Week 3 – Systems and Toolkits

Yasmin AlNoamany

L&S 88-2

University of California, Berkeley

Announcements

- Assignment 1 grades posted
- No lecture next time
  - readings and discussions
  - assignment 3
- Semester Project - Link

Previous lecture

- Types of Web data
- Introduction to Web Mining
  - Web content mining
  - Web usage mining
  - Web structure mining
- Getting data from the Web
  - Working with APIs
    - Climate change API
    - NYT API
    - Twitter API

And what if there is no available API?

And what if there is no available API?

Scraping!

Outlines

- Getting Data from the Web
  - Web scrapping
- Choosing Tools to Visualize Data
- Data preprocessing
  - Data Cleaning
- Tools for visualizing data
  - Intro to R and RStudio
  - D3 and Tableau, next lecture.

Resources

- Data Analysis and Visualization in R from Data Carpentry lessons
- Producing Simple Graphs with R https://www.harding.edu/fmccown/r/
- Visualize This, Ch 2 (Data Scraping)
- Web Scraping in Python (notebook)
- What I Learned Recreating One Chart Using 24 Tools, by Lisa Charlotte

Objectives

- Identify how to scrape data from Web
- Understand how HTML works with your browser to display a website
- Identify HTML tags and attributes
- Differentiate between different visualization tools and describe the type of tasks for which each tool might be most appropriate
- Use R and RStudio for data analysis and visualization

Web Scraping

- Extracting data from websites by extracting it directly from the HTML source code

- Web scraping a Web page involves
  - fetching a Web page (Web crawling)
  - Extracting data from it

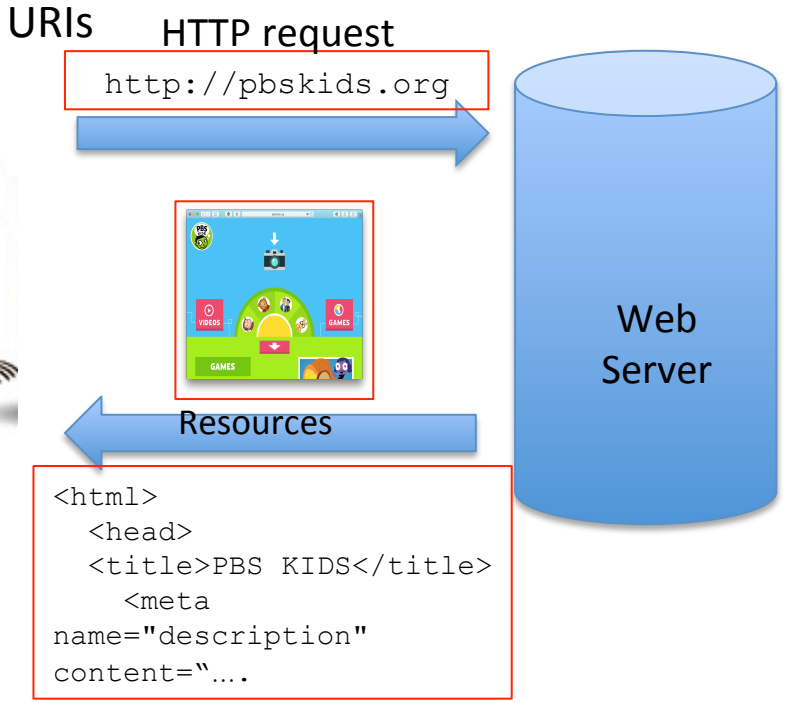
Why Web scrape

Tons of Web data useful for social scientists and humanists

- social media
- news media
- government publications
- organizational records

Basic Process

- Make a GET request
- Identify the patterns and pa
  HTML response
  - web page layout
- Iterate
  - visit each web page and grab the relevant data
  - if there are a lot of pages, this might take some time
- Store the data



Some disclaimers

- Check a site's terms and conditions before scraping
  - Some sites limit the rate of web requests you can make from a single IP
  - Be nice - don't hammer the site's server

You don’t want to get blocked!

- Web scraping etiquette
  - Only one connection should be open to any given host at a time
  - A waiting time of a few seconds should occur between successive requests to a host

Robot.txt

- Robots exclusion standard
- Web site owners use the /robots.txt file to give instructions about their site to Web robots; this is called The Robots Exclusion Protocol.
  - http://www.robotstxt.org/
- Examples:
  - https://www.facebook.com/robots.txt
  - http://www.cnn.com/robots.txt
  - http://www.ebay.com/robots.txt

```
User-agent: *
Disallow: /
The "User-agent: *" means this section applies to all robots. The
"Disallow: /" tells the robot that it should not visit any pages on
the site.
```

Slide from Michael L. Nelson at Old Dominion University

Web scraping benefits and challenges

- Web scraping benefits
  - Any content that can be viewed on a webpage can be scraped.
  - No API needed
  - No rate-limiting or authentication (usually)

- Web scraping challenges
  - Sometimes the web site structure is difficult to parse
    - might want to download the HTML pages to a local disk before parsing
  - Sites change their layout all the time. Your scraper will break.

Web scraping benefits and challenges

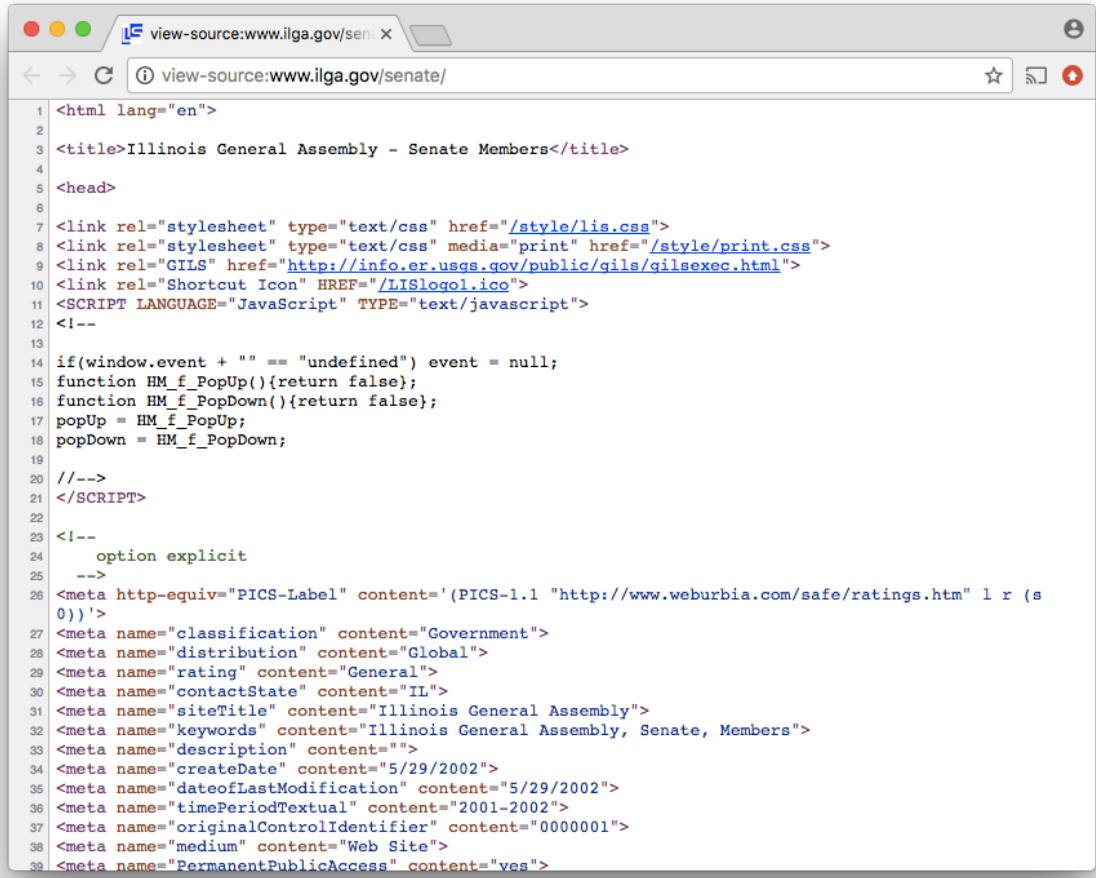
- Web scraping benefits
  - Any content that can be viewed on a webpage can be scraped.
  - No API needed
  - No rate-limiting or authentication (usually)

- Web scraping challenges
  - Sometimes the web site structure is difficult to parse
    - might want to download the HTML pages to a local disk before parsing
  - Sites change their layout all the time. Your scraper will break.

Check for API first. If not available, scrape!

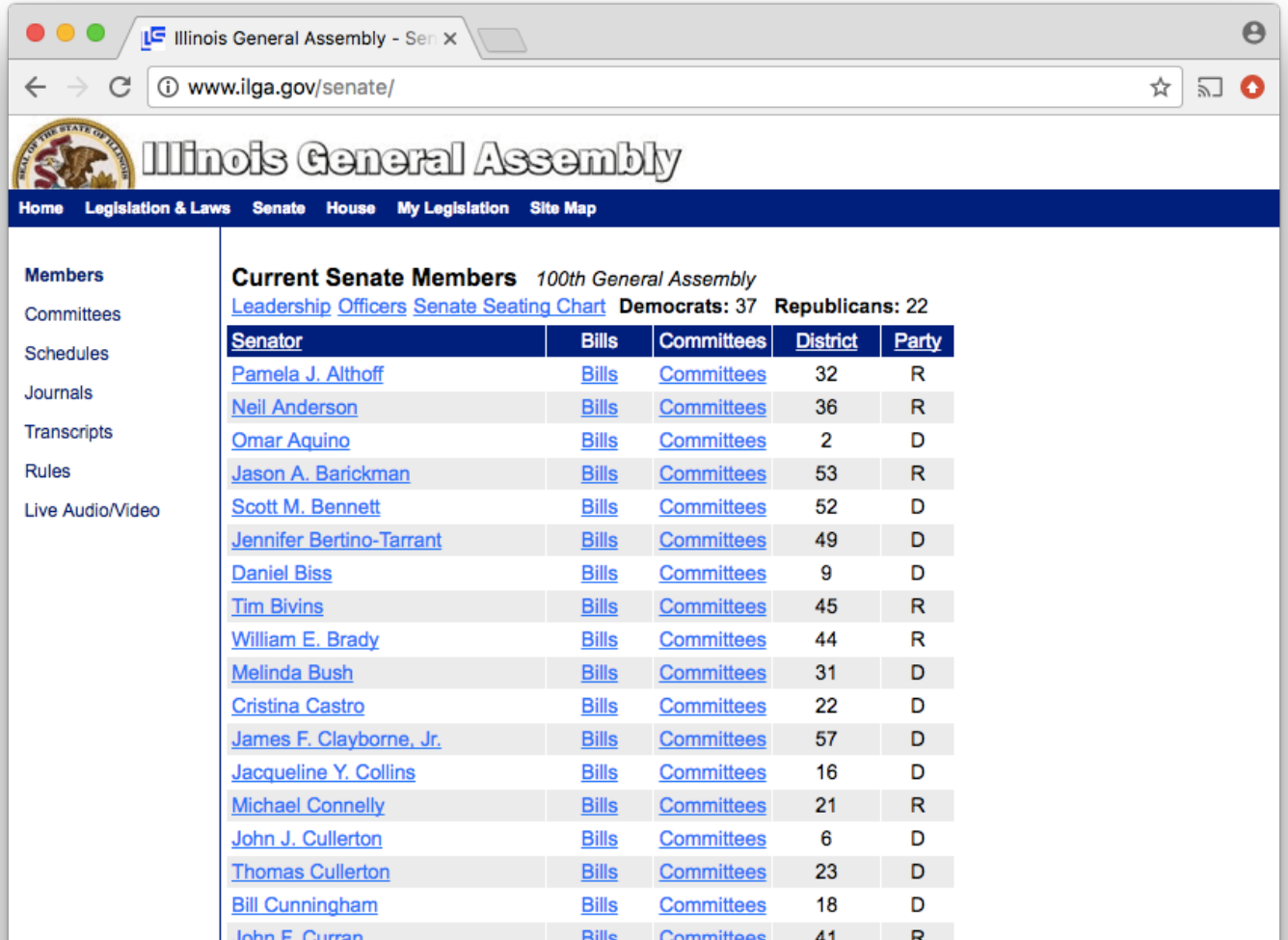
What is a Website

- Some combination of codebase, database
- The "front end" product is HTML + Cascading Style Sheets (CSS) + javascript



What is a Website

- Your browser turns that into a tidy layout



http://www.ilga.gov/senate/

Basic strategy of Web scraping:

- Find out what kind of HTML element your data is in (use your browser‘s “inspector”)
- Think about how you can differentiate those elements from other, similar elements in the webpage using CSS
- Use Python and add-on modules like BeautifulSoup to extract just that data

HTML: Basic structure

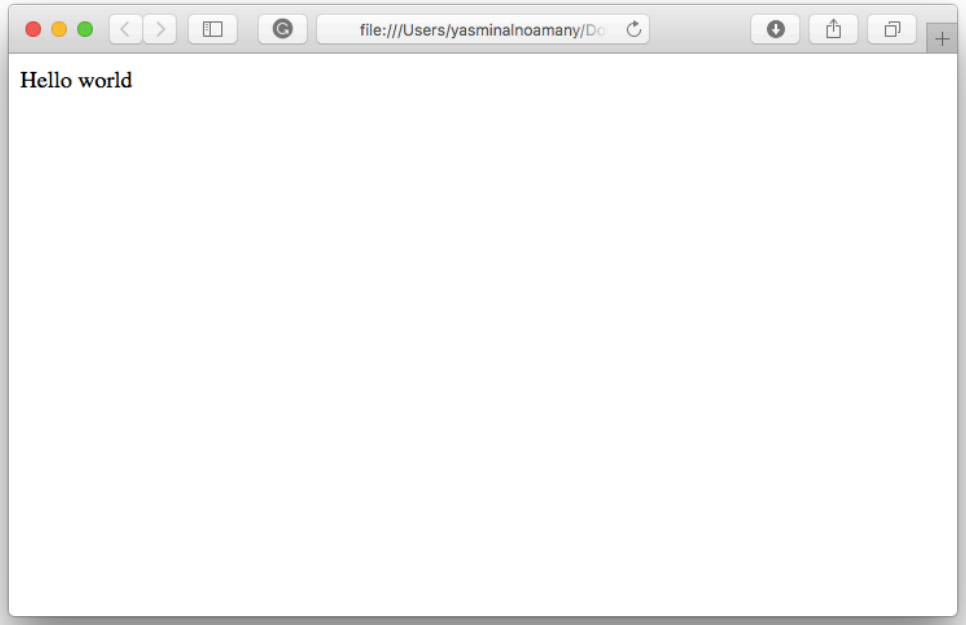
```
<DOCTYPE html>
<html>
    <head>
        <title>Page title</title>
    </head>

    <body>
        <p>Hello world</p>
    </body>
</html>
```

HTML: Basic structure

```
<DOCTYPE html>
<html>
    <head>
        <title>Page title</title>
    </head>

    <body>
        <p>Hello world</p>
    </body>
</html>
```



https://github.com/Data-on-the-Mind/2017-summer-workshop

HTML Elements

HTML element has three components:

- Tags (starting and ending the element)
- Attributes (giving information about the element)
- Text, or Content (the text inside the element)

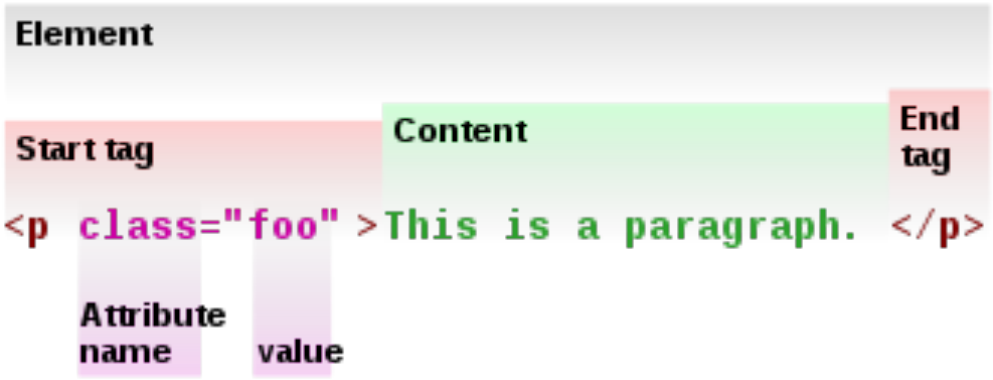


Image source: https://github.com/dlab-berkeley/python-data-from-web/blob/master/2_HTML_CSS/2_HTML_notes.md

Web Scraping Example - Check Nathan Yau, Visualize This, Ch 2

- Let's grab data from the Weather Underground
  - http://wunderground.com

- Want temperature data for Berkeley for 2011
  - starting point: http://www.wunderground.com/history/airport/ORF/ 2011/1/1/DailyHistory.html

- Tools
  - Python
  - Beautiful Soup - Python script for reading web pages (http://www.crummy.com/software/BeautifulSoup)

Example

Data Visualization Tools

Excel

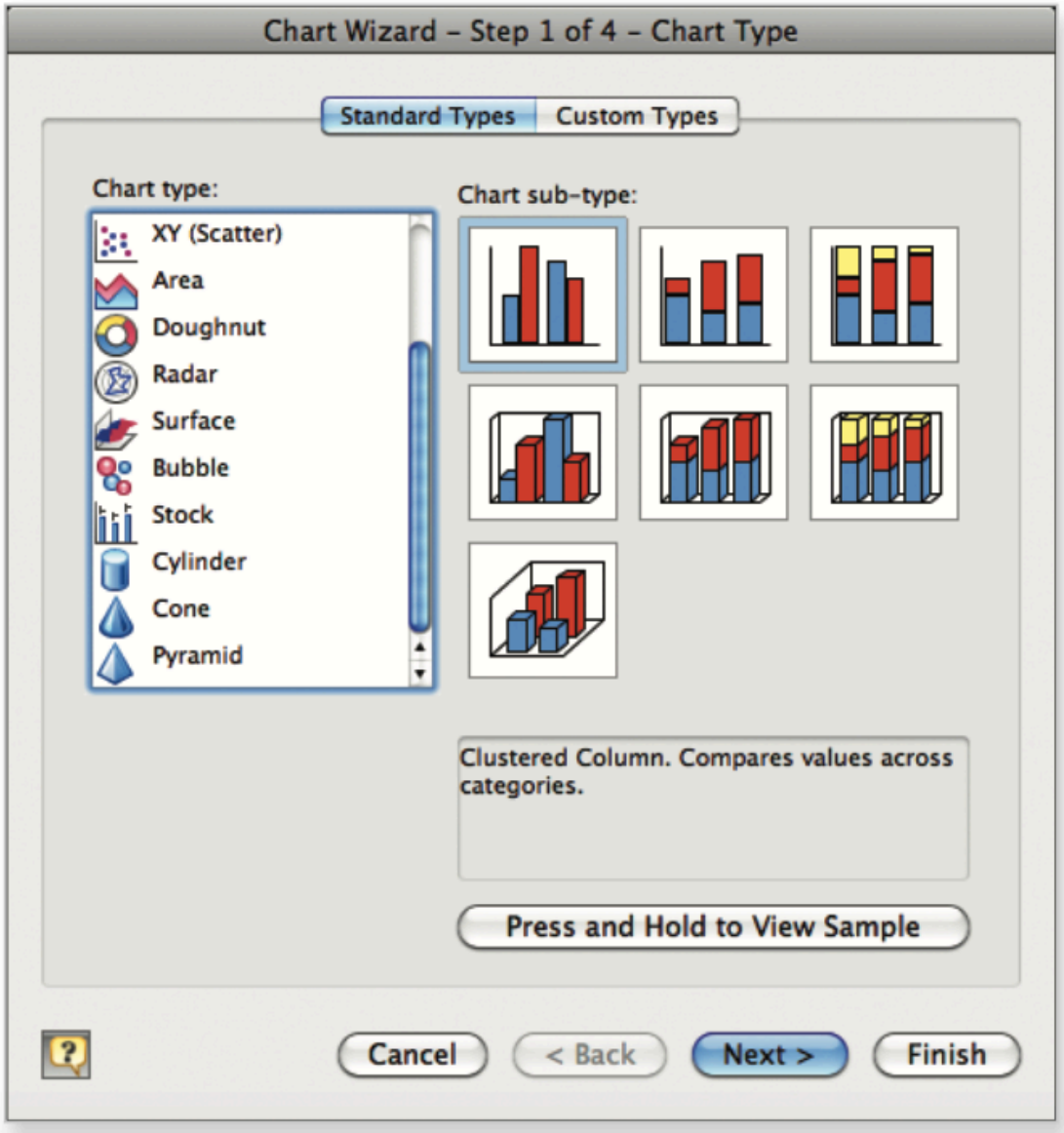
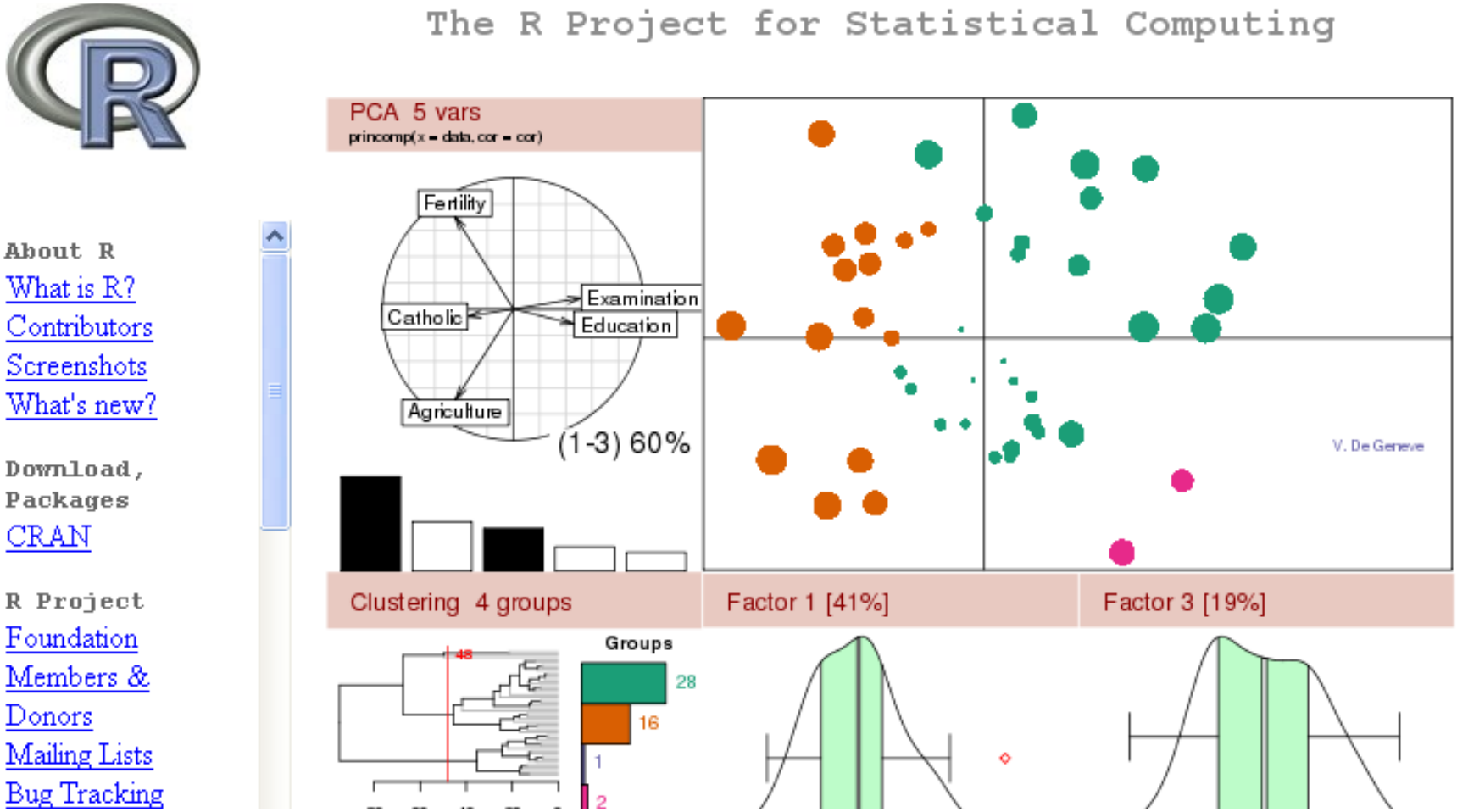


Chart Wizard – Step 1 of 4 – Chart Type
Standard Types
Custom Types
Chart type:
XY (Scatter)
Area
Doughnut
Radar
Surface
Bubble
Stock
Cylinder
Cone
Pyramid
Chart sub-type:
Clustered Column. Compares values across categories.
Press and Hold to View Sample
Cancel
< Back
Next >
Finish

R



http://pluto.huji.ac.il/~msby/StatThink/Screenshots_WinXP/1-R-project-HomePage.PNG

Python - Matplotlib

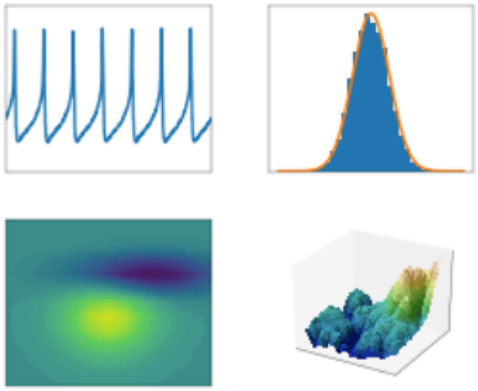
matplotlib

Fork me on GitHub

home | examples | gallery | pyplot | docs » modules | index

Introduction

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.



Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For a sampling, see the screenshots, thumbnail gallery, and examples directory

Depsy 100th percentile

Travis-CI: build error

Support matplotlib

Support NumFOCUS

Quick search

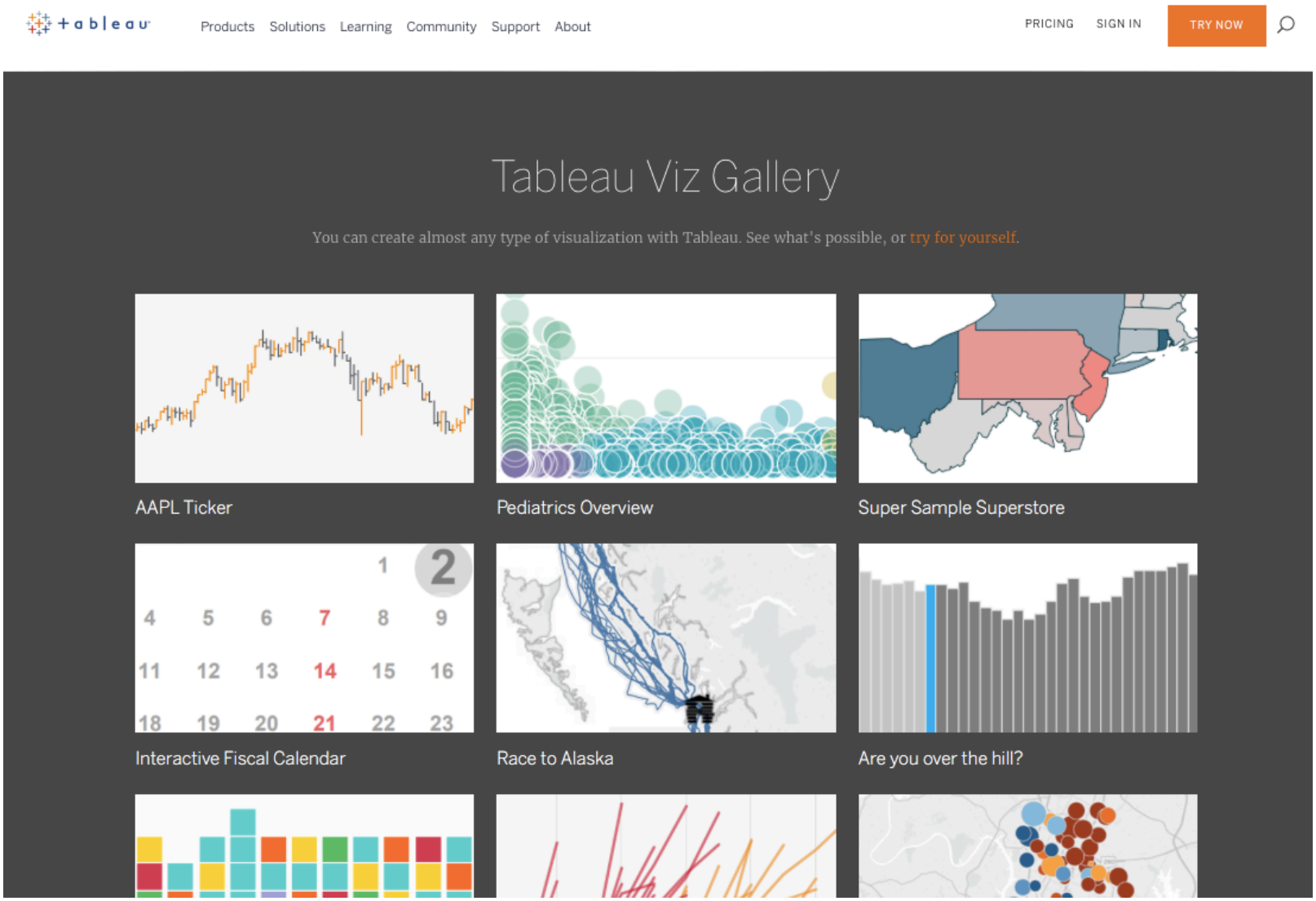
Go

https://matplotlib.org

D3

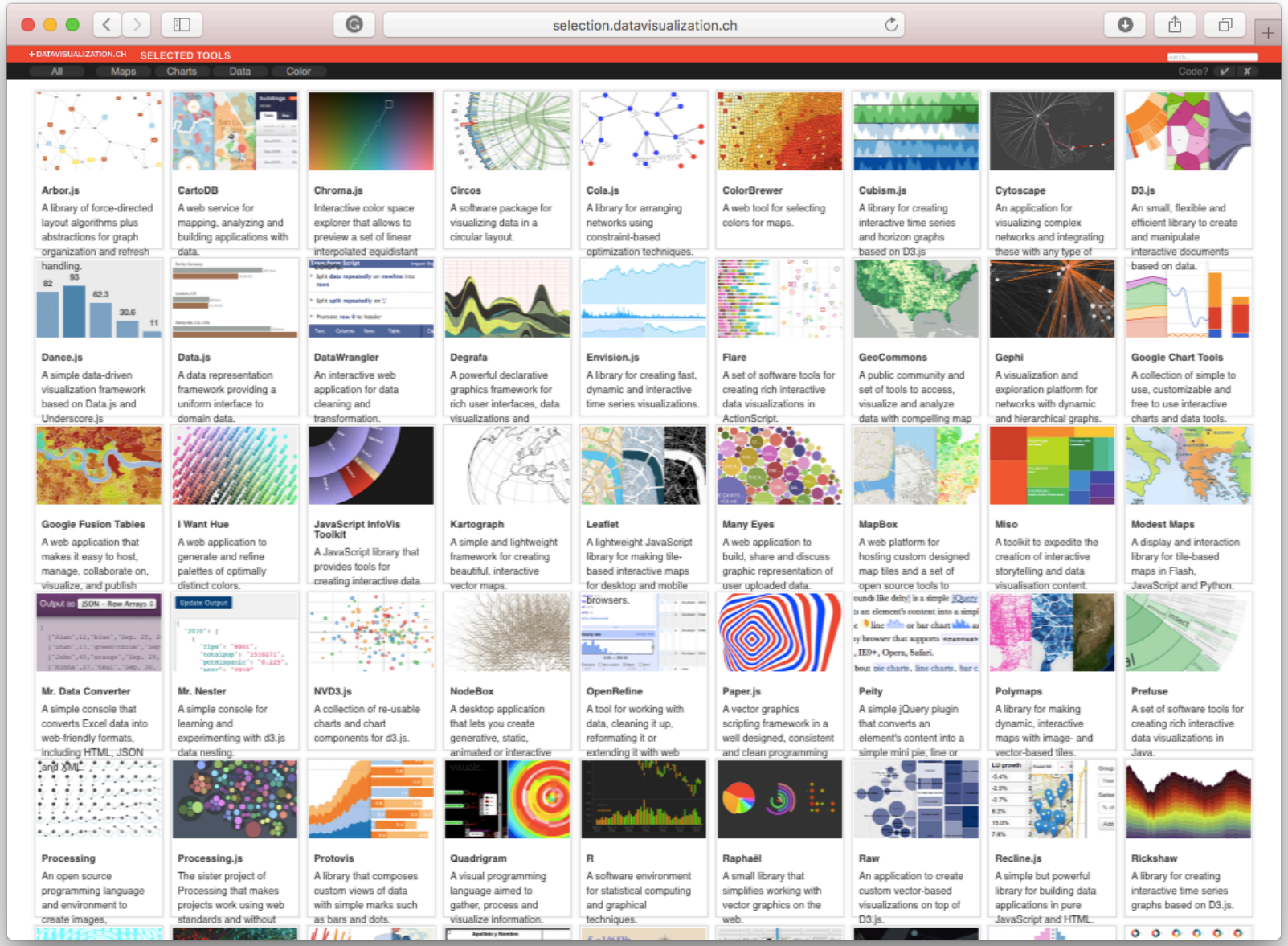
https://d3js.org

Tableau



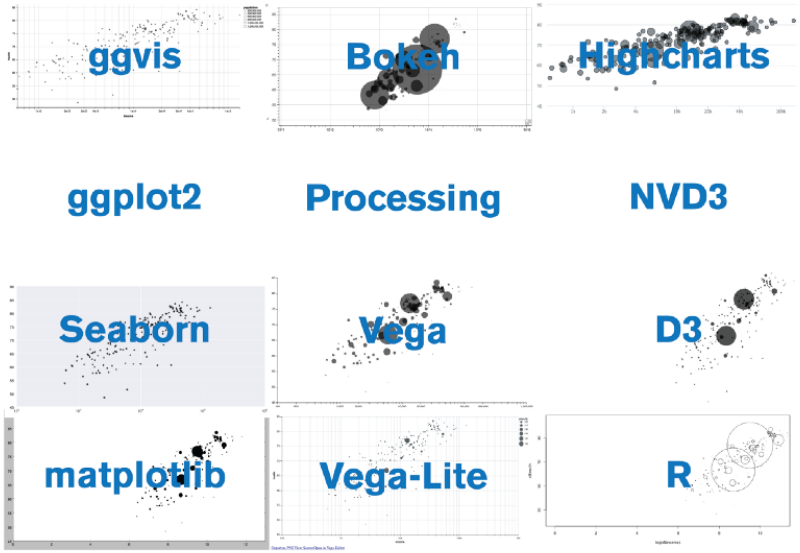
https://www.tableau.com/solutions/gallery

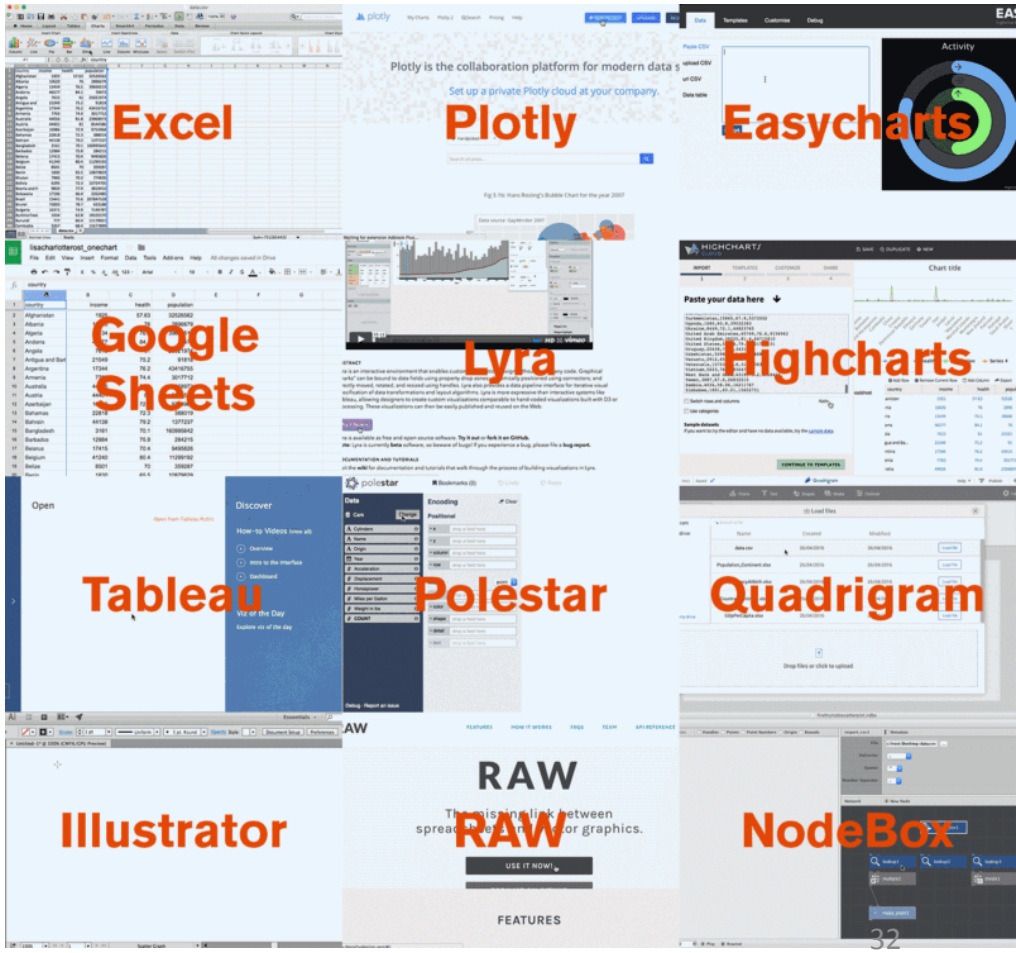
Many, many others



http://selection.datavisualization.ch

What I Learned Recreating One Chart Using 24 Tools

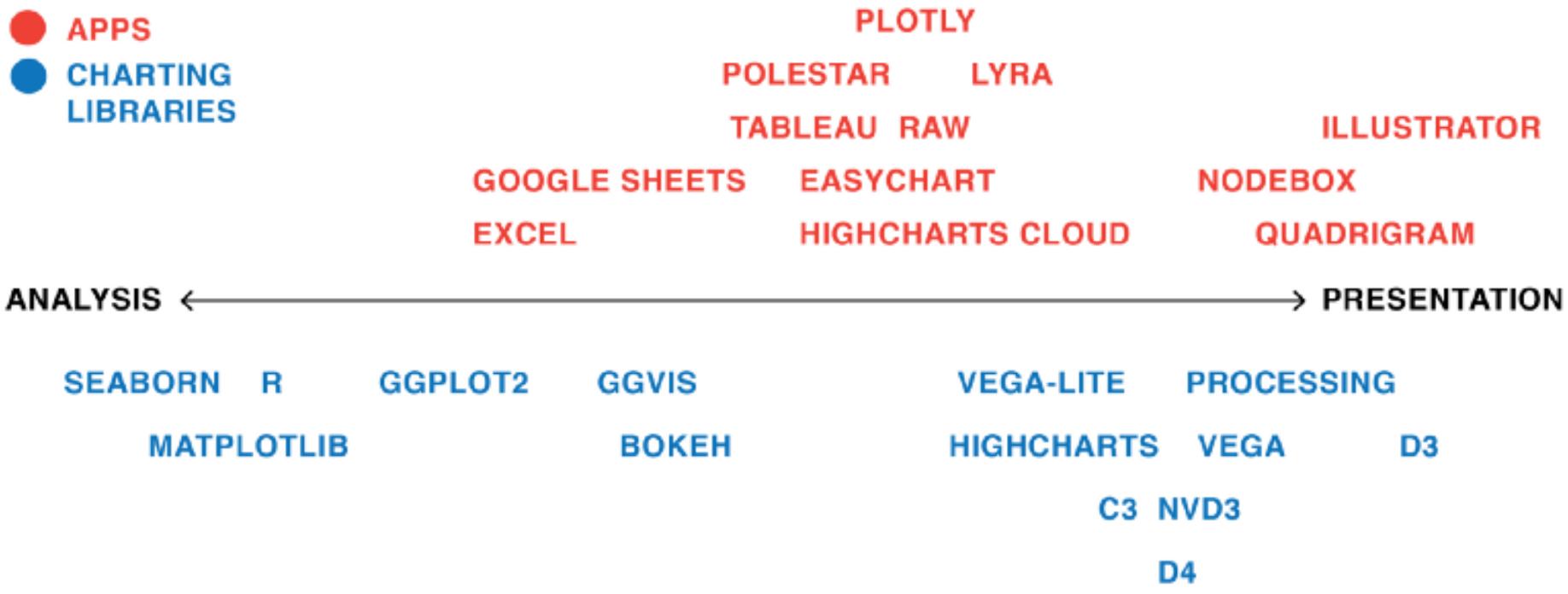




https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

Finding the best tool means thinking hard about your goals and needs.

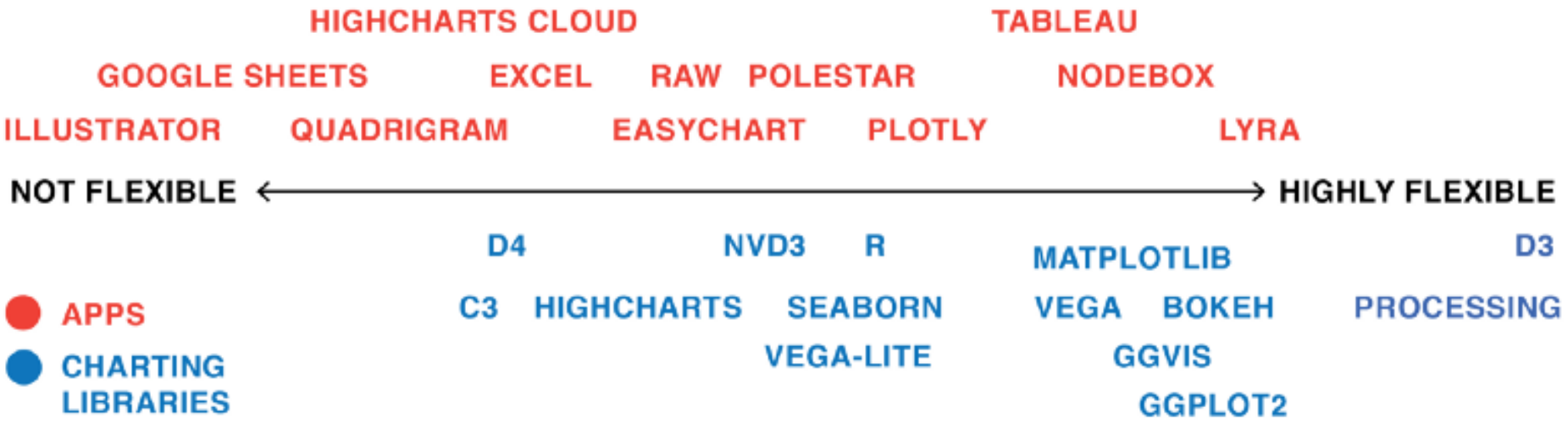
Tools for Analysis vs. Presentation



Do you want to use your tool to explore the data (R, Python) or do you want to build visualizations for the public (D3.js, Illustrator)?

https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

Flexibility of Tools



Do you just need basic chart types like a bar chart or line chart (Highcharts, Excel); or do you want to create crazy chart magic (D3.js)?

https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

Data Management

- Will you change your data (eg. changing one or all values or adding rows or columns) while creating a data visualization?
  - The least convenient in some apps (like Illustrator) you need to start building the graph all over again when changing the data only slightly
  - The medium-convenient way is to change the data outside of the tool, then import it (again) and update the visualization with the new data. D3.js is an example for this way
  - Other apps make it even more convenient than code to deal with data. Once you import your data, you can change it or add new, transformed columns directly in the tool (examples: Plotly and Lyra)

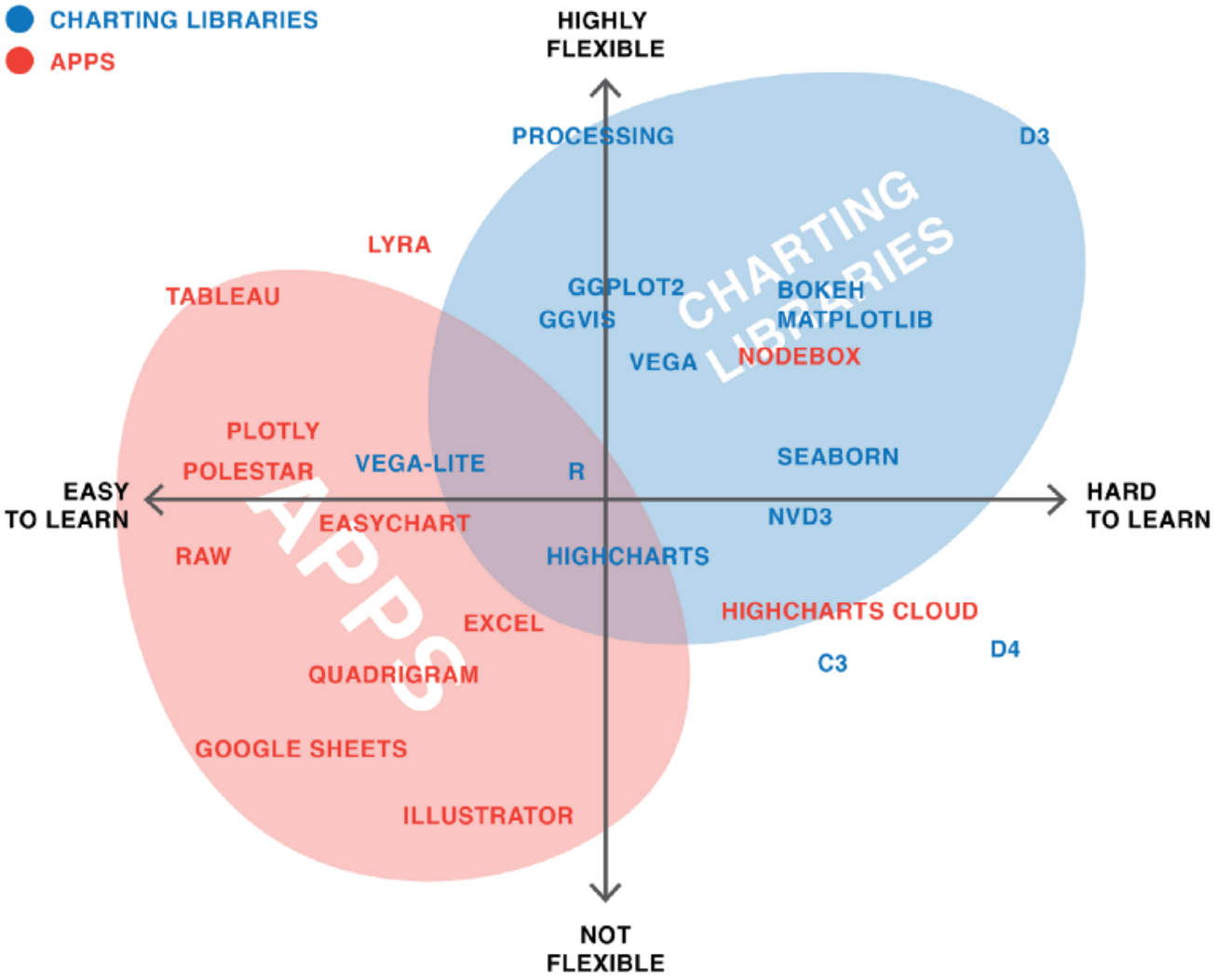
https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

Interactivity vs. static

	STATIC	WEB - INTERACTIVE
APPS	ILLUSTRATOR, NODEBOX, EXCEL, POLESTAR, RAW	HIGHCHARTS CLOUD, QUADRIGRAM, EASYCHRT, DATAWRAPPER, TABLEAU, PLOTLY, GOOGLE SHEETS
CHARTING LIBRARIES	GGPLOT2, MATPLOTLIB, R, SEABORN, BOKEH, PROCESSING	D3, D4, C3, NVD3, GGVIS, HIGHCHARTS, SHINY, VEGA, VEGA-LITE

https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

We Still Live in an "Apps Are for the Easy Stuff, Code Is for the Good Stuff" World



There Are No Perfect Tools, Just Good Tools for People with Certain Mindsets

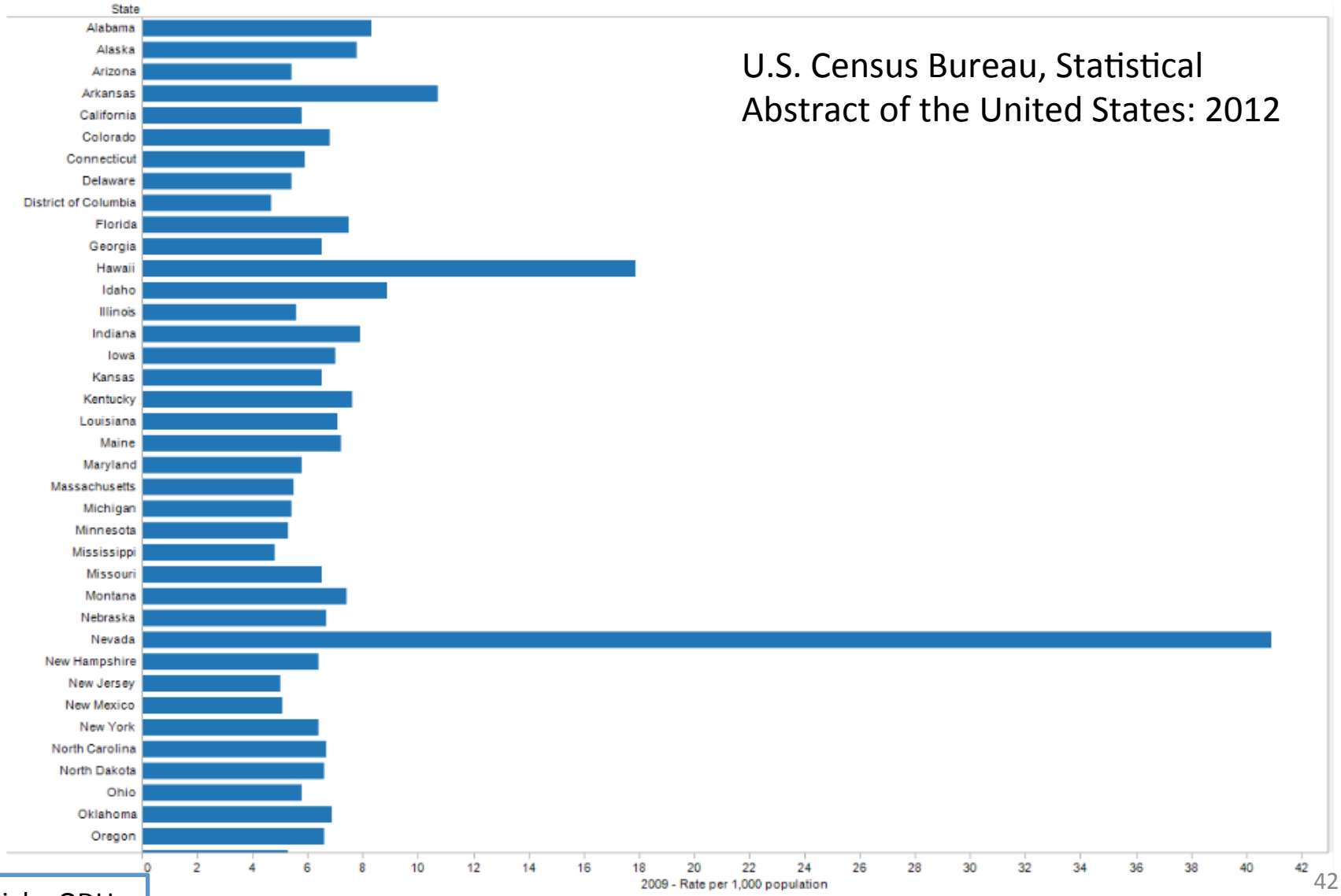
https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

Data Cleaning

Data cleaning

- Data in the real world tend to be:
  - Noisy: containing errors or outliers
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - Inconsistent: containing discrepancies in codes or names
- Data cleaning tend to fill in missing values, smooth out noise while identifying outliers, and correct inconsistences in data
- Before you can visualize something, you have to get the data into a form that you can work with

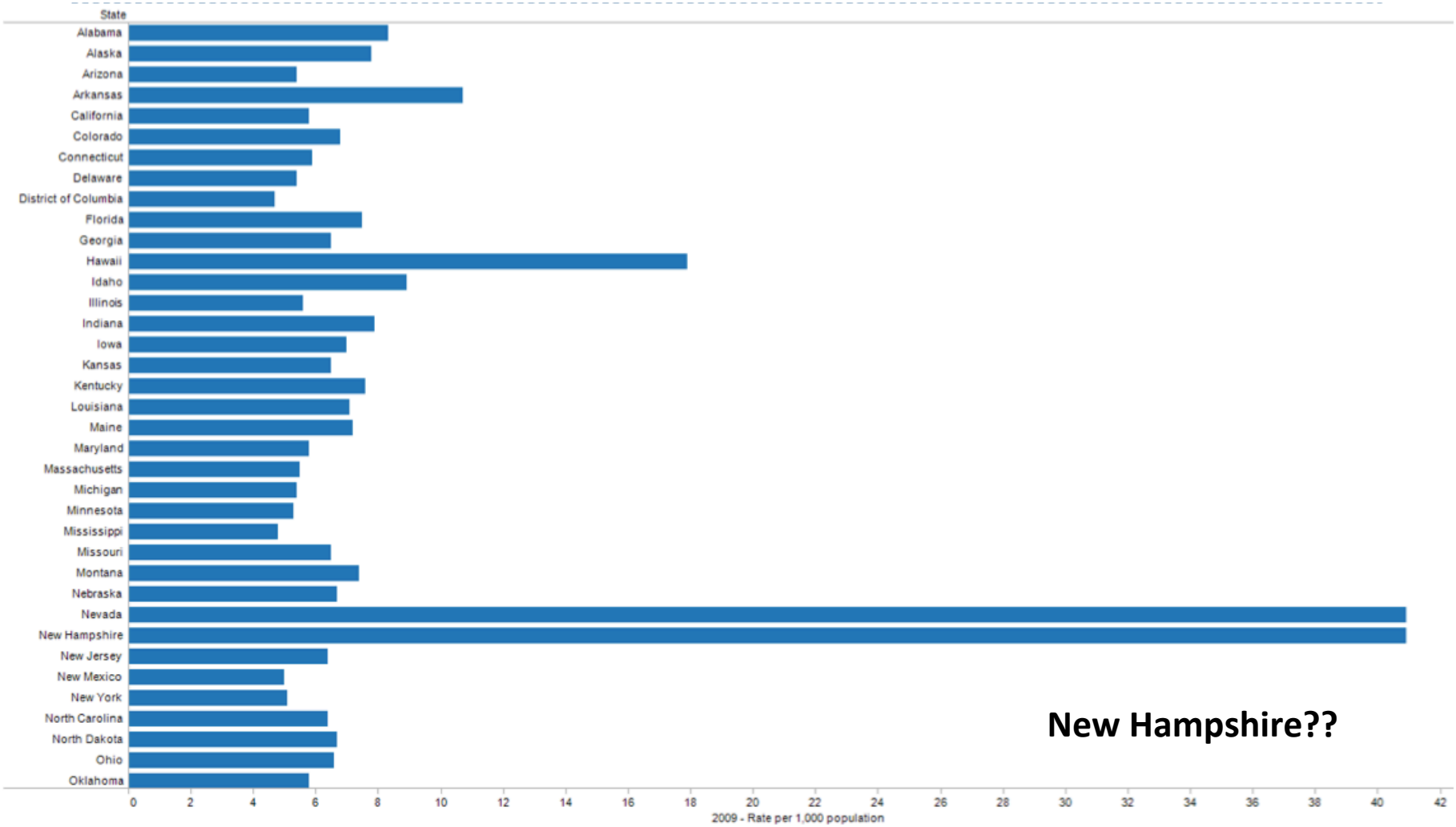
Marriage Rate per State



U.S. Census Bureau, Statistical Abstract of the United States: 2012

http://www.census.gov/compendia/statab/2012/tables/12s0133.pdf

Graphing the raw data



Now, let's look at the PDF

	Number (1,000)			Rate per 1,000 population [2]		
	1990	2000	2009	1990	2000	2009
Maryland	46.3	40.0	32.4	9.7	7.5	5.8
Massachusetts	47.7	37.0	36.7	7.9	5.8	5.5
Michigan	76.1	66.4	53.1	8.2	6.7	5.4
Minnesota	33.7	33.4	28.4	7.7	6.8	5.3
Mississippi	24.3	19.7	14.5	9.4	6.9	4.8
Missouri	49.1	43.7	39.8	9.6	7.8	6.5
Montana	6.9	6.6	7.1	8.6	7.3	7.4
Nebraska	12.6	13.0	12.5	8.0	7.6	6.7
Nevada	120.6	144.3	108.2	99.0	72.2	40.9
New Hampshire	10.5	11.6	8.5	9.5	9.4	6.4
New Jersey	58.7	50.4	46.3	7.6	6.0	5.0
New Mexico [5]	13.3	14.5	10.2	8.8	8.0	5.1
New York [5]	154.8	162.0	120.1	8.6	7.1	6.4
North Carolina	51.9	65.6	65.8	7.8	8.2	6.7
North Dakota	4.8	4.6	4.3	7.5	7.2	6.6
Ohio	98.1	88.5	64.8	9.0	7.8	5.8
Oklahoma	33.2	15.6	23.5	10.6	(NA)	6.9
Oregon	25.3	26.0	23.5	8.9	7.6	6.6
Pennsylvania	84.9	73.2	64.2	7.1	6.0	5.3
Rhode Island	8.1	8.0	6.5	8.1	7.6	5.9

http://www2.census.gov/library/publications/2011/compendia/statab/131ed/tables/vitstat.pdf

Marriage Rate per State

	Rate 1,000 population \2						
	1990	2000	2005	2006	2007	2008	2009
Maryland	9.7	5.8	6.9	6.6	6.5	6.0	5.8
Massachusetts	7.9	6.7	6.2	5.9	5.9	5.7	5.5
Michigan	8.2	6.8	6.0	5.9	5.9	5.5	5.4
Minnesota	7.7	6.9	5.9	6.0	6.0	5.5	5.3
Mississippi	9.4	7.8	5.8	5.7	5.7	5.2	4.8
Missouri	9.6	7.3	7.0	6.9	6.9	6.8	6.5
Montana	8.6	7.6	7.4	7.5	7.5	7.7	7.4
Nebraska	8.0	72.2	7.0	6.8	6.8	6.9	6.7
Nevada	99.0	9.4	57.8	52.6	52.6	43.1	40.9
New Hampshire	9.5	6.0	7.2	7.1	7.1	6.8	40.9
New Jersey	7.6	6.0	5.7	5.5	5.5	5.4	6.4
New Mexico \5	8.8	8.0	6.6	6.9	6.9	4.0	5.0
New York \5	8.6	7.1	6.8	6.8	6.8	6.5	5.1
North Carolina	7.8	8.2	7.3	7.3	7.3	7.0	6.4
North Dakota	7.5	7.2	6.9	6.8	6.8	6.7	6.7
Ohio	9.0	7.8	6.5	6.3	6.3	6.0	6.6
Oklahoma	10.6	(NA)	7.3	7.3	7.3	7.1	5.8
Oregon	8.9	7.6	7.3	7.2	7.2	6.9	6.9
Pennsylvania	7.1	6.0	5.8	5.7	5.7	5.6	5.3
Rhode Island	8.1	7.6	7.0	6.5	6.5	6.2	5.9

http://www2.census.gov/library/publications/2011/compendia/statab/131ed/tables/12s0133.xls

Bottom Line

- If you see something weird in your graph that you can't explain, go back and double-check your data

- Even if you didn't make an error, maybe someone else did

How to clean data

- Manually!
- Interactively with data wrangling tools
- Batch processing through scripting

Data Wrangling Tools

- **Open Refine (was Google refine)**
  - http://openrefine.org
  - video: http://www.youtube.com/watch? v=yNccGtn3Wb0&feature=player_ embedded
- **Mr. People**
  - http://people.ericson.net/
  - formats lists of names
- **Mr. Data Converter**
  - http://shancarter.com/ data_converter/
- **Data Wrangler**
  - http://vis.stanford.edu/wrangler/
  - video: http://vimeo.com/19185801
- **Data Science Toolkit**
  - http://www.datasciencetoolkit.org/
  - lots of quick conversion tools

Detect inconsistency in WDV survey data

- Look at the data carefully
  1. Identify all of the inconsistencies in the major column.
  2. Identify other problem you see in the data.
  3. What problems might these inconsistencies/incompleteness/ noise cause when trying to analyze or visualize the data?
  4. If the complete dataset was only these 40 entries, how would you go about cleaning it? (Assuming that you had either a CSV or Excel file.) Would your strategy change if you had the full dataset of 2128 entries?

- Add you answer in this document
  https://goo.gl/ejsCac

Data visualization in R

References

- Data Analysis and Visualization in R from Data Carpentry lessons http://www.datacarpentry.org/R-ecology-lesson/index.html
- Producing Simple Graphs with R https://www.harding.edu/fmccown/r/