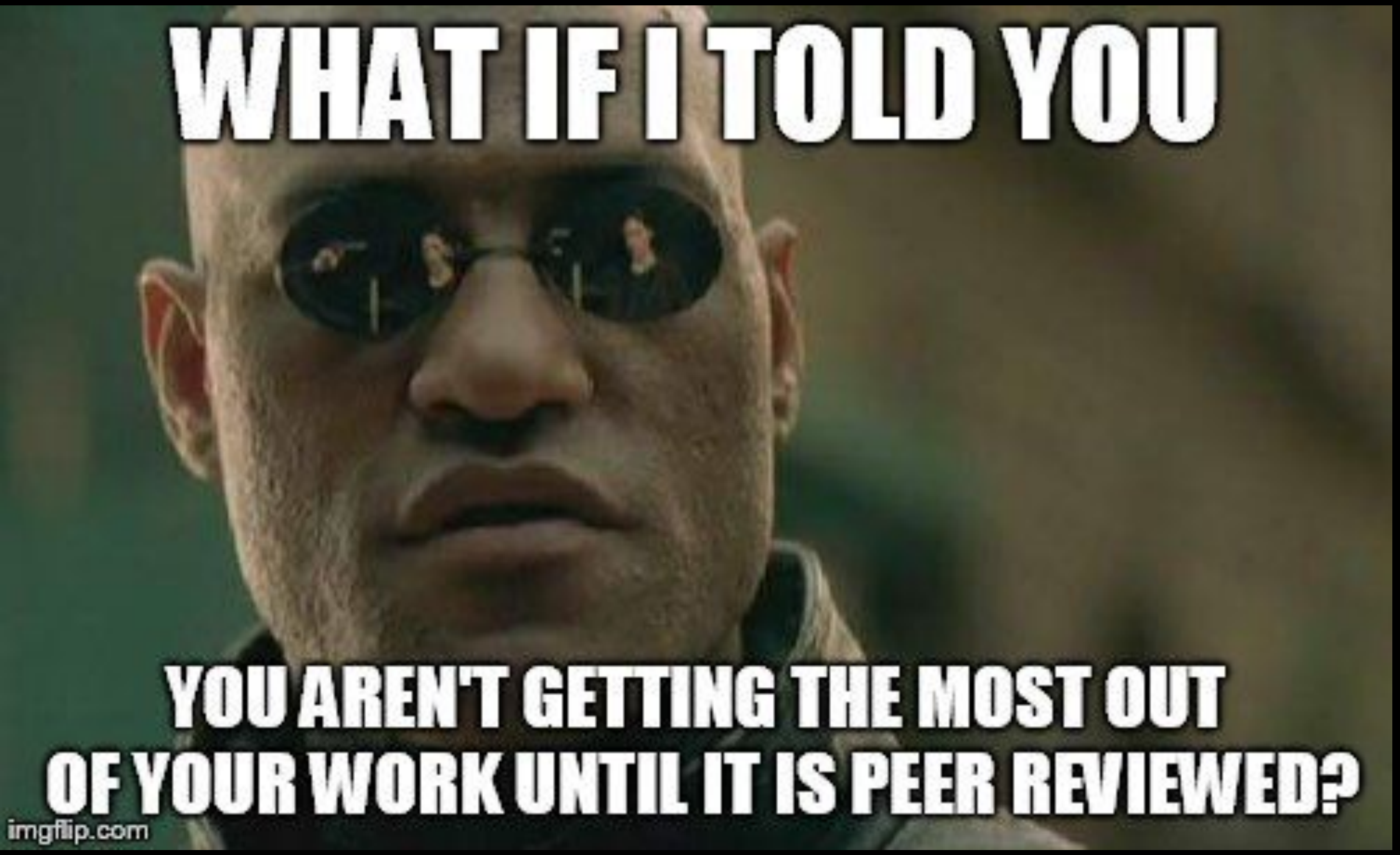
TABULAR DATA

KIRELL BENZI, PH.D

@KirellBenzi

Inspired by Lex

www.kirellbenzi.com


WHAT IF I TOLD YOU
YOU AREN'T GETTING THE MOST OUT OF YOUR WORK UNTIL IT IS PEER REVIEWED?
imgflip.com


REVIEW COMPLETED
TOAST TO YOU AND YOUR THOUGHTS
memecrunch.com


"ALL I WANTED WAS
SOME REVIEWS...."
imgflip.com

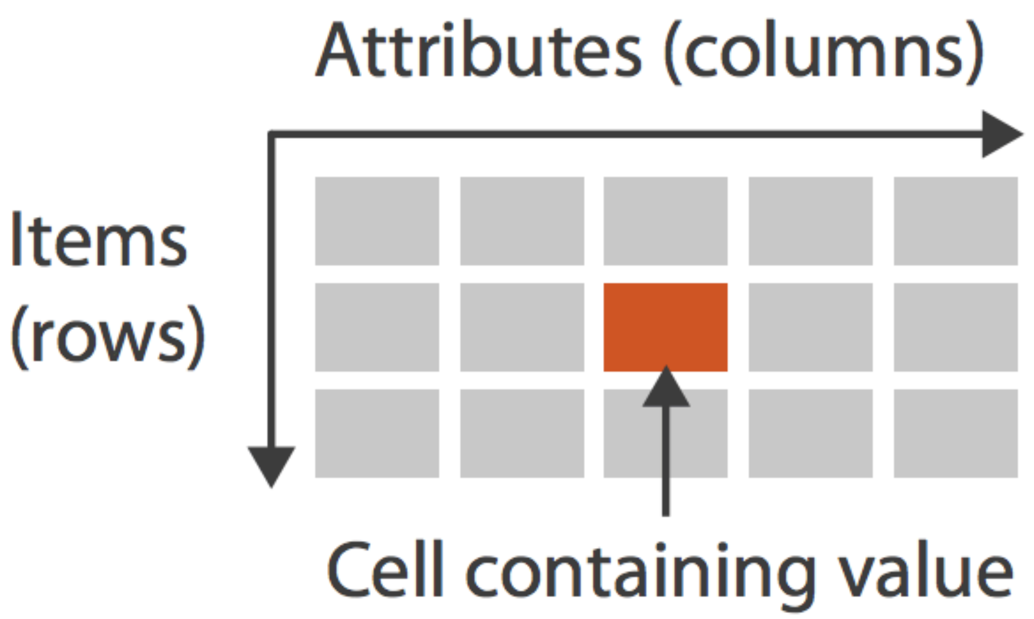

LET'S REVIEW


I WILL FIND YOU
AND I WILL THANK YOU
memegenerator.net

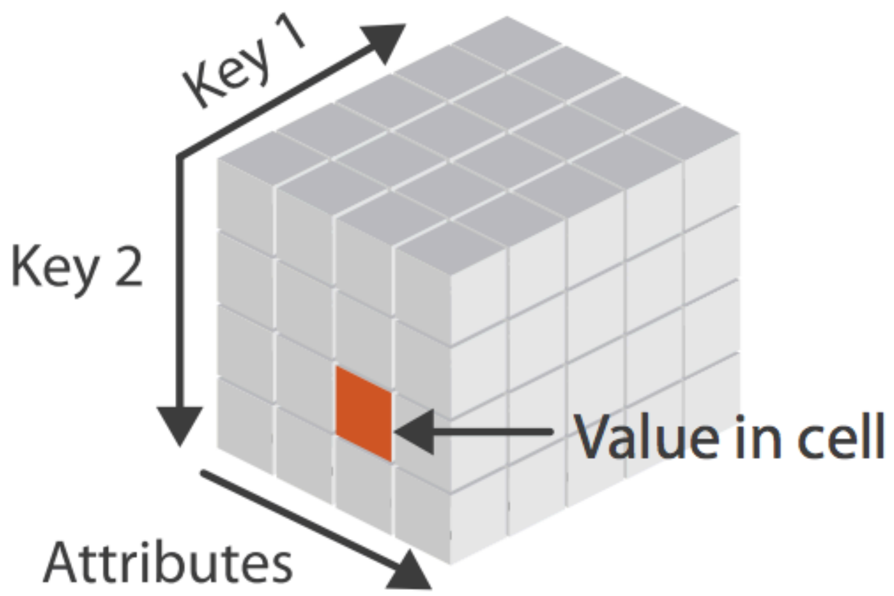

MID YEAR REVIEW...
WHO CARES?
memegenerator.net

Tabular data

→ Tables



→ *Multidimensional Table*



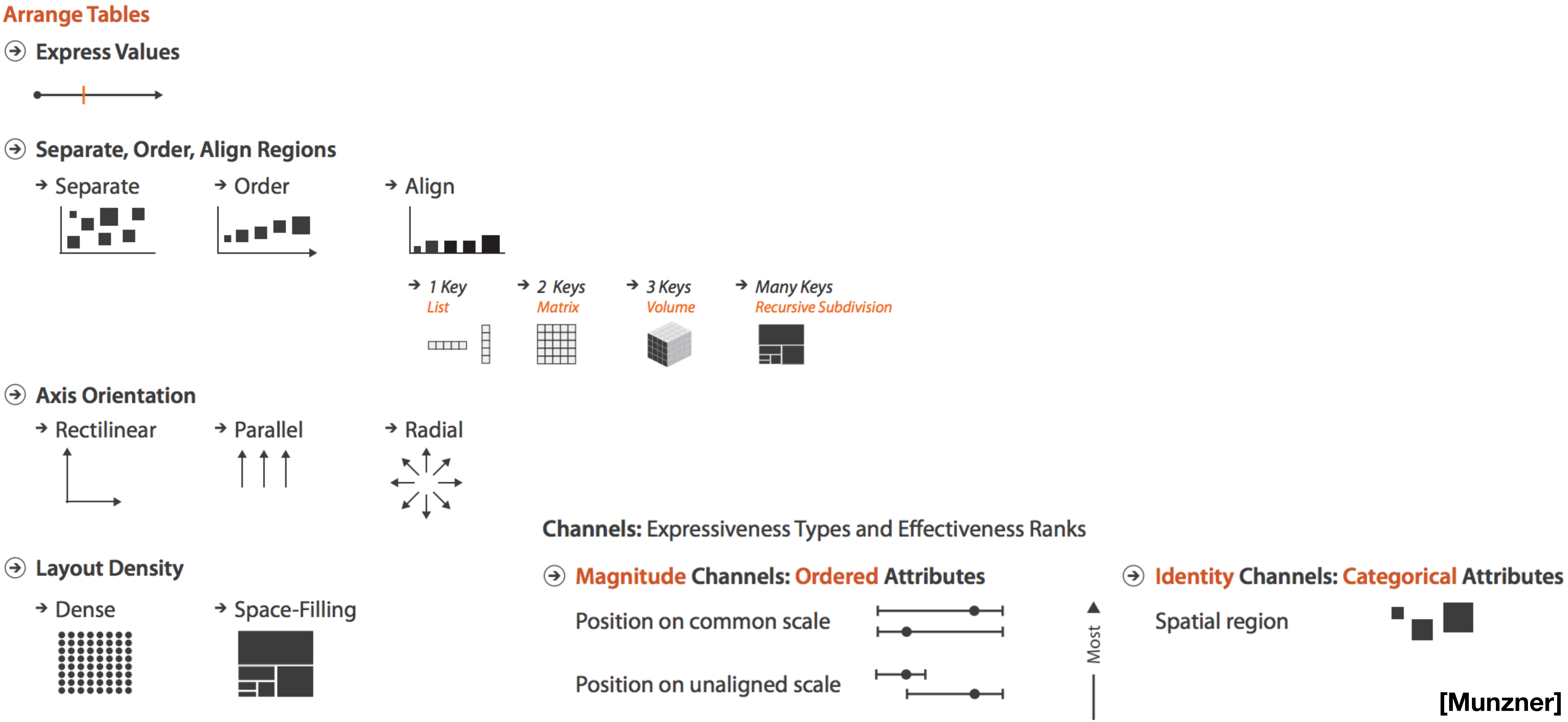
19 columns, **500+** rows

23.03 MB, Comma-Separated, UTF-8

album_id	album_comments	album_date_created	album_date_released	album_engineer	album_favorites	album_handle
1	0	11/26/2008 01:44:45 AM	1/05/2009		4	AWOL_-_A_Way_Of_Life
100	0	11/26/2008 01:55:44 AM	1/09/2009		0	On_Opaque_Things
1000	0	12/04/2008 09:28:49 AM	10/26/2008		0	DMBQ_Live_at_2008_Record_
10000	0	9/05/2011 04:42:57 PM			0	Live_at_CKUT_on_Montreal_
10001	0	9/06/2011 12:02:58 AM	1/01/2006		0	Grounds_Dream_Cosmic_Love
10002	0	9/06/2011 12:53:01 AM			0	Trust_Now
10003	0	9/06/2011 05:46:22 AM	9/06/2011		1	Or_Up_We_Fall

[Munzner]

Arranging tables overview



[Munzner]

Table viz scalability

Normal tables can be visualized “as is”

~ 50 attributes ~ 1000 records

High-dim tables need analytical methods

~ 1000 attributes >> 10’000 records

19 columns, **500+** rows

23.03 MB, Comma-Separated, UTF-8

album_id	album_comments	album_date_created	album_date_released	album_engineer	album_favorites	album_handle
1	0	11/26/2008 01:44:45 AM	1/05/2009		4	AWOL_-_A_Way_Of_Life
100	0	11/26/2008 01:55:44 AM	1/09/2009		0	On_Opaque_Things
1000	0	12/04/2008 09:28:49 AM	10/26/2008		0	DMBQ_Live_at_2008_Record_
10000	0	9/05/2011 04:42:57 PM			0	Live_at_CKUT_on_Montreal_
10001	0	9/06/2011 12:02:58 AM	1/01/2006		0	Grounds_Dream_Cosmic_Love
10002	0	9/06/2011 12:53:01 AM			0	Trust_Now
10003	0	9/06/2011 05:46:22 AM	9/06/2011		1	Or_Up_We_Fall

Table homogeneity?

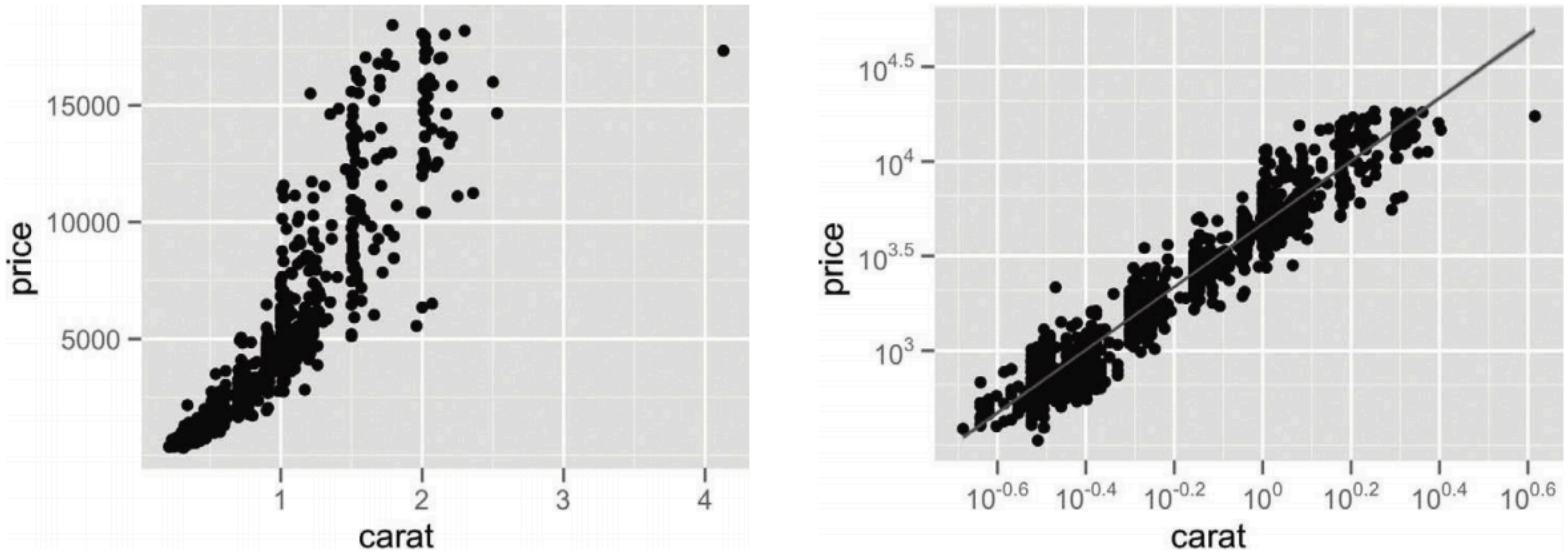
Scatterplots

Express Values

Show values directly

No keys

Derived attributes can be more expressive

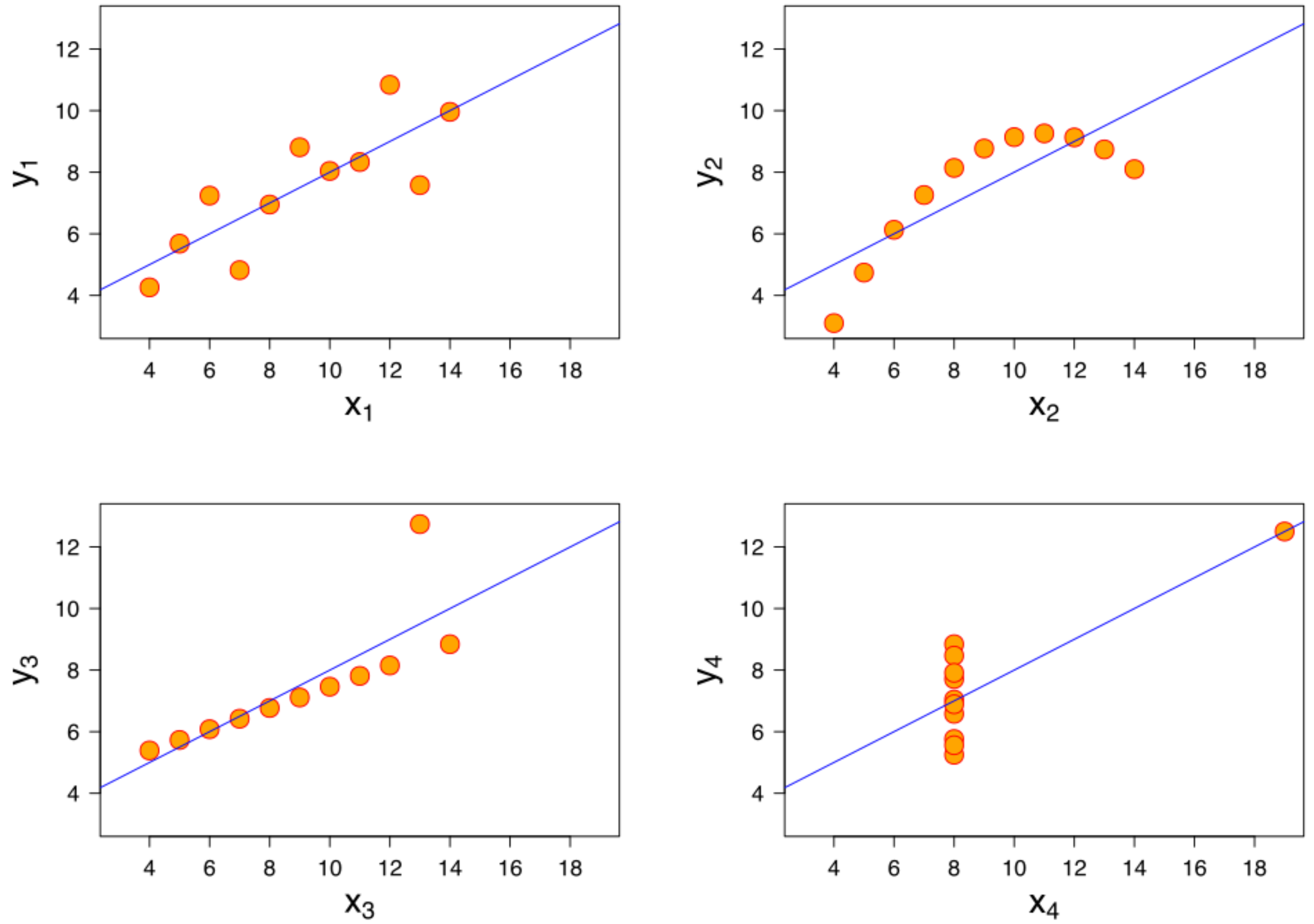


[Wickham 10]

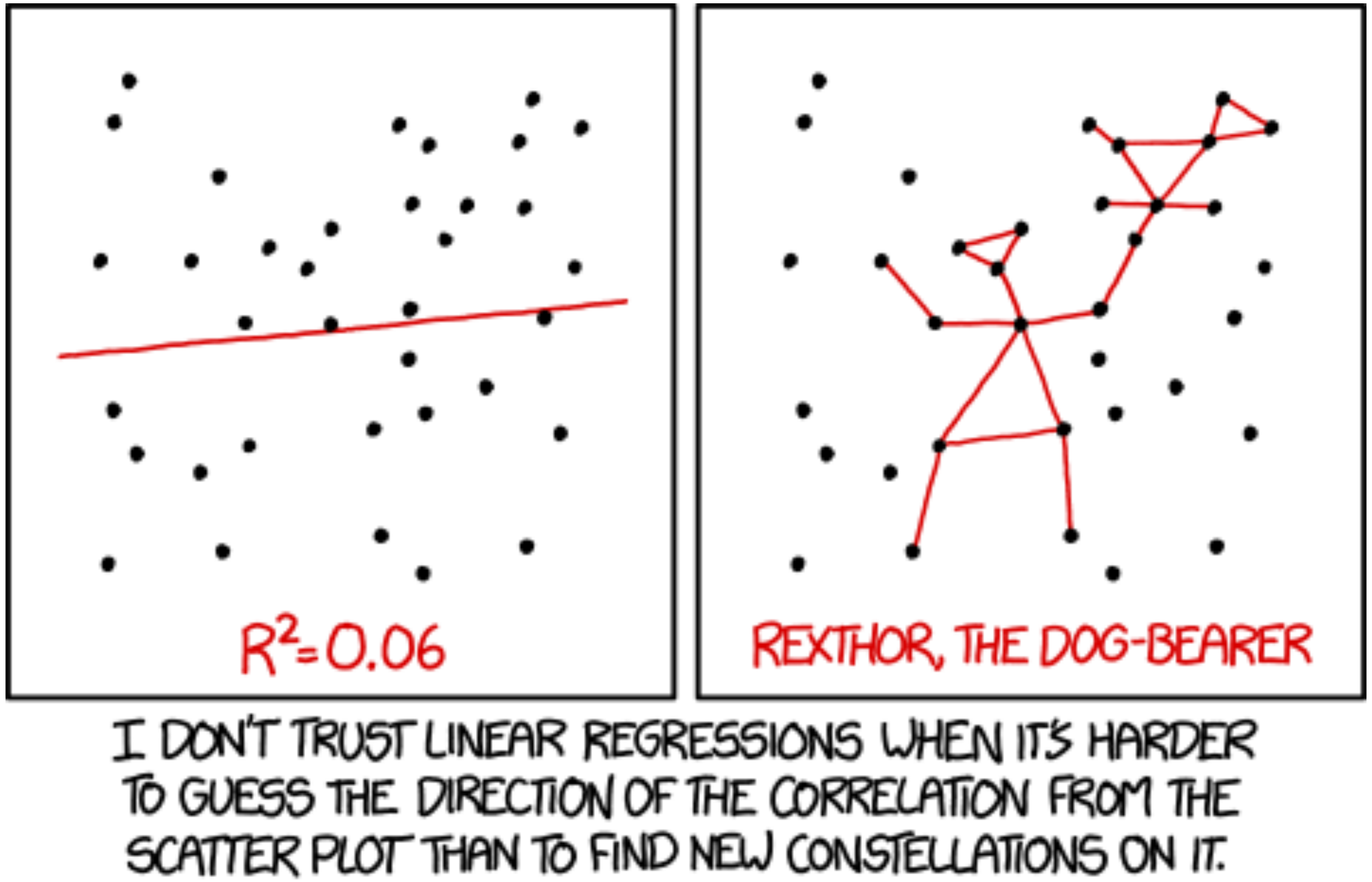
Showing trends

Helps showing trends at a quick glance

Most often use a linear regression



Anscombe's quartet



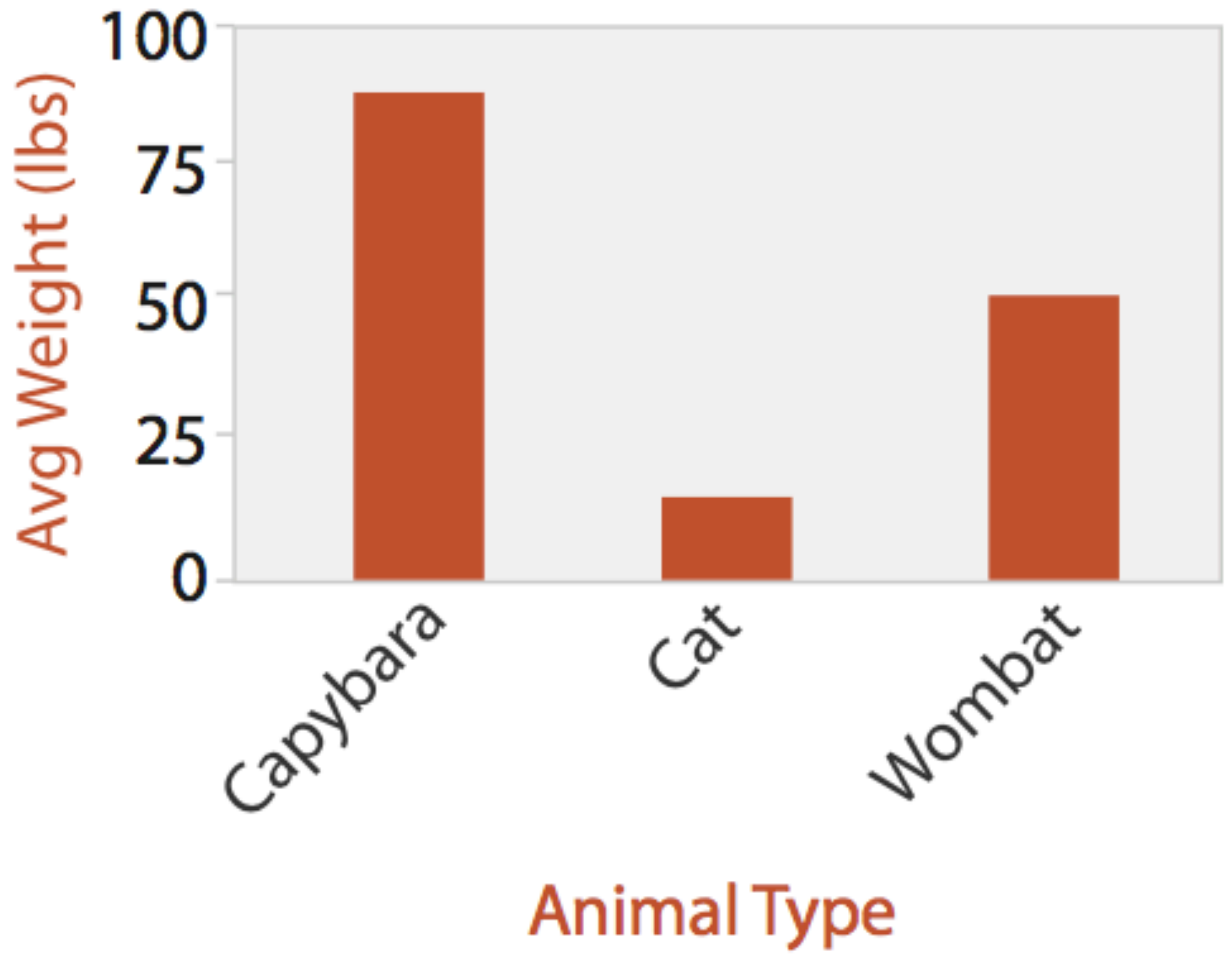
[xkcd]

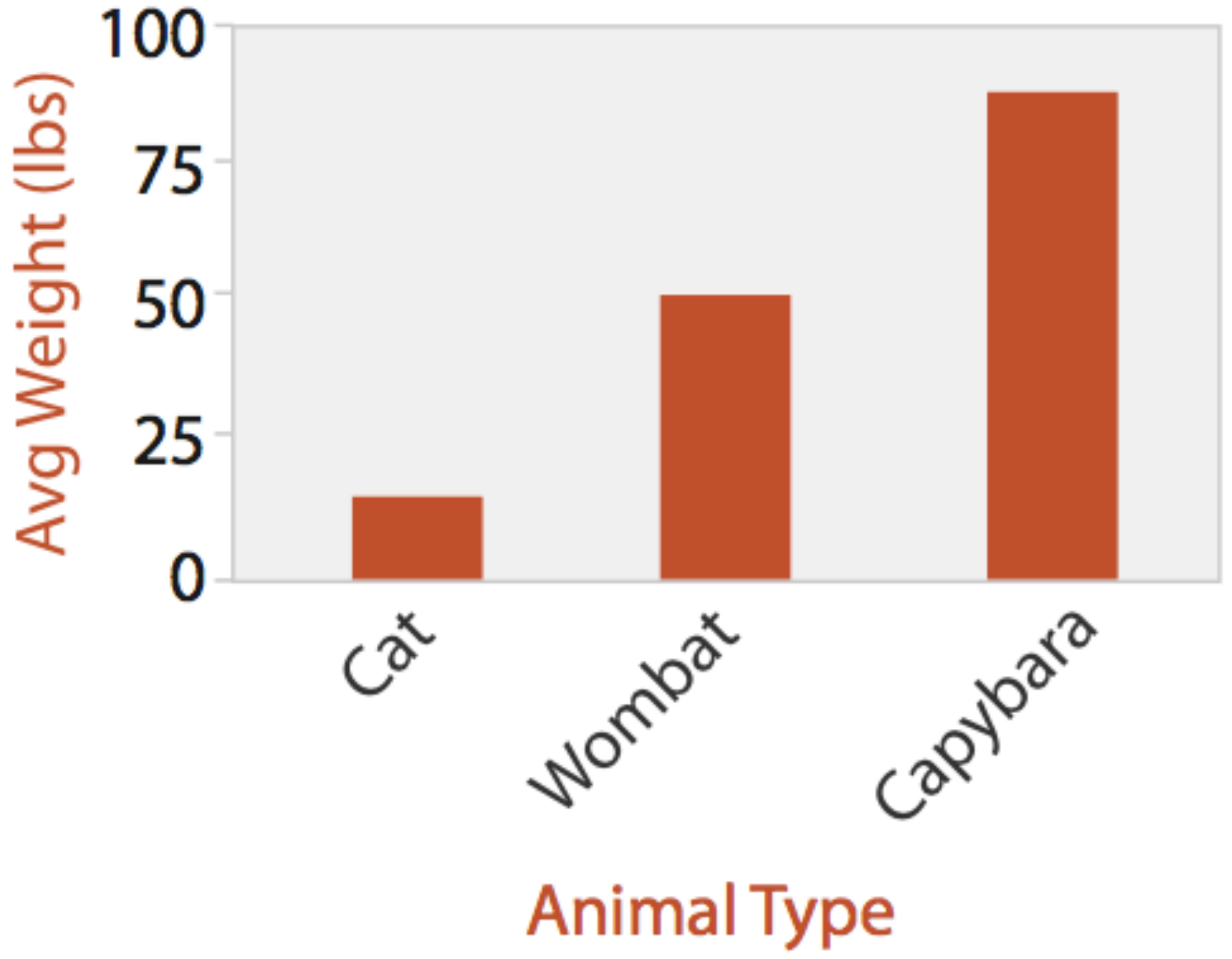
Bubble plots

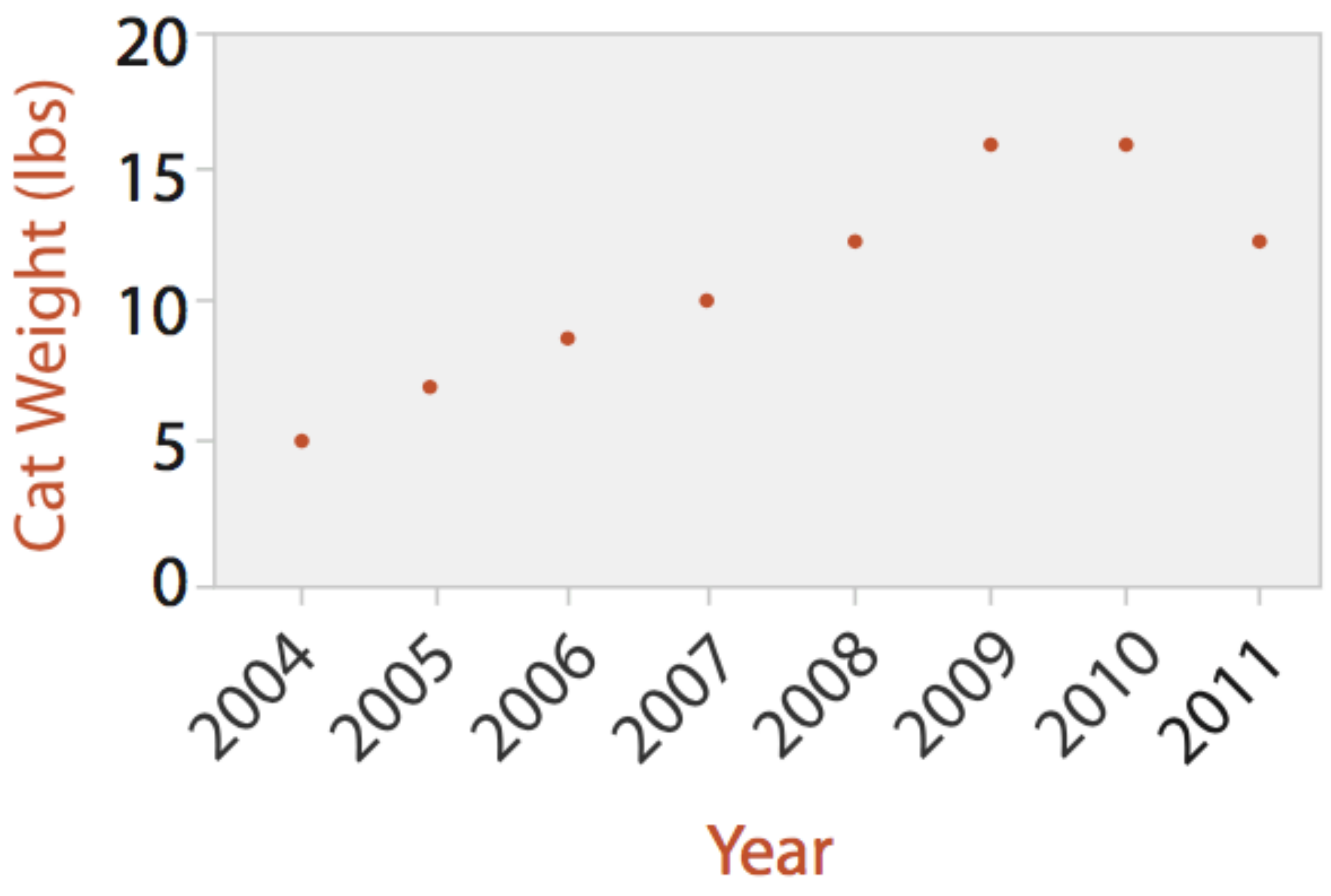
[Robertson et al. 08]

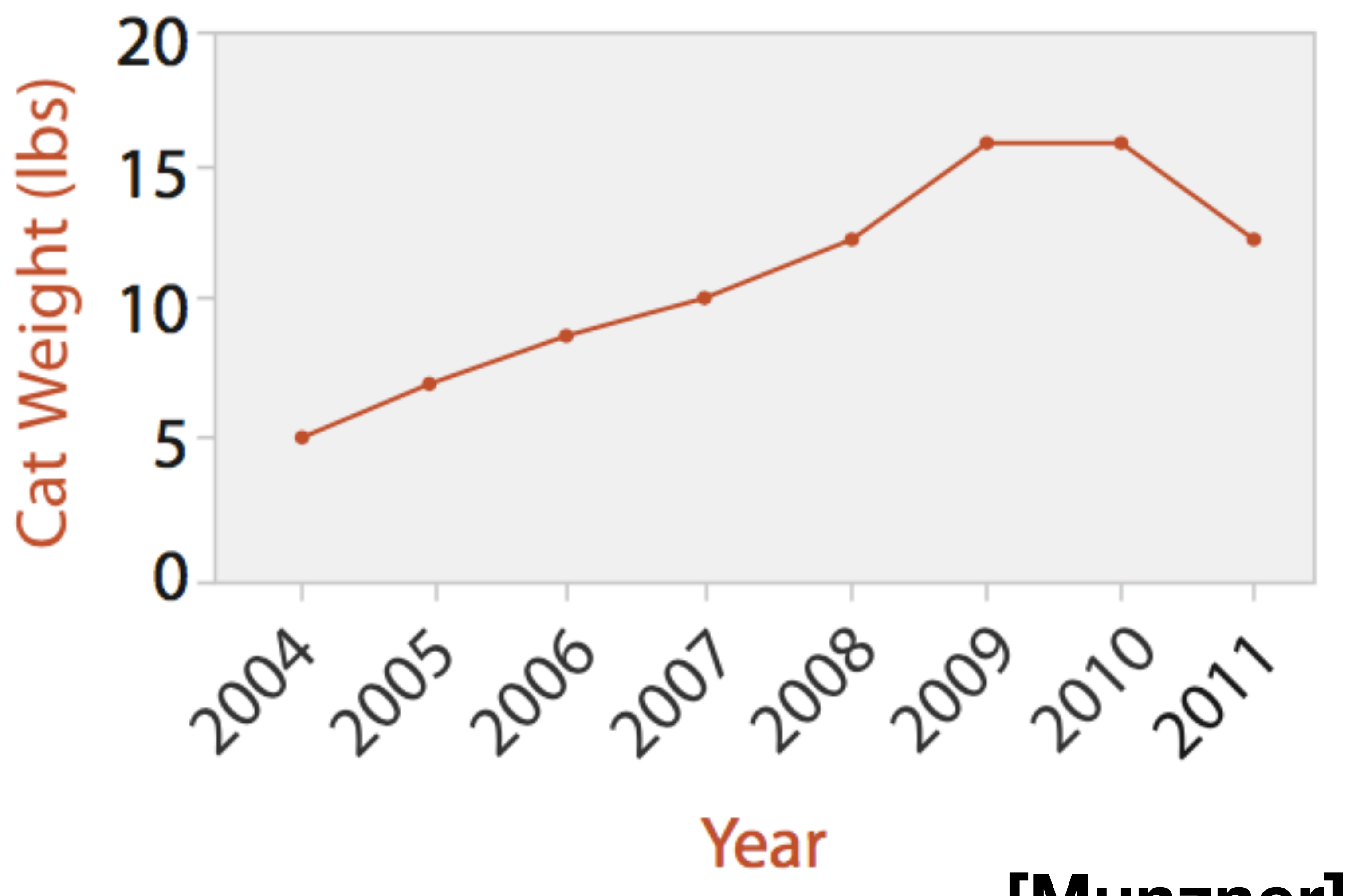
Simple bar, line, dot charts

Encode a single key attribute







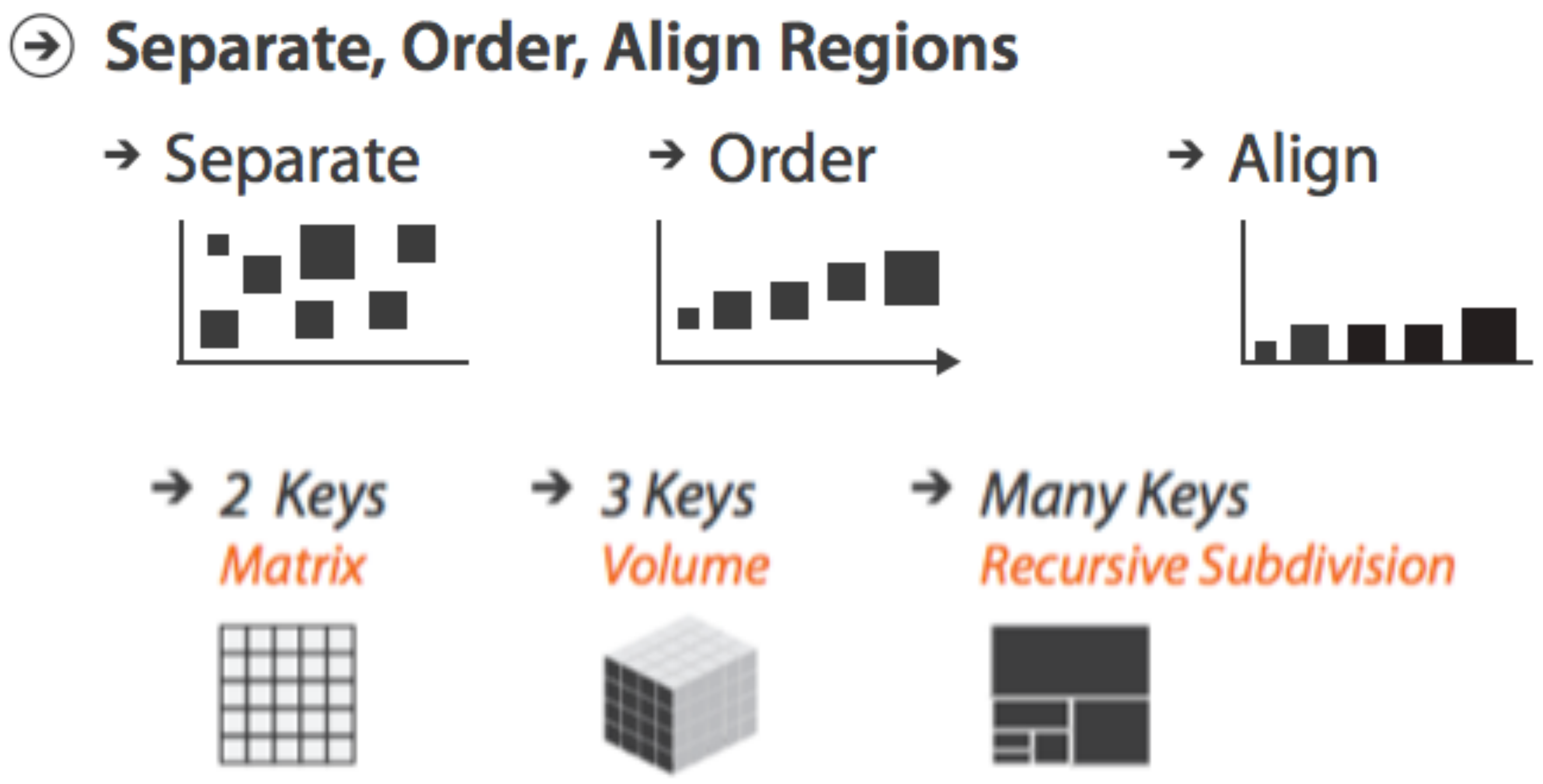


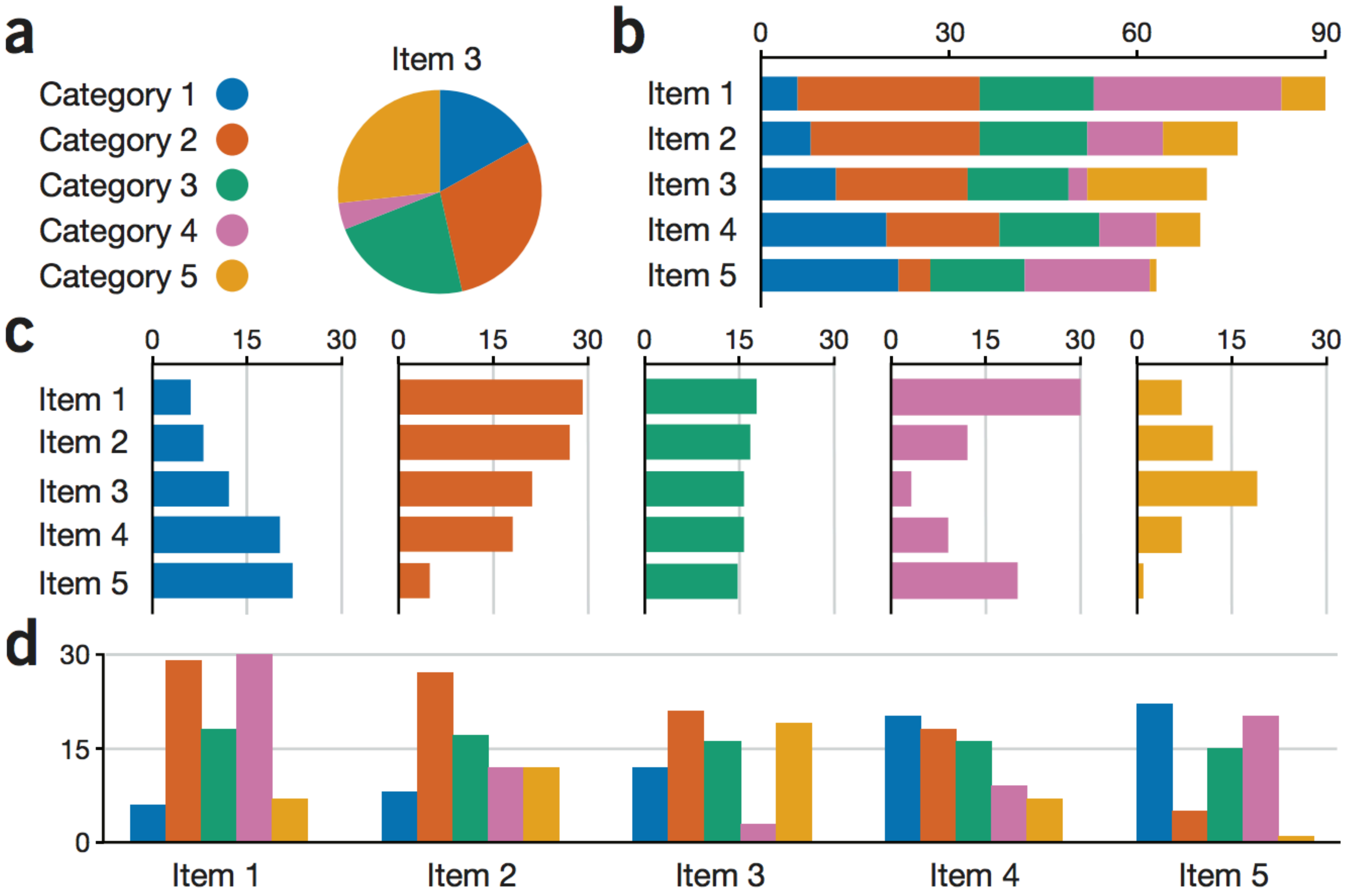
[Munzner]


Capybara

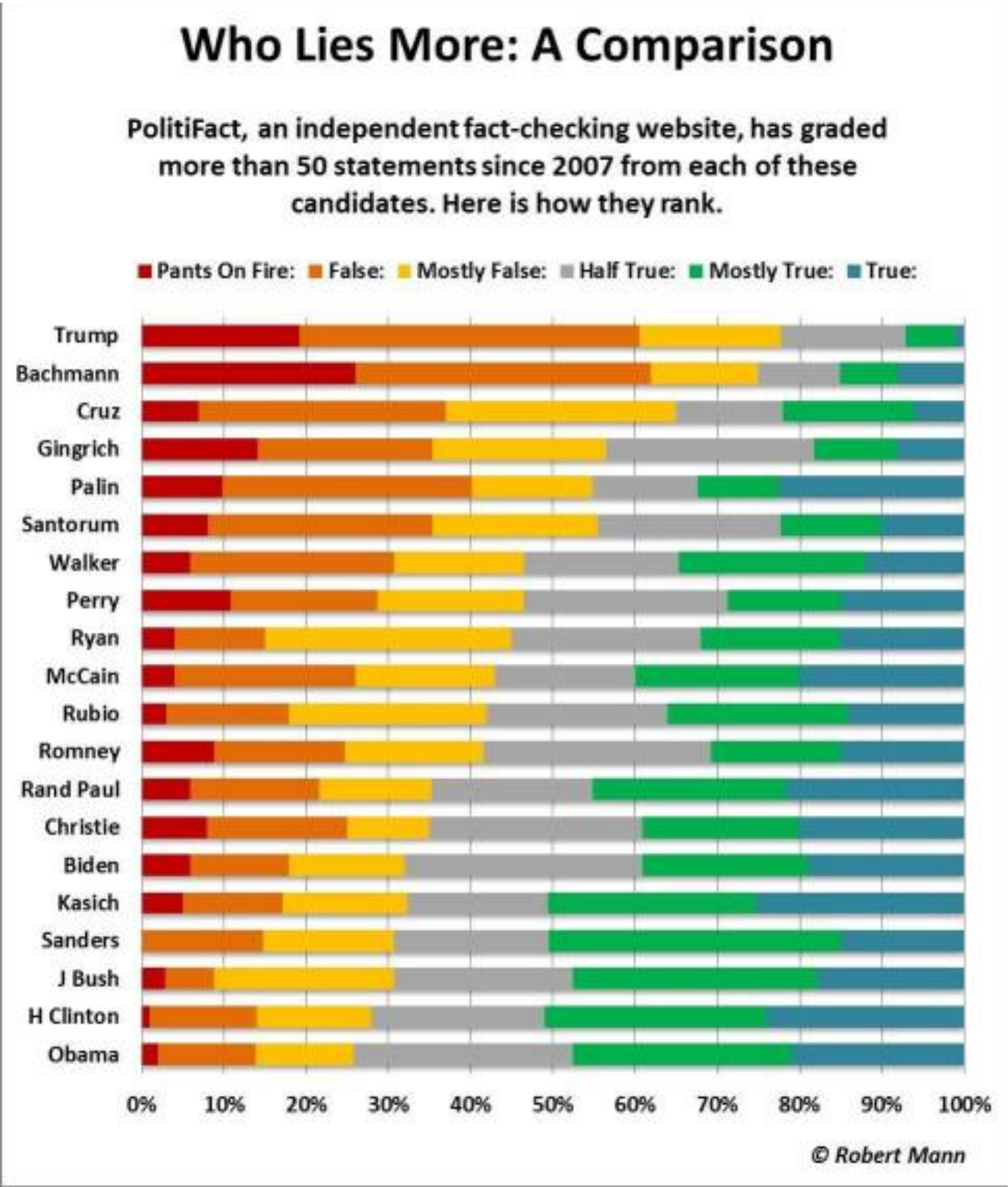

Wombat

Stacked bar charts

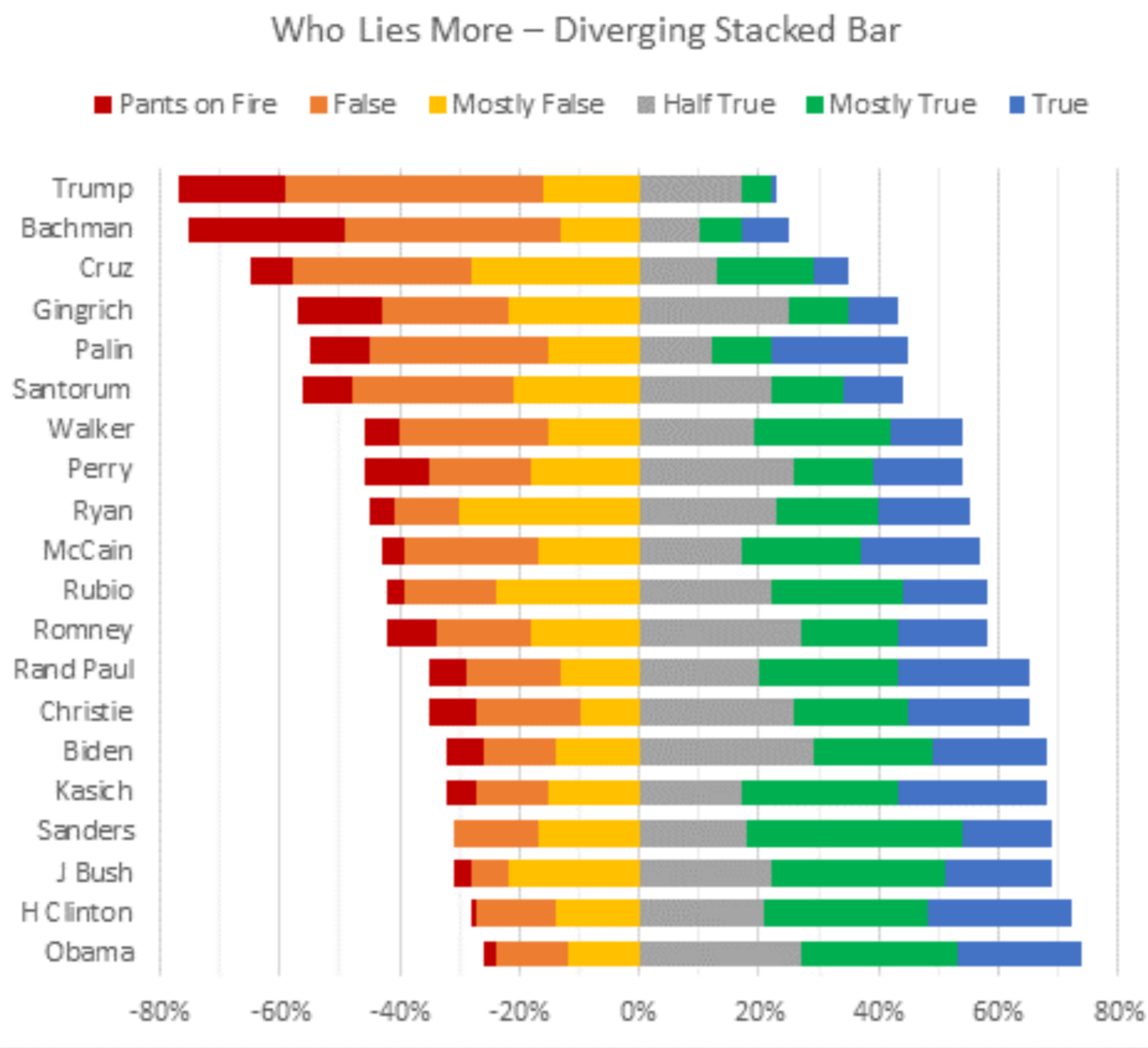




[Streit & Gehlenborg]



Who Lies More: A Comparison
PolitiFact, an independent fact-checking website, has graded more than 50 statements since 2007 from each of these candidates. Here is how they rank.
Pants On Fire: False: Mostly False: Half True: Mostly True: True:
Trump
Bachmann
Cruz
Gingrich
Palin
Santorum
Walker
Perry
Ryan
McCain
Rubio
Romney
Rand Paul
Christie
Biden
Kasich
Sanders
J Bush
H Clinton
Obama
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
© Robert Mann



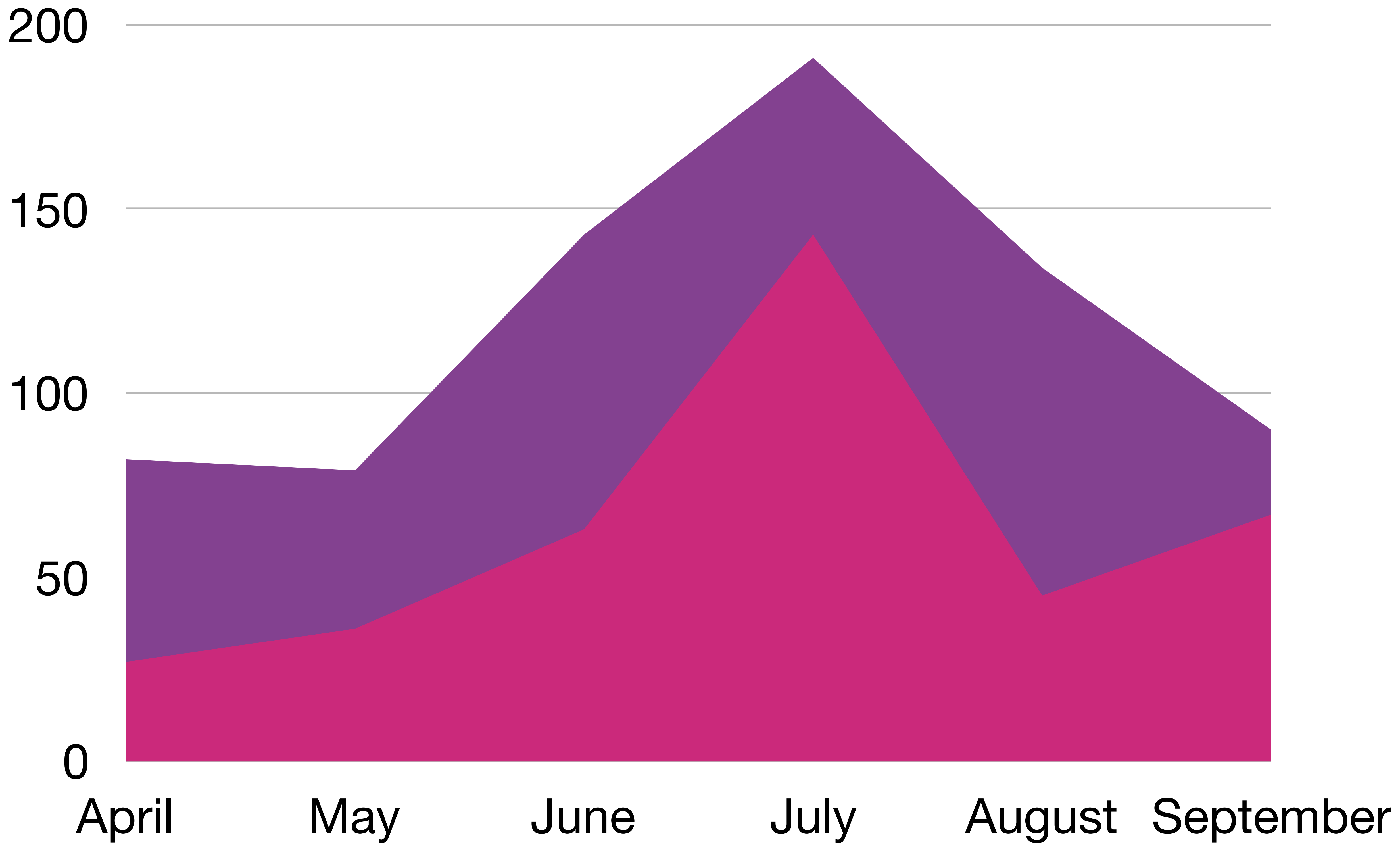
Who Lies More – Diverging Stacked Bar
Pants on Fire False Mostly False Half True Mostly True True
Trump
Bachman
Cruz
Gingrich
Palin
Santorum
Walker
Perry
Ryan
McCain
Rubio
Romney
Rand Paul
Christie
Biden
Kasich
Sanders
J Bush
H Clinton
Obama
-80% -60% -40% -20% 0% 20% 40% 60% 80%

Stacked area chart

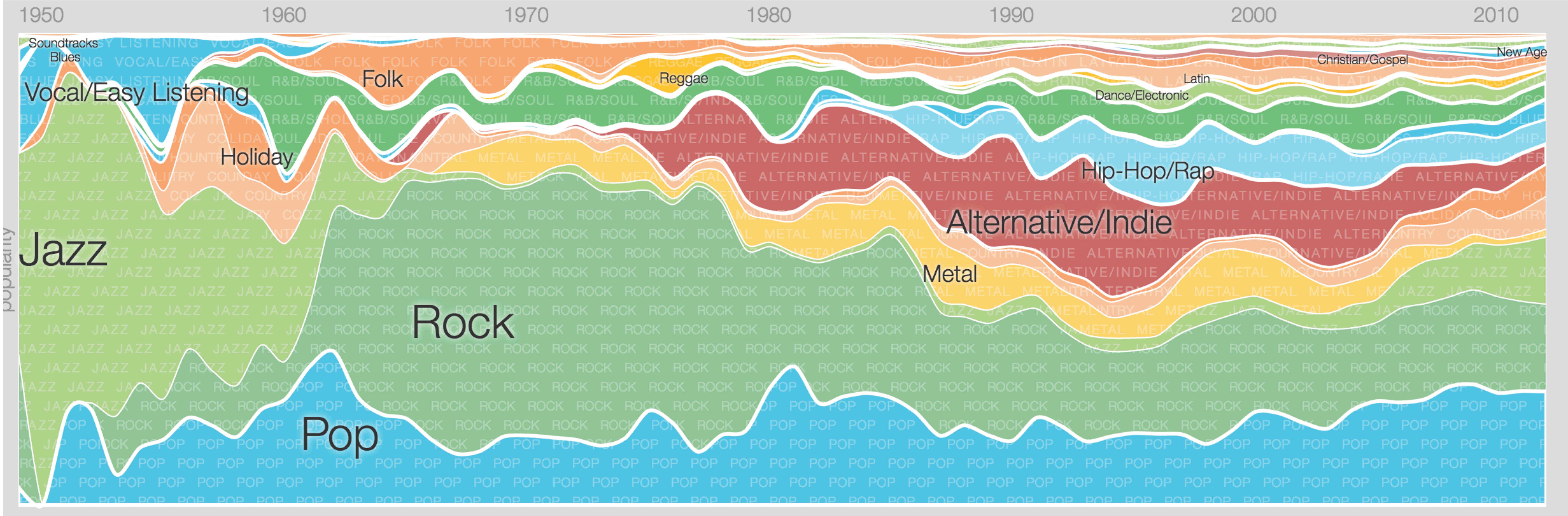
Color the area below the line that connects the values from the dataset

Display the development of quantitative values over an interval

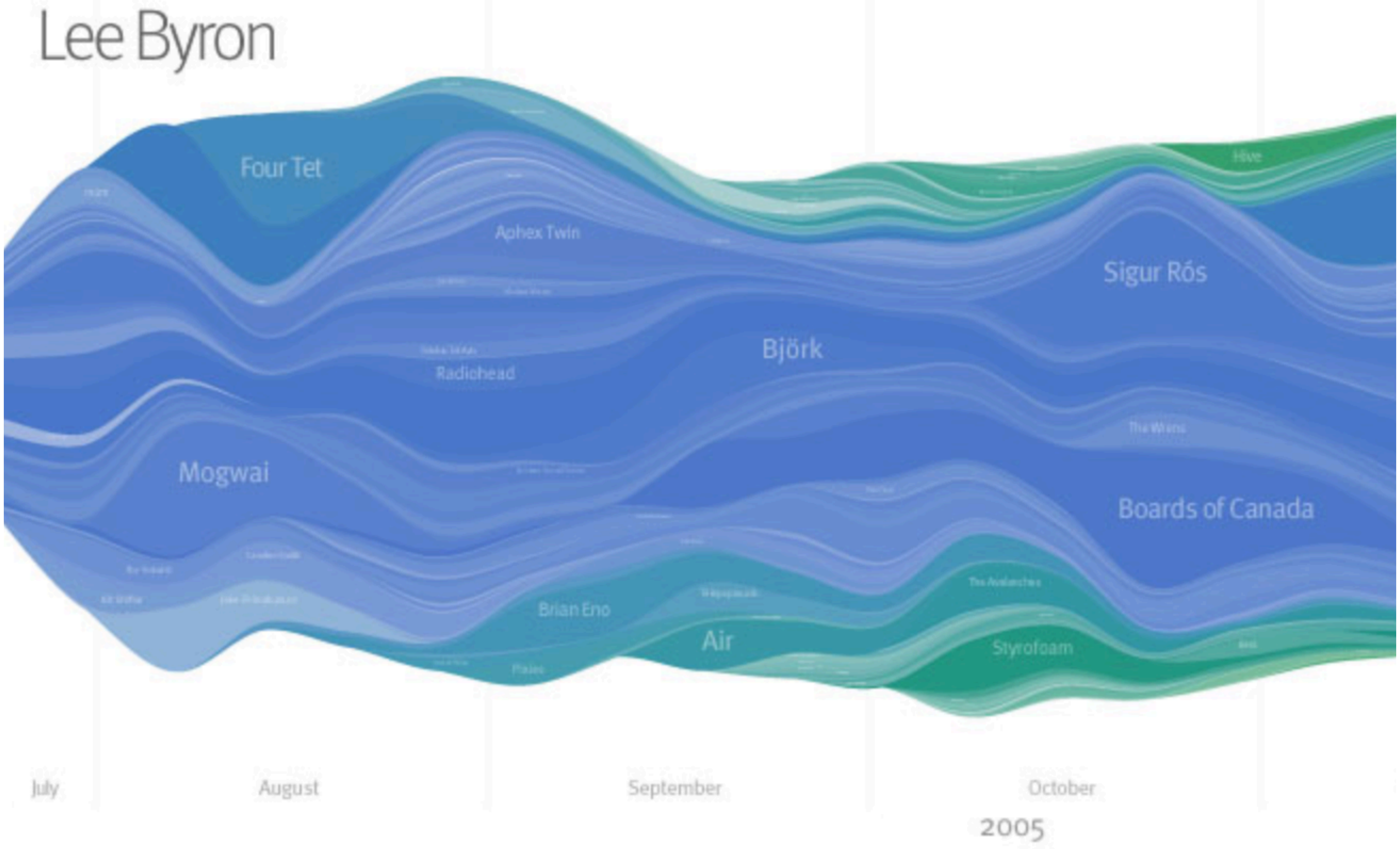
Ordering is paramount



100% stacked area chart



Streamgraph



Link to original paper

5 Algorithms for Stacked Graph Design

There are four main ingredients that determine a generalized stacked graph. The shape of the overall silhouette is the first ingredient; this shape is critical since it determines the overall slopes and curvature of the individual layers. The second important parameter is the ordering of the layers, which may be chosen to conform to different aesthetic criteria. As with any visualization, labels are important as well—but the organic forms of a Streamgraph mean that labels

lowing figures, we use a synthetic data set with randomly assigned

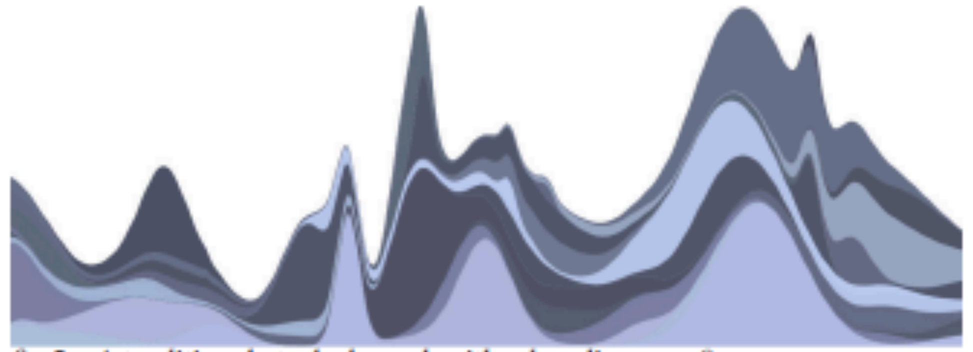


fig 5 – A traditional stacked graph with a baseline $g_0 = 0$

colors to distinguish layers.) In this layout the size of the sum of the series is easy to read, potentially at the expense of the legibility of the individual layers as mentioned in design issue (D).

A simple alternative layout was suggested by Havre et al in the ThemeRiver system. They used a layout symmetric around the x-axis. Mathematically, this can be expressed as:

$$g_0 + g_n = 0$$

or, from the definition of g_n,

$$2g_0 + \sum_{i=1}^{n} f_i = 0$$

which yields the ThemeRiver (*fig 6*) solution for g_0:

$$g_0 = -\frac{1}{2}\sum_{i=1}^{n} f_i$$

This is a simple enough definition, but there is another way to look at this formula which suggests some generalizations. What is special about the symmetric layout? Aside from a certain aesthetic quality, this layout has the effect of minimizing some important quantities. In particular, at each point, the silhouette is as close as possible to the x-axis, and in addition the slopes of the top and bottom of the silhouette are in a sense as small as possible (in the sense of total sums of squares). This directly addresses design issue (C) by making the overall graph much less "spiky" thus greatly reducing the horizontal space needed to satisfy Cleveland's principal. To see this, recall the following fact about averages. For any set of real numbers $\{a_1,\ldots, a_n\}$, the value of x that minimizes

$$\sum_{i=1}^{n}(x + a_i)^2$$

is

$$x = -\frac{1}{n}\sum_{i=1}^{n} a_i$$

From this fact it follows that the value of g_0 that gives the symmetric ThemeRiver layout minimizes the sum of squares of the top and bottom of the silhouette of the graph (at each point in [0,1]):

$$silhouette(g_0) = g_0^2 + g_n^2$$

since

$$g_0^2 + g_n^2 = g_0^2 + (g_0 + \sum_{i=1}^{n} f_i)^2$$

A similar calculation shows that the ThemeRiver layout also min-

bottom, as in the case *i=0*, we take the sum to be empty and equal to zero.)

We might also be interested in minimizing the sum of squares of the slopes at each value of x:

$$wiggle(g_0) = \sum_{i=0}^{n} g_i'^2 = \sum_{i=0}^{n}(g_0' + \sum_{j=1}^{i} f_j')^2$$

By the same logic, the deviation quantity is minimized when

$$g_0 = -\frac{1}{n+1}\sum_{i=0}^{n}\sum_{j=1}^{i} f_j = -\frac{1}{n+1}\sum_{i=1}^{n}(n-i+1)f_i$$

Moreover, differentiating this yields:

$$g_0' = -\frac{1}{n+1}\sum_{i=0}^{n}\sum_{j=1}^{i} f_j'$$

which minimizes the *wiggle* measure as well. In other words, this choice of g_0 has the effect of simultaneously minimizing both distance from the x-axis and variation in slope. Moreover, the formula is extremely efficient to calculate. As *fig 7* shows, it creates a more "even" layout compared to the ThemeRiver graph. This minimization directly addresses design issue (A) by attempting to reduce the effect of middle "wiggles" on the surrounding layers.

Nonetheless, the layout can be improved further. In particular, it is potentially problematic that very thin layers receive the same treatment as very thick ones. After all, the thick layers are visually more important. Thus we might want instead to optimize the following quantity, which represents an average of the squares of slopes between the midpoints of each layer, weighted by layer thickness:

$$weighted\_wiggle(g_0) =$$

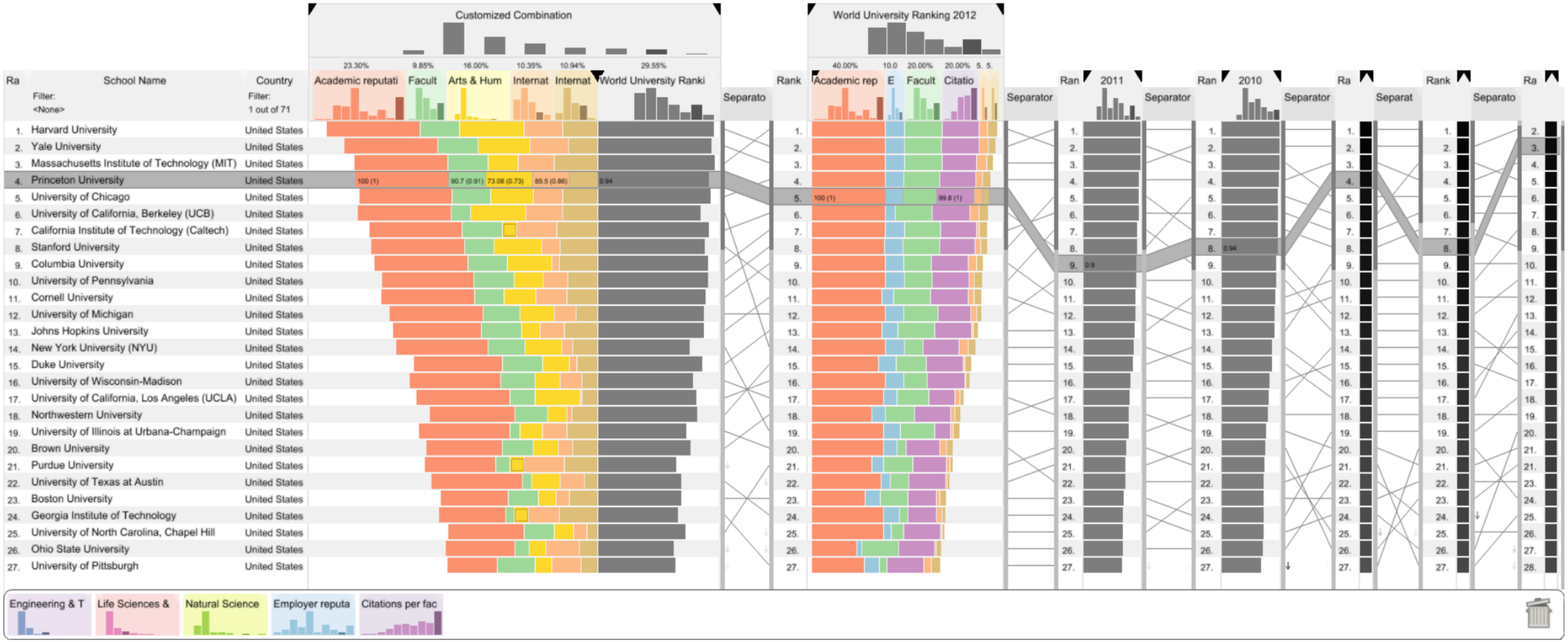
$$\sum_{i=1}^{n}(\frac{1}{2}(g_i' + g_{i-1}'))^2 f_i = \sum_{i=1}^{n}(g_0' + \frac{1}{2}f_i' + \sum_{j=1}^{i-1} f_j')^2 f_i$$

By the properties of weighted averages, this is minimized at each point in [0,1] when

$$g_0' = -\frac{1}{\sum f_i}\sum_{i=0}^{n}(\frac{1}{2}f_i' + \sum_{j=1}^{i-1} f_j')f_i$$

which can be integrated numerically to yield a solution for g_0. Indeed, it turns out to be equivalent to the algorithm used in the Streamgraph method, as portrayed in Listening History and Box Office Revenue graph. While this formula, which minimizes weighted wiggle, is not

LineUp: Rankings



[Gratzl et al.]

Demos Evaluation Window
World University Ranking 2012
Ra
School Name
Country
Academic reput
Employer reputa
Faculty/student r
Citations per fac
International fac
International stu
Filter: <None>
Filter: <None>
Aalborg University Denmark
2. Aalto University Finland
3. Aarhus University Denmark
4. Aberystwyth University United Kingdom
5. Abo Akademi University Finland
6. Al-Farabi Kazakh National University Kazakhstan
7. Al-Imam Mohamed Ibn Saud Islamic University Saudi Arabia
8. American University in Cairo Egypt
9. American University of Beirut (AUB) Lebanon
10. American University of Sharjah (AUS) United Arab Emi
11. Aristotle University of Thessaloniki Greece
12. Arizona State University United States
13. Aston University United Kingdom
14. Ateneo de Manila University Philippines
15. Auckland University of Technology New Zealand
16. Australian National University Australia
17. Bandung Institute of Technology (ITB) Indonesia
18. Bangor University United Kingdom
19. Bauman Moscow State Technical University Russia
20. Beijing Institute of Technology China
21. Beijing Normal University China
22. Ben Gurion University of The Negev Israel
23. Bilkent University Turkey
24. Birkbeck College, University of London United Kingdom
25. Bond University Australia
26. Boston College United States
27. Boston University United States
28. Brandeis University United States
29. Brown University United States
30. Brunel University United Kingdom
31. California Institute of Technology (Caltech) United States
32. Cardiff University United Kingdom
33. Carnegie Mellon University Australia
34. Case Western Reserve University United States
35. Chalmers University of Technology Sweden
36. Chang Gung University Taiwan
37. Charles University Czech Republic
38. Chiba University Japan
39. Christian-Albrechts-Universität zu Kiel Germany
40. Chulalongkorn University Thailand
41. City University London United Kingdom
42. City University of Hong Kong Hong Kong S.A.
43. City University of New York United States
0:36 / 4:57
HD

Pixel-based display (dense)

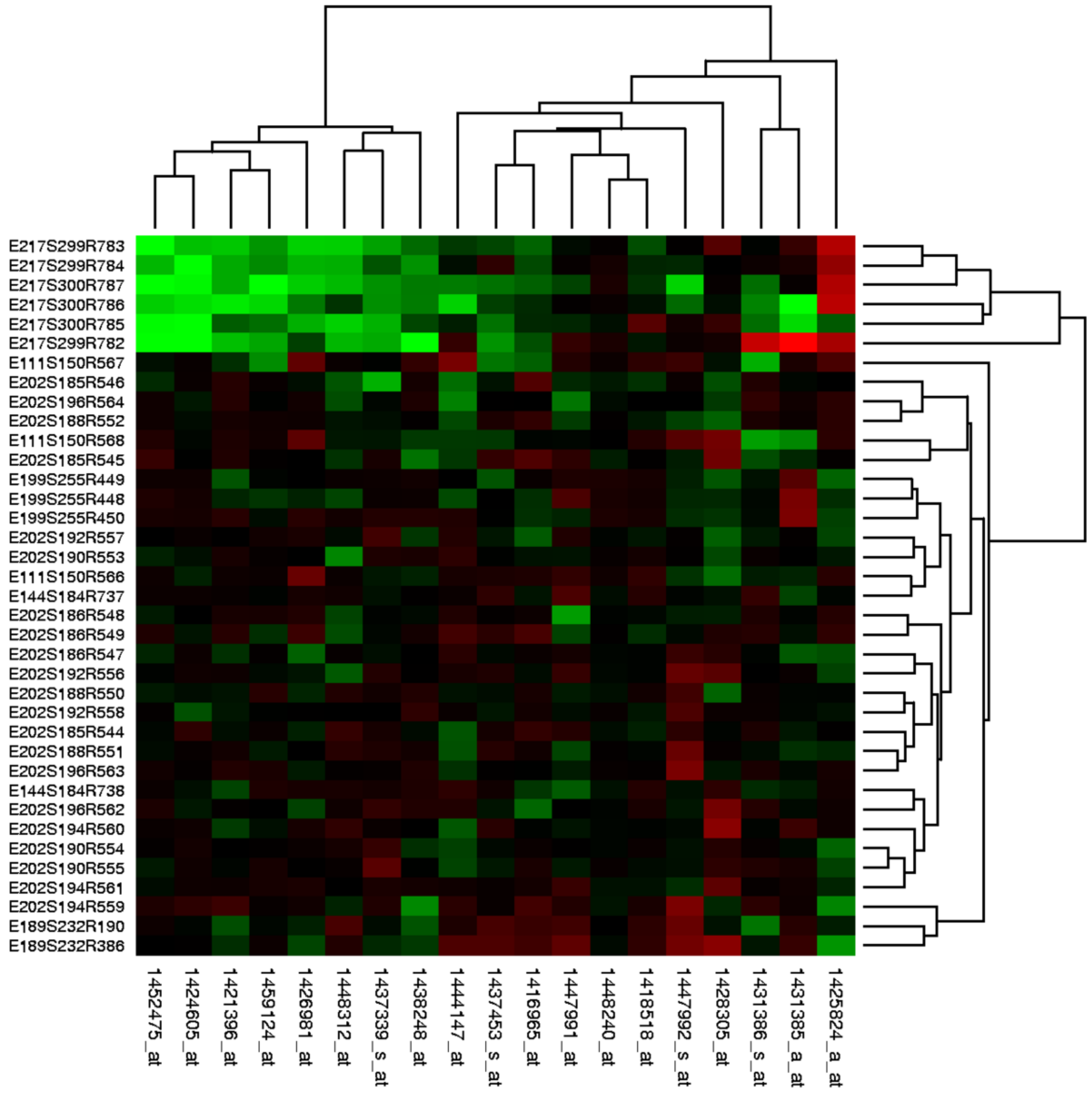
Each pixel encodes a value as a color

Pros

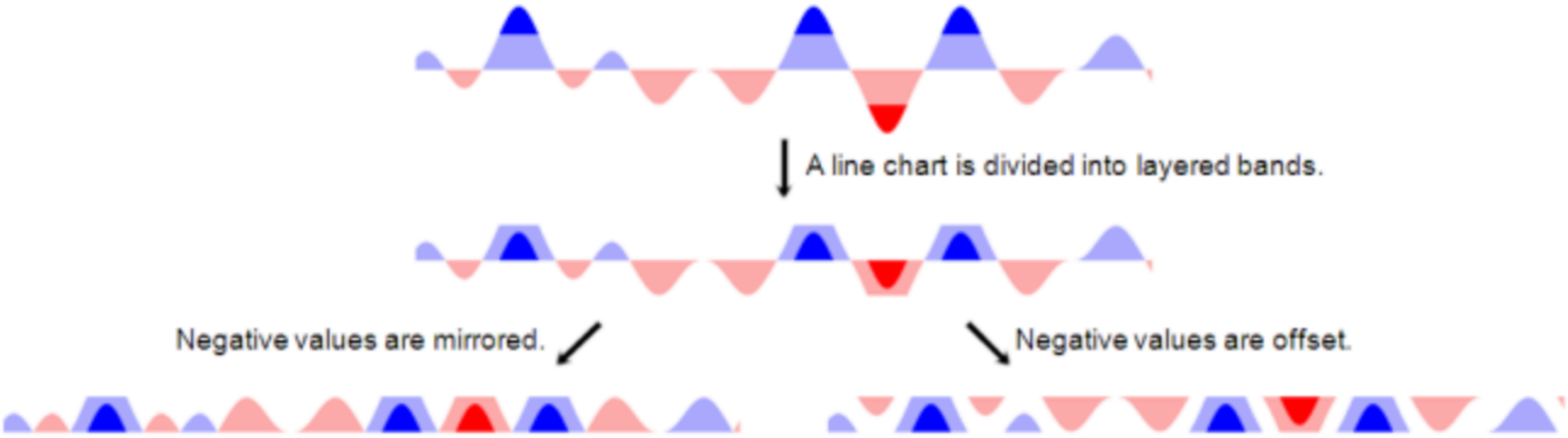
- Scalability
- Good for same type of data

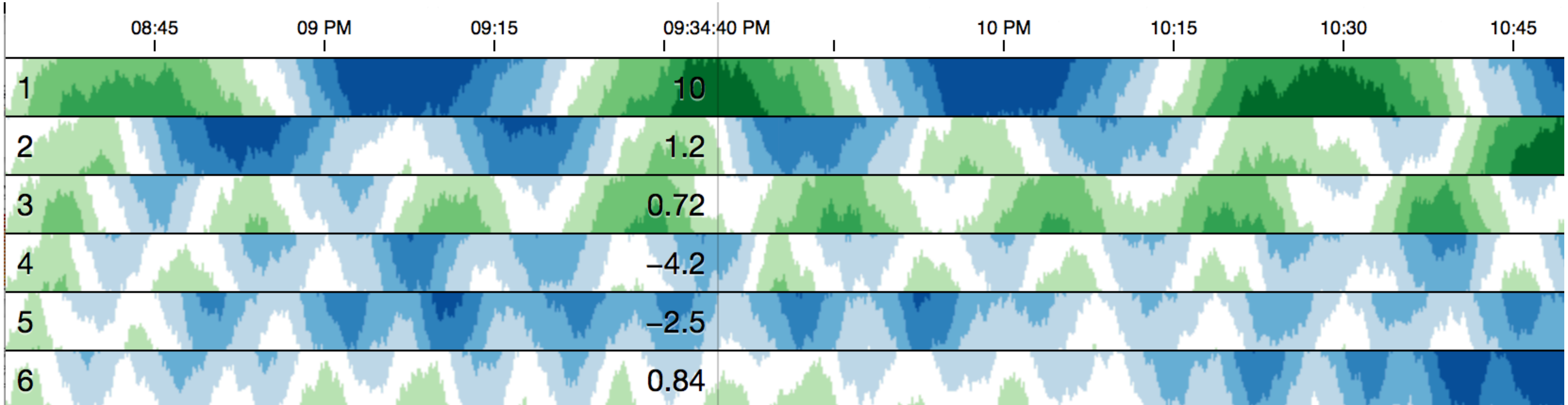
Cons

- ordering is critical
- a good color map is important

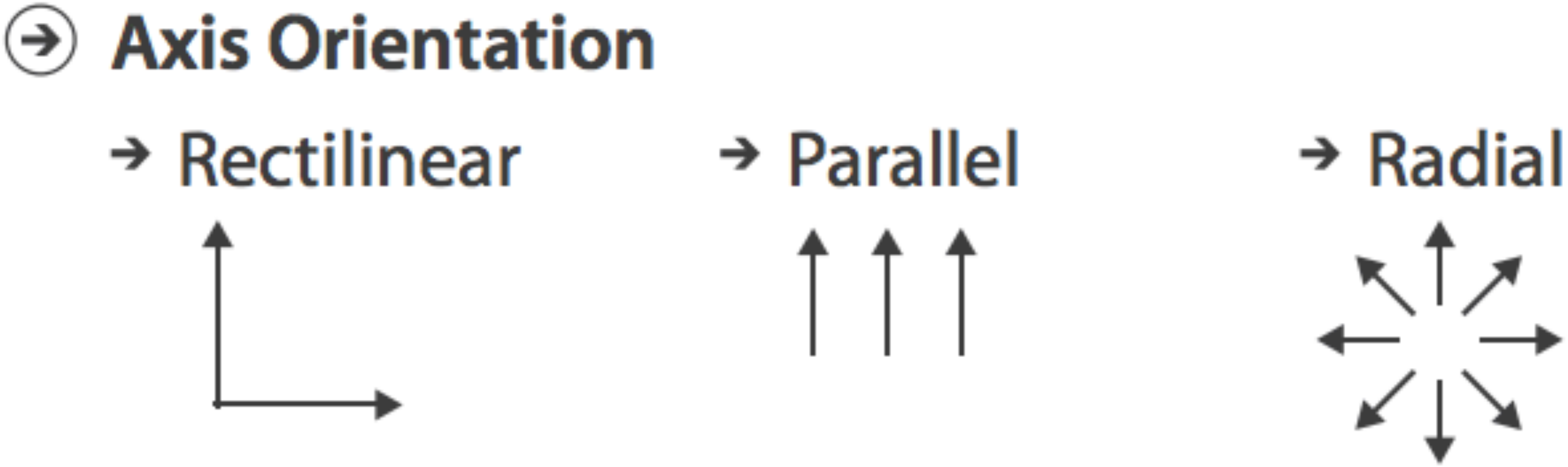


Multiple Horizon Charts





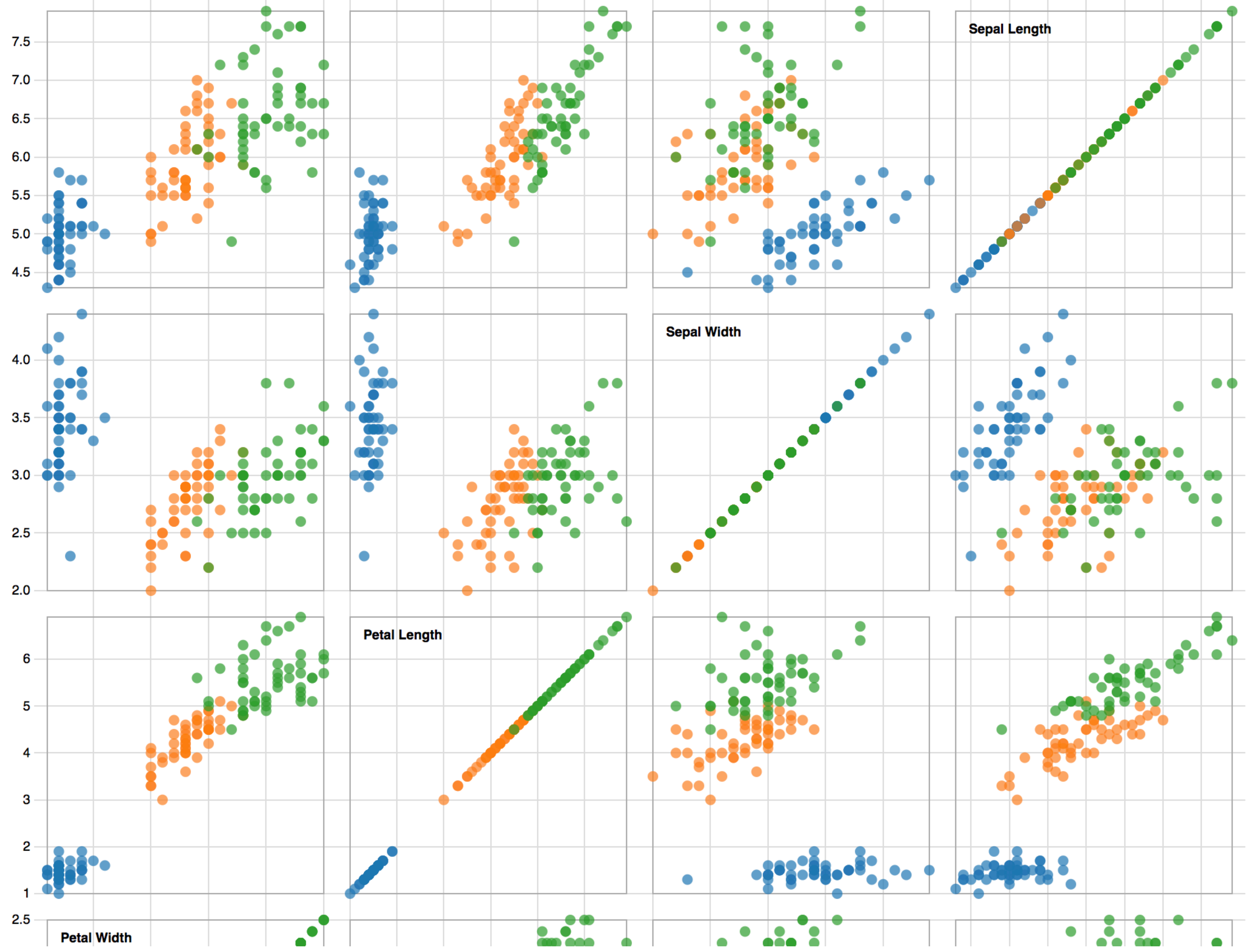
Spatial axis orientation



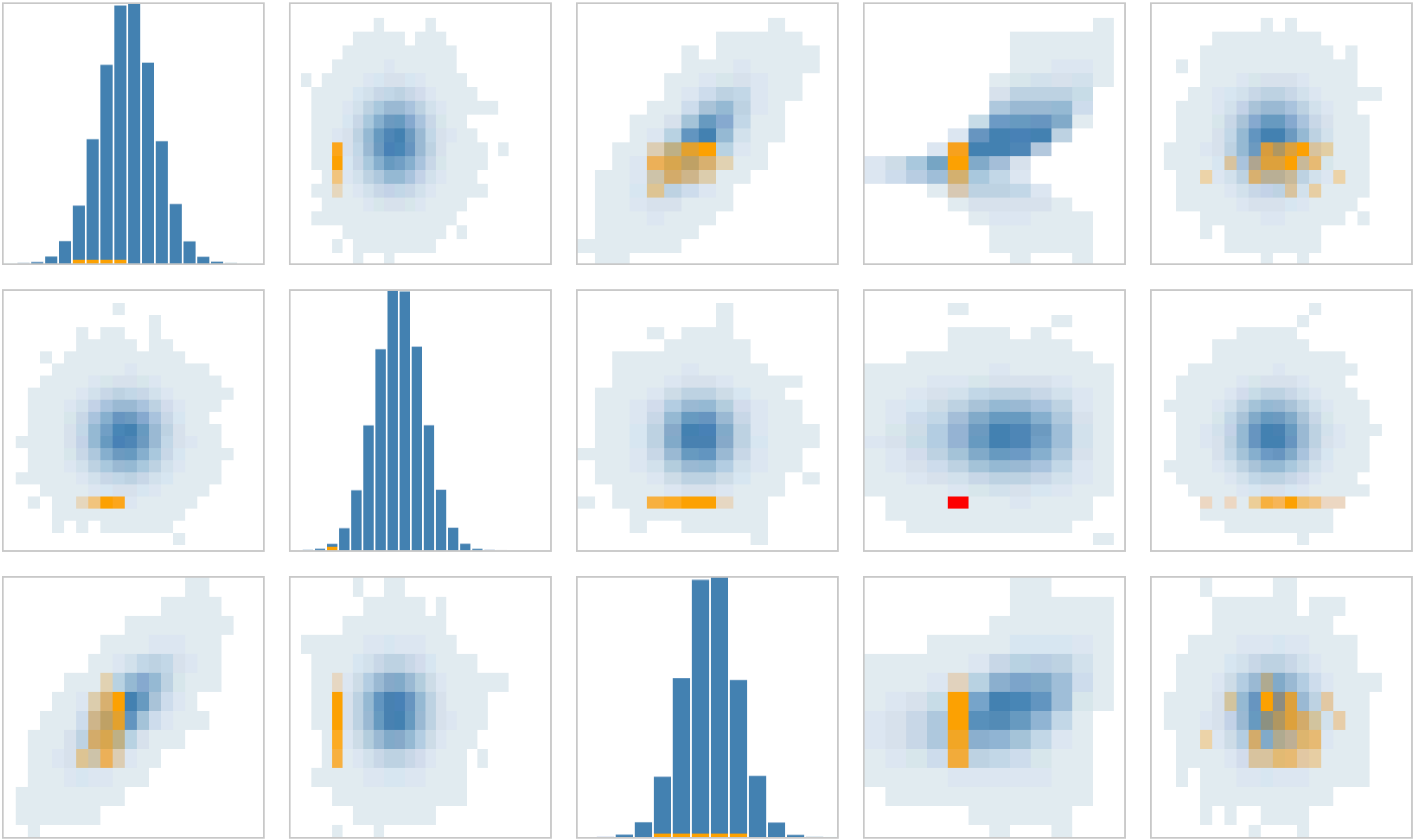
Scatterplot matrix (SPLOM)

Shows all possible pairwise combinations of attributes

Original attributes are used as rows and columns.



Interactive Binned Scatterplot Matrix Dimensions: 5 Bins: 20 Data Points: 100k



SPLOM

To be effective SPLOM needs to:

have not too many dimensions ~ 20 dim

have brushing capabilities

use a F+C technique

If the dataset is too complex:

Use clustering and aggregation

Choose interesting dimension and order (via exploratory analysis)

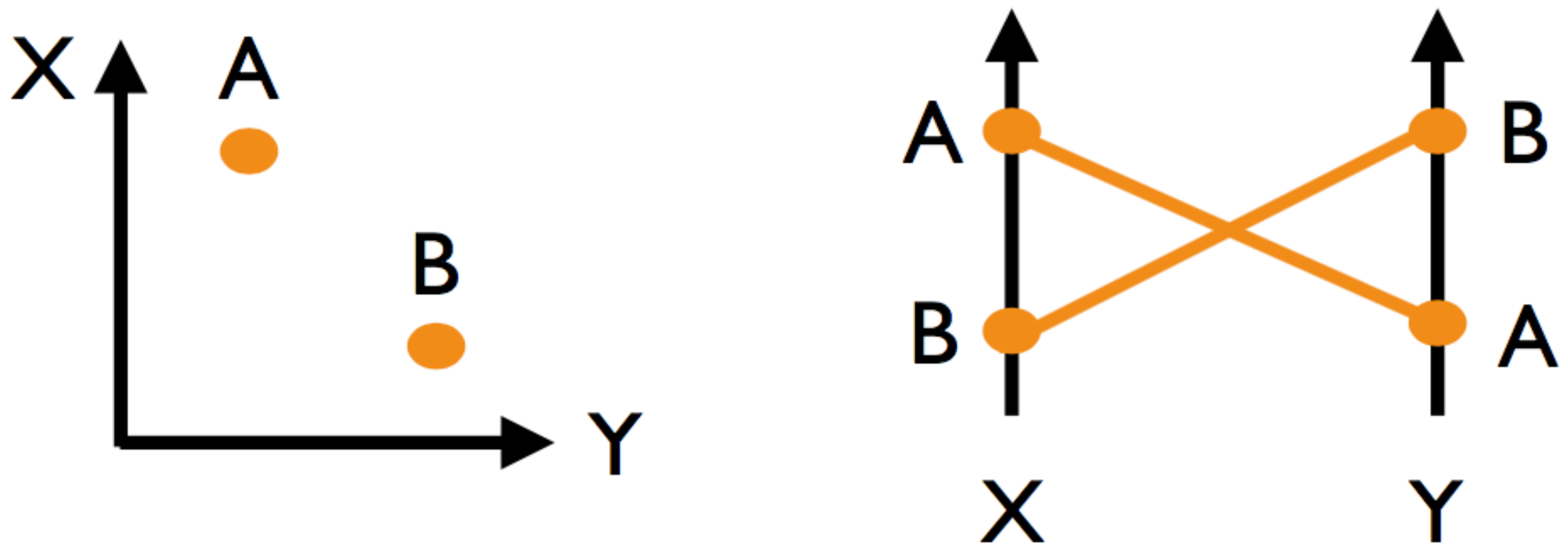
Parallel Coordinates (PC)

Each axe represents an attribute

Values are connected if them belong to the same record

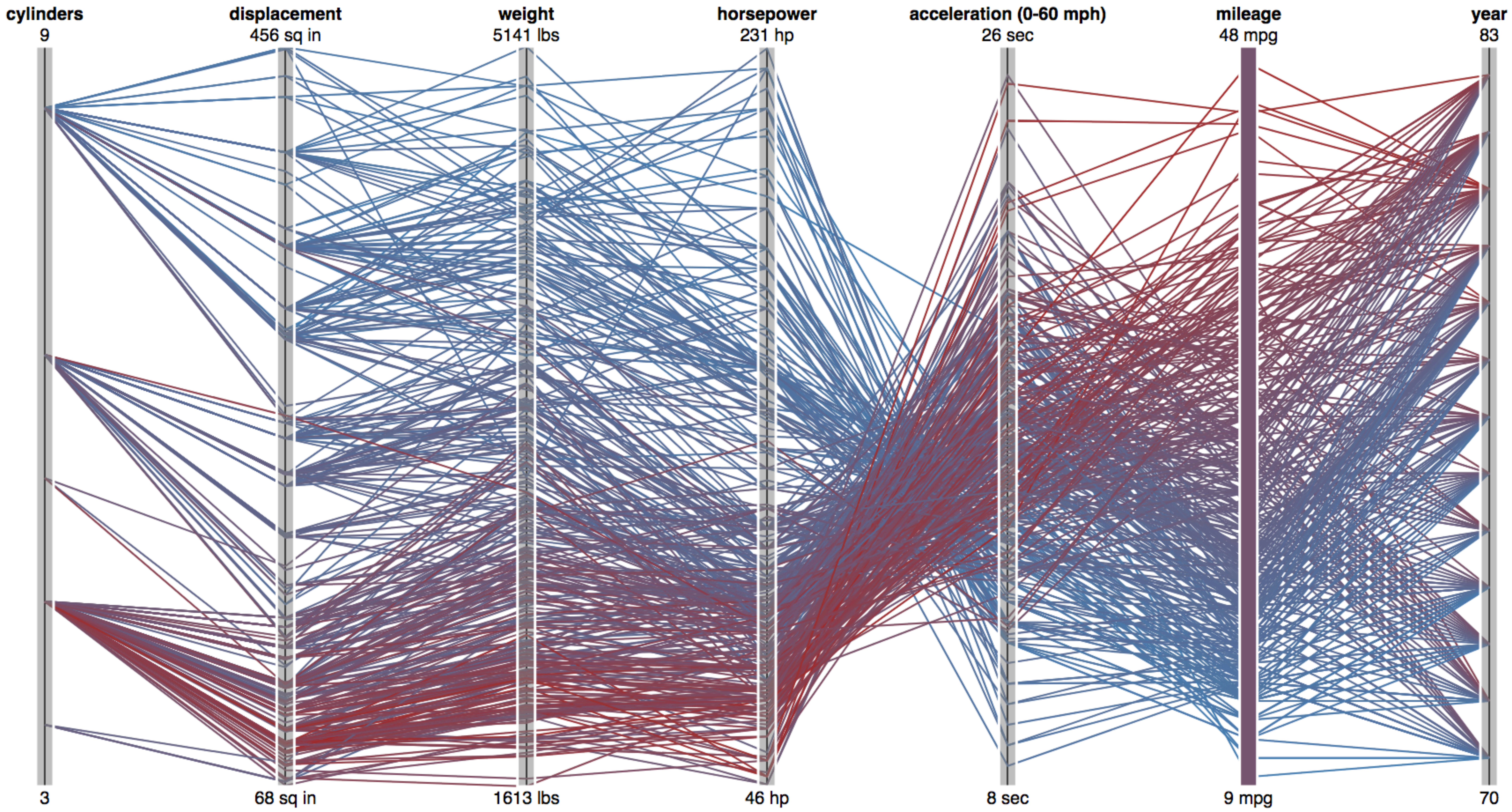
Works with heterogeneous data

Correlations only between adjacent axes

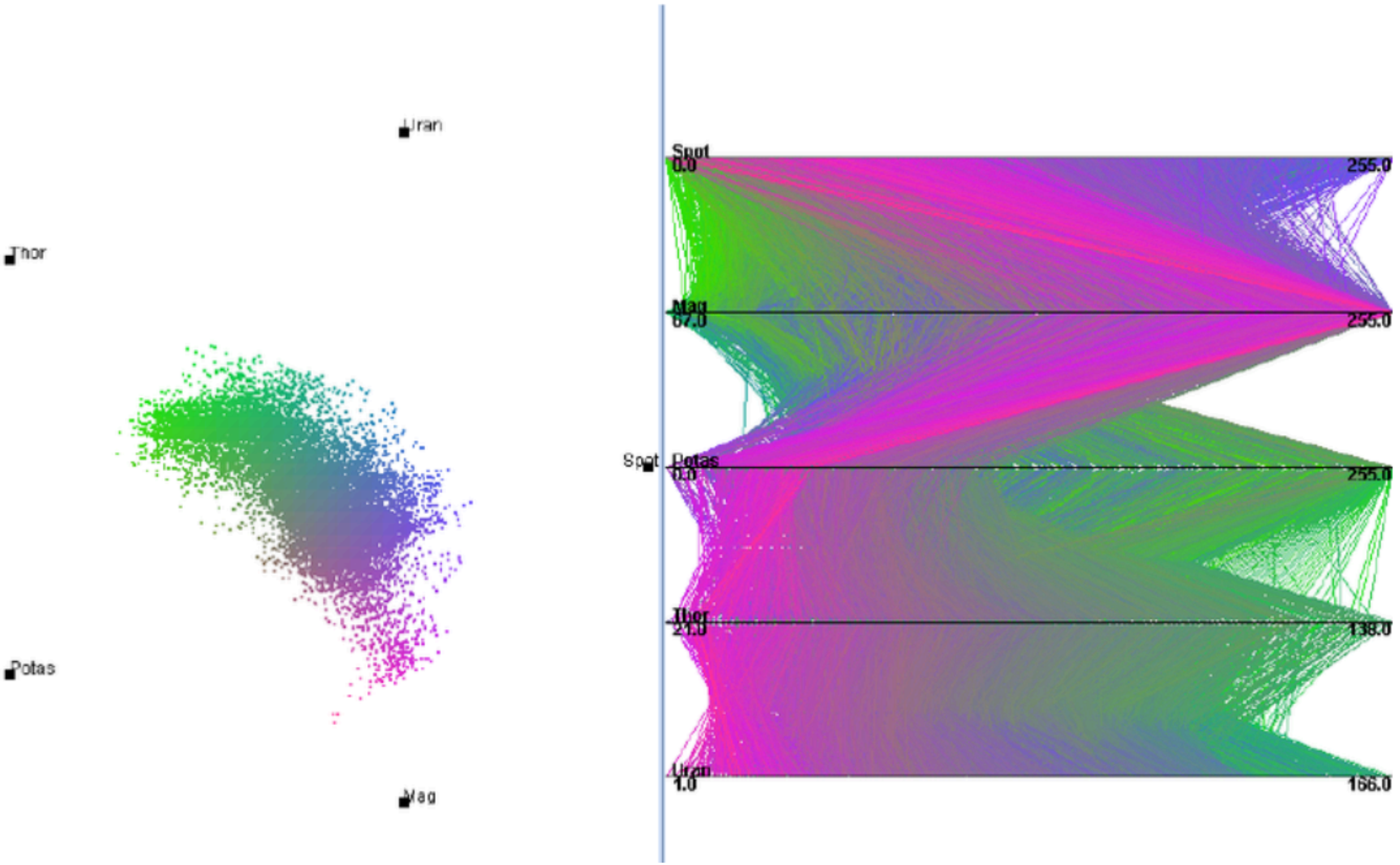


[Lex]

Parallel Coordinates



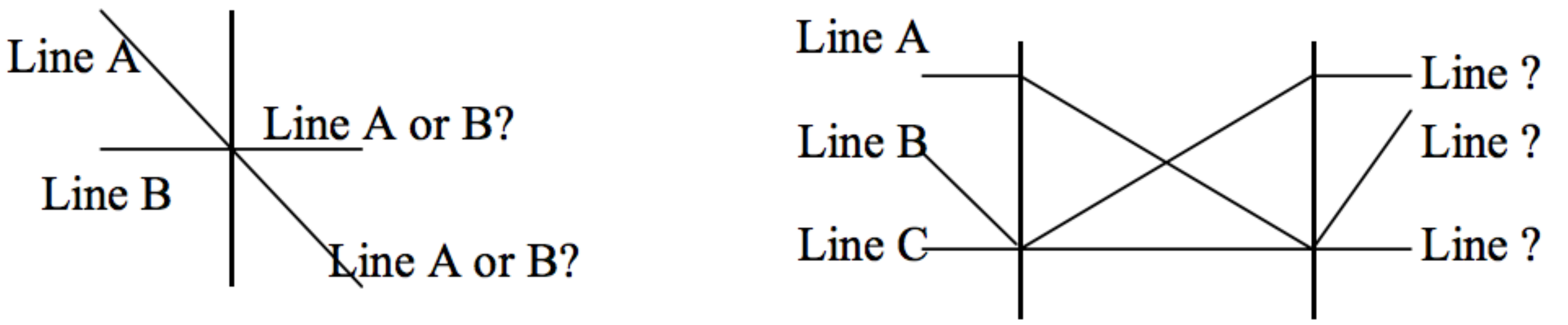
Issues?



Uran
Thor
Potas
Mag
Spot
Spot
0.0
255.0
Mag
67.0
255.0
Potas
0.0
255.0
Thor
21.0
138.0
Uran
1.0
166.0

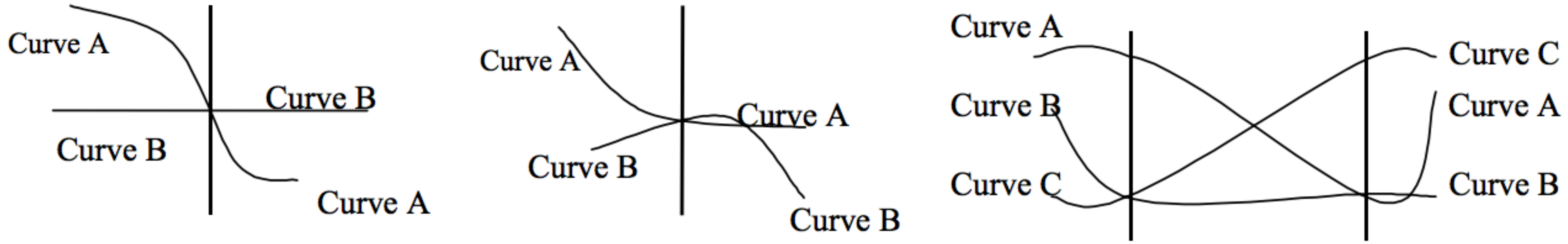
PC scalability

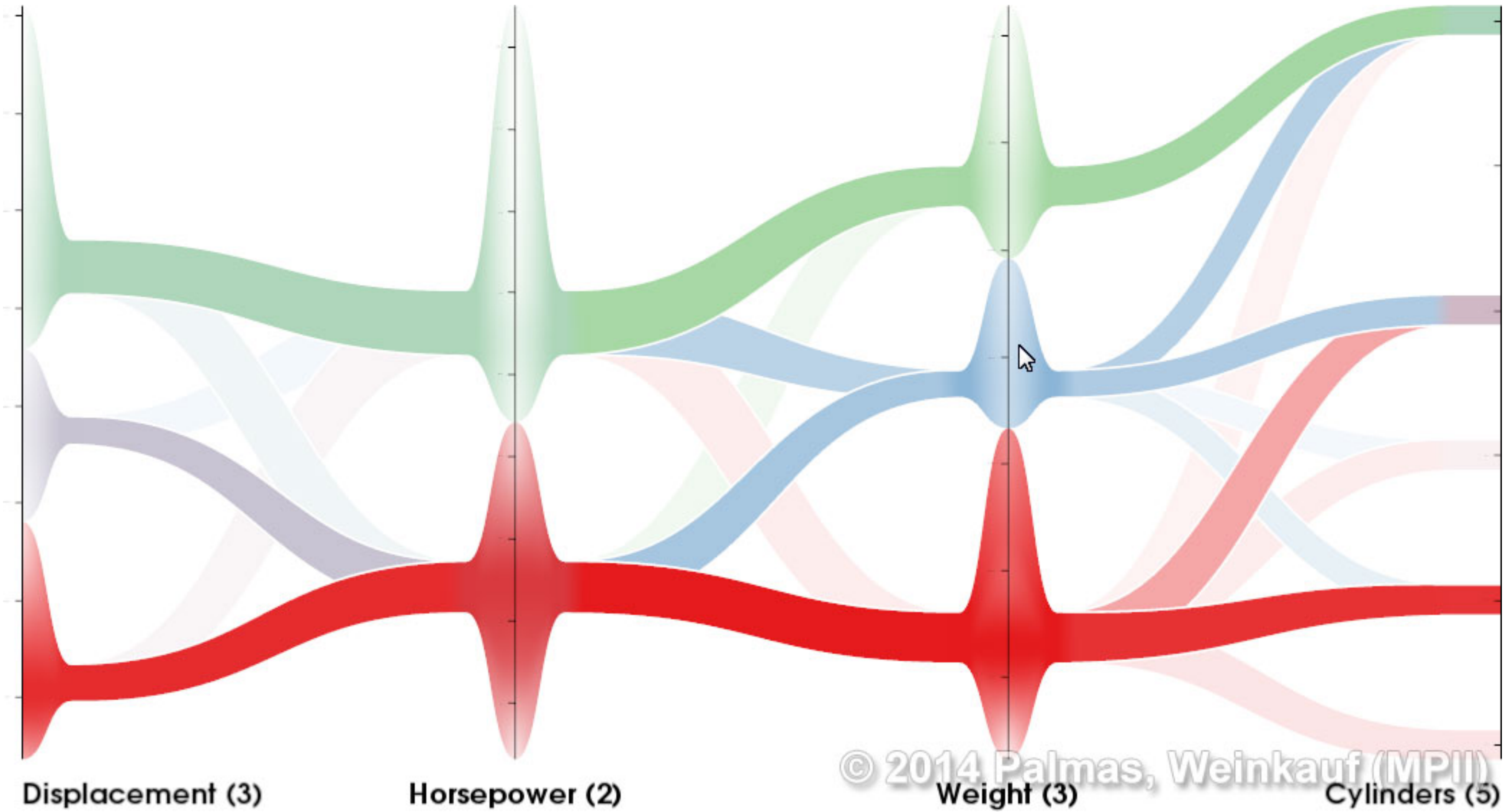
[Graham & Kennedy]



To scale PC we can use:

- Transparency
- Bundling
- Clustering
- Sampling
- Interaction: brusing, reorder axis
- Filtering



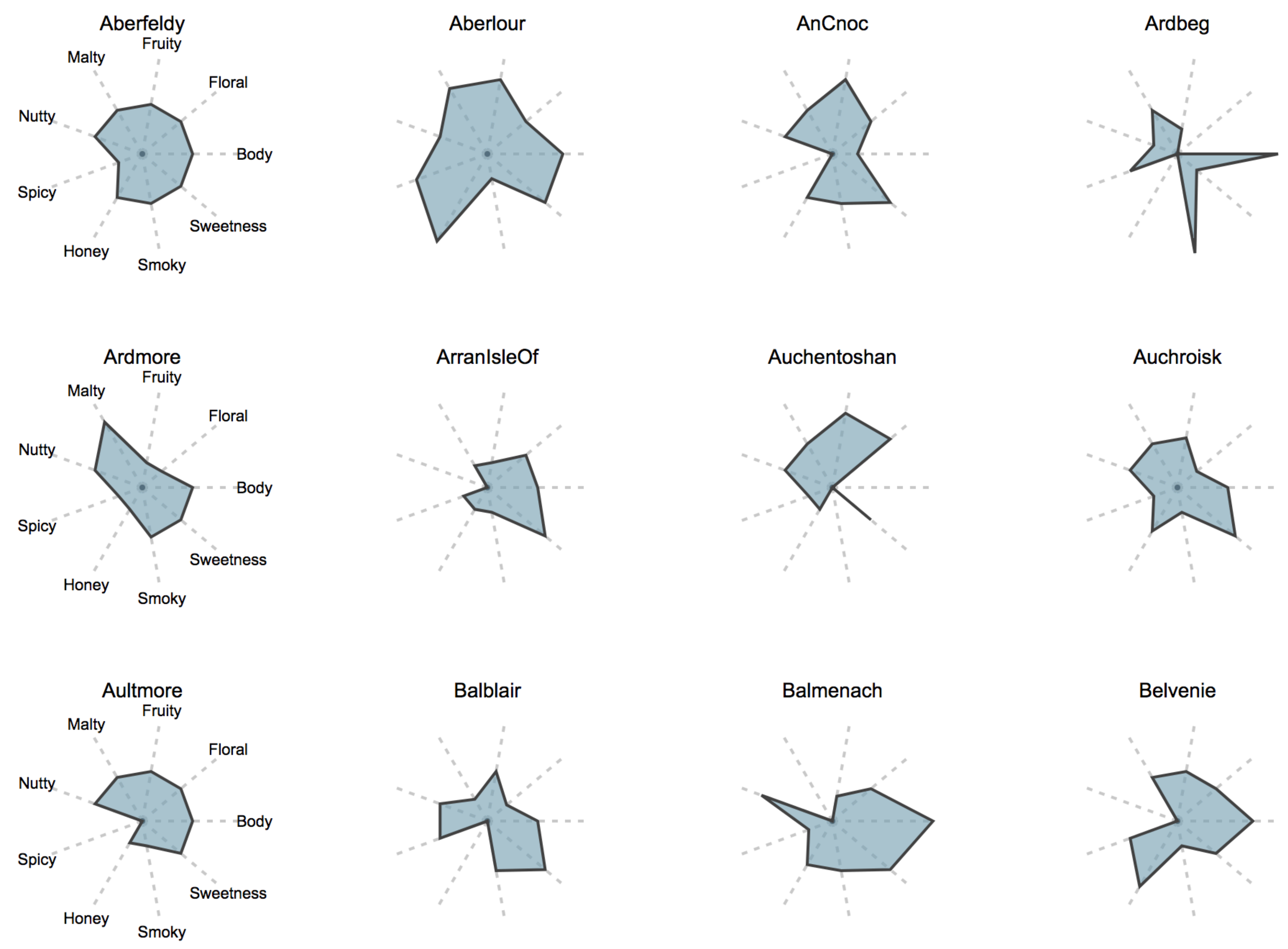


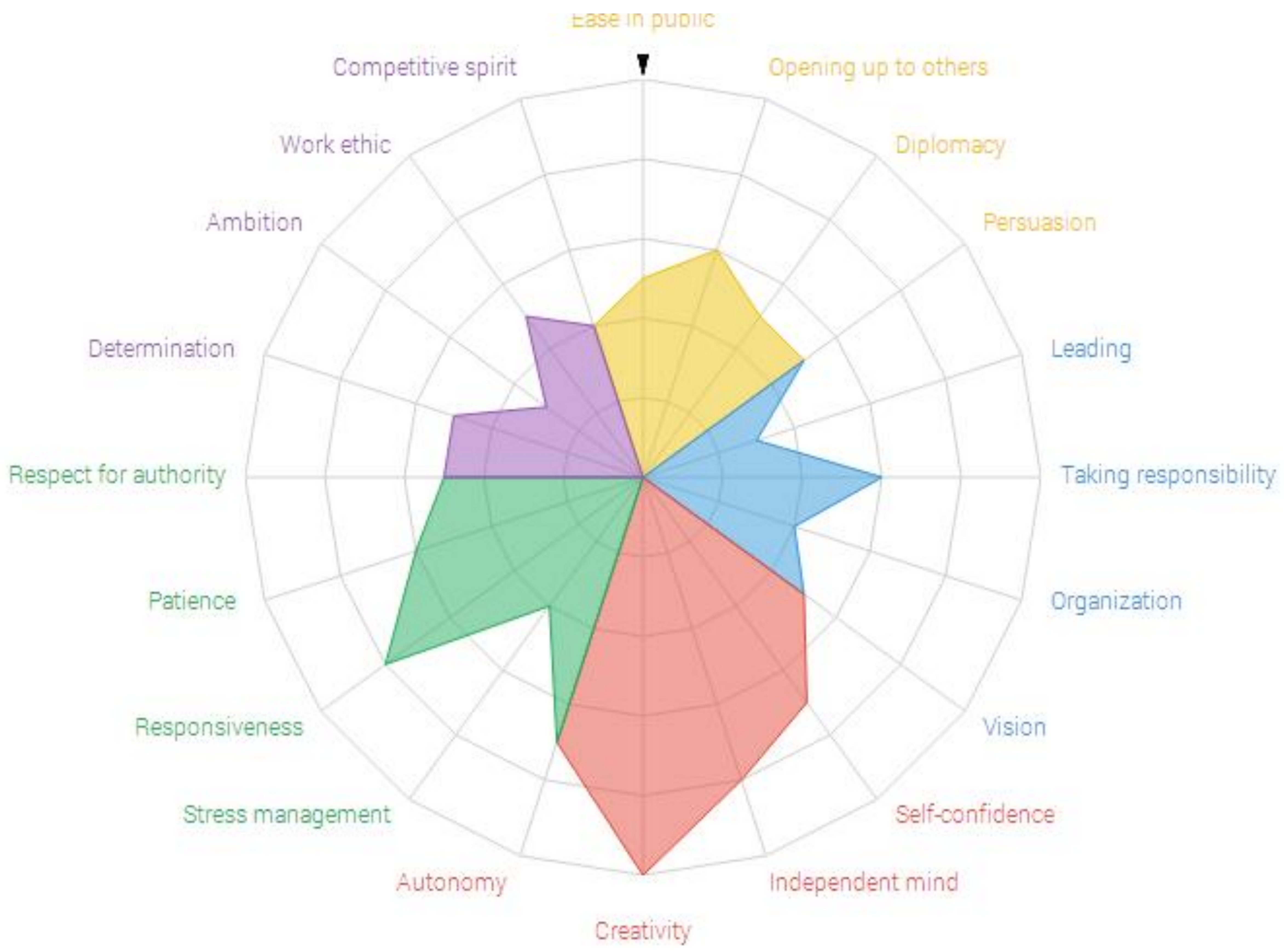

economy (mpg)
power (hp)
cylinders
displacement (cc)
weight (lb)
0-60 mph (s)
year
0
5
10
15
20
25
30
35
40
45
0
20
40
60
80
100
120
140
160
180
200
220
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
100
150
200
250
300
350
400
450
2,000
2,500
3,000
3,500
4,000
4,500
5,000
8
10
12
14
16
18
20
22
24
70
71
72
73
74
75
76
77
78
79
80
81
82

Star plot

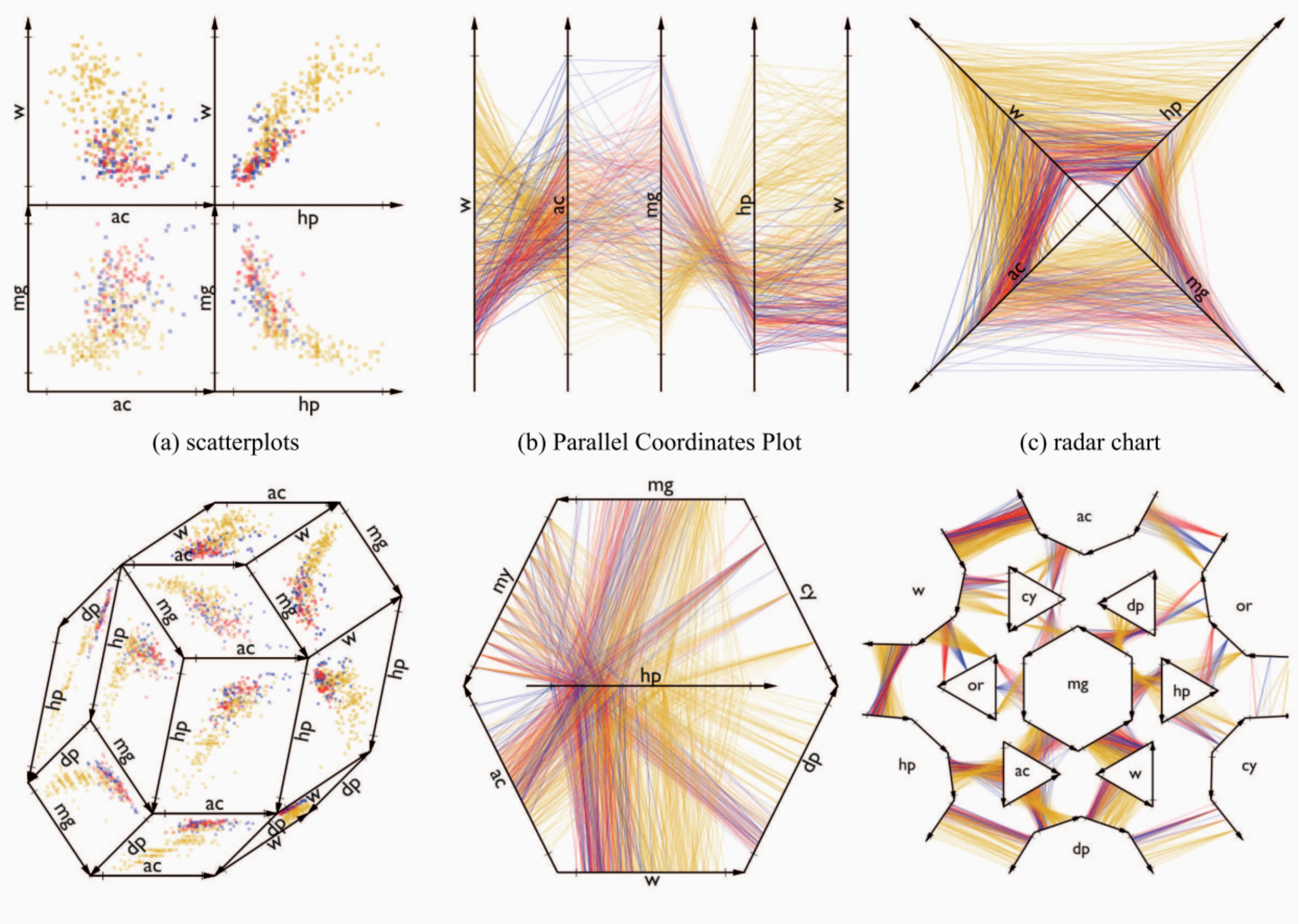
Similar to parallel coordinates

Radial layout from the center





Flexible Linked Axes (FLINA)

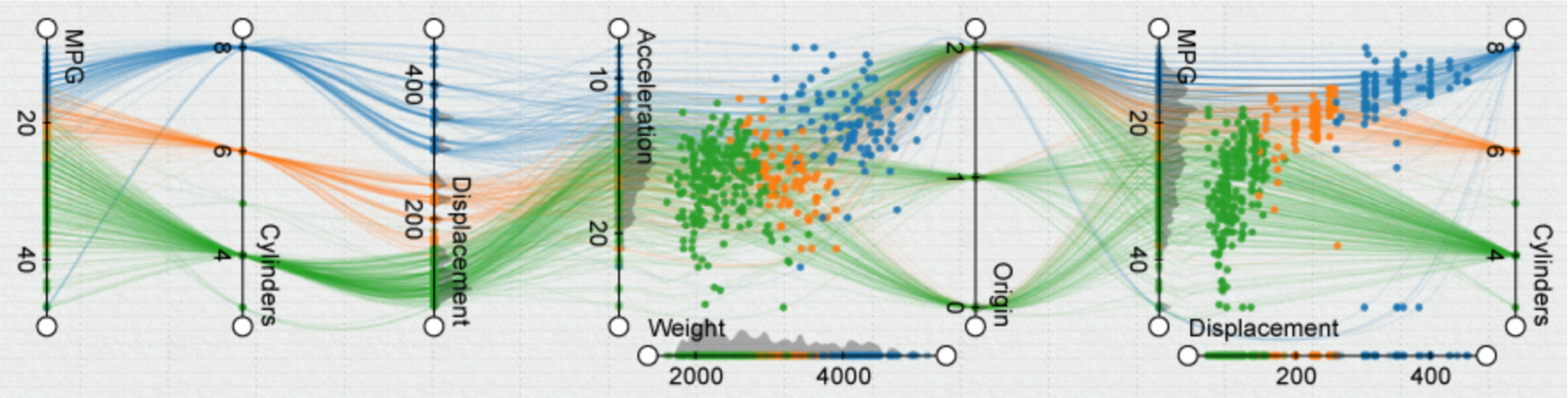


(a) scatterplots (b) Parallel Coordinates Plot (c) radar chart

(d) Hyperbox (e) Time Wheel (f) Many-to-many PCP

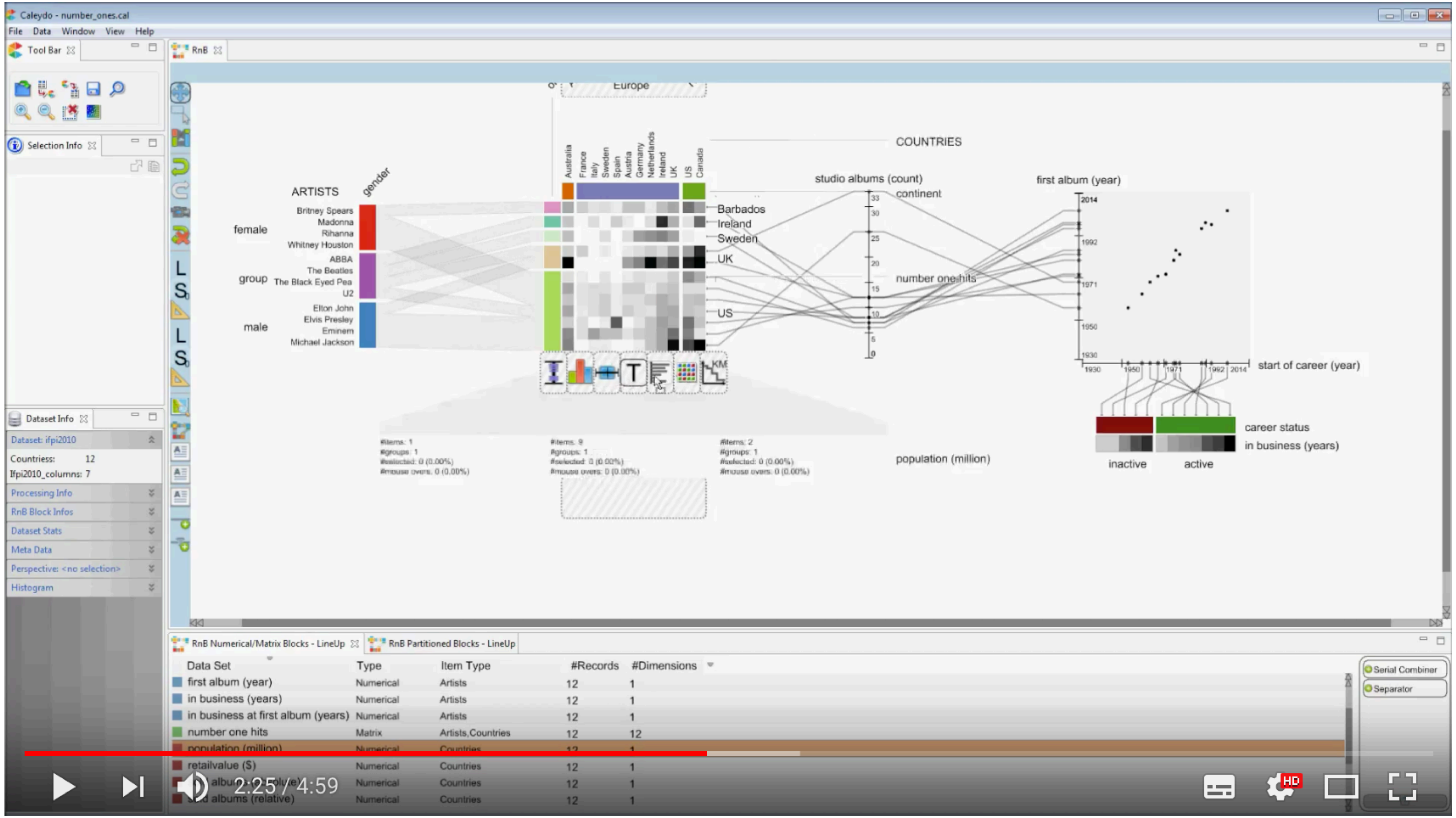
Claessen & van Wijk 2011

Visualization Assembly Line



Web-based implentation of FLINA

Domino


Caleydo - number_ones.cal
File Data Window View Help
Tool Bar
Selection Info
Dataset Info
Dataset: ifpi2010
Countries: 12
Ifpi2010_columns: 7
Processing Info
RnB Block Infos
Dataset Stats
Meta Data
Perspective: <no selection>
Histogram
RnB
Europe
COUNTRIES
ARTISTS
gender
female
group
male
Britney Spears
Madonna
Rihanna
Whitney Houston
ABBA
The Beatles
The Black Eyed Pea
U2
Elton John
Elvis Presley
Eminem
Michael Jackson
Australia
France
Italy
Sweden
Spain
Austria
Germany
Netherlands
Ireland
UK
US
Canada
Barbados
Ireland
Sweden
UK
US
studio albums (count)
continent
number one hits
first album (year)
start of career (year)
career status
in business (years)
inactive
active
population (million)
RnB Numerical/Matrix Blocks - LineUp
RnB Partitioned Blocks - LineUp
Data Set
Type
Item Type
#Records
#Dimensions
first album (year)
in business (years)
in business at first album (years)
number one hits
retailvalue ($)
Numerical
Matrix
Artists
Artists,Countries
Countries
Serial Combiner
Separator
2:25 / 4:59

Parallel set