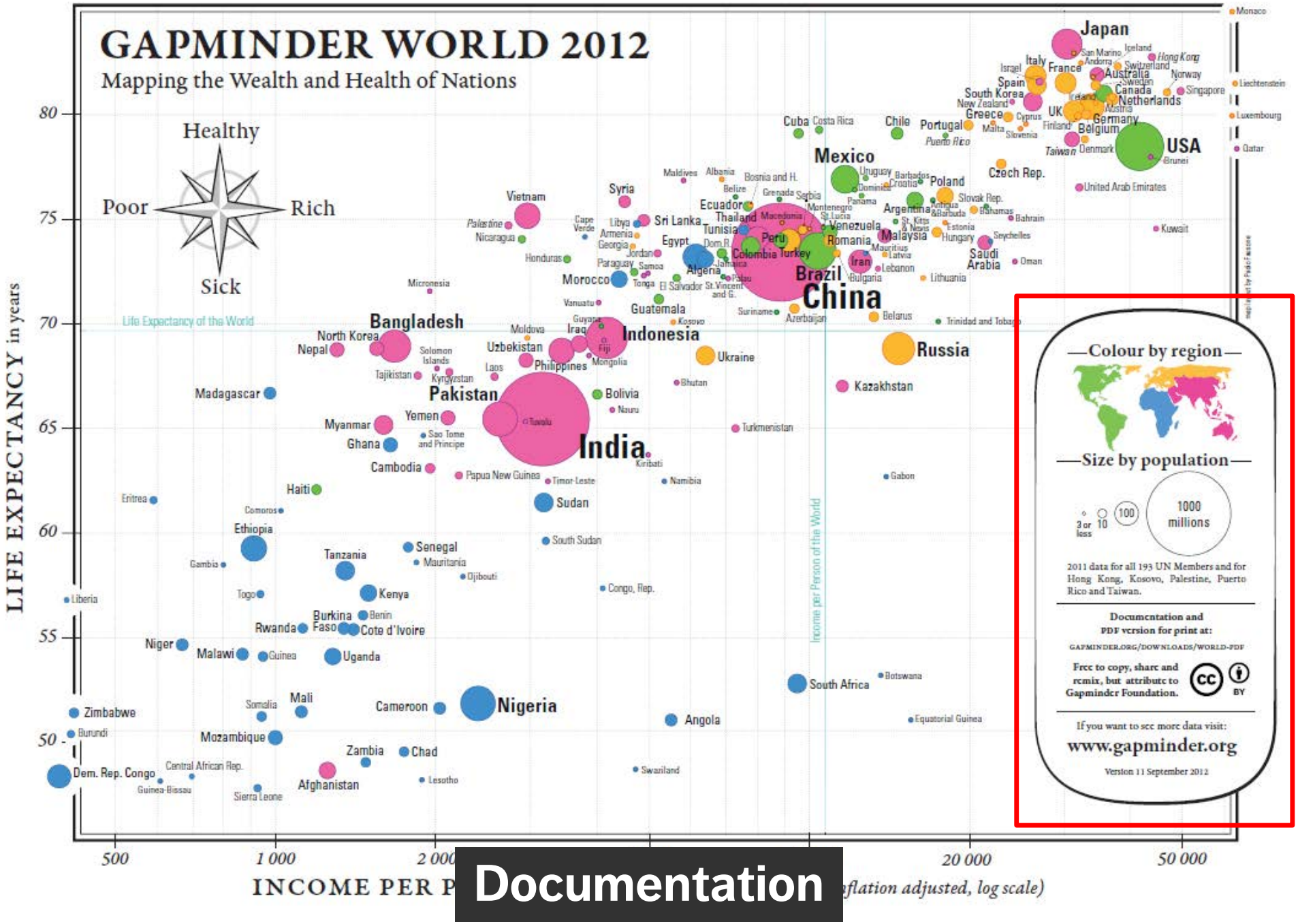
Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



Underlying Structure and Multivariate Links

[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



PRESENTING ANALYSIS RESULTS

Graphics should be clear and engaging.

Not every pretty picture tells a story, but if a story can't be told with pretty pictures, perhaps it's time to re-think the story…

Graphical representation techniques appear regularly – it's too early to tell which ones will stand the test of time.

Don't be afraid to try something new if it helps **convey the message**.

DESIGN ELEMENTS

Is the point getting across? Integrated data helps convey the message.

Not all **retinal variables** are equally effective when it comes to convey or represent information. Experiment as needed to find the optimal choice for the given context.

Adding design elements can enhance our understanding of the data.

How we spot patterns affect what we get out of data presentations.

Data displays are not just about picking a random visualization method. The result varies depending on the structure of the data and the (combinations of) questions.

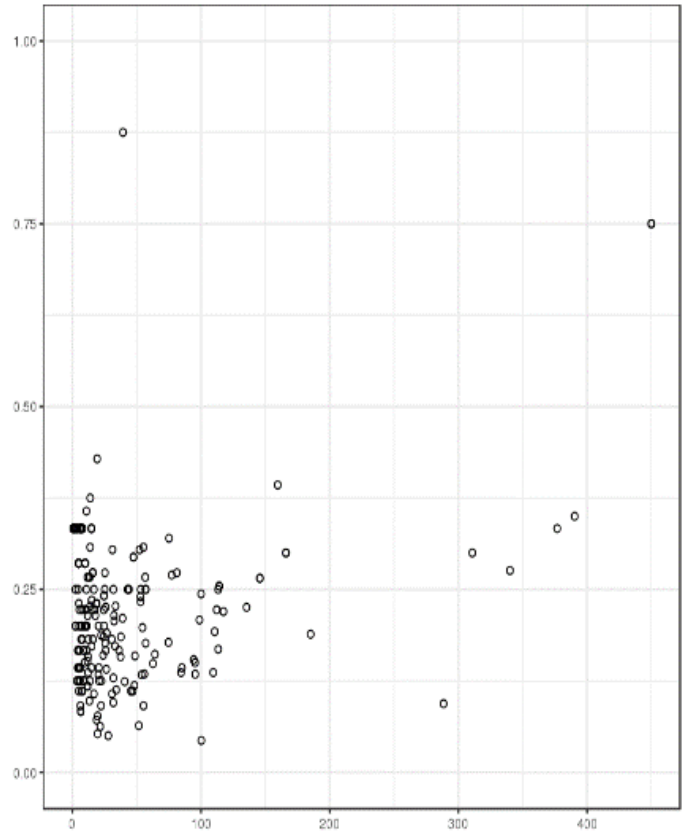
[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

REPRESENTING MULTIVARIATE OBSERVATIONS

2 variables can be represented by position in the plane. Additional factors can be depicted with:

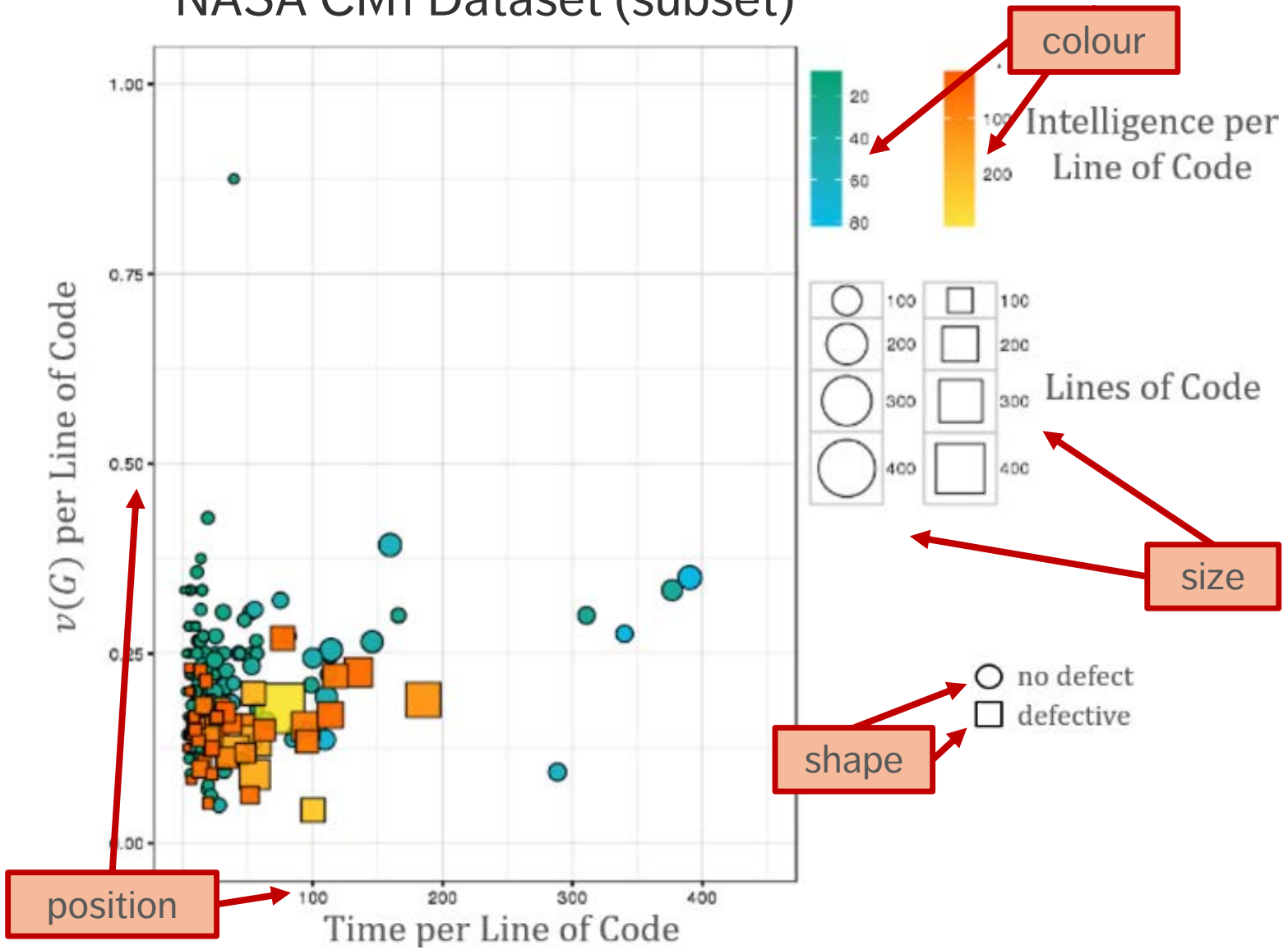
- size
- color
- value
- texture
- line orientation
- shape
- (motion?)

NASA CM1 Dataset (subset)



[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

NASA CM1 Dataset (subset)



DISCUSSION

What are some data visualization rules you have picked up over the years?

Are any of them hard rules?

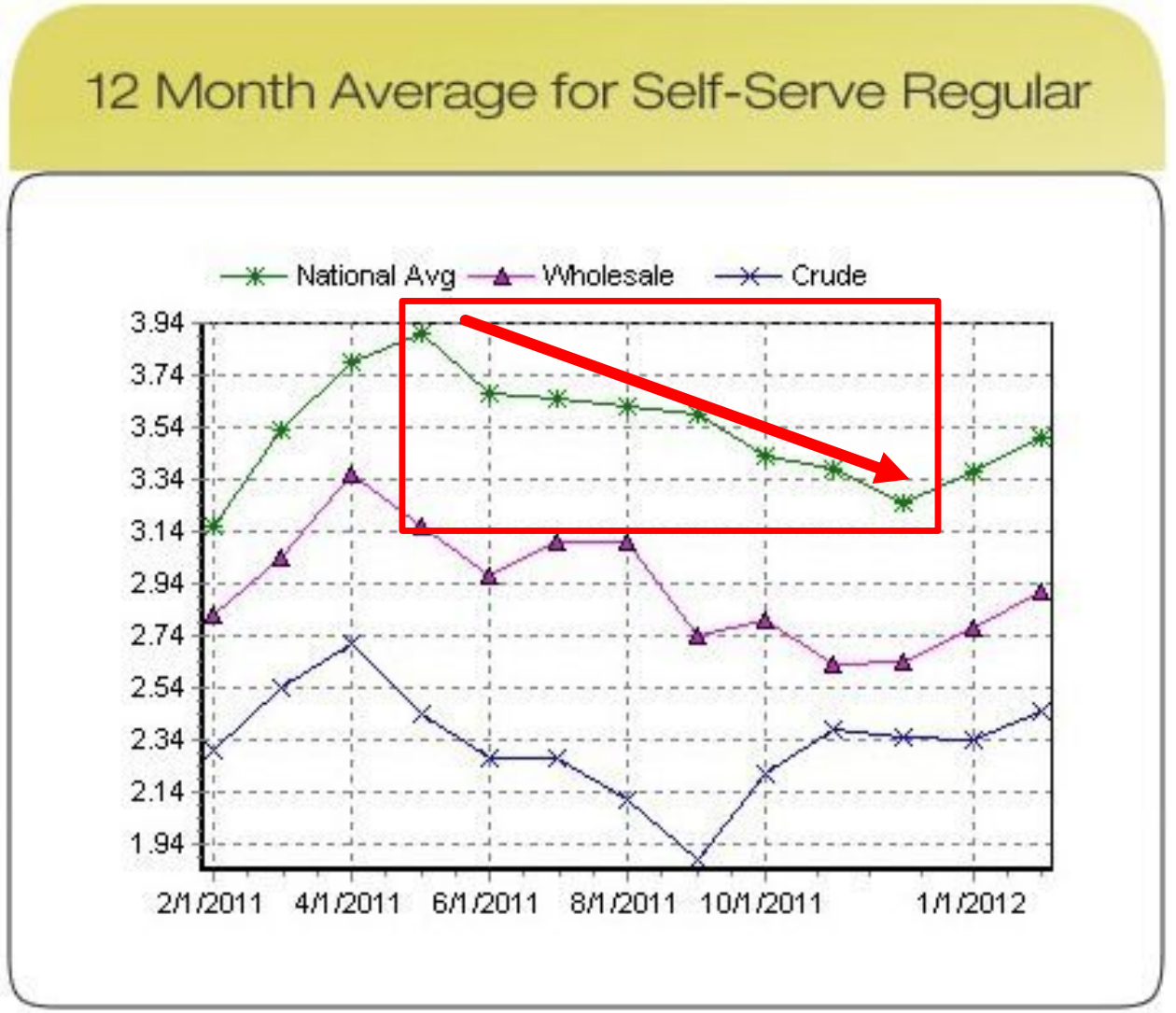
Which of the rules/principles shown here seem more fundamental?

HALL-OF-FAME / HALL-OF-SHAME

[Fox News, February 20, 2012, mediamatters.org, simplystatistics.org, badgraphs.tumblr.com]

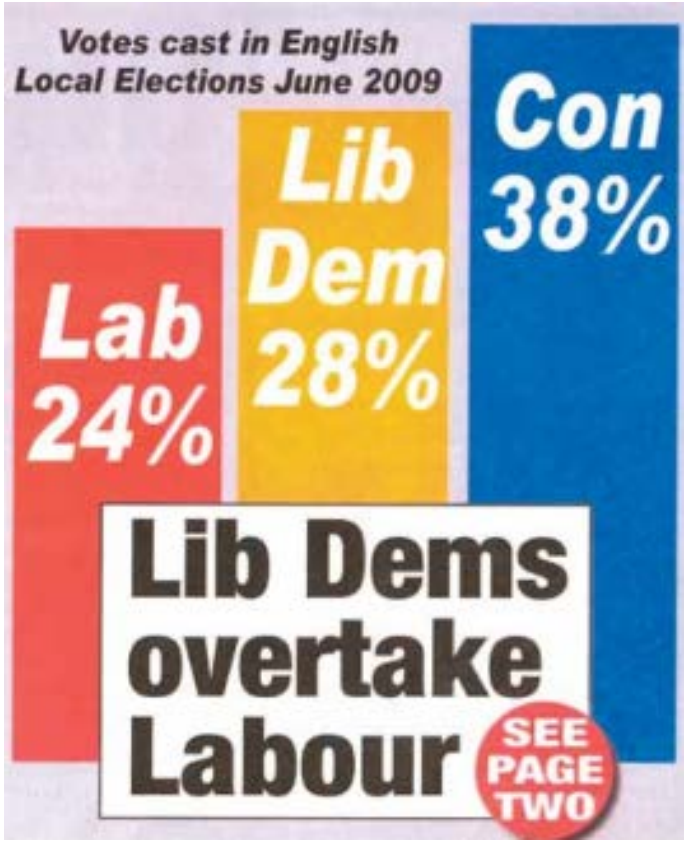
MISLEADING CHARTS

[AAA Gas Prices, found on mediamatters.org]



[Norwich North Electoral District, mediamatters.org, simplystatistics.org, badgraphs.tumblr.com]

MISLEADING CHARTS



MISLEADING CHARTS

Problems: disingenuous, selective and/or incompetent reporting

Solutions:

- Consistent scales and units of comparison
- Full time series
- No cherry picking the data range
- Cutting off -axis will exaggerate some effects
- Numbers must add up

WHAT TO WATCH FOR

Some methods yield visually striking, yet misleading, charts.

Be on the lookout for:

- tampering with axes and linear scales
- scaling effects, when representing data points as shapes or volumes
- cherry-picking by omitting certain data points

For low-dimensional datasets, a **tabular display** may provide as much information and be less likely to mislead.

YOU BE THE JUDGE

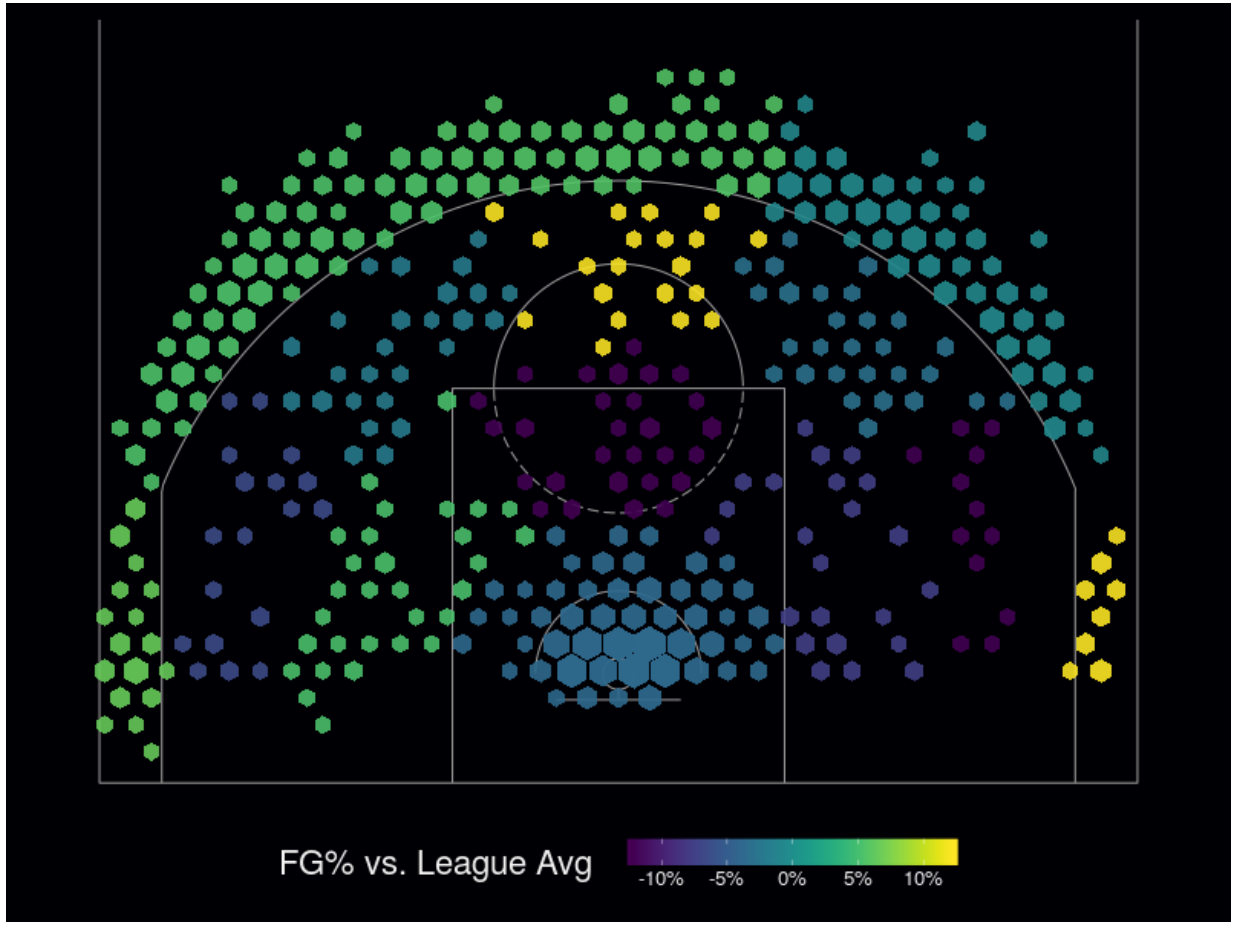
Some of the following are (arguably) good visualizations. But some are not!

Which are which? You be the judge...

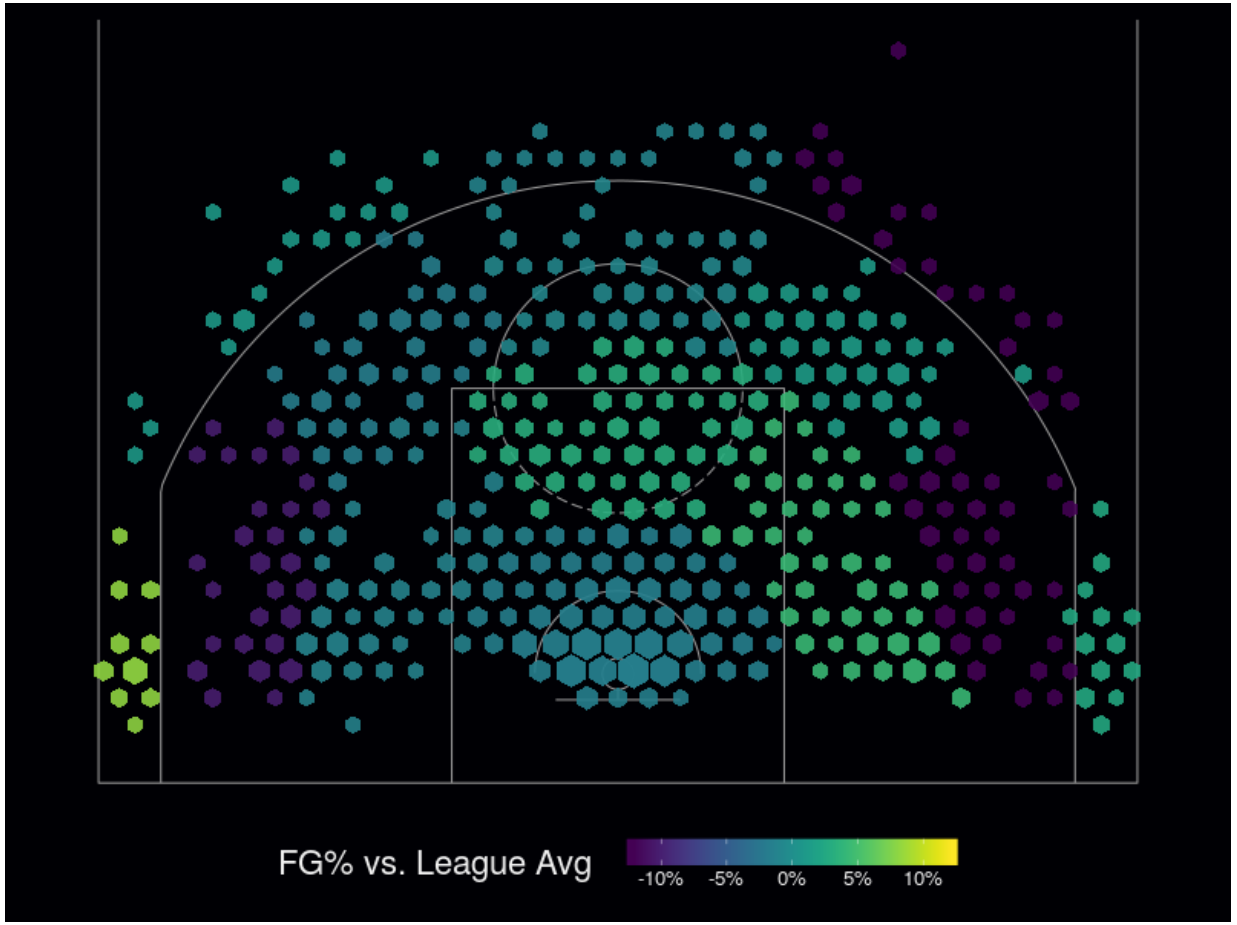
[T.W. Schneider]

NBA FG% Against League Average ('15–'16)

Kyle Lowry



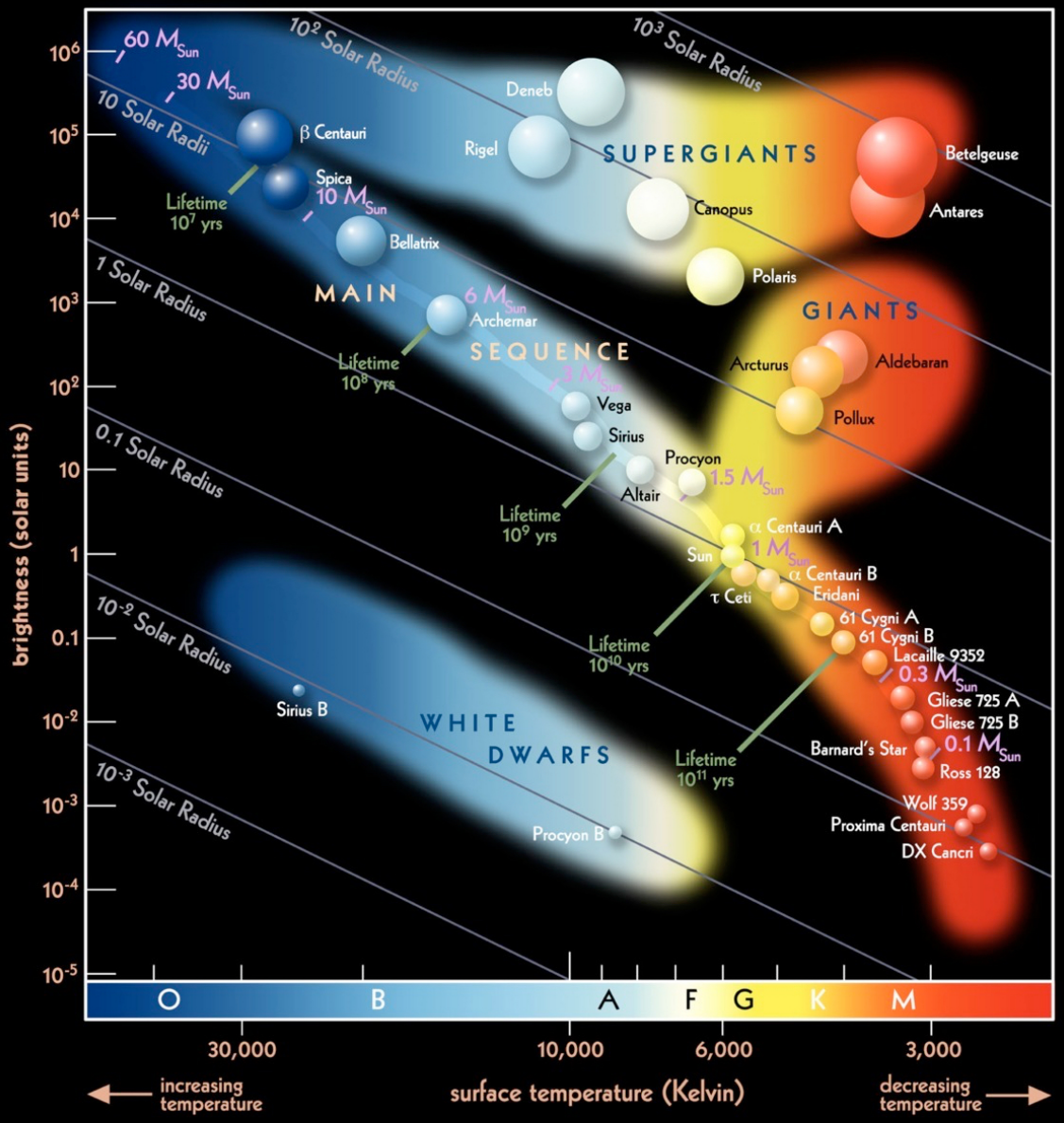
DeMar DeRozan



What comparisons can you make? Do you understand the encoding? The context?

[European Southern Observatory]

Hertzprung-Russell Diagram



Data Elements

- star radius (x 2)
- surface temperature (x 2)
- spectral class
- brightness
- mass
- lifetime
- name

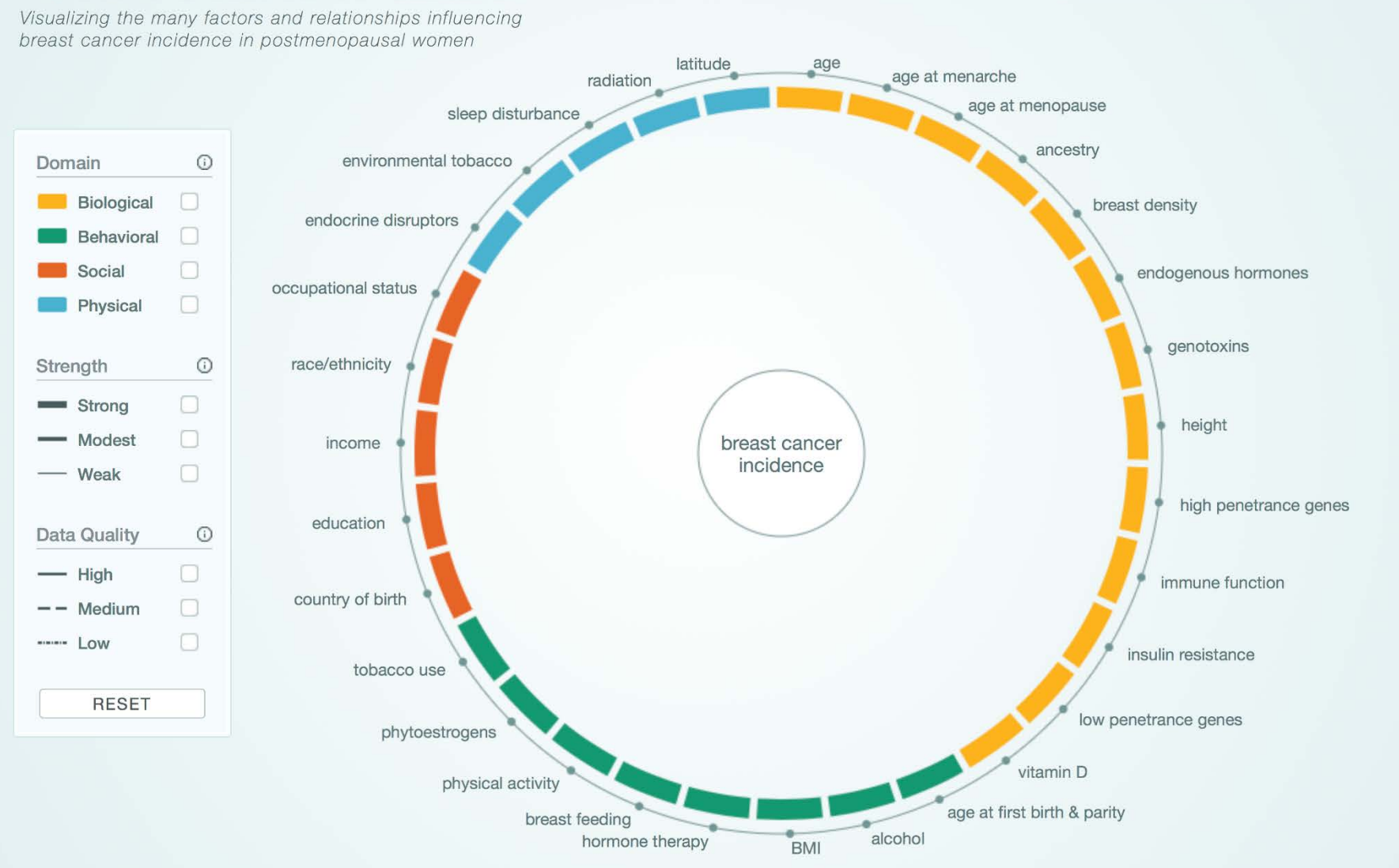
Underlying Structure

- 4 clusters/group
- lifetime, mass and radius are related to brightness and surface temperature on the Main Sequence

Only a subset of all the stars is shown in the HR diagram.

[http://www.cbcrp.org/causes]

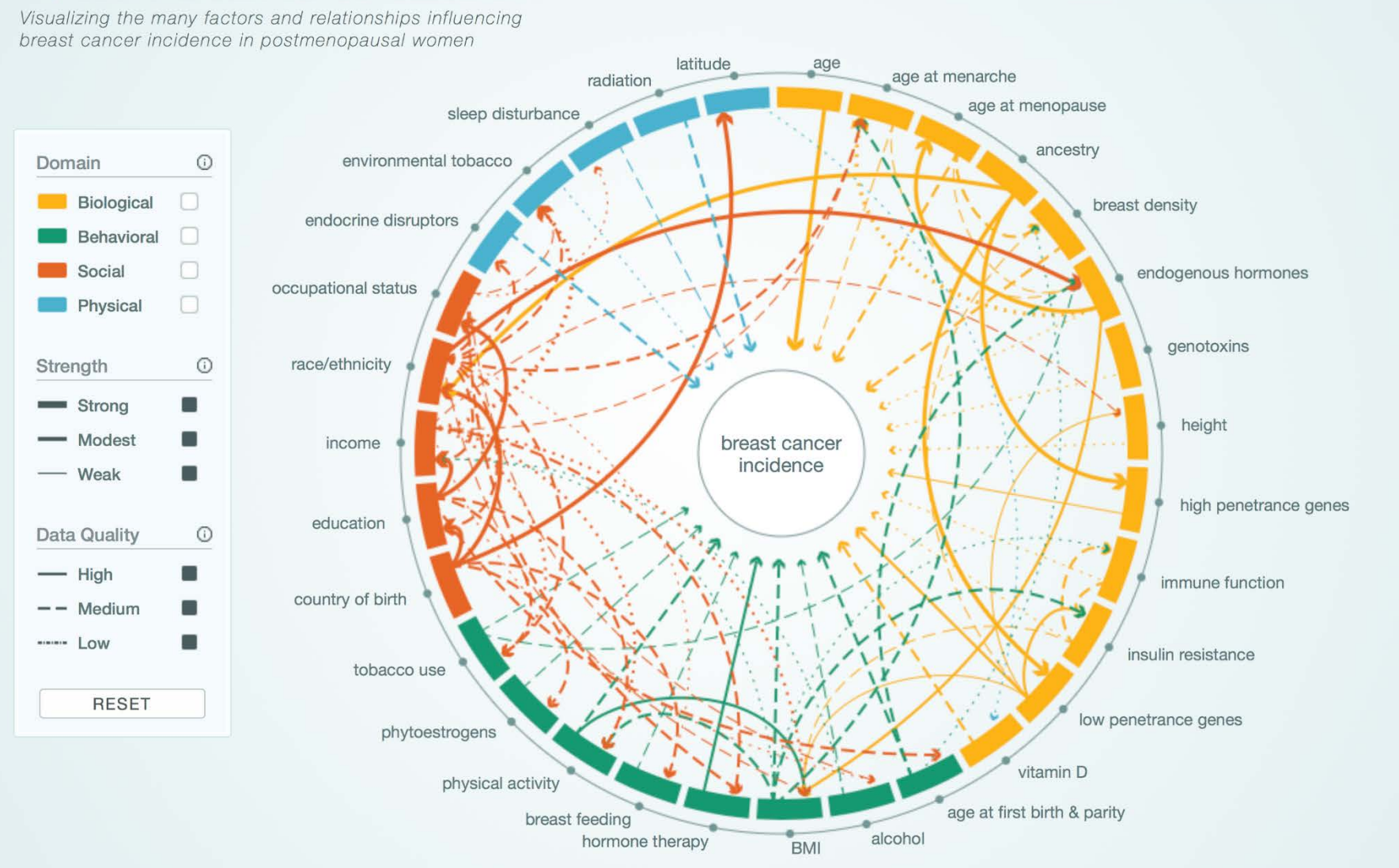
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

[http://www.cbcrp.org/causes]

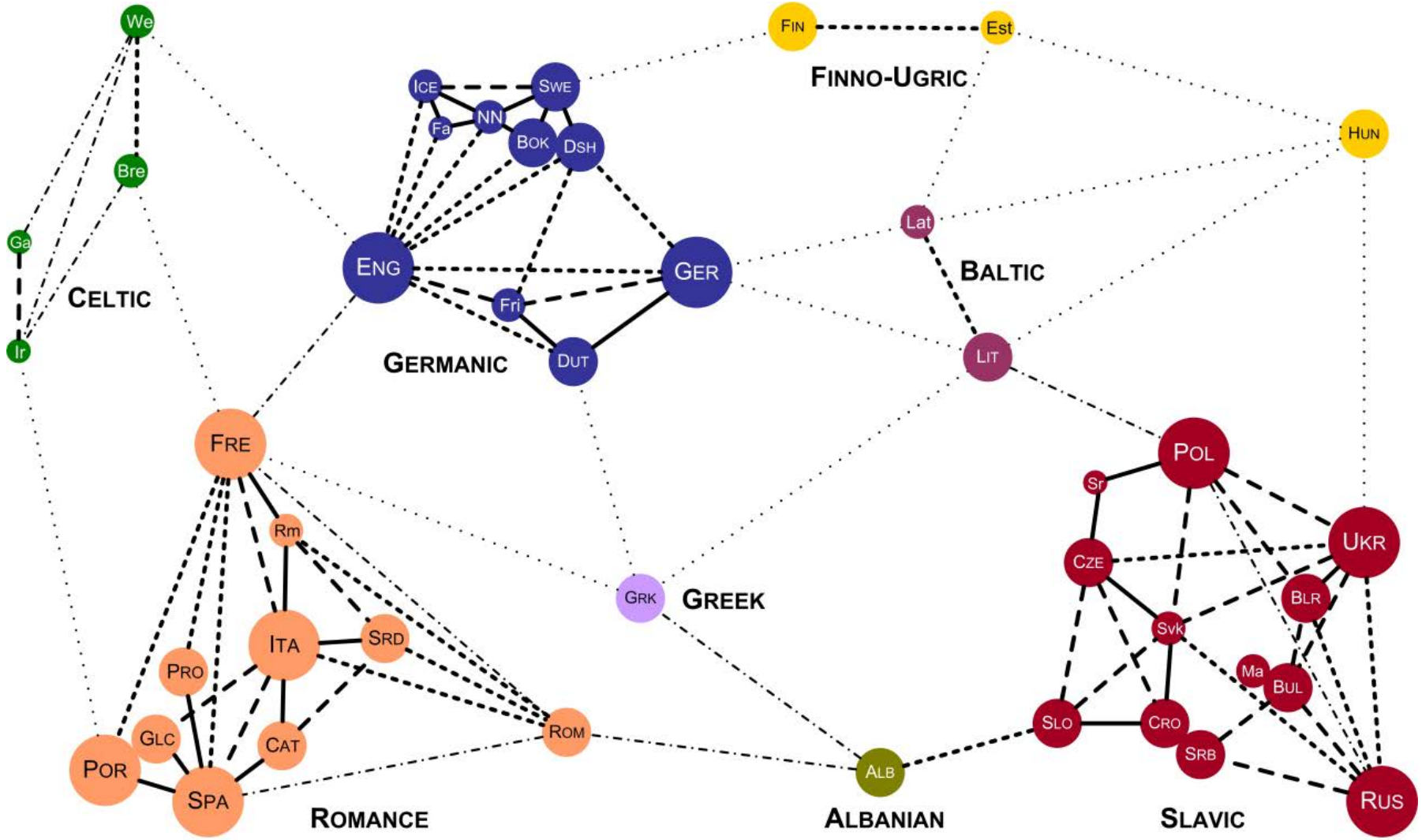
A Model of Breast Cancer Causation



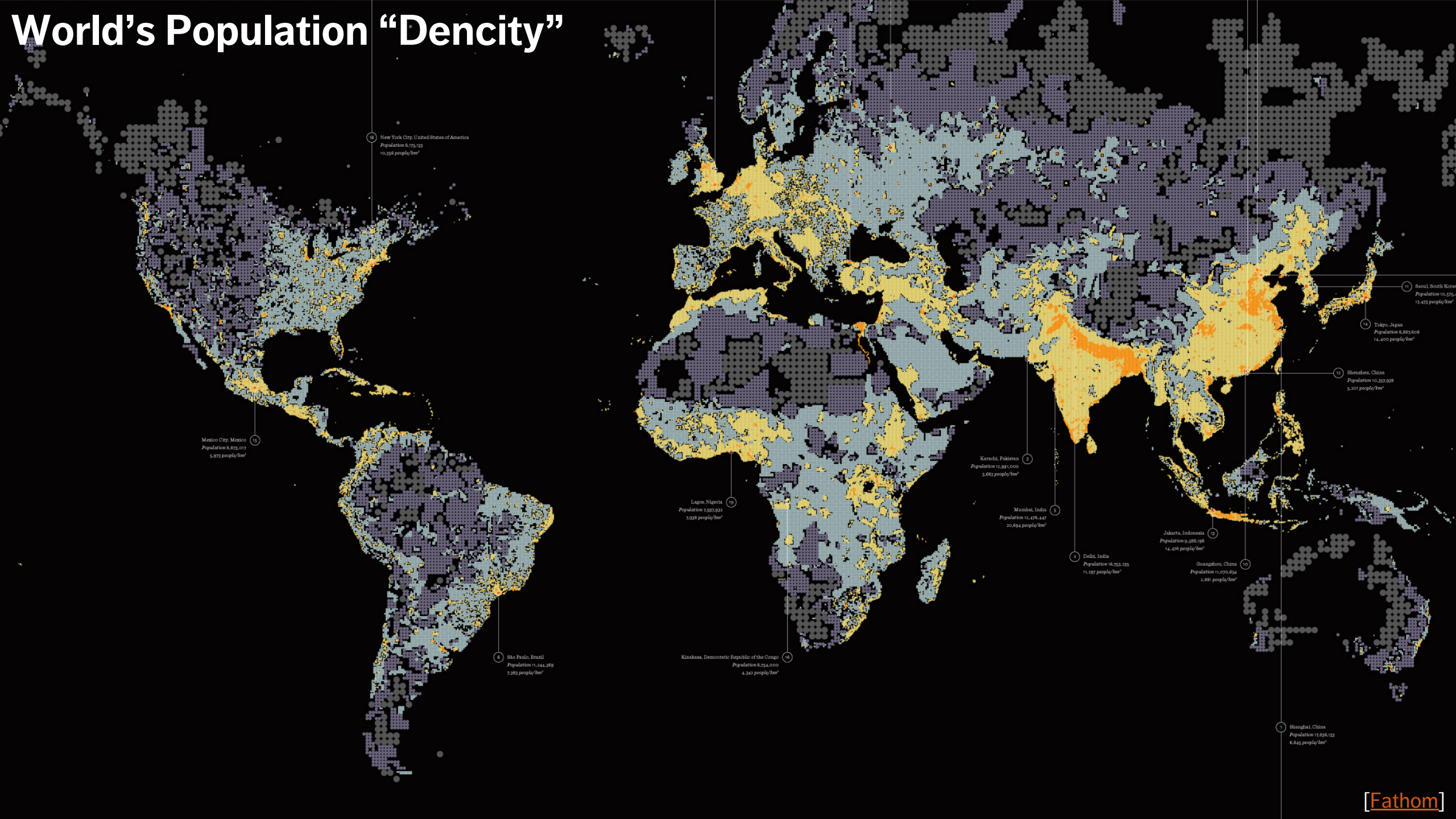
Can you infer causality from this diagram?

Lexical Distance Among the Languages of Europe

[T.Elms, Etymologikon]



World's Population "Dencity"



[Fathom]

[https://thinkprogress.org, found on http://viz.wtf]

[https://thinkprogress.org, found on http://viz.wtf]

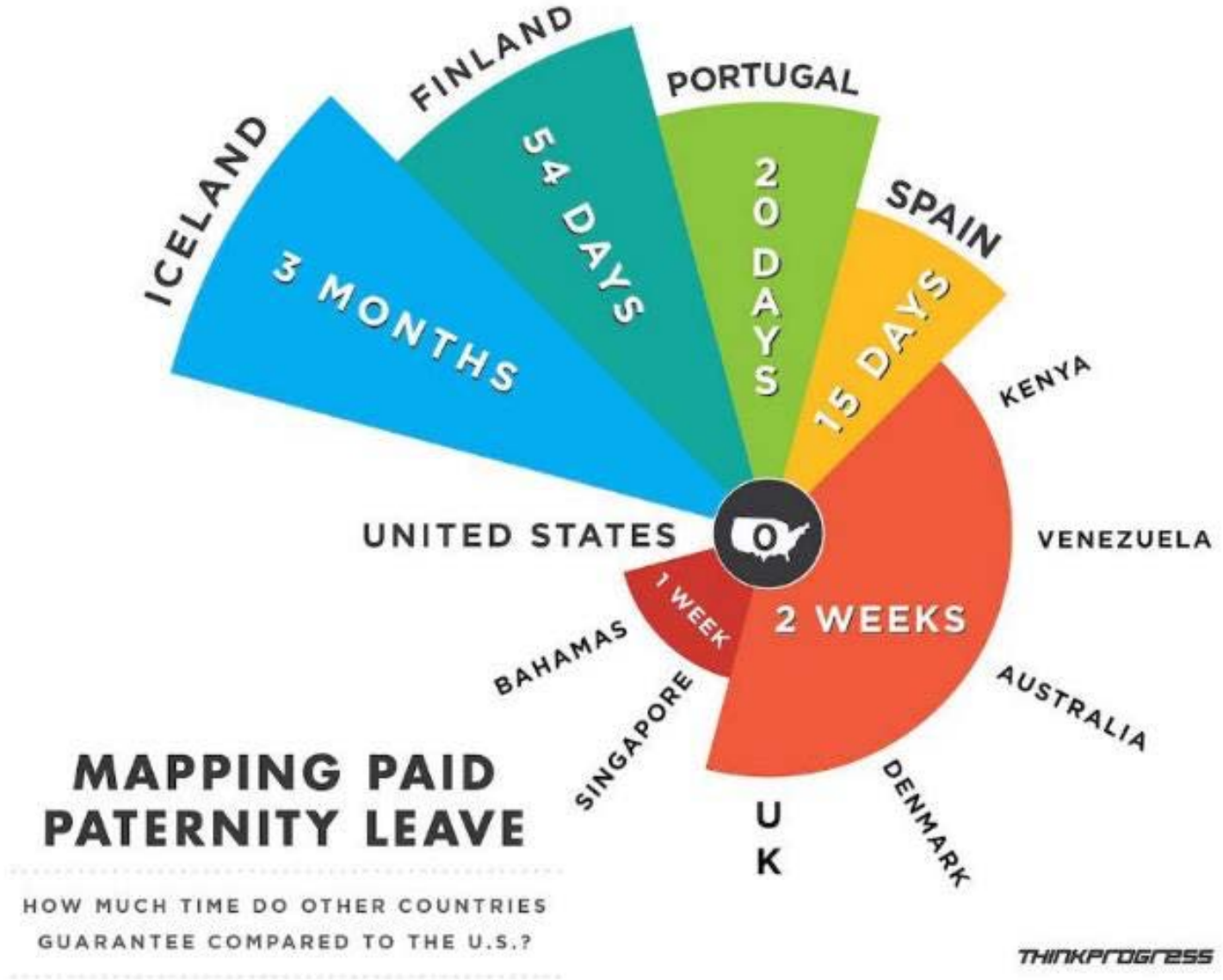
Low data density

High chartjunk ratio

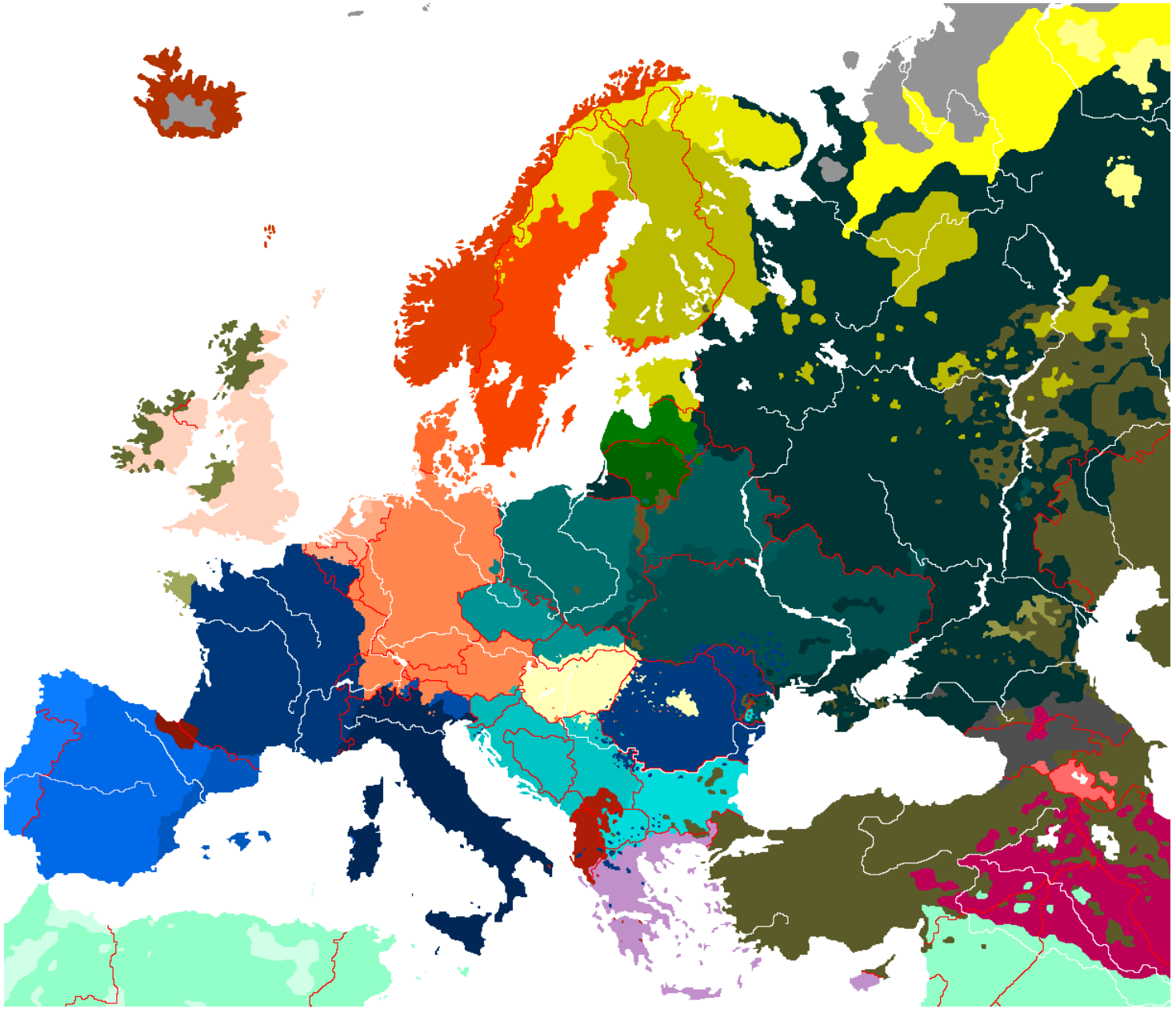
Scaling effects

Cherry-picking

Why not use a bar chart or a tabular display instead?



[Found on Wikipedia, author unknown]



[Found on Wikipedia, author unknown]

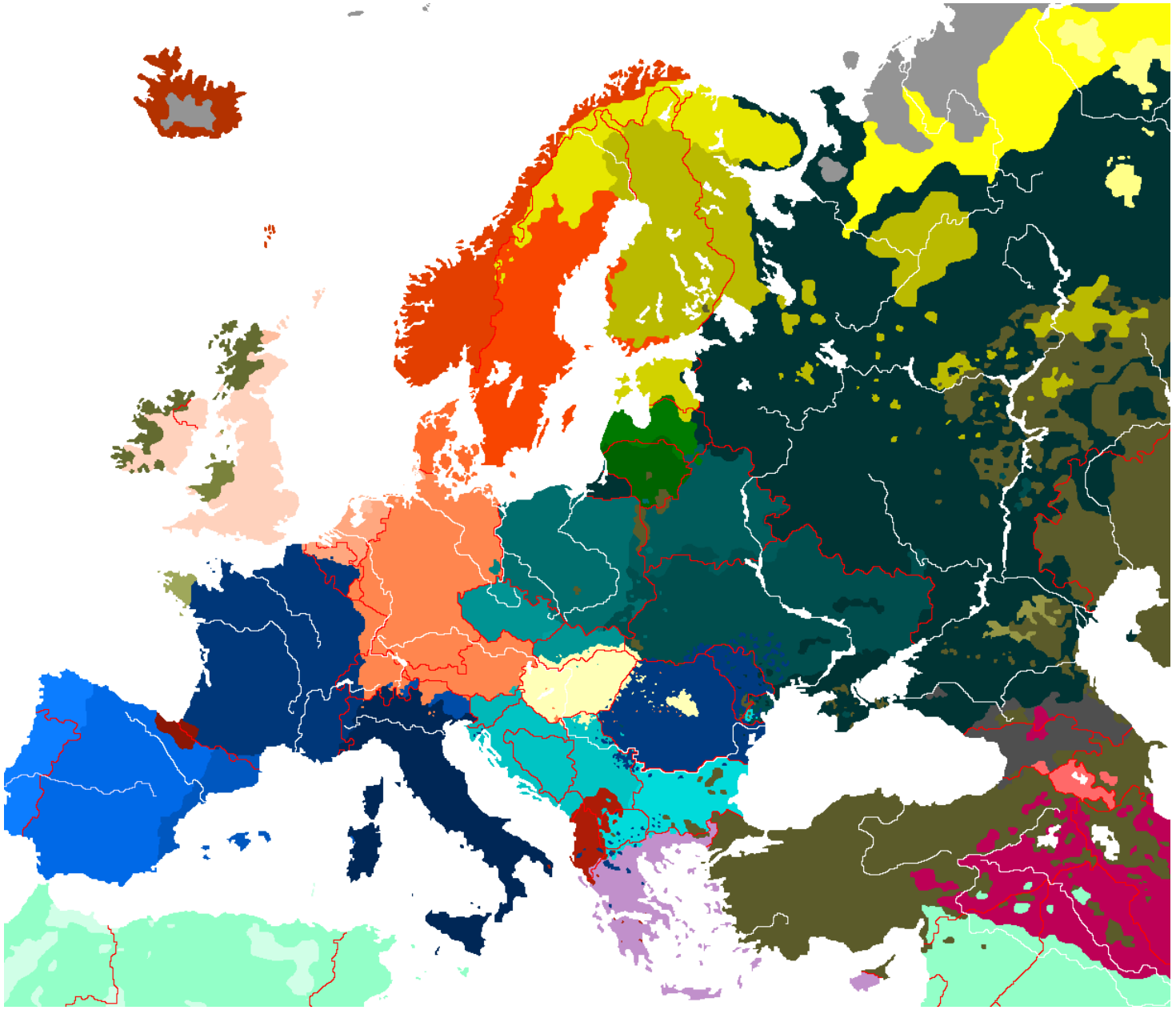
Encoding?

Population density?

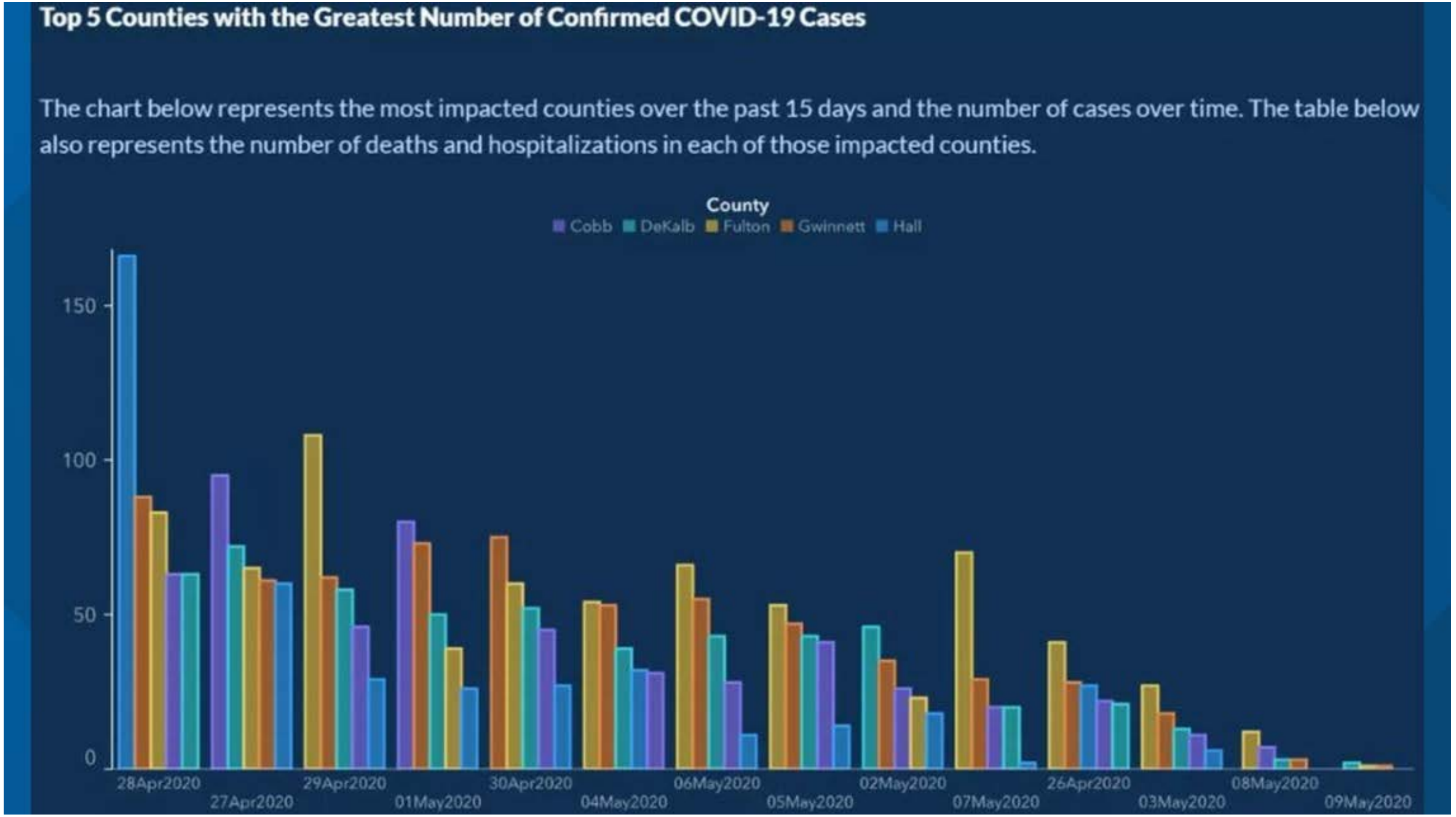
Secondary languages?

Rivers?

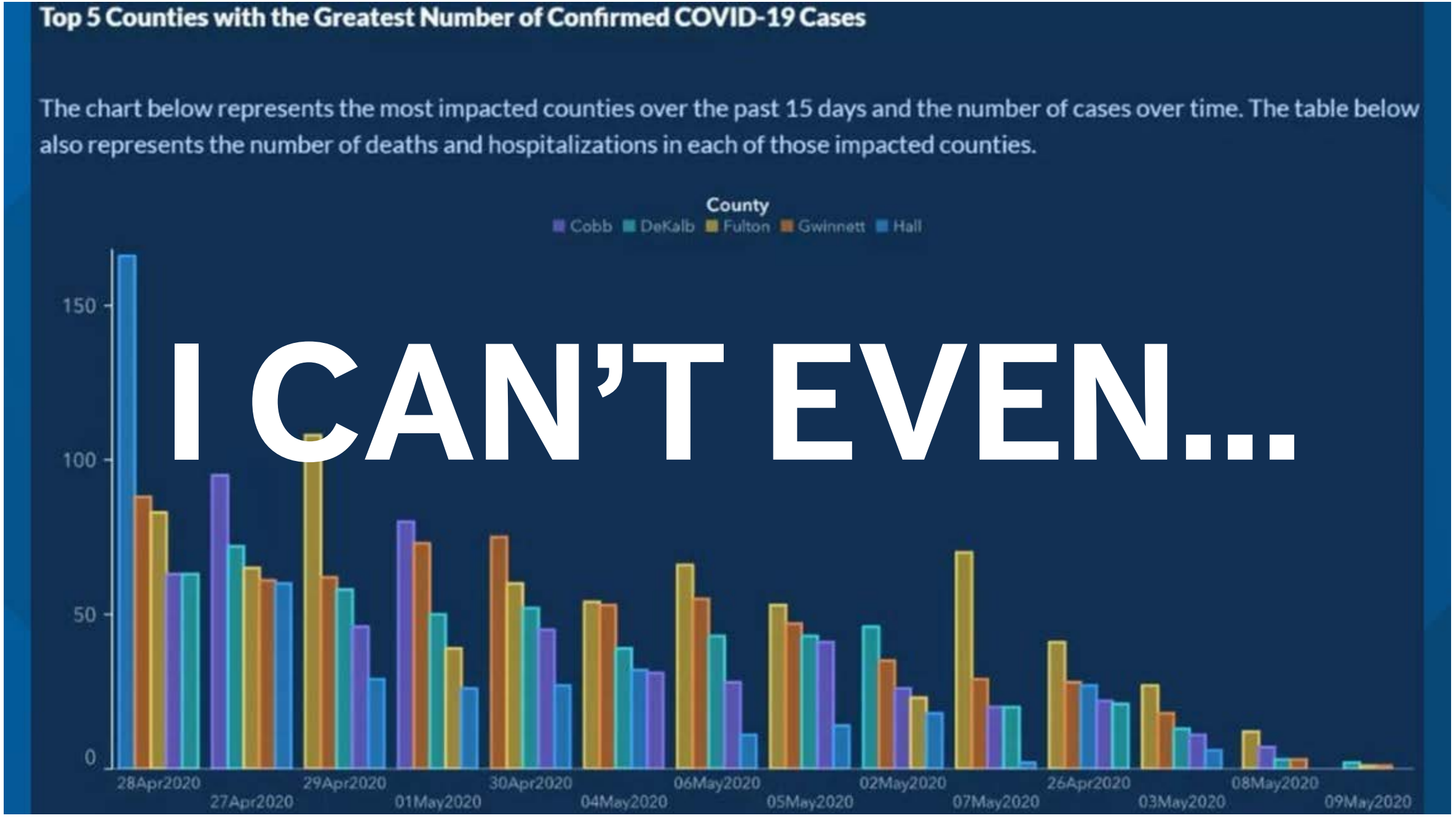
No data source



[Georgia Department of Public Health]



[Georgia Department of Public Health]



TAKE-AWAYS

Effective data visualizations **provide insights** and **facilitate understanding**.

The basic principles can guide your visualization design and consumption.

Be **creative**, but keep your data and your representations **honest**.

Be mindful of attempts to distort trends and conclusions with flashy visuals.

Data and code should be made available along with the displays.