VISUALIZATION: A WHIRLWIND TOUR

SOUMYA BANERJEE

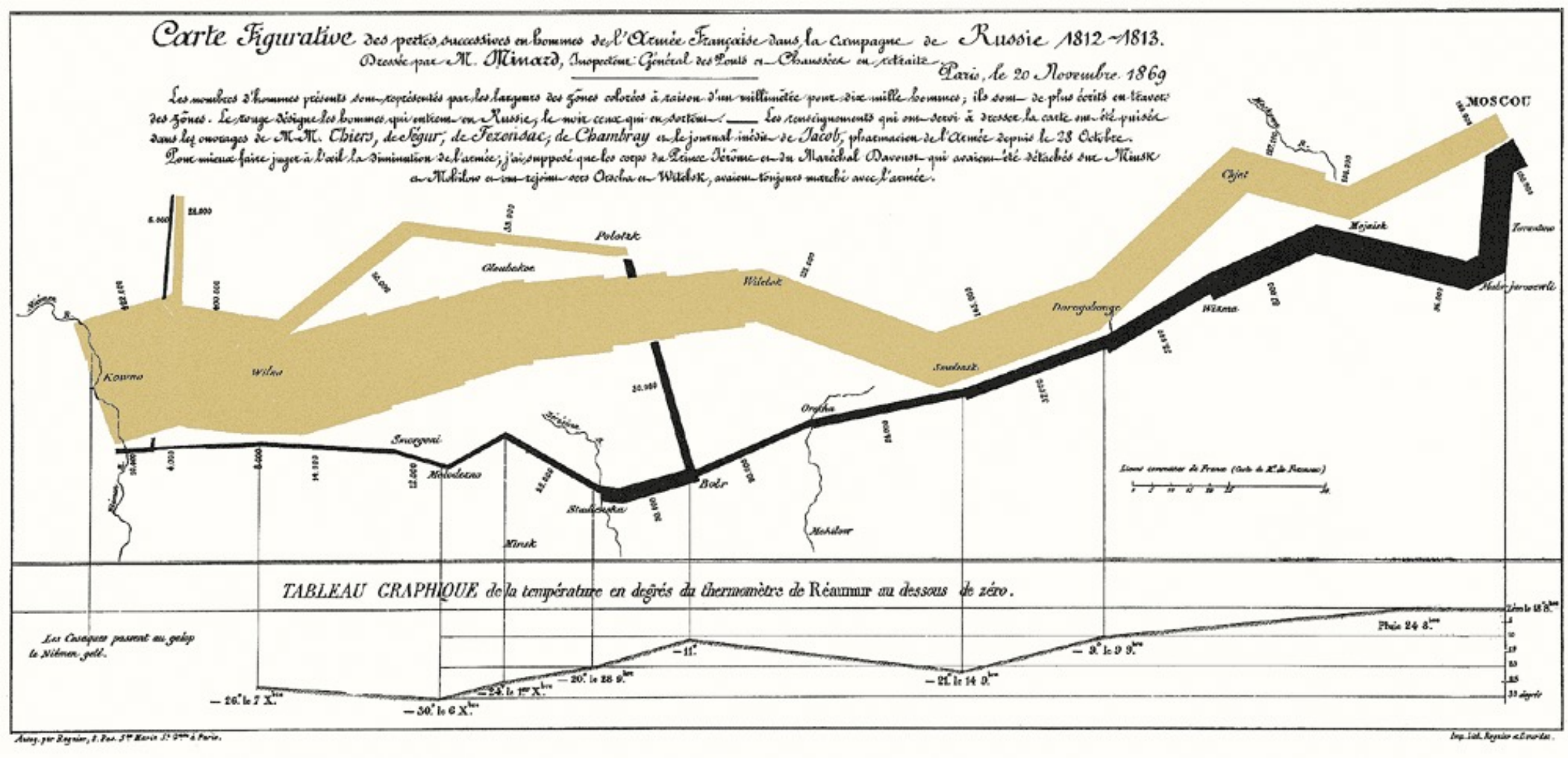
VISUALIZATION AS DEBUGGING

1. DATA SCIENCE AS (VISUAL) DEBUGGING

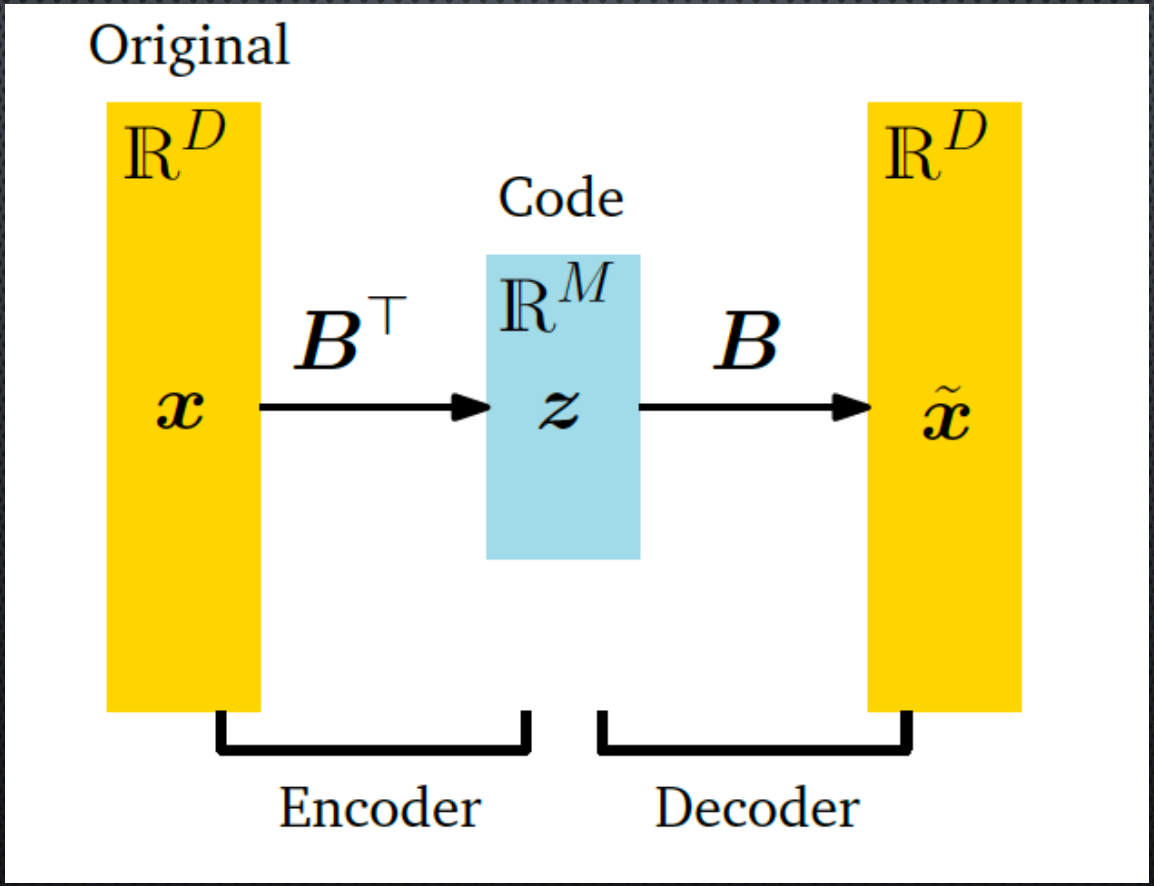
2. DATA STORYTELLING TO WORK CLOSELY WITH DOMAIN EXPERTS

3. CASE STUDIES
    A. GENOMIC DATA
    B. FORECASTING GDP
    C. MENTAL HEALTH OUTCOME PREDICTION

THE BEST STATISTICAL GRAPHIC EVER (EDWARD TUFTE)



IMPORTANT CONCEPT



INFORMATION BOTTLENECK

GENERALIZATIONS OF THIS IDEA

tSNE

Autoencoder (non-linear loss function)

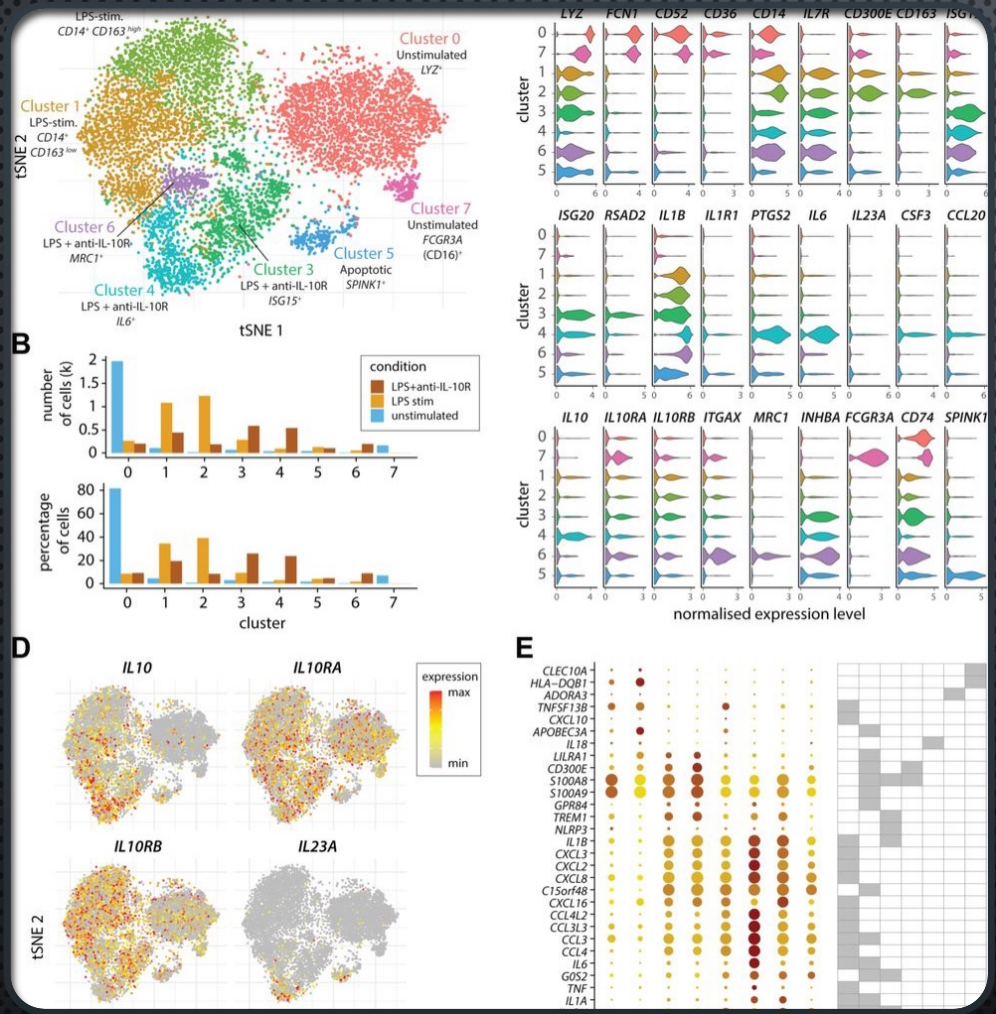
PITFALLS

tSNE can be used for hypothesis generation.

However There are many pitfalls

https://distill.pub/2016/misread-tsne/

some pitfalls: distances not preserved. For example, a 2D map is a projection from 3D



CASE STUDY

a. stochastic
b. distances not preserved
c. difficult to communicate to non-technical experts

VISUALIZATIONS CAN BE MISLEADING

1. DISTANCES NOT PRESERVED IN TSNE

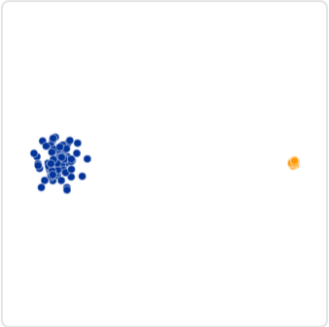
   HTTPS://DISTILL.PUB/2016/MISREAD-TSNE/

2. CLUSTER SIZES DO NOT MATTER

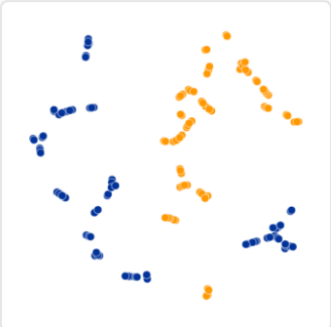
3. YOU CAN SEE SOME SHAPES SOMETIMES

4. RANDOM DOES NOT ALWAYS LOOK RANDOM

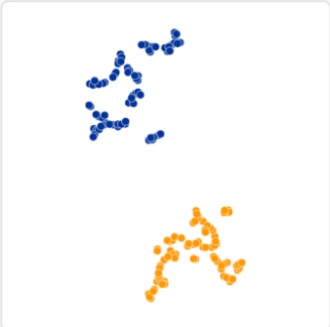
VISUALIZATIONS CAN BE MISLEADING



Original

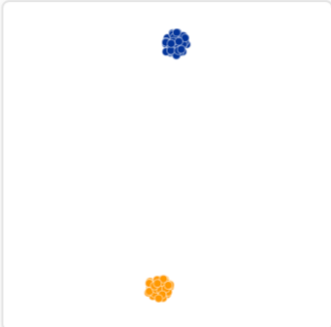


Perplexity: 2
Step: 5,000

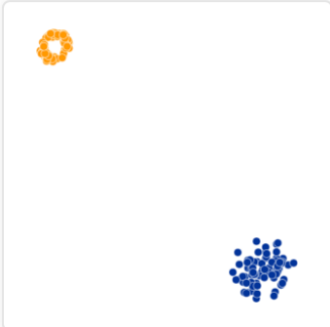


Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



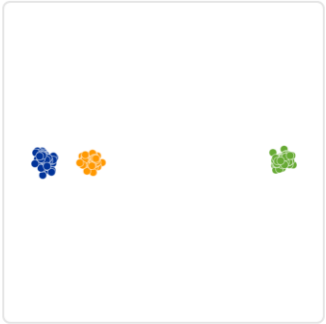
Perplexity: 100
Step: 5,000

CLUSTER SIZES DO NOT MEAN ANYTHING

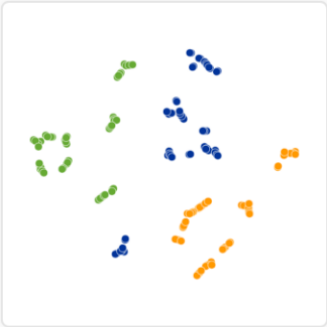
BUT WHAT IF THE TWO CLUSTERS HAVE DIFFERENT STANDARD DEVIATIONS, AND SO DIFFERENT SIZES? (BY SIZE WE MEAN BOUNDING BOX MEASUREMENTS, NOT NUMBER OF POINTS.) BELOW ARE T-SNE PLOTS FOR A MIXTURE OF GAUSSIANS IN PLANE, WHERE ONE IS 10 TIMES AS DISPERSED AS THE OTHER.

SURPRISINGLY, THE TWO CLUSTERS LOOK ABOUT SAME SIZE IN THE T-SNE PLOTS. WHAT'S GOING ON? THE T-SNE ALGORITHM ADAPTS ITS NOTION OF "DISTANCE" TO REGIONAL DENSITY VARIATIONS IN THE DATA SET. AS A RESULT, IT NATURALLY EXPANDS DENSE CLUSTERS, AND CONTRACTS SPARSE ONES, EVENING OUT CLUSTER SIZES.

VISUALIZATIONS CAN BE MISLEADING

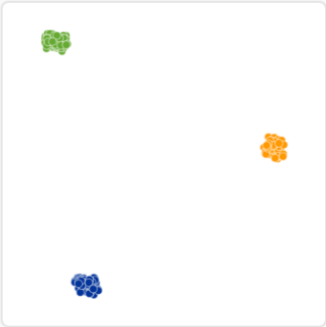


Original

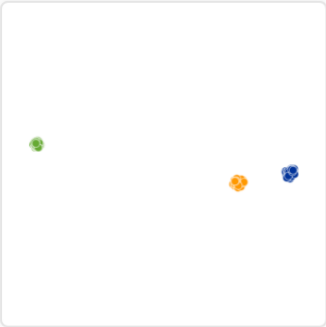


Perplexity: 2
Step: 5,000

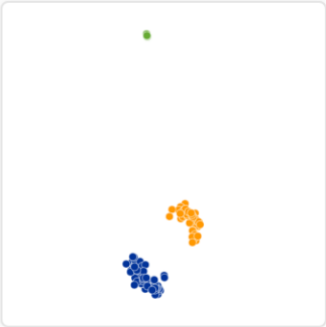
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

DISTANCES BETWEEN CLUSTERS MAY NOT MEAN ANYTHING

DIAGRAMS SHOW THREE GAUSSIANS OF 50 POINTS EACH, ONE PAIR BEING 5 TIMES AS FAR APART AS ANOTHER PAIR.

AT PERPLEXITY 50, THE DIAGRAM GIVES A GOOD SENSE OF THE GLOBAL GEOMETRY. FOR LOWER PERPLEXITY VALUES THE CLUSTERS LOOK EQUIDISTANT. WHEN THE PERPLEXITY IS 100, WE SEE THE GLOBAL GEOMETRY FINE, BUT ONE OF THE CLUSTER APPEARS, FALSELY, MUCH SMALLER THAN THE OTHERS. SINCE PERPLEXITY 50 GAVE US A GOOD PICTURE IN THIS EXAMPLE, CAN WE ALWAYS SET PERPLEXITY TO 50 IF WE WANT TO SEE GLOBAL GEOMETRY?

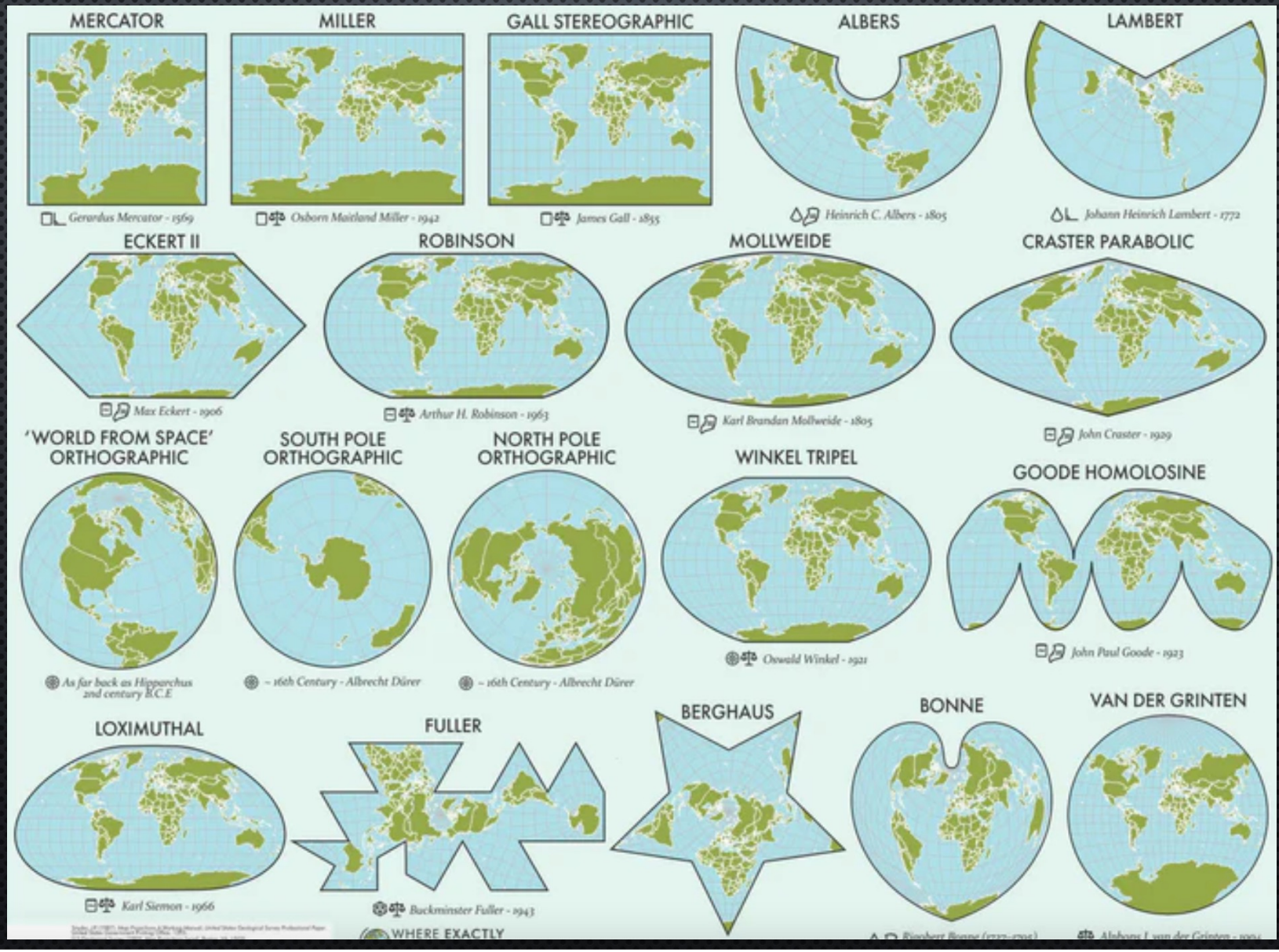
THE BASIC MESSAGE IS THAT DISTANCES BETWEEN WELL-SEPARATED CLUSTERS IN A T-SNE PLOT MAY MEAN NOTHING.

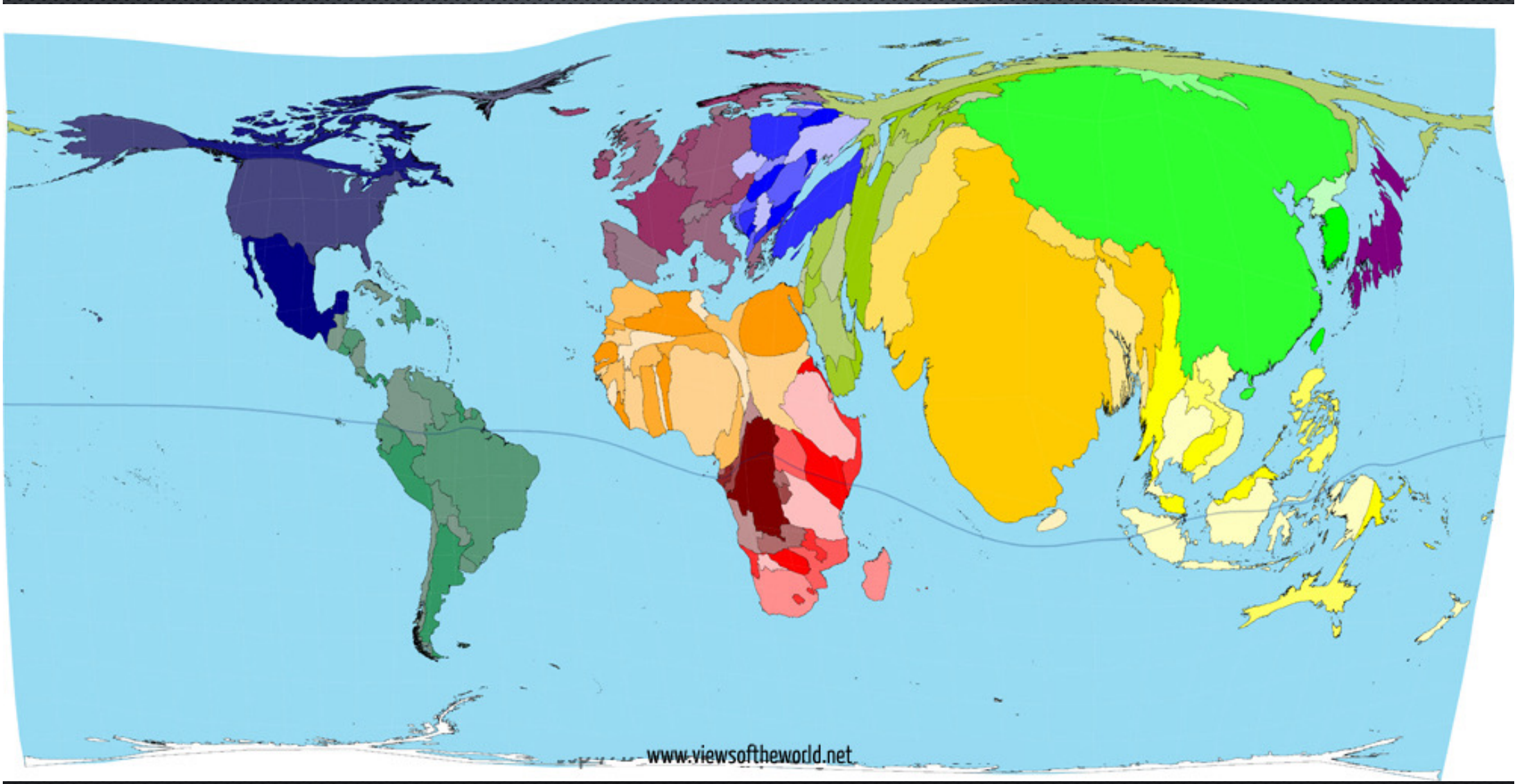
VISUALIZATIONS CAN BE MISLEADING

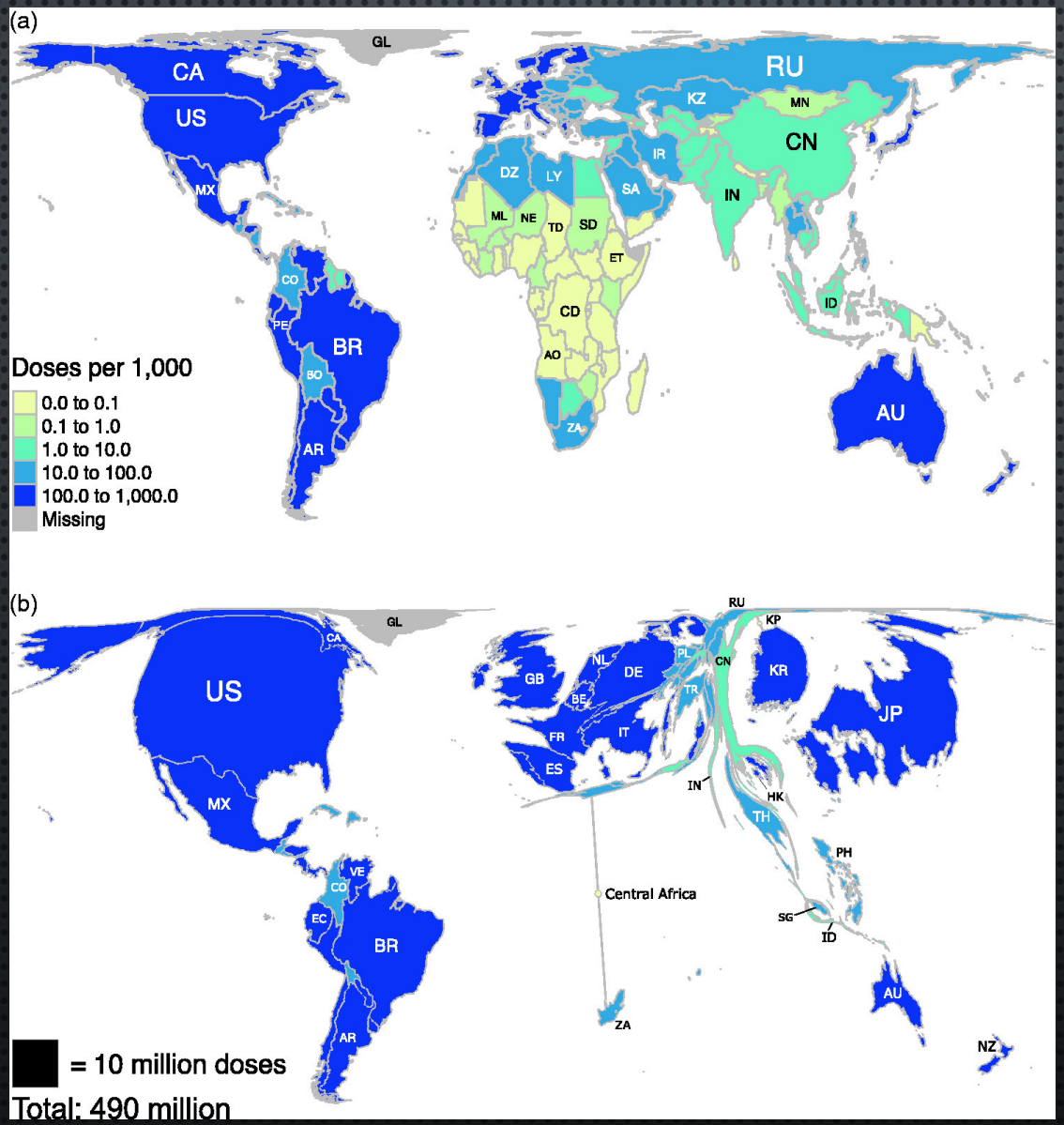
1. Gorilla experiment (selective attention bias)
2. Visualization as a hypothesis generation tool
3. Communicate clearly with stakeholders and validate these findings

VISUALIZING HIGH-DIMENSIONS

1. PROBLEMS WITH COMMUNICATING HIGH DIMENSIONAL DATA
2. HIGH DIMENSIONS ARE DIFFICULT TO VISUALIZE


MERCATOR
Gerardus Mercator - 1569
MILLER
Osborn Maitland Miller - 1942
GALL STEREOGRAPHIC
James Gall - 1855
ALBERS
Heinrich C. Albers - 1805
LAMBERT
Johann Heinrich Lambert - 1772
ECKERT II
Max Eckert - 1906
ROBINSON
Arthur H. Robinson - 1963
MOLLWEIDE
Karl Brandan Mollweide - 1805
CRASTER PARABOLIC
John Craster - 1929
'WORLD FROM SPACE' ORTHOGRAPHIC
As far back as Hipparchus 2nd century B.C.E
SOUTH POLE ORTHOGRAPHIC
~ 16th Century - Albrecht Dürer
NORTH POLE ORTHOGRAPHIC
~ 16th Century - Albrecht Dürer
WINKEL TRIPEL
Oswald Winkel - 1921
GOODE HOMOLOSINE
John Paul Goode - 1923
LOXIMUTHAL
Karl Siemon - 1966
FULLER
Buckminster Fuller - 1943
BERGHAUS
BONNE
VAN DER GRINTEN
WHERE EXACTLY


www.viewsoftheworld.net



Area proportional to number of vaccine doses

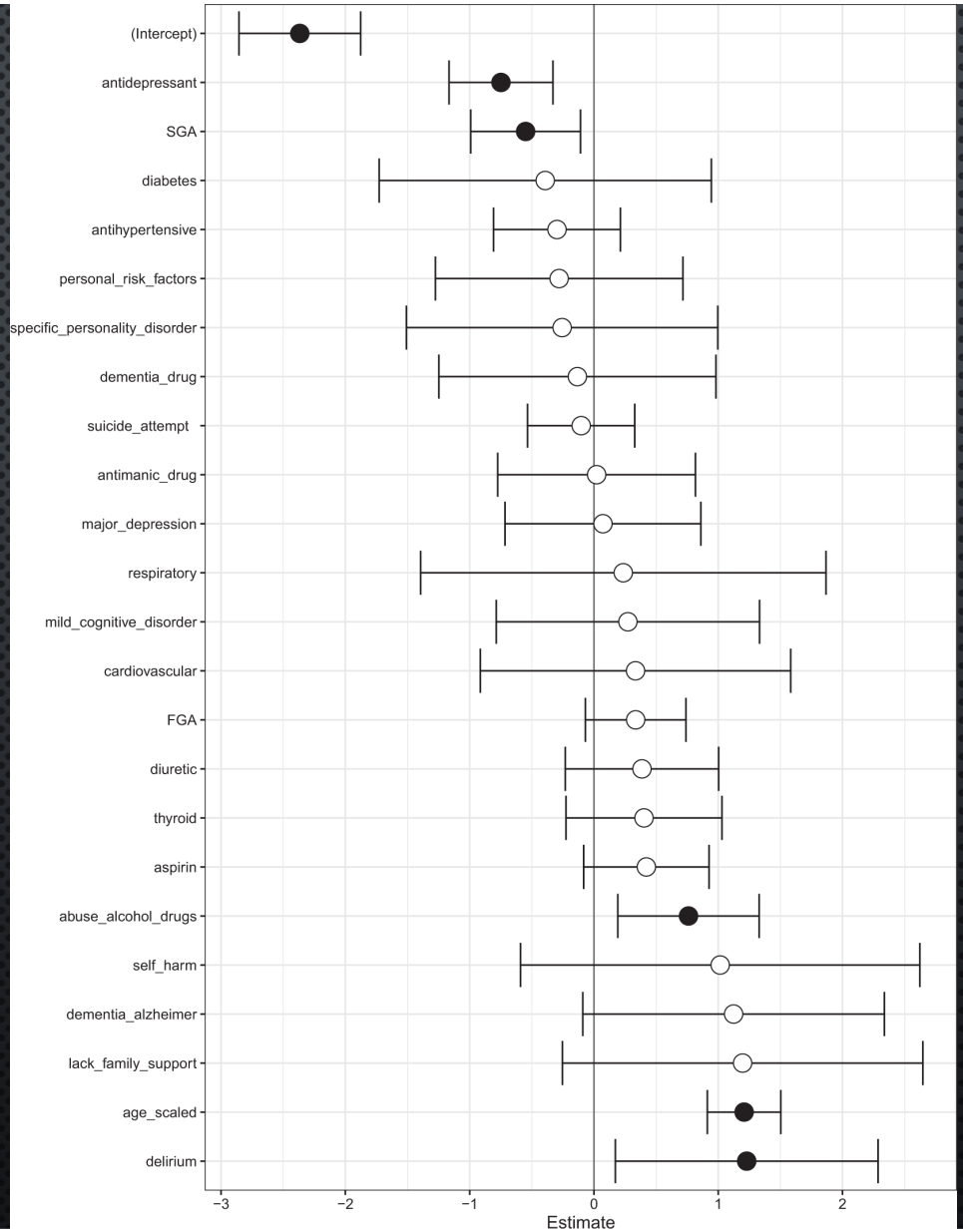
CARTOGRAMS

https://go-cart.io/tutorial

VISUALIZATION FOR DIAGNOSTICS

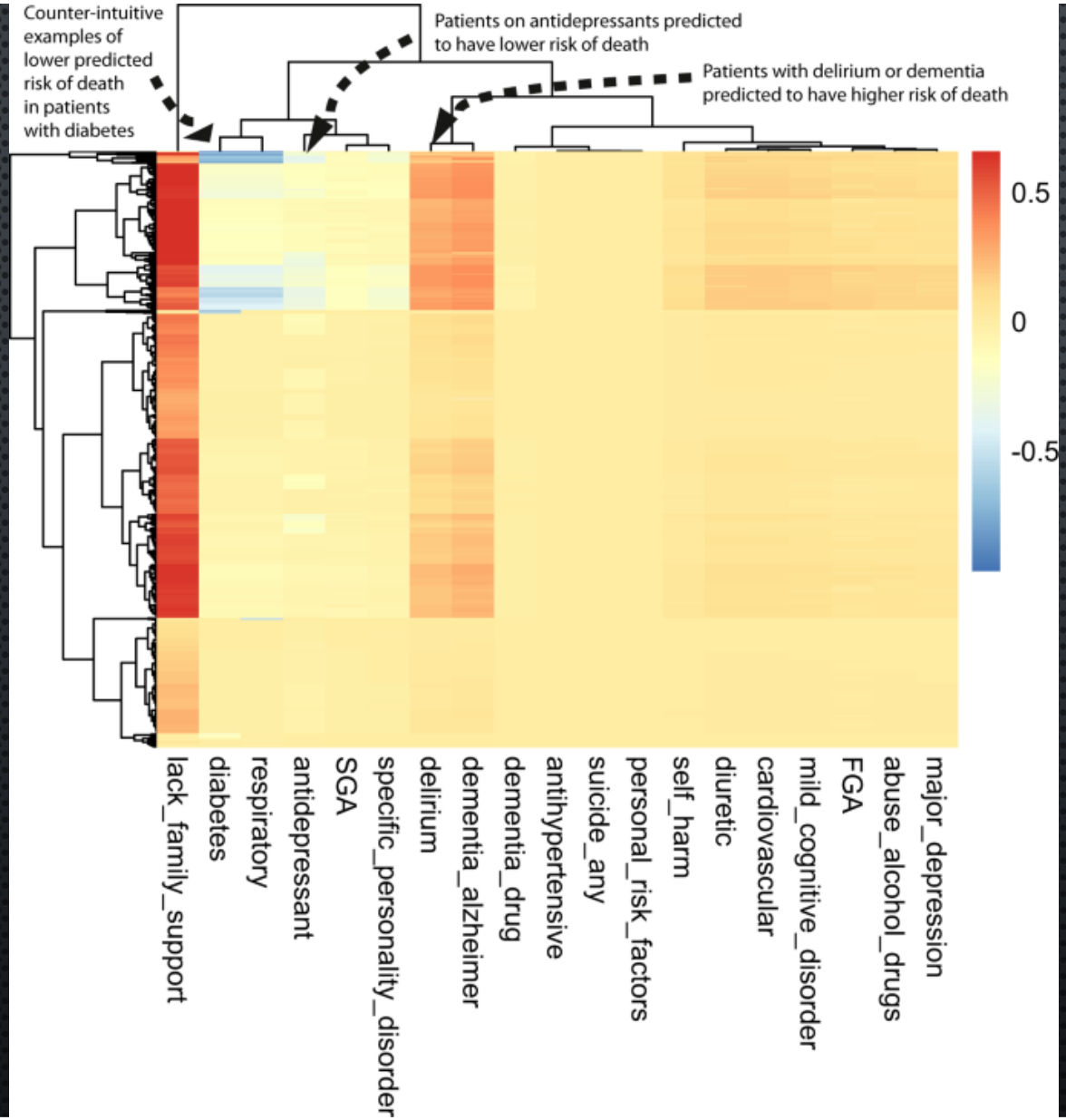
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

GENERALIZED LINEAR MODEL

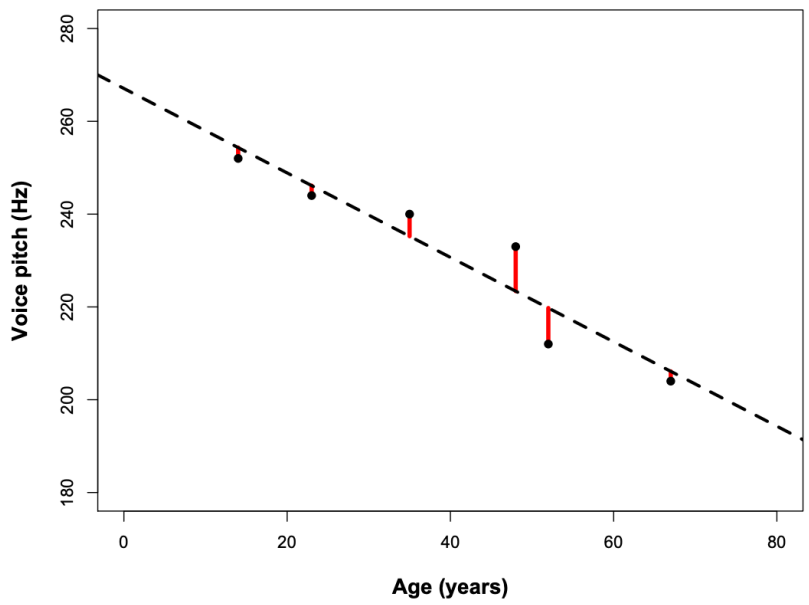


VISUALIZATION FOR DIAGNOSTICS

VISUALIZATION FOR DIAGNOSTICS

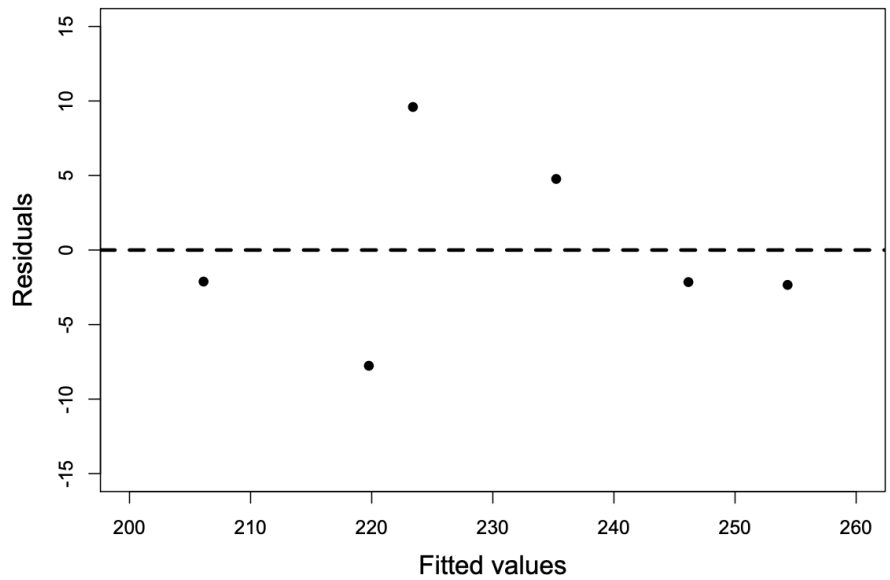


ASSUMPTIONS: LINEARITY



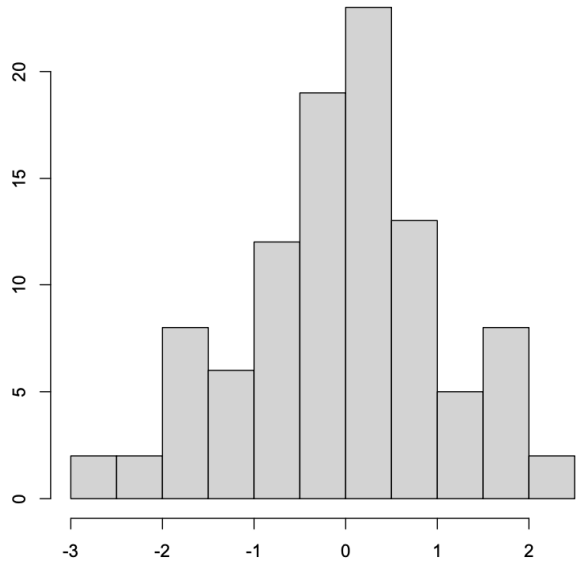
The red lines indicate the residuals, which are the deviations of the observed data points from the predicted values (the so-called "fitted values"). In this case, the residuals are all fairly small … which is because the line that represents the linear model predicts our data very well, i.e., all points are very close to the line.

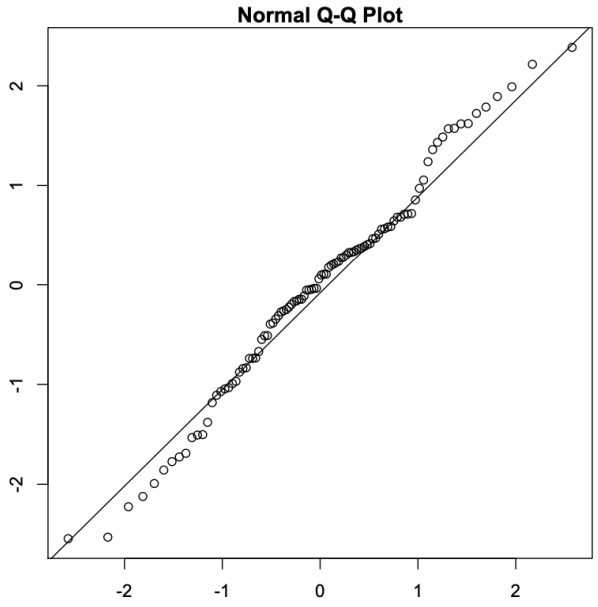
ASSUMPTIONS: LINEARITY



This is a residual plot. The fitted values (the predicted means) are on the horizontal line (at y=0). The residuals are the vertical deviations from this line. This view is just a rotation of the actual data (compare the residual plot with the scatterplot to see this). To construct the residual plot for yourself, simply type:

```
plot(fitted(xmdl),residuals(xmdl))
```
[2]

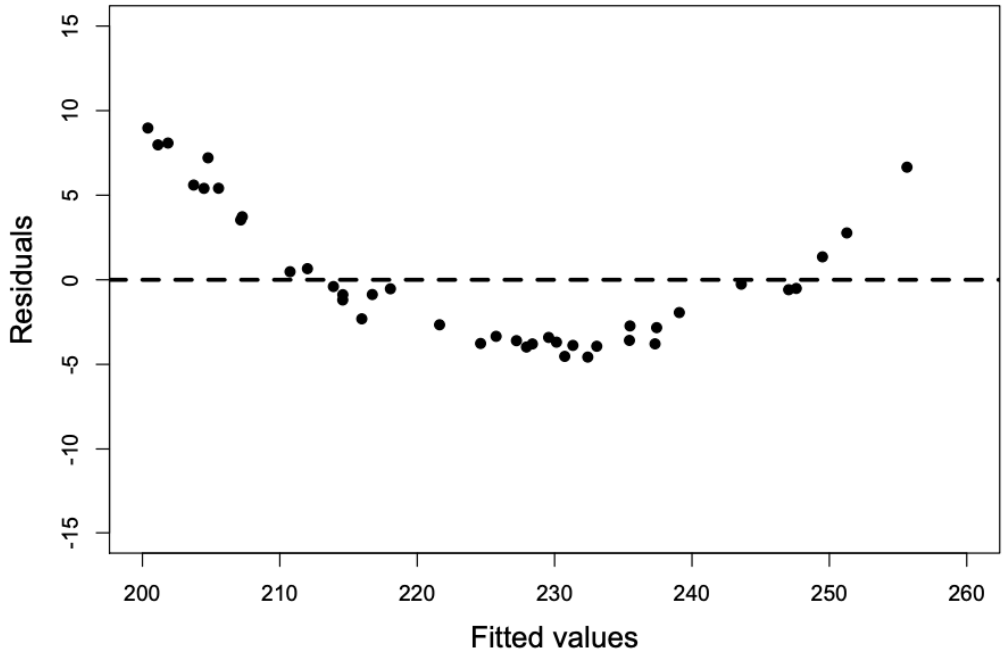




These look good. The histogram is relatively bell-shaped and the Q-Q plot indicates that the data falls on a straight line (which means that it's similar to a normal distribution). Here, we would conclude that there are no obvious violations of the normality assumption.

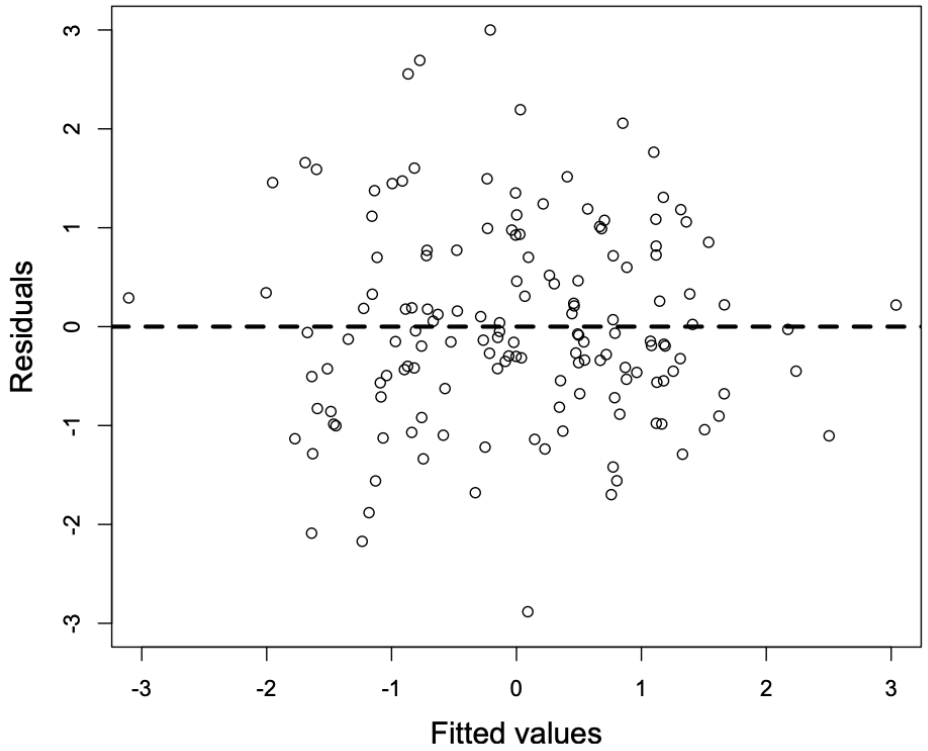
ASSUMPTIONS

1. NORMALITY OF RESIDUALS
2. Q-Q PLOT
3. HTTPS://STACKOVERFLOW.COM/QUESTIONS/13865596/QUANTILE-QUANTILE-PLOT-USING-SCIPY



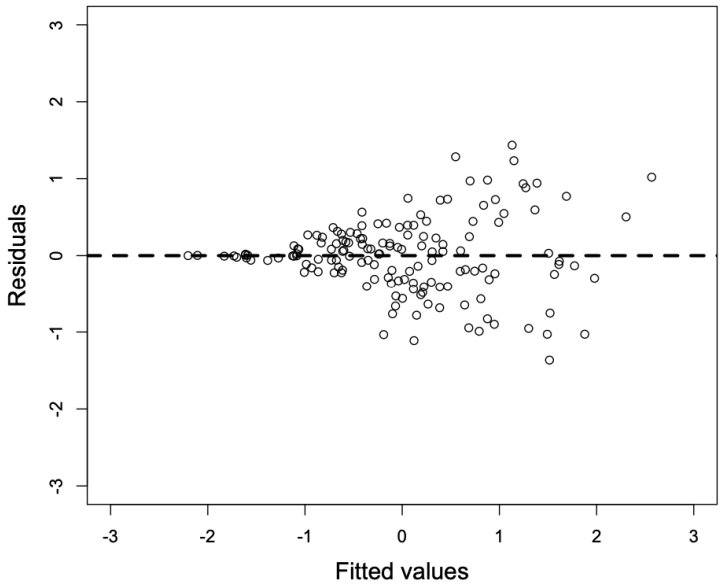
ASSUMPTIONS: LINEARITY

1. Other predictors
2. Transformation of response
3. Transformation of features



ASSUMPTIONS

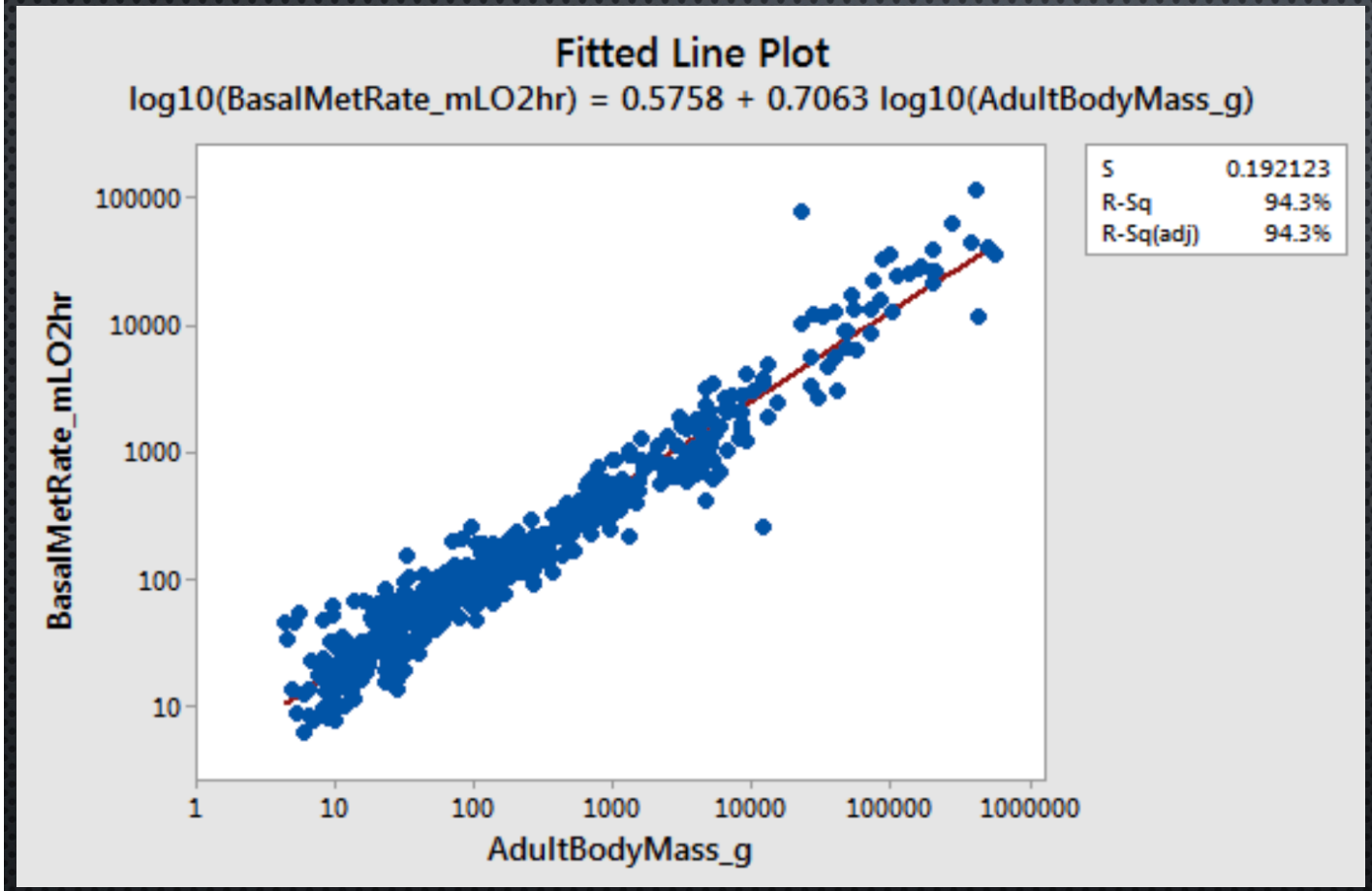
1. HOMOSKEDASTICITY
2. THE VARIABILITY OF YOUR DATA SHOULD BE APPROXIMATELY EQUAL ACROSS THE RANGE OF YOUR PREDICTED VALUES



In this plot, higher fitted values have larger residuals … indicating that the model is more "off" with larger predicted means. So, the variability is not homoscedastic: it's smaller in the lower range and larger in the higher range.

ASSUMPTIONS

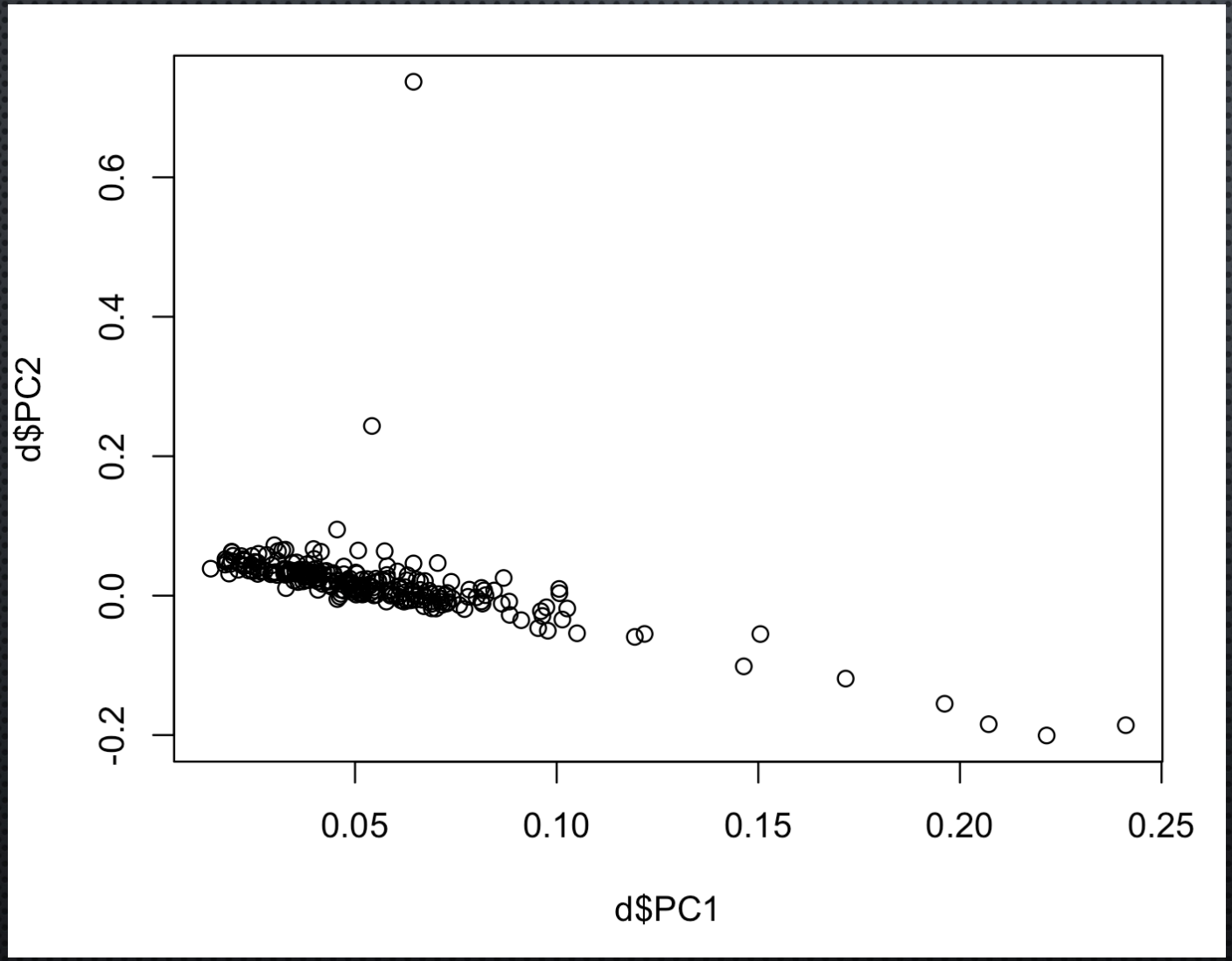
1. HOMOSKEDASTICITY
2. WHAT TO DO?
3. TRANSFORM YOUR DATA
4. DIFFERENT MODEL



DATA TRANSFORMATIONS

1. LOG-LOG PLOTS
2. CHECK THESE FOR NORMALITY. FOR EXAMPLE, IF THE DISTRIBUTION IS NOT NORMAL, THEN MAYBE YOU CAN TRANSFORM THE DATA SOMEHOW.

https://statisticsbyjim.com/regression/log-log-plots/



ABSENCE OF INFLUENTIAL DATA POINTS

1. Removing outliers and then some downstream processing

2. t-test

3. Run the analysis and without and with the data point

RESOURCE

Code for visualization in R, model diagnostics and linear mixed effects models

https://github.com/neelsoumya/anova_linear_mixed_effects_examples

TIME SERIES DATA

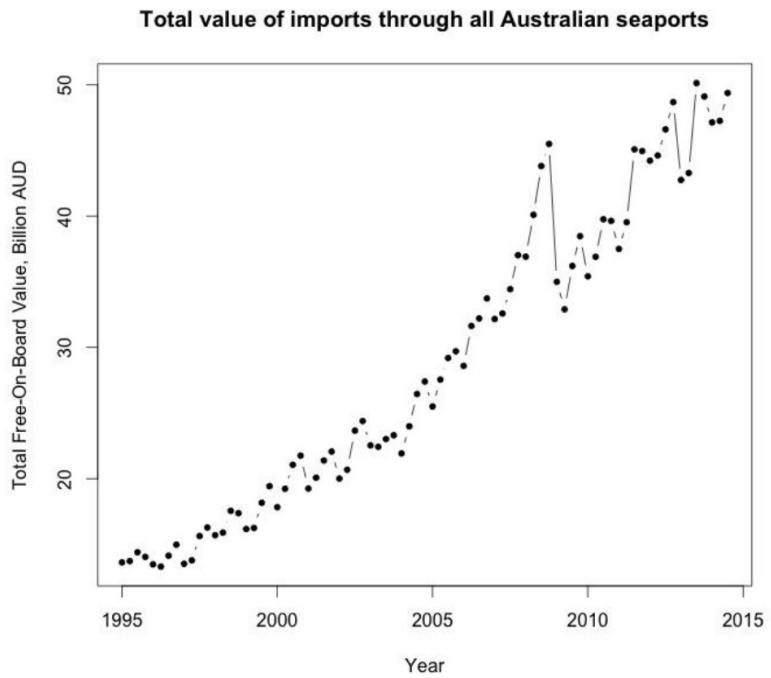
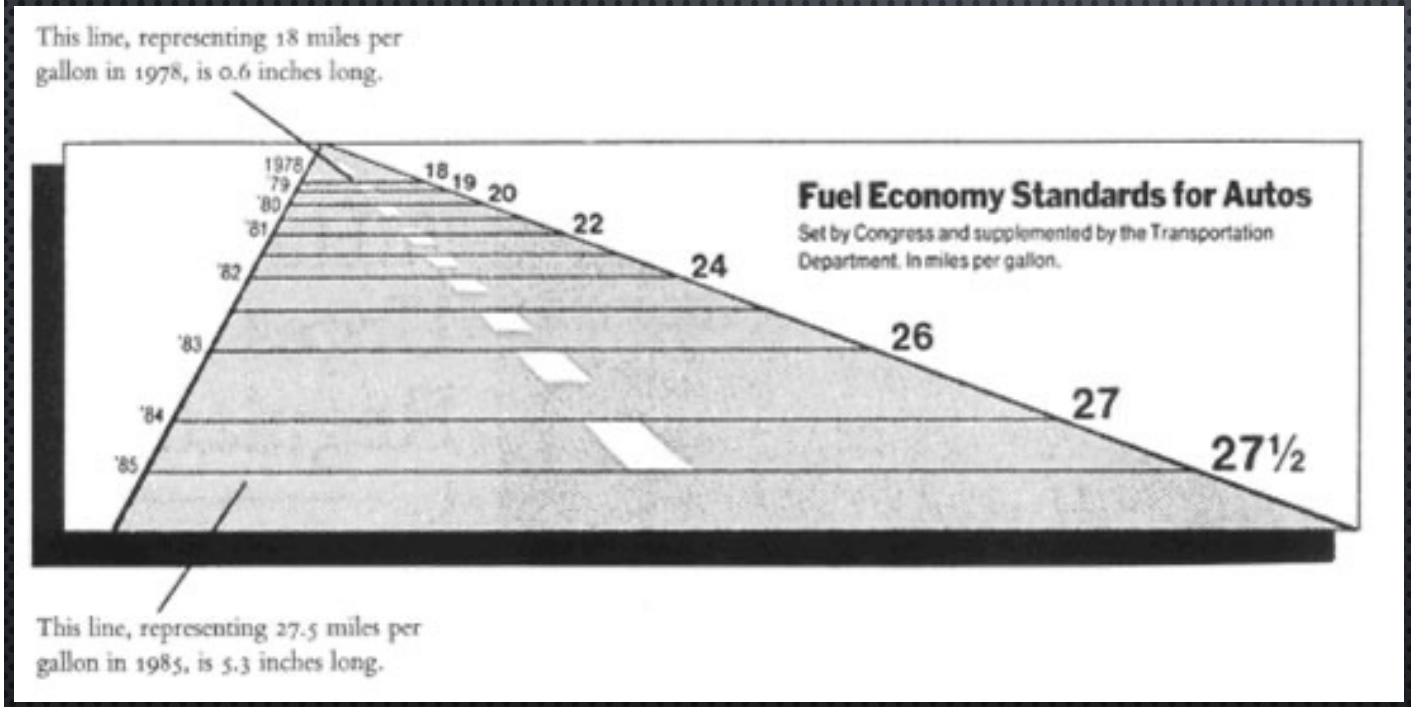
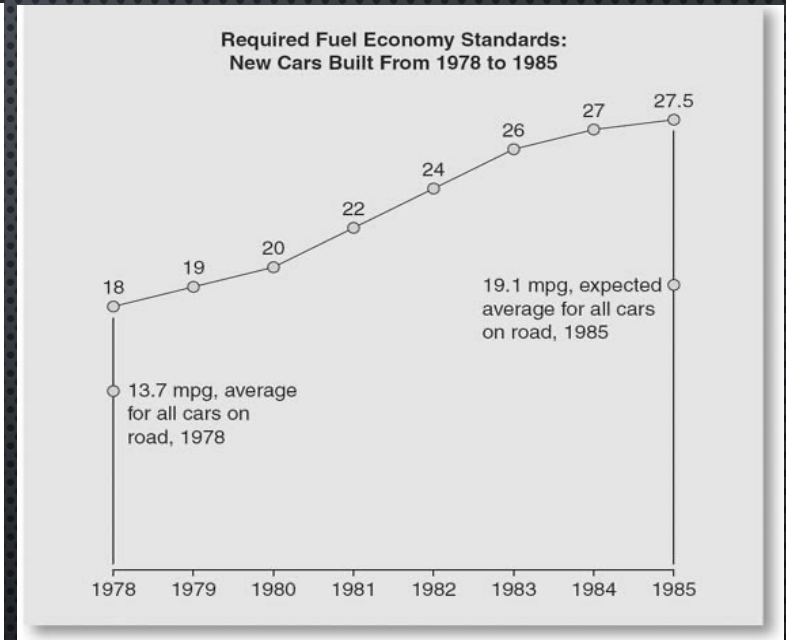


Figure 1. Total Free-On-Board value of Australian imports by sea through all ports.

1. SEASONALITY
2. AUTOCORRELATIONS
3. EXTERNAL SHOCKS

TIME SERIES DATA VISUALIZATION

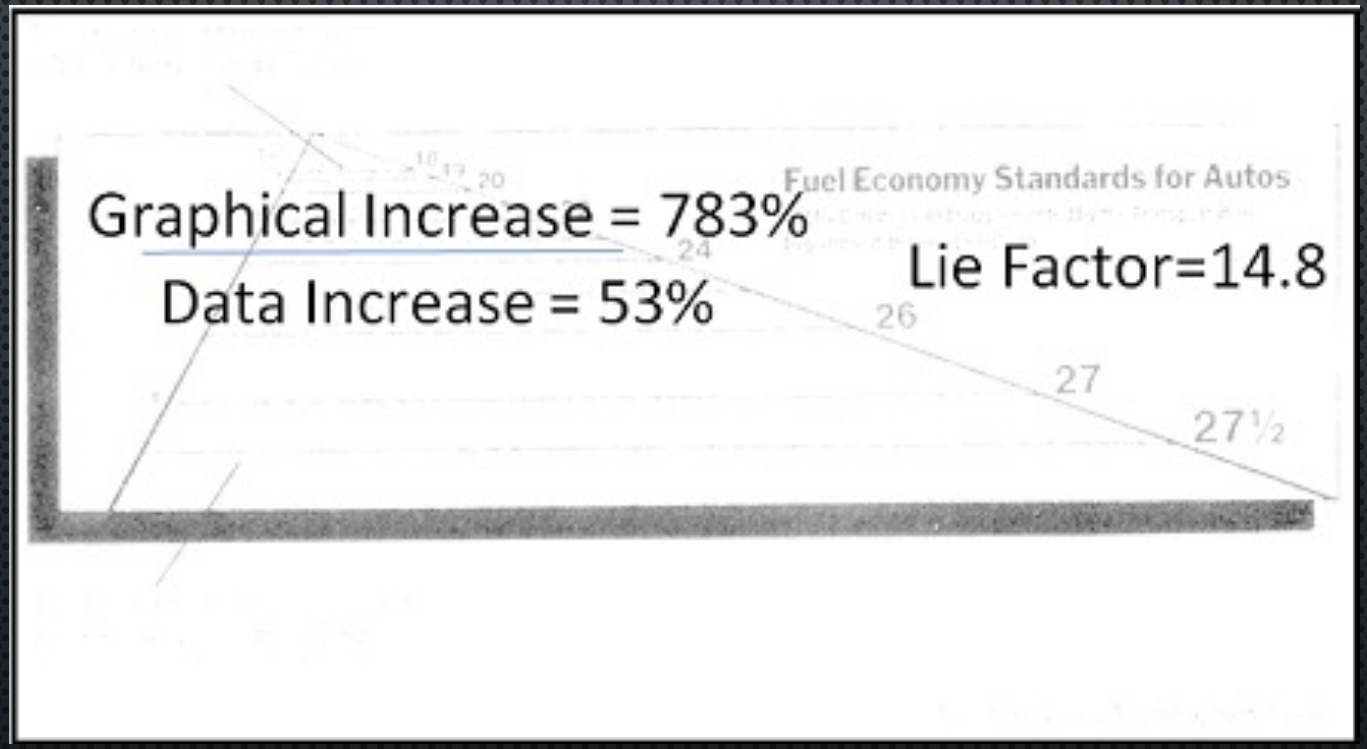




SOURCE: Tufte (2001). Used by permission from Graphics Press.

Design effect vs. data effect

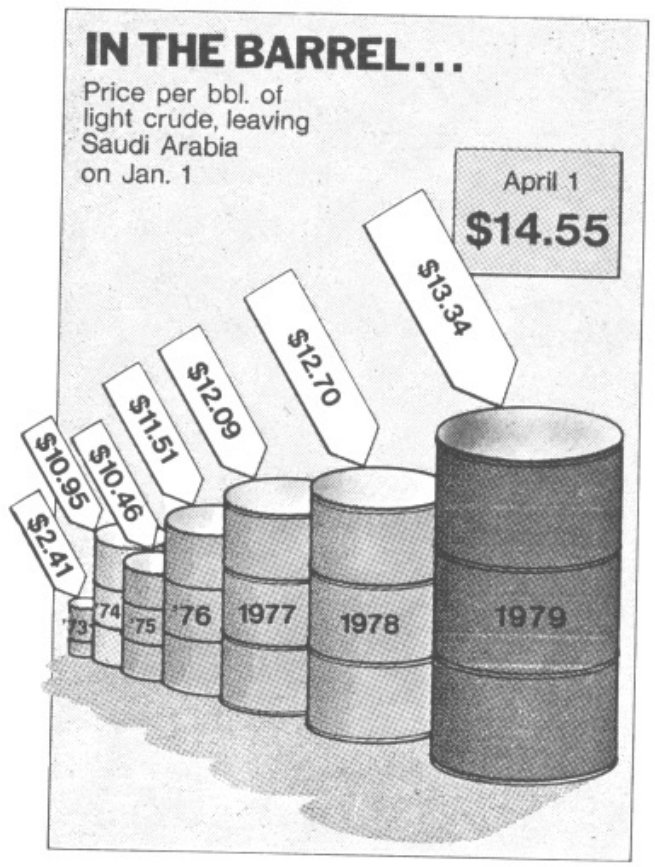
VISUALIZATIONS CAN BE MISLEADING



1. Area and volumes

2. Design effect vs. data effect

TIME SERIES DATA VISUALIZATION



1. Area and volumes

2. Design effect vs. data effect

3. The number of information carrying dimensions depicted should not exceed the number of dimensions of the data

SLOPE GRAPHS

1. slope-graph/table graphic
2. viewing architecture

SLOPE GRAPHS

1. https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0003nk
2. https://charliepark.org/slopegraphs/
3. Slope graphs in R
    1. https://www.r-bloggers.com/2018/06/creating-slopegraphs-with-r/
    1. https://github.com/ibecav/CGPfunctions
    1. https://github.com/leeper/slopegraph
4. Slope graphs in python
    1. https://dataviz.unhcr.org/tools/python/python_slope_chart.html

APPS FOR RAPID PROTOTYPING AND COMMUNICATION

- SHINY APPS FOR VISUALIZATION AND COMMUNICATION
- HTTPS://NEELSOUMYA.SHINYAPPS.IO/ACCIDENT_PREDICTION/

- OBSERVABLE (D3.JS)
  - HTTPS://OBSERVABLEHQ.COM

- REPRODUCIBLE ANALYSIS
- HTTPS://GITHUB.COM/NEELSOUMYA/TEACHING_REPRODUCIBLE_SCIENCE_R

WHY, WHEN AND HOW OF VISUALIZATION

- Visualization for diagnostics, picking models
- Visualization for data storytelling and for communication
- Visualization not just at the end of the data science pipeline but throughout
- Communicate well with people

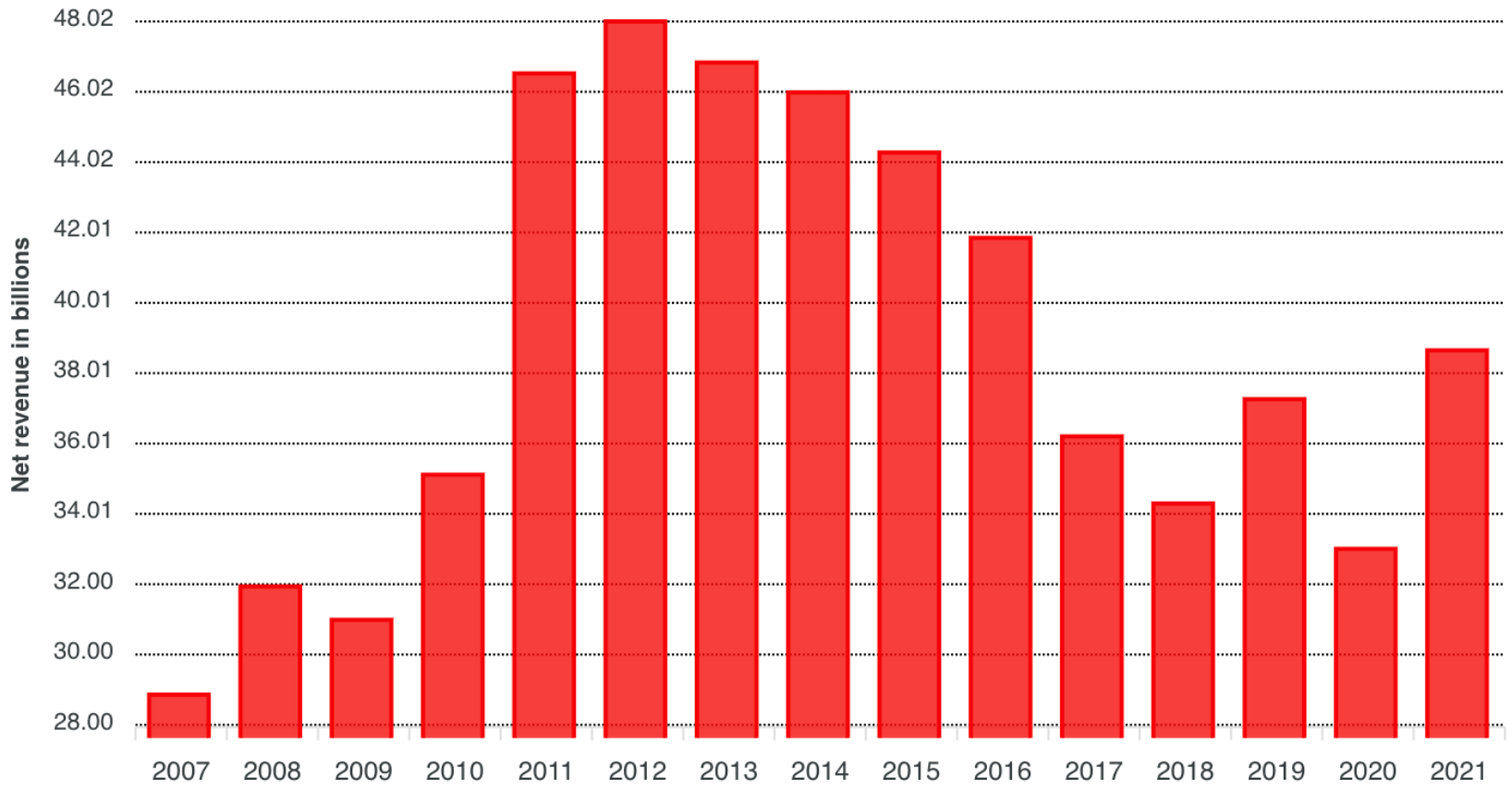
PRINCIPLES OF DATA VISUALIZATION (EDWARD TUFTE)

1. LESS IS MORE. “ABOVE ALL ELSE SHOW THE DATA” (***INK SPACE MAXIMIZE***)

2. KEEP IT PROPORTIONAL! “LIE FACTOR = SIZE OF EFFECT SHOWN IN GRAPHIC DIVIDED BY SIZE OF EFFECT IN DATA”

https://jeffhale.medium.com/five-takeaways-from-the-visual-display-of-quantitative-information-dd36dae35299

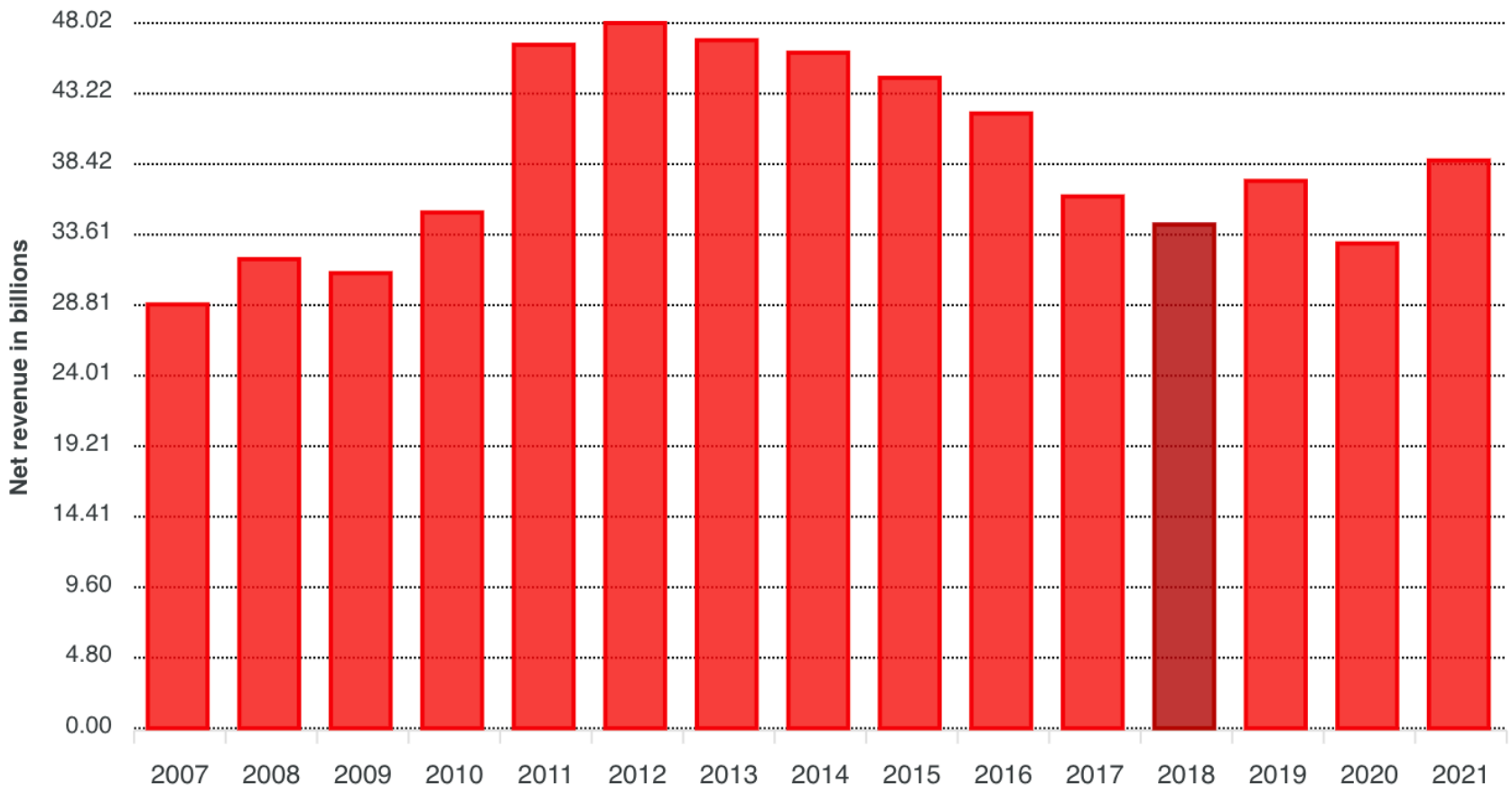
ACTIVITIES



Coca cola revenues

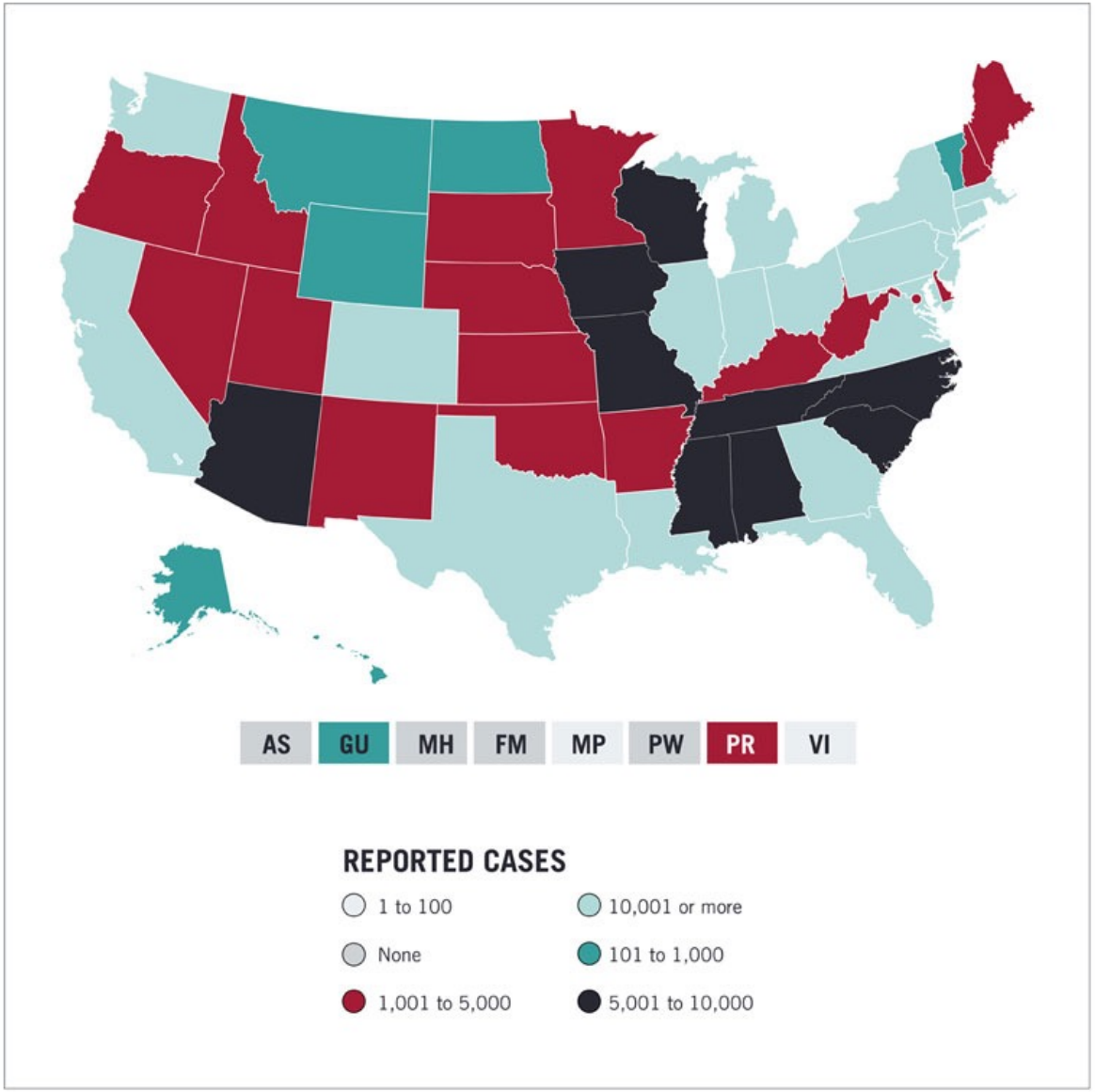
https://wpdatatables.com/misleading-data-visualization-examples/

ACTIVITIES



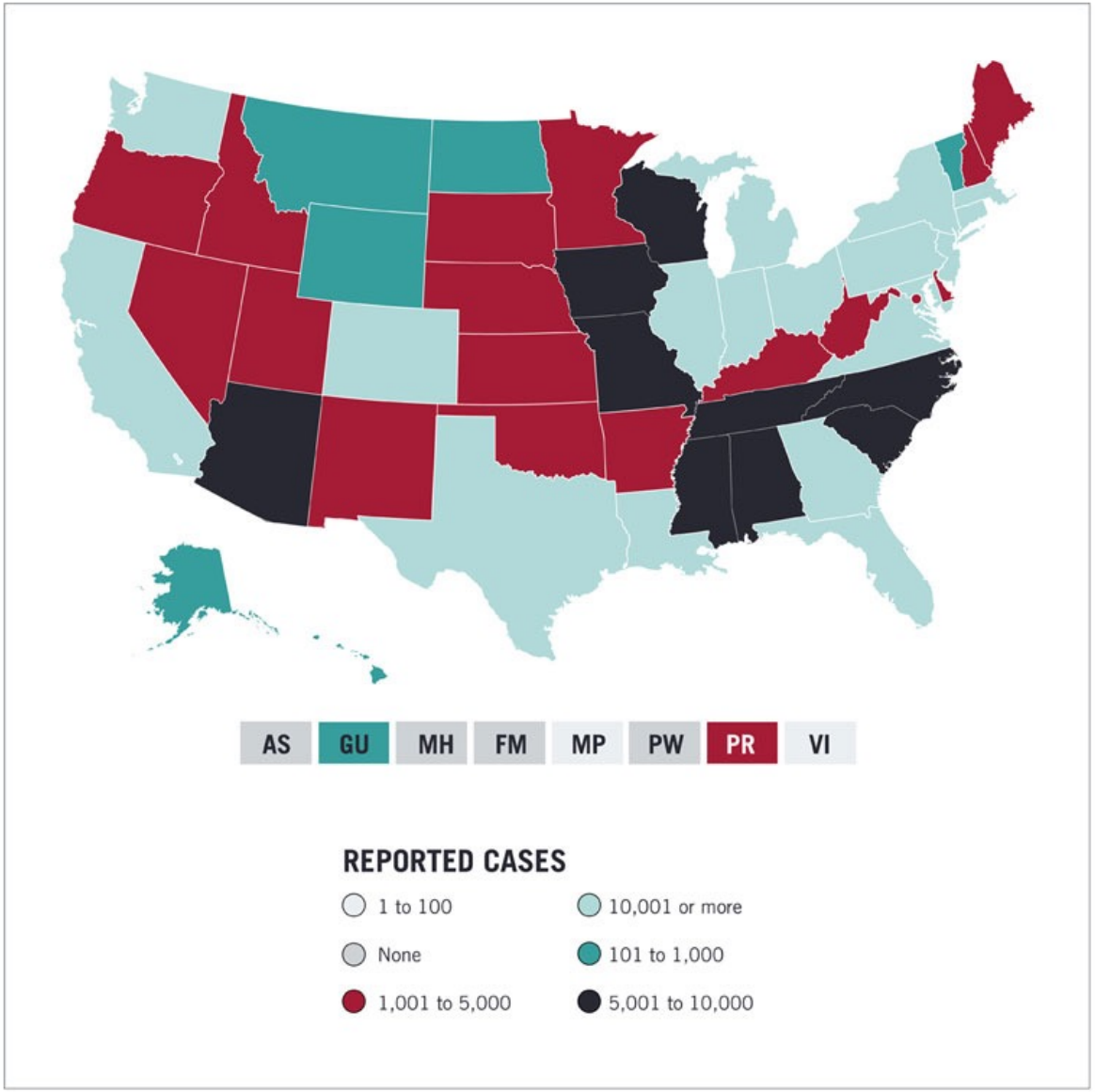
Coca cola revenues

https://wpdatatables.com/misleading-data-visualization-examples/



Color scales are important

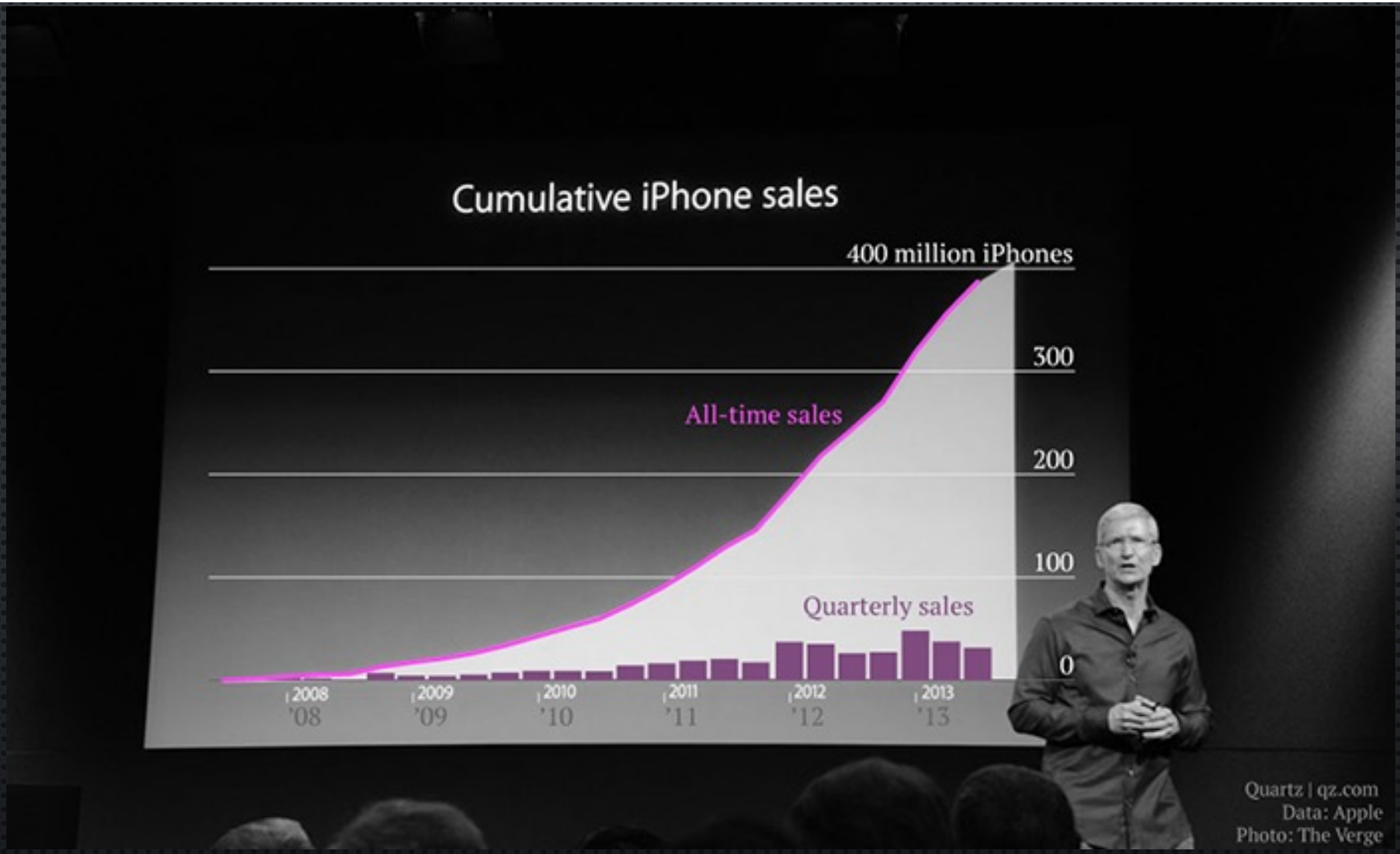
https://wpdatatables.com/misleading-data-visualization-examples/



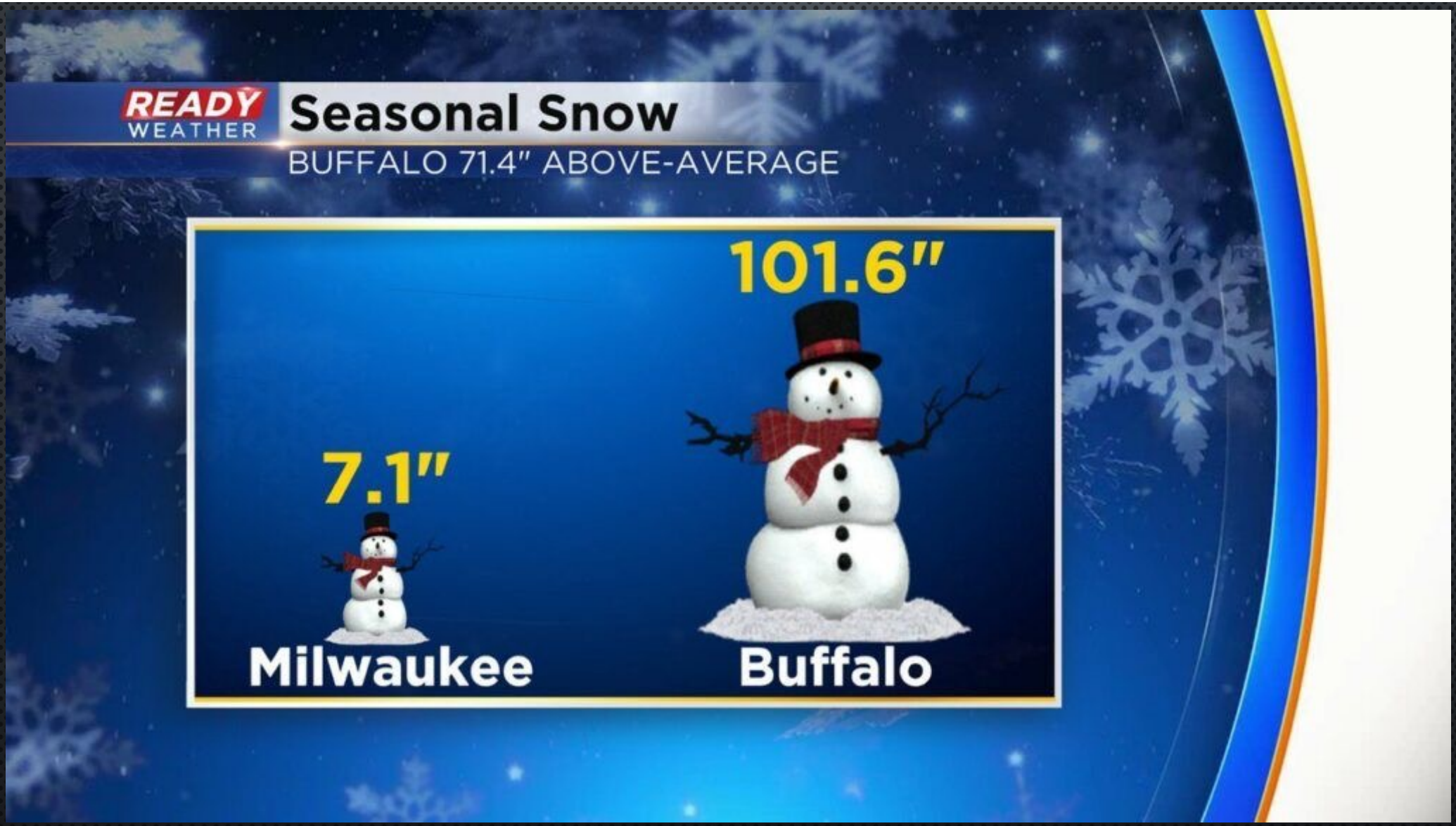
Color scales are important

https://wpdatatables.com/misleading-data-visualization-examples/

https://www.techjunkie.com/tim-cook-trying-prove-meaningless-chart/

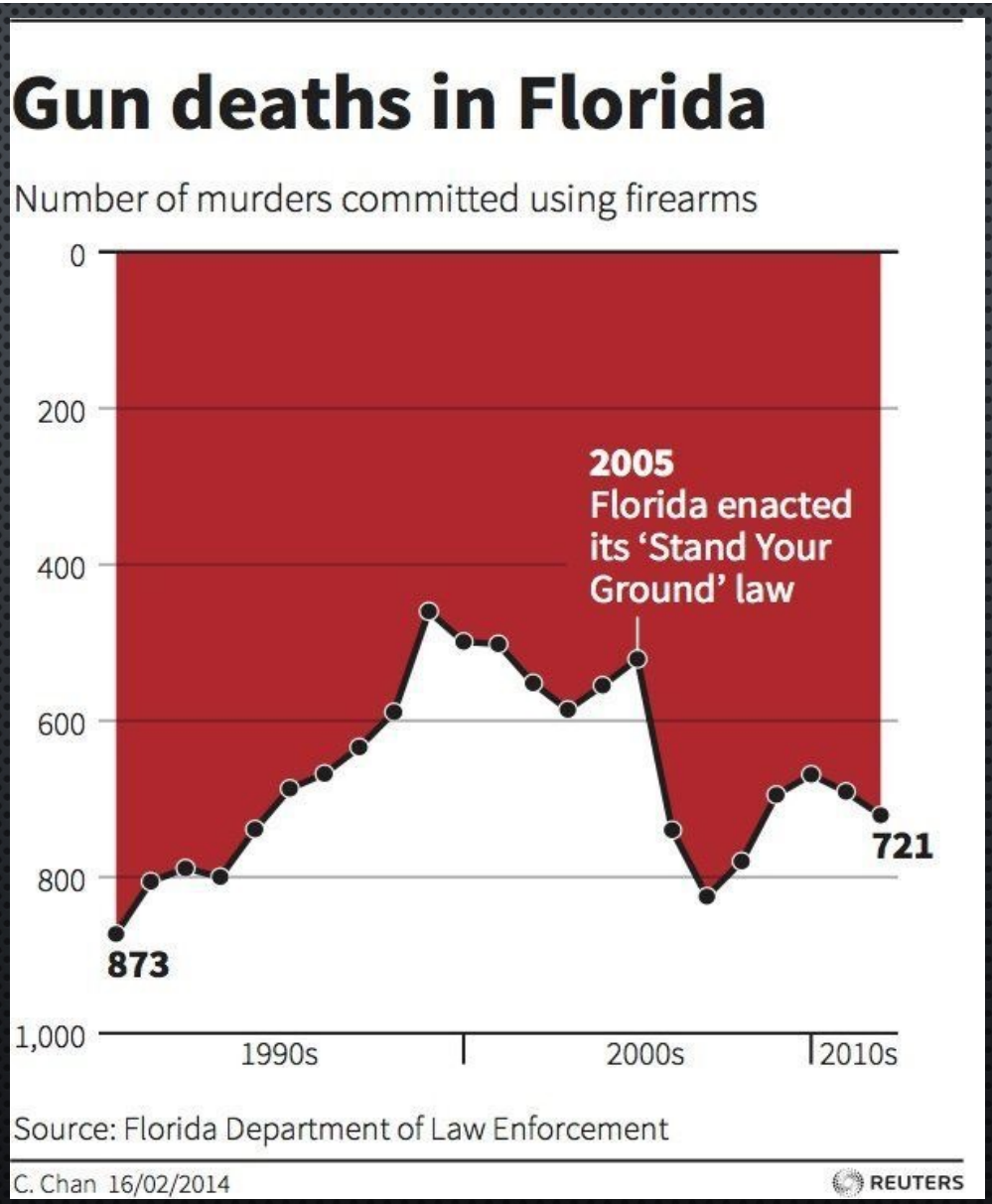


https://www.techjunkie.com/tim-cook-trying-prove-meaningless-chart/

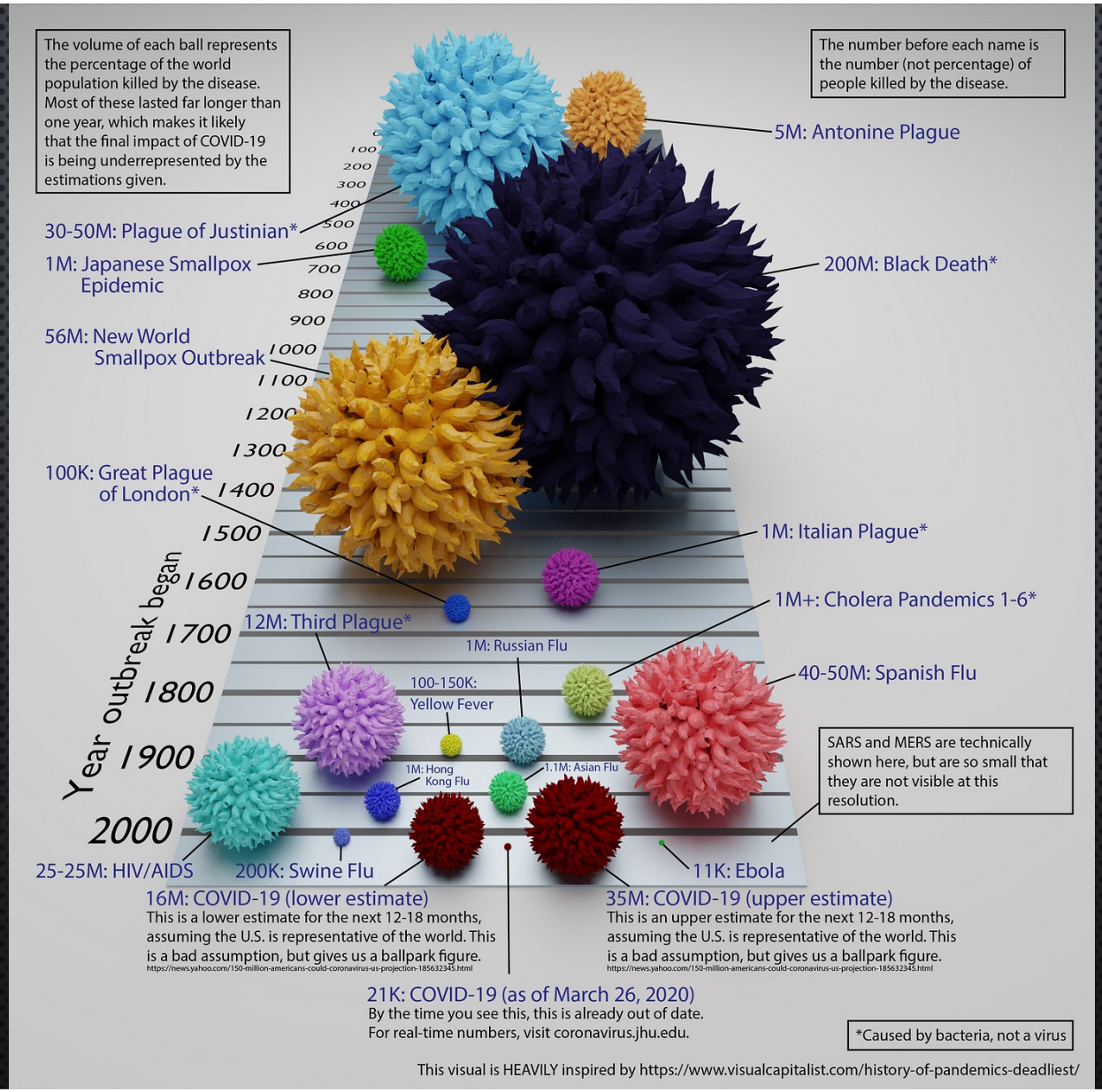


https://policyviz.com/2023/02/07/10-ways-to-mislead-with-data-visualization/

https://policyviz.com/2023/02/07/10-ways-to-mislead-with-data-visualization/



https://policyviz.com/2023/02/07/10-ways-to-mislead-with-data-visualization/



https://towardsdatascience.com/why-is-this-chart-bad-5f16da298afa

MATERIAL

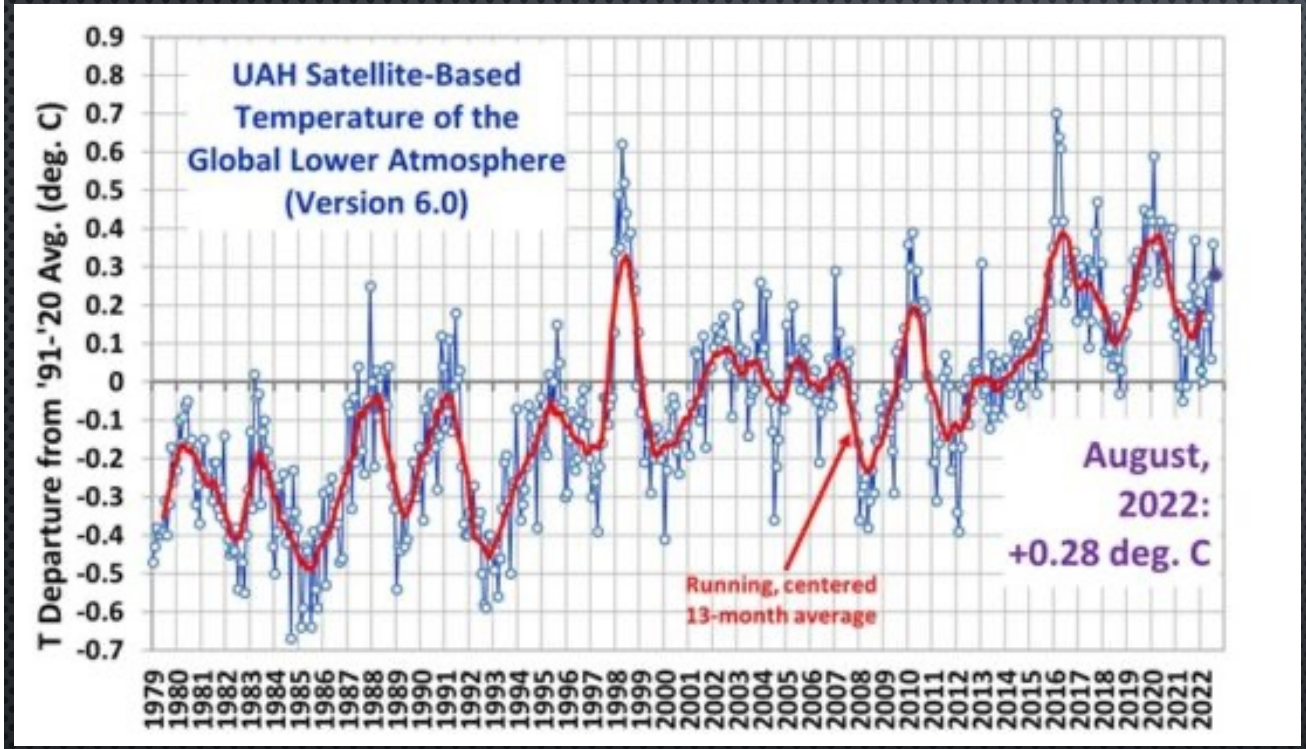
Material, slides, code, exercises, activities

https://github.com/neelsoumya/visualization_lecture

Derivations and technical details

https://github.com/neelsoumya/visualization_lecture/blob/main/mathematics_data_science.pdf

https://osf.io/mnh8d/



Steve Milloy
@JunkScience

August NASA satellite temps in.

No warming in almost 8 years -- despite lots more CO2.

August 2022 same temp as August 1998.

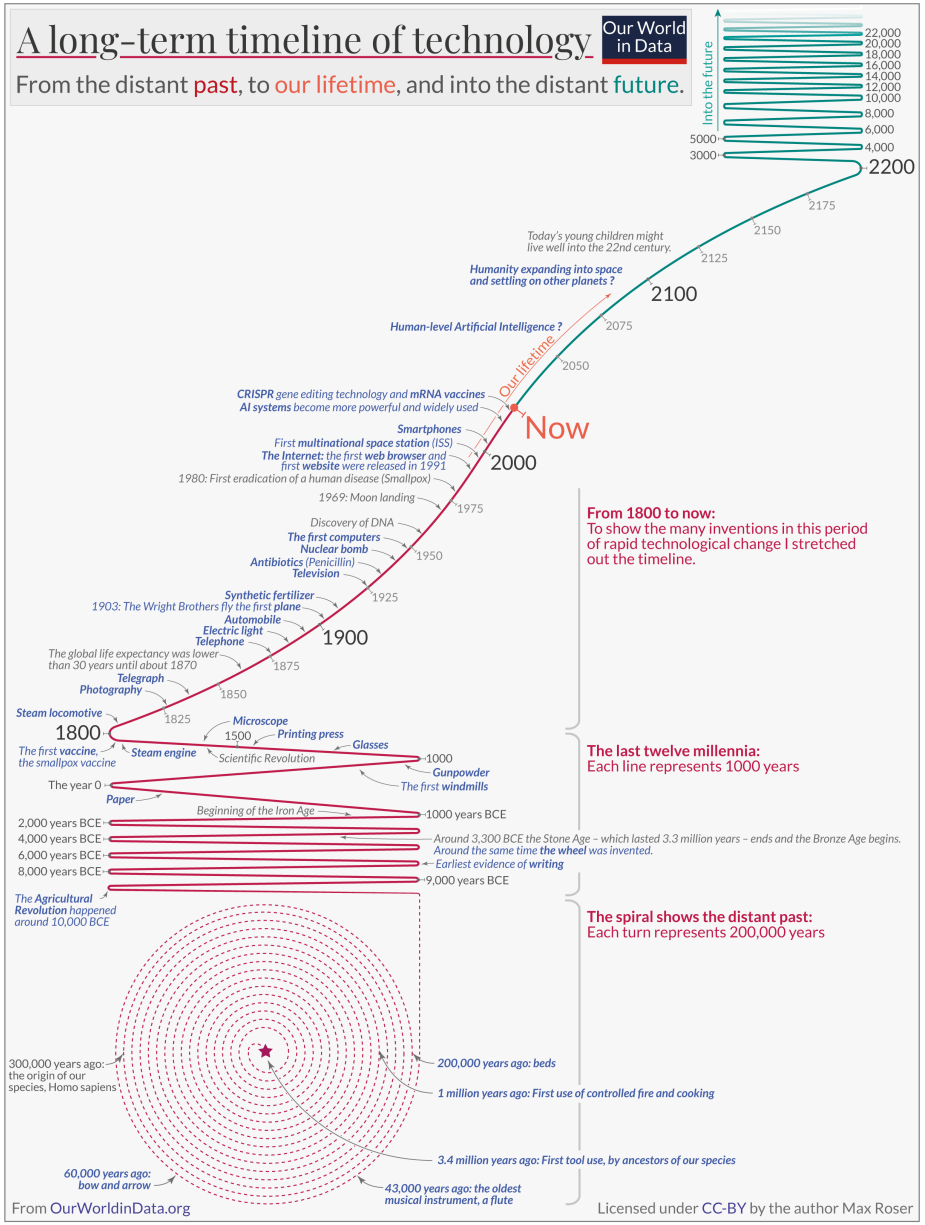
Climate is a hoax.

https://www.drroyspencer.com/2022/09/uah-global-temperature-update-for-august-2022-0-28-deg-c/

A long-term timeline of technology

Our World in Data

From the distant past, to our lifetime, and into the distant future.



From OurWorldinData.org

Licensed under CC-BY by the author Max Roser

https://ourworldindata.org/technology-long-run

ACTIVITIES

North korea missile range animation

Critique

https://nagix.github.io/nk-missile-tests/

INTERACTIVE DATA STORYTELLING

1. https://pudding.cool/projects/heat-records-map/

2. https://pudding.cool/2022/12/yard-sale/

3. https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1

ACTIVITIES

Create cartogram online

https://go-cart.io/tutorial

ACTIVITIES

CREATE INTERACTIVE DATA VISUALIZATION

HTTPS://WWW.GAPMINDER.ORG/TOOLS/#$CHART-TYPE=BUBBLES&URL=V1

HTTPS://OURWORLDINDATA.ORG/GRAPHER/ARTIFICIAL-INTELLIGENCE-TRAINING-COMPUTATION?TIME=EARLIEST..2023-10-11

ACTIVITIES

https://wpdatatables.com/data-visualization-examples/

https://wpdatatables.com/misleading-data-visualization-examples/

https://policyviz.com/2023/02/07/10-ways-to-mislead-with-data-visualization/

TIME SERIES DATA AND AUTOCORRELATIONS

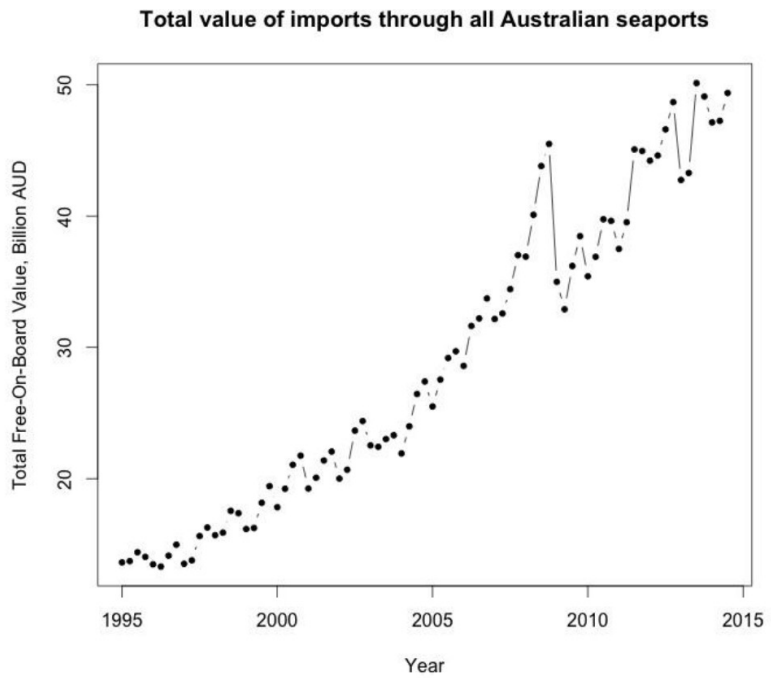


Figure 1. Total Free-On-Board value of Australian imports by sea through all ports.

1. SEASONALITY
2. AUTOCORRELATIONS

TIME SERIES DATA

Visualizing the data (for example, time-series data) can reveal what kinds of models would be appropriate. For example, if time series data has some seasonality, then a seasonal auto-regressive model (SARIMA) may be appropriate.

Visualization may also reveal if the underlying model/assumptions may have changed after a certain time. For example, in financial time-series data, there usually is a change after 2008 due to the global financial crisis.

This may suggest that a new model or more data is required.

TIME SERIES DATA

$$y_t = C + \sum_{i=1}^{p} \phi_i \cdot y_{t-i} - \sum_{i=1}^{q} \theta_i \cdot \epsilon_{t-i}$$

1. GENERALIZED LINEAR MODEL

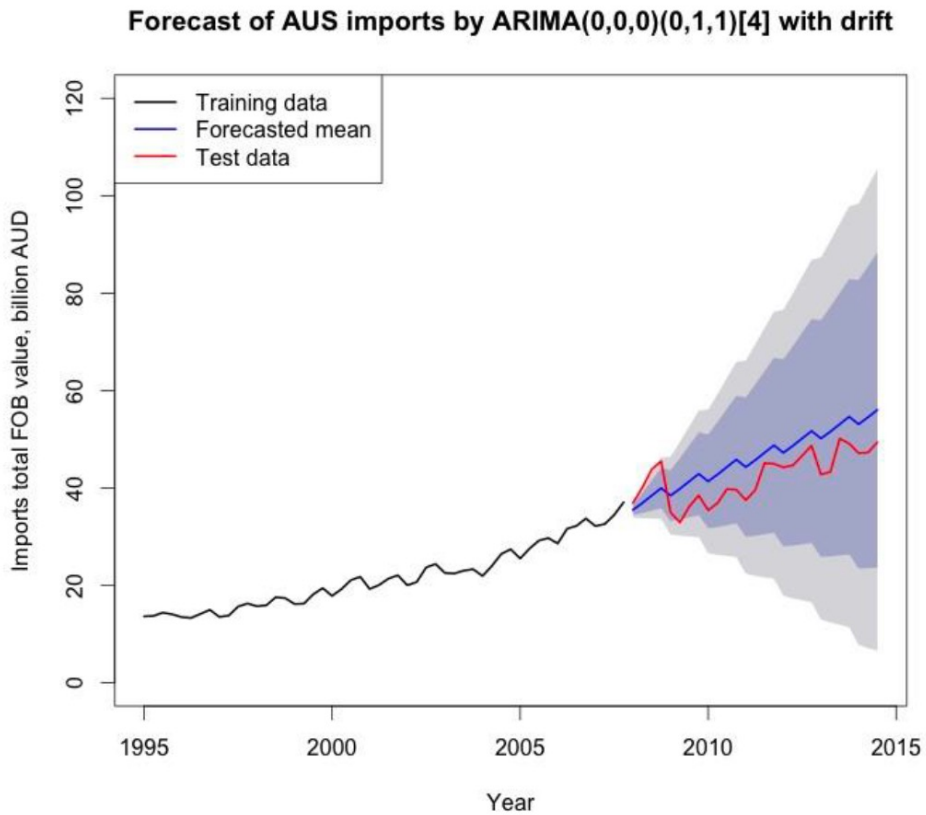


Figure 2. Forecast of port throughput (imports) by best fitting Seasonal ARIMA (SARIMA) model. In-sample (training set) AIC=37.66, out-of-sample (test set) Root Mean Squared Error RMSE=10190.

Vector Auto-Regressive (VAR) model:

This is a multivariate model, which is capable of modelling the joint dependencies between the throughput and the supporting GDP data, and uses these dependencies to forecast the imports along with the GDP in the future.

A p-th order VAR(p) is represented by the following equation:

$$\mathbf{y_t} = \mathbf{c} + \sum_{i=1}^{p} A_i \cdot \mathbf{y_{t-i}} + \epsilon_\mathbf{t}$$

The variables (port throughput and GDP of all countries) are subsumed in the vector $y_!$, and $y_{!!!}$ is the ith lag of $y_!$. The coefficient matrices $A_!$ are time-invariant and represent a set of model parameters, ϵ is a vector of error terms with mean 0 and covariance Σ, and c is a vector of constant intercept terms. Fitting the VAR model involves estimating the matrix of interactions A_i, vector c and the covariance matrix Σusing the training data.

In this work, two different types of VAR model classes are considered: a two dimensional class which incorporates the imports time series, and the GDP of Australia only; and a seven dimensional model class that in addition to the port throughput data and Australian GDP

DESIGN MATRIX

GLM
CORRELATED PERIODIC

EXPONENTIAL
HIOW TO PICK
HOW DO I GET TO THIS OUT OF DATA
HOW DO I COMMUNICATE THIS TO STAKEHOLDERS

MORE TOOLS FOR DATA VISUALIZATION
GENERATE DATA
VISUALIZE IT
HIGH DUMENOSIONAL DATA
RESULTS VISUALIZE
COMMUNICATE THIS

CANCER WORING LINK WINTER ENVIRONMENTAL ACTIIVIS RACHEL CARSON SILENT SPRING
PITFALLS
PROB VISUALIZE

UN LOGO

TIME SERIES DATA

ASSUMPTIONS

SIMILAR TO LINEAR REGRESSION/GLM (VAR MODEL/ARIMA)

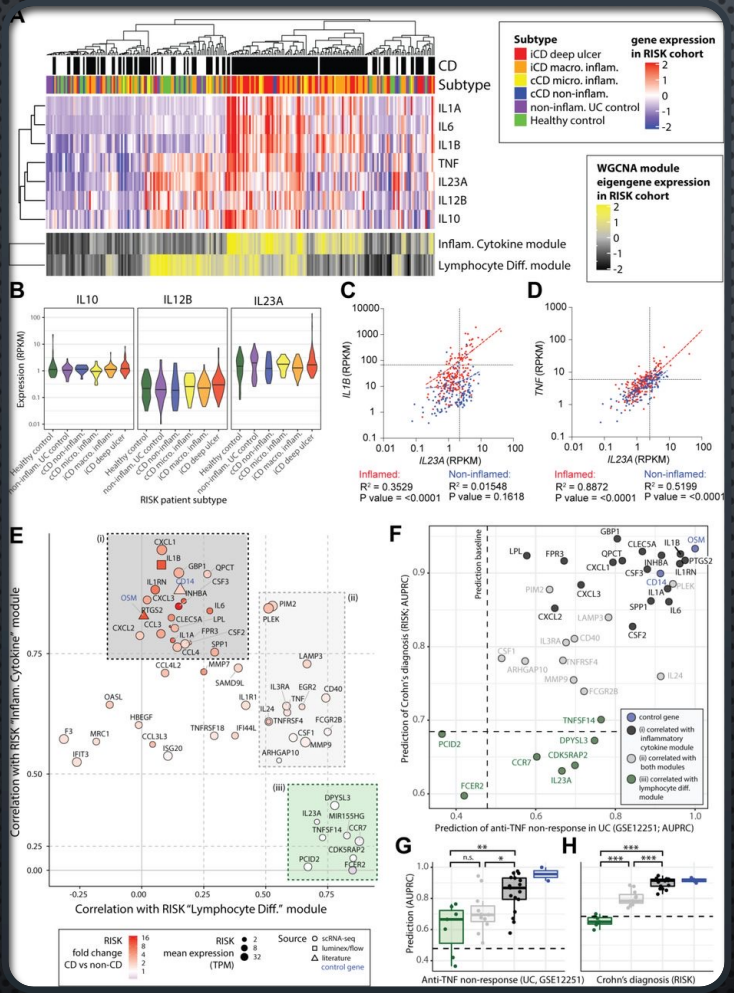
LOOK AT CORRELATIONS

SIMPLER MODELS ARE BETTER (MOVING AVERAGES BETTER THAN ANYTHING FANCY/ML)

DATA BIAS

Gorilla experiment (selective attention bias)

Hermann Hesse quote seeking is different to finding

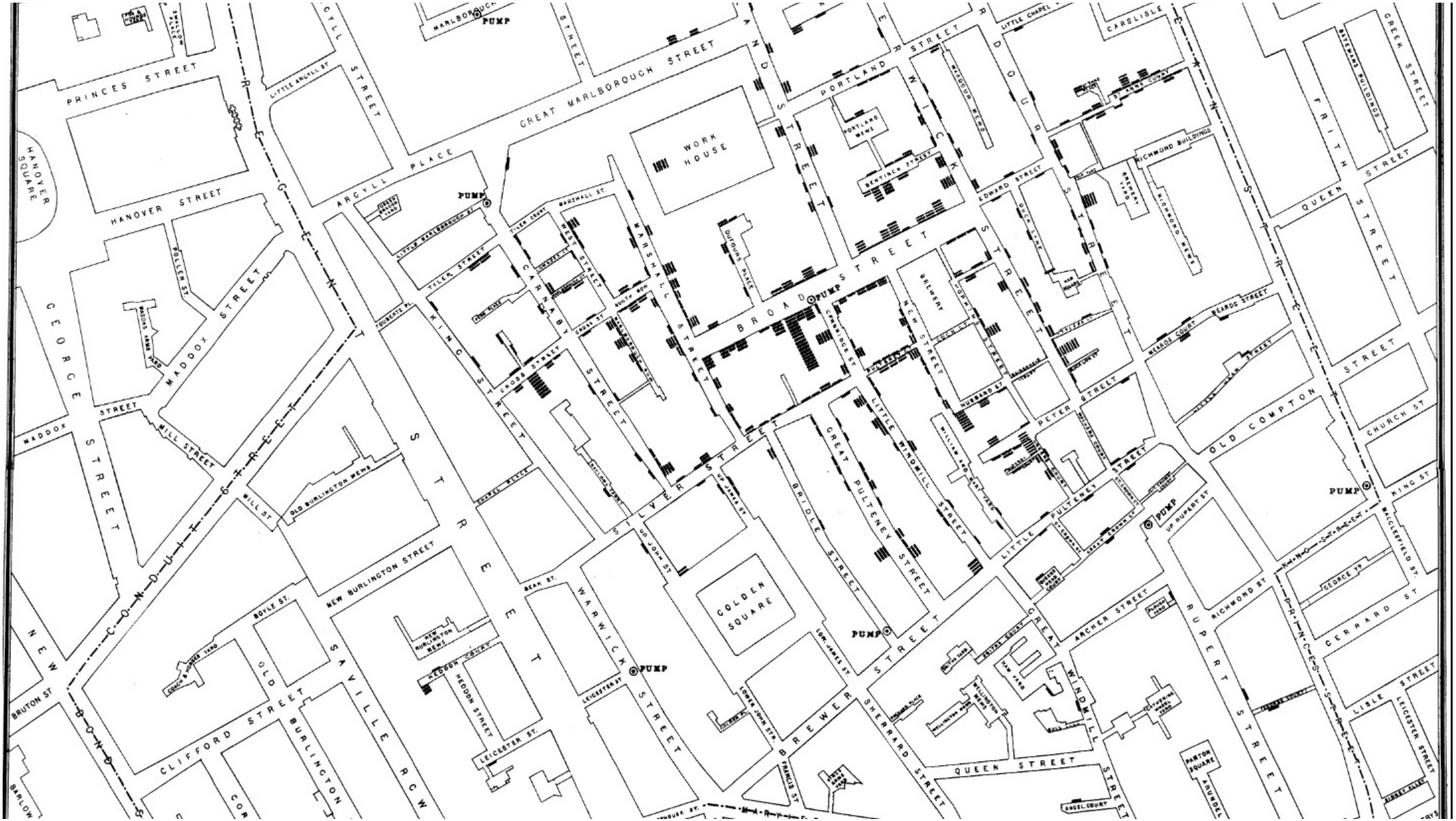

A
CD
Subtype
IL1A
IL6
IL1B
TNF
IL23A
IL12B
IL10
Inflam. Cytokine module
Lymphocyte Diff. module
Subtype
iCD deep ulcer
iCD macro. inflam.
cCD micro. inflam.
cCD non-inflam.
non-inflam. UC control
Healthy control
gene expression in RISK cohort
WGCNA module eigengene expression in RISK cohort
B
IL10
IL12B
IL23A
Expression (RPKM)
RISK patient subtype
C
IL1B (RPKM)
IL23A (RPKM)
Inflamed:
R² = 0.3529
P value = <0.0001
Non-inflamed:
R² = 0.01548
P value = 0.1618
D
TNF (RPKM)
IL23A (RPKM)
Inflamed:
R² = 0.8872
P value = <0.0001
Non-inflamed:
R² = 0.5199
P value = <0.0001
E
Correlation with RISK "Inflam. Cytokine" module
Correlation with RISK "Lymphocyte Diff." module
RISK fold change CD vs non-CD
RISK mean expression (TPM)
Source
scRNA-seq
luminex/flow
literature
control gene
F
Prediction baseline
Prediction of Crohn's diagnosis (RISK; AUPRC)
Prediction of anti-TNF non-response in UC (GSE12251; AUPRC)
control gene
(i) correlated with inflammatory cytokine module
(ii) correlated with both modules
(iii) correlated with lymphocyte diff. module
G
Prediction (AUPRC)
Anti-TNF non-response (UC, GSE12251)
H
Crohn's diagnosis (RISK)

CASE STUDY

BAD VISUALIZATIONS

BASICS AND PITFALLS IN DATA VISUALIZATION

1. KNOW YOUR AUDIENCE
2. PICK VISUALIZATION BASED ON AUDIENCE
3. VISUALIZE DATA AND THEN PICK MODELS
4. ADD NARRATIVE


PRINCES STREET
HANOVER SQUARE
HANOVER STREET
GEORGE STREET
MADDOX STREET
MILL STREET
NEW BURLINGTON STREET
BOYLE ST.
OLD BURLINGTON STREET
CLIFFORD STREET
BURLINGTON
SAVILLE ROW
BRUTON ST
BARLOW
ARGYLL PLACE
ARGYLL STREET
GREAT MARLBOROUGH STREET
MARLBOROUGH
PUMP
WORK HOUSE
BROAD STREET
PORTLAND STREET
PORTLAND MEWS
BENTINCK STREET
WELLS STREET
EDWARD STREET
BREWERY
NEW STREET
MARSHALL STREET
CARNABY STREET
KINGLY STREET
GREAT PULTENEY STREET
LITTLE PULTENEY STREET
BRIDLE STREET
LITTLE WINDMILL STREET
GREAT WINDMILL STREET
PETER STREET
HUSBAND ST
WARWICK STREET
GOLDEN SQUARE
BREWER STREET
SHERRARD STREET
QUEEN STREET
ANGEL COURT
ARCHER STREET
RICHMOND ST
RUPERT STREET
PRINCES STREET
MACCLESFIELD ST.
OLD COMPTON STREET
QUEEN STREET
FRITH STREET
CHURCH ST
KING ST
GERRARD ST
LISLE STREET
LEICESTER STREET
CARLISLE
GREEK STREET
RICHMOND BUILDINGS
RICHMOND MEWS
NEEDOE COURT
NEEDOE STREET
HEDDON STREET
LEICESTER ST.
BEAK ST.
UP JOHN ST.
LOWER JOHN ST.
FRANCIS ST
UP RUPERT ST
PARTON SQUARE
ARUNDEL
GEORGE YD
CHURCH PLACE
CHOBB STREET
TILLER STREET
TILLER COURT
MARSHALL ST.
DUFOURS PLACE
CARNABY STREET
SOUTH MEWS
CROSS ST
KINGLY STREET
FOUBERT'S PL.
POLLEN ST
MADDOX STREET
OLD BURLINGTON MEWS
NEW BURLINGTON MEWS
COACH & HORSES YARD
LITTLE ARGYLL ST.
LITTLE MARLBOROUGH ST.
LITTLE CHAPEL
ST ANNE'S COURT
BASTERBAS BUILDINGS
WARDOUR MEWS
WARDOUR STREET
DUCK LANE
NEW COURT
BREWERY YARD
RICHMOND MEWS
NEEDOE COURT
NEEDOE STREET
WILLIAM AND MARY YARD
HUSBAND ST
PETER STREET
LITTLE PULTENEY
CARPENTER ST
BEAK COURT
PULTENEY COURT
PULTENEY STREET
OLD COMPTON STREET
GEORGE YD
CHURCH PLACE
BRIDLE COURT
BRIDLE YARD
NEW YARD
WELLINGTON MEWS
WELLINGTON YARD
BELL YARD
ARCHER STREET
PUMP
GREAT WINDMILL STREET
TELEVISION WORKS YARD
PARTON SQUARE
ARUNDEL
LISLE STREET
LEICESTER STREET
PRINCES STREET

ASSUMPTIONS

Linearity (linear relationship between data points and lower dimensional representation)

Loss function/reconstruction error (squared loss)

Uses the dot product (one type of inner product)

TIME SERIES DATA

Visualizing the data (for example, time-series data) can reveal what kinds of models would be appropriate.

For example, if time series data has some seasonality, then a seasonal auto-regressive model (SARIMA) may be appropriate.

$$y_t = C + \sum_{i=1}^{p} \phi_i \cdot y_{t-i} - \sum_{i=1}^{q} \theta_i \cdot \epsilon_{t-i}$$

Visualization may also reveal if the underlying model/assumptions may have changed after a certain time. For example, in financial time-series data, there usually is a change after 2008 due to the global financial crisis.

This may suggest that a new model or more data is required.

1. GENERALIZED LINEAR MODEL

```
> dfbeta(xmdl)
  (Intercept)         age
1  -3.3645662  0.06437573
2  -1.6119656  0.02736278
3   1.5481303 -0.01456709
4  -0.0259835  0.05092767
5   0.8707699 -0.06479736
6   1.8551808 -0.06622744
```

For each coefficient of your model (including the intercept), the function gives you the so-called DFbeta values. These are the values with which the coefficients have to be adjusted if a particular data point is excluded (sometimes called "leave-one-out diagnostics"). More concretely, let's look at the age column in the data frame above. The first row means that the coefficient for age (which, if you remember, was -0.9099) has to be adjusted by 0.06437573 if data point 1 is excluded. That means that the coefficient of the model without the data point

ABSENCE OF INFLUENTIAL DATA POINTS

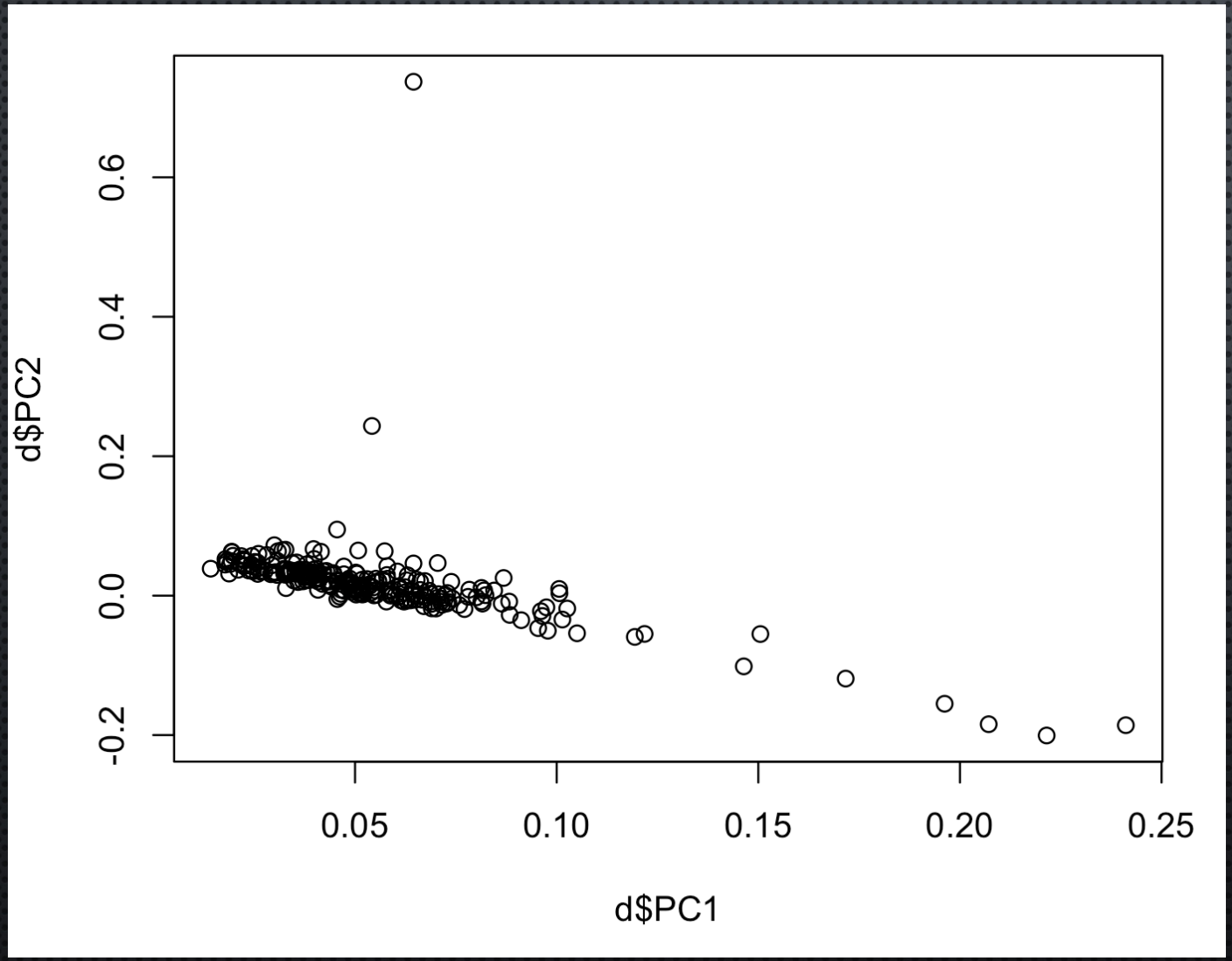
1. RUN THE ANALYSIS *WITH* AND *WITHOUT* INFLUENTIAL DATA POINTS

CASE STUDIES

REMOVING OUTLIERS IN GENOMIC DATA USING PCA.

FREQUENTLY IN GENOMIC DATA WE MAY HAVE TO REMOVE OUTLIERS. THESE OUTLIERS MAY BE DUE TO TECHNICAL/BATCH EFFECTS OR UNKNOWN REASONS NOT CONNECTED TO BIOLOGY.

THIS HAS IMPLICATIONS FOR ANY TESTS PERFORMED DOWNSTREAM. FOR EXAMPLE, T-TESTS CAN BE PERFORMED DOWNSTREAM AFTER PERFORMING PCA. IF THERE ARE OUTLIERS, IT MAY AFFECT THE RESULTS OF THE T-TEST. APPLICATION TO BULK AND SINGLE-CELL SEQUENCING DATA.



ABSENCE OF INFLUENTIAL DATA POINTS

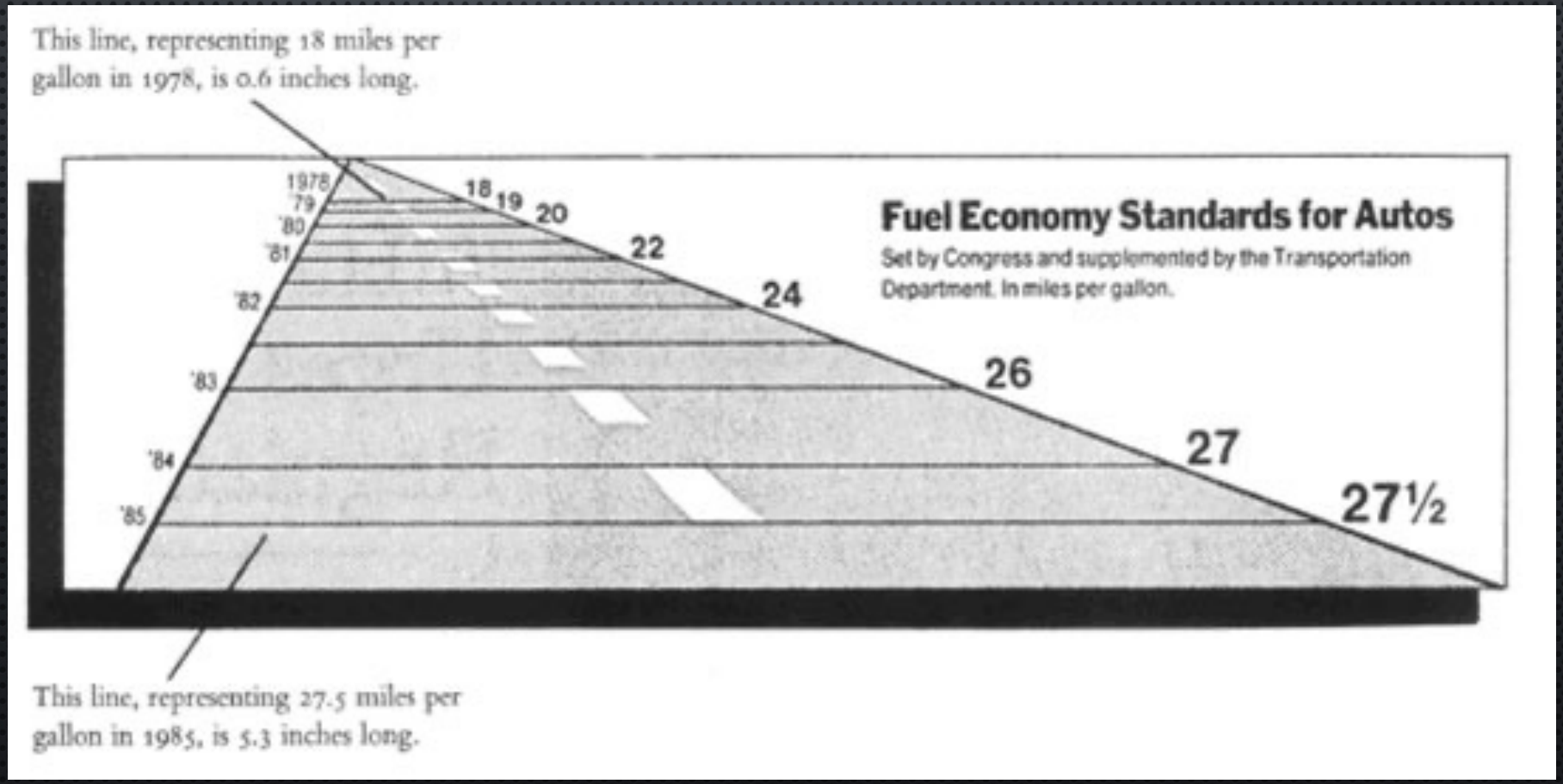
1. Removing outliers and then some downstream processing

2. t-test

3. Run the analysis and without and with the data point

Study 1			Study 2		
Subject	**Sex**	**Voice.Pitch**	**Subject**	**Age**	**Voice.Pitch**
1	female	233 Hz	1	14	252 Hz
2	female	204 Hz	2	23	244 Hz
3	female	242 Hz	3	35	240 Hz
4	male	130 Hz	4	48	233 Hz
5	male	112 Hz	5	52	212 Hz
6	male	142 Hz	6	67	204 Hz

INDEPENDENCE

1. Study design
2. Mixed effects models

VISUALIZATIONS CAN BE MISLEADING