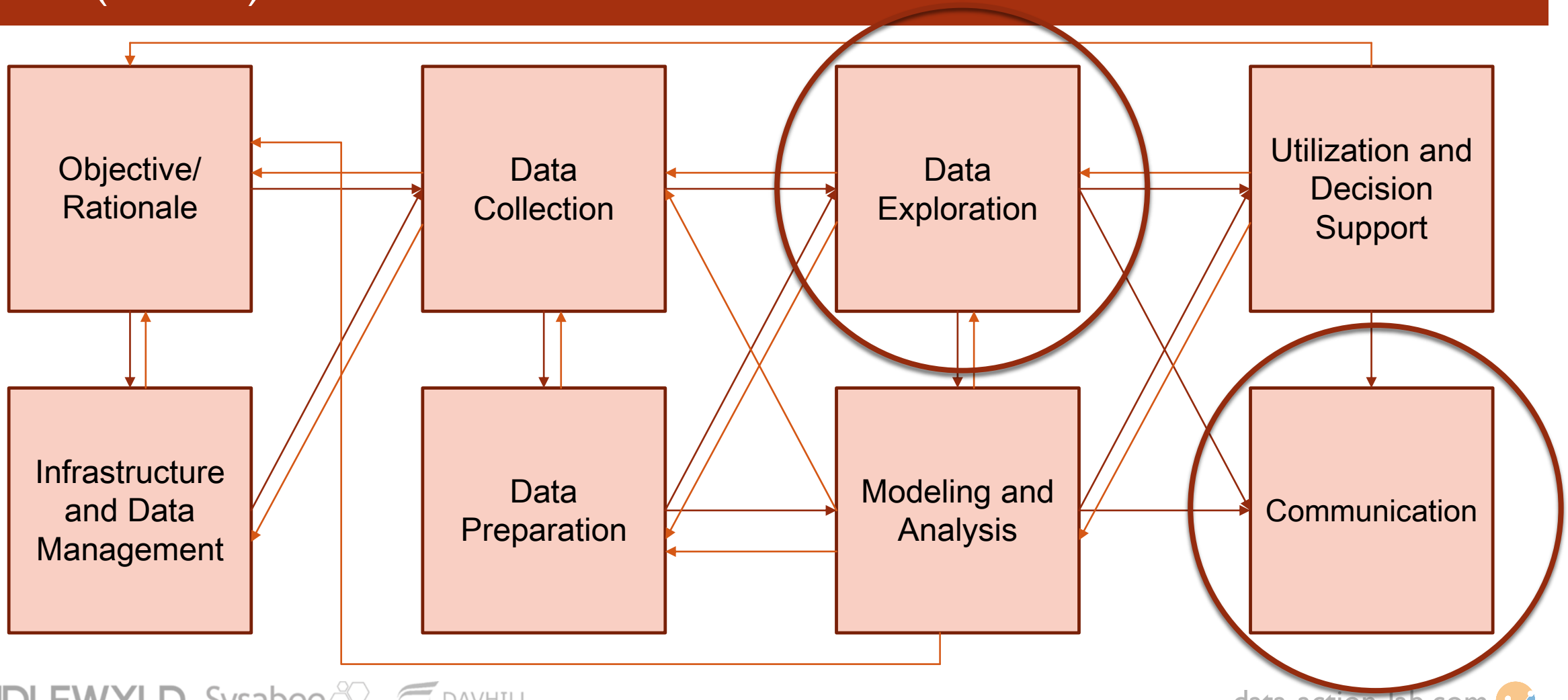
DATA EXPLORATION AND DATA VISUALIZATION

ADVANCED DATA SCIENCE TRAINING I

“Discovery is no longer limited by the collection and processing of data, but rather management, analysis, and visualization.”

@DamianMingle

THE (MESSY) ANALYSIS PROCESS



OUTLINE

1. Data Exploration
2. Pre-Analysis Data Visualization
3. Post-Analysis Data Visualization
4. Visualization Catalogue
5. Hall-of-Fame / Hall-of-Shame

LEARNING OBJECTIVES

Understand the different roles of data visualization in the data analysis process.

Increase your understanding of how to represent simultaneously multiple dimensions.

Improve your ability to judge how many dimensions are being represented in a chart.

Understand some of the strategies and considerations for creating good post-analysis visualizations.

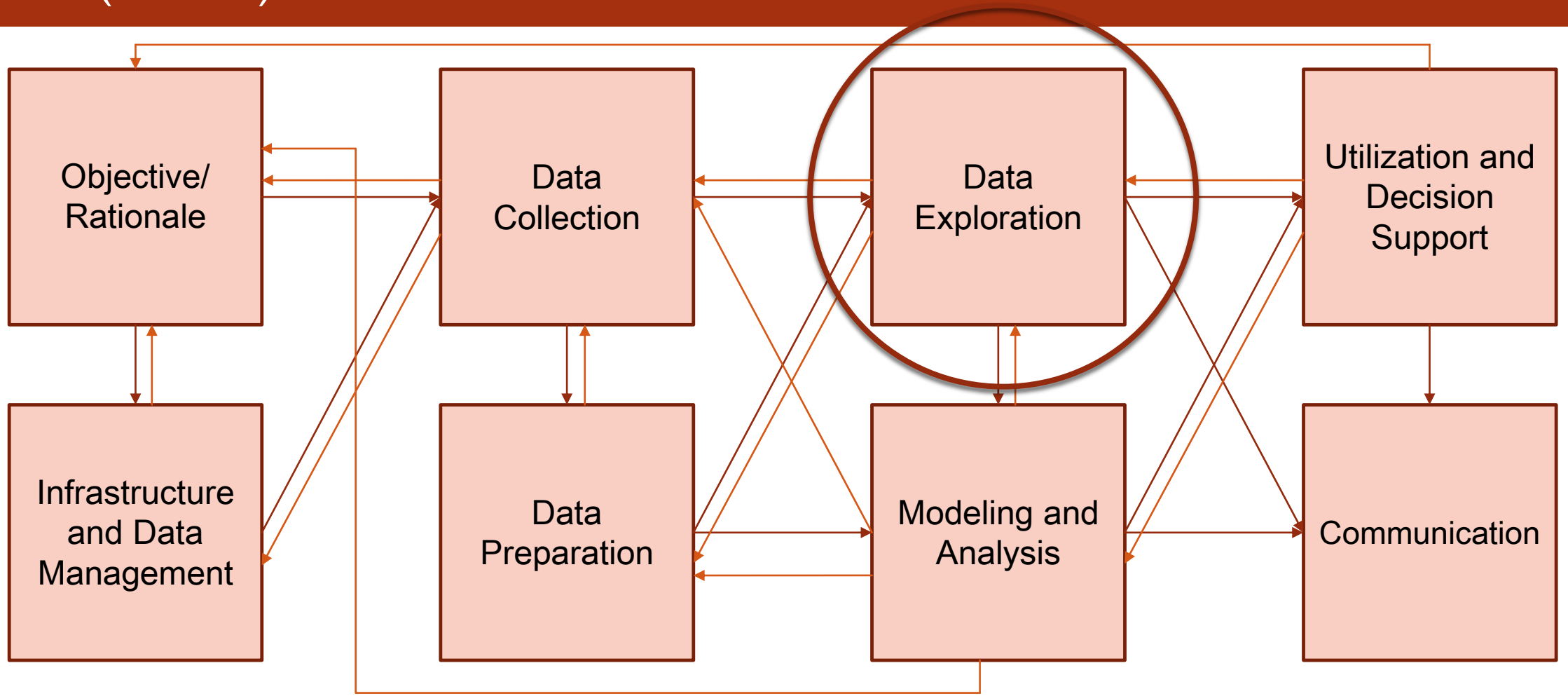
Understand the difference between a visualization and an info graphic.

Improve your judgment about the quality of data visualizations.

DATA EXPLORATION

DATA EXPLORATION AND DATA VISUALIZATION

THE (MESSY) ANALYSIS PROCESS



GETTING TO KNOW YOUR DATA

Hi dataset, I'm Morgan!

Hi Morgan, I'm dataset,
nice to meet you!

SOME BASIC QUESTIONS

What system does your data represent – objects, attributes, relationships?

How does it represent this system – i.e. the data model?

Who made this dataset? When? For what purpose?

Assuming a flat file – what do the rows represent? What do the columns represent?

Do you even have enough information (e.g. **metadata**) to answer these questions? Where can you find more information?

NON-VISUALIZATION BASED SUMMARIES OF YOUR DATASET

```
       Cl                NO3                NH4
 Min.   :  0.222   Min.   : 0.000   Min.   :    5.00
 1st Qu.: 10.994   1st Qu.: 1.147   1st Qu.:   37.86
 Median : 32.470   Median : 2.356   Median :  107.36
 Mean   : 42.517   Mean   : 3.121   Mean   :  471.73
 3rd Qu.: 57.750   3rd Qu.: 4.147   3rd Qu.:  244.90
 Max.   :391.500   Max.   :45.650   Max.   :24064.00
 NA's   :16        NA's   :2        NA's   :2

    season
 Length:340
 Class :character      autumn spring summer winter
 Mode  :character          80     84     86     90
```

PRE-ANALYSIS DATA VISUALIZATION

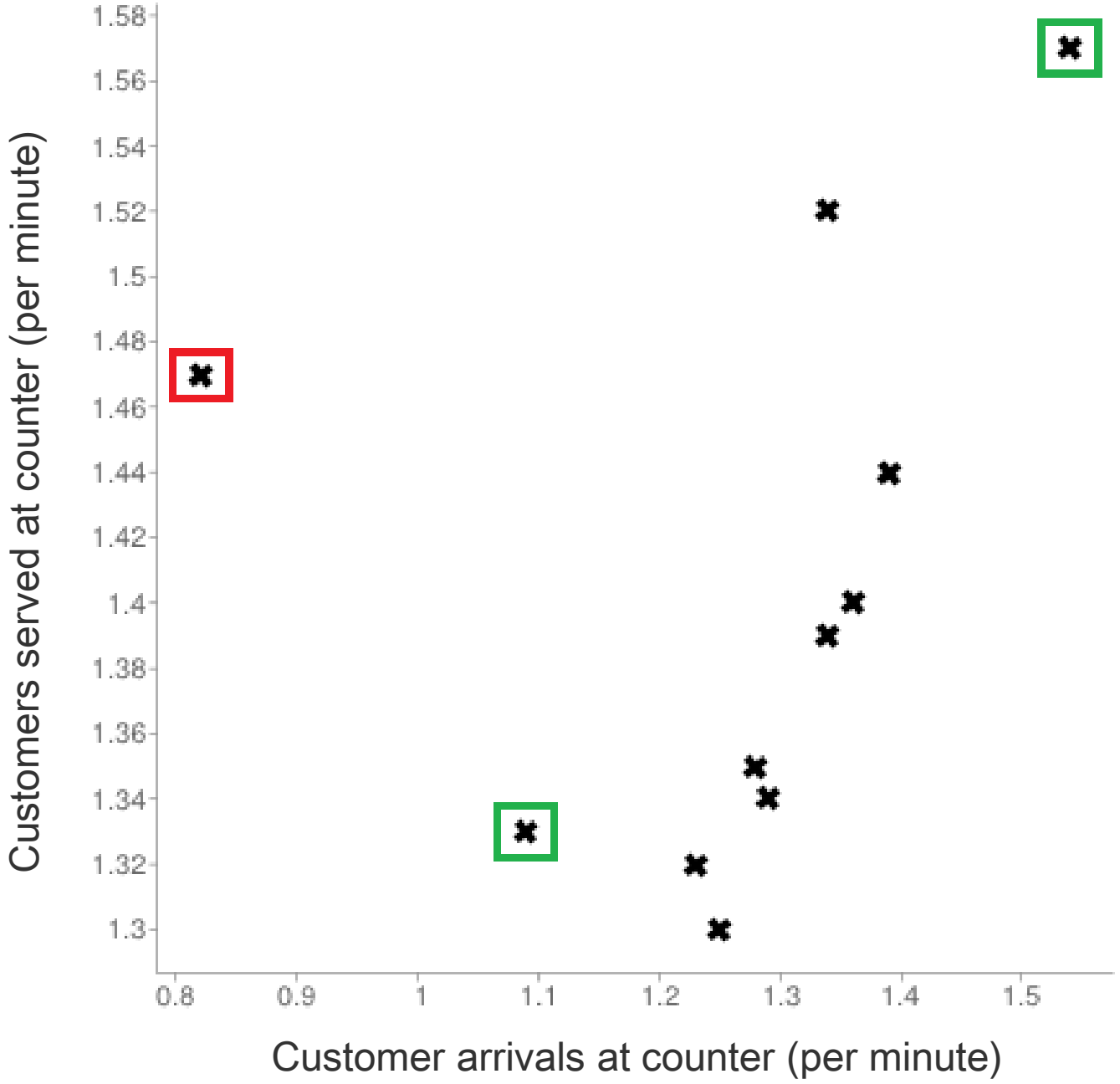
DATA EXPLORATION AND DATA VISUALIZATION

PRE-ANALYSIS USE

Data visualization can be used to set the stage for analysis:

- **detecting anomalous entries**
  invalid entries, missing values, outliers
- **shaping the data transformations**
  binning, standardization, Box-Cox transformations, PCA-like transformations
- **getting a sense for the data**
  data analysis as an art form, exploratory analysis
- **identifying hidden data structure**
  clustering, associations, patterns informing the next stage of analysis

[Personal dataset]

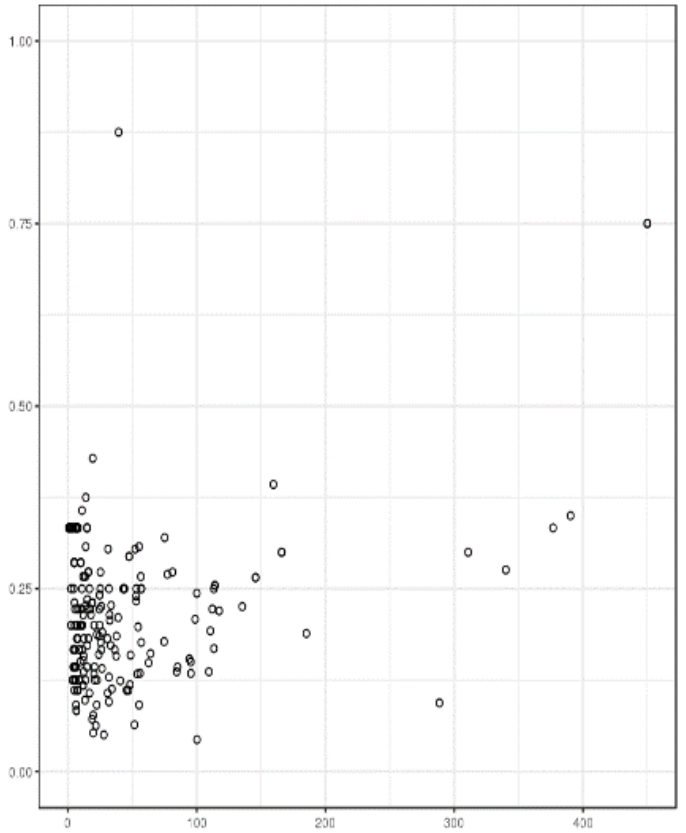


[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

REPRESENTING MULTIVARIATE OBSERVATIONS

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

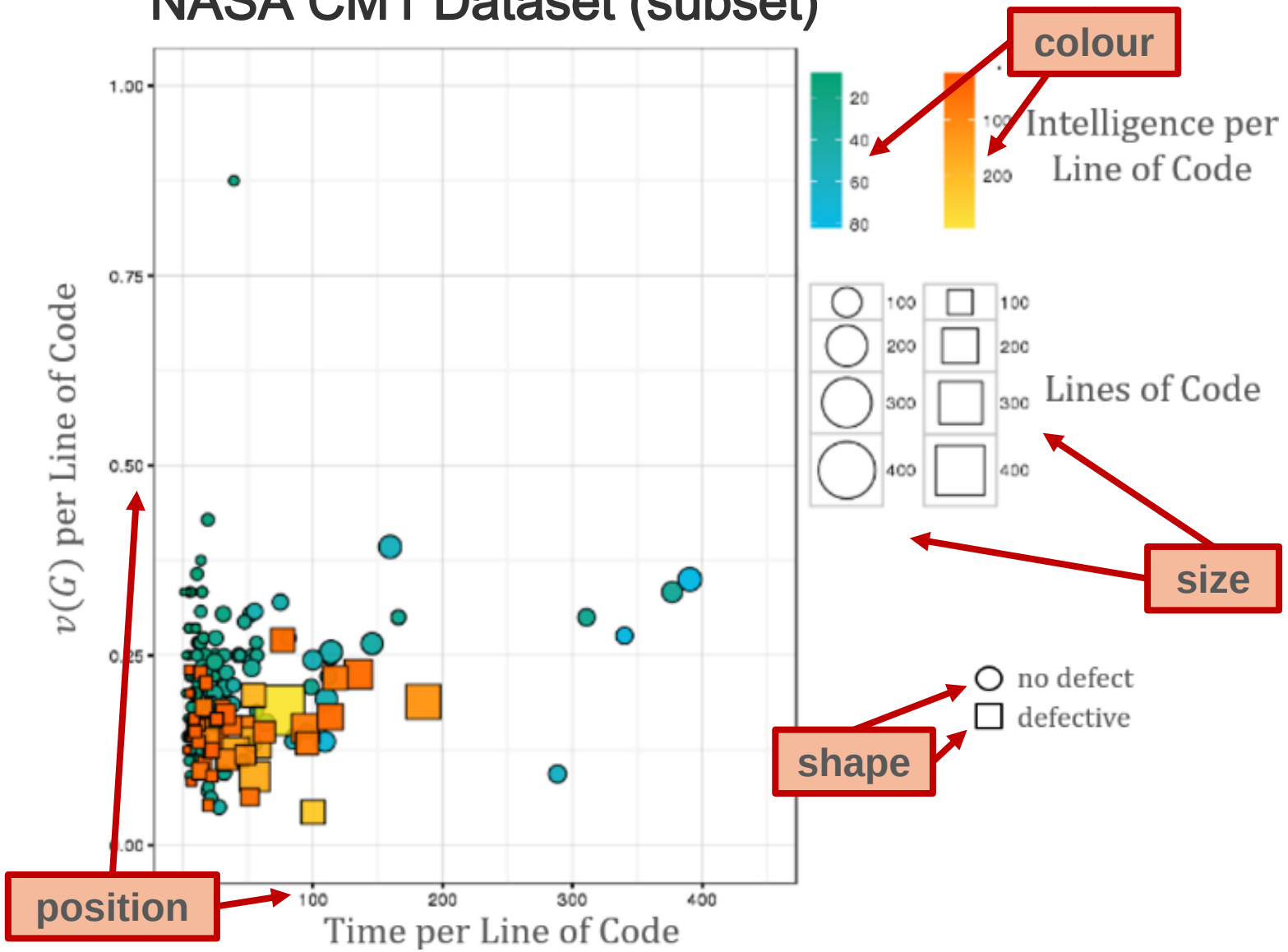
- size
- color
- value
- texture
- line orientation
- shape
- (motion?)



NASA CM1 Dataset (subset)

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

NASA CM1 Dataset (subset)



WORKHORSE DATA EXPLORATION VISUALIZATIONS

Line Chart/Rug Chart/Number Line

Histogram

(Boxplots)

Line Graph

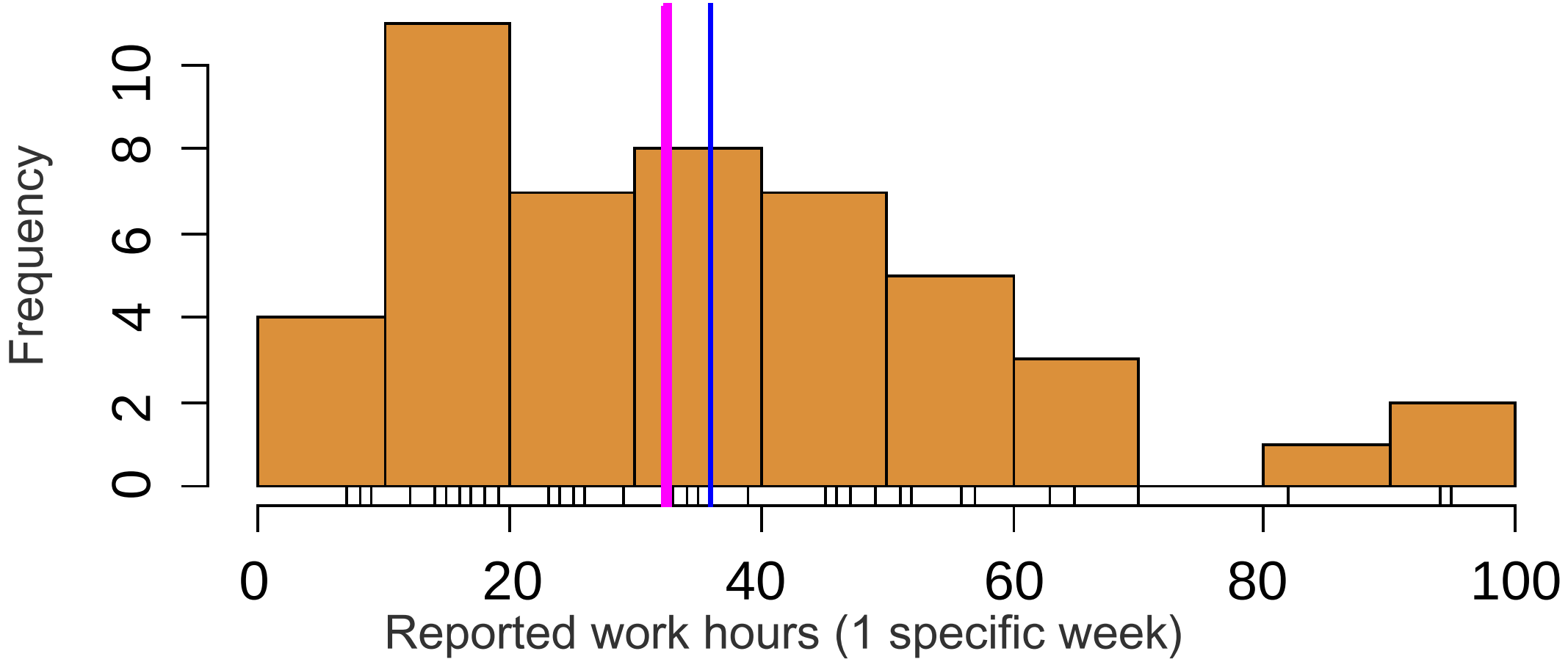
Bar Chart

Scatterplot

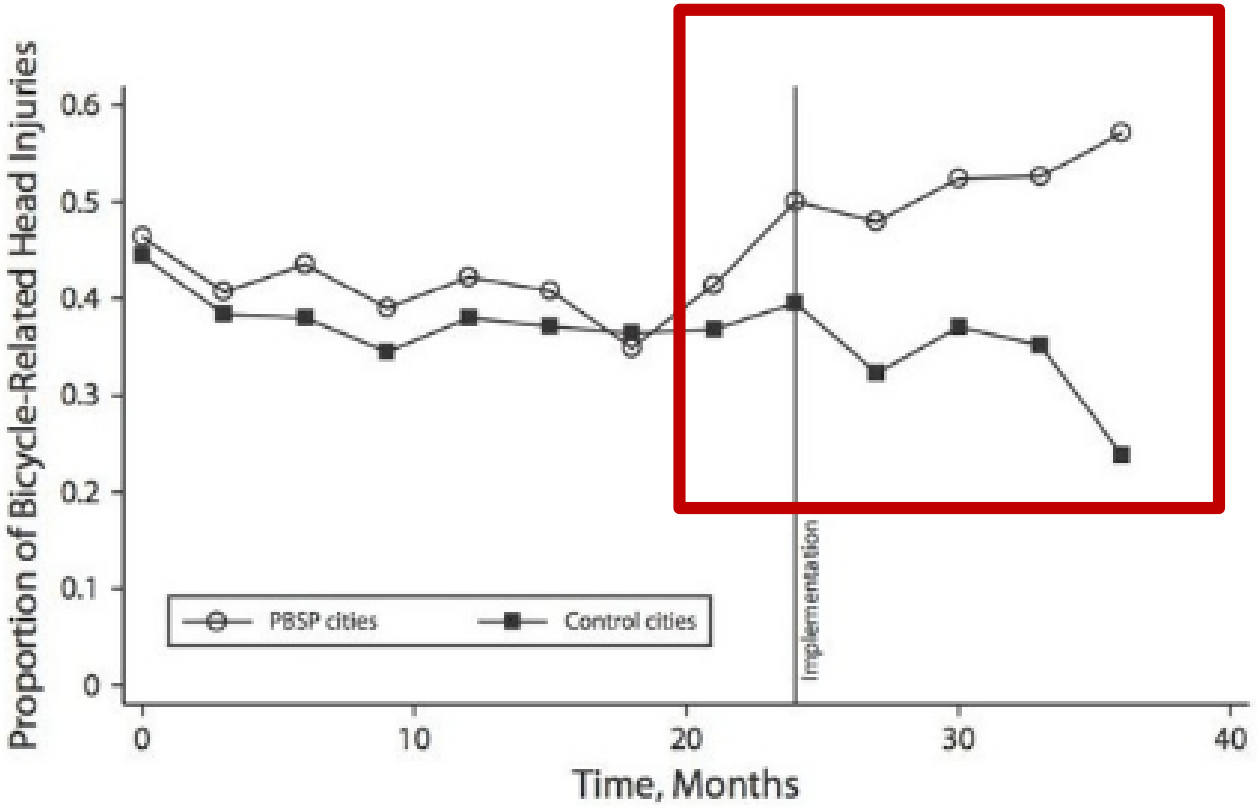
LINE CHART/RUG CHART

- Gaps in the number line indicate an absence of those numeric values in the dataset
- Remember: this is (possibly) different from the order that values appear in the dataset – since it is a number line, it shows where the values fall numerically
- If values are exactly the same, they will be on top of each other.

HISTOGRAMS


Frequency
0
2
4
6
8
10
0
20
40
60
80
100
Reported work hours (1 specific week)

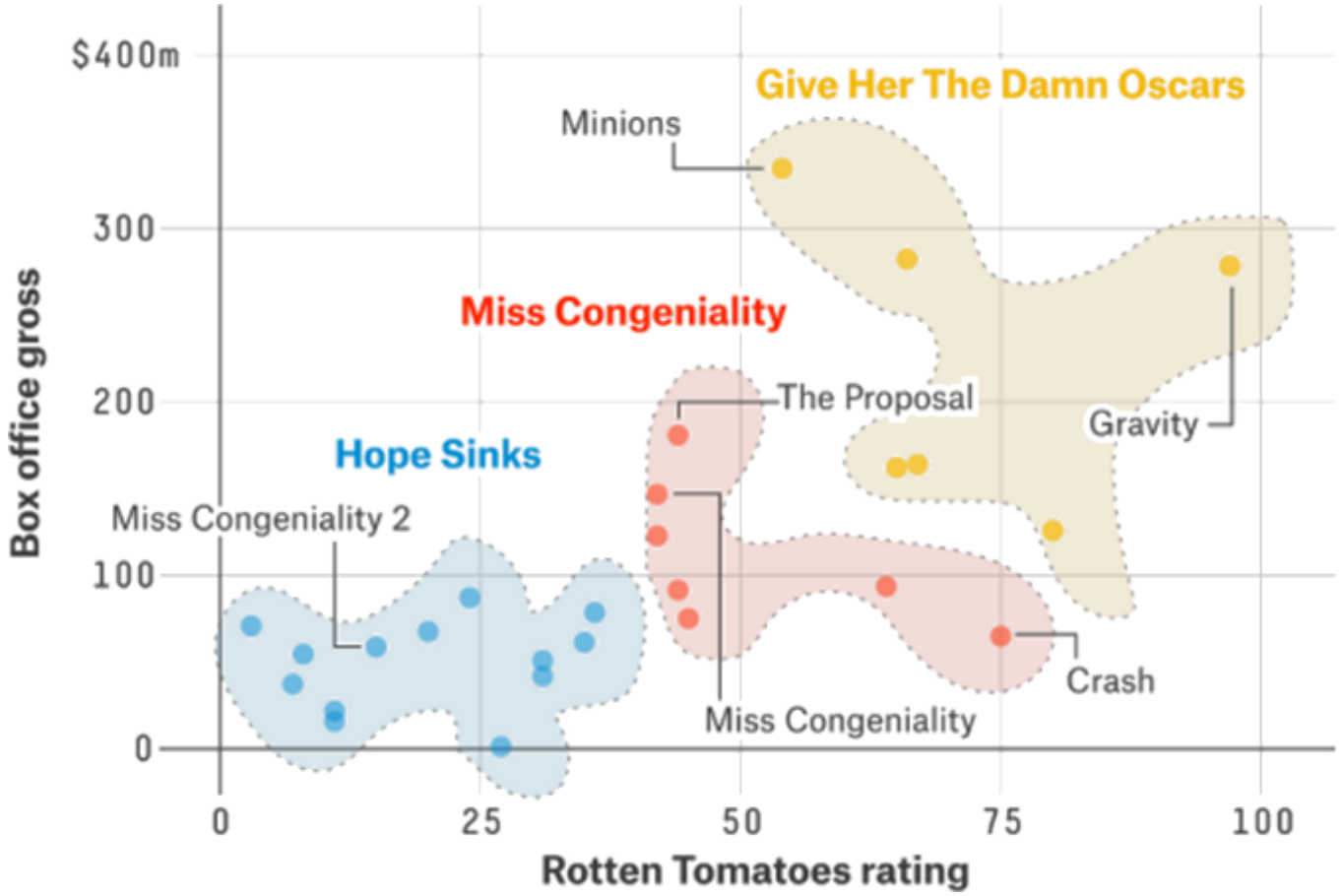
LINE GRAPHS

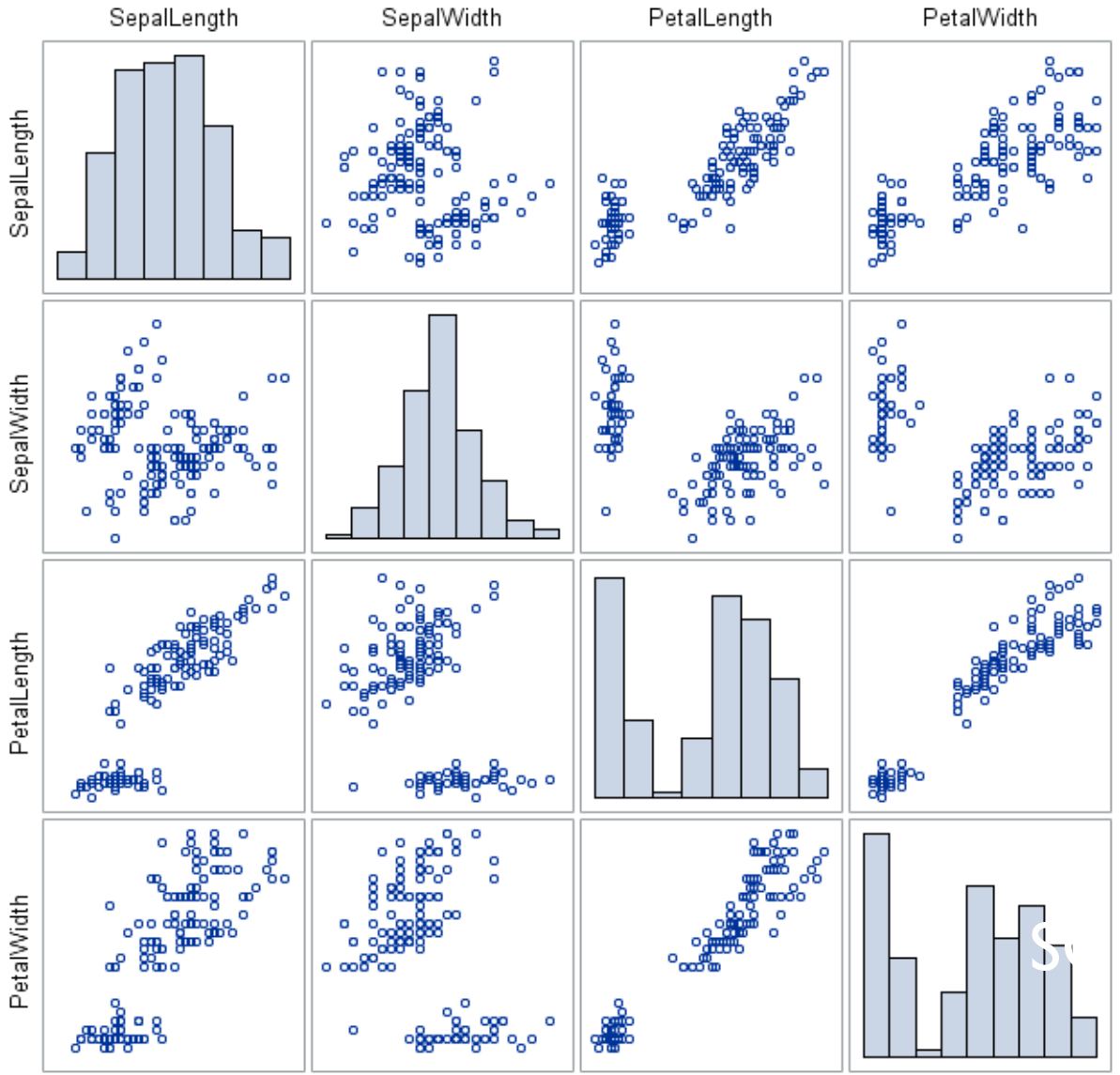


Proportion of all bicycle-related injuries that were classified as head injuries among cities with public bike share programs and control cities, centered on intervention date (vertical line); North America.

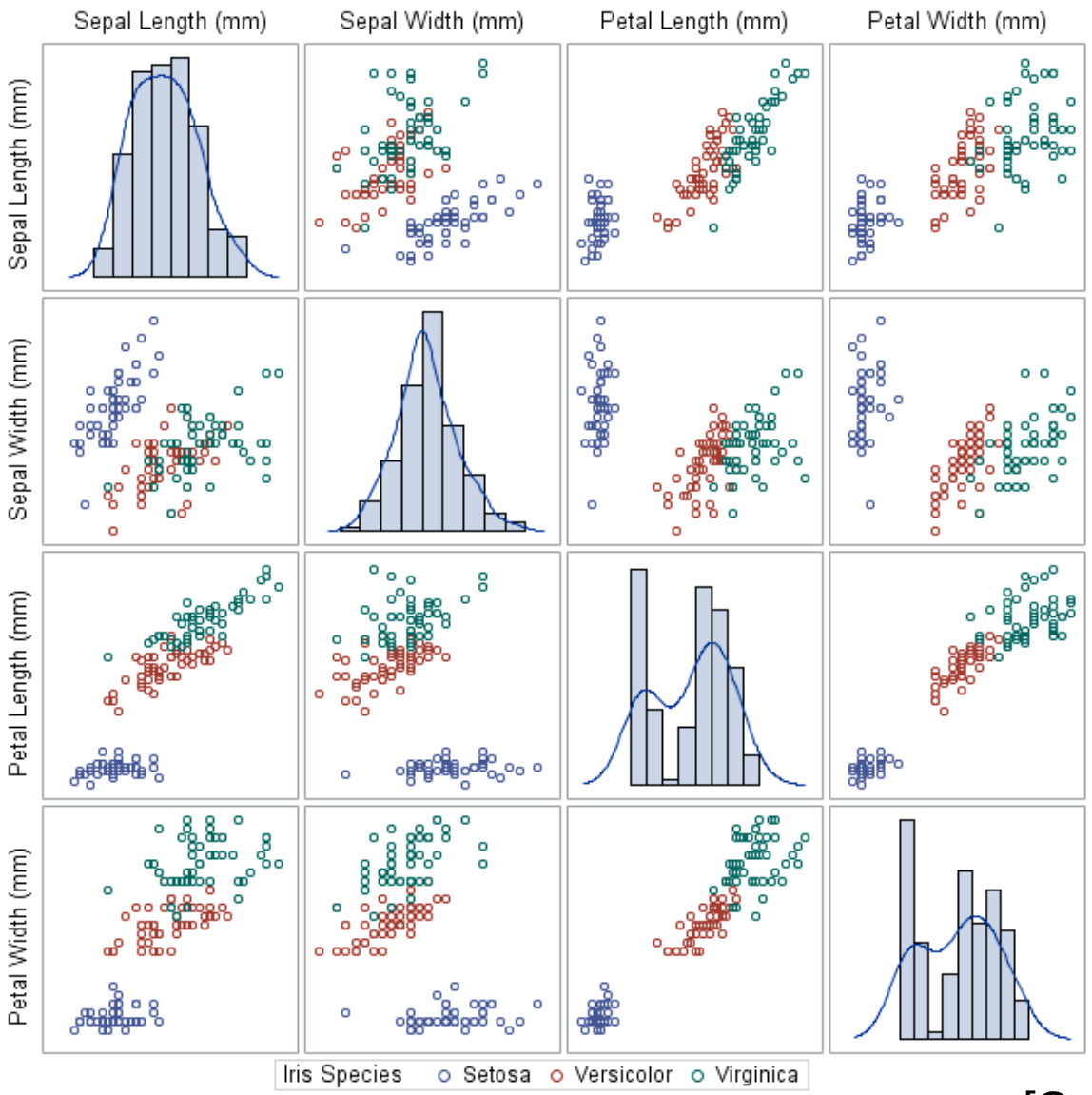
[Graves et al., *Am.J.Phys.Health*, 2014]

SCATTERPLOT

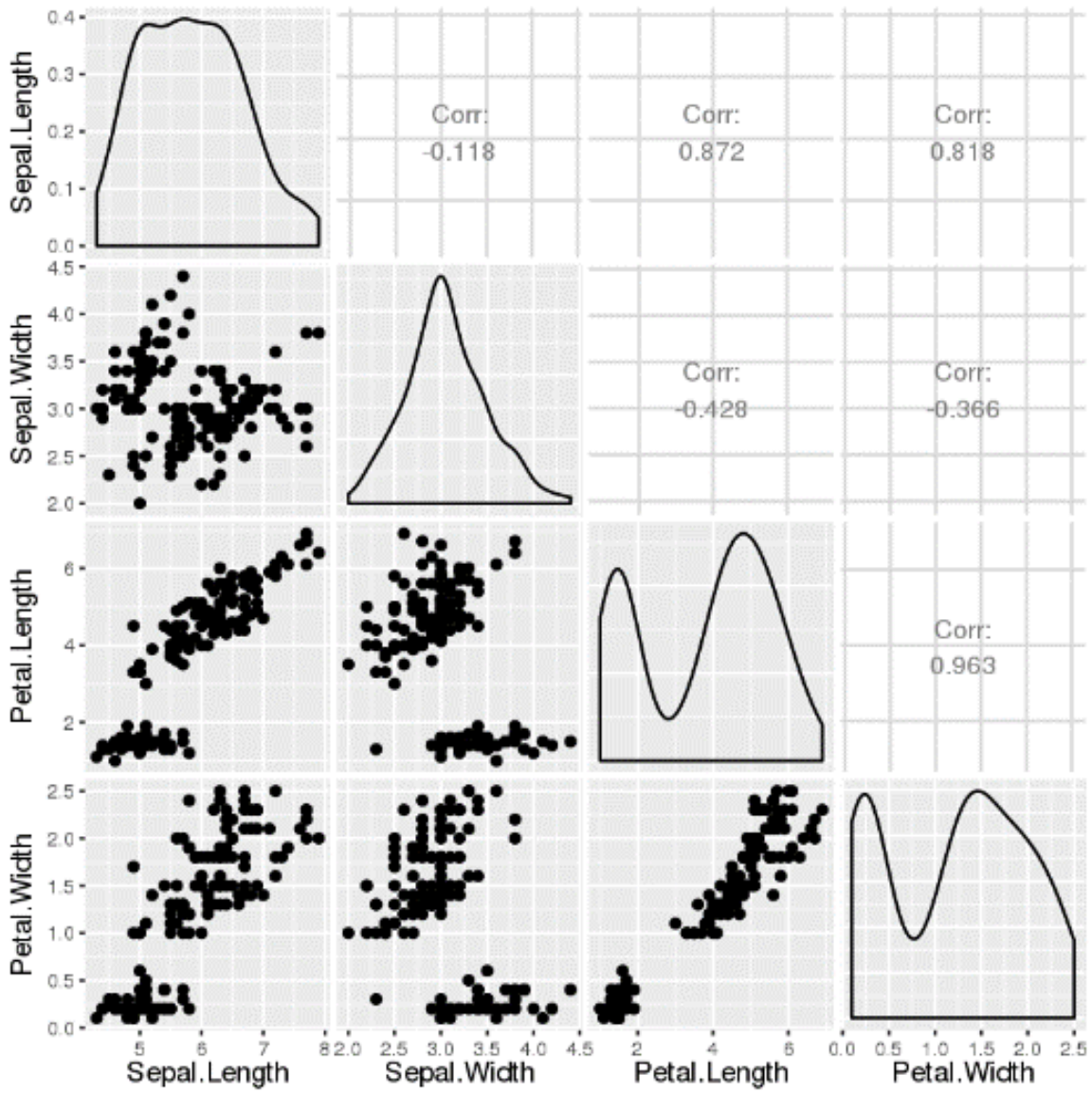




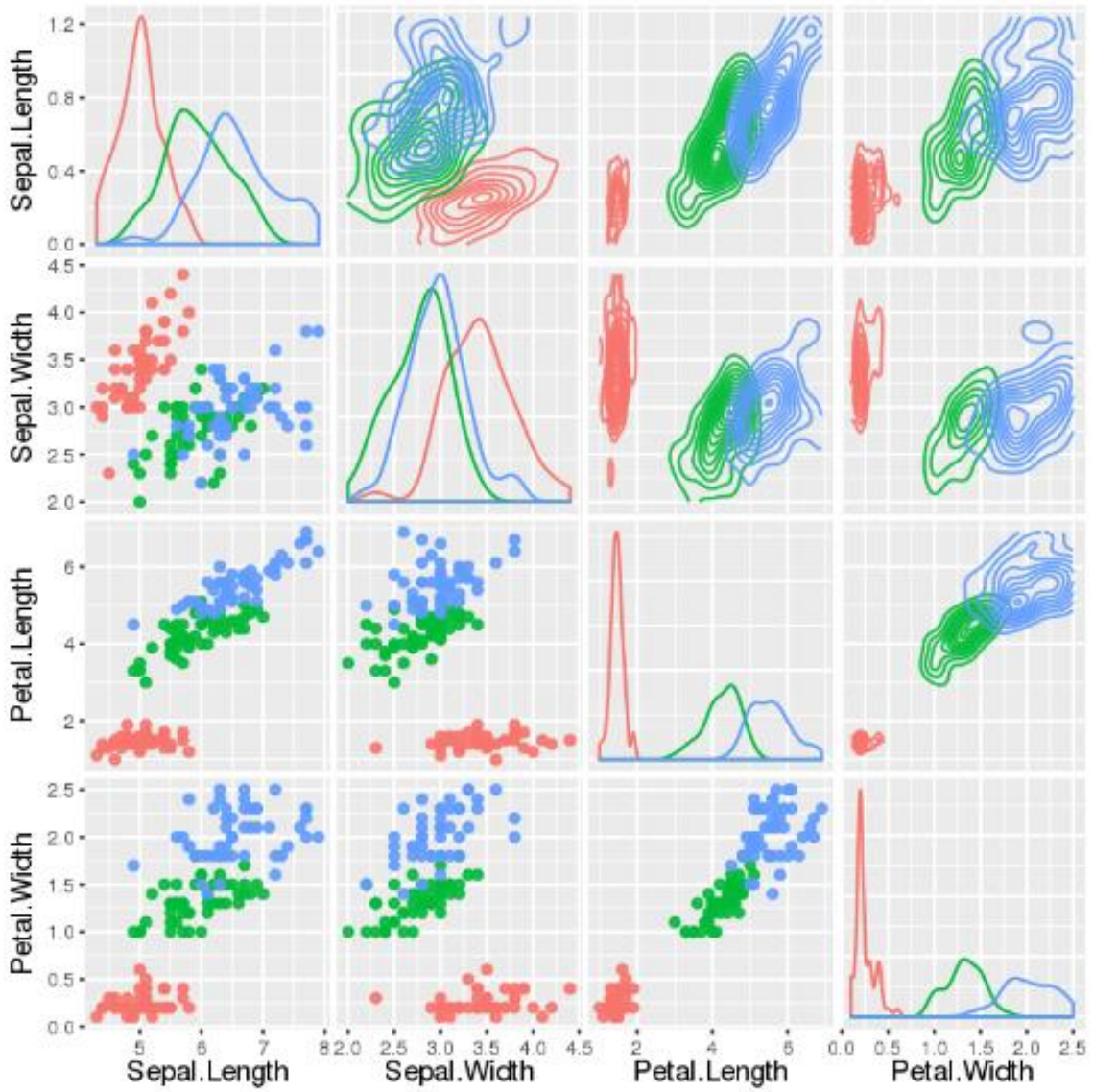
[Created using SAS proc corr]



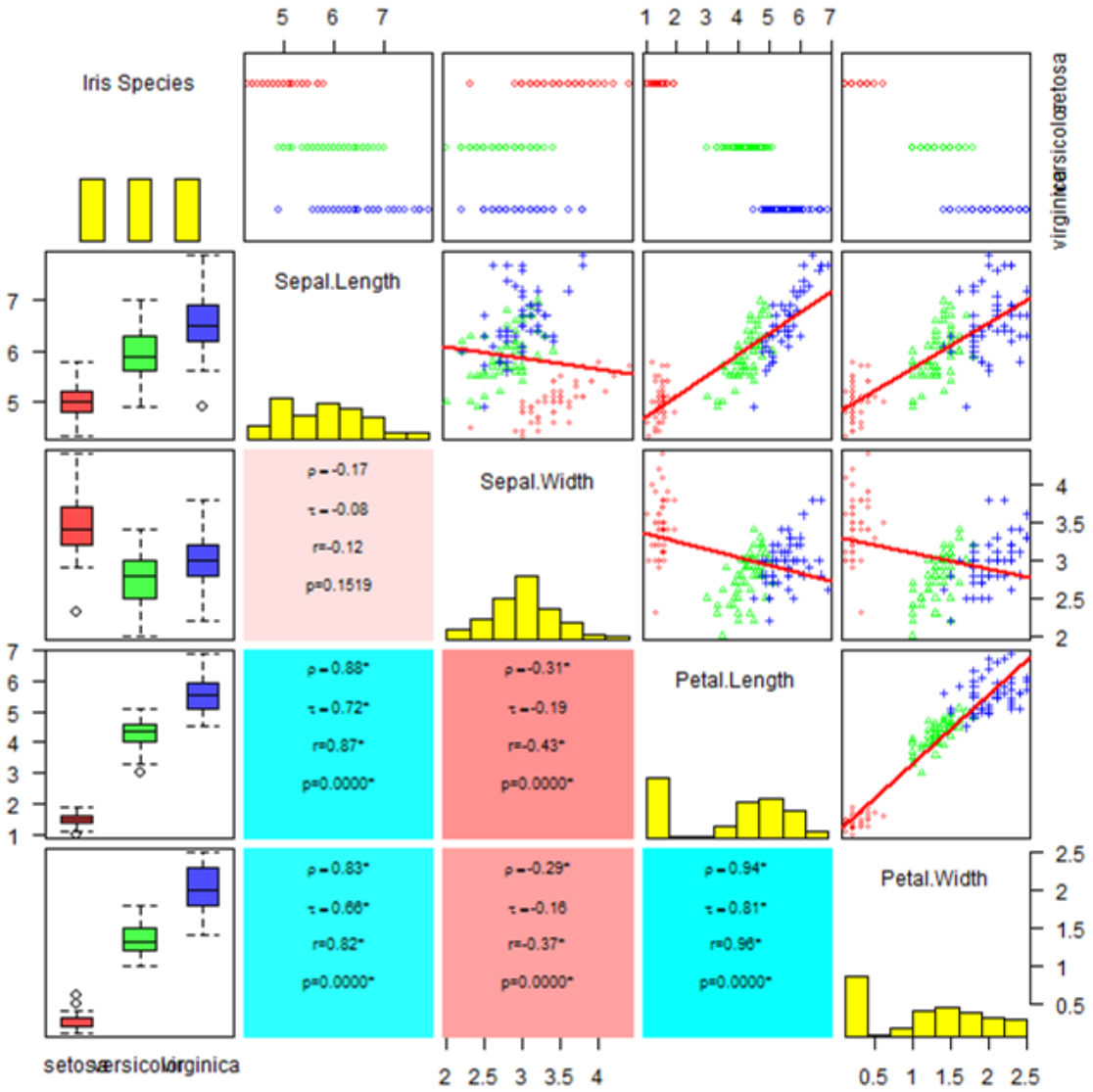
[Created using SAS proc sgscatter]



[Created using R command ggpairs]



[Created using R command ggpairs]



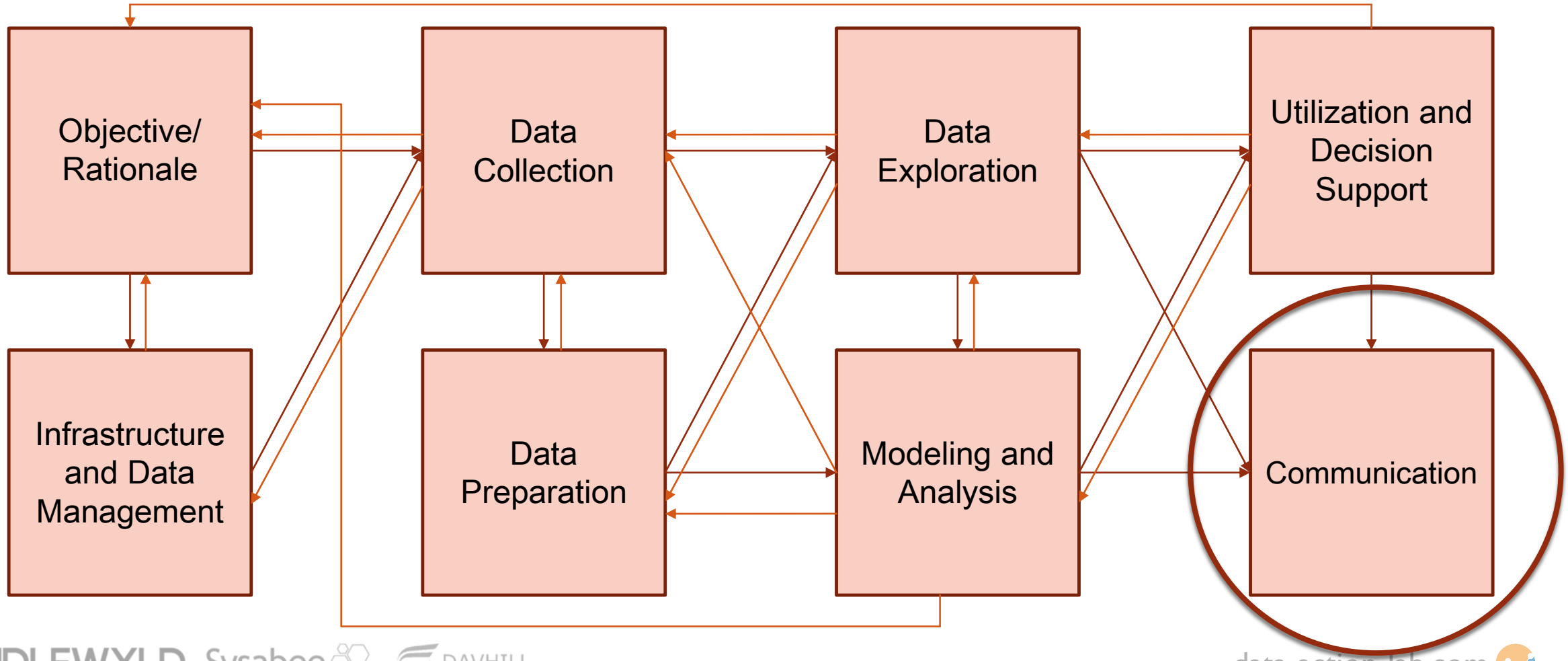
Is this starting to get too cluttered?

[Author unknown, found online]

POST-ANALYSIS DATA VISUALIZATION

DATA EXPLORATION AND DATA VISUALIZATION

THE (MESSY) ANALYSIS PROCESS



[E. Tufte, *Beautiful Evidence*]

FUNDAMENTAL PRINCIPLES OF ANALYTICAL DESIGN

Reasoning and communicating our thoughts are intertwined with our lives in a causal and dynamic multivariate Universe.

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

ACCESSIBILITY

A table can be translated to Braille fairly easily, but that's not always possible for charts.

Describing the features and emerging structures in a visualization is a possible solution… **if they can be spotted**.

Analysts must produce clear and meaningful visualizations, but they must also describe them and their features in a fashion that allows all to "see" the insights.

ACCESSIBILITY

Analysts need to have “seen” all the insights, which is not necessarily the case (if at all possible).

Data Perception:

- texture-based representations
- text-to-speech
- use of sounds/music
- odor-based or taste-based representations (?!?)

[J.P. Blackard]

INFOGRAPHICS

Created for **story-telling** purposes (**subjective**)

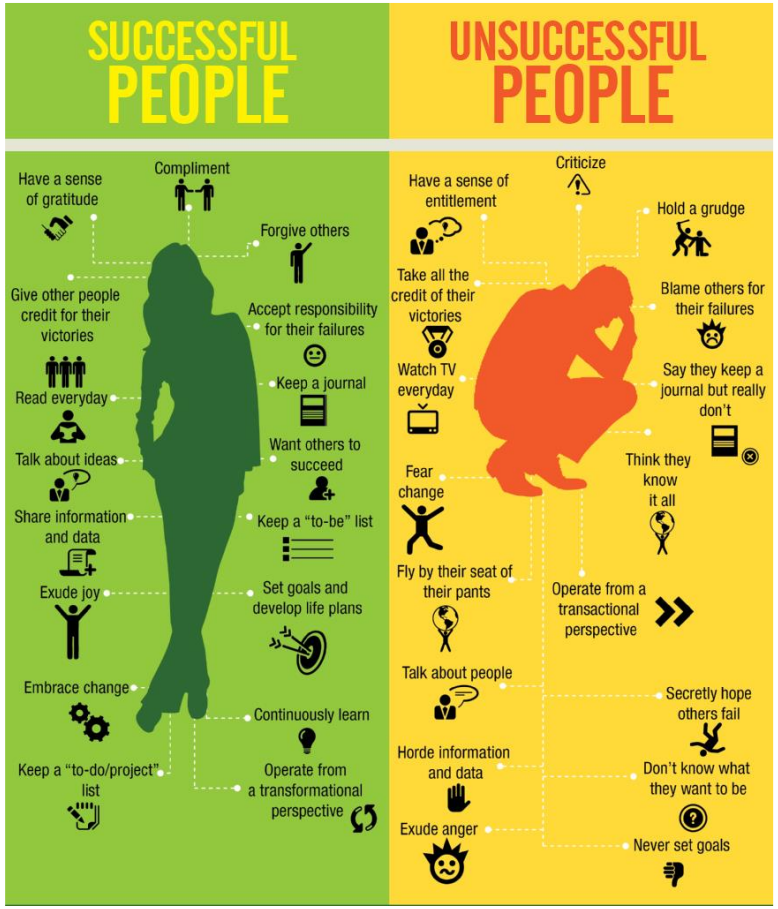
Intended for a **specific** audience

Self-contained and discrete

Graphic design aspect is key

Cannot usually be re-used with other data

Can incorporate **unquantifiable** information



DATA VISUALIZATION

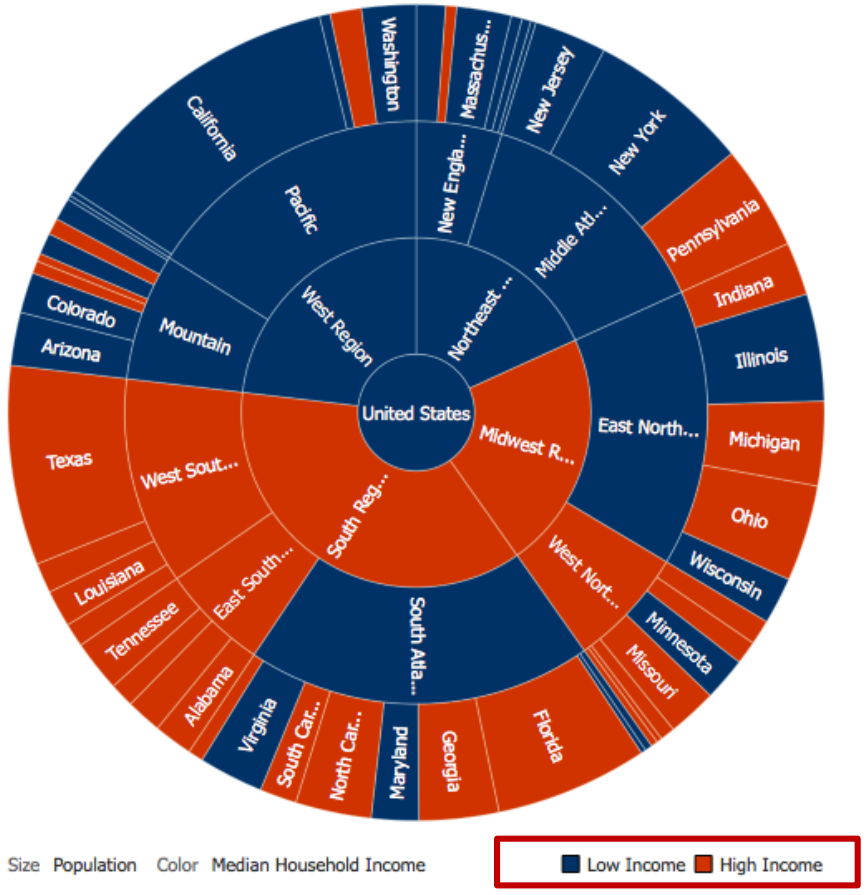
A **method**, as well as an item (**objective**)

Typically focuses on the **quantifiable**

Used to make sense of the data or to make it **accessible** (datasets can be massive and unwieldy)

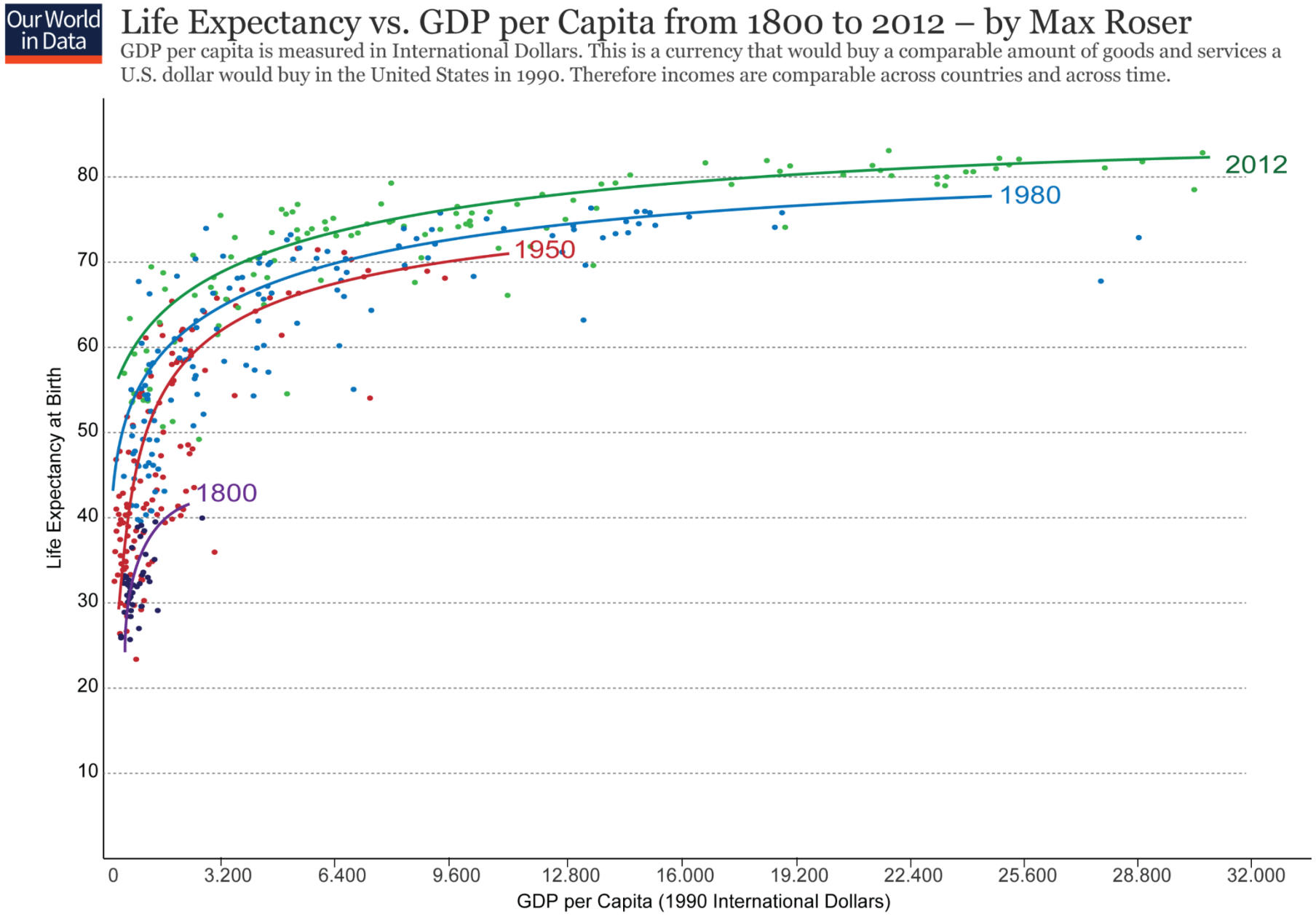
May be generated automatically

The look and feel are less important than the **insights conveyed** by the data



?

[https://ourworldindata.org/life-expectancy]

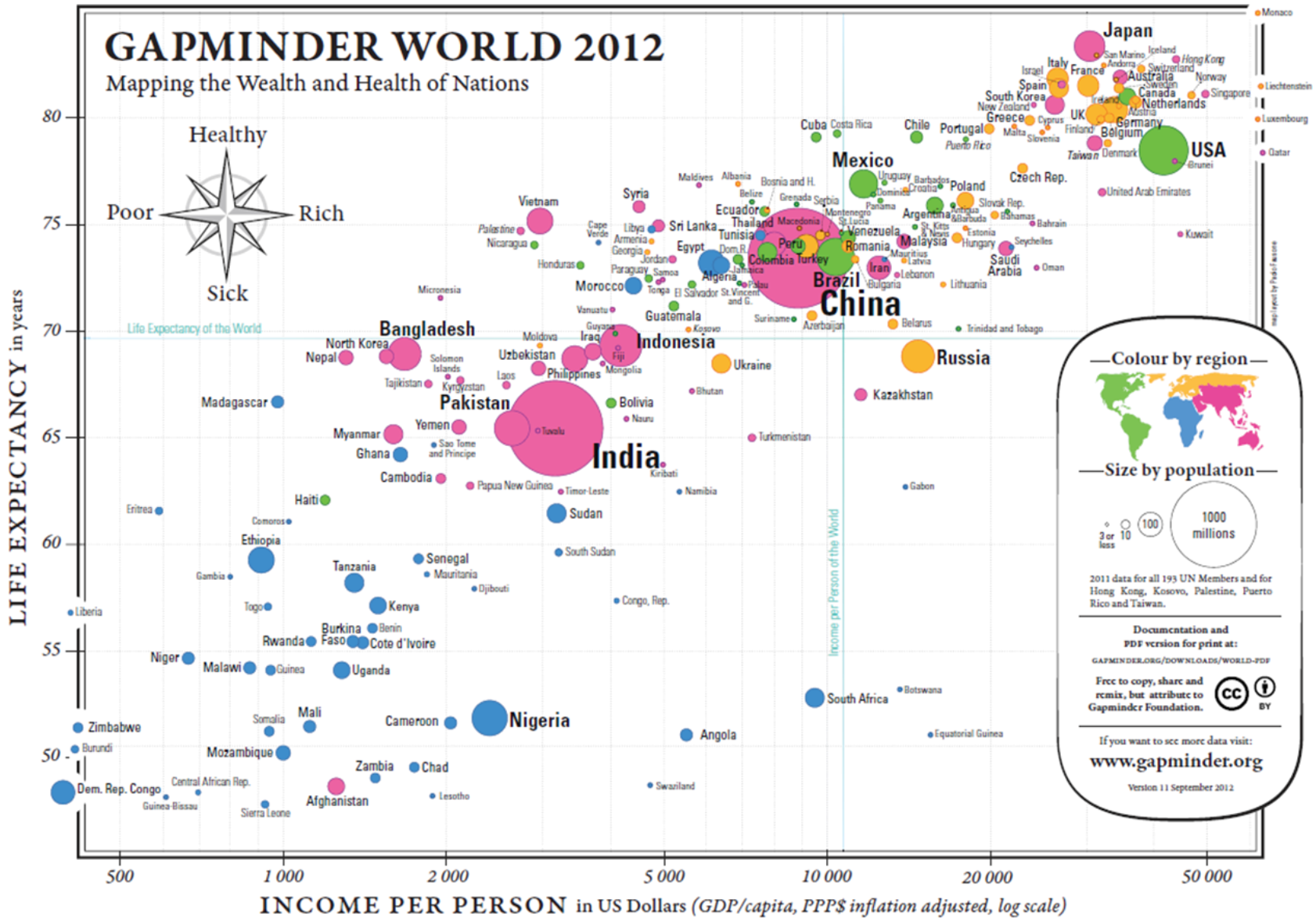


This graph displays the correlation between life expectancy and GDP per capita.

Countries with higher GDP have a higher life expectancy, in general.

The relationship seems to follow a **logarithmic trend**: the unit increase in life expectancy per unit increase in GDP decreases as GDP per

[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**.

VISUAL PROCESSING

Perception is **fragmented** – eyes are continuously scanning.

Visual thinking seeks patterns

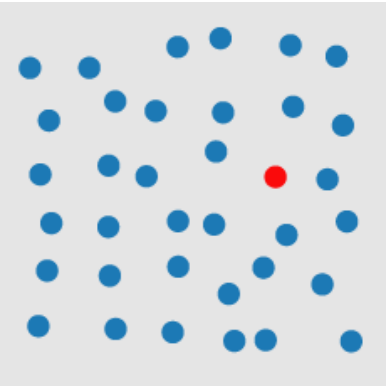
- **Pre-attentive processes:** fast, instinctive, efficient, multitasking gather information and build patterns:

  features → patterns → objects

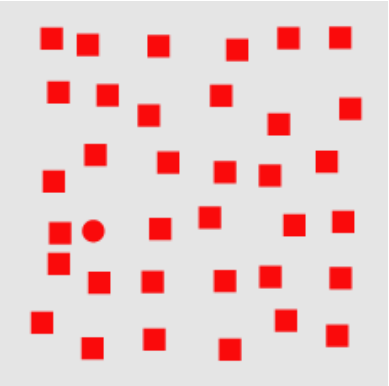
- **Attentive process:** slow, deliberate, focused discover features in the patterns:

  objects → patterns → features

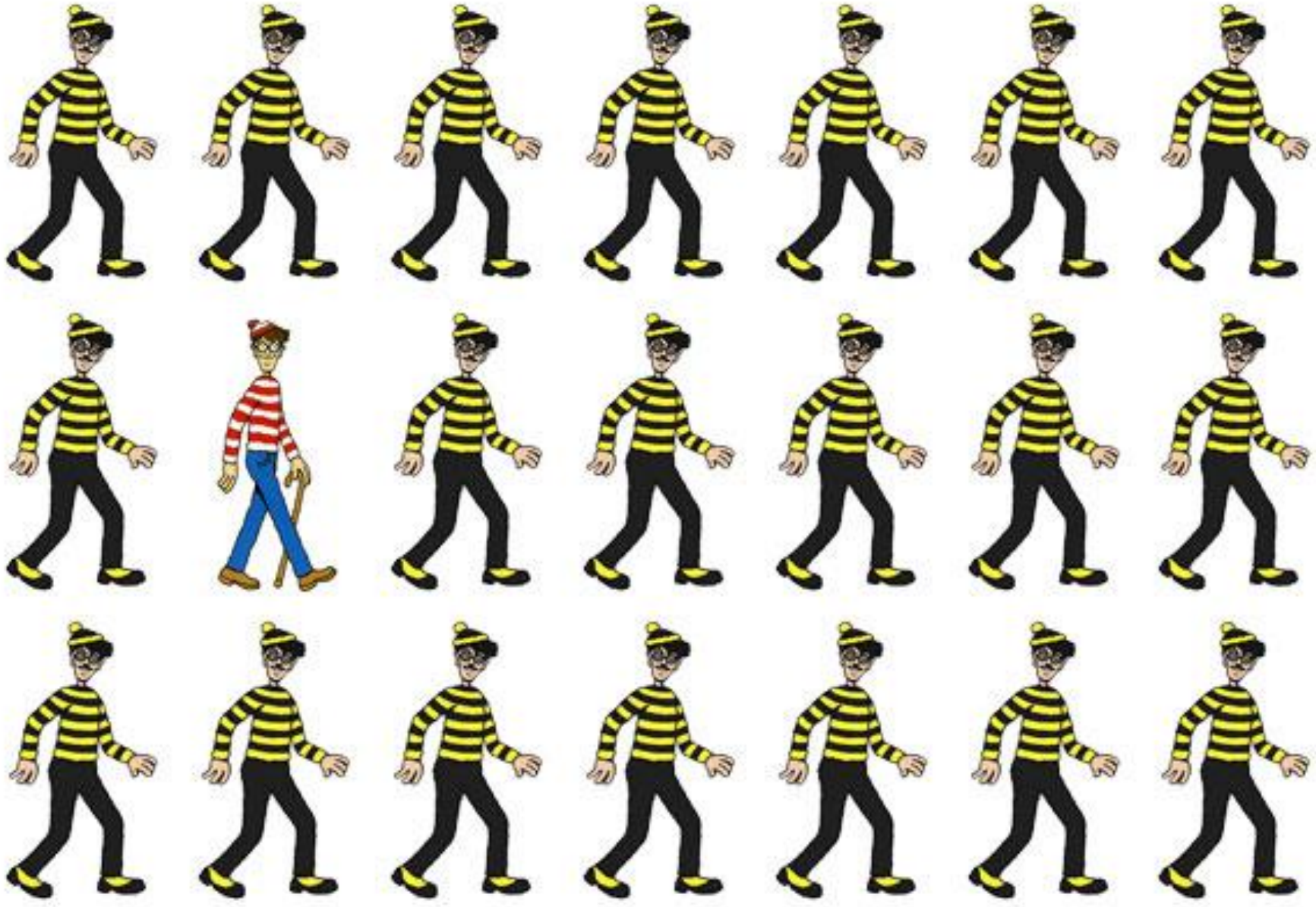
pre-attentive



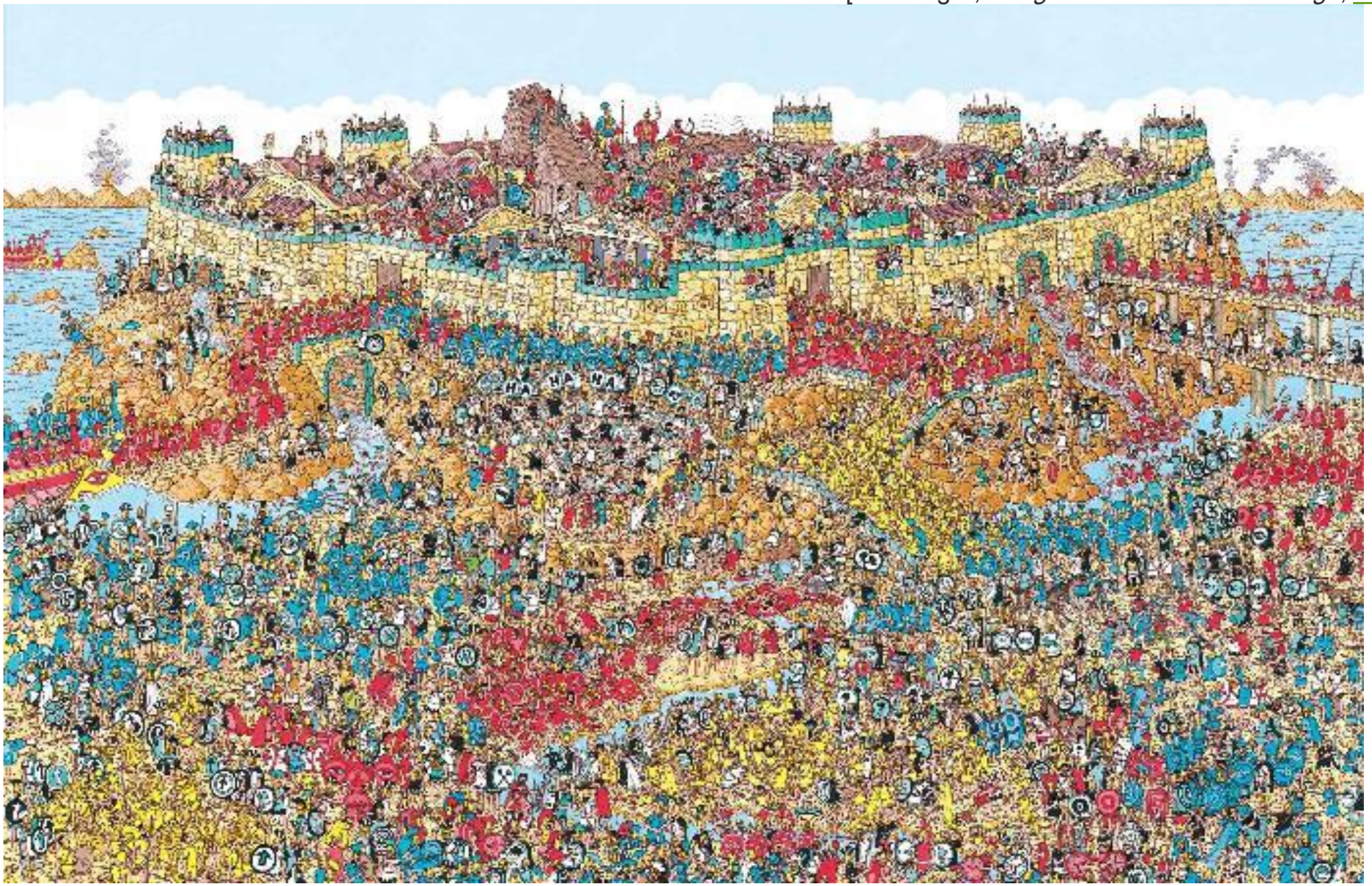
attentive



[L. Groeger, *A Big Article About Wee Things*, ProPublica.com]



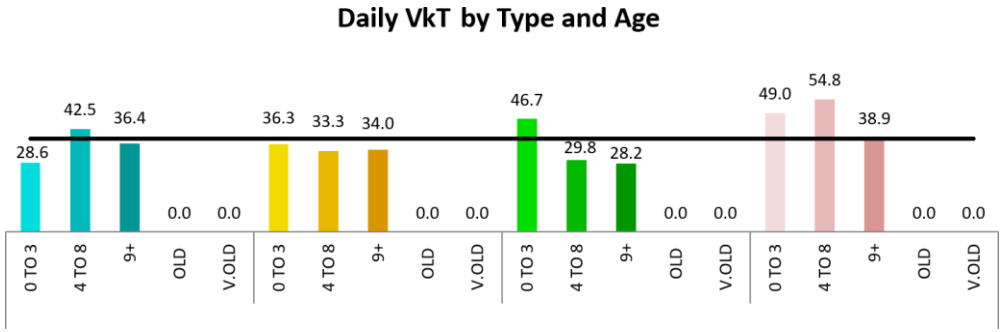
[L. Groeger, *A Big Article About Wee Things*, ProPublica.com]



BASIC RULES

1. Check the data
outliers, spikes, anomalies

2. Explain encoding
don't assume the reader knows what everything means

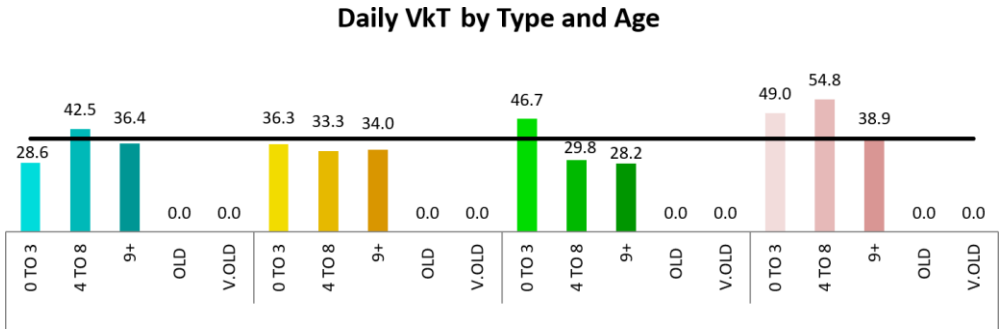


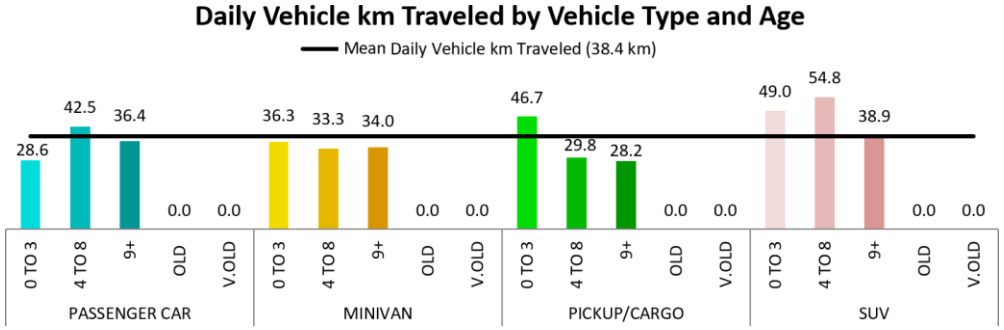
3. Label axes
knowing the scale is important

BASIC RULES

1. Check the data
outliers, spikes, anomalies

2. Explain encoding
don't assume the reader knows what everything means





3. Label axes
knowing the scale is important

[http://flowingdata.com/tag/pitfalls]

BASIC RULES

4. Include units
eliminate the need for guesswork

5. Keep your geometry in check
circles and 2D shape are sized by area, bar charts by length

6. Include your sources
protect yourself, and let those who want to dig deeper do so

7. Consider your audience
a poster can be wordy, a presentation should be minimalist

[http://flowingdata.com/tag/pitfalls]

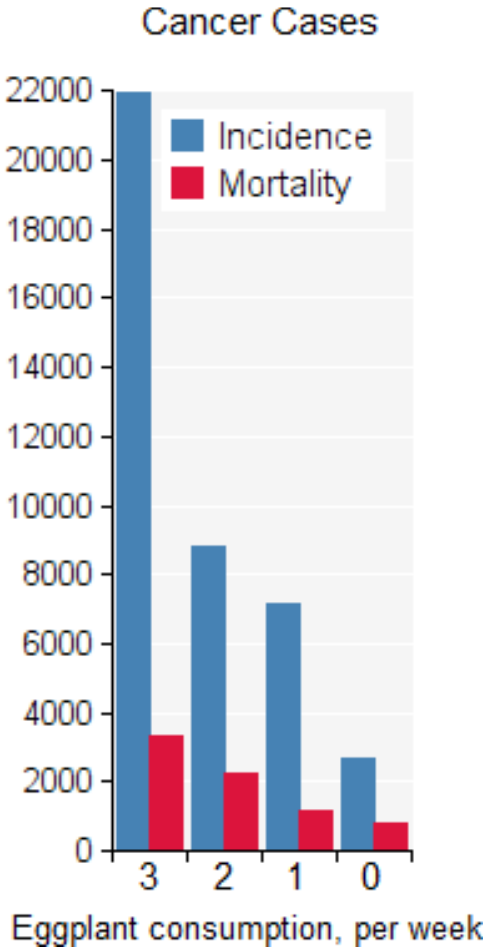
BASIC RULES

4. Include units
eliminate the need for guesswork

5. Keep your geometry in check
circles and 2D shape are sized by area, bar charts by length

6. Include your sources
protect yourself, and let those who want to dig deeper do so

7. Consider your audience
a poster can be wordy, a presentation should be minimalist



[http://flowingdata.com/tag/pitfalls]

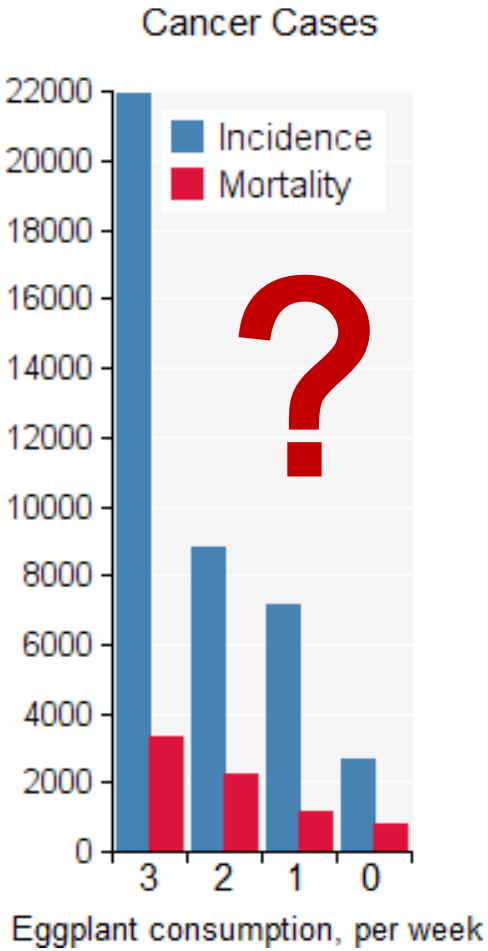
BASIC RULES

4. Include units
eliminate the need for guesswork

5. Keep your geometry in check
circles and 2D shape are sized by area, bar charts by length

6. Include your sources
protect yourself, and let those who want to dig deeper do so

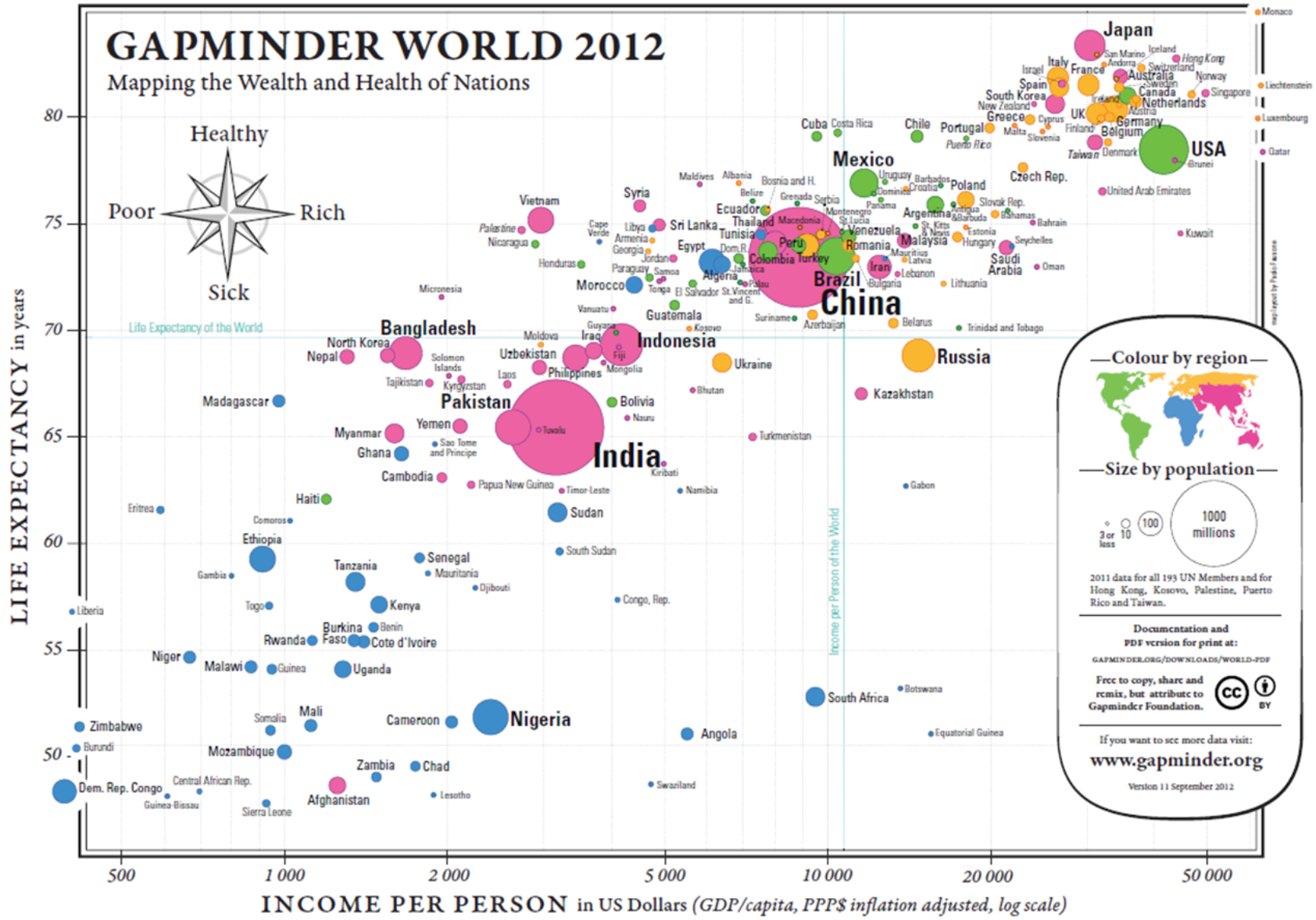
7. Consider your audience
a poster can be wordy, a presentation should be minimalist



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

Group Exercise

- How does this visualization help its audience understand the data?
- What are some interesting patterns you can see in this visualization?



DISCUSSION

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

In *Semiology of Graphics*, Bertin suggests that **not all retinal variables are equally effective** when it comes to convey or represent information. You may need to experiment 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.

VISUALIZATION CATALOGUE

[J. Nelson, IDV Solutions]

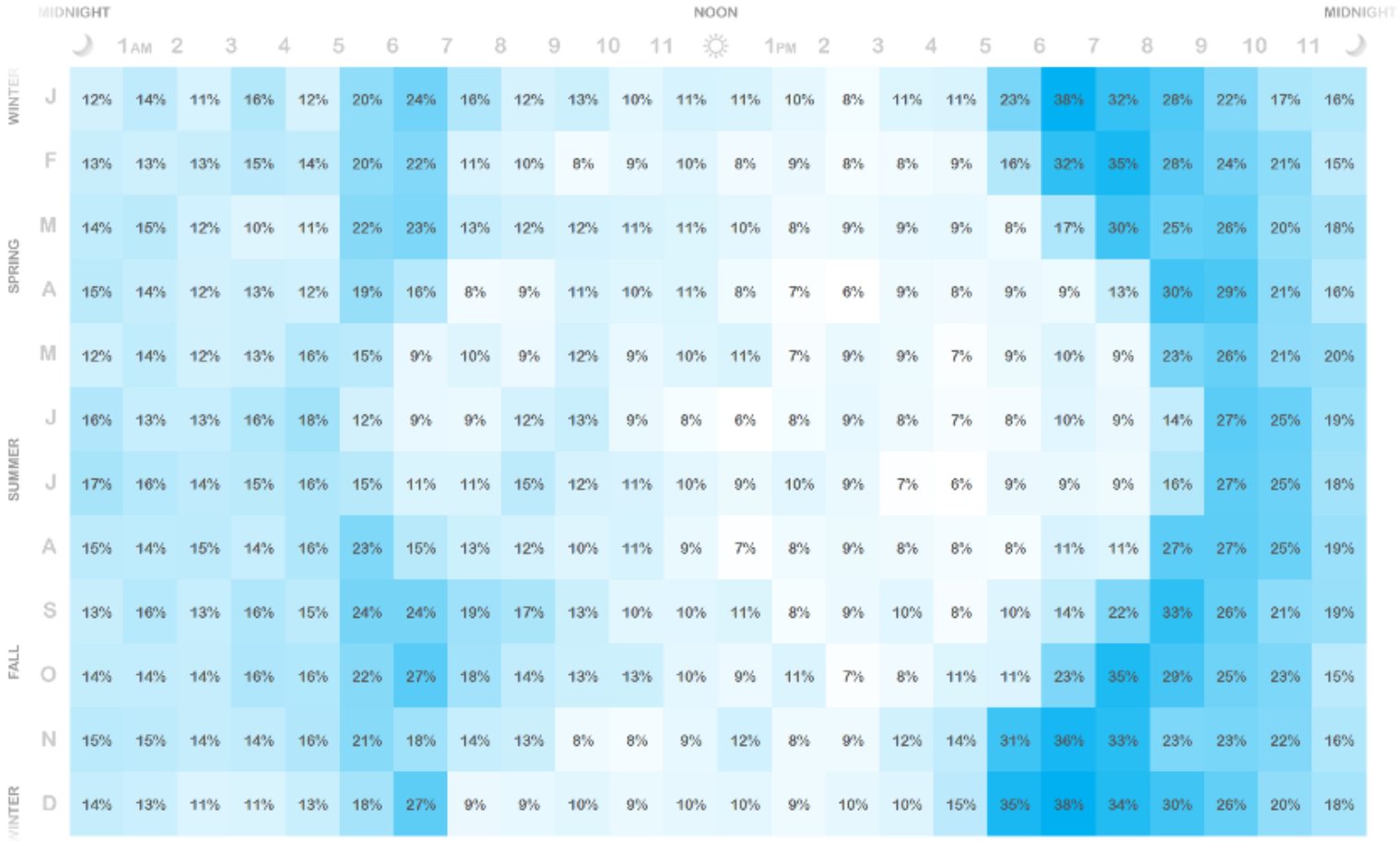
HEAT MAPS

The Horizon of Pedestrian Risk

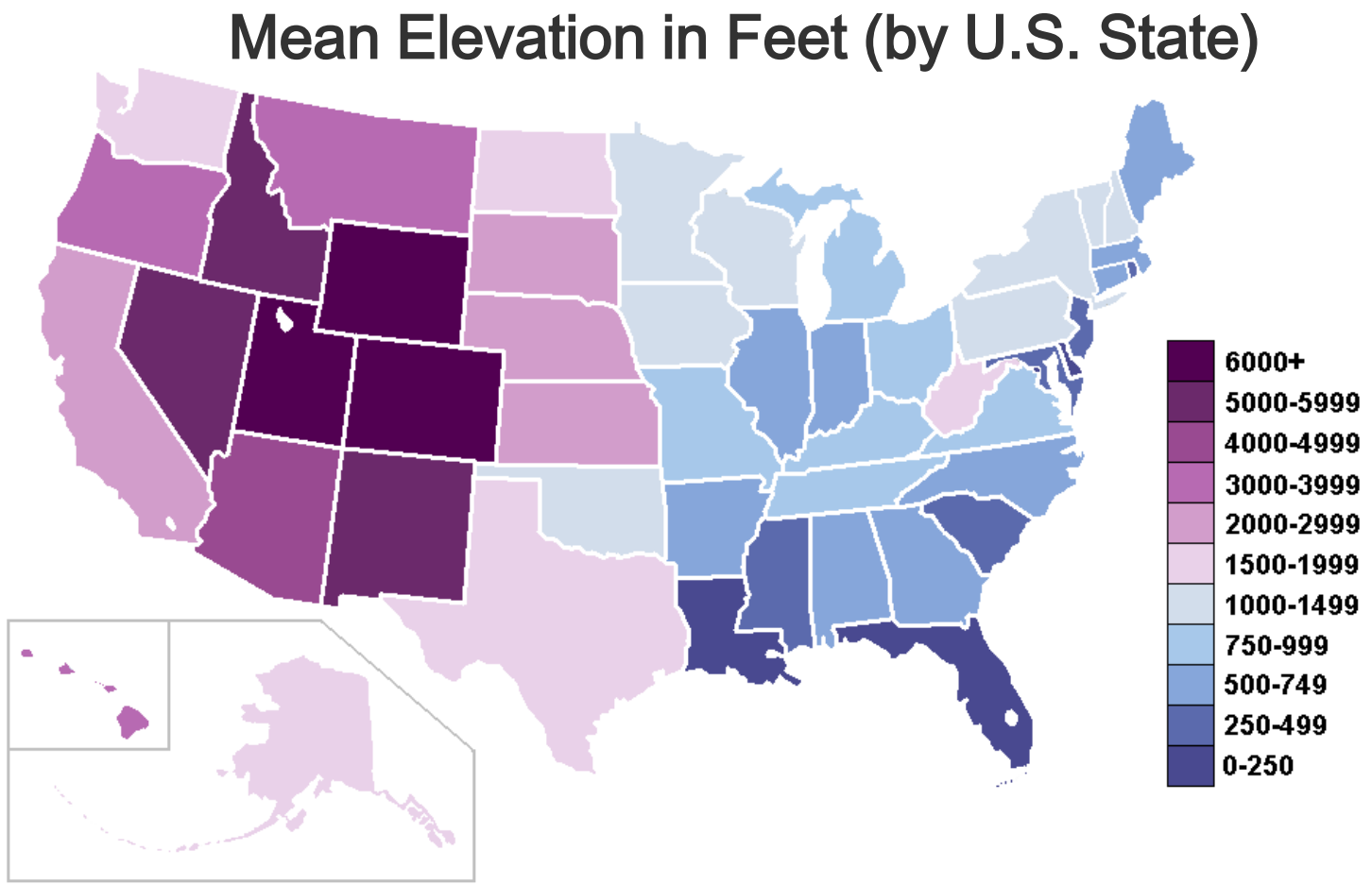
The rate of fatal traffic incidents involving pedestrians, each hour of the day, throughout the seasons of the year.

The seasonal shift of our setting sun traces an ark of elevated risk – an echo of the curve of the Earth, itself (**Note:** ???).

Source: Fatality Analysis Reporting System (NHTSA 2006-2010)



HEAT MAPS (CHOROPLETHS)



HEAT MAPS

Ideal to look at the relationship between 3 or 4 variables

- if one of them represents a percentage or a value within a set range (in order to fix the colour scale, for comparison purposes)
- and the other can act as categorical variables / size variables

Better to **bin the data**, even if the axes variables are continuous (decreases the number of required observations for usefulness)

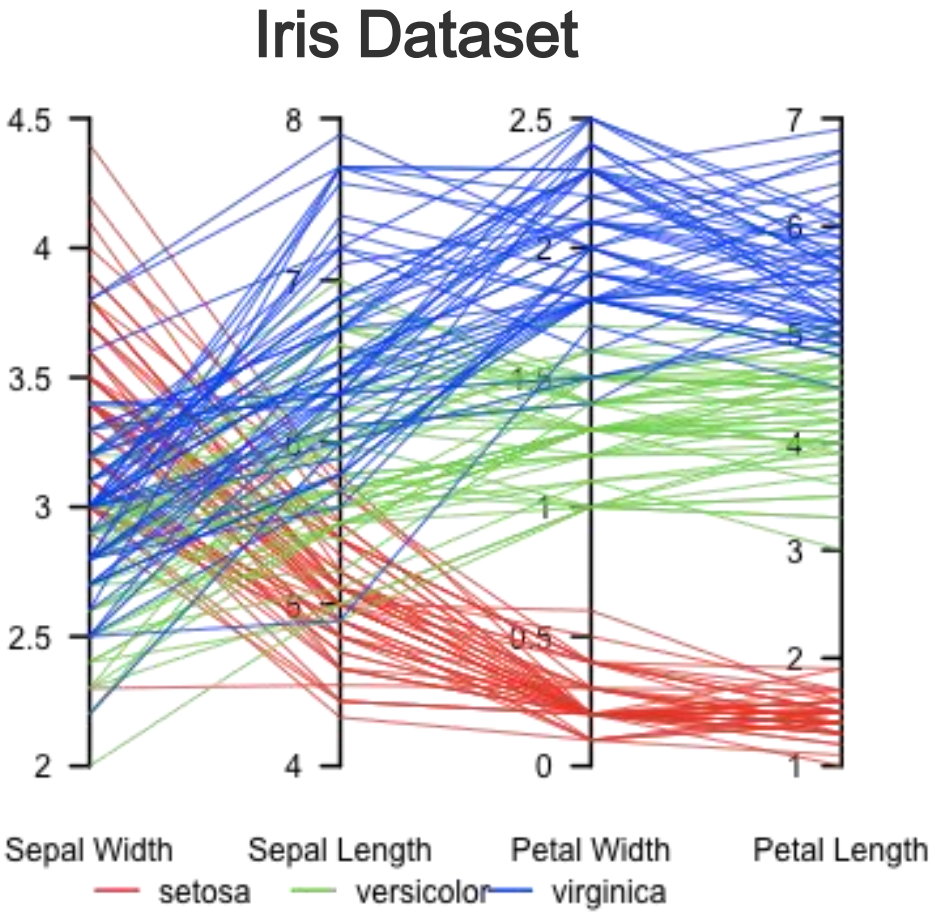
Easier to read if colours are selected along natural colour gradients, such as

Red → Green or Red → Yellow → Green

for instance (but that's not ideal if colour blind)

[Wikipedia entry on Parallel Coordinates]

PARALLEL COORDINATES

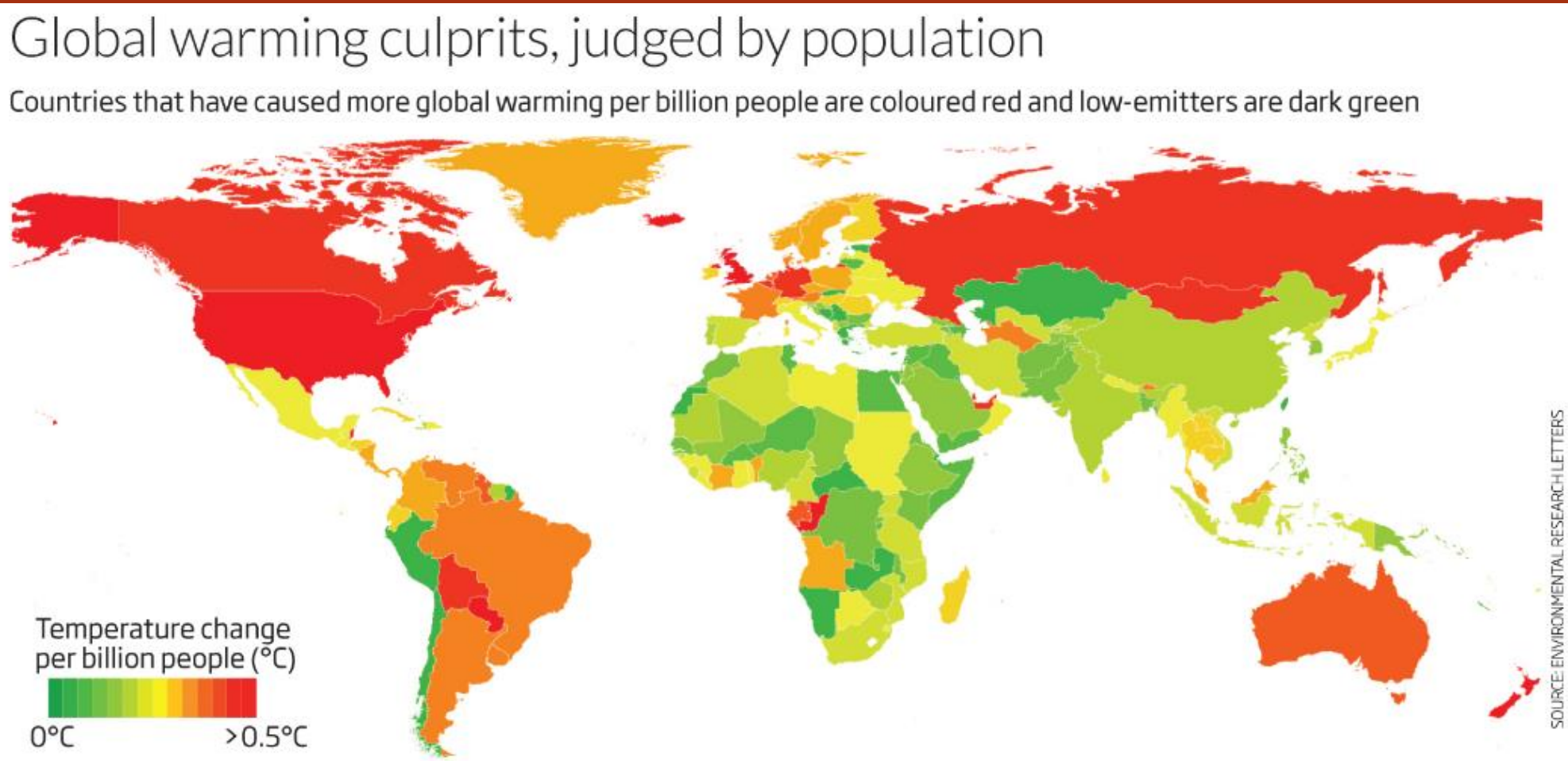


Fine for just a few subjects, or if behaviour differs markedly across groups.

Overlapping trajectories can make them difficult to read.

[New Scientist, Jan 2014]

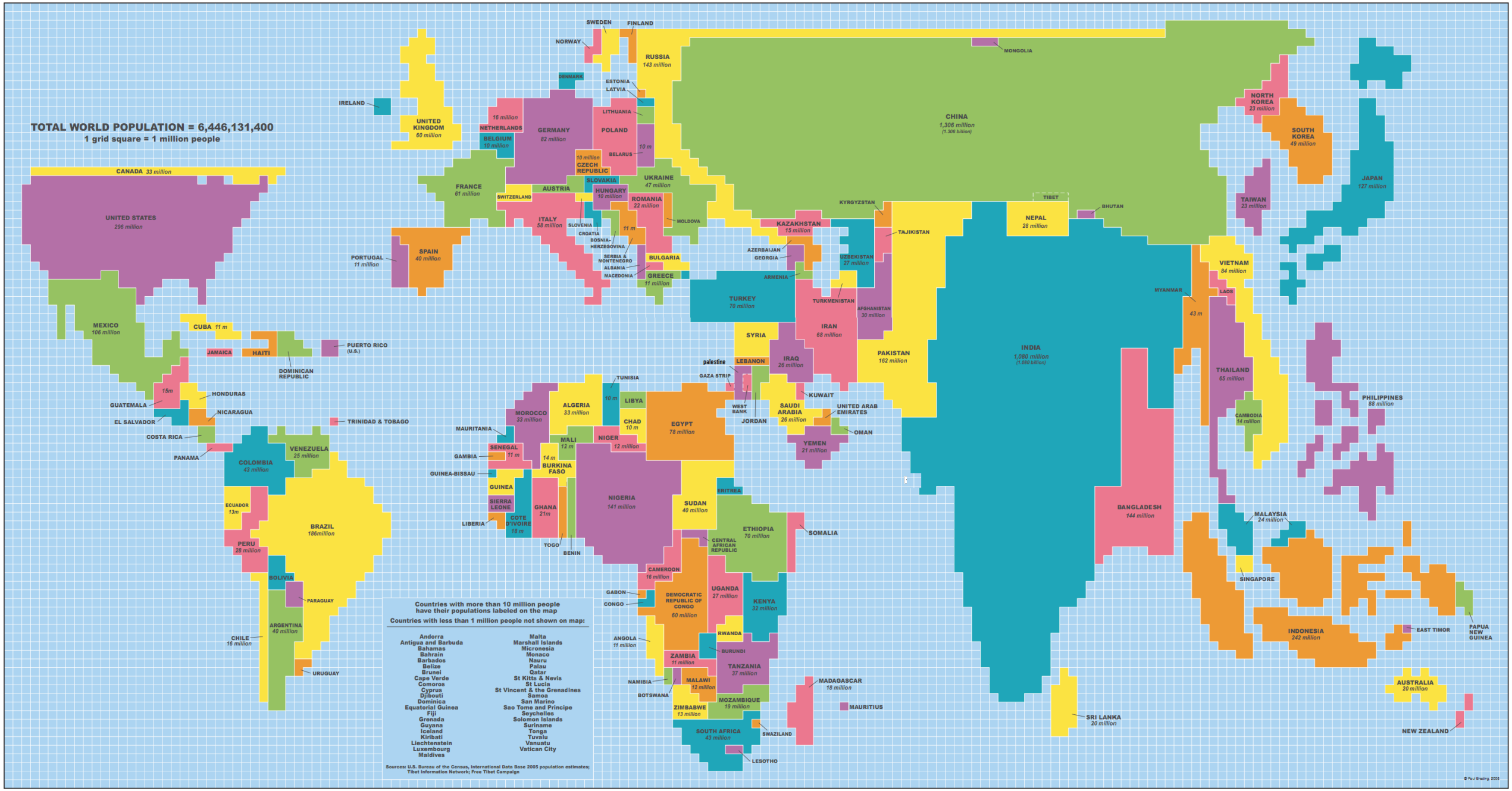
GEOGRAPHICAL MAPS



[New Scientist, Jan 2014]

GEOGRAPHICAL MAPS

[Paul Breding]



[Upside Down Maps]

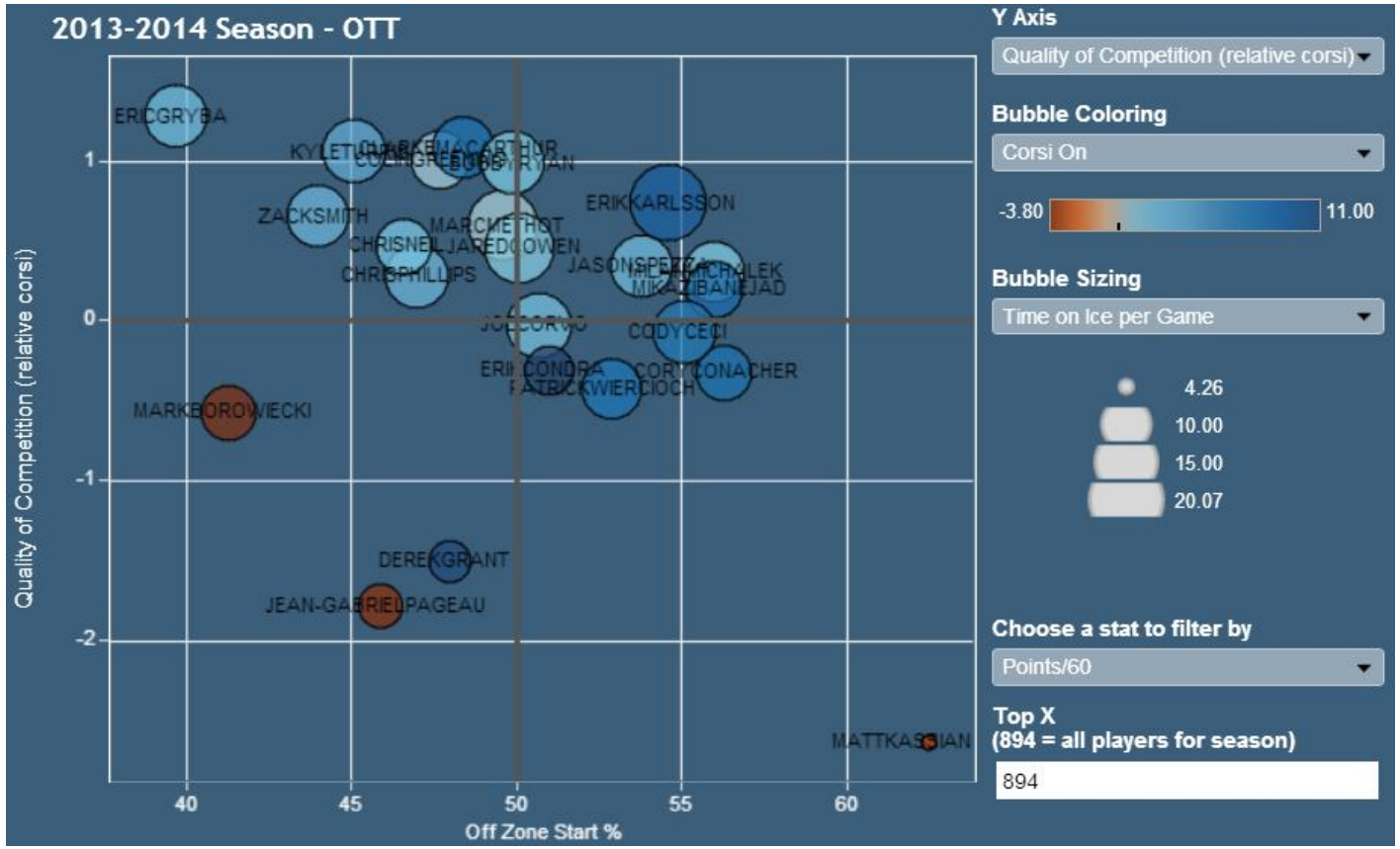
MAPS

Most of us are quite familiar with geographical maps, so they tend to be easier to interpret.

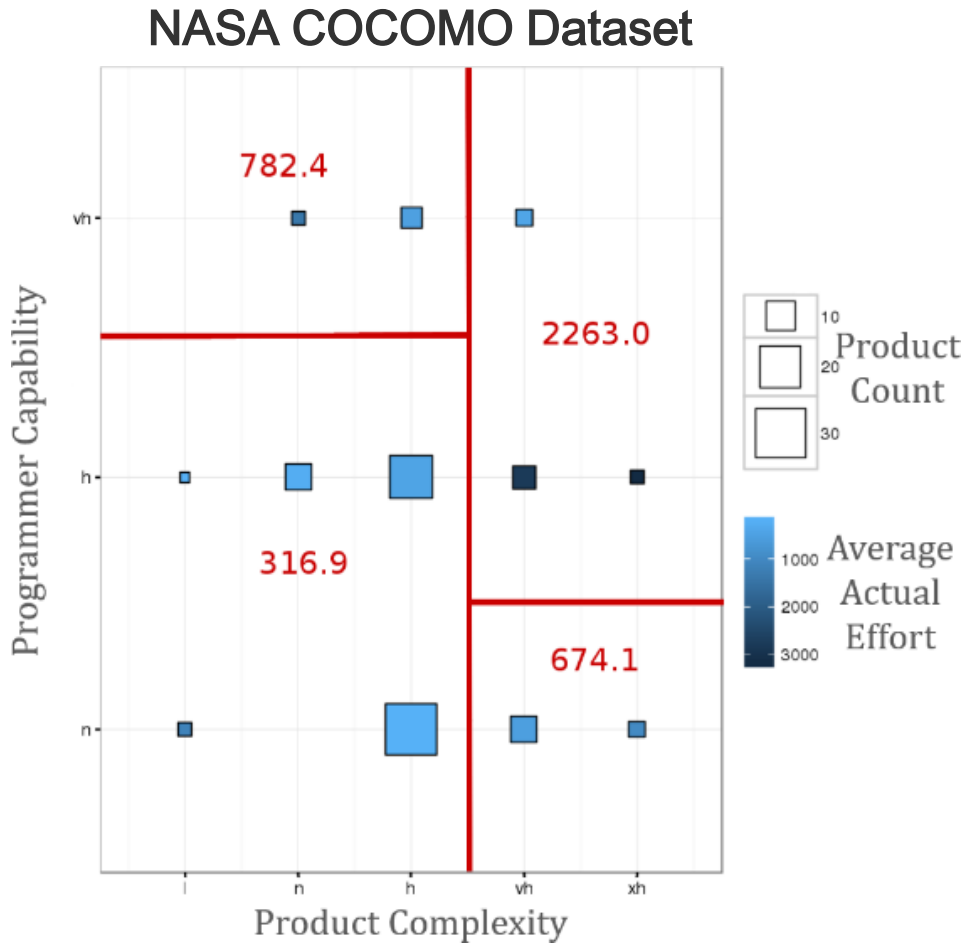
Can produce a striking effect when the data visualization shows **un-expected results**

- which may mask significant information
- or lack of significant information
- or change the way you view things

BUBBLE CHARTS

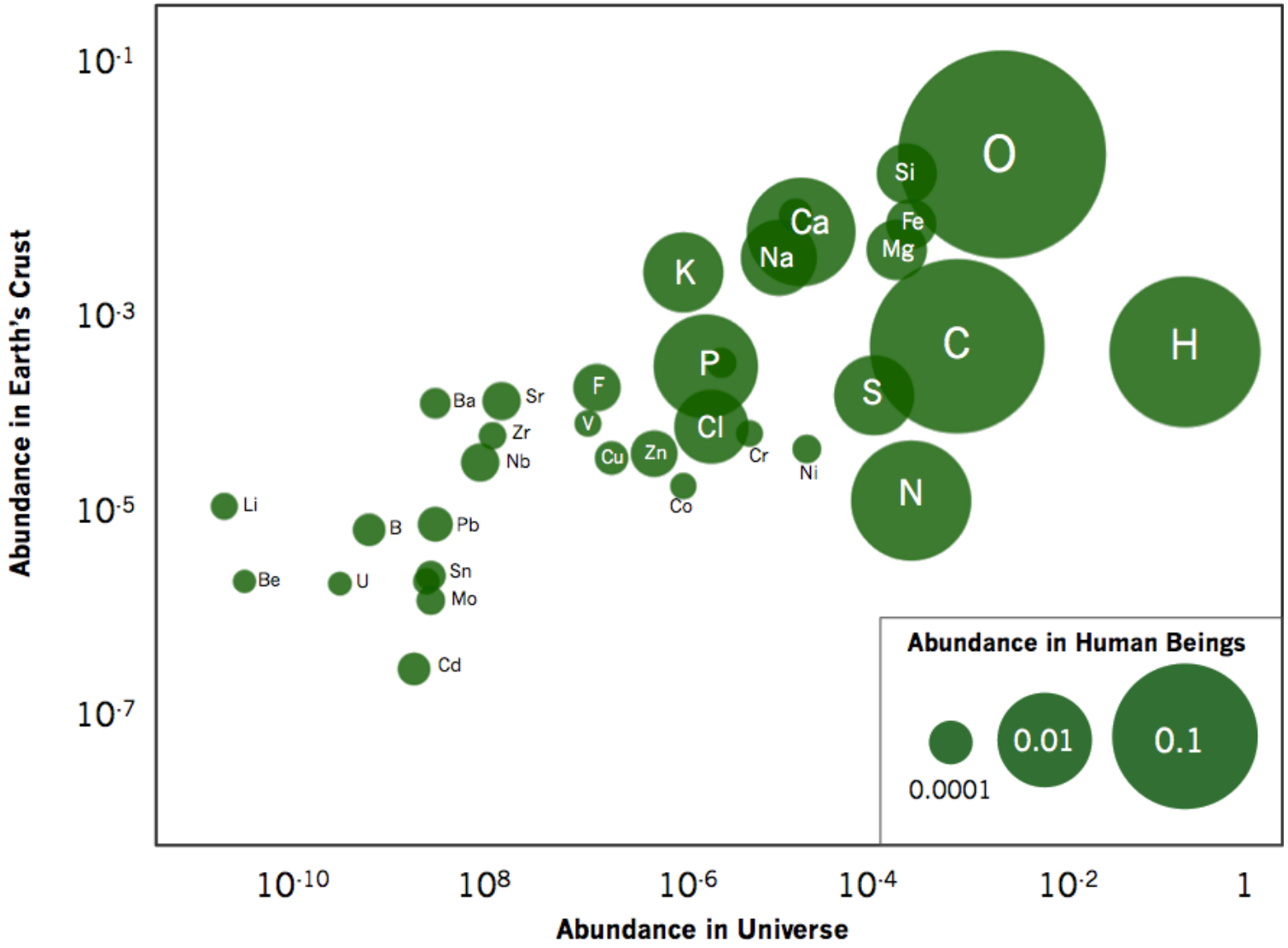


NHL Player Usage (Ottawa Senators)



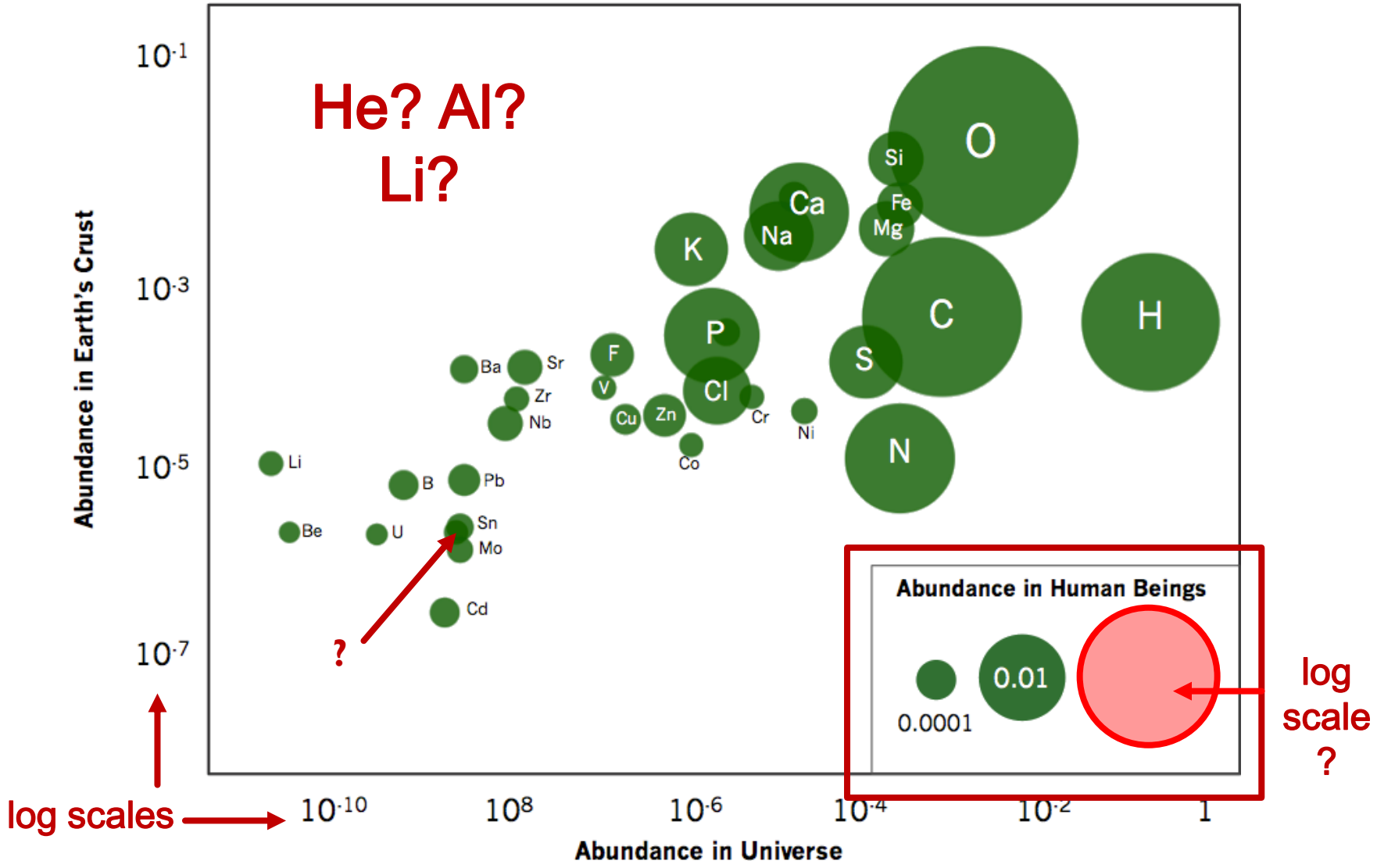
[M. Twardos, The Information Diet]

Abundance of Chemical Elements



[M. Twardos, The Information Diet]

Abundance of Chemical Elements



BUBBLE CHARTS

Colour + **geometry** allow us to plot (at least) 2 extra variables on a 2D scatter plot

May need to re-scale or bin the available data

A movie could be used to visualize an additional ordinal variable

Text can also be added to visualize an additional categorical variable

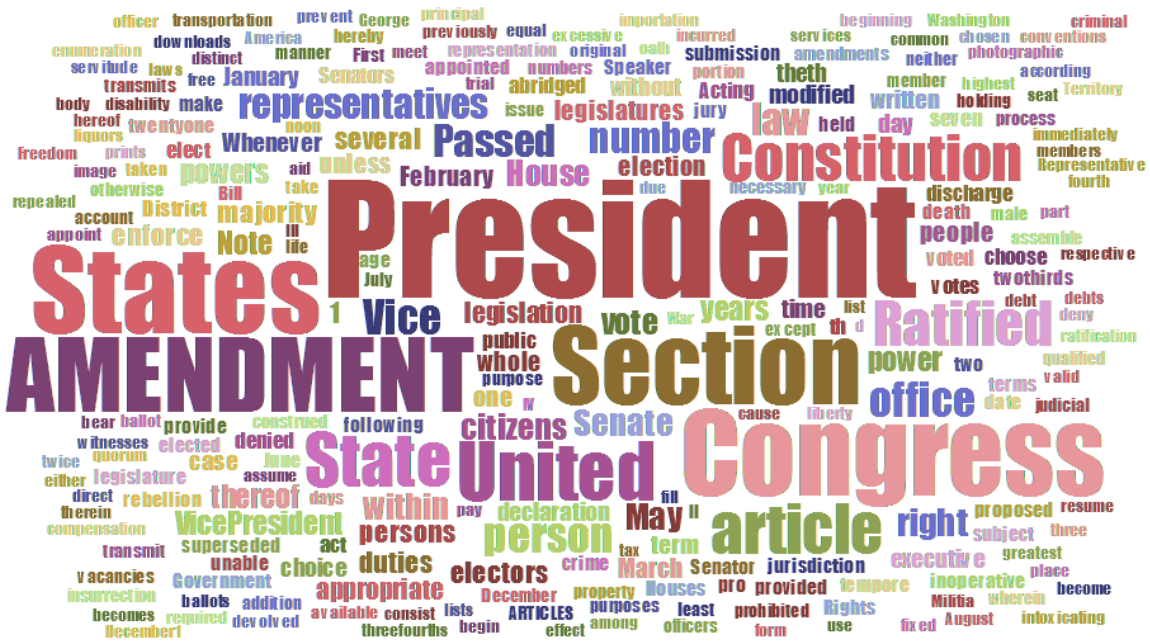
Works best when chart is **not too encumbered**

A **personal favourite** – a good mixture of traditional and modern features

[Created with J. Davies's WordCloud app]

WORD CLOUDS

U.S. Constitution vs. Canadian Charter of Rights



WORD CLOUDS

For maximal impact, font size should be a function of frequency

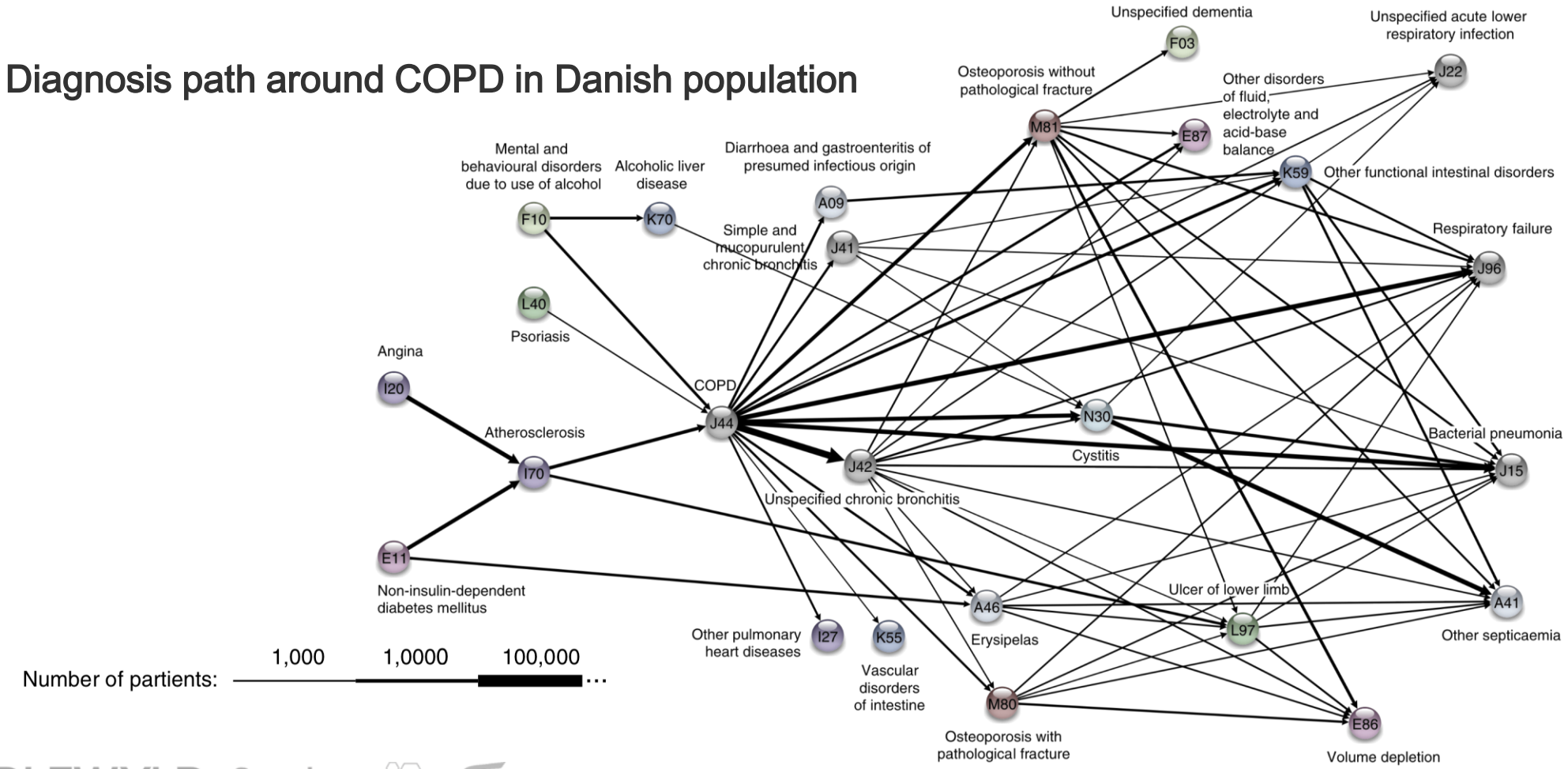
Typically used for univariate categorical data, but **small multiples**, **cloud shape**, **word placement**, **colour**, and **hue** could be used to integrate more variables

Word placement and colour choice algorithm are “hidden”

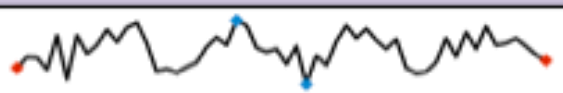
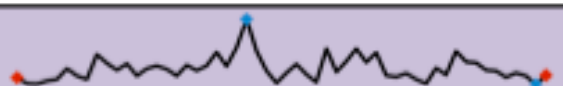
Could be used to answer authorship questions

[A.B. Jensen et al., *Nature Communications*]

NETWORK DIAGRAMS

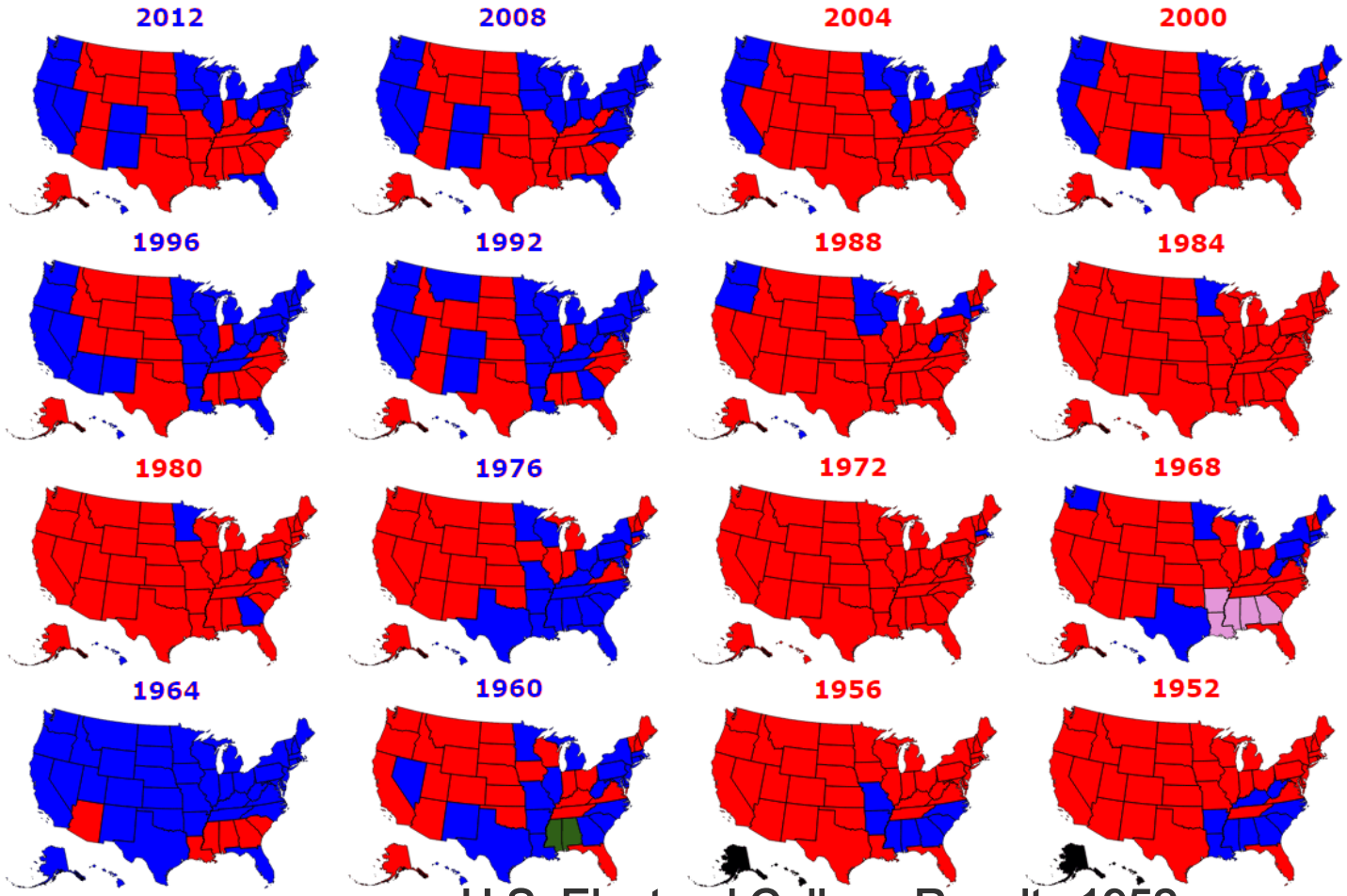


SPARKLINES AND SMALL MULTIPLES

	Start	Monthly Number of Cases	End	Low	High	Mean	Std Dev	Blanks	Zeros	Trend
Total	19502		17265	15150	25072	19903	2612	0.0	0.0	379.2
Hospital #1	46		19	3	46	19	9	0.0	0.0	-1.6
Hospital #2	156		240	101	326	194	60	0.0	0.0	9.7
Hospital #3	16		11	2	76	15	15	0.0	0.0	-2.9
Hospital #4	3		13	0	105	9	15	0.0	0.4	-1.8
Hospital #5	42		50	25	91	61	16	0.0	0.0	1.2
Hospital #6	48		53	34	169	67	25	0.0	0.0	0.6
Hospital #7	0		N.A.	0	0	0	0	2.2	9.8	0.0
Hospital #8	56		104	34	150	73	25	0.0	0.0	4.6

...

SMALL MULTIPLES



U.S. Electoral College Results 1952 – 2012

INTERACTIVE AND ANIMATED VISUALIZATIONS

Animation **does not always** improve a visualization.

What insights can interactivity provide? That depends on the data & on visualization.

Examples:

- The Clubs That Connect the World Cup, NY Times, 2014
- Who Marries Whom, Bloomberg, 2016
- Hipparcos Star Mapper, European Space Agency, 2016
- The Internet of Things – a Primer, Information is Beautiful, 2016
- The Genealogy and History of Popular Music Genres, Musicmap, 2016

INTERACTIVE AND ANIMATED VISUALIZATIONS

Examples (continued):

- Sequences Sunburst, Kerry Rodden, 2015
- Health and Wealth of Nations, Gapminder Foundation
- Mobius Transformations Revealed, Arnold D.N, Rogness, J, 2007
- Visualizing the Riemann Function and Analytic Continuation, 3Blue1Brown, 2016
- Small Arms and Ammunition – Imports and Exports, Google, 2012
- The Evolution of the Web, Google, Hyperakt, Vizzuality, 2012
- peoplemovin, Carlo Zapponi, 2012

DISCUSSION AND TAKE-AWAYS

"There is always a danger that if certain types of visualization techniques take over, the kinds of questions that are particularly well-suited to providing data for these techniques will come to dominate the landscape, which will then affect data collection techniques, data availability, future interest, and so forth."

(P. Boily)

Take-Aways:

- explore the data
- try different methods

HALL-OF-FAME / HALL-OF-SHAME

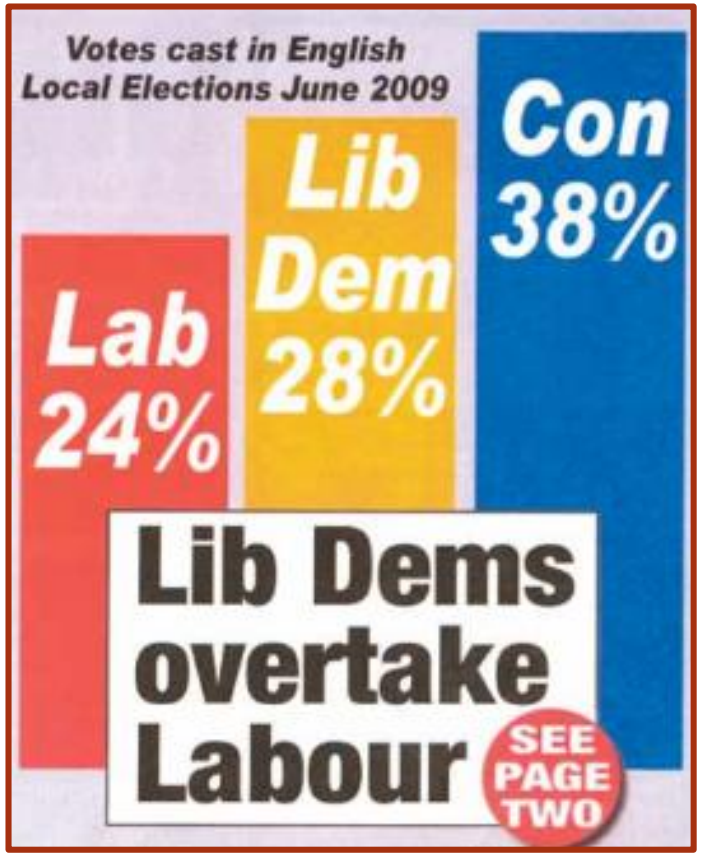
DATA EXPLORATION AND DATA VISUALIZATION

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

MISLEADING CHARTS

[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.

WHAT TO WATCH FOR

Several ways to quantify the misleading level of a chart:

- **Lie factor:** ratio of size of the effect shown on the graph by the size of the effect in the data
- **Data density:** number of observations by chart area
- **Chartjunk ratio:** ratio of area required to convey the data insight by chart area

Typically, the lie factor and chartjunk ratios should be close to 1, while the data density should be “high” (within reason).

YOU BE THE JUDGE

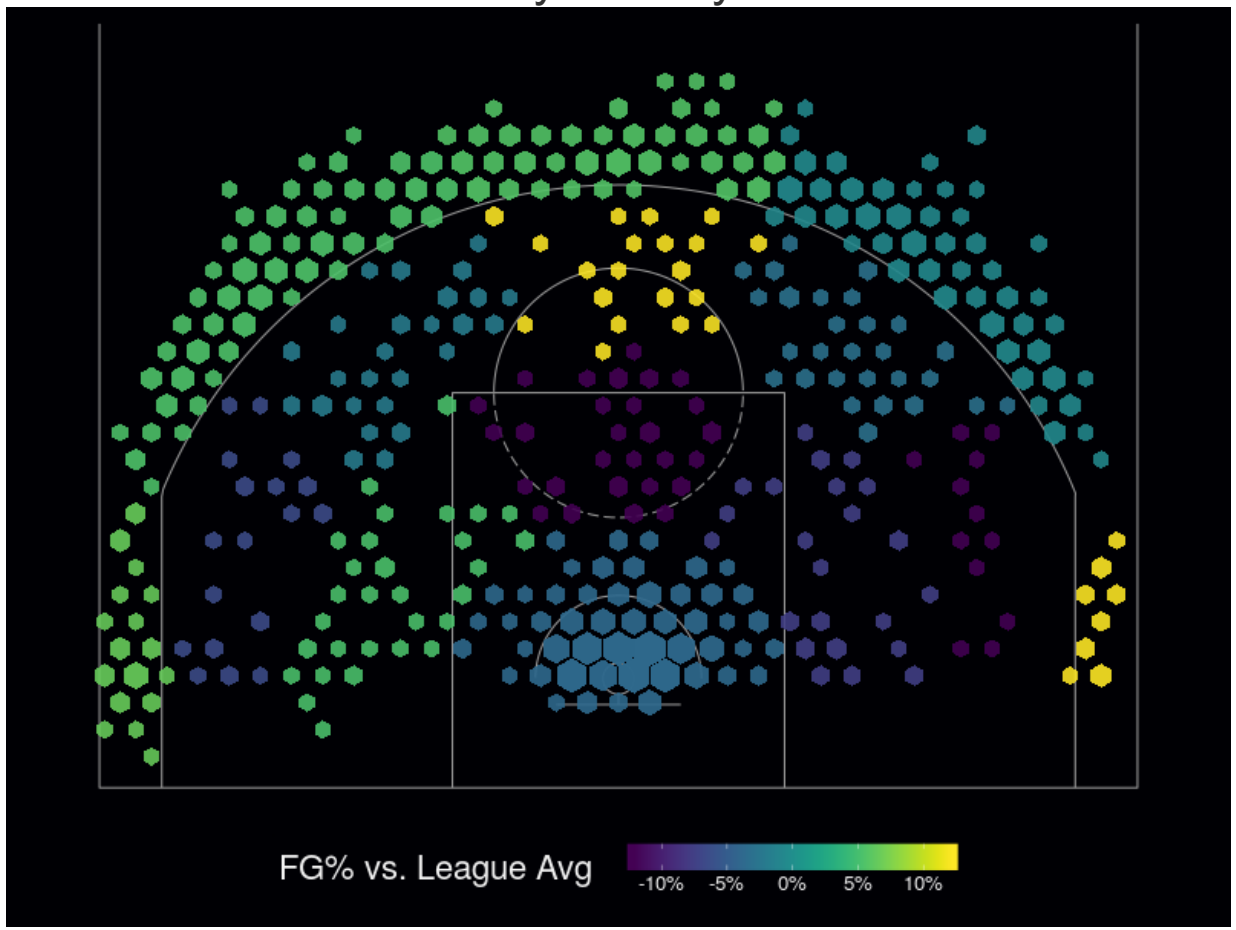
Some of the following are (arguably) good visualizations. But some are not!

Which are which? You be the judge…

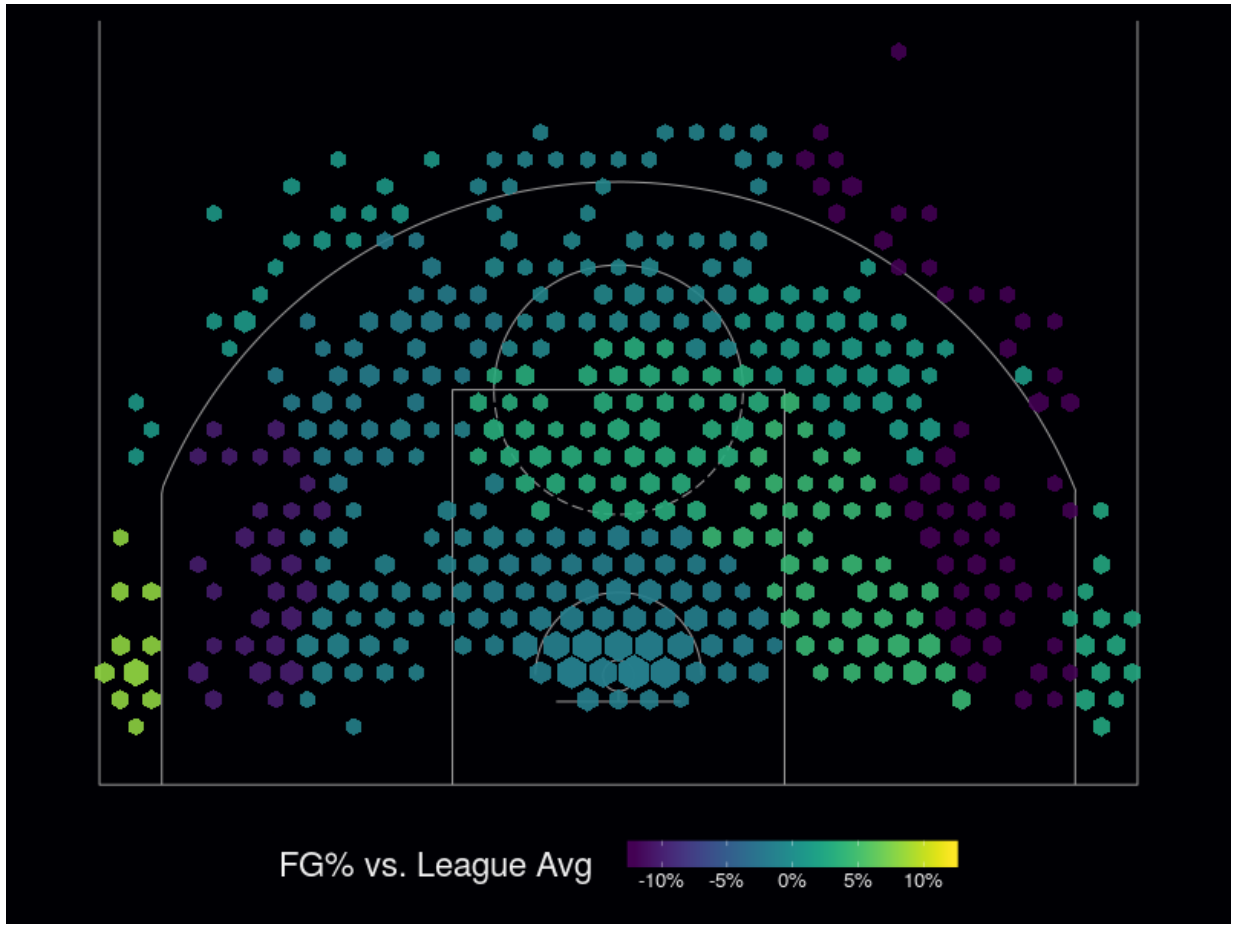
NBA FG% Against League Average (‘15–’16)

[T.W. Schneider]

Kyle Lowry



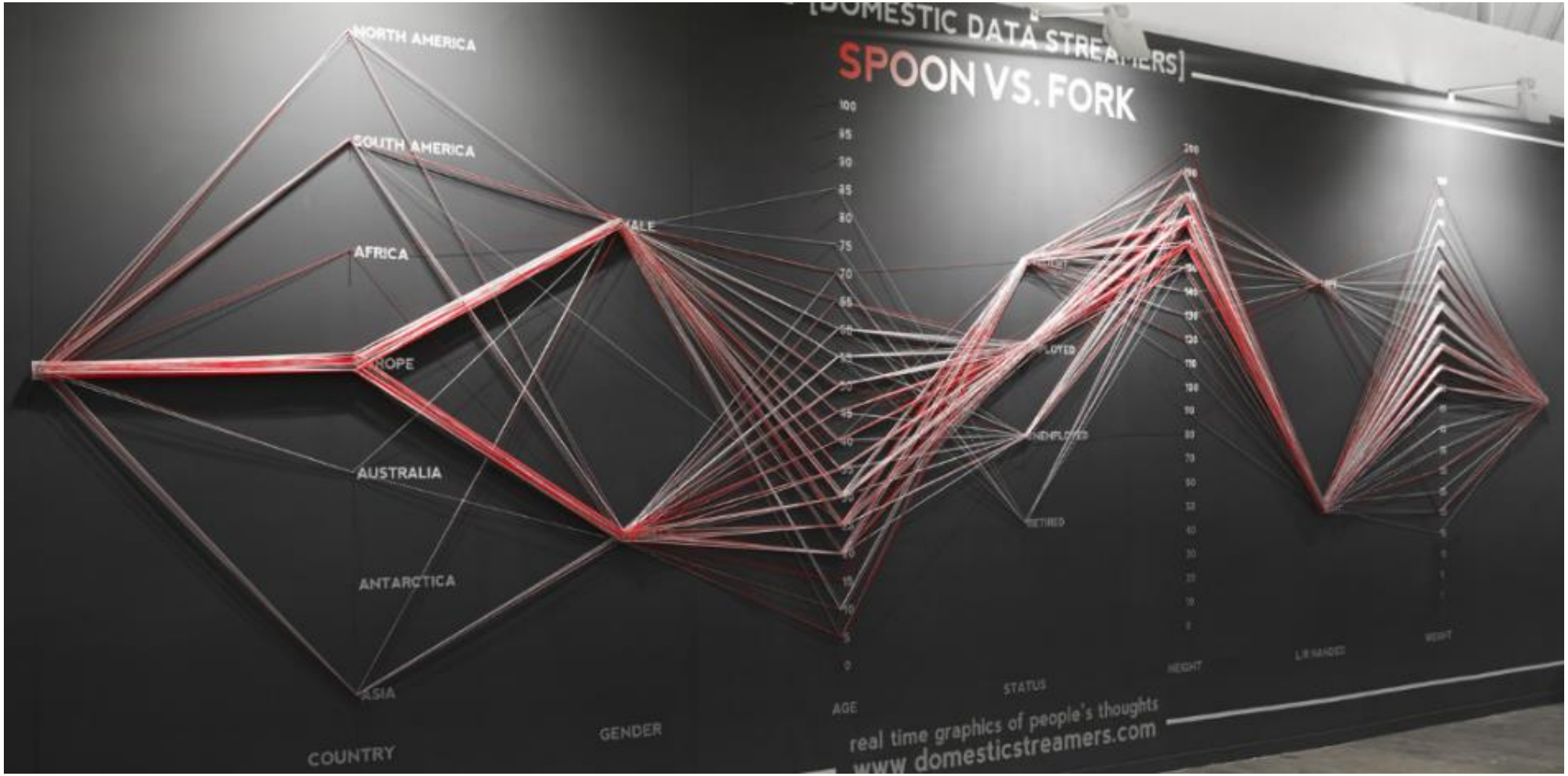
DeMar DeRozan



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

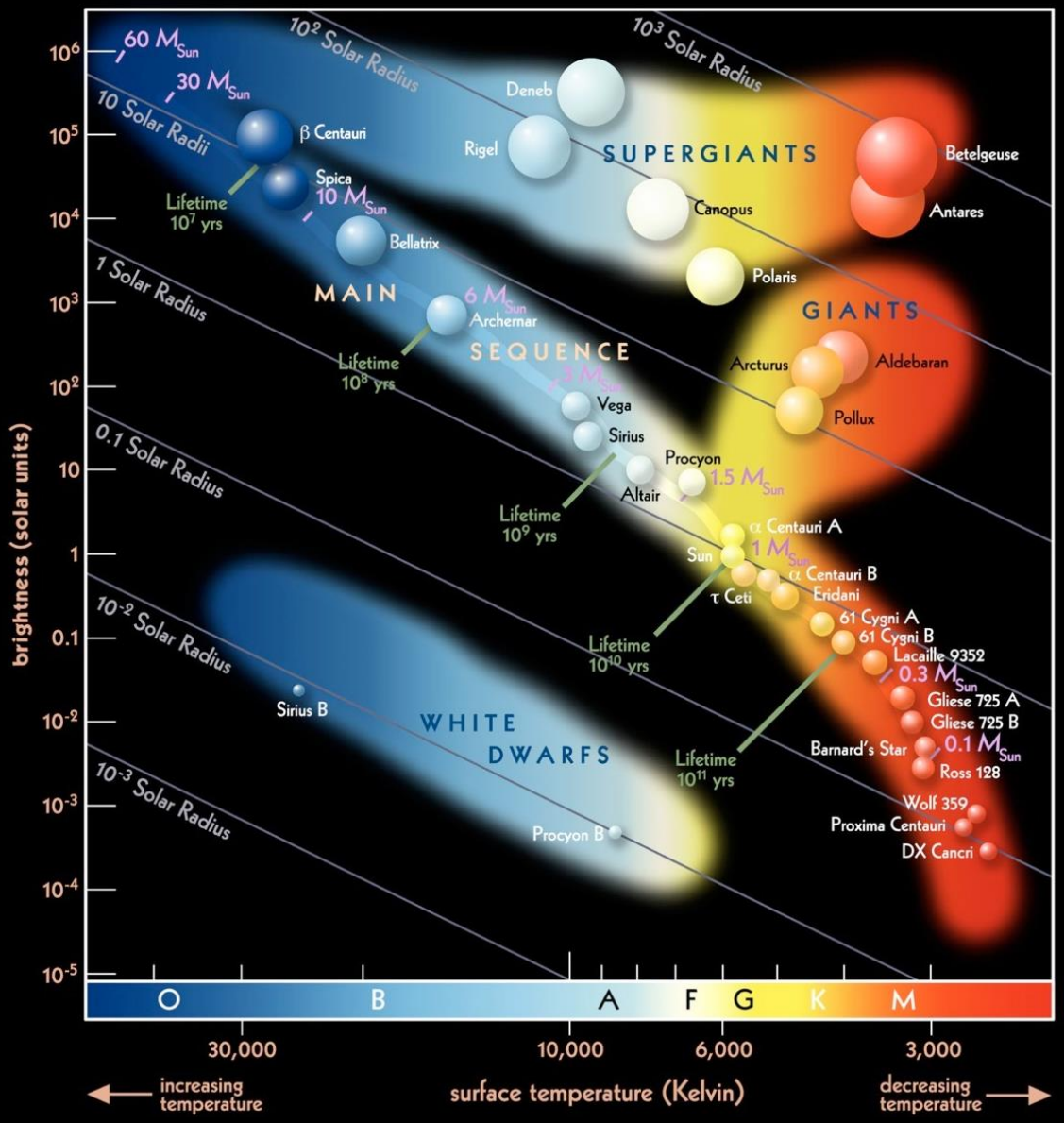
[domesticstreamers.com]

Spoon vs. Fork



Are there any issues with data collection? Where do you think this event took place? Is the spoon/fork question a red herring?

Hertzprung-Russell Diagram



[European Southern Observatory]

Data Elements

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

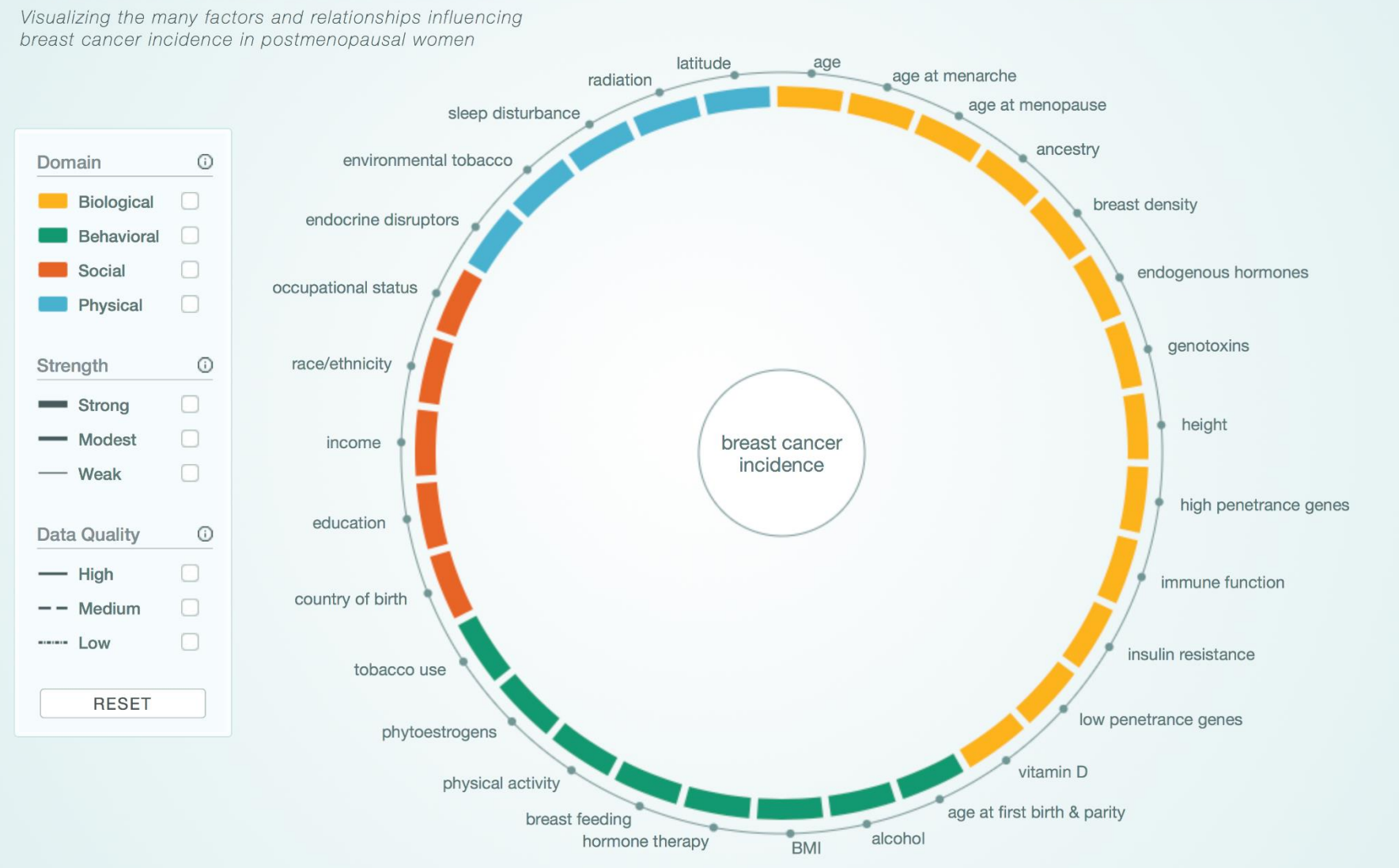
Underlying Structure

- 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 diagram.

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

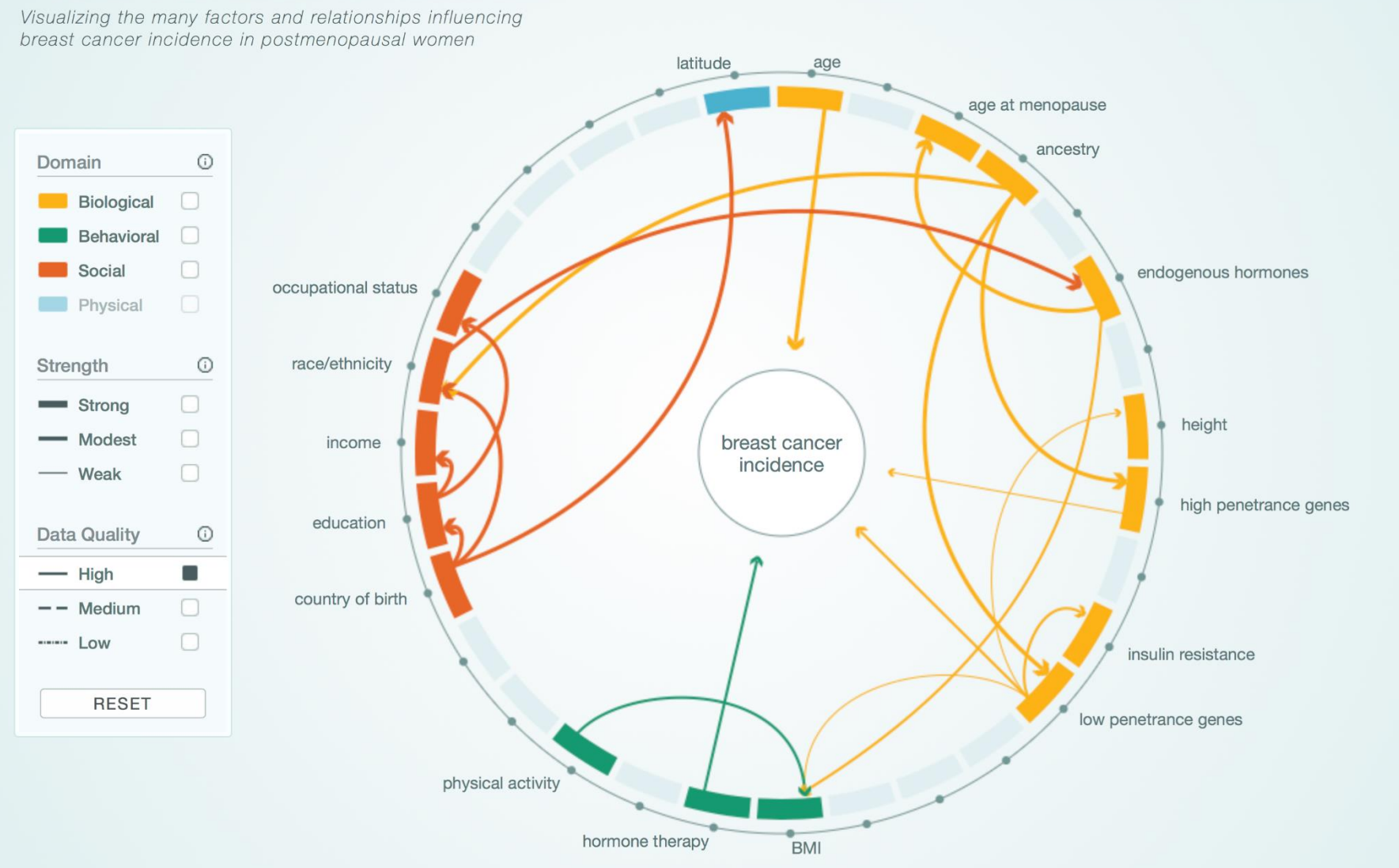
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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

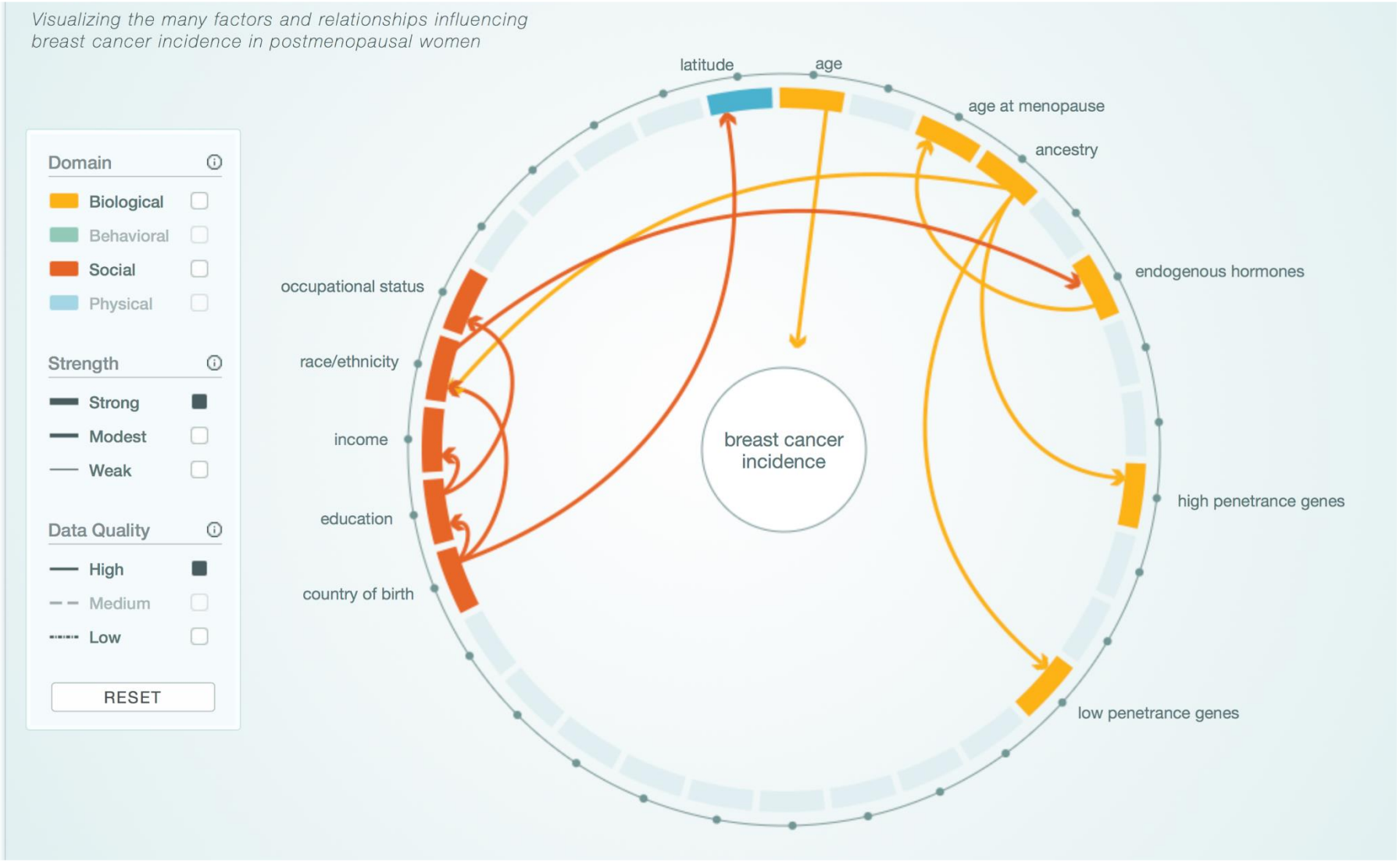
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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

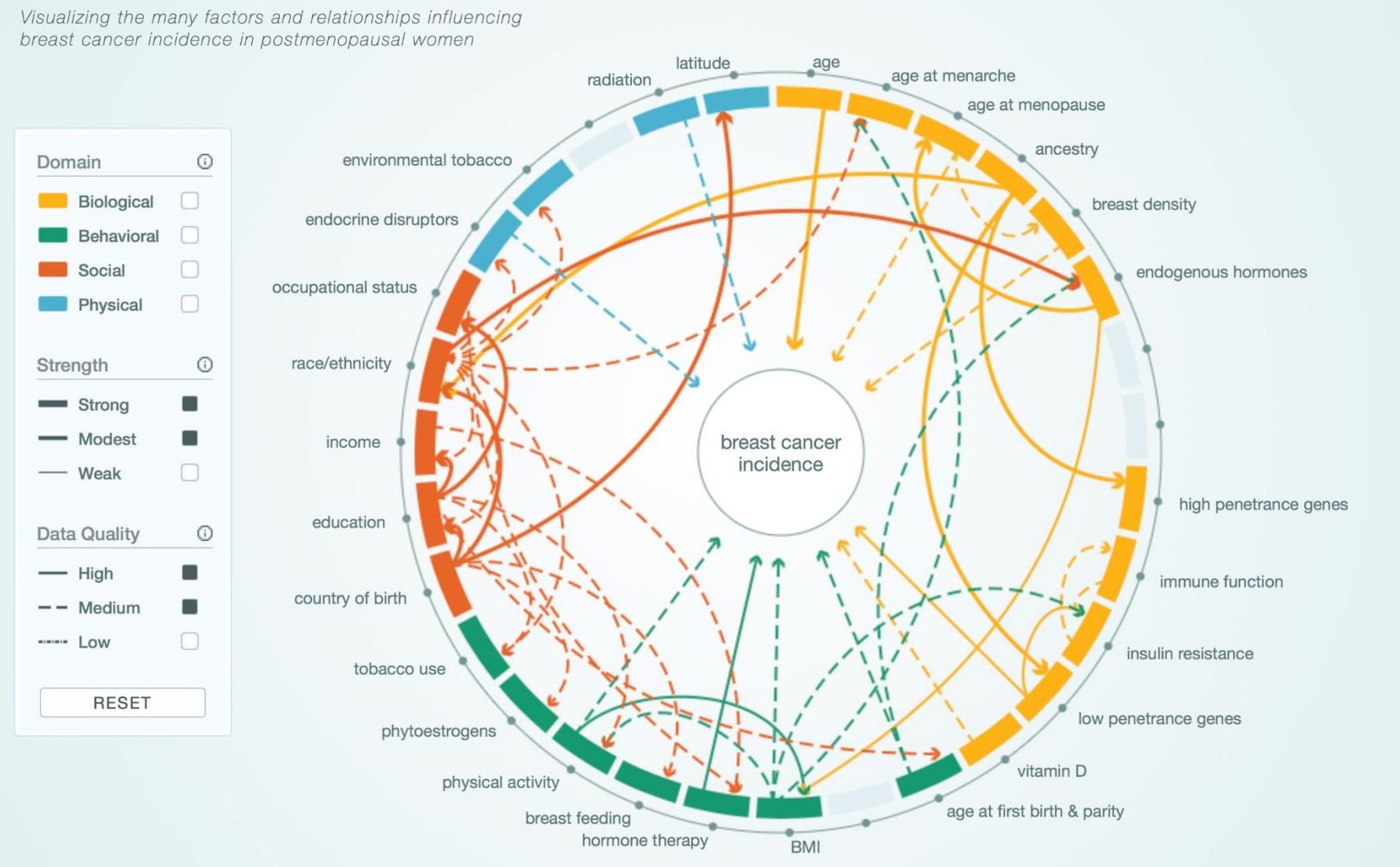
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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

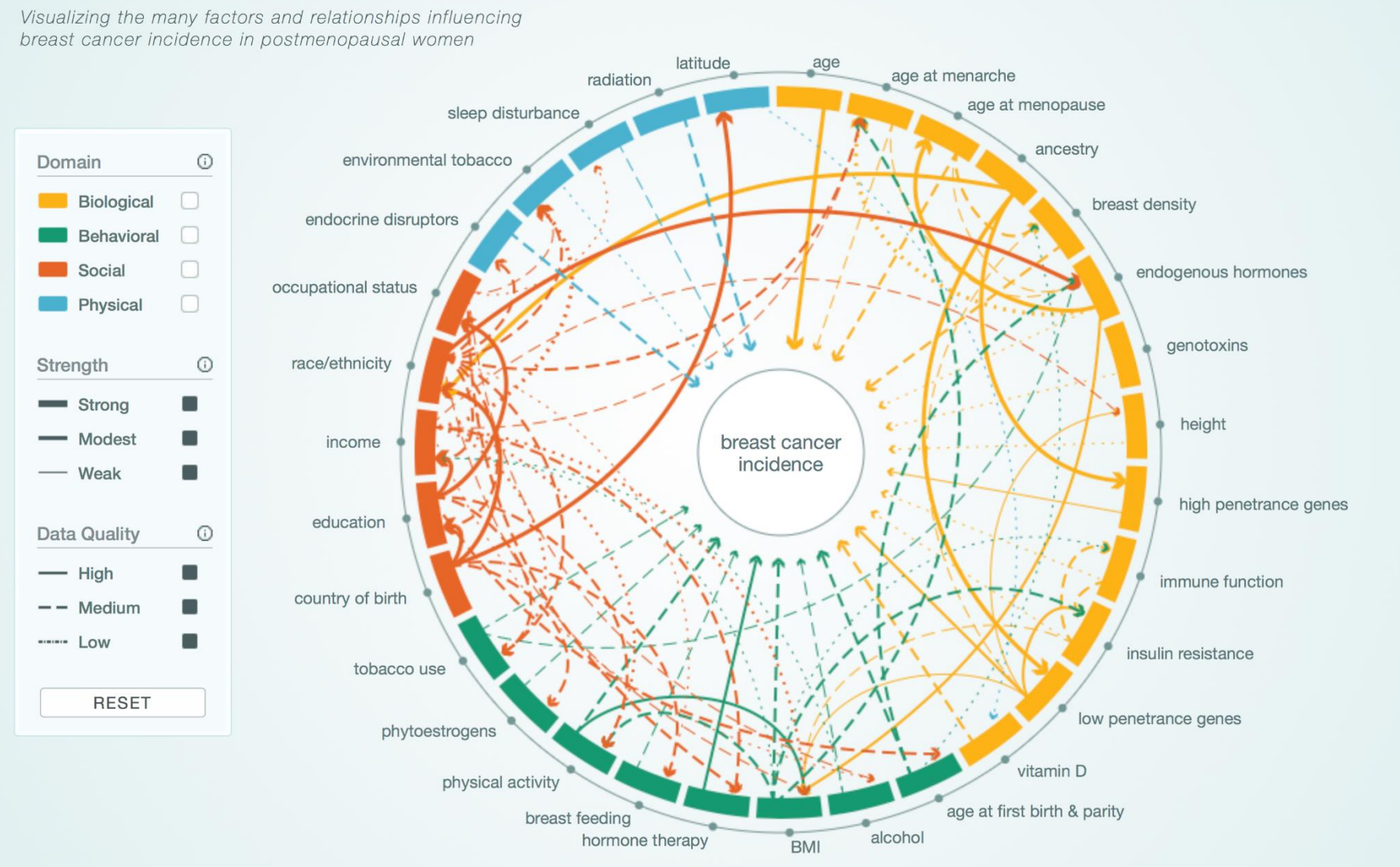
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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

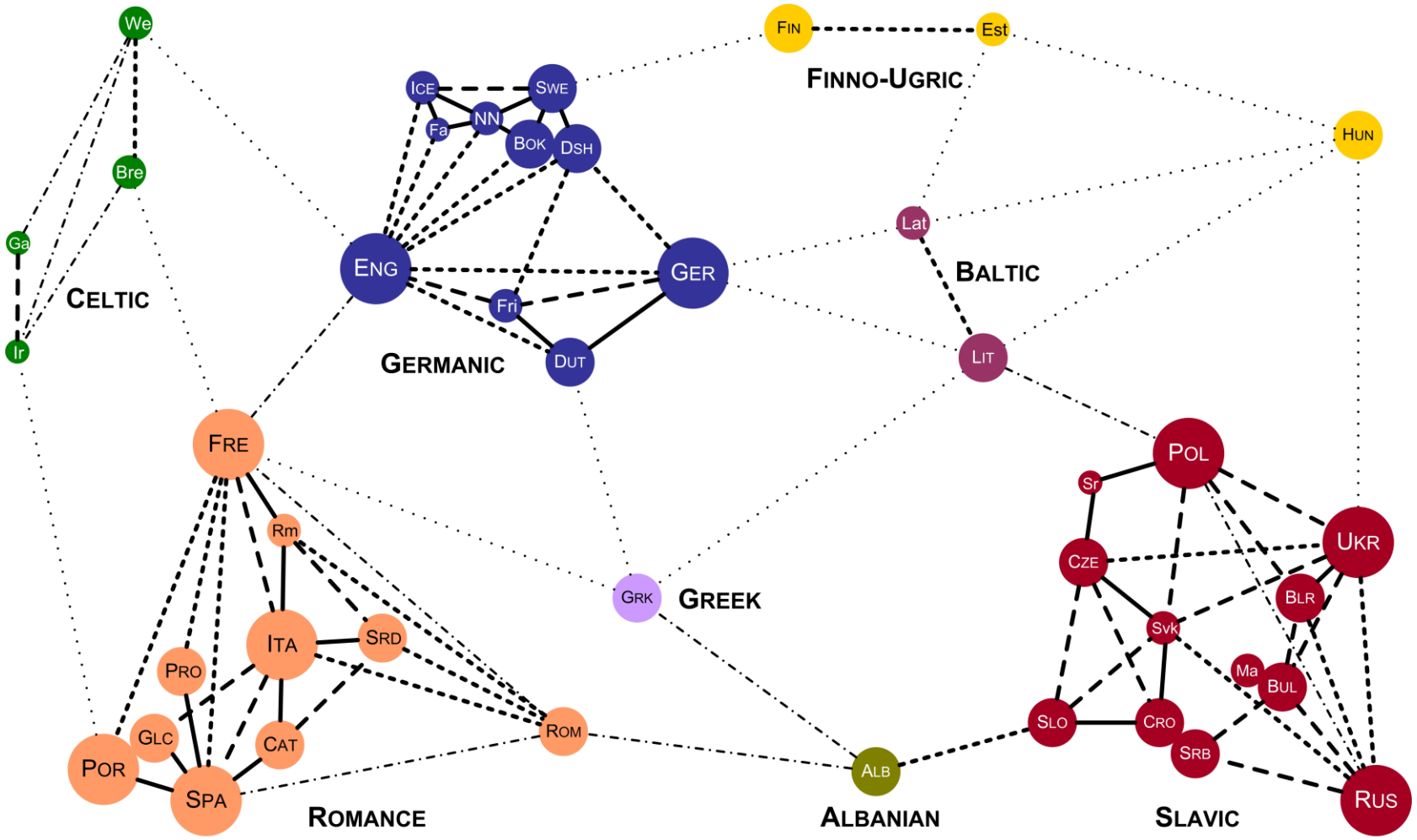
A Model of Breast Cancer Causation

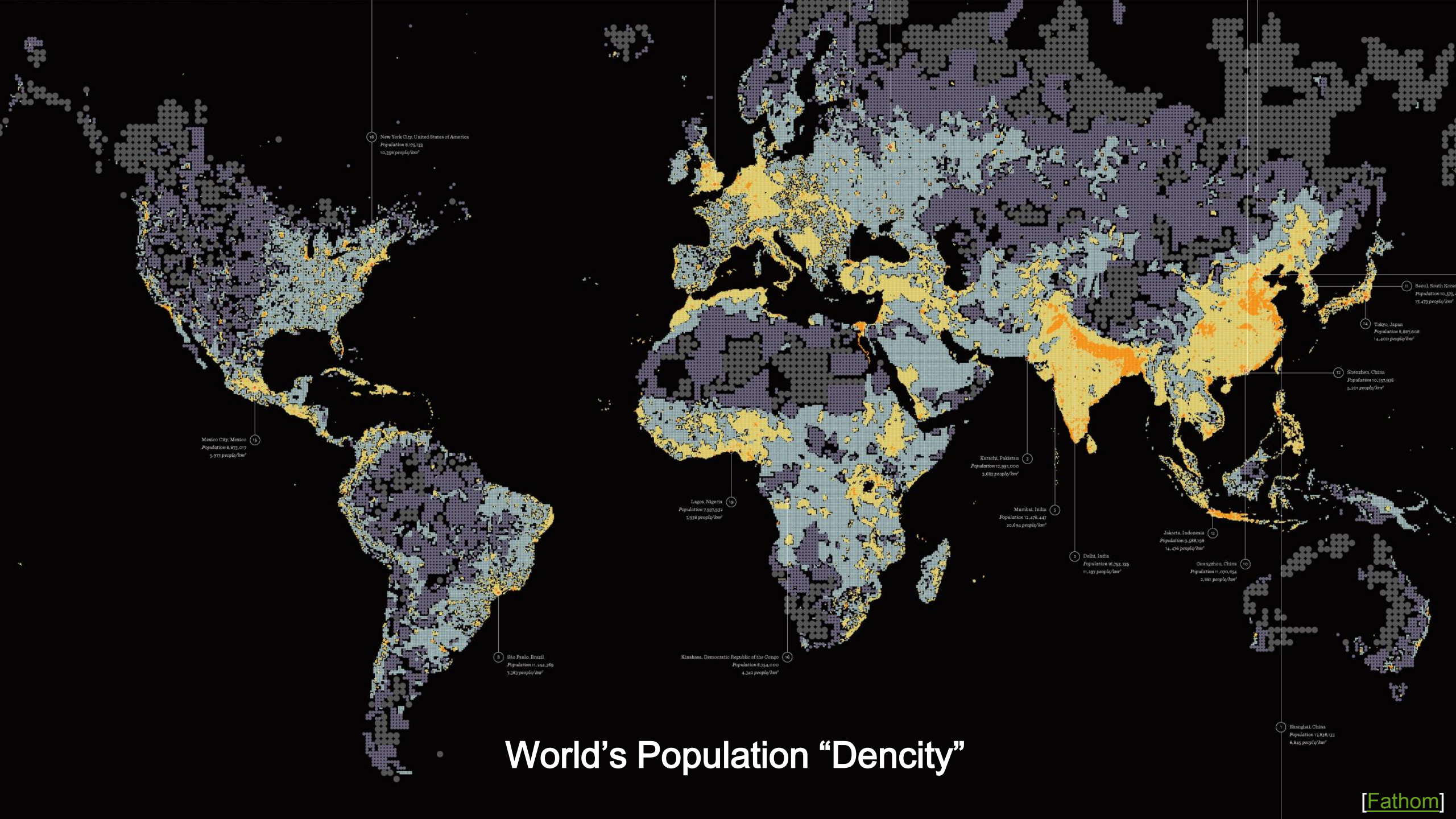


Can you infer causality from this diagram?

[T.Elms, Etymologikon]

Lexical Distance Among the Languages of Europe





World's Population "Dencity"

[Fathom]

[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]

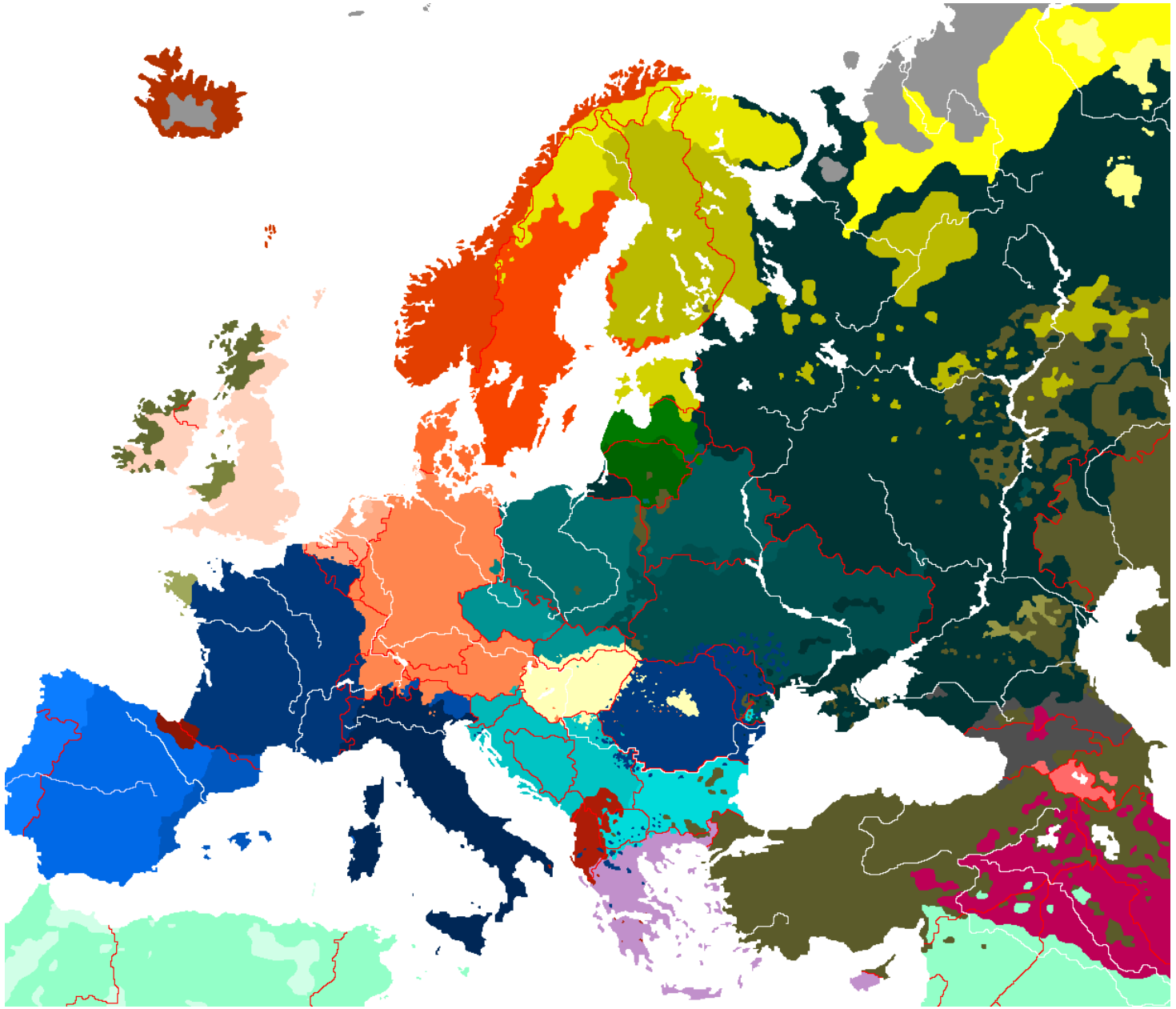
Encoding?

Population density?

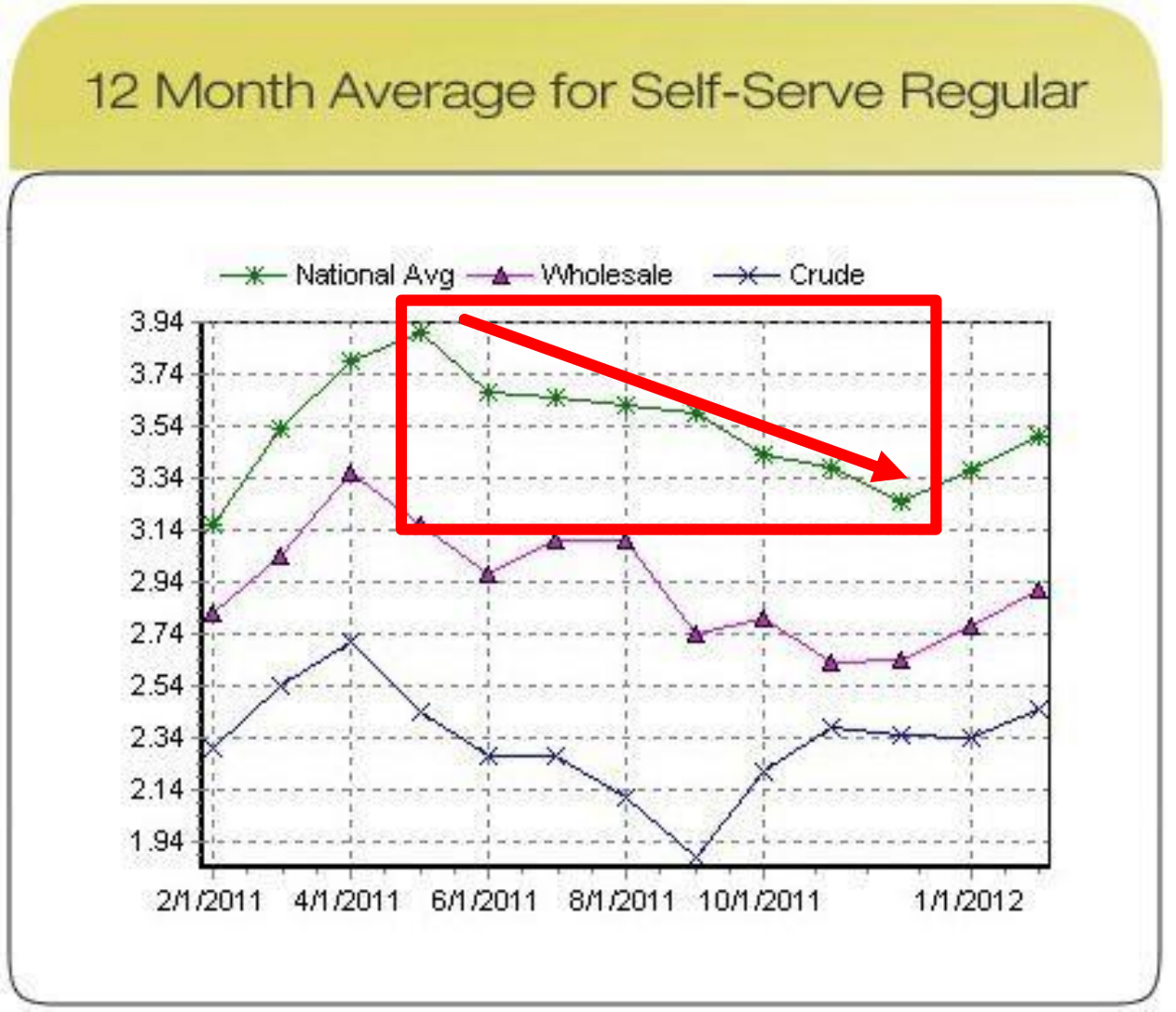
Secondary languages?

Rivers?

No data source



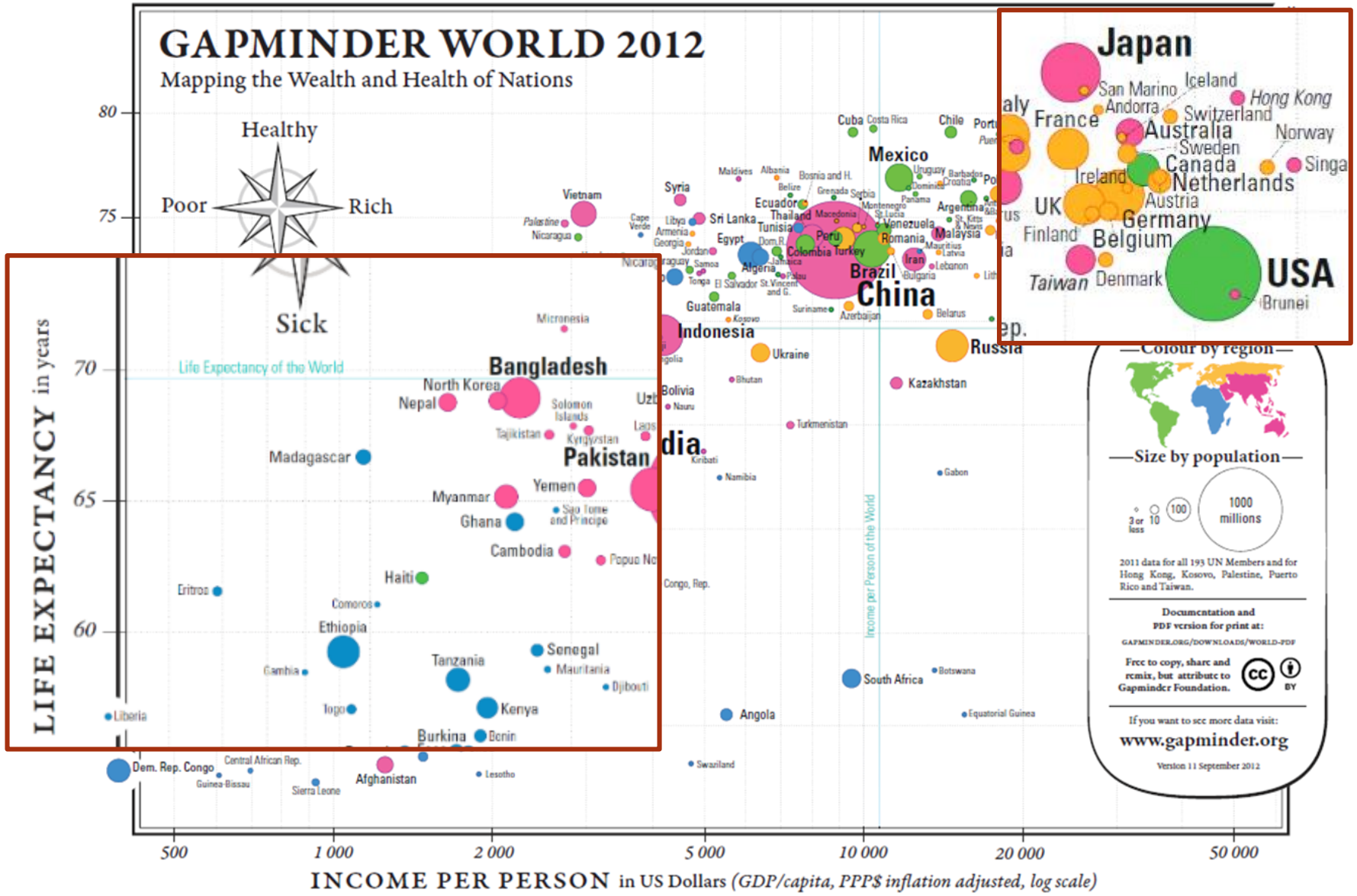
[AAA Gas Prices, found on mediamatters.org]



SUPPLEMENTAL MATERIAL AND EXERCISE ANSWERS

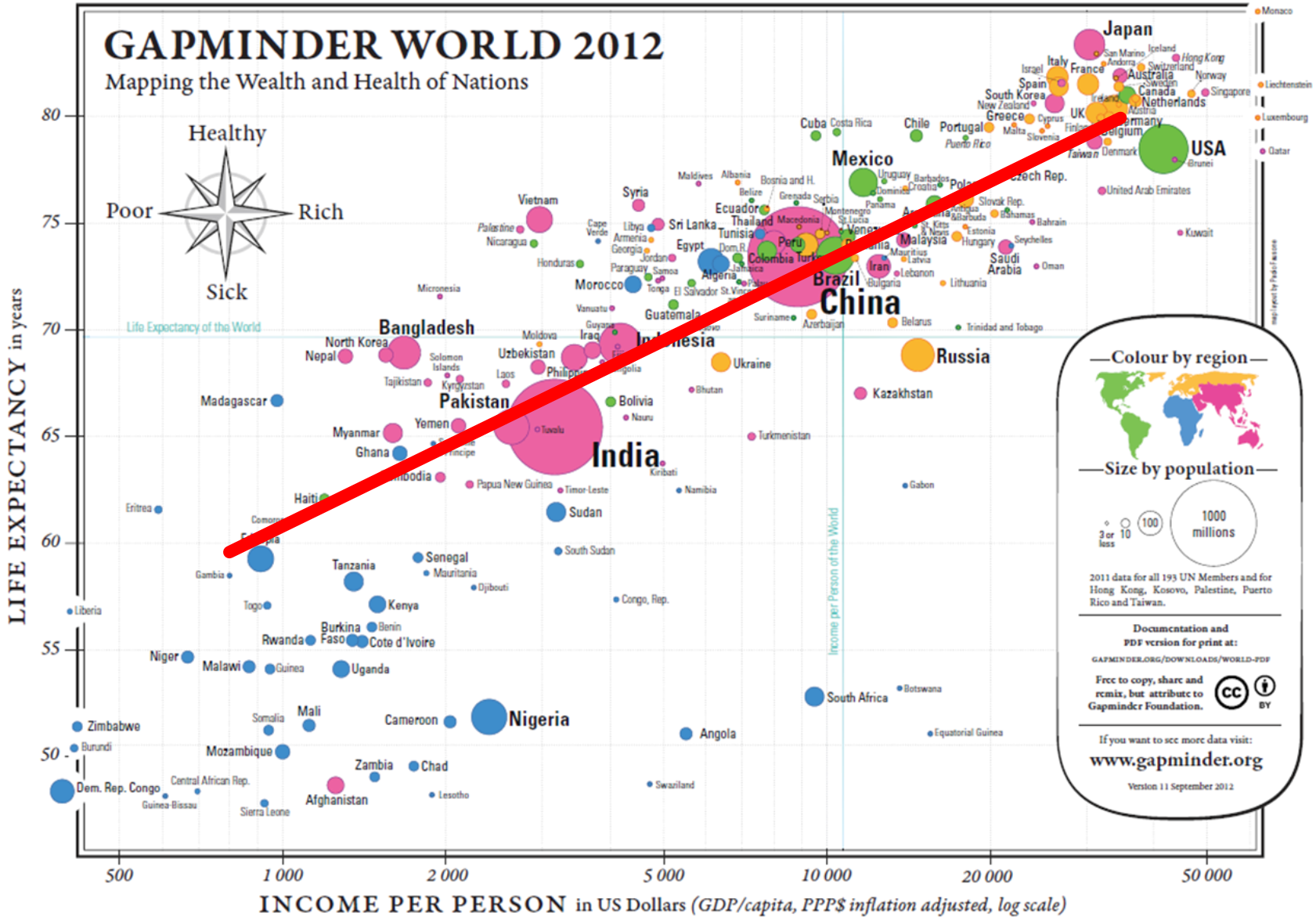
[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]

Meaningful Comparisons



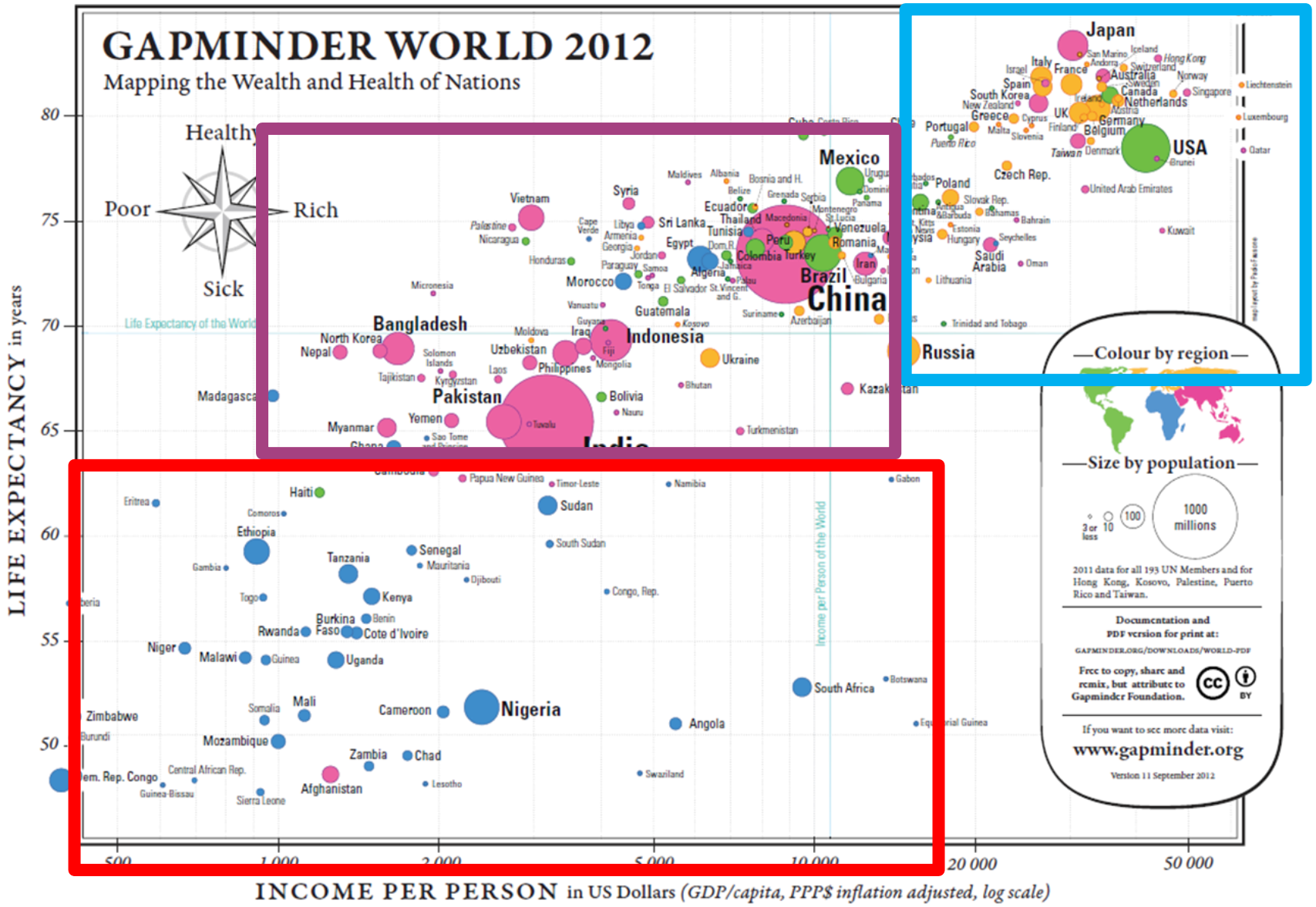
[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]

Underlying Structure and Multivariate Links



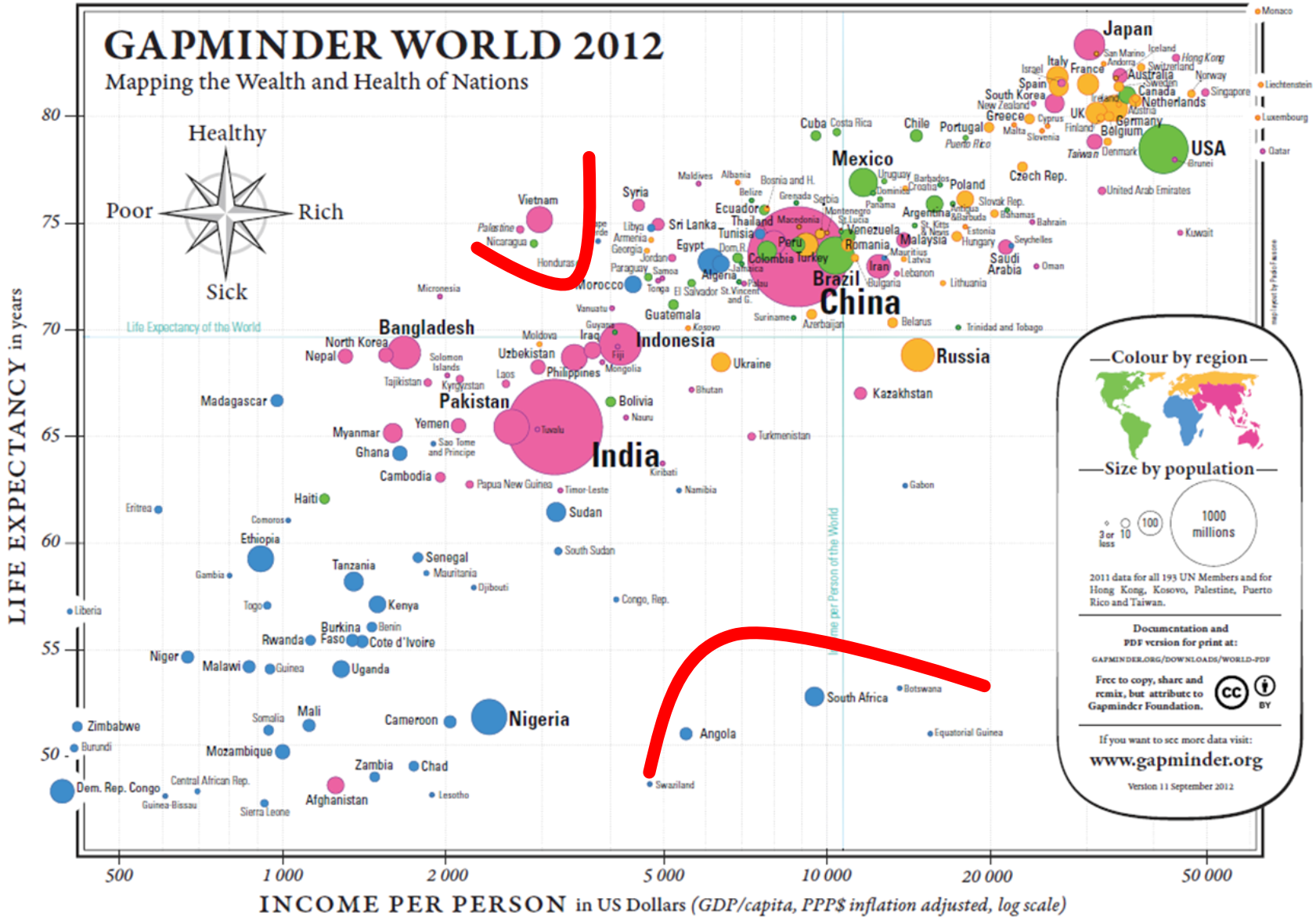
[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]

Underlying Structure and Multivariate Links



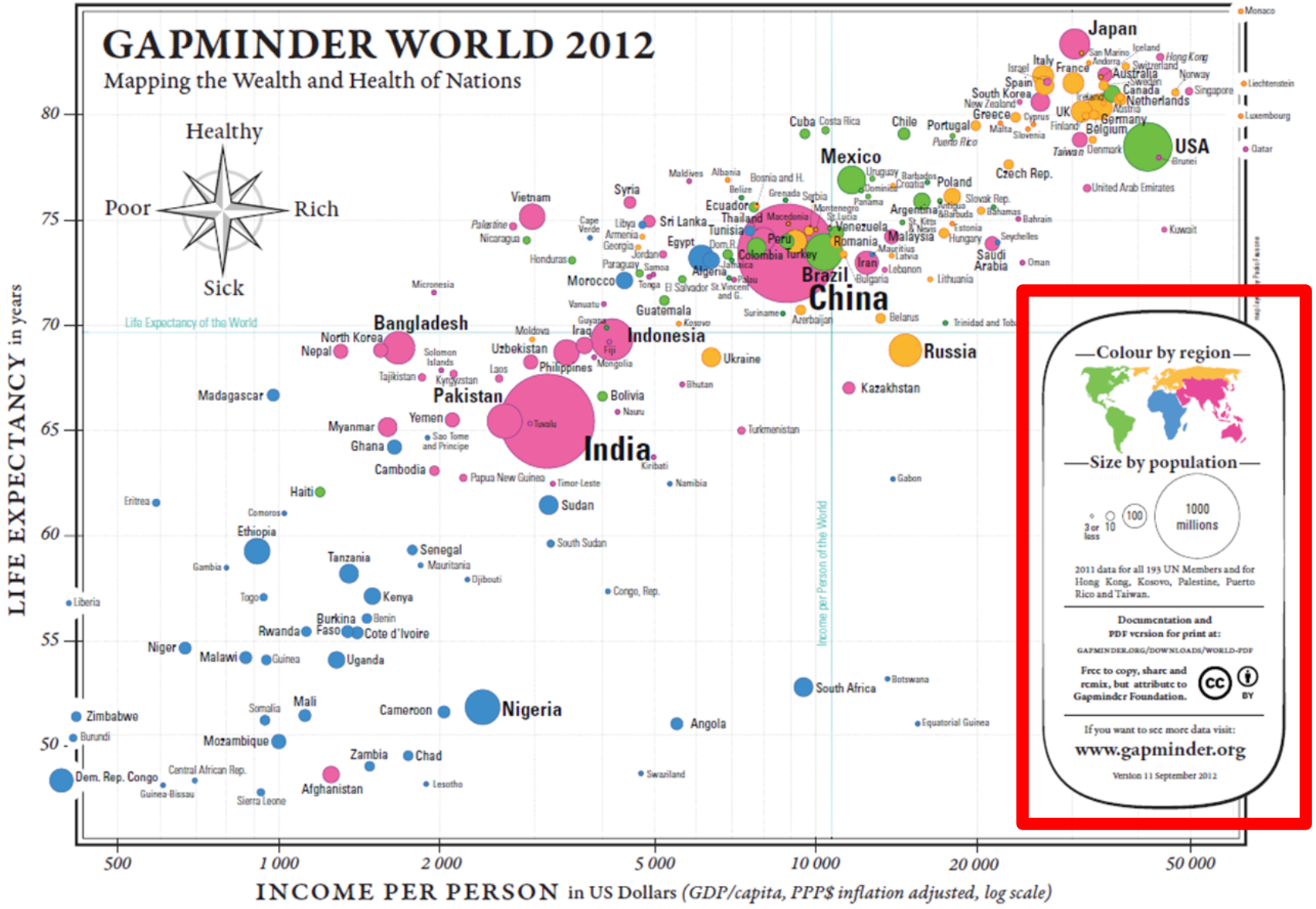
[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]

Underlying Structure and Multivariate Links



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

Documentation



HALL OF FAME/HALL OF SHAME CHARTS

Here's our judgment – do you agree?

- Cost of Gas: Thumbs down
- Lib Dems Overtake Labour: Thumbs down
- NBA FG% Against League Average ('15–'16): Thumbs up
- Spoon vs. Fork: Partial thumbs up
- Hertzprung-Russell Diagram: Thumbs up
- A Model of Breast Cancer Causation: Thumbs up
- Lexical Distance Among the Languages of Europe: Partial thumbs up
- World's Population "Dencity": Thumbs up
- Mapping Paid Paternity Leave: Thumbs down
- Mystery Colour Map: Partial thumbs down
- 12 Month Average For Self Serve Regular: Thumbs up

EXERCISES, READINGS, AND REFERENCES

DATA EXPLORATION AND DATA VISUALIZATION

EXTRA EXERCISES

Run the provided Jupyter Notebooks to see how visualizations can be generated programmatically

Read through the ggplot2 primer – ggplot2 is a well known R visualization library

See the resources channel on Slack for more visualization resources!

REFERENCES

Understanding Graphics

Krygier, J., Wood, D., [2016], *Making Maps: A Visual Guide to Map Design for GIS*, Guilford Press

Interactive Data Visualization on Wikipedia

Is animation an effective tool for data visualization?, NASA

Perception in Visualization, C.G. Healey (very cool!)

Data Physicalizations

Tufte, E. [2001], *The Visual Display of Quantitative Information*, Graphics Press.

Hu, D. [1954], *How to Lie With Statistics*, Norton

Tufte, E. [2008], *Beautiful Evidence*, Graphics Press

REFERENCES

Nussbaumer Knaflic, C. [2015], *Storytelling with Data*, Wiley

Cairo, A. [2013], *The Functional Art*, New Riders

Cairo, A. [2016], *The Truthful Art*, New Riders

Meireilles, I. [2013], *Design for Information*, Rockport

50 Great Examples of Data Visualization: http://www.webdesignerdepot.com

Visualising Data

Nathan Yau's FlowingData

Data Visualization on Wikipedia

Misleading Graphs on Wikipedia

REFERENCES

Prabhakaran, S., Top 50 ggplot2 Visualizations (with Master List R Code).

Miller, M. [2017], *The problem with Interactive graphics*, Co.Design

Wickham, H. [2016], *ggplot2: Elegant Graphics for Data Analysis* (2nd ed), Springer.

Gorelik, B., Data Visualization (blog).

Chang, W. [2013], R Graphics Cookbook, O'Reilly.

Wickham, H. [2009], A Layered Grammar of Graphics, Journal of Computational and Graphical Statistics 19:3–28.

Horton, N.J., Kleinman, K. [2016], Using R and RStudio for Data Management, Statistical Analysis, and Graphics, 2nd ed., CRC Press.

Healey, K. [2018], Data Visualization: A Practical Introduction.

REFERENCES

Kabacoff, R.I. [2011], R in Action, Second Edition: Data analysis and graphics with R, Live.

Maindonald, J.H. [2008], Using R for Data Analysis and Graphics: Introduction, Code and Commentary.

Tyner, S., Briatte, F., Hofmann, H. [2017], Network Visualization with ggplot2, The R Journal, vol. 9(1).

Broman, K. [2016], Data Visualization with ggplot2.

Robinson, D., Visualizing Data Using ggplot2, on varianceexplained.org.

Manipulating, analyzing and exporting data with tidyverse, on datacarpentry.org.

Wickham, H. [2014], Tidy Data, Journal of Statistical Software, v59, n10.

Gashim, E., Boily, P. [2018], A ggplot2 Primer, data-action-lab.com