

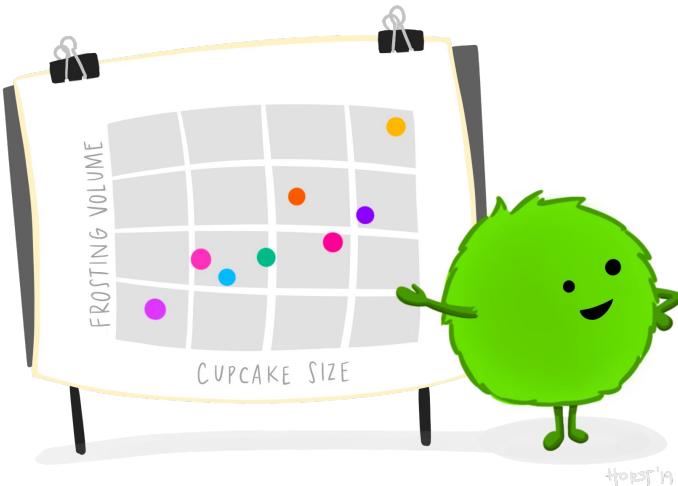
# PRIMER 3

## Presenting data & results in a computational biology project

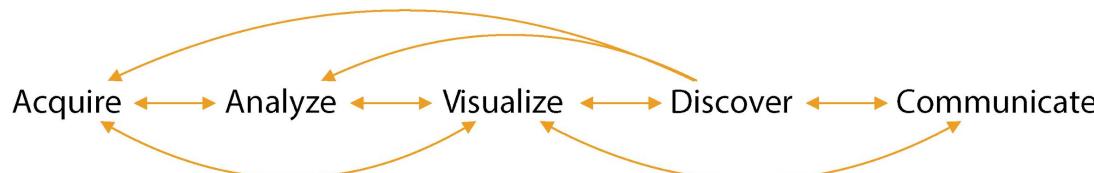
- The purpose of data visualization
- Choice of visualization
- Typical issues in aesthetics & data handling
- Improving visualization for clarity
- How to write a paper

# Purpose of data visualization

- Illustrate important findings.



- A critical component of your analysis and research!
- Allow the reader to confirm that the statistical analysis is appropriate for the study design.
- Allow the reader to critically evaluate the data.



# Choice of visualization

- Choosing among many options
- Barplots
- Boxplots



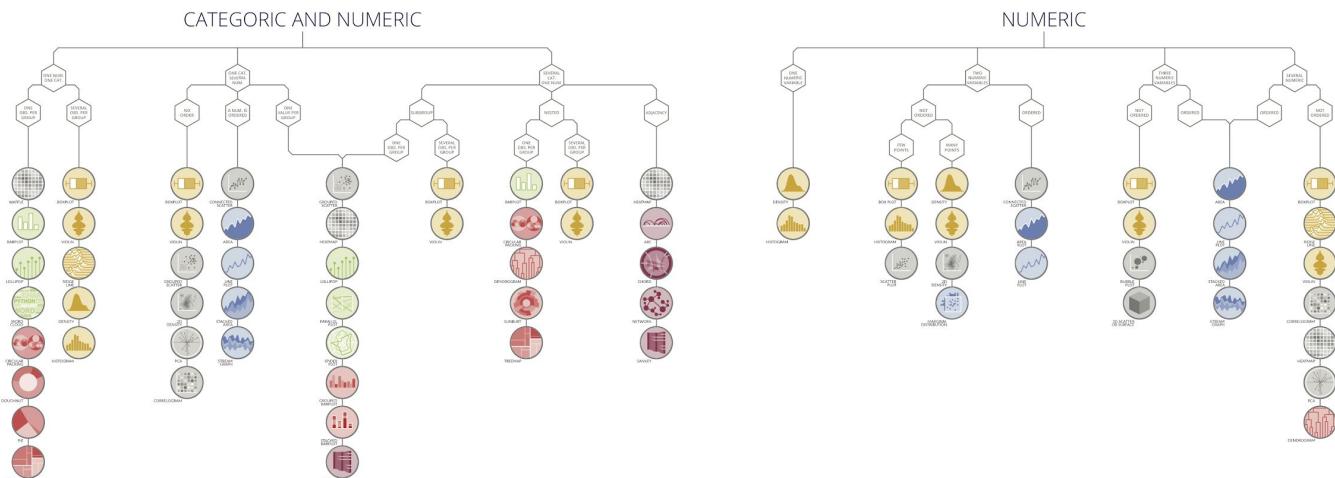
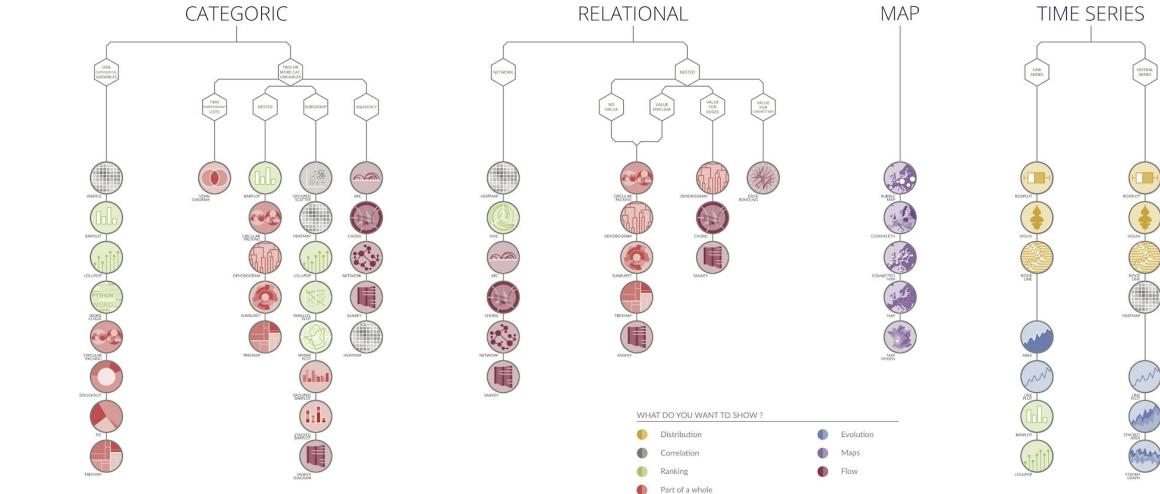
# from Data to Viz

**'From Data to Viz'** is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

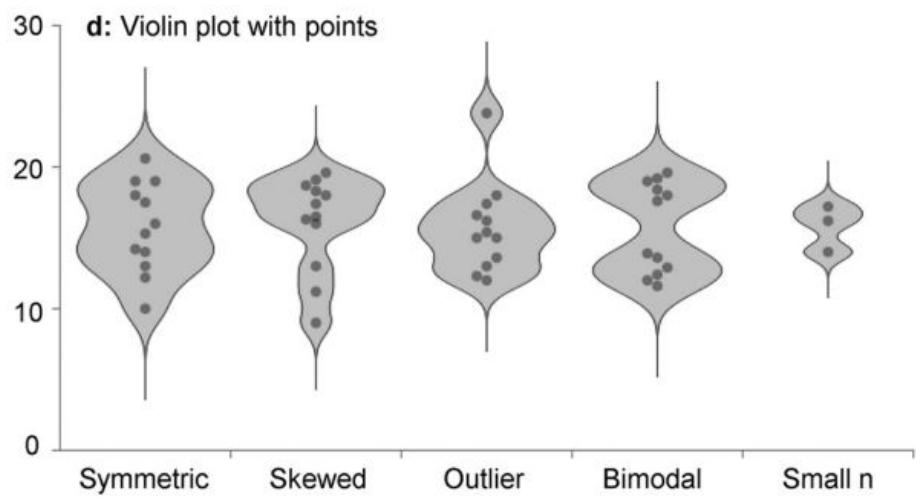
- 1 Identify what type of data you have.
  - 2 Go to the corresponding decision tree and follow it down to a set of possible charts.
  - 3 Choose the chart from the set that will suit your data and your needs best.

Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

[data-to-viz.com](http://data-to-viz.com)

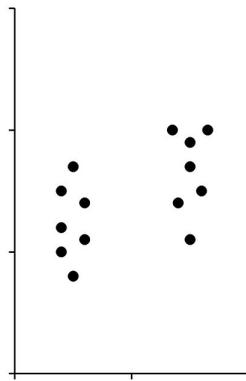


# Bar plots w/ error bars – Almost never a good idea

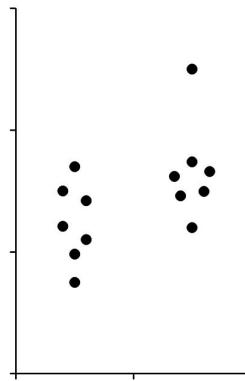


# Bar plots w/ error bars – Almost never a good idea

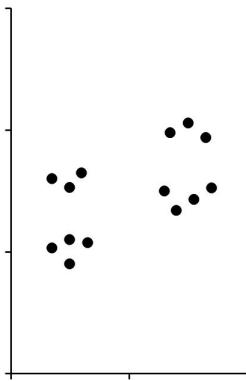
Symmetric



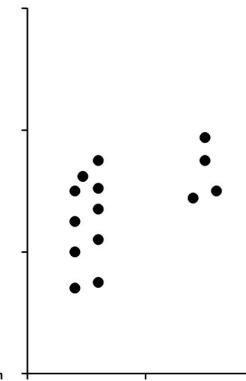
Outlier



Bimodal



Unequal n



Test

p value

	T-test: Equal var.	T-test: Unequal var.	Wilcoxon	
T-test: Equal var.	0.035	0.050	0.054	0.026
T-test: Unequal var.	0.035	0.050		0.026
Wilcoxon		0.073		0.128

Parametric vs.  
Non-parametric test

differences in transfection efficiency. Each data point is the mean of triplicate samples  $\pm$  the standard error; the data presented are repre-

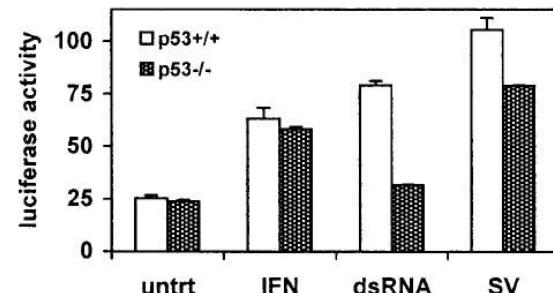
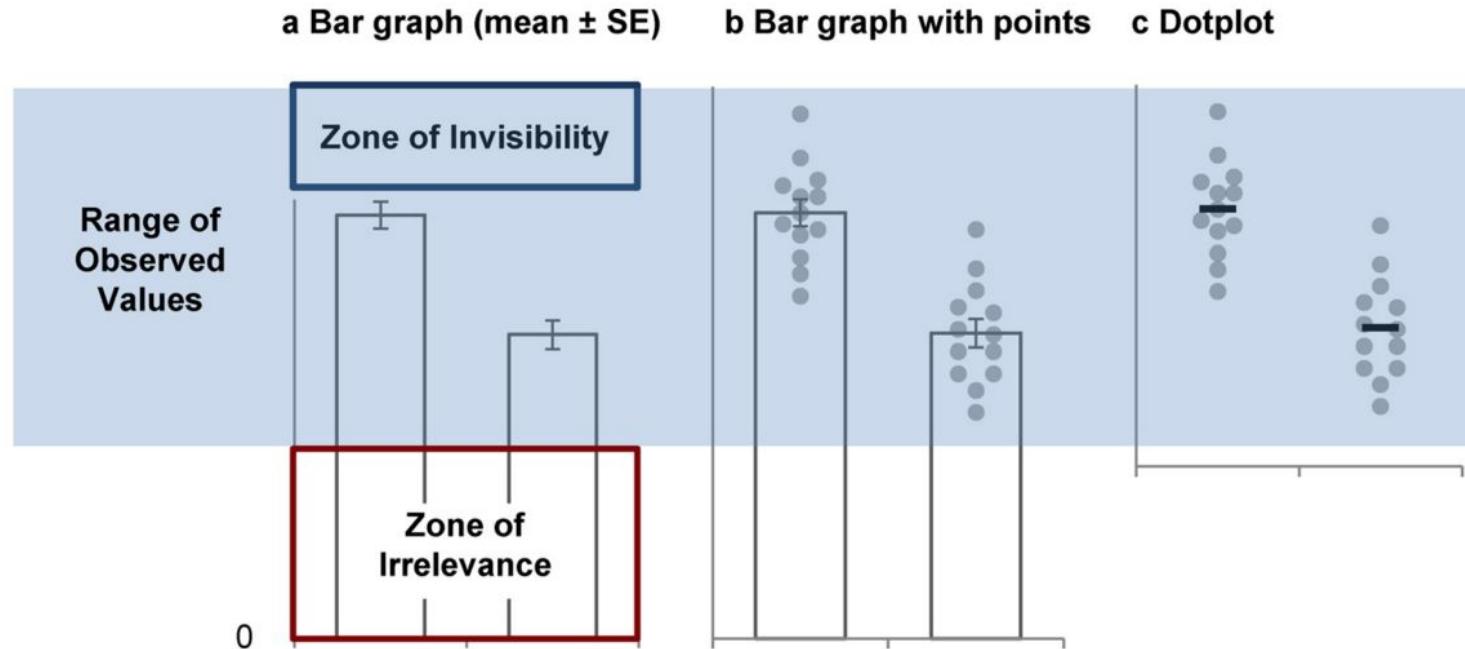
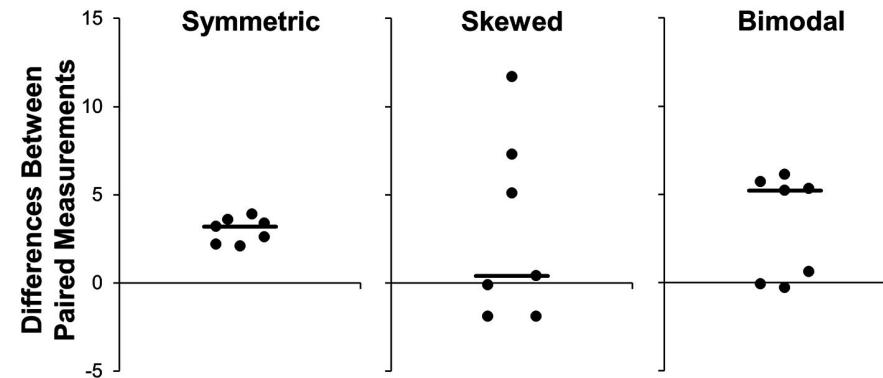
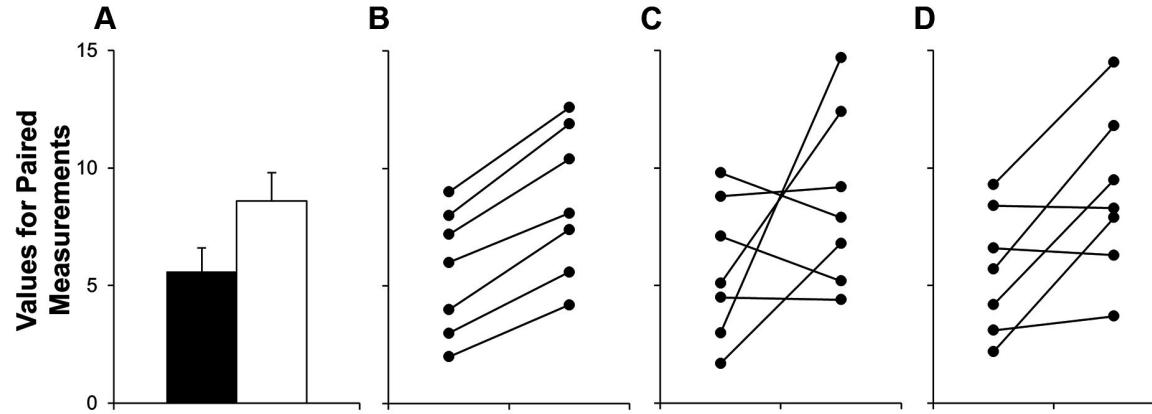


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells ( $6 \times 10^5$  HCT 116) were

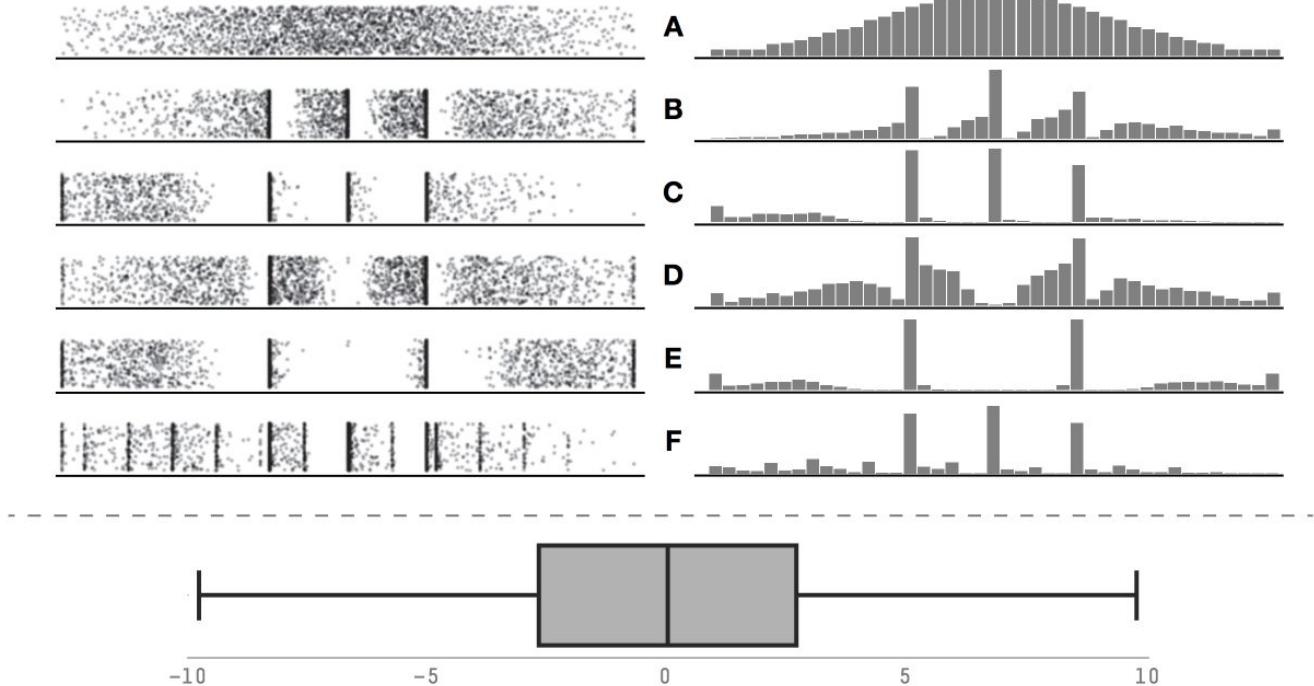
# Bar plots w/ error bars – Almost never a good idea



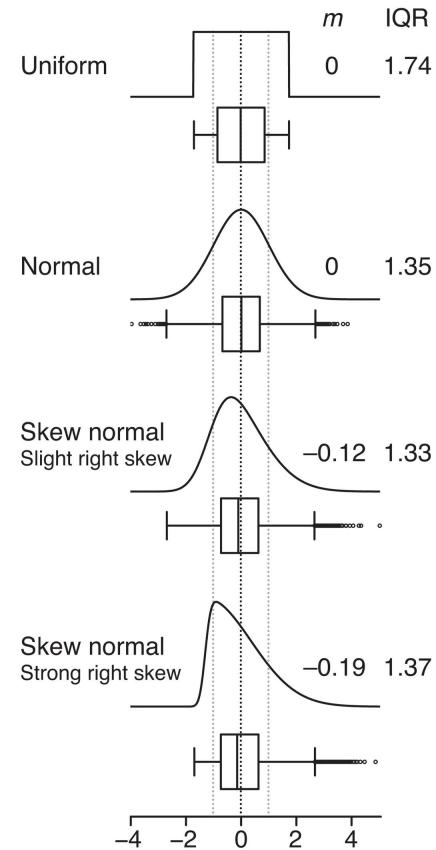
# Bar plots w/ error bars – Almost never a good idea



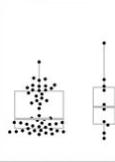
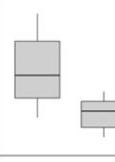
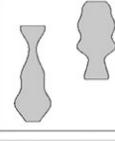
# Boxplots can confound different distributions



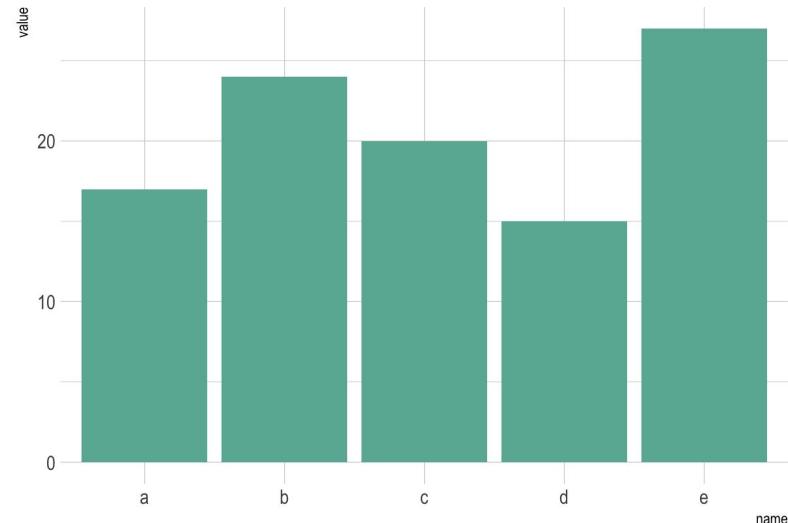
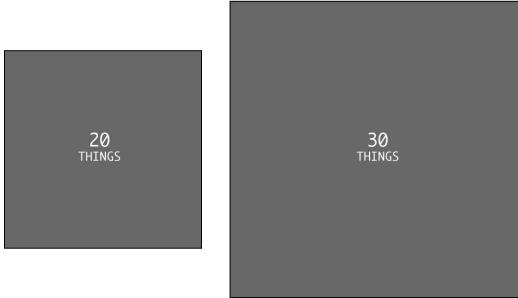
# Boxplots can confound different distributions



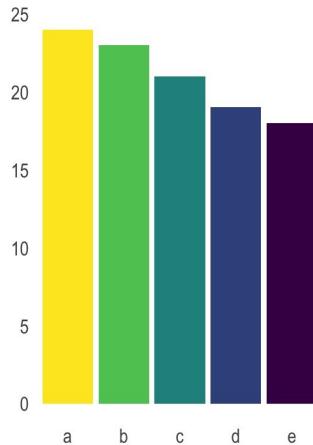
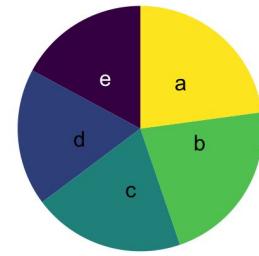
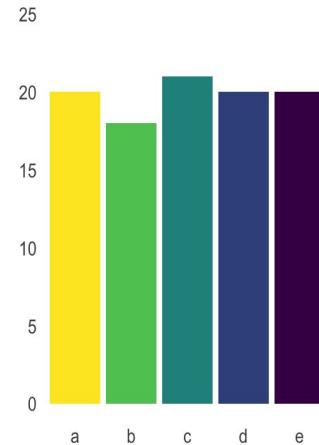
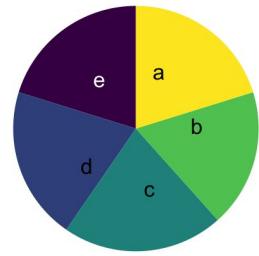
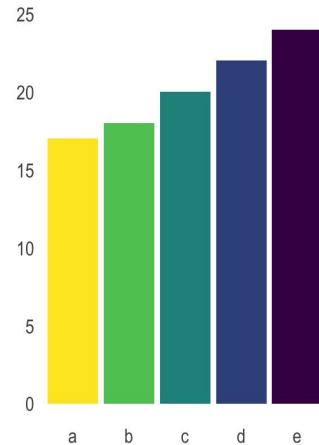
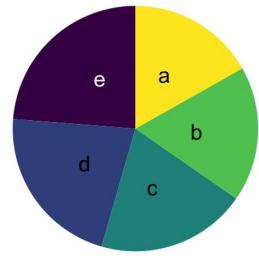
# Pick the plot to show different distributions

Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	<ul style="list-style-type: none"> <li>Make all data points visible - use symmetric jittering</li> <li>Many groups: Increase white space between groups, emphasize summary statistics &amp; de-emphasize points</li> <li>Only add error bars if the sample size is large enough to avoid creating a false sense of certainty</li> <li>Avoid "histograms with dots"</li> </ul>
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	<ul style="list-style-type: none"> <li>Make all data points visible (symmetric jittering)</li> <li>Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n</li> <li>Larger n: Emphasize box plot and de-emphasize points</li> </ul>
Box plot		Continuous	Horizontal lines on box: 75 <sup>th</sup> , 50 <sup>th</sup> (median) and 25 <sup>th</sup> percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	<ul style="list-style-type: none"> <li>List sample size below group name on x-axis</li> <li>Specify what whiskers represent in legend</li> </ul>
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	<ul style="list-style-type: none"> <li>List sample size below group name on x-axis</li> <li>The violin plot should not include biologically impossible values</li> </ul>
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	<ul style="list-style-type: none"> <li><b>Do not use for continuous data</b></li> </ul>

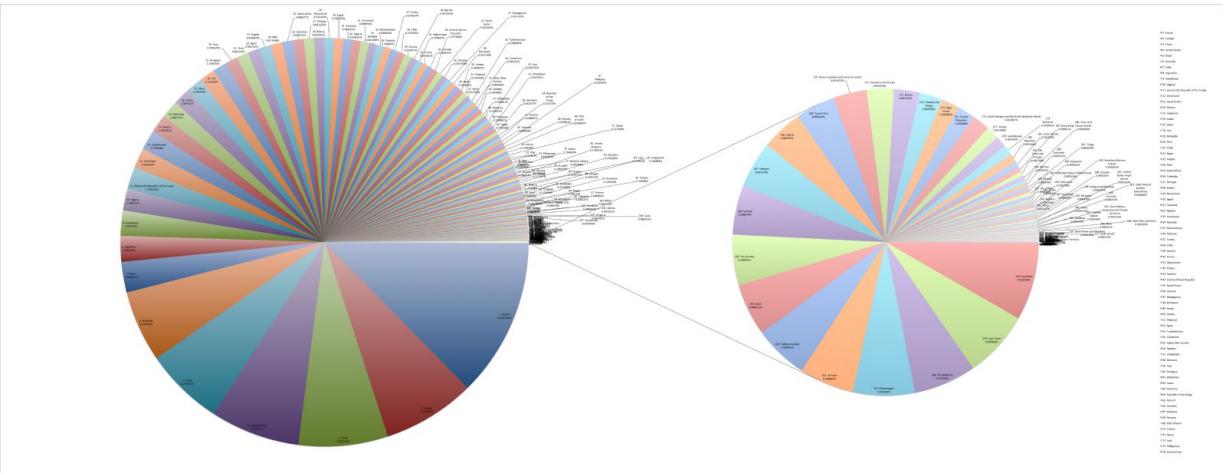
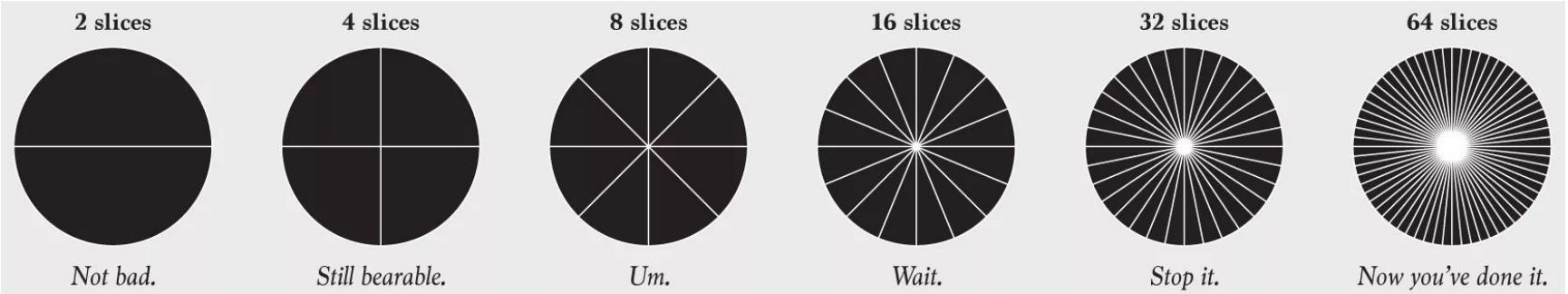
# Area is a poor choice for dimension



# Pie charts – almost never a good idea



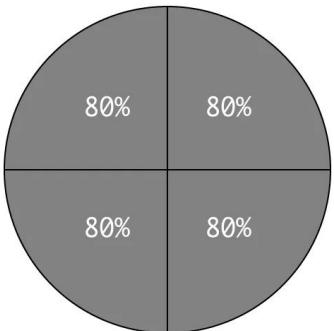
# Pie charts – almost never a good idea



<https://flowingdata.com/2015/08/11/real-chart-rules-to-follow/>

[https://commons.wikimedia.org/wiki/File:Pie\\_chart\\_of\\_countries\\_by\\_area.png](https://commons.wikimedia.org/wiki/File:Pie_chart_of_countries_by_area.png)

# ... especially when things don't add up!

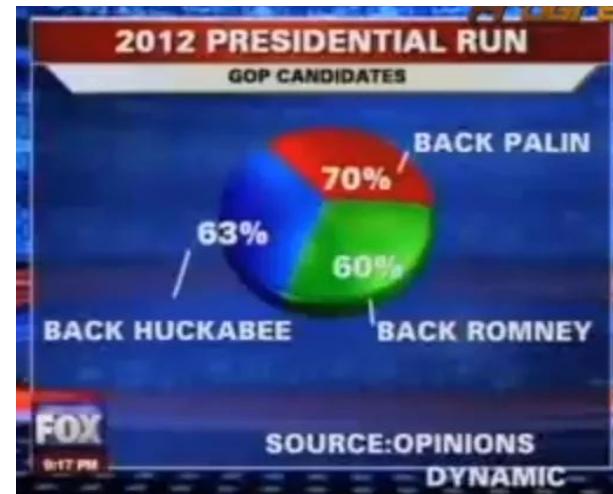
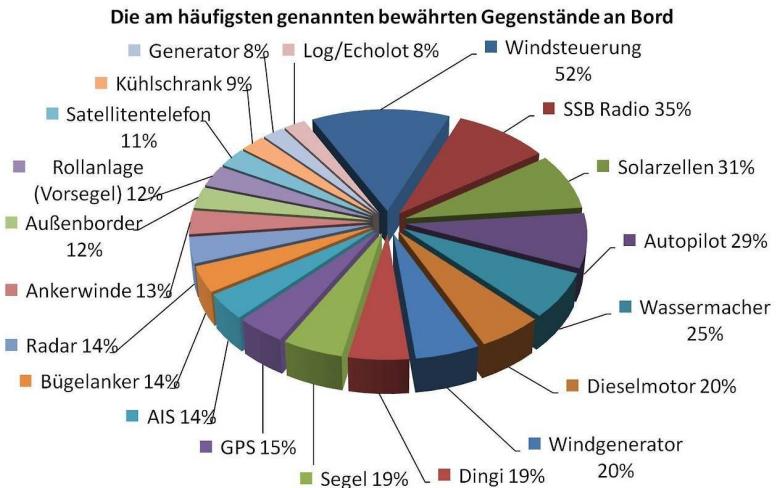


think with Google

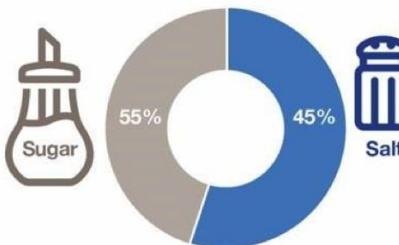
41 %

of people who own a voice-activated speaker say it feels like talking to a friend or another person.

Source: Google/Peerless Insights, "Voice & Voice-Activated Speakers: People's Lives Are Changing," n=1,642 U.S. voice-activated speaker owners who use their device monthly, A18+, Aug. 2017.

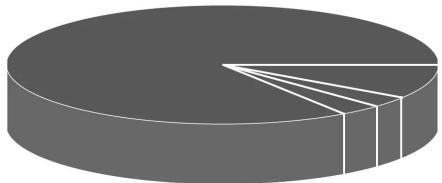


Last Week's Results  
Which of these would you have a harder time giving up, salt or sugar?

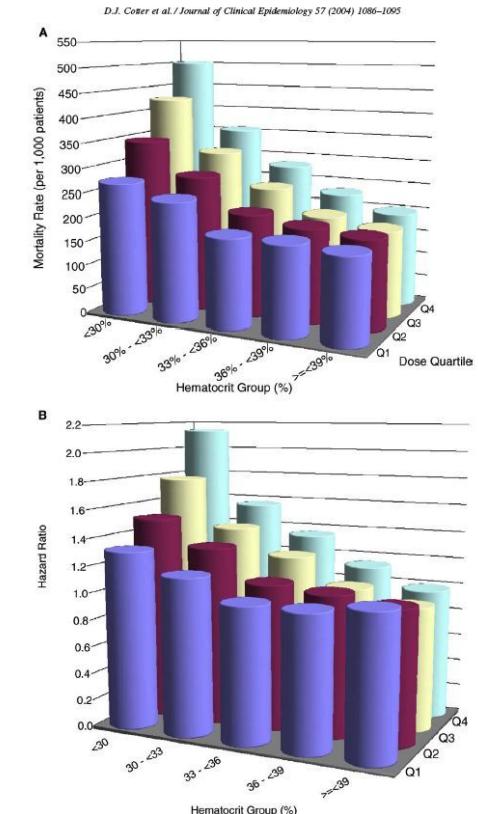
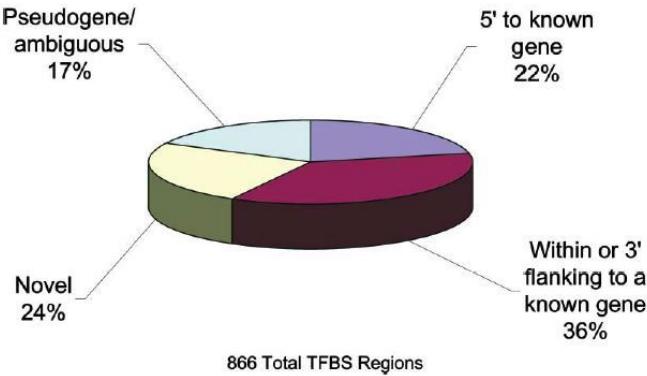


<https://www.data-to-viz.com/caveat/>  
<https://flowingdata.com/2017/02/09/how-to-spot-visualizations-lies/>

# No 3D!



## Distribution of All TFBS Regions

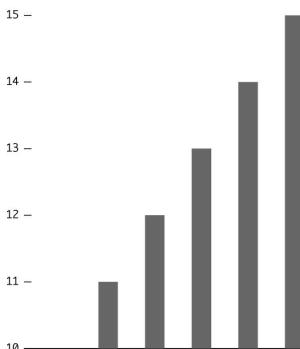


# Typical issues in aesthetics & data handling

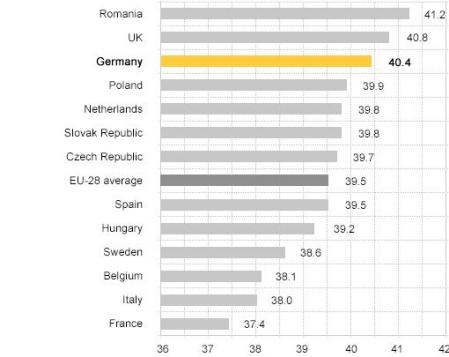
- Plot axes
- Data scales
- Data scope
- Data binning

# Avoid truncated axis

The value axis starts at ten. Liar, liar, pants on fire.

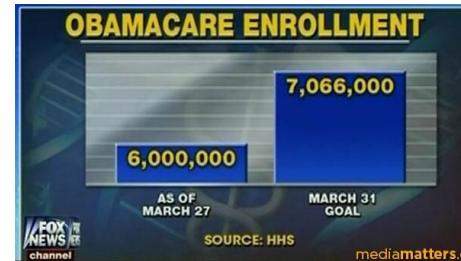
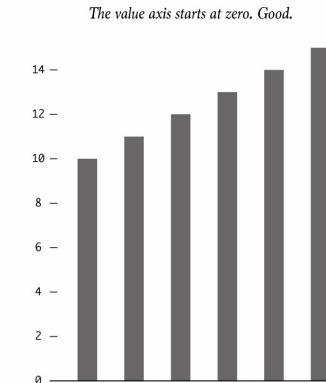


Average number of actual weekly hours of work in main job, full-time employees, 2013



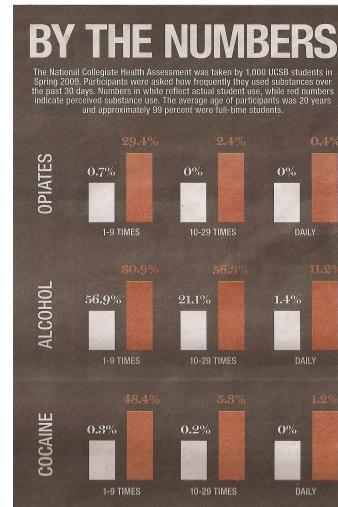
Source: Eurofound 2014

The value axis starts at zero. Good.



## BY THE NUMBERS

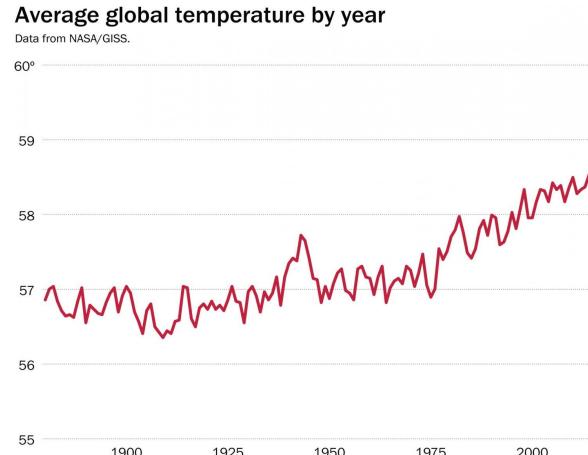
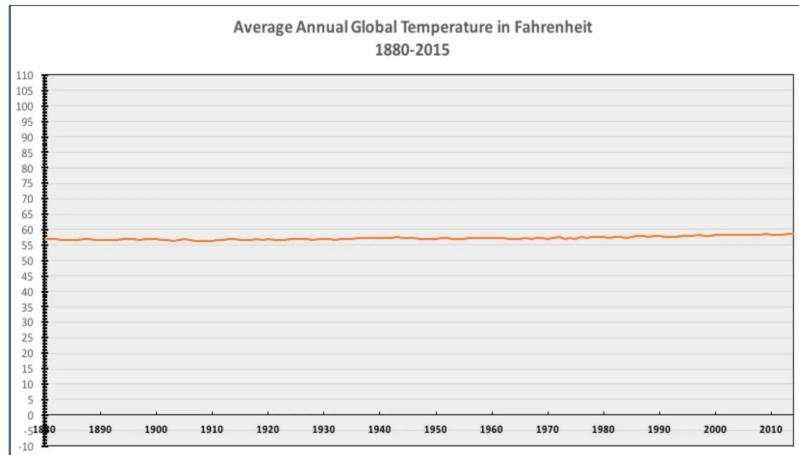
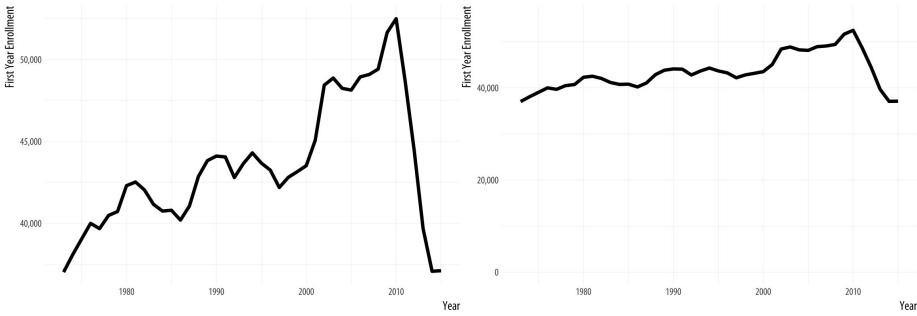
The National Collegiate Health Assessment was taken by 1,000 ICGB students in Spring 2009. Participants were asked how frequently they used substances over the past 30 days. Numbers in white reflect actual student use, while red numbers indicate perceived substance use. The average age of participants was 20 years and approximately 50 percent were full-time students.



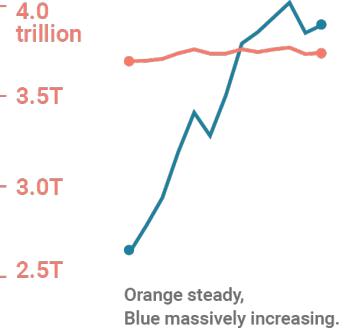
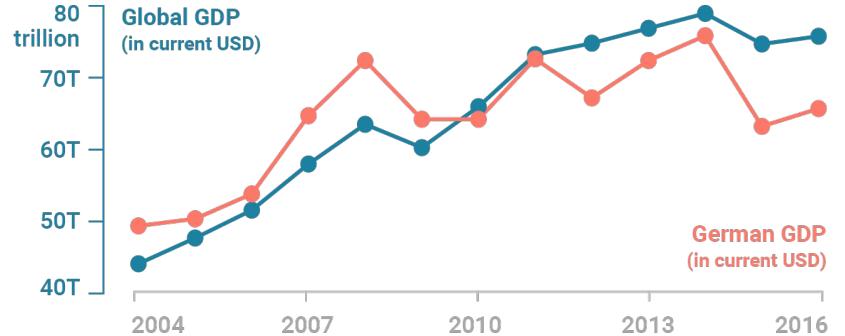
<https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/>

<http://socviz.co/lookatdata.html>; [https://www.callingbullshit.org/tools/tools\\_misleading\\_axes.html](https://www.callingbullshit.org/tools/tools_misleading_axes.html)

# Line graphs are better off without zero



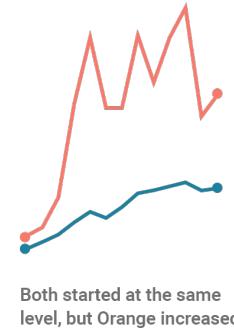
# No dual axes



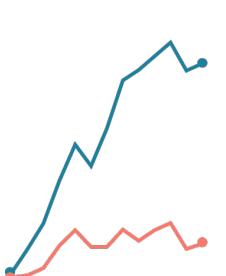
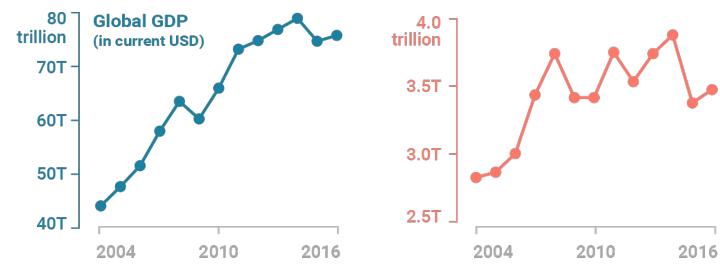
Orange steady,  
Blue massively increasing.



Blue steady,  
Orange increasing.



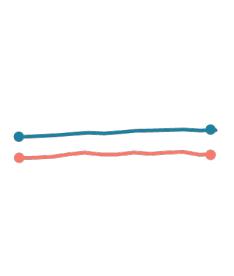
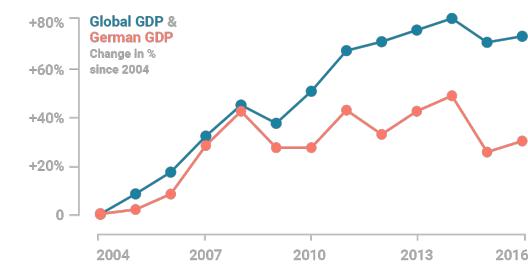
Both started at the same  
level, but Orange increased  
far more than Blue.



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

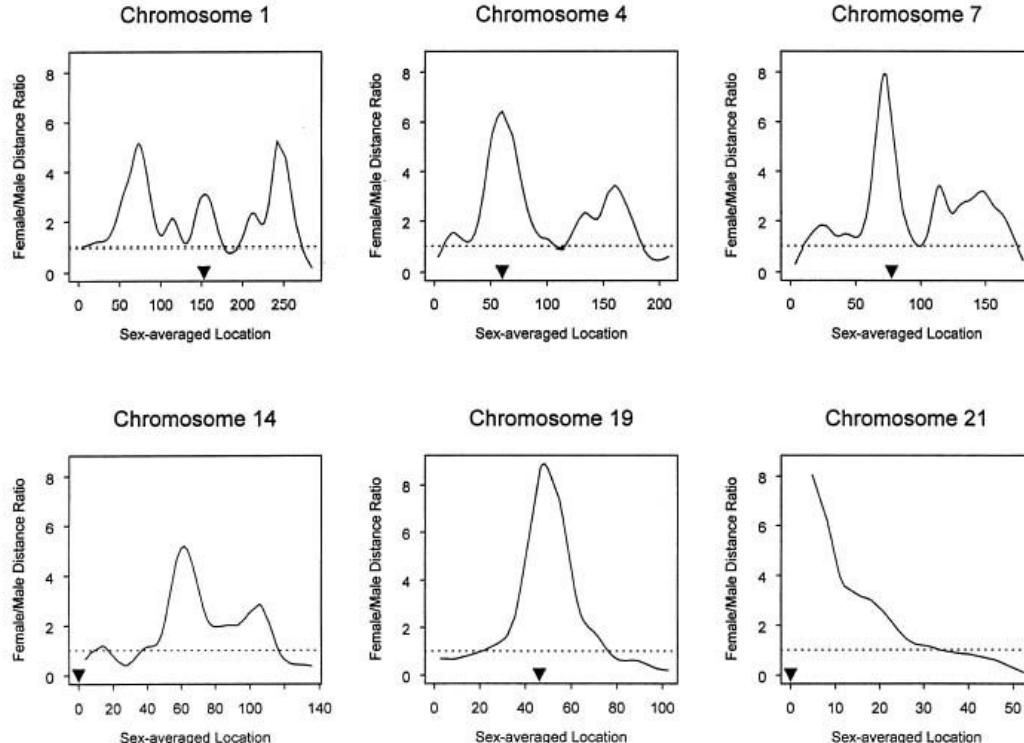


Both started with the  
same increase, then Blue  
raced to the top.



Both steady.

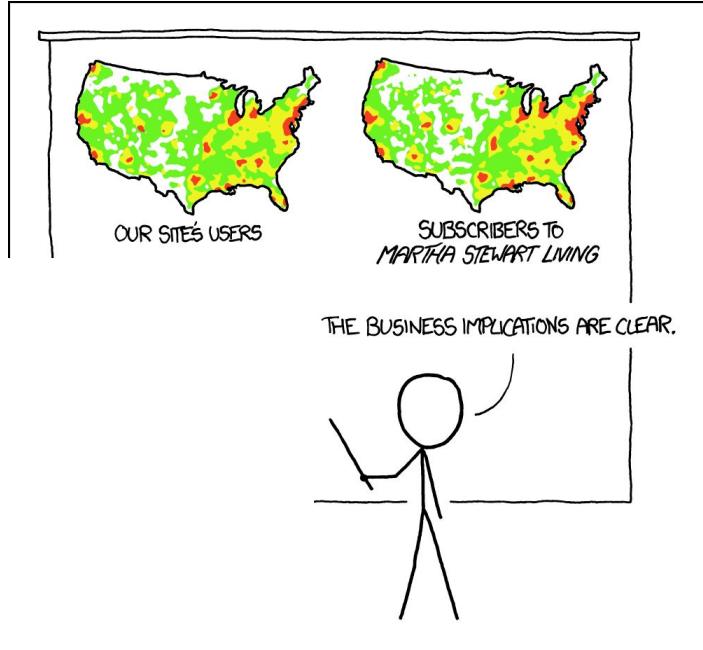
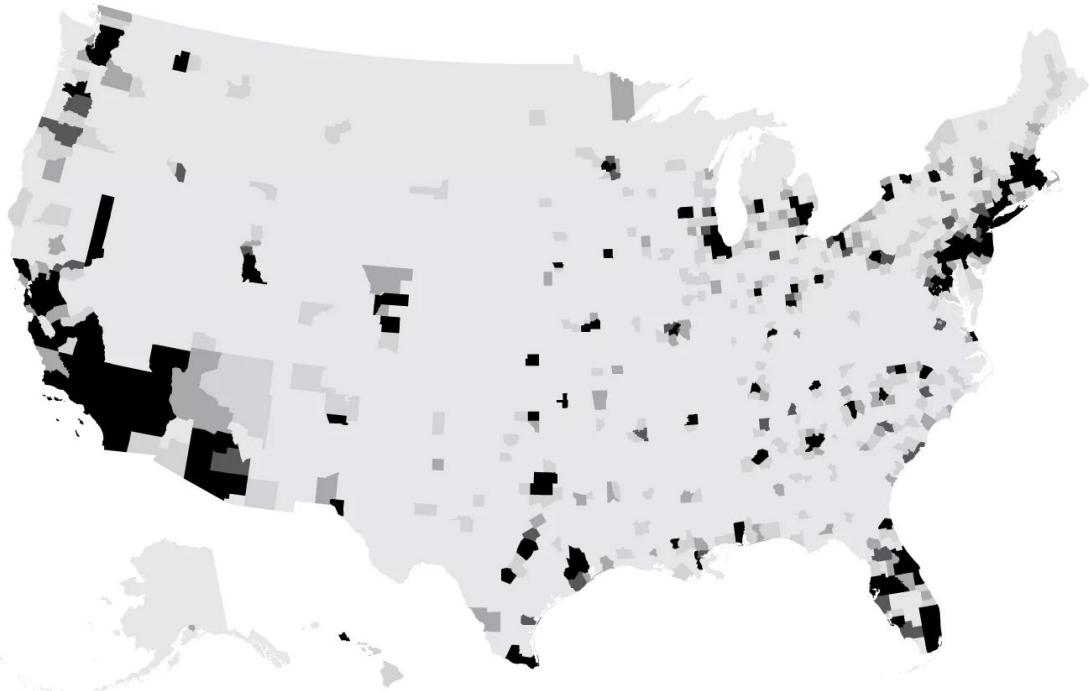
# Pay attention to inappropriate axes



**Figure 1** Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

# Beware of absolutes vs. relative values

*This is just population. When comparing across places, categories, or groups, you must compare fairly and consider relative values.*

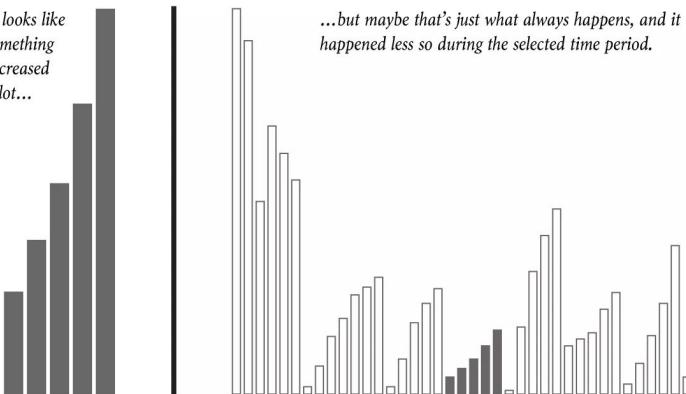


PET PEEVE #208:  
GEOGRAPHIC PROFILE MAPS WHICH ARE  
BASICALLY JUST POPULATION MAPS

# Showing full scope of the data is important

It looks like something increased a lot...

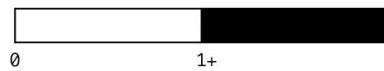
...but maybe that's just what always happens, and it happened less so during the selected time period.



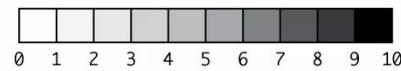
# Data binning matters

Two bins. What's really in the 1+ category?

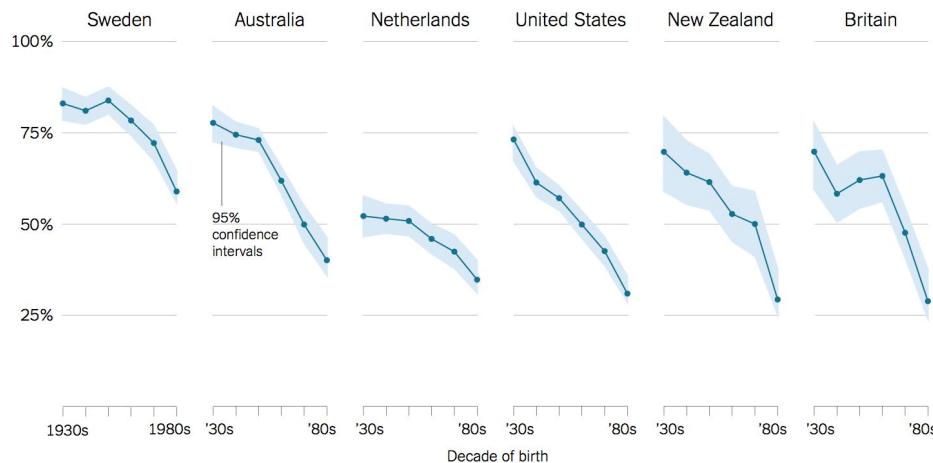
Might be hiding something.



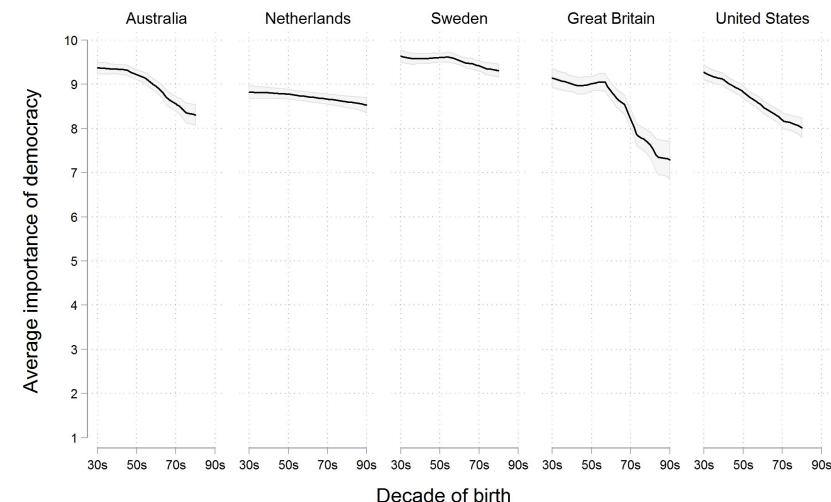
That's better. It can show more variation.



## Percentage of people who say it is “essential” to live in a democracy



Source: Yascha Mounk and Roberto Stefan Foa, “The Signs of Democratic Deconsolidation,” Journal of Democracy | By The New York Times



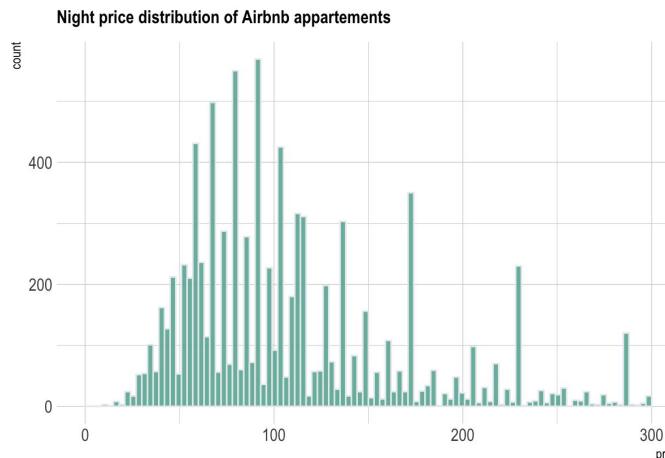
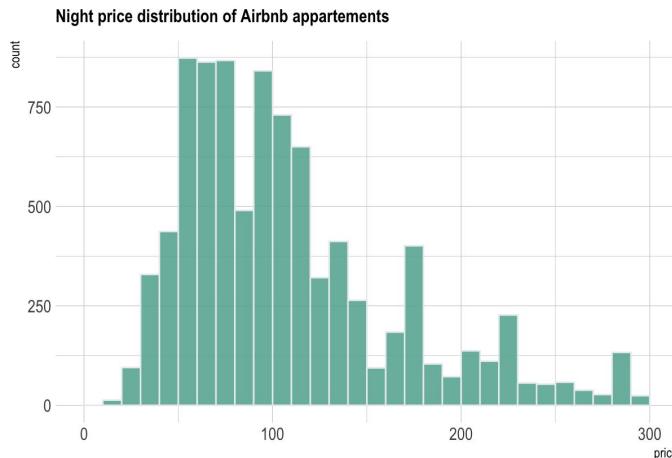
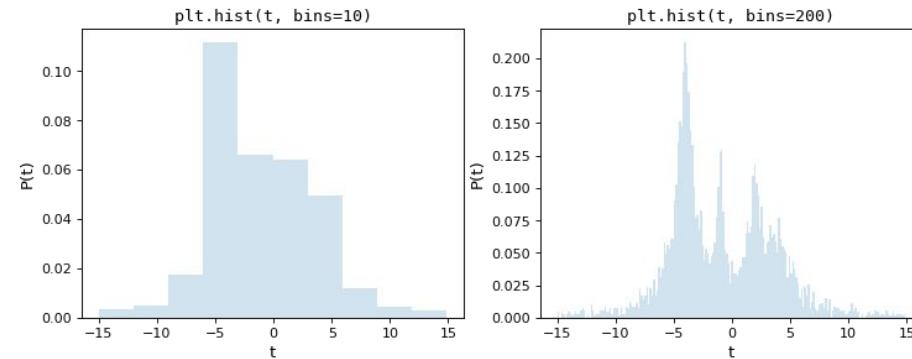
Graph by Erik Voeten, based on WVS 5

# Data binning matters

Two bins. What's really in the 1+ category?  
Might be hiding something.



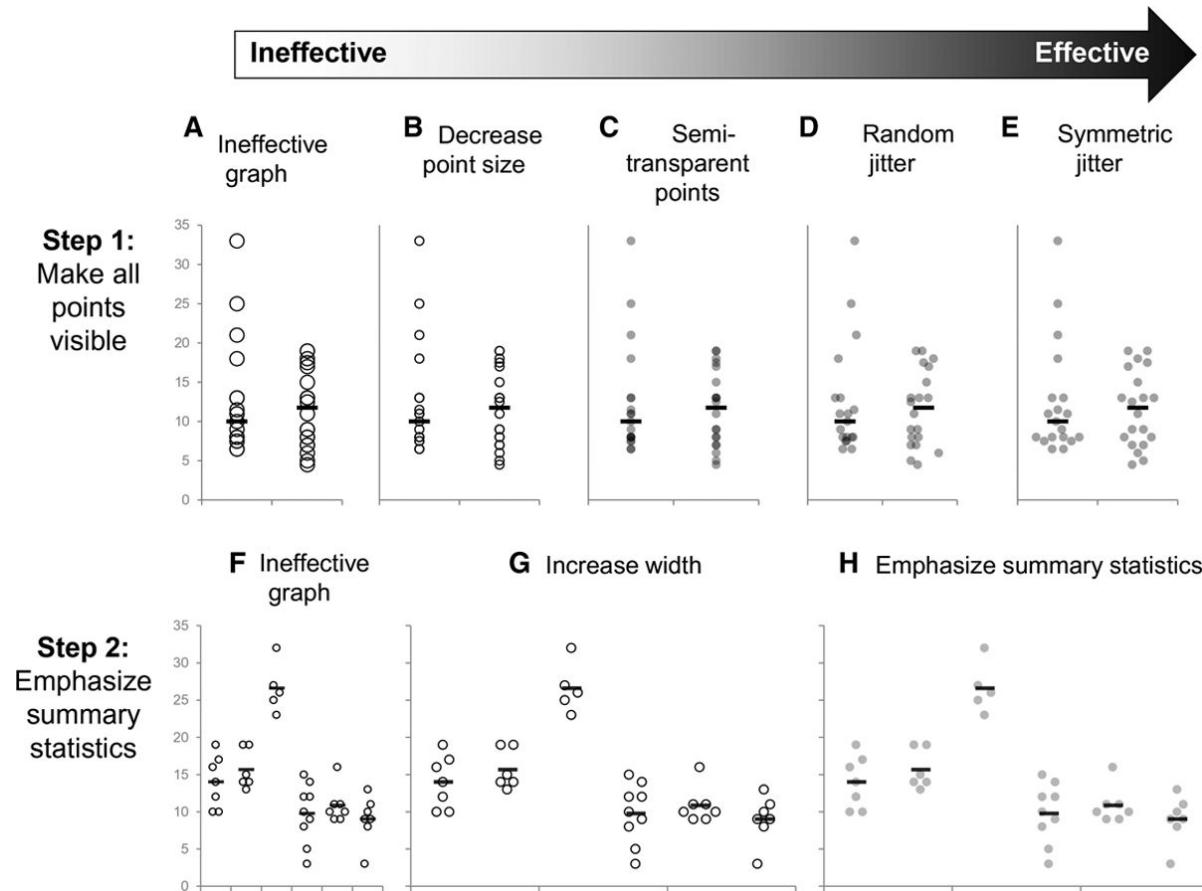
That's better. It can show more variation.



# Improving visualization for clarity

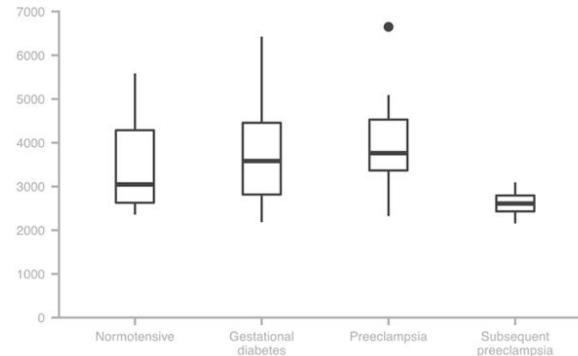
- Revealing underlying data as much as possible
- Changing emphasis based on data
- Reflecting study design / analysis
- Organizing & decluttering plots
- Choosing colors

# Simple changes can make plots effective

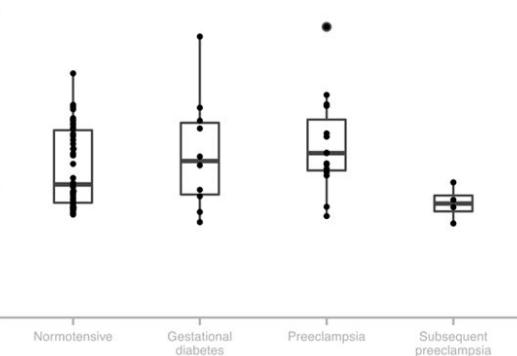


# Change emphasis based on data

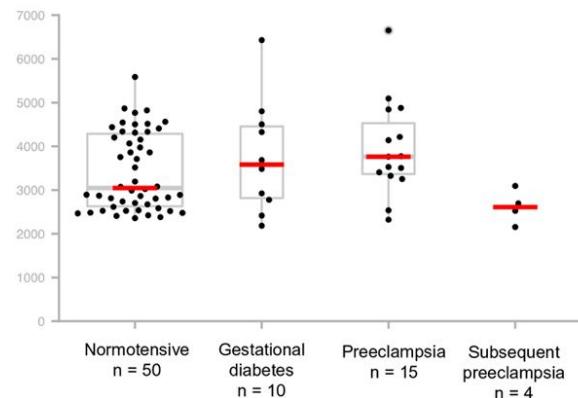
**A** Box plot



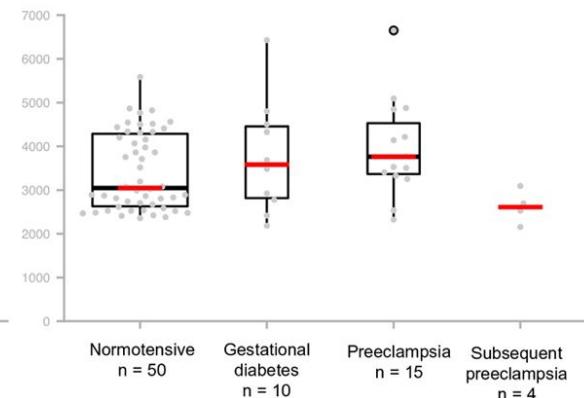
**B** Box plot withunjittered dot plot (strip plot)



**C** Emphasizing the dot plot



**D** Emphasizing the box plot



# Make the plot reflect the study-design / analysis

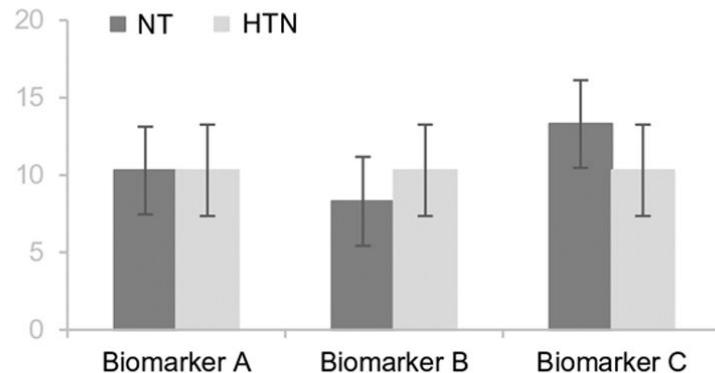
**Experimental goal:** Compare normotensive (NT) vs. hypertensive (HTN) patients

**Statistical analysis:** t-tests were used to compare values for each dependent variable (biomarker A, B and C)

**A**

## Sending mixed messages

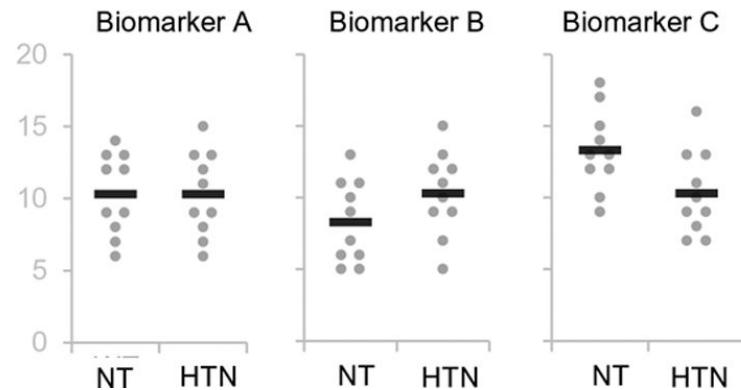
Figure structure erroneously suggests that authors also intended to compare biomarkers A, B and C



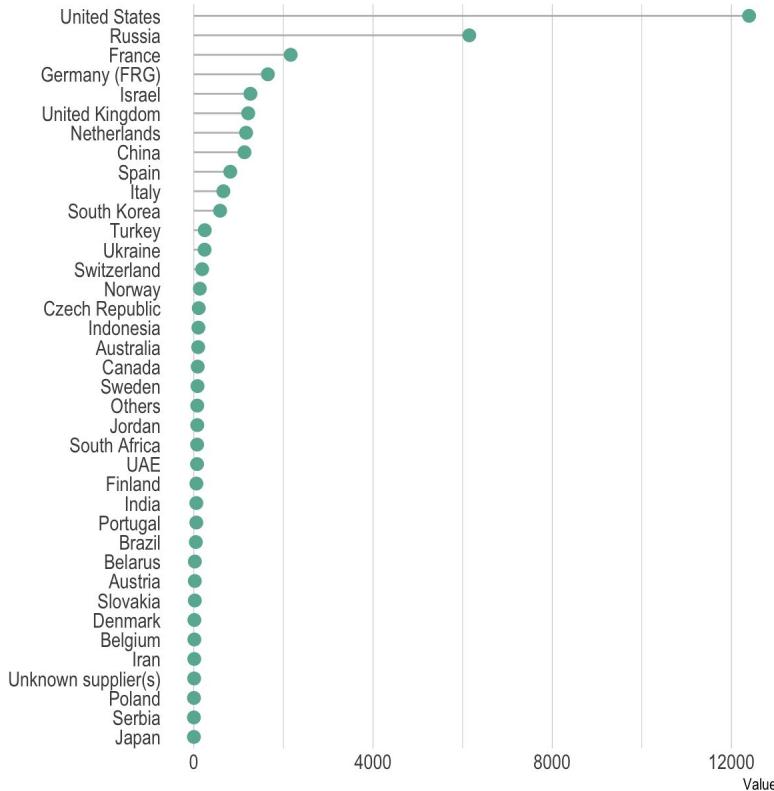
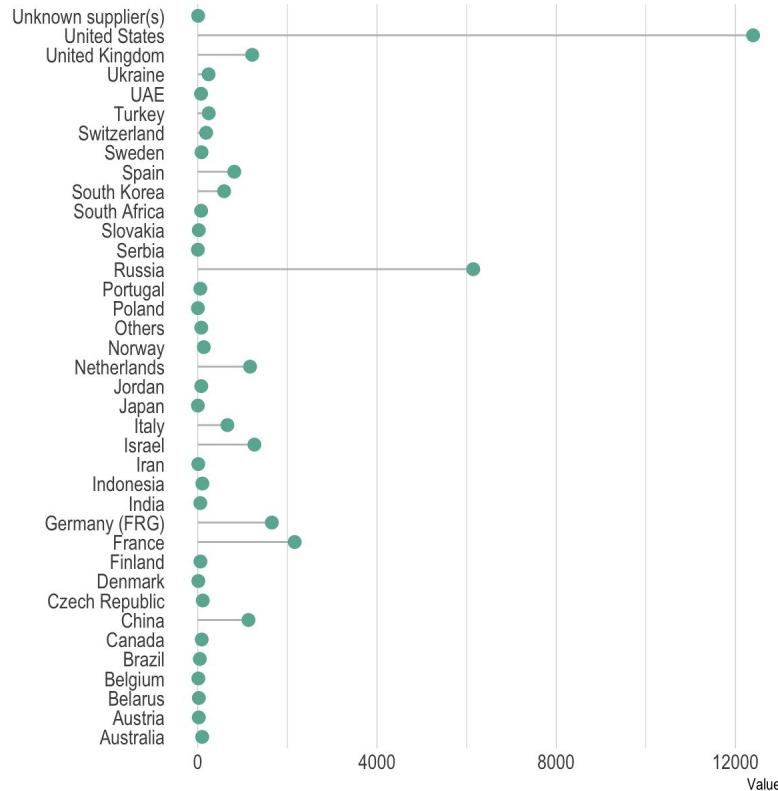
**B**

## Clear communication

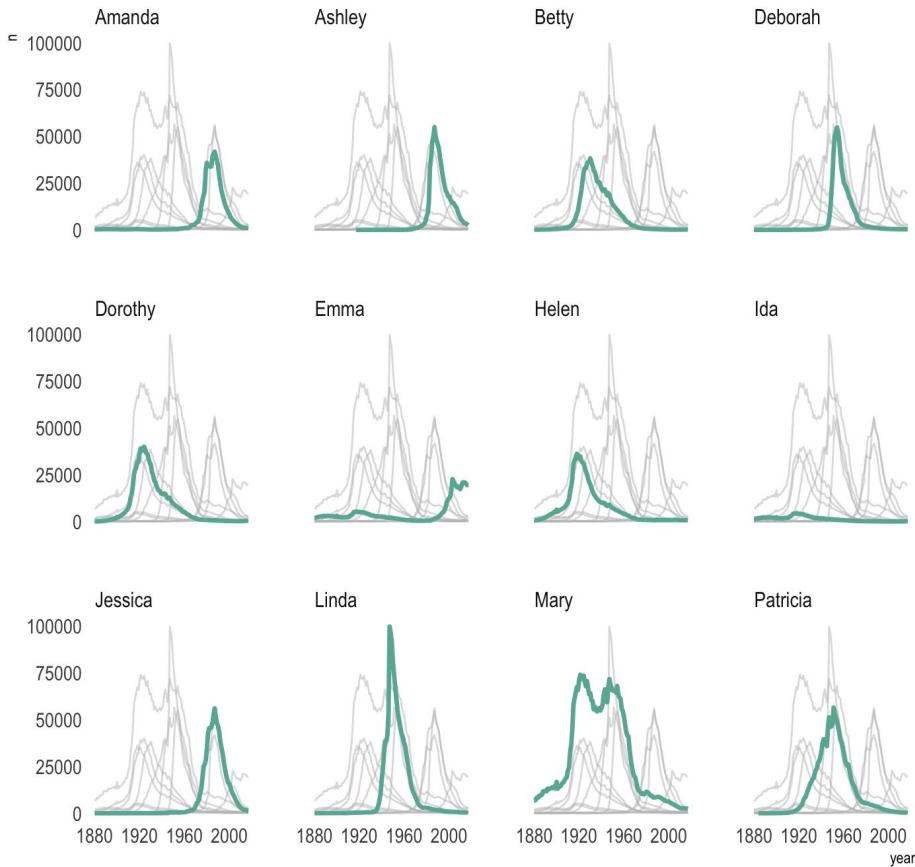
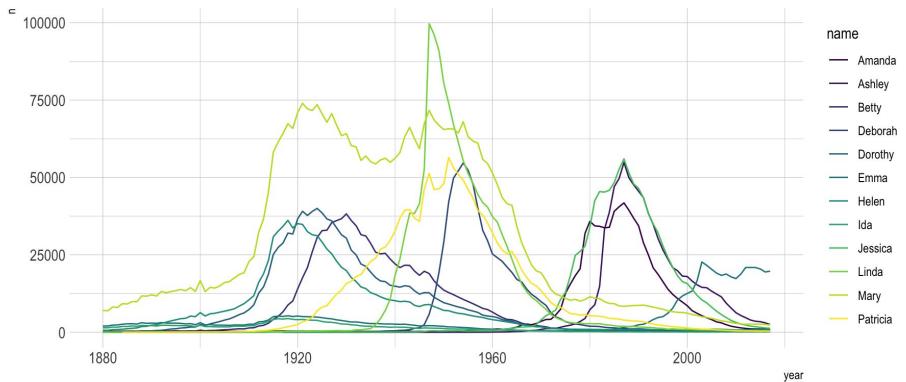
Figure structure matches study design & analysis, shows that the authors did not intend to compare biomarkers



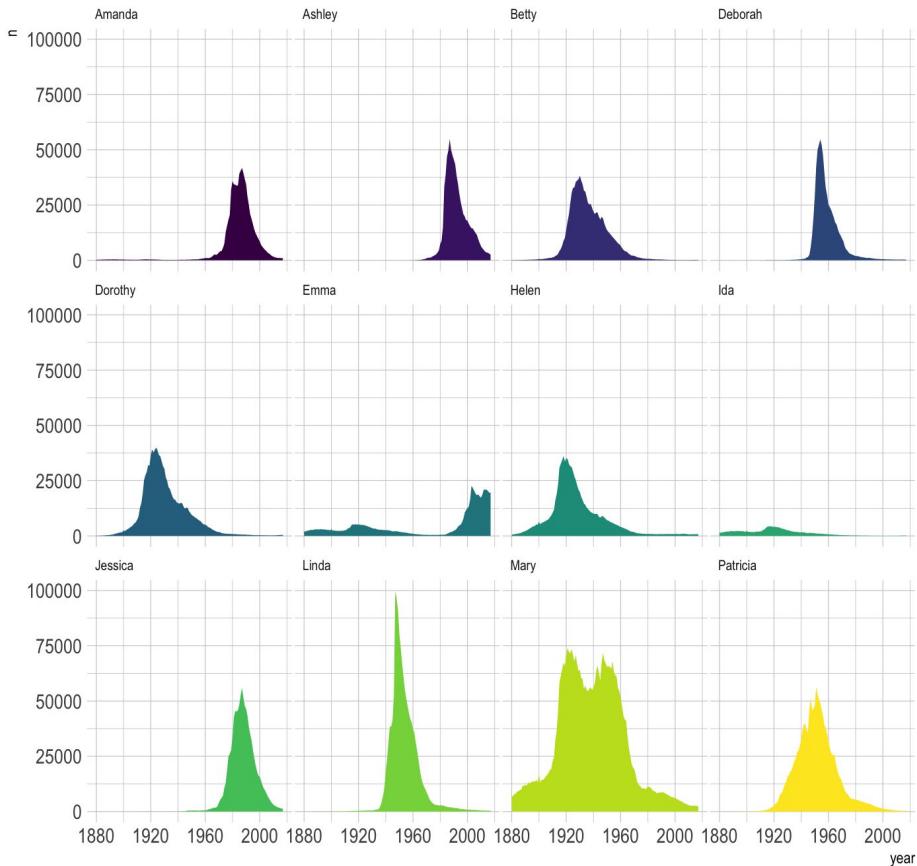
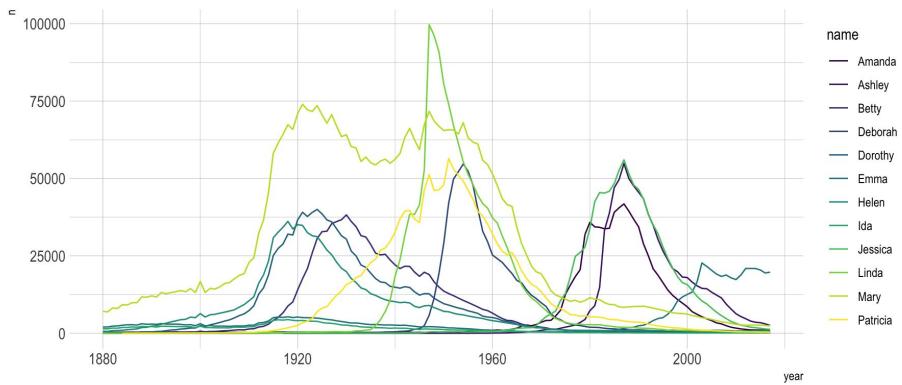
# Reorder to make comparison clearer



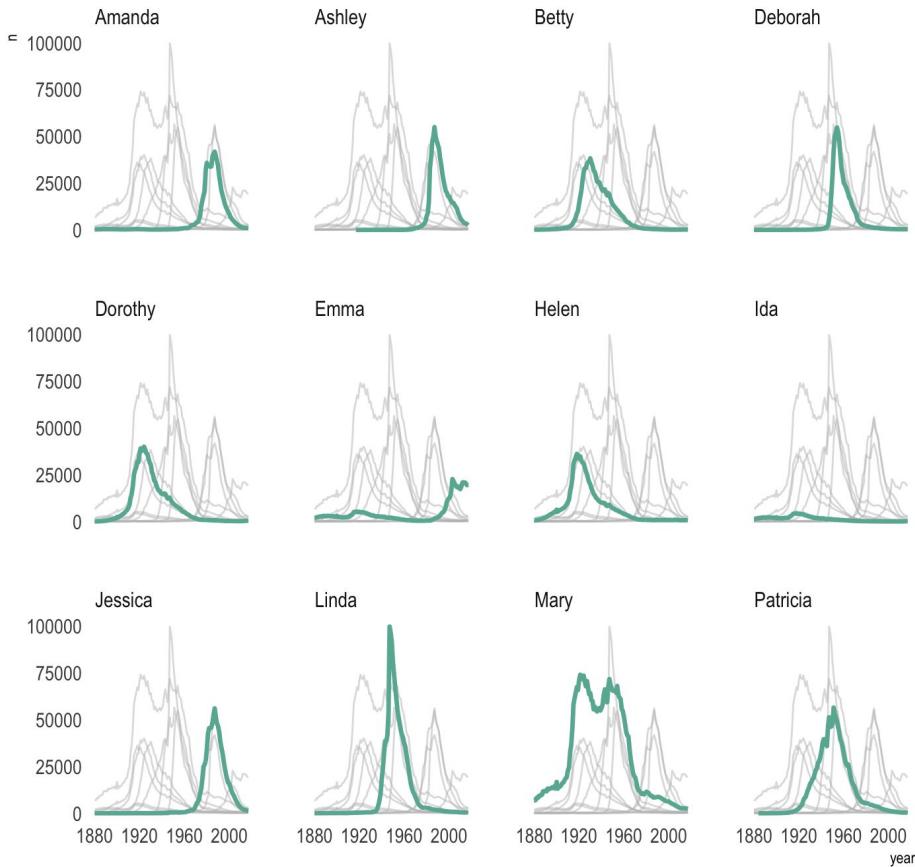
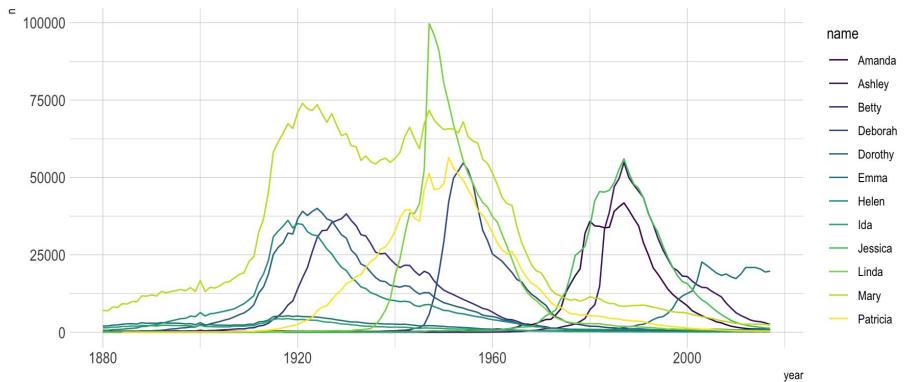
# Split plots if complicated



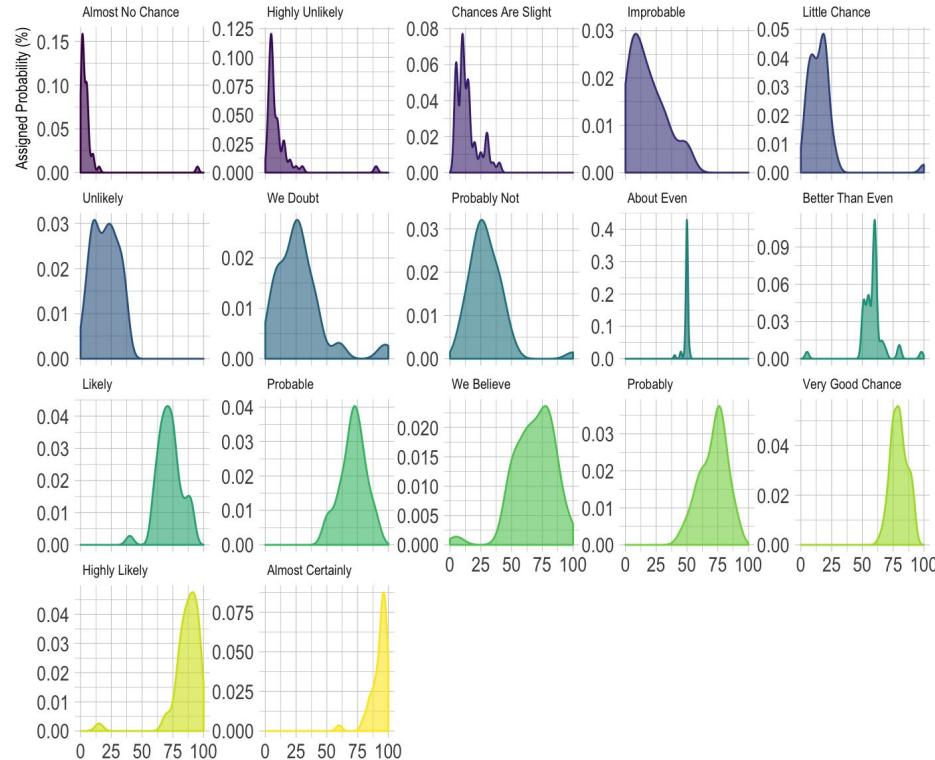
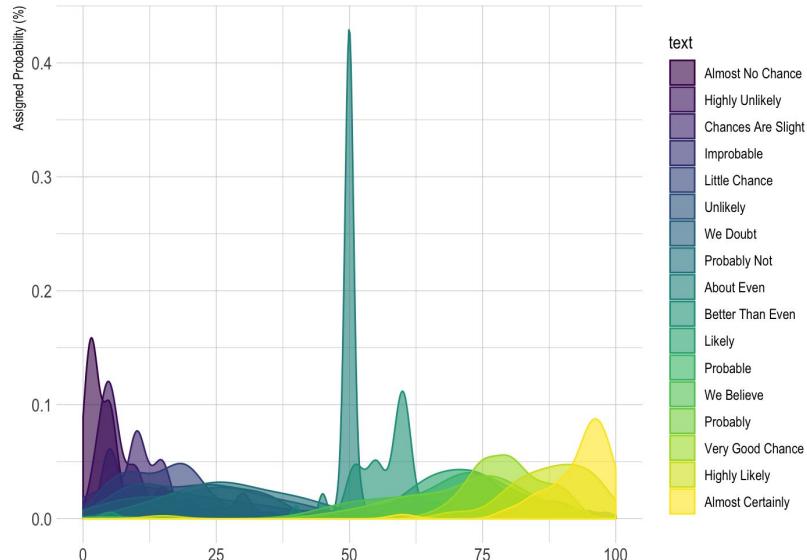
# Split plots if complicated



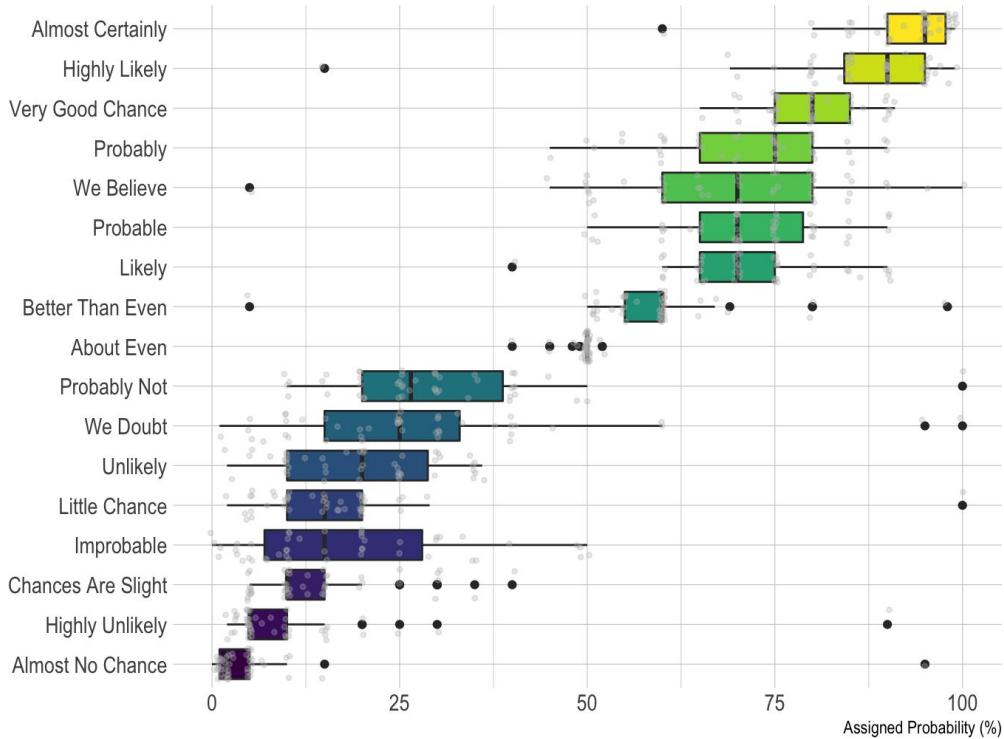
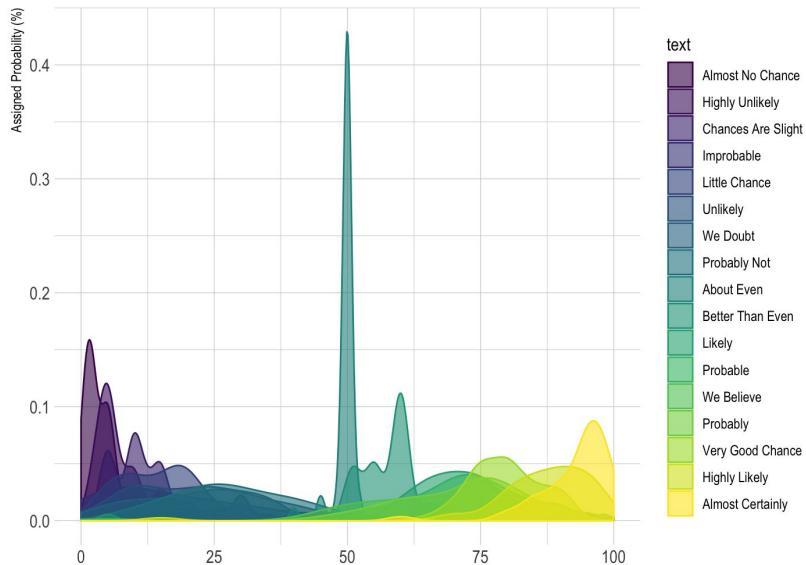
# Split plots if complicated



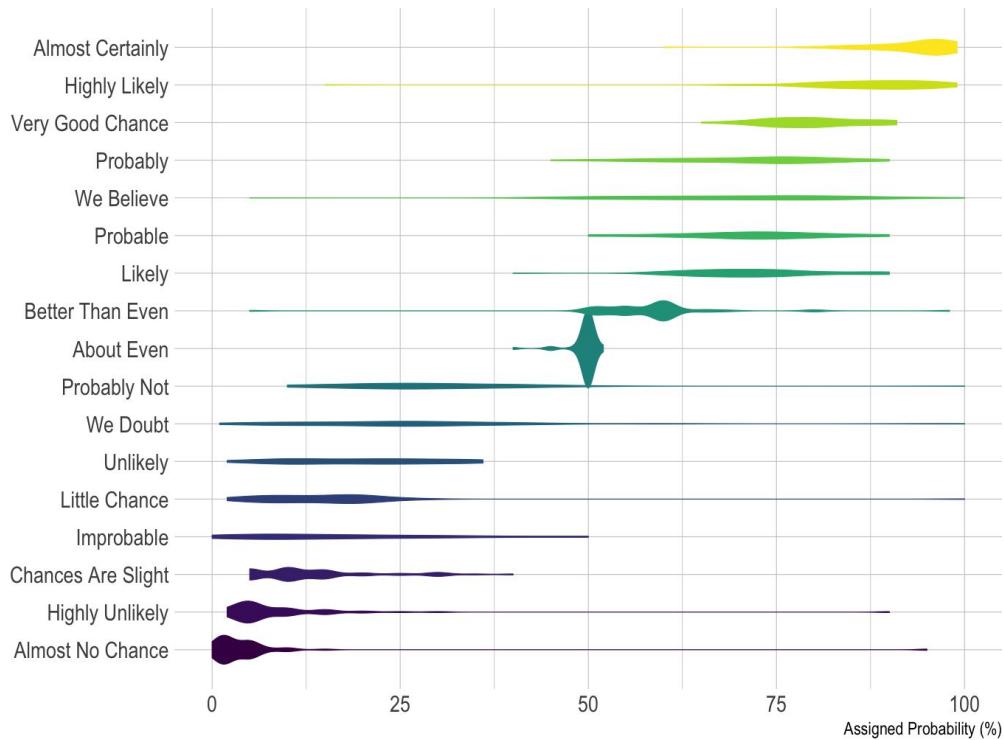
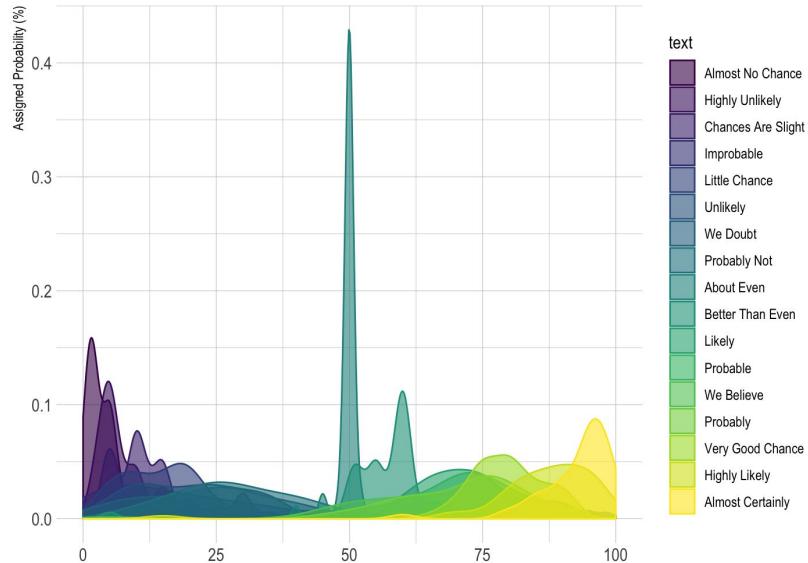
# Split plots if complicated



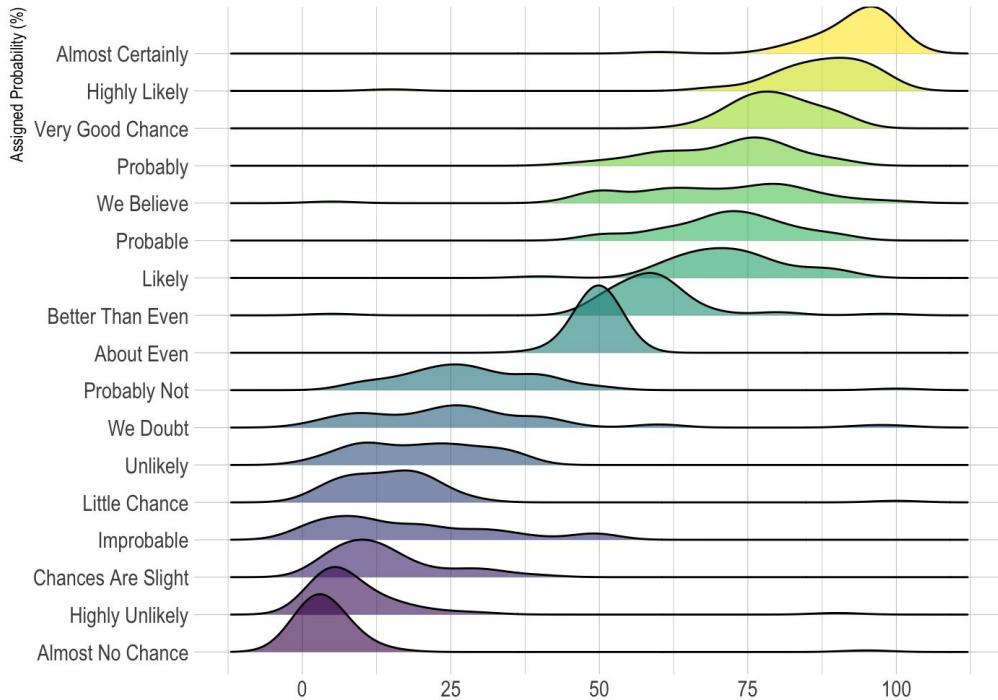
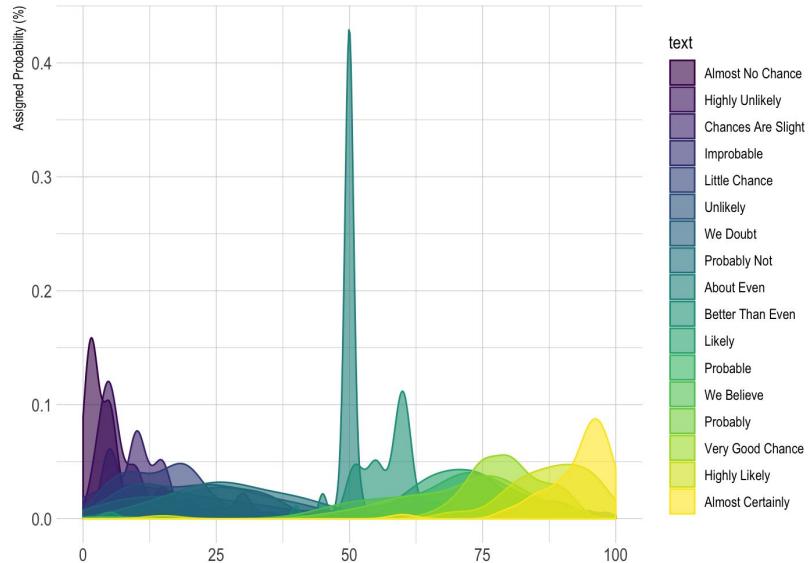
# Split plots if complicated



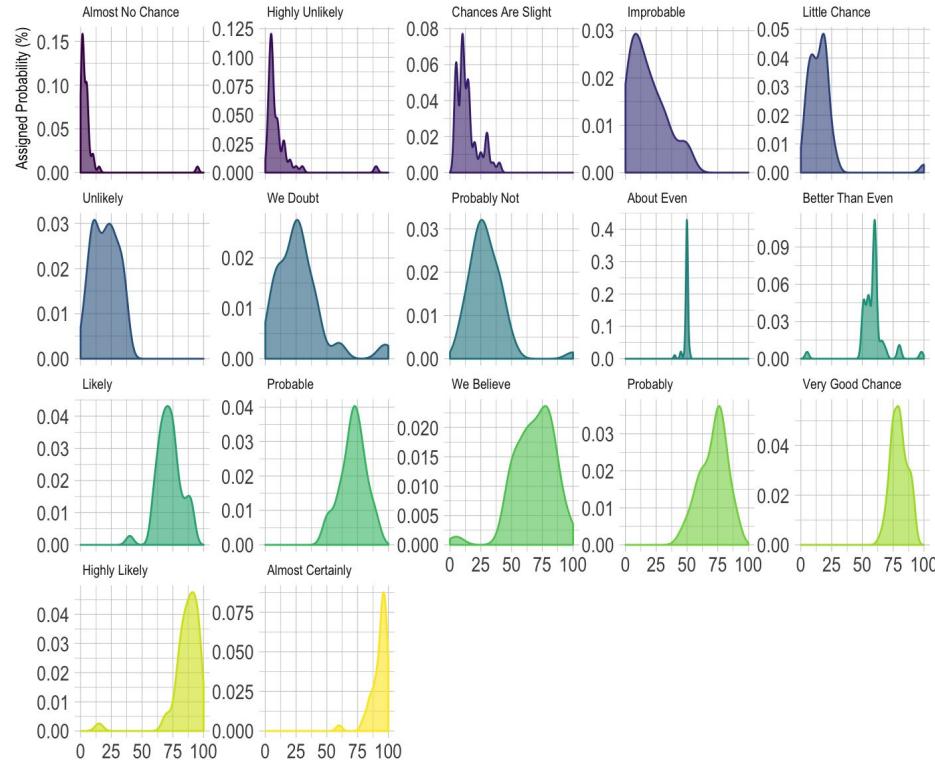
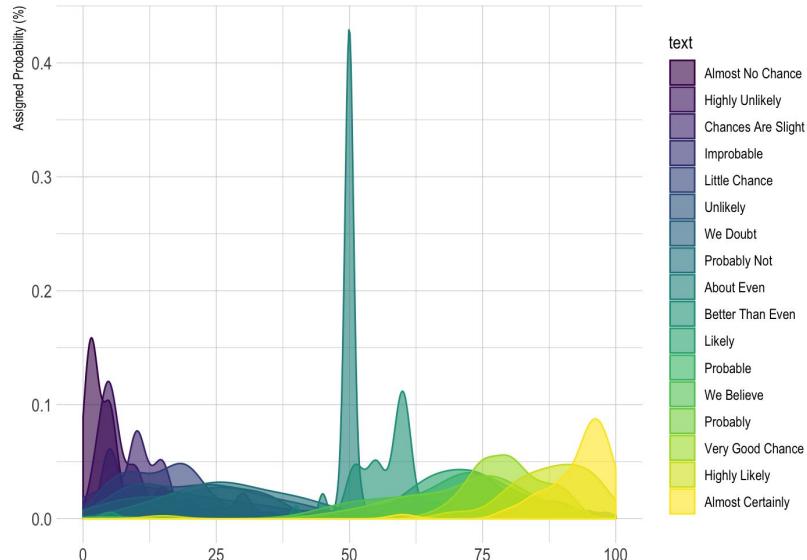
# Split plots if complicated



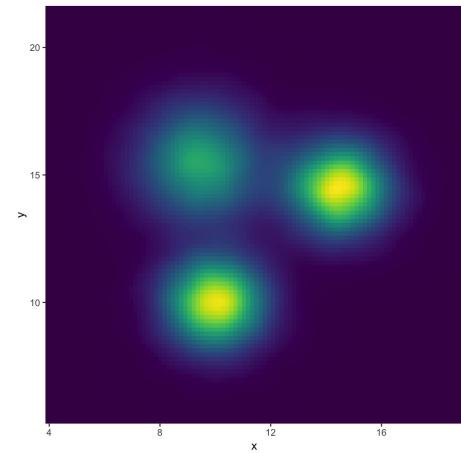
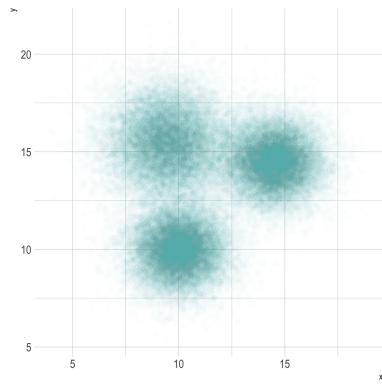
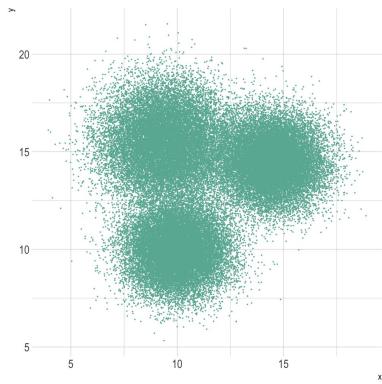
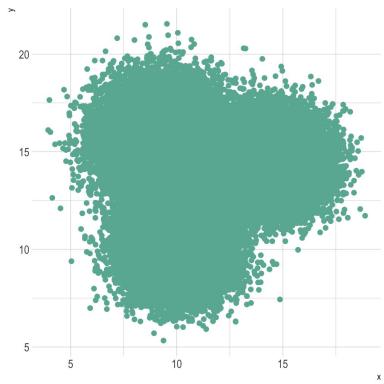
# Split plots if complicated



# Split plots if complicated

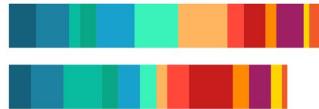


# Avoid overplotting



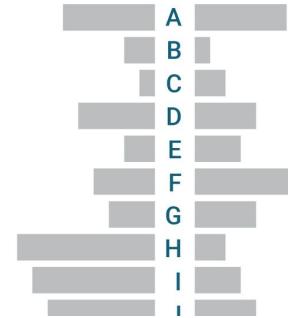
# Choose colors carefully

NOT IDEAL



BETTER

BETTER



COLOR KEY



CONTRAST

CONTRAST RATIOS

1.0	Choose if you dislike readers.	That's bad.	That's bad.	Horrible.
1.1	That's bad.	Not ideal.	That's bad.	My eyes!
1.5	Ok in 1% of the cases.	Not ideal.	Not ideal.	That's bad.
2.5	Can be a good choice.	Ok.	Not ideal.	That's bad.
4.5	Safe choice.	Great.	Ok.	Not ideal.

NOT IDEAL



SHARE OF PEOPLE IN CHINA AND GERMANY



BETTER





# from Data to Viz

'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

**1** Identify what type of data you have.

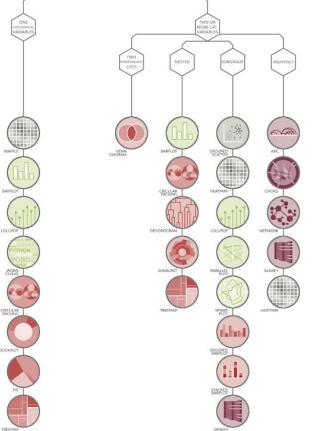
**2** Go to the corresponding decision tree and follow it down to a set of possible charts.

**3** Choose the chart from the set that will suit your data and your needs best.

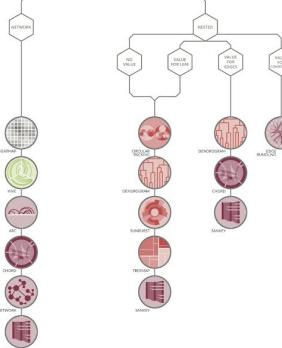
Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

[data-to-viz.com](http://data-to-viz.com)

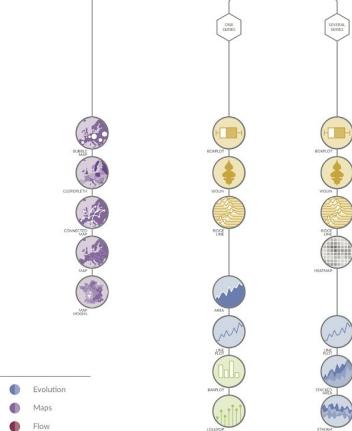
## CATEGORIC



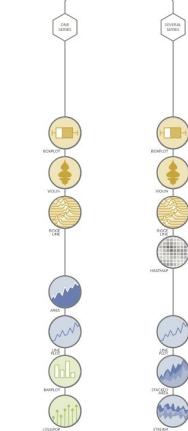
## RELATIONAL



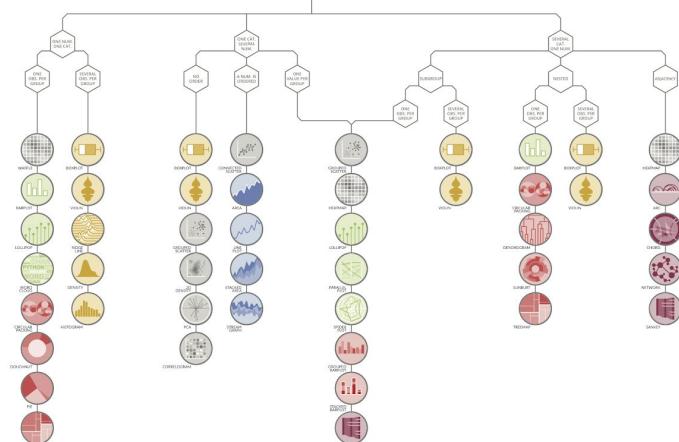
## MAP



## TIME SERIES



## CATEGORIC AND NUMERIC



## NUMERIC

