Visualization:
Principles & Software

Good Visualization?

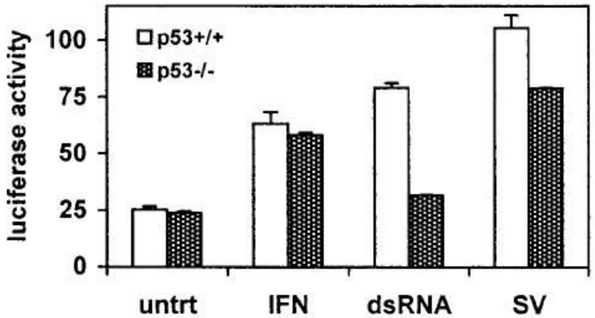


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were seeded in 32-mm plates and allowed to attach overnight. Cells were transfected with 500 ng of pGL3/ISG15-Luc, 50 ng of pRL null (Promega), and 450 ng of pcDNA3 for carrier DNA by using Lipofectamine Plus (Life Technologies) following the manufacturer's instructions. Twenty-four hours posttransfection, the medium was aspirated and replaced with medium containing either 1,000 U of IFN-α/ml, 50 μg of dsRNA/ml, or Sendai virus (multiplicity of infection, 10). Cells were incubated for 12 h and then lysed, and luciferase assays were performed. Luciferase activity was assessed on 20 μl of each lysate as directed by the supplier (Dual Luciferase Kit, Promega) using a TD 20/20 luminometer (Turner Designs). Luciferase activity is presented as the ratio of firefly activity to renilla activity to control for differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are representative of four independent experiments.

Hummer BT, Li XL, Hassel BA (2001) Role for p53 in gene induction by double-stranded RNA. *J Virol* 75:7774-7777, Figure 4

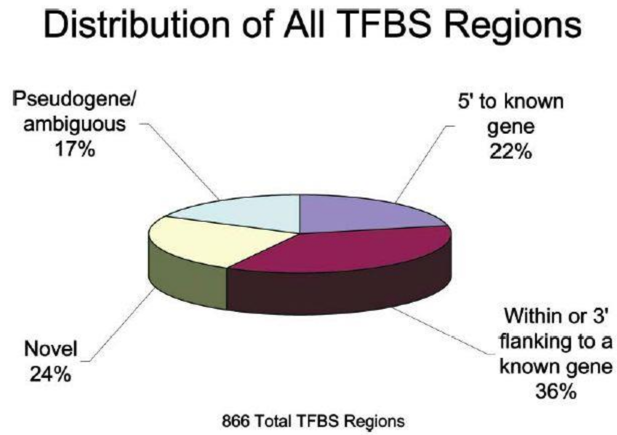


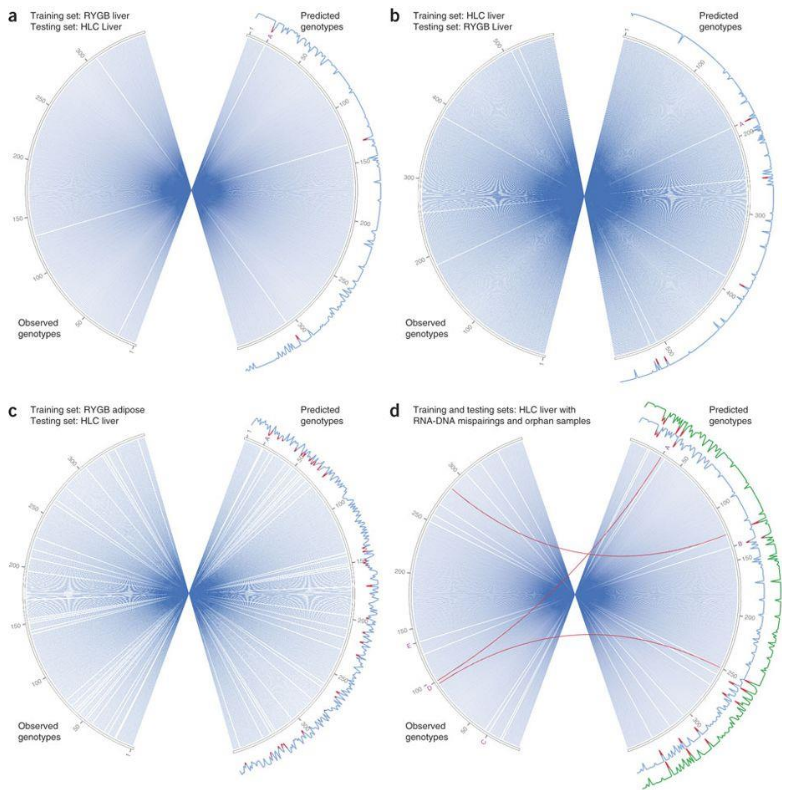
Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

Cawley S, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116:499-509, Figure 1

https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

Good Visualization?

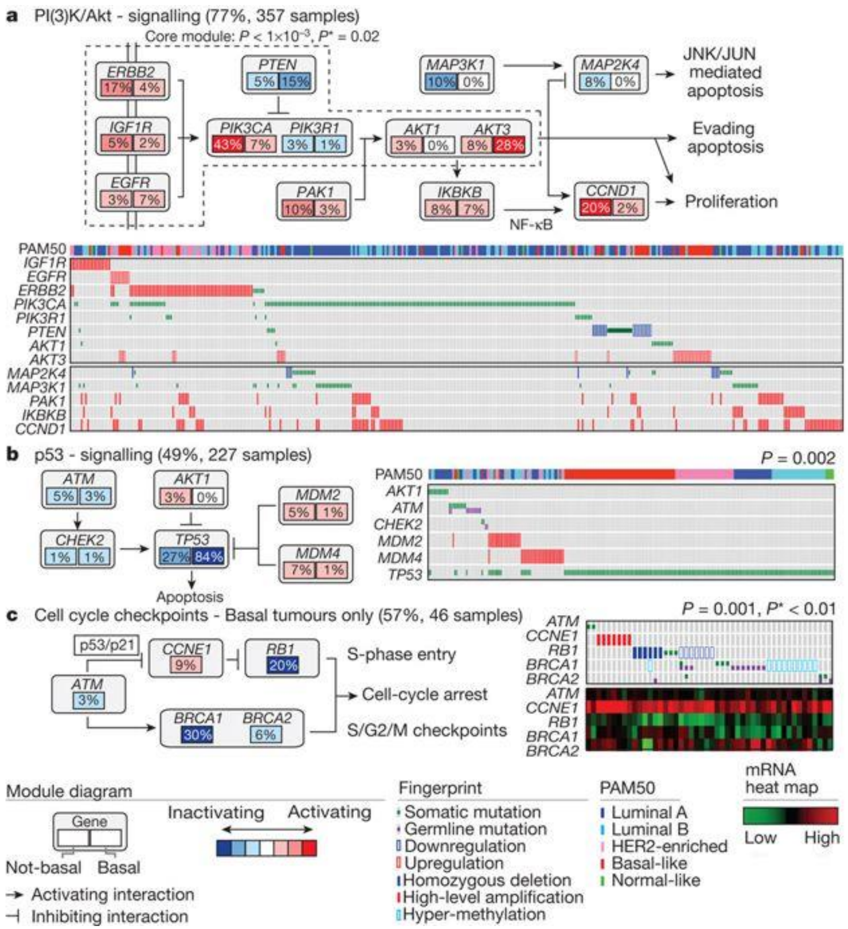


(**a**–**c**) Sample IDs were sorted for each semicircle (right, predicted genotypes; left, observed genotypes; numbers on the outside of the semicircles represent indexed sample numbers). Results are shown for experiments in which RYGB liver was used as the training set for HLC liver (**a**), HLC liver was used as the training set for RYGB liver (**b**) and RYGB adipose was used as the training set for HLC liver (**c**). In the case of a correct pairing (with adjusted minimum $P_{i,j}$ of $<1 \times 10^{-5}$), the connection between the semicircles was a straight line passing the circle center (blue lines). In the case that no match for a given individual was identified, no line existed: for example, tick A in **a**–**c**. The blue curves outside of the right semicircles denote adjusted minimum $P_{i,j}$ ($-\log_{10}$ transformed) for matching predicted genotype vectors to observed genotype vectors. For convenience, this value was capped at 16. If the value was <5, the curve is shown in red, indicating lack of statistical support for any match. (**d**) Matching was performed in the HLC liver set to which RNA-DNA mispairing and orphan samples had been added. In the case of a mispairing detected at adjusted minimum $P_{i,j}$ of $<1 \times 10^{-5}$, the line connecting the semicircles will not be straight (red connections). The predicted genotype of subject 31 (tick A) best matches the observed genotype of subject 98 (tick D). There was no line connecting the observed genotype of subject 31 (tick C). In the case of orphan RNA (for example, subject 137), there was no connection between the predicted genotype (tick B) and observed genotype (tick E). The green curve outside the right semicircle show adjusted $-\log_{10}(P_{i,i})$.

Bayesian method to predict individual SNP genotypes from gene expression data
Schadt, E.E., et al. Nature (2012)

http://jasonya.com/wp/science-figures-we-could-do-without/

Good Visualization?

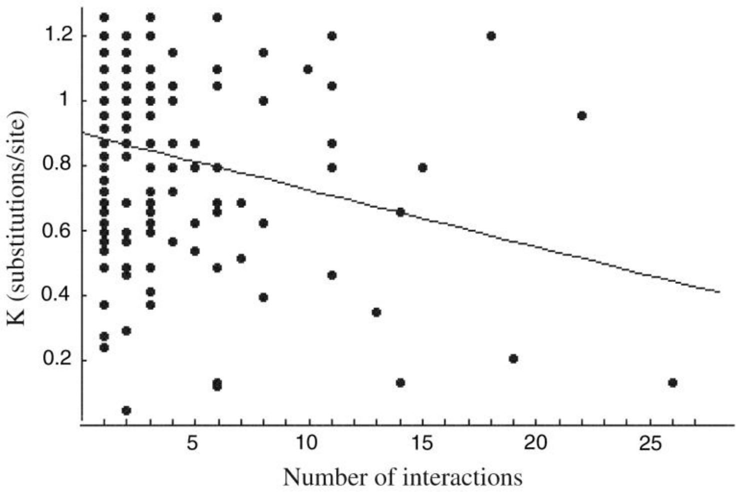


Mutual exclusivity modules are represented by their gene components and connected to reflect their activity in distinct pathways. For each gene, the frequency of alteration in basal-like (right box) and non-basal (left box) is reported. Next to each module is a fingerprint indicating what specific alteration is observed for each gene (row) in each sample (column). **a**, MEMo identified several overlapping modules that recapitulate the RTK–PI(3)K and p38–JNK1 signalling pathways and whose core was the top-scoring module. **b**, MEMo identified alterations to TP53 signalling as occurring within a statistically significant mutually exclusive trend. **c**, A basal-like only MEMo analysis identified one module that included ATM mutations, defects at BRCA1 and BRCA2, and deregulation of the RB1 pathway. A gene expression heat map is below the fingerprint to show expression levels.

TCGA
Nature (2012)

http://jasonya.com/wp/science-figures-we-could-do-without/

Good Visualization?

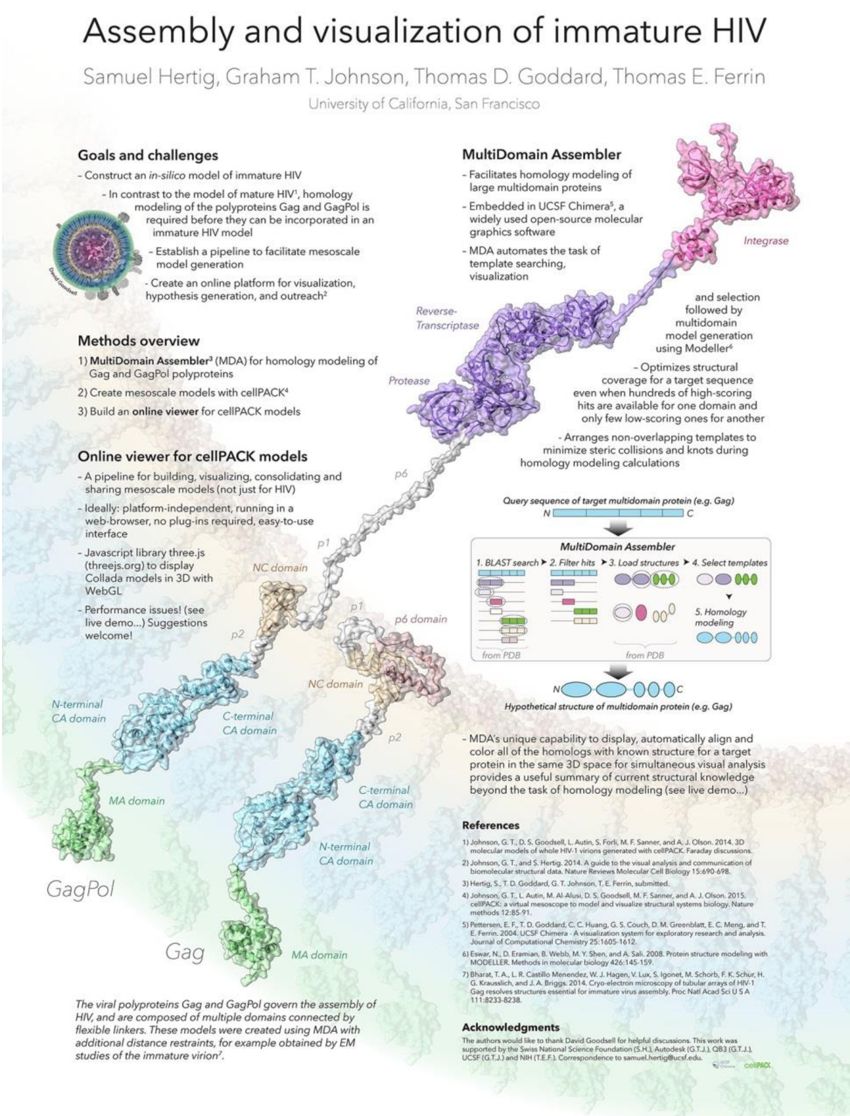


The relation between the number of protein-protein interactions (*I*) in which a yeast protein participates and that protein's evolutionary rate, as estimated by the evolutionary distance (*K*) to the protein's well-conserved ortholog in the nematode *C. elegans*.

Evolutionary Rate in the Protein Interaction Network
Fraser, H.B., et al. Science (2002)

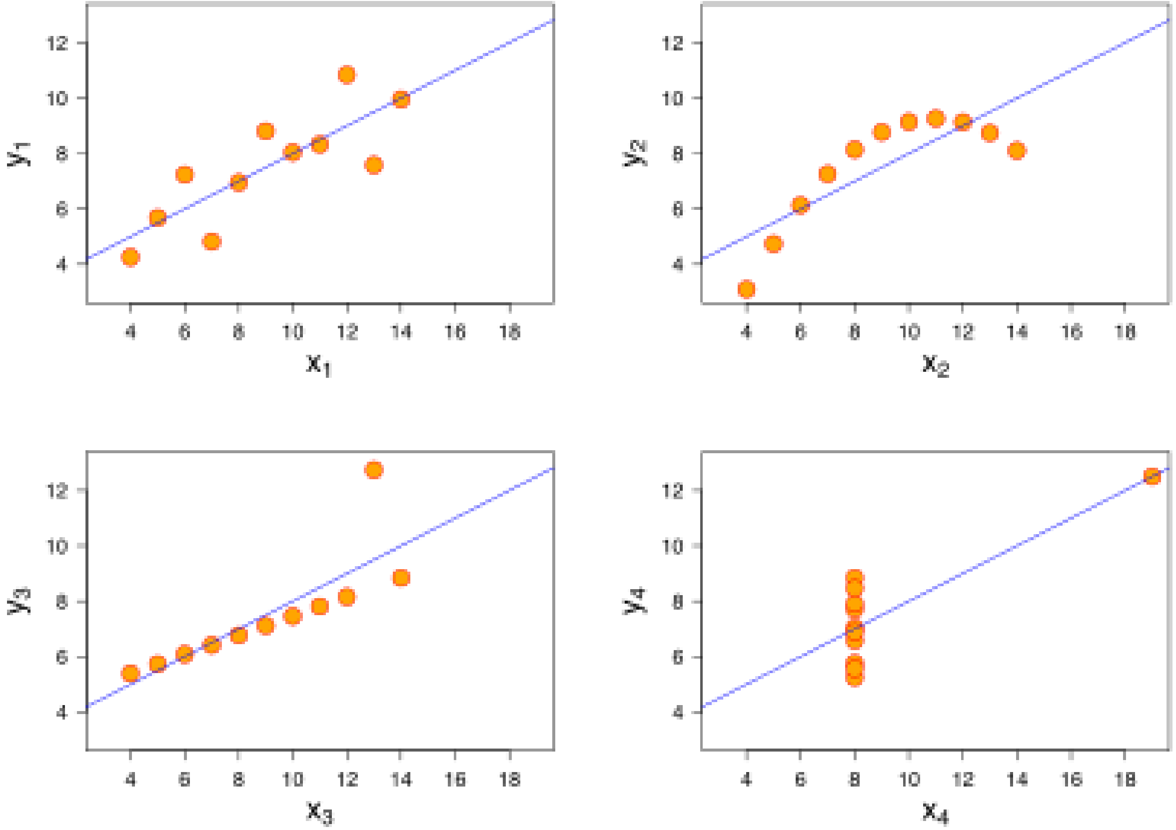
http://jasonya.com/wp/science-figures-we-could-do-without/

Visualizing Biological Data (VizBi)



vizbi.org
Best Poster (2015)

Ascombe's Quartet



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

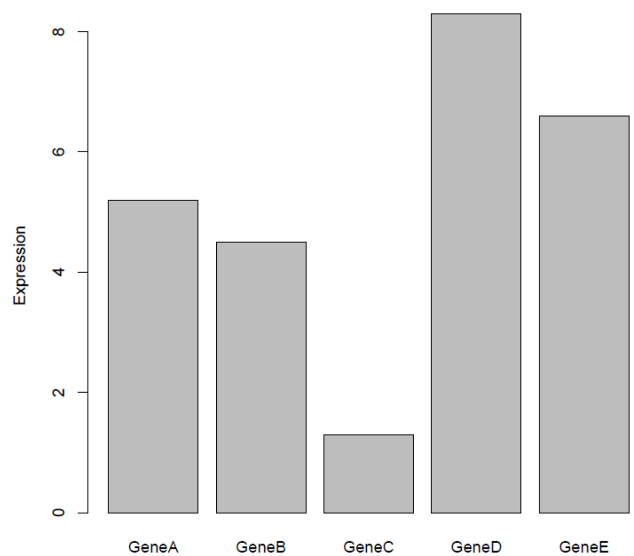
Visualization

- Common Misconceptions
  - ➢ Goal is to impress (wow!)
  - ➢ Visualization == Imaging
  - ➢ Easy
- Goals
  - ➢ Record: raw data
  - ➢ Analyze: reveal patterns or trends
  - ➢ Communicate

O'Donoghue, S.I, et al. *Visualization of Biomedical Data*. Annual Rev of Biomed Data Sci 1:275-304 (2018)

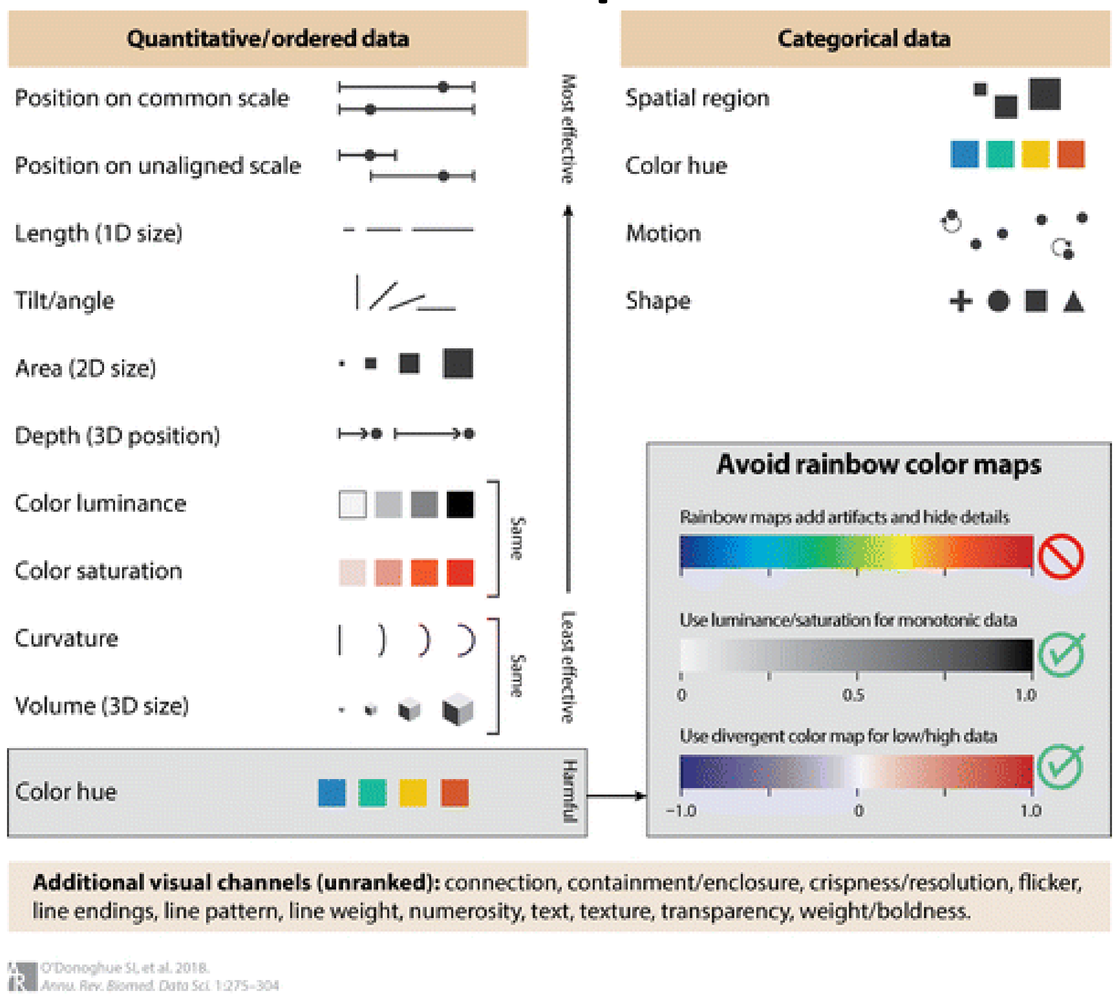
Visualization: Principles

- How do you encode information/data?
  - Marks: basic geometric elements
    e.g. circle, square
  - Channels: control the appearance of the marks
    e.g. color, size, orientation/direction, etc.



Marks: lines
Channels: length (of the lines)

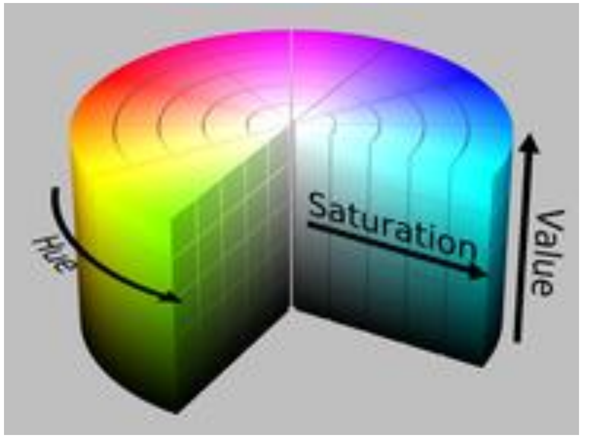
Visualization: Principles



O'Donoghue, S.I, et al. *Visualization of Biomedical Data*. Annual Rev of Biomed Data Sci 1:275-304 (2018)

Visualization: Color

- Hue
- Saturation
- Luminescence or Brightness (Value)



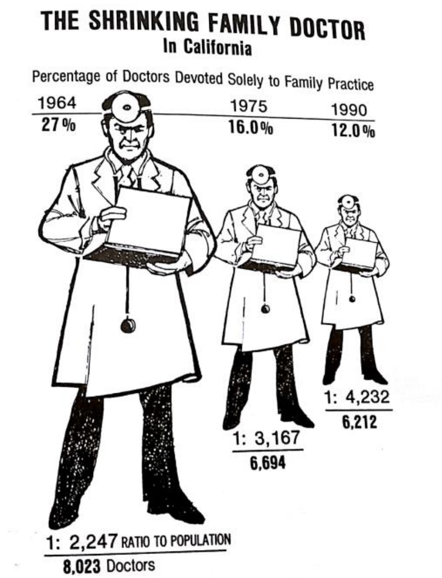
https://en.wikipedia.org/wiki/HSL_and_HSV

Visualization: Interaction

- Overview first, Zoom/Filter for details (e.g. Google Maps, IGV, 'hairball' network diagram, 3D protein viewer)
- Alternative: Details first, overview last
- Animation: Use to show change, especially over time. Often used ineffectively!

Visualization: Tufte's Principles

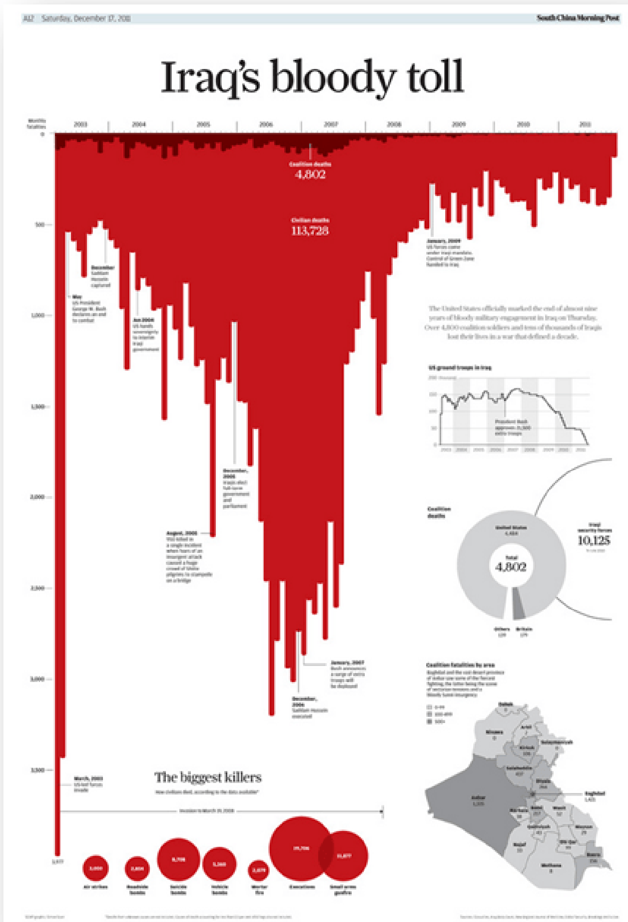
- Graphical integrity: maintain credibility
- Maximize data-ink ratio: avoid "chart junk"

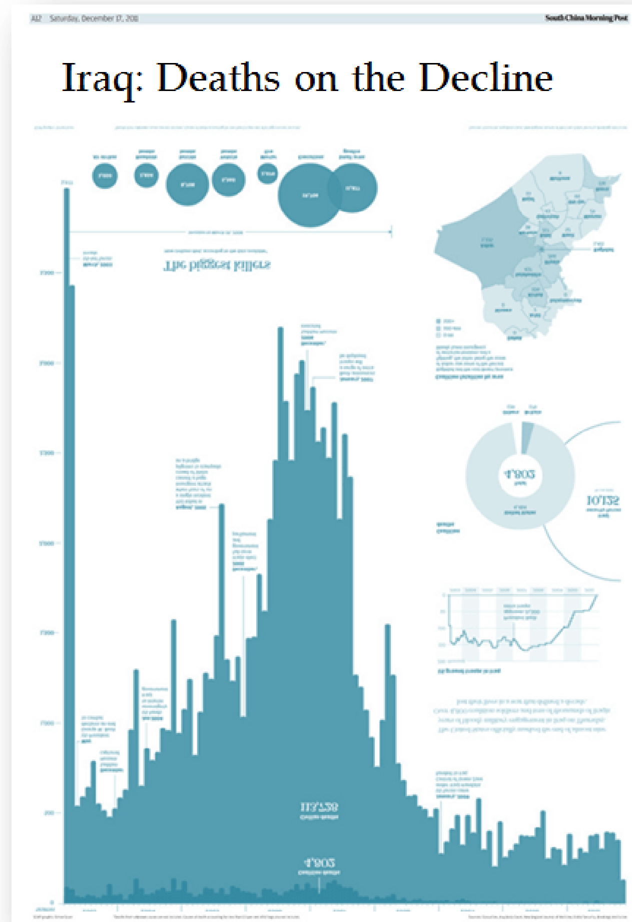


L.A. Times (Aug 5, 1979)

Tufte, E., *The Visual Display of Quantitative Information, 2nd Ed. (2001)*

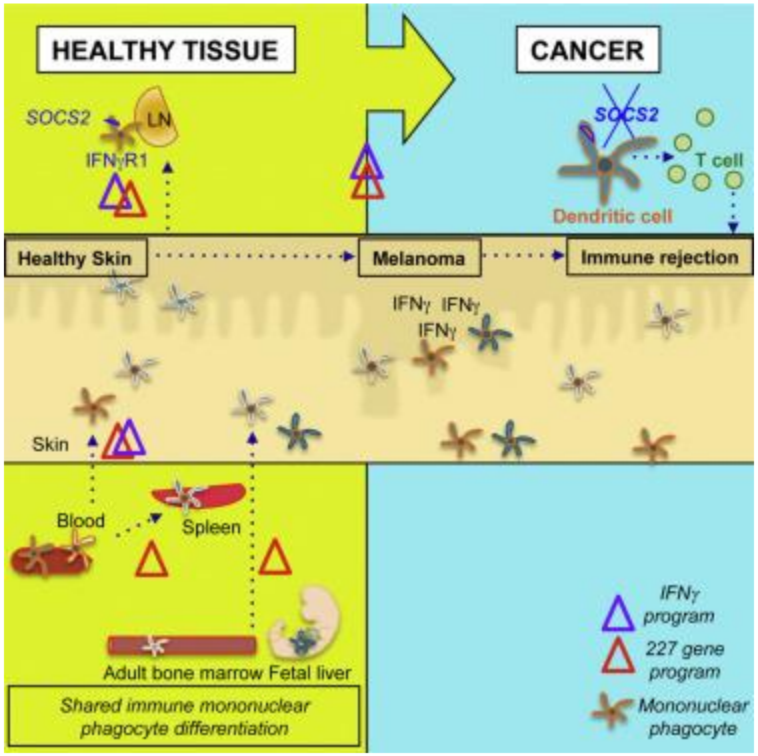
Visualization: Communication



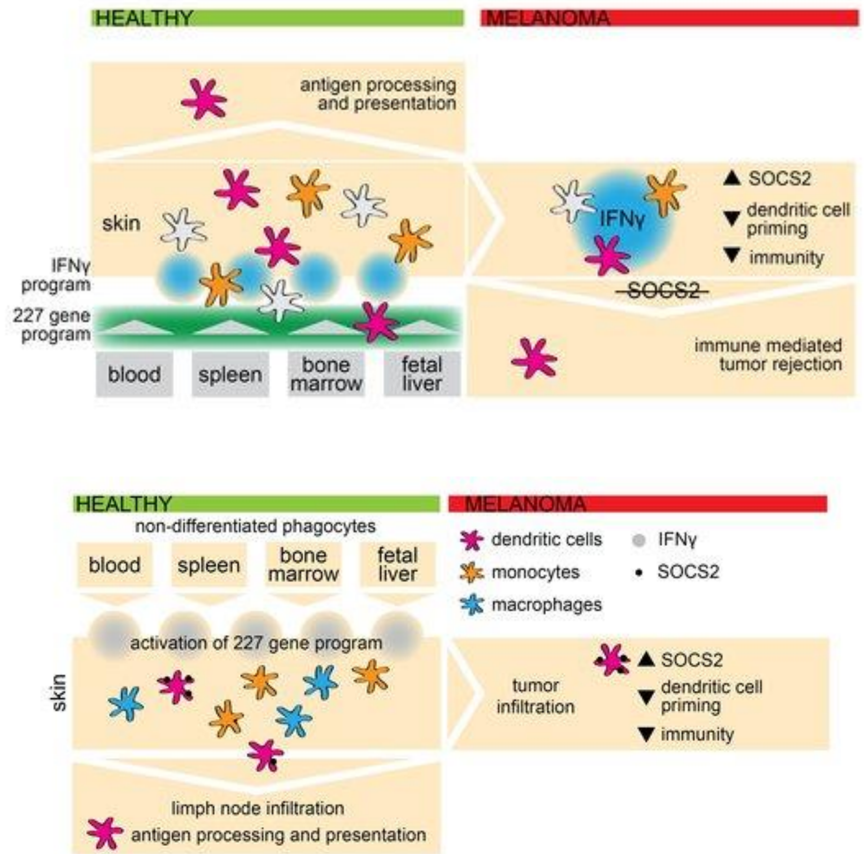


Andy Cotgreave blog, Tableau

Visualization: Graphical Abstracts



Nirschl, C.J. *et al.* (2017)
https://www.cell.com/cell/fulltext/S0092-8674(17)30699-2



https://www.gabrielaplucinska.com/blog/2017/9/7/graphicalabstract

Table 1. Comparison between different types of graphical representations

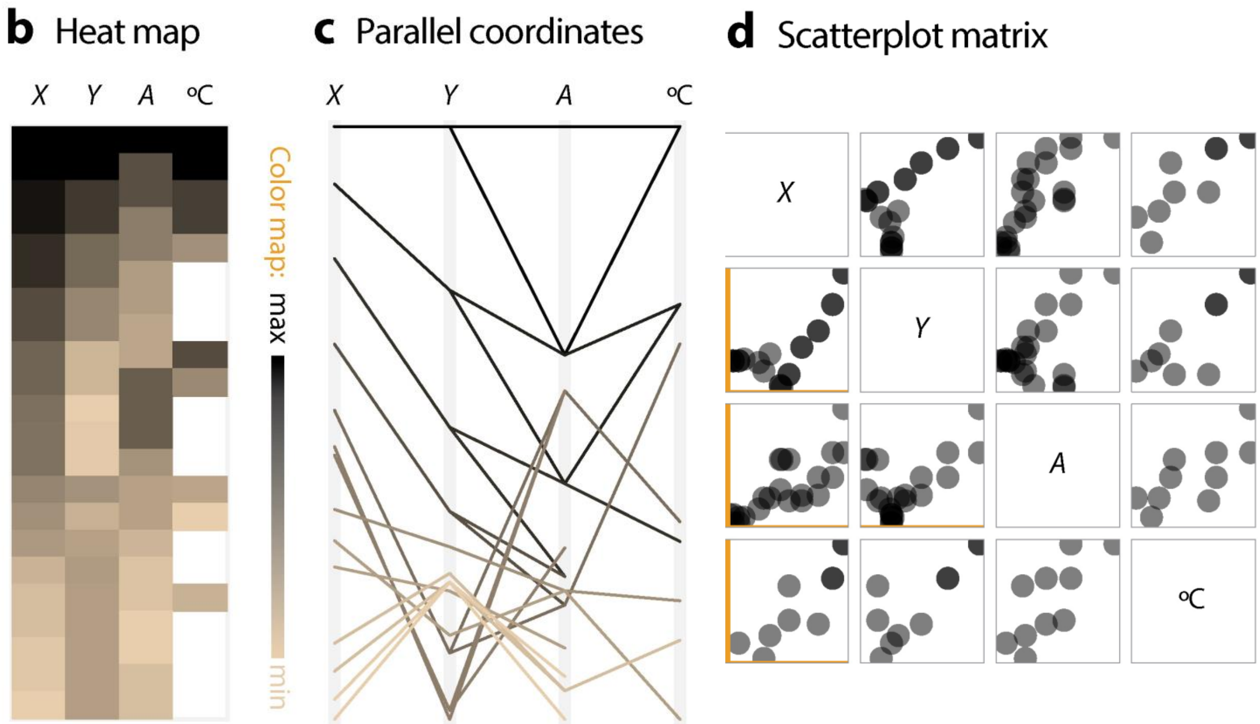
Type of visual display	Utility and pros	Cons
Graphical representation to illustrate data on overall survival or progression-free survival		
Kaplan-Meier curves	Allows estimation of survival and comparison of two treatment groups based on selected categories	Univariate analysis, which may be confounded by censoring differences between groups
Graphical representations of treatment effect		
Forest plots	Helps determine behaviors of different subgroups within a larger dataset	Subject to error if there are only small number of data points within subgroup analysis resulting in false interpretation
Funnel plots	Scatter plots of the effect estimates that can give an indication of heterogeneity	Shape of the plot is dependent on number of patients recruited in different risk groups
Violin plots	Indication of clusters within the data that highlight the variation in distribution	Does not allow easy comparison across different datasets
Graphical representations of tumor response		
Waterfall plots	Summarizes the typical response size and the fraction of patients experiencing benefit. Reveals interpatient heterogeneity of response	Only shows one measurement in time, and tumor response size may not represent actual patient benefit in terms of overall survival or progression-free survival
Spider plots	Allows visualization of data points across time rather than at a specified time point	Does not allow for formal statistical inference, difficult to interpret if large number of data points
Swimmer plots	Tumor response and timeframe of response displayed	May become cluttered and uninformative if too many subjects are included or too many variables are included
Graphical representations to illustrate cancer genotypes and phenotypes		
Heat maps	Allows complex data to be grouped according to thousands of individual data points, thereby allowing patterns within the data to be visualized	Clustering is based on multiple data points, which may dilute the effects of individual data points such that it is lost within the volume of data
Circos plots	Allows visualizing complex genome data in one plot, allows visualization of the interaction between genomic regions in addition to genome gains/losses	Highly complex plots without ability to focus on specific genomic regions
Graphical representations to illustrate connectedness and relatedness in cancer		
Subway diagrams	Visual simplification of successive steps in a complex pathway	Does not quantify impact or efficacy of each step in the pathway
Network analysis graphs	The vertex represents each factor that is being studied, and size of the vertex is proportional to the efficacy of the factor	Unable to quantify degree of effect other than via thickness of the links drawn in the diagram

Chia, P.L, et al. *Current and Evolving Methods to Visualize Biological Data in Cancer Research* J Nat Cancer Inst 108:8 (2016)

Relationships

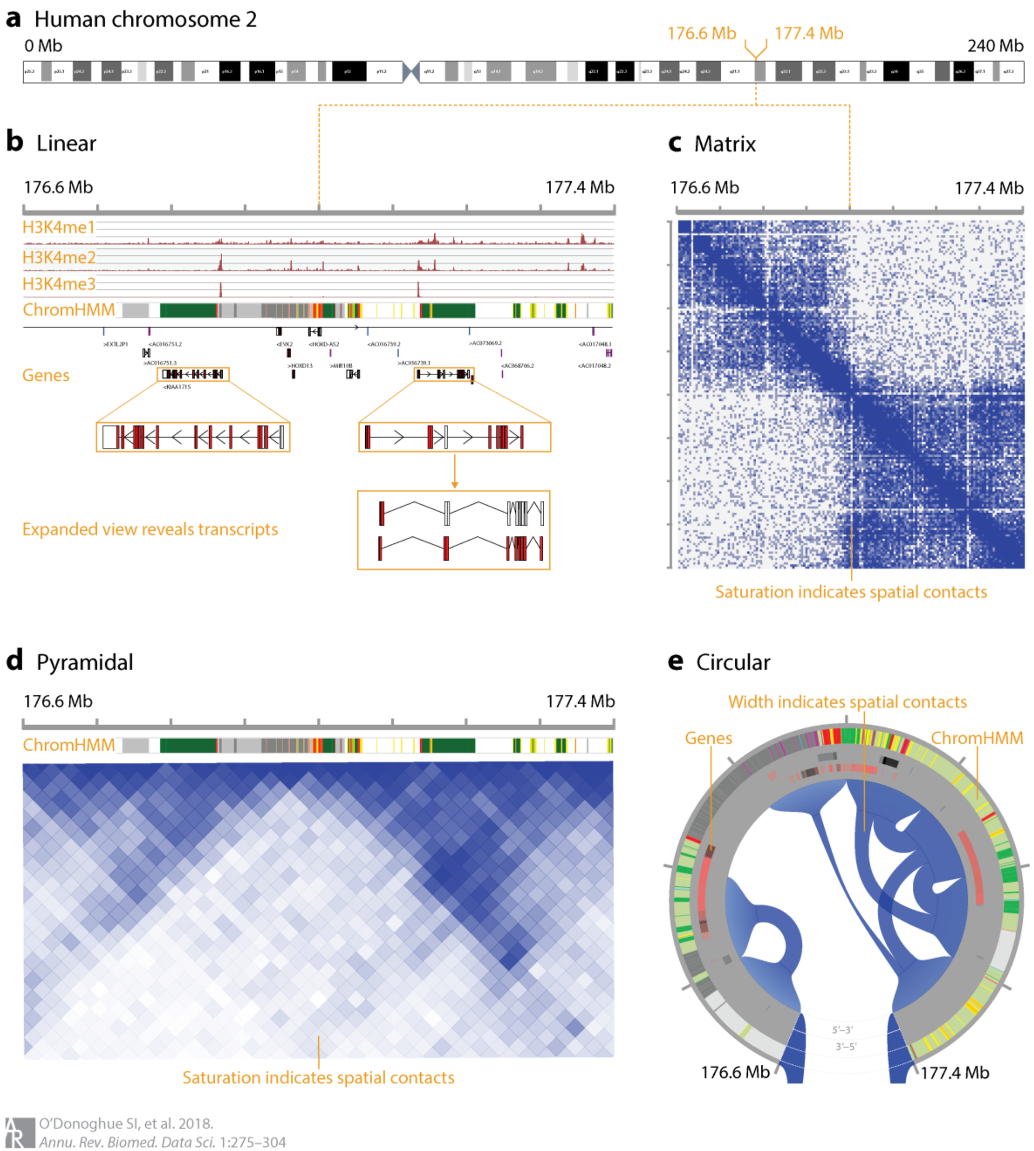
a Data matrix

X	*Y*	*A*	°C
0.768	0.30	87000	0
0.768	0.30	55000	0
0.700	0.26	55000	−11
0.700	0.26	37000	−11
0.612	0.23	37000	−26
0.612	0.23	24000	
0.511	0.21	24000	
0.511	0.21	20000	
0.433	0.18	20000	−14
0.433	0.18	50000	−25
0.390	0.16	50000	
0.380	0.16	50000	
0.380	0.16	28000	
0.316	0.20	22000	−30
0.279	0.18	22000	−38
0.248	0.19	14000	
0.158	0.20	8000	
0.125	0.19	8000	−33
0.125	0.19	4000	
0.092	0.19	4000	
0.092	0.19	10000	
0.068	0.19	10000	



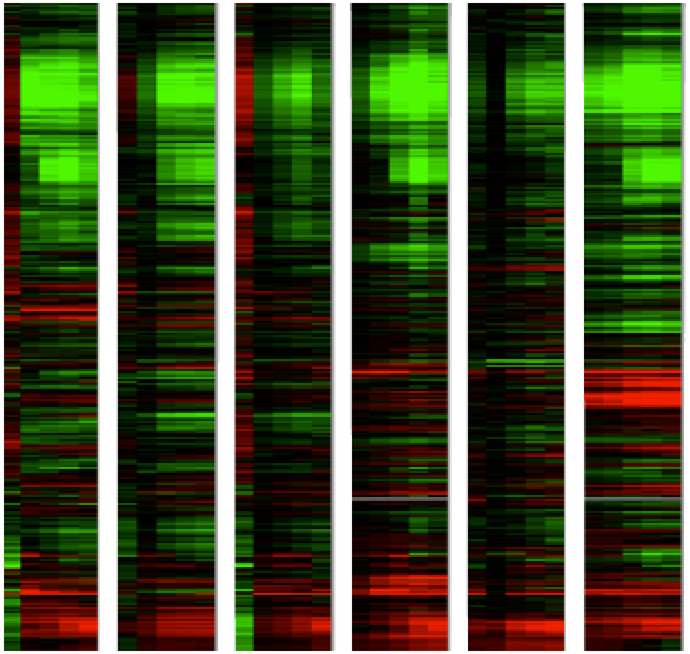
O'Donoghue, S.I, et al. *Visualization of Biomedical Data*. Annual Rev of Biomed Data Sci 1:275-304 (2018)

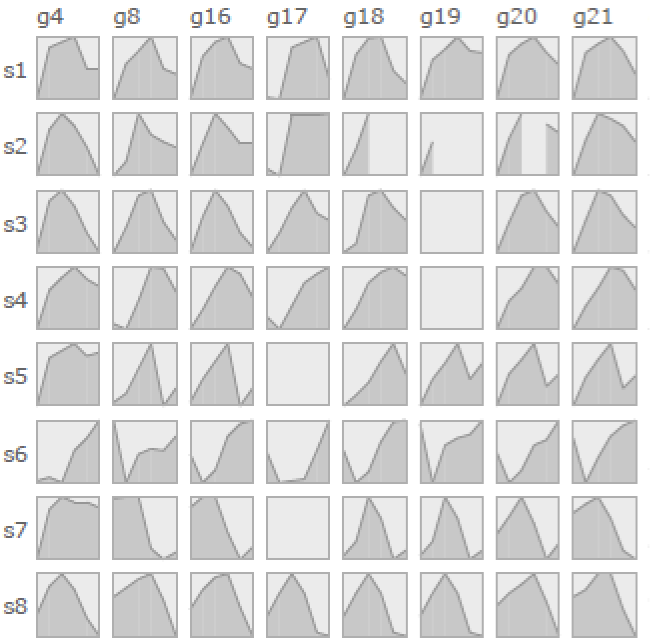
Genomic Features and Interactions

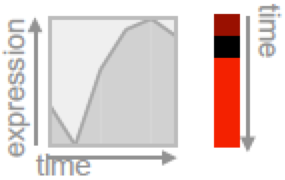


O'Donoghue, S.I, et al. *Visualization of Biomedical Data*. Annual Rev of Biomed Data Sci 1:275-304 (2018)

Heatmap vs Curvemap







Muzner, T. Vizbi 2011 Workshop of Visualizing Biological Data

Heatmap: Color Perception

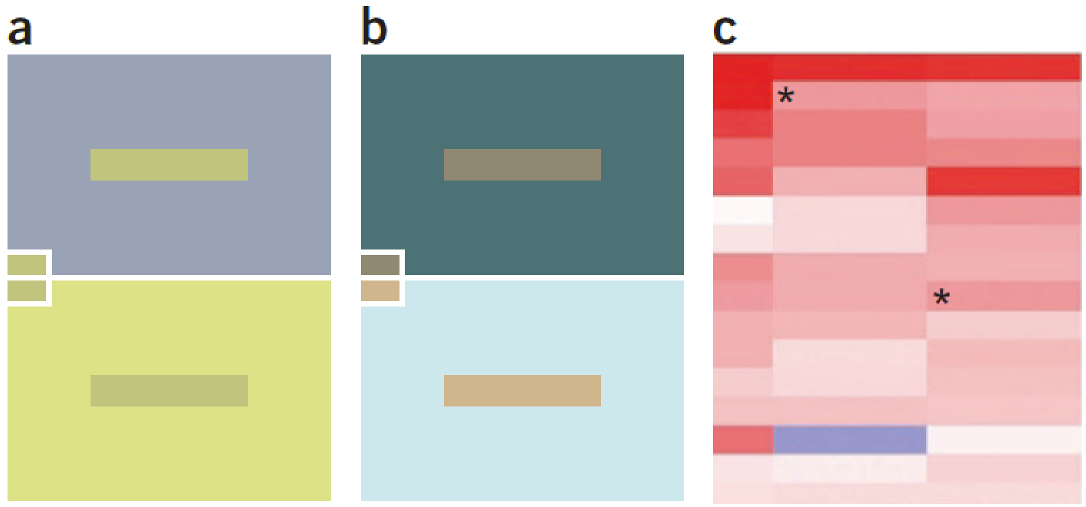
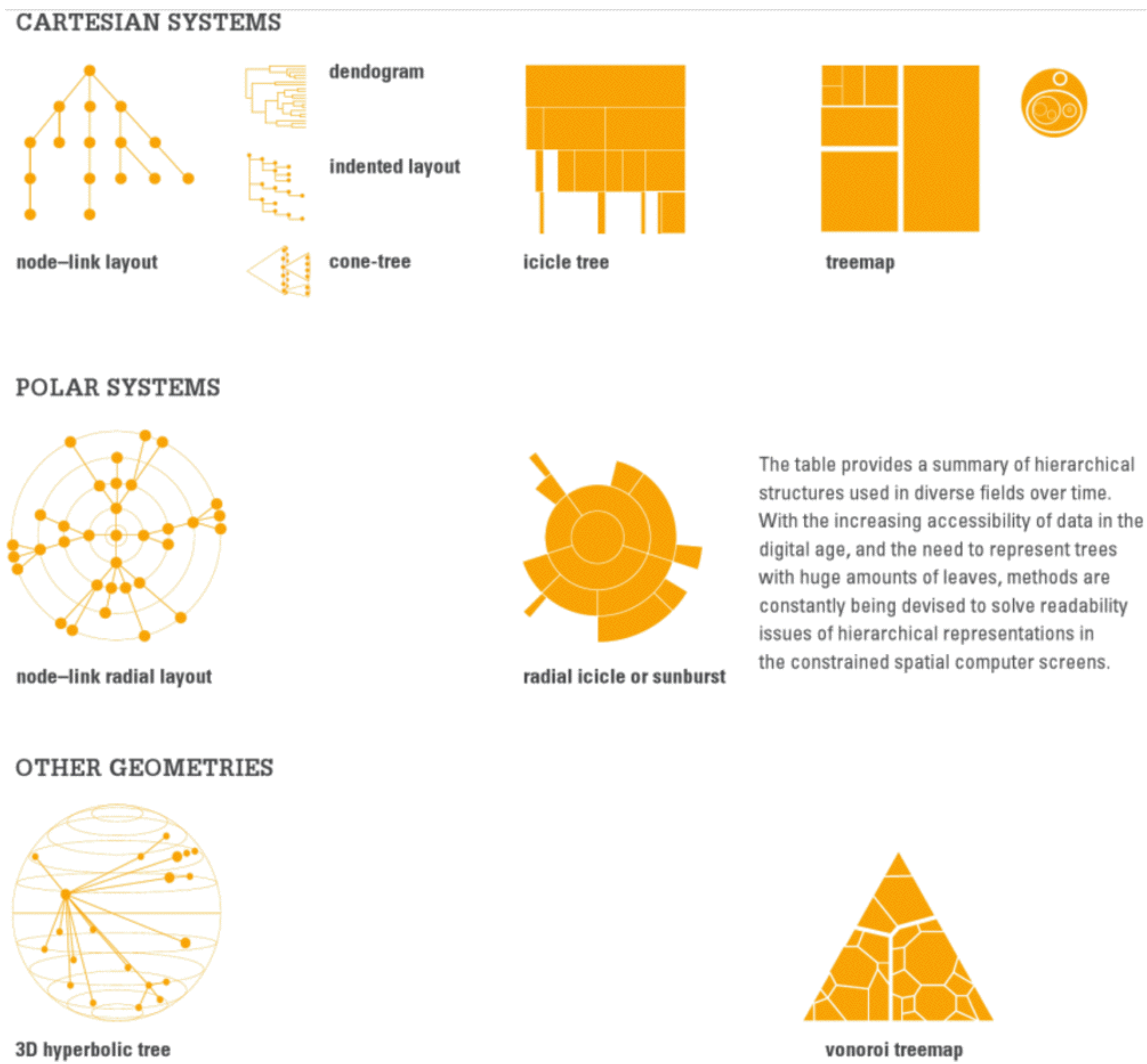


Figure 1 | Perception of color can vary. (**a**,**b**) The same color can look different (**a**), and different colors can appear to be nearly the same by changing the background color (**b**)[1]. (**c**) The rectangles in the heat map indicated by the asterisks (*) are the same color but appear to be different.

Wong, B. *Color Coding* Nature Methods 7:8 (2010)

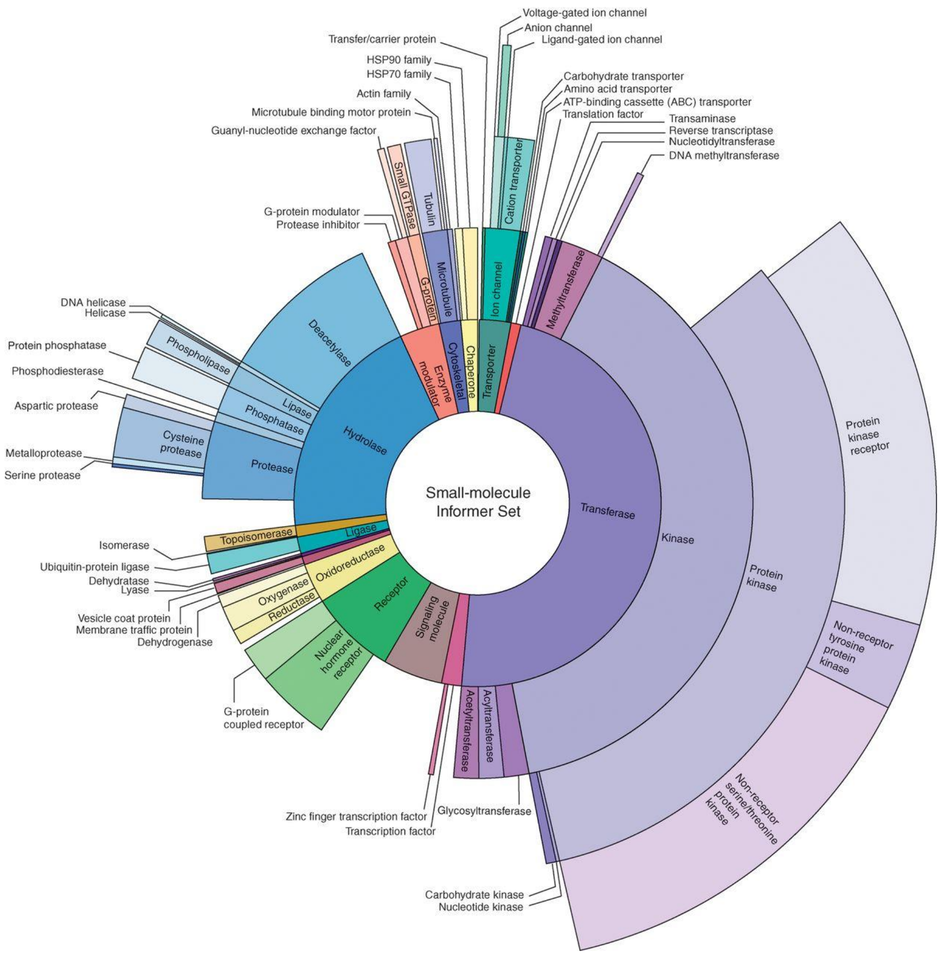
Hierarchies



http://phylonetworks.blogspot.com/2014/03/trees-treemaps-and-networks.html

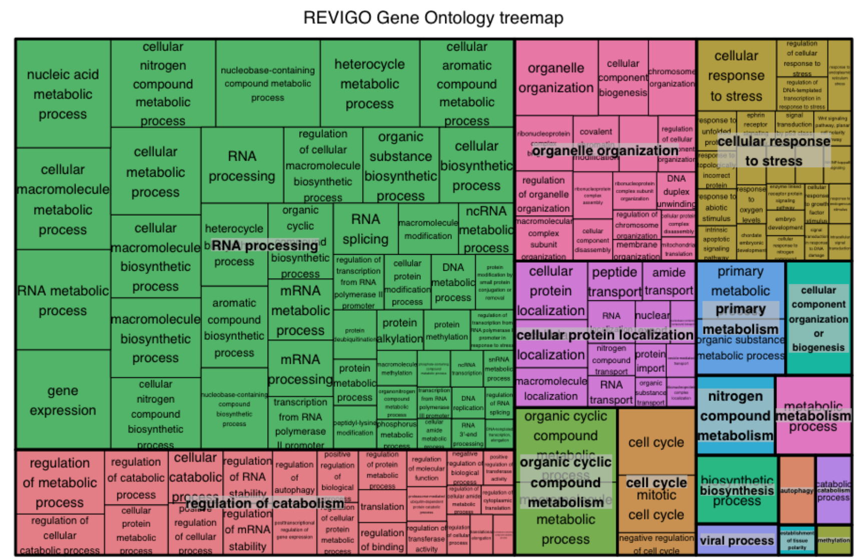
Hierarchies: Examples

Sunburst Diagram



http://cancerdiscovery.aacrjournals.org/content/5/11/1210

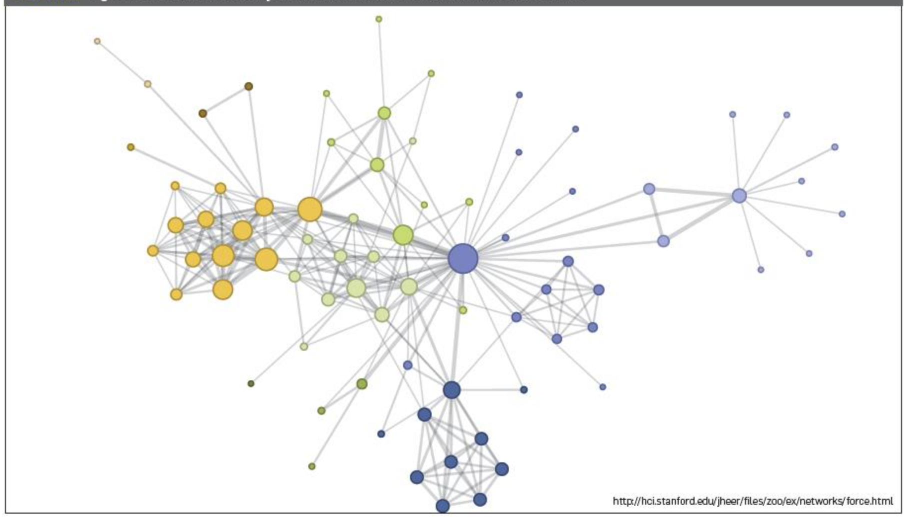
Treemap



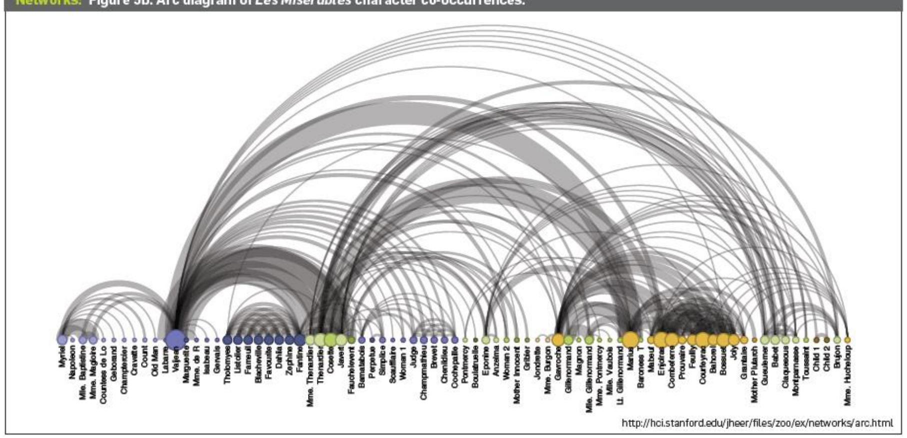
https://hbctraining.github.io/DGE_workshop/lessons/functional_analysis_other_methods.html

Networks

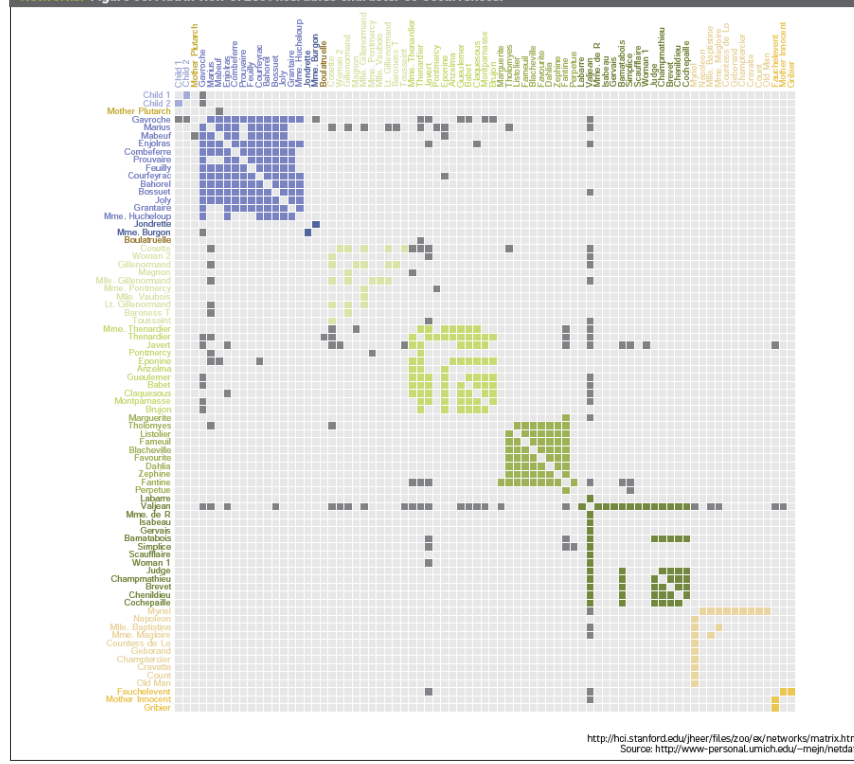
Networks: Figure 5a. Force-directed layout of *Les Misérables* character co-occurrences.



Networks: Figure 5b. Arc diagram of *Les Misérables* character co-occurrences.



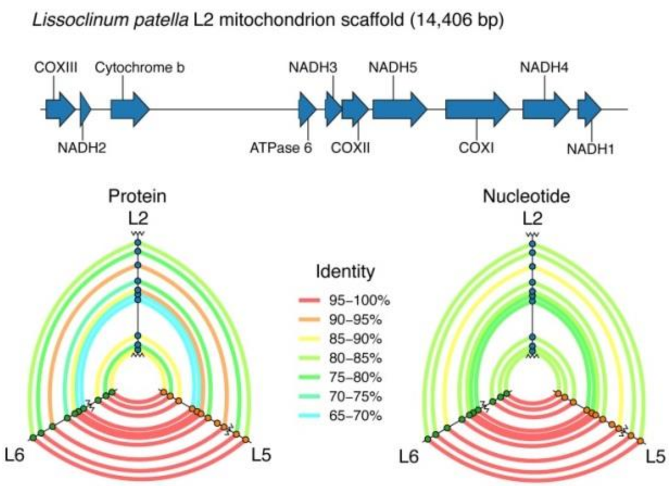
Networks: Figure 5c. Matrix view of *Les Misérables* character co-occurrences.



Heer, J. et al. *A Tour Through the Visualization Zoo* ACM Queue 53:6 p.59-67 (2010)

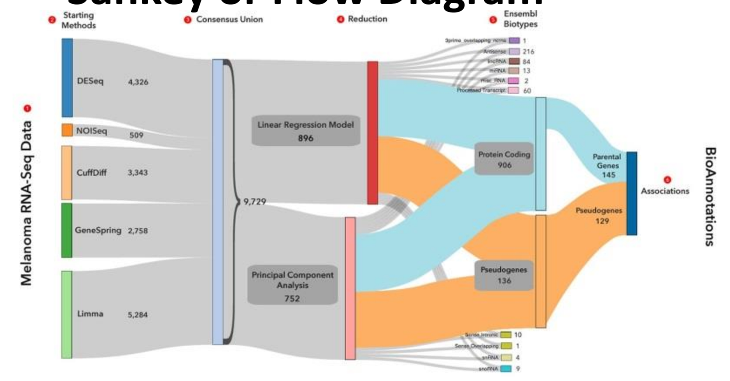
Visualization Tour: Others

Hive Plots



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0095850

Sankey or Flow Diagram



Notes

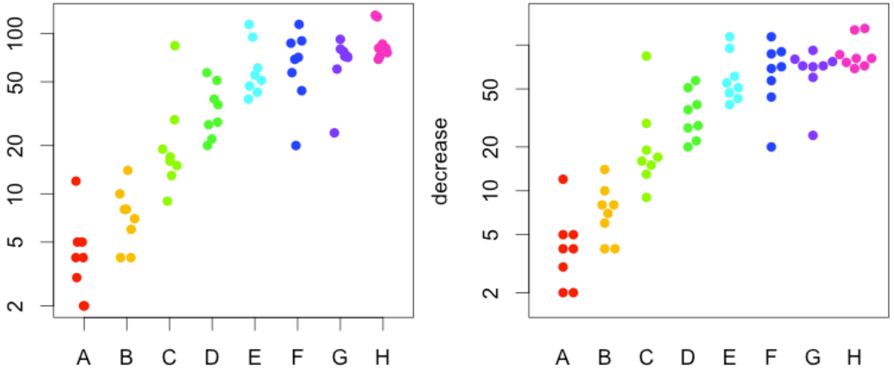
The flow refers to differentially expressed biotypes measured by different methods.

1. 103 Melanoma samples from The Cancer Genome Atlas (TCGA), and 2 Melanocyte control samples from the Gene Expression Omnibus (GEO)
2. Algorithms used for detecting differential expression
3. Consensus of differentially expressed bioentities
4. Reduction in size by linear regression model (LRM) and principal component analysis (PCA)
5. Final decomposition of biotypes
6. Particular focus goes to the association of pseudogenes and parental protein coding genes

* Non-coding Biotypes
Ensembl Biotypes such as miRNA and other short non-codings are just the byproducts of the detection protocol. They are reported here as an outcome of the read-mapping uncertainty in the alignment step.

https://www.nature.com/articles/s41598-017-17337-7

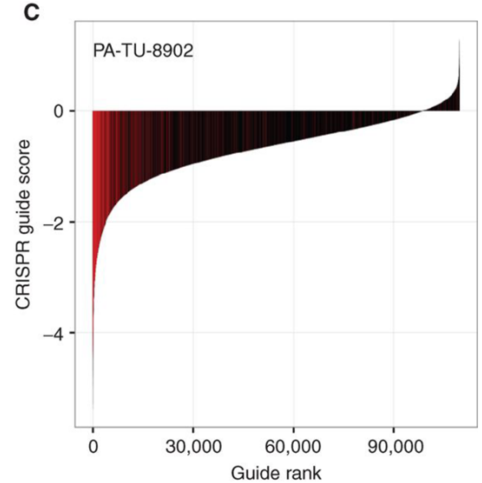
Stripchart and Beeswarm



http://www.cbs.dtu.dk/~eklund/beeswarm/

Waterfall Plot

c

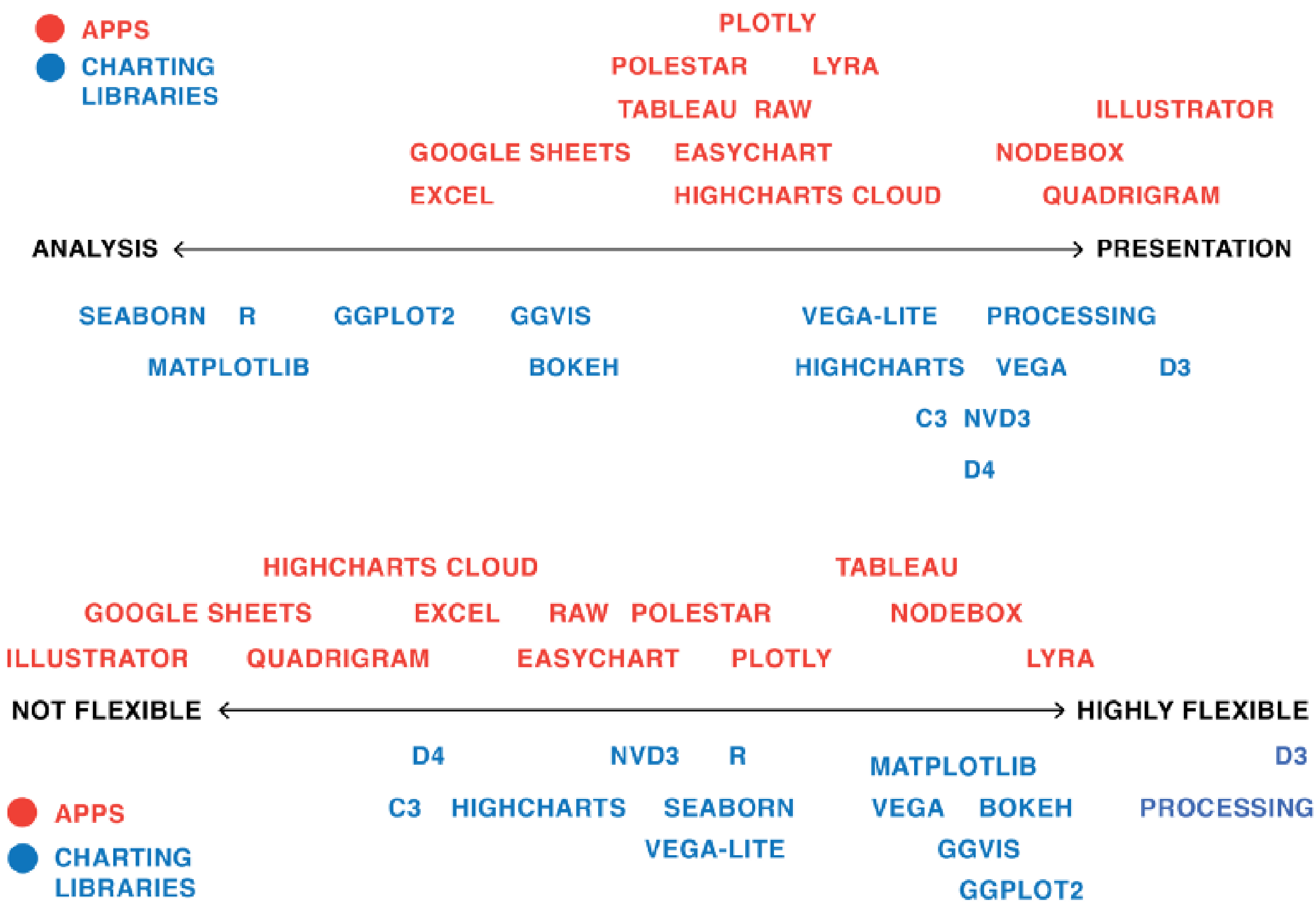


http://cancerdiscovery.aacrjournals.org/content/6/8/914

Resource	Description	URL
Discovery[b]		
Excel[c]	Everyday tool for generic visualization of smaller data sets	**http://microsoft.com/excel**
Plotly	Online tool for fast data visualization	**https://plot.ly/create/**
Tableau[c]	For interactive visualizations, including web based	**http://tableau.com**
Spotfire[c]	For visual analysis of larger data sets and tool generation	**https://spotfire.tibco.com/**
Origin[c,d]	For visual analysis of larger data sets	**http://originlab.com**
Mathematica[c]	For visual analysis of data sets and mathematical functions	**http://wolfram.com**
MATLAB[c]	For visual analysis of data sets and mathematical functions	**http://mathworks.com**
Matplotlib	For tailored visualizations of data sets in Python (115)	**http://matplotlib.org**
ggplot2	For tailored visualizations of large, complex data sets in R (116)	**http://ggplot2.org**
D3.js	For tailored, interactive web-based visualizations	**https://d3js.org**
Communication		
Photoshop[c]	For editing imaging data	**http://adobe.com/photoshop**
GIMP	Free, open-source alternative to Photoshop	**http://www.gimp.org**
Illustrator[c]	For creating and editing vector graphics	**http://adobe.com/illustrator**
Inkscape	Free, open-source alternative to Illustrator	**http://inkscape.org**
MolecularMaya	Molecular structure plug-in for Autodesk Maya[c] animation suite	**http://bit.ly/molmaya**
BioBlender	Molecular structure plug-in for Blender animation suit	**http://bioblender.org**
Utilities		
Color Brewer	Web tool for selecting contrasting color maps	**http://colorbrewer2.org**
Adobe Color	Web tool for designing sets of colors	**http://color.adobe.com**
Paletton	Web tool for designing sets of colors	**http://paletton.com**
General Resources		
BioVis	Computer science publications on biological visualizations	**http://biovis.net**
Clarafi[c]	Training guides for biomedical visualization tools	**http://clarafi.com**
Information is Beautiful	Showcase of charts and infographics for a wide variety of data	**http://bit.ly/Info_Beauty**
Visual Complexity	Catalog of tailored visualizations for complex data	**http://visualcomplexity.com**
VIZBI	Collected videos and posters on tailored biological visualizations	**http://vizbi.org**
Exemplars		
PDB101	Outstanding visual explanations of protein function and structure	**https://pdb101.rcsb.org**
Roche pathway	Tailored visualization showing ~3,000 metabolic reactions (72)	**http://bit.ly/RochePathway**
WEHI.tv	Collection of inspiring, informative biomedical animations	**http://wehi.tv**

O'Donoghue, S.I, et al. *Visualization of Biomedical Data*. Annual Rev of Biomed Data Sci 1:275-304 (2018)

Visualization: Software



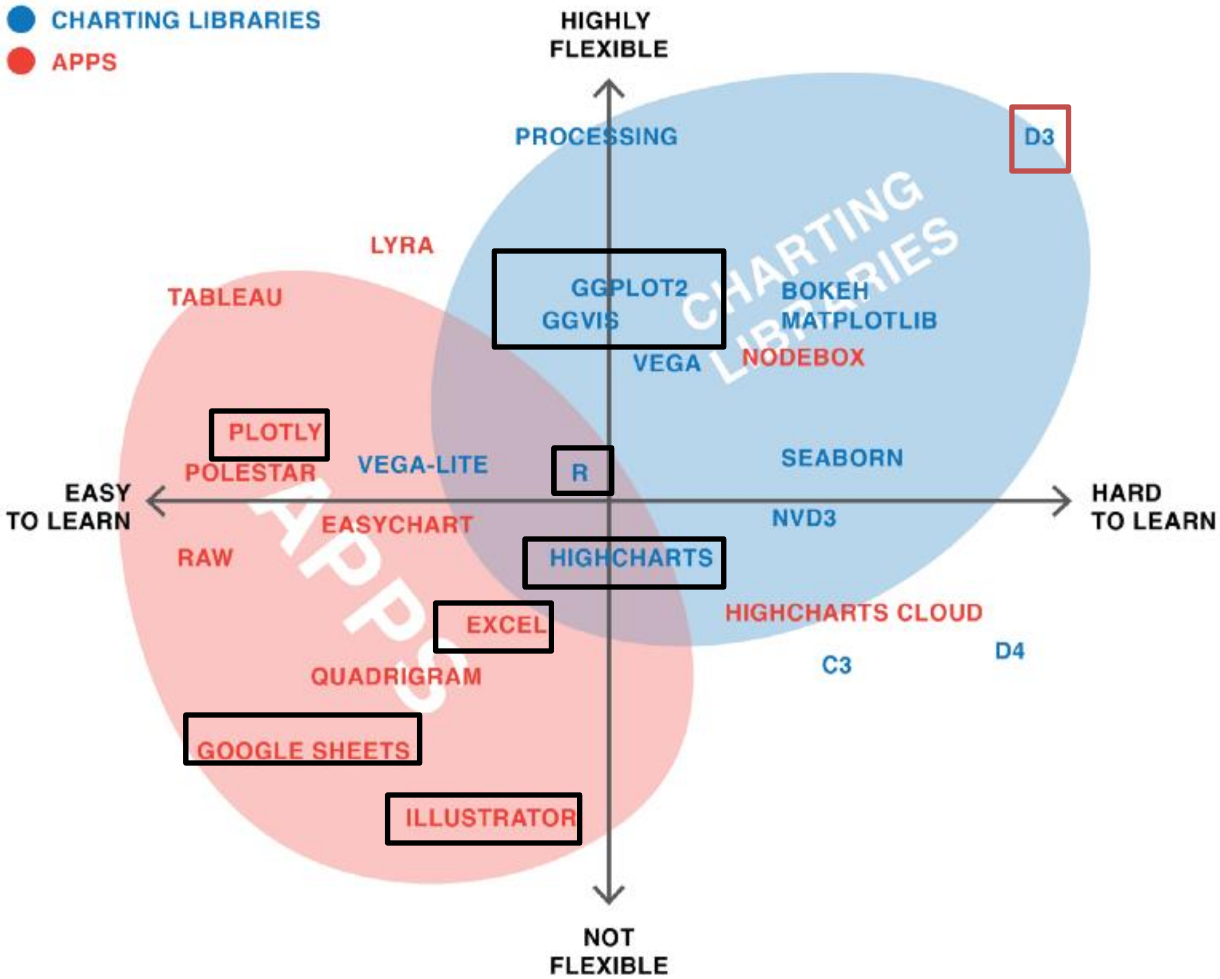
https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

Visualization: Static vs Interactive Software

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

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

Visualization: Software



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

Additional Reading

- Ten Simple Rules for Better Figures (PLOS)
  - Rougier, N.P, et al.
- Fundamentals of Data Viz.

https://serialmentor.com/dataviz/

- Points of View (Nature Methods)

http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html