ARC²S Group

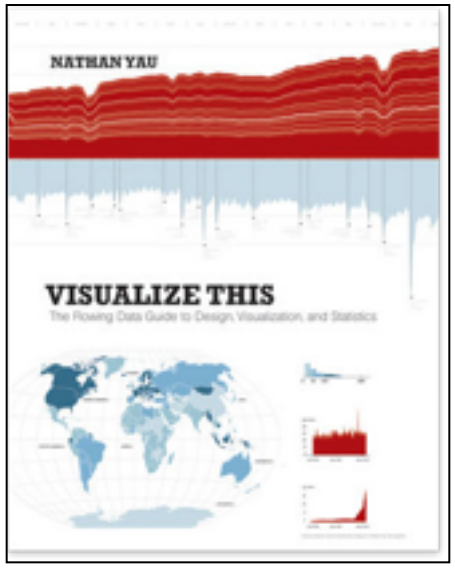
Applied Research on Computational Complex Systems

Multivariate data visualization

Prof. Giancarlo **Ruffo**

"Analisi e Visualizzazione di Reti Complesse" (9 credits)
Laurea Magistrale in **Informatica**
Università degli Studi di Torino
A.A. 2018/19

@giaruffo



Nathan Yau

Visualize This
The FlowingData Guide to Design, Visualization, and Statistics
Wiley, 2011

Chapter 7

http://flowingdata.com/

Outline

- Bubble Chart
- Heatmap
- Scatter plot Matrix
- Small Multiples
- Parallel Coordinates
- Spider Plots

- *Encoding Data*
- *Examples from Book*
- *Examples from D3*
- *Examples from R*
- *Examples from Web*
- *Discussion*

DATA TYPES

Revisited

Data Types

- Numerical
  - Continuous (-11.2, -1.3, 0.4, 4.8, 14.9, …)
  - Discrete (-2, -1, 0, 1, 2, …)
- Categorical
  - Ordered (January, February, March, April, …)
  - Unordered (Red, Blue, Green, Purple, …)

Data Types

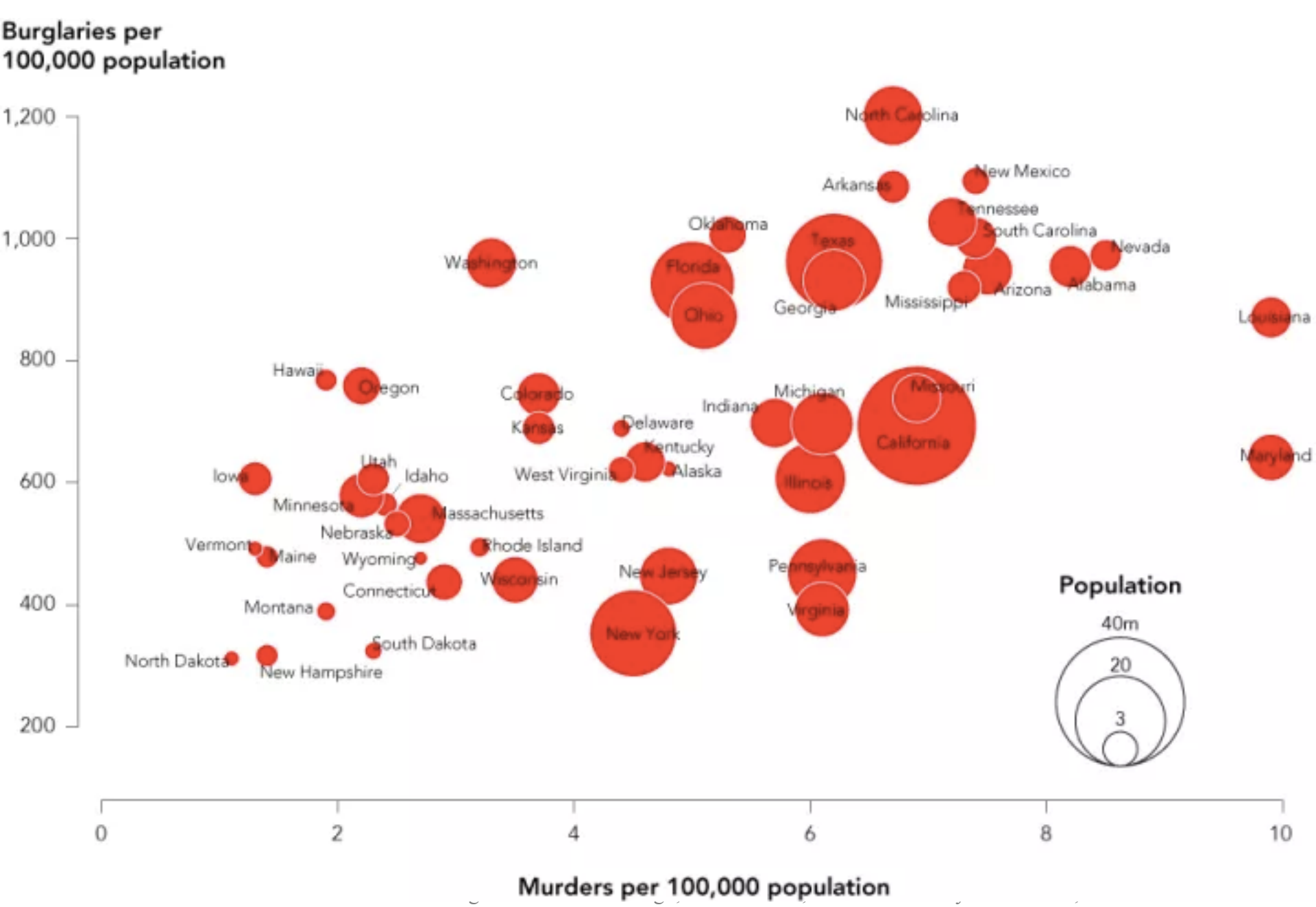
- Special/Structured
  - Time (1998-01-27 12:00, 1999-04-13 22:10, …)
  - GPS (37.77679 latitude, -122.45117 longitude,…)
  - Social Security Numbers (555-55-5555,…)
- Unstructured/Semi-Structured
  - Free-Form Text (Twitter Feed, Screenplay,…)
  - Mixed Types

BUBBLE CHART

Encoding Data

- Horizontal position
  - Continuous data
- Vertical position
  - Continuous data
- Circle area
  - Numerical data
- Circle color
  - Numerical or categorical data

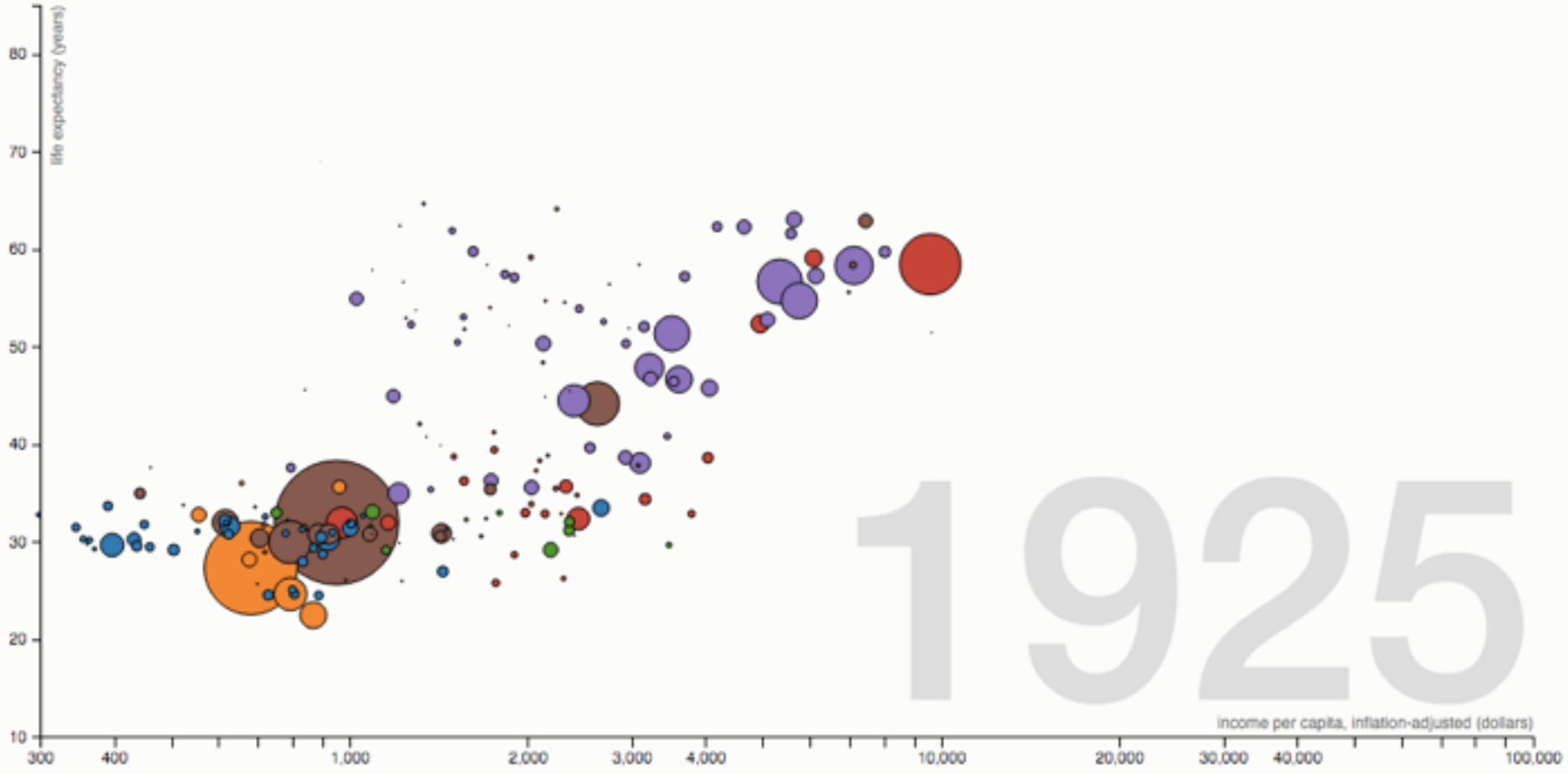
"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



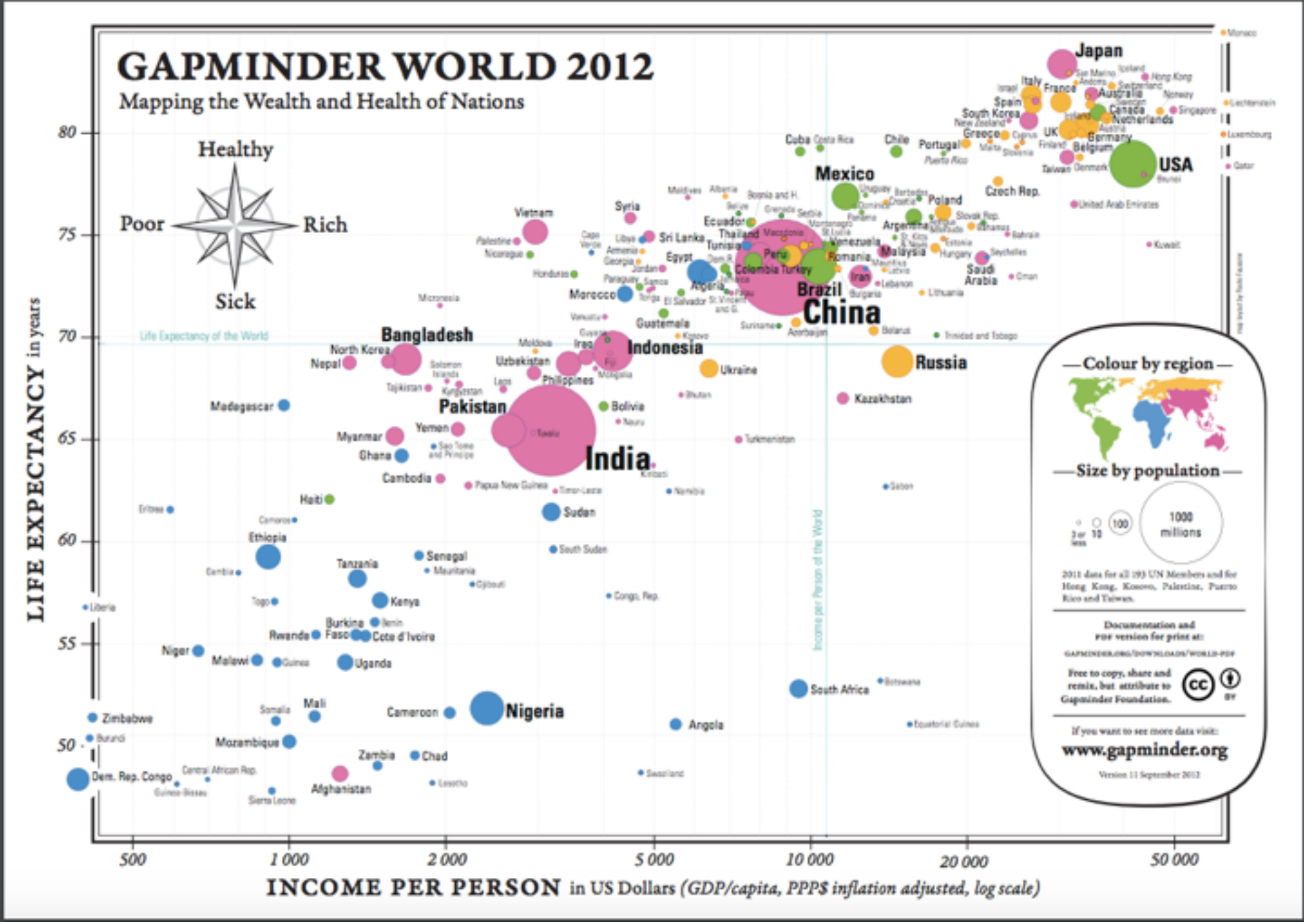
https://flowingdata.com/2010/11/23/how-to-make-bubble-charts/

March 13, 2012 / Mike Bostock

The Wealth & Health of Nations



http://bost.ocks.org/mike/nations/



http://www.gapminder.org/downloads/world-pdf/

Discussion

- 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 increase with density

HEATMAP

Encode Data

- Horizontal position
  - Column from dataset
- Vertical position
  - Row from dataset
- Box color
  - Value from row, column in dataset
  - Numerical or categorical

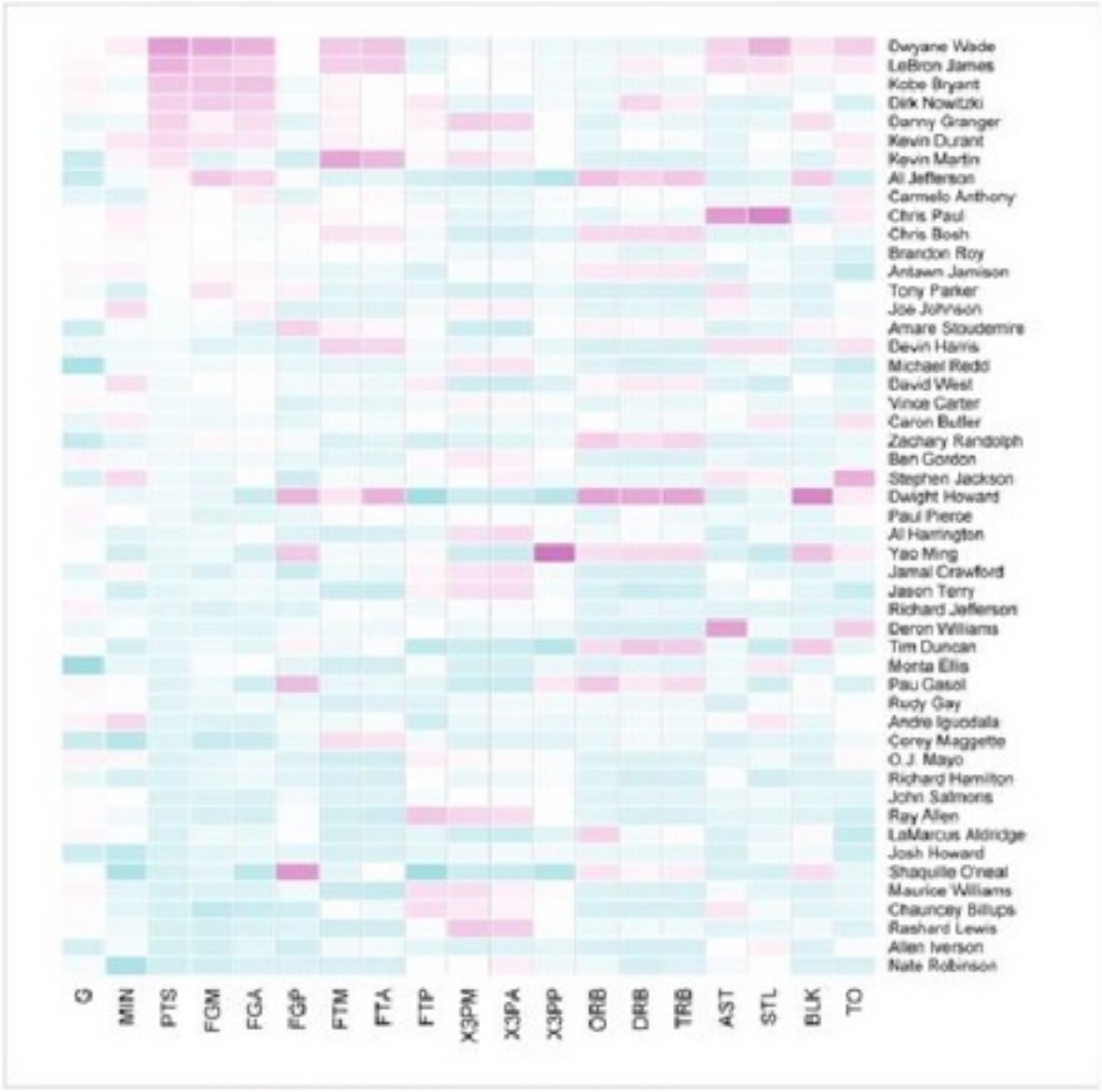
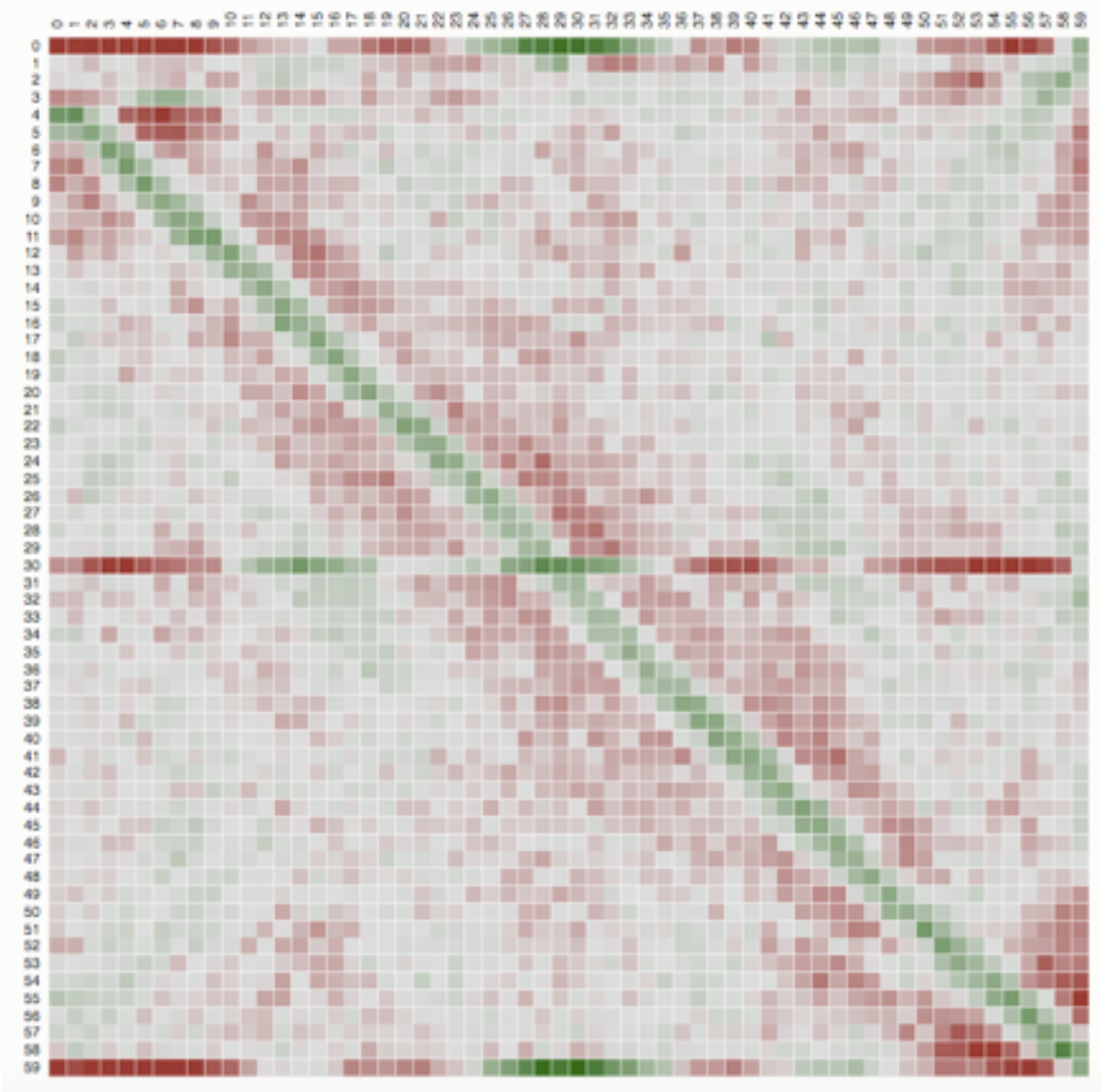


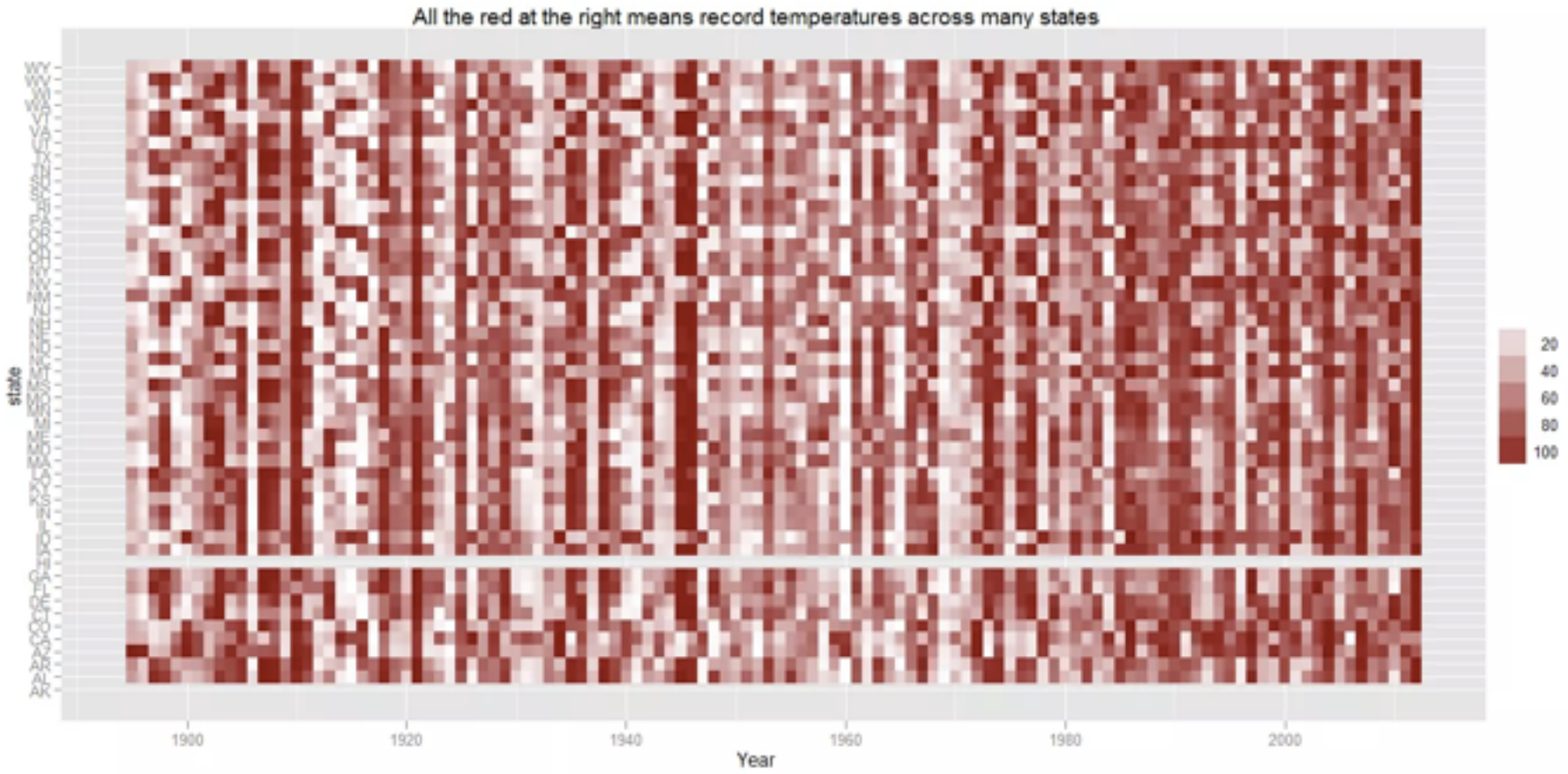
FIGURE 7-3 Default heatmap ordered by points per game

"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011

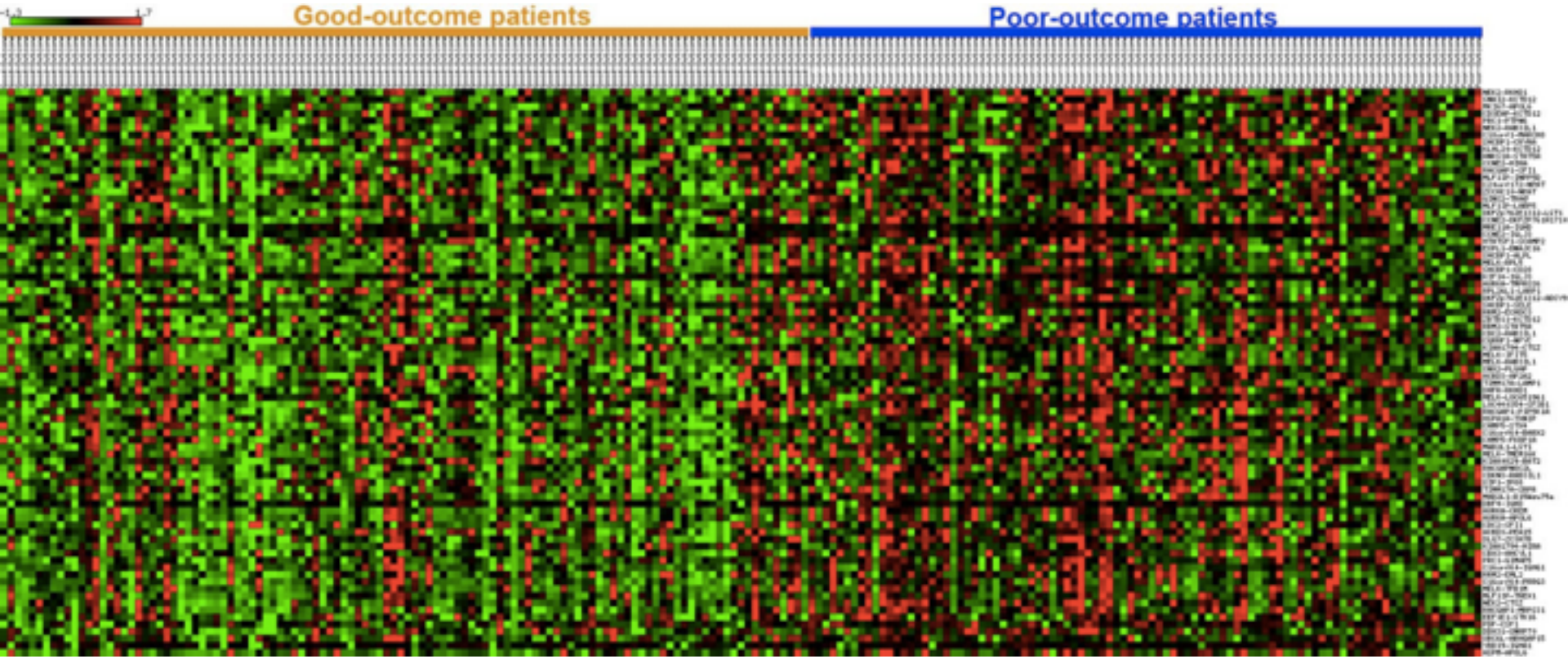
https://flowingdata.com/2010/01/21/how-to-make-a-heatmap-a-quick-and-easy-solution/



http://bost.ocks.org/mike/shuffle/compare.html



http://www.r-bloggers.com/118-years-of-us-state-weather-data/



http://www.biomedcentral.com/1471-2105/9/125

Discussion

- Essentially showing the dataset, but replaces numbers by color values
- Good for certain types of data
  - Continuous data well-suited
  - Unordered categorical limited to 7 categories
- Encourages comparison and pattern finding
  - Sorting changes patterns!

SCATTERPLOT MATRIX

Encode Data

- Scatter plot
  - Horizontal position maps to one variable
  - Vertical position maps to another variable
- Matrix of scatter plots
  - Each scatter plot focuses on one pair
  - Which pair is determined by row and column

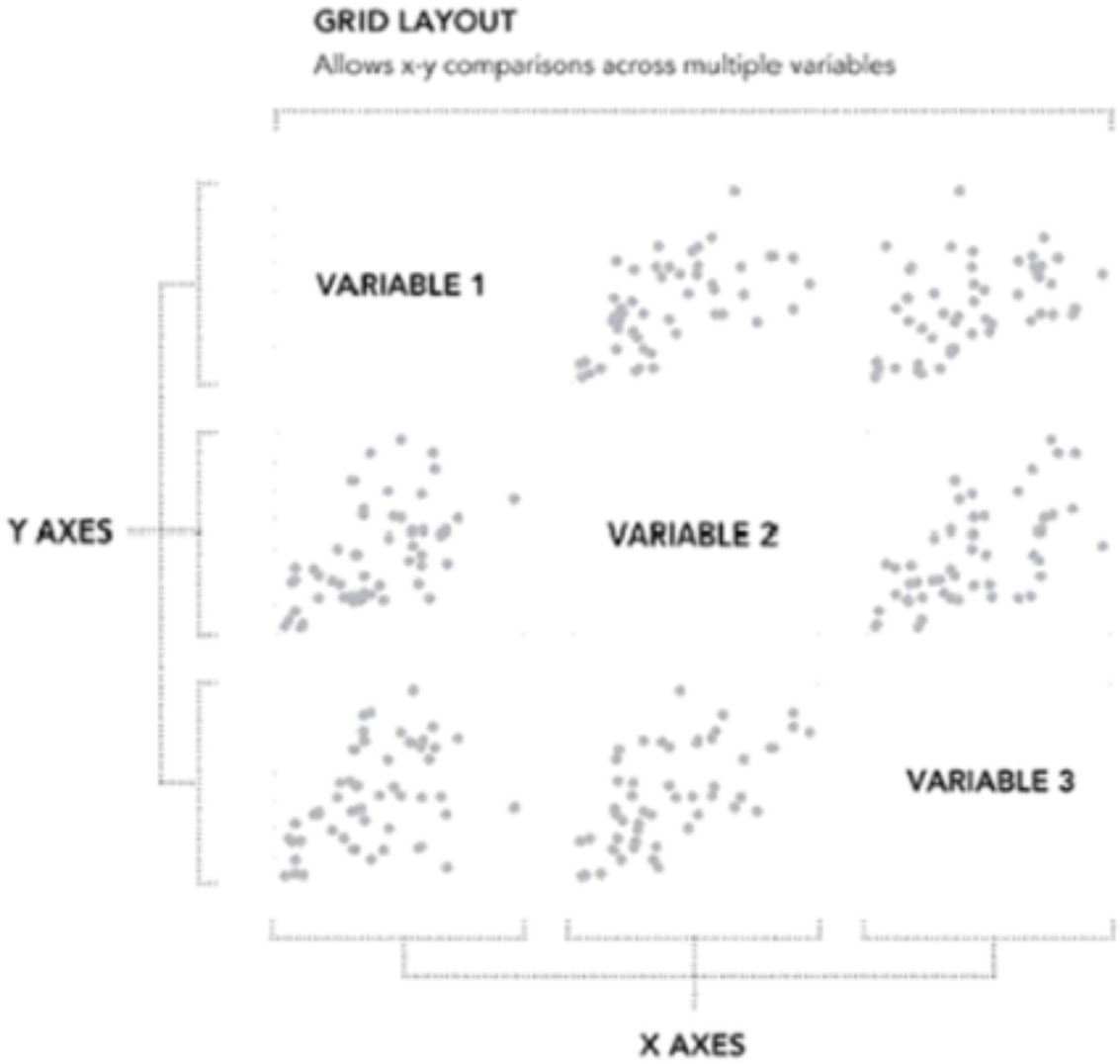


Figure 6-8: a scatter plot matrix
"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

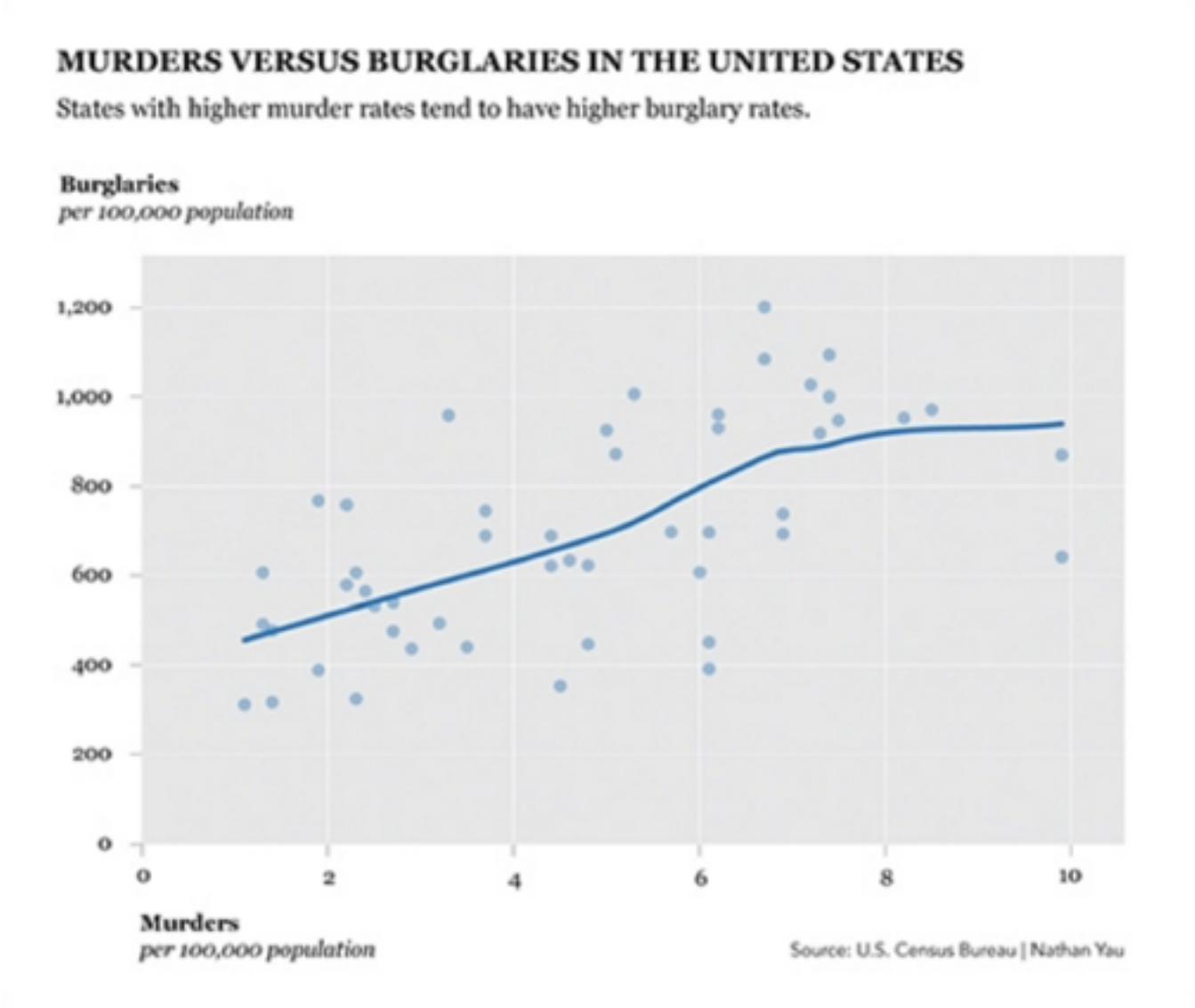
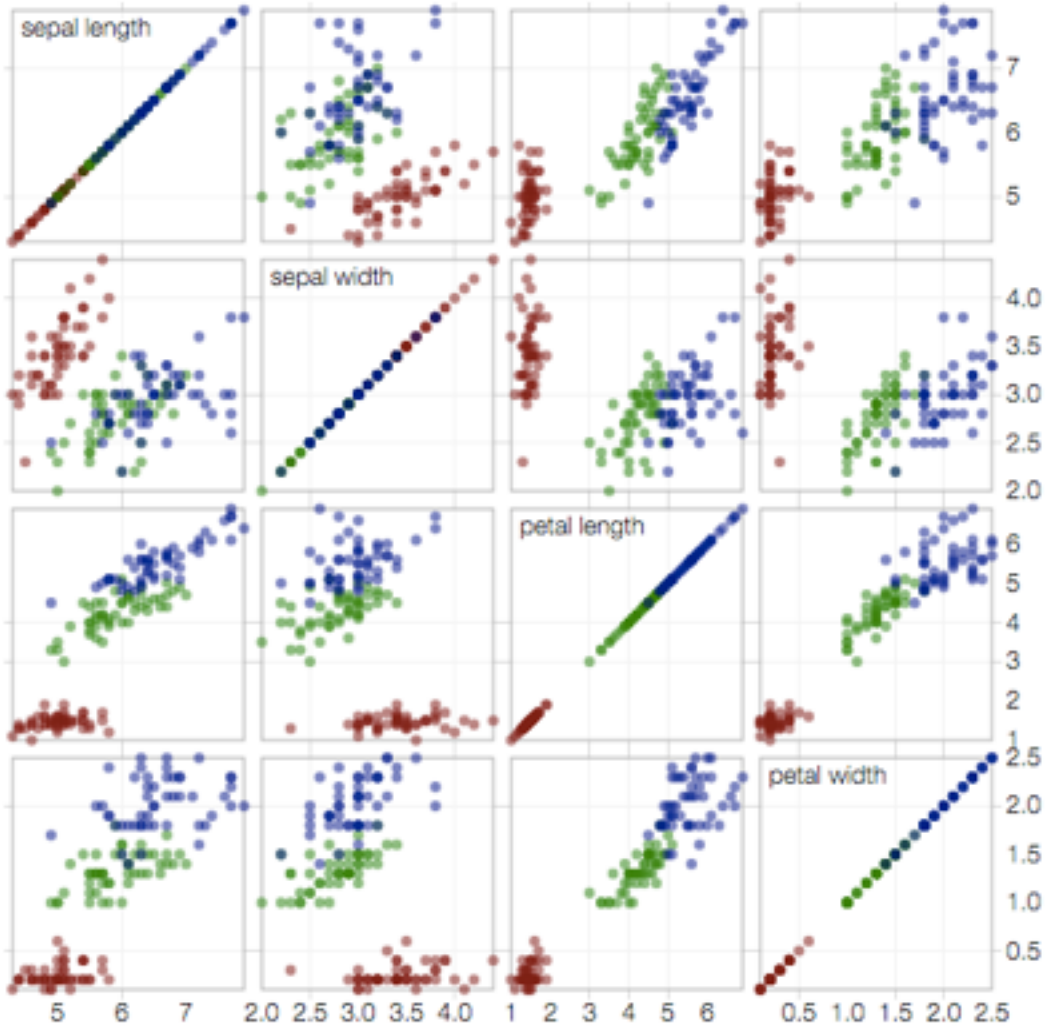


Figure 6-7
"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

FIGURE 6-9 Scatterplot matrix of crime rates

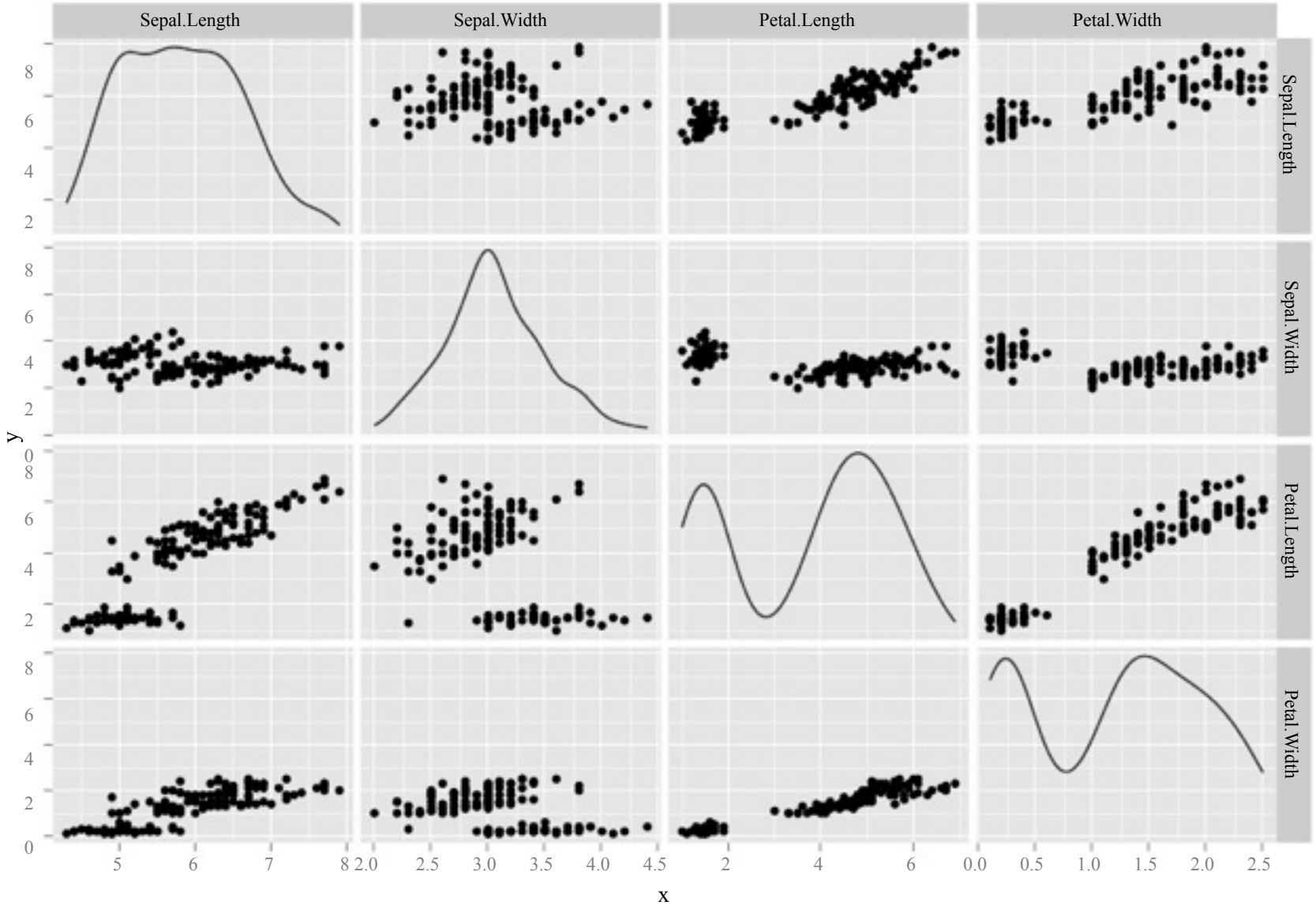
"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



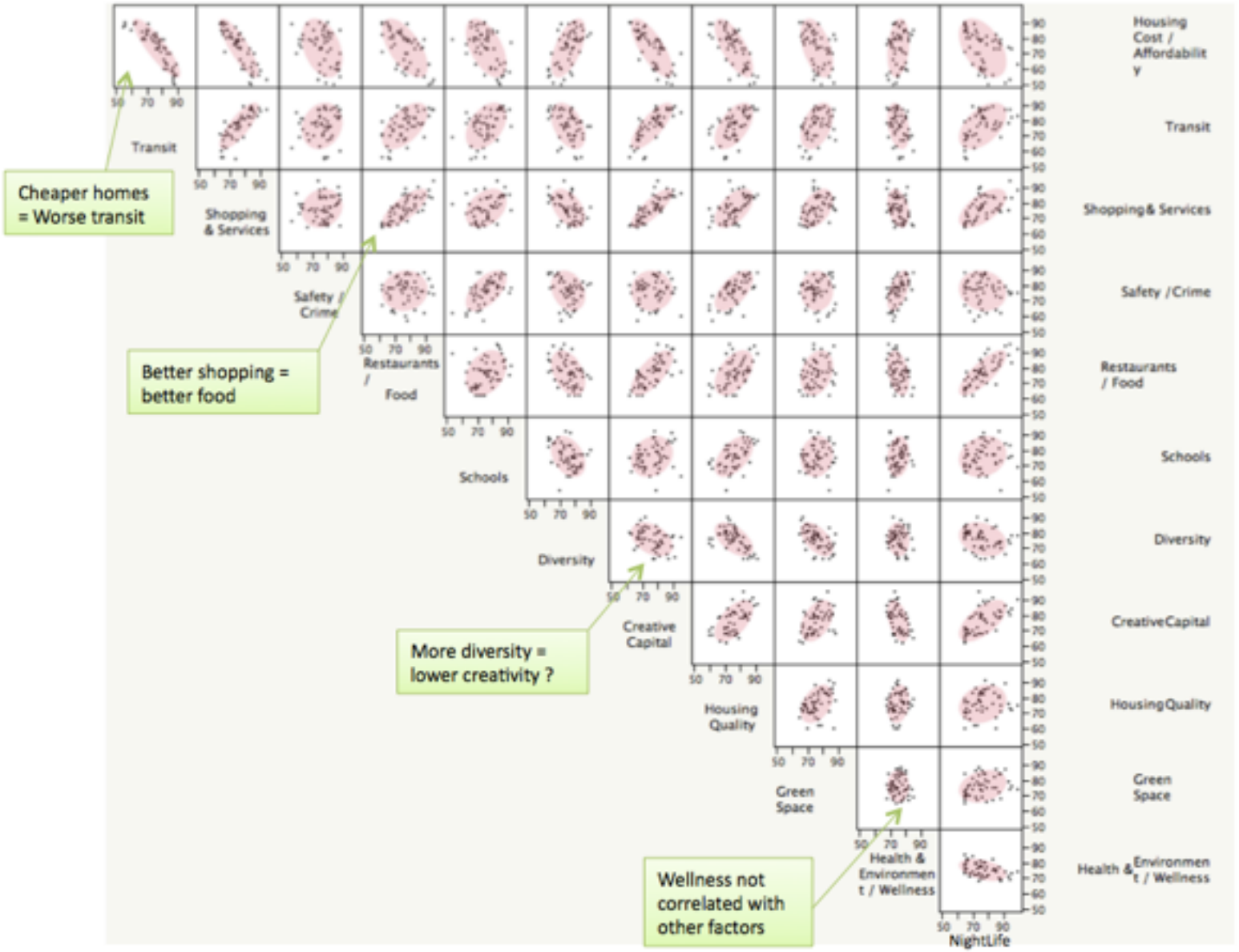
- *Iris setosa*
- *Iris versicolor*
- *Iris virginica*

Edgar Anderson's *Iris* data set
scatterplot matrix

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



http://gettinggeneticsdone.blogspot.com/2011/07/scatterplot-matrices-in-r.html



http://junkcharts.typepad.com/junk_charts/2010/06/the-scatterplot-matrix-a-great-tool.html

Discussion

- Good for exploration and comparison
  - Can be a little overwhelming at first
- Works with numerical or ordered data
- Form of small-multiples plot
  - Multiple scatter plots

SMALL MULTIPLES

Encode Data

- Group data by a variable to divide it into subsets
- Create a small plot for each subset
- Show all subset plots on same page

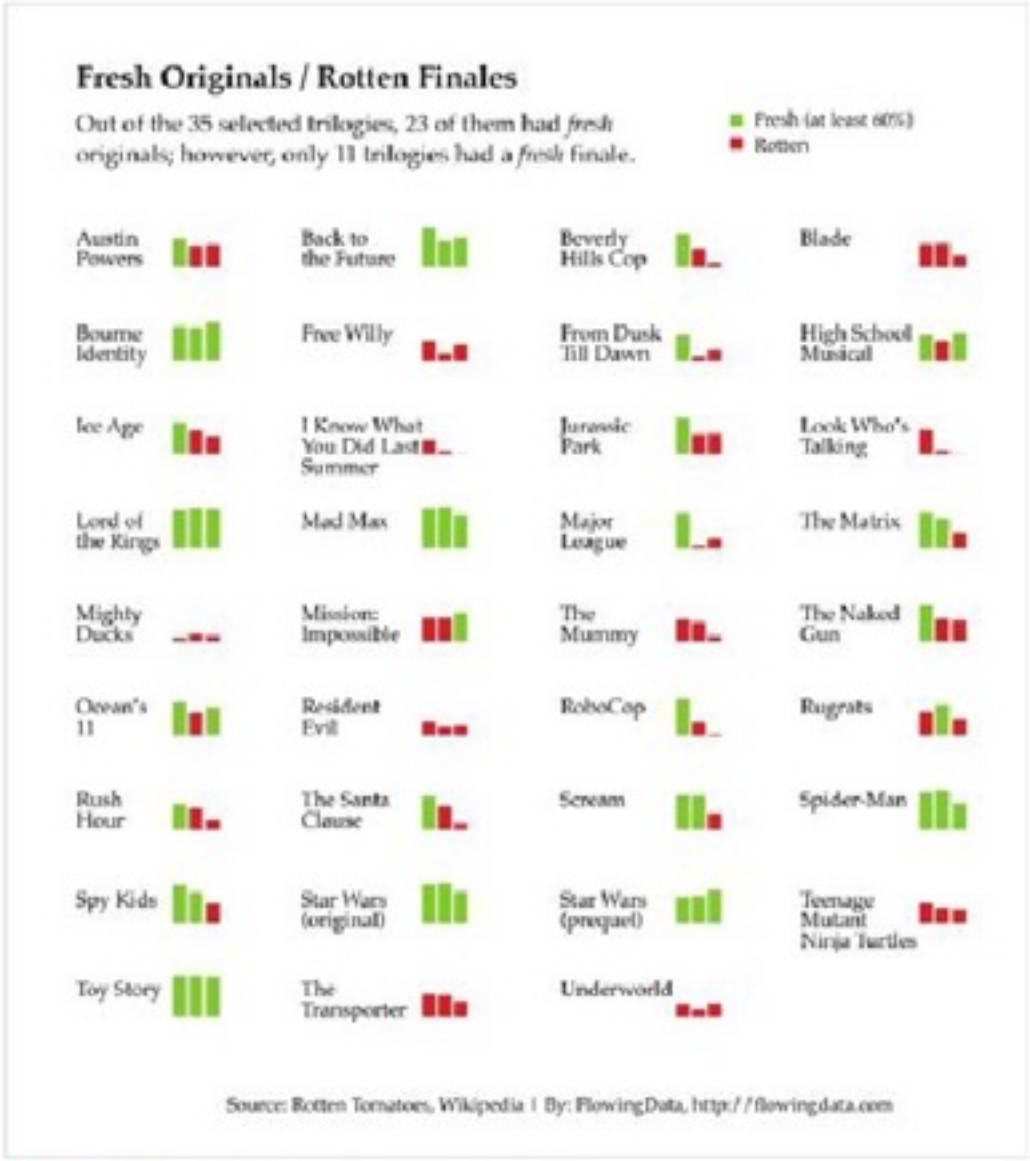
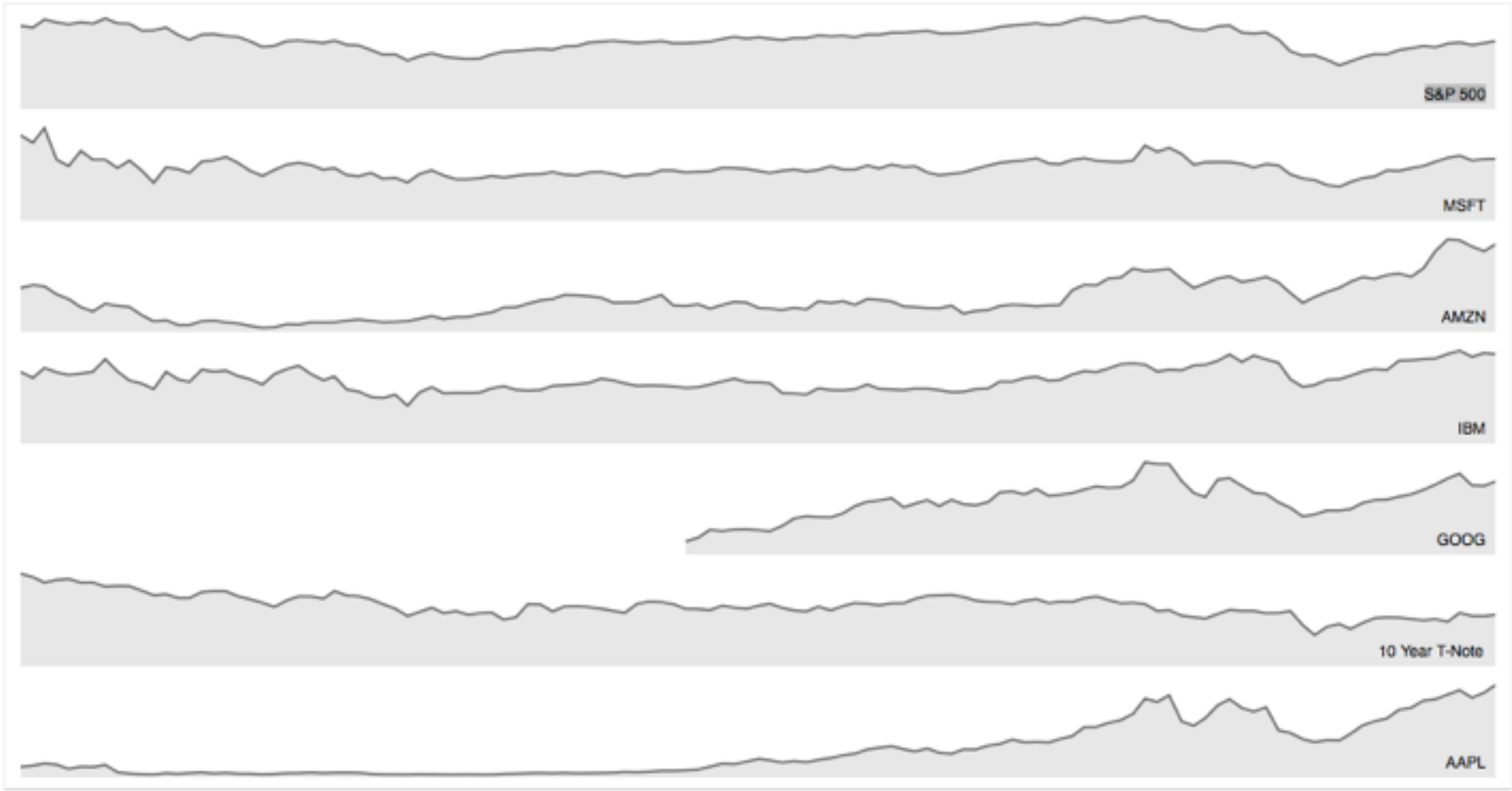
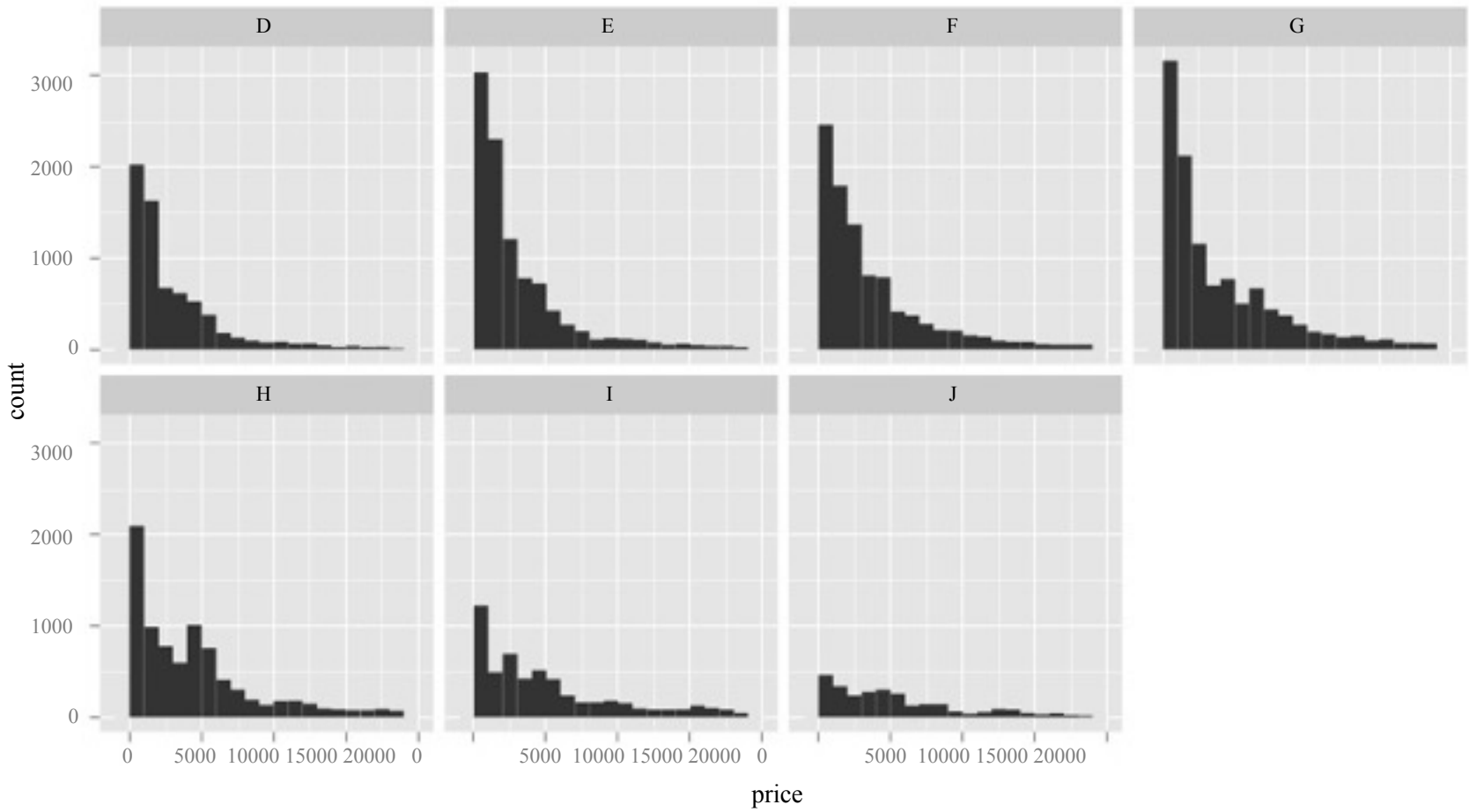


FIGURE 6-40 Small multiples for ratings of trilogies

"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



http://bl.ocks.org/1157787



http://docs.ggplot2.org/current/facet_wrap.html

Discussion

- Excellent for comparison
  - Depends on how to place the multiples!
- Requires a variable to use for grouping
  - Discrete or categorical data
- Can use with any type of plot
- Harder to tell exact values

PARALLEL COORDINATES

Encode Data

- Create one vertical line for every column
  - Numerical or ordered data
- Plot every row
  - x position is column
  - y position is value for that column
  - line connects values for a single row
- Picture an x y scatter plot, but putting both axis lines vertically

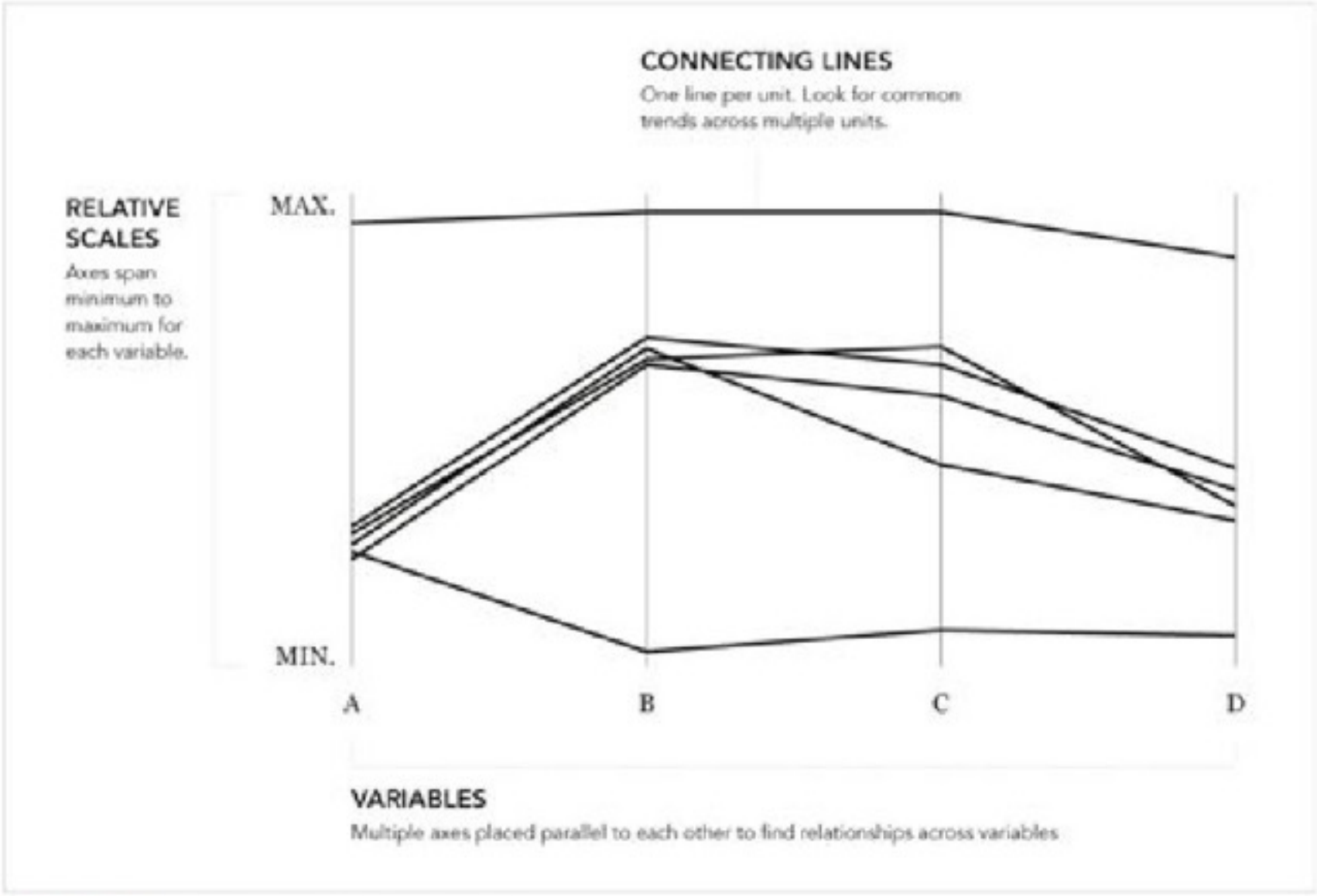


FIGURE 7-20 Parallel coordinates framework

"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011

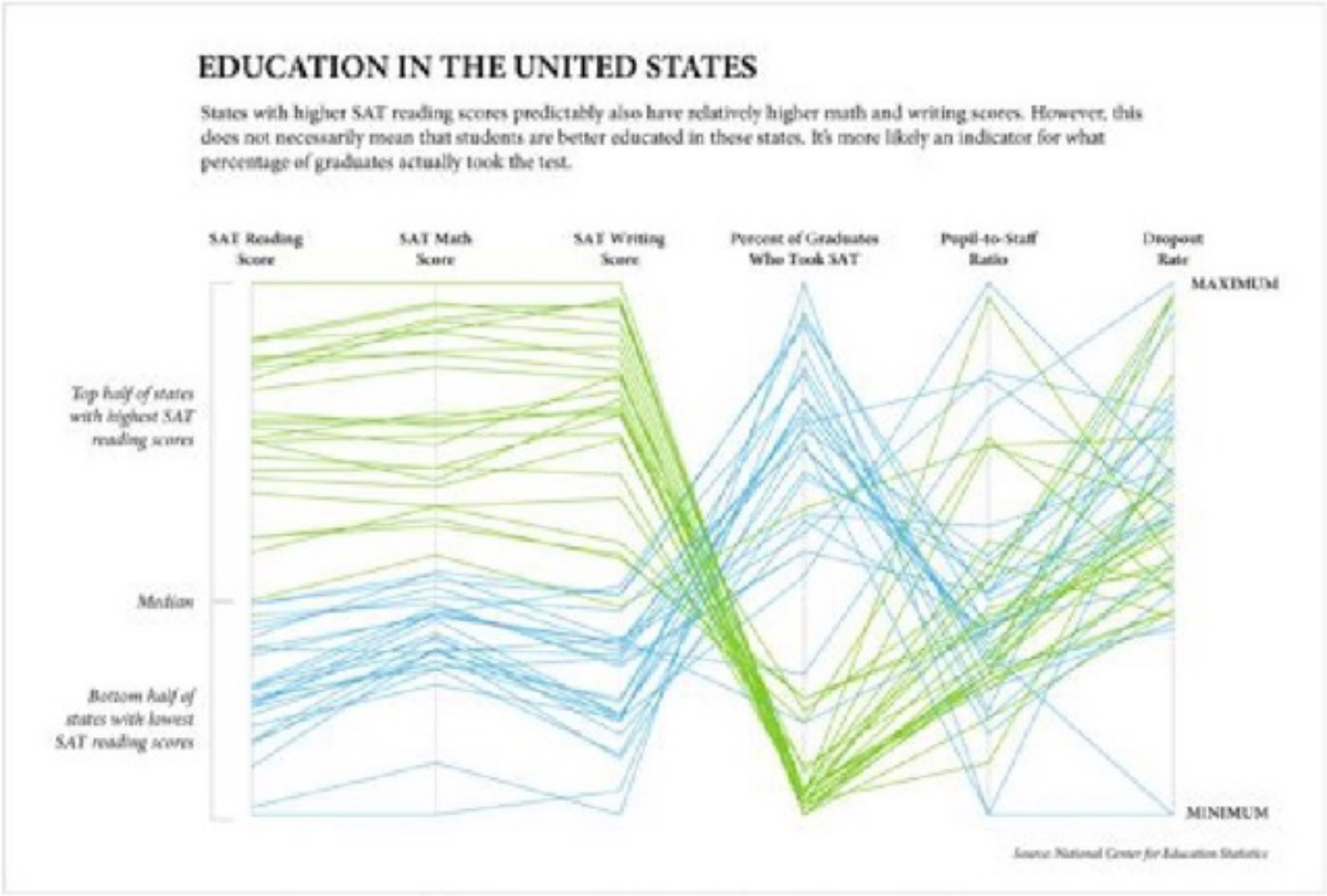
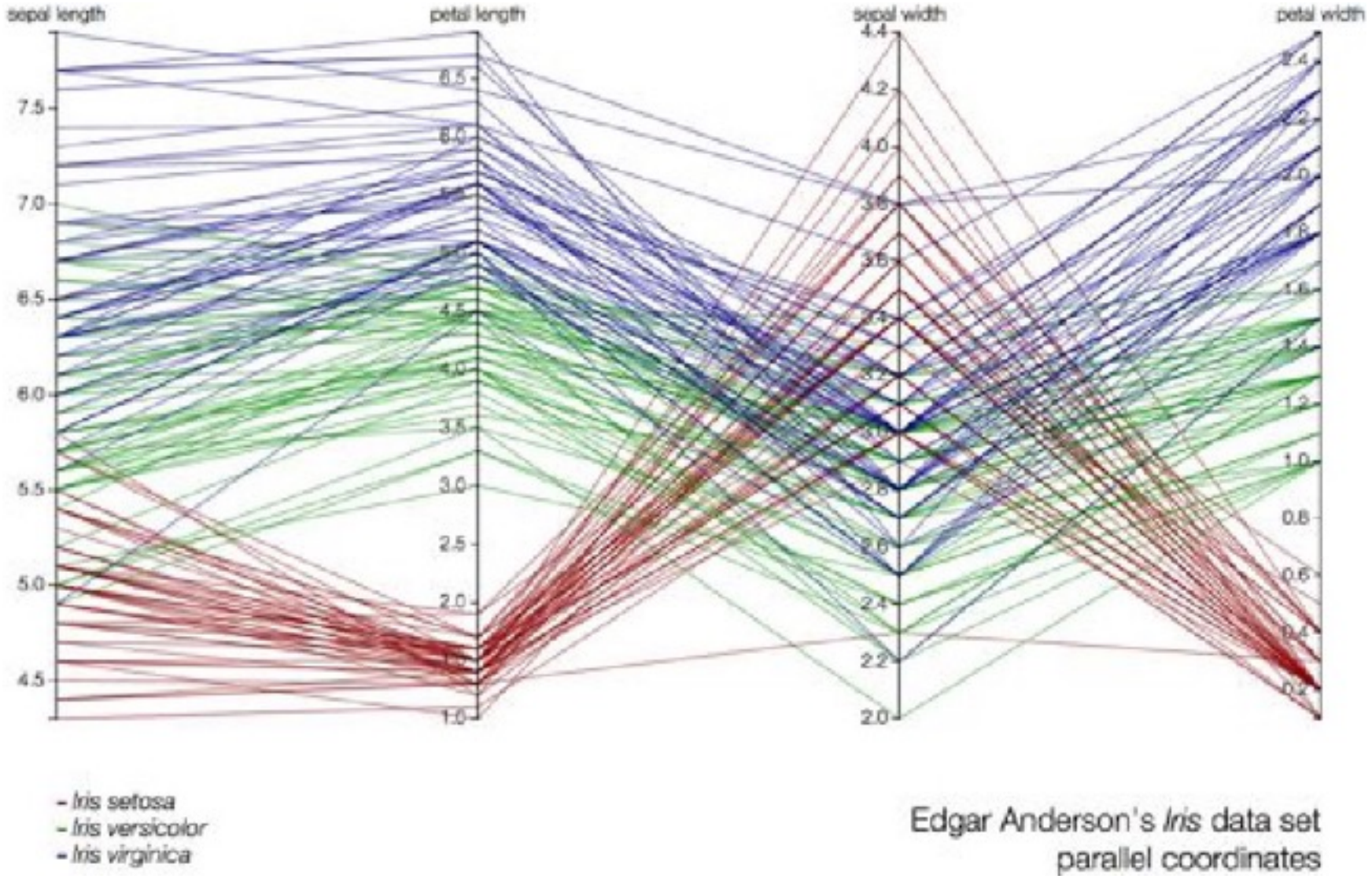
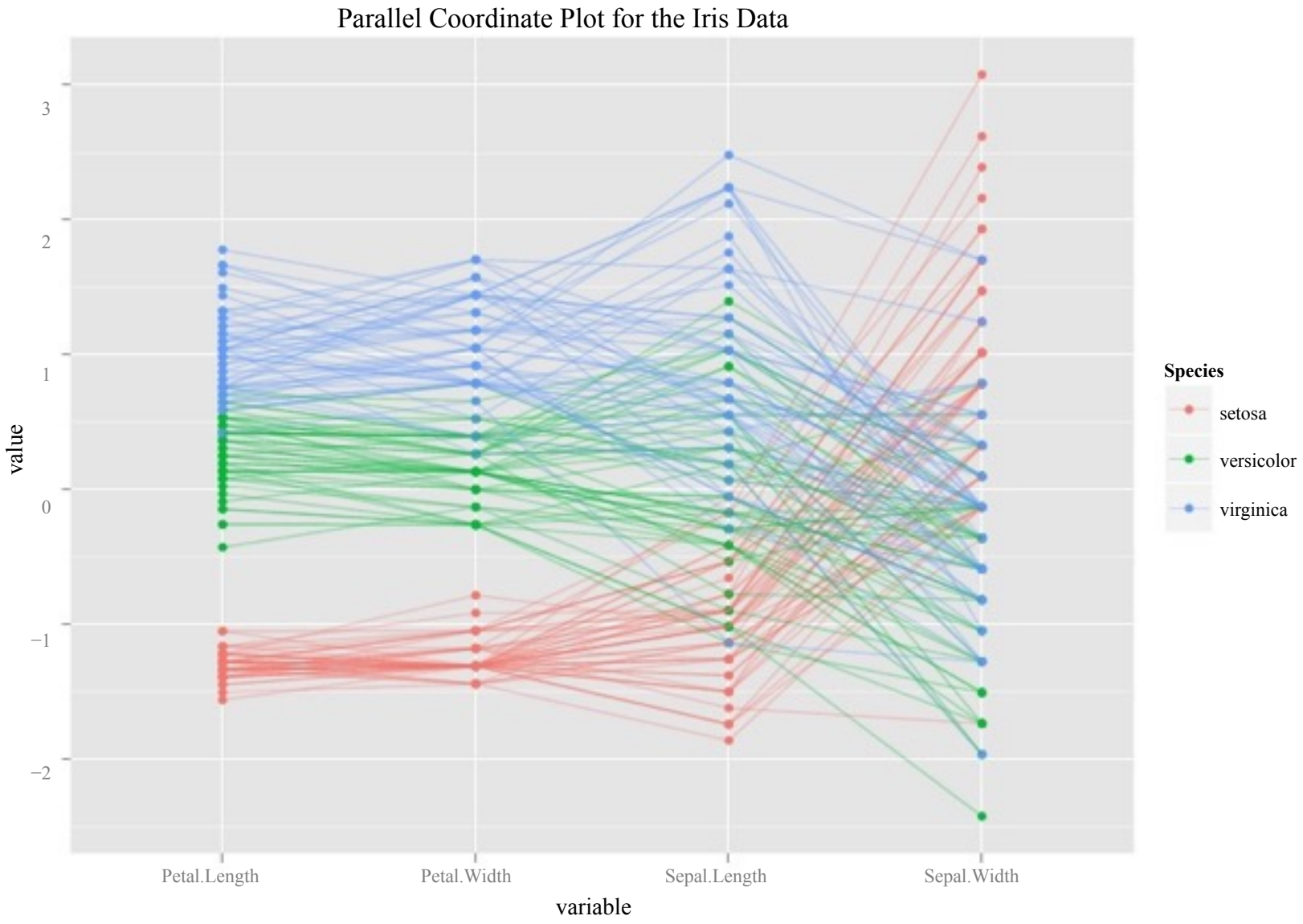


FIGURE 7-26 Standalone parallel coordinates plot on SAT scores

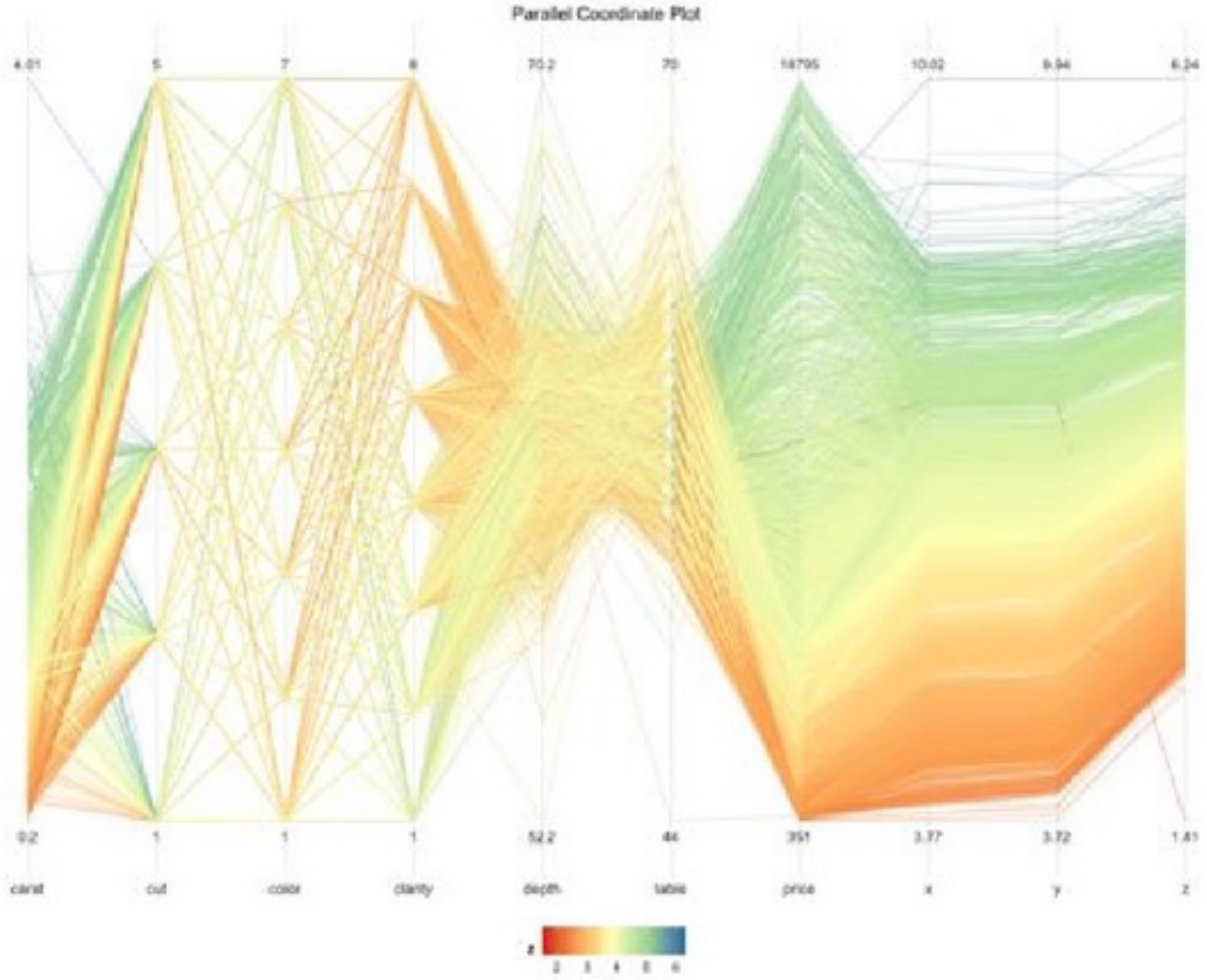
"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



http://mbostock.github.com/d3/talk/20111116/iris-parallel.html



http://www.inside-r.org/packages/cran/GGally/docs/ggparcoord



Custom Rand ggplot2 Implementation

Discussion

- Long startup time
- Almost always requires interactivity
  - Choose column to color by
  - Choose how to sort
  - Highlighting (brushing)
  - Clustering
- Very high density and data ink ratio, low lie factor

RADAR/STAR PLOT

Encode Data

- Same as a parallel coordinate plot, but drawn radially

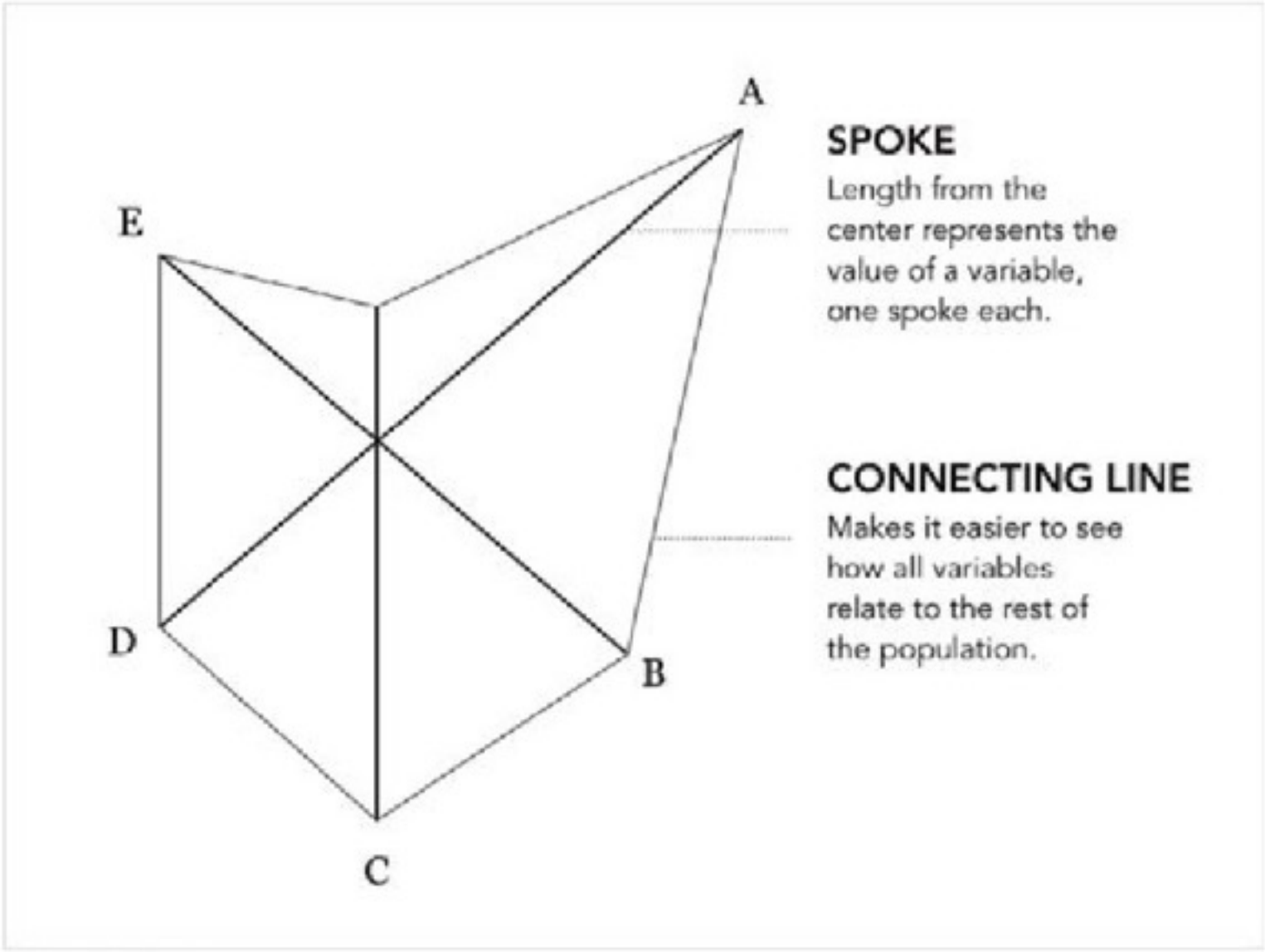


FIGURE 7-14 Star chart framework

"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

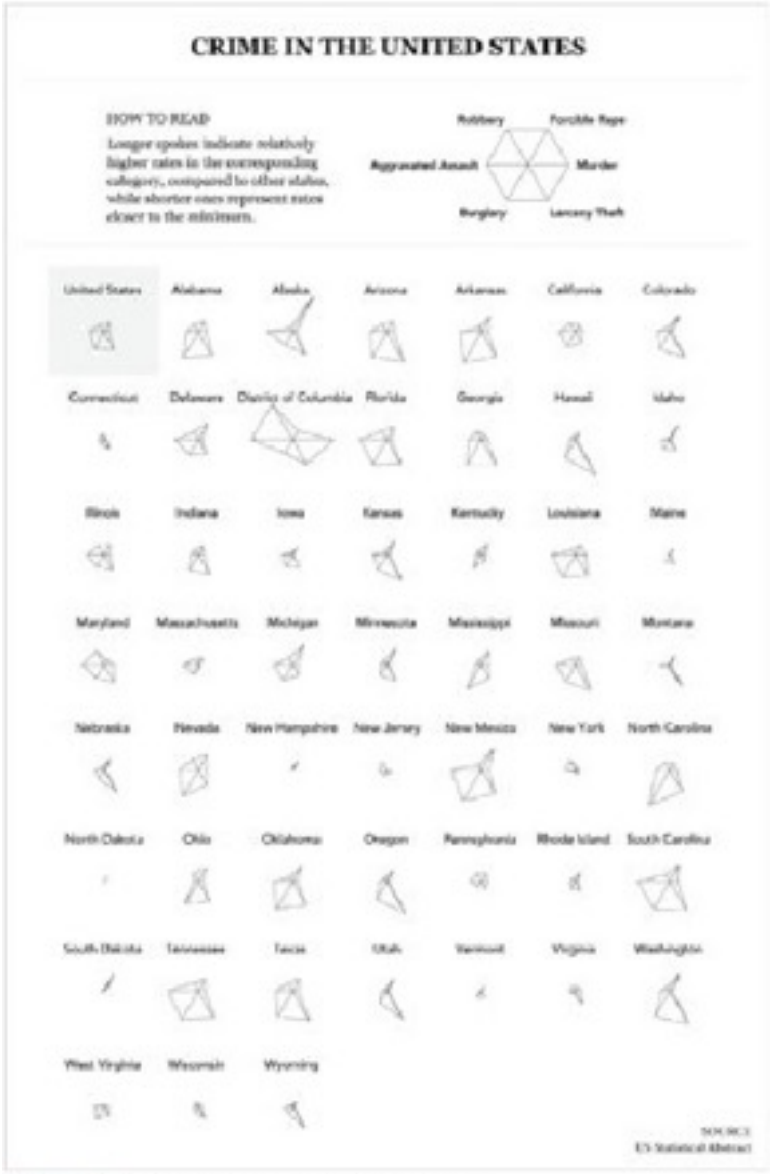
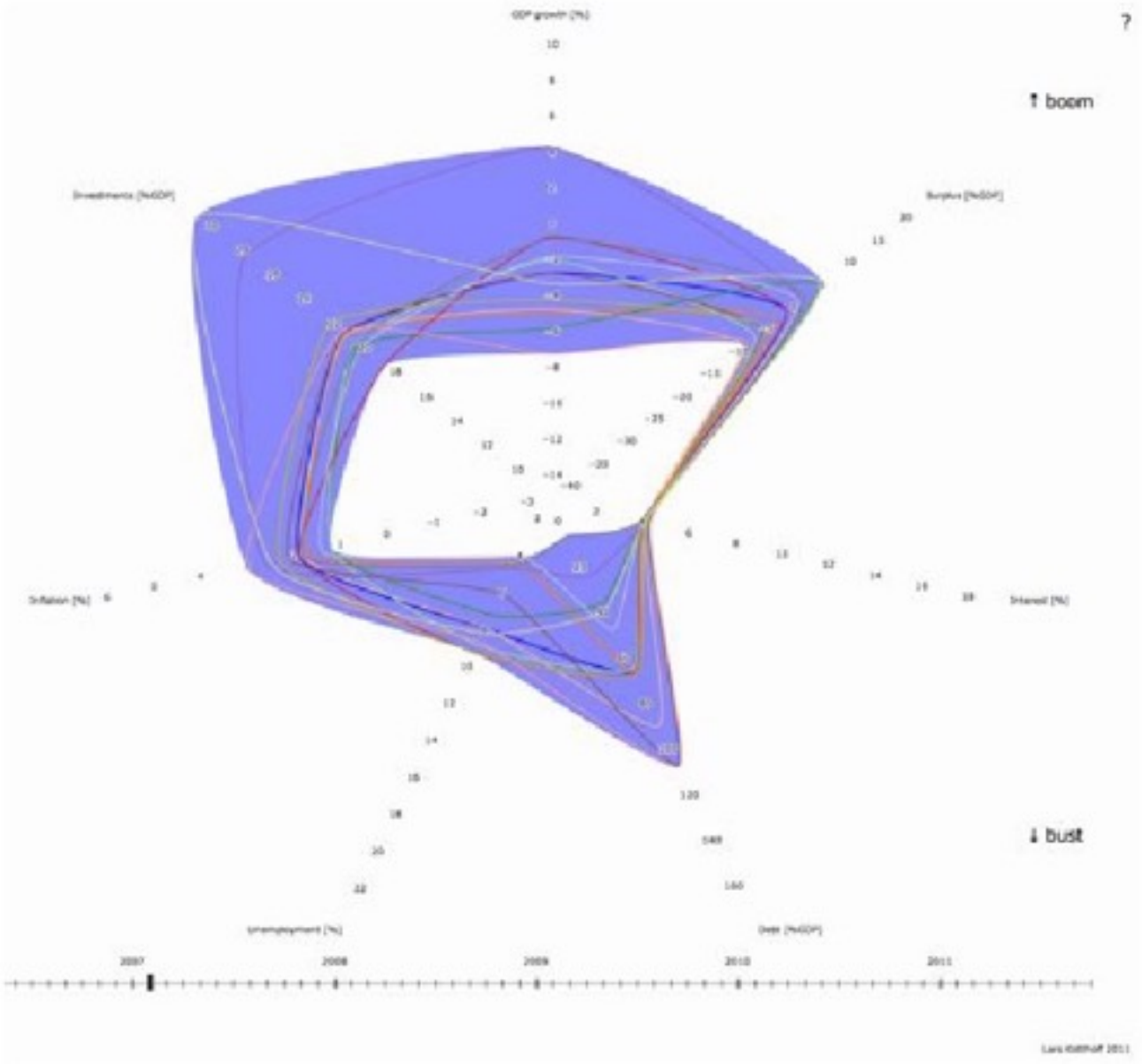
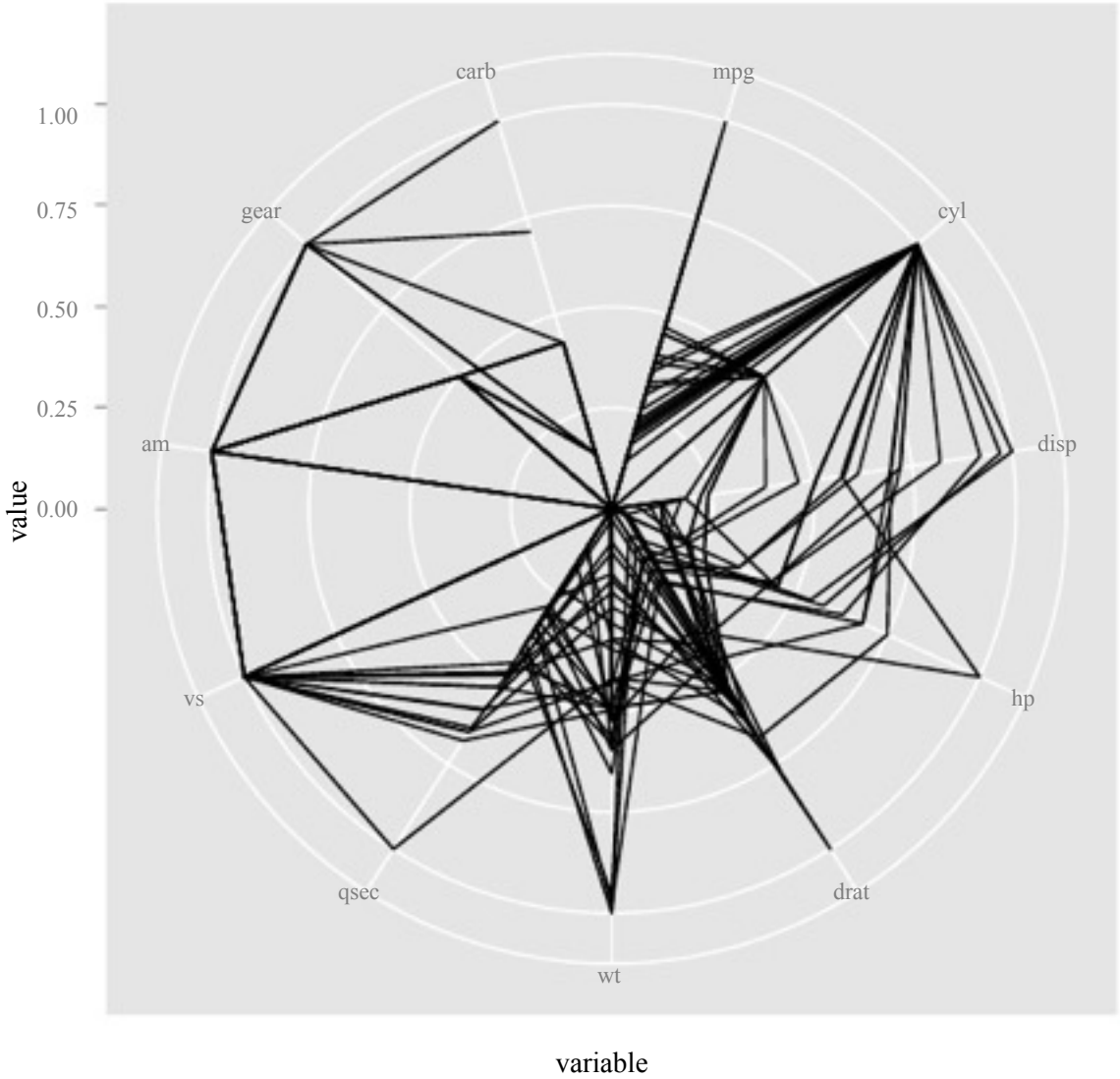


FIGURE 7-19 Series of star charts showing crime by state

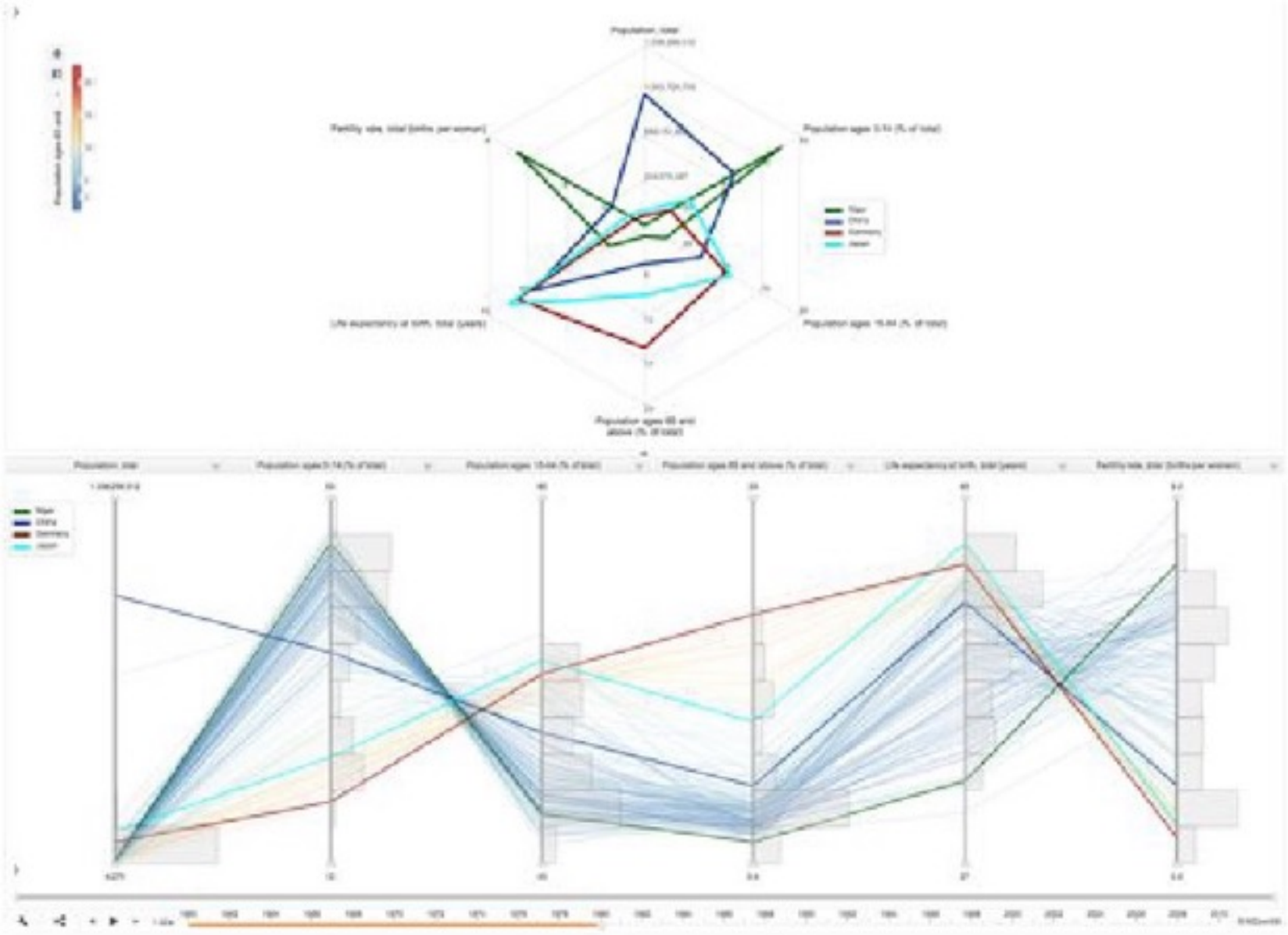
"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



http://www.larsko.org/v/euc/



https://github.com/hadley/ggplot2/issues/516



http://www.ncomva.se/html5/dynamic/index.html?layout=(radarchart,pcp)

Discussion

- Gives a shape to the data
- When plotted on top of each other, must have some ability to highlight and filter
- Works decently for small multiples?
  - Requires small number of rows
- Better for data with a circular aspect?
  - Monthly temperature time series

IRIS Dataset Five Ways

Multivariate data visualization case study

jupyter notebook: dv2_irisDataset

The Iris Dataset

The **Iris flower data set** or Fisher's Iris data set is a multivariate data set introduced by Ronald Fisher in his 1936 paper

It is sometimes called **Anderson's Iris data set** because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species

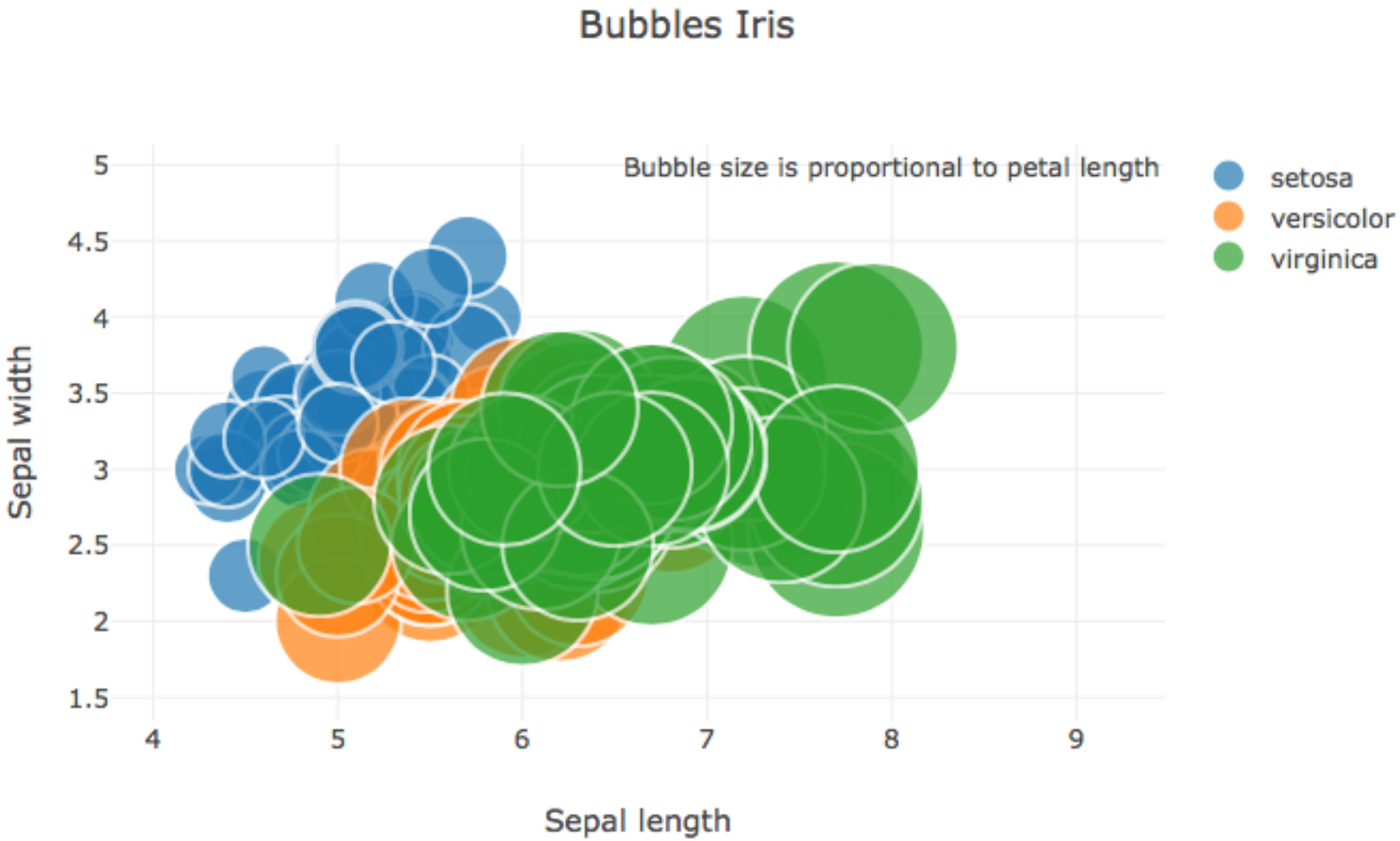
R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics 7 (2): 179–188

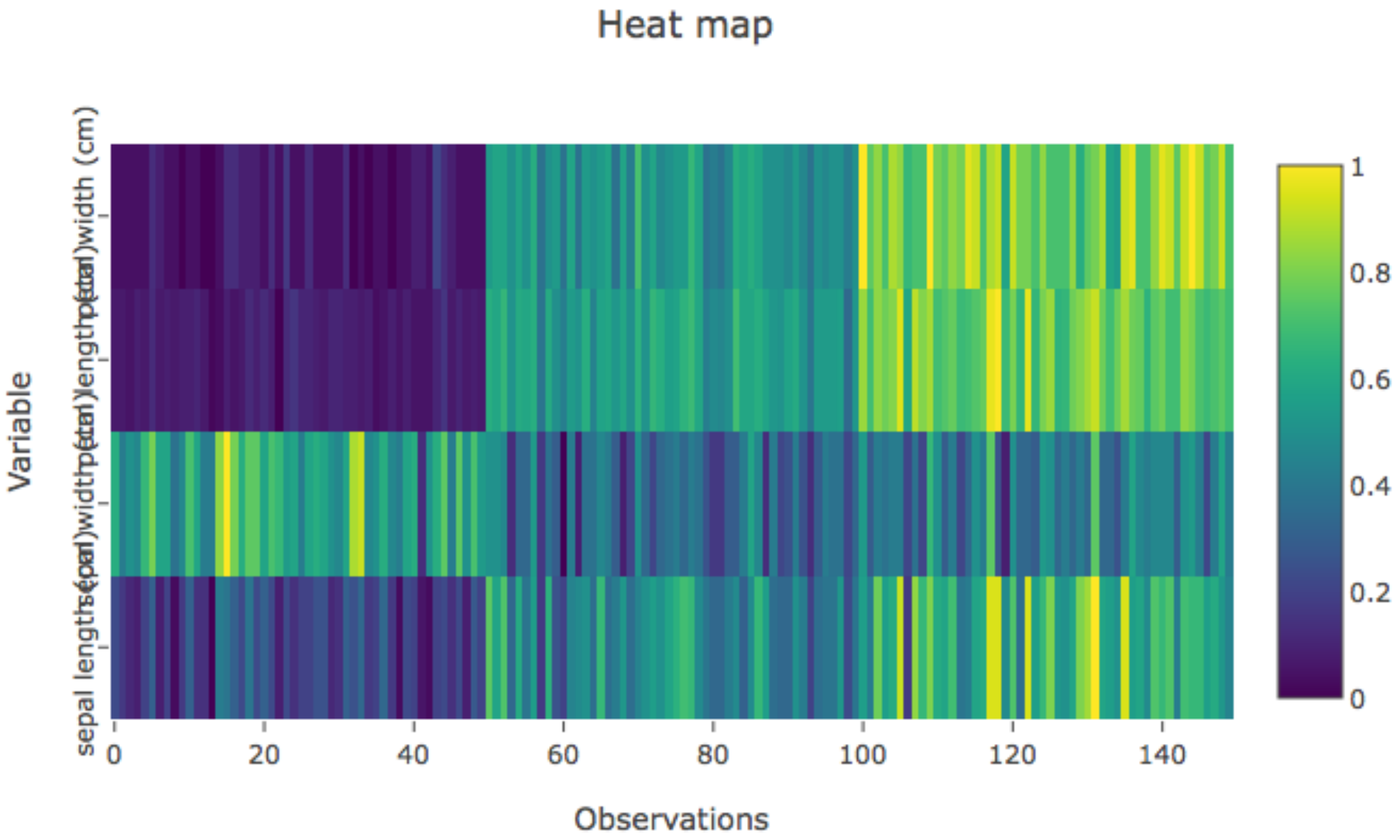
The Iris Dataset (2)

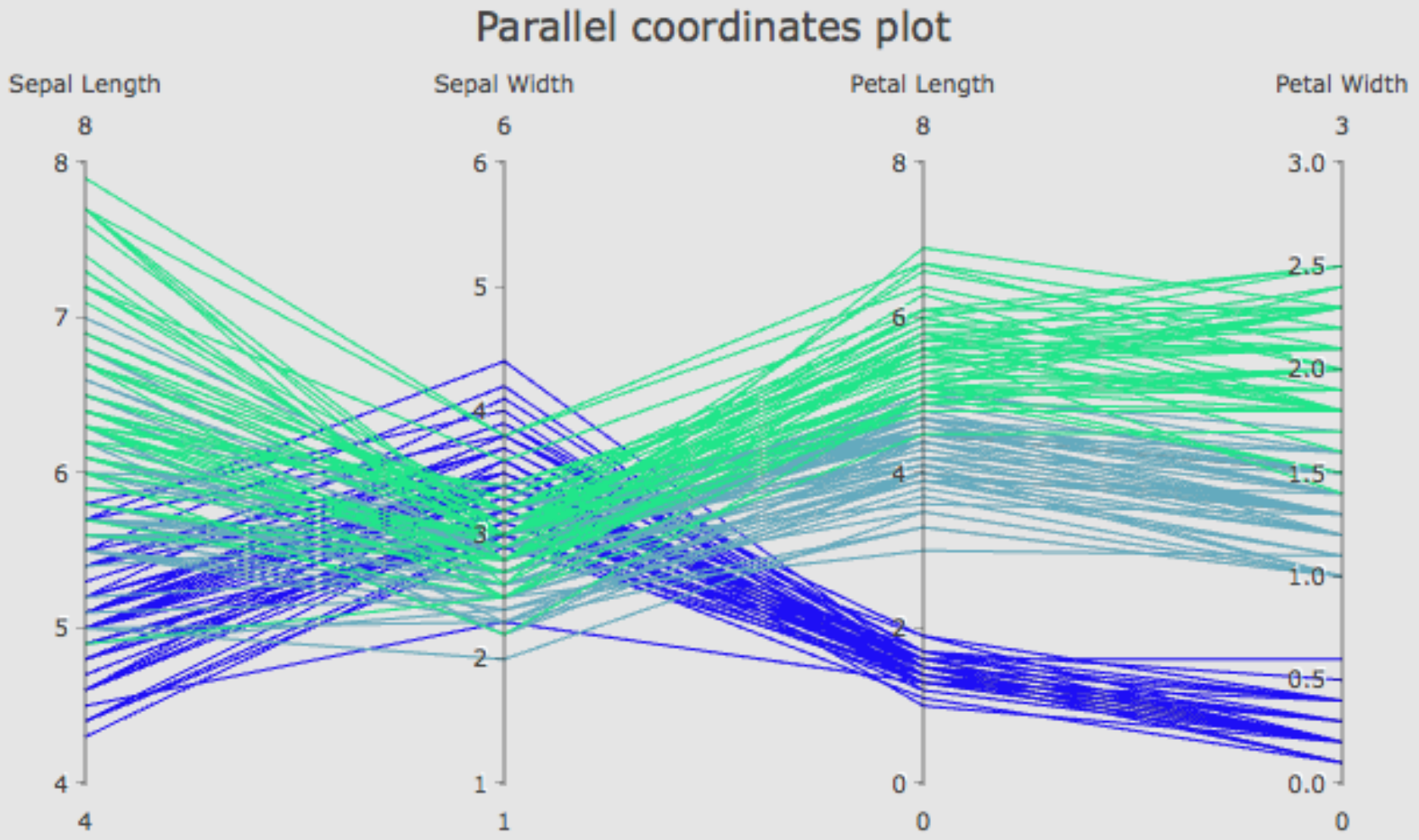
Two of the three species were collected in the Gaspé Peninsula "*all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus*"

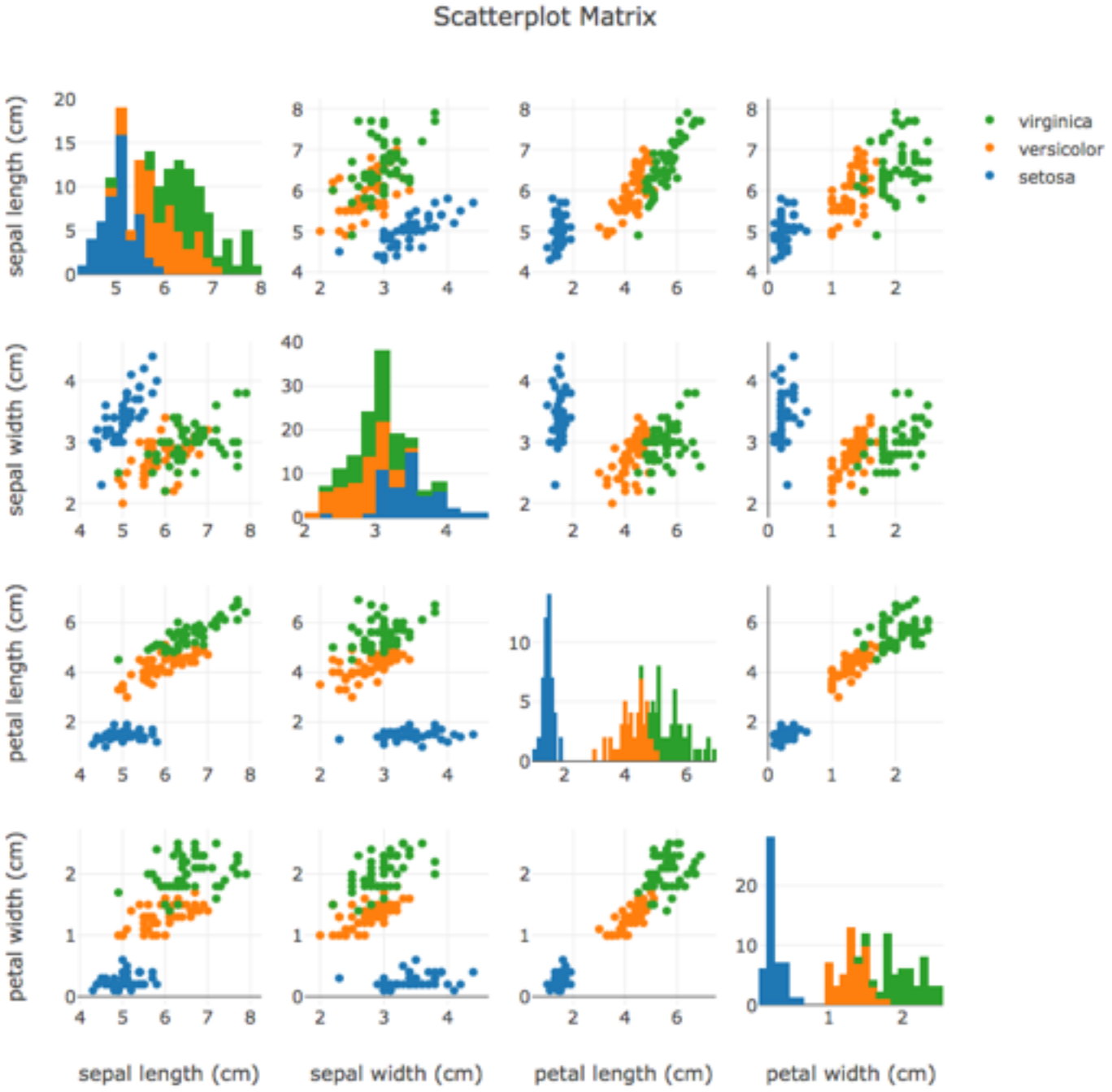
The data set consists of **50 samples** from each of three species of Iris (*Iris setosa, Iris virginica and Iris versicolor*). **Four features** were measured from each sample: *the length and the width of the sepals and petals*, in centimetres.

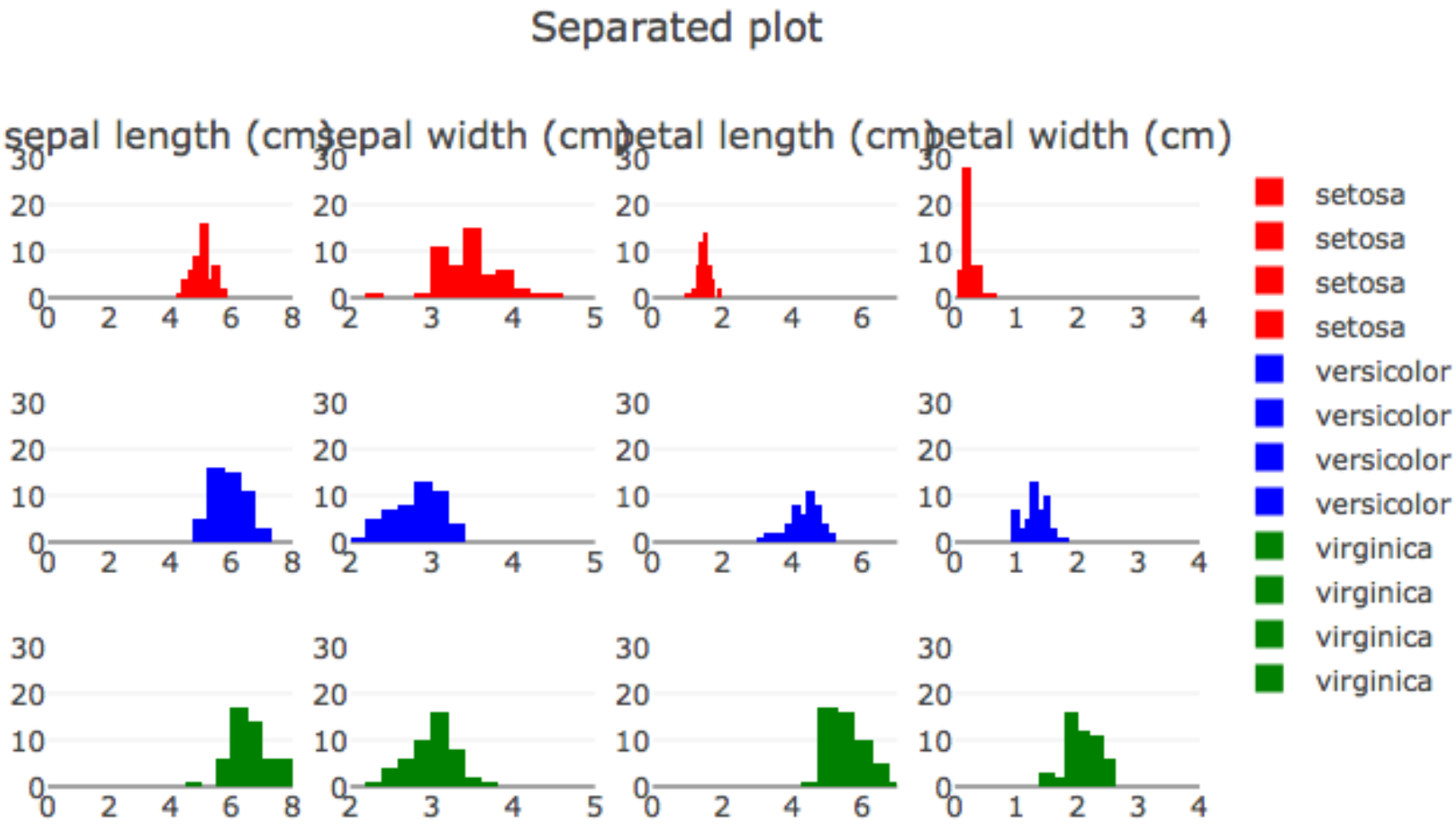
Based on the combination of these four features, Fisher developed a **linear discriminant model (LDA)** to distinguish the species from each other.


Bubbles Iris
Bubble size is proportional to petal length
setosa
versicolor
virginica
Sepal width
Sepal length
5
4.5
4
3.5
3
2.5
2
1.5
4
5
6
7
8
9


Heat map
Variable
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
0
20
40
60
80
100
120
140
Observations
1
0.8
0.6
0.4
0.2
0


Parallel coordinates plot
Sepal Length
Sepal Width
Petal Length
Petal Width
8
6
8
3
8
6
8
3.0
7
5
6
2.5
6
4
4
2.0
5
3
2
1.5
4
2
1.0
2
0.5
1
0
0.0
4
1
0
0


Scatterplot Matrix
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
virginica
versicolor
setosa


Separated plot
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
setosa
setosa
setosa
setosa
versicolor
versicolor
versicolor
versicolor
virginica
virginica
virginica
virginica

CONCLUSION

Other Techniques

- Parallel Sets
  - http://eagereyes.org/parallel-sets
- Rose Diagrams
  - http://dd.dynamicdiagrams.com/2008/01/nightingales-rose/

References

- Nathan Yau, Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics, Wiley Publishing, 2011.
- Mike Bostock, Data Driven Documents (D3.js)
  - https://github.com/mbostock/d3/wiki/Gallery
  - http://bl.ocks.org/mbostock
- Other references sourced on slides

Questions?

Thanks to
Sophie J. Engle
San Francisco University
for ideas, suggestions, slides, links, and much other stuff