Visualizing associations

HUGEN 2073

Genomic Data Visualization and Integration

Slides borrowed/adapted from H.J. Park with permission

Some statistical tests for associations/trends…
… and some natural complementary visualizations

Common Statistical Tests

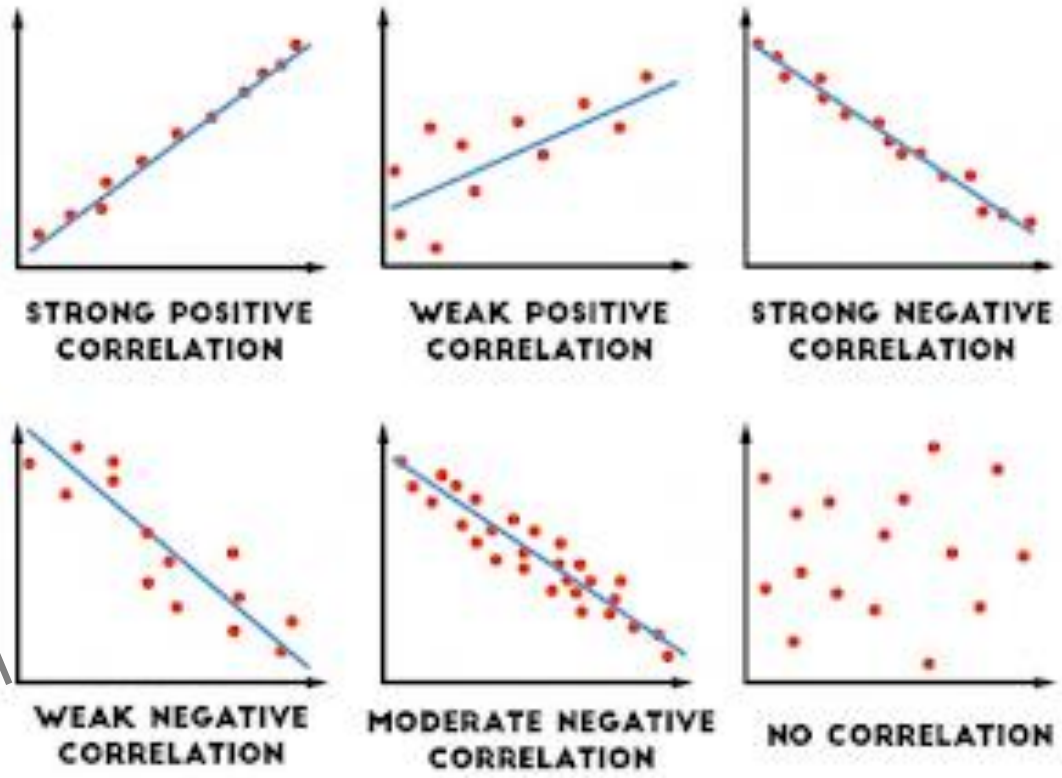
Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	*look for the difference between the means of variables*
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	*assess if change in one variable predicts change in another variable*
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	*are used when the data does not meet assumptions required for parametric tests*
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

Visualizing correlations

Common Statistical Tests

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	*look for the difference between the means of variables*
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	*assess if change in one variable predicts change in another variable*
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	*are used when the data does not meet assumptions required for parametric tests*
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

Correlation is a linear relationship



https://medium.com/@dipti.rohan.pawar/correlation-statistical-analysis-9471411f0431

Visualizing a difference in means

Common Statistical Tests

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	*look for the difference between the means of variables*
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	*assess if change in one variable predicts change in another variable*
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	*are used when the data does not meet assumptions required for parametric tests*
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

Group mean difference implies difference in two distributions

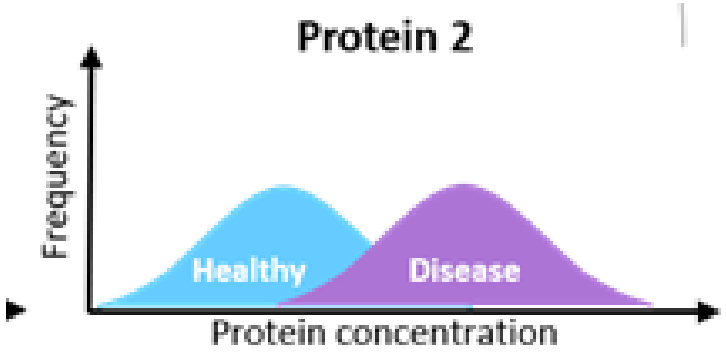


Figure 2. Overlapping histogram plots for concentrations of protein 2 in different populations.

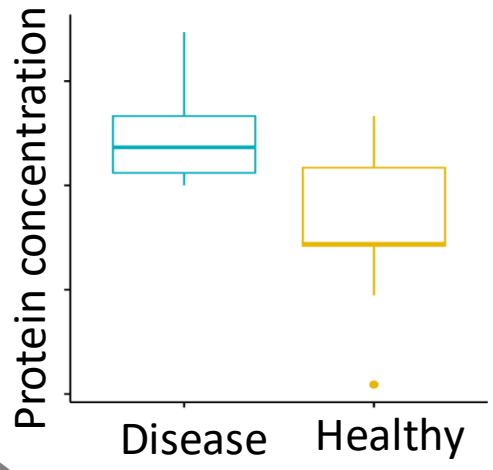
https://raybiotech.com/learning-center/t-test-anova/

Visualizing a difference in means (or medians)

Common Statistical Tests

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	*look for the difference between the means of variables*
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	*assess if change in one variable predicts change in another variable*
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	*are used when the data does not meet assumptions required for parametric tests*
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

Group mean difference implies difference in two distributions



Adapted from http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r

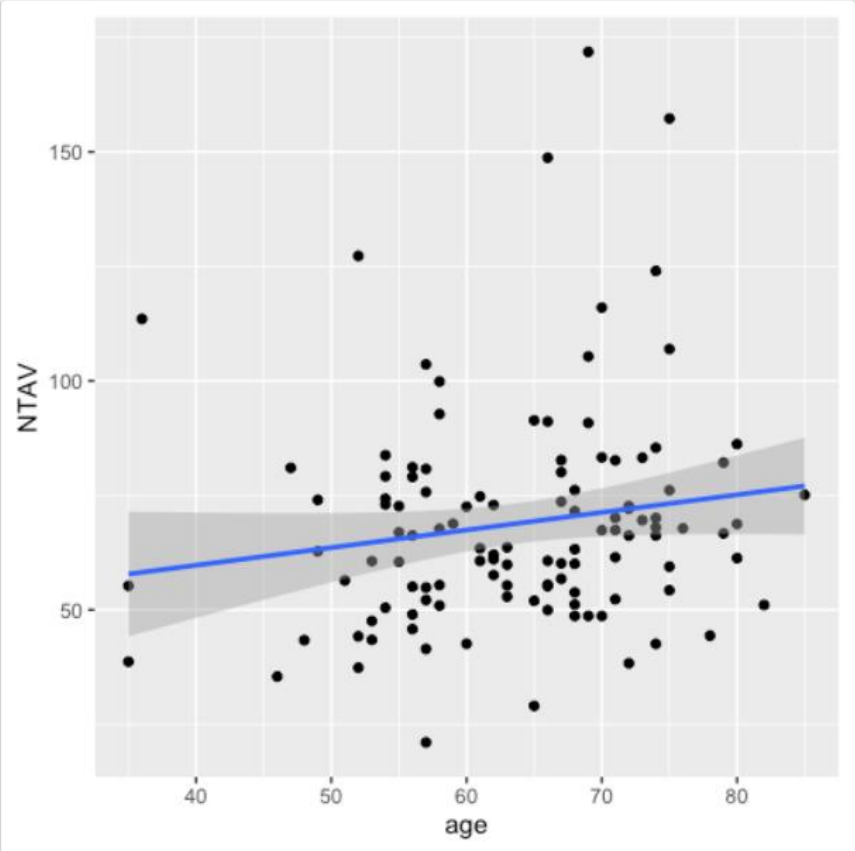
Visualizing a regression

Common Statistical Tests

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	*look for the difference between the means of variables*
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	*assess if change in one variable predicts change in another variable*
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	*are used when the data does not meet assumptions required for parametric tests*
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

Linear regression implies linear relationship

```
require(ggplot2)
ggplot(radial,aes(y=NTAV,x=age))+geom_point()+geom_smooth(method="lm")
```



https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html

Visualizing (multivariate) regression

Common Statistical Tests

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	*look for the difference between the means of variables*
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	*assess if change in one variable predicts change in another variable*
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	*are used when the data does not meet assumptions required for parametric tests*
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

Linear regression with three independent variables

```
fit4=lm(NTAV~age*weight*HBP,data=radial)
summary(fit4)
```

```
ggPredict(fit4,interactive = TRUE)
```

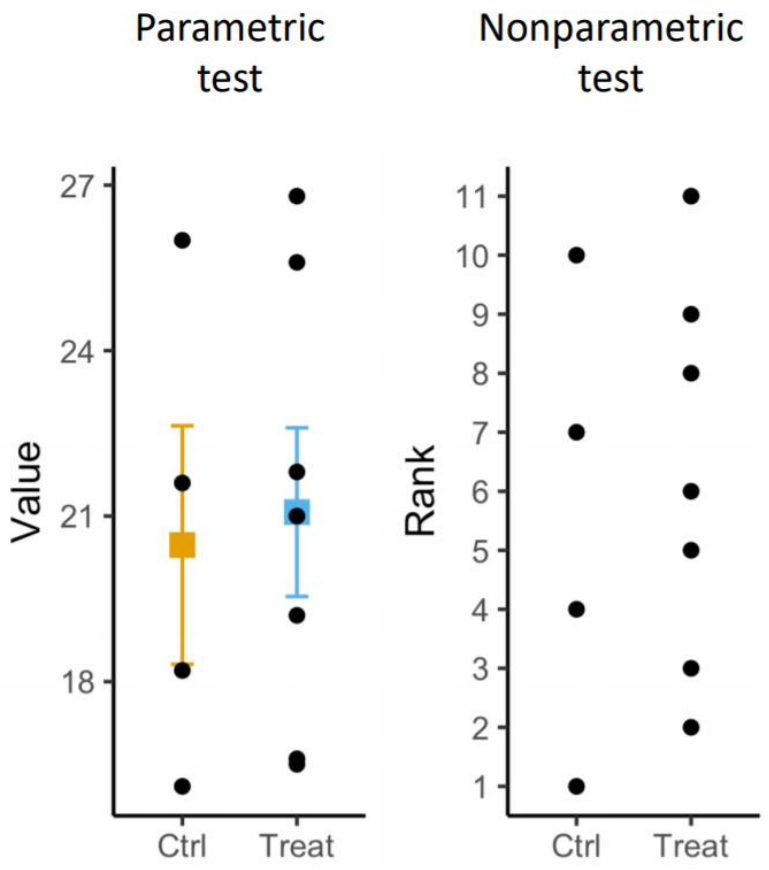
The third independent variable can be encoded in color

https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html

Visualizing non-parametric analyses

Common Statistical Tests

Type of Test:	Use:
Correlational	These tests look for an association between variables
Pearson correlation	Tests for the strength of the association between two continuous variables
Spearman correlation	Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data)
Chi-square	Tests for the strength of the association between two categorical variables
Comparison of Means:	*look for the difference between the means of variables*
Paired T-test	Tests for difference between two related variables
Independent T-test	Tests for difference between two independent variables
ANOVA	Tests the difference between group means after any other variance in the outcome variable is accounted for
Regression:	*assess if change in one variable predicts change in another variable*
Simple regression	Tests how change in the predictor variable predicts the level of change in the outcome variable
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable
Non-parametric:	*are used when the data does not meet assumptions required for parametric tests*
Wilcoxon rank-sum test	Tests for difference between two independent variables - takes into account magnitude and direction of difference
Wilcoxon sign-rank test	Tests for difference between two related variables - takes into account magnitude and direction of difference
Sign test	Tests if two related variables are different – ignores magnitude of change, only takes into account direction

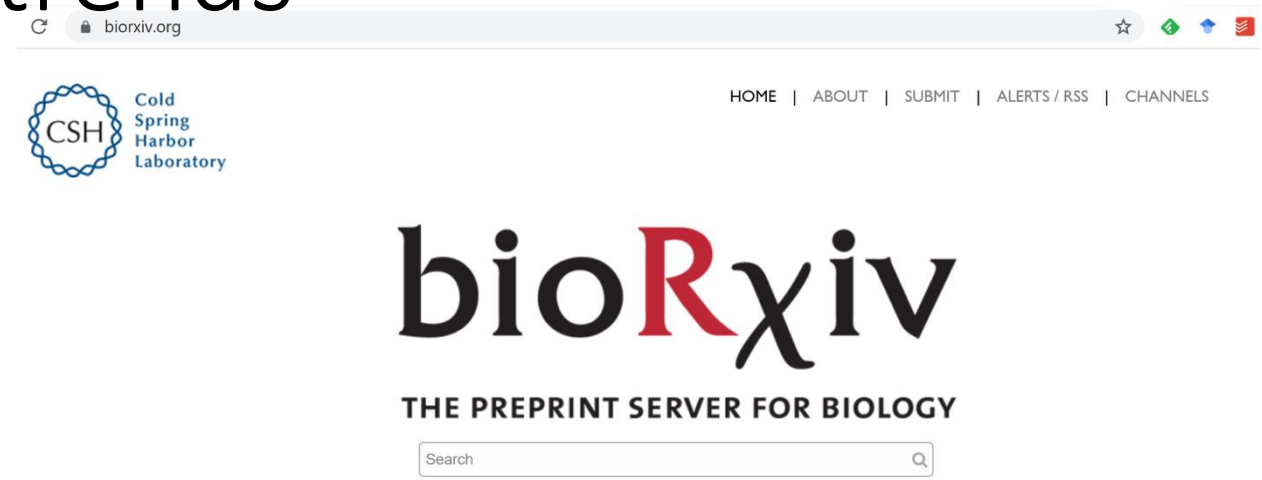


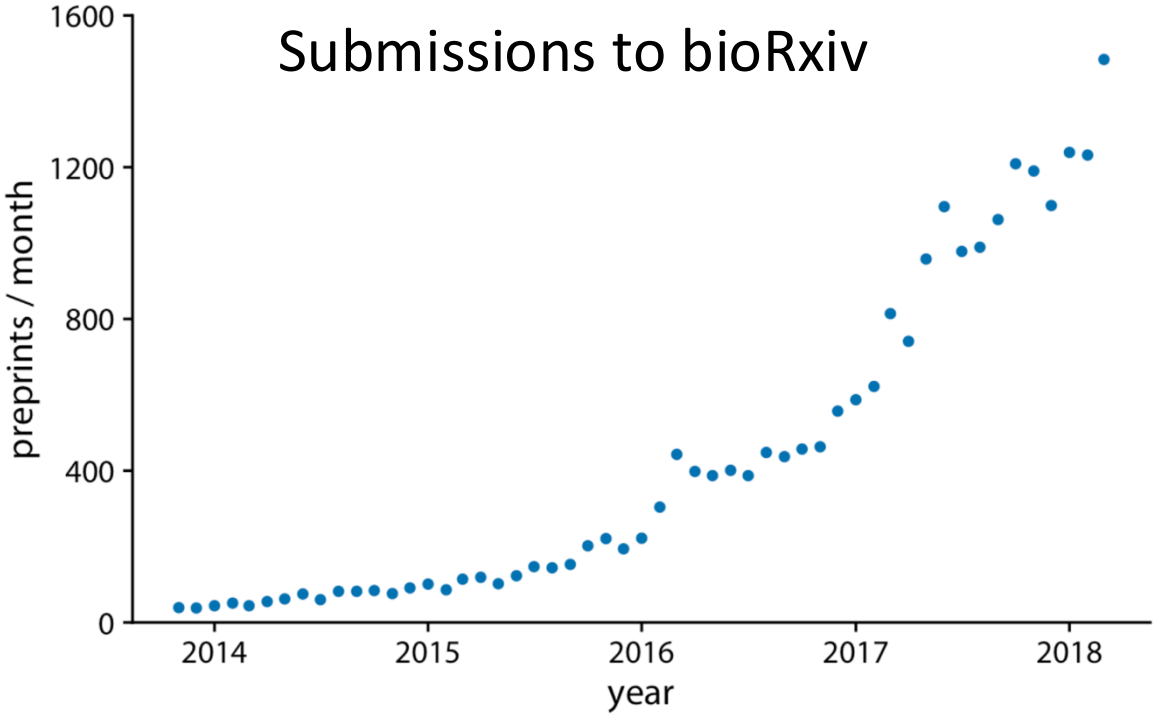
http://www.compbio.dundee.ac.uk/user/mgierlinski/talks/pvalues2/p-values5.pdf

Scatterplots *can* show trends

Scatterplot

- Plots the combination of x-/y-axis values
- Can also represent the relationship across the x-axis





Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

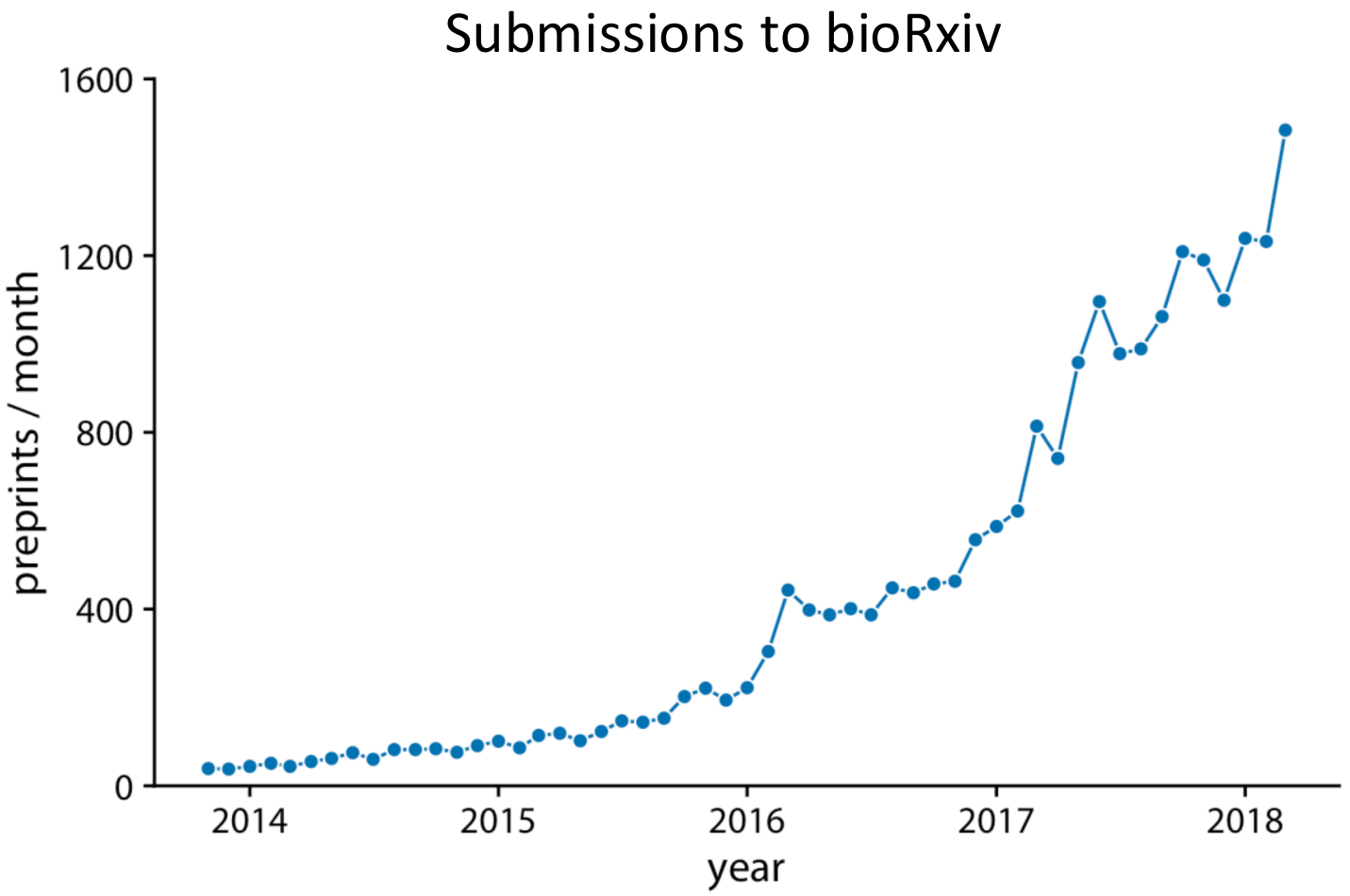
Line graphs are often better

Like scatterplot:

- Data plotted against x-/y-axis

Unlike scatterplot: Dots

- Have exactly one left and right neighbor
- E.g., time points or in a controlled experiment where a treatment is set to different values



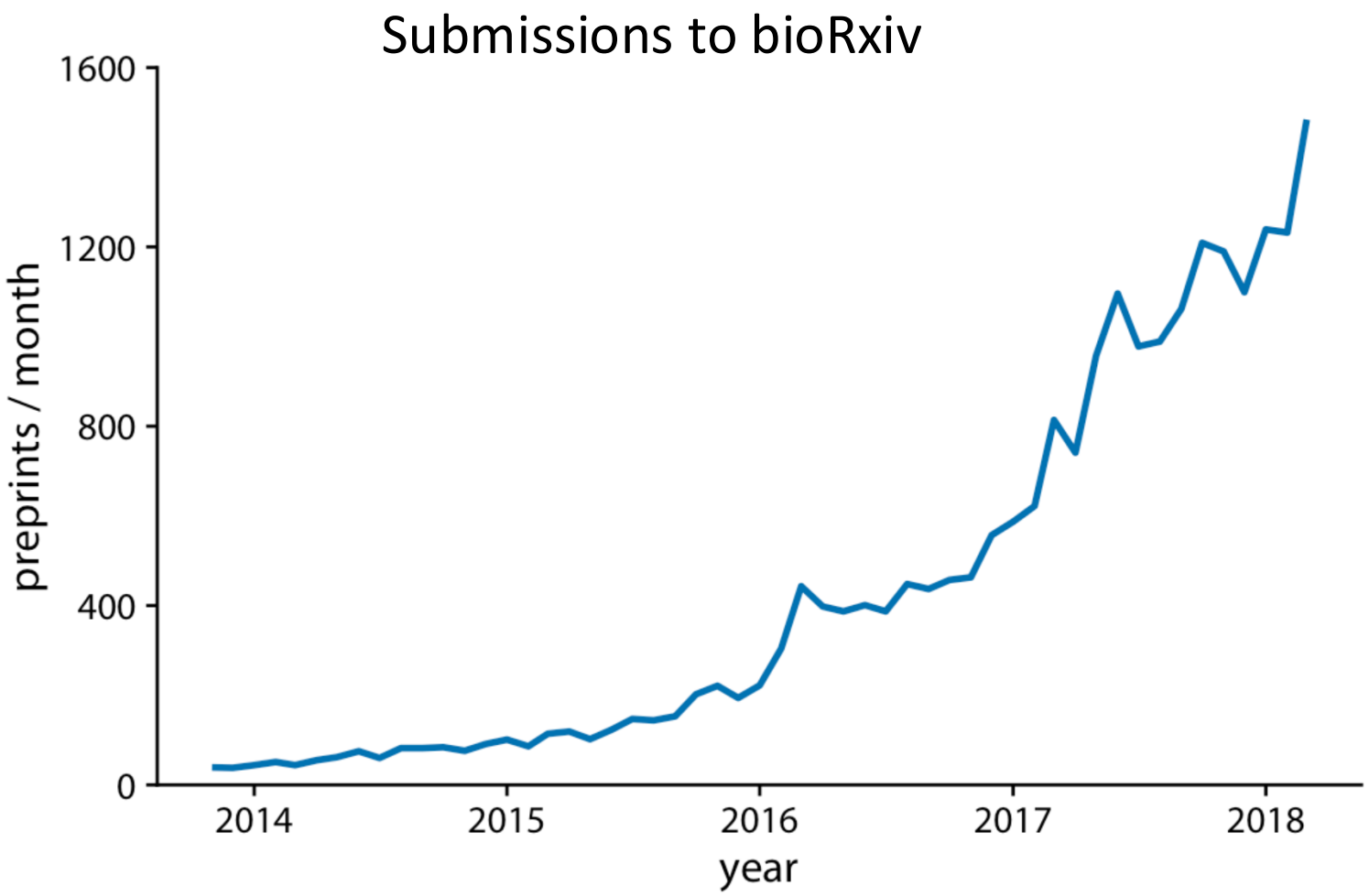
Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

A line graph without datapoints?

- Pros: less busy
- Cons: lack of actual data points

➡ OK to use if

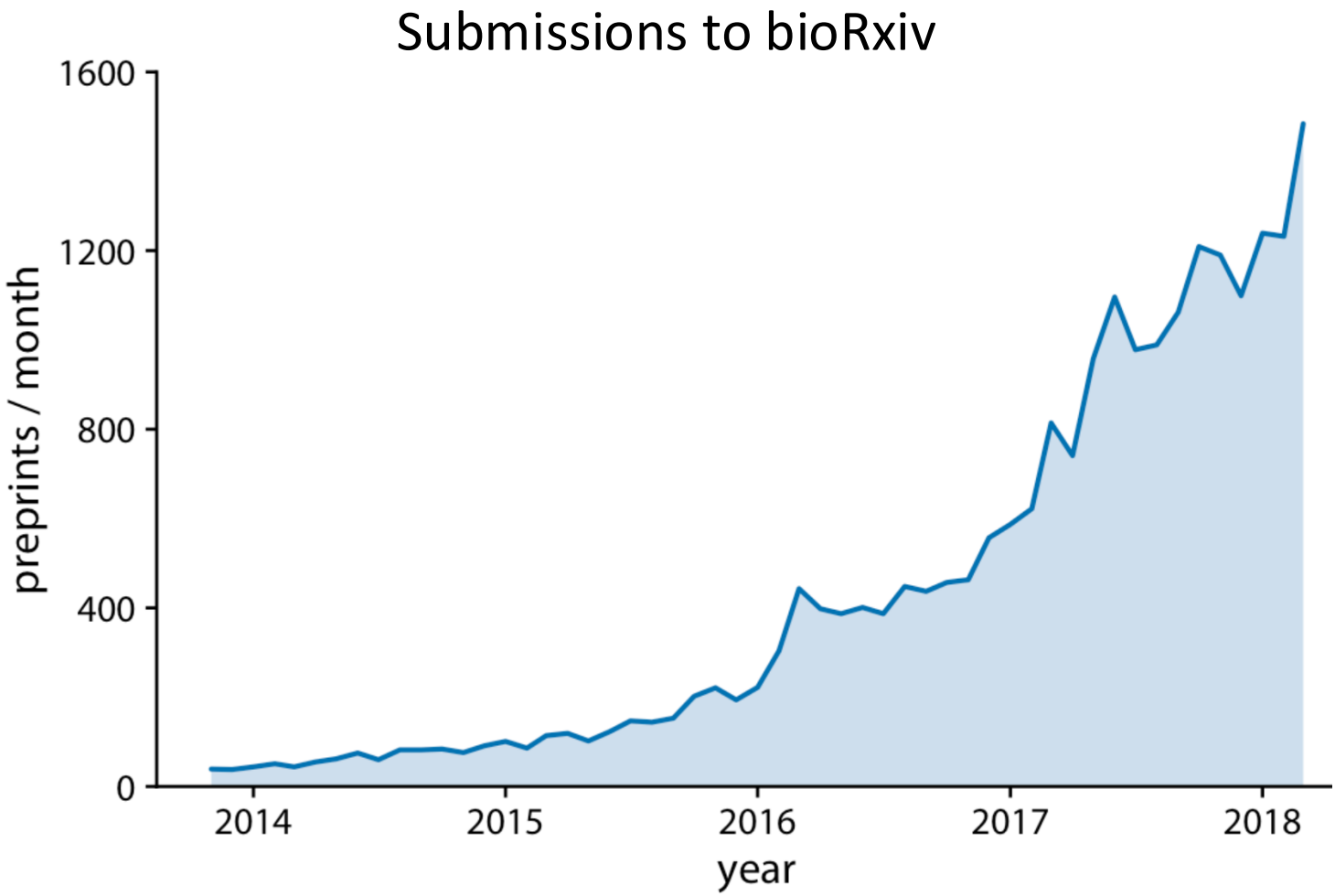
- Data points are dense
- You want to emphasize trend



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Shaded line graph to further highlight a trend?

- Shaded area perceived to be more “real” or physical?
- To represent “physicality”, y-axis needs to start at 0

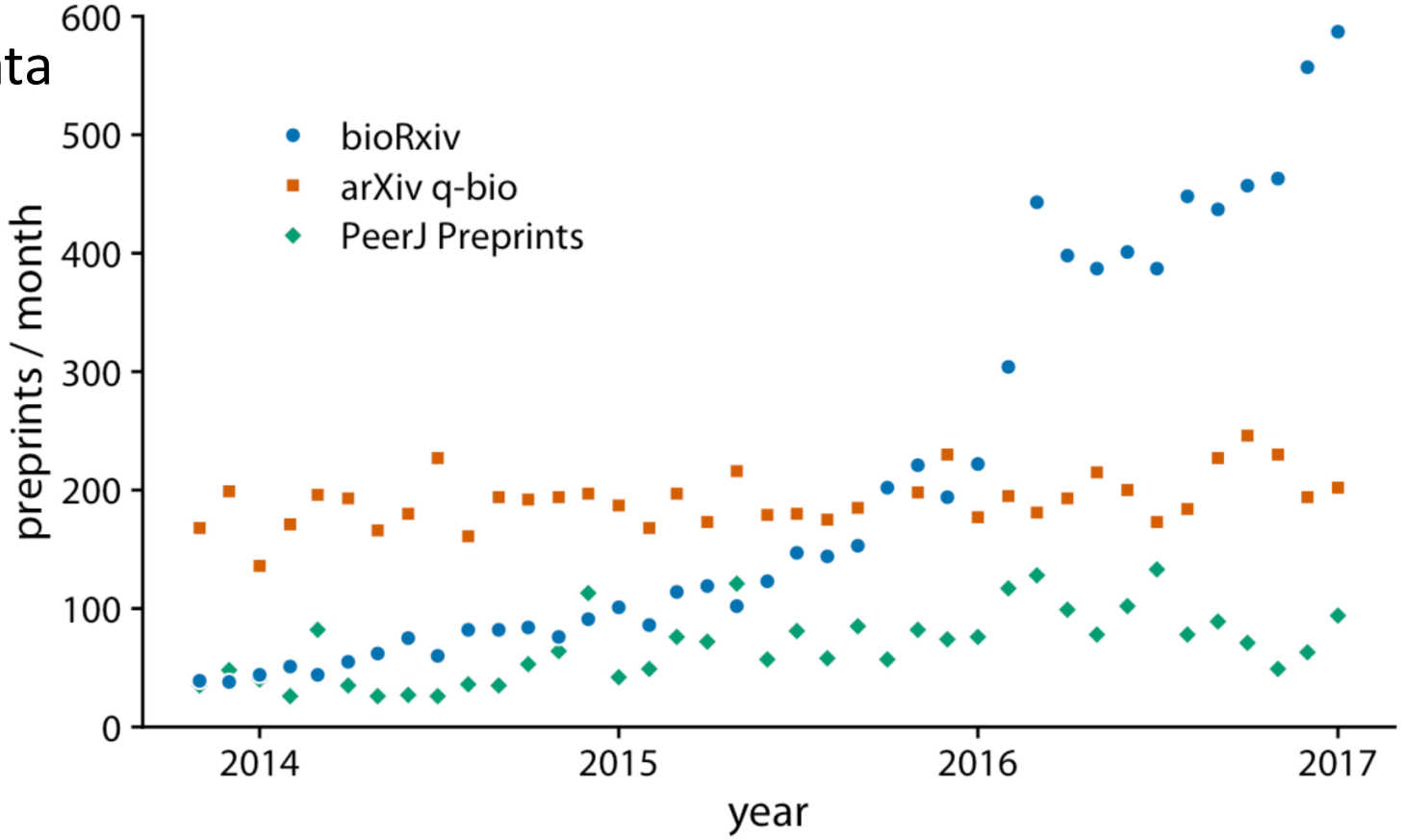


Fundamentals of Data Visualization, Wilke, O’Reilly, 1st Ed.

Scatterplot to show trend in multiple time-series?

- Showing **many** independent data points in three groups

Monthly submissions to three preprint servers covering biomedical research

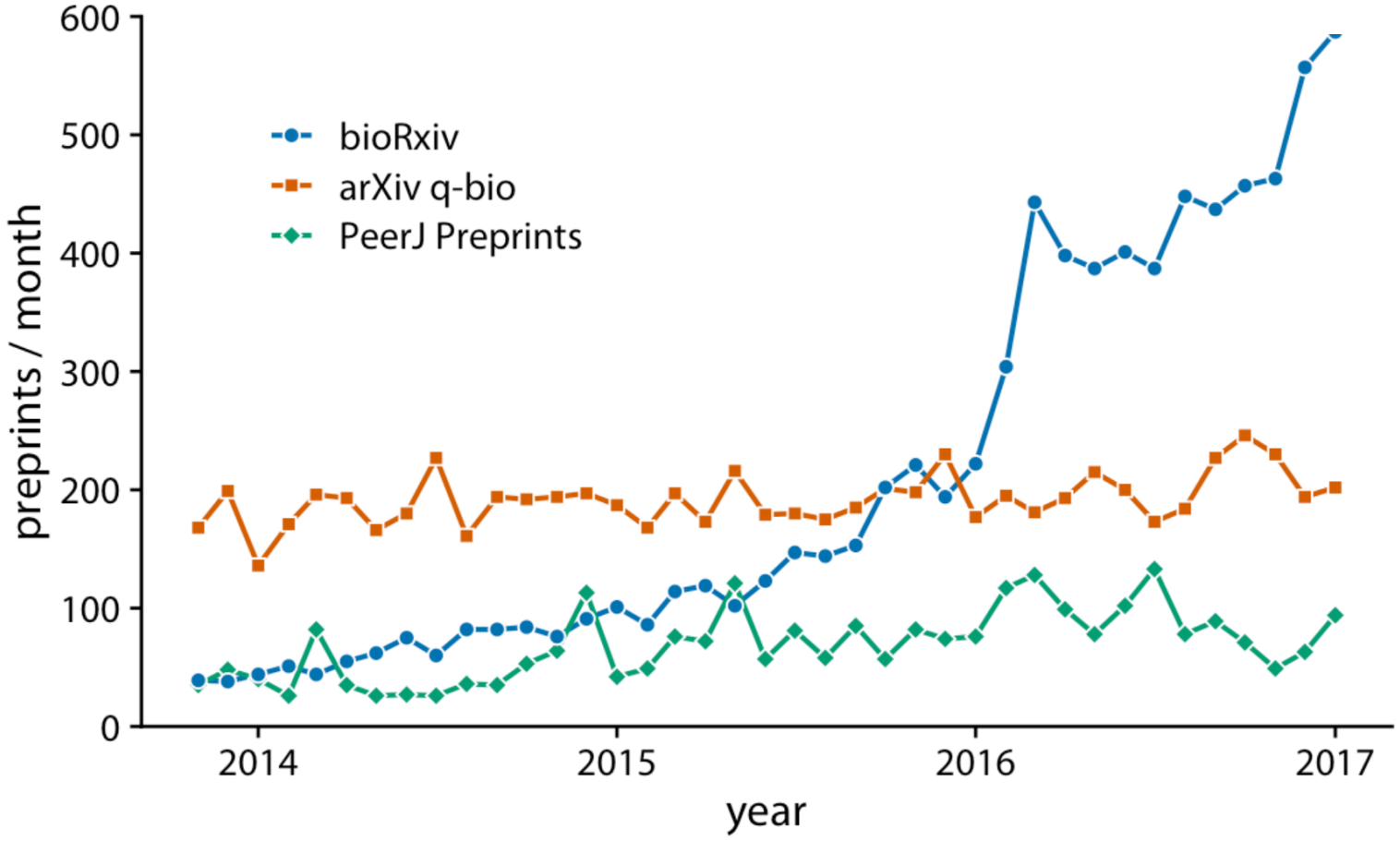


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Line graph to show trend in multiple time-series

- Showing **three** independent trends

Monthly submissions to three preprint servers covering biomedical research



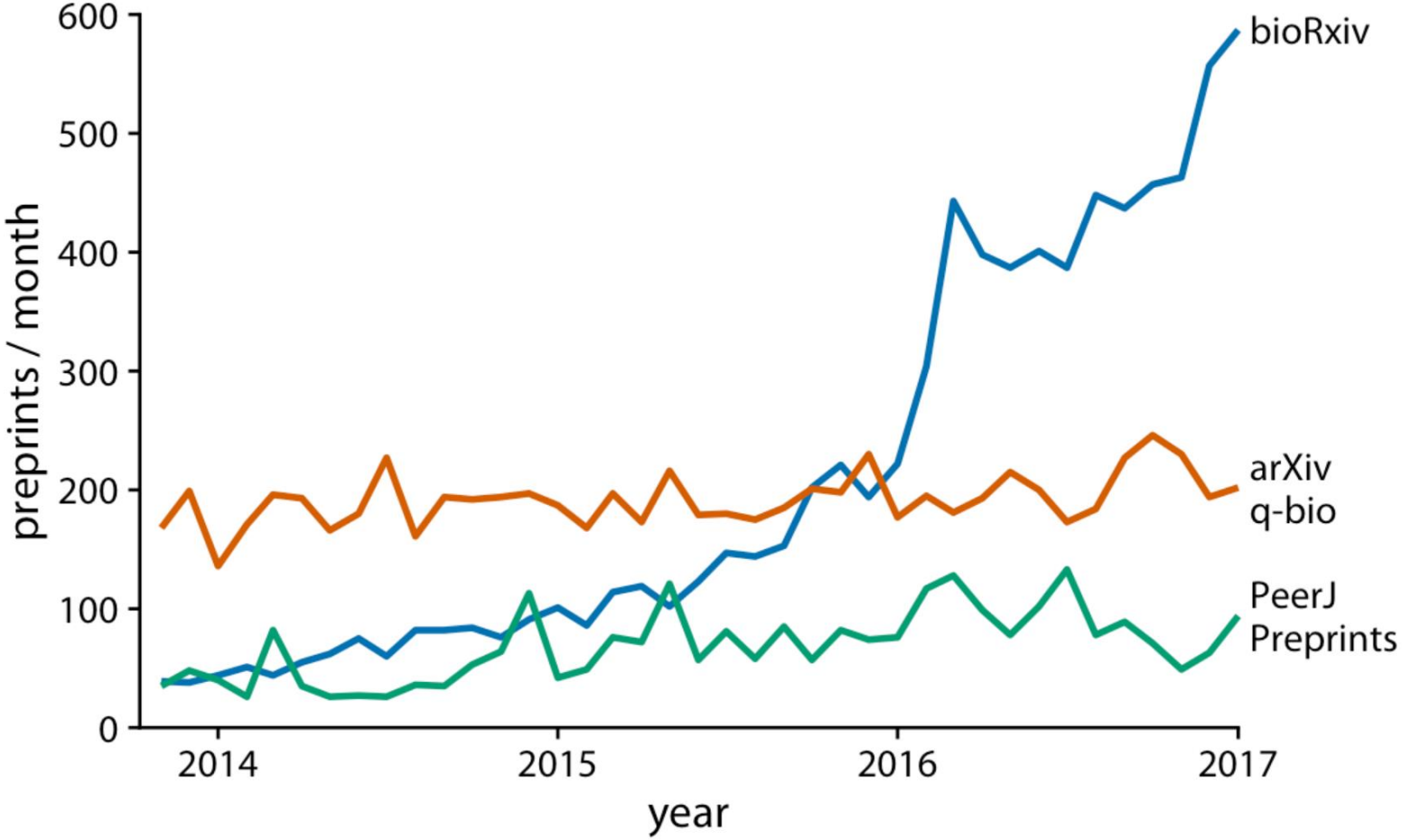
Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Line graph to show trend in multiple time-series

Improvements (?):

1. Data points dropped?
2. Legend replaced

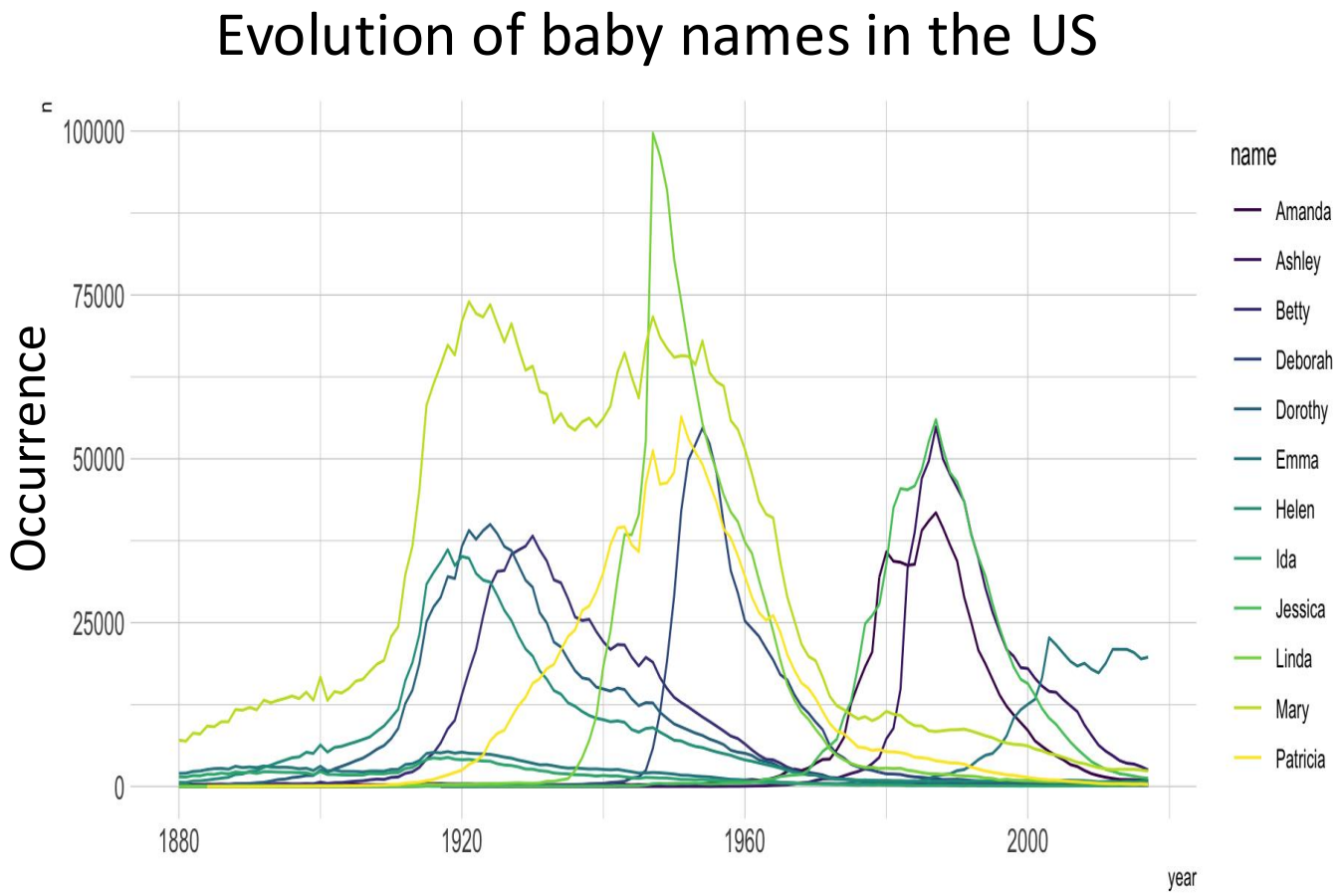
Monthly submissions to three preprint servers covering biomedical research



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

The spaghetti plot problem

- Each additional trendline might not seem like “too much”
- But the total effect can be a “spaghetti plot”
- Difficult to locate a specific line



https://www.data-to-viz.com/caveat/spaghetti.html

Multiples to avoid the spaghetti plot problem?

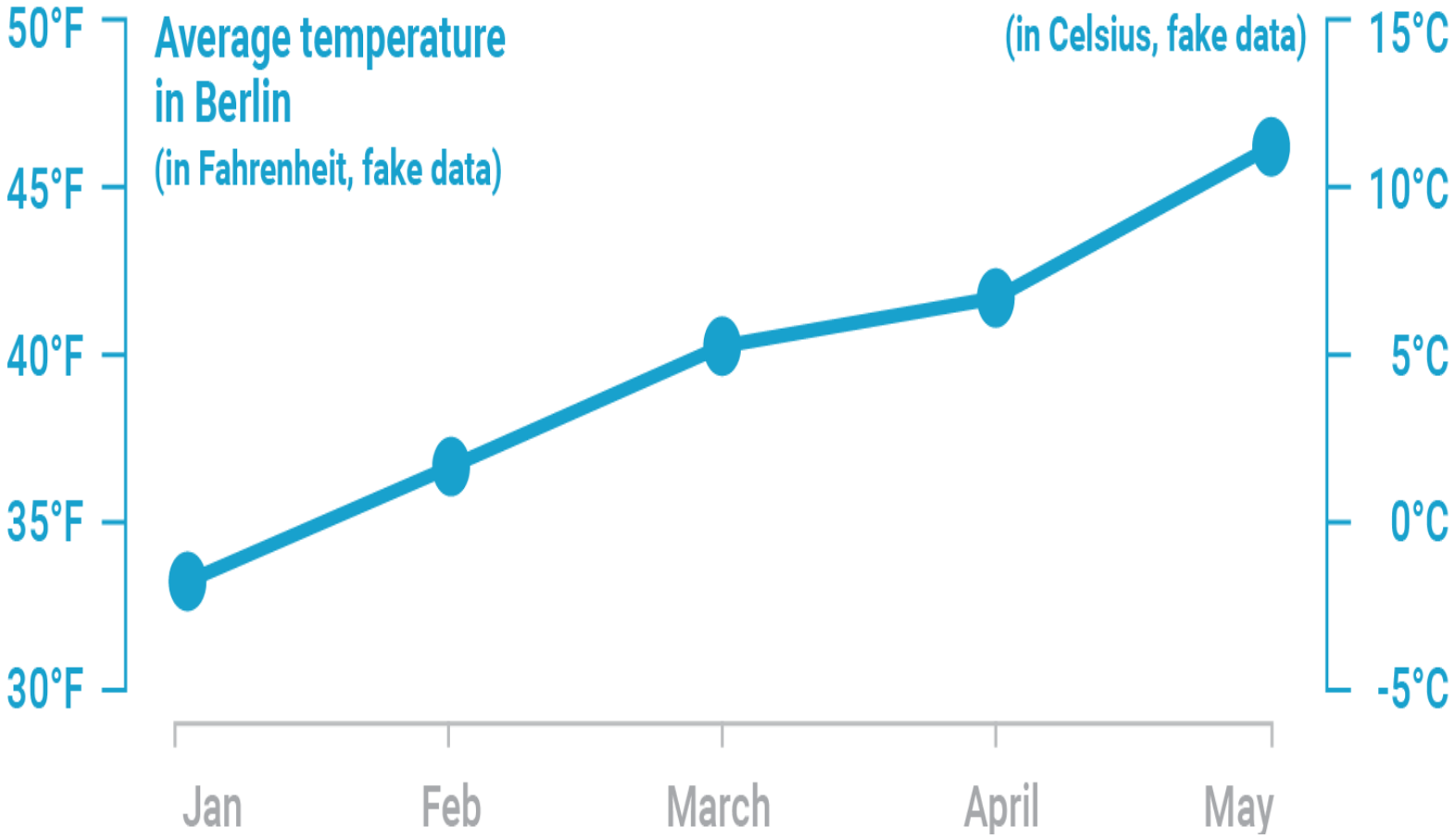
Pros:
Any trend can be easily seen in detail

Cons:
Trends can’t be directly compared

Evolution of baby names in the US

https://www.data-to-viz.com/caveat/spaghetti.html

Dual axis chart to show different scale

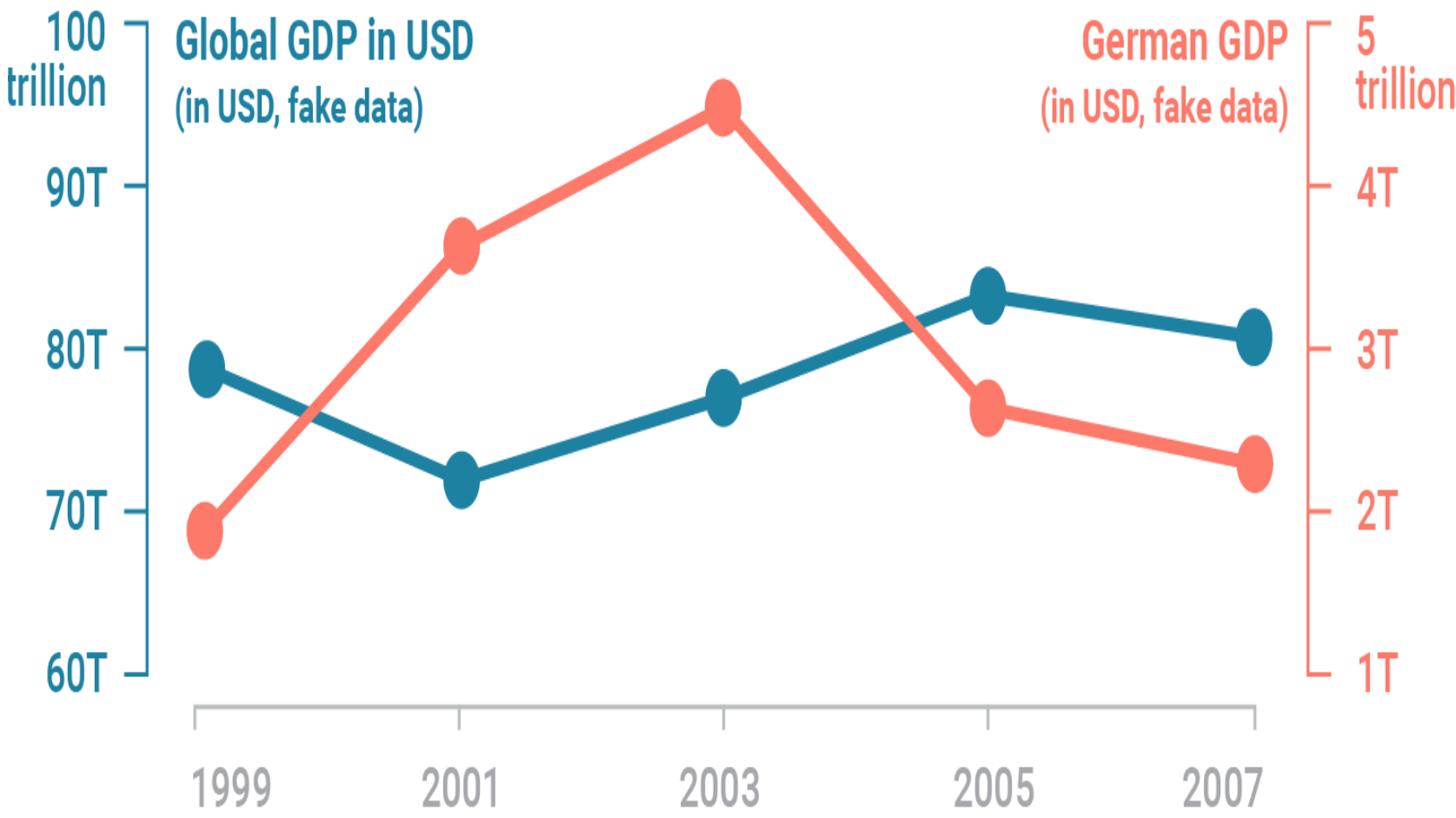


https://blog.datawrapper.de/dualaxis/

Dual axis chart can show association

Pros: helps to easily see

- No association
- Association
  - Positive
  - Negative

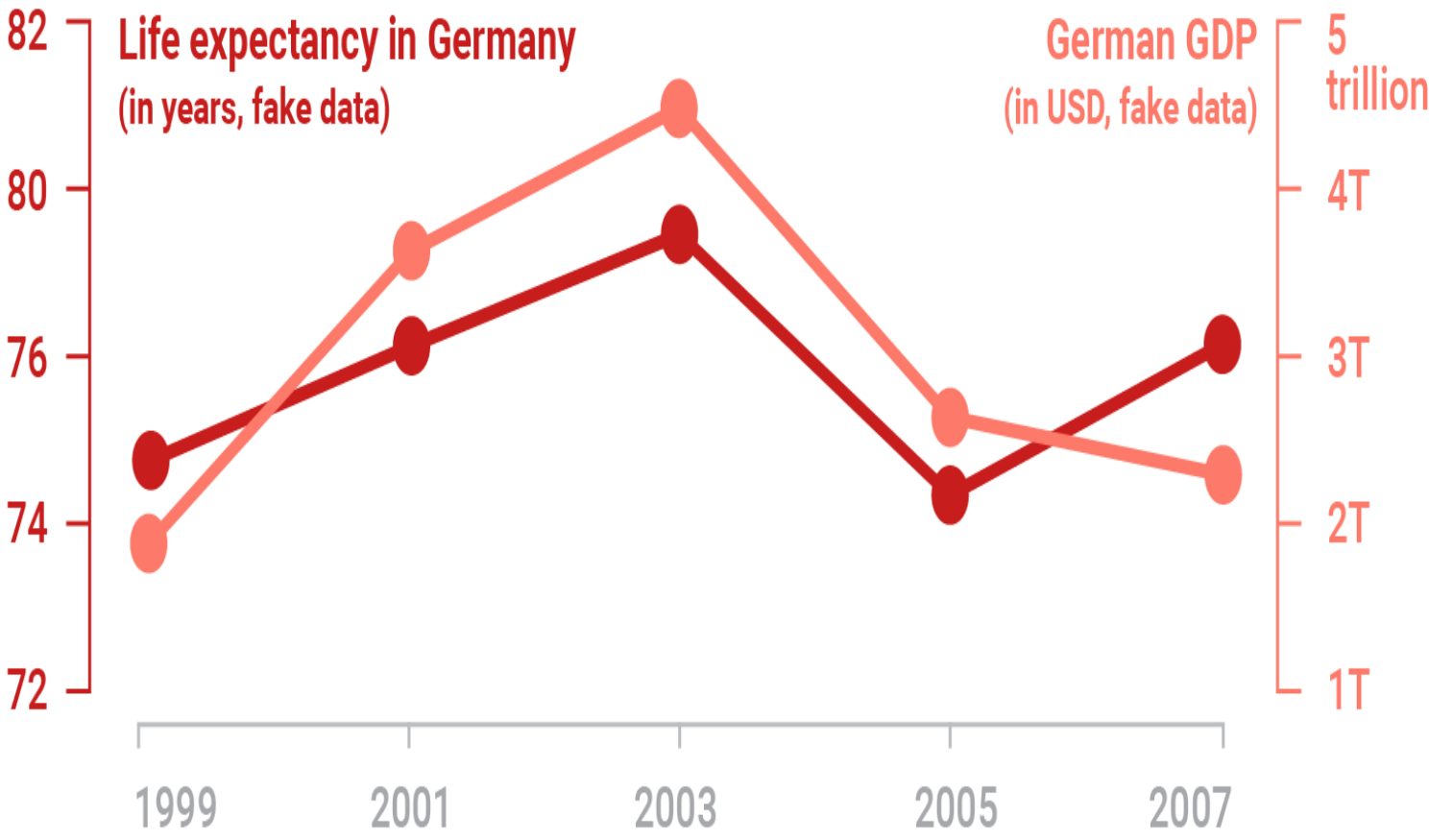


https://blog.datawrapper.de/dualaxis/

Dual axis chart can show association

Pros: helps to easily see

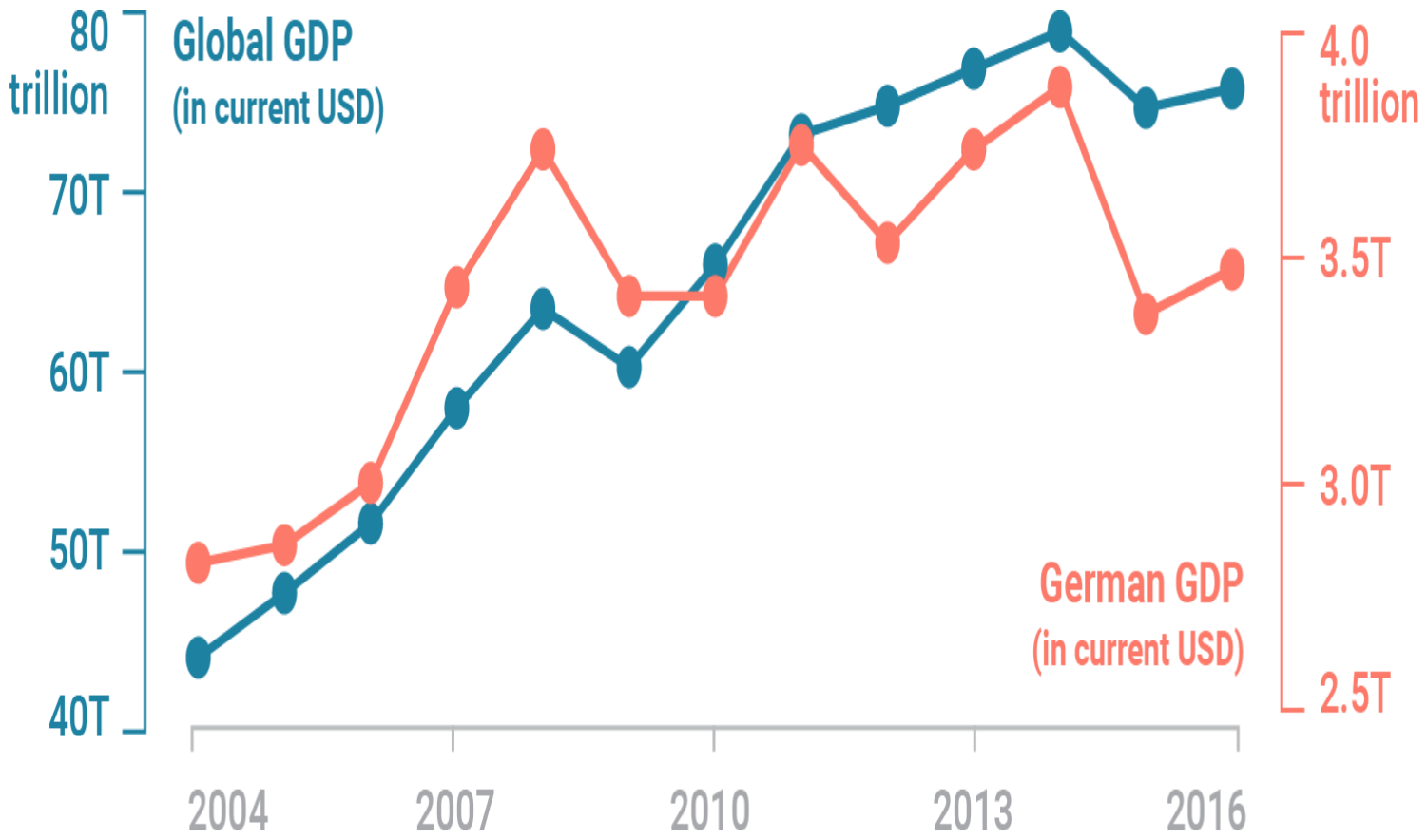
- No association
- Association
  - Positive
  - Negative



https://blog.datawrapper.de/dualaxis/

Dual axis chart can mislead

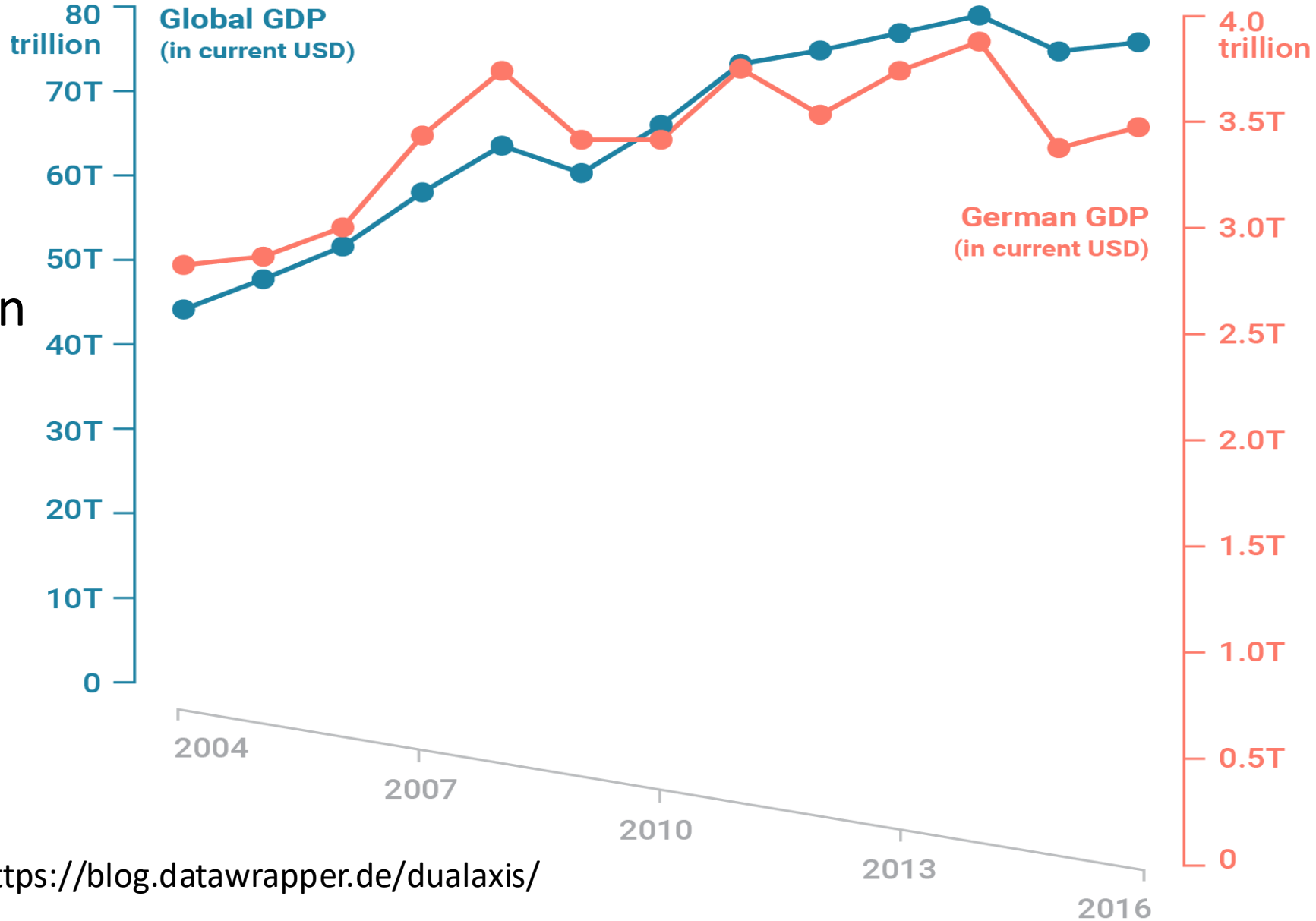
- The proportions of the two scales can be different



https://blog.datawrapper.de/dualaxis/

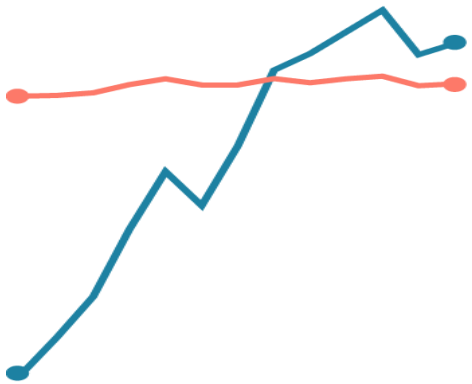
Dual axis chart can mislead

- The proportions of the two scales can be different
- If the axes extend down to zero, ..

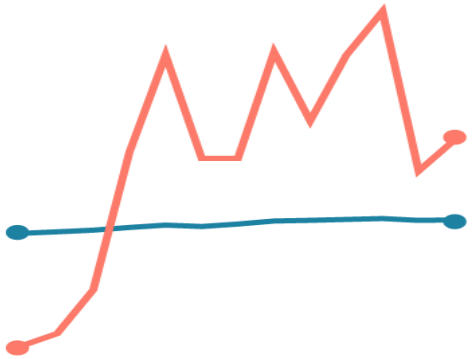


https://blog.datawrapper.de/dualaxis/

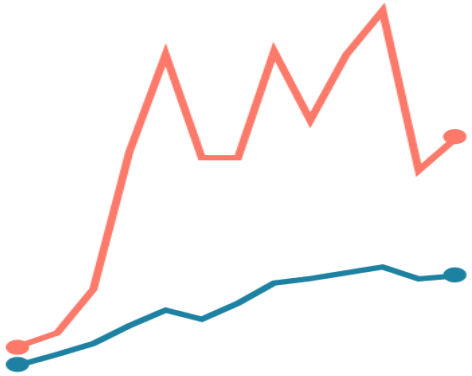
Different looks with different axes



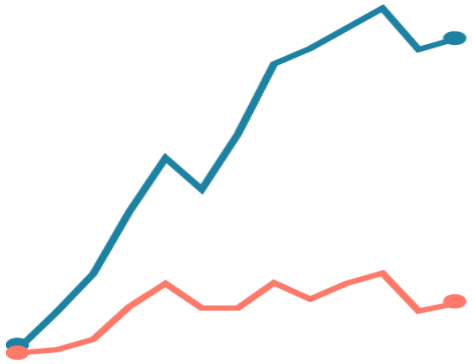
Orange steady,
Blue massively increasing.



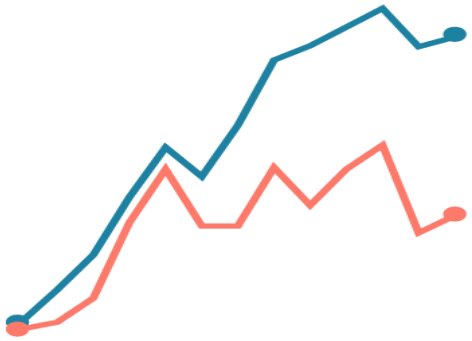
Blue steady,
Orange increasing.



Both started at the same level, but Orange increased far more than Blue.



Both started at the same level, but Blue increased far more than Orange.



Both started with the same increase, then Blue raced to the top.

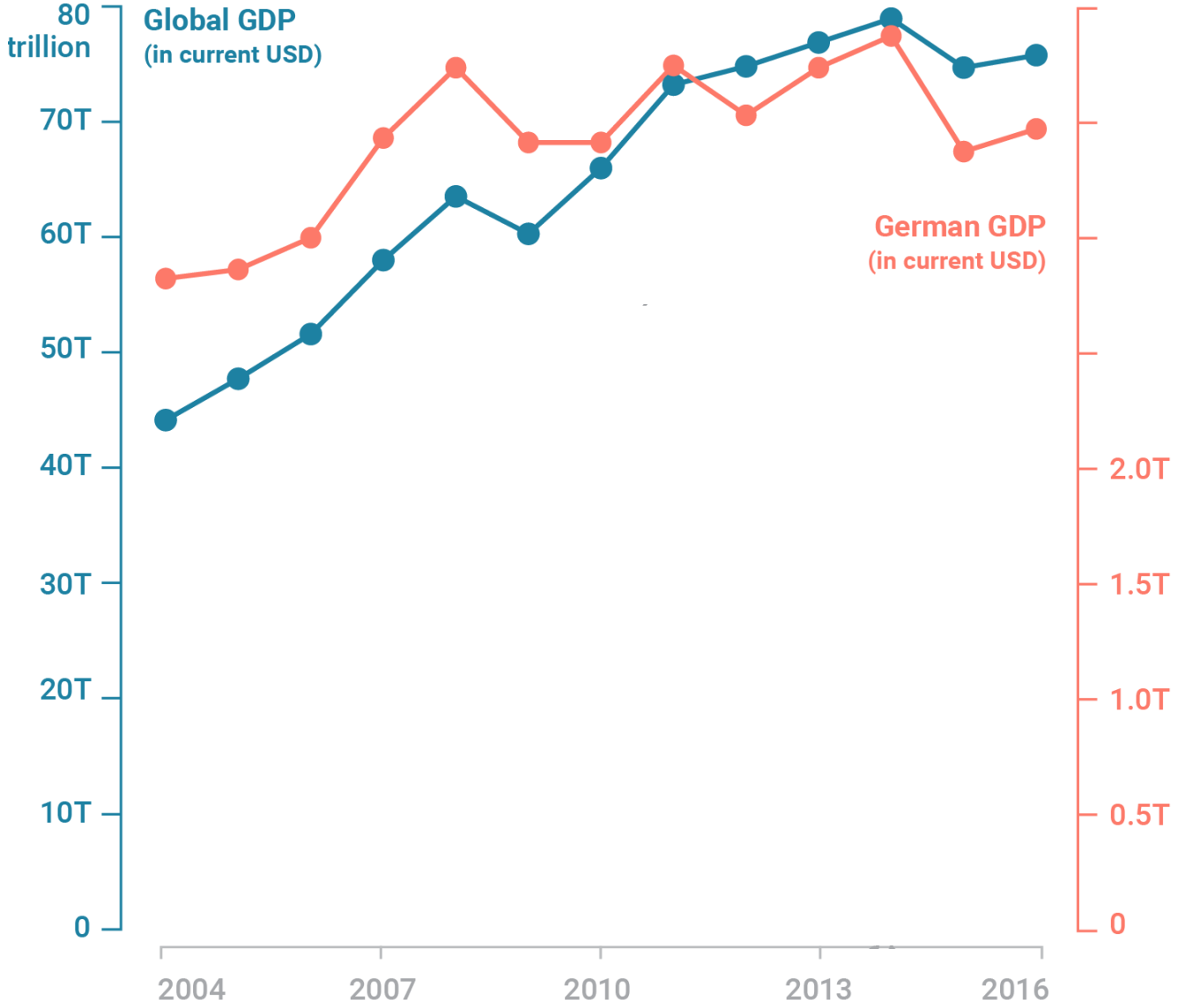
Both steady.

https://blog.datawrapper.de/dualaxis/

Zero baselines at the same height can mislead

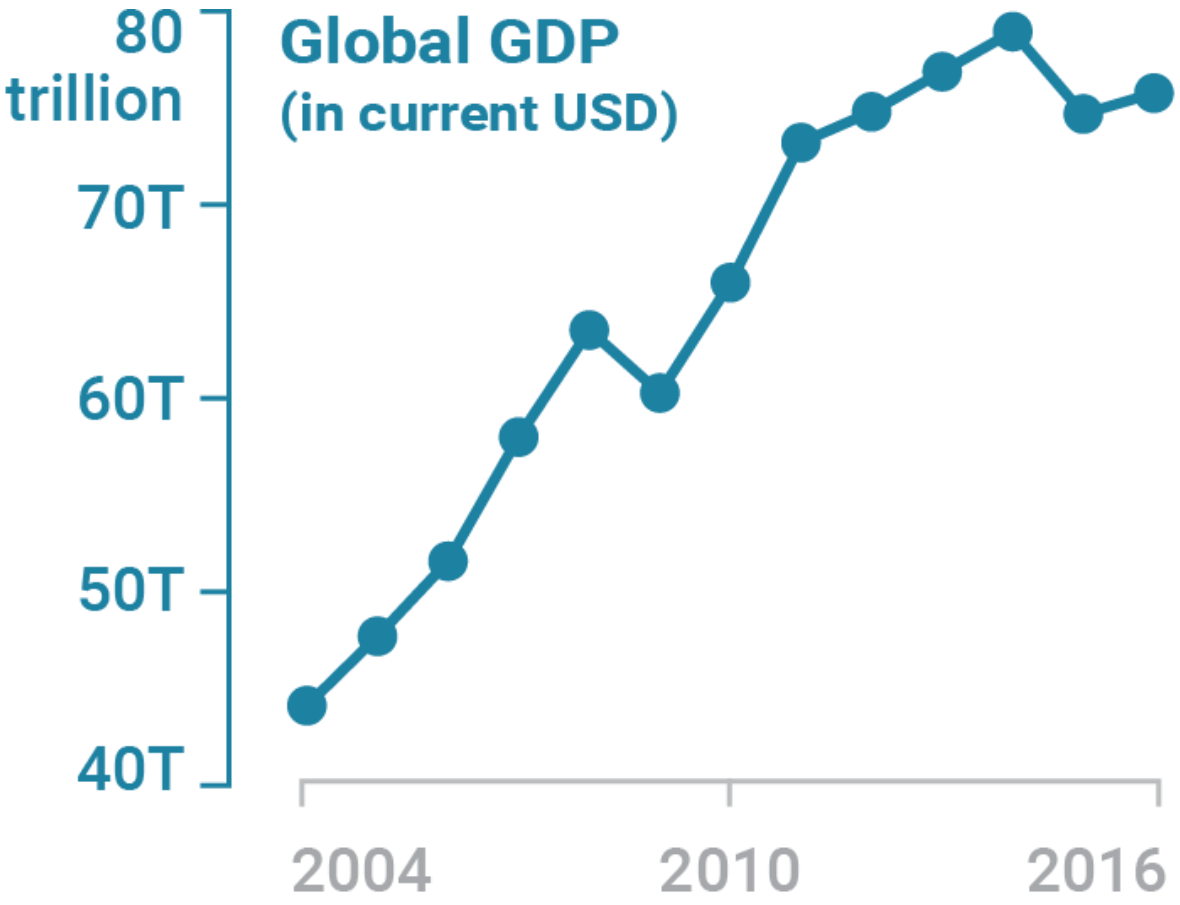
Potential misinterpretations

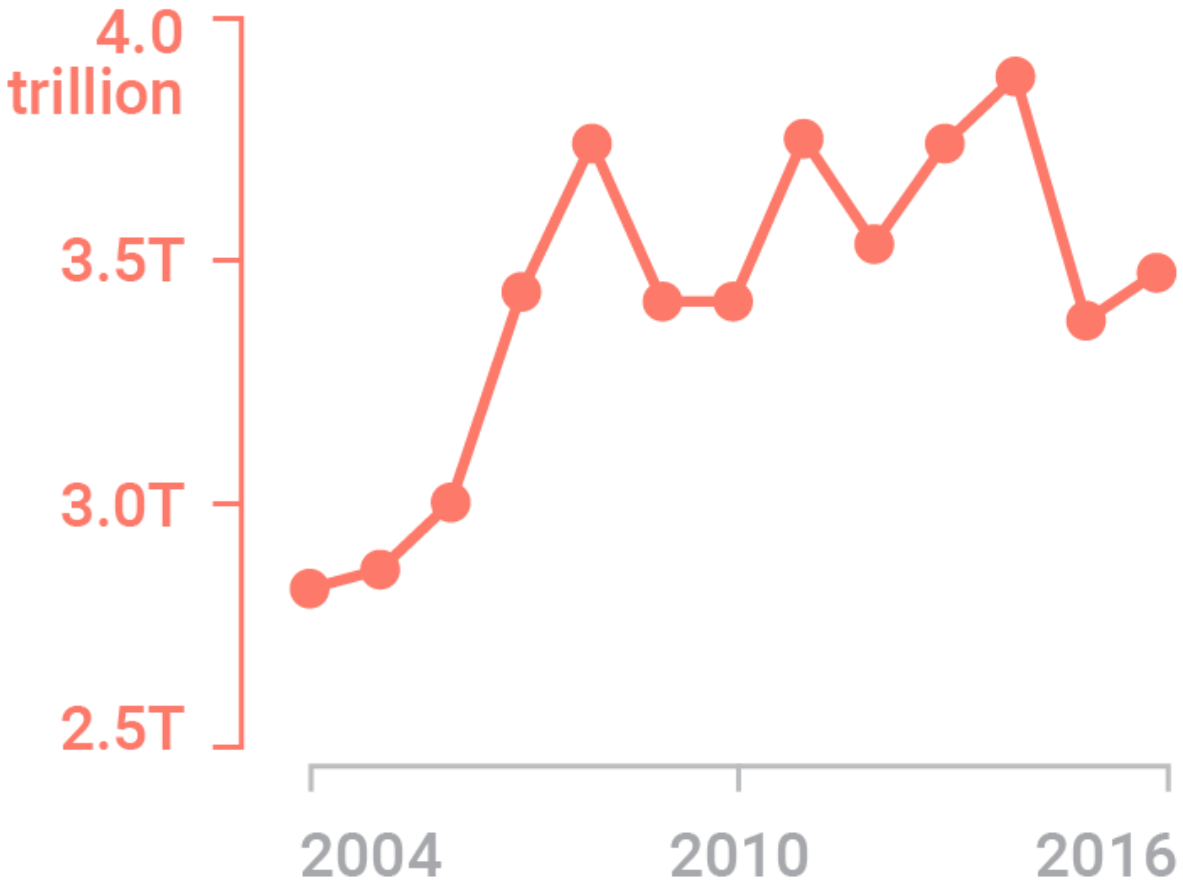
1. German GDP higher until 2012? (10% of chance)
2. Around the same at 2012? (10% of chance)
3. German GDP slightly lower after then? (10% of chance)
4. Waste of space



Adapted from https://blog.datawrapper.de/dualaxis/

Solution 1. Side-by-side charts



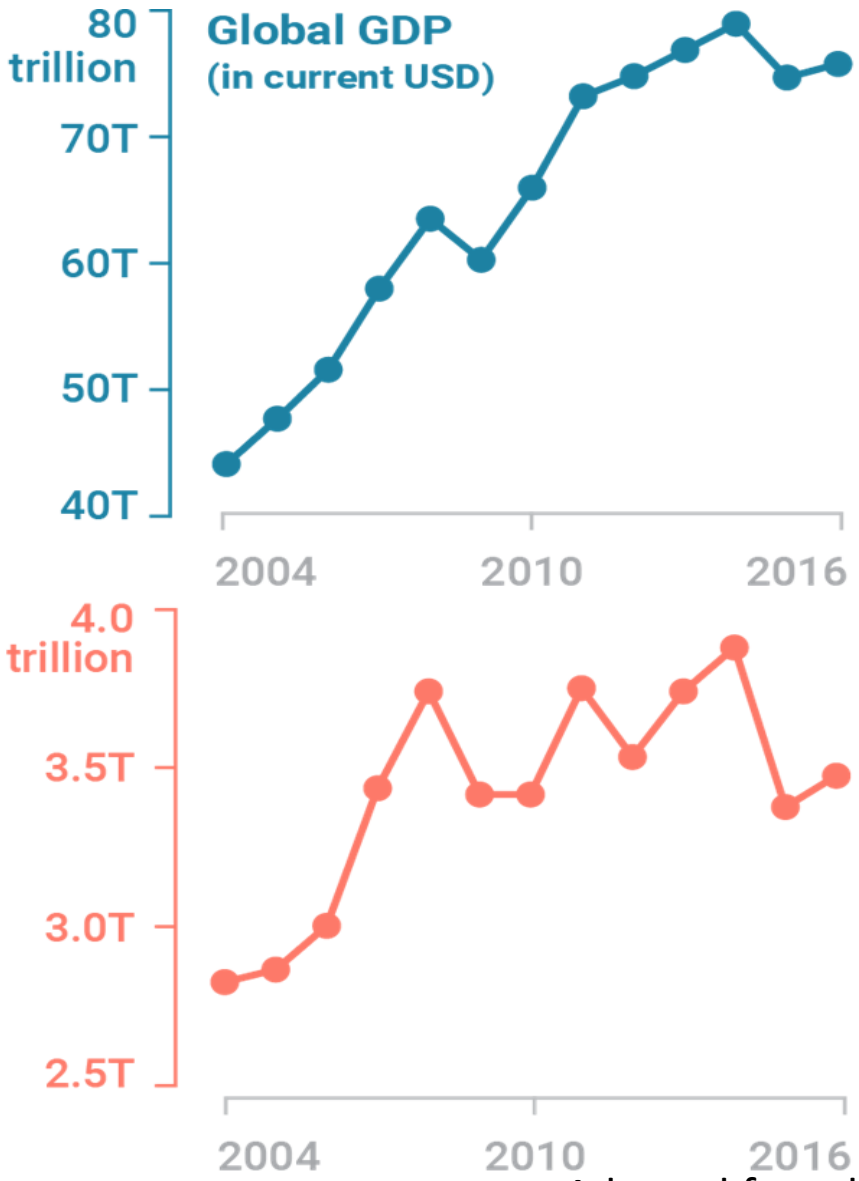


Imposing independence

Adapted from https://blog.datawrapper.de/dualaxis/

Solution 1. Side-by-side charts

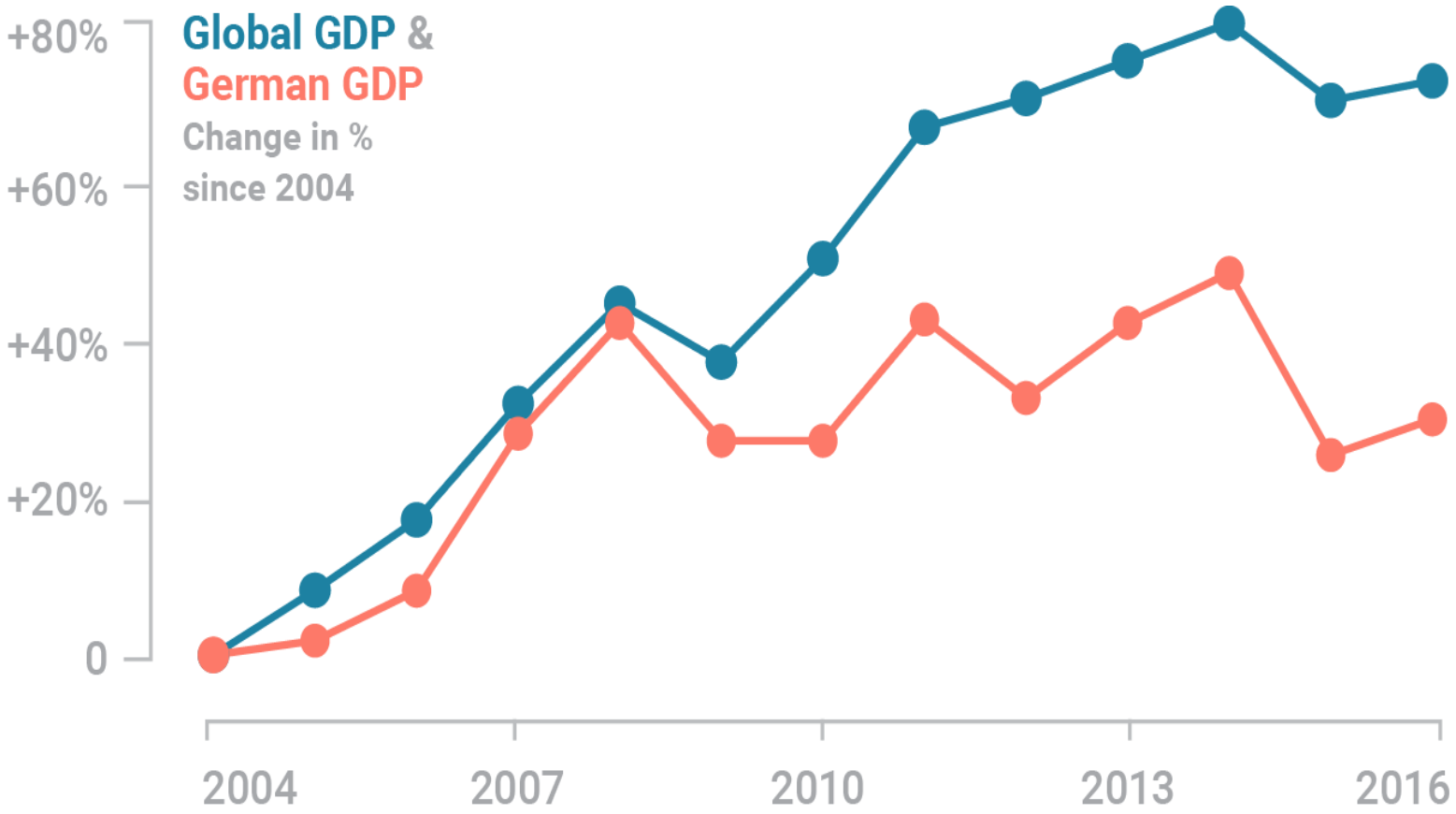
Pros: better if you want to compare them by year



Adapted from https://blog.datawrapper.de/dualaxis/

Solution 2. Indexed charts

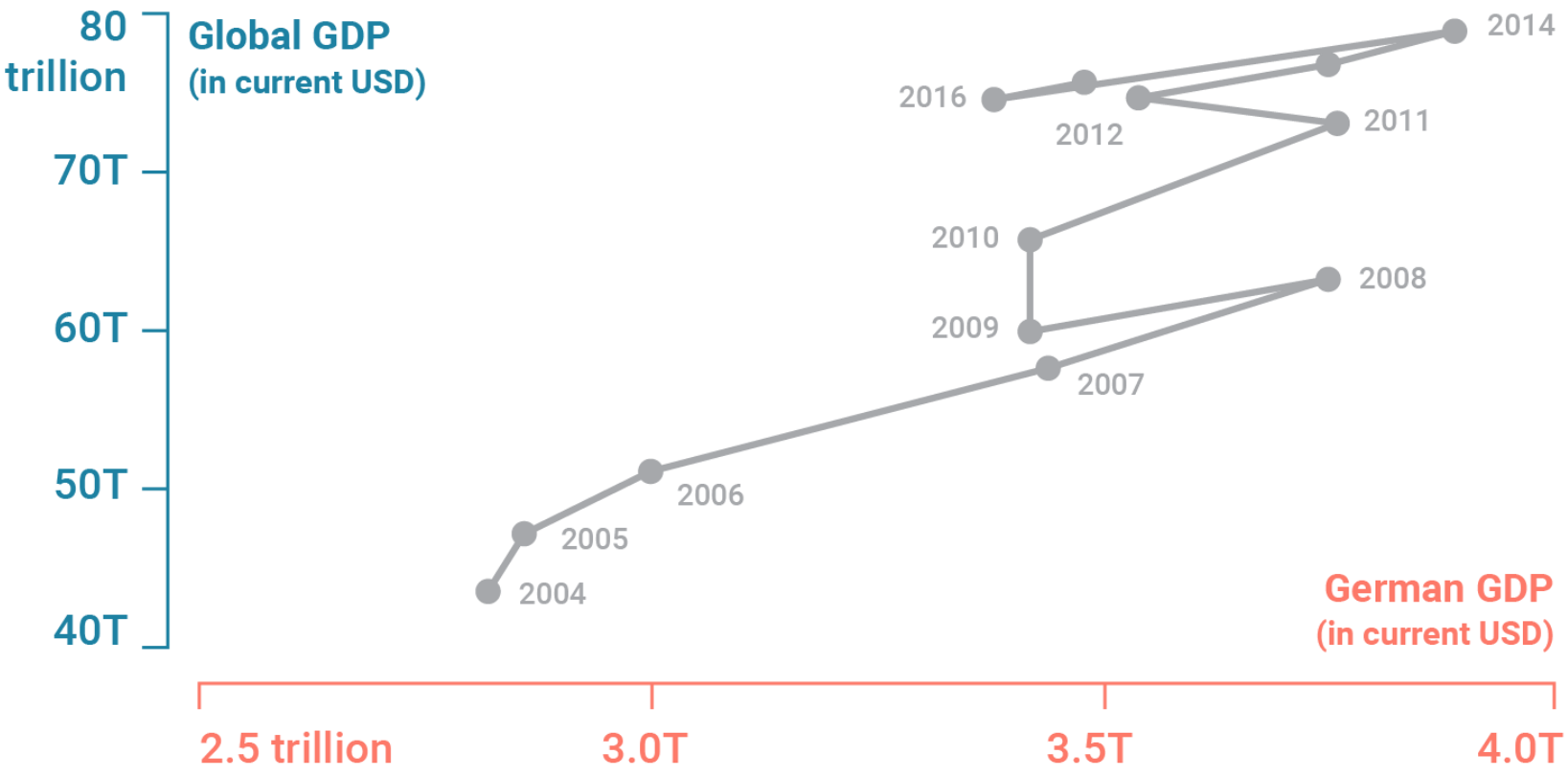
Pros: works for data series with a similar rate of change
Cons: more indirect than the raw GDP values



Adapted from https://blog.datawrapper.de/dualaxis/

Solution 3. Connected scatter plot

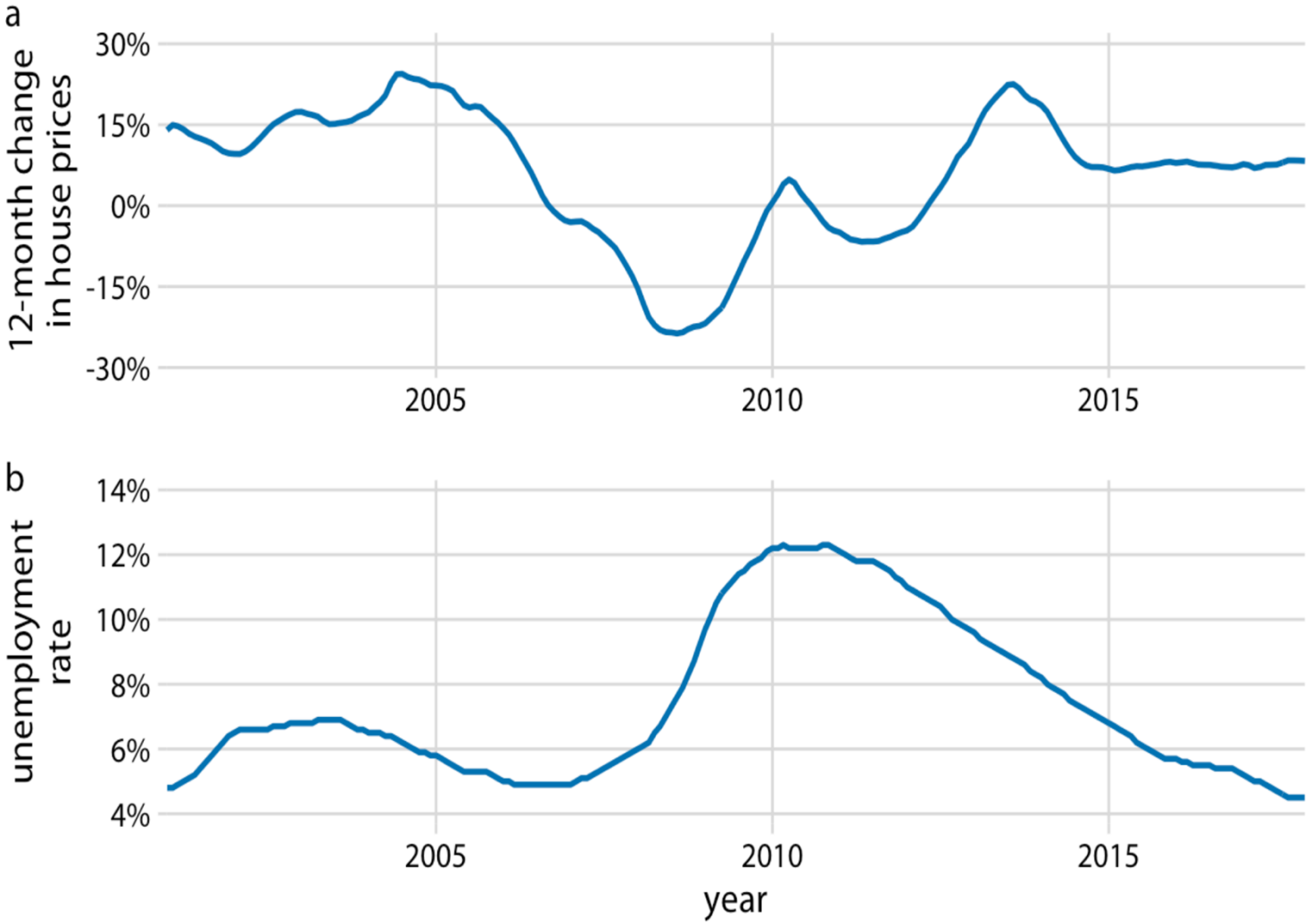
- Unintuitive, but fun
- Sometimes, overkill



Adapted from https://blog.datawrapper.de/dualaxis/

Line graph to show trend of ≥ 2 variables

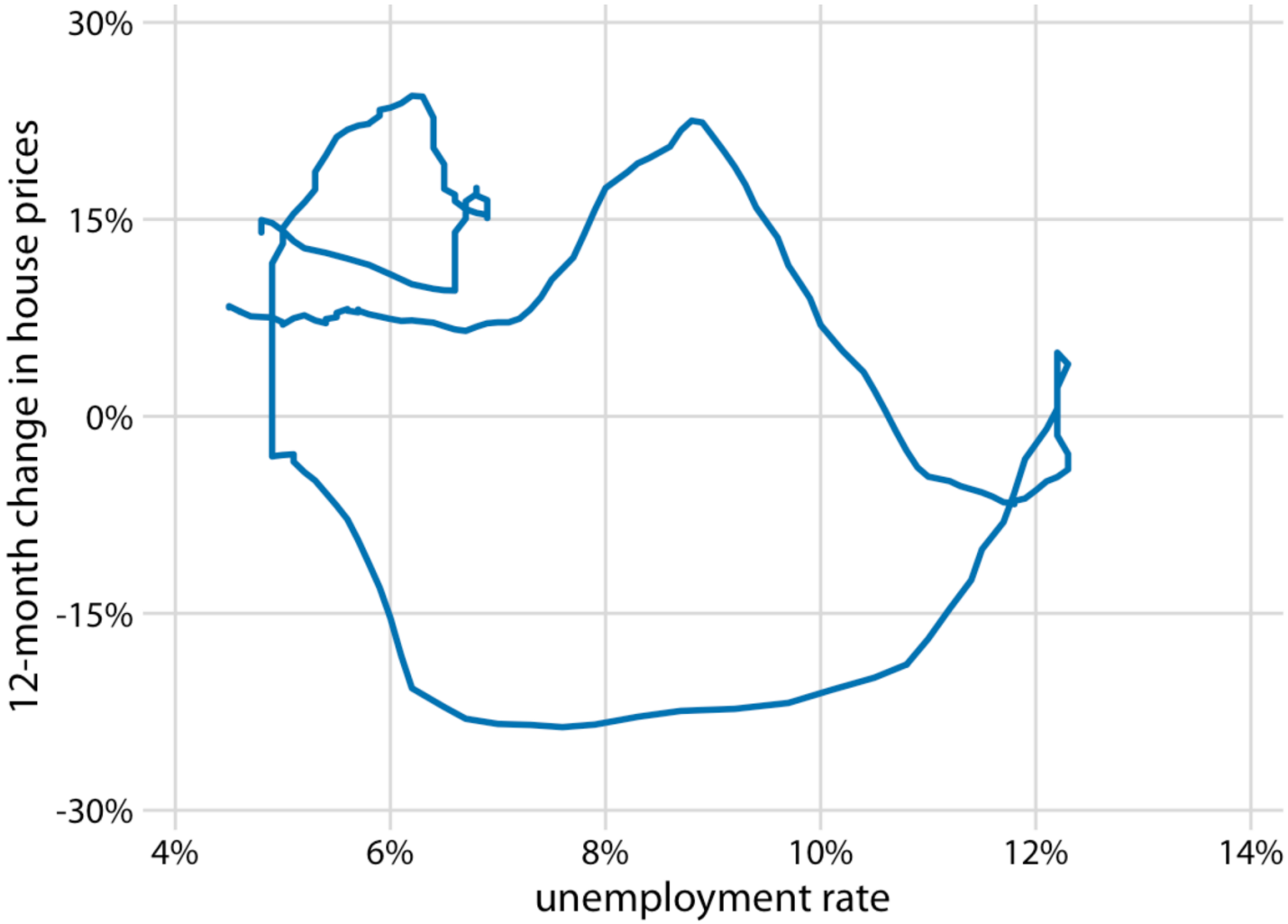
- E.g., consider house prices in relation to unemployment rate
- Interested in temporal regions when both variables move in the same or in opposite directions.
- *Could* use a dual-axis chart to show both



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Connected scatter plot to show trend of ≥ 2 variables

- Like scatterplots: plotting two variables against each other
- Unlike scatterplots: connecting neighboring points from the earliest time point to the latest



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Connected scatter plot to show trend of ≥ 2 variables

- Important to indicate both the direction and the temporal scale of the data
- E.g., encode time with color gradation or arrows

Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

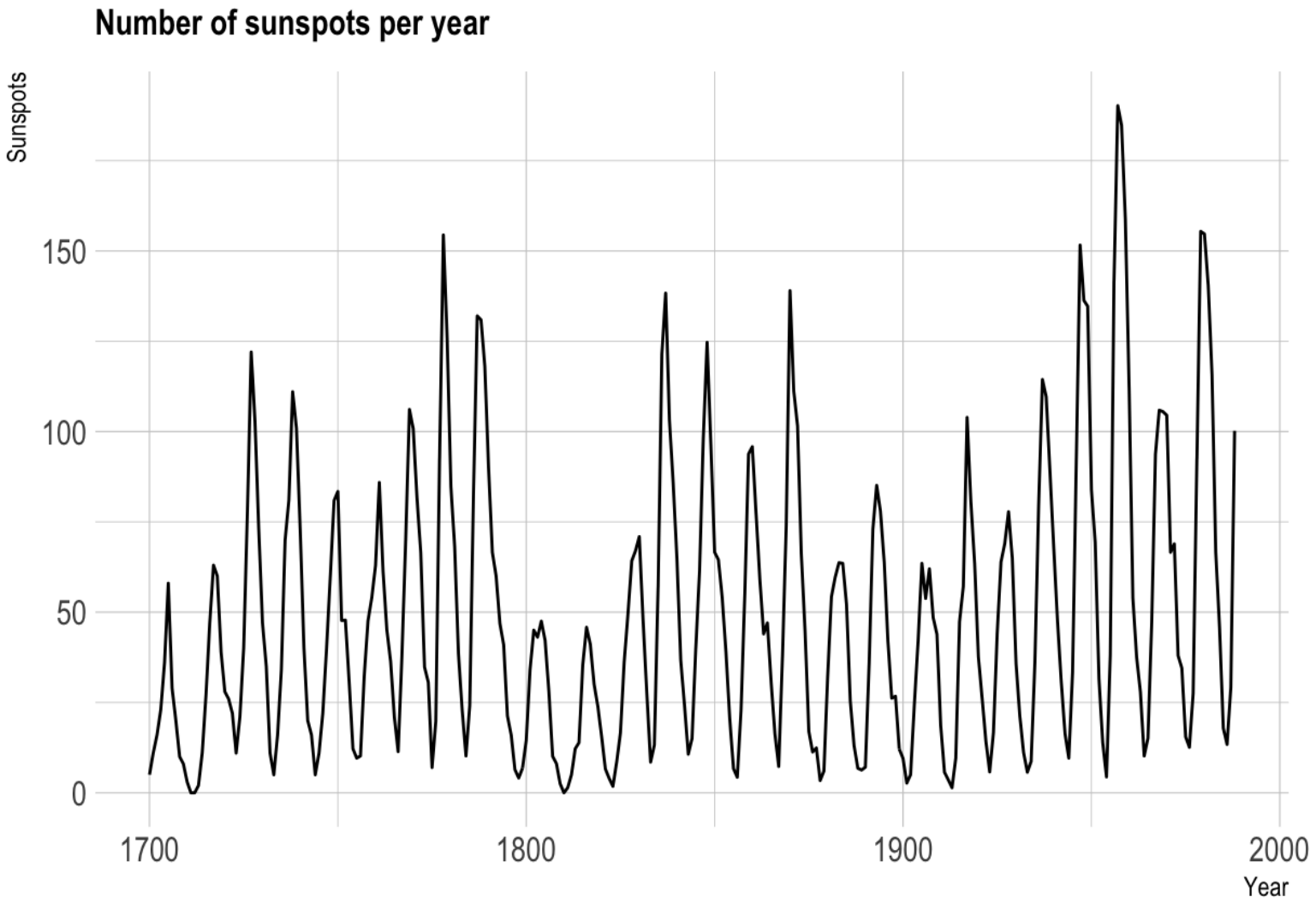
Connected scatter plot to show trend of ≥ 2 response variables

- Positive/negative correlation -> lines lying in the "northeast" or "southeast" directions
- Cyclic relationship → loops
- Transition of the center of the circles?

Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Different aspect ratios show different trends

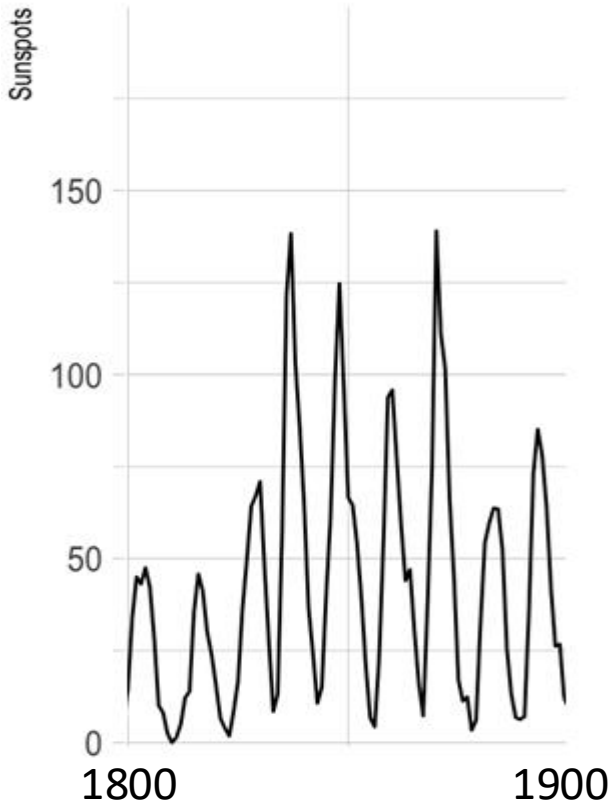
- The sunspot numbers follow a periodic 11-year cycle
- Some spikes are higher than others

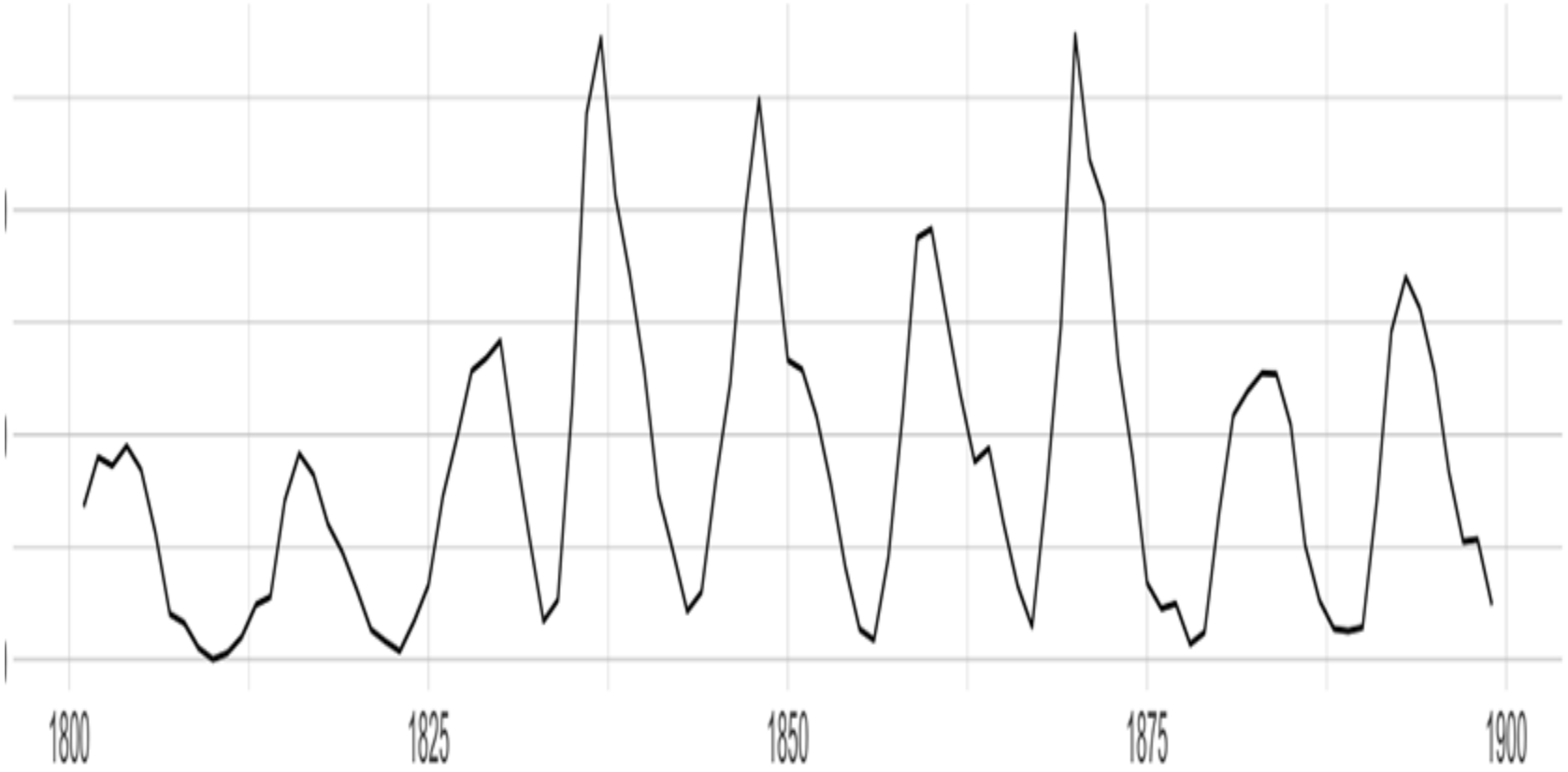


https://www.data-to-viz.com/caveat/aspect_ratio.html

Different aspect ratios show different trends

- Aspect ratio: the ratio of the width to the height of a graphic
- the sunspot # increases quicker than it decreases in a wider aspect ratio

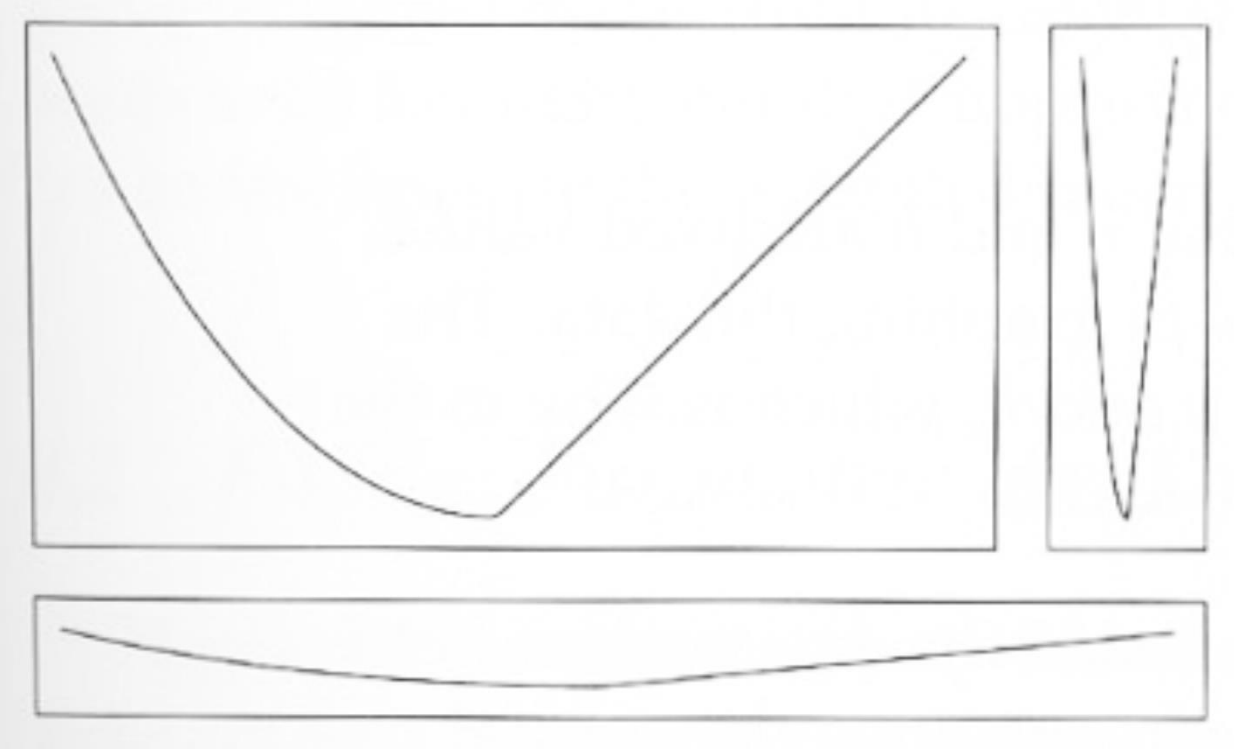




https://www.data-to-viz.com/caveat/aspect_ratio.html

Different aspect ratios show different trends

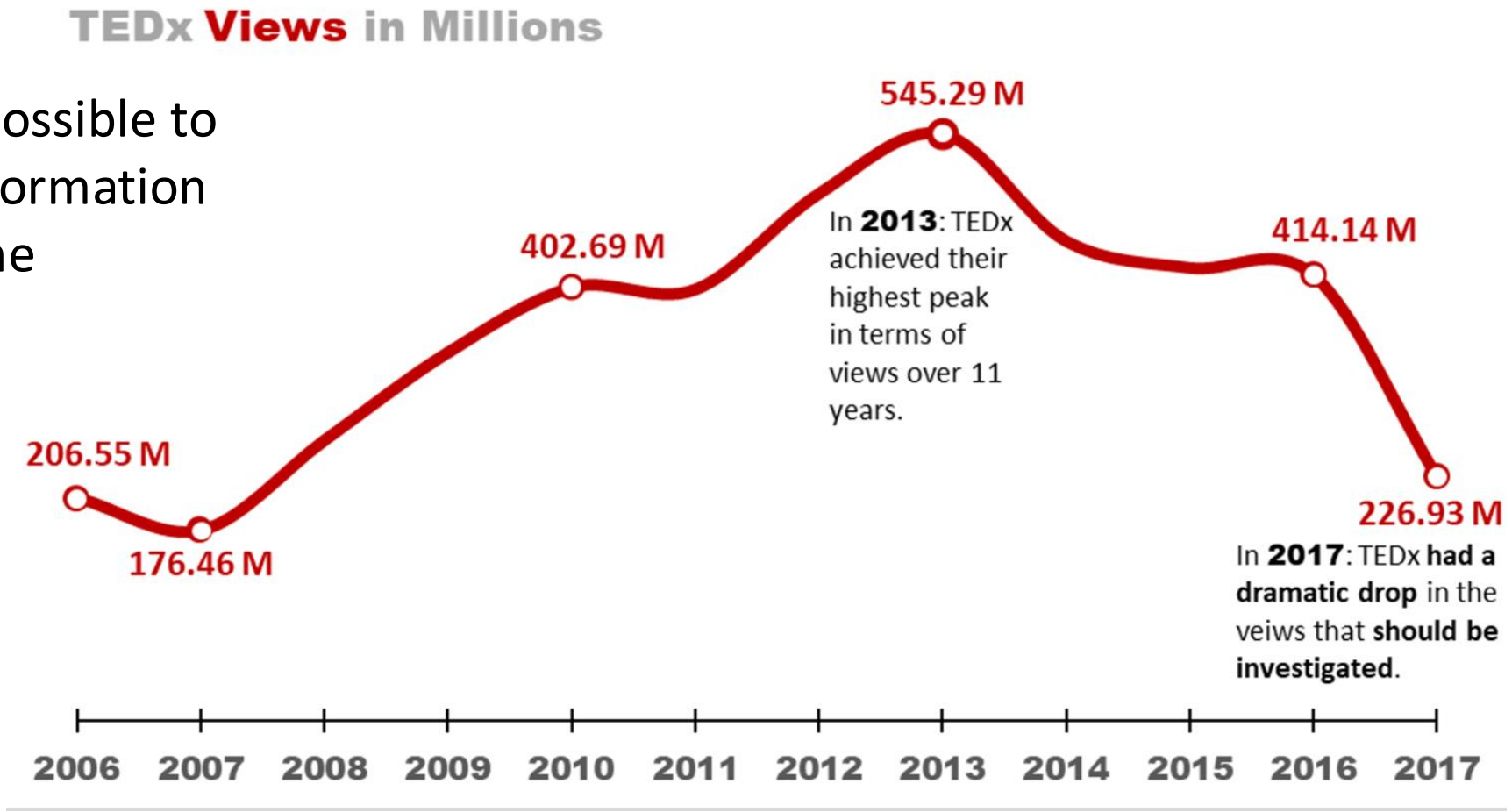
- The signal should be placed in the 2/3rd portion of the y-axis
- The pattern of your interest should target a 45° angle (?)



W.S. Cleveland. 1993. Visualizing Data.

Annotated line graph

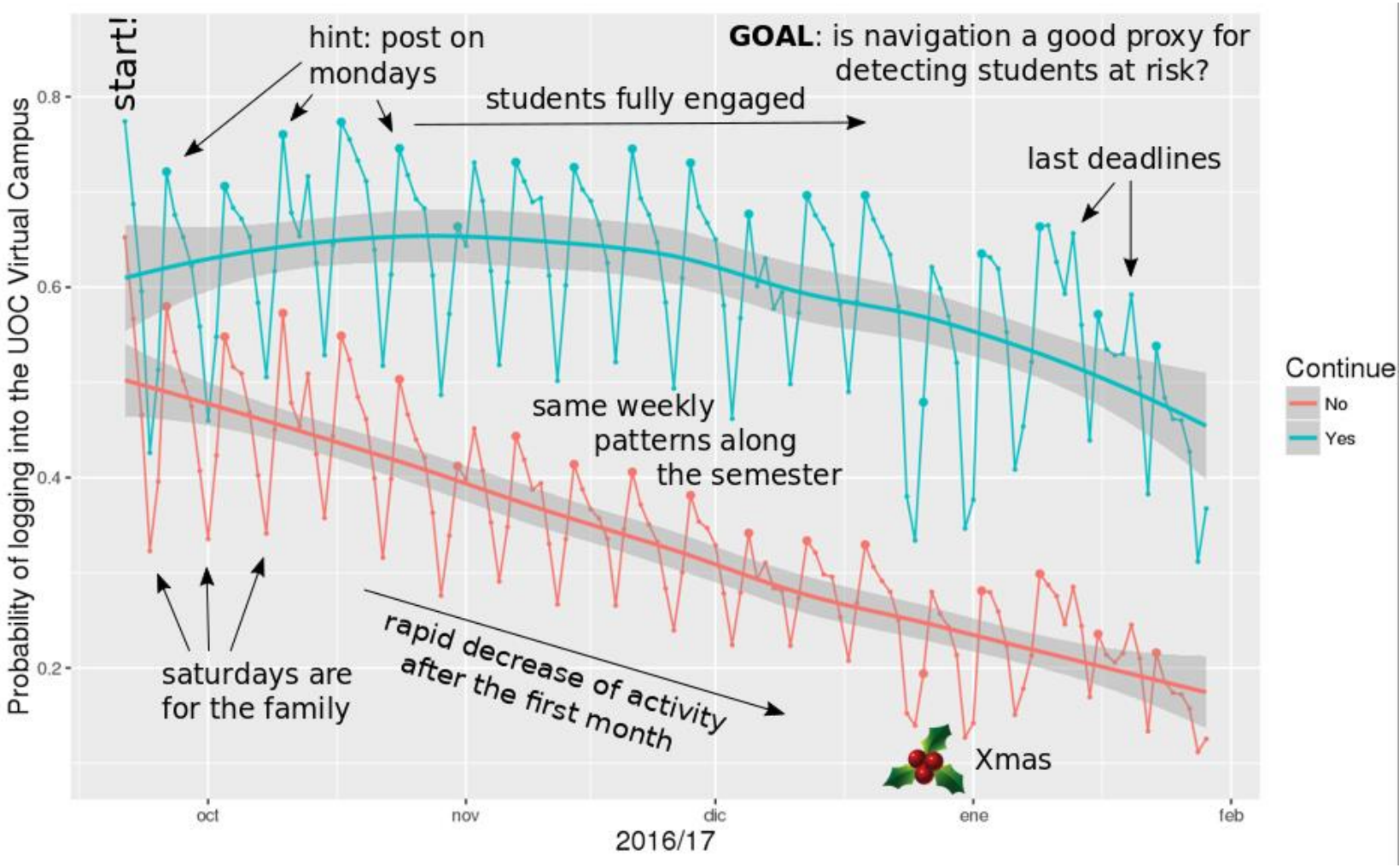
- Due to the simplicity, possible to annotate additional information
- Critical to determine the appropriate amount of information.



http://www.storytellingwithdata.com/blog/2018/1/22/88-annotated-line-graphs

Line graph with too much annotation

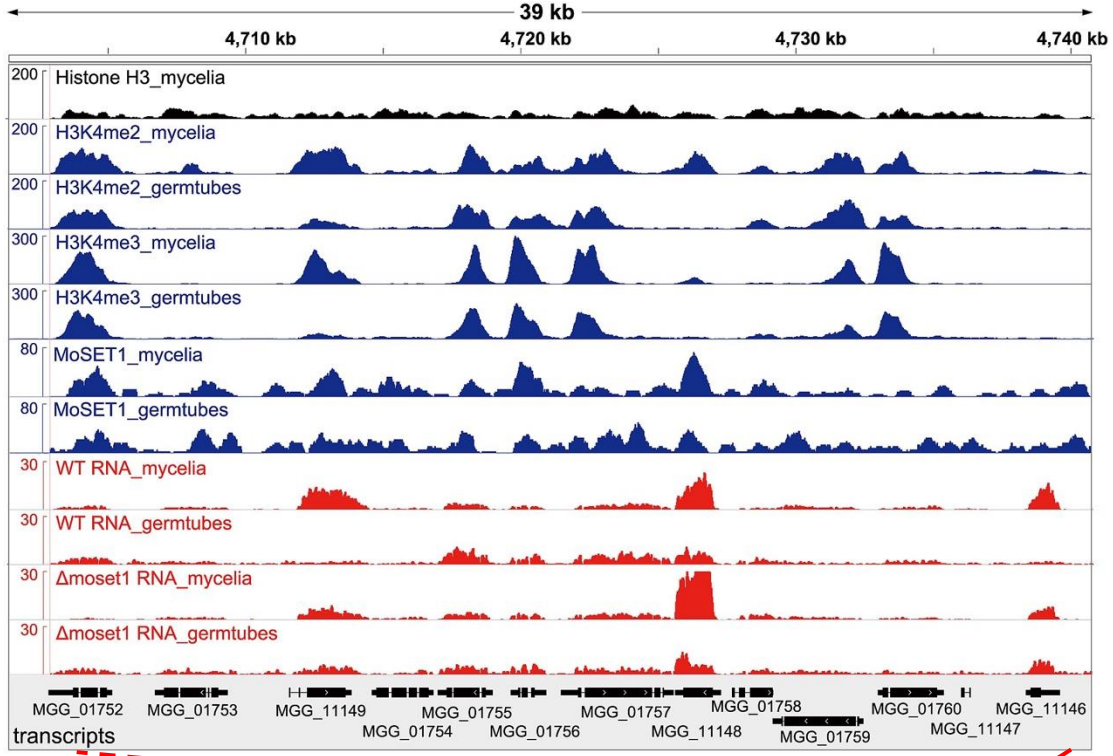
- Each additional label might not seem like “too much”
- Eventually you can end up with too much information
- → Critical to determine the appropriate amount of info.



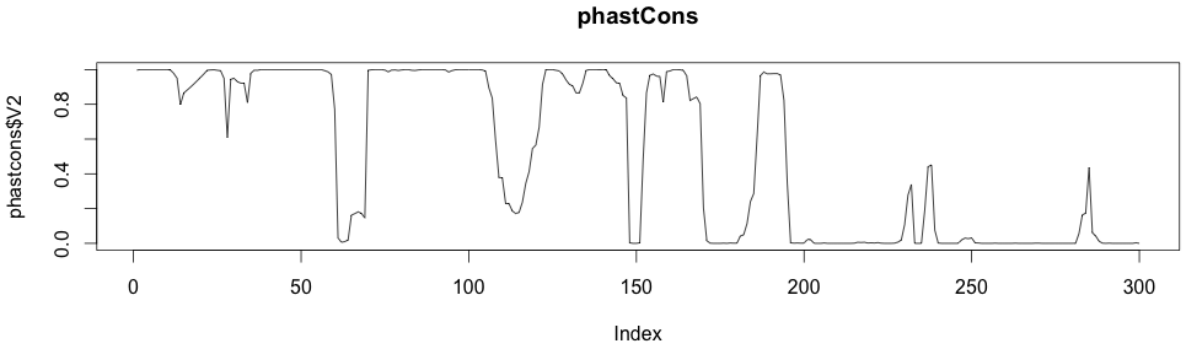
http://www.storytellingwithdata.com/blog/2018/1/22/88-annotated-line-graphs

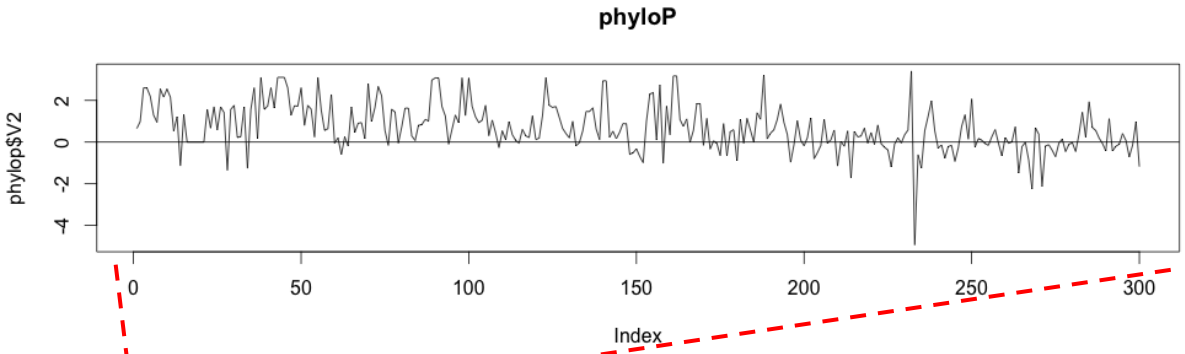
Complex biological signals in line plots

Genetic/epigenetic signal from sequencing data



Conservation signal on genomics data



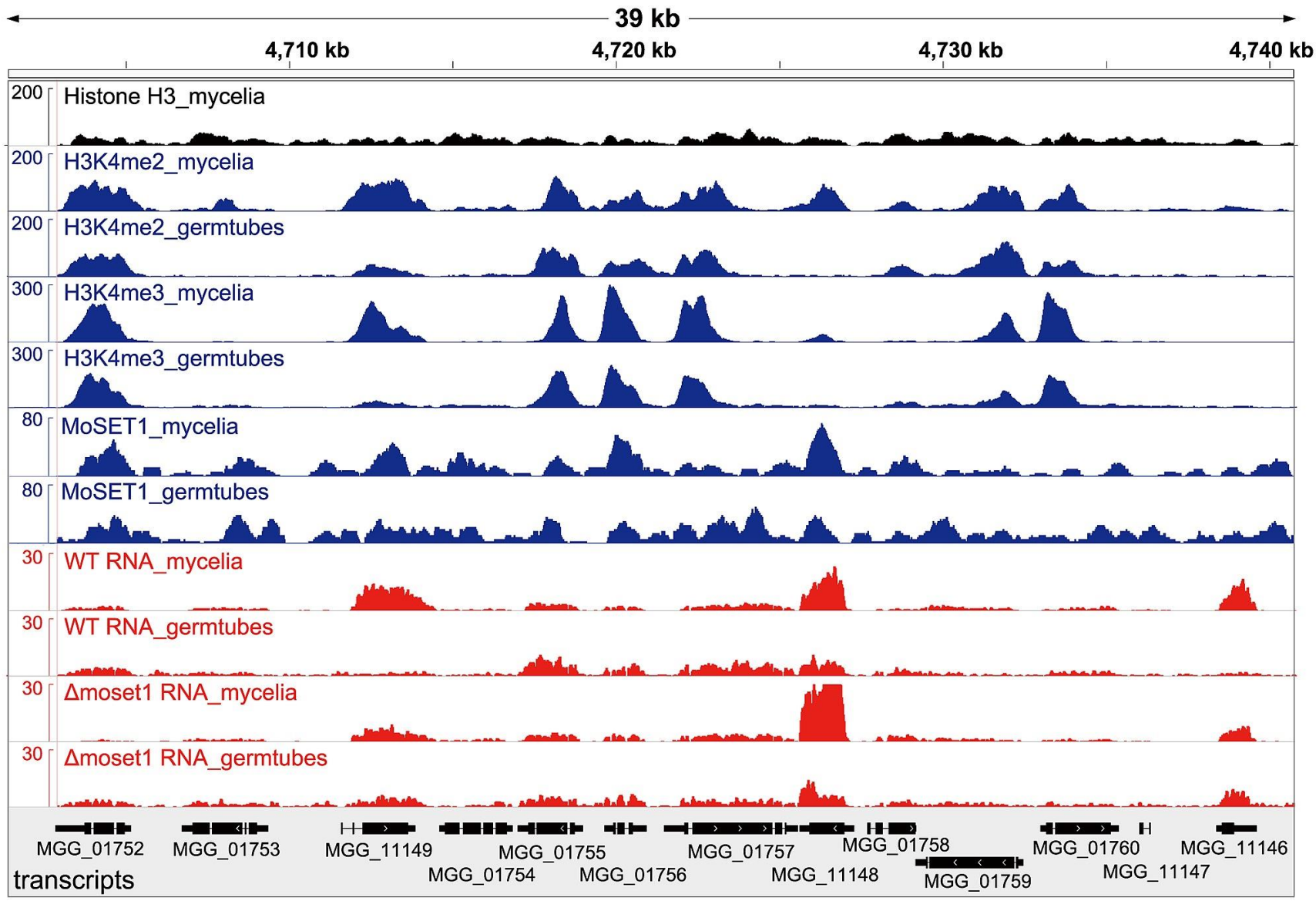


https://davetang.org/muse/2014/08/28/finding-sequence-conservation/

- Dynamic changes across the whole genome (e. g., 6.4 billion BP for human)

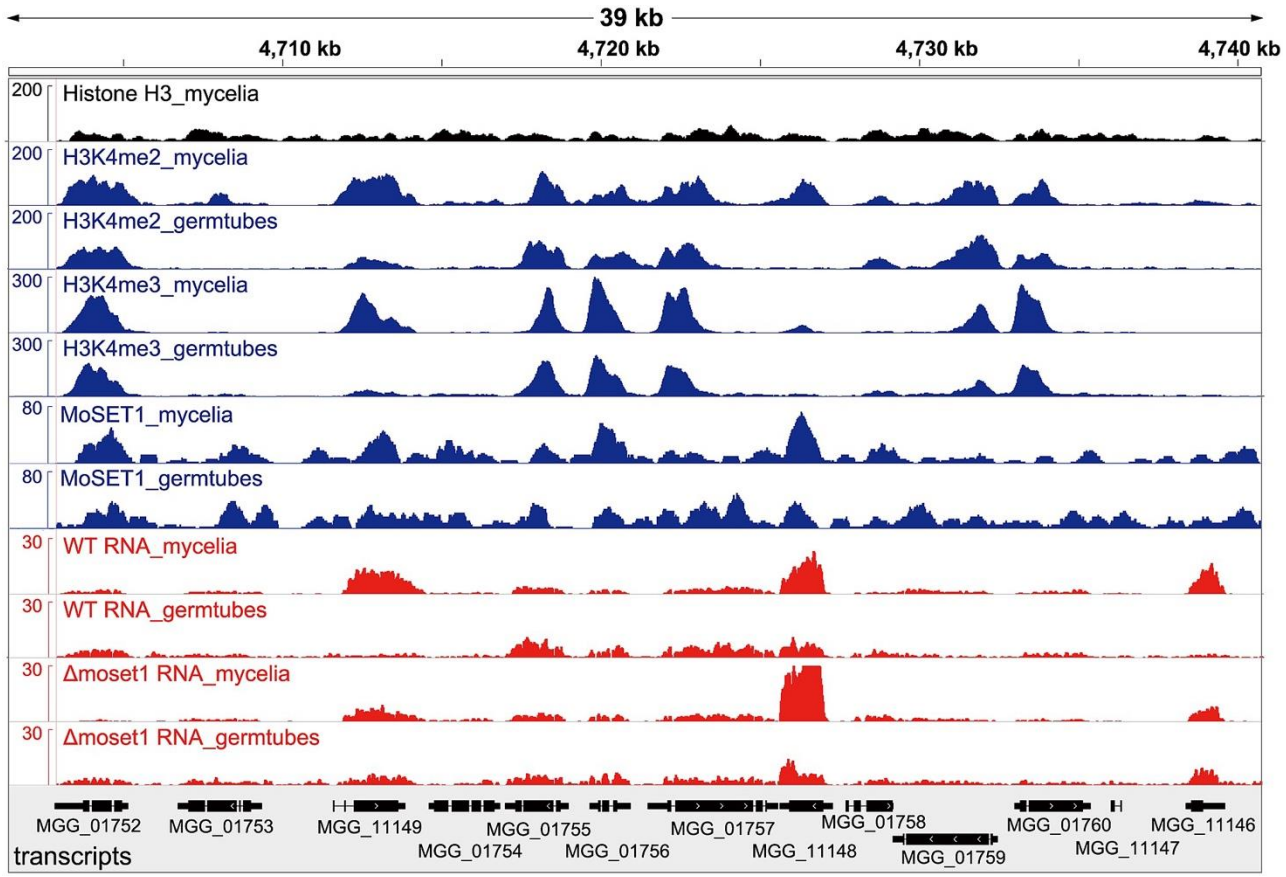
Shaded line plots for each case

- To discuss in detail when covering Genome Browser

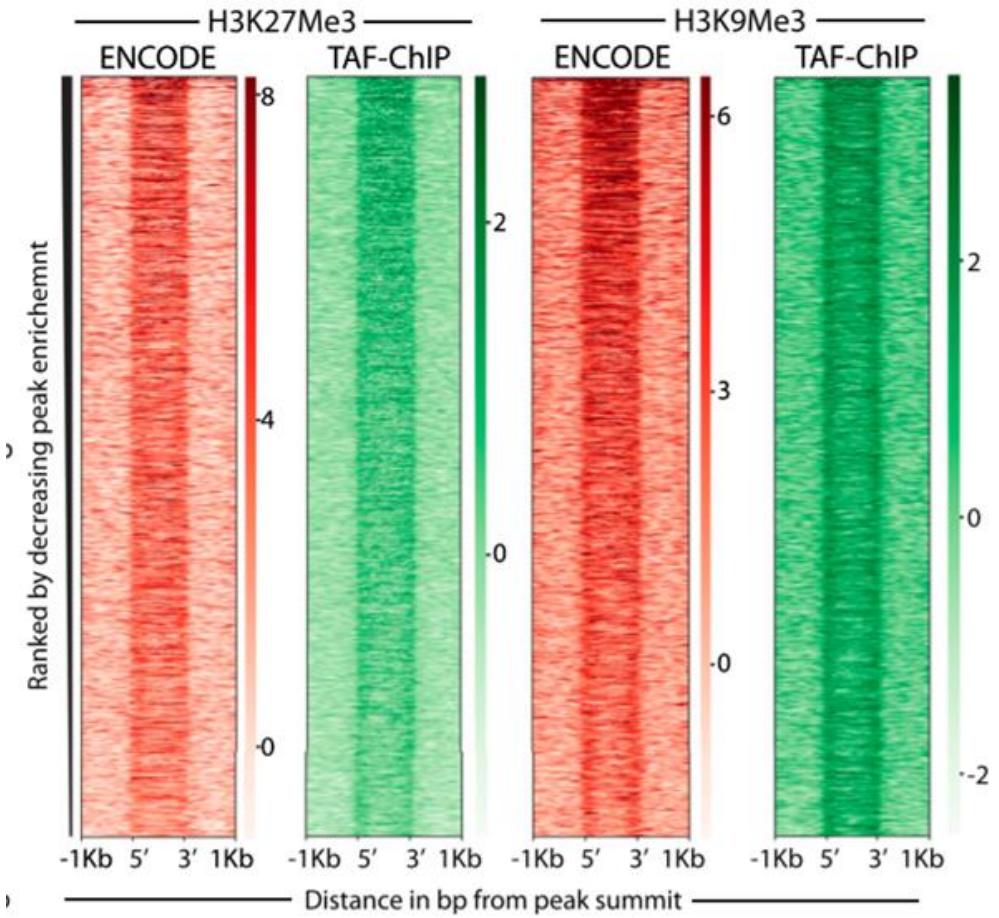


Pham et al. PLOS Genetics. 2015

Heatmaps to show many signals

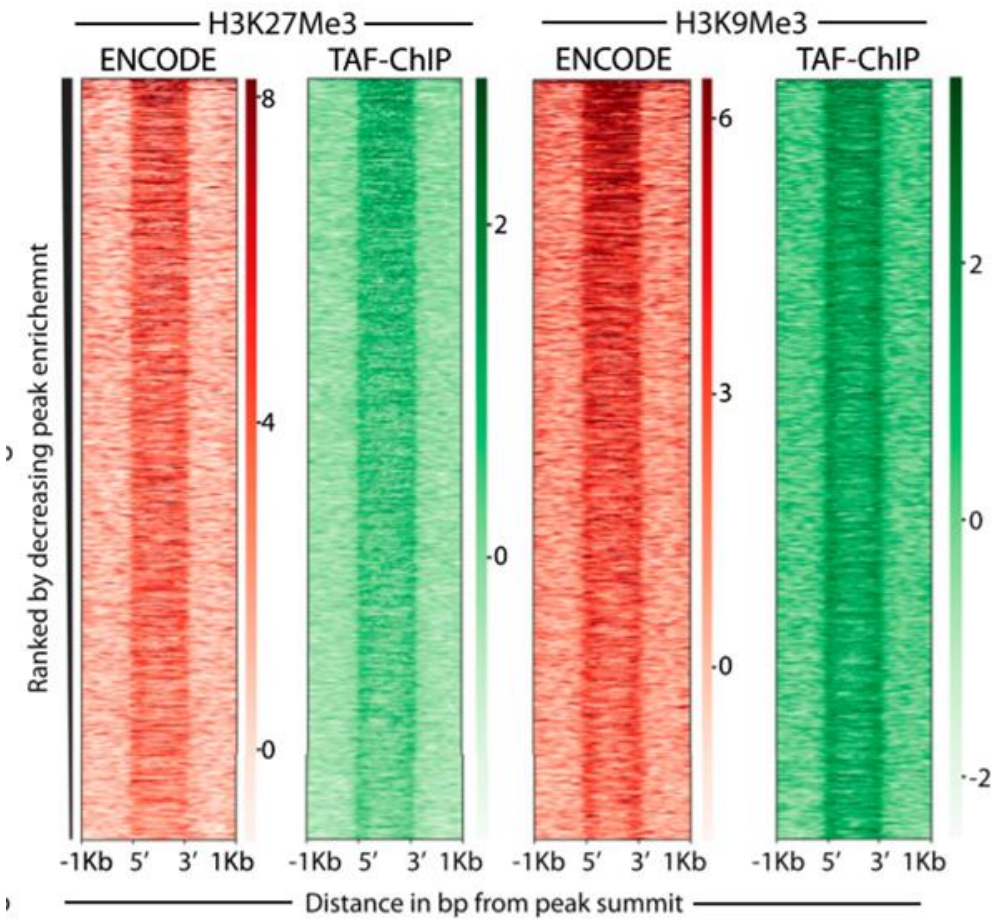


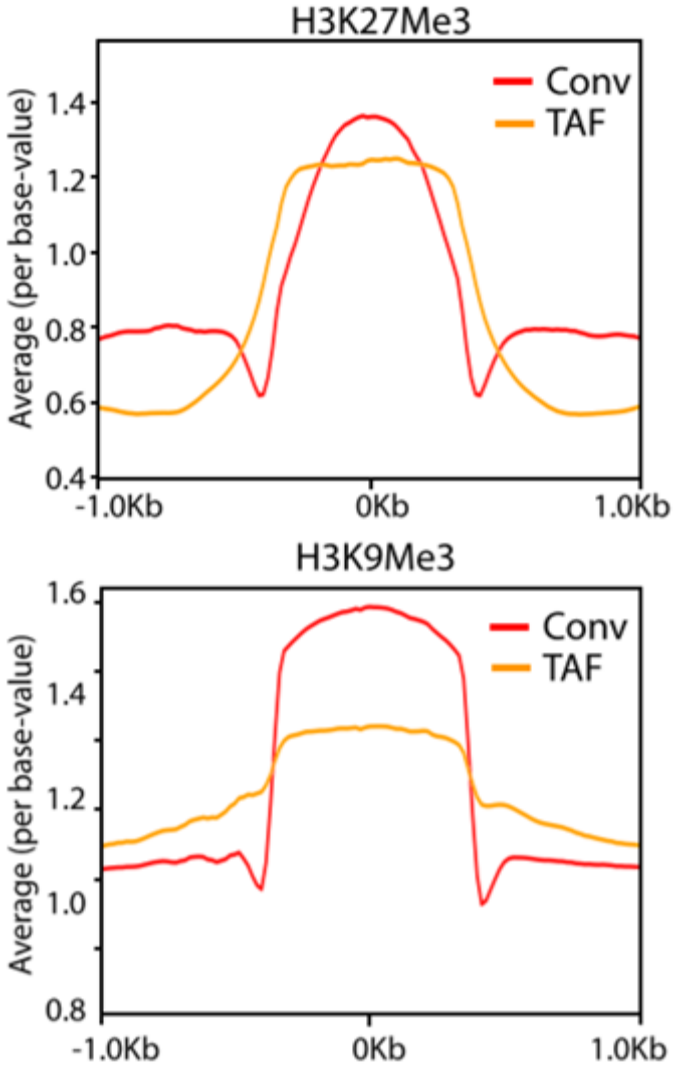
Multiple regions



Akhtar et al. Life Science Alliance, 2019

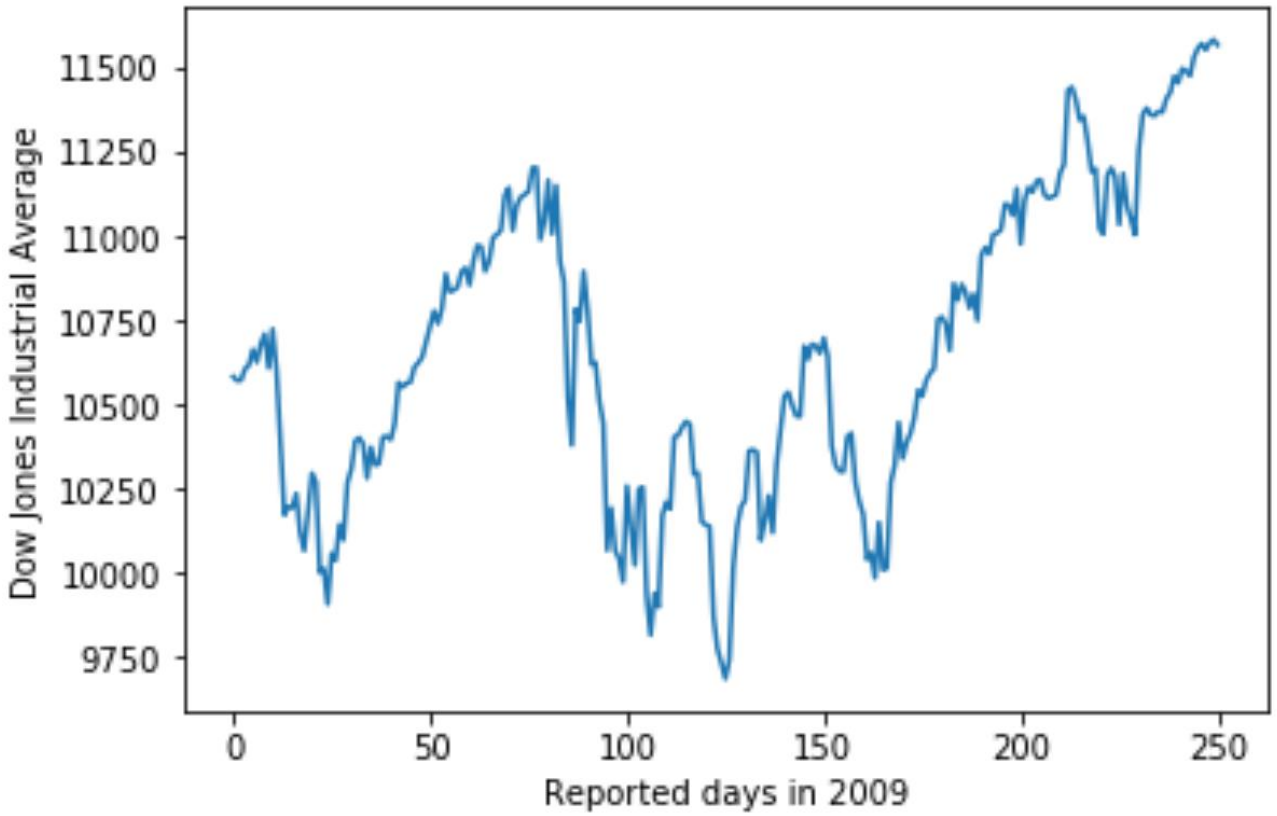
Lines plots as summary of MANY signals





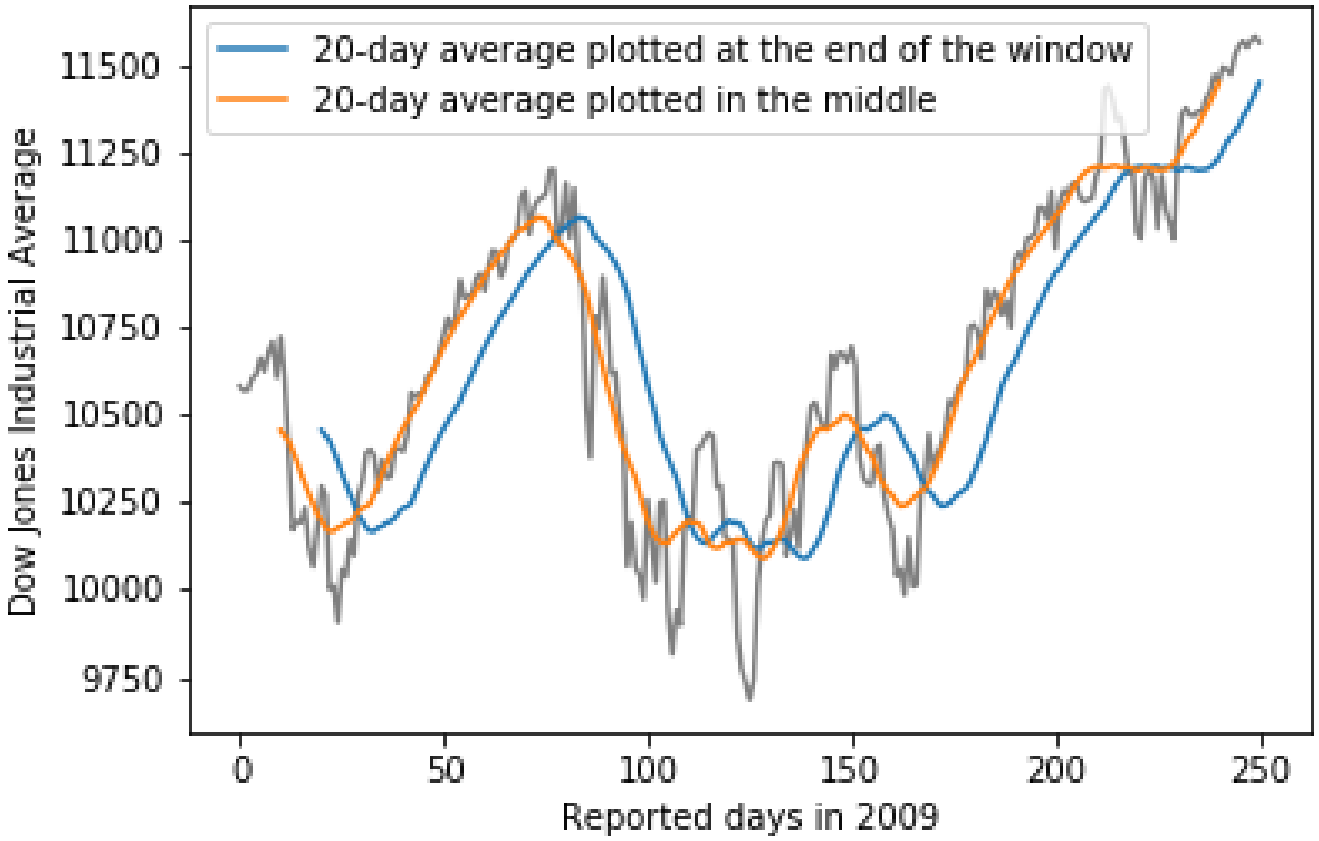
Akhtar et al. Life Science Alliance, 2019

Sliding window to show trends in complex signal shape



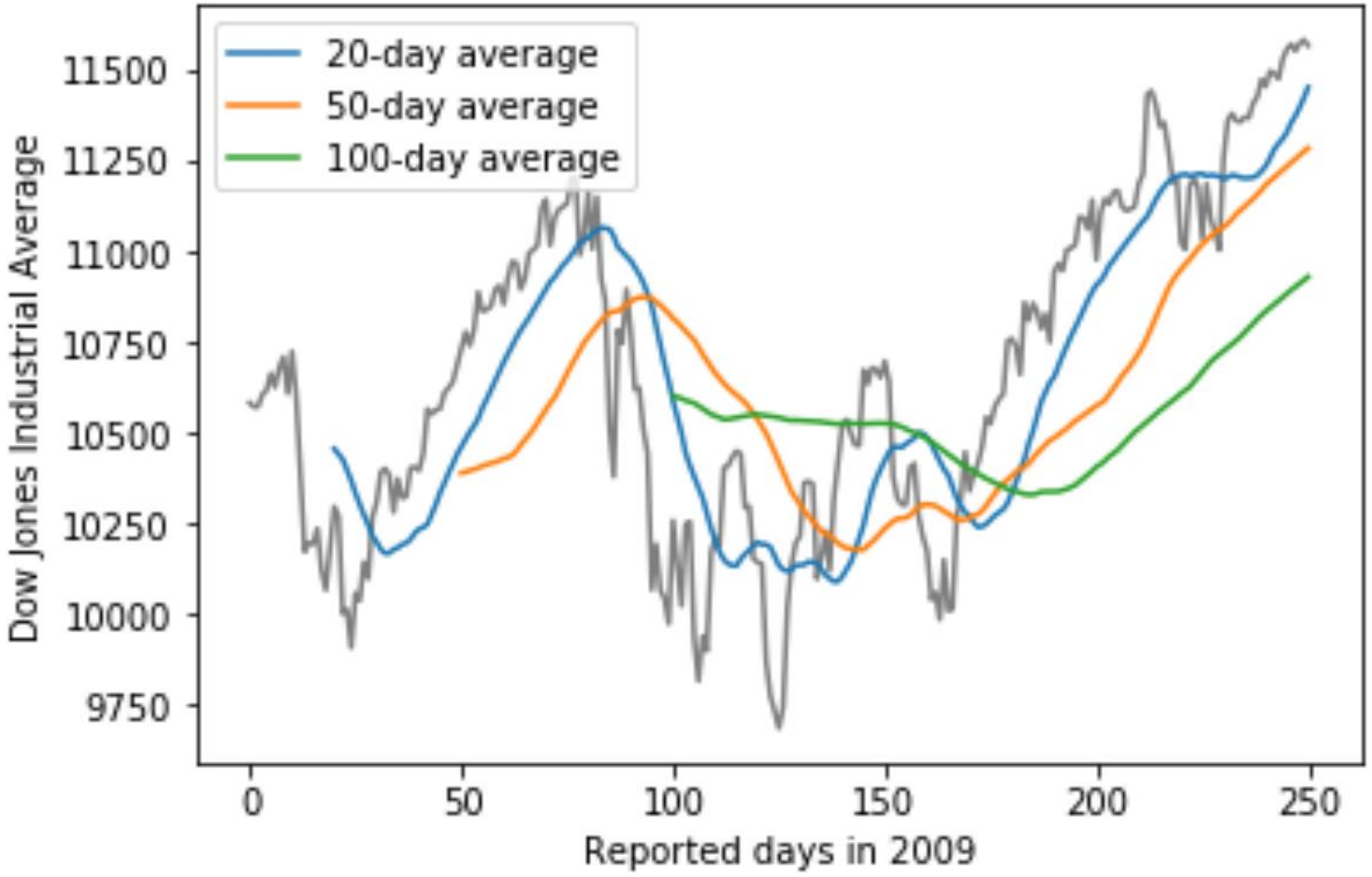
Sliding window to show trends in complex signal shape

- To capture key patterns while removing irrelevant minor detail or noise, one can calculate moving averages
- Depending on where the window is "centered"/how it's smoothed, the smoothing curve may generate a lag



Sliding window to show trends in complex signal shape

- With short-range window, small smoothing effect
- With wide-range window, can focus on a broader pattern
- As with bin width and band width (for histograms/density plots), one can manually check the patterns and determine the range



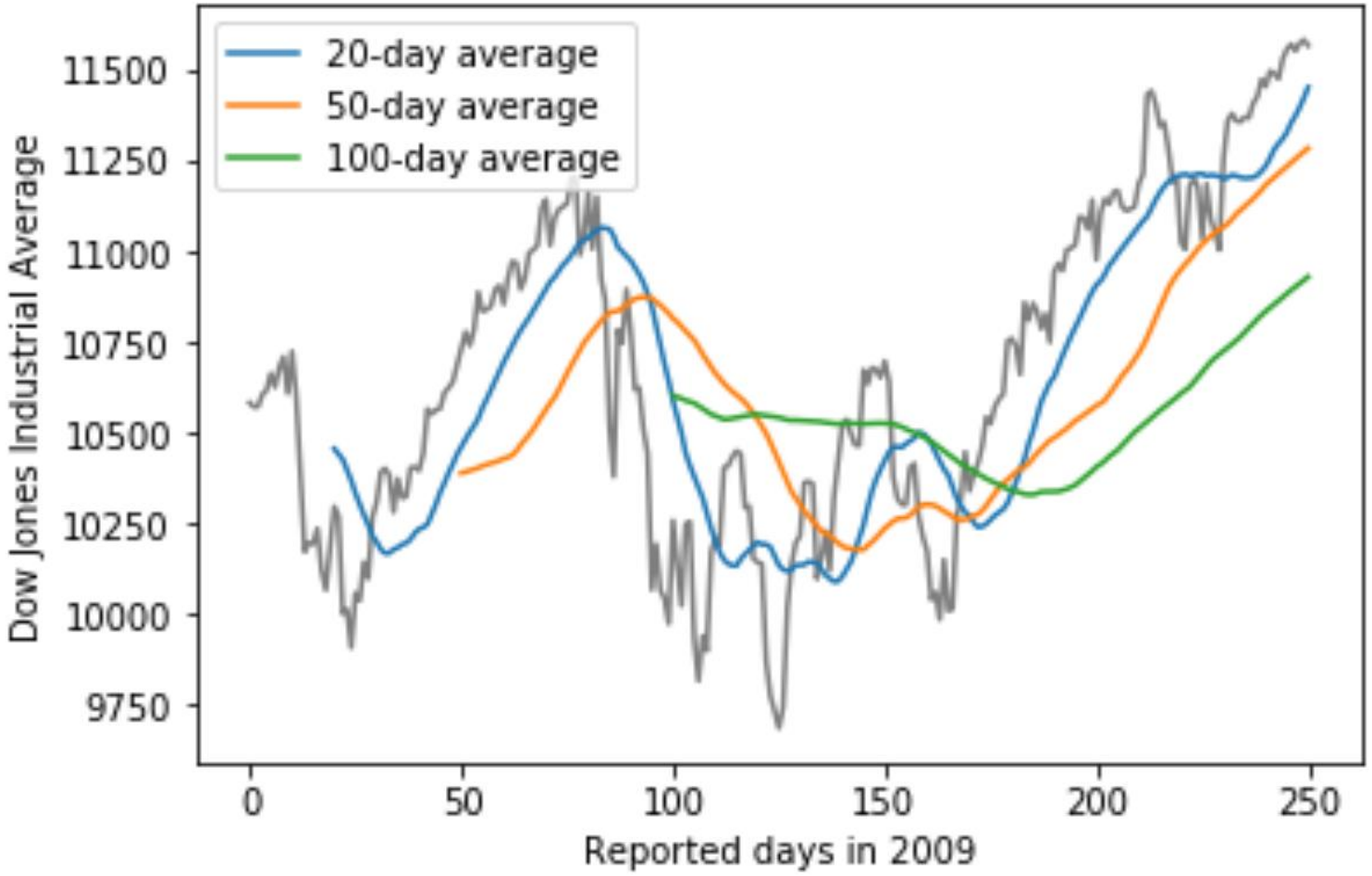
Sliding window to show trends in complex signal shape

Pros

- Simple to implement and interpret
- Often used to do AI on signal data

Cons

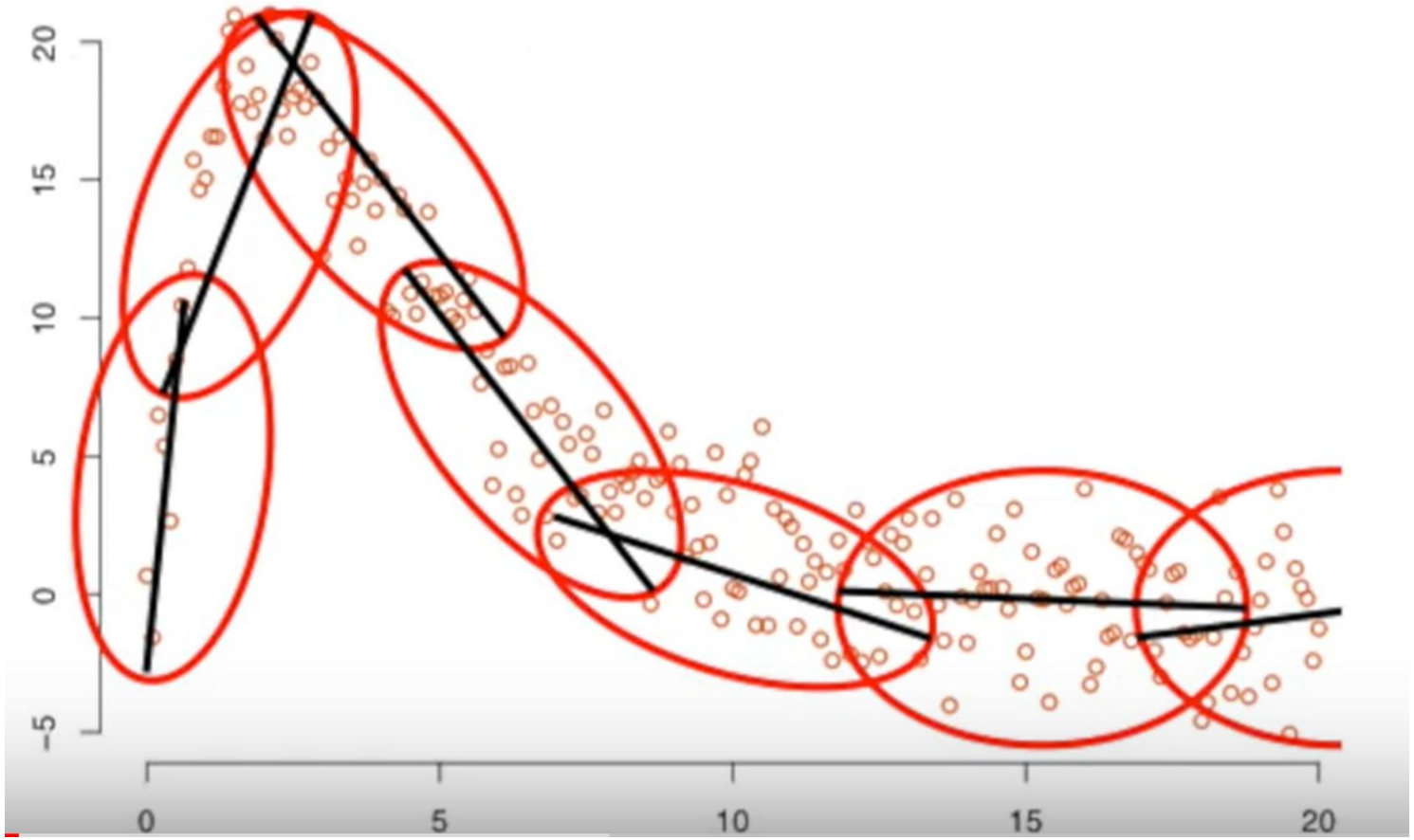
- Shorter than the original curve. The more smoothing, the shorter the curve
- How to interpret with lags?
- Still not that smooth?



Locally Weighted Scatterplot Smoothing (LOWESS)

Like moving window, use a sliding window to divide the data into smaller chunks

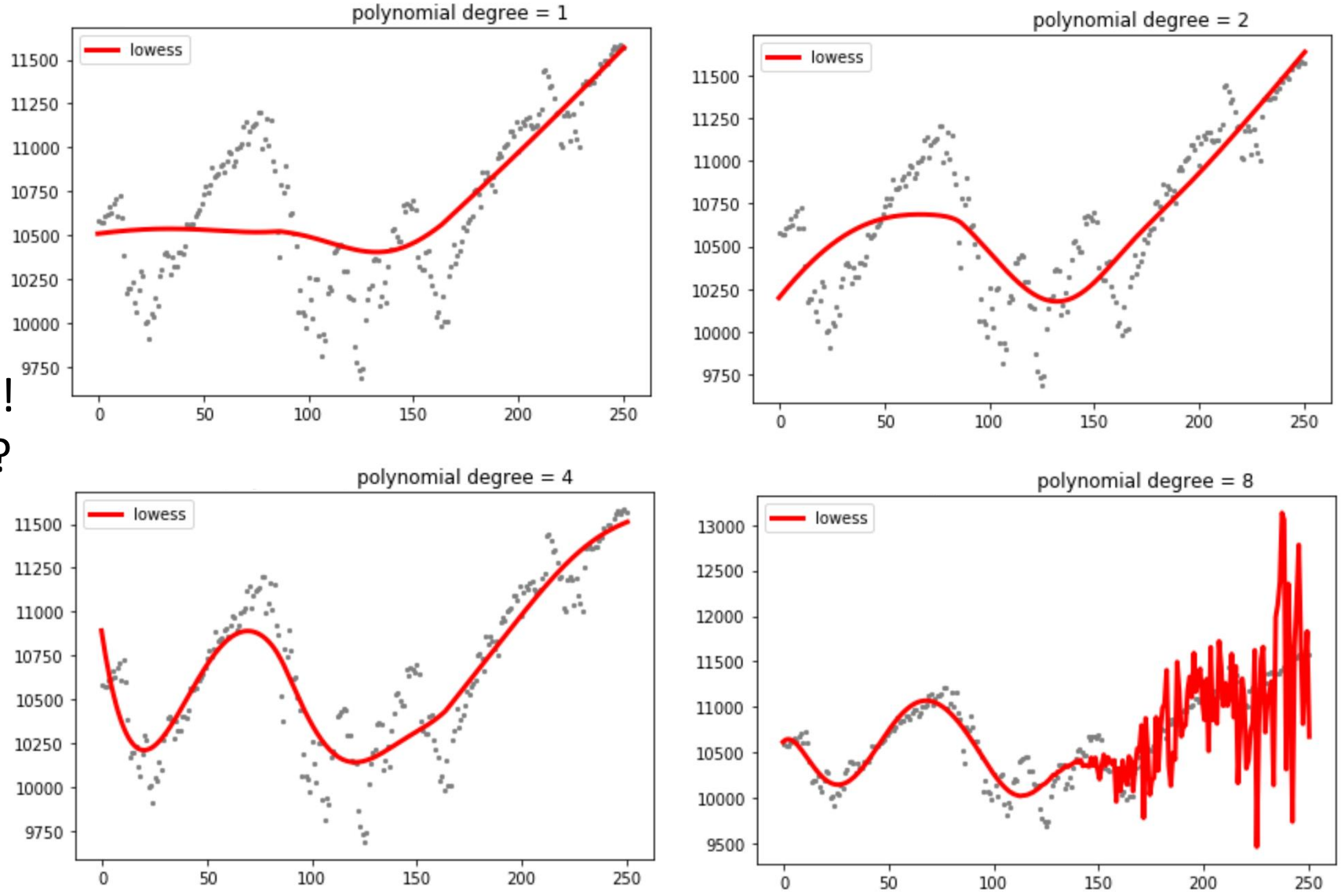
Unlike a moving window, it fits a polynomial regression in each window, where center points in the window are weighted more heavily



Adapted from https://www.youtube.com/watch?v=Vf7oJ6z2LCc

LOWESS with different fitting parameters

- Polynomial degree matters!
- Implications for prediction?

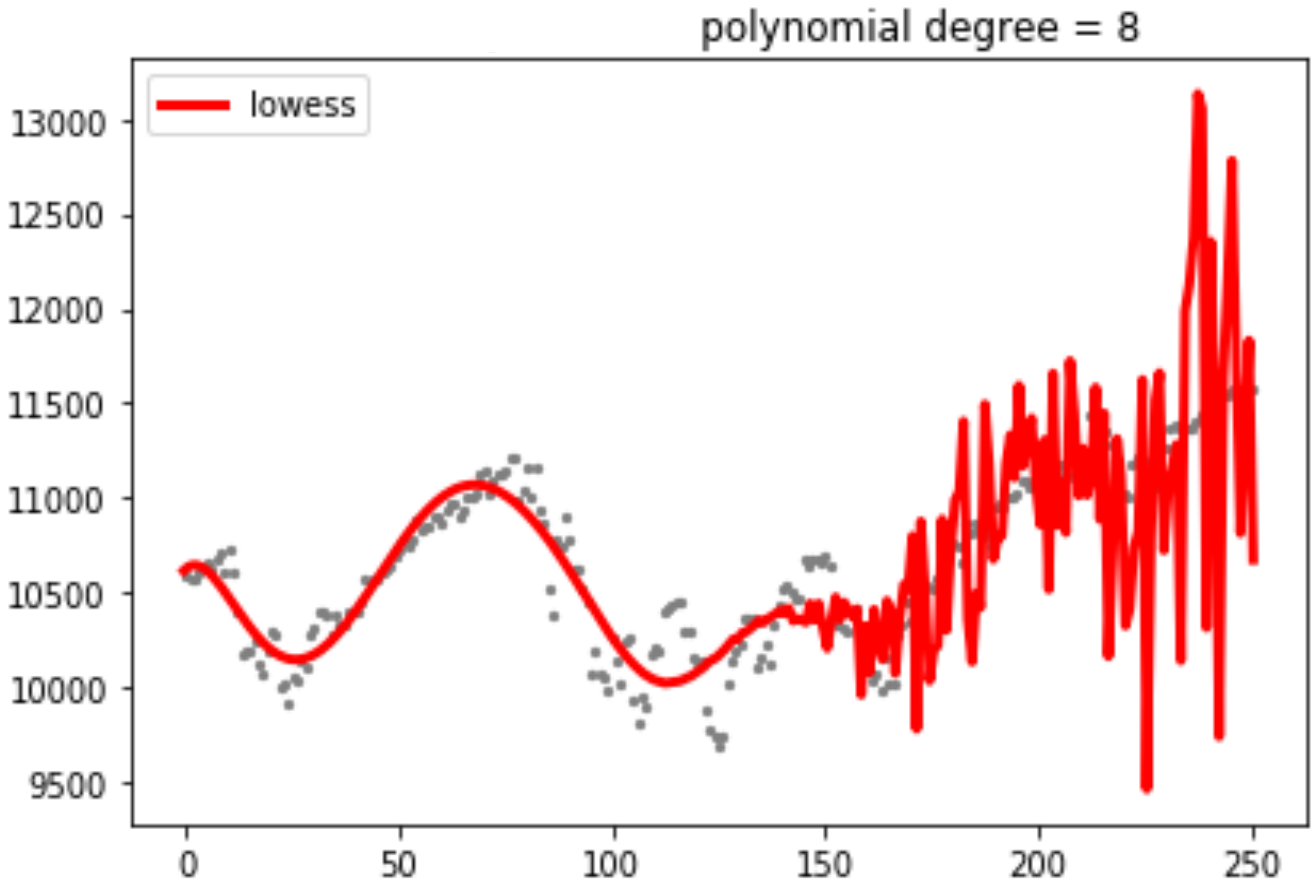


High polynomial degree may lead to overfitting in LOESS

Overfitting fits the model to the non-signal parts of data that the resulting model

- Does not represent the underlying signal
- Cannot predict future observations reliably

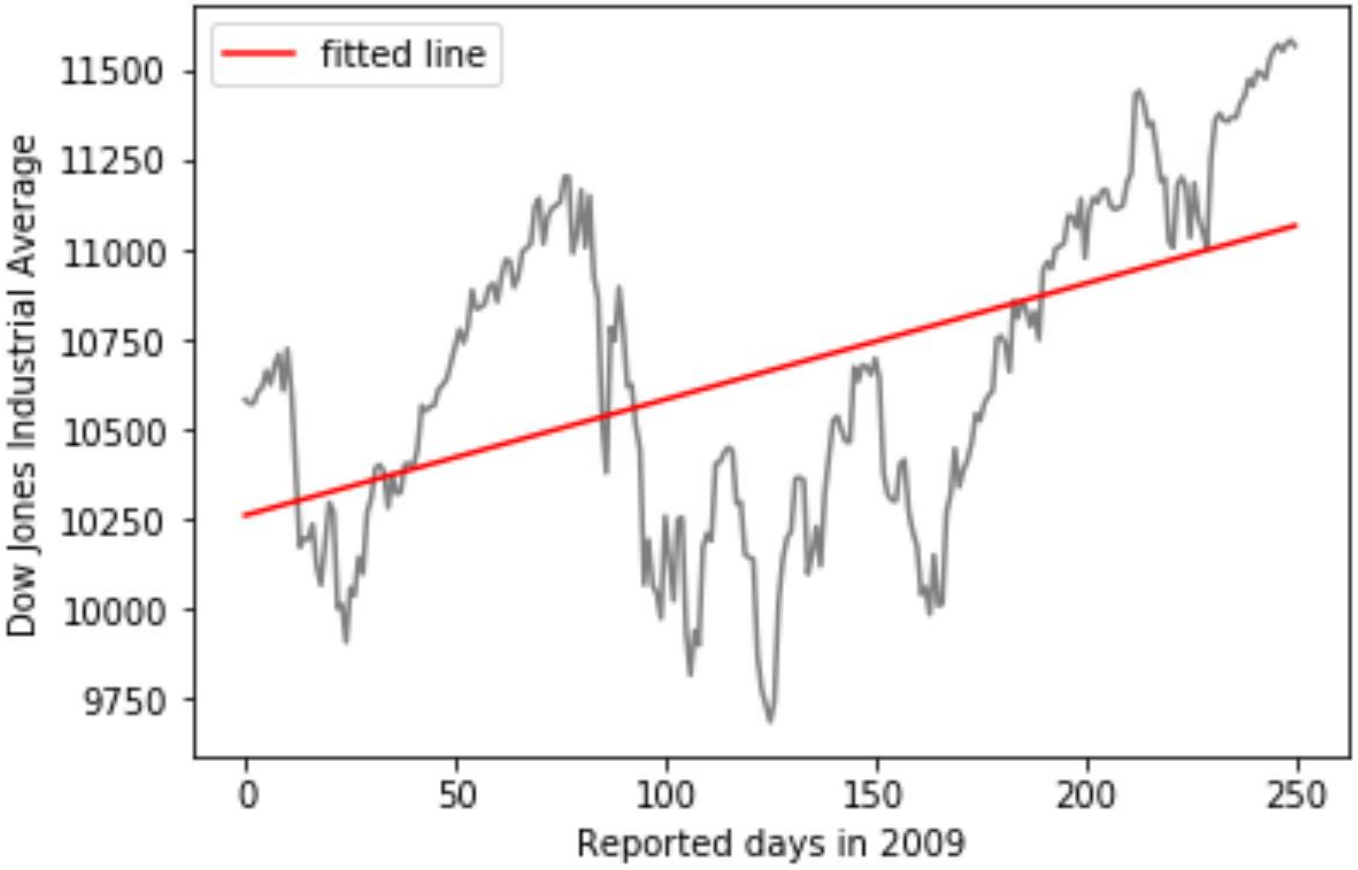
- However, LOESS is not for prediction



Linear least square regression to predict based on linear trend

(A, m)=argmin($\sum_x (y_x - F(x))2$),

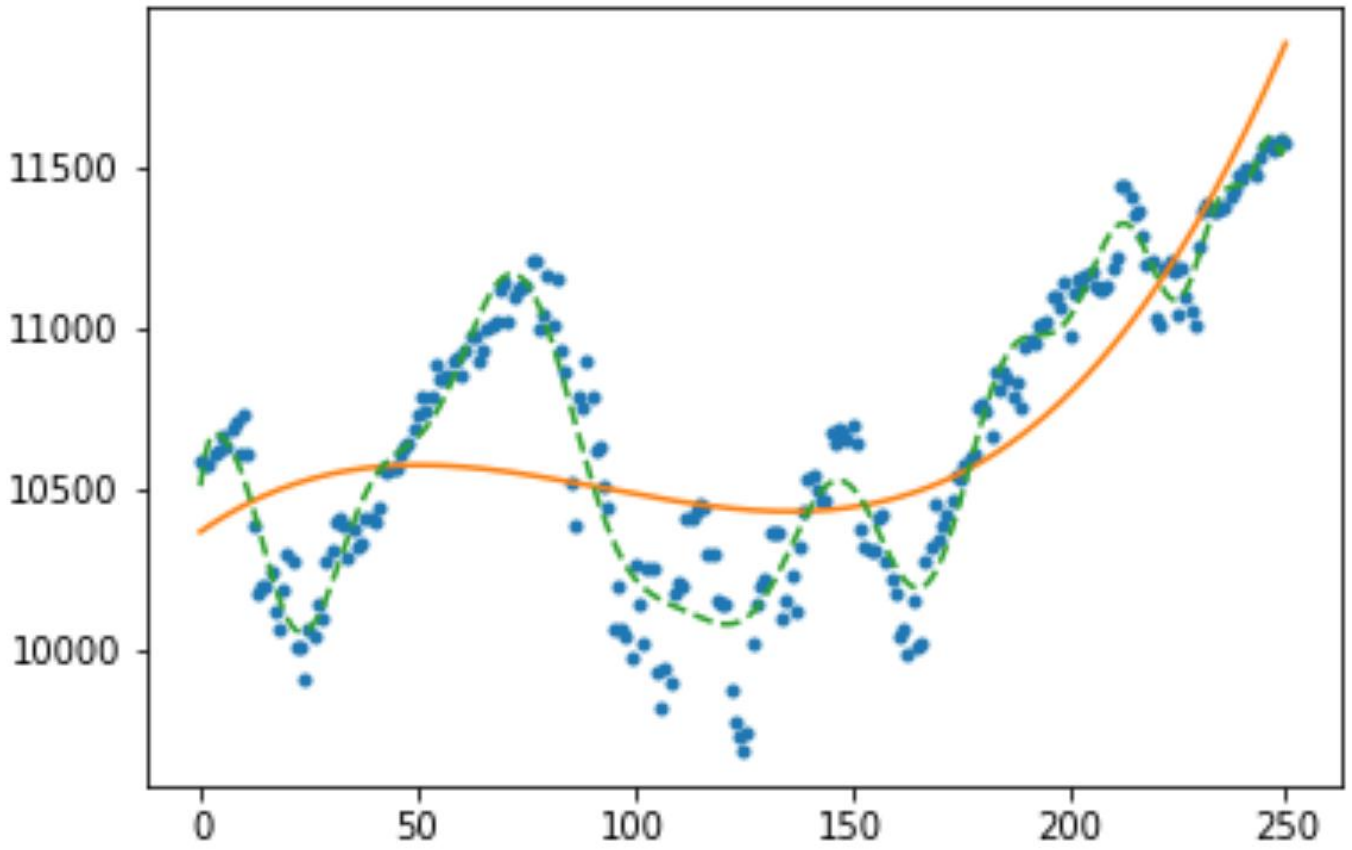
where F(x)=A+mx



Quadratic least square regression to predict based on nonlinear trend

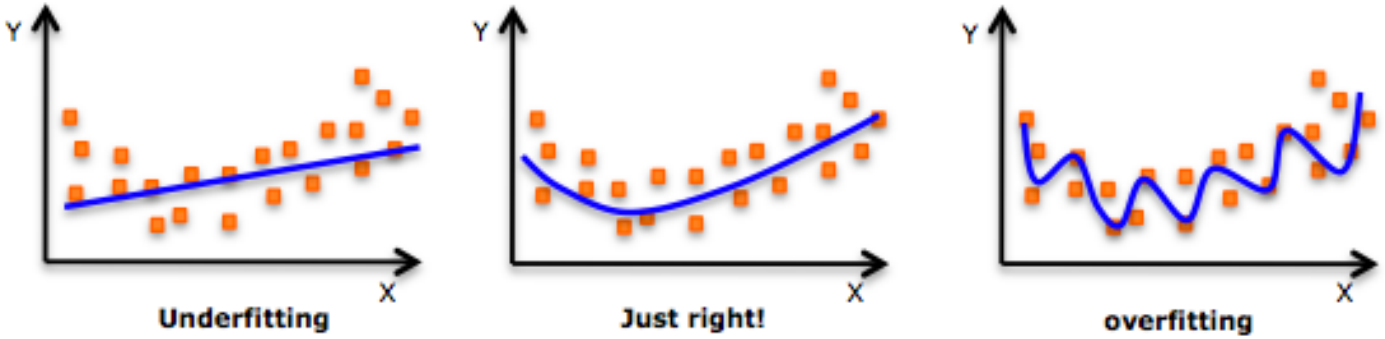
(A, m)=argmin($\sum_x (y_x - F(x))2$),

, where F(x) $= ax^2 + bx + c$



Systematic ways to overcome the overfit

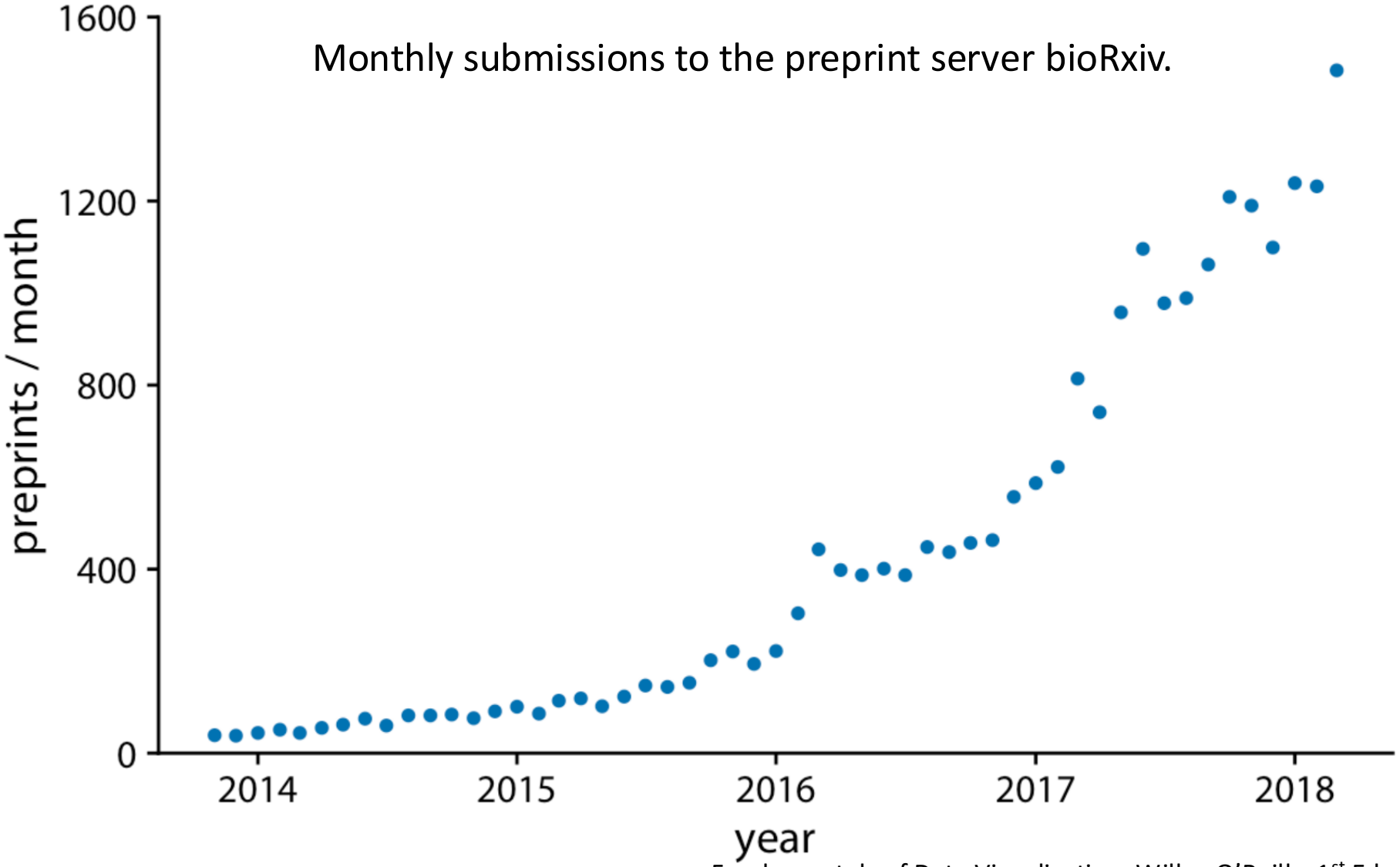
- The higher the degree is (the more variables it has), the better it fits to the data
- What is the right degree (or complexity or the number of variables) of the model for the data?
- Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), or variable selection (e. g., via penalization)



https://www.quora.com/What-are-the-key-trade-offs-between-overfitting-and-underfitting

Transforming axes may help explain the underlying relationship

- $\Delta y(t) \propto y(t-1)$
- E. g., the increase in submissions in a month

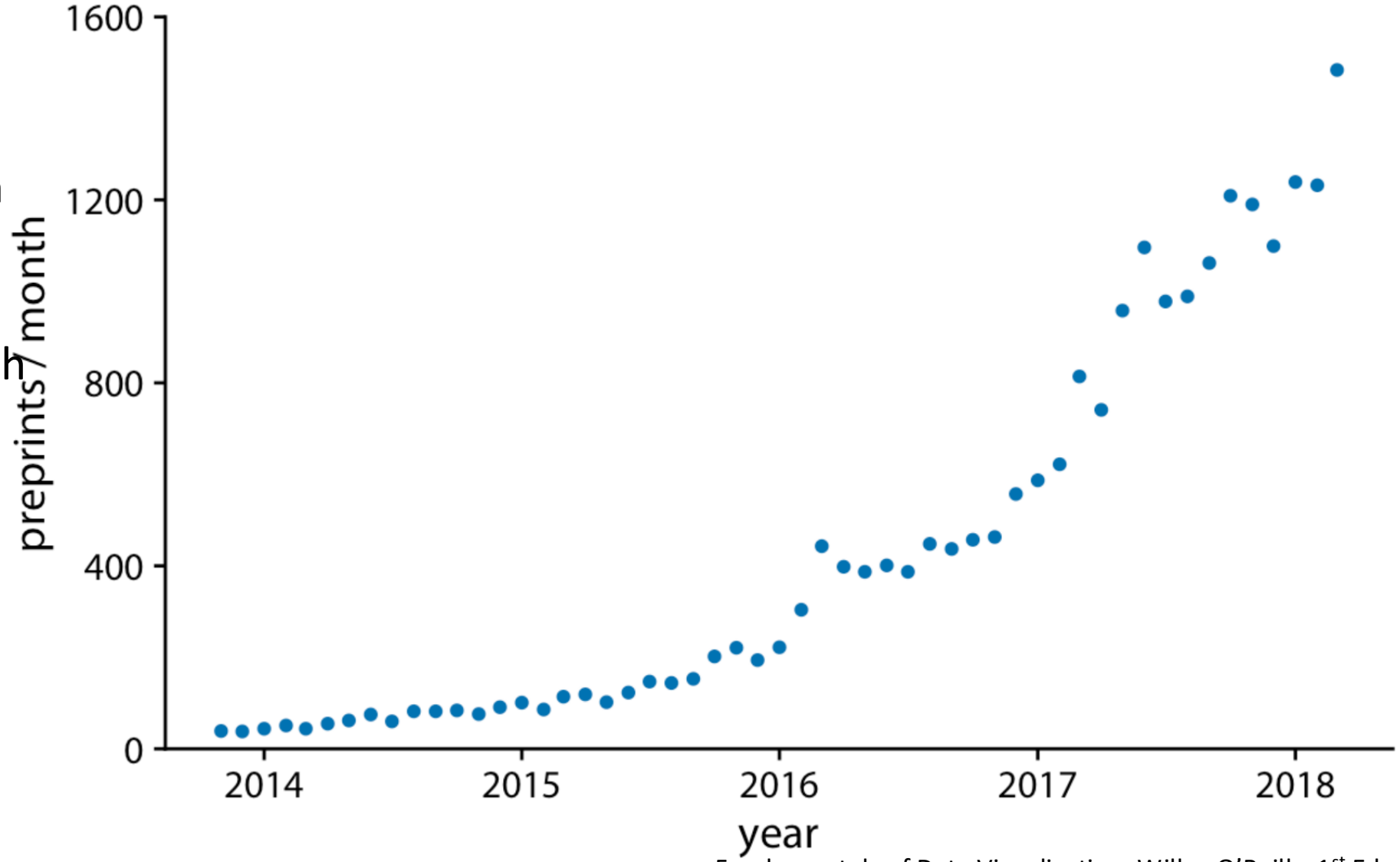


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Transforming axes may help explain the underlying relationship

- $\Delta y(t) \propto y(t-1)$
- E. g., the increase in submissions in a month
- **Exponential growth function** grows at a constant percent growth rate
- $y=\exp(A+mx)$

Monthly submissions to the preprint server bioRxiv.



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

How do I select a distribution to try fitting

- Log-linear
  ➡ log(y)=A+mx

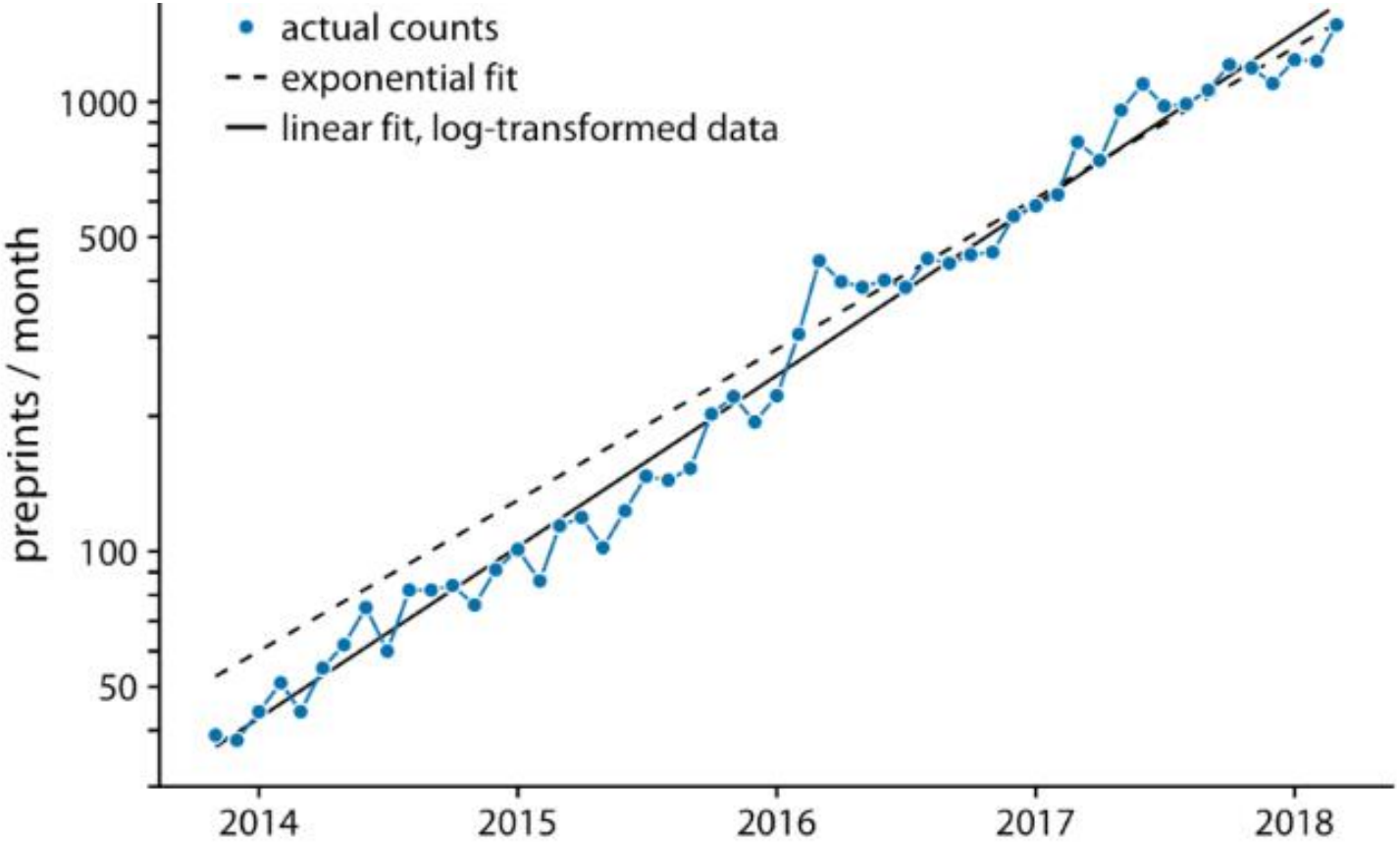
- log-log
  ➡ $y \sim x^{\alpha}$

  ➡ power laws

- Linear-log
  ➡ y~log(x)

- cover a wide range of real-world applications

Monthly submissions to the preprint server bioRxiv.

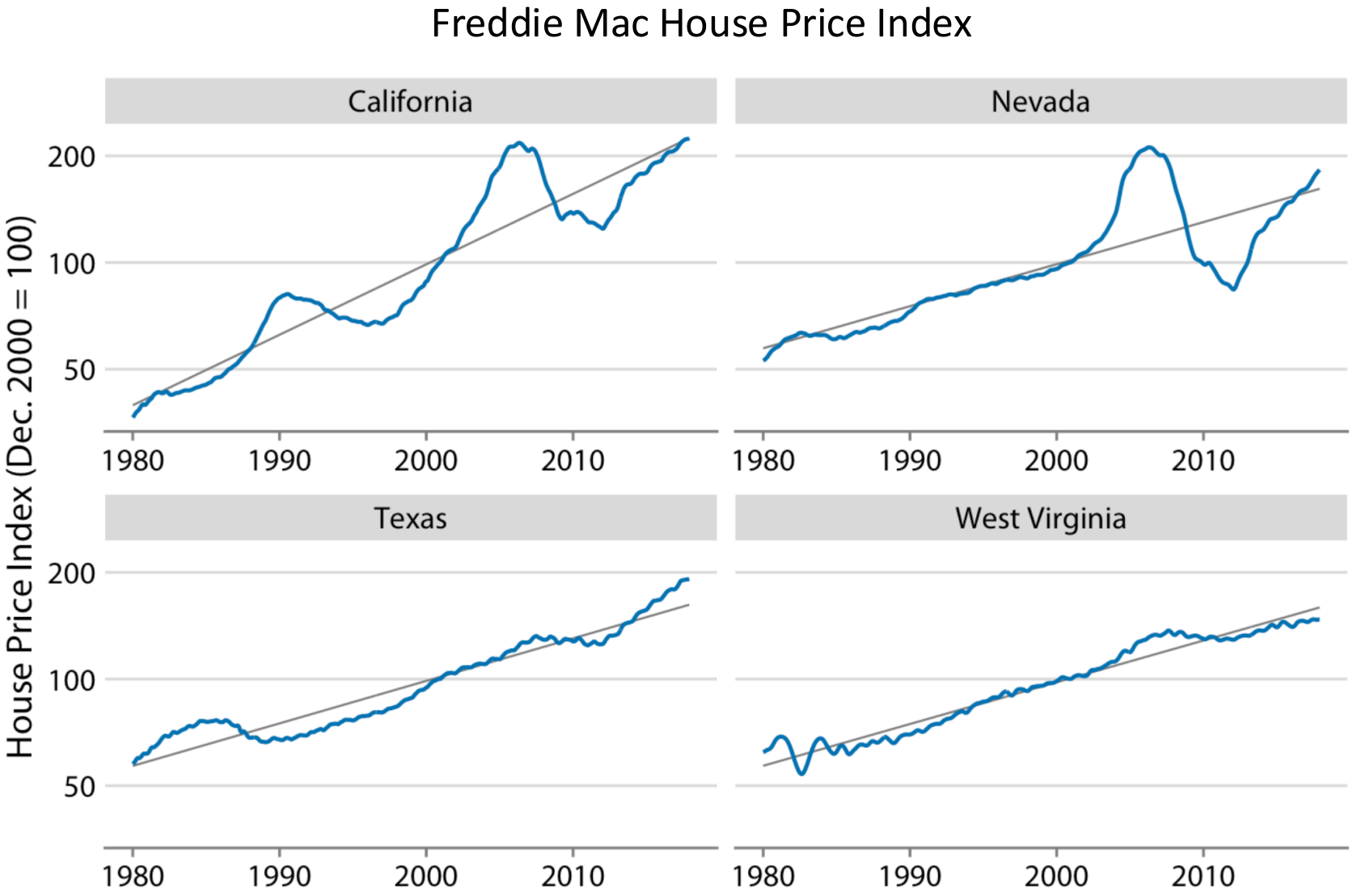


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Detrending to catch signals beyond the “first” trend

House Price = annual growth
+ house bubbles

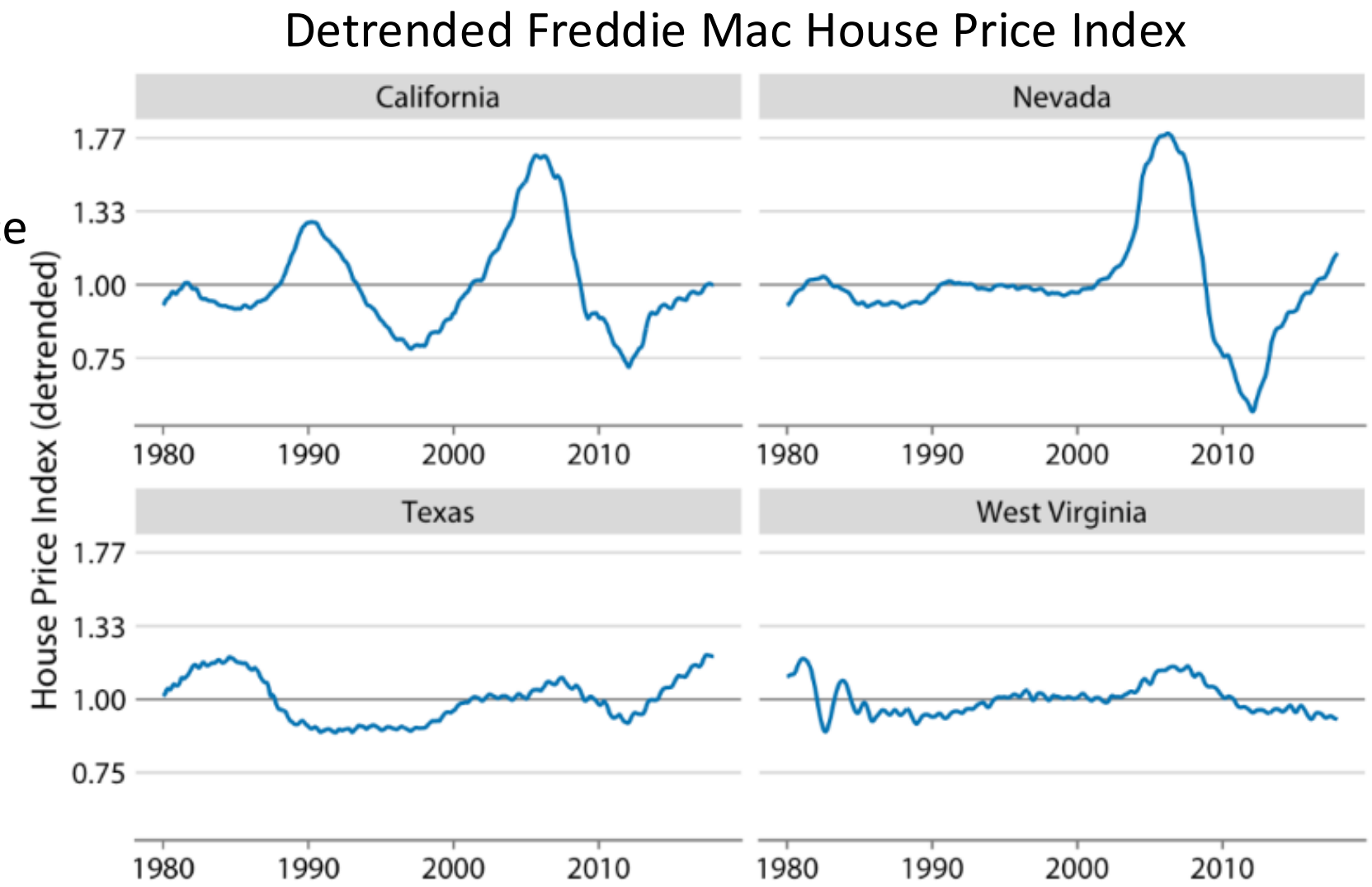
- 4.7% annual price increase with two bubbles in California
- 2.8% in the other states



Fundamentals of Data Visualization, Wilke, O’Reilly, 1st Ed.

Detrending signals for patterns beyond trend

- House bubbles = house price - annual growth
- A way to check patterns beyond a prominent trend

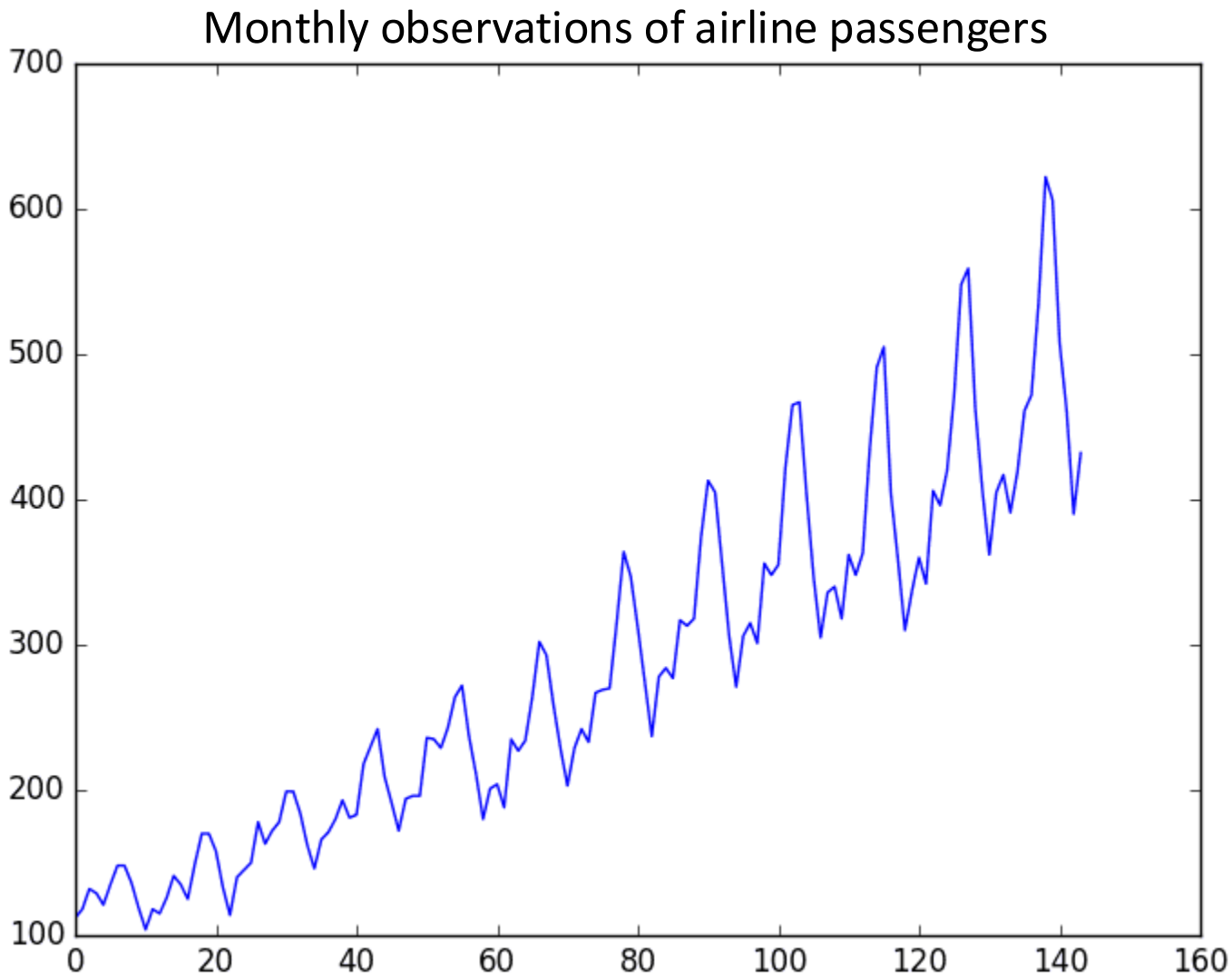


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Detrending time-series data into multiple factors

A time series consists of

1. Long-term,
2. Cyclical variations of various intervals
3. Etc (random noise+unique external events)



https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

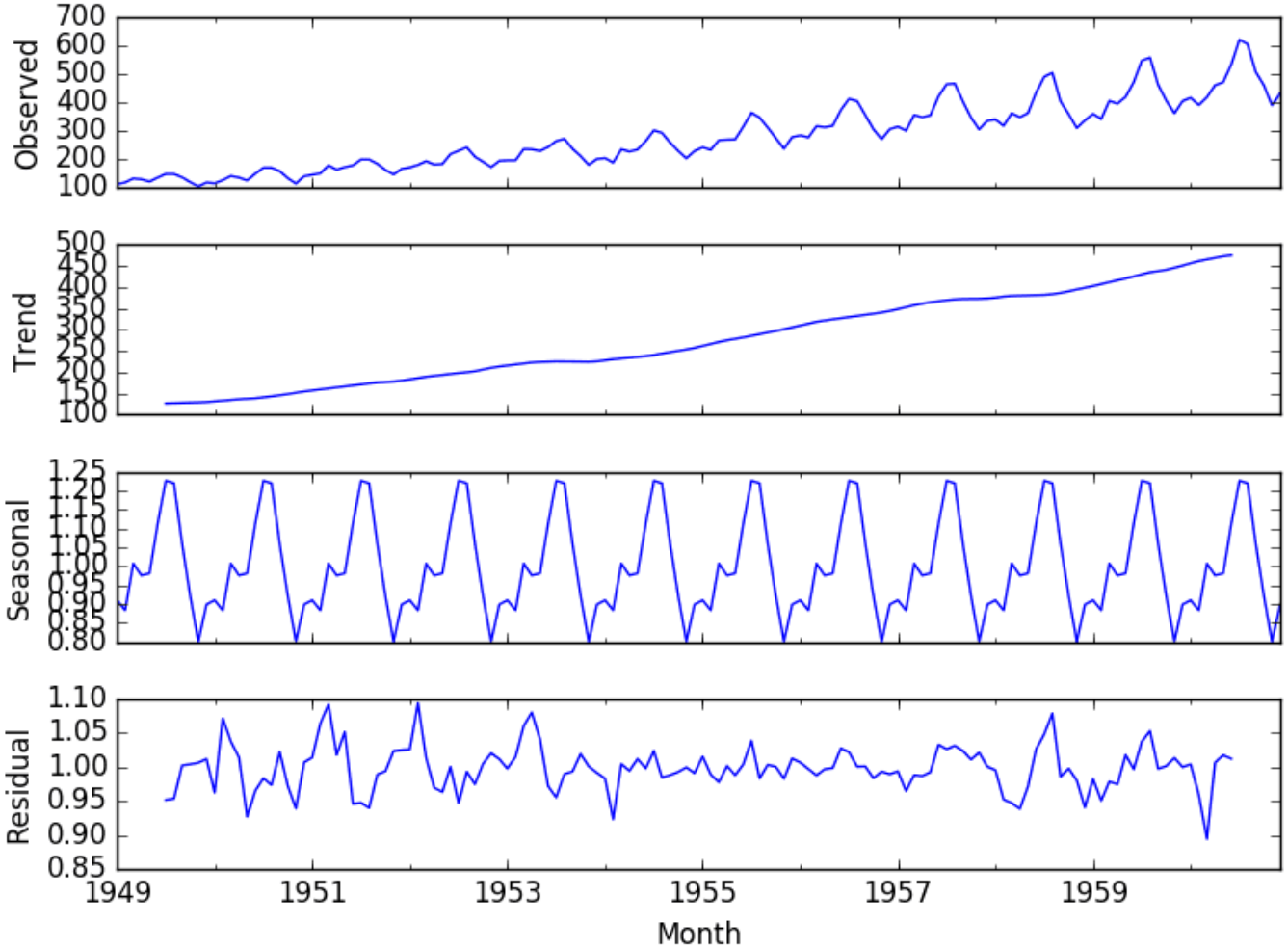
Detrending time-series data into multiple factors

A time series consists of

1. Long-term,
2. Cyclical variations of various intervals
3. Methods available. E. g.,in python

```
from pandas import read_csv
from matplotlib import pyplot
from statsmodels.tsa.seasonal import seasonal_decompose
series = read_csv('airline-passengers.csv', header=0, index_col=0)
result = seasonal_decompose(series, model='multiplicative')
result.plot()
pyplot.show()
```

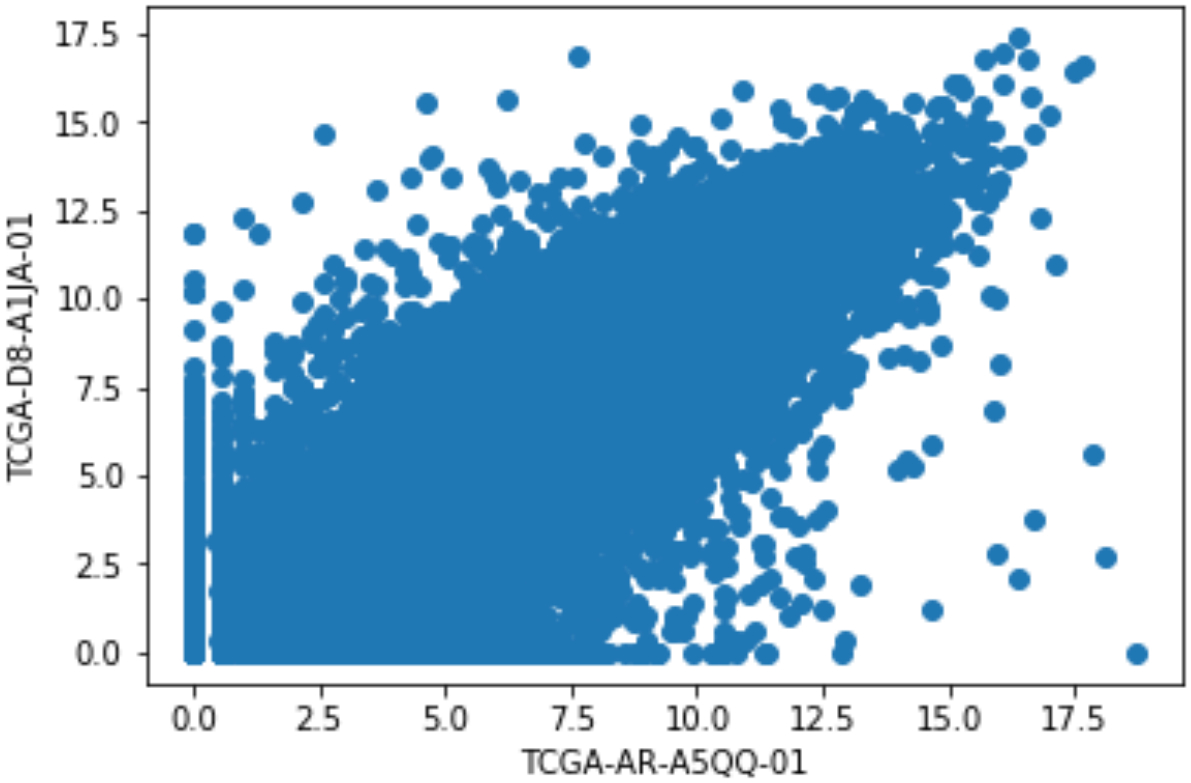
Monthly observations of airline passengers



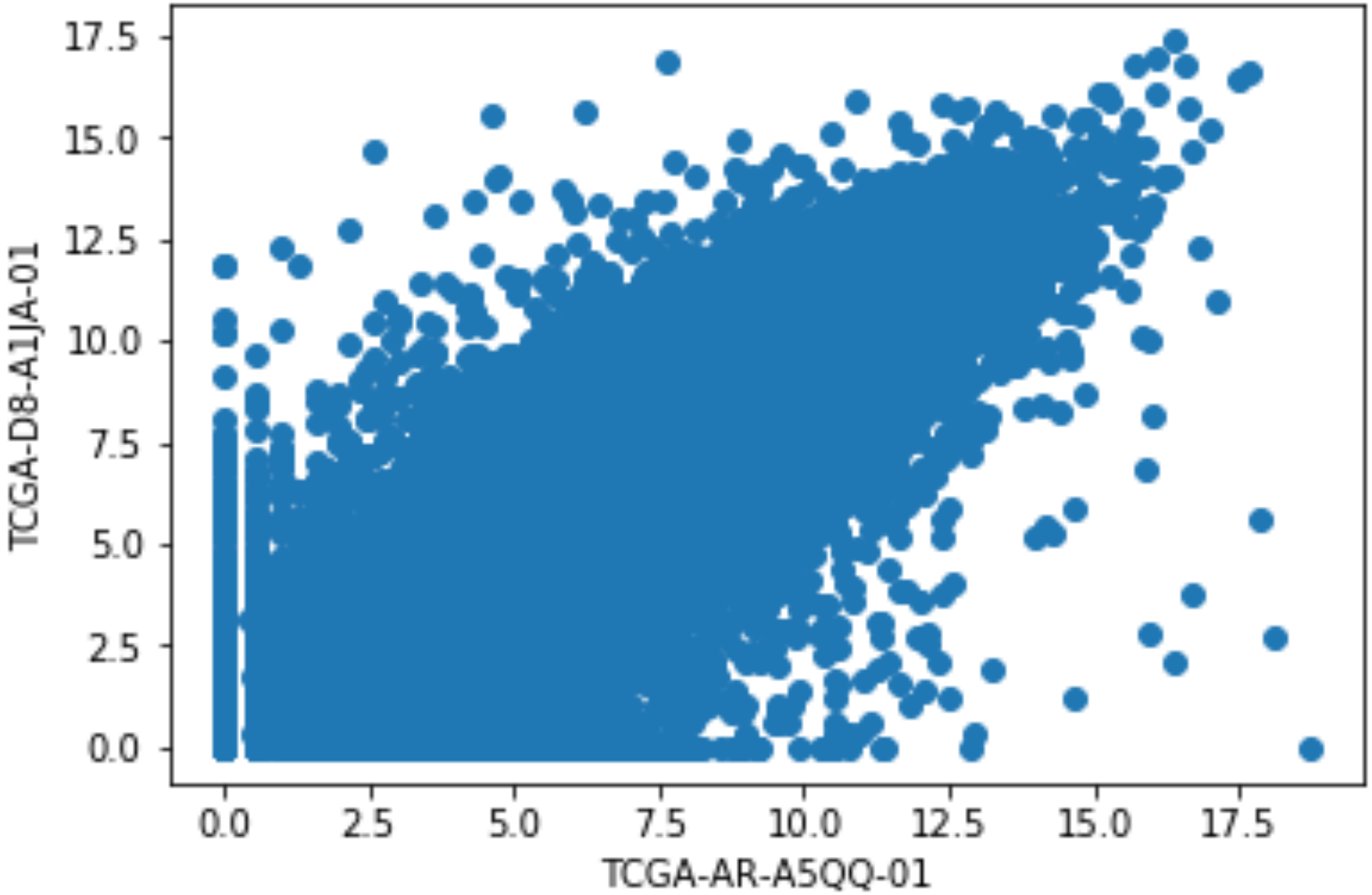
https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

Scatterplots

- Associate two samples on the same variable
  - E. g., two tumor samples on the same set of genes
  - Can show the genes expressed in either
- Or associate two different variables across many samples



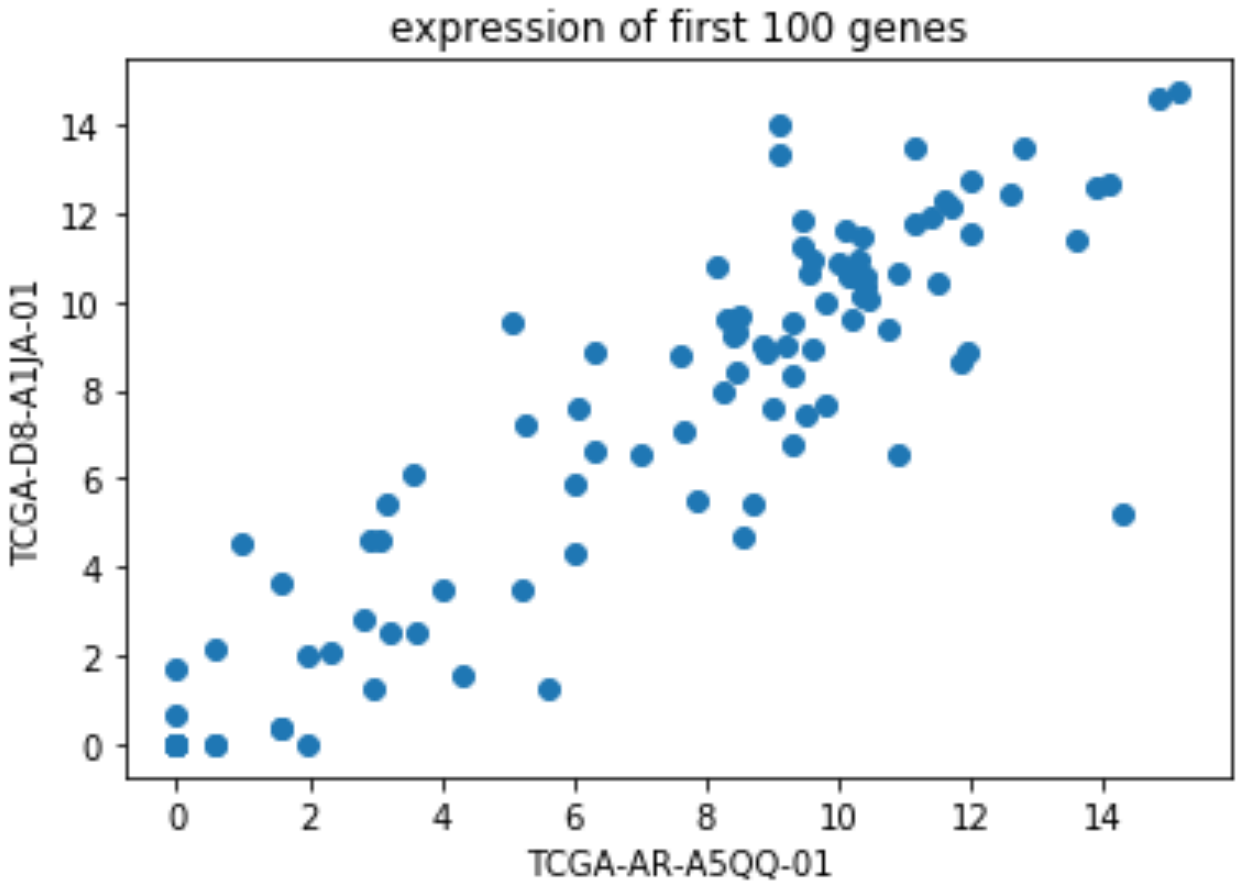
Overplotting



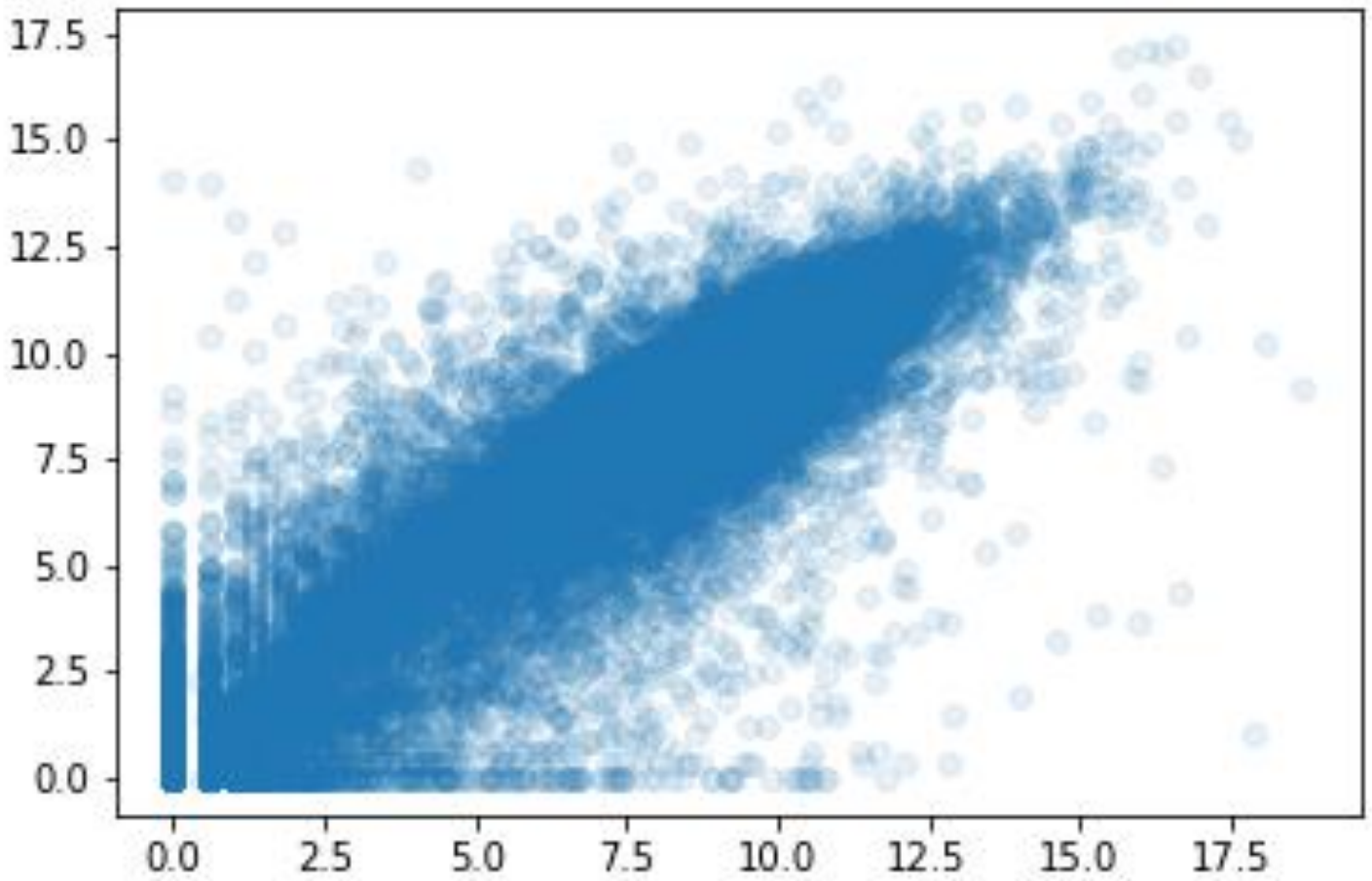
Down-sampling to address overplotting?

Not a solution!

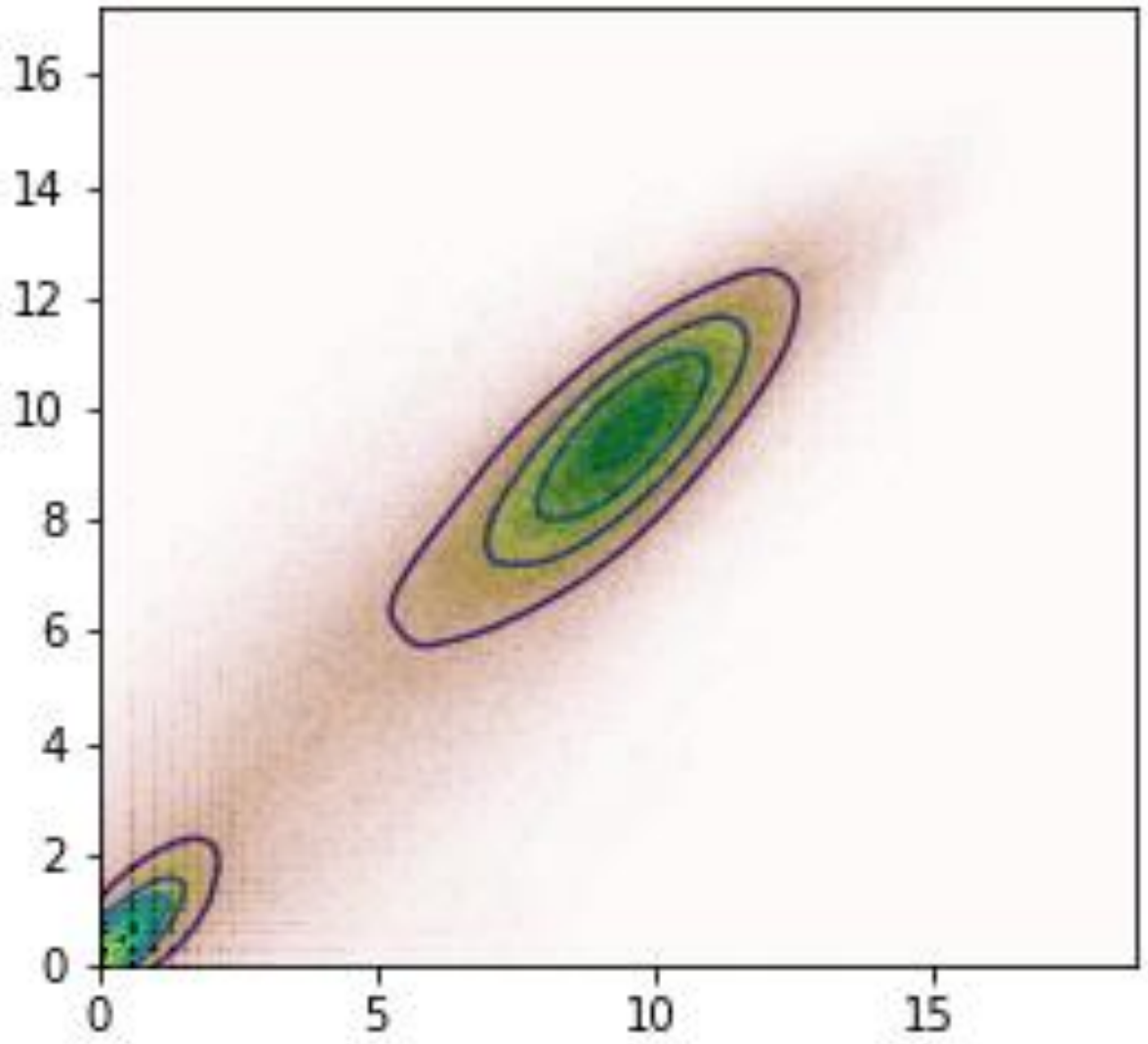
- A random sampling might be OK for exploration
- Sampling represents a portion of the input data.
- Any sampling approaches need to be done multiple times for statistical inference (e.g., bootstrapping)
- Not recommended to remove data



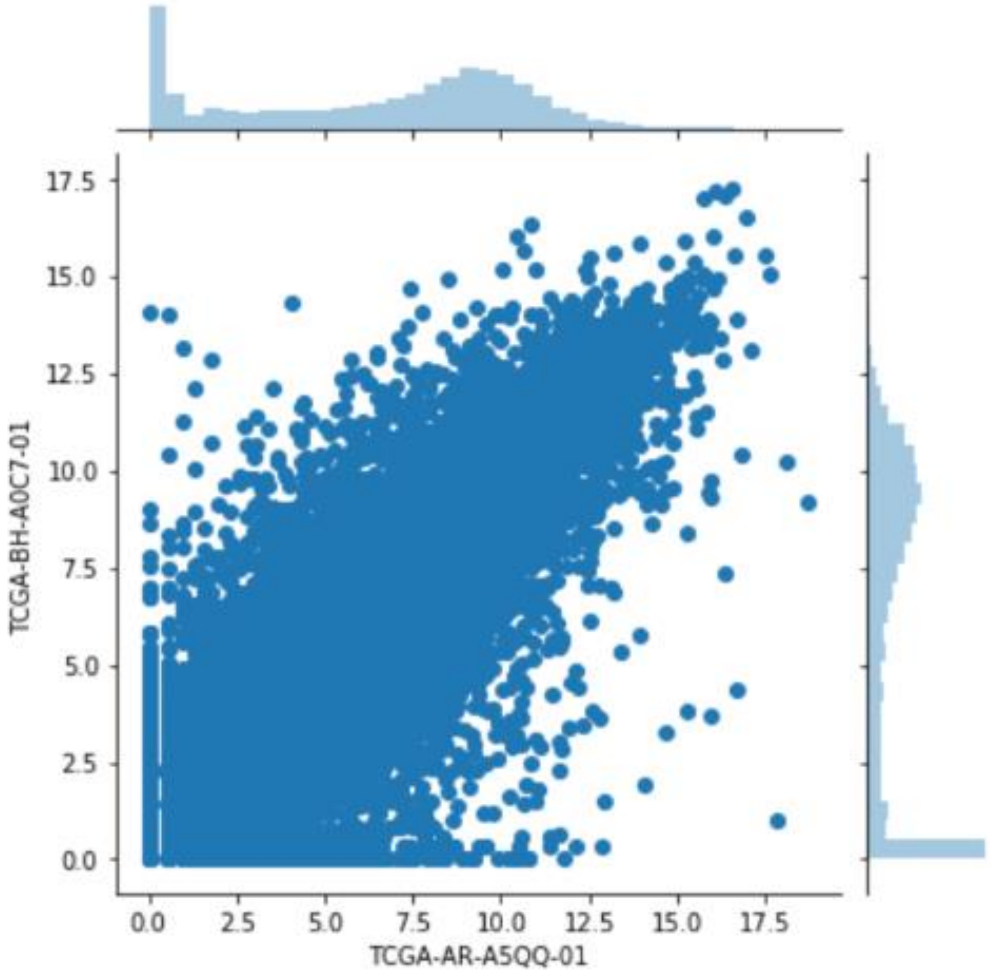
Transparency to overcome overplotting?



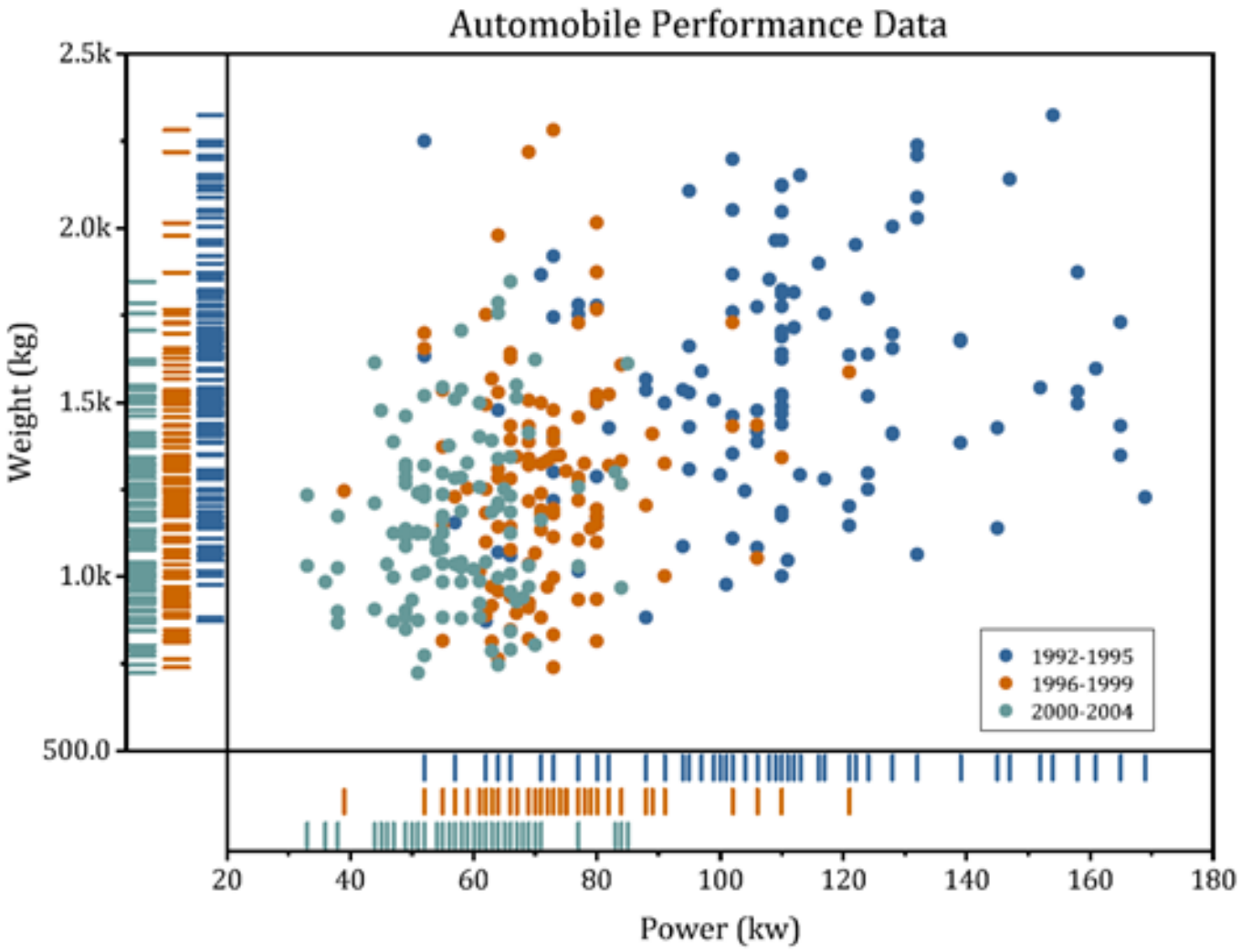
Contour lines to overcome overplotting?



Plotting marginal distribution to address overplotting?

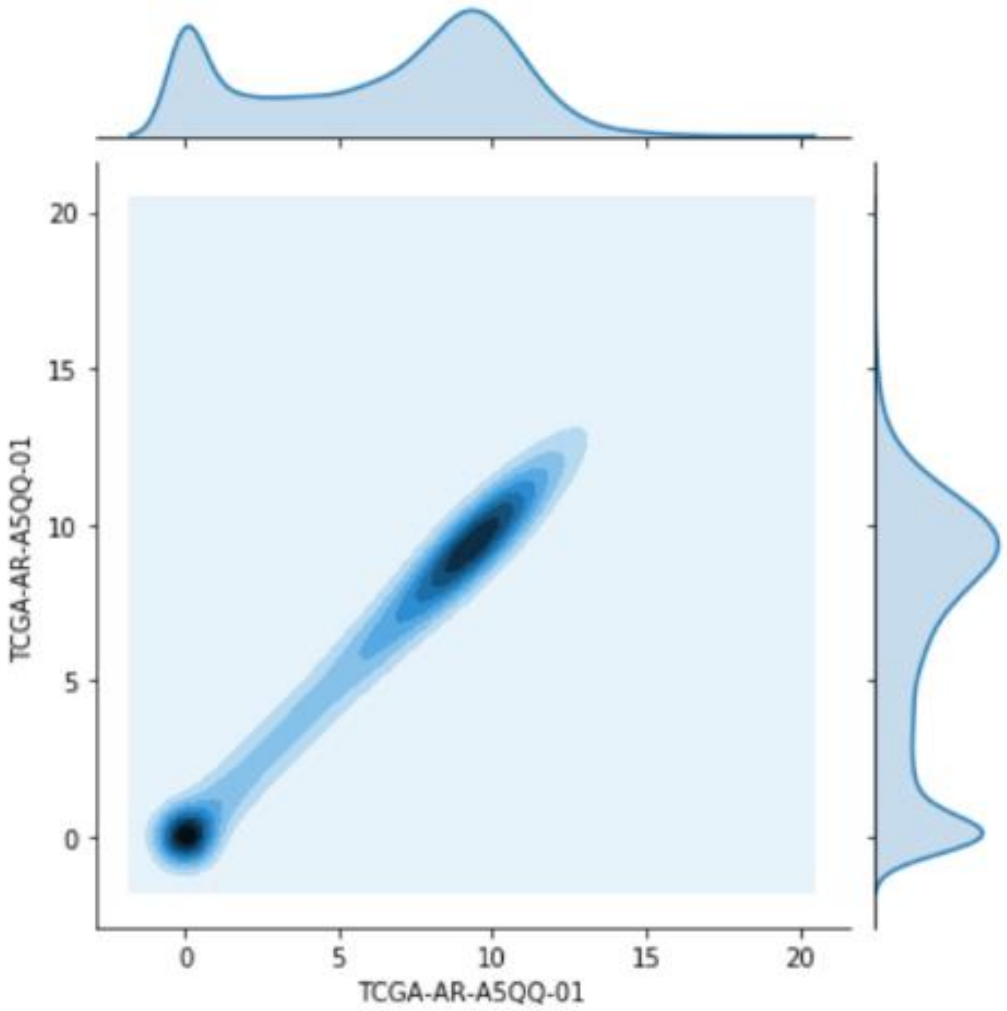


Similar: rug plots to view marginal dist’n.



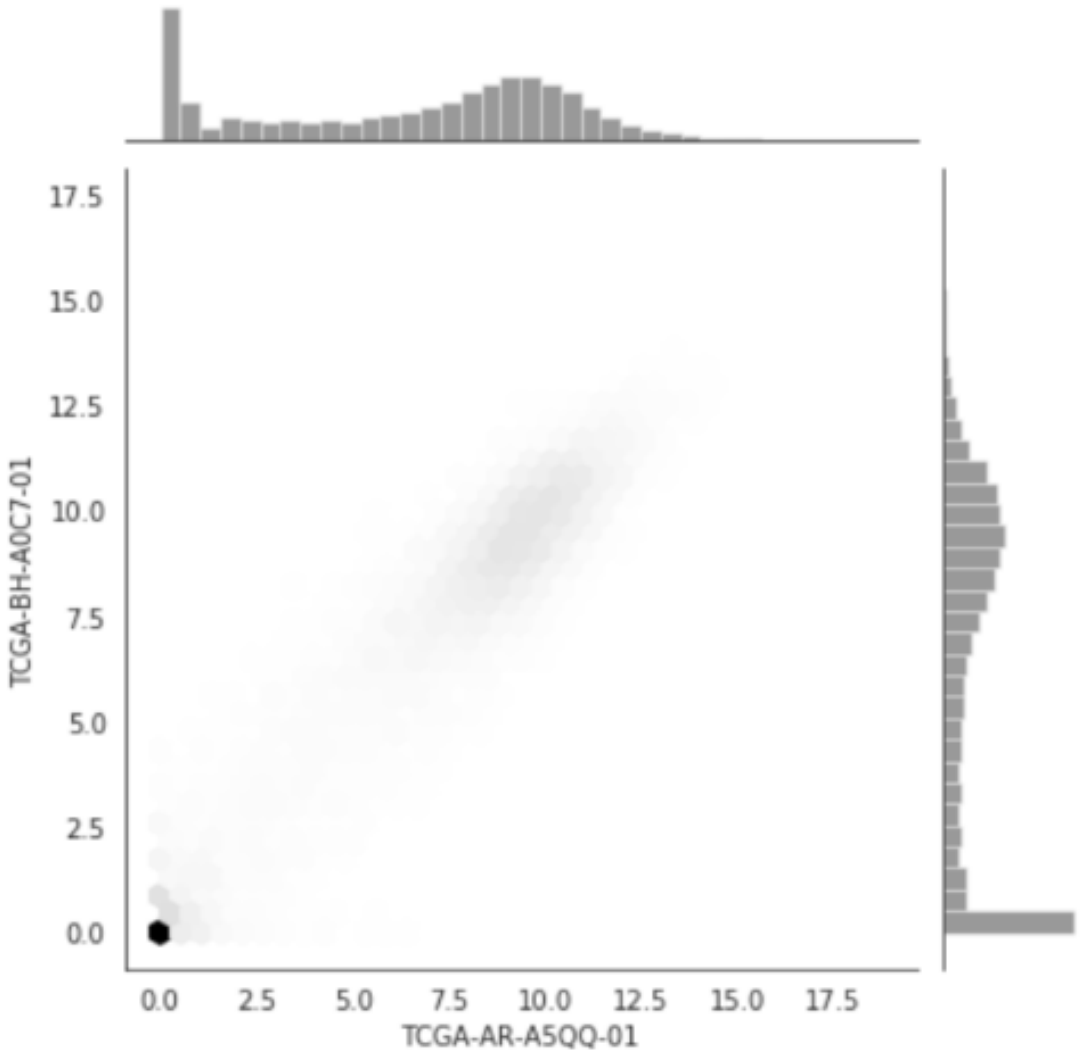
Signal density plots with marginal distribution to address overplotting?

Signal density estimation smoothes the peak at the low expression level



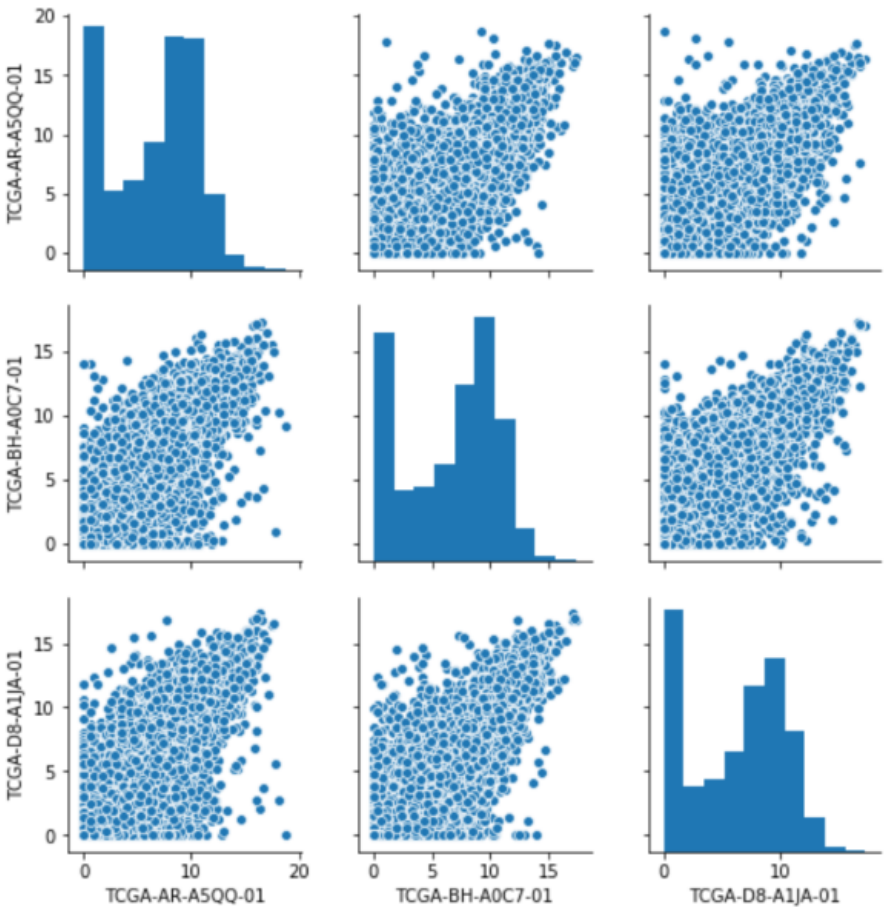
Hexabin plots with marginal distribution to address overplotting?

Hexabin = the bivariate analogue of a bin in a histogram



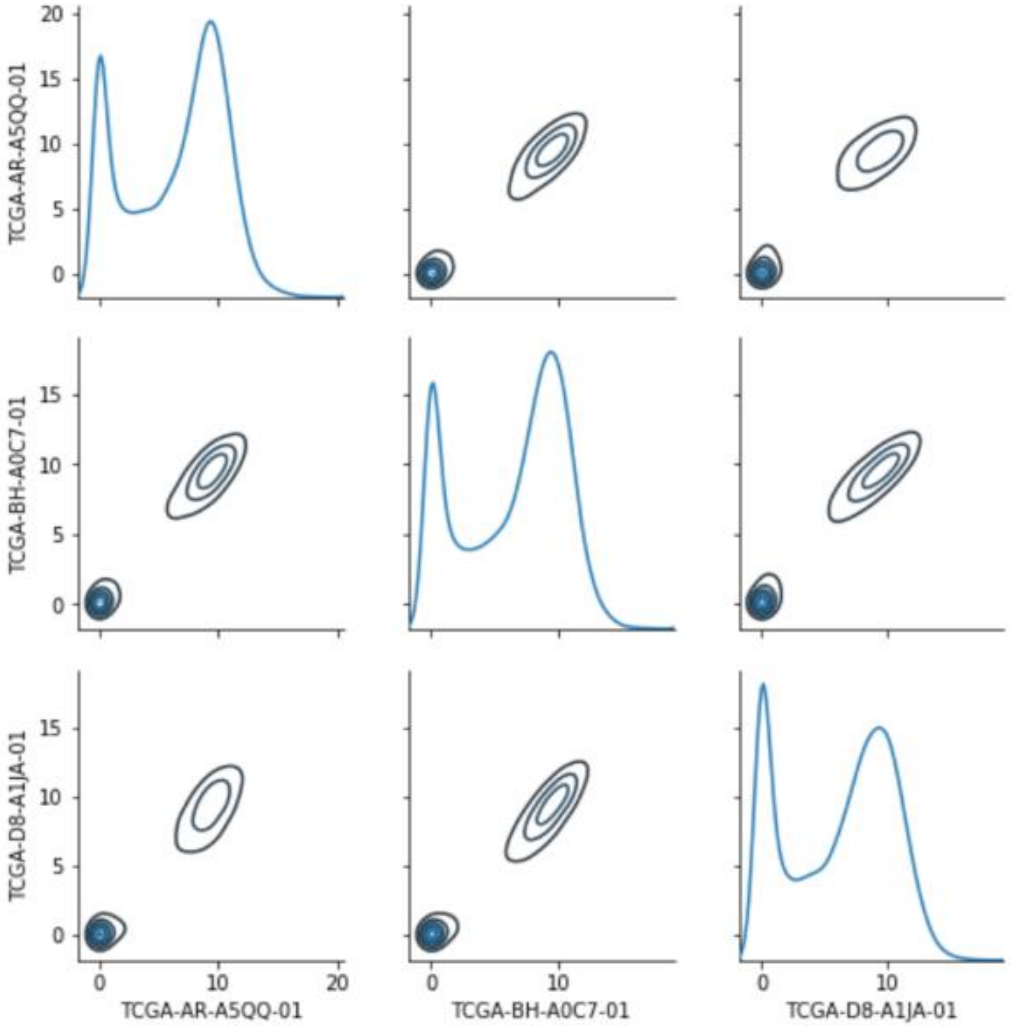
Pairwise plots to visualize pairwise relationships in a multidimensional dataset?

- Plotting a bivariate distribution in a figure
- Aligning the figures in a structured way while assuming no order among the conditions
- In R, use ggpairs from GGally package



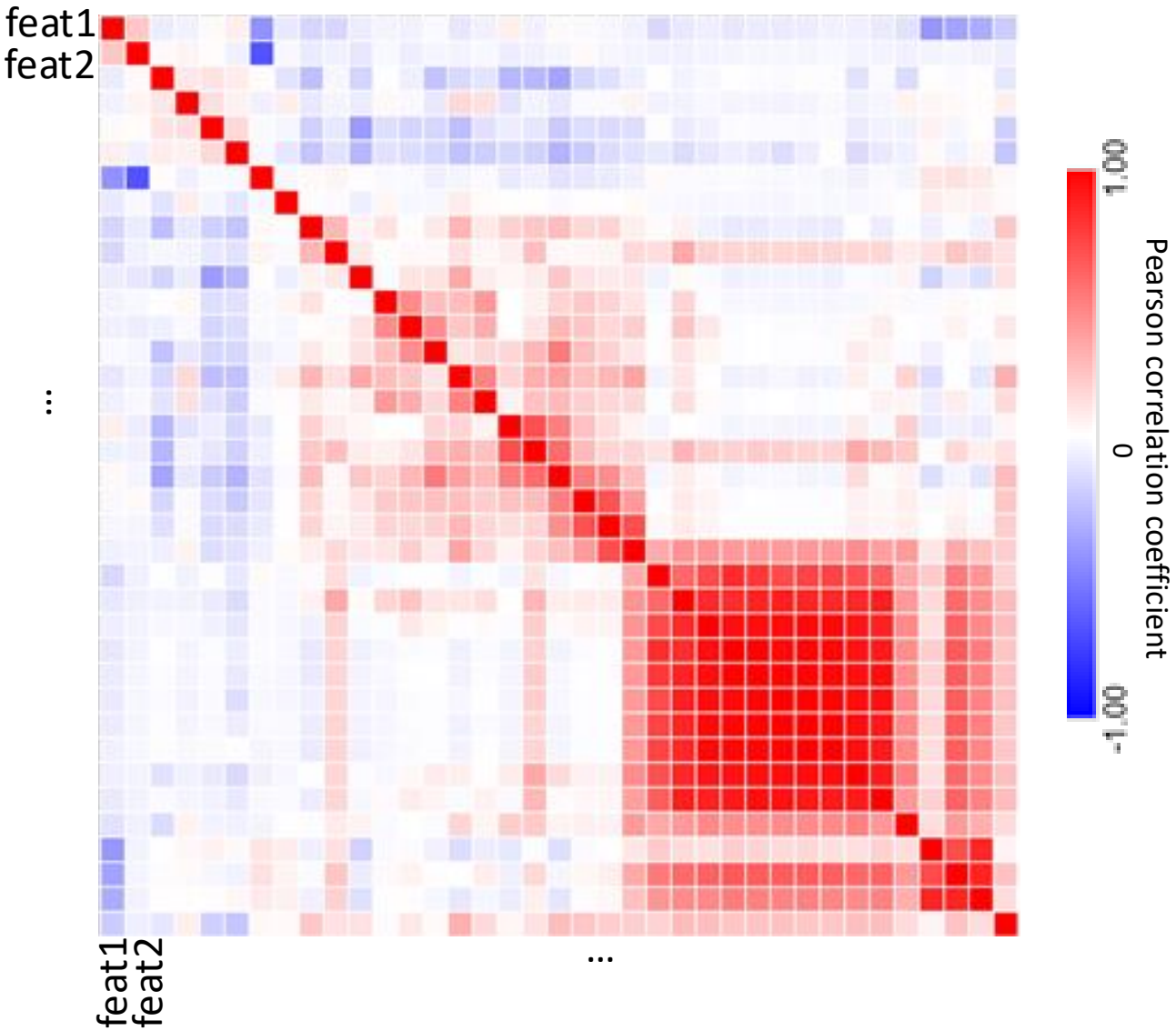
Pairplots to visualize pairwise relationships in a multidimensional dataset?

- Plotting a bivariate distribution in a figure
- Aligning the figures in a structured way
- Good to use if all the variables are equally important
- Better if used for exploration



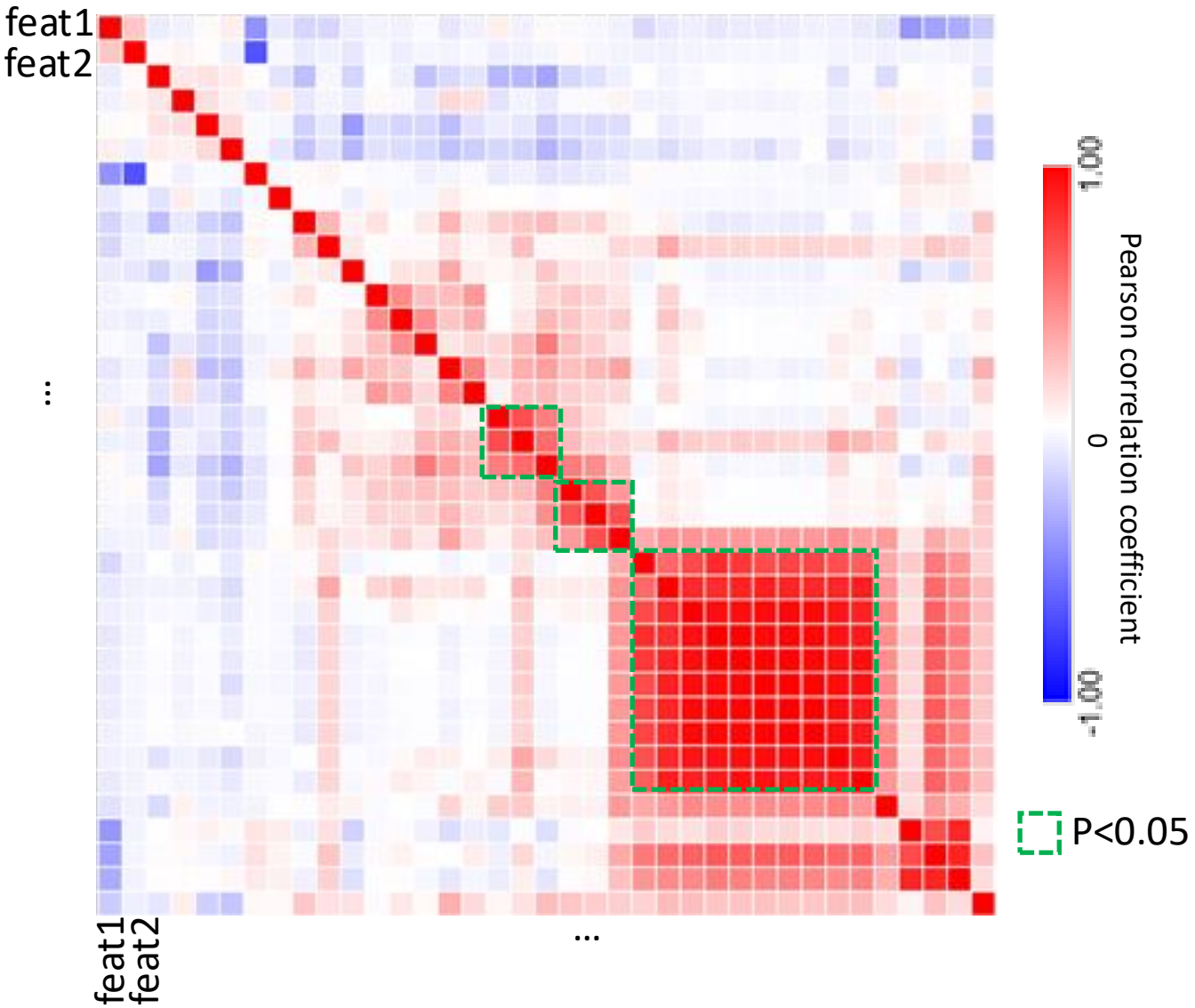
Heatmap for correlation across many variables

- Quantitative analyses use two measures, effect size and p-value, but a heatmap can show only either.
  Using Effect Size—or Why the *P* Value Is Not Enough, PMID: 23997866
- E. g., Pearson correlation coefficients between clinical features of patients in my collaboration project
- Another heatmap with p-value needs to be presented together for comprehensive understanding, but usually not used



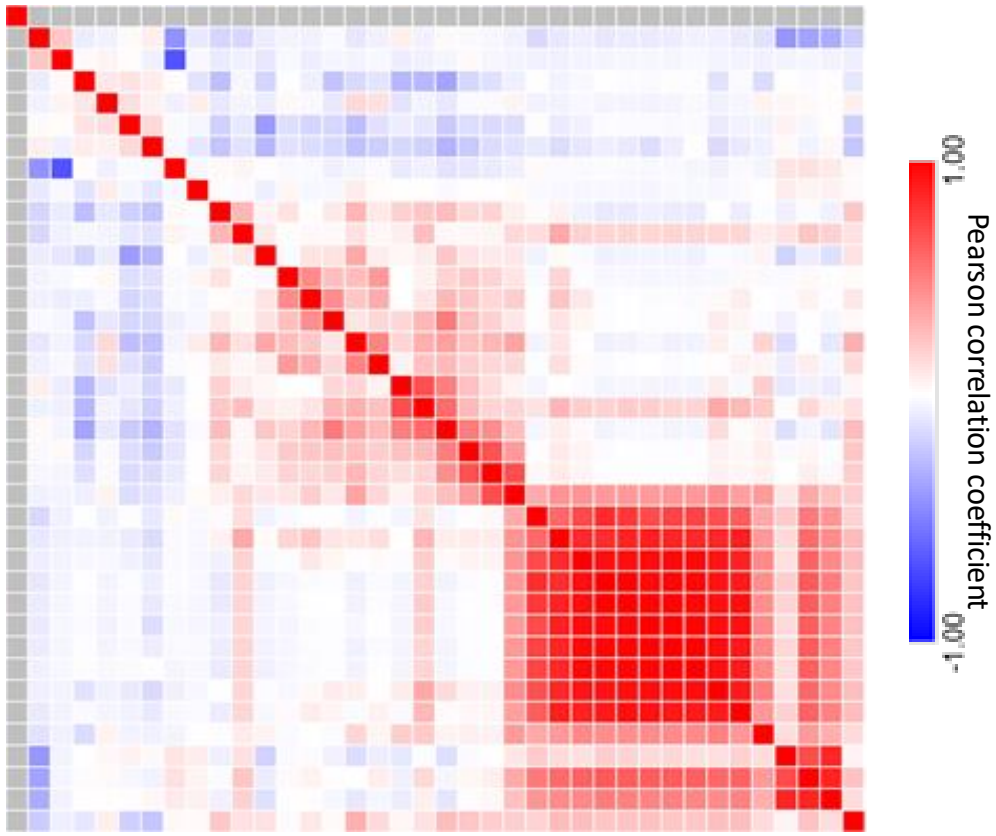
A heatmap shows a single measure

- Quantitative analyses use two measures, effect size and p-value, but a heatmap can show only either.
  Using Effect Size—or Why the *P* Value Is Not Enough, PMID: 23997866
- E. g., Pearson correlation coefficients between clinical features of patients in my collaboration project
- Another heatmap with p-value needs to be presented together for comprehensive understanding, but usually not used
- Instead indicate significance on top?

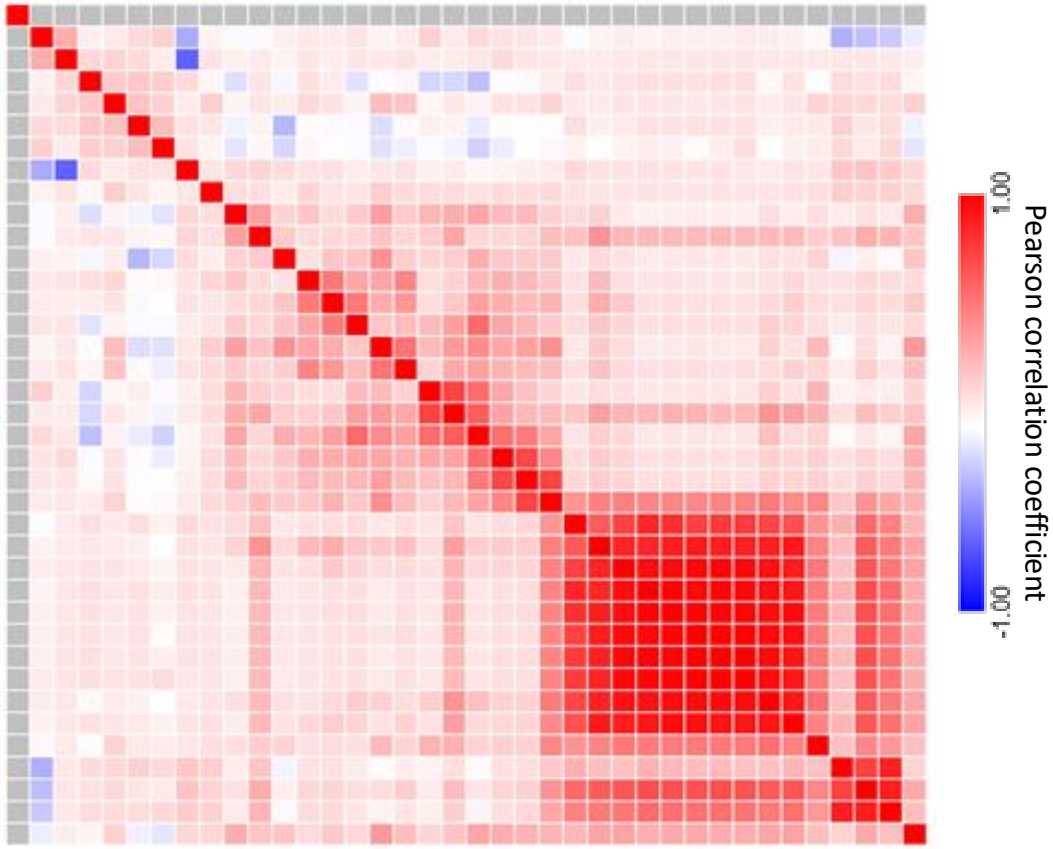


Heatmaps are sensitive to color assignment!

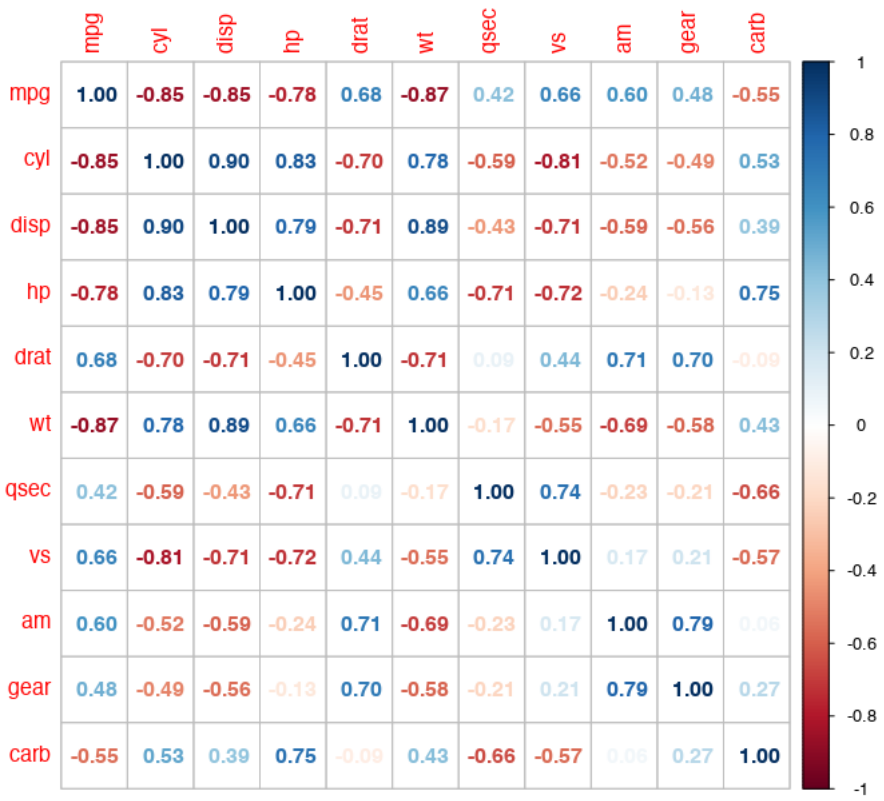
- Heatmaps on the same data can make different impressions

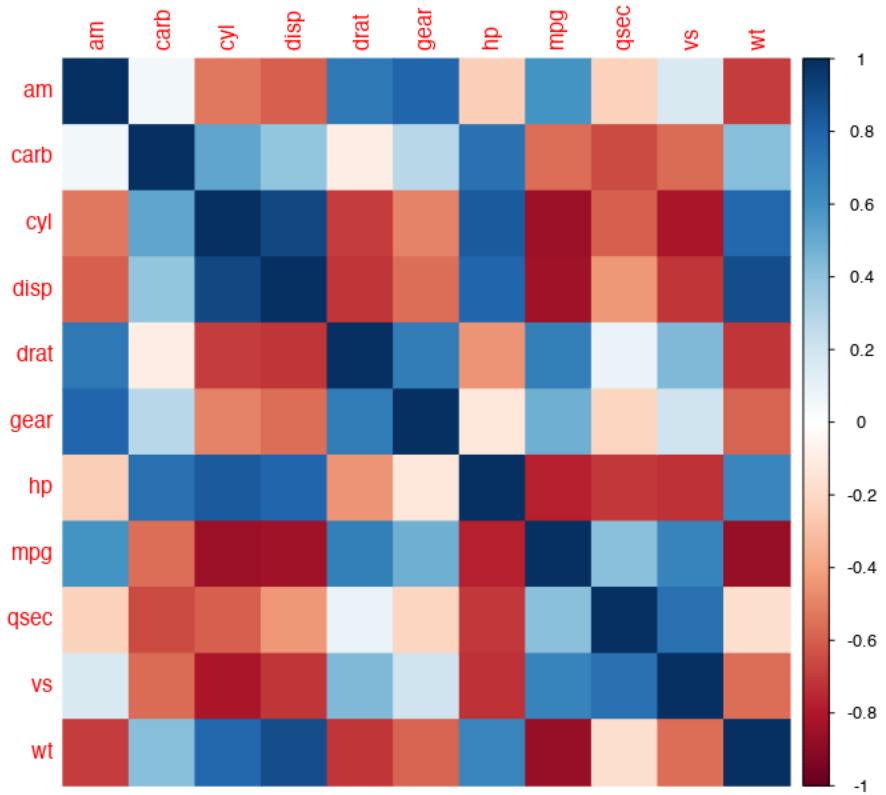


=



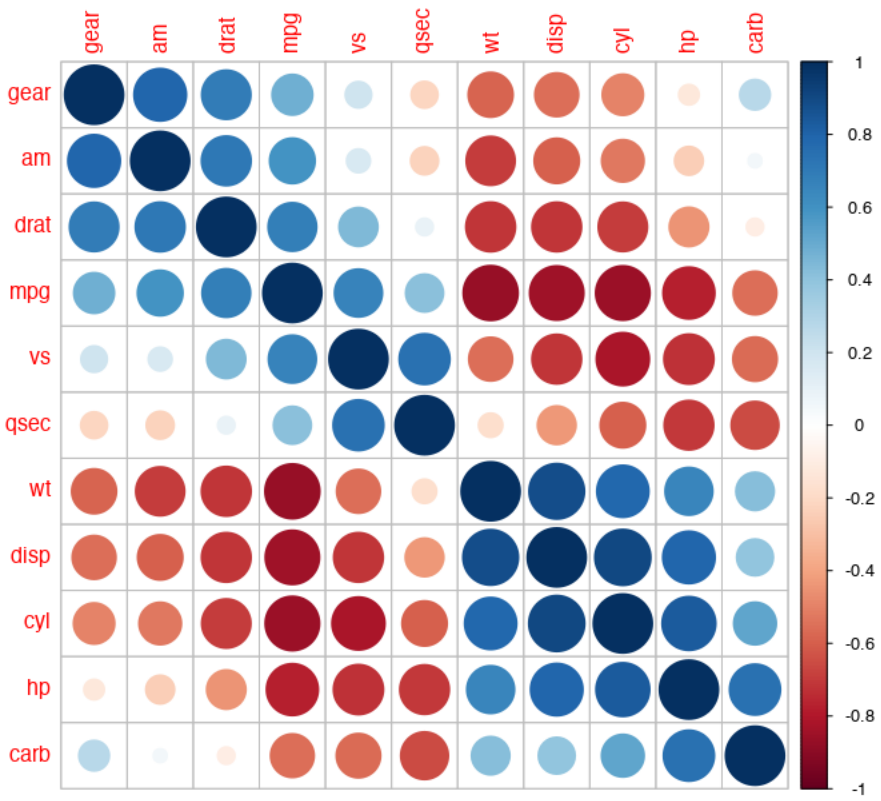
Other correlation heatmaps

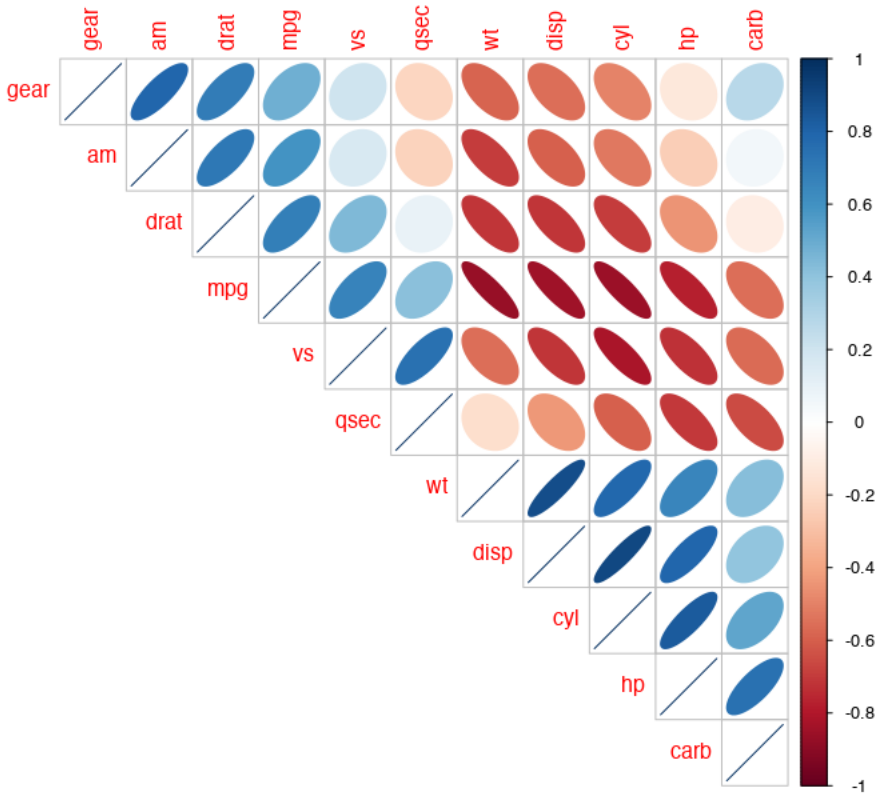




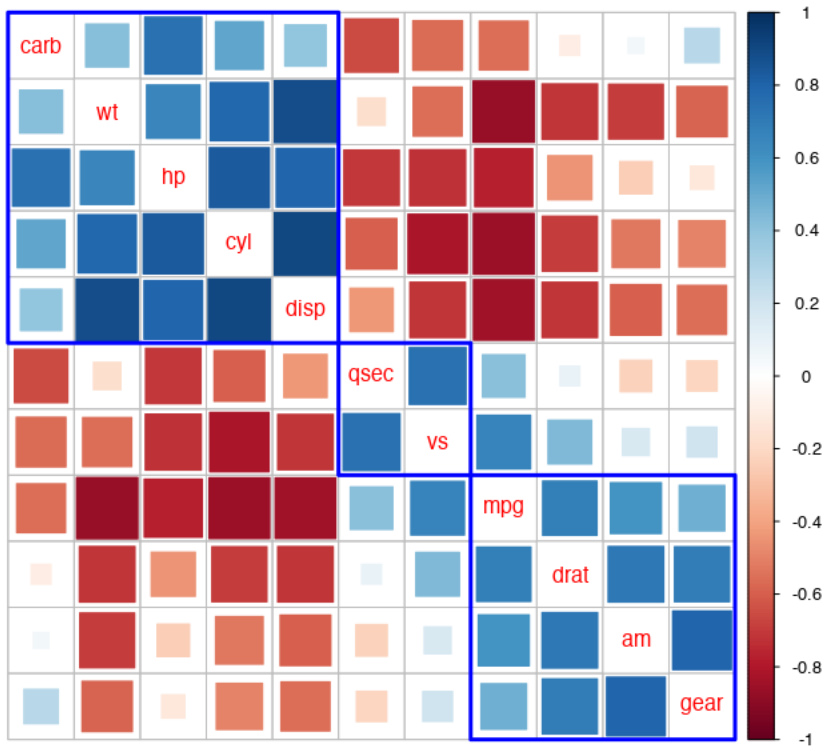
https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

Other correlation heatmaps





Other correlation heatmaps

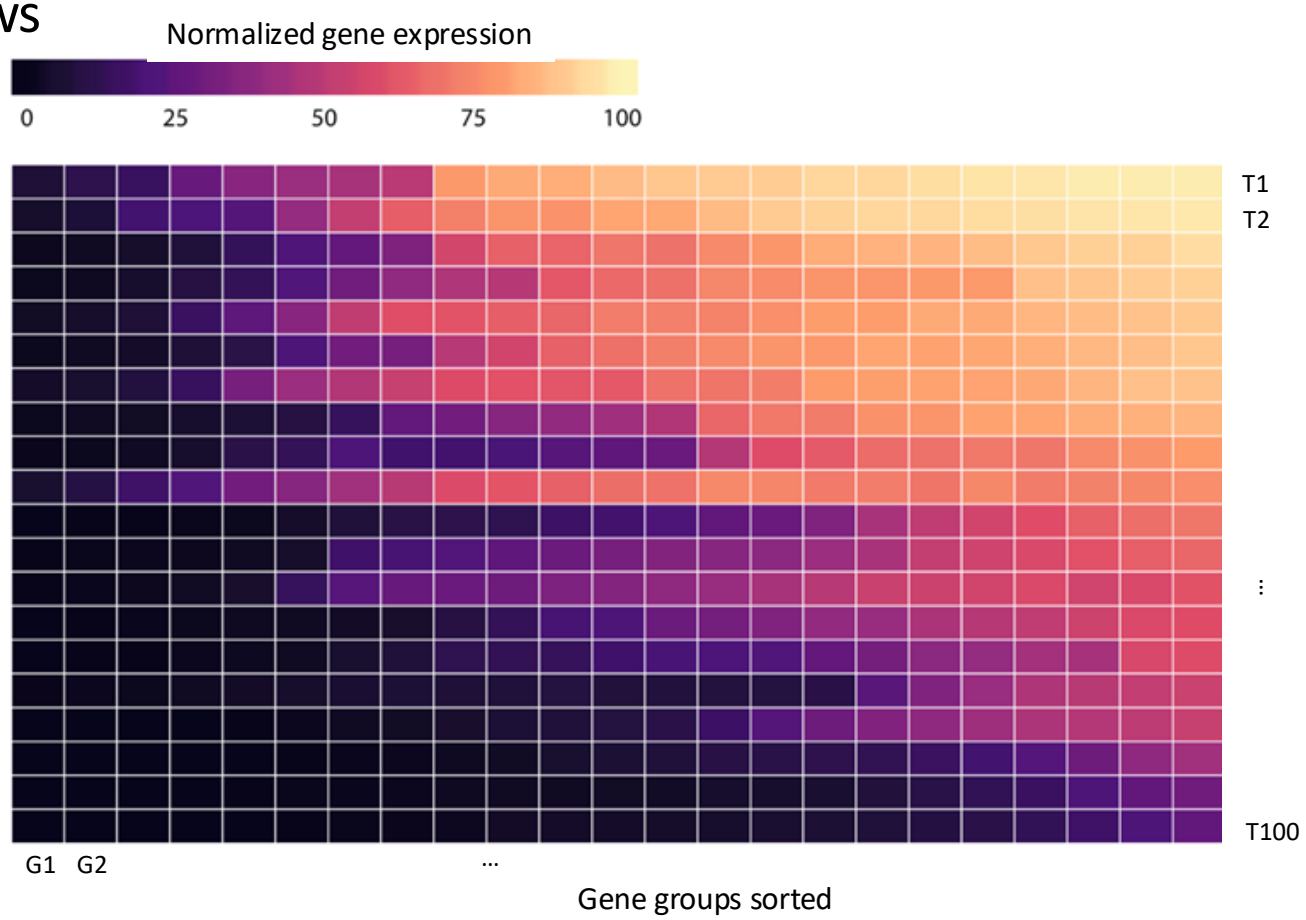


Heatmap can visualize trends/amounts in large datasets

Hypothetical example with sorted gene expressions in multiple breast tumor samples (T1, T2,…, T100)

Heatmaps map data values onto colors in rows (samples) and columns (variables)

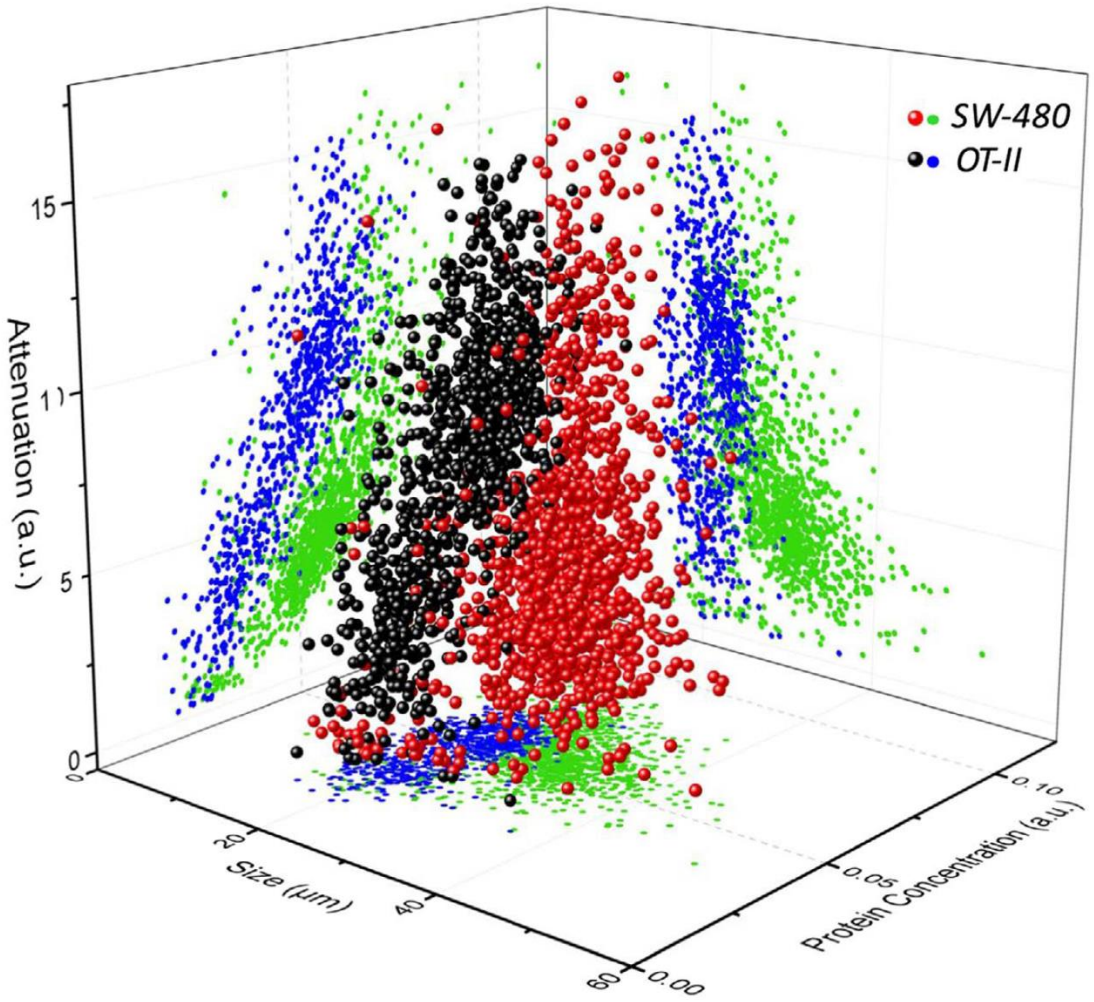
- Pros: highlighting broader trends
- Cons:
  - difficult to read the exact data values
  - shows only a single measure
  - sensitive to color assignment



Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Scatterplot to show landscape?

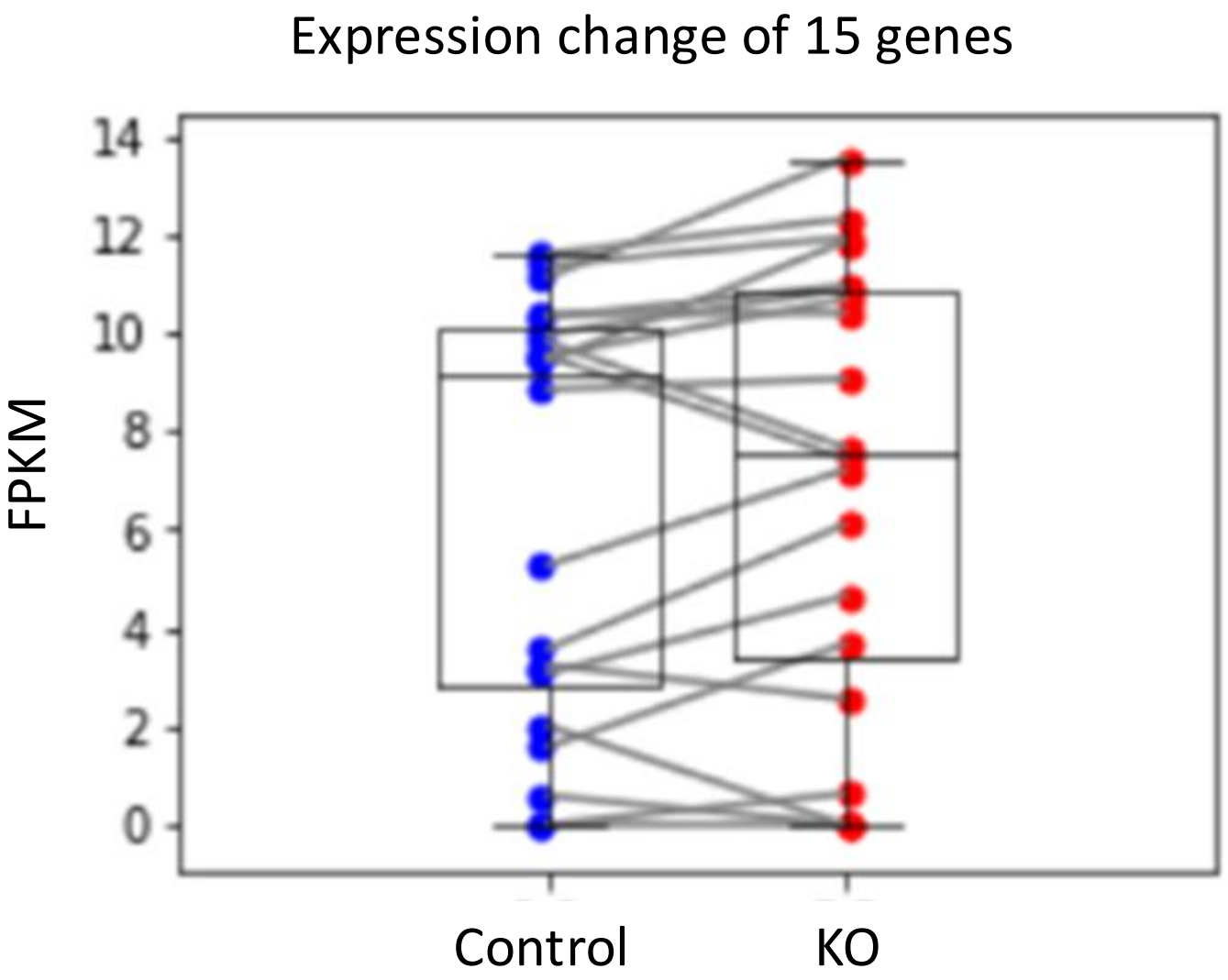
Deep Learning in Label-free Cell Classification



Chen et al. Sci. Reports 2016

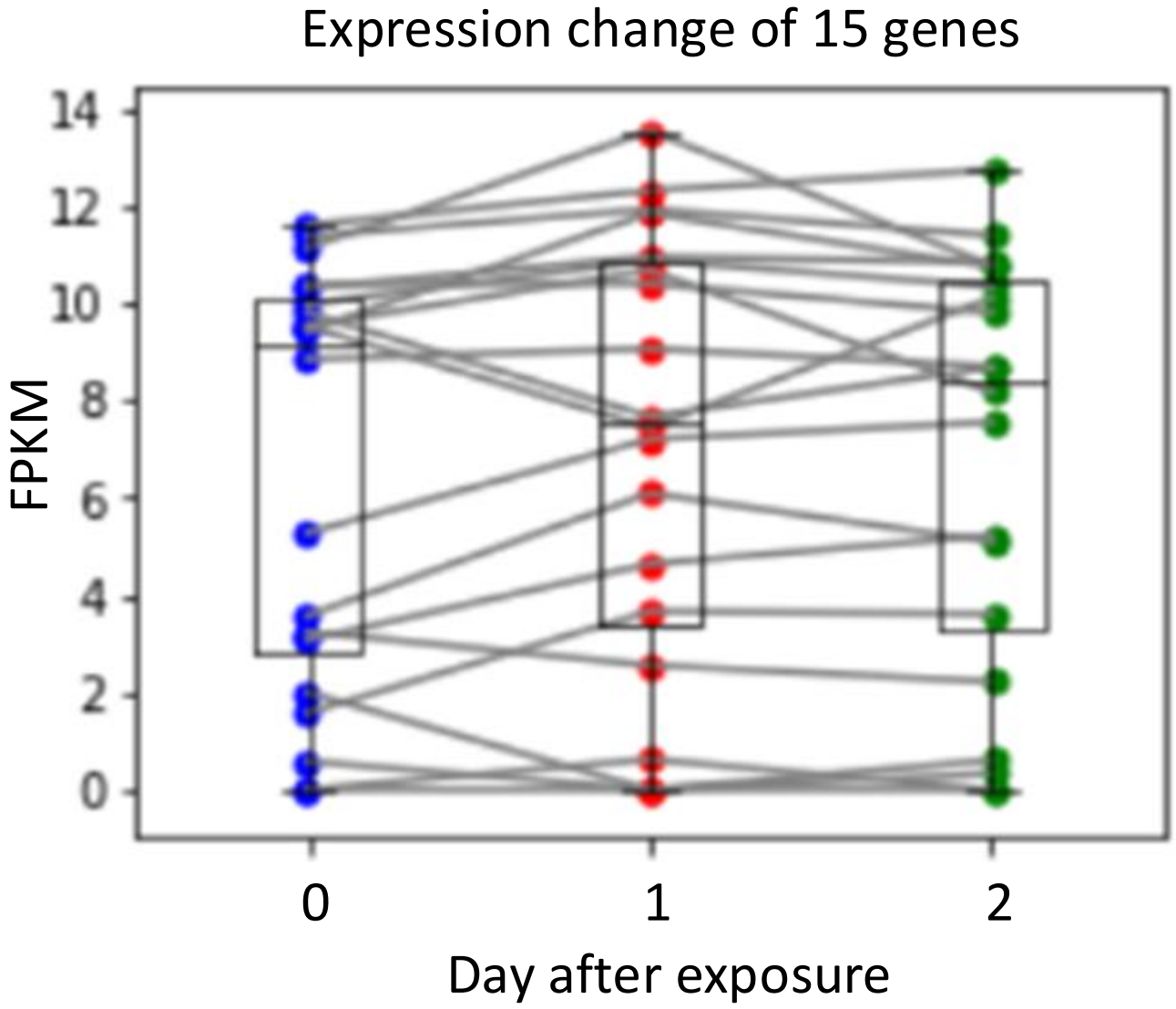
Slopegraph for changes over two variables

- Indicates individual-level measurements across time/conditions
- Good when interested in the identity/trajectory of each individual case
- Works well for a small number of observations



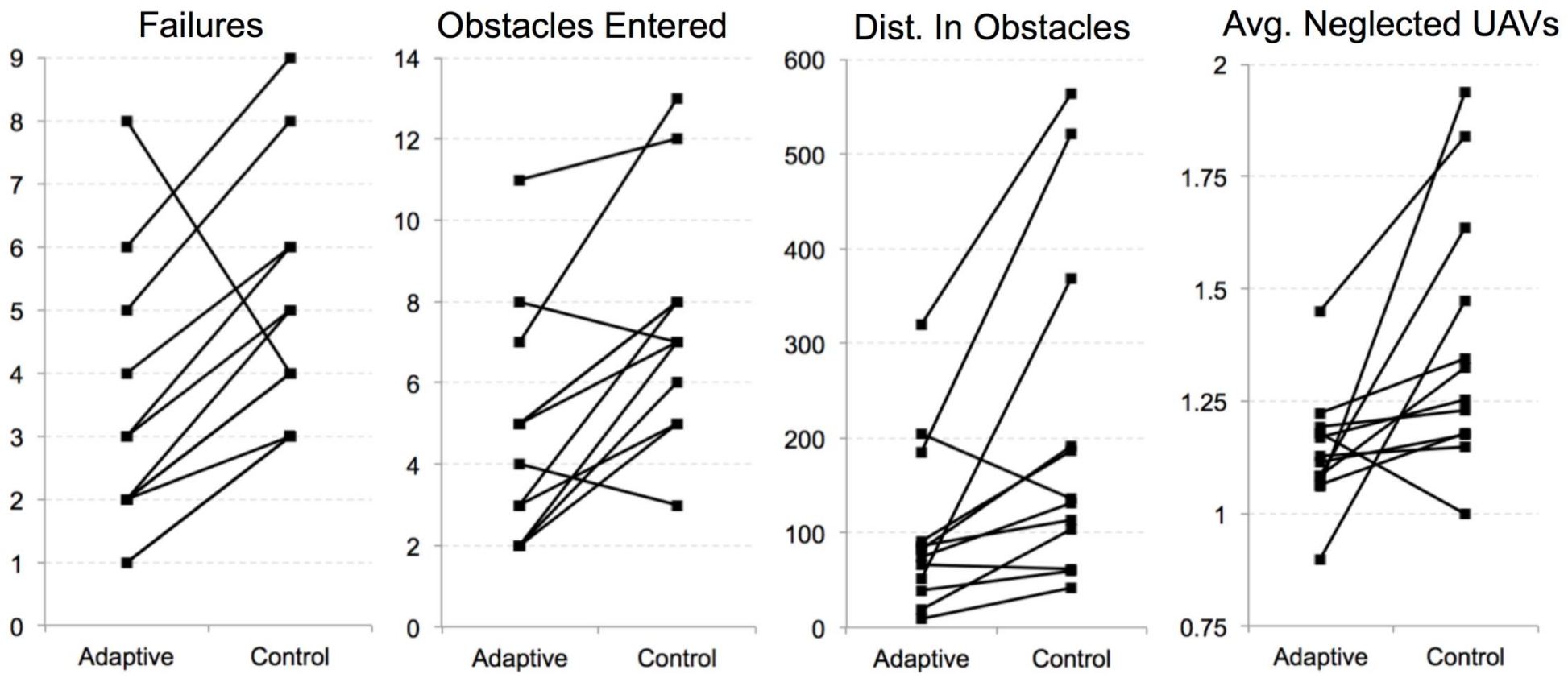
Slopegraph for changes over multiple variables

- assumes an order among the conditions
- Note the boxplots



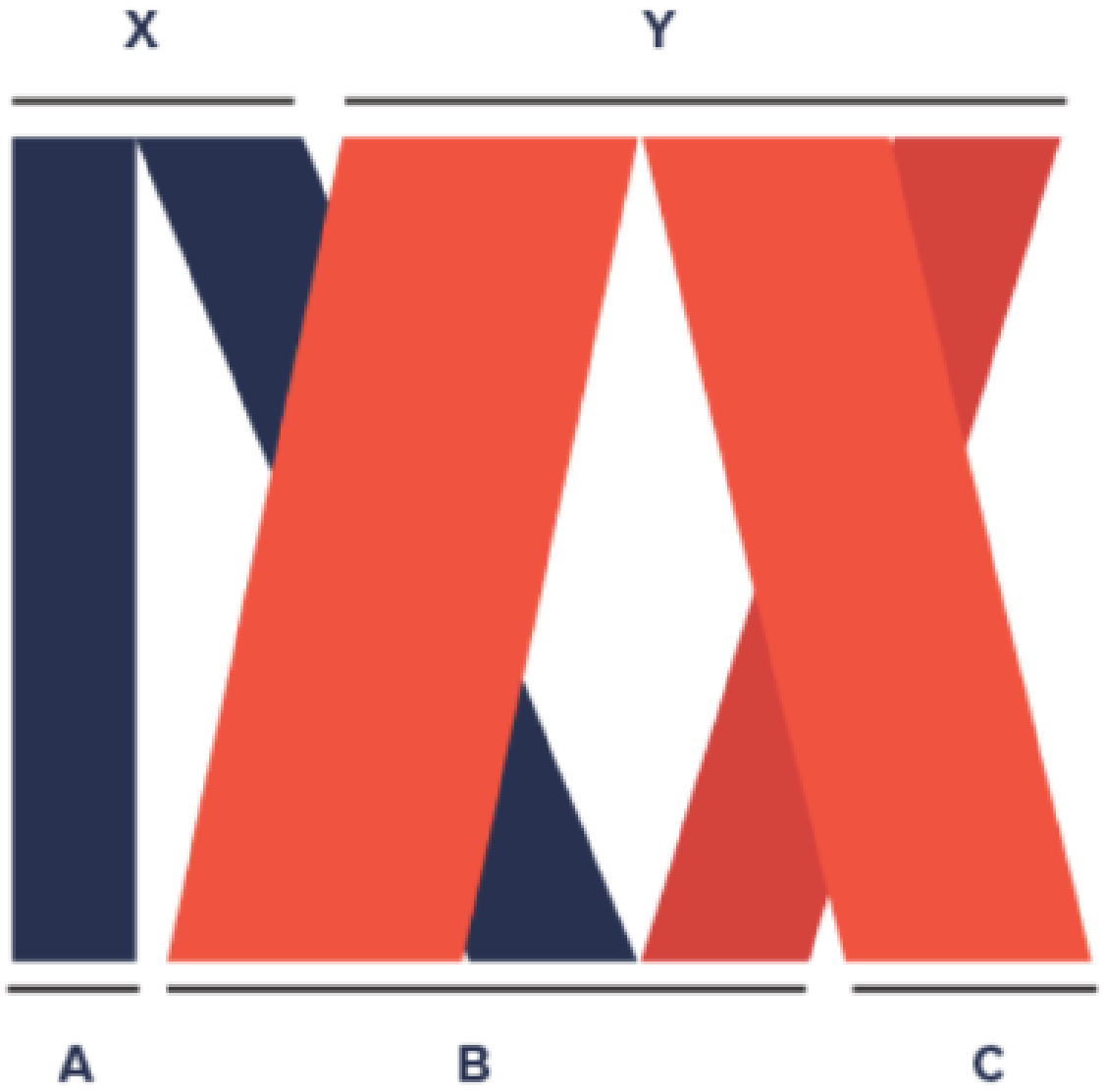
Slopegraph to compare

Dynamic Difficulty Using Brain Metrics of Workload

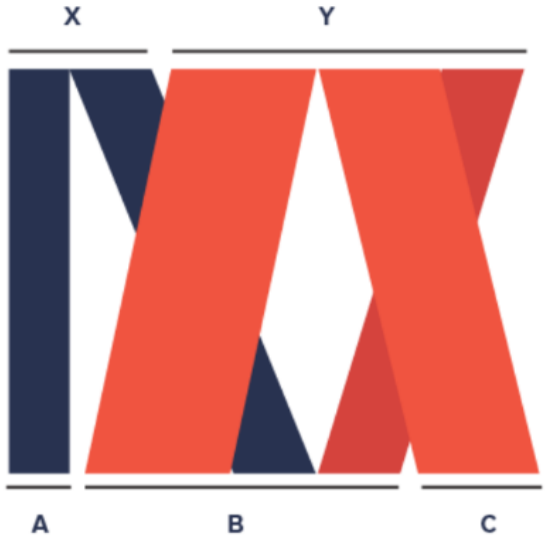


Afergan et al. Proceeding in CHI 2014

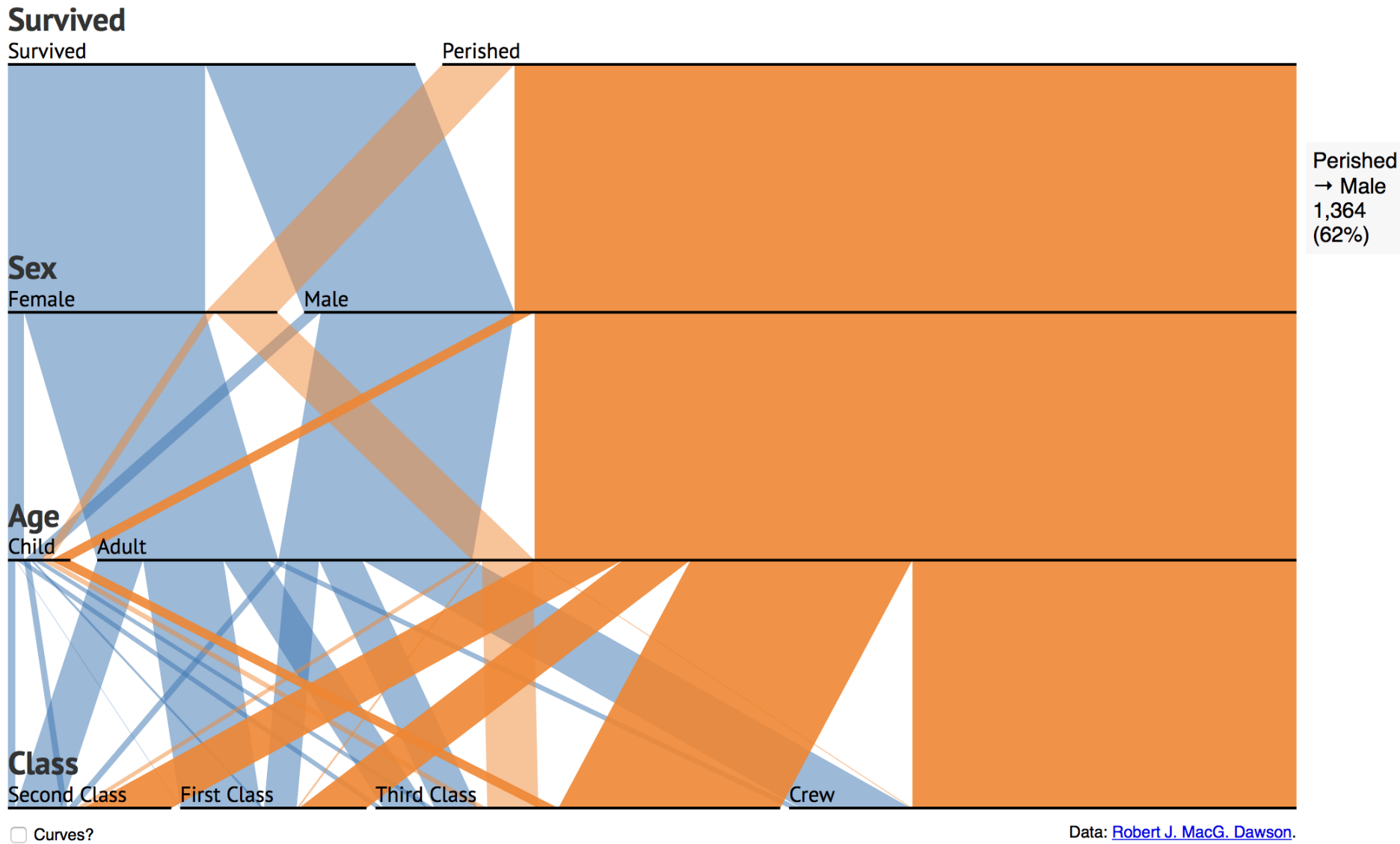
Parallel sets to visualize association between two sets of categorical variables



https://datavizproject.com/

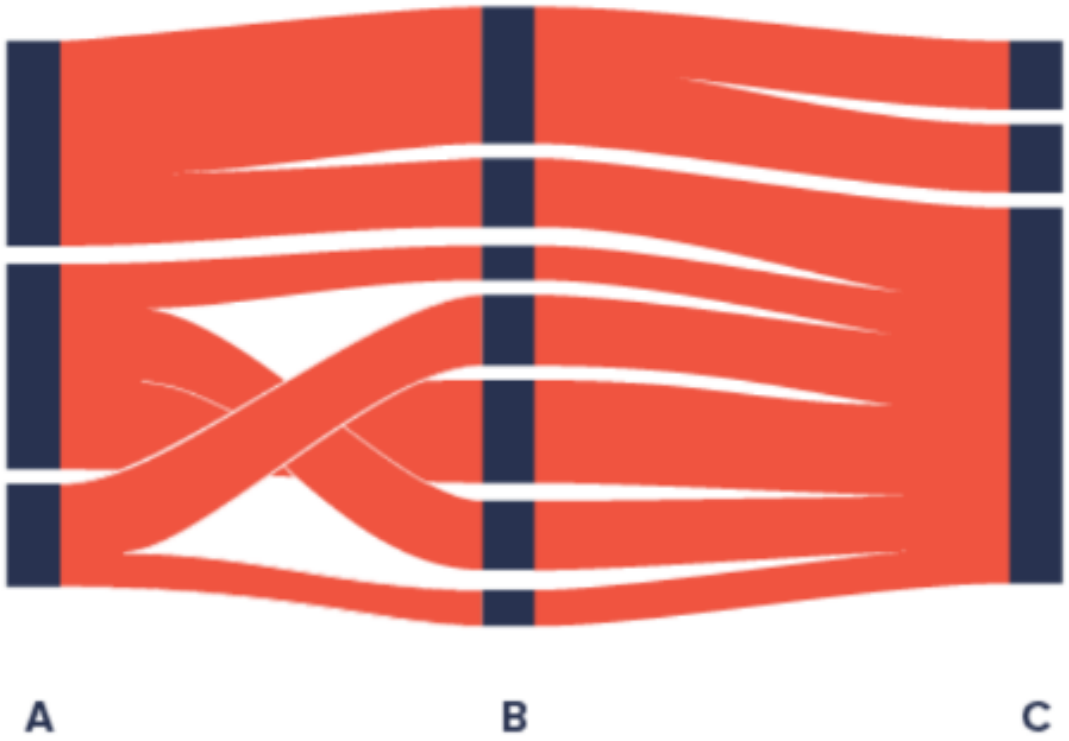


Titanic Survivors



Alluvial for changes over multiple categorical variables

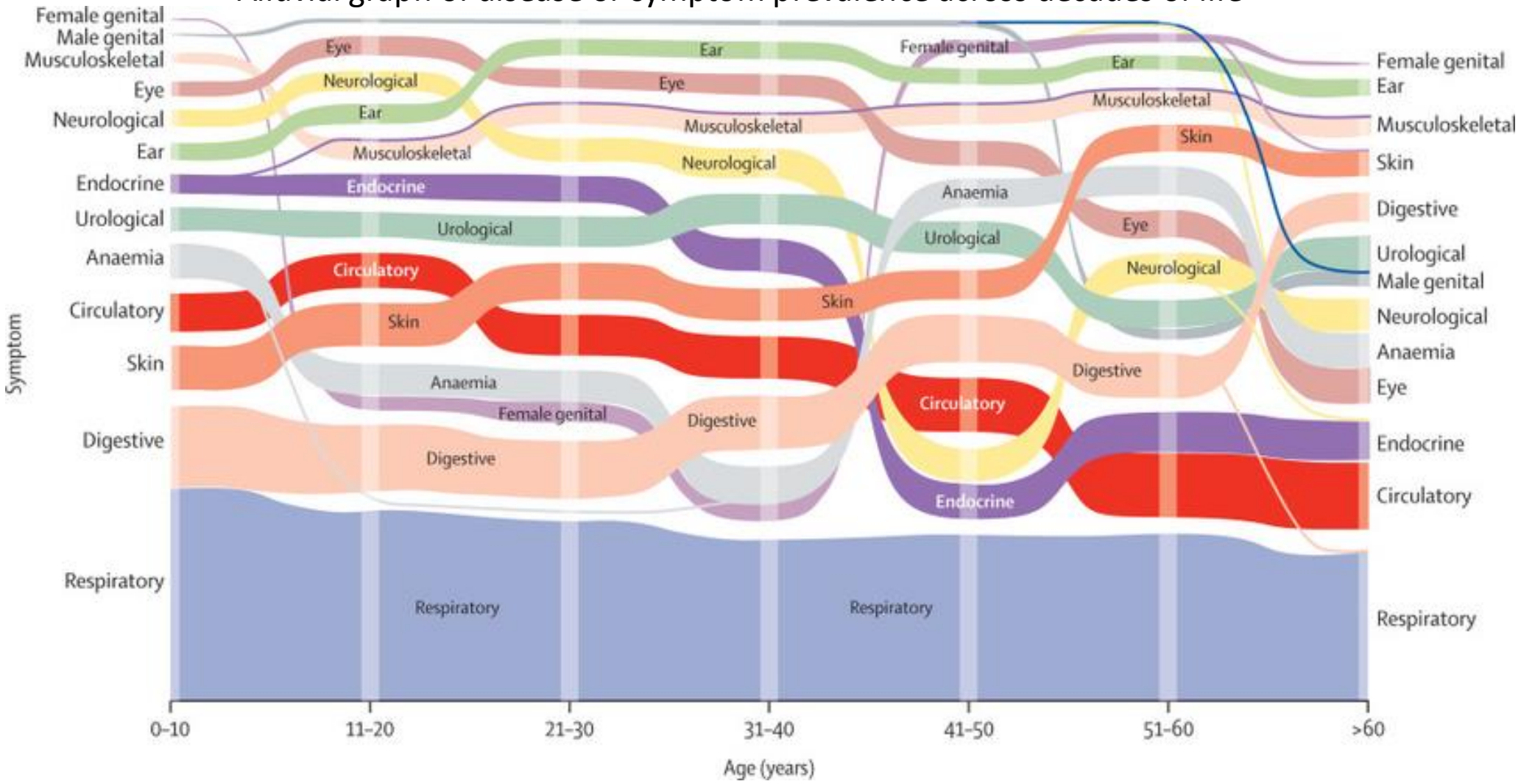
https://geographyeducation.org/2017/12/21/alluvial-fans-2/



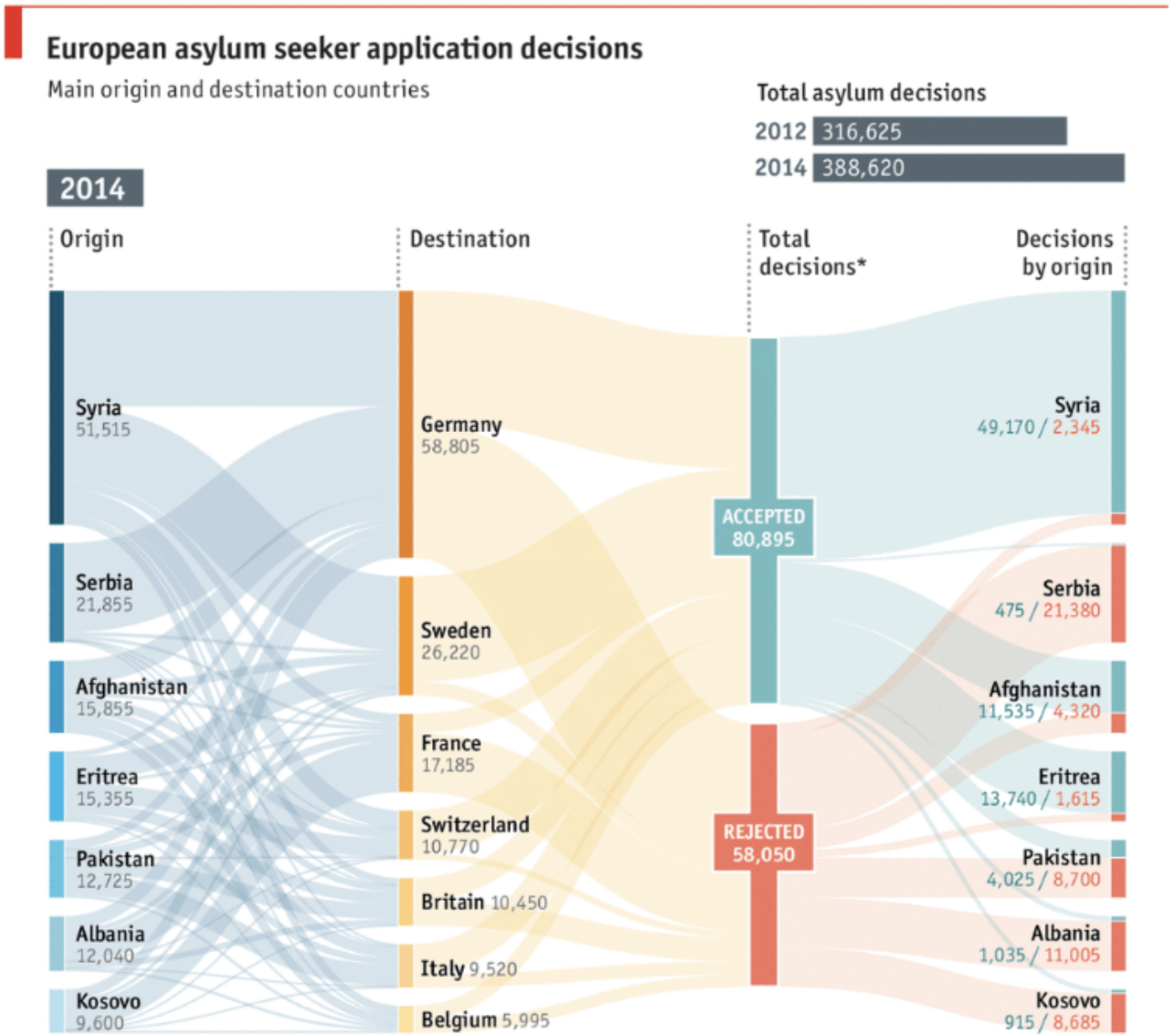
https://datavizproject.com/

Alluvial for changes over multiple categorical variables

Alluvial graph of disease or symptom prevalence across decades of life



Salvi et al. 10.1016/S2214-109X(15)00152-7



https://datavizcatalogue.com/blog/sankey-diagrams-parallel-sets-alluvial-diagrams-whats-the-difference/