

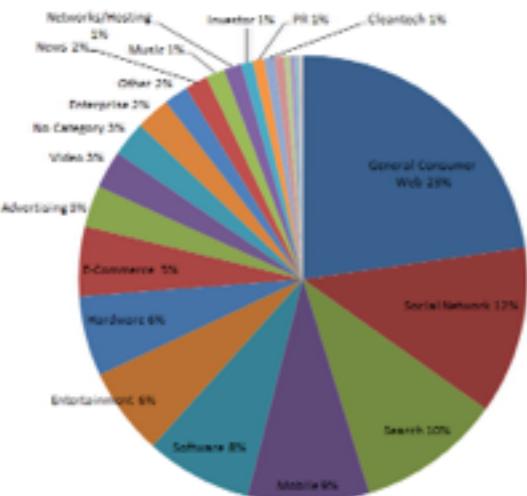
# CS 109: Data Science

## Visualization of Multi-Dim Data, Maps and Text

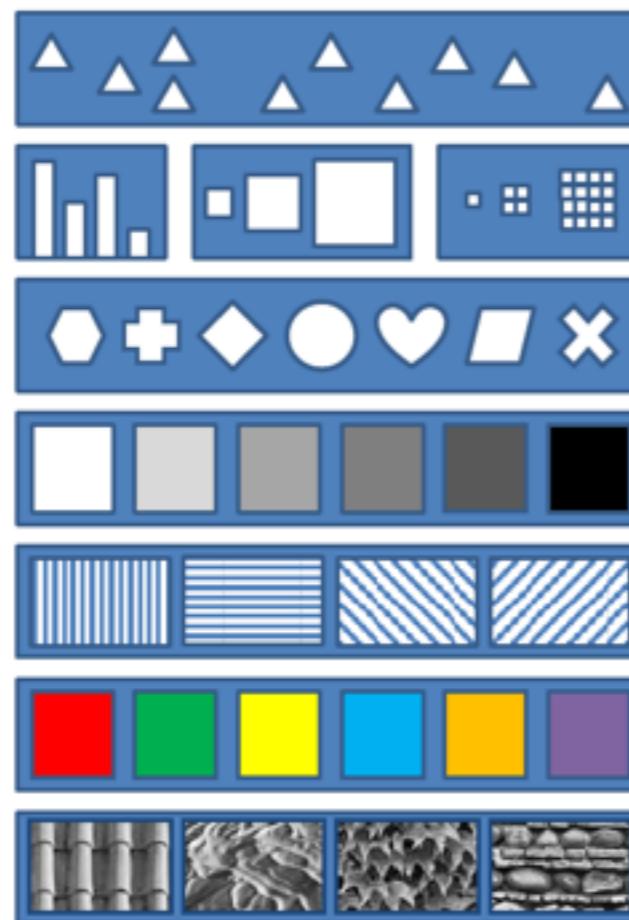
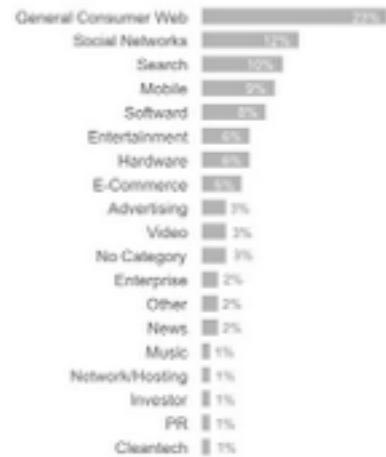
Marc Streit

[mstreit@seas.harvard.edu](mailto:mstreit@seas.harvard.edu)

# Last Week

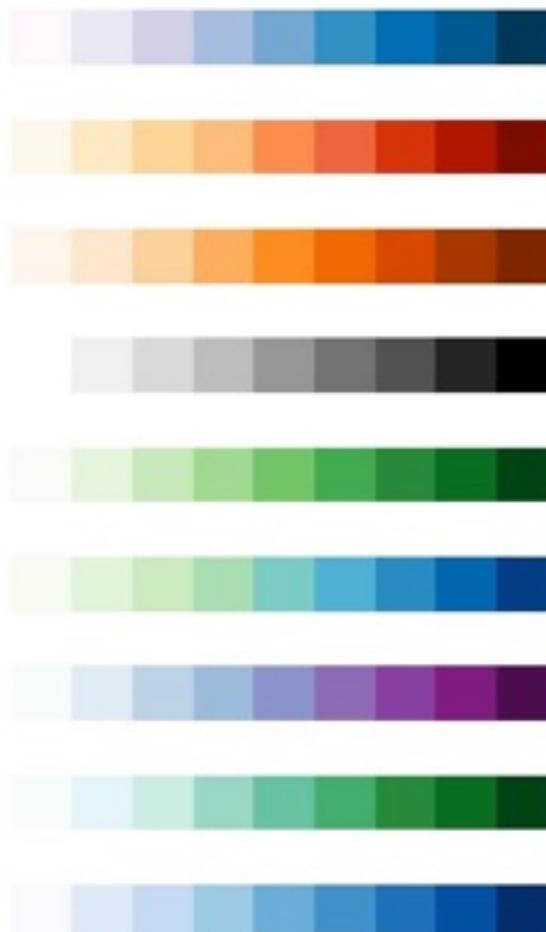


TechCrunch Coverage: 2005 - 2011  
Bars are best!

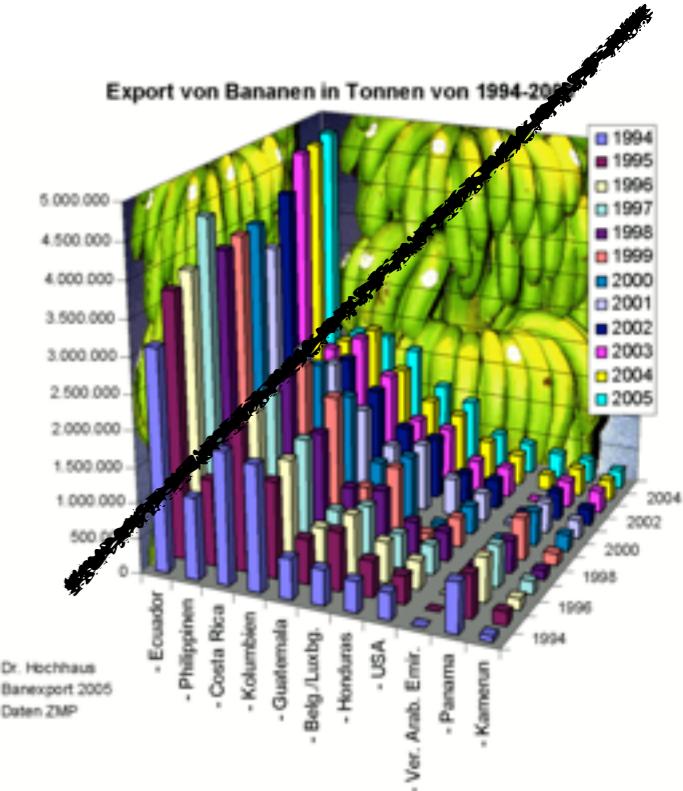


Effective vs.  
Non Effective  
Visualizations

Visual  
Attributes

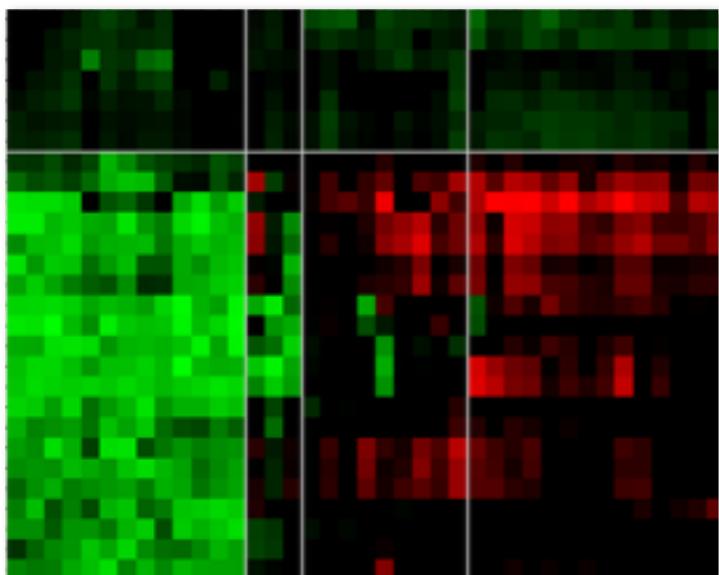
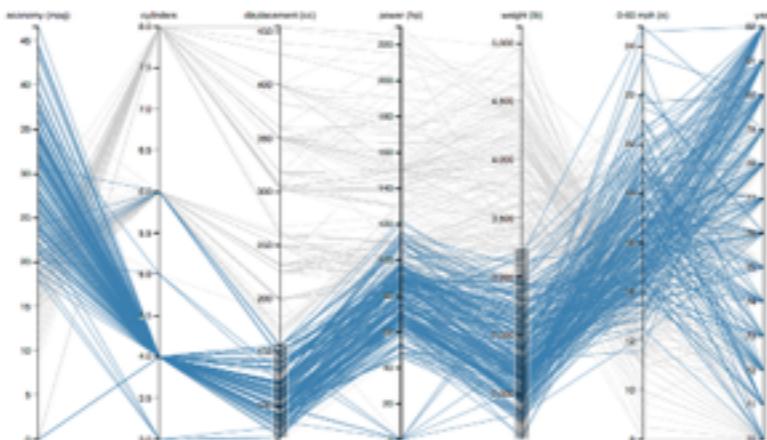


Color

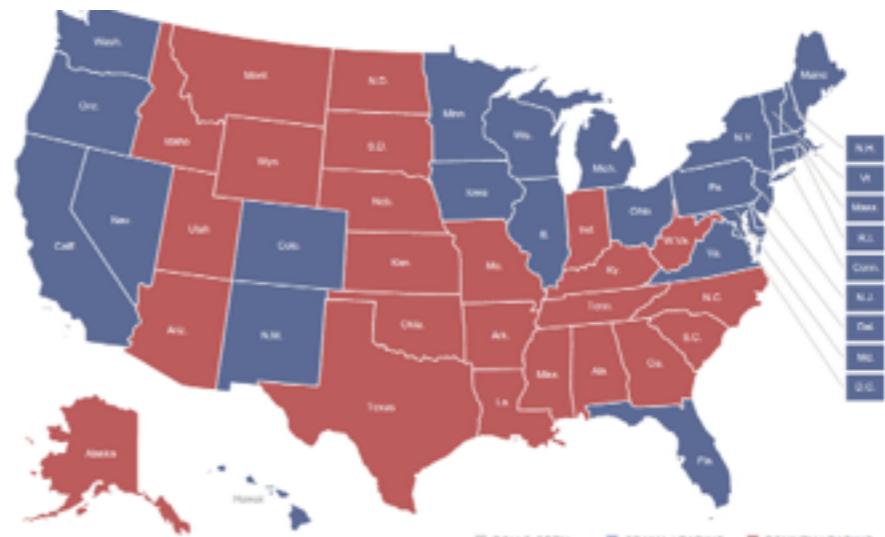


Design  
Principles

# Today



Multi-Dimensional  
Data Visualization



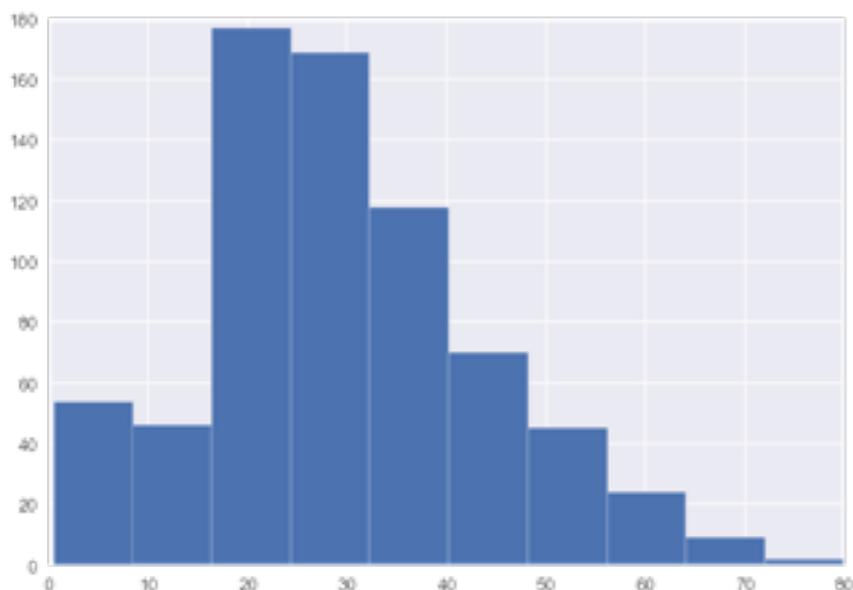
Map & Text  
Visualization

# **Multi-Dimensional Data Visualization**

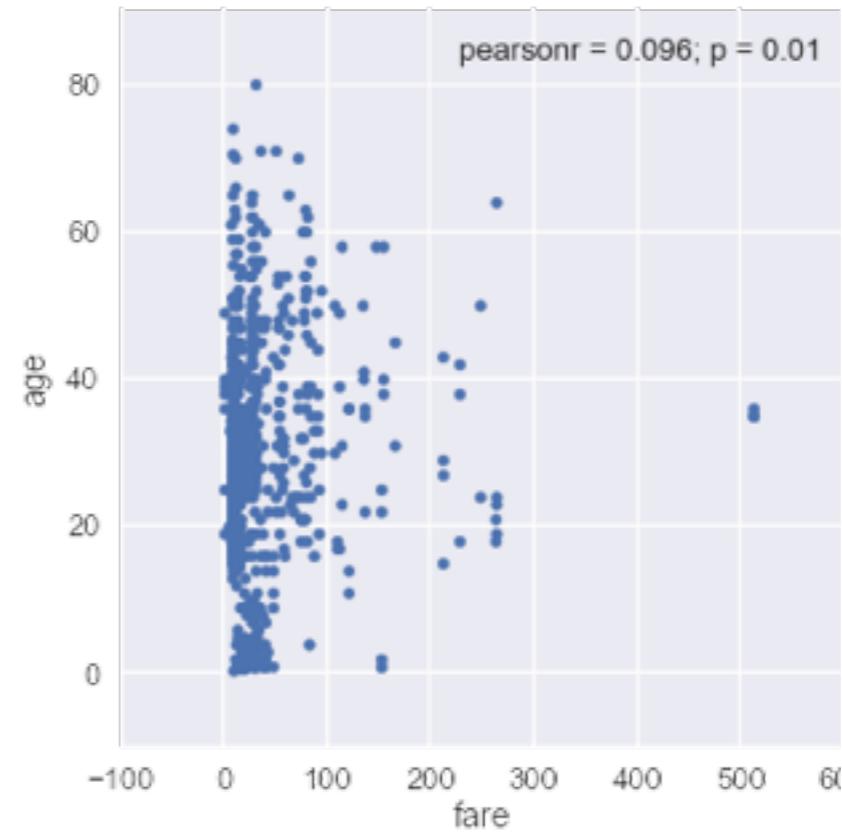
survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	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

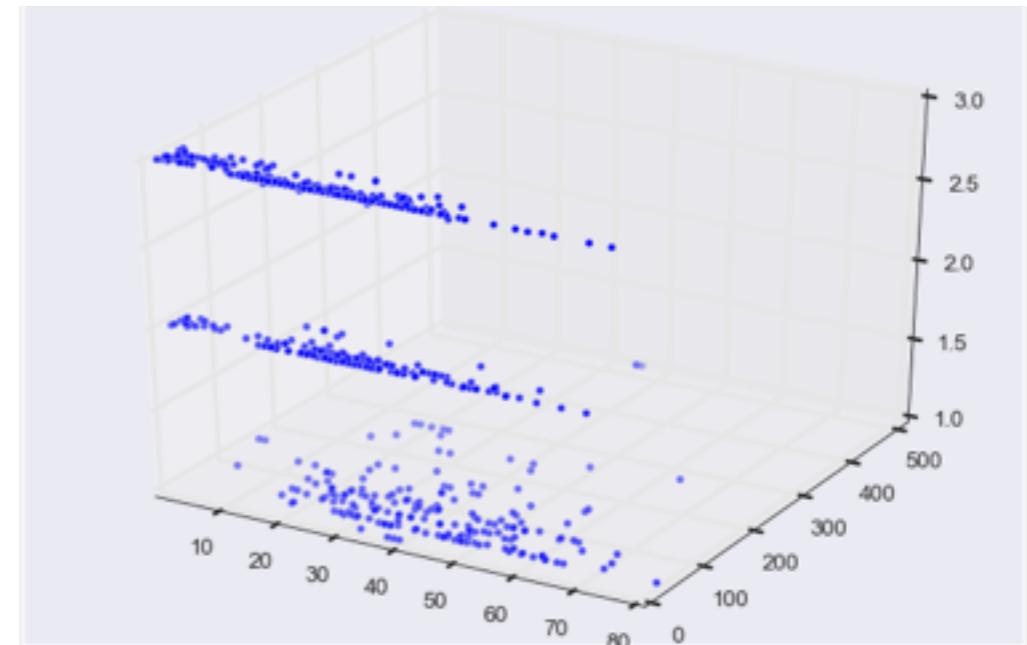
# ID



# 2D



# 4D?



# 3D

# What is “high” dimensional?

How many dimensions (attributes)?

- ~50 – tractable with “just” vis
- ~1,000 – need analytical methods

How many items?

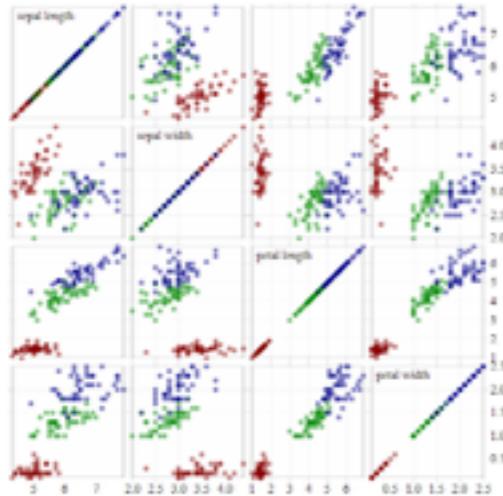
- ~ 1,000 – “just” vis is fine
- >> 10,000 – need analytical methods

## Homogeneity

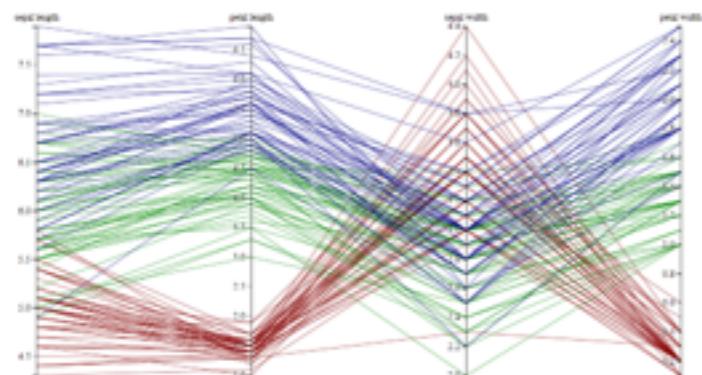
Same data type?

Same scales?

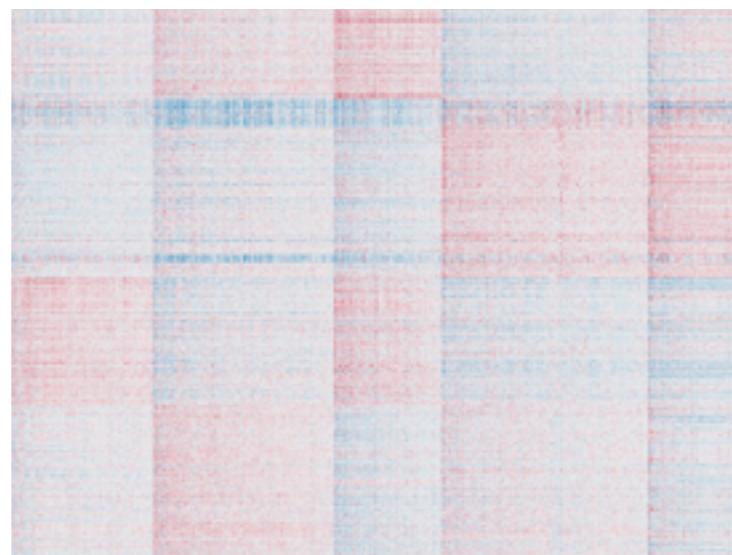
# Analytic Component



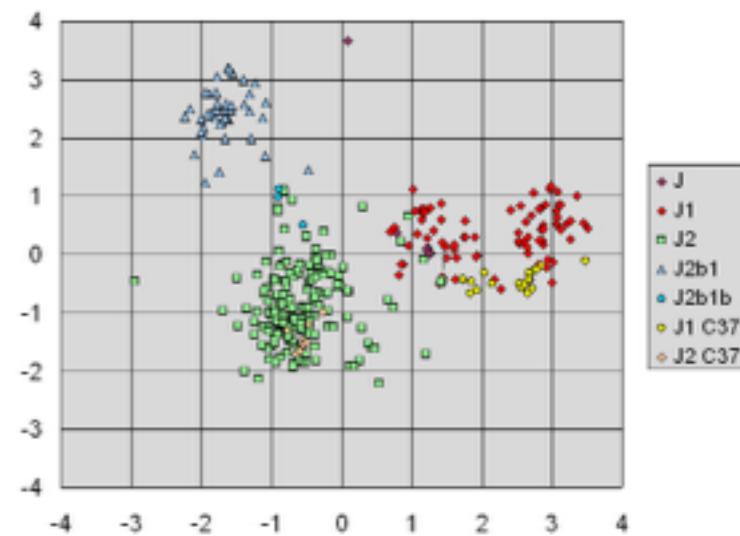
Scatterplot Matrices



Parallel Coordinates



Pixel-based Visualizations /  
Heat Maps



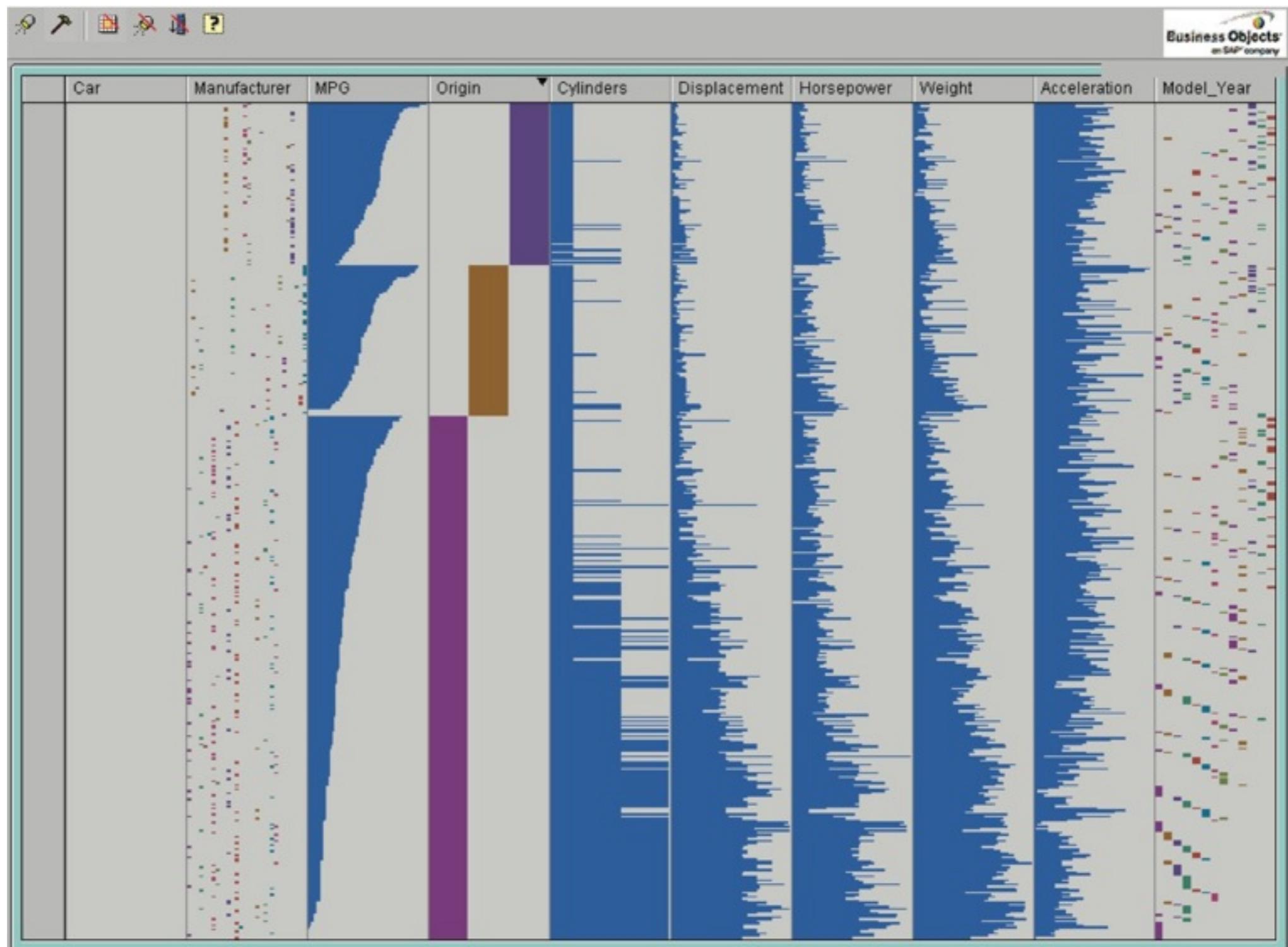
Dimensionality  
Reduction  
(e.g., [PCA](#))



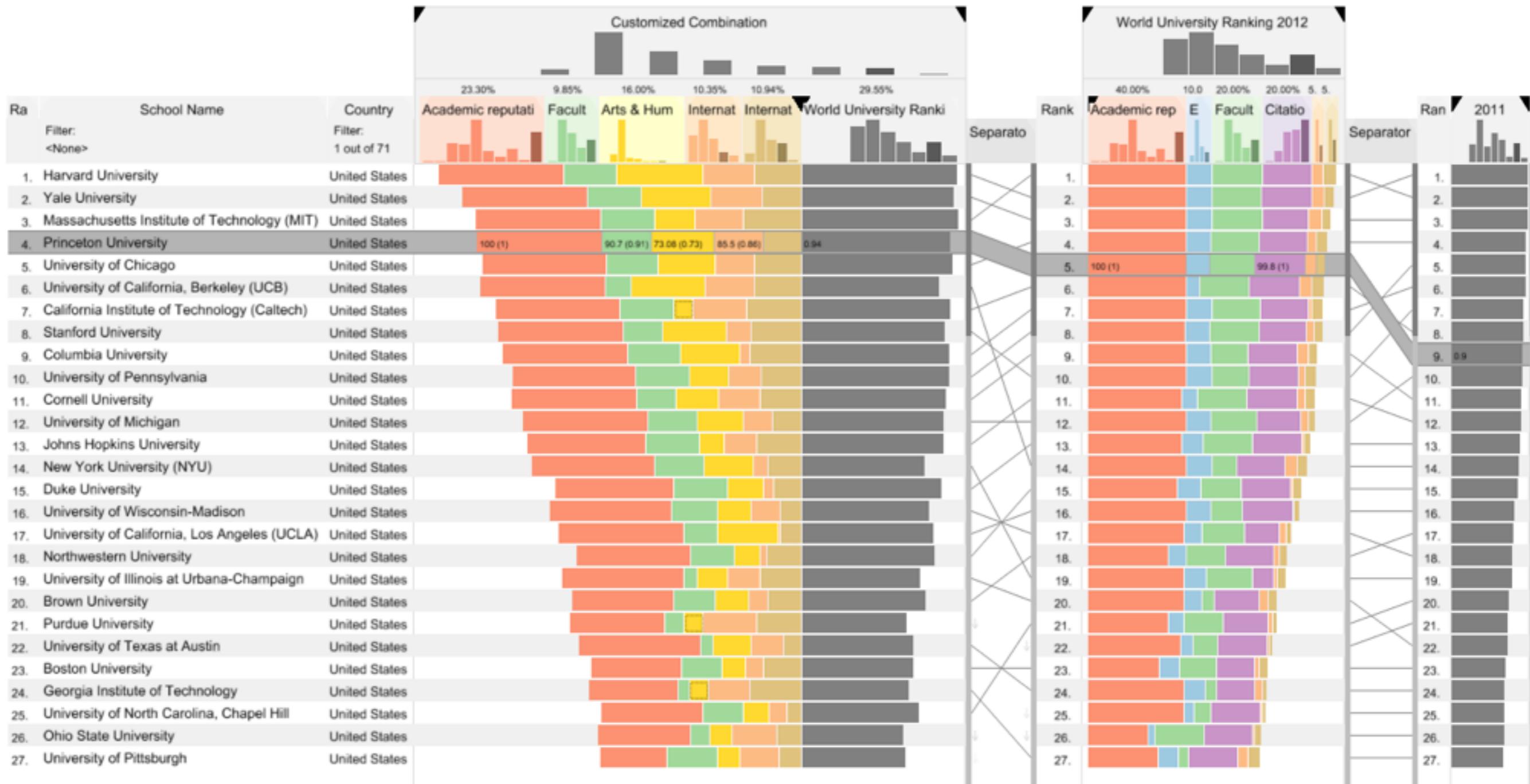
no / little analytics

strong analytics  
component

# Table Lens



# LineUp



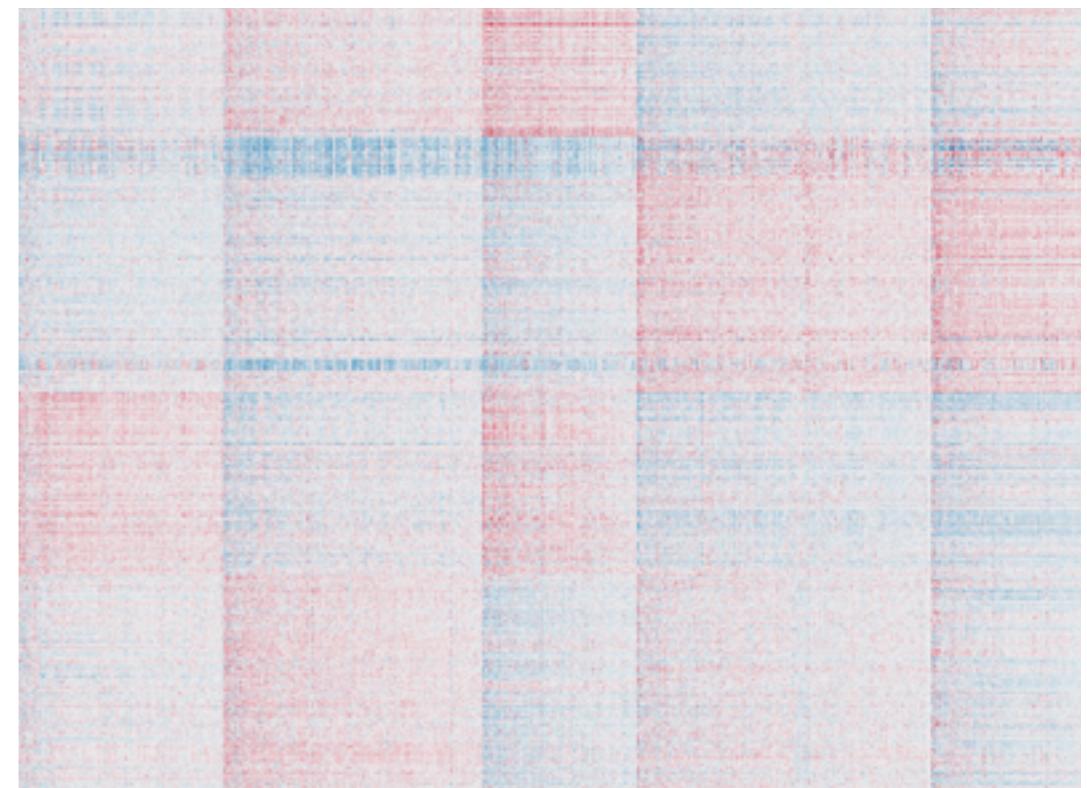
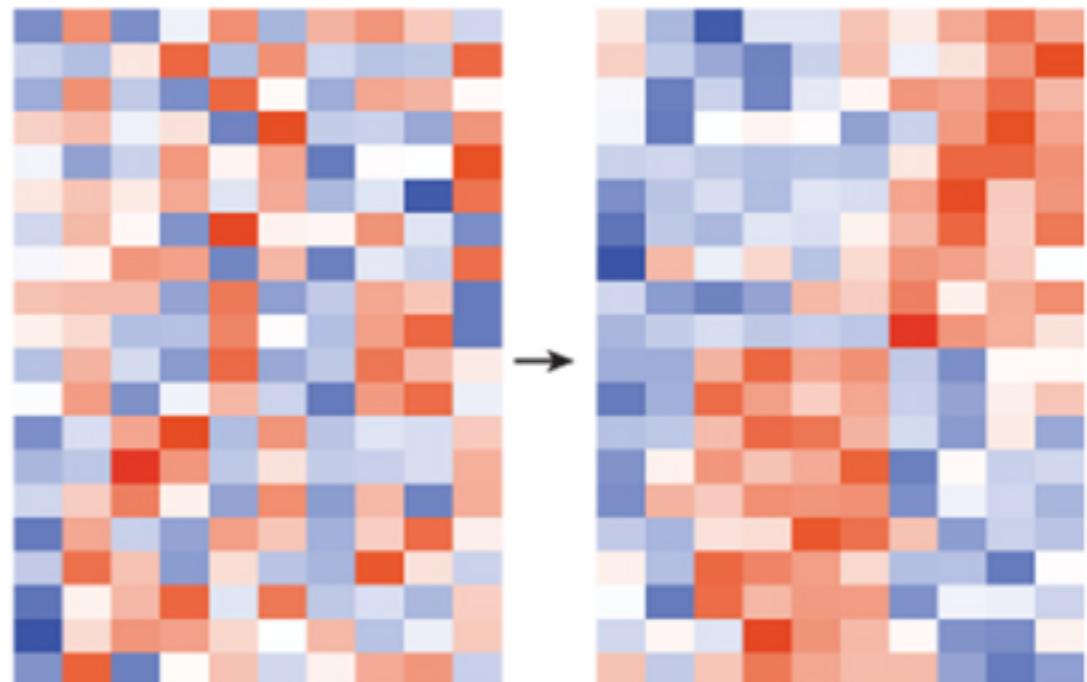
Video at <http://lineup.caleydo.org>

# Heat Map

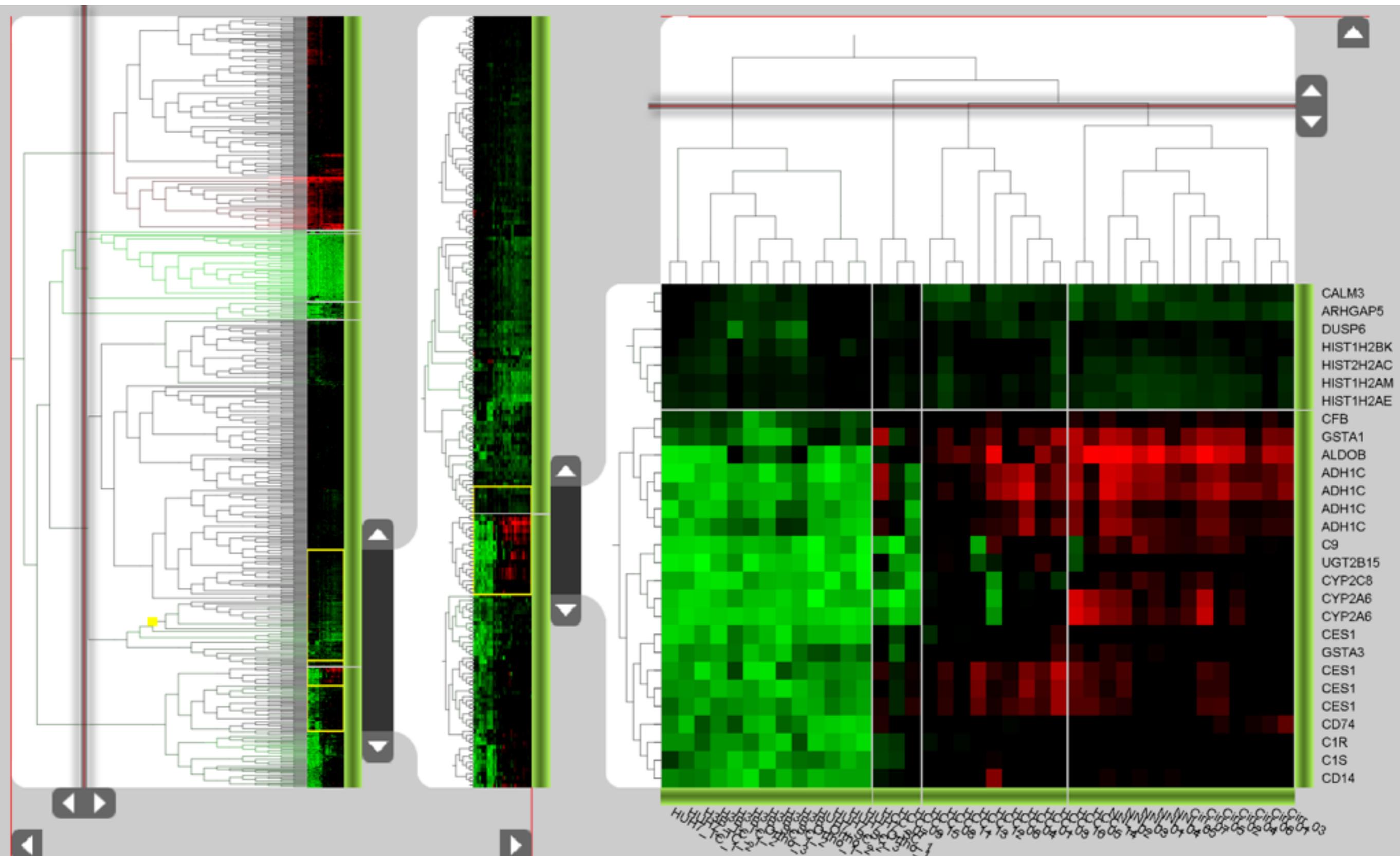
Each cell is a “pixel”, value encoded using color

Meaning derived from ordering  
If no ordering inherent,  
clustering is used

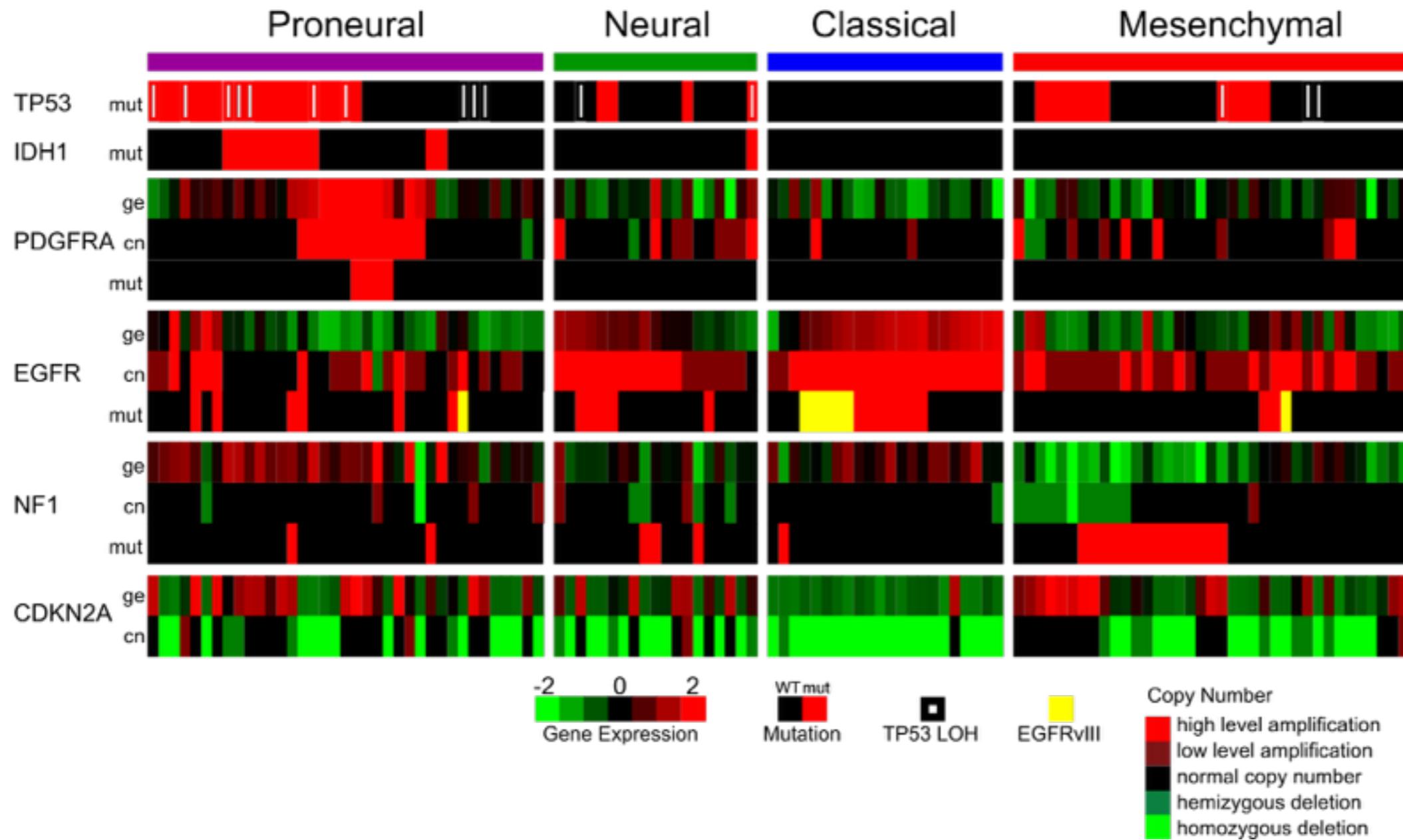
Scalable – 1 px per item  
Good for homogeneous data

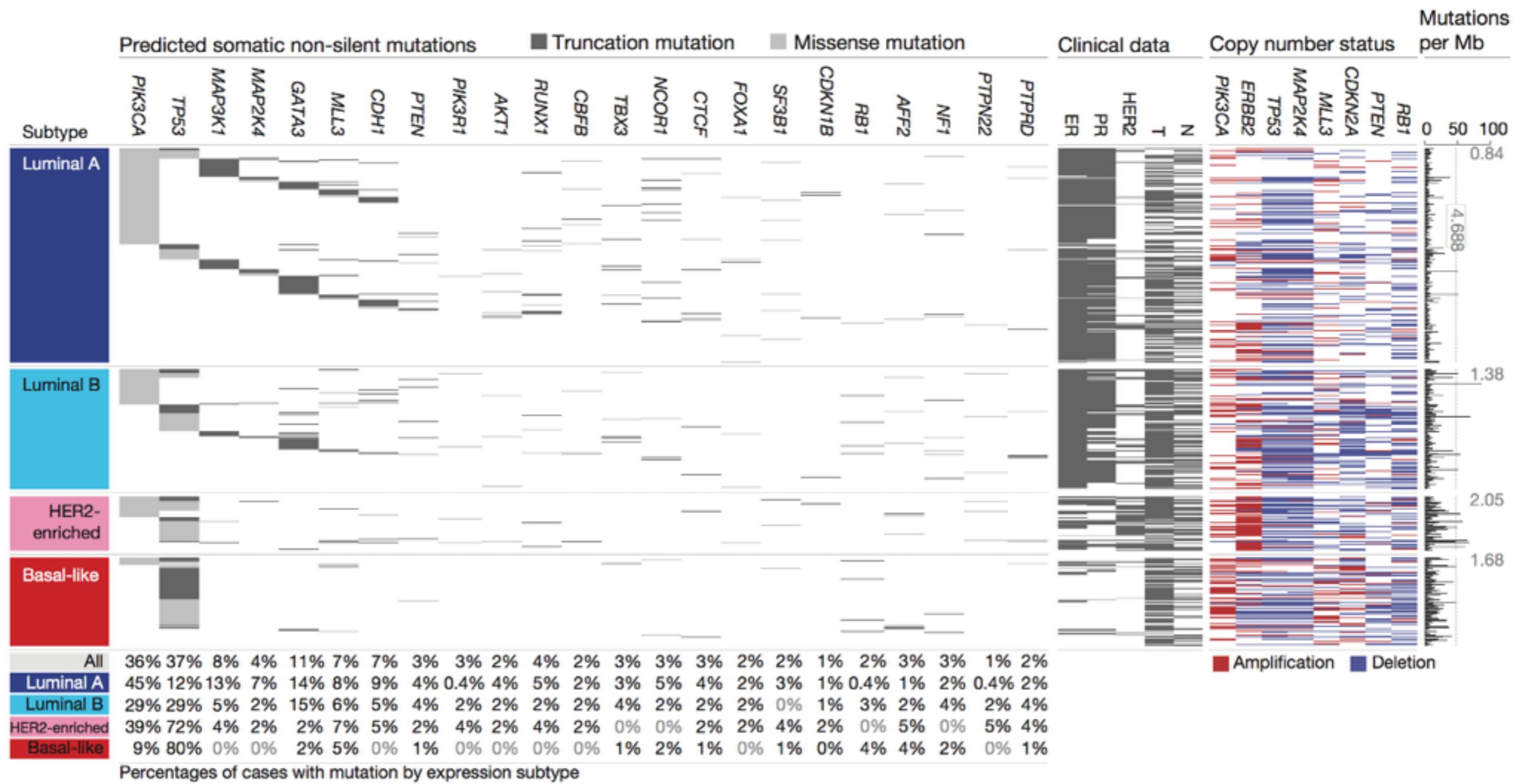


# Clustered Heatmap + Dendrograms



# Heterogeneous Data?





# Heterogeneous Heatmap

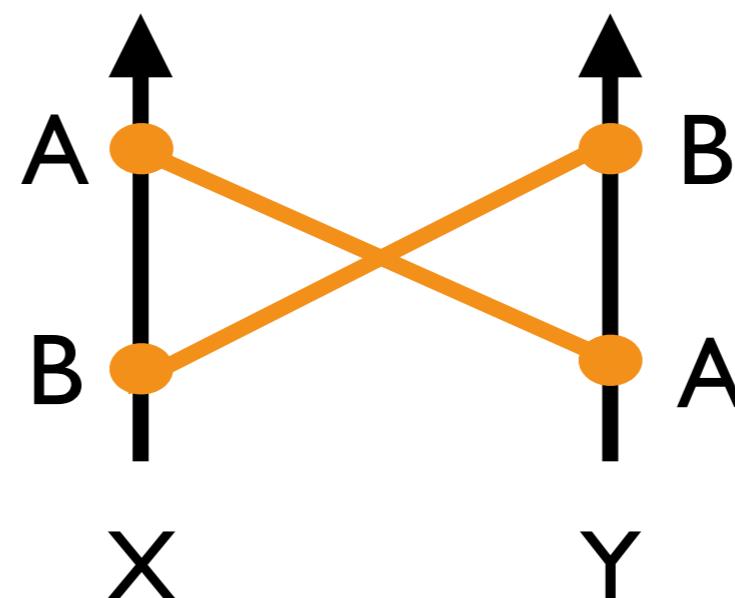
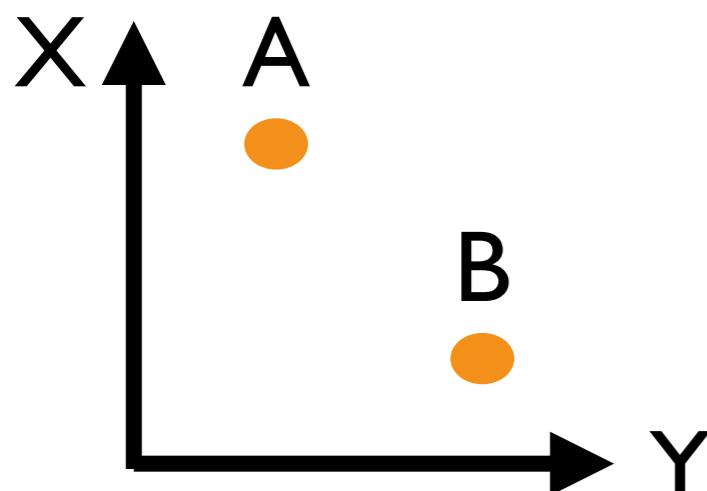
TCGA Paper, Nature 2012

# Parallel Coordinates (PC)

Inselberg 1985

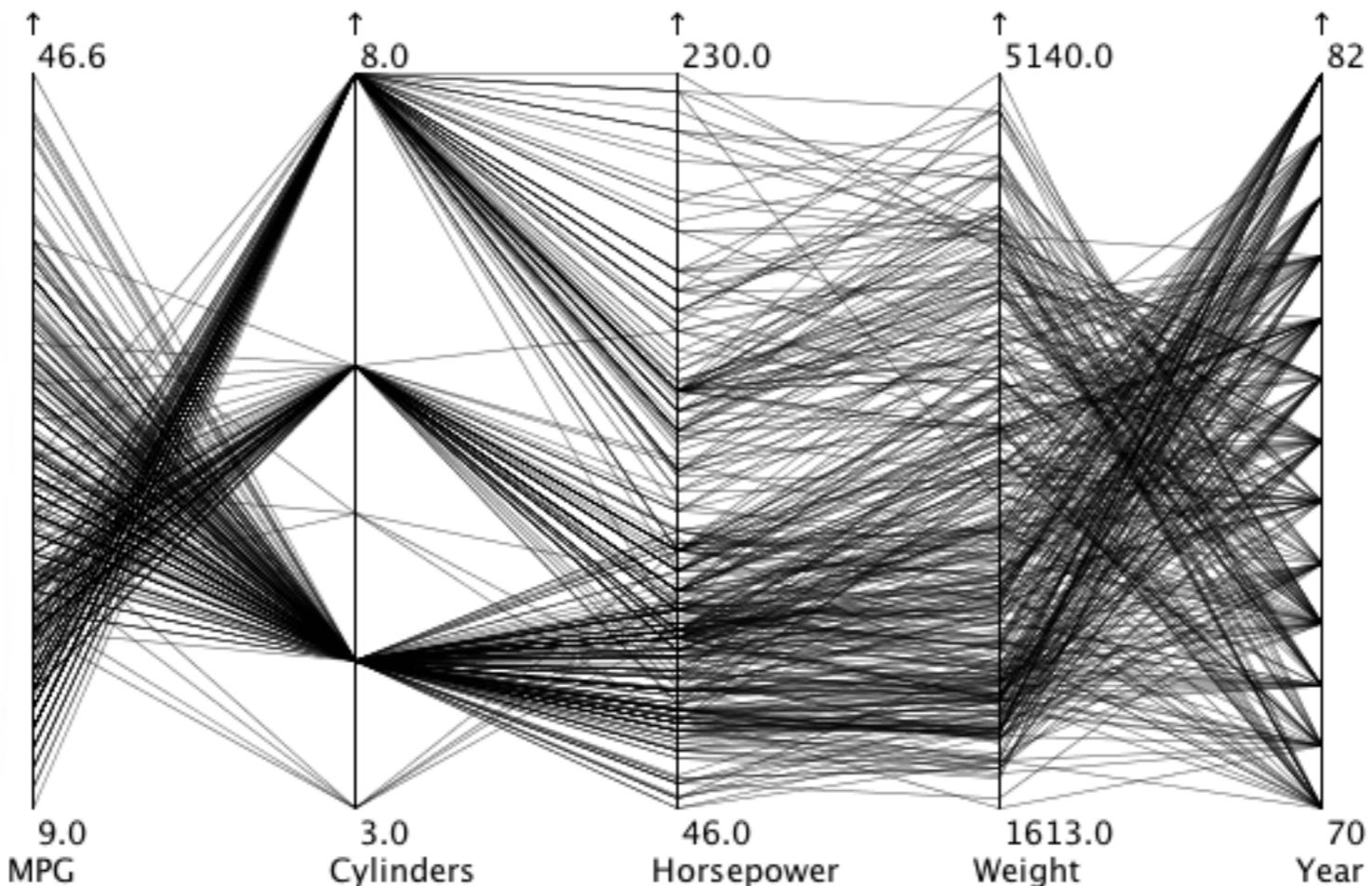
Axes represent attributes

Lines connecting axes represent items



# Example: Cars Dataset

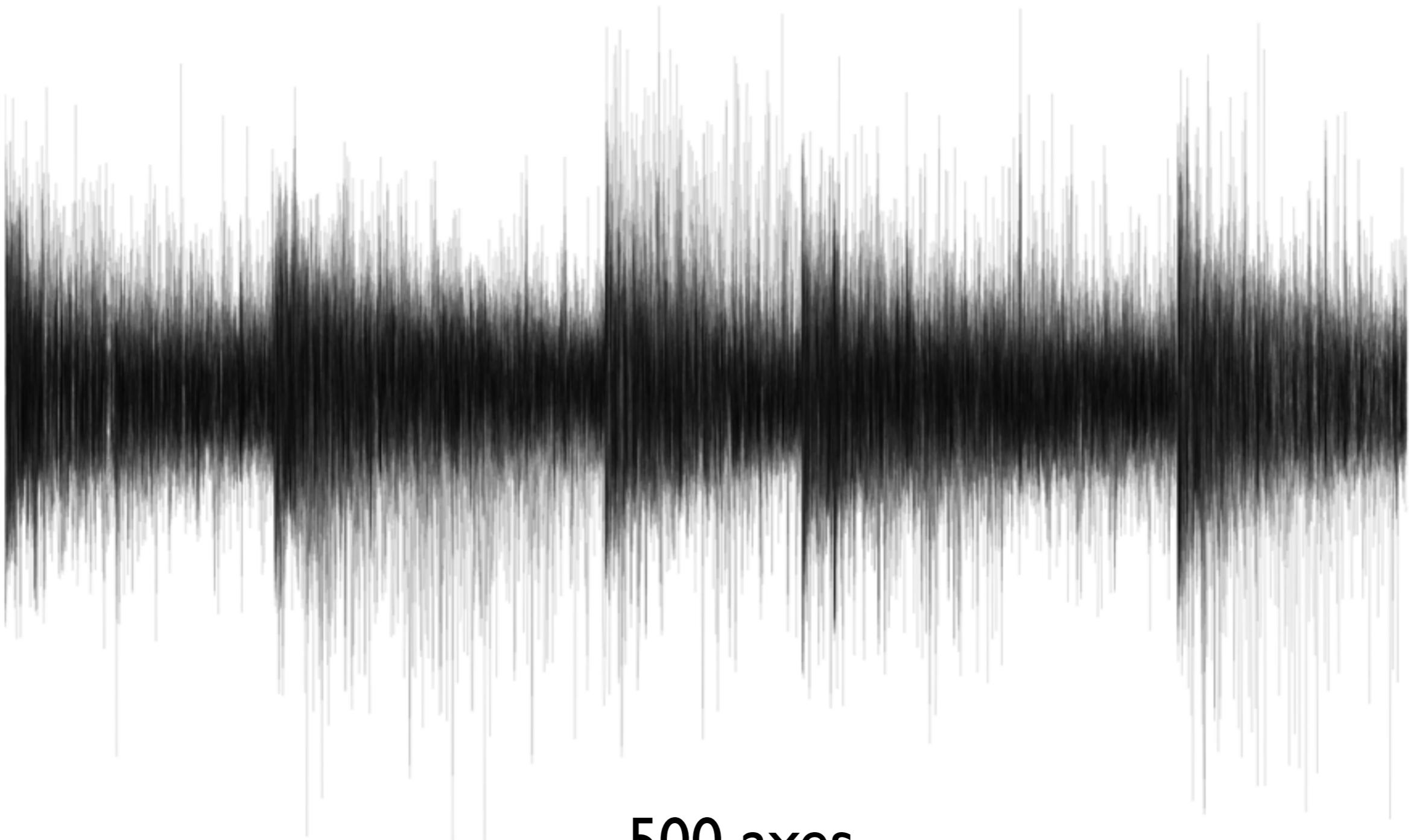
1	MPG	Cylinders	Horsepower	Weight	Acceleration	Year	Origin
2	18	8	130	3504	12	70	USA
3	15	8	165	3693	11.5	70	USA
4	18	8	150	3436	11	70	USA
5	16	8	150	3433	12	70	USA
6	17	8	140	3449	10.5	70	USA
7	15	8	198	4341	10	70	USA
8	14	8	220	4354	9	70	USA
9	14	8	215	4312	8.5	70	USA
10	14	8	225	4425	10	70	USA
11	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

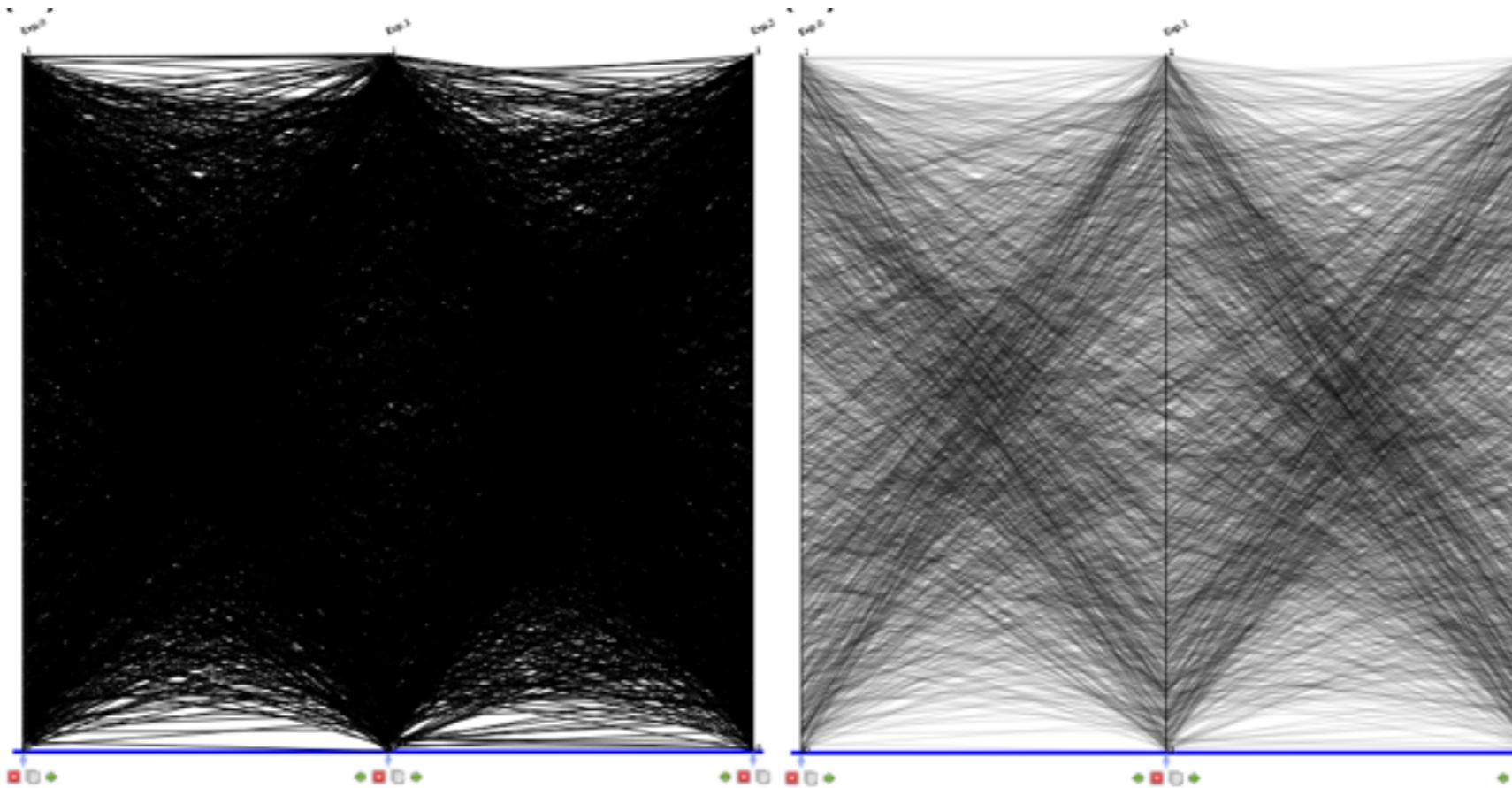
## Scalability to Many Dimensions



500 axes

# PC Limitations

## Scalability to Many Items



Solutions:

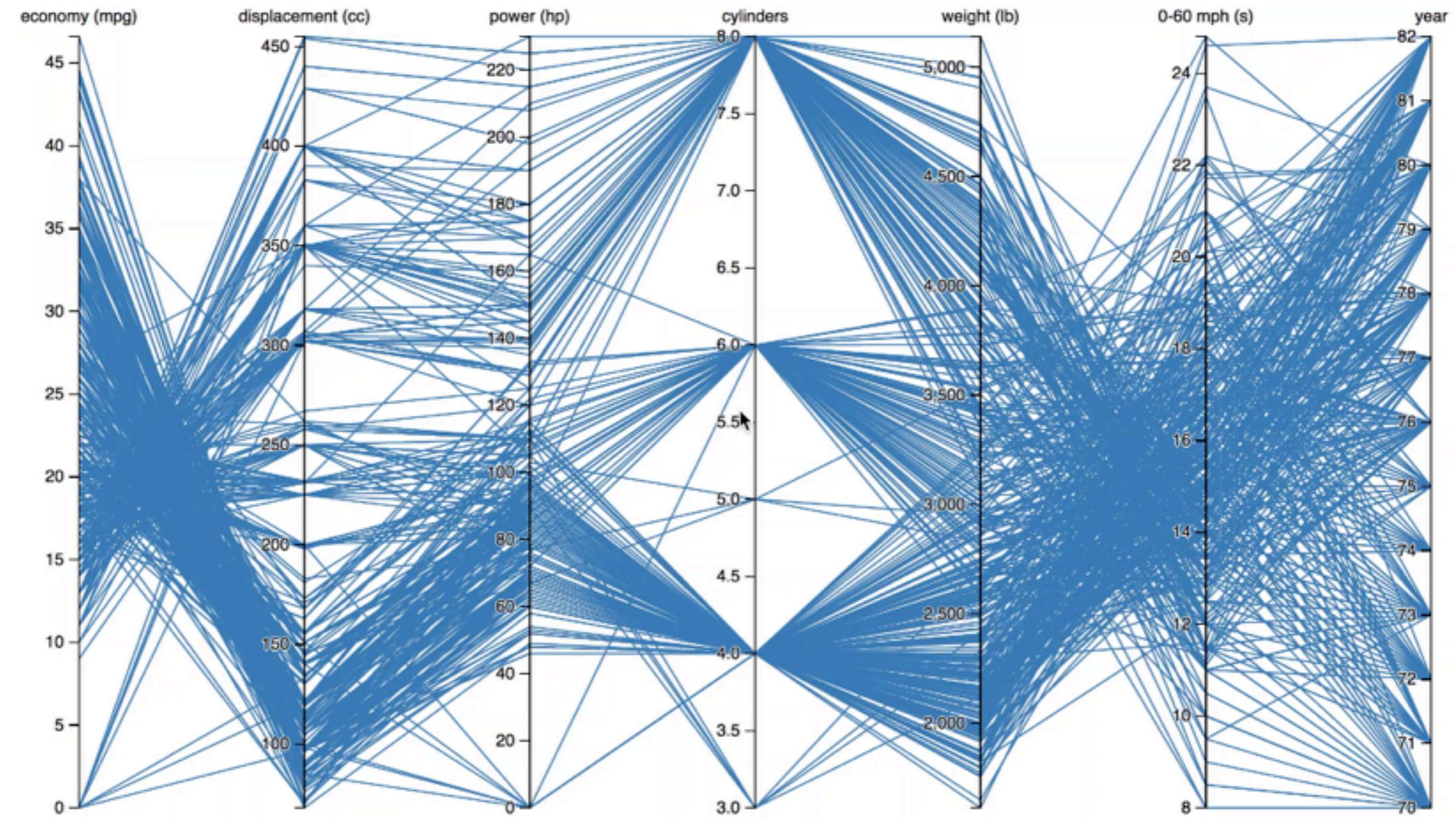
Transparency

Bundling, Clustering

Sampling

# PC Limitations

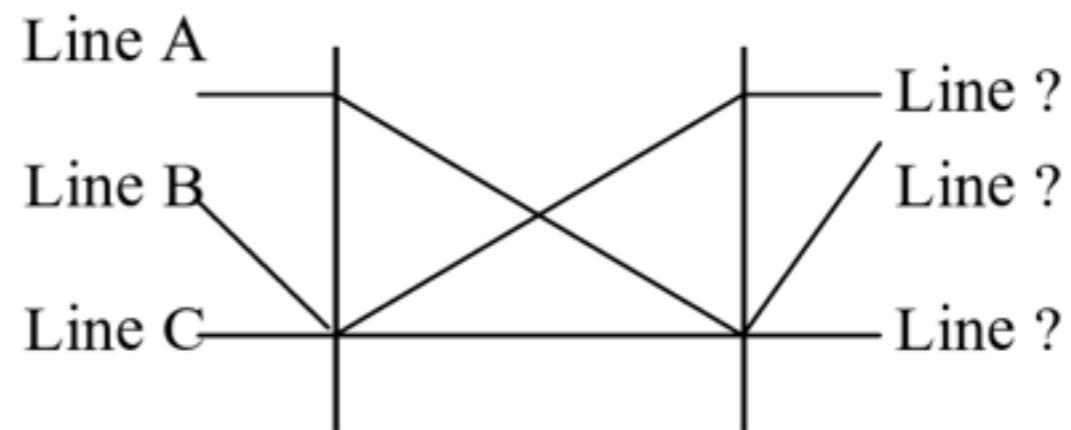
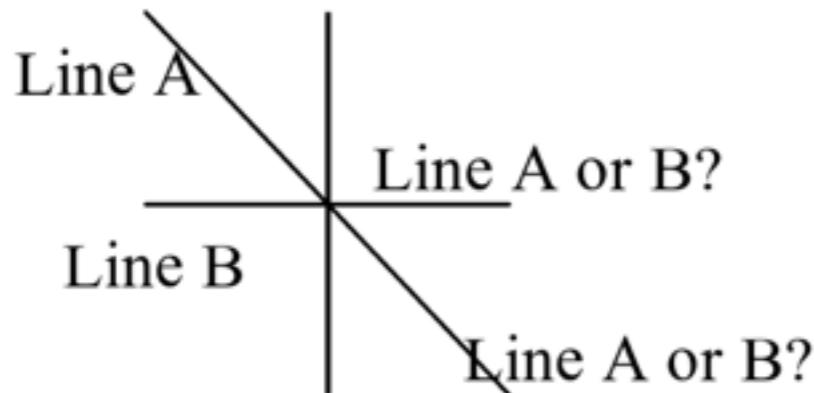
## Correlations only between adjacent axes



Solution: Let user change order

# PC Limitations

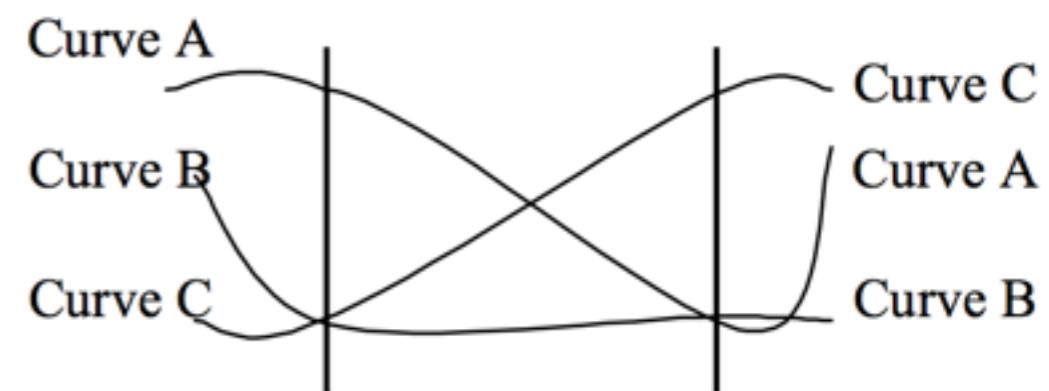
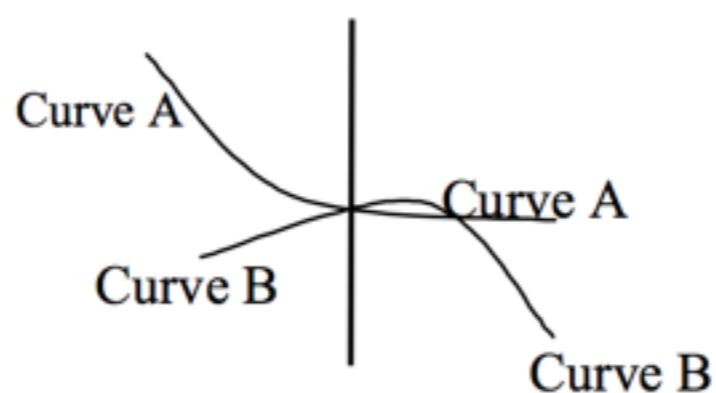
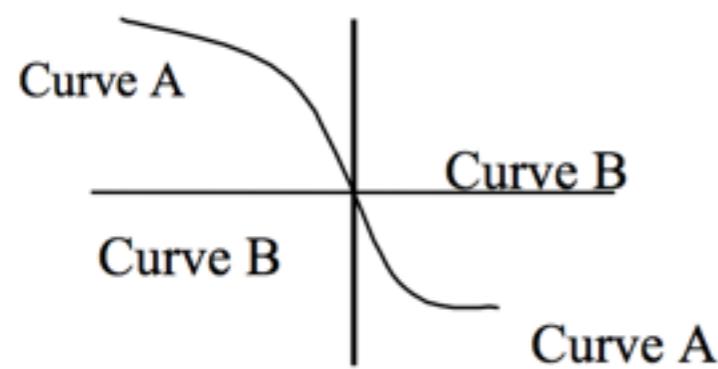
## Ambiguity



**Solutions:**

Interactive highlighting

Curves

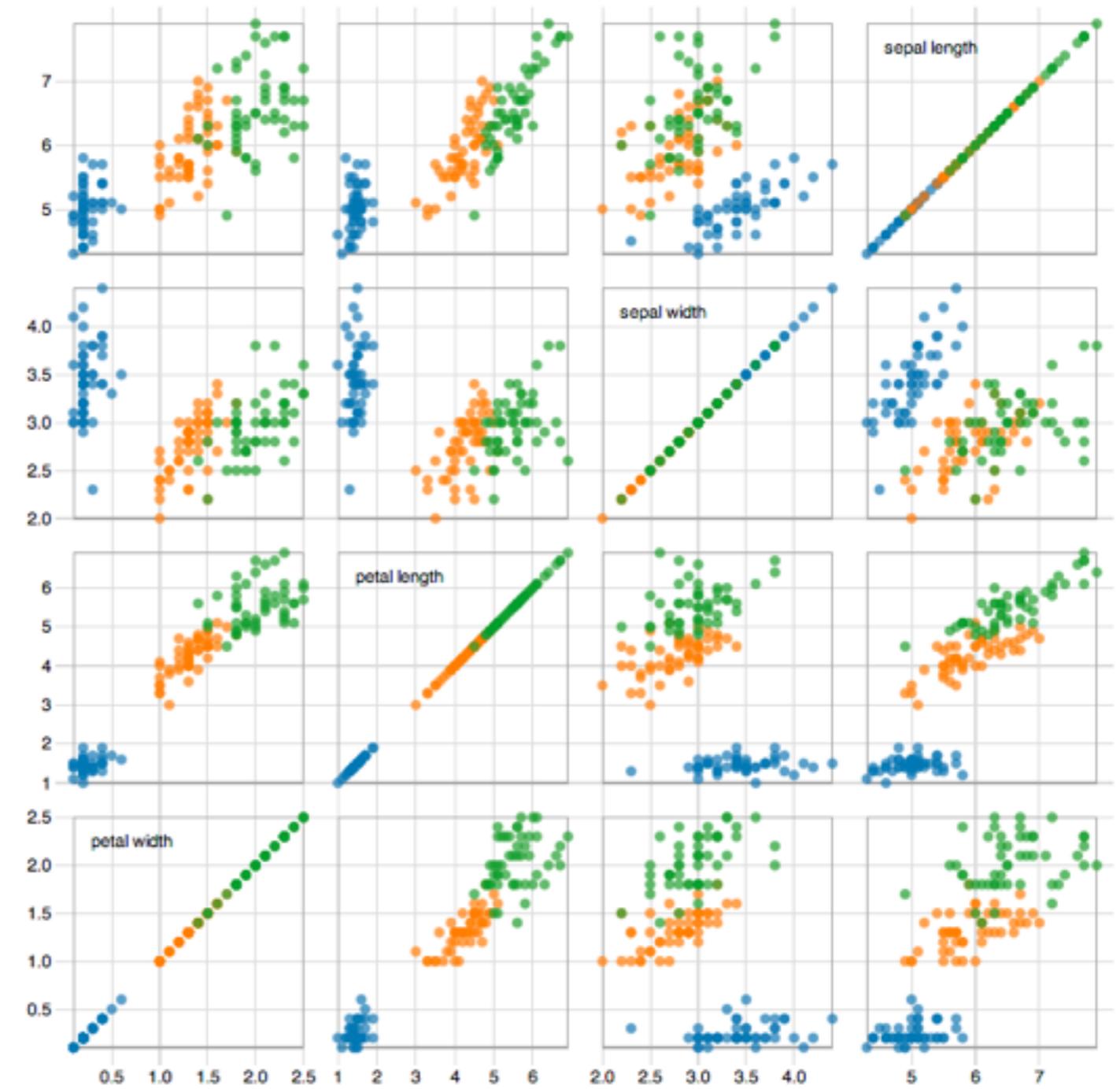


# Scatterplot Matrix (SPLOM)

N dimensions

$N^2$  scatterplots

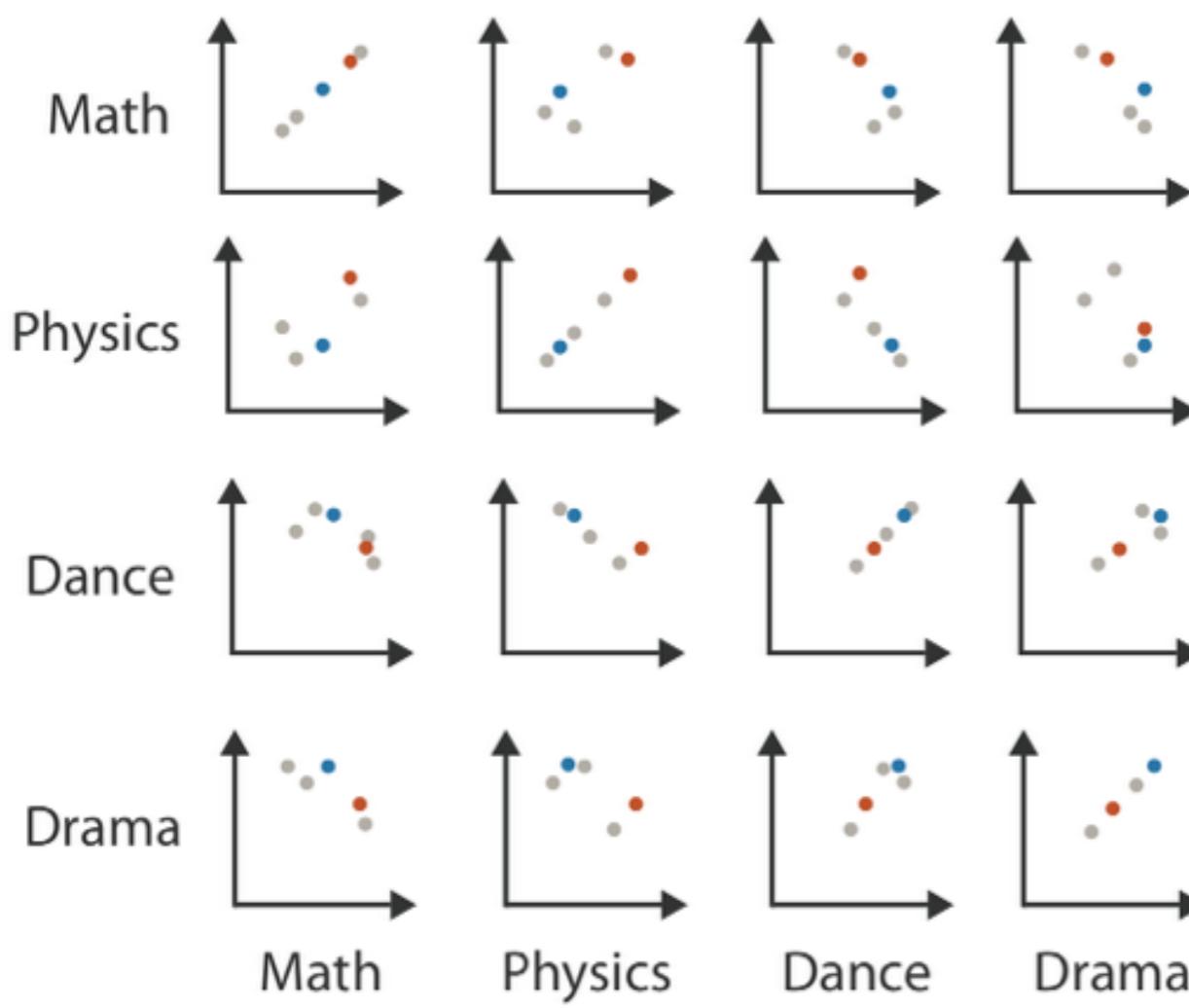
Limited scalability  
(~20 dims,  
~500-1k items)



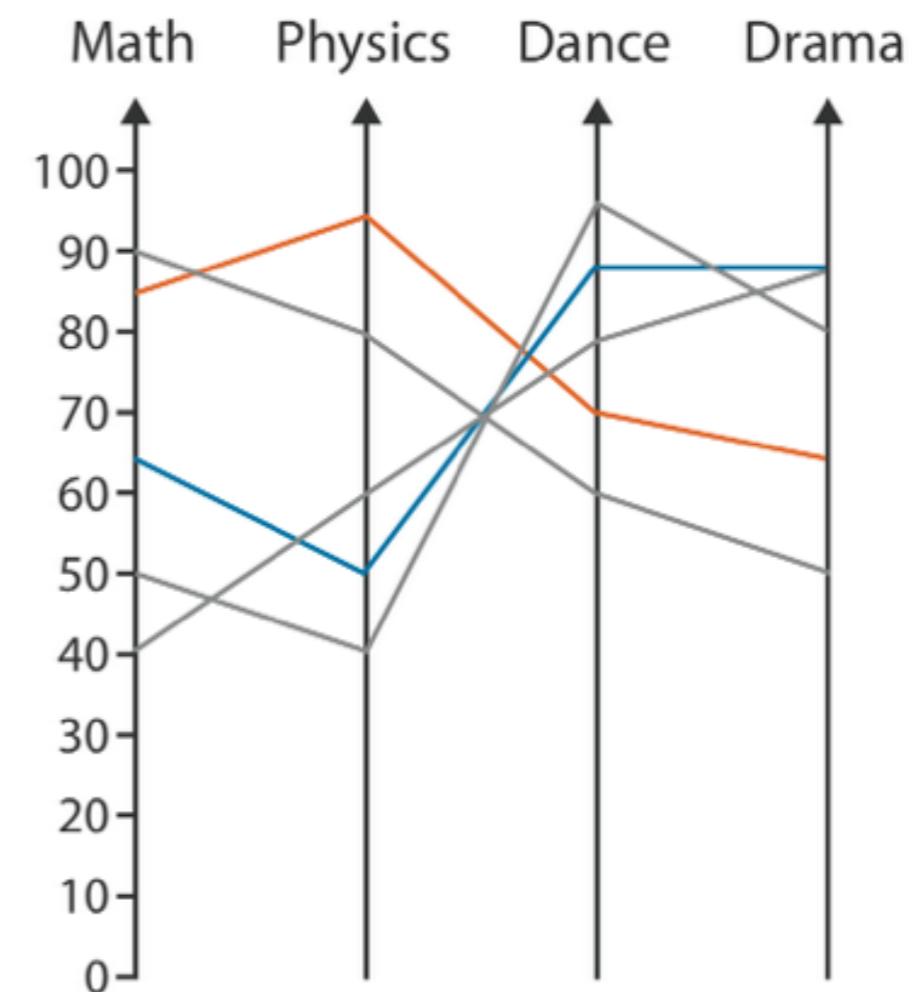
# Table

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

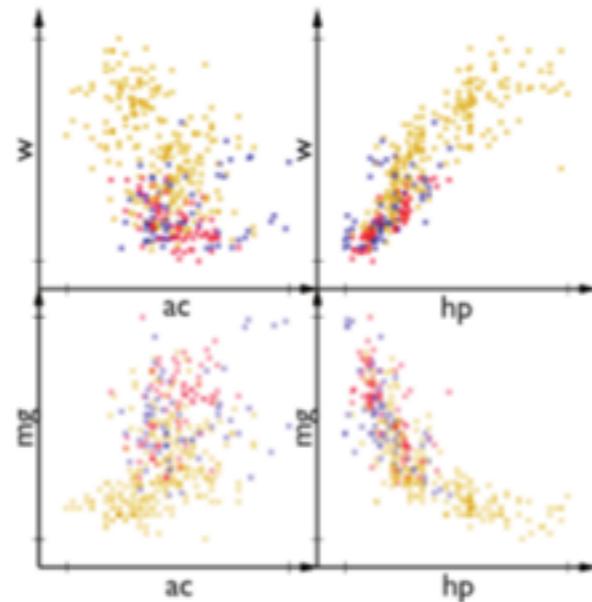
## Scatterplot Matrix



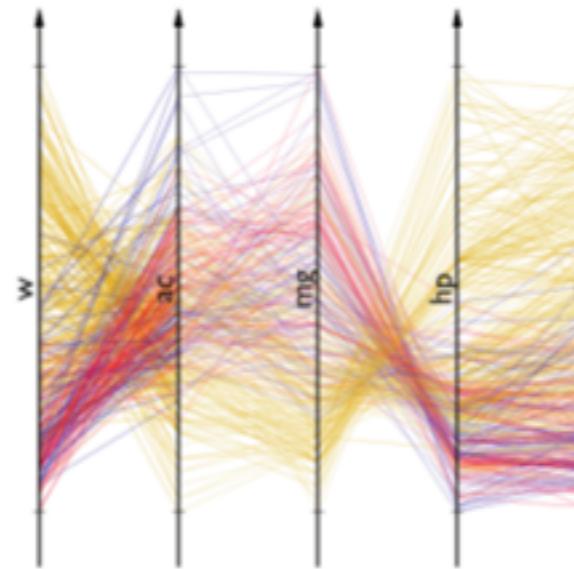
## Parallel Coordinates



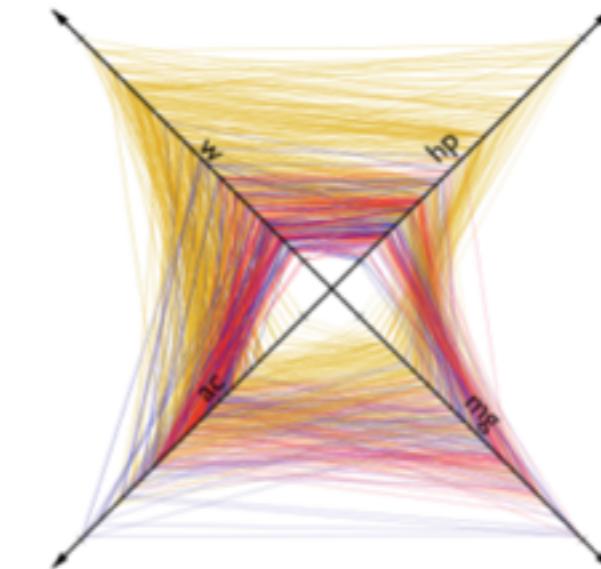
# Flexible Linked Axes (FLINA)



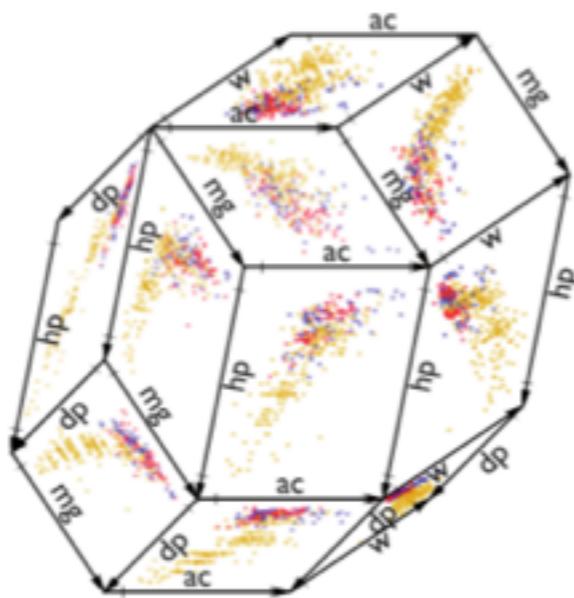
(a) scatterplots



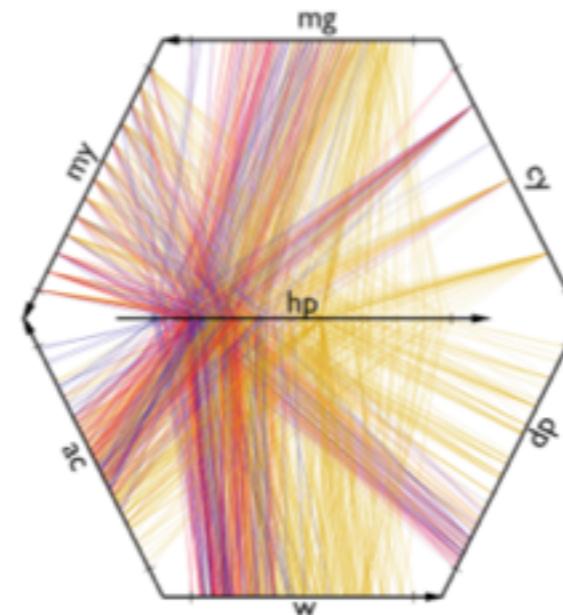
(b) Parallel Coordinates Plot



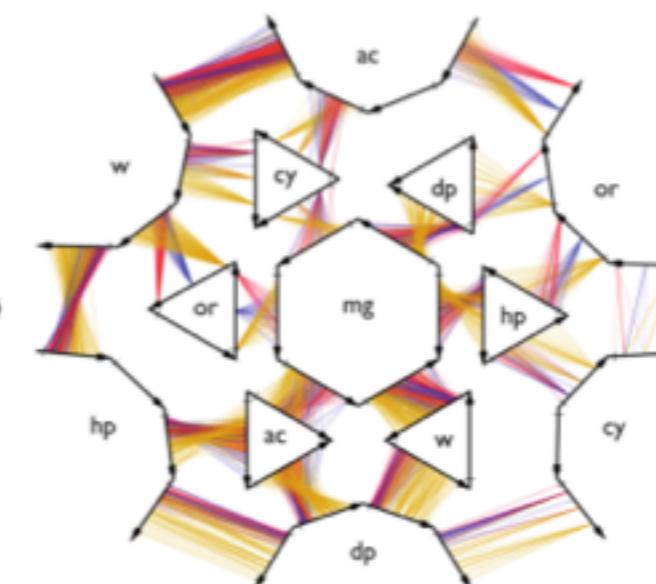
(c) radar chart



(d) Hyperbox



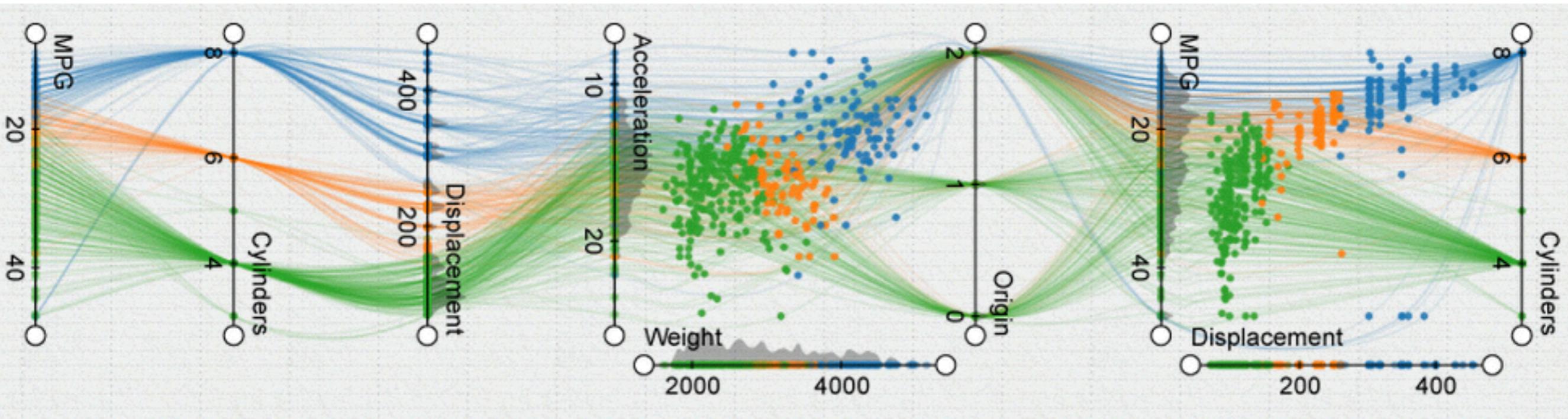
(e) Time Wheel



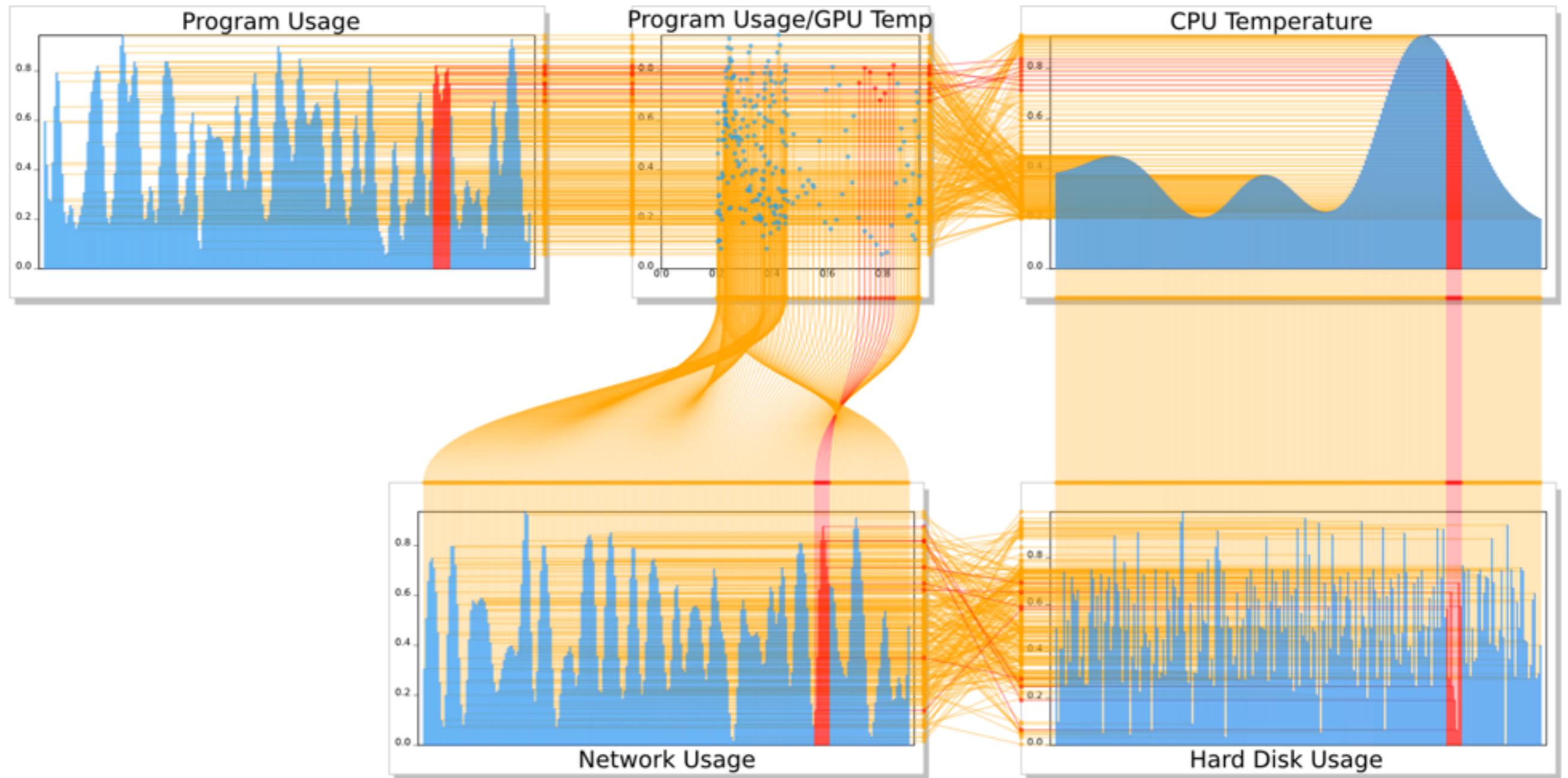
(f) Many-to-many PCP

# Web-based implementation of FLINA concept

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

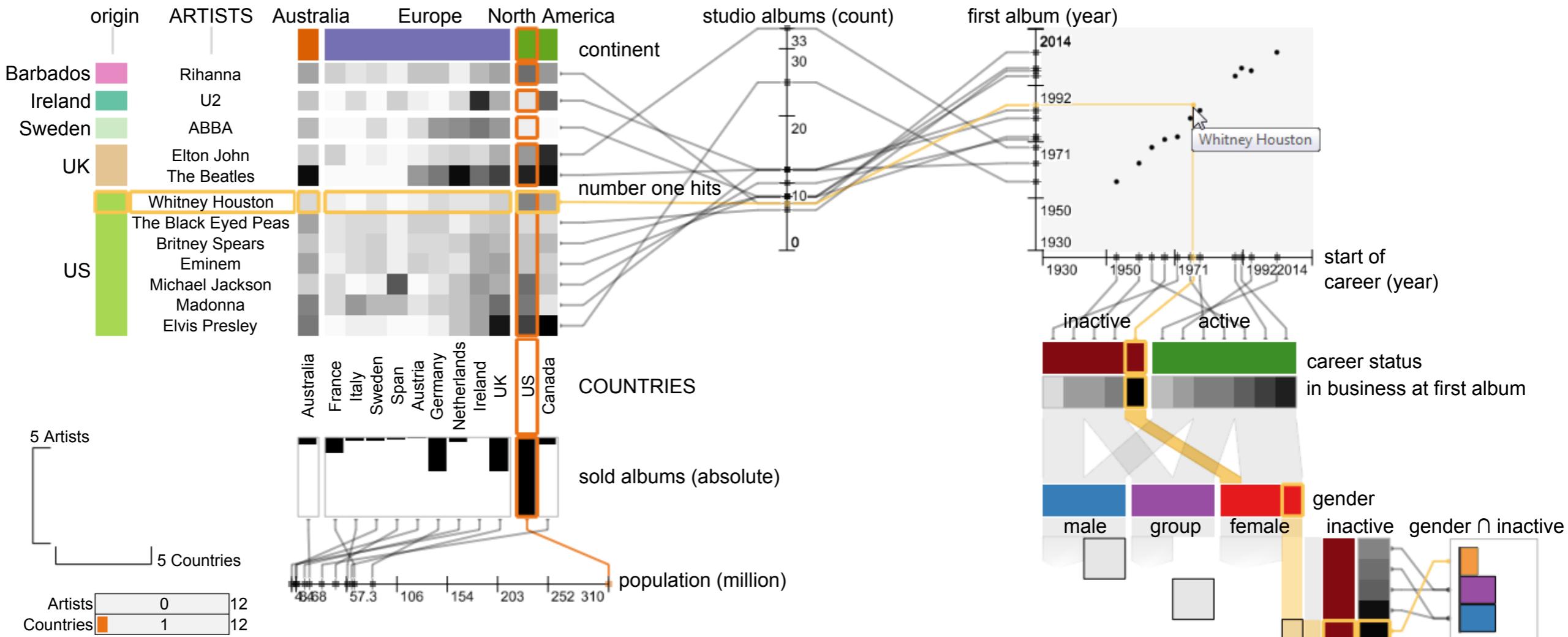


# Connected Charts



# Domino

<http://domino.caleydo.org>



# Domino

## Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets

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



**HARVARD**  
School of Engineering  
and Applied Sciences



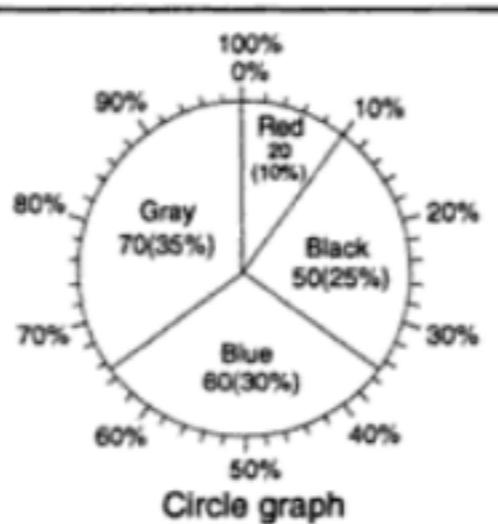
**HARVARD**  
MEDICAL SCHOOL

Audio: "The Long Goodbye" by John Pazdan / CC BY 2.5

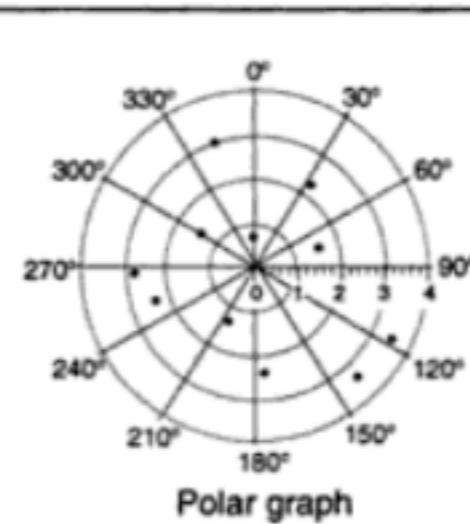
# Radial Axis Techniques

Similar to parallel coordinates

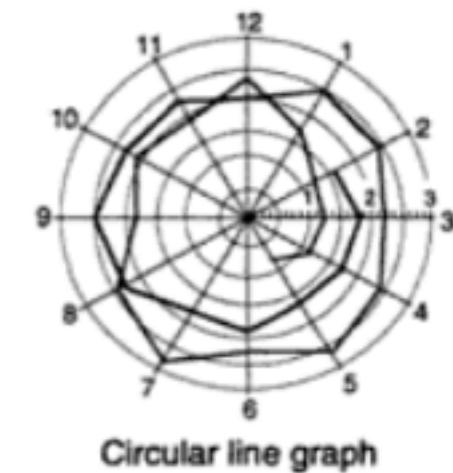
Axes radiate from a common origin



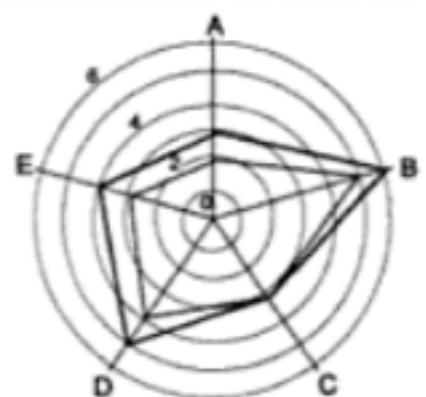
Circle graph  
Shows the relationship of the size of the parts to one another and to the whole



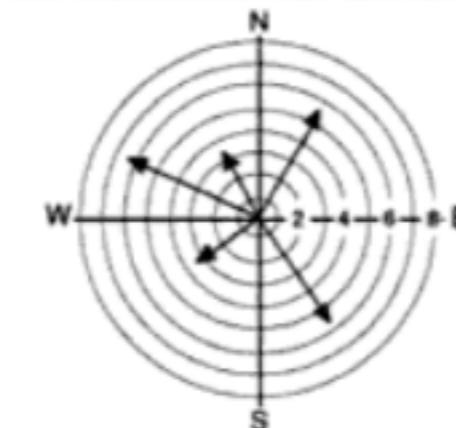
Polar graph  
Used for plotting data that have a value and an angle associated with them



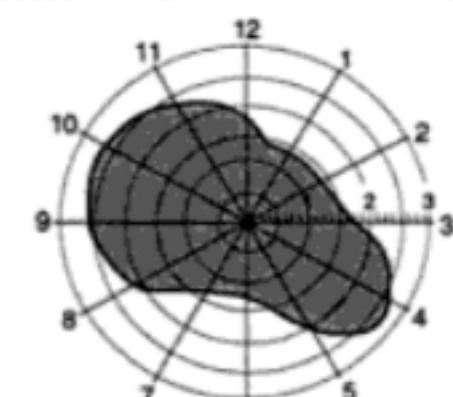
Circular line graph  
Used for plotting data that have a repetitive cycle, particularly those with long cycles such as a day or week



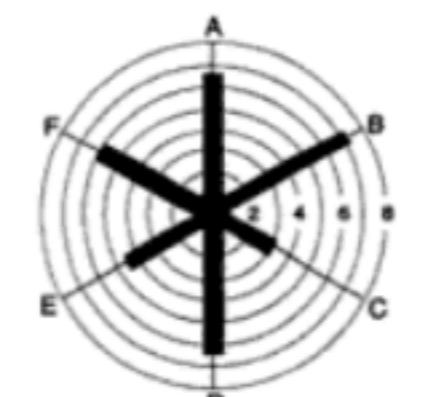
Radar, spider, or star graph  
Uses polygons to compare things with regards to multiple characteristics



Vector graph  
Vectors show direction (angle of arrow) as well as magnitude (length of arrow) of each data point



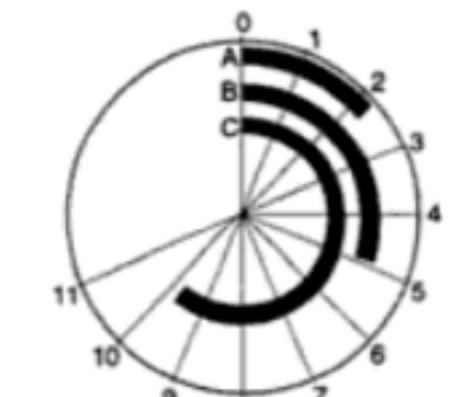
Circular area graph  
Used when the area under the curve is equally or more important than specific data points



Circular column graph  
Used for the same purposes as rectangular column graphs. Sometimes offer an advantage when the data is repetitive

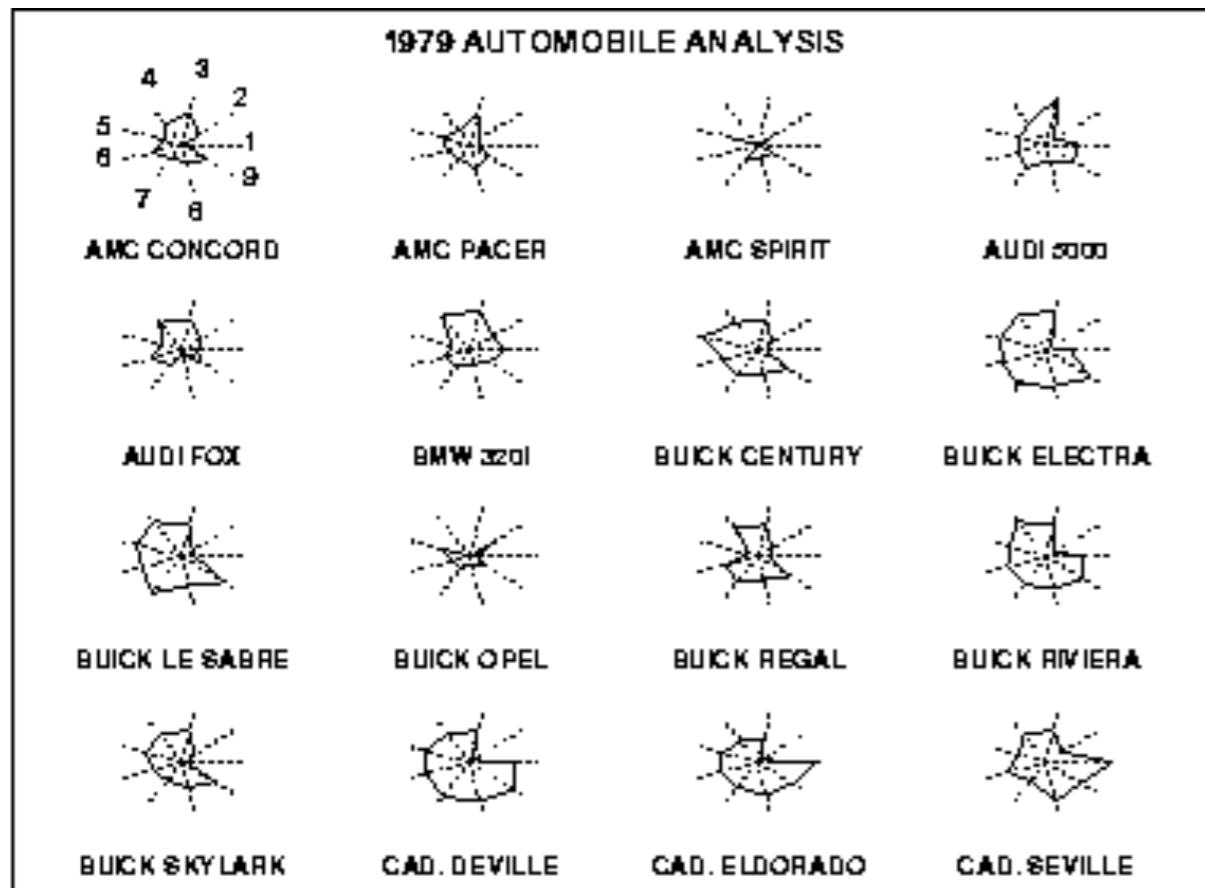


Sector graph  
Similar to the circular column graph with the added option of making sector areas proportional to values they represent



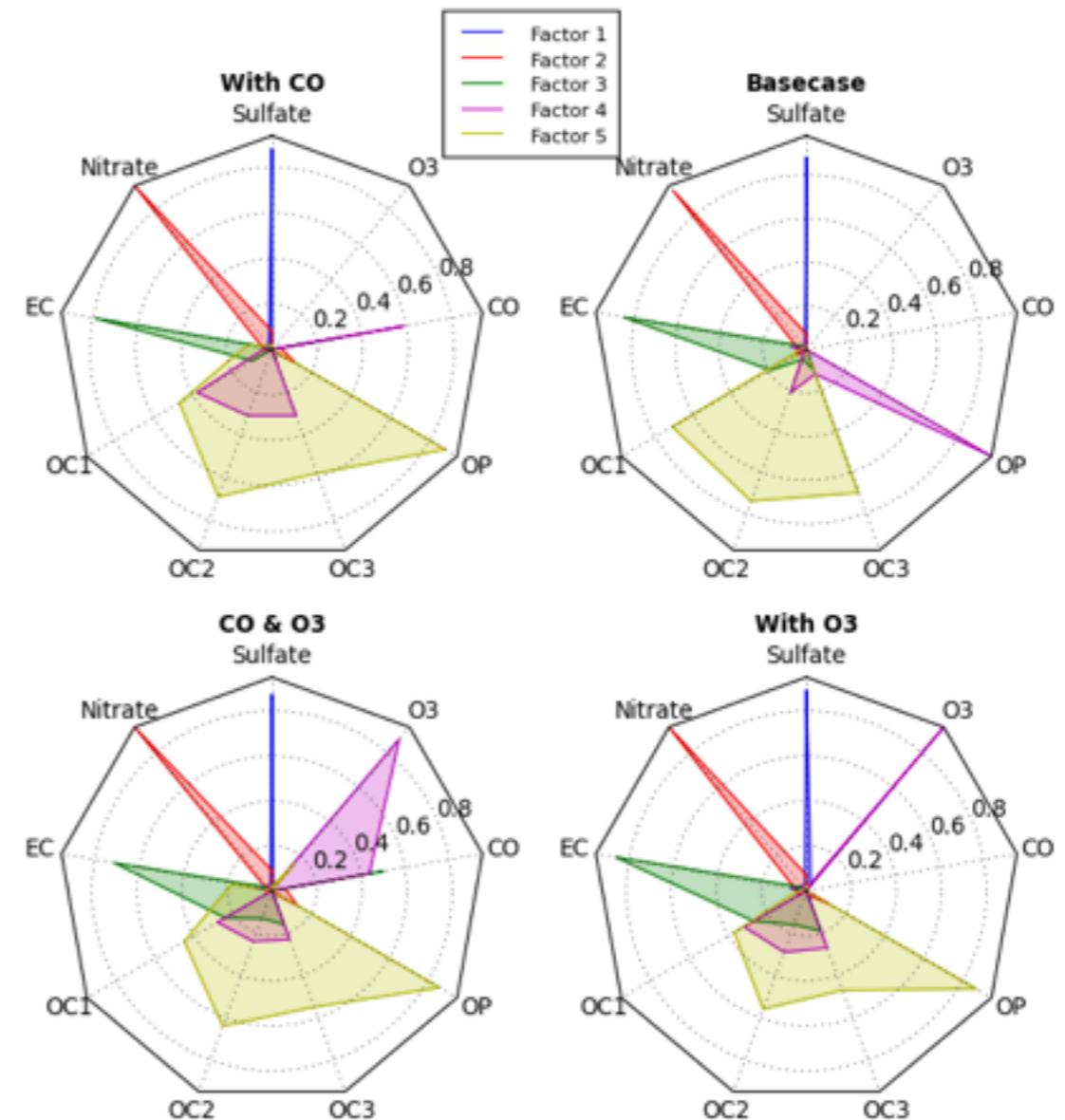
Circular bar graph  
Used primarily for its aesthetic value. Sometimes used with a time series scale to display repetitive events

# Star Plot



Coekin 1969

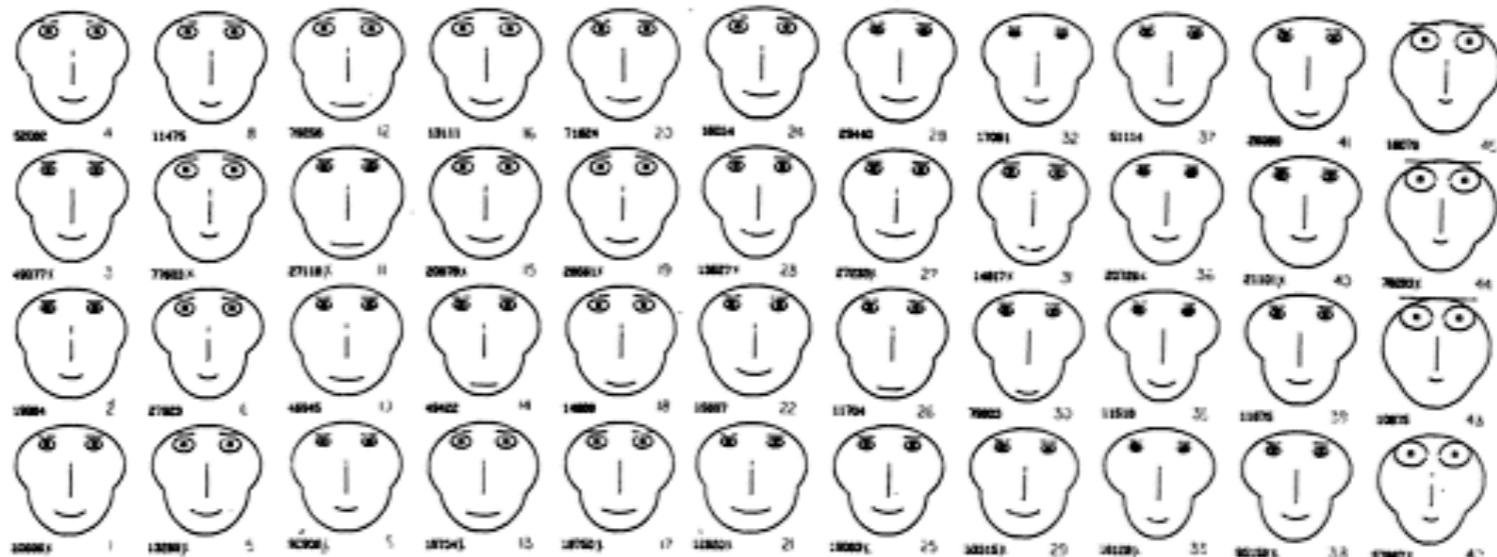
**5-Factor Solution Profiles Across Four Scenarios**



# 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



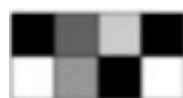
**PROFILE GLYPHS**



**STARS AND  
METROGLYPHS**



**STICKS AND TREES**



**AUTOGLYPH/BOX GLYPH**



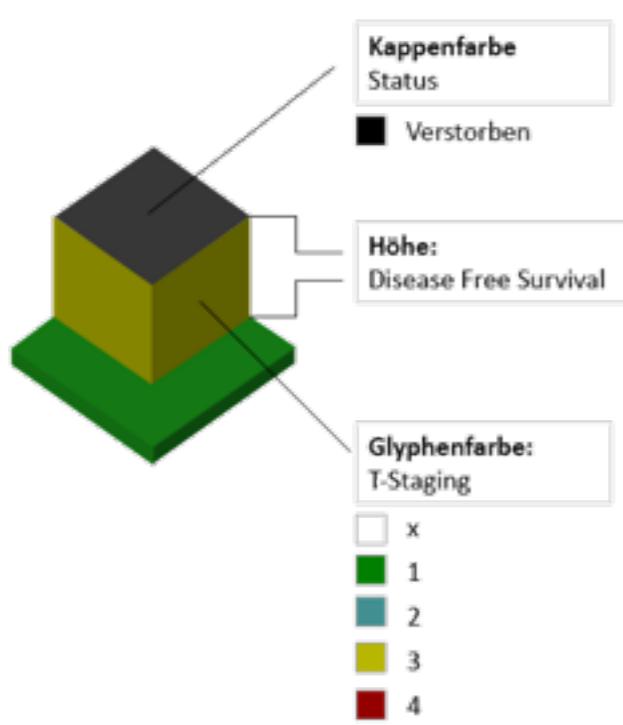
**FACE GLYPHS**



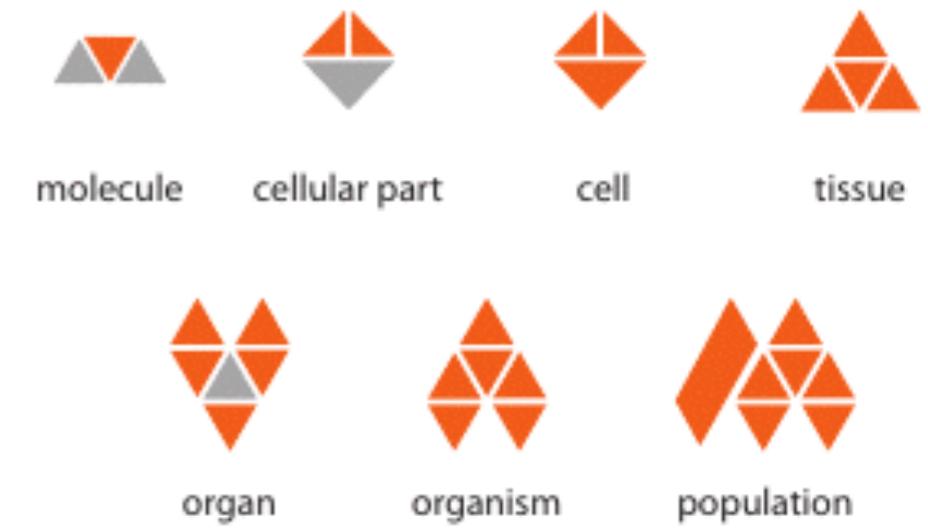
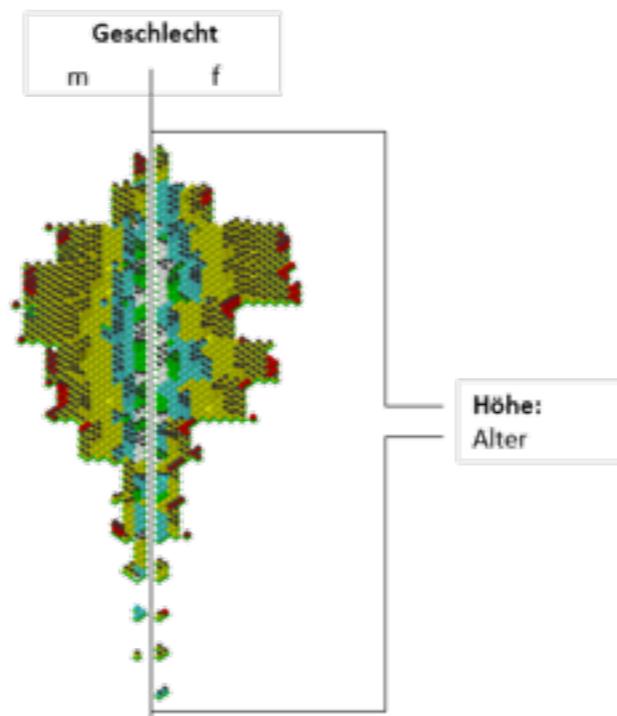
**ARROWS/WEATHERVANES**



# Glyph Examples

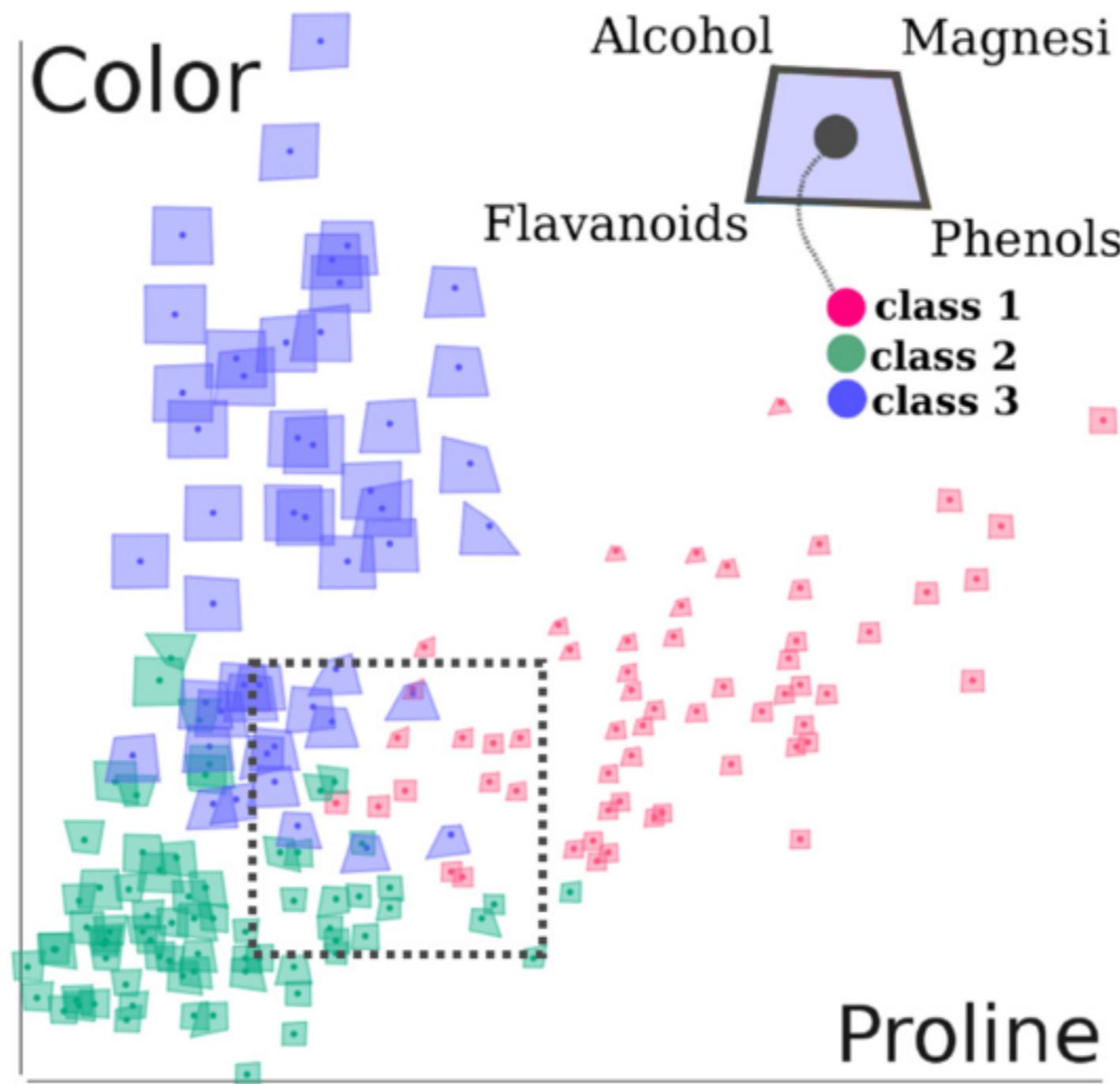


Mueller et al. 2007

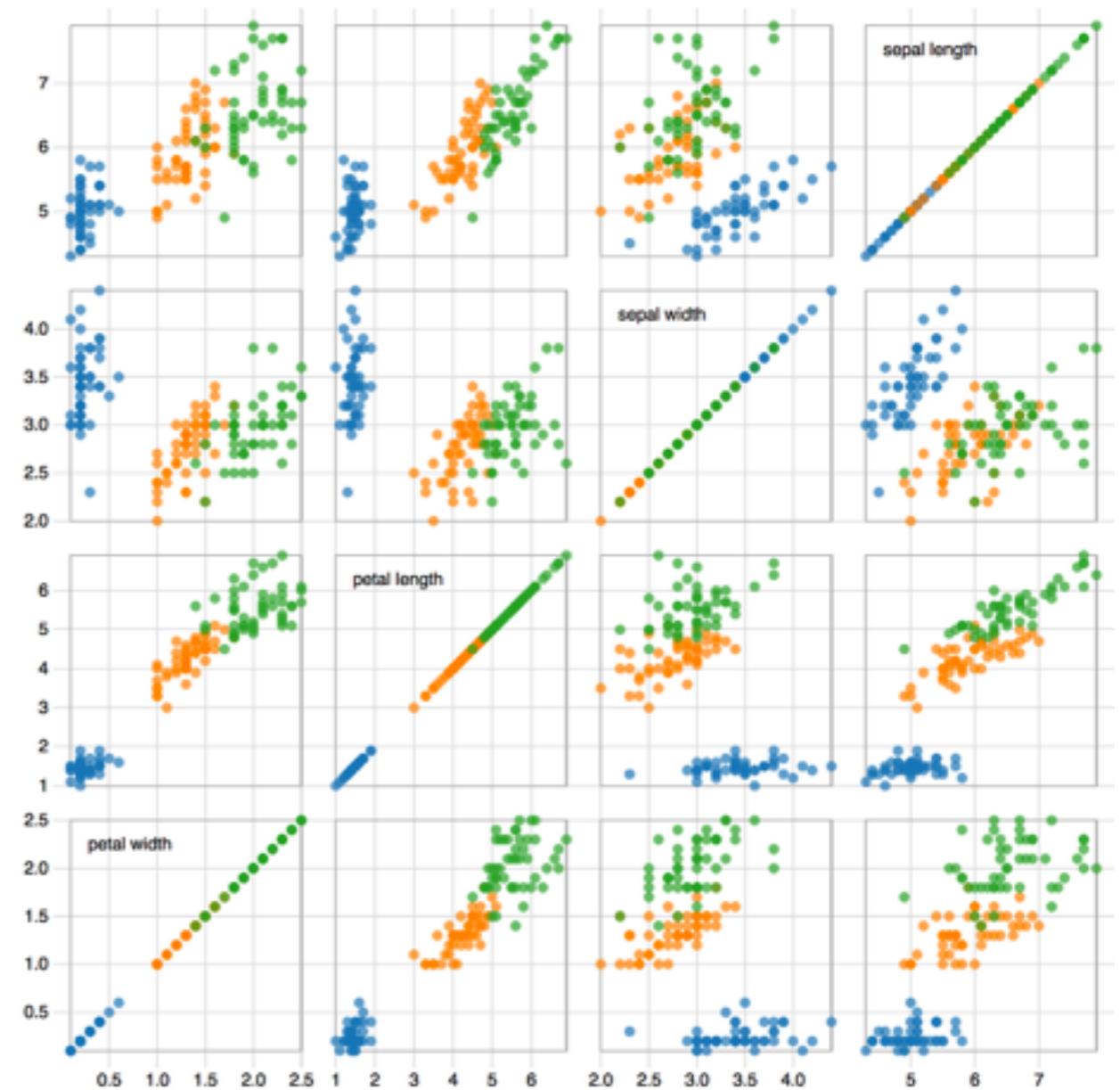
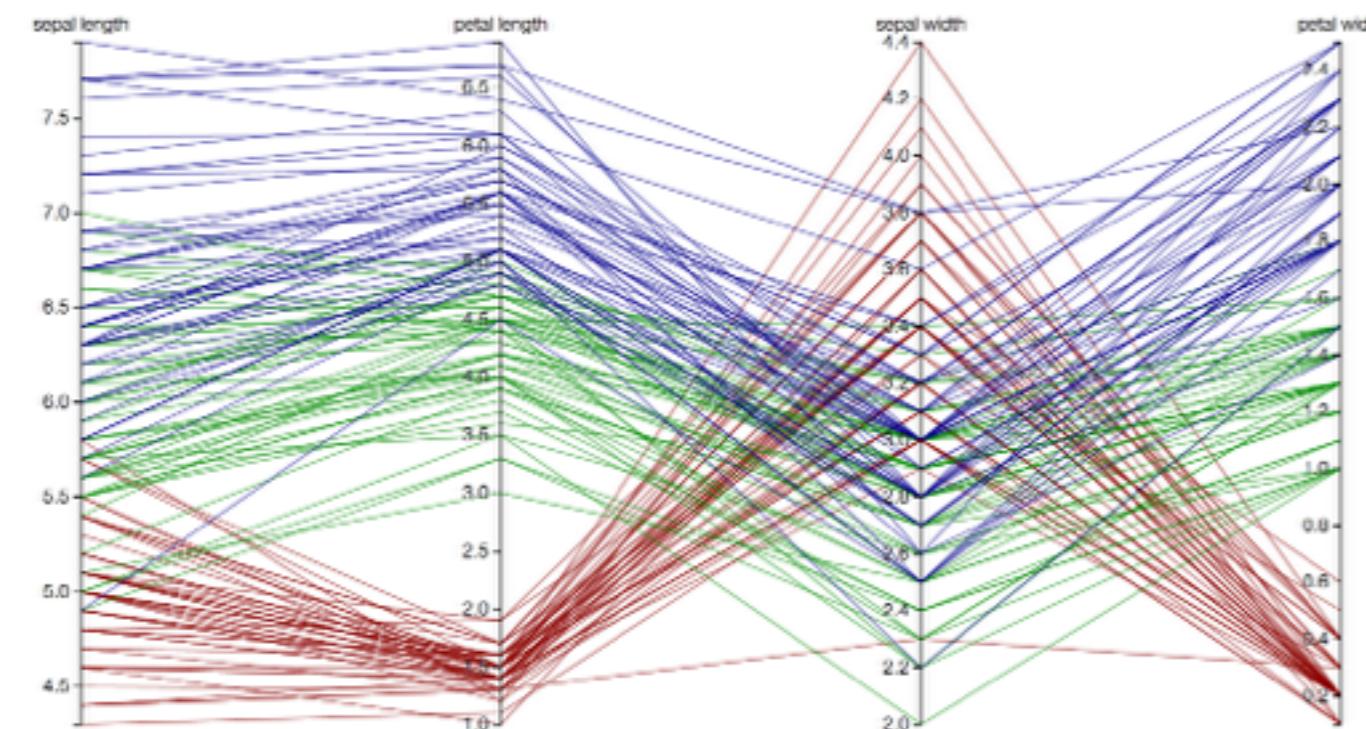


Maguire et al. 2012

# Glyph as Marks in Scatterplot

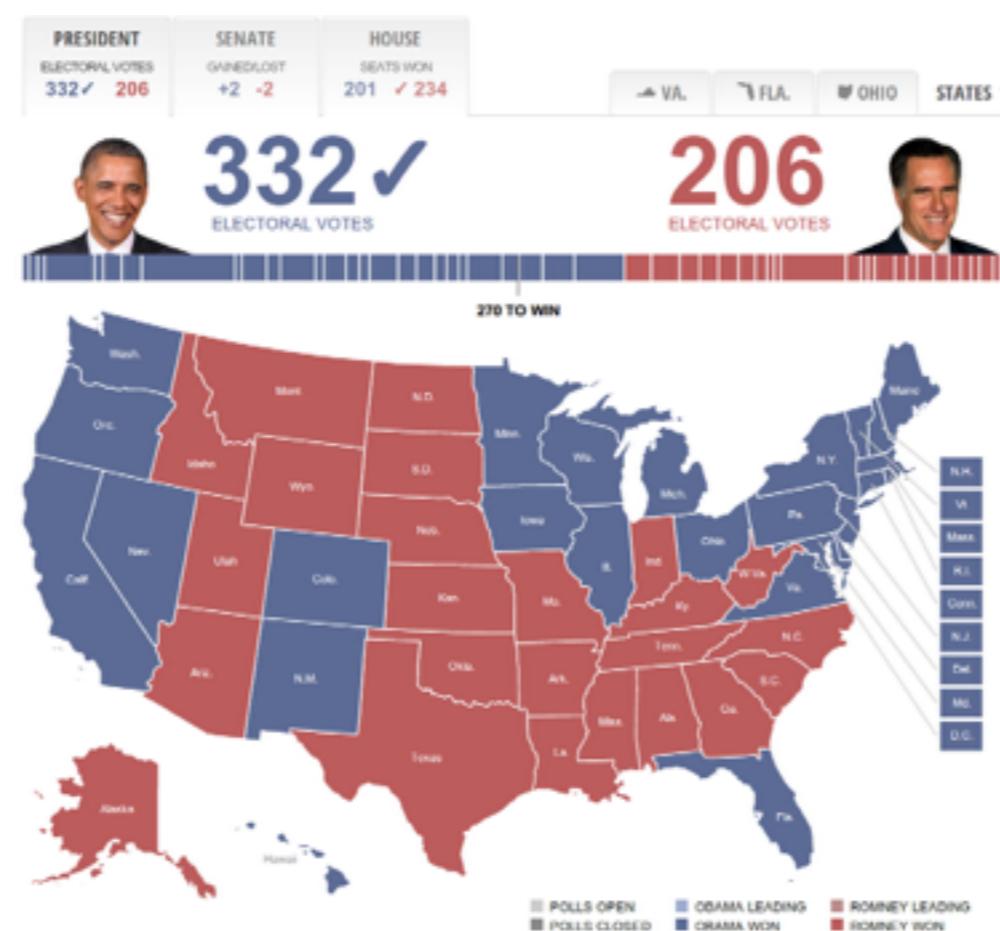


# Single Complex Visualization



# Multiple Simple Visualizations

# Map Visualization



# Map Projections



Cylinder  
projection



Plane  
projection



Cone  
projection

Projection Overview

<http://www.jasondavies.com/maps/>

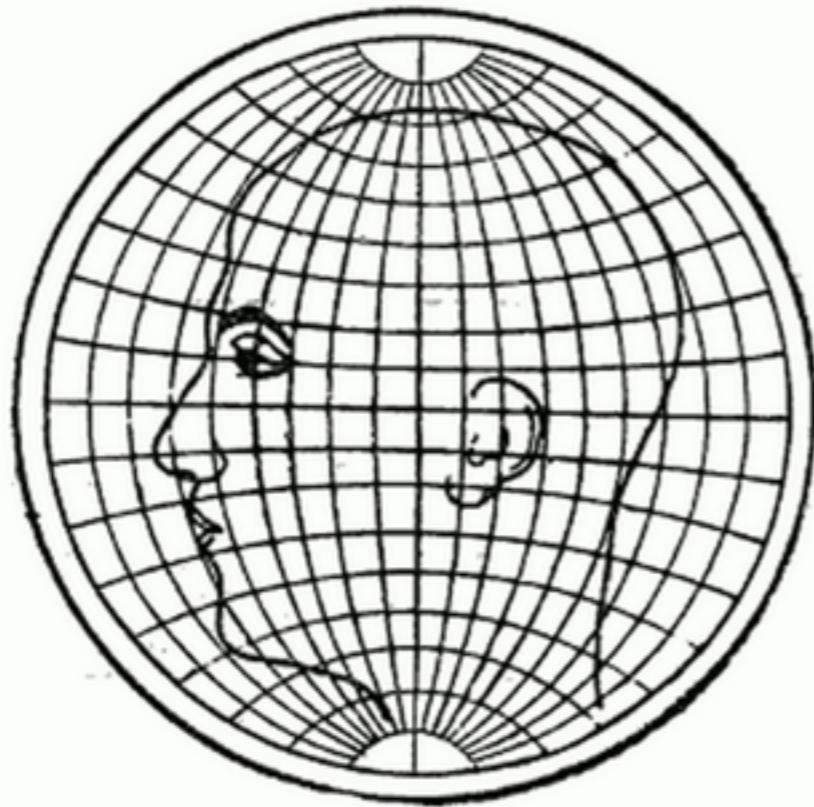


FIG. 42.—Man's head drawn on globular projection.

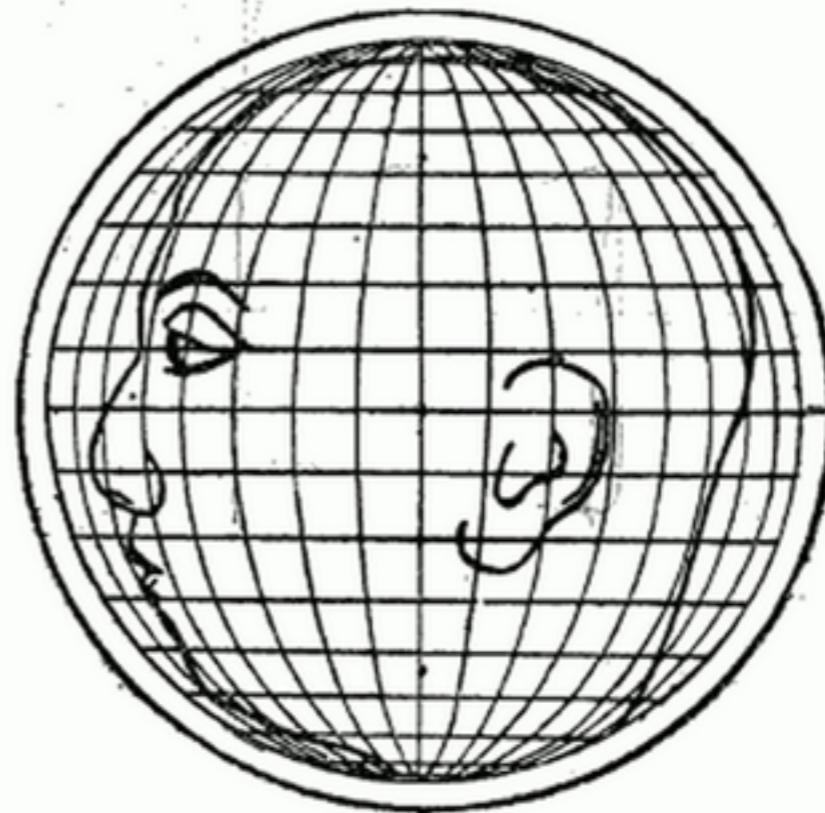


FIG. 43.—Man's head plotted on orthographic projection.

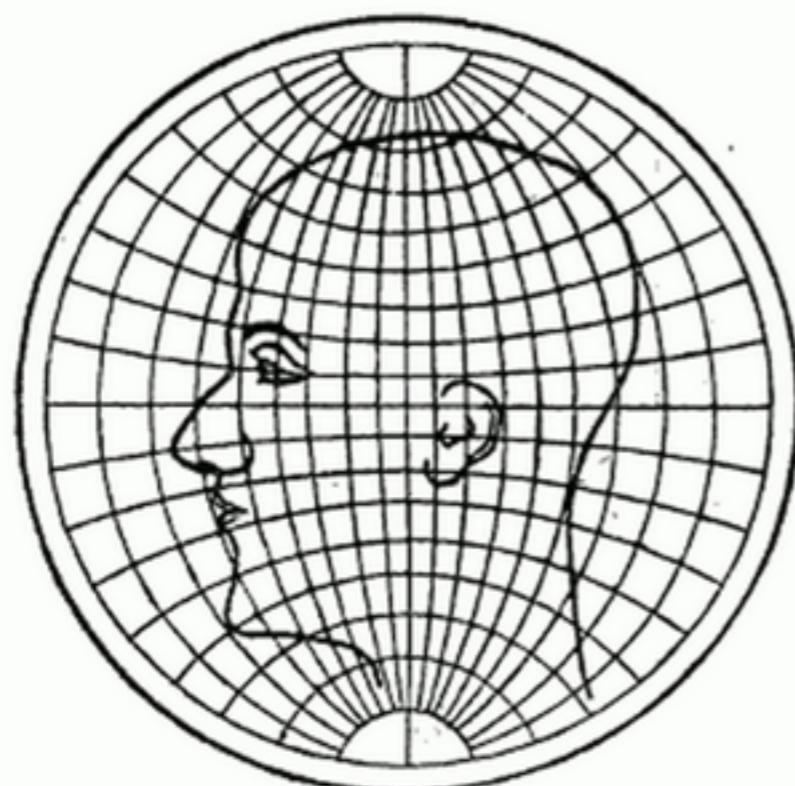


FIG. 44.—Man's head plotted on stereographic projection.

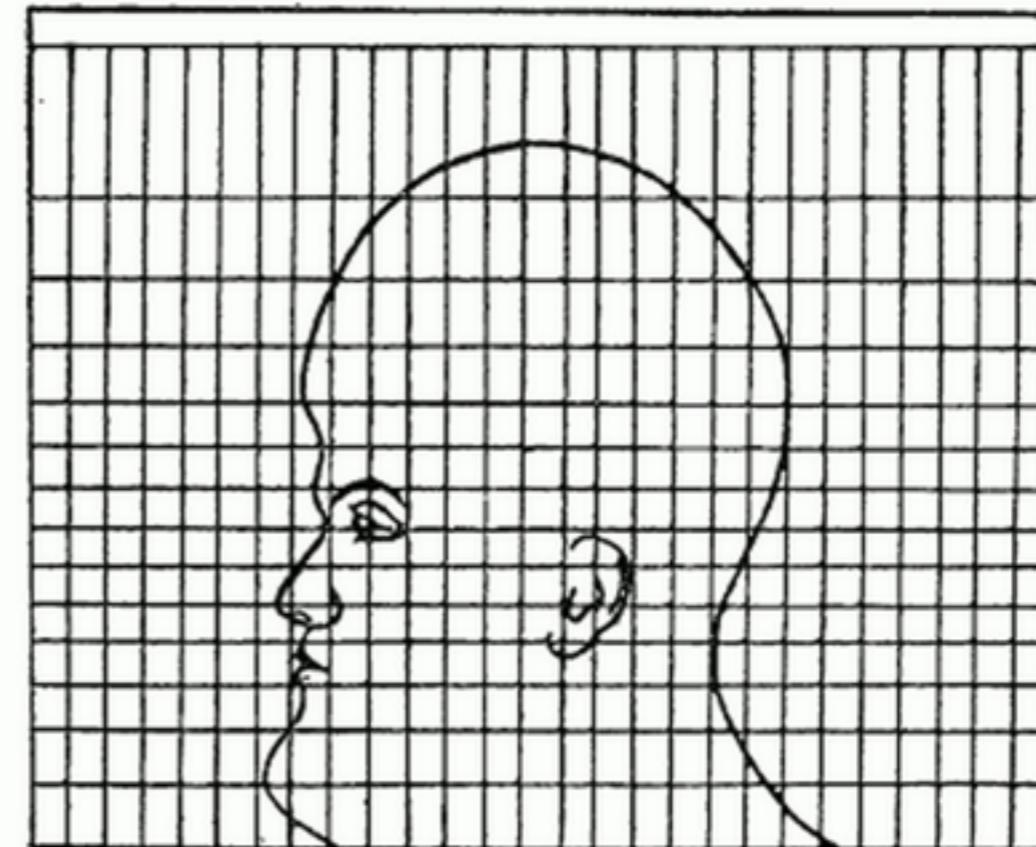
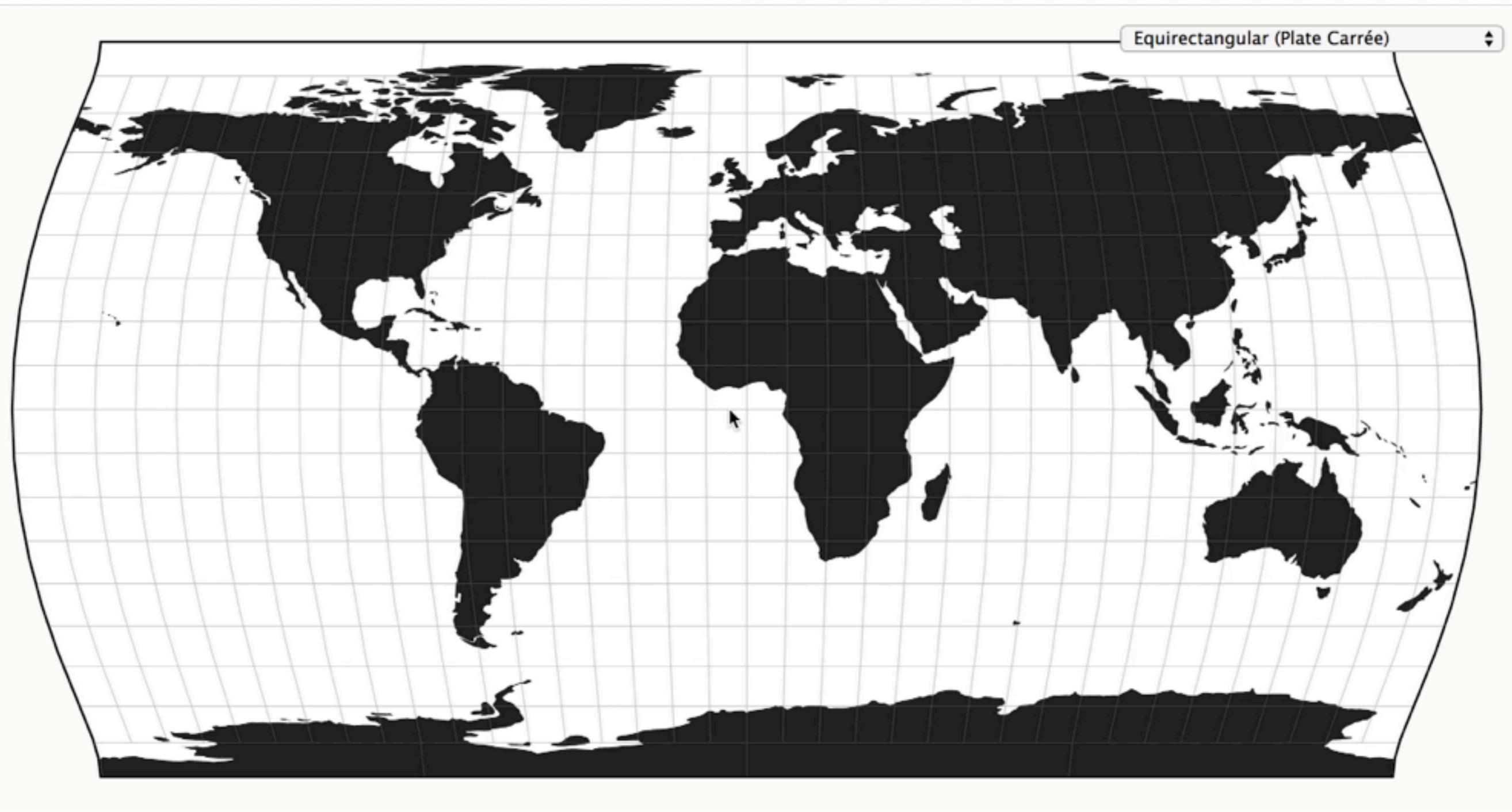
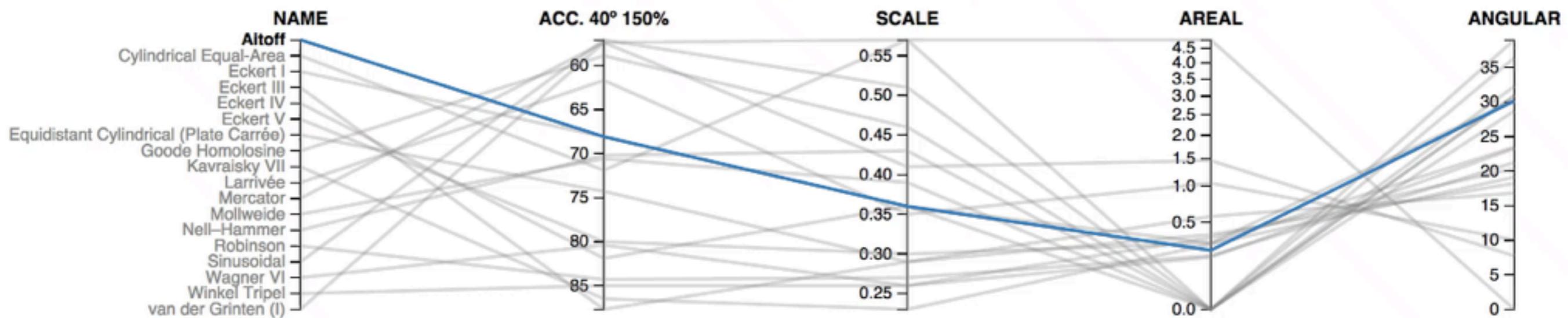
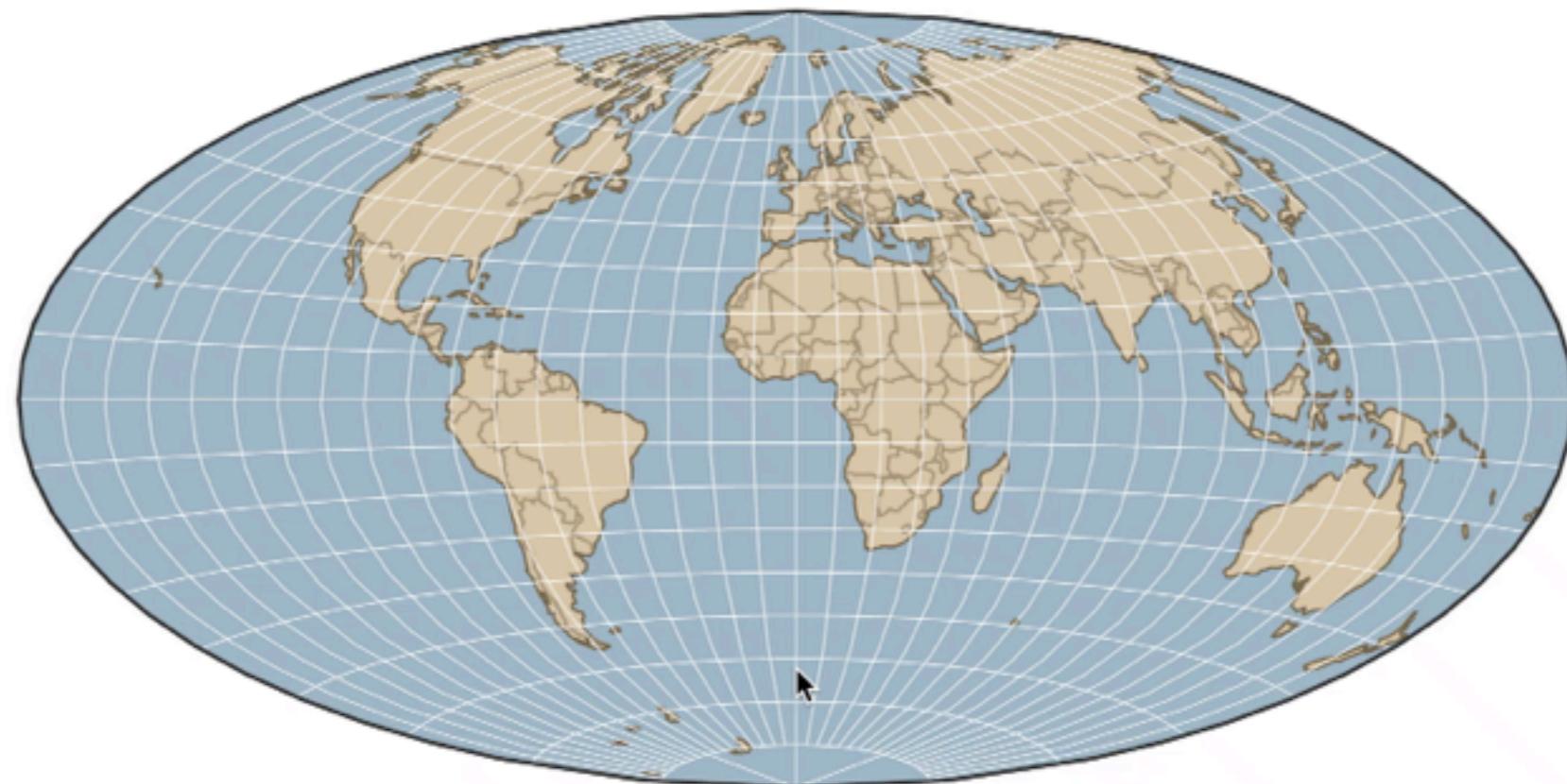


FIG. 45.—Man's head plotted on Mercator projection.

# Map Projection Demo



# Map Projection Demo



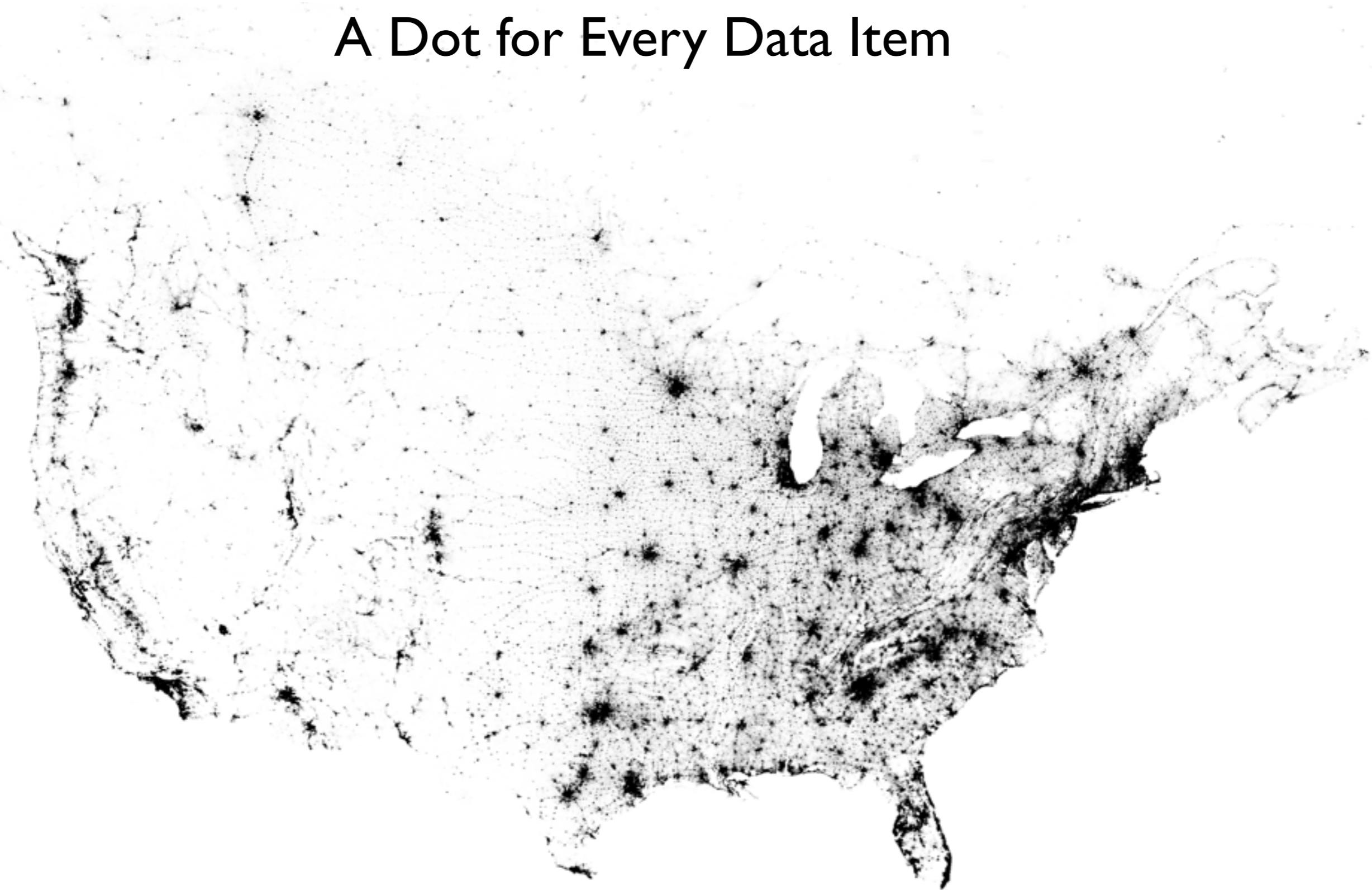
# Visual Variables for Spatial Data

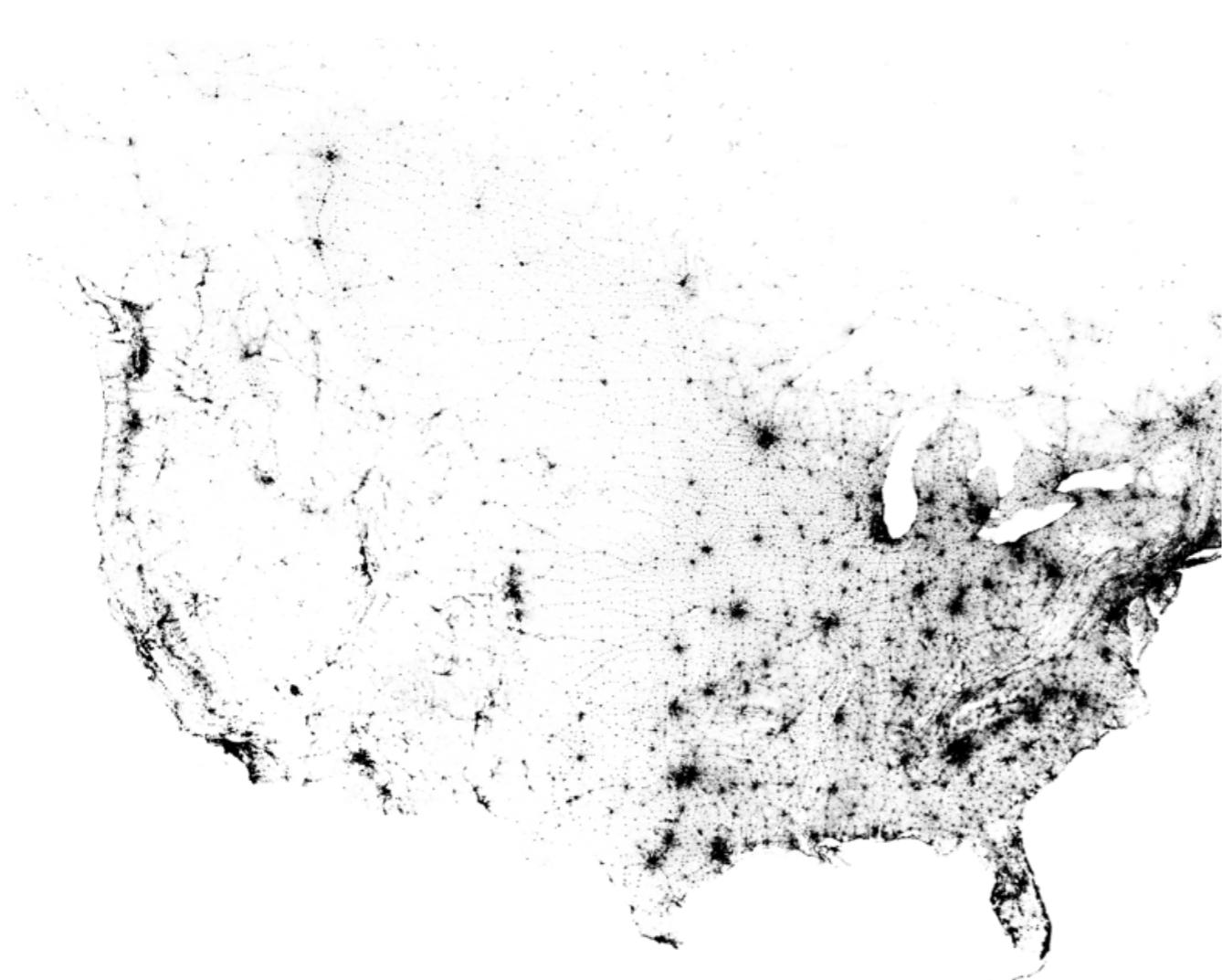
	Size	Shape	Brightness	Color	Orientation	Spacing	Perspective height	Arrangement
Point								
Linear								
Areal								

Slocum 1999

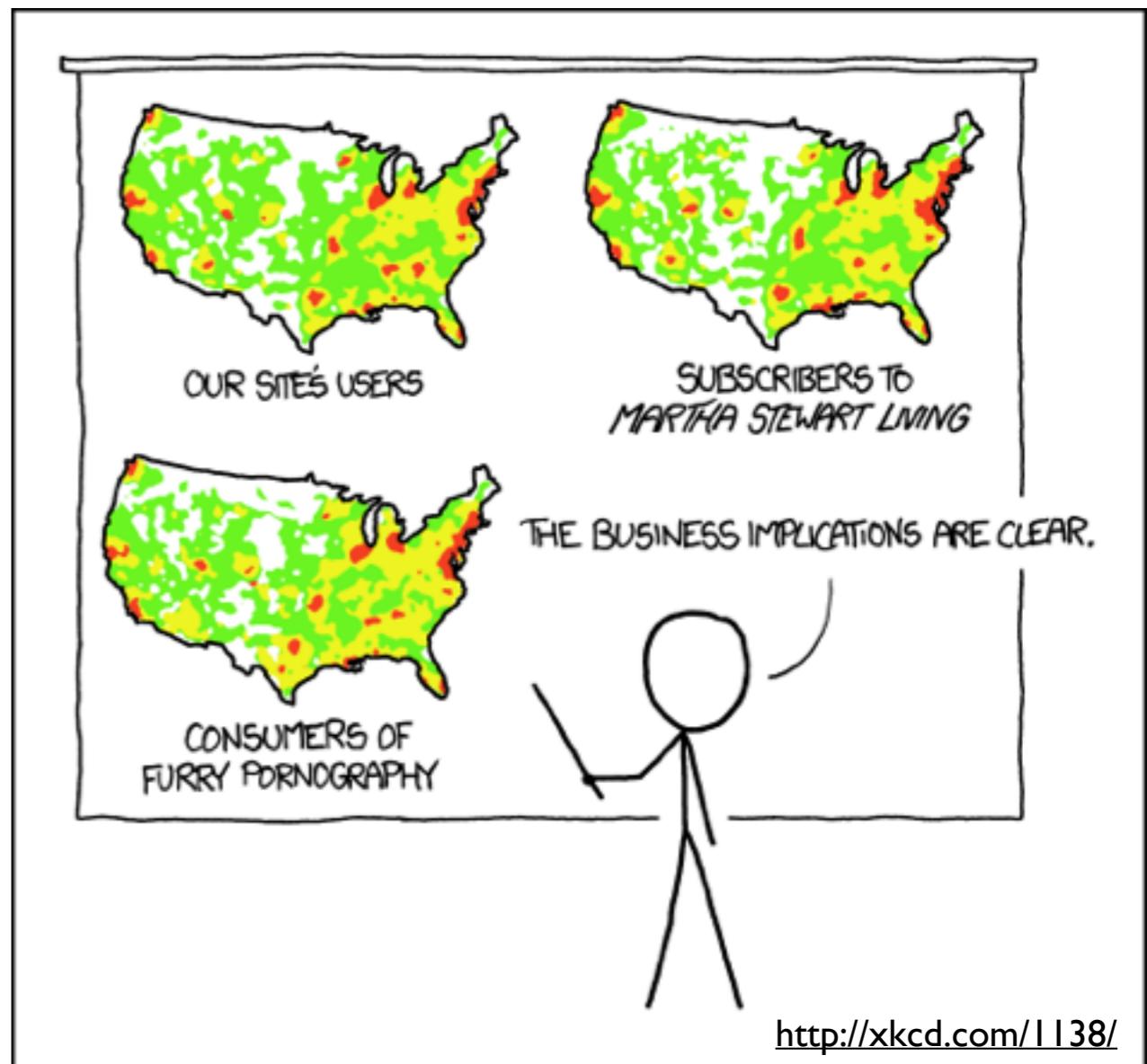
# Dot Map

A Dot for Every Data Item





# Maps can lie, too!



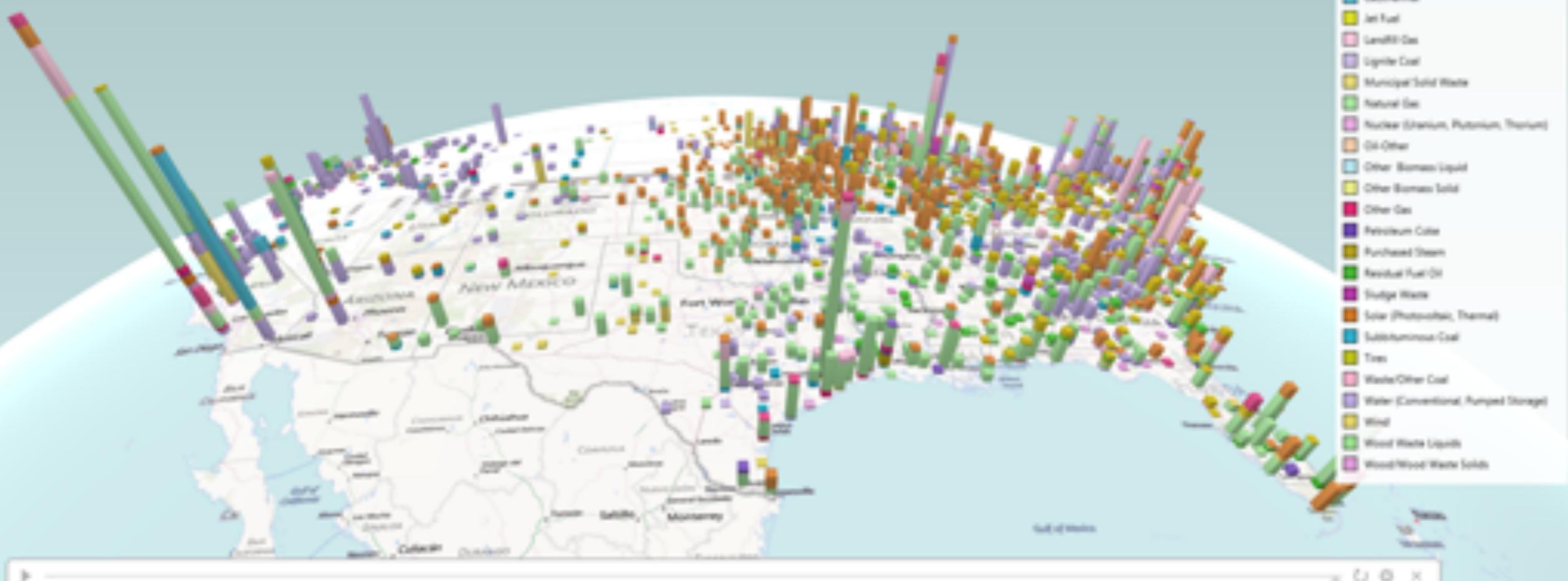
PET PEEVE #208:  
GEOGRAPHIC PROFILE MAPS WHICH ARE  
BASICALLY JUST POPULATION MAPS

## Power Stations across the US 1900-2008

The growth of power stations across the US from 1900 to 2008 and their energy sources

12/1/2008 12:00 AM

Energy Source
#icks
Agriculture Crop
Anthracite Coal, Bituminous Coal
Black Liquor
Blast Furnace Gas
Diesel Fuel Oil
Geothermal
Jet Fuel
Liquid Gas
Lignite Coal
Municipal Solid Waste
Natural Gas
Nuclear (Uranium, Plutonium, Thorium)
Oil - Other
Other Biomass/Liquid
Other Biomass/Solid
Other Gas
Petroleum Coke
Purchased Steam
Residual Fuel Oil
Ridge Waste
Solar (Photovoltaic, Thermal)
Subbituminous Coal
Tires
Waste/Oil/Gas
Water (Conventional, Pumped Storage)
Wind
Wood/Wood Liquids
Wood/Wood Waste Solids



READY

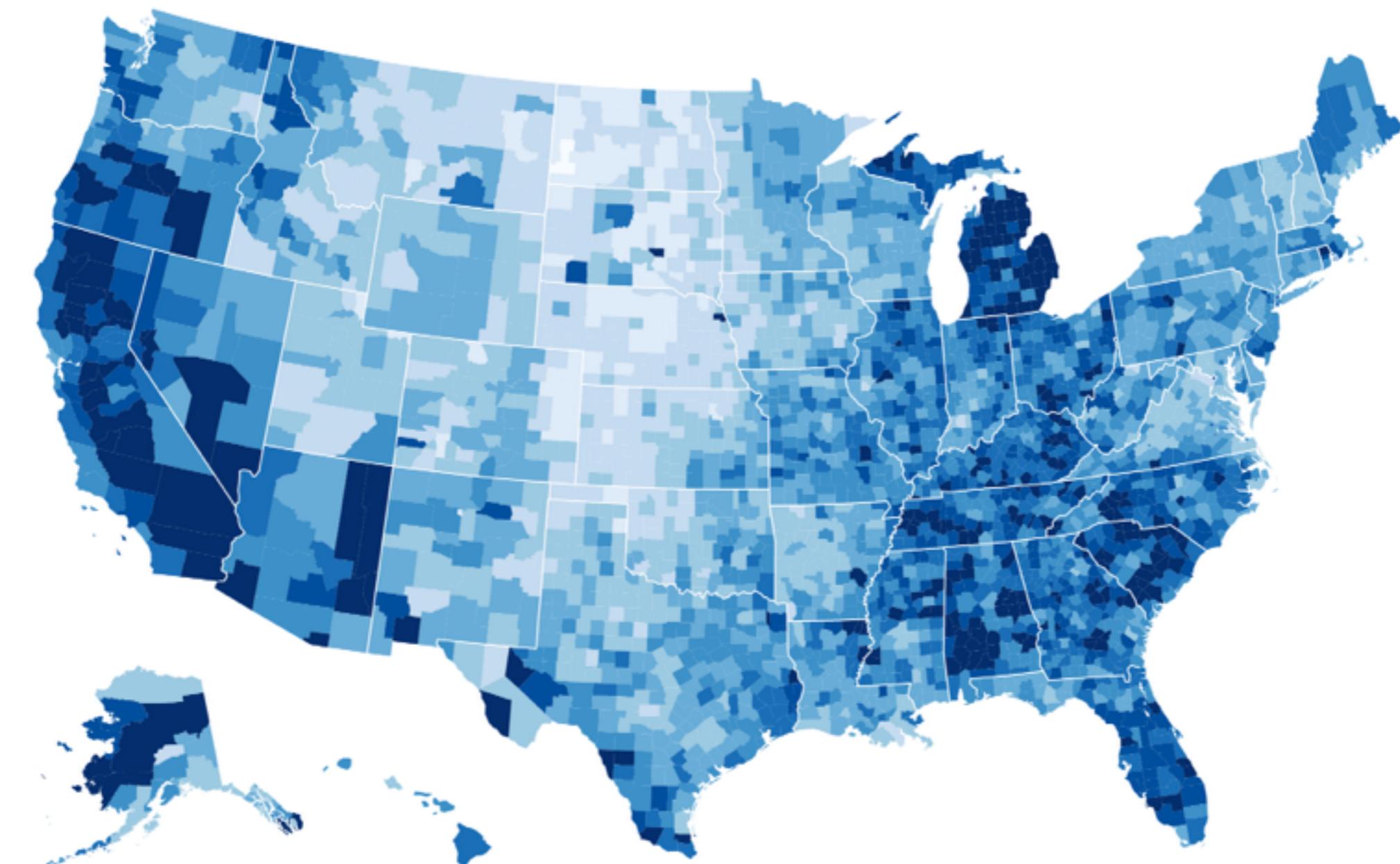
PROCESSING 782 OF 1100 VALUES

8.000 Times

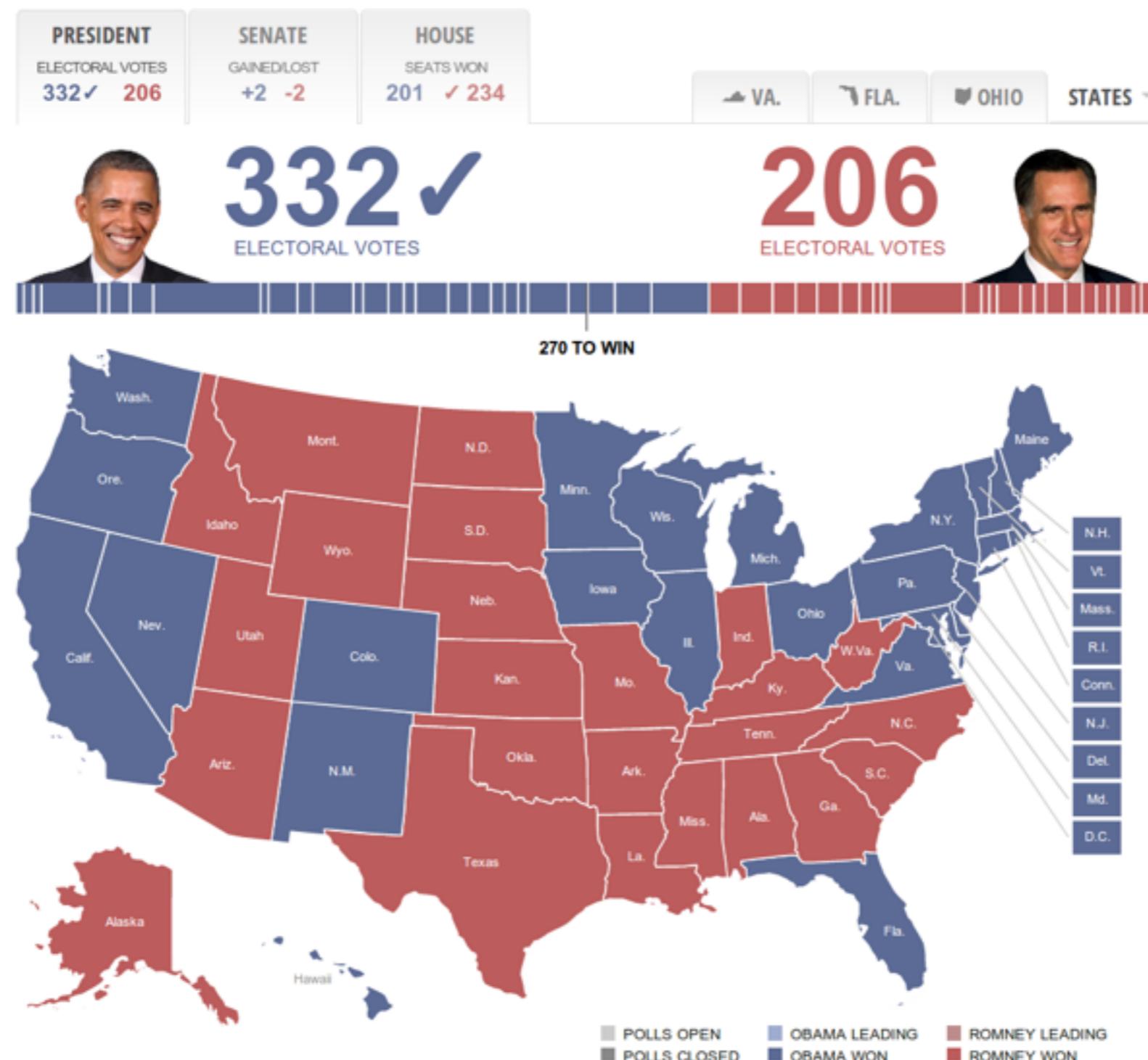
# Microsoft GeoFlow – Part of Excel 2013

# Choropleth Map

Attribute uniformly distributed in region



# Misleading Coropleth Map



# Better Version by NYT

## In a Decisive Victory, Obama Reshapes the Electoral Map

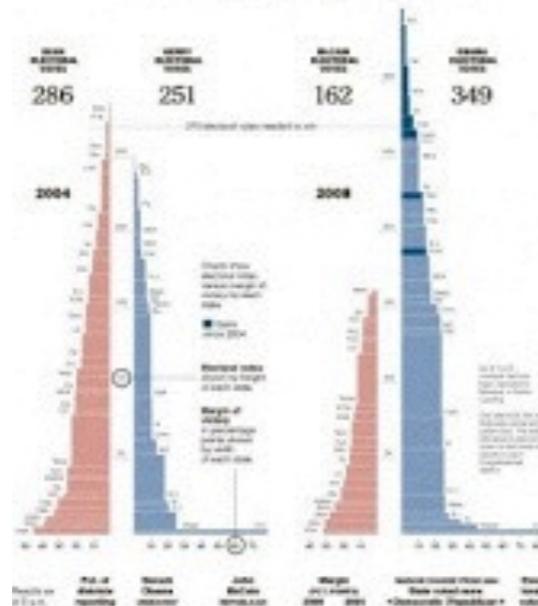
Barack Obama's historic win, with at least 349 electoral votes to John McCain's 162, can be attributed to his victories in several high-profile states, like Florida, Virginia and Ohio, that George W. Bush won handily in 2004.

The struggling economy, especially in core

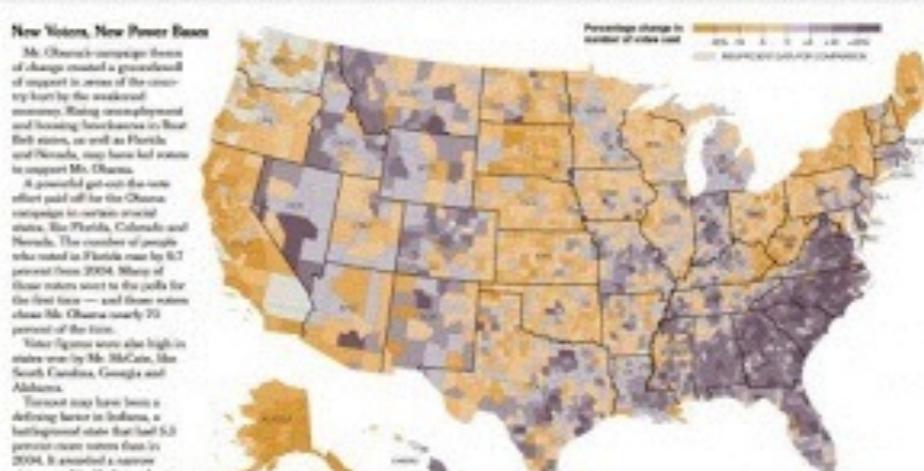
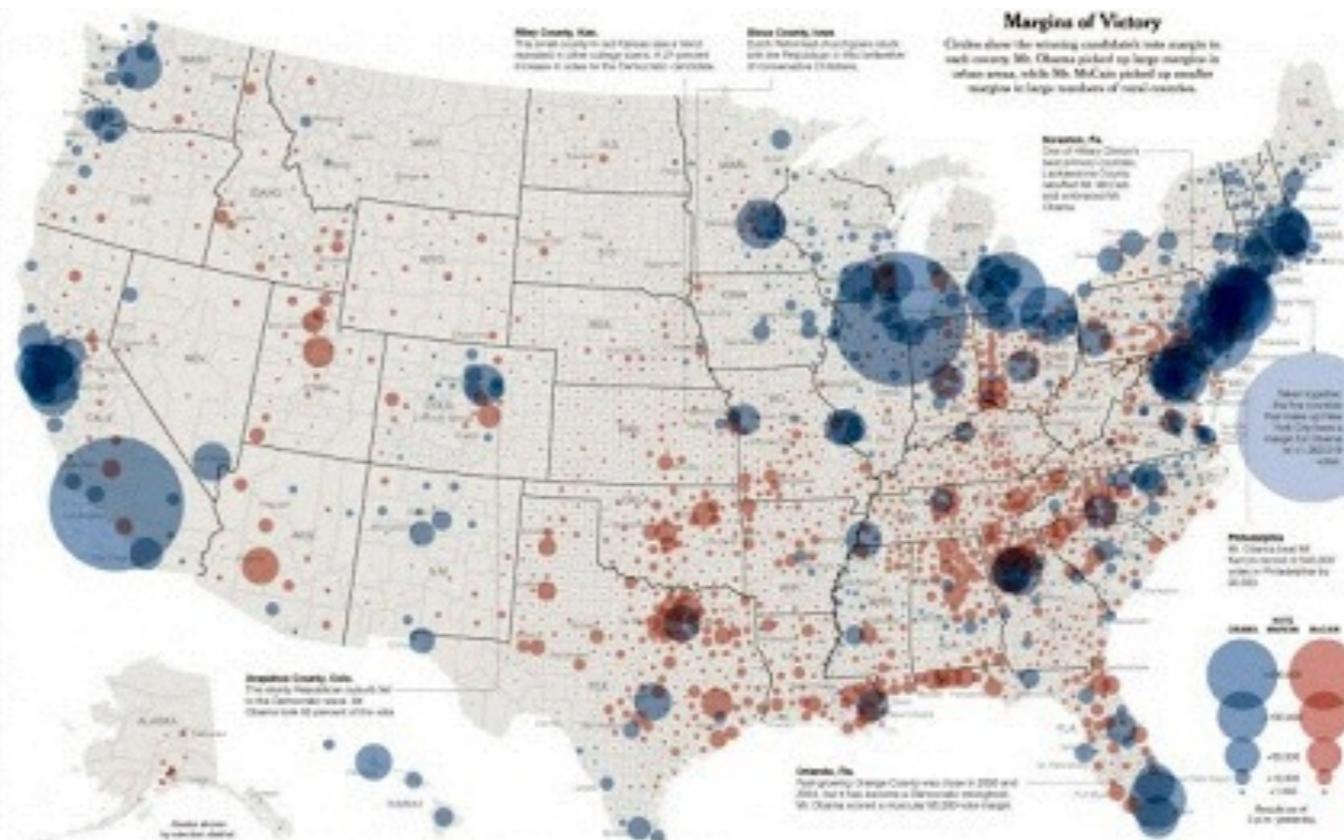
industrial states, and high numbers of new voters helped flip key swing states to blue.

Even where Mr. McCain beat Mr. Obama, he won by slimmer margins, as much of the electorate — across age, race and income lines — moved toward the Democratic Party.

By Eric Alterman, Joe Romm, Balazs Caparos, Martha Frizzell, Howard Portnoy, Fred Razzouk, Hansen Park and Andrew Zick

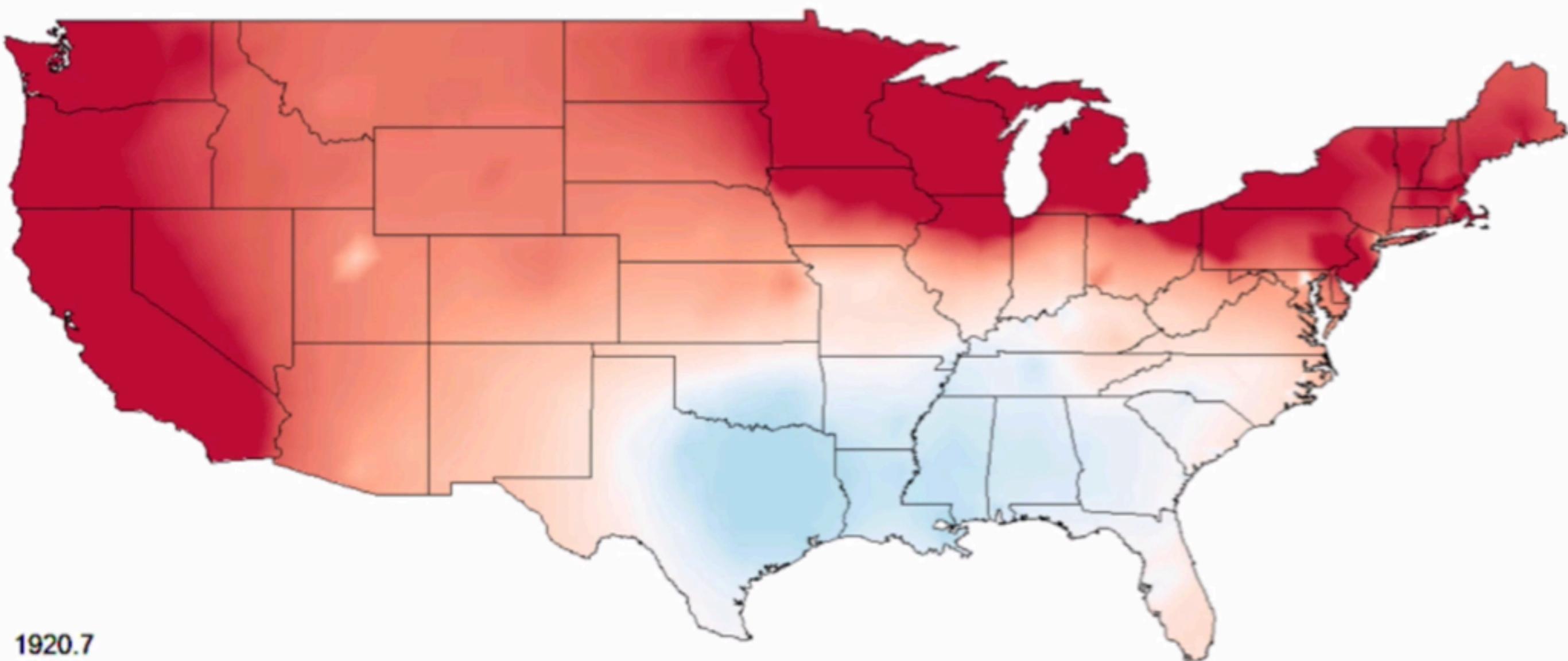


State	McCain (Red)	Obama (Blue)	Total Elect. Votes	Margin of Victory
Alabama	100%	0%	29	-100%
Alaska	100%	0%	3	-100%
Arizona	100%	0%	11	-100%
Arkansas	100%	0%	5	-100%
California	100%	0%	55	-100%
Colorado	100%	0%	9	-100%
Connecticut	100%	0%	4	-100%
Delaware	100%	0%	1	-100%
District of Columbia	100%	0%	3	-100%
Florida	100%	0%	29	-100%
Georgia	100%	0%	16	-100%
Hawaii	100%	0%	4	-100%
Idaho	100%	0%	3	-100%
Illinois	100%	0%	21	-100%
Indiana	100%	0%	11	-100%
Iowa	100%	0%	6	-100%
Kansas	100%	0%	4	-100%
Louisiana	100%	0%	6	-100%
Maine	100%	0%	4	-100%
Maryland	100%	0%	8	-100%
Massachusetts	100%	0%	11	-100%
Michigan	100%	0%	14	-100%
Minnesota	100%	0%	10	-100%
Mississippi	100%	0%	2	-100%
Missouri	100%	0%	10	-100%
Montana	100%	0%	1	-100%
Nevada	100%	0%	3	-100%
New Hampshire	100%	0%	4	-100%
New Jersey	100%	0%	14	-100%
New Mexico	100%	0%	3	-100%
New York	100%	0%	29	-100%
North Carolina	100%	0%	15	-100%
North Dakota	100%	0%	1	-100%
Ohio	100%	0%	16	-100%
Oklahoma	100%	0%	3	-100%
Oregon	100%	0%	4	-100%
Pennsylvania	100%	0%	21	-100%
Rhode Island	100%	0%	2	-100%
South Carolina	100%	0%	6	-100%
South Dakota	100%	0%	1	-100%
Tennessee	100%	0%	11	-100%
Texas	100%	0%	34	-100%
Utah	100%	0%	3	-100%
Vermont	100%	0%	1	-100%
Virginia	100%	0%	13	-100%
Washington	100%	0%	11	-100%
West Virginia	100%	0%	5	-100%
Wisconsin	100%	0%	10	-100%
Wyoming	100%	0%	1	-100%
Total	550,000,000 (50%)	500,000,000 (47%)	1,050,000,000	-100%



# Isarithmic Map

Color coding continuous phenomena



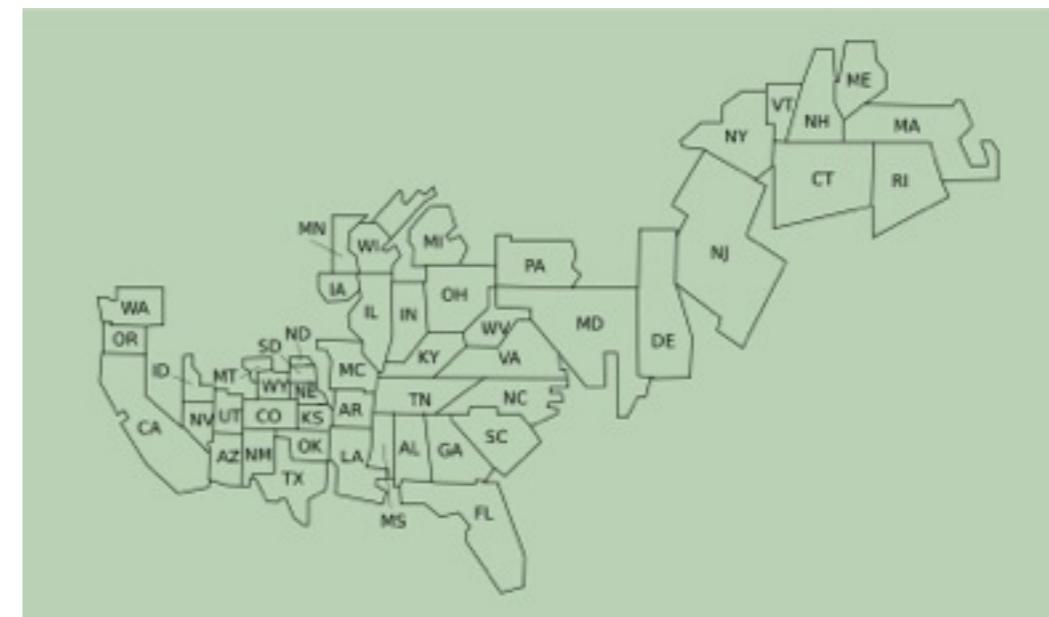
1920.7

# Cartograms

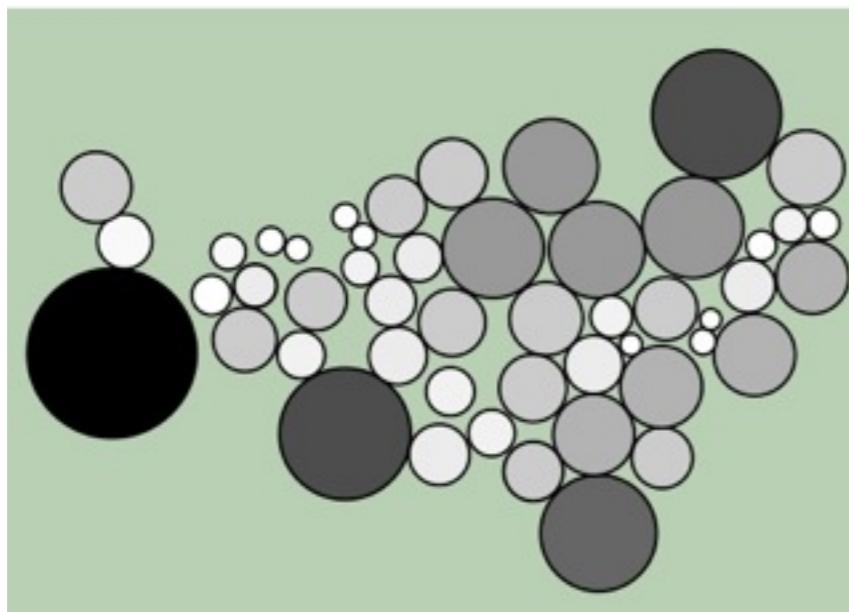
Size of region scaled to attribute value



Noncontinuous cartogram



Noncontiguous cartogram

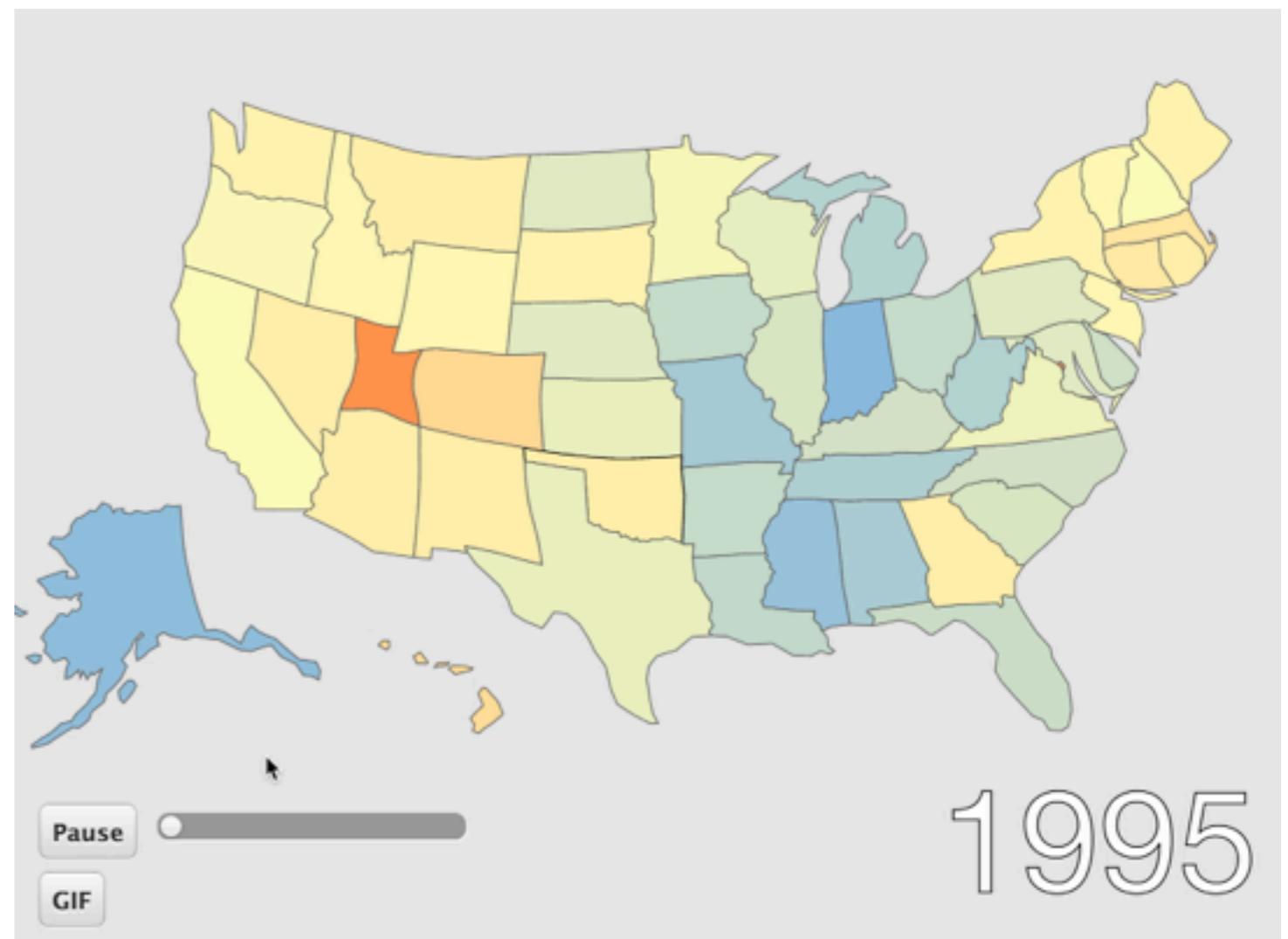
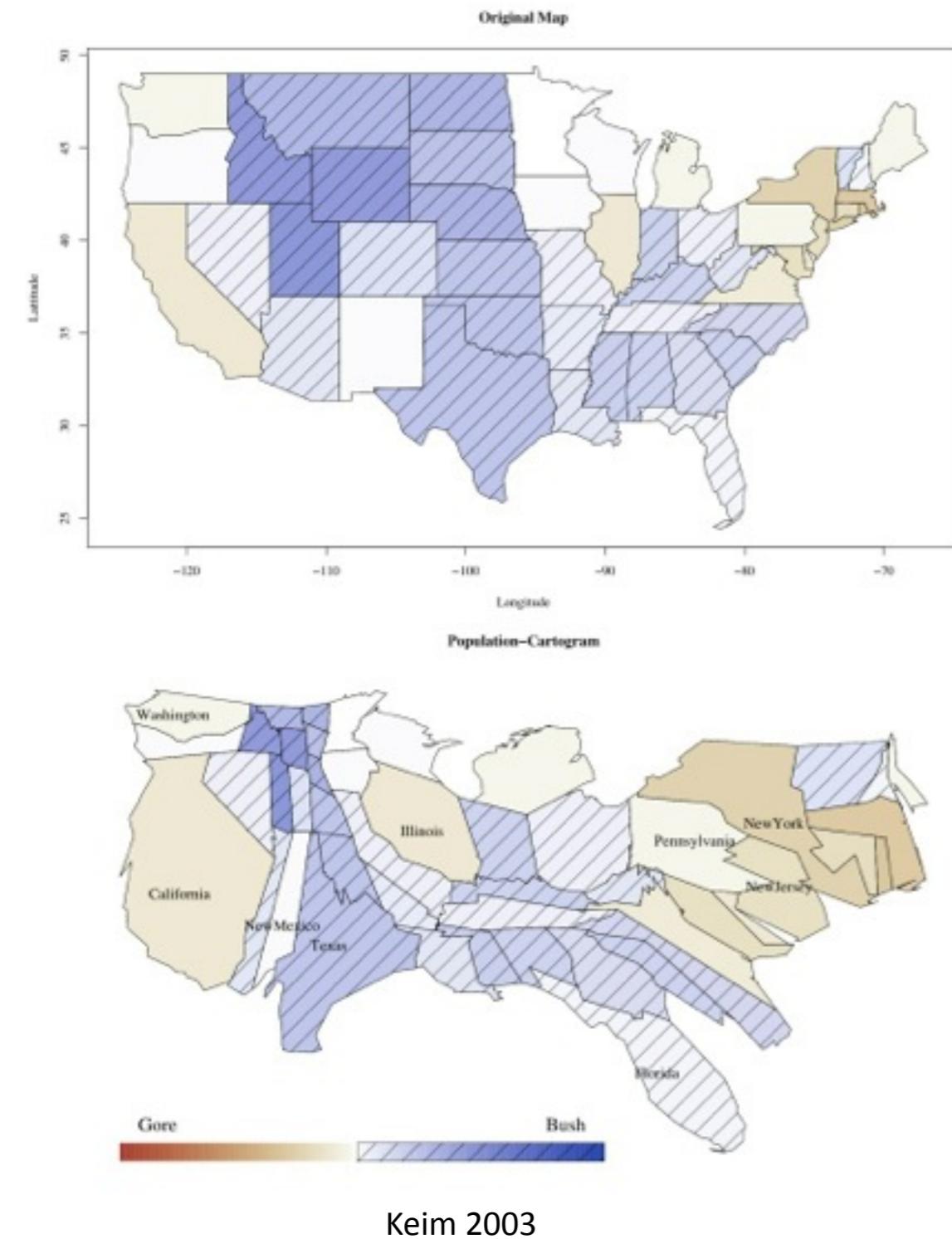


Circular cartogram

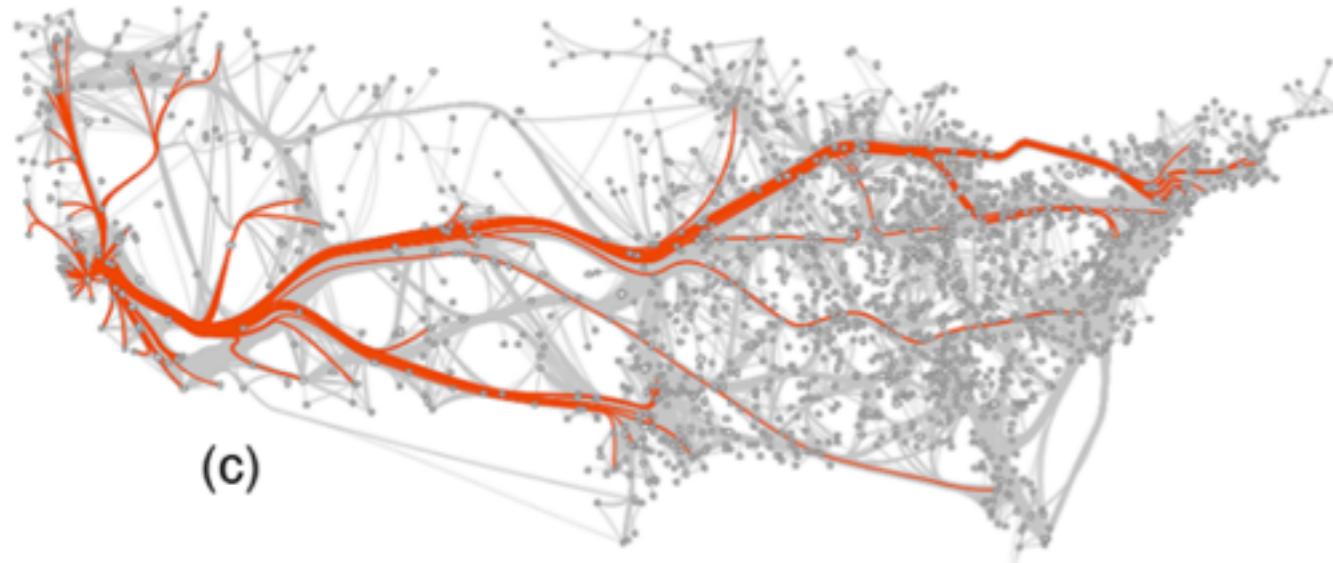


Continuous cartogram

# Continuous Cartogram Example



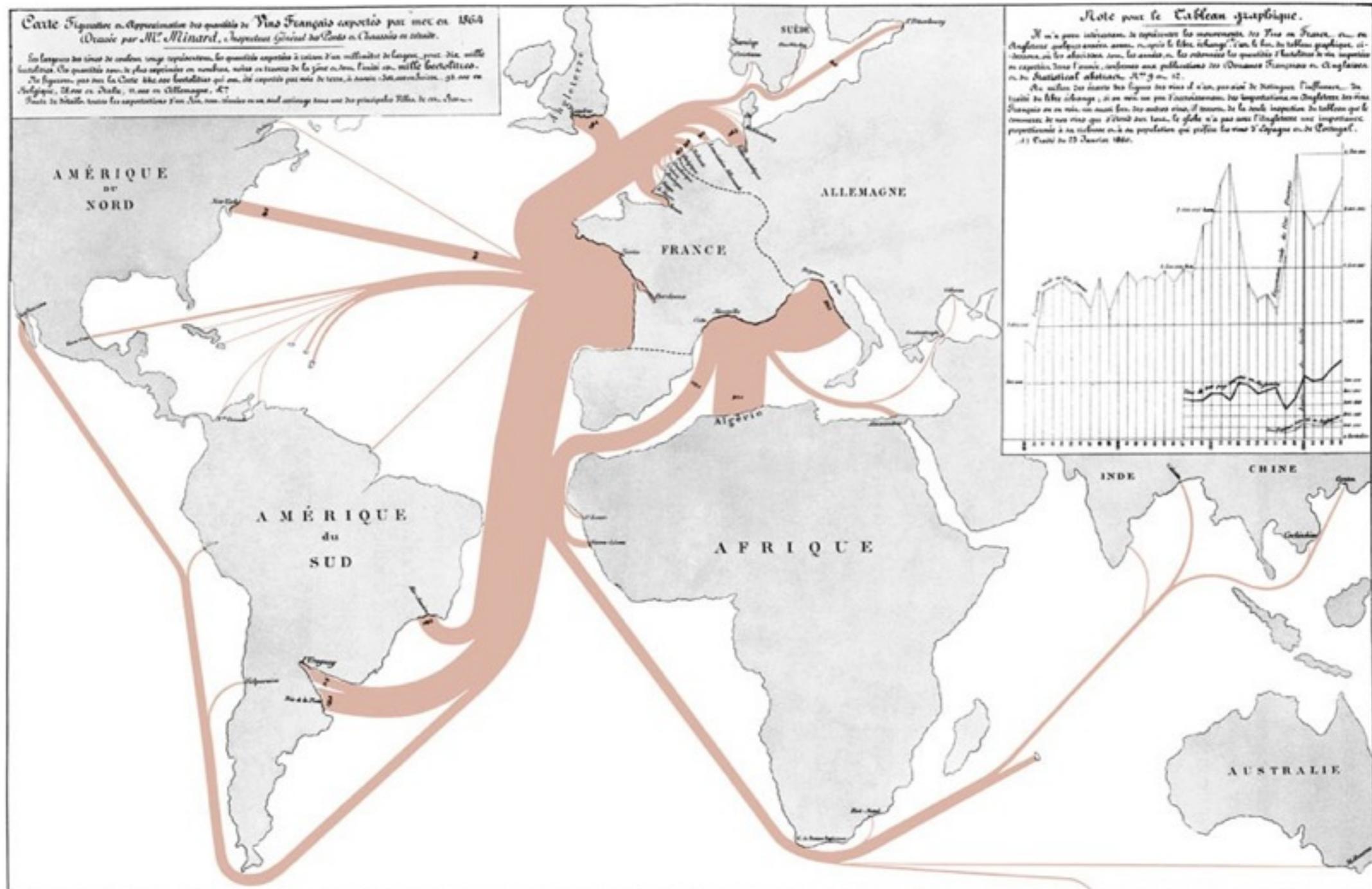
# Edges Augmented Onto Map



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

Institutional freshman facilities learning established  
freshman Henry Technology Aid well MIT 19 institution facilities learning established  
Information external See body including Financial Robert 31 Handbook North culture  
February 2001 Search Ivy January 2007 Washington Medical 2003 many Admissions  
House Libraries b Sciences Science Largest Yale Alumni Sports league expanded  
Degrees Big Science Main David Arts York United High William  
Ranking John River Game Museums Wikipedia New hockey Princeton Health  
Professional ms 42 Museum teams September Center General Teaching Brown  
Foundation dated Library October Association 100 Universities ISBN class Top Office  
Identifiers Study w School ha Graduate Century Universities Charles Bay  
South history history 10 world Times Harvards 2009 million  
system links News Mens Institute academic First 1 August 28 5 State potentially  
links News Mens Institute Cambridge page July 8 articles Yard One Billion Three  
Great among Large 18 NCAA colleges 14 also Retrieved 2 Boston College v programs universitys  
among Large 18 NCAA colleges 14 also Retrieved 2 Boston College v programs universitys  
Admission President Higher 85 Early 9 Education 2 students Hall Years 3 2010 Women Report Making  
Faculty Program containing campus 15 s 2014 2013 Art March Red athletics  
85 Early 9 Education 2 students Hall Years 3 2010 Women Report Making  
cheating several Undergraduate 16 1636 Radcliffe 4 Public Institutions became  
two staff Business Team won Court 16 1636 Radcliffe 4 Public Institutions became  
Columbia athletic Home Penn statements endowment National  
Texas 2005 expansion Canadian Medicine references Cornell 2006 Member  
Western 2005 expansion Curriculum 1936 1986 Americas Quincy Rivalry throughout  
undergraduates

Enter a URL below, or paste some text.

[http://en.wikipedia.org/wiki/Harvard\\_University](http://en.wikipedia.org/wiki/Harvard_University)

Go!

Spiral:  Archimedean  Rectangular

Scale:   $\log n$    $\sqrt{n}$    $n$

Font: Impact

5 orientations from 0 ° to 0 °



Number of words: 250

One word per line

Download: [SVG](#) | [PNG](#)

President Bush, January 29, 2002

President Obama, January 25, 2011

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

# Spark Clouds

Lee 2010

Convey trends between multiple tag clouds over time

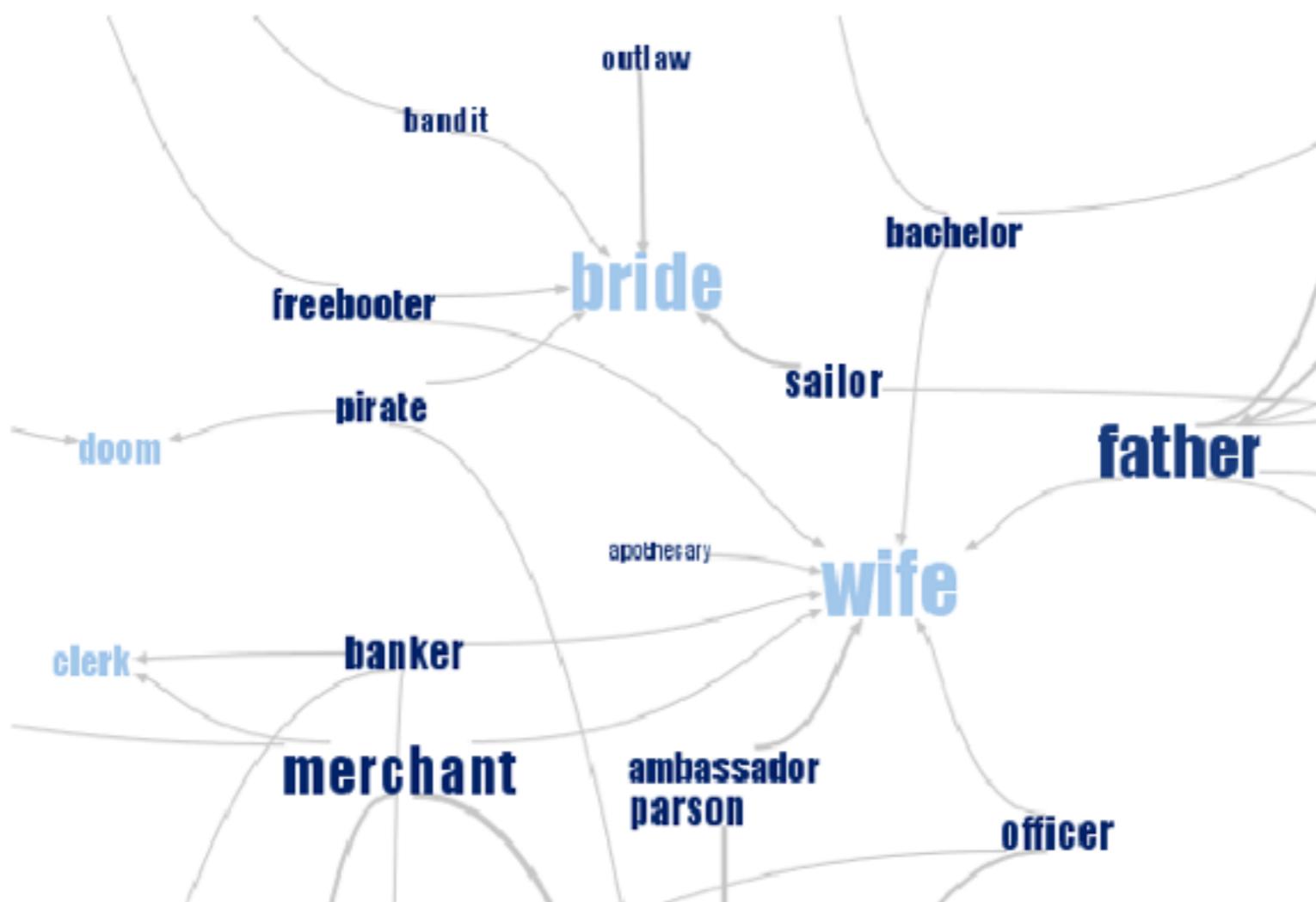


# Example: Phrase Net

van Ham et al. 2009

Visual overviews of unstructured text

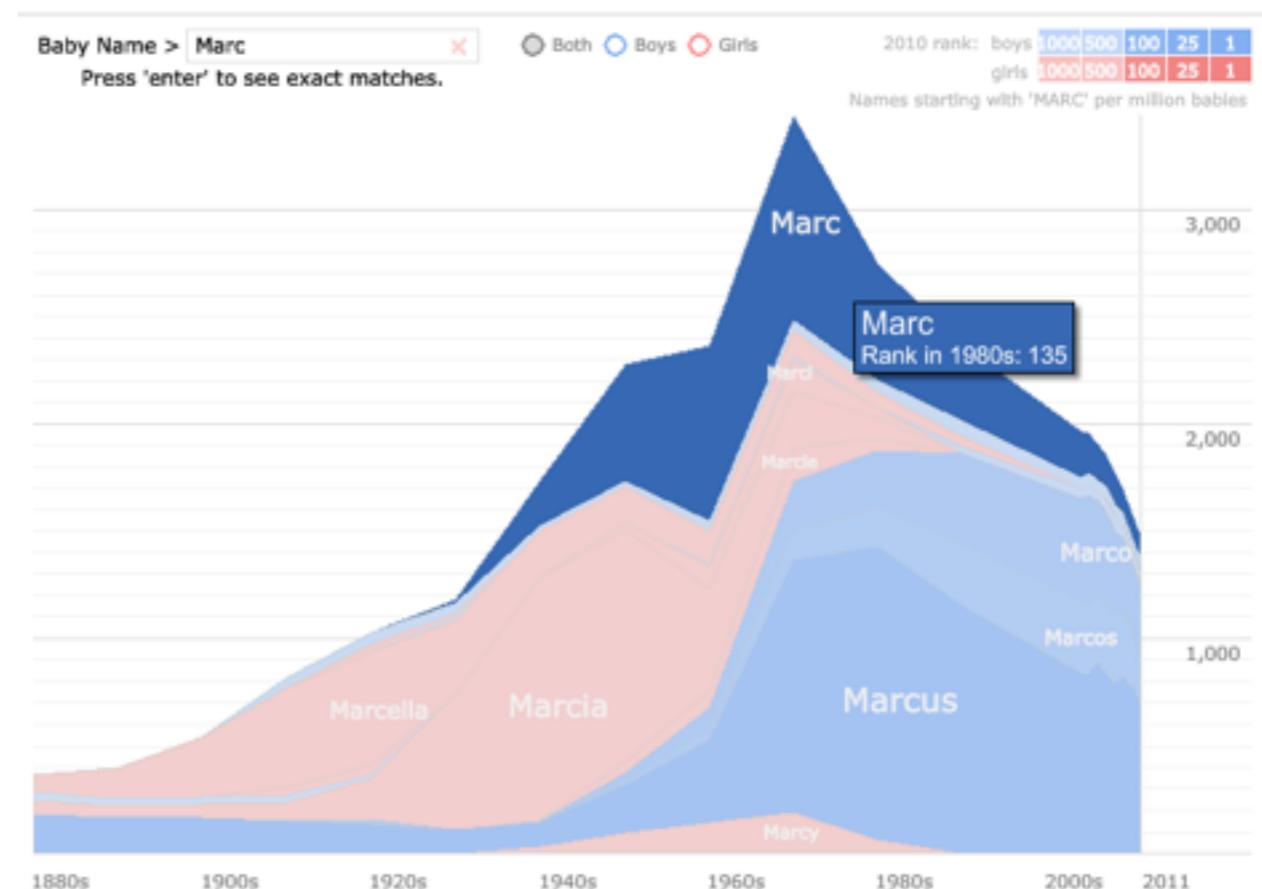
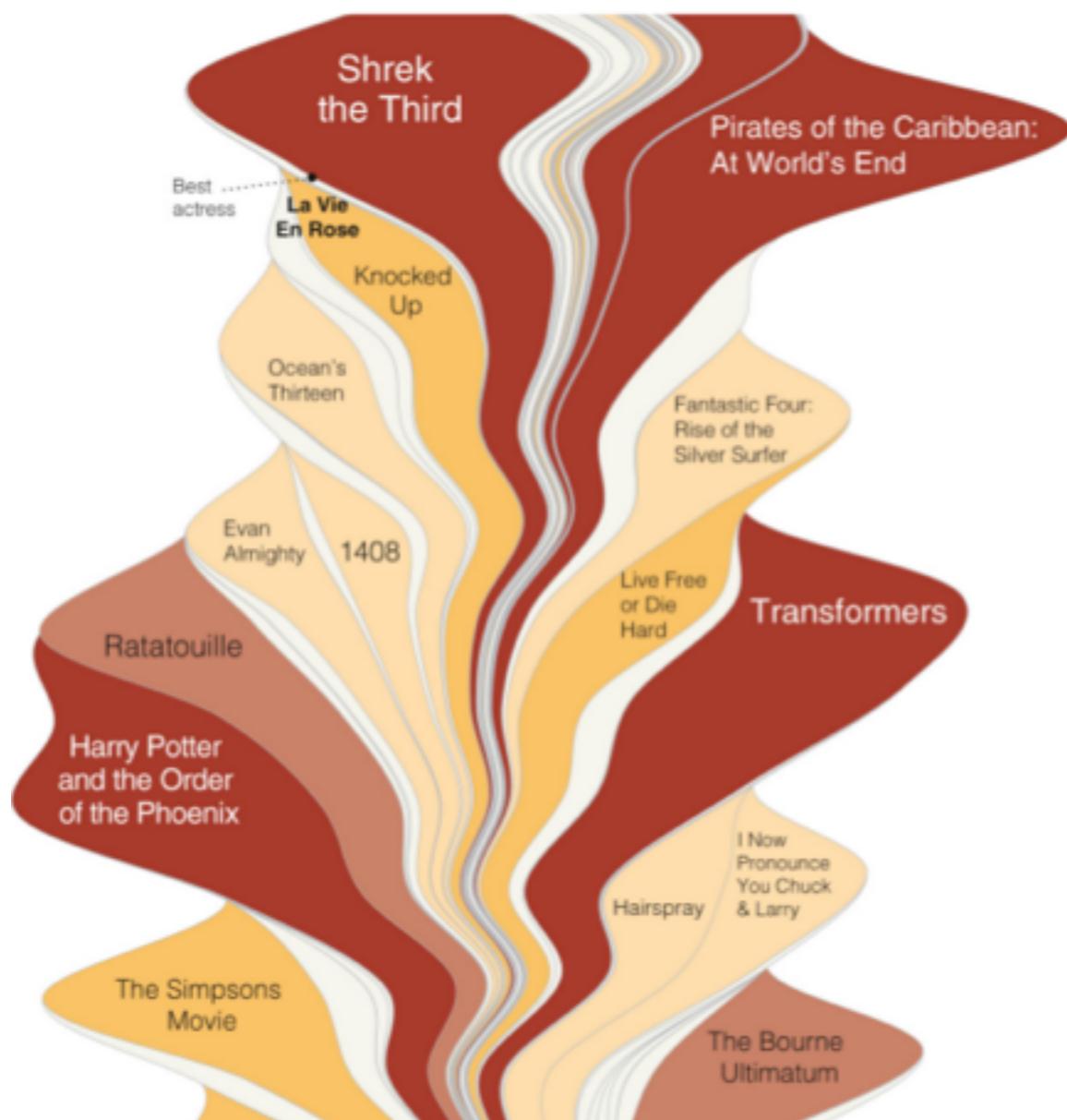
Graph; nodes = words; edges = linked by user specified relation



Relations in 18th and 19th century novels

# Theme River / Stream Graph

## Thematic changes over time. Height = frequency

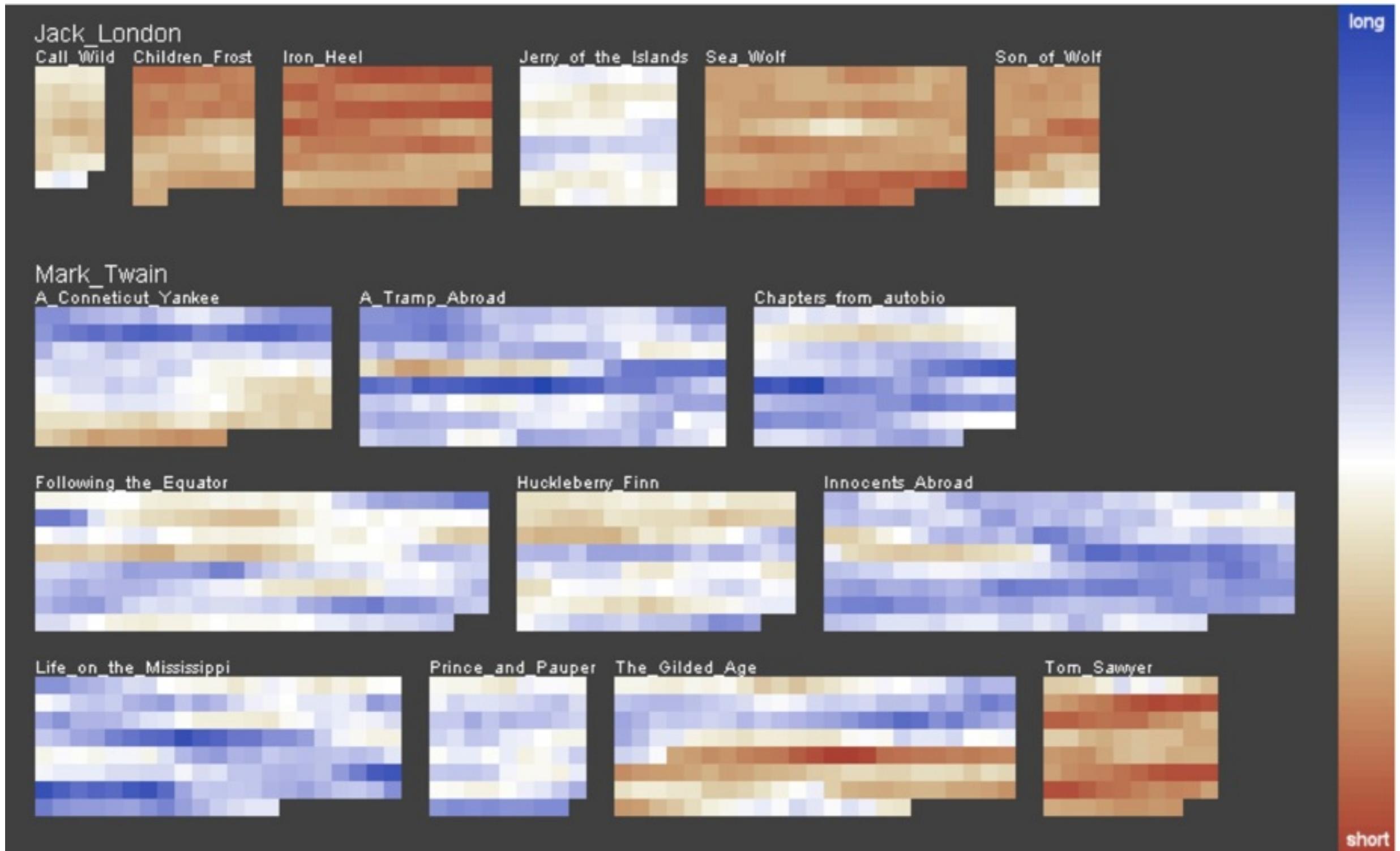


Baby Name Voyager

<http://www.babynamewizard.com/>

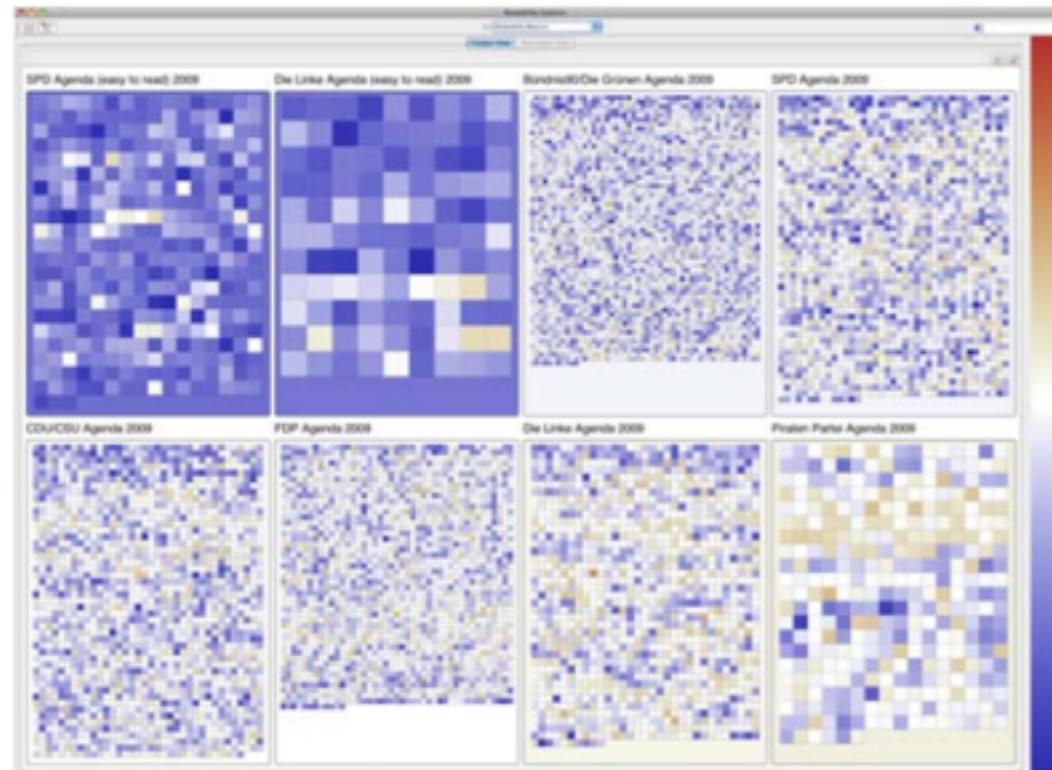
Wattenberg 2005

# Example: Literature Fingerprinting

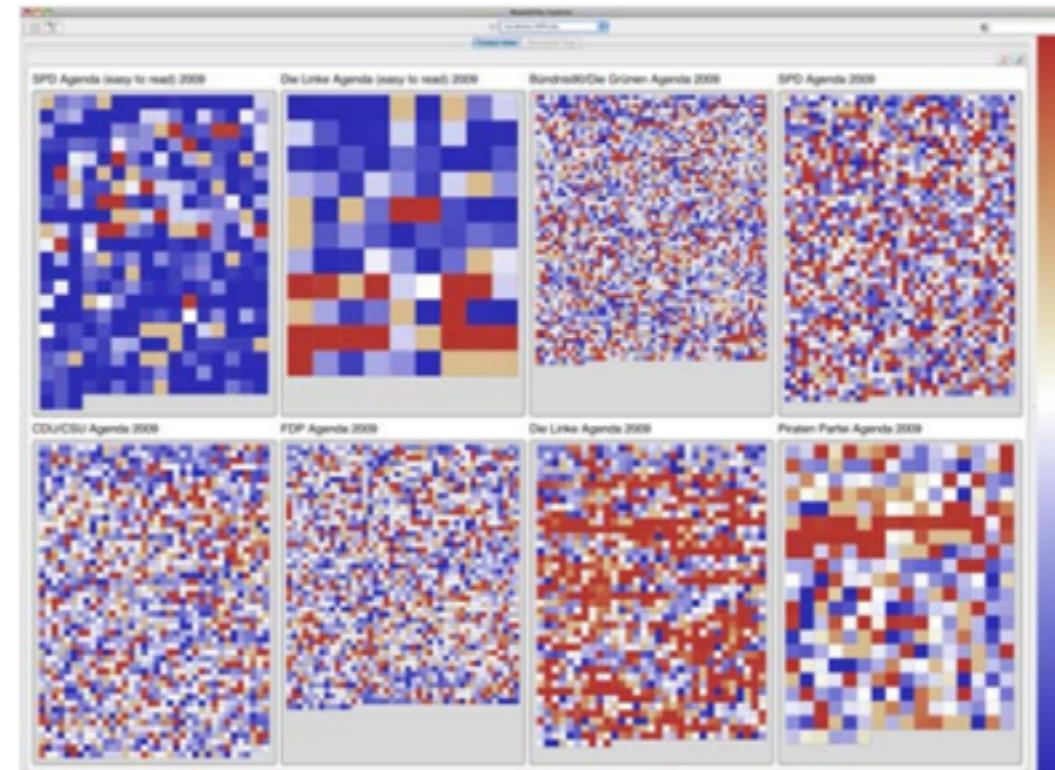


Visualize several text measures to discriminate between authors.  
Pixel = text block, Group = book, color = average sentence length

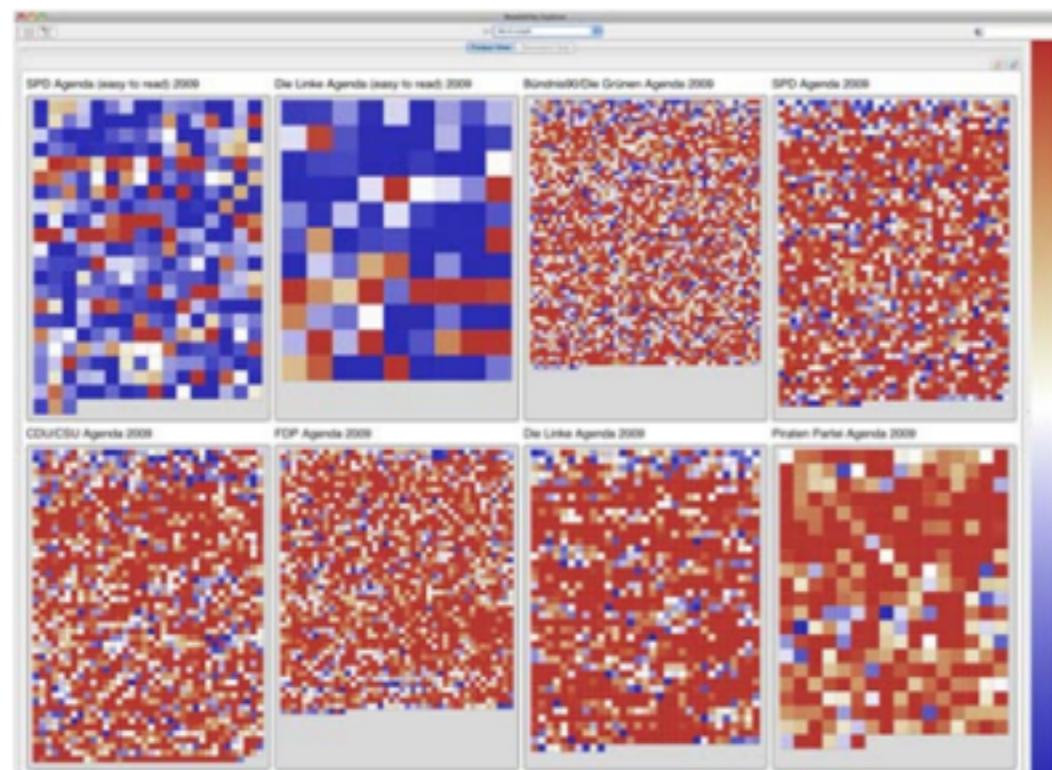
# Example: Readability Analysis



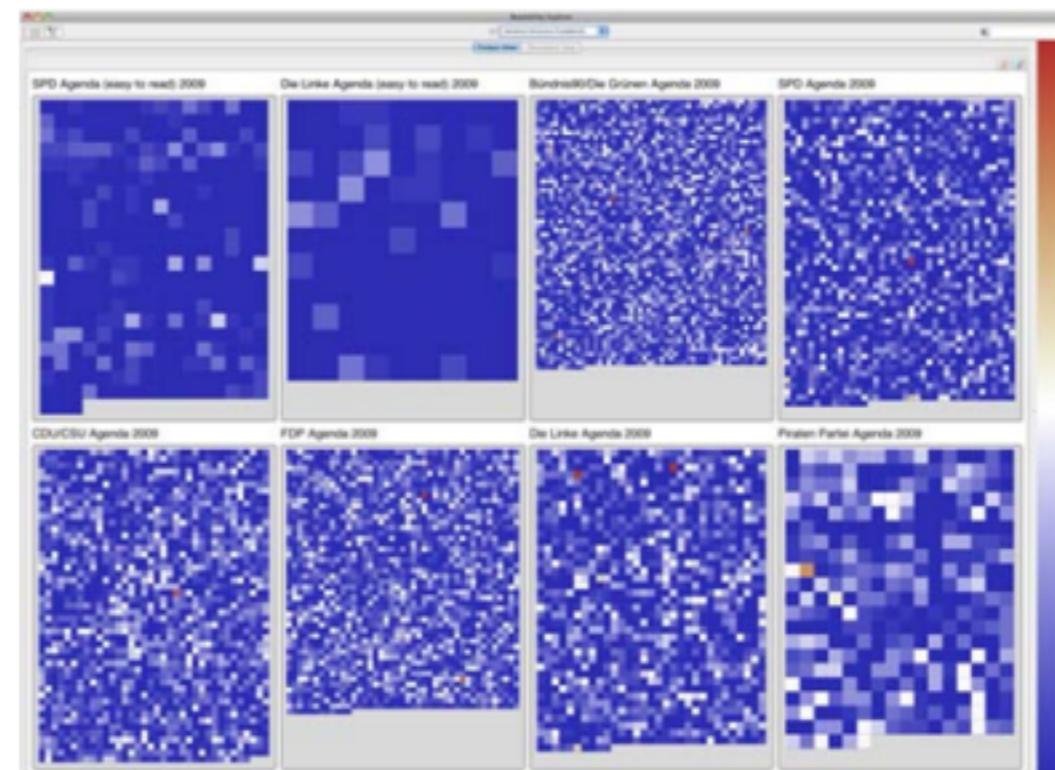
(a) Average Readability Score



(b) Feature: Vocabulary Difficulty



(c) Feature: Word Length



(d) Feature: Sentence Structure Complexity

# 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/>

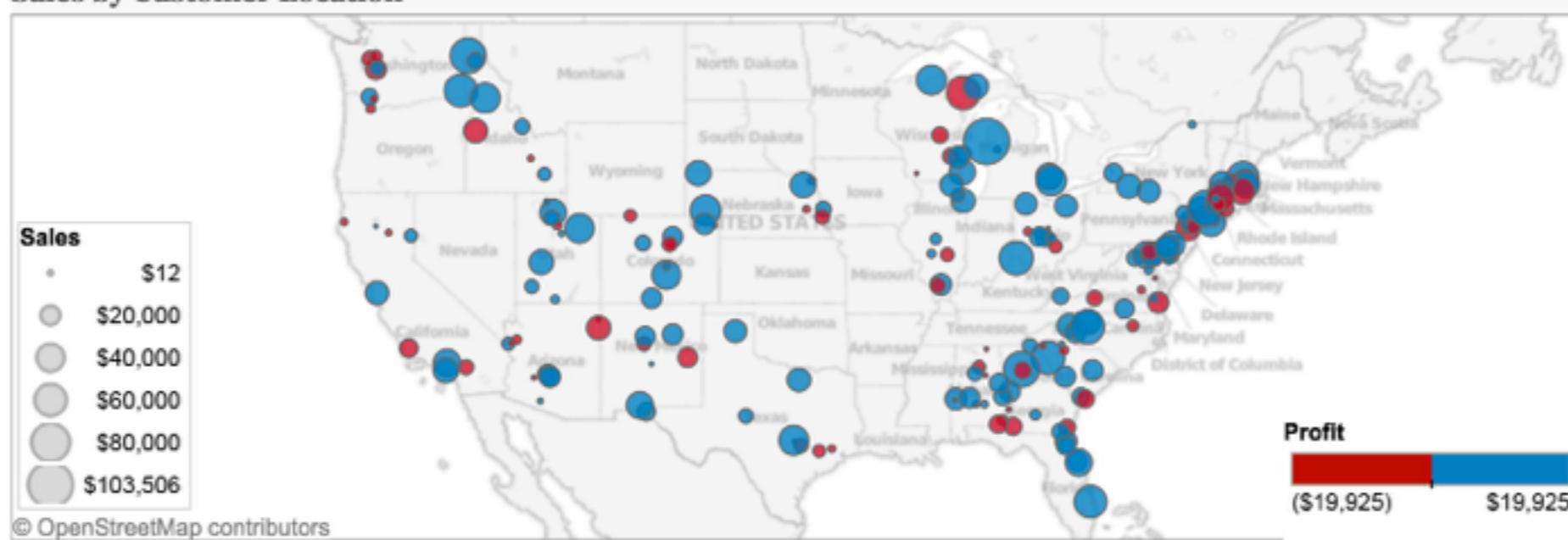
# Software



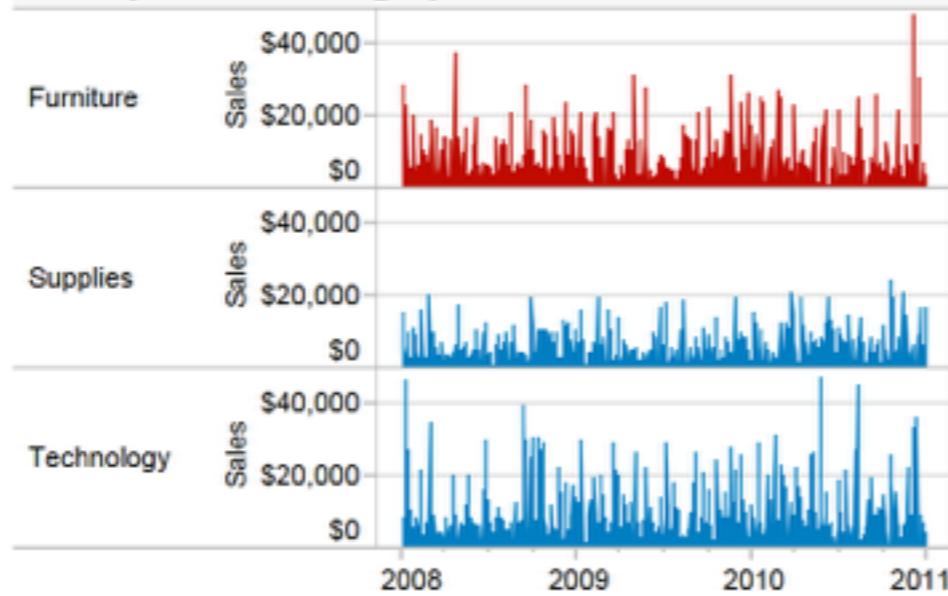
+ a b | e a u®  
S O F T W A R E

<http://www.tableausoftware.com/>

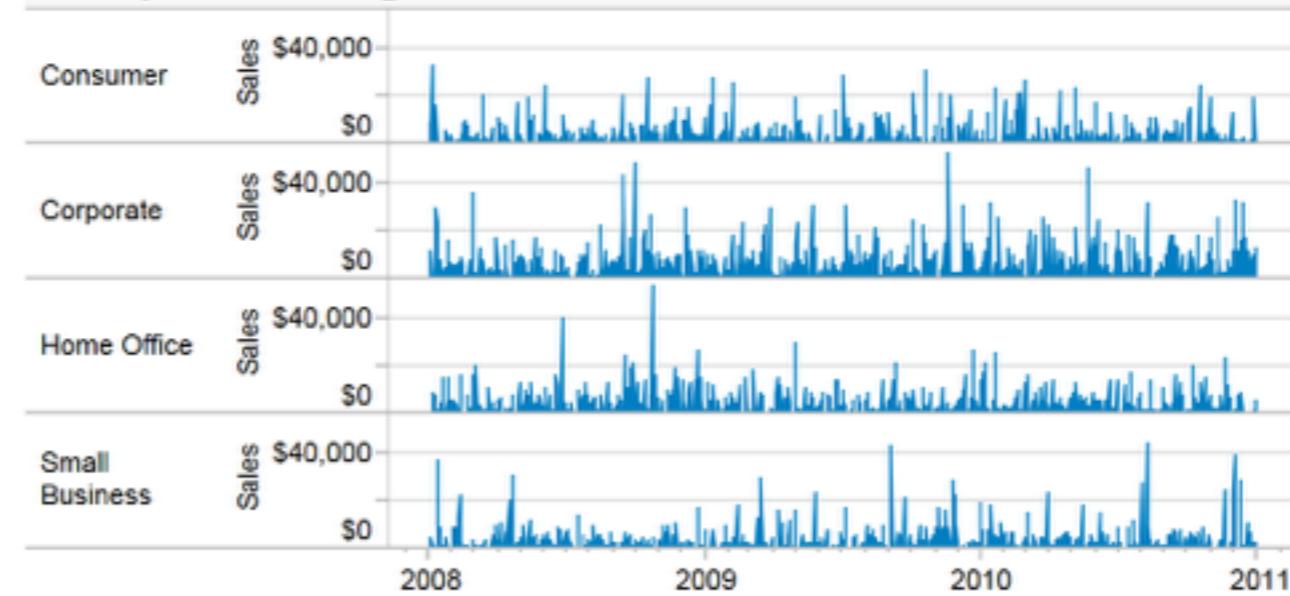
Sales by Customer Location



Sales by Product Category



Sales by Customer Segment



# Data-Driven Documents (D3.js)

<http://d3js.org>

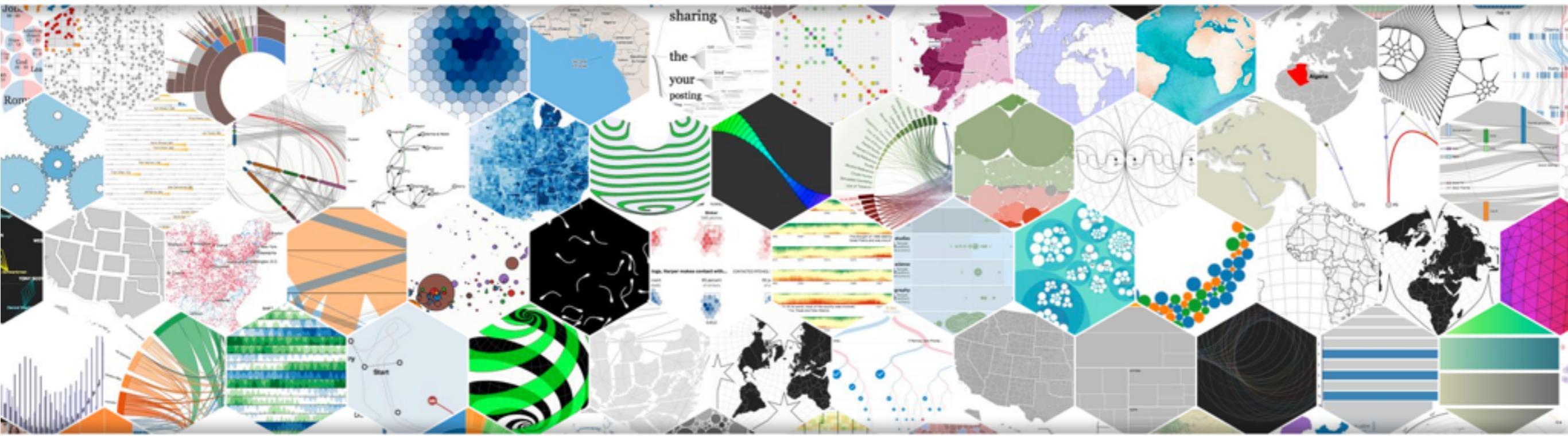
# Java Script library

# Technologies: CSS3, HTML5, SVG

## Bind arbitrary data to a Document Object Model (DOM)

Then apply data-driven transformations to the document

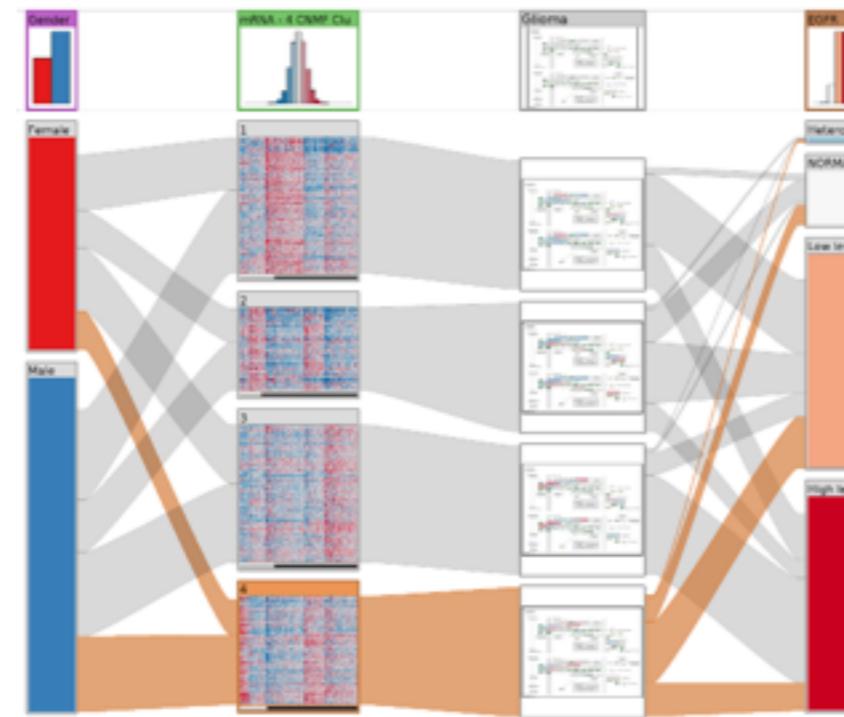
Well documented / lots of examples & tutorials / free books



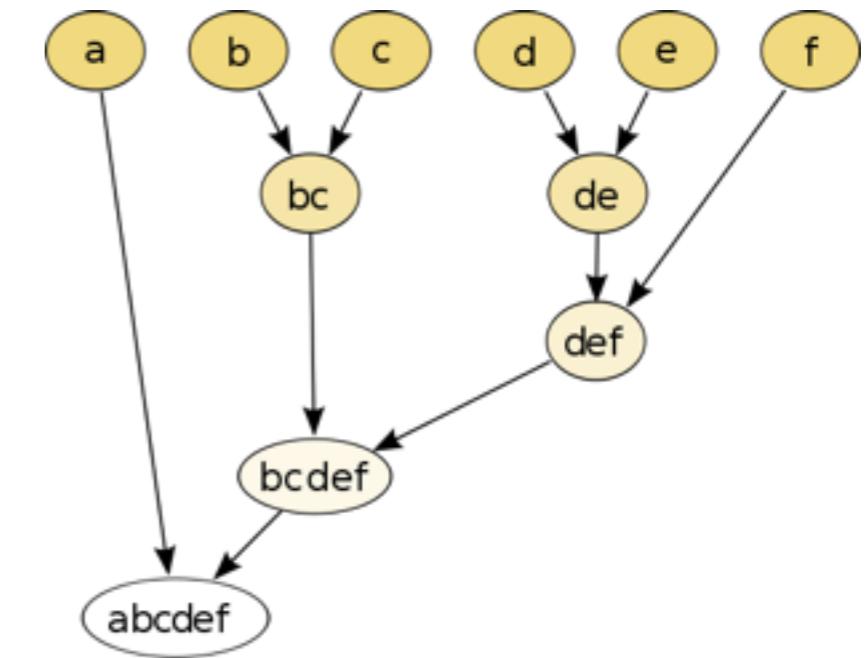
# On Thursday



Big Data  
Visualization



Case Study



Clustering,  
Dimensionality  
Reduction

# Questions

