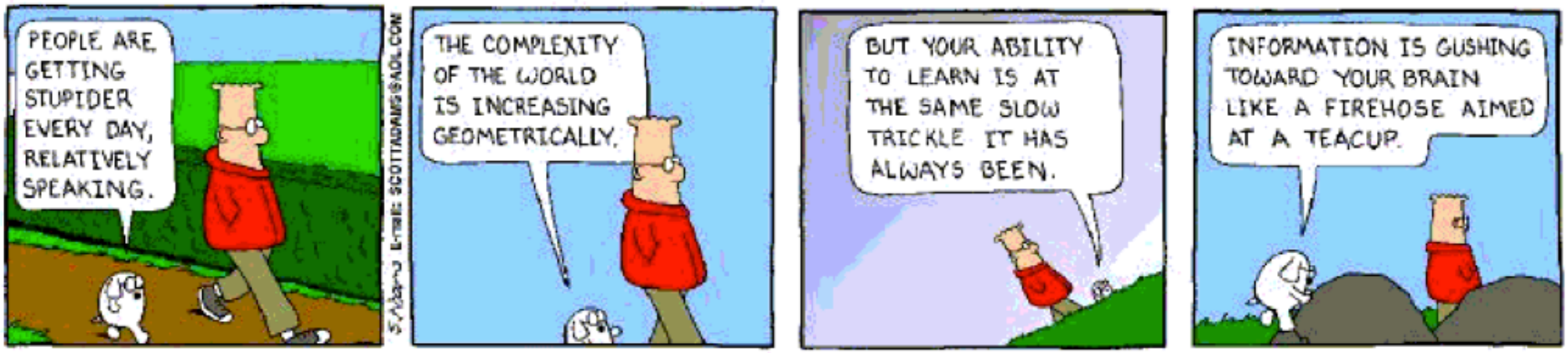
Introduction to Data Visualization

Naren Ramakrishnan

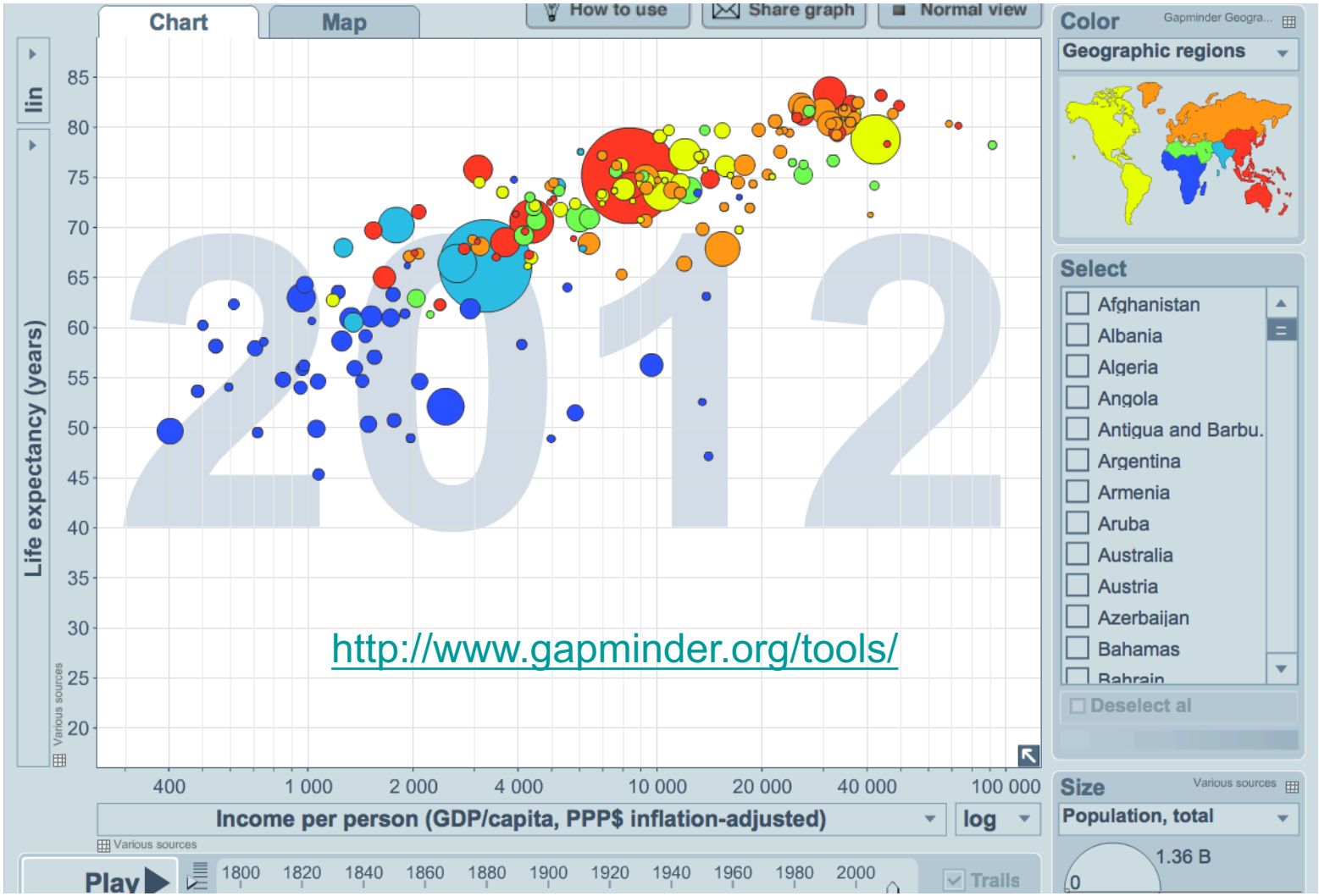
What is Data Visualization?

- **Data visualization** seeks to provide interactive visual representations of abstract datasets, designed to support human cognition and help people carry out analysis tasks more effectively.

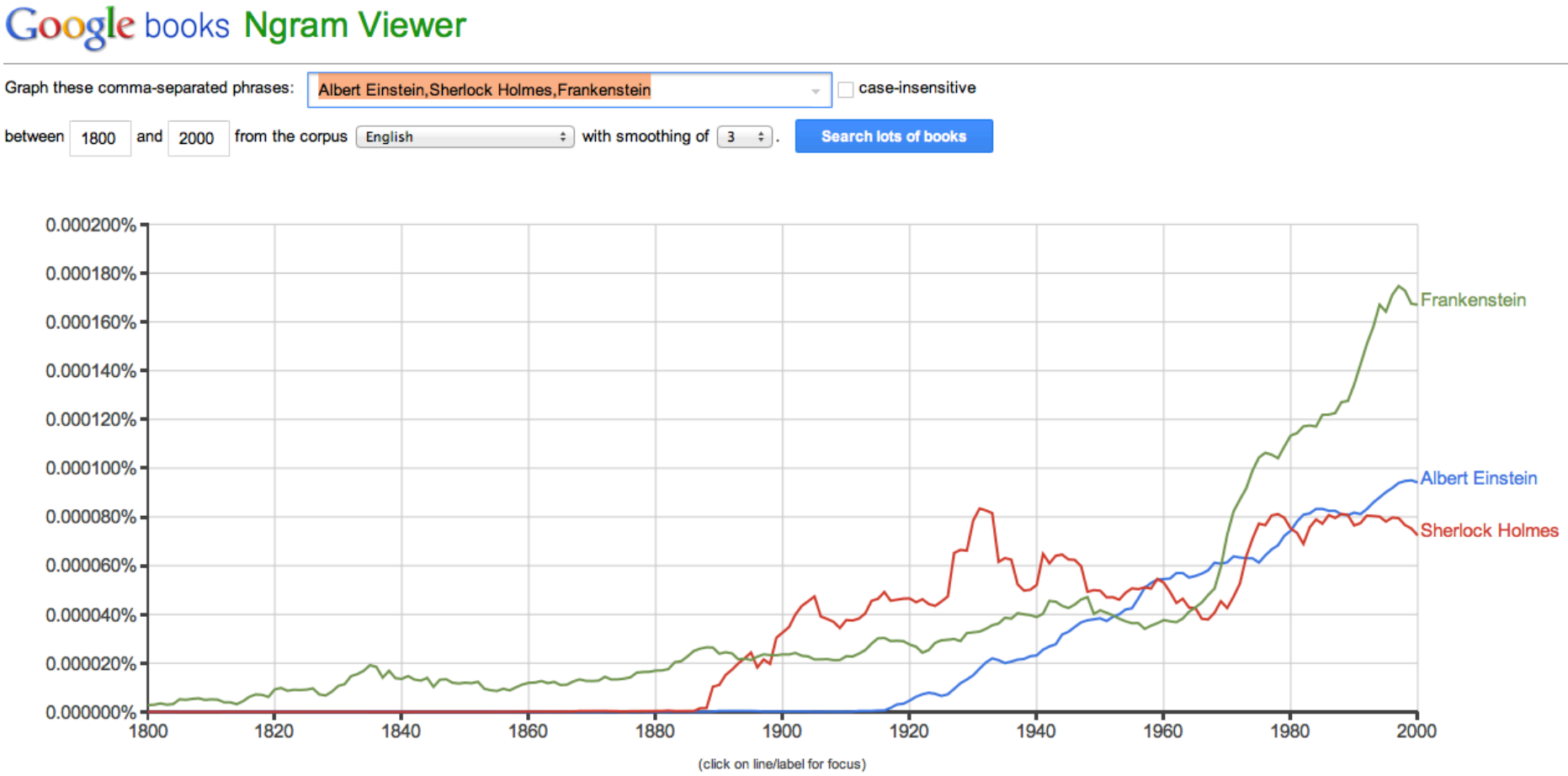
  --adapted from: T. Munzner in Visualization Analysis & Design



What is Data Visualization?



https://books.google.com/ngrams



Why Visual Representations?

- Which state has the highest Income? Average? Distribution?
- Relationship between Income and Education? Outliers?

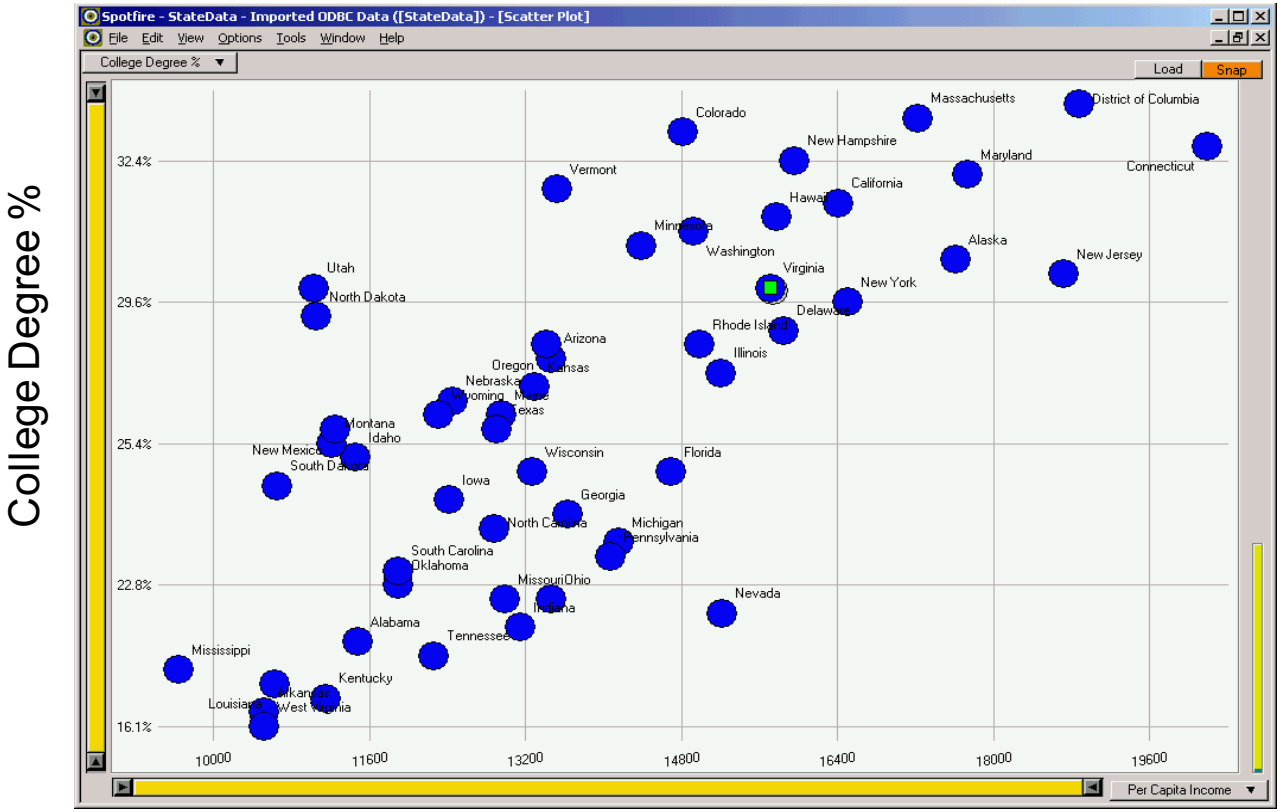
Table - StateData ()

Load Snap

State	College Degree %	Per Capita Income
Alabama	20.6%	11486
Alaska	30.3%	17610
Arizona	27.1%	13461
Arkansas	17.0%	10520
California	31.3%	16409
Colorado	33.9%	14821
Connecticut	33.8%	20189
Delaware	27.9%	15854
District of Columbia	36.4%	18881
Florida	24.9%	14698
Georgia	24.3%	13631
Hawaii	31.2%	15770
Idaho	25.2%	11457
Illinois	26.8%	15201
Indiana	20.9%	13149
Iowa	24.5%	12422
Kansas	26.5%	13300
Kentucky	17.7%	11153
Louisiana	19.4%	10635
Maine	25.7%	12957
Maryland	31.7%	17730
Massachusetts	34.5%	17224
Michigan	24.1%	14154
Minnesota	30.4%	14389
Mississippi	19.9%	9648
Missouri	22.3%	12989
Montana	25.4%	11213
Nebraska	26.0%	12452
Nevada	21.5%	15214
New Hampshire	32.4%	15959
New Jersey	30.1%	18714
New Mexico	25.5%	11246
New York	29.6%	16501
North Carolina	24.2%	12885
North Dakota	28.1%	11051
Ohio	22.3%	13461
Oklahoma	22.8%	11893
Oregon	27.5%	13418
Pennsylvania	23.2%	14068
Rhode Island	27.5%	14981
South Carolina	23.0%	11897
South Dakota	24.6%	10661
Tennessee	20.1%	12255
Texas	25.5%	12904
Utah	30.0%	11029
Vermont	31.5%	13527
Virginia	30.0%	15713
Washington	30.9%	14923
West Virginia	16.1%	10520
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

Why Visual Representations?

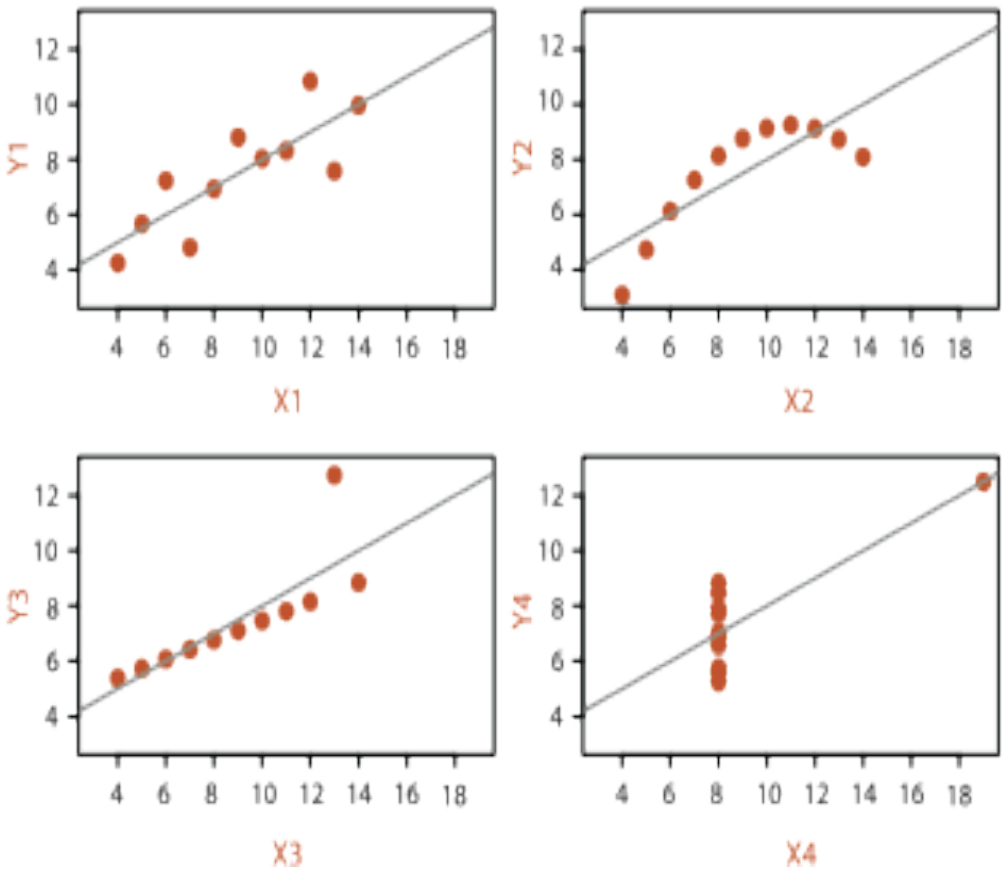
- Which state has highest Income? Average? Distribution?
- Relationship between Income and Education? Outliers?



Why show Data in Detail?

Anscombe's Quartet

Identical statistics	
x mean	9
x variance	11
y mean	7.5
y variance	4.1
correlation	0.8
regression	y = 3 + 0.5x



It is about “Insight”

- Easy stuff:
  - Reduce to only 1 data item or value
  - Stats: Min, max, average, %
  - Search: known item

Tables can do this

- Hard stuff:
  - Require seeing the whole
  - Patterns: distributions, trends, frequencies, structures
  - Outliers: exceptions
  - Relationships: correlations, multi-way interactions
  - Tradeoffs: combined min/max
  - Comparisons: choices (1:1), context (1:M), sets (M:M)
  - Clusters: groups, similarities
  - Anomalies: data errors
  - Paths: distances, ancestors, decompositions, …

Visualization can do this

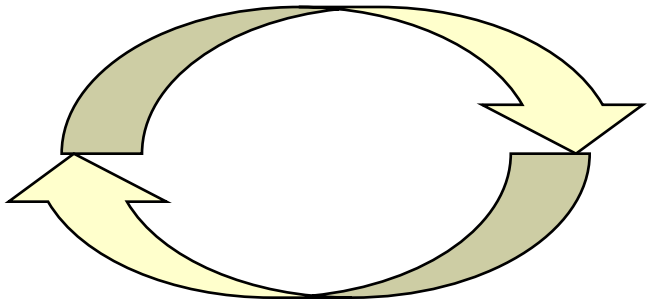
It is about “Insight”

Two types of Insight

Informal

Seeing the whole
Complex patterns
High-level understanding
Unforeseen
Domain relevant
Hard to measure

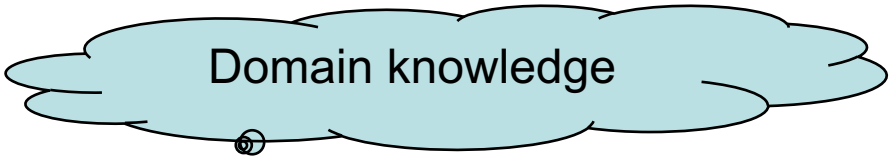
Usability:
Efficiently transition



Hypothesis generation

Formal

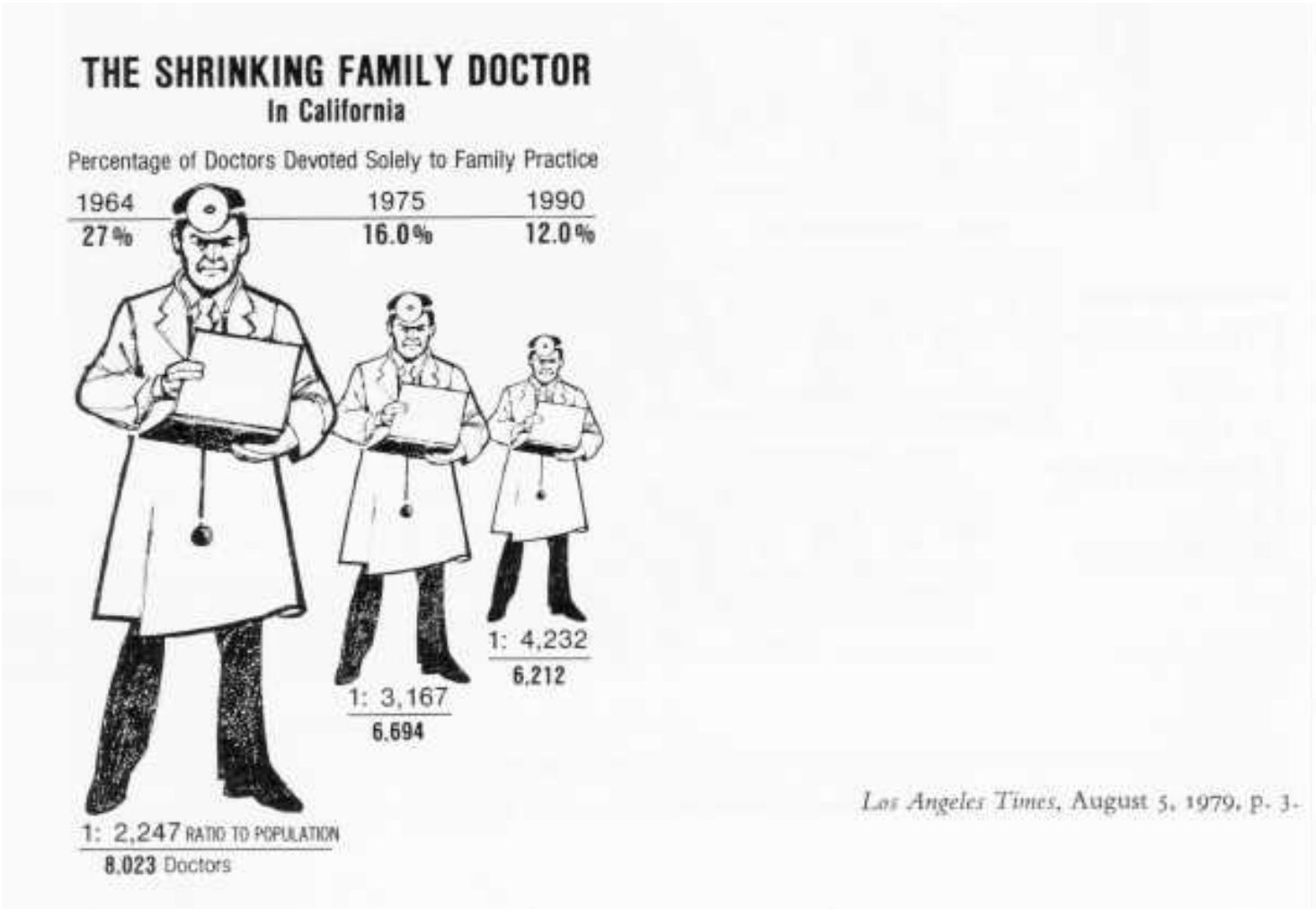
Aggregated, specific
Symbolic
Low-level tasks
Predefined
Domain independent
Measurable



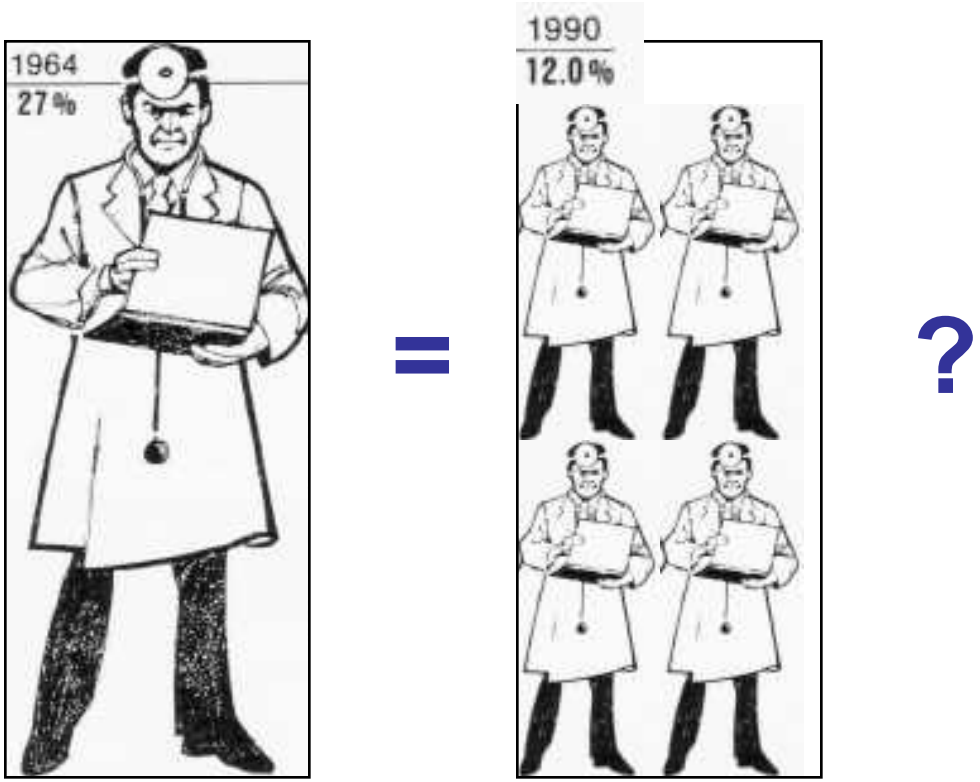
Visual Illiteracy

$$Lie\ Factor = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

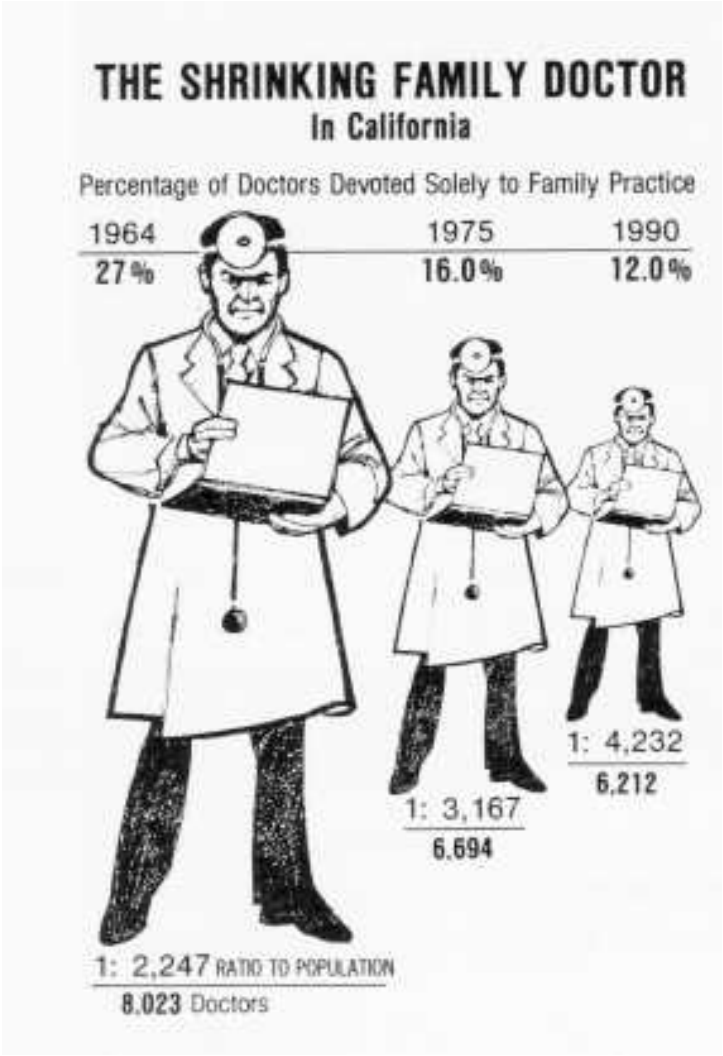
Visual Illiteracy

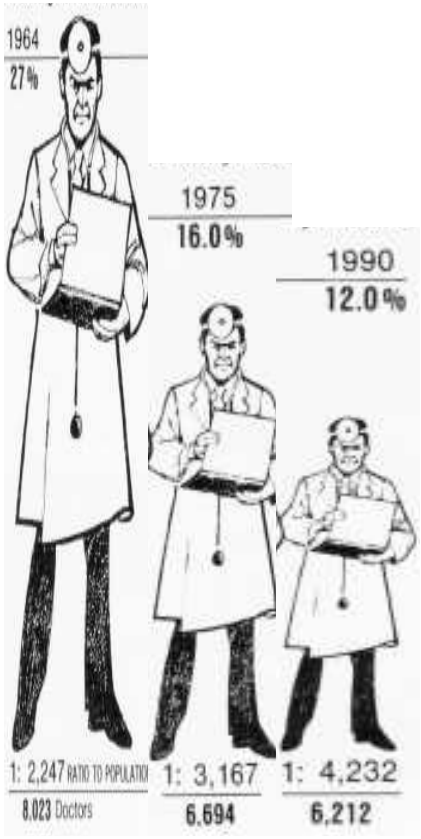


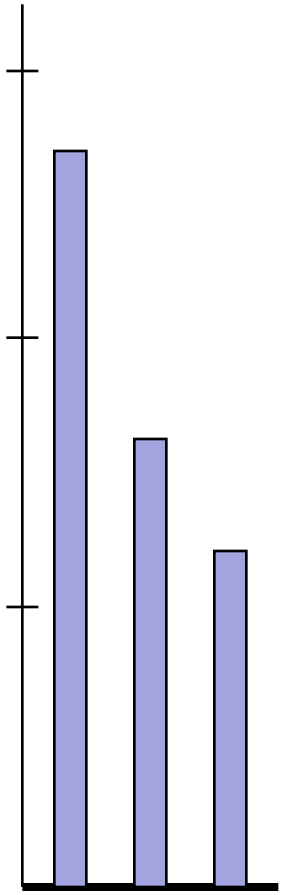
Size Encoding: 2D Area?



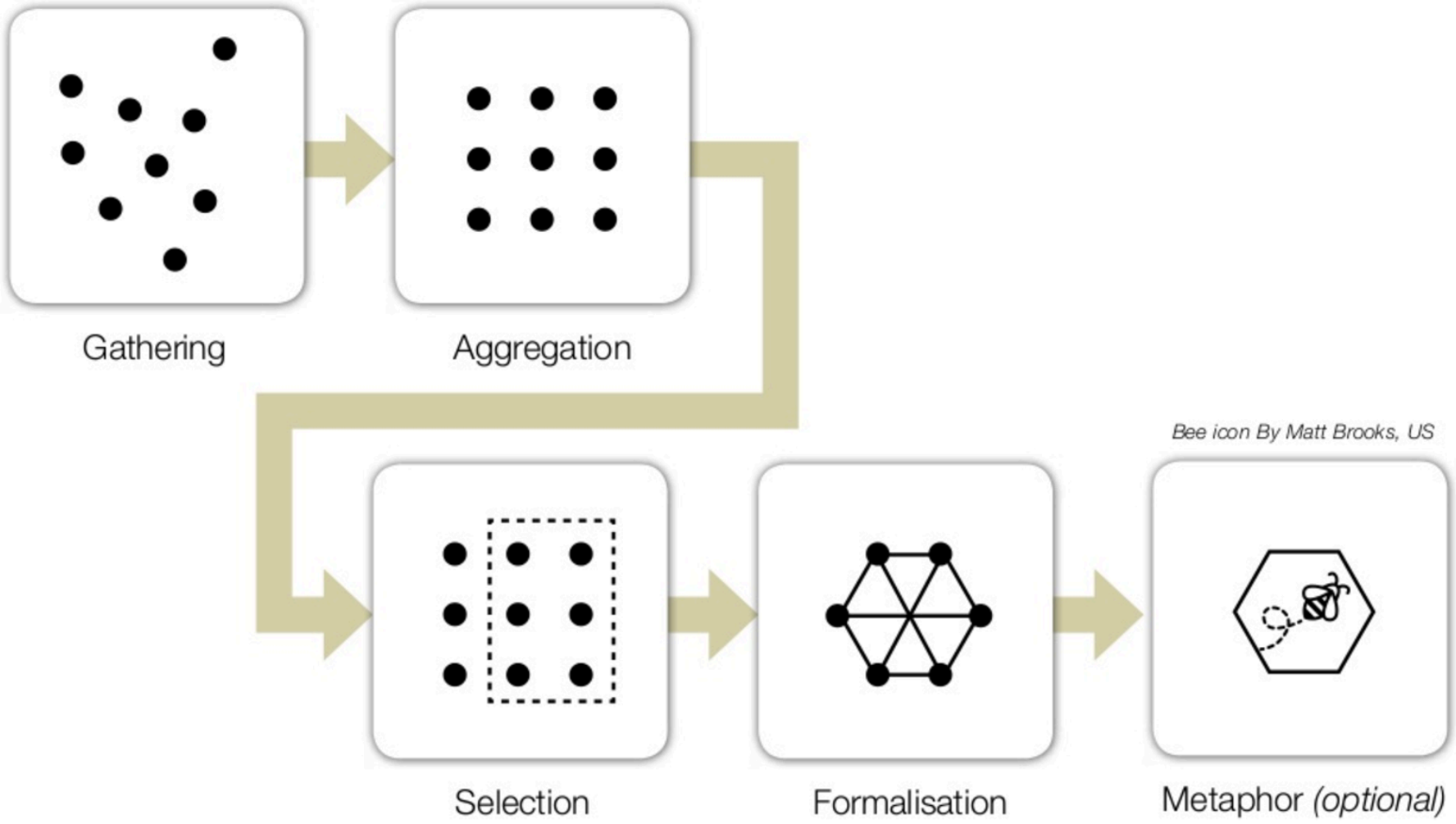
Size Encoding: 1D Height?







Data Visualization Process



Credit: Davide Genco

A Tour Through The Visualization Jungle

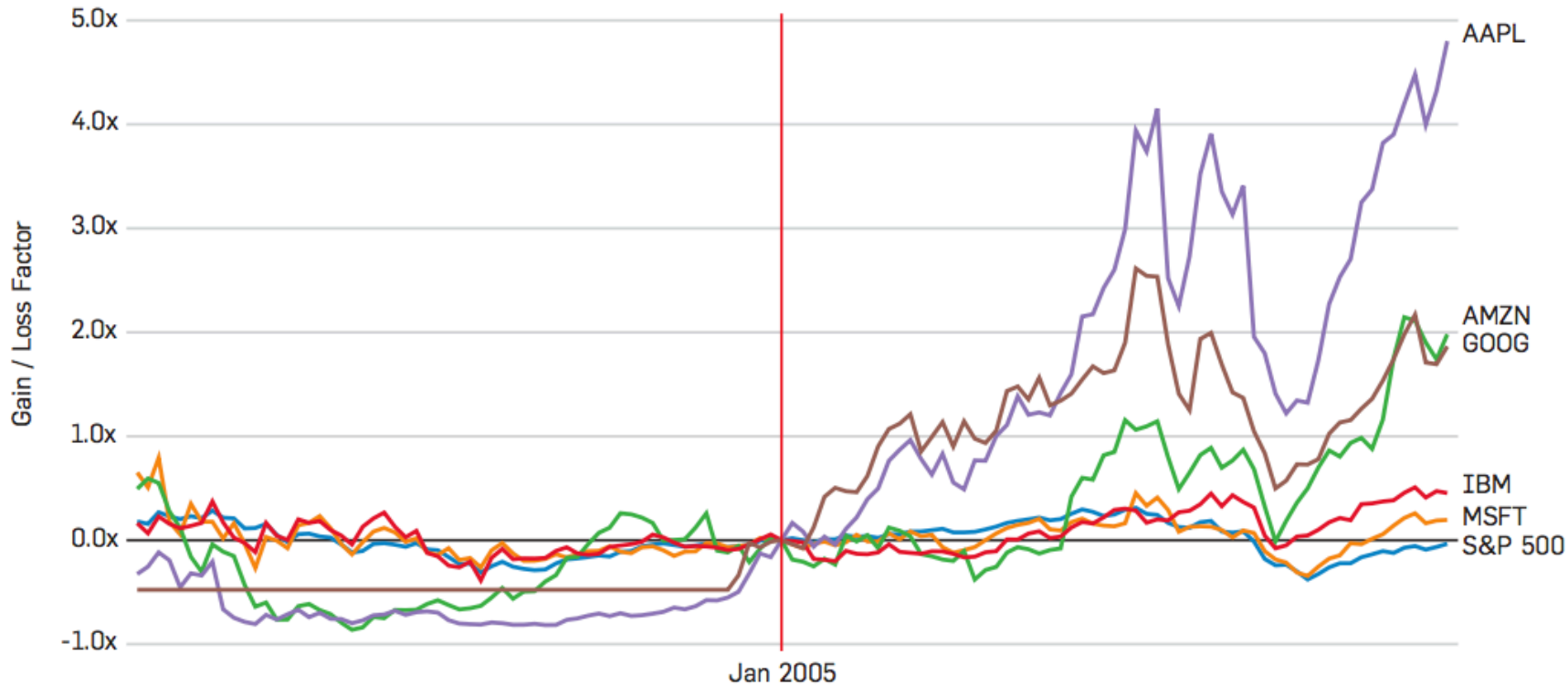
A tour through the visualization Jungle
By Jeffrey Heer et al.

Time Series Data

- Sets of values changing over time
- Central to many domains such as
  - finance (stock prices, exchange rates)
  - science (temperatures, pollution levels, electric potentials), and
  - public policy (crime rates) etc.
- One often needs to compare a large number of time series simultaneously

Index Charts

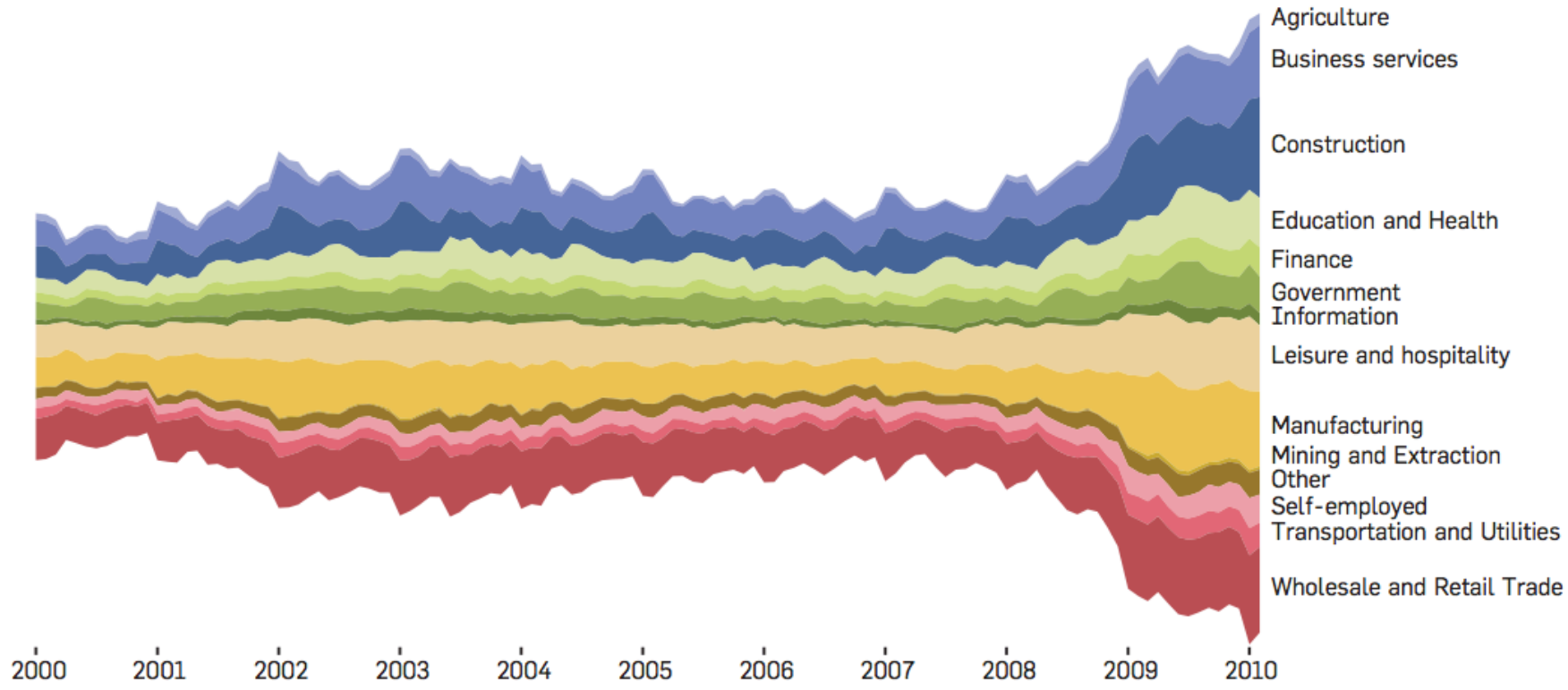
Time-Series Data: **Figure 1a. Index chart of selected technology stocks, 2000–2010.**



Source: Yahoo! Finance; http://hci.stanford.edu/jheer/files/zoo/ex/time/index-chart.html

Stacked Graphs

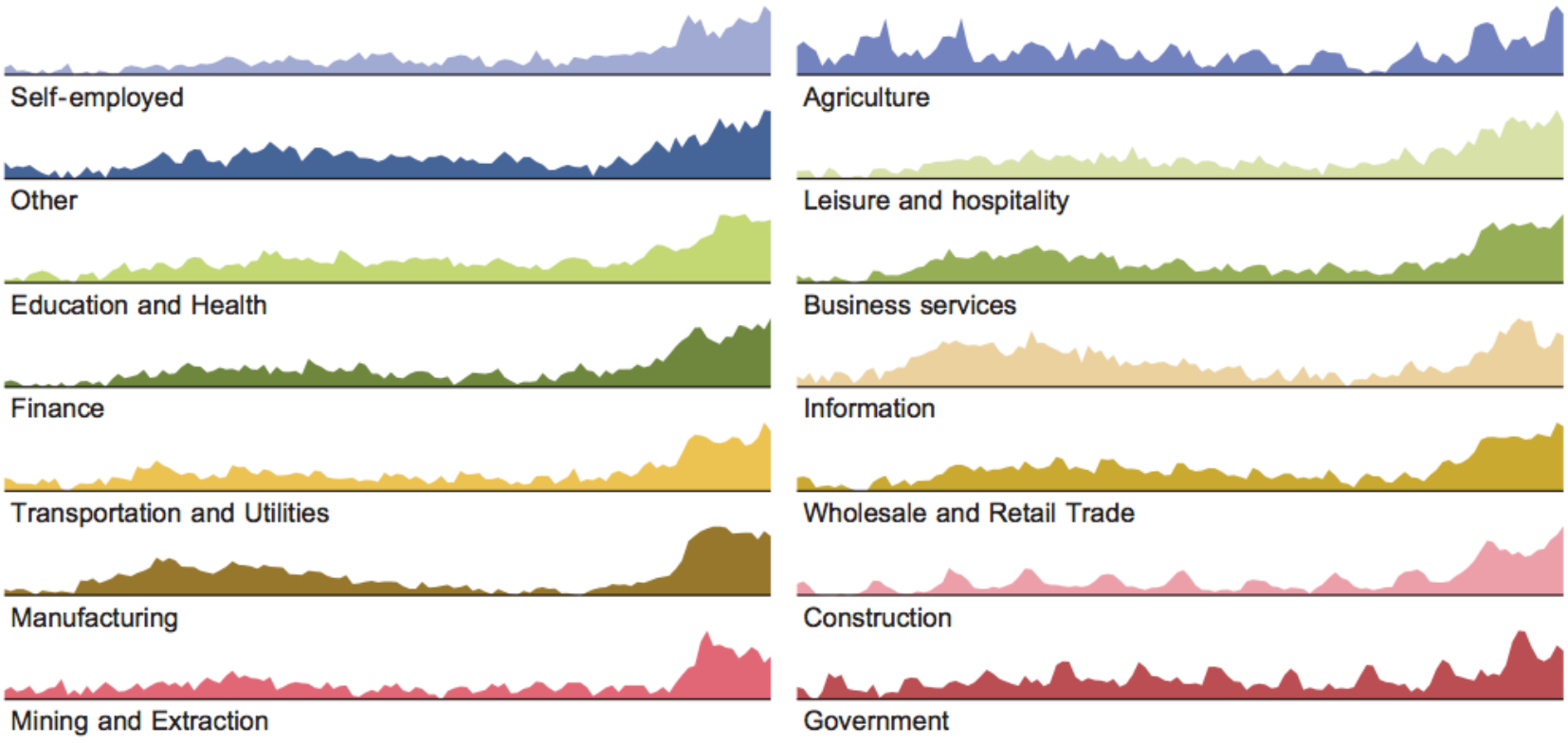
Time-Series Data: **Figure 1b. Stacked graph of unemployed U.S. workers by industry, 2000–2010.**



Source: U.S. Bureau of Labor Statistics; http://hci.stanford.edu/jheer/files/zoo/ex/time/stack.html

Small Multiples

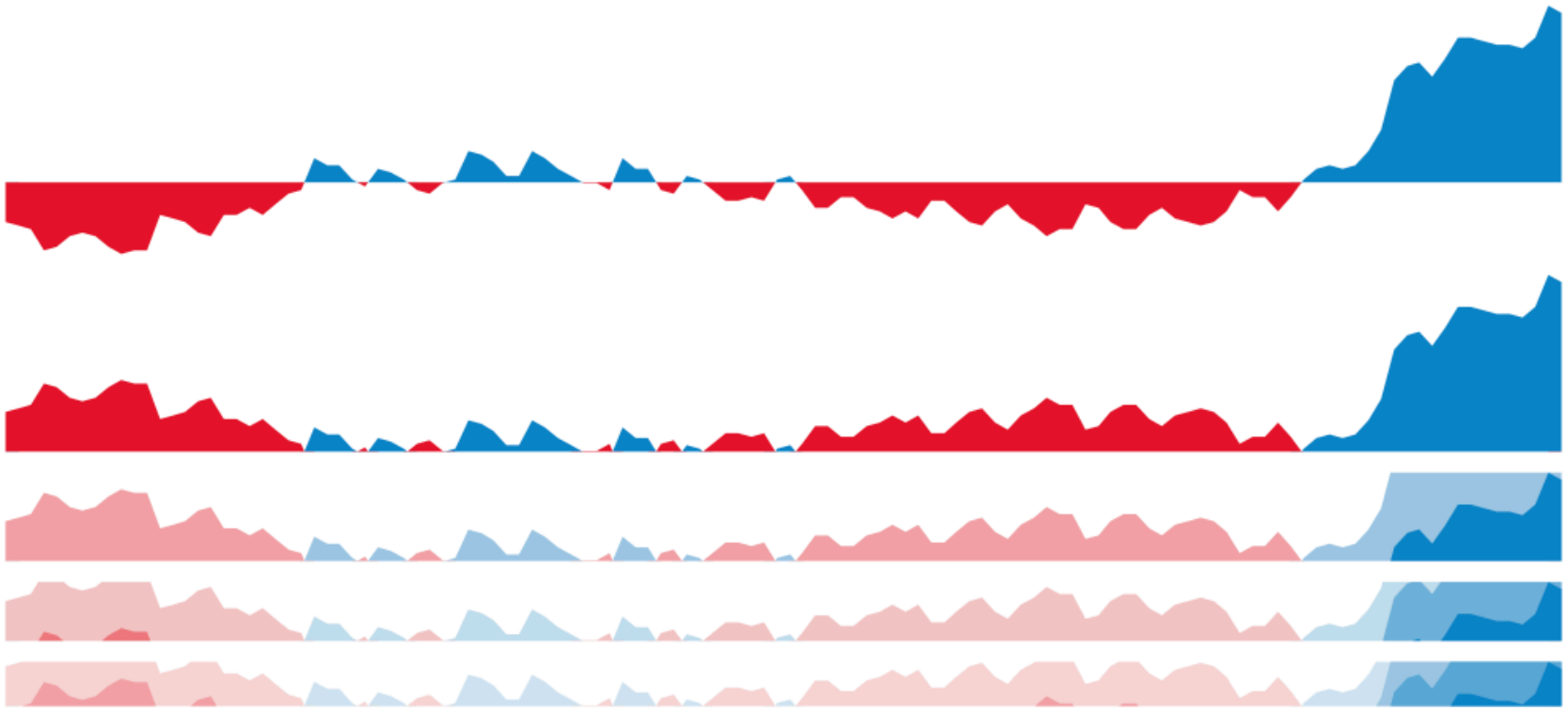
Time-Series Data: **Figure 1c. Small multiples of unemployed U.S. workers, normalized by industry, 2000–2010.**



Source: U.S. Bureau of Labor Statistics; http://hci.stanford.edu/jheer/files/zoo/ex/time/multiples.html

Horizon Graphs

Time-Series Data: **Figure 1d. Horizon graphs of U.S. unemployment rate, 2000–2010.**



Source: U.S. Bureau of Labor Statistics; http://hci.stanford.edu/jheer/files/zoo/ex/time/horizon.html

Statistical Distributions

- Visualizations designed to reveal how a set of numbers is distributed and thus help an analyst better understand the statistical properties of the data
- Important use of visualizations is exploratory data analysis – gaining insight into how data is distributed to inform data transformation and modeling decisions
- Common techniques include:
  - histogram, which shows the prevalence of values grouped into bins, and
  - box-and-whisker plot, which can convey statistical features such as the mean, median, quartile boundaries, or extreme outliers

Stem and Leaf Plots

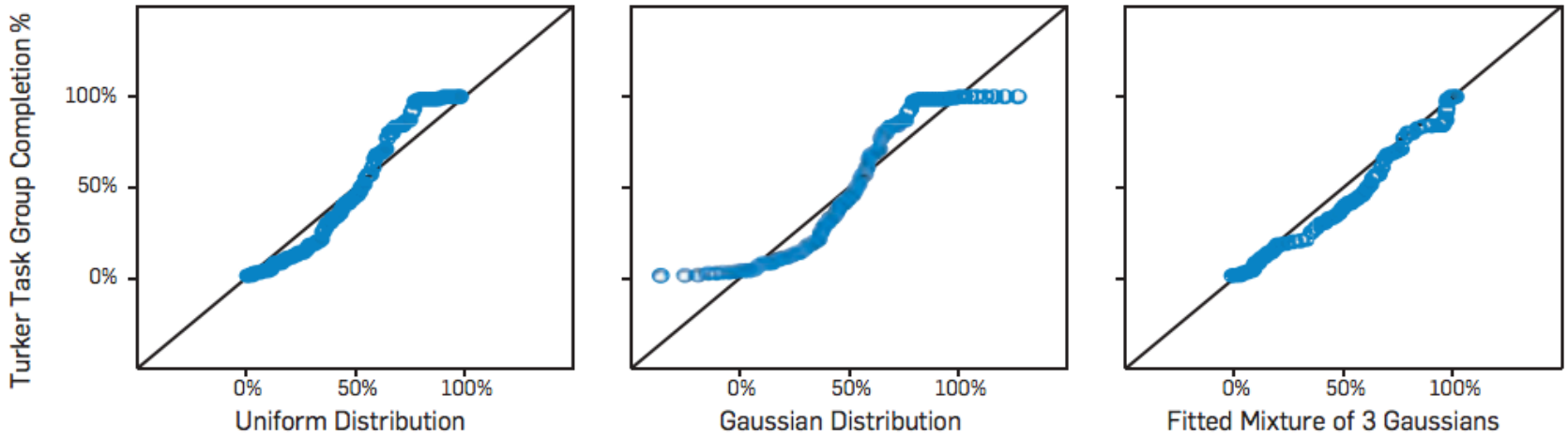
Statistical Distributions: Figure 2a. Stem-and-leaf plot of Mechanical Turk participation rates.

0	1	1	1	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5	6	7	8	8	8	8	8	8	9
1	0	0	0	0	1	1	1	1	2	2	3	3	3	3	4	4	4	4	5	5	6	7	7	8	9	9	9	9	9			
2	0	0	1	1	1	5	7	8	9																							
3	0	0	1	2	3	3	3	4	6	6	8	8																				
4	0	0	1	1	1	1	3	3	4	5	5	5	6	7	8	9																
5	0	2	3	5	6	7	7	7	9																							
6	1	2	6	7	8	9	9	9																								
7	0	0	0	1	6	7	9																									
8	0	0	1	2	3	4	4	4	4	4	4	4	5	6	7	7	7	9														
9	1	3	3	5	7	8	8	8	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9						
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0													

Source: Stanford Visualization Group; http://hci.stanford.edu/jheer/files/zoo/ex/stats/stem-and-leaf.html

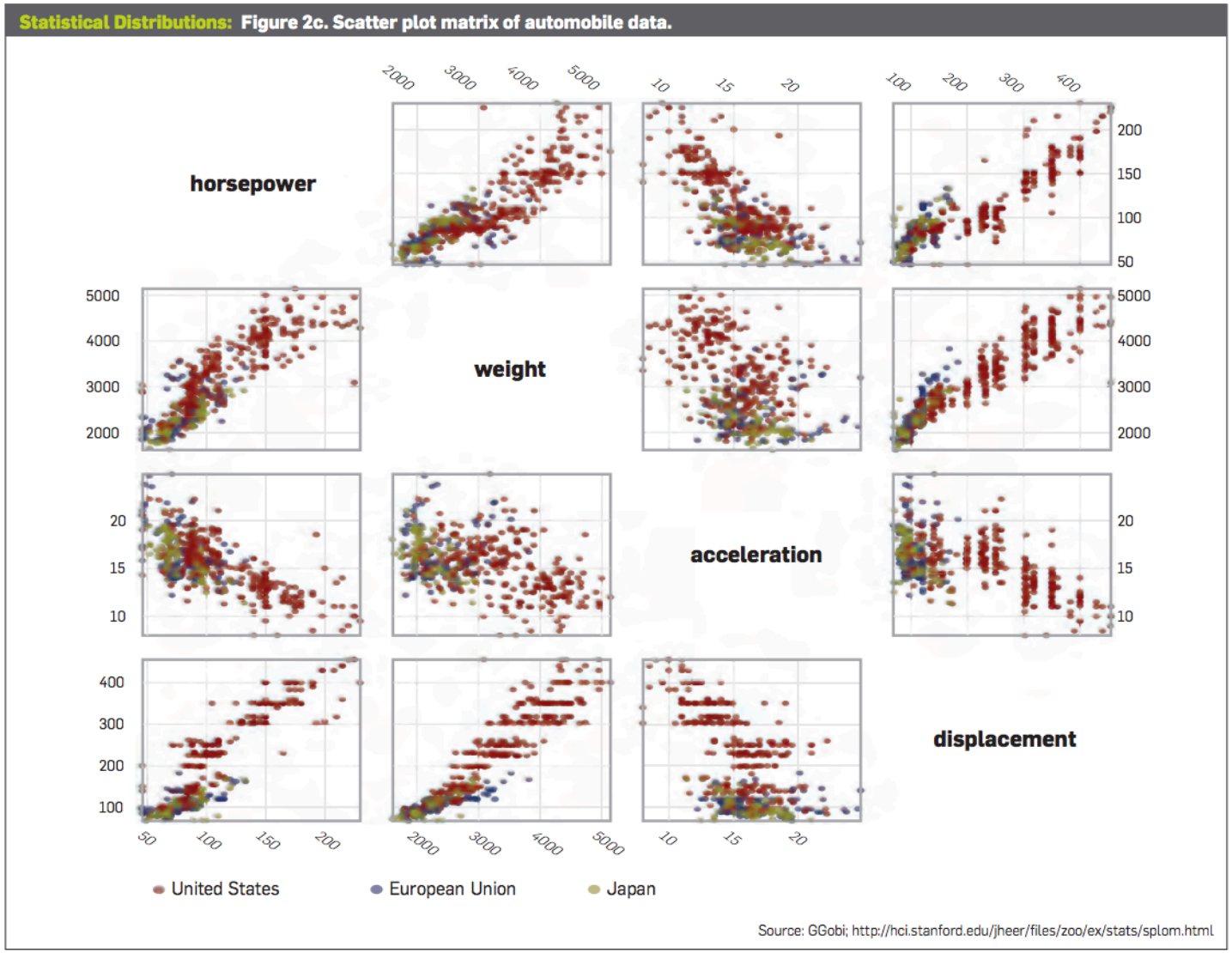
QQ Plots

Statistical Distributions: **Figure 2b. Q-Q plots of Mechanical Turk participation rates.**

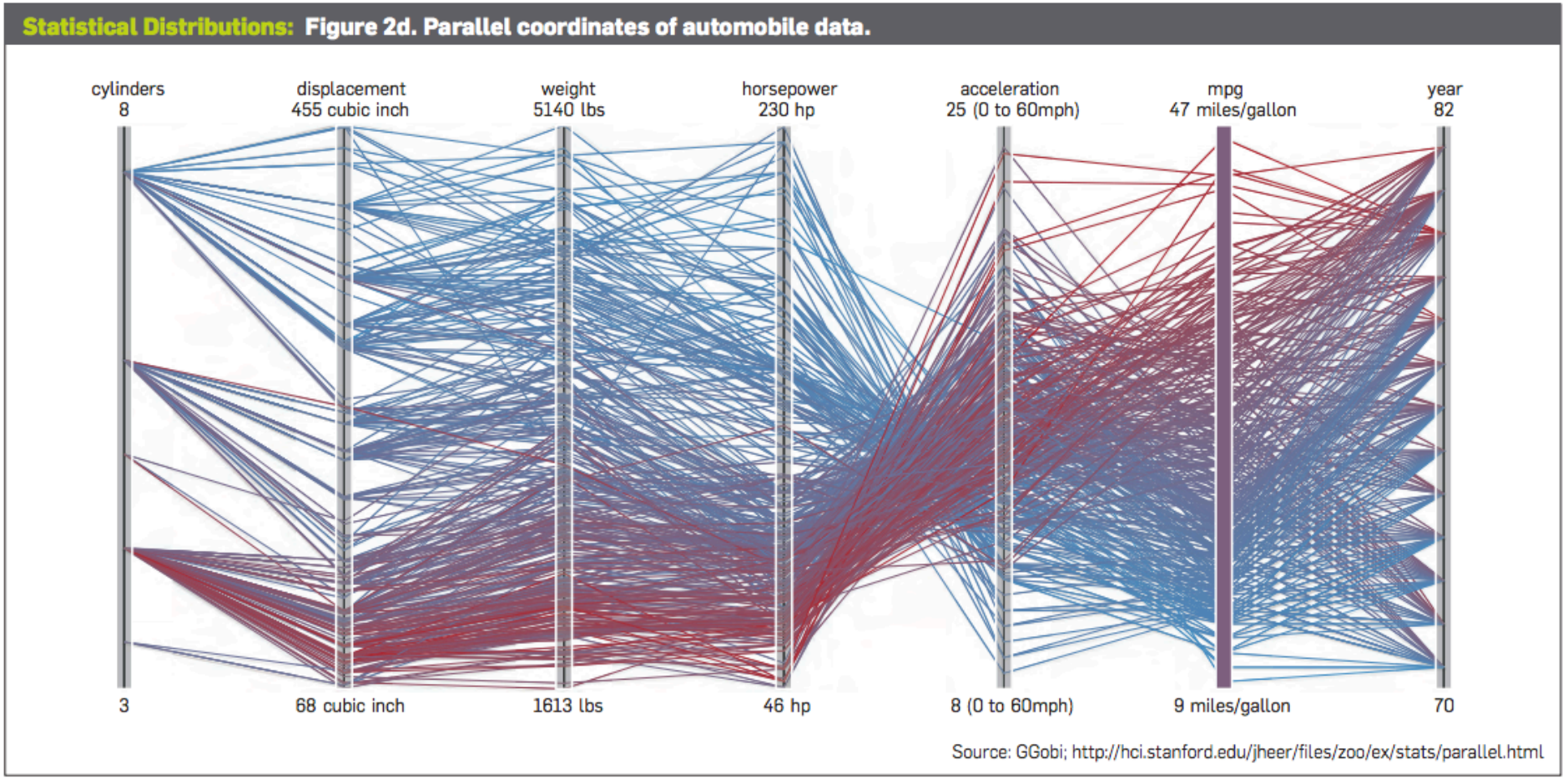


Source: Stanford Visualization Group; http://hci.stanford.edu/jheer/files/zoo/ex/stats/qqplot.html

SPLOM (Scatter Plot Matrix)



Parallel Coordinates



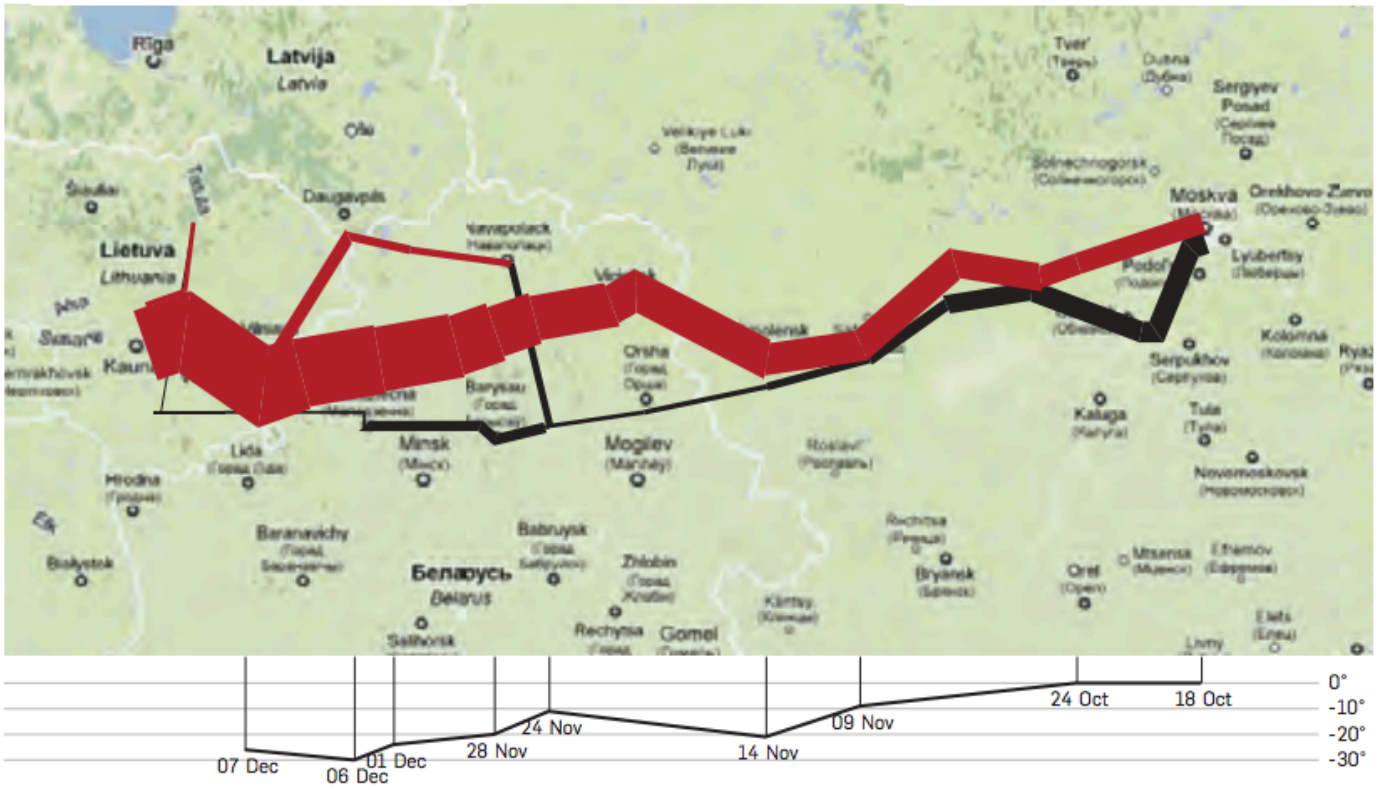
Maps

- Many maps are based upon a cartographic projection:
  - a mathematical function that maps the 3D geometry of the Earth to a 2D image
- Other maps knowingly distort or abstract geographic features to tell a richer story or highlight specific data

Good example of map distortion

Flow Maps

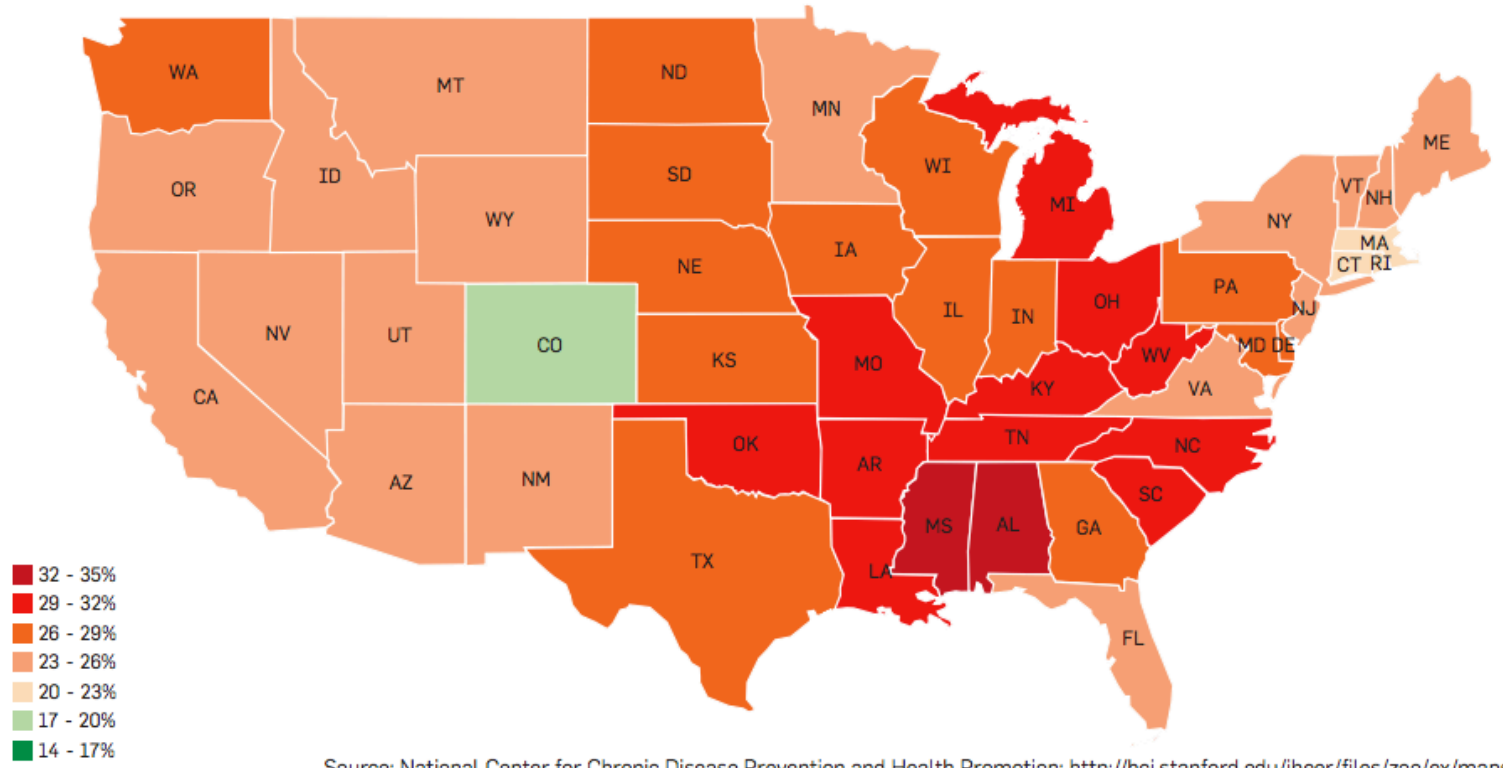
Maps: Figure 3a. Flow map of Napoleon's March on Moscow, based on the work of Charles Minard.



http://hci.stanford.edu/jheer/files/zoo/ex/maps/napoleon.html

Choropleth Maps

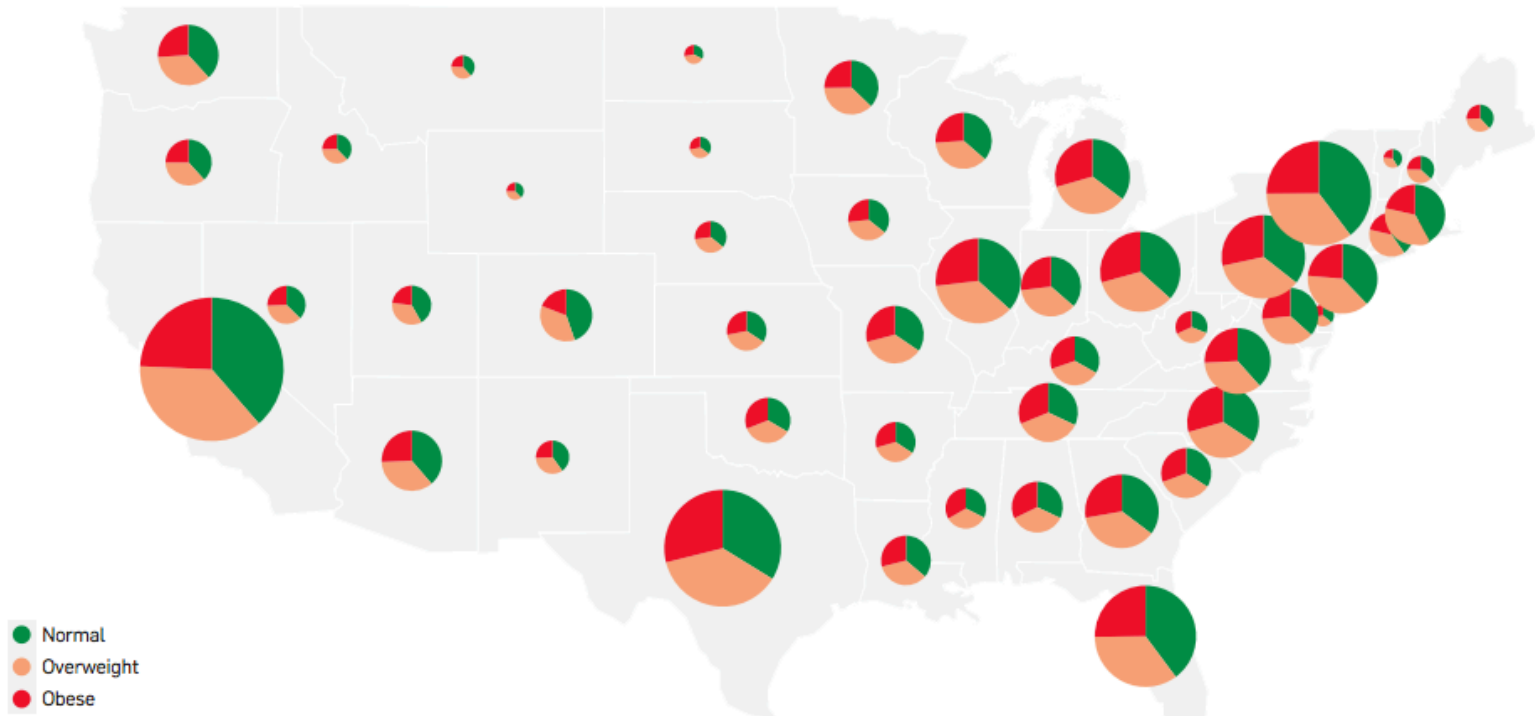
Maps: Figure 3b. Choropleth map of obesity in the U.S., 2008.



Source: National Center for Chronic Disease Prevention and Health Promotion; http://hci.stanford.edu/jheer/files/zoo/ex/maps/choropleth.html

Graduated Symbol Maps

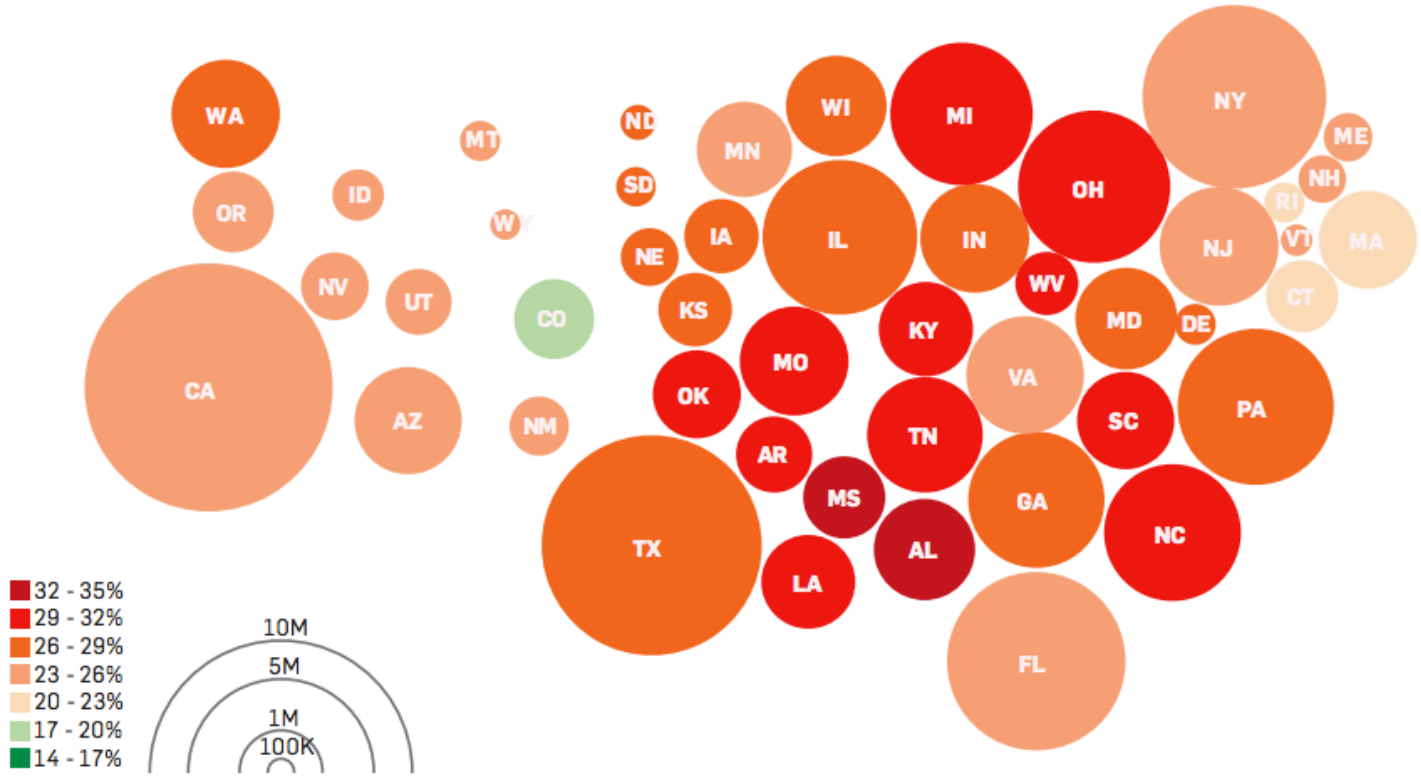
Maps: Figure 3c. Graduated symbol map of obesity in the U.S., 2008.



Source: National Center for Chronic Disease Prevention and Health Promotion; http://hci.stanford.edu/jheer/files/zoo/ex/maps/symbol.html

Cartograms

Maps: Figure 3d. Dorling cartogram of obesity in the U.S., 2008.

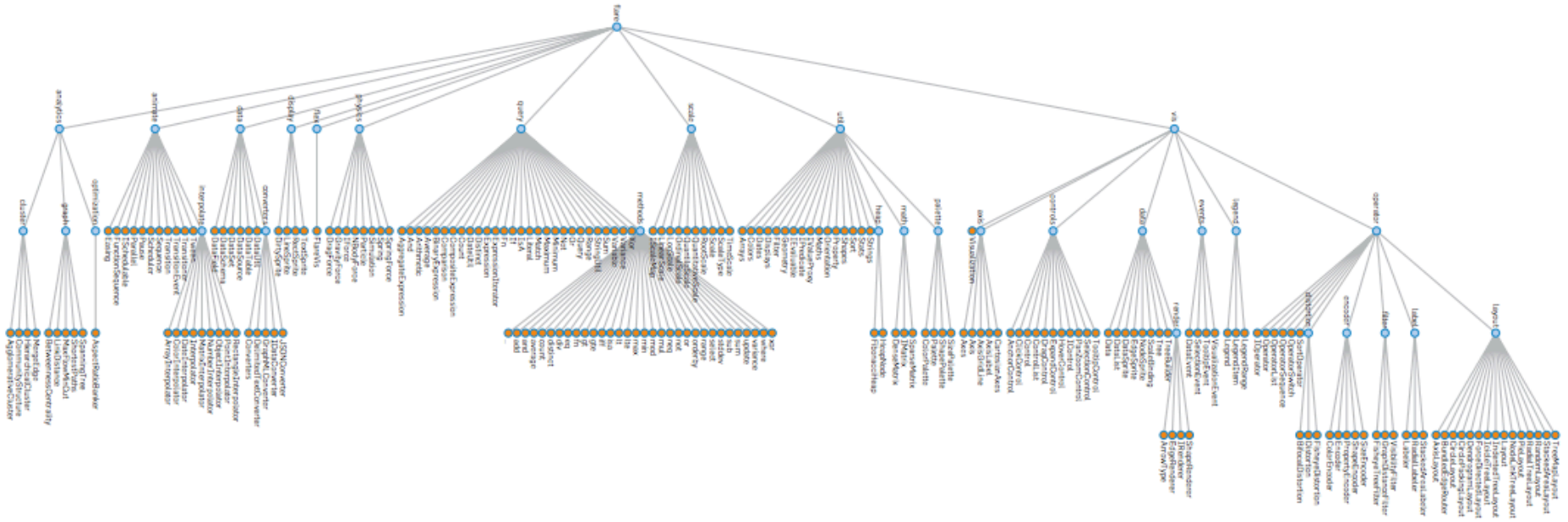


Source: National Center for Chronic Disease Prevention and Health Promotion; http://hci.stanford.edu/jheer/files/zoo/ex/maps/cartogram.html

Hierarchies

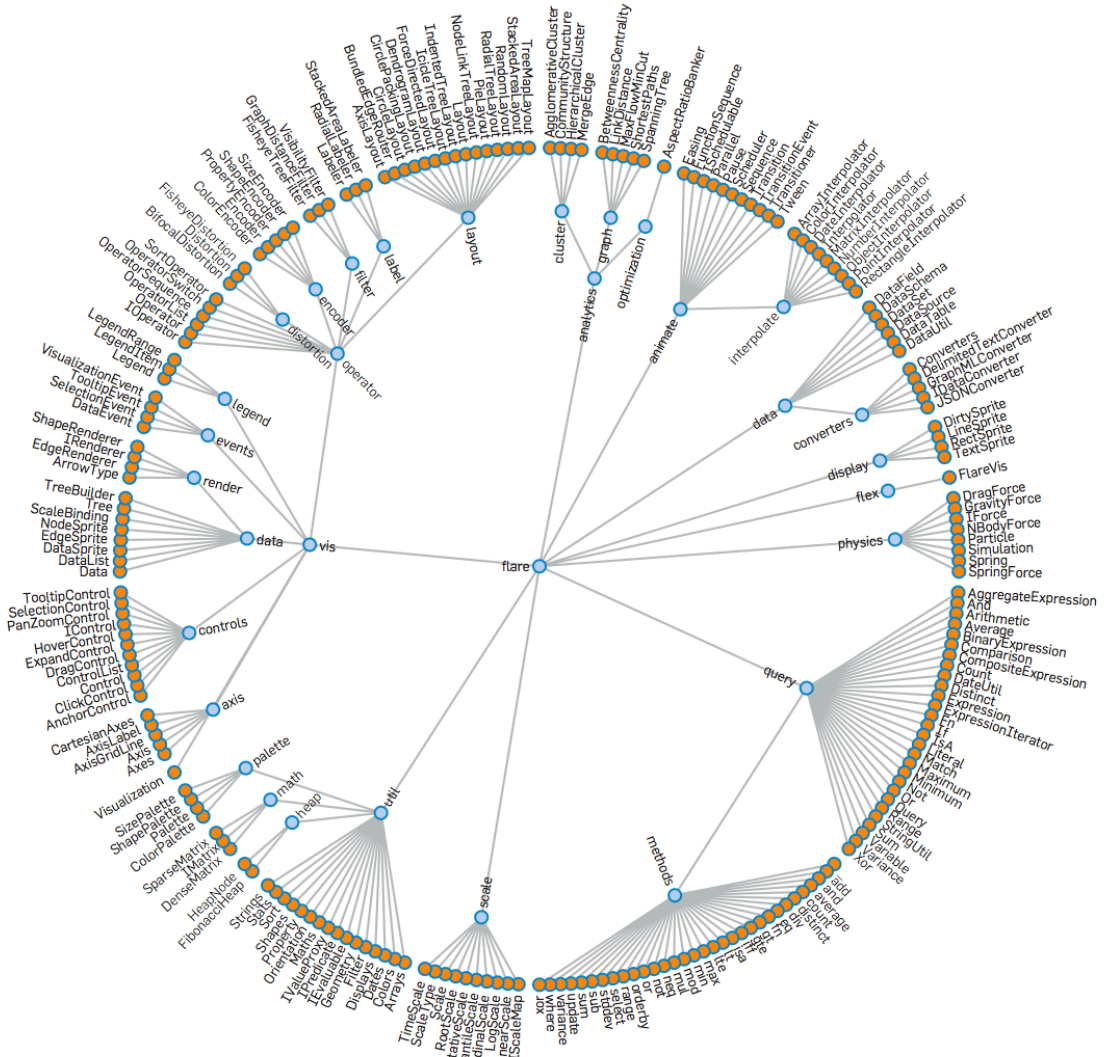
- Some data can be organized into natural hierarchies:
  - Spatial entities, such as counties, states, and countries
  - Command structures for businesses and governments
  - Software packages
  - Phylogenetic trees
- Even for data with no apparent hierarchy, statistical methods (for example, k-means clustering) may be applied to organize data empirically

Cartesian Node Link Diagram



http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/tree.html

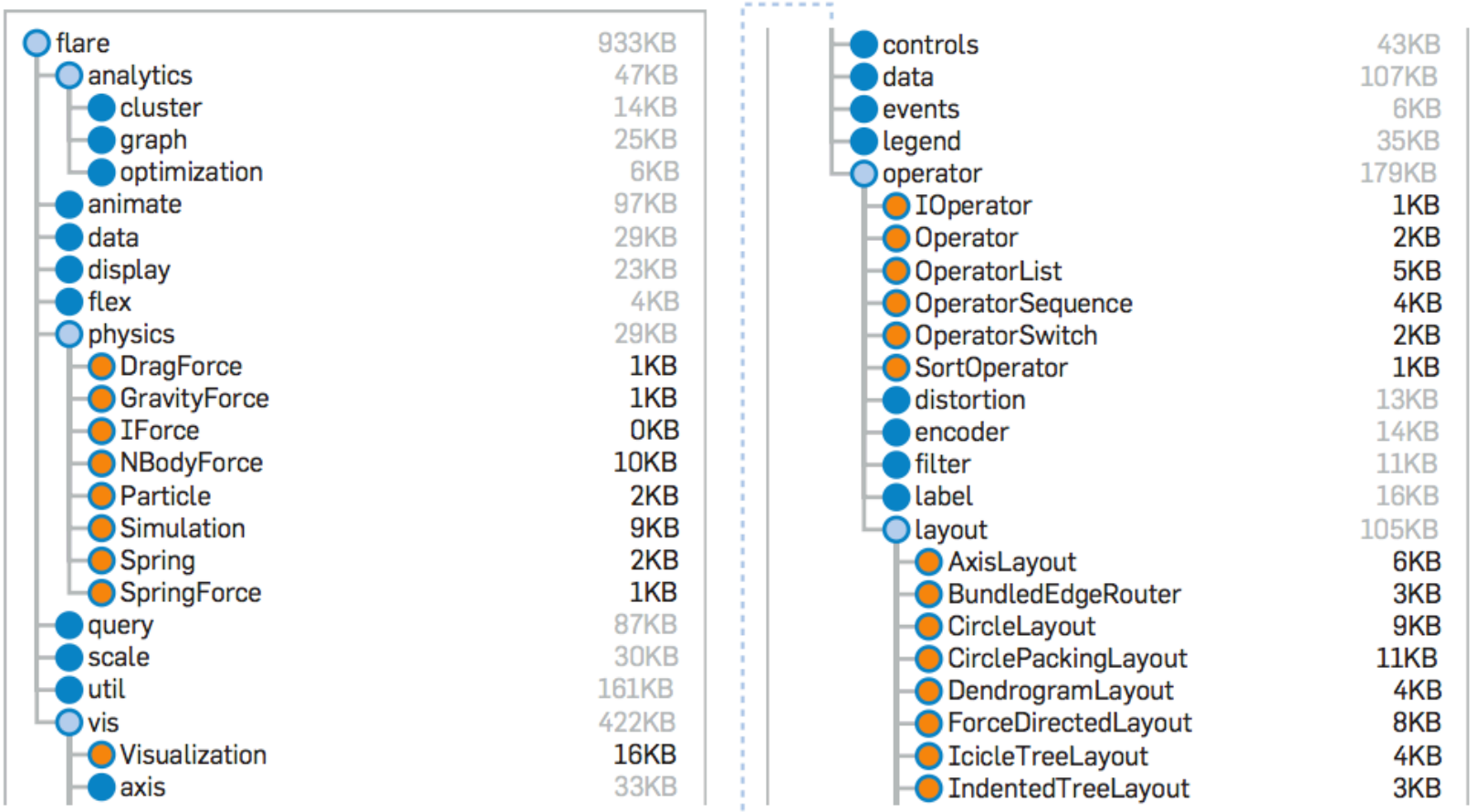
Radial Node Link Diagram



http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/cluster-radial.html

Indented Tree Layout

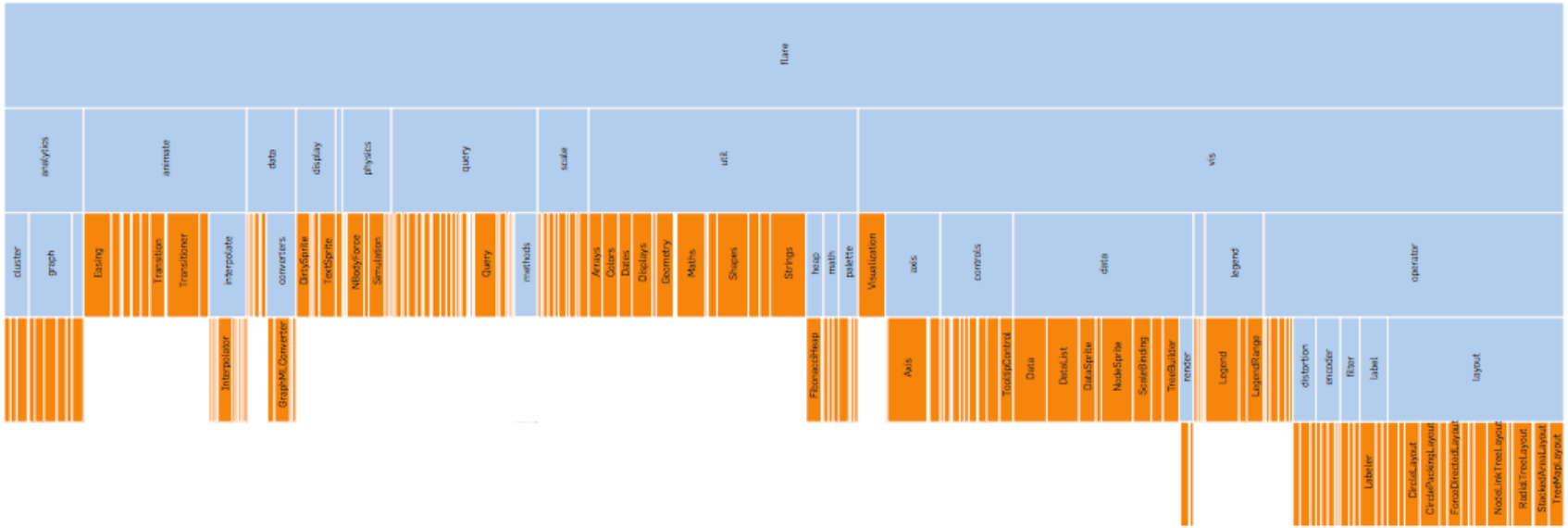
Hierarchies: **Figure 4c. Indented tree layout of the Flare package hierarchy.**



http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/indent.html

Icicle Tree Layout

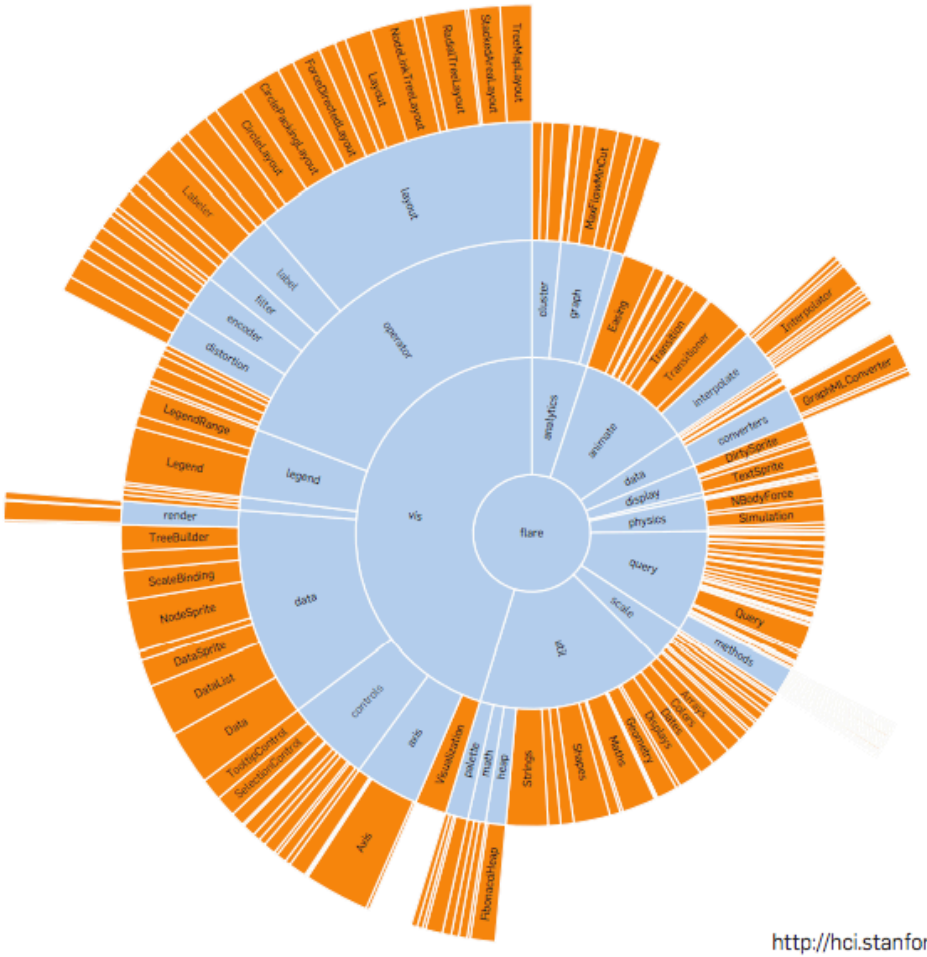
Hierarchies: Figure 4d. Icicle tree layout of the Flare package hierarchy.



http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/icicle.html

Sunburst Layout

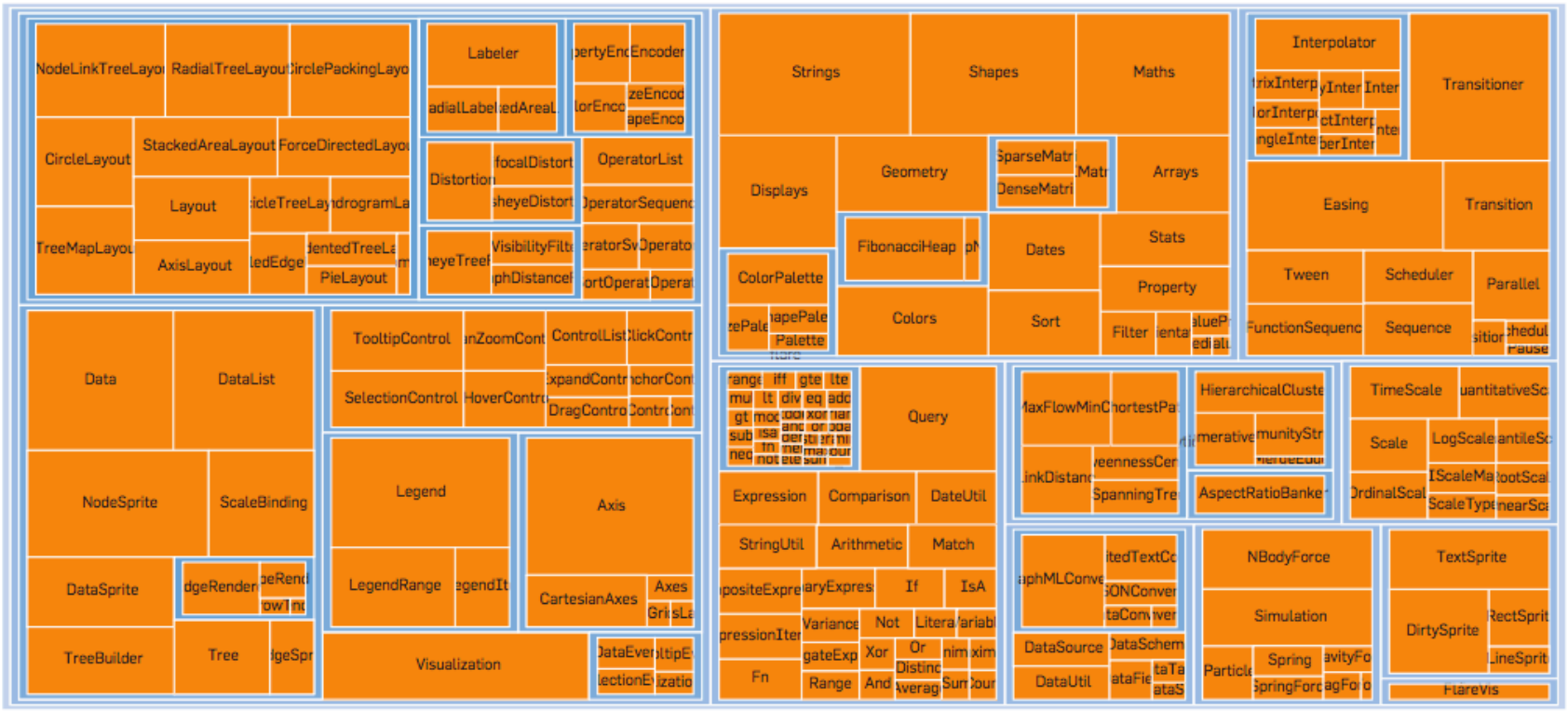
Hierarchies: **Figure 4e. Sunburst (radial space-filling) layout of the Flare package hierarchy.**



http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/sunburst.html

Treemap Layout

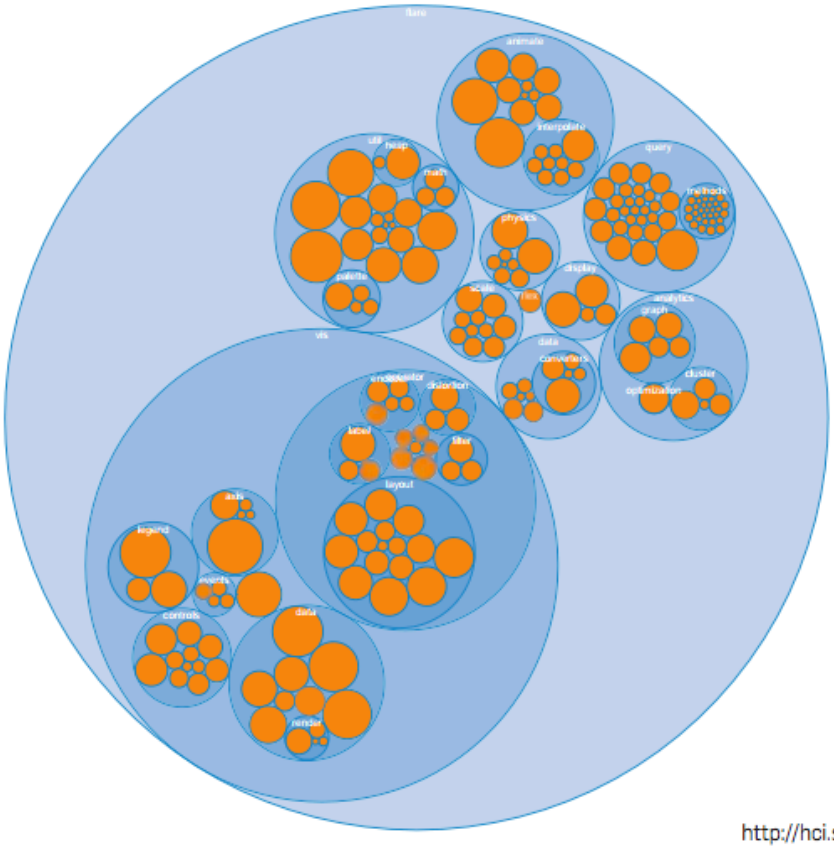
Hierarchies: **Figure 4f. Treemap layout of the Flare package hierarchy.**



http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/treemap.html

Nested Circles Layout

Hierarchies: Figure 4g. Nested circles layout of the Flare package hierarchy.

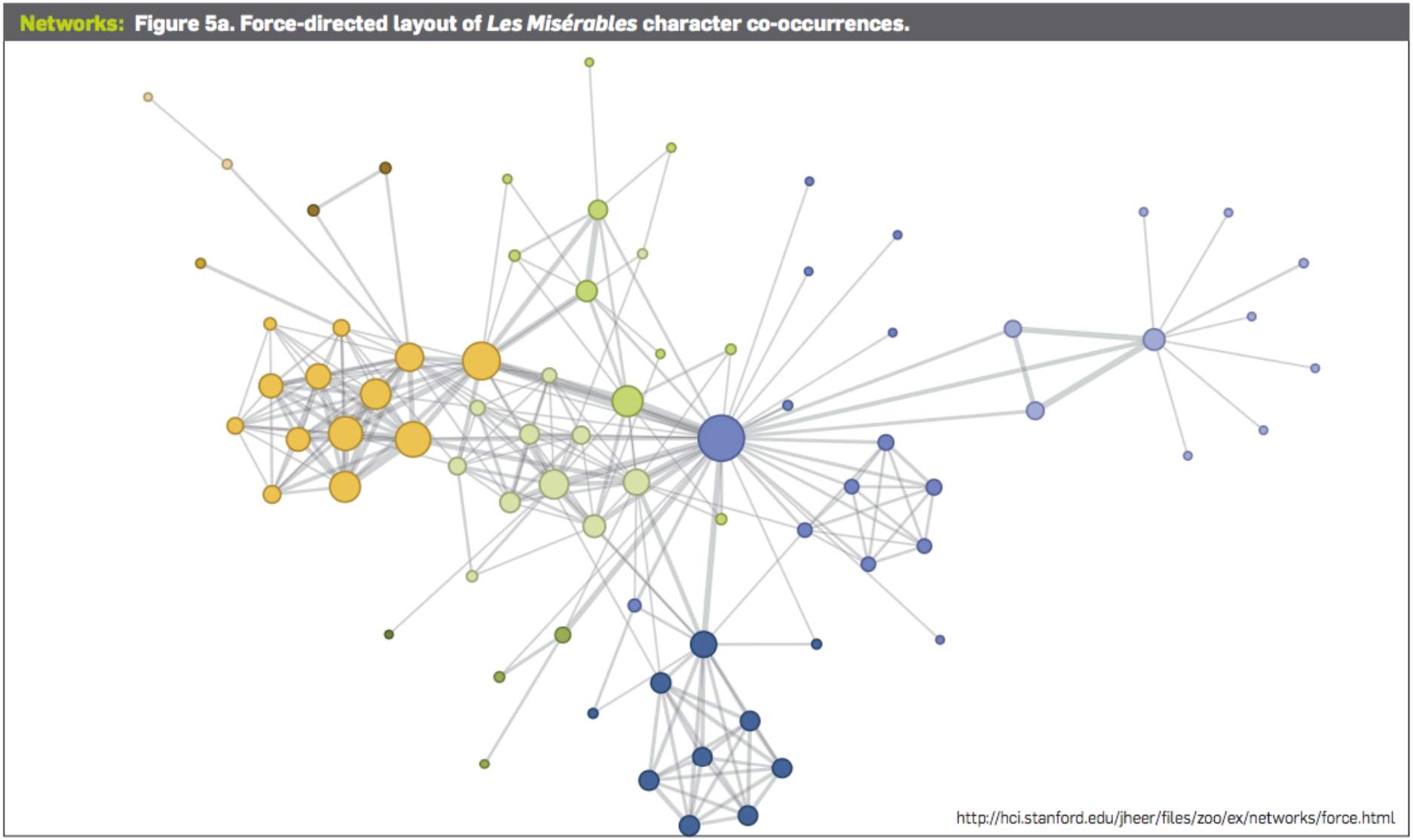


http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/pack.html
Source: The Flare Toolkit http://flare.prefuse.org

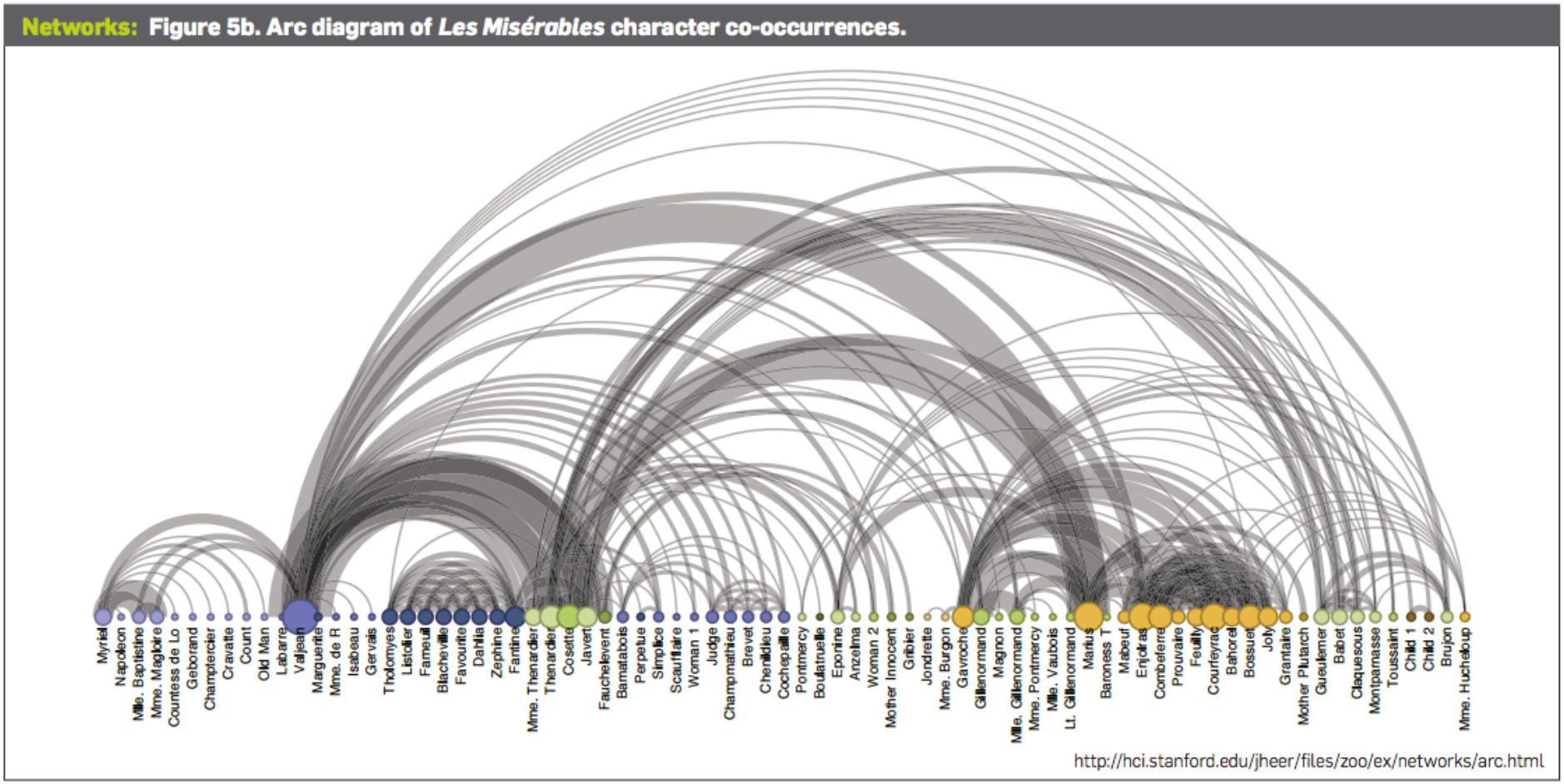
Networks

- Relationships
  - given a social network, who is friends with whom?
  - Who are the central players?
  - What cliques exist?
  - Who, if anyone, serves as a bridge between disparate groups?
- Abstractly, a hierarchy is a specialized form of network:
  - Node Link diagrams can be used for networks
- Mathematicians use the formal term graph to describe a network

Force-Directed Layout

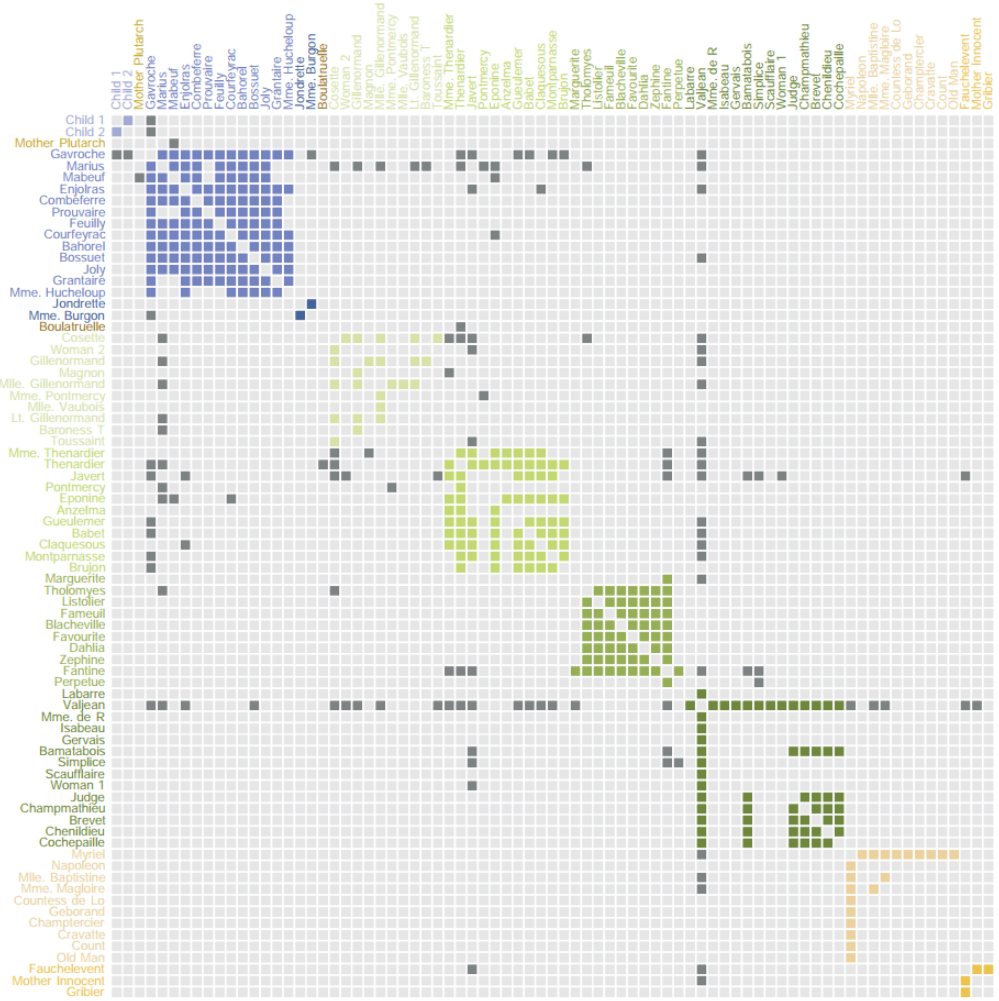


Arc Diagram


Networks: Figure 5b. Arc diagram of *Les Misérables* character co-occurrences.
Myriel
Napoleon
Mlle. Baptistine
Mme. Magloire
Countess de Lo
Geborand
Champtercier
Cravatte
Count
Old Man
Labarre
Valjean
Marguerite
Mme. de R
Isabeau
Gervais
Tholomyes
Listolier
Fameuil
Blacheville
Favourite
Dahlia
Zephine
Fantine
Mme. Thenardier
Thenardier
Cosette
Javert
Fauchelevent
Bamatabois
Perpetue
Simplice
Scaufflaire
Woman 1
Judge
Champmathieu
Brevet
Chenildieu
Cochepaille
Pontmercy
Boulatruelle
Eponine
Anzelma
Woman 2
Mother Innocent
Gribier
Jondrette
Mme. Burgon
Gavroche
Gillenormand
Magnon
Mlle. Gillenormand
Mme. Pontmercy
Mlle. Vaubois
Lt. Gillenormand
Marius
Baroness T
Mabeuf
Enjolras
Combeferre
Prouvaire
Feuilly
Courfeyrac
Bahorel
Bossuet
Joly
Grantaire
Mother Plutarch
Gueulemer
Babet
Claquesous
Montparnasse
Toussaint
Child 1
Child 2
Brujon
Mme. Hucheloup
http://hci.stanford.edu/jheer/files/zoo/ex/networks/arc.html

Matrix View

Networks: **Figure 5c. Matrix view of *Les Misérables* character co-occurrences.**



http://hci.stanford.edu/jheer/files/zoo/ex/networks/matrix.html
Source: http://www-personal.umich.edu/~mejn/netdata

Ten Simple Rules for Better Visualizations

Ten Simple Rules for Better Figures

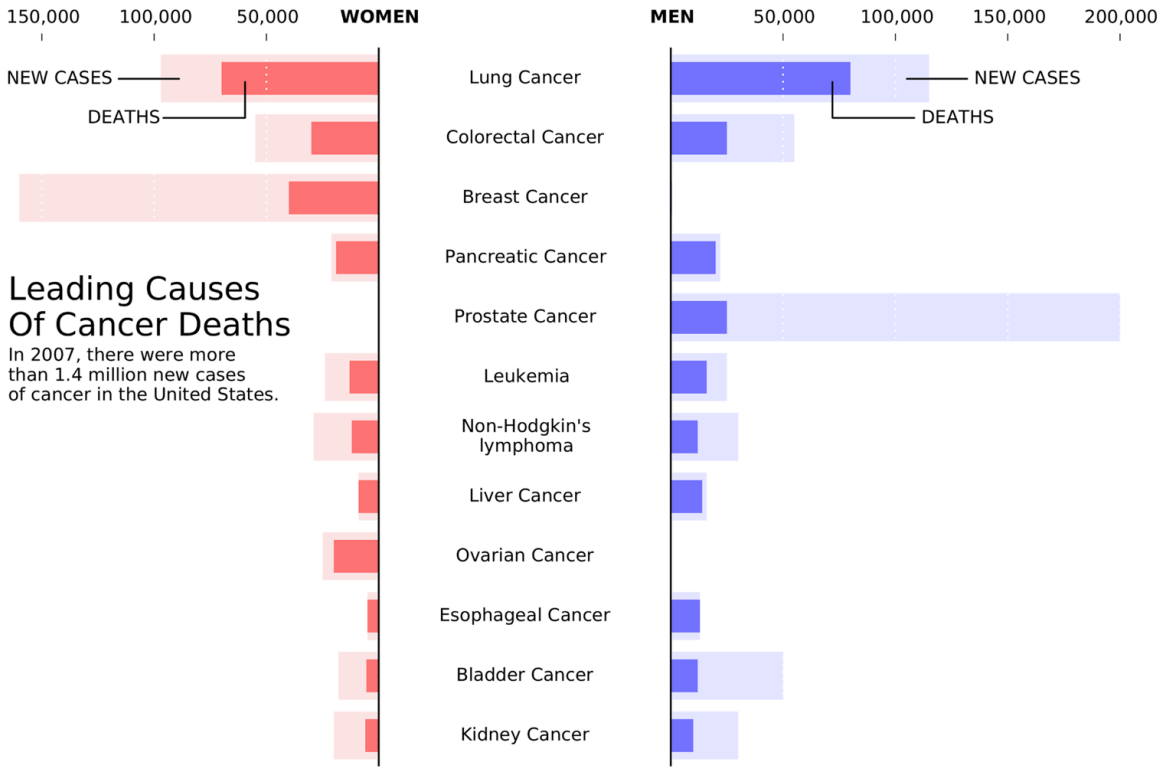
Rougier NP, Droettboom M, Bourne PE (2014) Ten Simple Rules for Better Figures. PLoS Comput Biol 10(9): e1003833. doi:10.1371/journal.pcbi.1003833

Link:
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833

Rule 1: Know Your Audience

- Problems arise when how a visual is perceived differs significantly from the intent of the conveyer
- If you are making a figure for yourself and your direct collaborators, you can possibly skip a number of steps in the design process, because each of you knows what the figure is about.
- However, if you intend to publish a figure in a scientific journal, you should make sure your figure is correct and conveys all the relevant information to a broader audience.

Rule 1: Know Your Audience

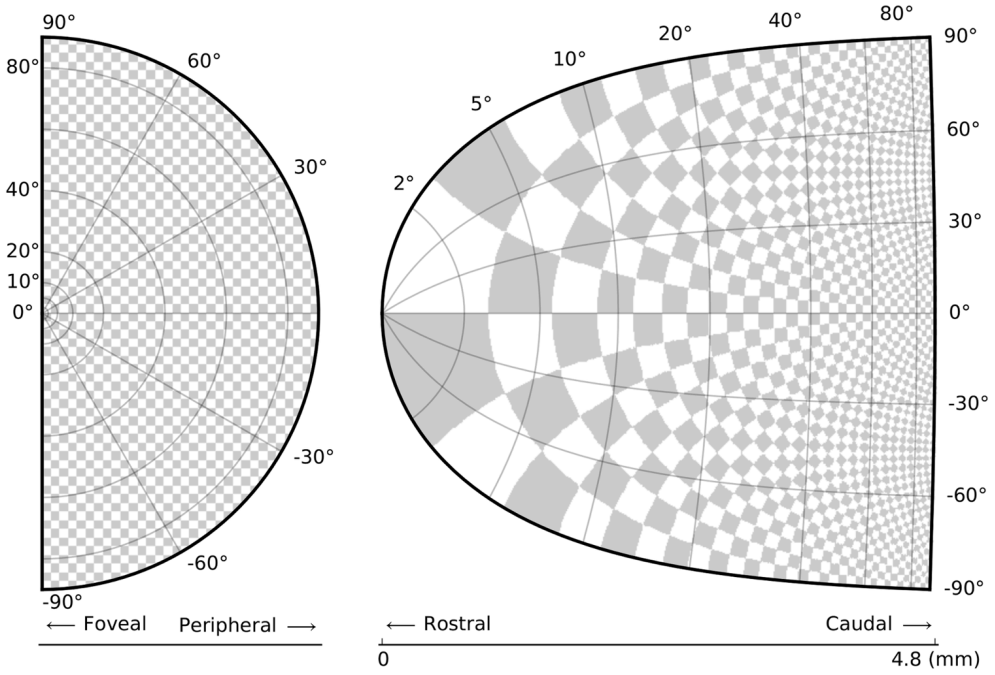


This is a remake of a figure that was originally published in the New York Times (NYT) in 2007. This is a self-contained figure that delivers a clear message on cancer deaths. However, it is not precise. The chosen layout makes it actually difficult to estimate the number of kidney cancer deaths because of its bottom position and the location of the labeled ticks at the top.

Rule 2: Identify Your Message

- A figure is meant to express an idea or introduce some facts or a result that would be too long (or nearly impossible) to explain only with words
- In this context, it is important to clearly identify the role of the figure, i.e., what is the underlying message and how can a figure best express this message?

Rule 2: Identify Your Message

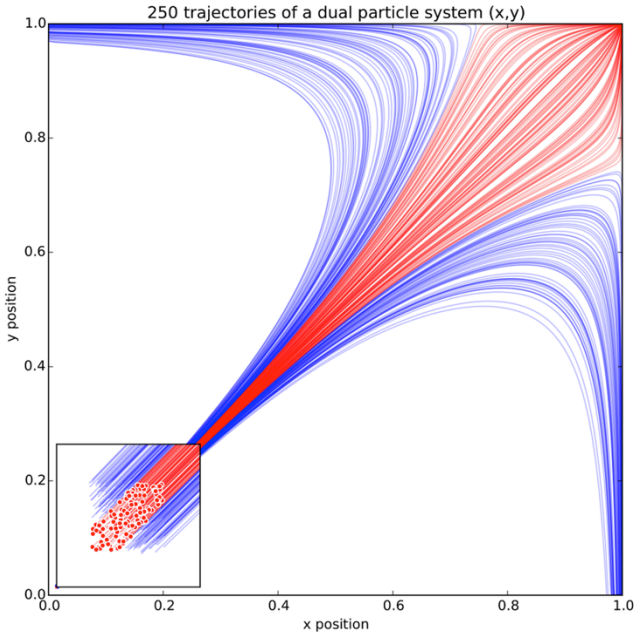


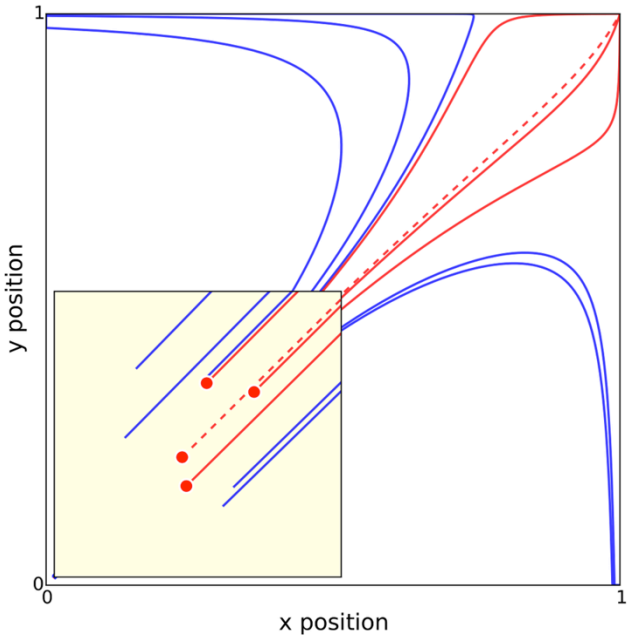
The superior colliculus (SC) is a brainstem structure at the crossroads of multiple functional pathways. Several neurophysiological studies suggest that the population of active neurons in the SC encodes the location of a visual target that induces saccadic eye movement. The projection from the retina surface (on the left) to the collicular surface (on the right) is based on a standard and quantitative model in which a logarithmic mapping function ensures the projection from retinal coordinates to collicular coordinates. This logarithmic mapping plays a major role in saccade decision. To better illustrate this role, an artificial checkerboard pattern has been used, even though such a pattern is not used during experiments. This checkerboard pattern clearly demonstrates the extreme magnification of the foveal region, which is the main message of the figure.

Rule 3: Adapt the Figure to the Support Medium

- A figure can be displayed on a variety of media, such as a poster, a computer monitor, a projection screen (as in an oral presentation), or a simple sheet of paper (as in a printed article).
- Each of these media represents different physical sizes for the figure, but more importantly, each of them also implies different ways of viewing and interacting with the figure.
- For example, during an oral presentation, a figure will be displayed for a limited time. Thus, the viewer must quickly understand what is displayed and what it represents while still listening to your explanation.

Rule 3: Adapt the Figure to the Support Medium





These two figures represent the same simulation of the trajectories of a dual-particle system, where each particle interacts with the other. Depending on the initial conditions, the system may end up in three different states. The left figure has been prepared for a journal article where the reader is free to look at every detail. The red color has been used consistently to indicate both initial conditions (red dots in the zoomed panel) and trajectories (red lines). Line transparency has been increased in order to highlight regions where trajectories overlap (high color density). The right figure has been prepared for an oral presentation. Many details have been removed (reduced number of trajectories, no overlapping trajectories, reduced number of ticks, bigger axis and tick labels, no title, thicker lines) because the time-limited display of this figure would not allow for the audience to scrutinize every detail.

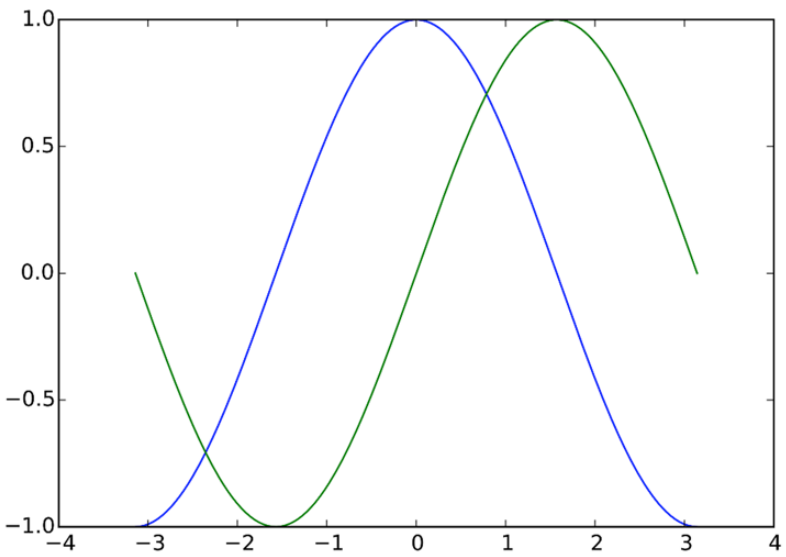
Rule 4: Captions Are Not Optional

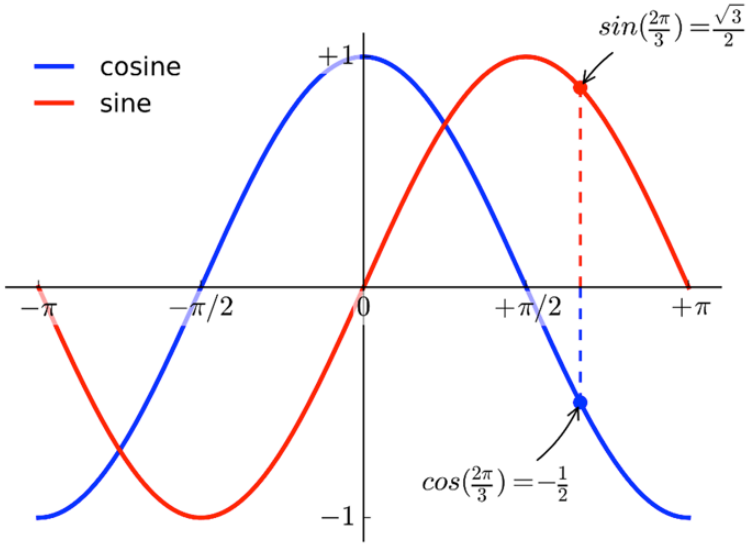
- Whether describing an experimental setup, introducing a new model, or presenting new results, you cannot explain everything within the figure itself—a figure should be accompanied by a caption.
- The caption explains how to read the figure and provides additional precision for what cannot be graphically represented.
- This can be thought of as the explanation you would give during an oral presentation, or in front of a poster, but with the difference that you must think in advance about the questions people would ask.
- Similarly, if there is a point of interest in the figure (critical domain, specific point, etc.), make sure it is visually distinct but do not hesitate to point it out again in the caption.

Rule 5: Do Not Trust the Defaults

- Any plotting library or software comes with a set of default settings. When the end-user does not specify anything, these default settings are used to specify size, font, colors, styles, ticks, markers, etc.
- Since these settings are to be used for virtually any type of plot, they are not fine-tuned for a specific type of plot.
- In other words, they are good enough for any plot but they are best for none.
- All plots require at least some manual tuning of the different settings to better express the message.

Rule 5: Do Not Trust the Defaults





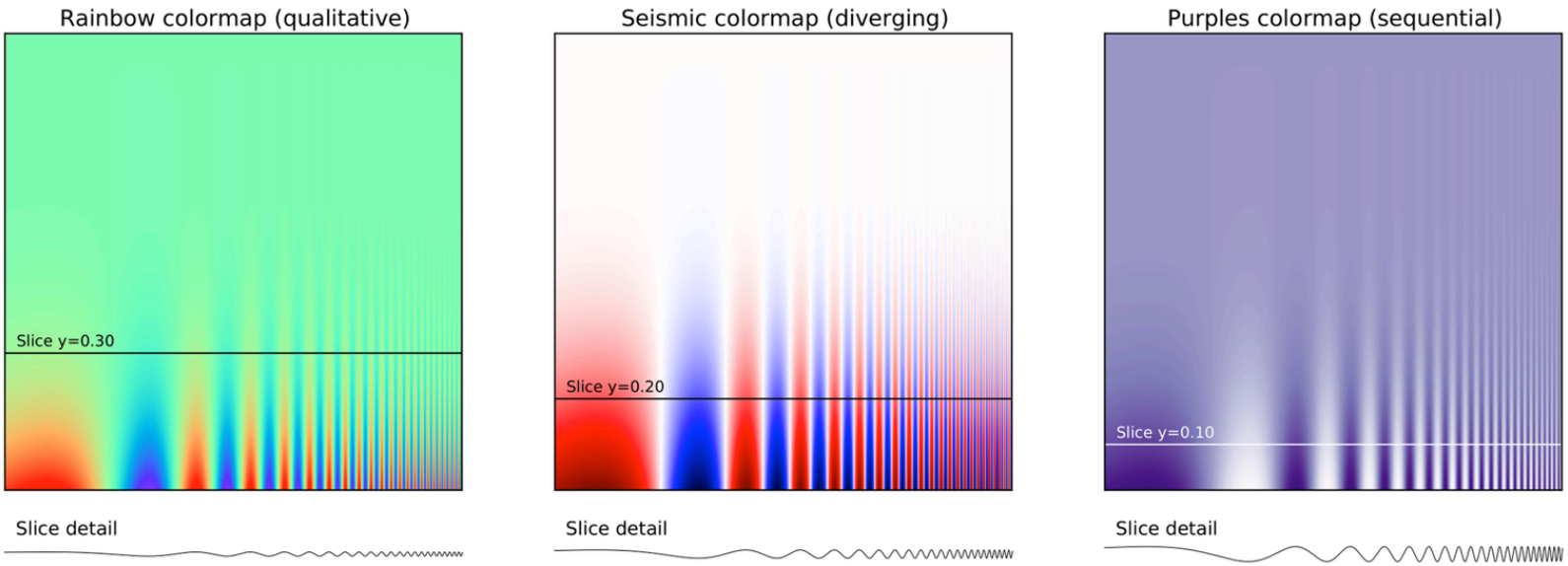
The left panel shows the sine and cosine functions as rendered by matplotlib using default settings. While this figure is clear enough, it can be visually improved by tweaking the various available settings, as shown on the right panel.

Rule 6: Use Color Effectively

- Color is an important dimension in human vision and is consequently equally important in the design of a scientific figure.
- However, as explained by Edward Tufte [1], color can be either your greatest ally or your worst enemy if not used properly.
- If you decide to use color, you should consider which colors to use and where to use them.
- Avoid using too many similar colors since color blindness may make it difficult to discern some color differences

[1] Tufte EG (1983) The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphics Press.

Rule 6: Use Color Effectively

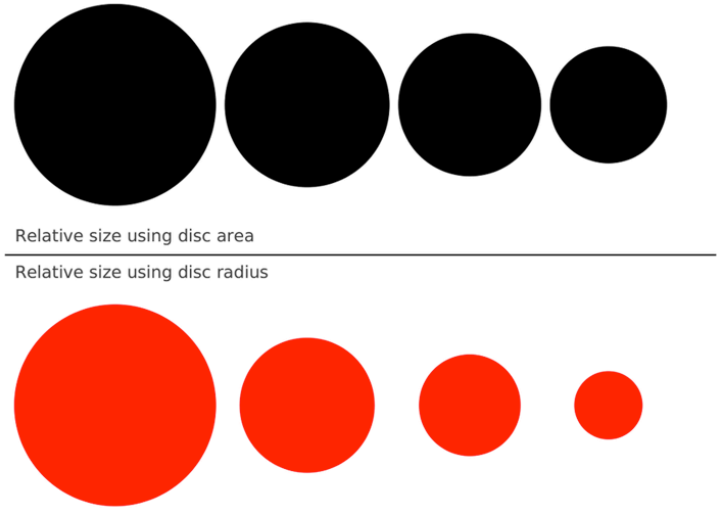


This figure represents the same signal, whose frequency increases to the right and intensity increases towards the bottom, using three different colormaps. The rainbow colormap (qualitative) and the seismic colormap (diverging) are equally bad for such a signal because they tend to hide details in the high frequency domain (bottom-right part). Using a sequential colormap such as the purple one, it is easier to see details in the high frequency domain.

Rule 7: Do Not Mislead the Reader

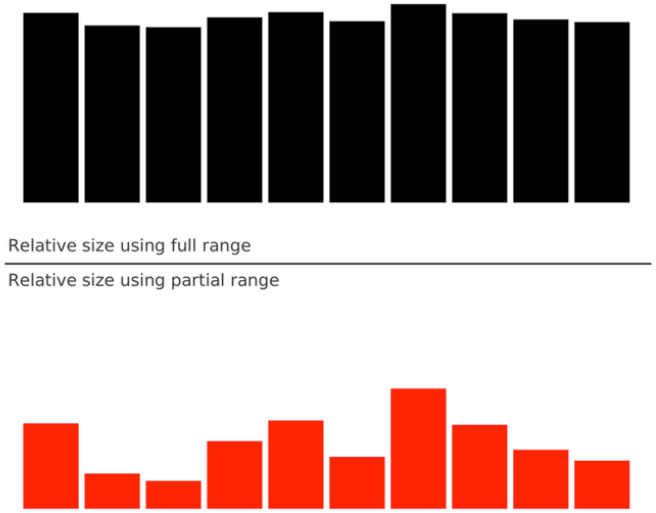
- What distinguishes a scientific figure from other graphical artwork is the presence of data that needs to be shown as objectively as possible.
- However, representing results objectively is not always straightforward (see next slide).
- As a rule of thumb, make sure to always use the simplest type of plots that can convey your message and make sure to use labels, ticks, title, and the full range of values when relevant.
- Lastly, do not hesitate to ask colleagues about their interpretation of your figures.

Rule 7: Do Not Mislead the Reader



- We represented a series of four values: 30, 20, 15, 10
- On the upper left part, we used the disc area to represent the value, while in the bottom part we used the disc radius. Results are visually very different.
- In the latter case (red circles), the last value (10) appears very small compared to the first one (30), while the ratio between the two values is only 3 : 1.
- This situation is actually very frequent in the literature because the command (or interface) used to produce circles or scatter plots (with varying point sizes) offers to use the radius as default to specify the disc size
- It thus appears logical to use the value for the radius, but this is misleading.

Rule 7: Do Not Mislead the Reader



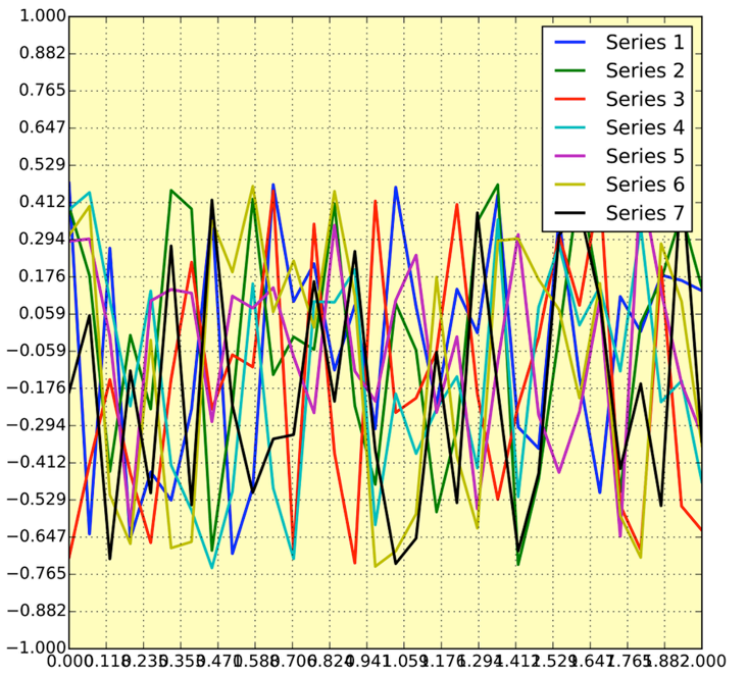
- We display a series of same ten values both at the top and bottom, and explicitly did not label the y-axis
- The top one uses the full range for values (y axis goes from 0 to 100), while the bottom one uses a partial range (y axis goes from 80 to 100)
- The visual perception of the two series is totally different. In the top part (black series), we tend to interpret the values as very similar, while in the bottom part, we tend to believe there are significant differences.
- Even if we had used labels to indicate the actual range, the effect would persist because the bars are the most salient information on these figures.

Rule 8: Avoid “Chartjunk”

- Chartjunk refers to all the unnecessary or confusing visual elements found in a figure that do not improve the message (in the best case) or add confusion (in the worst case).
- For example, chartjunk may include the use of too many colors, too many labels, gratuitously colored backgrounds, useless grid lines, etc.

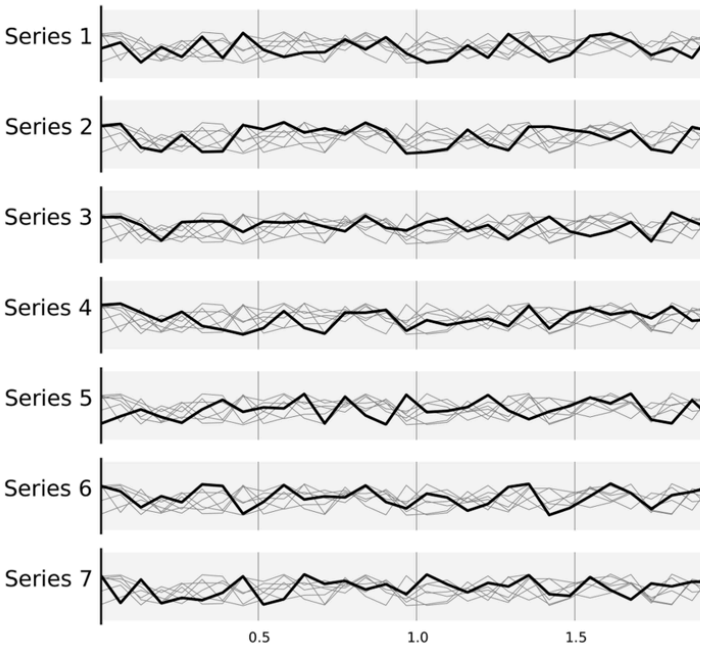
Rule 8: Avoid “Chartjunk”

- We have seven series of samples that are equally important, and we would like to show them all in order to visually compare them (exact signal values are supposed to be given elsewhere).
- The figure demonstrates what is certainly one of the worst possible designs. All the curves cover each other and the different colors (that have been badly and automatically chosen by the software) do not help to distinguish them.
- The legend box overlaps part of the graphic, making it impossible to check if there is any interesting information in this area.
- There are far too many ticks: x labels overlap each other, making them unreadable, and the three-digit precision does not seem to carry any significant information.
- Finally, the grid does not help because (among other criticisms) it is not aligned with the signal, which can be considered discrete given the small number of sample points.



Rule 8: Avoid “Chartjunk”

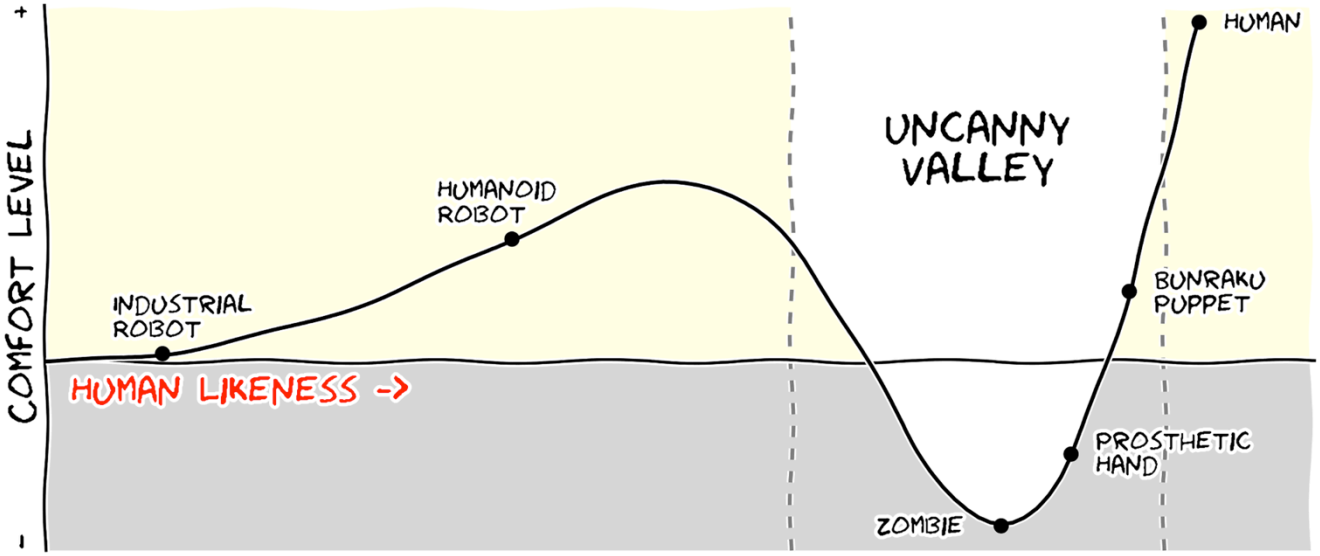
- This figure shows the exact same data but adopts a radically different layout while using the same area on the sheet of paper.
- Series have been split into seven plots, each of them showing one series, while other series are drawn very lightly behind the main one.
- Series labels have been put on the left of each plot, avoiding the use of colors and a legend box.
- The number of x ticks has been reduced to three, and a thin line indicates these three values for all plots.
- Finally, y ticks have been completely removed and the height of the gray background boxes indicate the [−1,+1] range (this should also be indicated in the figure caption if it were to be used in an article).



Rule 9: Message Trumps Beauty

- Often you may need to design a brand-new figure, because there is no standard way of describing your research.
- In such a case, browsing the scientific literature is a good starting point. If some article displays a stunning figure to introduce results similar to yours, you might want to try to adapt the figure for your own needs.
- You have to be very careful. There exists a myriad of online graphics in which aesthetic is the first criterion and content comes in second place. Even if a lot of those graphics might be considered beautiful, most of them do not fit the scientific framework

Rule 9: Message Trumps Beauty



This figure is an extreme case where the message is particularly clear even if the aesthetic of the figure is questionable. The uncanny valley is a well-known hypothesis in the field of robotics that correlates our comfort level with the human-likeness of a robot. To express this hypothetical nature, hypothetical data were used and the figure was given a sketched look (xkcd filter on matplotlib) associated with a cartoonish font that enhances the overall effect. Tick labels were also removed since only the overall shape of the curve matters. Using a sketch style conveys to the viewer that the data is approximate, and that it is the higher-level concepts rather than low-level details that are important.

Rule 10: Get the Right Tool

- There exist many tools that can make your life easier when creating figures
- Depending on the type of visual you're trying to create, there is generally a dedicated tool that will do what you're trying to achieve.
- It is important to understand at this point that the software or library you're using to make a visualization can be different from the software or library you're using to conduct your research and/or analyze your data. You can always export data in order to use it in another tool.

Rule 10: Get the Right Tool

- **Matplotlib** is a python plotting library, primarily for 2-D plotting, but with some 3-D support, which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. It comes with a huge gallery of examples that cover virtually all scientific domains (http://matplotlib.org/gallery.html).
- **R** is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible.
- **Inkscape** is a professional vector graphics editor. It allows you to design complex figures and can be used, for example, to improve a script-generated figure or to read a PDF file in order to extract figures and transform them any way you like.

Rule 10: Get the Right Tool

- **TikZ** and **PGF** are TeX packages for creating graphics programmatically. TikZ is built on top of PGF and allows you to create sophisticated graphics in a rather intuitive and easy manner, as shown by the Tikz gallery.
- **GIMP** is the GNU Image Manipulation Program. It is an application for such tasks as photo retouching, image composition, and image authoring. If you need to quickly retouch an image or add some legends or labels, GIMP is the perfect tool.
- **D3.js** (or just D3 for Data-Driven Documents) is a JavaScript library that offers an easy way to create and control interactive data-based graphical forms which run in web browsers, as shown in the gallery

Rule 10: Get the Right Tool

- **Cytoscape** is a software platform for visualizing complex networks and integrating these with any type of attribute data. If your data or results are very complex, cytoscape may help you alleviate this complexity.
- **Circos** was originally designed for visualizing genomic data but can create figures from data in any field. Circos is useful if you have data that describes relationships or multilayered annotations of one or more scales.
- **NetworkX** is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.
- **Seaborn** is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.