VISUALIZATION: A WHIRLWIND TOUR

SOUMYA BANERJEE

VISUALIZATION AS DEBUGGING

1. DATA SCIENCE AS (VISUAL) DEBUGGING

2. DATA STORYTELLING

3. HAVE YOU DONE PCA YET? (HAVE YOU TURNED IT OFF AND ON AGAIN?)

4. WORK CLOSELY WITH DOMAIN EXPERTS
    UGANDA VISUALIZATION TOOL (NO FANCY DATA SCIENCE/ML REQUIRED)

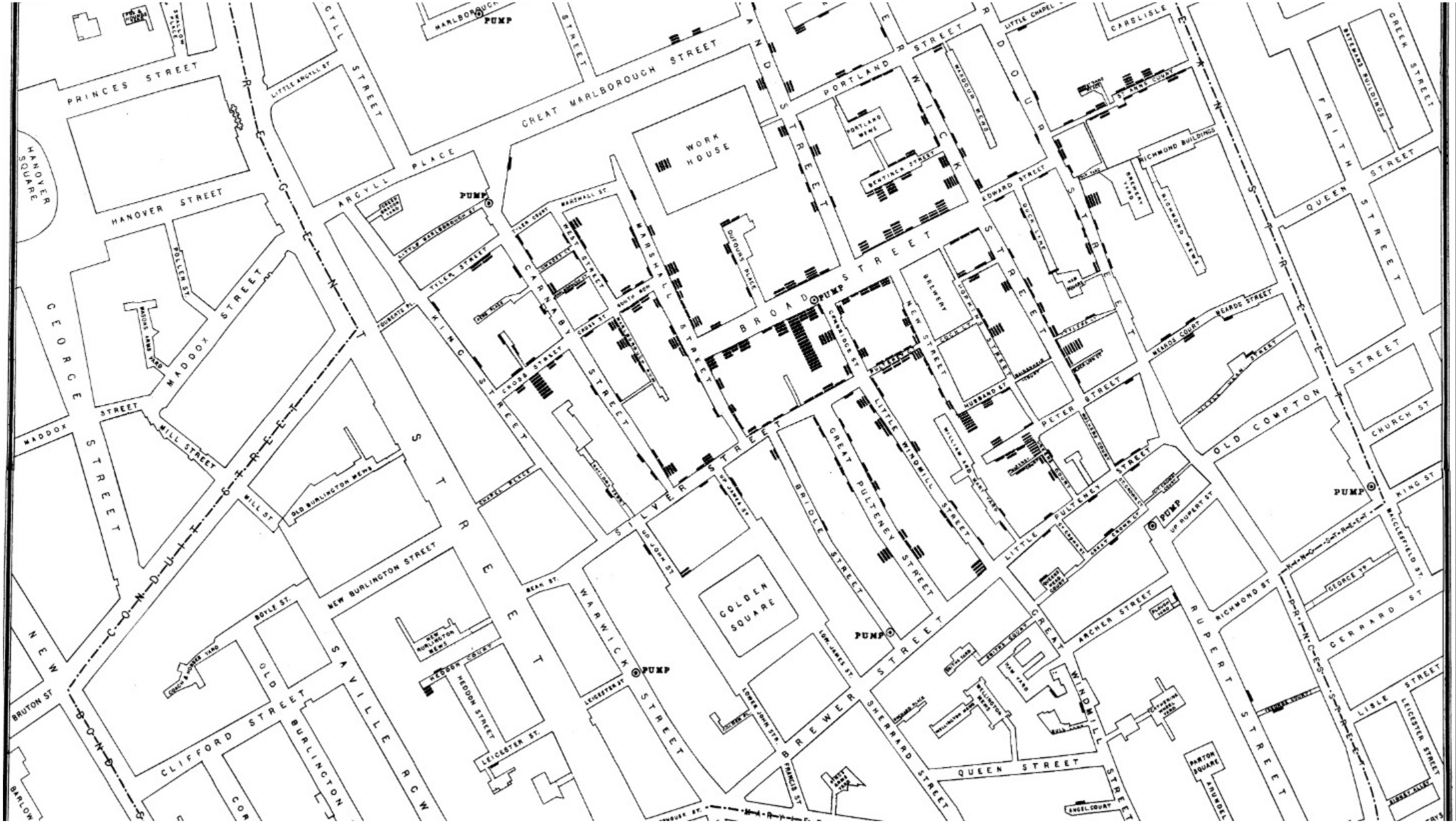
PRINCIPLES OF DATA VISUALIZATION (EDWARD TUFTE)

1. LESS IS MORE. "ABOVE ALL ELSE SHOW THE DATA"

2. "GRAPHICAL EXCELLENCE CONSISTS OF COMPLEX IDEAS COMMUNICATED WITH CLARITY, PRECISION, AND EFFICIENCY."

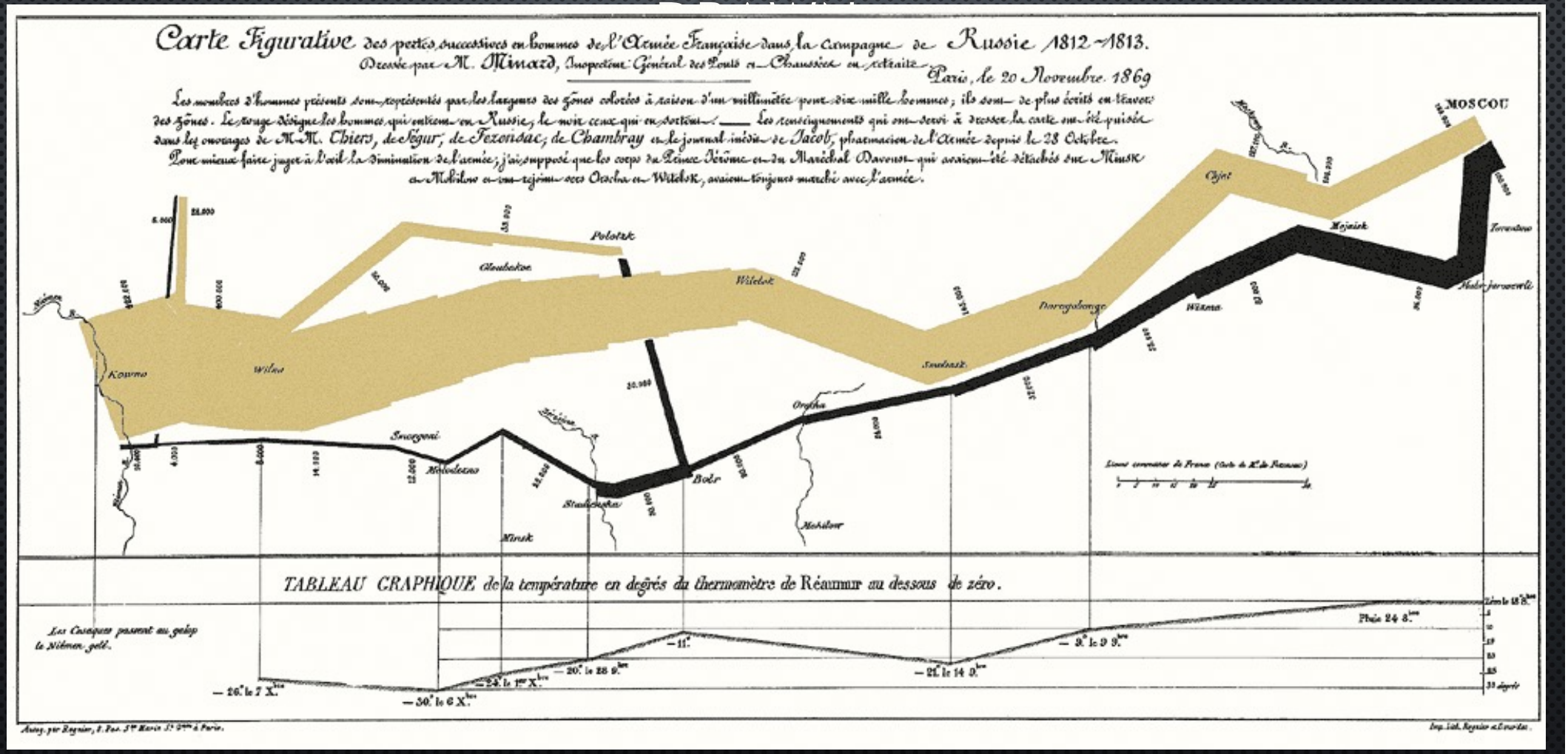
3. KEEP IT PROPORTIONAL! "LIE FACTOR = SIZE OF EFFECT SHOWN IN GRAPHIC DIVIDED BY SIZE OF EFFECT IN DATA"

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



PUMP
PRINCES STREET
HANOVER SQUARE
HANOVER STREET
GREAT MARLBOROUGH STREET
ARGYLL PLACE
LITTLE ARGYLL ST
LITTLE MARLBOROUGH ST
WORK HOUSE
PORTLAND STREET
PORTLAND MEWS
BENTINCK STREET
MARSHALL ST
BROAD STREET
DUFOURS PLACE
CARNABY STREET
KING STREET
TYLER STREET
CROSS STREET
SOUTH ROW
EDWARD STREET
DEAN STREET
WARDOUR STREET
RICHMOND BUILDINGS
RICHMOND MEWS
ST ANNS COURT
CARLISLE STREET
LITTLE CHAPEL ST
FRITH STREET
GREEK STREET
QUEEN STREET
BATEMANS BUILDINGS
BREWERY
HOPKINS STREET
NEW STREET
HUSBAND ST
PETER STREET
BEARDS STREET
MEARDS COURT
OLD COMPTON STREET
CHURCH STREET
KING ST
MACCLESFIELD ST
GEORGE YD
GERRARD ST
LITTLE WINDMILL STREET
GREAT PULTENEY STREET
LITTLE PULTENEY STREET
WILLIAM AND MARY YARD
SILVER STREET
BRIDLE STREET
GOLDEN SQUARE
GEORGE STREET
MADDOX STREET
MILL STREET
MILL ST
OLD BURLINGTON MEWS
NEW BURLINGTON STREET
BOYLE ST
CONDUIT STREET
NEW BOND STREET
BRUTON ST
CLIFFORD STREET
OLD BURLINGTON STREET
SAVILLE ROW
REGENT STREET
NEW BURLINGTON MEWS
HEDDON COURT
HEDDON STREET
LEICESTER ST
WARWICK STREET
BEAK ST
UP JAMES ST
UP JOHN ST
LOW JAMES ST
LOWER JOHN STR
BREWER STREET
SHERRARD STREET
FRANCIS ST
QUEEN STREET
GREAT WINDMILL STREET
ARCHER STREET
HAM YARD
WELLINGTON MEWS
RUPERT STREET
UP RUPERT ST
RICHMOND ST
PRINCES STREET
PARTON SQUARE
ARUNDEL
LISLE STREET
LEICESTER STREET
ANGEL COURT

THE BEST STATISTICAL GRAPHIC EVER



PRINCIPLES OF DATA VISUALIZATION

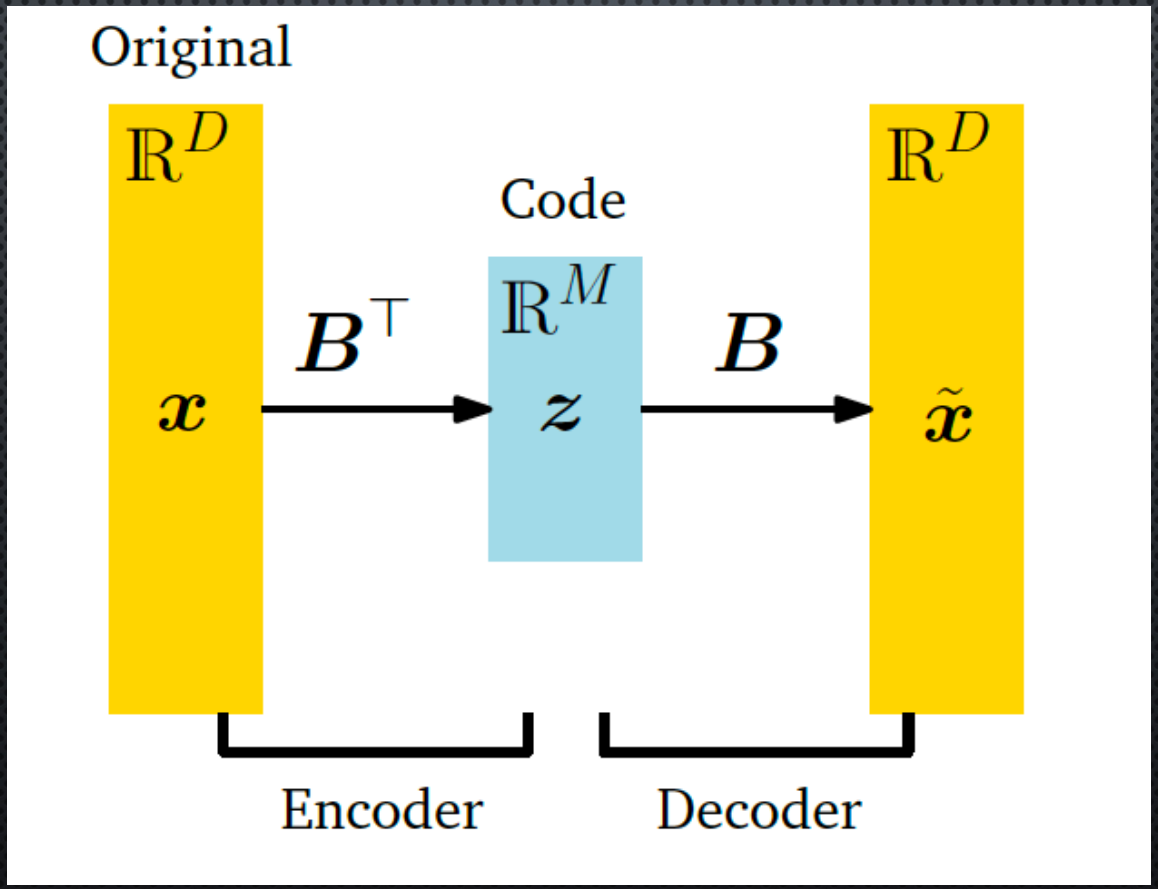
- HEATMAPS
- TSNE
- COMMUNICATE
- DATA STORYTELLING
- DISTANCES NOT PRESERVED
- STOCHASTIC
- DIFFICULT TO COMMUNICATE TO NON-TECHNICAL EXPERTS

VISUALIZING HIGH-DIMENSIONS

1. PROBLEMS WITH COMMUNICATING HIGH DIMENSIONAL DATA

2. HIGH DIMENSIONS ARE DIFFICULT TO VISUALIZE

IMPORTANT CONCEPT



INFORMATION BOTTLENECK

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)

GENERALIZATIONS OF THIS IDEA

- tSNE
- Autoencoder (non-linear loss function)

PITFALLS

Distances not preserved

tSNE can be used for hypothesis generation. There are many pitfalls to this.

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

Other 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.

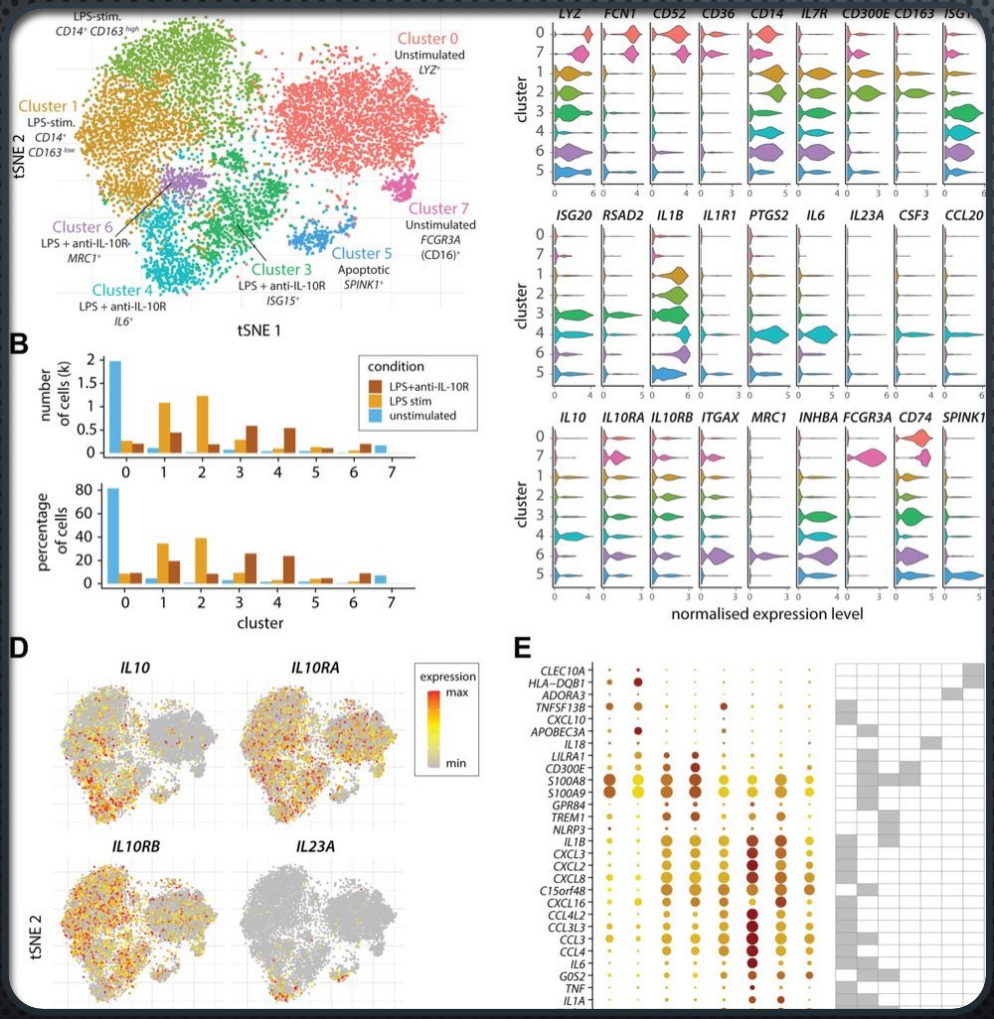
Case studies

Visualization/dimensionality reduction pitfalls

Pitfalls

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

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



CASE STUDY

Color scales

VISUALIZATIONS CAN BE MISLEADING

1. DISTANCES NOT PRESERVED IN TSNE

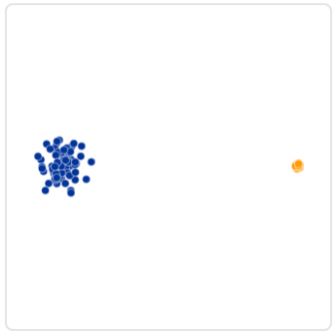
HTTPS://DISTILL.PUB/2016/MISREAD-TSNE/

2. CLUSTER SIZES DO NOT MATTER

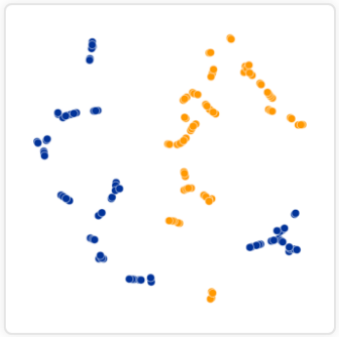
3. YOU CAN SEE SOME SHAPES SOMETIMES

4. RANDOM DOES NOT ALWAYS LOOK RANDOM

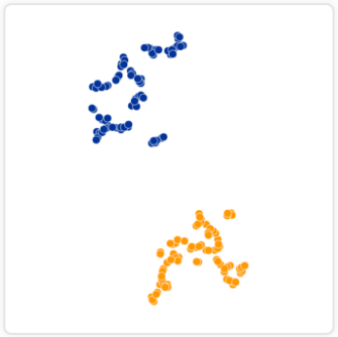
VISUALIZATIONS CAN BE MISLEADING



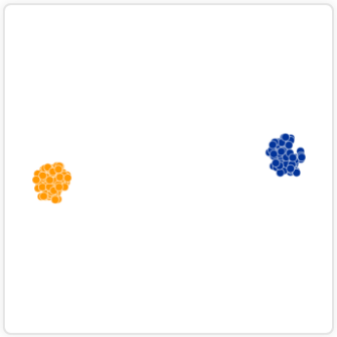
Original



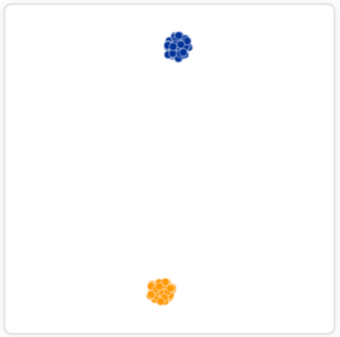
Perplexity: 2
Step: 5,000



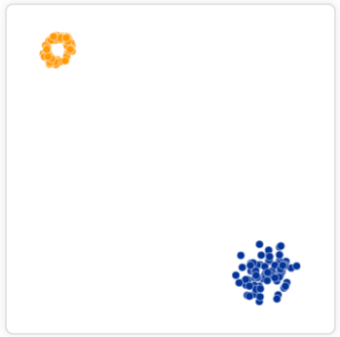
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000

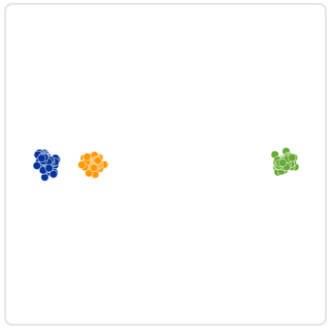


Perplexity: 100
Step: 5,000

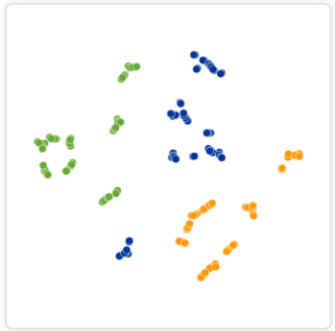
BUT WHAT IF THE TWO CLUSTERS HAVE DIFFERENT STANDARD DEVIATIONS, AND SO DIFFERENT SIZES? (BY SIZE WE MEAN BOUNDING BOX MEASUREMENTS, NOT NUMBER OF POINTS.) BELOW ARE T-SNE PLOTS FOR A MIXTURE OF GAUSSIANS IN PLANE, WHERE ONE IS 10 TIMES AS DISPERSED AS THE OTHER.

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.

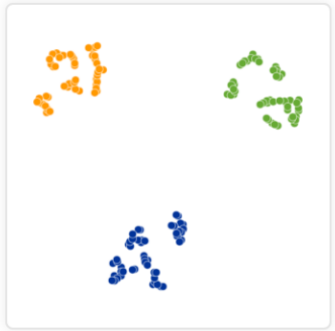
VISUALIZATIONS CAN BE MISLEADING



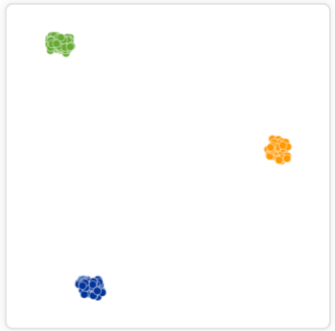
Original



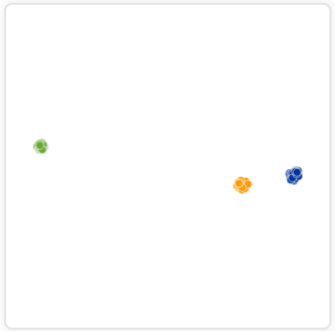
Perplexity: 2
Step: 5,000



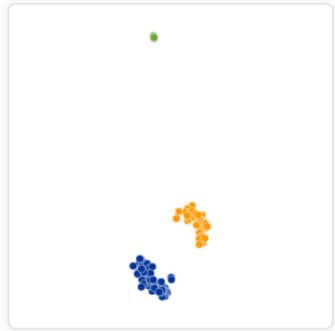
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000



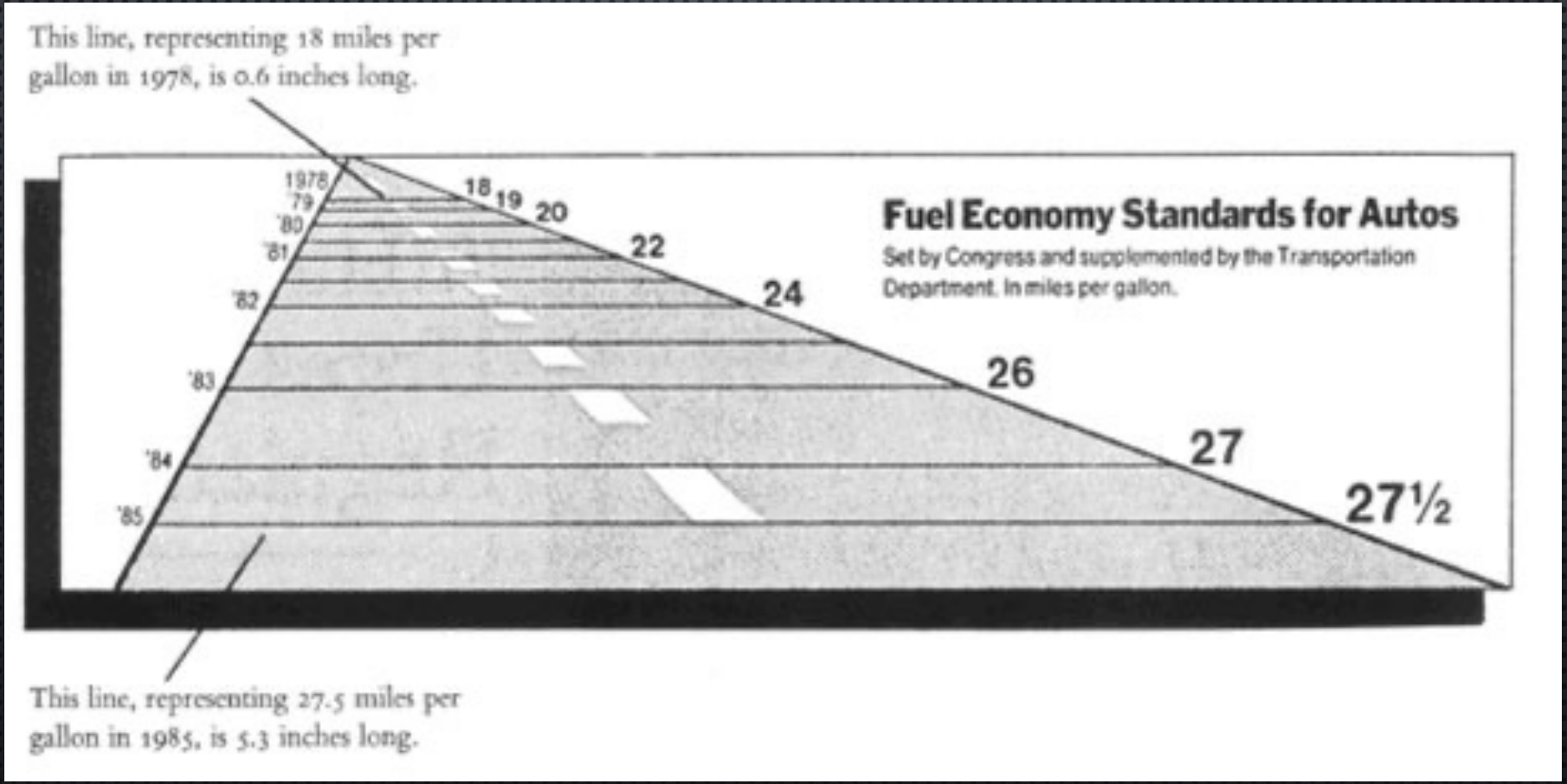
Perplexity: 100
Step: 5,000

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?

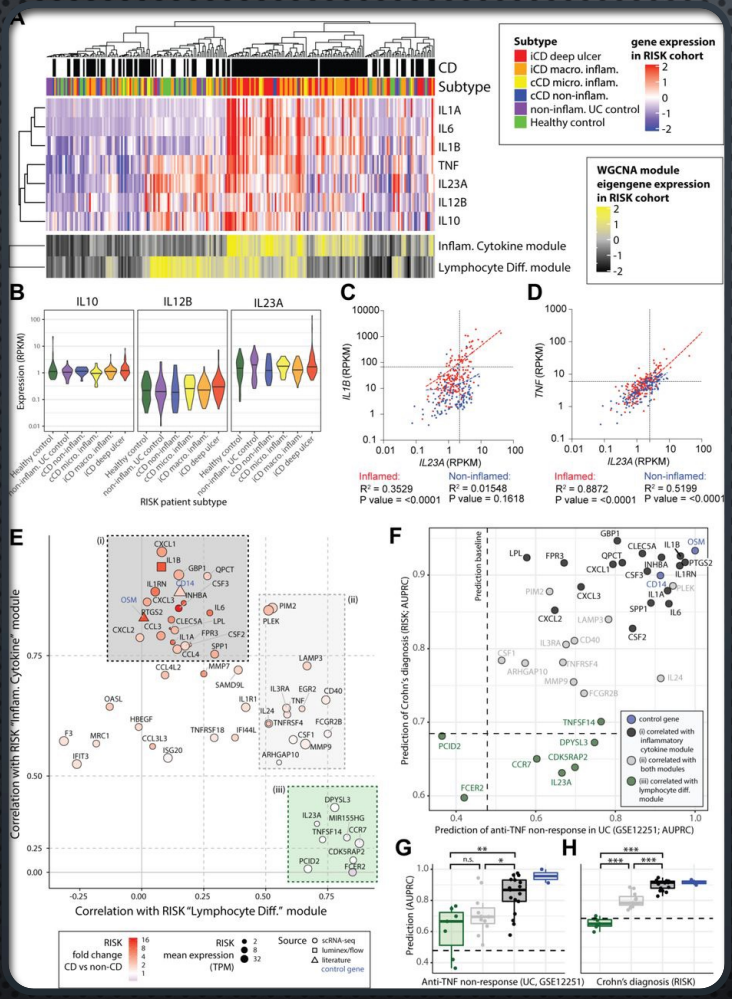
THE BASIC MESSAGE IS THAT DISTANCES BETWEEN WELL-SEPARATED CLUSTERS IN A T-SNE PLOT MAY MEAN NOTHING.

VISUALIZATIONS CAN BE MISLEADING

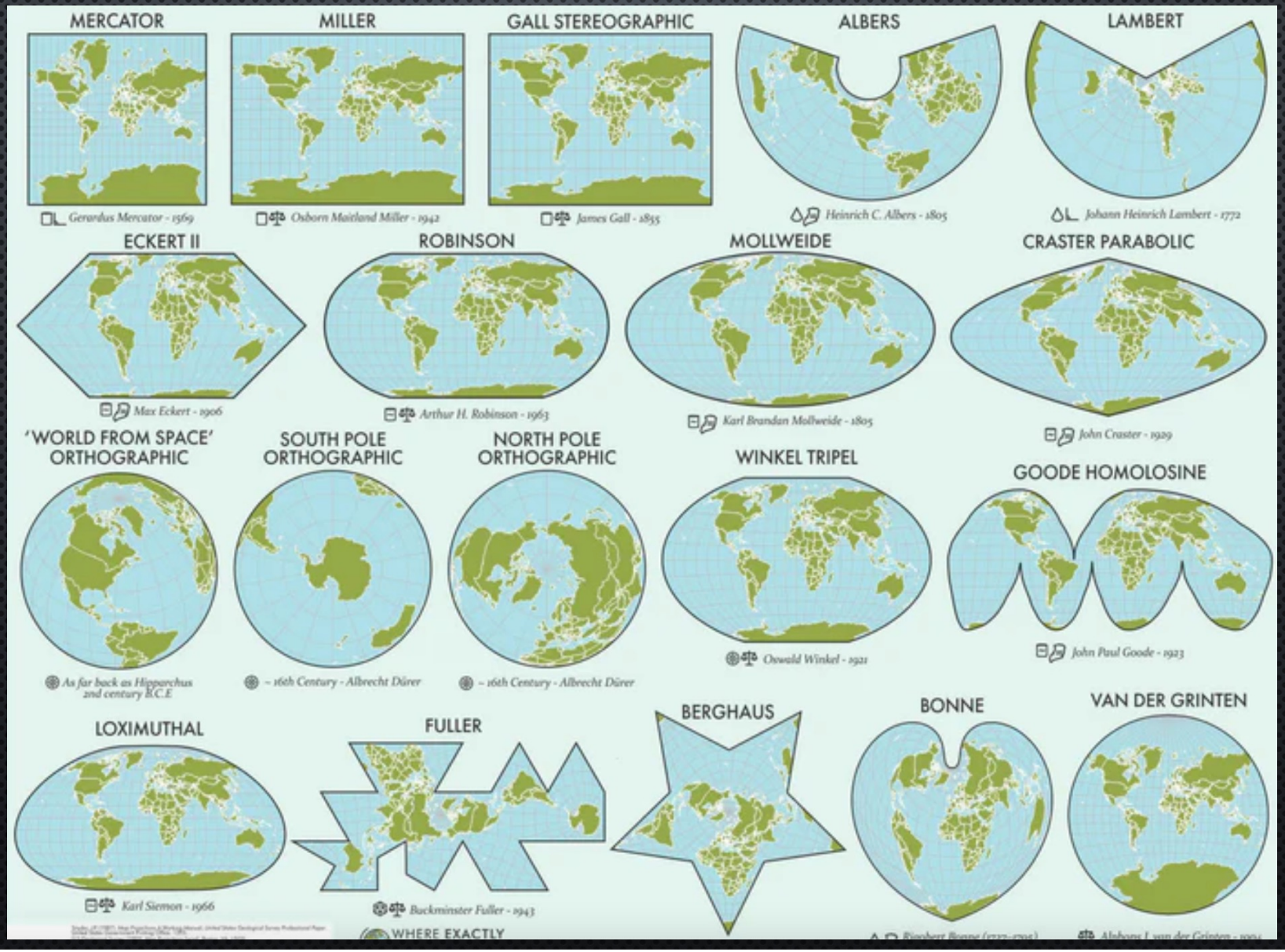


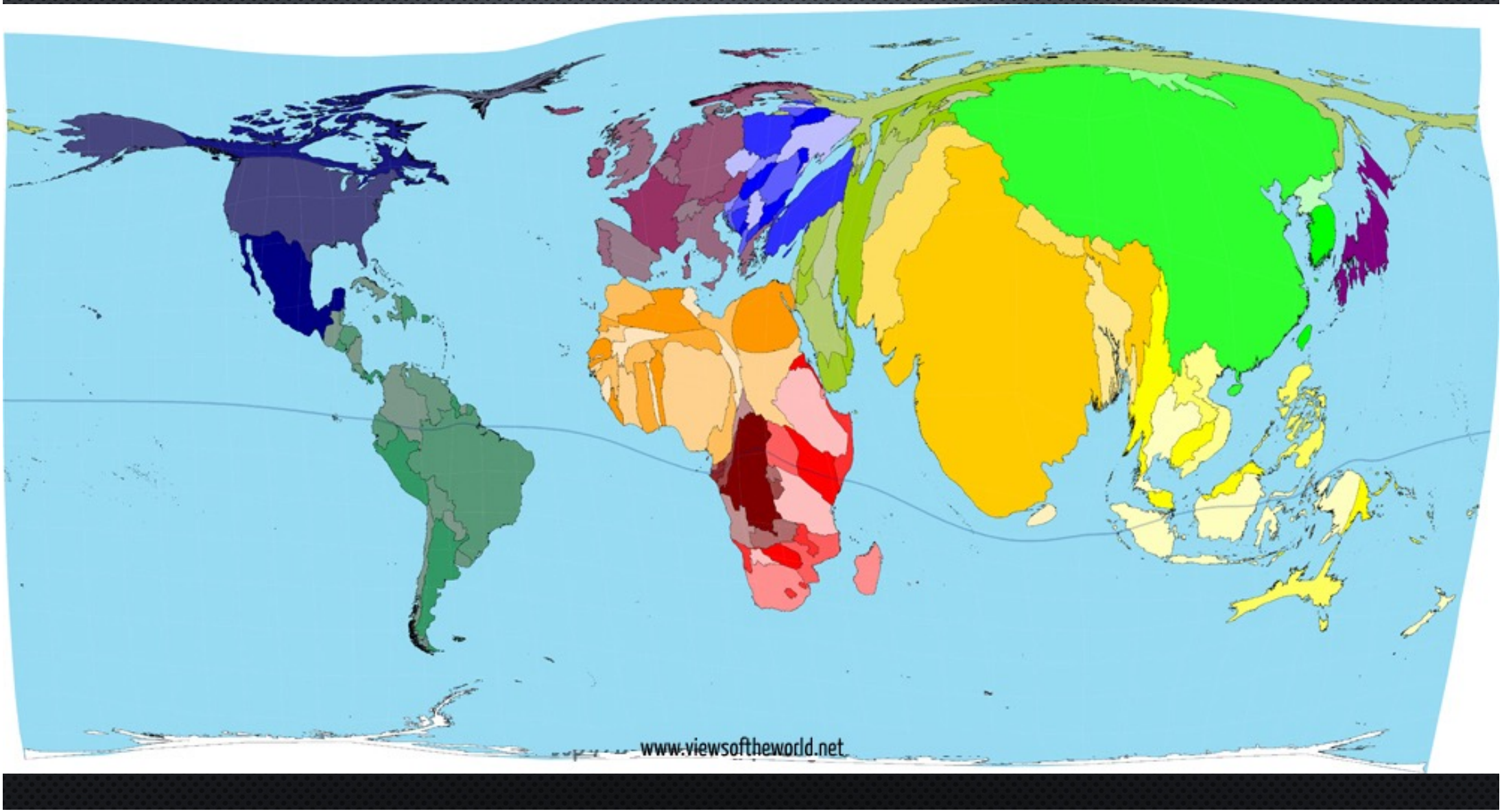
VISUALIZATIONS CAN BE MISLEADING

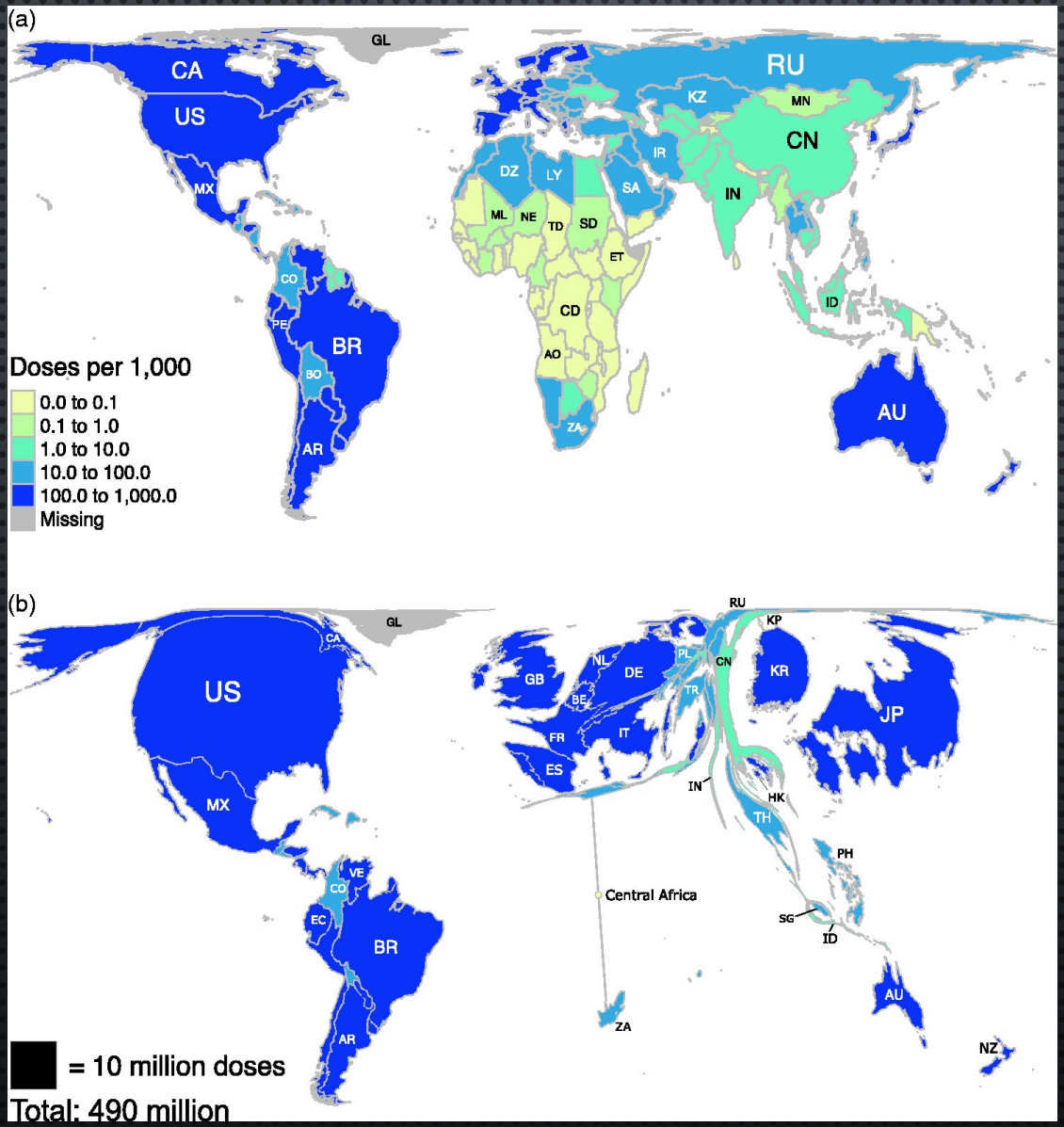
1) HENCE COMMUNICATE CLEARLY WITH STAKEHOLDERS AND VALIDATE THESE FINDINGS
2) SEE THE TSNE PLOTS AGAIN. WHAT PITFALLS DO YOU SEE IN COMMUNICATING THESE INSIGHTS TO EXPERIMENTALISTS?


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


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

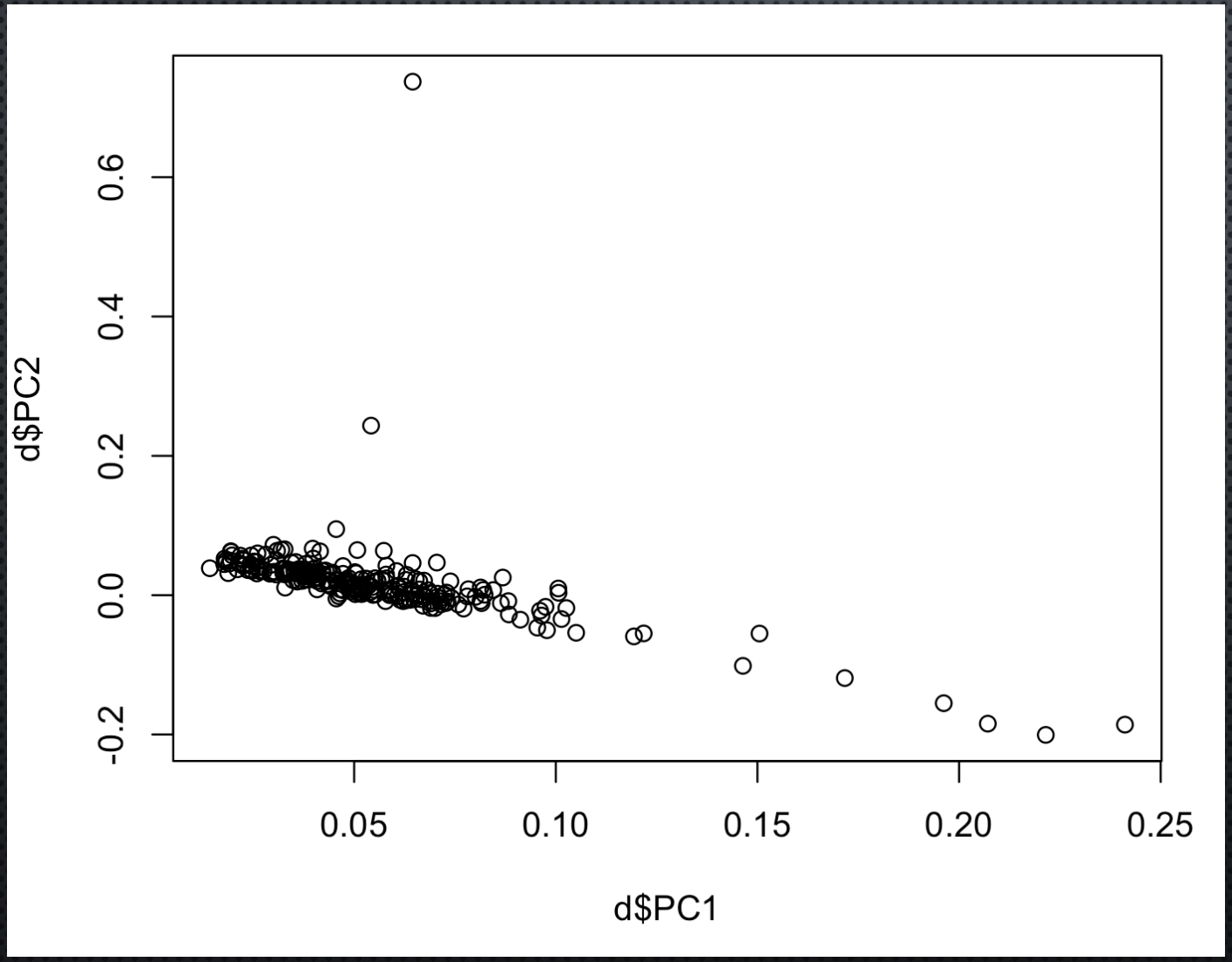

www.viewsoftheworld.net



Area proportional to number of vaccine doses

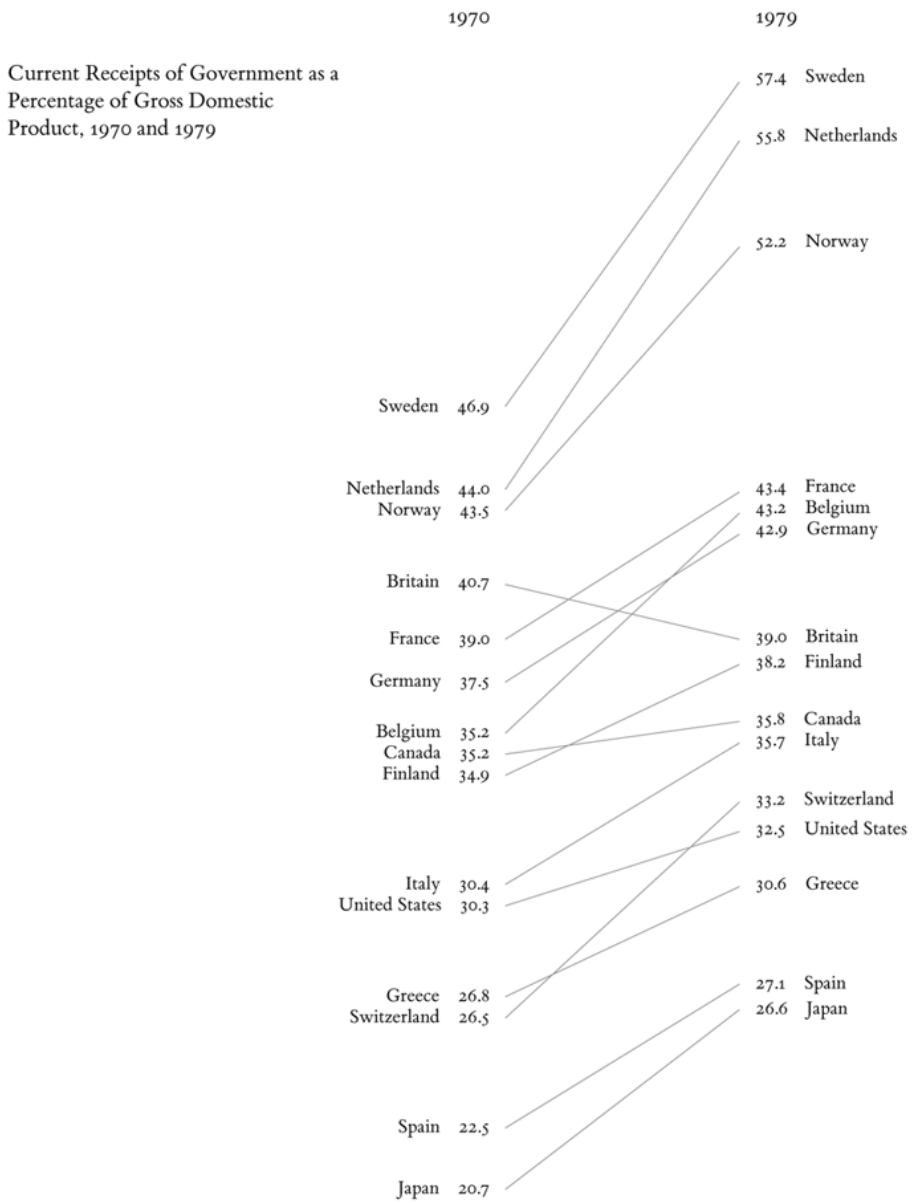
CARTOGRAMS

https://go-cart.io/tutorial



1. Removing outliers and then some downstream processing

2. t-test



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

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

BASICS AND PITFALLS IN DATA VISUALIZATION

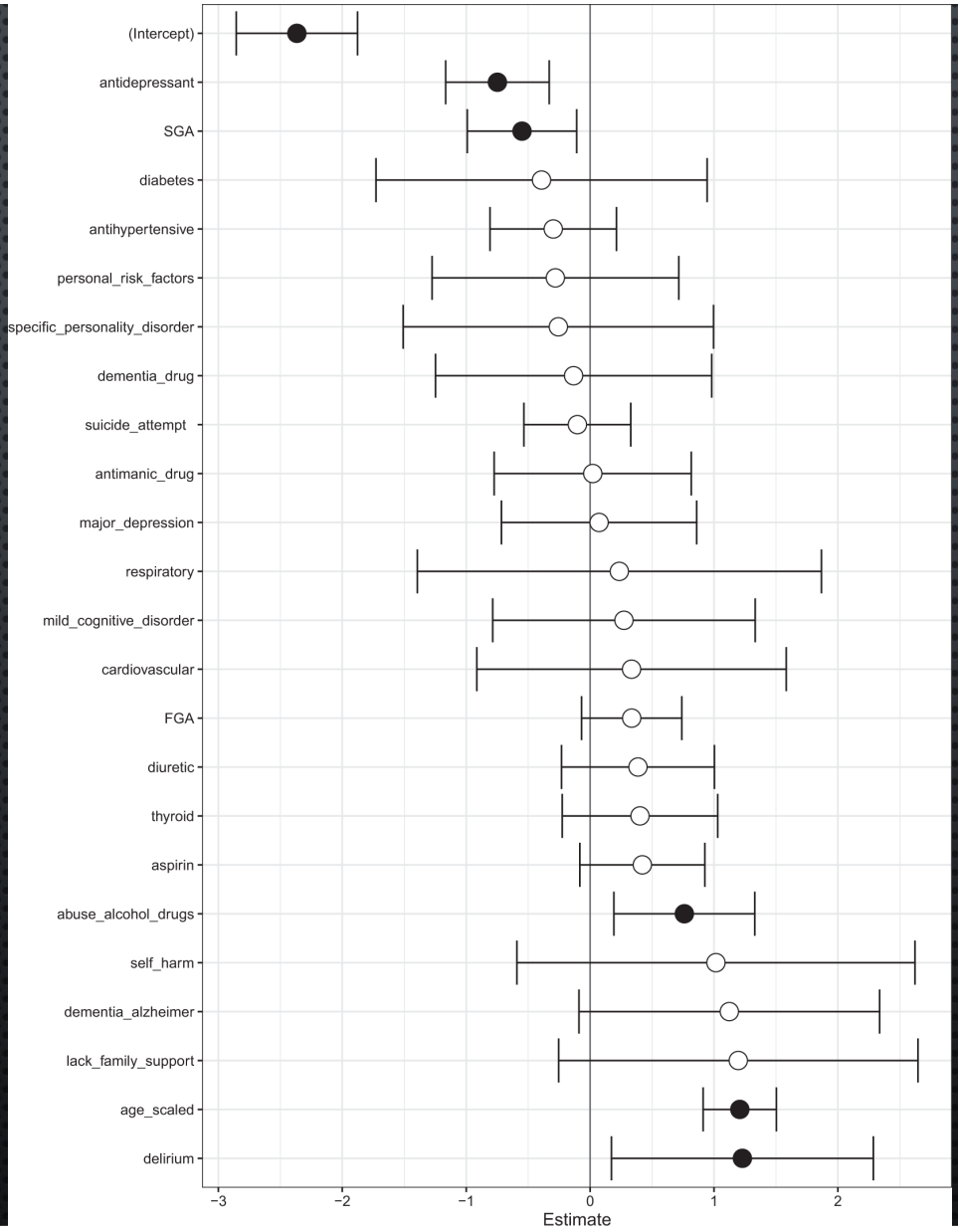
1. KNOW YOUR AUDIENCE
2. PICK VISUALIZATION BASED ON AUDIENCE
3. VISUALIZE DATA AND THEN PICK MODELS
4. ADD NARRATIVE

VISUALIZATION FOR DIAGNOSTICS

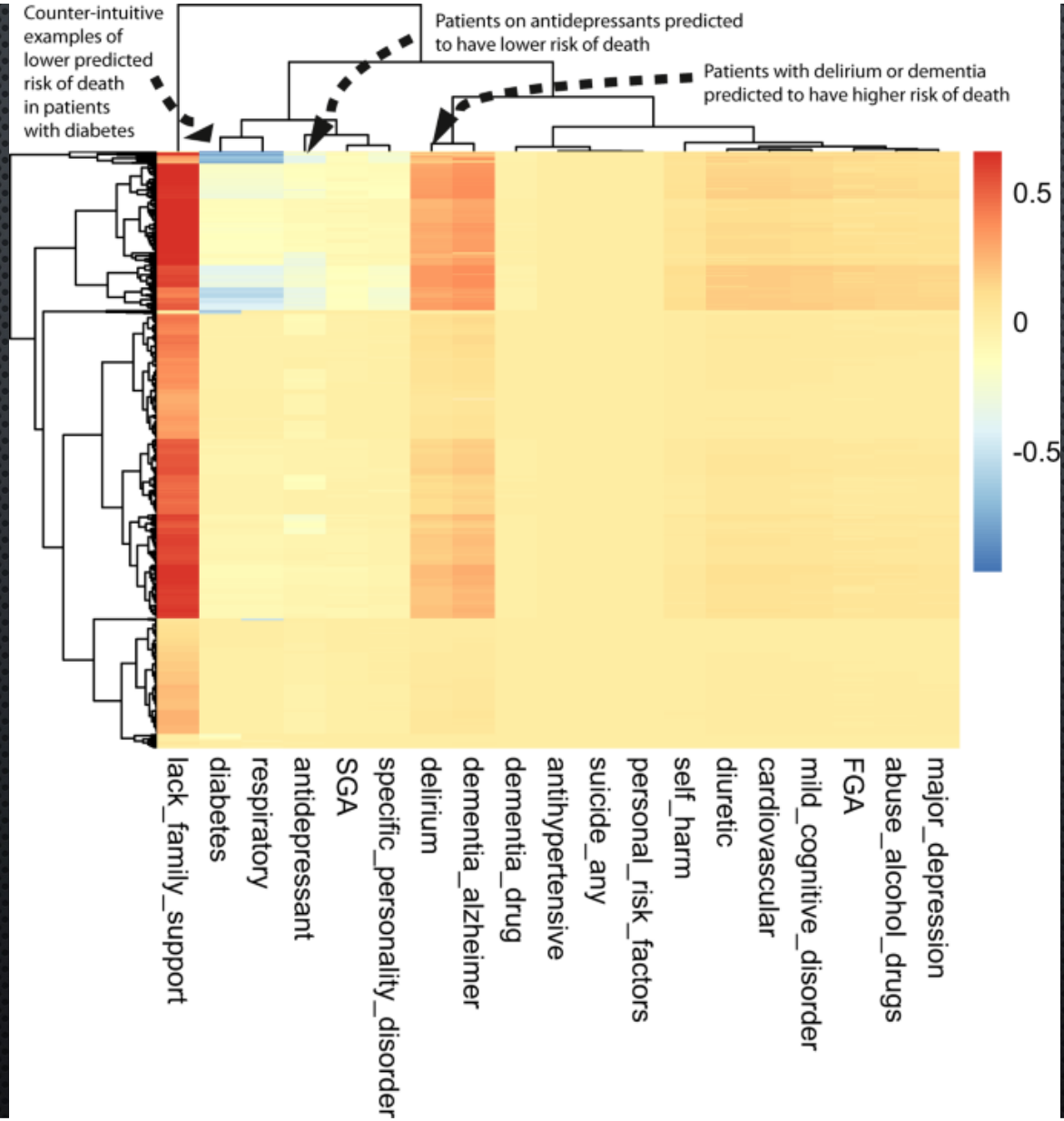
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

1. GENERALIZED LINEAR MODEL

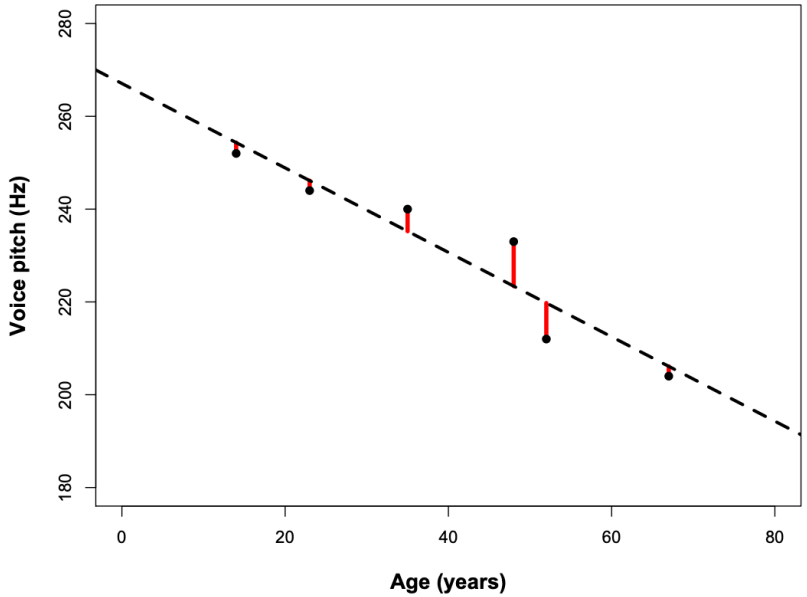
VISUALIZATION FOR DIAGNOSTICS



VISUALIZATION FOR DIAGNOSTICS

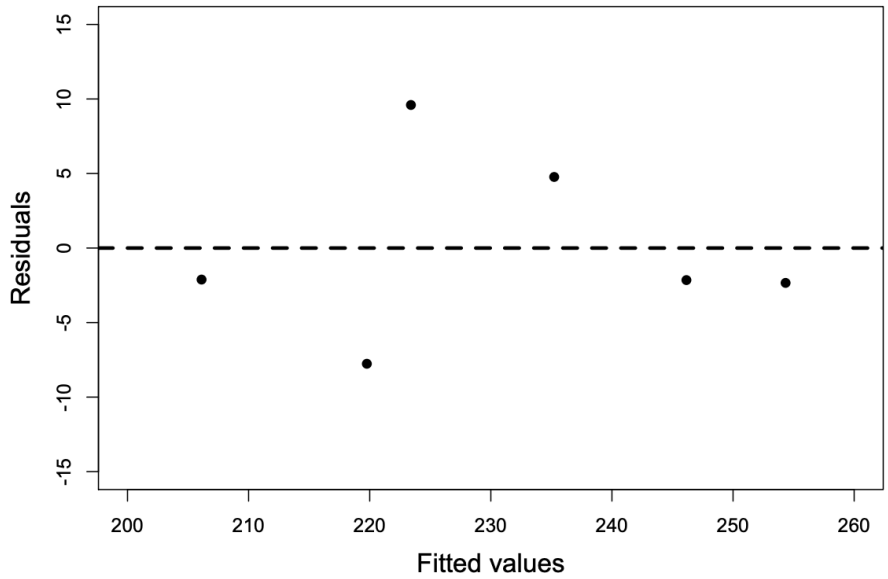


ASSUMPTIONS: LINEARITY



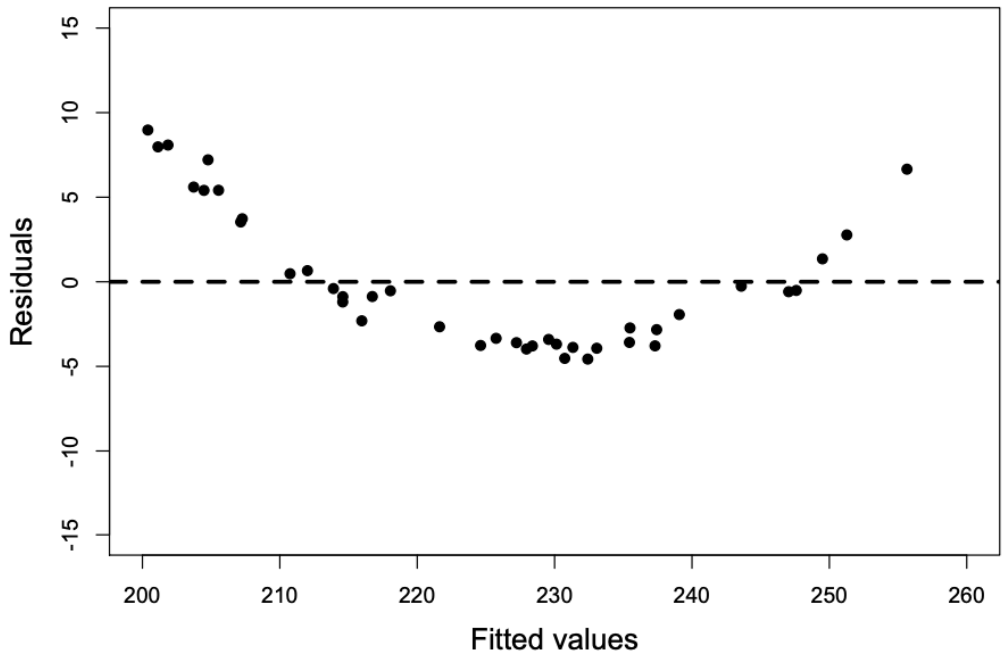
The red lines indicate the residuals, which are the deviations of the observed data points from the predicted values (the so-called "fitted values"). In this case, the residuals are all fairly small … which is because the line that represents the linear model predicts our data very well, i.e., all points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

ASSUMPTIONS: LINEARITY



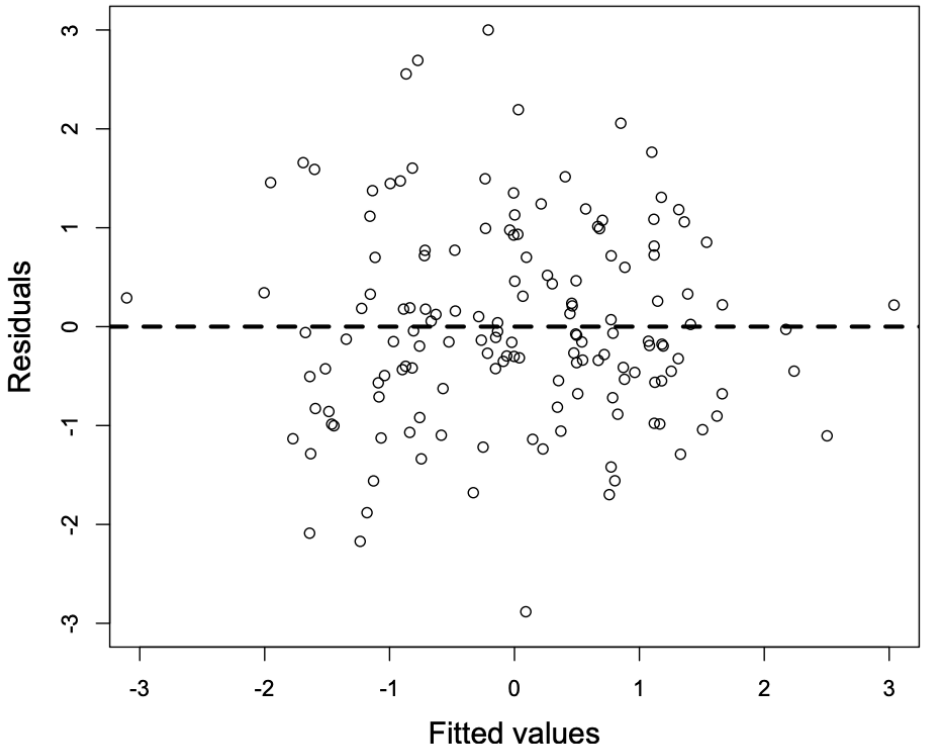
This is a residual plot. The fitted values (the predicted means) are on the horizontal line (at y=0). The residuals are the vertical deviations from this line. This view is just a rotation of the actual data (compare the residual plot with the scatterplot to see this). To construct the residual plot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

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



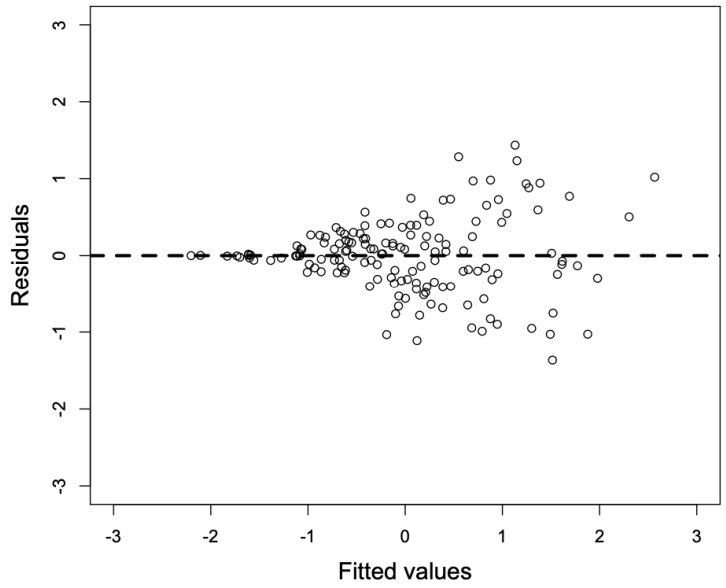
ASSUMPTIONS: LINEARITY

1. OTHER PREDICTORS
2. TRANSFORMATION OF RESPONSE
3. TRANSFORMATION OF FEATURES



ASSUMPTIONS

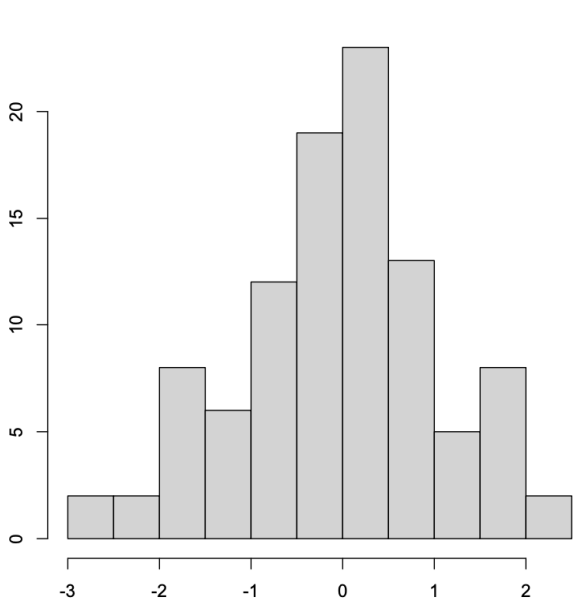
1. HOMOSKEDASTICITY
2. THE VARIABILITY OF YOUR DATA SHOULD BE APPROXIMATELY EQUAL ACROSS THE RANGE OF YOUR PREDICTED VALUES

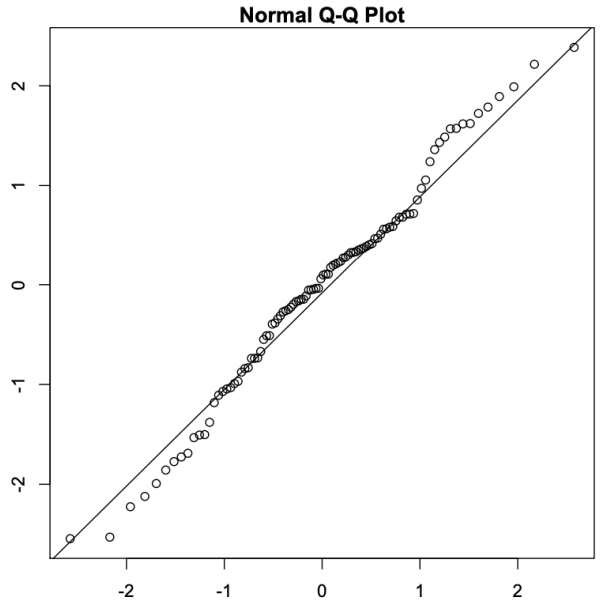


In this plot, higher fitted values have larger residuals … indicating that the model is more "off" with larger predicted means. So, the variability is not homoscedastic: it's smaller in the lower range and larger in the higher range.

ASSUMPTIONS

1. HOMOSKEDASTICITY
2. WHAT TO DO? TRANSFORM YOUR DATA

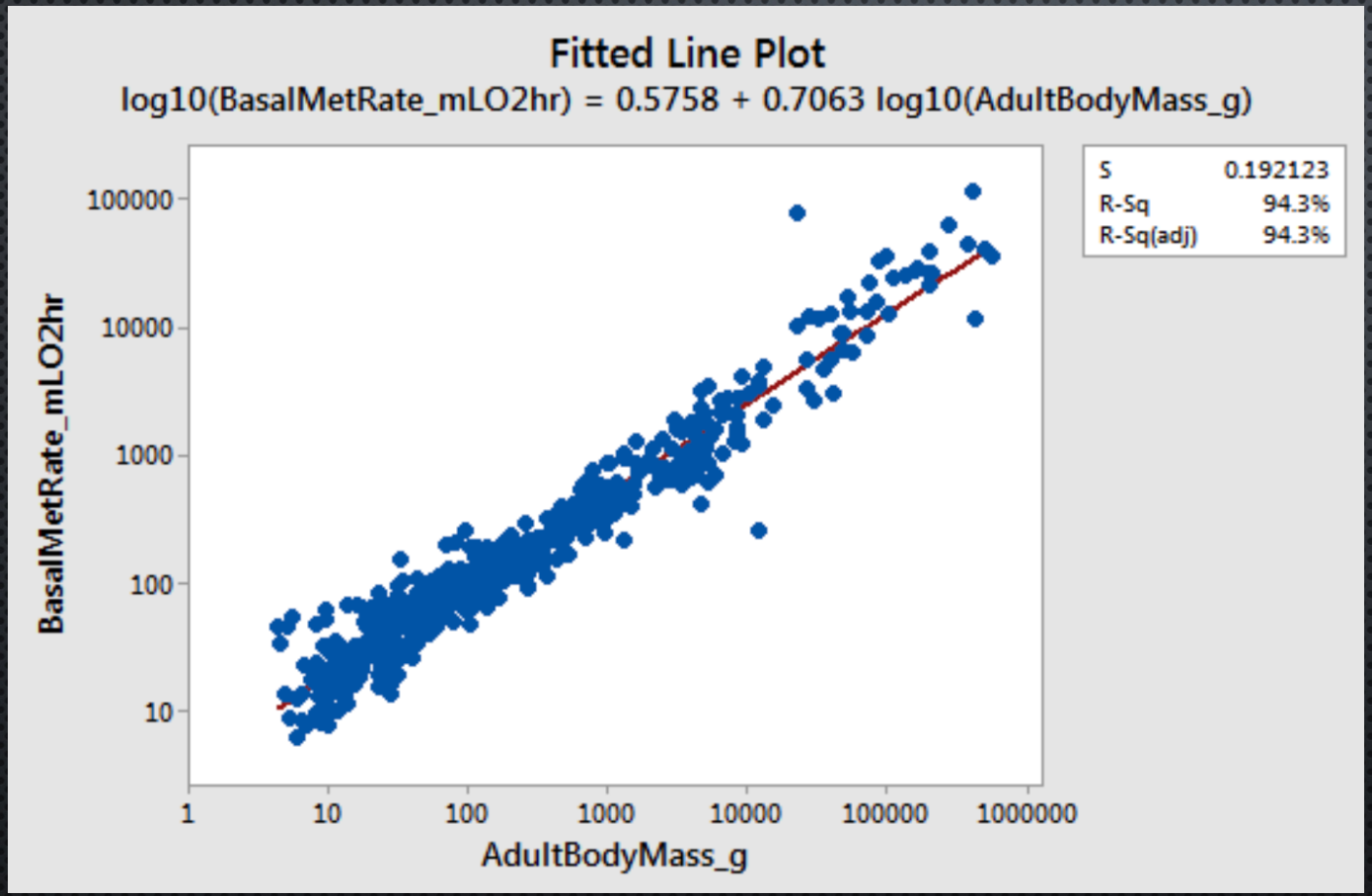




These look good. The histogram is relatively bell-shaped and the Q-Q plot indicates that the data falls on a straight line (which means that it's similar to a normal distribution). Here, we would conclude that there are no obvious violations of the normality assumption.

ASSUMPTIONS

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

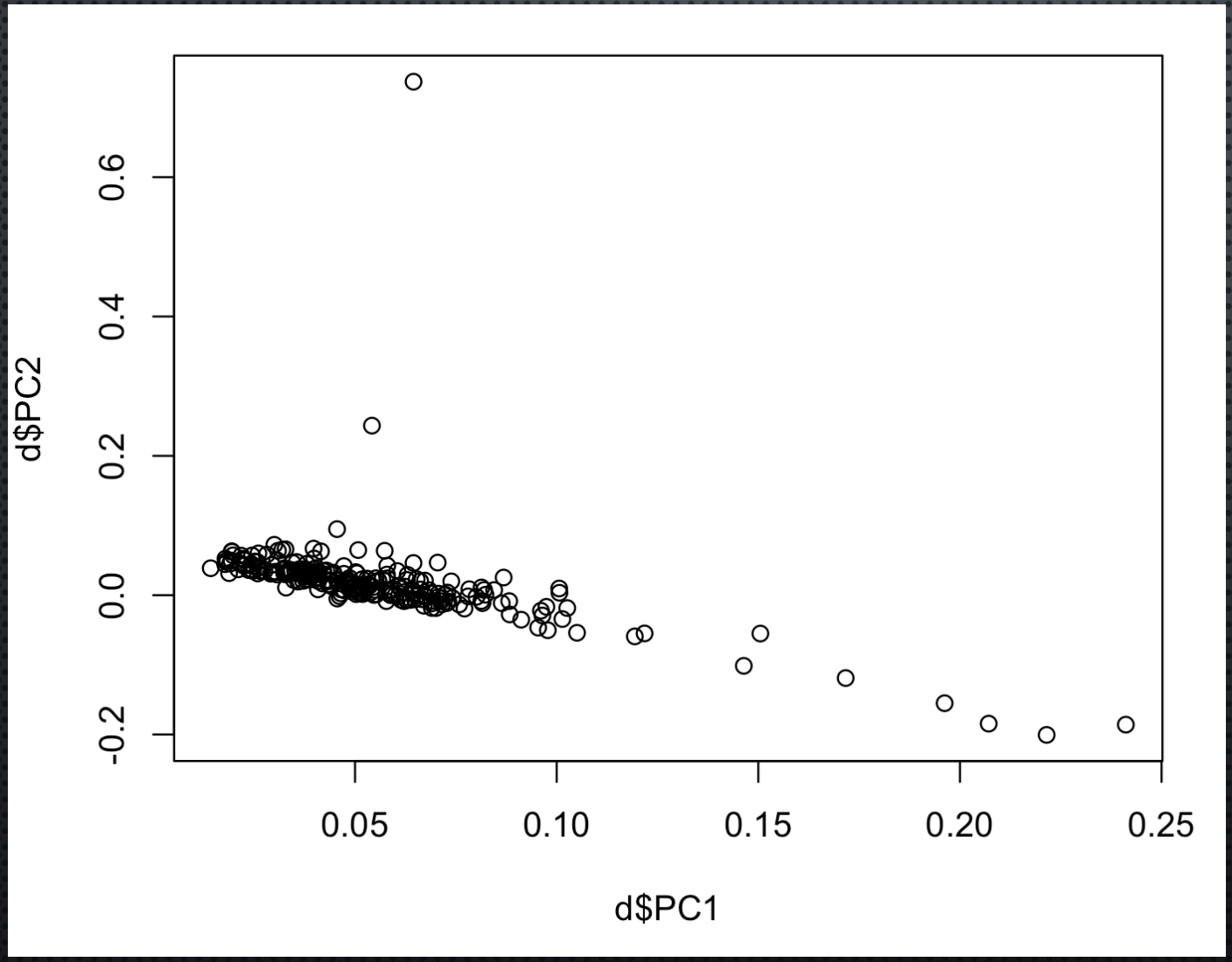
```
> dfbeta(xmdl)
  (Intercept)         age
1  -3.3645662  0.06437573
2  -1.6119656  0.02736278
3   1.5481303 -0.01456709
4  -0.0259835  0.05092767
5   0.8707699 -0.06479736
6   1.8551808 -0.06622744
```

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



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

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 AND AUTOCORRELATIONS

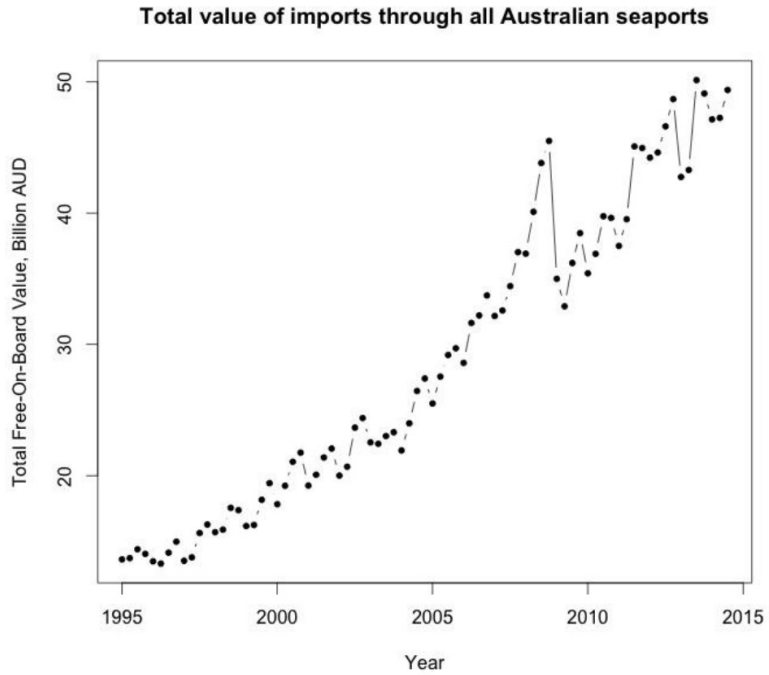


Figure 1. Total Free-On-Board value of Australian imports by sea through all ports.

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.

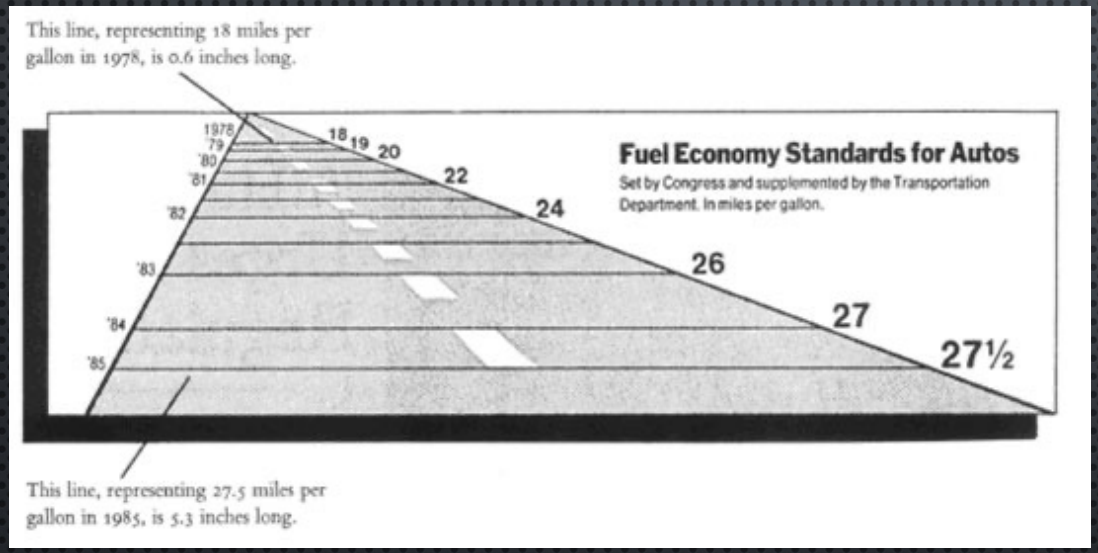
Visualization may also reveal if the underlying model/assumptions may have changed after a certain time. For example, in financial time-series data, there usually is a change after 2008 due to the global financial crisis.

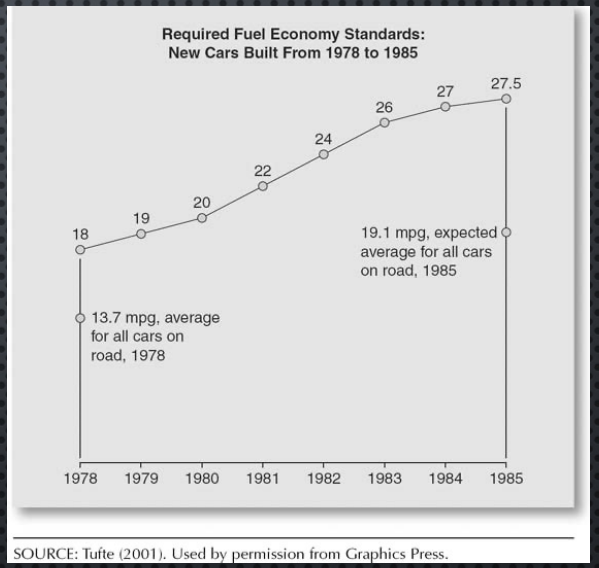
This may suggest that a new model or more data is required.

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

1. GENERALIZED LINEAR MODEL

TIME SERIES DATA VISUALIZATION

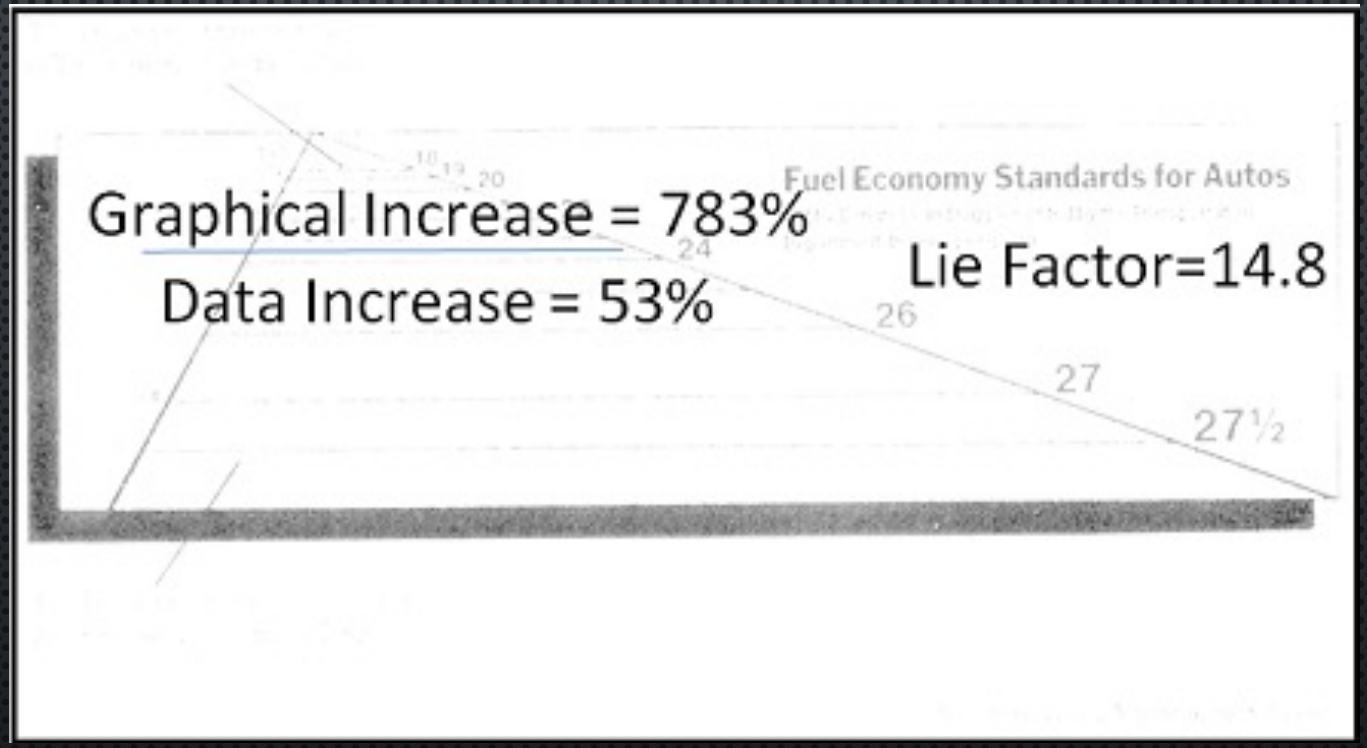




SOURCE: Tufte (2001). Used by permission from Graphics Press.

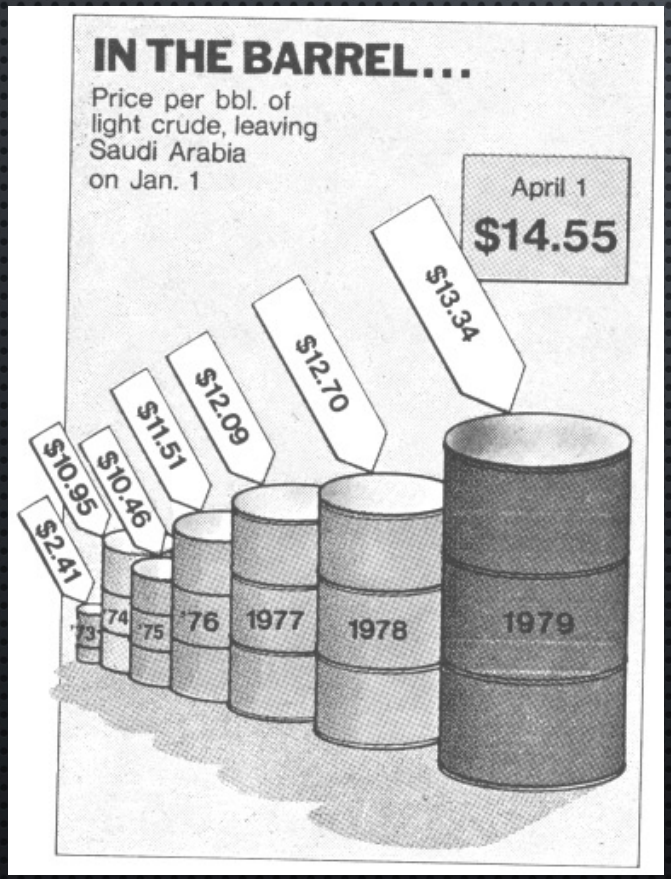
1. Area and volumes
2. Design effect vs. data effect

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/

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

APPS FOR RAPID PROTOTYPING AND COMMUNICATION

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

WHY AND WHEN 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

MATERIAL

Material, 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/

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

ACTIVITIES

North korea missile range animation

Critique

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

ACTIVITIES

Create cartogram online

https://go-cart.io/tutorial

ACTIVITIES

Create interactive data visualization

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

TIME SERIES DATA AND AUTOCORRELATIONS

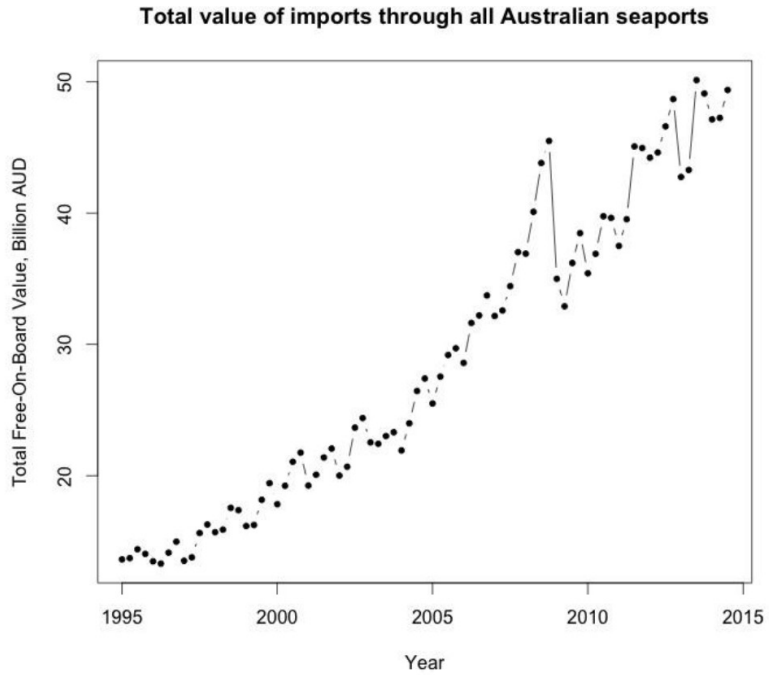


Figure 1. Total Free-On-Board value of Australian imports by sea through all ports.

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

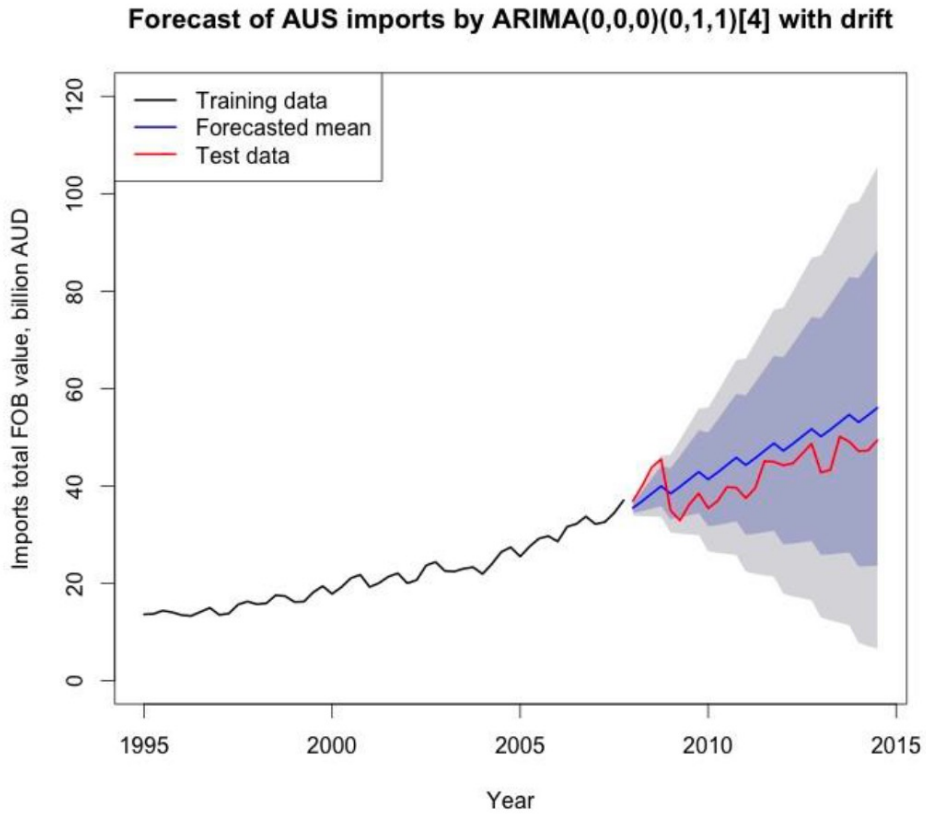


Figure 2. Forecast of port throughput (imports) by best fitting Seasonal ARIMA (SARIMA) model. In-sample (training set) AIC=37.66, out-of-sample (test set) Root Mean Squared Error RMSE=10190.

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:

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

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)