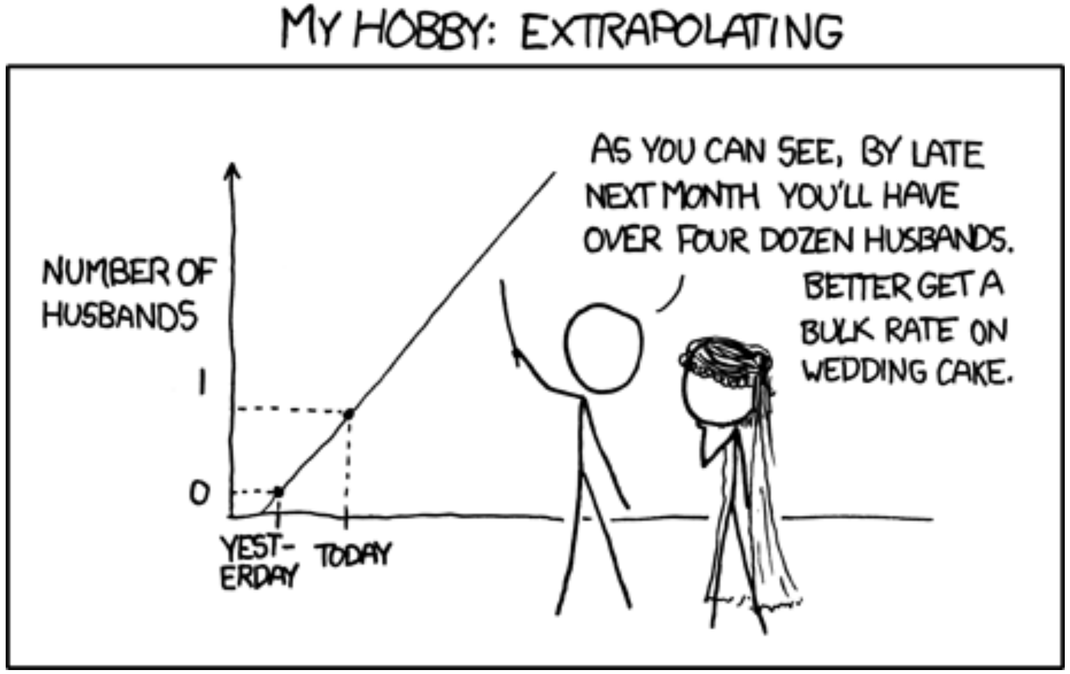
Applied Data Visualization
Tabular Data

Alexander Lex
alex@sci.utah.edu

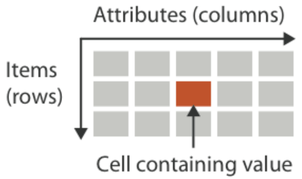
THE UNIVERSITY OF UTAH



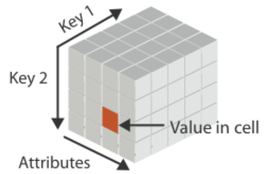
[xkcd]

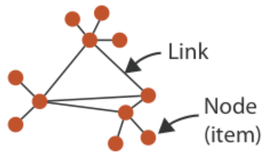
Dataset Types

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

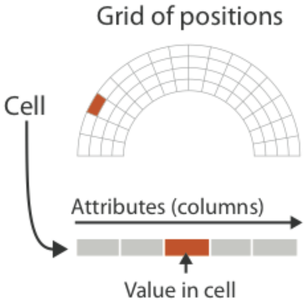


➔ *Multidimensional Table*





➔ *Trees*



Exercise: Sketch 2 Ways to Vis. Each Table

	Age	Best 100 m	Furthest Jump	Sex
Amy	16	13.2	5.2	F
Basil	18	12.4	4.2	F
Clara	14	14.1	2.5	F
Desmond	22	10.01	6.3	M
Charles	19	11.3	5.3	M

	BPM T1	BPM T2	BPM T3
Amy	90	130	150
Basil	70	110	109
Clara	60	140	141
Desmond	84	100	108
Charles	81	110	130

Scale of Tables

Need different approaches for “normal” and “high-dimensional” tables.

How many dimensions?

~50 – tractable with “just” vis

~1000 – need analytical methods

How many records?

~ 1000 – “just” vis is fine

>> 10,000 – need analytical methods

Homogeneity

Same data type?

Same scales?

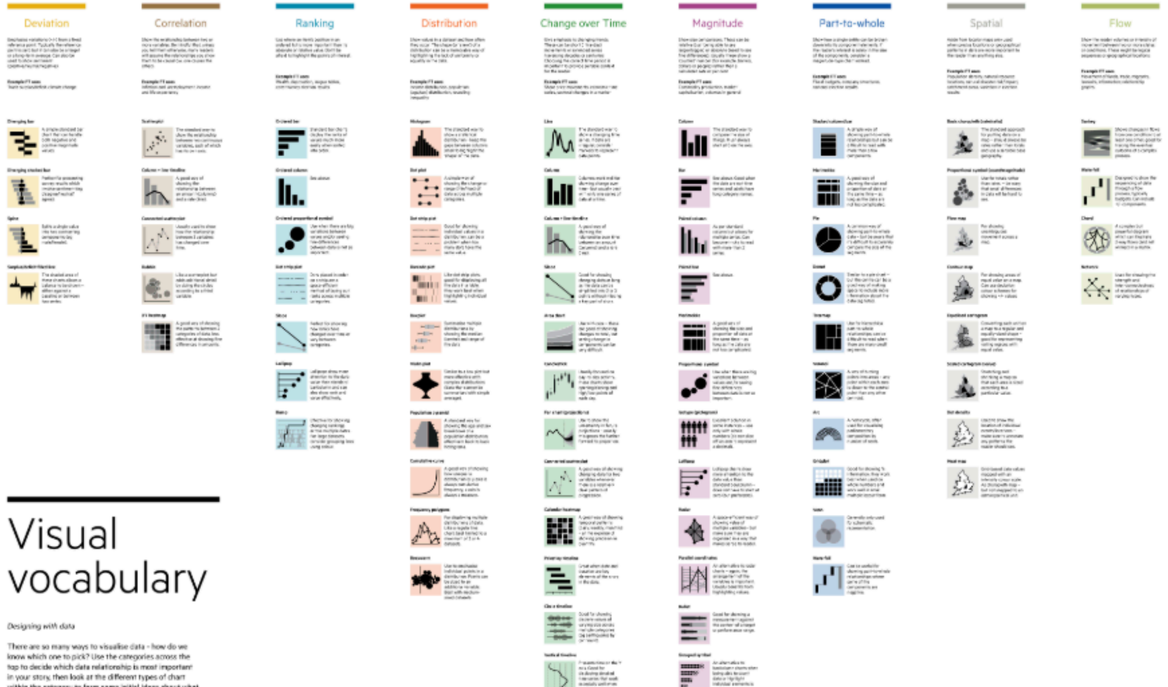
	Age	Gender	Height
Bob	*25*	*M*	*181*
Alice	*22*	*F*	*185*
Chris	*19*	*M*	*175*

	BPM 1	BPM 2	BPM 3
Bob	*65*	*120*	*145*
Alice	*80*	*135*	*185*
Chris	*45*	*115*	*135*

Techniques and Tasks

Magnitude Distribution Deviation Correlation

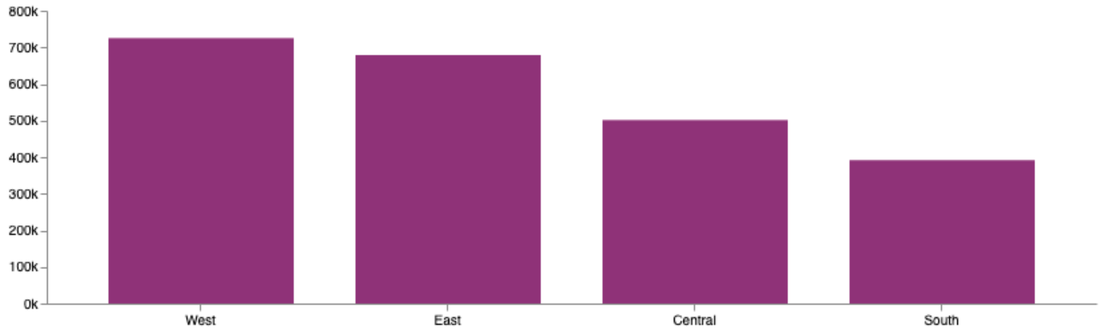
Ranking Part to whole Change over Time



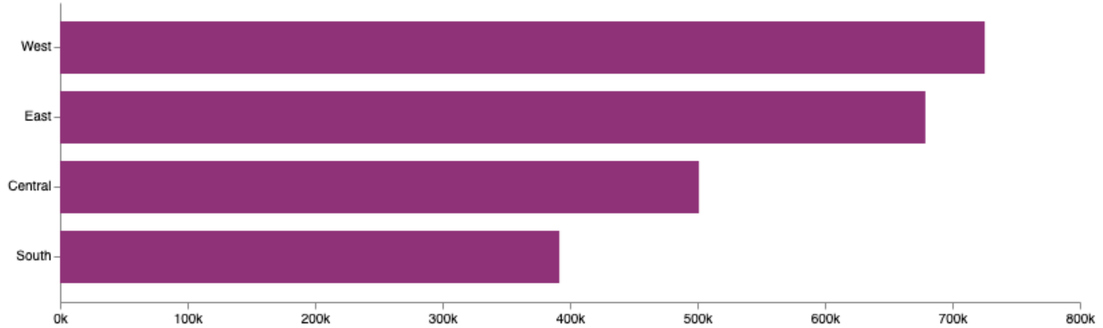
https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary
https://gramener.github.io/visual-vocabulary-vega/#/Magnitude/

Magnitude

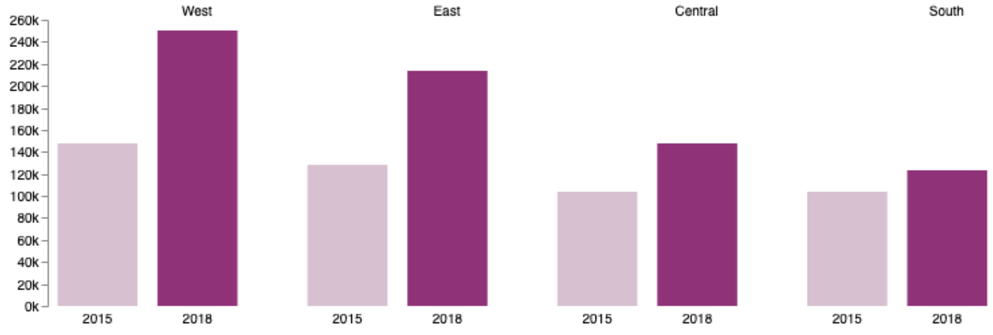
Bar Chart Variants



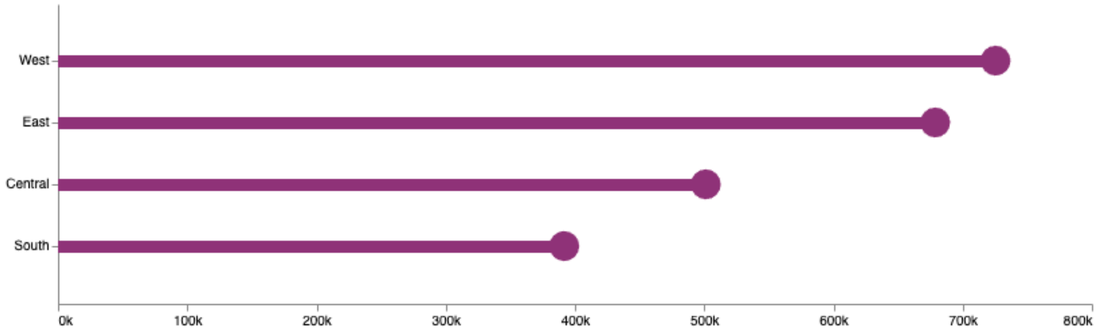
Vertical Bar Chart / Column Chart



Horizontal Bar Chart



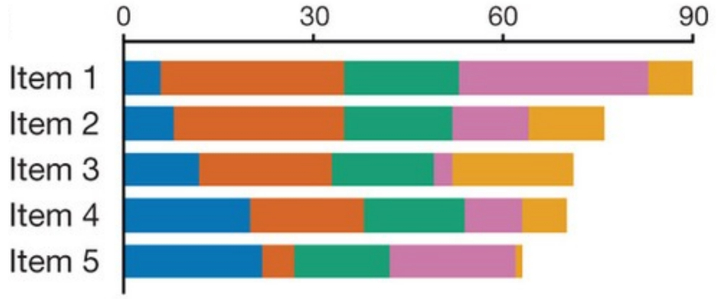
Grouped Bar Chart



Lollipop Chart

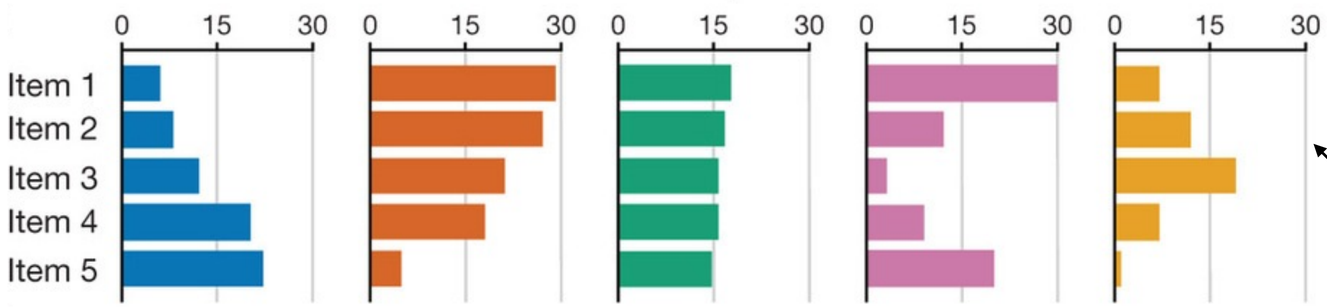
Comparison of bar chart types

Pie Chart

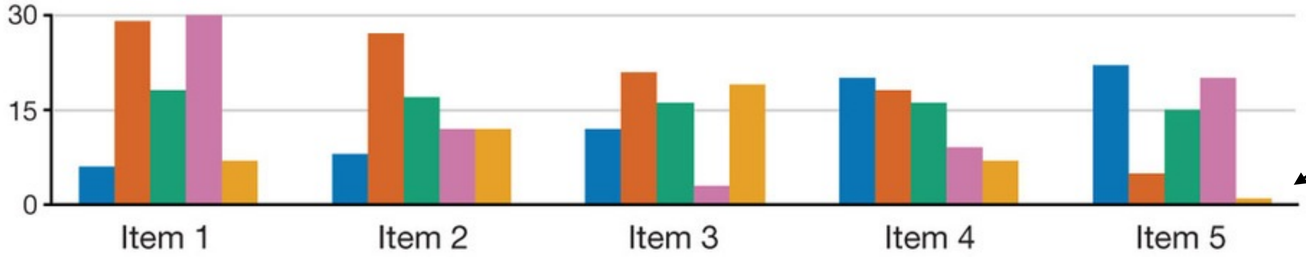


Stacked bar chart

Layered Bar Chart



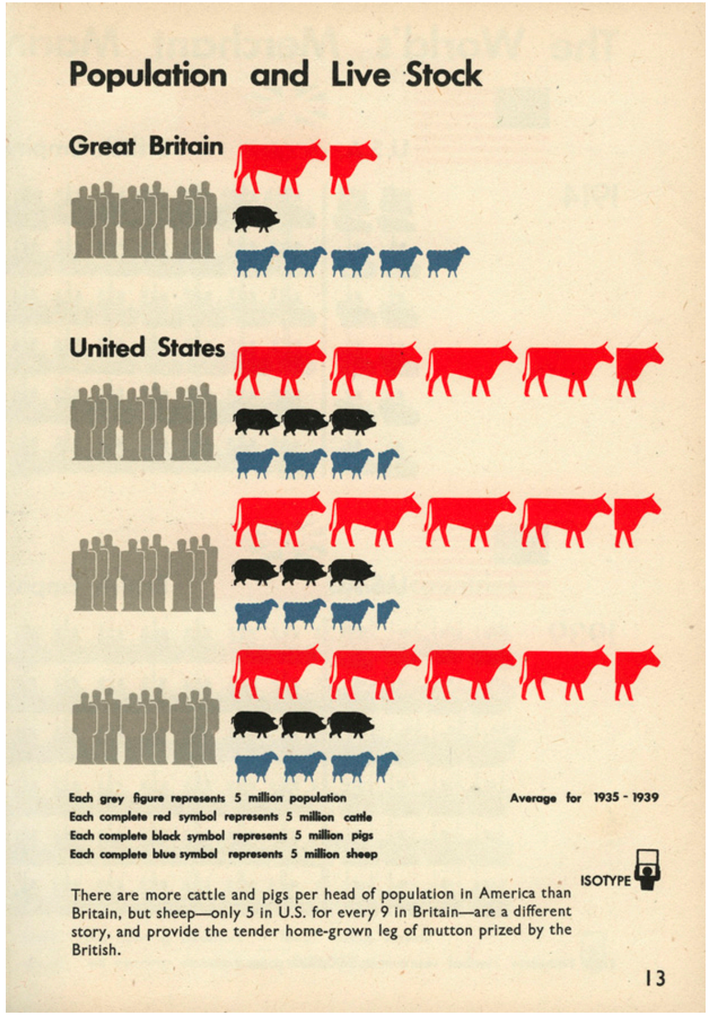
Grouped Bar Chart



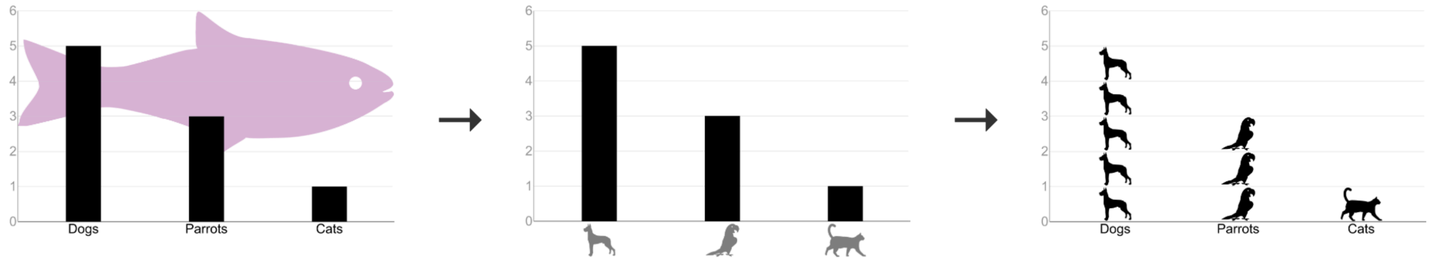
Small Multiples

Streit & Gehlenborg, PoV, Nature Methods, 2014

IsoType Visualization



Otto and Marie Neurath



http://steveharoz.com/research/isotype/

Part of Whole

Stacked Bar Chart

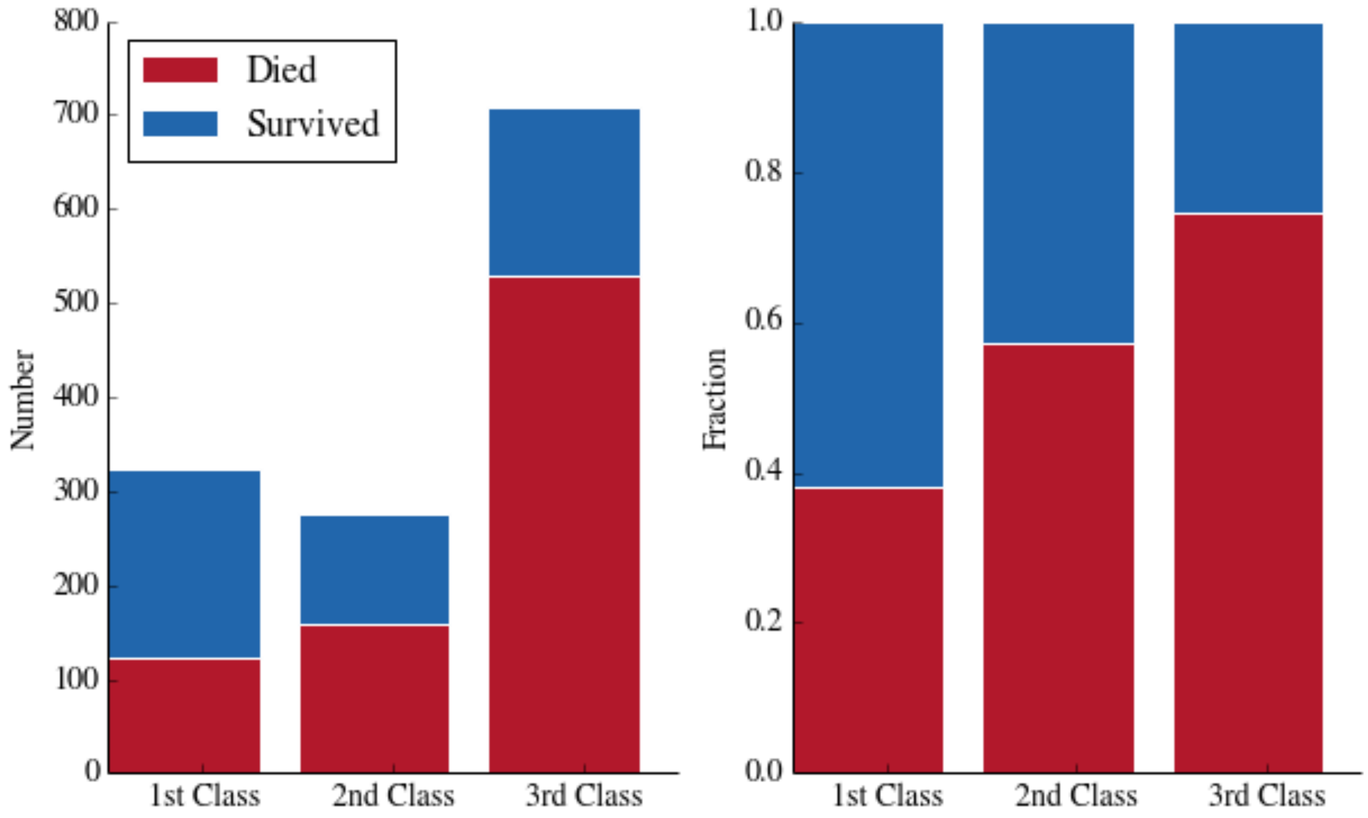
Keys: Class, Survival

Class is spatial

Survival is color

Left: absolute values

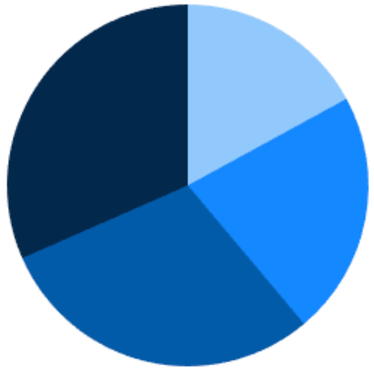
Right: proportional values



Pie and Donut Charts

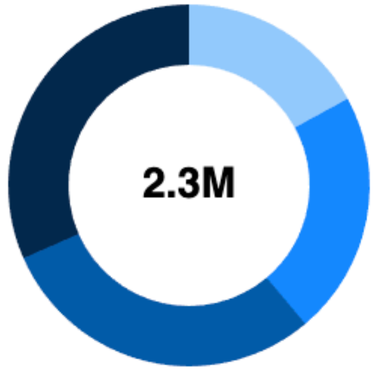
Pie Edit

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

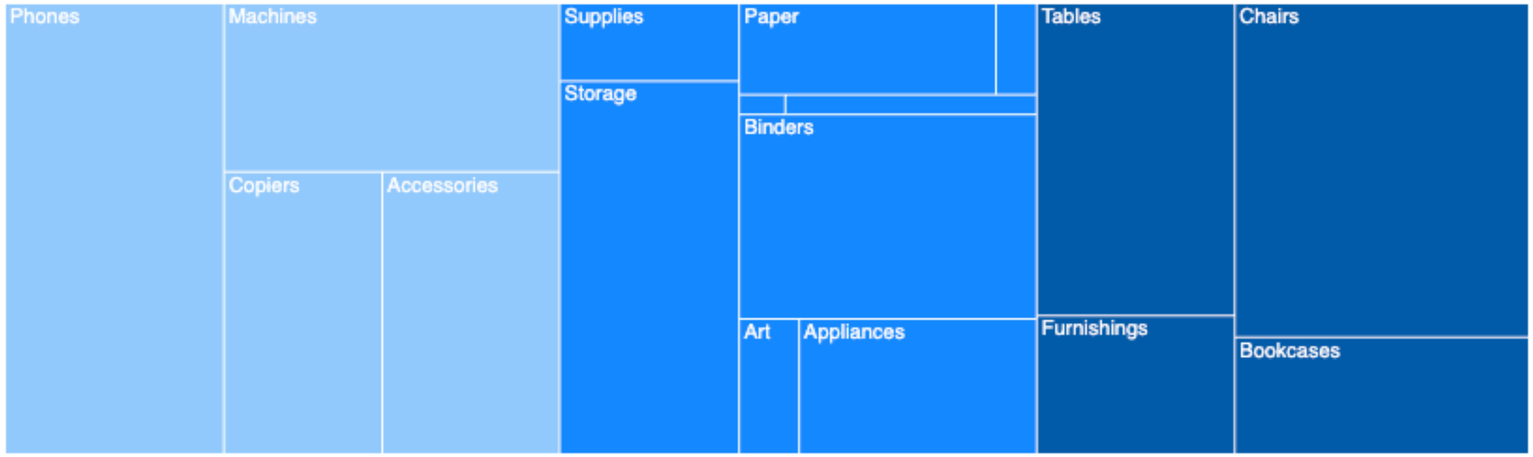
Similar to a pie chart - but the centre can be a good way of making space to include more information about the data (eg. total)



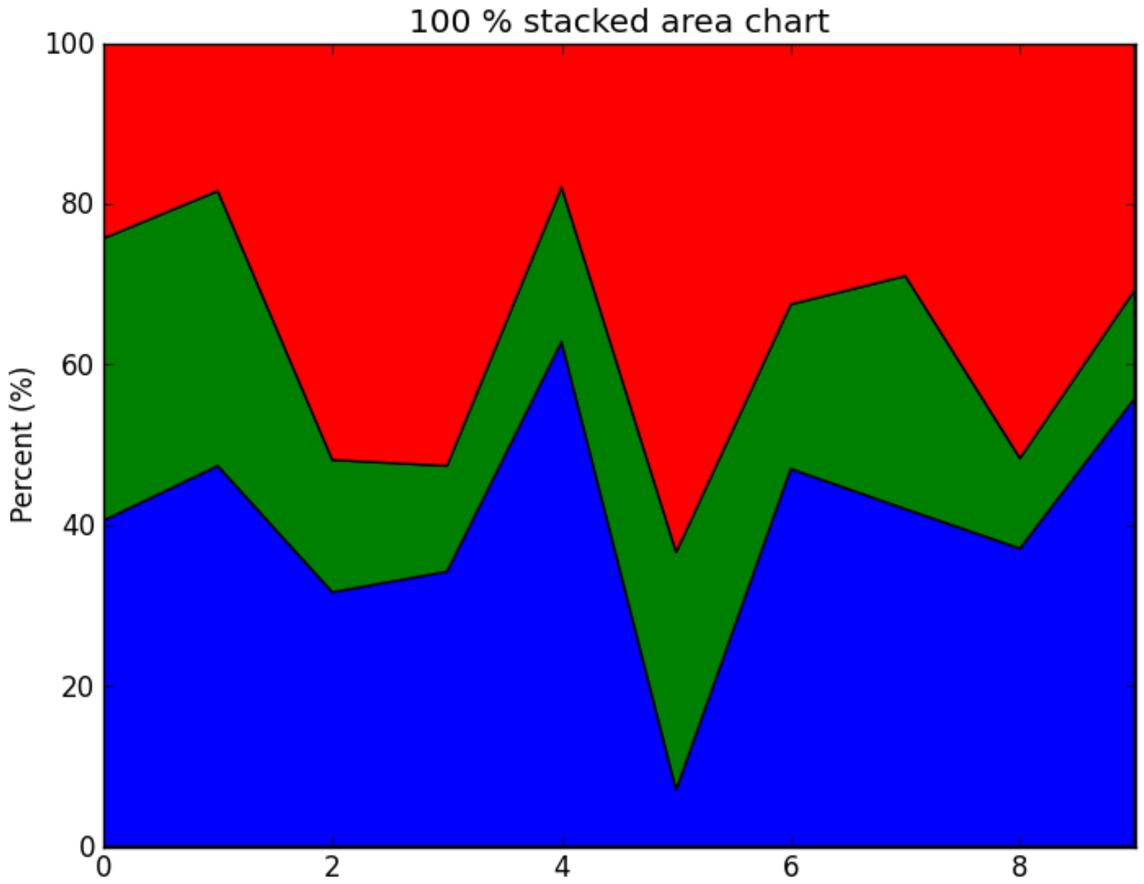
TreeMap

Treemap

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



Part of Whole for Time Series



Distribution

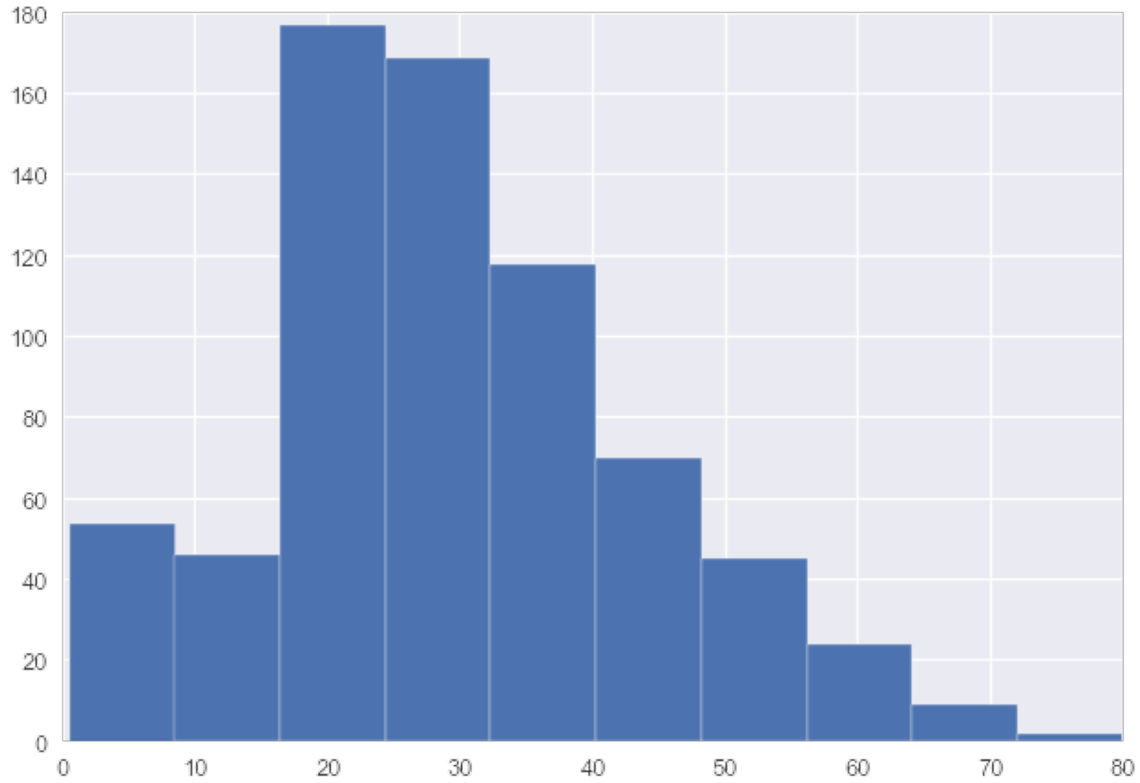
Aggregating Large Data Vectors

Instead of showing all data points, show a data's distribution

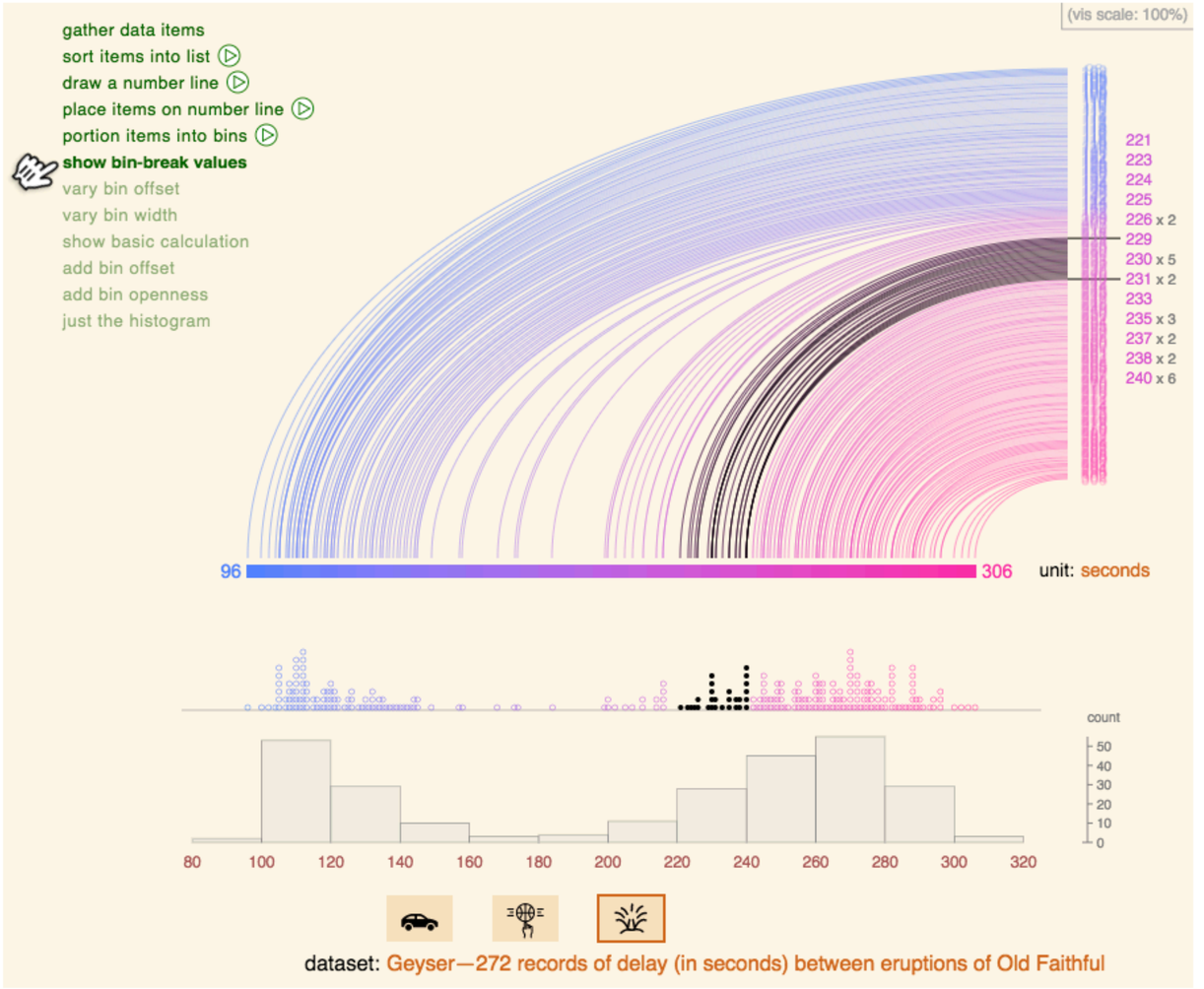
Pro: compact representation

Con: Works only if data is "well behaved" for the type of distribution visualization.

What’s a histogram?



Histograms Explained



http://tinlizzie.org/histograms/

Histogram

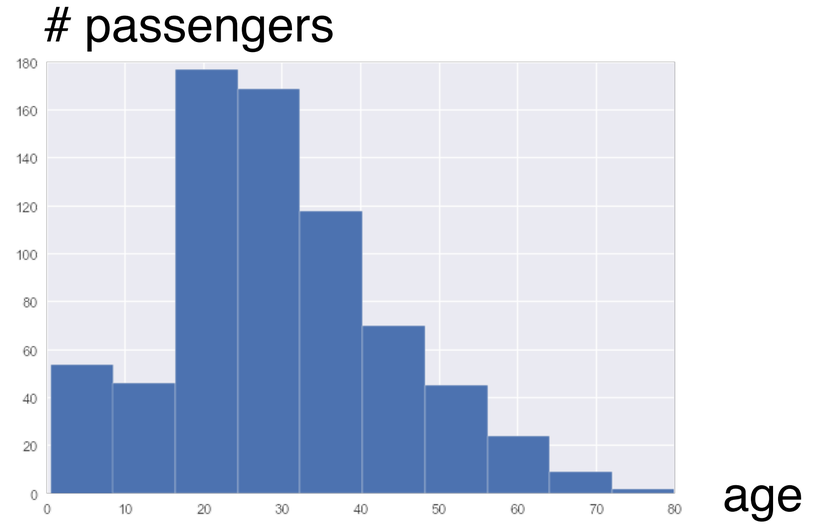
Good #bins hard to predict

make interactive!

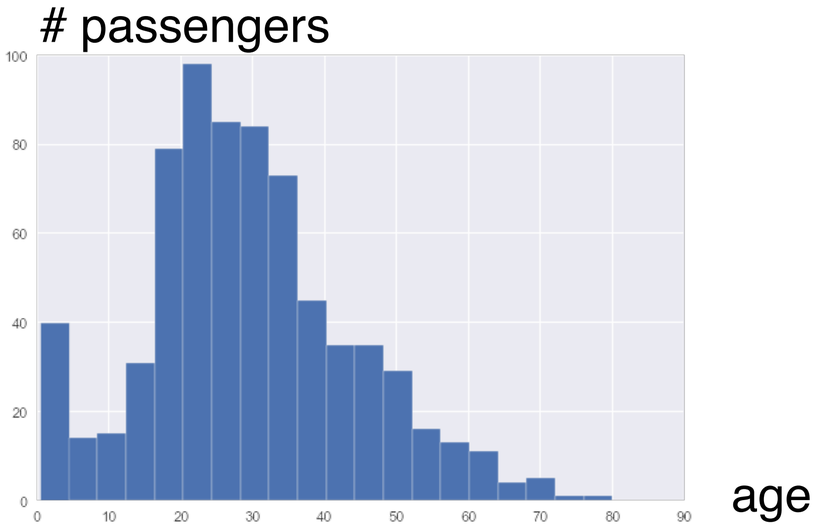
rules of thumb:

#bins = sqrt(n)

#bins = log2(n) + 1

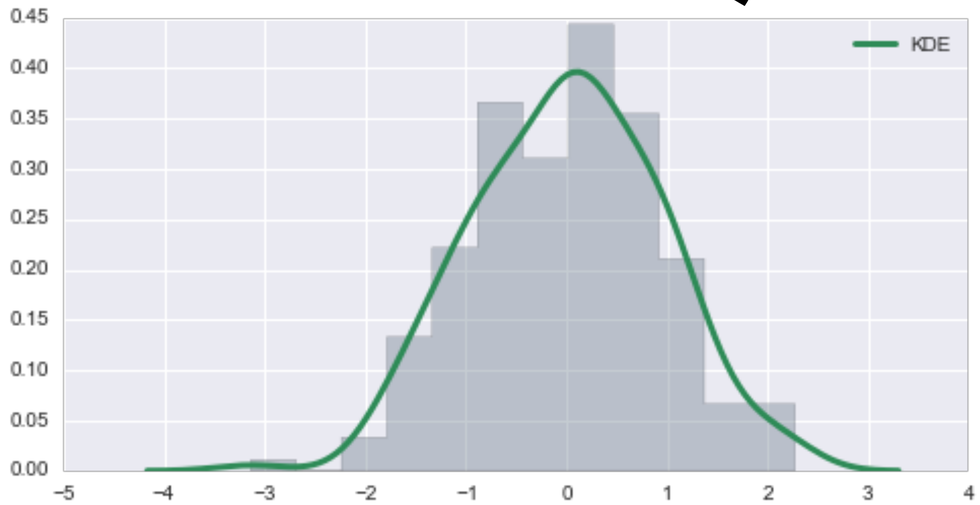


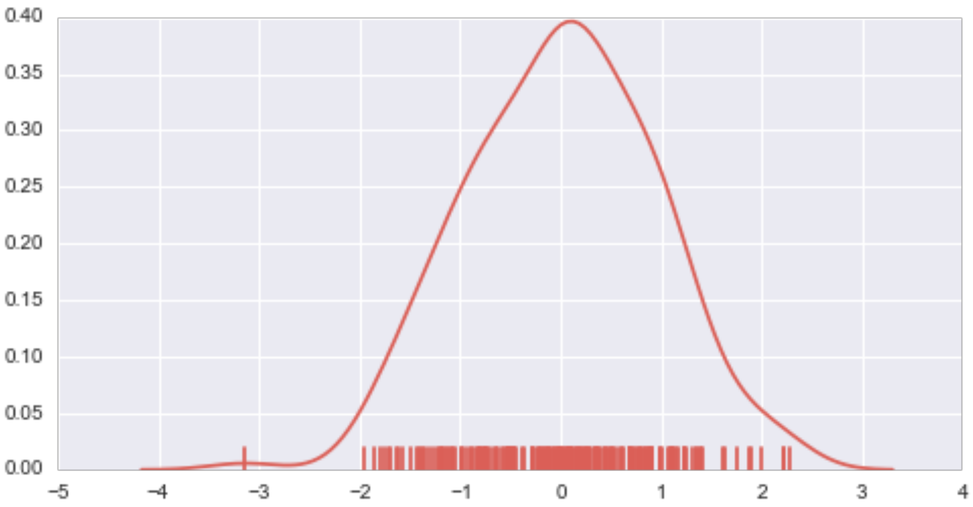
10 Bins

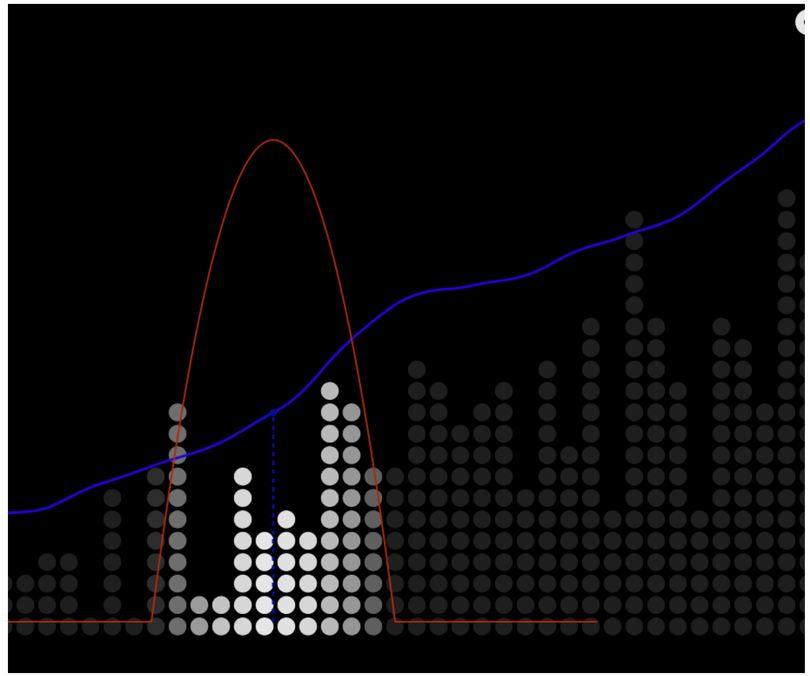


20 Bins

Density Plots (Kernel Density Estimation)







http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

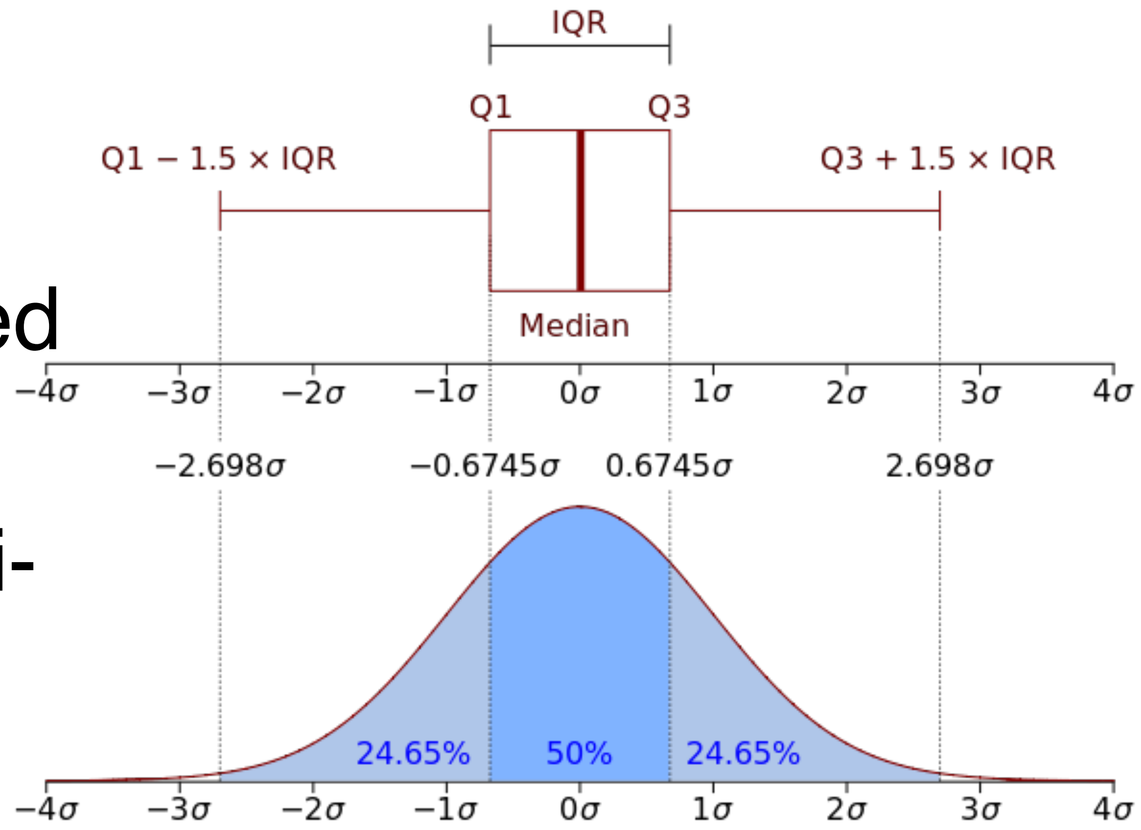
Box Plots

aka Box-and-Whisker Plot

Show outliers as points!

Bad for non-normal distributed data

Especially bad for bi- or multi-modal distributions



Wikipedia

One Boxplot, Four Distributions

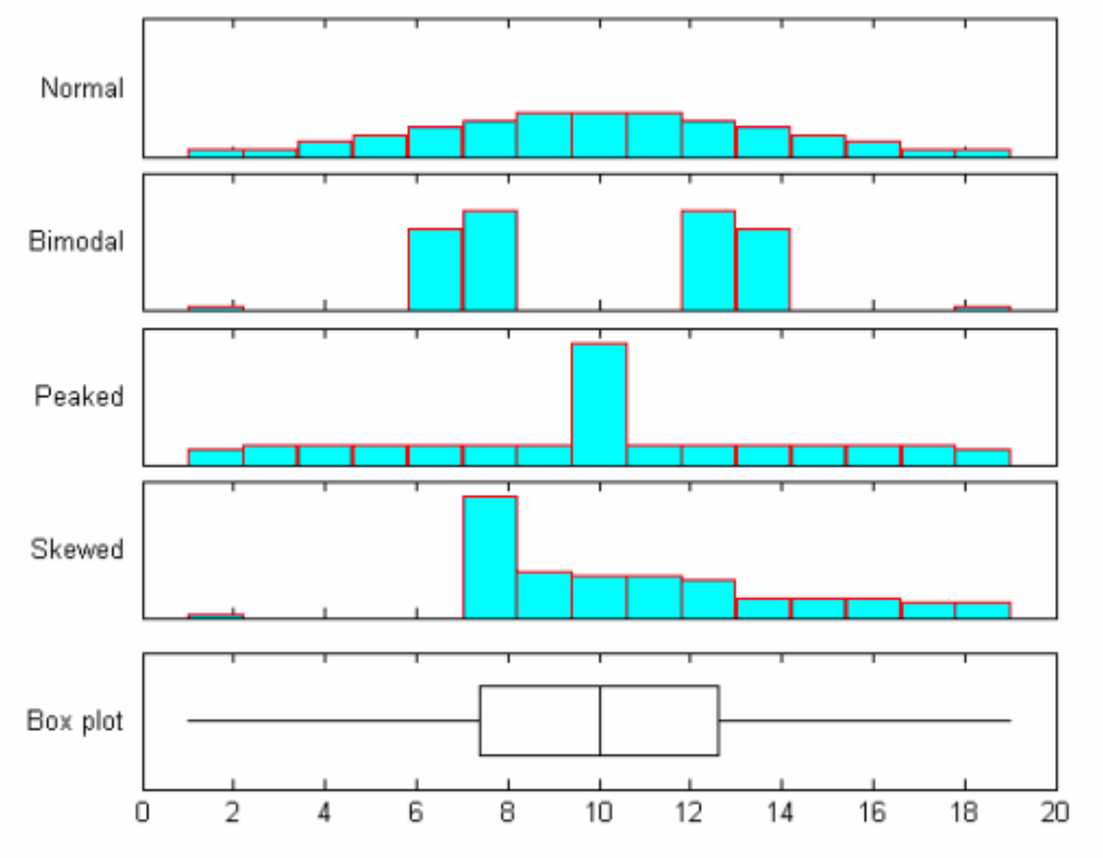
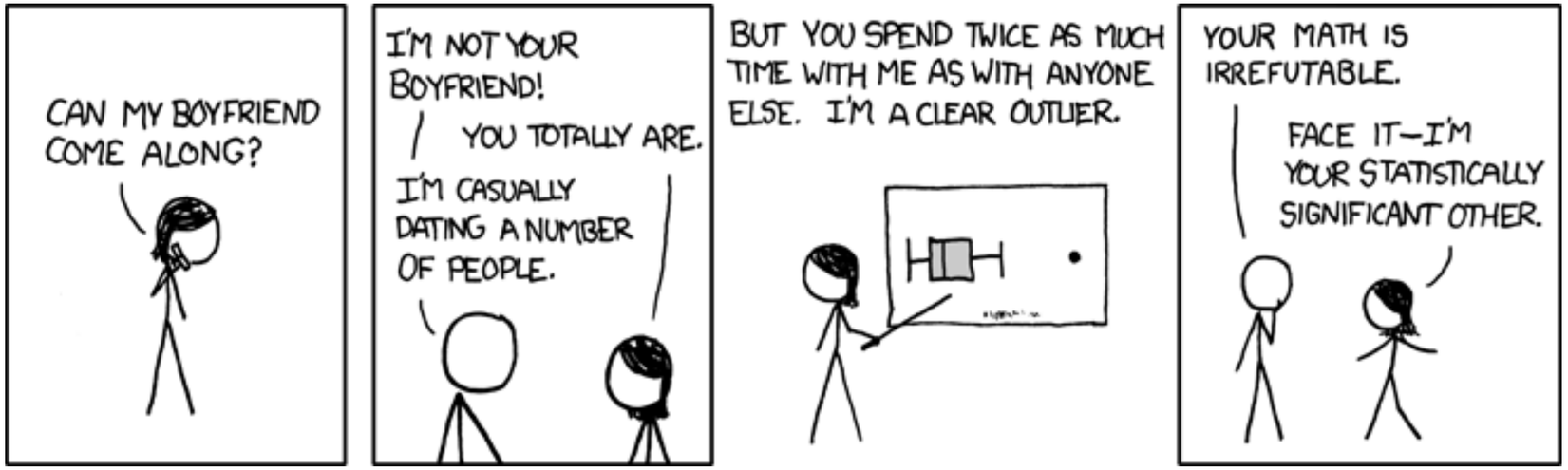


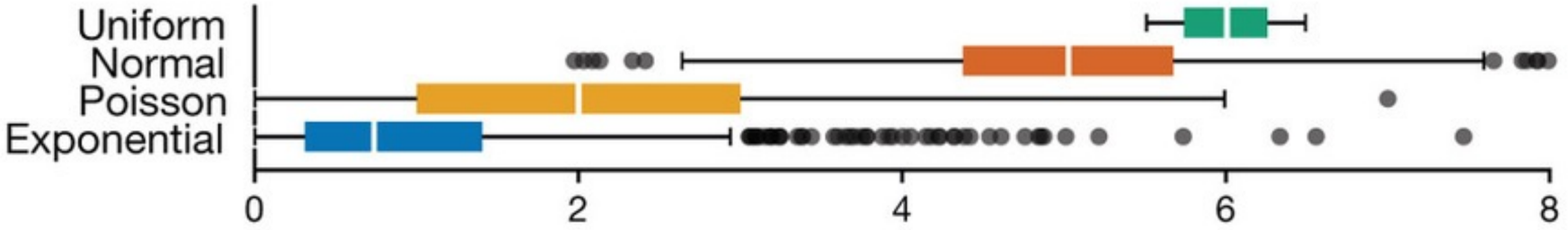
Figure 1: Histograms and box plot: four samples each of size 100

http://stat.mq.edu.au/wp-content/uploads/2014/05/Can_the_Box_Plot_be_Improved.pdf

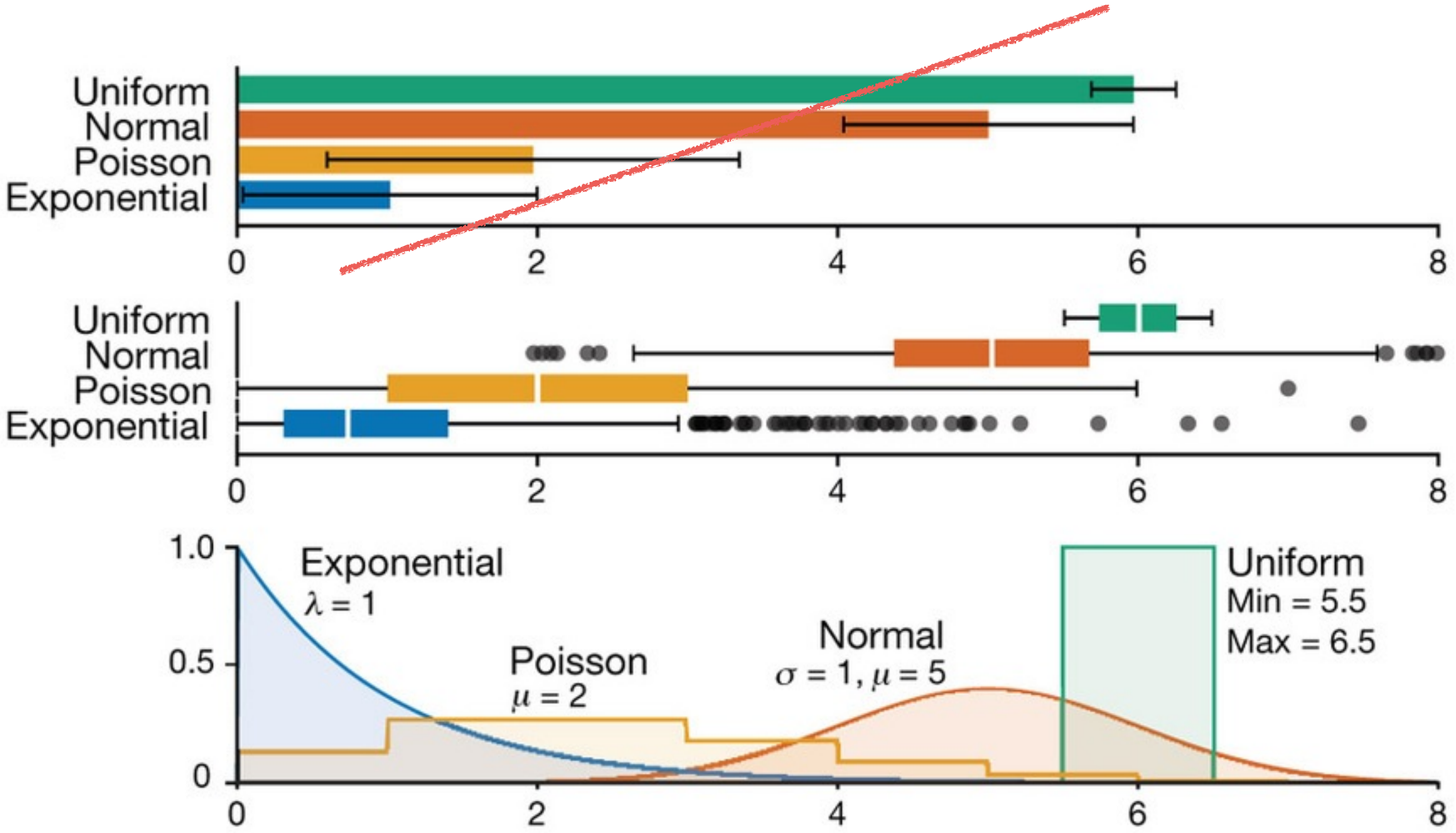
Box(and Whisker) Plots



http://xkcd.com/539/

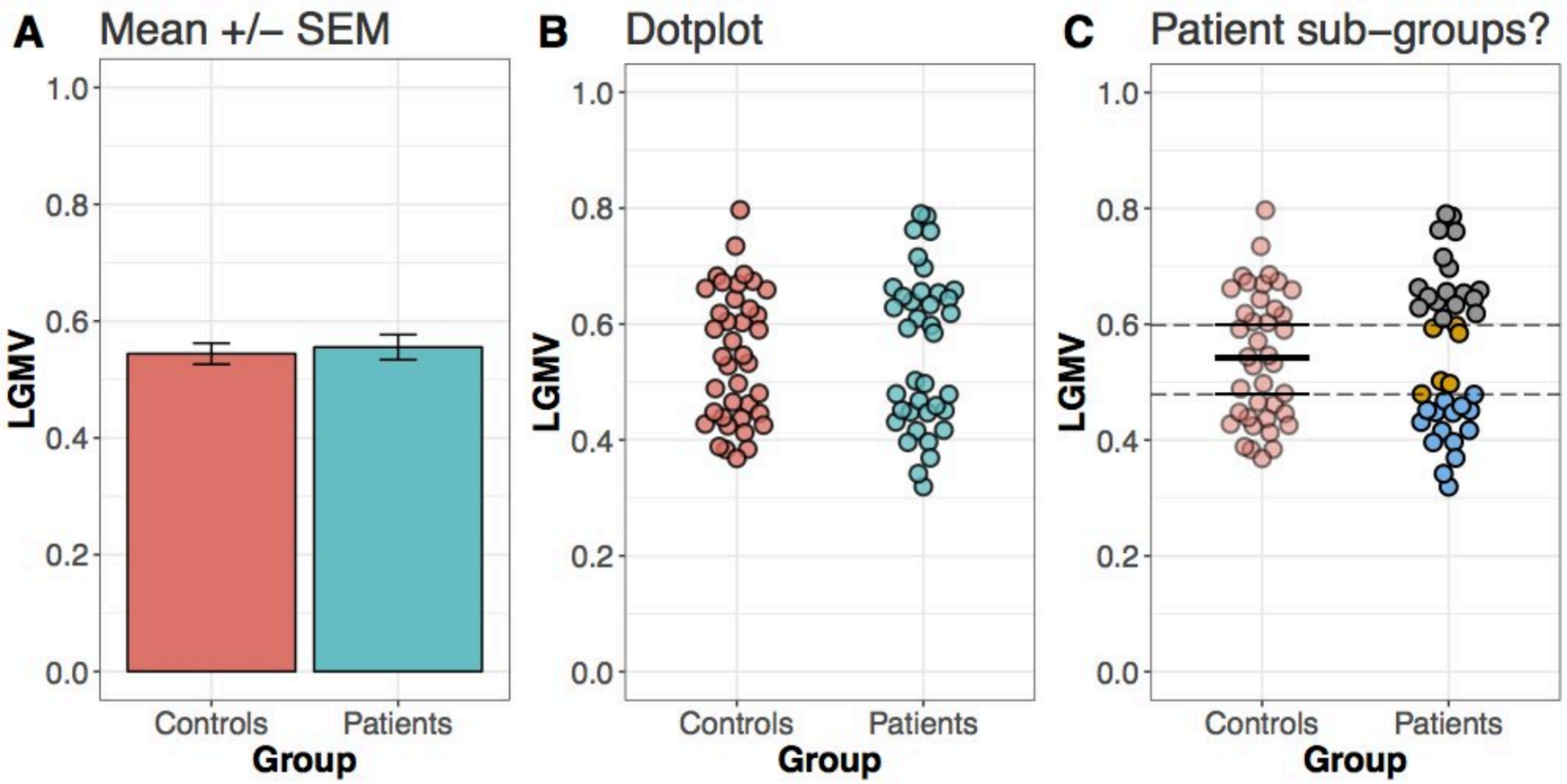


Comparison



Streit & Gehlenborg, PoV, Nature Methods, 2014

Bar Charts vs Dot Plots

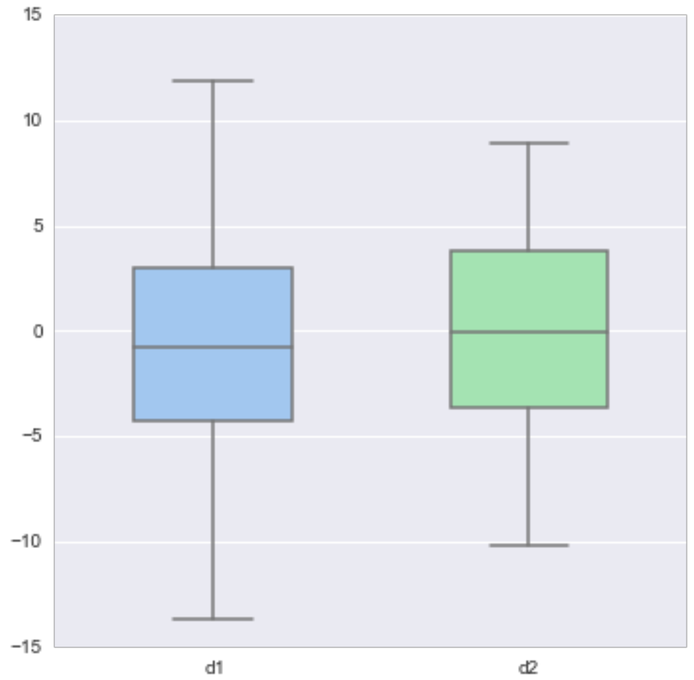


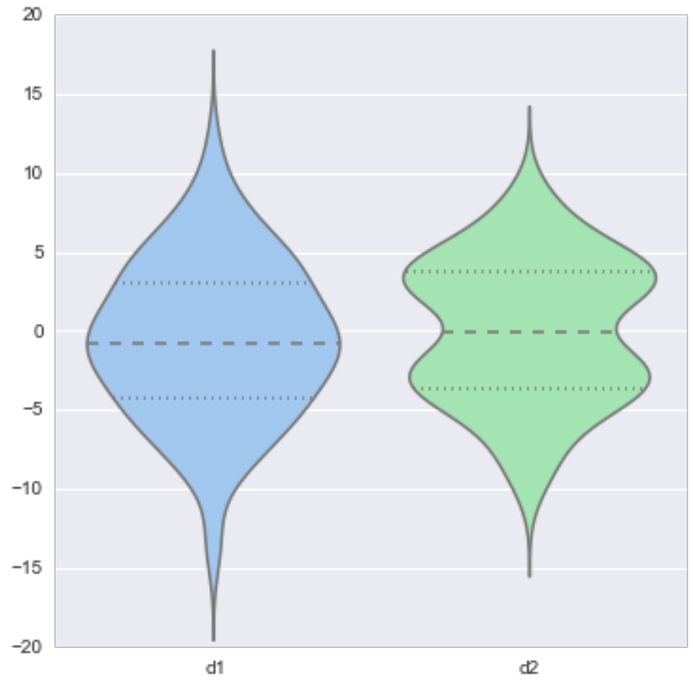
Data Source https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-10-67

https://twitter.com/robustgar/status/859318971920769024

Violin Plot

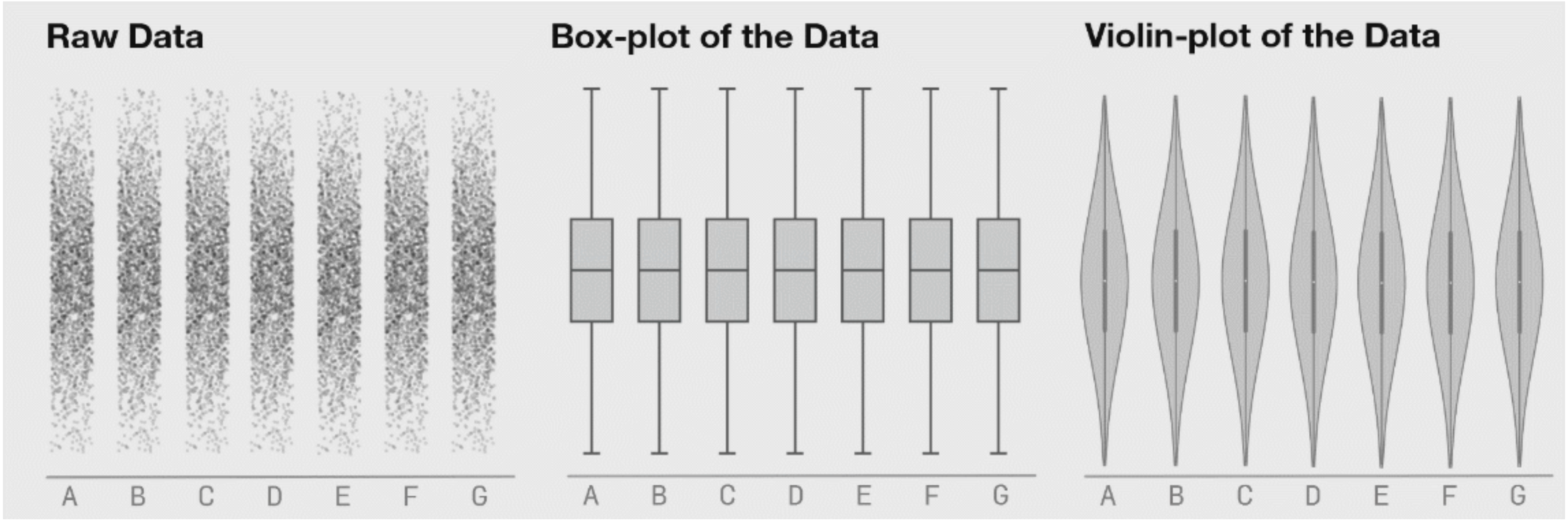
= Box Plot + Probability Density Function





http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

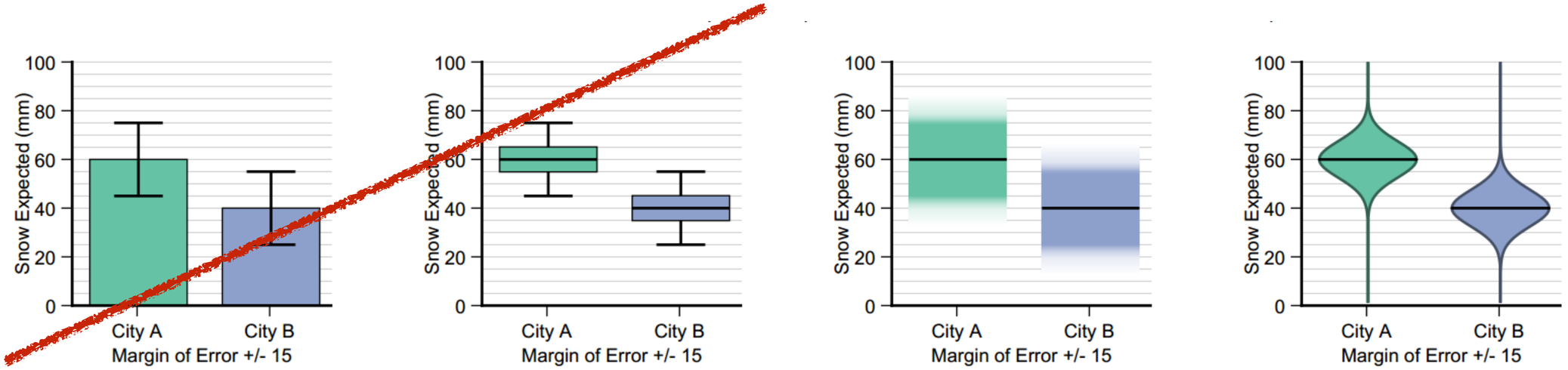
Different Distributions



https://blog.bioturing.com/2018/05/16/5-reasons-you-should-use-a-violin-graph/

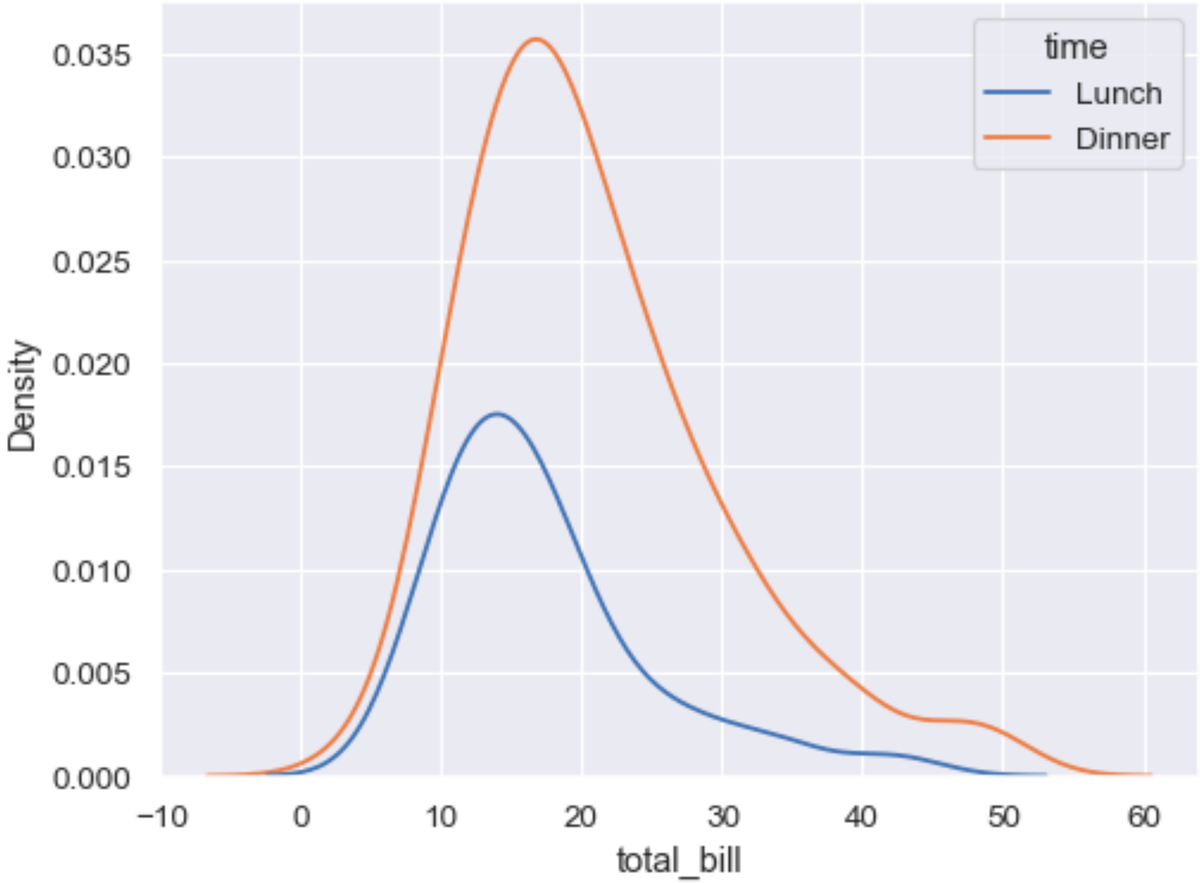
Showing Expected Values & Uncertainty

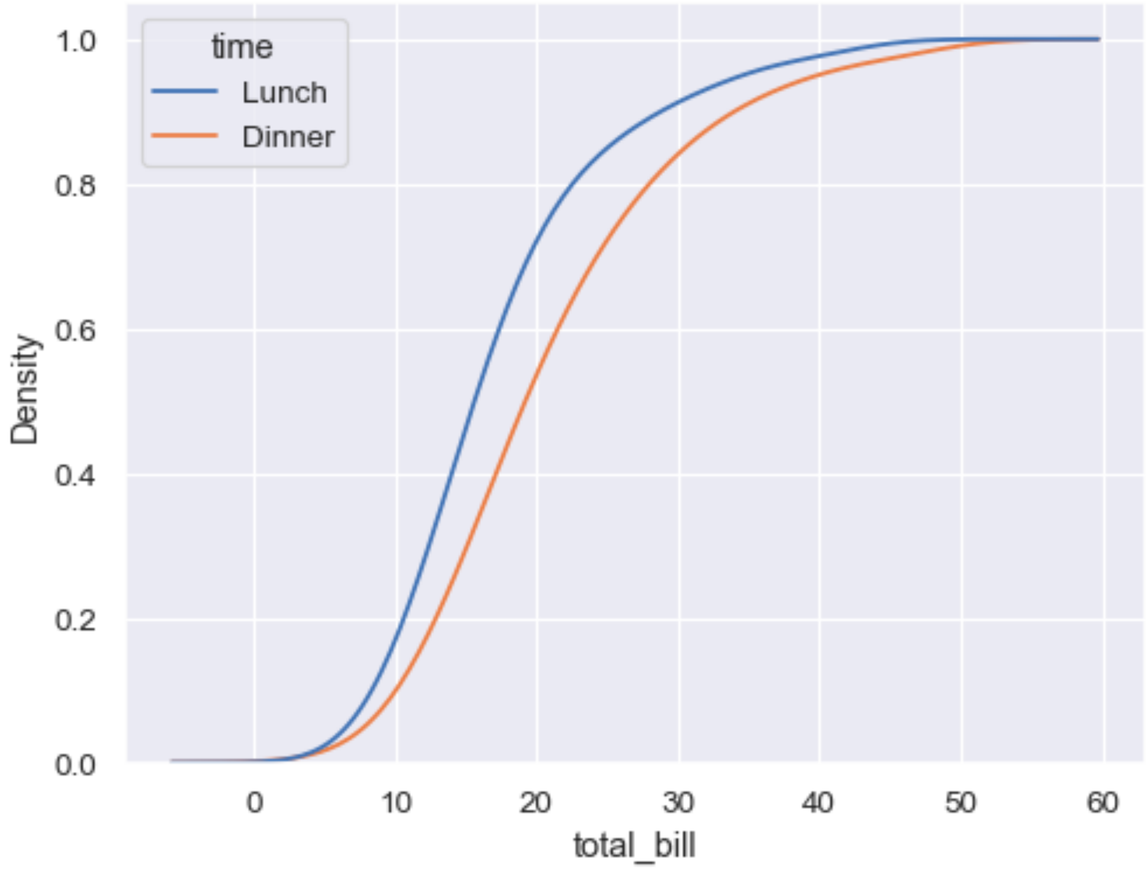
NOT a distribution!



Error Bars Considered Harmful:
Exploring Alternate Encodings for Mean and Error
Michael Correll, and Michael Gleicher

Cumulative KDE

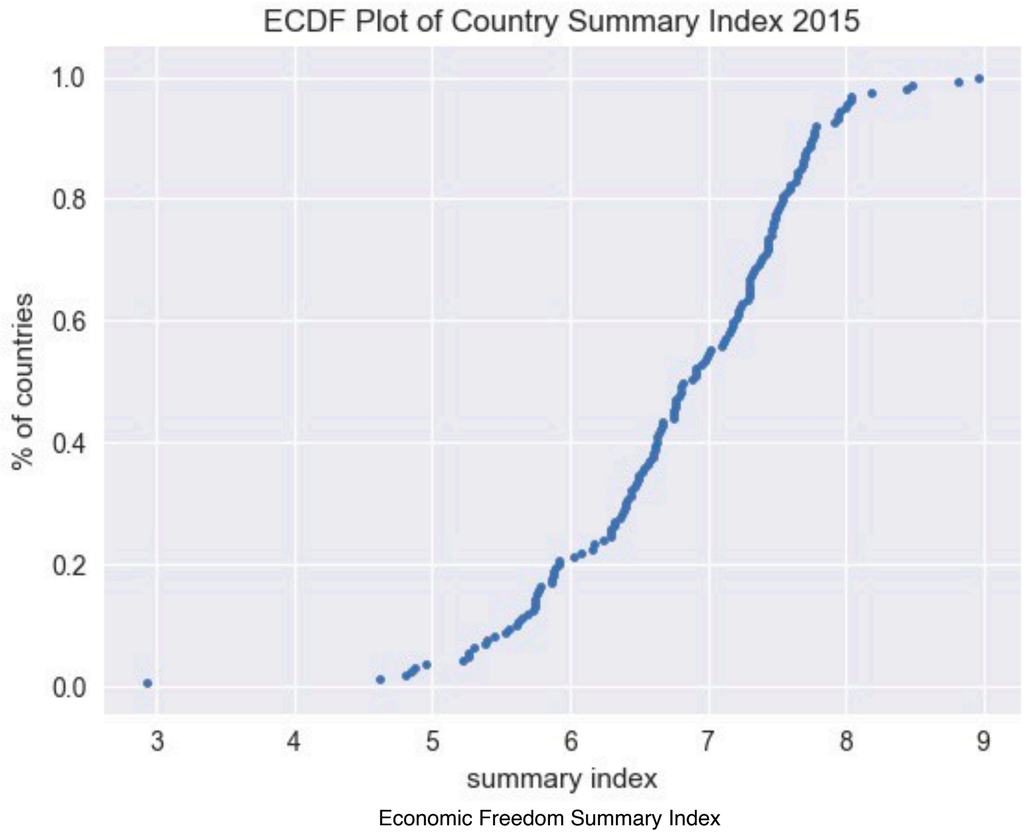




What percentage of bills are > 20?

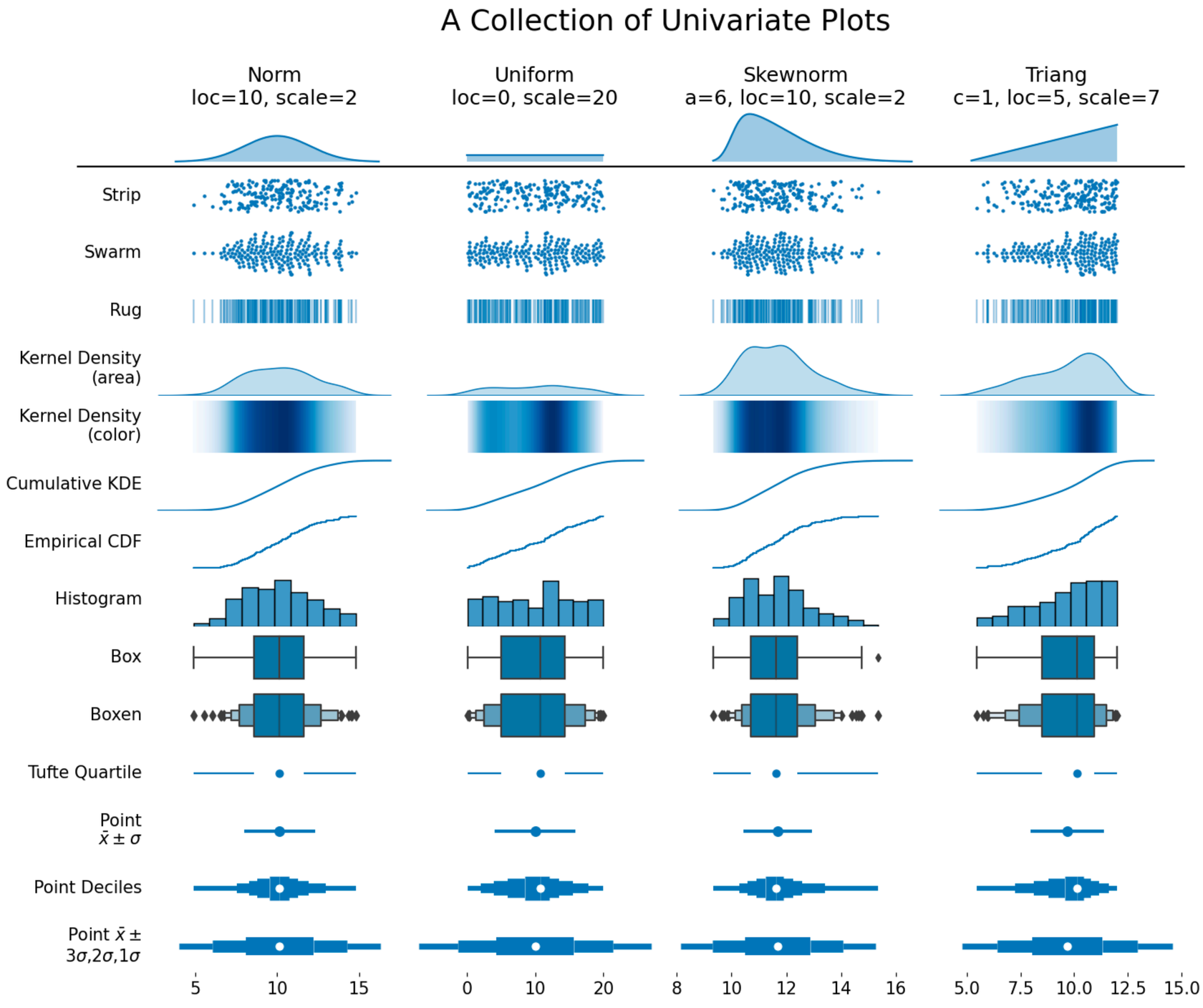
Empirical KDE

Similar to regular KDE, but observed points instead of function.

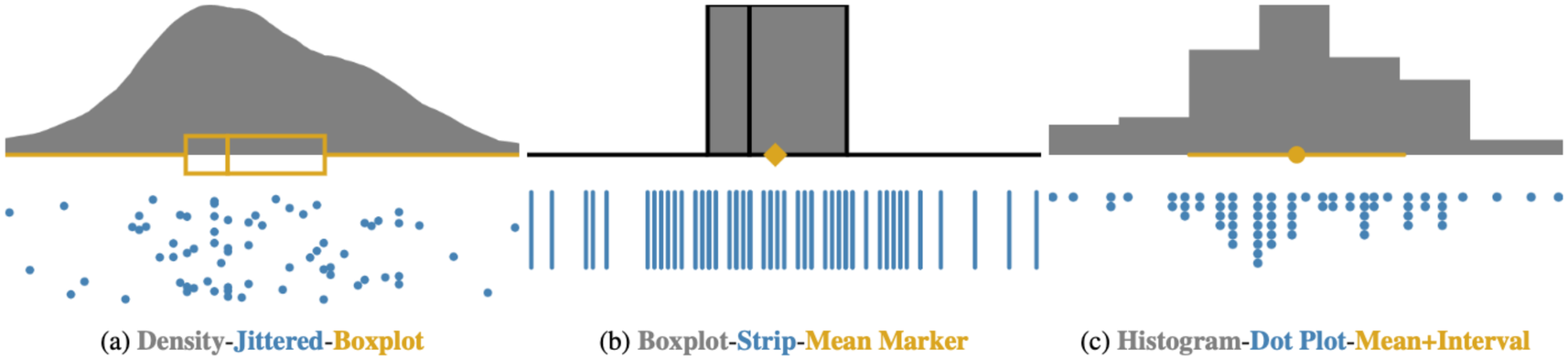


Economic Freedom Summary Index

1. What percentage of countries have a summary index less than 6?
2. What is an approximate percentage of the countries that have a summary index less than 8?


A Collection of Univariate Plots
Norm
loc=10, scale=2
Uniform
loc=0, scale=20
Skewnorm
a=6, loc=10, scale=2
Triang
c=1, loc=5, scale=7
Strip
Swarm
Rug
Kernel Density
(area)
Kernel Density
(color)
Cumulative KDE
Empirical CDF
Histogram
Box
Boxen
Tufte Quartile
Point
$\bar{x} \pm \sigma$
Point Deciles
Point $\bar{x} \pm$
$3\sigma,2\sigma,1\sigma$
5
10
15
0
10
20
8
10
12
14
16
5.0
7.5
10.0
12.5
15.0

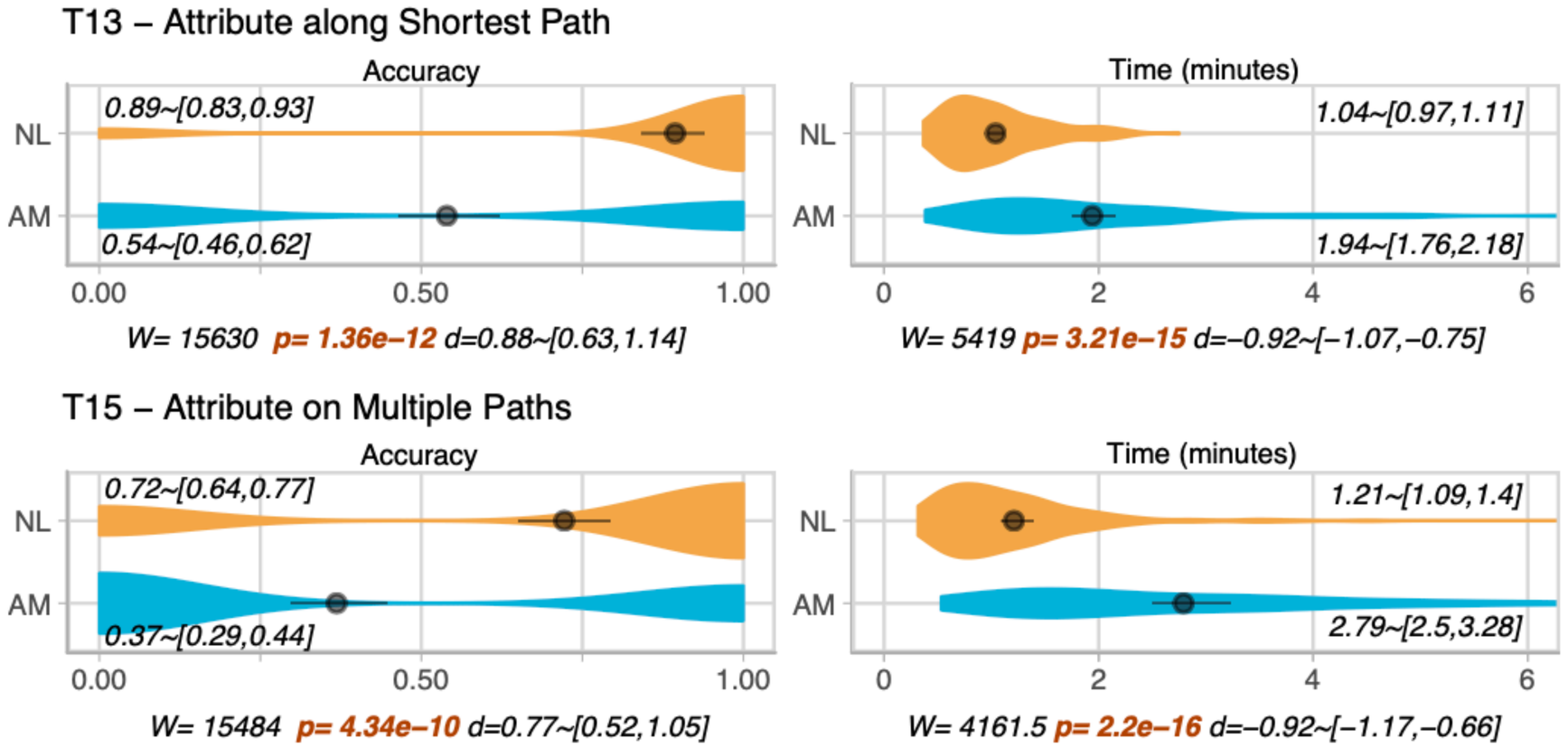
Combinations: Clouds, Rain and Lightning



Combine distribution, statistical information, and raw data

[Corell 2023]

How to Visually Report Distributions for Experiments



[Nobre 2019]

One of these things is not like the other…

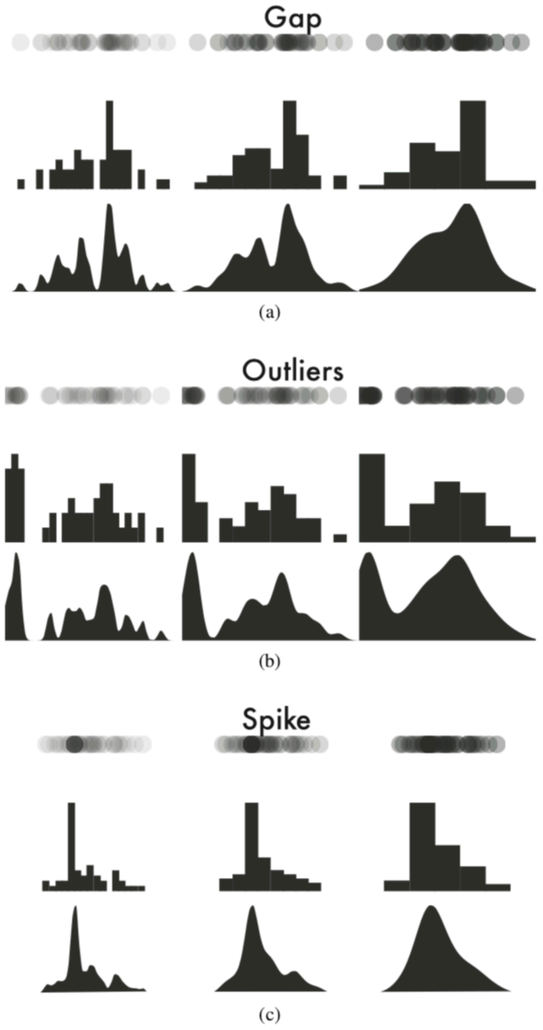
19 charts are random samples from a gaussian.
1 chart has 20% of samples with identical value

[Corell et al, InfoVis 2019]

Detecting Data Flaws

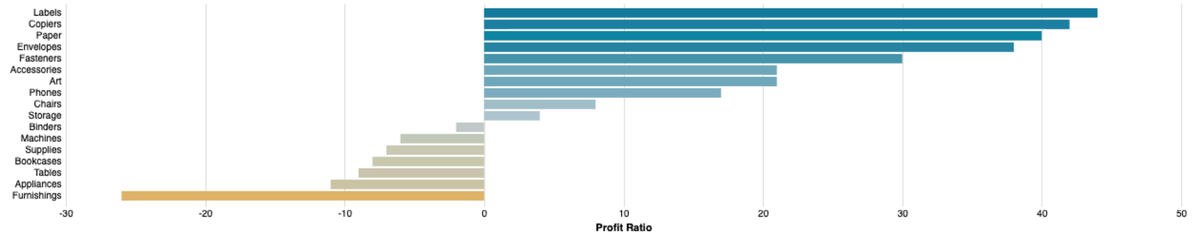
Tricky with aggregate visualization

Bin size / kernel type / bandwidth / visualization choice all affect different situations

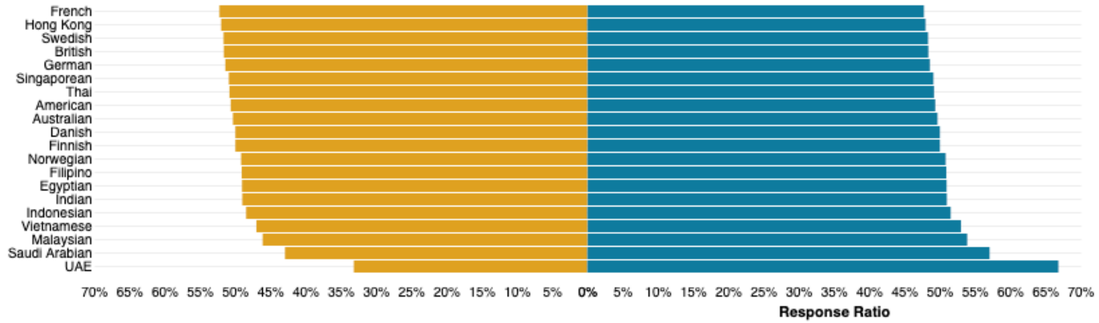


Deviation

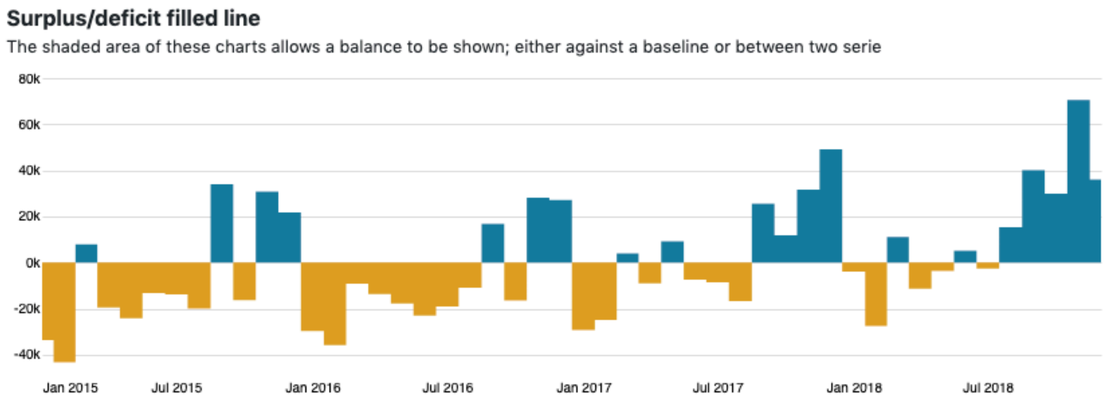
Comparison to Reference Point

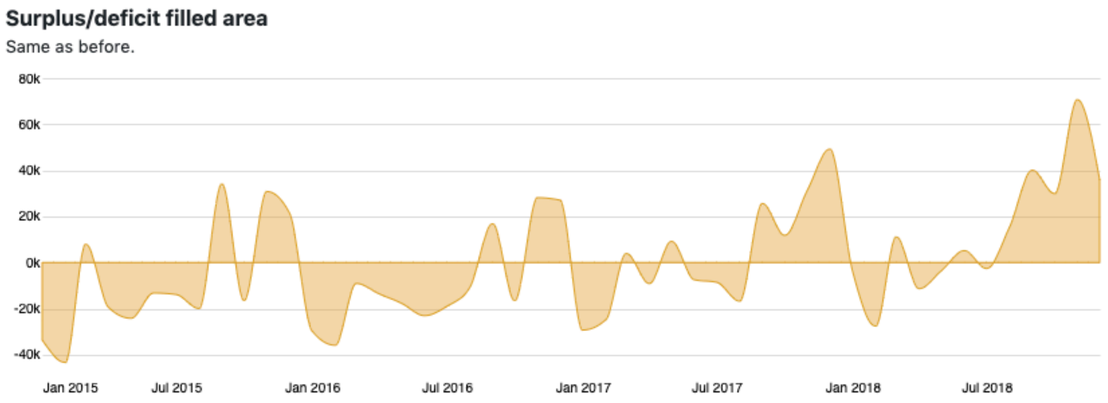


Diverging Bar Chart



Juxtaposing Two Variables (male/female)





Change over Time

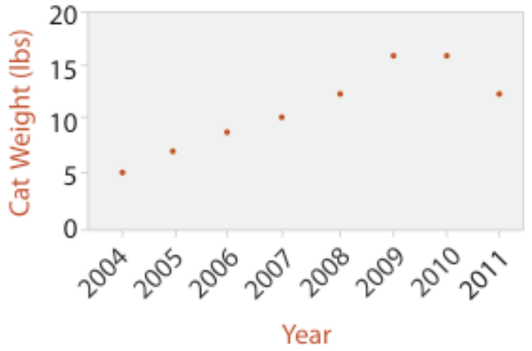
Line Chart

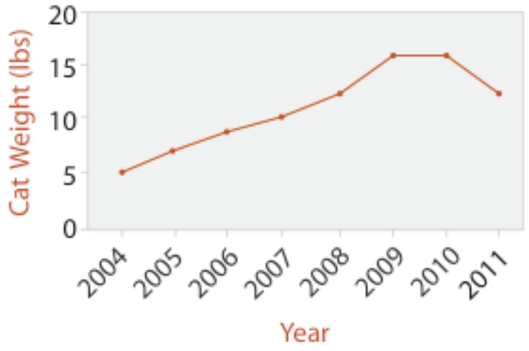
Simple

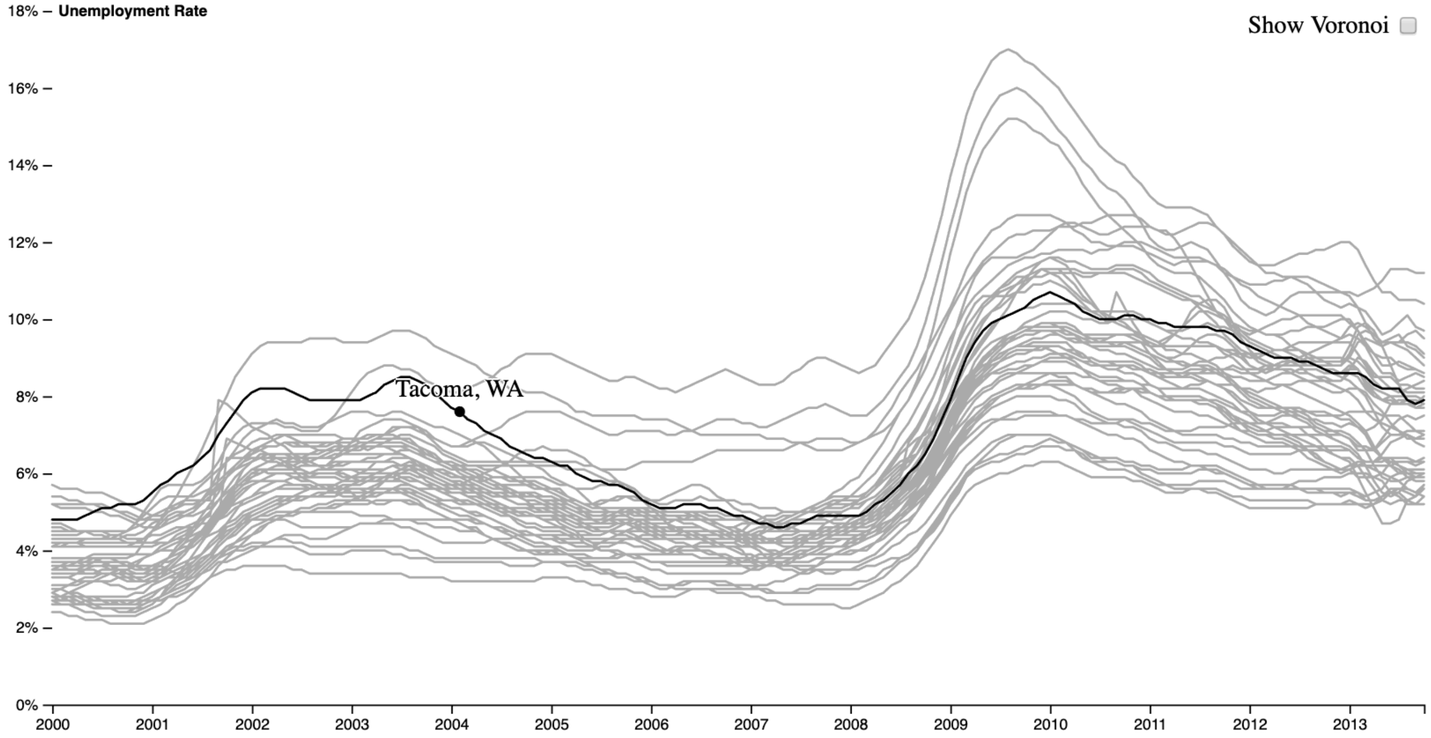
Familiar

Accurate

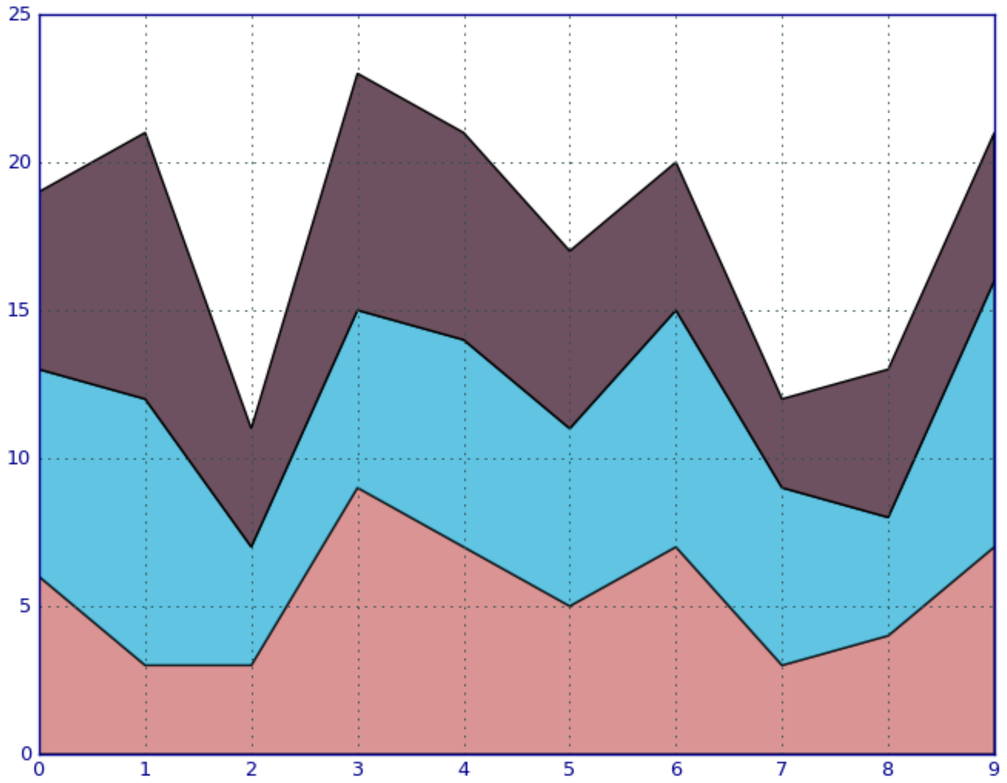
Fairly Scalable



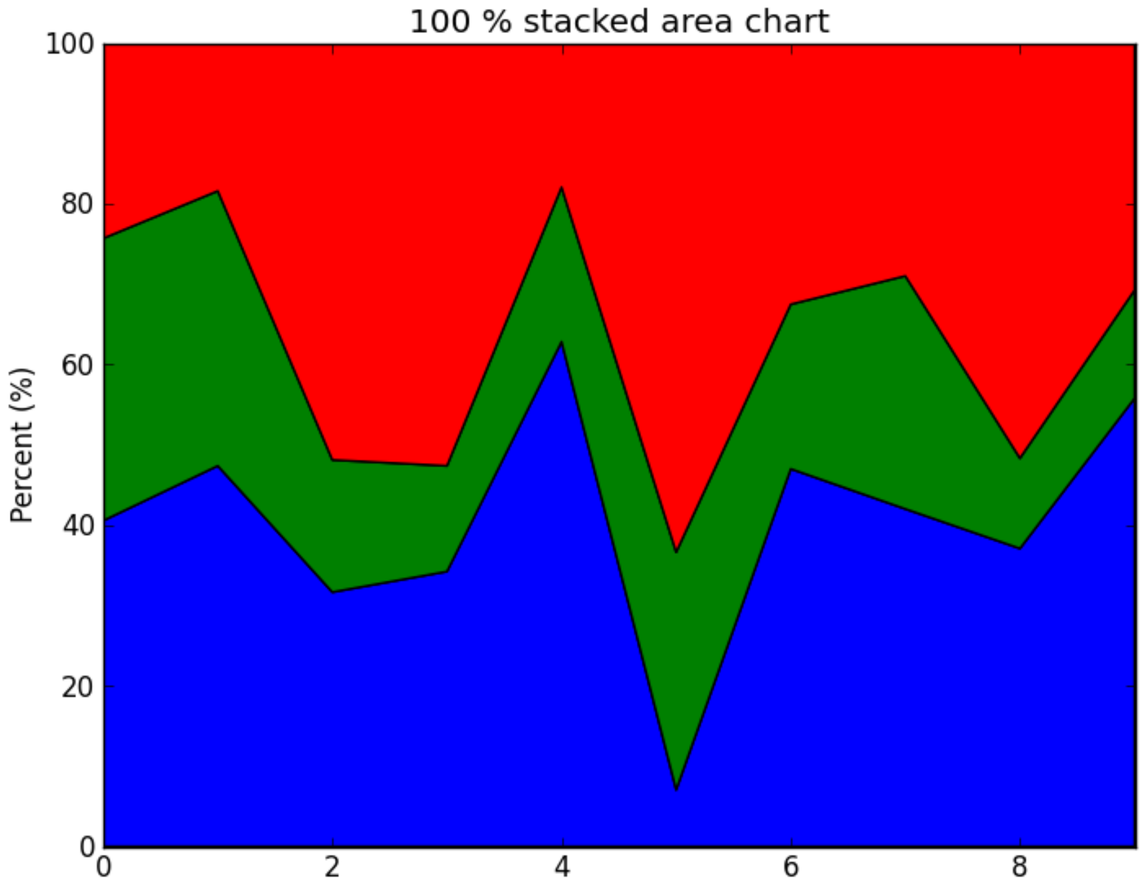




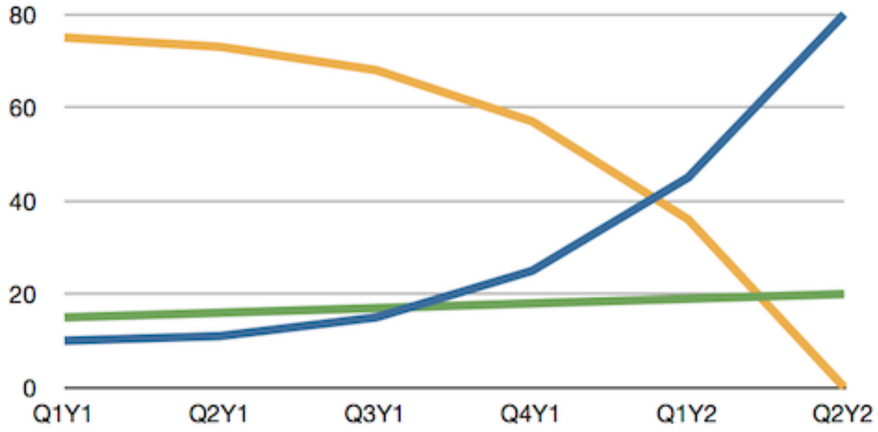
Stacked Area Chart



100% Stacked Area Chart

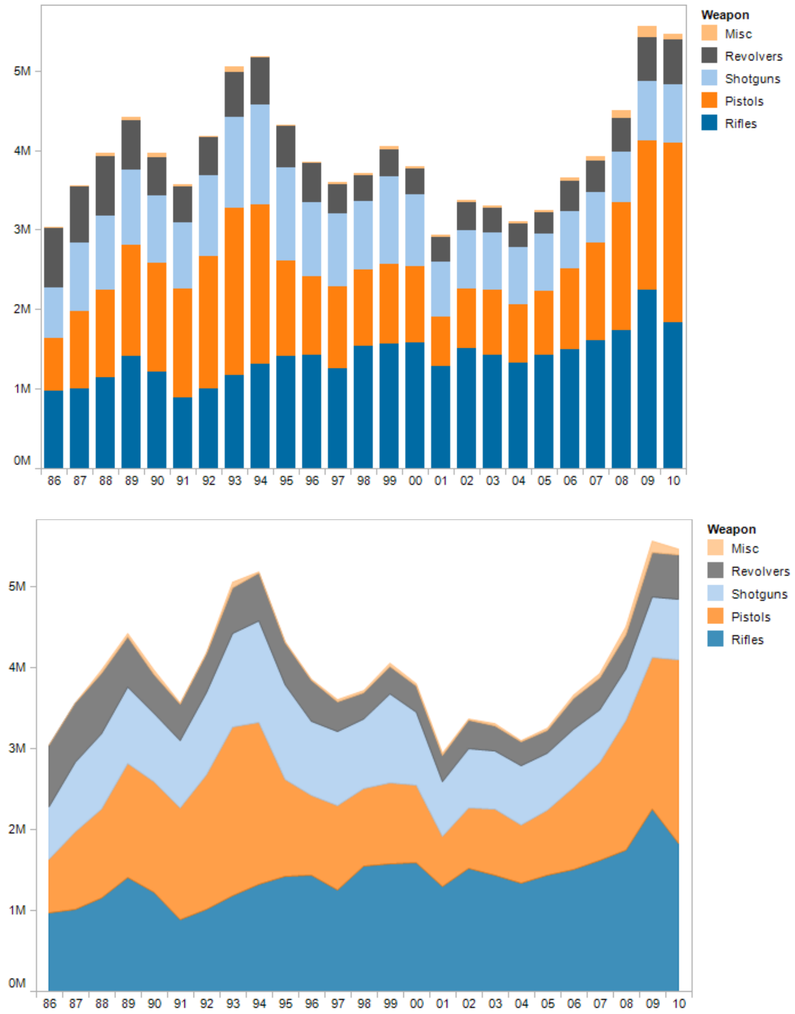


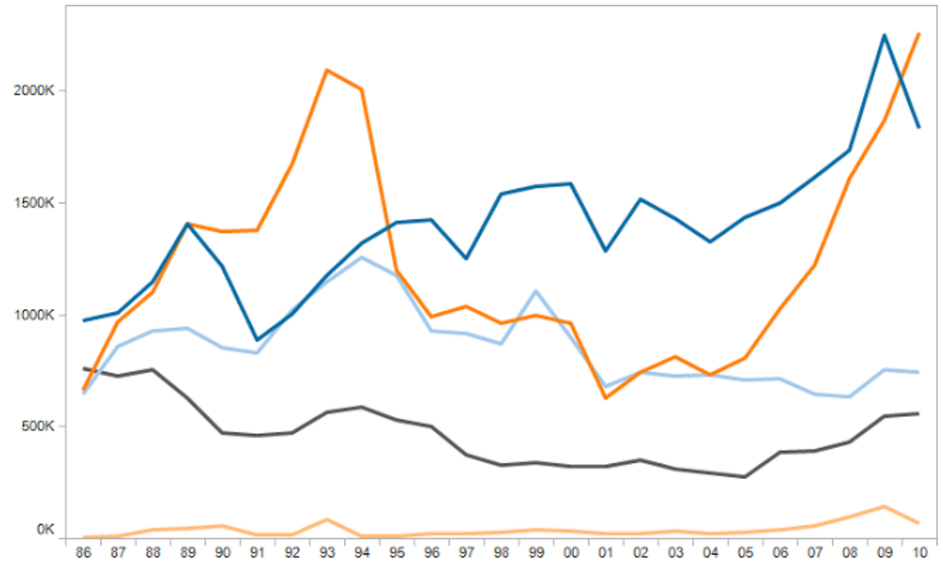
Stacked Area vs. Line Graphs



leancrew.com &
Practically Efficient

Can you spot the trends? Overall vs Individual Components.





VizWiz, A. Kriebel

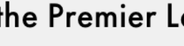
Sparklines

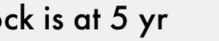
Small line charts

can be embedded in text

or part of a table

Mauricio Pochettino has lead Spurs on their best run 8TH  2ND in 24 years of the Premier League

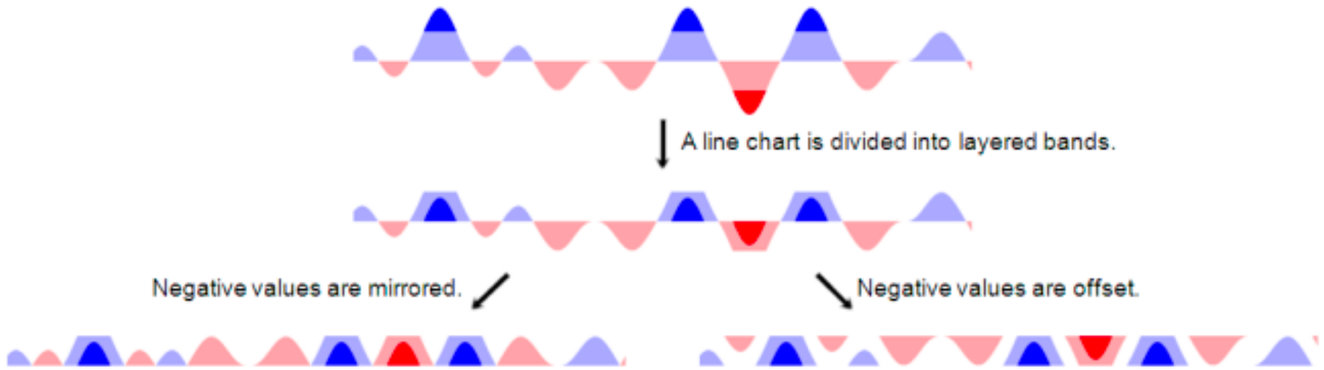
Alibaba stock is at 5 yr high 93.89  152.11 as of July 2017

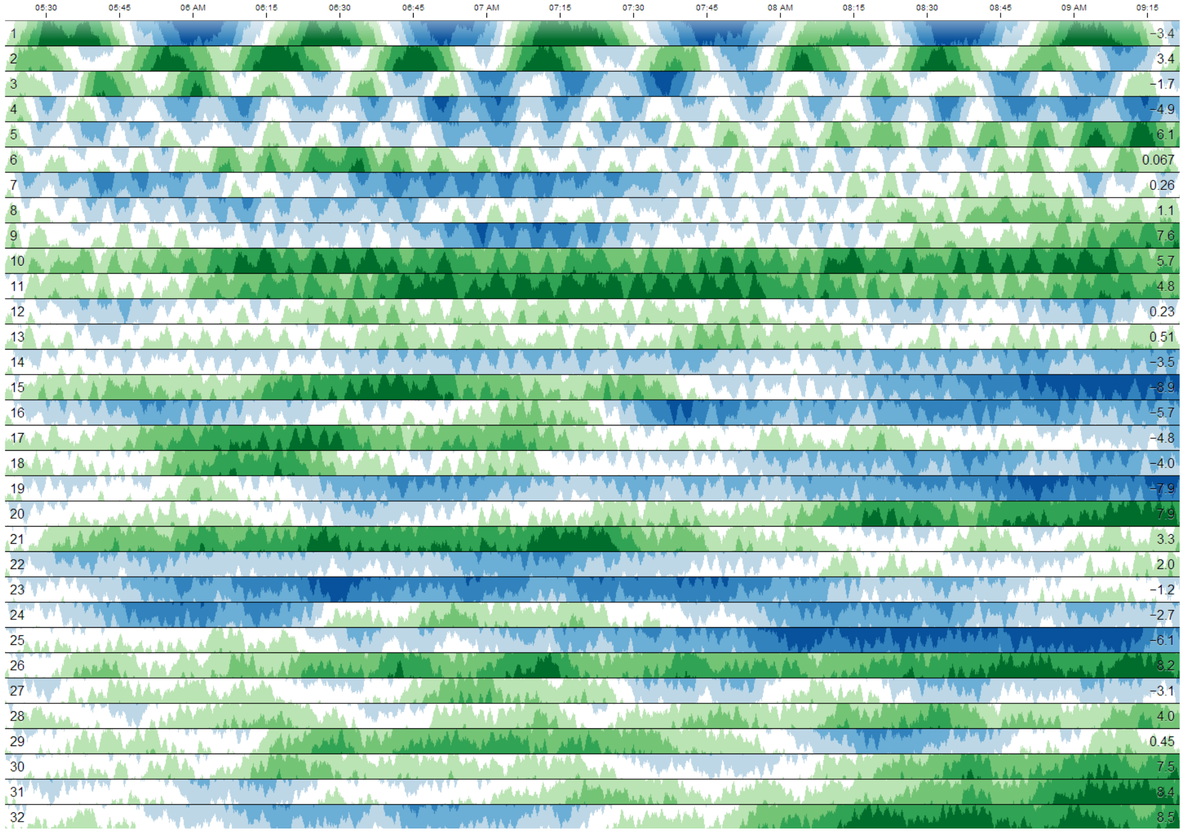
The FTSE100 Brexit bounce 5562  7501 continues one year on from the vote last summer

Symbol	Bid	Ask	Last	Change	T	Chart	Volume	High	Low	Value Change		Value	Gain	
DELL	89 3/4	89 13/16	89 3/4	+ 1 1/4			10,310,100	90 1/8	88 1/2	+1.41%	250	17,950	+273.72%	13,147
CPQ	48 7/16	48 9/16	48 7/16	- 13/16			25,628,700	51 1/4	1/4	-1.65%	-81	4,844	+60.79%	1,831
SDTI	26 1/4	26 3/8	26 3/8	+ 1/2			504,600	27 3/8	25 5/8	+1.93%	250	13,188	+133.15%	7,531
COMS	46 1/2	46 9/16	46 9/16	- 25/32			3,191,100	47 15/16	45 3/4	-1.65%	-102	6,053	+29.79%	1,389
LU	111 5/8	111 11/16	111 9/16	+ 1 9/16			5,104,600	112 5/8	110	+1.42%	78	5,578	+22.76%	1,034
YHOO	368 1/16	368 1/2	368 1/2	+ 17 1/4			3,787,800	381 3/16	280	+4.91%	431	9,213	-0.41%	-38
AOL	162 13/16	163	163	+ 8			10,008,500	164	158 1/2	+5.16%	280	5,705	+73.06%	2,408
CMGI	97 3/8	97 1/2	97 1/2	+ 5 7/8			1,323,800	98 1/2	93	+6.41%	705	11,700	+186.76%	7,620
SPLN	33 13/16	33 15/16	33 13/16	+ 7/16			300,200	34 3/4	33 5/8	+1.31%	88	6,763	+94.60%	3,288
BEAS	13 1/2	13 5/8	13 5/8	- 7/16			389,200	14 1/4	13 1/8	-3.11%	-44	1,363	-9.17%	-138
GNET	102	103 3/16	101 5/16	+ 6 1/8			307,600	108	97	+6.43%	613	10,131	+130.26%	5,731
RNWK	67	67 1/4	67	+ 2 3/4			1,233,900	69	64 15/16	+4.28%	275	6,700	+79.87%	2,975
MSFT	173 1/8	173 1/4	173 5/16	+ 1 3/4			13,284,500	174 7/16	170	+1.02%	175	17,331	+54.74%	6,131
INTC	133 3/4	133 13/16	133 13/16	- 3 1/8			8,094,300	137 1/2	133 3/8	-2.28%	-625	26,763	+65.20%	10,563
TOTAL								205,302	80,993	+1.63%	2,293	143,280	+79.41%	63,377

https://www.bram.us/2017/09/12/spark-a-typeface-for-creating-sparklines-in-text-without-code/

Horizon Graphs





http://square.github.io/cubism/

[Heer, Sizing the Horizon, 2009]

Horizon Chart Explanation

A Horizon Chart is a specialized type of chart for time series data. It is especially useful for showing data with large amplitudes in a short vertical space. The idea was introduced by Saito et al. in Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data. Panopticon commercialized and coined the term Horizon Chart. Like any novel visualization, one downside is the cost for your audience to learn and understand that chart. Therefore, I have built this interactive visualization to help make it easier to understand how Horizon Charts work.

Select Function: y = sin(x)

Mirror Negative Values

Include Bin Lines

Mod Height: 0.25

Baseline: 0

Container Width: 400

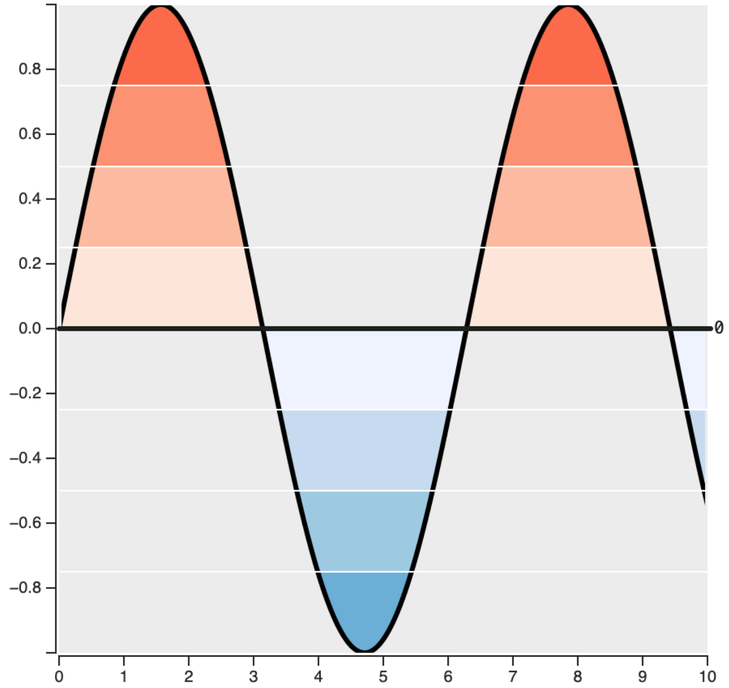
Row Height: 40

Match Row Height

Horizon Chart

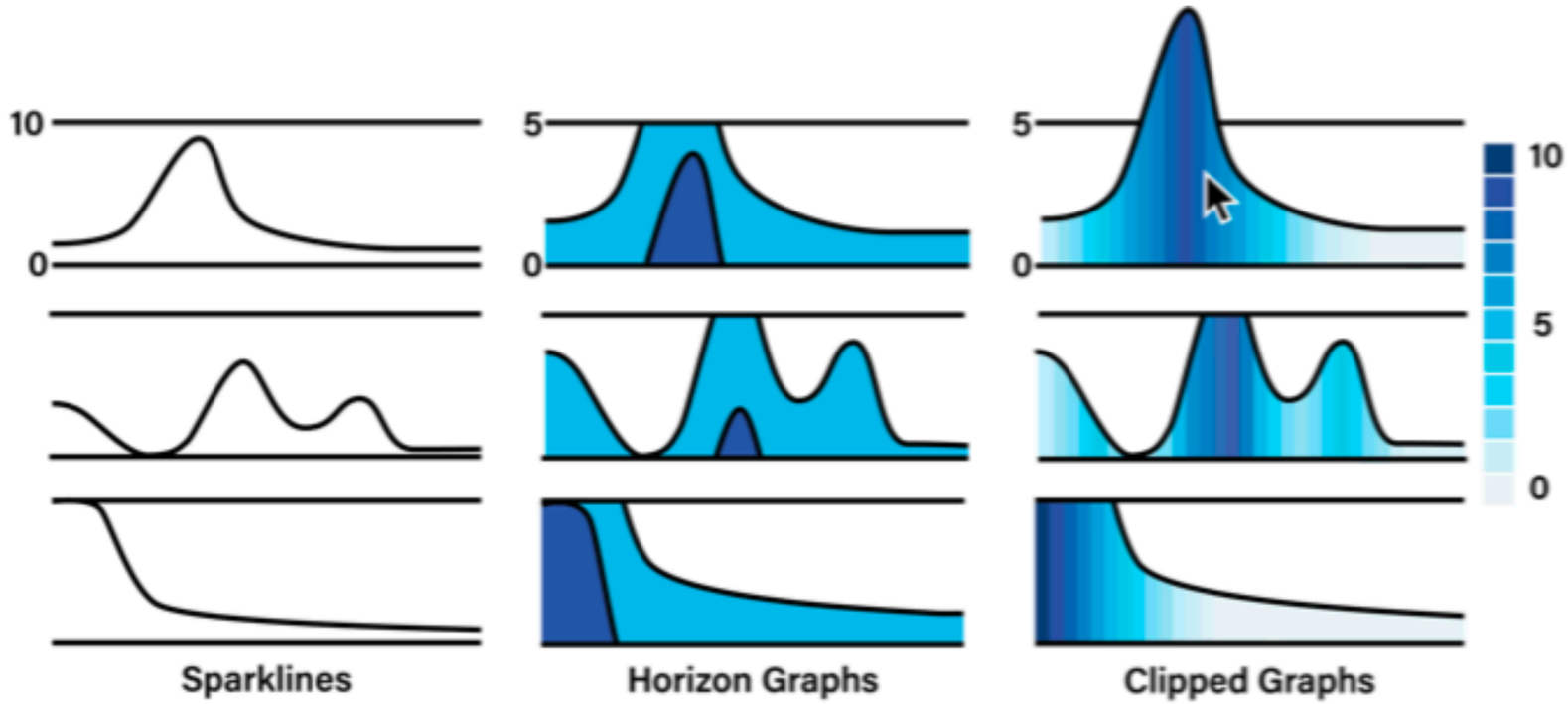
Press and hold to stack!

Explanation Chart



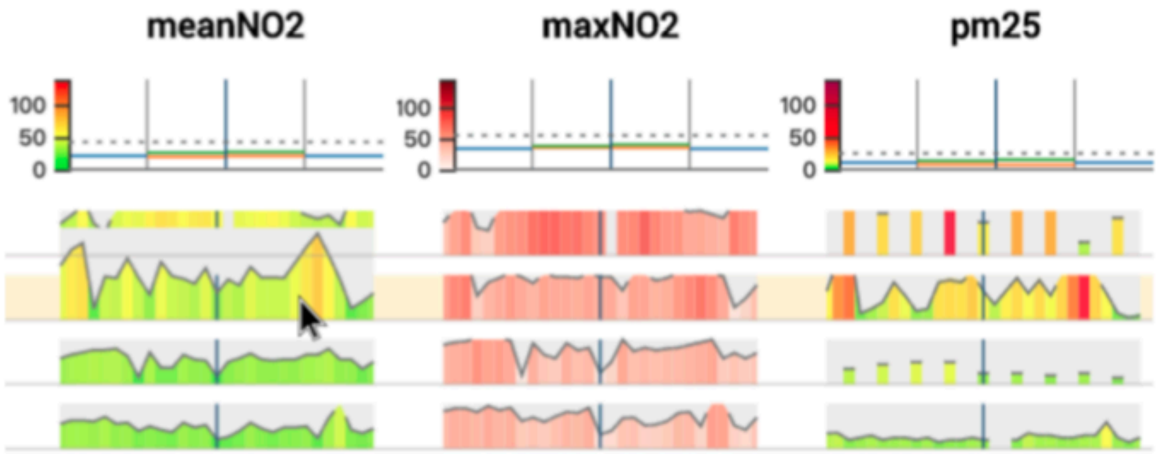
http://www.horizon-chart-explanation.devinlange.com/

Clipped Graphs



[Lin 2019]

Clipped Graphs

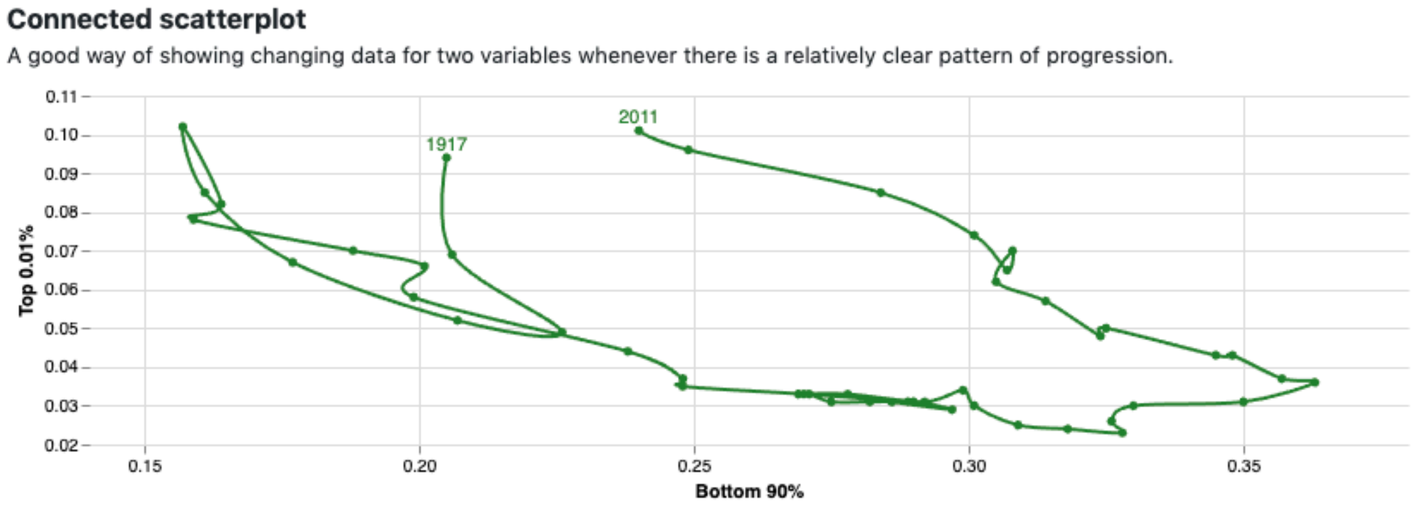


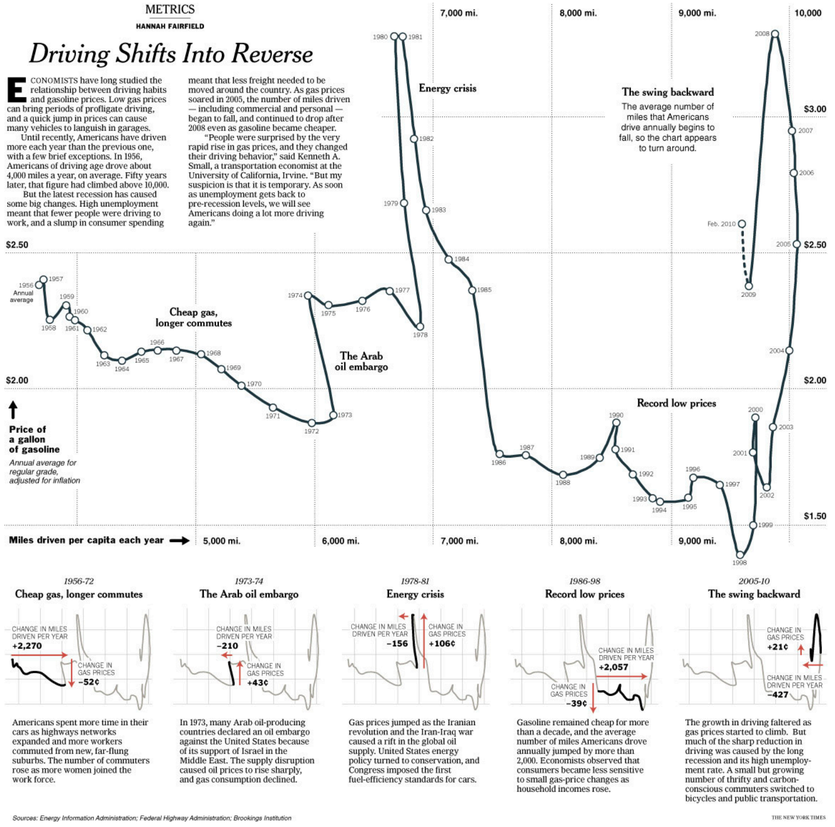
Connected Scatterplot

Two Variables + Time

Only one per Chart!

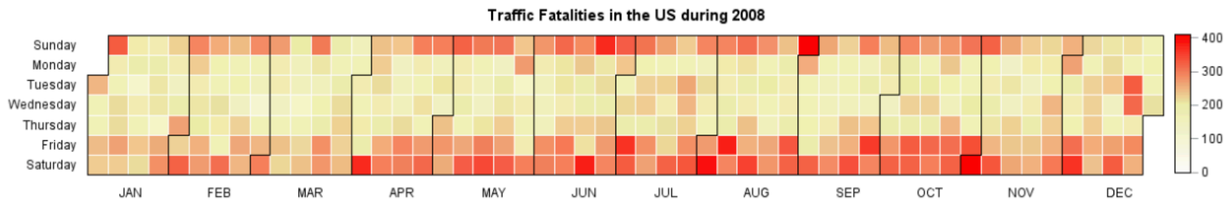
Labels important





http://www.thefunctionalart.com/2012/09/in-praise-of-connected-scatter-plots.html

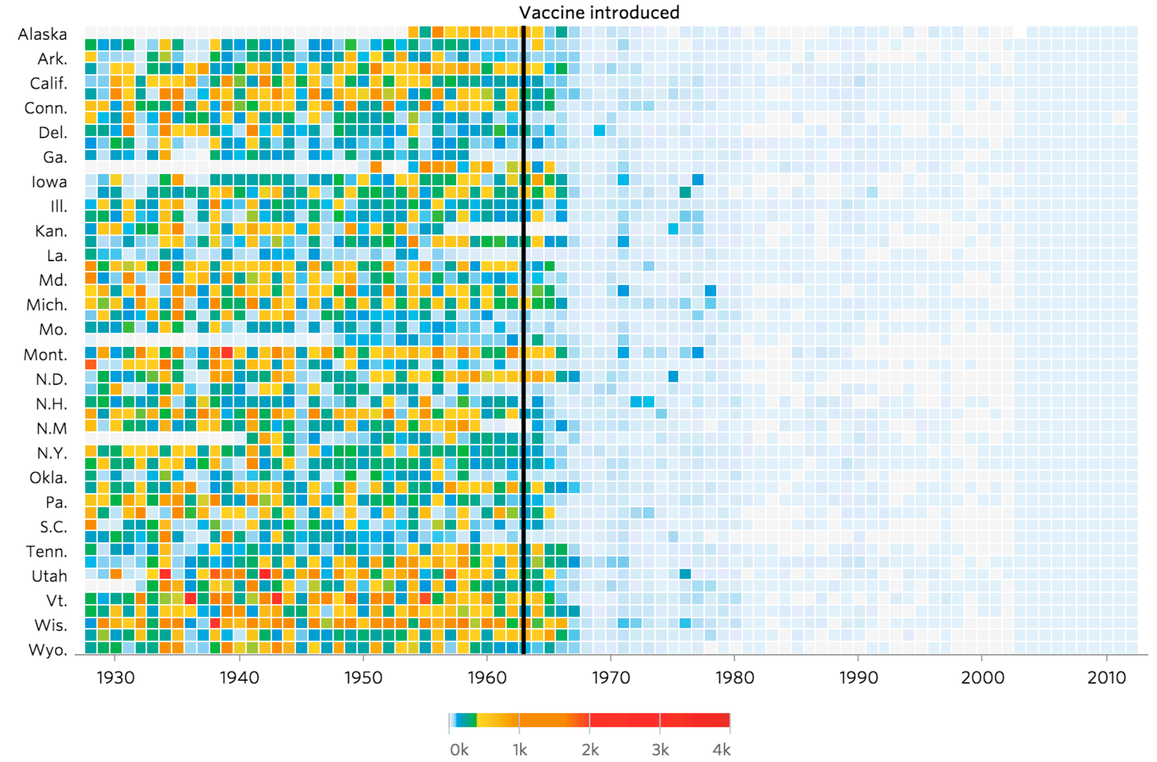
Heat Map and Calendar Heat Map



https://blogs.sas.com/content/graphicallyspeaking/2011/12/08/calendar-heatmaps-in-gtl/

The heat maps below show number of cases per 100,000 people.

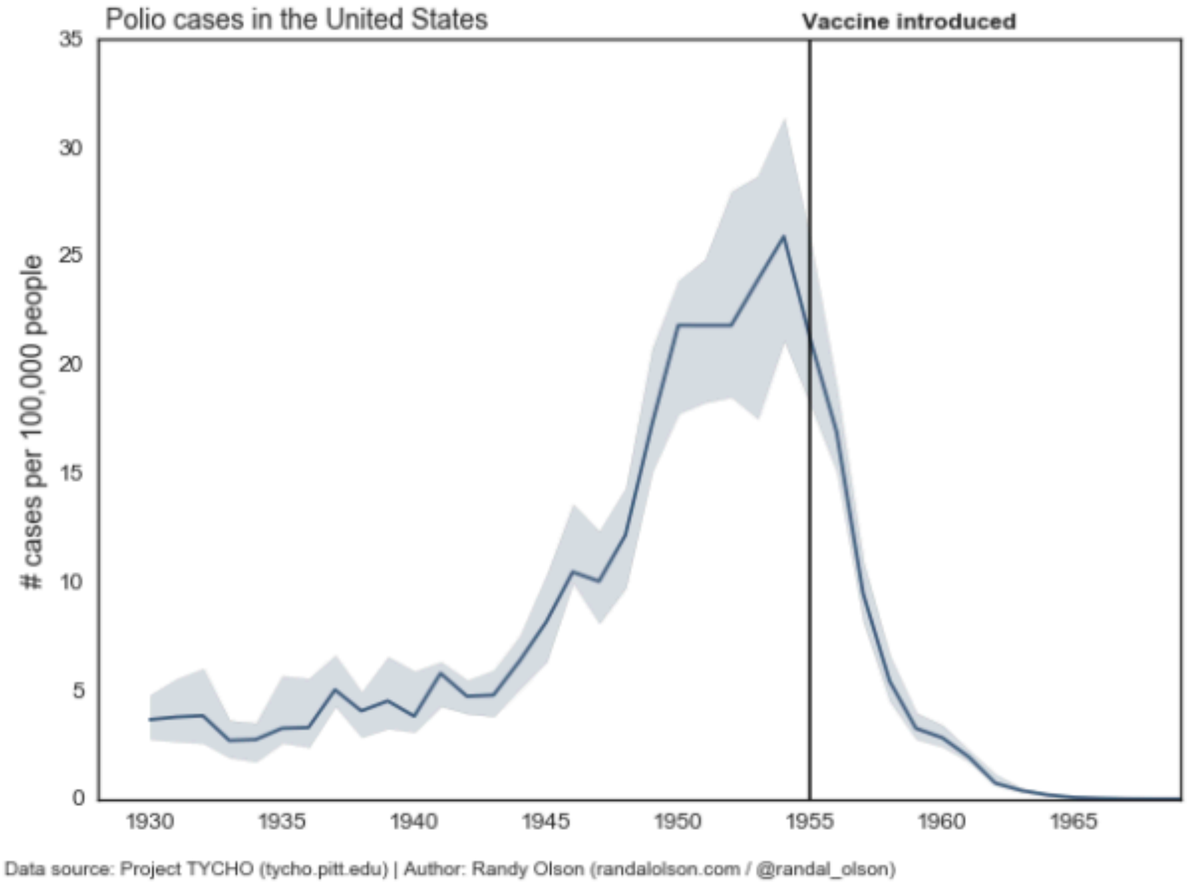
Measles



Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

https://www.informationisbeautifulawards.com/showcase/660-vaccines-and-infectious-diseases

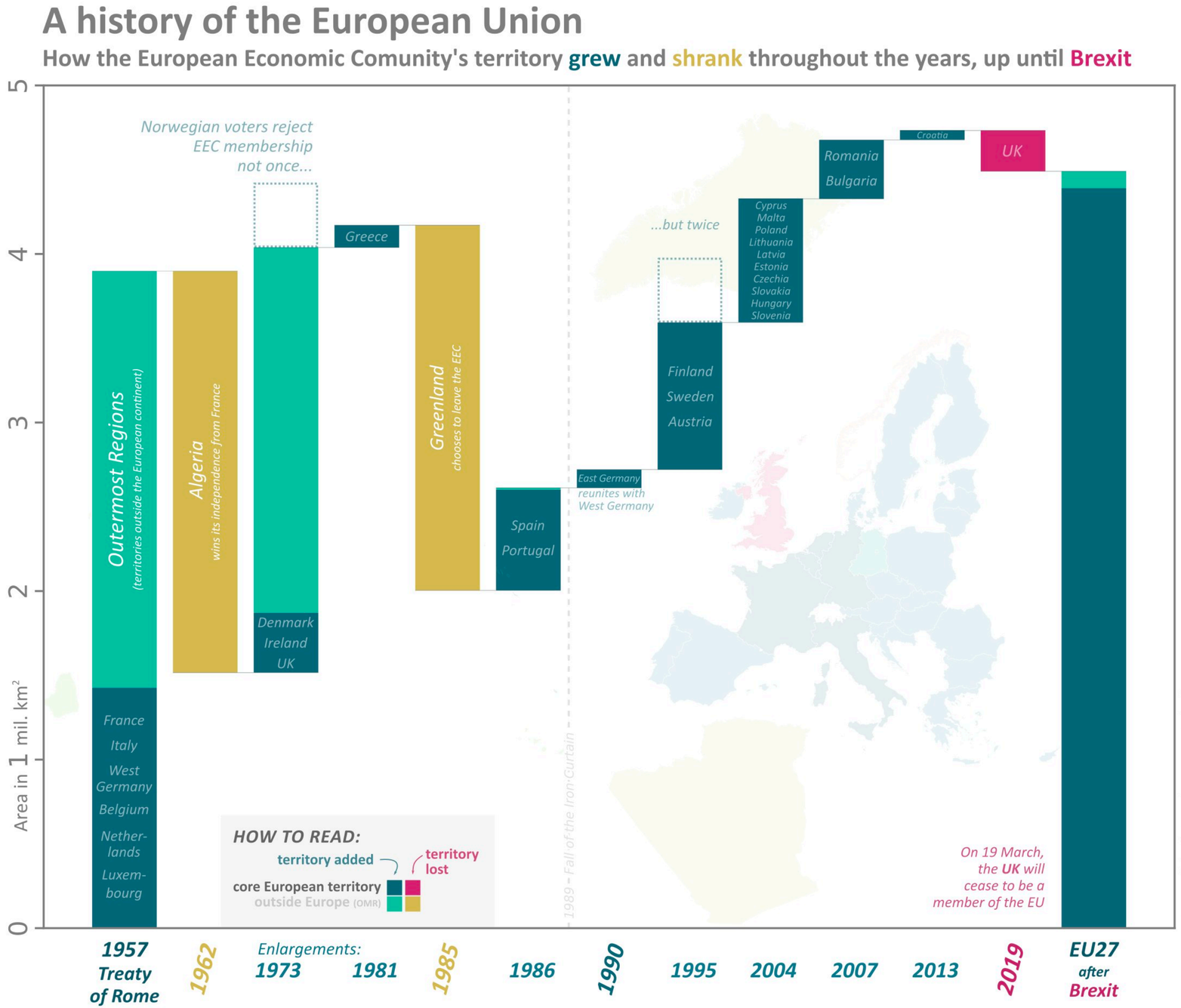
Sometimes you can Show Too Much Data



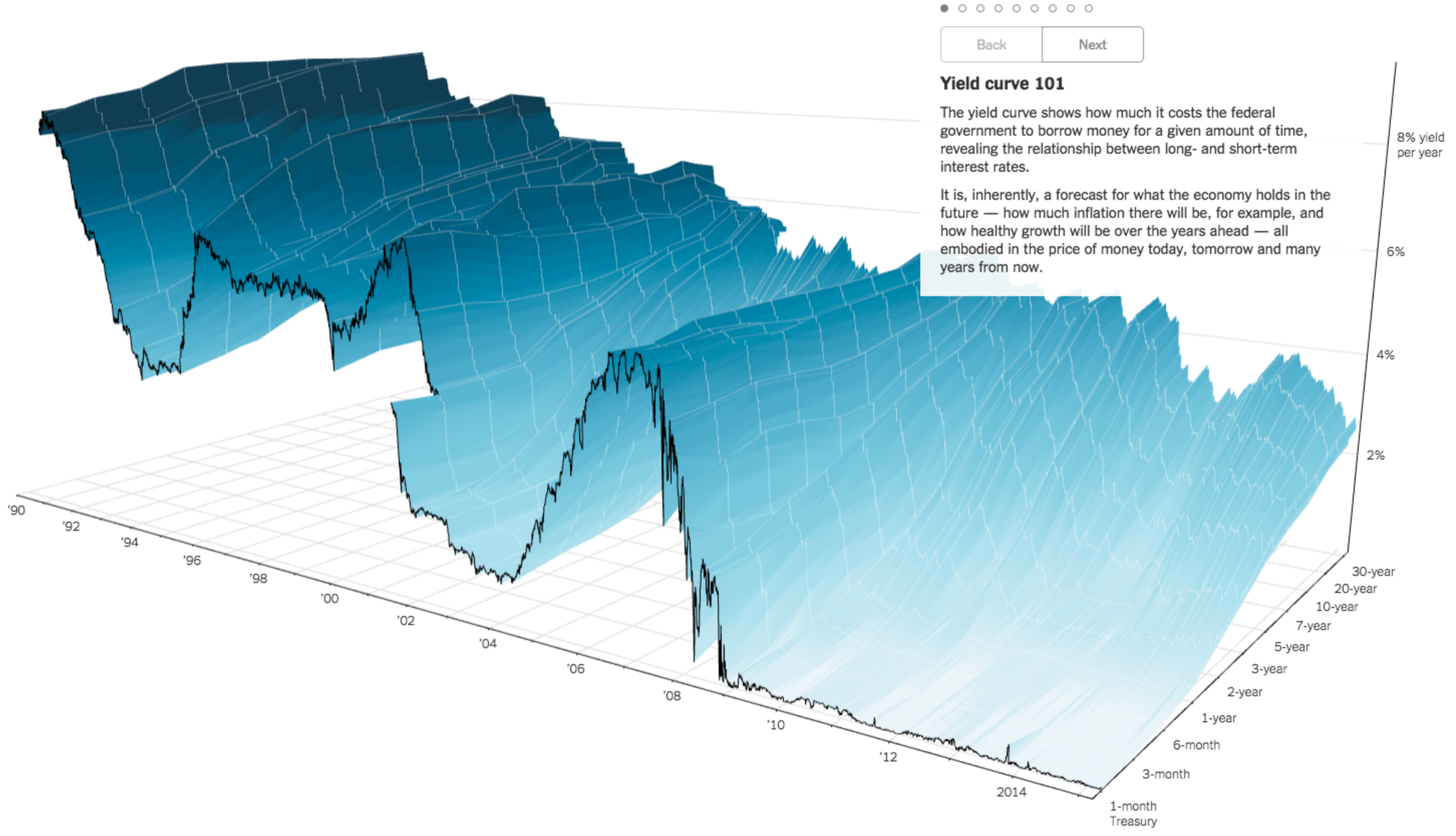
http://www.randalolson.com/2016/03/04/revisiting-the-vaccine-visualizations/

Waterfall Chart

Great way to show evolution of part of whole over time / events (non-linear time)



Design Critique



Document: https://goo.gl/W6w0iI
Website: http://goo.gl/D3mIsy

Context / Critiques

https://vimeo.com/127205447

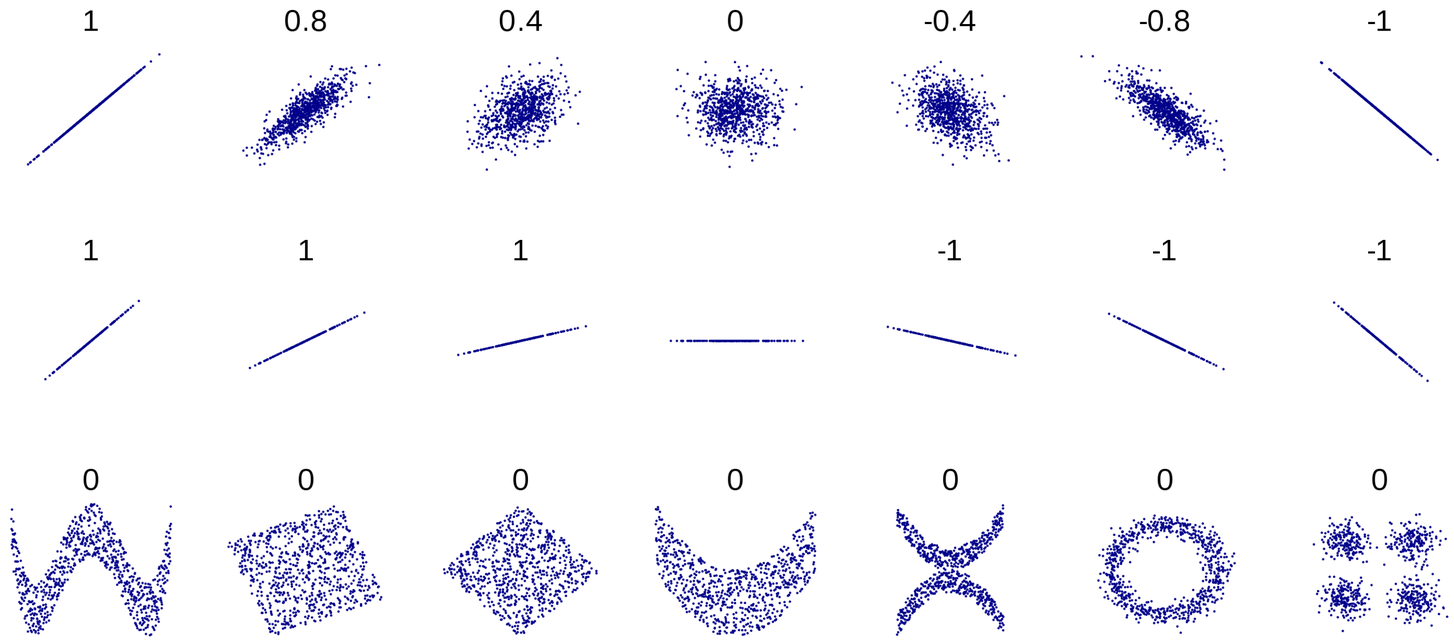
https://community.jmp.com/t5/JMP-Blog/Graph-makeover-3-D-yield-curve-surface/ba-p/30573

http://www.visualisingdata.com/2015/03/when-3d-works/

Correlation

What is Correlation

How do two or more variables behave relative to each other?



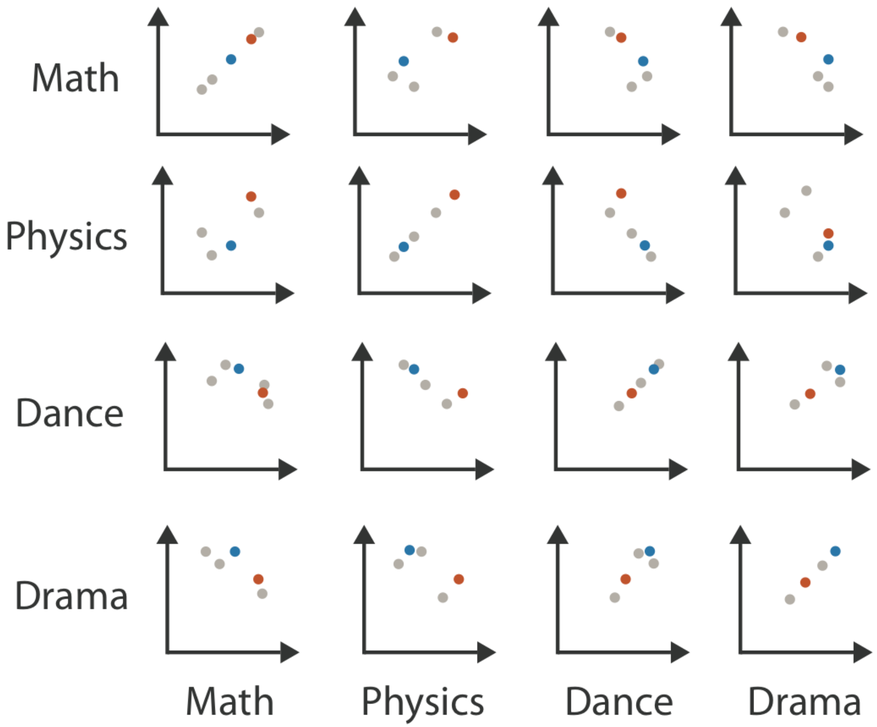
By DenisBoigelot, CC0, https://commons.wikimedia.org/w/index.php?curid=15165296

Axis-Based Techniques

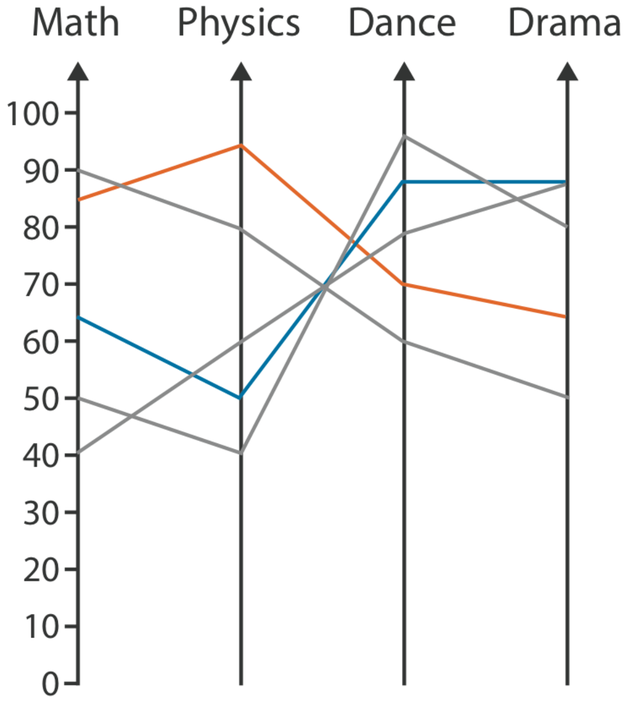
Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatterplot Matrix



Parallel Coordinates



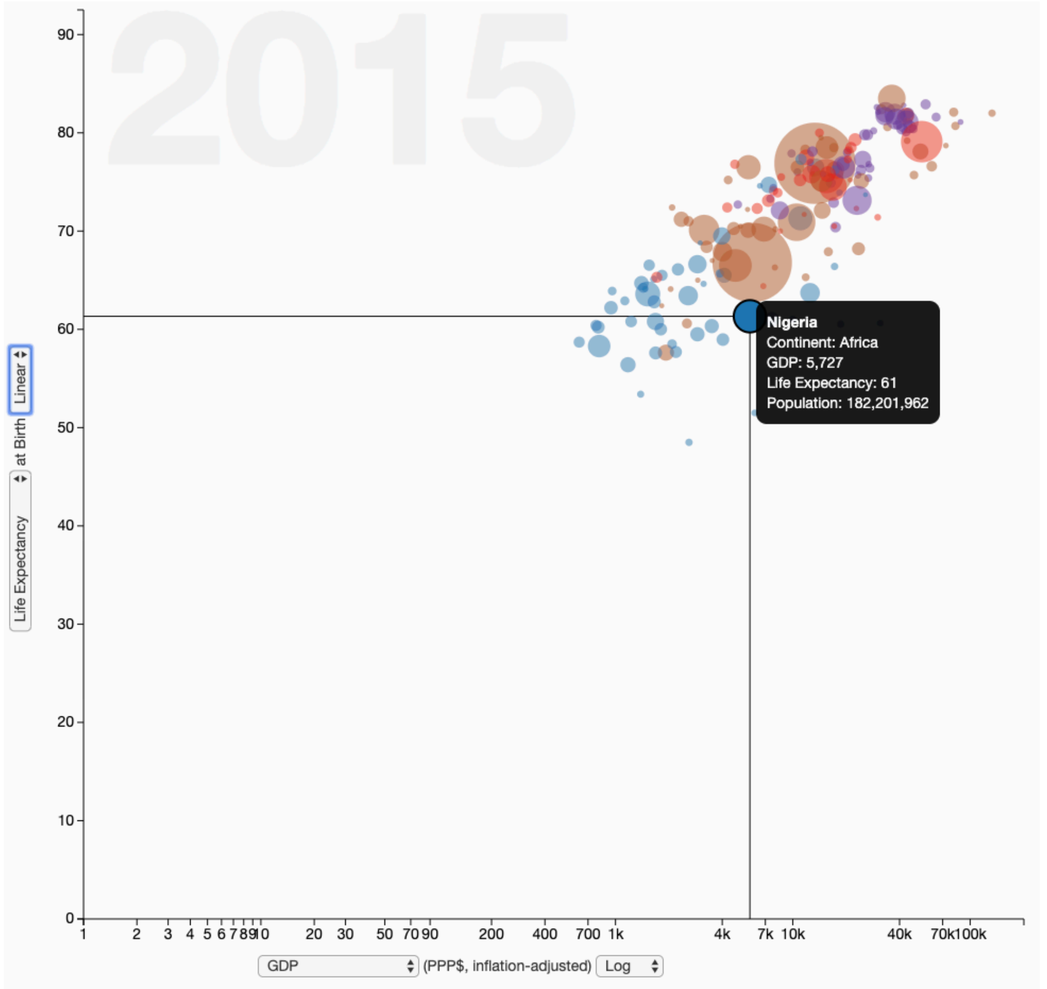
Scatterplots

Scatterplots

Two orthogonal axis visualizing one dimension each.

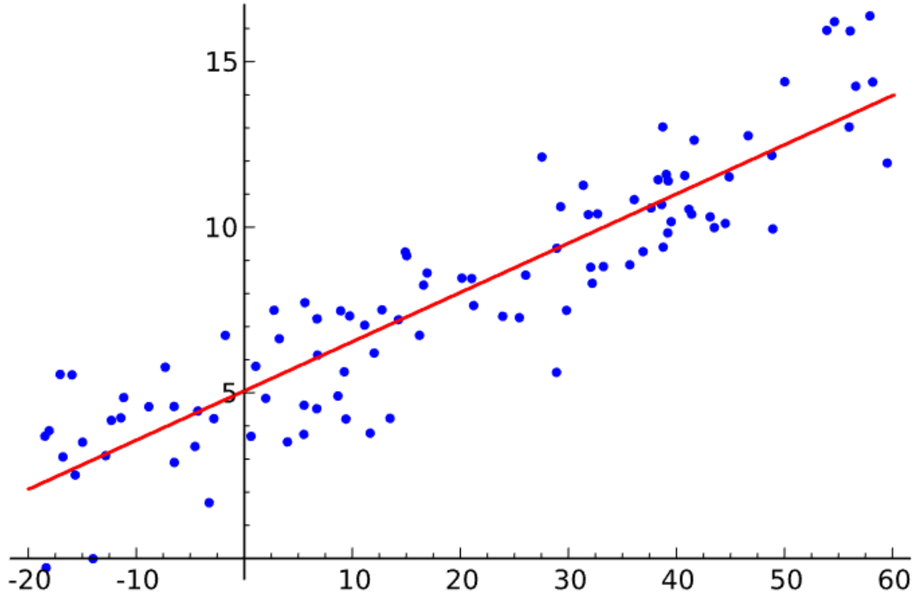
How to encode the mark?

How to deal with many points?



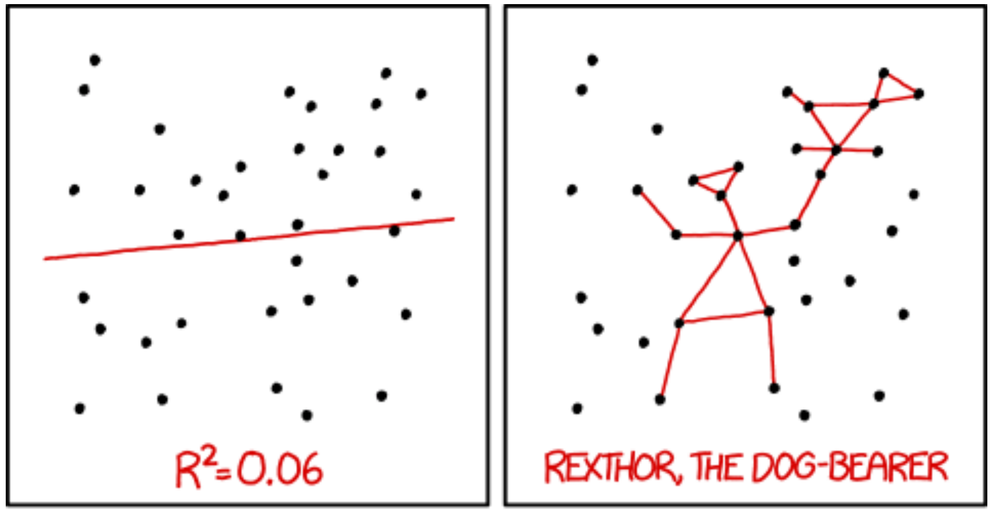
Regression Lines

$$y \sim \beta_0 + \beta_1 x$$

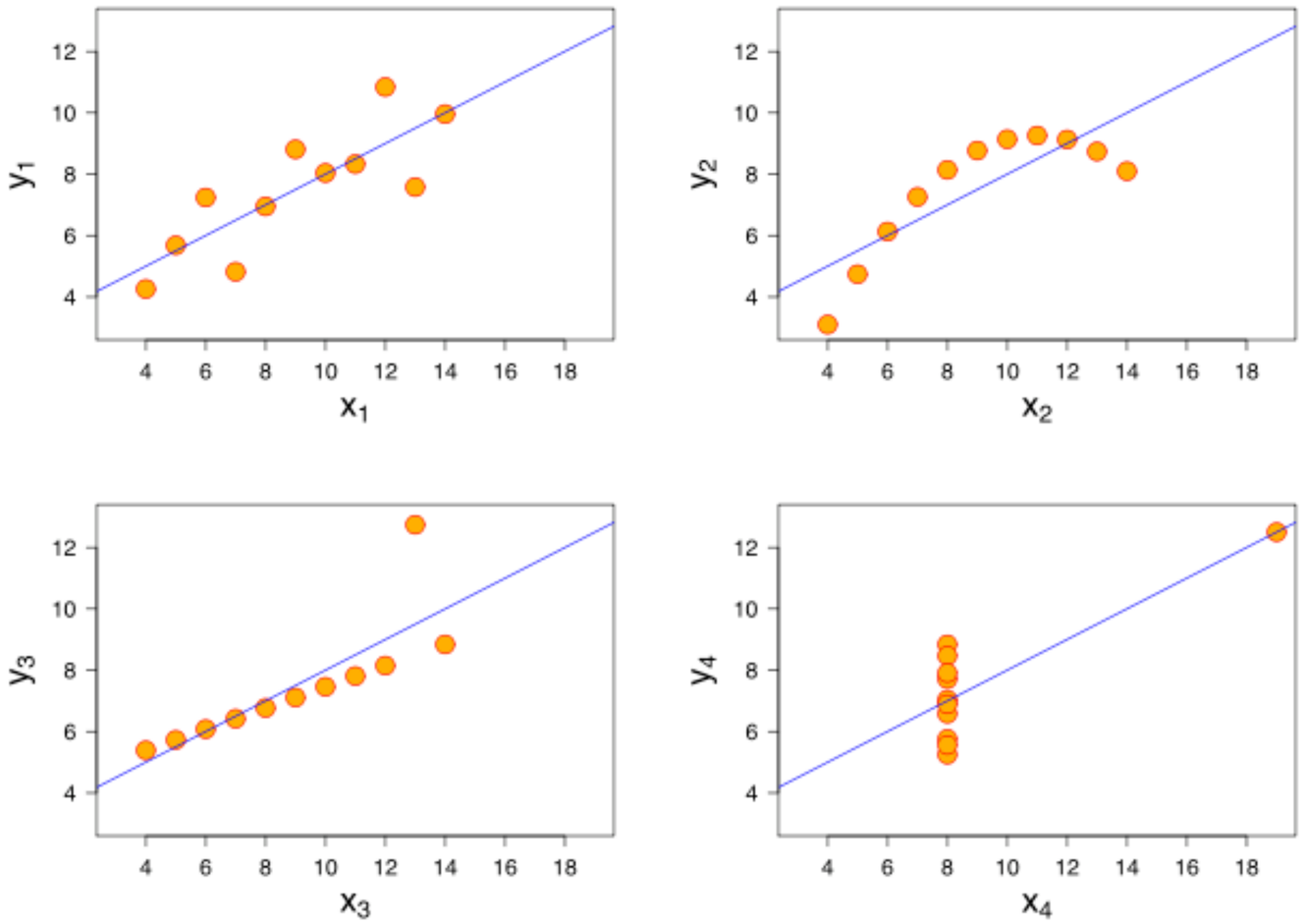


Goal: Find the best values of β_0 and β_1, denoted $\hat{\beta}_0$ and $\hat{\beta}_1$, so that the prediction $y = \hat{\beta}_0 + \hat{\beta}_1 x$ "best fits" the data.

Approach: use least squares to minimize the sum of the squares of the errors



Anscombe's Quartet

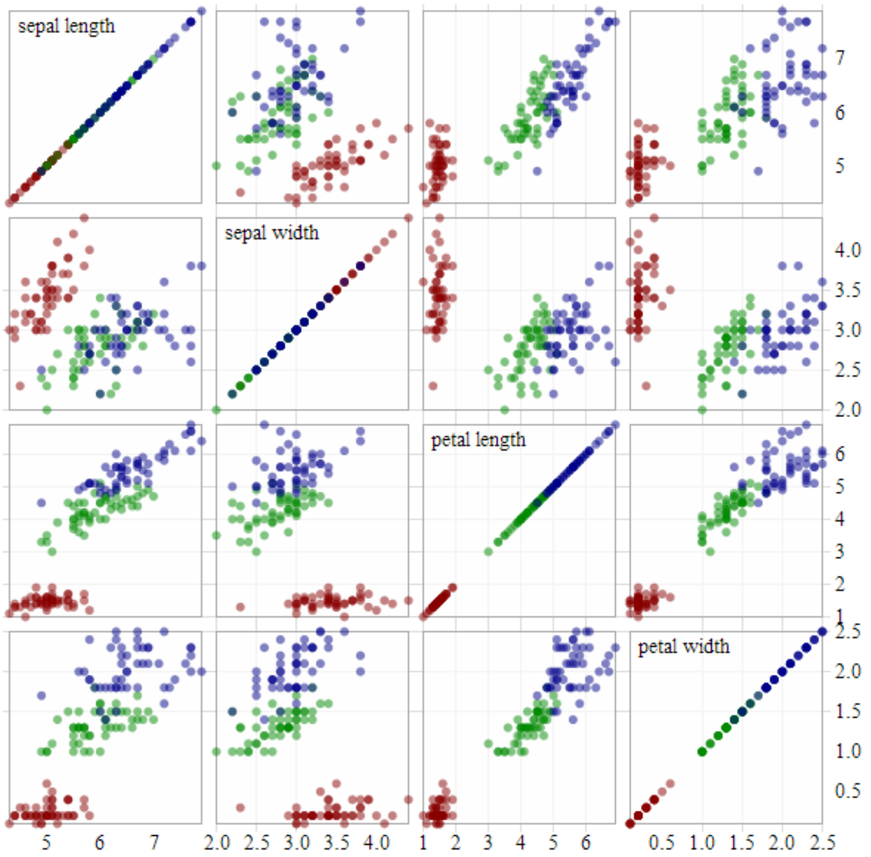


Scatterplot Matrices (SPLOM)

Matrix of size d*d

Each row/column is one dimension

Each cell plots a scatterplot of two dimensions



Scatterplot Matrices

Limited scalability (~20 dimensions, ~500-1k records)

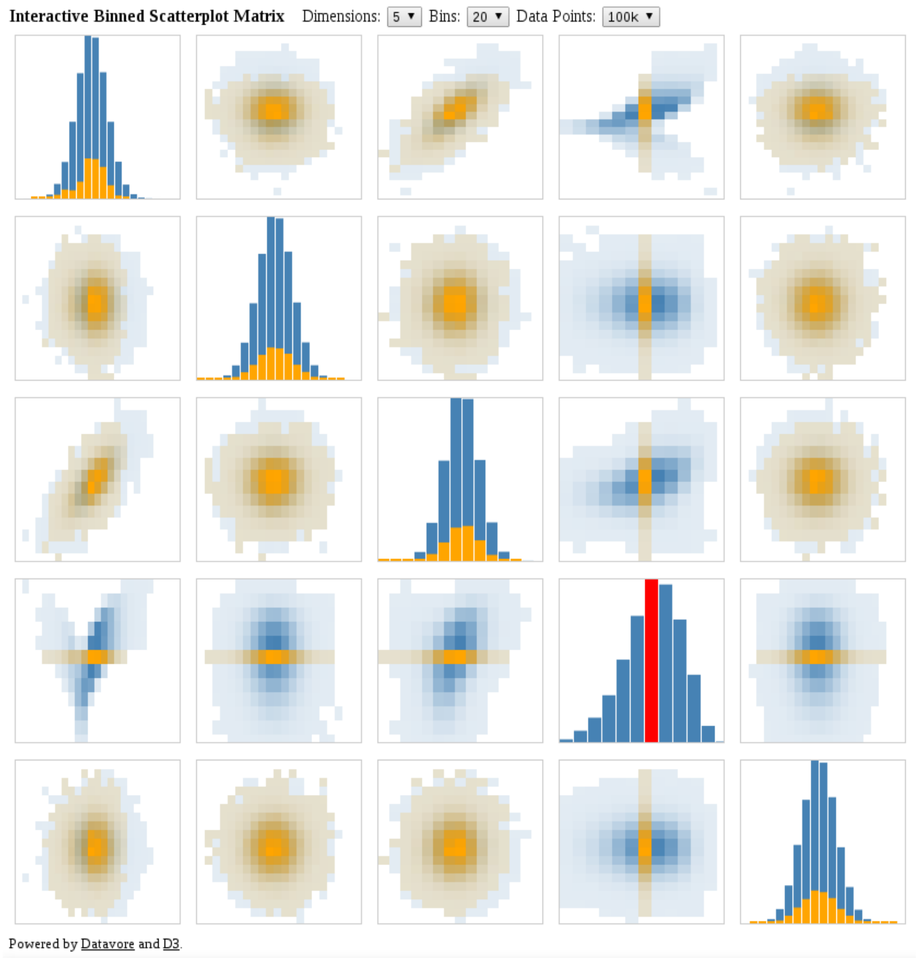
Brushing is important

Often combined with “Focus Scatterplot” as F+C technique

Algorithmic approaches:

- Clustering & aggregating records
- Choosing dimensions
- Choosing order

SPLOM Aggregation - Heat Map



Datavore: http://vis.stanford.edu/projects/datavore/splom/

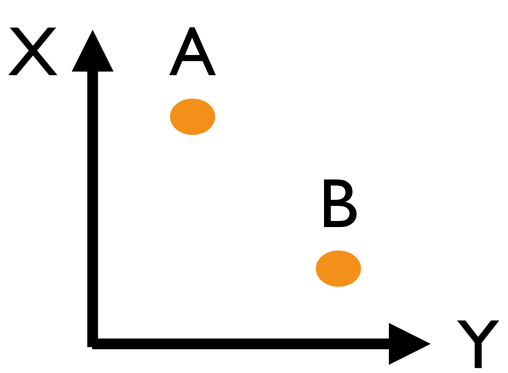
Parallel Coordinates

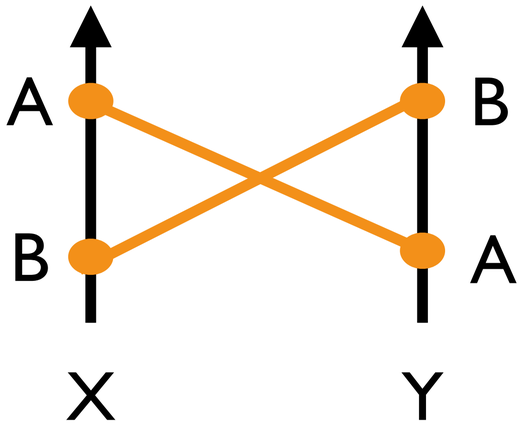
Parallel Coordinates (PC)

Inselberg 1985

Axes represent attributes

Lines connecting axes represent items





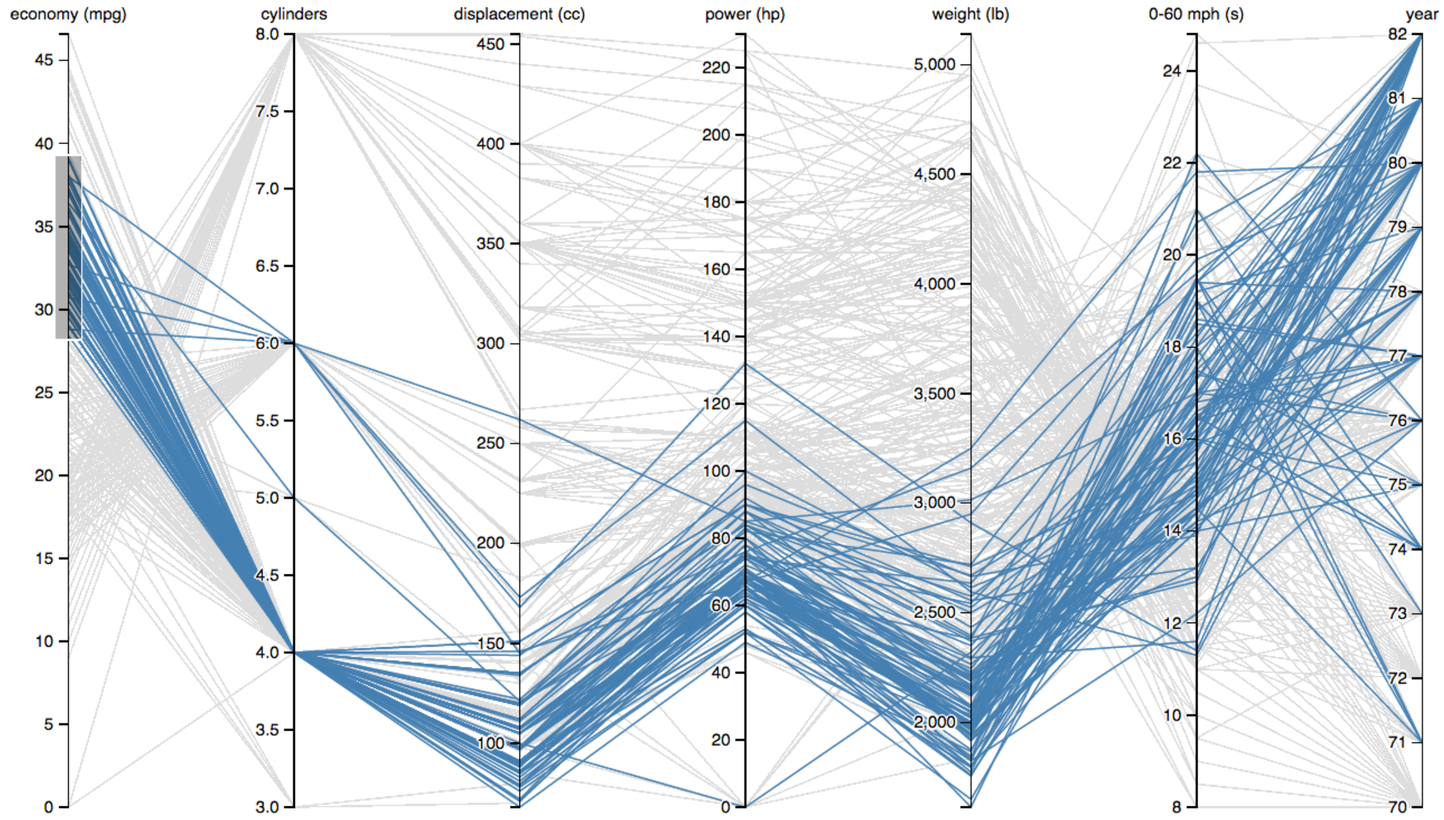
Parallel Coordinates

Each axis represents dimension

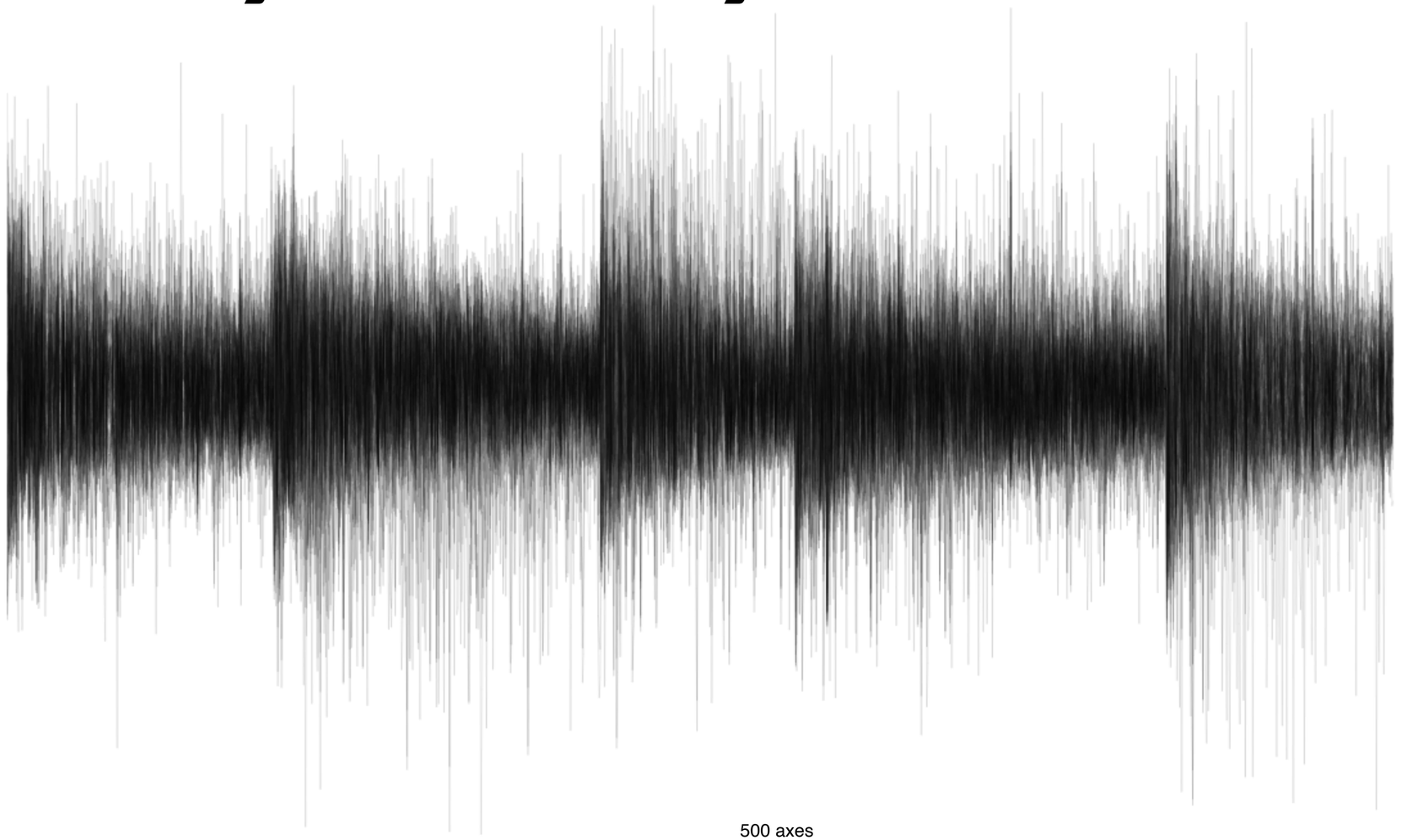
Lines connecting axis represent records

Suitable for

- all tabular data types
- heterogeneous data



PC Limitation:
Scalability to Many Dimensions

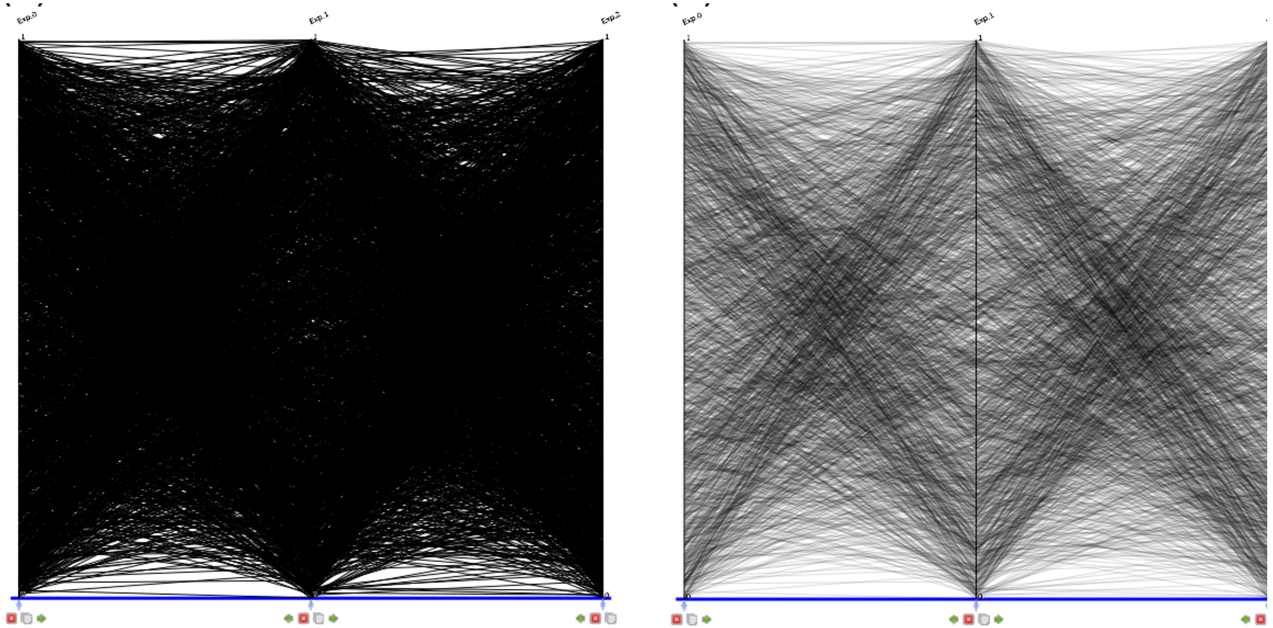


500 axes

PC Limitation: Scalability to Many Items

Solutions:

- Transparency
- Bundling, Clustering
- Sampling

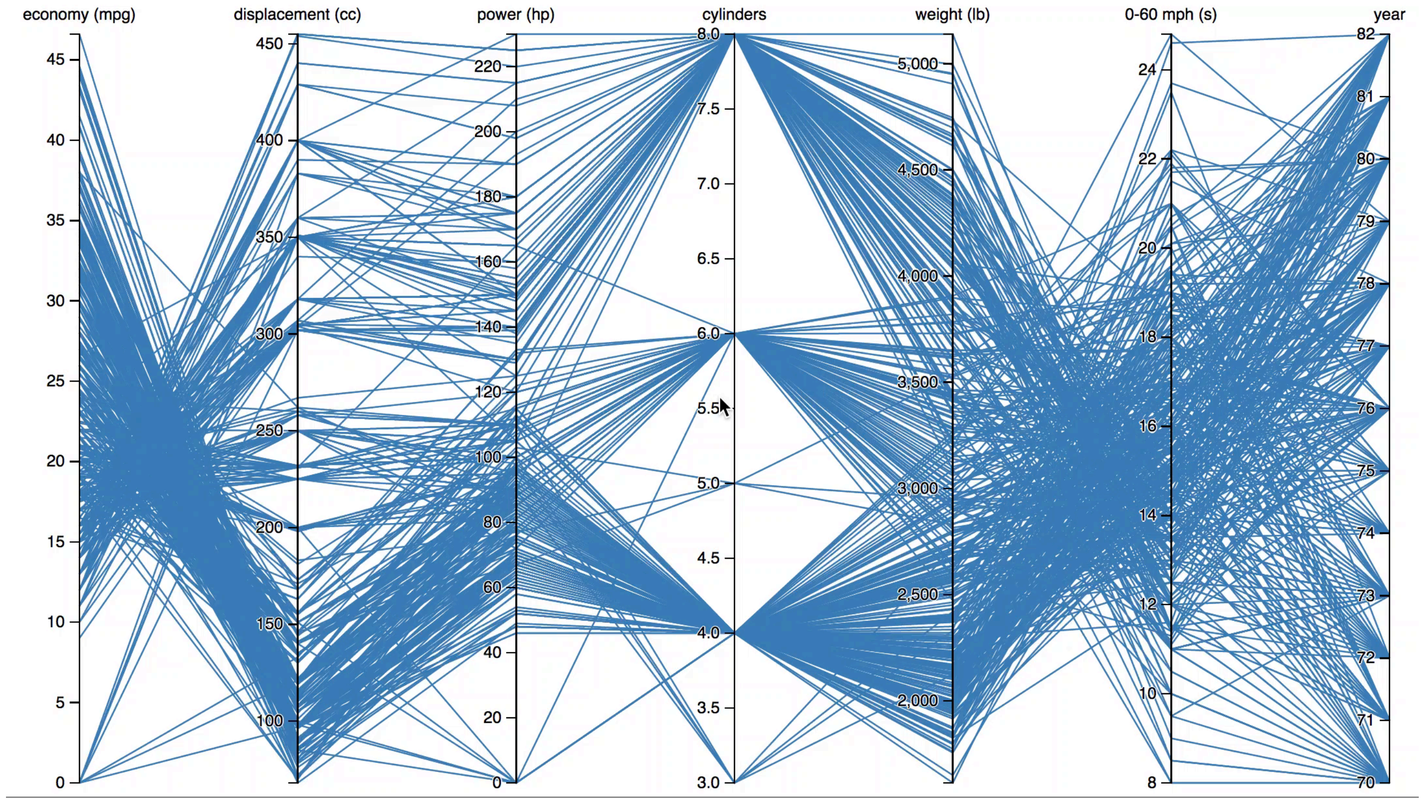


PC Limitations

Correlations only between adjacent axes

Solution: Interaction

- Brushing
- Let user change order



Parallel Coordinates

Shows primarily relationships between adjacent axis

Limited scalability (~50 dimensions, ~1-5k records)

- Transparency of lines

Interaction is crucial

- Axis reordering
- Brushing
- Filtering

Algorithmic support:

- Choosing dimensions
- Choosing order
- Clustering & aggregating records

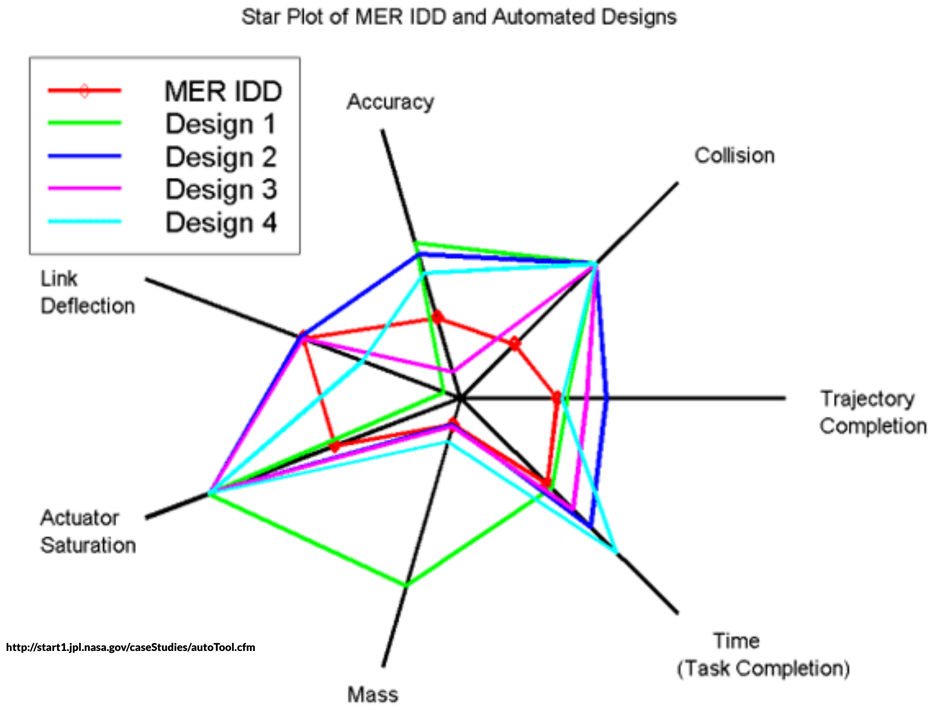
http://bl.ocks.org/jasondavies/1341281

Star Plot

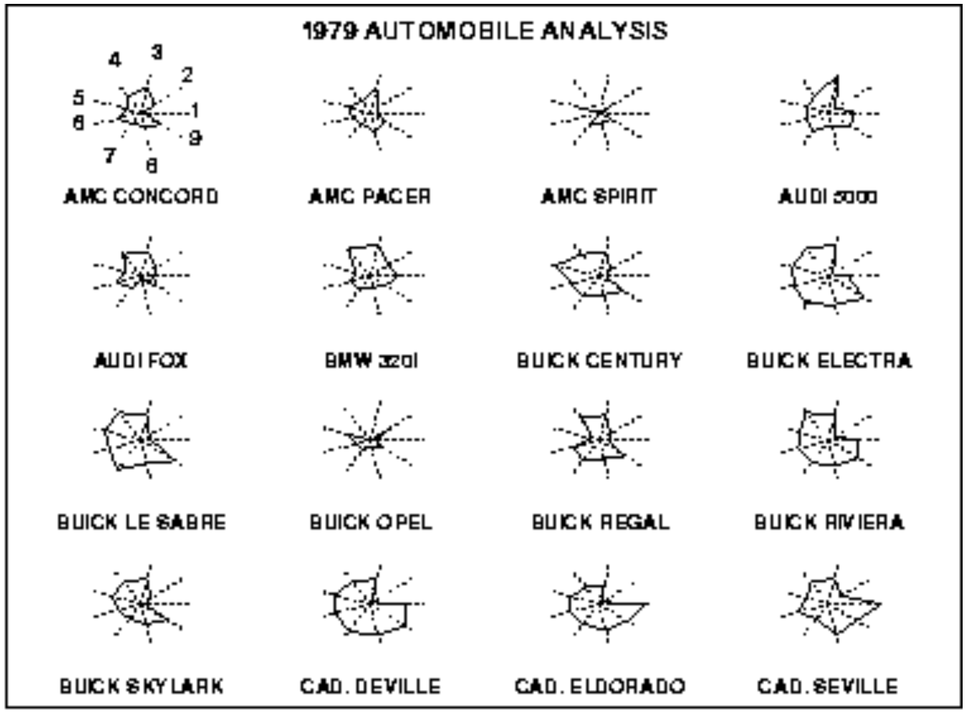
[Coekin1969]

Similar to parallel coordinates

Radiate from a common origin



http://start1.jpl.nasa.gov/caseStudies/autoTool.cfm



http://www.itl.nist.gov/div898/handbook/eda/section3/starplot.htm

http://bl.ocks.org/kevinschaul/raw/8833989/

Data Reduction

Sampling

Don’t show every element, show a (random) subset

Efficient for large dataset

Apply only for display purposes

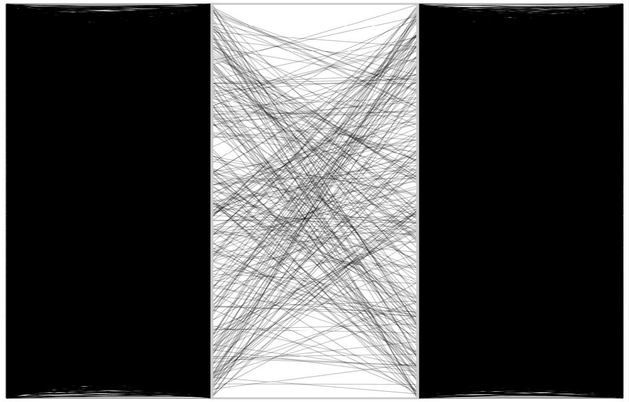
Outlier-preserving approaches

Filtering

Define criteria to remove data, e.g.,

- **minimum variability**
- **> / < / = specific value for one dimension**
- **consistency in replicates, …**

Can be interactive, combined with sampling



[Ellis & Dix, 2006]

Parallel Sets

Parallel Sets

builds on PC to better handle categorical data

- discrete
- small number of values
- no implied ordering between attributes

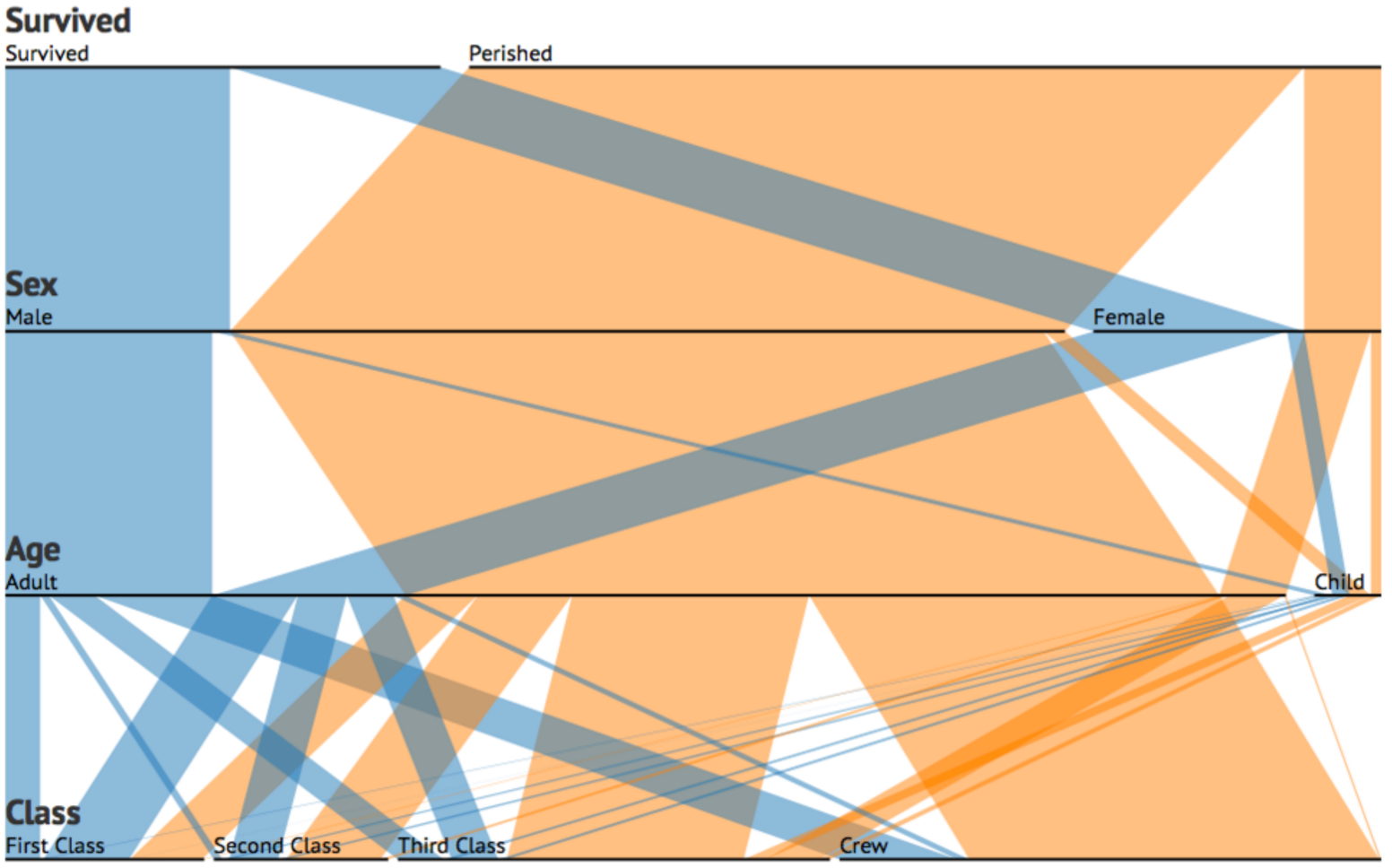
task: find relationship between attributes

interaction driven technique

Parallel Sets

A visualisation technique for multidimensional categorical data.

Titanic Survivors



Curves?

Data: Robert J. MacG. Dawson.

Tabular / Grid / Matrix - Based Representations

Tabular Representation

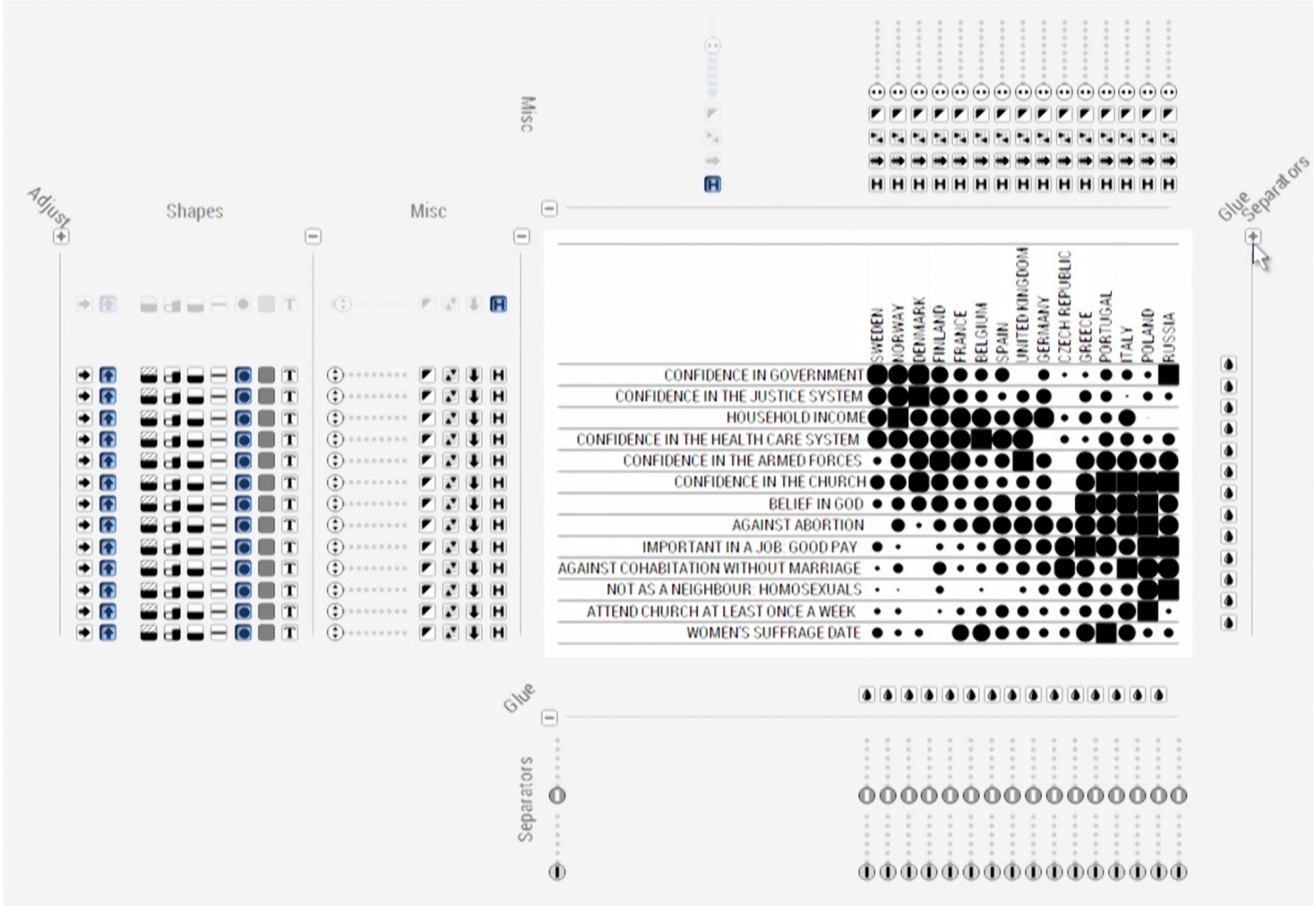
Like spreadsheet: each variable in it’s own column

Visual encodings to make it scalable

Bertifier

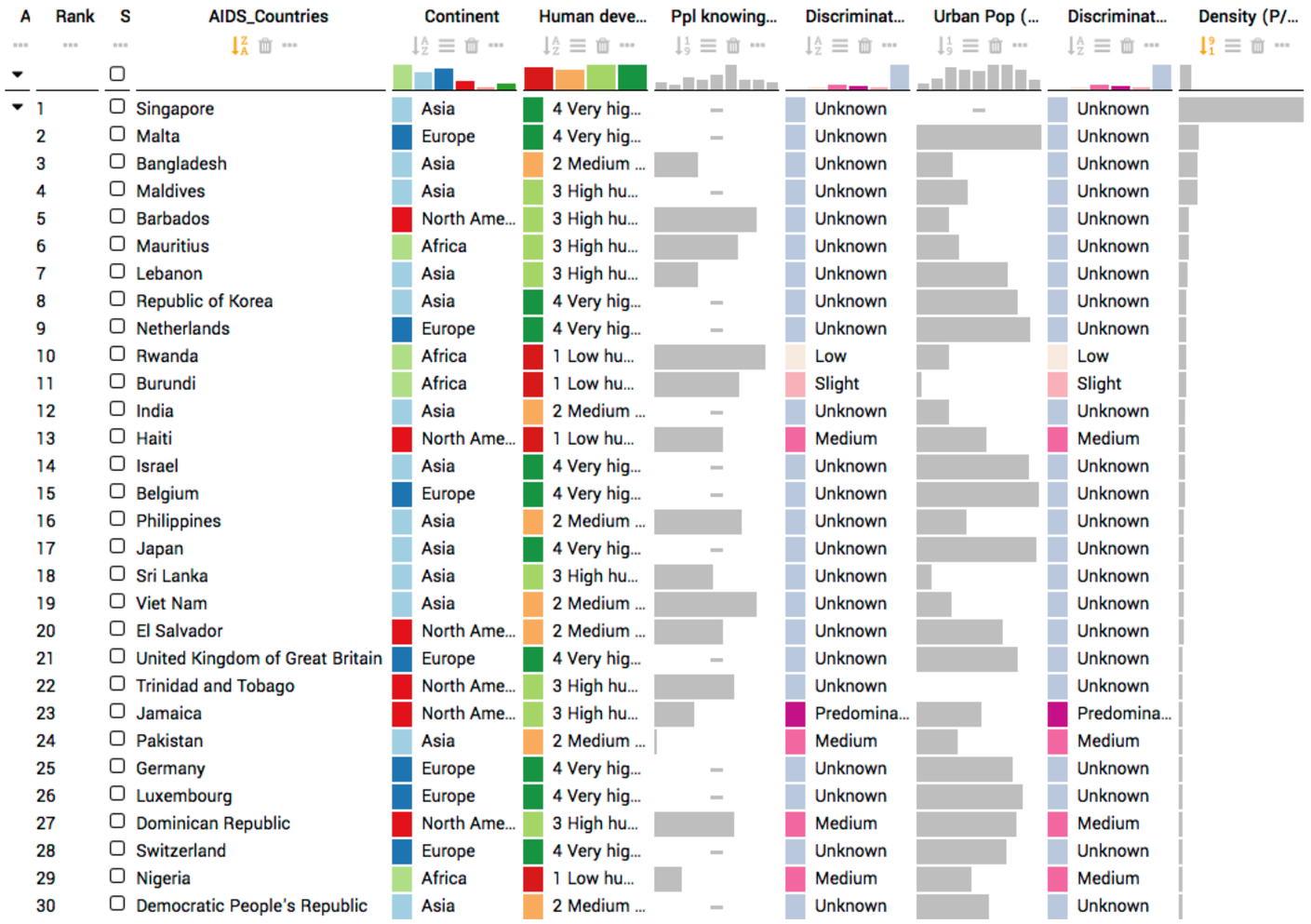
Matrix/Table representation

Authoring Interface



http://www.aviz.fr/bertifier
Charles Perin, Pierre Dragicevic and Jean-Daniel Fekete

Taggle



A	Rank	S	AIDS_Countries	Continent	Human deve...	Ppl knowing...	Discriminat...	Urban Pop (...	Discriminat...	Density (P/...
				Asia	4 Very hig...		Unknown		Unknown	
	1		Singapore	Asia	4 Very hig...		Unknown		Unknown	
	2		Malta	Europe	4 Very hig...		Unknown		Unknown	
	3		Bangladesh	Asia	2 Medium ...		Unknown		Unknown	
	4		Maldives	Asia	3 High hu...		Unknown		Unknown	
	5		Barbados	North Ame...	3 High hu...		Unknown		Unknown	
	6		Mauritius	Africa	3 High hu...		Unknown		Unknown	
	7		Lebanon	Asia	3 High hu...		Unknown		Unknown	
	8		Republic of Korea	Asia	4 Very hig...		Unknown		Unknown	
	9		Netherlands	Europe	4 Very hig...		Unknown		Unknown	
	10		Rwanda	Africa	1 Low hu...		Low		Low	
	11		Burundi	Africa	1 Low hu...		Slight		Slight	
	12		India	Asia	2 Medium ...		Unknown		Unknown	
	13		Haiti	North Ame...	1 Low hu...		Medium		Medium	
	14		Israel	Asia	4 Very hig...		Unknown		Unknown	
	15		Belgium	Europe	4 Very hig...		Unknown		Unknown	
	16		Philippines	Asia	2 Medium ...		Unknown		Unknown	
	17		Japan	Asia	4 Very hig...		Unknown		Unknown	
	18		Sri Lanka	Asia	3 High hu...		Unknown		Unknown	
	19		Viet Nam	Asia	2 Medium ...		Unknown		Unknown	
	20		El Salvador	North Ame...	2 Medium ...		Unknown		Unknown	
	21		United Kingdom of Great Britain	Europe	4 Very hig...		Unknown		Unknown	
	22		Trinidad and Tobago	North Ame...	3 High hu...		Unknown		Unknown	
	23		Jamaica	North Ame...	3 High hu...		Predomina...		Predomina...	
	24		Pakistan	Asia	2 Medium ...		Medium		Medium	
	25		Germany	Europe	4 Very hig...		Unknown		Unknown	
	26		Luxembourg	Europe	4 Very hig...		Unknown		Unknown	
	27		Dominican Republic	North Ame...	3 High hu...		Medium		Medium	
	28		Switzerland	Europe	4 Very hig...		Unknown		Unknown	
	29		Nigeria	Africa	1 Low hu...		Medium		Medium	
	30		Democratic People's Republic	Asia	2 Medium ...		Unknown		Unknown	

Pixel Based Displays

Each cell is a “pixel”, value encoded in color / value

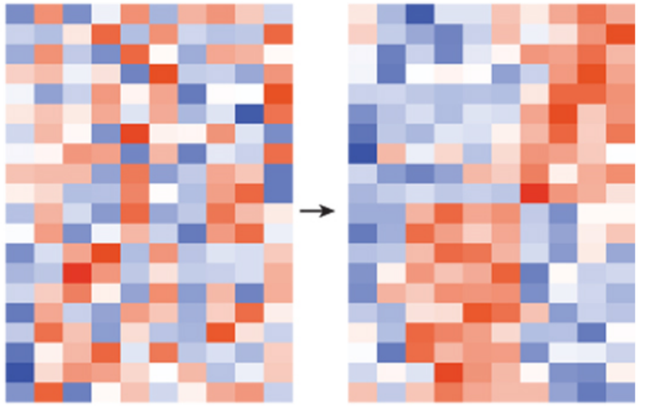
Ordering critical for interpretation

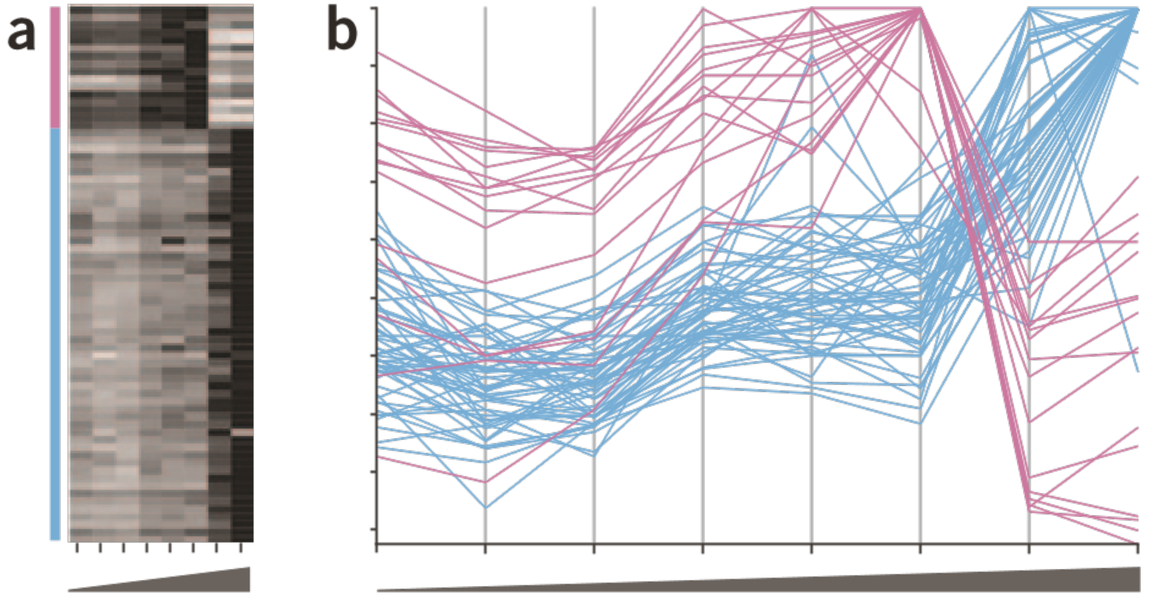
If no ordering inherent, clustering is used

Scalable – 1 px per item

Good for homogeneous data

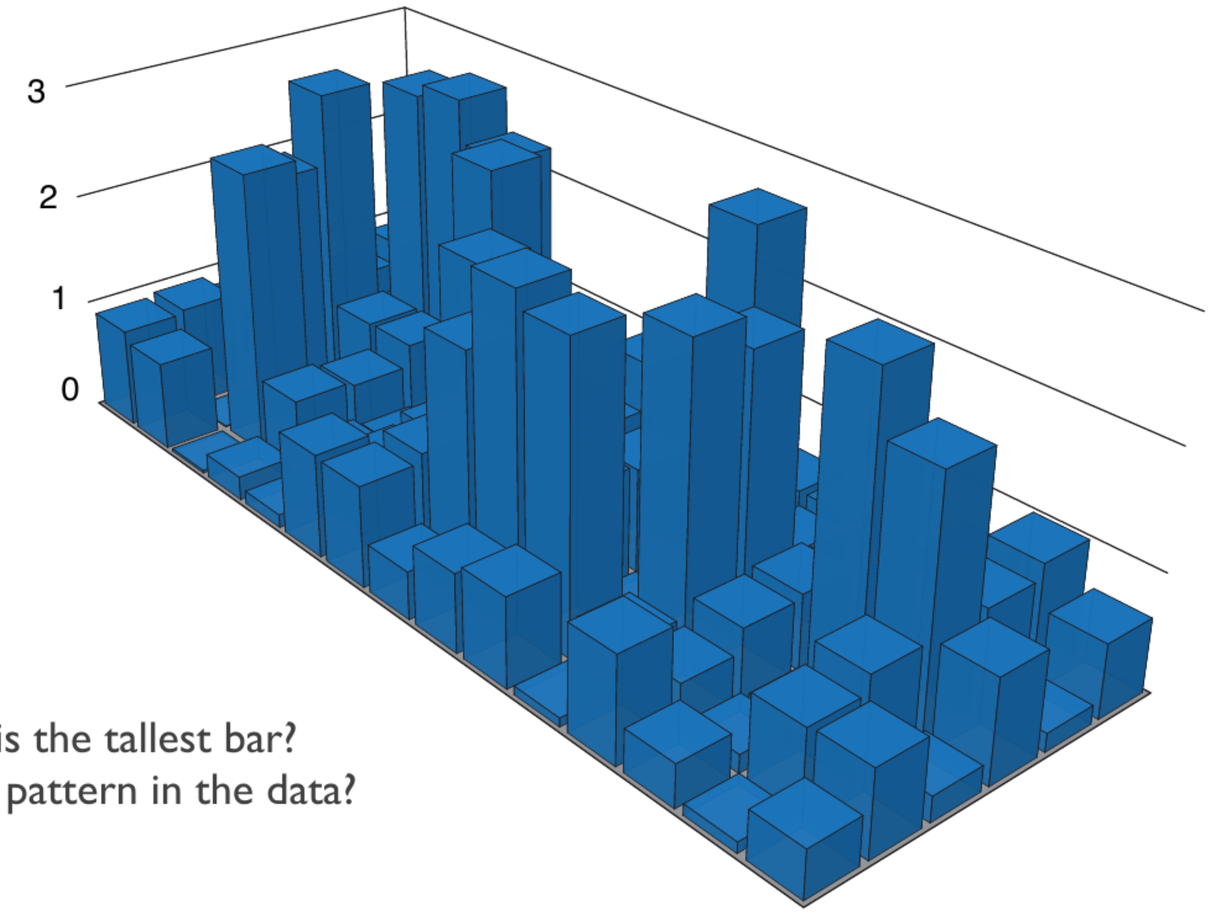
same scale & type





[Gehlenborg & Wong 2012]

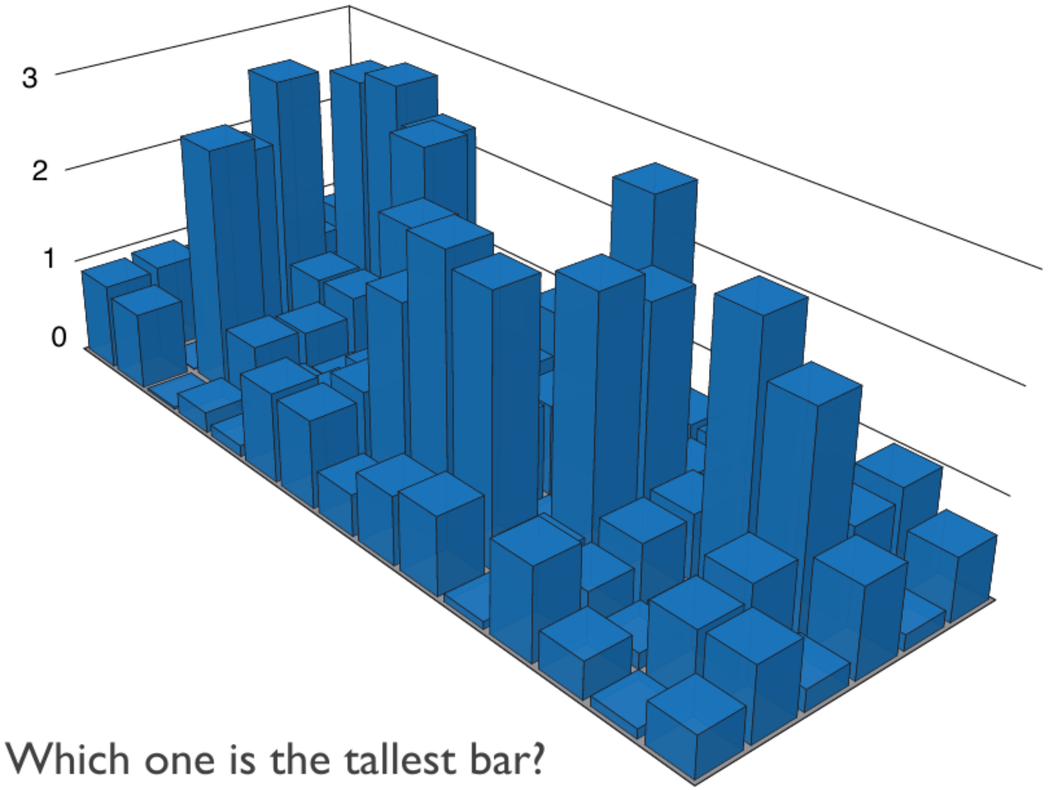
3D Pitfall: Occlusion & Perspective



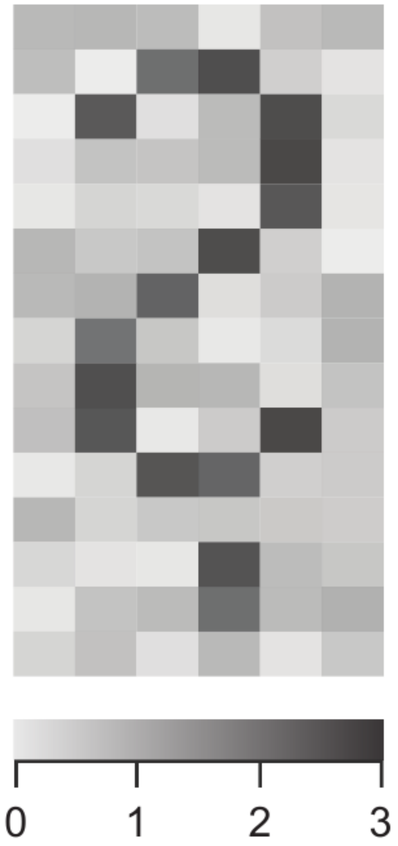
Which one is the tallest bar?
What is the pattern in the data?

[Gehlenborg and Wong, Nature Methods, 2012]

3D Pitfall: Occlusion & Perspective

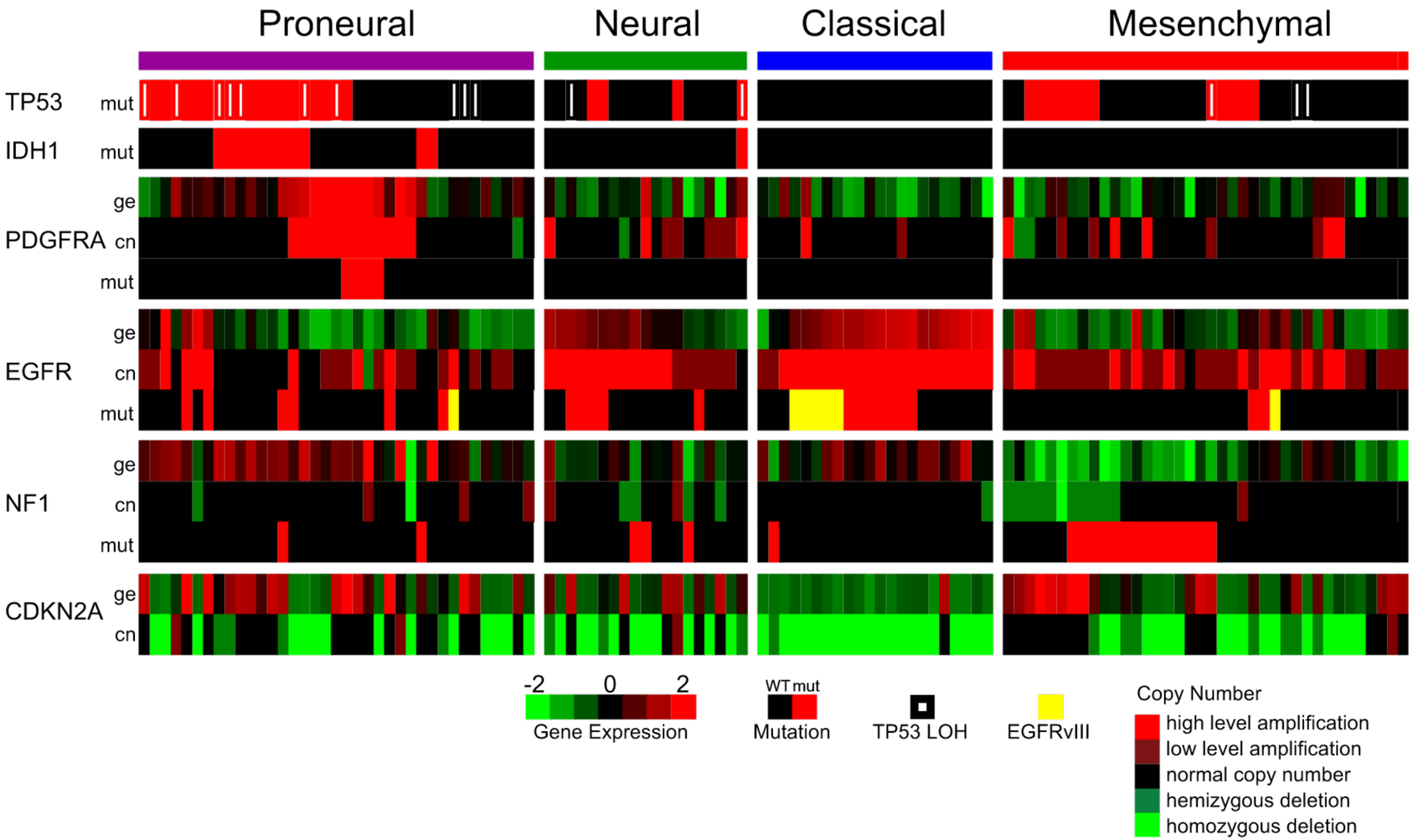


Which one is the tallest bar?
What is the pattern in the data?



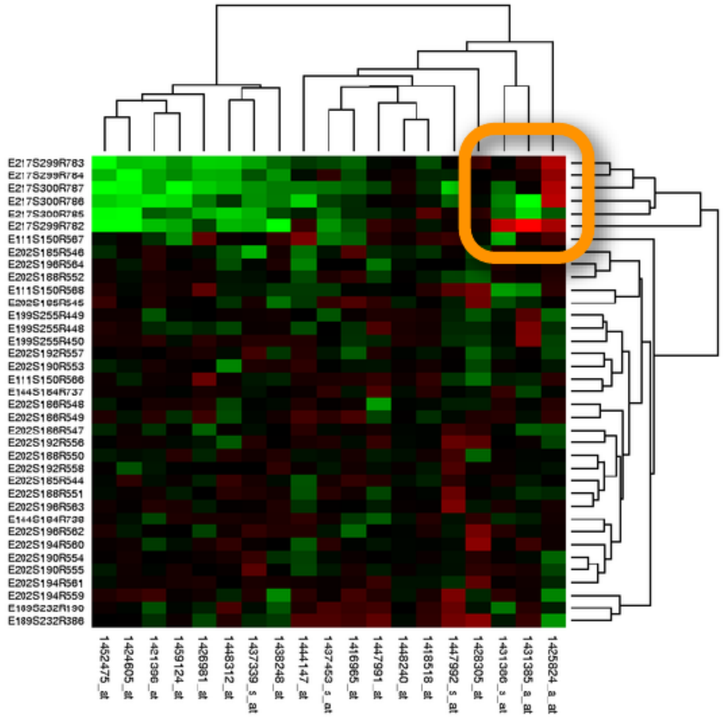
[Gehlenborg and Wong, Nature Methods, 2012]

Heterogeneous Data?

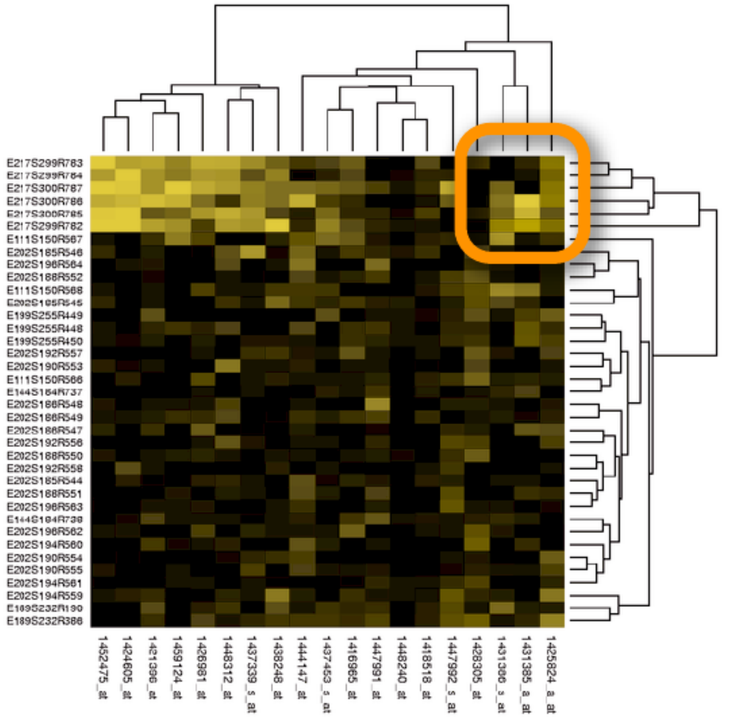


[Verhaak 2012]

Bad Color Mapping

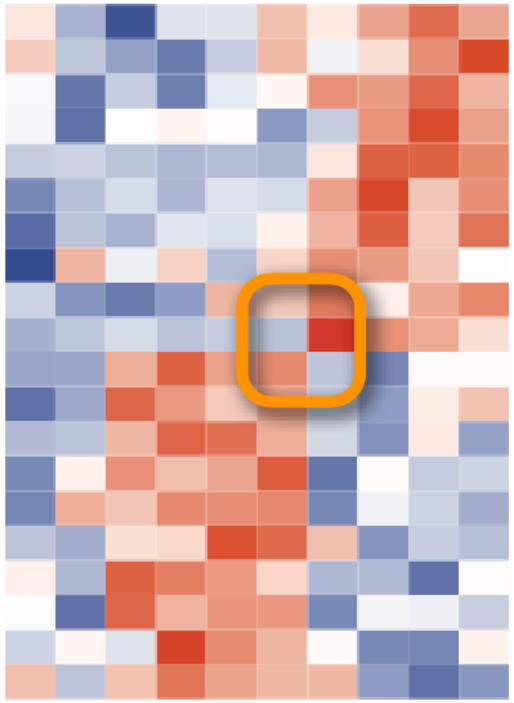


Normal Vision

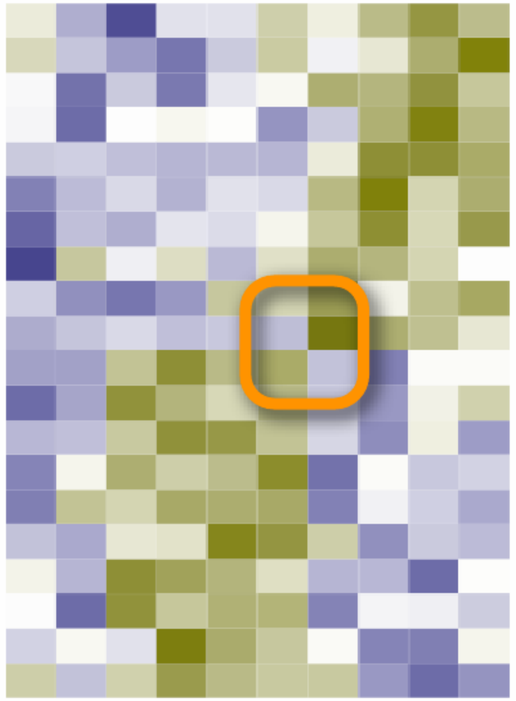


Deuteranope Vision
("Red-Green Blindness")

Good Color Mapping

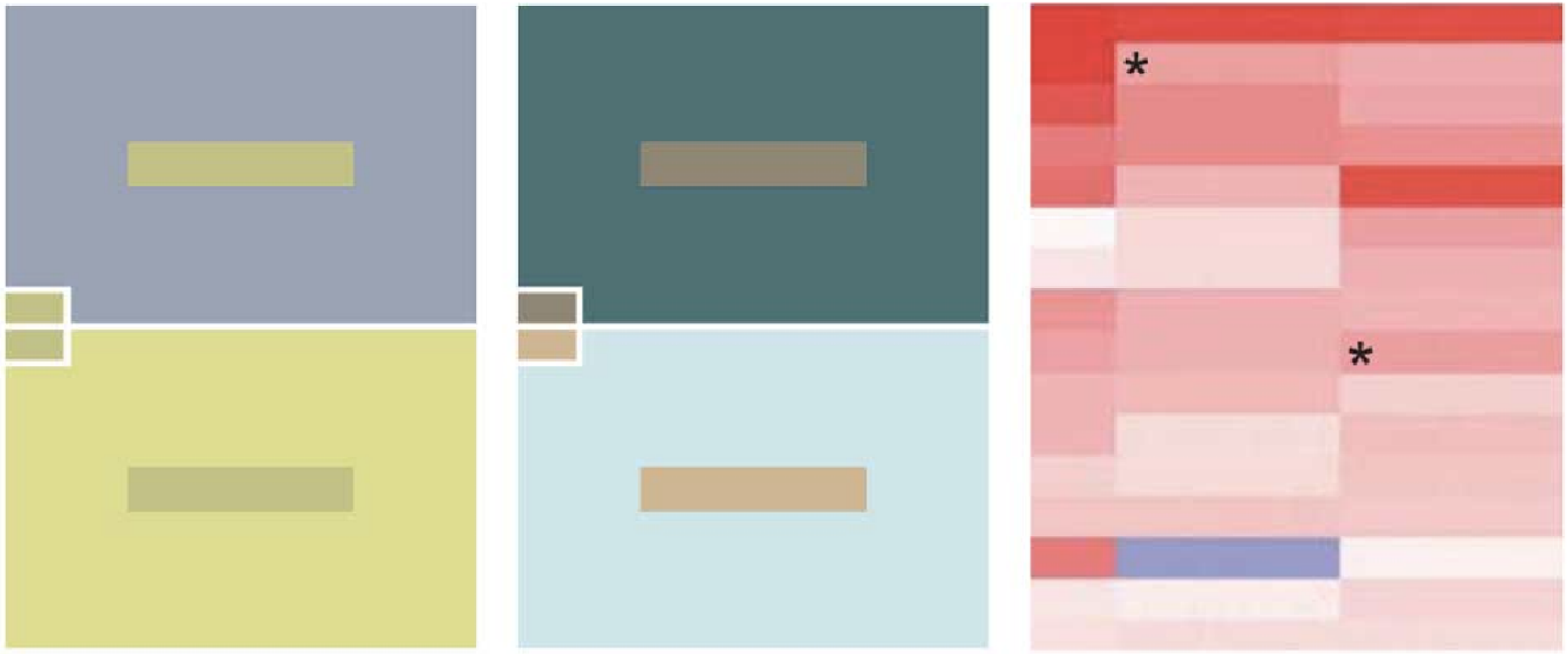


Normal Vision

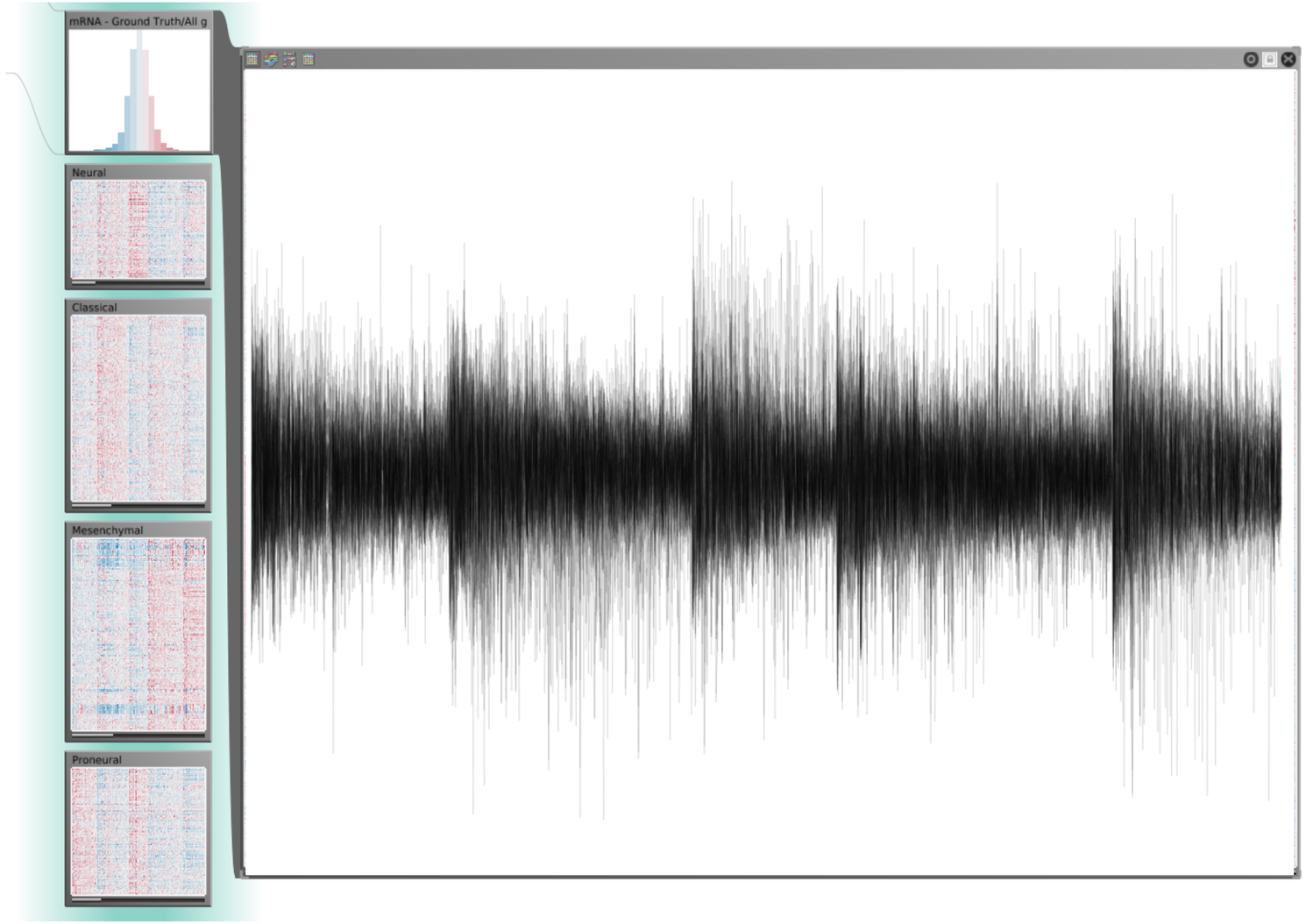


Deuteranope Vision
("Red-Green Blindness")

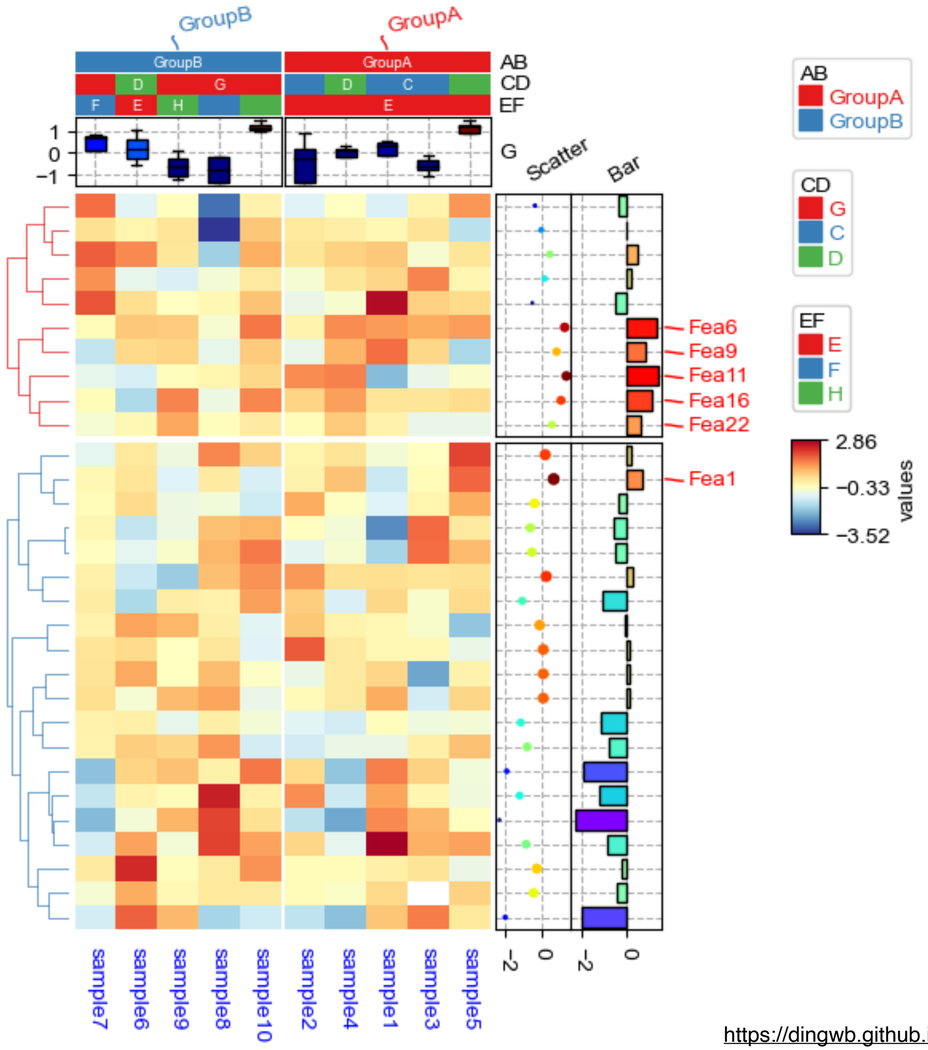
Color is relative!



Clustered Heat Map



Combining Various Charts



https://dingwb.github.io/PyComplexHeatmap/build/html/notebooks/advanced_usage.html#Add-selected-rows-labels

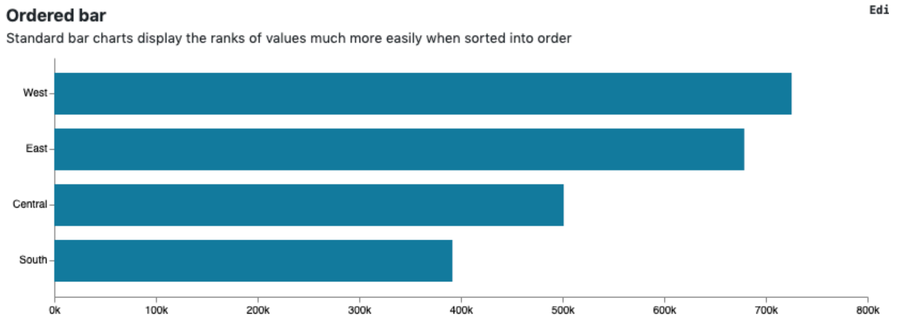
Ranking

Ranking Exercise

	1	2	3	4	5	6	7
Bavaria	8	6	2	4	2	1	3
Dortmund	1	1	5	2	3	8	8
Leipzig	2	2	1	1	1	2	4
Leverkusen	5	5	4	8	7	6	7
Moenchengladbach	10	7	8	7	6	5	1
Wolfsburg	6	4	3	5	8	7	2

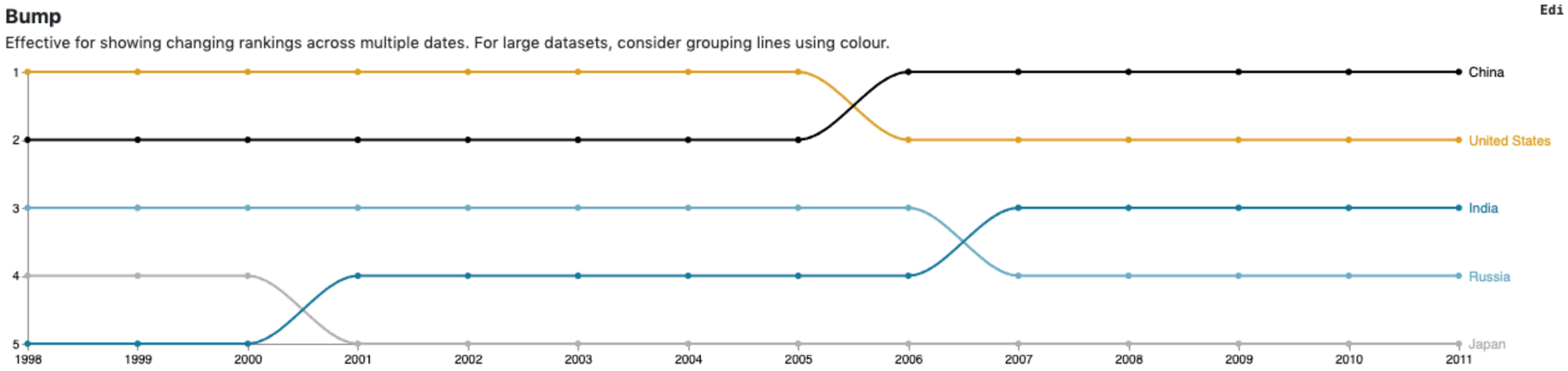
Design a visualization showing the ranking of these football clubs over time.

Ranking



Magnitude Visualization + Sorting

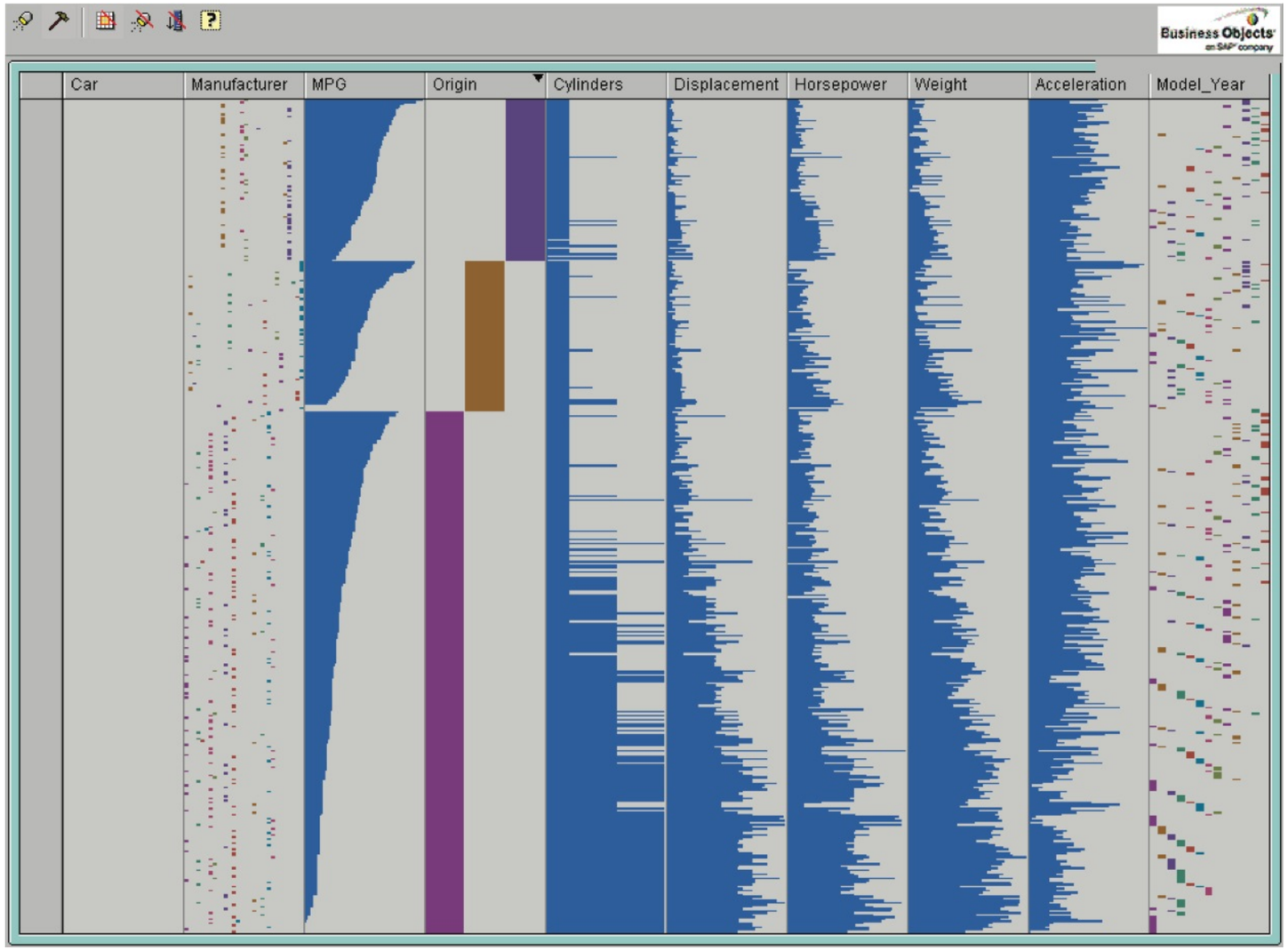
Bump Charts for Rankings over Time



https://gramener.github.io/visual-vocabulary-vega/#/Ranking/

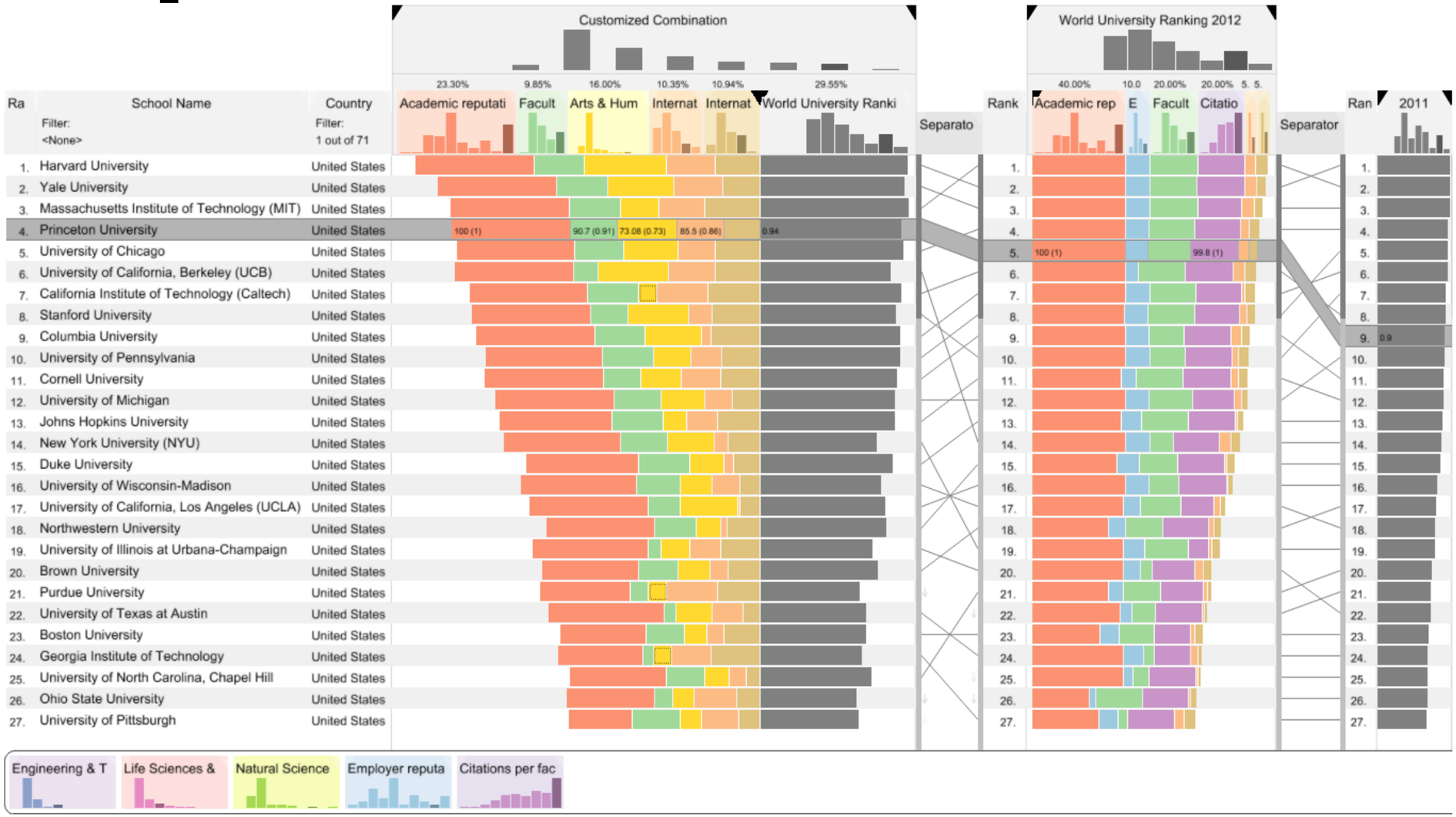
Table Lens

Interactive table-based representation



Rao & Card 1994

LineUp



Video at http://lineup.caleydo.org