MODULE 3: DATA ANALYSIS AND VISUAL STORYTELLING

CT ACADEMY | DATA ACTION LAB

8. DATA ANALYSIS

DATA ANALYSIS AND VISUAL STORYTELLING

CONTINGENCY/PIVOT TABLES

Contingency table: examines the relationship between two categorical variables via their relative (cross-tabulation).

Pivot table: a table generated by applying operations (sum, count, mean, etc.) to variables, possibly based on another (categorical) variable.

Contingency tables are special cases of pivot tables.

	Large	Medium	Small
Window	1	32	31
Door	14	11	0

Type	Count	Signal avg	Signal stdev
Blue	4	4.04	0.98
Green	1	4.93	N.A.
Orange	4	5.37	1.60

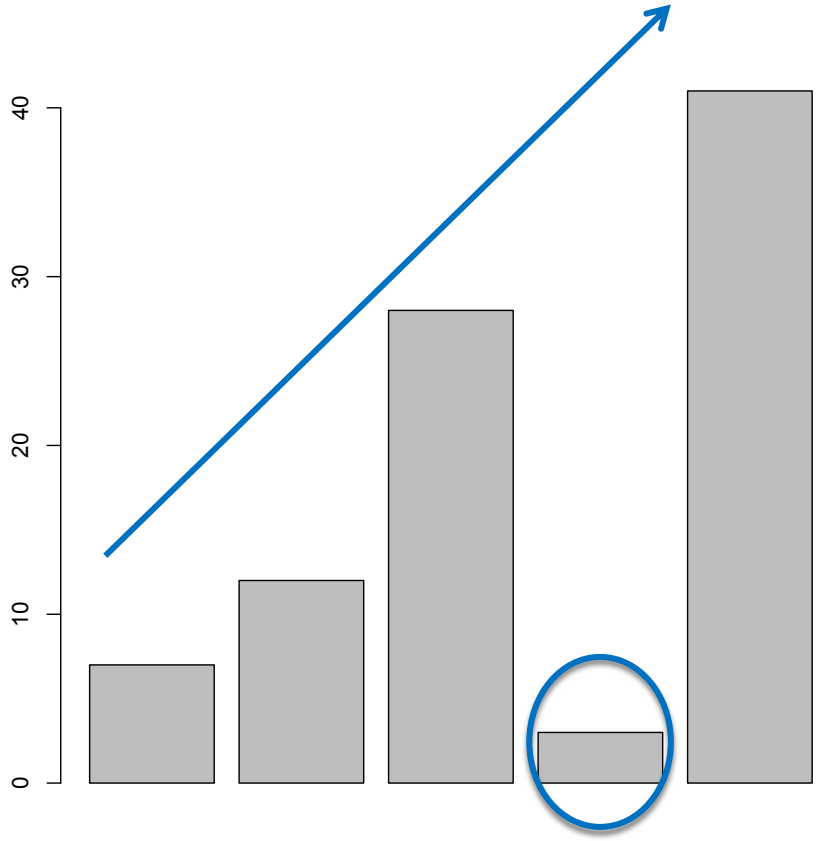
ANALYSIS THROUGH VISUALIZATION

Analysis (broad definition):

- identifying patterns or structure
- adding meaning to these patterns or structure by interpreting them in the context of the system.

Option 1: use analytical methods to achieve this.

Option 2: visualize the data and use the brain's analytic power (perceptual) to reach meaningful conclusions about these patterns.



NUMERICAL SUMMARIES

In a first pass, a variable can be described along 2 dimensions: **centrality** & **spread** (skew and kurtosis are also used).

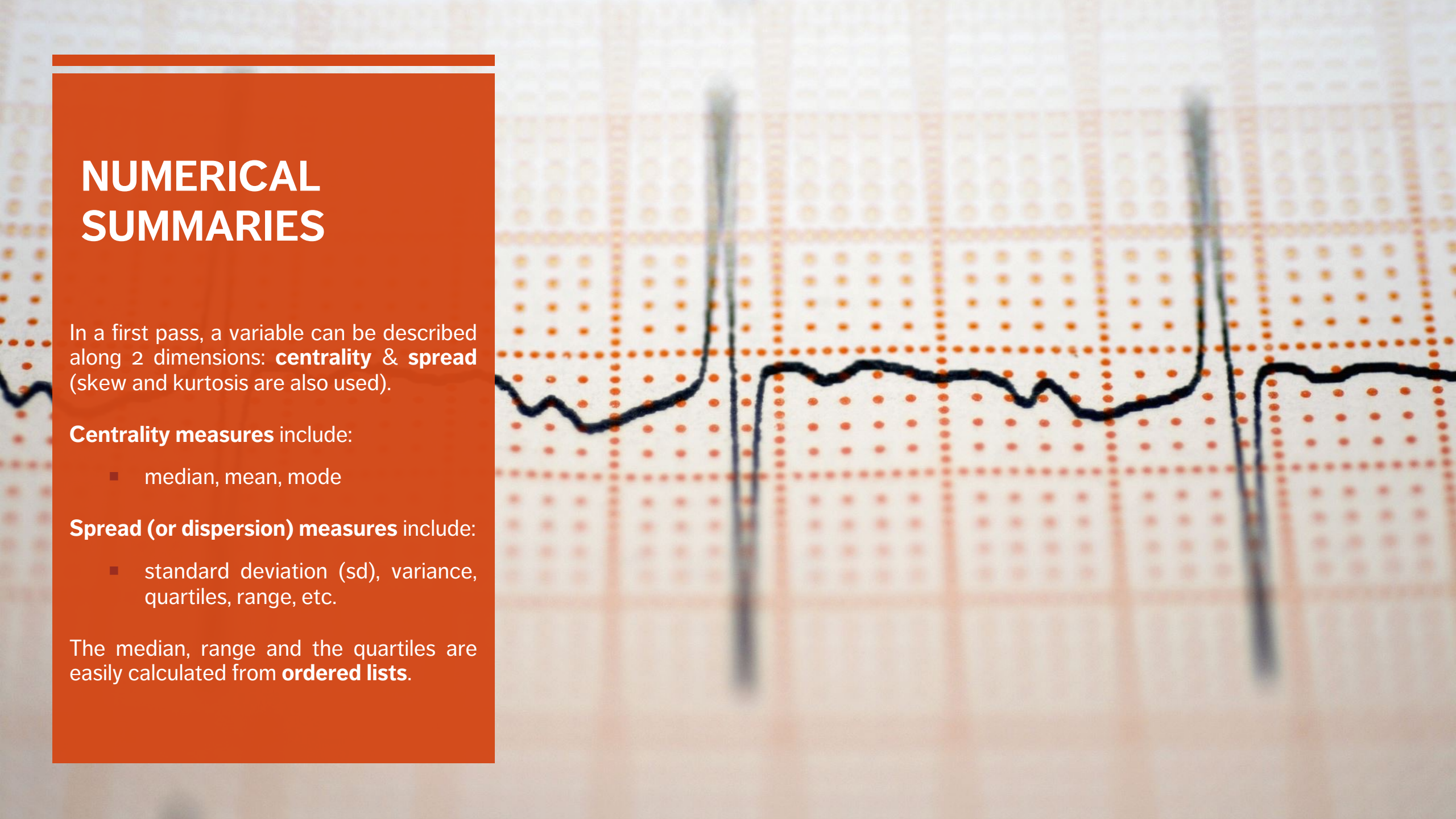
Centrality measures include:

- median, mean, mode

Spread (or dispersion) measures include:

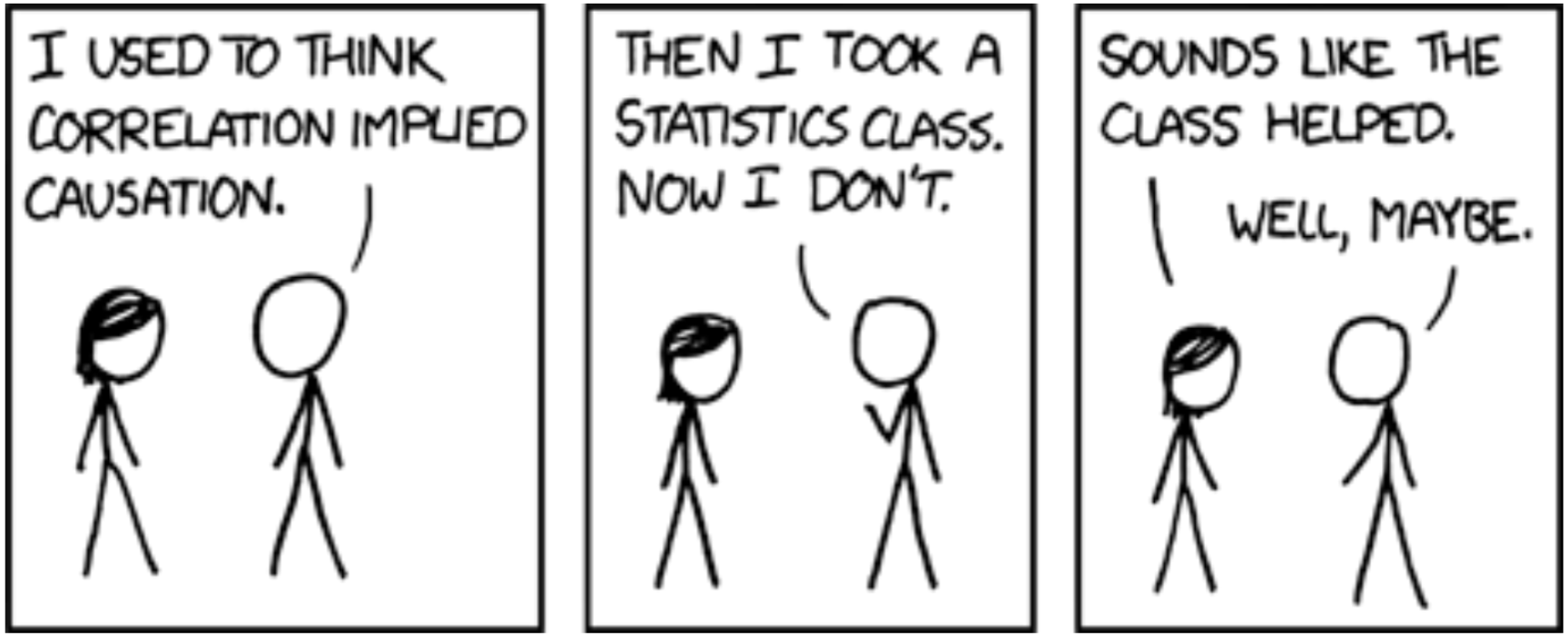
- standard deviation (sd), variance, quartiles, range, etc.

The median, range and the quartiles are easily calculated from **ordered lists**.



[https://xkcd.com/552]

CORRELATION



Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

LINEAR REGRESSION

The basic assumption of **linear regression** is that the dependent variable y can be approximated by a linear combination of the independent variables:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

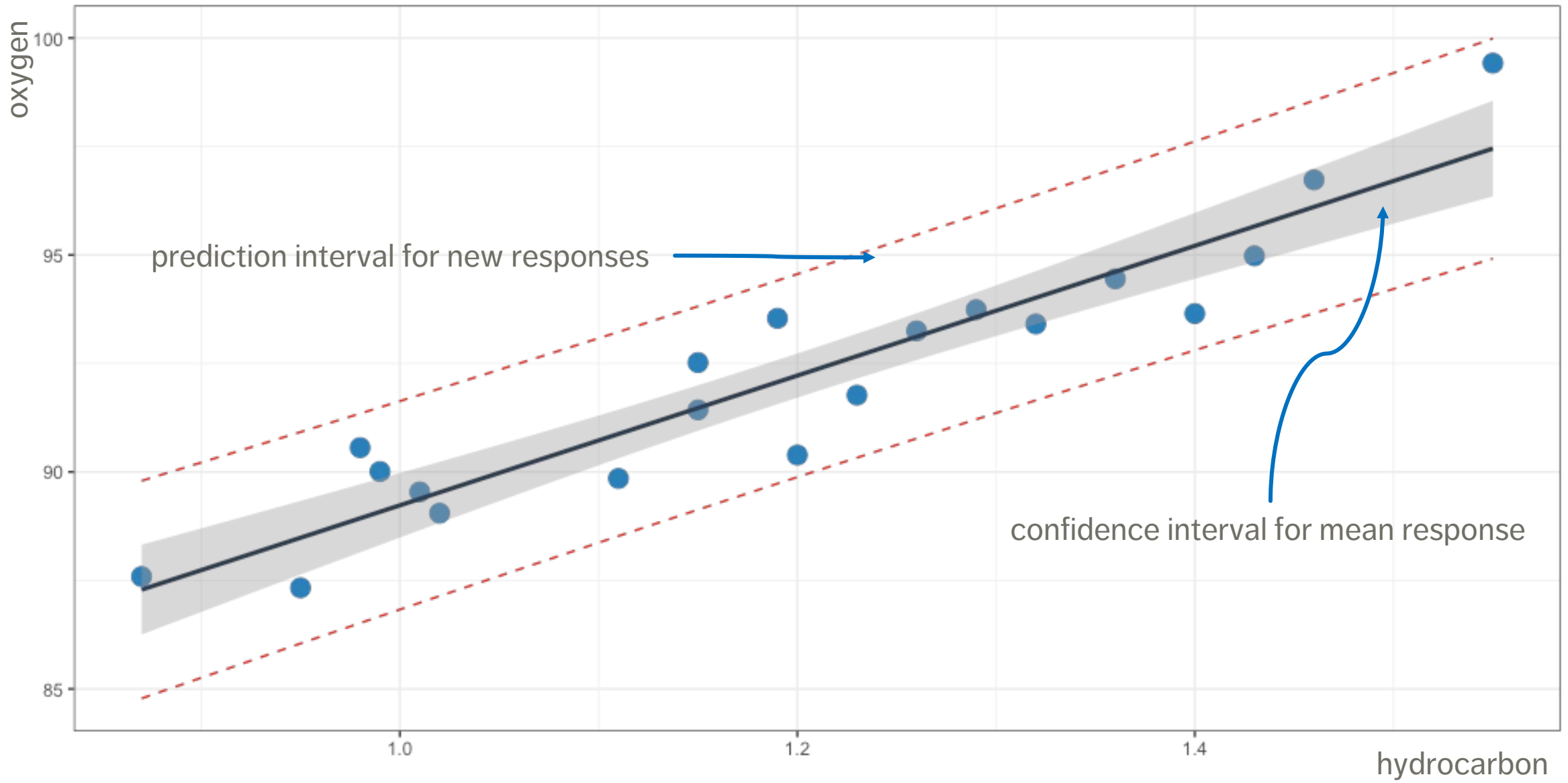
where $\boldsymbol{\beta} \in \mathbb{R}^p$ is to be determined based on the **training set**, and for which

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \qquad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\mathbf{X}) = \sigma^2\mathbf{I}.$$

Typically, the errors are also assumed to be **normally distributed**:

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

$$\text{oxygen} = 14.95 \times \text{hydrocarbon} + 74.28$$



MACHINE LEARNING TASKS

Classification, class probability estimation: which clients are likely to be repeat customers?

Clustering: do customers form natural groups?

Association rule discovery: what books are commonly purchased together?

Others: **value estimation** (how much is a client likely to spend in a restaurant); **profiling and behaviour description**; **link prediction**; **data reduction**; **influence/ causal modeling**; **similarity matching** (which prospective clients are similar to a company's best clients?), etc.

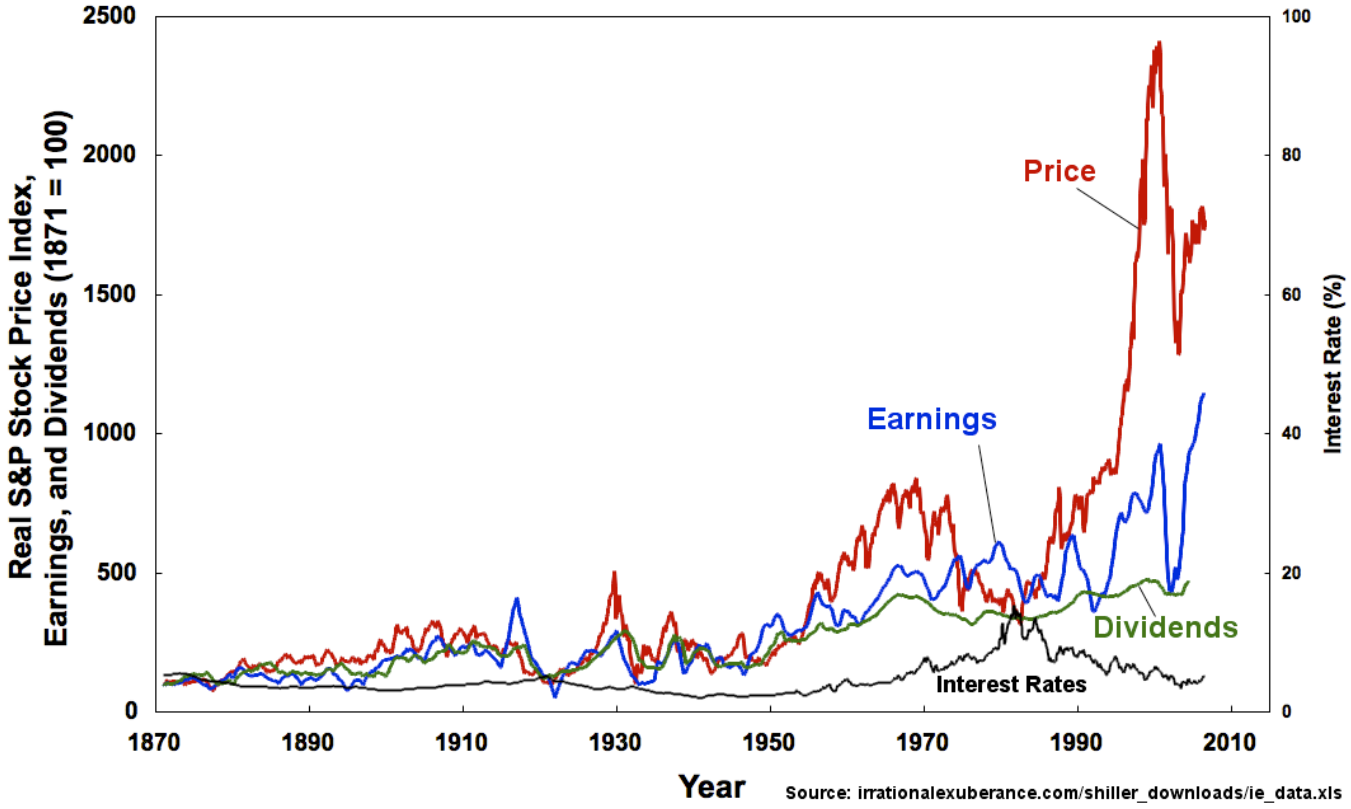
TIME SERIES ANALYSIS

A simple **time series:**

- has two variables: time + 2nd variable
- the second variable is *sequential*

What is the **pattern of behaviour** of this second variable over time? Relative to other variables?

Can we use this to **forecast the future behaviour** of the variable?



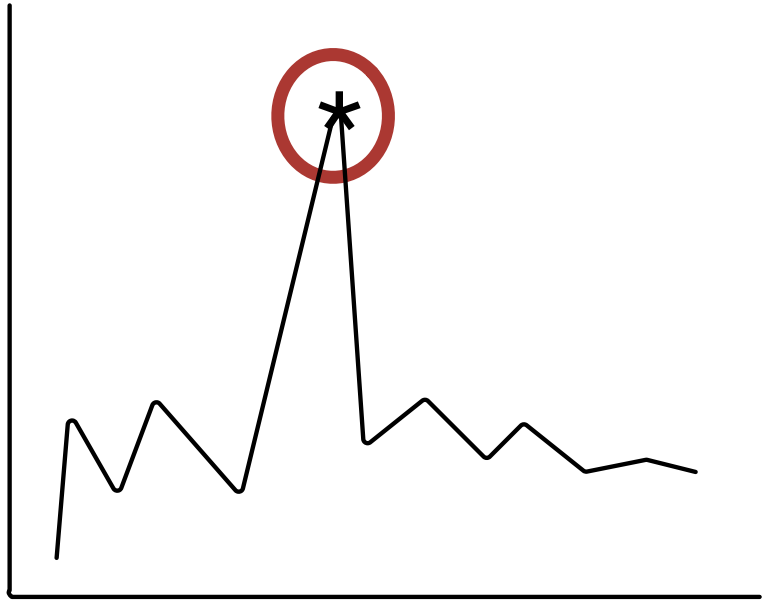
ANOMALY DETECTION

Anomaly: an unexpected, unusual, atypical or statistically unlikely event

Wouldn't it be nice to have a data analysis pipeline that alerted you when things were out of the ordinary?

Many different analytic approaches to take!

- clustering
- classification
- ensemble techniques, etc.



THE HOT MESS

"Data is messy, you know."
"Even after it's been cleaned?"
"*Especially* after it's been cleaned."

Data **cleaning**, **processing**, **wrangling** are essential aspects of data science projects; analysts may spend **up to 80%** of their time on **data preparation**.

DATA WRANGLING AND TIDY DATA

Tidy data has a specific structure:

- each variable is in a single column
- each observation is in a single row
- each type of observational unit is in a single table

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

vs.

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

DATA WRANGLING FUNCTIONALITY

Data wrangling functions should allow the analyst to:

- extract a subset of variables from the data frame
- extract a subset of observations from the data frame
- sort the data frame along any combination of variables in increasing or decreasing order
- to create new variables from existing variables
- to create (so-called) pivot tables, by observation groups
- database functionality (joins, etc.)
- etc.

APPROACHES TO DATA CLEANING

There are 2 **philosophical** approaches to data cleaning and validation:

- methodical
- narrative

The **methodical** approach consists of running through a **check list** of potential issues and flagging those that apply to the data.

The **narrative** approach consists of **exploring** the dataset and trying to spot unlikely and irregular patterns.

APPROACHES TO DATA CLEANING

Methodical (syntax)

- Pros: checklist is **context-independent**; pipelines **easy to implement**; common errors and invalid observations **easily identified**
- Cons: may prove **time-consuming**; cannot identify new types of errors

Narrative (semantics)

- Pros: process may simultaneously yield **data understanding**; false starts are (at most) as costly as switching to mechanical approach
- Cons: may miss important sources of errors and invalid observations for datasets with **high number of features**; domain knowledge may bias the process by neglecting uninteresting areas of the dataset

DATA SOUNDNESS

The ideal dataset will have as few issues as possible with:

- **validity:** data type, range, mandatory response, uniqueness, value, regular expressions
- **completeness:** missing observations
- **accuracy and precision:** related to measurement and data entry errors; target diagrams (accuracy as bias, precision as standard error)
- **consistency:** conflicting observations
- **uniformity:** are units used uniformly throughout?

Checking for data quality issues at an early stage can save headaches later in the analysis.

DATA SOUNDNESS

accurate and precise

precise but not accurate

accurate but not precise

neither accurate nor very precise

COMMON ERROR SOURCES

When dealing with **legacy**, **inherited** or **combined** datasets (i.e., datasets over which there is no collection and initial processing control):

- missing data given a code
- ‘NA’/‘blank’ given a code
- data entry error
- coding error
- measurement error
- duplicate entries
- heaping

[Bill Watterson, *Calvin and Hobbes*, 15 Sep. 1990]

DETECTING INVALID ENTRIES

Potentially invalid entries can be detected with the help of:

- **univariate descriptive statistics**
  count, range, z-score, mean, median, standard deviation, logic check
- **multivariate descriptive statistics**
  n-way table, logic check
- **data visualization**
  scatterplot, scatterplot matrix, histogram, joint histogram, etc.

DETECTING INVALID ENTRIES

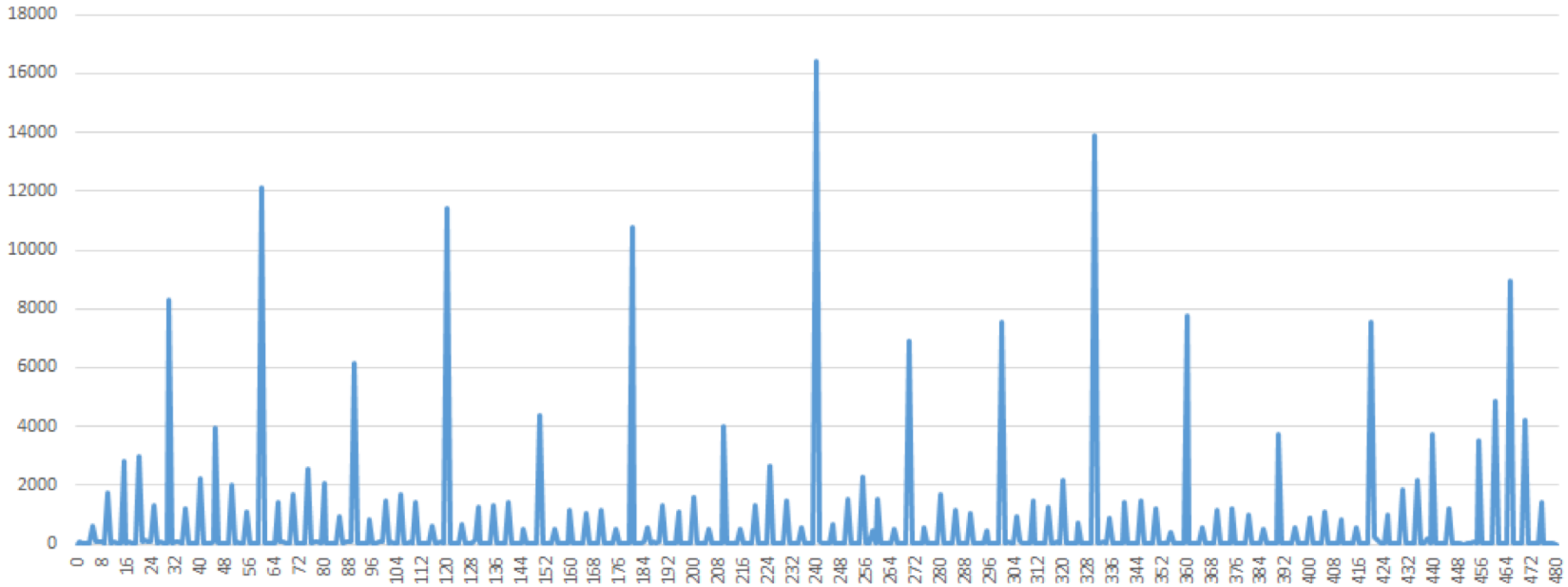
Univariate tests do not always tell the **whole** story.

This step might allow for the identification of potential outliers.

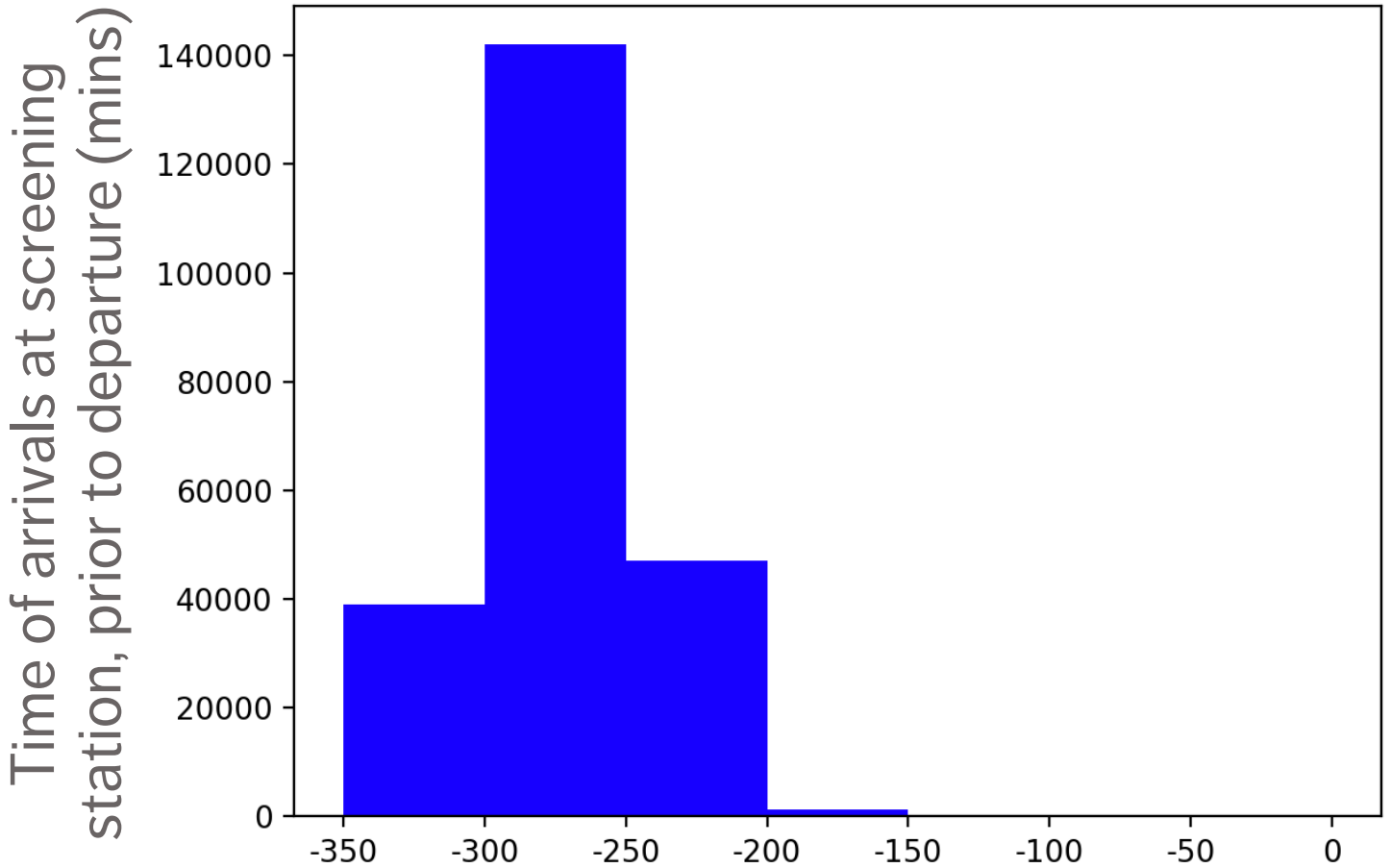
Failure to detect invalid entries ≠ all entries are valid.

Small numbers of invalid entries recoded as “missing.”

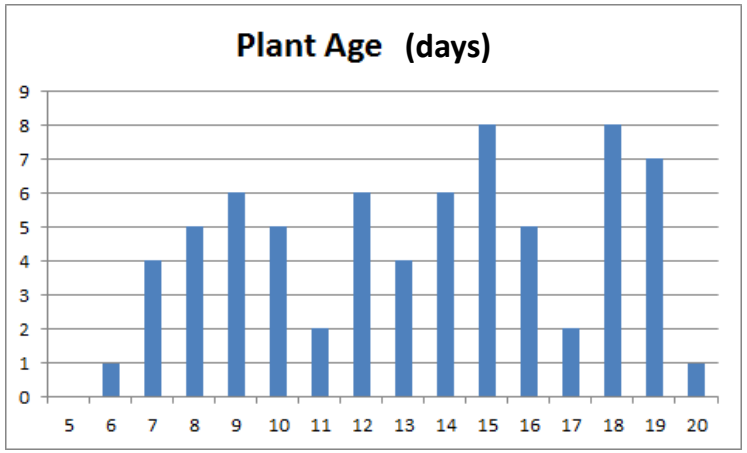
DETECTING INVALID ENTRIES

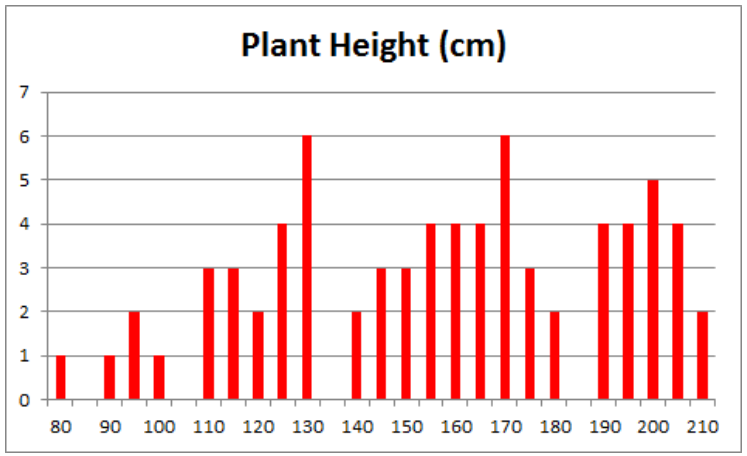


DETECTING INVALID ENTRIES

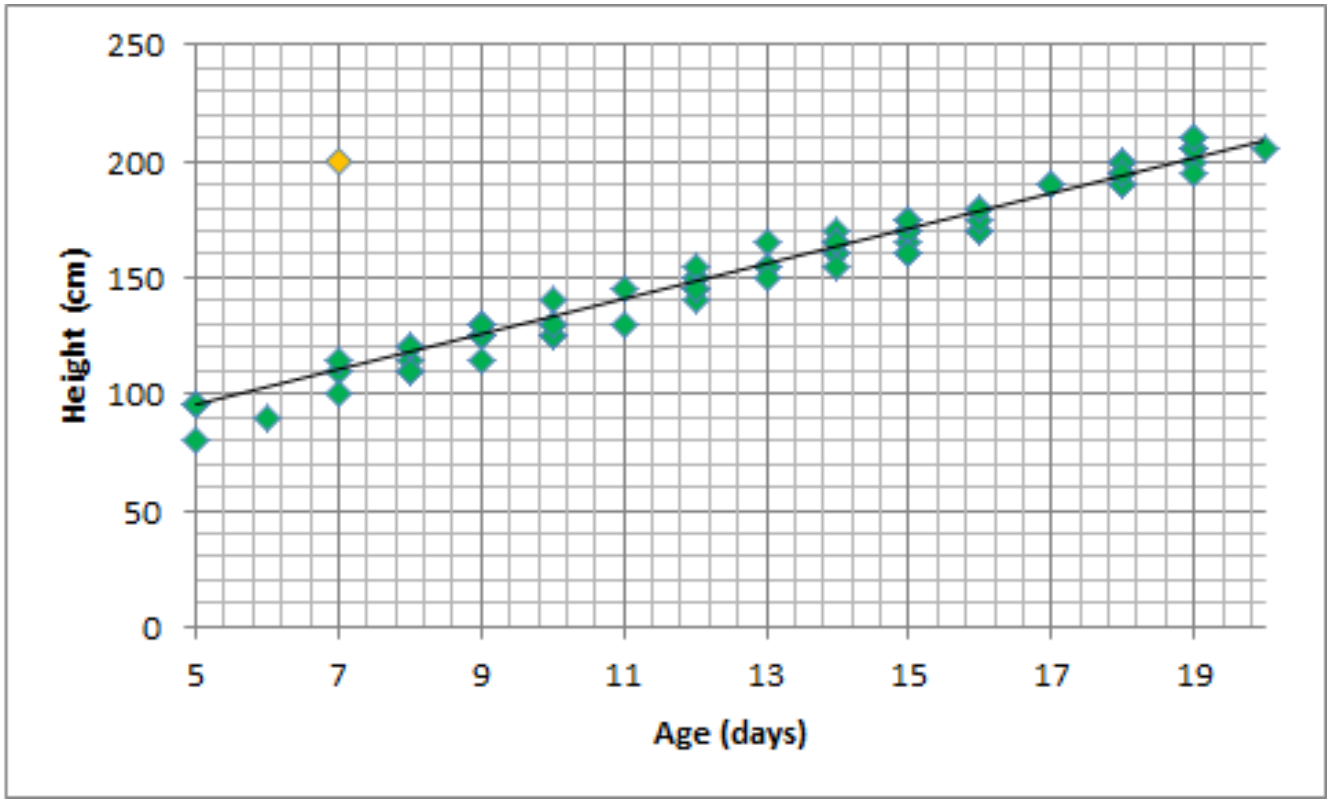


DETECTING INVALID ENTRIES



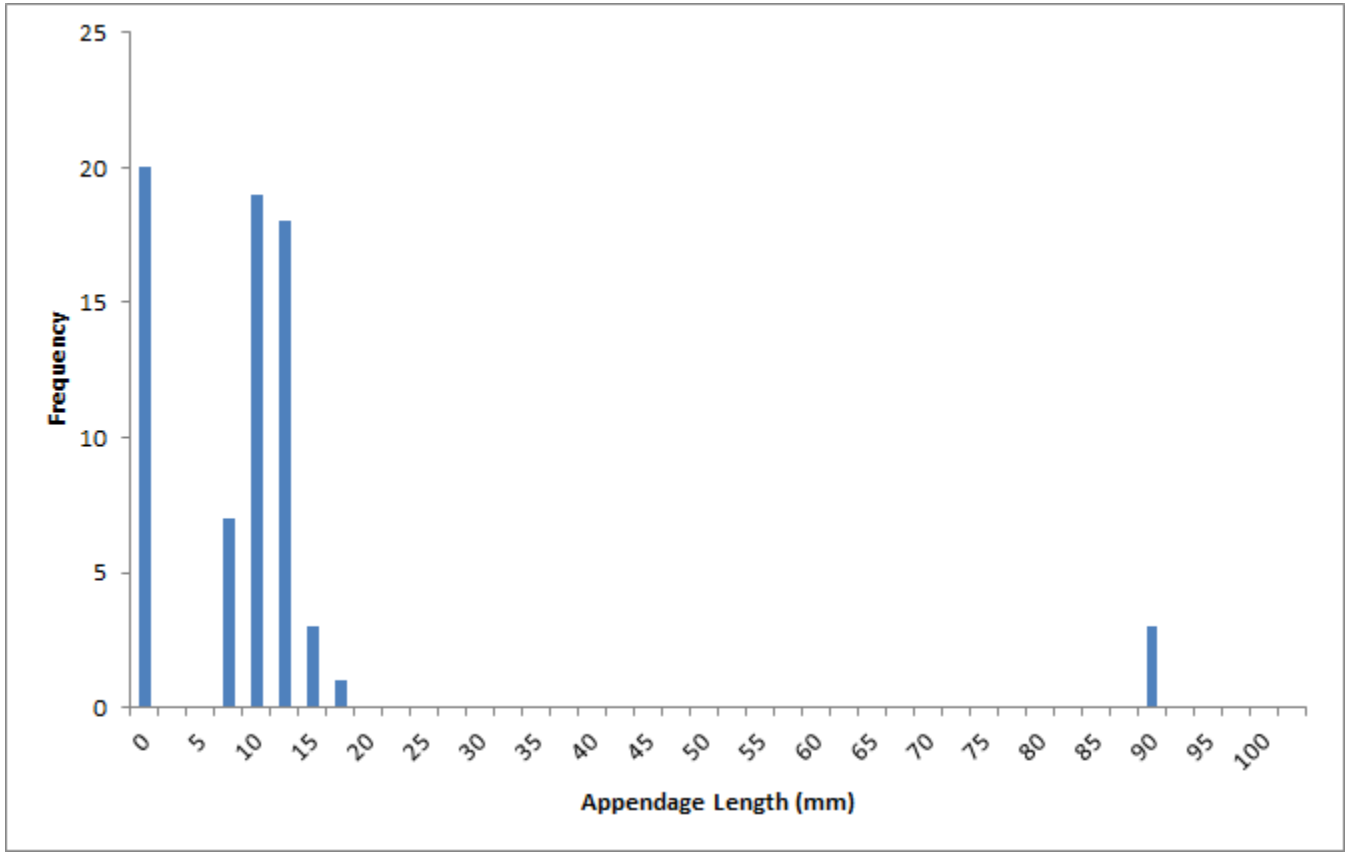


vs.



DETECTING INVALID ENTRIES

Appendage length (mm)	
Mean	10.35
Standard Deviation	16.98
Kurtosis	16.78
Skewness	4.07
Minimum	0
First Quartile	0
Median	8.77
Third Quartile	10.58
Maximum	88
Range	88
Interquartile Range	10.58
Mode	0
Count	71



TYPES OF MISSING OBSERVATIONS

Blank fields come in 4 flavours:

- **nonresponse**
  an observation was expected but none had been entered
- **data entry issue**
  an observation was recorded but was not entered in the dataset
- **invalid entry**
  an observation was recorded but was considered invalid and has been removed
- **expected blank**
  a field has been left blank, but expectedly so

TYPES OF MISSING OBSERVATIONS

Too many missing values (of the first three type) can be indicative of **issues with the data collection process** (more on this later).

Too many missing values (of the fourth type) can be indicative of **poor questionnaire design**.

Finding missing values can help you deal with other data science problems.

THE CASE FOR IMPUTATION

Not all analytical methods can easily accommodate missing observations:

- **discard** the missing observation
  - not recommended, unless the data is MCAR in the dataset as a whole
  - acceptable in certain situations (e.g., small number of missing values in a large dataset)
- come up with a **replacement (imputation) value**
  - main drawback: we never know what the true value would have been
  - often the best available option

MISSING VALUES MECHANISM

Missing Completely at Random (MCAR)

- item absence is independent of its value or of auxiliary variables
- **example:** an electrical surge randomly deletes an observation in the dataset

Missing at Random (MAR)

- item absence is not completely random; can be accounted by auxiliary variables with complete info
- **example:** if women are less likely to tell you their age than men for societal reasons, but not because of the age values themselves)

MISSING VALUES MECHANISM

Not Missing at Random (NMAR)

- reason for nonresponse is related to item value (also called **non-ignorable non-response**)
- **example:** if illicit drug users are less likely to admit to drug use than teetotallers

In general, the missing mechanism **cannot be determined** with any certainty; we may need to make assumptions (domain expertise can help).

IMPUTATION METHODS

- list-wise deletion
- mean or most frequent imputation
- regression or correlation imputation
- stochastic regression imputation
- last observation carried forward
- next observation carried backward
- k-nearest neighbours imputation
- multiple imputation
- etc.

IMPUTATION METHODS

List-wise deletion: remove units with at least one missing values

- **assumption:** MCAR
- **cons:** can introduce bias (if not MCAR), reduction in sample size, increase in standard error

Mean/most frequent imputation: substitute missing values by average/most frequent value

- **assumption:** MCAR
- **cons:** distortions of distribution (spike at mean) and relationships among variables

IMPUTATION METHODS

Regression/correlation imputation: substitute missing values using fitted values based on other variables with complete information

- **assumption:** MAR
- **cons:** artificial reduction in variability, over-estimation of correlation

Stochastic regression imputation: regression/correlation imputation with a random error term added

- **assumption:** MAR
- **cons:** increased risk of type I error (false positives) due to small std error

IMPUTATION METHODS

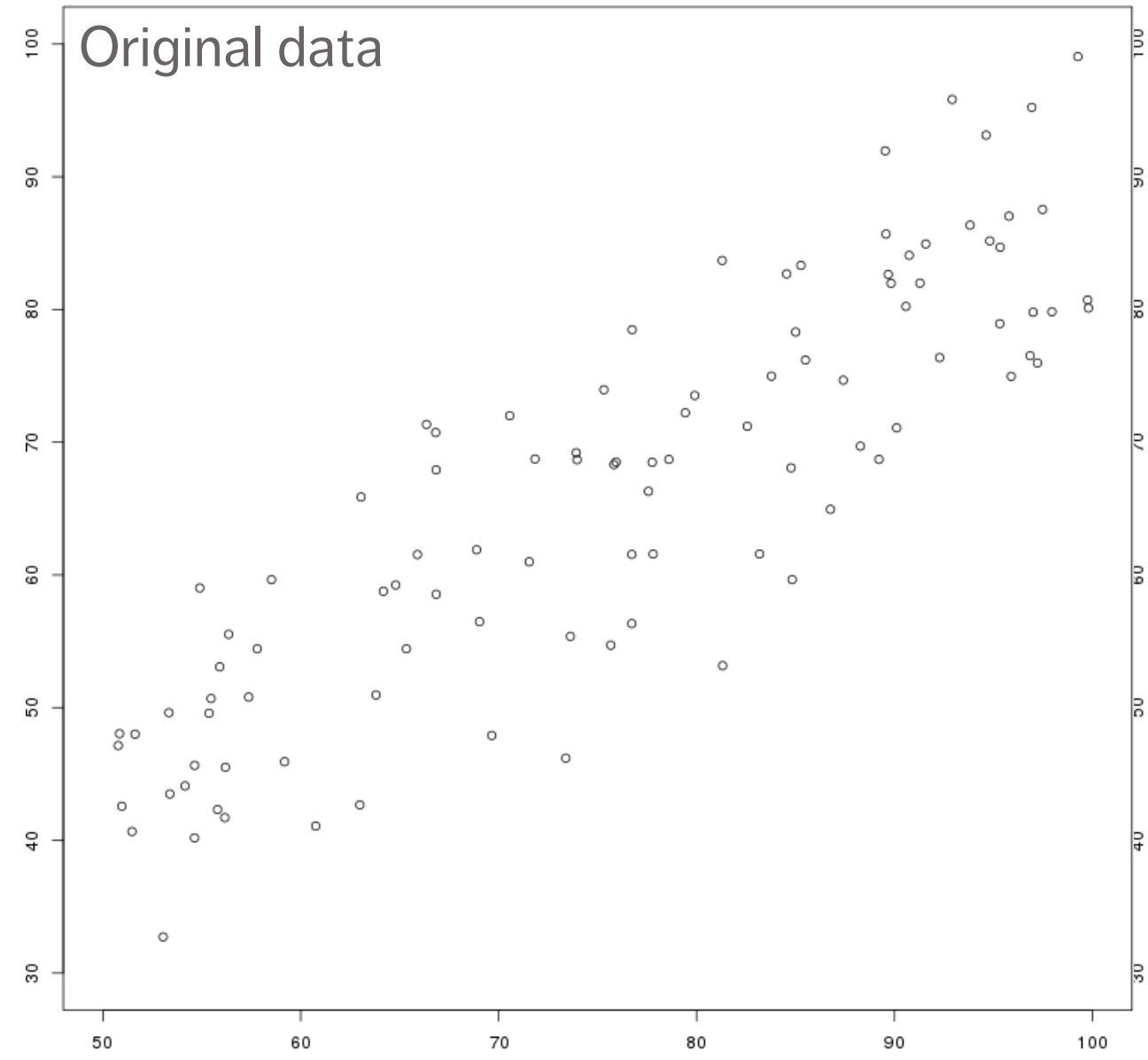
Last observation carried forward: substitute the missing values with latest previous values (in a longitudinal study)

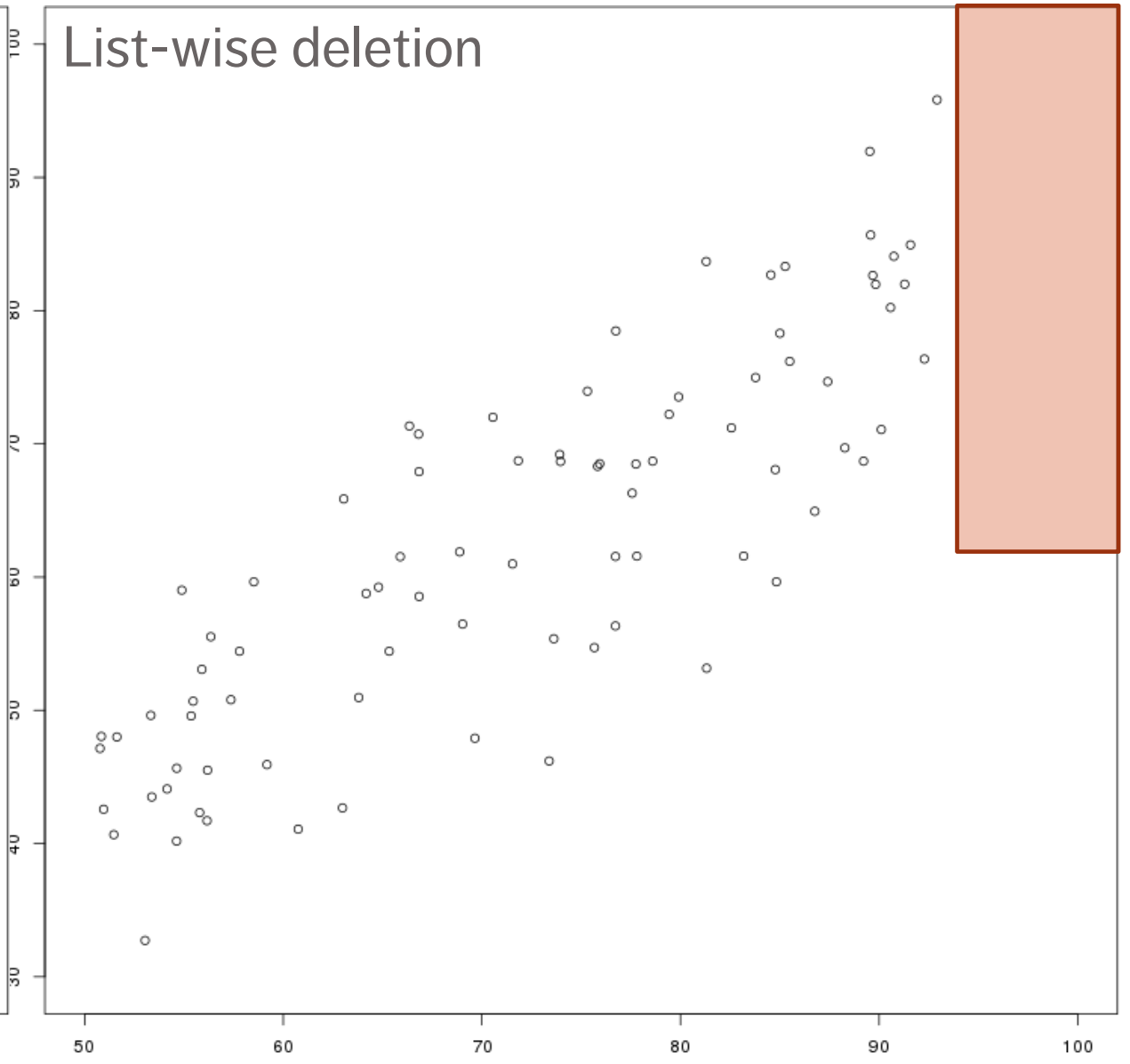
- **assumption:** MCAR, values do not vary greatly over time
- **cons:** may be too "generous", depending on the nature of study

k nearest neighbour imputation (kNN): substitute the missing entry with the average from the group of the k most similar complete cases

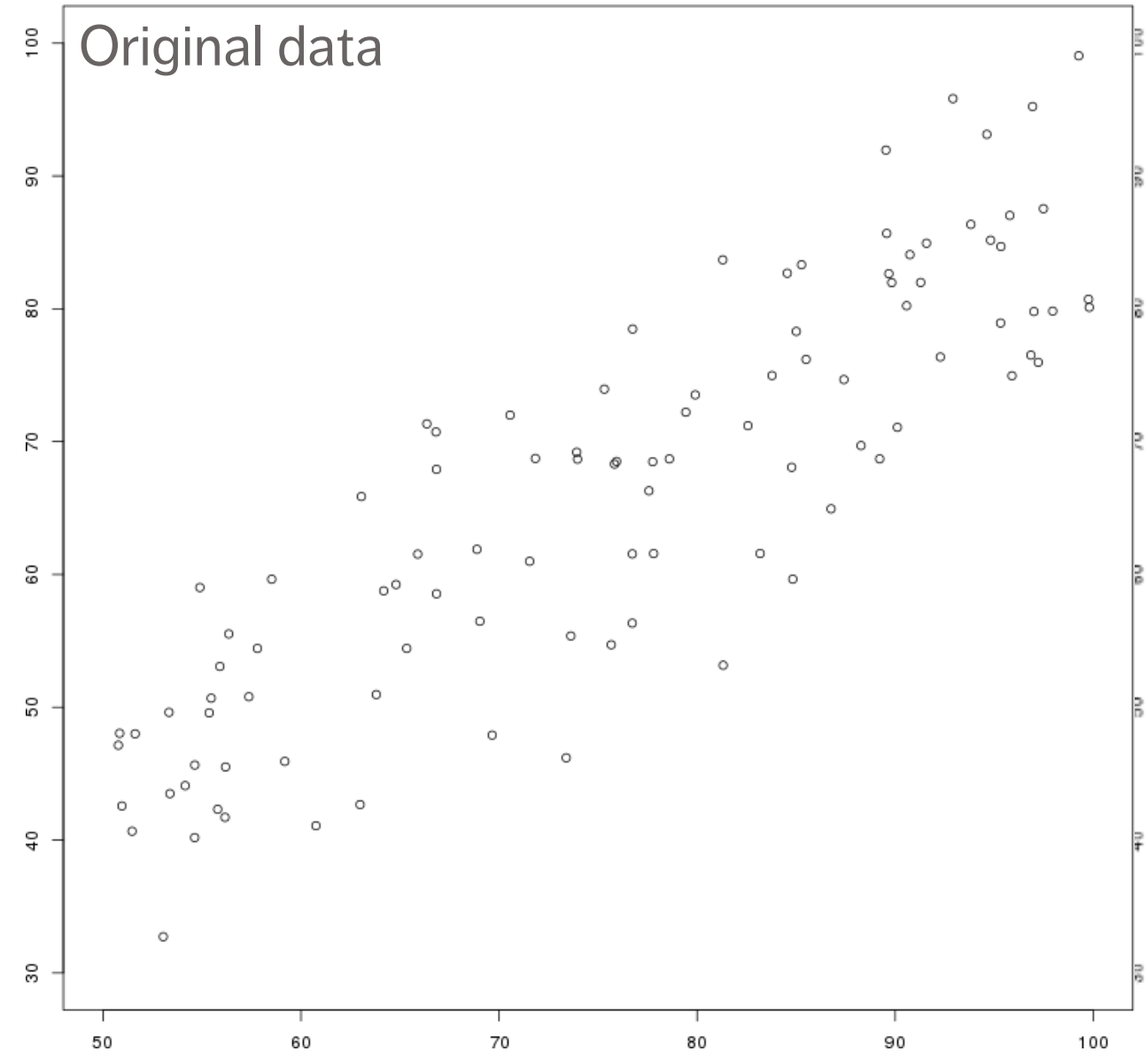
- **assumption:** MAR
- **cons:** difficult to choose appropriate value for k; possible distortion in data structure

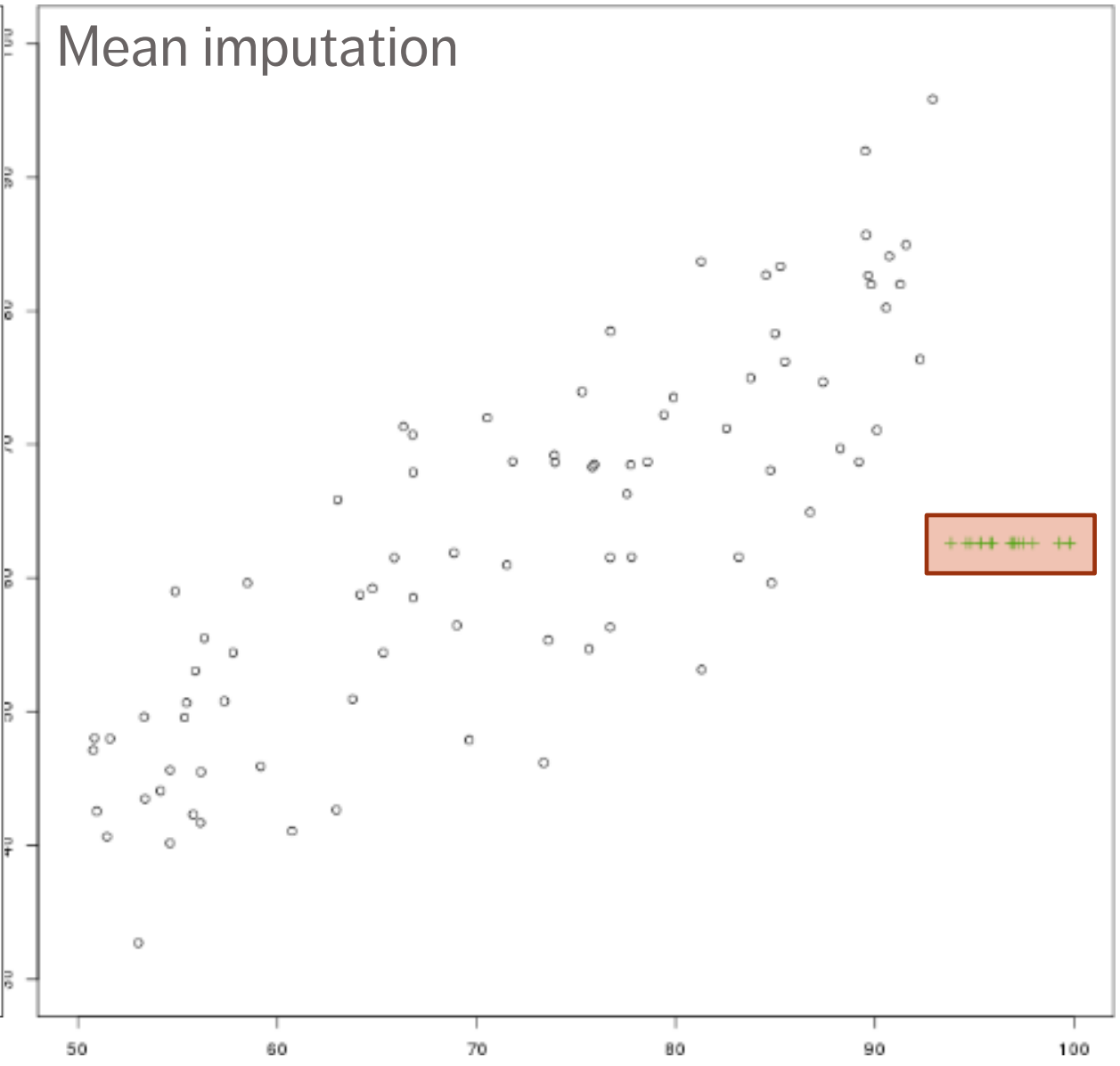
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



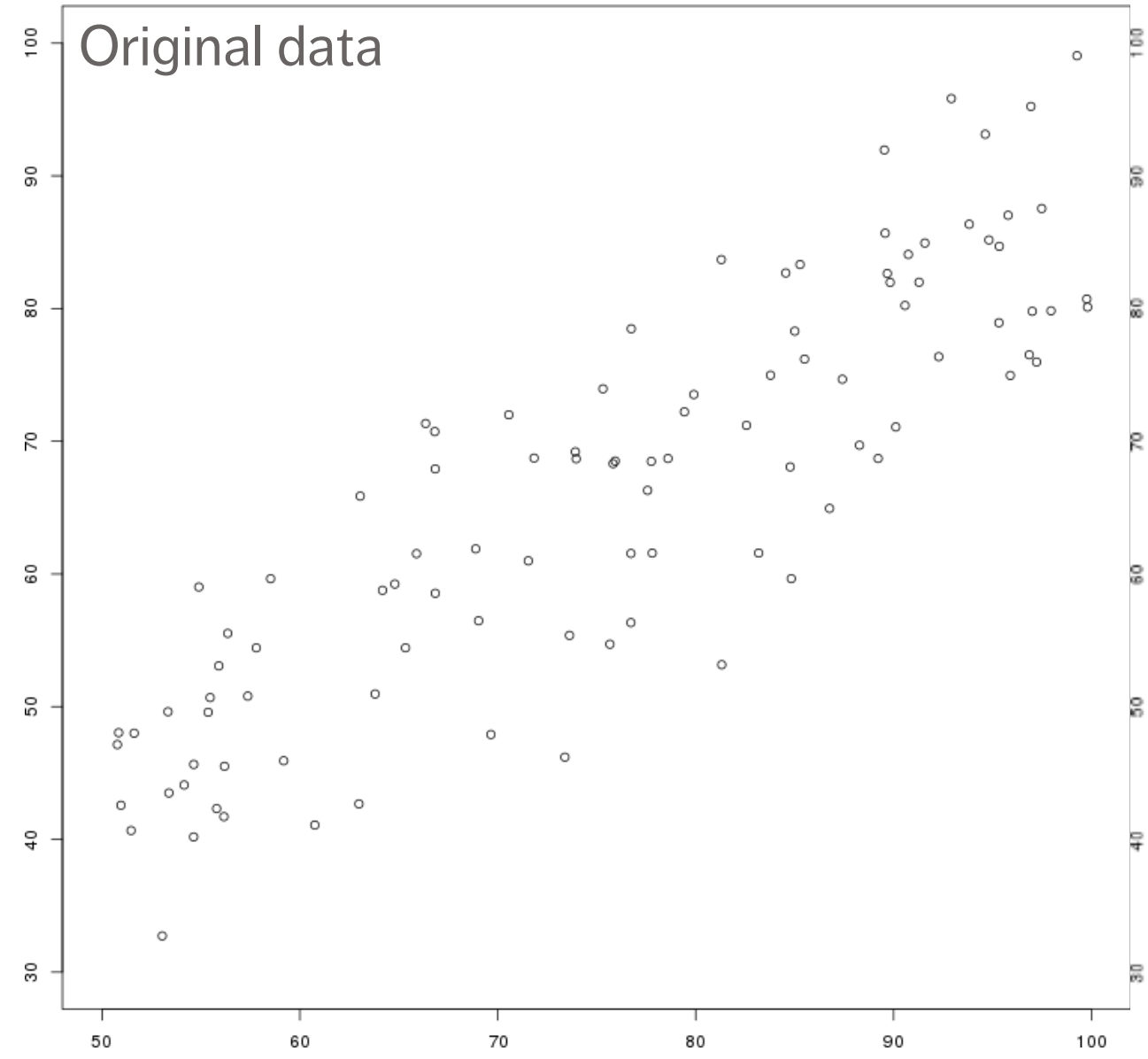


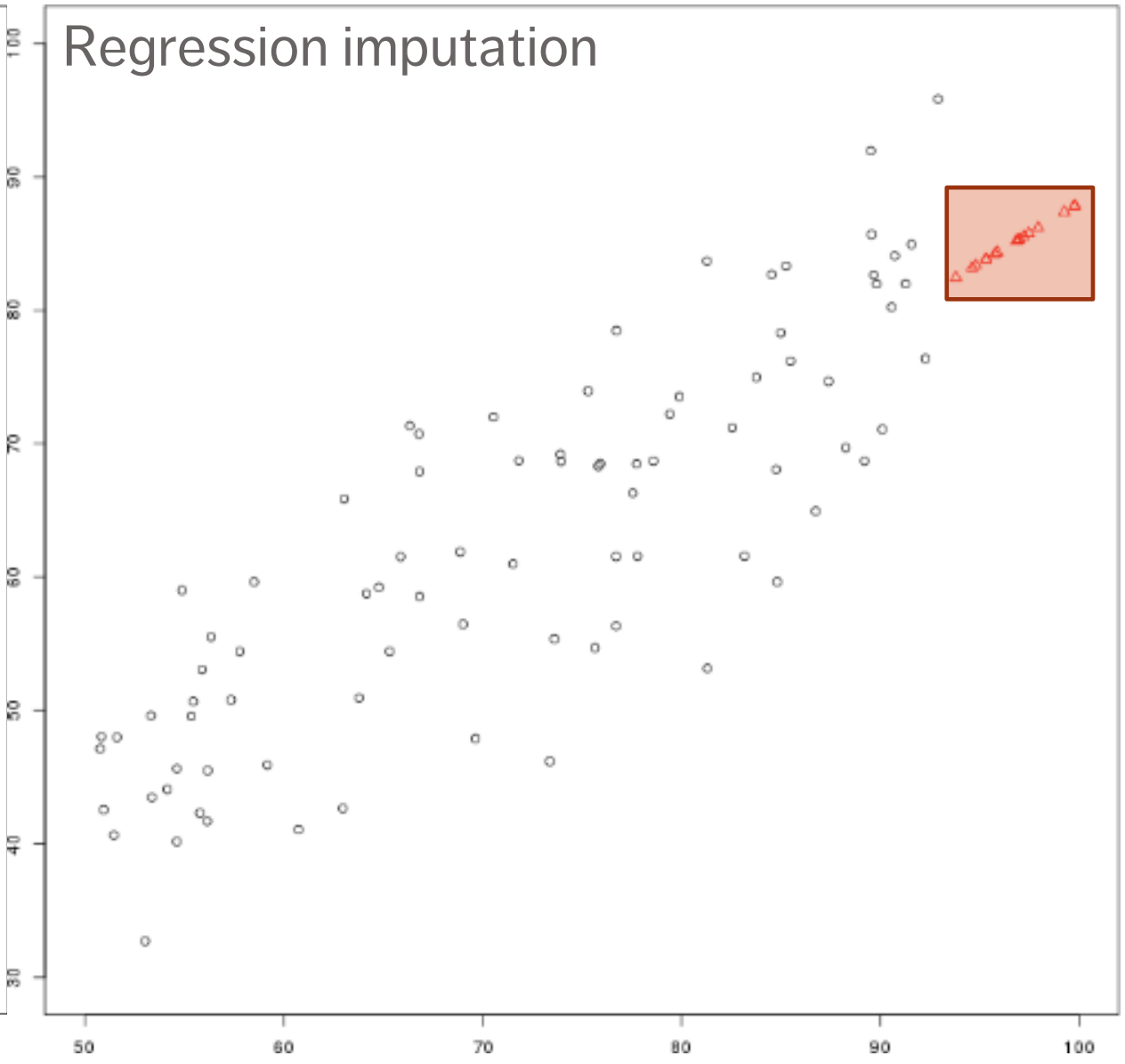
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



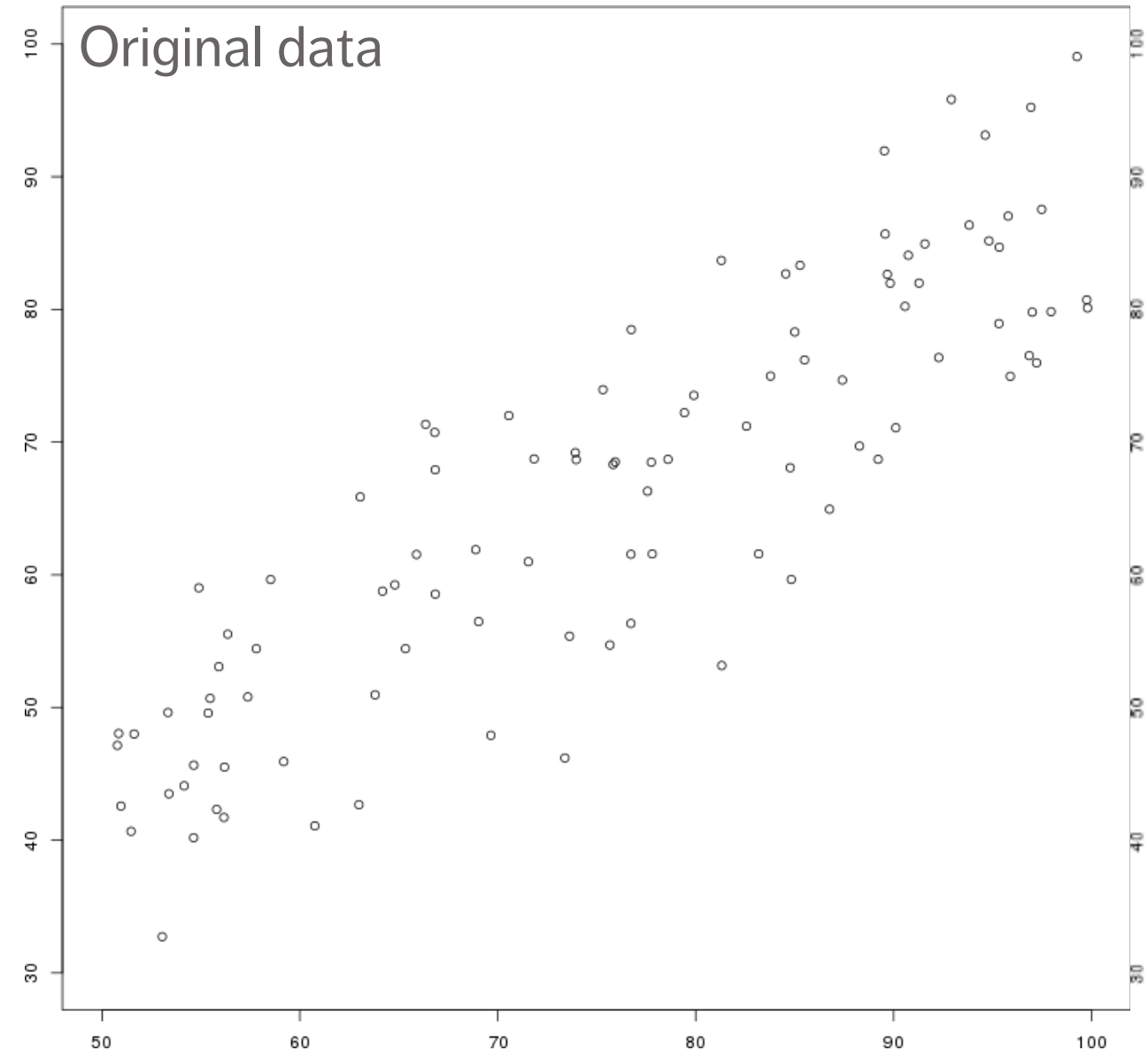


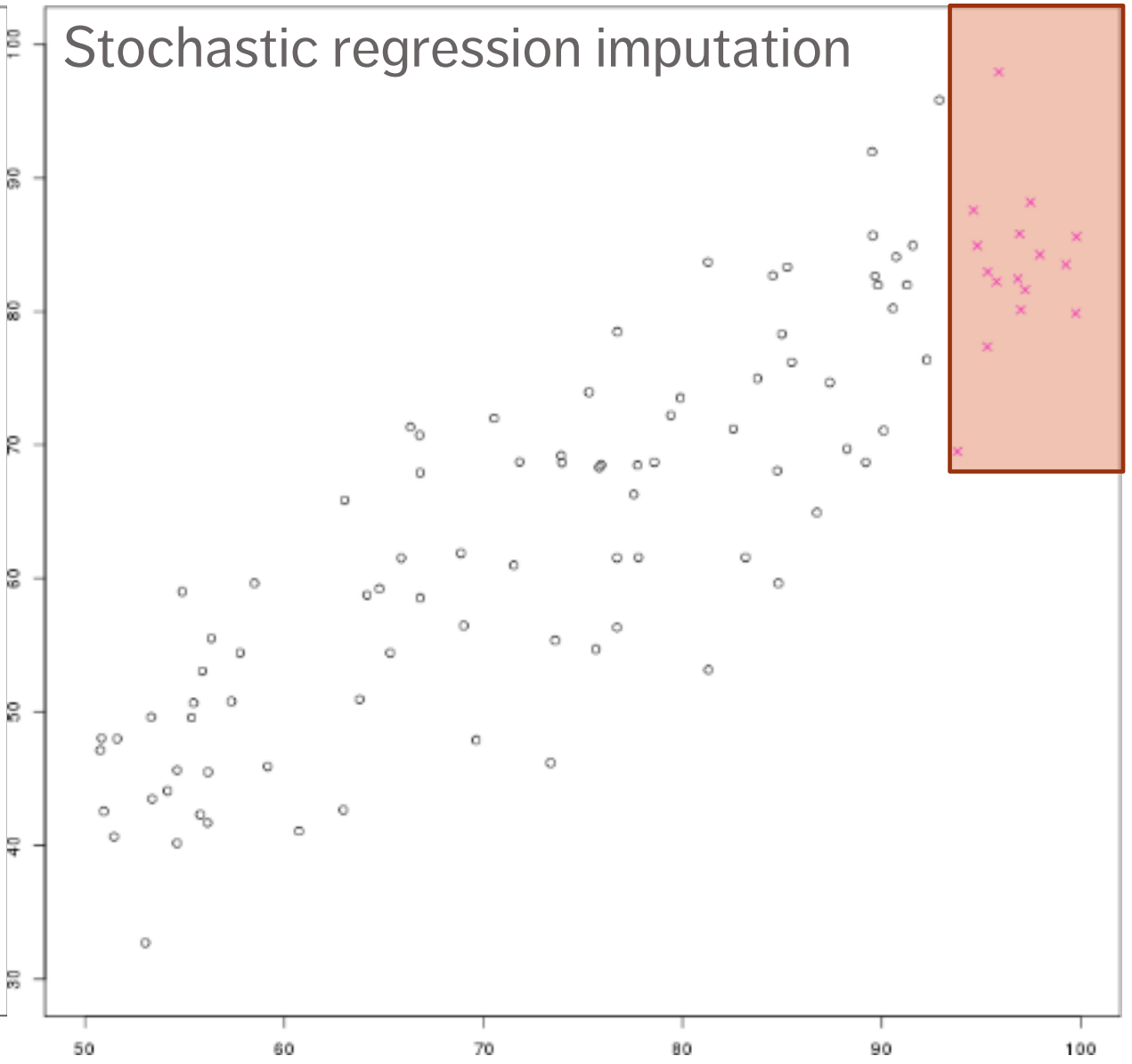
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





MULTIPLE IMPUTATION

Imputations increase the noise in the data.

In **multiple imputation**, the effect of that noise can be measured by consolidating the analysis outcome from multiple imputed datasets

Steps:

1. repeated imputation creates m versions of the dataset
2. each of these datasets is analyzed, yielding m outcomes
3. the m outcomes are pooled into a single result for which the mean, variance, and confidence intervals are known

MULTIPLE IMPUTATION

Advantages

- **flexible**; can be used in a various situations (MCAR, MAR, even NMAR in certain cases)
- accounts for **uncertainty** in imputed values
- fairly easy to implement

Disadvantages

- m may need to be fairly **large** when there are many missing values in numerous features, which slows down the analyses
- if the analysis output is not a single value but some complicated mathematical object, this approach is unlikely to be useful

TAKE-AWAYS

Missing values **cannot simply be ignored**.

The missing mechanism **cannot typically be determined** with any certainty.

Imputation methods work best when values are **MCAR** or **MAR**, but imputation methods tend to produce biased estimates.

In single imputation, imputed data is treated as the actual data; multiple imputation can help reduce the noise.

Is stochastic imputation best? In our example, yes – but … **No-Free Lunch theorem**!

ANOMALOUS OBSERVATIONS

In practice, an **anomalous observation** may arise as

- a **"bad" object/measurement:** data artifacts, spelling mistakes, poorly imputed values, etc.
- a **misclassified observation:** according to the existing data patterns, the observation should have been labeled differently;
- an observation whose measurements are found in the **distribution tails** of a large enough number of features;
- an **unknown unknown**: a completely new type of observations whose existence was heretofore unsuspected.

ANOMALOUS OBSERVATIONS

Observations could be anomalous in one context, but not in another:

- a 6-foot tall adult male is in the 86th percentile for **Canadian males** (tall, but not unusual);
- in **Bolivia**, the same man would be in the 99.9th percentile (very tall and unusual).

Anomaly detection points towards interesting questions for analysts and SMEs: in this case, **why is there such a large discrepancy** in the two populations?

OUTLIERS

Outlying observations are data points which are **atypical** in comparison to

- the unit's remaining features (*within-unit*),
- the field measurements for other units (*between-units*)

Outliers are observations which are **dissimilar to other cases** or which **contradict known dependencies** or rules.

Careful study is needed to determine whether outliers should be retained or removed from the dataset.

DETECTING ANOMALIES

Outliers may be anomalous along any of the unit's variables, or in combination.

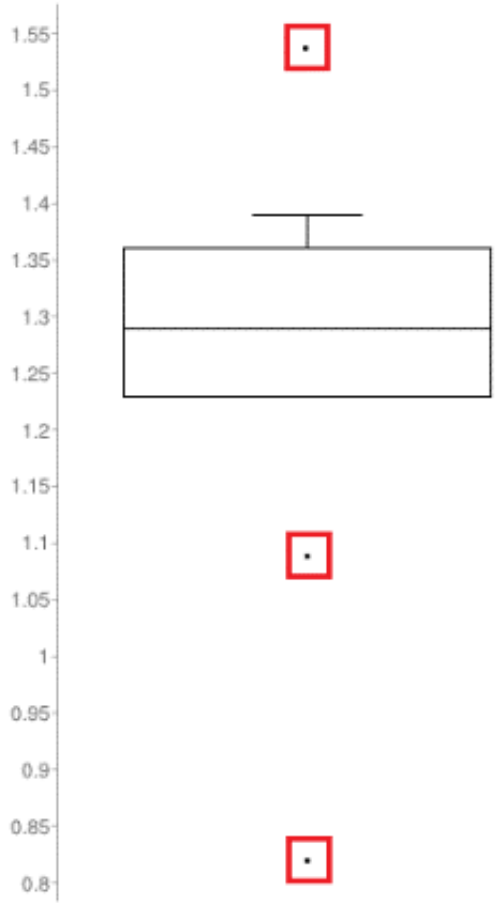
Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small sample sizes.

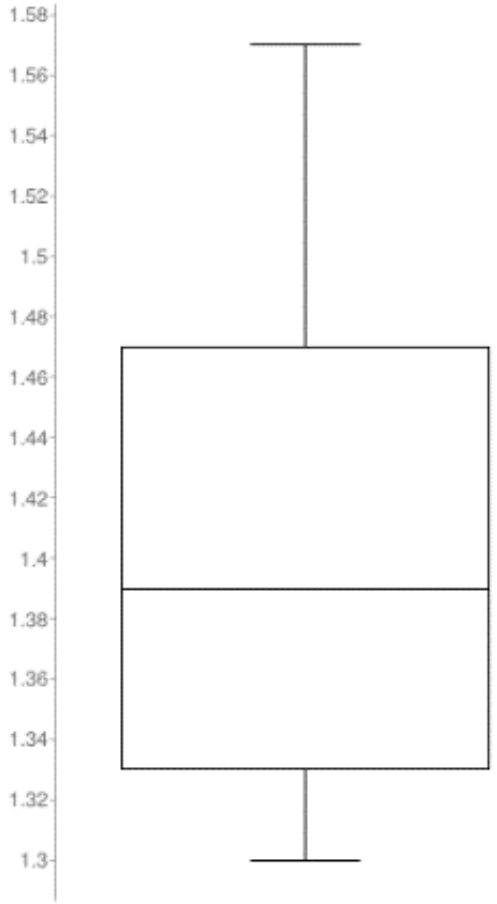
Differentiating anomalies from noise or data entry errors is **hard**.

Boundaries between normal and deviating units may be **fuzzy**.

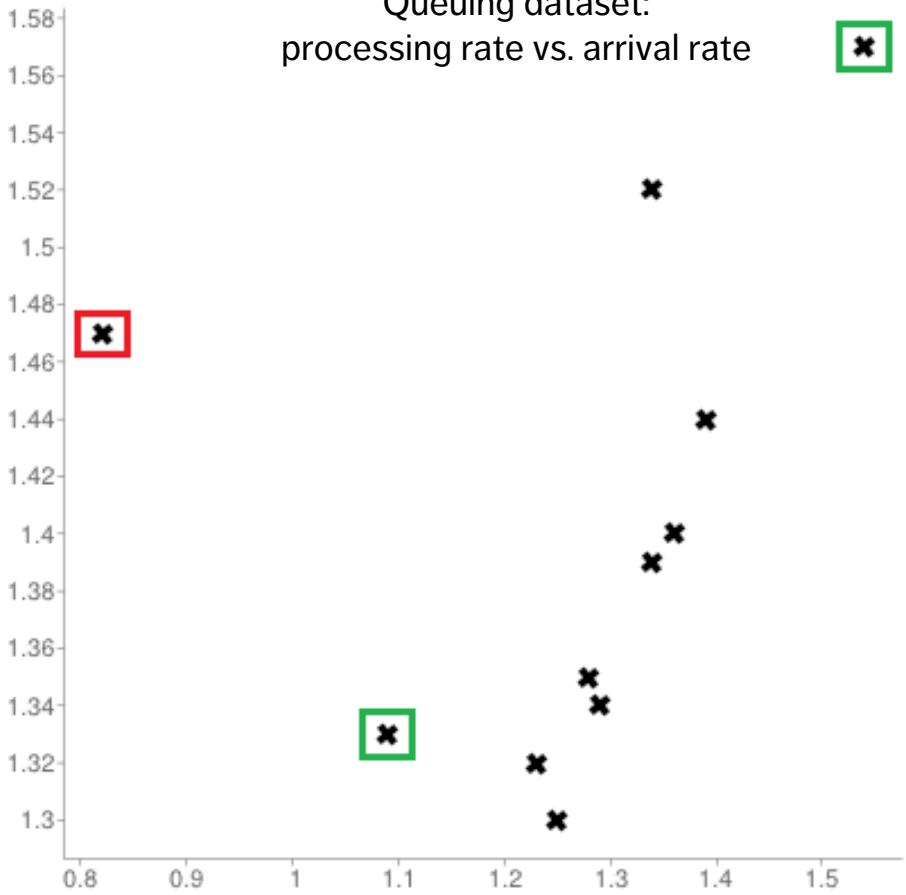
Anomalies associated with malicious activities are typically **disguised**.

VISUAL OUTLIER DETECTION





Queuing dataset:
processing rate vs. arrival rate



DETECTING ANOMALIES

Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used.

Graphical methods are easy to implement and interpret:

- **outlying observations**
  box-plots, scatterplots, scatterplot matrices, 2D tour, Cooke's distance, normal qq plots
- **influential data**
  some level of analysis must be performed (leverage)

Careful: once anomalous observations have been removed from the dataset, previously “regular" units may become anomalous.

ANOMALY DETECTION ALGORITHMS

Supervised methods use a historical record of labeled anomalous observations:

- domain expertise is required to tag the data
- classification or regression task
- rare occurrence problem

		Predicted Class	
		Normal	Anomaly
Actual Class	Normal	*TN*	*FP*
	Anomaly	*FN*	*TP*

Unsupervised methods don't use external information:

- traditional methods and tests
- can also be seen as a clustering or association rules problem

ANOMALY DETECTION ALGORITHMS

The mis-classification cost is often assumed to be symmetrical, which can lead to **technically correct but useless** outputs.

For instance, most (99.999+%) air passengers do not bring weapons with them on flights; a model that predicts that no passenger is smuggling a weapon would be 99.999+% accurate, but it would miss the point completely.

For the **security agency**, the cost of wrongly thinking that a passenger is:

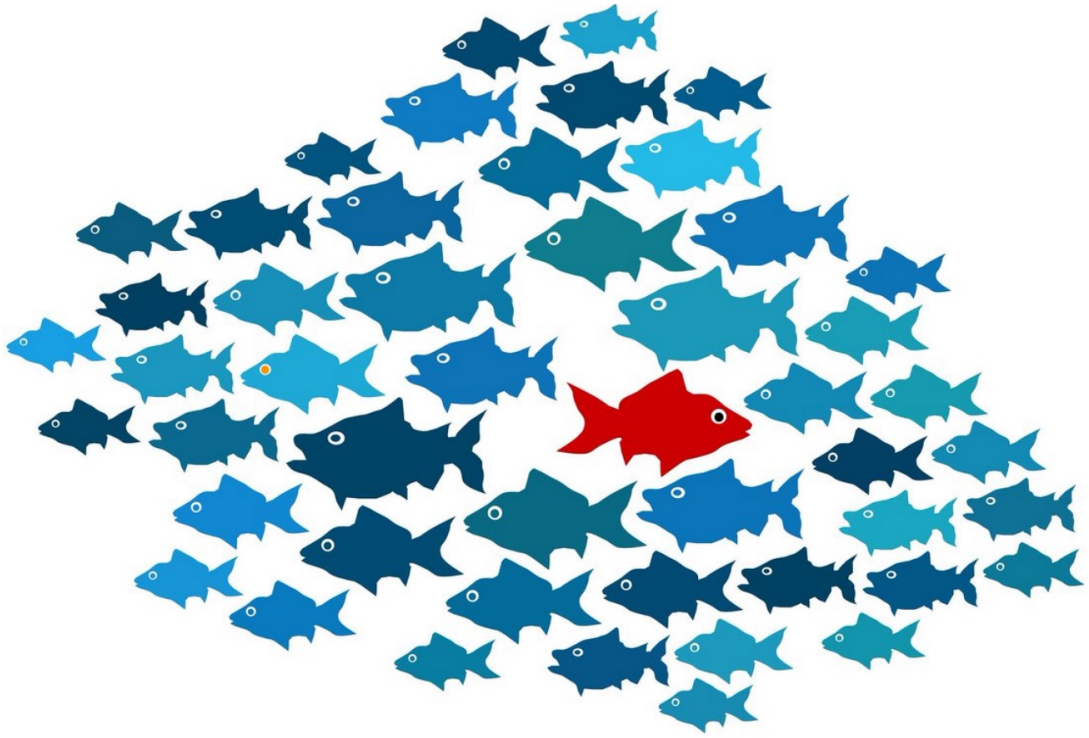
- smuggling a weapon $\Longrightarrow$ cost of a single search
- NOT smuggling a weapon $\Longrightarrow$ catastrophe (potentially)

But **wrongly targeted individuals** may have a different take on this!

ANOMALY DETECTION ALGORITHMS

If all participants in a workshop except for one can view the video conference lectures, then the one individual/internet connection/computer is **anomalous** – it behaves in a manner which is different from the others.

But this **DOES NOT MEAN** that the different behaviour is necessarily the one we are interested in...



SIMPLE OUTLIER TESTS

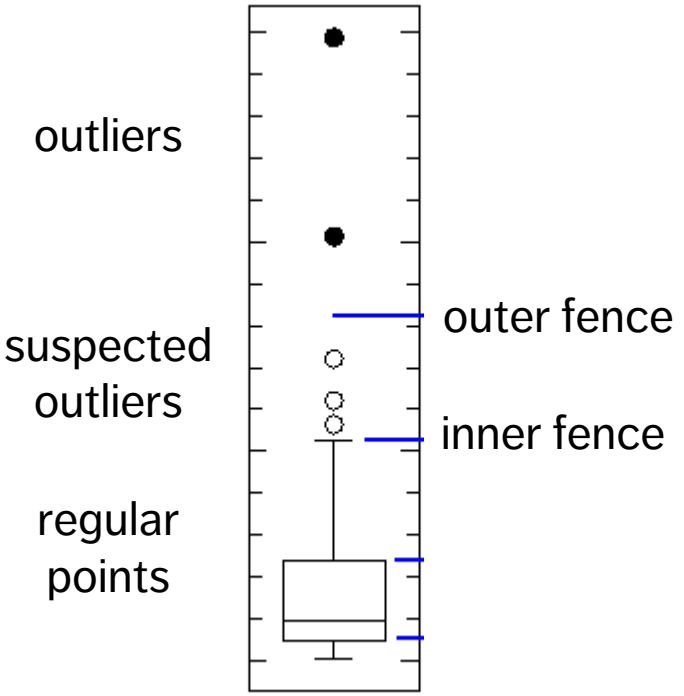
Tukey's Boxplot test: for normally distributed data, regular observations typically lie between the **inner fences**

$$Q_1 - 1.5\times(Q_3 - Q_1) \text{ and } Q_3 + 1.5\times(Q_3 - Q_1).$$

Suspected outliers lie between the **inner fences** and the **outer fences**

$$Q_1 - 3\times(Q_3 - Q_1) \text{ and } Q_3 + 3\times(Q_3 - Q_1).$$

Outliers lie beyond the **outer fences**.



SIMPLE OUTLIER TESTS

The **Dixon Q Test** is used in experimental sciences to find outliers in (extremely) small datasets (dubious validity).

The **Mahalanobis Distance** (linked to the leverage) can be used to find multi-dimensional outliers (when relationships are linear).

Other simple tests:

- **Grubbs** (univariate)
- **Tietjen-Moore** (for a specific # of outliers)
- **generalized extreme studentized deviate** (for unknown # of outliers)
- **chi-square** (outliers affecting goodness-of-fit)

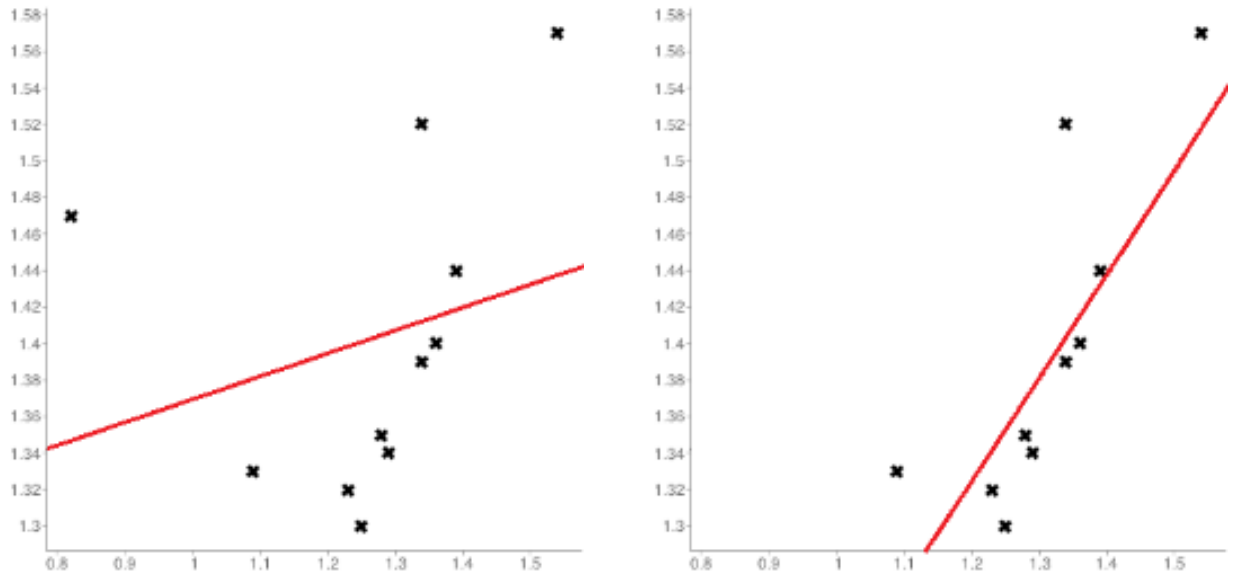
INFLUENTIAL OBSERVATIONS

Influential data points are observations whose absence leads to **markedly different** analysis results.

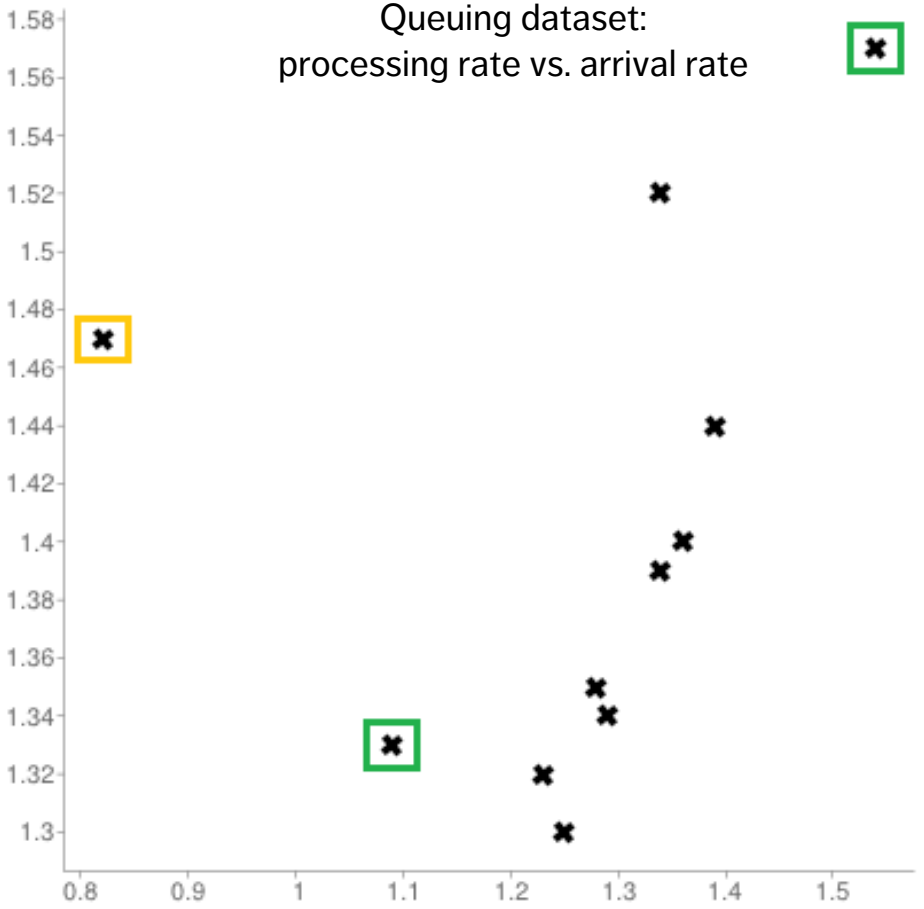
When influential observations are identified, **remedial measures** (such as data transformations) may be required to minimize their undue effects.

Outliers may be influential data points; influential data points need not be outliers (and *vice-versa*).

INFLUENTIAL OBSERVATIONS



Queuing dataset:
processing rate vs. arrival rate



ANOMALY DETECTION REMARKS

Identifying influential points is an **iterative process** as the various analyses have to be run numerous times.

Fully automated identification and removal of anomalous observations is **NOT recommended**.

Use data transformations if the data is **NOT normally distributed**.

Whether an observation is an outlier or not depends on **various factors**; what observations end up being influential data points depends on the **specific analysis to be performed**.

DIMENSIONALITY OF DATA

In data analysis, the **dimension** of the data is the number of attributes that are collected in a dataset, represented by the **number of columns**.

We can think of the number of variables used to describe each object (row) as a vector describing that object: the dimension is simply the **size** of that vector.

(**Note:** "dimension" is used differently in business intelligence contexts)

HIGH DIMENSIONALITY AND BIG DATA

Datasets can be “big” in a variety of ways:

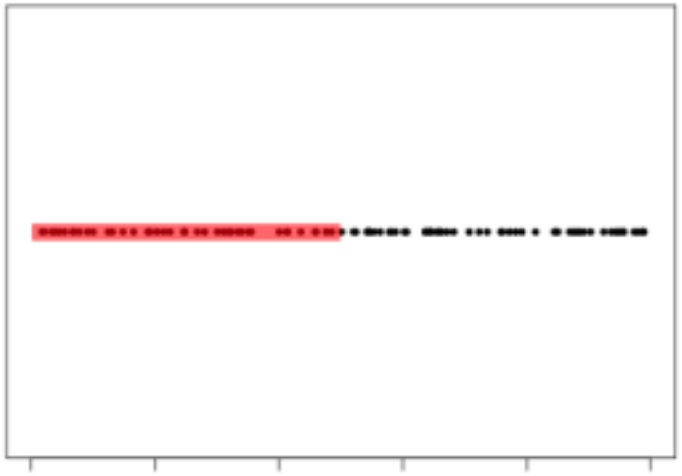
- too large for the **hardware** to handle (cannot be stored, accessed, manipulated properly due to # of observations, # of features, the overall size)
- dimensions can go against **modeling assumptions** (# of features ≫ # observations)

Examples:

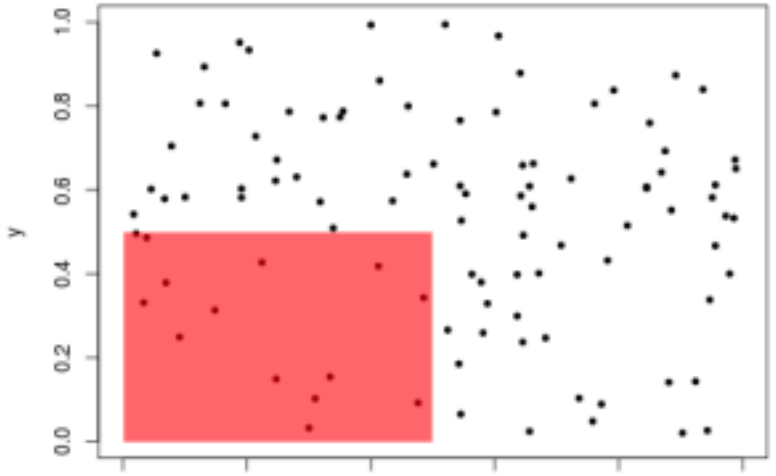
- multiple sensors recording 100+ observations per second in a large geographical area over a long time period = **very big dataset**
- in a corpus’ *Term Document Matrix* (cols = terms, rows = documents), the number of terms is usually substantially higher than the number of documents, leading to **sparse data**

CURSE OF DIMENSIONALITY

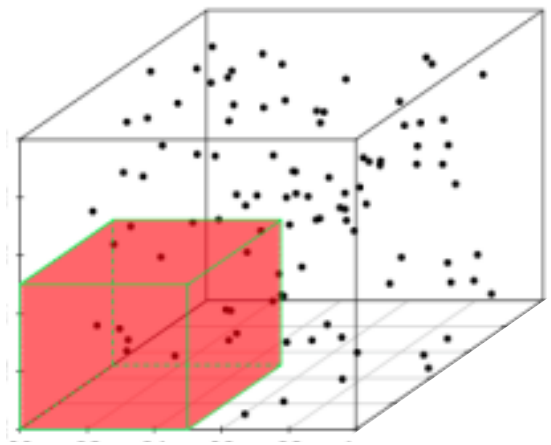
42% of data is captured



14% of data is captured



7% of data is captured



$N = 100$ observations, uniformly distributed on $[0,1]^d, d = 1, 2, 3.$
% of observations captured by $[0,1/2]^d, d = 1,2,3.$

SAMPLING OBSERVATIONS

Question: does every row of the dataset need to be used?

If rows are selected randomly (with or without replacement), the resulting sample might be **representative** of the entire dataset.

Drawbacks:

- if the signal of interest is rare, sampling might drown it altogether
- if aggregation is happening down the road, sampling will necessarily affect the numbers (passengers vs. flights)
- even simple operations on a large file (finding the # of lines, say) can be taxing on the memory – **prior information on the dataset structure can help**

FEATURE SELECTION

Removing **irrelevant/redundant** variables is a common data processing task.

Motivations:

- modeling tools do not handle these well (variance inflation due to multicolinearity, etc.)
- dimension reduction (# variables ≫ # observations)

Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

COMMON TRANSFORMATIONS

Models sometimes require that certain data assumptions be met (normality of residuals, linearity, etc.).

If the raw data does not meet the requirements, we can either:

- abandon the model
- attempt to **transform** the data

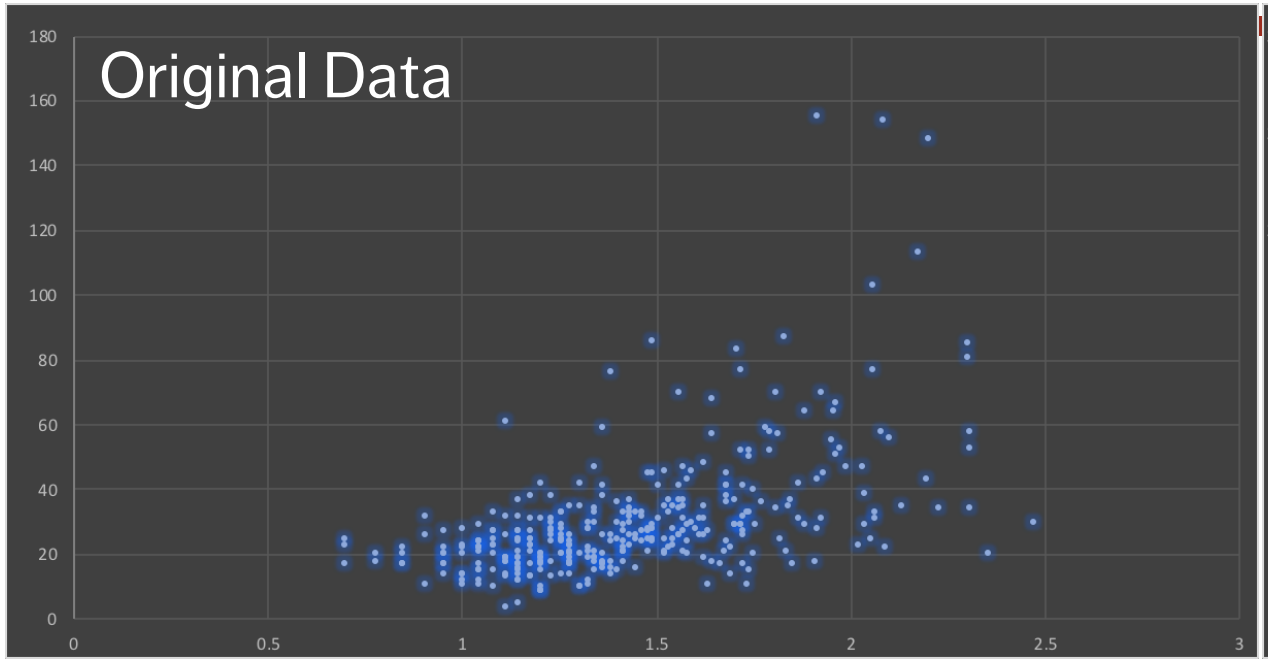
The second approach requires an **inverse transformation** to be able to draw conclusions about the **original data**.

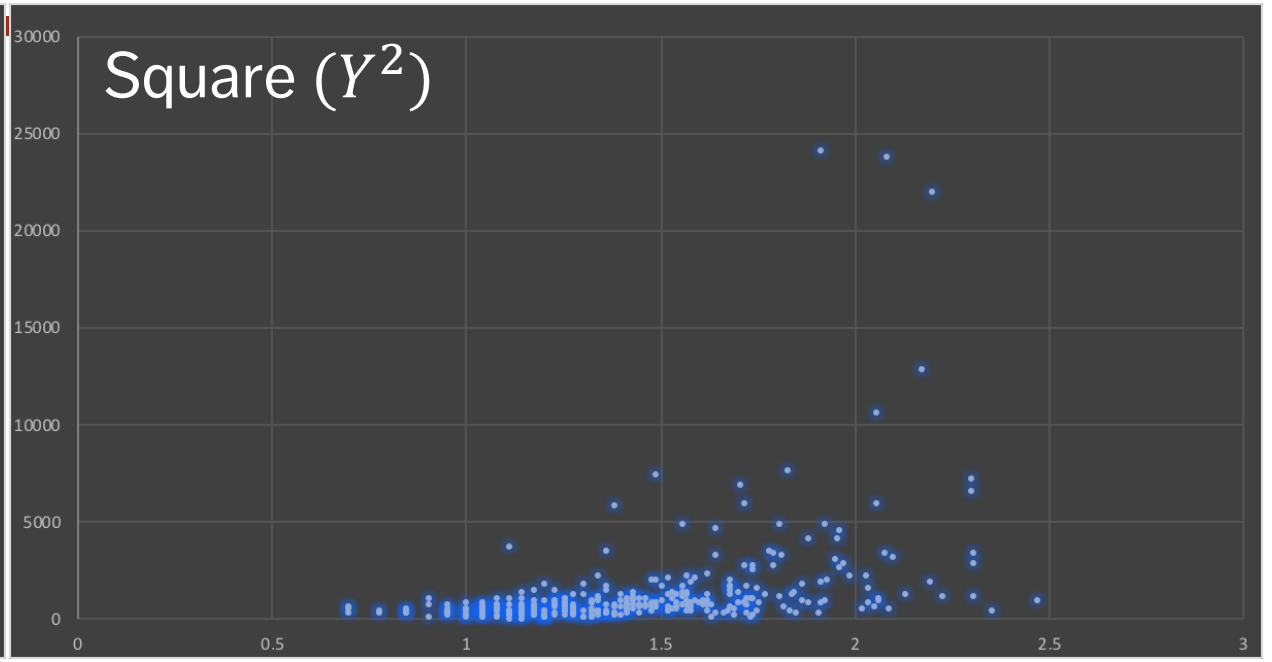
COMMON TRANSFORMATIONS

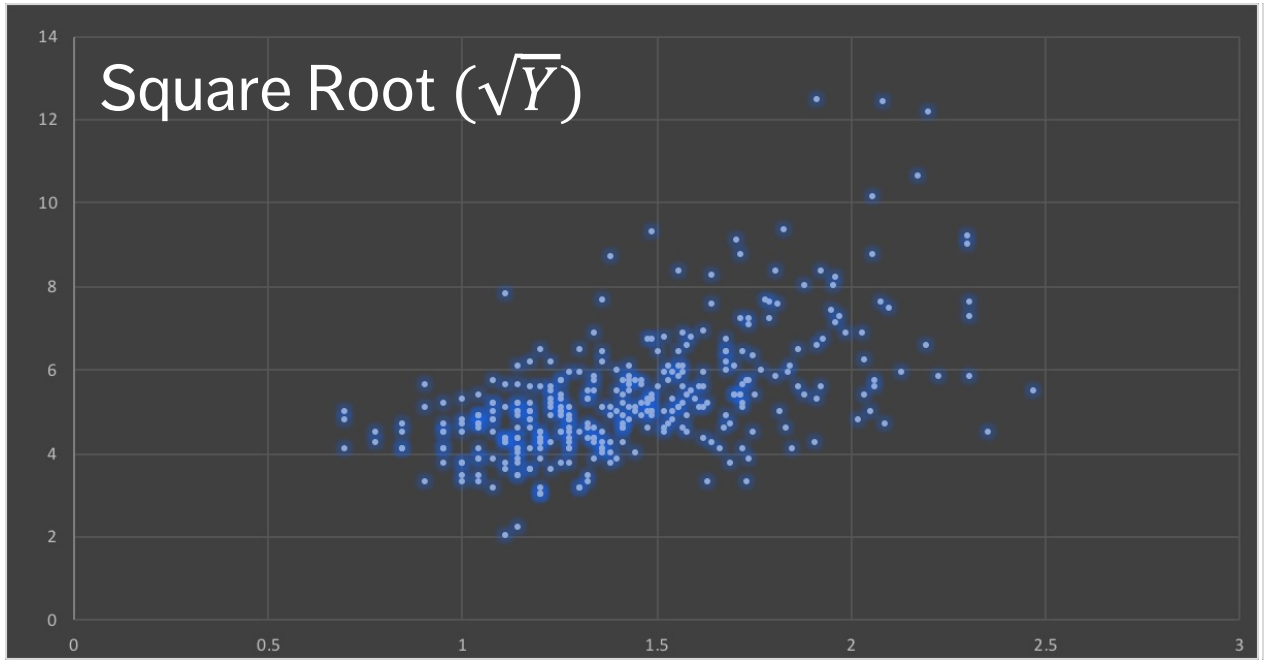
In the data analysis context, transformations are **monotonic:**

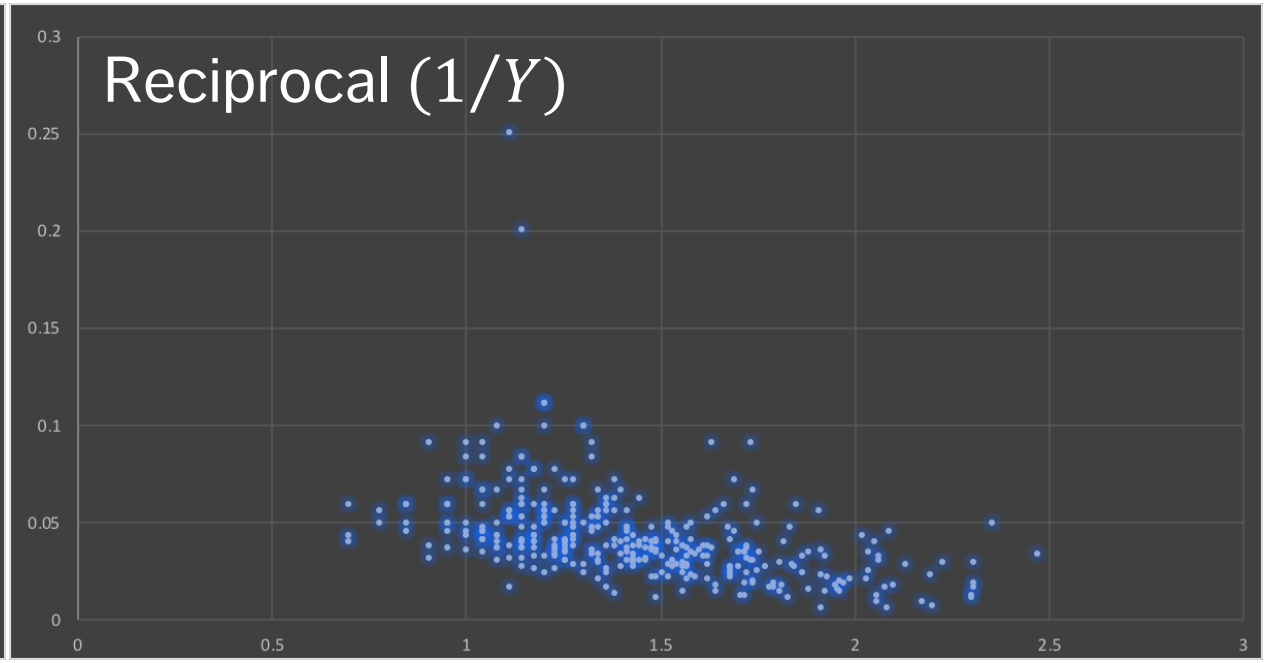
- logarithmic
- square root, inverse, power: W^k
- exponential
- Box-Cox, etc.

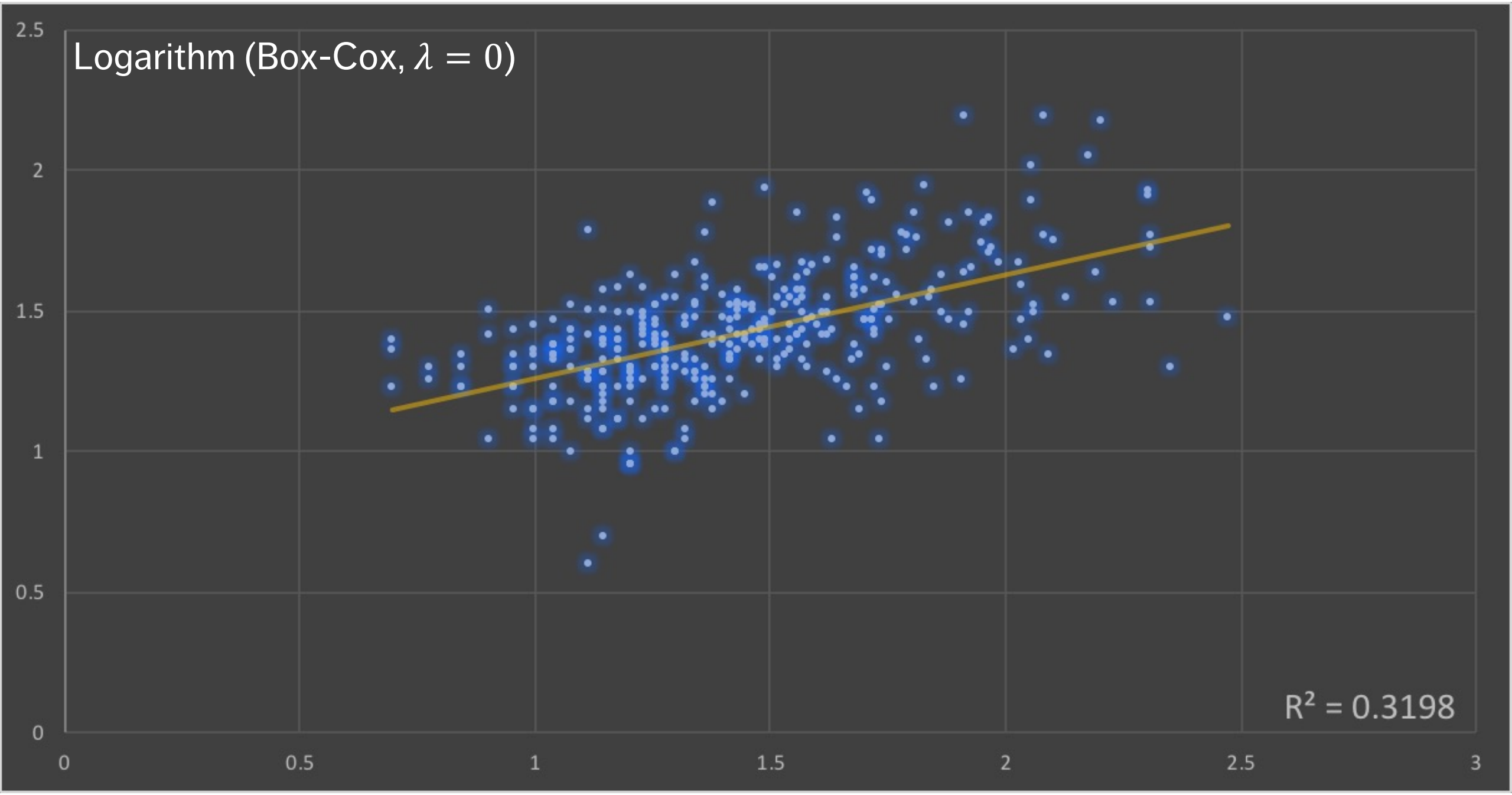
Transformations on X may achieve linearity, but usually at some price (correlations are not preserved, for instance). Transformations on Y can help with non-normality and unequal variance of error terms.


Original Data
180
160
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3


Square (Y^2)
30000
25000
20000
15000
10000
5000
0
0
0.5
1
1.5
2
2.5
3


Square Root ($\sqrt{Y}$)
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3


Reciprocal ($1/Y$)
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.5
1
1.5
2
2.5
3


Logarithm (Box-Cox, $\lambda = 0$)
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
$R^2 = 0.3198$

SCALING

Numeric variables may have different **scales** (i.e., weights and heights).

The variance of a large-range variable is typically greater than that of a small-range variable, introducing a bias (for instance).

Standardization creates a variable with mean 0 and std. dev. 1:

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

Normalization creates a new variable in the range $[0,1]$: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

DISCRETIZING

To reduce computational complexity, a numeric variable may need to be replaced by an **ordinal** variable (from *height* value to "*short*", "*average*", "*tall*", for instance).

Domain expertise can be used to determine the bins' limits (although that may introduce unconscious bias to the analyses)

In the absence of such expertise, limits can be set so that either

- the bins each contain the same number of observations
- the bins each have the same width
- the performance of some modeling tool is maximized

CREATING VARIABLES

New variables may need to be introduced:

- as **functional relationships** of some subset of available features
- because modeling tool may require **independence of observations**
- because modeling tool may require **independence of features**
- to simplify the analysis by looking at **aggregated summaries** (often used in text analysis)

Time dependencies ⟶ time series analysis (lags?)

Spatial dependencies ⟶ spatial analysis (neighbours?)

9. STORYTELLING AND VISUALIZATION

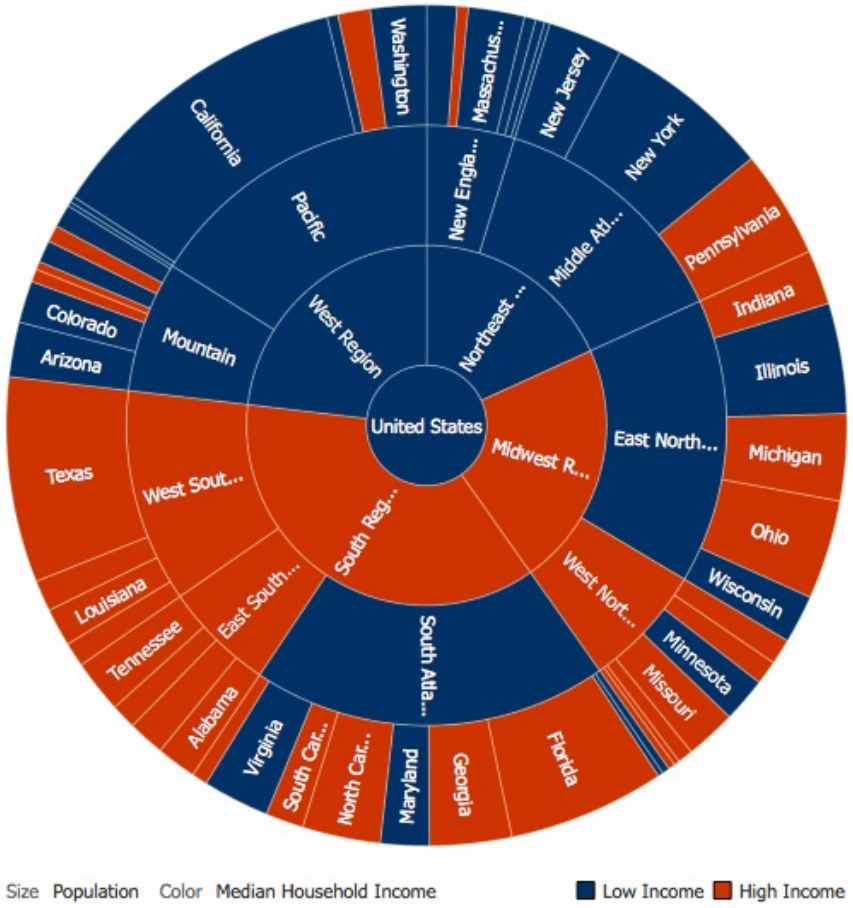
DATA ANALYSIS AND VISUAL STORYTELLING

[Author unknown]

DATA VISUALIZATION VS. INFOGRAPHICS

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**



data-action-lab.com

[J.P. Blackard]

DATA VISUALIZATION VS. INFOGRAPHICS

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

HISTORICAL CHARTS

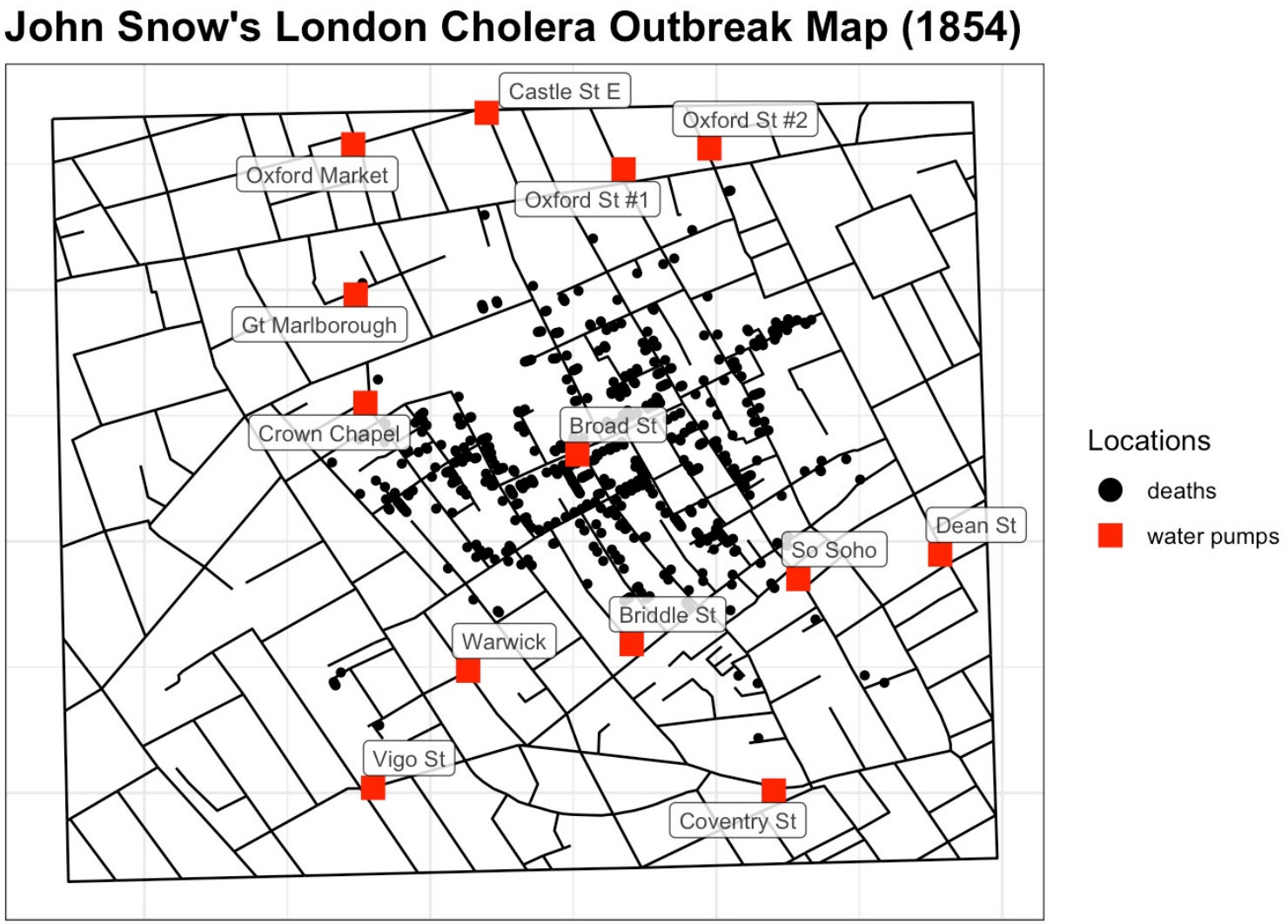
Data visualization is not confined to the recent past: charts have been used for many years to help **communicate information** and **tell stories**.

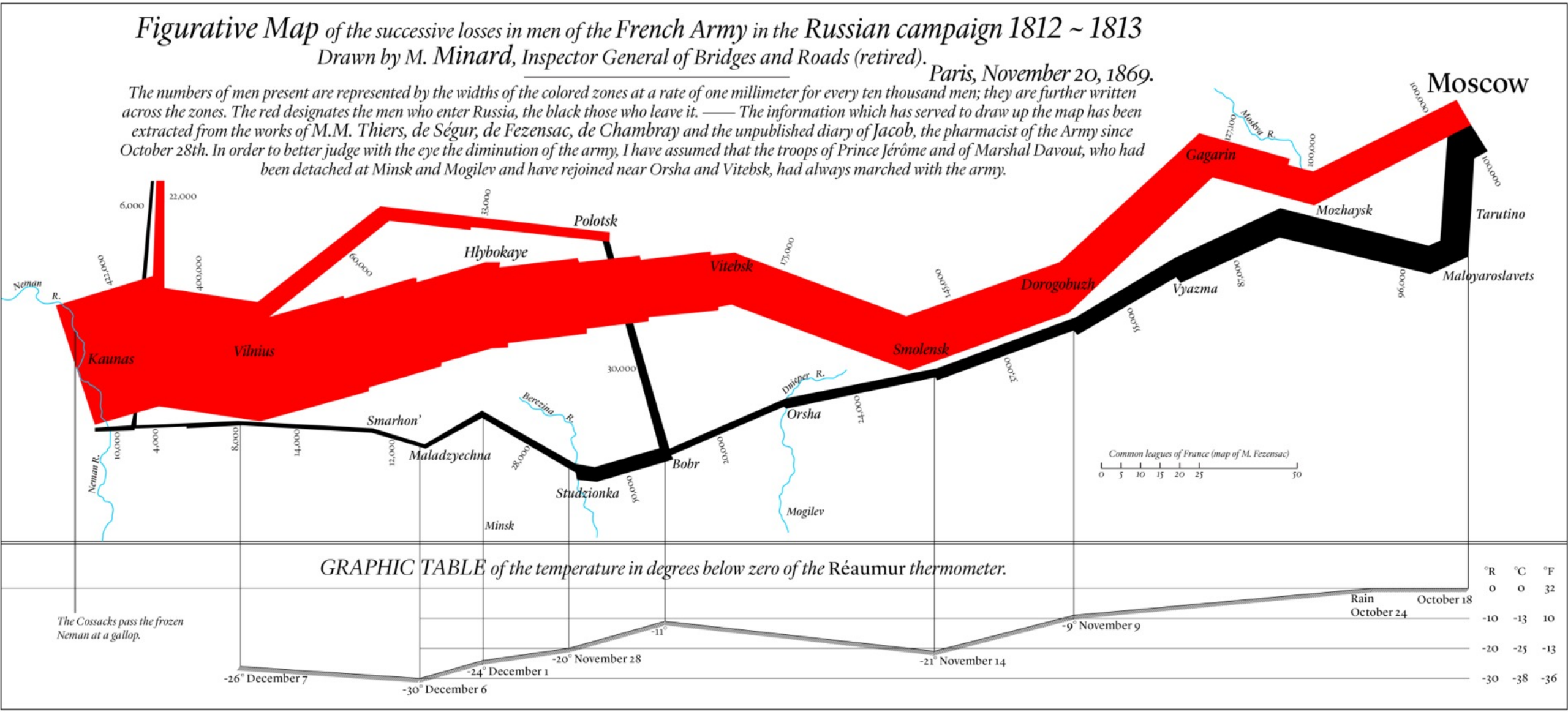
Due to the absence of technical tools, a lot of thought had to go into the design and creation of these visualizations.

Consequently, there is a lot we can (and **should**) learn to bring into the development of charts from a **design and storytelling perspective**.

London's Cholera Outbreak of 1854

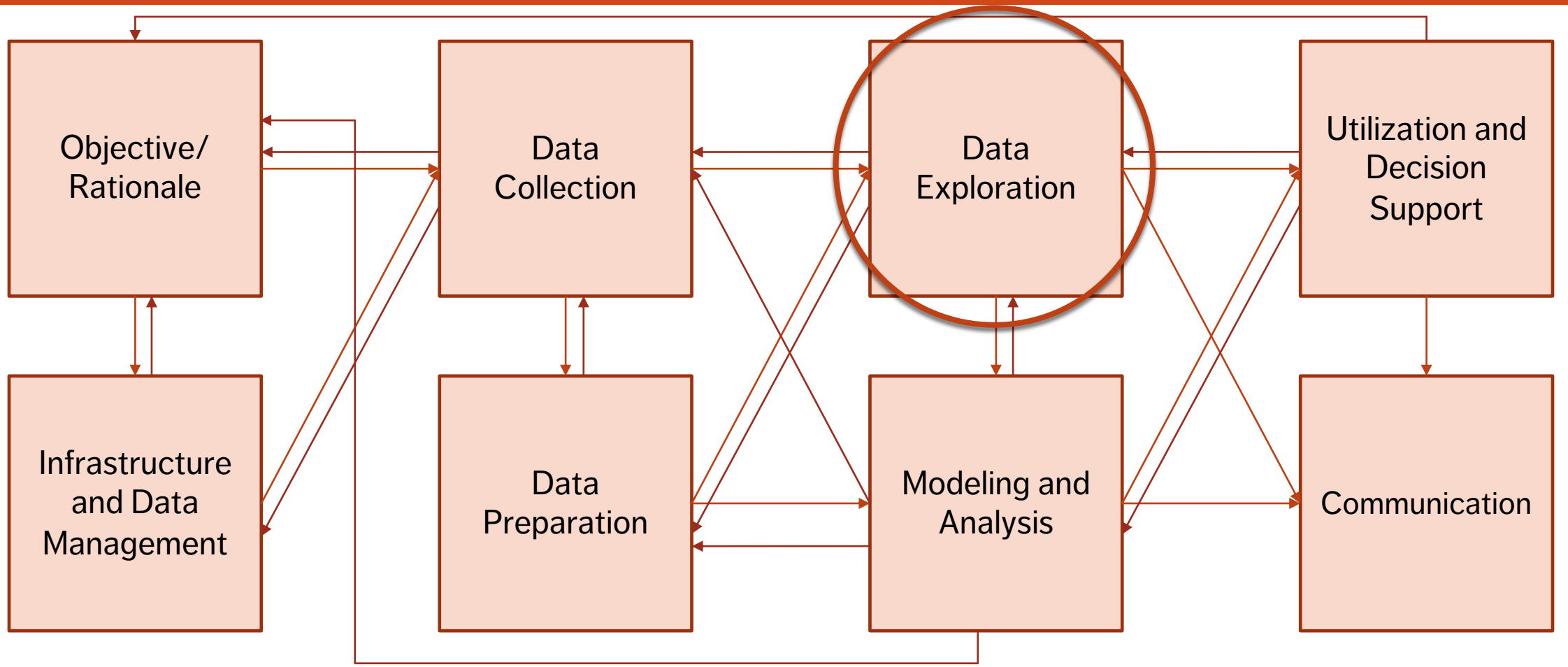
Physician John Snow links the outbreak to a contaminated well by plotting number of cases on a map, jump-starting the science of epidemiology.





Minard's March to Moscow

THE (MESSY) ANALYSIS PROCESS



NON-VISUALIZATION SUMMARIES

```
autumn spring summer winter
    80     84     86     90
```

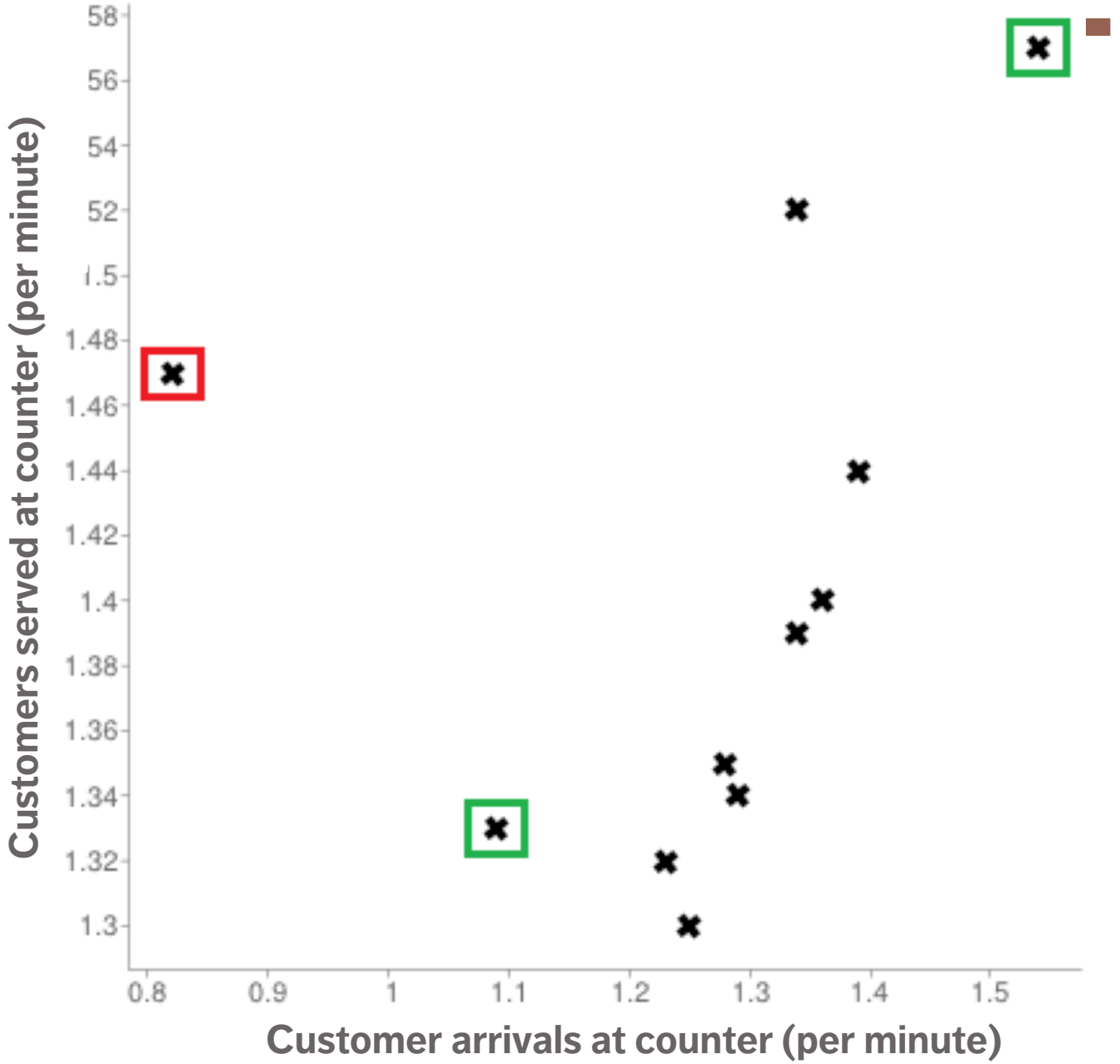
```
       Cl              NO3              NH4          
 Min.   :  0.222   Min.   : 0.000   Min.   :    5.00  
 1st Qu.: 10.994   1st Qu.: 1.147   1st Qu.:   37.86  
 Median : 32.470   Median : 2.356   Median :  107.36  
 Mean   : 42.517   Mean   : 3.121   Mean   :  471.73  
 3rd Qu.: 57.750   3rd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

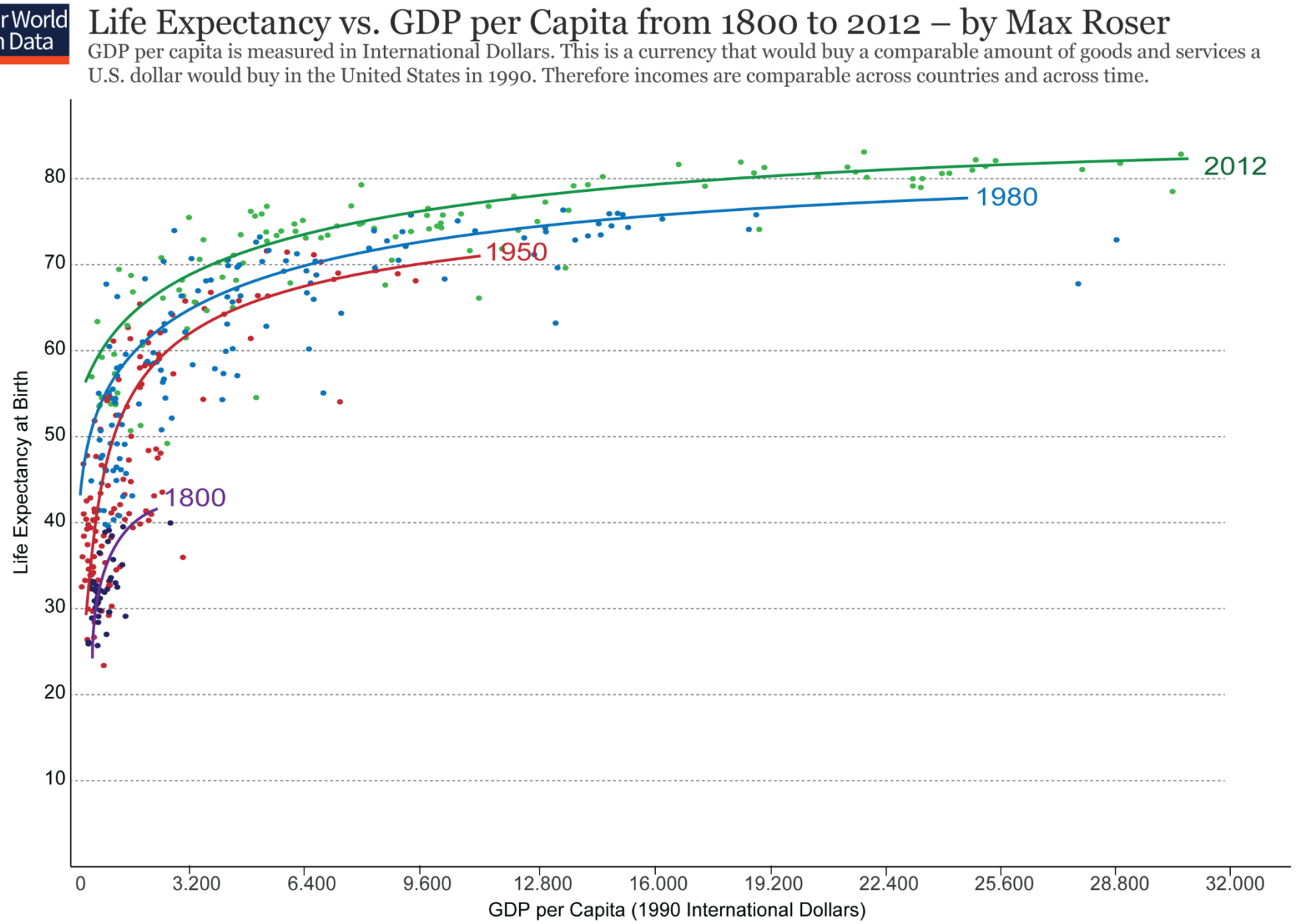
PRE-ANALYSIS USE

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


Customers served at counter (per minute)
58
56
54
52
1.5
1.48
1.46
1.44
1.42
1.4
1.38
1.36
1.34
1.32
1.3
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
Customer arrivals at counter (per minute)

[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.

data-action-lab.com

LINE CHART/RUG CHART

Gaps in the number line: **absence** of those numeric values in the data.

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.

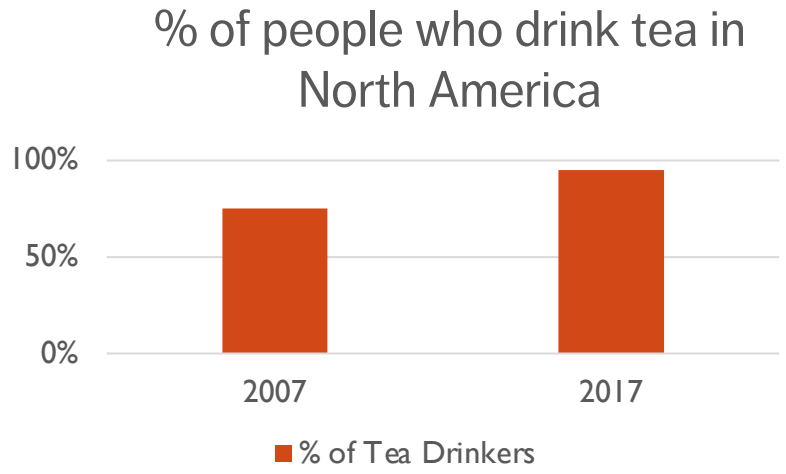
If some values are identical, they lie on top of each other (use **jitter**?).

SIMPLE TEXT

One or two numbers to focus on.

Good at “setting the scene”.

Draws focus to an area of the report.



95% of the population

drinks tea today compared to

75% in 2007

TABLES

Tables interact with our **verbal** system, which means we **read** them:

- used to compare values
- audiences will look for their rows

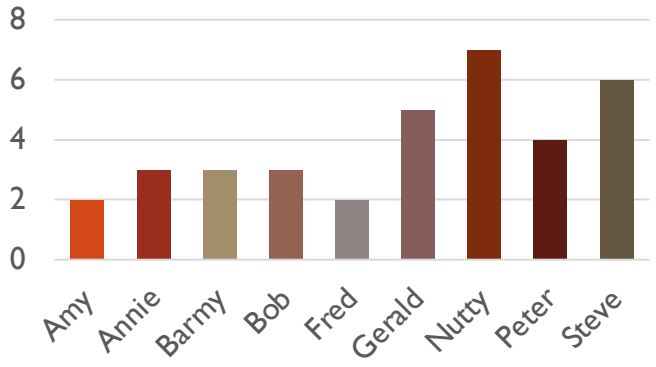
Table design needs to **blend** into background

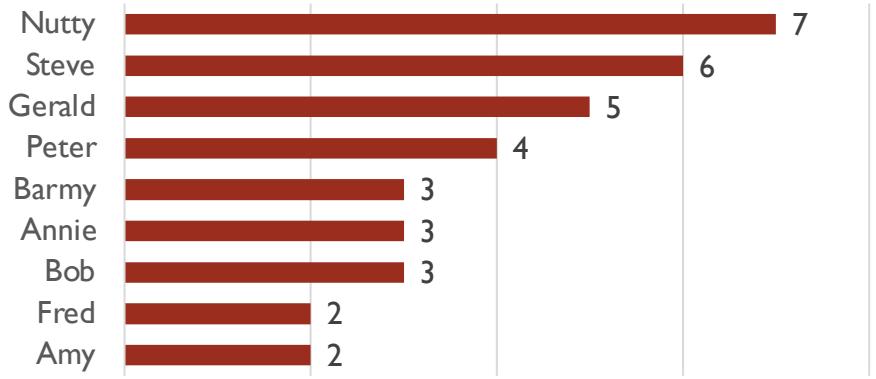
- the data should stand out, not the borders
- dense table/data: use alternating row colour

Name	Last Year	This Year
Bob	20	30
Fred	30	40
George	10	15

Name	Last Year	This Year
Bob	20	30
Fred	30	40
George	10	15

BAR CHARTS





Very versatile and useful.

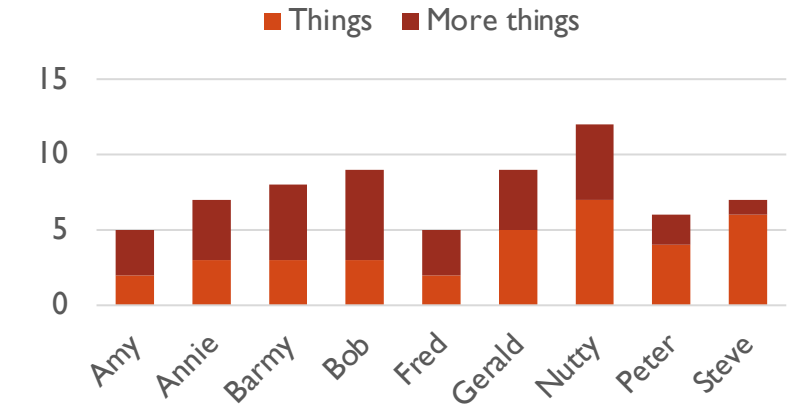
ALWAYS (?) have a zero baseline.

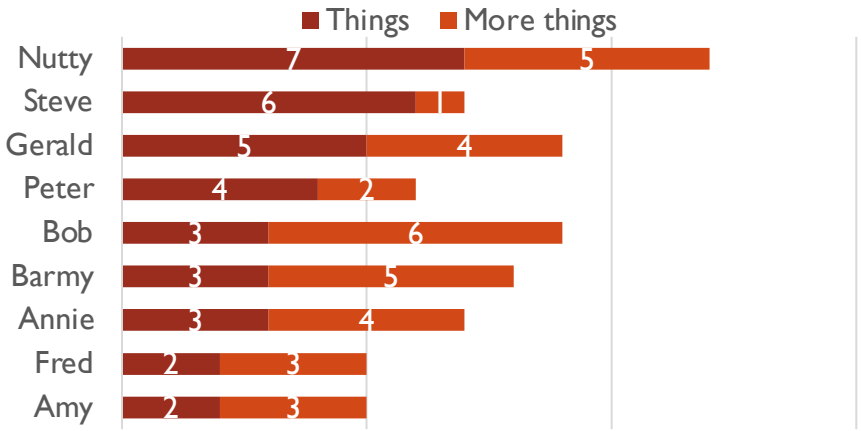
Use graph axis OR data labels. Axis for broad statements, data labels for more detail.

Horizontal charts are apparently **easier to read** (according to many studies).

Think about the ordering of categories.

STACKED BAR CHARTS



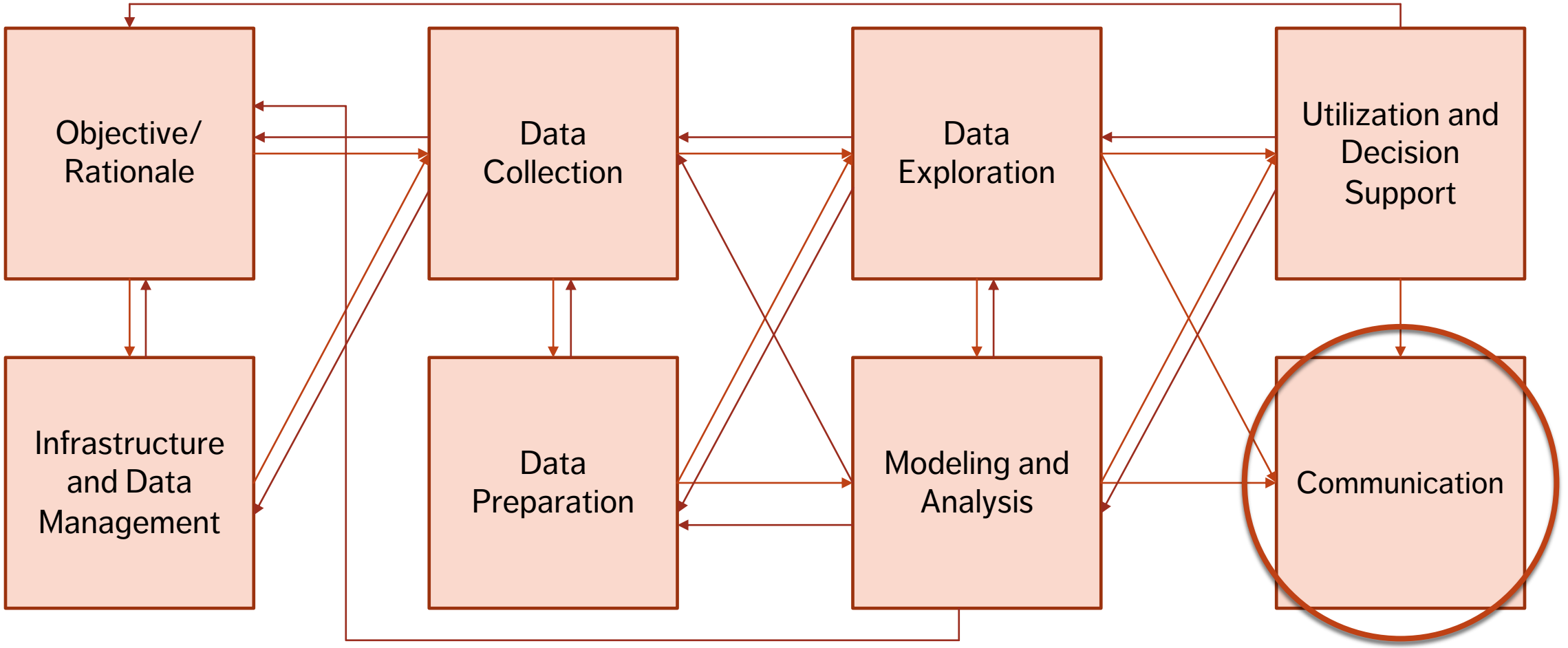


Designed for **comparing totals**, but can quickly become **overwhelming**.

Hard to sort / order.

Filtering is complicated in Power BI (what do you click on & how the chart responds when filter is clicked on?)

THE (MESSY) ANALYSIS PROCESS



[E. Tufte, Beautiful Evidence]

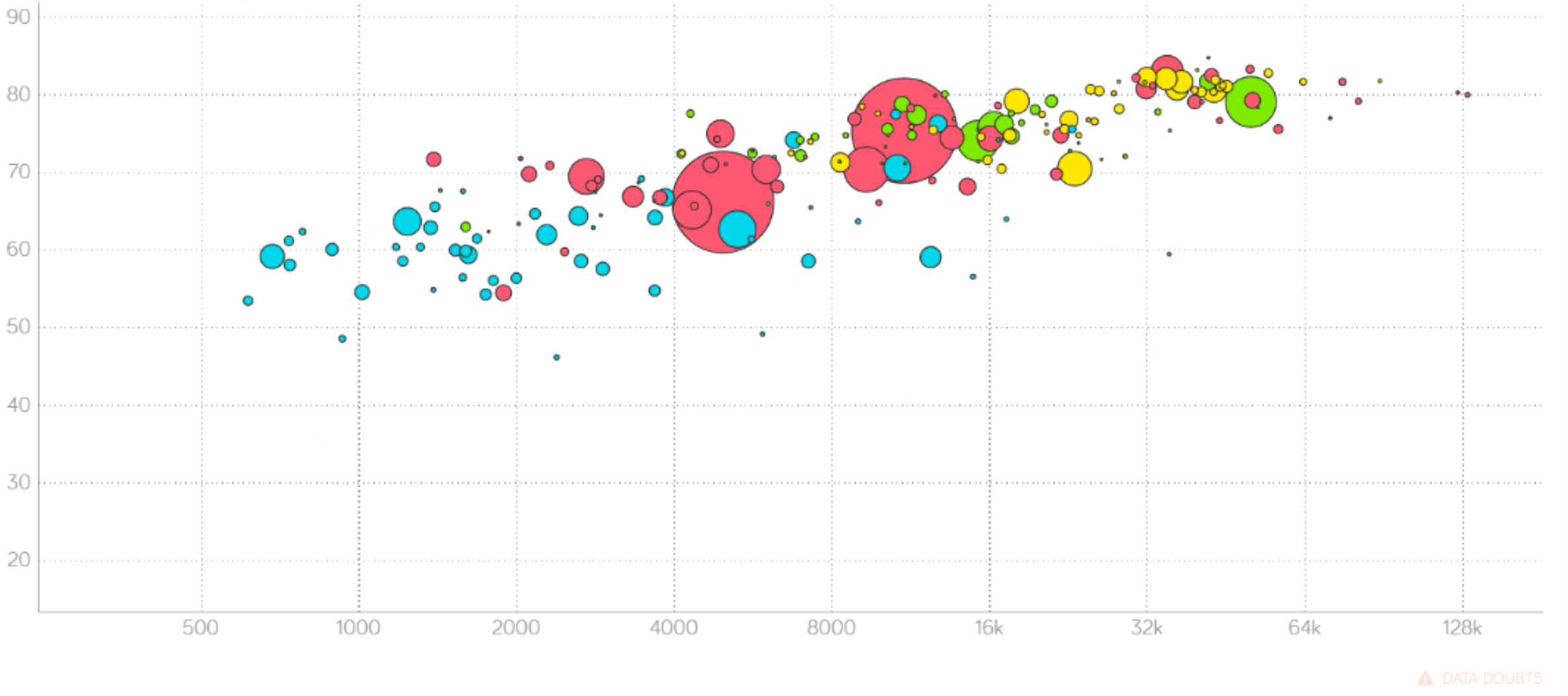
PRINCIPLES OF ANALYTICAL DESIGN

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

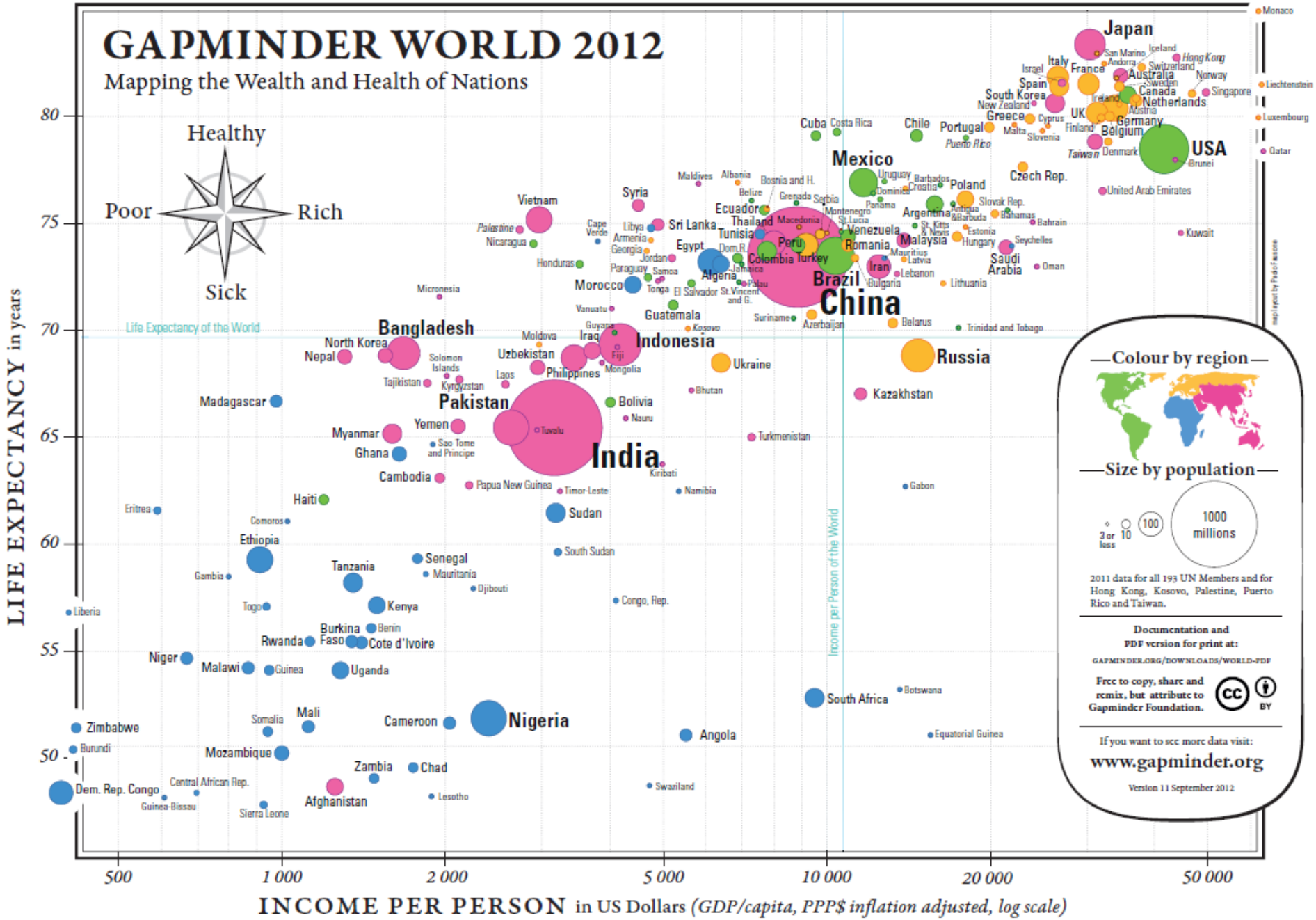
[Gapminder Foundation]



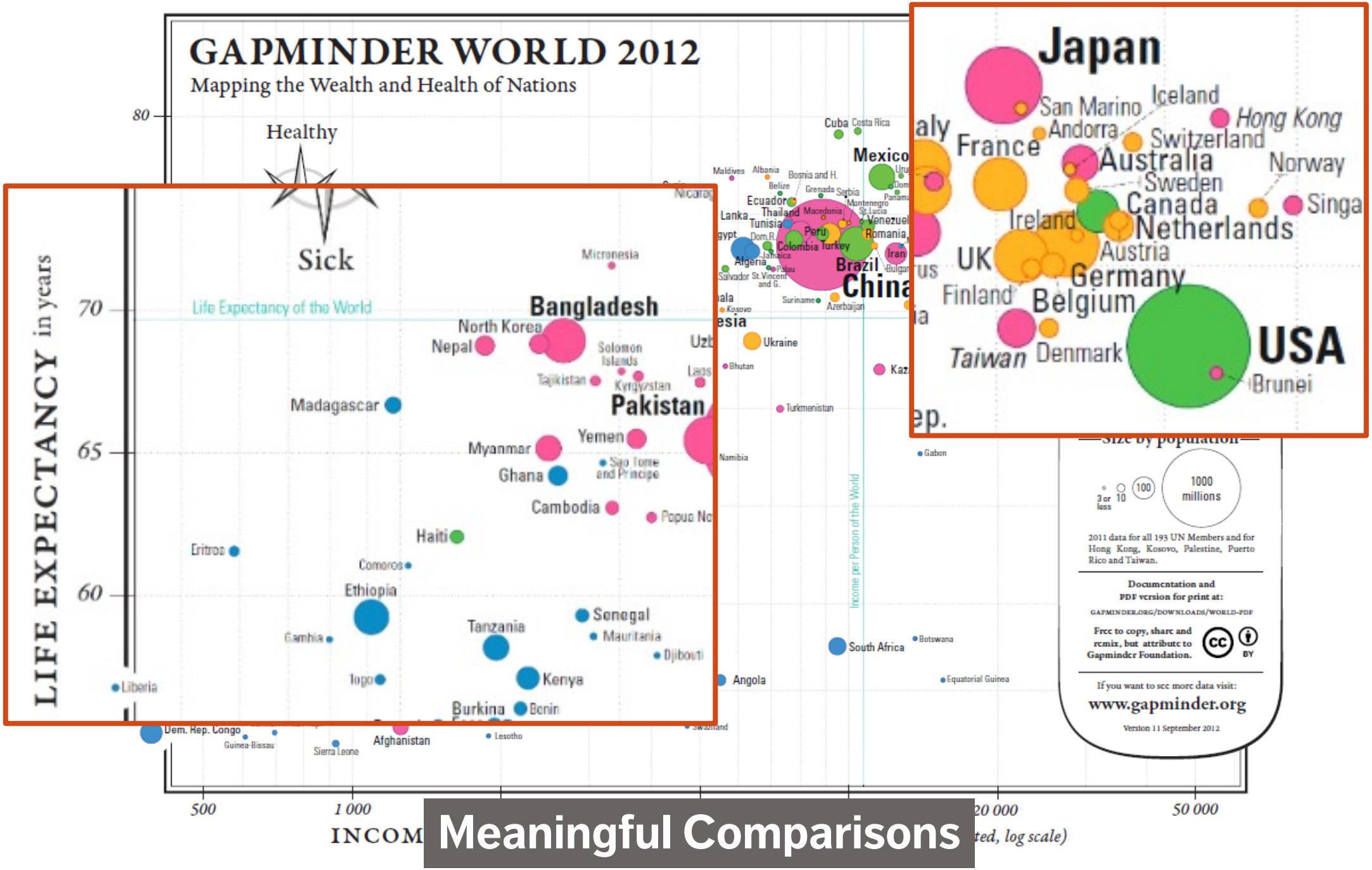
Non-Integrated Data

data-action-lab.com

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

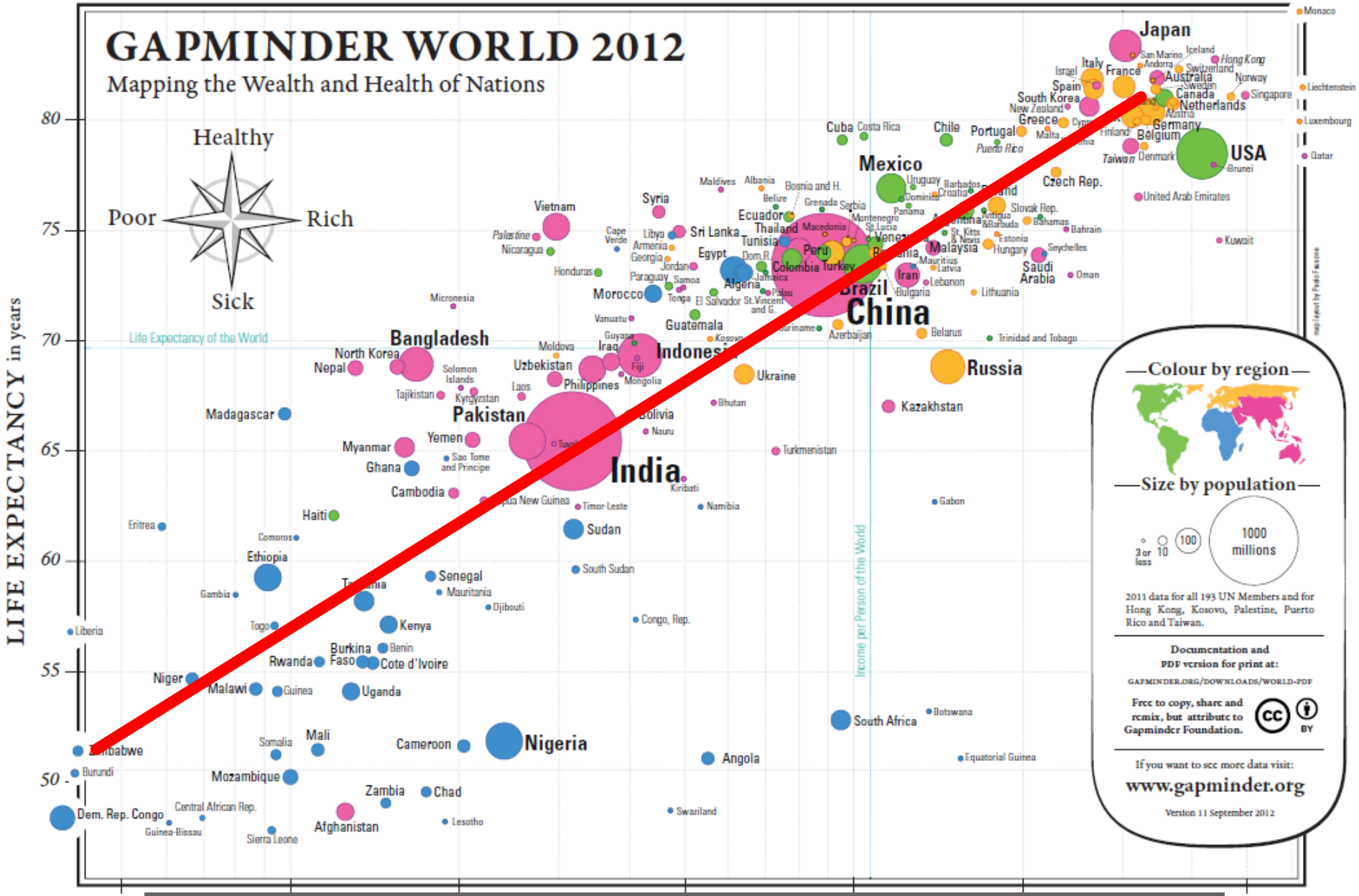


[Gapminder Foundation, https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen]



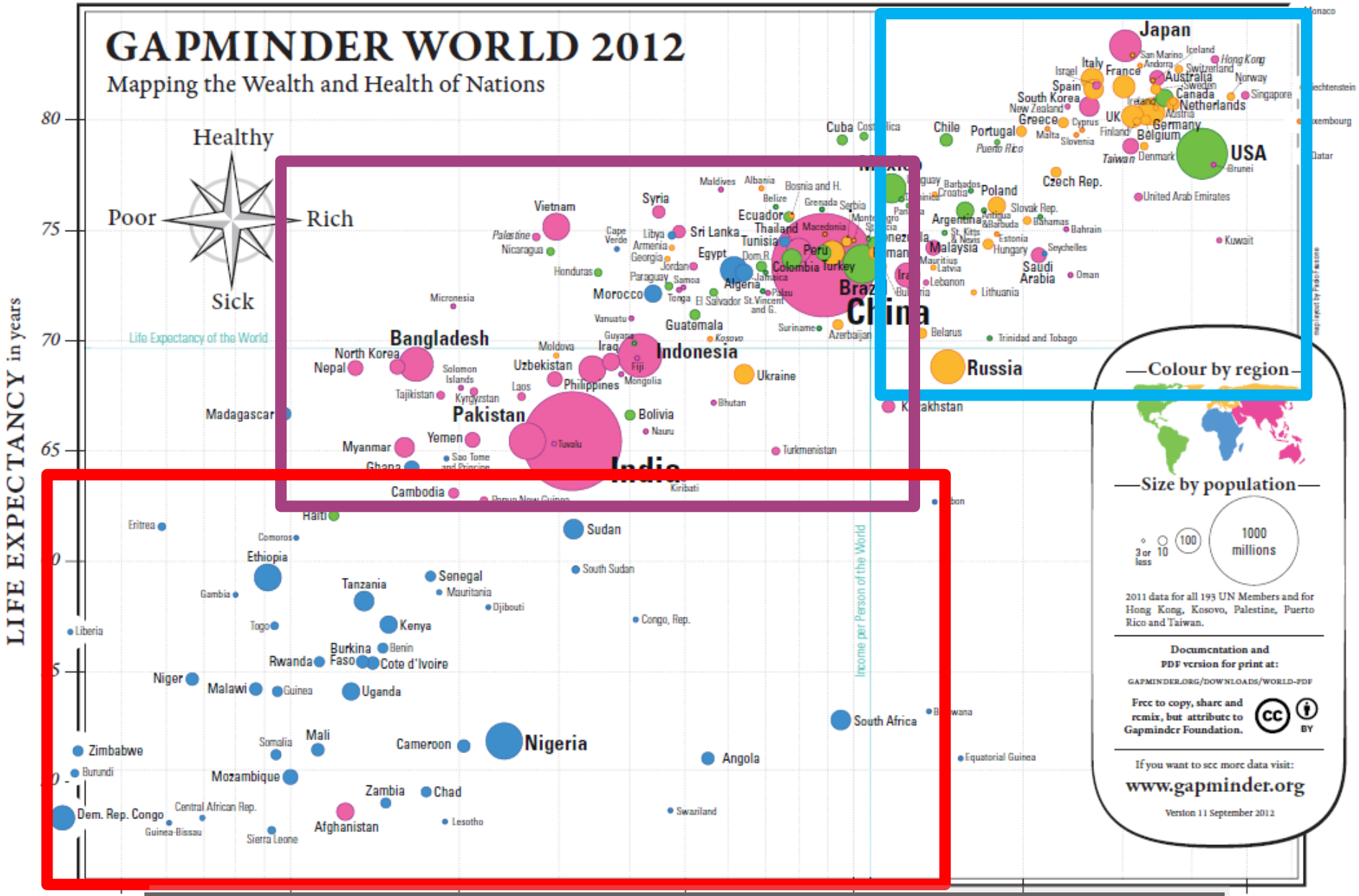
Meaningful Comparisons

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



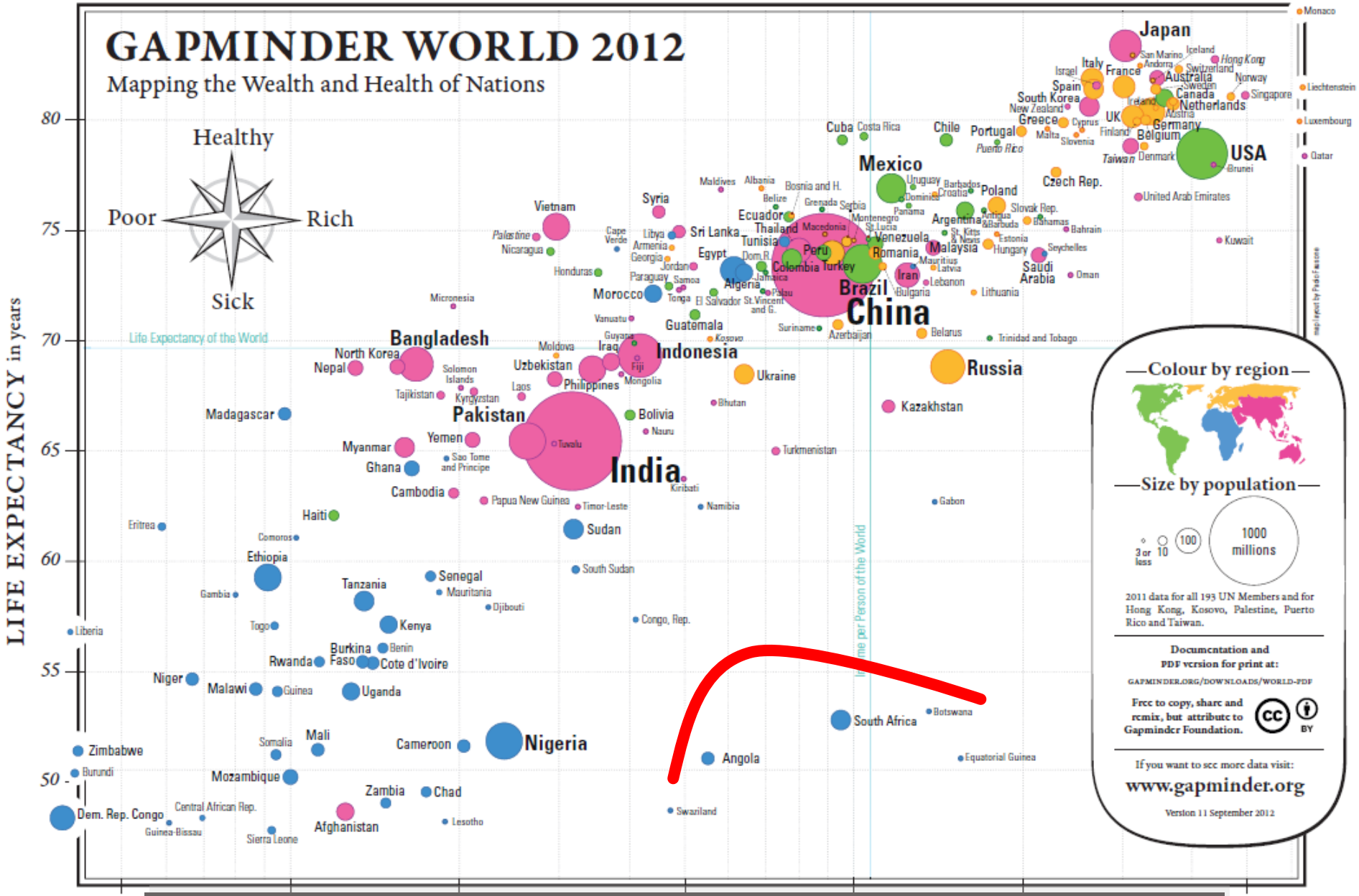
Underlying Structure and Multivariate Links

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



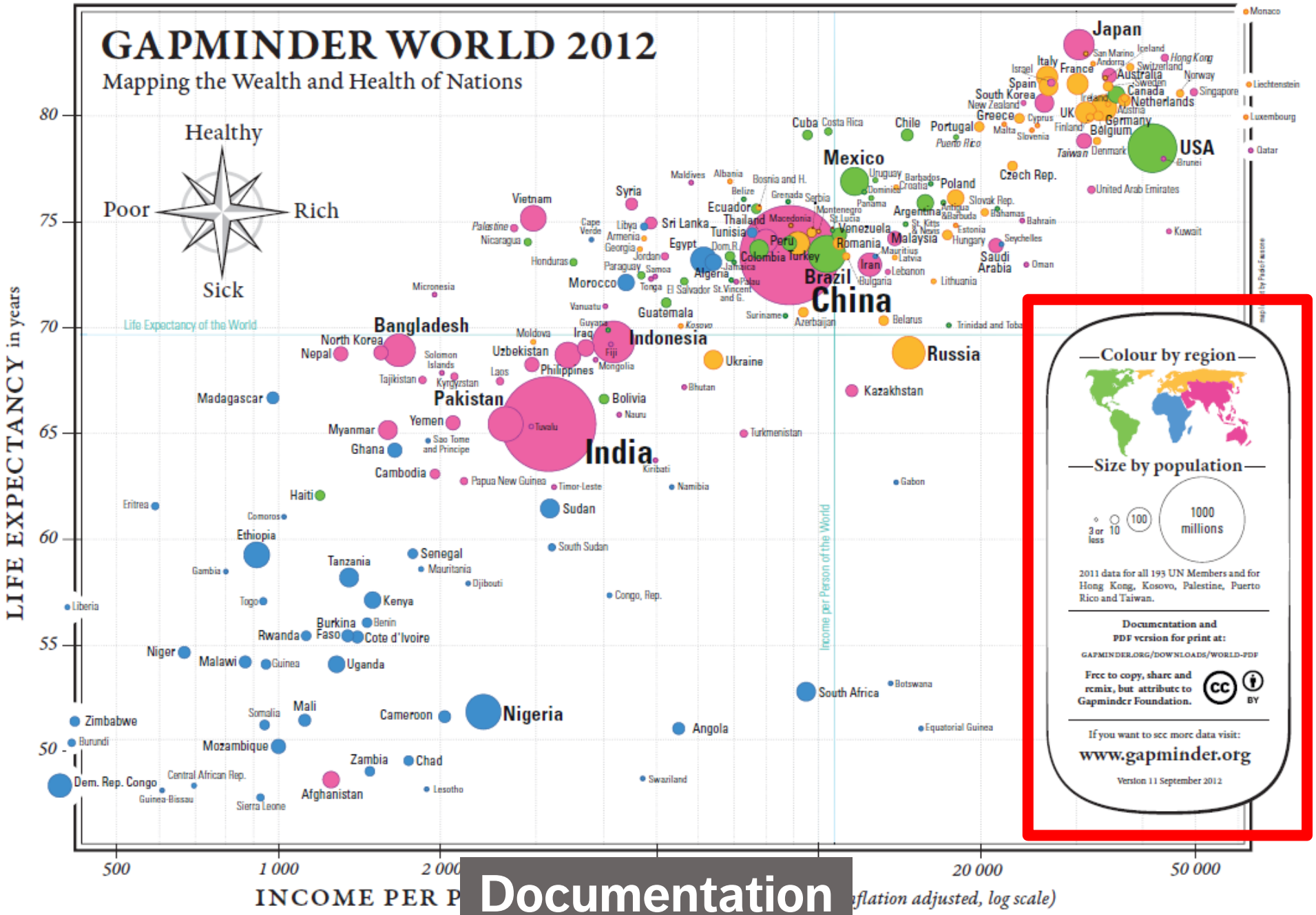
Underlying Structure and Multivariate Links

[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]



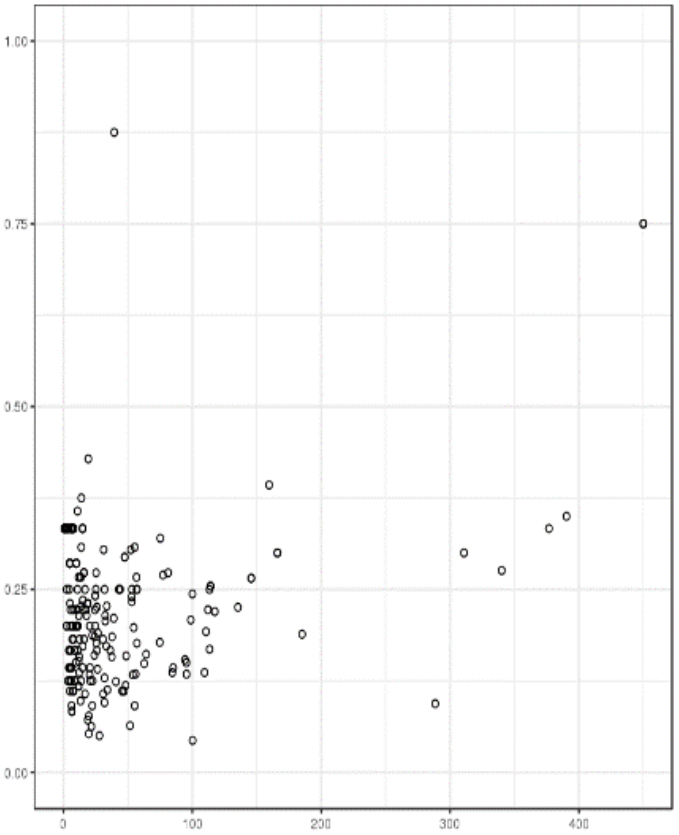
Documentation

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

REPRESENTING MULTIVARIATE DATA

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

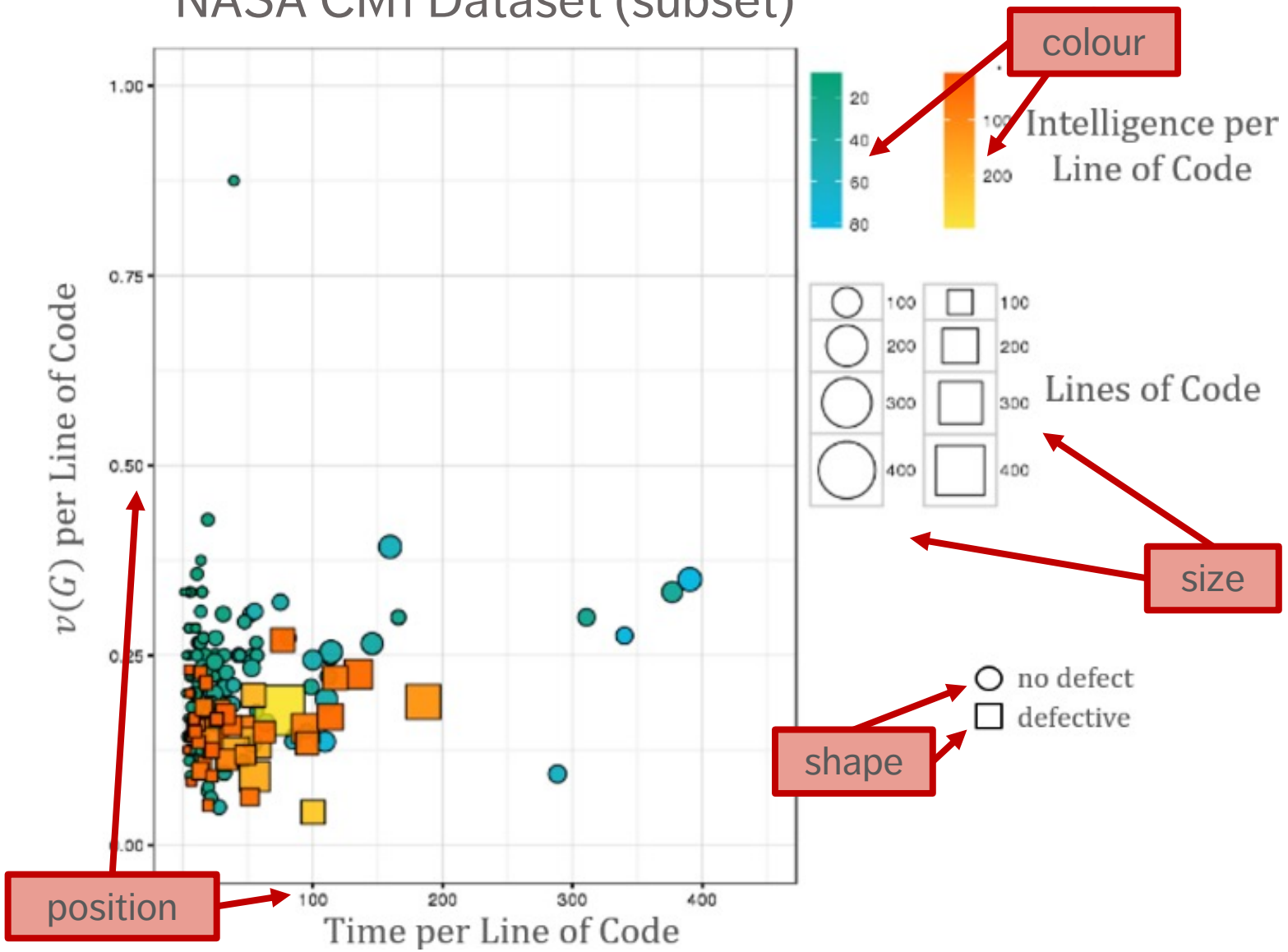
- size
- color
- value
- texture
- line orientation
- shape
- (motion?)



NASA CM1 Dataset (subset)

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

NASA CM1 Dataset (subset)



A CLASSIFICATION OF CHART TYPES

Data comparison charts

Comparison

Bars

Dot plot

Bullet

1D Scatterplot

Heat map

Slope

Alert

Composition

Pie

Pareto

Multidimensional Pie

Distribution

Histogram

1D Scatterplot

Boxplot

Data reduction charts

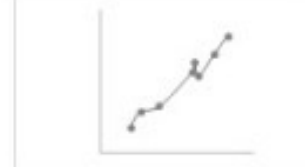
Evolution

Line

Horizon

Step

Connected Scatterplot



Relationship

Scatterplot

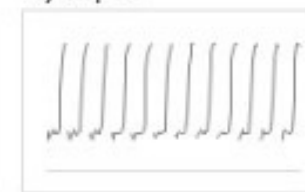
Connected Scatterplot

Bubble

Profiling

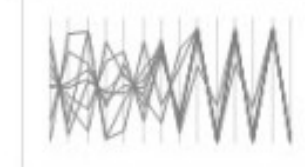
Grouped bars

Cycle plot



Reorderable matrix

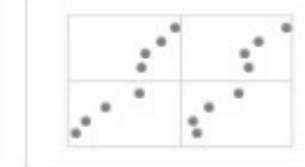
Parallel Plot



Scatterplot matrix

Horizon

Trellis

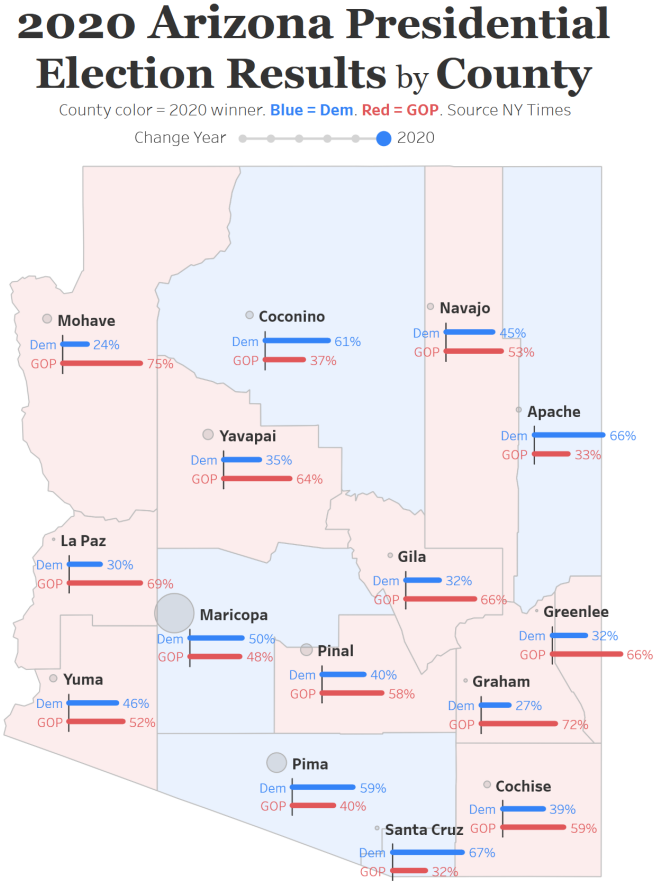


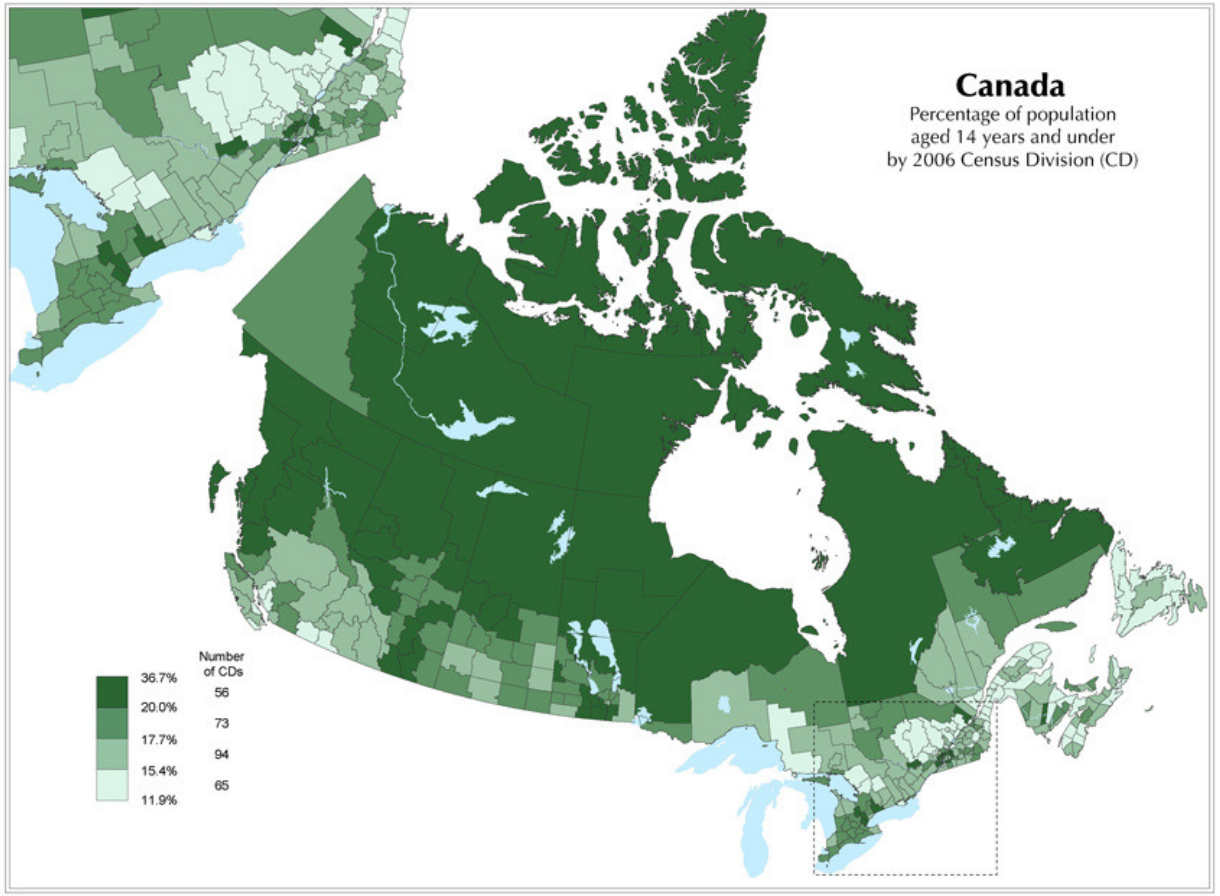
v 0.9

excelcharts.com

[left: A.E. McCann, right: author unknown]

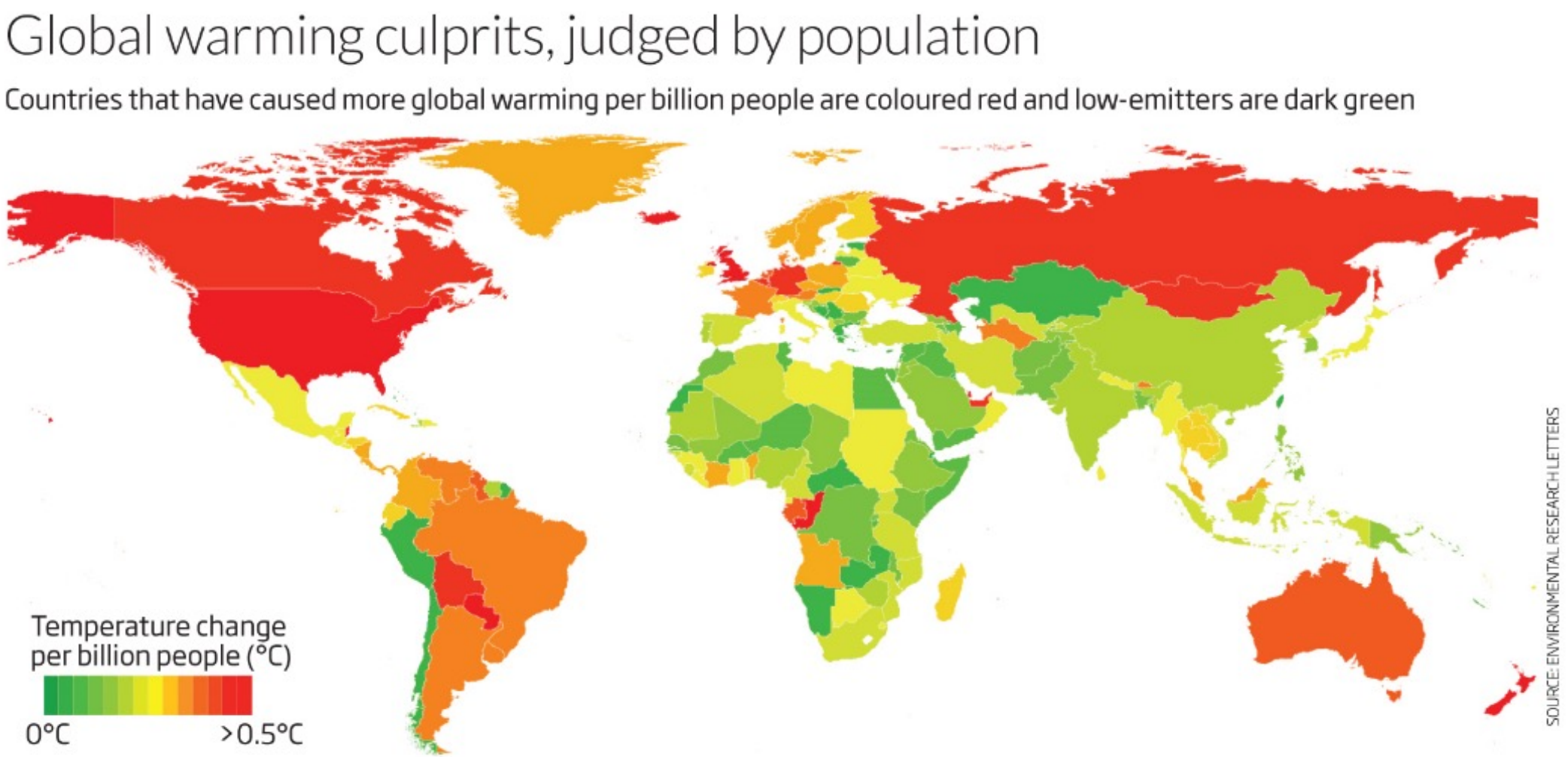
HEAT MAPS (CHOROPLETHS)





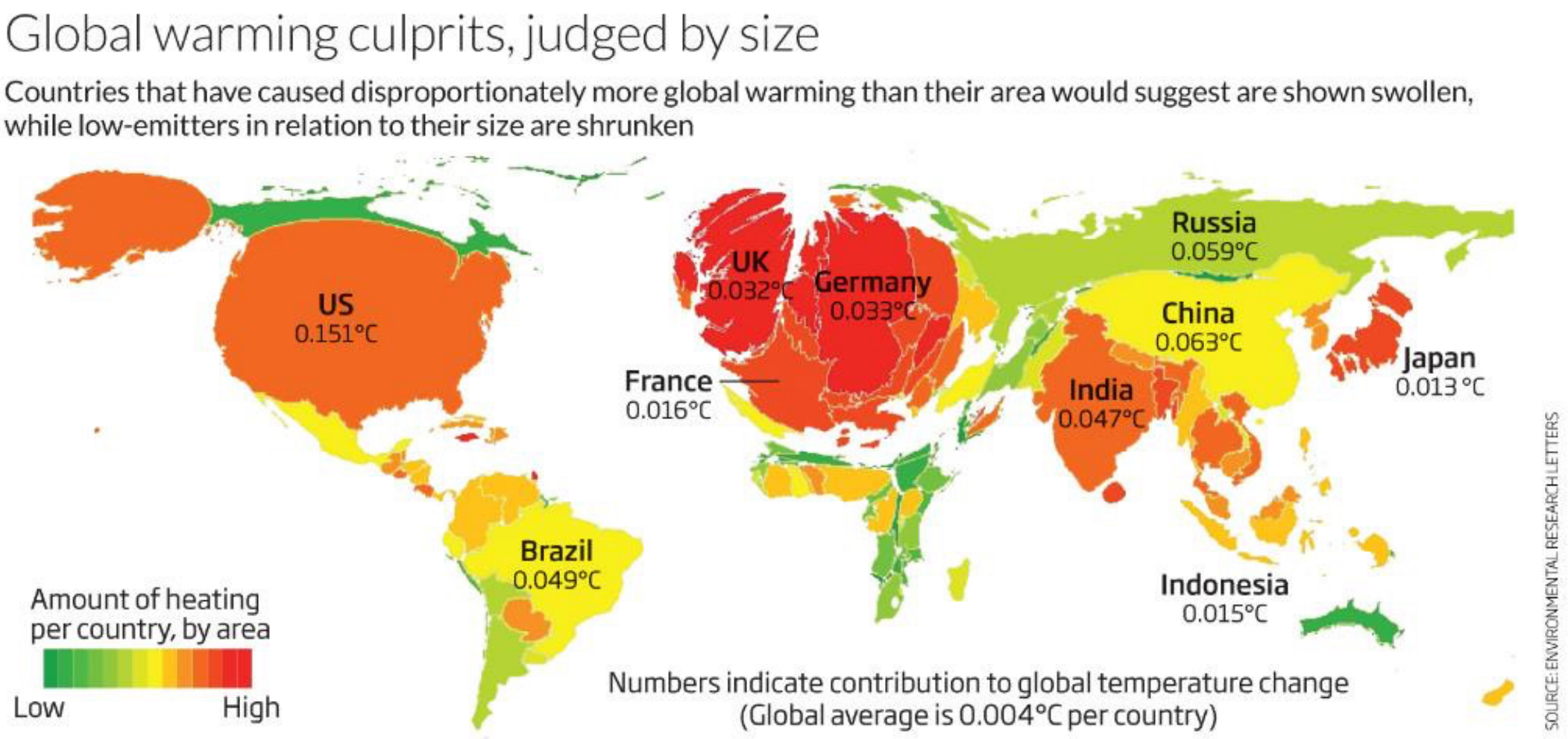
[New Scientist, Jan 2014]

GEOGRAPHICAL MAPS



[New Scientist, Jan 2014]

DISTORTED GEOGRAPHICAL MAPS

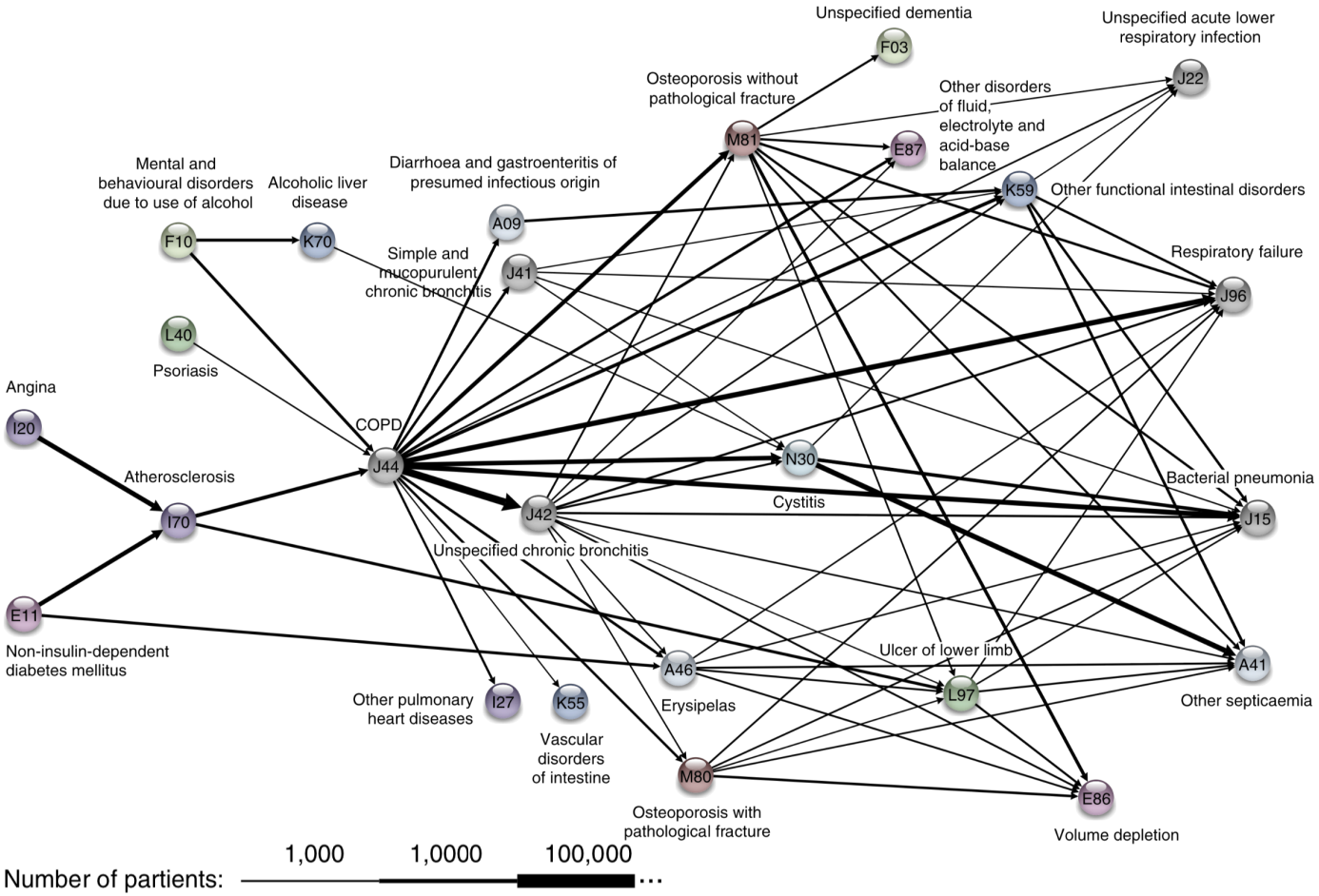


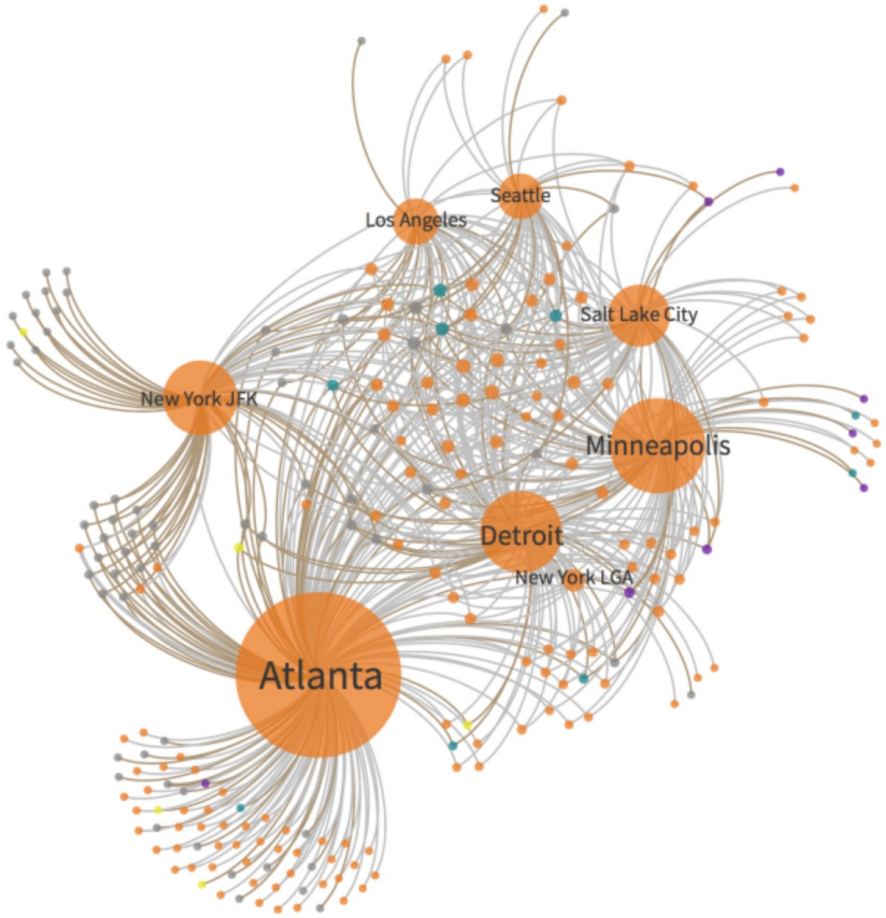
[A.B. Jensen et al., *Nature Communications*]

[Author unknown]

NETWORK DIAGRAMS

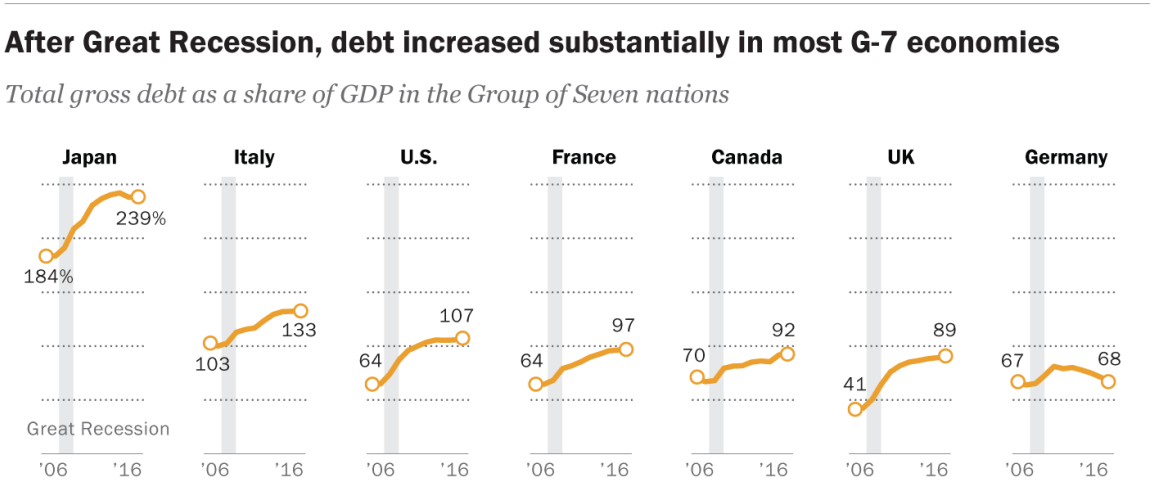
Diagnosis path around COPD in Danish population

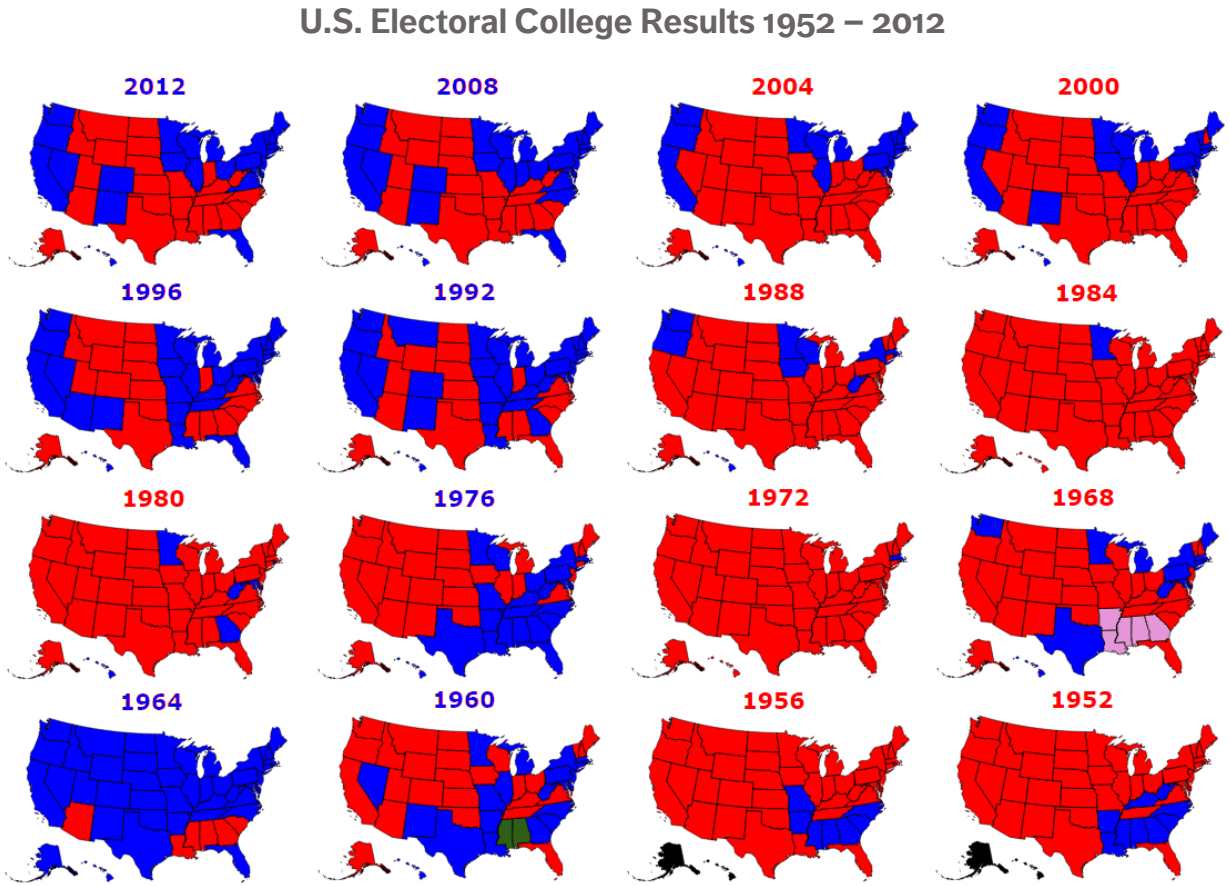




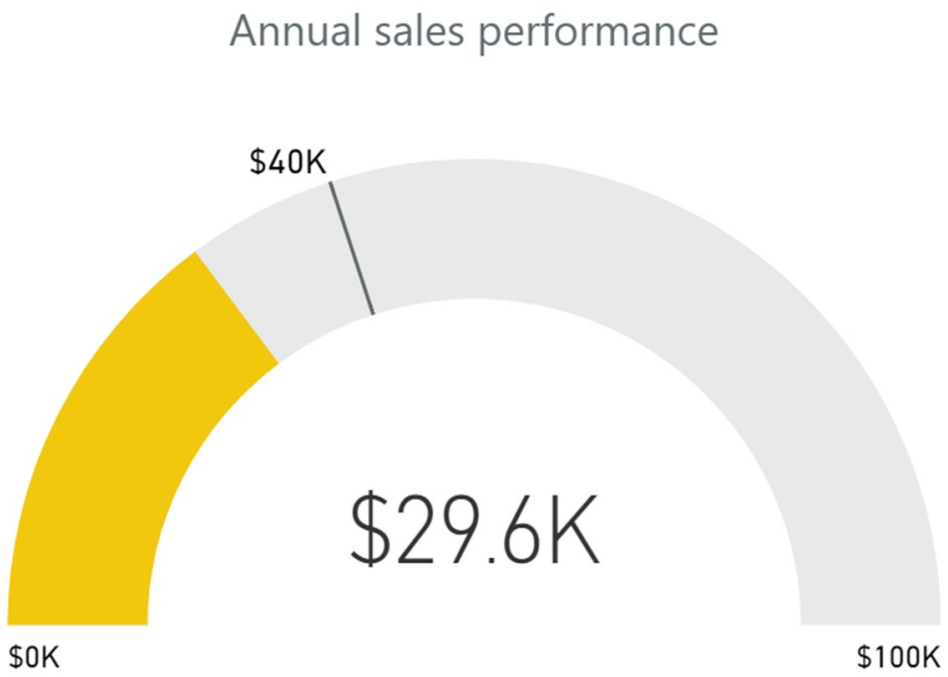
data-action-lab.com

SMALL MULTIPLES





GAUGE CHARTS



Often used as a dashboard component (with or without needle).

Displays single value measures towards goal / KPI.

Great to show progress (a bit of a management fad, though...)

Displays information that can be quickly **scanned** and **understood**.

INTERACTIVE & ANIMATED VISUALIZATIONS

"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 on." (P. Boily)

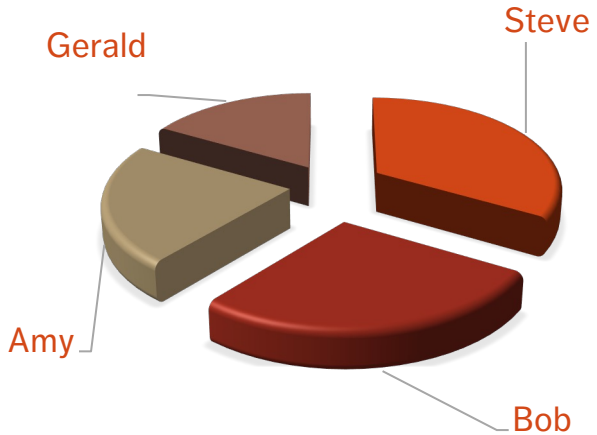
Even when done well, 85% of users don't bother with interactive viz (NY Times).

Take-Away: explore the data and try different methods

CHARTS TO AVOID

ANYTHING with an arc (except gauge)

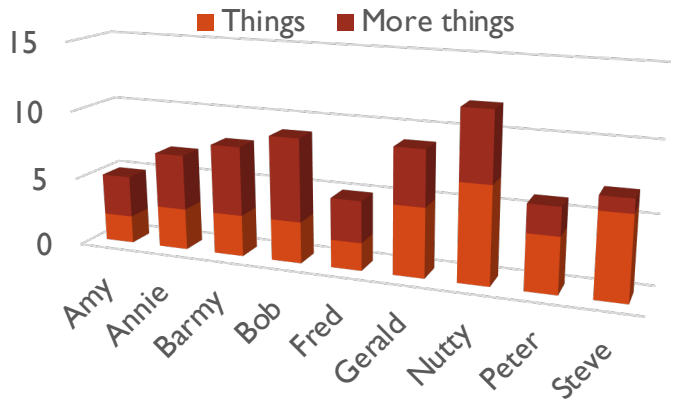
- pie
- donut



Brain cannot compare arcs and they can be misleading: how different are Steve & Bob in the pie chart?

ALL 3D IS EVIL!

- as with arc, we cannot easily visually compare data series
- adds way too much clutter



[www.interaction-design.org/literature/topics/gestalt-principles]

GESTALT PRINCIPLES

The **Gestalt principles** are the “laws” of human perception.

They describe how humans group similar elements, recognize patterns and simplify complex images when they perceive objects.

Designers use them to organize content on charts, dashboards, websites, and other interfaces so that they be **aesthetically pleasing/easy to understand**.

[www.interaction-design.org/literature/topics/gestalt-principles]

GESTALT PRINCIPLES

"Gestalt" is German for "unified whole".

The first principles were devised in the 1920s by German psychologists Wertheimer, Koffka ("the whole is greater than the sum of the parts"), Kohler.

Aim: understand how we gain meaning from the chaotic stimuli around us.

The Gestalt principles are a set of "laws" which address the natural compulsion to find order in disorder. According to this, the mind "informs" what the eye sees by **perceiving a series of individual elements as a whole**.

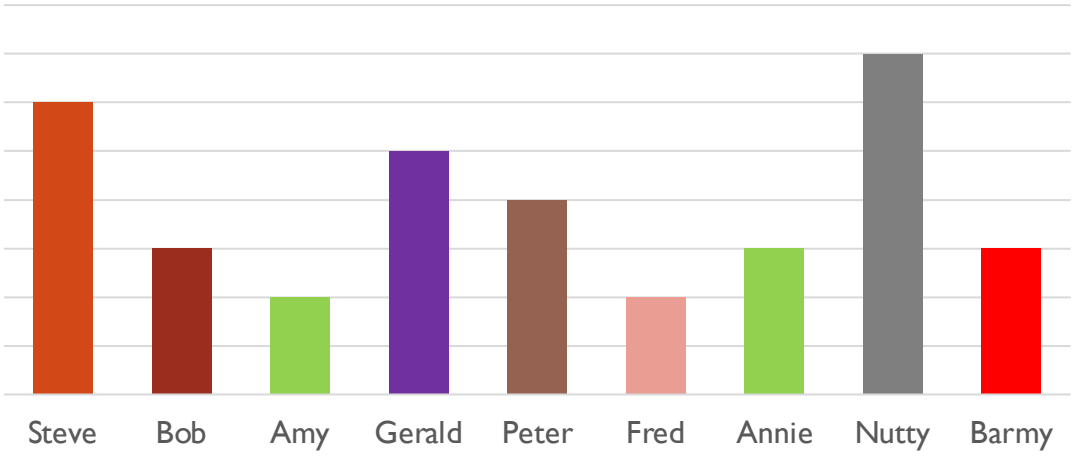
GESTALT PRINCIPLES

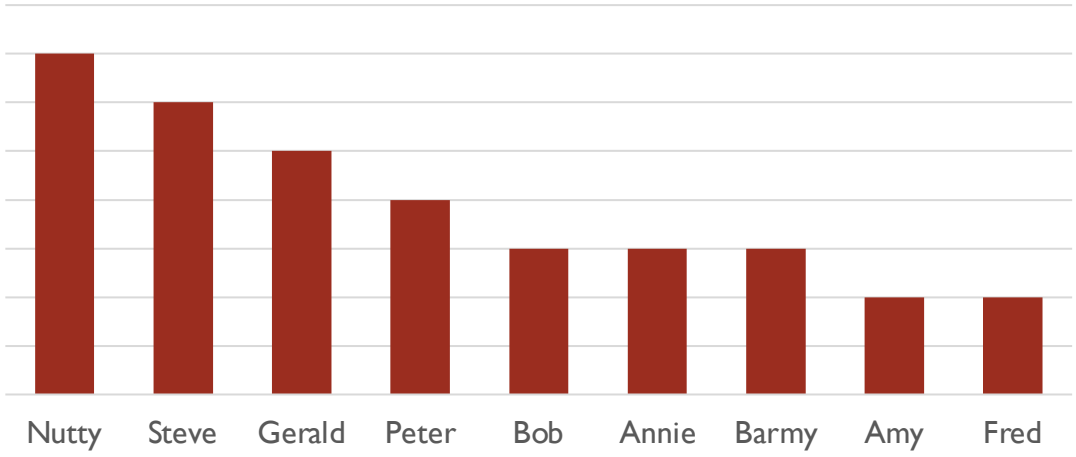
- **simplicity**
- continuation
- **proximity**
- **similarity (invariance)**
- **focal point**
- isomorphic correspondence
- **figure / ground duality**
- common fate
- closure
- uniform connectedness

GESTALT PRINCIPLES – SIMPLICITY

The brain has a preference for **simplicity** – it tends to process simple patterns faster than patterns that are more complex.

Lesson: arrange data simply and logically wherever possible.





GESTALT PRINCIPLES – PROXIMITY

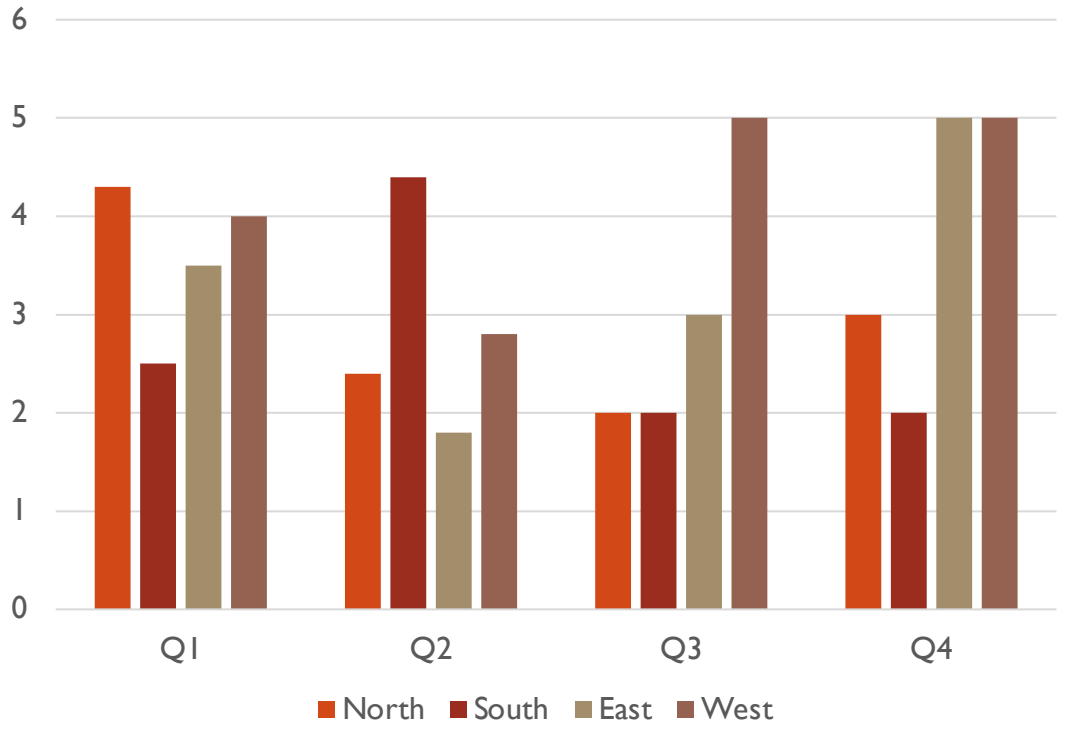
Objects/shapes that are in **proximity** (close) appear to form **groups**.

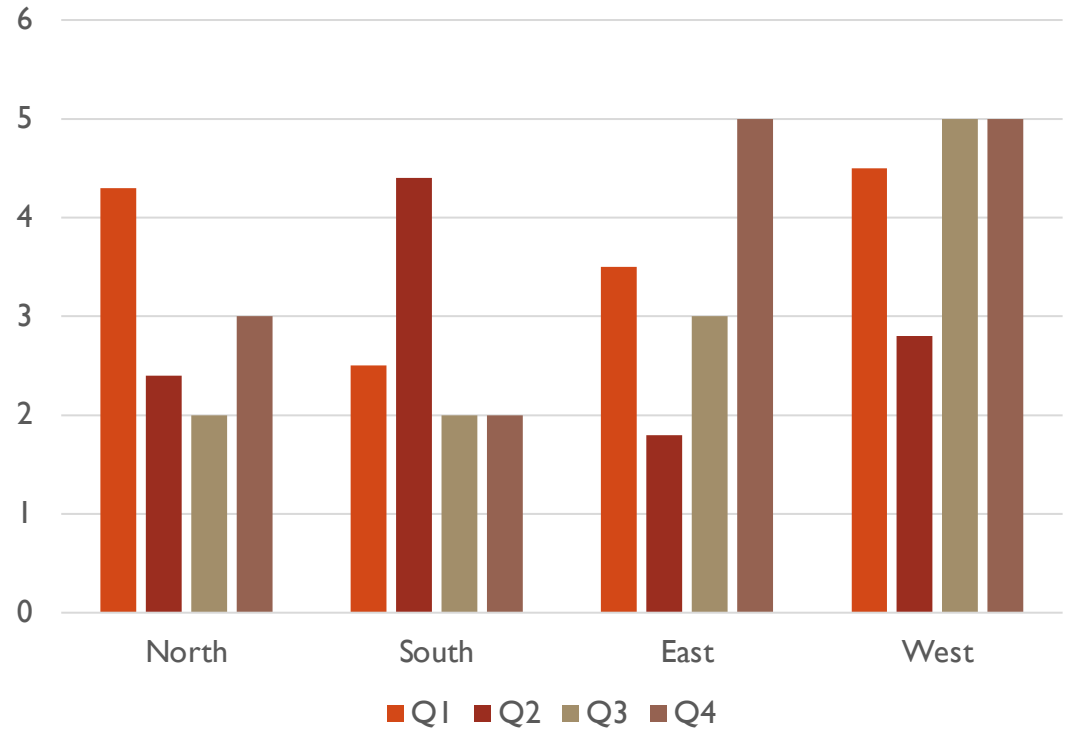
The effect generated by the collected group is more “powerful” than that generated by separate elements.

Elements which are grouped together create the **illusion** of shapes/planes in space, even if the elements are not touching.

Lesson: understand the chart’s priorities and create groupings through proximity that support those priorities.

GESTALT PRINCIPLES – PROXIMITY





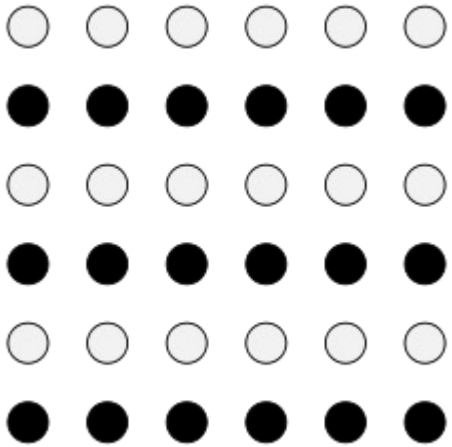
GESTALT PRINCIPLES – SIMILARITY

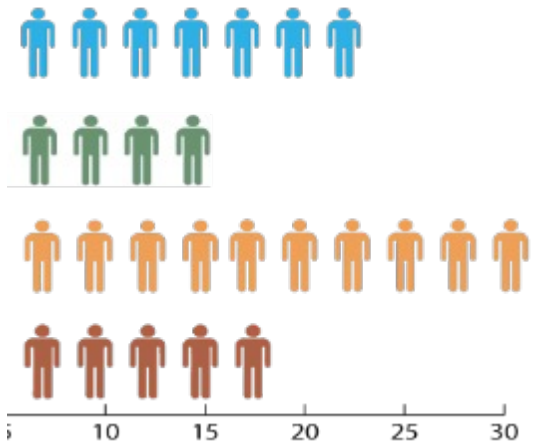
Stimuli that physically resemble each other are viewed as **part of the same object**; stimuli that don't are viewed as part of a different object.

Similarity and proximity often come together to form a **visual hierarchy**. Either principle can dominate the other, depending on their application and combination.

Lesson: use similar characteristics to establish relationships and to encourage groupings of objects.

GESTALT PRINCIPLES – SIMILARITY





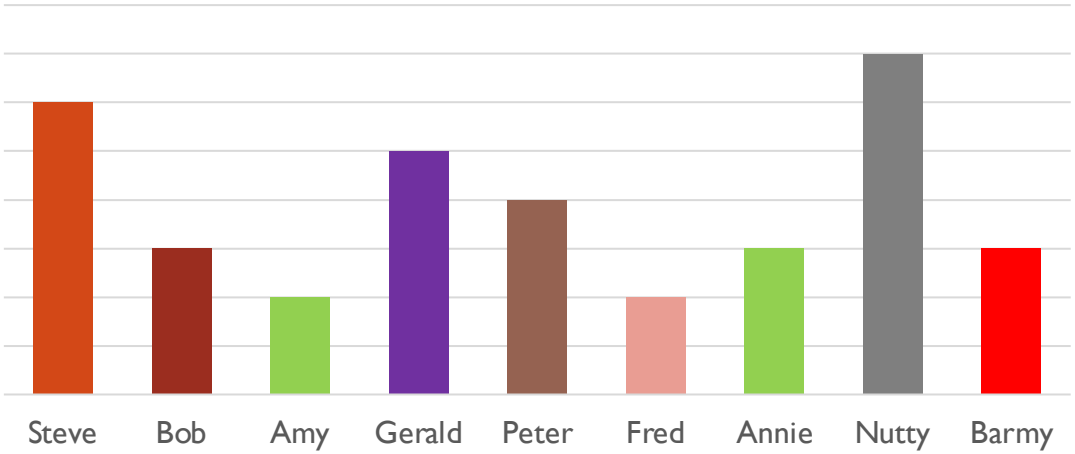
In these examples, similarity dominates over proximity: we see rows before we see columns.

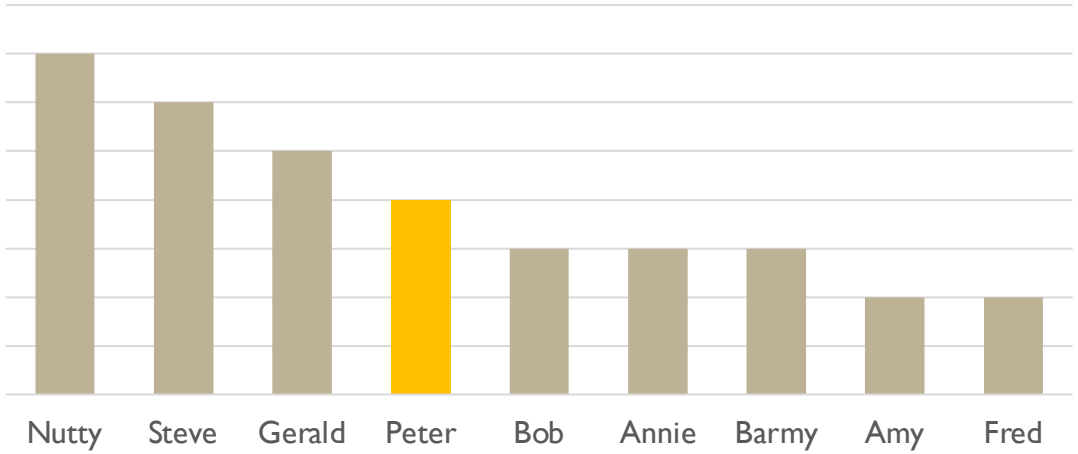
GESTALT PRINCIPLES – FOCAL POINT

In opposition to similarity, the **focal point** principle states that distinctive-looking objects can create a focal point.

To highlight one salesperson's performance, make their bar graph color different.

Lesson: use different characteristics to highlight and create focal points.





GESTALT PRINCIPLES – DUALITY

Chart elements are either perceived as **figures** (focus) or as (back)**ground**.

Foreground objects are **promoted** by the brain, background objects are **demoted**.

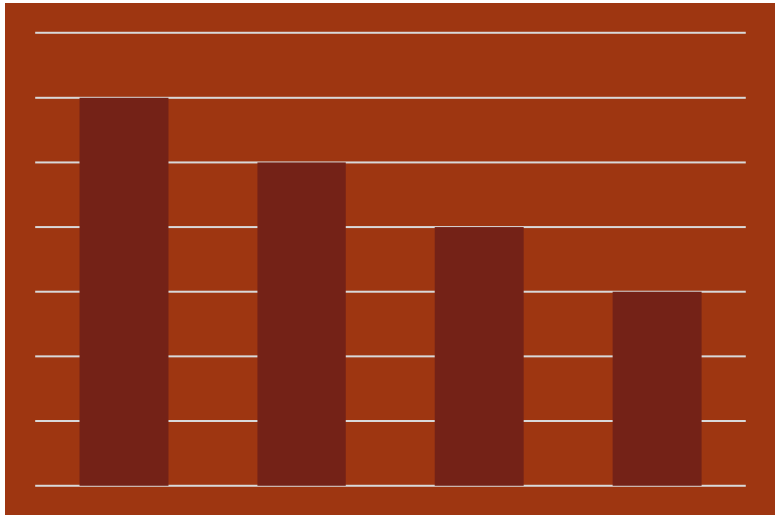
Strong contrast makes it easier to distinguish between the two types of objects.

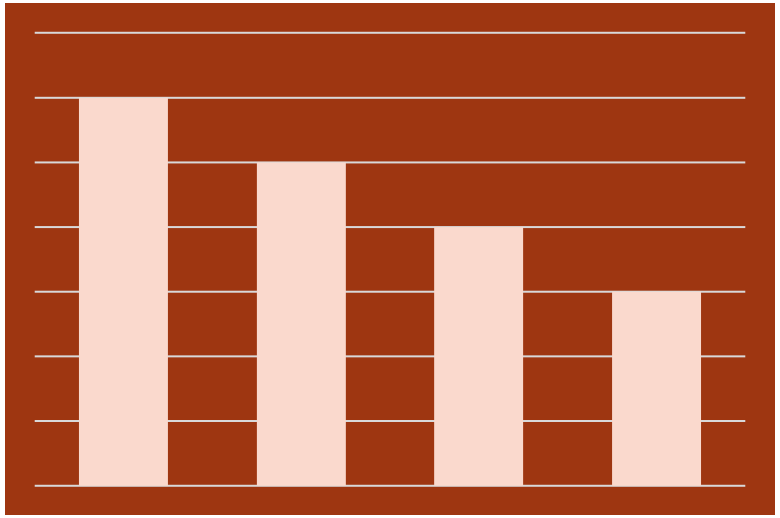
Lesson: ensure there is enough contrast between the chart foreground (figures) and their background.

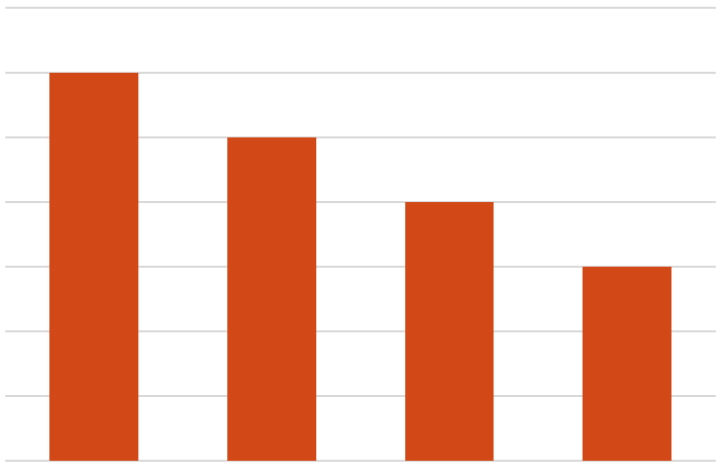
GESTALT PRINCIPLES – DUALITY

Because of the **low contrast** between the figure and background in the chart on the left, there is an **additional cognitive load**.

Increasing the contrast on the right improves readability.







DECLUTTERING

Clutter is the enemy!

Every element on a page adds cognitive load

- identify and remove anything that isn't adding value
- think of cognitive load as mental effort required to process information (lower is better)

Tufte refers to the **data-to-ink ratio** – "the larger the share of a graphic's ink devoted to data, the better"

In *Resonate*, Duarte refers to this as "**maximizing the signal-to-noise ratio**" where the signal is the information or the story we want to communicate.

DECLUTTERING

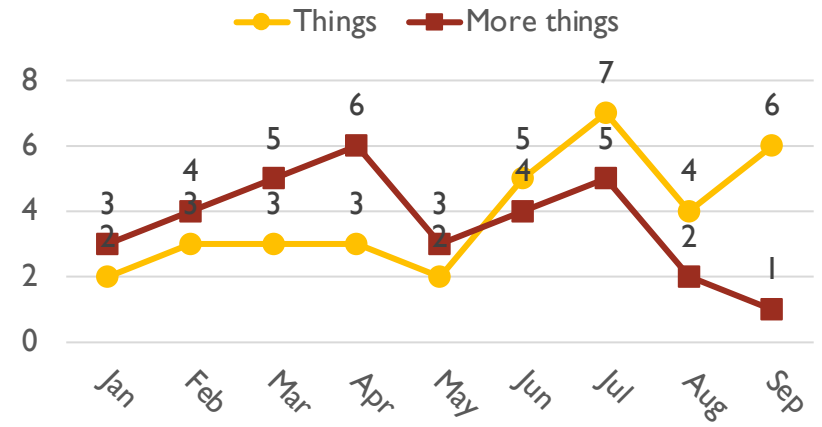
Use the Gestalt Principles to **organize/highlight** data in the chart.

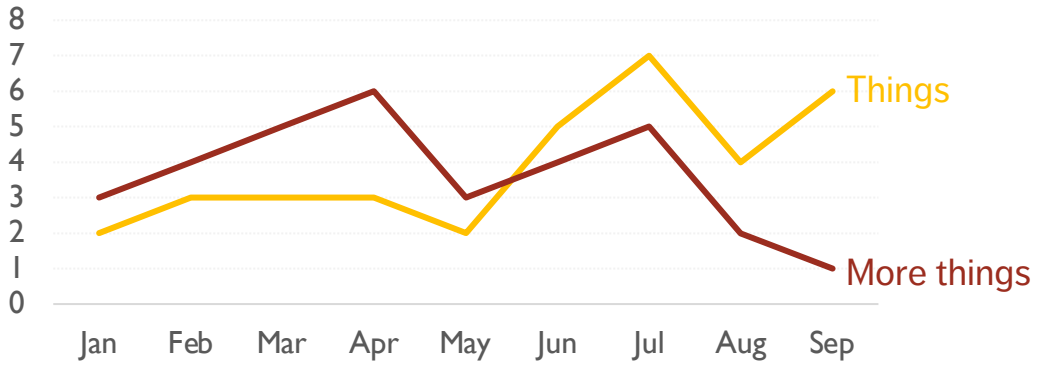
Align all elements (graphs, text, lines, etc.):

- don't rely on eye, use position boxes and values

Charts:

- remove border, gridlines, data markers
- clean up axis labels
- label data directly





[L.C. Muth, https://blog.datawrapper.de/fonts-for-data-visualization]

DECLUTTERING

Use **consistent** fonts, font size, colour and alignment.

Don't rotate text to anything other than 0 or 90 degrees (however: English/French incompatibility with vertical text).

Use **white space**:

- margins should remain free of text and visuals
- don't stretch visuals to edge of page or too close to other visuals
- think of white space as a border

COLOUR SCHEMES

Achromatic (colourless, using only blacks, whites and grays)

Monochromatic (1-colour schemes)

Complementary (colours directly across from each other on the colour wheel)

Split complementary (2 of the 3 colors are adjacent; 1 of the colours is opposite)

COLOUR TIPS

When it comes to colour, **less is more**: use it sparingly (graphic designers are taught to “get it right, in black and white”).

Based on the Gestalt Principles, **monochrome** schemes can be effective.

When appropriate, pick corporate identity scheme (this maximizes buy in).

Create a **template** (and stick to it).

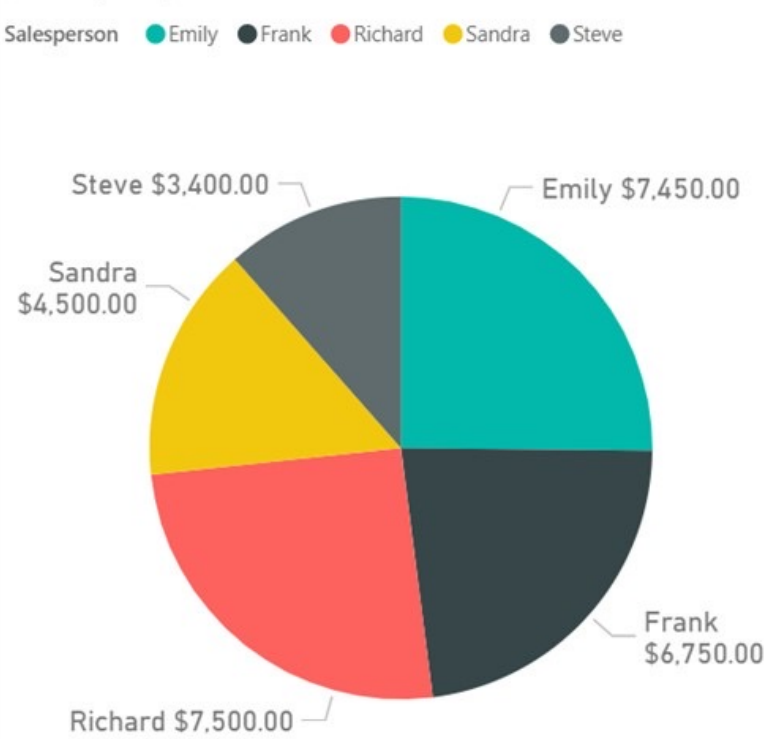
Upload images to see what charts look like for flavours of colourblindness:

- https://www.color-blindness.com/coblis-color-blindness-simulator (not the only tool)

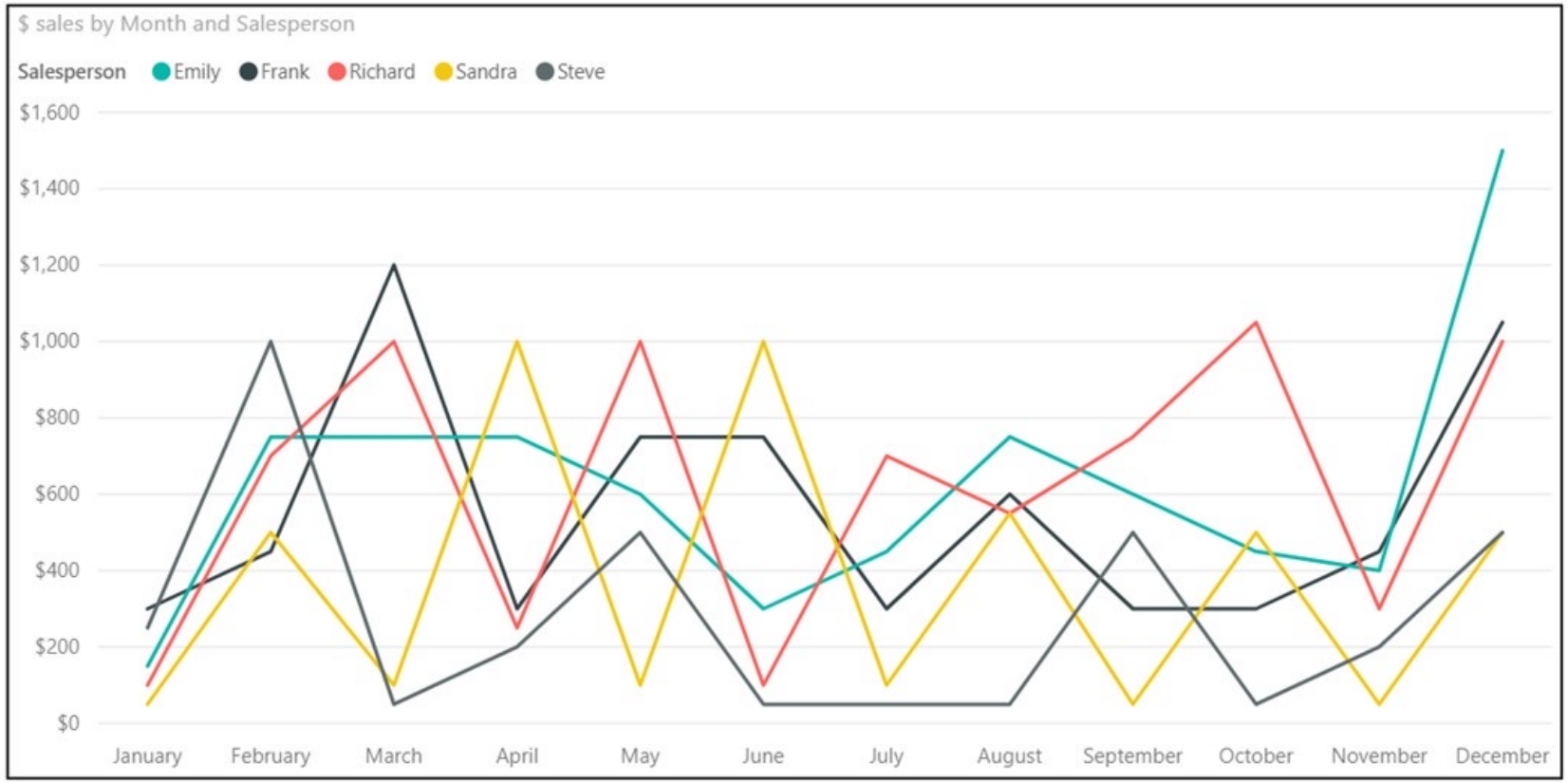
Sales Dashboard

$ sales

$ sales by Salesperson



$ sales by Month and Salesperson



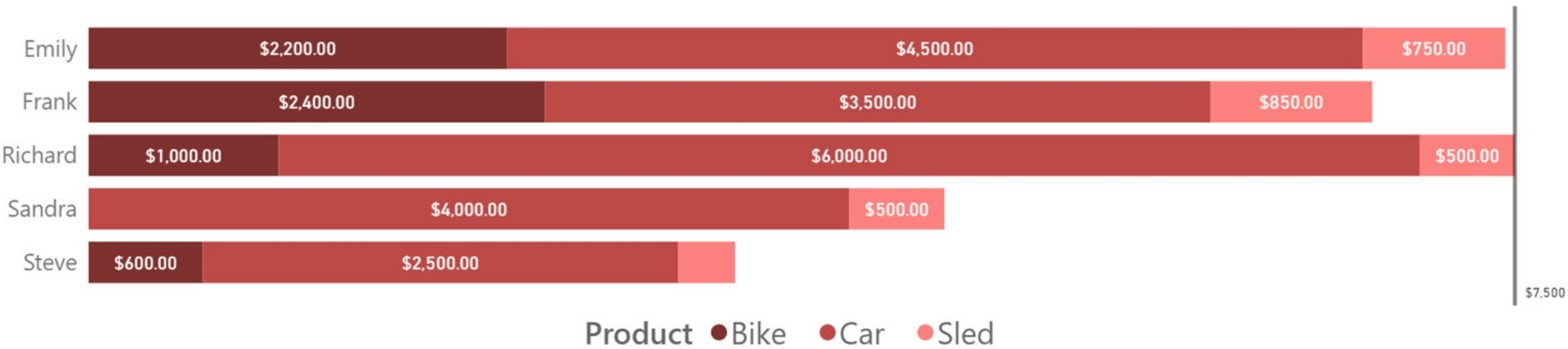
$ sales by Product and Salesperson

Sales Dashboard

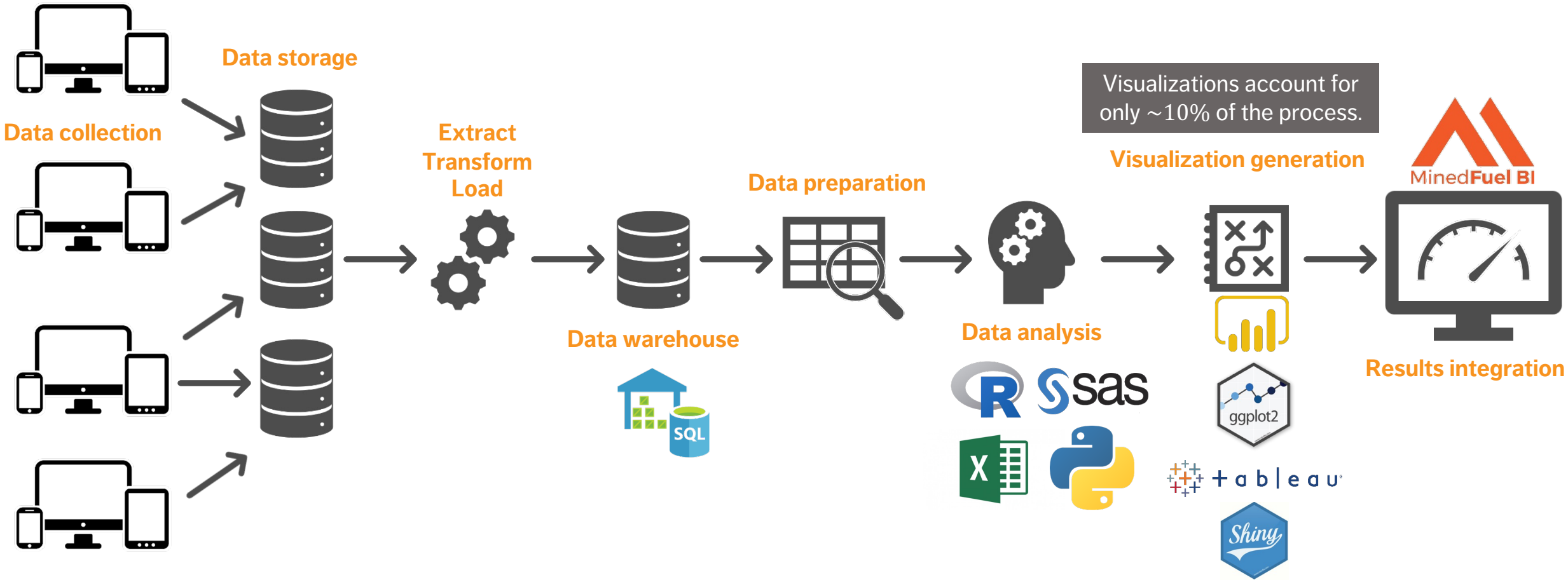
Annual Sales for 2017

Total Sales

$29.6K



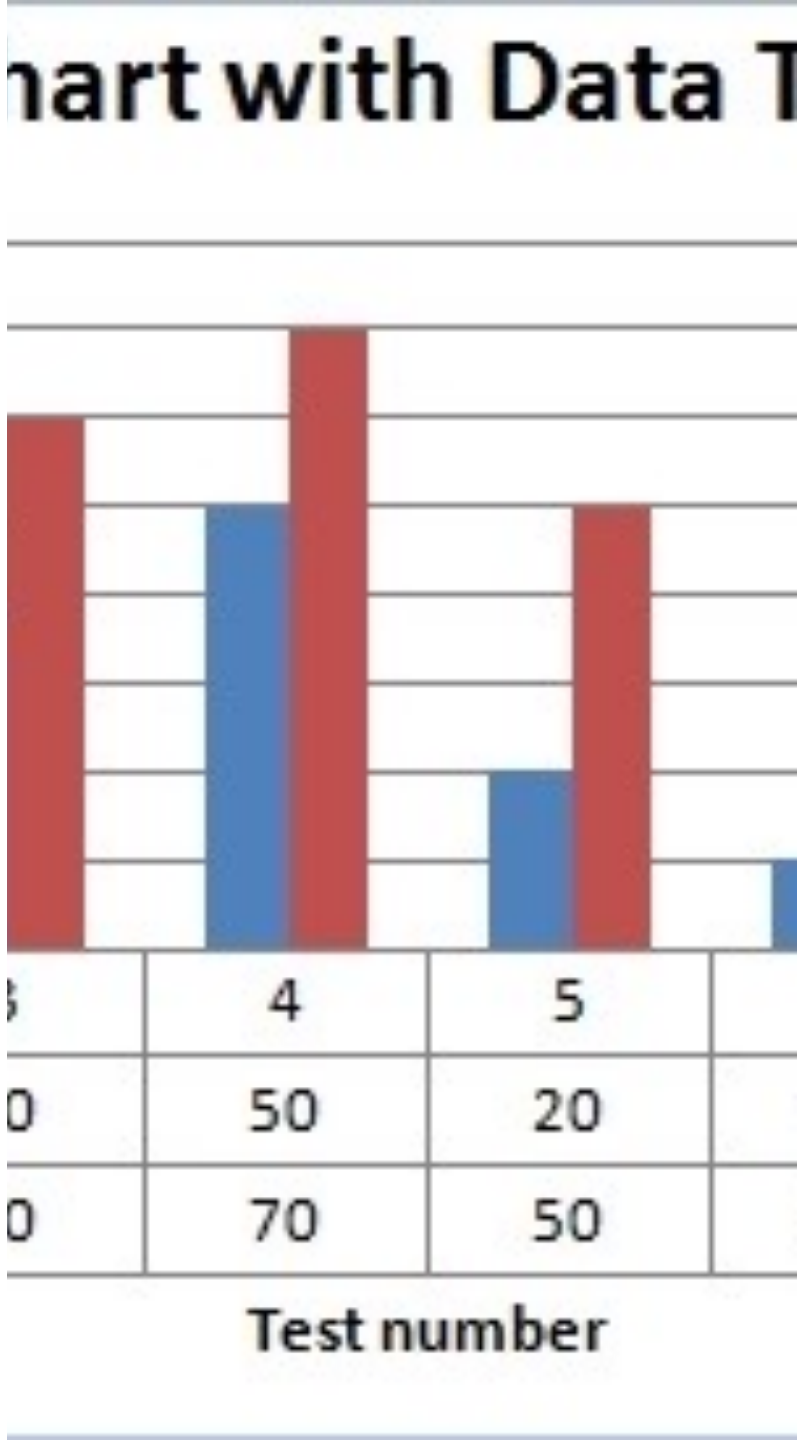
DATA ENVIRONMENT



OVERVIEW

The past is **data-driven**:

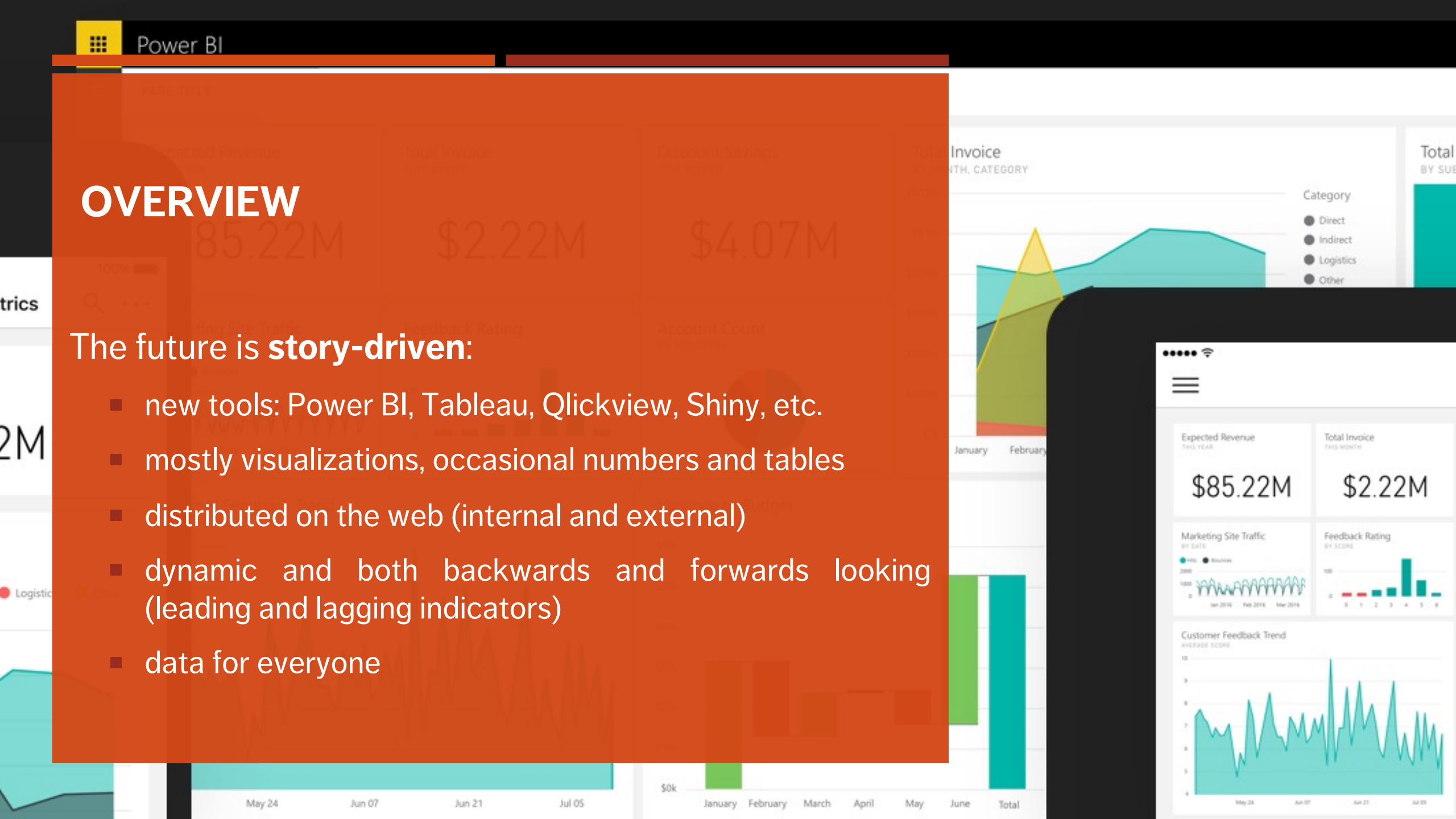
- mostly Excel (or reporting tools like Cognos)
- mostly numbers, tables and non-interactive graphs
- distributed on desktop computers, by email, in PowerPoint presentation
- static, mostly backwards looking (lagging indicators)
- KPIs and dashboards were somewhat contrived



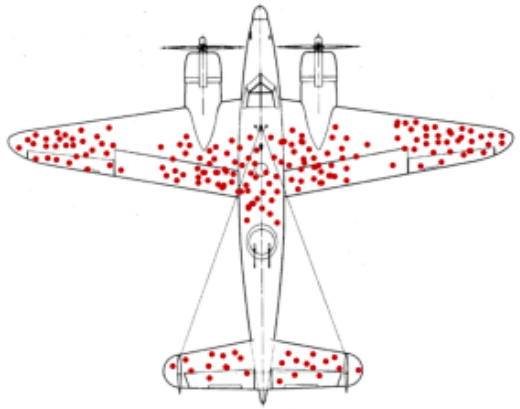
OVERVIEW

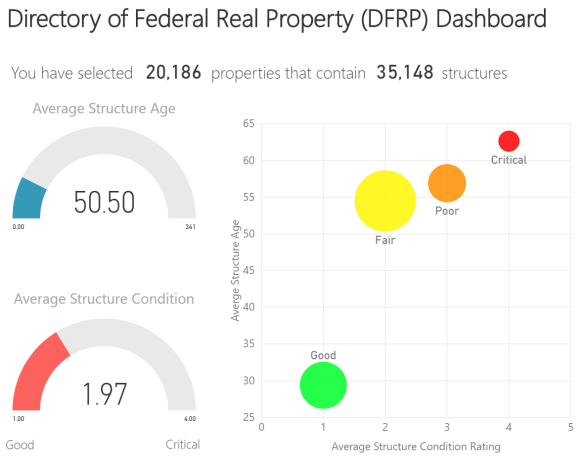
The future is **story-driven**:

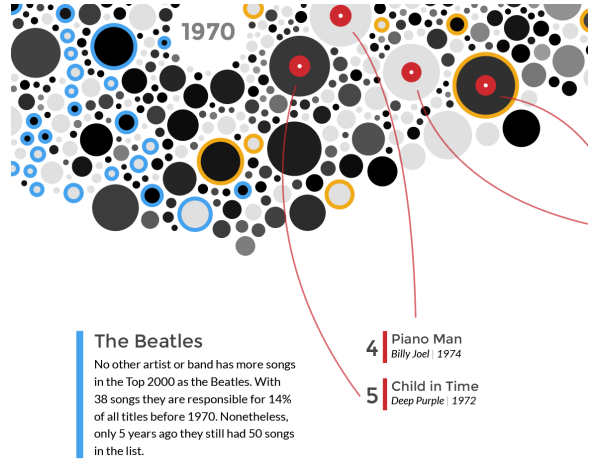
- new tools: Power BI, Tableau, Qlickview, Shiny, etc.
- mostly visualizations, occasional numbers and tables
- distributed on the web (internal and external)
- dynamic and both backwards and forwards looking (leading and lagging indicators)
- data for everyone



DEFINING CONTEXT







Seconds | **Minutes** | **Fraction of Hour** | **Hours**

← **Infographics/Data Viz** →

← **Dashboards** →

← **Reports and Exploration** →

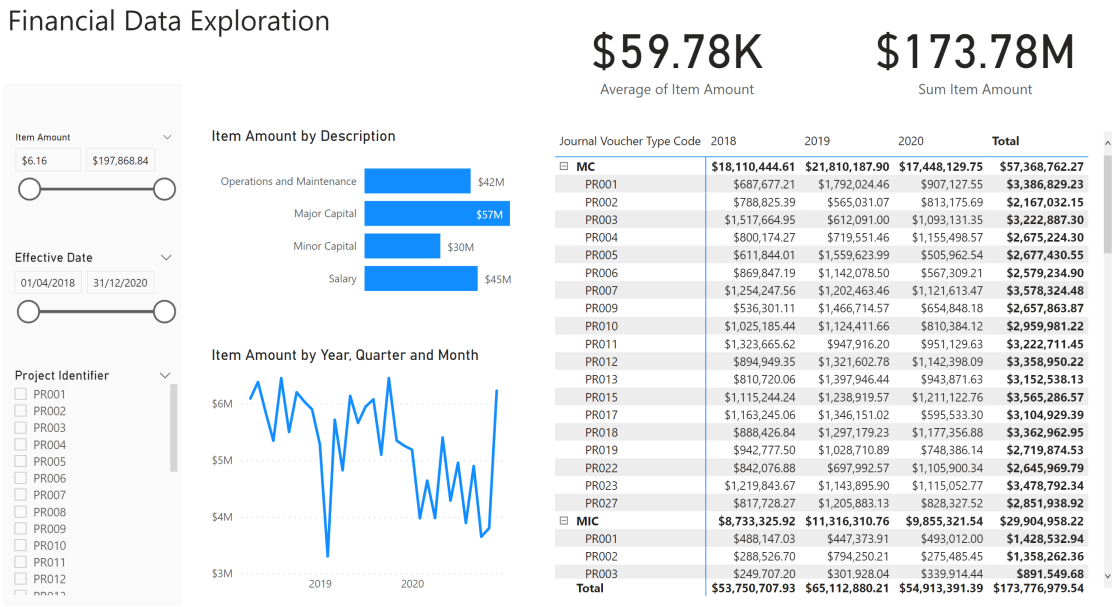
← **Data Art** →

data-action-lab.com

DASHBOARD TYPES

Exploration: using visualizations as a tool to explore/understand the data

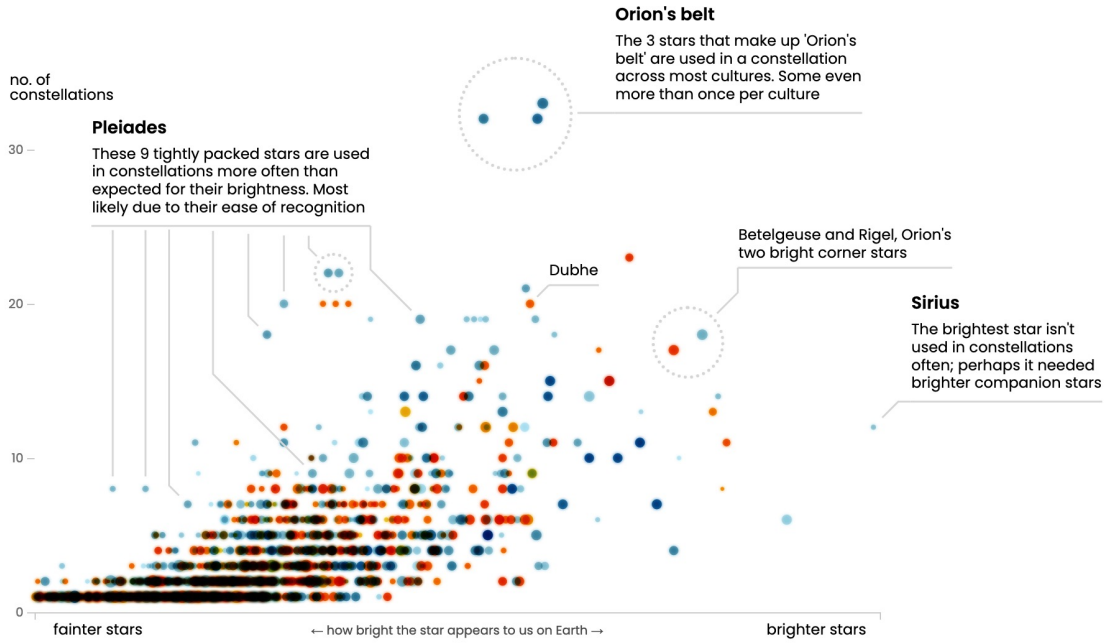
- high level of interactivity
- high level of detail
- **all** aspects of data should be represented (tables, columns, calculations etc.)
- no annotations or explanations required



DASHBOARD TYPES

Storybook: using visualizations as a tool to explain the data and communicate the story

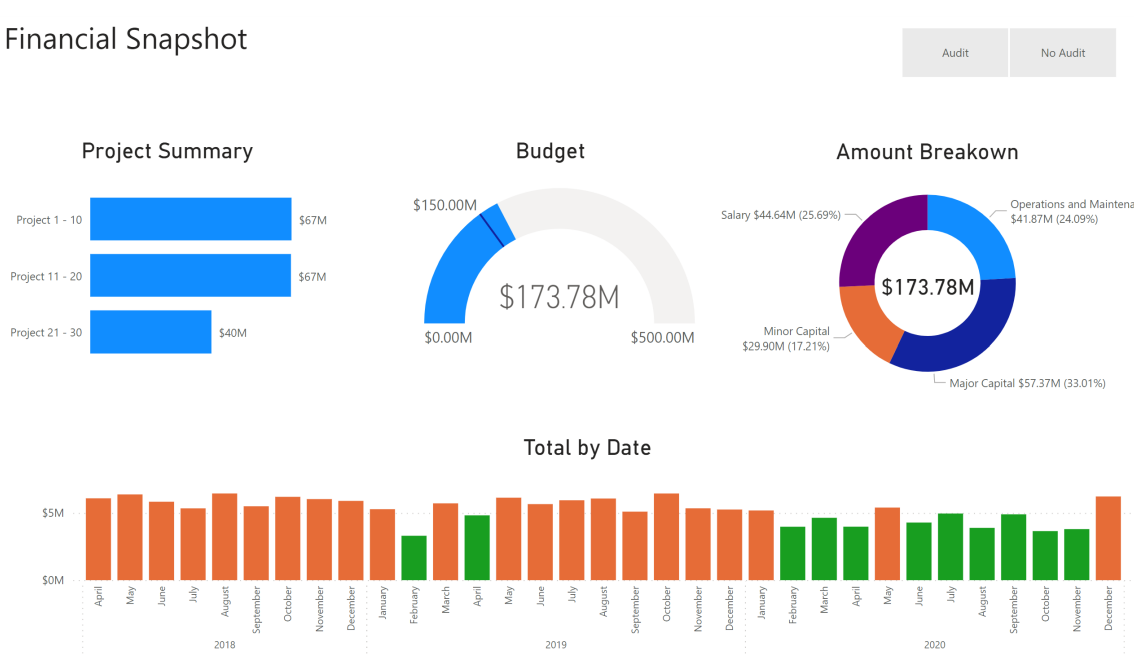
- low level of interactivity
- low level of detail
- key aspects of data should be represented
- annotations and explanations drive the “story”



DASHBOARD TYPES

Situational Awareness: using visualizations as a tool to provide a snapshot of the data

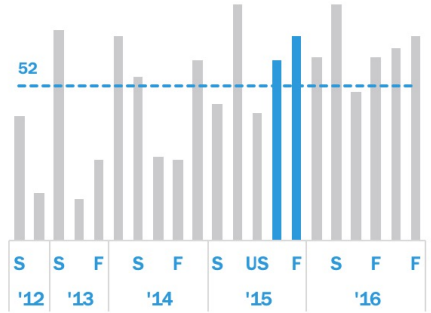
- medium level of interactivity
- not “scripted” but well organized (e.g., categorized)
- summary data should be represented
- anomalies are highlighted
- often used for internal presentations



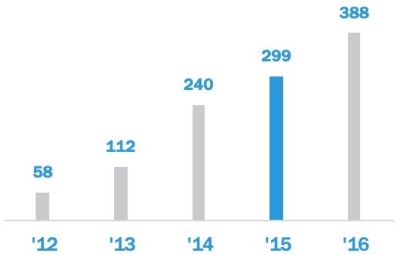
Course Metrics

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

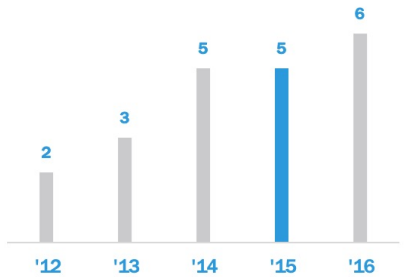
Students



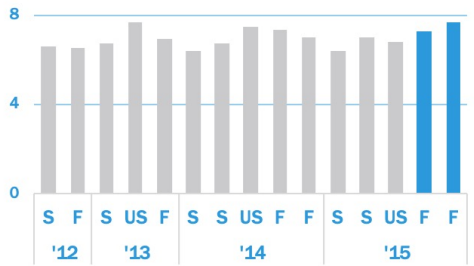
Enrollments



Classes



Ratings



1097

Total Students in five years

687

Total Students in 2015-2016

21

Total Classes in five years

7.7 of **8**

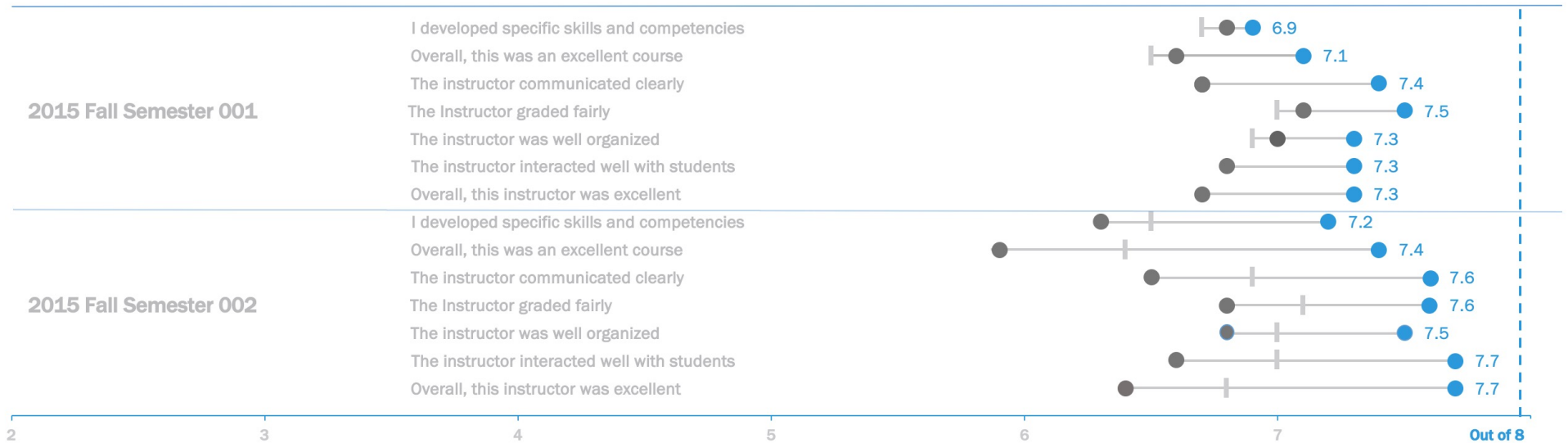
Most recent instructor rating (out of 8.0)

Semesters

Questions

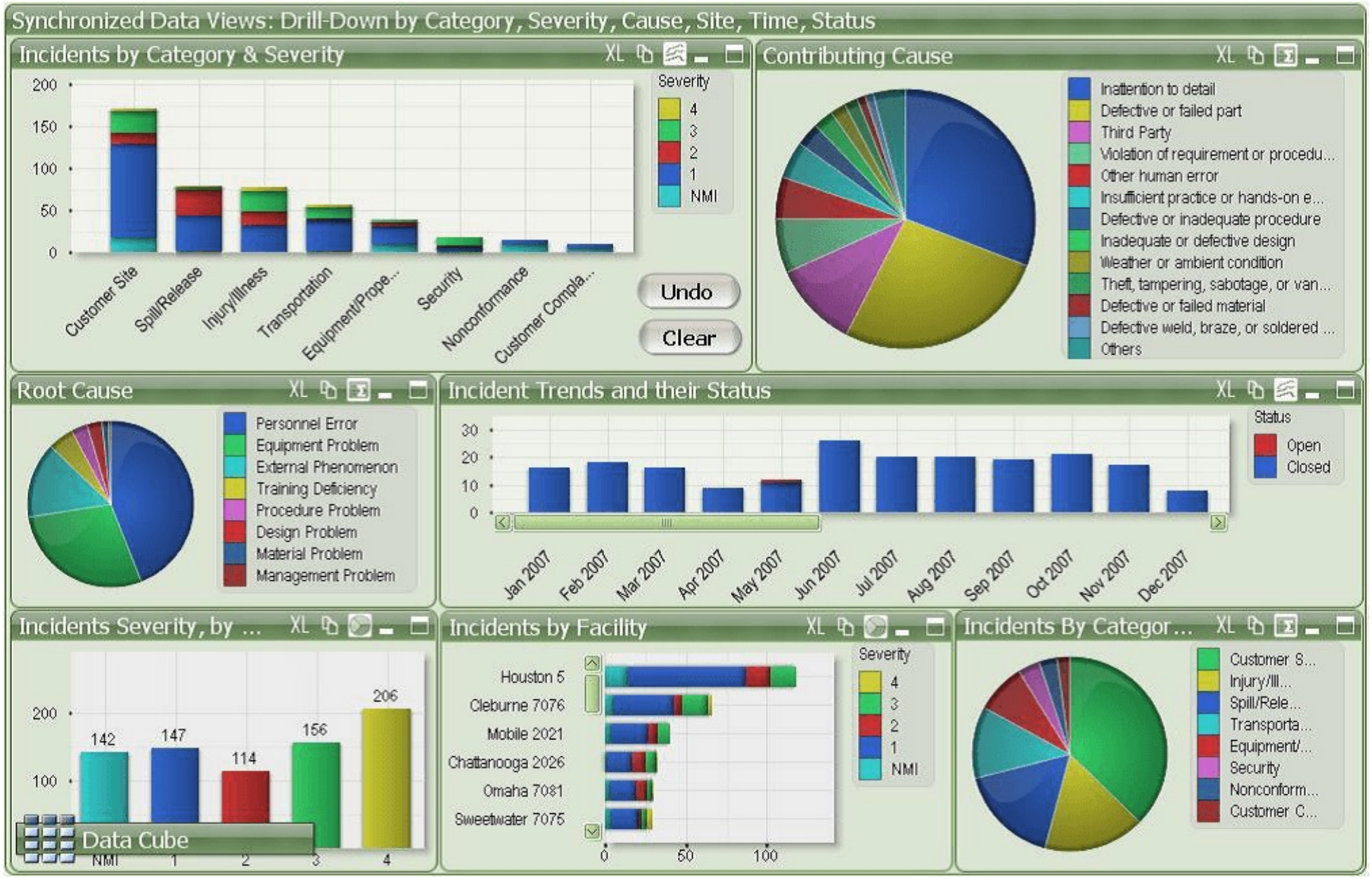
BANA College Shaffer

Ratings



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

[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]

PRACTICAL DEFINITION OF A STORY

To paraphrase U.S. judge Potter Stewart: “I may not be able to define what a story is, but I know one when I see one”.

We could say that a **story** consists of:

- context,
- series of events, and
- outcome, result, consequence, or resolution.

STORYTELLING GOALS

Cultural Stories

- entertain, inform, teach, explore, shock

Data (Scientific) Stories

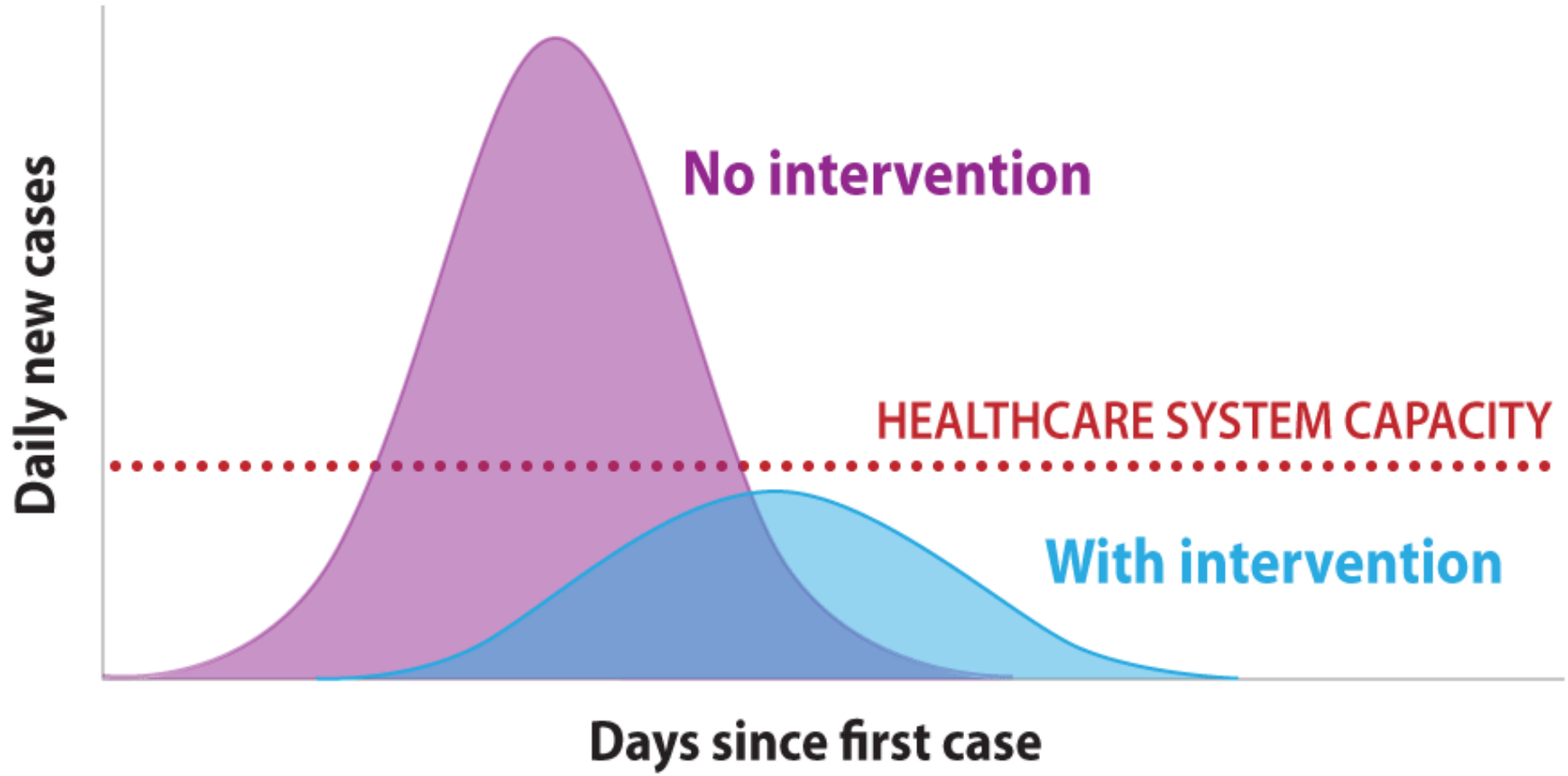
- describe, diagnose, predict, prescribe, persuade

Any overlap?

Anything missing?

FLATTENING THE CURVE

A look at the importance of slowing the spread of a virus, so that the rate of infection doesn't outpace the resources to fight against it.



SOURCE: CDC

THE CANADIAN PRESS

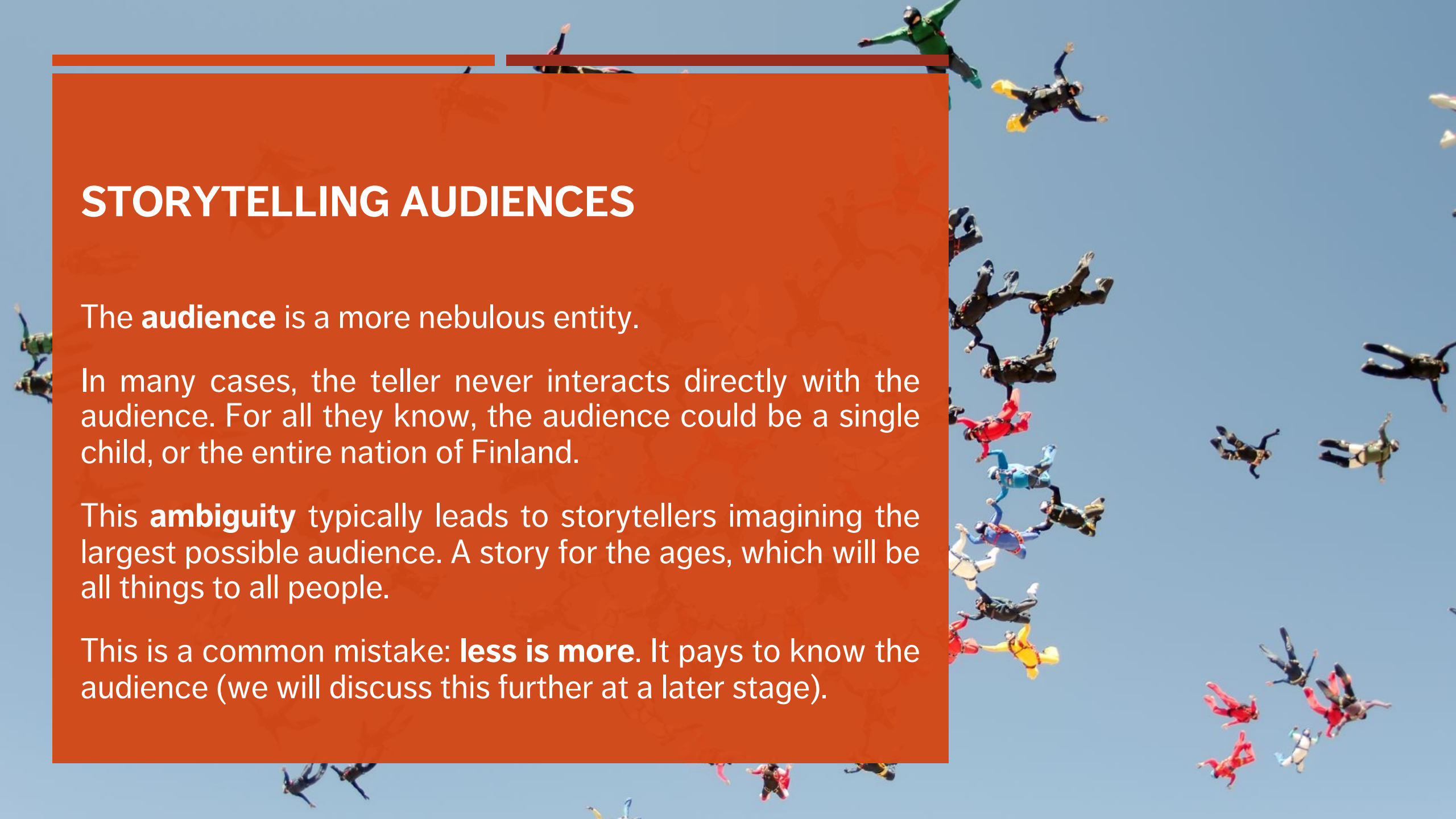
STORYTELLING AUDIENCES

Storytelling requires a **teller** and a **story**, but also an **audience.**

The **teller**'s job is to convince the audience to accept:

1. the premise ("I'm about to tell you a really interesting story, so listen up!")
2. the contents ("All these things happened, honest!")
3. the conclusion ("And that's why you should never put peanut butter in your laundry.")

The **story**'s must first and foremost not come in the way of the teller's job.



STORYTELLING AUDIENCES

The **audience** is a more nebulous entity.

In many cases, the teller never interacts directly with the audience. For all they know, the audience could be a single child, or the entire nation of Finland.

This **ambiguity** typically leads to storytellers imagining the largest possible audience. A story for the ages, which will be all things to all people.

This is a common mistake: **less is more**. It pays to know the audience (we will discuss this further at a later stage).

STORYTELLING CONTEXT

A given action may be seen as positive or as negative by audiences with different pre-existing feelings/knowledge concerning the agent/situation.

- Would you be able to recognize nobility in a political enemy's actions?
- Could a fan of the Maple Leafs/Habs ever have something worthy to say about hockey?

Similarly, a story may have different **outcomes/impacts** in different contexts.

Trevor Greenway · Feb 9 · 5 min read

[https://www.lowdownonline.com/post/wakefield-nurse-fires-up-freedom-convoy]

Wakefield nurse fires up Freedom Convoy

Wakefield's Bethan Nodwell is known in the Gatineau Hills for many things: being the hospital's former head nurse, singing onstage at the Black Sheep Inn, and more recently, disseminating debatable facts and anti-vax sentiments on social media. Now she's running the main stage at the Freedom Convoy in downtown Ottawa, firing up the crowd as seen here Feb. 4. Trevor Greenway photo

Bethan Nodwell had thousands of demonstrators in Ottawa hanging onto her every word.

What might lead one to view the **subject** of this article in a positive light?

A negative light? A neutral light?

What might lead one to view the **author** of this article in a positive light?

A negative light? A neutral light?

[Illustrations: V. Jaskina]

STORYTELLING UNIVERSALITY

There once was a shepherd boy who was bored as he sat on the hillside watching the village sheep. To amuse himself he took a great breath and sang out, "Wolf! Wolf! The Wolf is chasing the sheep!"

The villagers came running up the hill to help the boy drive the wolf away. But when they arrived at the top of the hill, they found no wolf. The boy laughed at the sight of their angry faces. "Don't cry 'wolf', shepherd boy," said the villagers, "when there's no wolf!" They went grumbling back down the hill.

Later, the boy sang out again, "Wolf! Wolf! The wolf is chasing the sheep!" To his naughty delight, he watched the villagers run up the hill to help him drive the wolf away.

When the villagers saw no wolf they sternly said, "Save your frightened song for when there is really something wrong! Don't cry 'wolf' when there is NO wolf!"

[Illustrations: V. Jaskina]

STORYTELLING UNIVERSALITY

But the boy just grinned and watched them go grumbling down the hill once more.

Later, he saw a REAL wolf prowling about his flock. Alarmed, he leaped to his feet and sang out as loudly as he could, "Wolf! Wolf!" But the villagers thought he was trying to fool them again, and so they didn't come.

At sunset, everyone wondered why the shepherd boy hadn't returned to the village with their sheep. They went up the hill to find the boy. They found him weeping.

"There really was a wolf here! The flock has scattered! I cried out, "Wolf!" Why didn't you come?"

An old man tried to comfort the boy as they walked back to the village. "We'll help you look for the lost sheep in the morning," he said, putting his arm around the youth, "Nobody believes a liar ... **even when they are telling the truth**/**so don't get caught telling the same lie twice**."

[online.hbs.edu/blog/post/data-storytelling, C. Cote]

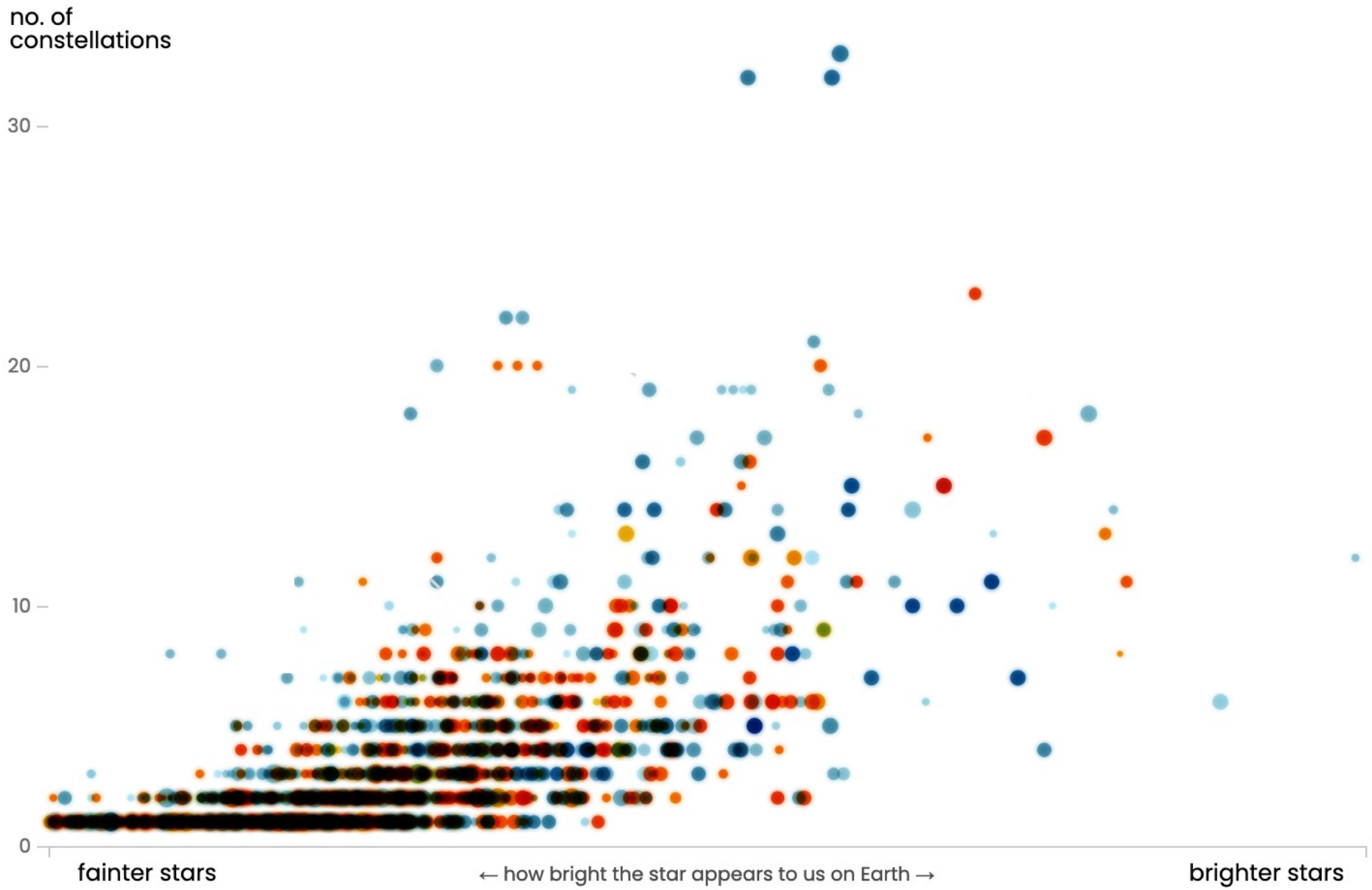
DATA STORIES

Data storytelling is the ability to effectively communicate insights from a dataset using narratives and visualizations. It can be used to put data insights into context for and inspire action from the audience.

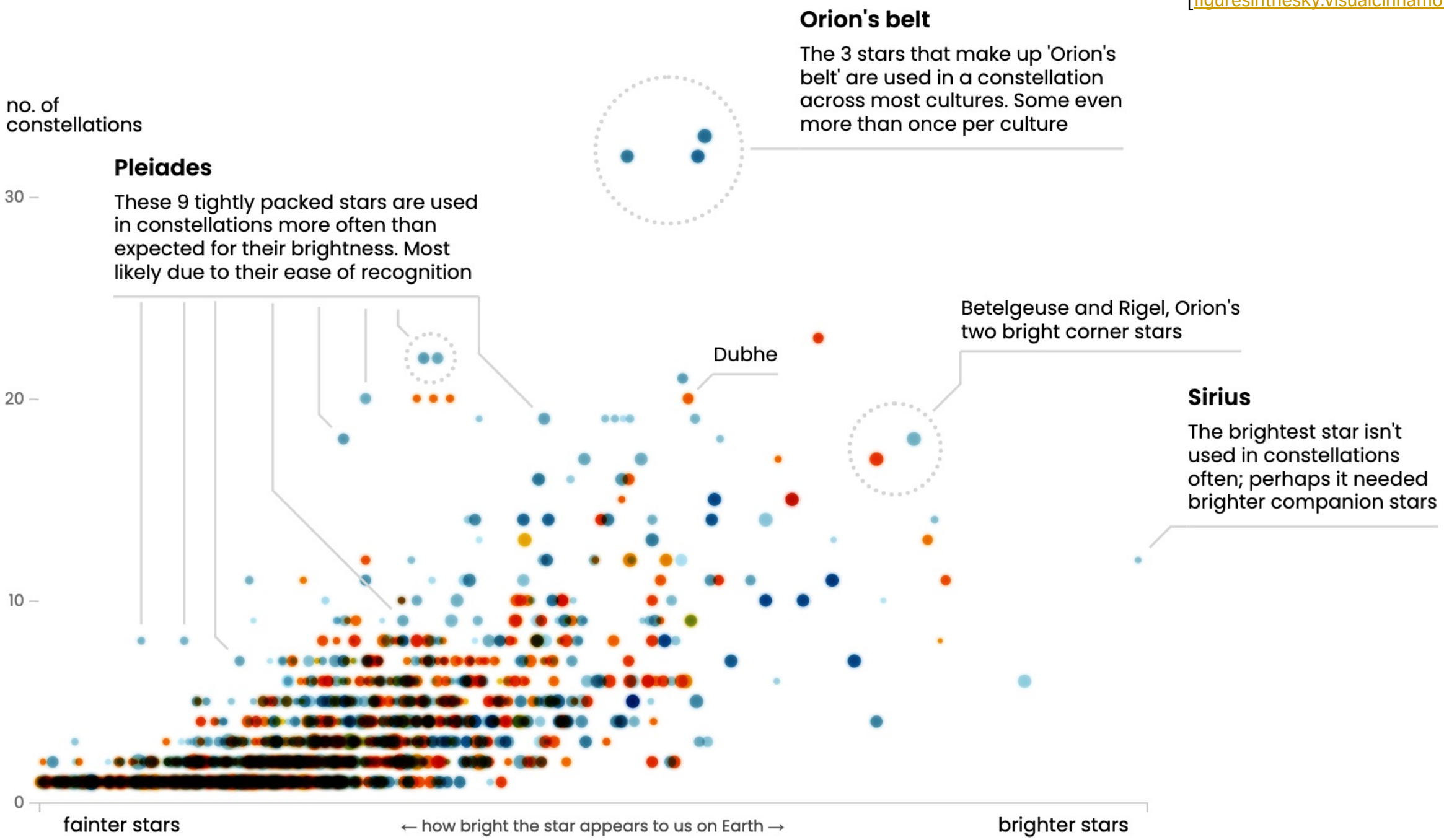
There are 3 key components:

1. **data:** foundation of data story (descriptive, diagnostic, predictive, prescriptive analysis)
2. **narrative:** storyline used to communicate the insights gleaned from data and context, and recommended actions
3. **visuals:** representations of data, analysis results, and narratives, which are used to communicate stories clearly and memorably (charts, graphs, diagrams, pictures, or videos)

[figuresinthesky.visualcinnamon.com]



[figuresinthesky.visualcinnamon.com]



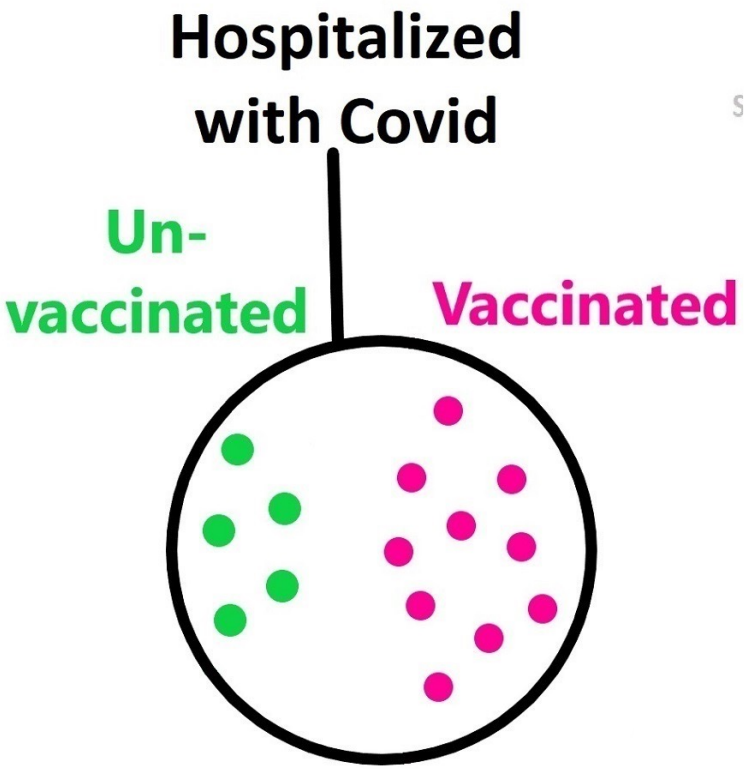
STORYTELLING RISKS

A good story can help shed insights on a situation, but storytelling requires **choices**; the outcome is affected by what is **included** and what is **omitted**.

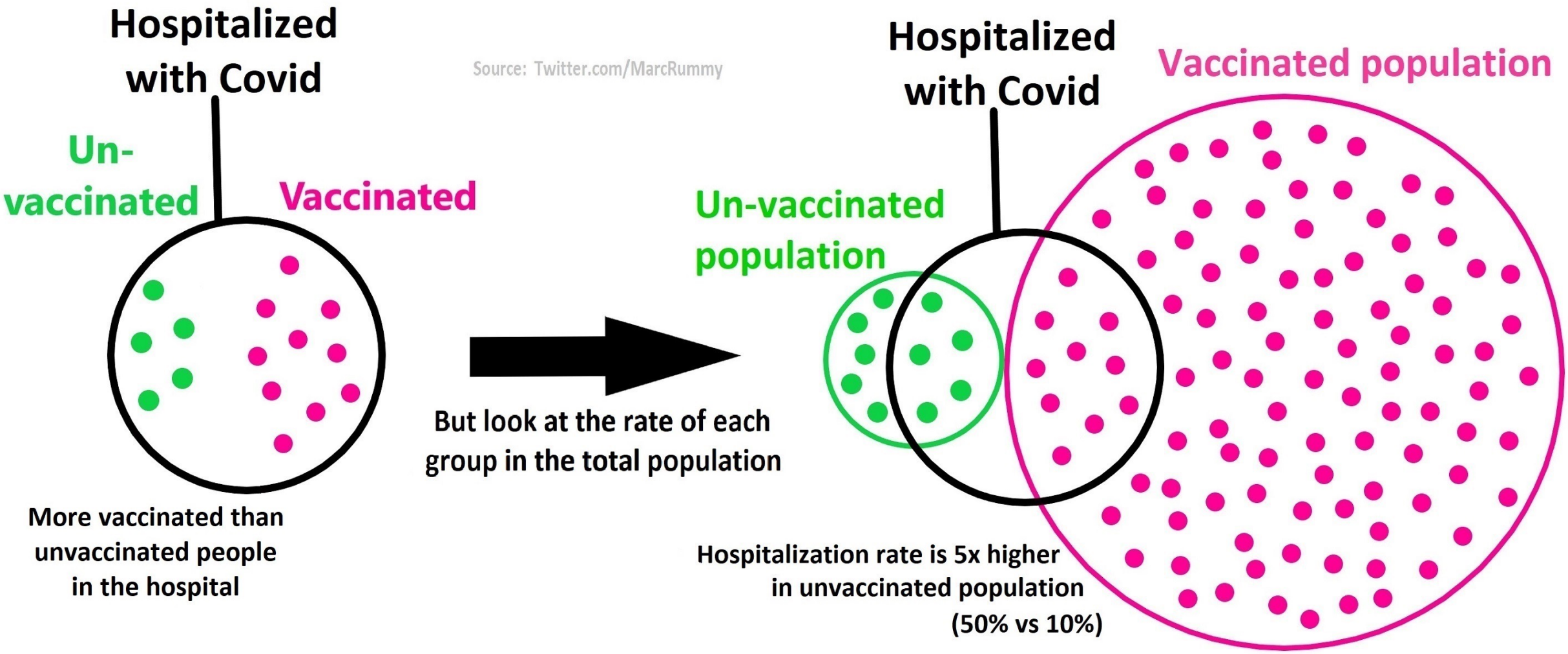
It is easy to mislead by **accident**; it is also easy to mislead by **design**.

With data stories, there is an additional complication: we usually only have access to the **available data**. The data that was not collected is, by definition, not available. Some of the data that was collected may also be unavailable for a variety of reasons.

This implicit bias can lead to compelling (yet **flawed**) data stories.



More vaccinated than unvaccinated people in the hospital



Note: The ratios presented are made to illustrate the concept of the base rate fallacy when the vaccination rate is high

[based on an idea from S. McCloud, *Making Comics*]

WORDS AND IMAGES

A picture is worth a thousand words (vs. a picture is worth 1000 words).

Words bring an unparalleled level of **specificity**. There is no image so vague that words cannot lock it into a **desired meaning**.

Some concepts and names can only be clearly expressed **through words**.

[S. McCloud, *Making Comics*]

"Look, it's Kelly Donovan, twin brother of the Xander actor on *Buffy the Vampire Slayer*, plus Humphrey Bogart wearing a Freddy Mercury mask, and a robot duplicate of former U.N. Secretary General Boutros Boutros-Ghali!"

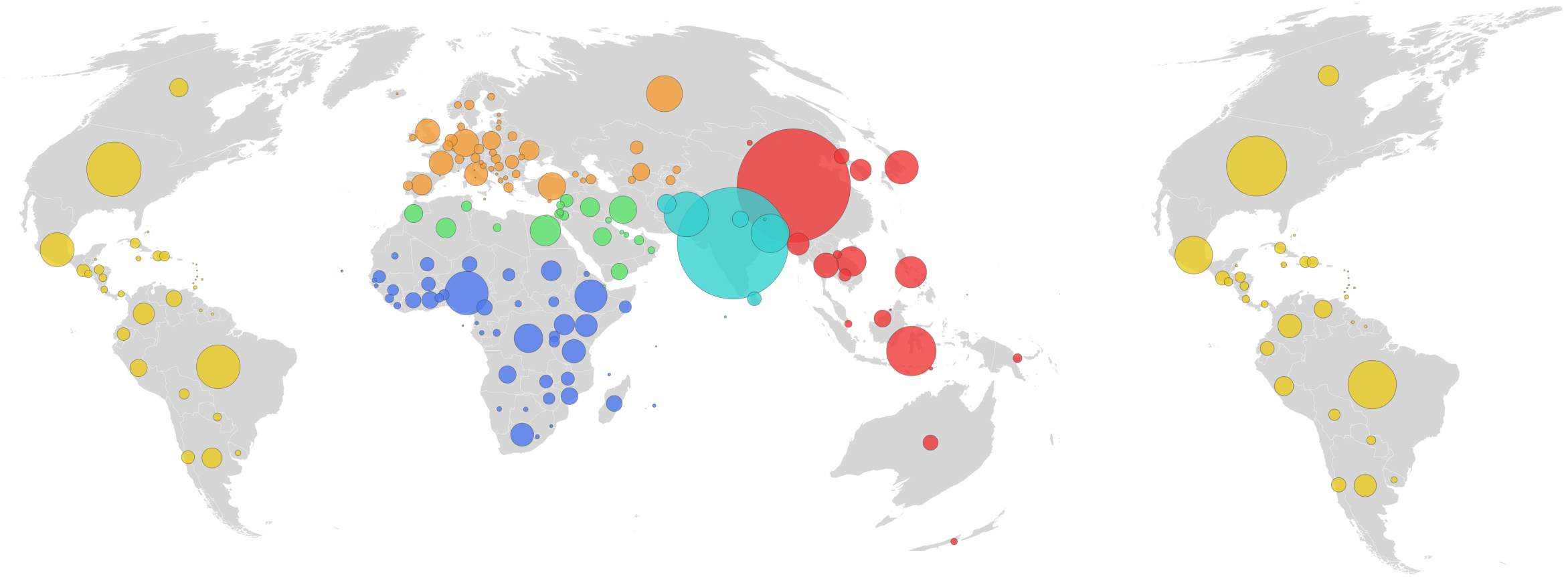
[based on an idea from S. McCloud, *Making Comics*]

VISUAL STORYTELLING CHOICES

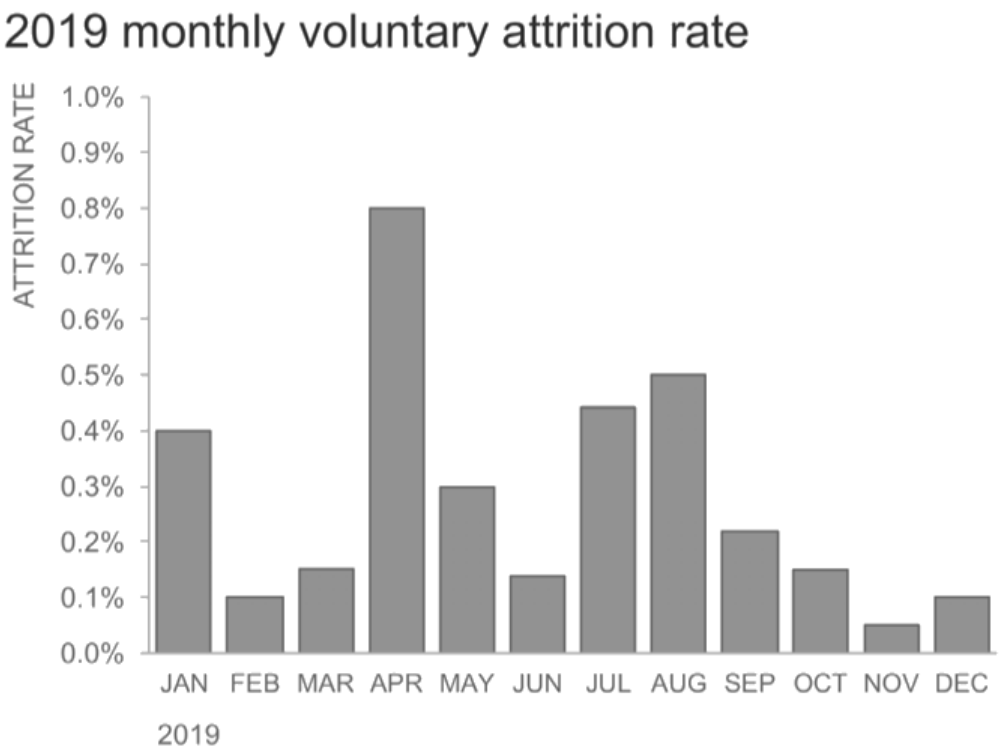
Communicating with **clarity** means that audience comprehension remains the **ultimate goal:**

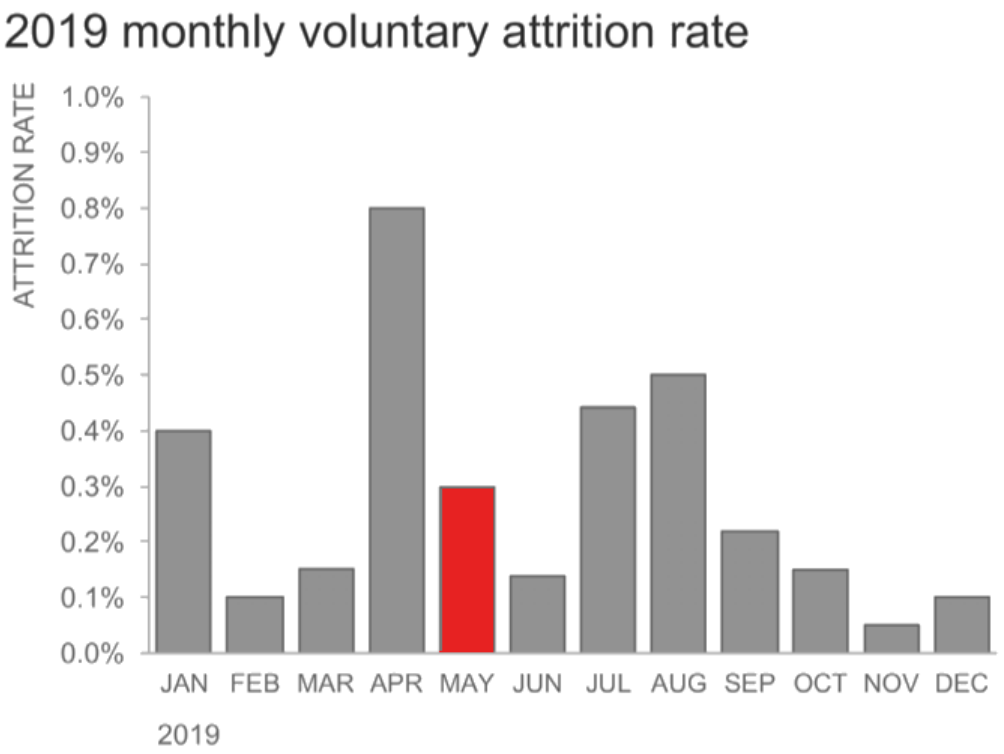
- choice of **moment** is 'connecting the dots', showing only what matters to the story;
- choice of **frame** is creating and directing the audience's focus;
- choice of **image** is selecting the right charts for the story, with emphasis on simplicity and ability to convey the message;
- choice of **word** is clearly and persuasively communicating ideas in seamless combination with the charts;
- choice of **flow** is guiding the audience from one chart to the next, from one page to the next, and creating a transparent and intuitive 'reading' experience, by arranging pages in a dashboard, charts on a page, and elements within charts intelligently.

CHOICE OF MOMENT

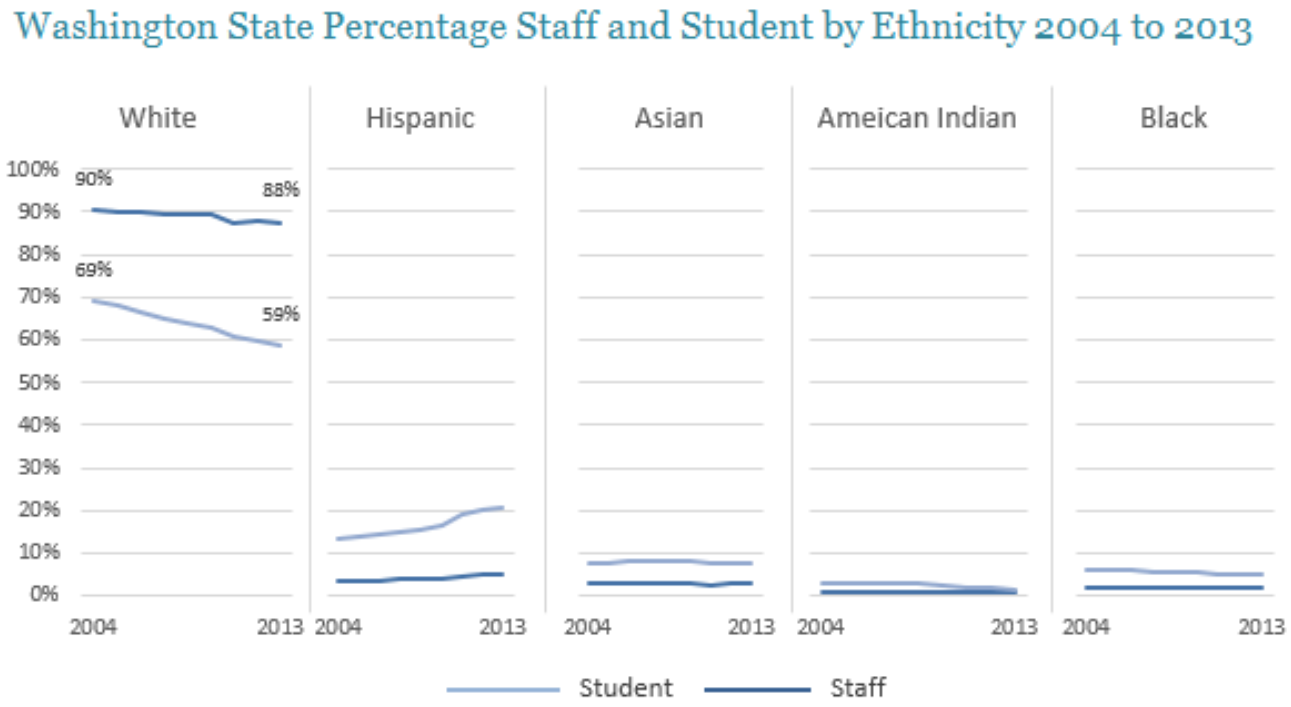


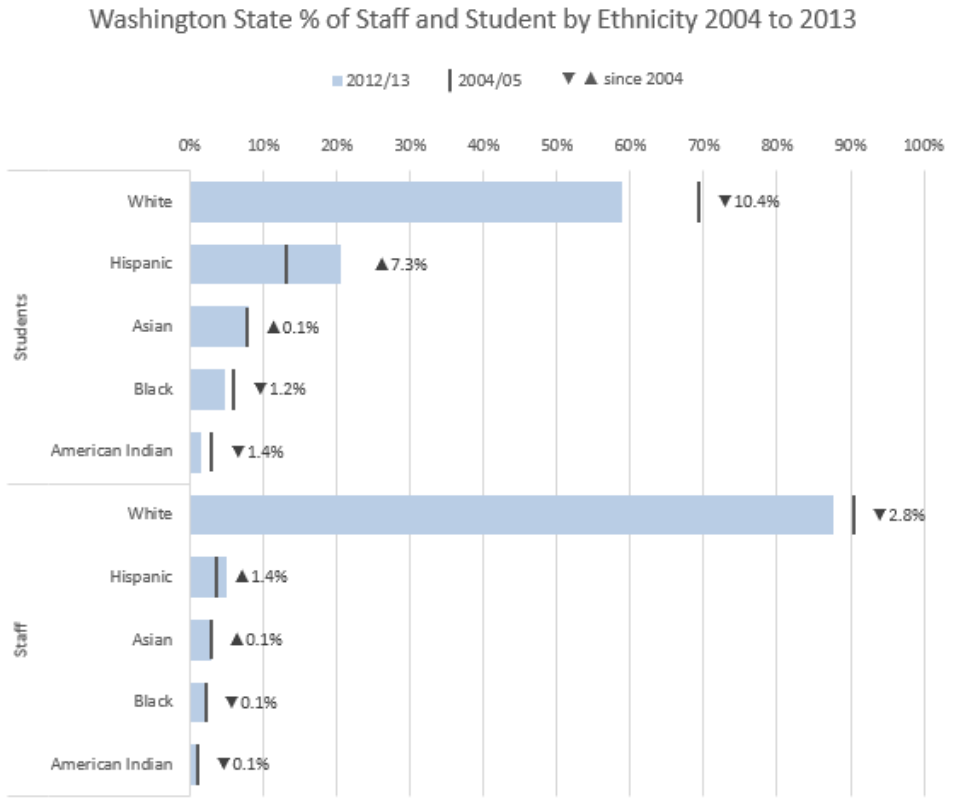
CHOICE OF FRAME



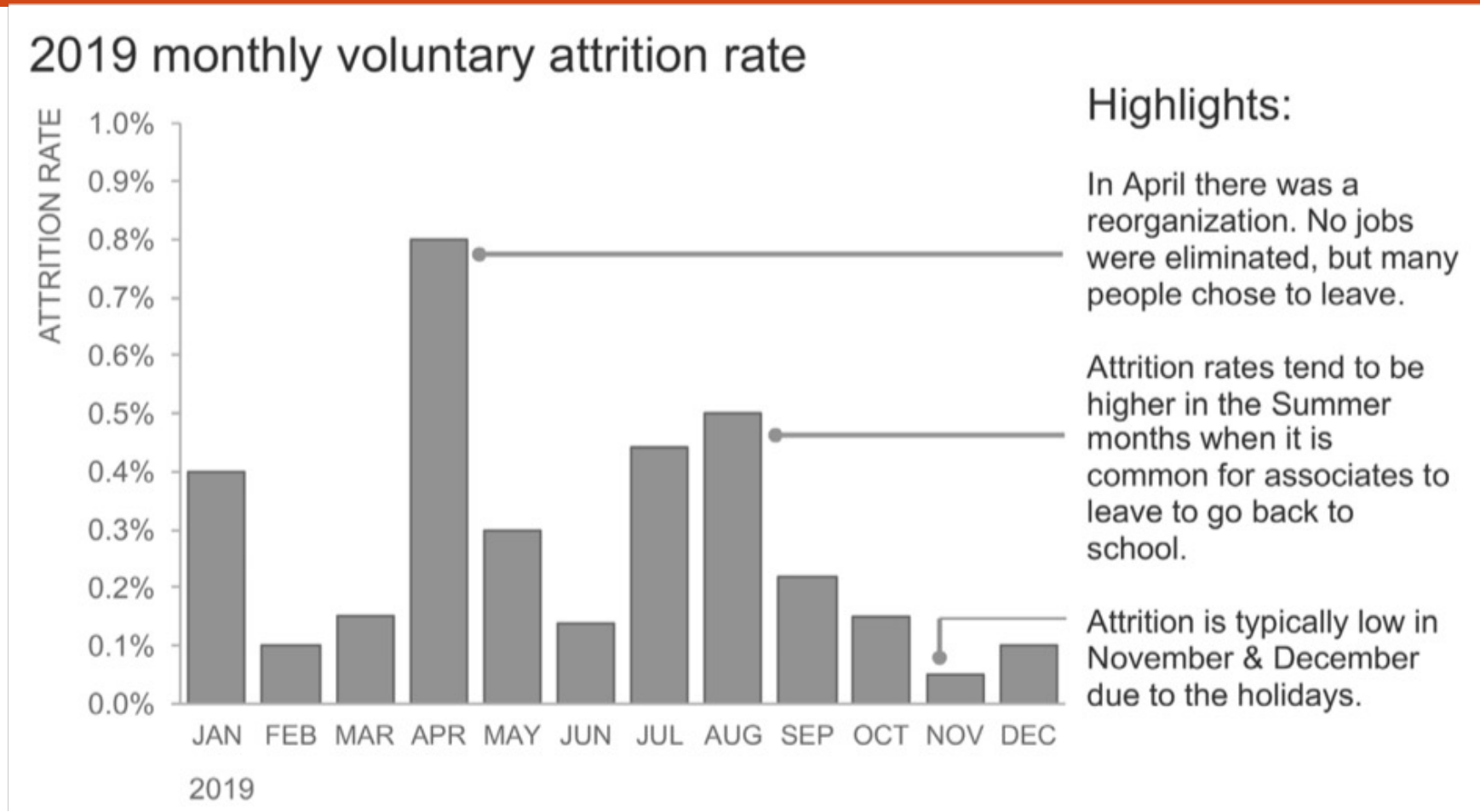


CHOICE OF IMAGE





CHOICE OF WORD



[based on an idea from S. McCloud, *Making Comics*]

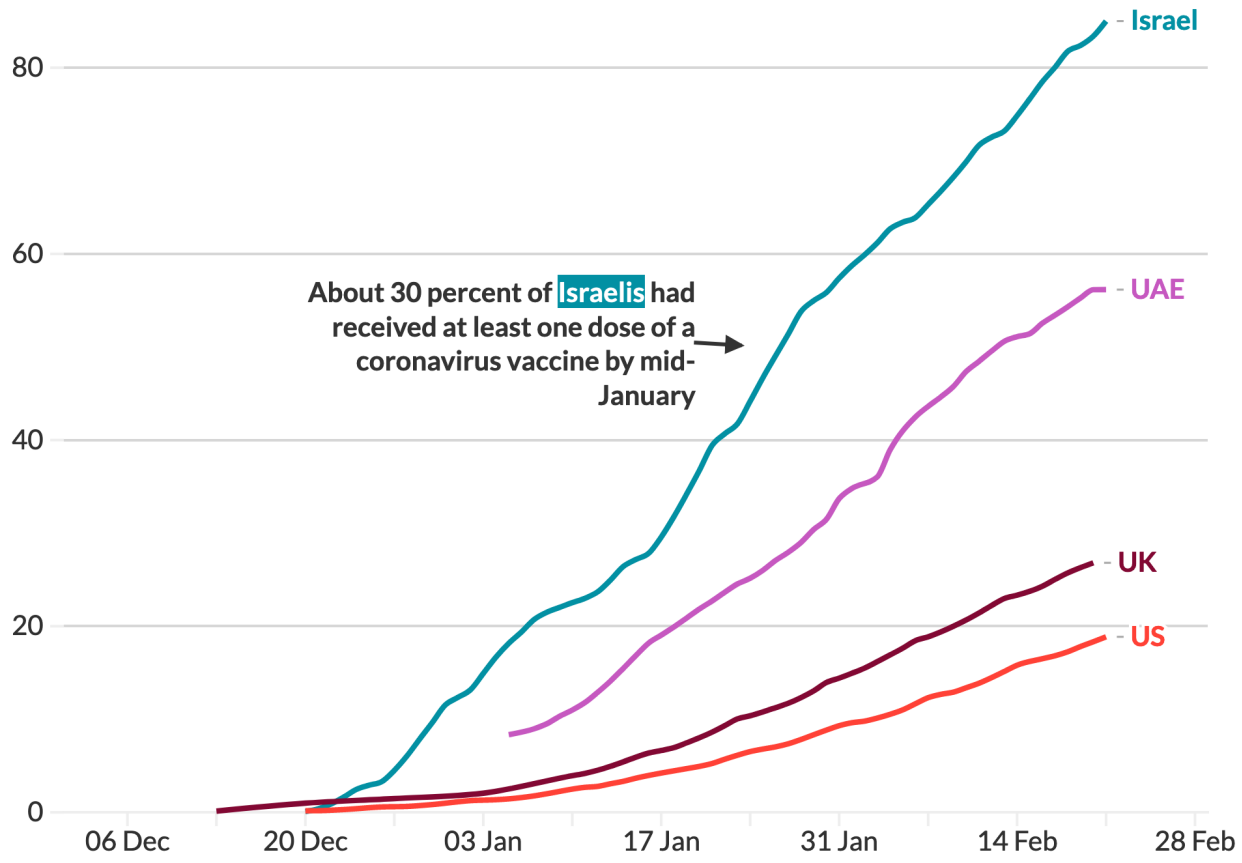
VISUAL STORYTELLING COMBINATIONS

- **text-specific**, where text provides all that is needed to know and the charts illustrate some aspects of the story that is described
- **chart-specific**, where the charts provide all that is needed to know and the text accentuates some aspects of the story that is shown
- **duo-specific**, where text and charts are both telling roughly the same story
- **intersecting**, where text and charts work together in some respects but also contribute to the story independently
- **interdependent**, where text and charts combine to convey an aspect of the story that neither could convey alone
- **parallel**, where words and charts follow seemingly different storylines, without intersecting

[https://flourish.studio/blog/accessible-chart-design]

Cumulative vaccination doses administered in Israel, UAE, UK and US

Cumulative doses administered per 100 residents • Data last updated 24 Feb



Source: ECDC/OWID • Graphic: Flourish • Embed this

[P. Tickell, The Story of the Impossible Train]

I have a story I'd like to tell you. It's about a train, and a group of people who live on that train and know of nothing else.

This train has been moving since anyone can remember. The people on the train can't imagine a time when the train wasn't moving, and when they were not on the train. Everyone works to keep the train moving. The train never stops.

It never stops. It cannot stop.

People on the train live in constant churn. The work to keep the train moving is hard, and inhumane. On the train, people are treated with cruelty and oppression. Some are treated worse than others. But nobody is truly living.

Sometimes they get breaks, but it is hard.

One day, a fire breaks out in one of the carriages of the train.

There is panic. The fire spreads throughout the whole train… Without getting off the train everyone is going to die.

Then the impossible happens.

The brakes no-one believed existed start to work. In the emergency, no-one notices how extraordinary it is that the train is stopping. They're too focused on the fire. The old rules go out the window.

For years on the train, the "worker class" of people have been dying from the awful conditions of the work they have to do on the train. They sleep in the aisles and sometimes have nowhere to sleep at all.

Suddenly, there are orders to house them and treat their ailments.

The train stops, and people begin to get off. Apart from the sound of the fire, suddenly there is a great silence.

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

A WORD ABOUT ACCESSIBILITY

A table can be translated to Braille, 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 their features in a fashion that allows all to "see" the insights.

But this requires them to have "seen" all the insights, which is not always necessarily the case (if at all possible).

A WORD ABOUT ACCESSIBILITY

Data Perception:

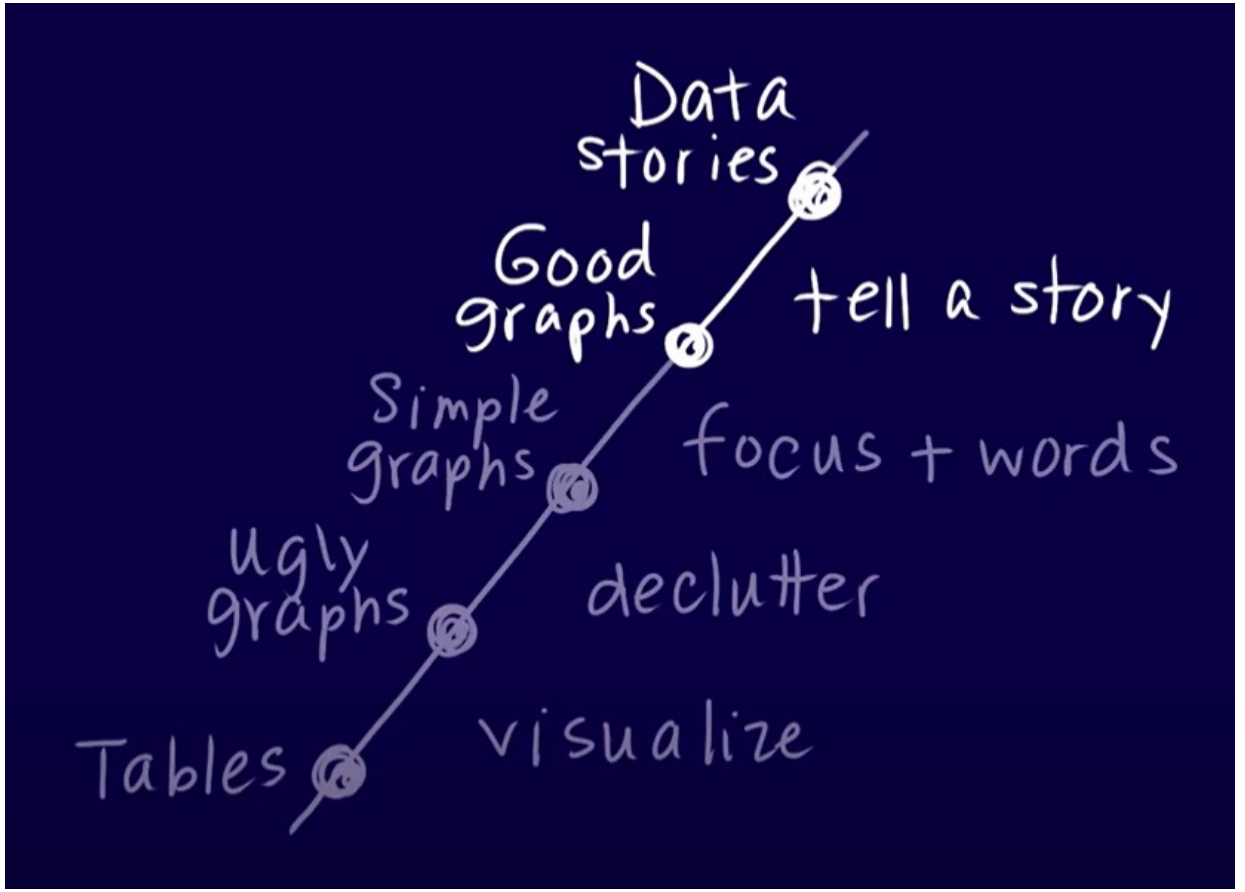
- texture-based representations
- text-to-speech
- sound/music
- odor-based representations (?)
- taste-based representations (?!?)

Sonifications:

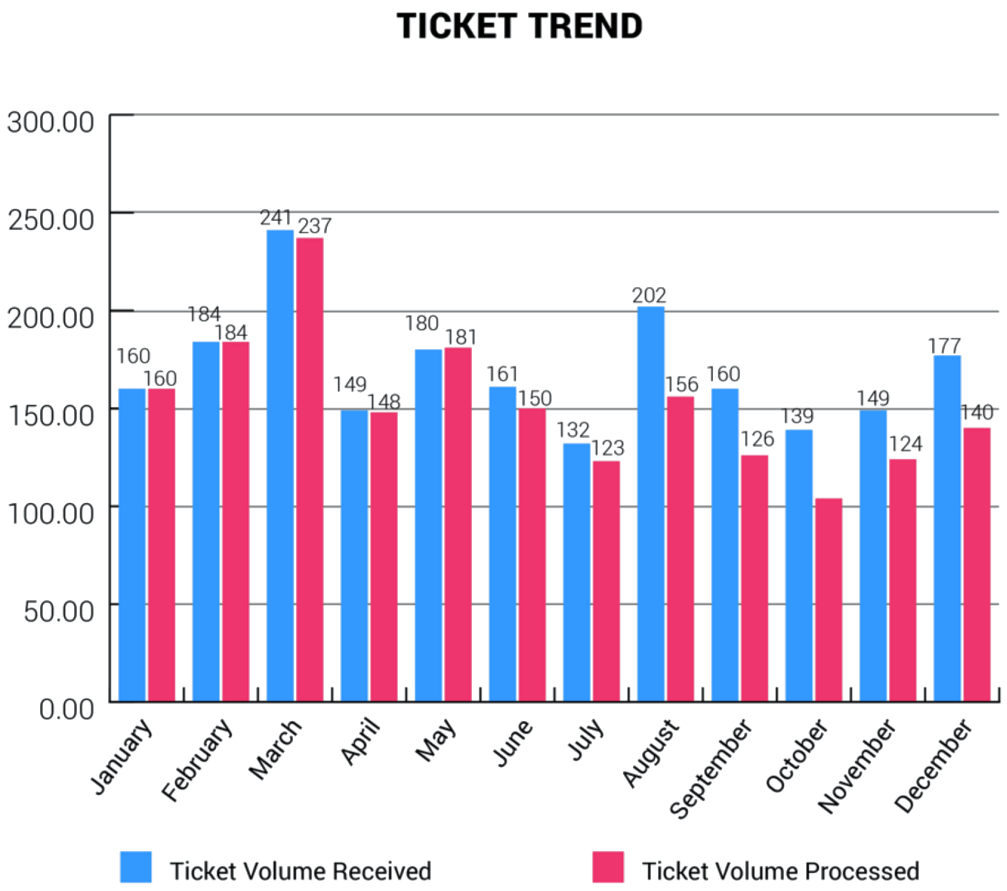
- TRAPPIST Sounds : TRAPPIST-1 Planetary System Translated Directly Into Music
- Listening to data from the Large Hadron Collider, L. Asquith

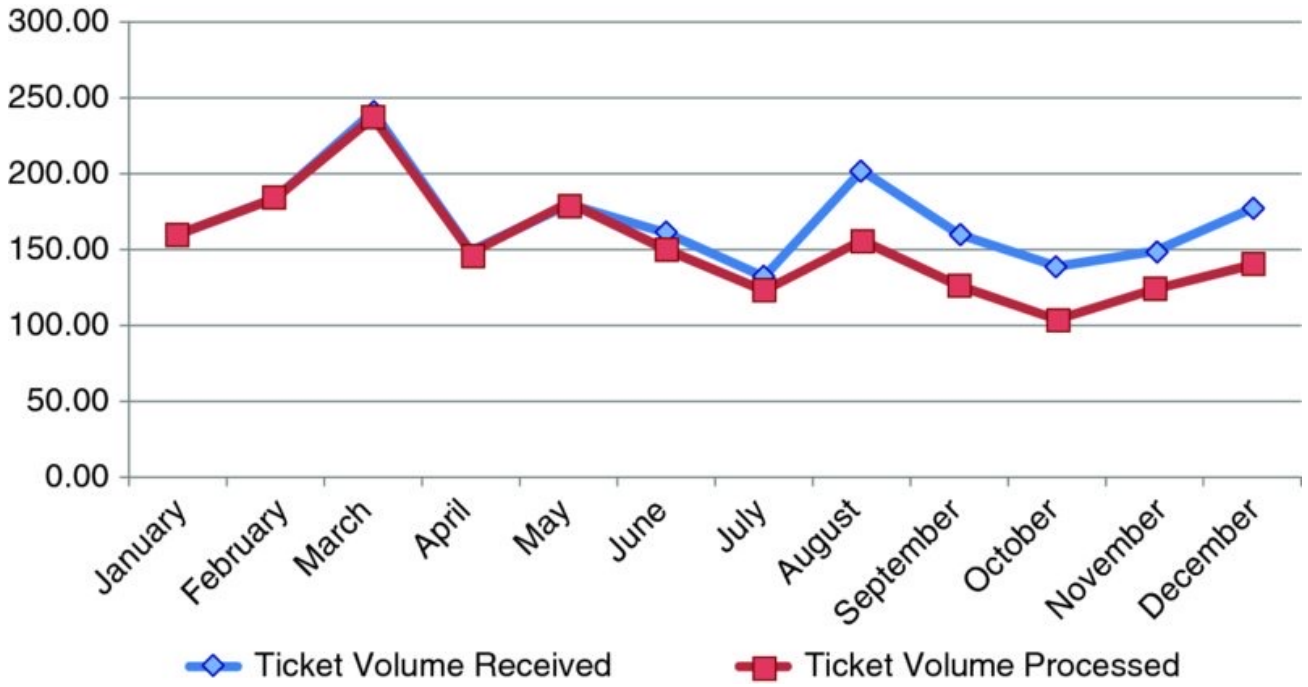
[C. Nussbaumer Knaflic, *Storytelling with Data*]

EVOLVING A VISUALIZATION



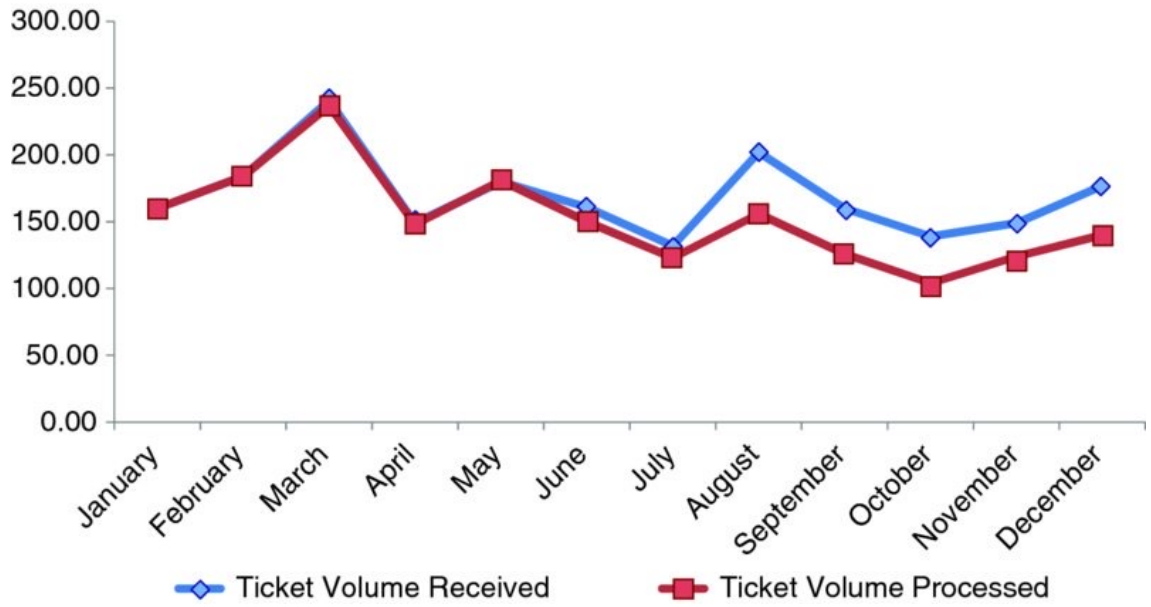
EXAMPLES

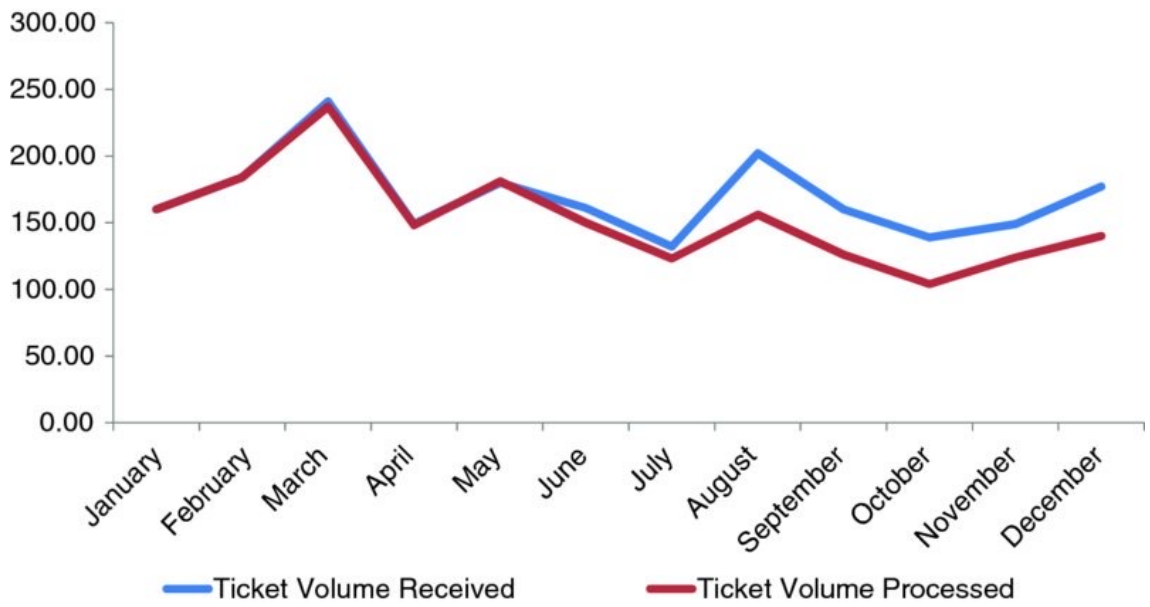




1. Change graph type (trope)

EXAMPLES

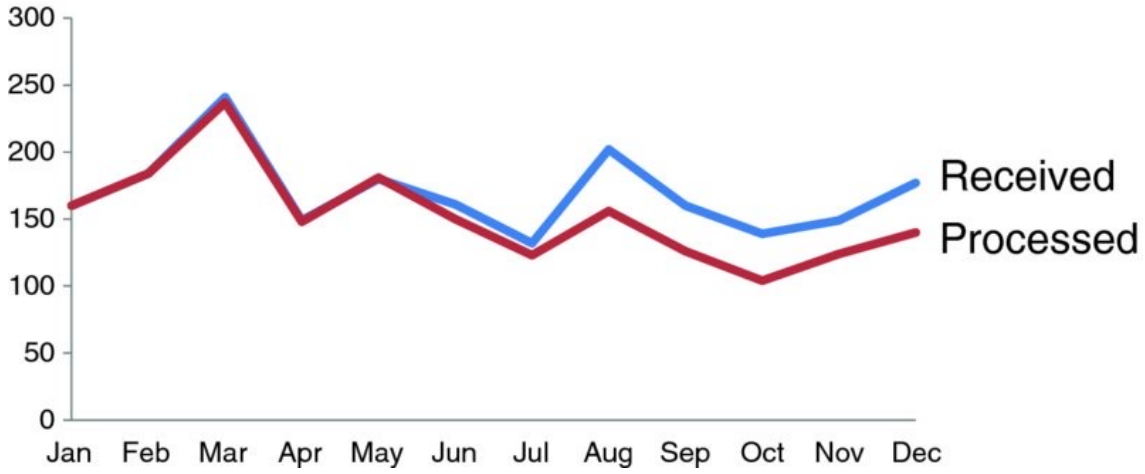




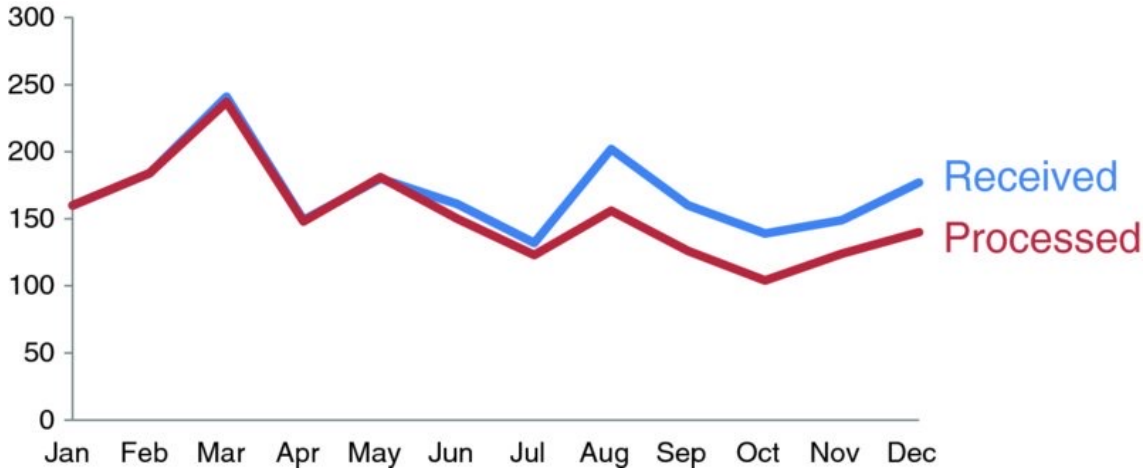
2. Remove border and gridlines

3. Remove markers

EXAMPLES

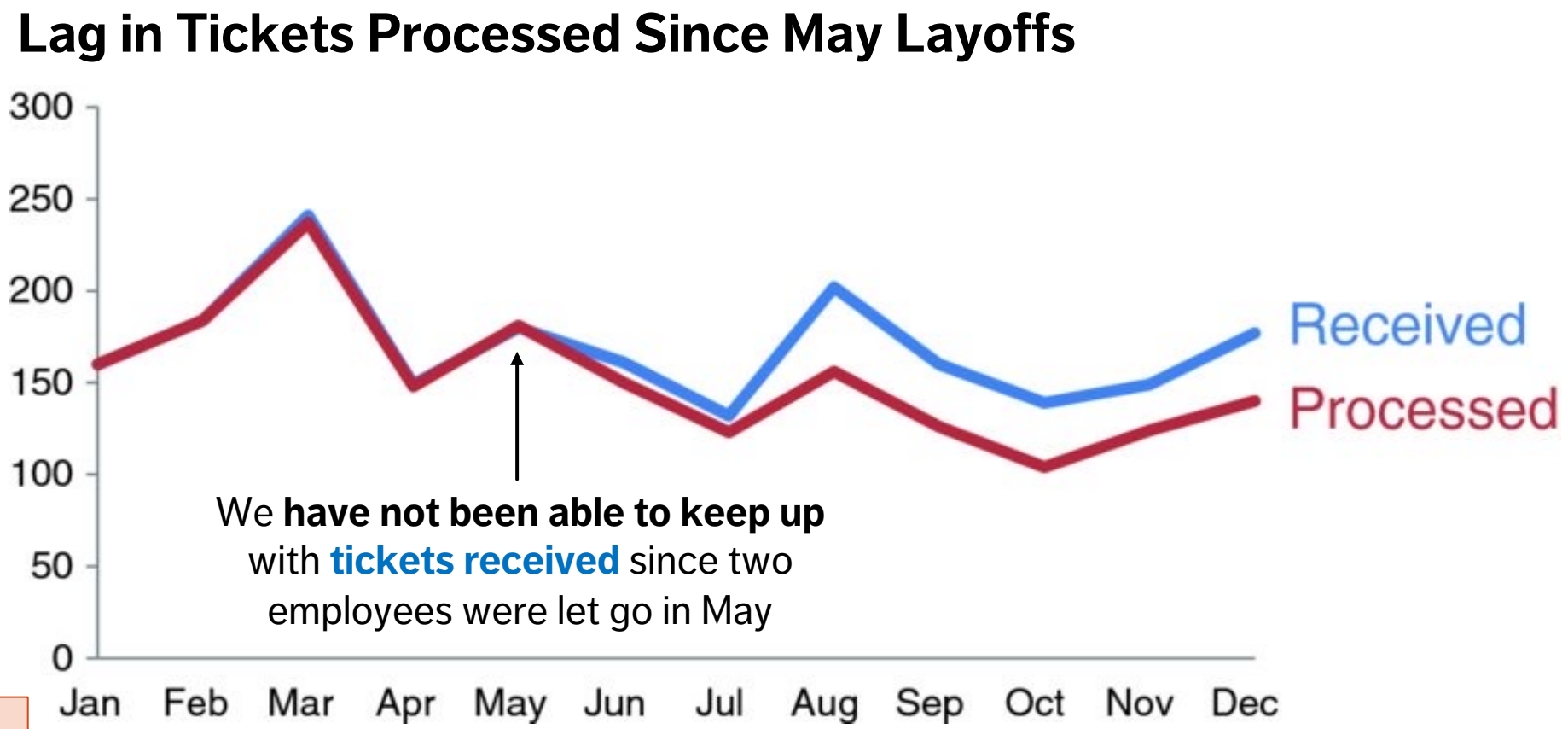


4. Clean-up axis labels and legend



5. Colour code the lines

EXAMPLES



6. Tell the story

DATA STORY TROPES

Some visualizations patterns are so familiar that they become **tropes:**

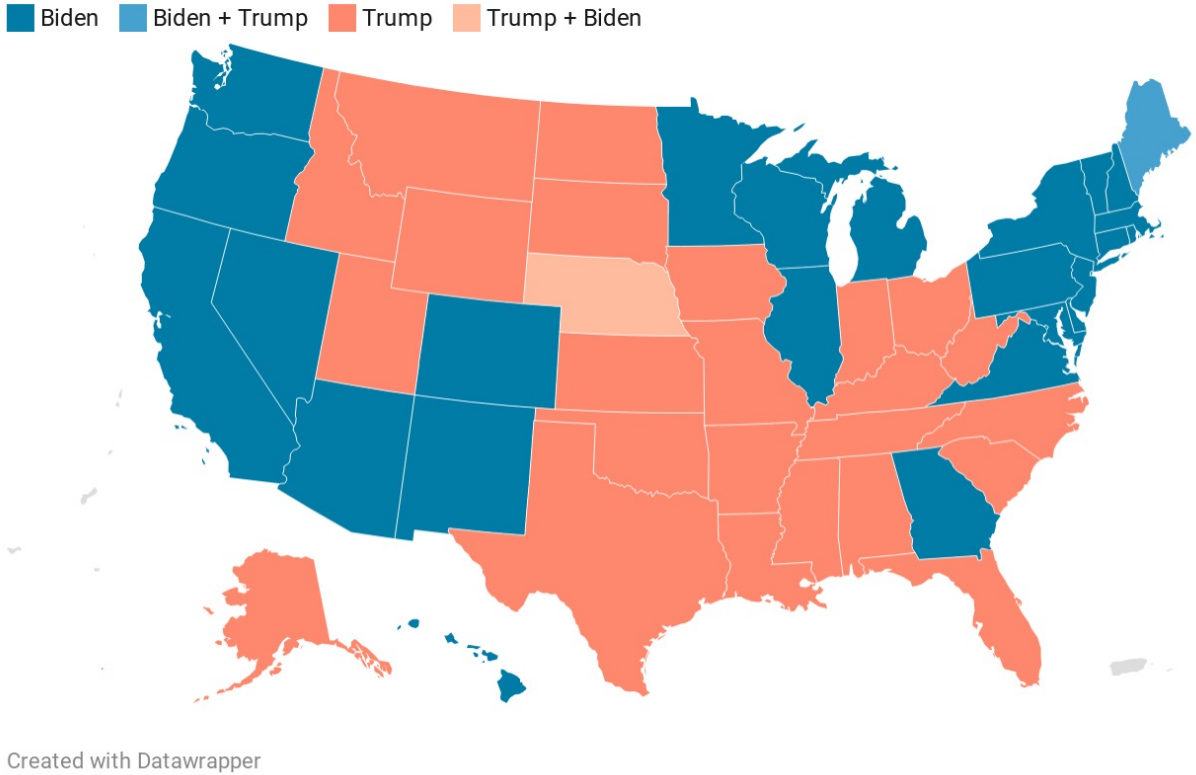
- a scatterplot with a trend line going straight up or straight down
- a cluster bar chart with two categories where one is always lower than the other
- a line chart with the two lines crossing in one place
- pie charts being used all over the place (to avoid)
- red for republican, blue for democrat (US); red for left-leaning, blue for right-leaning (ROW)
- using broken axes to exaggerate effects (sometimes justified…)
- etc.

[https://handsondataviz.org/spatial-bias.html]

DATA STORYTELLING TROPES – EXAMPLES

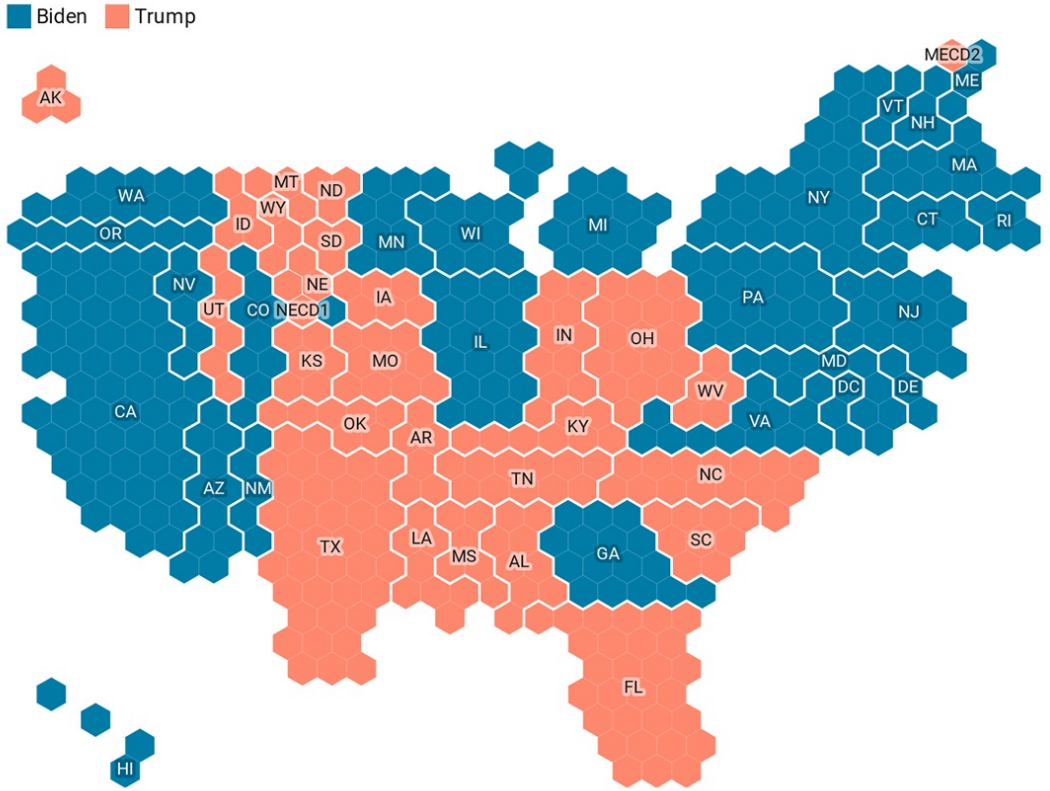
Conventional Map of 2020 US Presidential Election Results

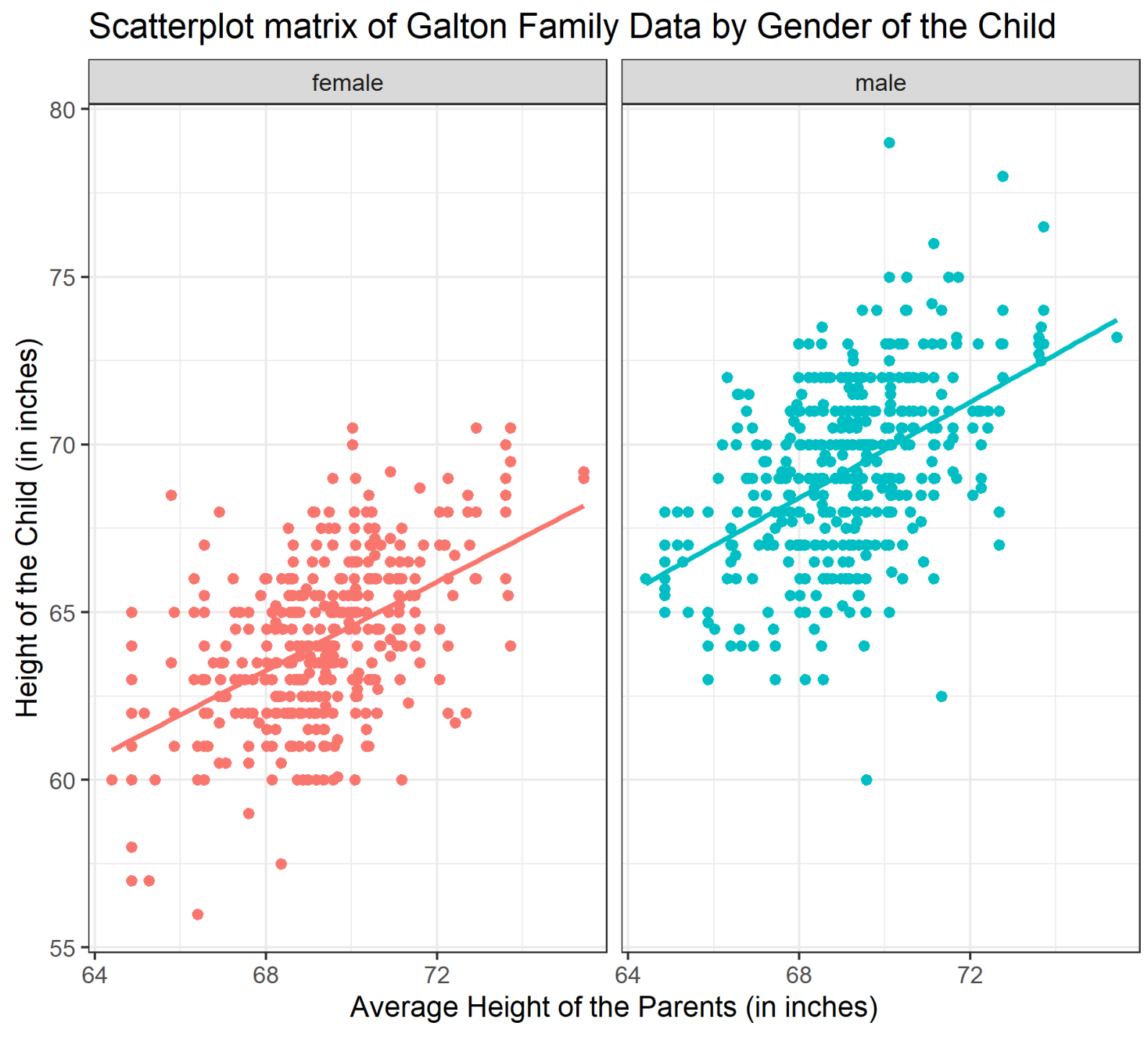
Maine and Nebraska allow some electoral votes to be split by district



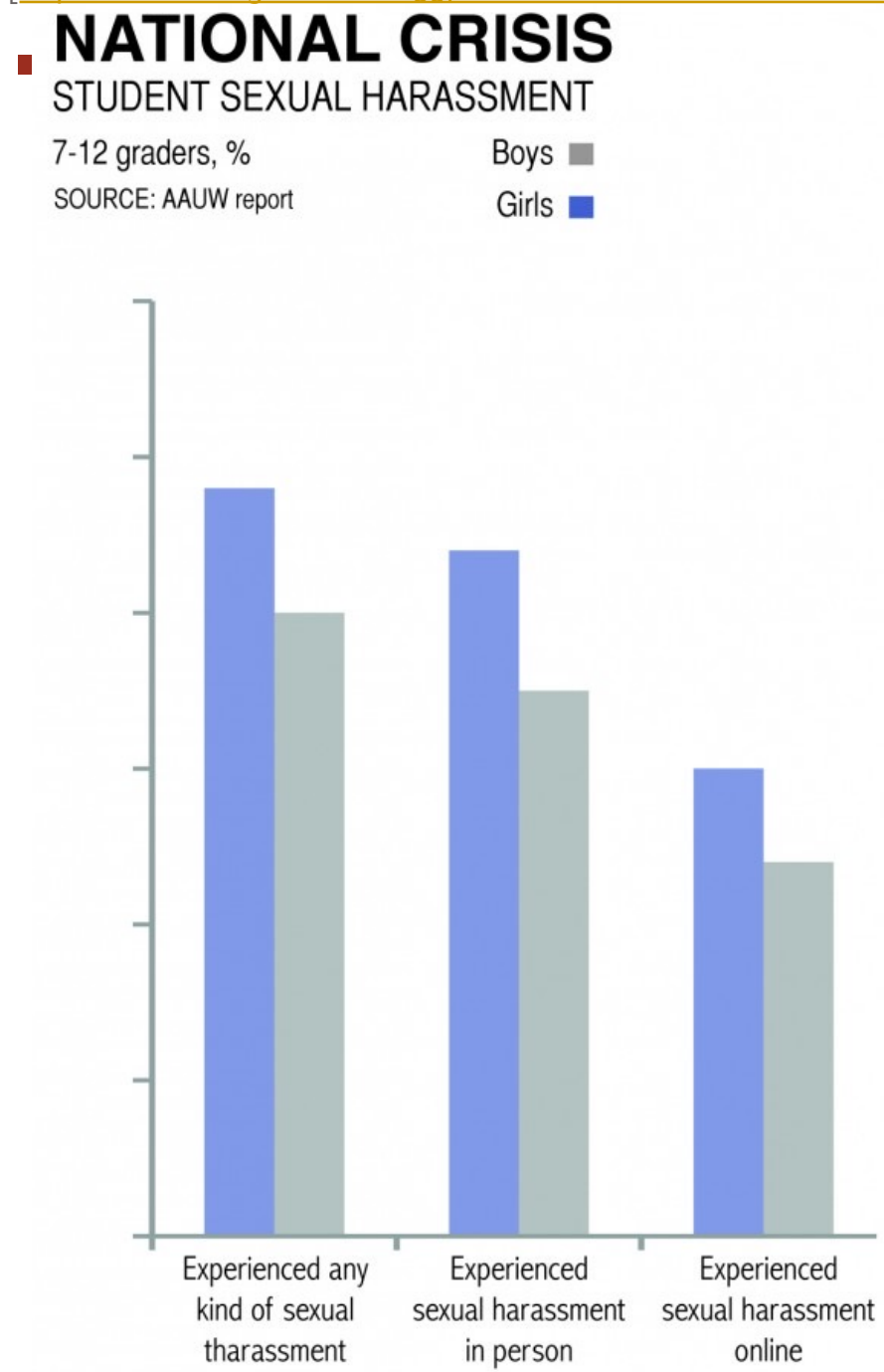
Cartogram of 2020 US Presidential Election Results

Each hexagon represents one electoral college vote





[https://www.chsglobe.com/13376/cover-stories/sexualharassment]



VISUAL PROCESSING

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

Visual thinking seeks **patterns**

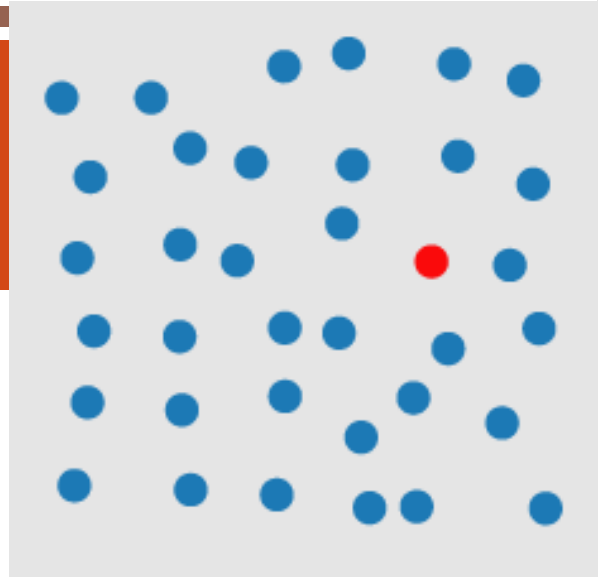
- **pre-attentive processes:** fast, instinctive, efficient, multitasking
  gather information and build patterns:

  features → patterns → objects

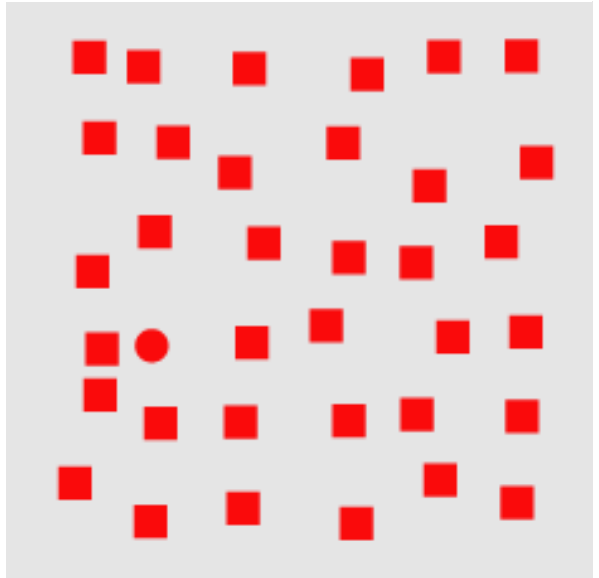
- **attentive process:** slow, deliberate, focused
  discover features in the patterns:

  objects → patterns → features

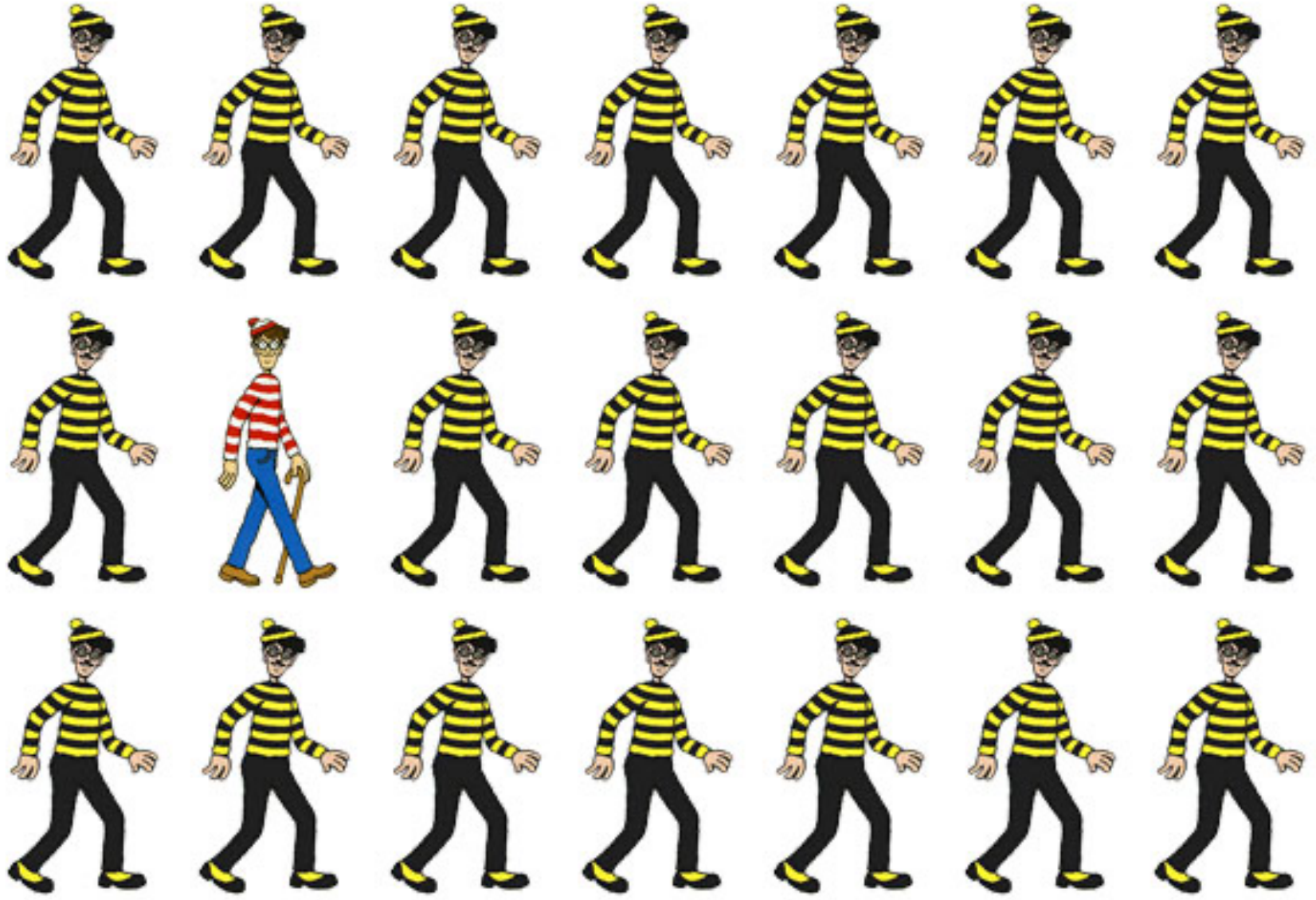
pre-attentive



attentive



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



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