DATA EXPLORATION AND DATA VISUALIZATION

SETTING THE STAGE

OUTLINE

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

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

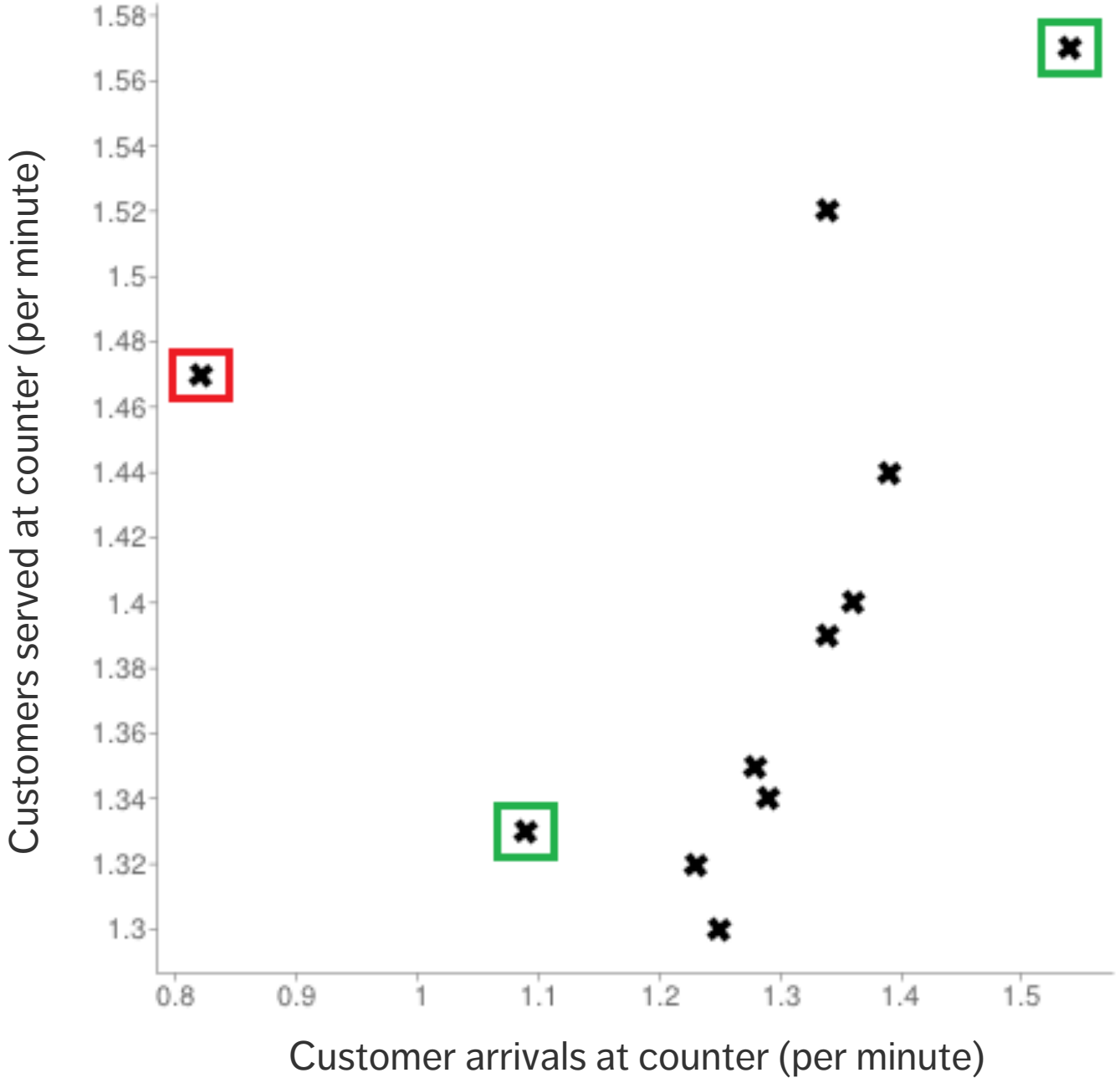
```
       Cl                NO3                NH4
 Min.   :  0.222   Min.   : 0.000   Min.   :    5.00
 1st Qu.: 10.994   1st Qu.: 1.147   1st Qu.:   37.86
 Median : 32.470   Median : 2.356   Median :  107.36
 Mean   : 42.517   Mean   : 3.121   Mean   :  471.73
 3rd Qu.: 57.750   3rd Qu.: 4.147   3rd 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 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]

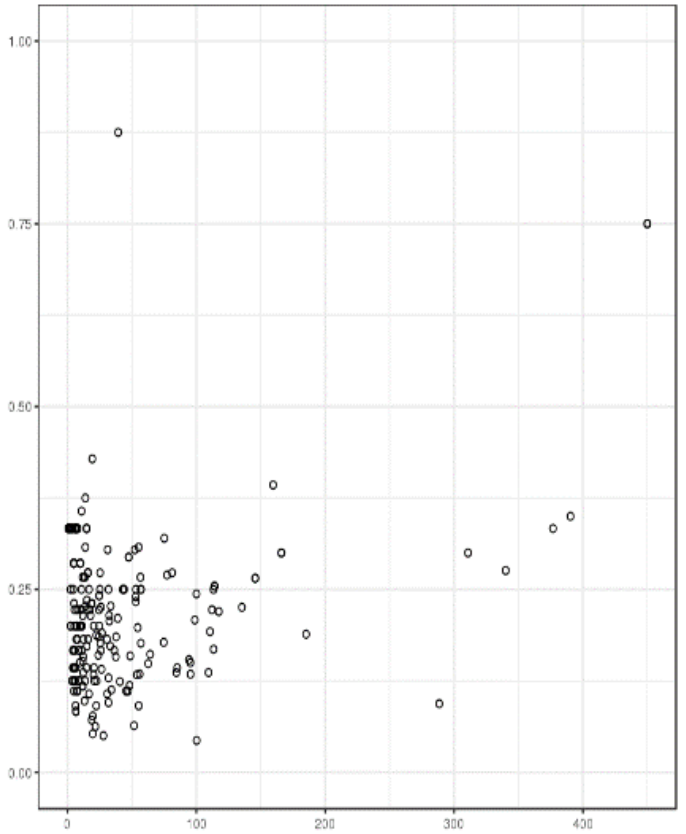


[http://promise.site.uottawa.ca/SERepository/datasets/cm1.arff]

REPRESENTING MULTIVARIATE OBSERVATIONS

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

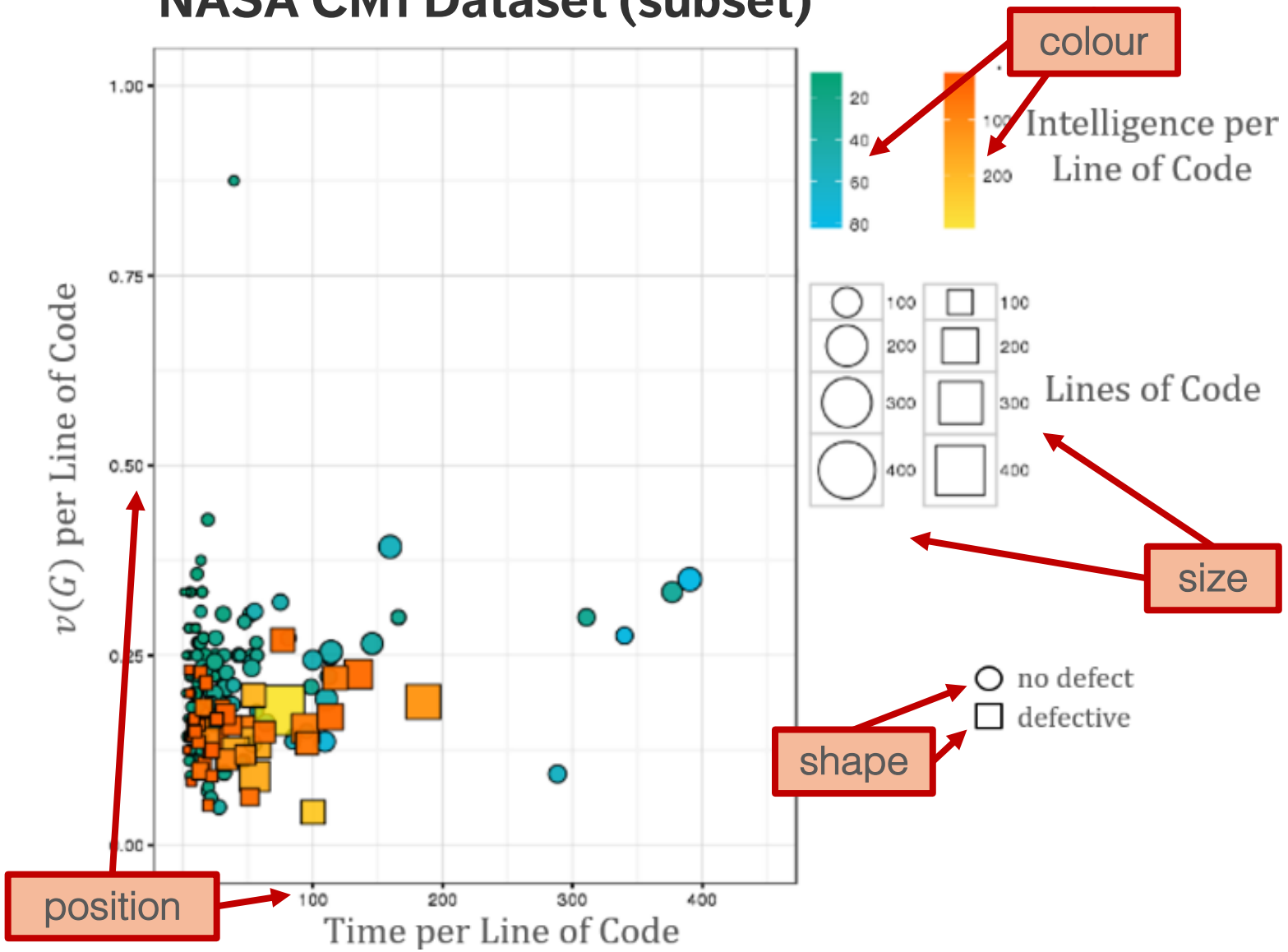
- size
- color
- value
- texture
- line orientation
- shape
- (motion?)



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

[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

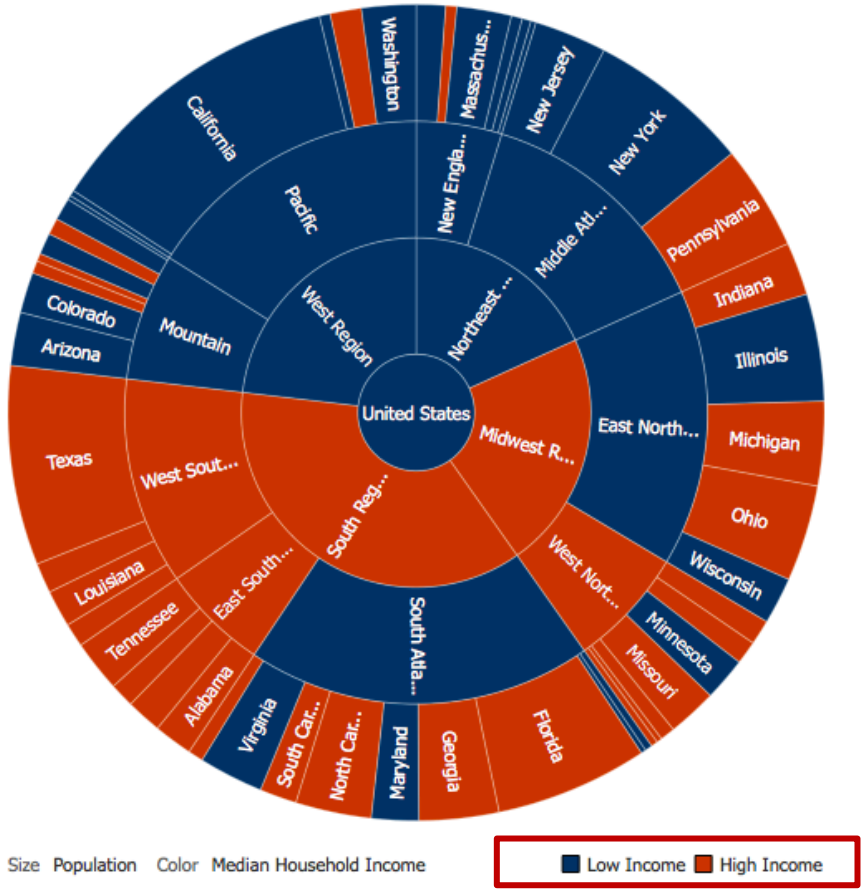
A **method**, as well as an item (**objective**)

Typically focuses on the **quantifiable**

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

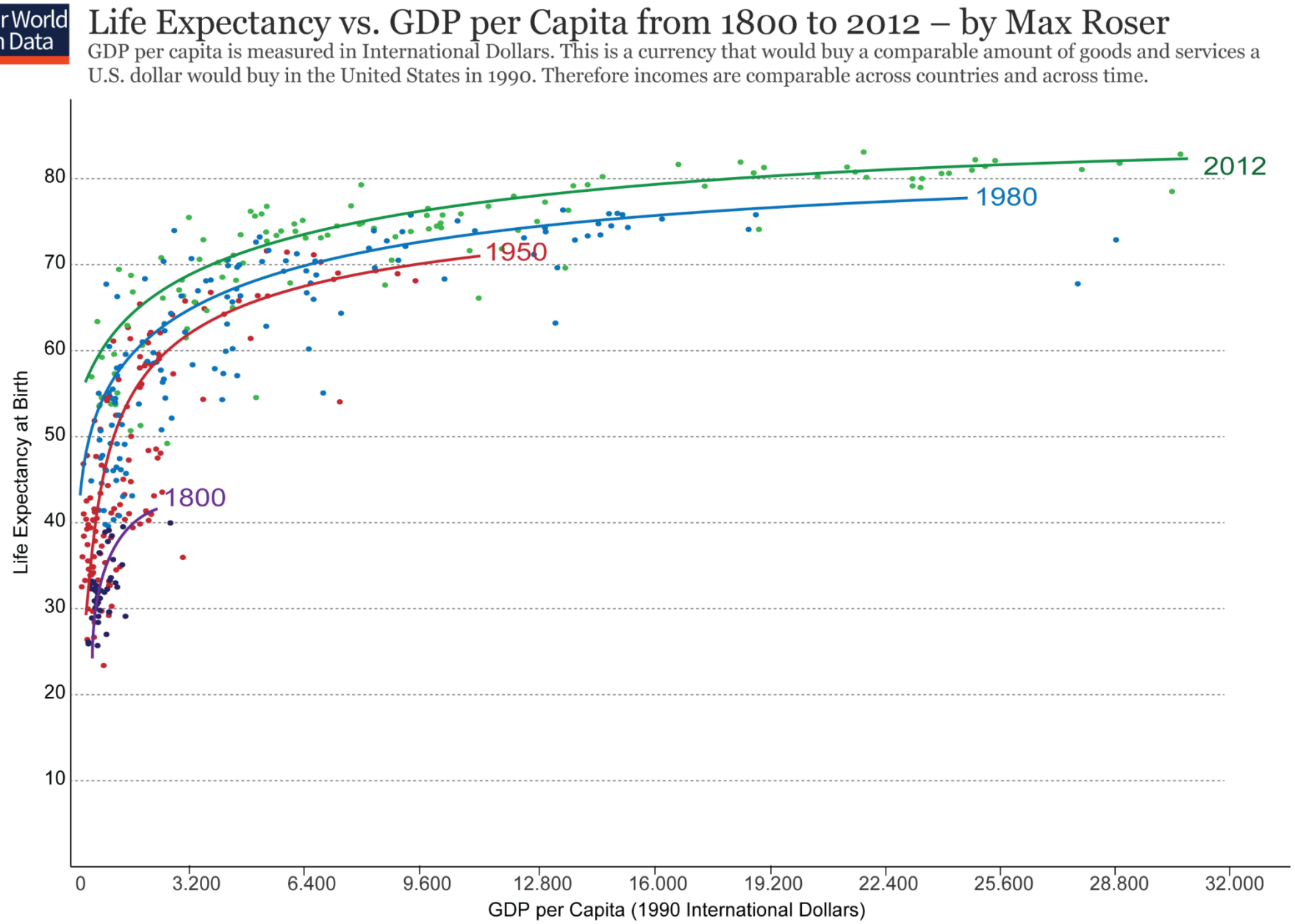
May be generated automatically

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



?

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



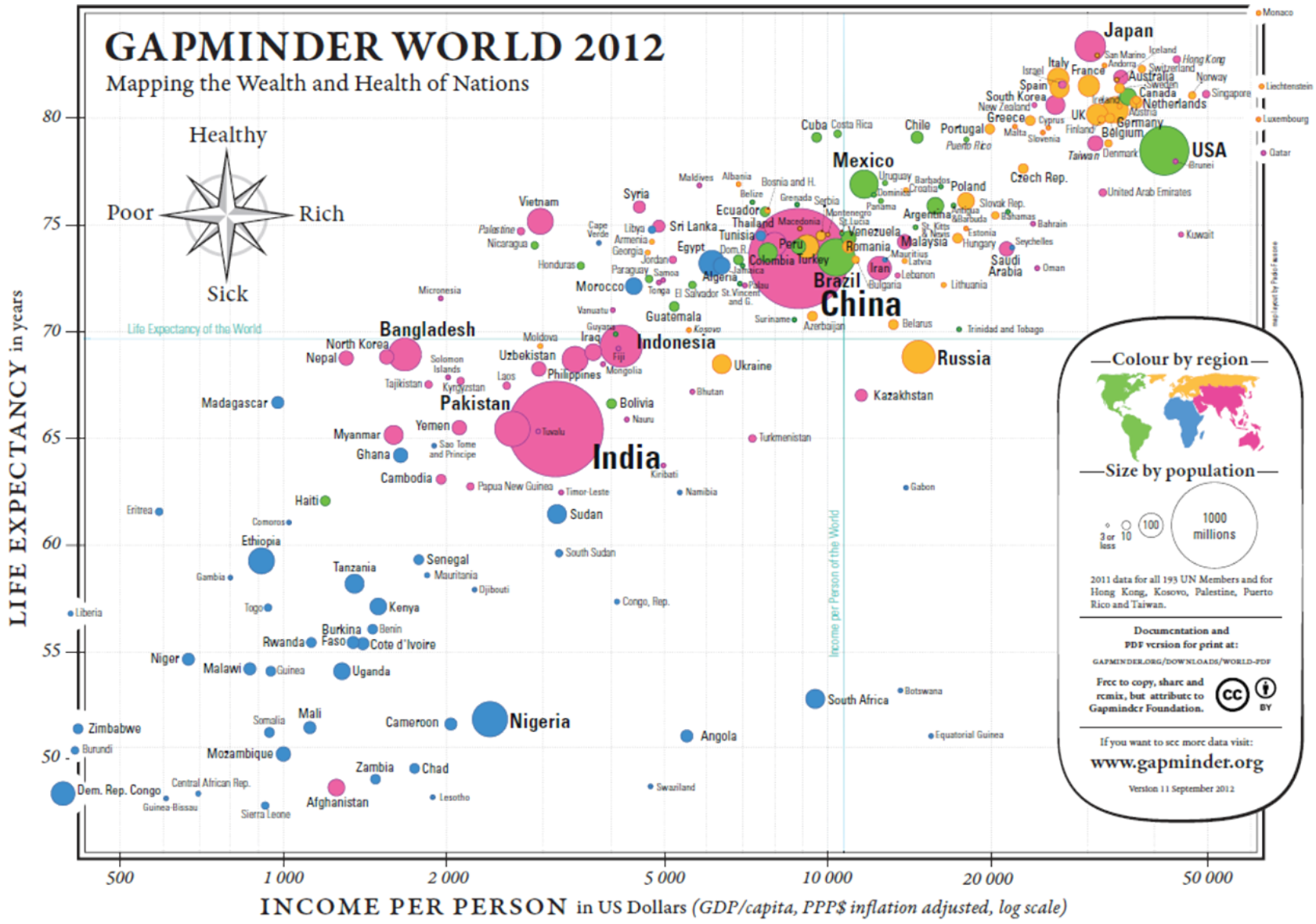
This graph displays the correlation between life expectancy and GDP per capita.

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 capita increases.

Data sources: Data on life expectancy are from Gapminder.org; data on GDP per capita are from the 'New Maddison Project Database'.
The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.

[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

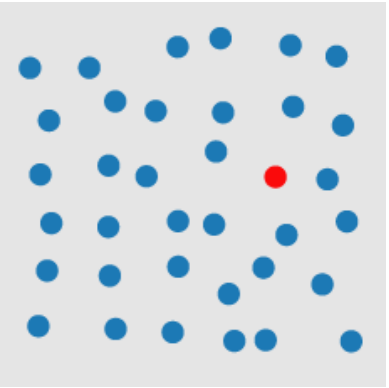
- **Pre-attentive processes:** fast, instinctive, efficient, multitasking gather information and build patterns:

  features → patterns → objects

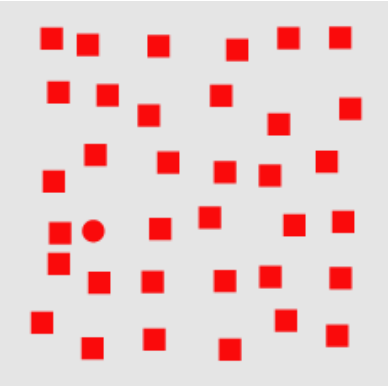
- **Attentive process:** slow, deliberate, focused discover features in the patterns:

  objects → patterns → features

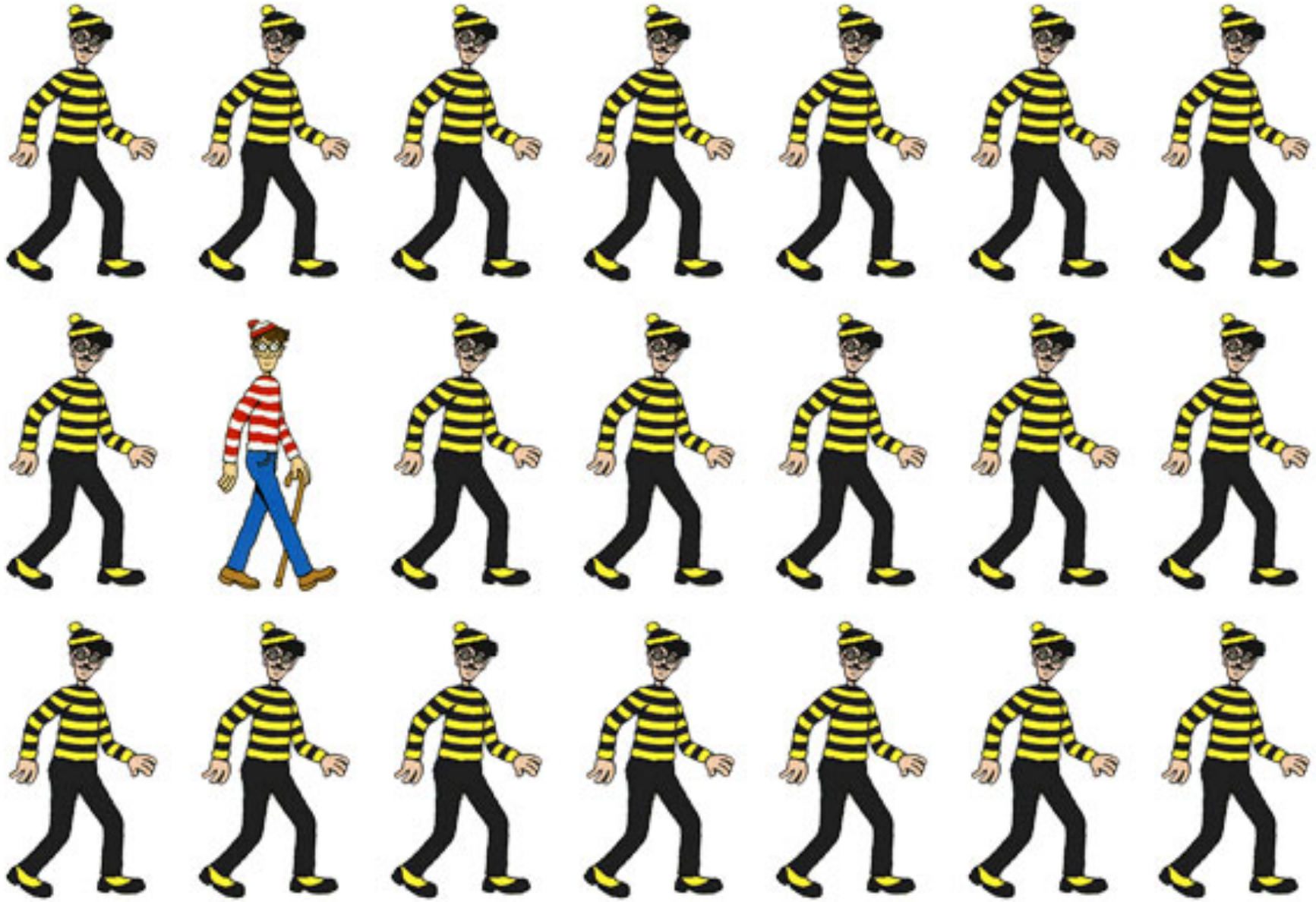
pre-attentive



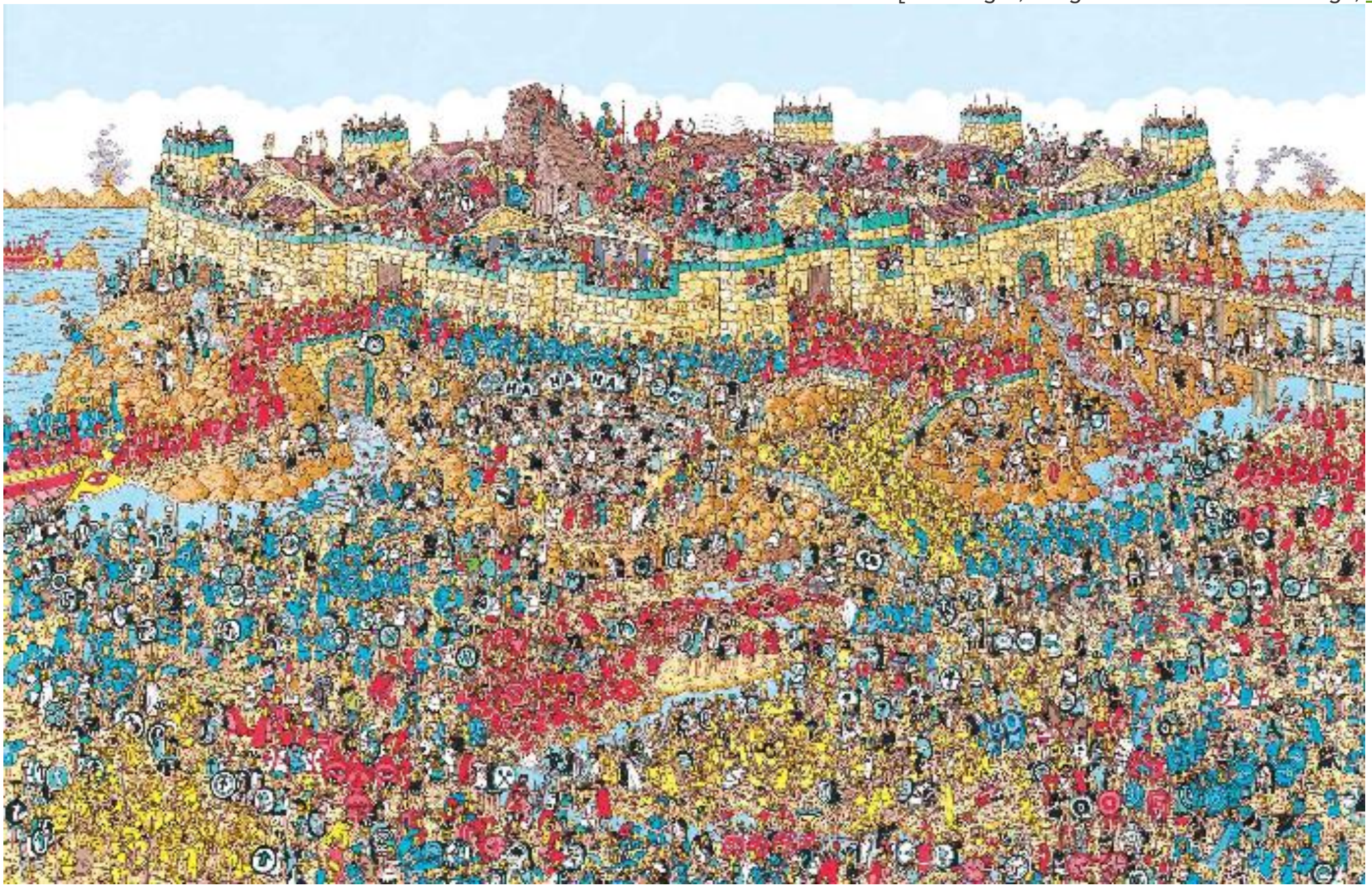
attentive



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



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



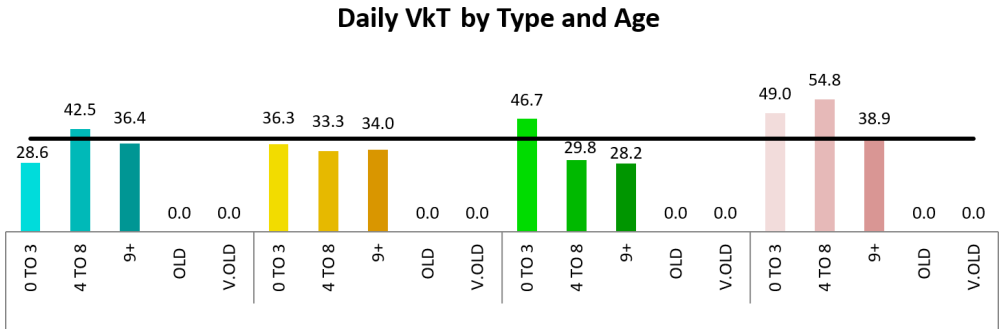
BASIC RULES

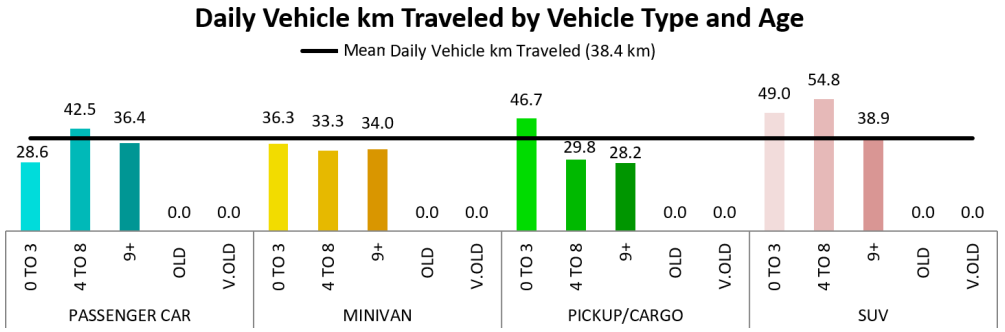
1. Check the data

outliers, spikes, anomalies

2. Explain encoding

don't assume the reader knows what everything means





3. Label axes

knowing the scale is important

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

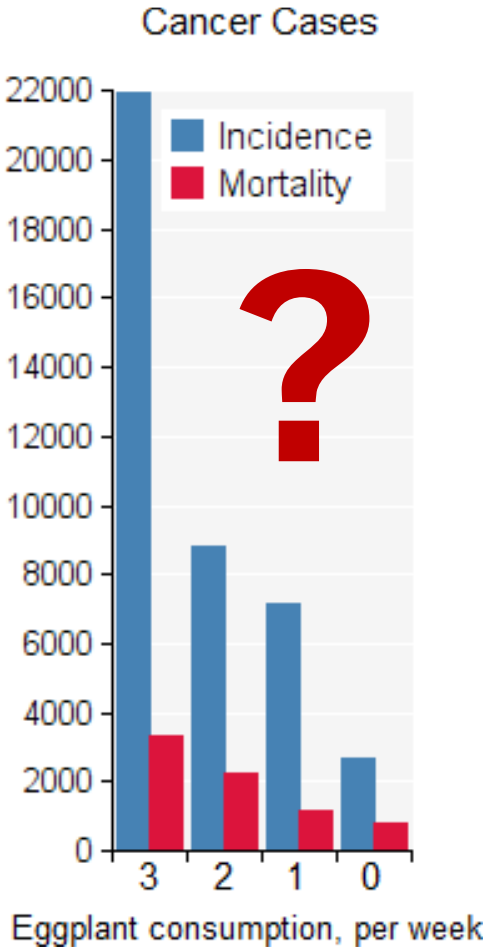
BASIC RULES

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



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

Heatmaps/Choropleths

Parallel Coordinates (Spaghetti Plots)

Maps and Distortions

Bubble Charts

Word Clouds and Other Text Visualizations

Network Diagrams

Sparklines

Small Multiples

Interactive Displays

Animated Charts

etc.

Consult the slide deck for examples.

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).

[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