CS 725/825
Information Visualization
Fall 2017

Intro to Info Vis

Dr. Michele C. Weigle

http://www.cs.odu.edu/~mweigle/CS725-F17/

Today

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

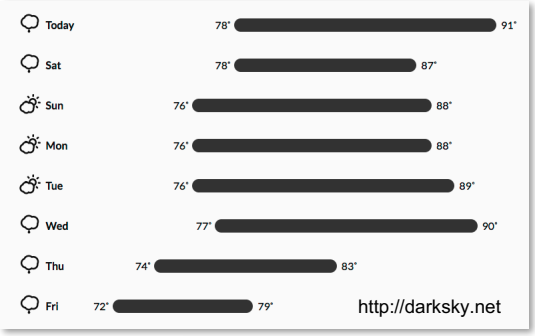
Topic Objectives

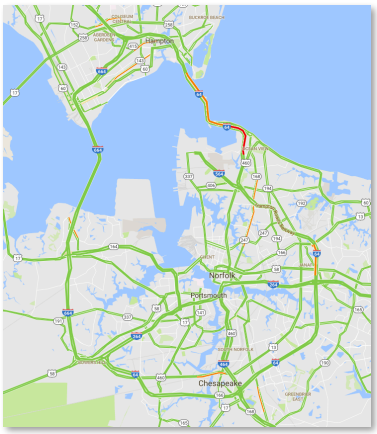
- Define visualization.
- Explain the importance of humans in the visualization process.
- Explain why human vision is particularly well-suited for information transfer.
- Give an example of a visualization idiom.
- Explain why it is best to consider multiple alternatives for vis before selecting a solution.
- 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.

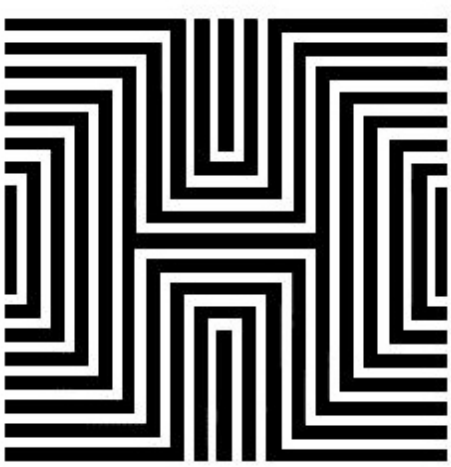
Munzner, pg. 1

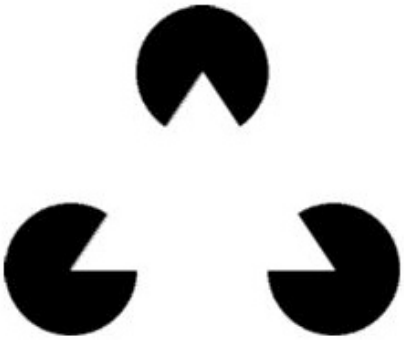
Why have a human in the loop?

- 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

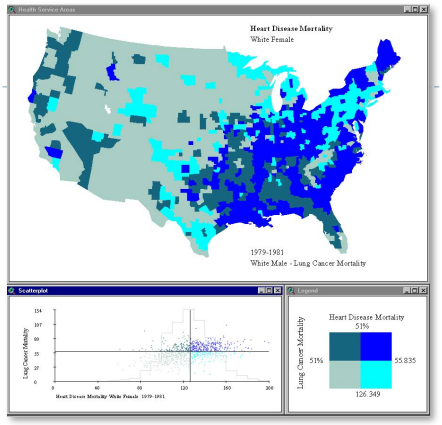




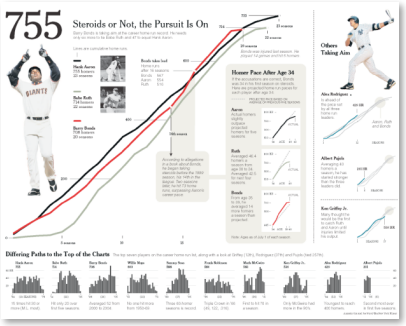
Create visualizations that lets computers do what computers do well and lets humans do what humans do well.

Uses of vis tools

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

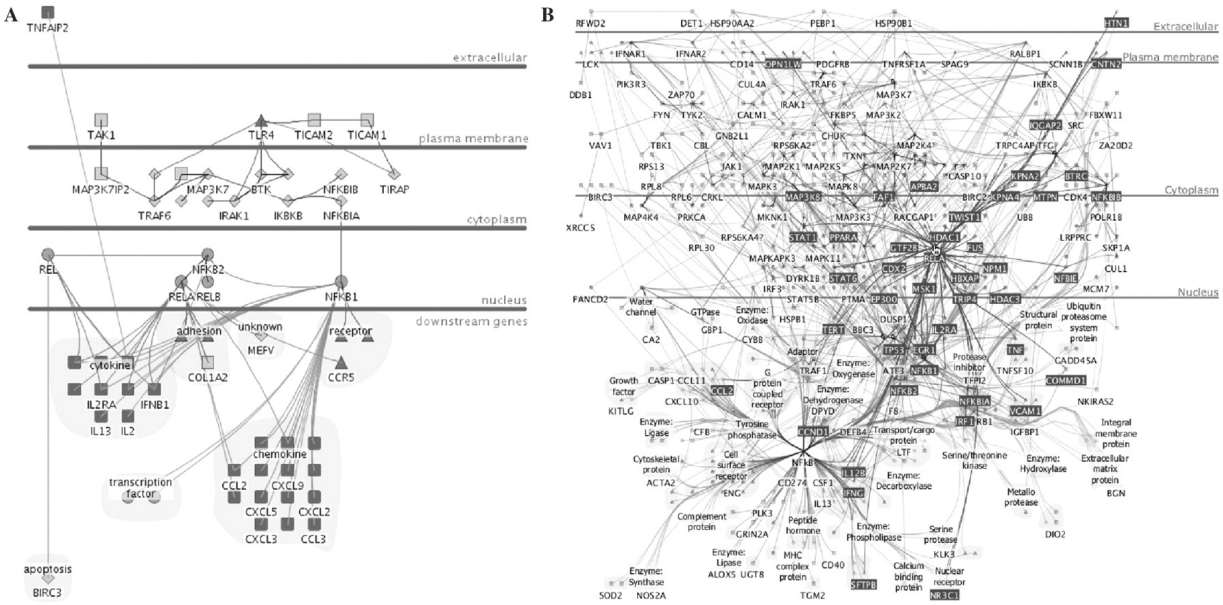


https://www.geovista.psu.edu/research/healthvisualization/



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

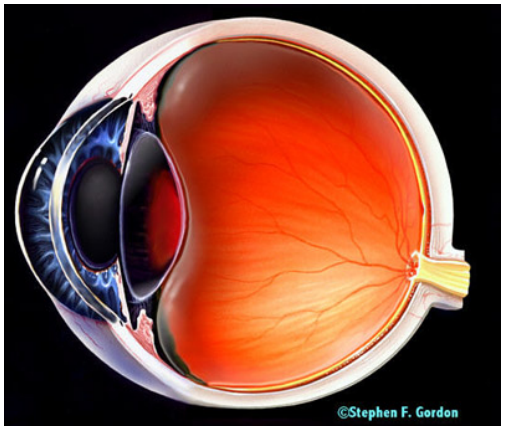
Visualization can extend your memory

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]

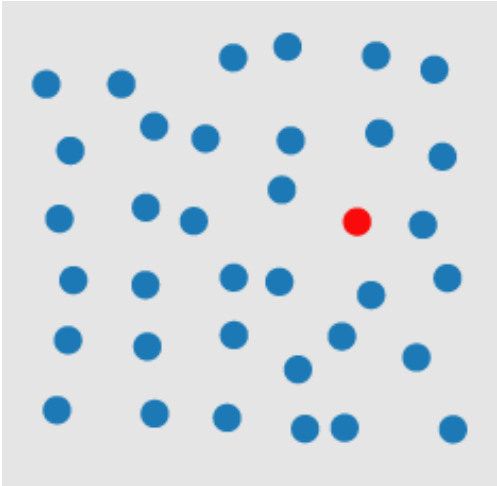
Example courtesy Tamara Munzner, Univ. of British Columbia

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

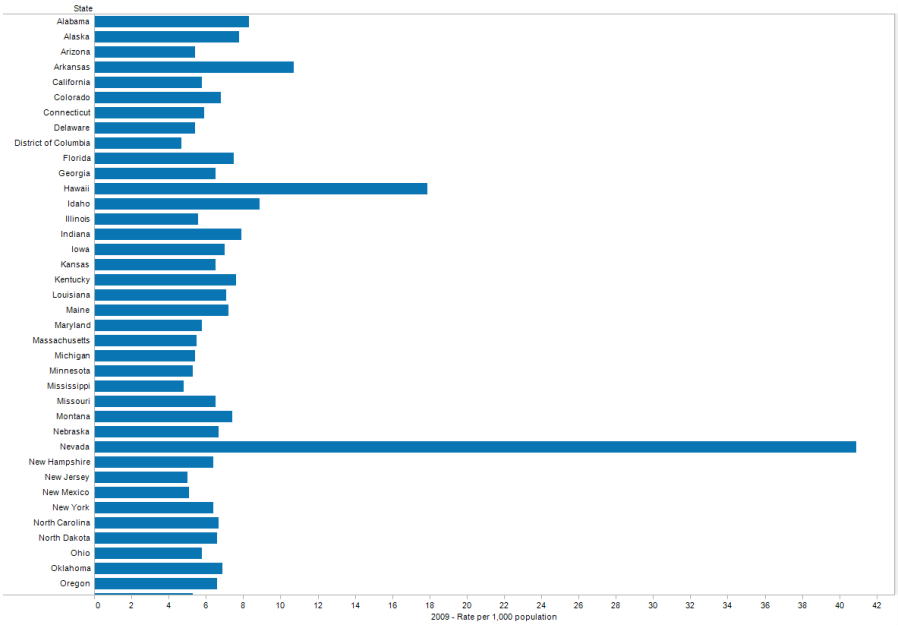
http://www.csc.ncsu.edu/faculty/healey/PP/index.html

Which state had the highest marriage rate?

augment human capabilities

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	8.9	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)	8.9
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
http://www.census.gov/compendia/statab/2012/tables/12s0133.pdf,
http://www.census.gov/compendia/statab/2012/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"*

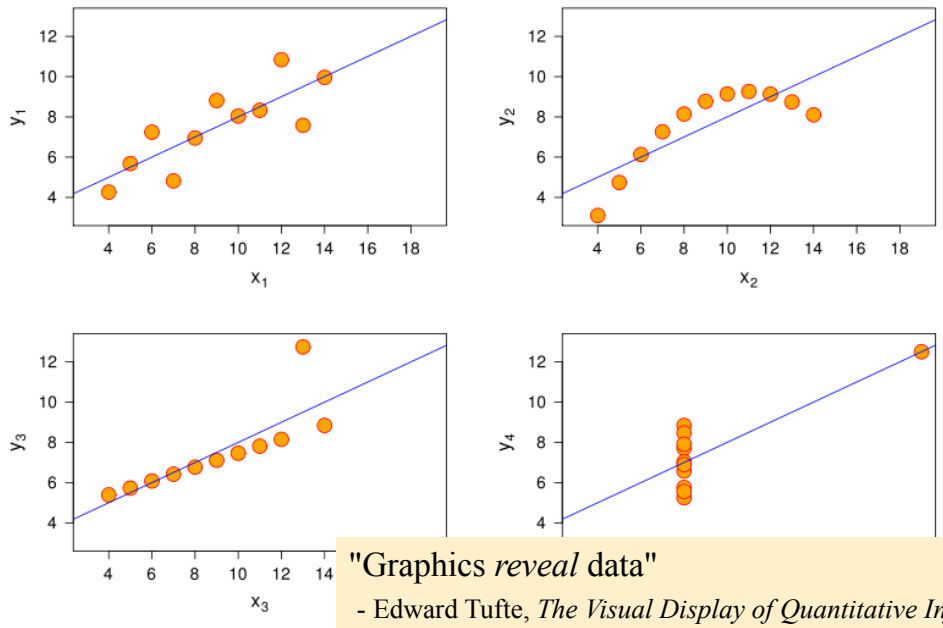
Anscombe's Quartet

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

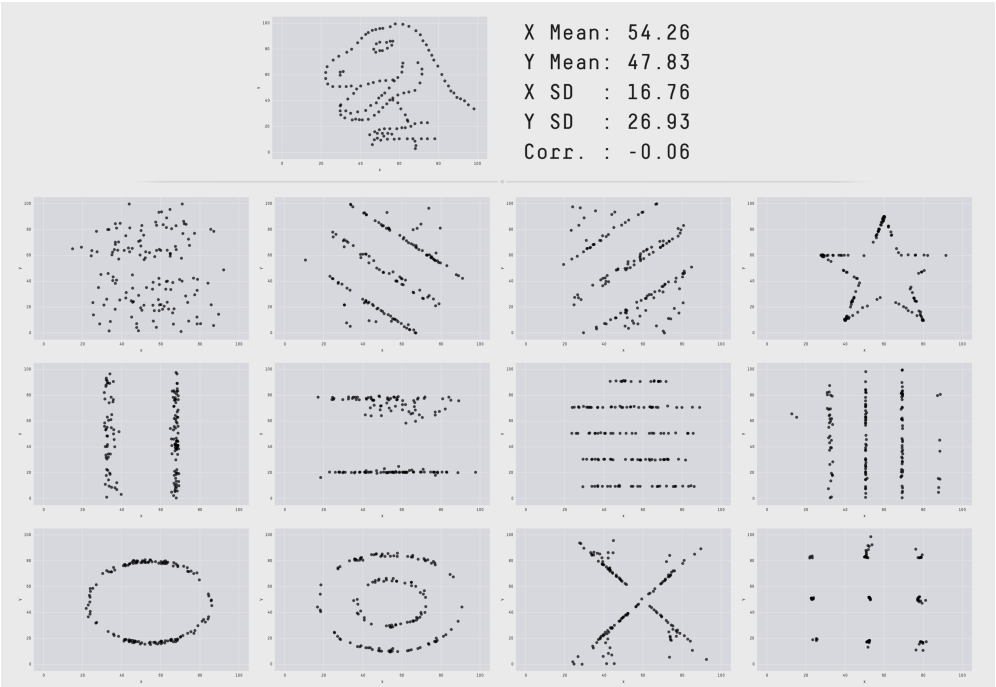


"Graphics *reveal* data"
- Edward Tufte, *The Visual Display of Quantitative Information*

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

Datasaurus Dozen

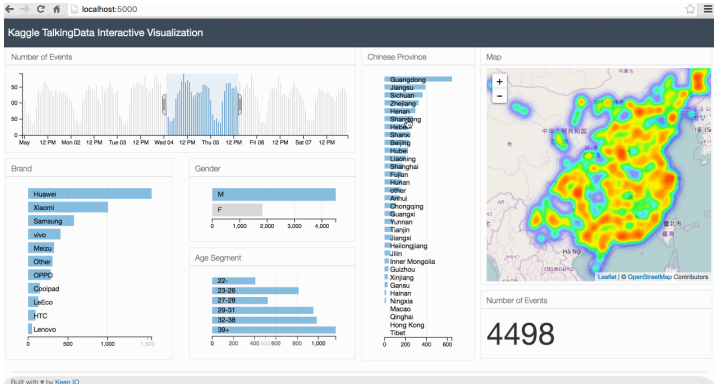
https://www.autodeskresearch.com/publications/samestats



Why use interactivity?

- 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.

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



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

Why is the idiom design space huge?

- 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 idioms - approaches to creating and manipulating visual representations

- Simple examples: scatterplots, bar charts, line charts

Visual Vocabulary

Deviation

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine chart

Splits a single value into 2 contrasting components (eg Male/Female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Correlation

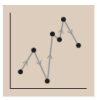
Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Line + Column

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot



Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts.

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Ranking

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop chart

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Barcode plot

Like dot strip plots, good for displaying all the data in a table,they work best when highlighting individual values.

Cumulative curve

A good way of showing how unequal a distribution is: y axis is always cumulative frequency, x axis is always a measure.

Change Over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader.

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time - but usually best with only one series of data at a time.

Line + column

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Stock price

Usually focused on day-to-day activity, these charts show opening/closing and hi/low points of each day.

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Fan chart (projections)

Use to show the uncertainty in future projections - usually this grows the further forward to projection.

Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline

Great when date and duration are key elements of the story in the data.

Circle timeline

Good for showing discrete values of varying size across multiple categories (eg earthquakes by contintent).

Seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data.

Part-to-Whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Stacked column

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Proportional stacked bar

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg. total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Sunburst

Another way of visualisaing hierarchical part-to-whole relationships. Use sparingly (if at all) for obvious reasons.

Arc

A hemicycle, often used for visualising political results in parliaments.

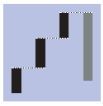
Gridplot

Good for showing % information, they work best when used on whole numbers and work well in multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall



Can be useful for showing part-to-whole relationships where some of the components are negative.

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Proportional stacked bar

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop chart

Lollipop charts draw more attention to the data value than standard bar/column – does not HAVE to start at zero (but preferable).

Radar chart

A space-efficient way of showing value pf multiple variables– but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrngement of the variables is important. Usually benefits from highlighting values.

Flow

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

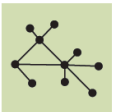
Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network



Used for showing the strength and inter-connectdness of relationships of varying types.

Spatial

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Used only when precise locations or geographical patterns in data are more important to the reader than anything else.

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitde)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Why focus on tasks?

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

- 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.

- 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.

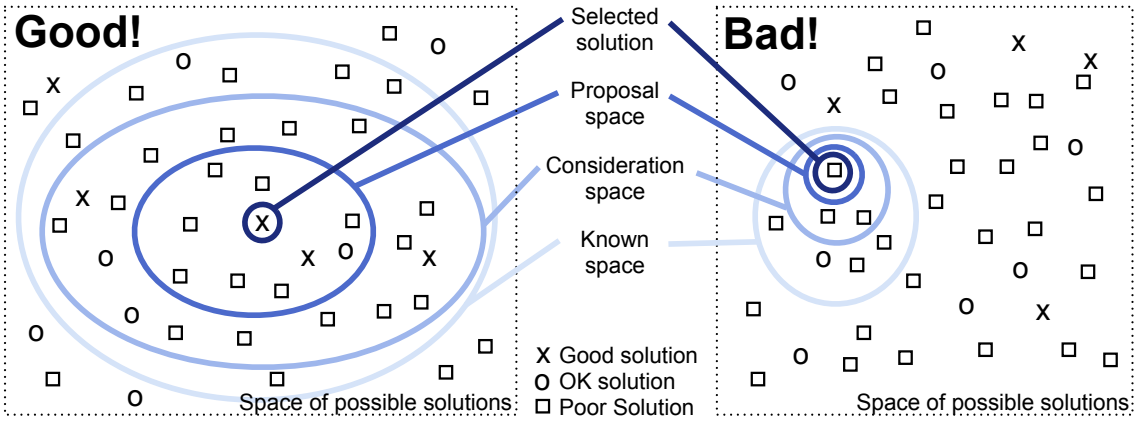
Why are most designs ineffective?

- 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.

- Design may not match with human perception

- Design may not match with intended task

Search space metaphor for vis design



Munzner, Fig 1.5

Why is validation difficult?

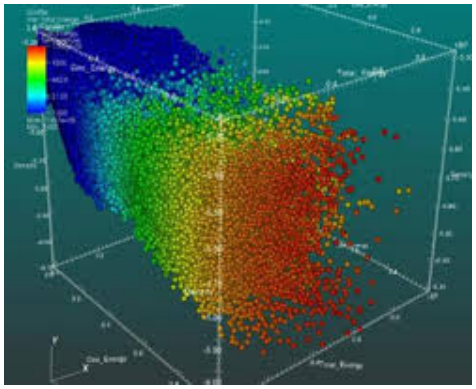
- Vis design is full of **tradeoffs**, and **most possibilities in the design space are ineffective** for a particular task, so **validating the effectiveness** of a design is both necessary and difficult.

- 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?

Why are there resource limitations?

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

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



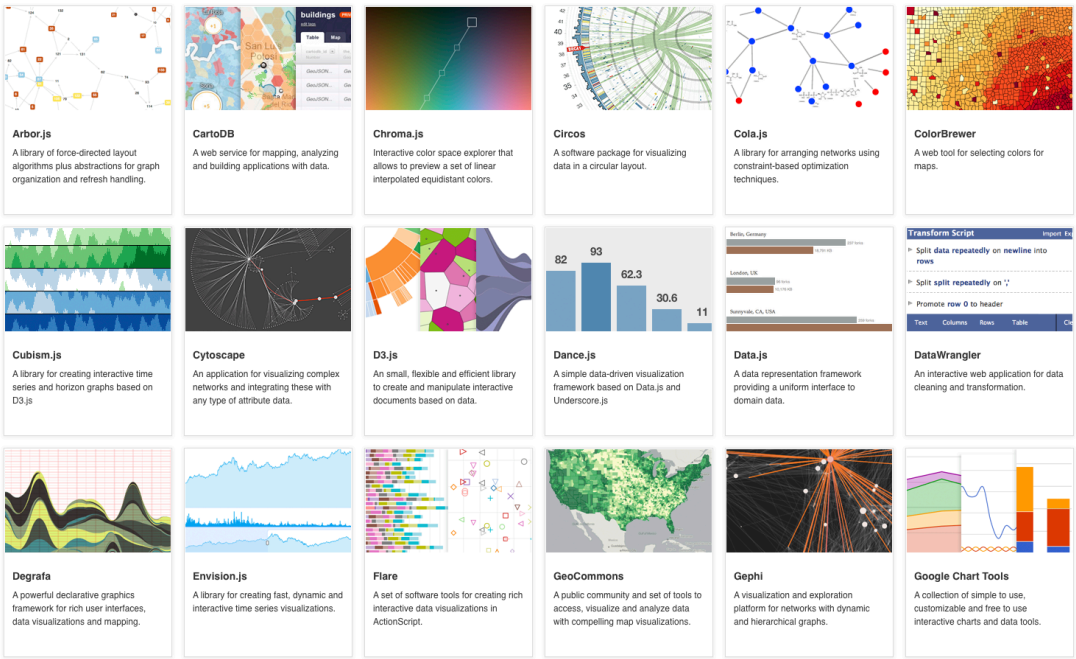
http://bigdata-madesimple.com/30-simple-data-visualization-tools/

Why analyze vis?

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

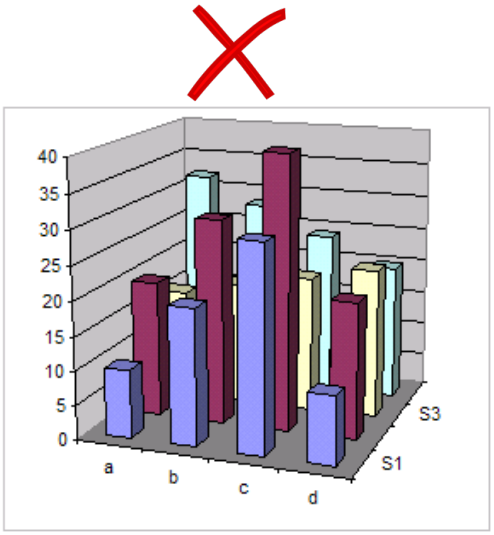
Tools

Tools of the Trade

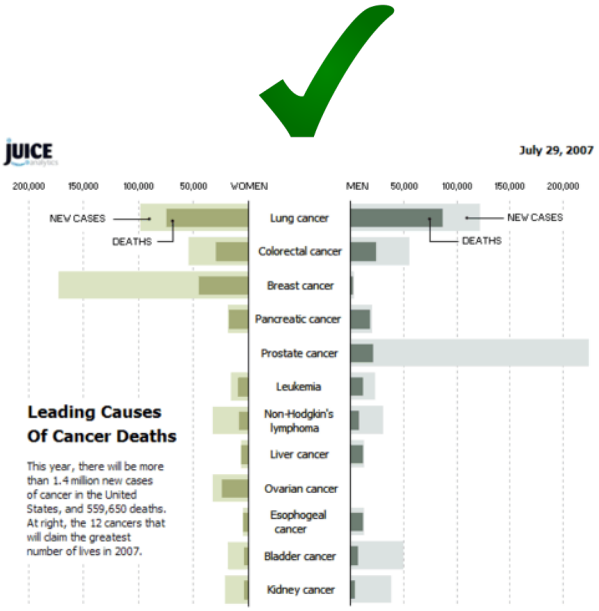


http://selection.datavisualization.ch/

Excel

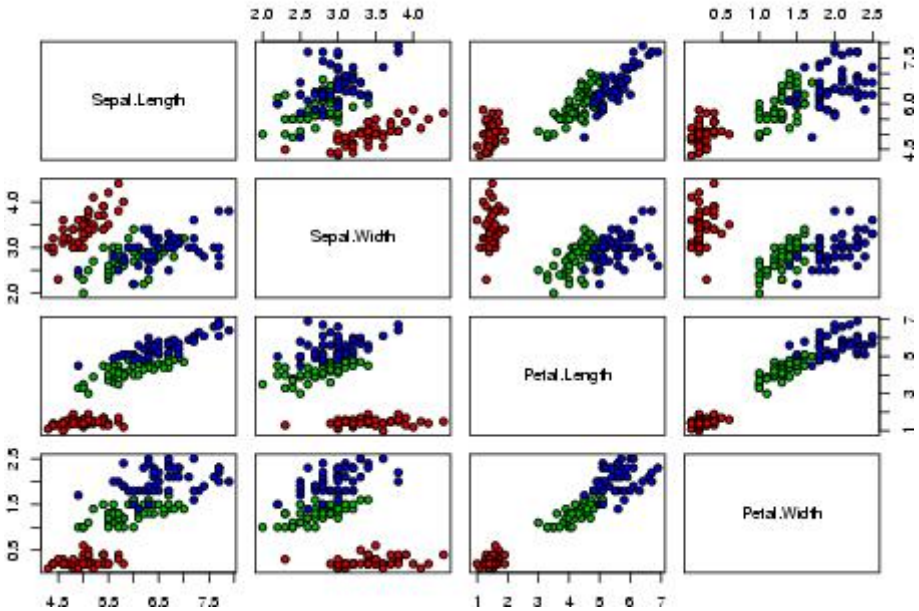


http://chandoo.org/wp/2008/09/03/6-charts-to-never-use/



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

http://www.r-project.org

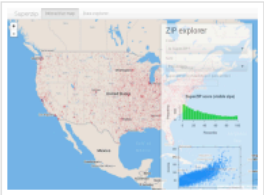


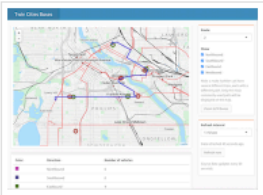
Shiny

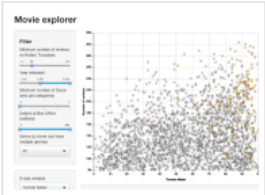
https://shiny.rstudio.com

Interactive visualizations

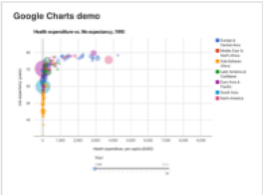
Shiny is designed for fully interactive visualization, using JavaScript libraries like d3, Leaflet, and Google Charts.


SuperZip example


Bus dashboard



Movie explorer

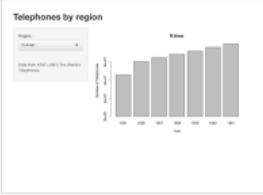


Google Charts

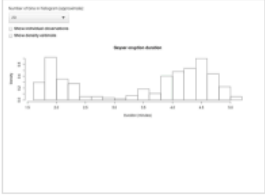
Start simple

If you're new to Shiny, these simple but complete applications are designed for you to study.

Kmeans example



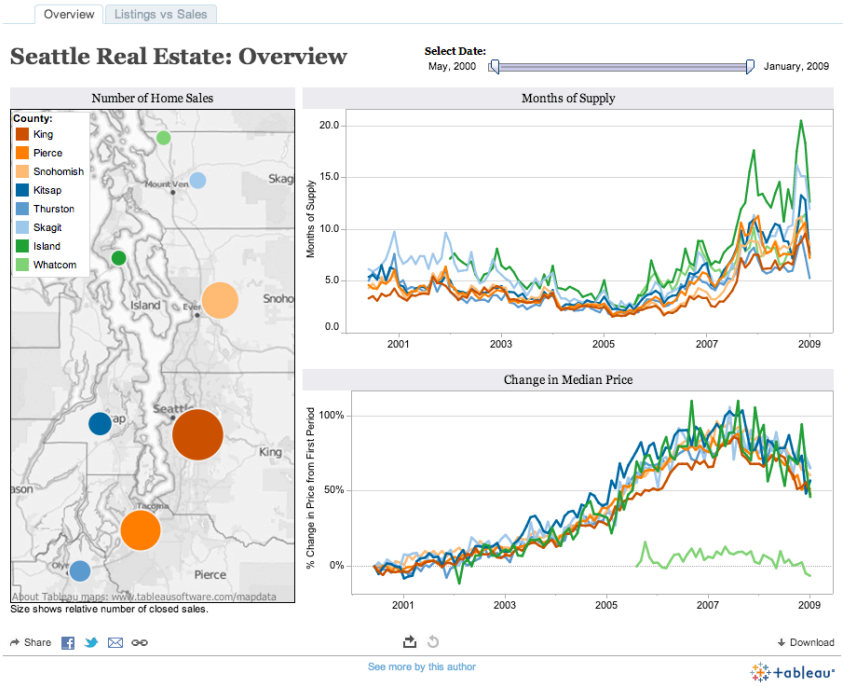
Telephones by region


Faithful

Word cloud

Tableau

http://www.tableausoftware.com

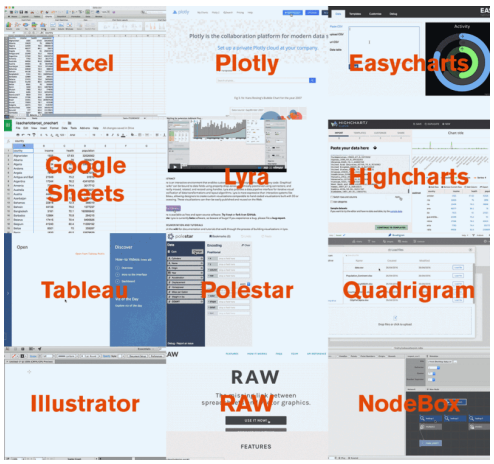


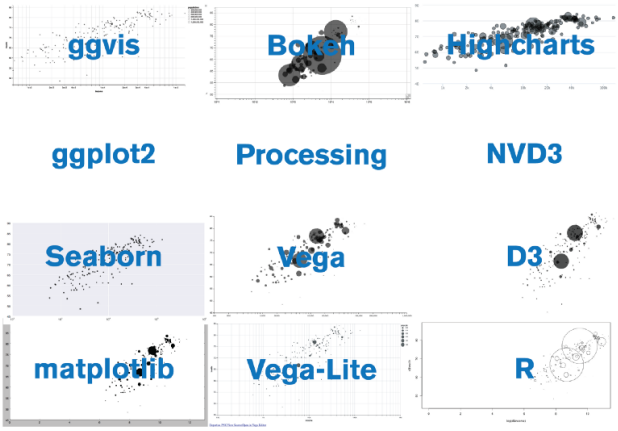
D3

http://d3js.org

Many, many others

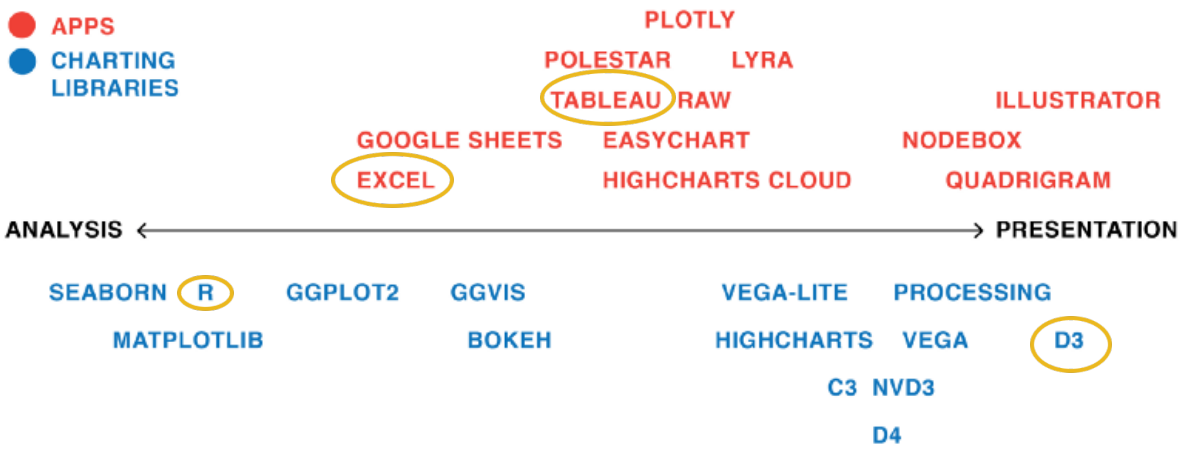
- "What I Learned Recreating One Chart Using 24 Tools" - *assigned reading*





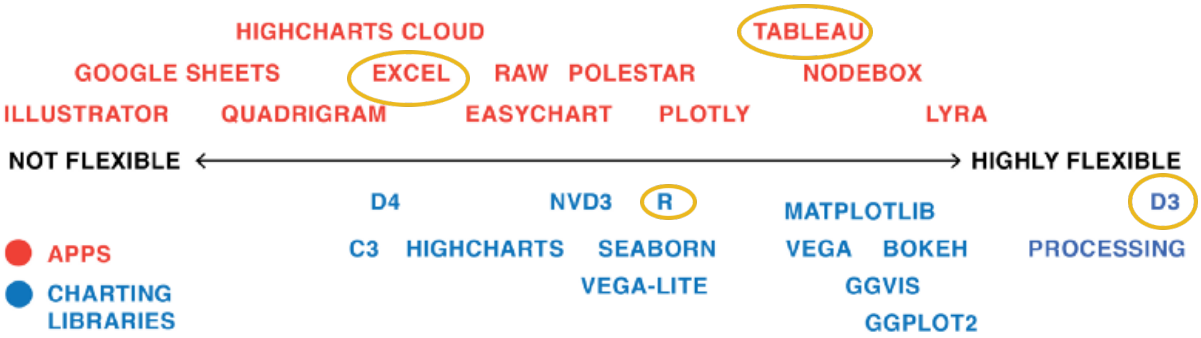
https://source.opennews.org/en-US/articles/what-i-learned-recreating-one-chart-using-24-tools/

Tools for Analysis vs. Presentation



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

Flexibility of Tools



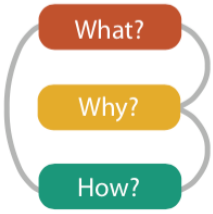
https://source.opennews.org/en-US/articles/what-i-learned-recreating-one-chart-using-24-tools/

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/

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.