DATA EXPLORATION AND DATA VISUALIZATION

CONCEPTS AND NOTIONS

DATA IN THE 20TH CENTURY

In the 20th century, data problems were mostly related to

- **engineering** (design of machines)
- **sciences** (formulation of theories)

Problems were solved **empirically**, **theoretically**, or through **computation**.

DATA IN THE 20TH CENTURY

Engineers equipped machines with sensors ⇒ used data to assess if the machines behaved as expected & to improve designs.

Scientists set up experiments ⇒ used data to test the validity of theories.

- Experiments are expensive; relatively few data points are generated.

Data contained additional information which is often ignored.

- Example: Mendel's experimental data, analyzed by Fisher, found to be too good to be true.

DATA IN THE 21ST CENTURY

In the 21st century, there is:

- there is **more data**
- it's mostly **digital**
- it's mostly **observed** (rather than generated by designed experiment)

Problems are solved **empirically**, **theoretically**, through **computation** and/or **data exploration/visualization**.

DATA IN THE 21ST CENTURY

Empirically: observe and describe what happens

Theoretically: generalize and build models and generalizations to understand what happens

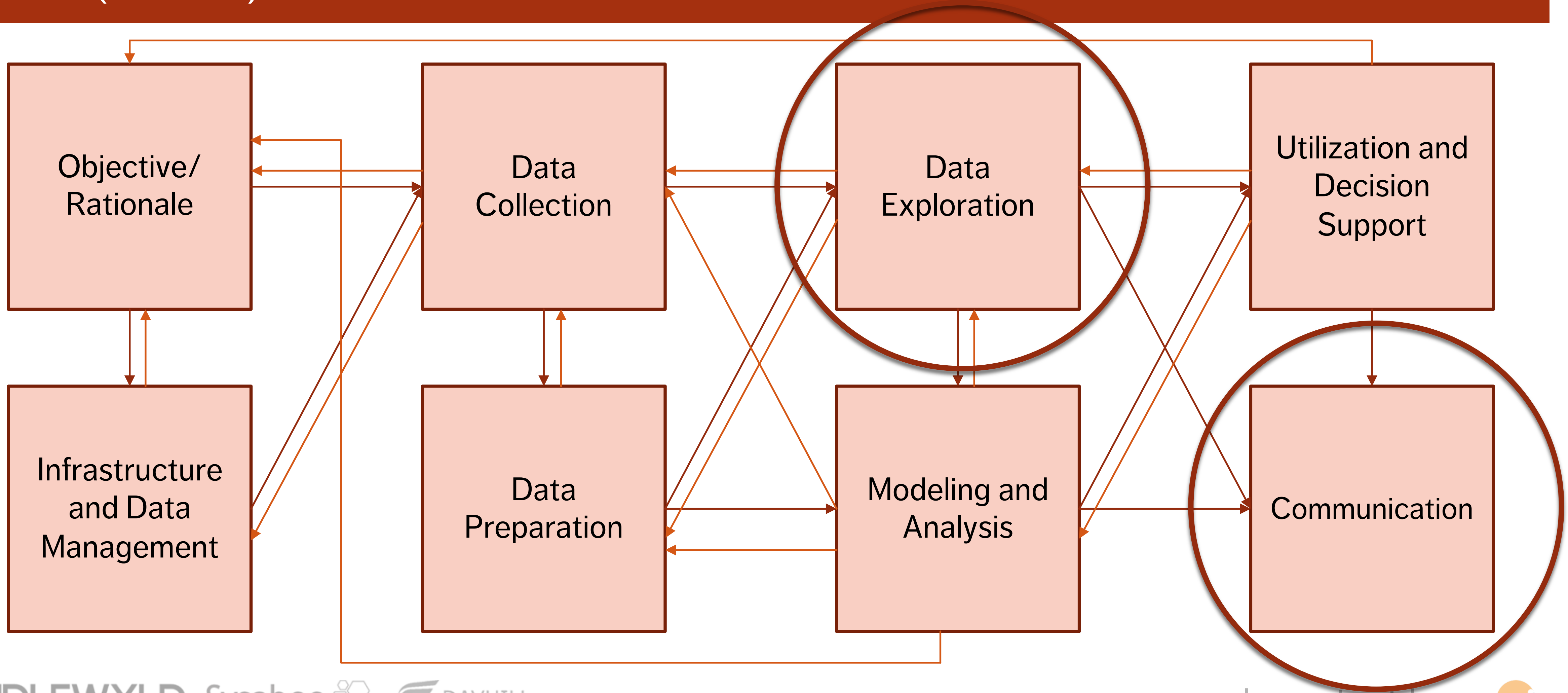
Computationally: design computer simulations to better understand what happens

Data Exploration/Visualization: the new approach to understanding

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

@DamianMingle

THE (MESSY) ANALYSIS PROCESS



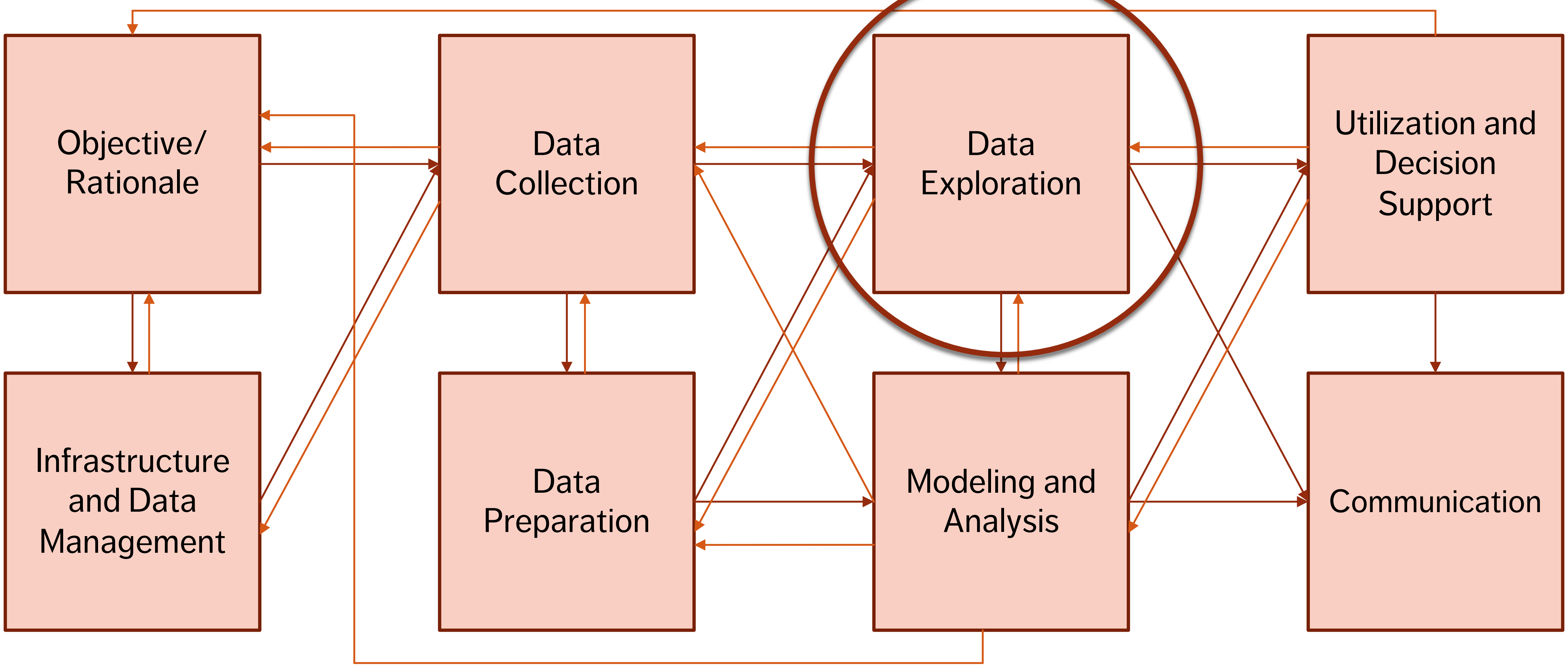
OUTLINE

1. Data Exploration
2. Pre-Analysis Data Visualization
3. Post-Analysis Data Visualization
4. Visualization Catalogue
5. Hall-of-Fame / Hall-of-Shame
6. The Grammar of Graphics
7. An Introduction to Dashboards

DATA EXPLORATION

DATA EXPLORATION AND DATA VISUALIZATION

THE (MESSY) ANALYSIS PROCESS



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

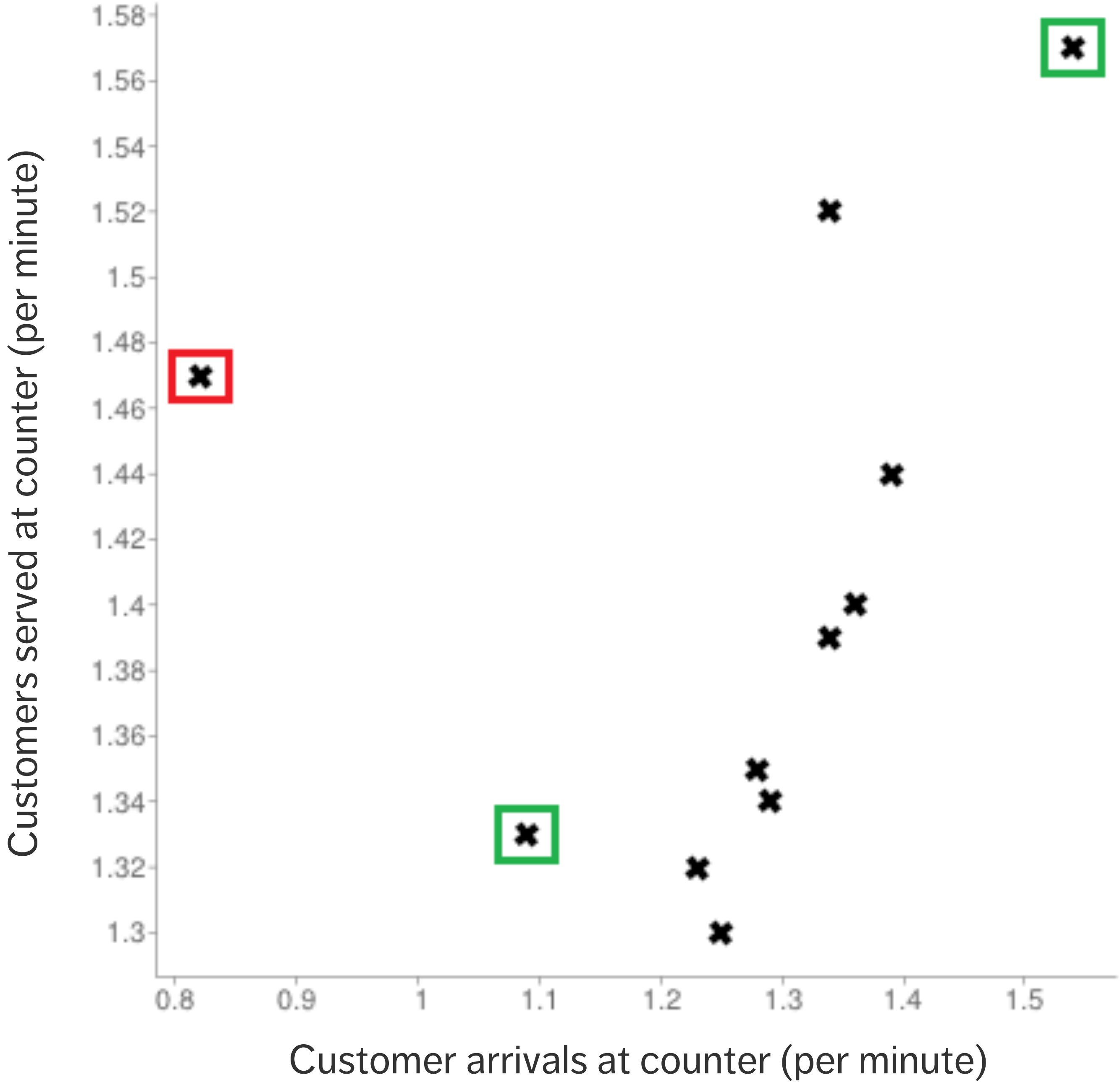
DATA EXPLORATION AND DATA VISUALIZATION

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]

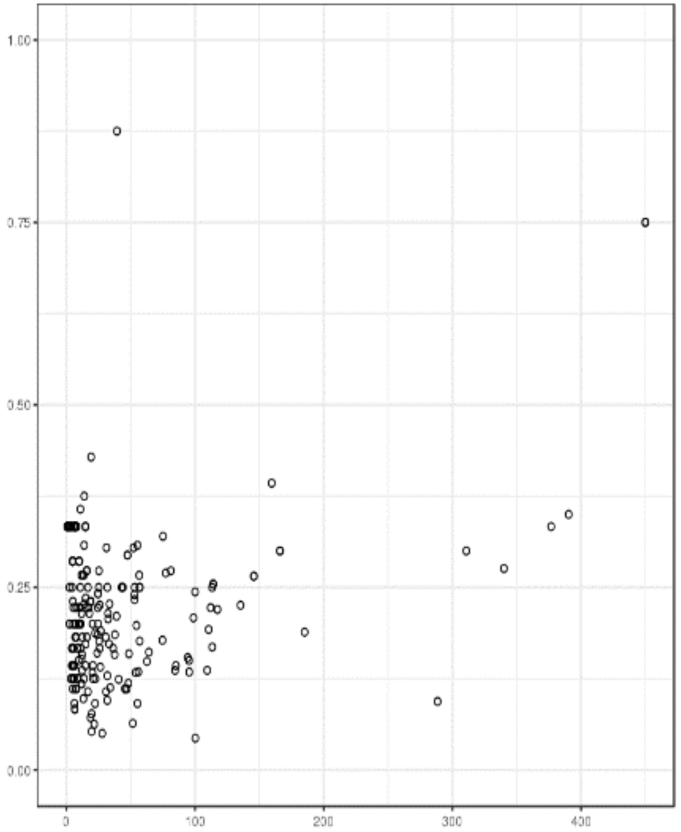


[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

REPRESENTING MULTIVARIATE OBSERVATIONS

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

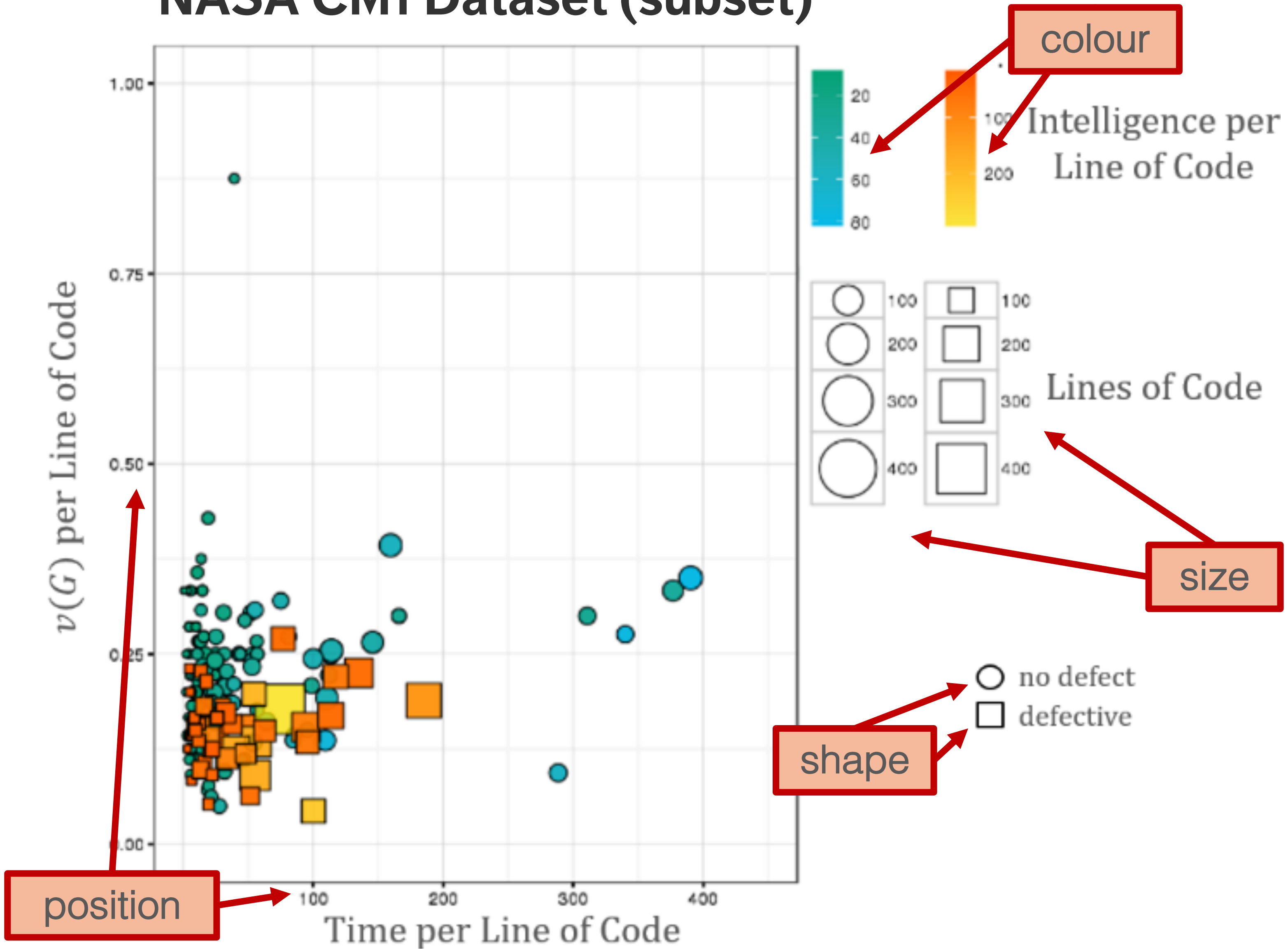
- size
- color
- value
- texture
- line orientation
- shape
- (motion?)



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)

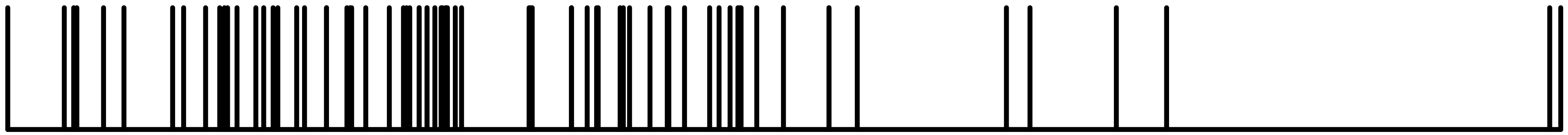
Line Graph

Bar Chart

Scatterplot

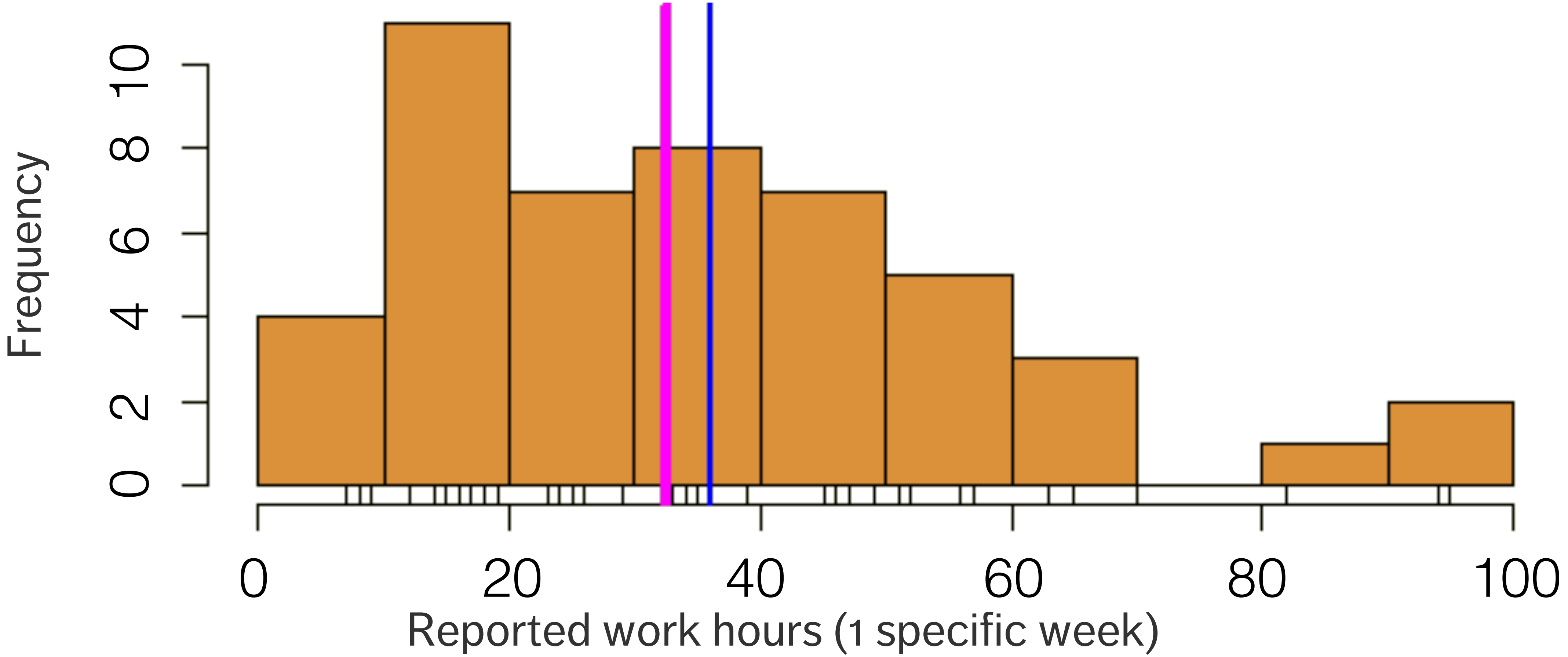
```
linestack(pageview$pages)
```

LINE CHART/RUG CHART

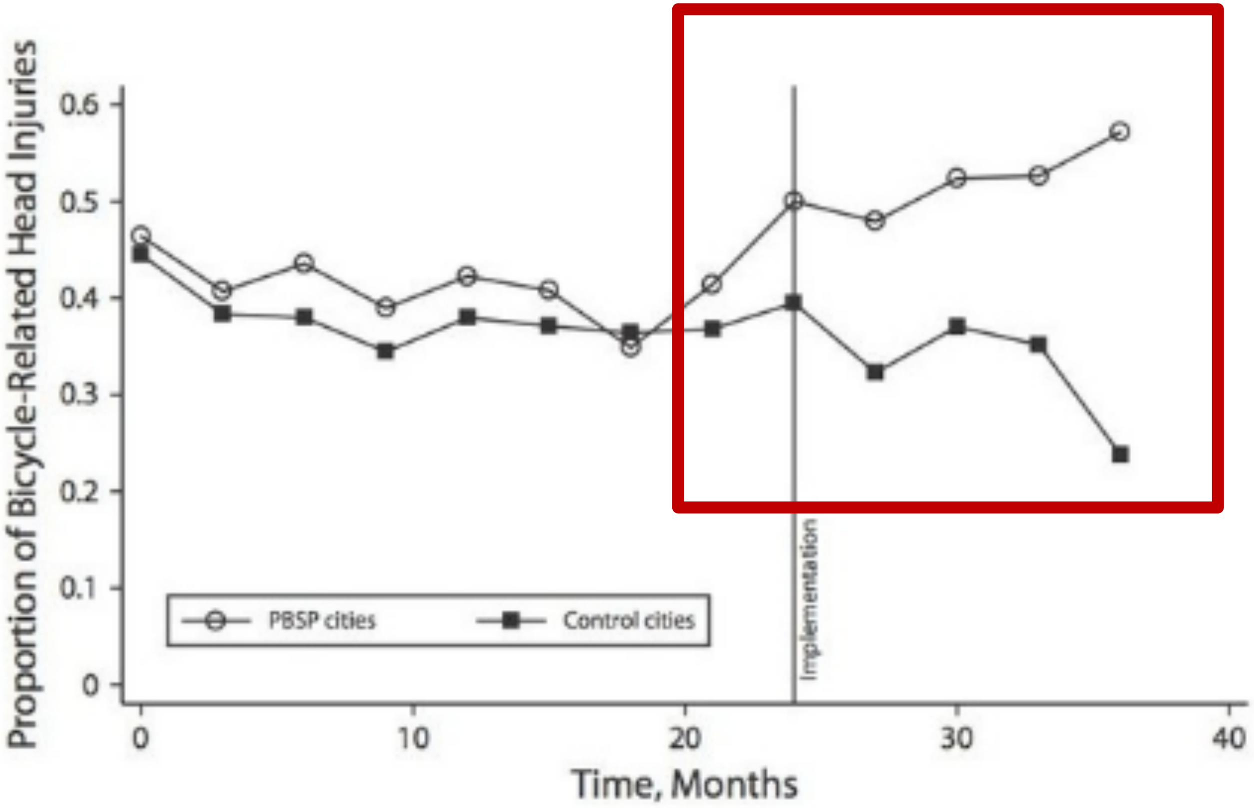


- Gaps in the number line indicate an absence of those numeric values in the dataset
- Remember: this is (possibly) different from the order that values appear in the dataset – since it is a number line, it shows where the values fall numerically
- If values are exactly the same, they will be on top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

HISTOGRAMS


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

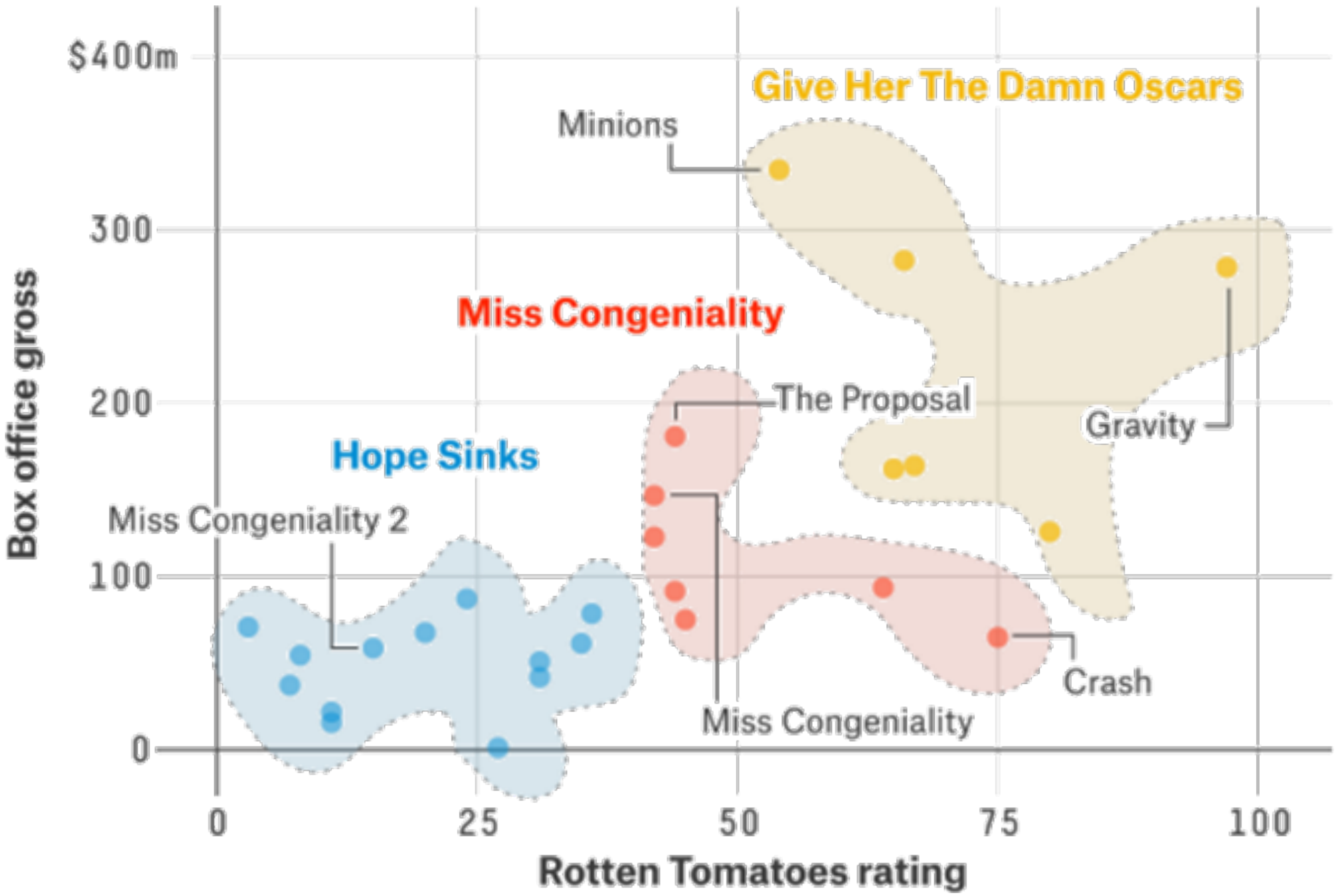
LINE GRAPHS

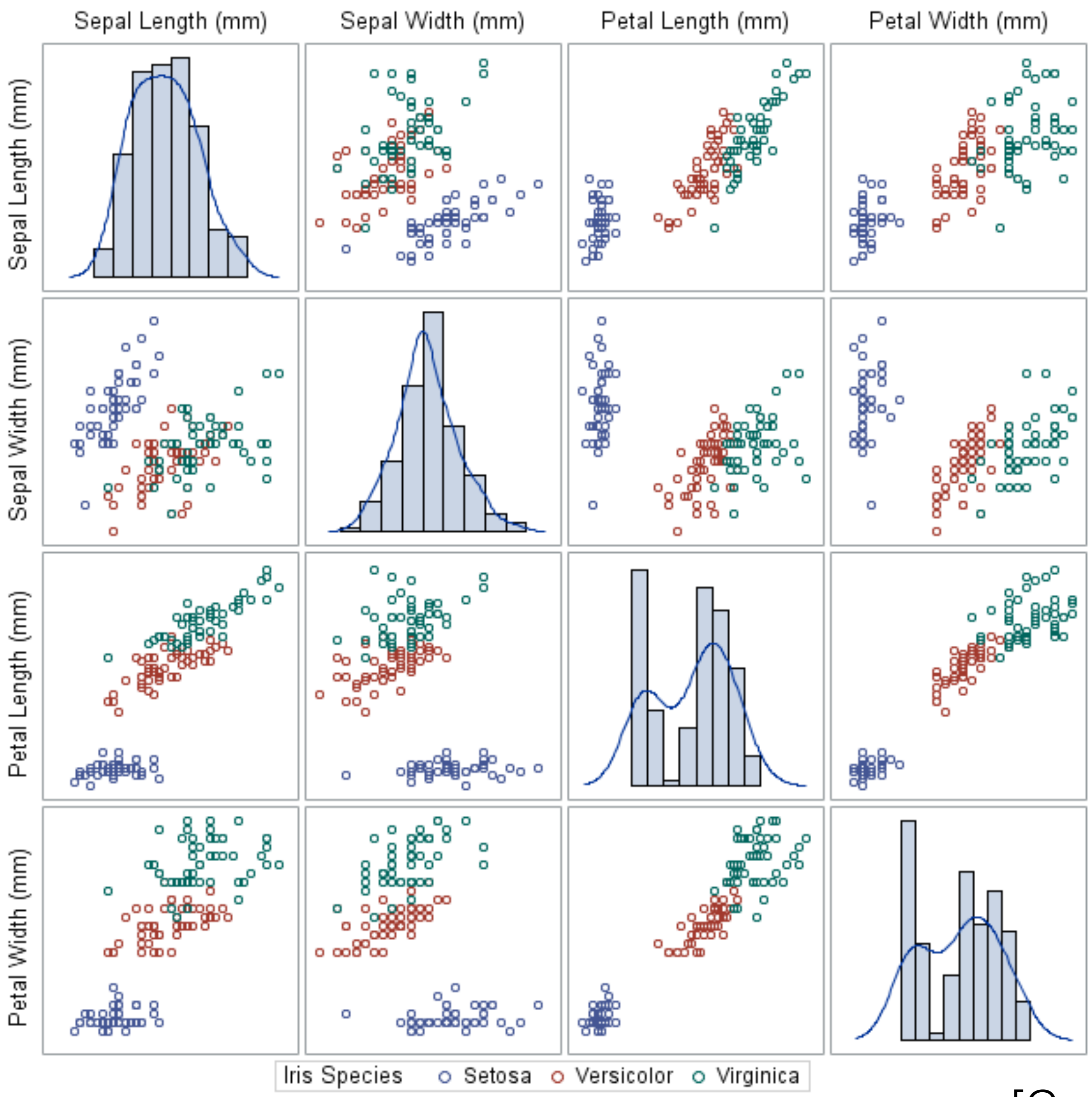


Proportion of all bicycle-related injuries that were classified as head injuries among cities with public bike share programs and control cities, centered on intervention date (vertical line); North America.

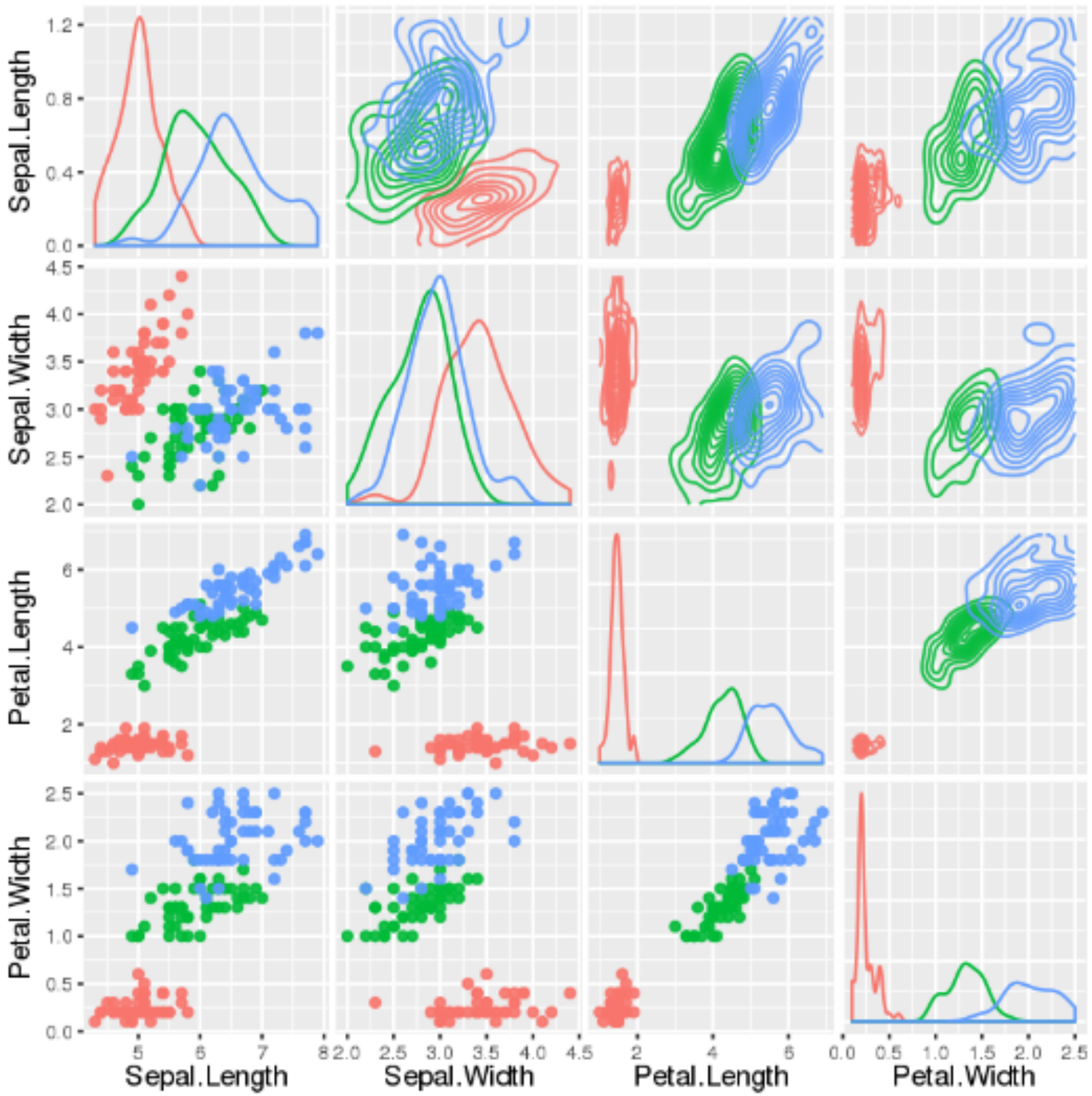
[Graves et al., *Am.J.Phys.Health*, 2014]

SCATTERPLOT

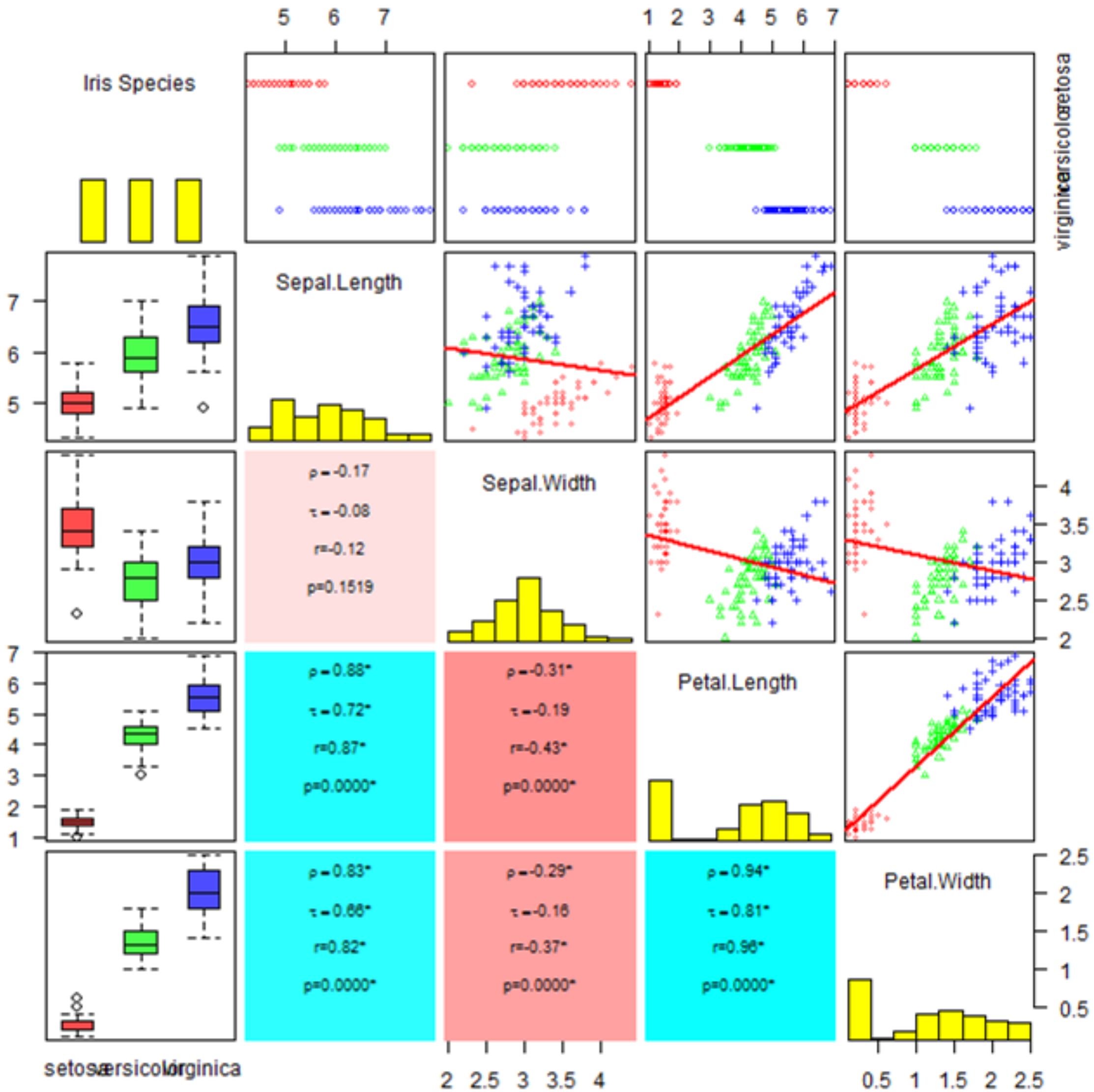




[Created using SAS proc sgscatter]



[Created using R command ggpairs]



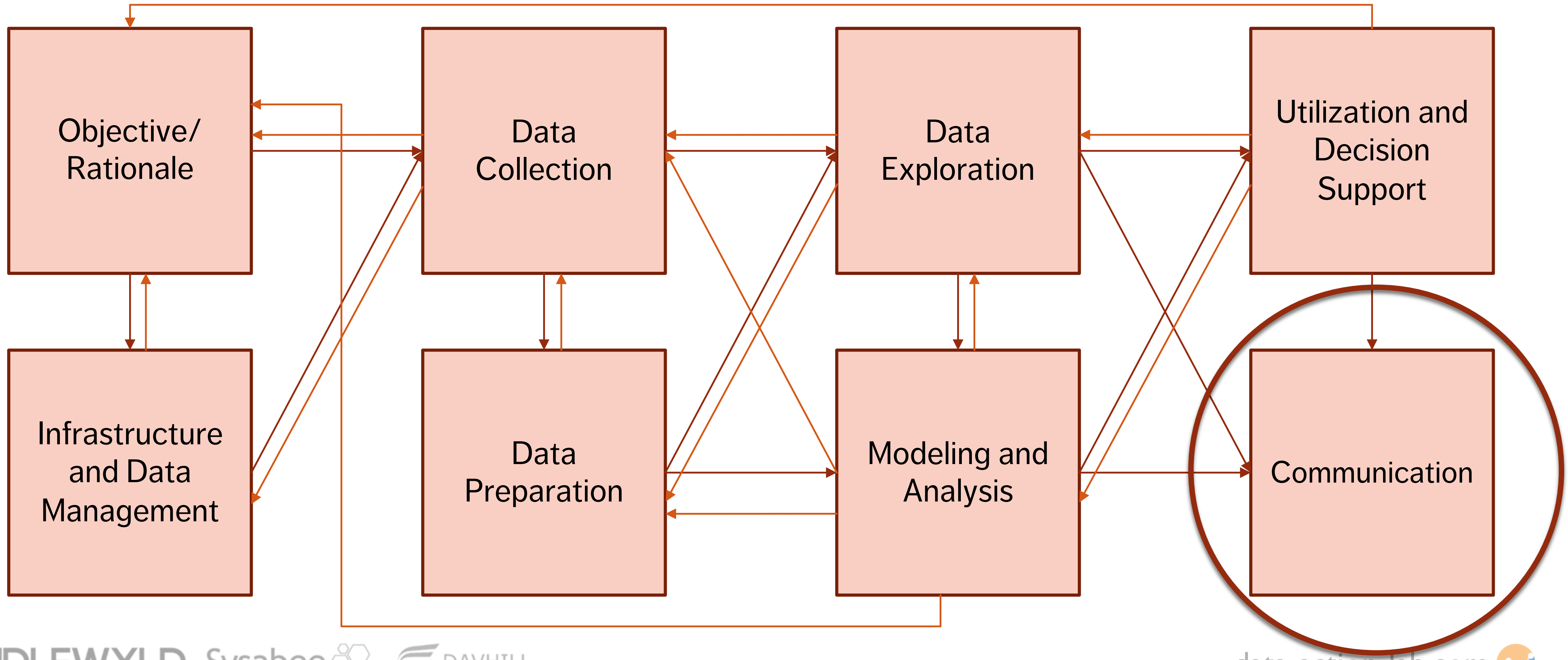
Is this starting to get too cluttered?

[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.

[http://dataphys.org/list/]

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

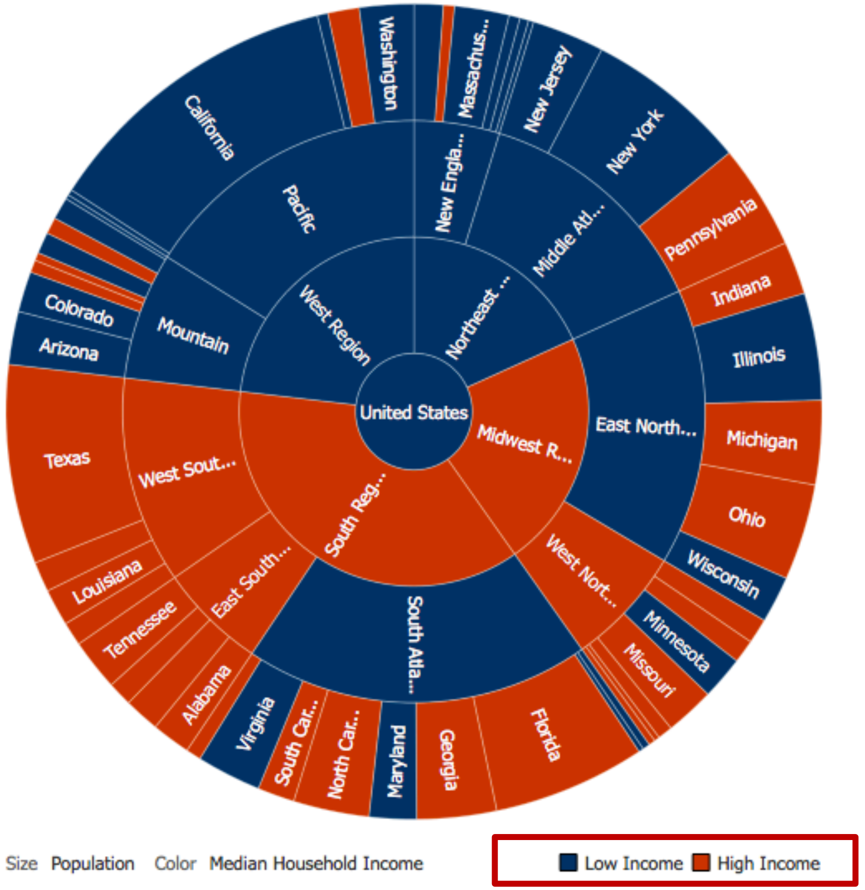
A **method**, as well as an item (**objective**)

Typically focuses on the **quantifiable**

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

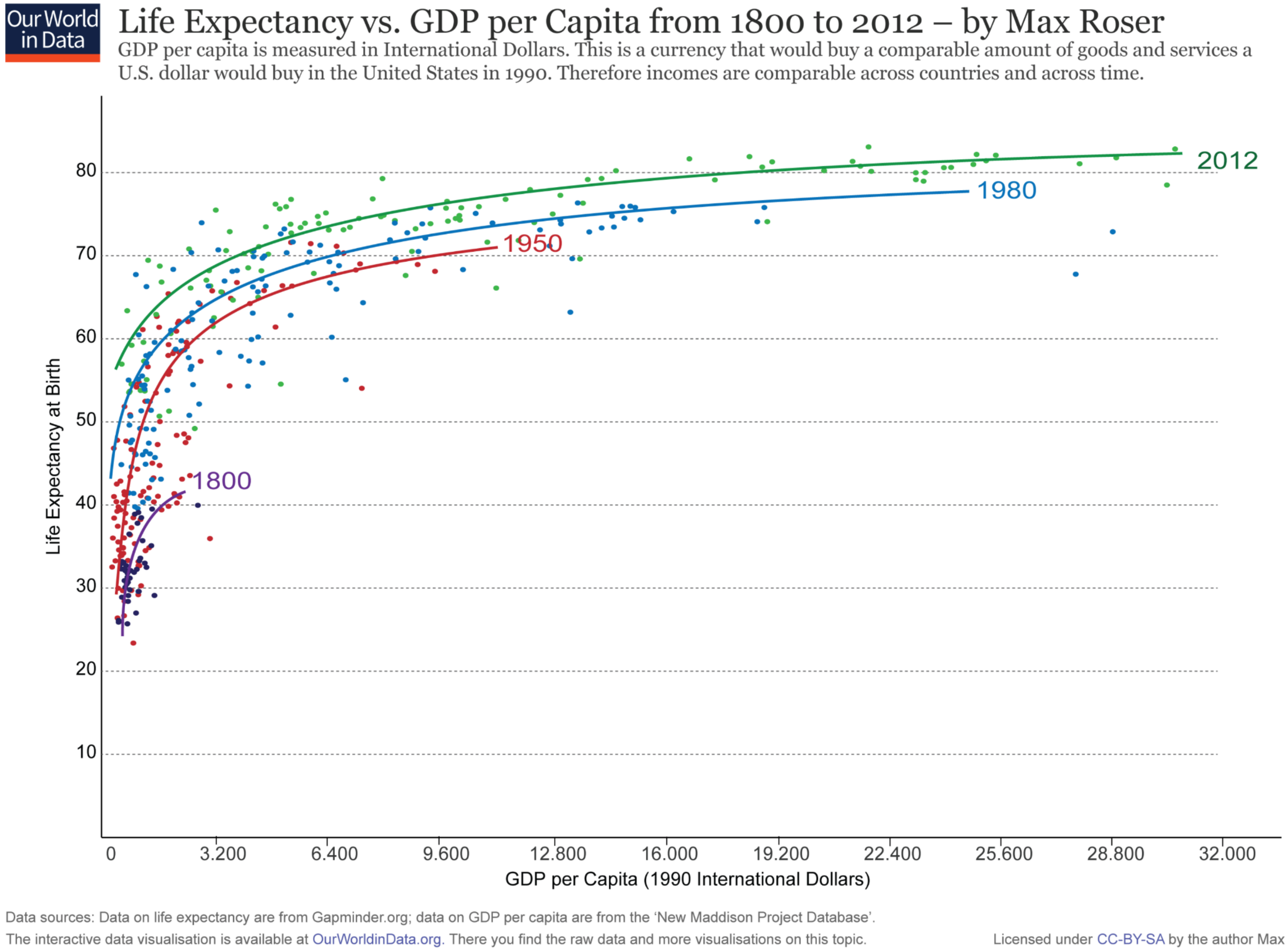
May be generated automatically

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



?

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

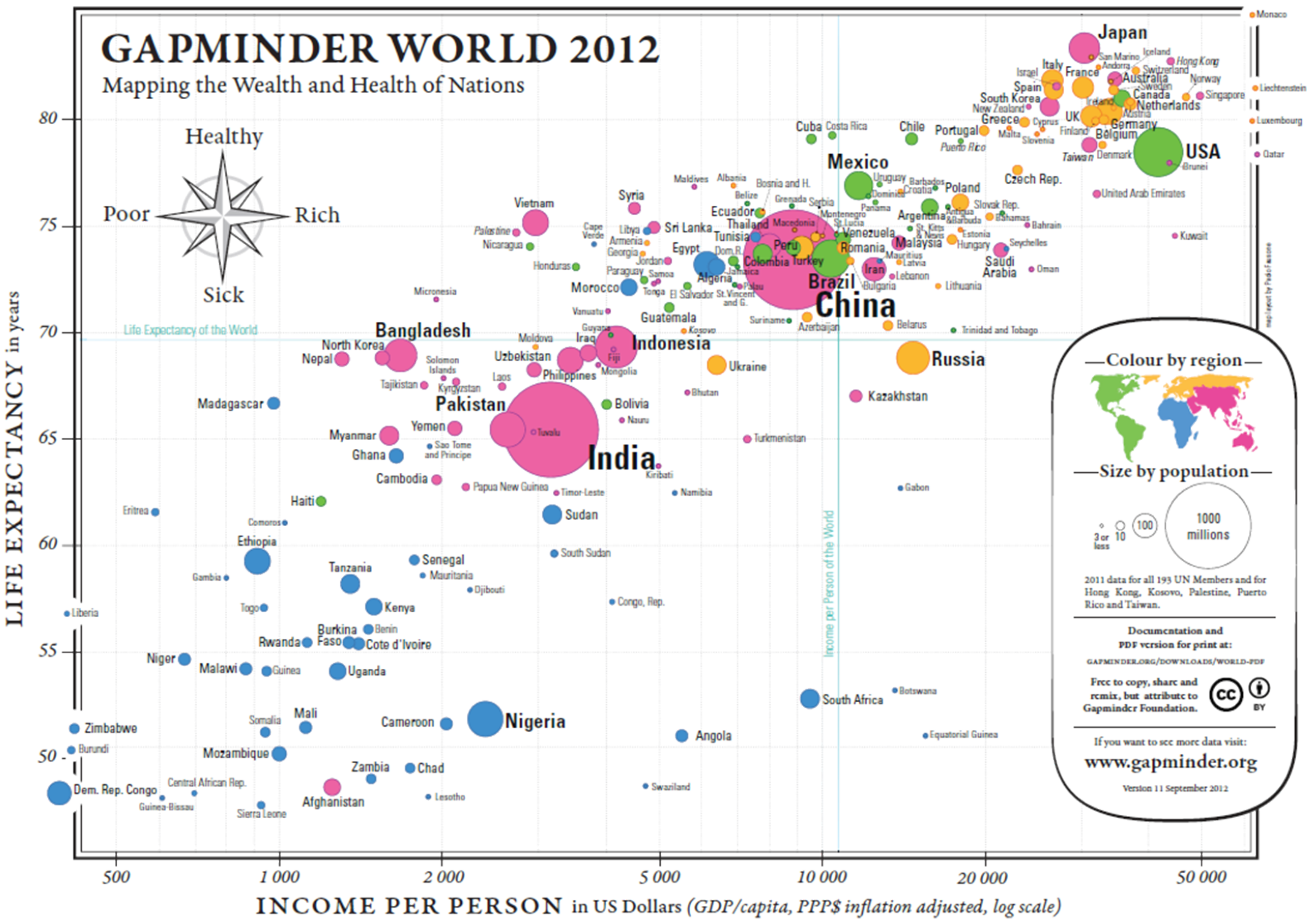


This graph displays the correlation between life expectancy and GDP per capita.

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 capita increases.

[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

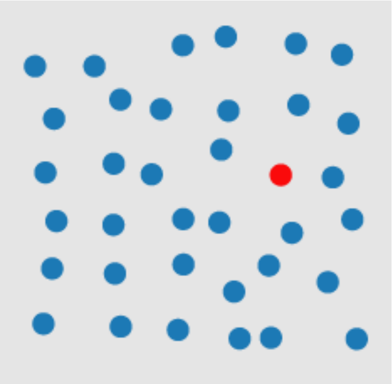
- **Pre-attentive processes:** fast, instinctive, efficient, multitasking
  gather information and build patterns:

  features → patterns → objects

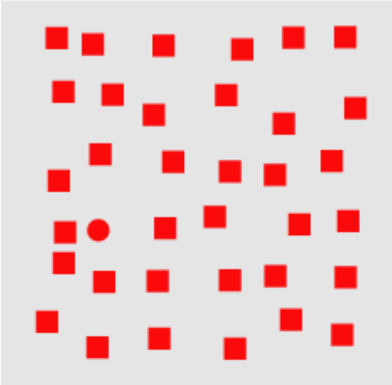
- **Attentive process:** slow, deliberate, focused
  discover features in the patterns:

  objects → patterns → features

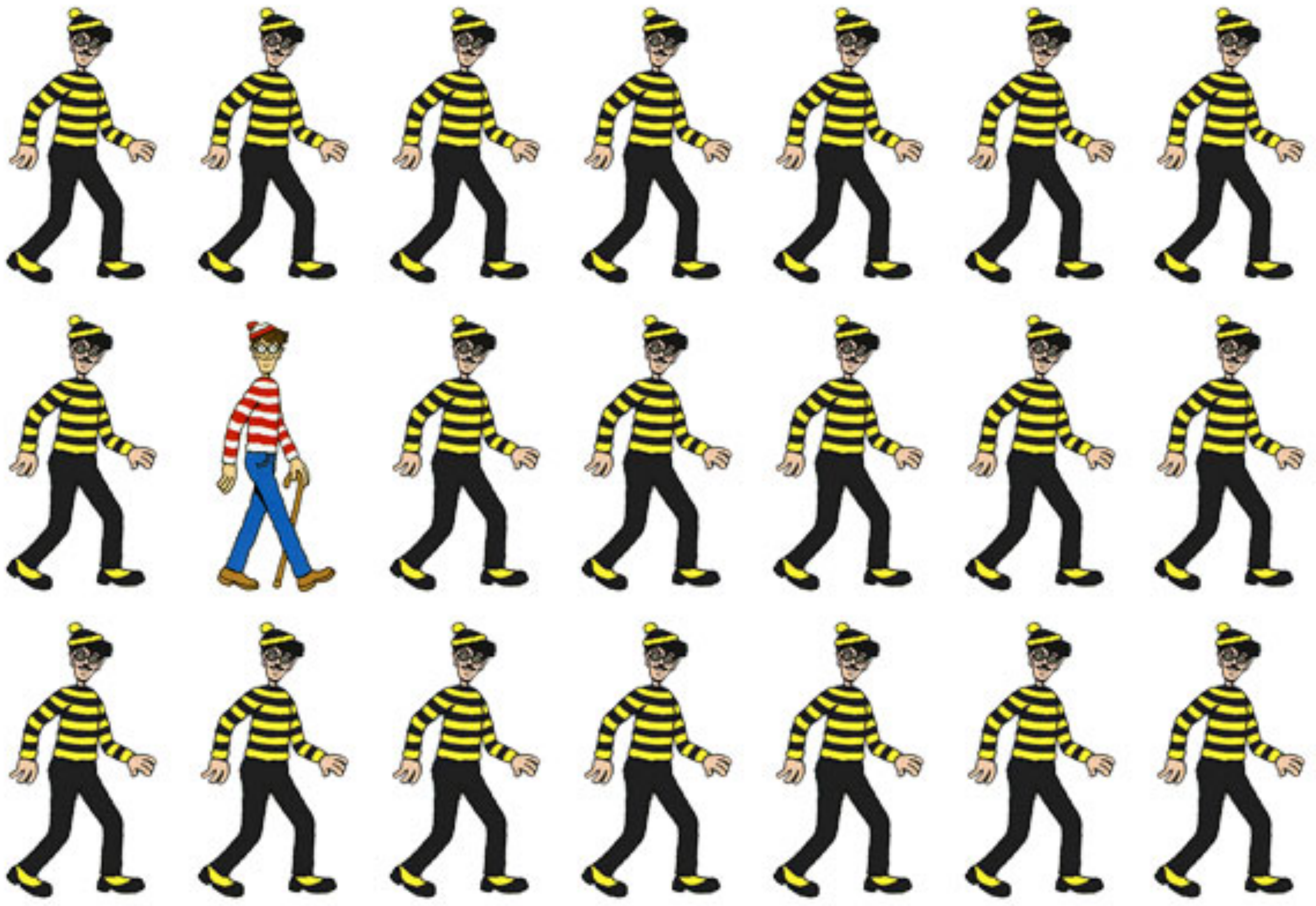
pre-attentive



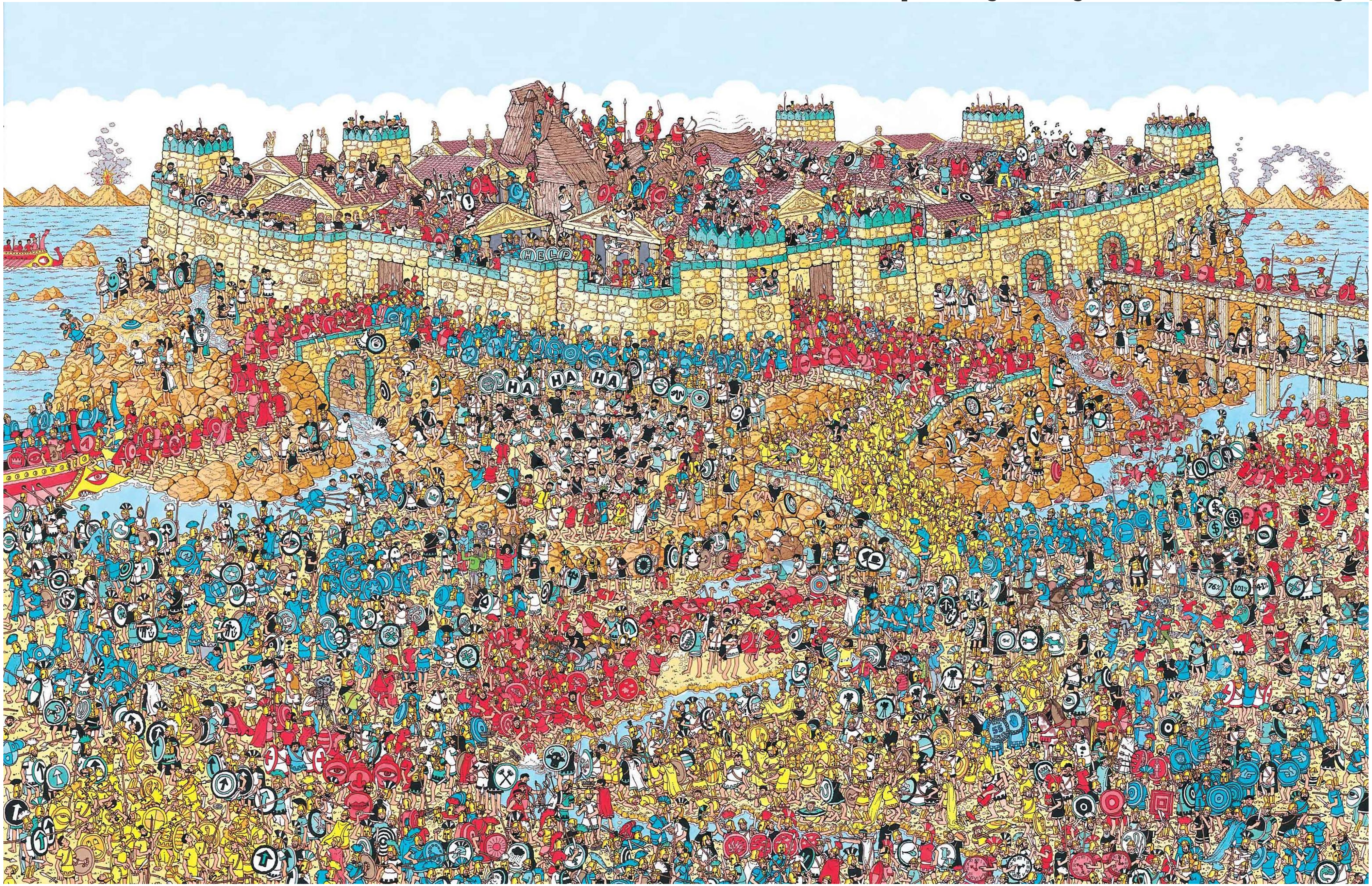
attentive



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



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



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

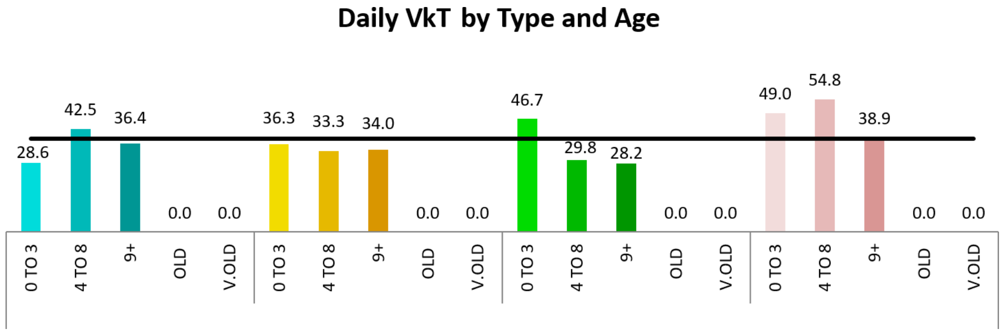
BASIC RULES

1. Check the data

outliers, spikes, anomalies

2. Explain encoding

don't assume the reader knows what everything means



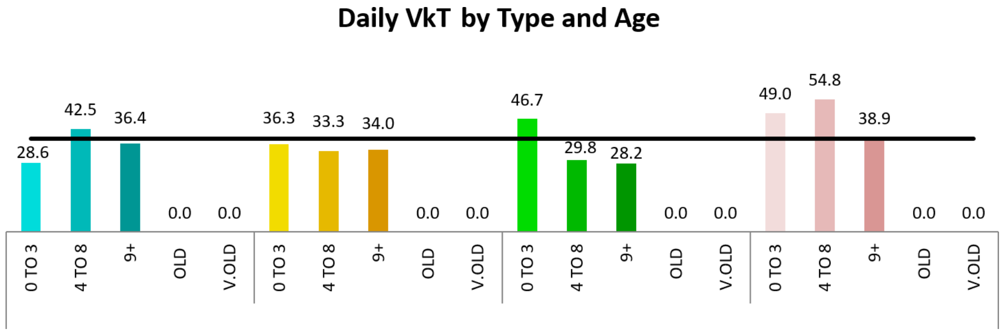
3. Label axes

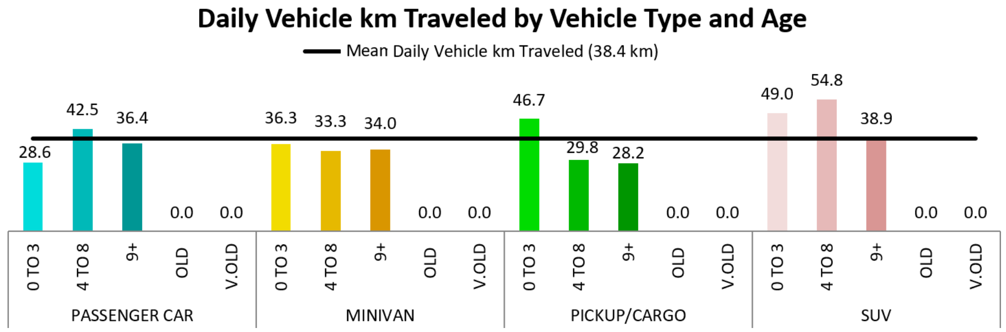
knowing the scale is important

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

BASIC RULES

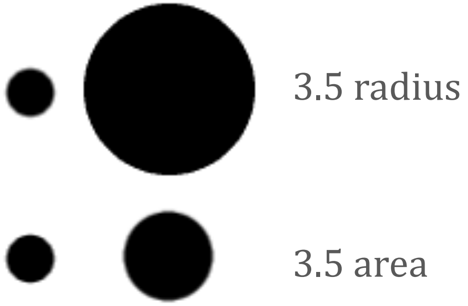
4. Include units
eliminate the need for guesswork

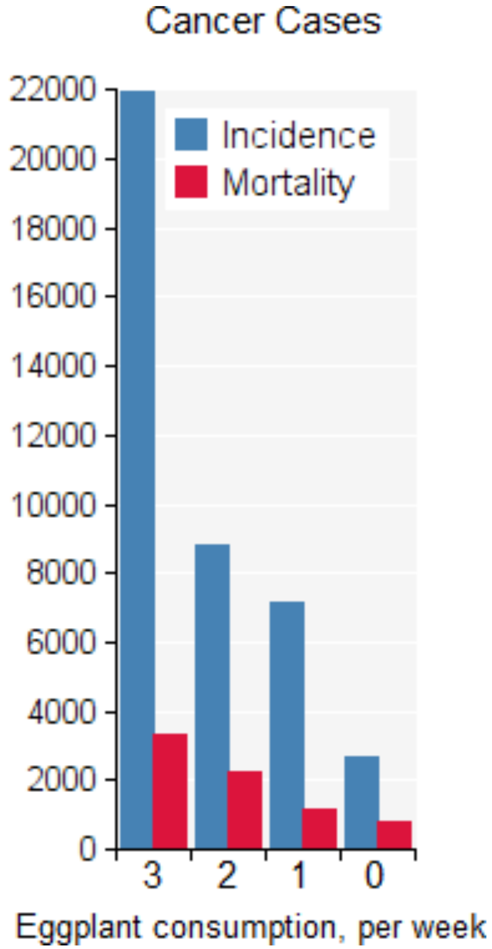
5. Keep your geometry in check
circles and 2D shape are sized by area, bar charts by length

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

BASIC RULES





4. Include units
eliminate the need for guesswork

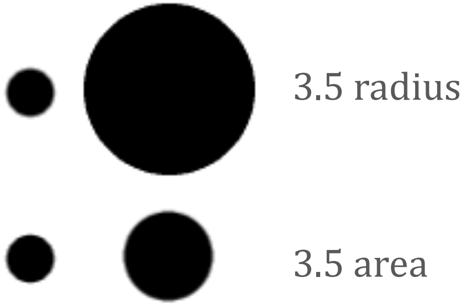
5. Keep your geometry in check
circles and 2D shape are sized by area, bar charts by length

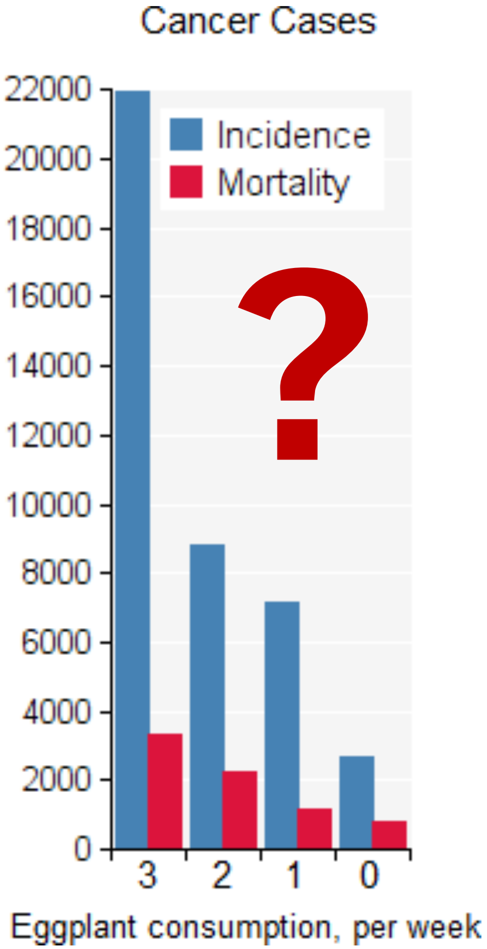
6. Include your sources
protect yourself, and let those who want to dig deeper do so

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

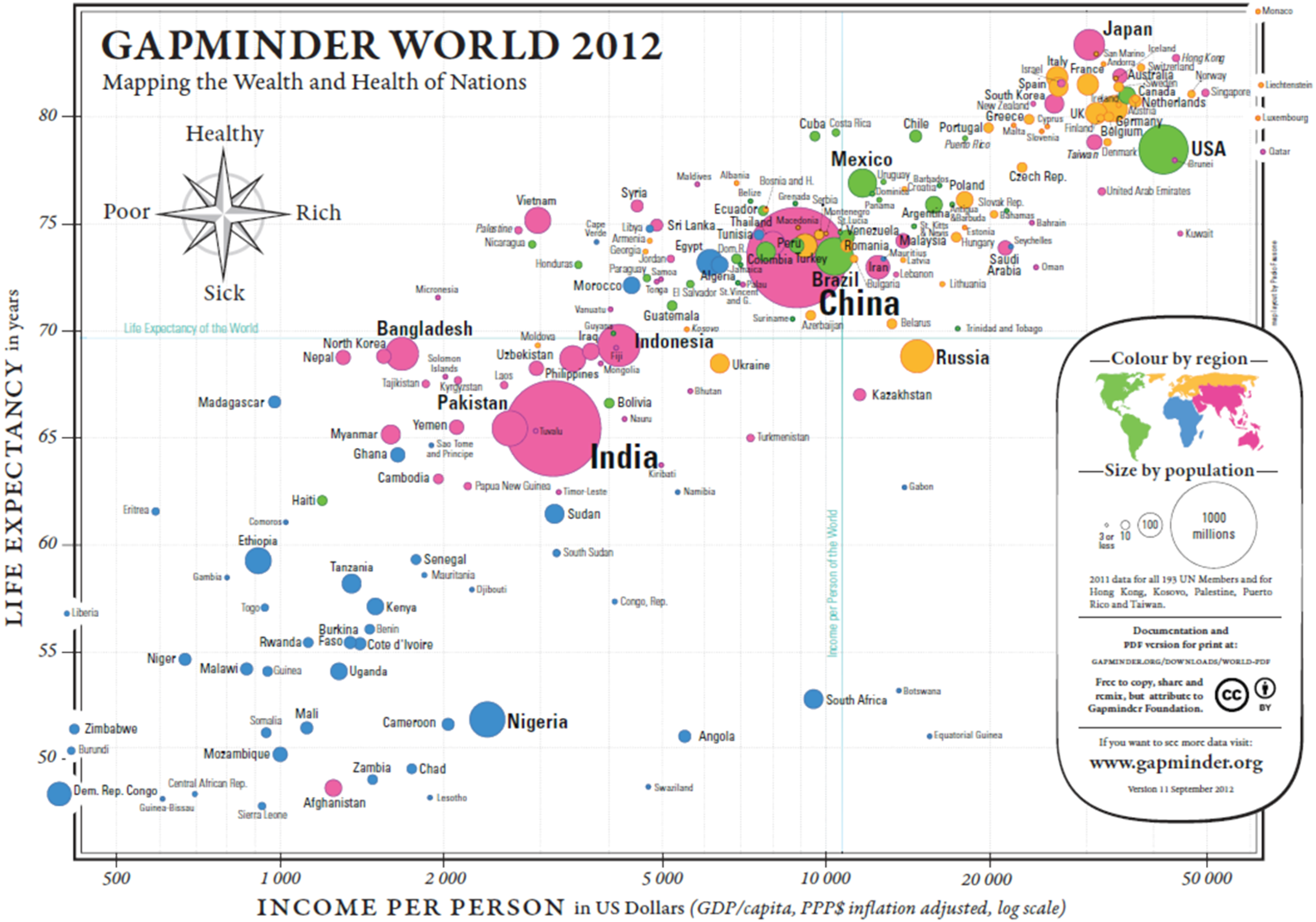
6. Include your sources
protect yourself, and let those who want to dig deeper do so

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]

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

DATA EXPLORATION AND DATA VISUALIZATION

[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.

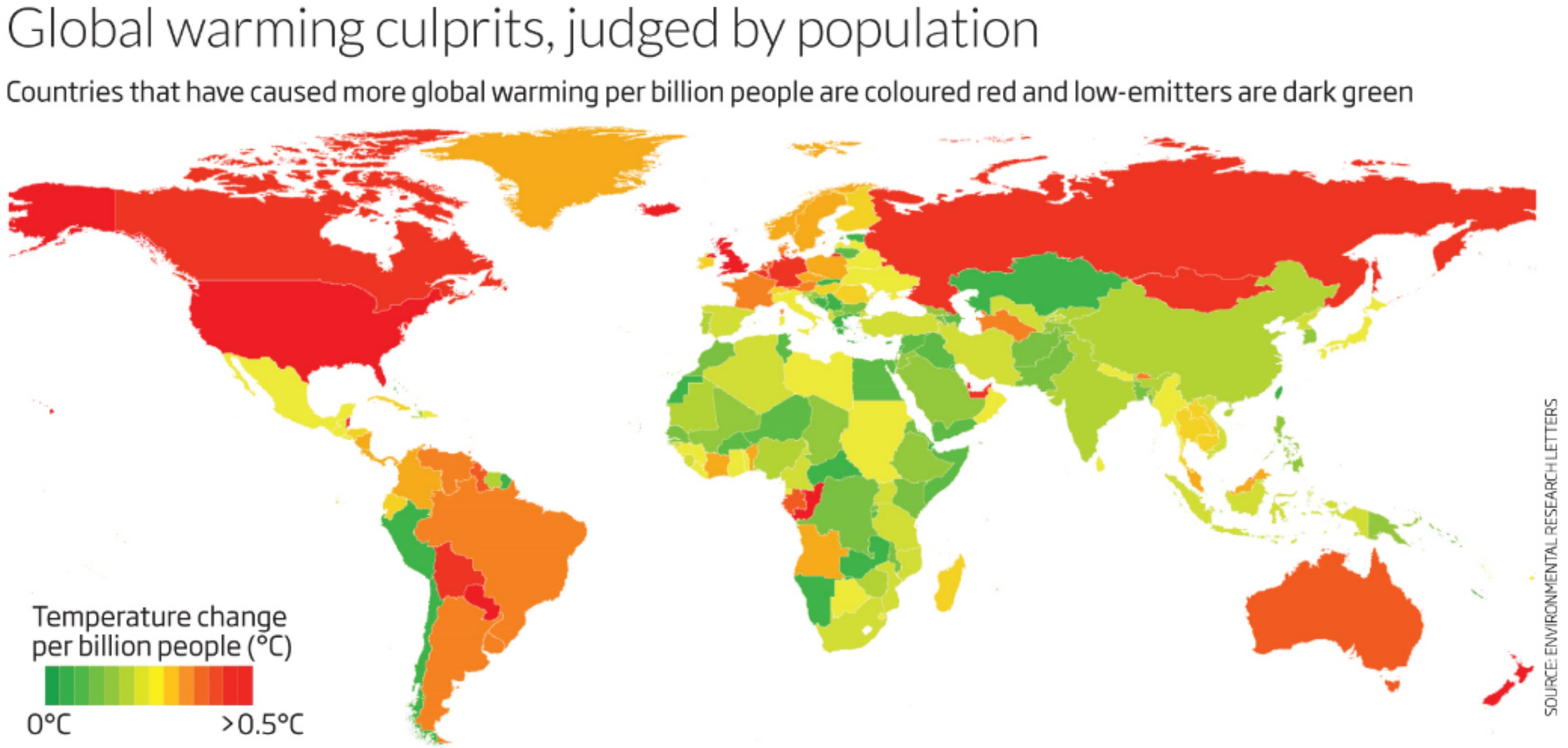
The seasonal shift of our setting sun traces an ark of elevated risk – an echo of the curve of the Earth, itself (**Note:** ???).

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

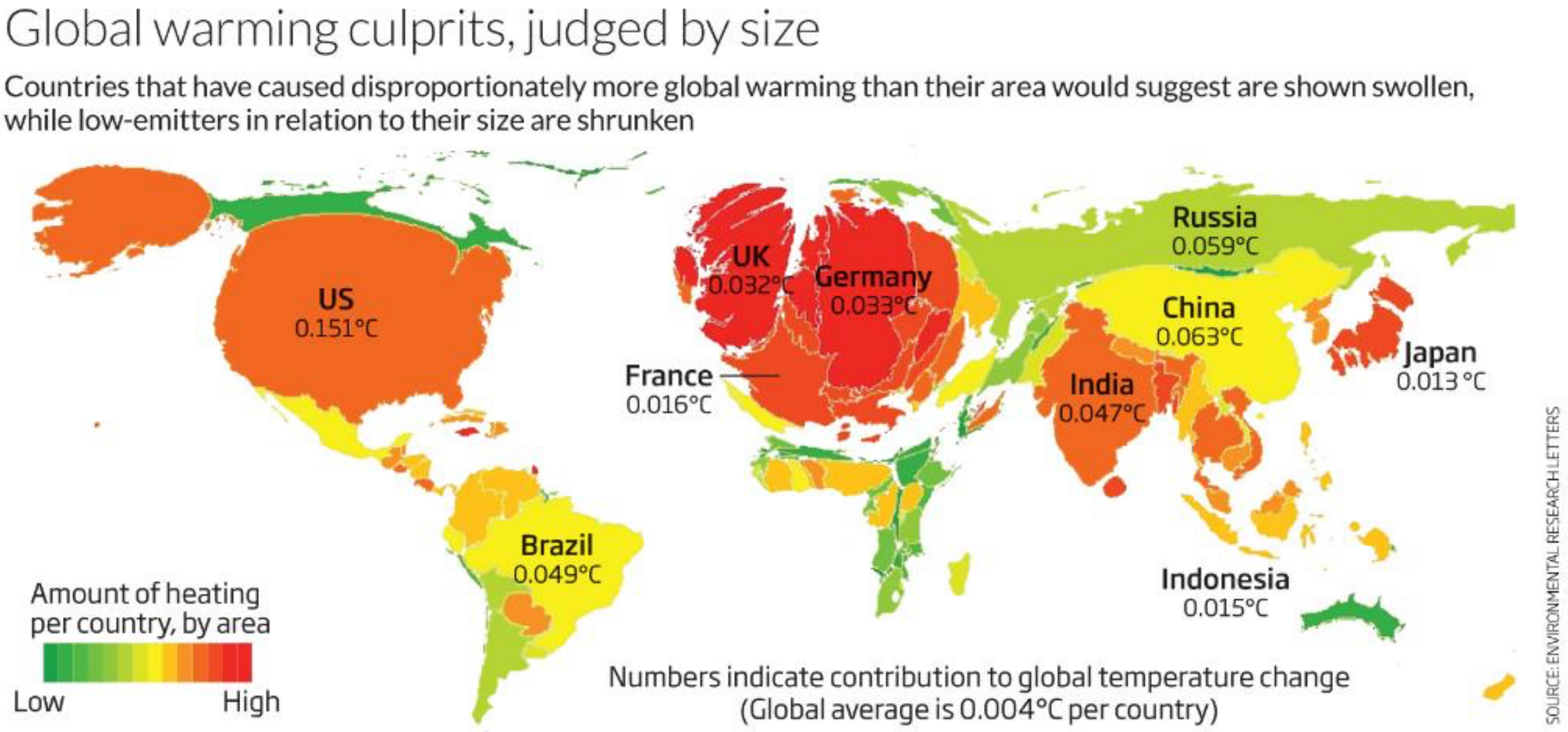
		MIDNIGHT	1 AM	2	3	4	5	6	7	8	9	10	11	NOON	1 PM	2	3	4	5	6	7	8	9	10	11
WINTER	J	12%	14%	11%	16%	12%	20%	24%	16%	12%	13%	10%	11%	11%	10%	8%	11%	11%	23%	38%	32%	28%	22%	17%	16%
	F	13%	13%	13%	15%	14%	20%	22%	11%	10%	8%	9%	10%	8%	9%	8%	8%	9%	16%	32%	35%	28%	24%	21%	15%
SPRING	M	14%	15%	12%	10%	11%	22%	23%	13%	12%	12%	11%	11%	10%	8%	9%	9%	9%	8%	17%	30%	25%	26%	20%	18%
	A	15%	14%	12%	13%	12%	19%	16%	8%	9%	11%	10%	11%	8%	7%	6%	9%	8%	9%	9%	13%	30%	29%	21%	16%
	M	12%	14%	12%	13%	16%	15%	9%	10%	9%	12%	9%	10%	11%	7%	9%	9%	7%	9%	10%	9%	23%	26%	21%	20%
SUMMER	J	16%	13%	13%	16%	18%	12%	9%	9%	12%	13%	9%	8%	6%	8%	9%	8%	7%	8%	10%	9%	14%	27%	25%	19%
	J	17%	16%	14%	15%	16%	15%	11%	11%	15%	12%	11%	10%	9%	10%	9%	7%	6%	9%	9%	9%	16%	27%	25%	18%
	A	15%	14%	15%	14%	16%	23%	15%	13%	12%	10%	11%	9%	7%	8%	9%	8%	8%	8%	11%	11%	27%	27%	25%	19%
FALL	S	13%	16%	13%	16%	15%	24%	24%	19%	17%	13%	10%	10%	11%	8%	9%	10%	8%	10%	14%	22%	33%	26%	21%	19%
	O	14%	14%	14%	16%	16%	22%	27%	18%	14%	13%	13%	10%	9%	11%	7%	8%	11%	11%	23%	35%	29%	25%	23%	15%
	N	15%	15%	14%	14%	16%	21%	18%	14%	13%	8%	8%	9%	12%	8%	9%	12%	14%	31%	36%	33%	23%	23%	22%	16%
WINTER	D	14%	13%	11%	11%	13%	18%	27%	9%	9%	10%	9%	10%	10%	9%	10%	10%	15%	35%	38%	34%	30%	26%	20%	18%

[New Scientist, Jan 2014]

GEOGRAPHICAL MAPS



GEOGRAPHICAL MAPS



[Upside Down Maps]

MAPS

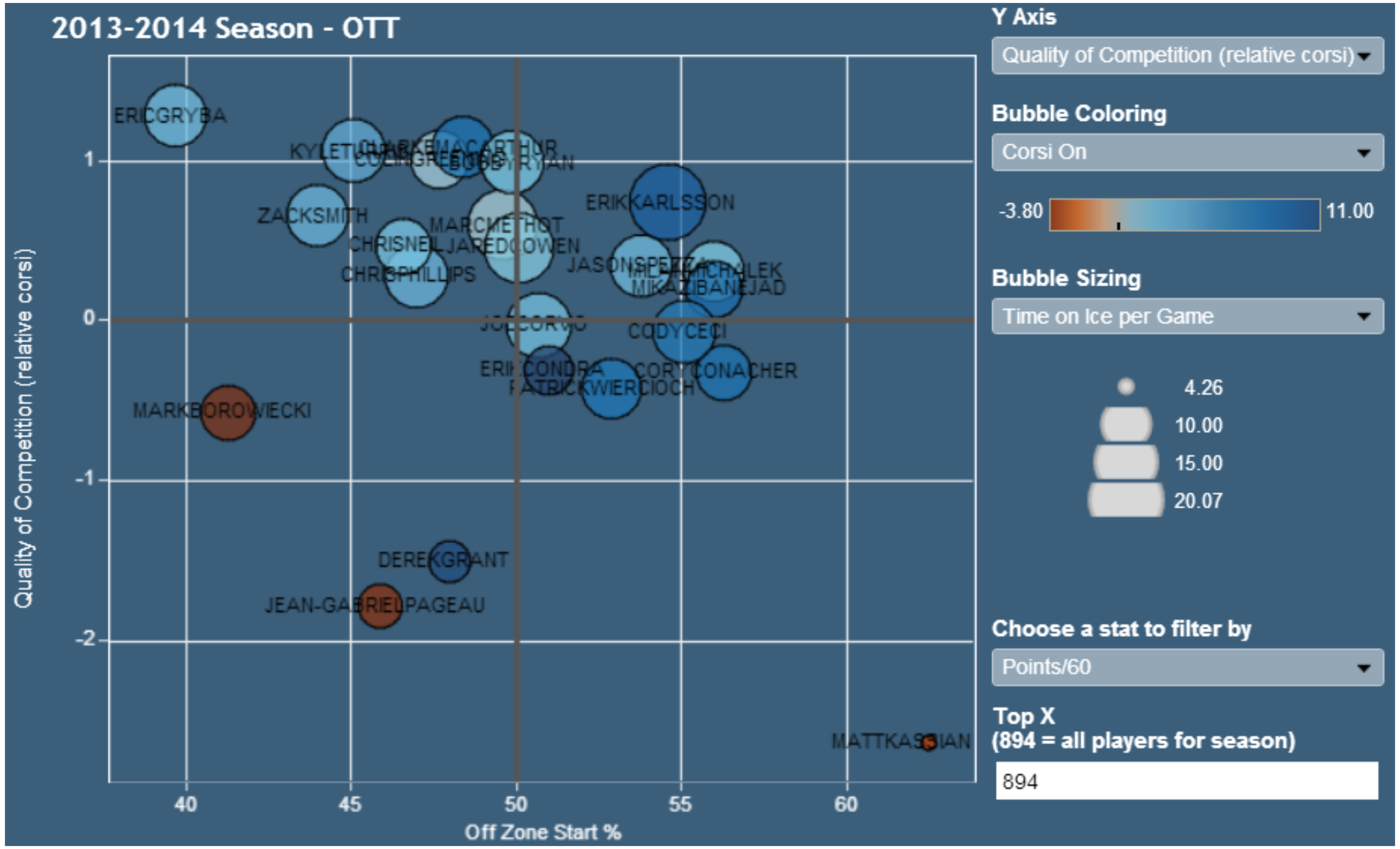
Most of us are quite familiar with geographical maps, so they tend to be easier to interpret.

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

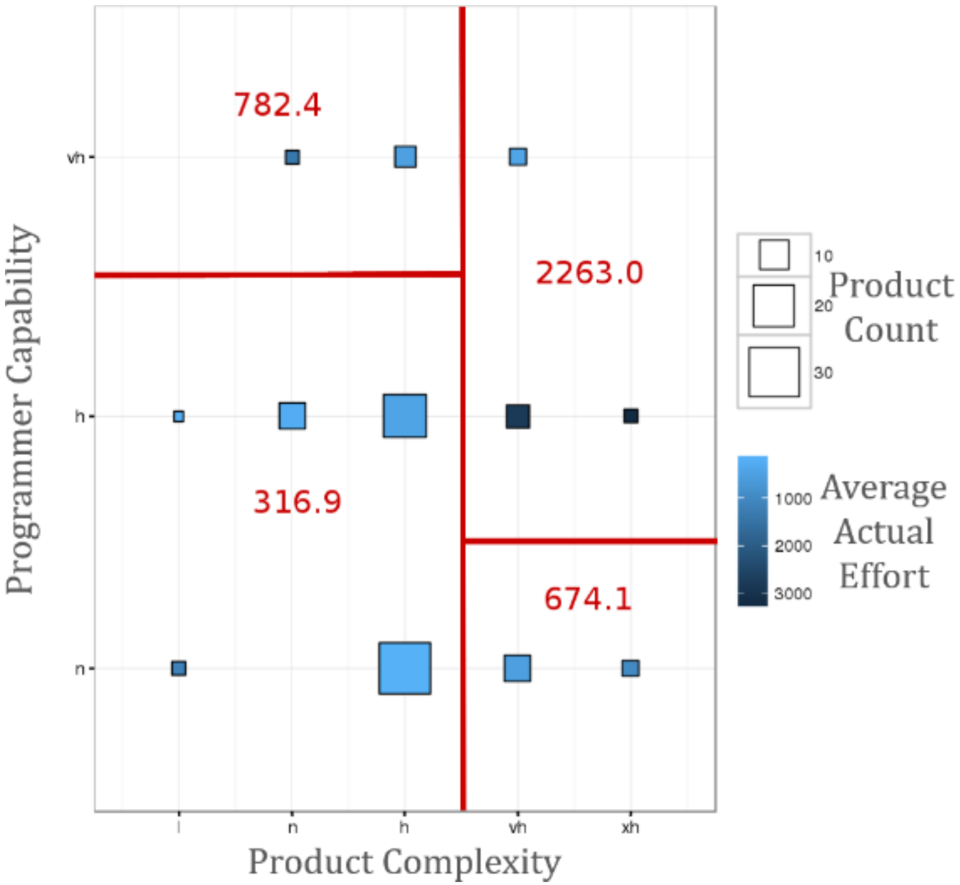
[Hockey Abstract] [Personal file]

BUBBLE CHARTS



NHL Player Usage (Ottawa Senators)

NASA COCOMO Dataset

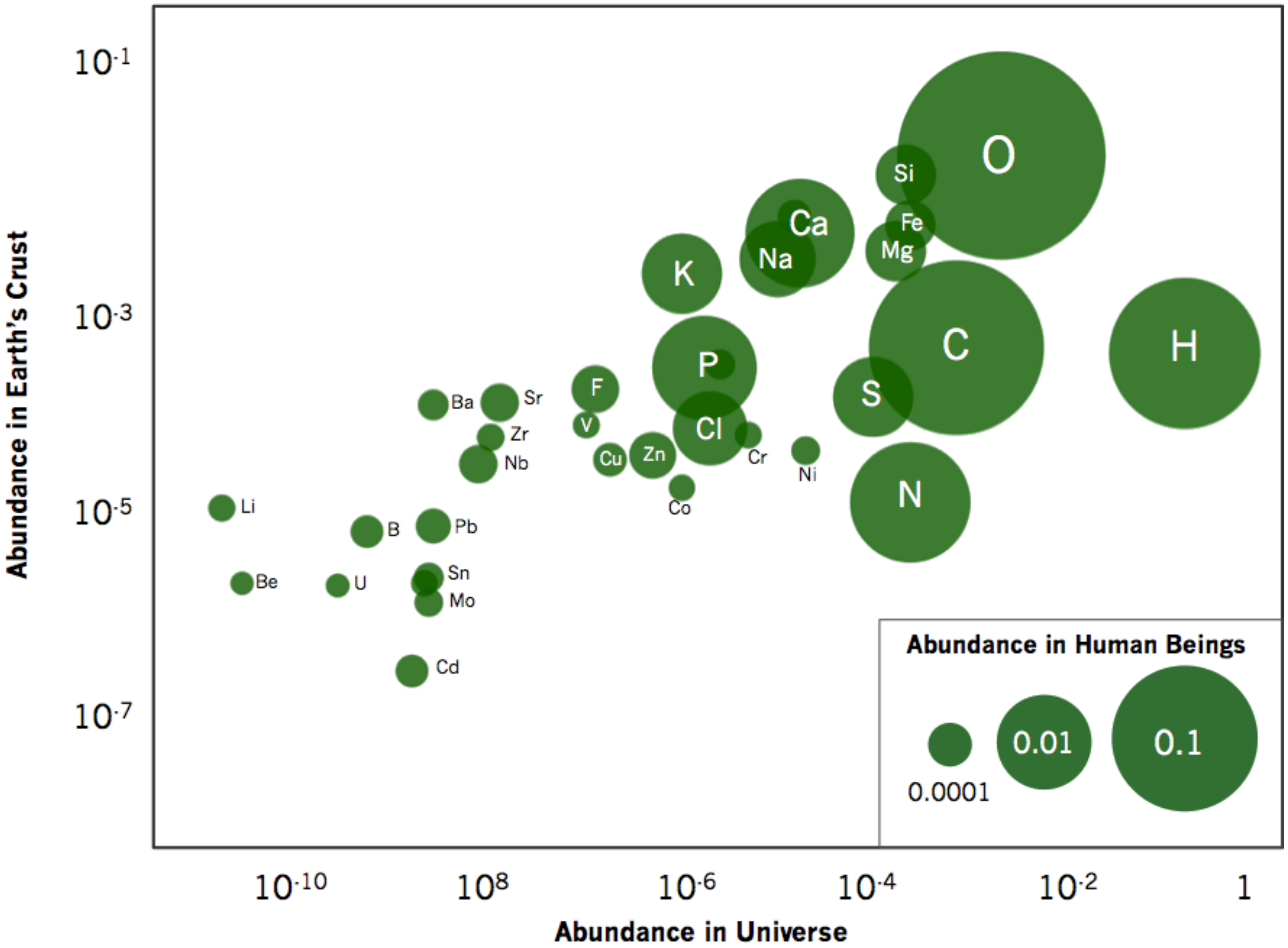


IDLEWYLD Sysabee DAVHILL

data-action-lab.com

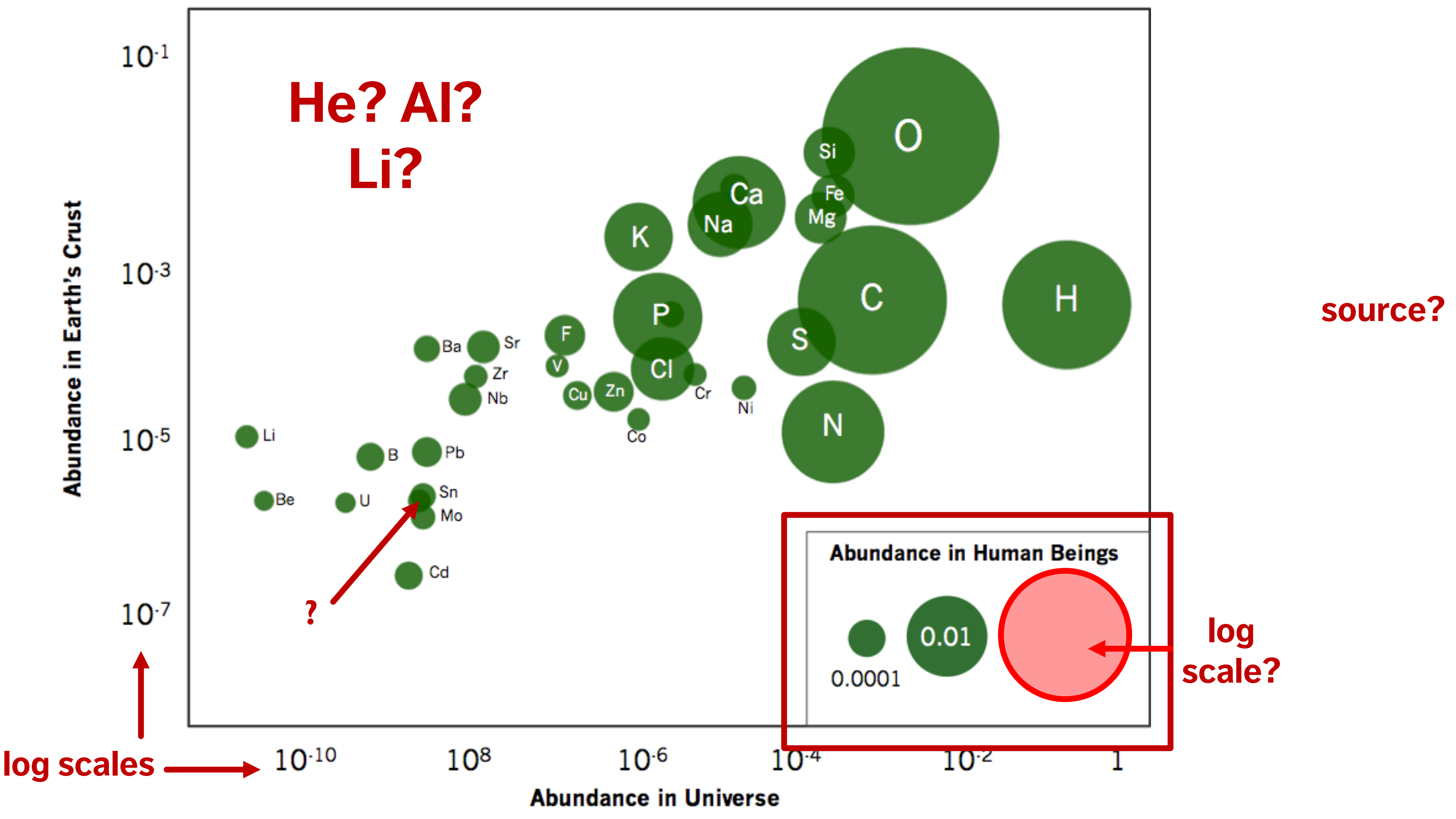
[M. Twardos, The Information Diet]

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

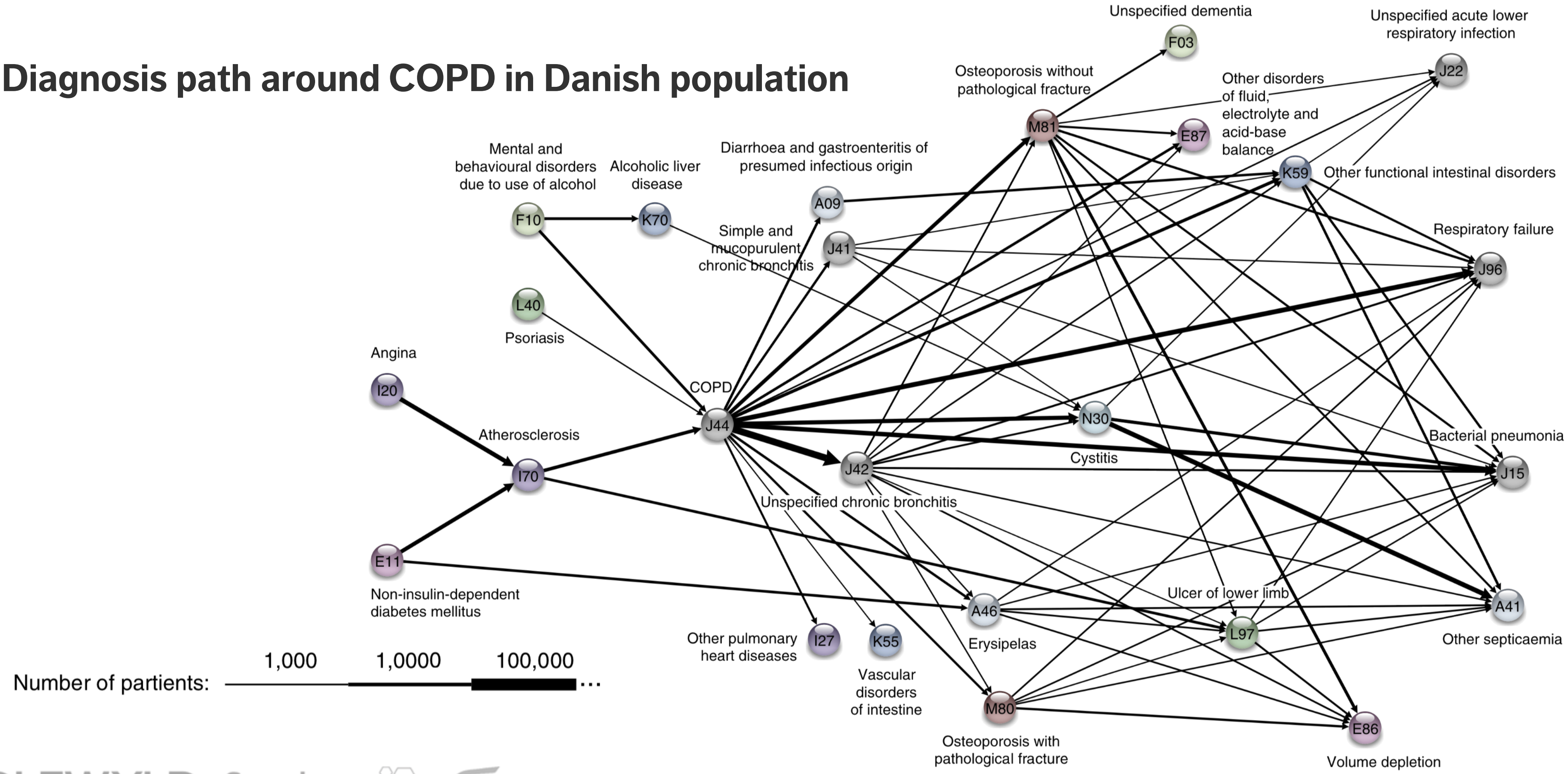
Works best when chart is **not too encumbered**

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

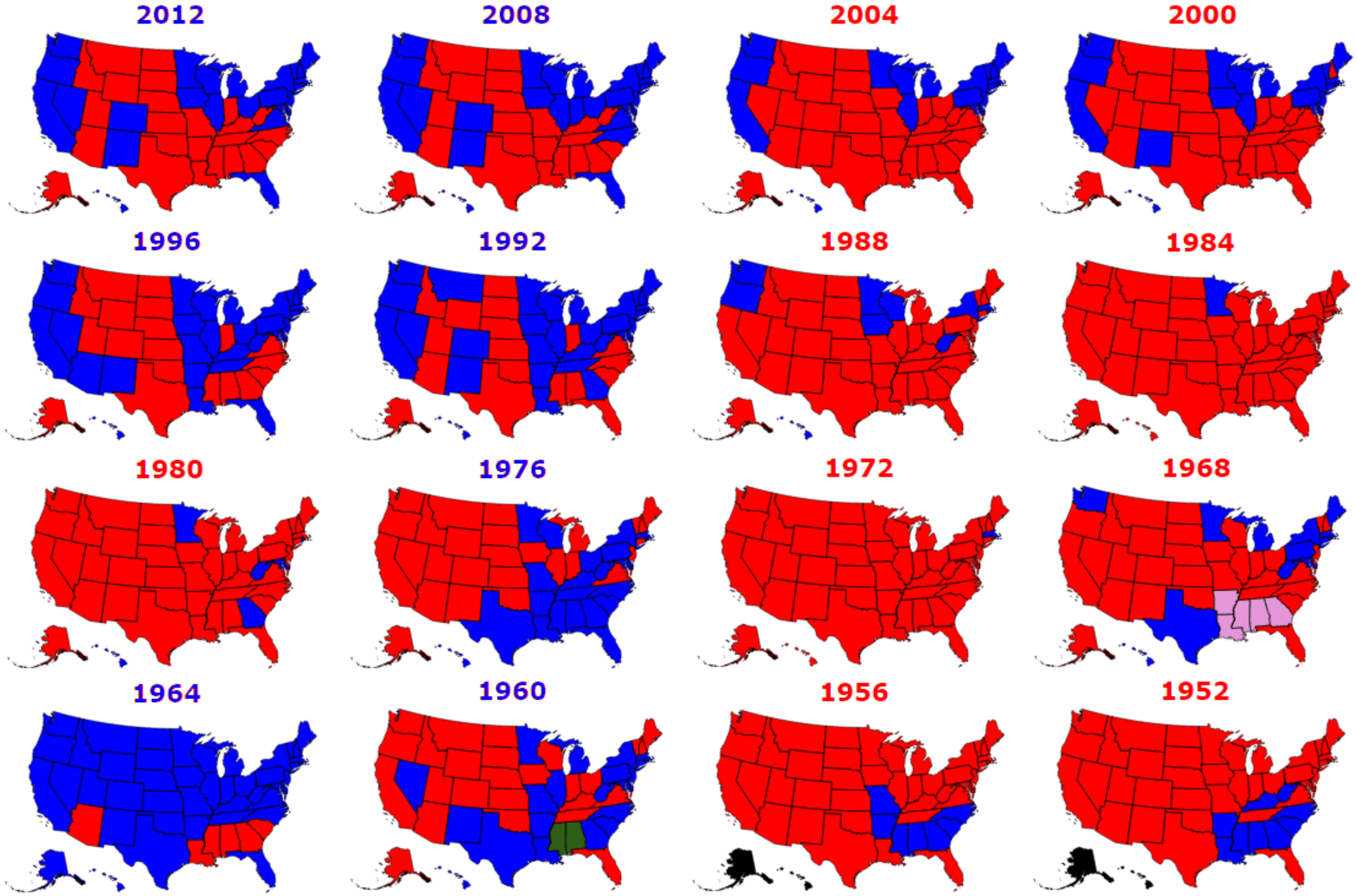
[A.B. Jensen et al., *Nature Communications*]

NETWORK DIAGRAMS

Diagnosis path around COPD in Danish population



SMALL MULTIPLES



U.S. Electoral College Results 1952 – 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)

Animation **does not always** improve a visualization. What insights can interactivity provide? That depends on the data & on visualization.

Take-Aways: explore the data and try different methods

HALL-OF-FAME / HALL-OF-SHAME

DATA EXPLORATION AND DATA VISUALIZATION

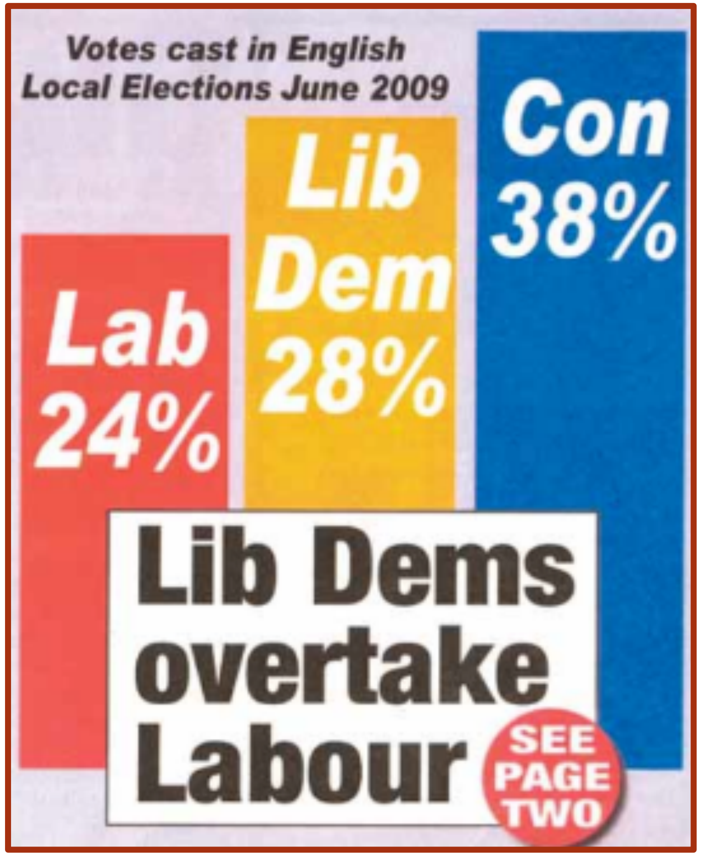
[Fox News, February 20, 2012, mediamatters.org, simplystatistics.org, badgraphs.tumblr.com]

MISLEADING CHARTS

[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.

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

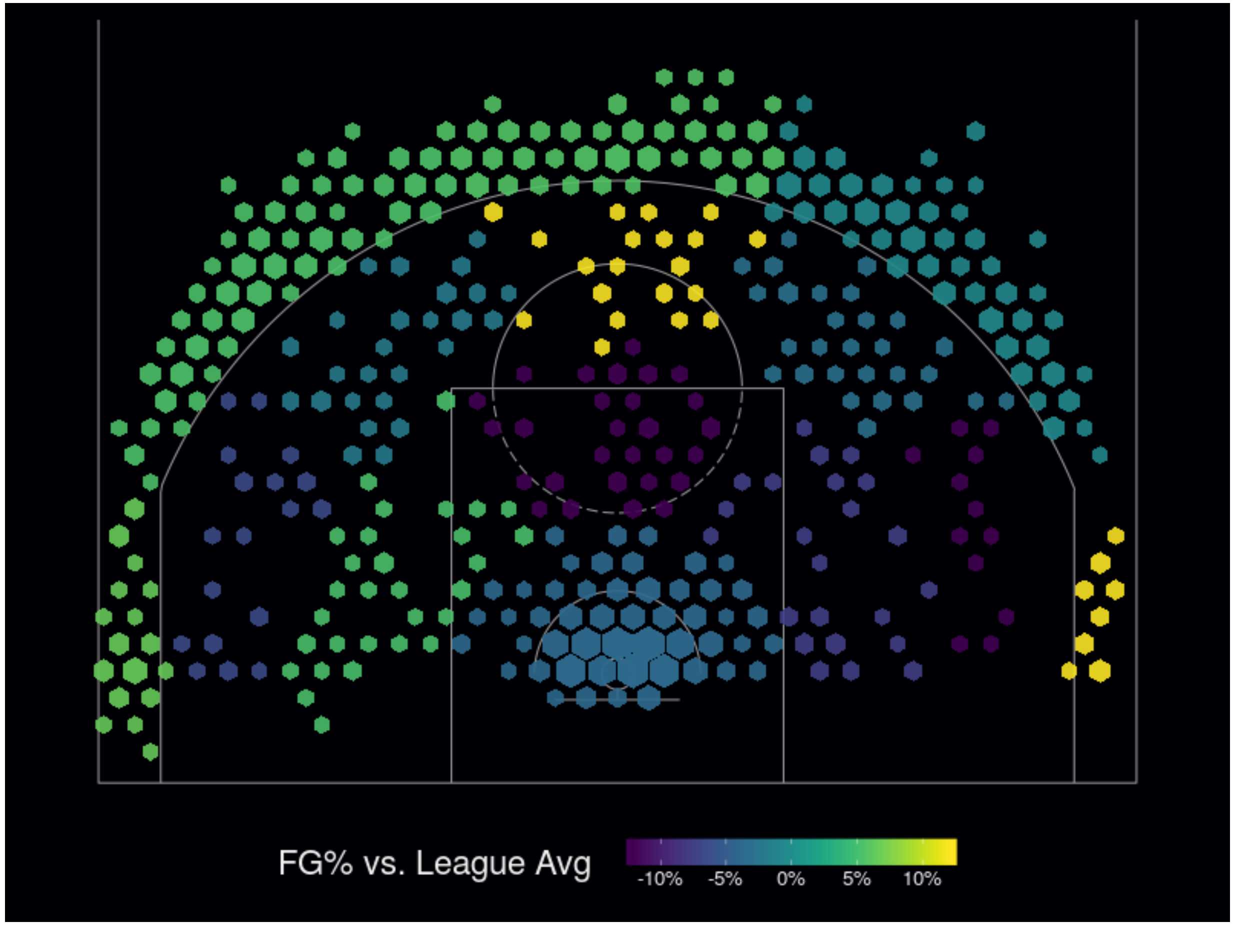
Some of the following are (arguably) good visualizations. But some are not!

Which are which? You be the judge…

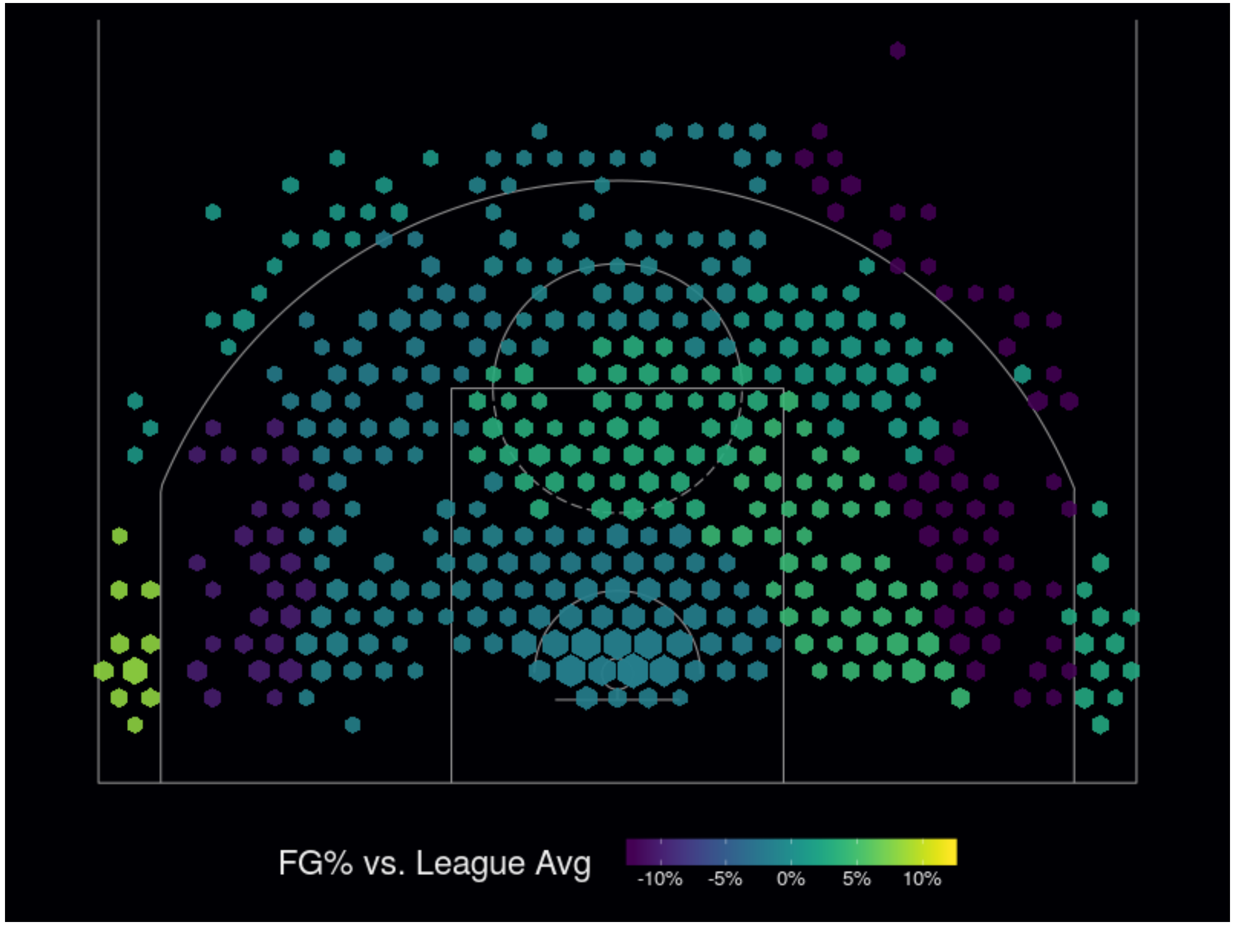
NBA FG% Against League Average ('15–'16)

[T.W. Schneider]

Kyle Lowry



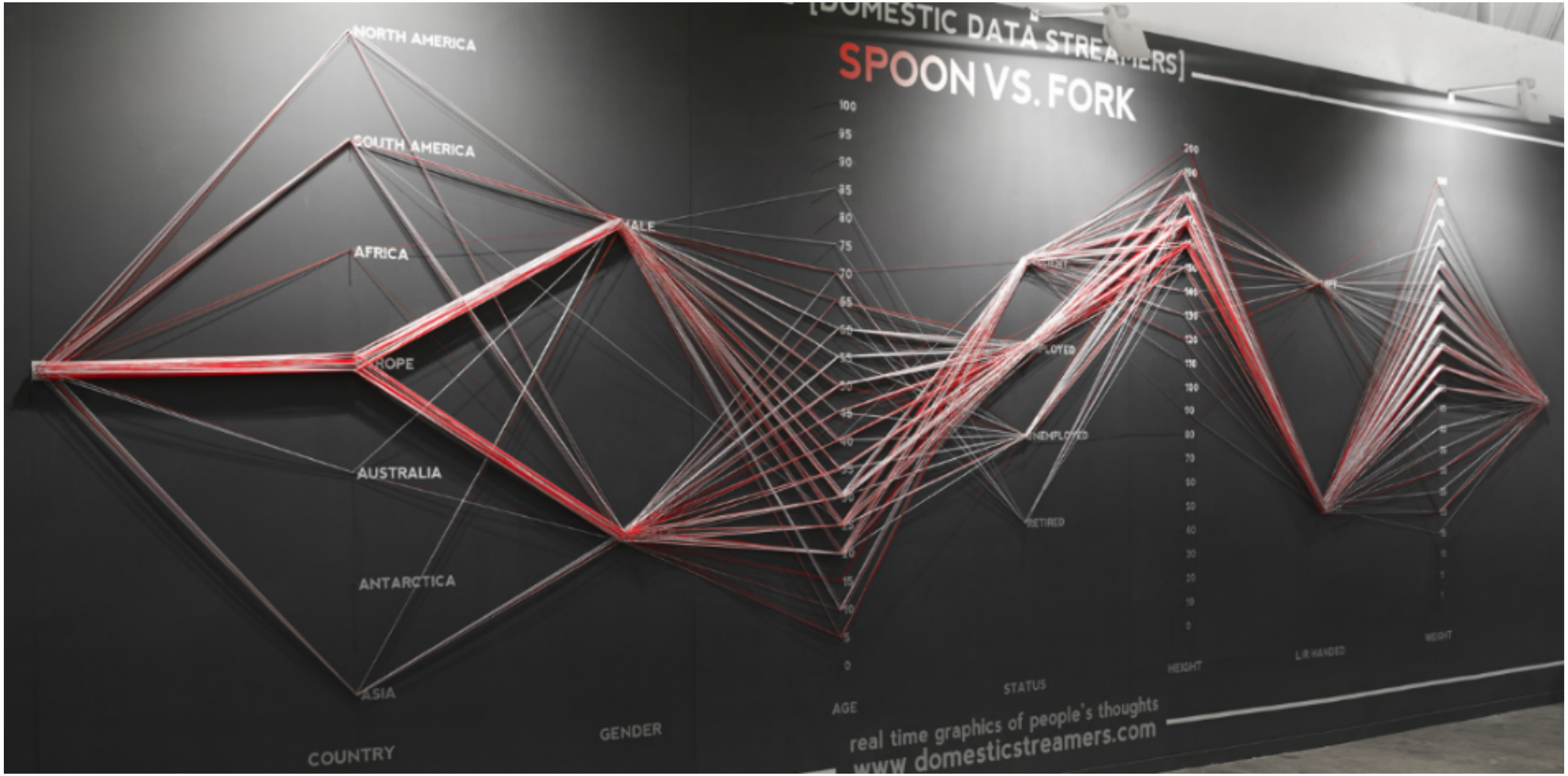
DeMar DeRozan



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

Spoon vs. Fork

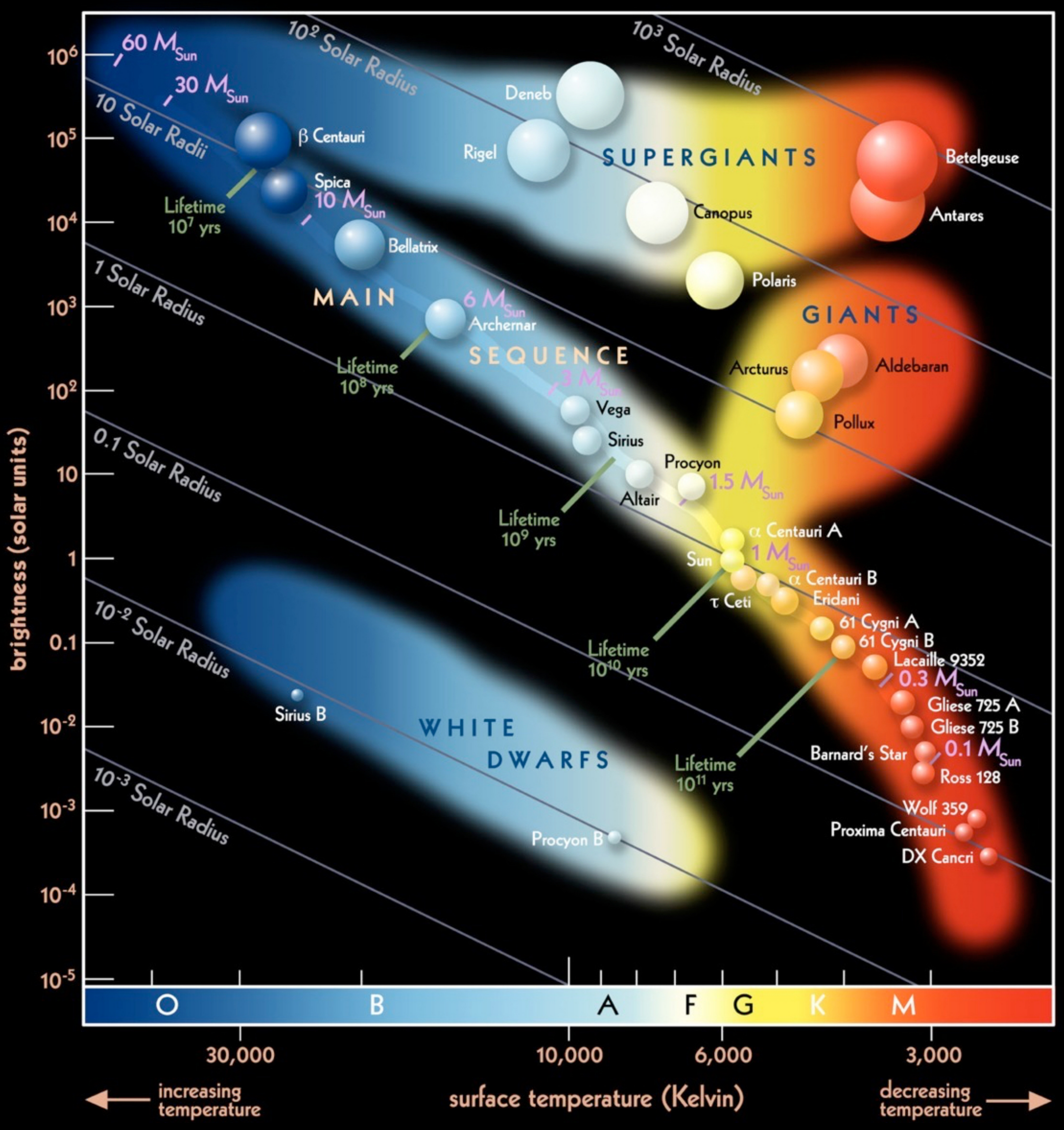
[domesticstreamers.com]



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

[European Southern Observatory]

Hertzprung-Russell Diagram



Data Elements

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

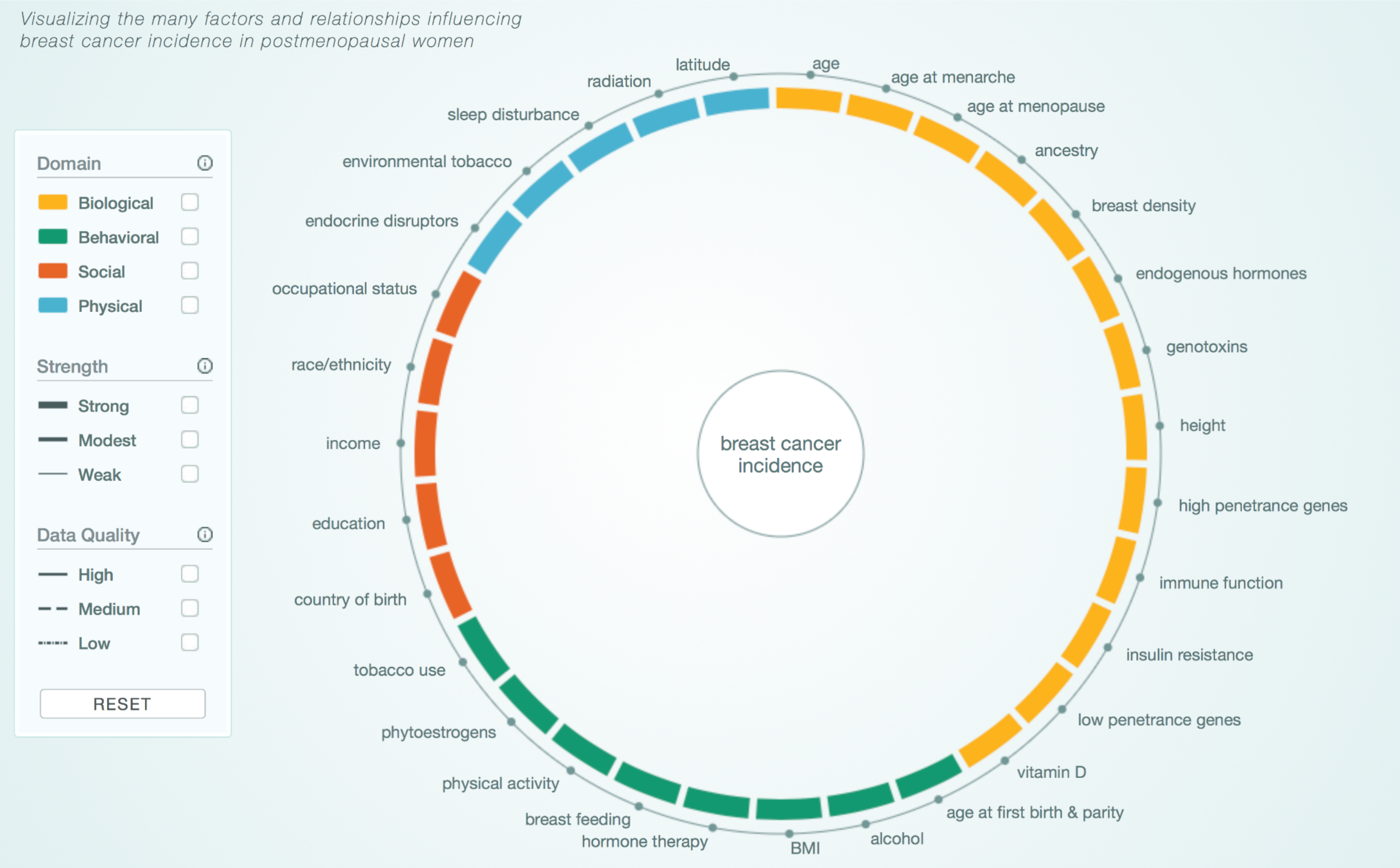
Underlying Structure

- 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 diagram.

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

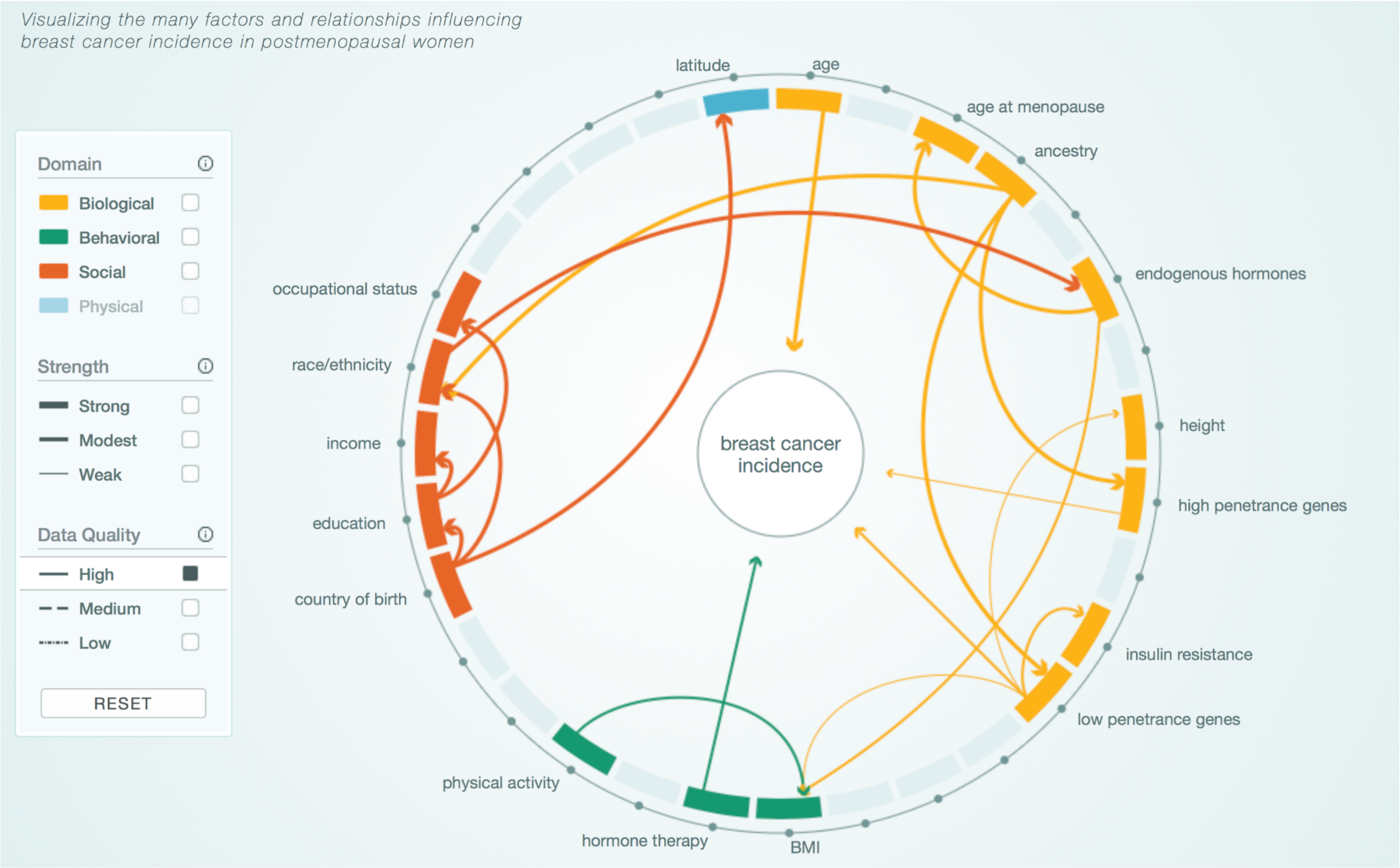
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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

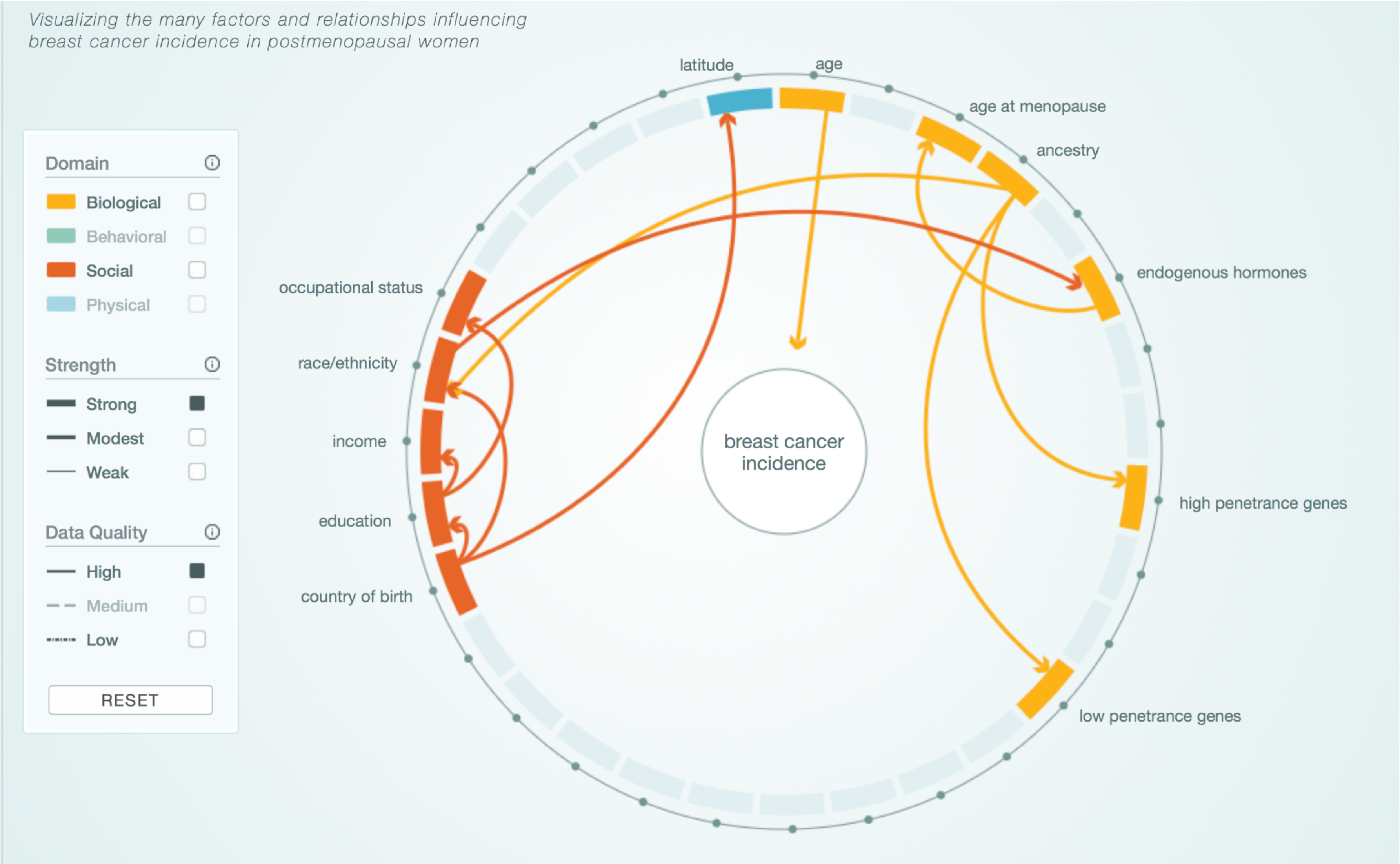
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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

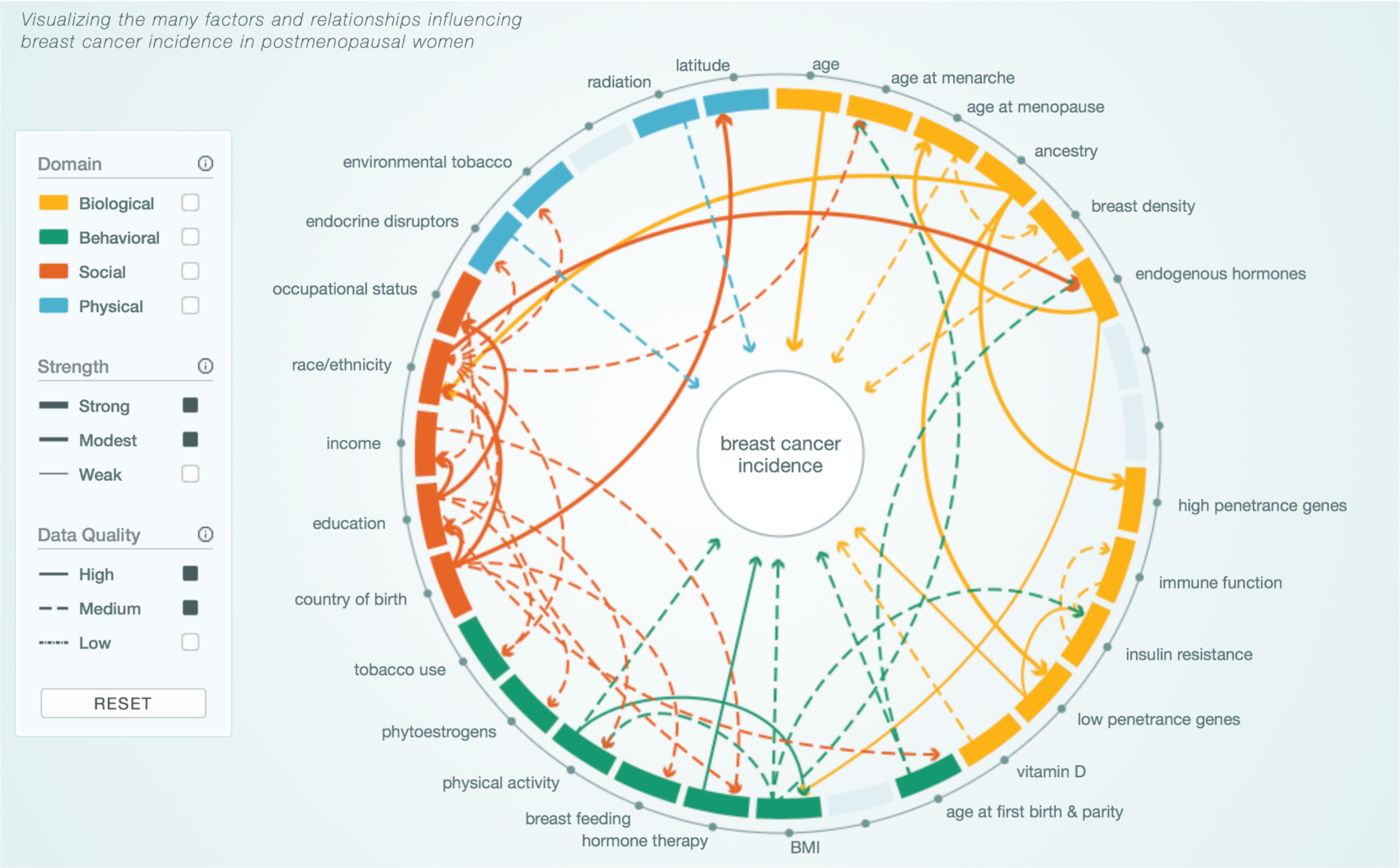
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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

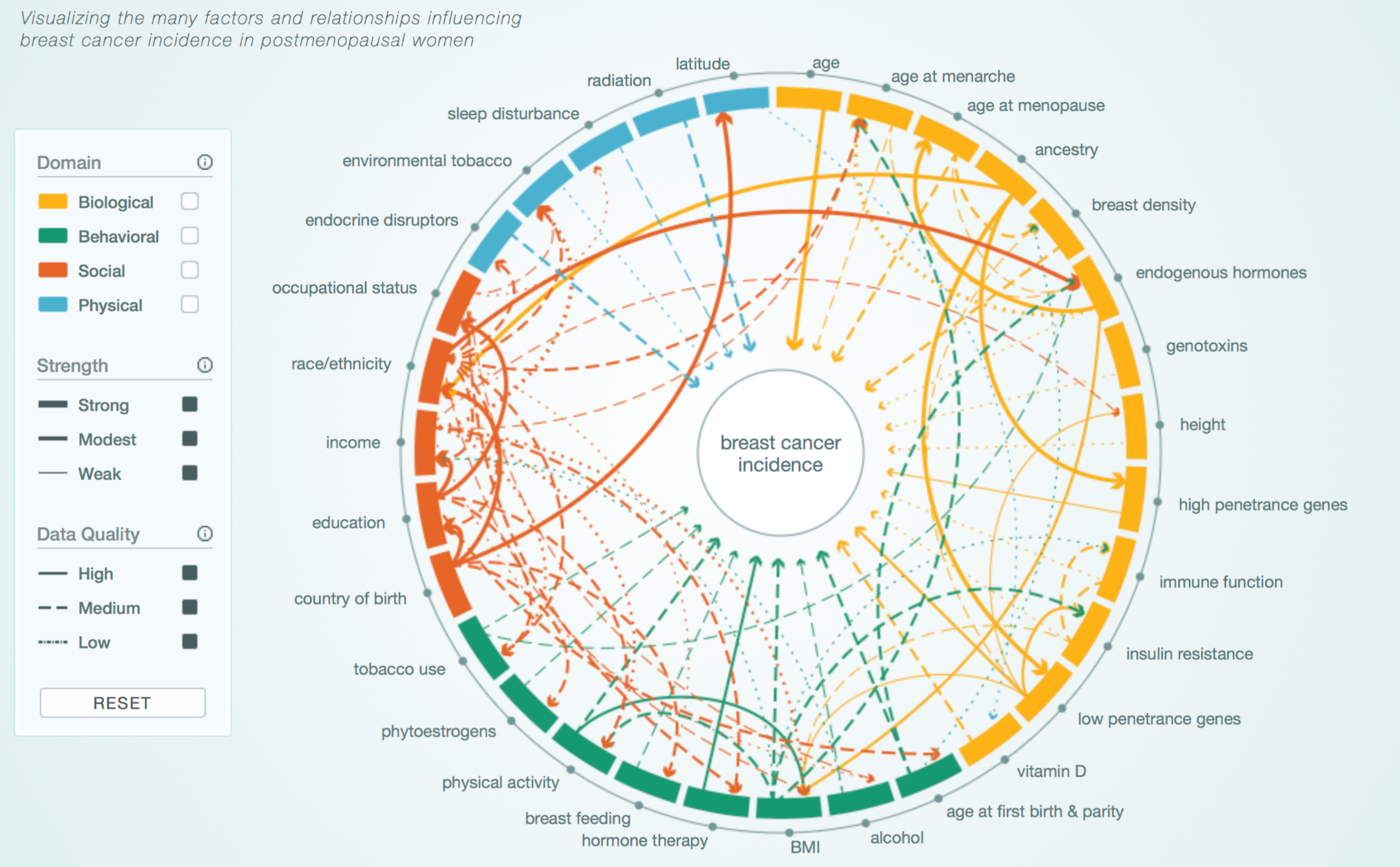
A Model of Breast Cancer Causation



Can you infer causality from this diagram?

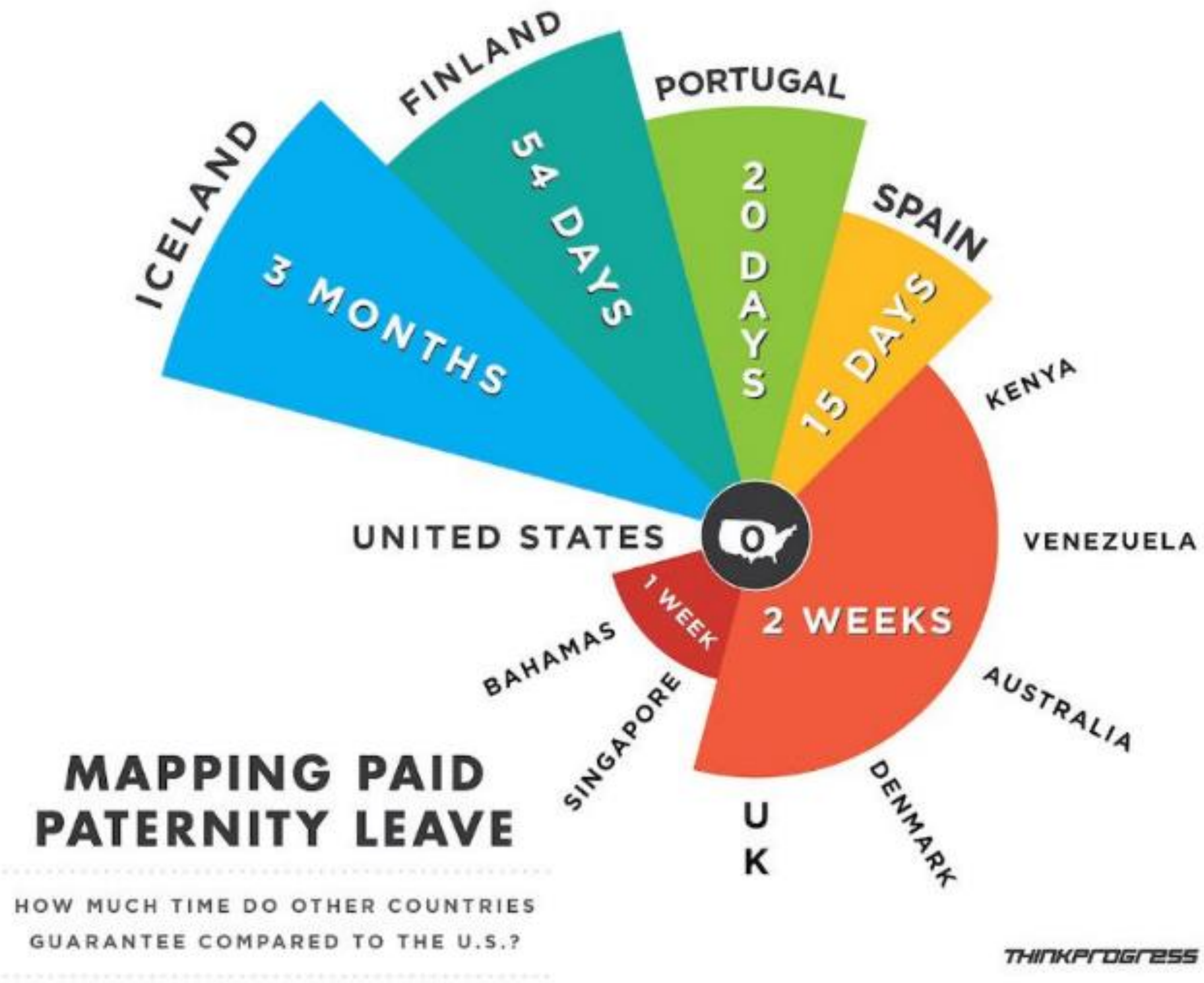
[http://www.cbcrp.org/causes/index.php]

A Model of Breast Cancer Causation



Can you infer causality from this diagram?

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



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

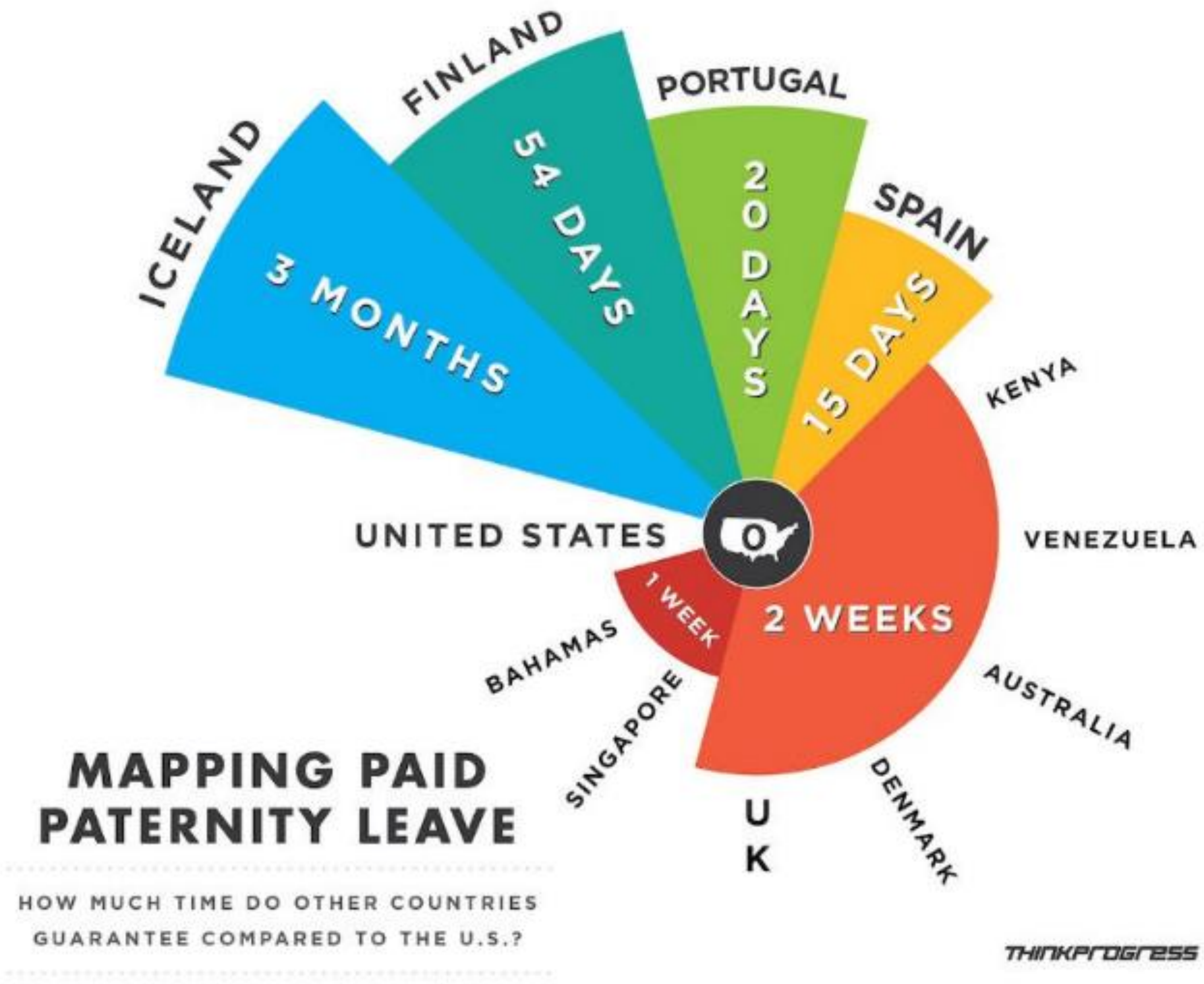
Low data density

High chartjunk ratio

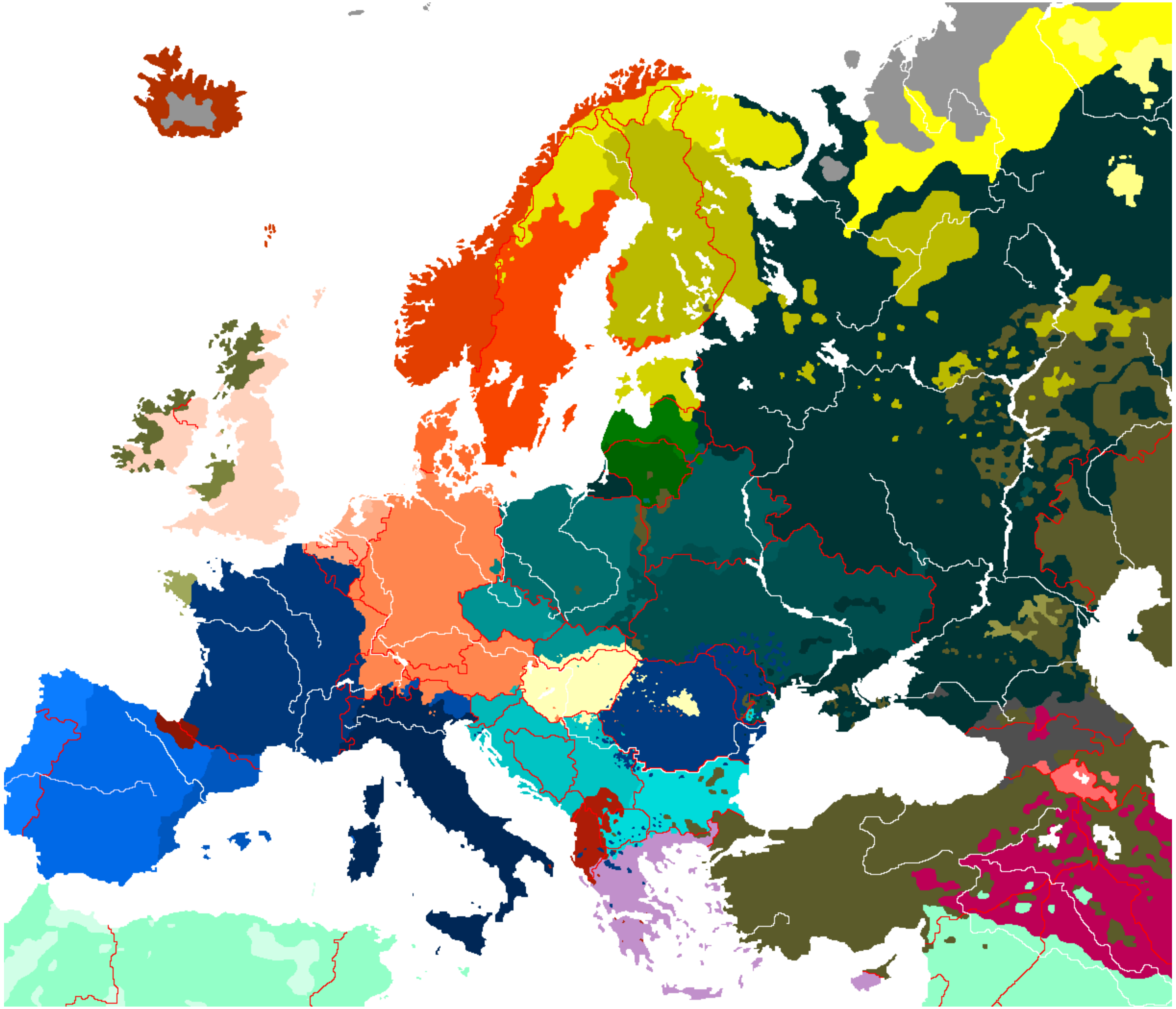
Scaling effects

Cherry-picking

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



[Found on Wikipedia, author unknown]



[Found on Wikipedia, author unknown]

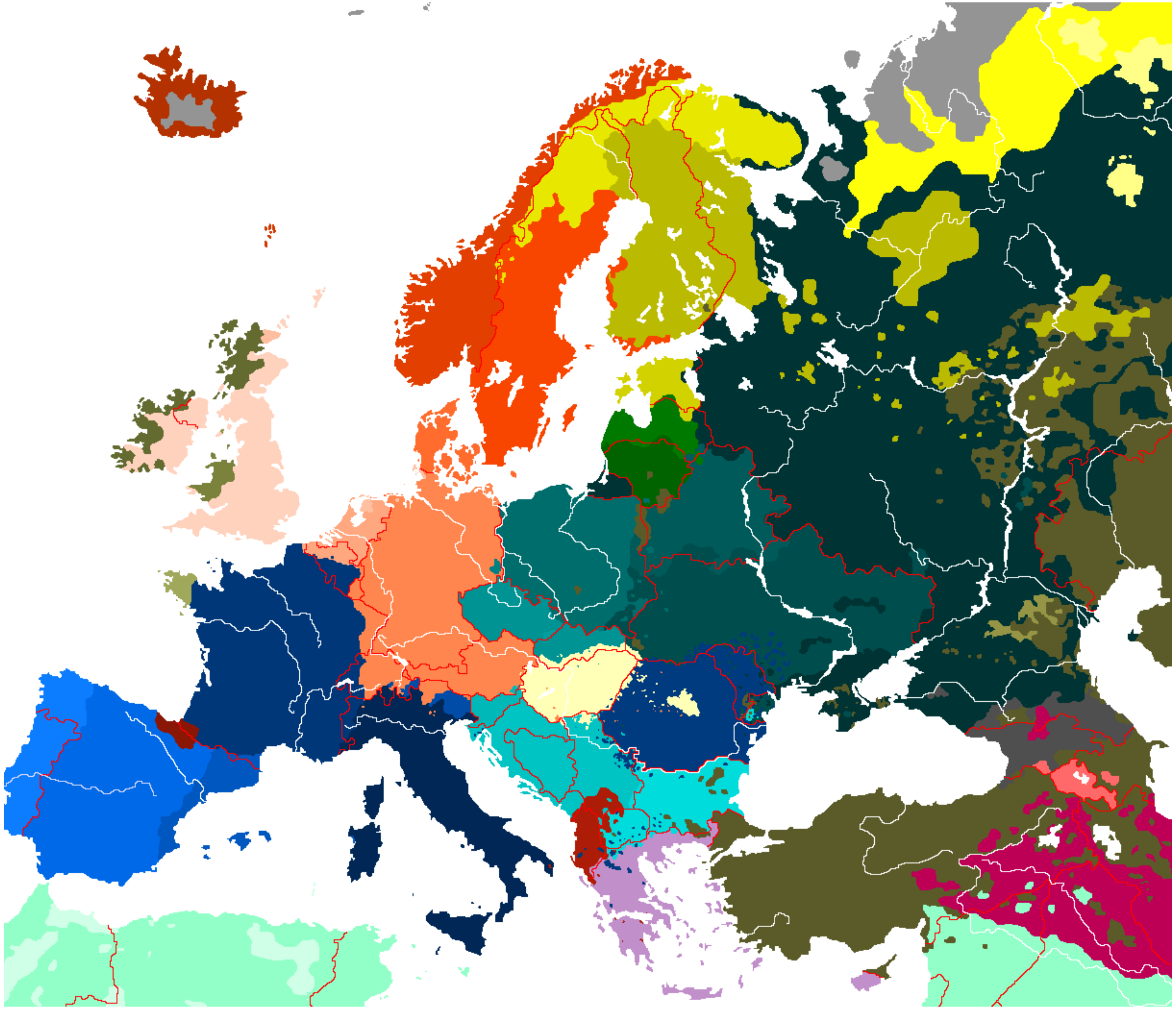
Encoding?

Population density?

Secondary languages?

Rivers?

No data source



THE GRAMMAR OF GRAPHICS

DATA EXPLORATION AND DATA VISUALIZATION

*not required

A GGPLOT2 PRIMER

ggplot2 is a set of tools that map data to visual display elements, and that allow the user to control the fine details of plot display.

Most important aspect: *ggplot2* can be used to think about the **logical structure** of the plot.

A *ggplot2* graph has 2 main components (and optional terms):

- aesthetic mappings (**aes** – connections between data and plot elems.)
- plot geometry (**geom** – specifies the type of plot)
- *facets, *coordinates, *scales, *labels, *guides, etc.

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

GGPLOT2 GRAMMAR

1. Tidy Data

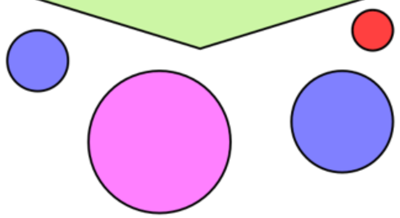
```
p <- ggplot(data = gapminder, ...
```

gdp	lifexp	pop	continent
340	65	31	Euro
227	51	200	Amer
909	81	80	Euro
126	40	20	Asia

2. Mapping

```
p <- ggplot(data = gapminder, mapping =
       aes(x = gdp, y = lifexp, size = pop,
       color = continent))
```

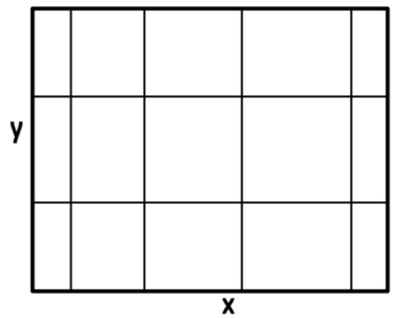
3. Geom



```
p + geom_point()
```

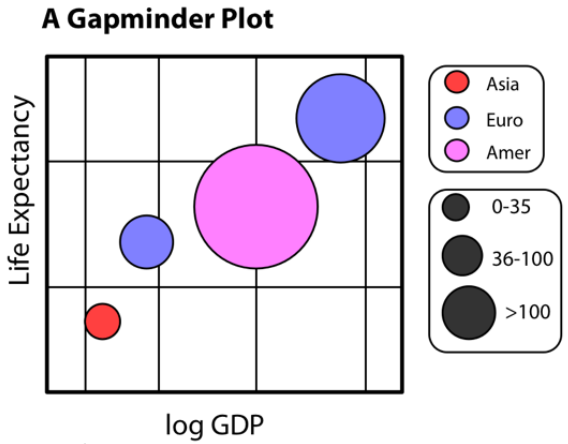
4. Co-Ordinates & Scales

```
p + coord_cartesian() + scale_x_log10()
```



5. Labels & Guides

```
p + labs(x = "log GDP", y = "Life
Expectancy", title = "A Gapminder Plot")
```



GGPLOT2 GRAMMAR – GEOMS

The data source and variables to be plotted are specified *via* `ggplot()`.

The various geom functions specify **how** these variables are to be visually represented

- using points, bars, lines, shaded regions, etc.

There are currently 37 available geoms.

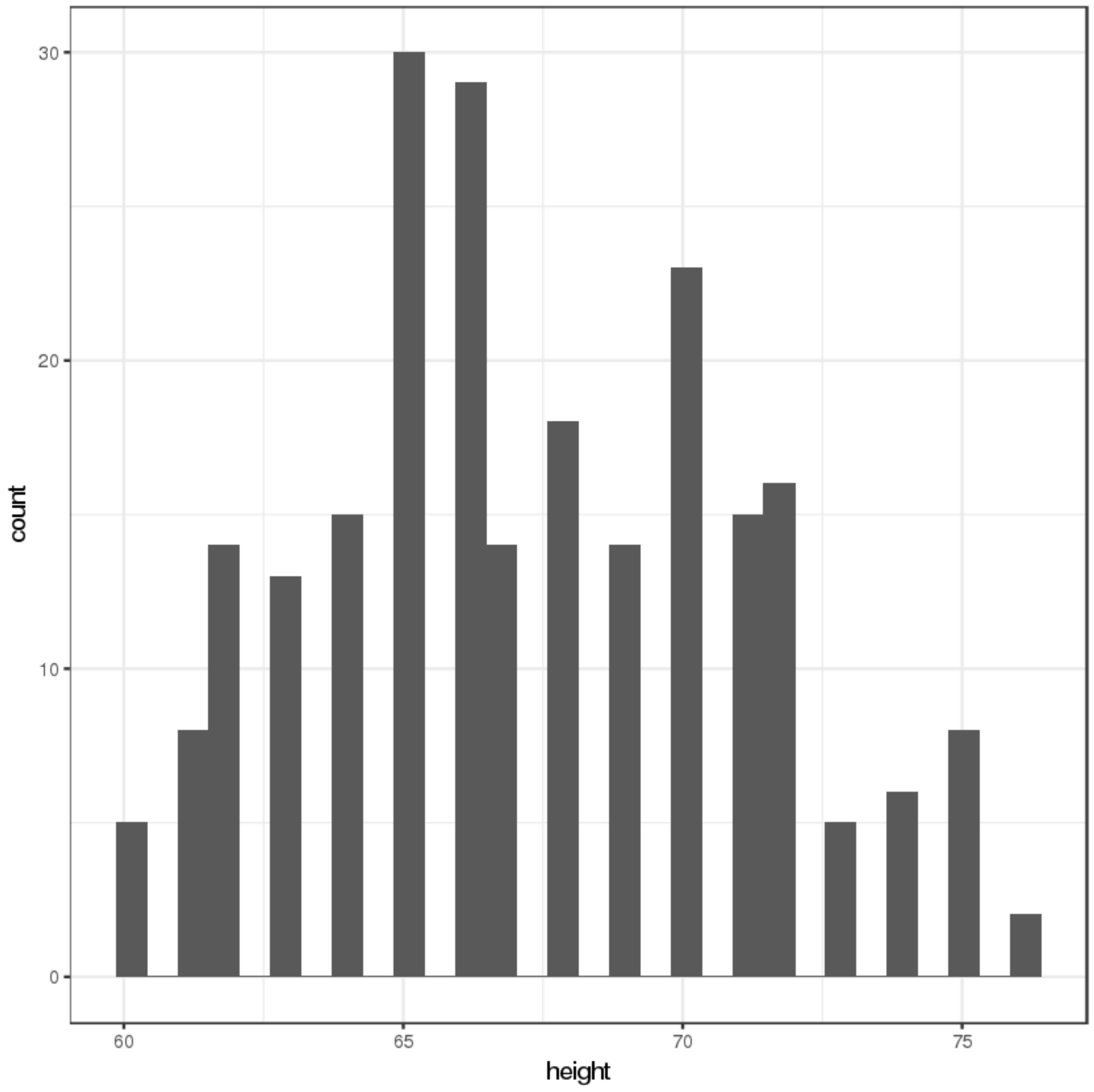
GGPLOT2 GRAMMAR – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
  geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
  geom_boxplot()
```

What do you expect the output to be?

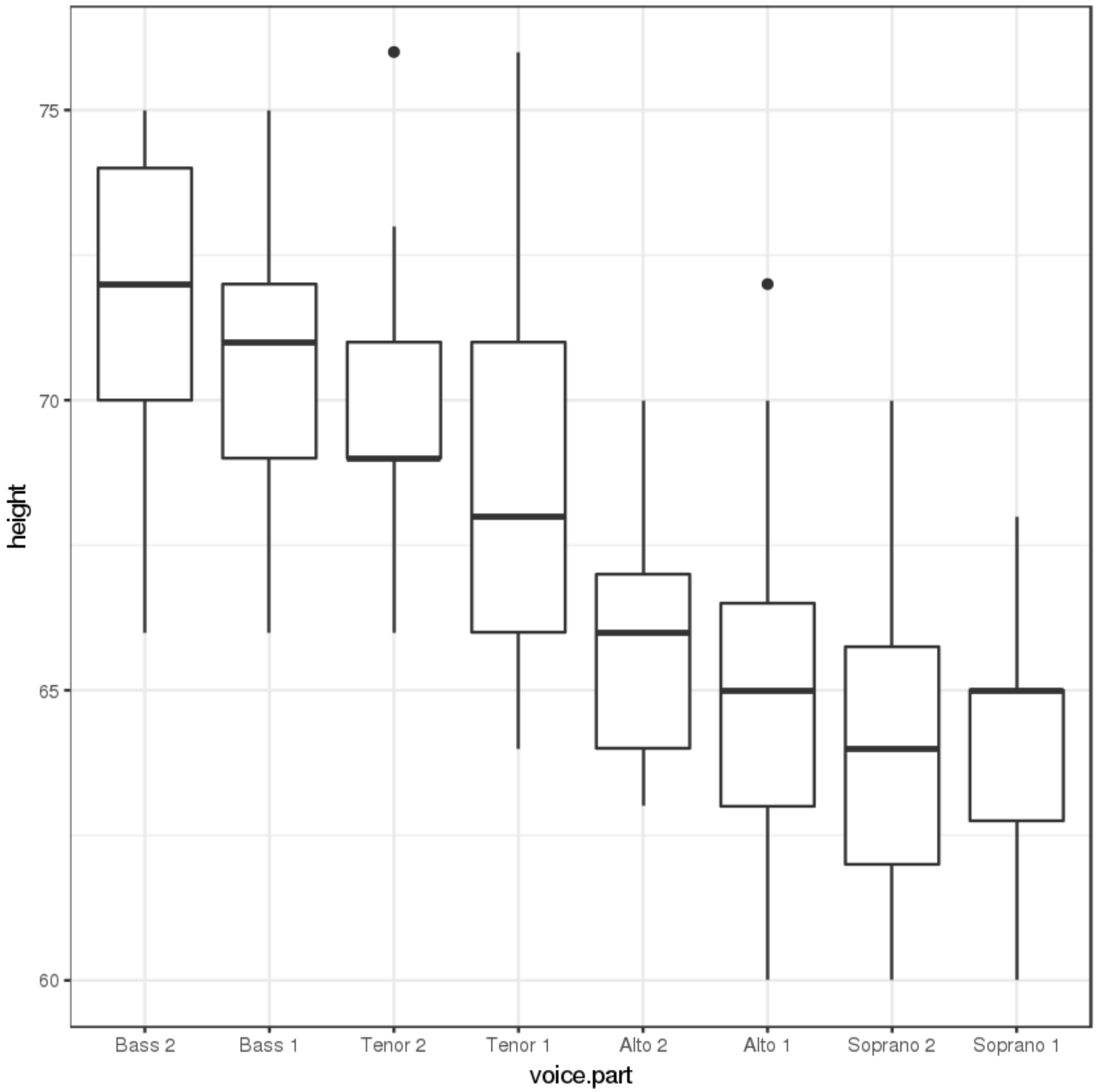
GGPLOT2 GRAMMAR – GEOM()

```
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
  geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
  geom_boxplot()
```



GGPLOT2 GRAMMAR – GEOM()

```r
library("ggplot2")
data(singer, package="lattice")
# Using data from the 1979 ed. of the
# New York Choral Society

# Histogram of heights
ggplot(singer, aes(x=height)) +
  geom_histogram()

# Boxplot of heights by voice part
ggplot(singer, aes(x=voice.part, y=height)) +
  geom_boxplot()
```



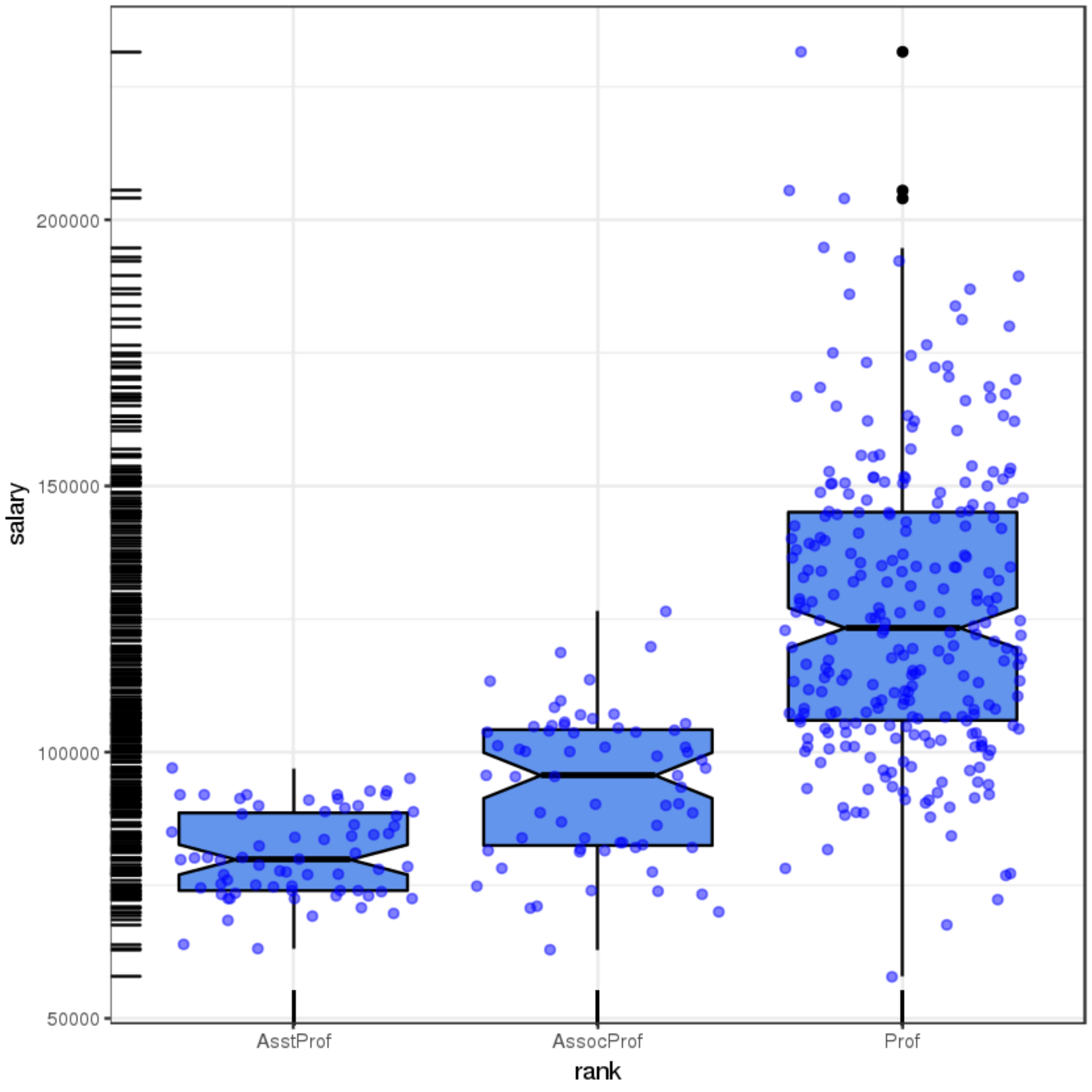
GGPLOT2 GRAMMAR – GEOM()

```r
library(ggplot2)
data(Salaries, package="car")
# Using data on salaries of a sample of
# US university professors (2018-2019)
# var: rank, sex, yrs.since.phd, yrs.service, salary

ggplot(Salaries, aes(x=rank, y=salary)) +
  geom_boxplot(fill="cornflowerblue",color="black", notch=TRUE) +
  geom_point(position="jitter", color="blue", alpha=.5) +
  geom_rug(side="l", color="black")
```

What do you expect the output to be?

GGPLOT2 GRAMMAR – GEOM()

```
library(ggplot2)
data(Salaries, package="car")
# Using data on salaries of a sample of
# US university professors (2018-2019)
# var: rank, sex, yrs.since.phd, yrs.service, salary

ggplot(Salaries, aes(x=rank, y=salary)) +
  geom_boxplot(fill="cornflowerblue",color="black", notch=TRUE) +
  geom_point(position="jitter", color="blue", alpha=.5) +
  geom_rug(side="l", color="black")
```



GGPLOT2 GRAMMAR – AESTHETICS

Aesthetics refer to the displayed attributes of the data.

They map the data to an attribute (such as the size or shape of a marker) and generate an appropriate legend.

Aesthetics are specified with the `aes()` function.

Aesthetics can be specified within the data function or within a geom. If they're specified within the data function then they apply to all specified geoms.

GGPLOT2 GRAMMAR – AESTHETICS

The aesthetics available to `geom_point()` (scatterplot), as an example, are:

- `x`, `y`, `alpha`, `color`, `fill`, `shape`, `size`

Important difference between specifying characteristics (like colour and shape) inside and outside the `aes()` function

- inside: assigned colour or shape automatically based on the data.
- outside: not mapped to data.

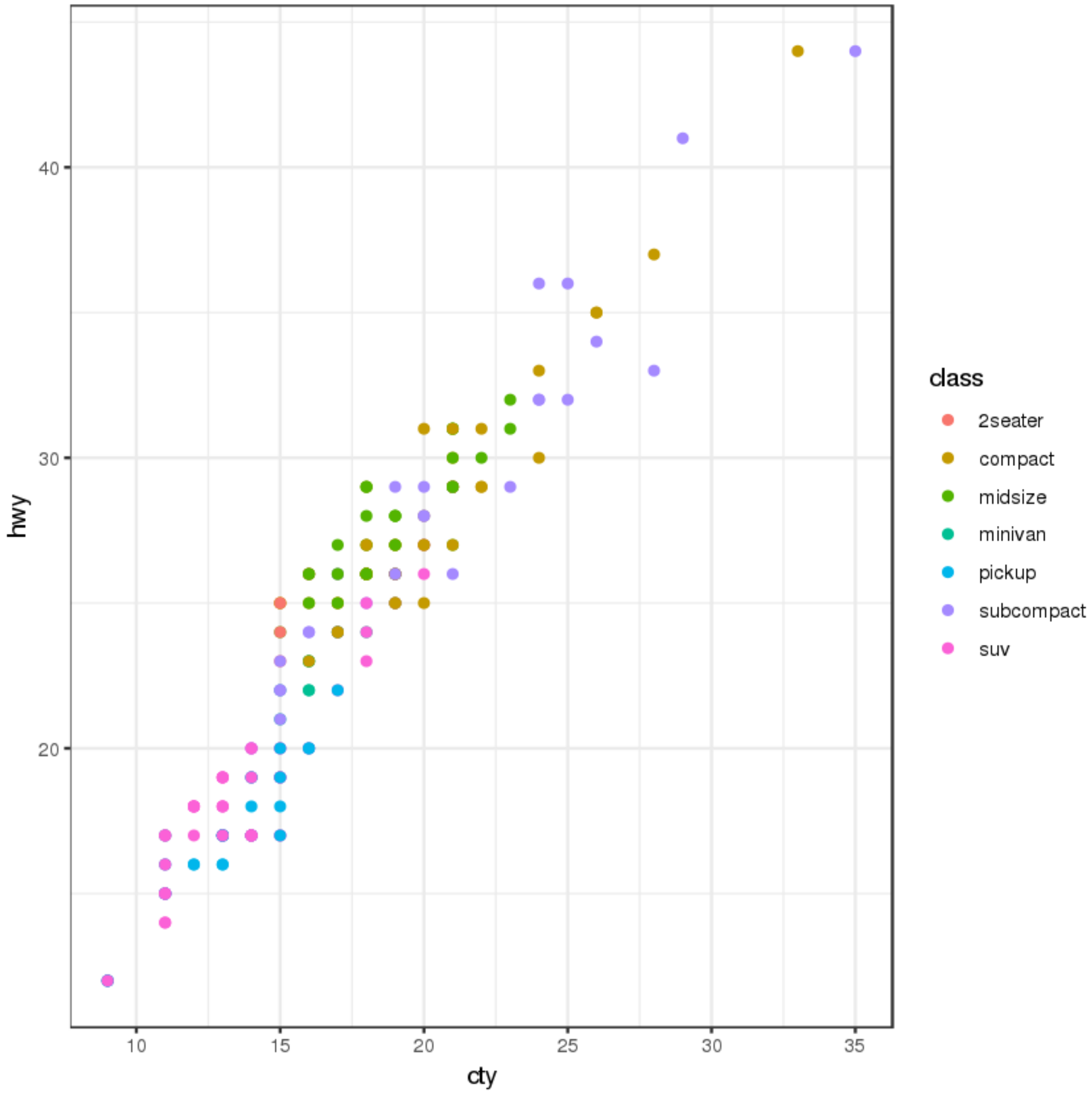
GGPLOT2 GRAMMAR – AES()

```
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```

What do you expect the output to be?

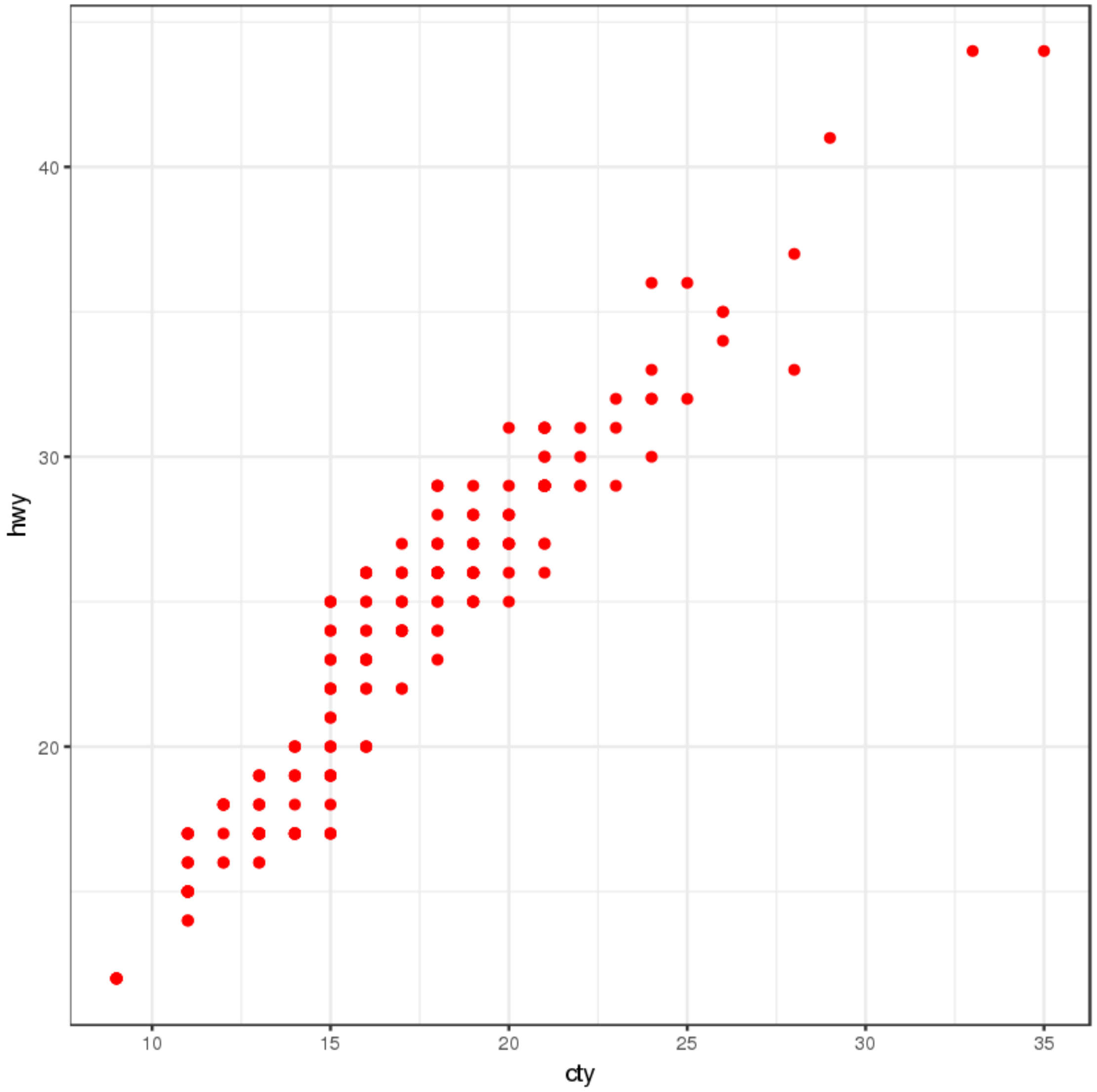
GGPLOT2 GRAMMAR – AES()

```r
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```



GGPLOT2 GRAMMAR – AES()

```r
library(ggplot2)
# Using the mpg dataset

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(aes(colour = class))

# specifying characteristics inside aes()
ggplot(mpg, aes(cty, hwy)) +
  geom_point(colour = "red")
```



AN INTRODUCTION TO DASHBOARDS

DATA EXPLORATION AND DATA VISUALIZATION

REPORTING AND DEPLOYMENT

An analysis can only be as good as how it is **communicated** and/or **deployed**.

Crucial Questions:

- Who is in receipt of the report(s)?
- How are the workflows deployed into production?
- Can data insights be turned into useful policies?

Automatic reporting should be audited and validated **regularly**.

REPORTING AND DEPLOYMENT

Communication should occur at various stages of the project, not solely upon completion:

- keep sponsors / clients aware of broad lines
- technical details may be avoided, but documented nonetheless

Ideal scenario: analysis software is also reporting software

- minimizes human error related to cut-and-paste
- removes the need for keeping analysis and reporting separate
- makes sharing the work with other project member easier

Simplify the process further by deploying directly to the Web.

DISCUSSION

What are your favourite reporting tools?

How much should you test a product before deployment?

What's the cost of deploying a faulty product?

[Wexler, Shaffer, Cotgreave, *The Big Book of Dashboards*]

DASHBOARDS

A **dashboard** is any visual display of data used to monitor conditions and/or facilitate understanding.

Examples:

- interactive display that allows people to explore motor insurance claims by city, province, driver age, etc.
- PDF showing key audit metrics that gets e-mailed to a Department's DG on a weekly basis.
- wall-mounted screen that shows call centre statistics in real-time.
- mobile app that allow hospital administrators to review wait times on an hourly- and daily-basis for the current year and the previous year.

SOME QUESTIONS TO CONSIDER

In a car's dashboard, a small number of **key indicators** (speed, gasoline level, lights, etc.) need to be understood **at a glance**. A dashboard design that does not take these two characteristics under consideration can have catastrophic consequences.

The following questions need to be answered prior to the dashboard being designed:

- Who is the dashboard's **consumer**?
- What **story** does the dashboard tell?
- What data (categories) will be used?
- What will **appear** on the dashboard?
- How can the dashboard **help** the consumer?

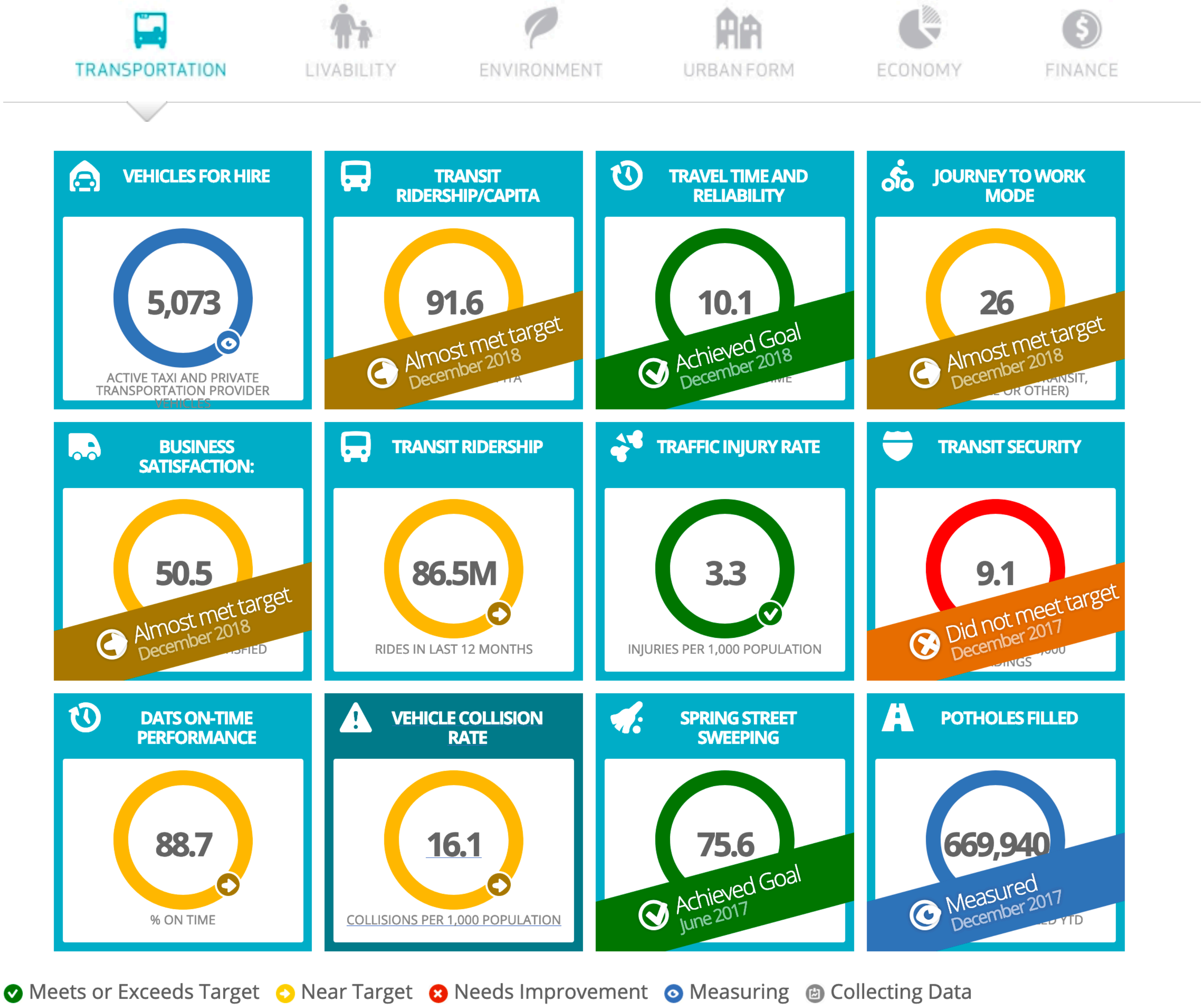
[https://www.geckoboard.com/blog/building-great-dashboards-6-golden-rules-to-successful-dashboard-design]

DASHBOARD DESIGN GUIDELINES

Nick Smith suggests the following 6 Golden Rules:

- **Consider the audience** (who are you trying to inform? does the DG really need to know that the servers are operating at 88% capacity?)
- **Select the right type of dashboard** (operational, strategic/executive, analytical)
- **Group data logically, use space wisely** (split functional areas: product, sales/marketing, finance, people, etc.)
- **Make the data relevant to the audience** (scope and reach of data, different dashboards for different departments, etc.)
- **Avoid cluttering the dashboard** (present the most important metrics only)
- **Refresh your data at the right frequency** (real-time, daily, weekly, monthly, etc.)

[https://dashboard.edmonton.ca]

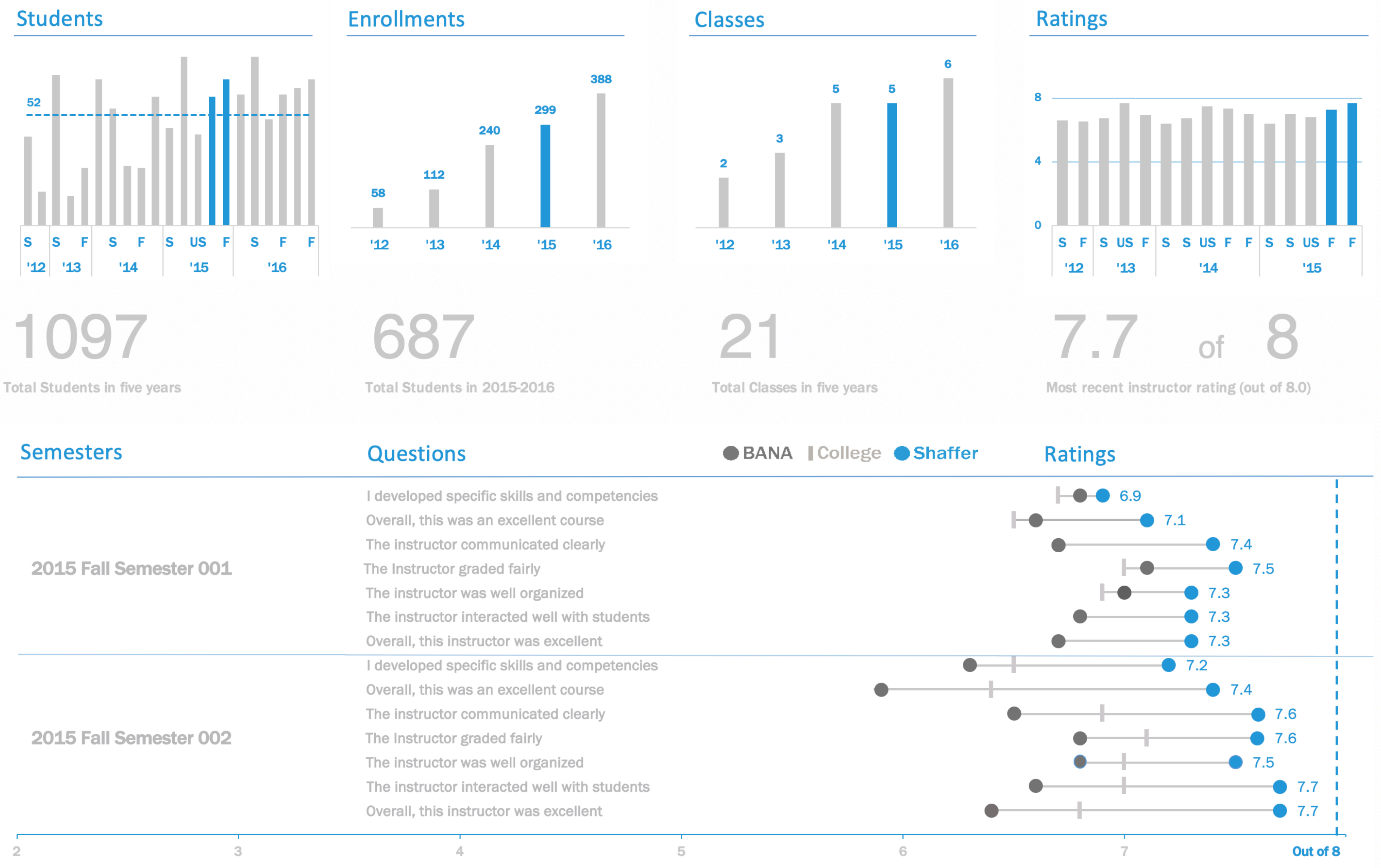


IDLEWYLD Sysabee DAVHILL

data-action-lab.com

Course Metrics

[https://bigbookofdashboards.com/dashboards.html]



Course Metrics Dashboard created by Jeffrey A. Shaffer. Data from University of Cincinnati Course Evaluations. Blue indicates the 2 most recent rating periods.

COURSE METRICS DASHBOARD – STRENGTHS

Easy-to-see key metrics

Simple color scheme

Potential to be static or interactive

Both overview and details are clear

DISCUSSION

There are no perfect dashboards – no collection of charts will ever suit everyone who encounters it.

All dashboards should be **truthful** and **functional**, but dashboards that are also **elegant** (delightful, enjoyable) will take you further.

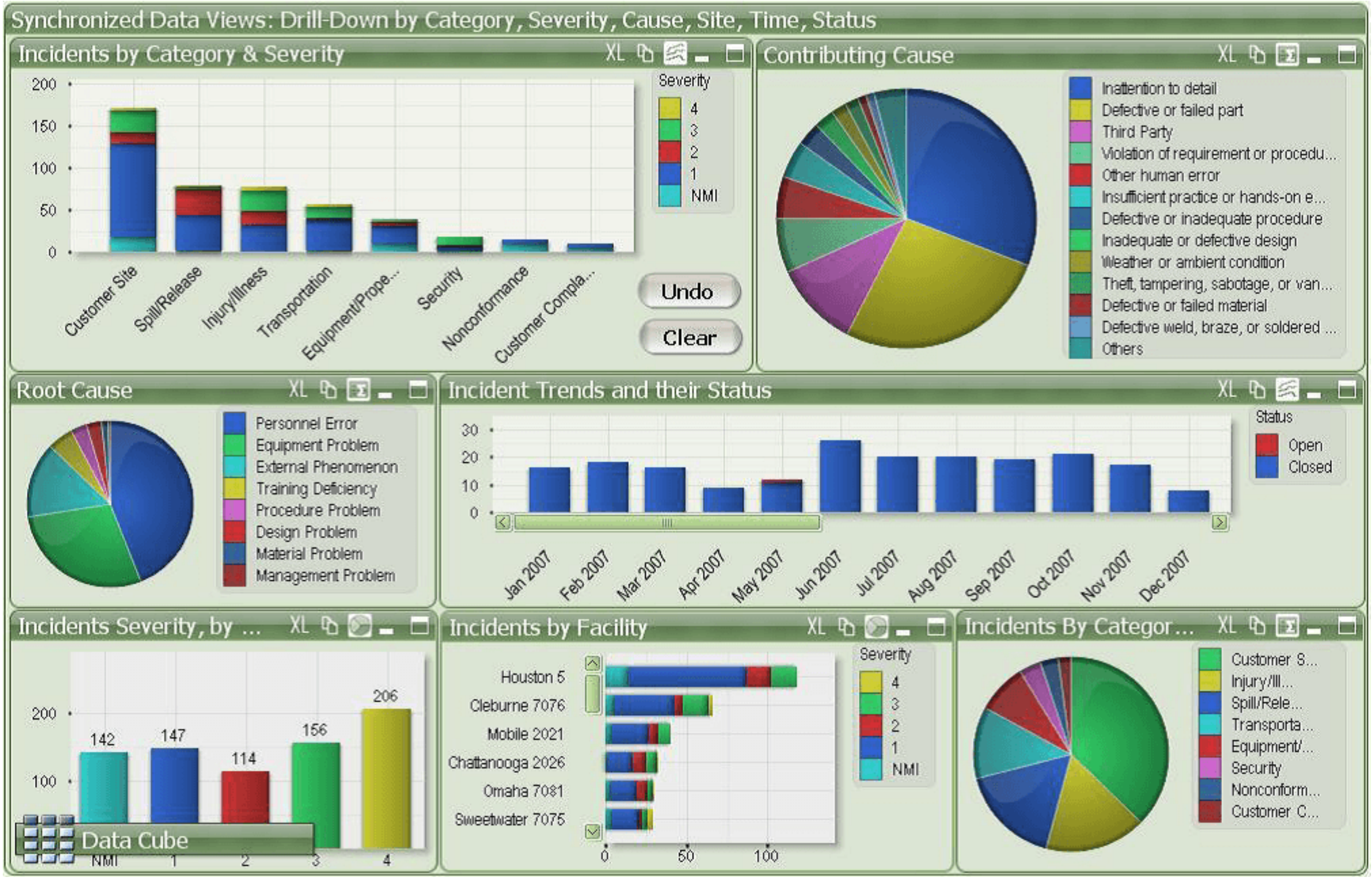
All dashboards are **incomplete**. Good dashboards will still lead to dead ends, but they should allow users to ask: "Why? What is the root cause of a problem?"

Tools: Excel, Power BI, Tableau, R + Shiny, Geckoboard, Matillion, etc.

EXERCISE

Consider the following dashboards. Can you figure out, at a glance, who their audience is? What are their strengths? What are their limitations? How could you improve them?

[https://www.matillion.com/wp-content/uploads/2014/11/qlikview-poor-use-of-dashboard-software.png]



[https://www.geckoboard.com/assets/2-terrible-dashboard-example-min.png]

WHAT'S WRONG?

Dashboard #1: not glanceable, overuse of colour, pie charts??

Dashboard #2: 3D visualizations, distracting borders and background, lack of filtered data, insufficient labels and context.