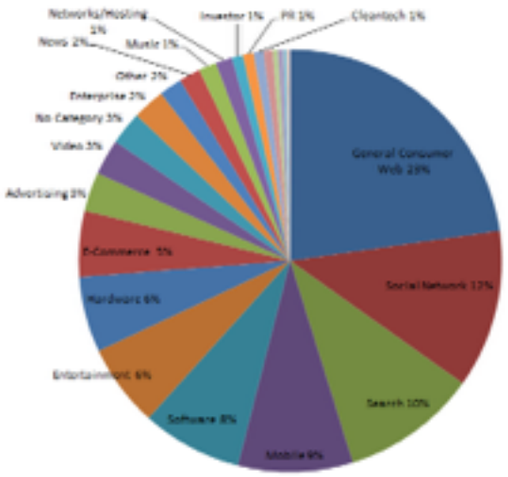
CS 109: Data Science

Visualization of Multi-Dim Data, Maps and Text

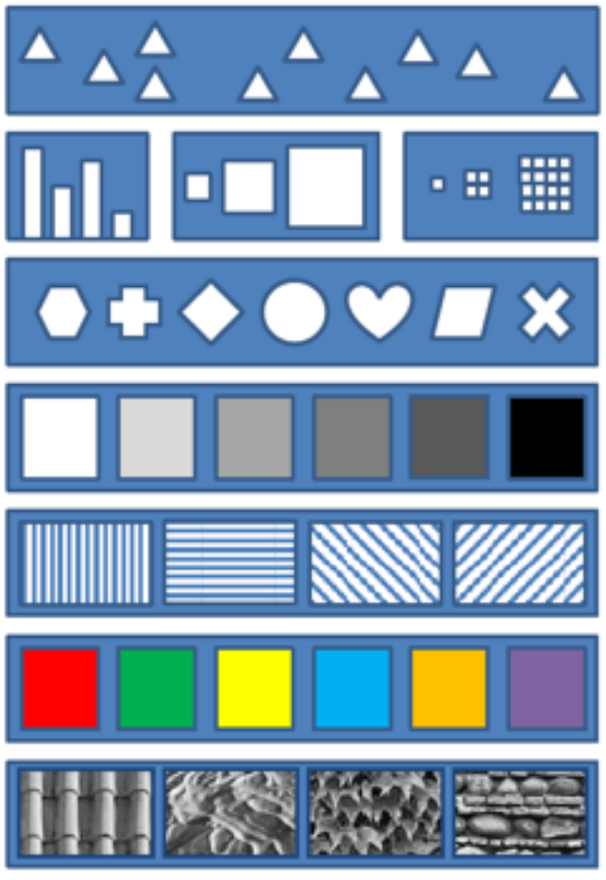
Marc Streit

mstreit@seas.harvard.edu

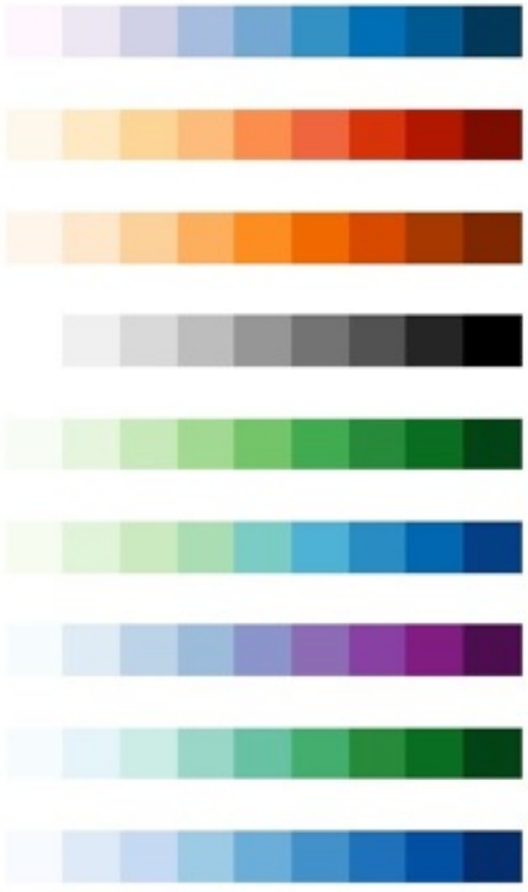
Last Week



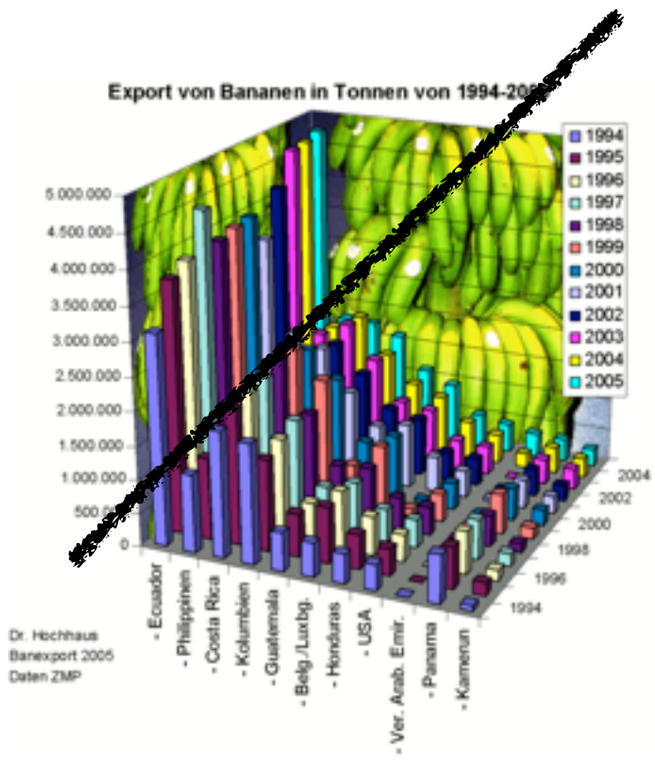
Effective vs. Non Effective Visualizations



Visual Attributes

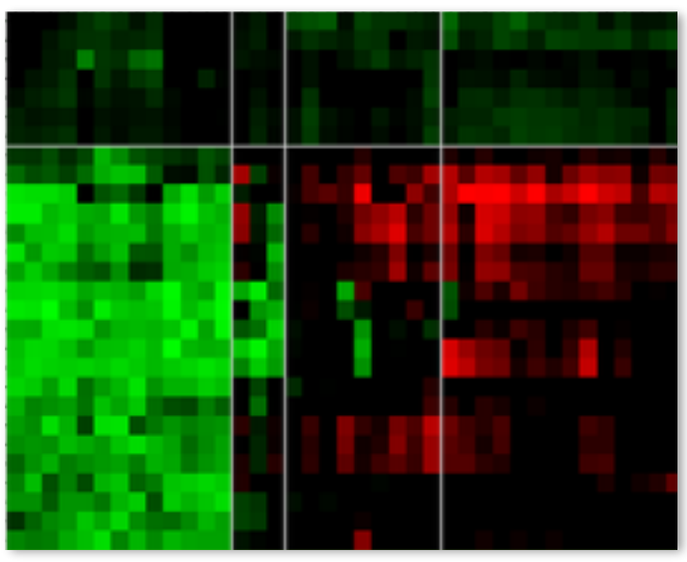


Color

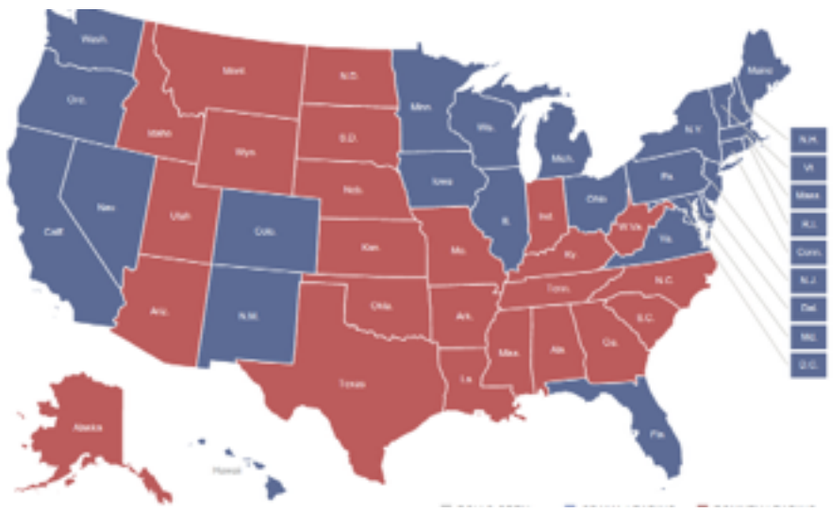


Design Principles

Today



Multi-Dimensional
Data Visualization



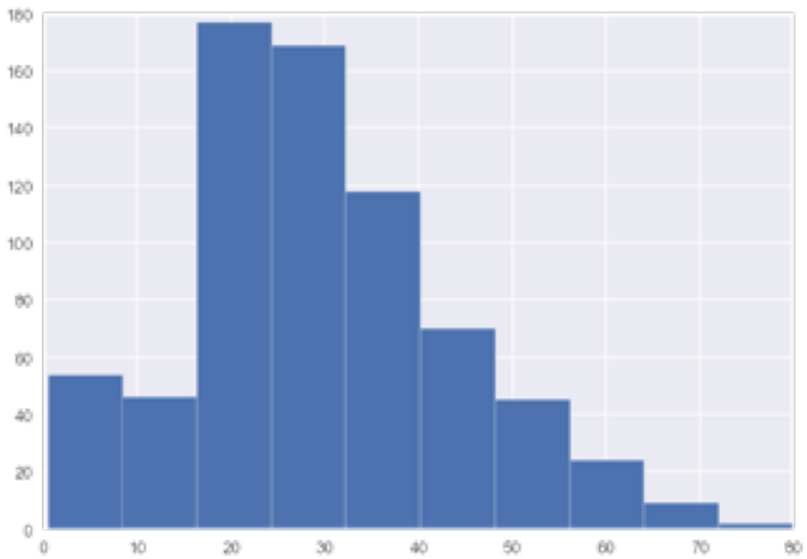
Map & Text
Visualization

Multi-Dimensional Data Visualization

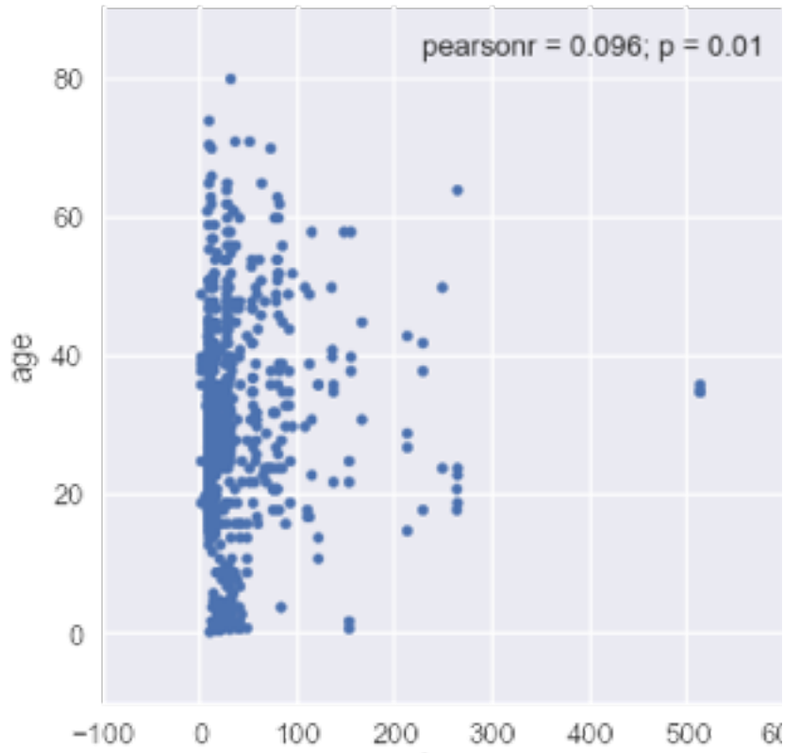
survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0	30.0708	C	Second	child	False		Cherbourg	yes	False
1	3	female	4.0	1	1	16.7	S	Third	child	False	G	Southampton	yes	False
1	1	female	58.0	0	0	26.55	S	First	woman	False	C	Southampton	yes	True
0	3	male	20.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male	39.0	1	5	31.275	S	Third	man	True		Southampton	no	False
0	3	female	14.0	0	0	7.8542	S	Third	child	False		Southampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

Example: Titanic Dataset

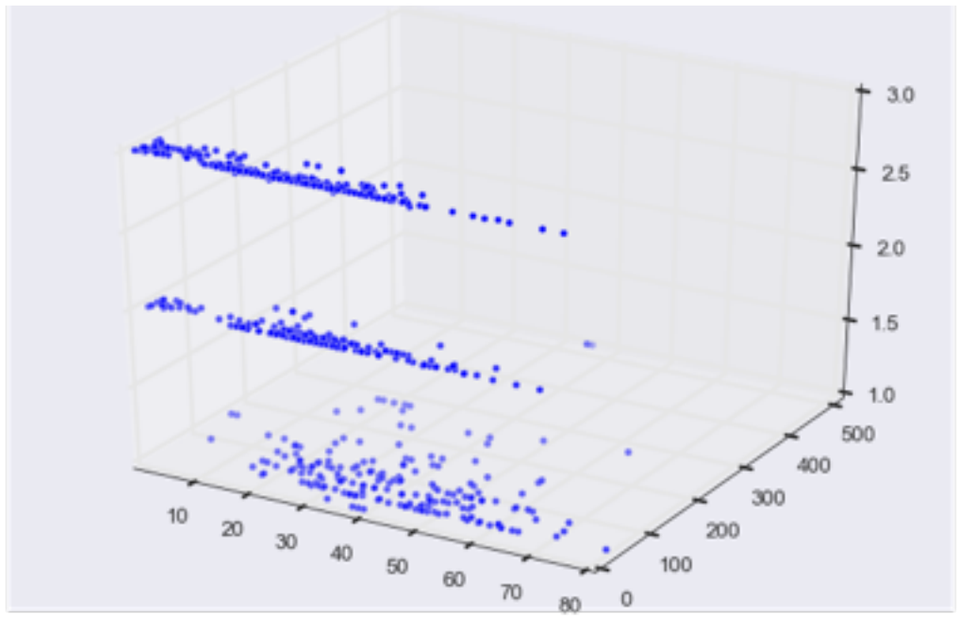
1D


160
140
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80

2D


pearsonr = 0.096; p = 0.01
80
60
40
20
0
age
−100
0
100
200
300
400
500

3D


fare
3.0
2.5
2.0
1.5
1.0
500
400
300
200
100
0
10
20
30
40
50
60
70
80

4D?

What is "high" dimensional?

How many dimensions (attributes)?

~50 – tractable with "just" vis

~1,000 – need analytical methods

How many items?

~ 1,000 – "just" vis is fine

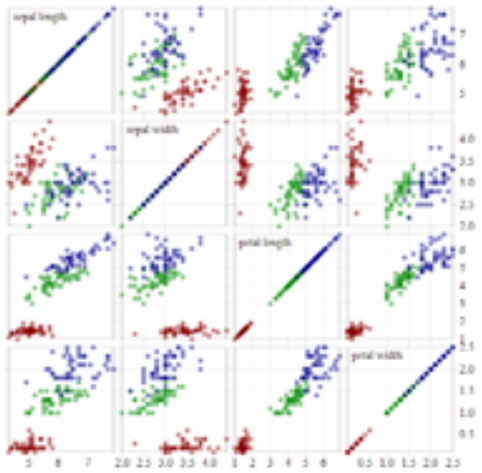
>> 10,000 – need analytical methods

Homogeneity

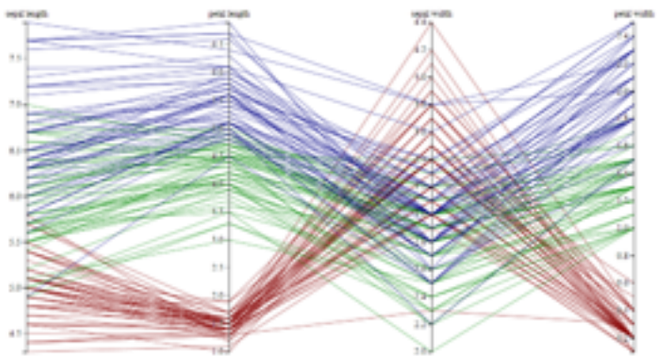
Same data type?

Same scales?

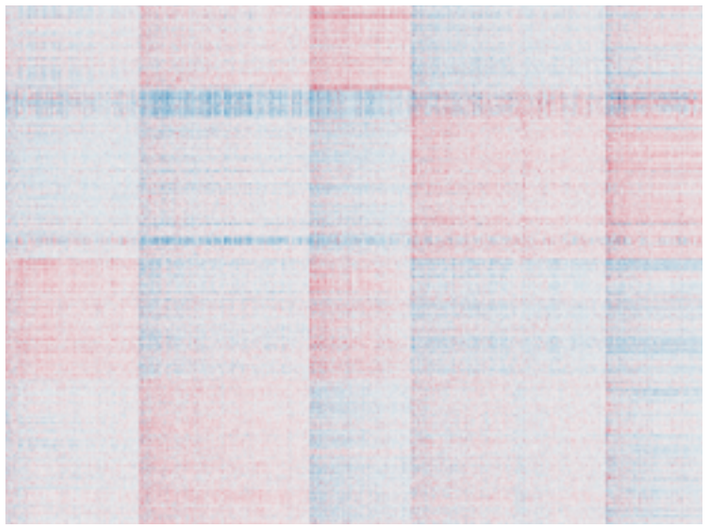
Analytic Component



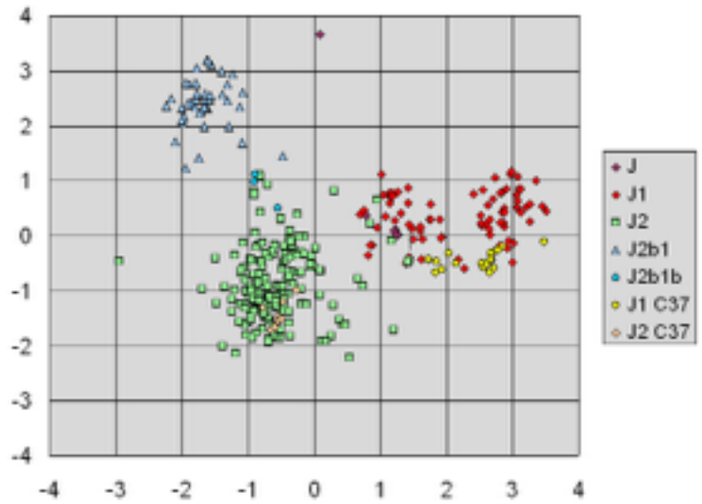
Scatterplot Matrices



Parallel Coordinates



Pixel-based Visualizations / Heat Maps

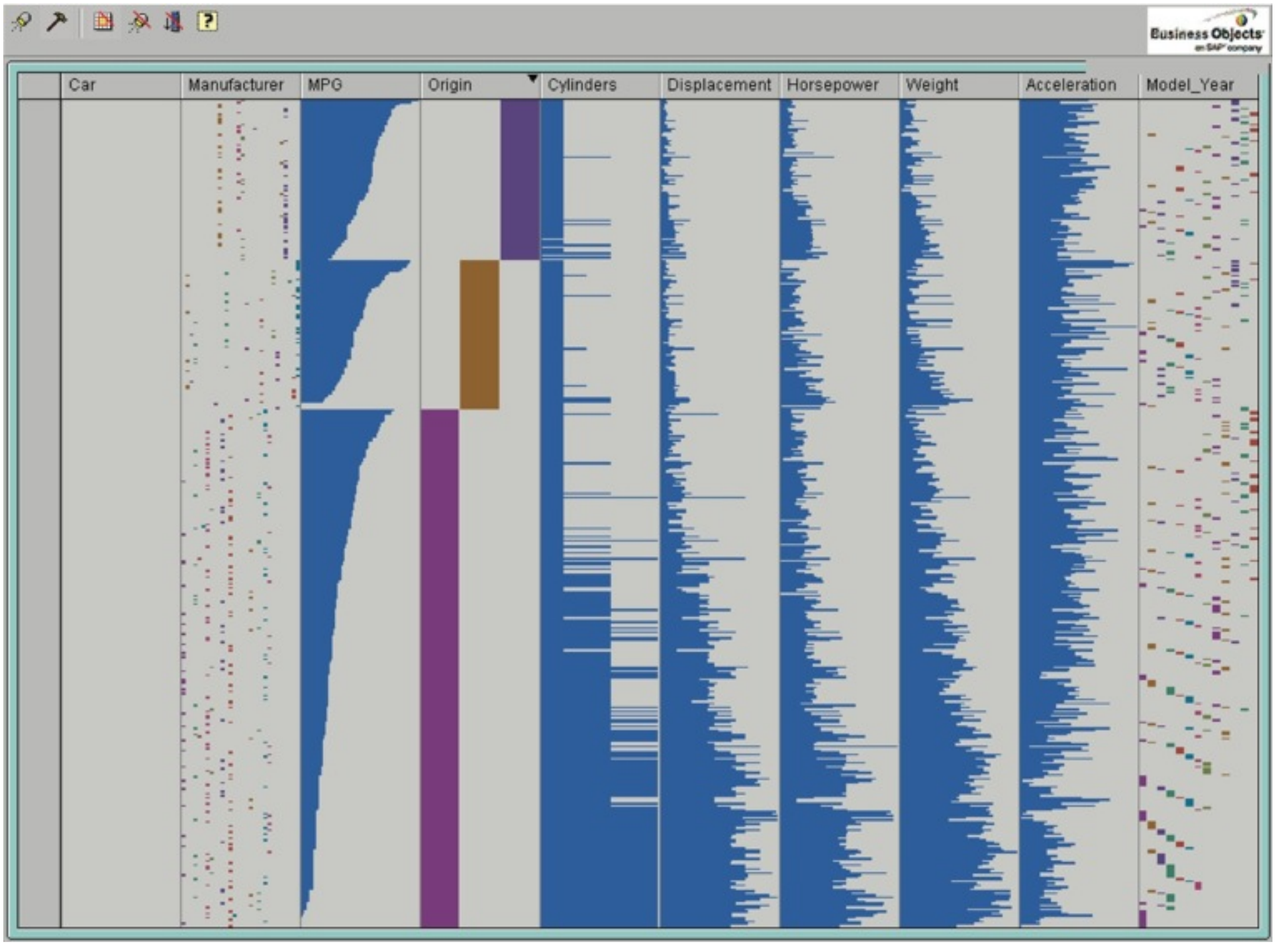


Dimensionality Reduction (e.g., PCA)

no / little analytics

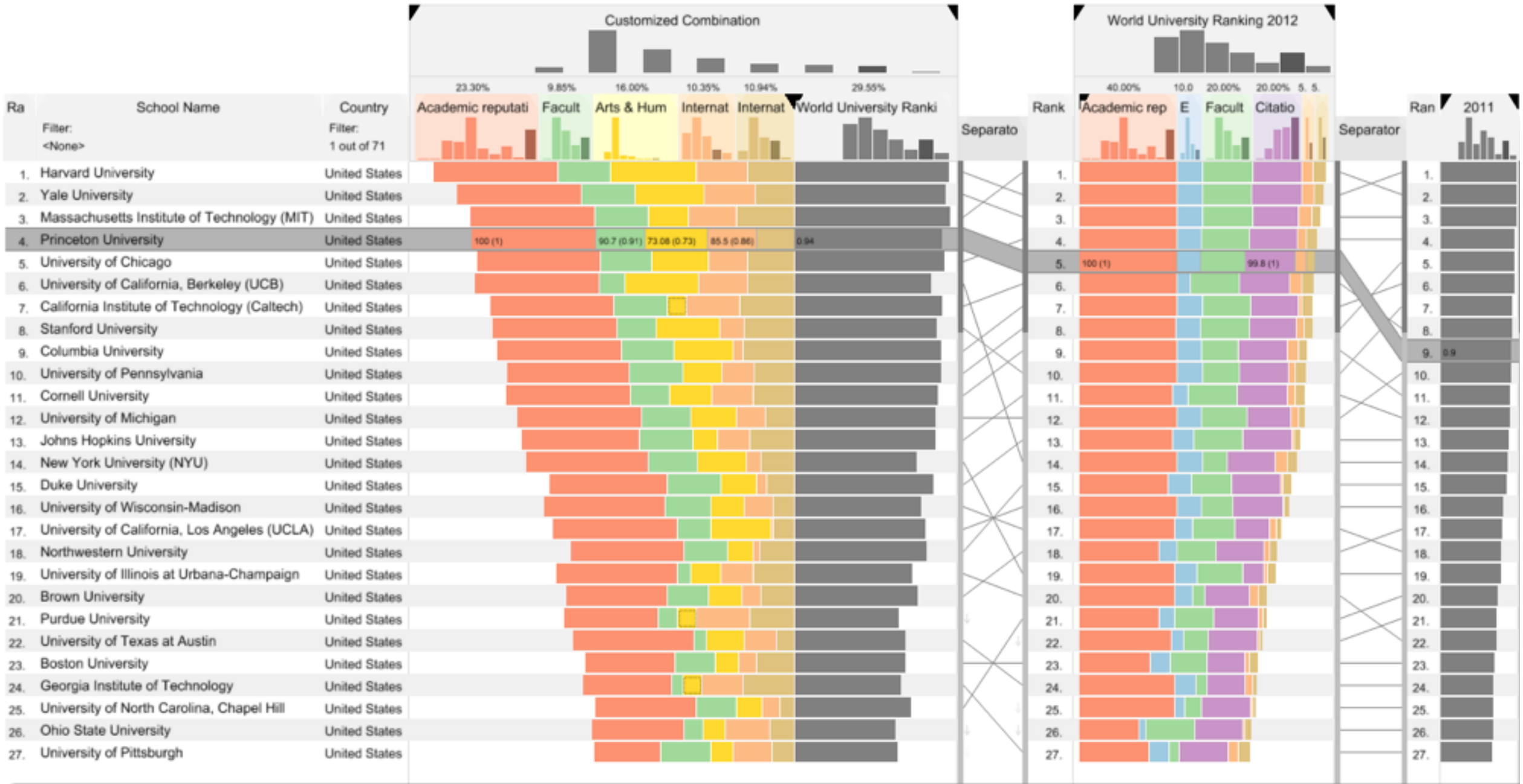
strong analytics component

Table Lens



Inxight 1994

LineUp



Video at http://lineup.caleydo.org

Heat Map

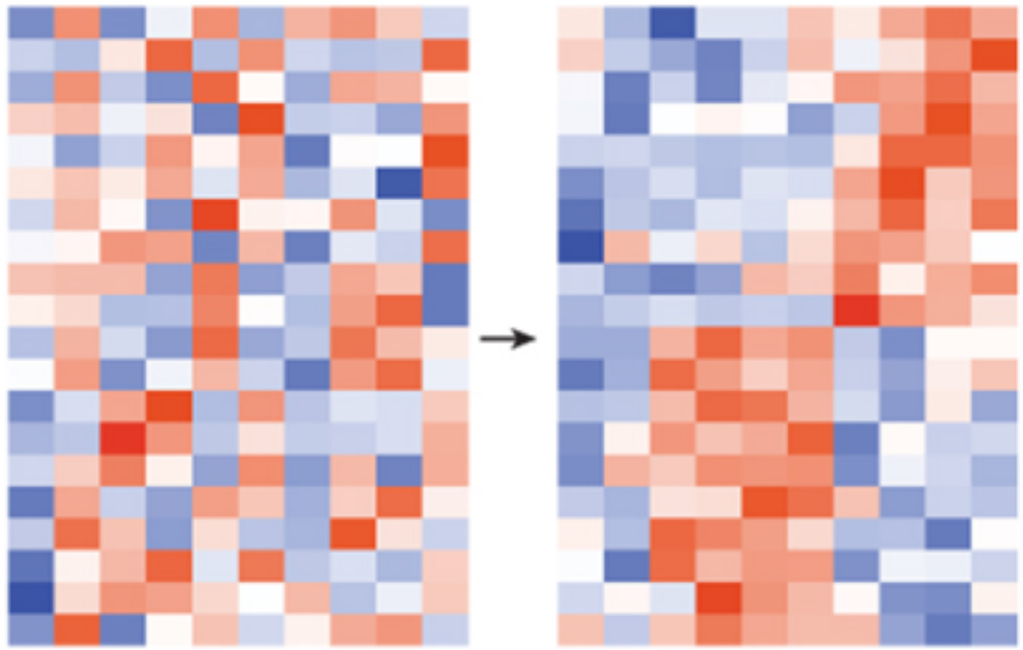
Each cell is a “pixel”, value encoded using color

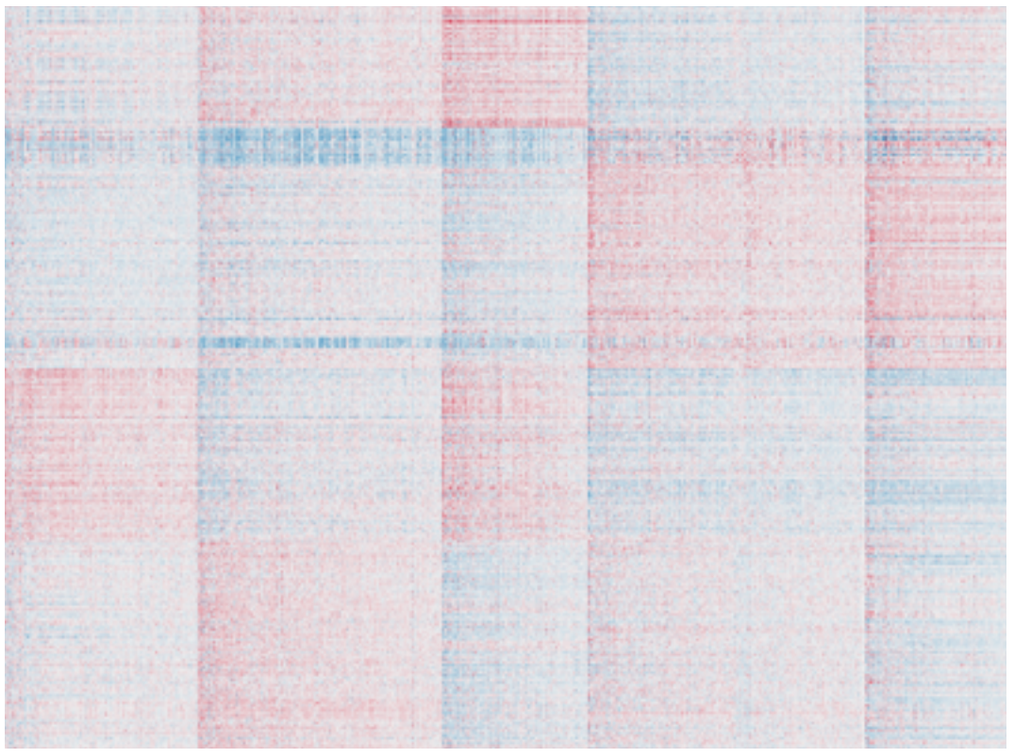
Meaning derived from ordering

If no ordering inherent, clustering is used

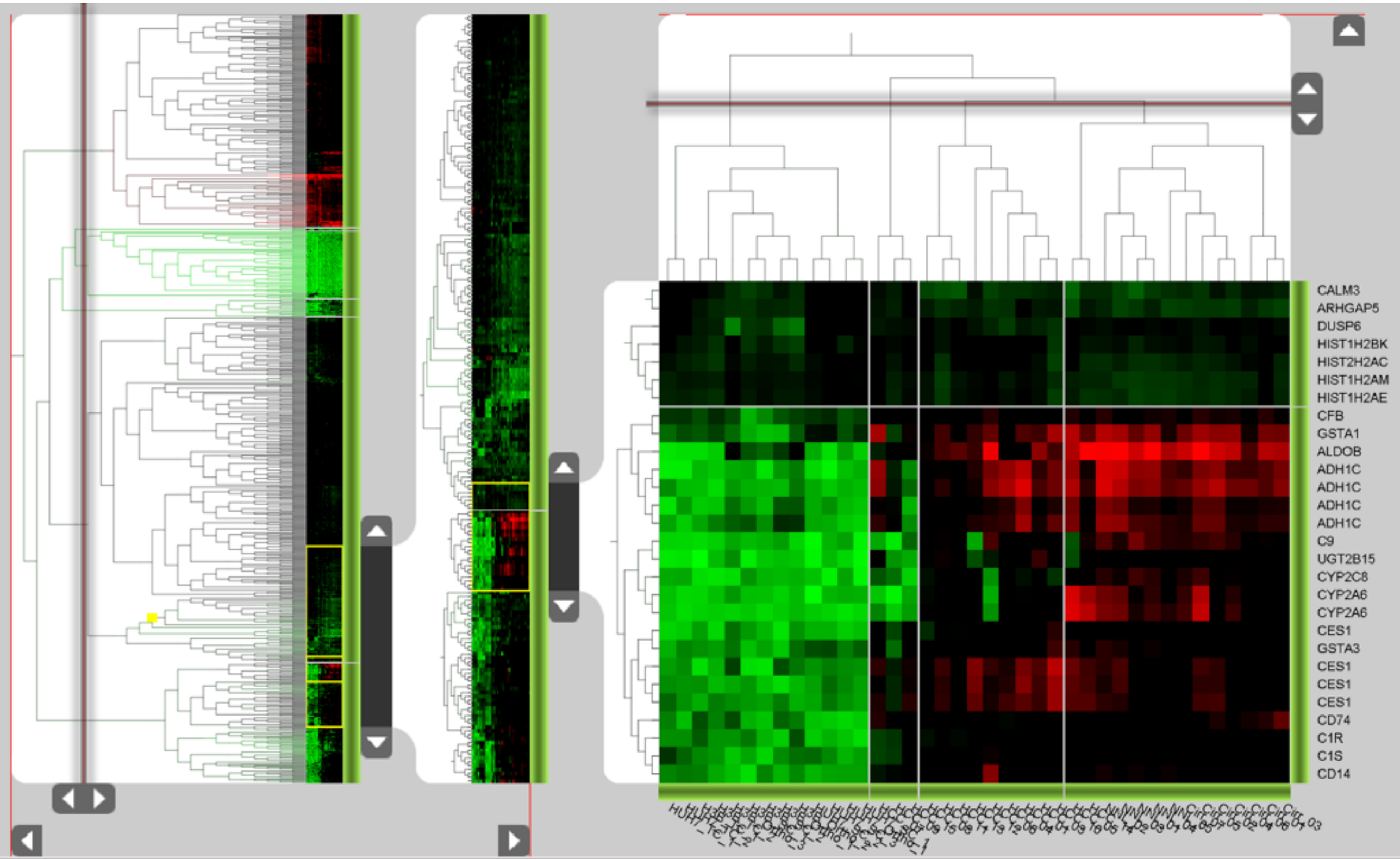
Scalable – 1 px per item

Good for homogeneous data



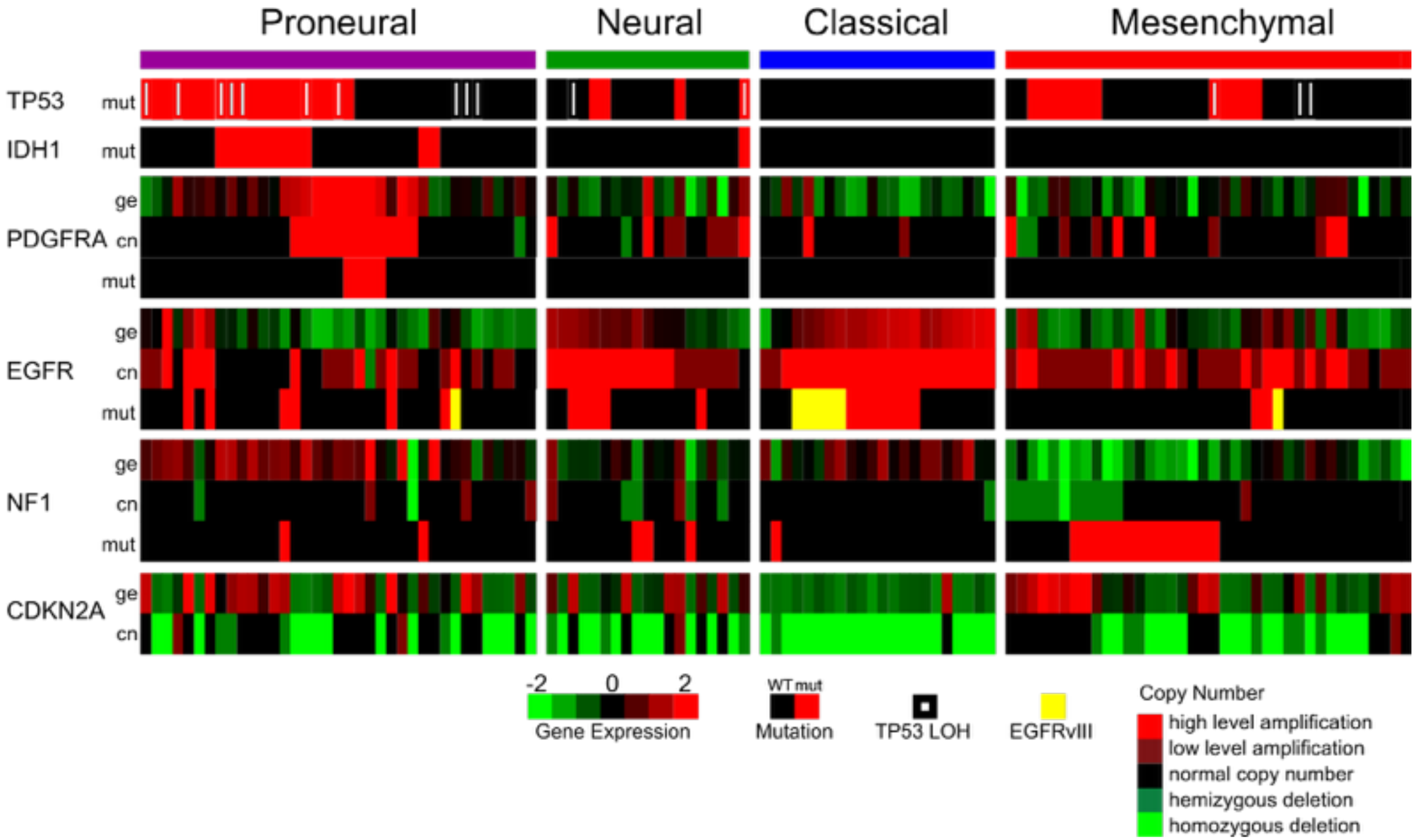


Clustered Heatmap + Dendrograms

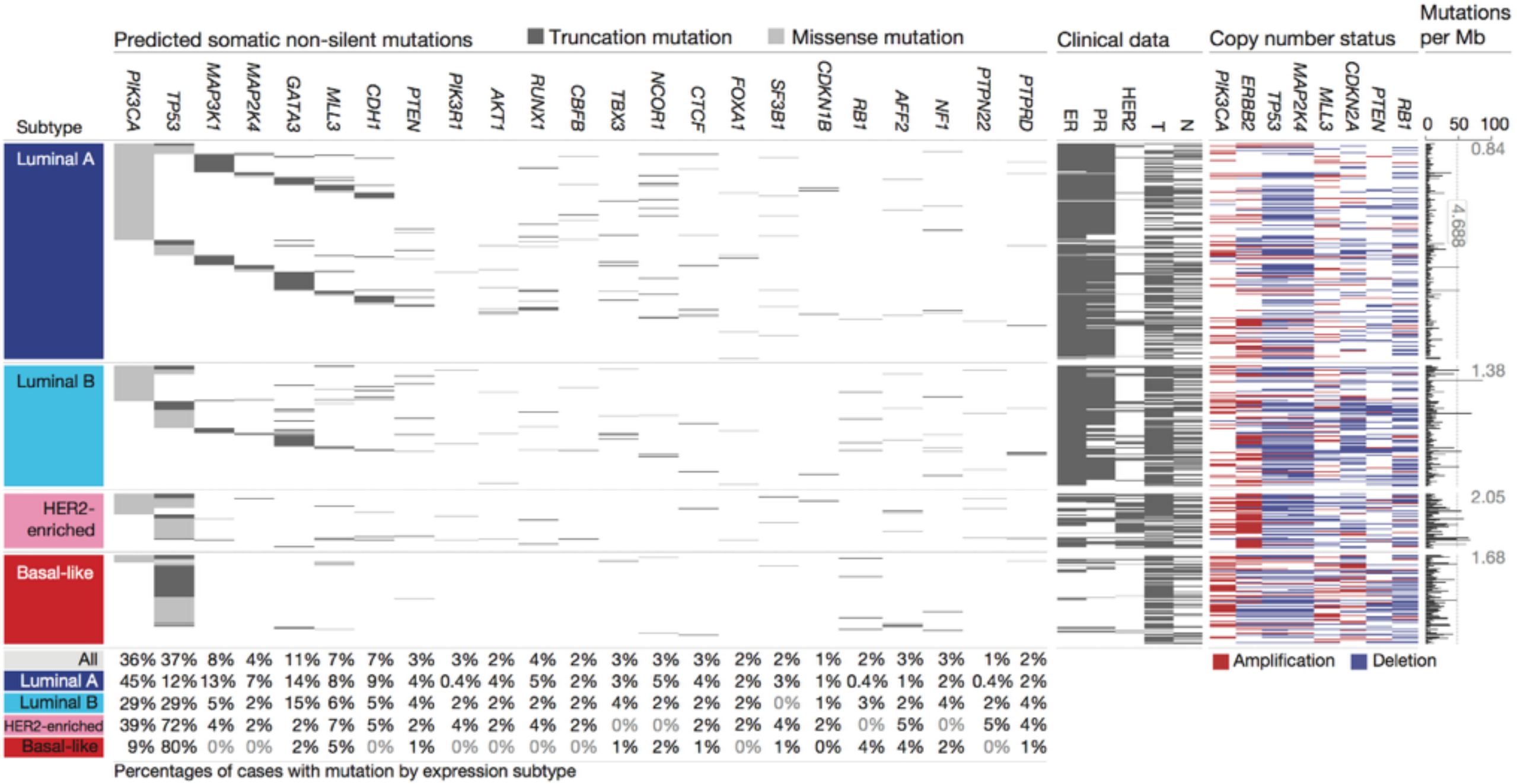


Lex et al. 2010

Heterogeneous Data?



Verhaak 2012



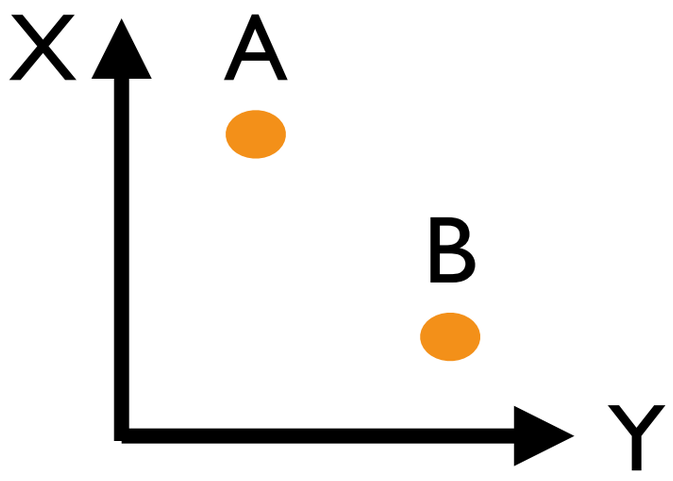
Heterogeneous Heatmap
TCGA Paper, Nature 2012

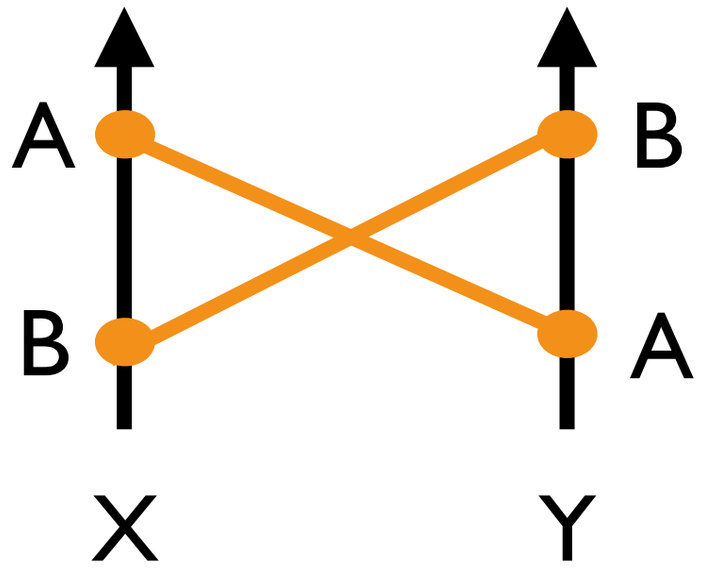
Parallel Coordinates (PC)

Inselberg 1985

Axes represent attributes

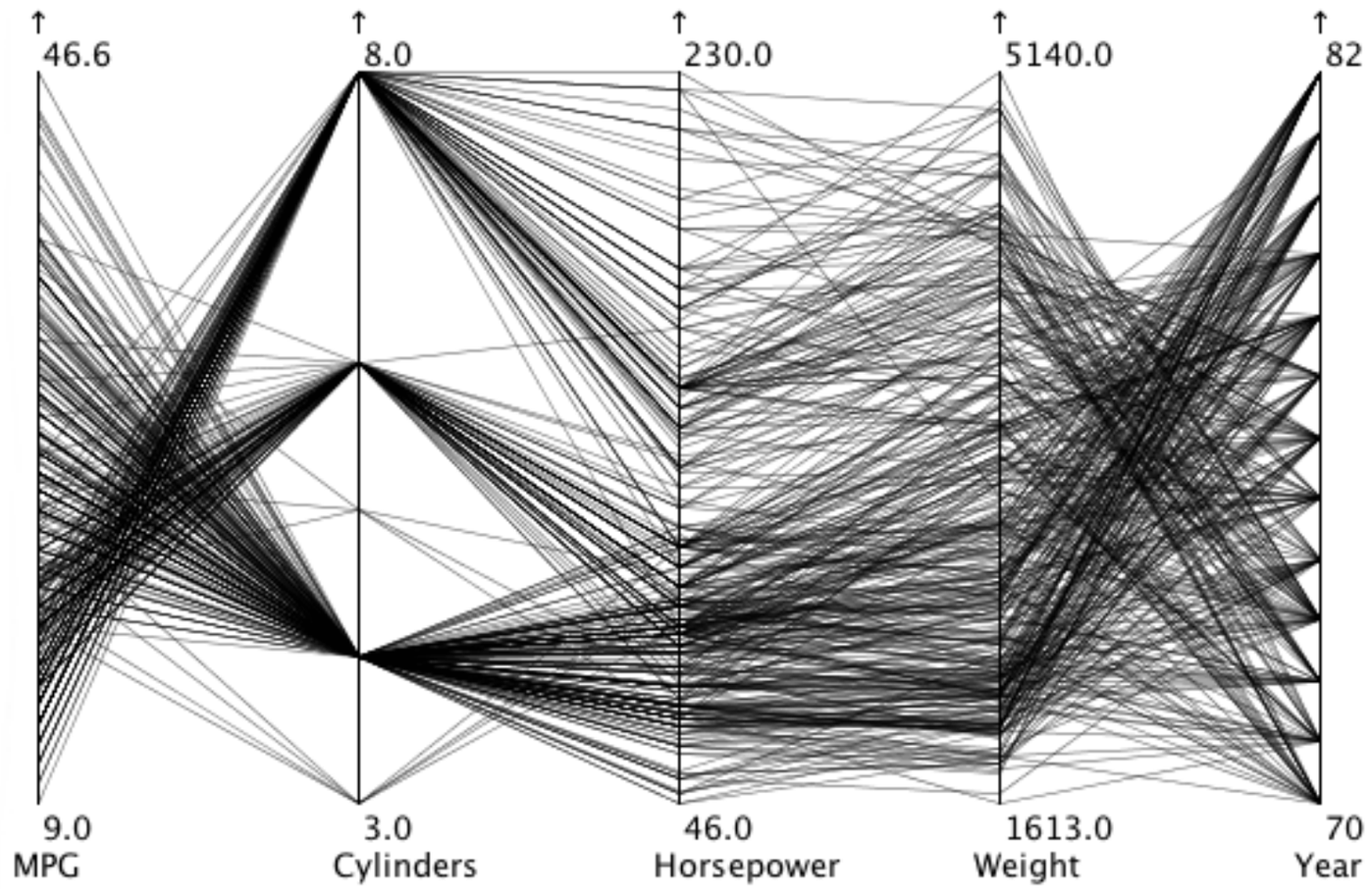
Lines connecting axes represent items





Example: Cars Dataset

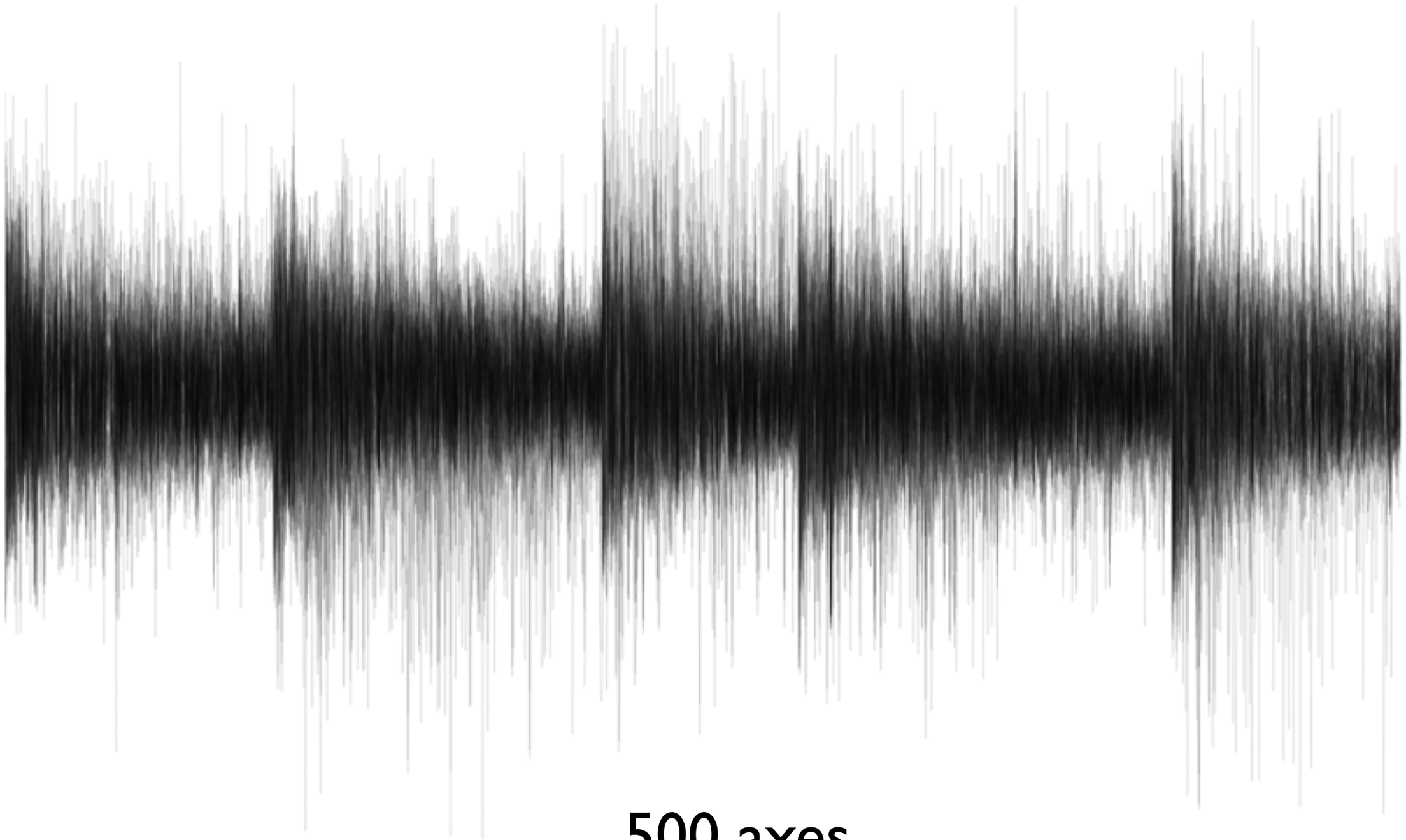
1	MPG	Cylinders	Horsepower	Weight	Acceleration	Year	Origin
2	18	8	130	3504	12	70	USA
3	15	8	165	3693	11.5	70	USA
4	18	8	150	3436	11	70	USA
5	16	8	150	3433	12	70	USA
6	17	8	140	3449	10.5	70	USA
7	15	8	198	4341	10	70	USA
8	14	8	220	4354	9	70	USA
9	14	8	215	4312	8.5	70	USA
10	14	8	225	4425	10	70	USA
11	15	8	190	3850	8.5	70	USA
12	15	8	170	3563	10	70	USA
13	14	8	160	3609	8	70	USA
14	15	8	150	3761	9.5	70	USA
15	14	8	225	3086	10	70	USA
16	24	4	95	2372	15	70	Europe
17	22	6	95	2833	15.5	70	USA
18	18	6	97	2774	15.5	70	USA
19	21	6	85	2587	16	70	USA



Limitations of PC?

PC Limitations

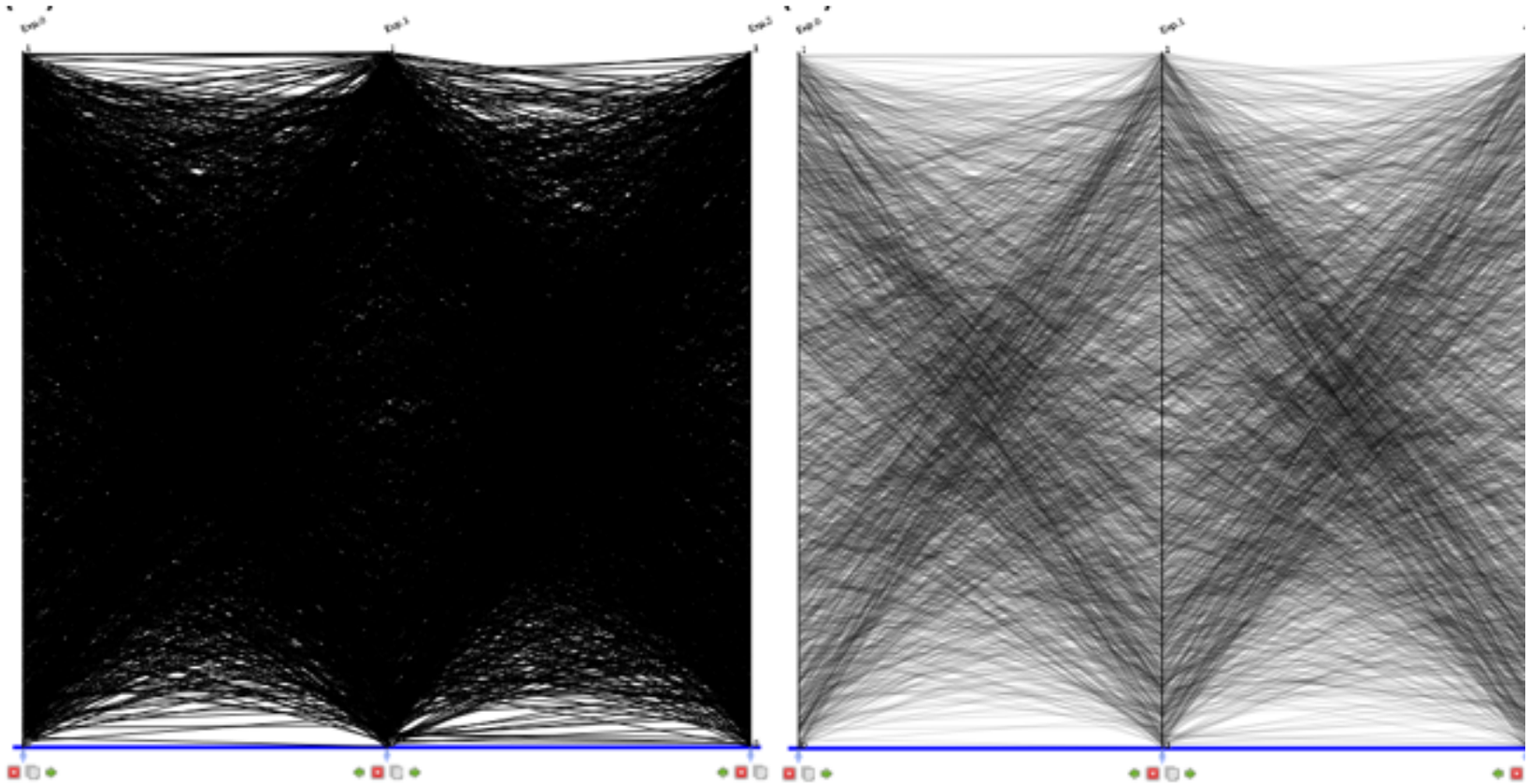
Scalability to Many Dimensions



500 axes

PC Limitations

Scalability to Many Items

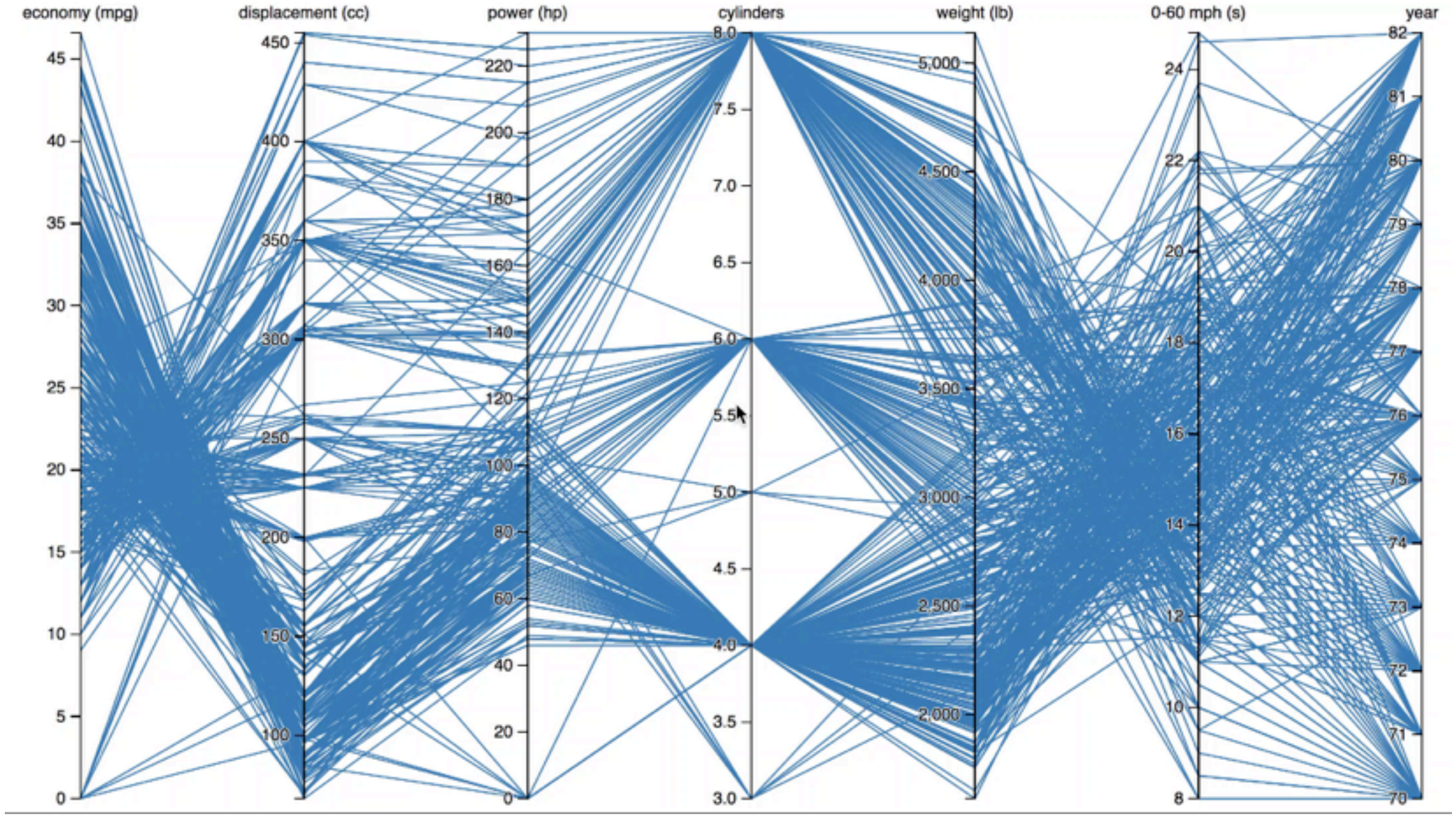


Solutions:

- Transparency
- Bundling, Clustering
- Sampling

PC Limitations

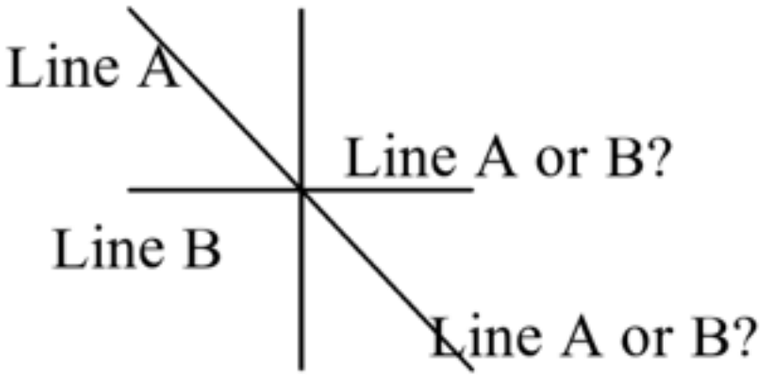
Correlations only between adjacent axes

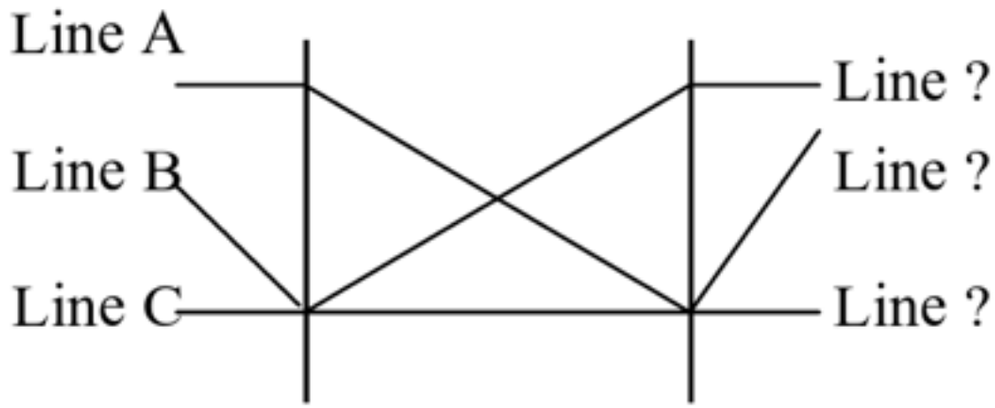


Solution: Let user change order

PC Limitations

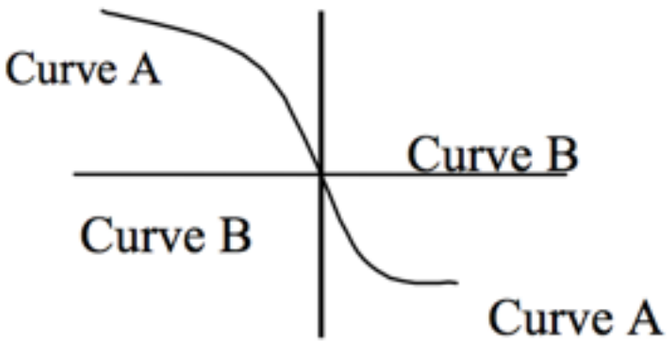
Ambiguity

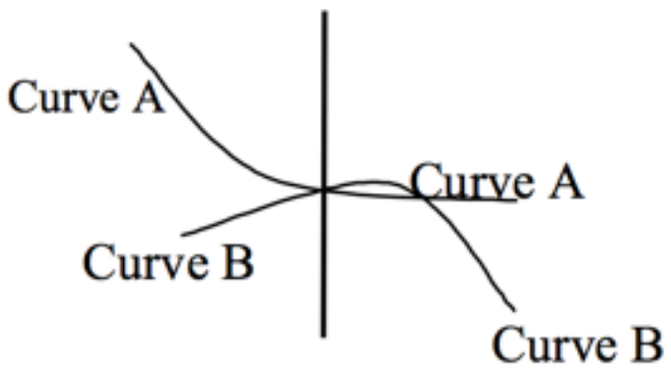


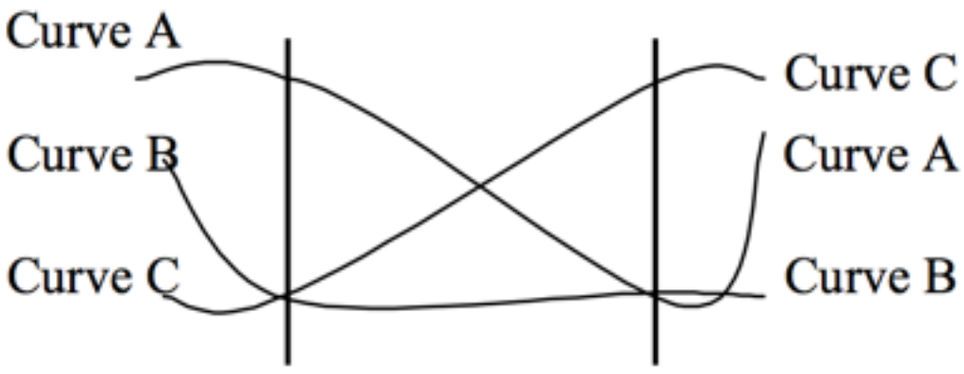


Solutions:

- Interactive highlighting
- Curves







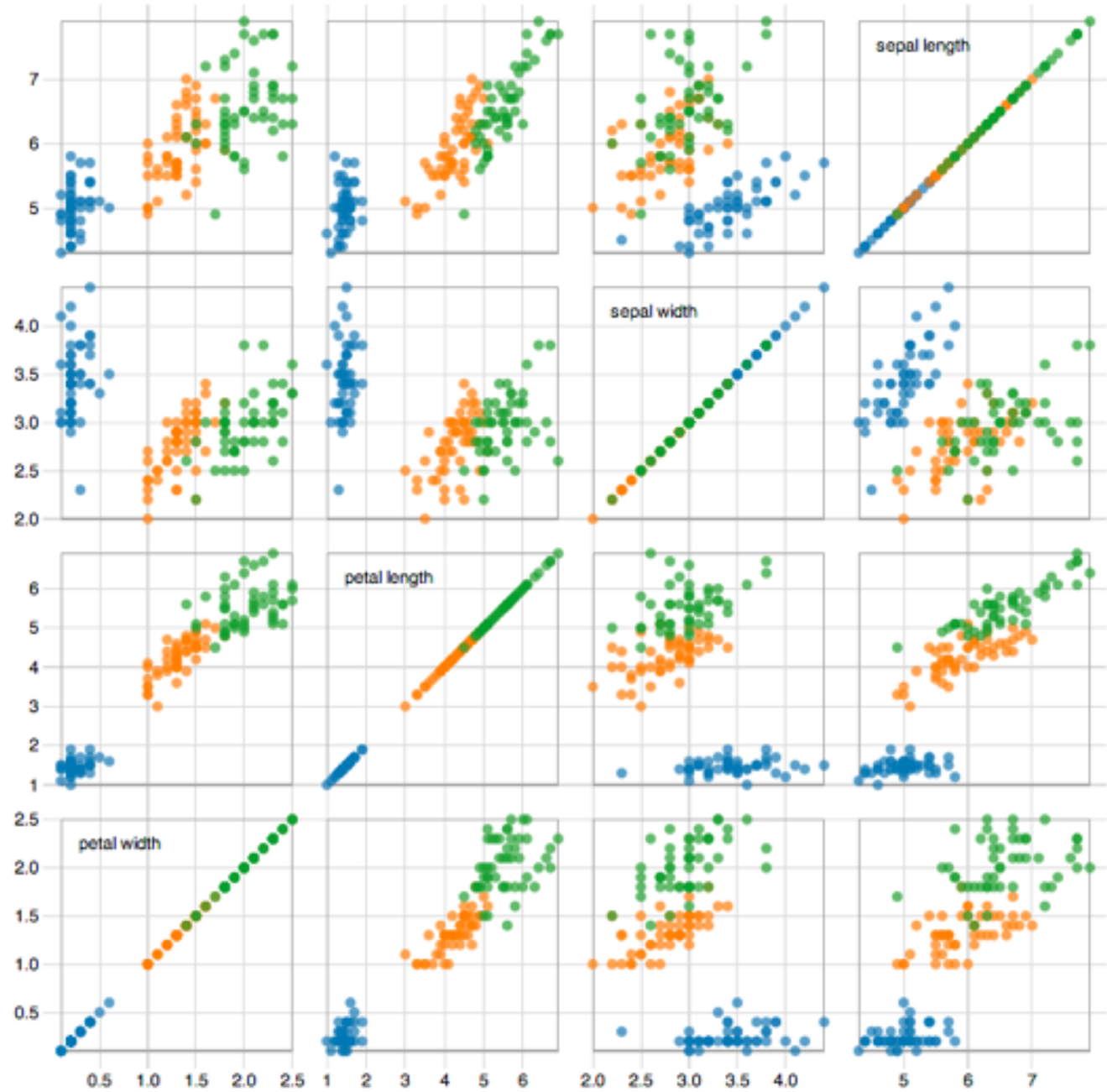
Graham and Kennedy 2003

Scatterplot Matrix (SPLOM)

N dimensions

N^2 scatterplots

Limited scalability
(~20 dims,
~500-1k items)

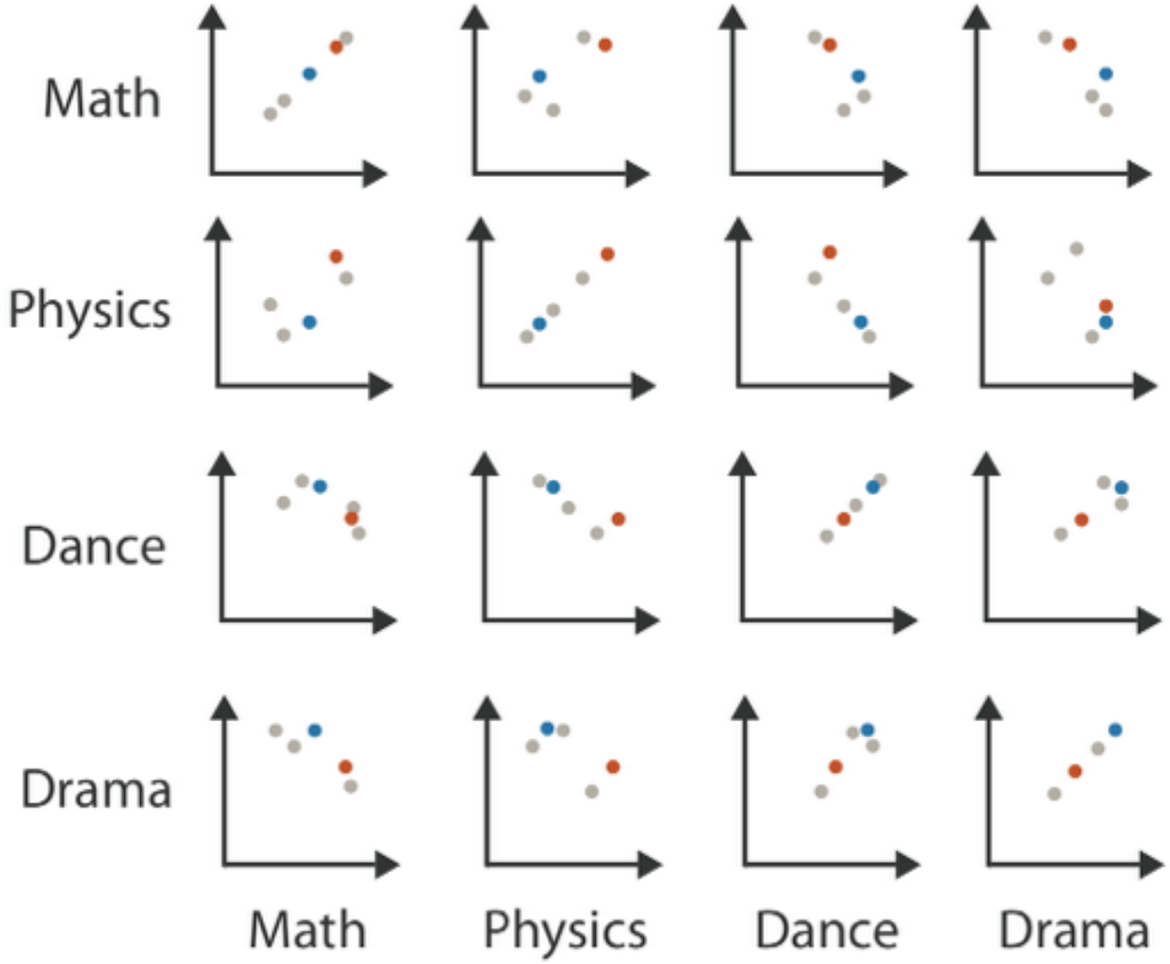


http://bl.ocks.org/mbostock/4063663

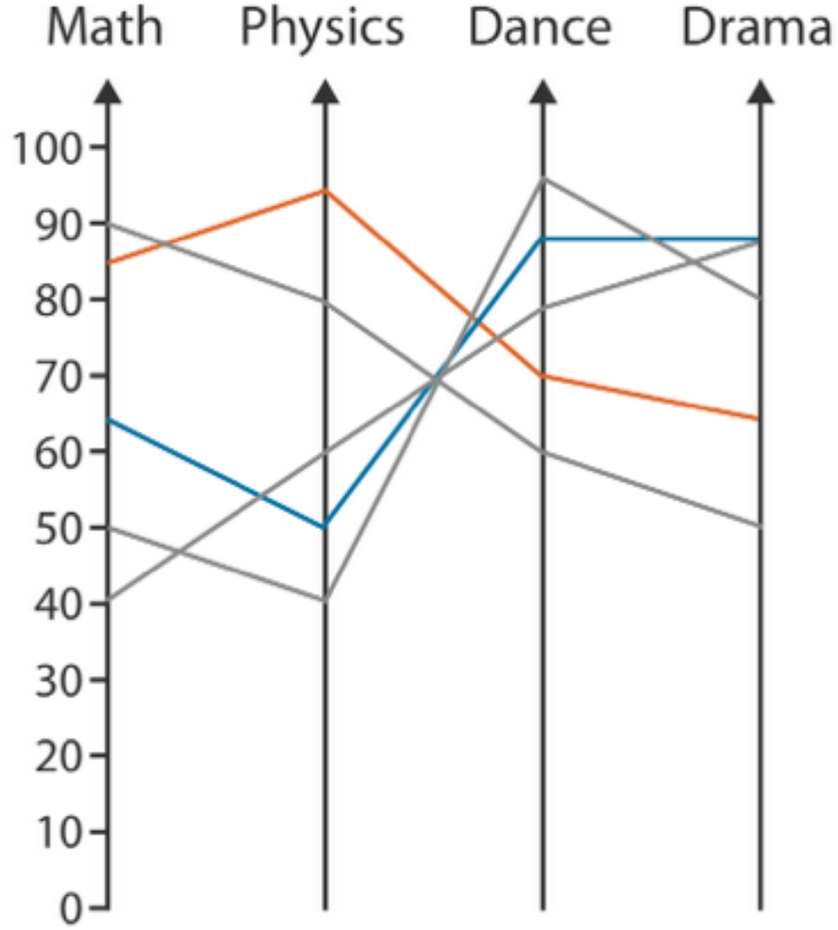
Table

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

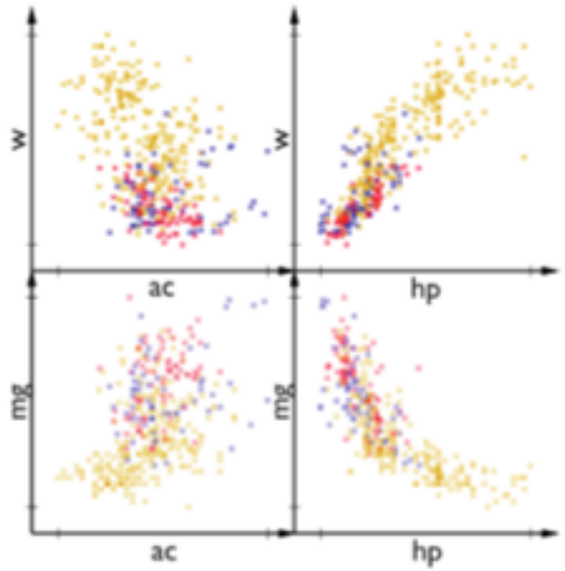


Parallel Coordinates

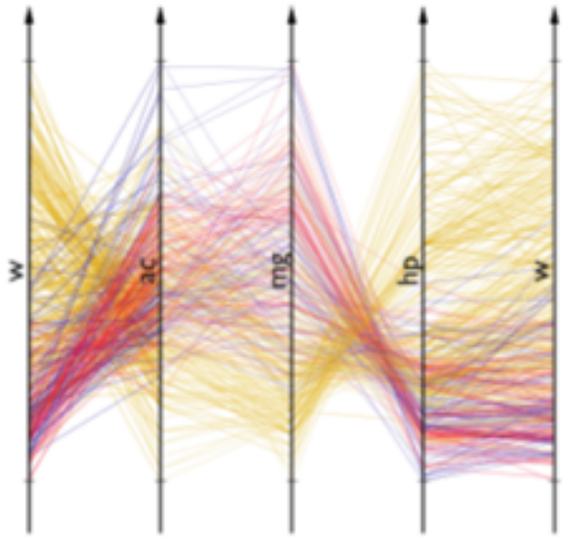


Munzner 2014

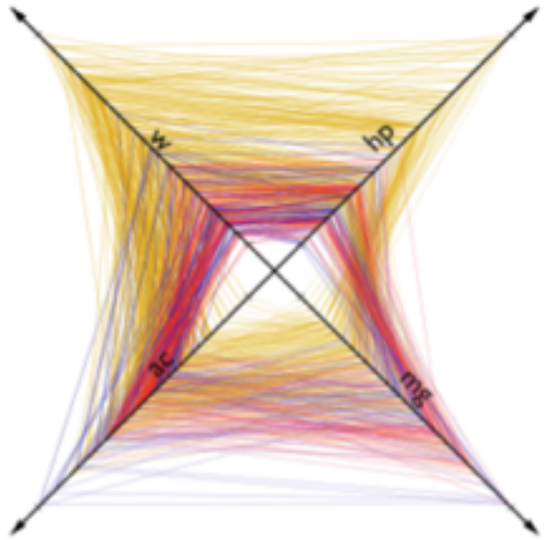
Flexible Linked Axes (FLINA)



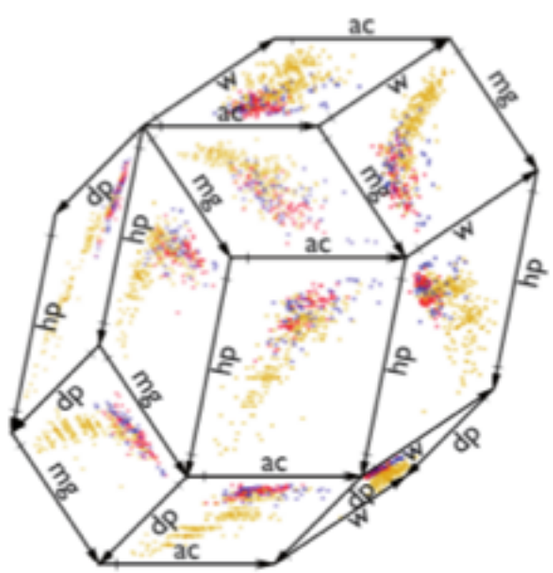
(a) scatterplots



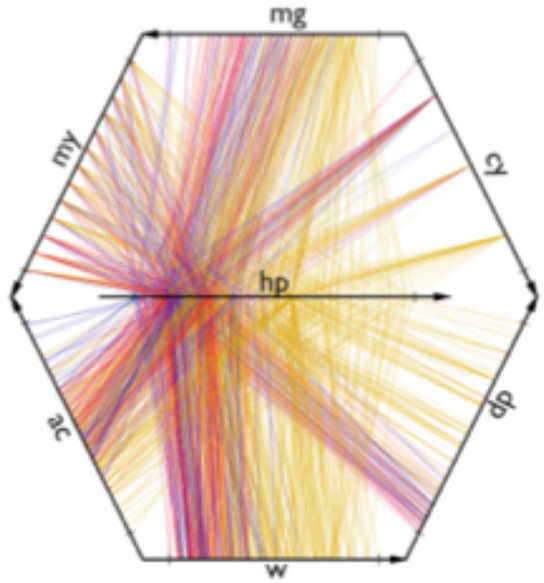
(b) Parallel Coordinates Plot



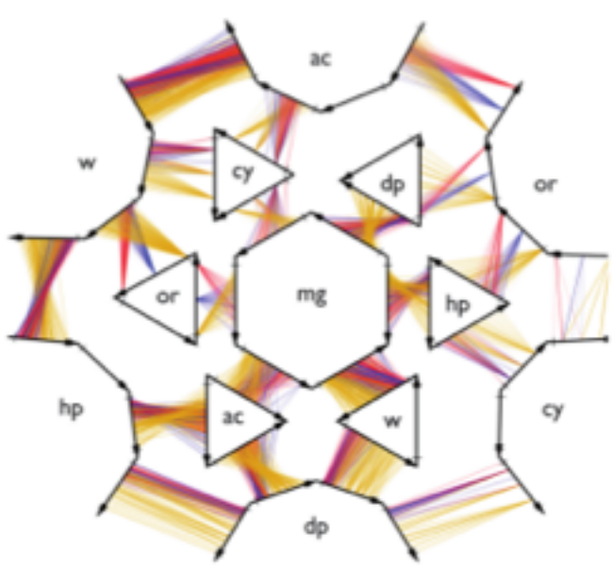
(c) radar chart



(d) Hyperbox



(e) Time Wheel

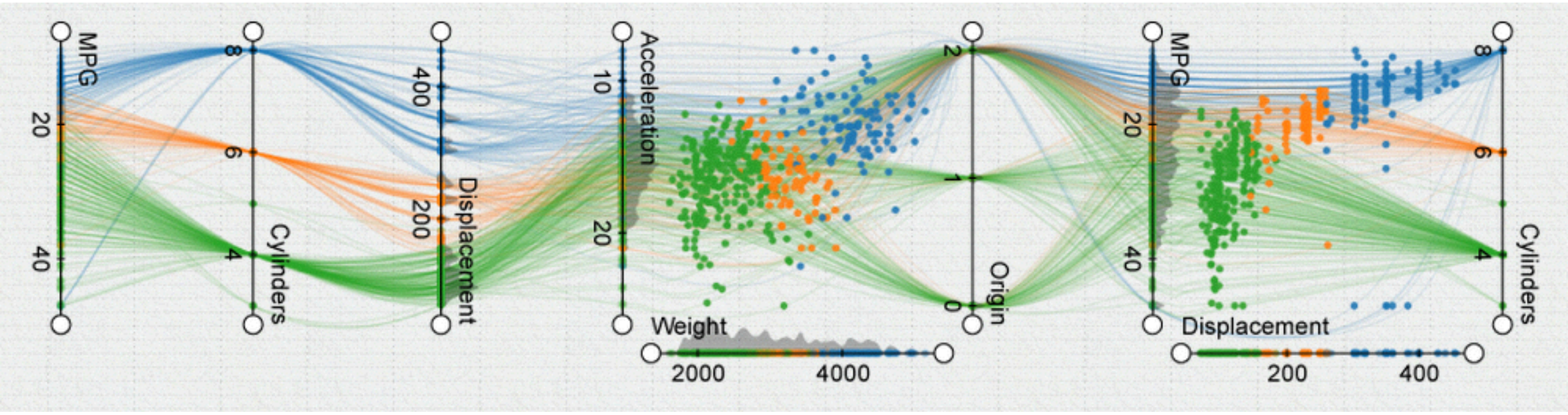


(f) Many-to-many PCP

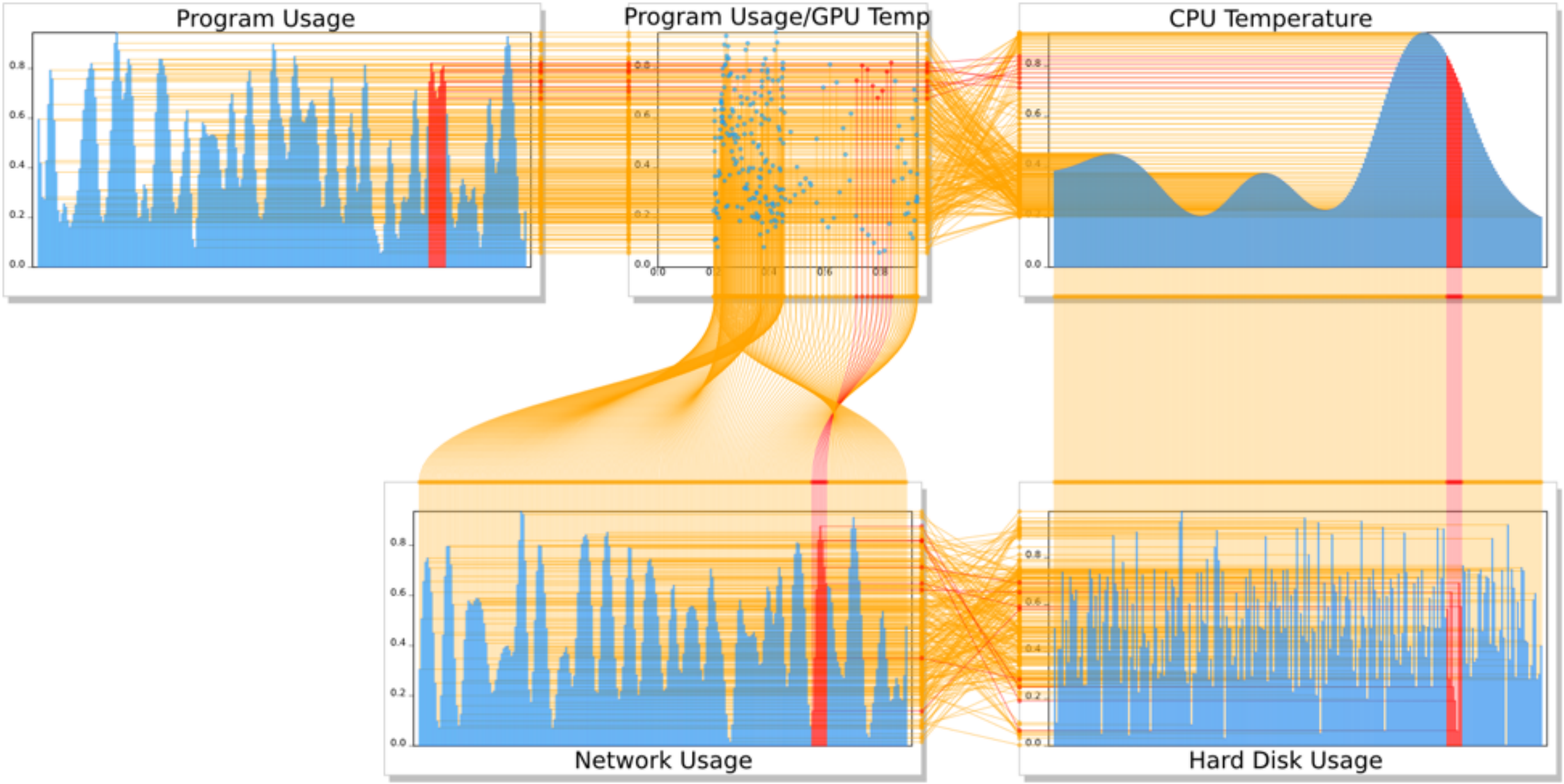
Claessen & van Wijk 2011

Web-based implementation of FLINA concept

http://vis.pku.edu.cn/mddv/val/



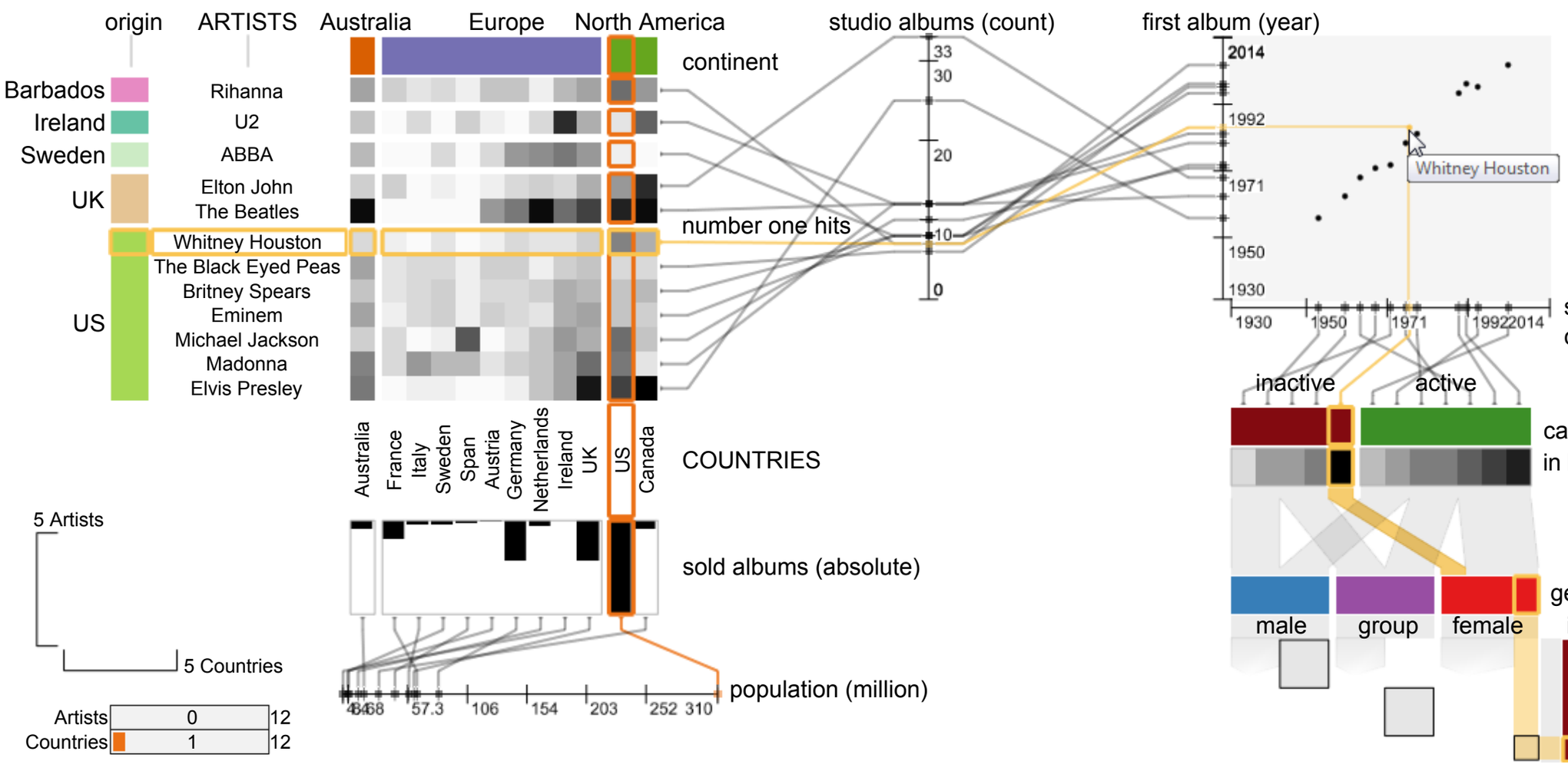
Connected Charts



Viau & McGuffin 2012

Domino

http://domino.caleydo.org



Gratzl et al. 2014

Domino

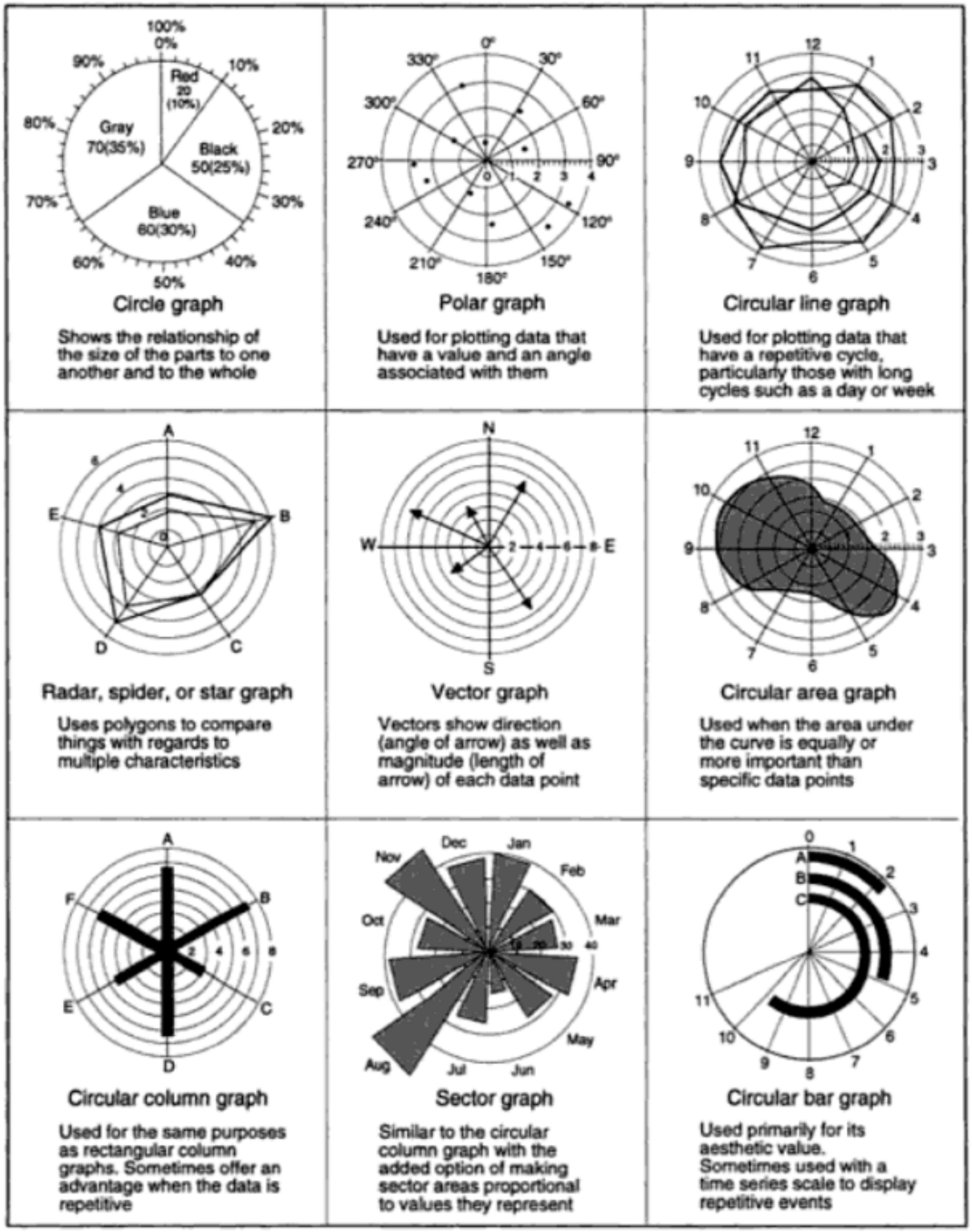
Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets

Samuel Gratzl, Nils Gehlenborg, Alexander Lex, Hanspeter Pfister, and Marc Streit

Radial Axis Techniques

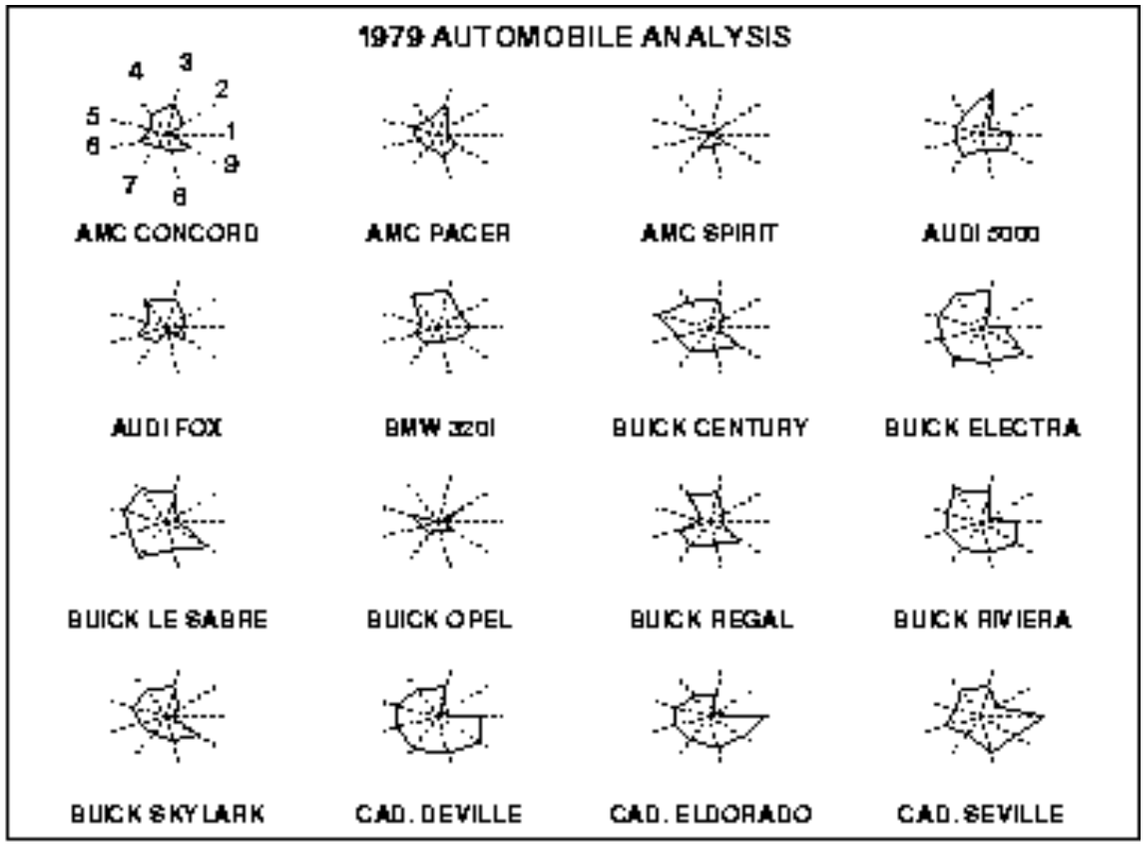
Similar to parallel coordinates

Axes radiate from a common origin

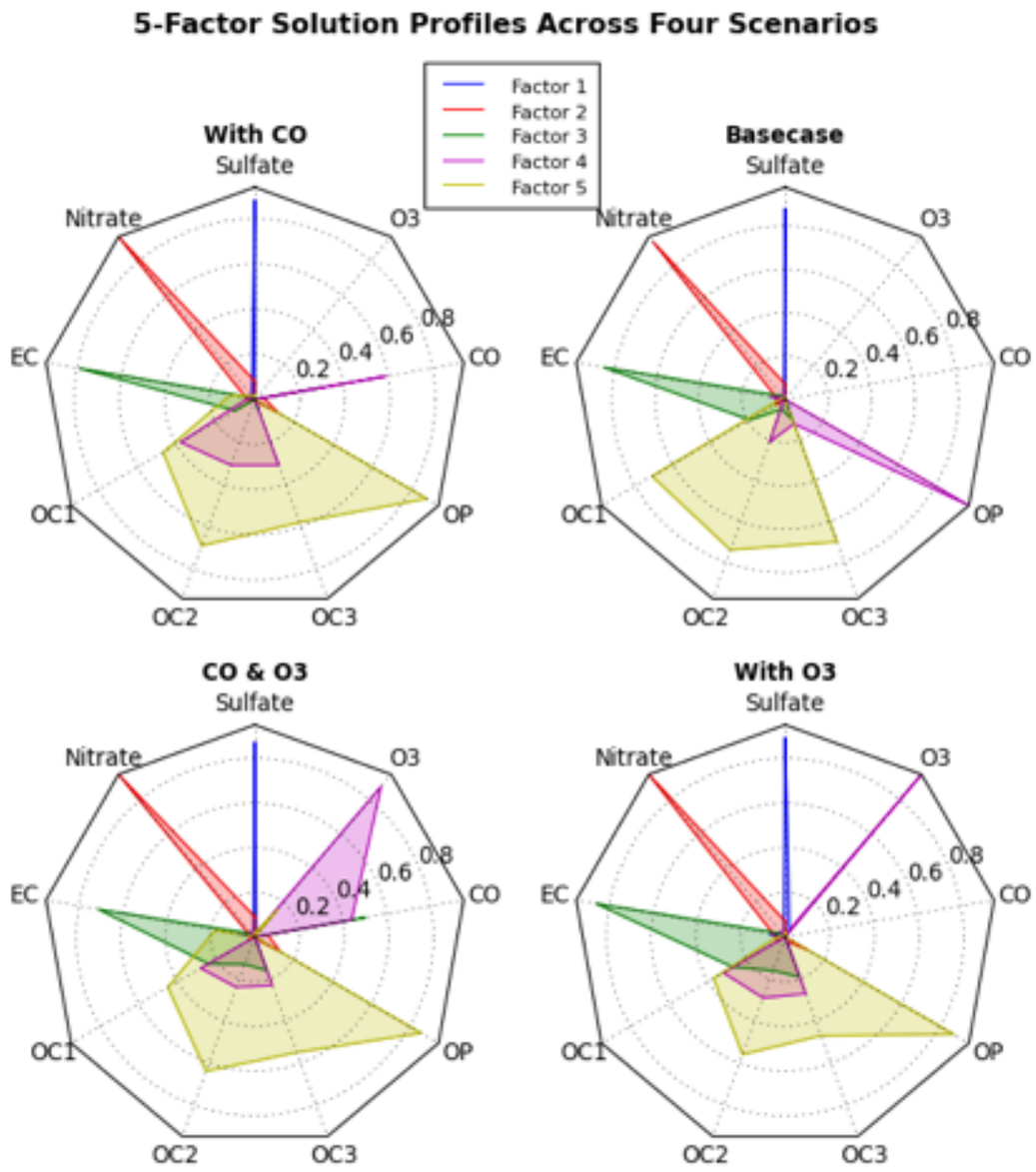


Harris 1999

Star Plot



Coekin 1969

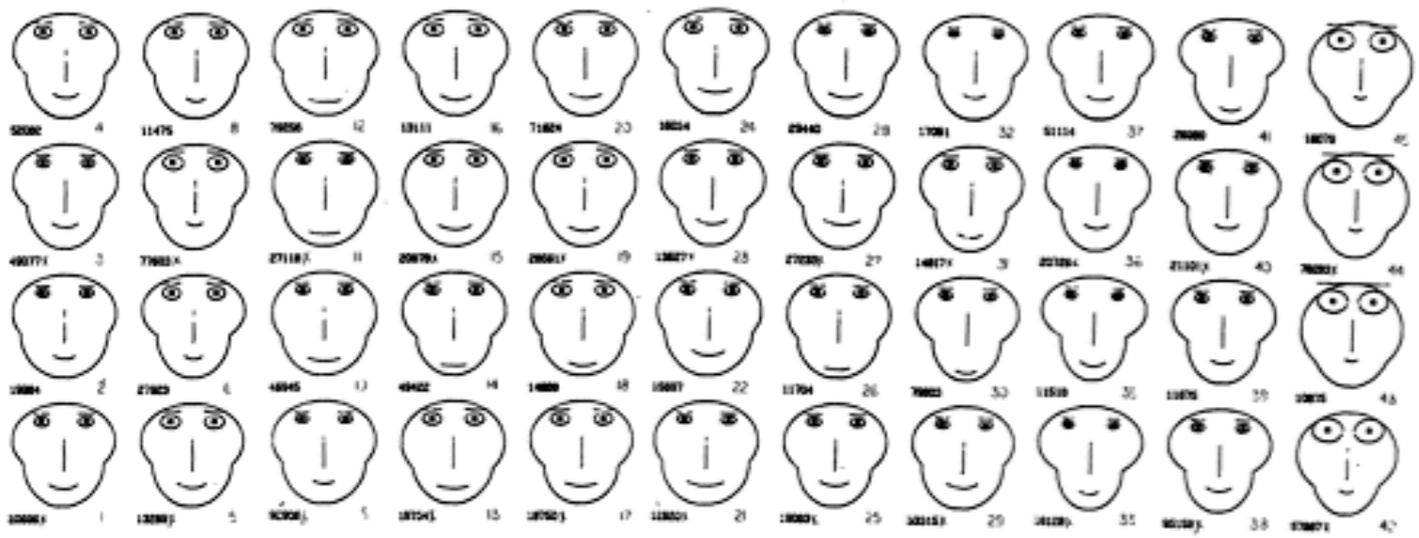


http://matplotlib.org/examples/api/radar_chart.html

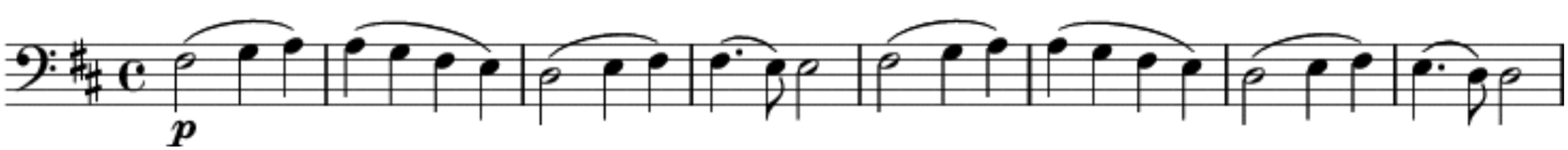
Chernoff Faces

Used visual attributes:

- size of face
- curvature of face
- position of the eyes
- length of the nose
- position of the mouth



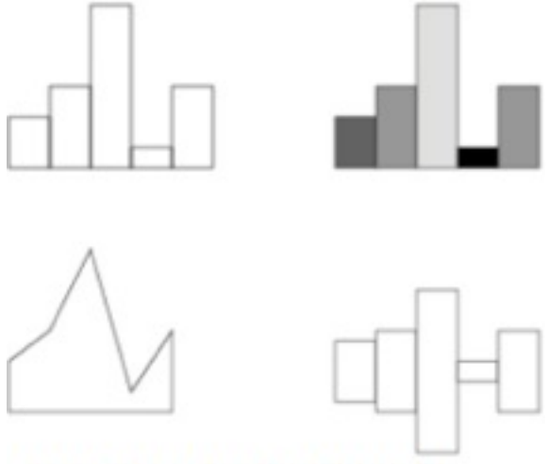
Does not work!



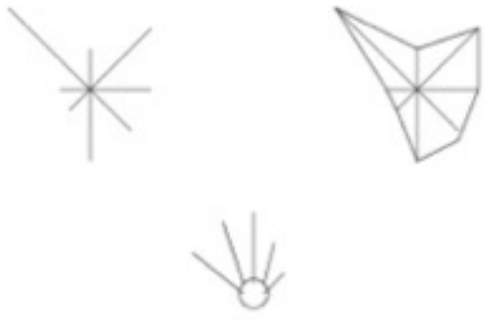
Glyphs / Icons

Visual encoding of multiple data values
in an object or symbol

Mapping of data values to visual attributes



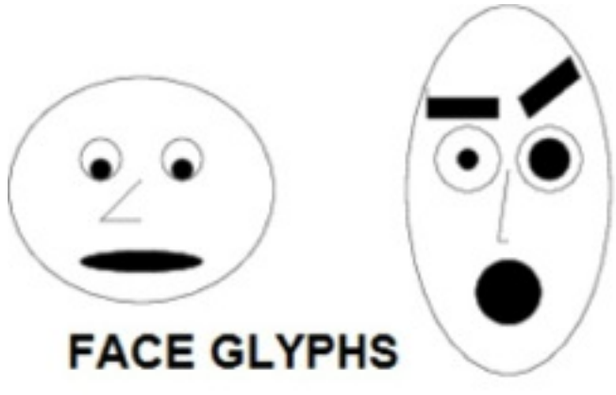
PROFILE GLYPHS



STARS AND
METROGLYPHS

STICKS AND TREES

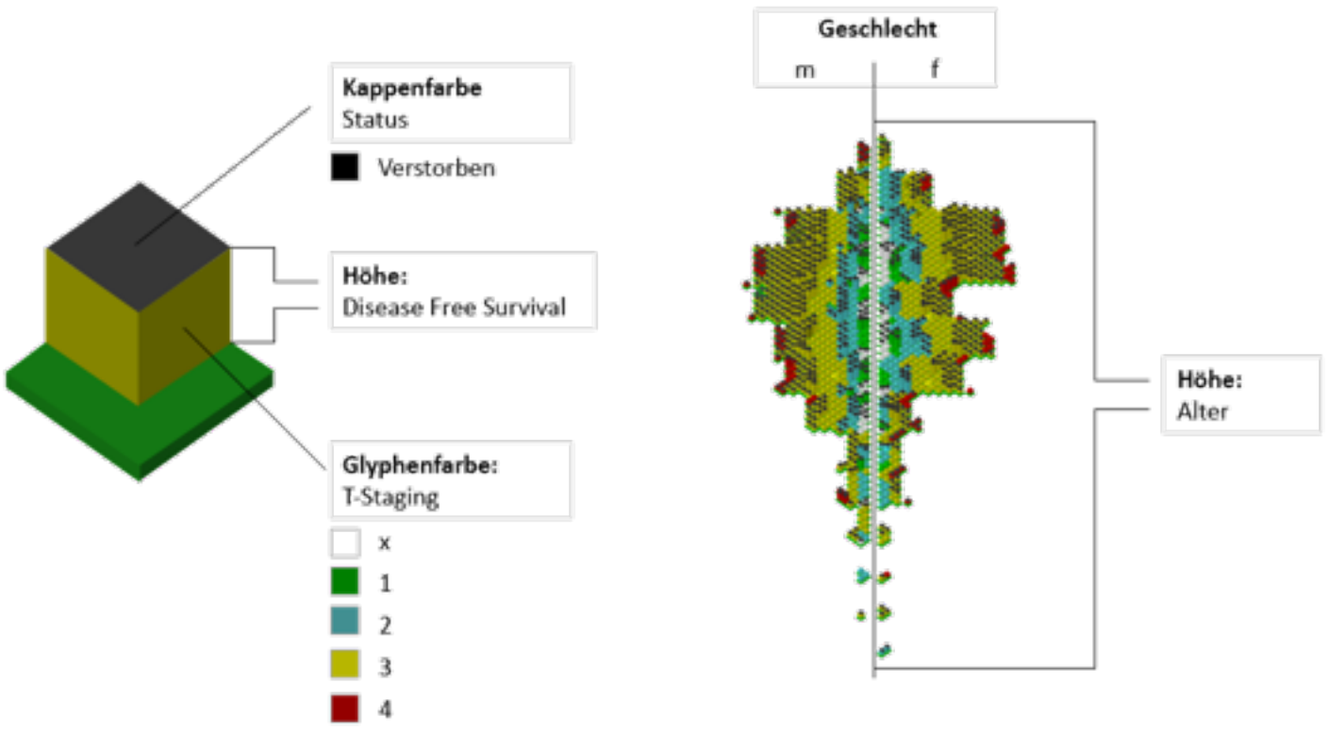
AUTOGLYPH/BOX GLYPH



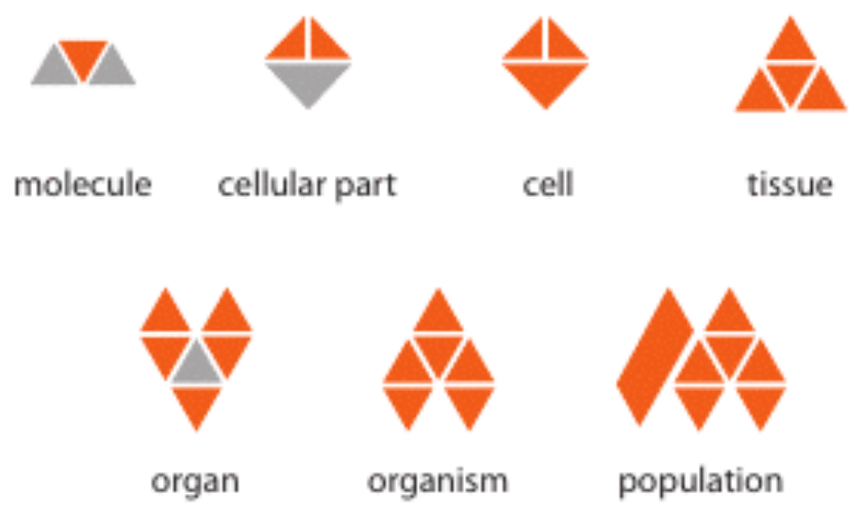
FACE GLYPHS

ARROWS/WEATHERVANES

Glyph Examples

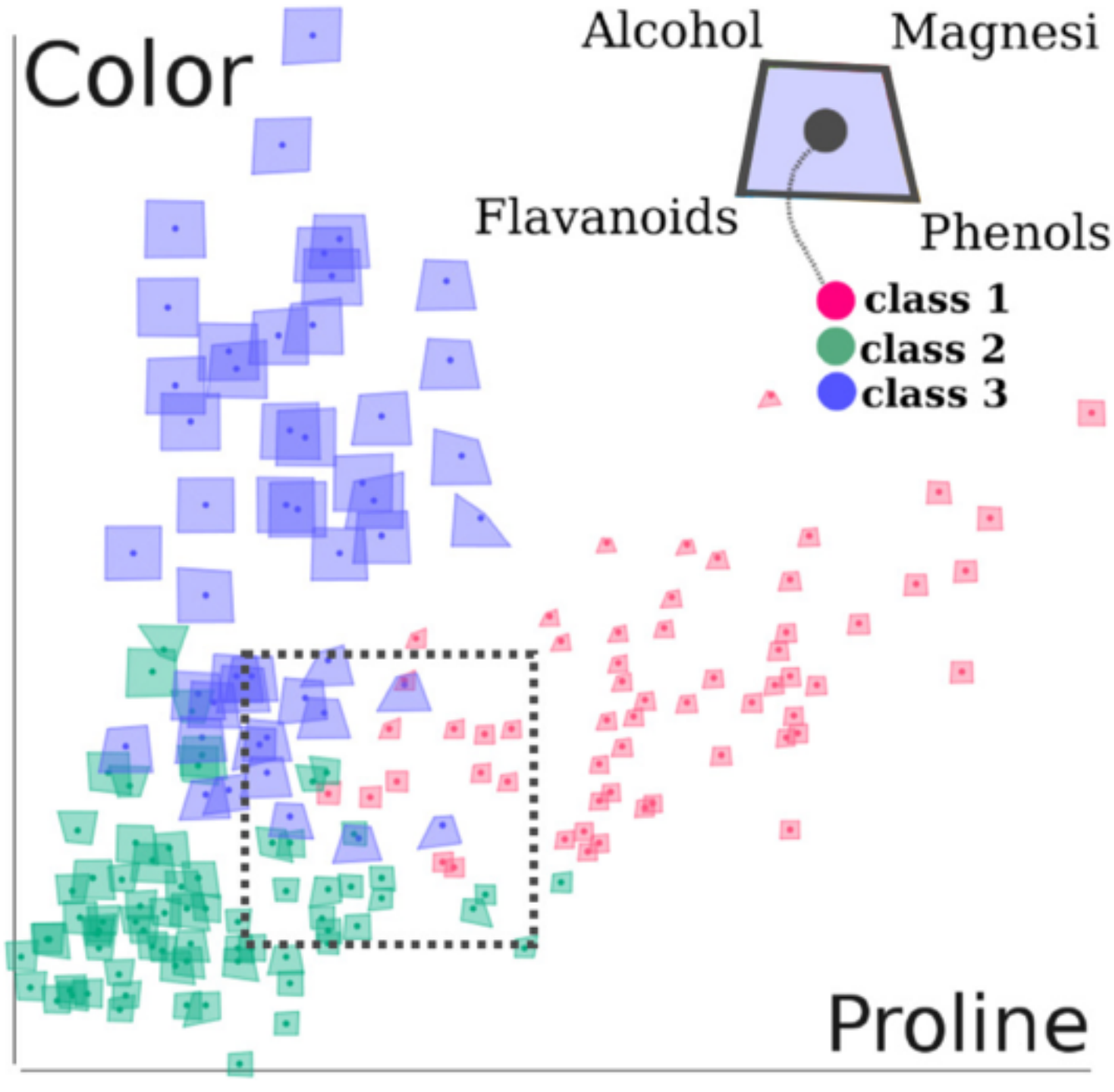


Mueller et al. 2007



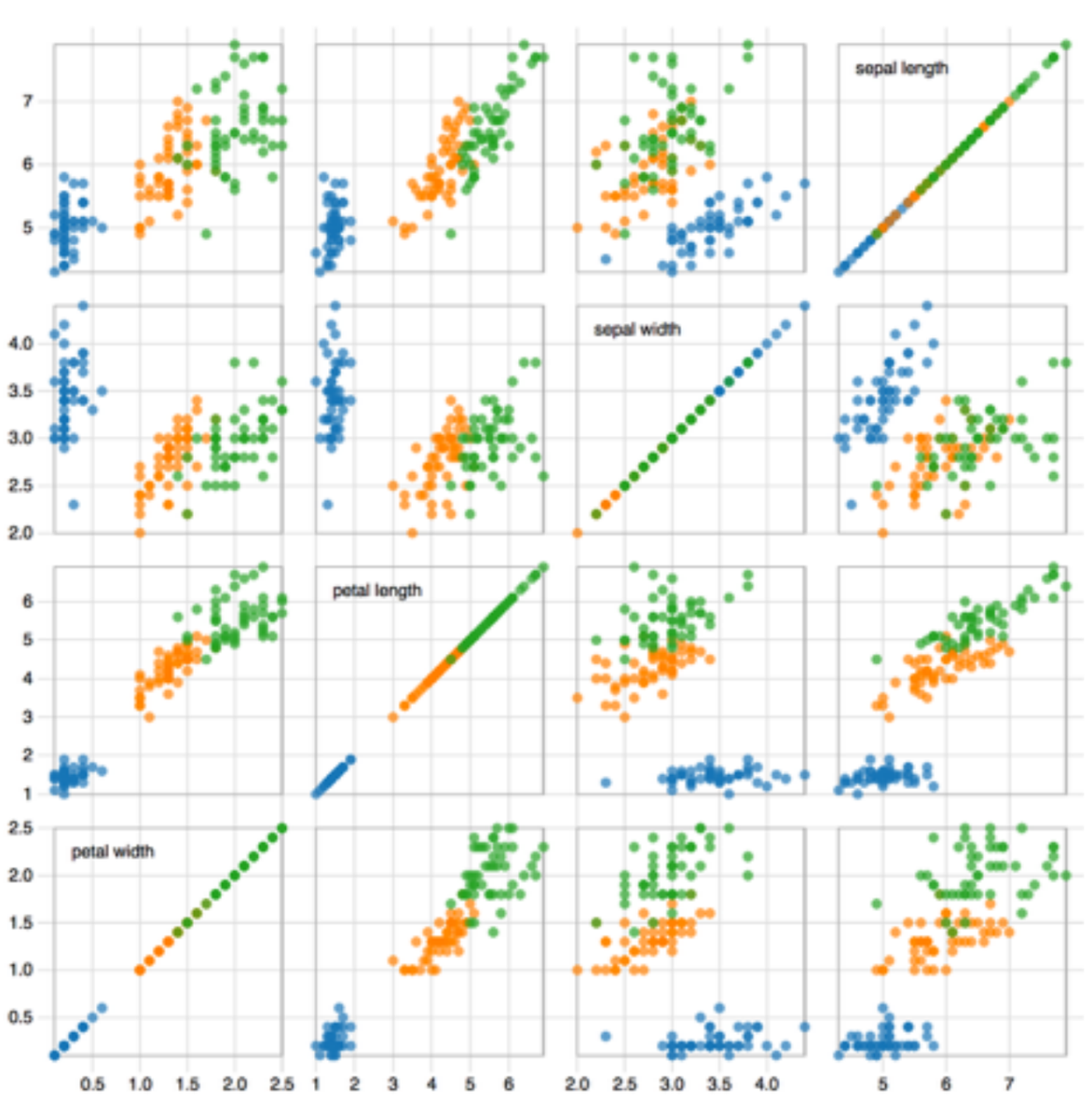
Maguire et al. 2012

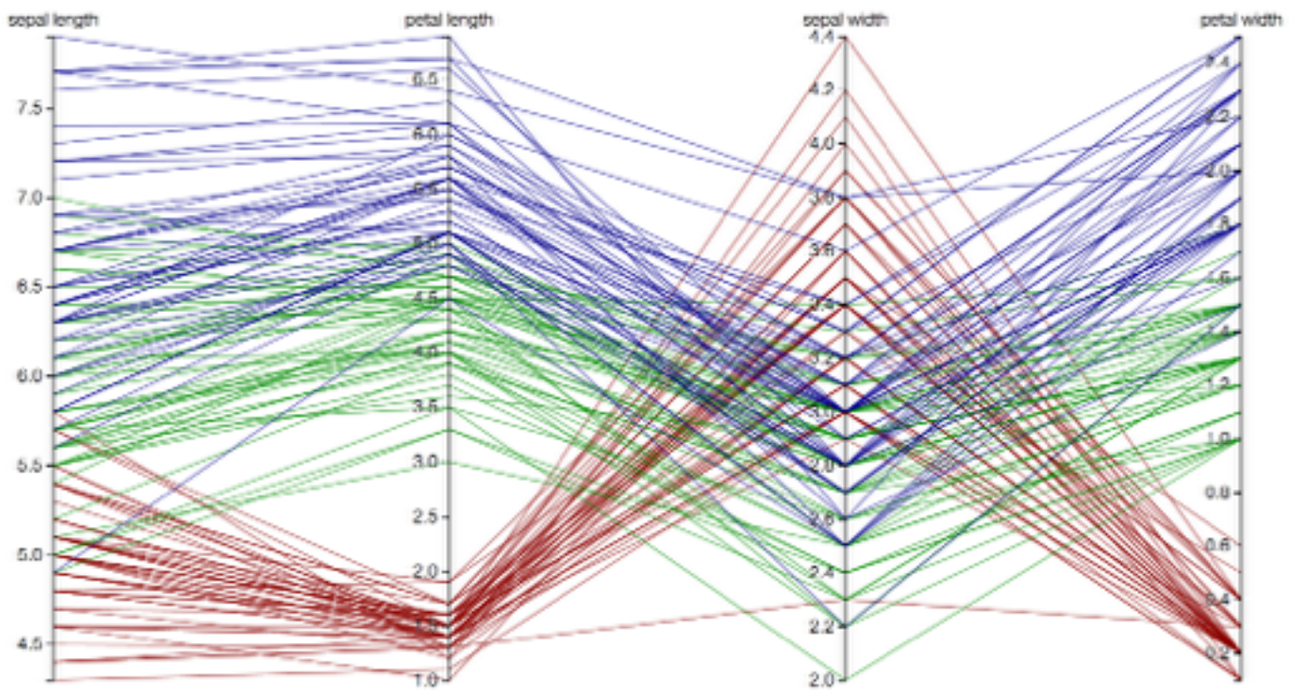
Glyph as Marks in Scatterplot



Chan et al. 2013

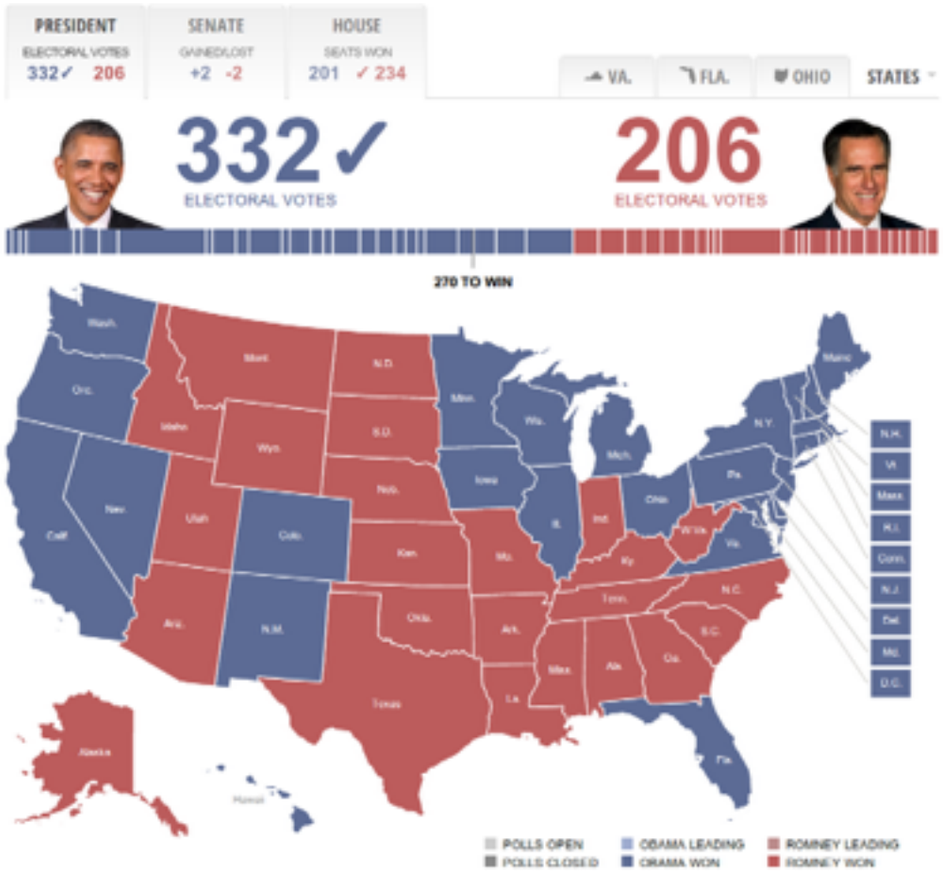
Single
Complex
Visualization





Multiple
Simple
Visualizations

Map Visualization



Map Projections

Cylinder
projection

Plane
projection

Cone
projection

Projection Overview
http://www.jasondavies.com/maps/

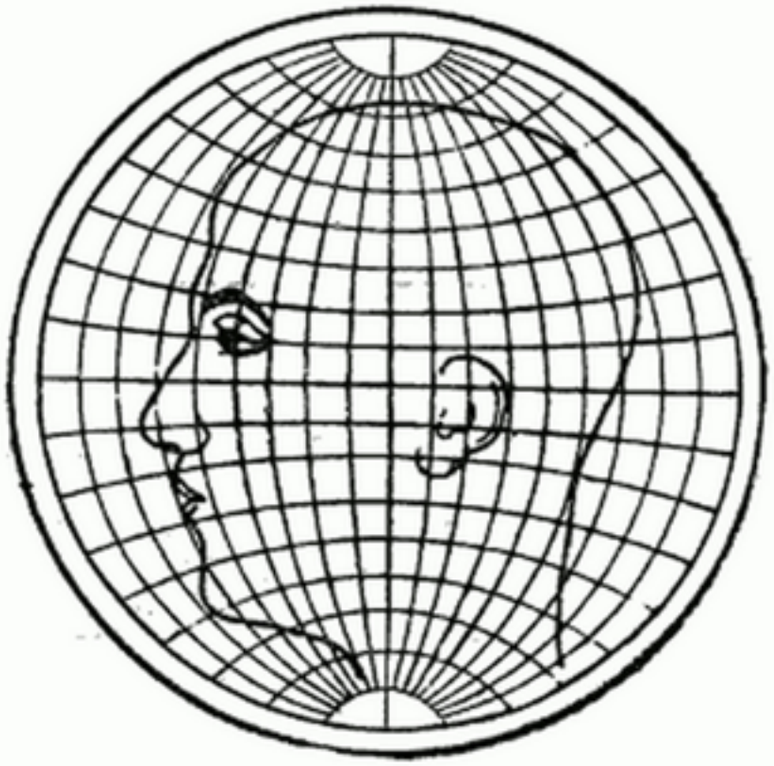


FIG. 42.—Man's head drawn on globular projection.

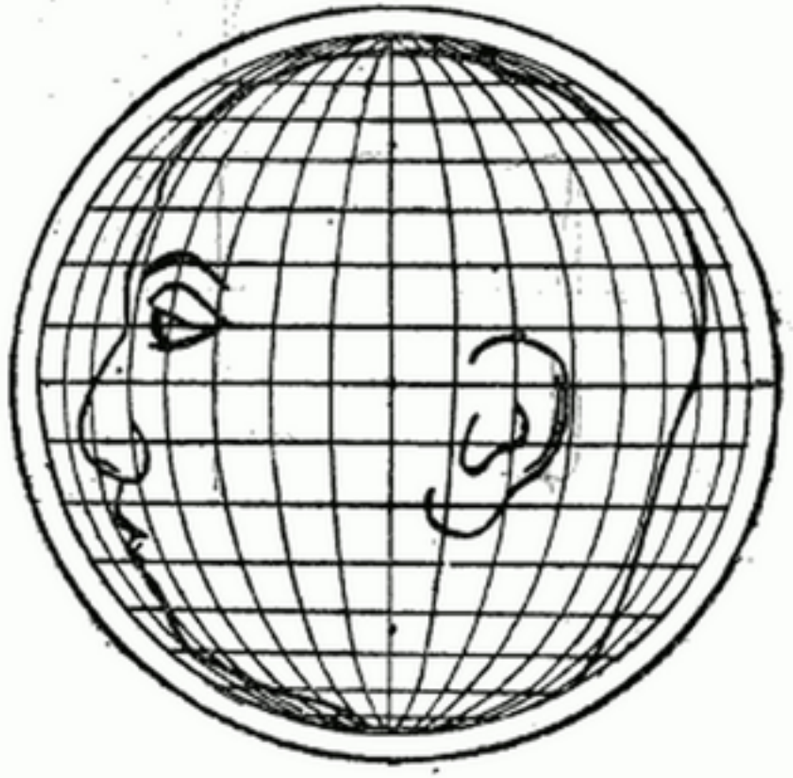


FIG. 43.—Man's head plotted on orthographic projection.

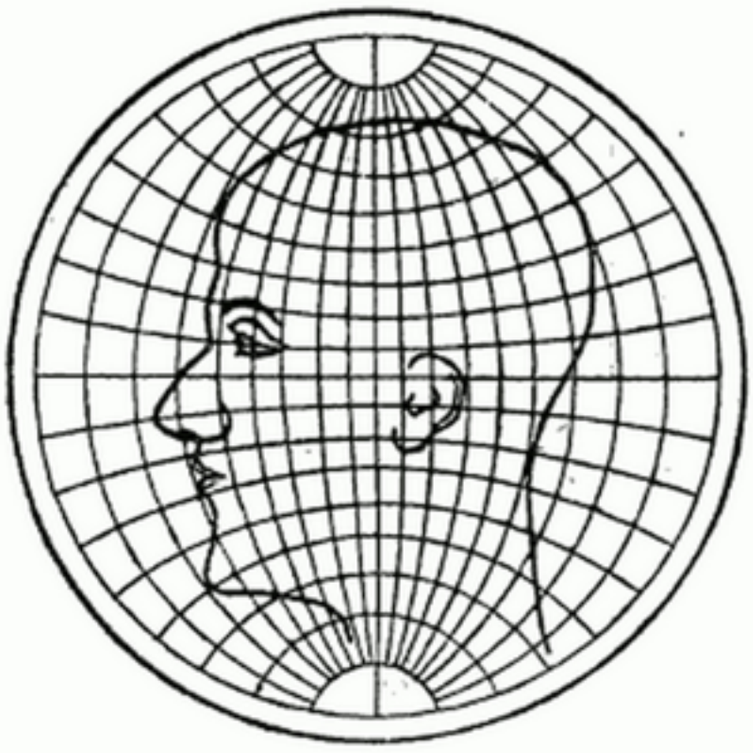


FIG. 44.—Man's head plotted on stereographic projection.

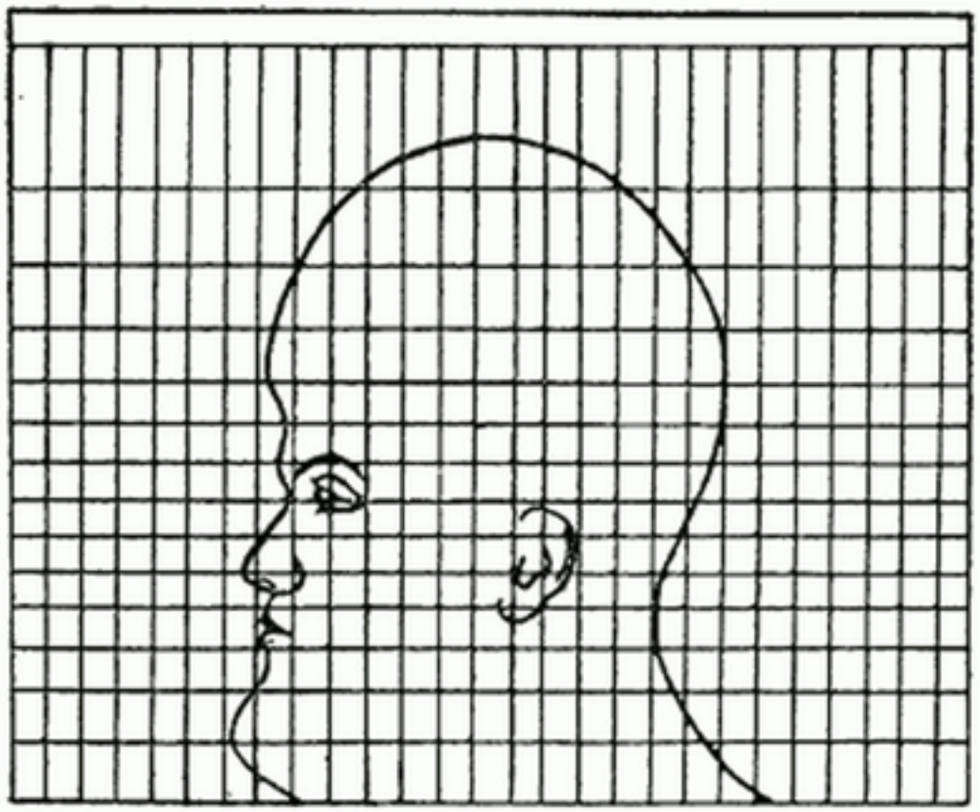
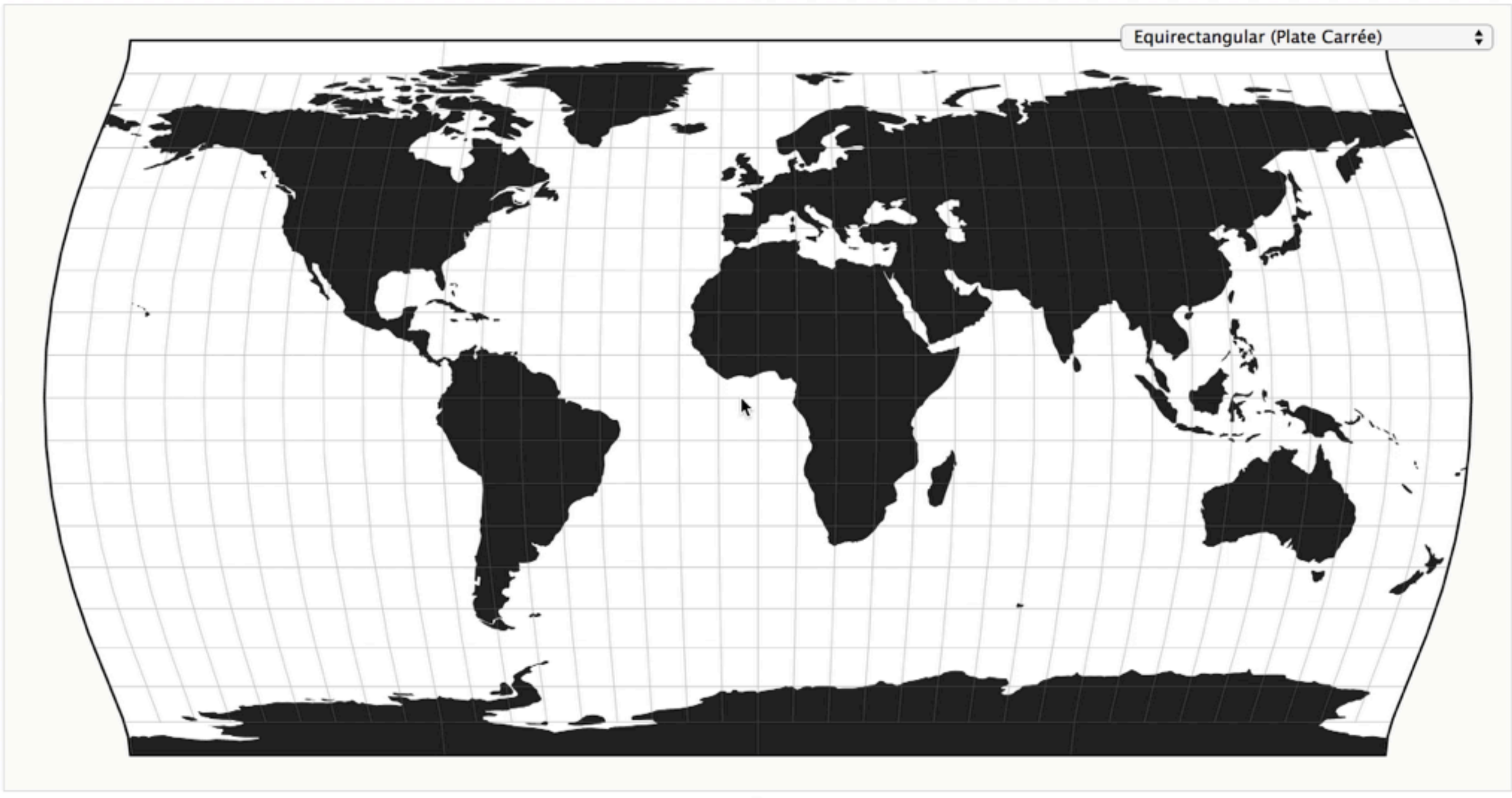


FIG. 45.—Man's head plotted on Mercator projection.

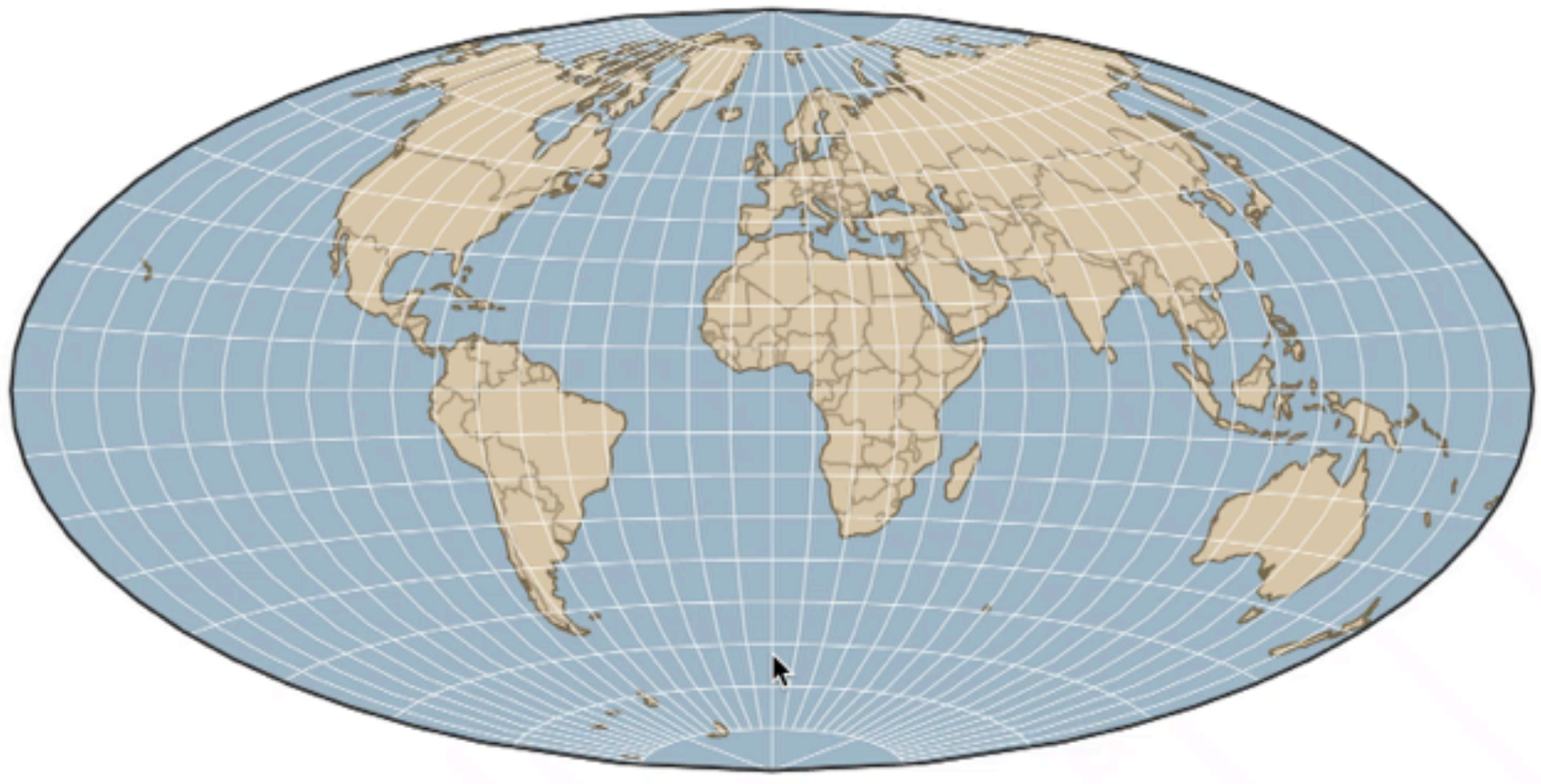
Deetz & Adams 1921

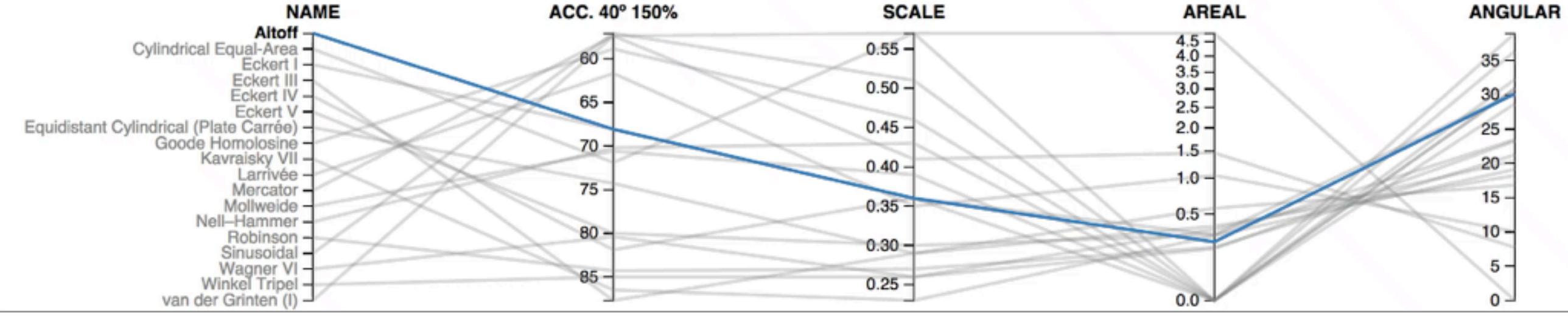
Map Projection Demo



http://bl.ocks.org/mbostock/3711652

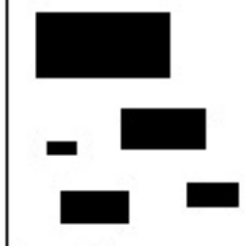
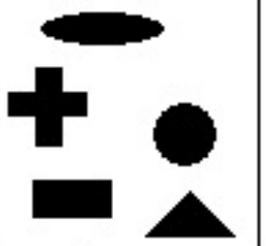
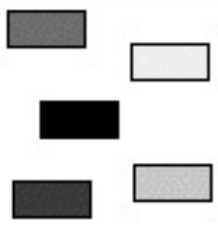
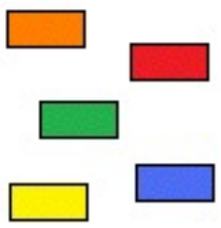
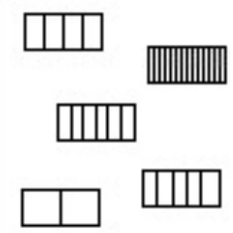
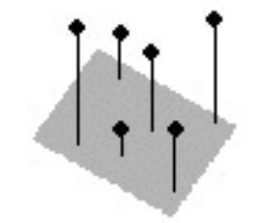
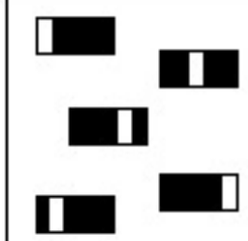
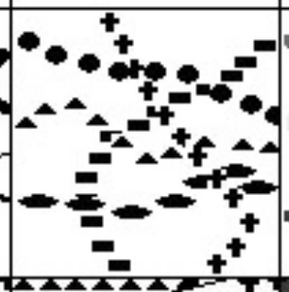
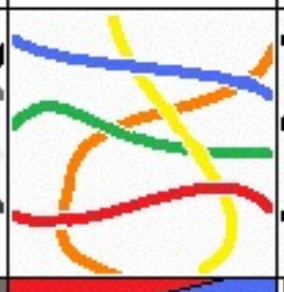
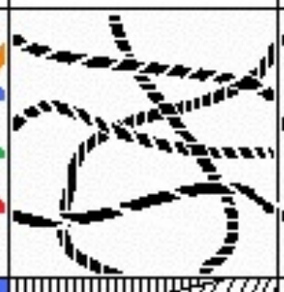
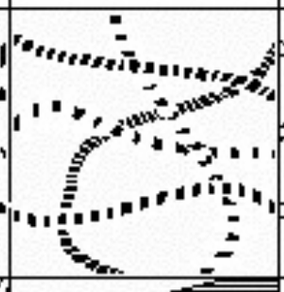
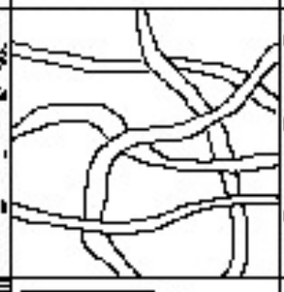
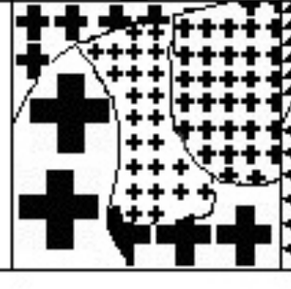
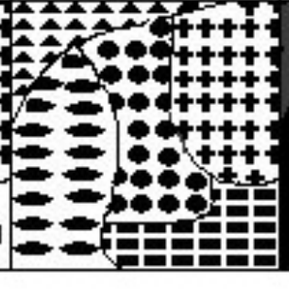
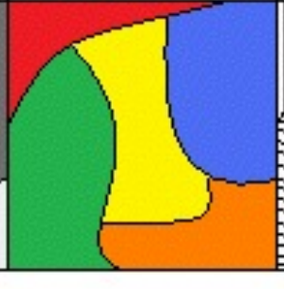
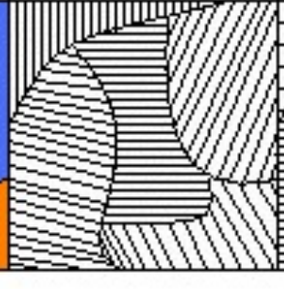
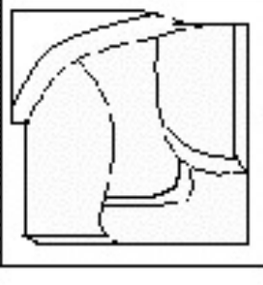
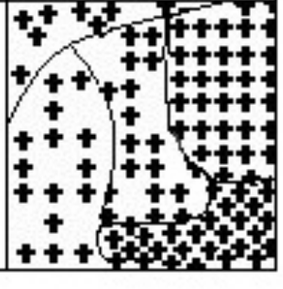
Map Projection Demo





http://bl.ocks.org/syntagmatic/3711245

Visual Variables for Spatial Data

	Size	Shape	Brightness	Color	Orientation	Spacing	Perspective height	Arrangement
Point								
Linear								
Areal								

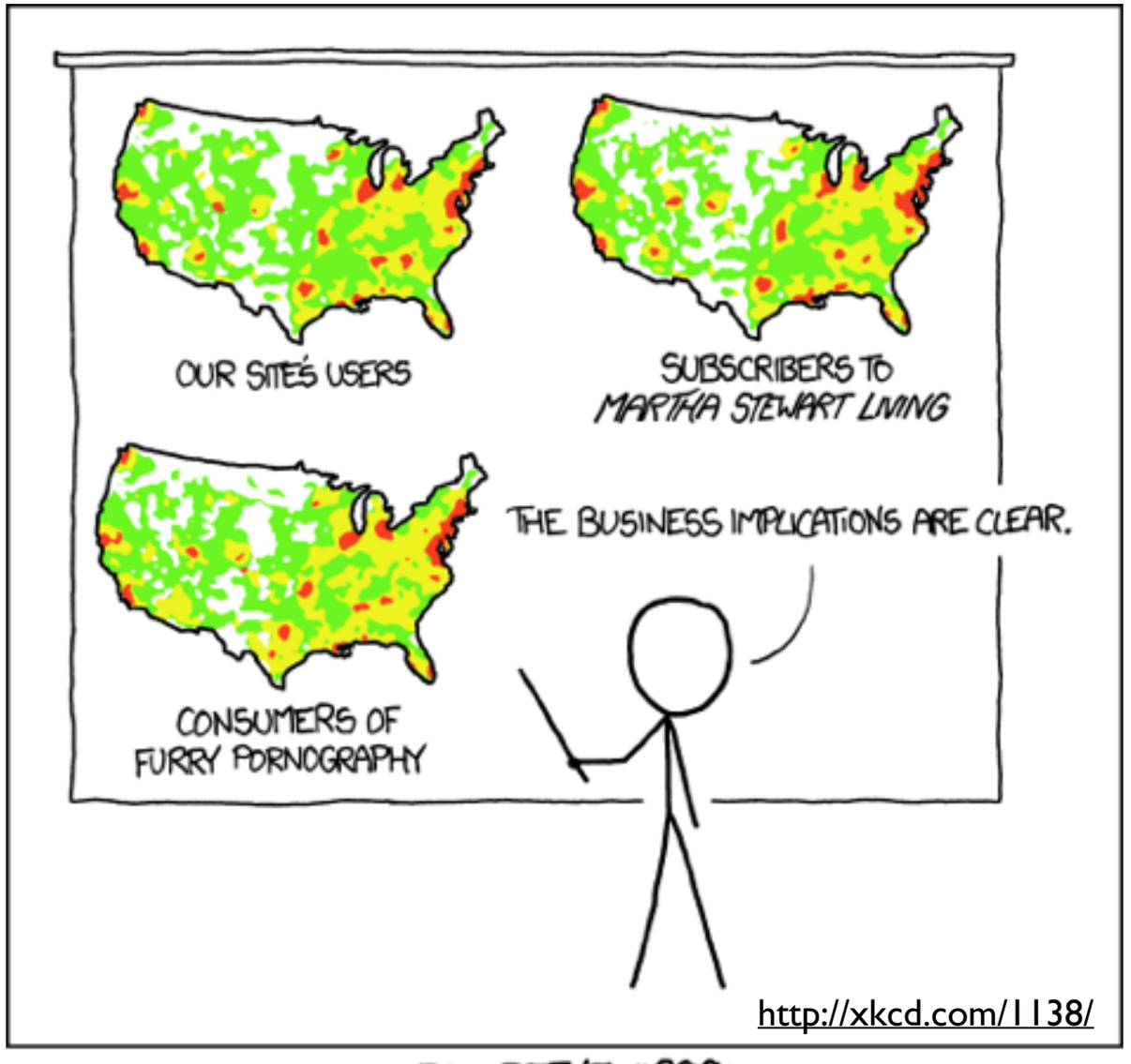
Slocum 1999

Dot Map

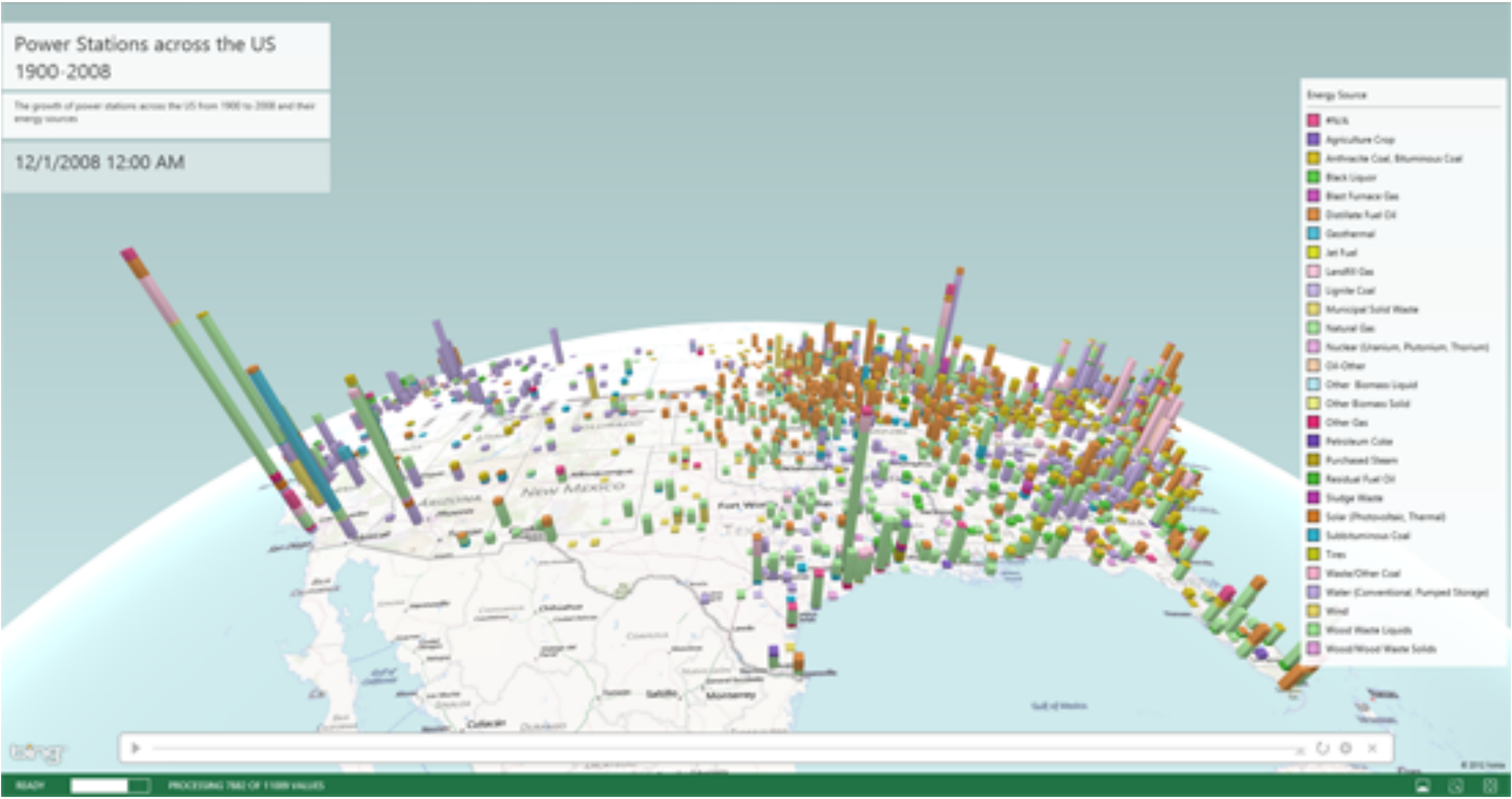
A Dot for Every Data Item

http://www.cairco.org/news/those-scary-dots-population-america

Maps can lie, too!



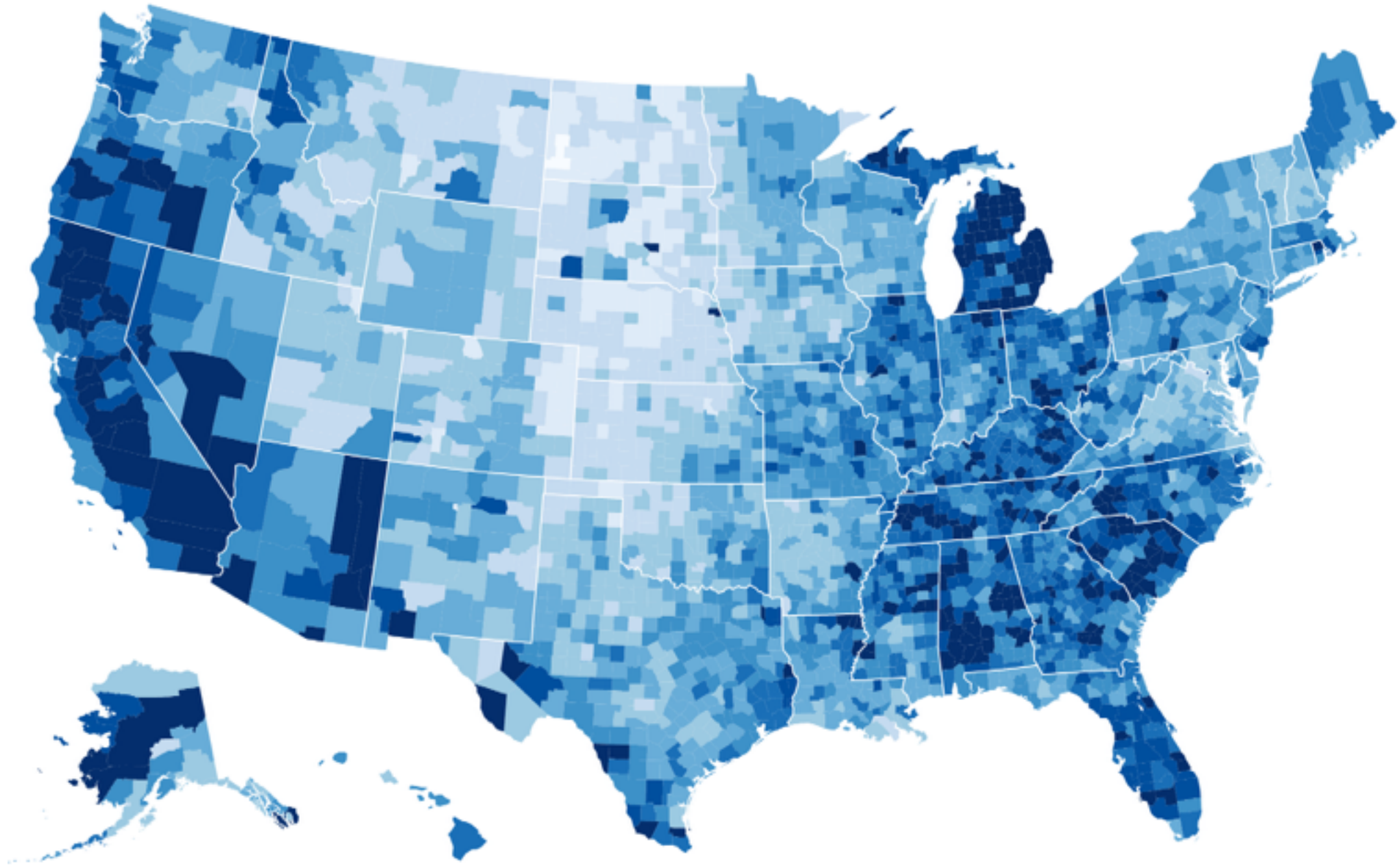
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS



Microsoft GeoFlow – Part of Excel 2013

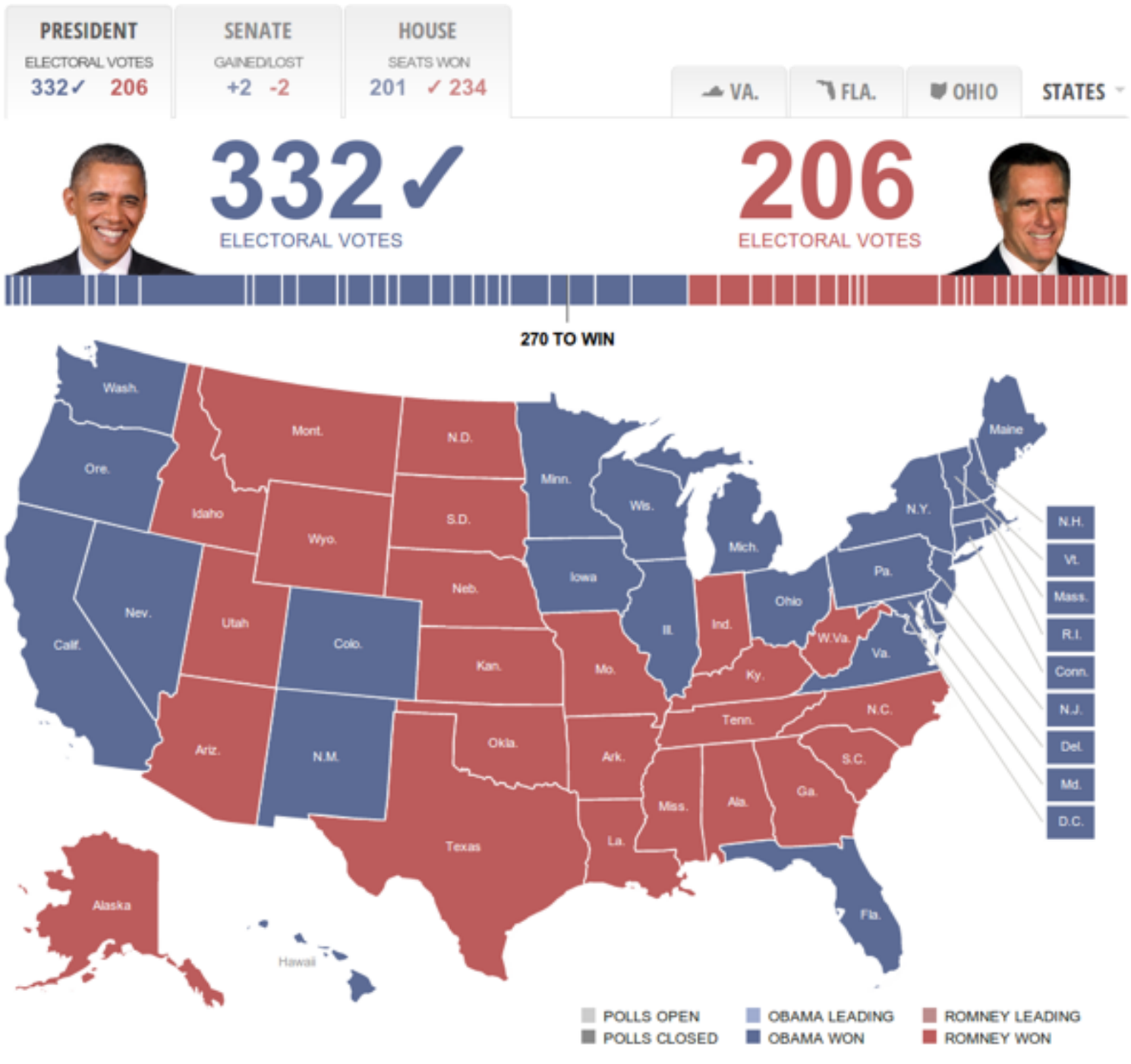
Choropleth Map

Attribute uniformly distributed in region



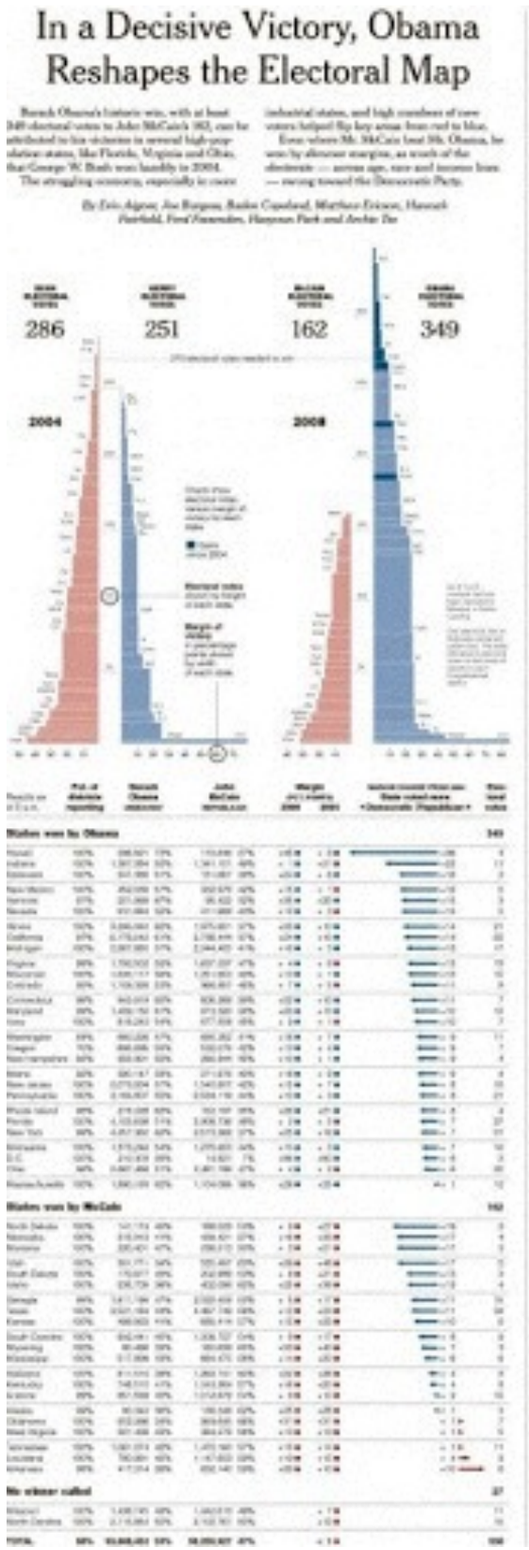
http://bl.ocks.org/mbostock/4060606

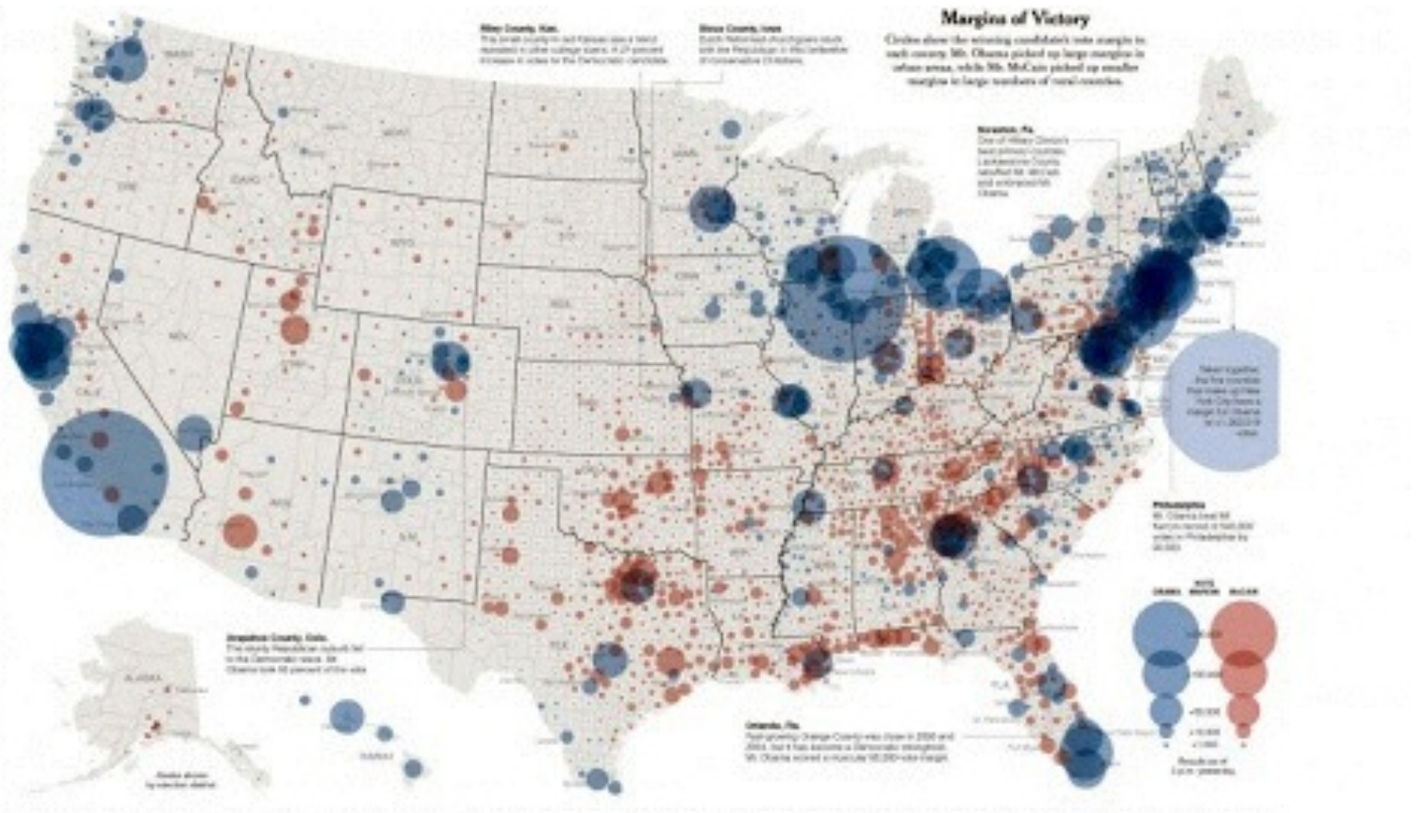
Misleading Coropleth Map

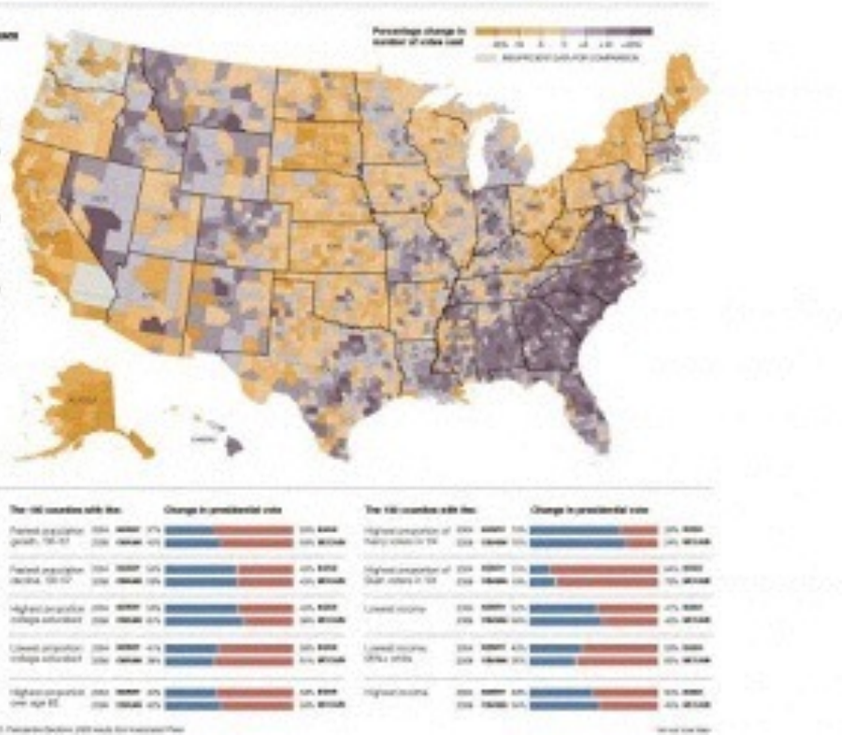


Huffington Post, http://elections.huffingtonpost.com/2012/results

Better Version by NYT



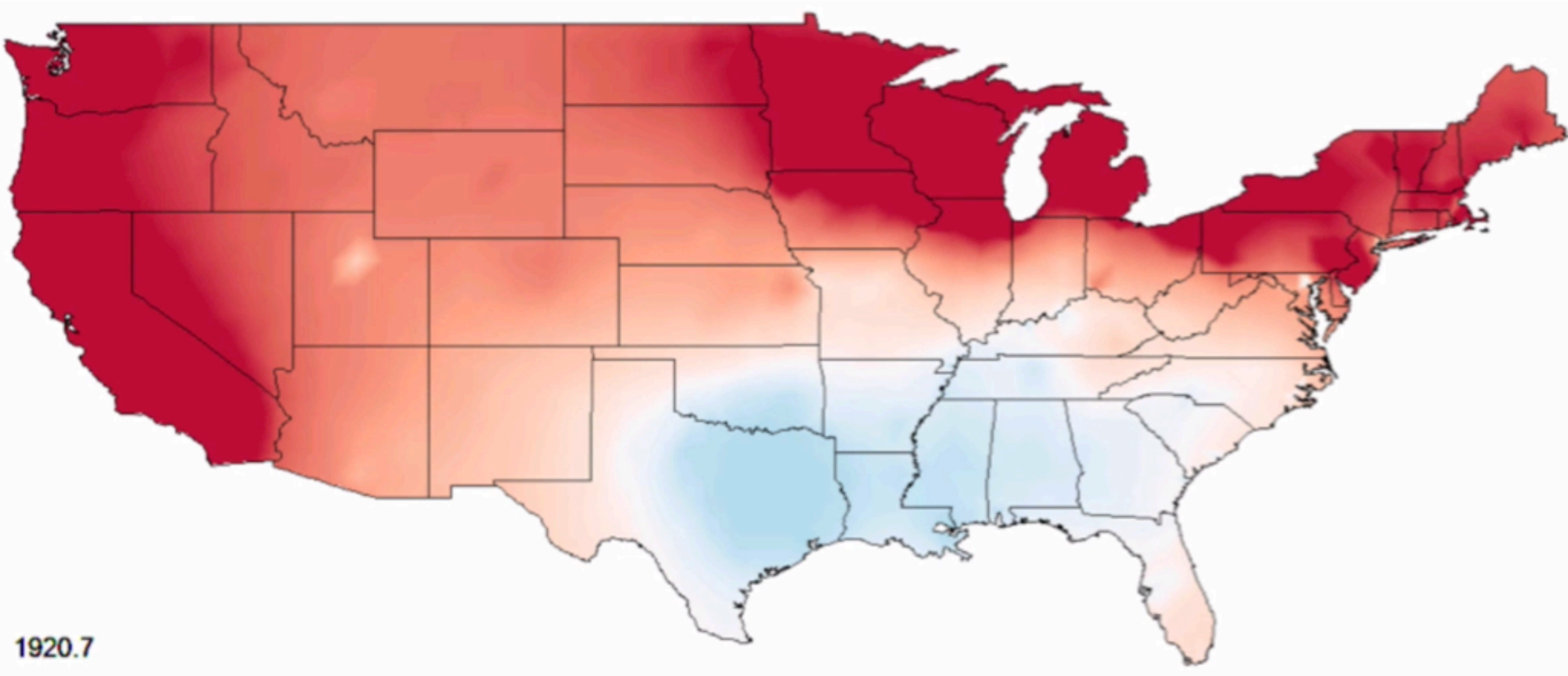




http://www.thefunctionalart.com/2012/11/when-choropleth-maps-deceive.html

Isarithmic Map

Color coding continuous phenomena

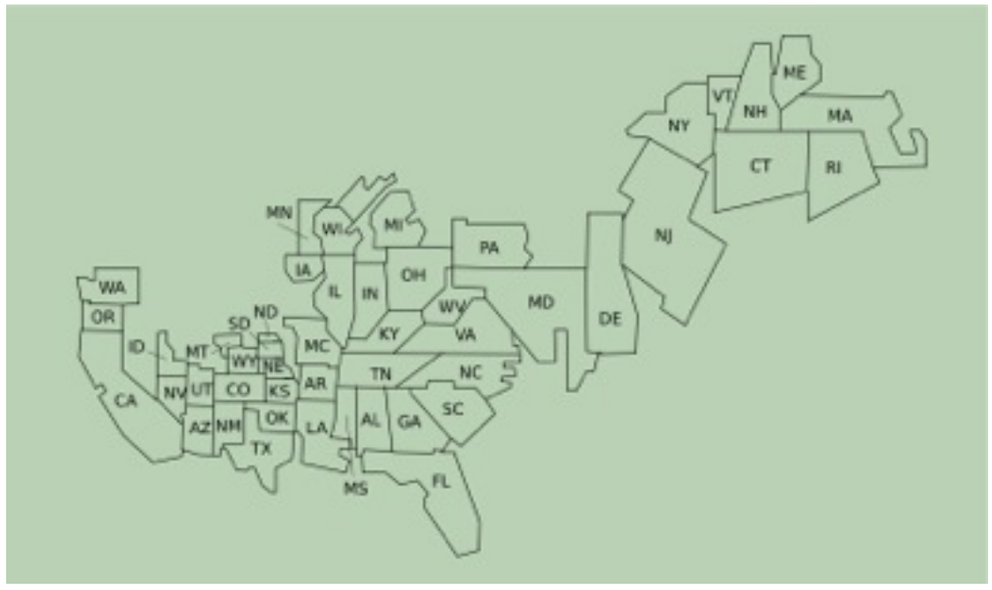


http://ecopolitology.org/2010/11/23/stunning-animated-time-lapse-map-of-red-blue-america-1920-2008/

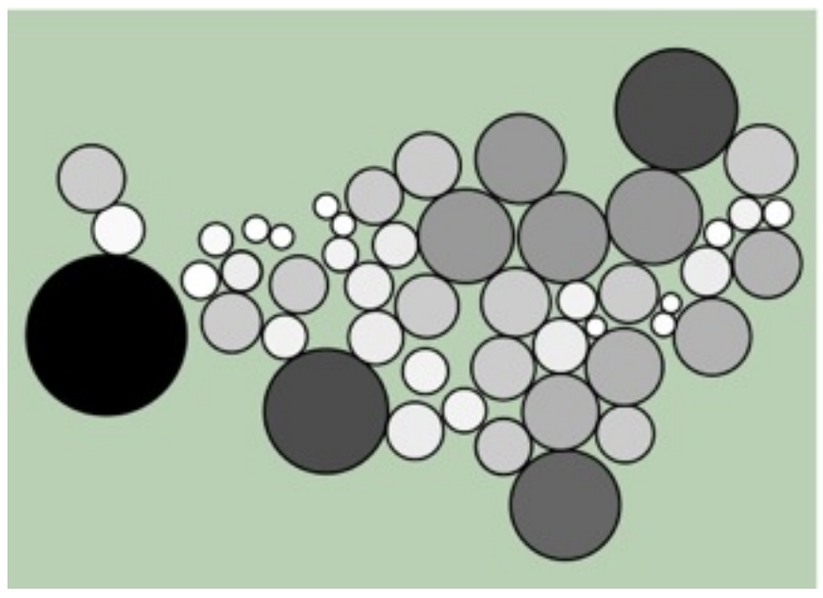
Cartograms

Size of region scaled to attribute value

Noncontinuous cartogram



Noncontiguous cartogram

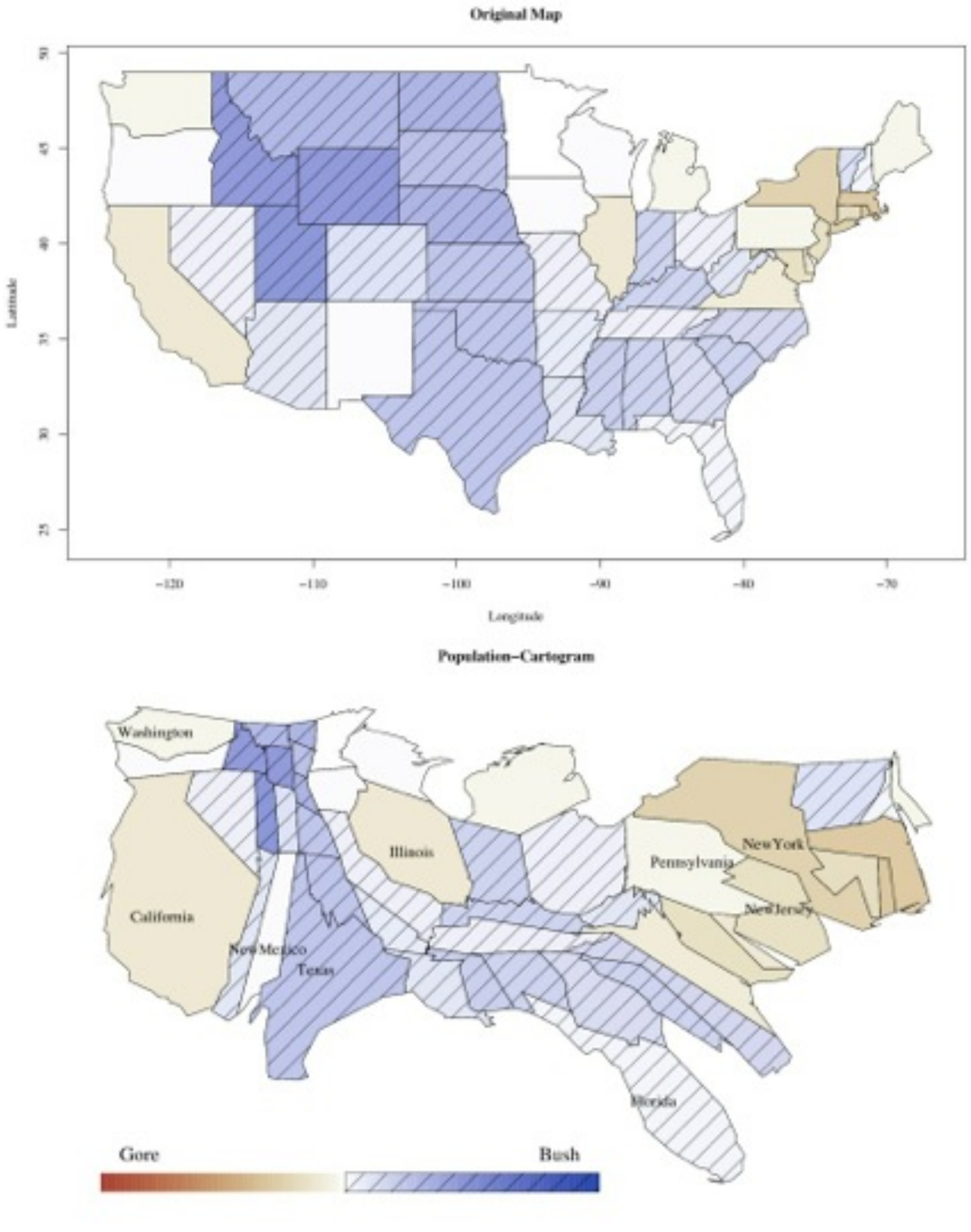


Circular cartogram

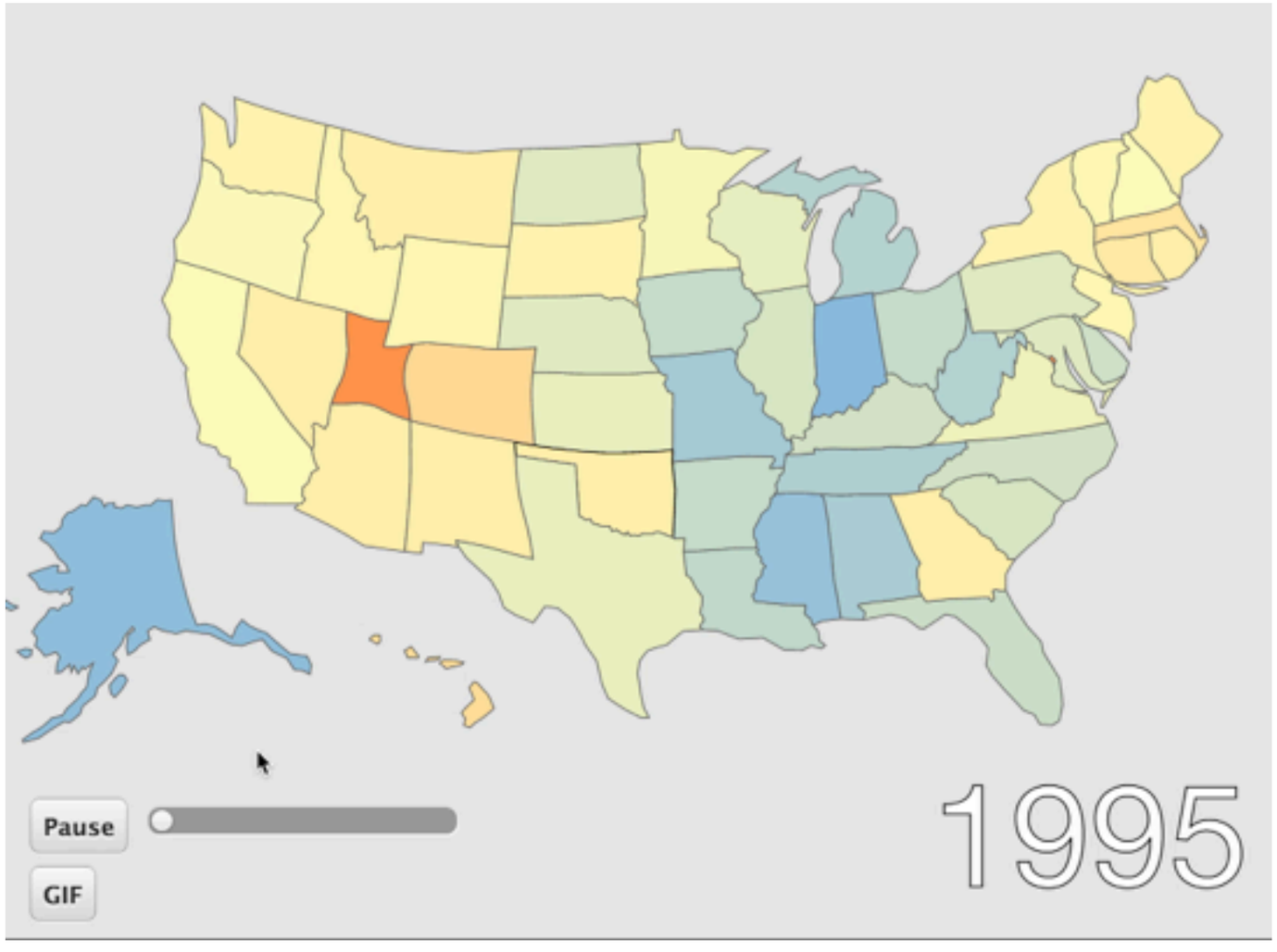
Continuous cartogram

Keim 2005

Continuous Cartogram Example

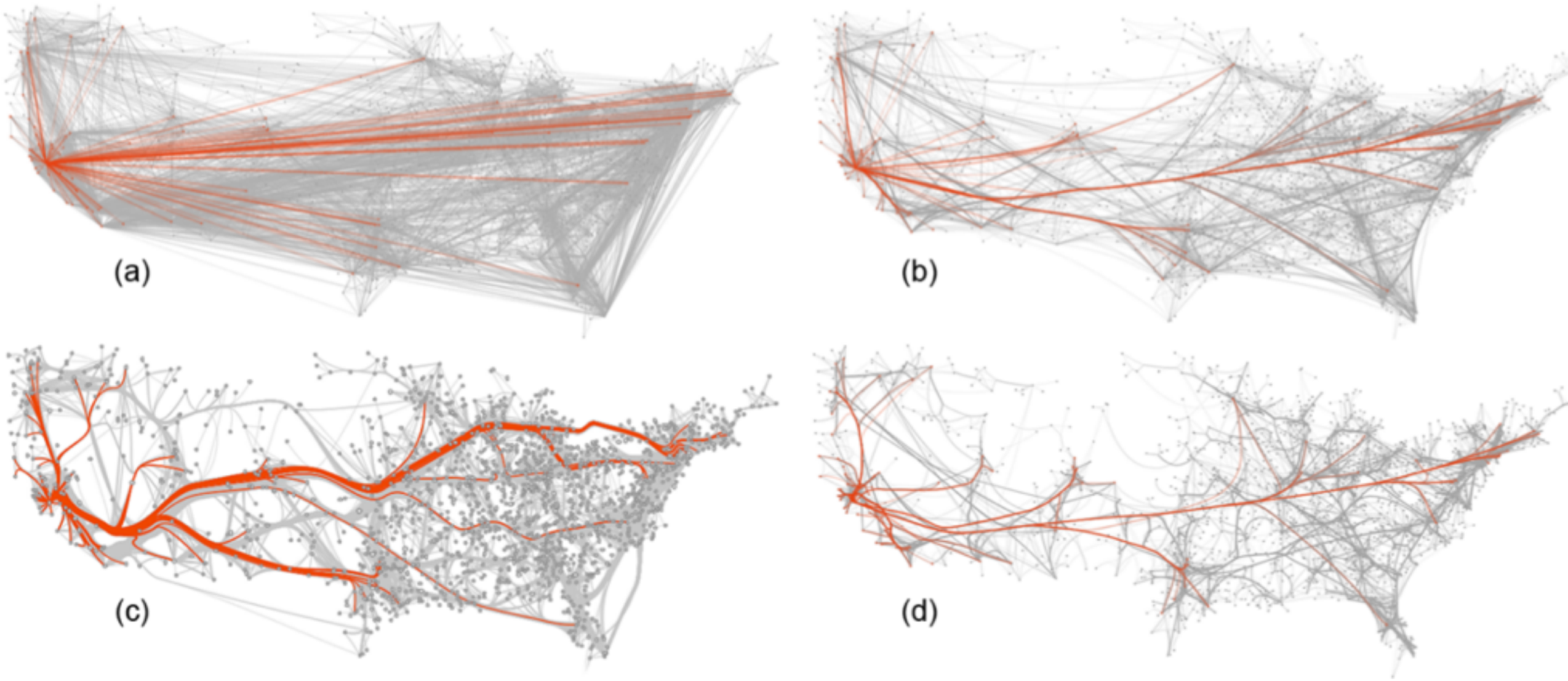


Keim 2003



http://tributary.io/inlet/5651806

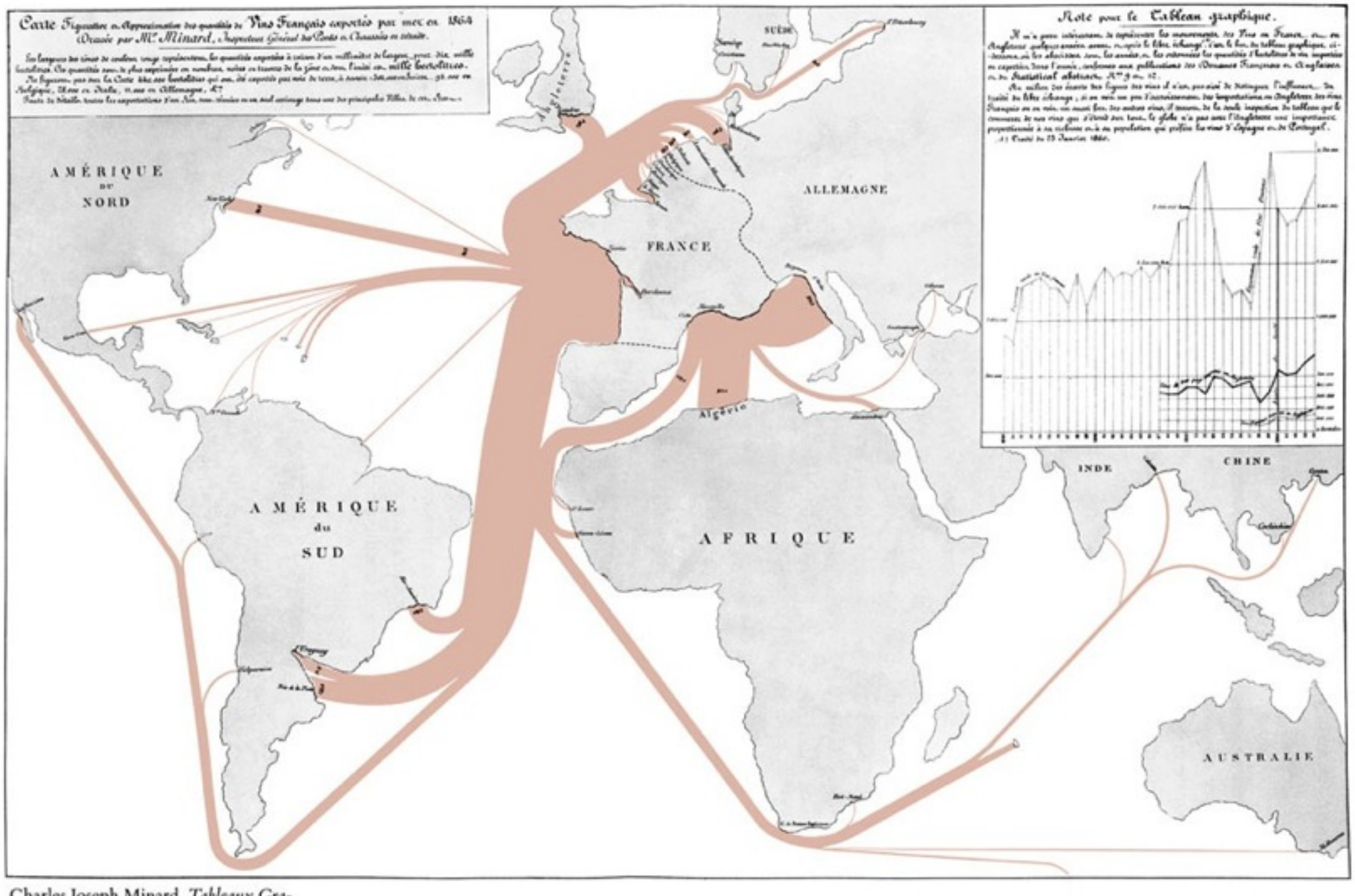
Edges Augmented Onto Map



Holten 2009

Force Directed Edge Bundles
US migration graph (1715 nodes, 9780 edges)

Flow Maps



Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard*, 1845–1869, a portfolio of his work held by the Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris.

Text Visualization

$59,413,405,476,974

The outstanding United States public debt as of Sept 15, 2014.

Tag Cloud / Word Cloud / Wordle

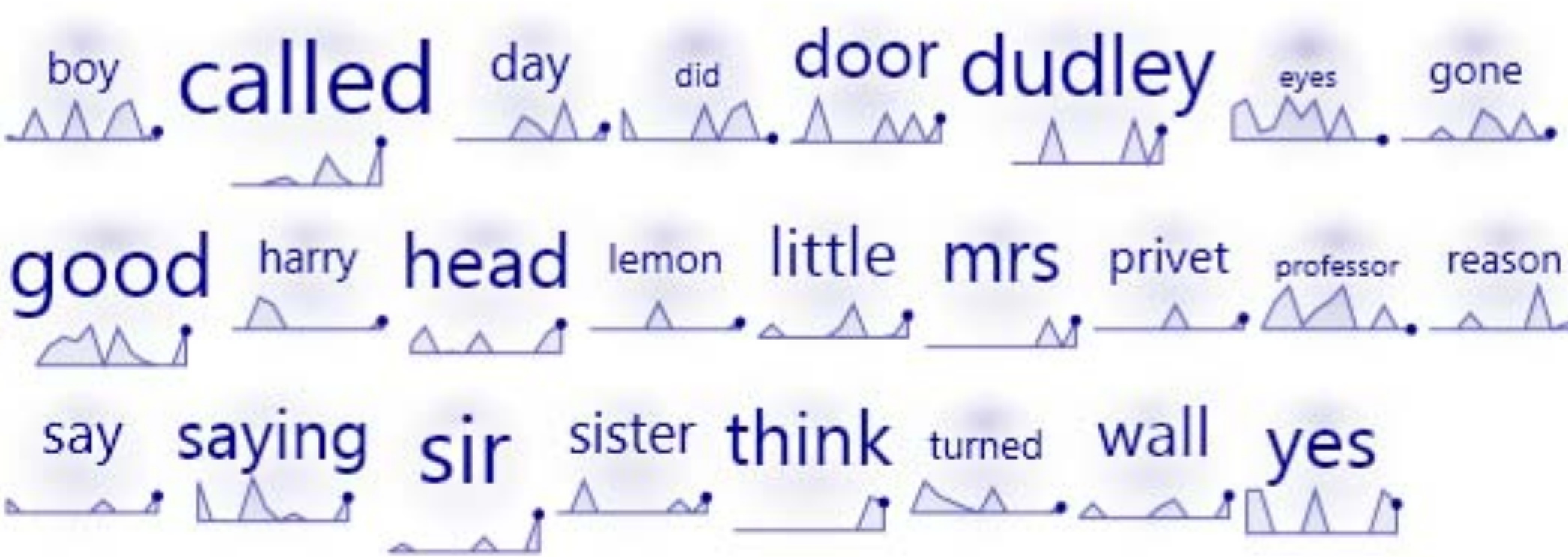
Change word size/color by frequency

http://www.jasondavies.com/wordcloud/

Spark Clouds

Lee 2010

Convey trends between multiple tag clouds over time



http://research.microsoft.com/en-us/um/redmond/groups/cue/infovis/

Example: Phrase Net

van Ham et al. 2009

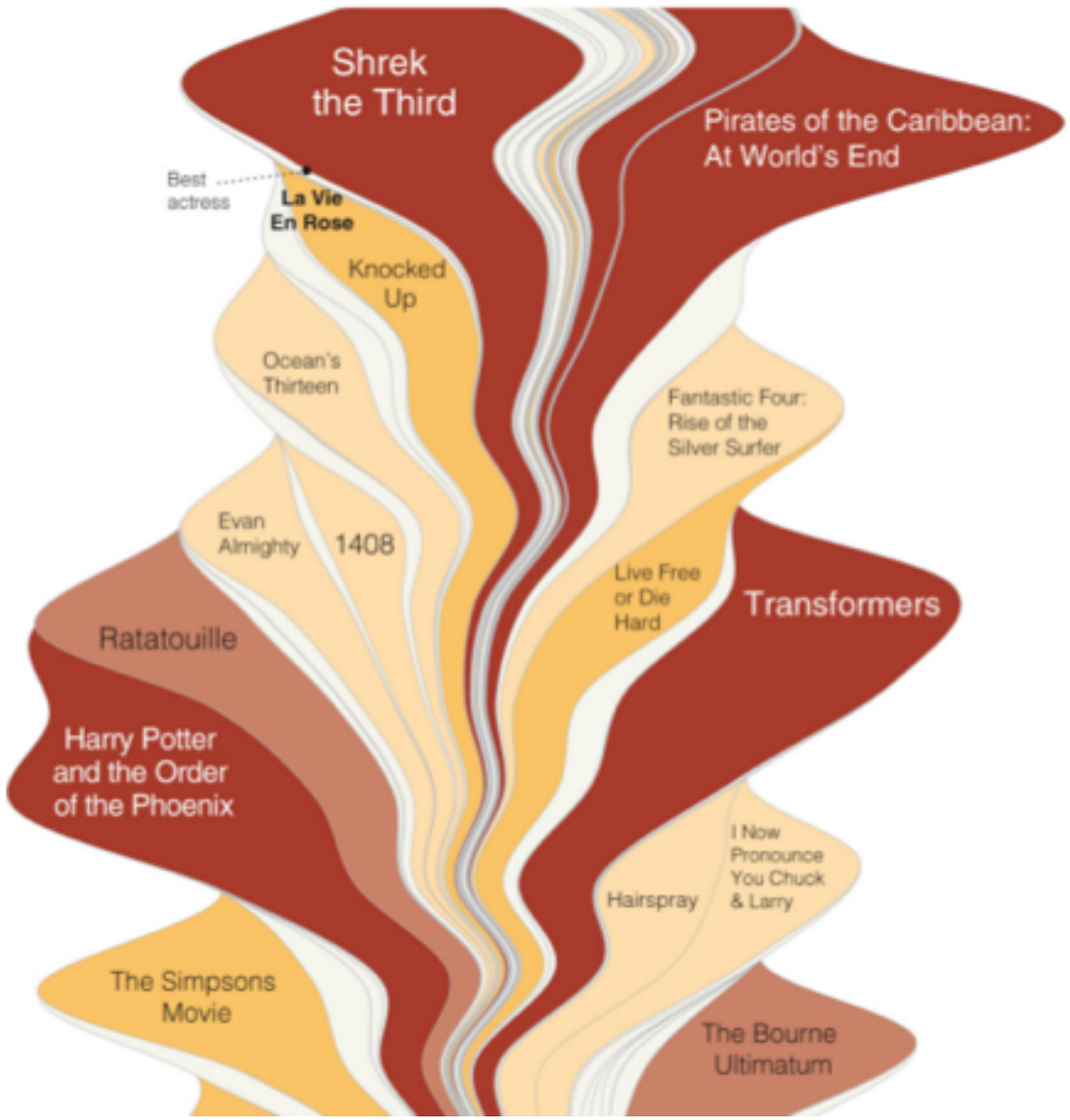
Visual overviews of unstructured text

Graph; nodes = words; edges = linked by user specified relation

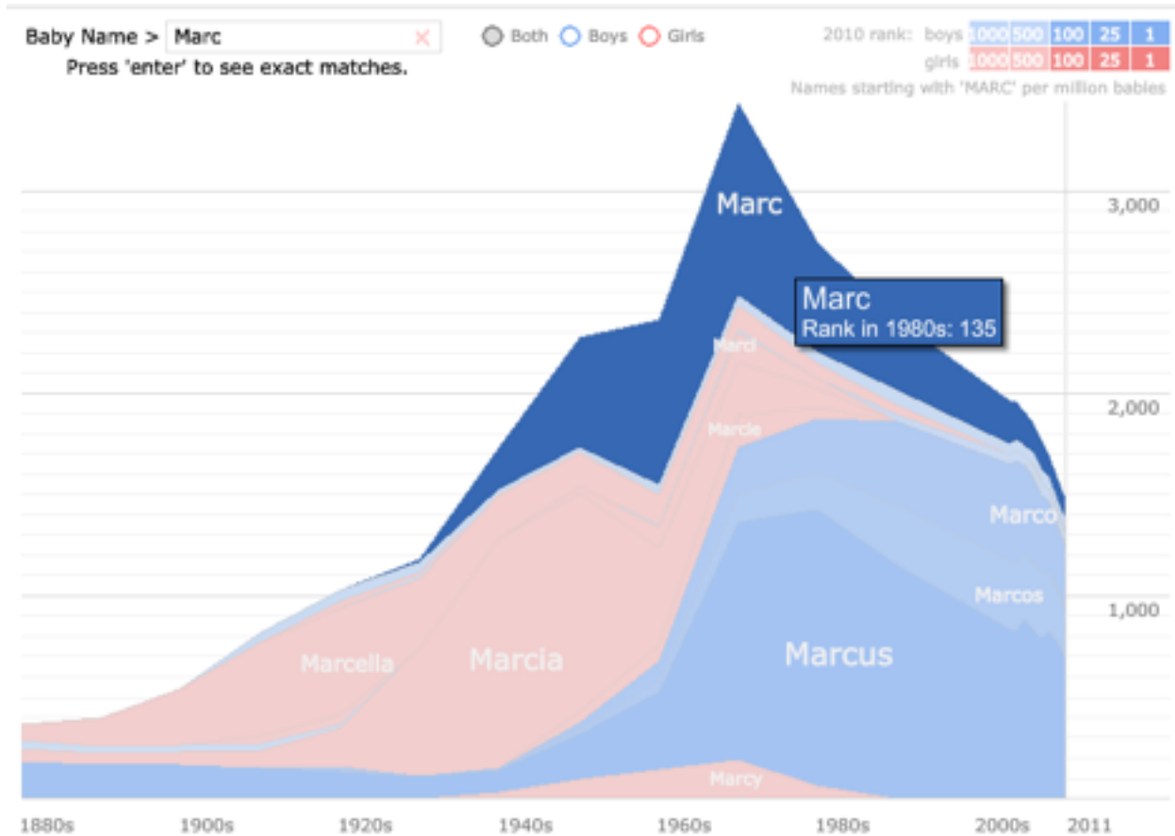
Relations in 18th and 19th century novels

Theme River / Stream Graph

Thematic changes over time. Height = frequency

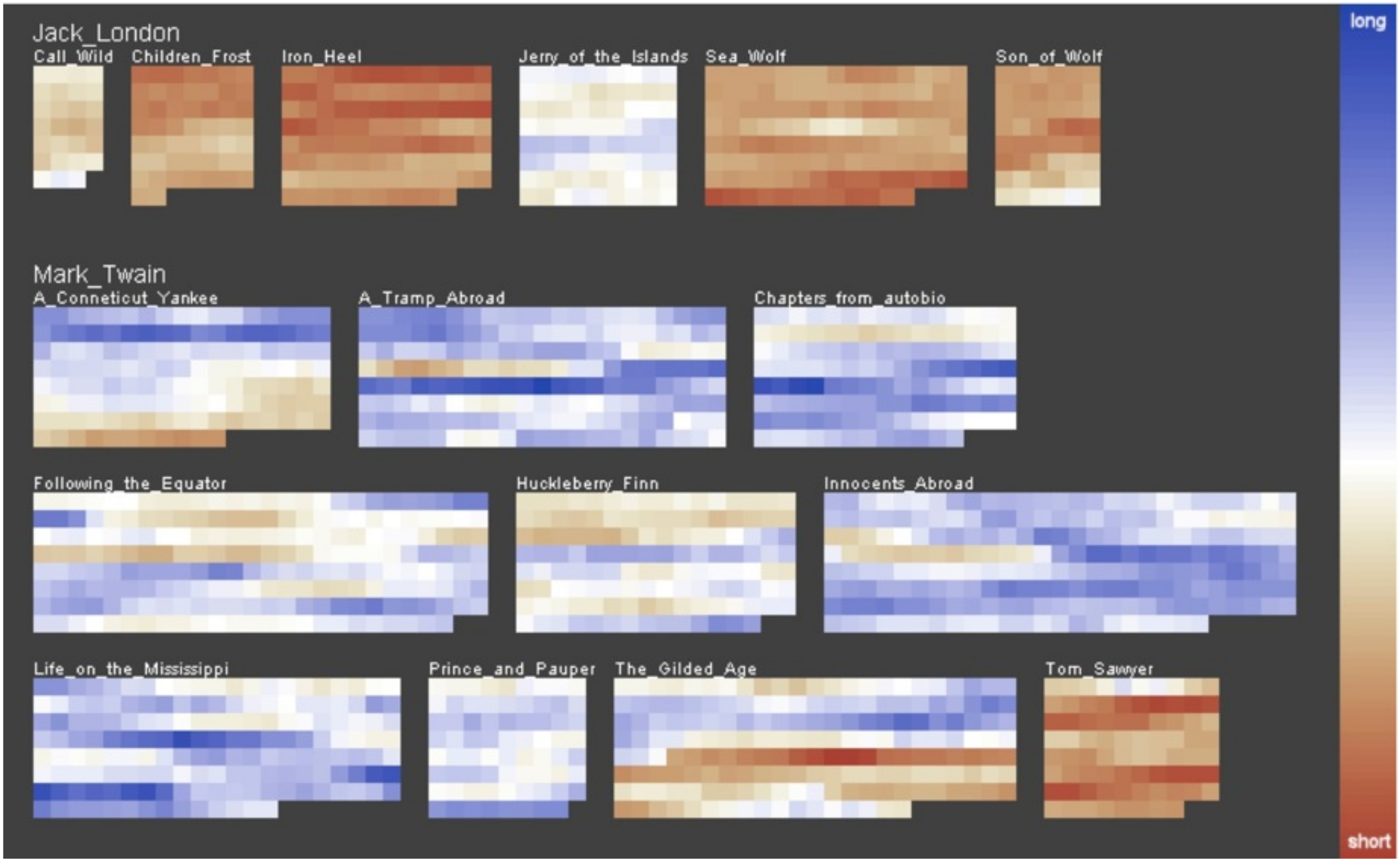


StreamGraph: Films from summer 2007
[Byron & Wattenberg 2008]



Baby Name Voyager
http://www.babynamewizard.com/
Wattenberg 2005

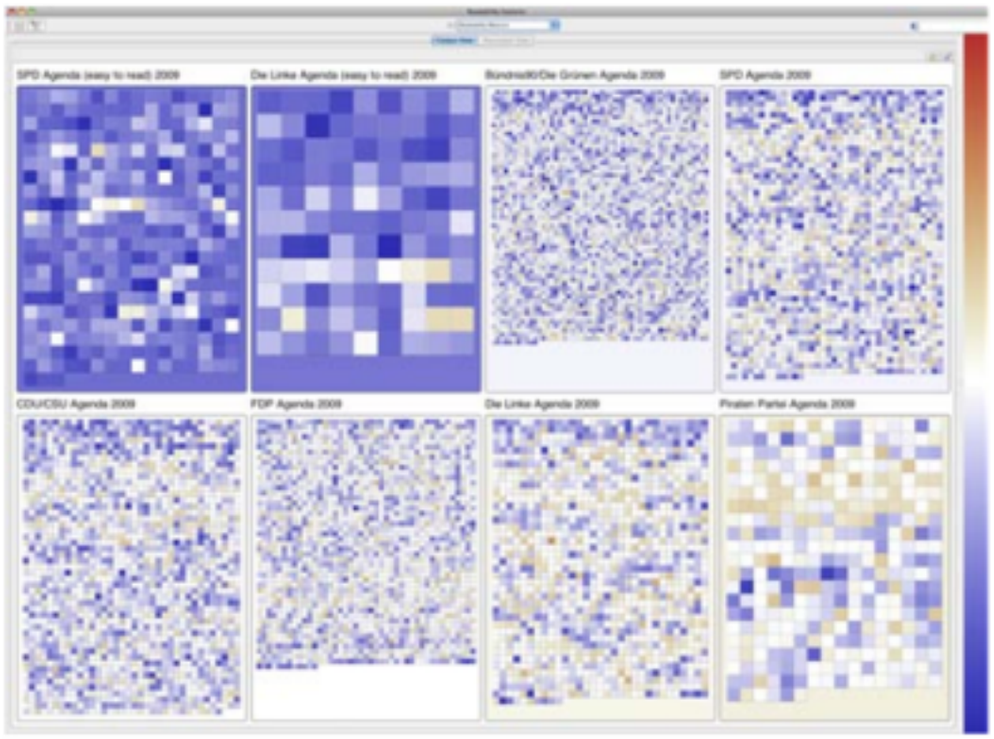
Example: Literature Fingerprinting



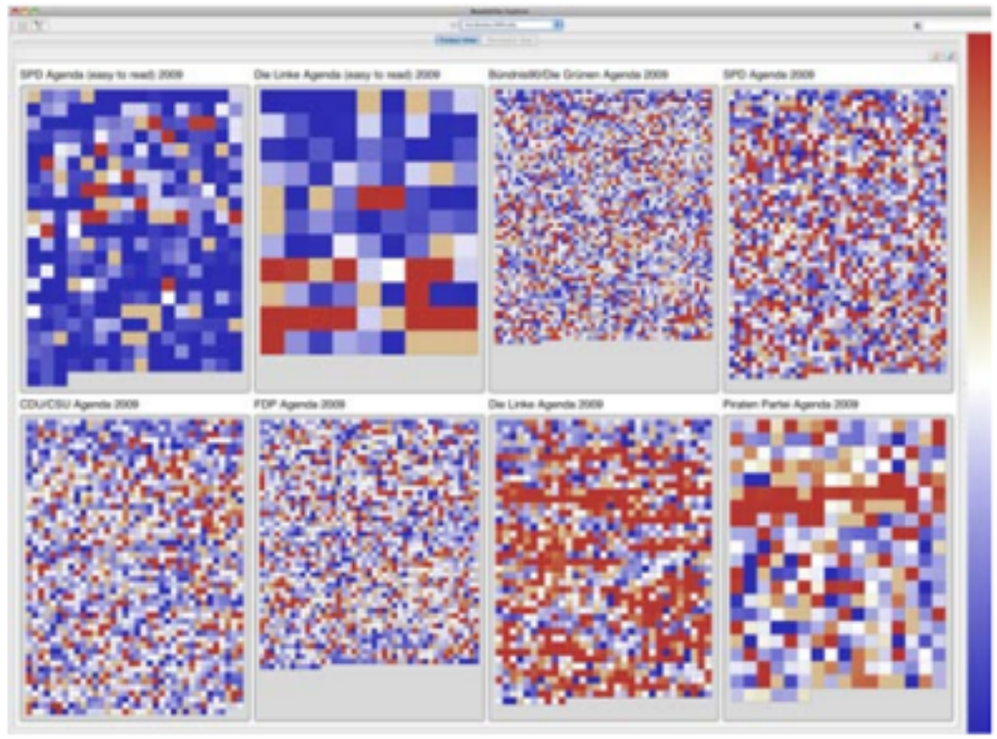
Visualize several text measures to discriminate between authors.
Pixel = text block, Group = book, color = average sentence length

Keim and Oelke 2007

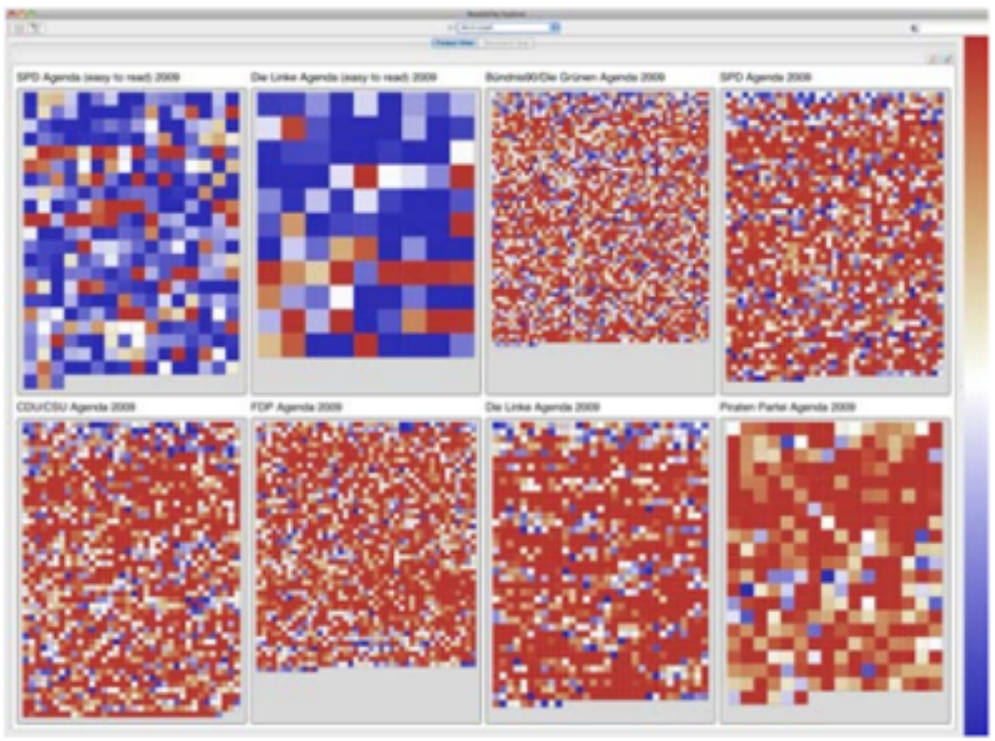
Example: Readability Analysis



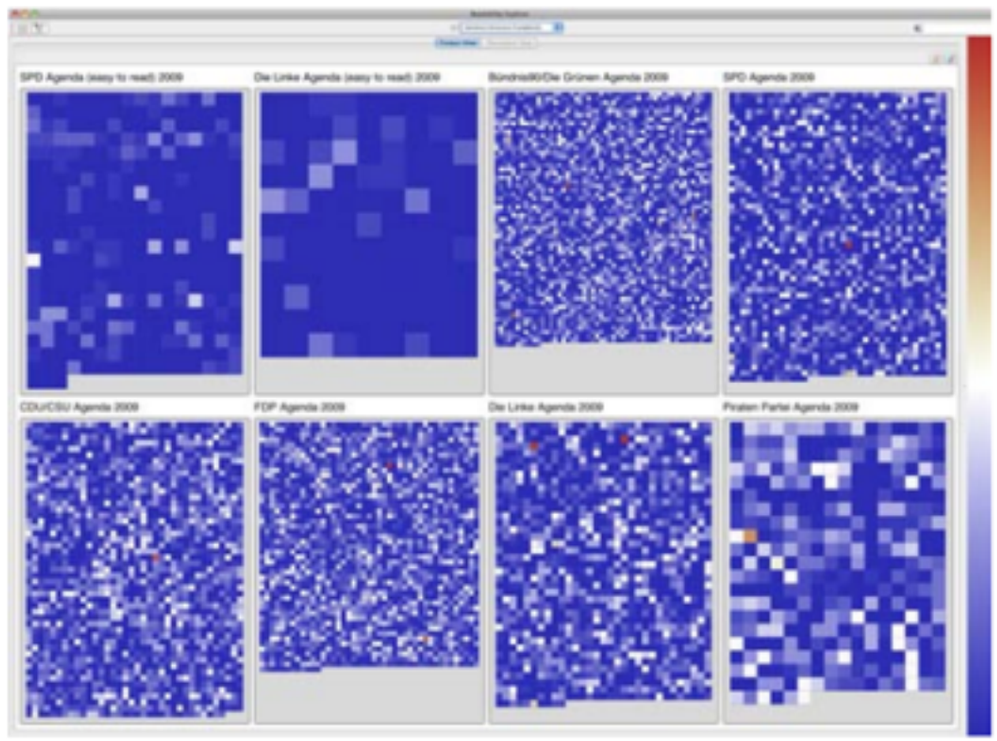
(a) Average Readability Score



(b) Feature: Vocabulary Difficulty



(c) Feature: Word Length



(d) Feature: Sentence Structure Complexity

Oelke 2007

WordsEye

Automatic Text-to-Scene Conversion System

The lawn mower is 5 feet tall. John pushes the lawn mower. The cat is 5 feet behind John. The cat is 10 feet tall.

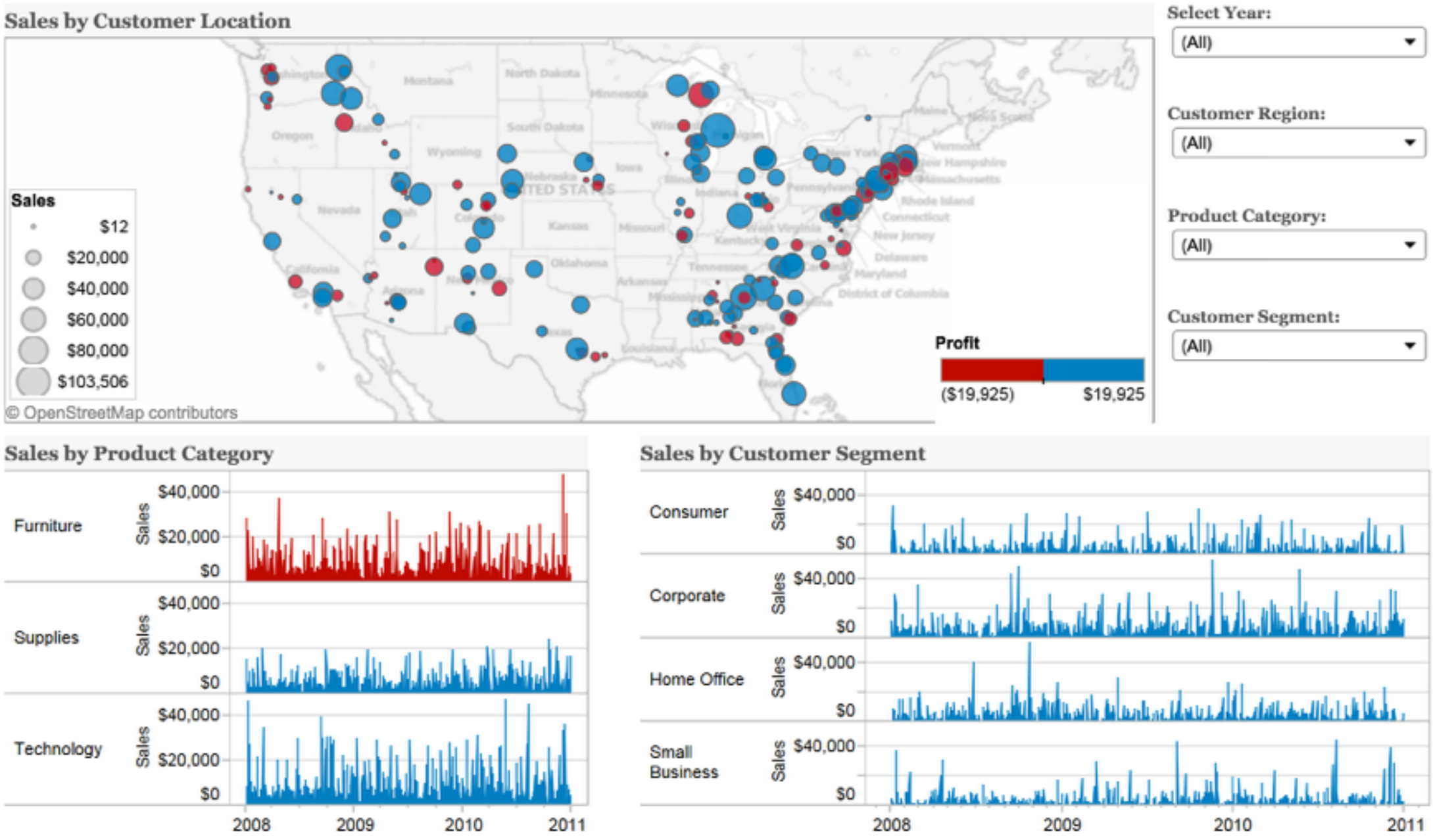
http://www.wordseye.com/

Coyne and Sproat 2001

Software

tableau SOFTWARE

http://www.tableausoftware.com/



Data-Driven Documents (D3.js)

http://d3js.org

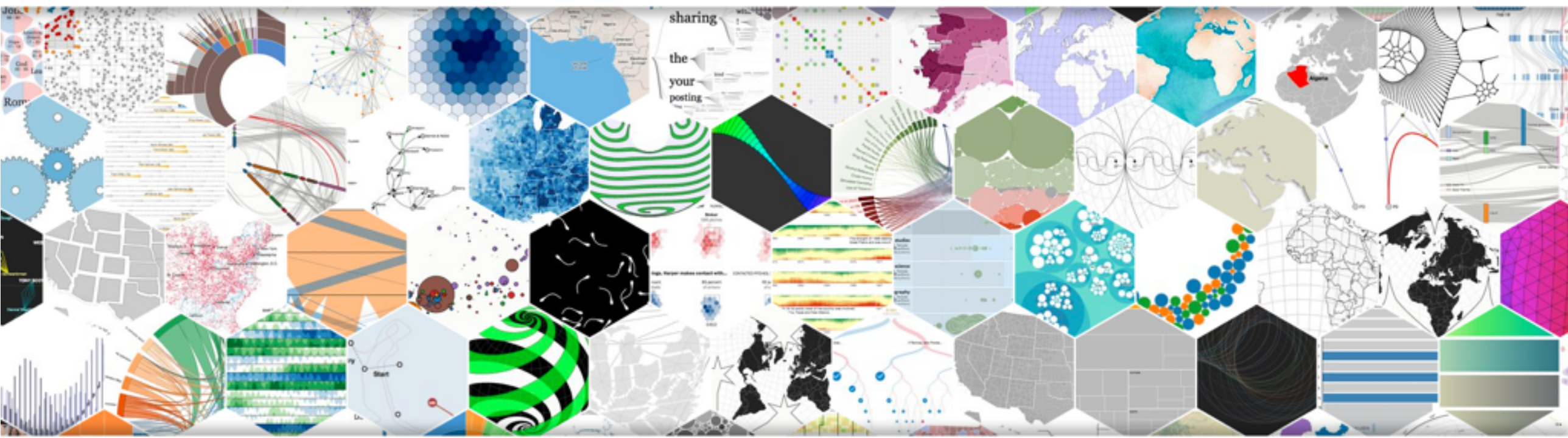
Java Script library

Technologies: CSS3, HTML5, SVG

Bind arbitrary data to a Document Object Model (DOM)

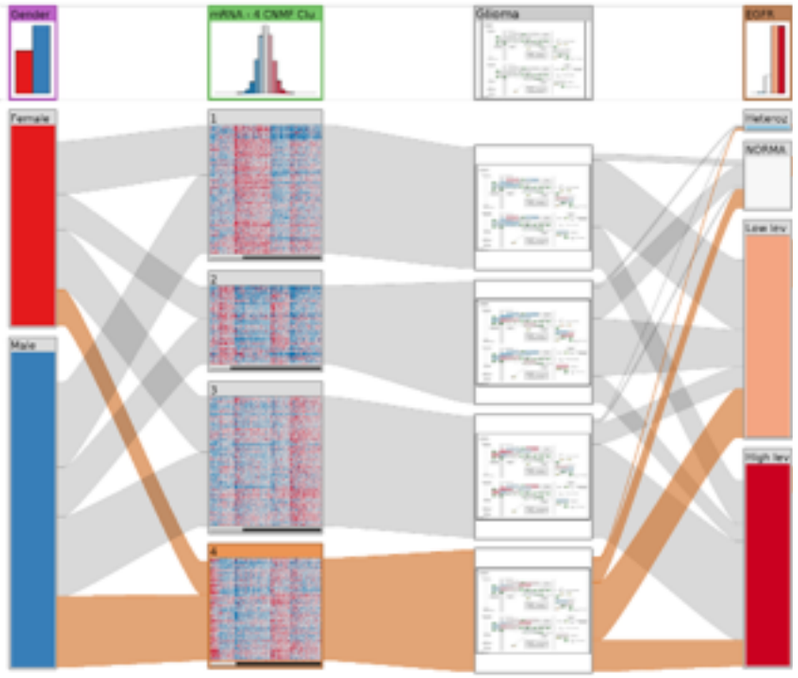
Then apply data-driven transformations to the document

Well documented / lots of examples & tutorials / free books

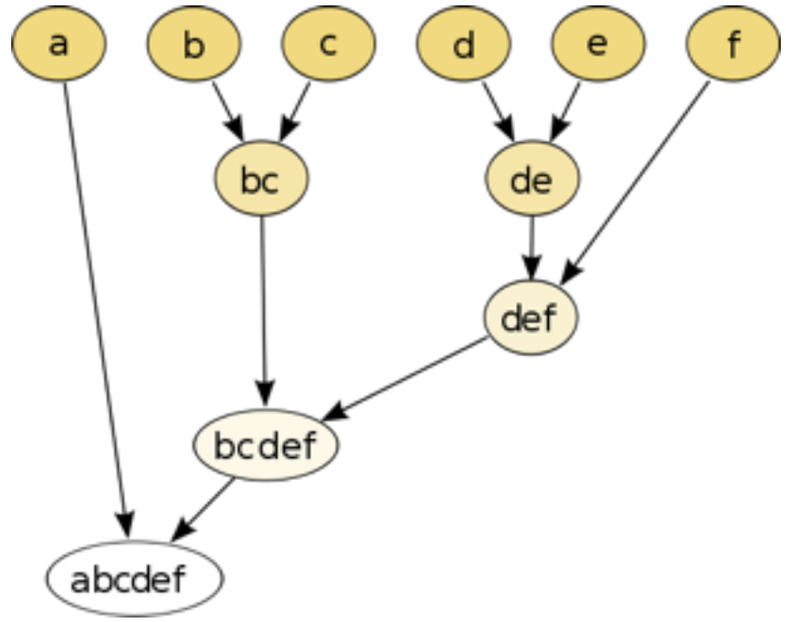


On Thursday

Big Data Visualization



Case Study



Clustering, Dimensionality Reduction

Questions