

VISUALIZATION: A WHIRLWIND TOUR

SOUMYA BANERJEE

VISUALIZATION AS DEBUGGING

1. DATA SCIENCE AS (VISUAL) DEBUGGING
2. DATA STORYTELLING
3. WORK CLOSELY WITH DOMAIN EXPERTS

THE BEST STATISTICAL GRAPHIC EVER

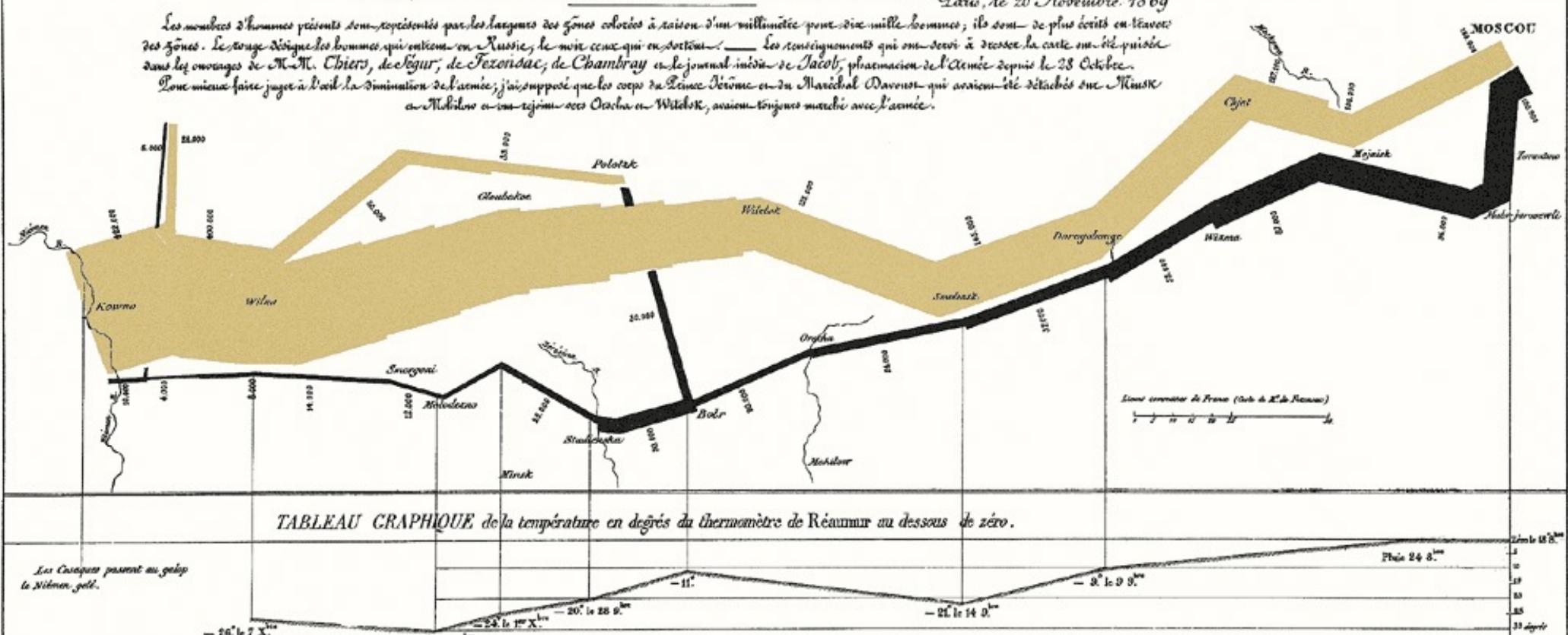
Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

Dessiné par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres sur ces zones. Le rouge désigne les hommes qui restent en Russie; le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte sont tirés dans les ouvrages de M. M. Chiers, de Léger, de Tocozac, de Chambray et le journal intime de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

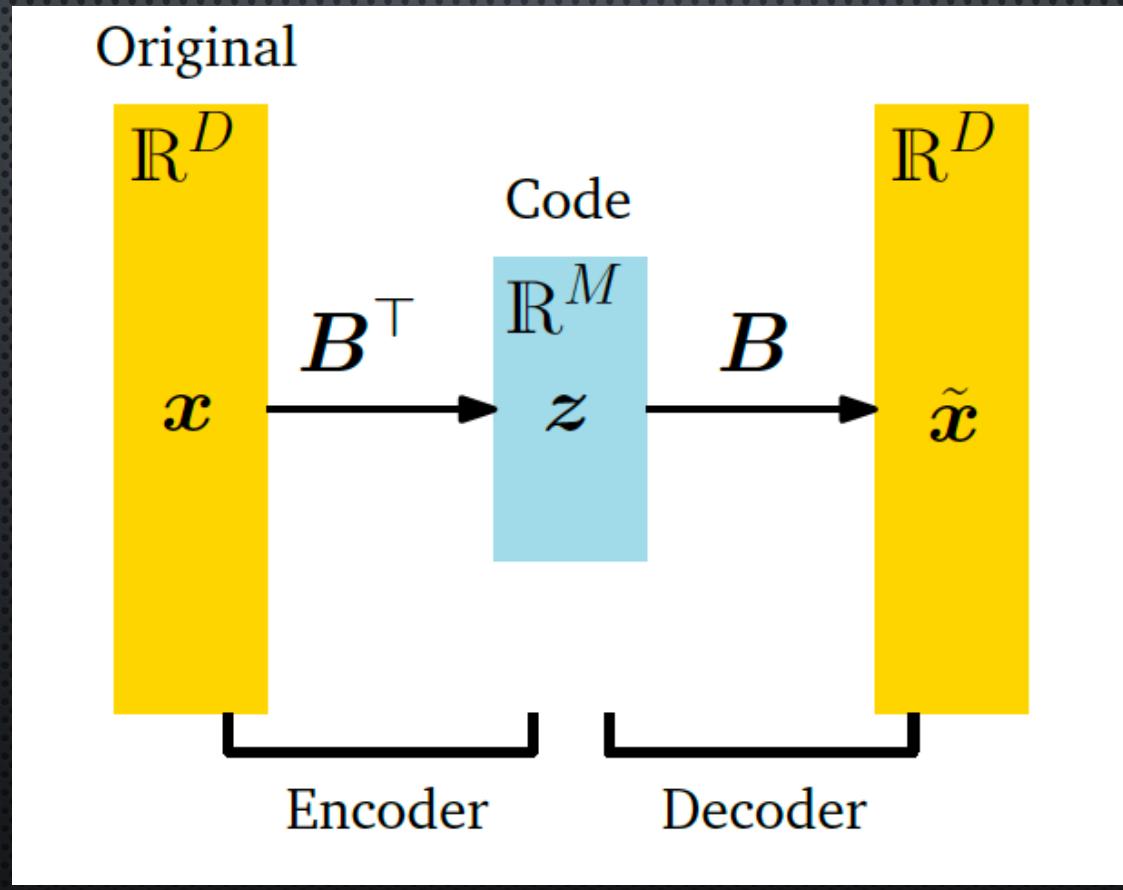
Pour mieux faire juger à quel point la diminution de l'armée, j'ai supposé que les corps de l'armée détruite en du Maréchal Davout qui avaient été détachés sur Nischni Novgorod et qui rejoignirent Orléans en Witebsk, avaient toujours marché avec l'armée.



DATA VISUALIZATION

1. COMMUNICATE
2. DATA STORYTELLING
3. CASE STUDIES (TSNE)
 - A. STOCHASTIC
 - B. DISTANCES NOT PRESERVED
 - C. DIFFICULT TO COMMUNICATE TO NON-TECHNICAL EXPERTS

IMPORTANT CONCEPT



GENERALIZATIONS OF THIS IDEA

TSNE

AUTOENCODER (NON-LINEAR LOSS FUNCTION)

PITFALLS

TSNE CAN BE USED FOR HYPOTHESIS GENERATION. THERE ARE MANY PITFALLS TO THIS.

[HTTPS://DISTILL.PUB/2016/MISREAD-TSNE/](https://distill.pub/2016/misread-tsne/)

SOME PITFALLS: DISTANCES NOT PRESERVED. FOR EXAMPLE, A 2D MAP IS A PROJECTION FROM 3D

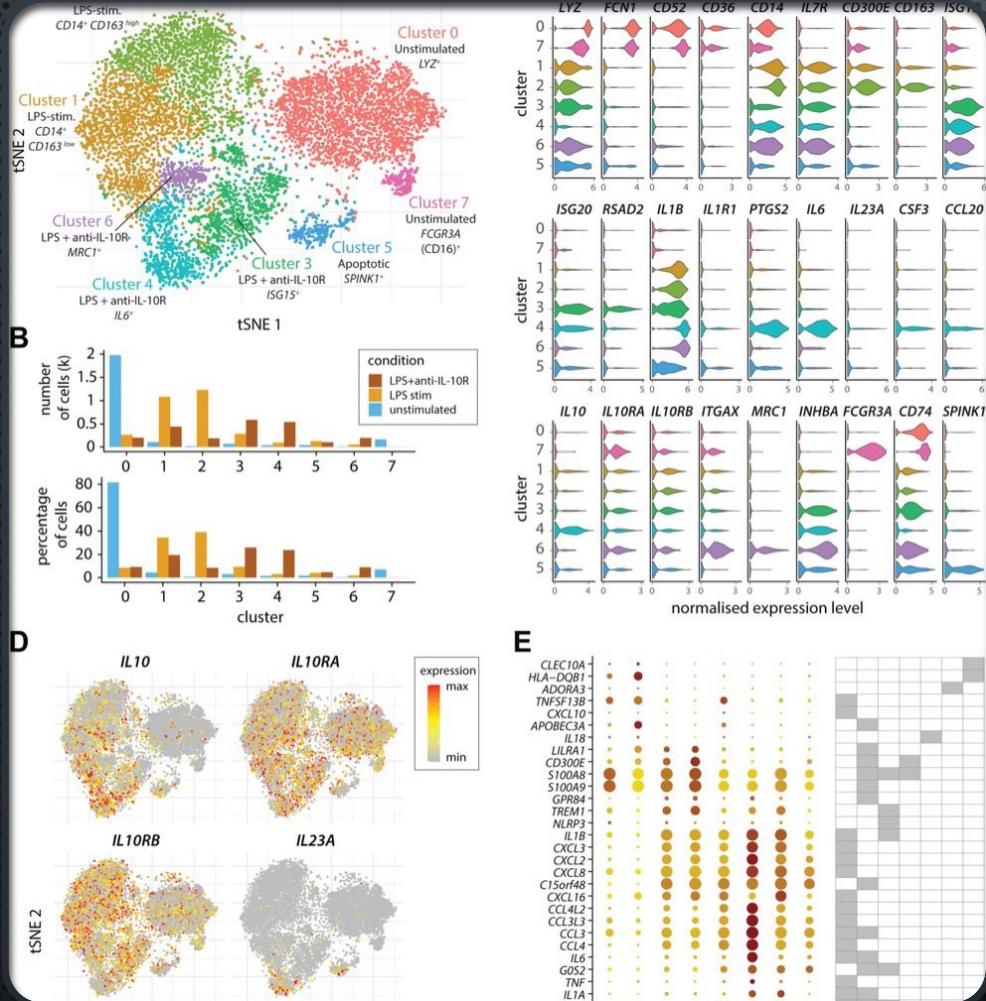
CASE STUDIES

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.

CASE STUDY



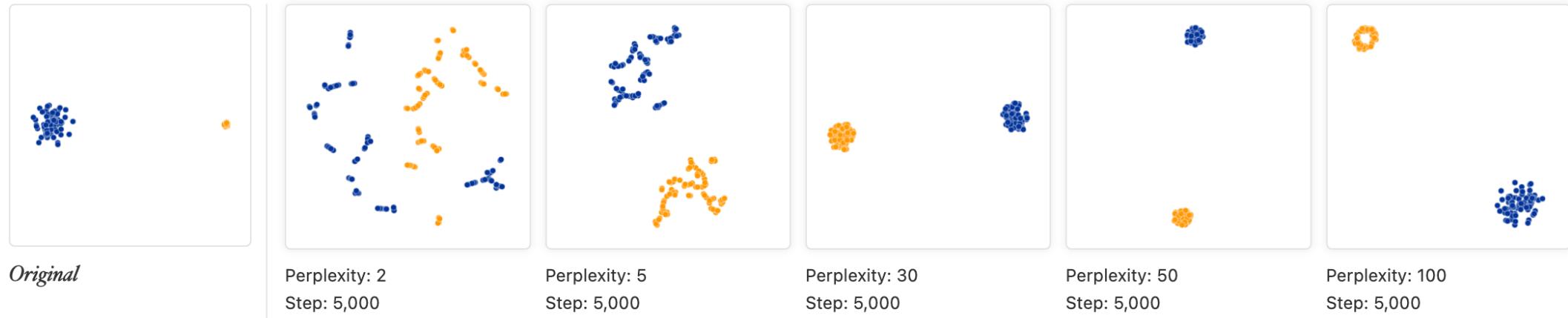
VISUALIZATIONS CAN BE MISLEADING

1. DISTANCES NOT PRESERVED IN TSNE

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

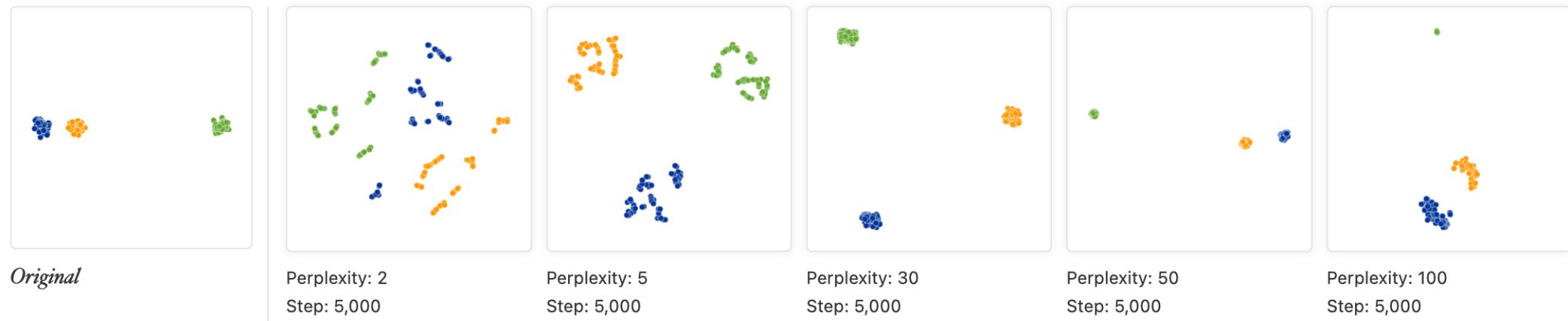


CLUSTER SIZES DO NOT MEAN ANYTHING

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



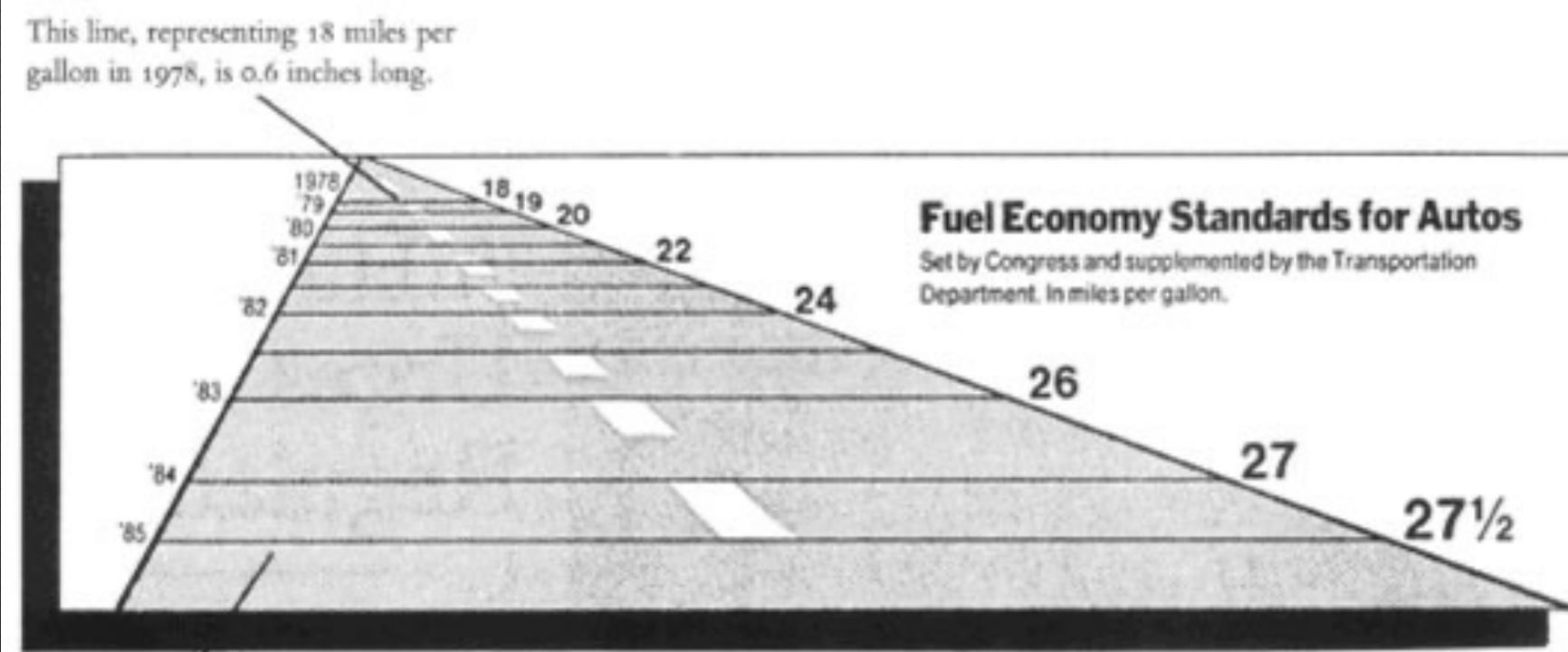
DISTANCES BETWEEN CLUSTERS MAY NOT MEAN ANYTHING

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



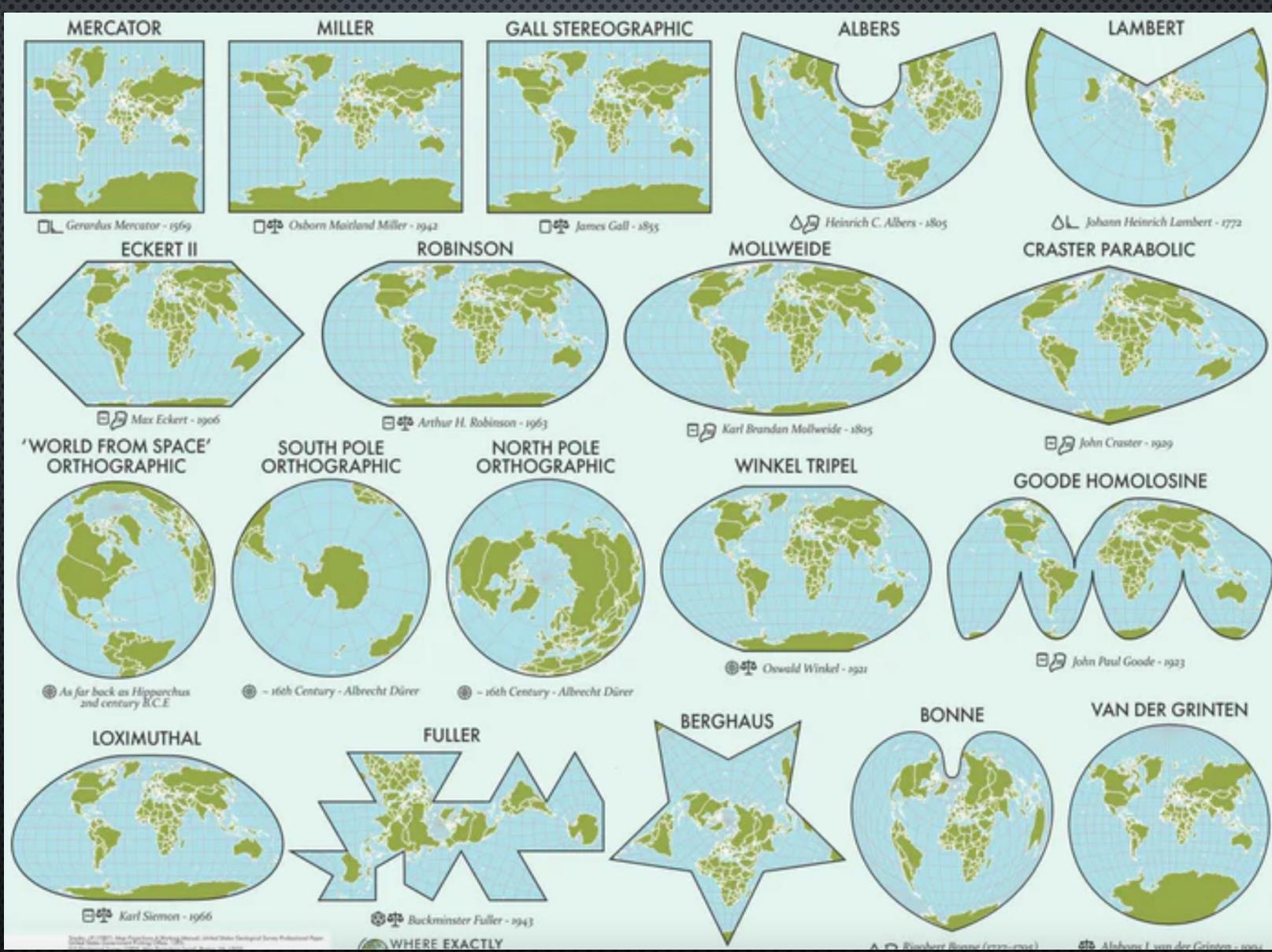
This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

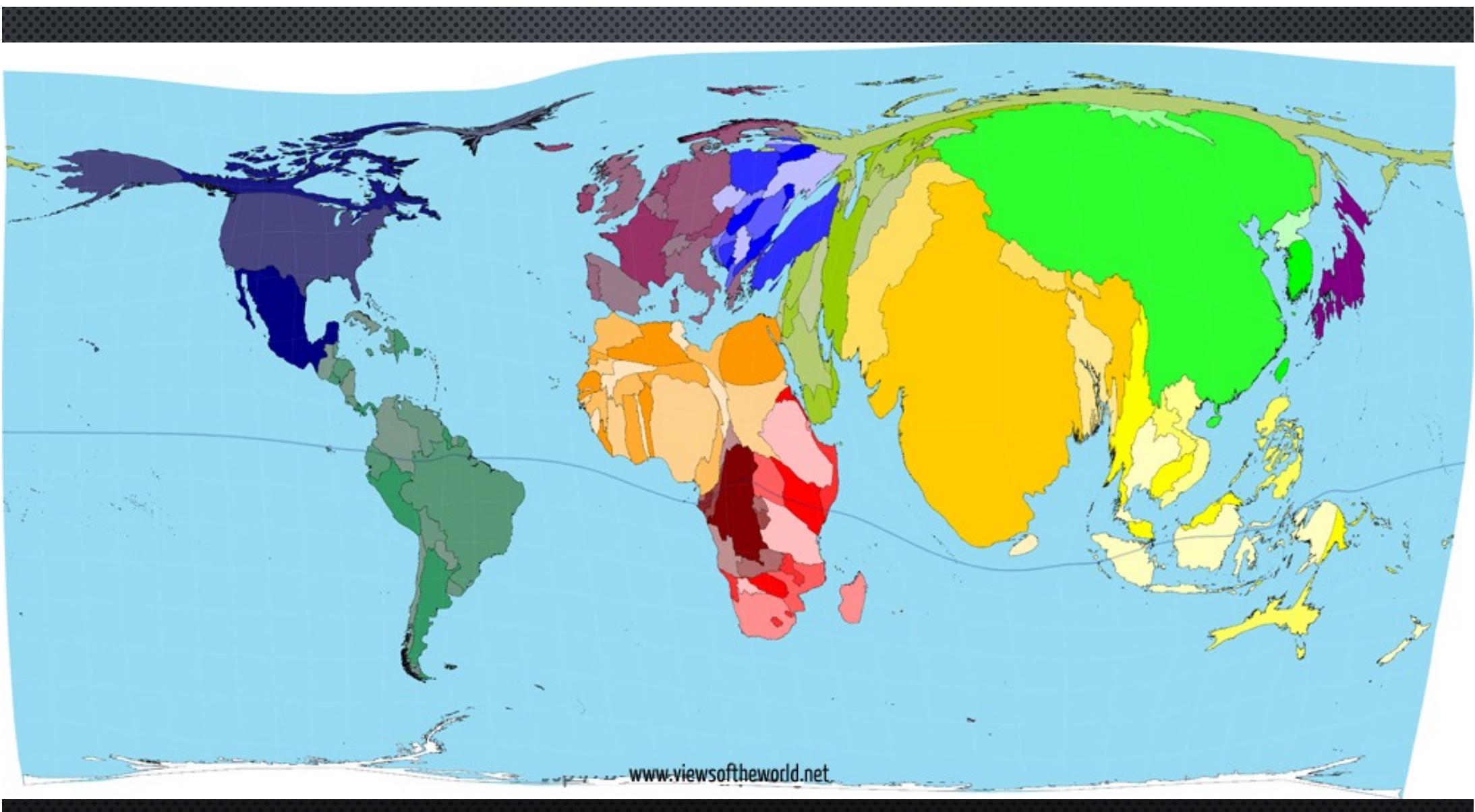
VISUALIZATIONS CAN BE MISLEADING

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

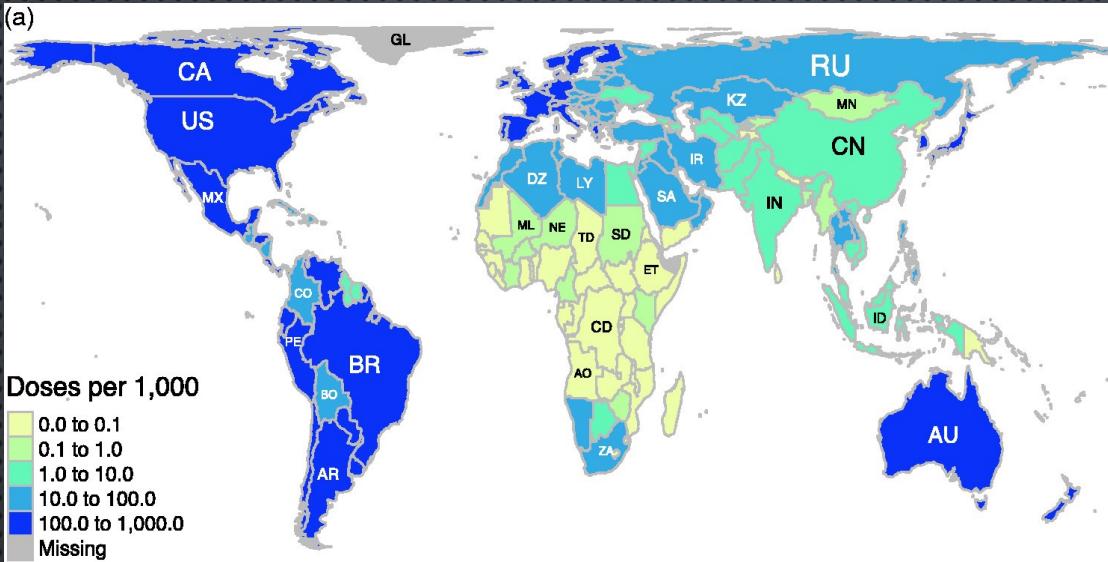
VISUALIZING HIGH-DIMENSIONS

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

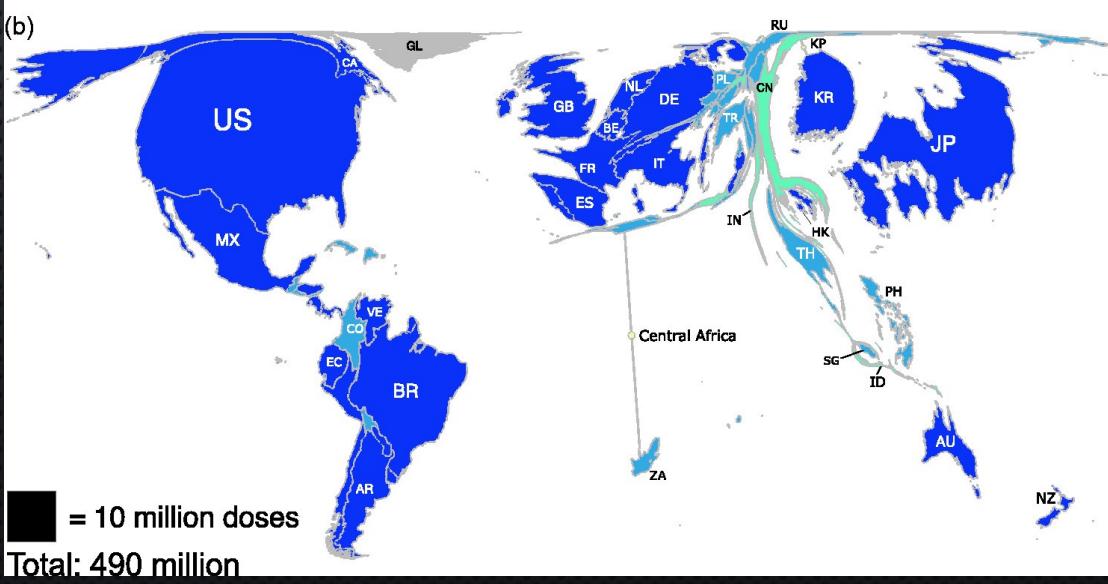




www.viewsoftheworld.net



Area proportional to number
of vaccine doses

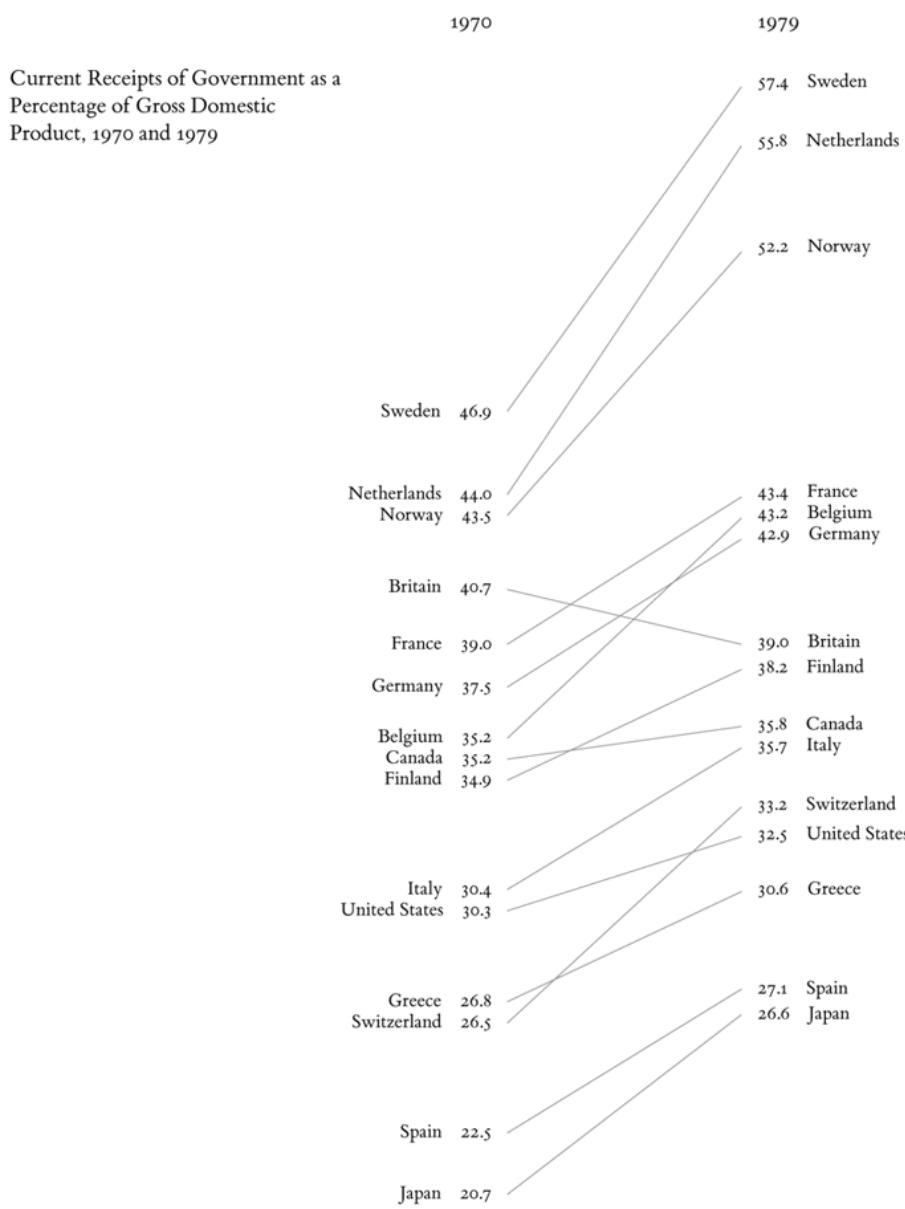


CARTOGRAMS

<https://go-cart.io/tutorial>

SLOPE GRAPHS

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



SLOPE GRAPHS

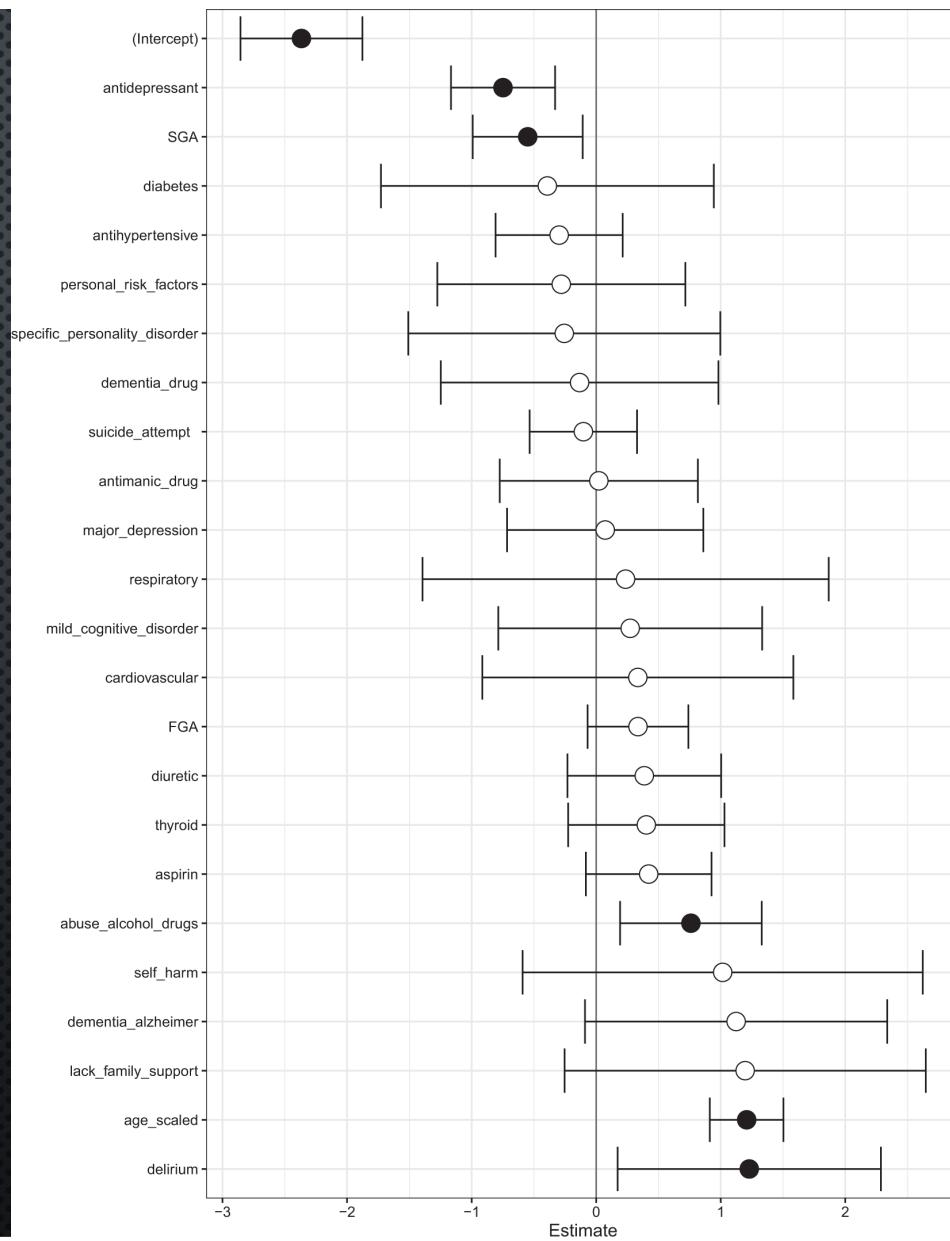
1. [https://www.edwardtufte.com/bboard/q-and-a-fetch-msg\ id=0003nk](https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg\ id=0003nk)
2. <https://charliepark.org/slopegraphs/>
3. Slope graphs in R
 1. <https://www.r-bloggers.com/2018/06/creating-slopegraphs-with-r/>
 1. <https://github.com/ibecav/CGPfunctions>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

VISUALIZATION FOR DIAGNOSTICS

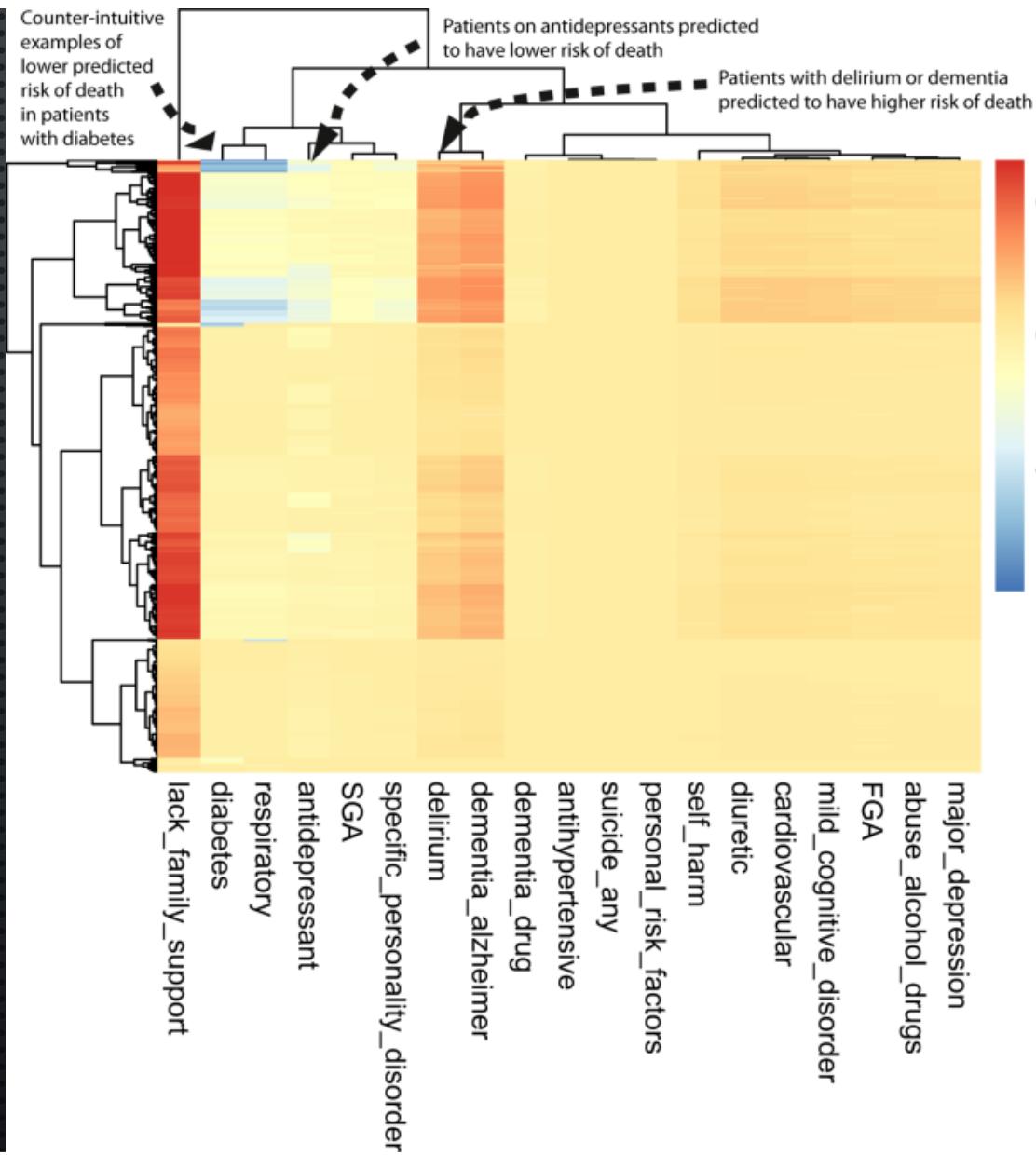
$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X.$$

1. GENERALIZED LINEAR MODEL

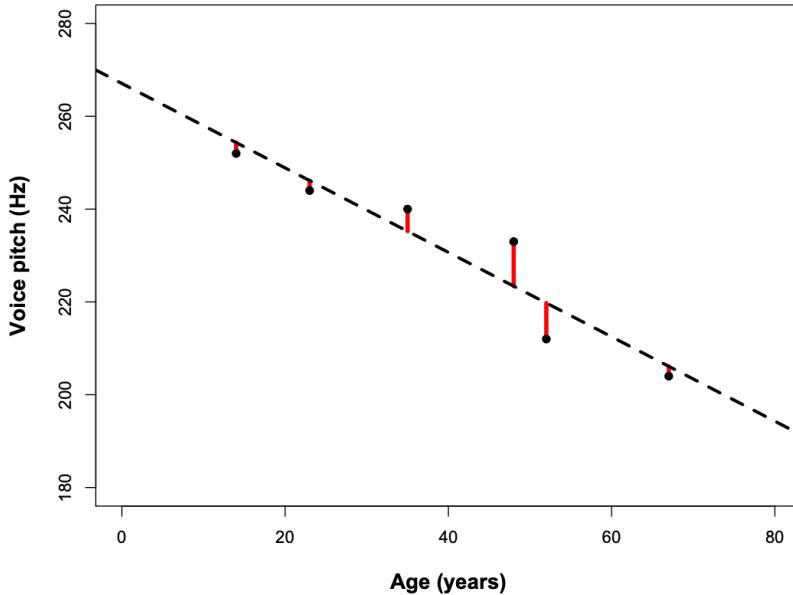
VISUALIZATION FOR DIAGNOSTICS



VISUALIZATION FOR DIAGNOSTICS

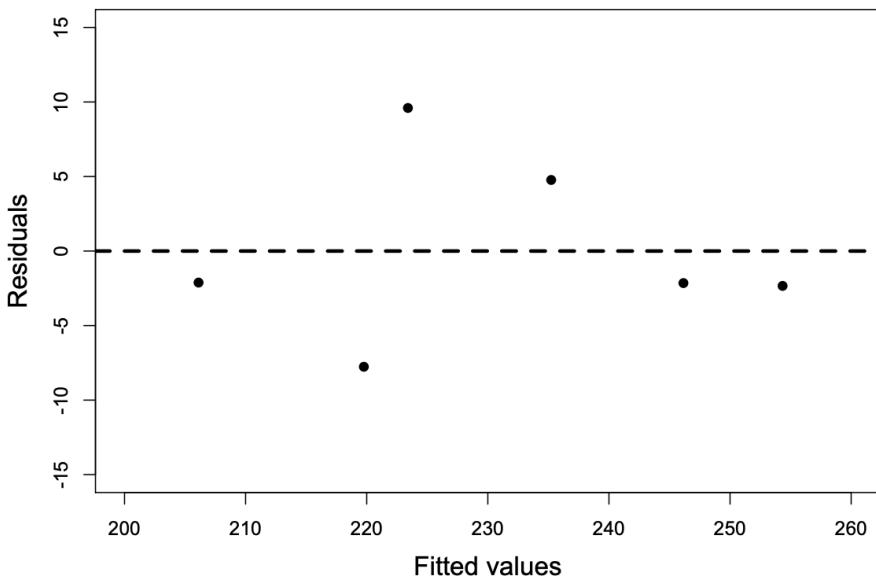


ASSUMPTIONS: LINEARITY



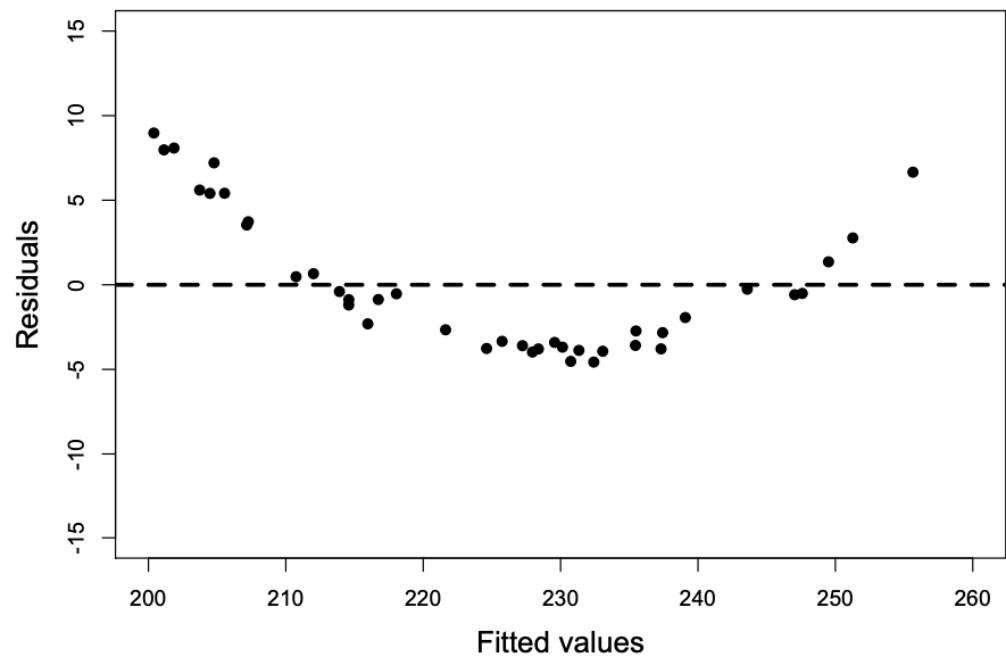
The red lines indicate the residuals, which are the deviations of the observed data points from the predicted values (the so-called “fitted values”). In this case, the residuals are all fairly small ... which is because the line that represents the linear model predicts our data very well, i.e., all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

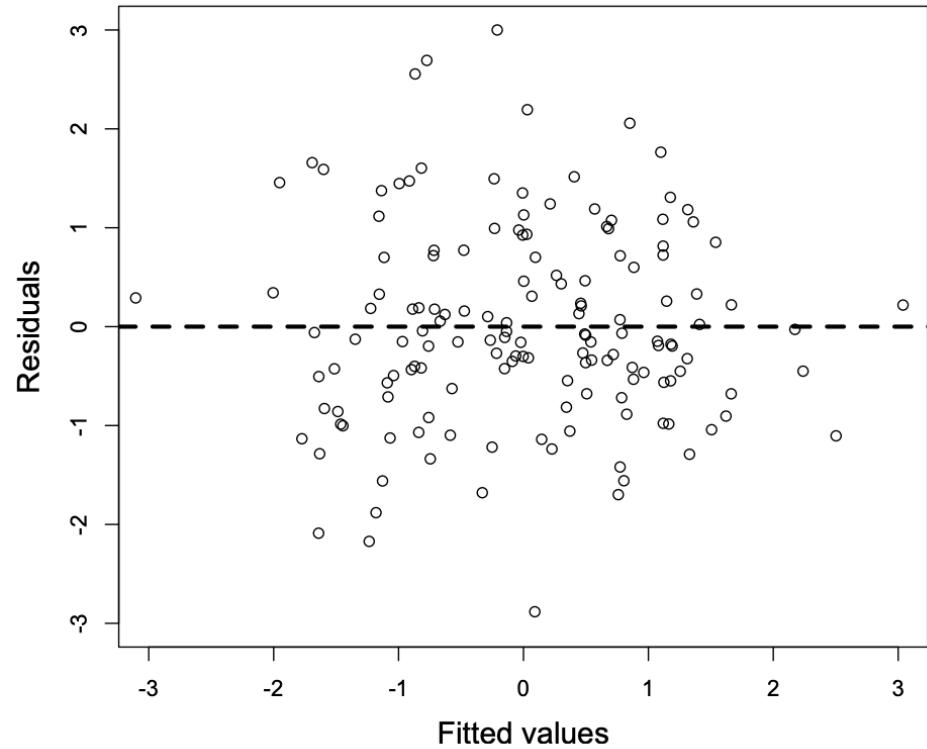
ASSUMPTIONS: LINEARITY



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:

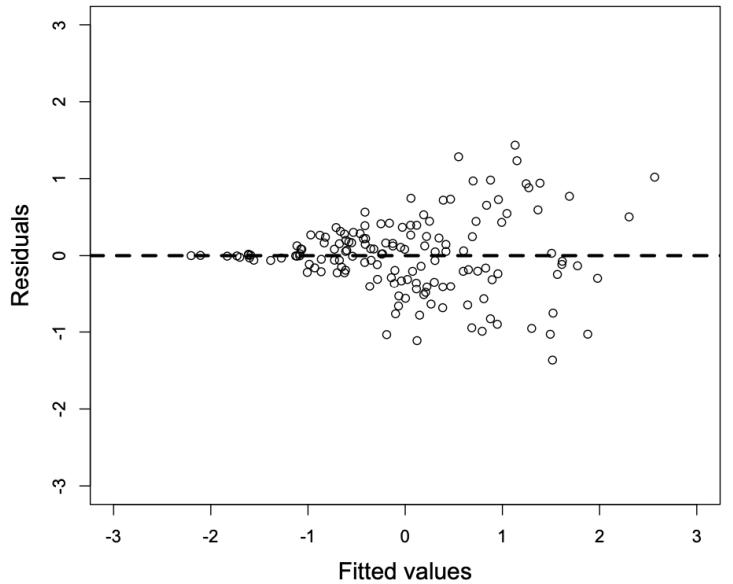
```
plot(fitted(xmdl),residuals(xmdl))2
```





ASSUMPTIONS

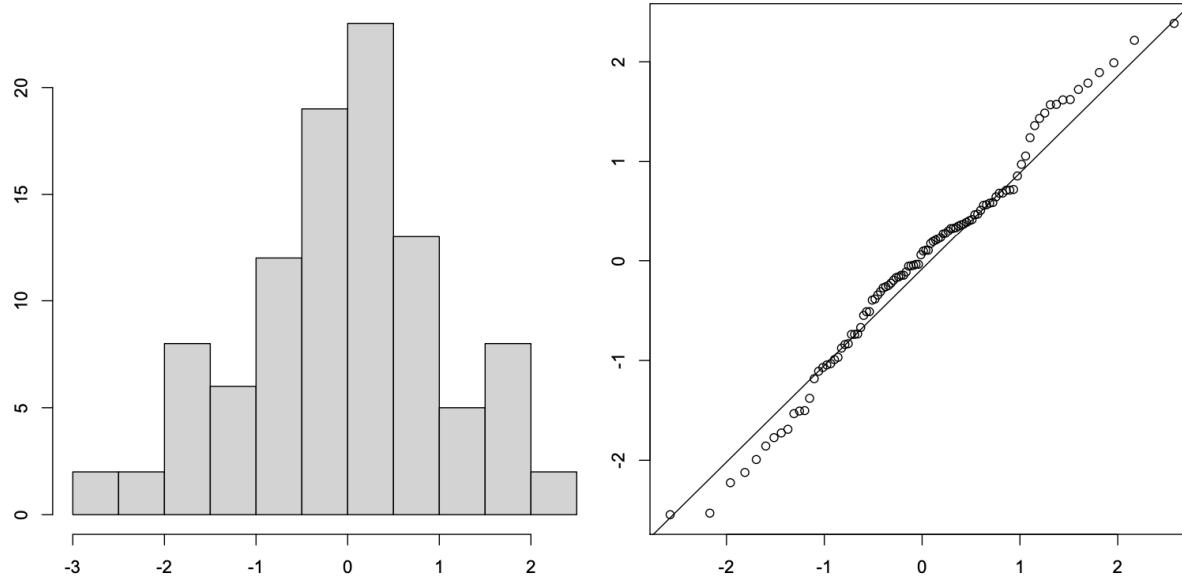
1. HOMOSKEDASTICITY
2. THE VARIABILITY OF YOUR DATA SHOULD BE APPROXIMATELY EQUAL ACROSS THE RANGE OF YOUR PREDICTED VALUES



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? TRANSFORM YOUR DATA



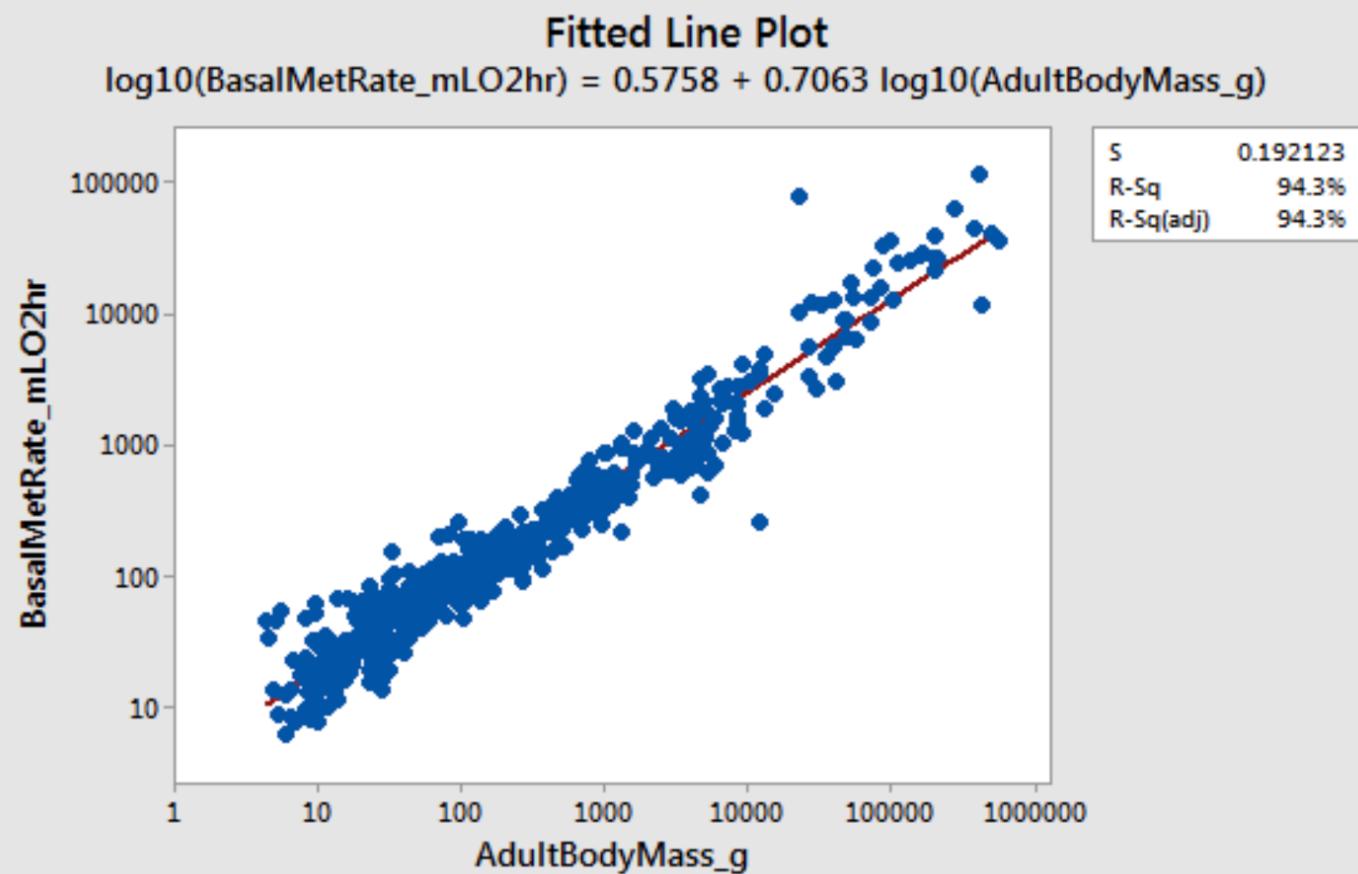
These look good. The histogram is relatively bell-shaped and the Q-Q plot indicates that the data falls on a straight line (which means that it's similar to a normal distribution). Here, we would conclude that there are no obvious violations of the normality assumption.

ASSUMPTIONS

1. NORMALITY OF RESIDUALS
2. Q-Q PLOT
3. [HTTPS://STACKOVERFLOW.COM/QUESTIONS/13865596/QUANTILE- PLOT-USING-SCIPY](https://stackoverflow.com/questions/13865596/quantile-plot-using-scipy)

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/>

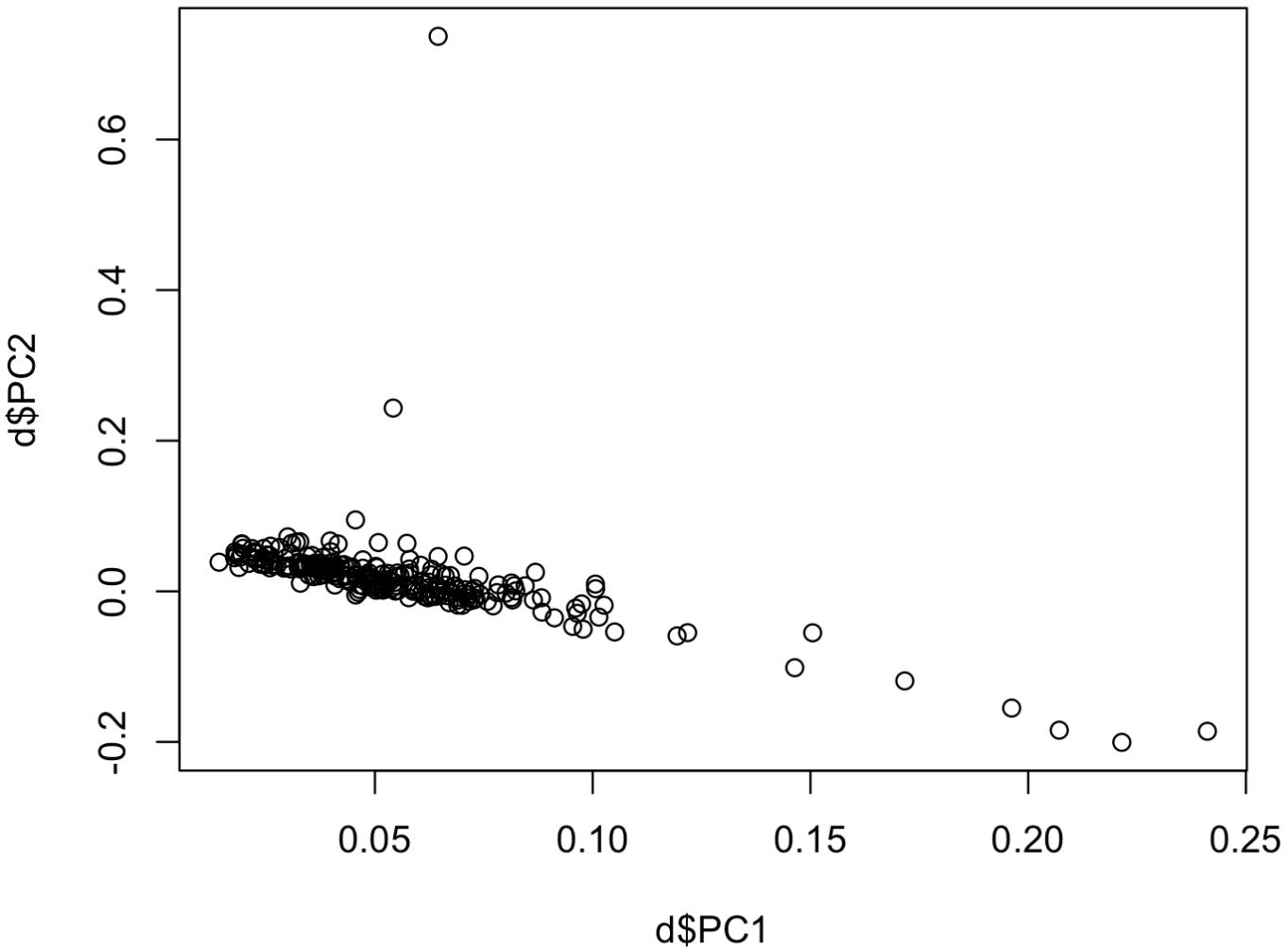
```
> dfbeta(xmdl)
   (Intercept)      age
1  -3.3645662  0.06437573
2  -1.6119656  0.02736278
3   1.5481303 -0.01456709
4  -0.0259835  0.05092767
5   0.8707699 -0.06479736
6   1.8551808 -0.06622744
```

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

ABSENCE OF INFLUENTIAL DATA POINTS



1. Removing outliers and then some downstream processing
2. t-test

Study 1		
Subject	Sex	Voice.Pitch
1	female	233 Hz
2	female	204 Hz
3	female	242 Hz
4	male	130 Hz
5	male	112 Hz
6	male	142 Hz

Study 2		
Subject	Age	Voice.Pitch
1	14	252 Hz
2	23	244 Hz
3	35	240 Hz
4	48	233 Hz
5	52	212 Hz
6	67	204 Hz

INDEPENDENCE

1. STUDY DESIGN
2. MIXED EFFECTS MODELS

RESOURCE

CODE FOR VISUALIZATION IN R, MODEL DIAGNOSTICS AND LINEAR MIXED EFFECTS MODELS

[HTTPS://GITHUB.COM/NEELSOUMYA/ANOVA_LINEAR_MIXED_EFFECTS_EXAMPLES](https://github.com/neelsoumya/ANOVA_LINEAR_MIXED_EFFECTS_EXAMPLES)

TIME SERIES DATA AND AUTOCORRELATIONS

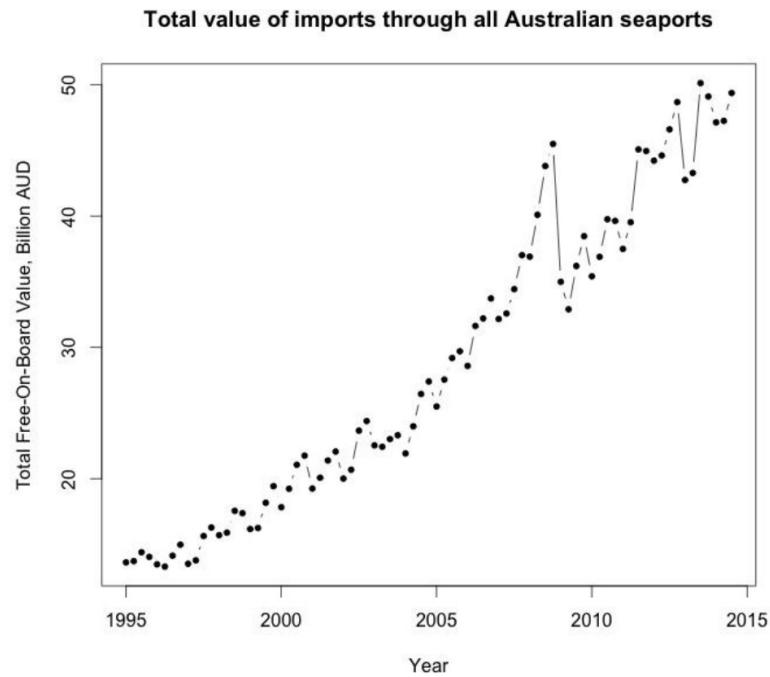


Figure 1. Total Free-On-Board value of Australian imports by sea through all ports.

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.

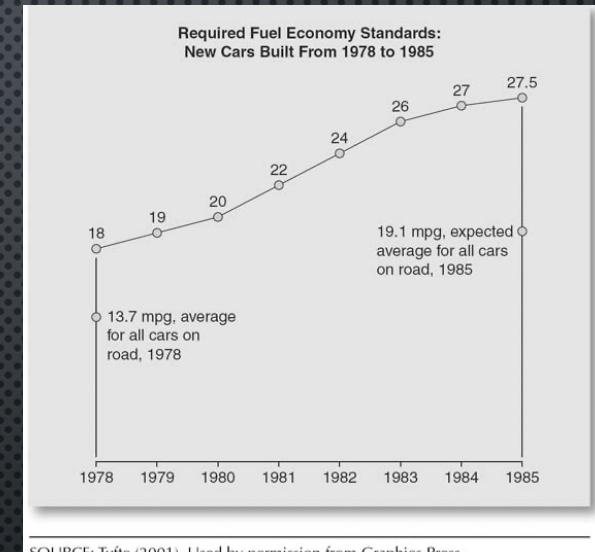
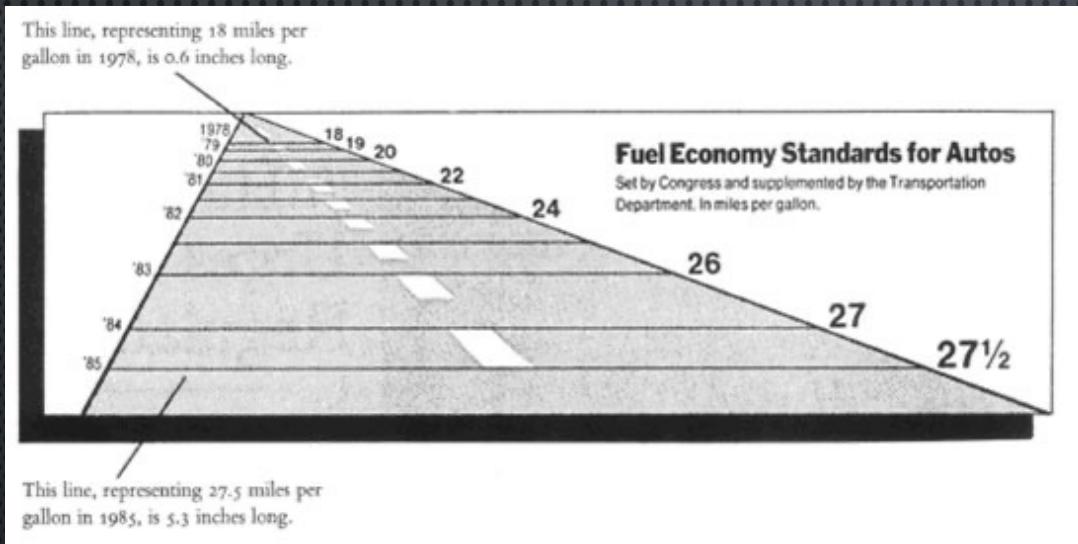
Visualization may also reveal if the underlying model/assumptions may have changed after a certain time. For example, in financial time-series data, there usually is a change after 2008 due to the global financial crisis.

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

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

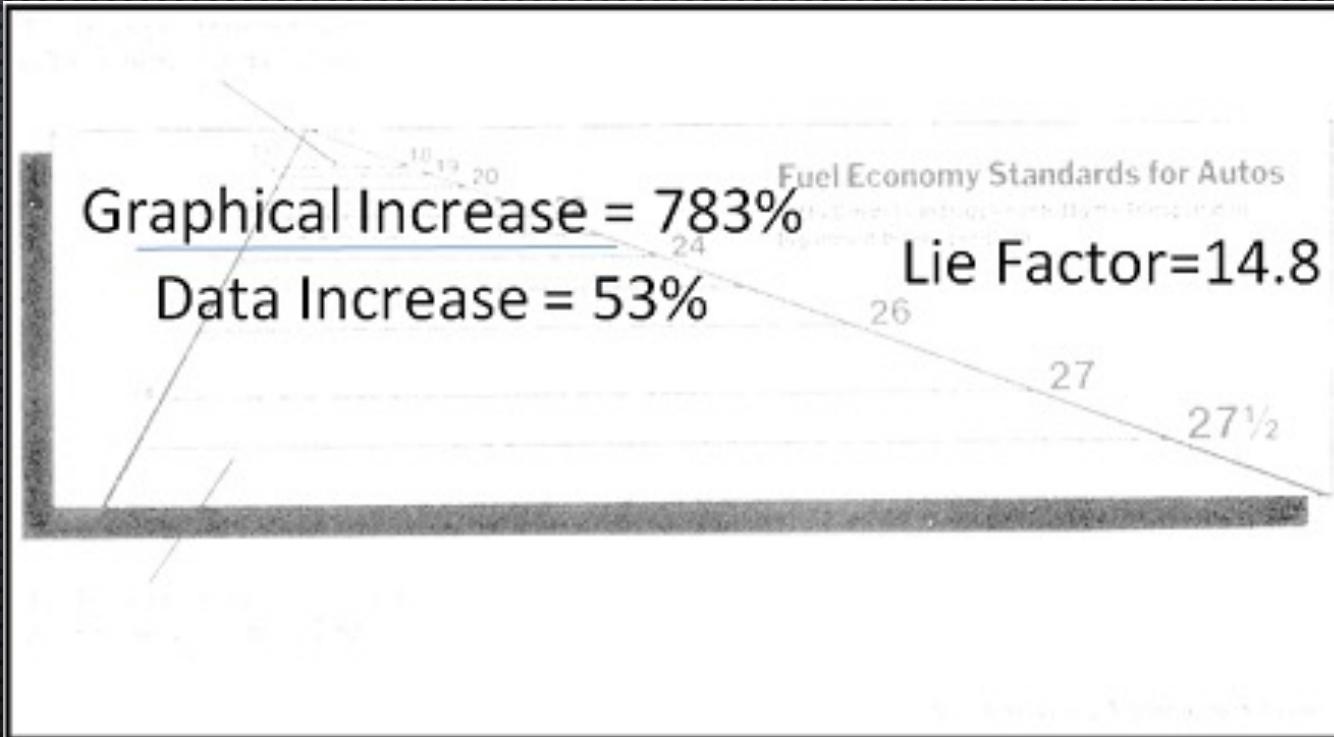
1. GENERALIZED LINEAR MODEL

TIME SERIES DATA VISUALIZATION



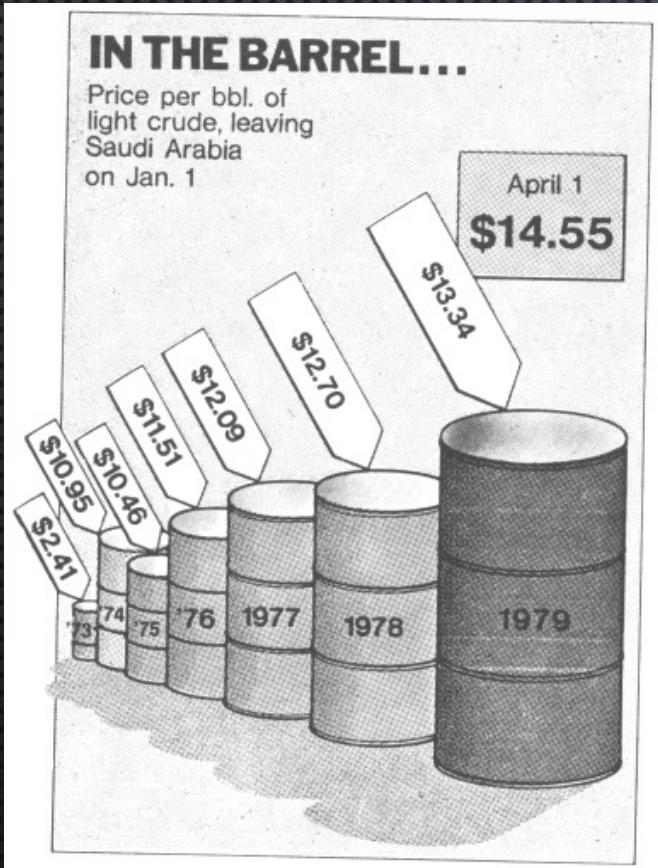
1. Area and volumes
2. Design effect vs. data effect

TIME SERIES DATA VISUALIZATION



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

APPS FOR RAPID PROTOTYPING AND COMMUNICATION

- SHINY APPS FOR VISUALIZATION AND COMMUNICATION
- [HTTPS://NEELSOUMYA.SHINYAPPS.IO/ACCIDENT_PREDICTION/](https://neelsoumya.shinyapps.io/accident_prediction/)
- REPRODUCIBLE ANALYSIS
- [HTTPS://GITHUB.COM/NEELSOUMYA/TEACHING_REPRODUCIBLE_SCIENCE_R](https://github.com/neelsoumya/teaching_reproducible_science_R)

APPS FOR RAPID PROTOTYPING AND COMMUNICATION

- OBSERVABLE (D3.JS)
 - [HTTPS://OBSERVABLEHQ.COM](https://observablehq.com)

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

MATERIAL

MATERIAL, SLIDES, CODE, EXERCISES, ACTIVITIES

[HTTPS://GITHUB.COM/NEELSOUMYA/VISUALIZATION_LECTURE](https://github.com/neelsoumya/visualization_lecture)

DERIVATIONS AND TECHNICAL DETAILS

[HTTPS://GITHUB.COM/NEELSOUMYA/VISUALIZATION_LECTURE/BLOB/MAIN/MATHEMATICS_DATA_SCIENCE.PDF](https://github.com/neelsoumya/visualization_lecture/blob/main/mathematics_data_science.pdf)

[HTTPS://OSF.IO/MNH8D/](https://osf.io/mnh8d/)

DATA STORYTELLING

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

ACTIVITIES

NORTH KOREA MISSILE RANGE ANIMATION

CRITIQUE

[HTTPS://NAGIX.GITHUB.IO/NK-MISSILE-TESTS/](https://nagix.github.io/nk-missile-tests/)

ACTIVITIES

CREATE CARTOGRAM ONLINE

[HTTPS://GO-CART.IO/TUTORIAL](https://go-cart.io/tutorial)

ACTIVITIES

CREATE INTERACTIVE DATA VISUALIZATION

[HTTPS://WWW.GAPMINDER.ORG/TOOLS/#\\$CHART-TYPE=BUBBLES&URL=v1](https://www.gapminder.org/tools/#$CHART-TYPE=BUBBLES&URL=v1)

[HTTPS://OURWORLDINDATA.ORG/GRA](https://ourworldindata.org/grapher/artificial-intelligence-training-computation?time=earliest..2023-10-11)
PER/ARTIFICIAL-INTELLIGENCE-TRAINING-
COMPUTATION?TIME=EARLIEST..2023-10-11

TIME SERIES DATA AND AUTOCORRELATIONS

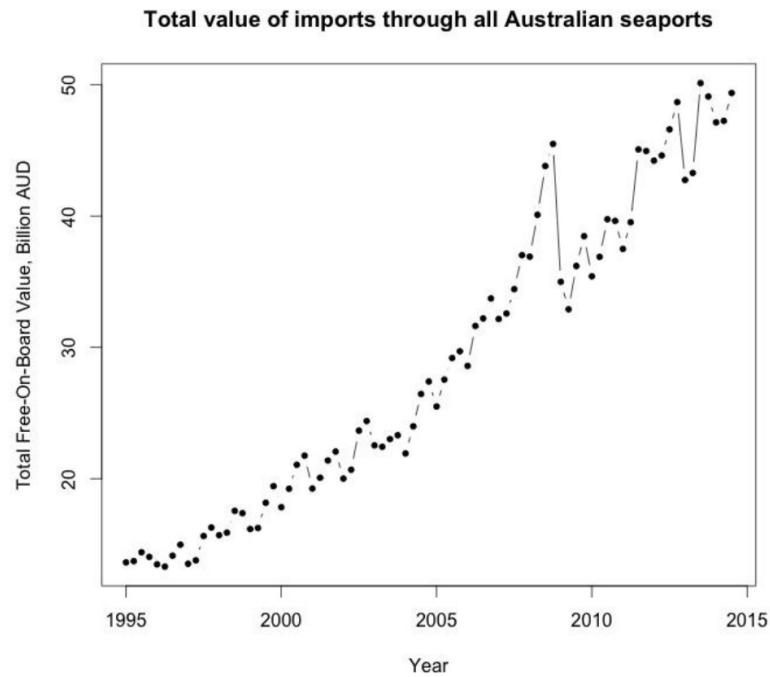


Figure 1. Total Free-On-Board value of Australian imports by sea through all ports.

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.

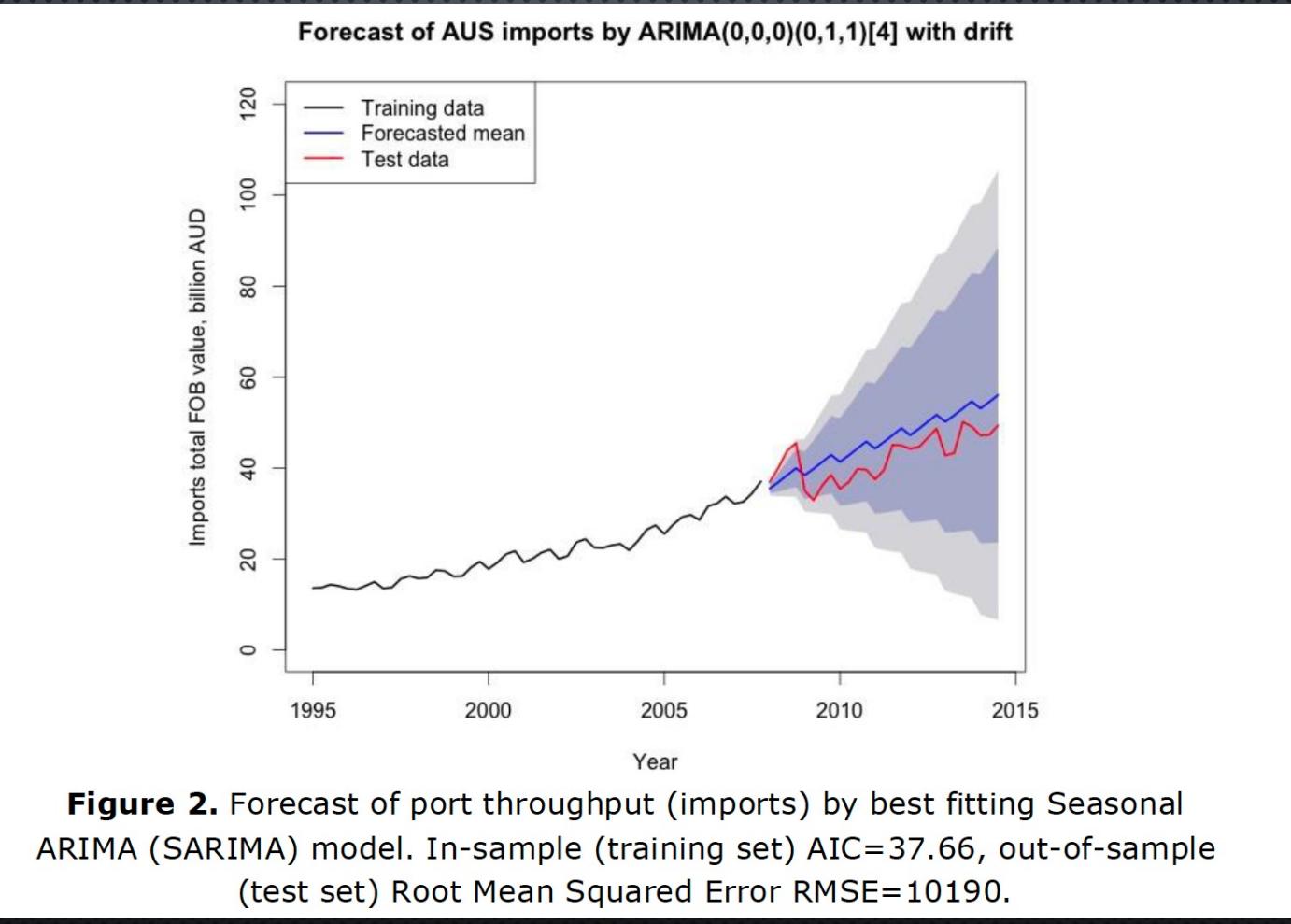
Visualization may also reveal if the underlying model/assumptions may have changed after a certain time. For example, in financial time-series data, there usually is a change after 2008 due to the global financial crisis.

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



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:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^p A_i \cdot \mathbf{y}_{t-i} + \epsilon_t$$

The variables (port throughput and GDP of all countries) are subsumed in the vector y_t , and y_{t-i} is the i th lag of y_t . The coefficient matrices A_i 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 Σ using 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

HOW 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 DIMENSIONAL DATA

RESULTS VISUALIZE

COMMUNICATE THIS

CANCER WORKING LINK WINTER ENVIRONMENTAL ACTIVISTS RACHEL CARSON SILENT SPRING

PITFALLS

PROB VISUALIZE

UN LOGO



TIME SERIES DATA



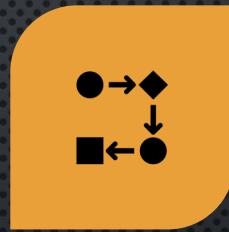
ASSUMPTIONS



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



LOOK AT CORRELATIONS



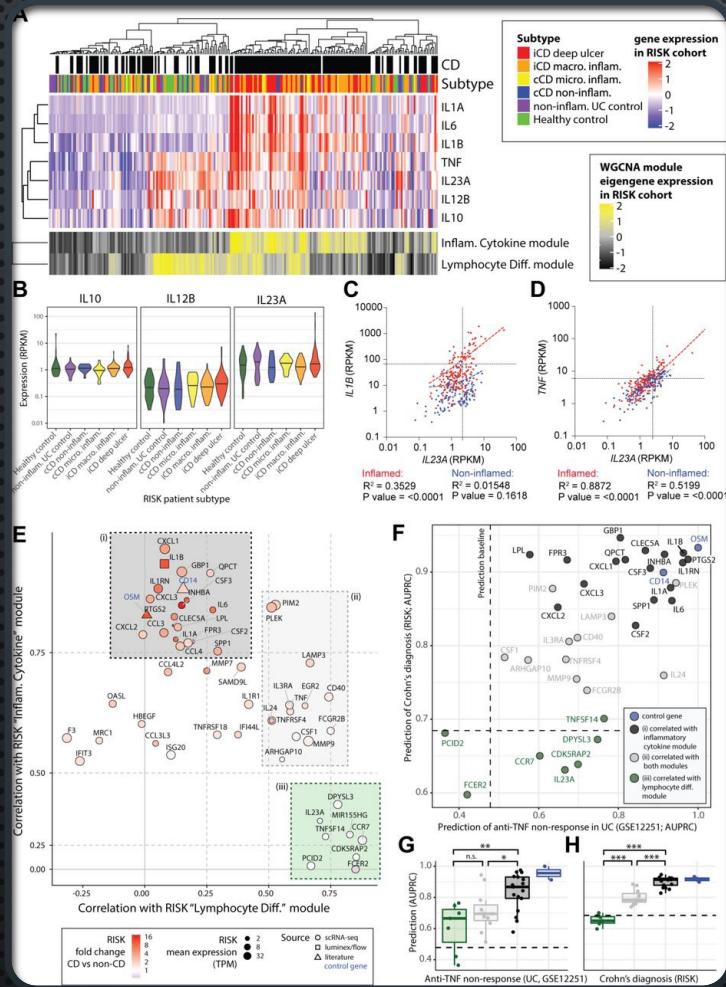
SIMPLER MODELS ARE
BETTER (MOVING
AVERAGES BETTER THAN
ANYTHING FANCY/ML)

DATA BIAS

Gorilla experiment (selective attention bias)

Hermann Hesse quote seeking is different to finding

CASE STUDY



PRINCIPLES OF DATA VISUALIZATION (EDWARD TUFTE)

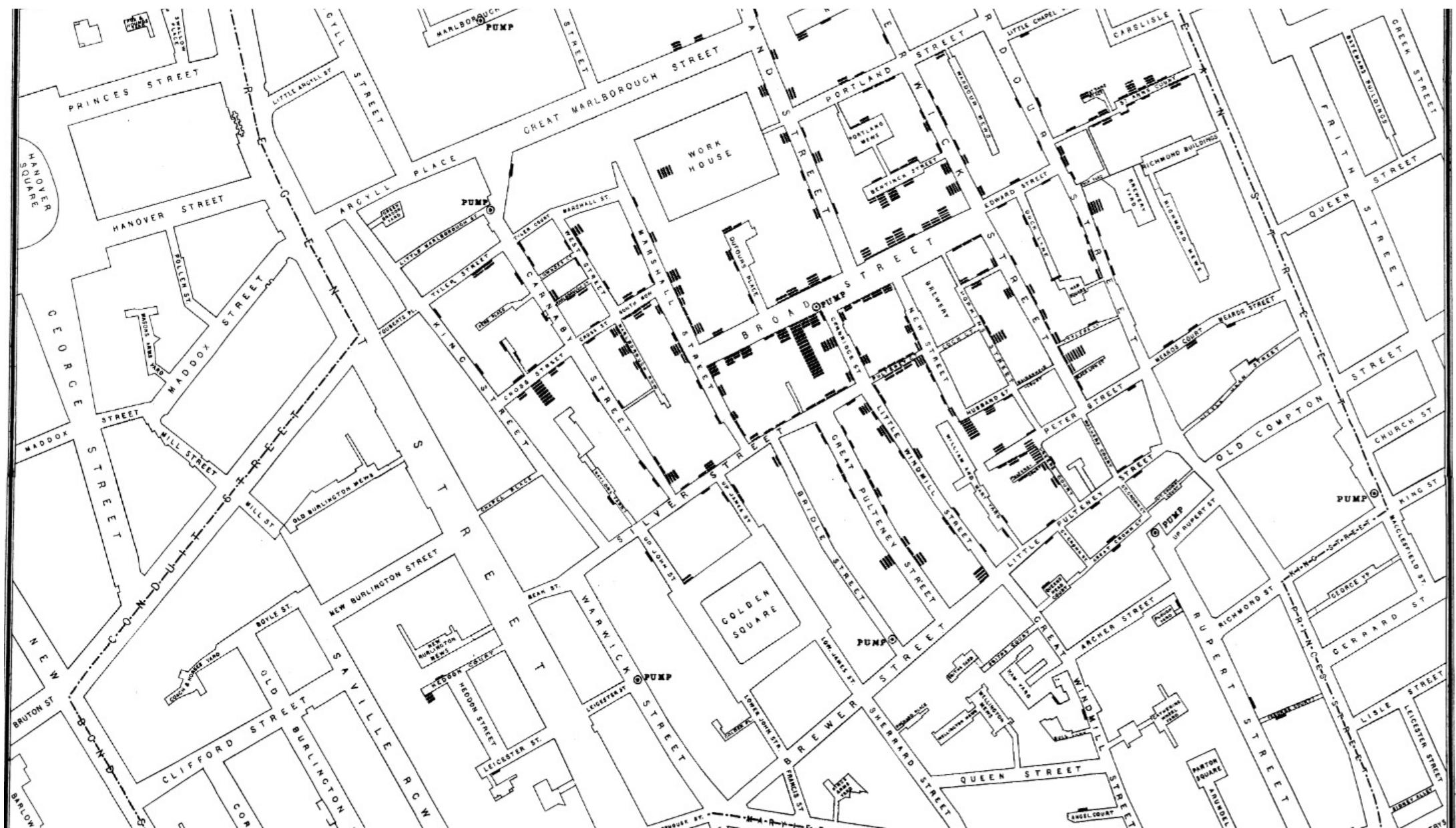
1. LESS IS MORE. "ABOVE ALL ELSE SHOW THE DATA"
2. "GRAPHICAL EXCELLENCE CONSISTS OF COMPLEX IDEAS COMMUNICATED WITH CLARITY, PRECISION, AND EFFICIENCY."
3. KEEP IT PROPORTIONAL! "LIE FACTOR = SIZE OF EFFECT SHOWN IN GRAPHIC DIVIDED BY SIZE OF EFFECT IN DATA"

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

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



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)