Visualizing data

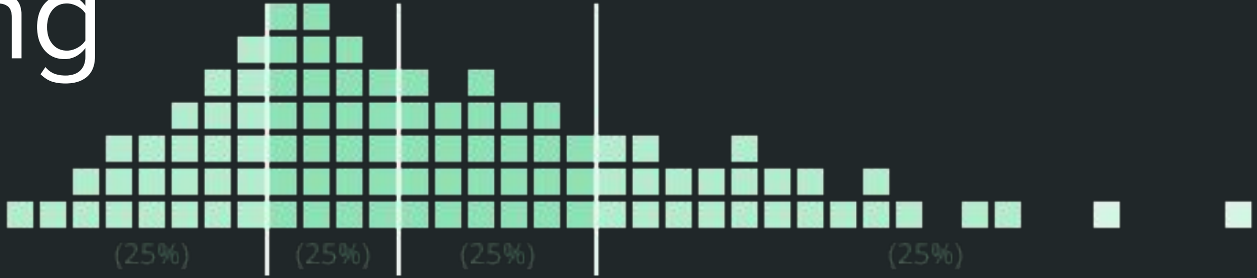


BILD 62

Objectives for today

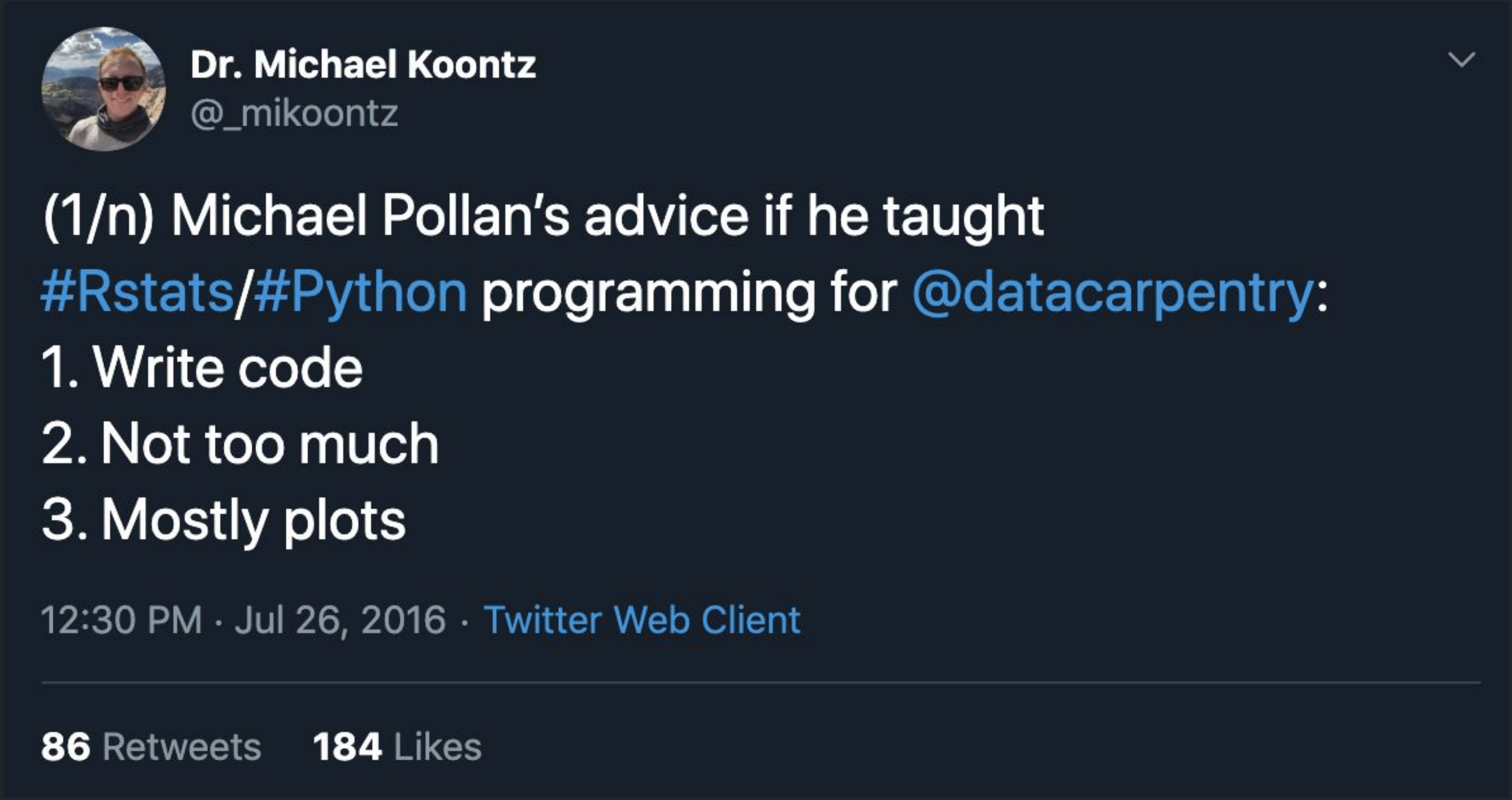
- Describe best practices for data visualization
- Introduce the tools you can use to plot in Python
- Dive into a notebook to plot our inflammation data

Data visualization is an important step for sharing big insights about your data

Your plots should be **clear** & **concise**.

all axes, groups, & trendlines are labeled

make a point with the least amount of visual information



Dr. Michael Koontz
@_mikoontz

(1/n) Michael Pollan's advice if he taught #Rstats/#Python programming for @datacarpentry:

1. Write code
2. Not too much
3. Mostly plots

12:30 PM · Jul 26, 2016 · Twitter Web Client

86 Retweets **184** Likes

When should you use the following types of graphs?

- Histogram
- Line graph
- Box plot
- Scatter plot
- Heatmap
- Bar graph
- Pie chart

Image: Klipfolio

When should you use the following types of graphs?

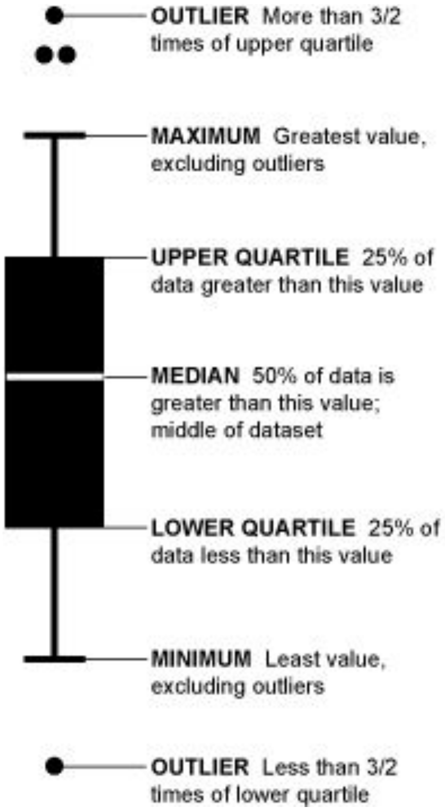
- **Histogram**: to see the distribution of your data
- **Line graph**: continuous data (e.g., over time or distance)
- **Box/violin plot**: to compare different categorical groups when you have information about the variability
- **Scatter plot**: compare continuous data for two groups
- **Heatmap:** when you'd like to show complex data that has three dimensions; often comparing two categorical *or* continuous variables (often when variability is *less* important to show)

Interpreting a box & whisker plot

1. Draw the median at 50% of the data points
2. Divide the top & bottom half into quartiles. These represent an additional 25% of the data.
3. Outliers are 3/2 (or 1.5X) the bottom/top quartile.

Note: Box & whisker plots can include the mean, too!

Note #2: Works best with 5+ data points.


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

See also https://www.nature.com/articles/nmeth.2813; https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/stats-box-whisker-plots/v/reading-box-and-whisker-plots

Scatter plots are useful to inspect relationships between two variables…

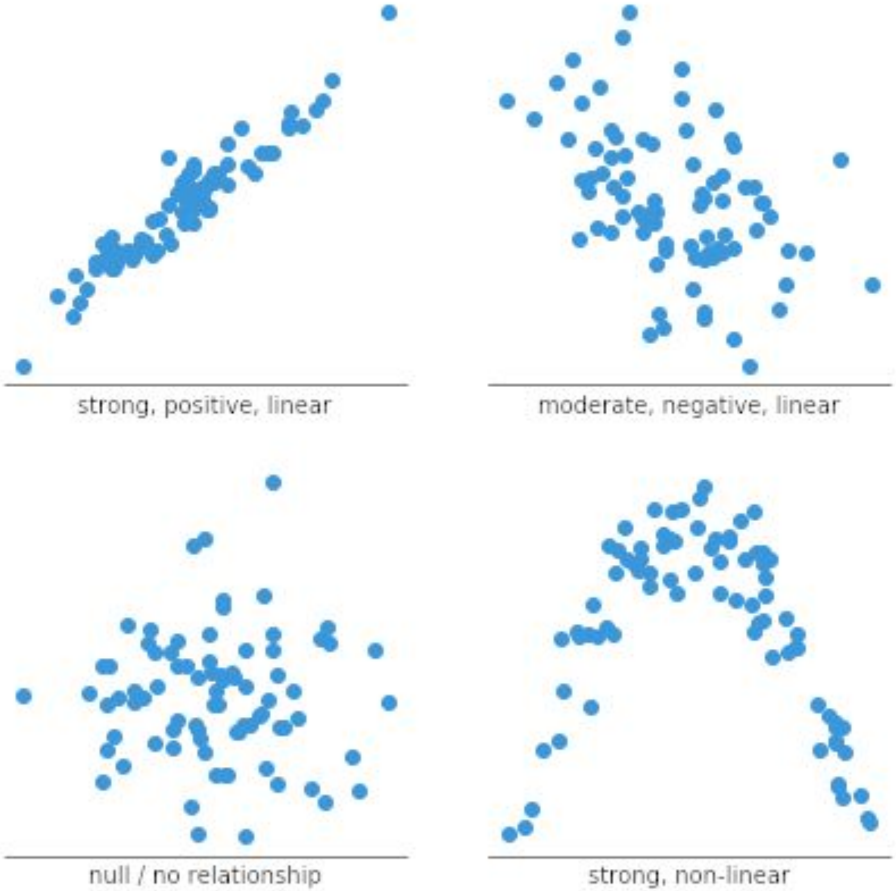


Image: ChartIO

… as well as identifying different patterns in the data

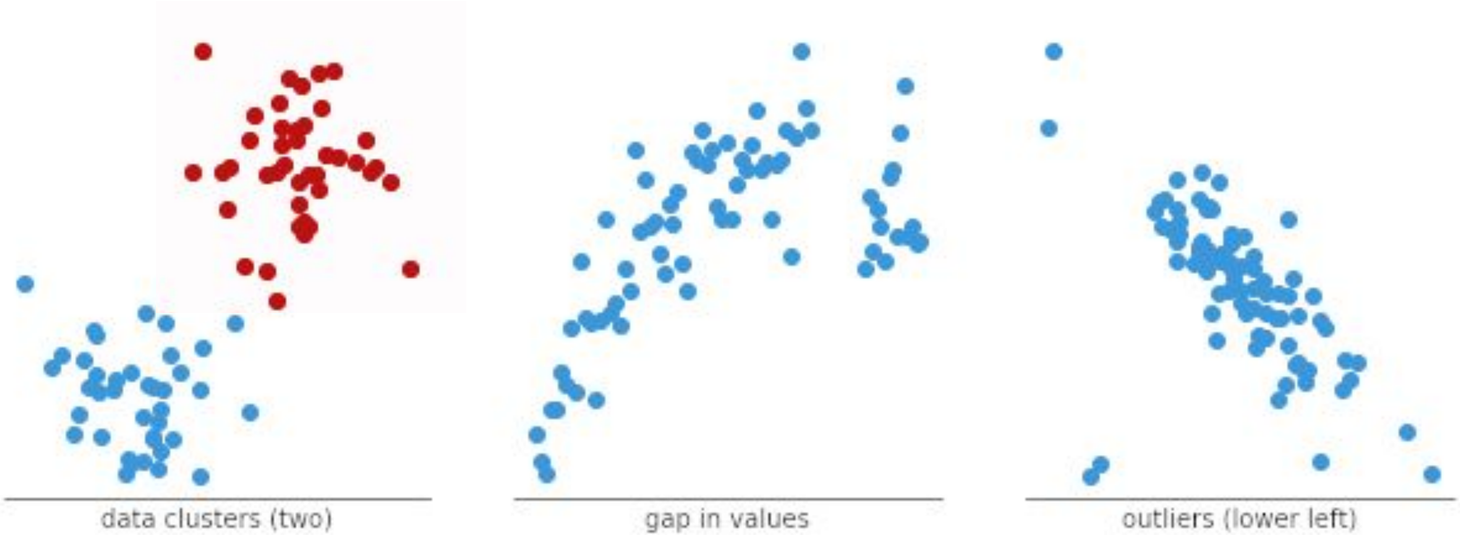
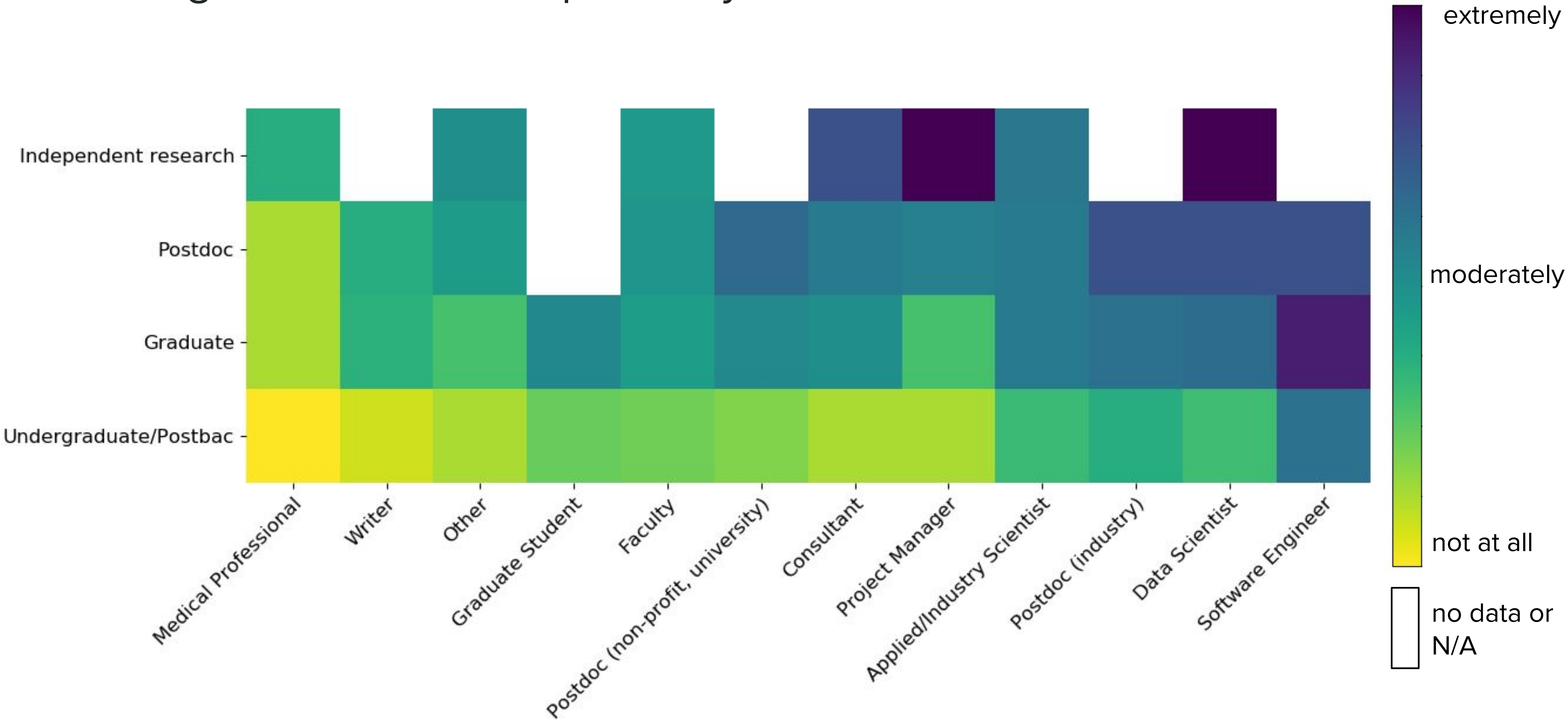


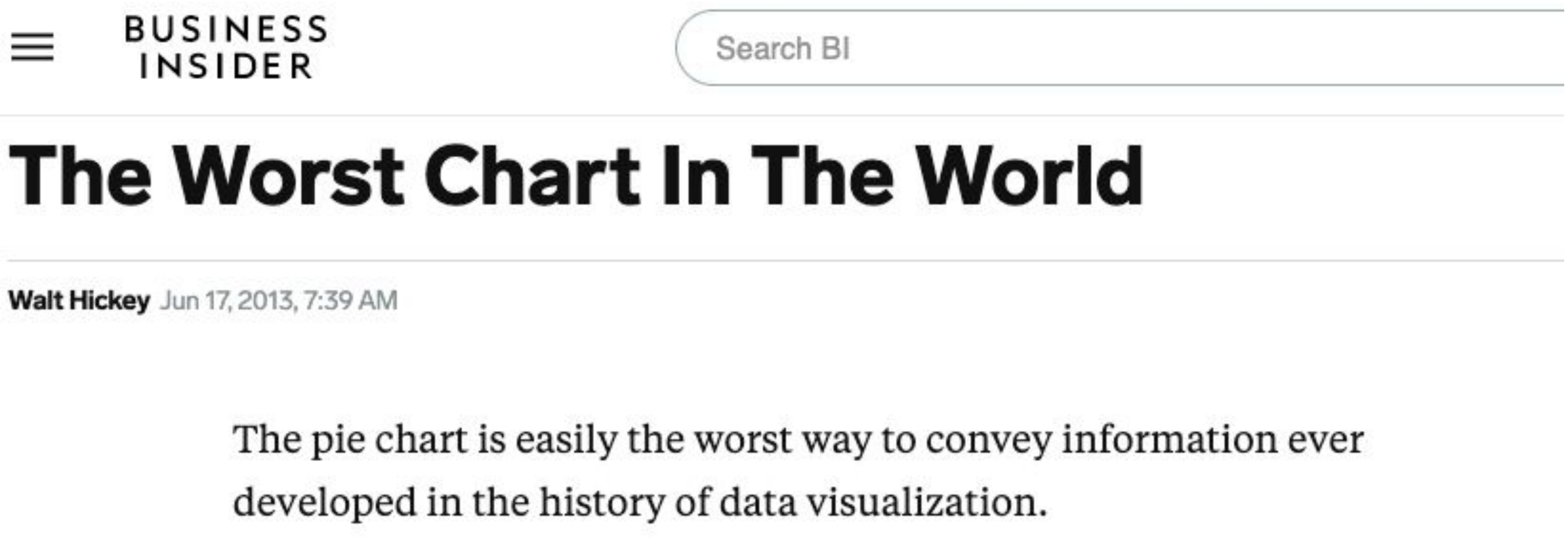
Image: ChartIO

Heatmap example: How comfortable did/do you feel working with code at this point in your career?



Try to avoid using bar graphs and pie charts

- **Bar graph**: acceptable for preliminary data visualization or to show data for which you do not have information about the variability (e.g., # of observations, percentages)
- **Pie chart**: only if you're showing 2-3 groups that are *very different*

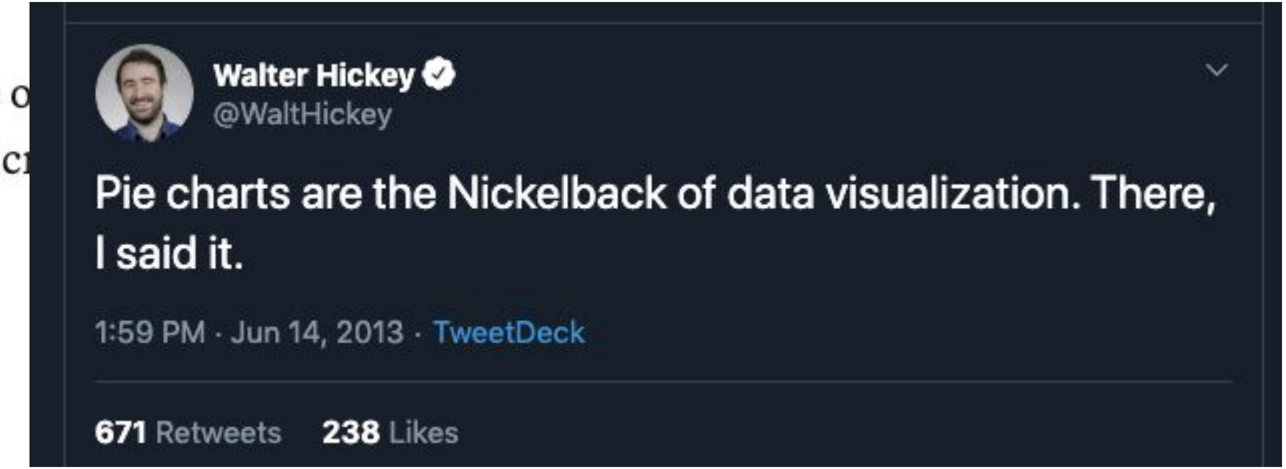


The Worst Chart In The World

Walt Hickey Jun 17, 2013, 7:39 AM

The pie chart is easily the worst way to convey information ever developed in the history of data visualization.

Sure, there are o
none have the c
has.



Walter Hickey @WaltHickey

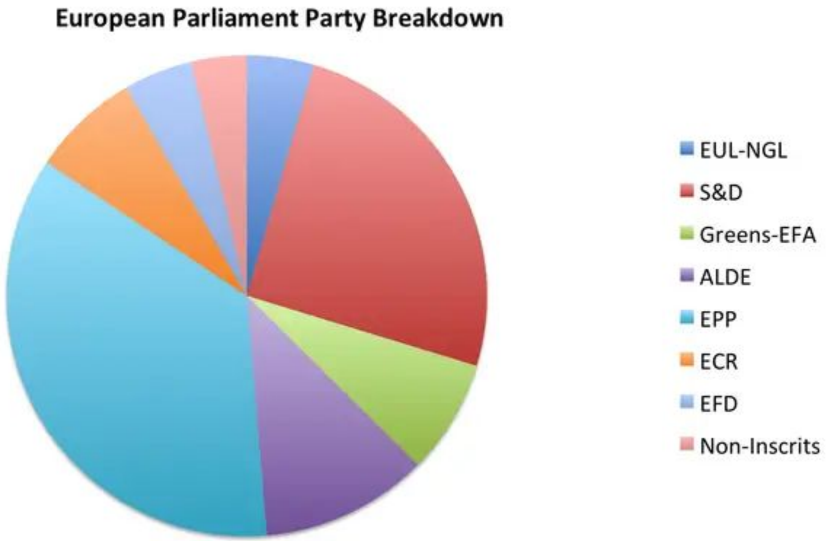
Pie charts are the Nickelback of data visualization. There, I said it.

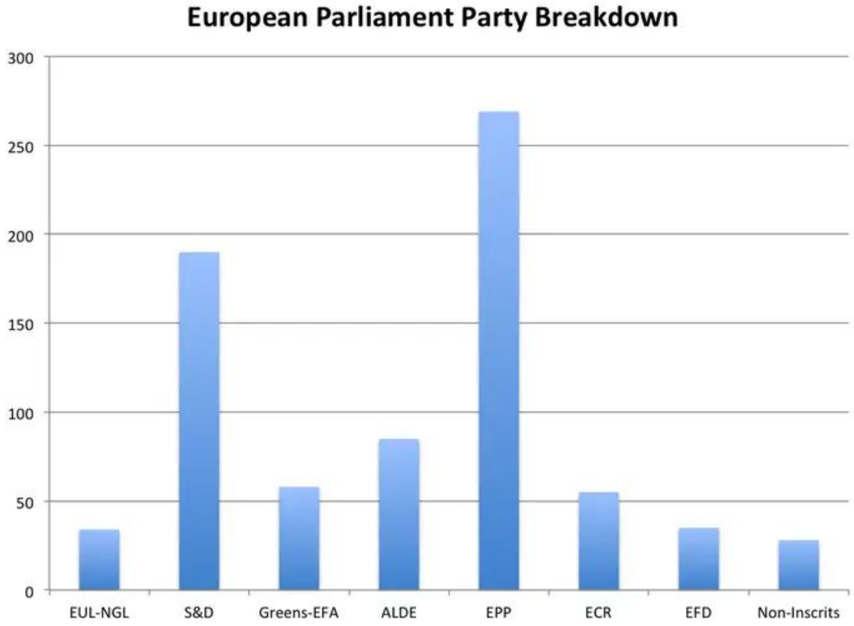
1:59 PM · Jun 14, 2013 · TweetDeck

671 Retweets **238** Likes

https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

Which chart makes a clearer point?

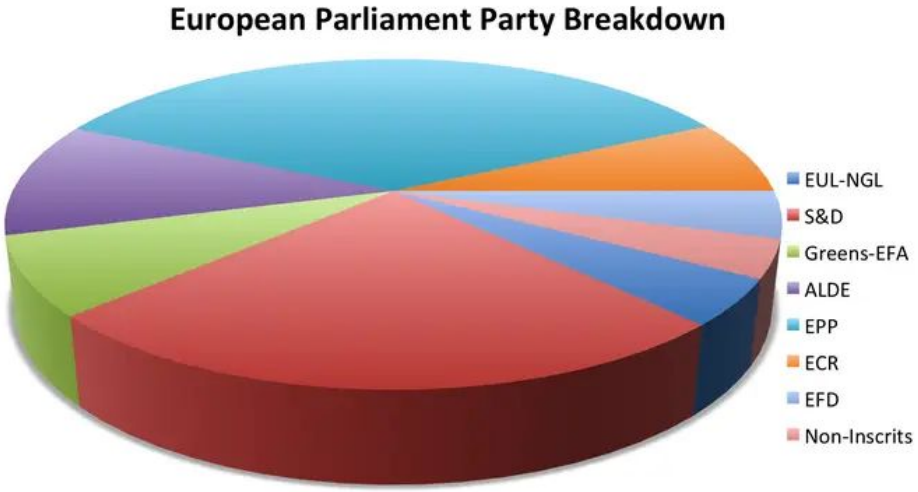


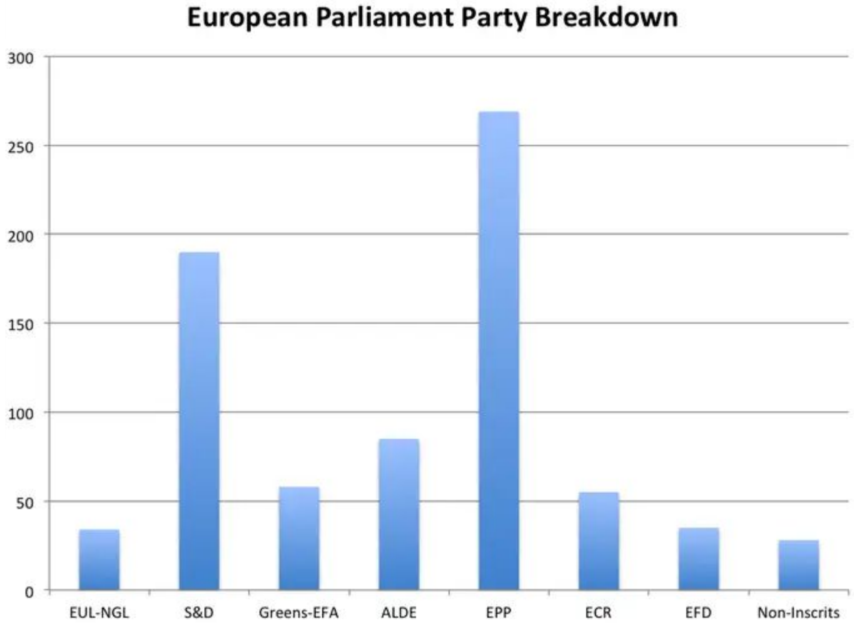


Walter Hickey / Business Insider

https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

3D pie charts are even worse!





Walter Hickey / Business Insider
https://www.businessinsider.com/pie-charts-are-the-worst-2013-6

What kind of graph would you use for the following:

a. You have recorded 10 data points for tumor volume with treatment A, and would like to look at the distribution of these data points.

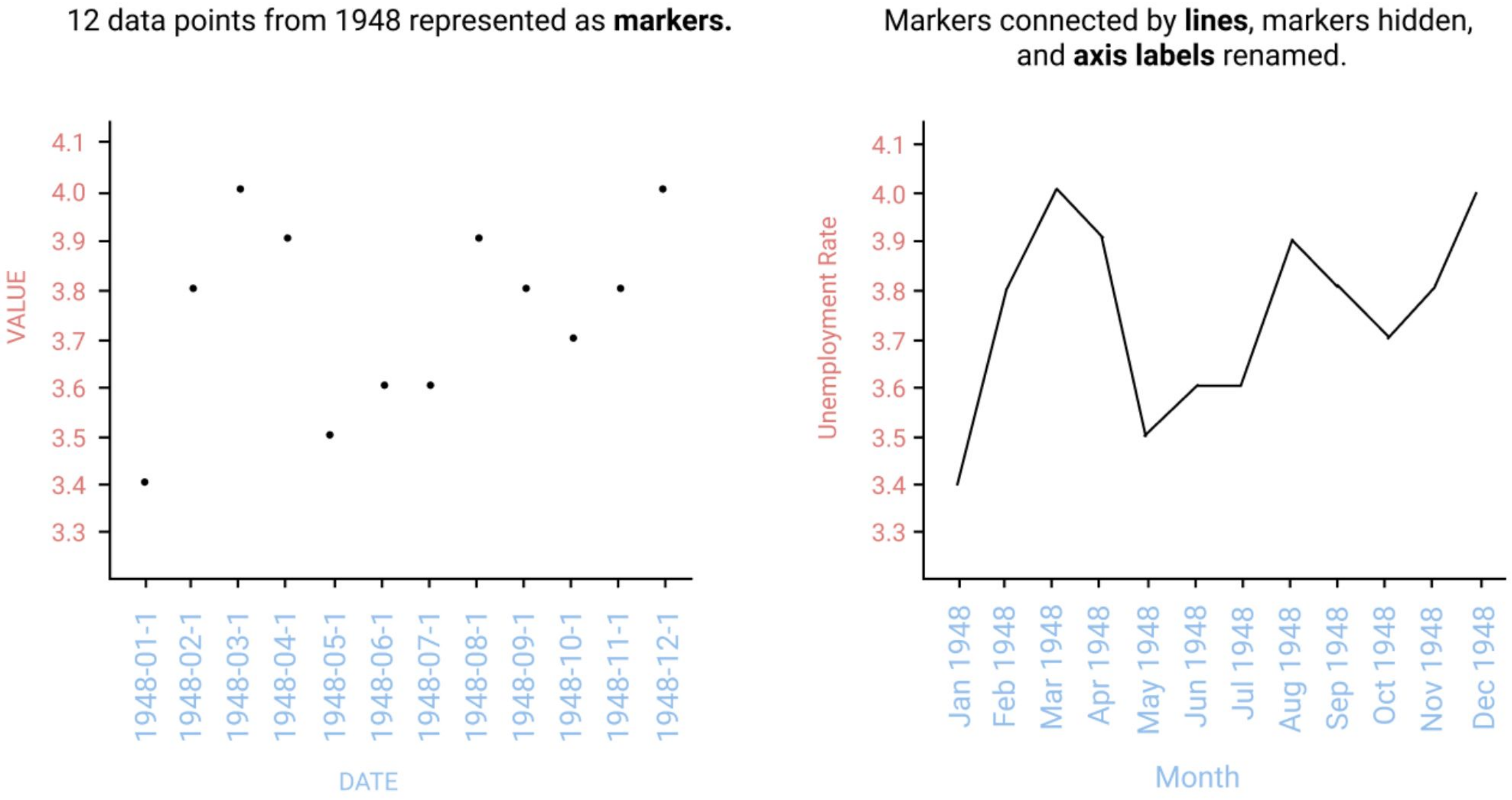
b. You have recorded 10 data points of tumor volume with treatment A and B. After looking at the underlying distributions, you'd like to plot the data to clearly show that the treatment is working.

c. You have recorded 10 data points for tumor volume *and* survival rate for tumors in treatment A and B, and you'd like to see whether tumor volume and survival rate are related.

On the whiteboard, draw your hypothetical plots.

Guidelines for data visualization in science

1. If you have raw data or distributions, inspect them *first*.
2. If you have **variability**, show it.
3. If you're making a comparison, it should be clear what you're comparing.
4. Don't connect dots unless the data is continuous.



12 data points from 1948 represented as **markers.**

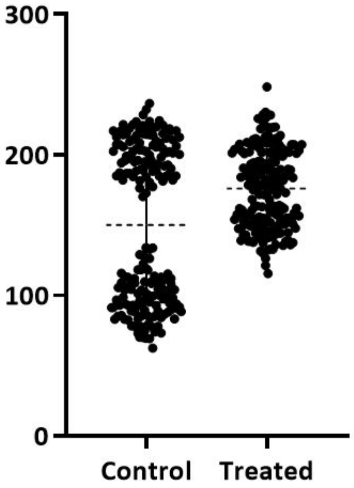
Markers connected by **lines**, markers hidden, and **axis labels** renamed.

Image: DataQuest

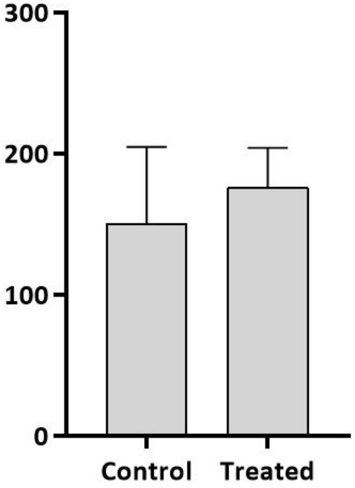
Guidelines for data visualization in science

1. If you have raw data or distributions, inspect them *first*.
2. If you have **variability**, show it.
3. If you’re making a comparison, it should be clear what you’re comparing.
4. Don’t connect dots unless the data is continuous.
5. Use **consistent colors** across multiple graphs, especially to link groups.
6. Be intentional about your color choices (see full guide here)
7. If you don’t need something on your graph to make your point, *remove it*.

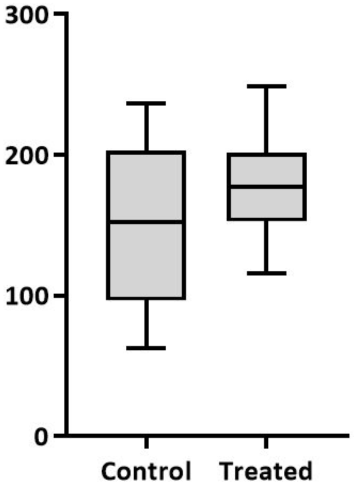
Advanced plots



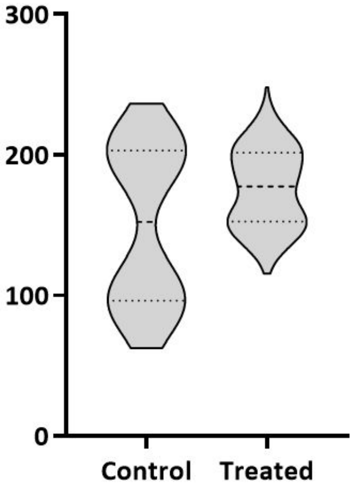
Scatter/dot plot
(good for few observations)



Bar plot



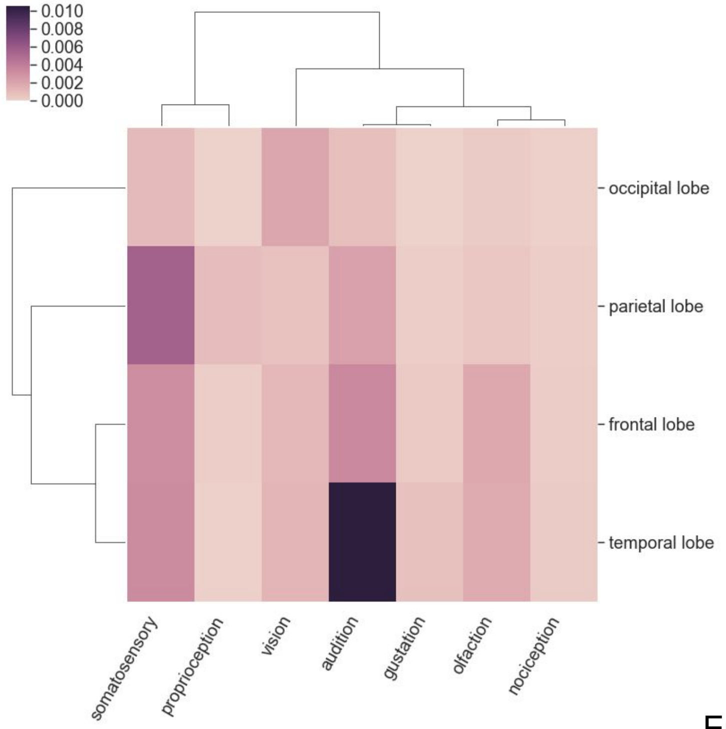
Box plot
(good for 5+ observations)



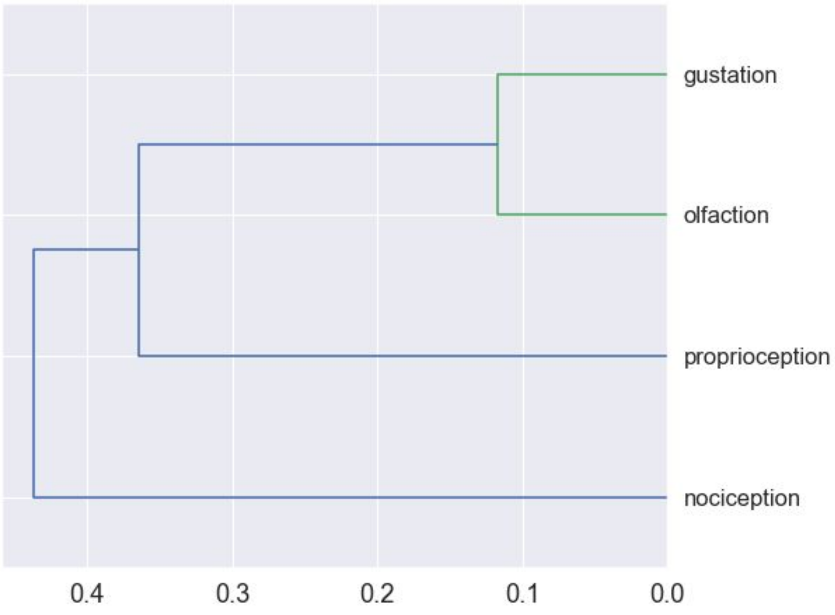
Violin plot
(good for many observations)

Advanced plots (continued)

Clustermap



Dendrogram



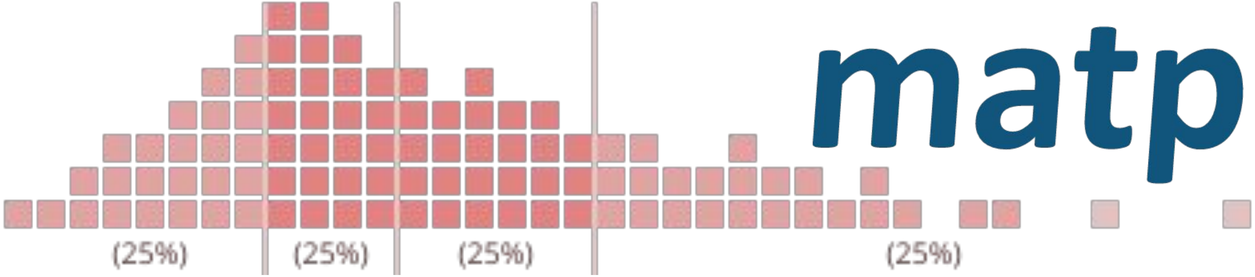
From https://lisc-tools.github.io/lisc/auto_tutorials/index.html

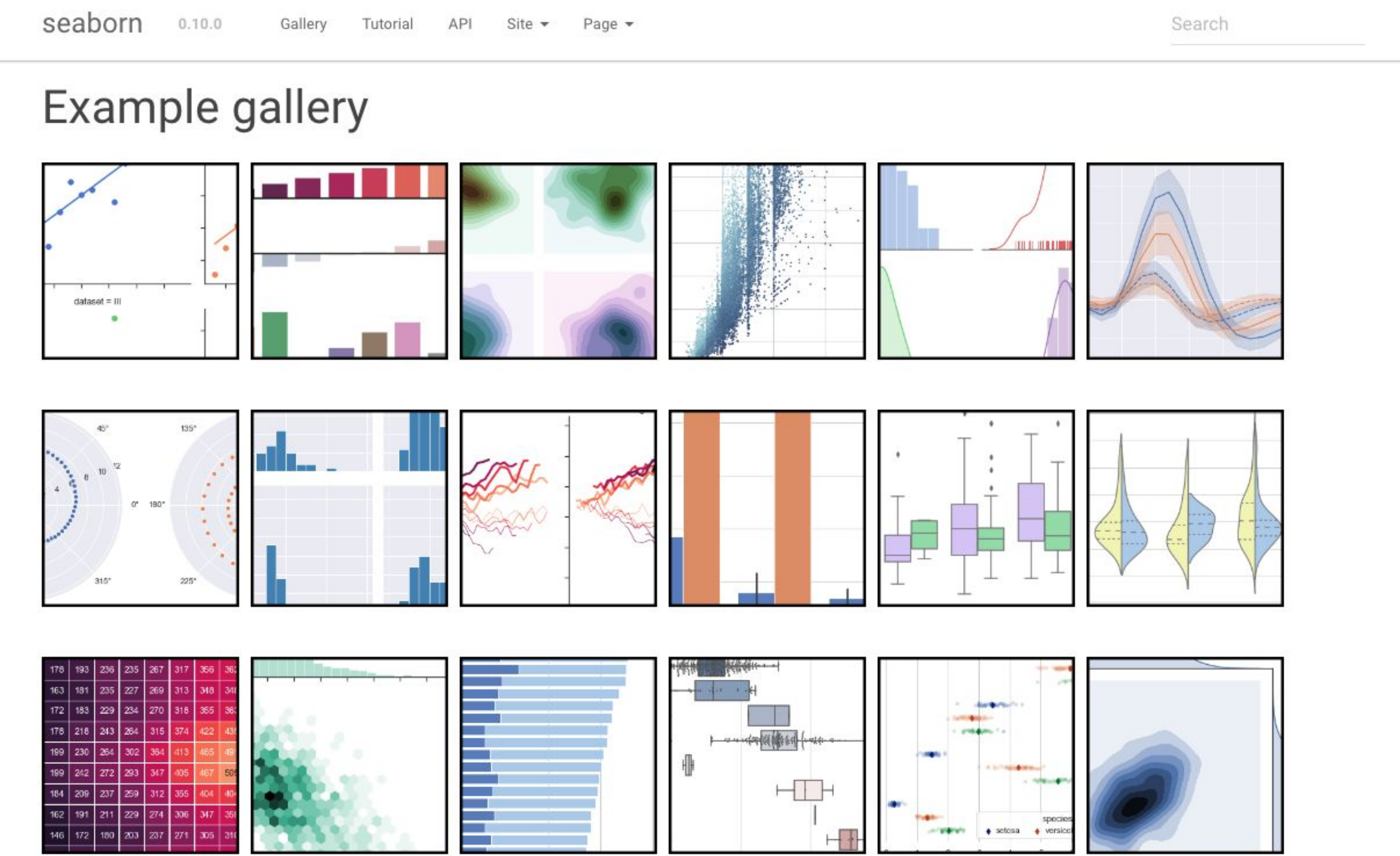
Objectives for today

- Describe best practices for data visualization
- **Introduce the tools you can use to plot in Python**
- Dive into a notebook to plot our inflammation data

There are multiple ways to plot in Python

- Matplotlib (https://matplotlib.org/index.html)
  - Call to `pyplot` module
  - Through pandas (which uses pyplot)
- Seaborn (built on top of Matplotlib; https://seaborn.pydata.org/)
  - Loved by many #dataviz folks

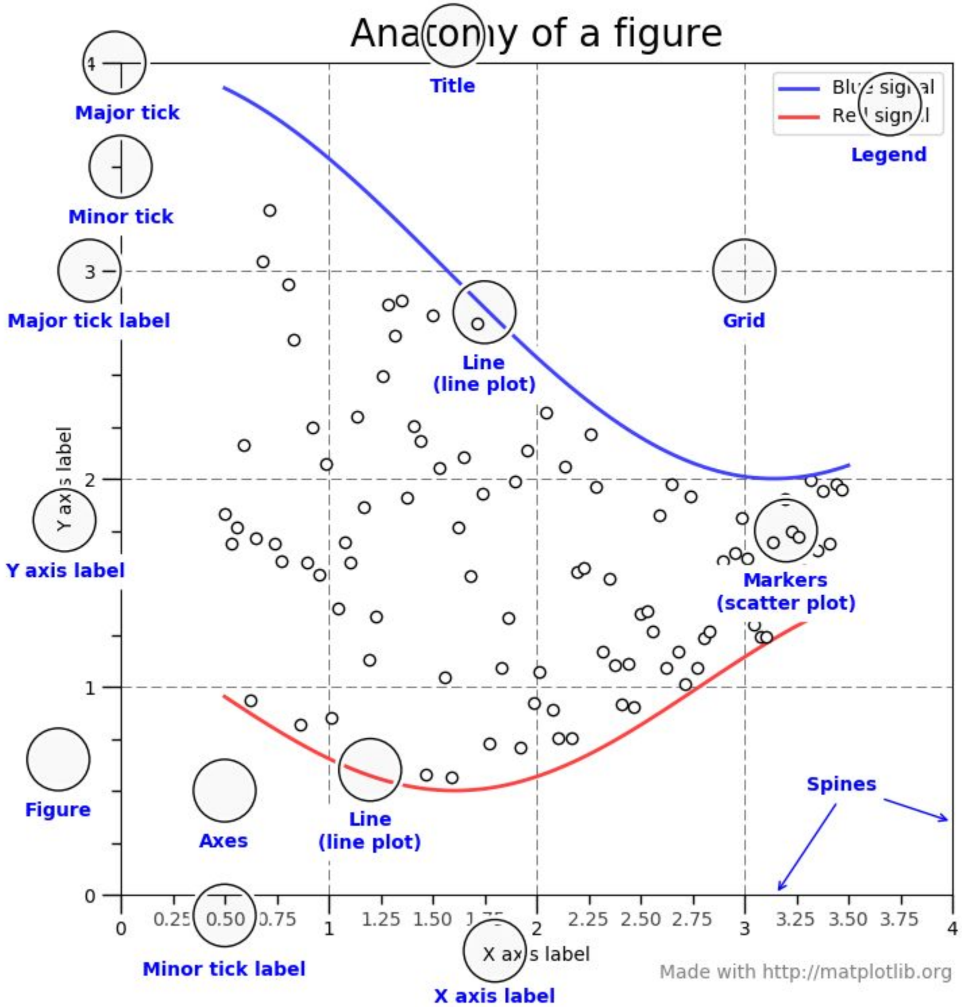




Example gallery

There are (almost) endless things you can customize on your plot

… and once you write code to do so, you can reuse it!



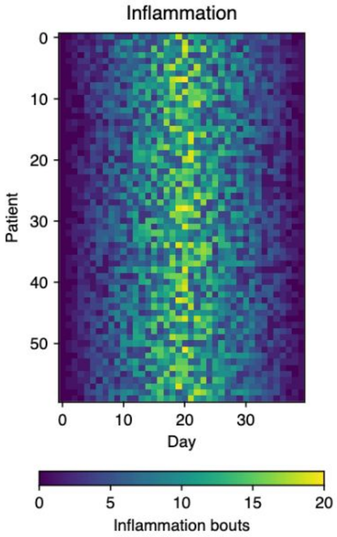
Objectives for today

- Describe best practices for data visualization
- Introduce the tools you can use to plot in Python
- **Dive into a notebook to plot our inflammation data**

Inflammation data

Patients	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
👤	0	0	1	3	1	2	4
👤	0	1	2	1	2	1	3
👤	0	1	1	3	3	2	6
👤	0	0	2	0	4	2	2
👤	0	1	1	3	3	1	3

Analysis



Conclusion

How does the medication affect patients?

Today, visualizing our data as **line charts** and a **heatmap** will help us explore trends in the data

Resources

Matplotlib Tutorial

Tableau “What is data visualization”?

Top 50 Matplotlib Data Visualizations

Towards Data Science: Python Plotting Basics