Paul Rosen
paul.rosen@utah.edu
@paulrosenphd
https://cspaul.com

Visualization for Data Science
DS-4630 / CS-5630 / CS-6630

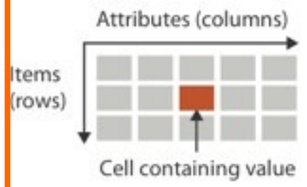
Tabular Data Visualization

dataset types

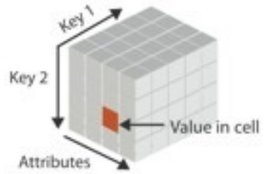
Tables

Items

Attributes



→ *Multidimensional Table*

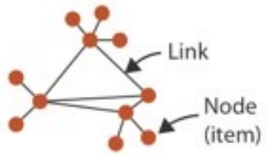


Networks & Trees

Items (nodes)

Links

Attributes



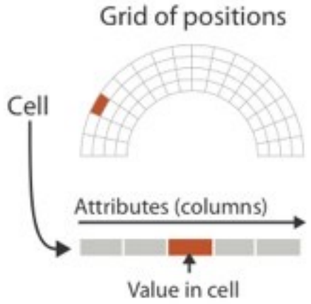
→ *Trees*

Fields

Grids

Positions

Attributes



Geometry

Items

Positions

Clusters, Sets, Lists

Items

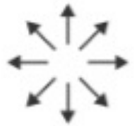
Arrange Tables

- **Express Values**

- **Separate, Order, Align Regions**
  - Separate
  - Order
  - Align

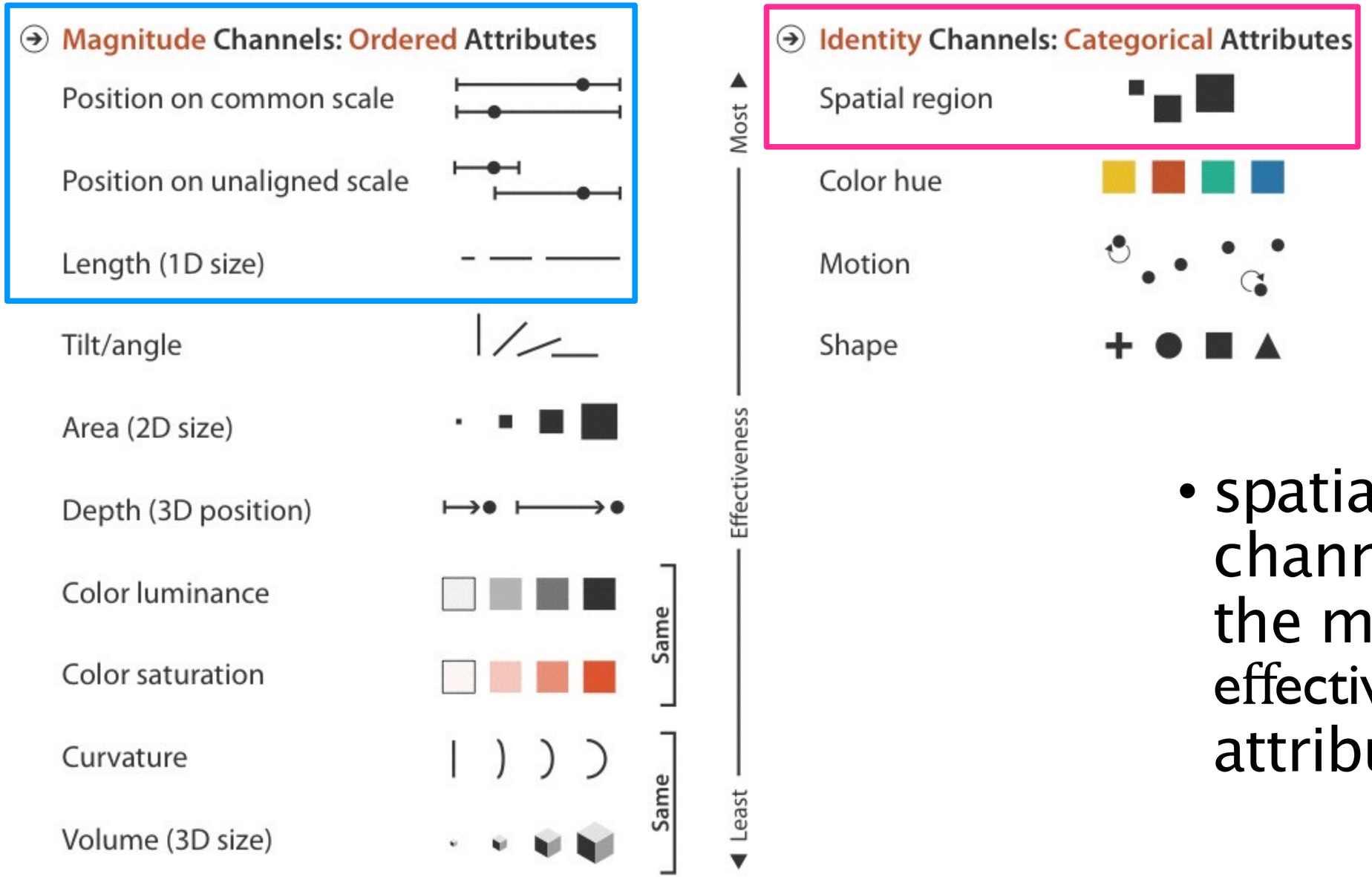
  - *1 Key* — *List*
  - *2 Keys* — *Matrix*
  - *3 Keys* — *Volume*
  - *Many Keys* — *Recursive Subdivision*

- **Axis Orientation**
  - Rectilinear
  - Parallel
  - Radial



- **Layout Density**
  - Dense
  - Space-Filling

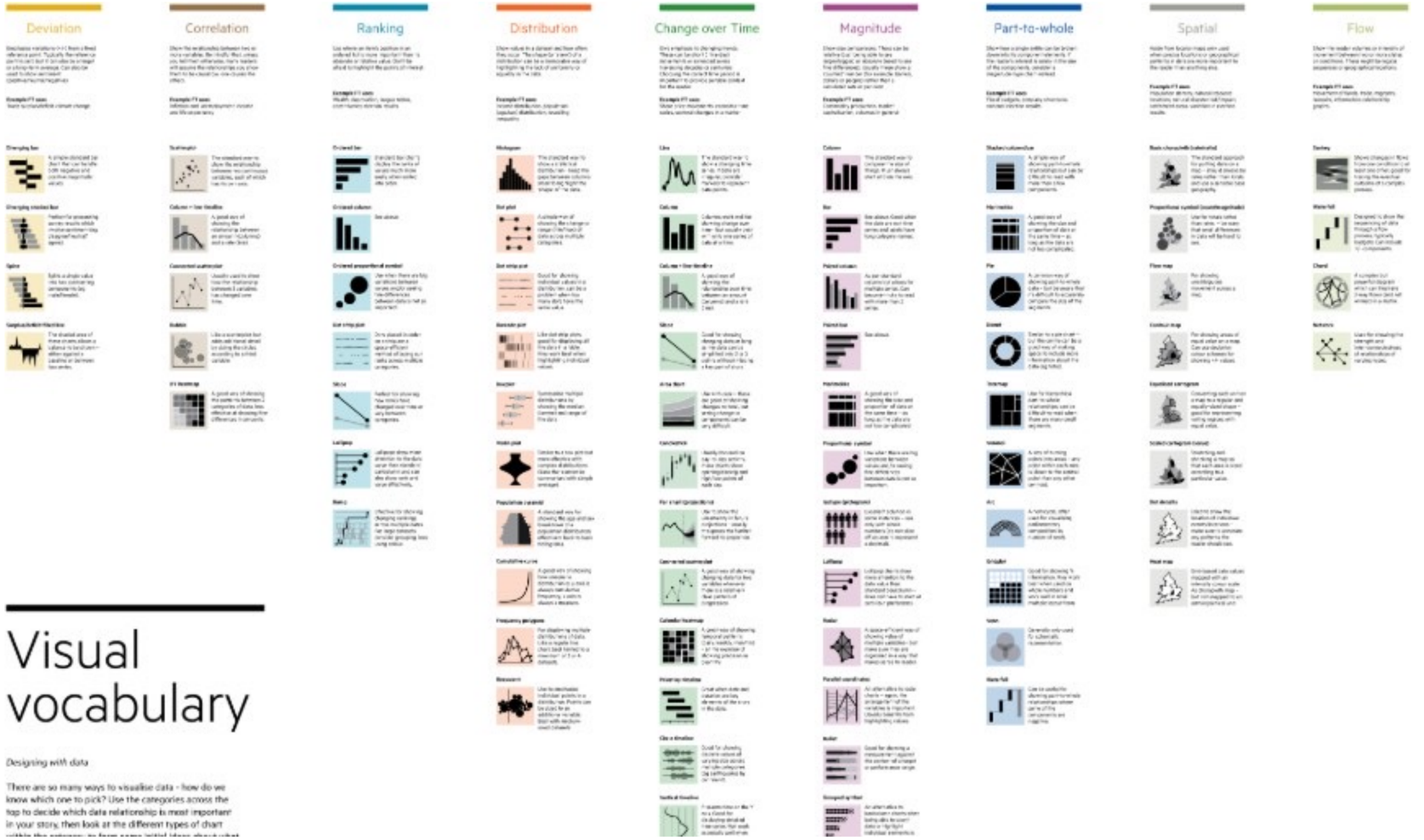
arrange is the focus of all four design choices for tabular data



- spatial channels are the most effective for all attribute types

Techniques and Tasks

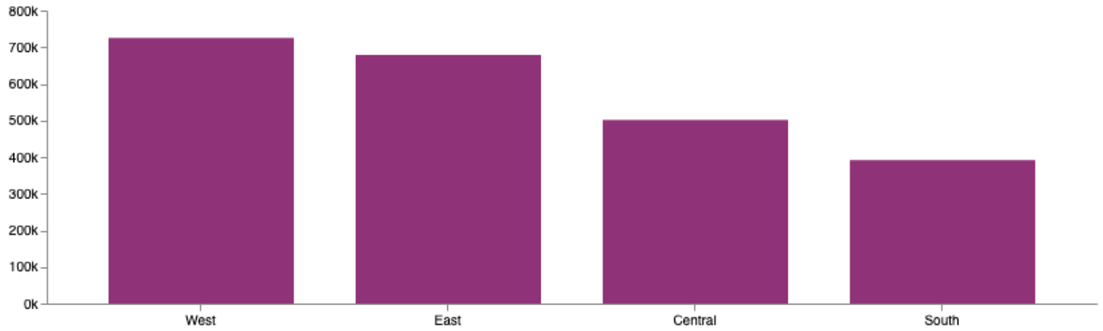
- Magnitude
- Part to whole
- Distribution
- Deviation
- Change over Time
- Ranking
- Correlation



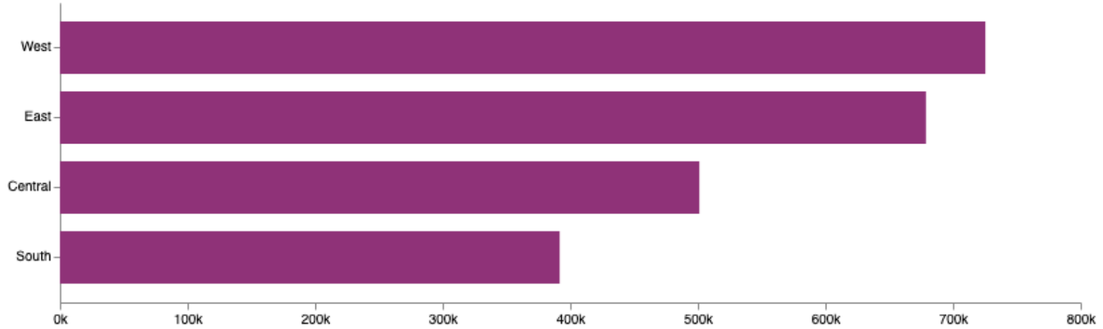
https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary
https://gramener.github.io/visual-vocabulary-vega/#/Magnitude/

Magnitude

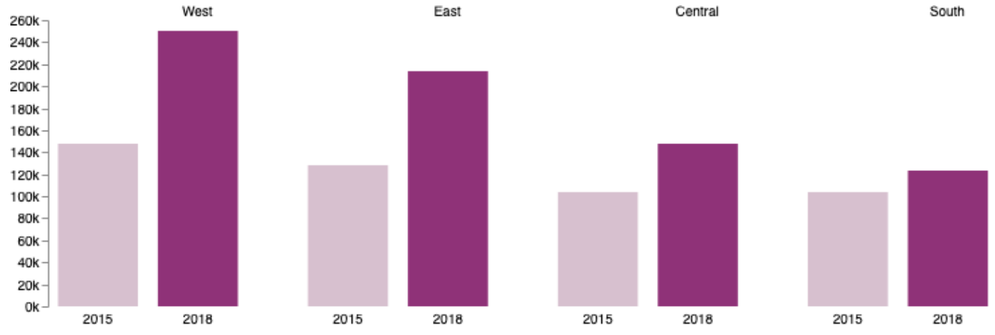
Bar Chart Variants



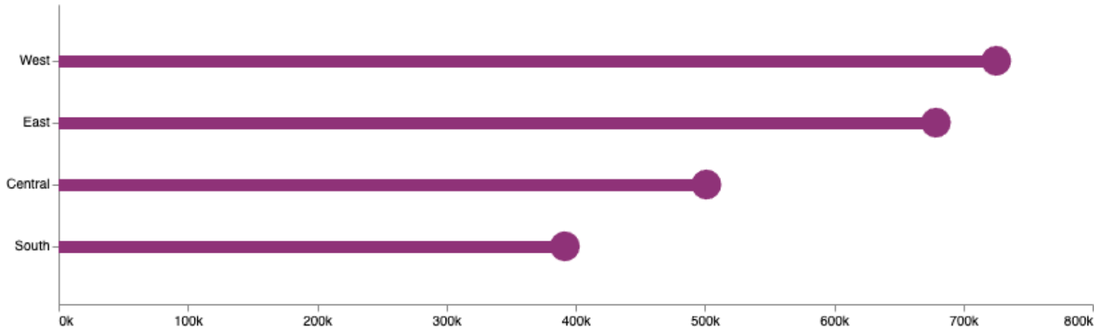
Vertical Bar Chart / Column Chart



Horizontal Bar Chart

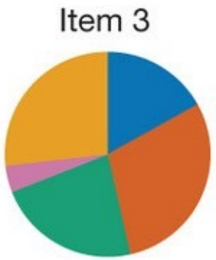


Grouped Bar Chart

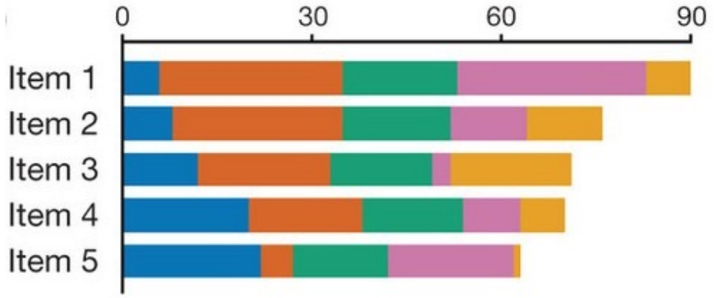


Lollipop Chart

Comparison of bar chart types

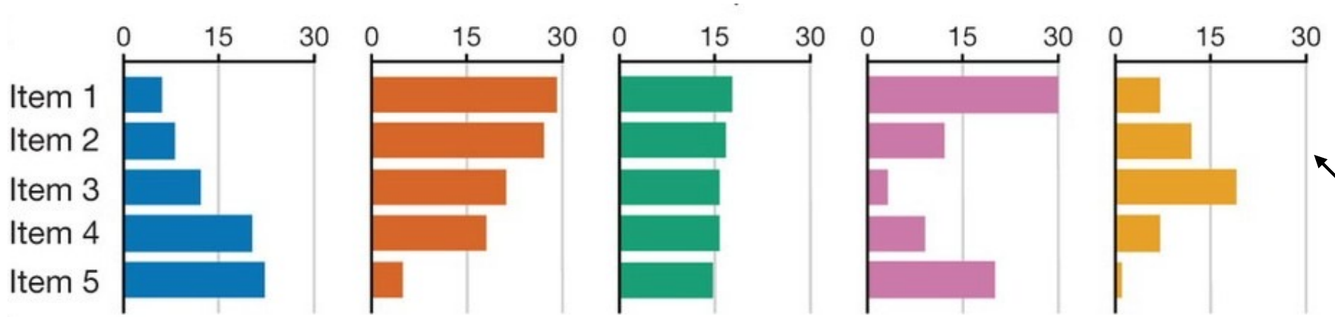


Pie Chart

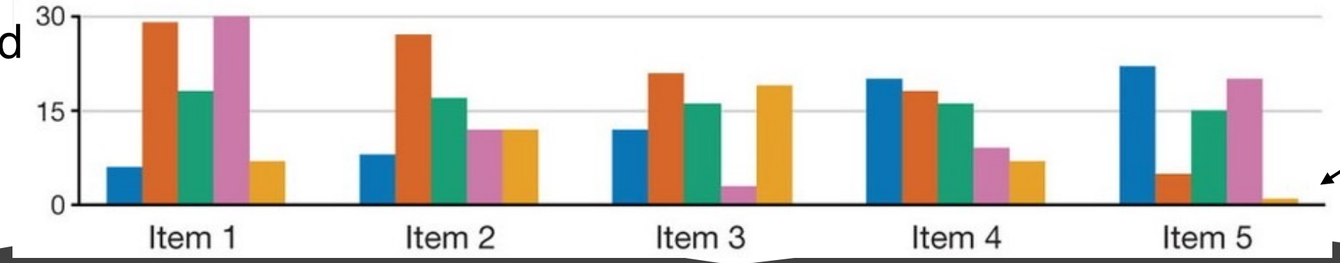


Stacked bar chart

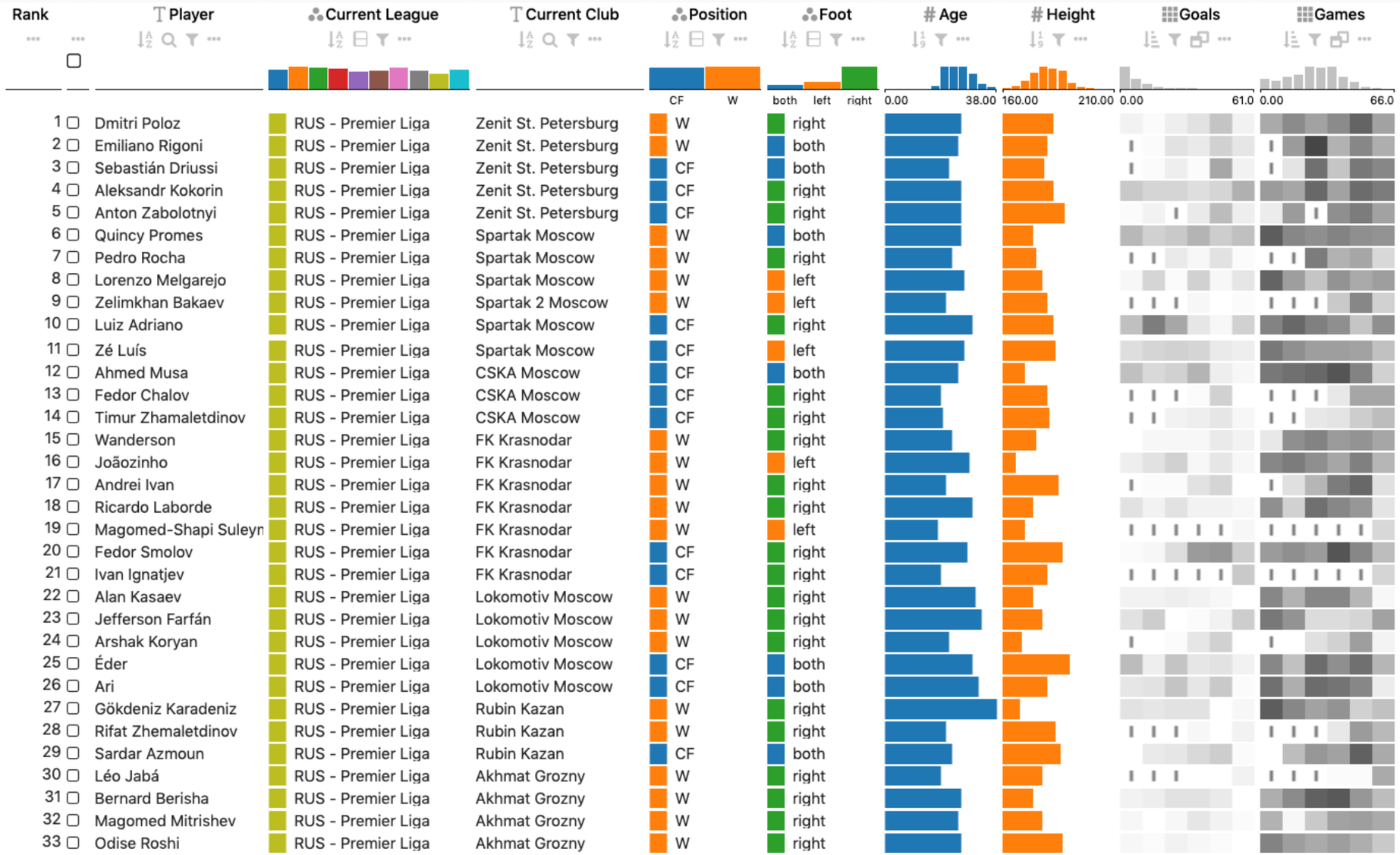
Layered Bar Chart



Grouped Bar Chart

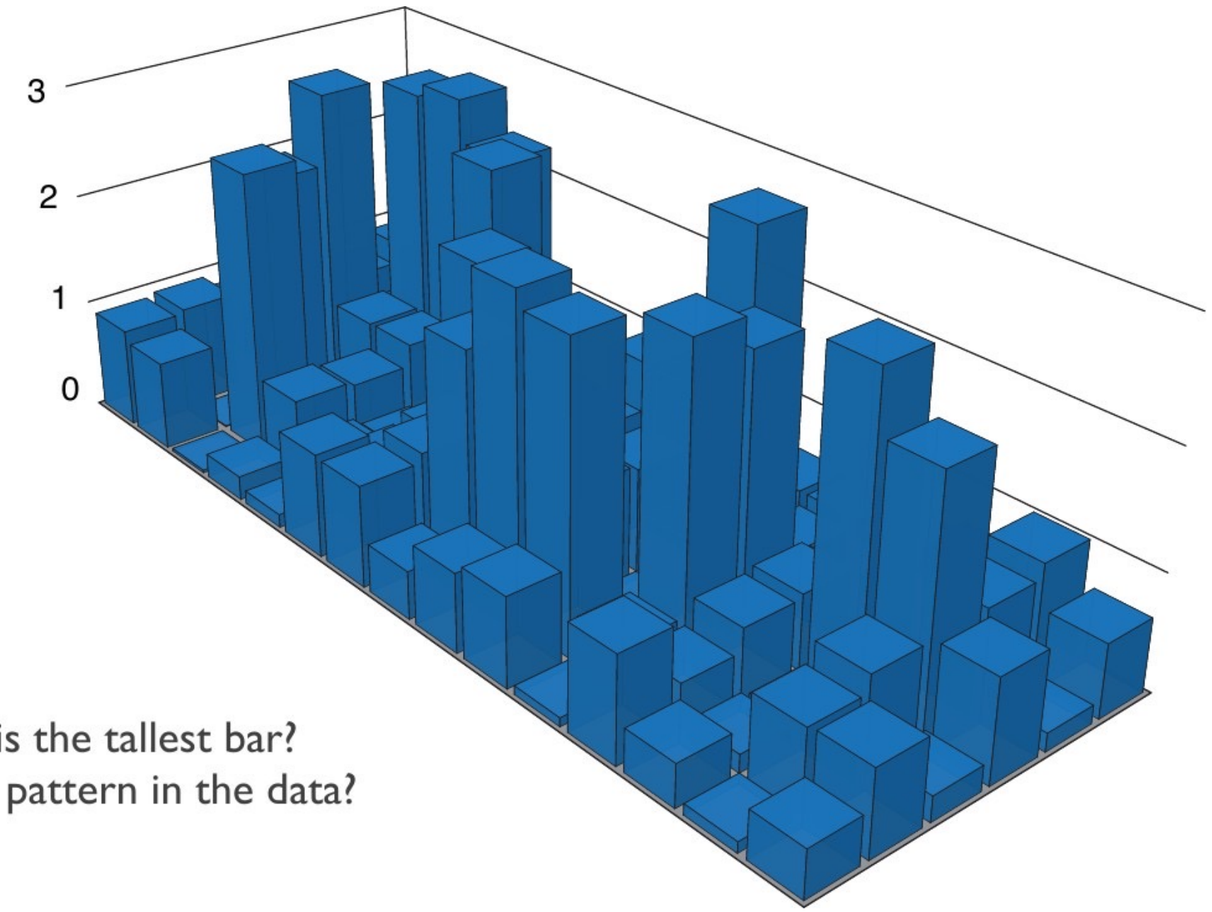


Small Multiples



Rank	Player	Current League	Current Club	Position	Foot	Age	Height	Goals	Games
				CF W	both left right	0.00 38.00	160.00 210.00	0.00 61.0	0.00 66.0
1	Dmitri Poloz	RUS - Premier Liga	Zenit St. Petersburg	W	right				
2	Emiliano Rigoni	RUS - Premier Liga	Zenit St. Petersburg	W	both				
3	Sebastián Driussi	RUS - Premier Liga	Zenit St. Petersburg	CF	both				
4	Aleksandr Kokorin	RUS - Premier Liga	Zenit St. Petersburg	CF	right				
5	Anton Zabolotnyi	RUS - Premier Liga	Zenit St. Petersburg	CF	right				
6	Quincy Promes	RUS - Premier Liga	Spartak Moscow	W	both				
7	Pedro Rocha	RUS - Premier Liga	Spartak Moscow	W	right				
8	Lorenzo Melgarejo	RUS - Premier Liga	Spartak Moscow	W	left				
9	Zelimkhan Bakaev	RUS - Premier Liga	Spartak 2 Moscow	W	left				
10	Luiz Adriano	RUS - Premier Liga	Spartak Moscow	CF	right				
11	Zé Luís	RUS - Premier Liga	Spartak Moscow	CF	left				
12	Ahmed Musa	RUS - Premier Liga	CSKA Moscow	CF	both				
13	Fedor Chalov	RUS - Premier Liga	CSKA Moscow	CF	right				
14	Timur Zhamaletdinov	RUS - Premier Liga	CSKA Moscow	CF	right				
15	Wanderson	RUS - Premier Liga	FK Krasnodar	W	right				
16	Joãozinho	RUS - Premier Liga	FK Krasnodar	W	left				
17	Andrei Ivan	RUS - Premier Liga	FK Krasnodar	W	right				
18	Ricardo Laborde	RUS - Premier Liga	FK Krasnodar	W	right				
19	Magomed-Shapi Suleyn	RUS - Premier Liga	FK Krasnodar	W	left				
20	Fedor Smolov	RUS - Premier Liga	FK Krasnodar	CF	right				
21	Ivan Ignatjev	RUS - Premier Liga	FK Krasnodar	CF	right				
22	Alan Kasaev	RUS - Premier Liga	Lokomotiv Moscow	W	right				
23	Jefferson Farfán	RUS - Premier Liga	Lokomotiv Moscow	W	right				
24	Arshak Koryan	RUS - Premier Liga	Lokomotiv Moscow	W	right				
25	Éder	RUS - Premier Liga	Lokomotiv Moscow	CF	both				
26	Ari	RUS - Premier Liga	Lokomotiv Moscow	CF	both				
27	Gökdeniz Karadeniz	RUS - Premier Liga	Rubin Kazan	W	right				
28	Rifat Zhemaletdinov	RUS - Premier Liga	Rubin Kazan	W	right				
29	Sardar Azmoun	RUS - Premier Liga	Rubin Kazan	CF	both				
30	Léo Jabá	RUS - Premier Liga	Akhmat Grozny	W	right				
31	Bernard Berisha	RUS - Premier Liga	Akhmat Grozny	W	right				
32	Magomed Mitrishev	RUS - Premier Liga	Akhmat Grozny	W	right				
33	Odise Roshi	RUS - Premier Liga	Akhmat Grozny	W	right				

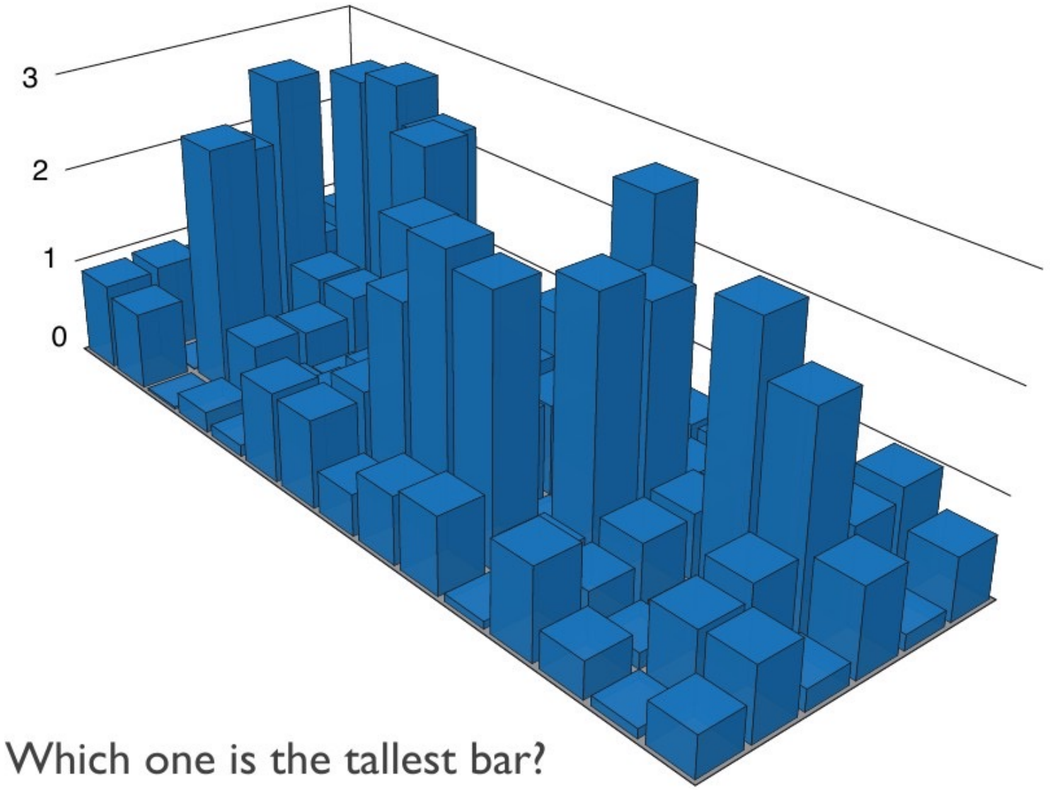
3D Pitfall: Occlusion & Perspective



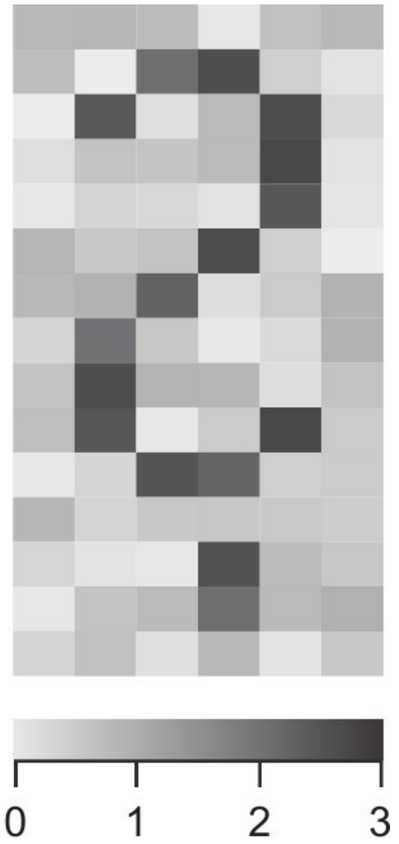
Which one is the tallest bar?
What is the pattern in the data?

[Gehlenborg and Wong, Nature Methods, 2012]

3D Pitfall: Occlusion & Perspective

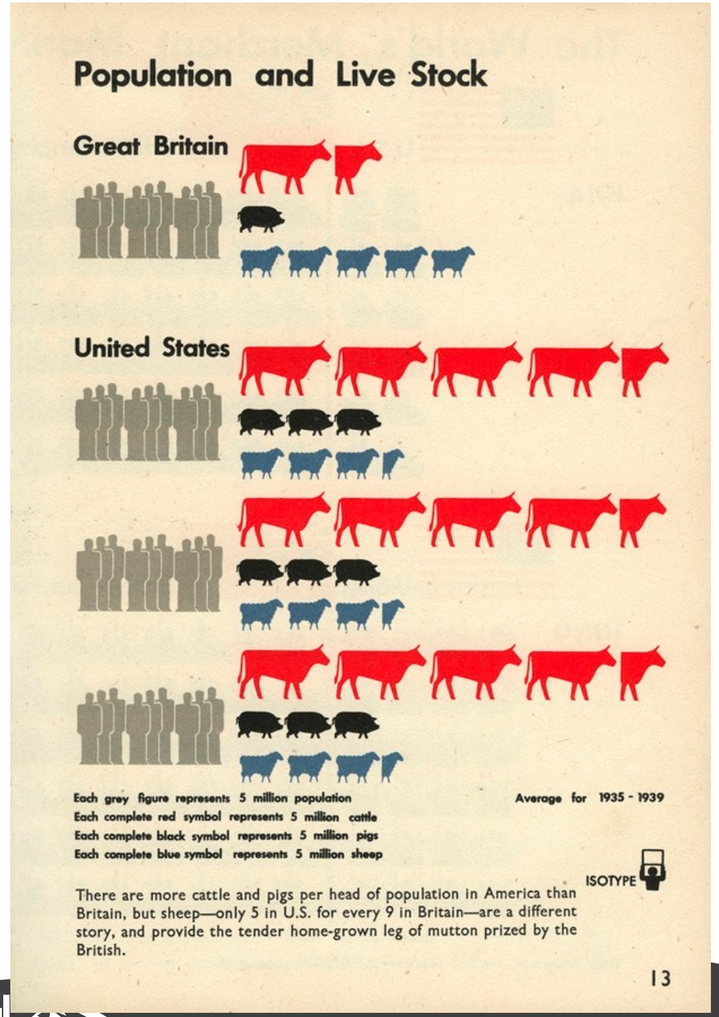


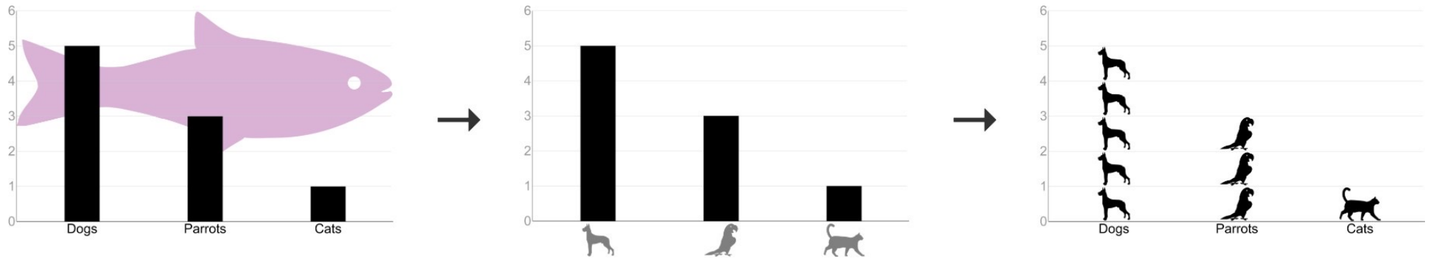
Which one is the tallest bar?
What is the pattern in the data?



[Gehlenborg and Wong, Nature Methods, 2012]

IsoType Visualization





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

Otto and Marie Neurath

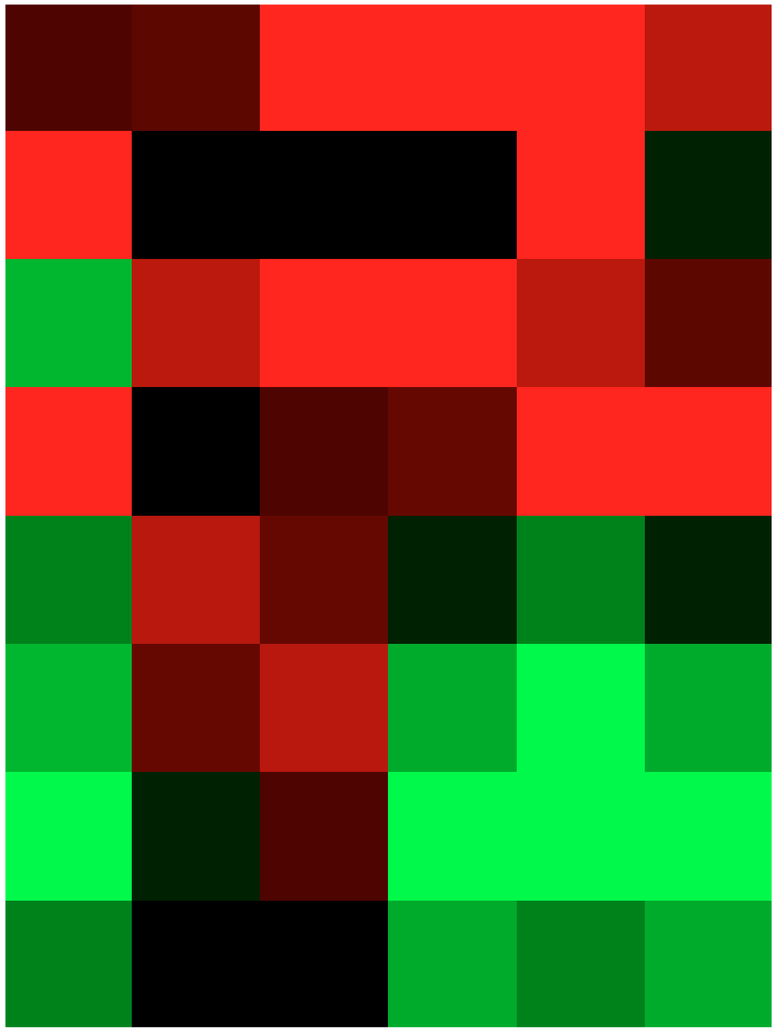
heatmap

- uses heatmap representation
  - matrix layout using keys
  - encode values with color
- often augmented with clustering

0.2	0.4	I	I	I	0.8
I	0	0	0	I	I
0.7	0.8	I	I	0.8	0.6
I	0	0.2	0.5	I	I
0.5	0.8	0.5	0.3	0.5	0.8
0.7	0.5	0.8	0.7	I	I
I	0.3	0.4	I	I	I
0.5	0	0	0.7	0.5	0.3

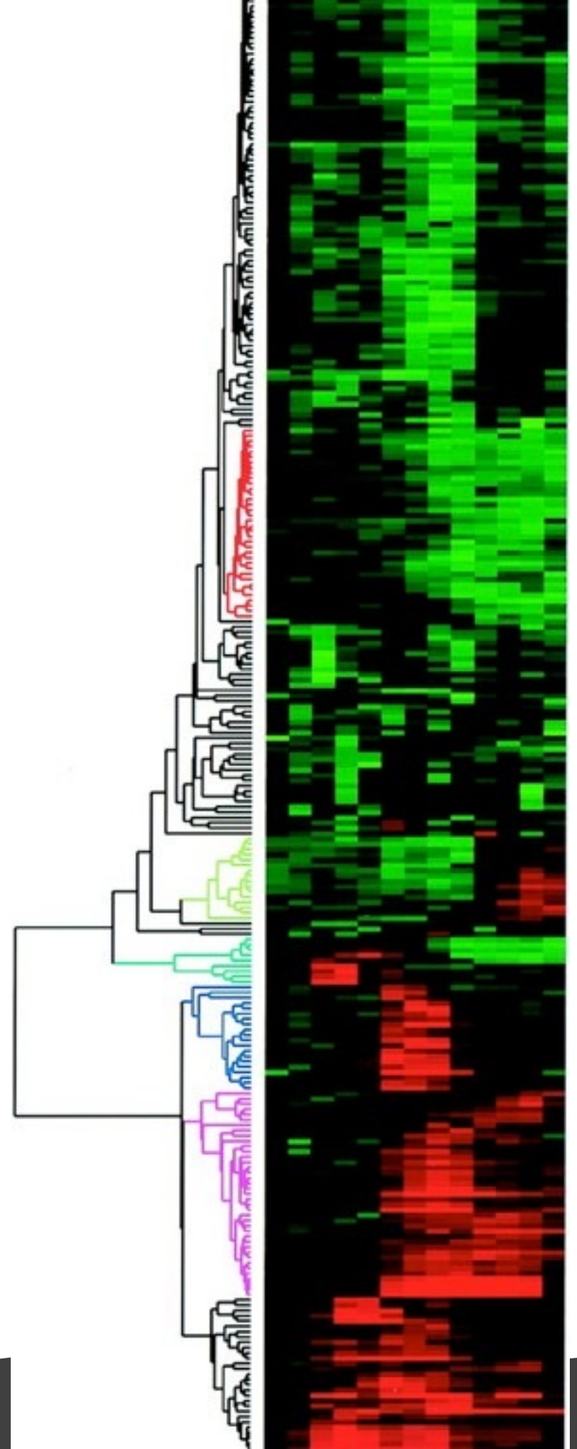
heatmap

- uses heatmap representation
  - matrix layout using keys
  - encode values with color
- often augmented with clustering



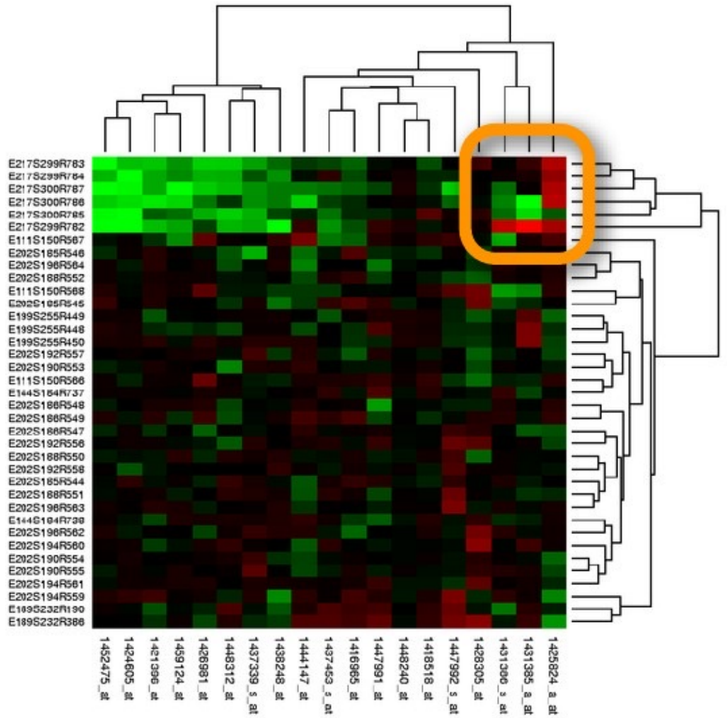
heatmap

- uses heatmap representation
  - matrix layout using keys
  - encode values with color
- often augmented with clustering
- here, used on genomic data

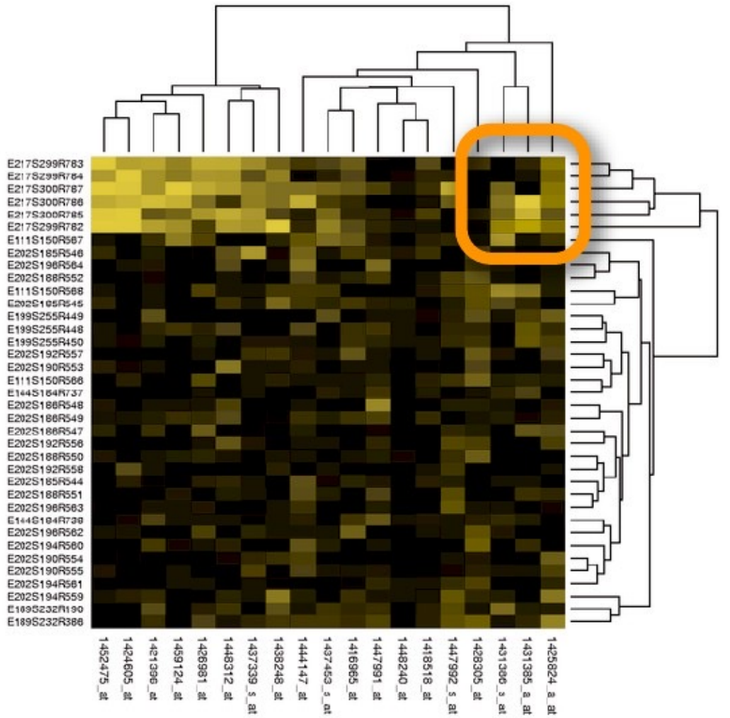


Eisen 1998

Bad Color Mapping

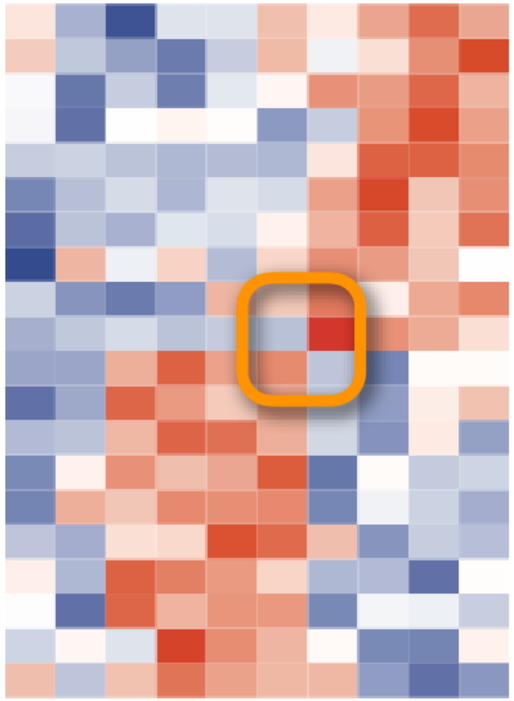


Normal Vision

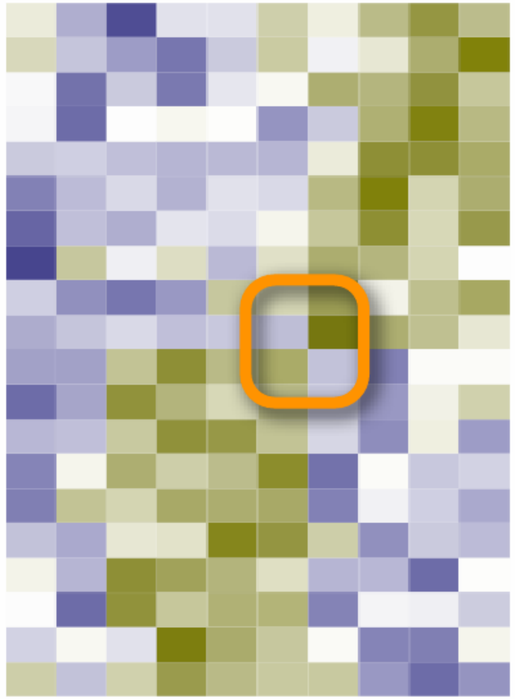


Deuteranope Vision
("Red-Green Blindness")

Good Color Mapping

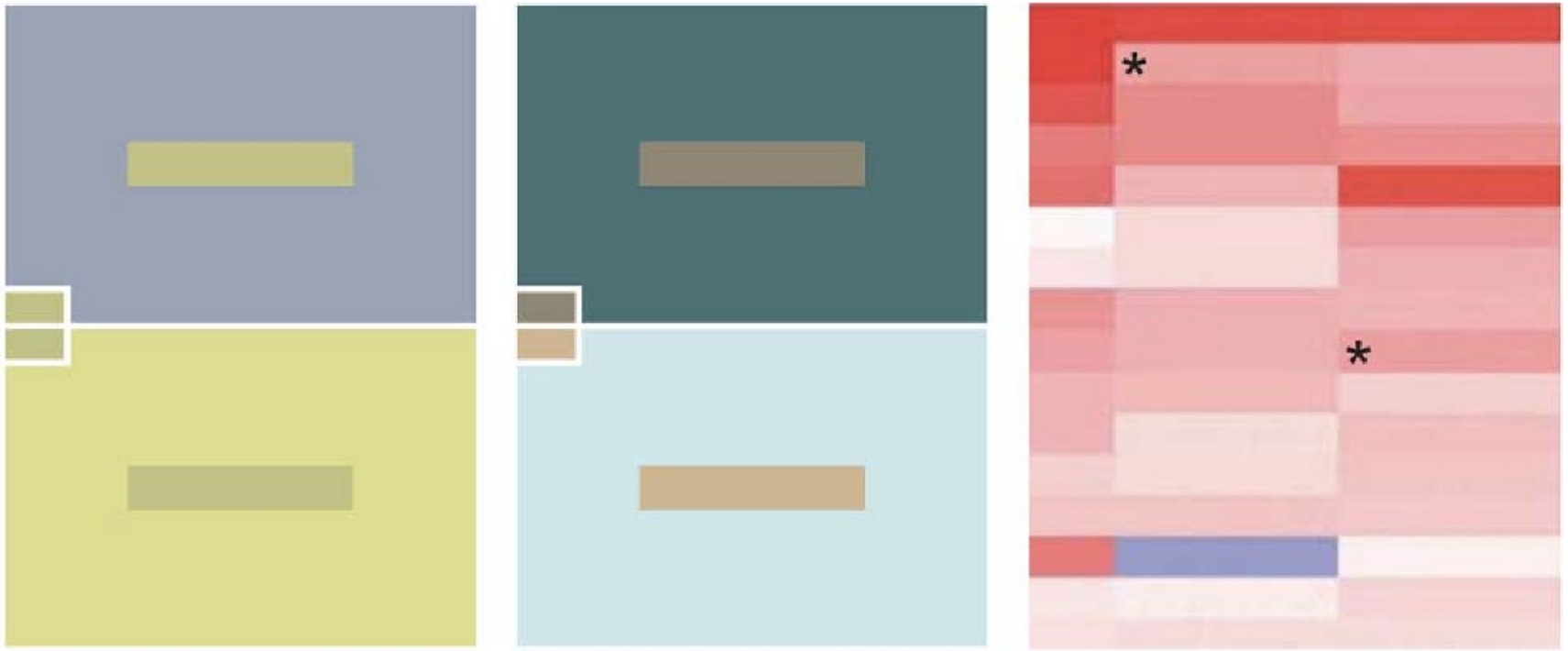


Normal Vision



Deuteranope Vision
("Red-Green Blindness")

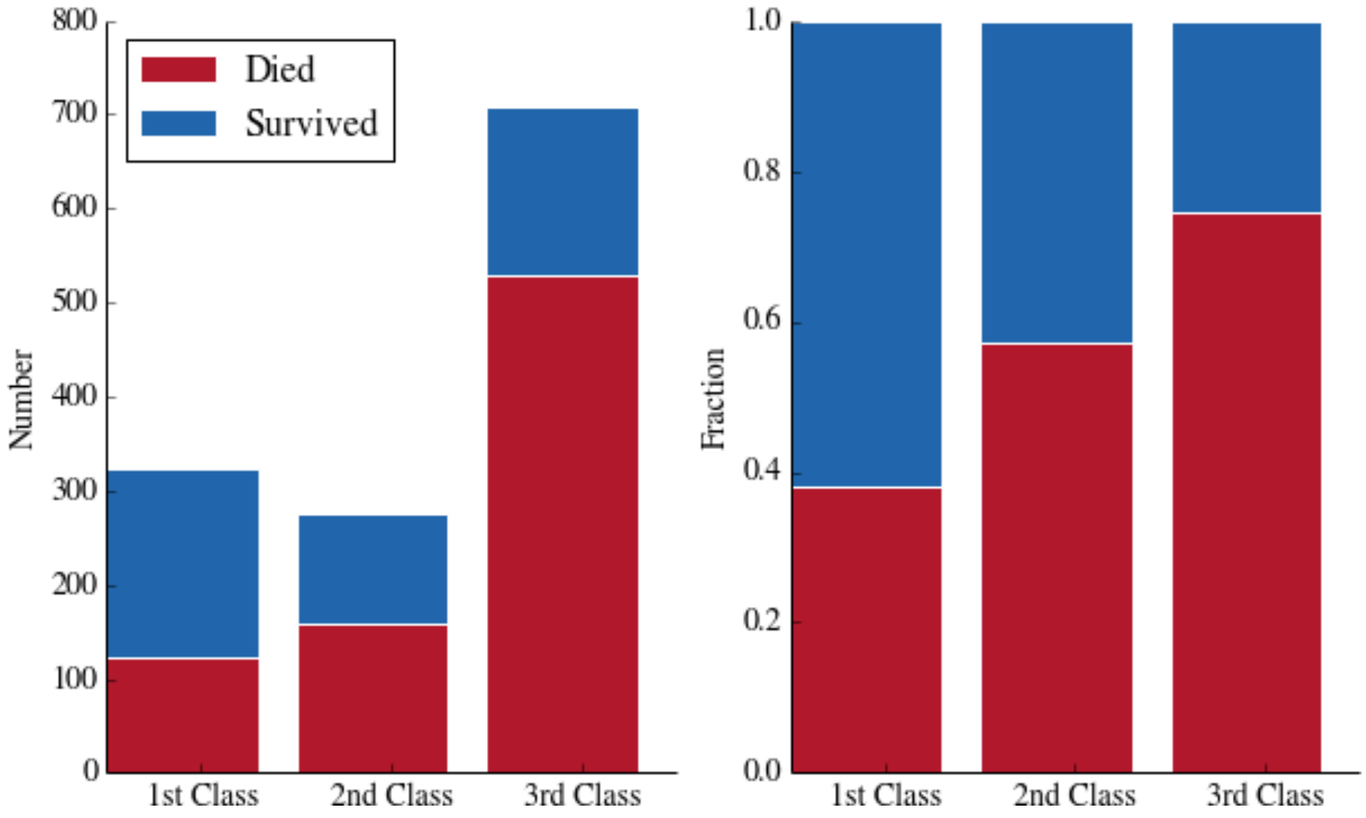
Color is relative!



Part of Whole

Stacked Bar Chart

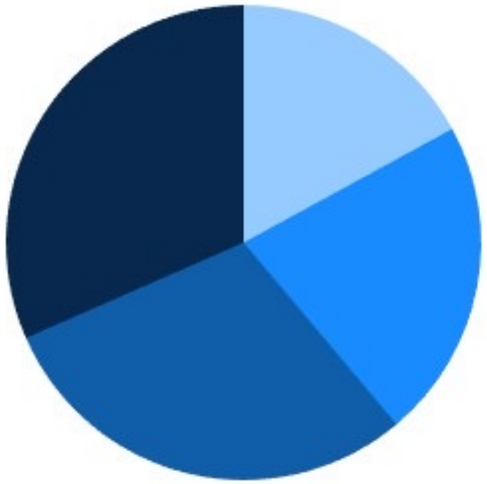
- Keys: Class, Survival Class is spatial Survival is color
- Left: absolute values
- Right: proportional values



Pie and Donut Charts

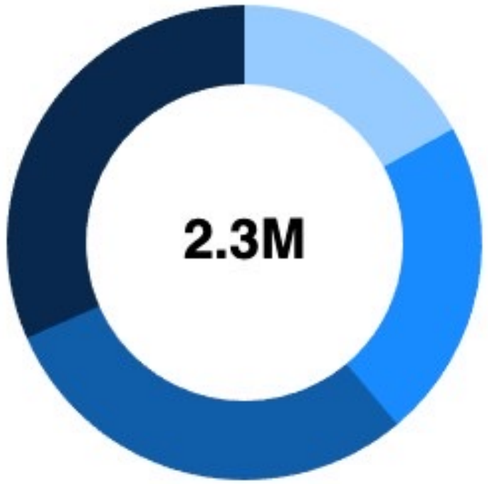
Pie Edit

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

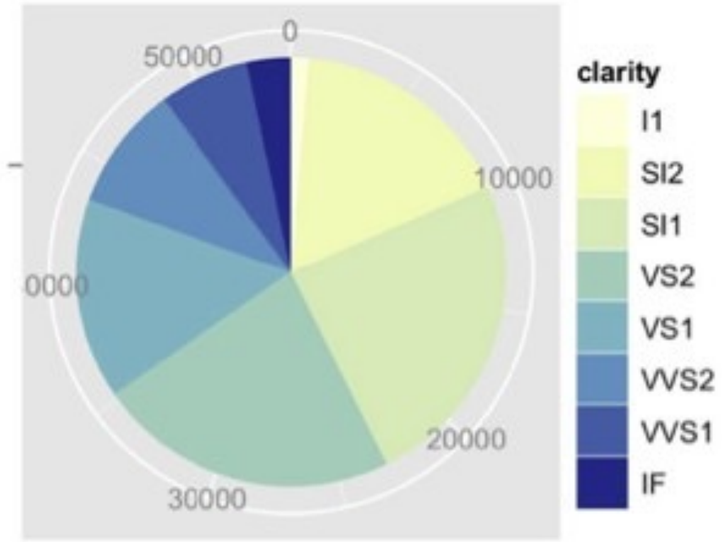


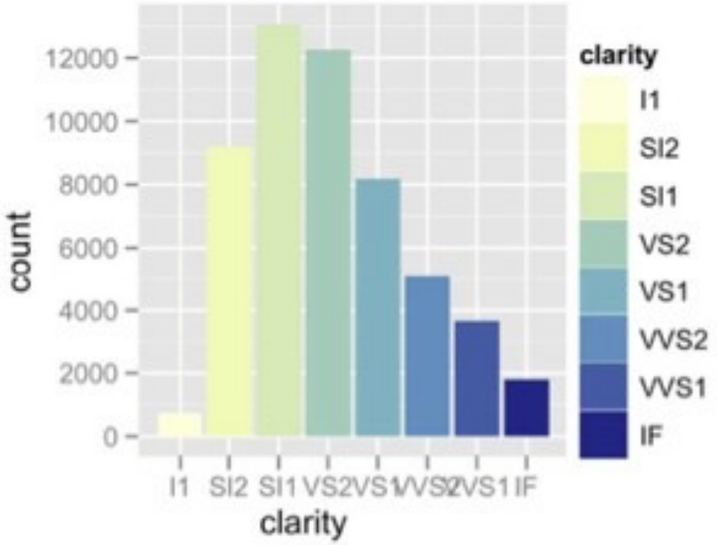
Donut

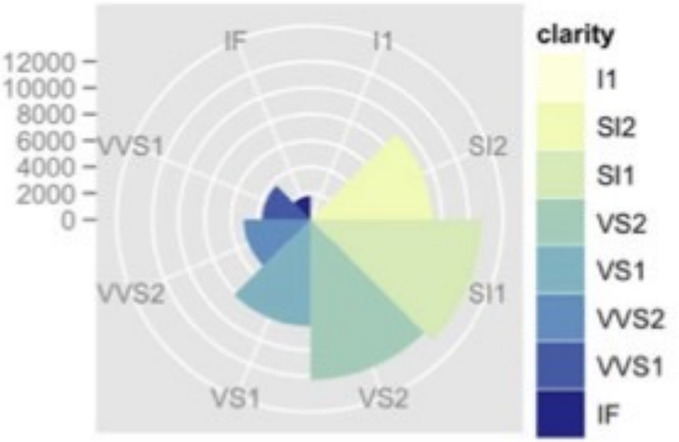
Similar to a pie chart - but the centre can be a good way of making space to include more information about the data (eg. total)



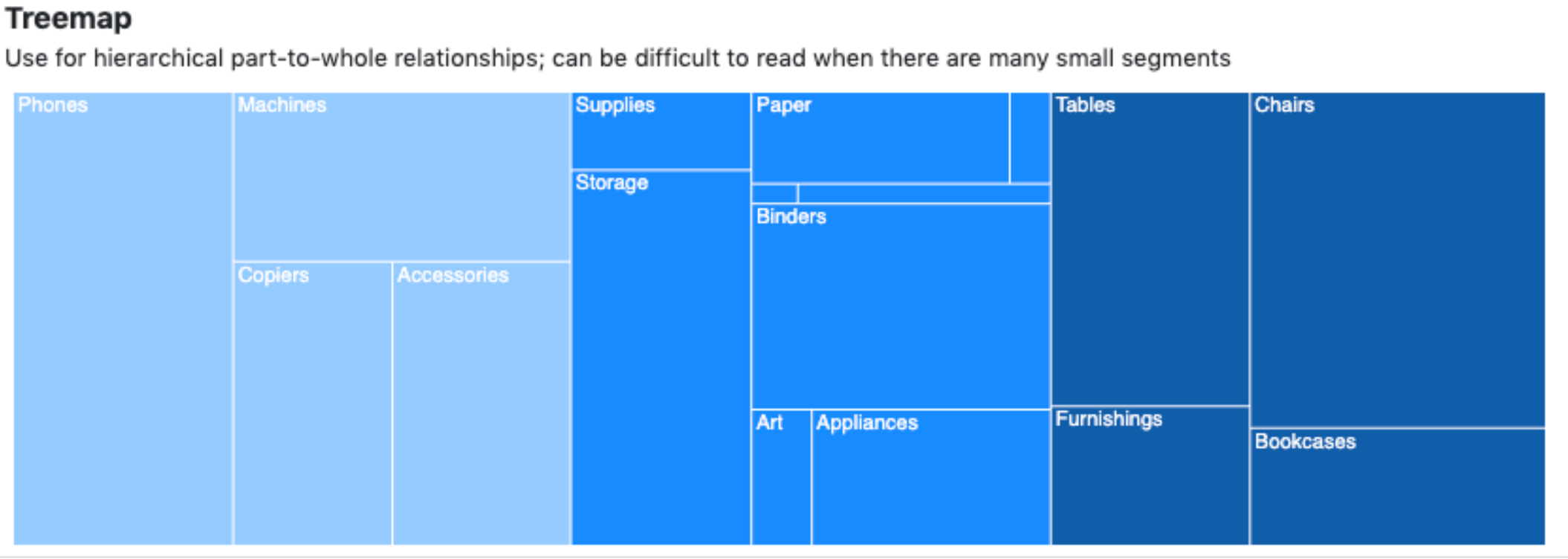
pie charts: take care with accuracy



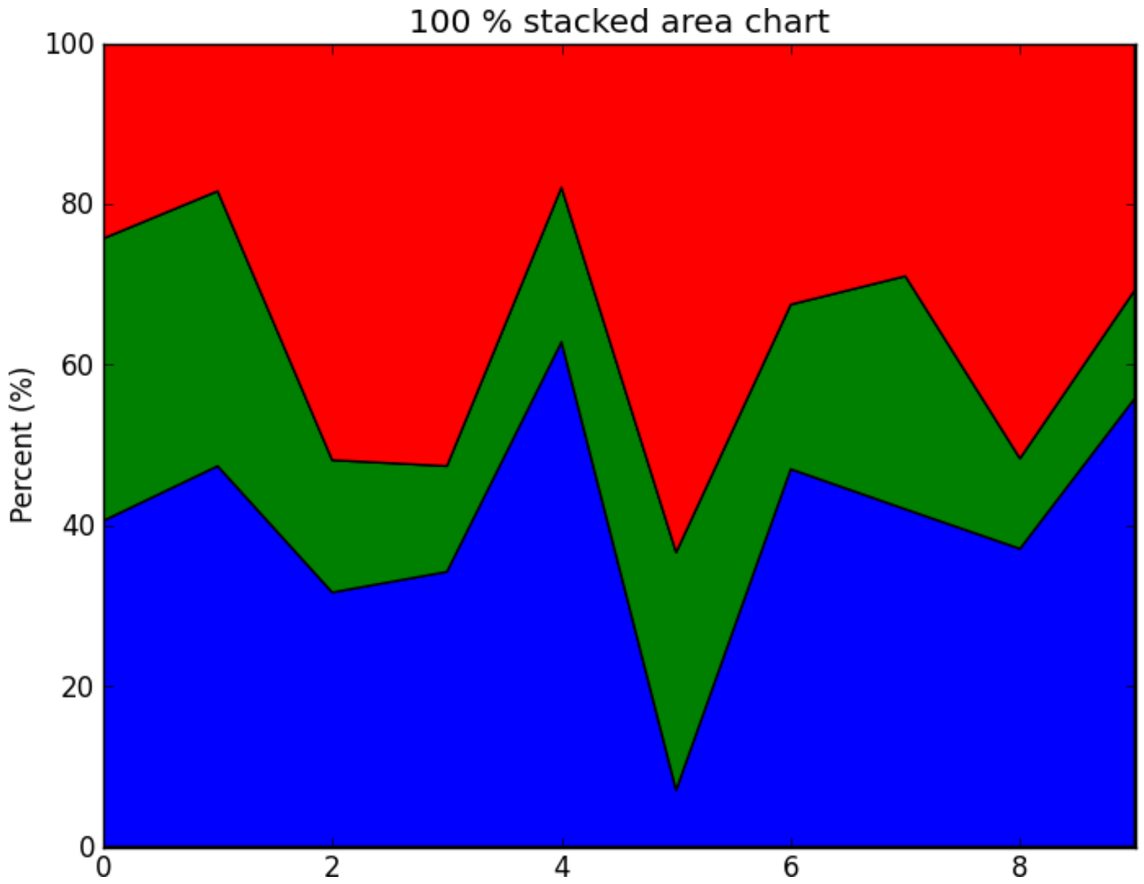




TreeMap


Treemap
Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments
Phones
Machines
Copiers
Accessories
Supplies
Storage
Paper
Binders
Art
Appliances
Tables
Furnishings
Chairs
Bookcases

Part of Whole for Time Series



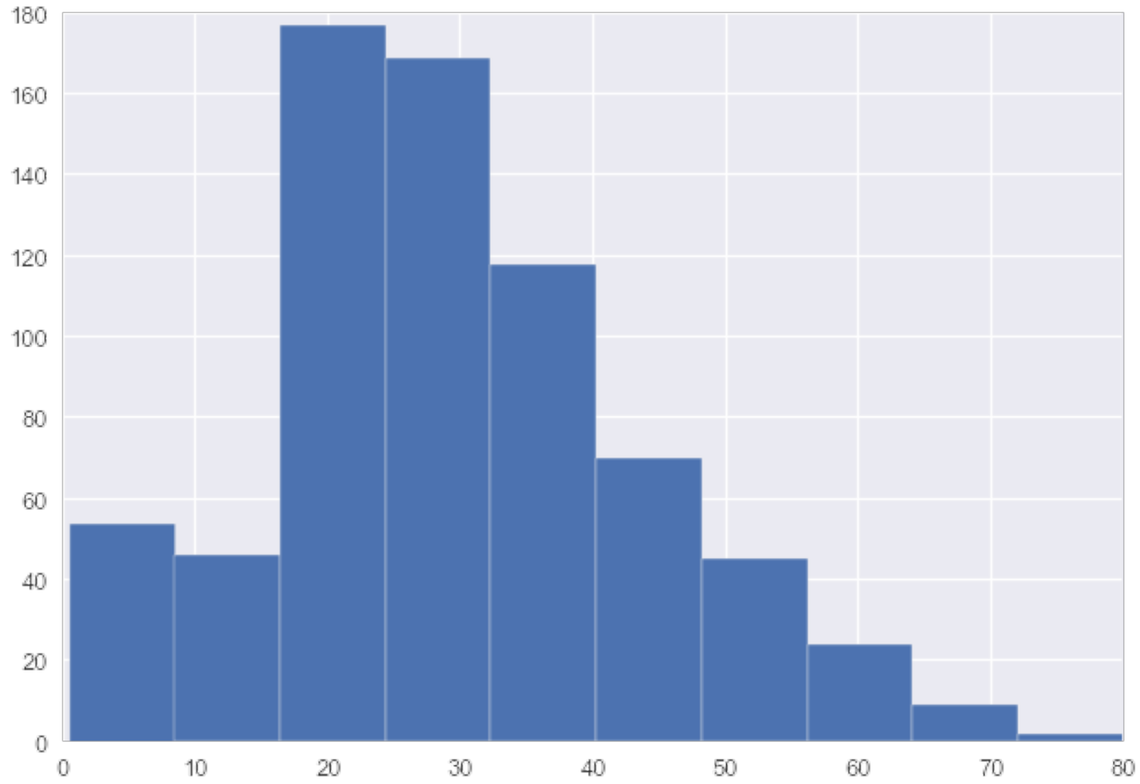
Distribution

Aggregating Large Data Vectors

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

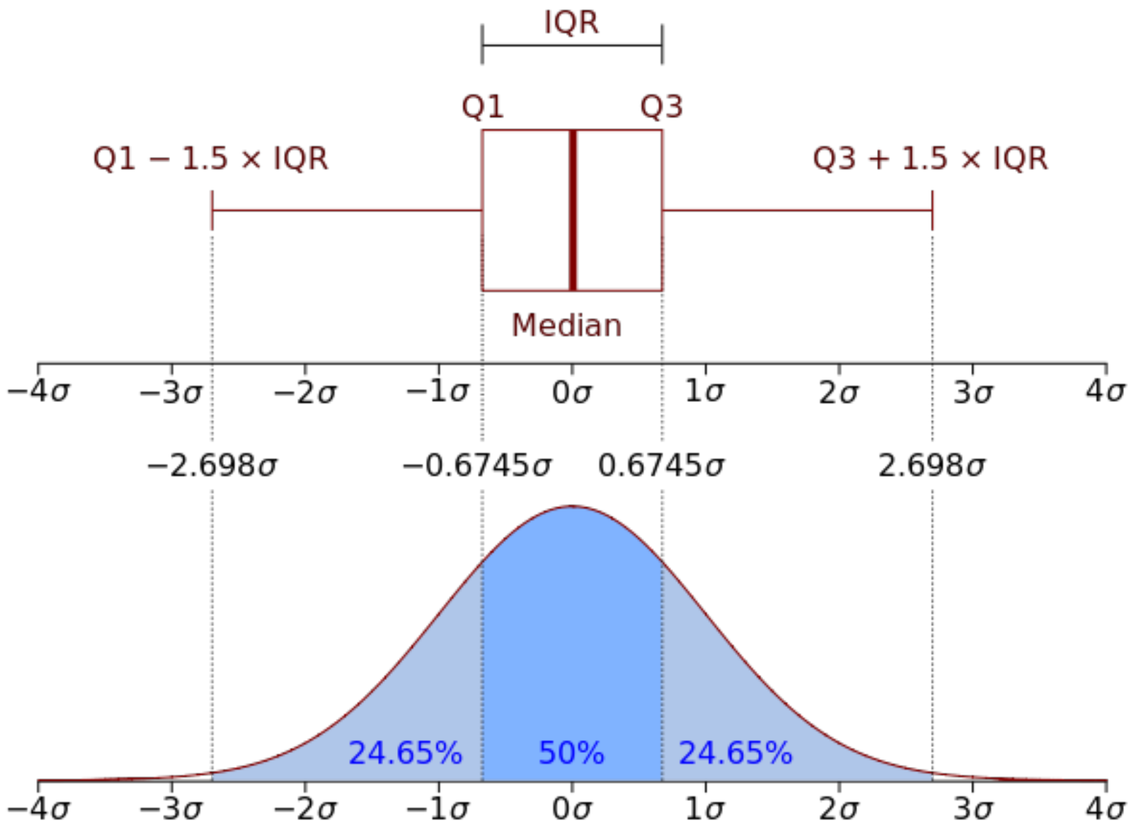
- Pro: compact representation
- Con: Works only if data is "well behaved" for the type of distribution visualization.

Histogram



Box Plots

- aka Box-and-Whisker Plot
- Bad for non-normal distributed data
- Especially bad for bi- or multi-modal distributions



One Boxplot, Four Distributions

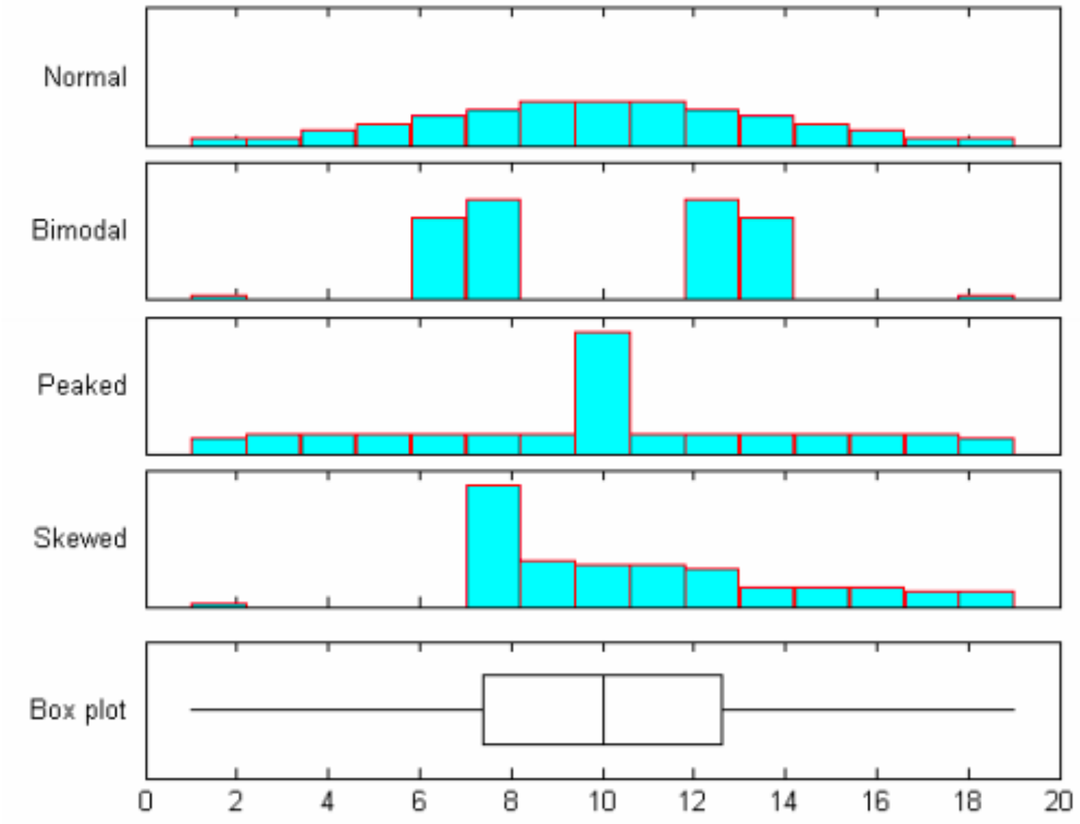
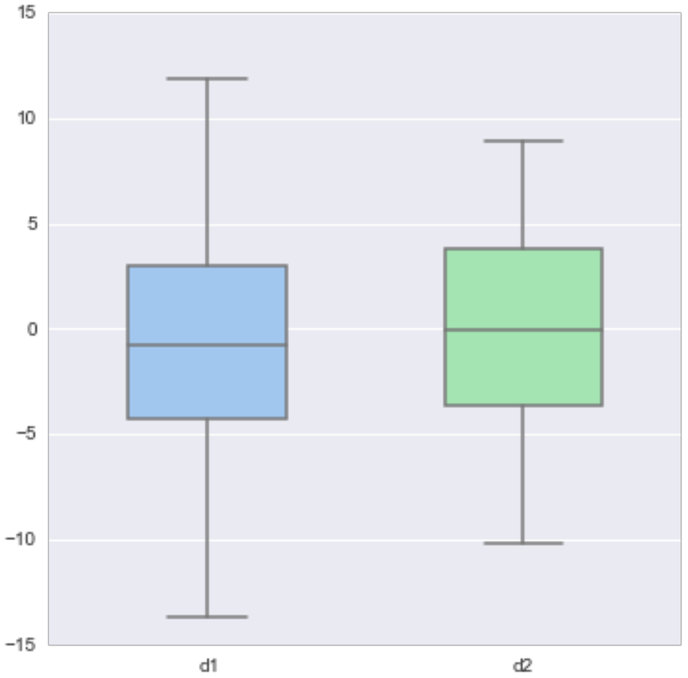


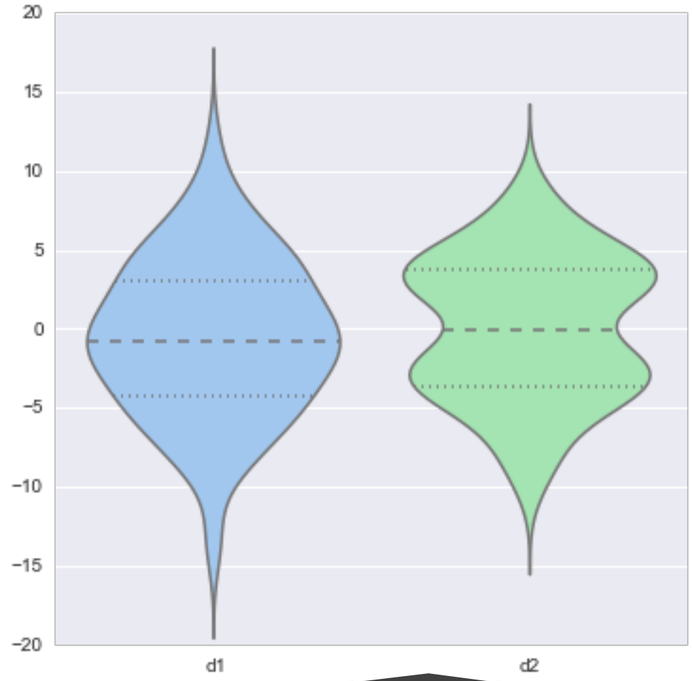
Figure 1: Histograms and box plot: four samples each of size 100

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

Violin Plot

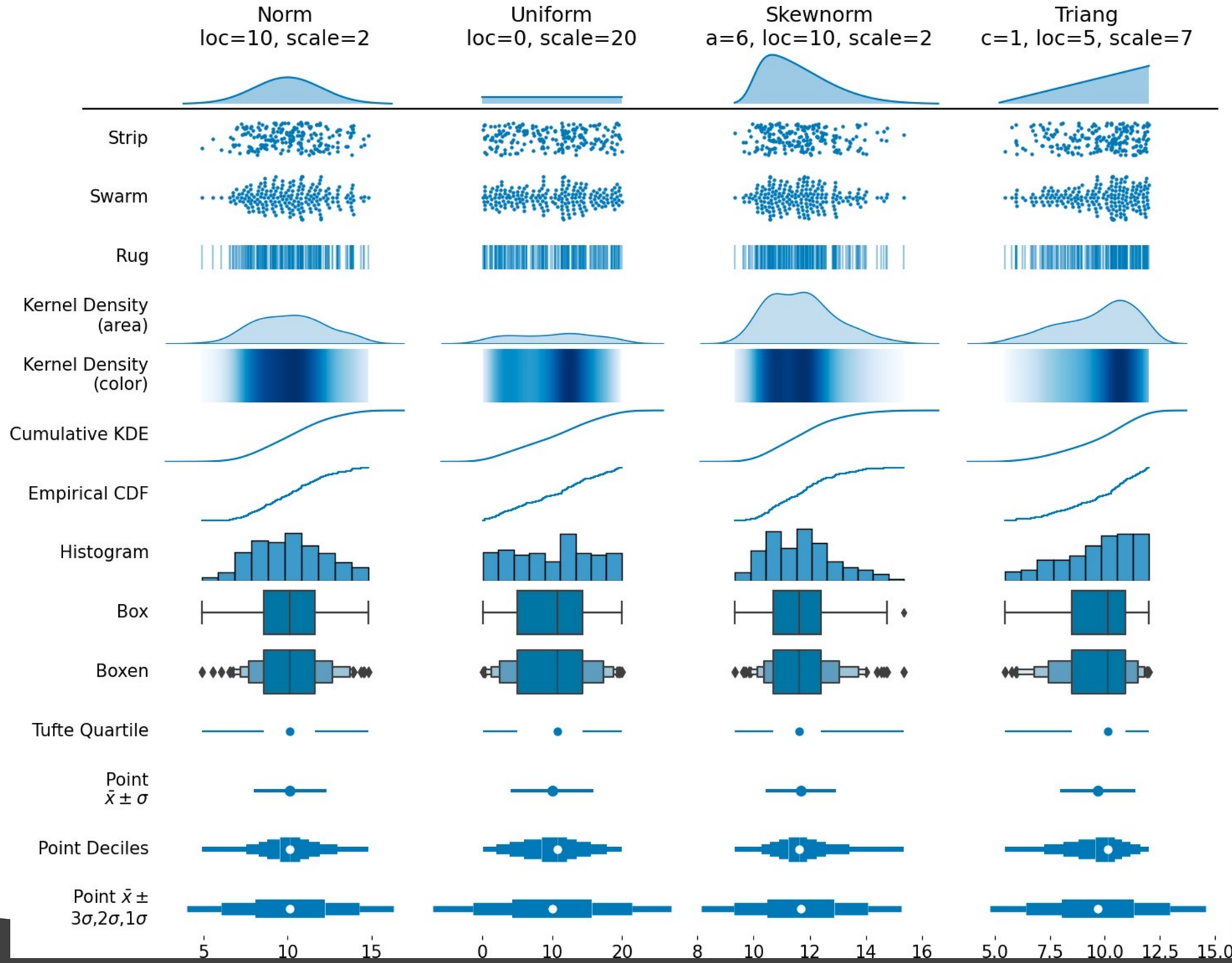
- = Box Plot + Probability Density Function





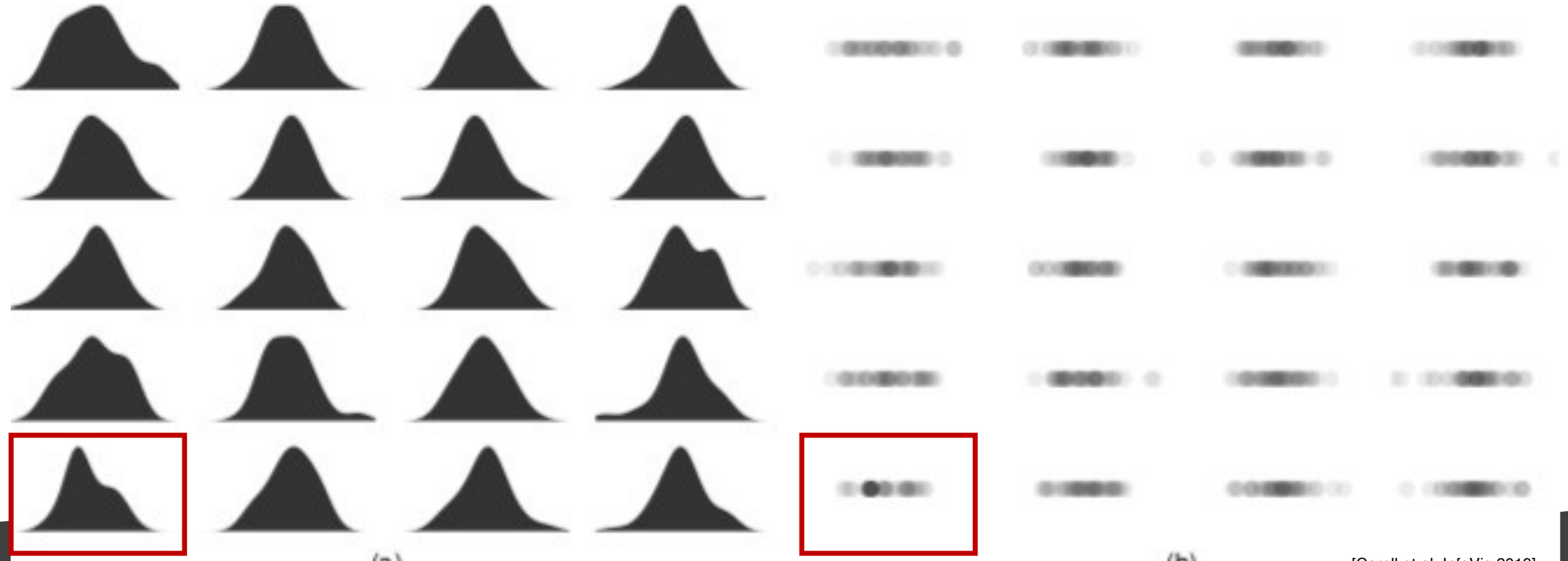
http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

A Collection of Univariate Plots



One of these things is not like the other…

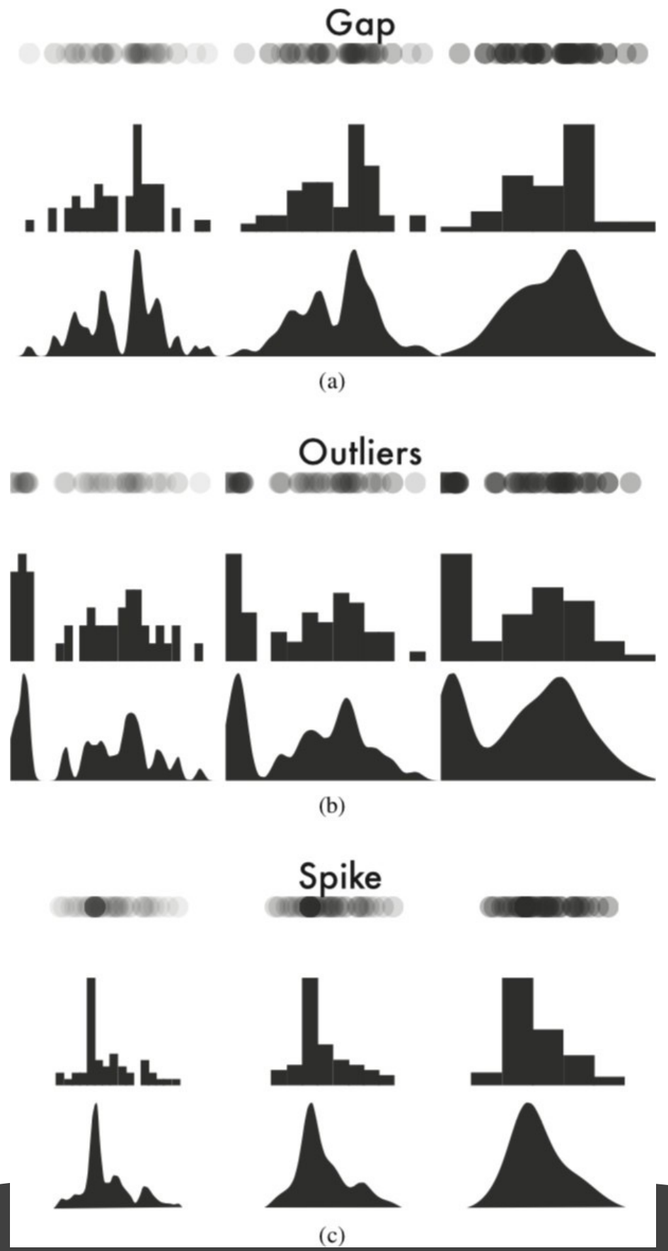
- 19 charts are random samples from a gaussian
- 1 chart has 20% of samples with identical value



[Corell et al, InfoVis 2019]

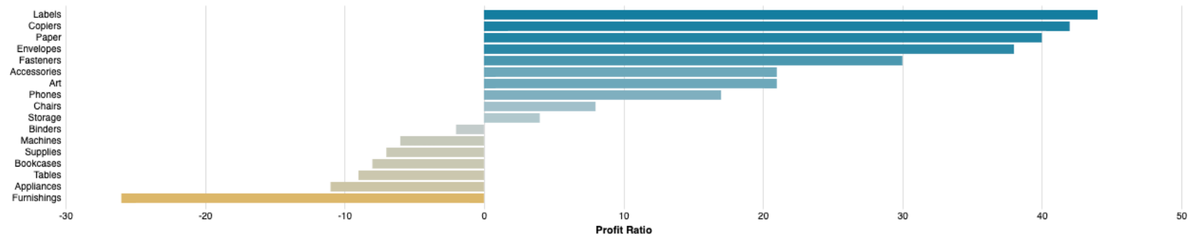
Detecting Data Flaws

- Tricky with aggregate visualization
- Bin size / kernel type / bandwidth / visualization choice all affect different situations

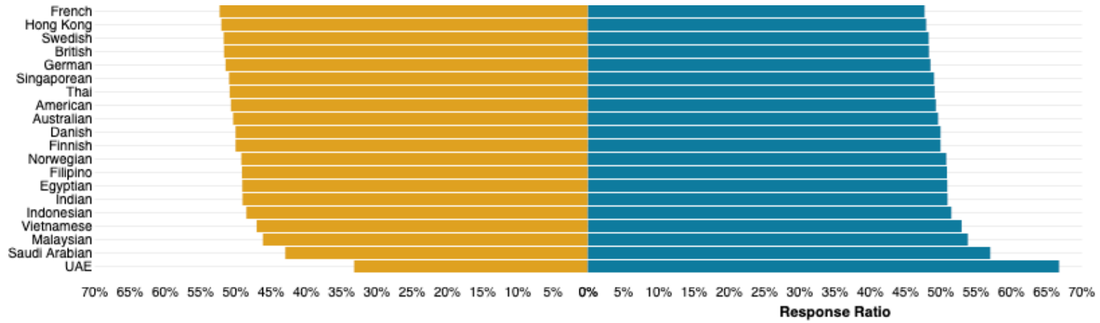


Deviation

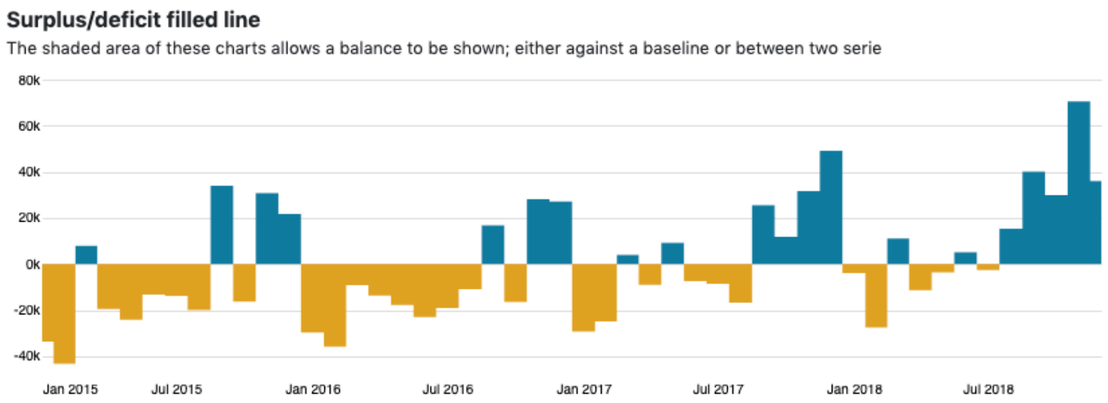
Comparison to Reference Point

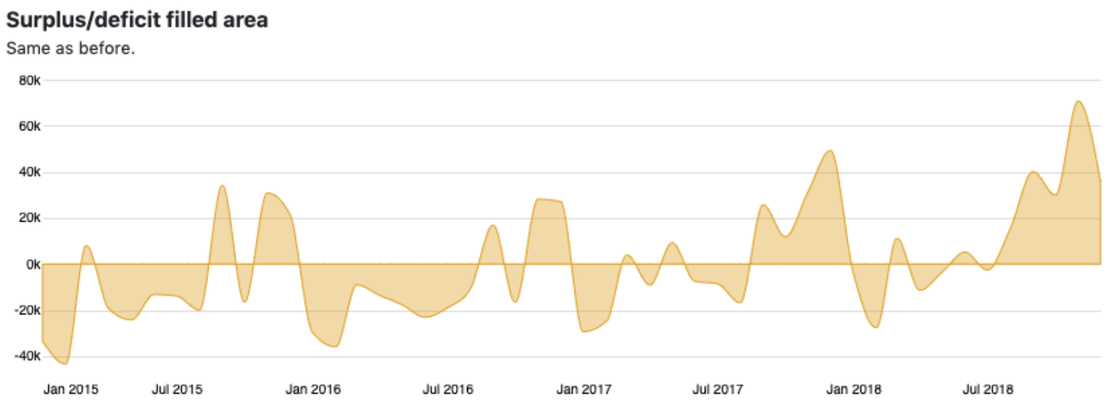


Diverging Bar Chart



Juxtaposing Two Variables (male/female)

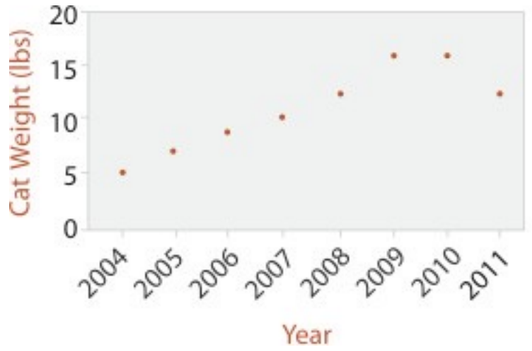


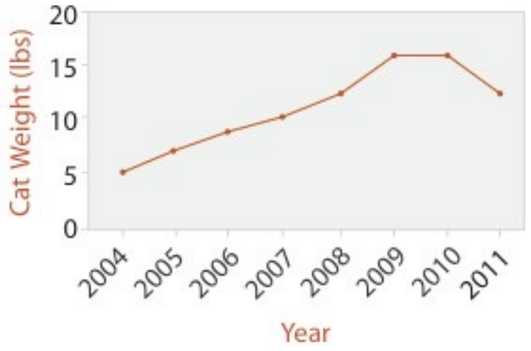


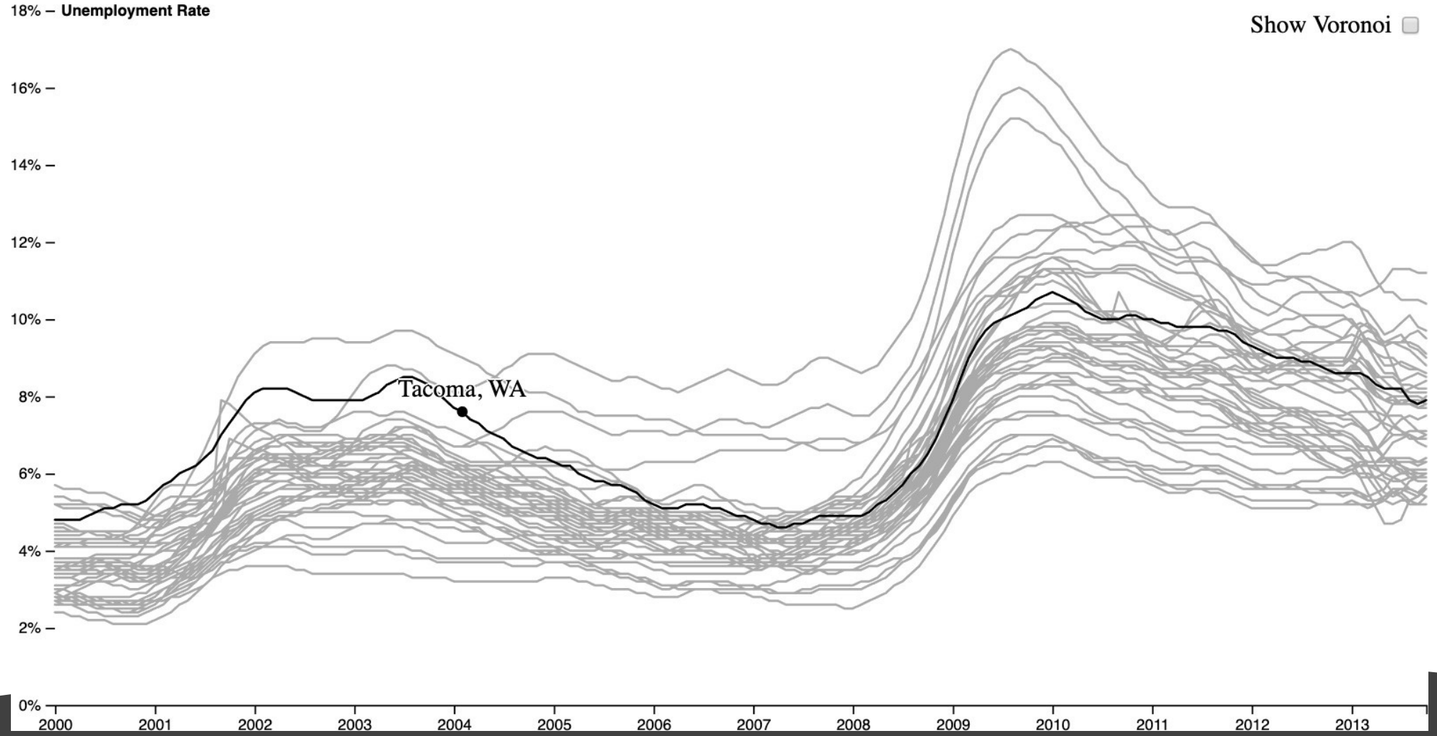
Change over Time

Line Chart

- Simple
- Familiar
- Accurate
- Fairly Scalable

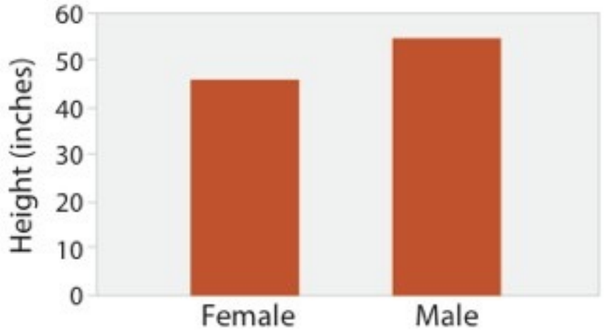






don’t use line charts for categorical attributes!

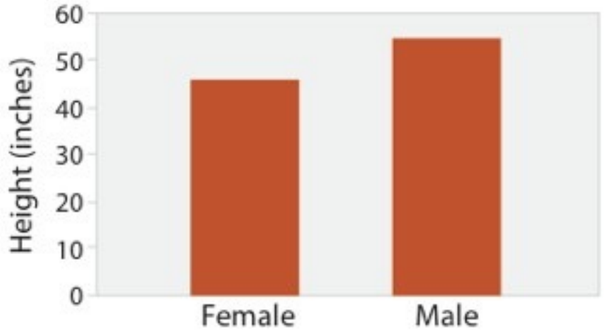
ok: “Men are taller than women (on average)”



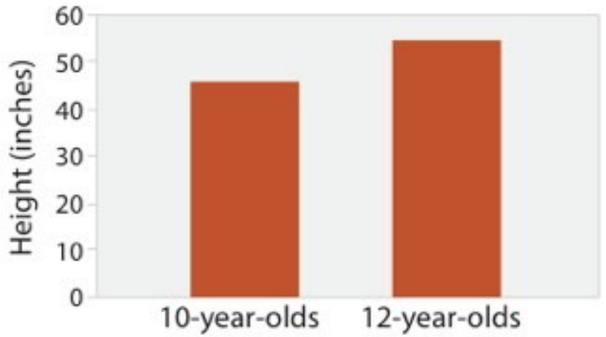
bad: “The more male a person is, the taller he/she is”

don’t use line charts for categorical attributes!

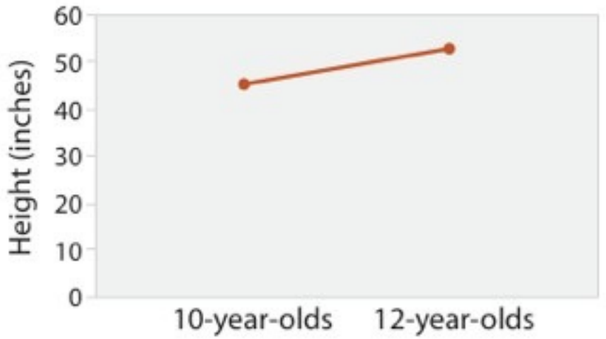
ok: “Men are taller than women (on average)”



bad: “The more male a person is, the taller he/she is”

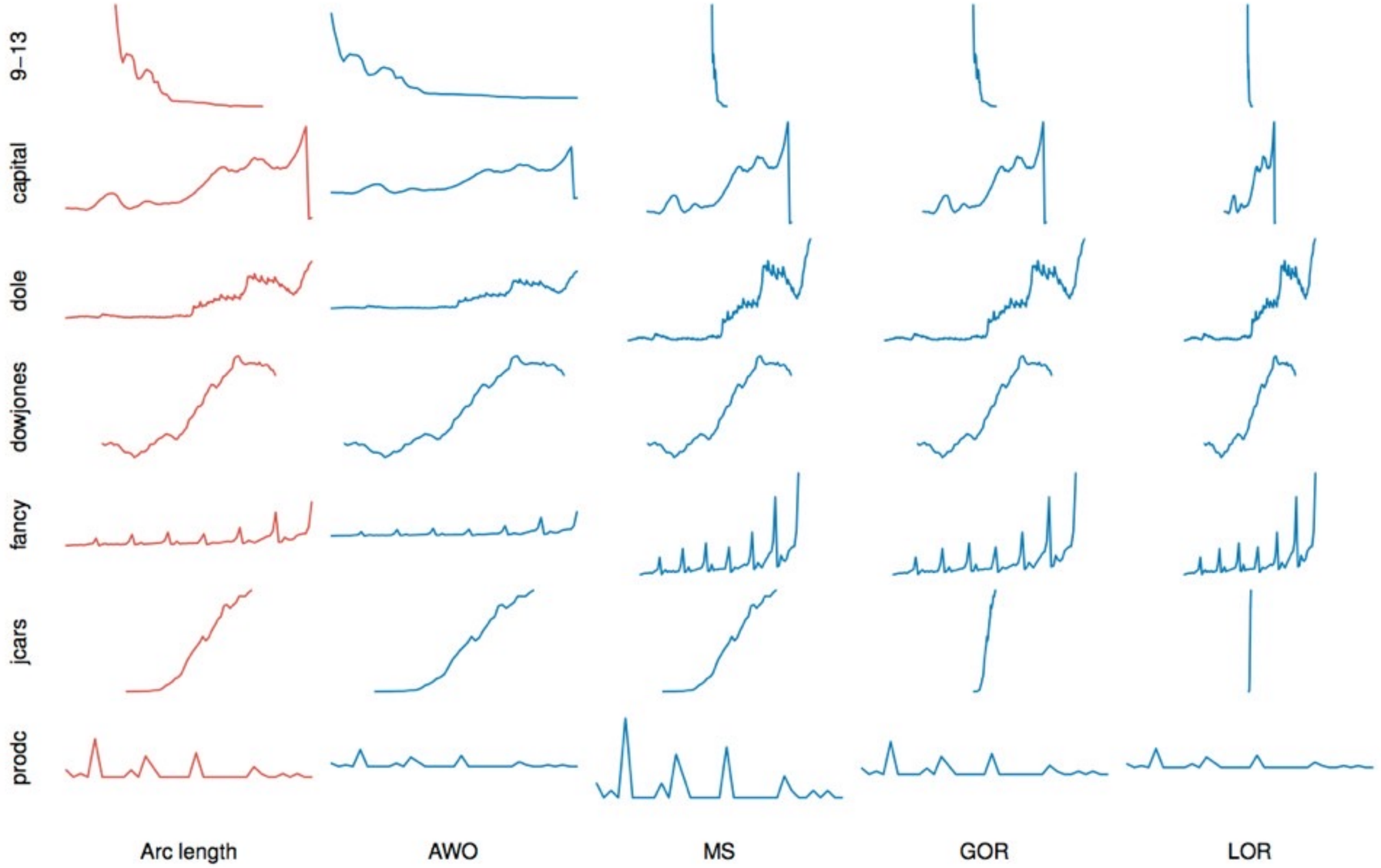


ok: “Twelve year olds are taller than ten year olds”



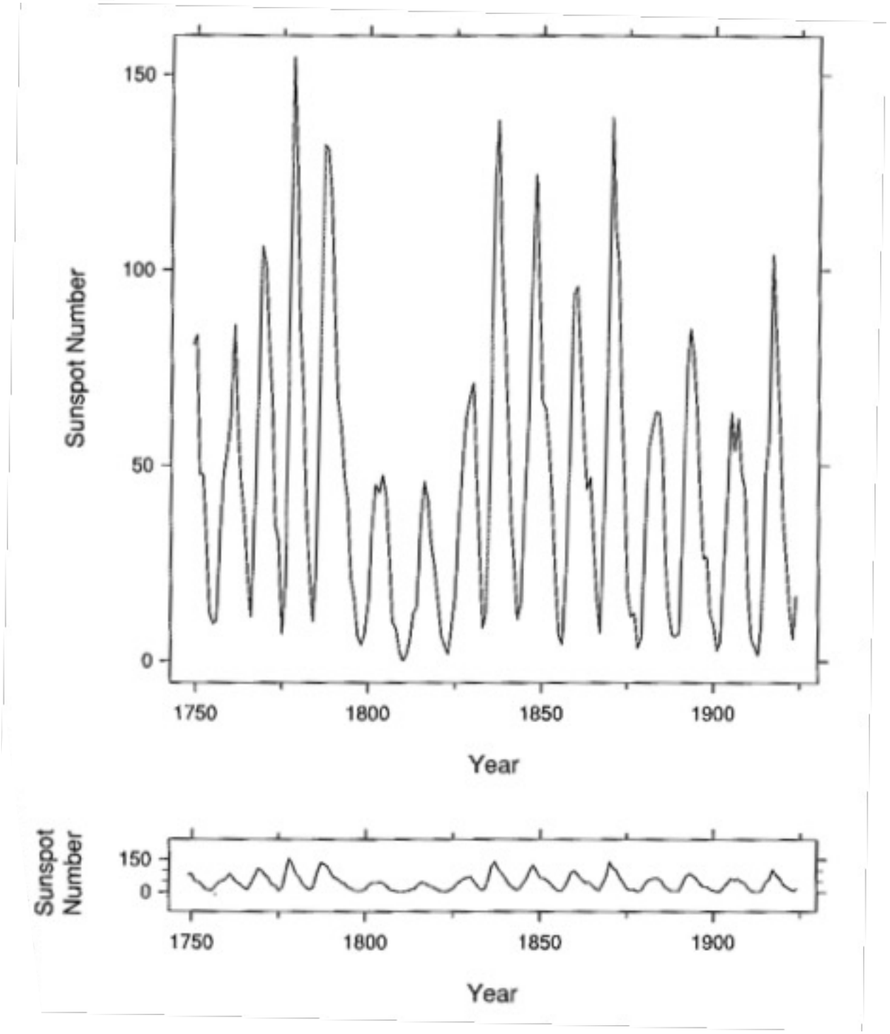
ok: “Height increases with age”

Aspect ratio



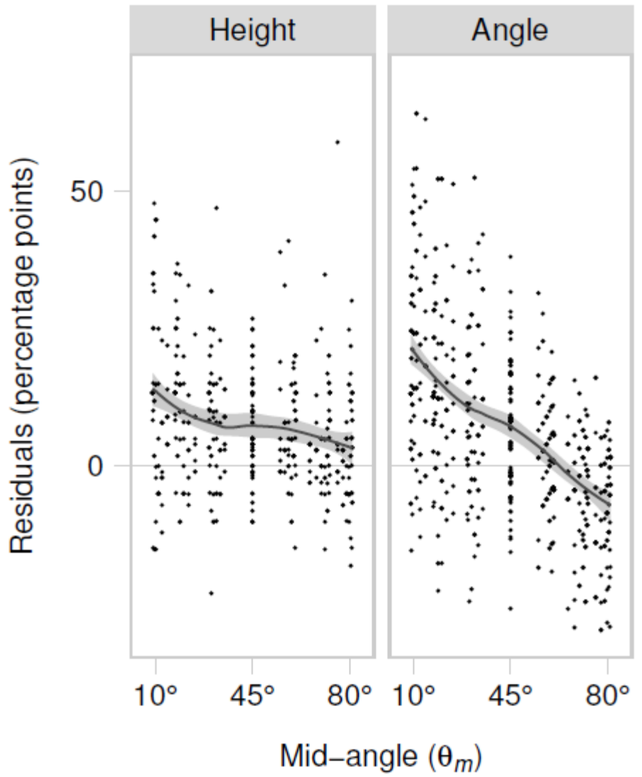
Aspect Ratio [Cleveland 1994]

- Bank to 45°
  - The aspect ratio of a graph is an important factor for judging rate of change.
  - perceptual principle: most accurate angle judgment is at 45°



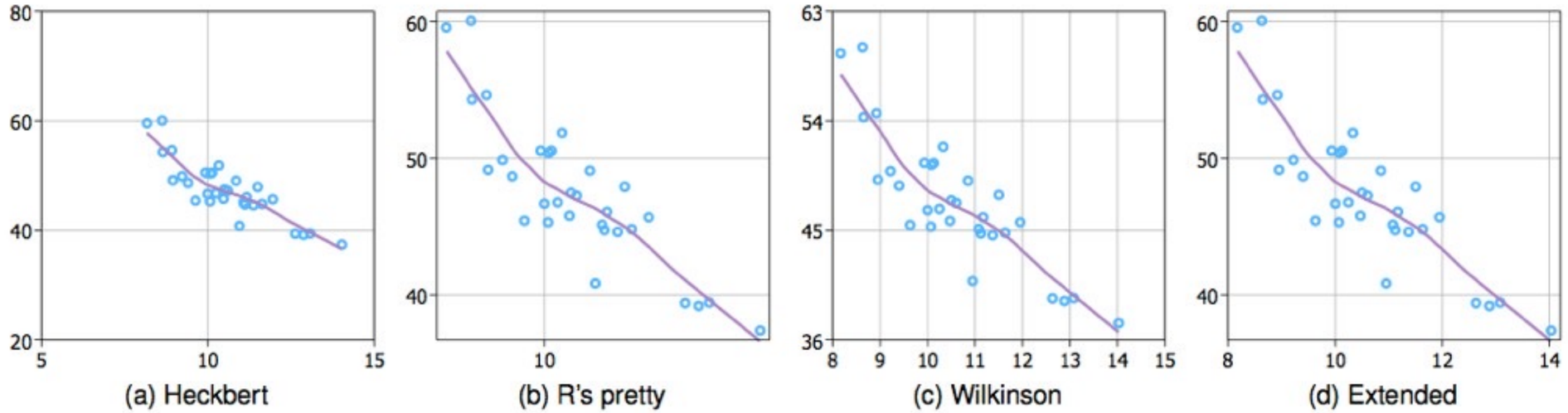
Counter-Point: Talbot 2012

- people use two different strategies to estimate slope—angle and height
- slope angle accuracy NOT minimized at 45° (closer to 60°)

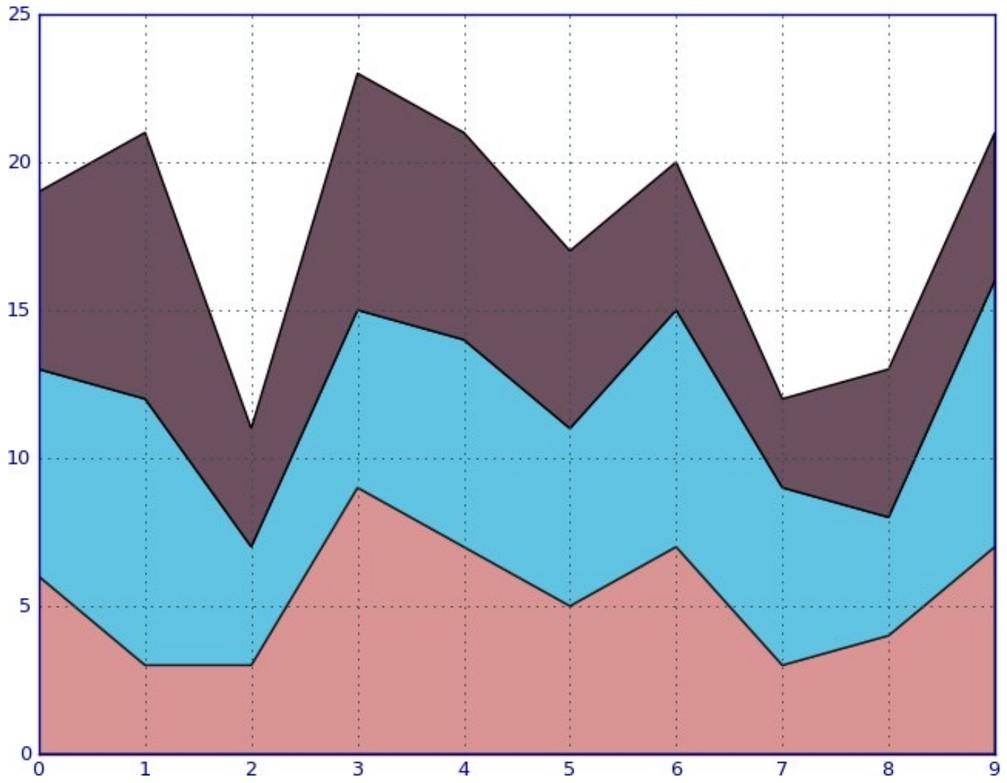


Tick Placement

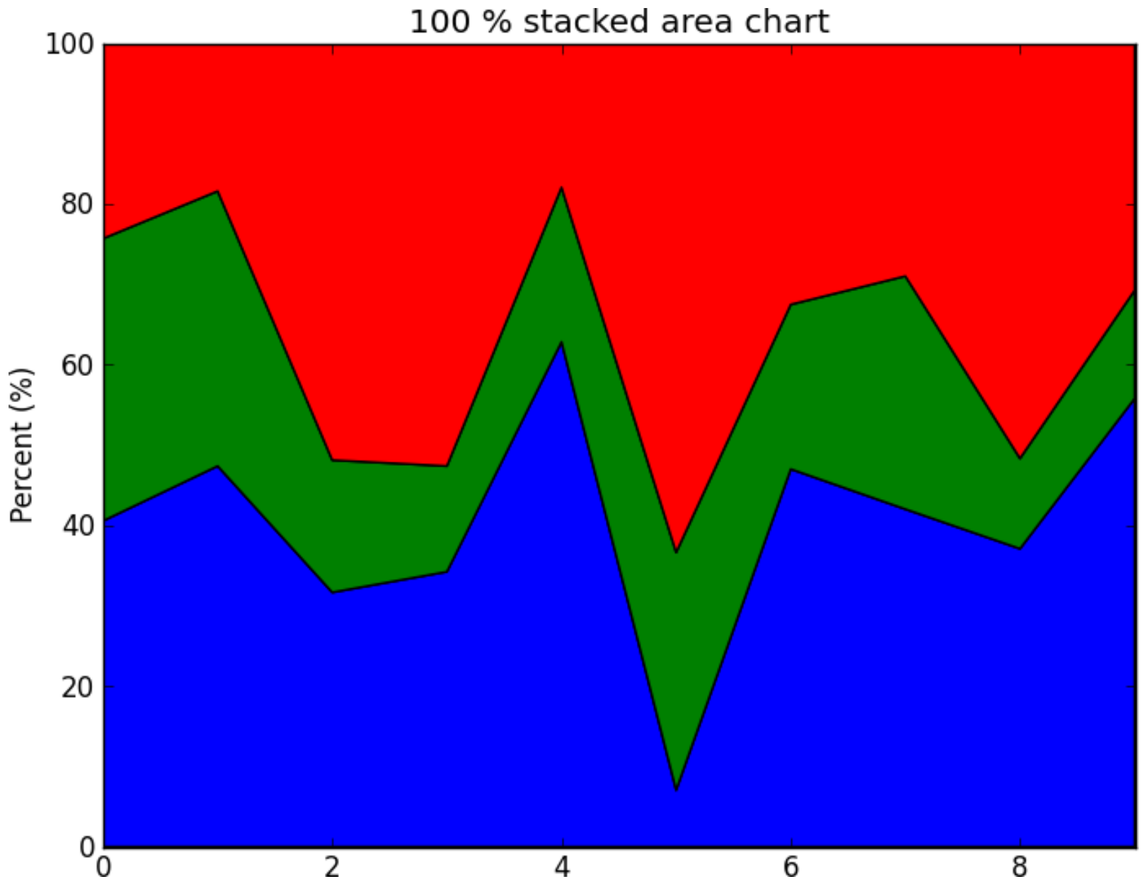
- Ticks help in user interpretation of data, but too much may hinder
- Automatic optimization of label formatting, font size, and orientation
  - placement based on simplicity, coverage, granularity, and legibility



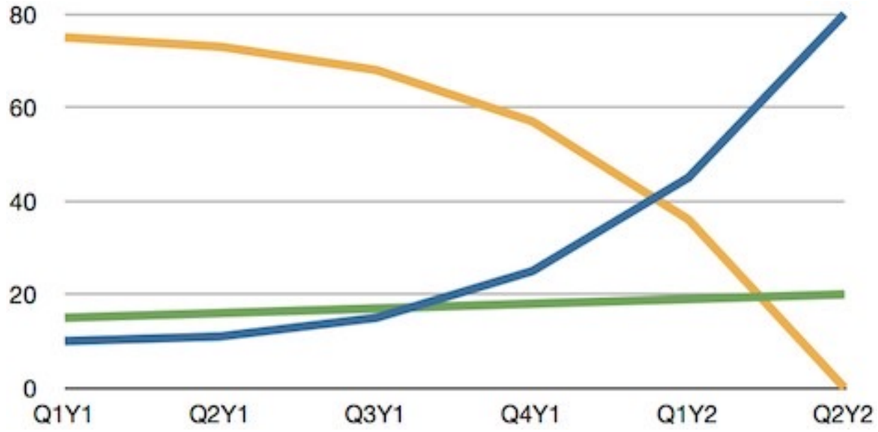
Stacked Area Chart



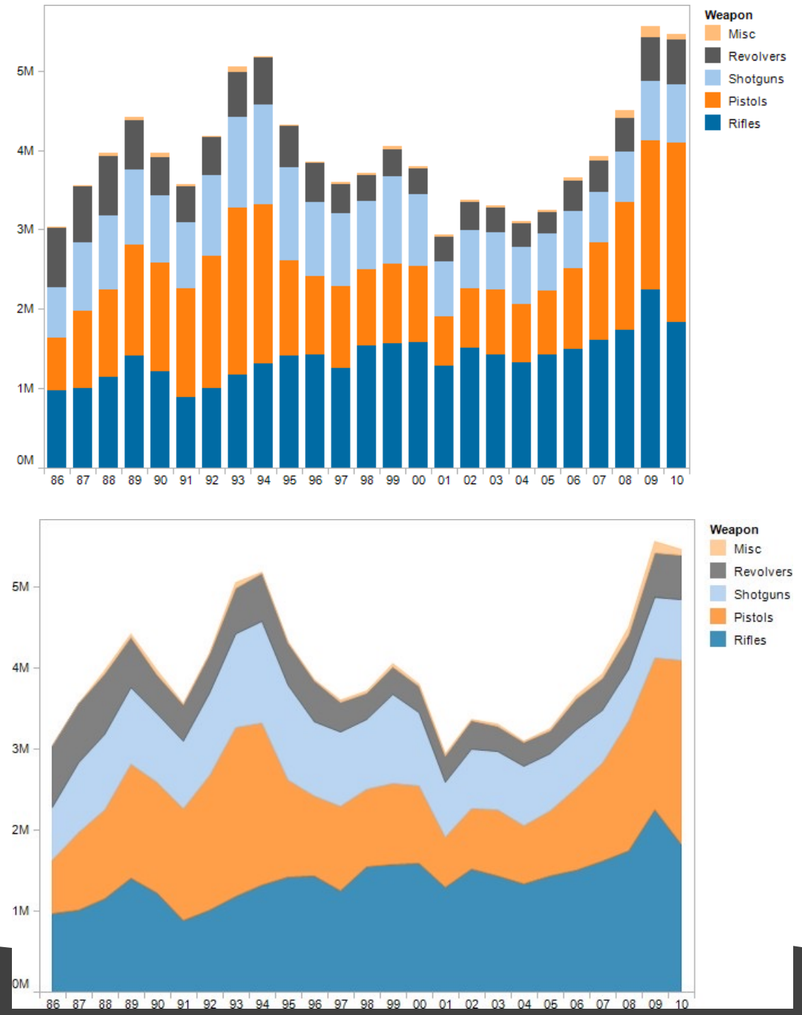
100% Stacked Area Chart

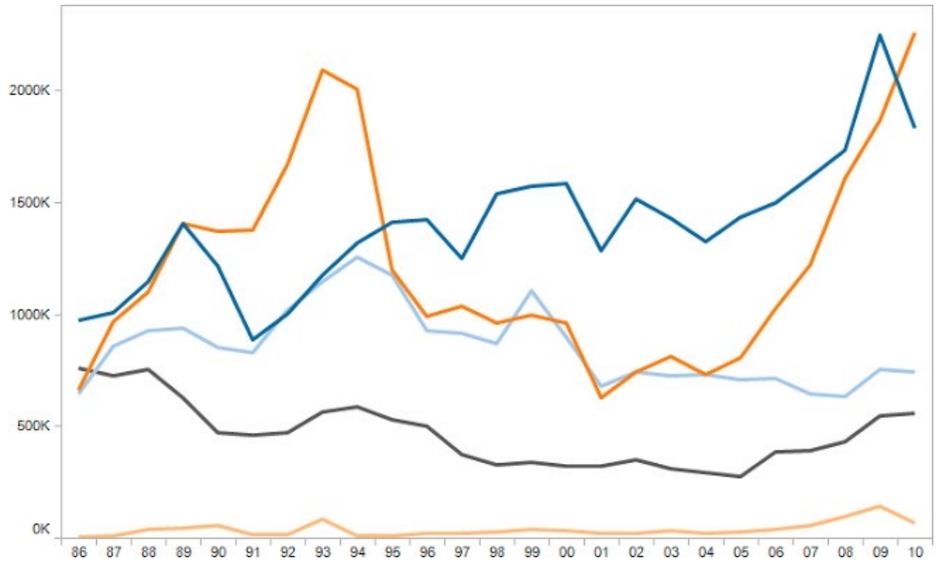


Stacked Area vs. Line Graphs



Can you spot the trends?
Overall vs Individual Components





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

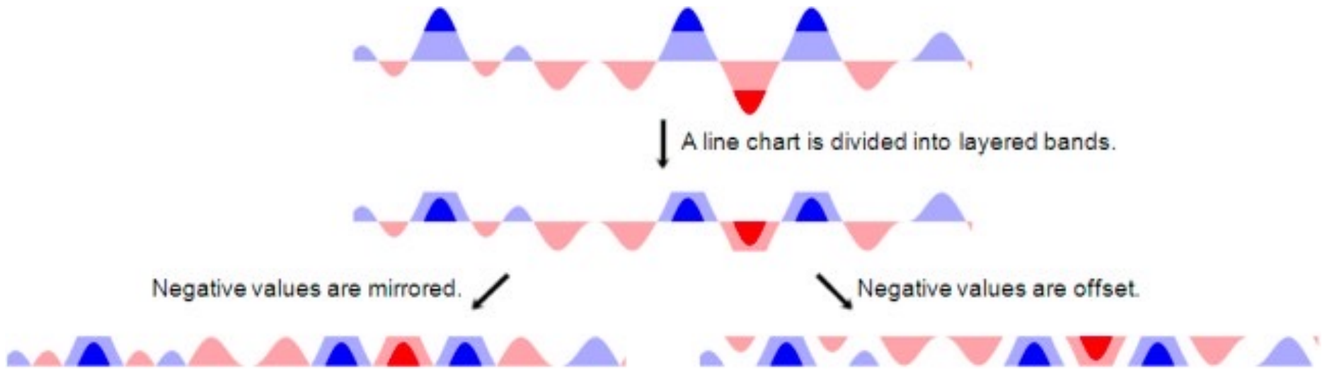
The FTSE100 Brexit bounce 5562 7501 continues one year on from the vote last summer

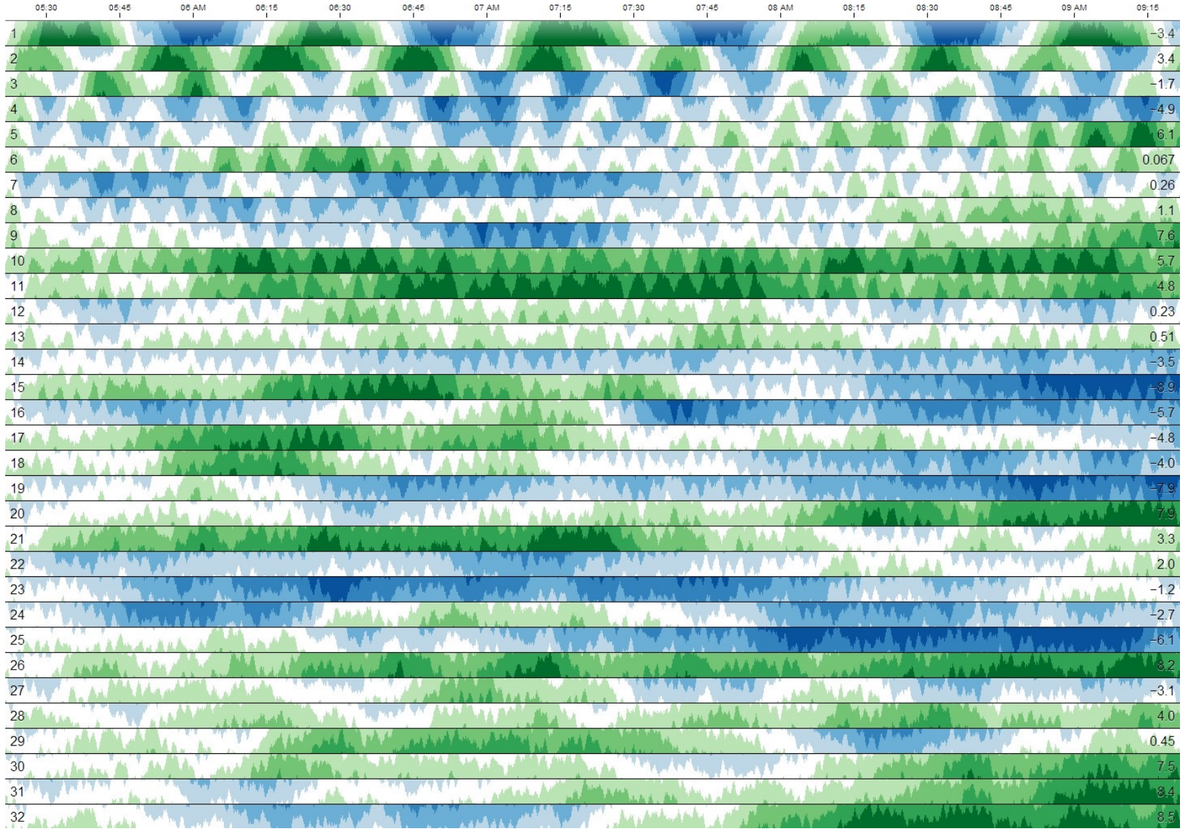
Symbol	Bid	Ask	Last	Change	T	Chart	Volume	High	Low	Value Change		Value	Gain	
DELL	89 3/4	89 13/16	89 3/4	+ 1 1/4			10,310,100	90 1/8	88 1/2	+1.41%	250	17,950	+273.72%	13,147
CPQ	48 7/16	48 9/16	48 7/16	- 13/16			25,628,700	51 1/4	1/4	-1.65%	-81	4,844	+60.79%	1,831
SDTI	26 1/4	26 3/8	26 3/8	+ 1/2			504,600	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/

By Peter Zelchenko

Horizon Graphs





[Heer, Sizing the Horizon, 2009]

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

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

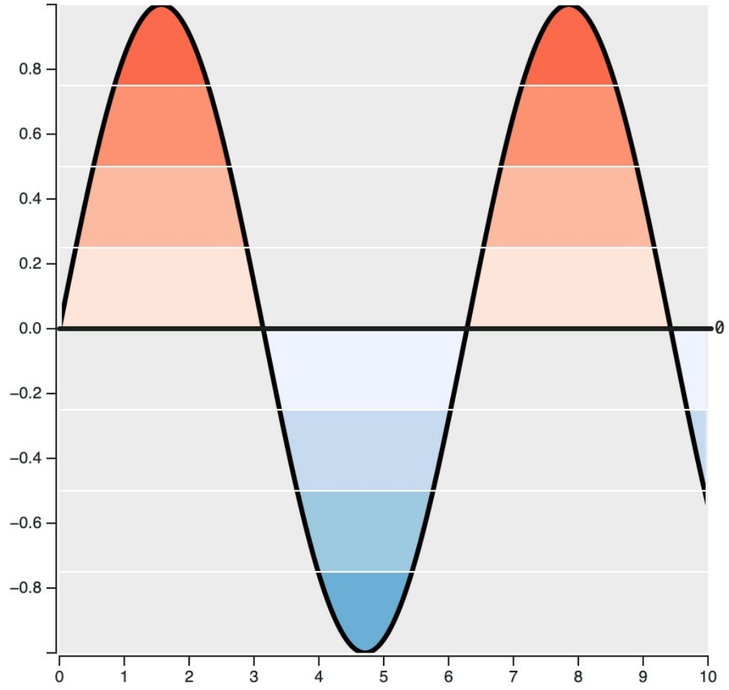
Row Height: 40

Match Row Height

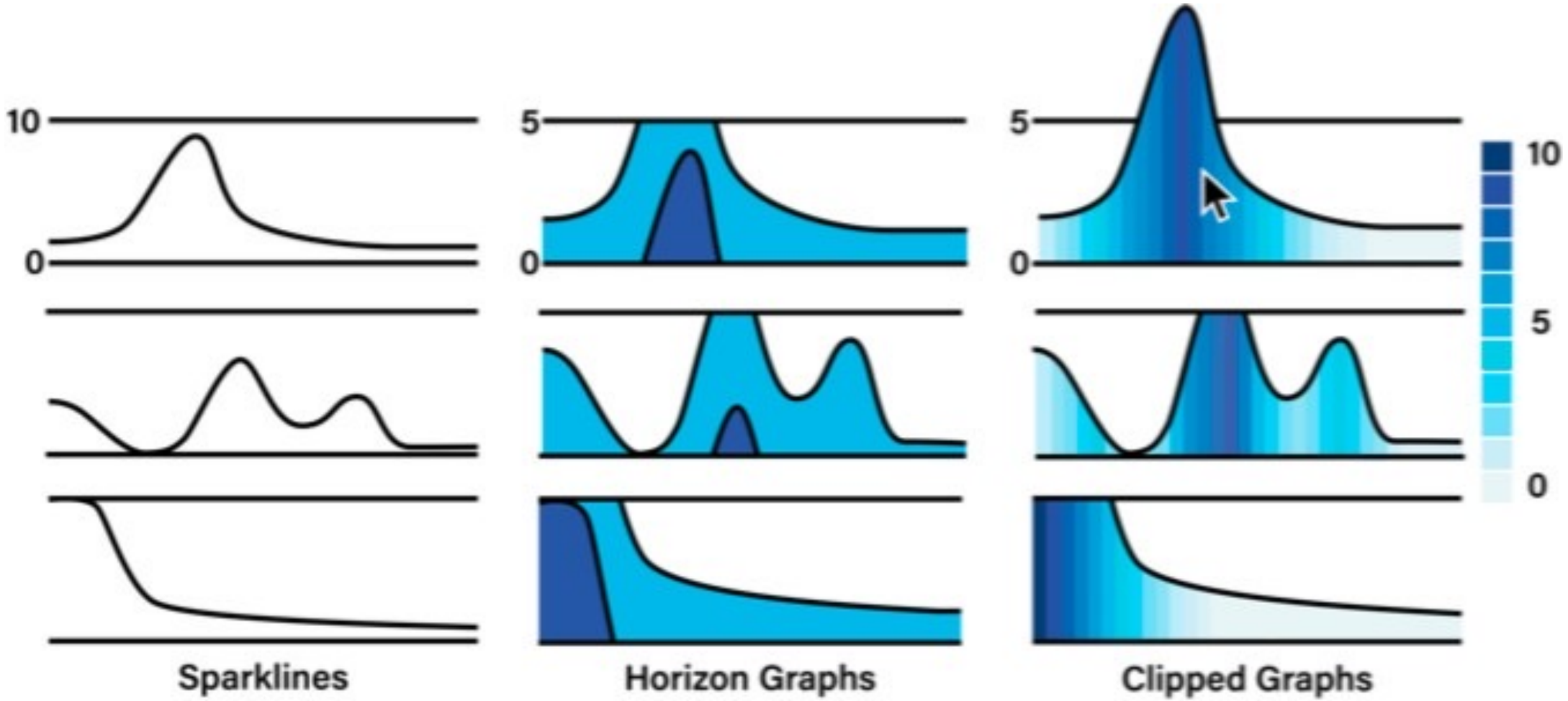
Horizon Chart

Press and hold to stack!

Explanation Chart

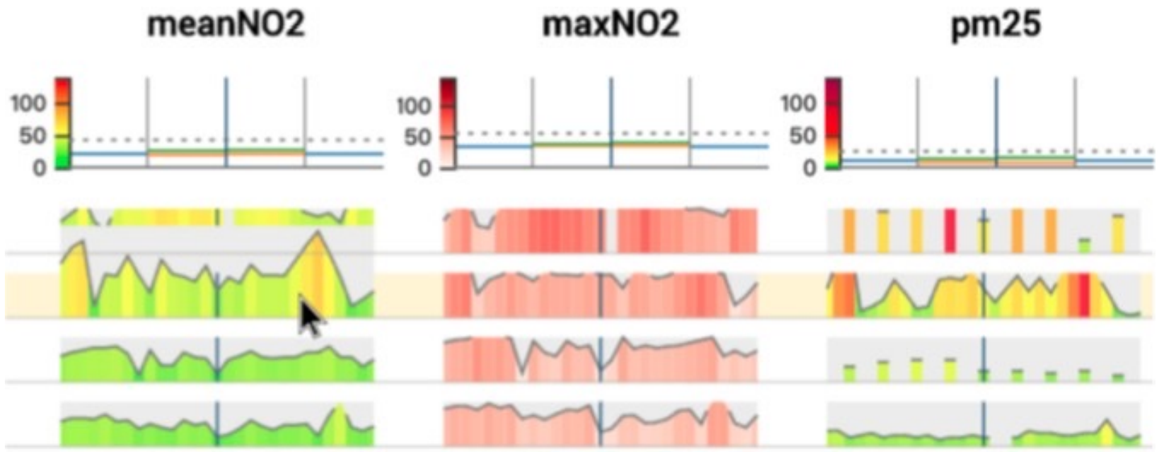


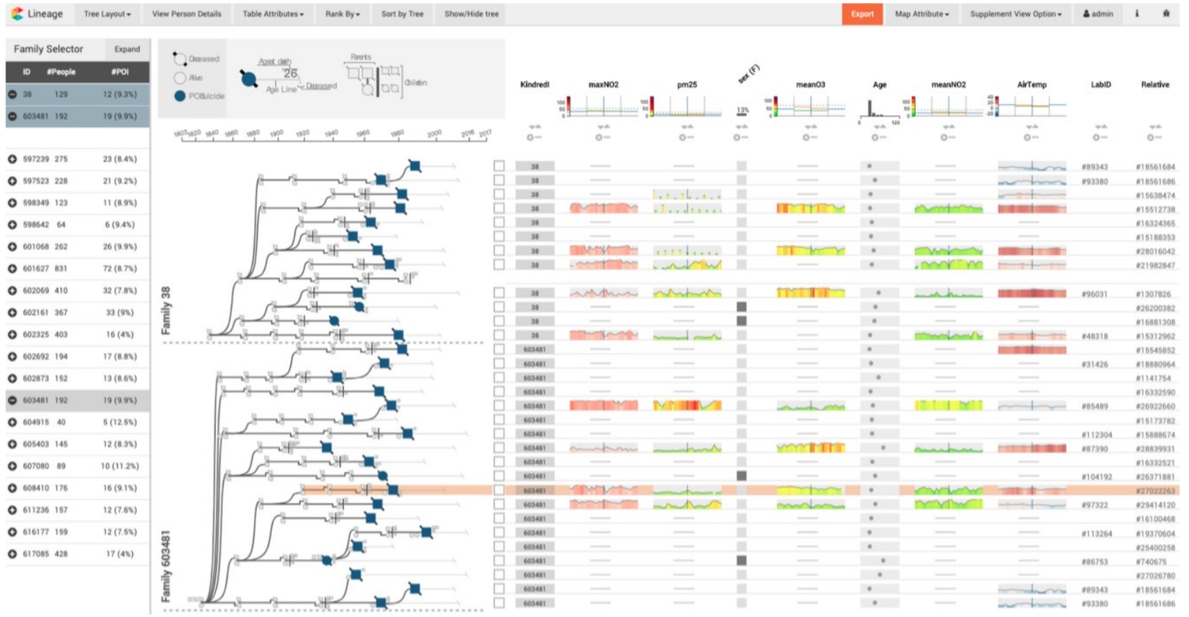
Clipped Graphs



[Lin 2019]

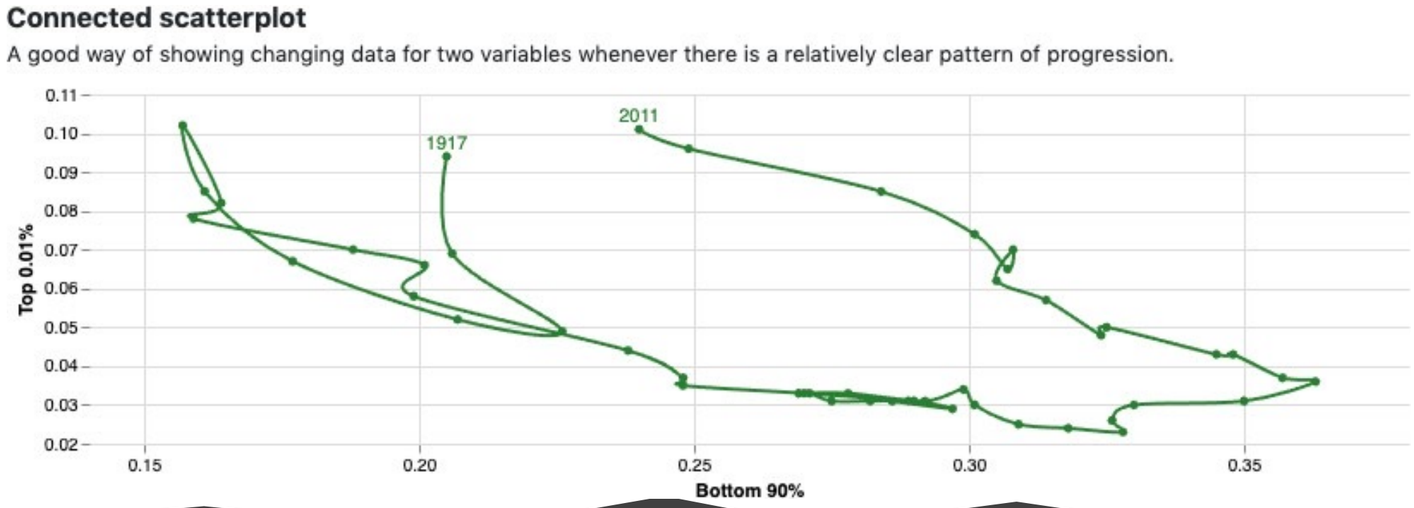
Clipped Graphs

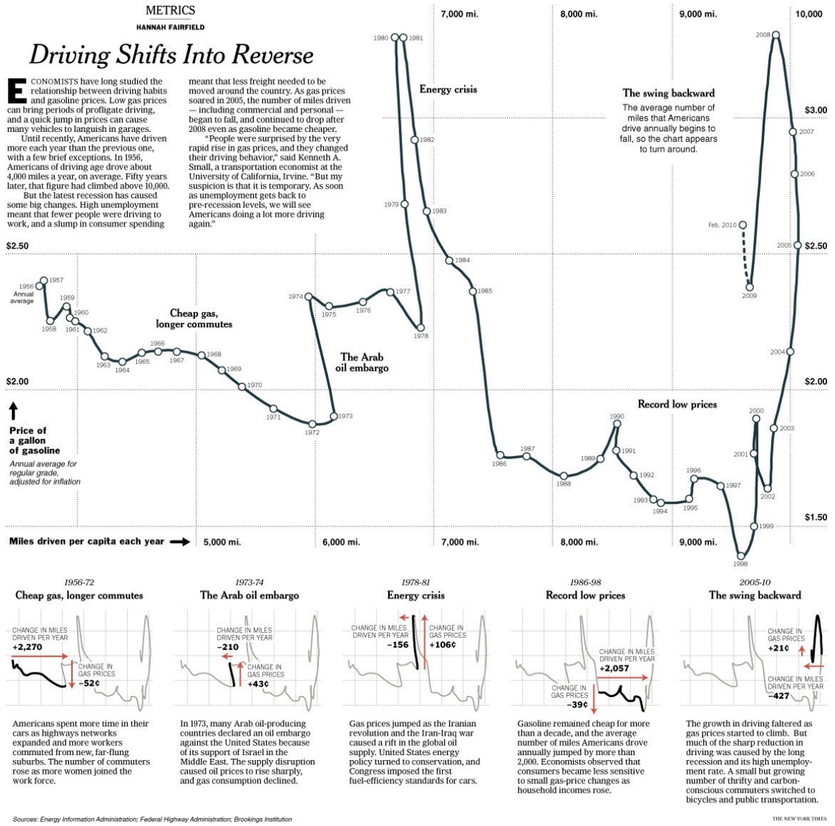




Connected Scatterplot

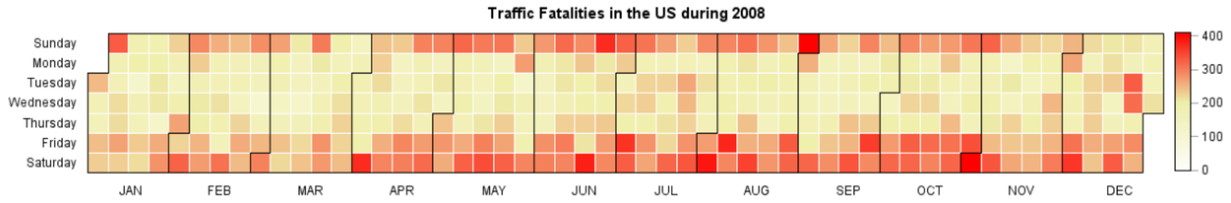
- Two Variables + Time Only one per Chart!
- Labels important





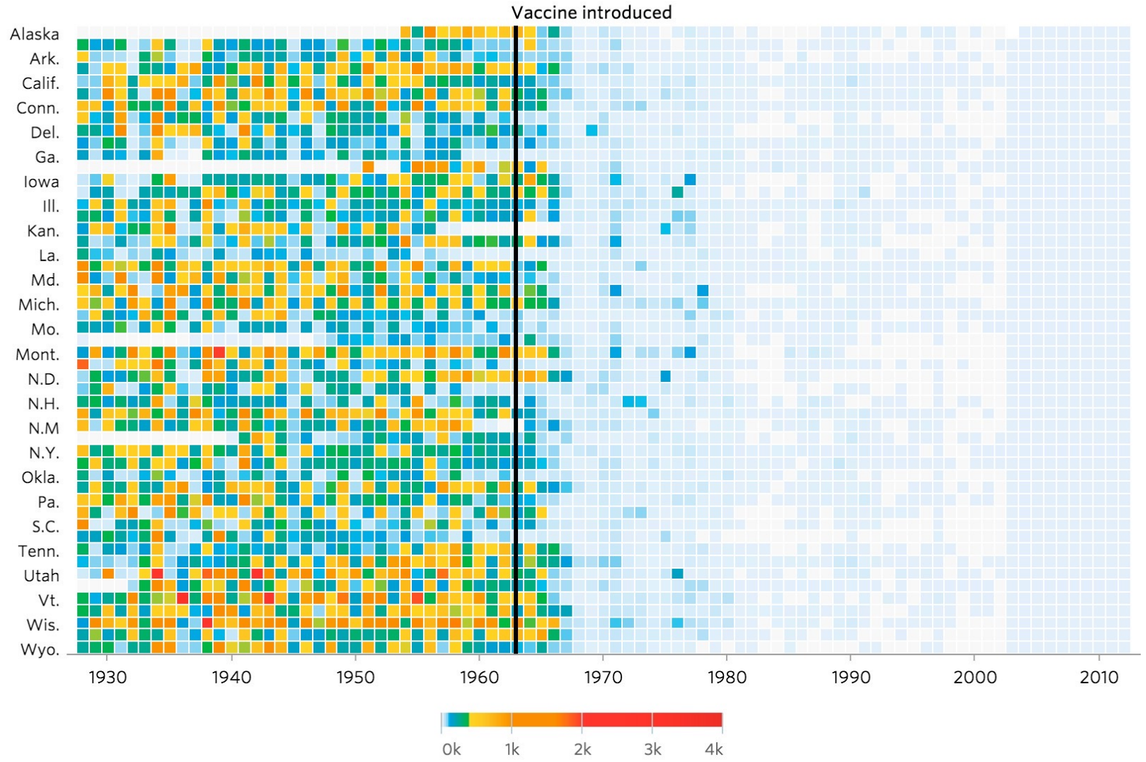
http://www.thefunctionalart.com/2012/09/in-praise-of-connected-scatter-plots.html

Heat Map and Calendar Heat Map



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

Measles



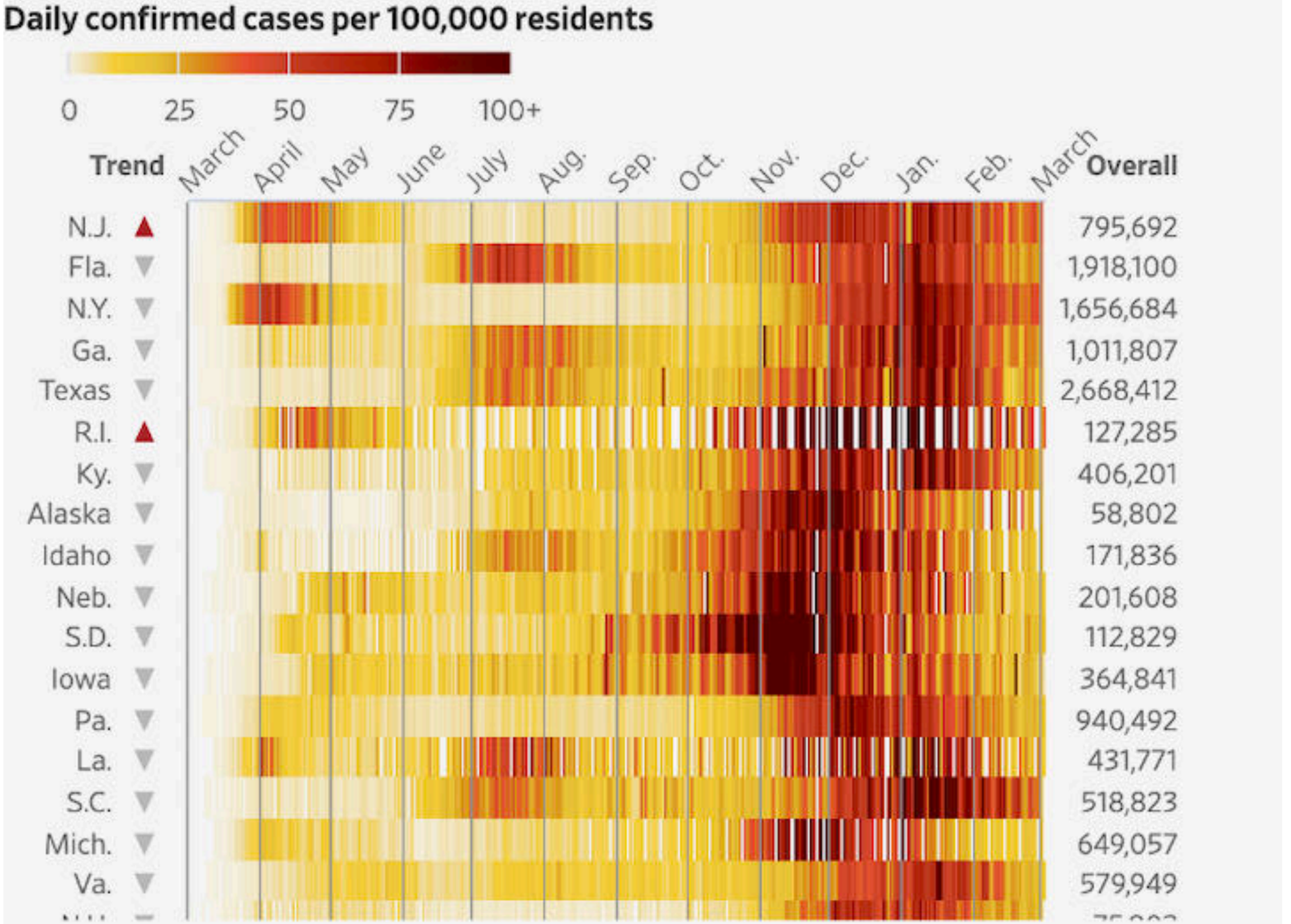
Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

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

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

Monitoring the U.S. Outbreak

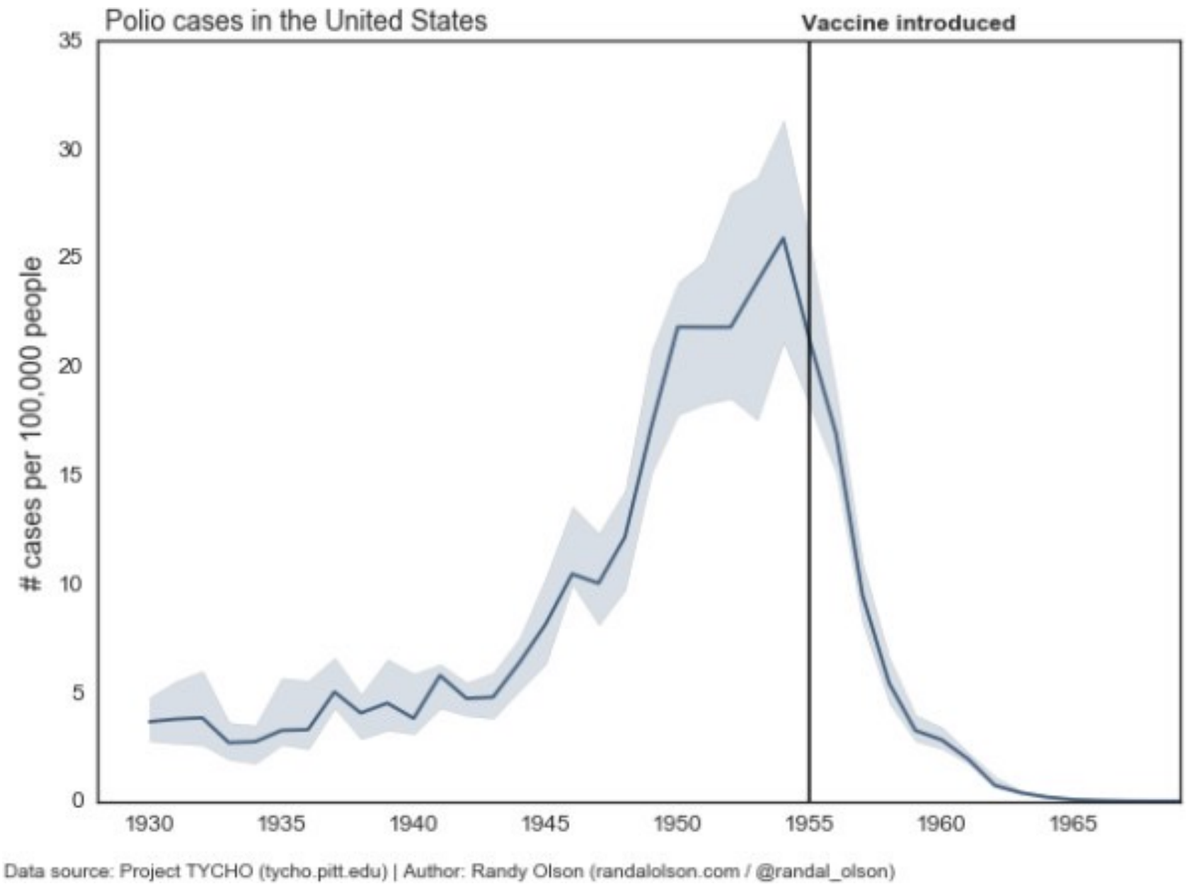
Confirmed cases by state, ranked by latest full-day count



Note: Trend indicates whether a state had an increase or decrease in total number of cases in the past seven days compared with previous seven days. Last updated March 3, at 1:56 p.m.

Sources: Johns Hopkins Center for Systems Science and Engineering; the Lancet; Associated Press; U.S.

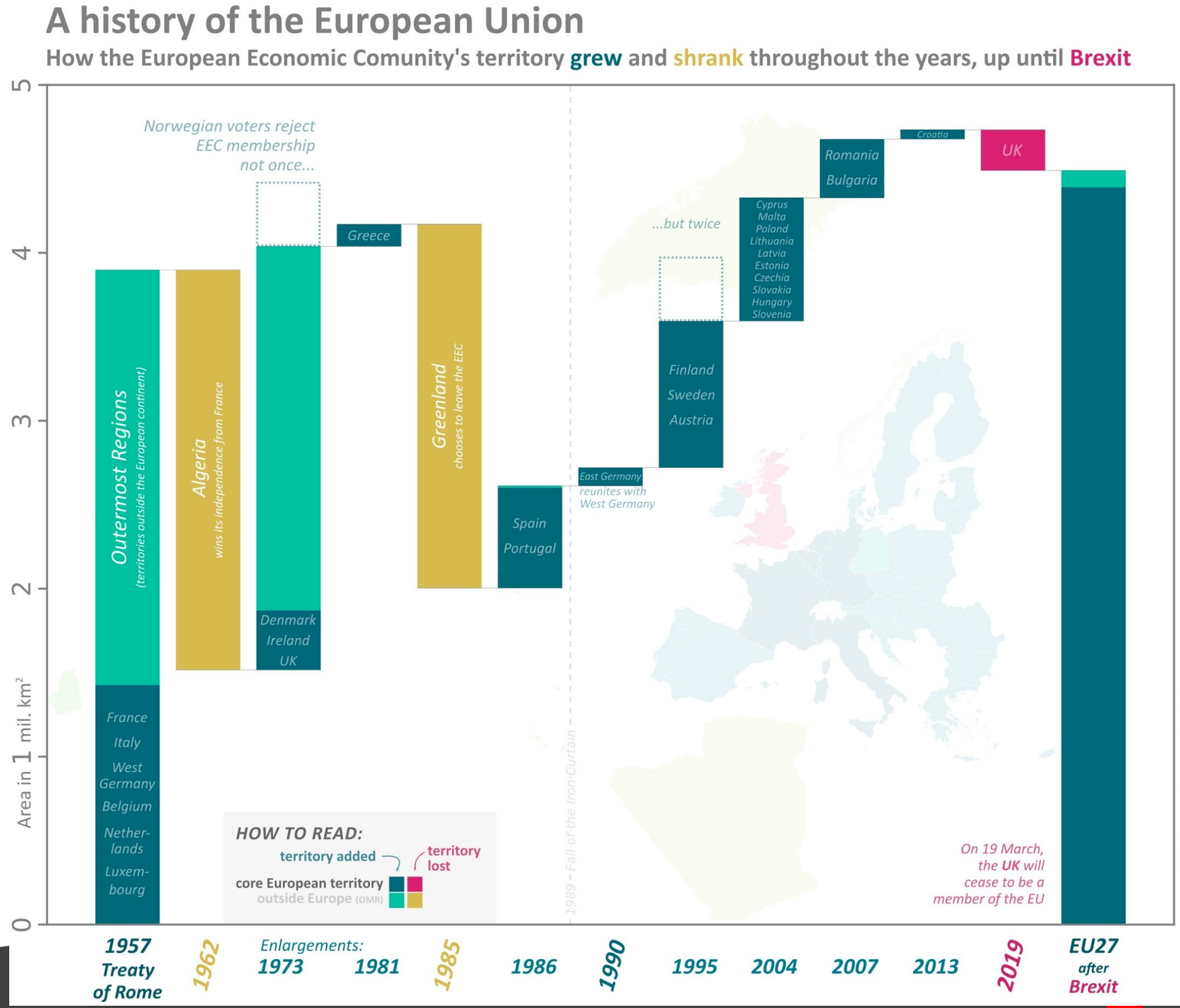
Sometimes you can Show Too Much Data



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)

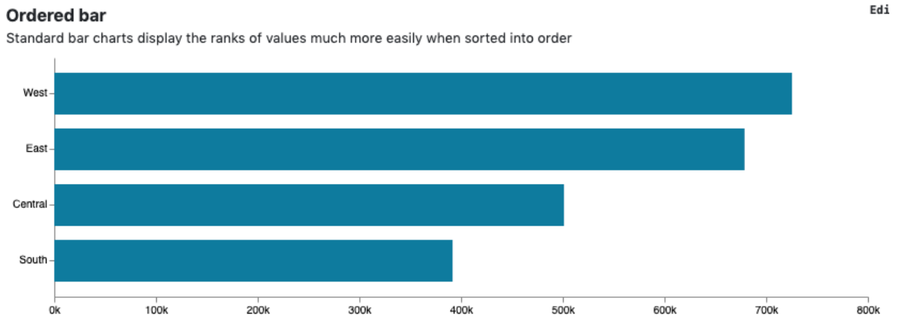


Ranking

Rankings are Popular

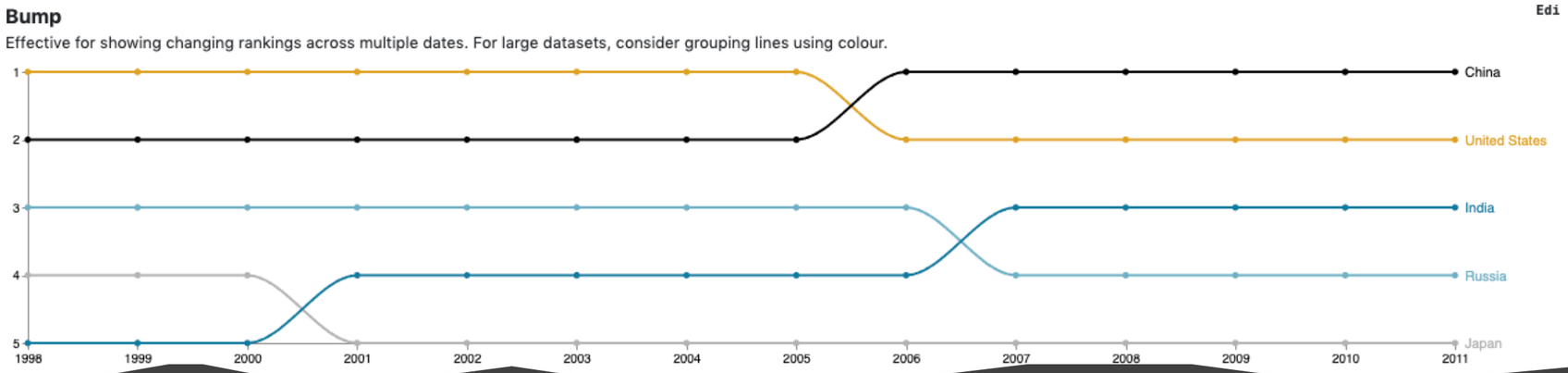
Rank	University	Score
1.	MIT	
2.	Harvard	
3.	Princeton	
4.	Cambridge	
5.	Oxford	

Ranking



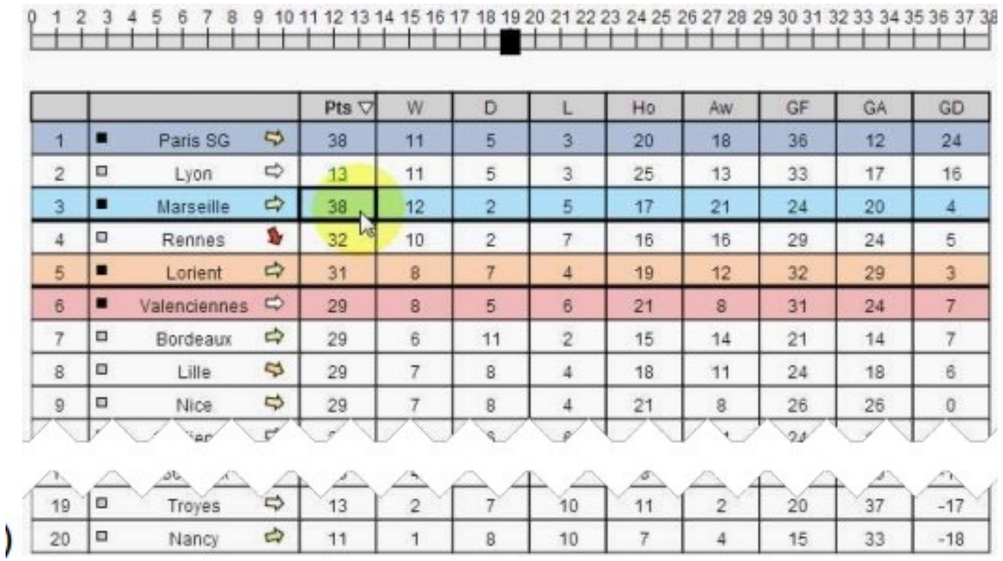
Magnitude Visualization + Sorting

Bump Charts for Rankings over Time



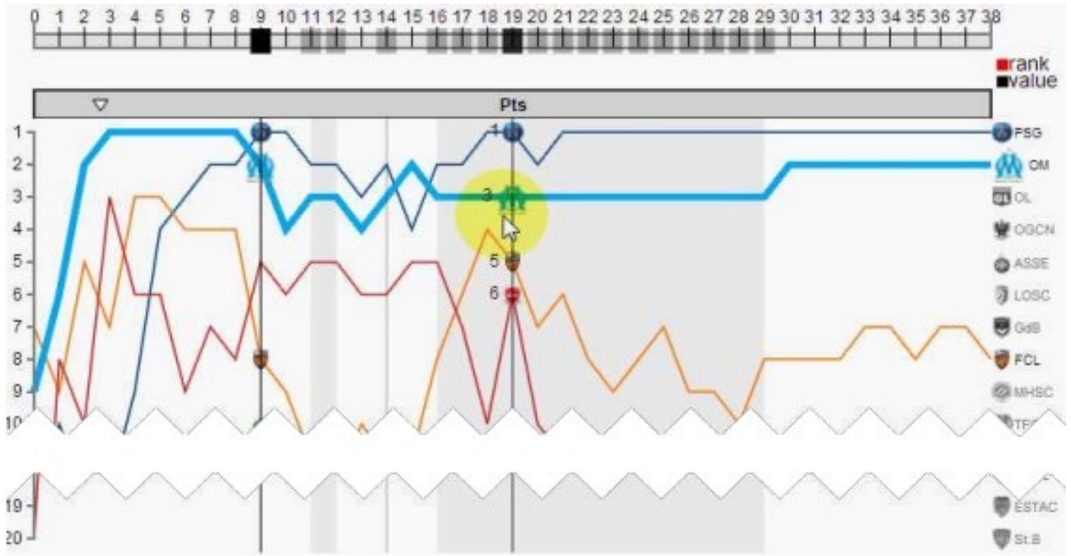
https://gramener.github.io/visual-vocabulary-vega/#/Ranking/

Temporal Rankings



		Pts	W	D	L	Ho	Aw	GF	GA	GD
1	Paris SG	38	11	5	3	20	18	36	12	24
2	Lyon	13	11	5	3	25	13	33	17	16
3	Marseille	38	12	2	5	17	21	24	20	4
4	Rennes	32	10	2	7	16	16	29	24	5
5	Lorient	31	8	7	4	19	12	32	29	3
6	Valenciennes	29	8	5	6	21	8	31	24	7
7	Bordeaux	29	6	11	2	15	14	21	14	7
8	Lille	29	7	8	4	18	11	24	18	6
9	Nice	29	7	8	4	21	8	26	26	0
19	Troyes	13	2	7	10	11	2	20	37	-17
20	Nancy	11	1	8	10	7	4	15	33	-18

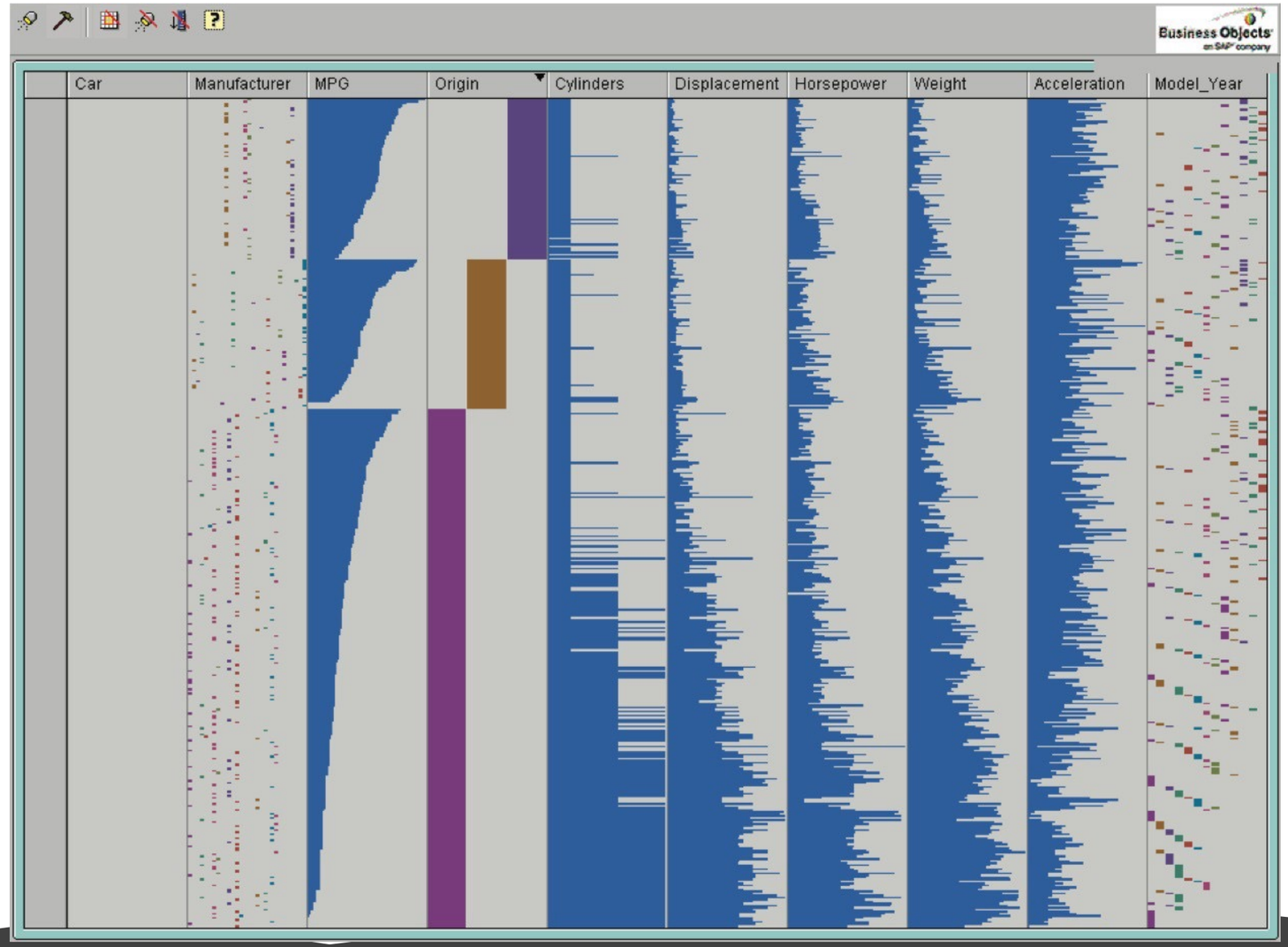
(b)



[Perin, Vuillemot, Fekete, CHI 2014]

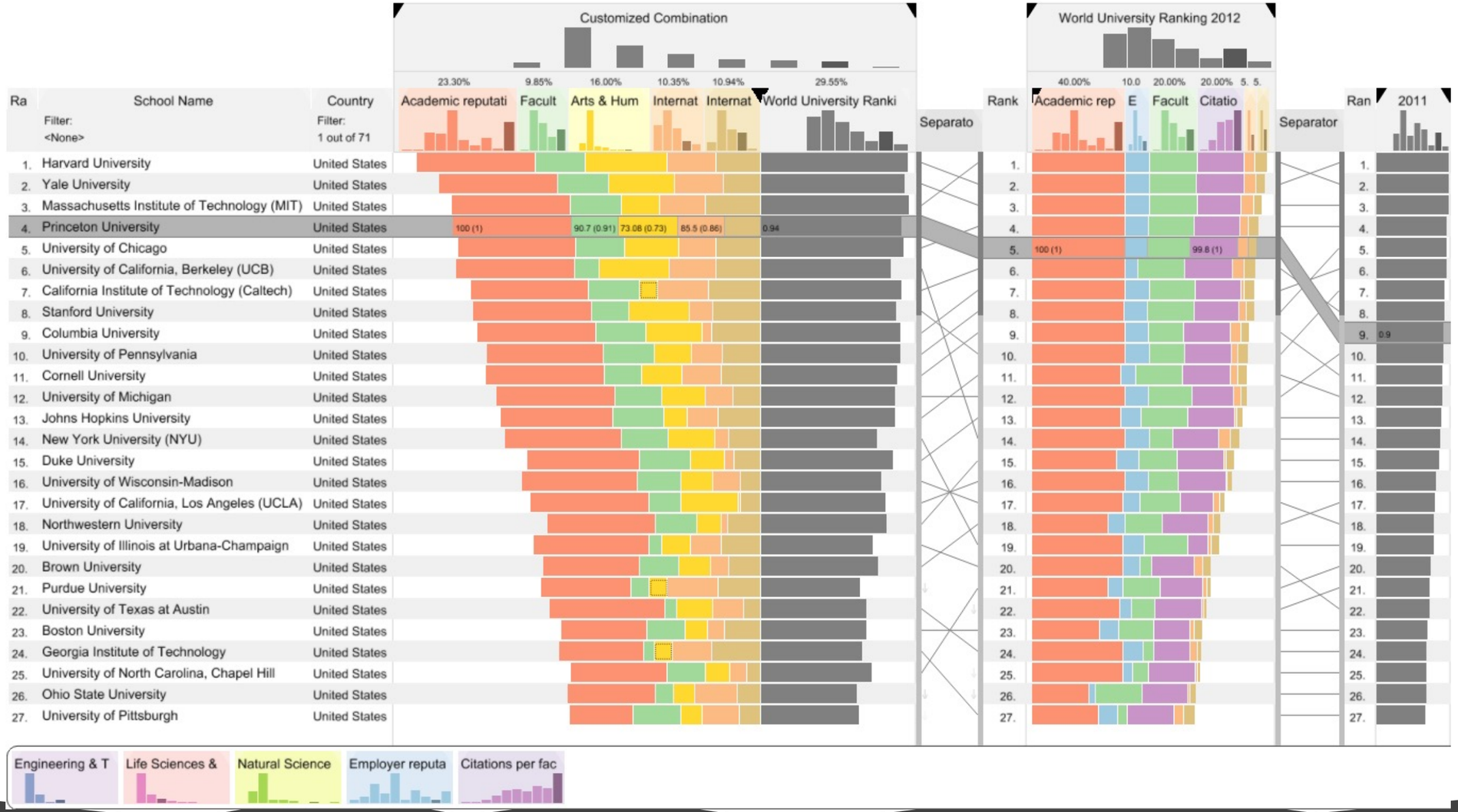
Table Lens

- Interactive table-based representation



Rao & Card 1994

LineUp

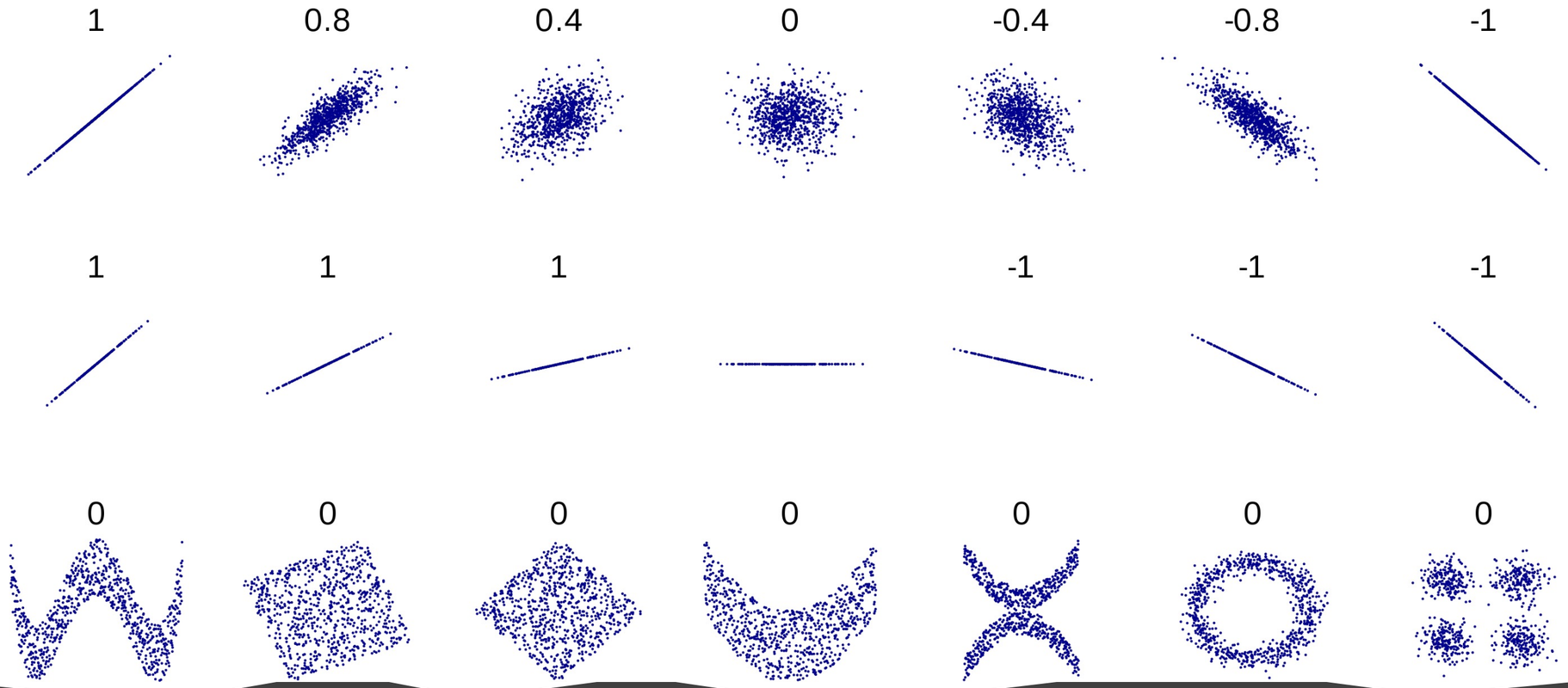


Video at http://lineup.caleydo.org

Correlation

What is Correlation

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

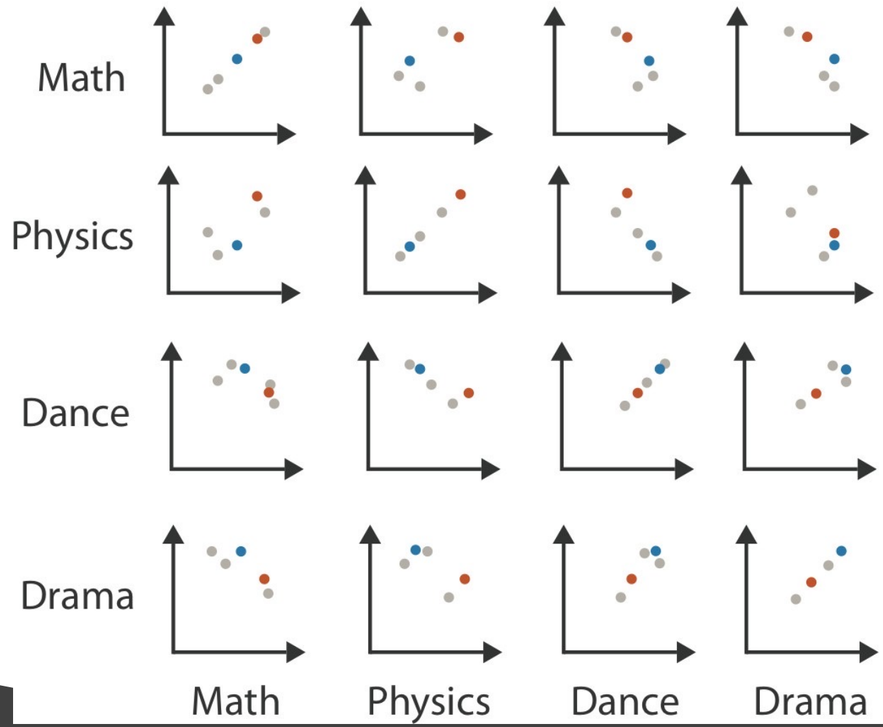


Axis-Based Techniques

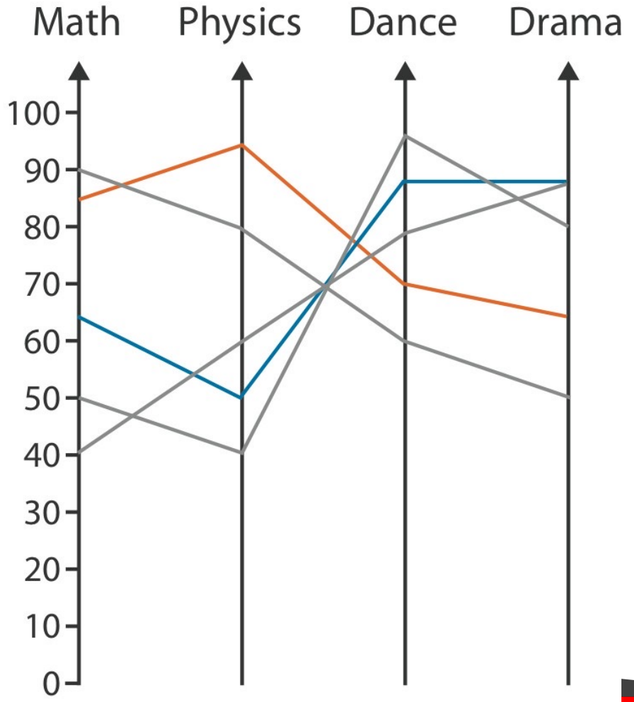
Table

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



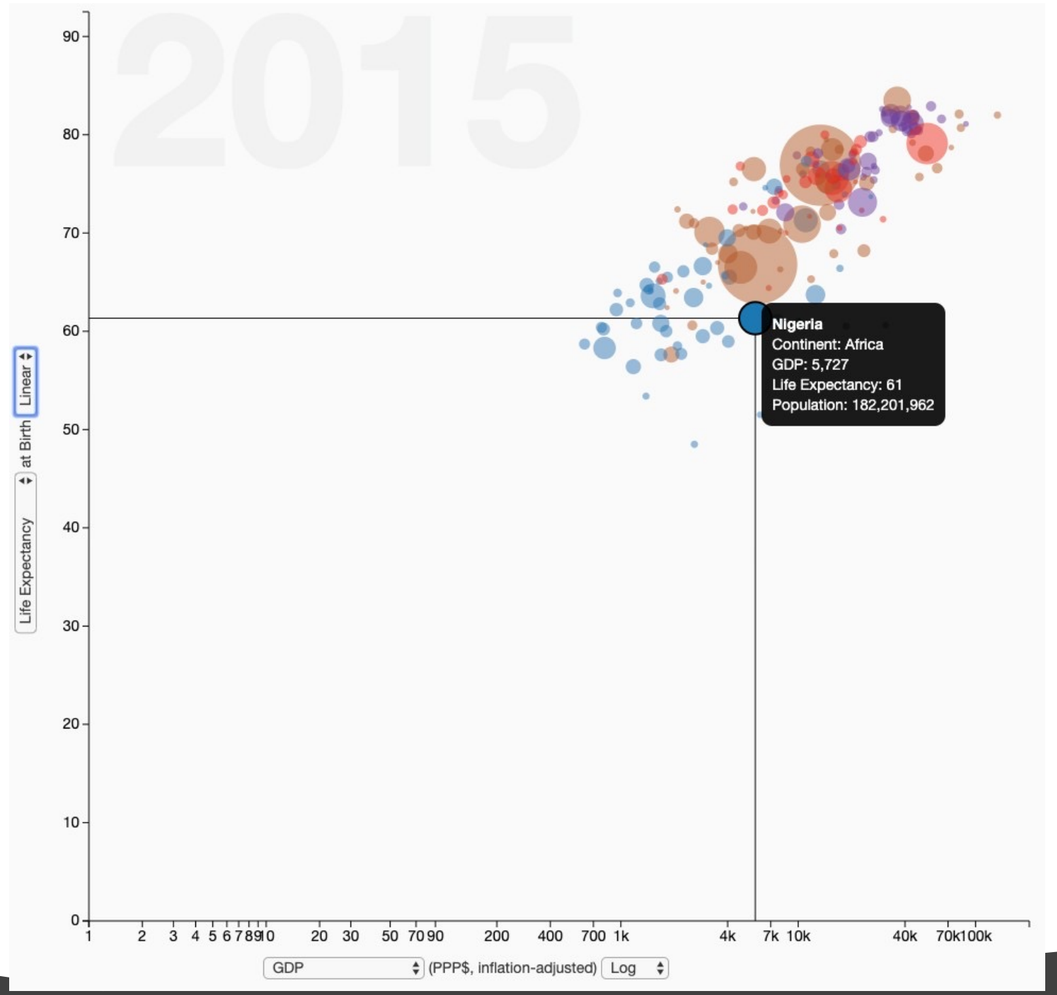
Parallel Coordinates



Scatterplots

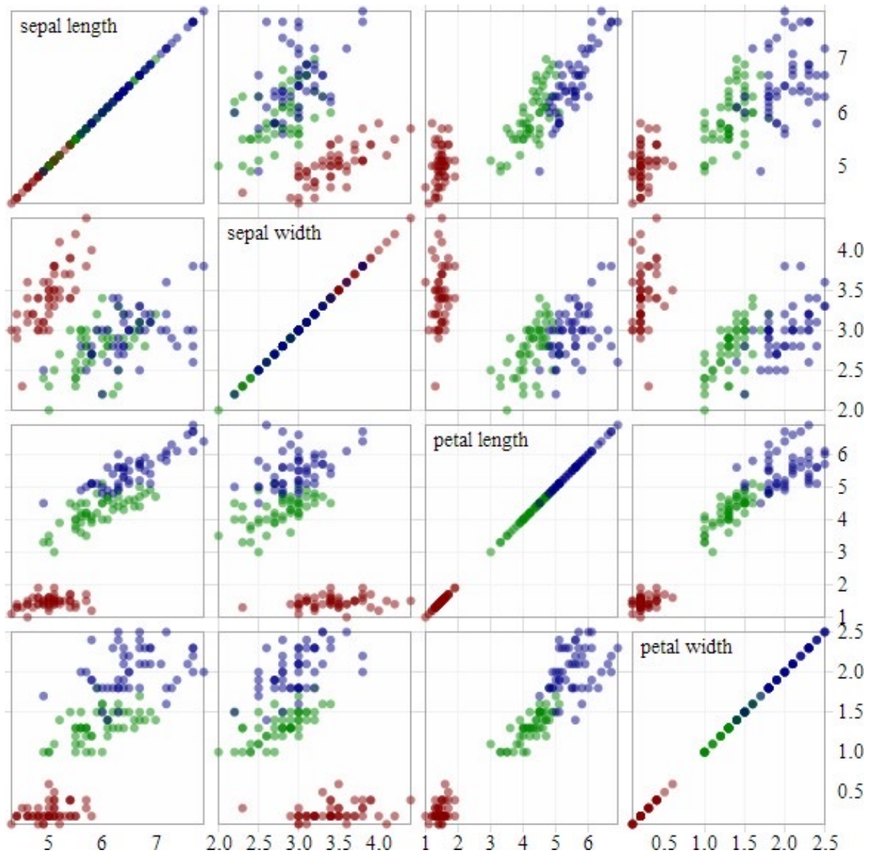
Scatterplots

- Two orthogonal axis visualizing one dimension each.
- (see prior lecture)



Scatterplot Matrices (SPLOM)

- Matrix of size d*d
- Each row/column is one dimension
- Each cell plots a scatterplot of two dimensions

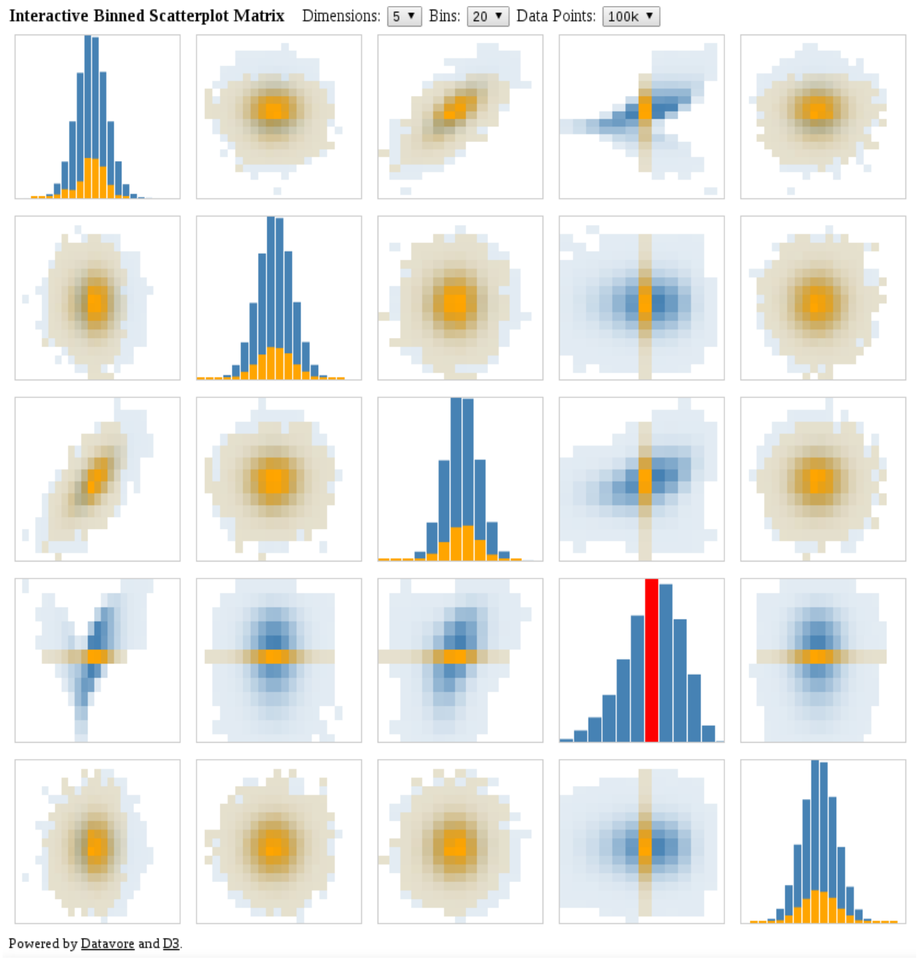


Scatterplot Matrices

- Limited scalability (~20 dimensions, ~500-1k records)
- Brushing is important
- Often combined with “Focus Scatterplot” as F+C technique

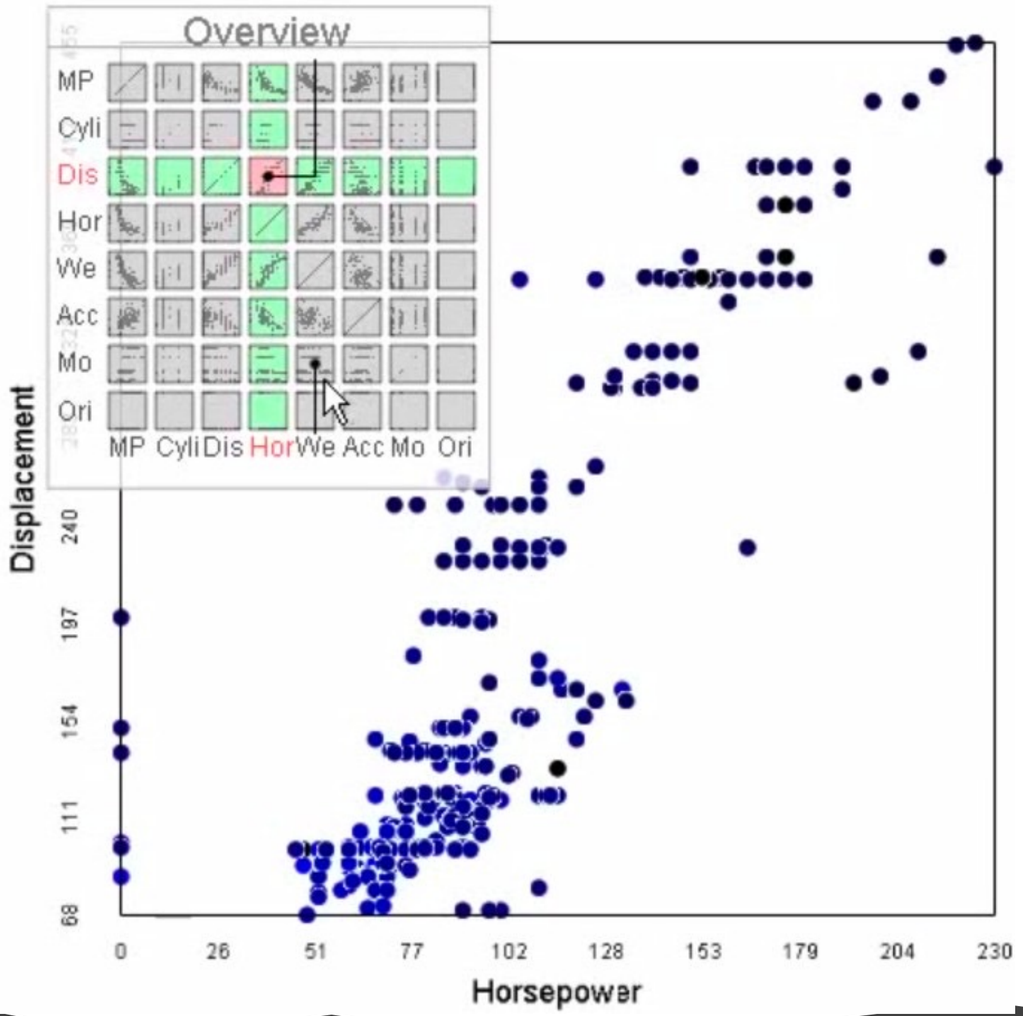
- Algorithmic approaches:
  - Clustering & aggregating records
  - Choosing dimensions
  - Choosing order

SPLOM Aggregation - Heat Map



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

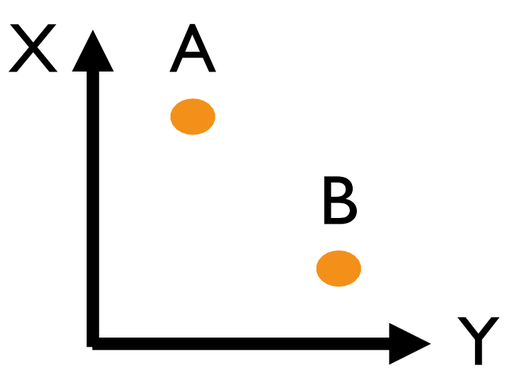
SPLOM F+C, Navigation

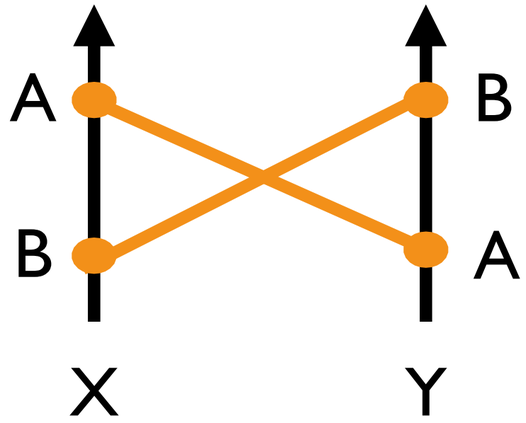


Parallel Coordinates

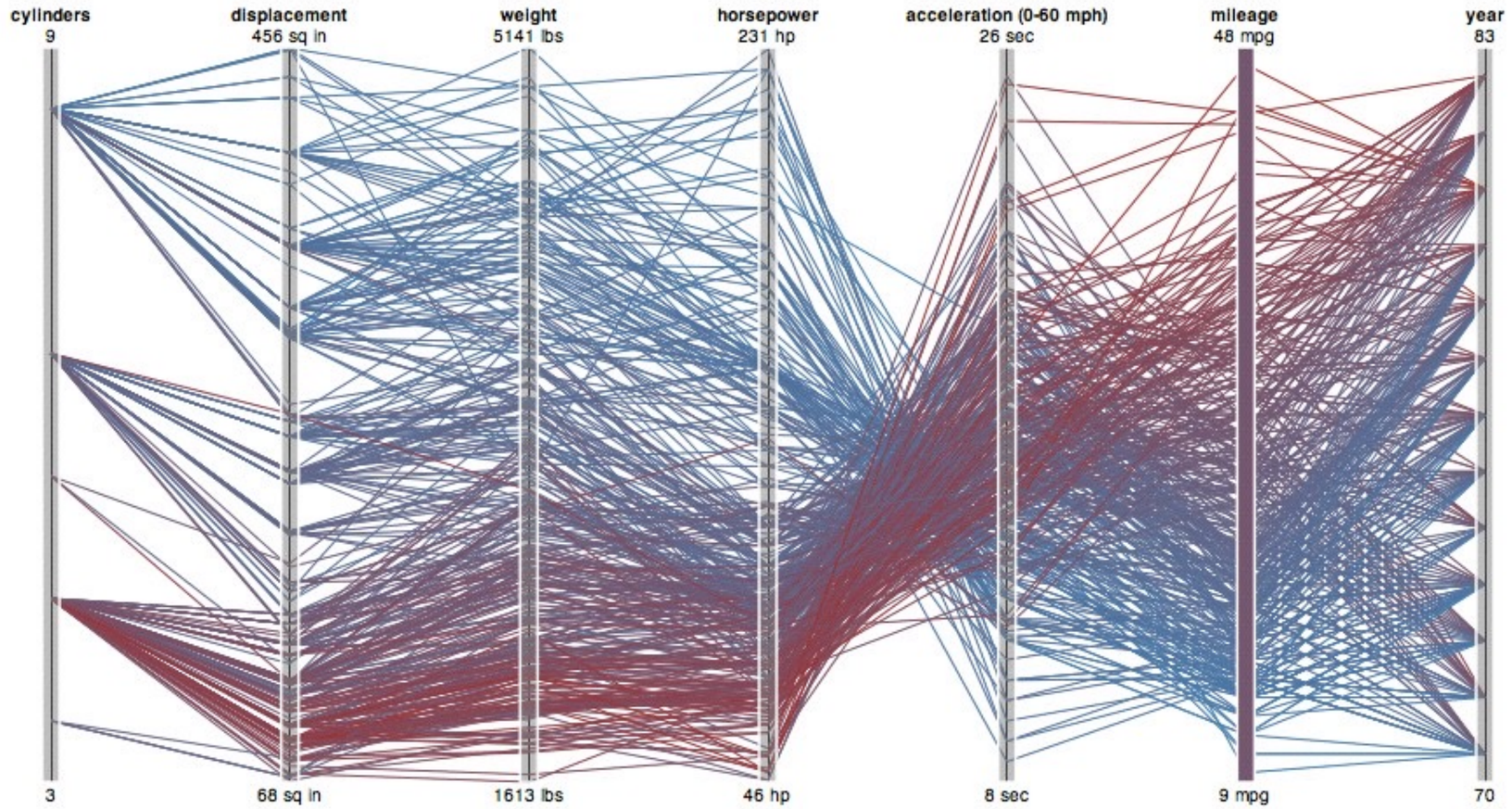
Parallel Coordinates (PC)

- Axes represent attributes
- Lines connecting axes represent items



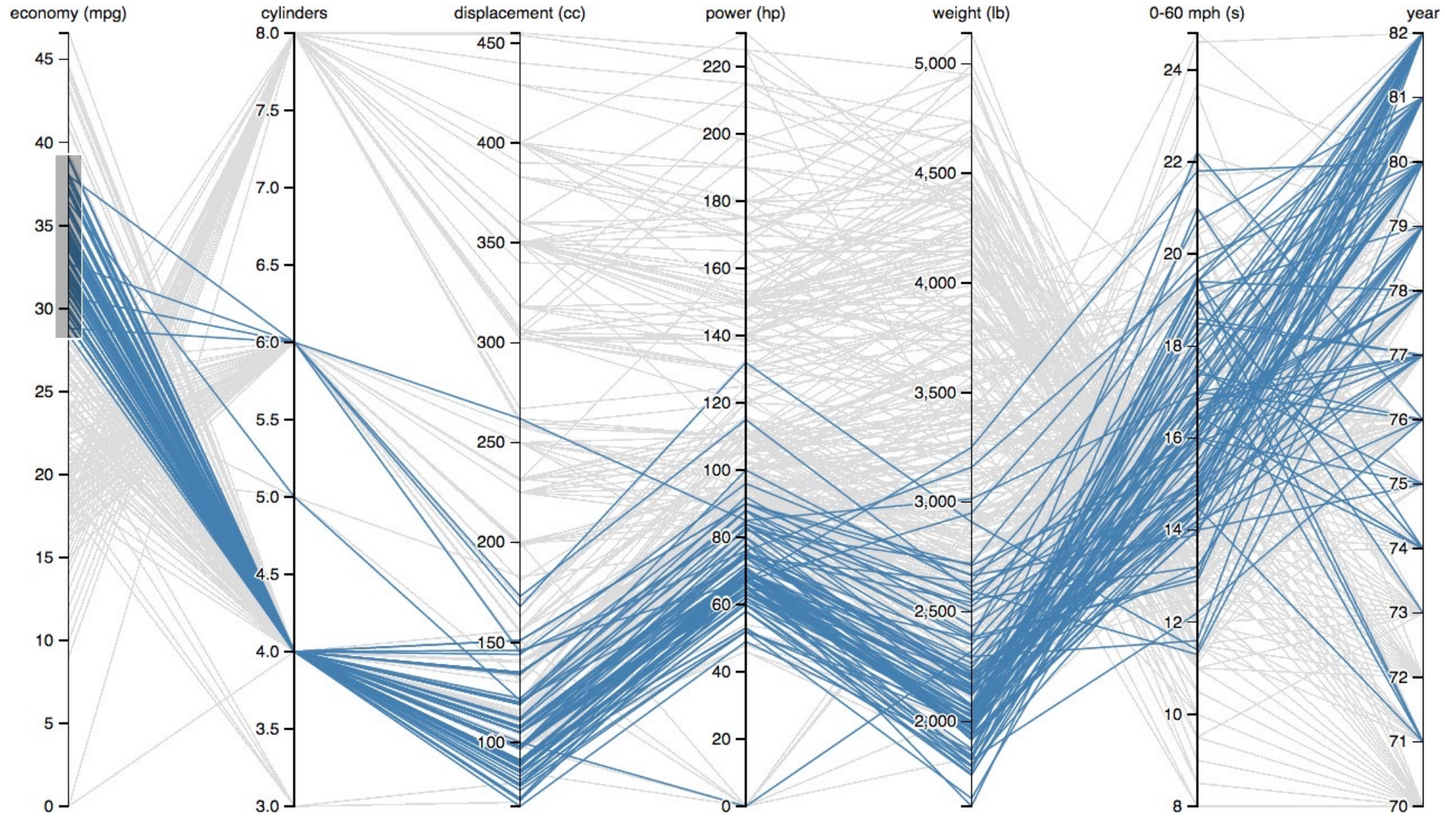


parallel coordinates



Parallel Coordinates

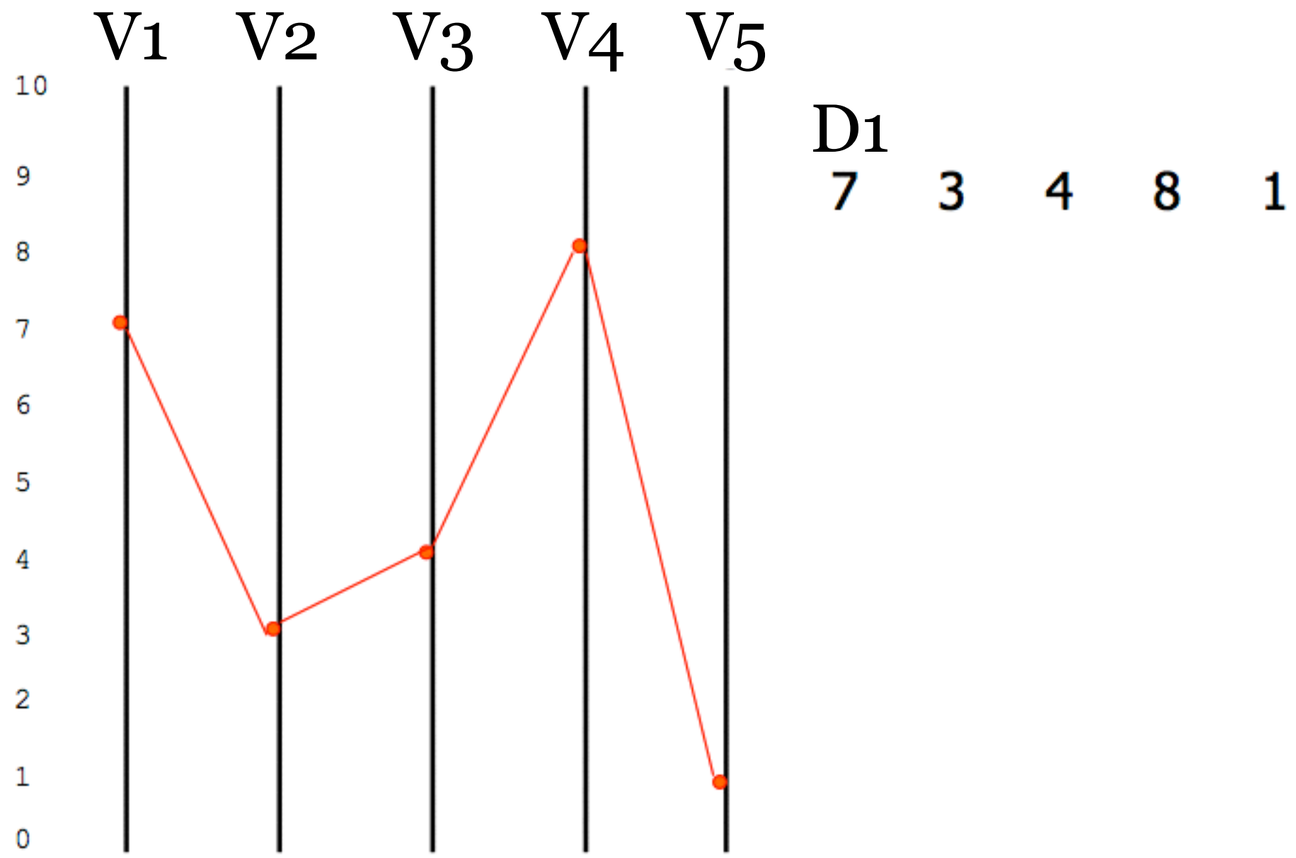
- Each axis represents dimension
- Lines connecting axis represent records
- Suitable for
  - all tabular data types
  - heterogeneous data



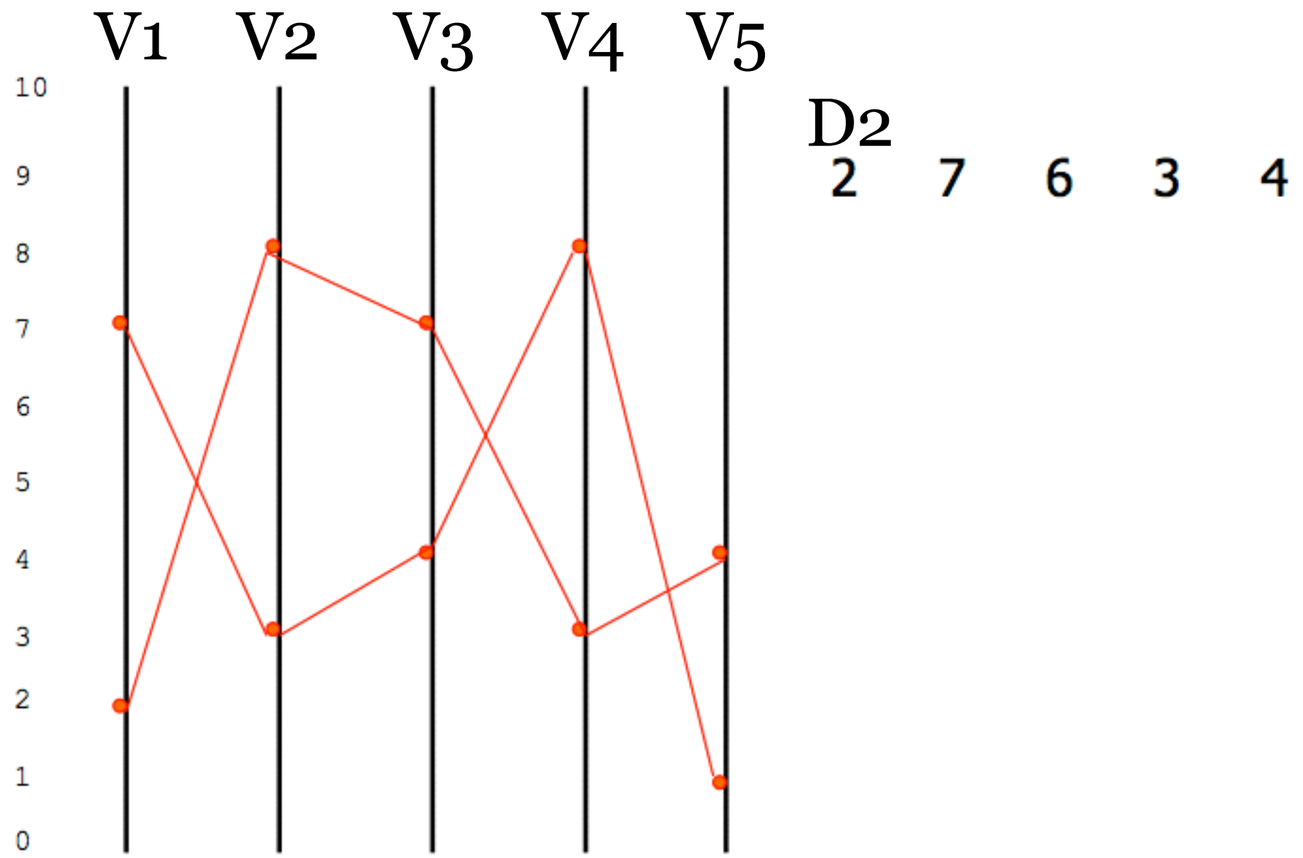
EXAMPLE

	V1	V2	V3	V4	V5
D1	7	3	4	8	1
D2	2	7	6	3	4
D3	9	8	1	4	2

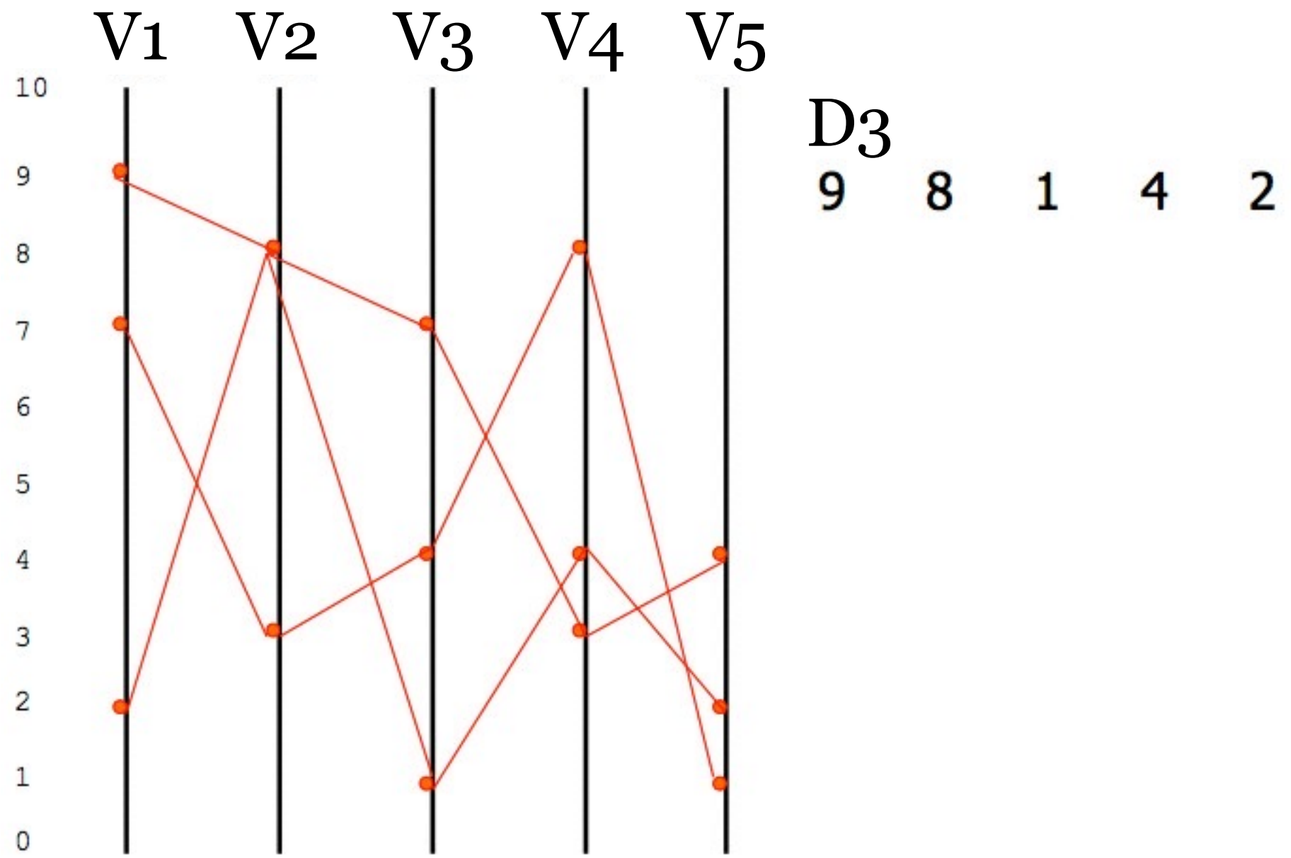
EXAMPLE



EXAMPLE



EXAMPLE



show correlation

- positive correlation: straight lines
- negative correlation: lines cross at a single point

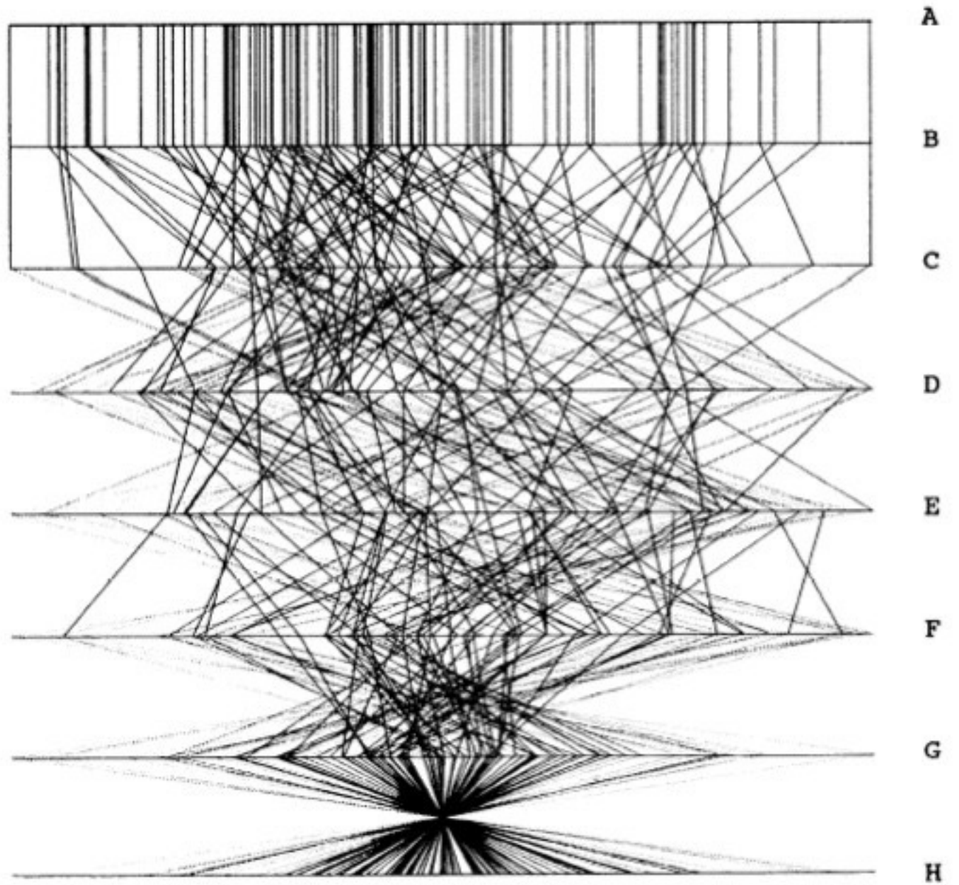
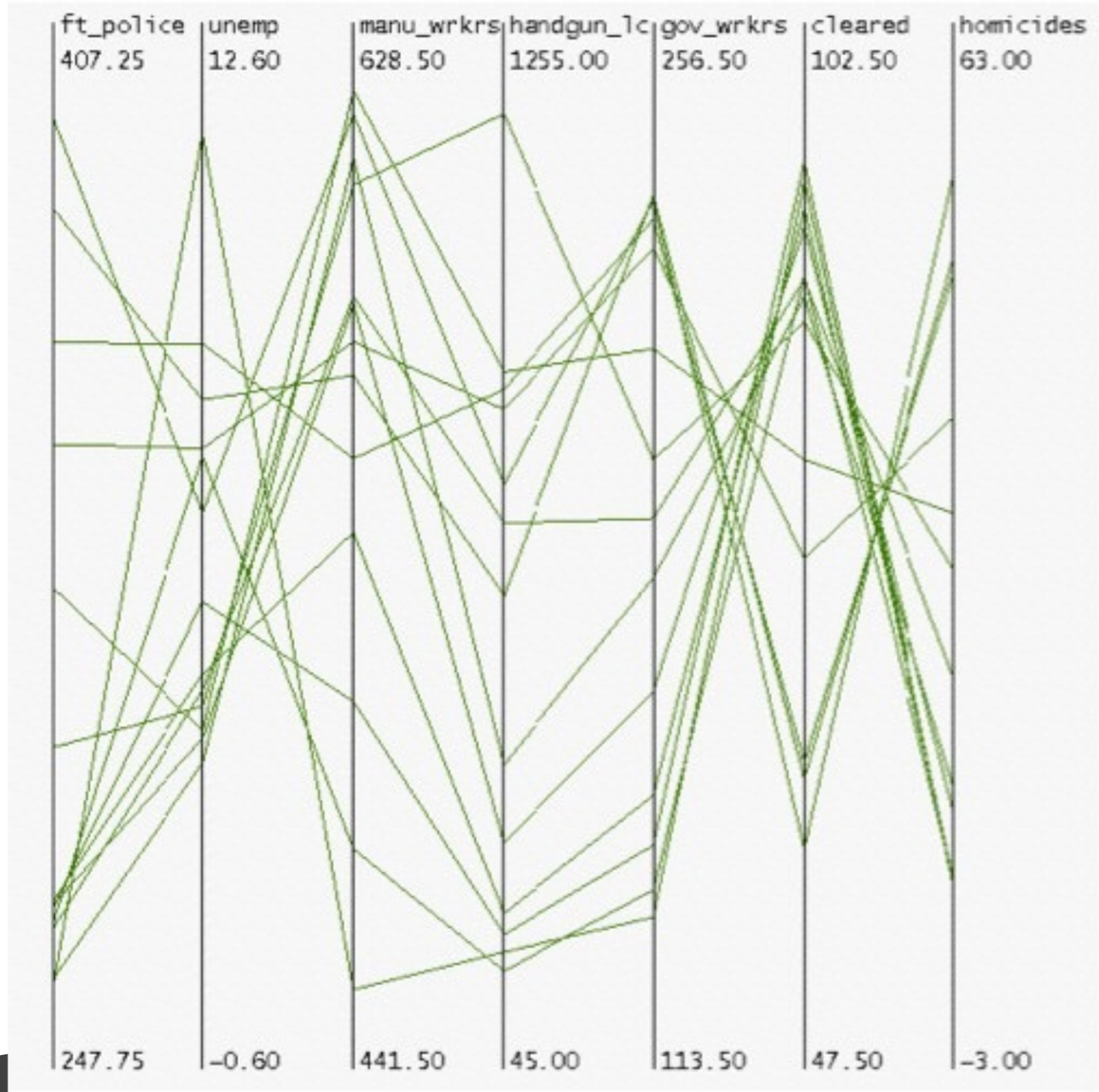


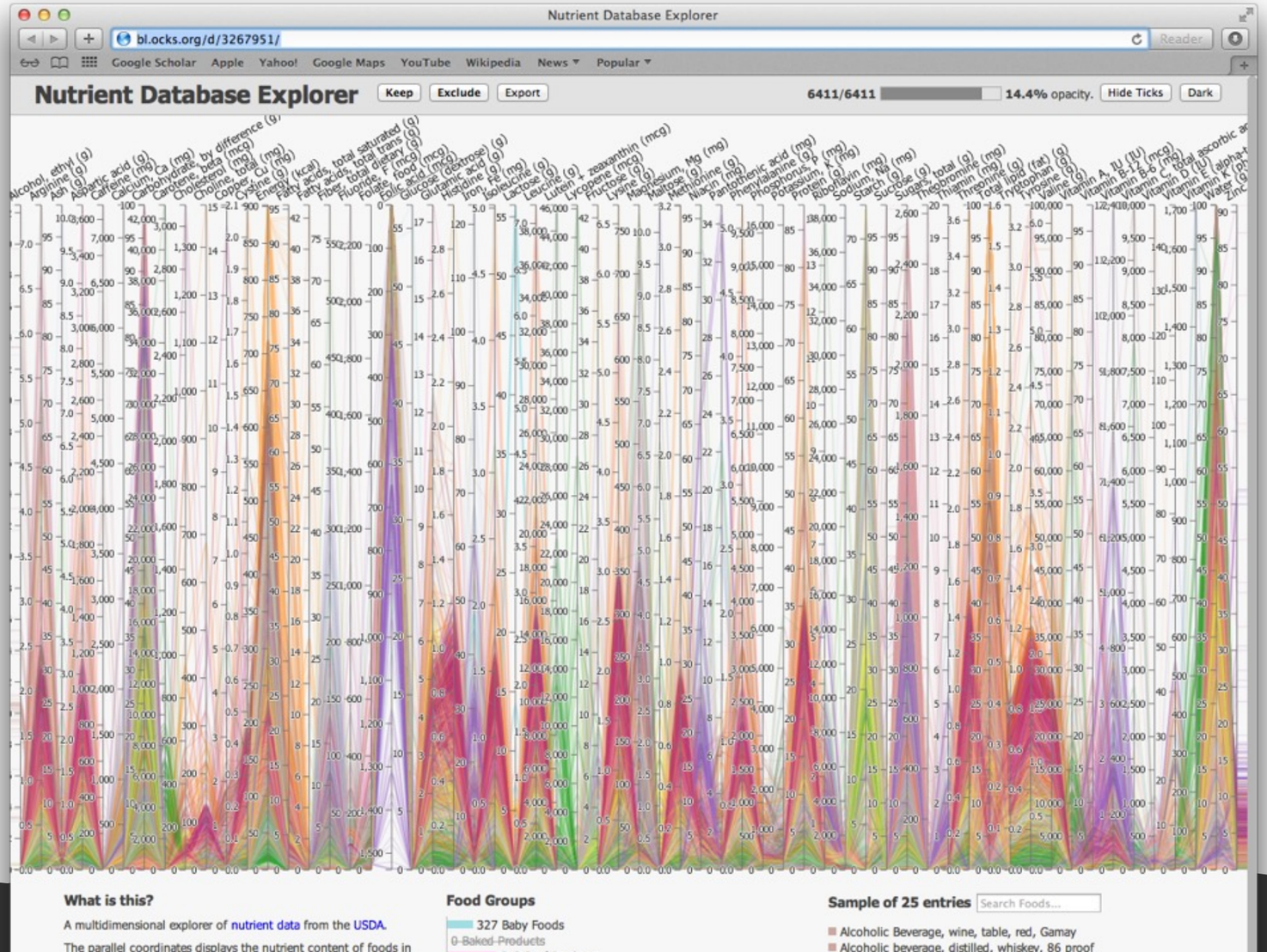
Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of $\rho = 1, .8, .2, 0, -.2, -.8,$ and -1.

- Wegman 1990

do you see any correlations?

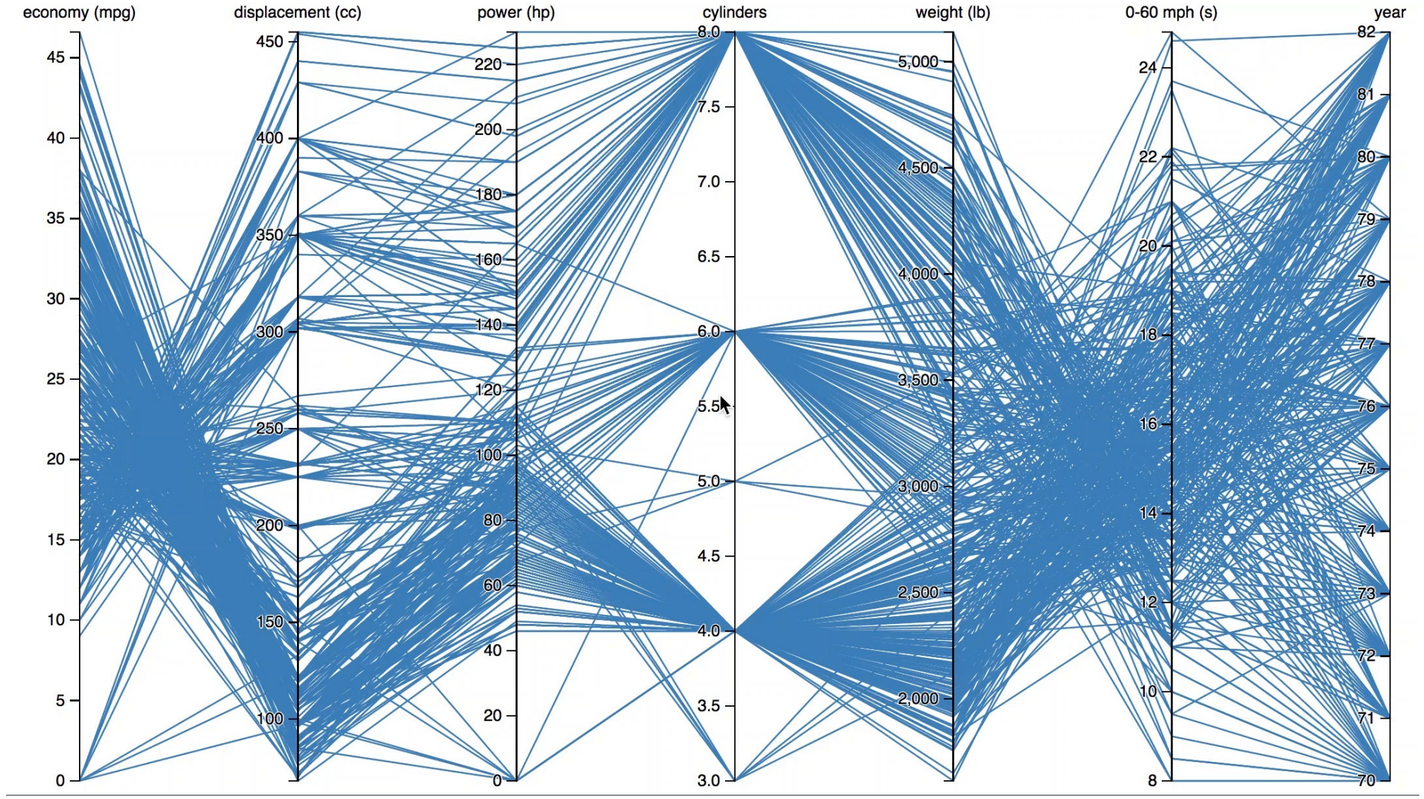


PC Limitation:
Scalability to Many Dimensions



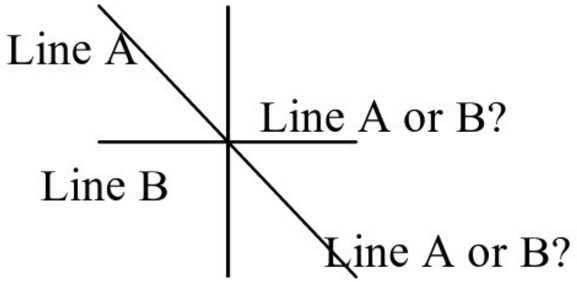
PC Limitations: Correlations only between adjacent axes

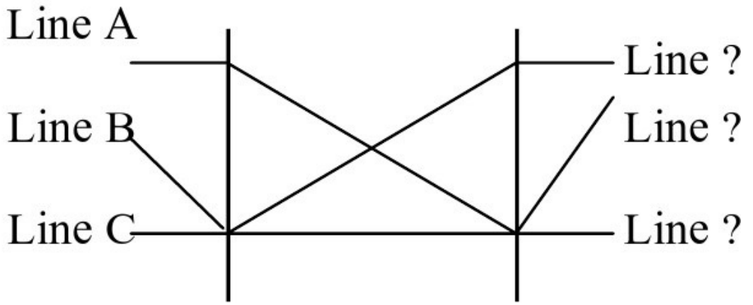
- Solution: Interaction
  - Brushing
  - Let user change order

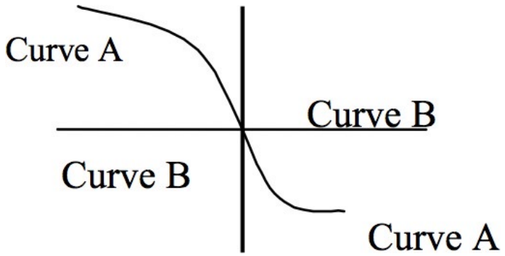


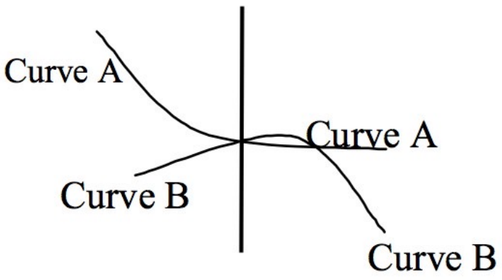
PC Limitation: Ambiguity

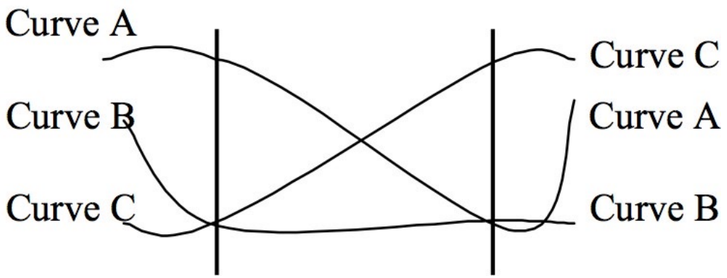
- Solutions:
  - Brushing
  - Curves







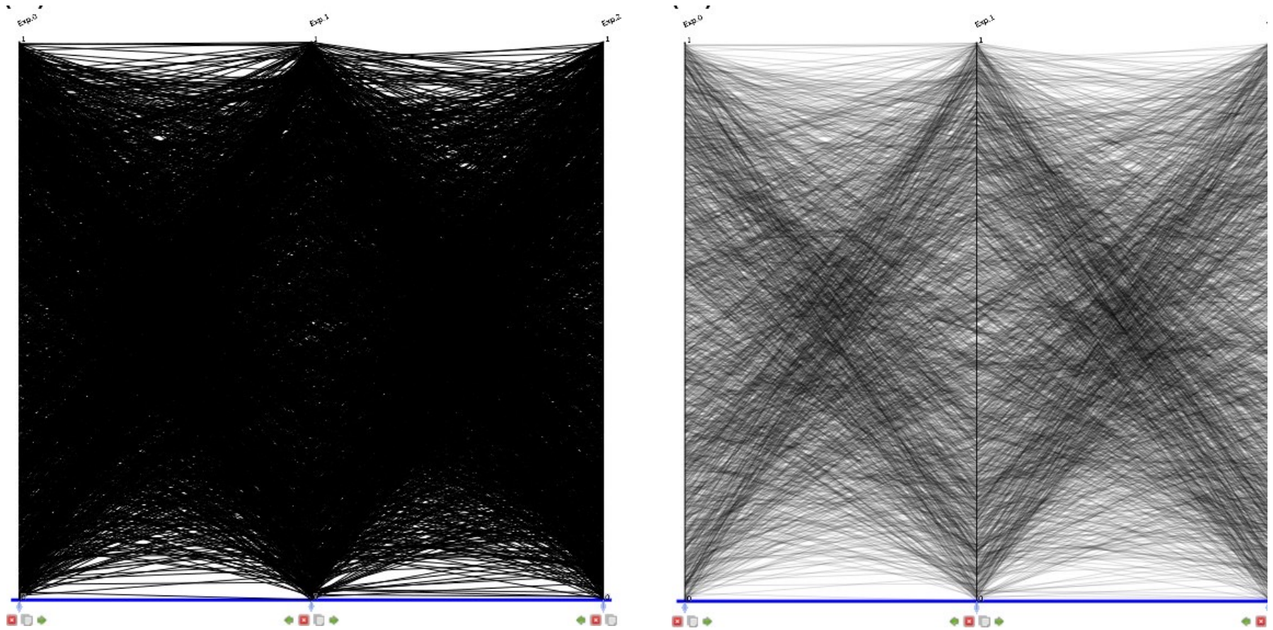




Graham and Kennedy 2003

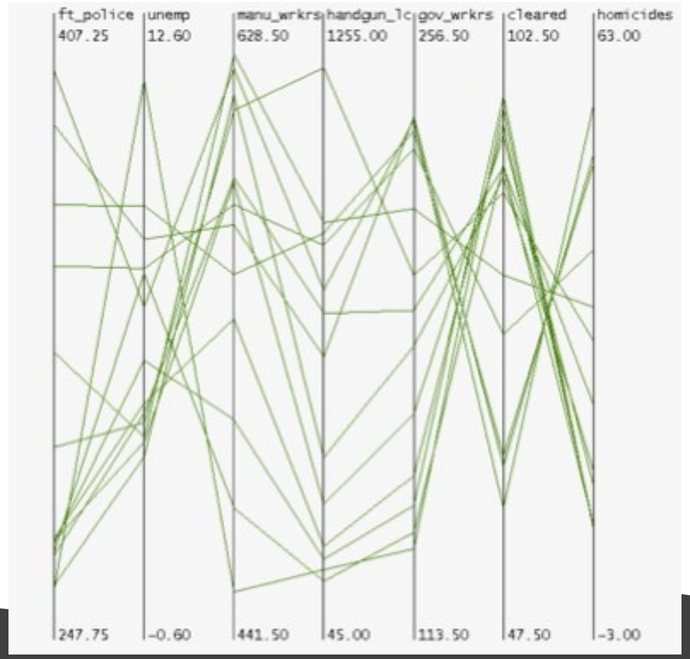
PC Limitation:
Scalability to Many Items

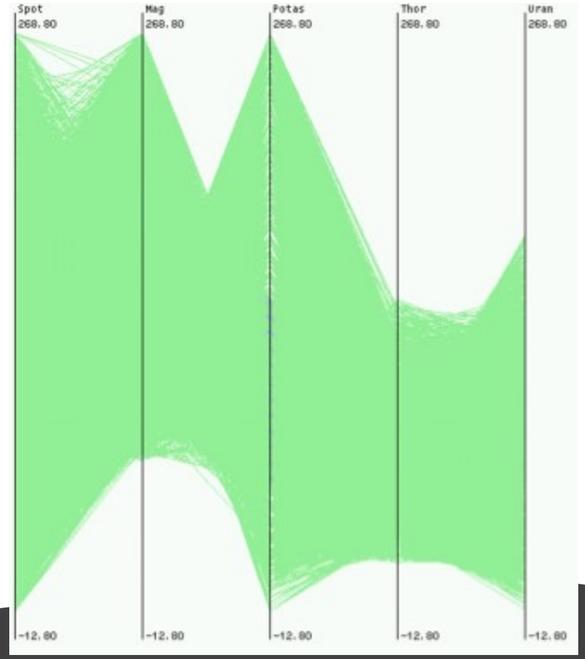
- Solutions:
  - Transparency
  - Bundling
  - Clustering
  - Sampling



HIERARCHICAL PARALLEL COORDINATES

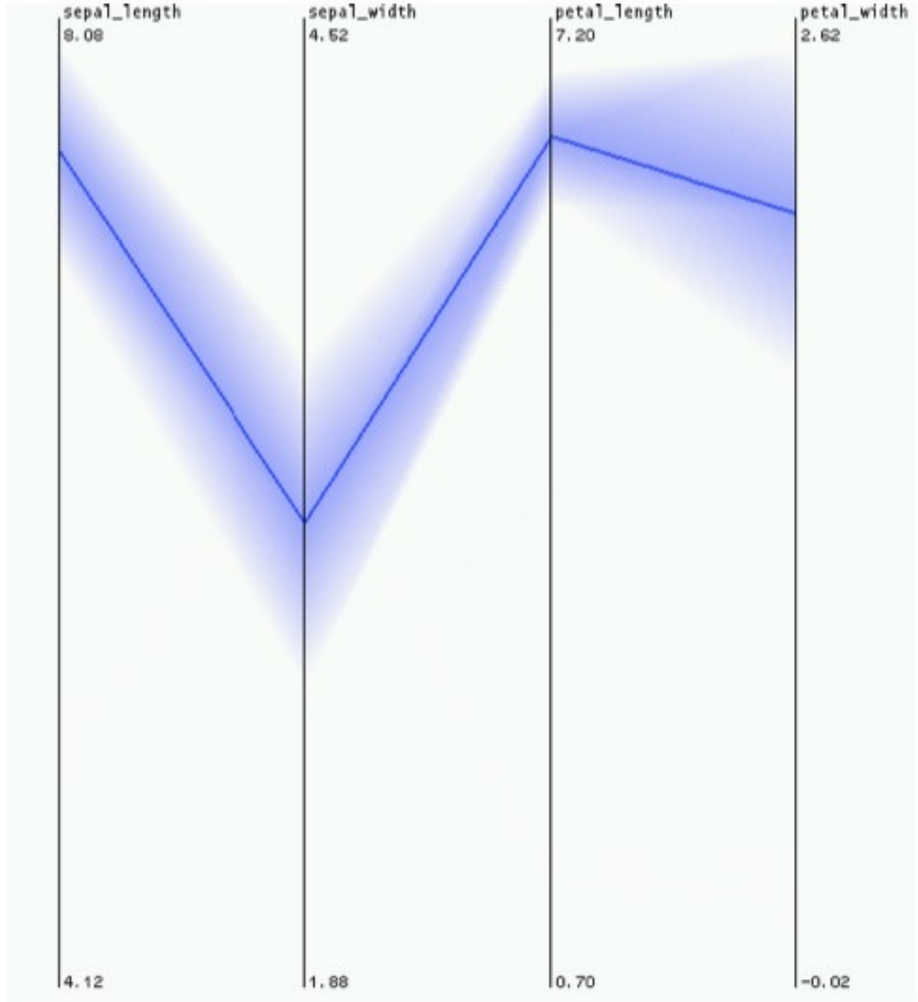
- goal: scale up parallel coordinates to large datasets
  - challenge: overplotting/occlusion





HPC: ENCODING DERIVED DATA

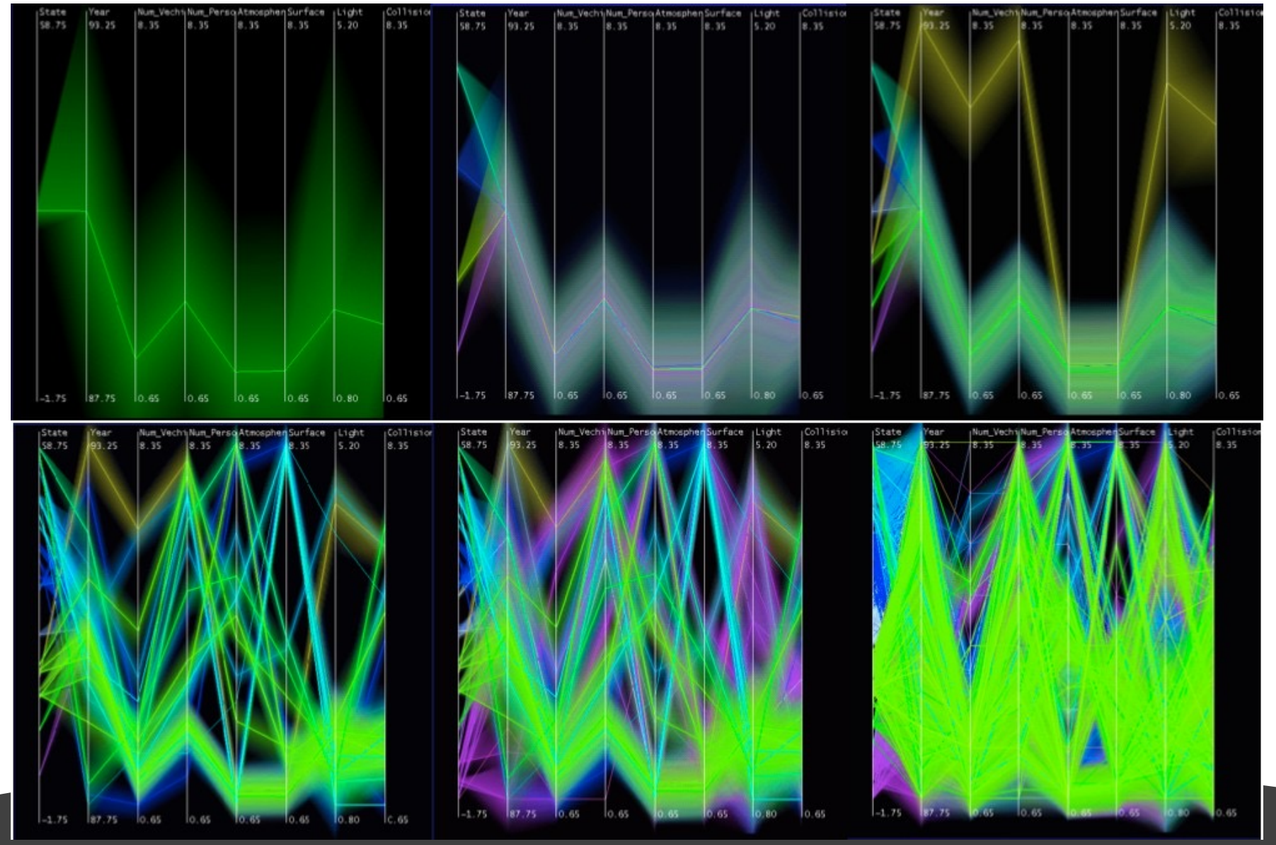
- visual representation: variable-width opacity bands
- show whole cluster, not just single item
- min / max: spatial position cluster density: transparency mean: opaque



Fua 1999

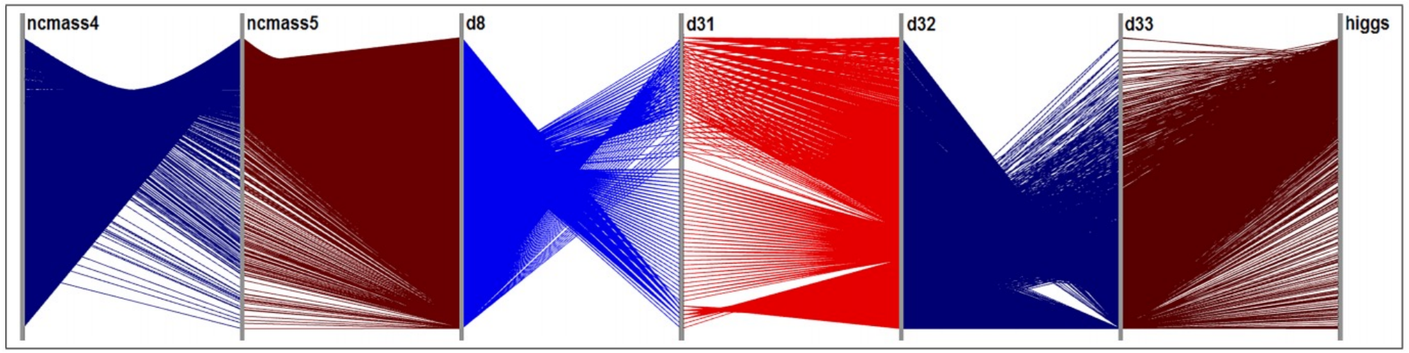
HPC: INTERACTING WITH DERIVED DATA

- interactively change level of detail to navigate cluster hierarchy

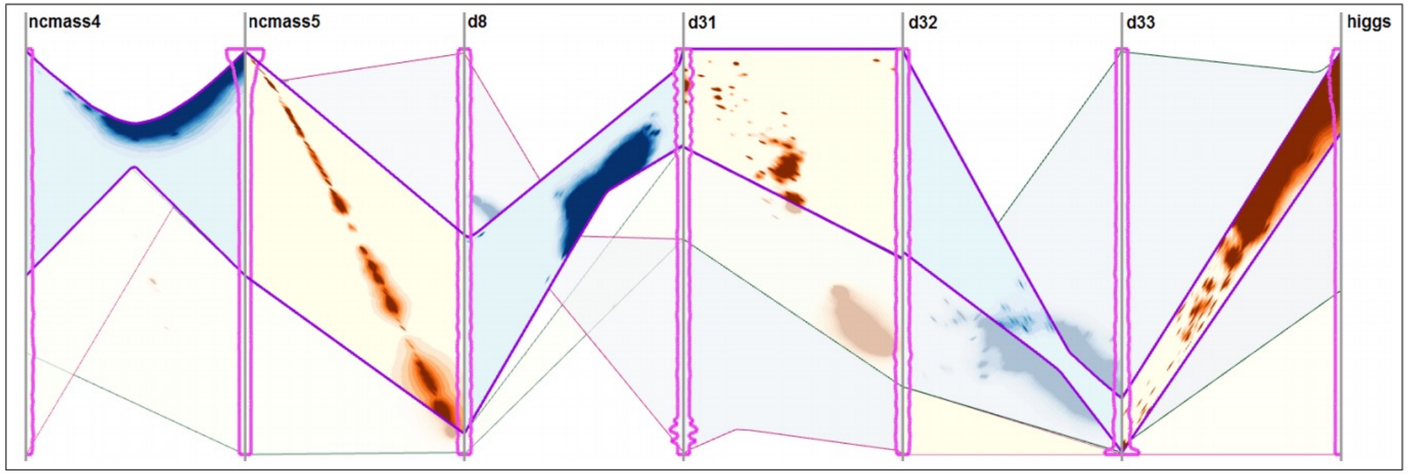


Data-Scalable Parallel Coordinates

- Cluster into groups of homogeneous behavior and represent positive and negative correlations directly



(a) Conventional PCPs



(b) DSPCP using K-means clustering

Parallel Coordinates

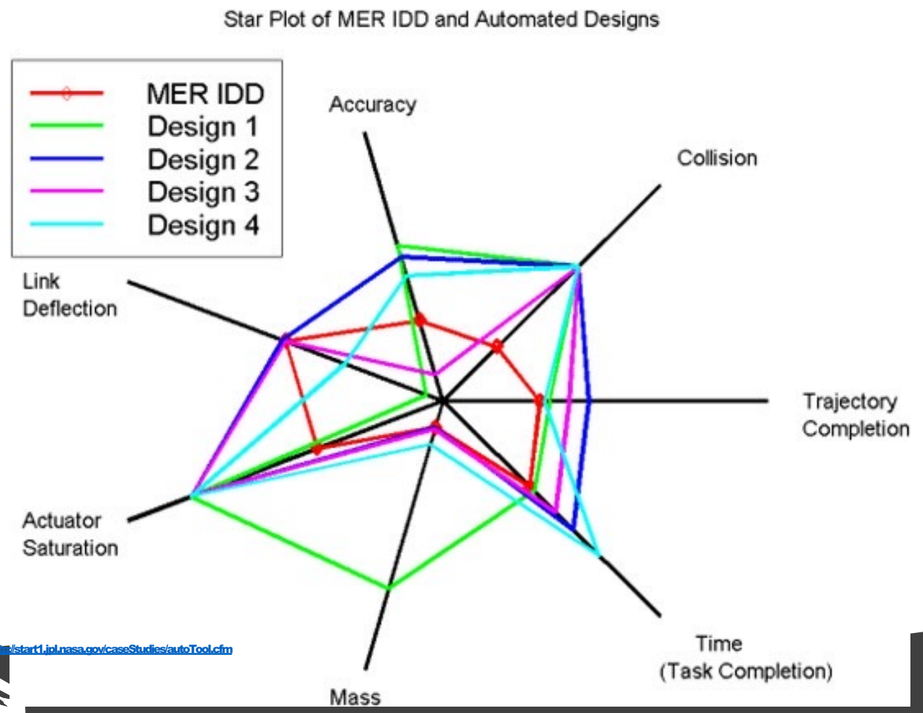
- Shows primarily relationships between adjacent axis
- Limited scalability (~50 dimensions, ~1-5k records)
  - Transparency of lines
- Interaction is crucial
  - Axis reordering
  - Brushing
  - Filtering

- Algorithmic support:
  - Choosing dimensions
  - Choosing order
  - Clustering & aggregating records

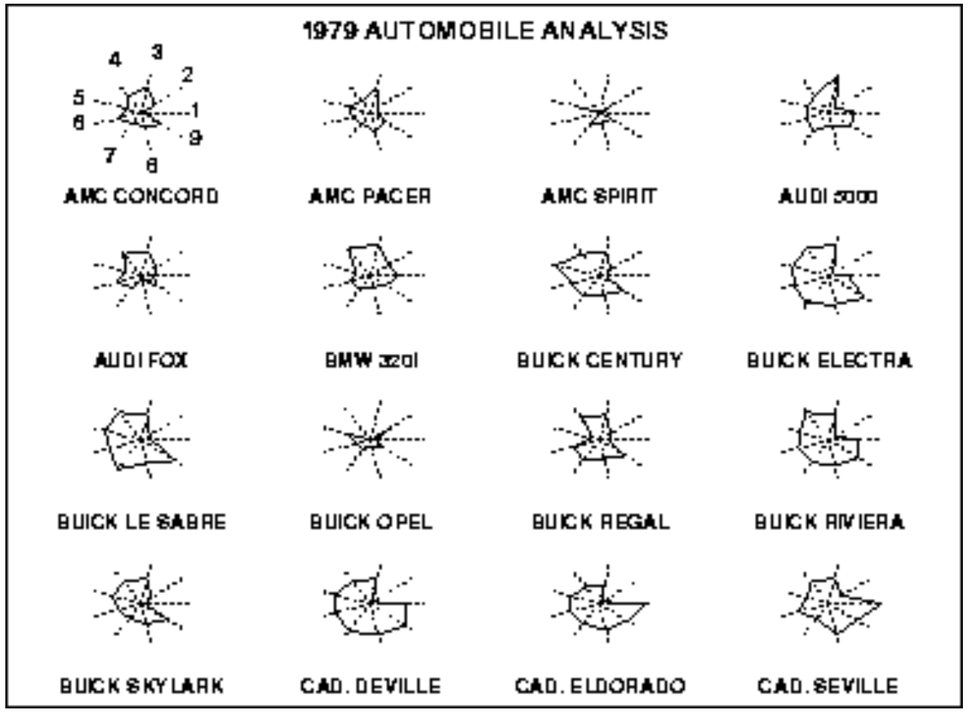
http://bl.ocks.org/jasondavies/1341281

Star Plot

- Similar to parallel coordinates
- Radiate from a common origin



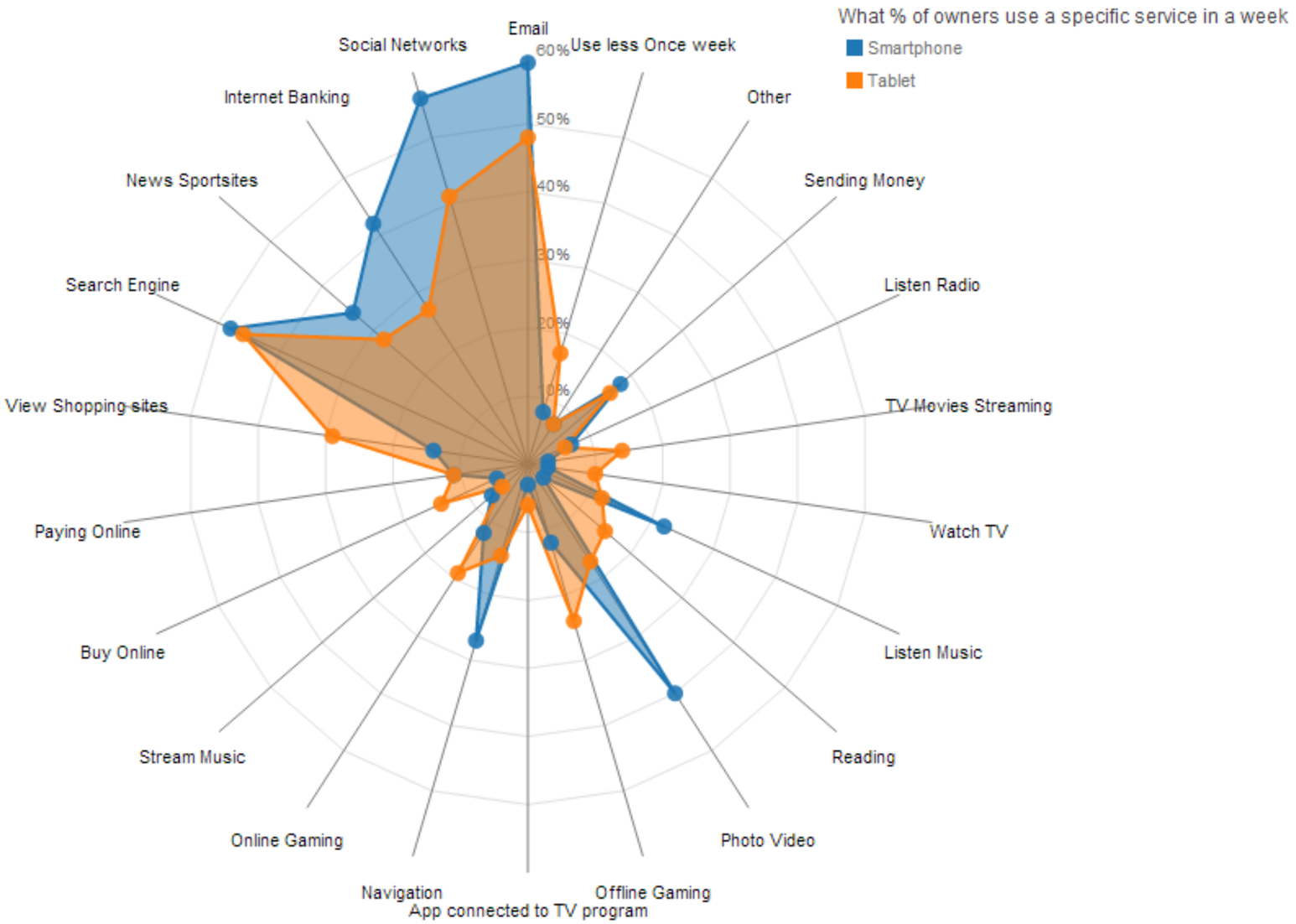
http://start1.jpl.nasa.gov/caseStudies/autoTool.cfm



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

[Coekin1969]

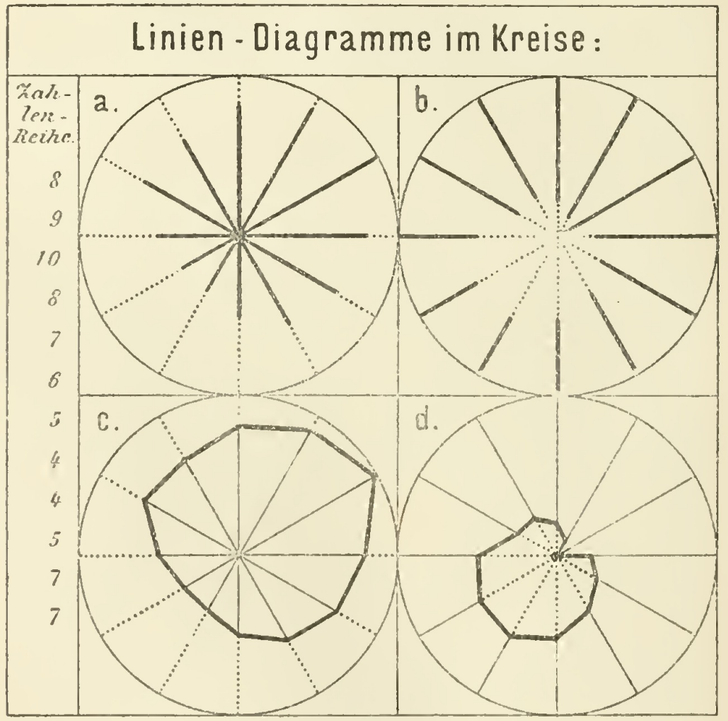
What do you think?



78 I. Abschnitt.

Zahlenergebnissen proportional ist. Auch können Verlängerungen der Radien über die Peripherie hinaus hiezu benützt werden. Zweckmäßig wird auch hier die lineare Verbindung der Endpunkte der betreffenden Geraden vorgenommen.

Beispiele von Linien-Diagrammen im Kreise sind in der folgenden Fig. 4 gegeben. Bei a und c bildet der Mittelpunkt, bei b und d die Peripherie den Ausgangspunkt der



Figur 4.

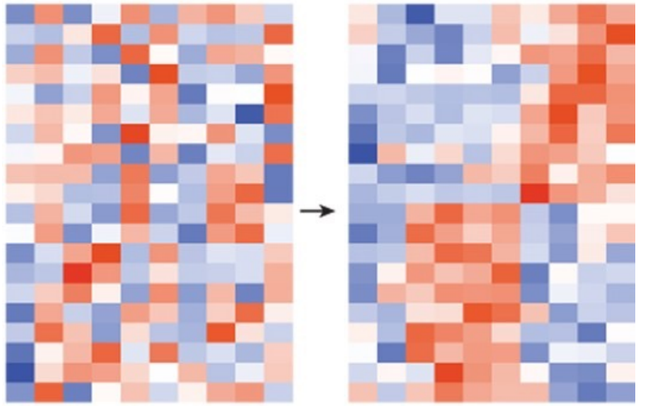
Geraden, welche als Radientheile von differenter Größe die Zahlenverschiedenheiten der statistischen Reihe darstellen. Bei a und b ist die Veranschaulichung lediglich durch

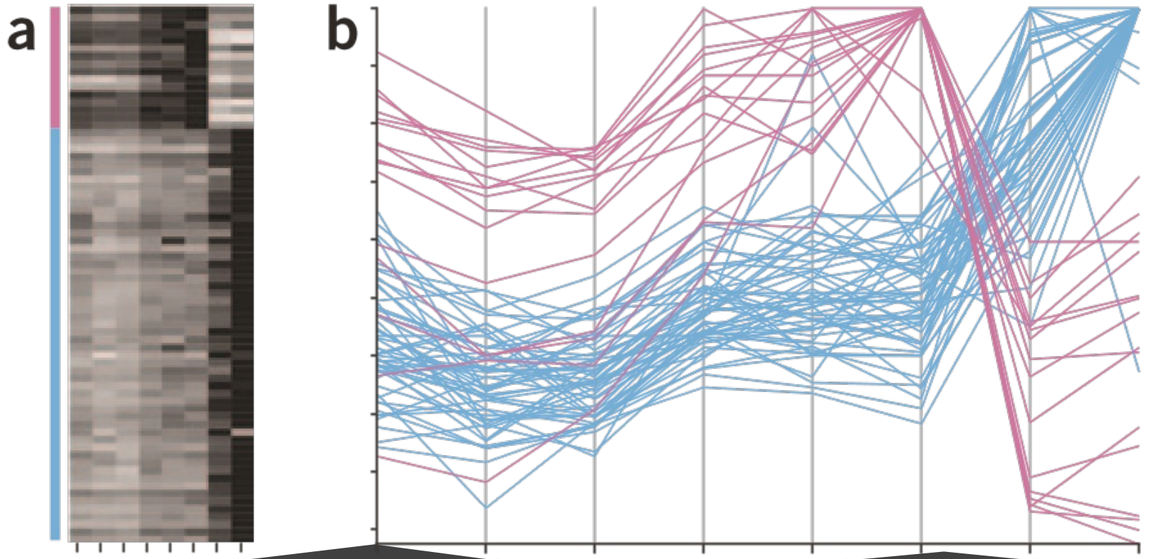
http://www.visualcinnamon.com/2013/09/making-d3-radar-chart-look-bit-better.html

Pixel-based Representations

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]

HiVE example: London property

partitioning attributes

house type
neighborhood
sale time

encoding attributes

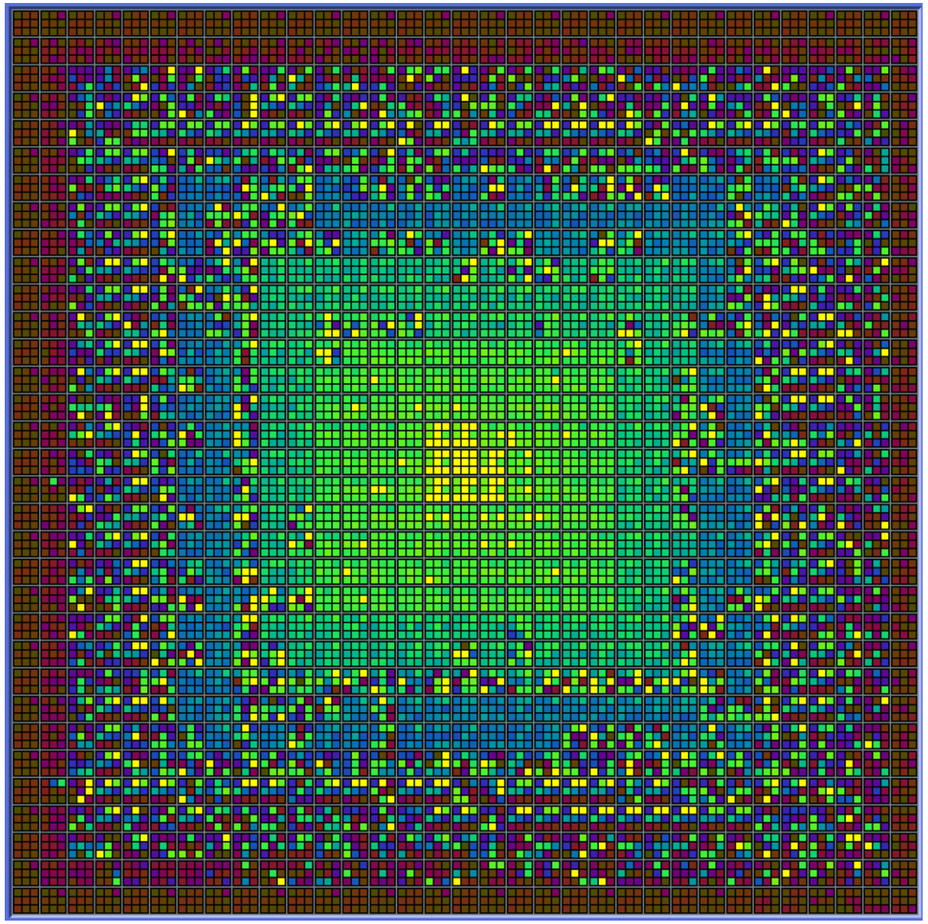
average price (color)
number of sales (size)

results

between neighborhoods, different housing distributions
within neighborhoods, similar prices

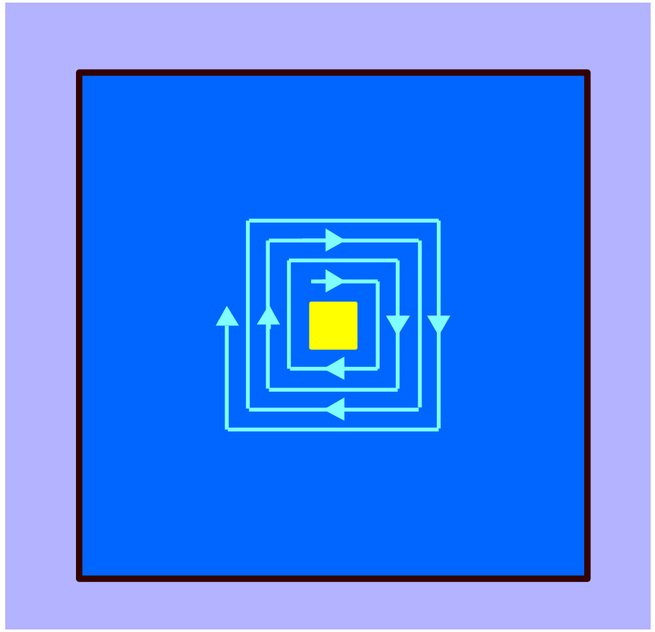
Dense pixel display: VisDB

- represent each data item, or each attribute in an item as a single pixel
- can fit as many items on the screen as there are pixels, on the order of millions
- relies heavily on color coding challenge: what's the layout?

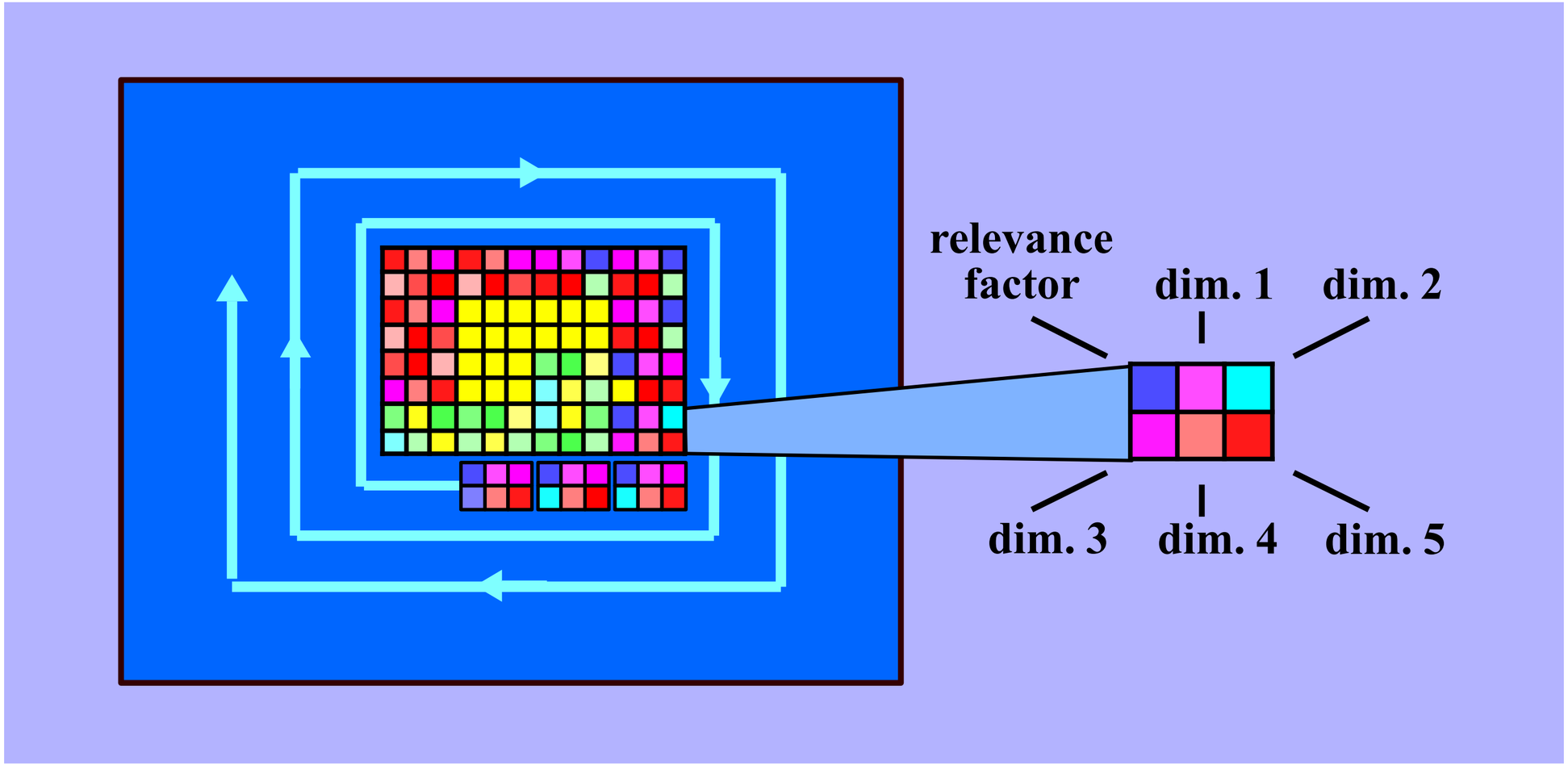


The data…

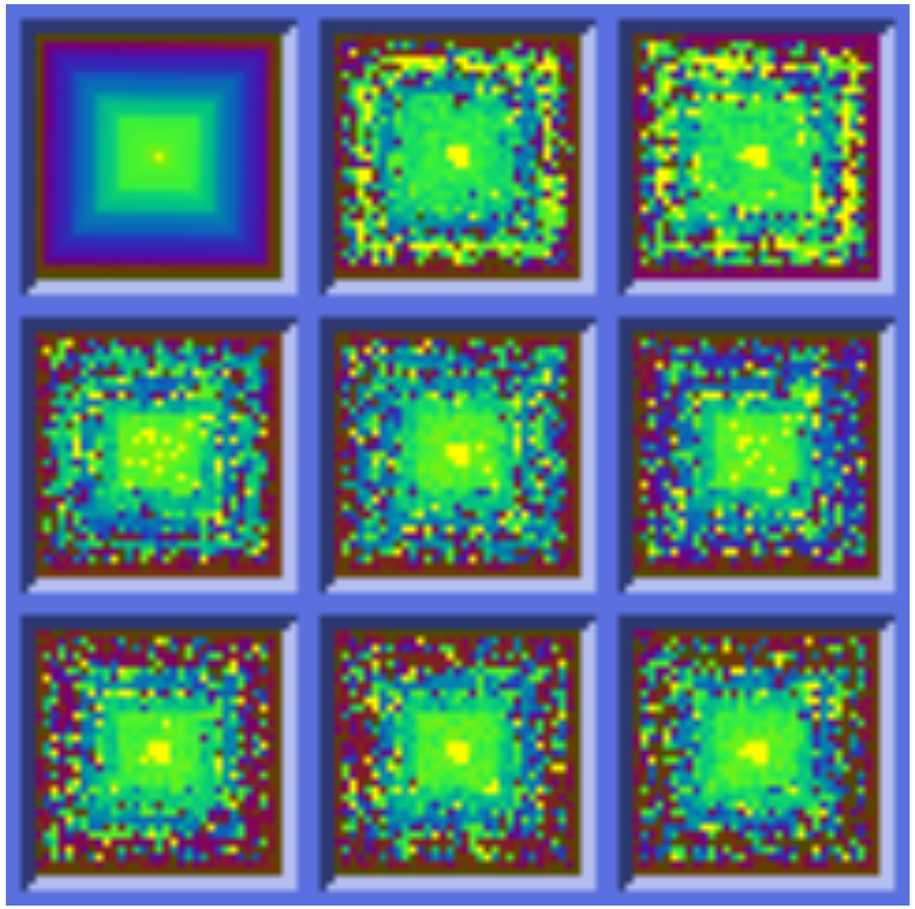
- large database where each item has multiple attributes (on the order of 10)
- goal: visualize the relevance of set of items which satisfy a query
- plot out data items in a spiral pattern, ordered by relevance

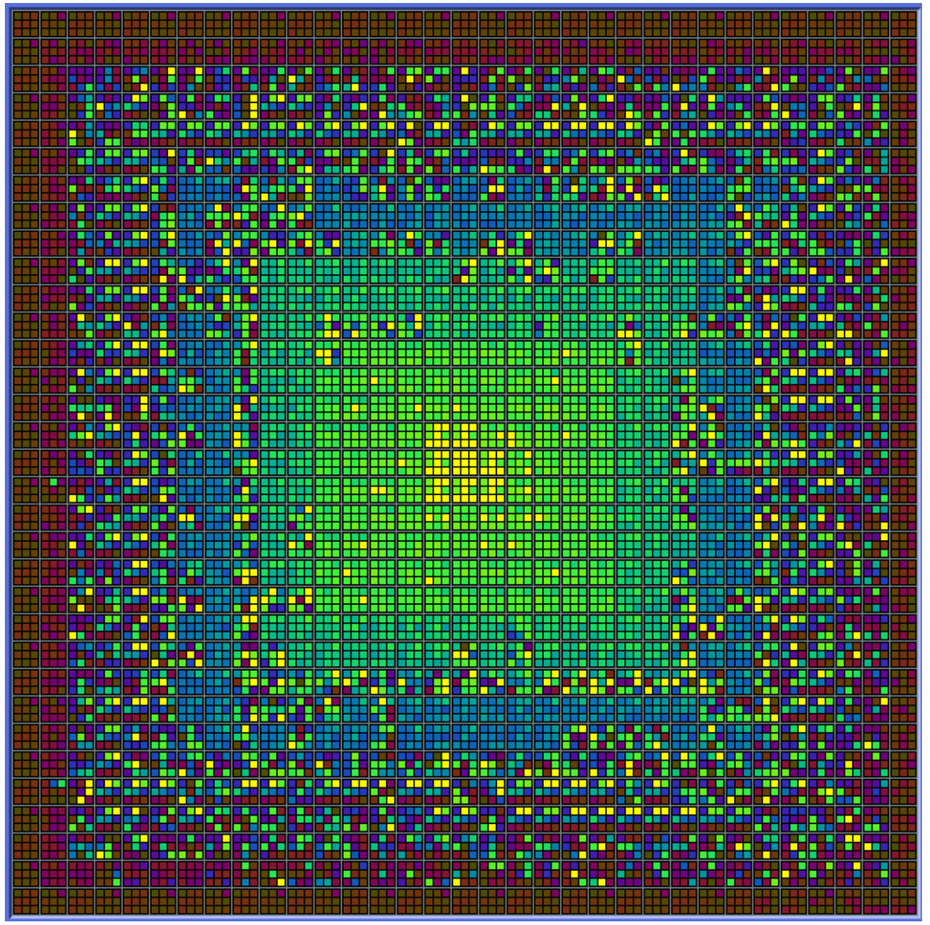


Keim, Kreigel, 1994



Keim, Kreigel, 1994





Keim, Kreigel, 1994