CS 625
Data Visualization
Fall 2019

Intro to Data Visualization

Dr. Michele C. Weigle

http://www.cs.odu.edu/~mweigle/CS625-F19/

Today

- *Before class*
  - Reading: Ch 1 - What's Vis, and Why Do It?
- *During class*
  - Highlight and discuss Ch 1
  - ***We will not cover everything that you are responsible for during class time.***

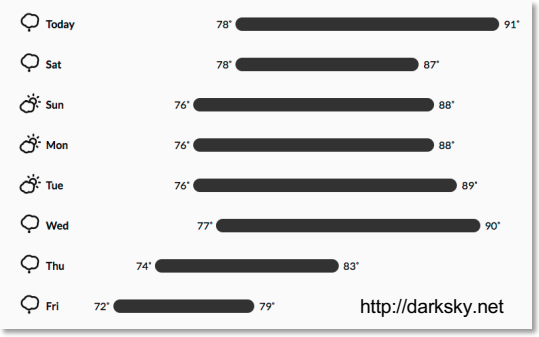
Topic Objectives

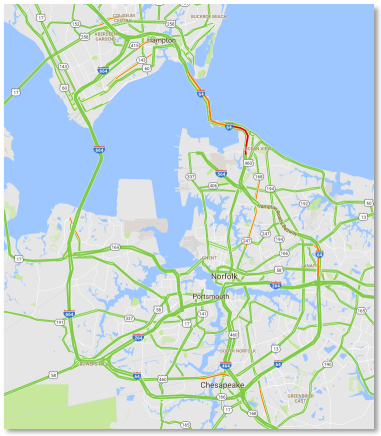
- Define visualization.
- Explain the importance of humans in the visualization process.
- Explain why human vision is particularly well-suited for information transfer.
- Give an example of a visualization idiom.
- Explain why it is best to consider multiple alternatives for vis before selecting a solution.
- Explain at a high-level the "what-why-how" framework for analyzing visualization use.
- Differentiate between R, D3, and Tableau and describe the type of tasks for which each tool might be most appropriate.

What is visualization?

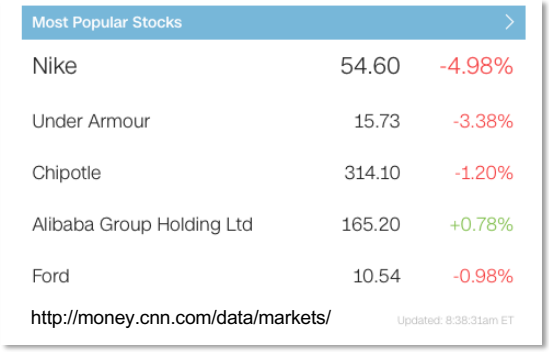
- "The communication of information using graphical representations"
  - Ward, Grinstein, Keim
- "The use of computer-supported *interactive* visual representations of data to amplify *cognition*"
  - Card, Mackinlay, Shneiderman, *Readings in Information Visualization: Using Vision to Think*
- "The purpose of visualization is insight, not pictures."
  - Ben Shneiderman

Where have you seen a visualization today?





http://google.com/maps/



Most Popular Stocks		
Nike	54.60	-4.98%
Under Armour	15.73	-3.38%
Chipotle	314.10	-1.20%
Alibaba Group Holding Ltd	165.20	+0.78%
Ford	10.54	-0.98%

http://money.cnn.com/data/markets/

Updated: 8:38:31am ET

What's vis?

- Visualization is suitable when there is a **need to augment human capabilities** rather than replace people with computational decision-making methods.
- The **design space of possible vis idioms is huge**, and includes the considerations of both **how to create** and **how to interact** with visual representations.
- Vis design is full of **tradeoffs**, and **most possibilities in the design space are ineffective** for a particular task, so **validating the effectiveness** of a design is both necessary and difficult.
- Vis designers must take into account three very different kinds of **resource limitations**: those of computers, of humans, and of displays.
- Vis usage can be **analyzed in terms of why** the user needs it, **what** data is shown**, and how** the idiom is designed.

Munzner, pg. 1

Visualization is suitable when there is a **need to augment human capabilities** rather than replace people with computational decision-making methods.

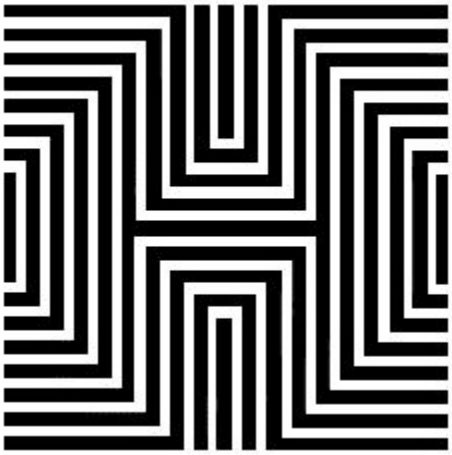
augment human capabilities

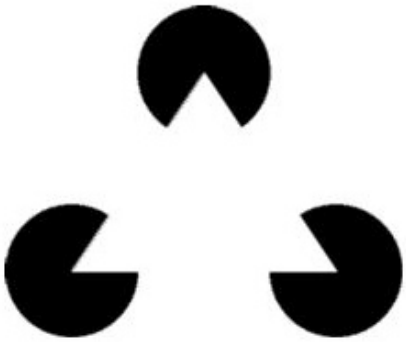
Why have a human in the loop?

- Vis allows people to analyze data when they don't know exactly what questions to ask in advance.

- Best path - put a human in the loop
  - exploit the pattern detection properties of human vision

Humans are great at pattern recognition

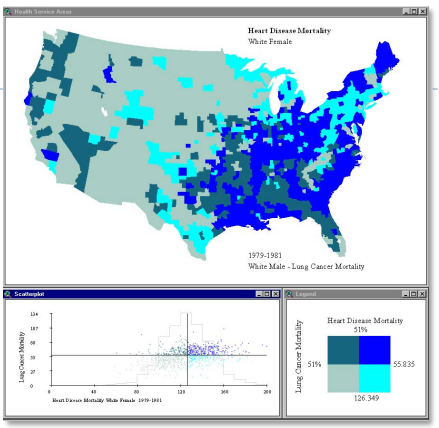




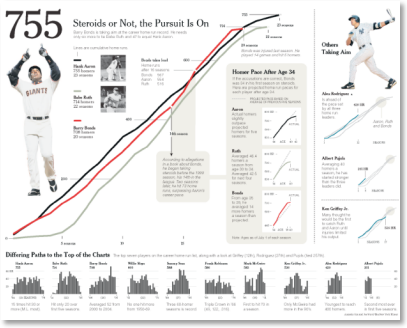
Create visualizations that lets computers do what computers do well and lets humans do what humans do well.

Uses of vis tools

- Transitional
  - vis works itself out of a job
- Long-term
  - exploratory analysis
- Presentation
  - visual explanations

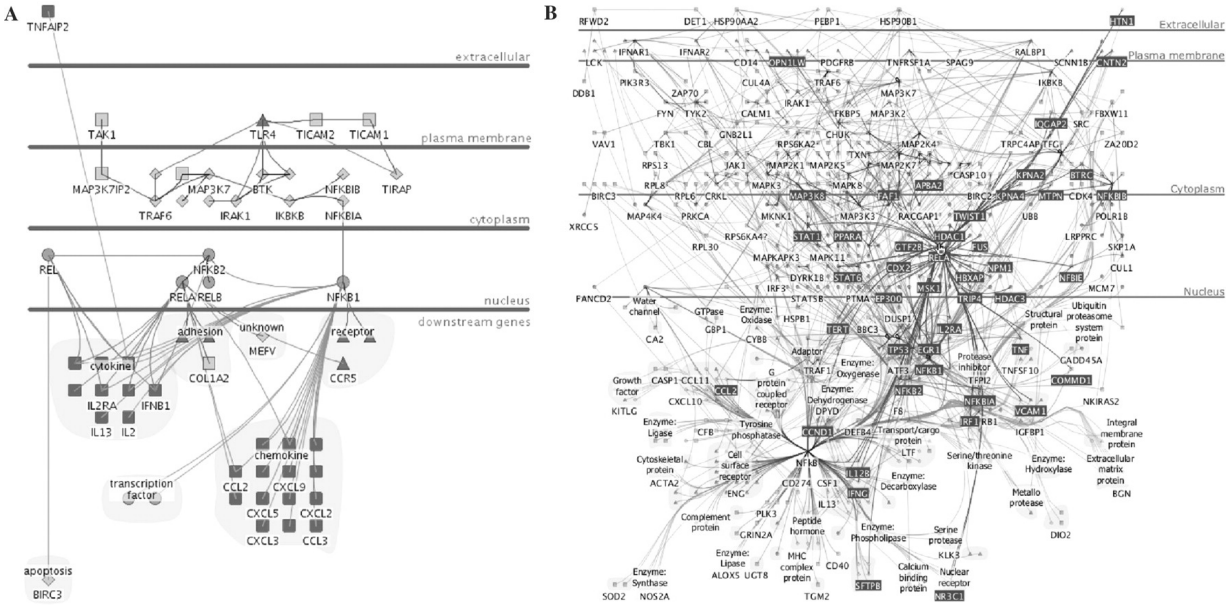


https://www.geovista.psu.edu/research/healthvisualization/



New York Times, 2006

Why have a computer in the loop?



Munzner, Fig 1.2 (Barsky et al., 2007)

Why use an external representation?

- Vis allows people to offload cognition and memory usage to make space for other operations.

- Diagrams as external representations
  - information can be organized by spatial location
    - search - grouping items needed for problem-solving in one location
    - recognition - grouping relevant info for one item in the same location

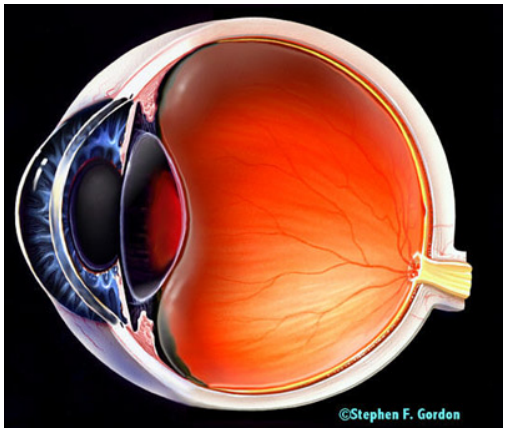
Visualization can extend your memory

What is 57 x 48?

paper	mental buffer
2	
57	[8 * 7 = 56]
x 48	[8 * 5 = 40 + 5 = 45]
1	
456	[4 * 7 = 28]
228	[4 * 5 = 20 + 2 = 22]
2736	[8 + 5 = 13]
	[4 + 2 + 1 = 7]

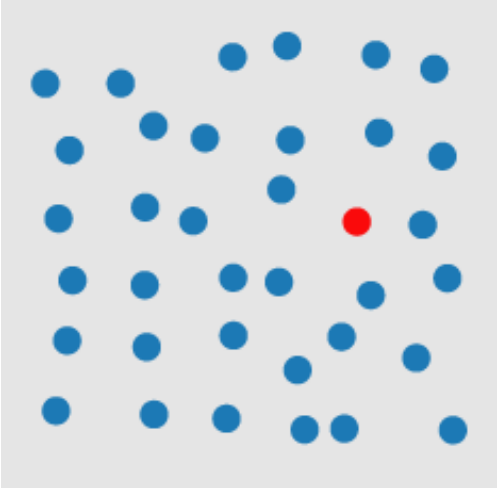
Example courtesy Tamara Munzner, Univ. of British Columbia

Why depend on vision?



- Visual system provides a high-bandwidth channel to our brains.
- Significant amount of visual information processing occurs in parallel at the pre-conscious level.

Can you find the red dot?



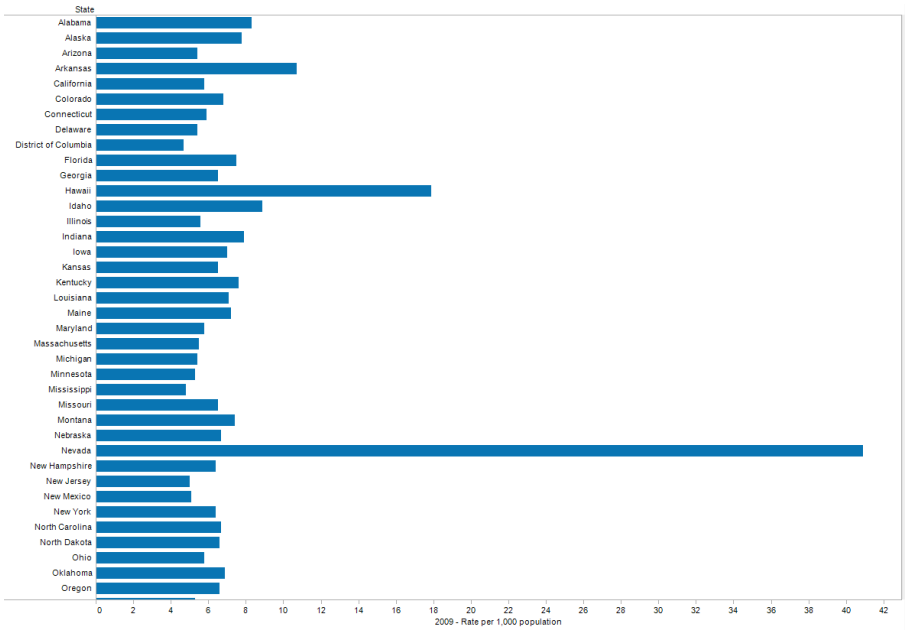
preattentive processing

http://www.csc.ncsu.edu/faculty/healey/PP/index.html

Which state had the highest marriage rate?

Which state had the highest marriage rate?

State	Rate per 1,000 population [2]		
	1990	2000	2009
United States [4]	**9.8**	**8.3**	**6.8**
Alabama	10.6	10.1	8.3
Alaska	10.2	8.9	7.8
Arizona [5]	10.0	7.5	5.4
Arkansas	15.3	15.4	10.7
California	7.9	5.8	5.8
Colorado	9.8	8.3	6.8
Connecticut	7.9	5.7	5.9
Delaware	8.4	6.5	5.4
District of Columbia	8.2	4.9	4.7
Florida	10.9	8.9	7.5
Georgia	10.3	6.8	6.5
Hawaii	16.4	20.6	17.9
Idaho	13.9	10.8	8.9
Illinois	8.8	6.9	5.6
Indiana	9.6	7.9	7.9
Iowa	9.0	6.9	7.0
Kansas	9.2	8.3	6.5
Kentucky	13.5	9.8	7.6
Louisiana	9.6	9.1	7.1
Maine	9.7	8.8	7.2
Maryland	9.7	7.5	5.8
Massachusetts	7.9	5.8	5.5
Michigan	8.2	6.7	5.4
Minnesota	7.7	6.8	5.3
Mississippi	9.4	6.9	4.8
Missouri	9.6	7.8	6.5
Montana	8.6	7.3	7.4
Nebraska	8.0	7.6	6.7
Nevada	99.0	72.2	40.9
New Hampshire	9.5	9.4	6.4
New Jersey	7.6	6.0	5.0
New Mexico [5]	8.8	8.0	5.1
New York [5]	8.6	7.1	6.4
North Carolina	7.8	8.2	6.7
North Dakota	7.5	7.2	6.6
Ohio	9.0	7.8	5.8
Oklahoma	10.6	(NA)	6.9
Oregon	8.9	7.6	6.6
Pennsylvania	7.1	6.0	5.3
Rhode Island	8.1	7.6	5.9
South Carolina	15.9	10.6	7.4
South Dakota	11.1	9.4	7.2
Tennessee	13.9	15.5	8.4
Texas	10.5	9.4	7.1
Utah	11.2	10.8	8.2
Vermont	10.9	10.0	8.7
Virginia	11.4	8.8	7.0
Washington	9.5	6.9	6.0
West Virginia	7.2	8.7	6.9
Wisconsin	7.9	6.7	5.3
Wyoming	10.7	10.0	8.2



U.S. Census Bureau, Statistical Abstract of the United States: 2012
https://www2.census.gov/library/publications/2011/compendia/statab/131ed/tables/vitstat.pdf,
http://www2.census.gov/library/publications/2011/compendia/statab/131ed/tables/12s0133.xls

Why show the data in detail?

- Vis tools can allow people to explore data to find patterns or to determine if a statistical model actually fits the data

- Look out for questionable data
  - "just because it's numbers doesn't mean it's true"
  - is it a typo or something interesting?
    - *"make sure you know which one it is"*

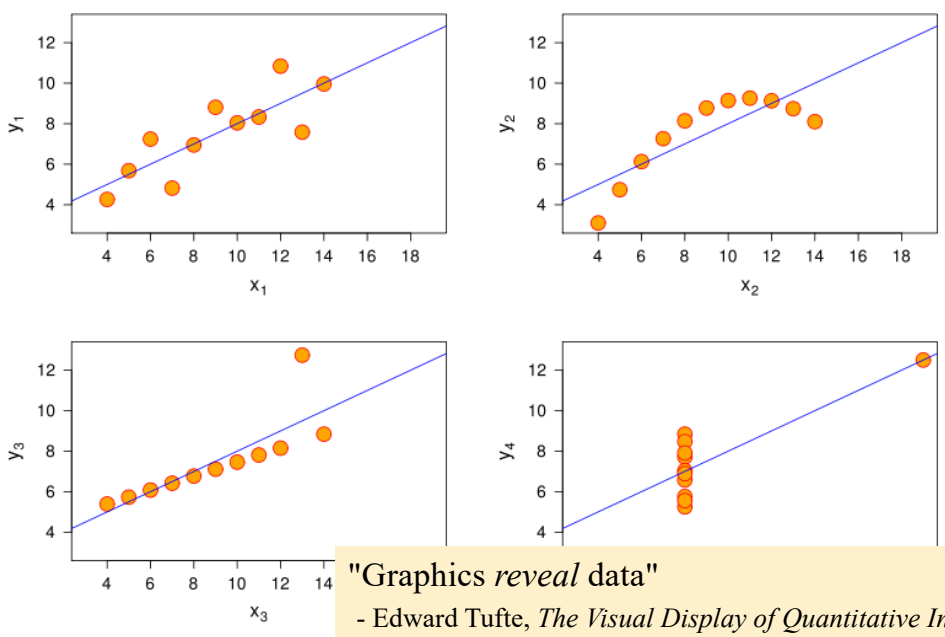
Anscombe's Quartet

Anscombe's Quartet: Raw Data

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	

F.J. Anscombe, "Graphs in Statistical Analysis", *American Statistician*, Feb 1973
Munzner, Figure 1.3

The four data sets are *not* the same

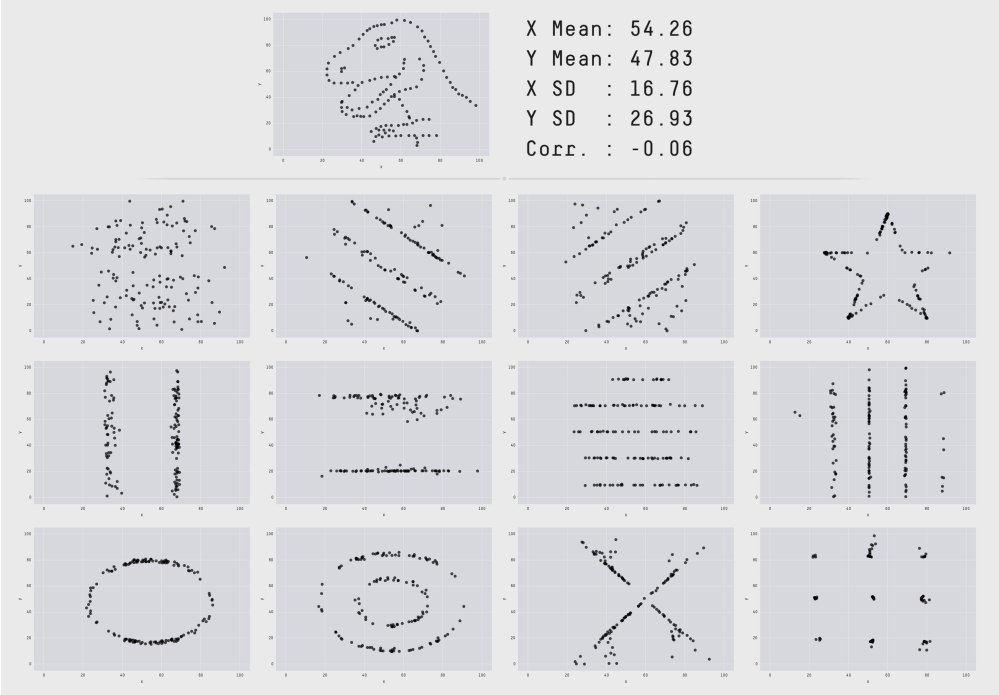


"Graphics *reveal* data"
- Edward Tufte, *The Visual Display of Quantitative Information*

http://en.wikipedia.org/wiki/File:Anscombe.svg

Datasaurus Dozen

https://www.autodeskresearch.com/publications/samestats

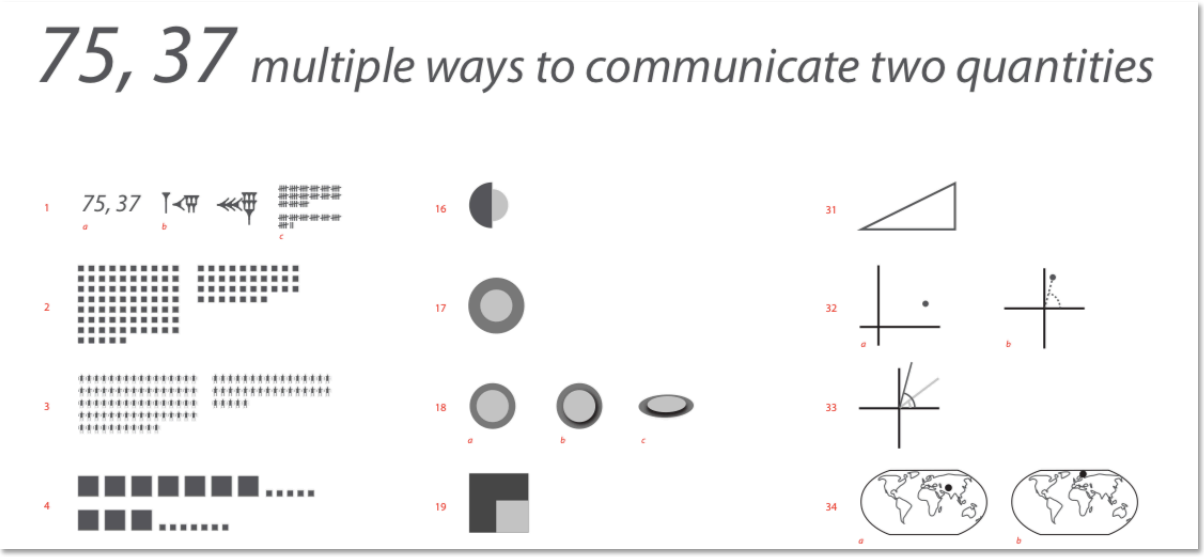


The **design space of possible vis idioms is huge**, and includes the considerations of both **how to create** and **how to interact** with visual representations.

Why is the idiom design space huge?

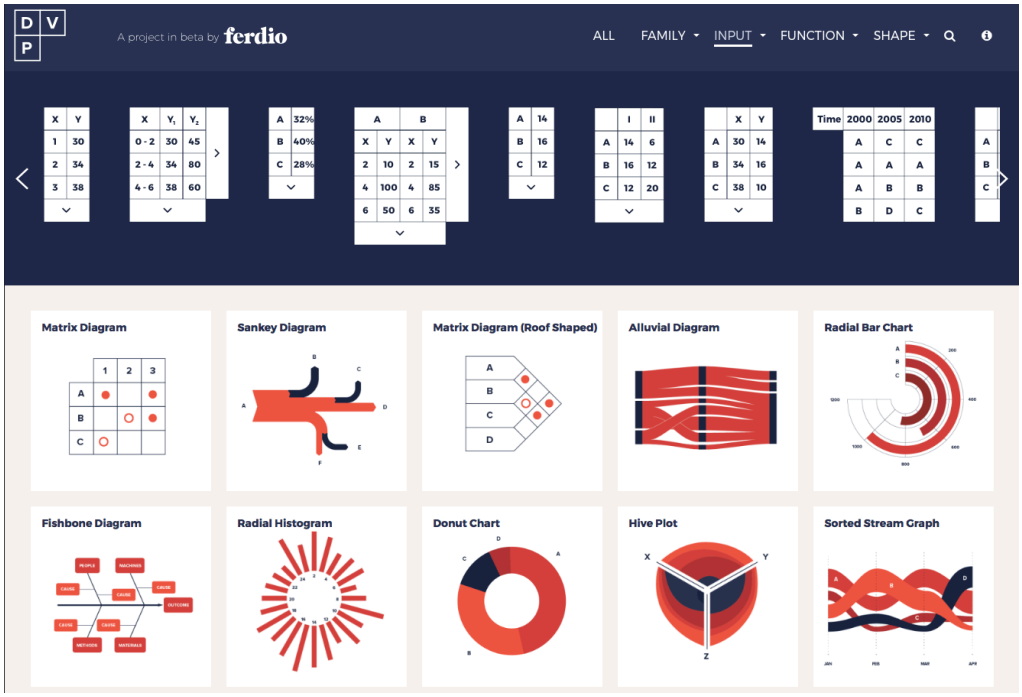
- Vis idioms - approaches to creating and manipulating visual representations
- Simple examples: scatterplots, bar charts, line charts

45 Ways to Communicate Two Quantities

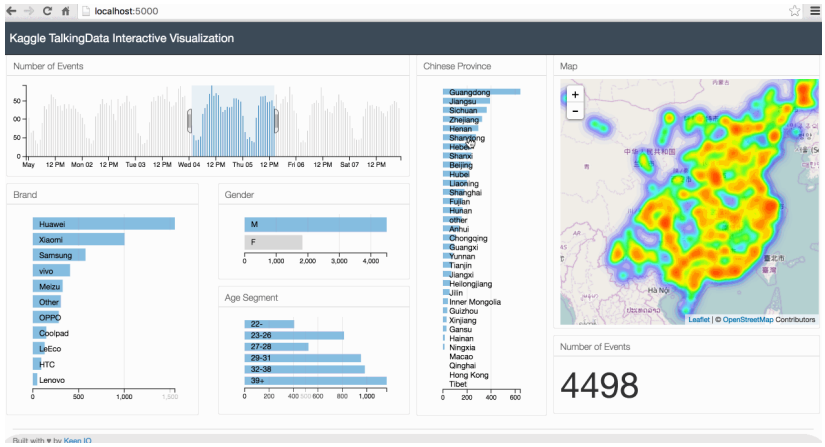


https://visual.ly/blog/45-ways-to-communicate-two-quantities/

datavizproject.com



Why use interactivity?



http://adilmoujahid.com/posts/2016/08/interactive-data-visualization-geospatial-d3-dc-leaflet-python/

- Interaction allows for
  - handling complexity
  - displaying multiple aspects of a dataset

Vis design is full of **tradeoffs**, and **most possibilities in the design space are ineffective** for a **particular task**, so **validating the effectiveness** of a design is both necessary and difficult.

…ineffective for a particular task

Why focus on tasks?

- The intended task is ***just as important*** as the data to be visualized.

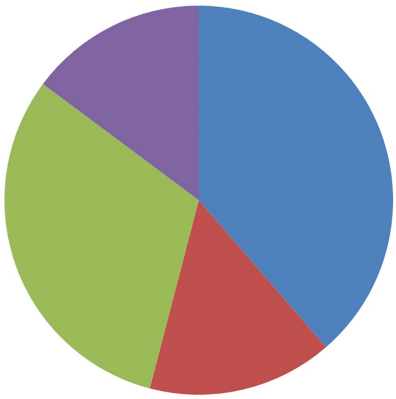
- Four categories of tasks
  - presentation
  - discovery
  - enjoyment of information
  - producing more information for later use

Why focus on effectiveness?

- Effectiveness is an important measure for understanding if the user task was supported.
  - "The purpose of visualization is insight, not pictures." -Ben Shneiderman
- But, no picture can tell the truth, the whole truth, and nothing but the truth.

Why are most designs ineffective?

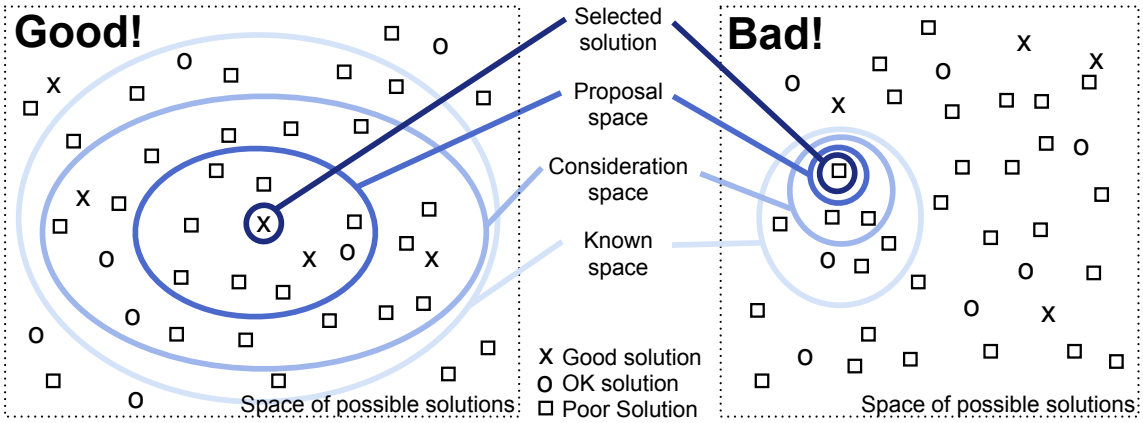
- Design may not match with human perception
- Design may not match with intended task



Which color comprises the greatest portion?

What is the percentage of the green region?

Search space metaphor for vis design



Munzner, Fig 1.5

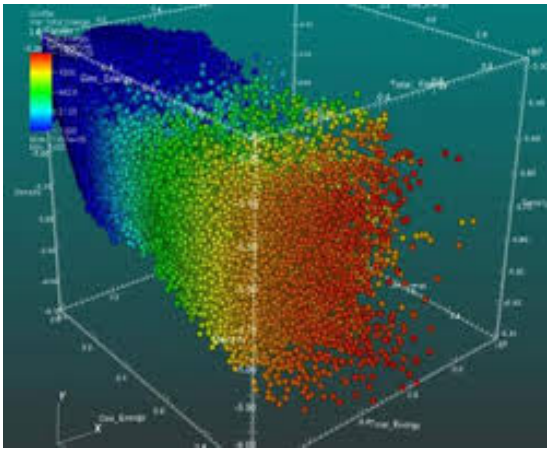
Why is validation difficult?

- How do you know if your visualization "works"?
  - How do you measure insight?
- How do you argue that one design is better than another?
  - What does "better" mean? faster? more fun? more effective?
  - What does "effectively" mean?

Vis designers must take into account three very different kinds of **resource limitations**: those of computers, of humans, and of displays.

Why are there resource limitations?

- computational capacity
- human perceptual and cognitive capacity
- display capacity



http://bigdata-madesimple.com/30-simple-data-visualization-tools/

Vis usage can be **analyzed in terms of why** the user needs it, **what** data is shown**, and how** the idiom is designed.

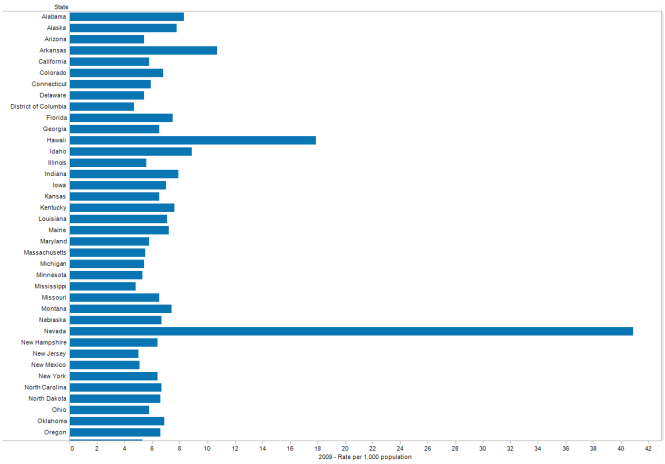
Why analyze vis?

- Analyzing existing systems is a good stepping stone to designing new ones.
- High-level framework for analyzing vis use
  - *what* data the user sees
  - *why* the user intends to use a vis tool
  - *how* the visual encoding and interaction idioms are constructed in terms of design choices

Tools

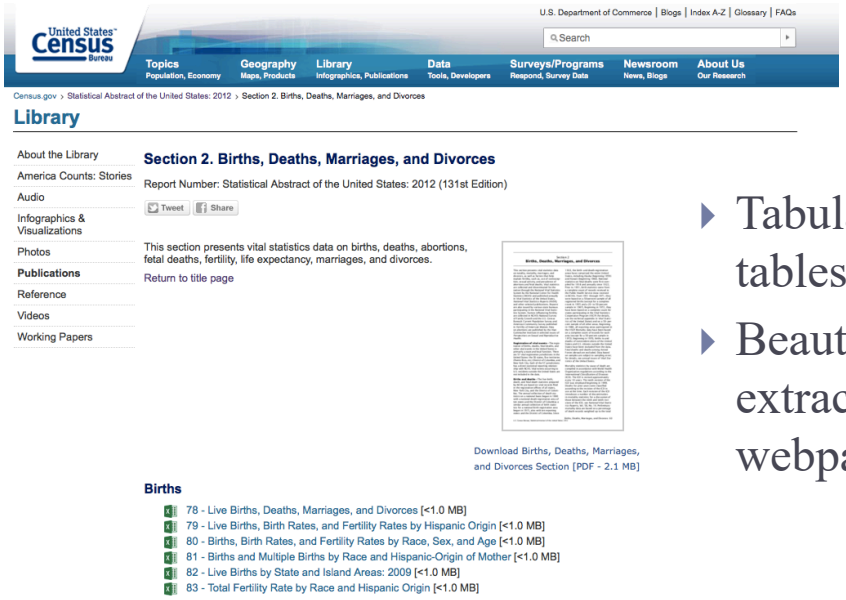
Workflow

- What
  - data gathering
  - data wrangling
- Why
  - developing questions
  - initial analysis
- How
  - charts for analysis
  - charts for presentation



What: Data Gathering

https://www.census.gov/library/publications/2011/compendia/statab/131ed/births-deaths-marriages-divorces.html



- Tabula – extract tables from PDFs
- Beautiful Soup – extract data from webpages

What: Data Wrangling

- Data is often messy
- Tools
  - Excel
  - OpenRefine – filter and clean data files
- *Much more on this next week*

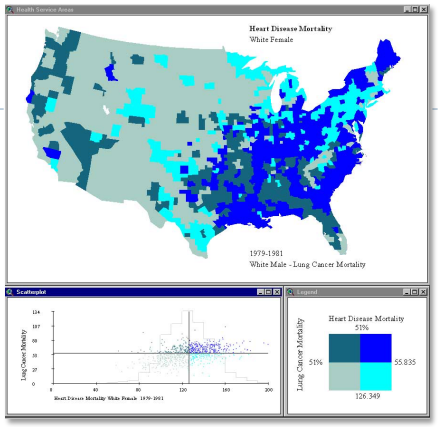
Why

- Developing questions
  - which states have the highest marriage rates?
  - which states have the highest divorce rates? is that correlated to marriage rate?
  - which states have the highest birth rates? is that correlated to marriage rate?

- Initial analysis
  - Excel
  - Google Sheets, Google Charts
  - Tableau
  - R

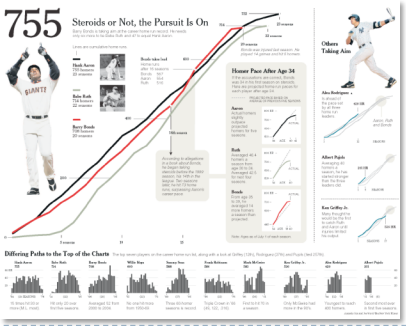
How

- Charts for analysis



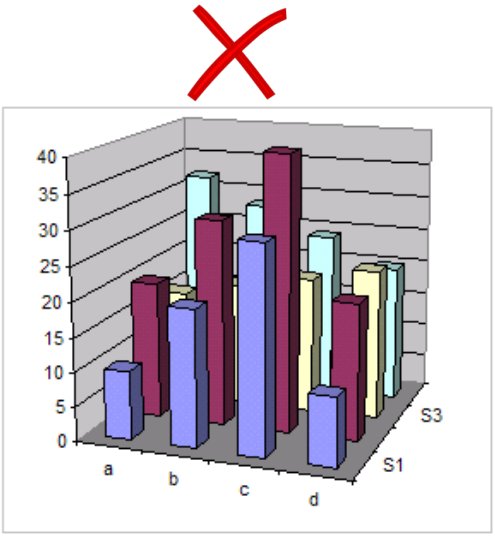
https://www.geovista.psu.edu/research/healthvisualization/

- Charts for presentation

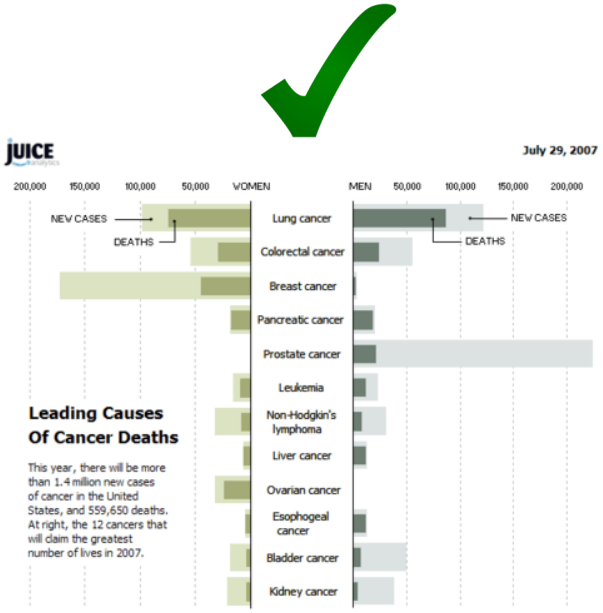


New York Times, 2006

Excel

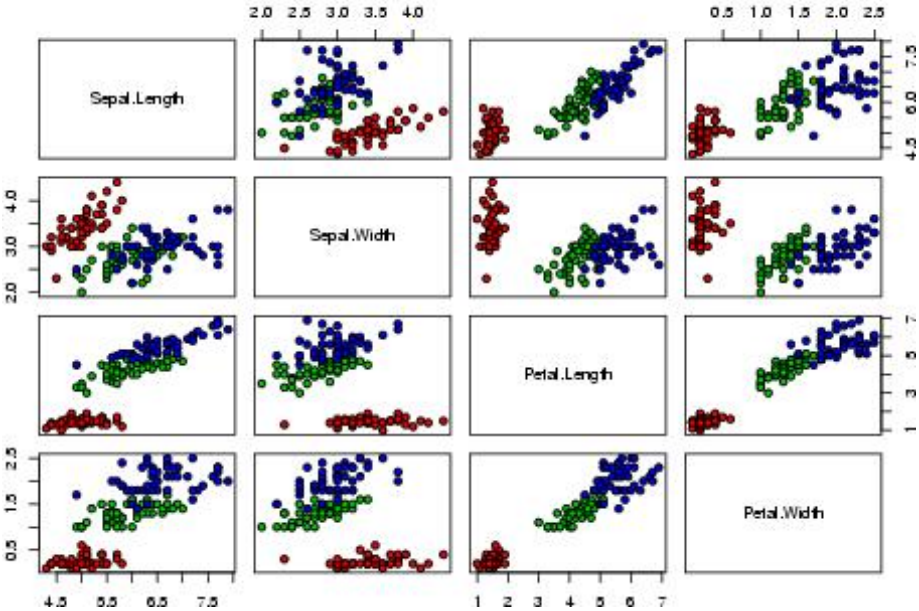


http://chandoo.org/wp/2008/09/03/6-charts-to-never-use/



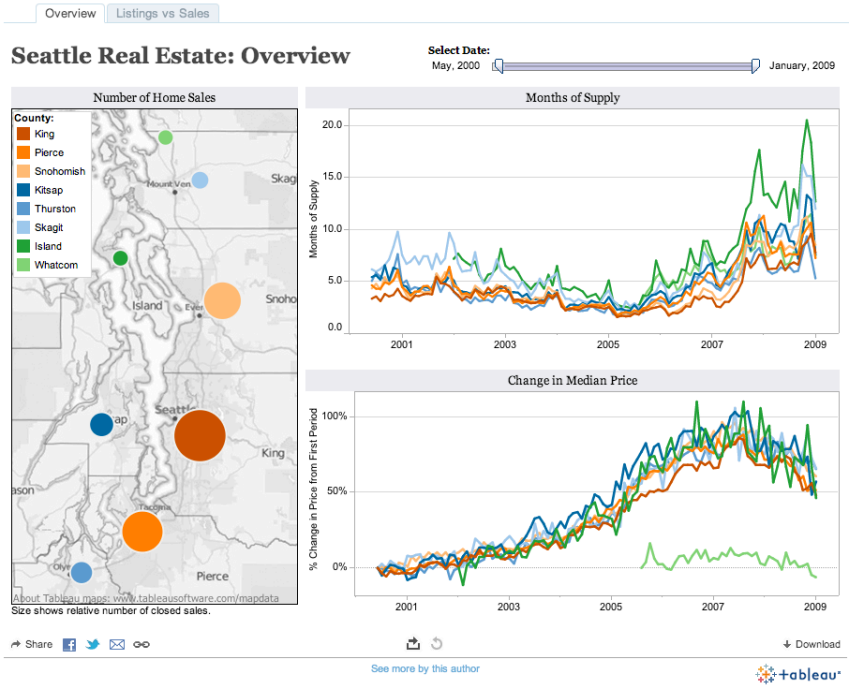
http://www.juiceanalytics.com/writing/recreating-ny-times-cancer-graph/

http://www.r-project.org



Tableau

https://www.tableau.com/academic/students



D3

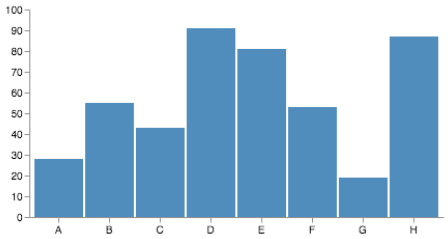
http://d3js.org

Vega and Vega-Lite

https://vega.github.io/vega/
https://vega.github.io/vega-lite/

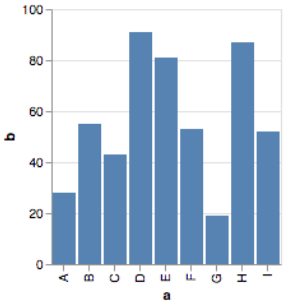
- Vega is a visualization grammar, built on top of D3

https://vega.github.io/vega/examples/bar-chart/



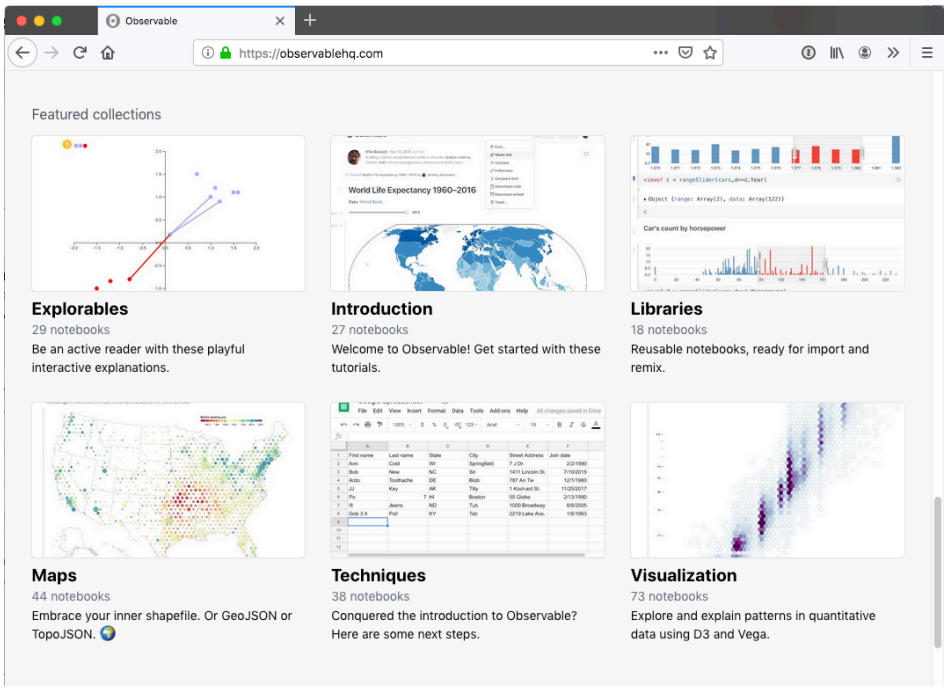
- Vega-Lite is a higher-level language built on top of Vega

https://vega.github.io/vega-lite/examples/bar.html



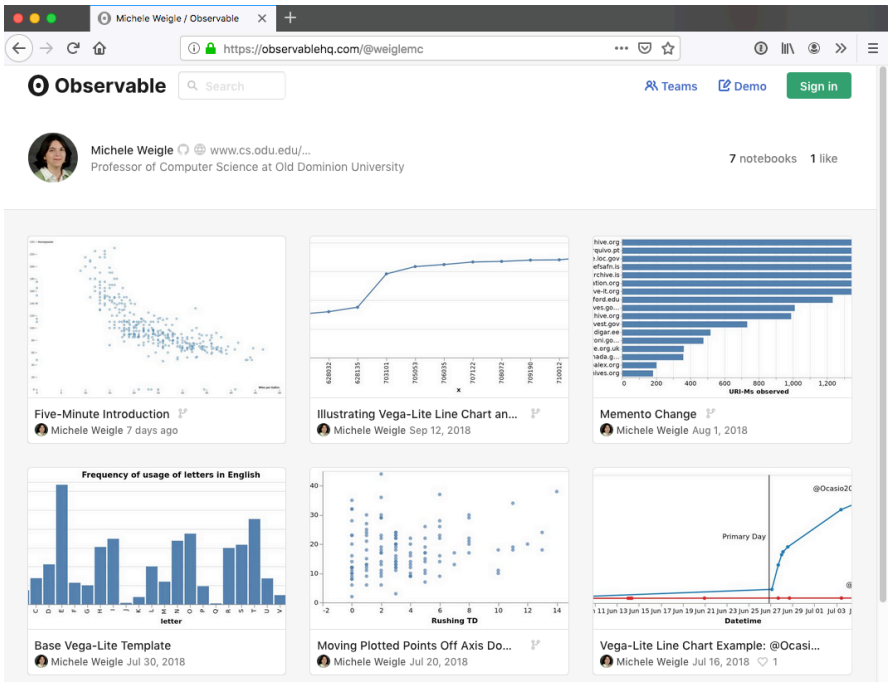
Observable

https://observablehq.com

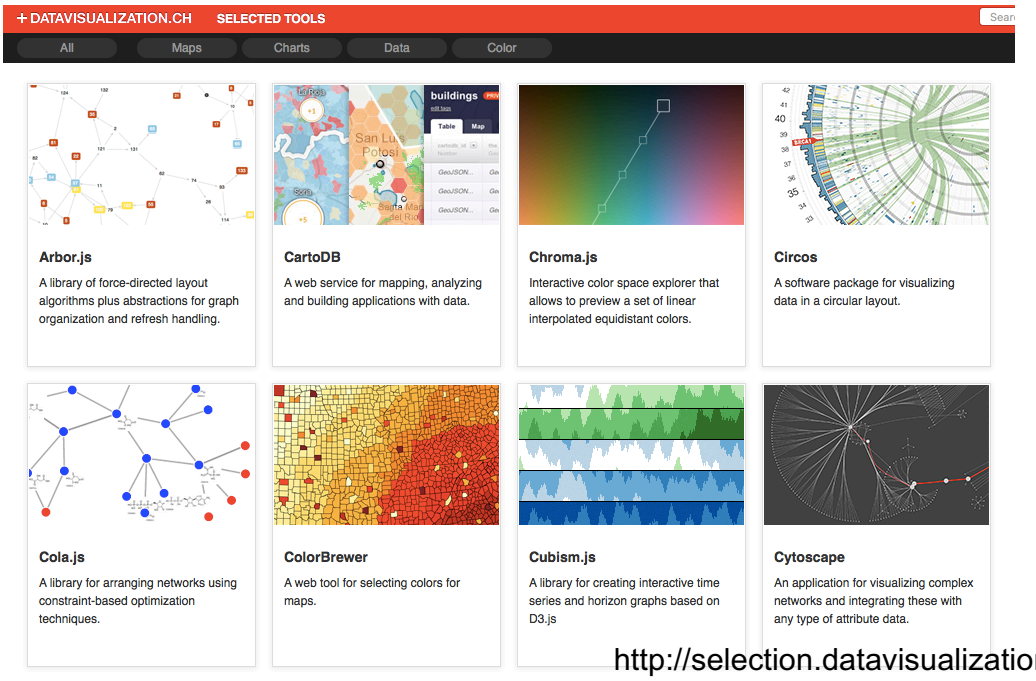


https://observablehq.com/@weiglemc

My Observable Notebooks

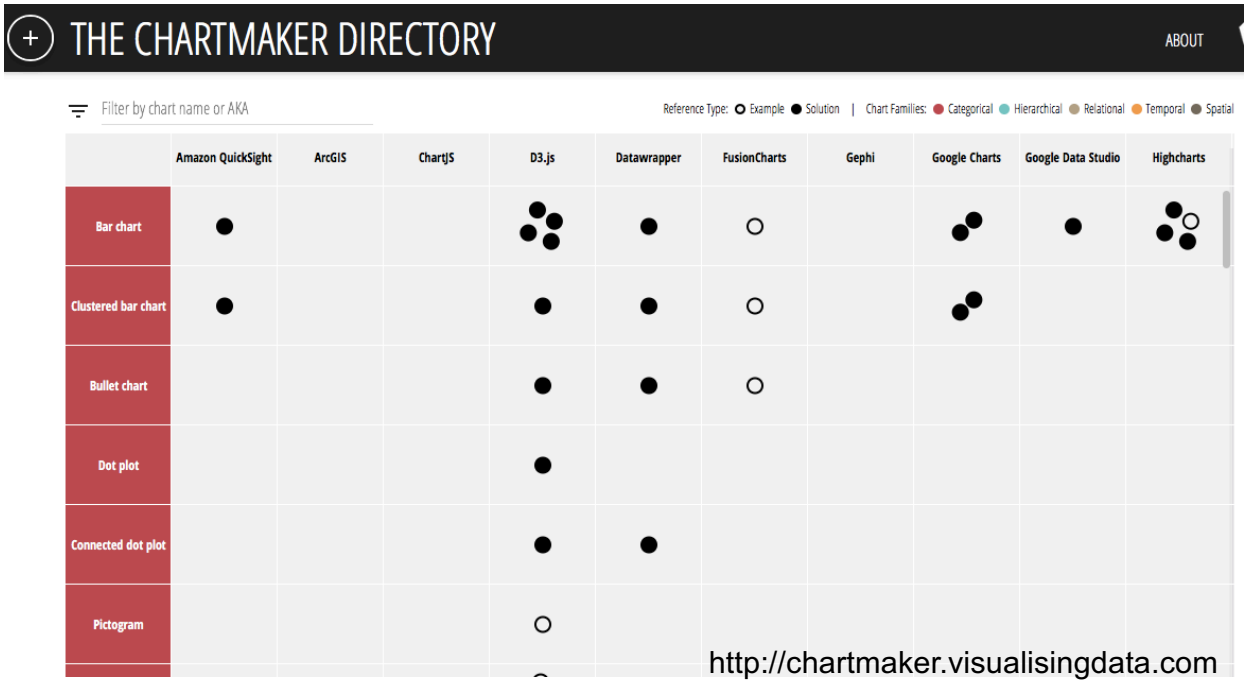


Choosing Tools – datavisualization.ch



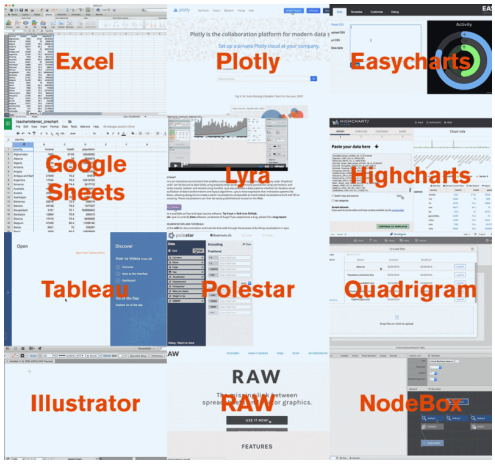
http://selection.datavisualization.ch

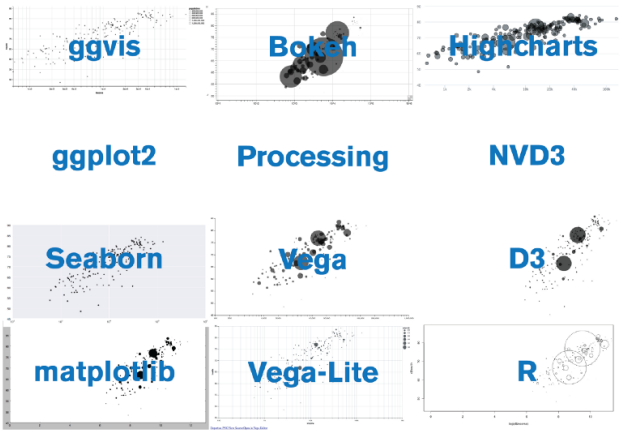
Choosing Tools – chartmaker directory



http://chartmaker.visualisingdata.com

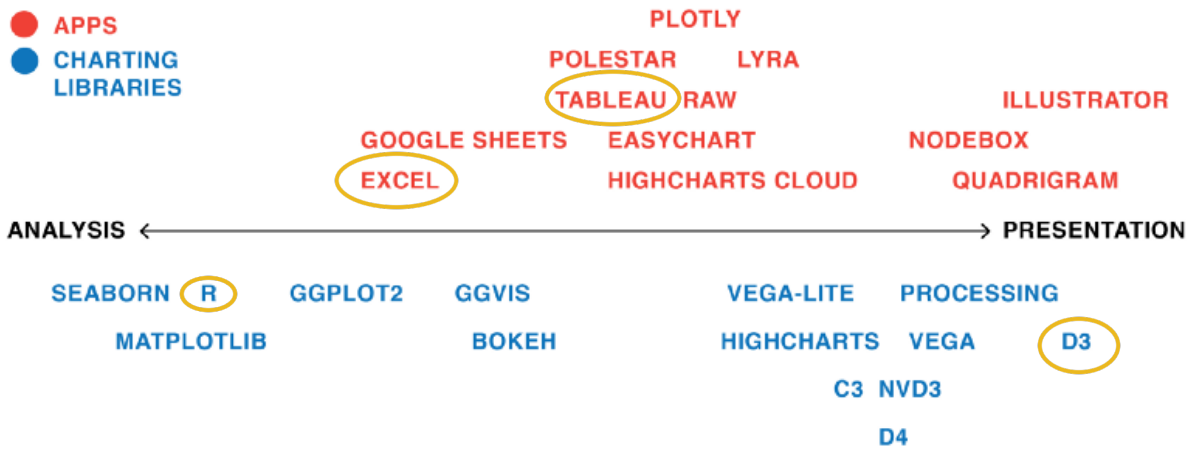
What I Learned Recreating One Chart Using 24 Tools





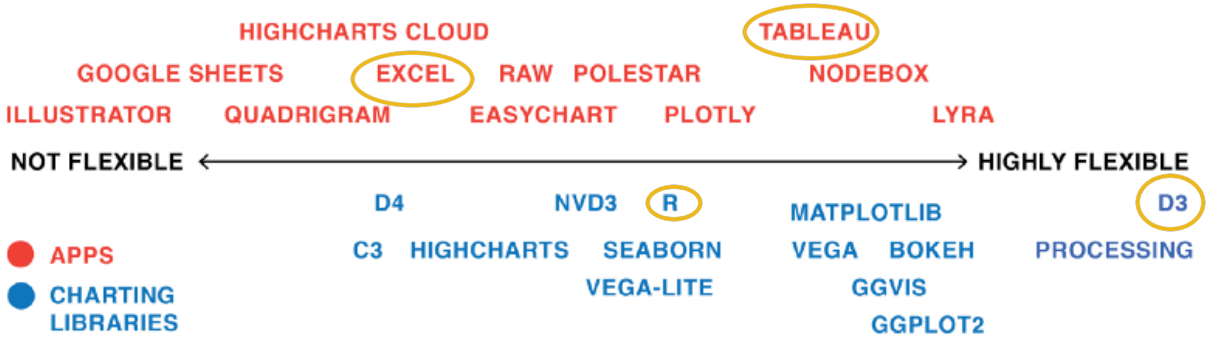
https://source.opennews.org/en-US/articles/what-i-learned-recreating-one-chart-using-24-tools/

Tools for Analysis vs. Presentation



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

Flexibility of Tools



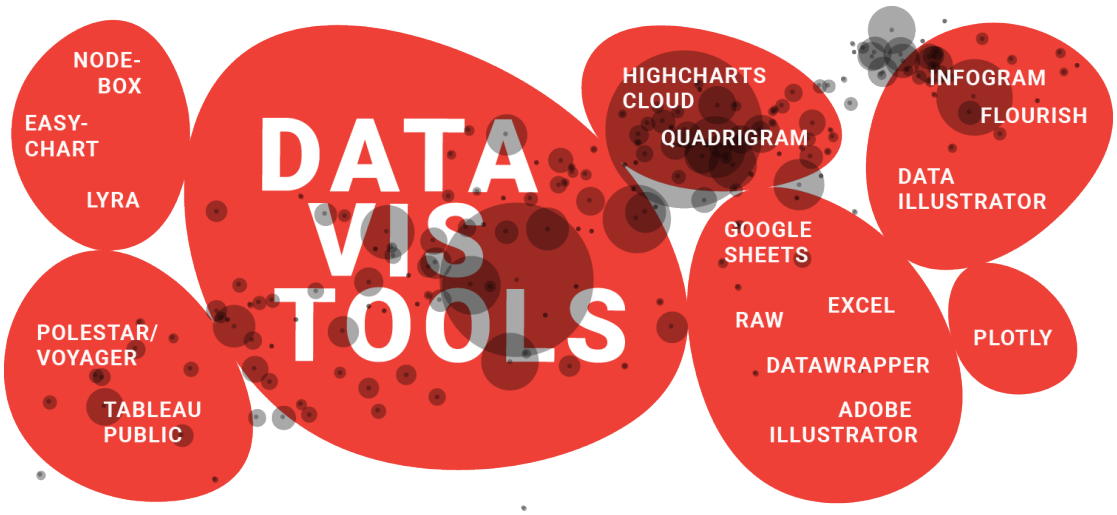
https://source.opennews.org/en-US/articles/what-i-learned-recreating-one-chart-using-24-tools/

Apps vs. Libraries and Static vs. Interactive

	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

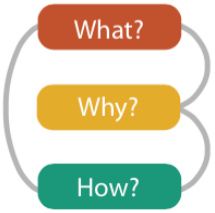
https://source.opennews.org/en-US/articles/what-i-learned-recreating-one-chart-using-24-tools/

One Chart, Nine Tools – Revisited



https://lisacharlotterost.github.io/datavistools-revisited

Topic Objectives



- Define visualization.
- Explain the importance of humans in the visualization process.
- Explain why human vision is particularly well-suited for information transfer.
- Give an example of a visualization idiom.
- Explain why it is best to consider multiple alternatives for vis before selecting a solution.
- Explain at a high-level the "what-why-how" framework for analyzing visualization use.
- Differentiate between R, D3, and Tableau and describe the type of tasks for which each tool might be most appropriate.