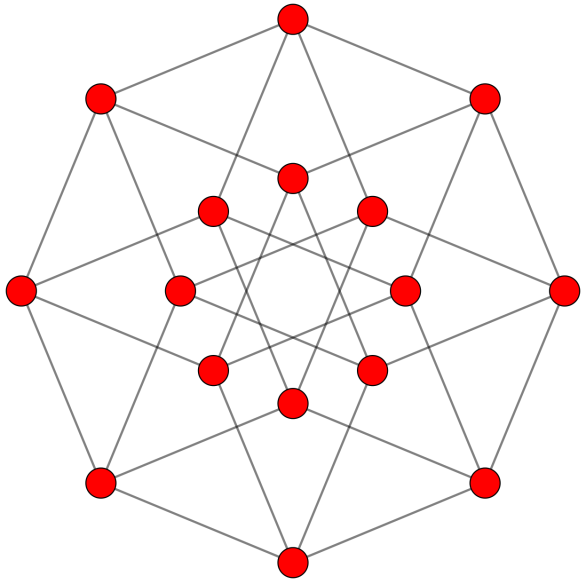
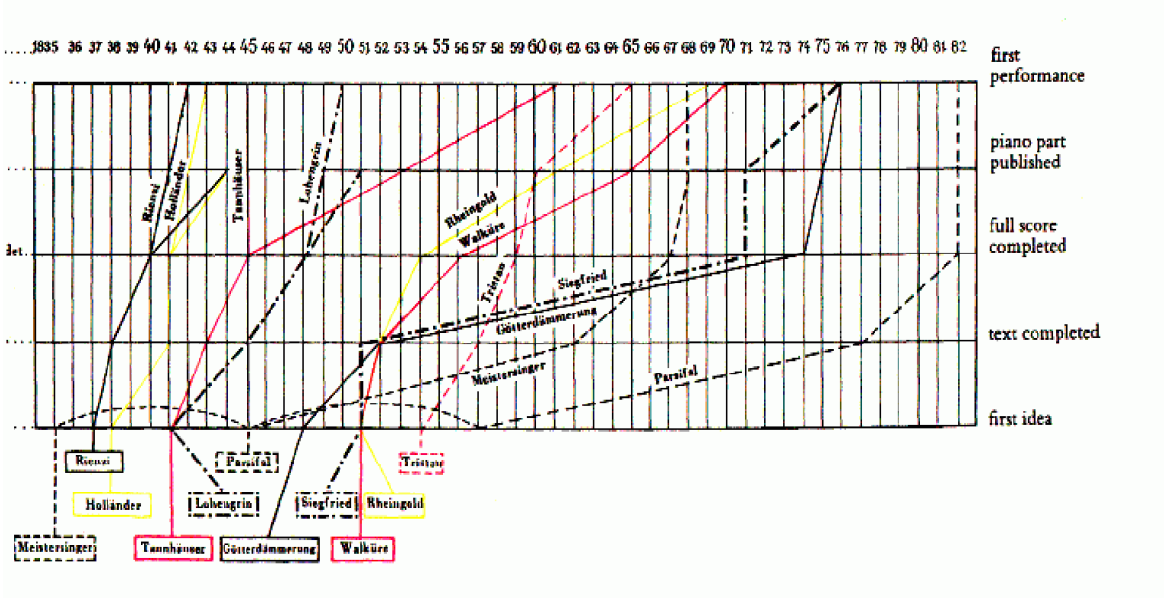
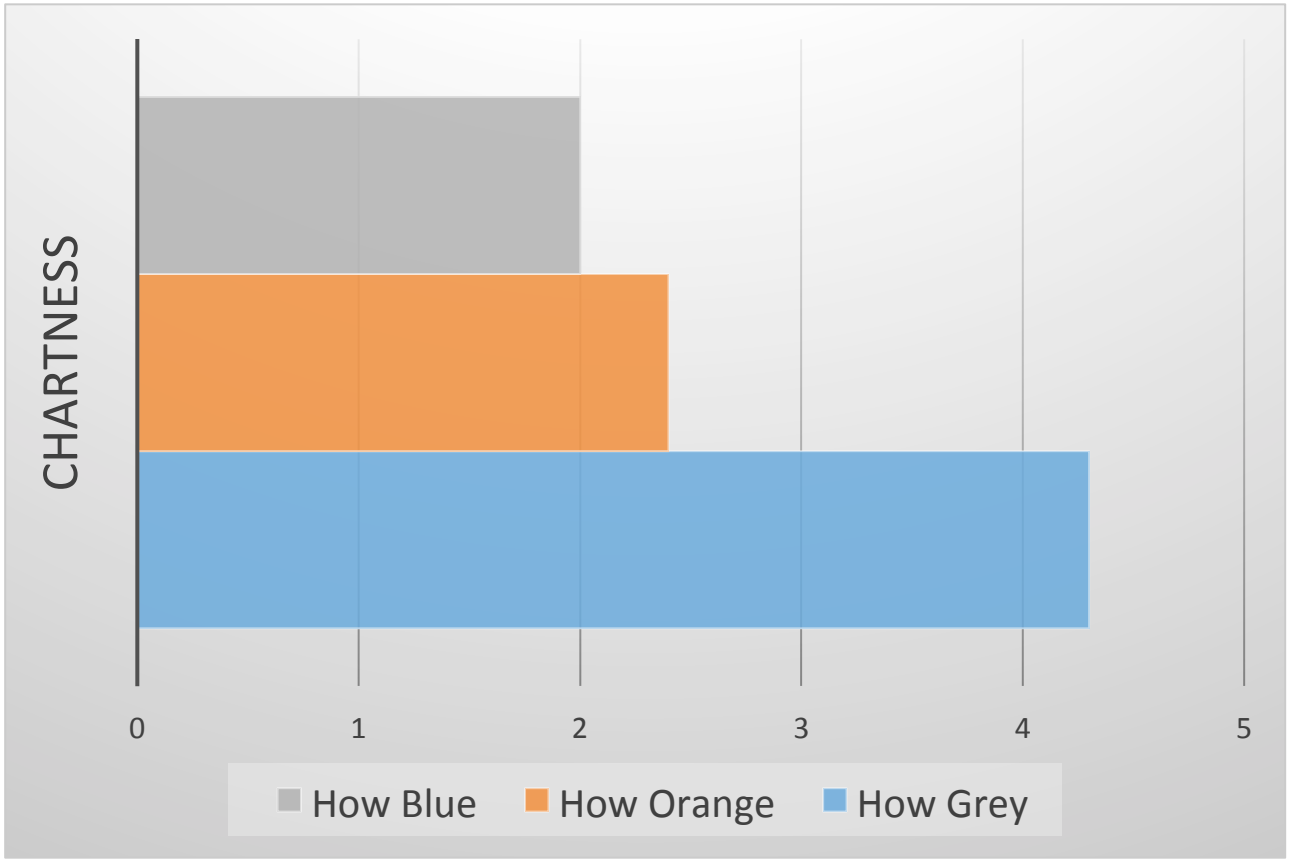
Multidimensional Data

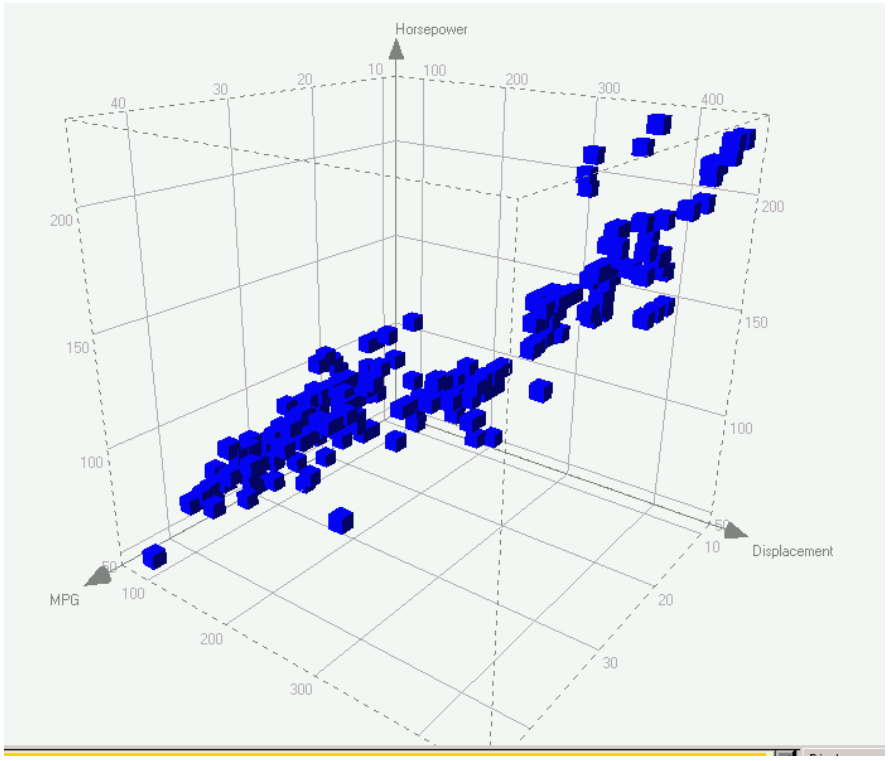


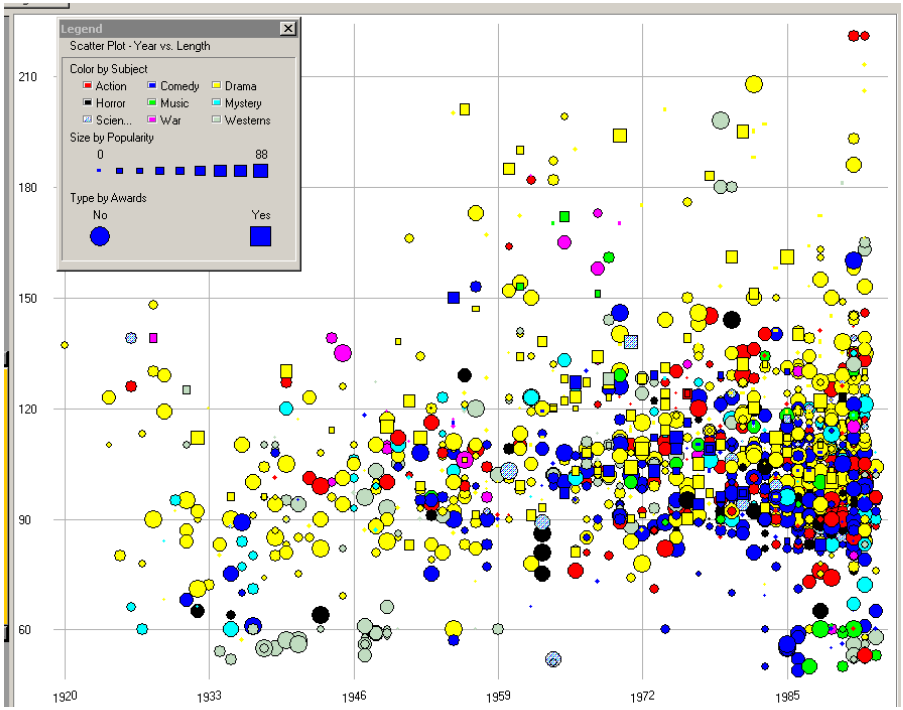


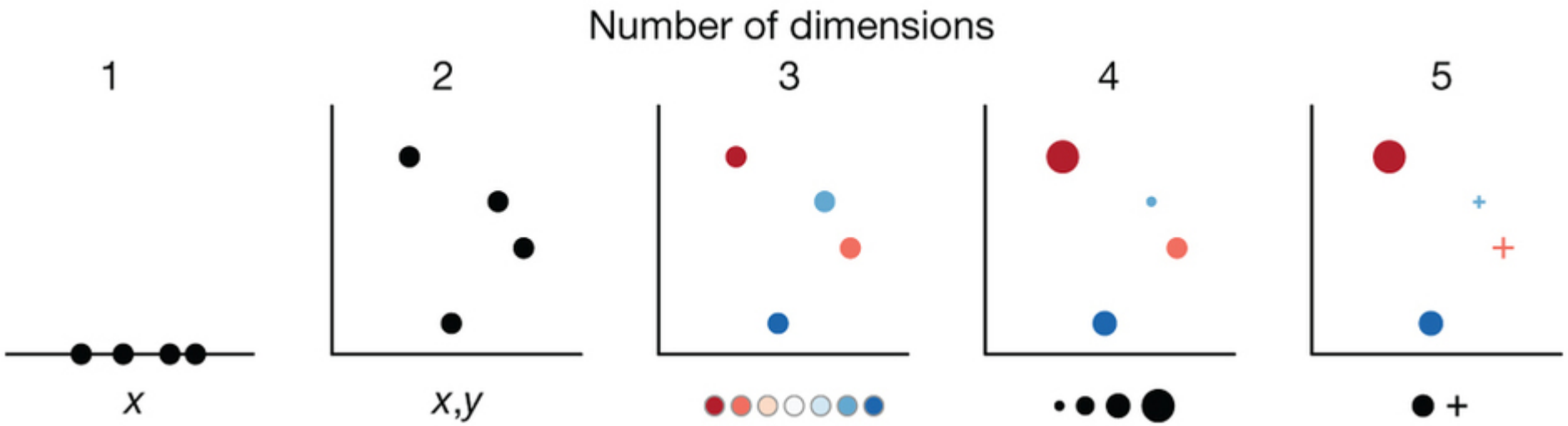
Data Visualization | Steven Bedrick & Jackie Wirz

Image from xkcd.com http://xkcd.com/503/

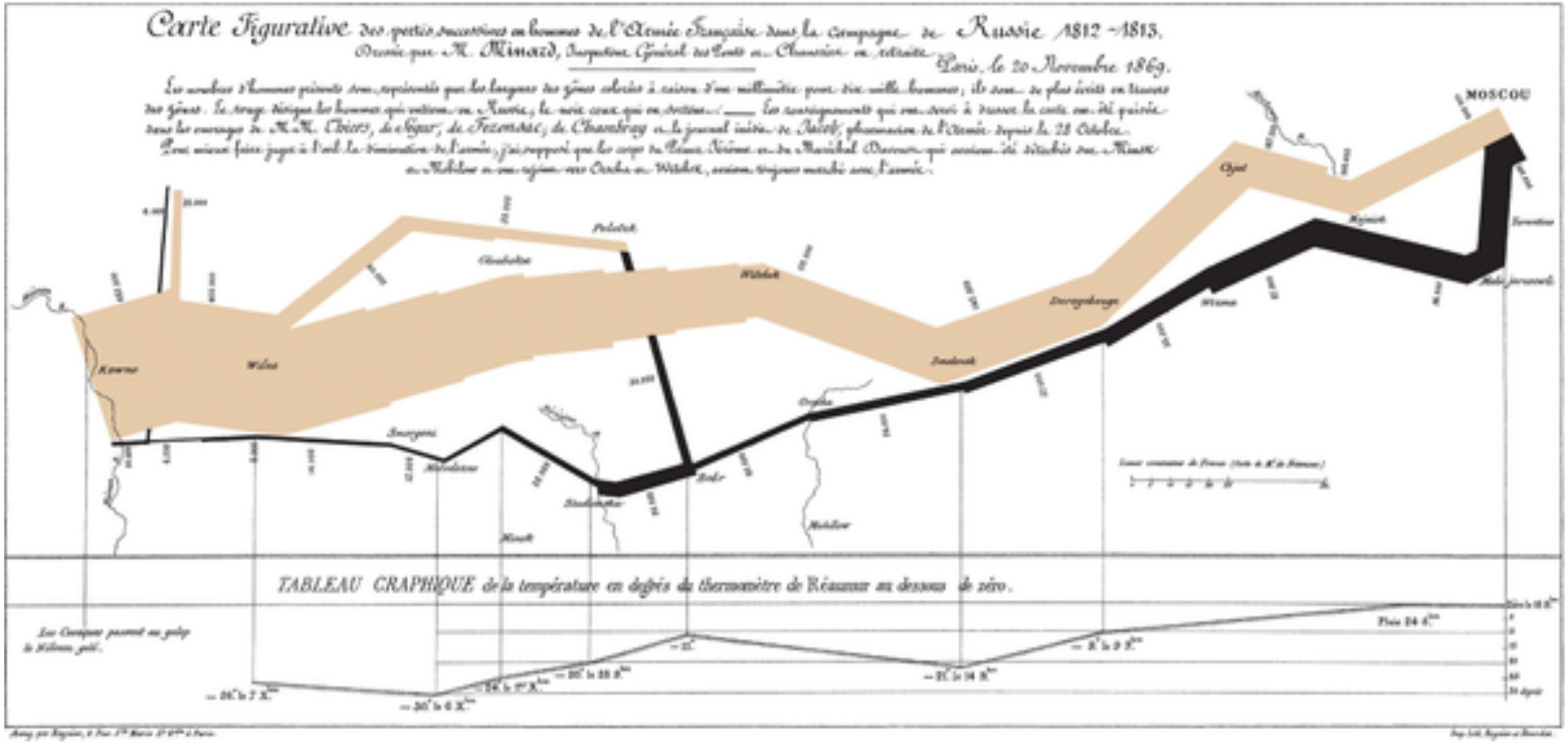

CHARTNESS
0
1
2
3
4
5
How Blue
How Orange
How Grey


Horsepower
10
100
200
300
400
40
30
20
200
150
100
50
MPG
100
200
300
Displacement
10
20
30
50
100
150
200

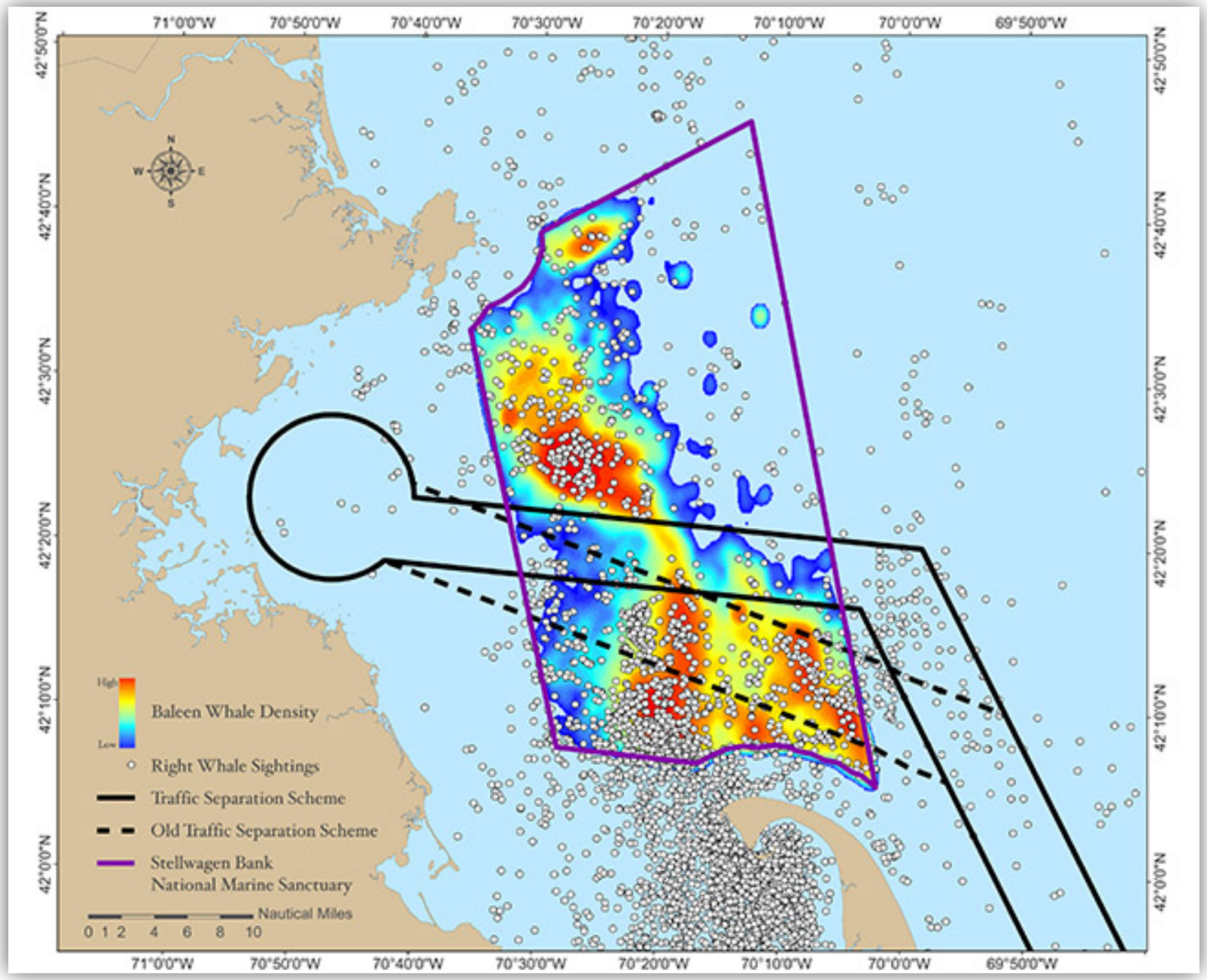

Legend
Scatter Plot - Year vs. Length
Color by Subject
Action
Comedy
Drama
Horror
Music
Mystery
Scien...
War
Westerns
Size by Popularity
0
88
Type by Awards
No
Yes
210
180
150
120
90
60
1920
1933
1946
1959
1972
1985



Krzywinski and Savig "Points of view: Multidimensional data" Nature Methods 10, 595 (2013) doi:10.1038/nmeth.2531



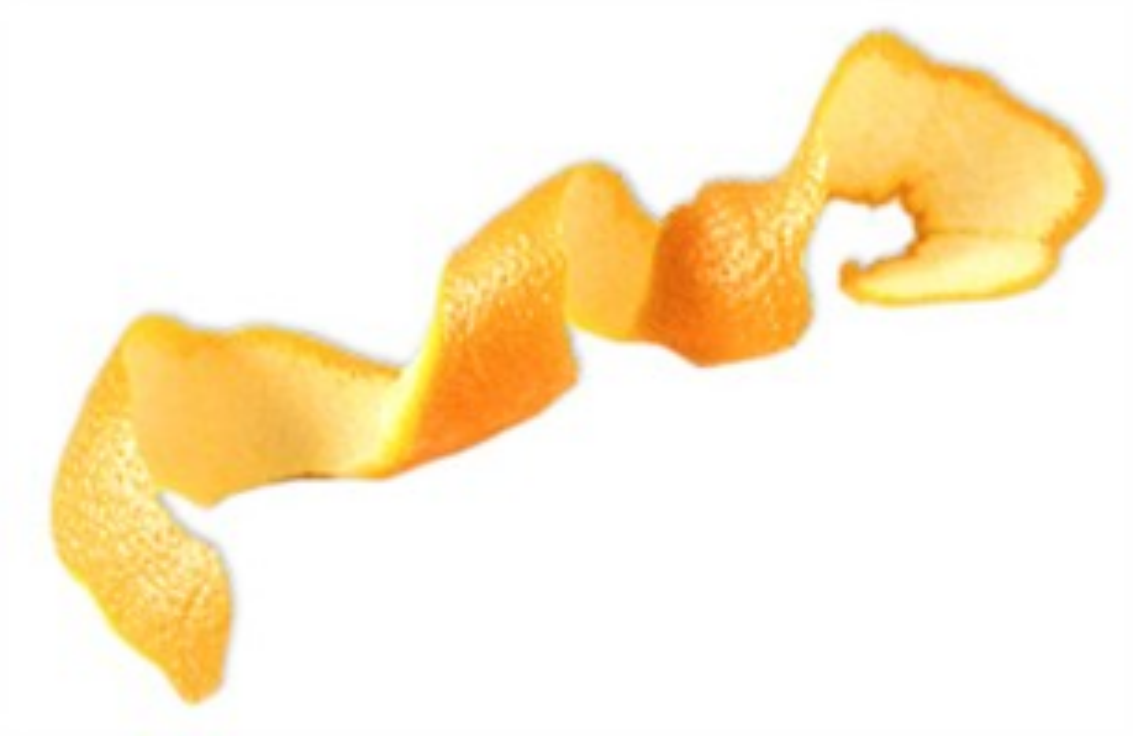
Charles Joseph Minard *map of Napoleon's disastrous Russian campaign of 1812* (1869)



Wiley, Thompson and Merrick: *Realigning the Boston Traffic Separation Scheme* (2006): http://scimaps.org/V.3

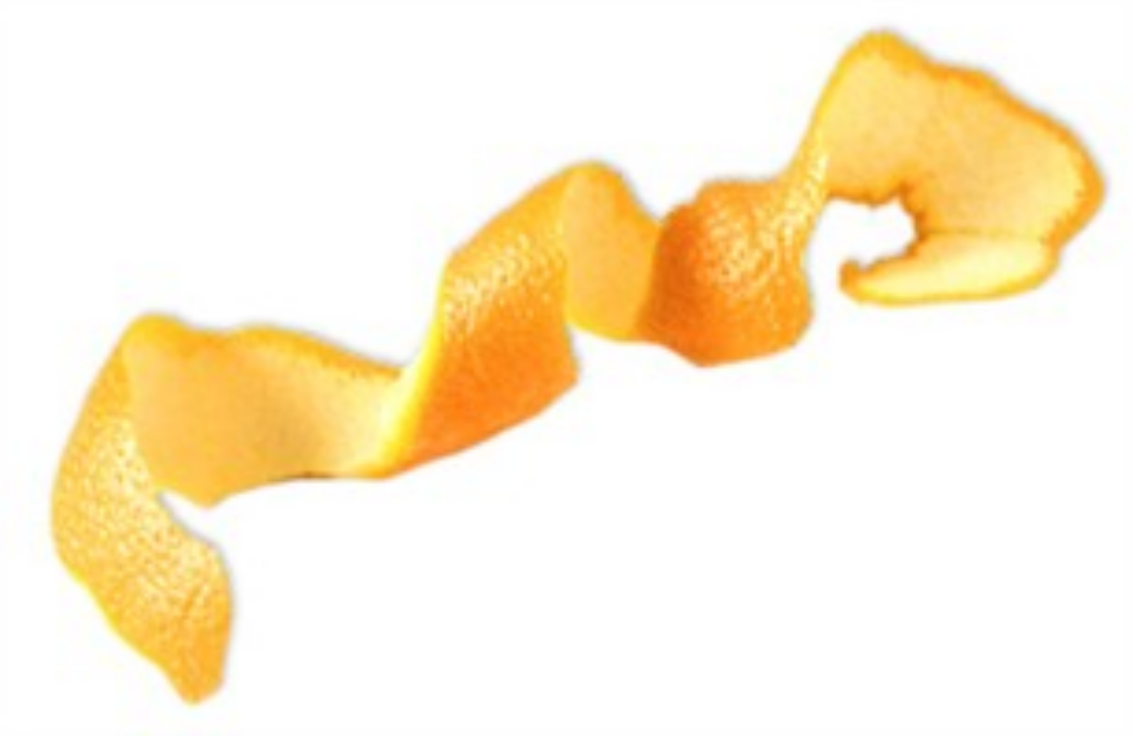
Multivariate

Too many dimensions



3D Spheres do not flatten well…

http://www.herbalwater.com/herbs-and-ingredients/orange-peel.aspx



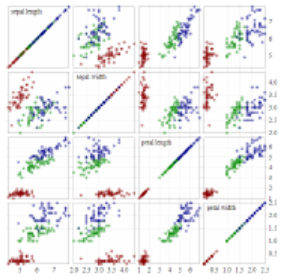
Nth Dimensional Data Needs to be Projected onto 2D

http://www.herbalwater.com/herbs-and-ingredients/orange-peel.aspx

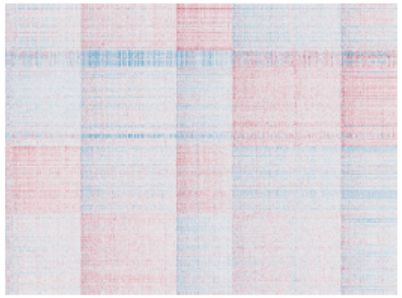
The Fundamental Problem

- Each attribute defines a dimension
- Small # of dimensions easy
    - Data mapping, Cleveland's rules
- What about many dimensional data? The *real* stuff that occurs in the *real* world?

The Fundamental Problem

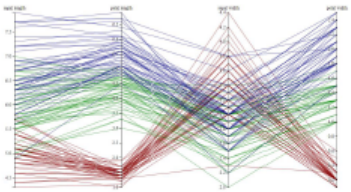


Scatterplot Matrices
[Bostock]

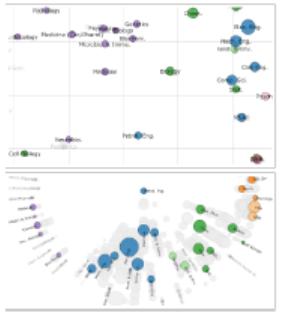


Pixel-based visualizations / heat maps

Multidimensional Scaling
[Doerk 2011]



Parallel Coordinates
[Bostock]



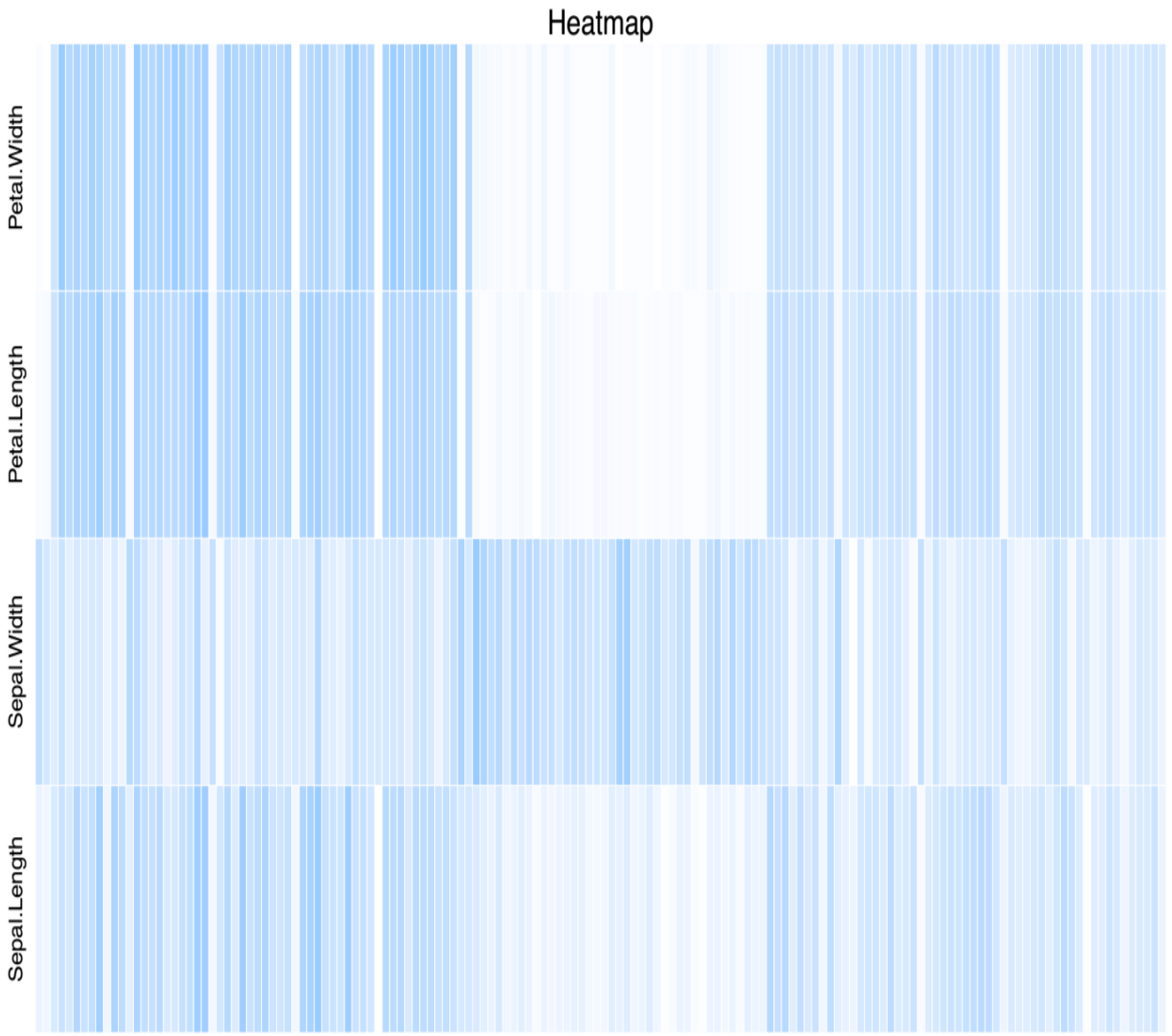
[Chuang 2012]

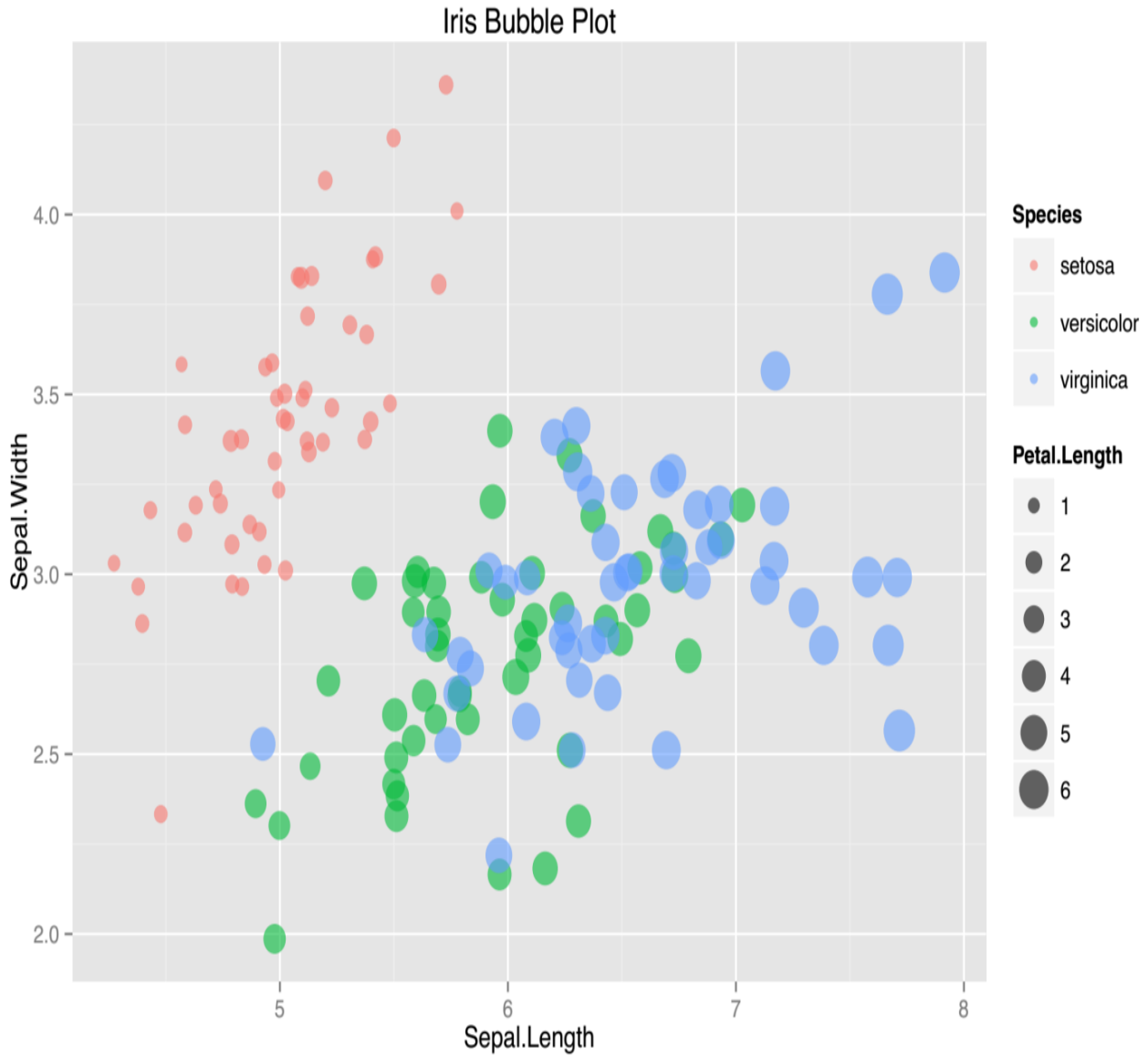
no / little analytics ⟷ **strong analytics component**

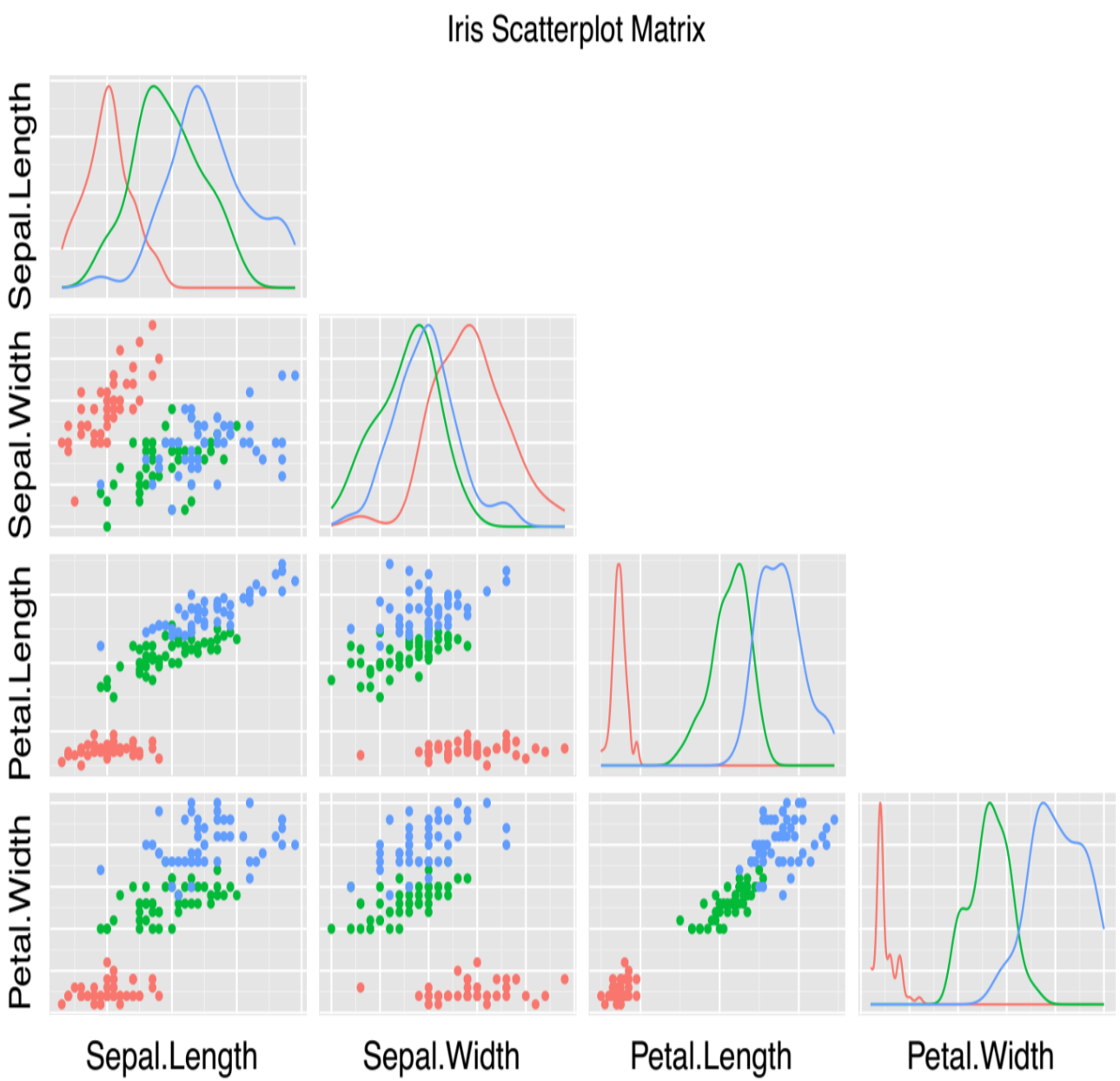
Slide modified from Alexander Lex, Harvard

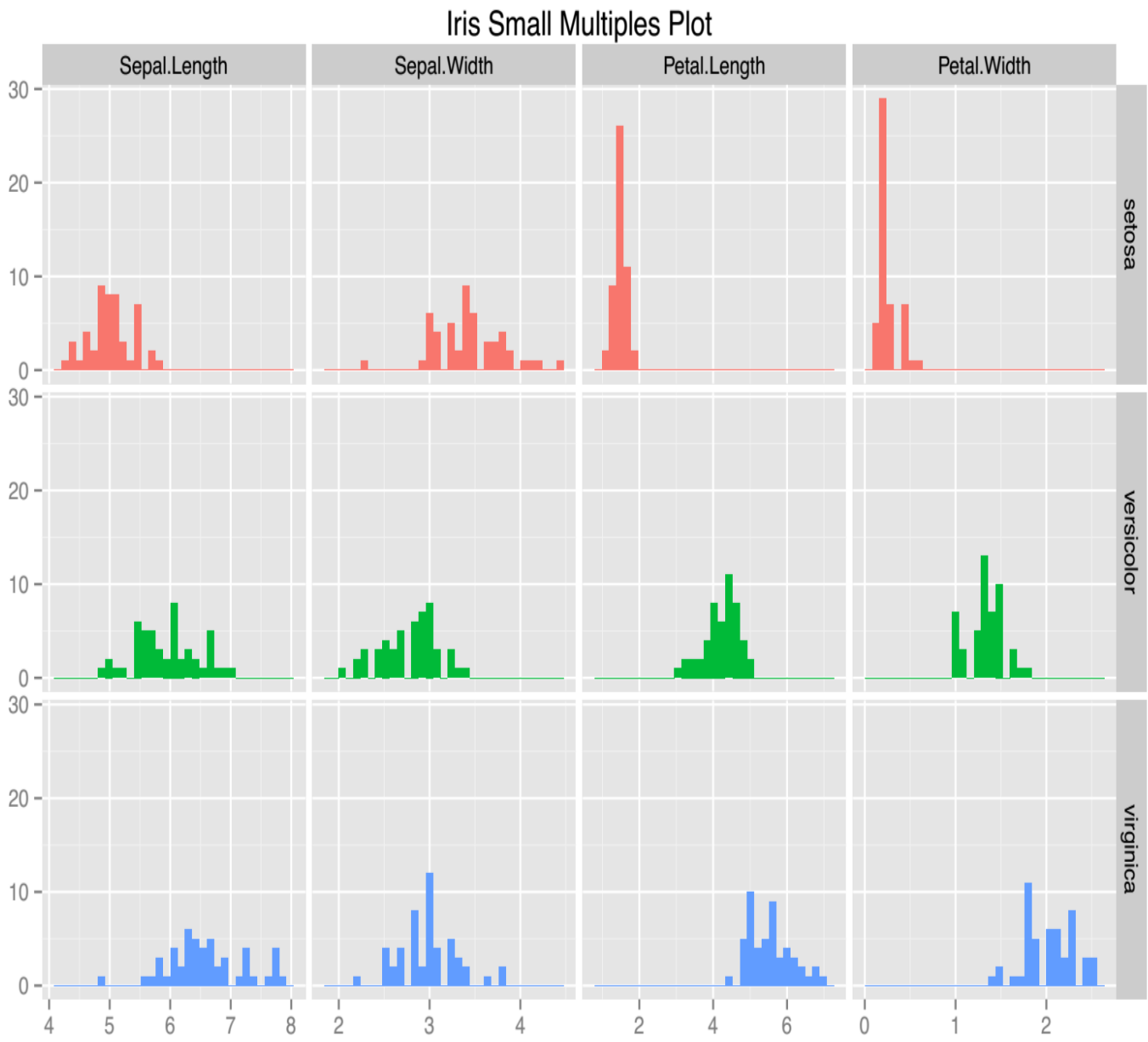
Iris Data Set

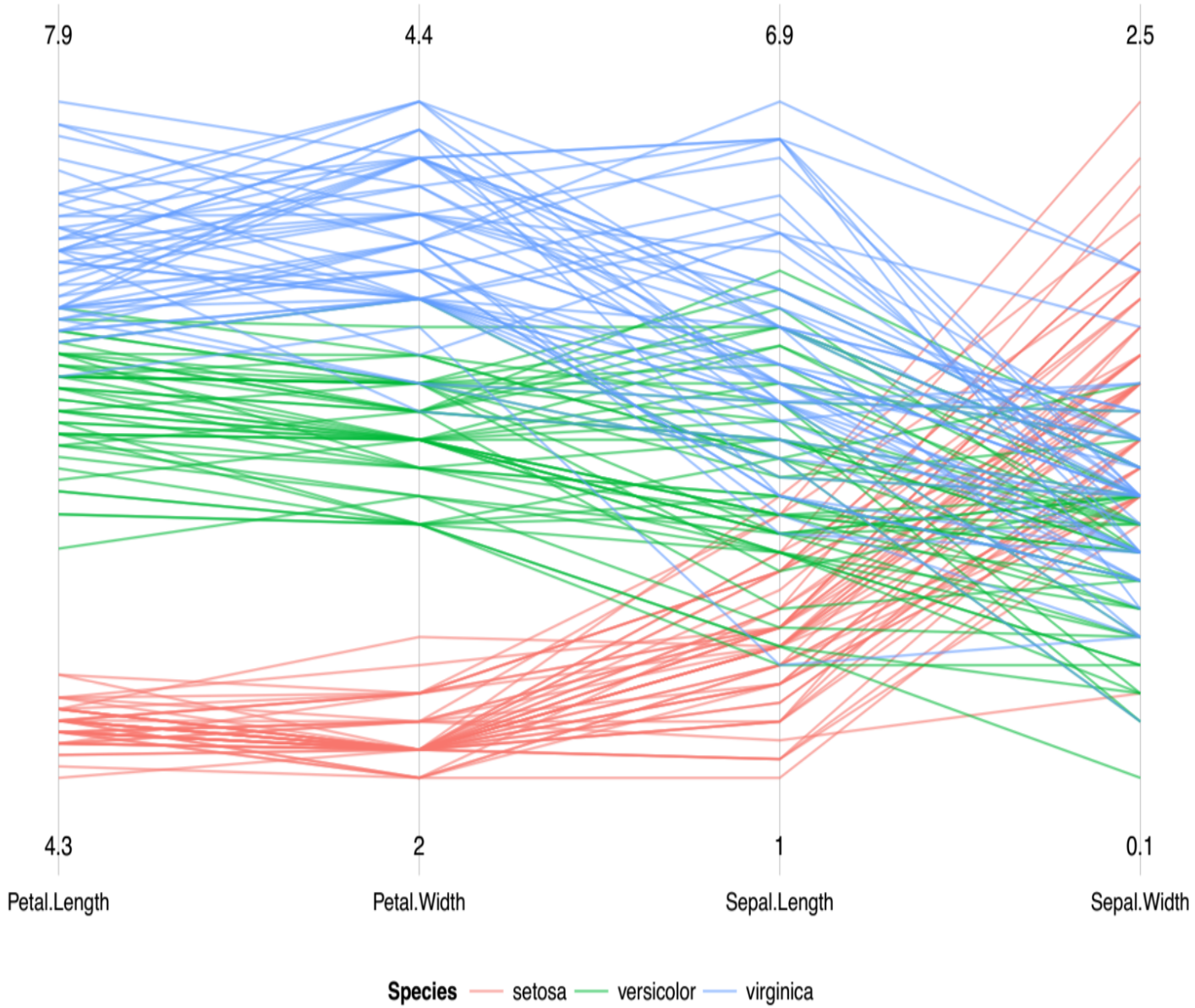
Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Fisher (1936) as an example of discriminant analysis


Heatmap
Petal.Width
Petal.Length
Sepal.Width
Sepal.Length


Iris Bubble Plot
Sepal.Width
4.0
3.5
3.0
2.5
2.0
5
6
7
8
Sepal.Length
Species
setosa
versicolor
virginica
Petal.Length
1
2
3
4
5
6


Iris Scatterplot Matrix
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width


Iris Small Multiples Plot
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
setosa
versicolor
virginica
30
20
10
0
4
5
6
7
8
2
3
4
2
4
6
0
1
2


7.9
4.4
6.9
2.5
4.3
2
1
0.1
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width
Species
setosa
versicolor
virginica

Bubbly

Symbolically Speaking

The Bubble Plot

Horizontal position: Continuous data

Vertical position: Continuous data

Circle area: Numerical data

Circle color: Numerical or **categorical data**

The Bubble Plot

Able to encode four dimensions of data

Ideal if one dimension is categorical (color)

Rough comparison possible

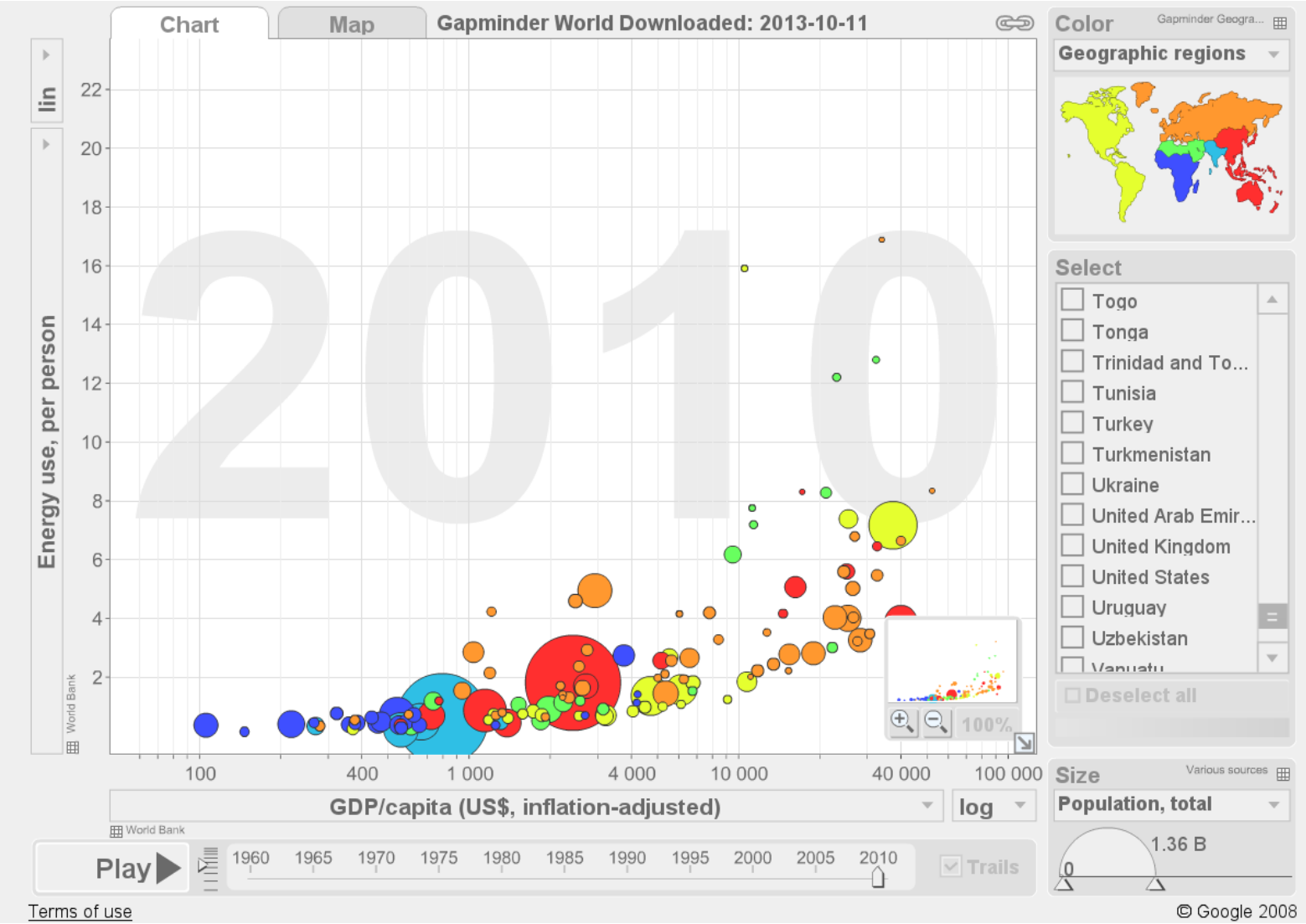
Beware comparing circle areas!

Obscuring data may be an issue

Large circles should be behind smaller ones

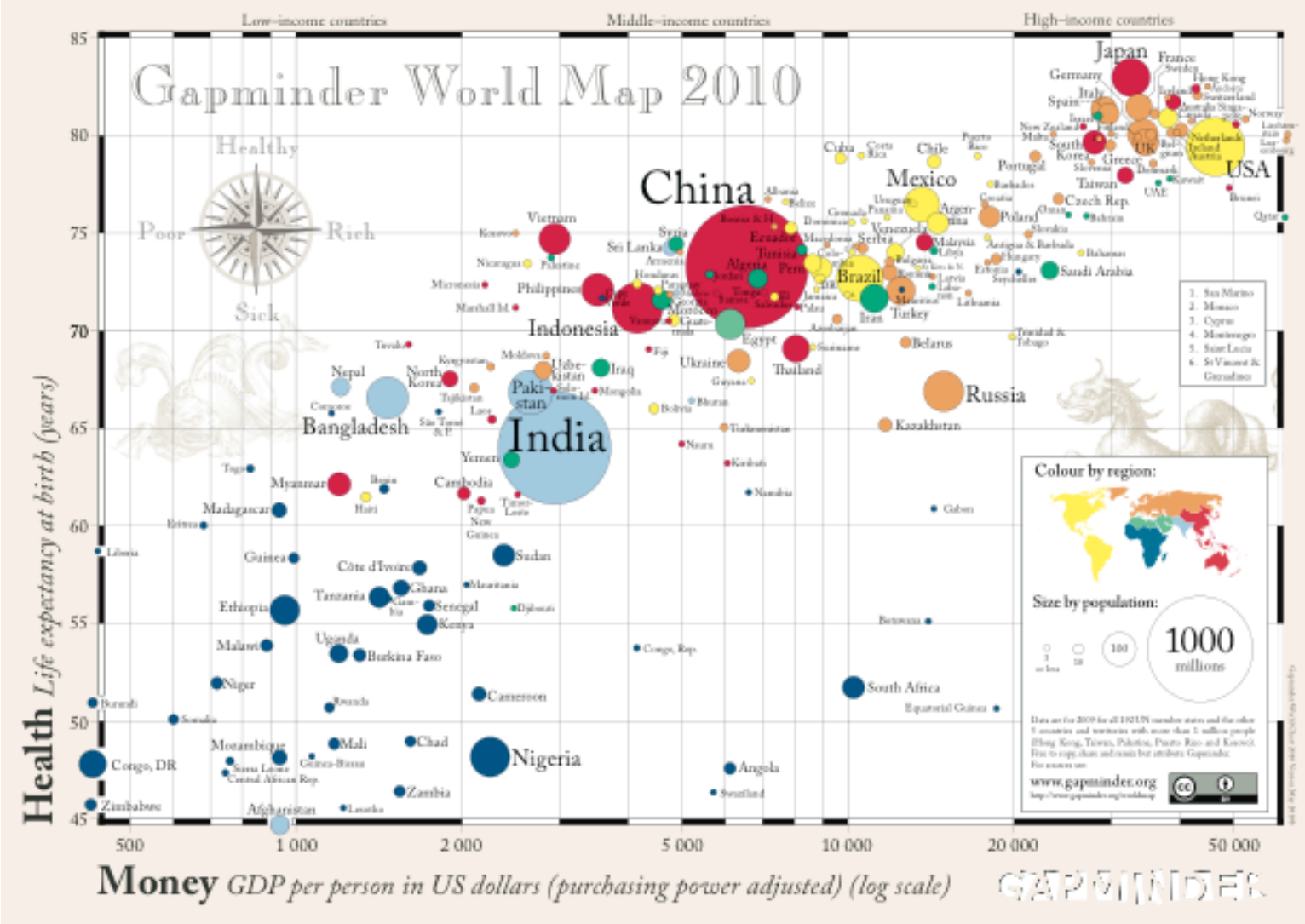
Issues increases with density

The (new classic) Bubble Plot



http://en.wikipedia.org/wiki/Trendalyzer#mediaviewer/File:Gapminder_world.png

The (new classic) Bubble Plot



http://betterevaluation.org/sites/default/files/BubbleChartImage1.png

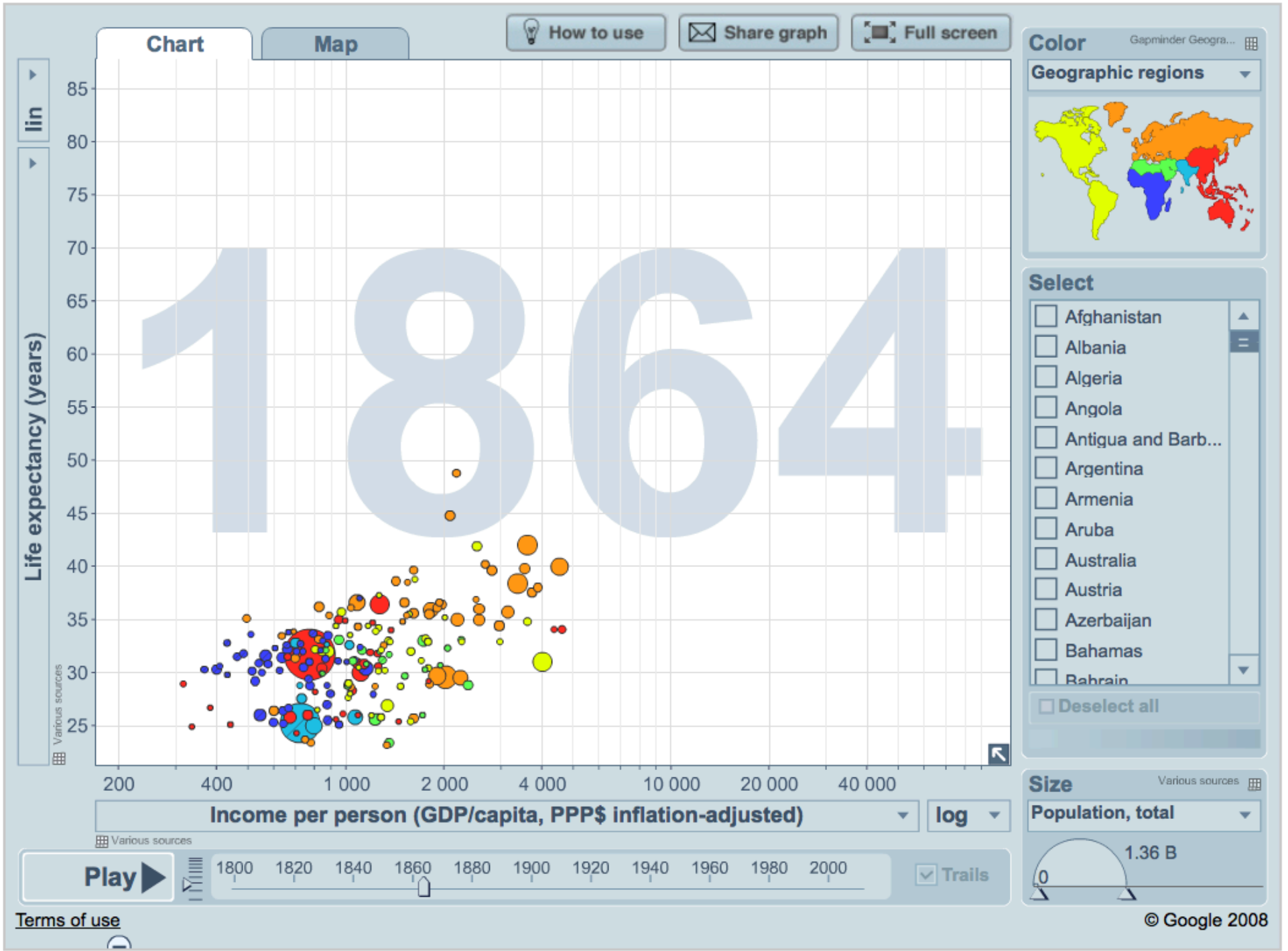


Chart
Map
How to use
Share graph
Full screen
Color
Gapminder Geogra...
Geographic regions
Select
Afghanistan
Albania
Algeria
Angola
Antigua and Barb...
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Deselect all
lin
Life expectancy (years)
Various sources
85
80
75
70
65
60
55
50
45
40
35
30
25
1864
200
400
1 000
2 000
4 000
10 000
20 000
40 000
Income per person (GDP/capita, PPP$ inflation-adjusted)
log
Various sources
Size
Various sources
Population, total
1.36 B
0
Play
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Trails
Terms of use
© Google 2008

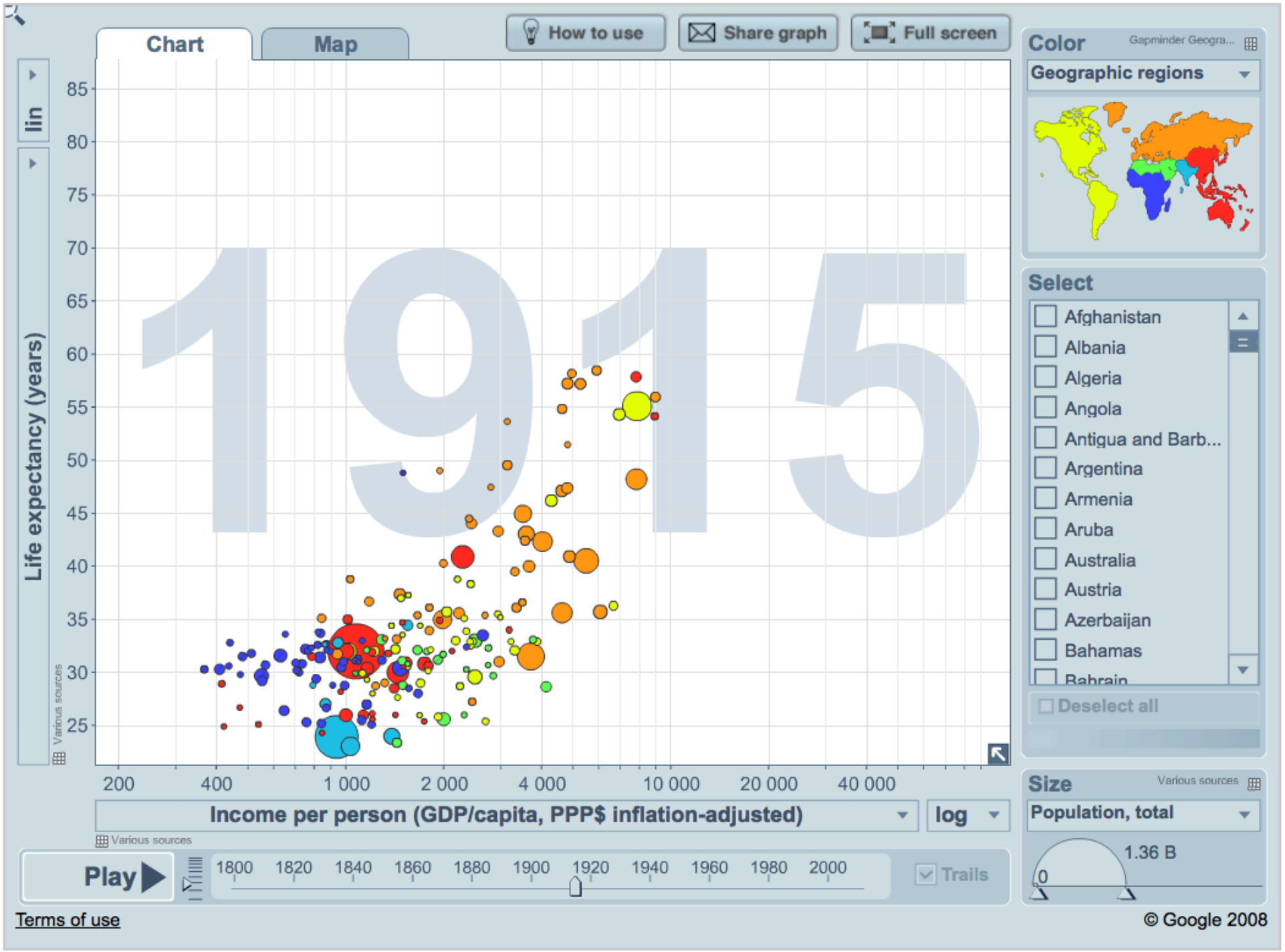


Chart
Map
How to use
Share graph
Full screen
Color
Gapminder Geogra...
Geographic regions
Select
Afghanistan
Albania
Algeria
Angola
Antigua and Barb...
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Deselect all
lin
Life expectancy (years)
Various sources
85
80
75
70
65
60
55
50
45
40
35
30
25
1915
200
400
1 000
2 000
4 000
10 000
20 000
40 000
Income per person (GDP/capita, PPP$ inflation-adjusted)
log
Various sources
Size
Various sources
Population, total
1.36 B
0
Play
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Trails
Terms of use
© Google 2008

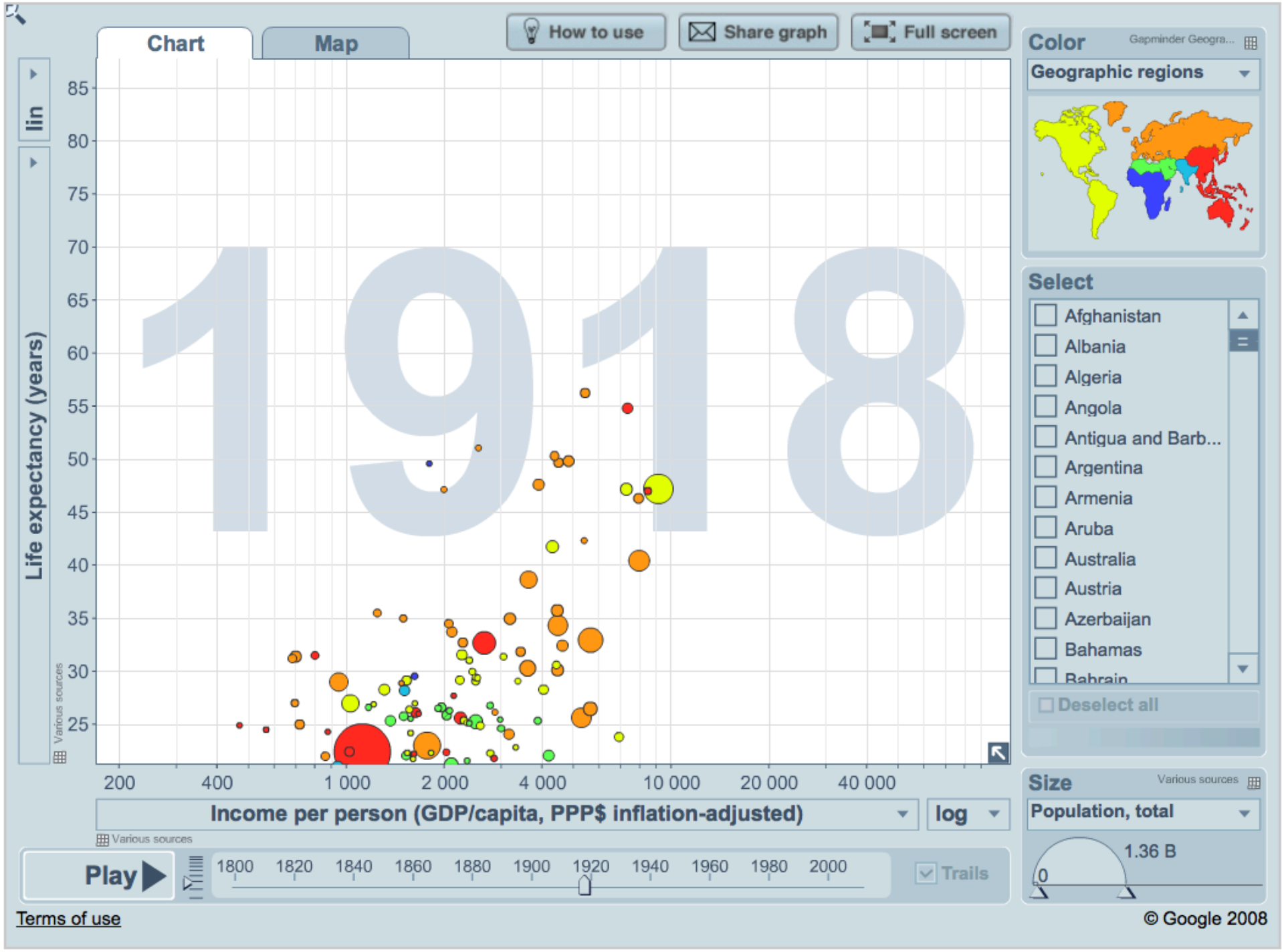


Chart
Map
How to use
Share graph
Full screen
Color
Gapminder Geogra...
Geographic regions
Select
Afghanistan
Albania
Algeria
Angola
Antigua and Barb...
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Deselect all
lin
Life expectancy (years)
Various sources
1918
85
80
75
70
65
60
55
50
45
40
35
30
25
200
400
1 000
2 000
4 000
10 000
20 000
40 000
Income per person (GDP/capita, PPP$ inflation-adjusted)
log
Various sources
Size
Various sources
Population, total
1.36 B
0
Play
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Trails
Terms of use
© Google 2008

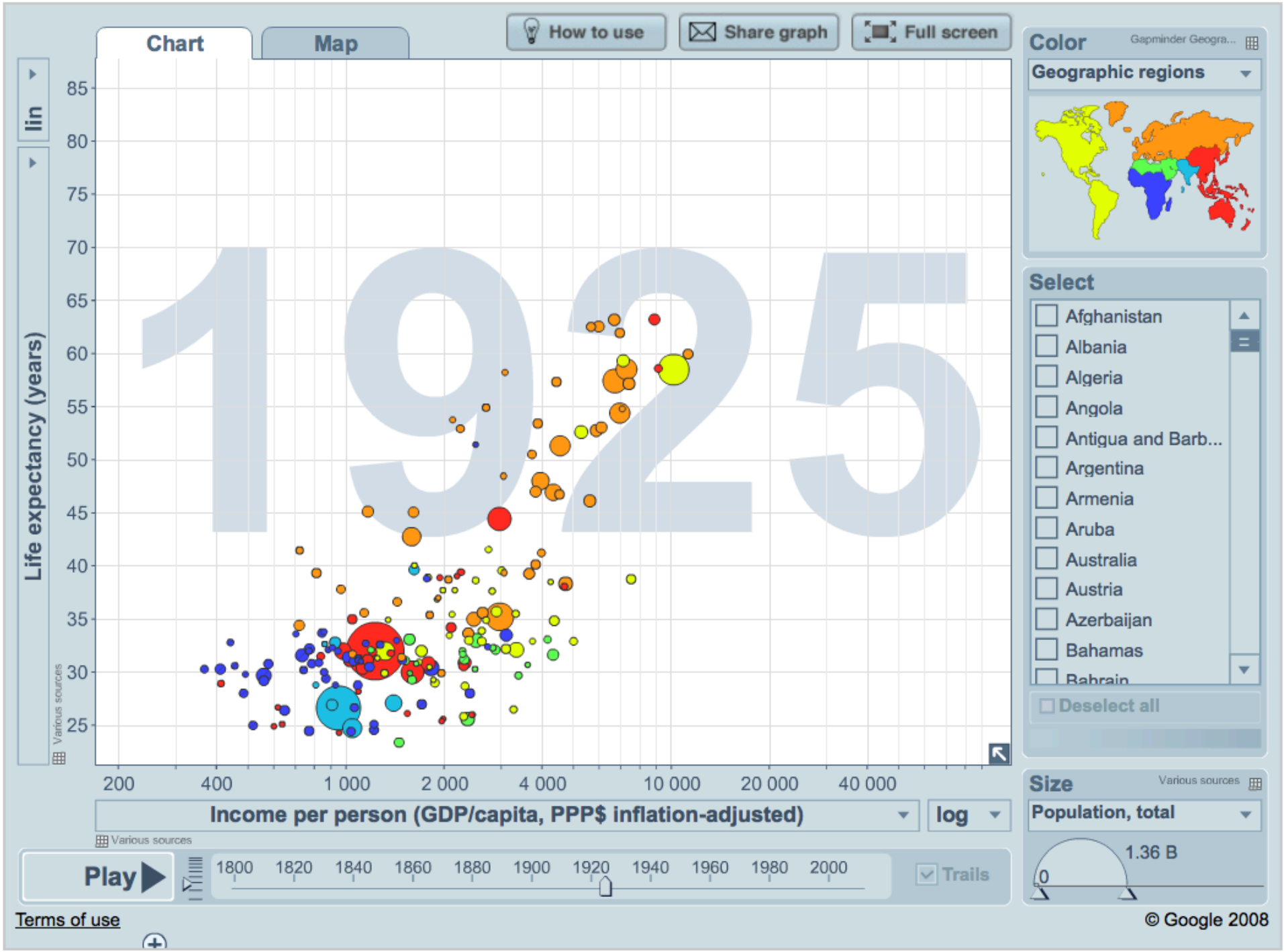
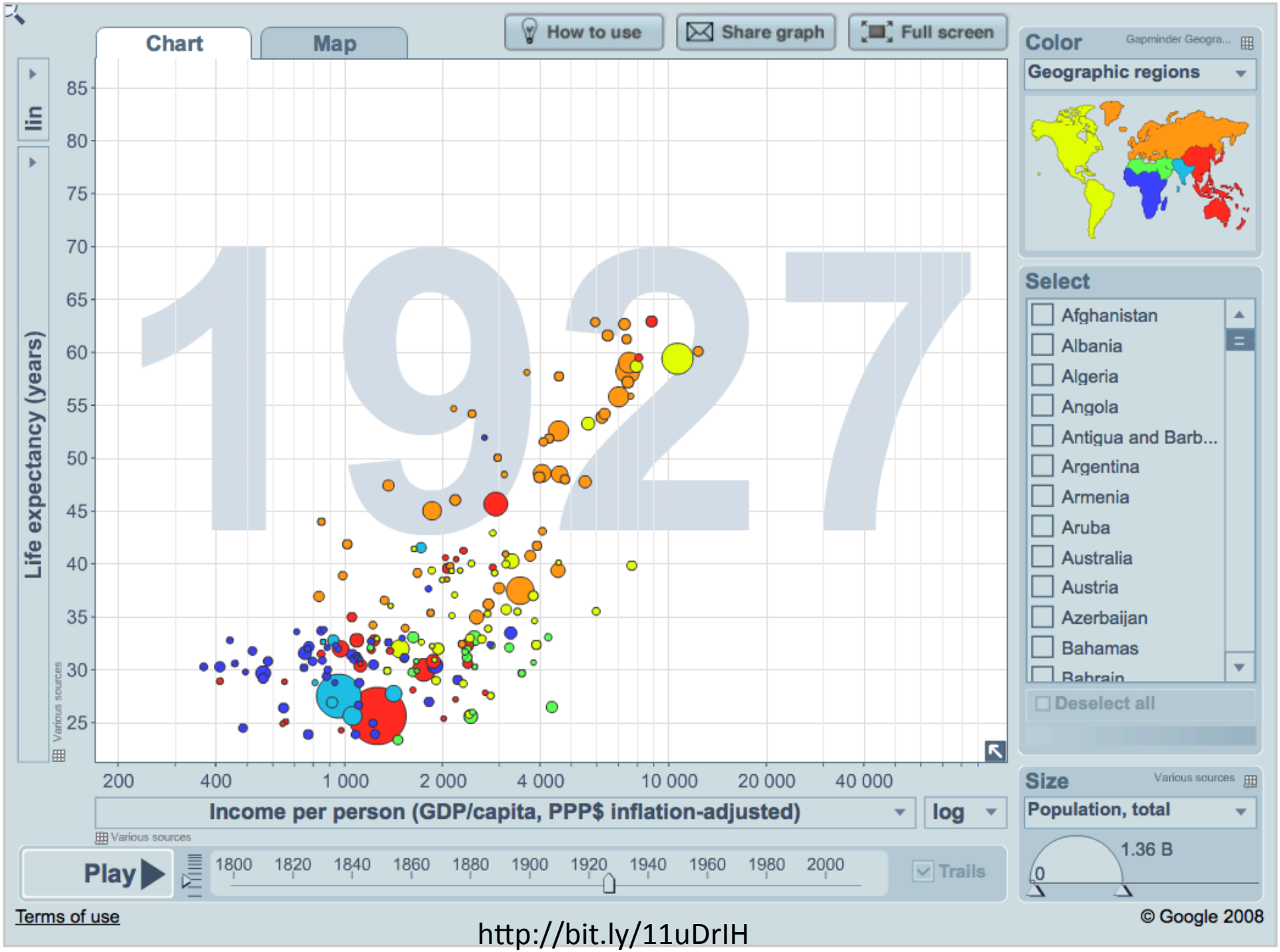
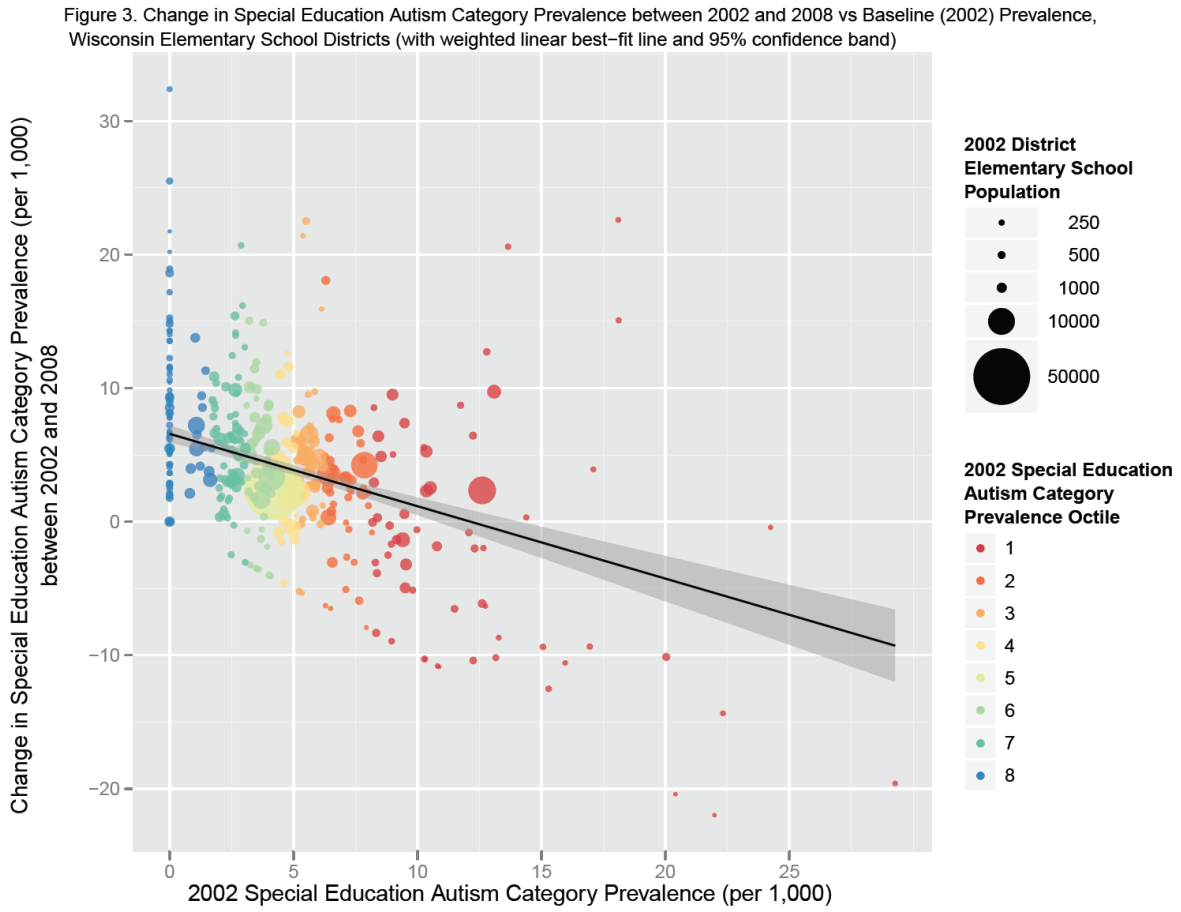


Chart
Map
How to use
Share graph
Full screen
Color
Gapminder Geogra...
Geographic regions
Select
Afghanistan
Albania
Algeria
Angola
Antigua and Barb...
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Deselect all
lin
Life expectancy (years)
Various sources
85
80
75
70
65
60
55
50
45
40
35
30
25
1925
200
400
1 000
2 000
4 000
10 000
20 000
40 000
Income per person (GDP/capita, PPP$ inflation-adjusted)
log
Various sources
Size
Various sources
Population, total
1.36 B
0
Play
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Trails
Terms of use
© Google 2008



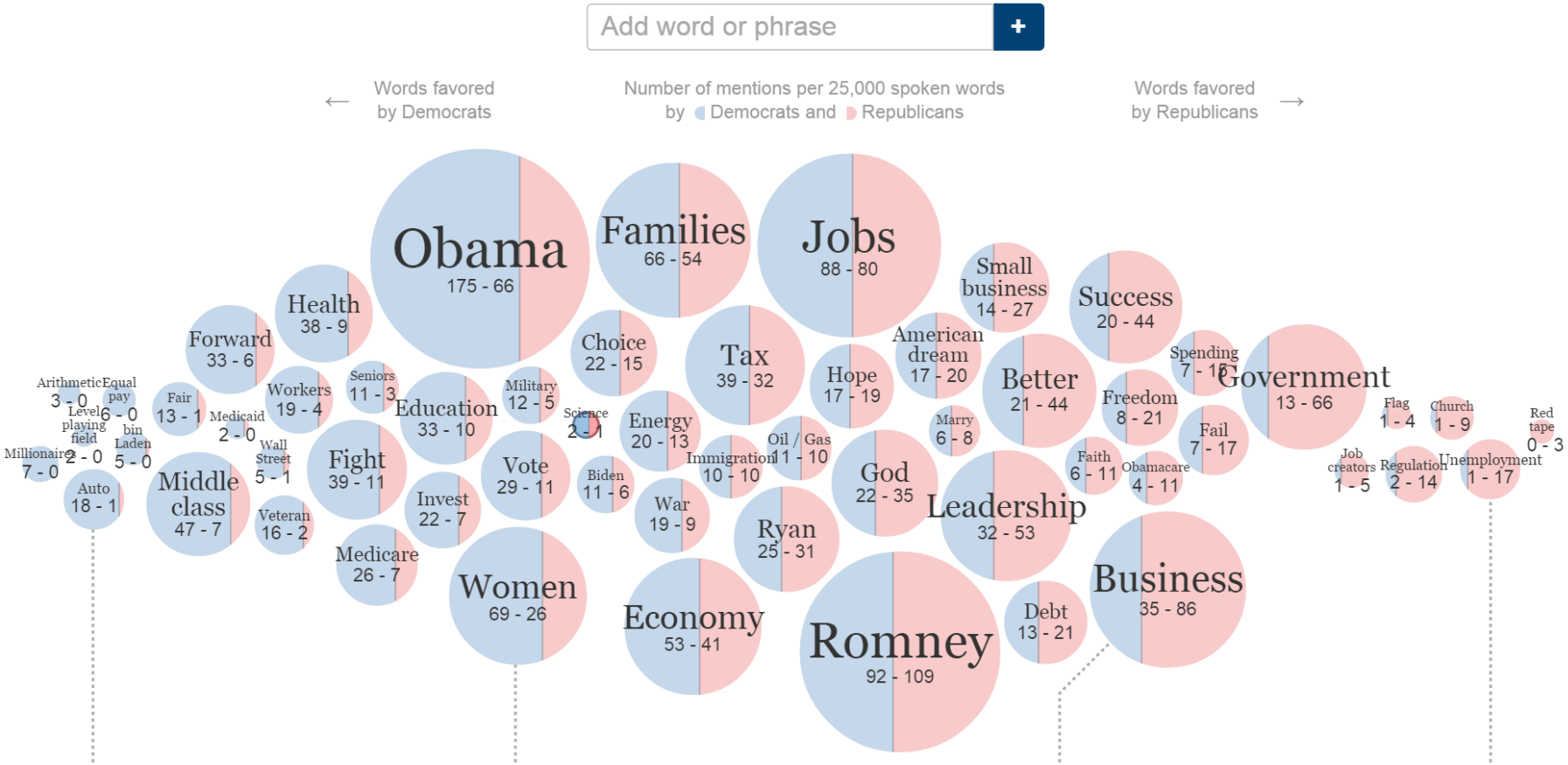
http://bit.ly/11uDrIH

The Academic Bubble Plot



http://www.matthewmaenner.com/blog/wp-content/uploads/2010/11/Fig3.png

The Bubble Plot (sort of)



http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html?_r=1&#science

Scatter Plots & Scatterplot Matrices

Symbolically Speaking

Scatters

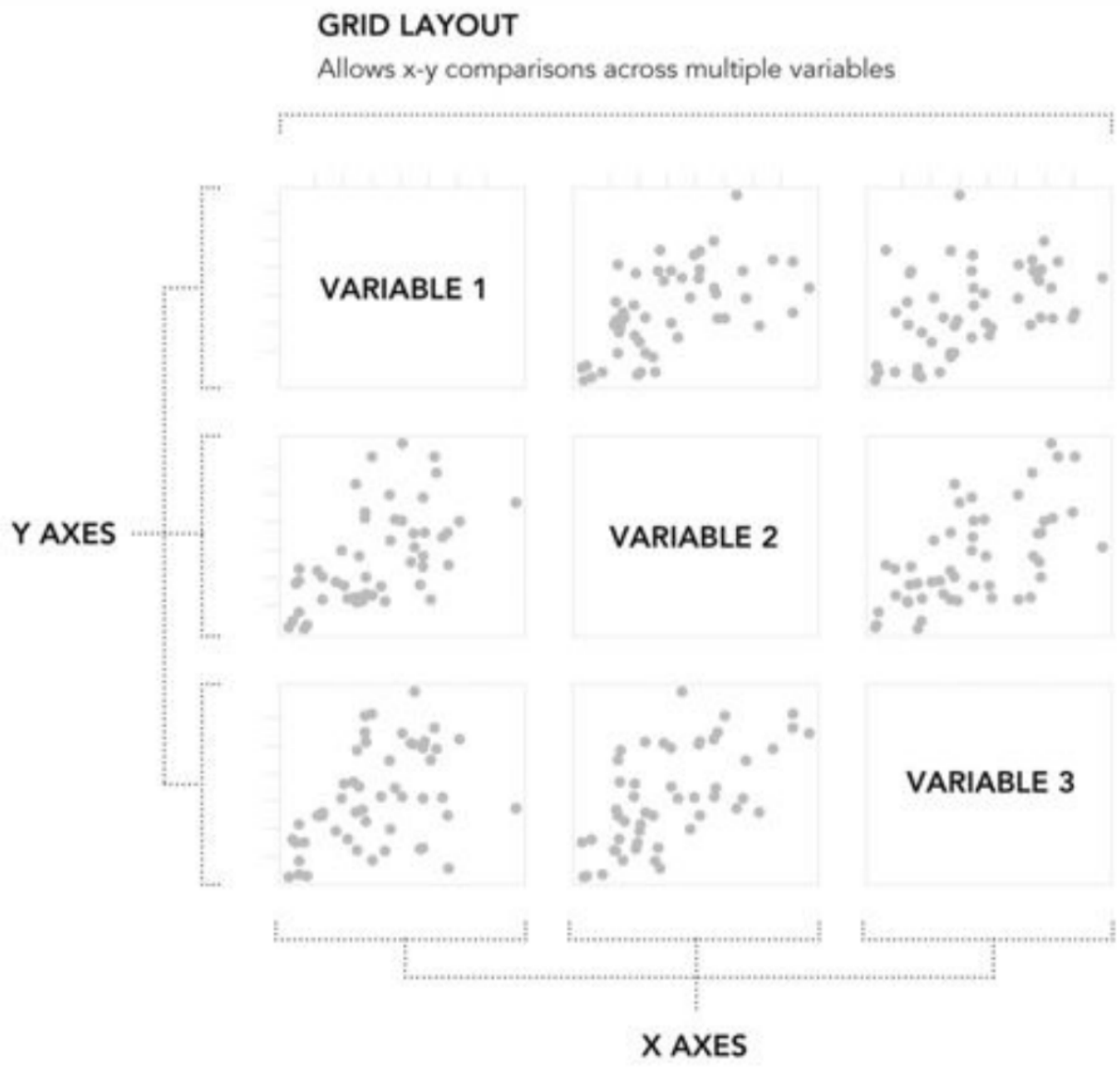
Scatterplot

- Horizontal position maps to one variable
- Vertical position maps to another variable

Matrix of scatterplots

- Each scatterplot focuses on one pair
- Which pair is determined by row and column
- Good for exploration and comparison
  - Can be a little overwhelming at first

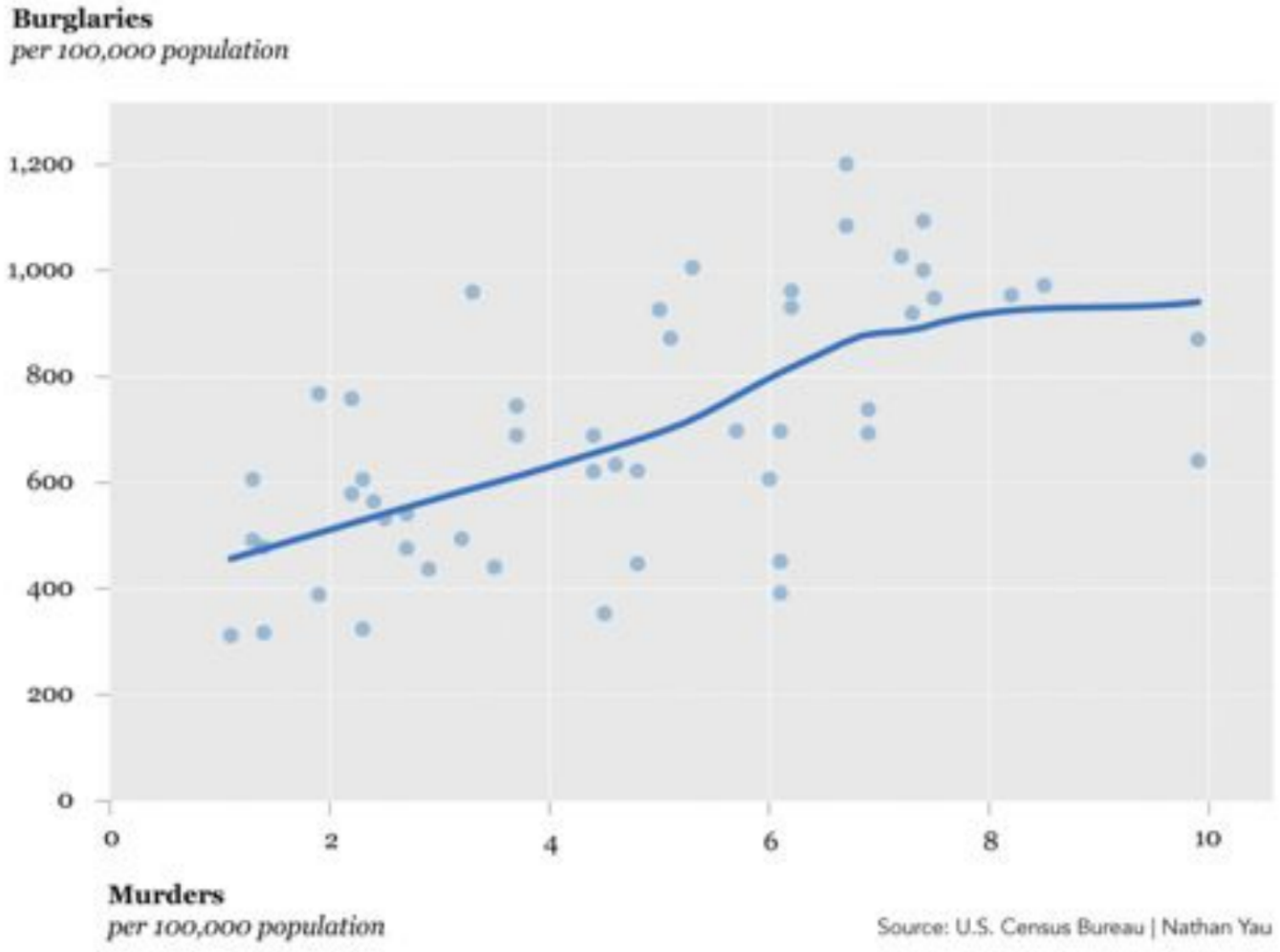
Limited scalability

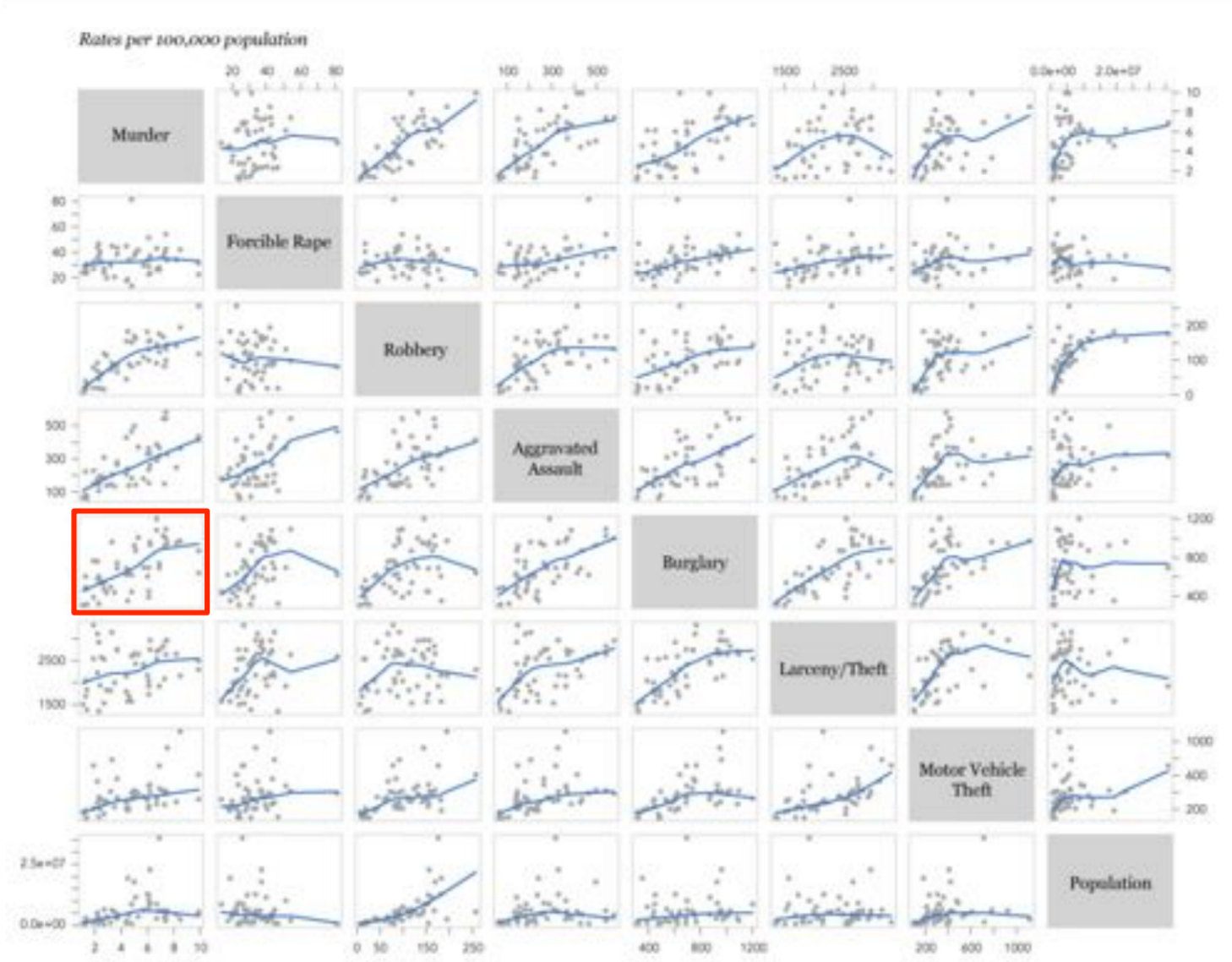


Nathan Yau “Visualize This” 2011

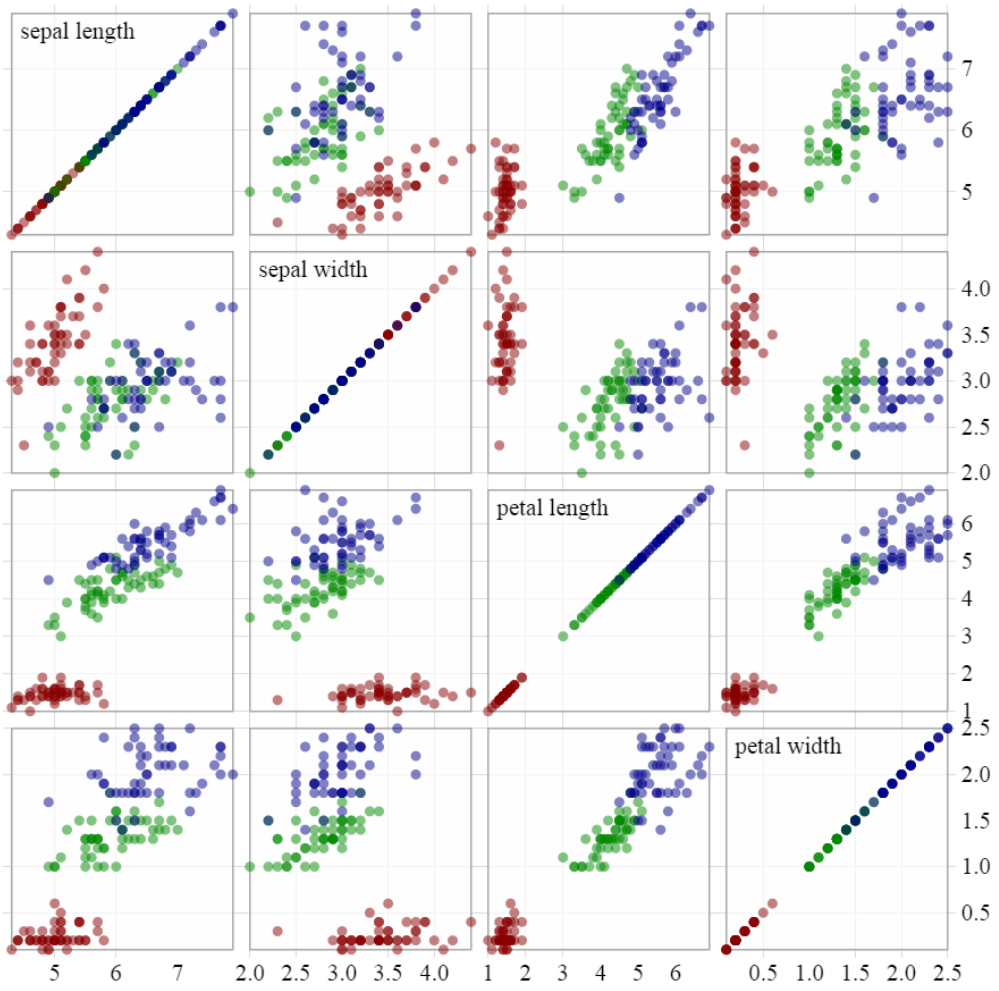
MURDERS VERSUS BURGLARIES IN THE UNITED STATES

States with higher murder rates tend to have higher burglary rates.





Nathan Yau “Visualize This” 2011 Chapter 6



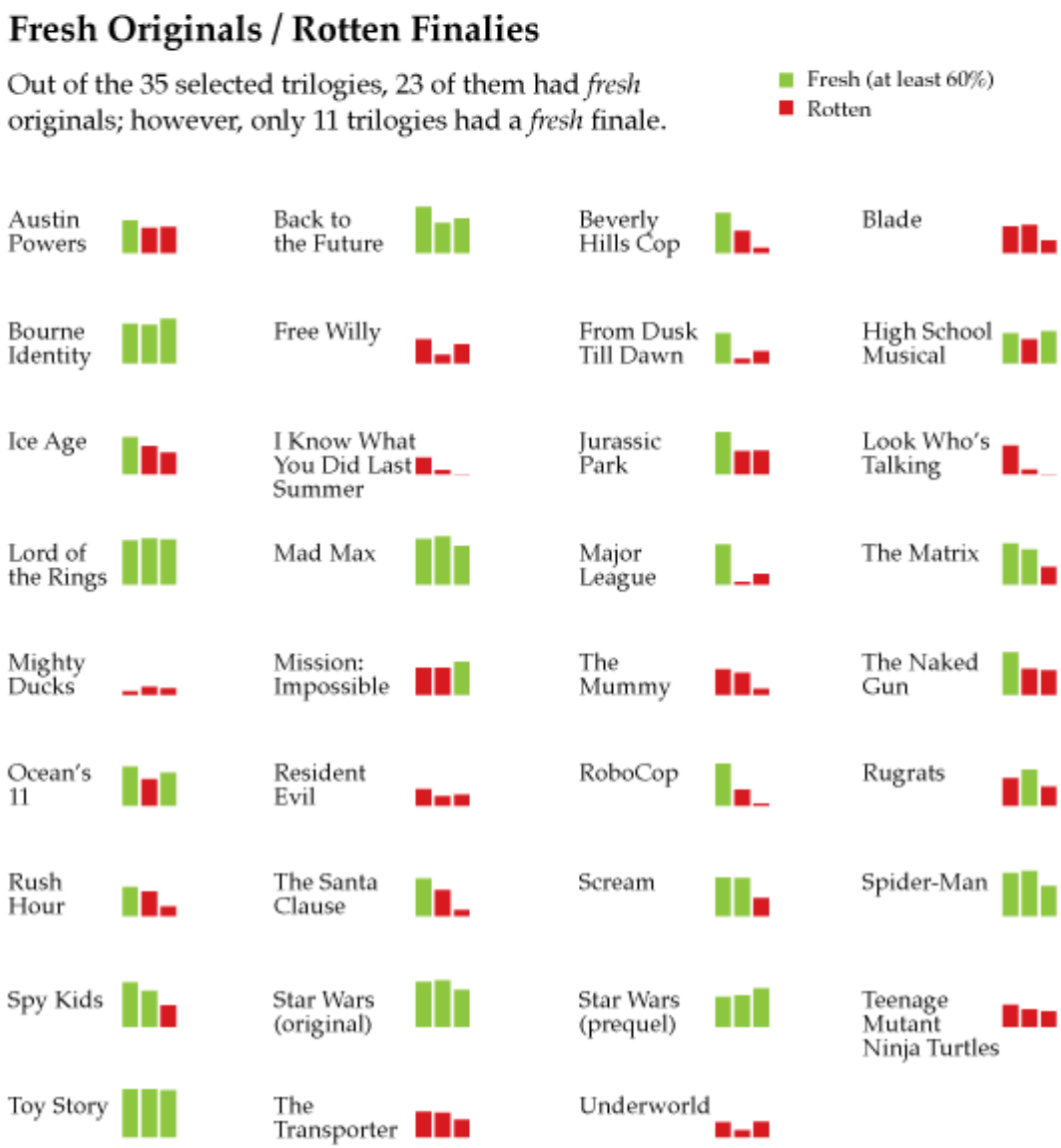
- *Iris setosa*
- *Iris versicolor*
- *Iris virginica*

Edgar Anderson's *Iris* data set
scatterplot matrix

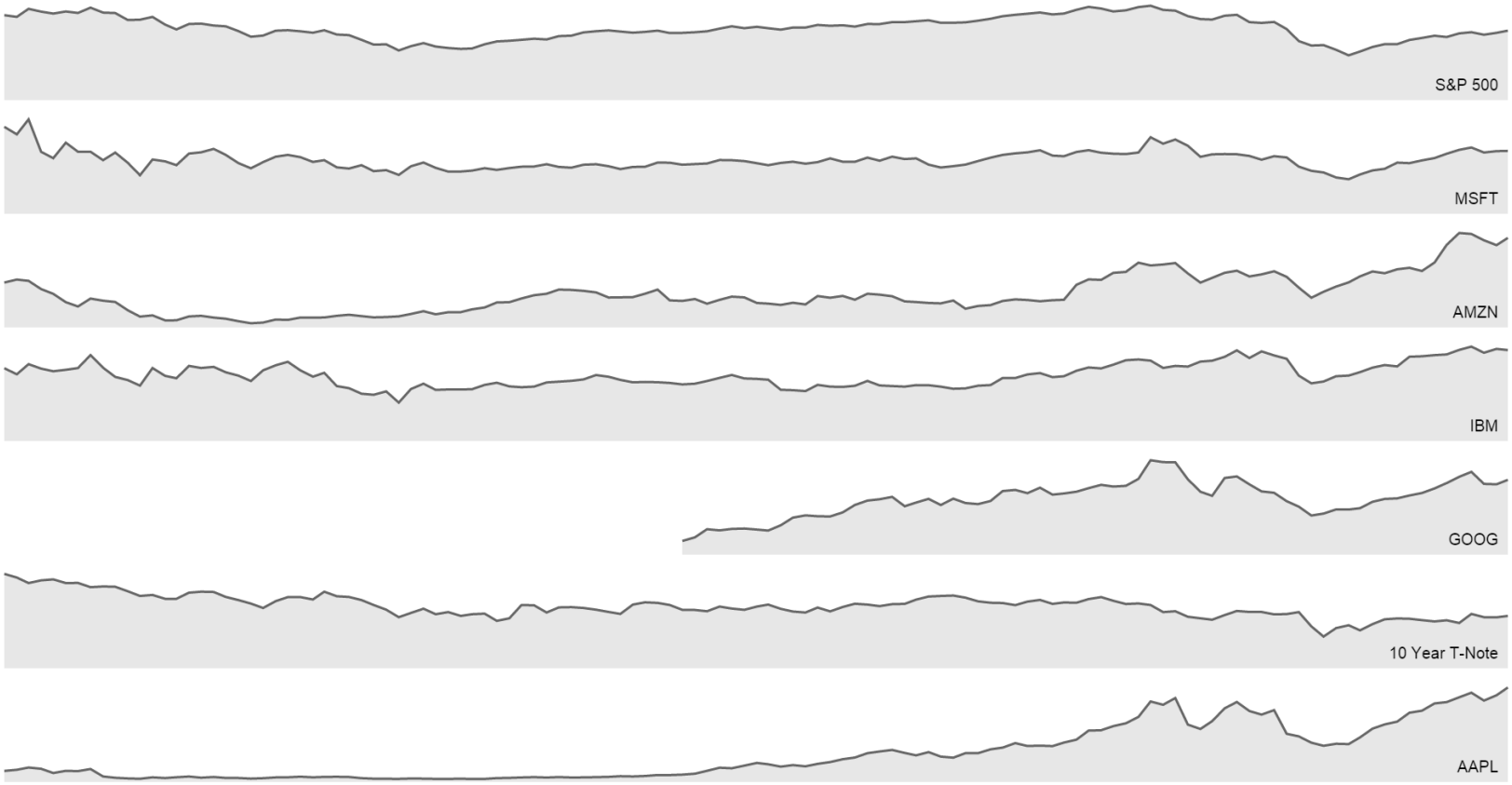
http://mbostock.github.io/d3/talk/20111116/iris-splom.html

Small Multiples…

Not scatter but you get the idea…



http://flowingdata.com/2010/06/28/do-movie-sequels-live-up-to-their-originals/



http://bl.ocks.org/mbostock/1157787

Glyph Plots

Symbolically Speaking

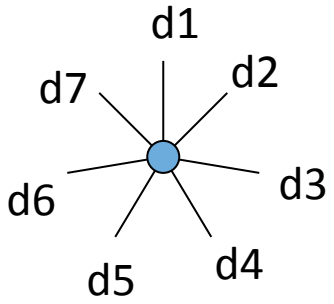
Star Plots

Add additional axes in a radial fashion

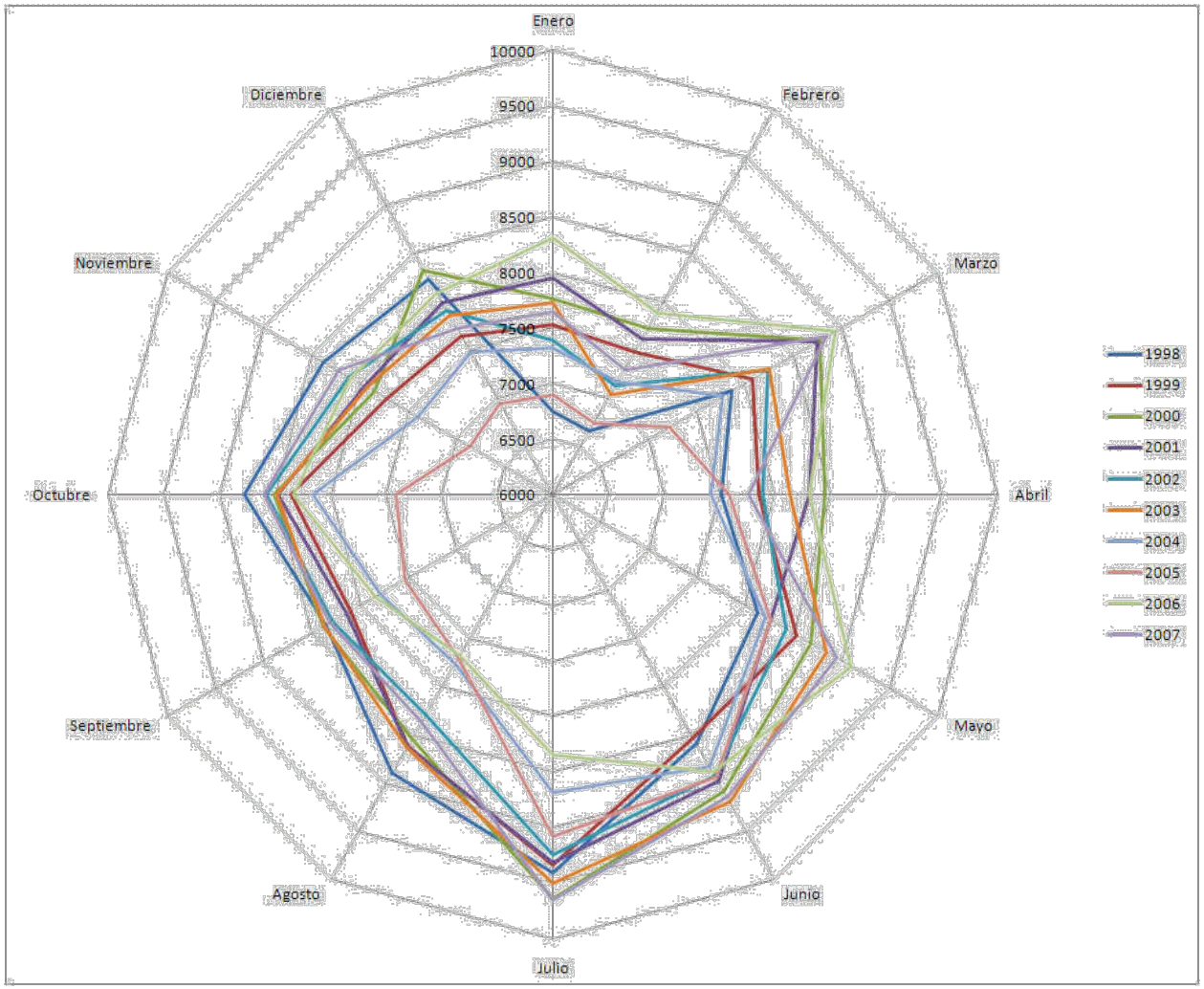
- Can be overwhelming; occlude data

Lends itself to “circular” relationships

- Time

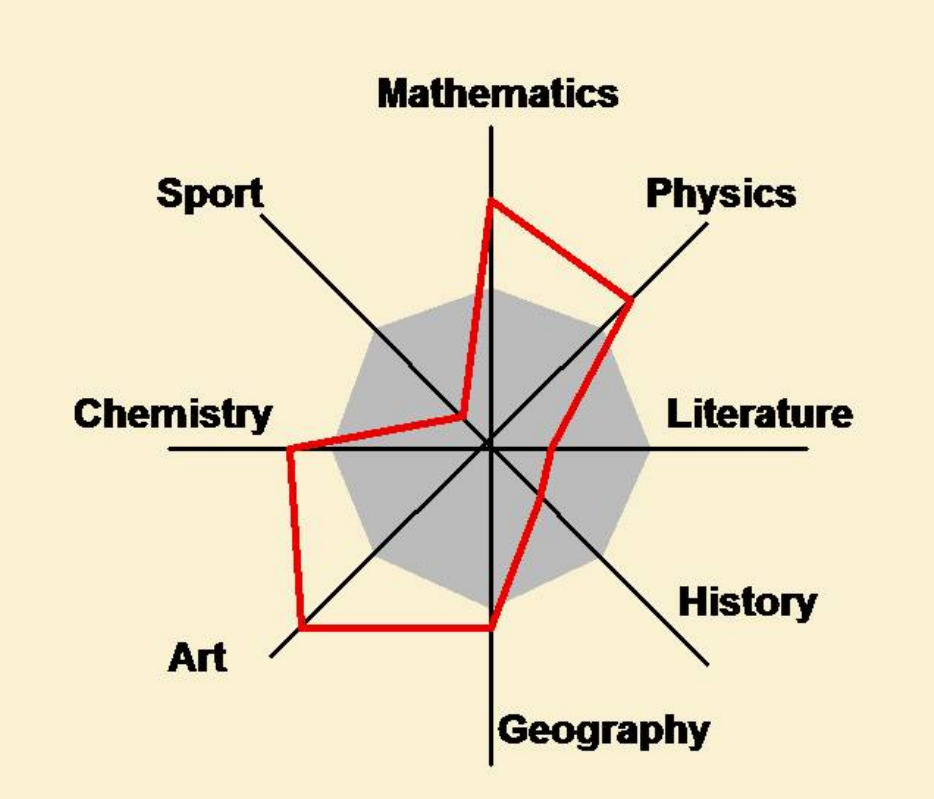


Star Plots

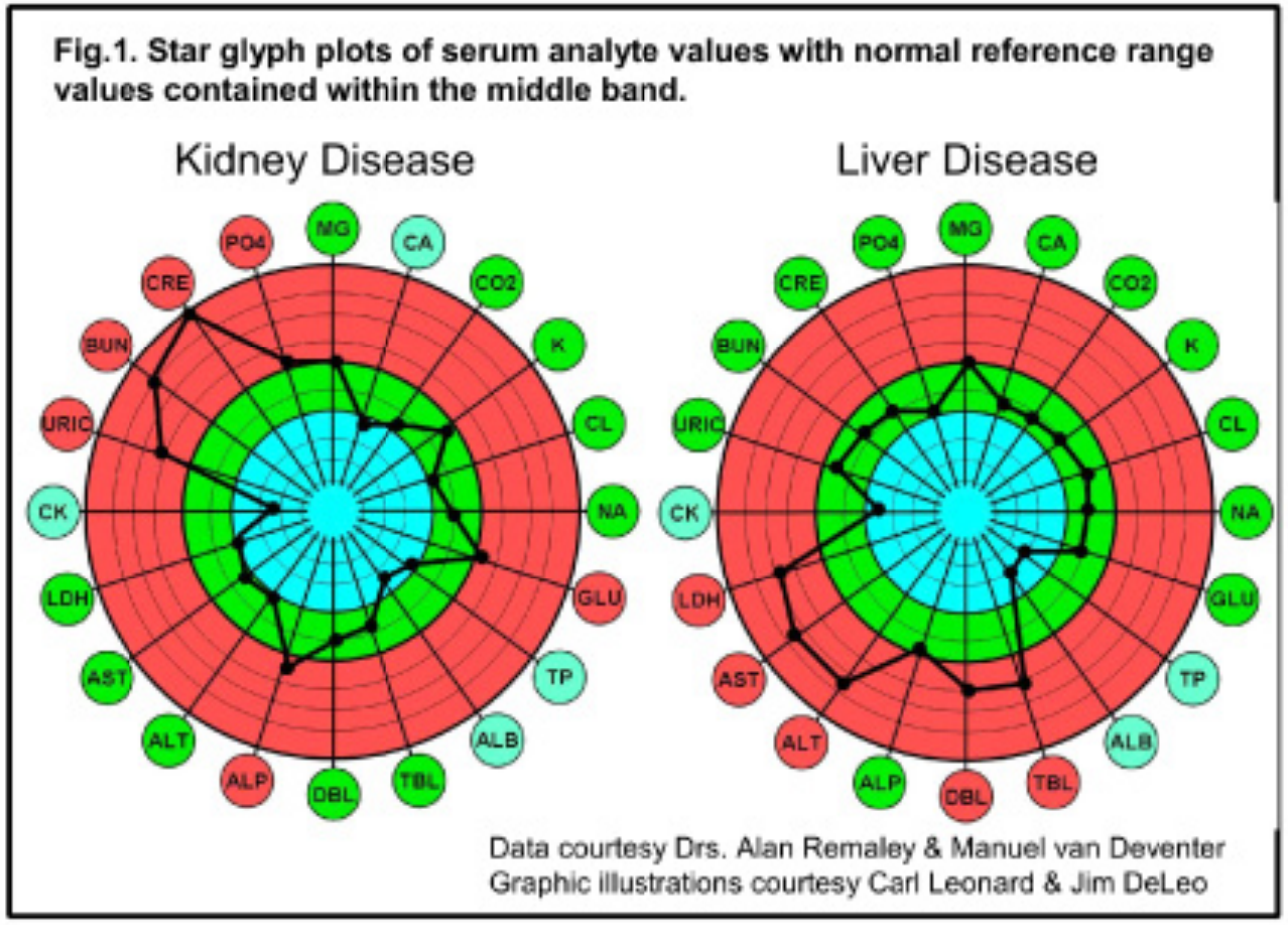


http://www.infovis.net/printMag.php?num=201&lang=2

Star Plots


Mathematics
Physics
Literature
History
Geography
Art
Chemistry
Sport

“Any reasonable number of values can be plotted in star glyph fashion”

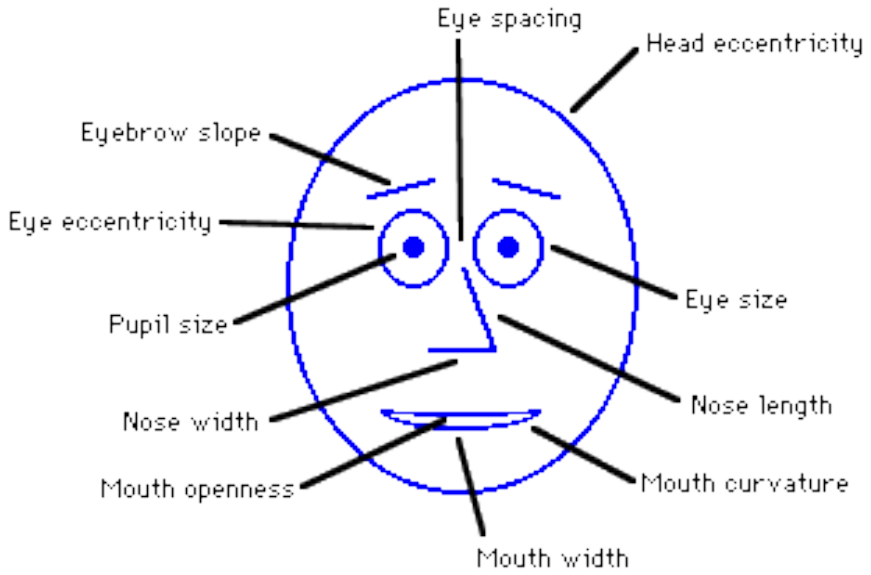


http://laboratory-manager.advanceweb.com/

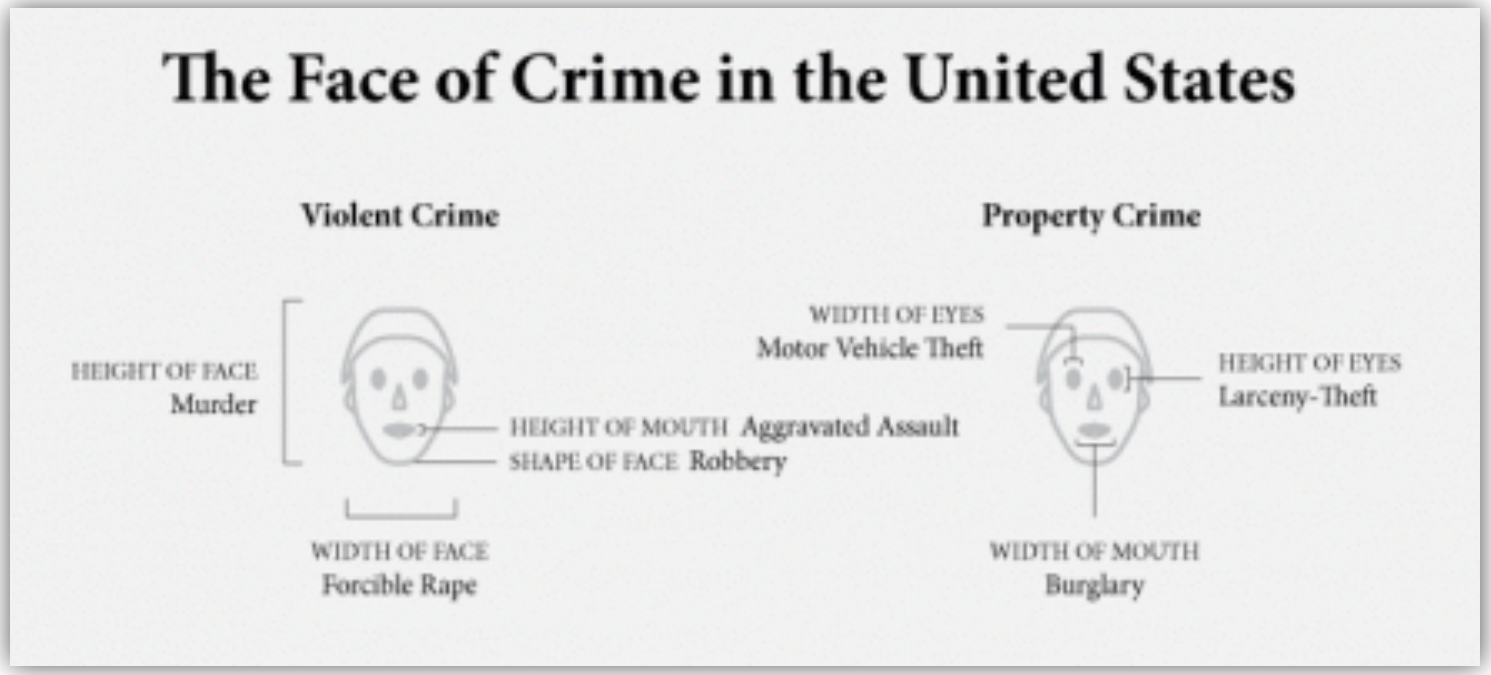
Chernoff Faces

10 Parameters:

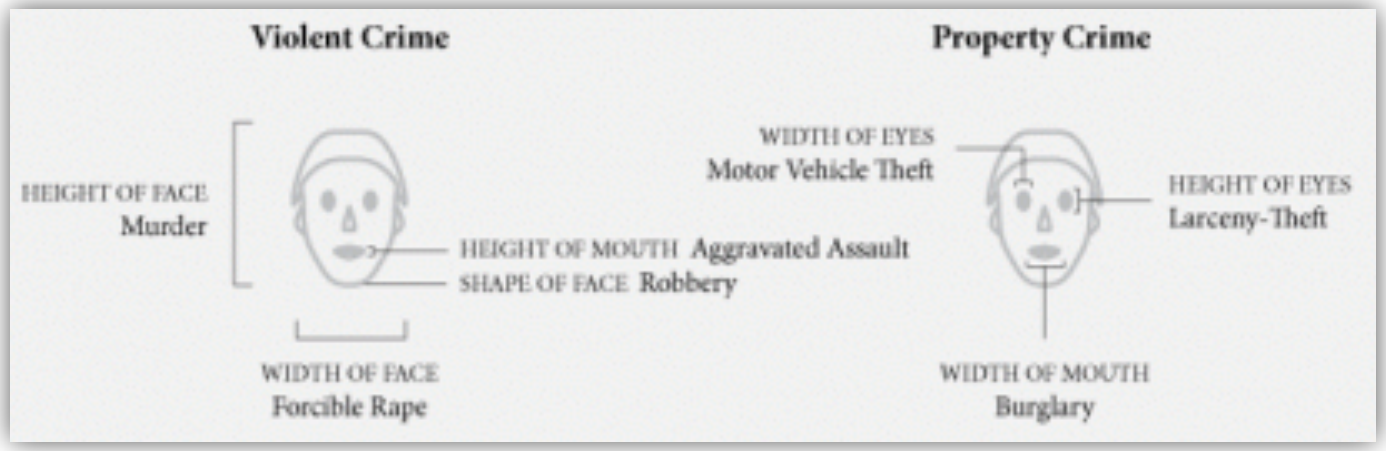
- Head Eccentricity
- Eye Eccentricity
- Pupil Size
- Eyebrow Slope
- Nose Size
- Mouth Vertical Offset
- Eye Spacing
- Eye Size
- Mouth Width
- Mouth Openness



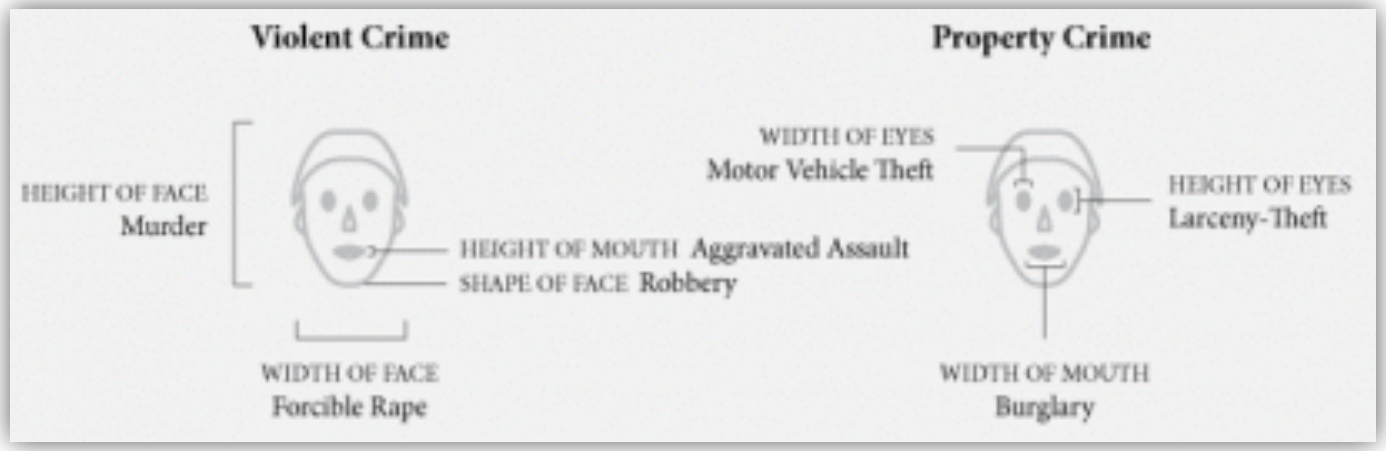
Chernoff Faces



http://flowingdata.com/2010/08/31/how-to-visualize-data-with-cartoonish-faces/

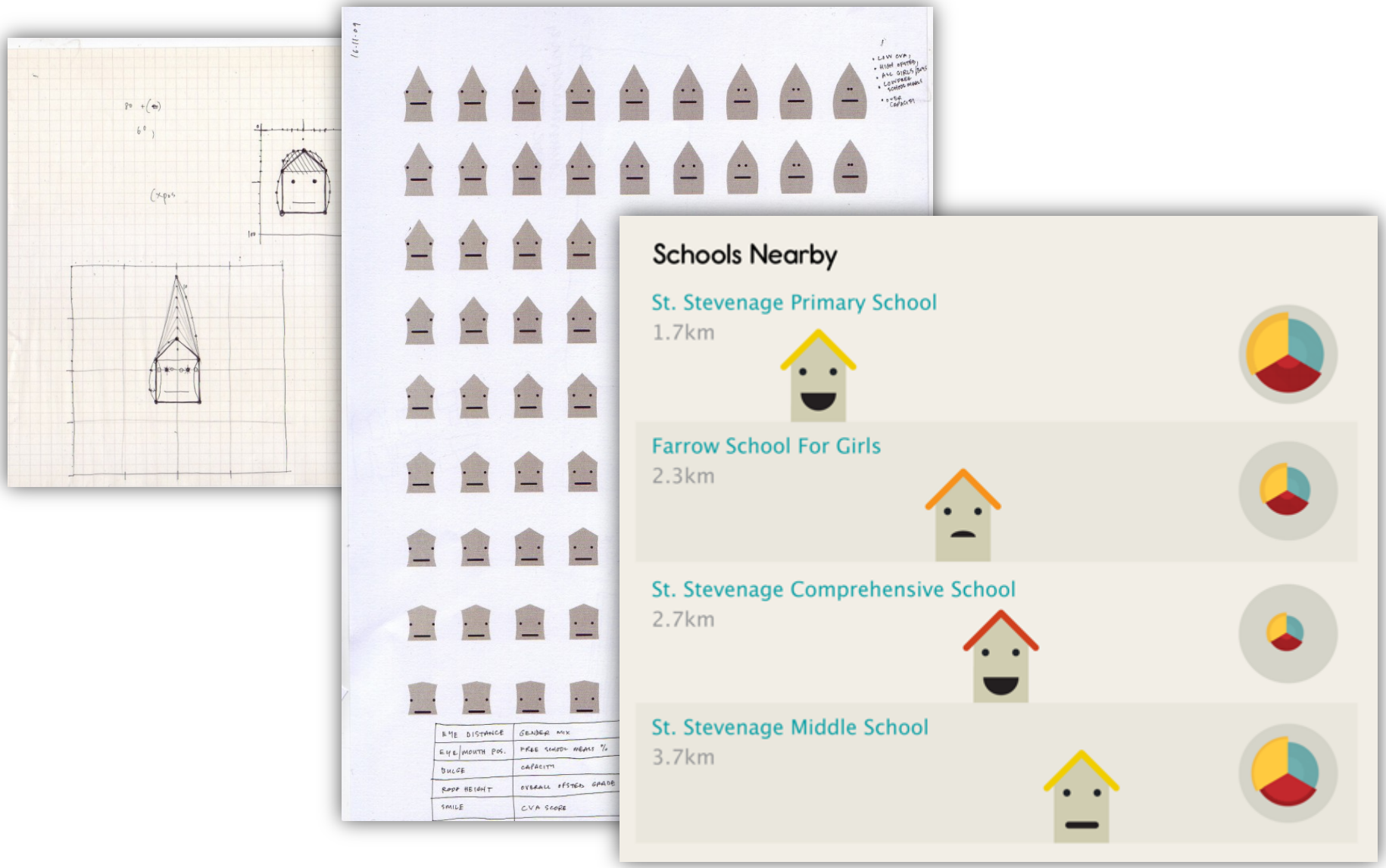


http://flowingdata.com/2010/08/31/how-to-visualize-data-with-cartoonish-faces/



http://flowingdata.com/2010/08/31/how-to-visualize-data-with-cartoonish-faces/

Chernoff Schools



http://berglondon.com/blog/tag/chernoff-faces/

Chernoff Faces
2005 National League

alexreisner.com/baseball/stats/chernoff

	PCT	H	HR	BB	SB
ARI	0.475	1419	191	606	67
ATL	0.556	1453	184	534	92
CHI	0.488	1506	194	419	65
CIN	0.451	1453	222	611	72
COL	0.414	1477	150	509	65
FLO	0.512	1499	128	512	96
HOU	0.549	1400	161	481	115
LAD	0.438	1374	149	541	58
MIL	0.500	1413	175	531	79
NYM	0.512	1421	175	486	153
PHI	0.543	1494	167	639	116
PIT	0.414	1445	139	471	73
SDP	0.506	1416	130	600	99
SFG	0.463	1427	128	431	71
STL	0.617	1494	170	534	83
WAS	0.500	1367	117	491	45

Sometimes less successful…

http://www.alexreisner.com/baseball/chernoff-faces

Parallel Coordinates

Not scatter but you get the idea…

Parallel Plots

Create one vertical line for every variable

Minimum and Maximum values / multiple scales

Plot every entity across each variable

Line connects values for a single entity

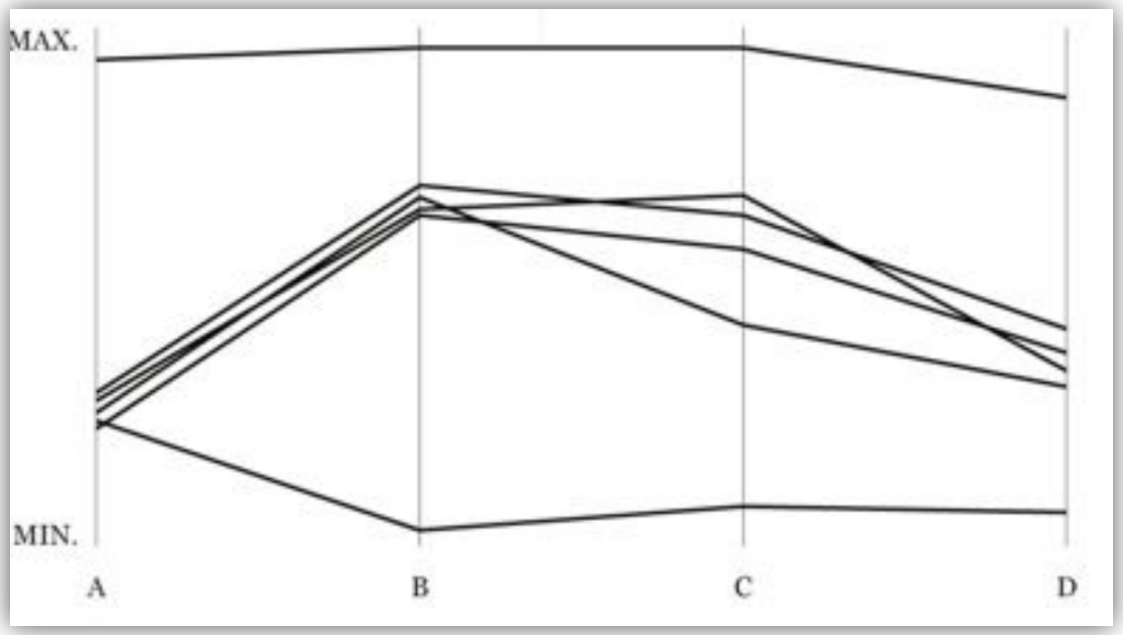
Picture an xy scatterplot, but putting both axis lines vertically

No zero point for singularities

Better balance of ink/data ratio, although can be abused

CONNECTING LINES

One line per unit; look for *common trends* across multiple units



RELATIVE SCALES

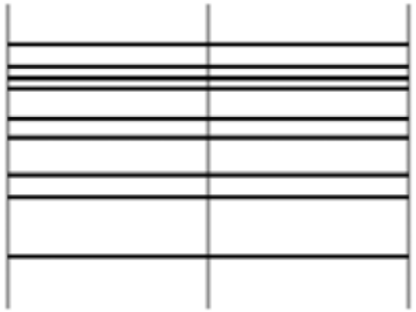
Axes span minimum to maximum for *each* variable.

VARIABLES

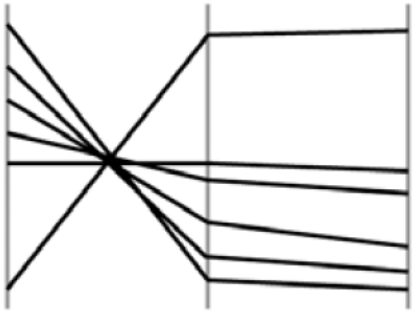
Multiple axes placed parallel to each other to find relationships *across variables*

Nathan Yau “Visualize This” 2011 and “Data Points” 2013

Positive correlation
Lines run parallel



Negative correlation
Lines cross consistently



Weak correlation
No clear direction

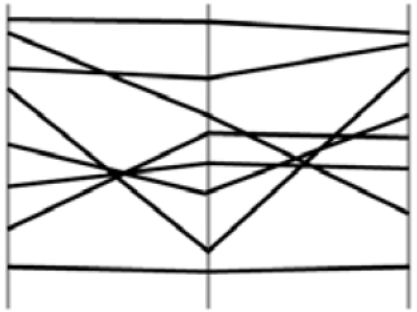
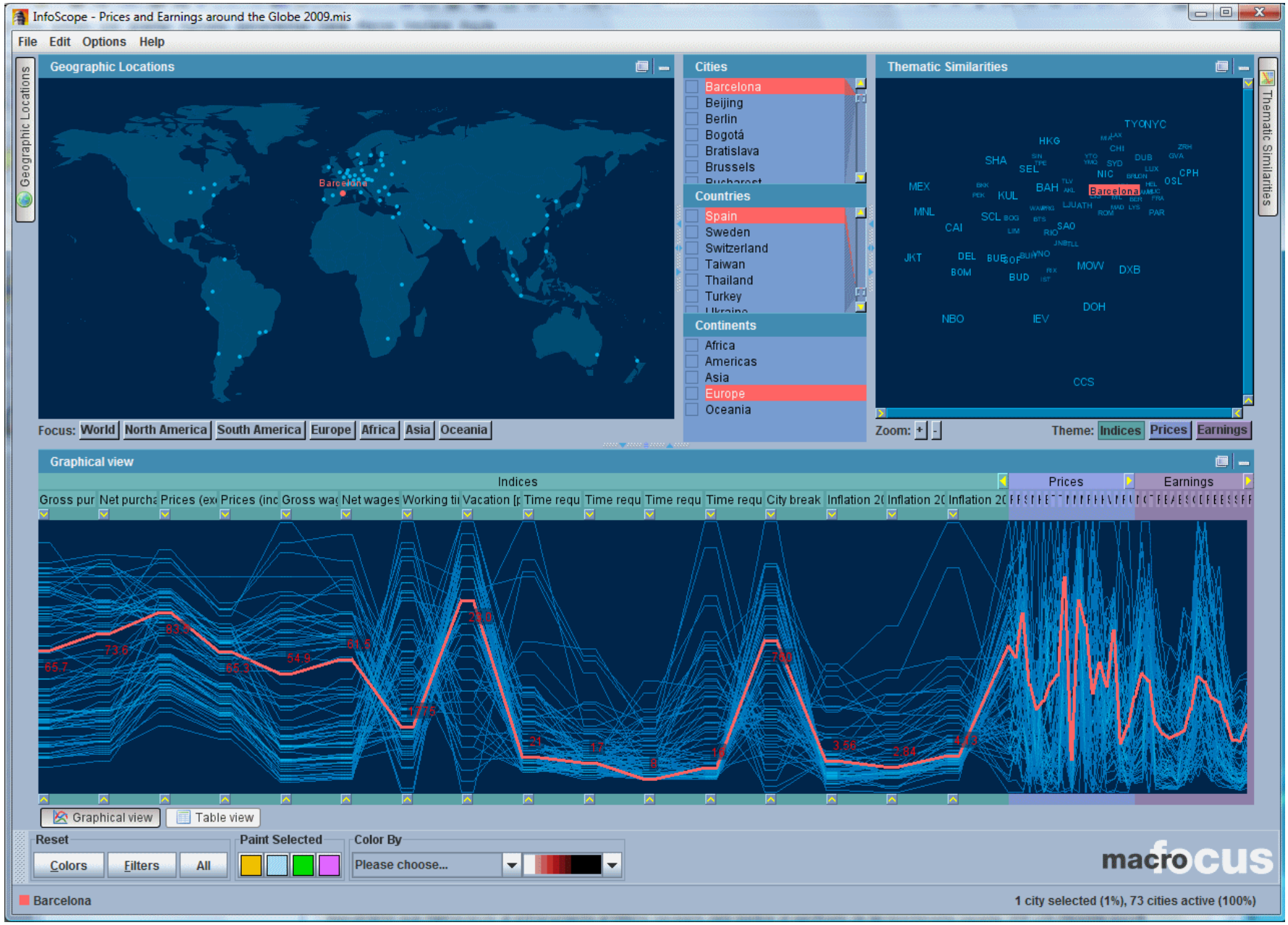
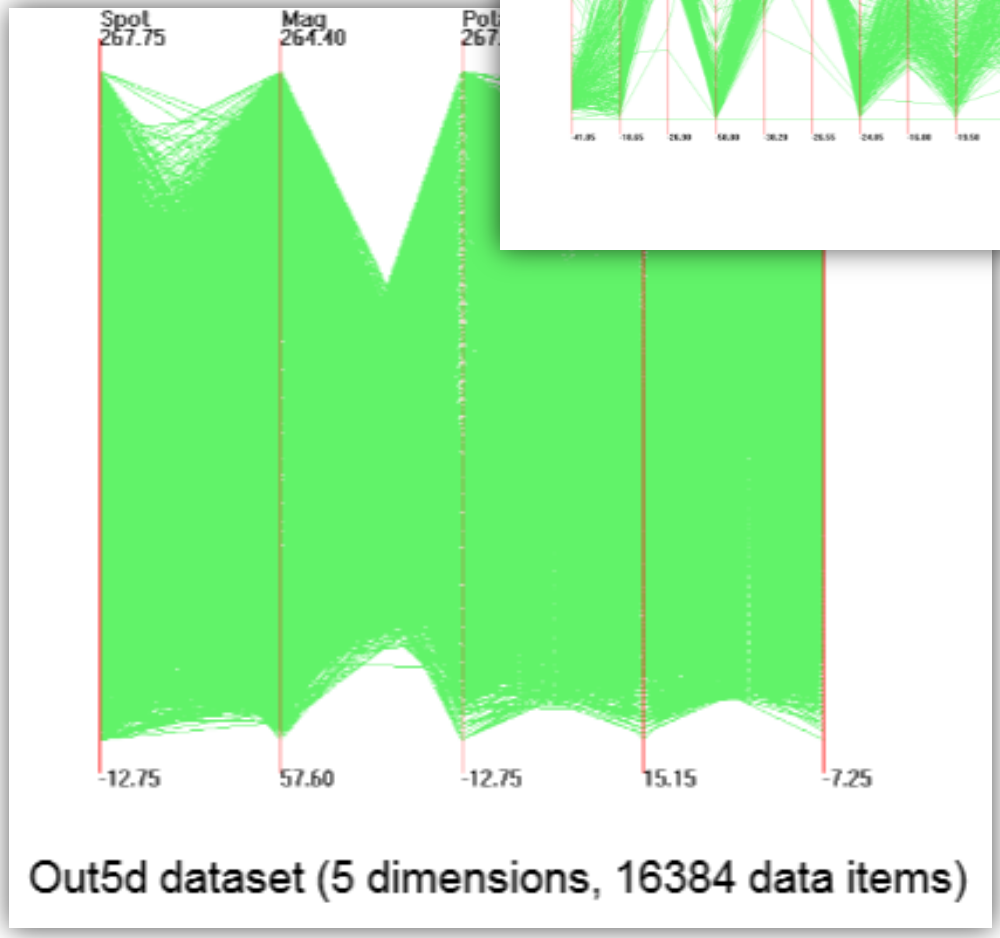


FIGURE 4-45 *Relationships with parallel coordinates plot*

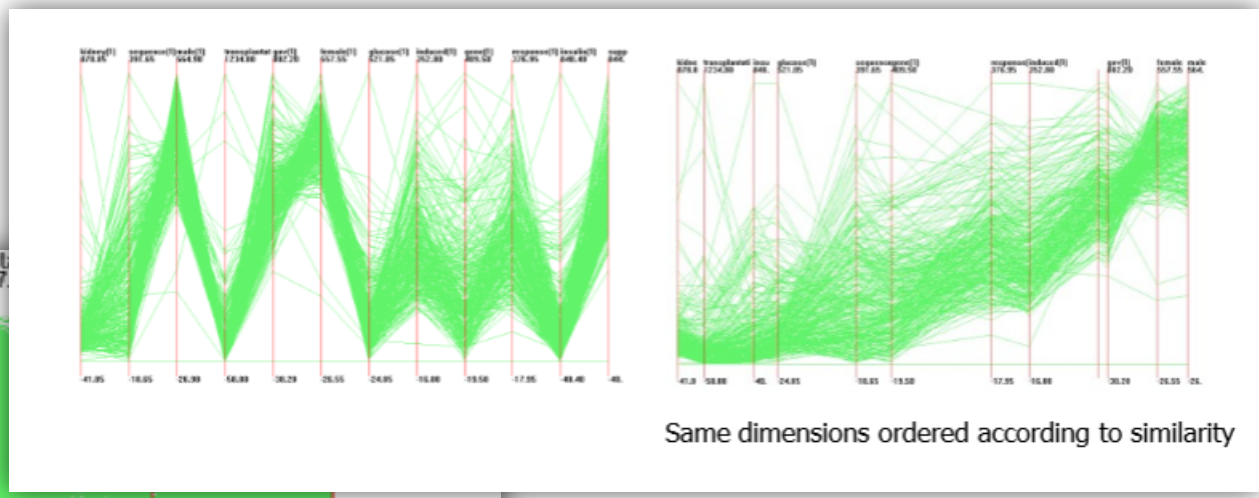
Nathan Yau “Visualize This” 2011 and “Data Points” 2013



http://www.infovis.net/imagenes/T1_N201_A1398_InfoScopeBcn.gif

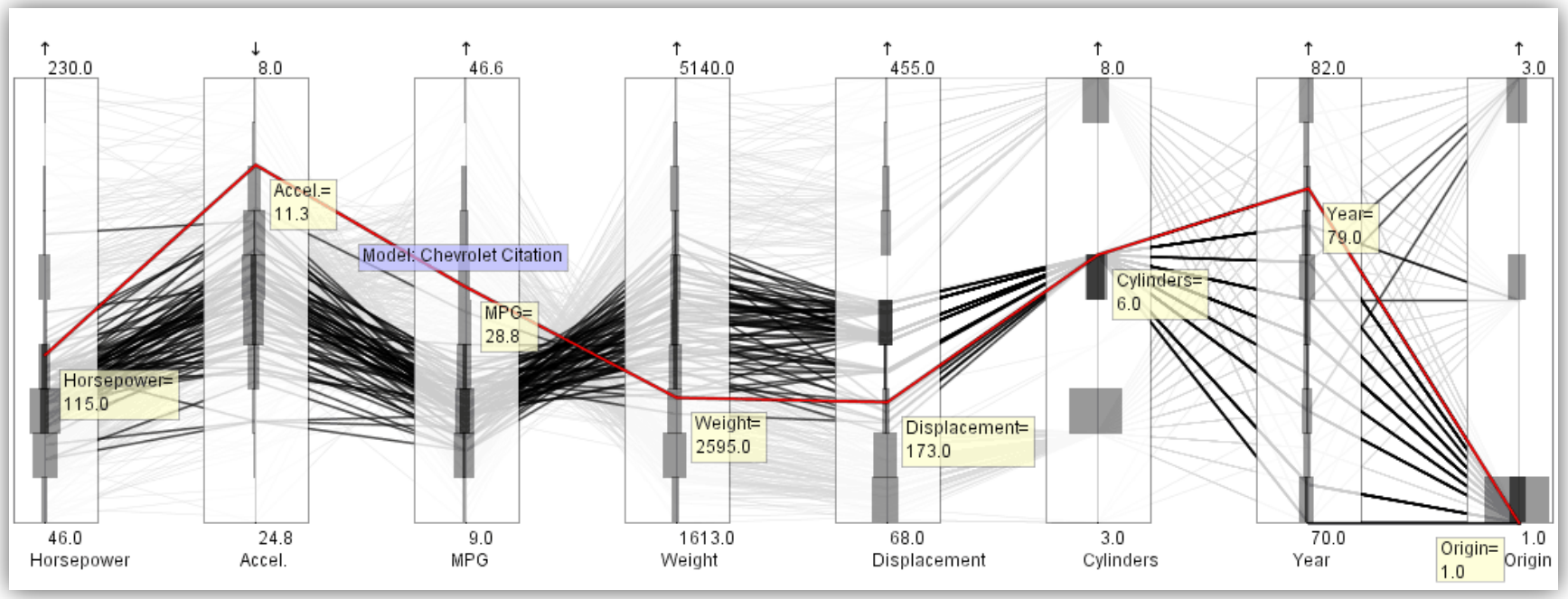


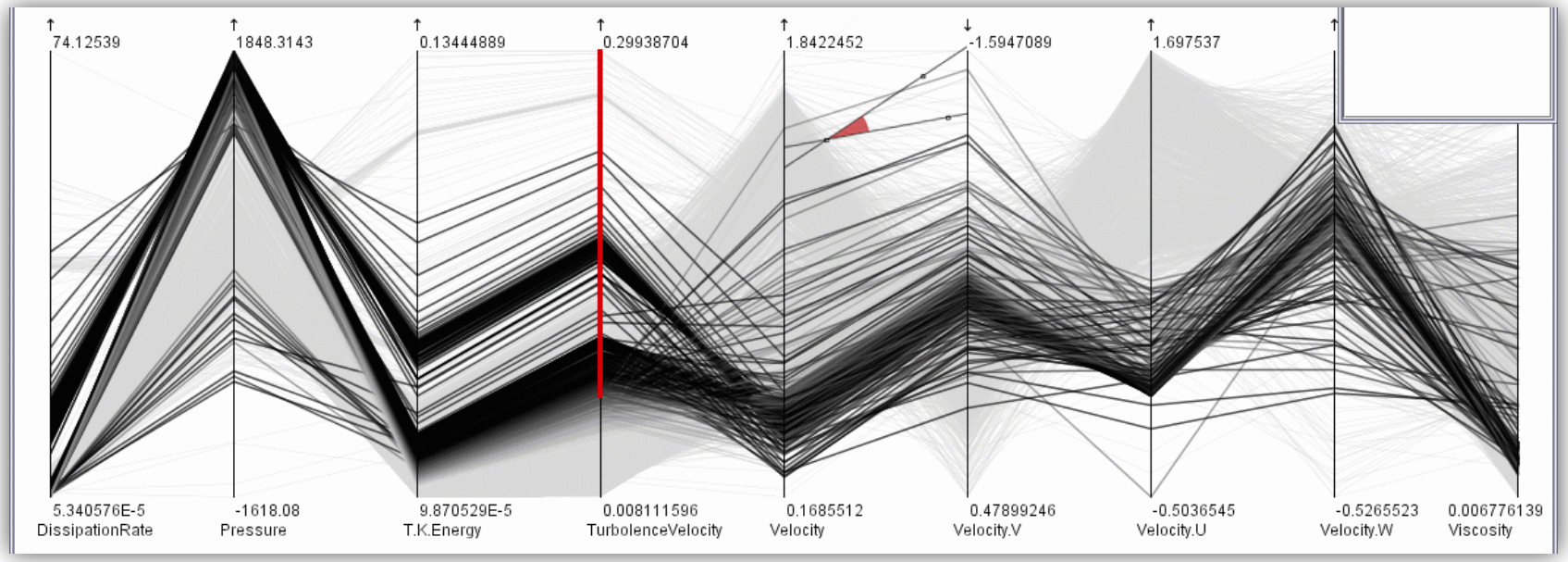
Out5d dataset (5 dimensions, 16384 data items)



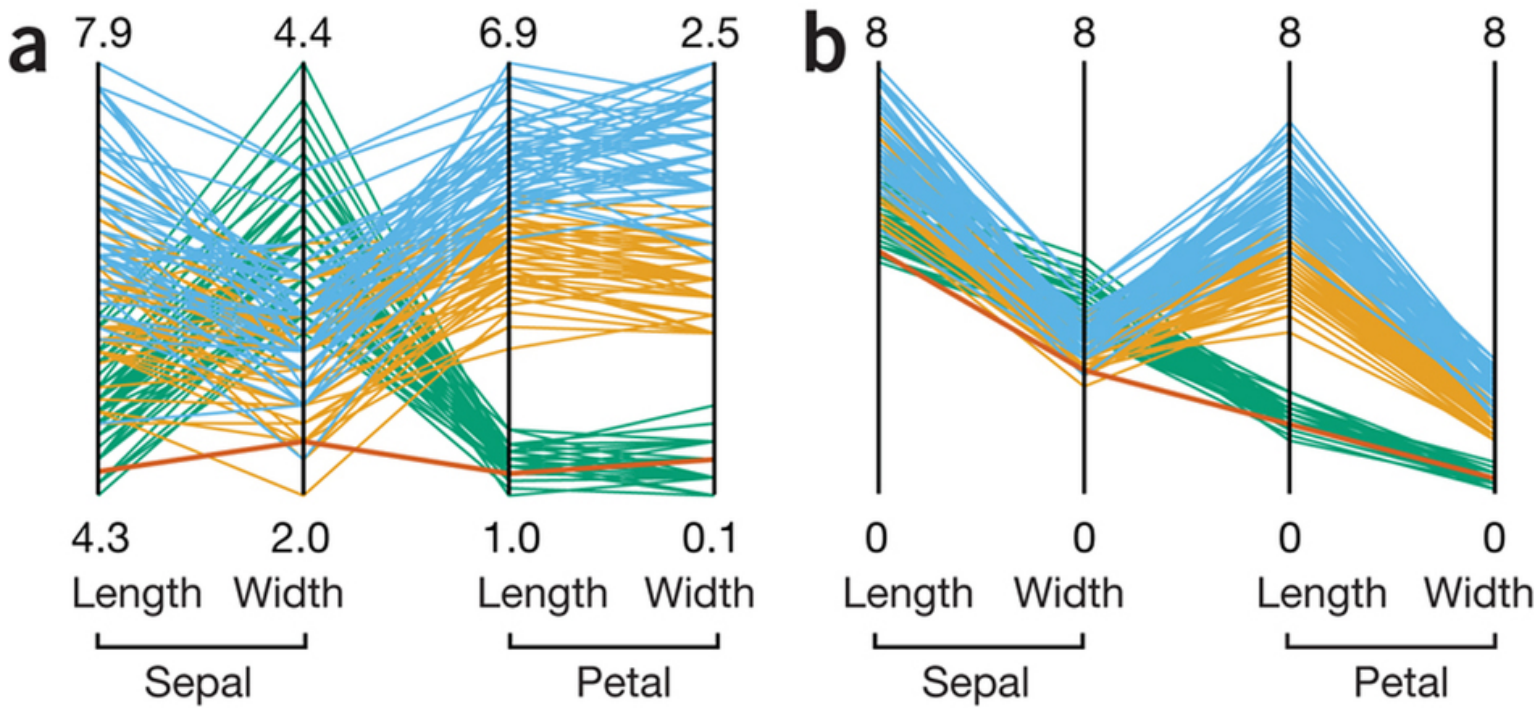
Same dimensions ordered according to similarity

http://www.cc.gatech.edu/~stasko/7450/Notes/multivar1.pdf courtesy J Yang



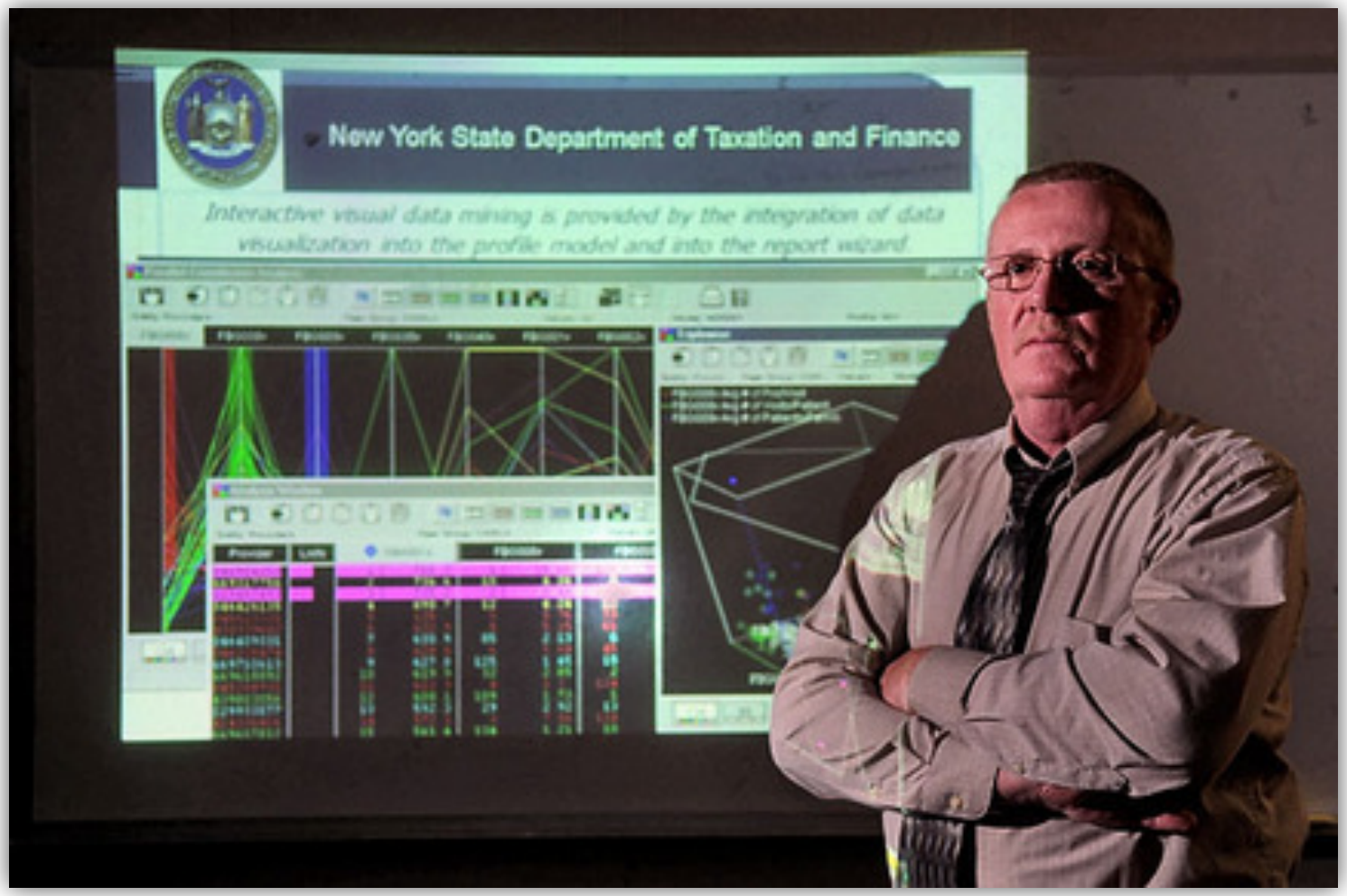


http://old.vrvis.at/via/research/ang-brush/teaser.gif



Krzywinski and Savig "Points of view: Multidimensional data" Nature Methods 10, 595 (2013) doi:10.1038/nmeth.2531

Fight Crime with Parallel Plots!



http://www.syracuse.com/news/index.ssf/2010/01/data_mining_helps_new_york_cat.html

Heatmaps/Matrix Plots

So colorful!

A common problem: 3-dimensional data

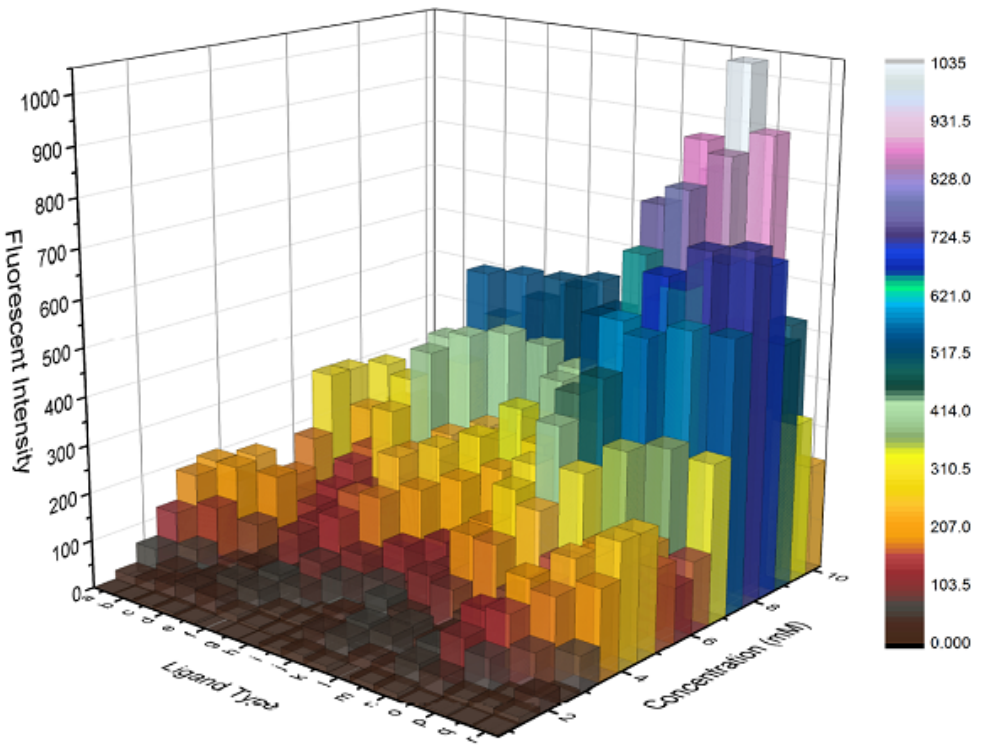
Data with three dimensions:

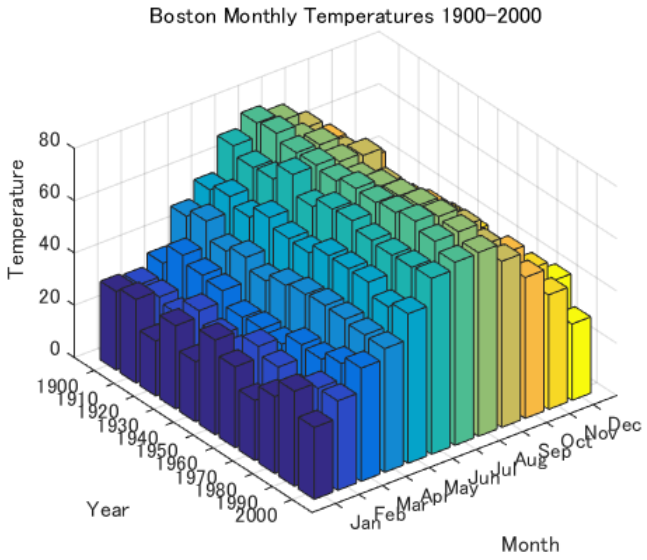
time-point, molarity, expression level

treatment group, location, intensity

X-Y location + some other continuous value

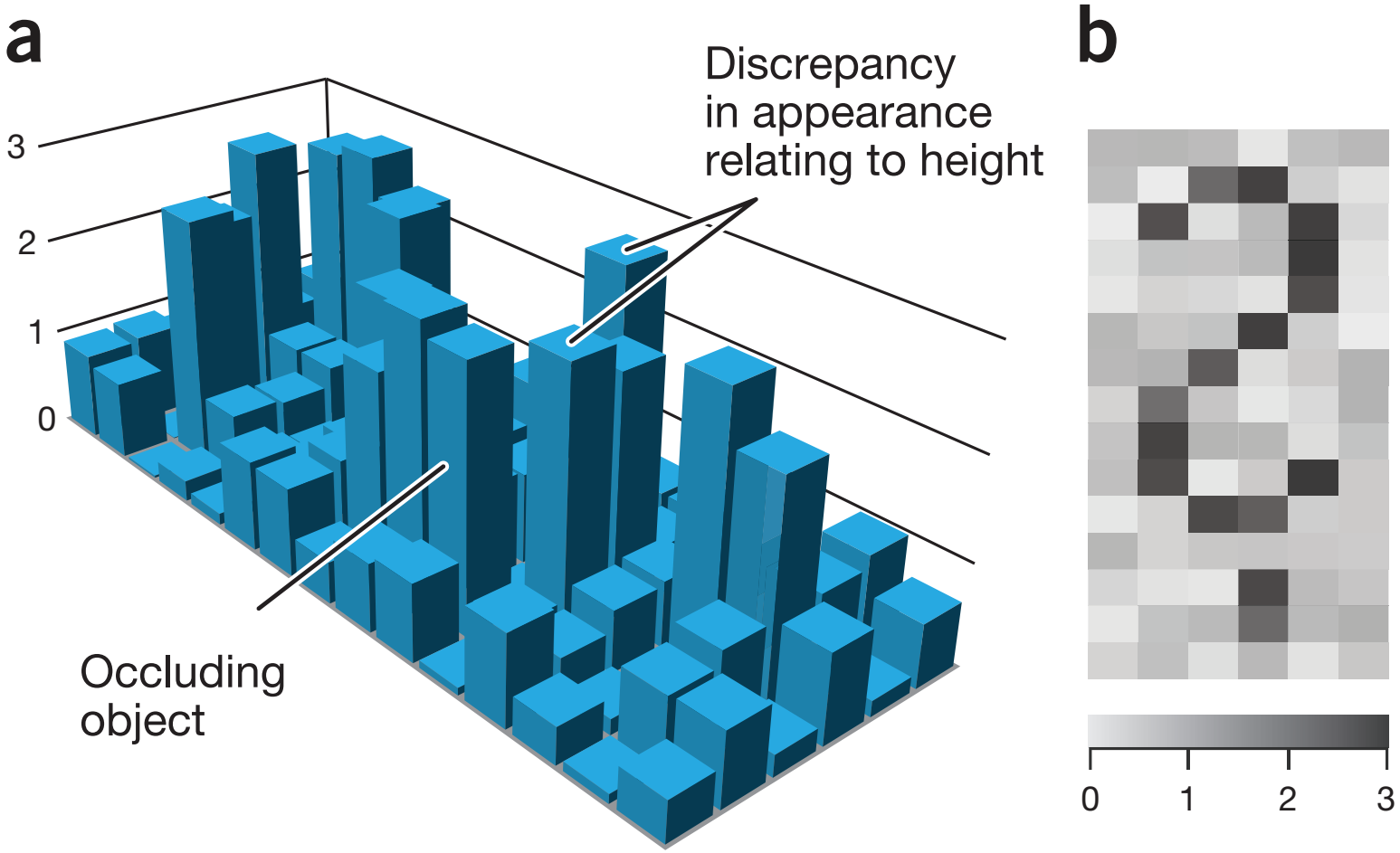
One“solution”: 3-dimensional bar graphs





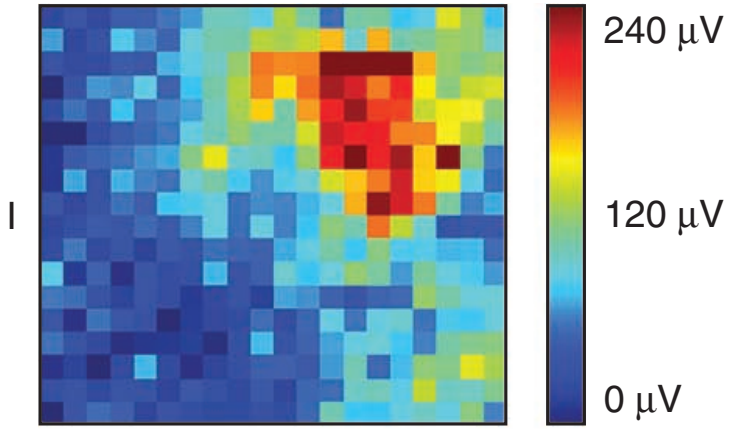
http://www.originlab.com/index.aspx?go=Products/Origin/Graphing
http://www.mathworks.com/matlabcentral/fileexchange/35274-matlab-plot-gallery-bar-graph-3d/content/html/Bar_Plot_3D.html

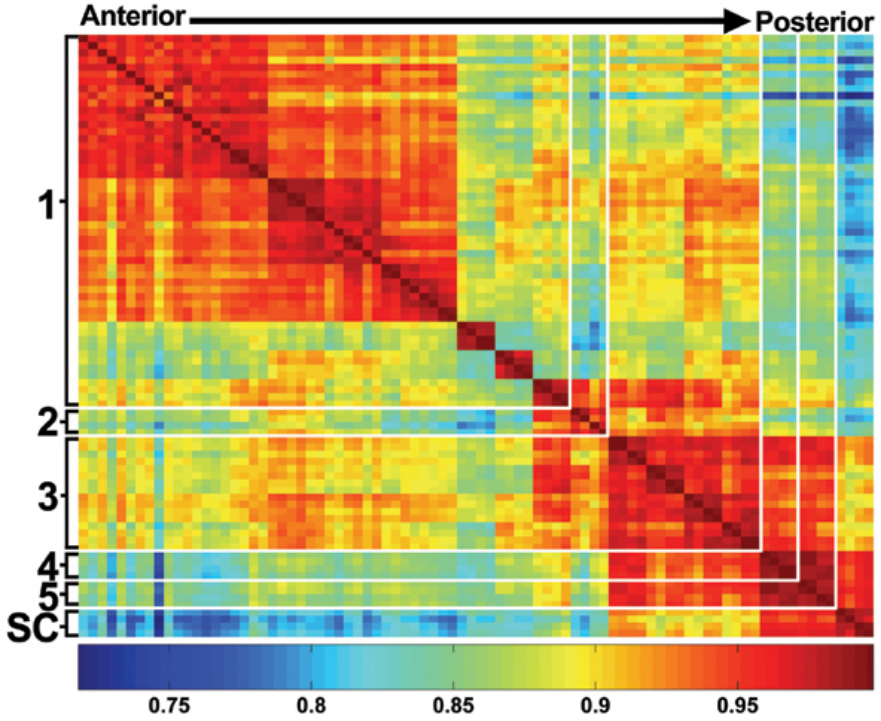
These are almost never a good idea.



Gehlenborg N, Wong B. Points of view: Into the third dimension. Nat Meth. 2012 Aug 30;9(9):851–1.

A better solution: Heatmaps

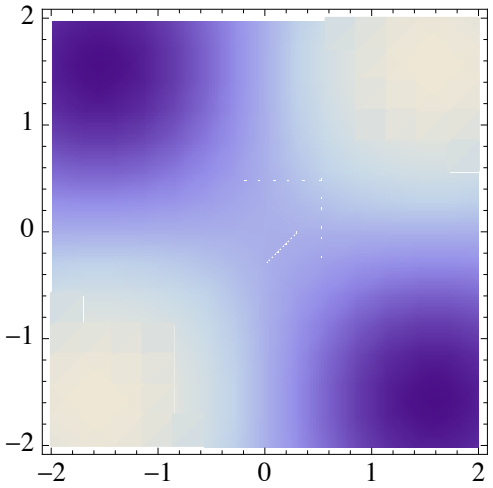


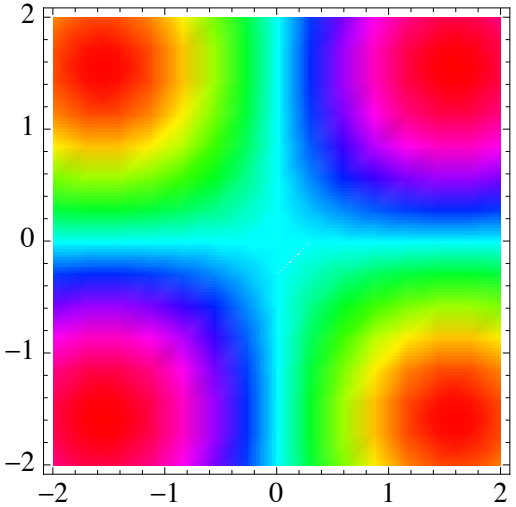


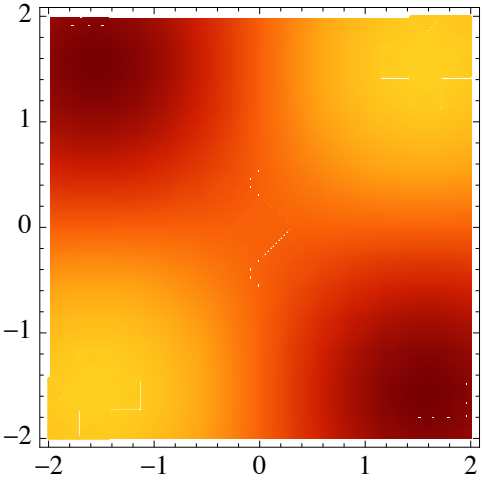
Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci USA. 2006 Dec 19;103(51):19430–5.

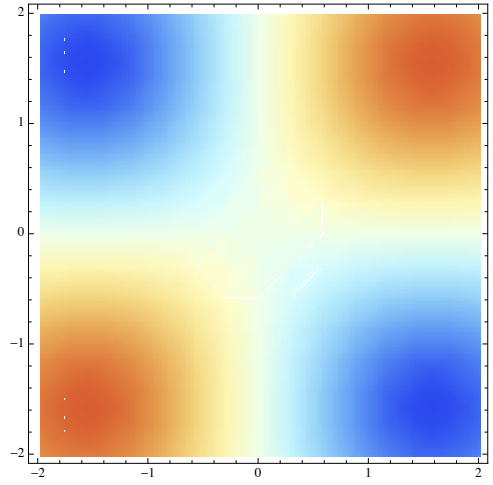
Viventi J, Kim D-H, Vigeland L, Frechette ES, Blanco JA, Kim Y-S, et al. Nat Neurosci. 2011 Nov 13;14(12):1599–605.

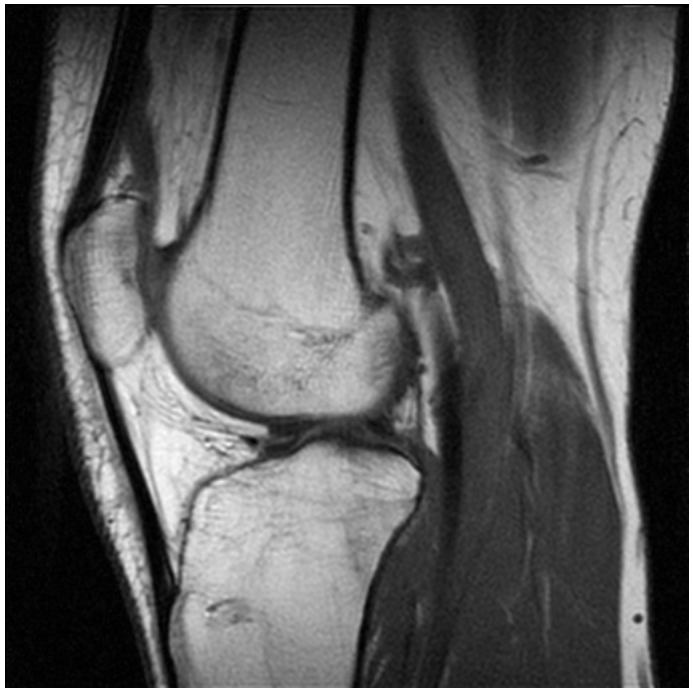
One important consideration: choice of color scale

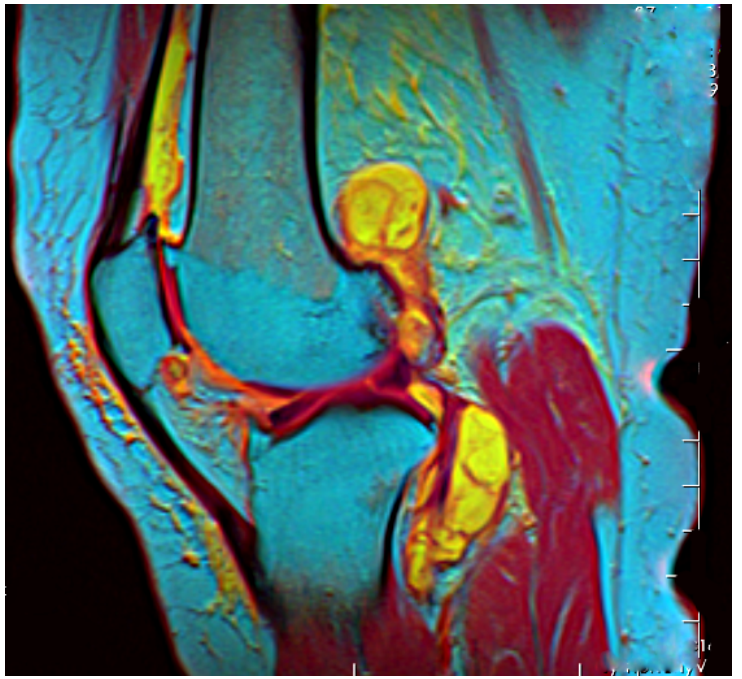


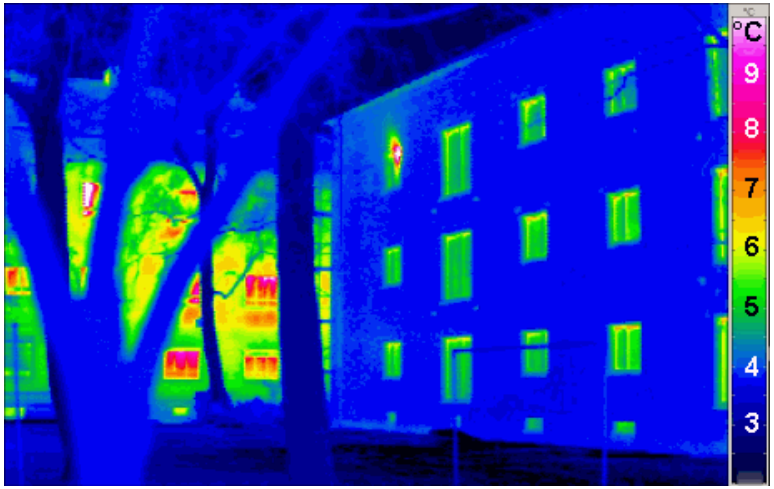










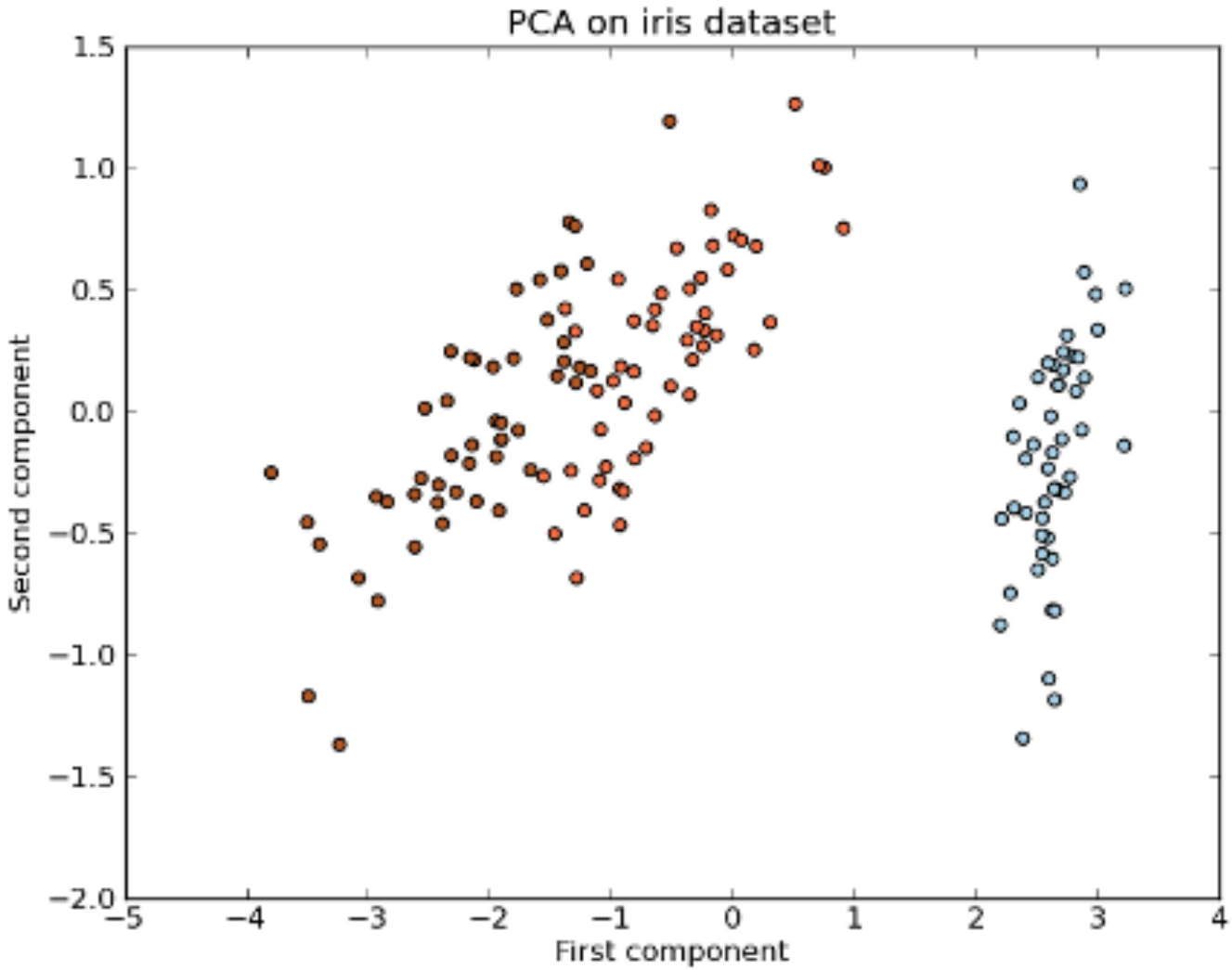


http://en.wikipedia.org/wiki/False_color

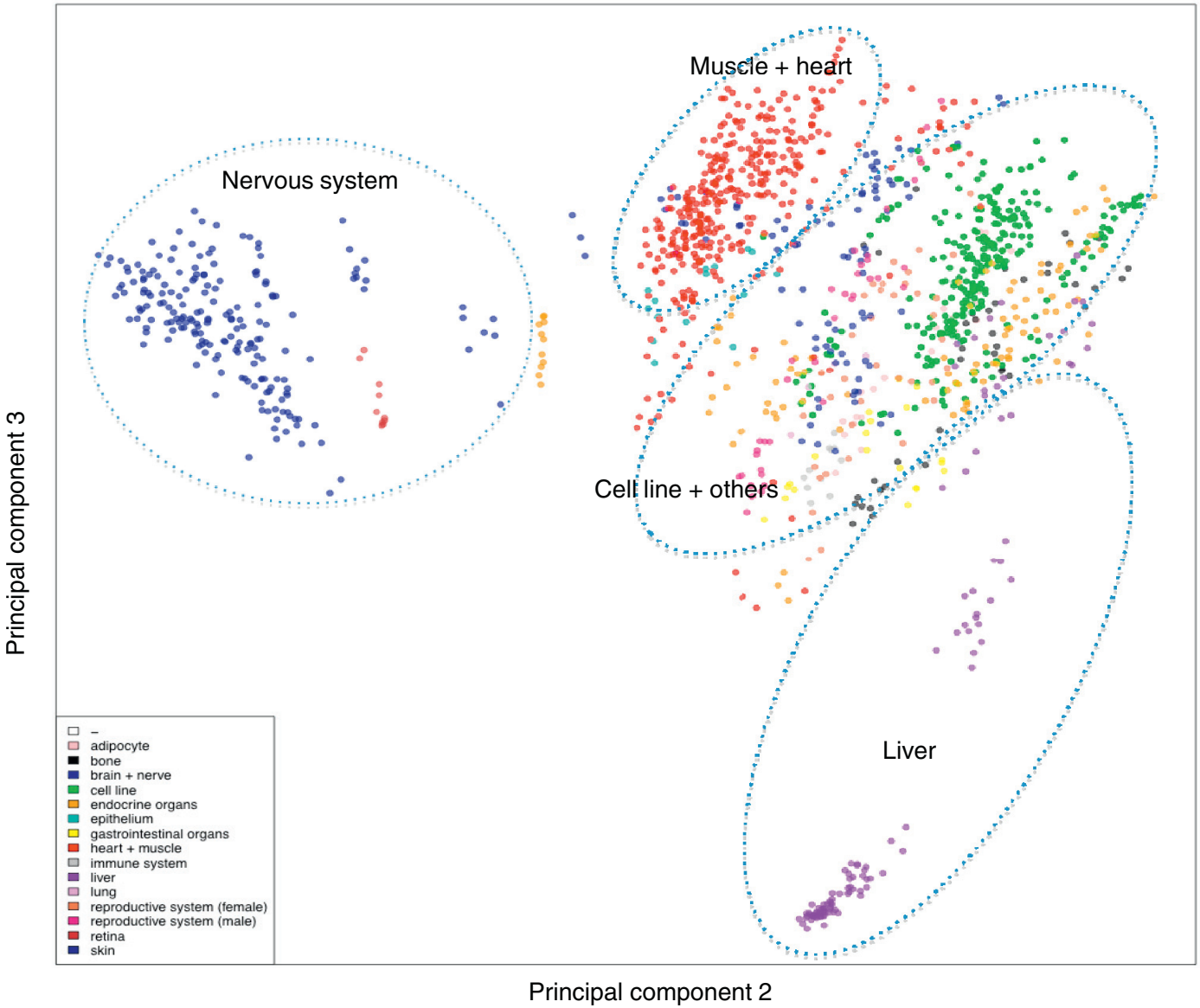
Another alternative: dimensionality reduction

Many techniques exist to take high-dimensional data and project it into fewer dimensions:

Principal component analysis, etc.



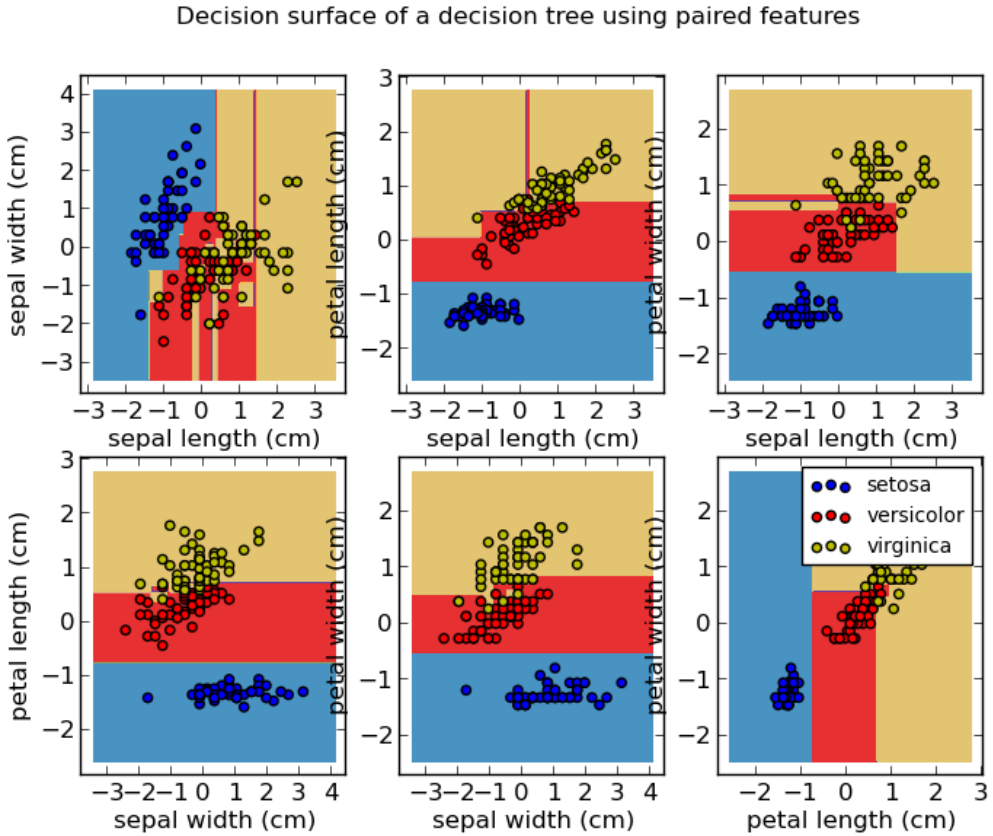
http://mlpy.sourceforge.net/docs/3.2/tutorial.html

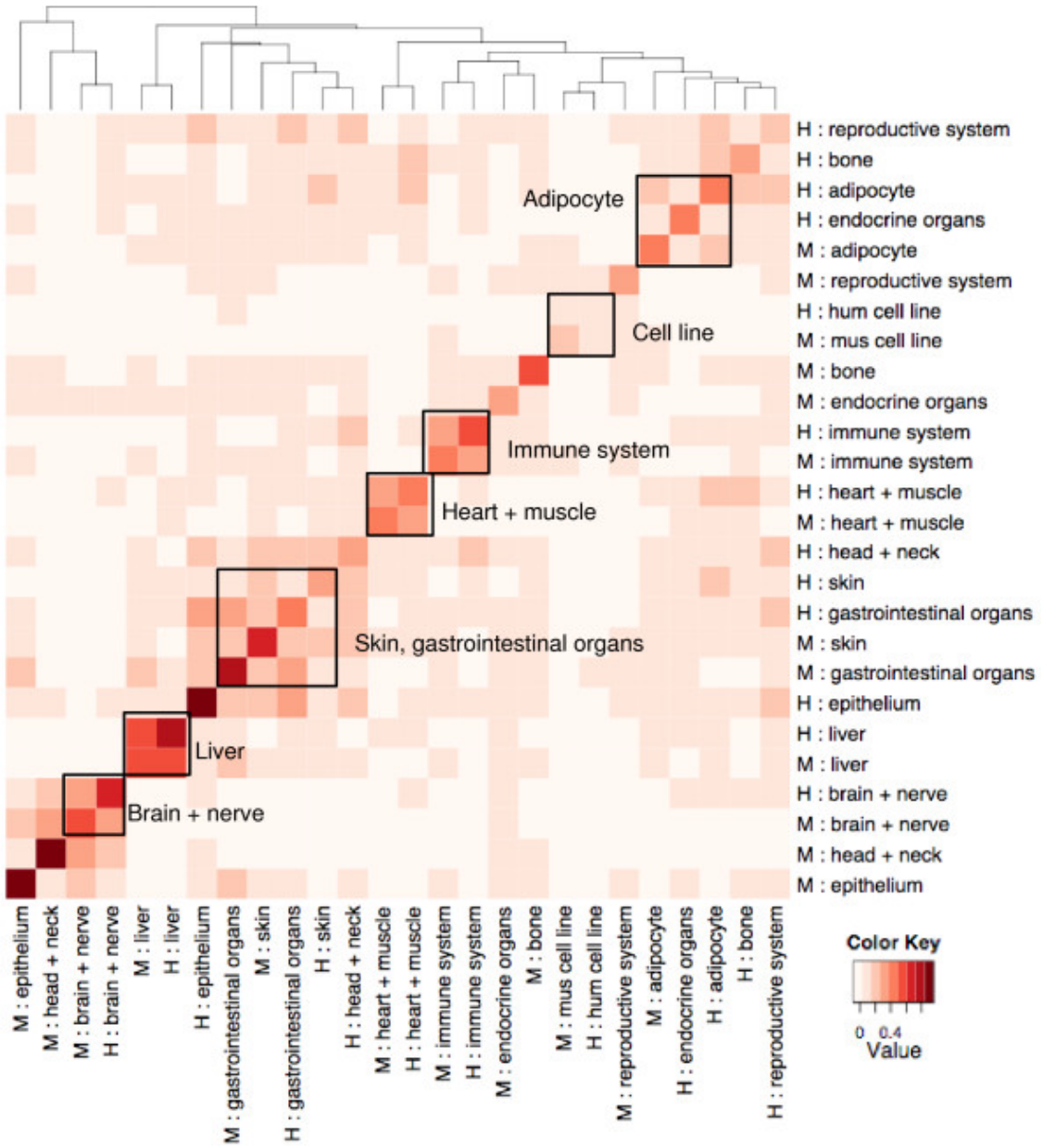


Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. Genome Biol. 2010;11(12):R124.

Don’t forget annotations!

Adding additional information (*not data*) to the graph can be very useful when many dimensions are involved.





Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. Genome Biol. 2010;11(12):R124.

NATURE METHODS | METHAGORA

Data visualization: A view of every Points of View column

30 Jul 2013 | 8:08 AM | Posted by Daniel Evanko | Category: Featured, Visualization

We've organized all the Points of View columns on data visualization published in *Nature Methods* and provide this as a guide to accessing this trove of practical advice on visualizing scientific data.

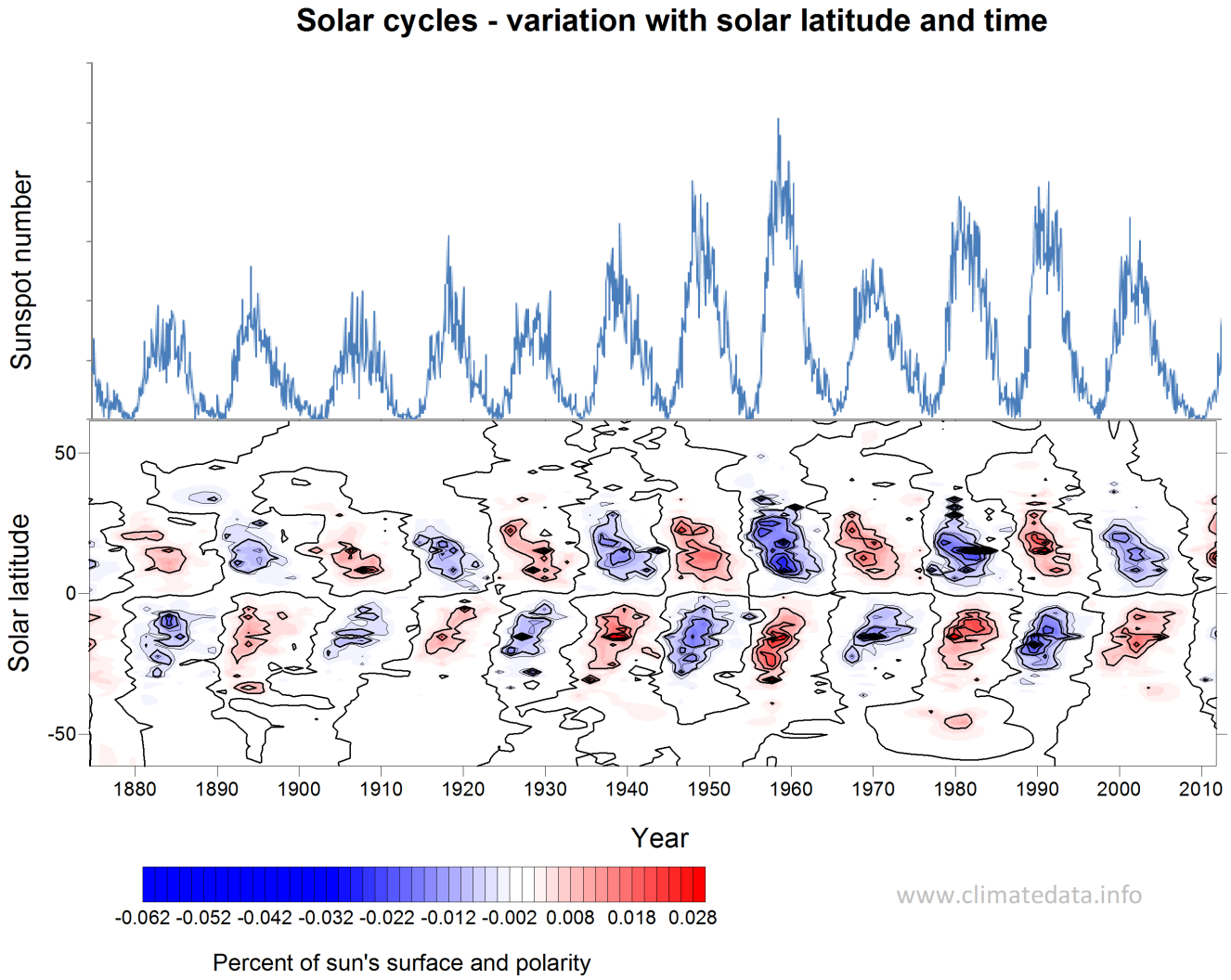
As of July 30, 2013 *Nature Methods* has published 35 Points of View columns written by Bang Wong, Martin Krzywinski and their co-authors: Nils Gehlenborg, Cydney Nielsen, Noam Shoresh, Rikke Schmidt Kjærgaard, Erica Savig and Alberto Cairo. As we prepare to launch a new column in our September issue we felt this would be a good time to collect and organize links to all the Points of View columns together in one place to make it easier to navigate this wonderful resource that the authors have provided us. For the month of August we will be making all the columns free to access so everyone can benefit from this practical advice on data visualization.

This should not be the end of the Points of View column though. We will be inviting new visualization experts to author columns on new topics that have not been covered so far or which can be expanded on. This page will be continuously updated whenever a new column is published so stay tuned. If you have a suggestion for a topic you would like to see covered in a future points of view column please comment below.

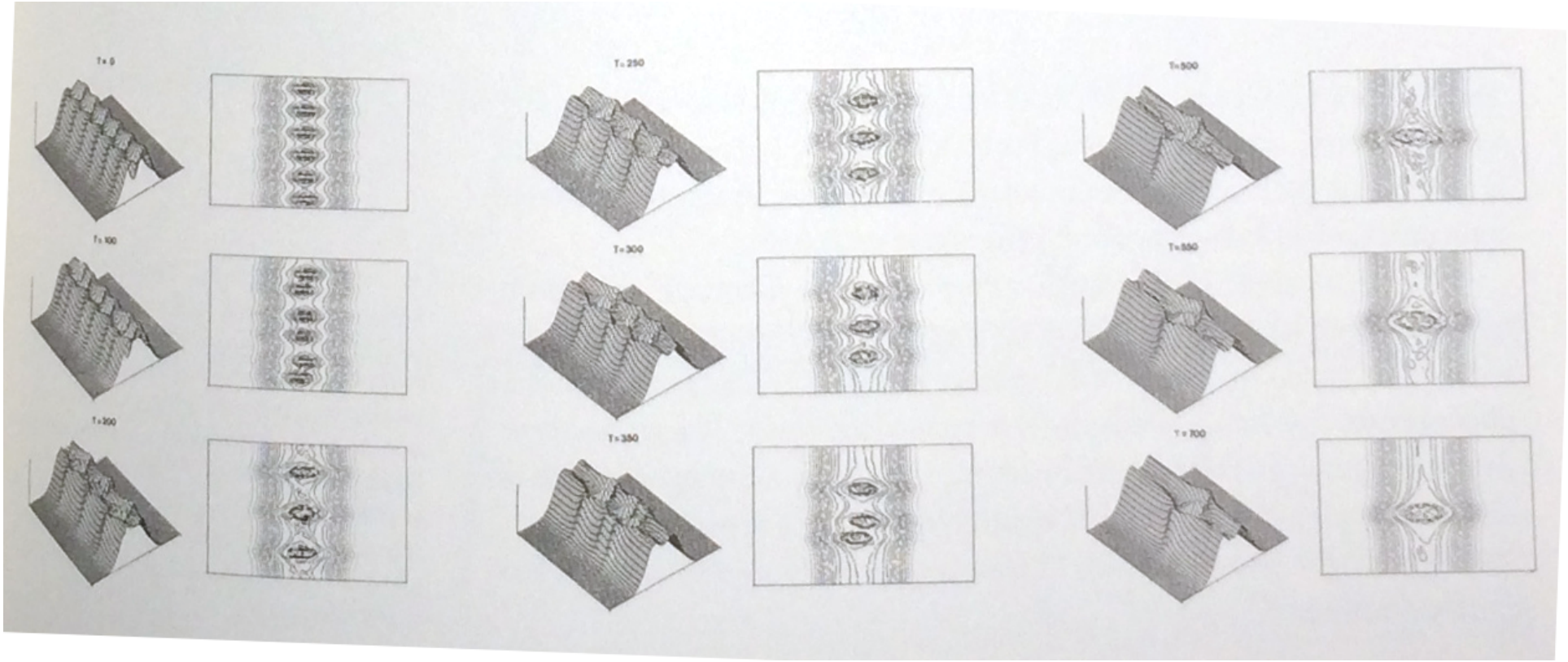
. . . . . .

http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html

Bonus!


Solar cycles - variation with solar latitude and time
Sunspot number
Solar latitude
50
0
-50
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
Year
-0.062 -0.052 -0.042 -0.032 -0.022 -0.012 -0.002 0.008 0.018 0.028
Percent of sun's surface and polarity
www.climatedata.info

Bonus!



Ghizzo A, Izrar B, Bertrand P, Fijalkow E. Stability of Bernstein–Greene–Kruskal plasma equilibria. Numerical experiments over a long time. Physics of Fluids 1988 31(1).