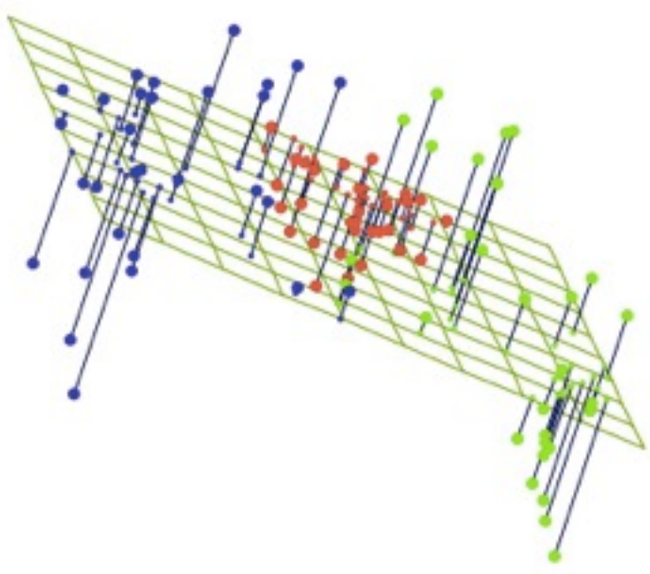
CS109 Data Science
High-Dimensional Data

Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



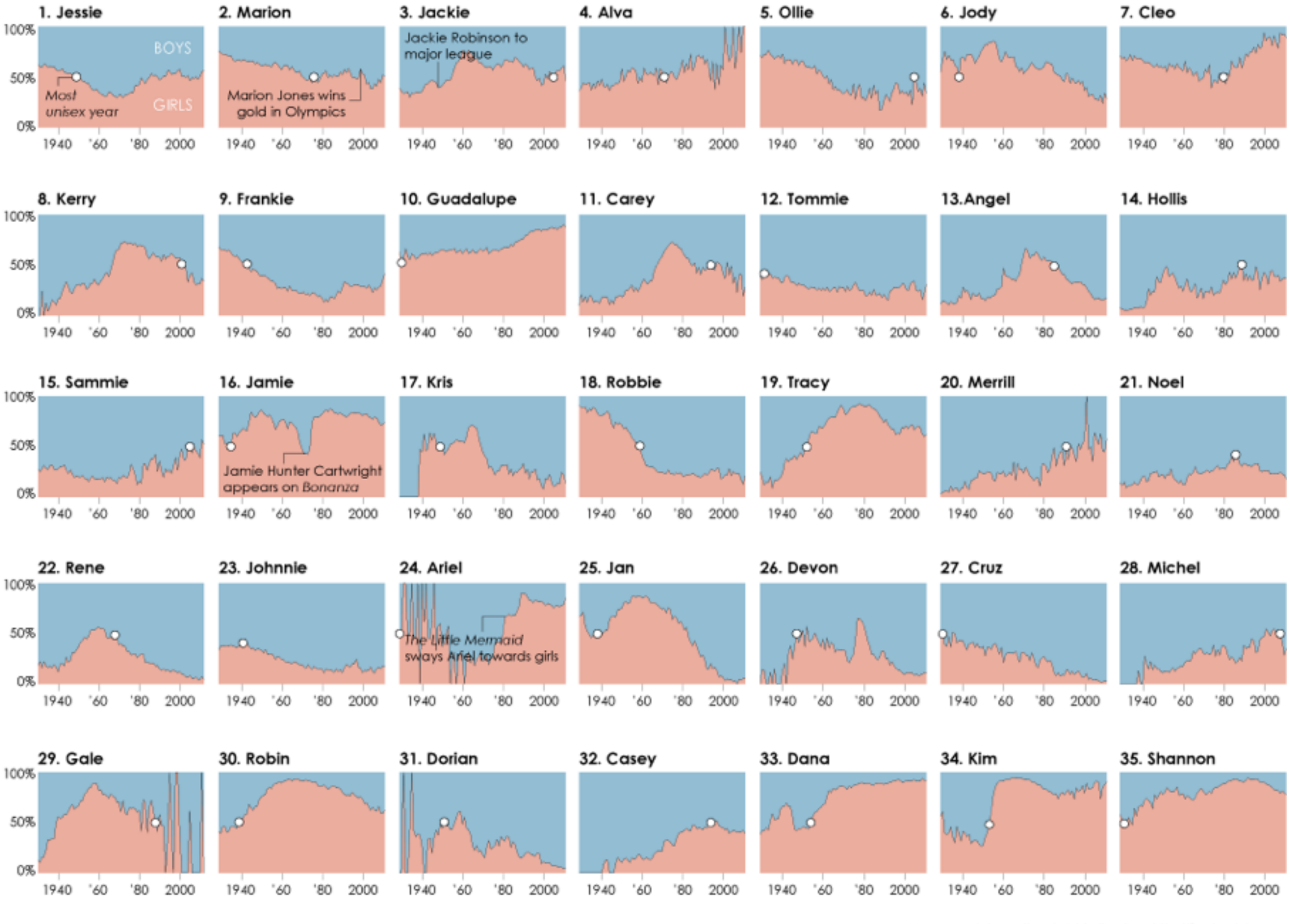
This Week

- HW1 solution (w/ screencast) on Piazza
- Fill in HW1 survey (bit.ly/feedback_hw1)
- HW2 - due Thursday, Oct 3, you should have started already!
- Friday lab **10-11:30 am** in MD G115
  - *Scikit-learn* with Rahul, Deqing, Johanna, and Ray
  - Linear regression, logistic regression, PCA

Vis of the Week

The most unisex names in US history

SEPTEMBER 25, 2013 I DATA UNDERLOAD



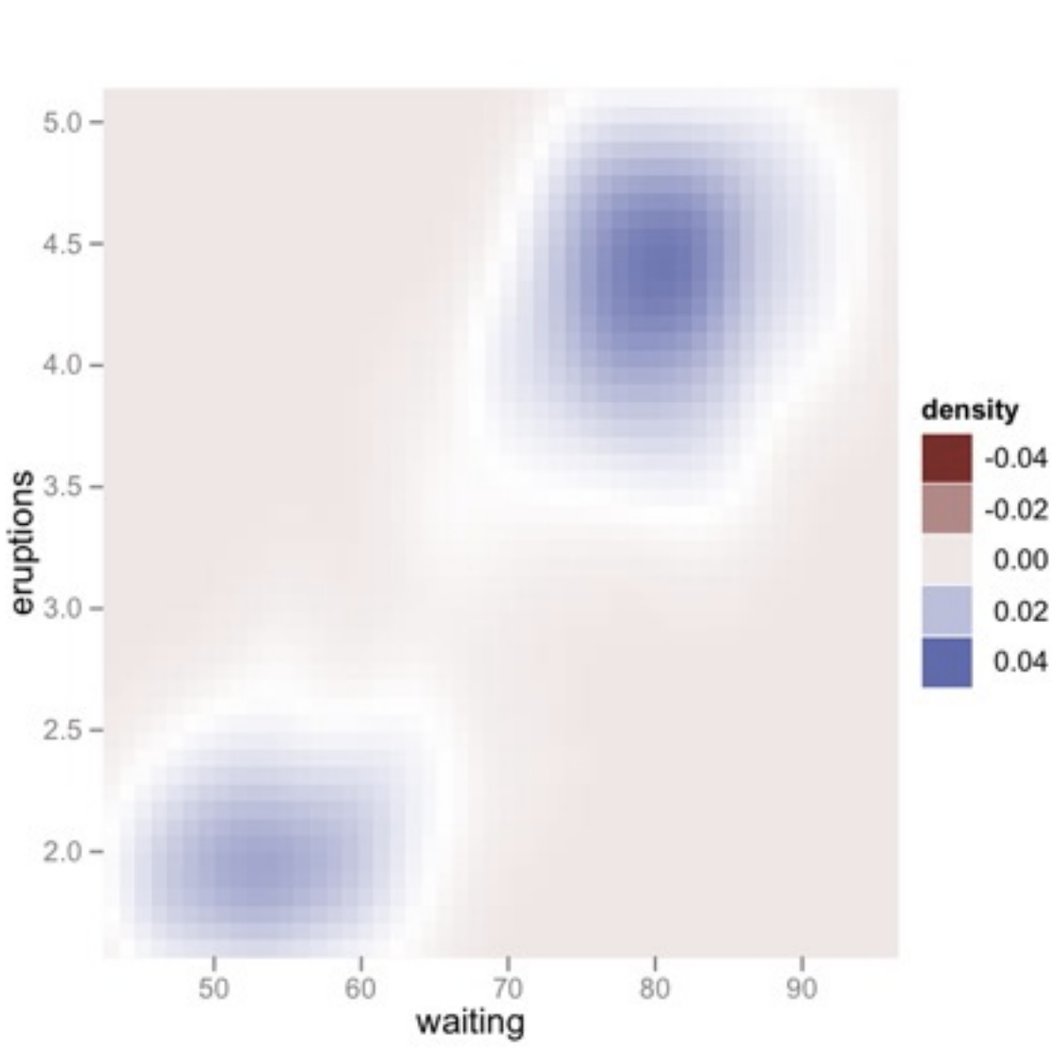
http://flowingdata.com/2013/09/25/the-most-unisex-names-in-us-history/

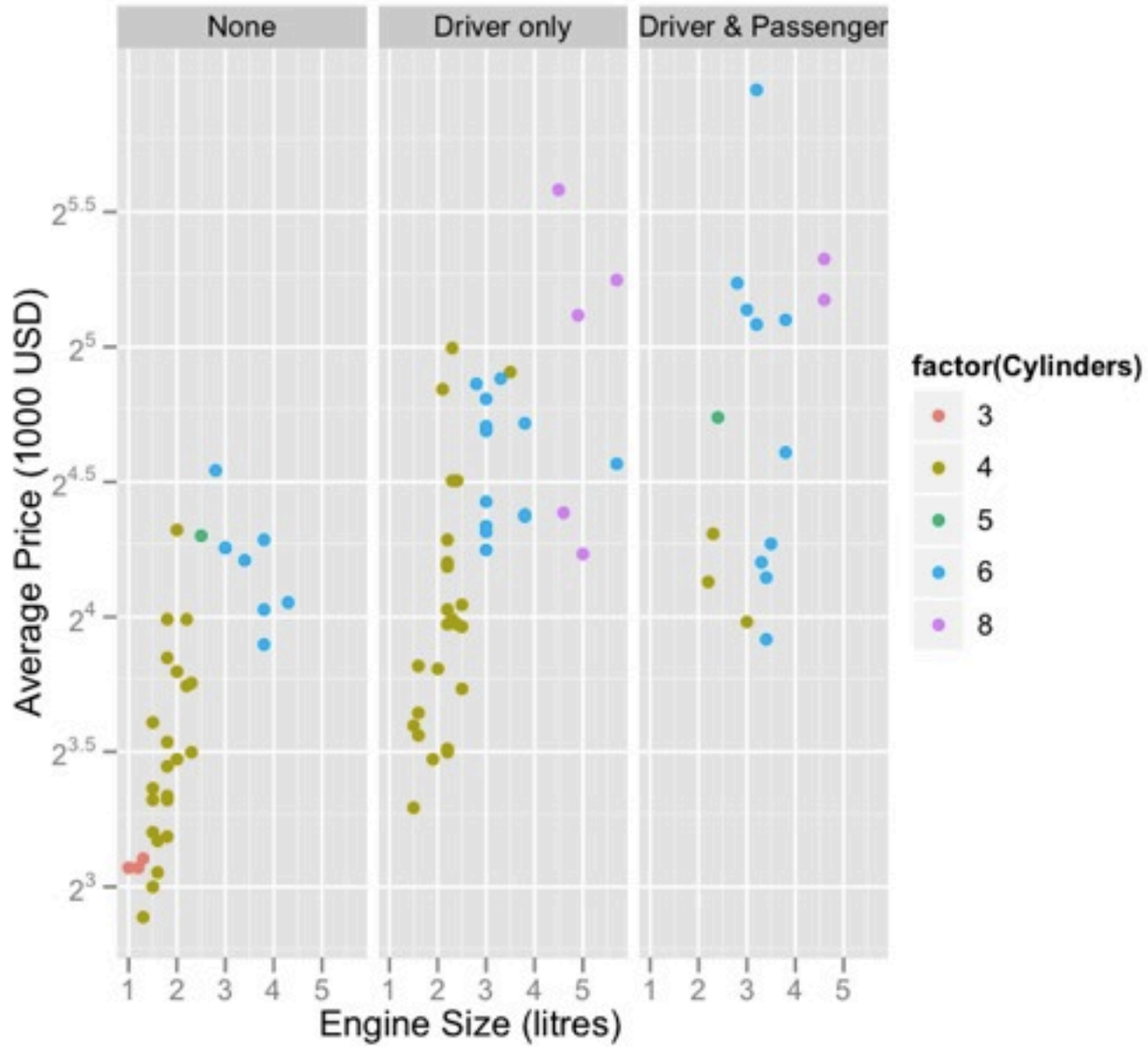
High-Dimensional Data

Taxonomy

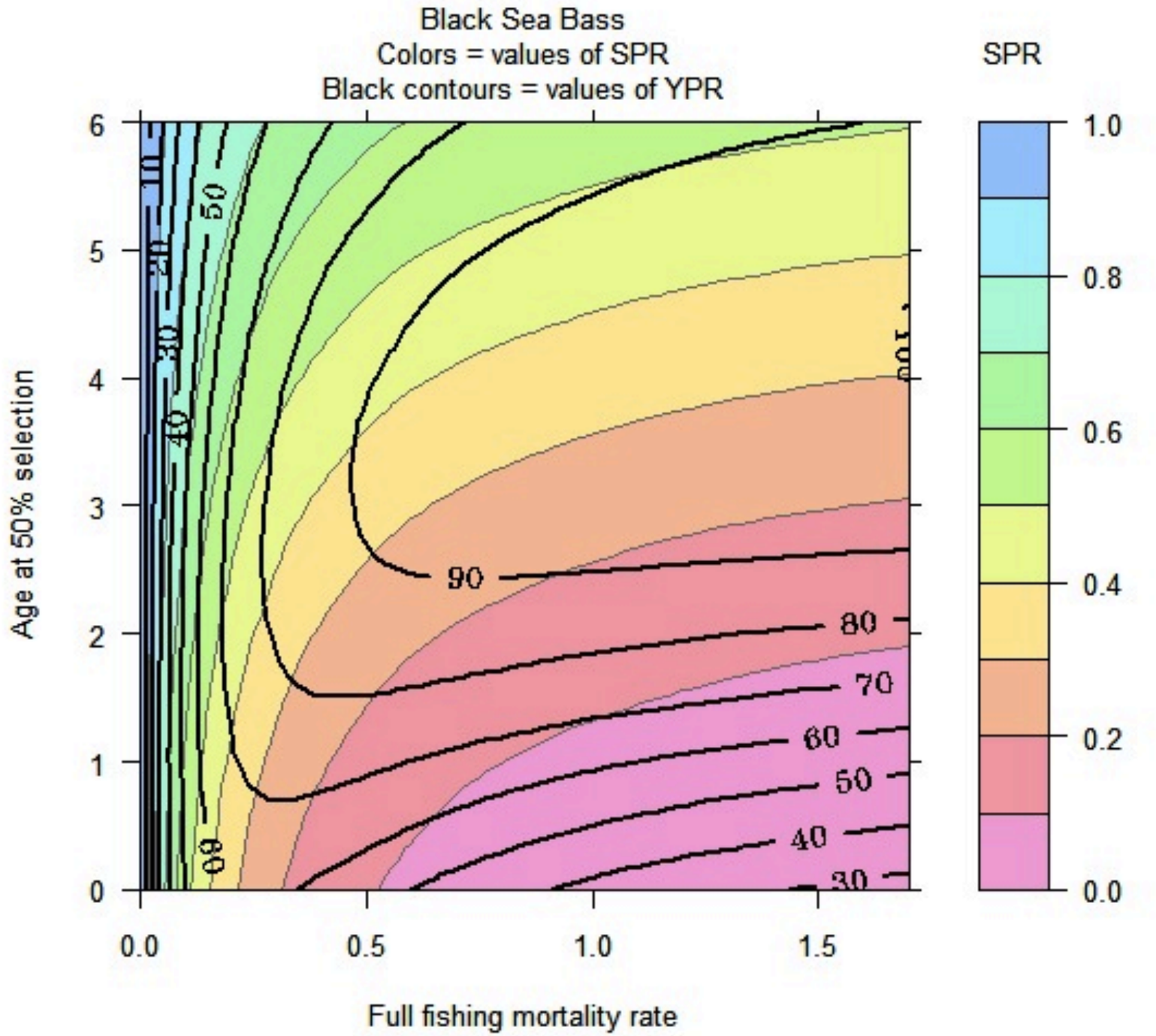
- Based on number of attributes
  - 1: Univariate
  - 2: Bivariate
  - 3: Trivariate
  - >3: Multivariate (or high-dimensional)

Multivariate Plots



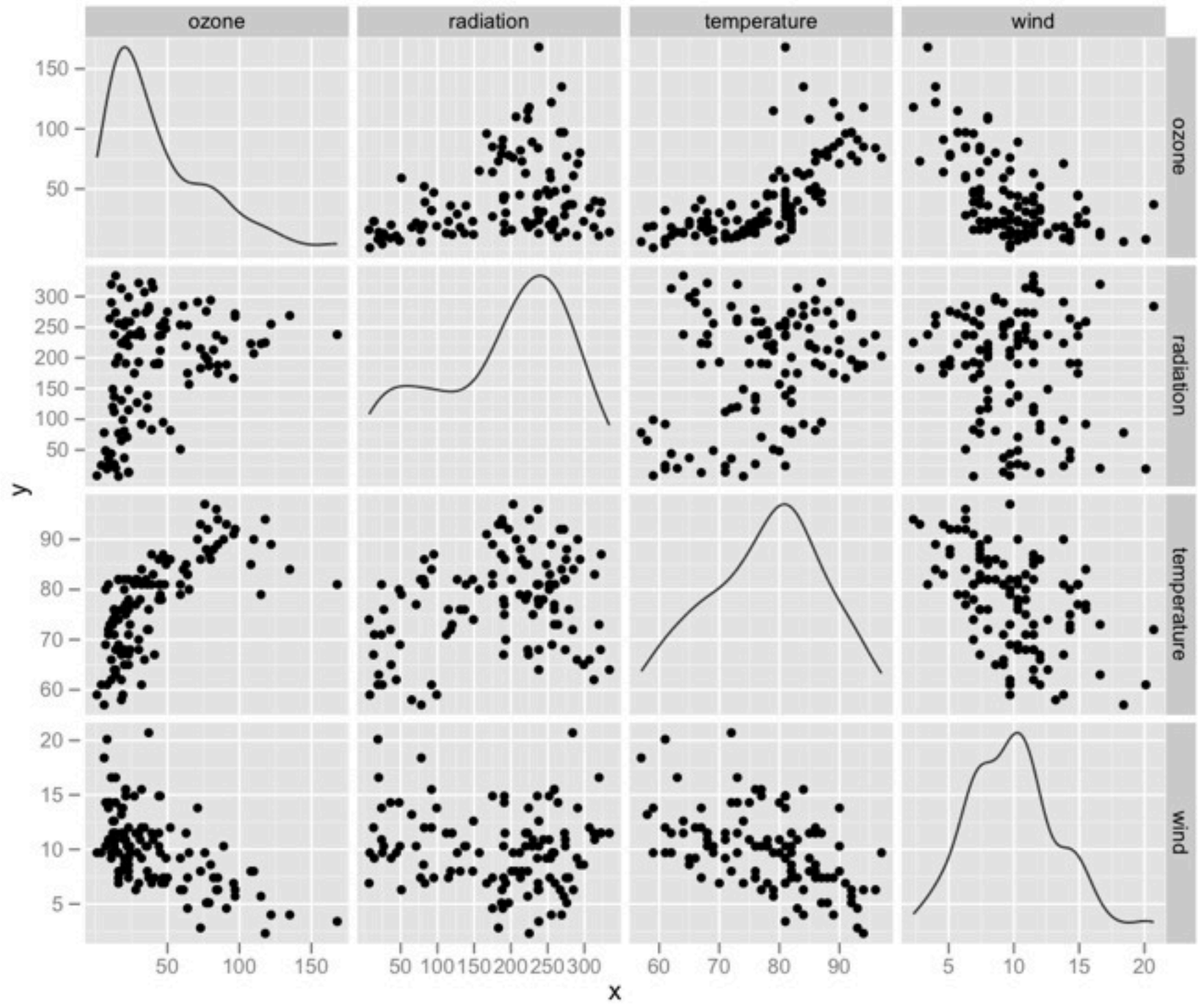


Multivariate Plots



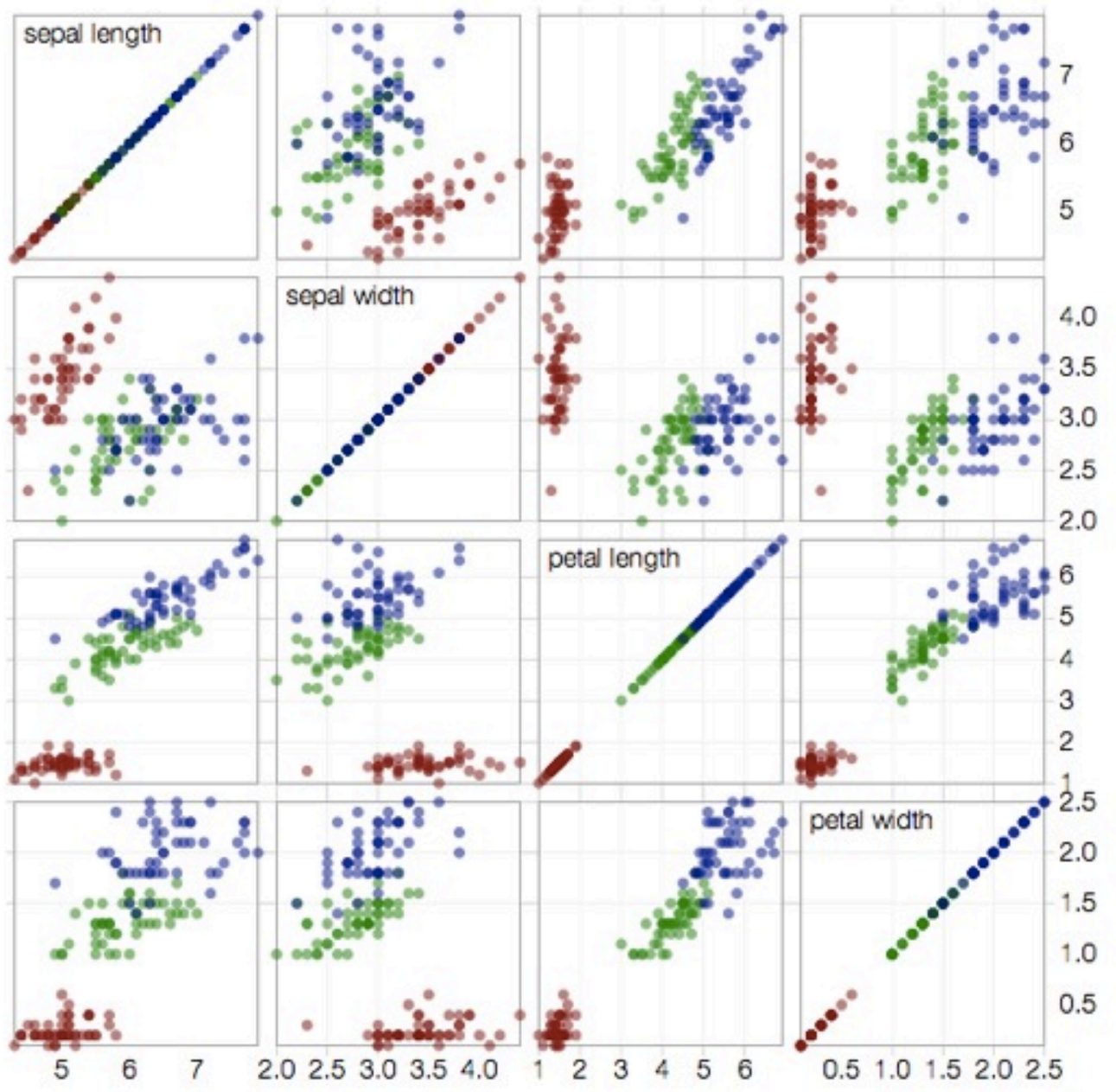
R

Scatterplot Matrix (SPLOM)



ggplot2

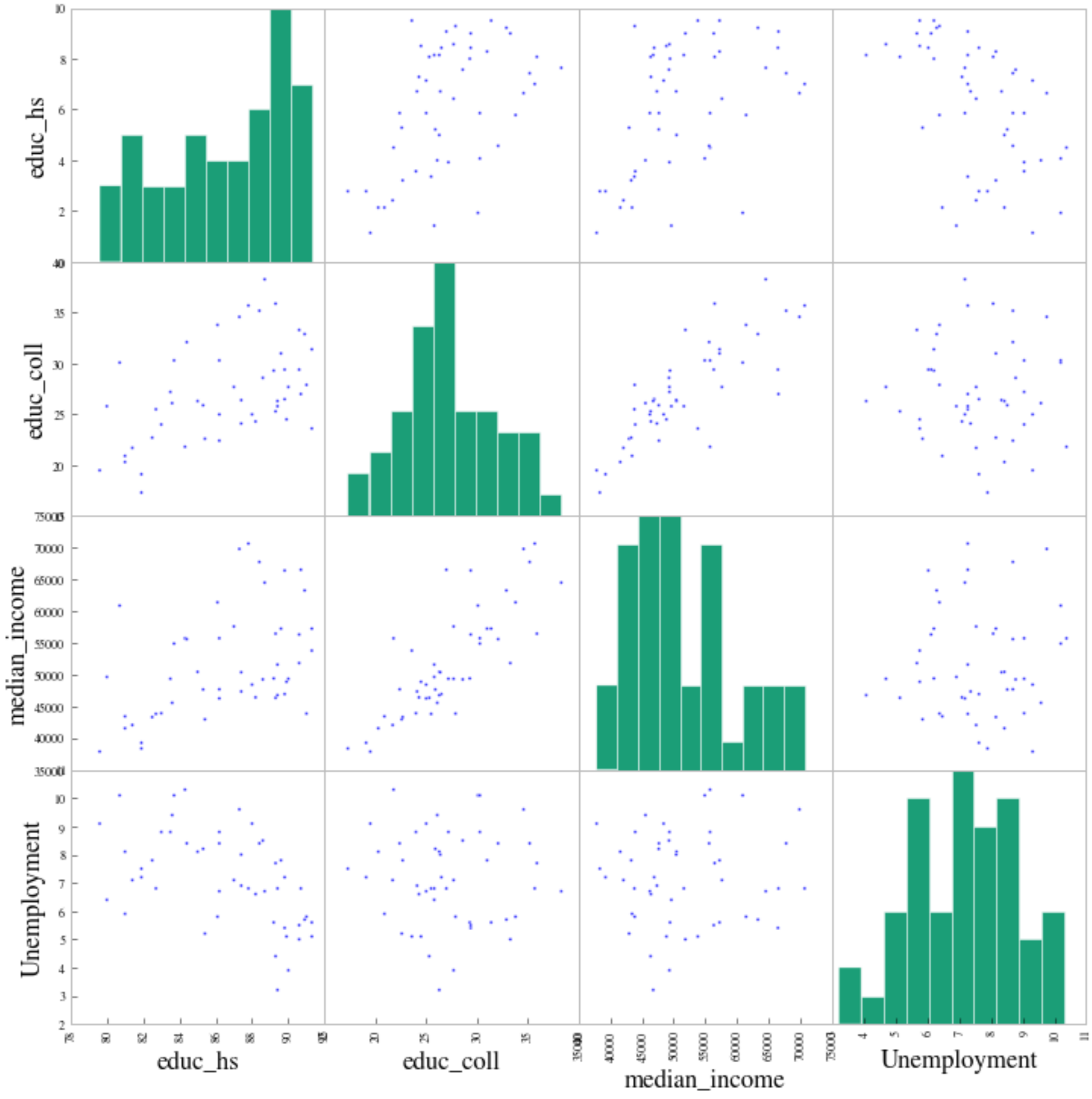
SPLOM



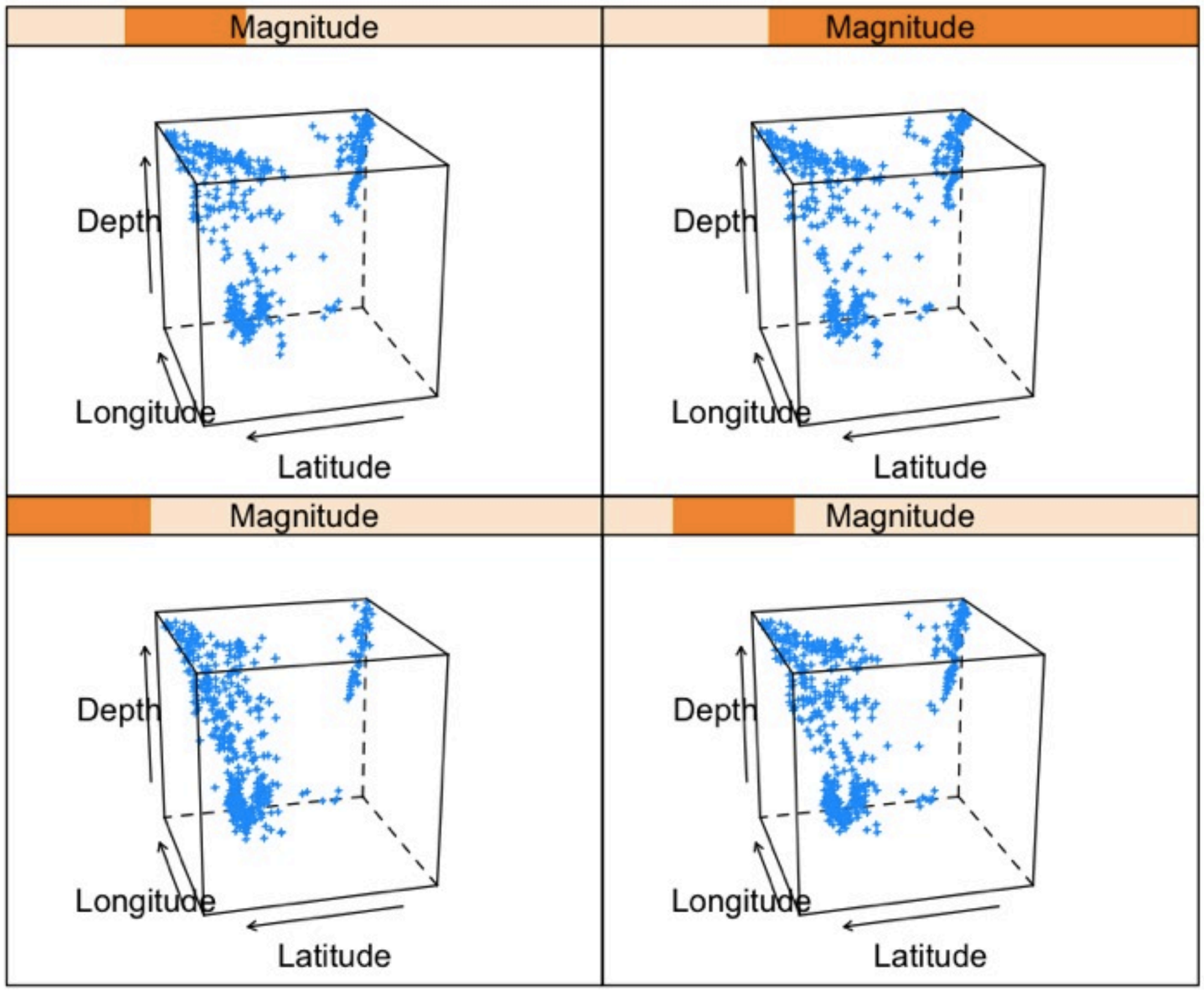
- *Iris setosa*
- *Iris versicolor*
- *Iris virginica*

Edgar Anderson's *Iris* data set
scatterplot matrix

HW2

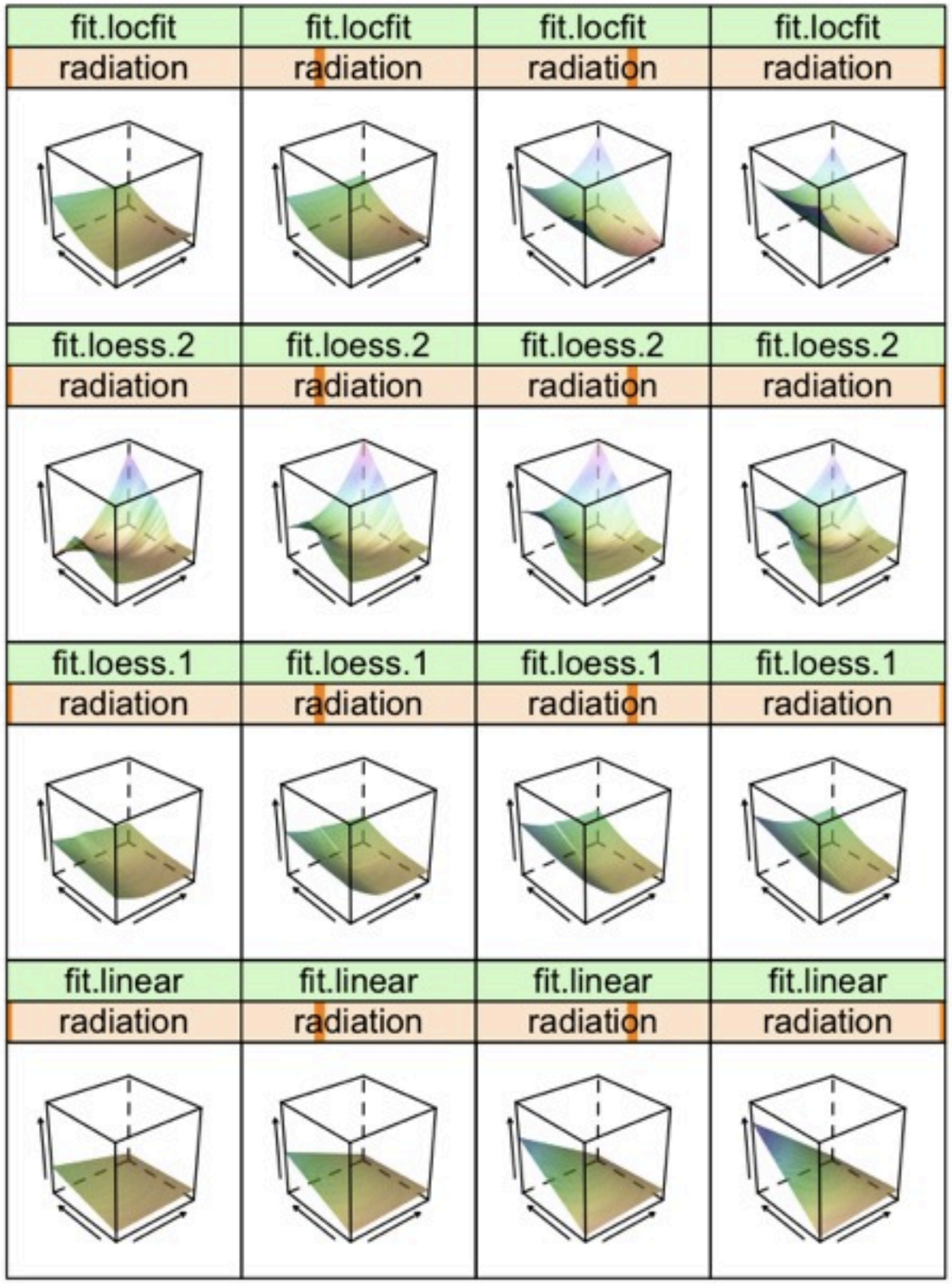


Don’t!



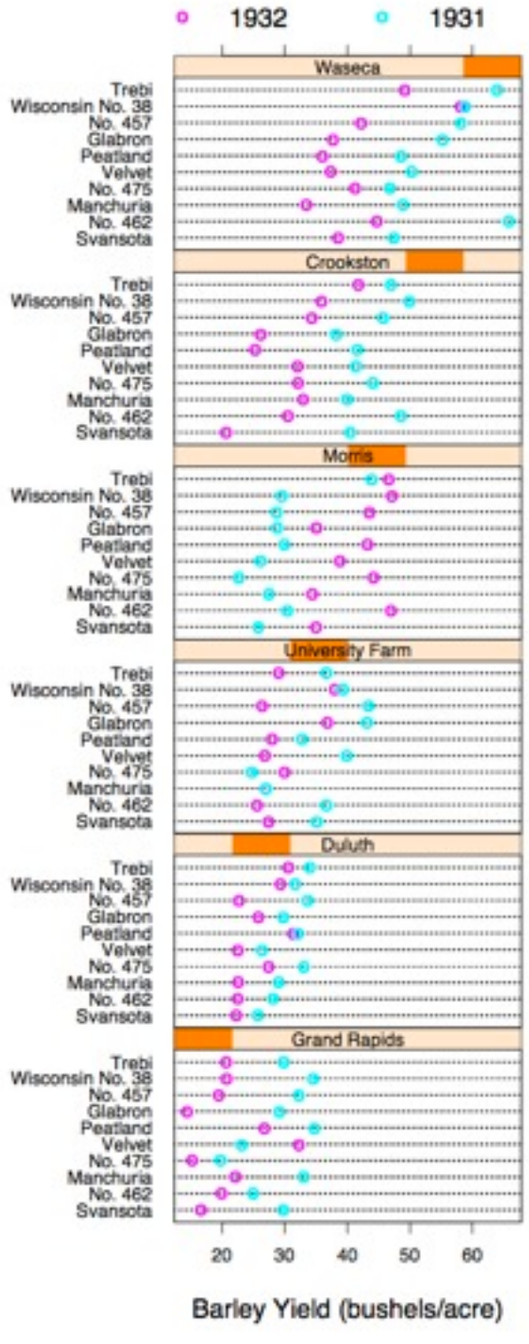
R, lattice

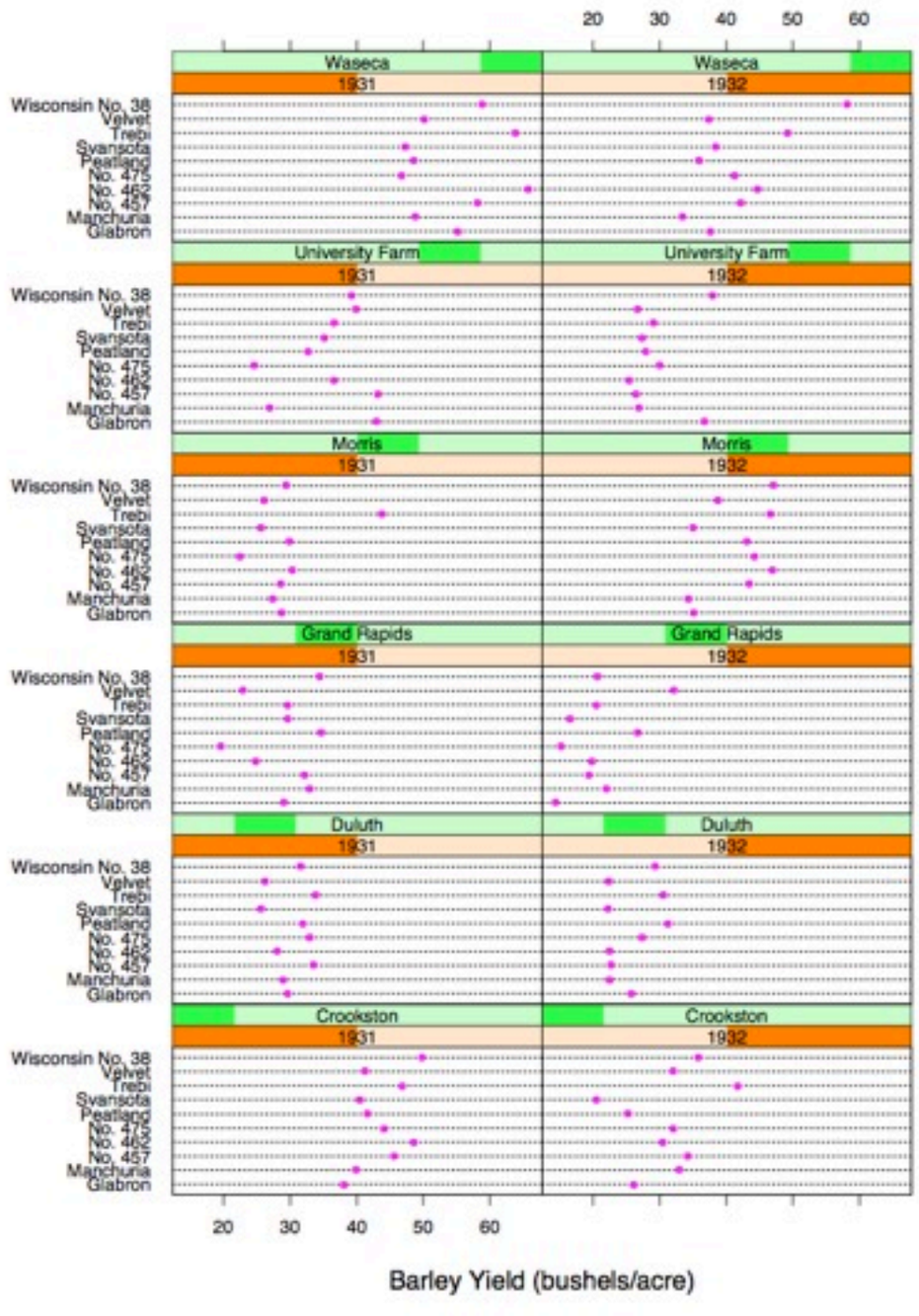
3D Surface Plots



R, lattice

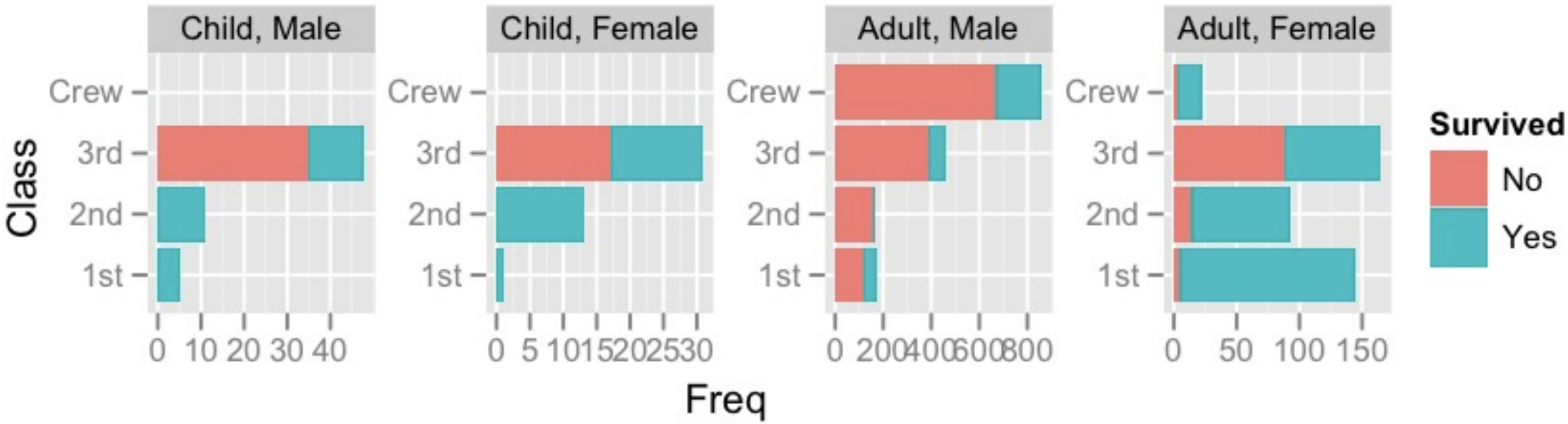
Lattice / Trellis Plots





Becker 1996

Small Multiples

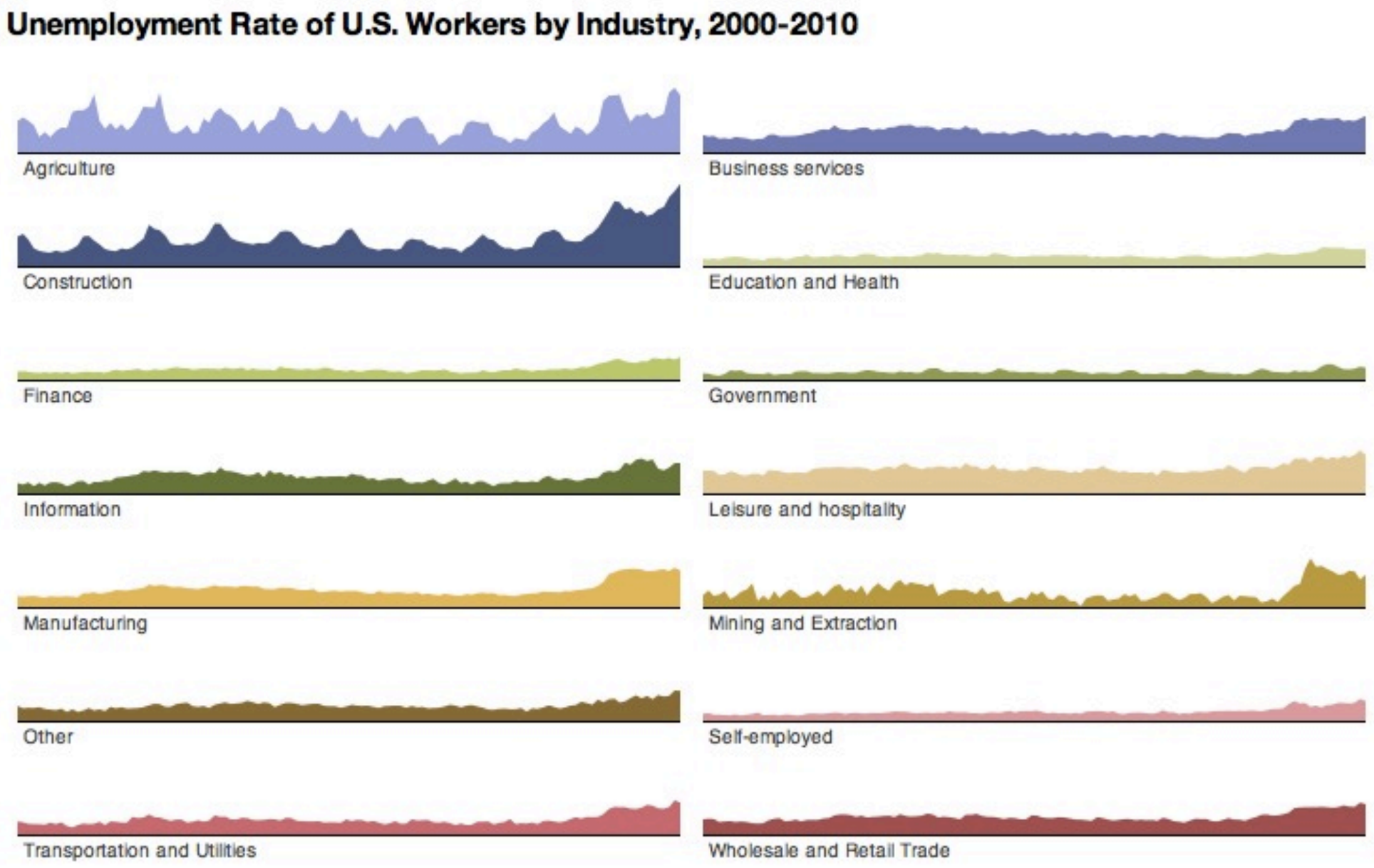


ggplot2

Small Multiples

Tableau

Small Multiples

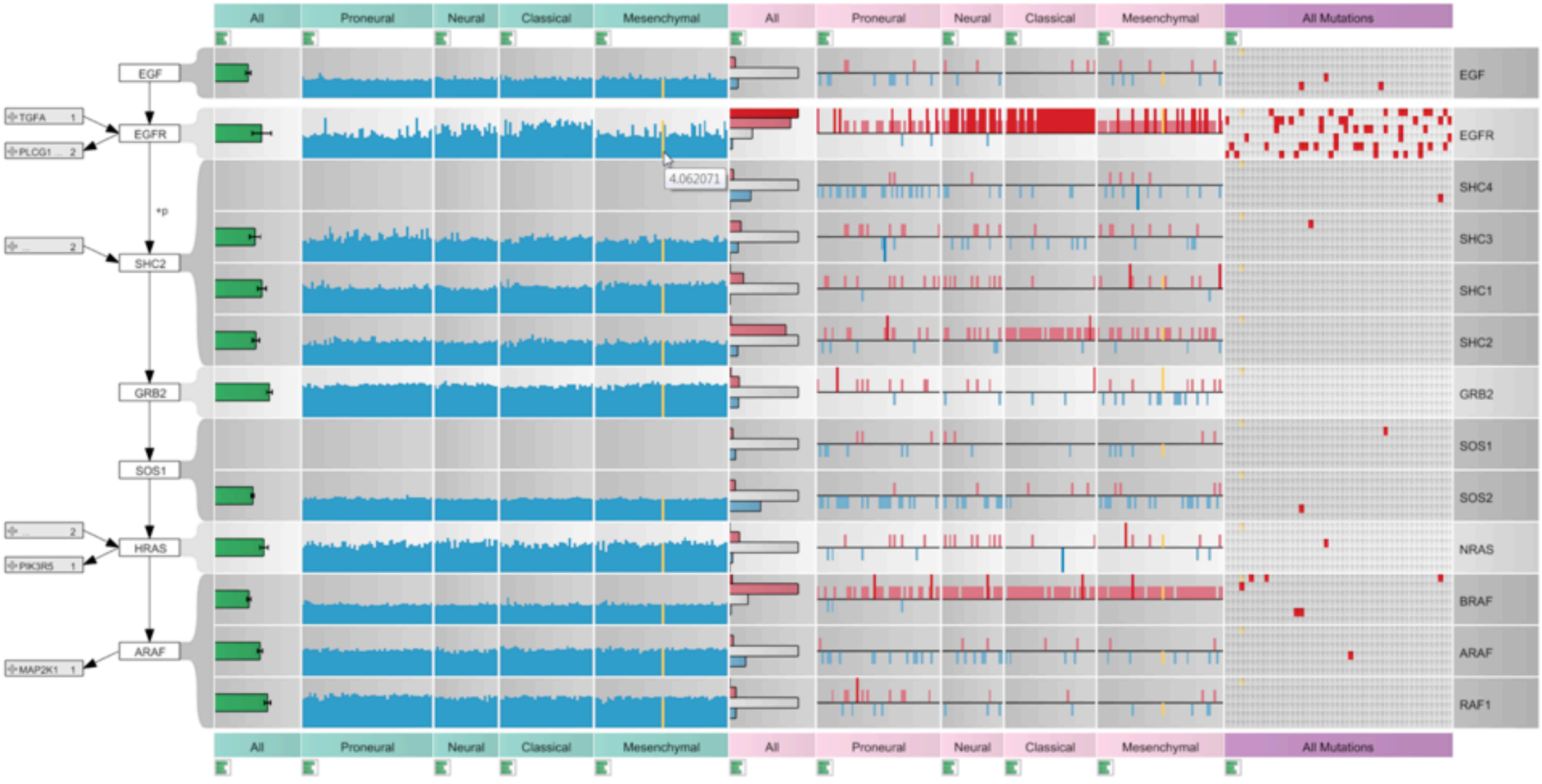


Unemployment rates of U.S. workers per industry.

Source: U.S. Bureau of Labor Statistics

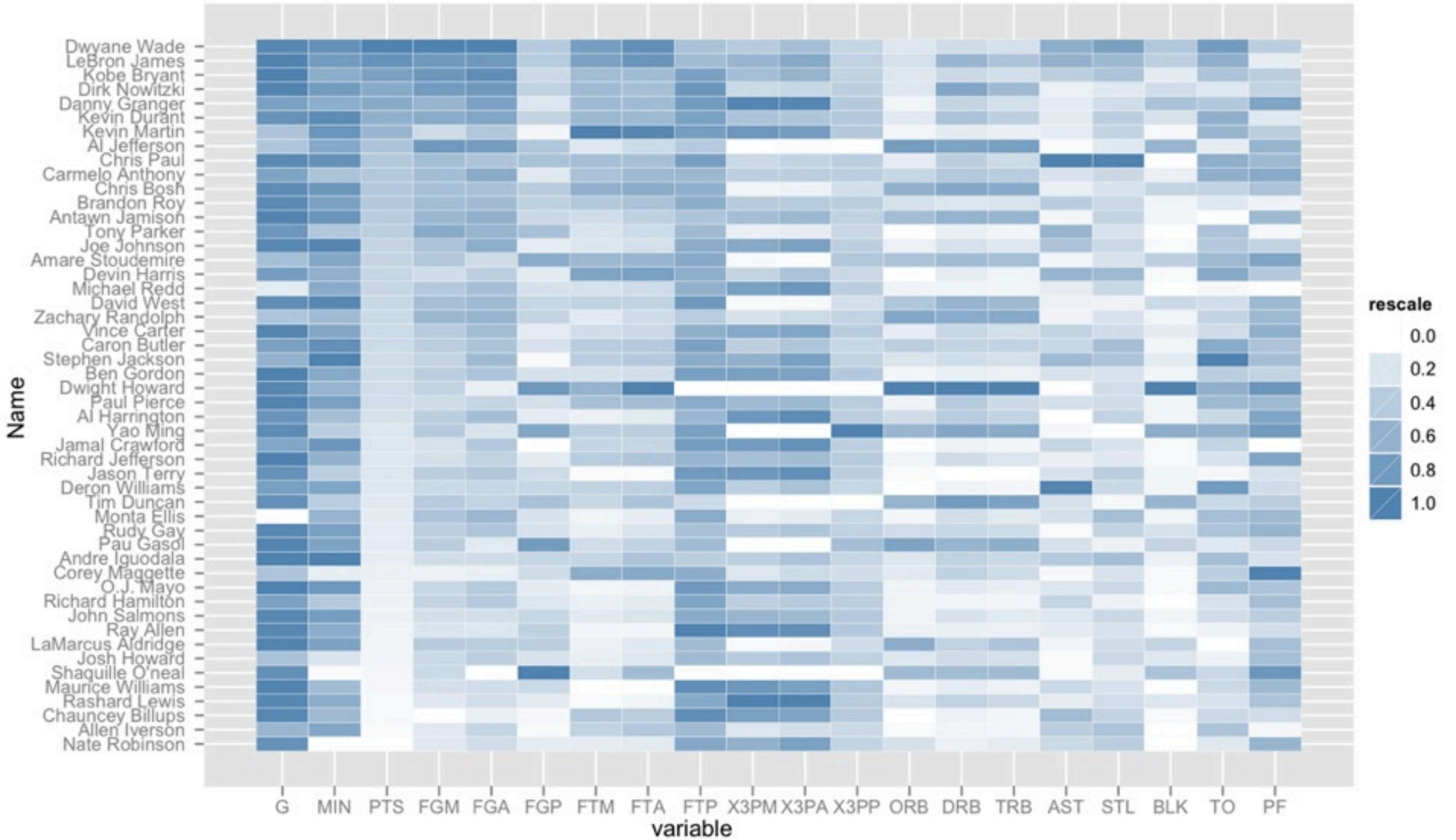
Protovis

EnRoute



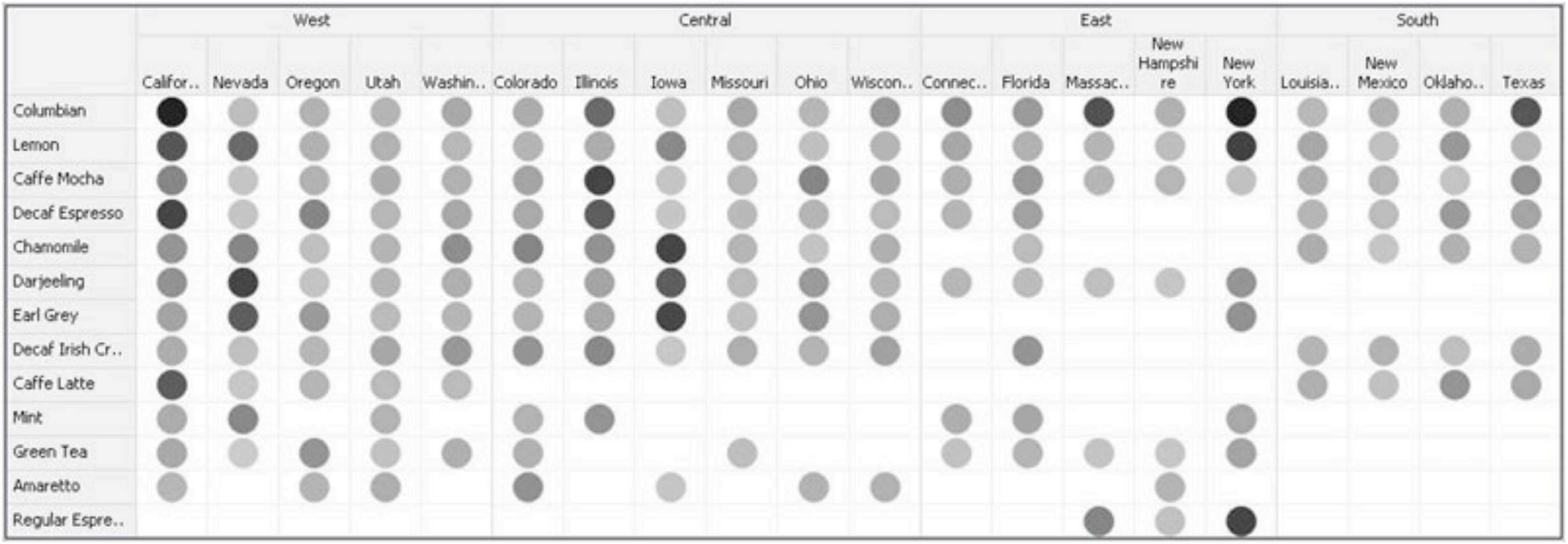
A. Lex

Heatmap



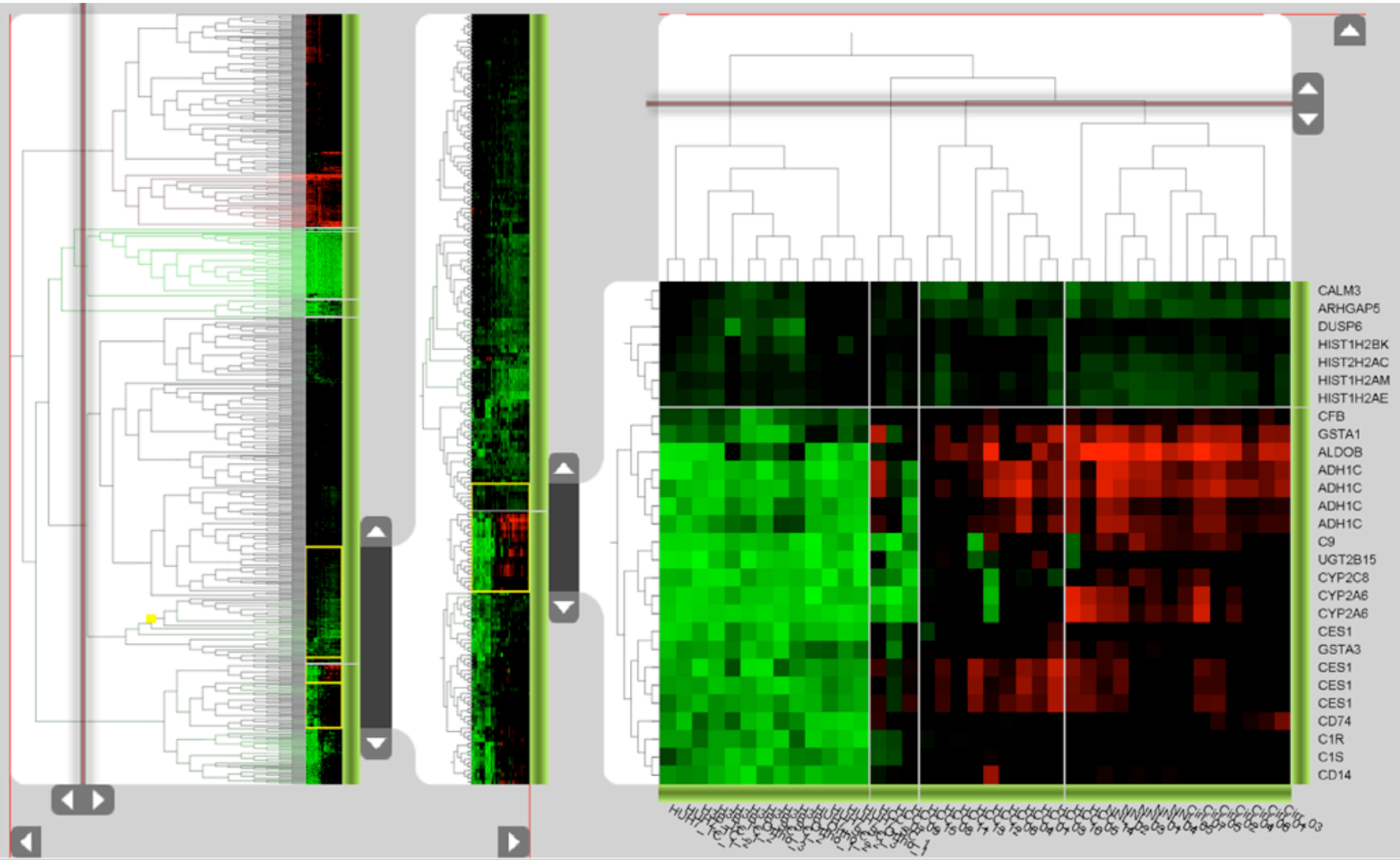
ggplot2

Heatmap



Tableau

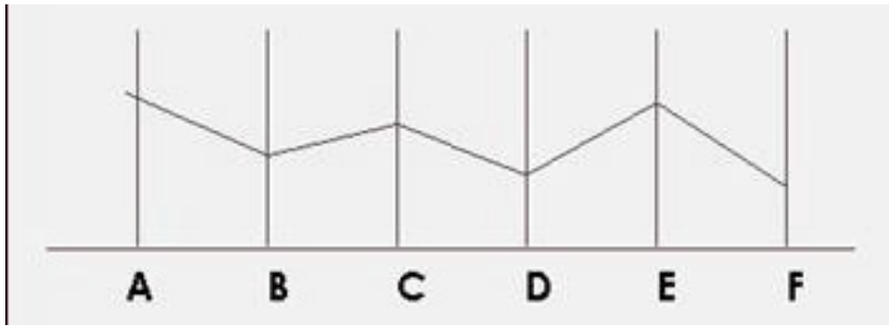
Hierarchical Heatmap

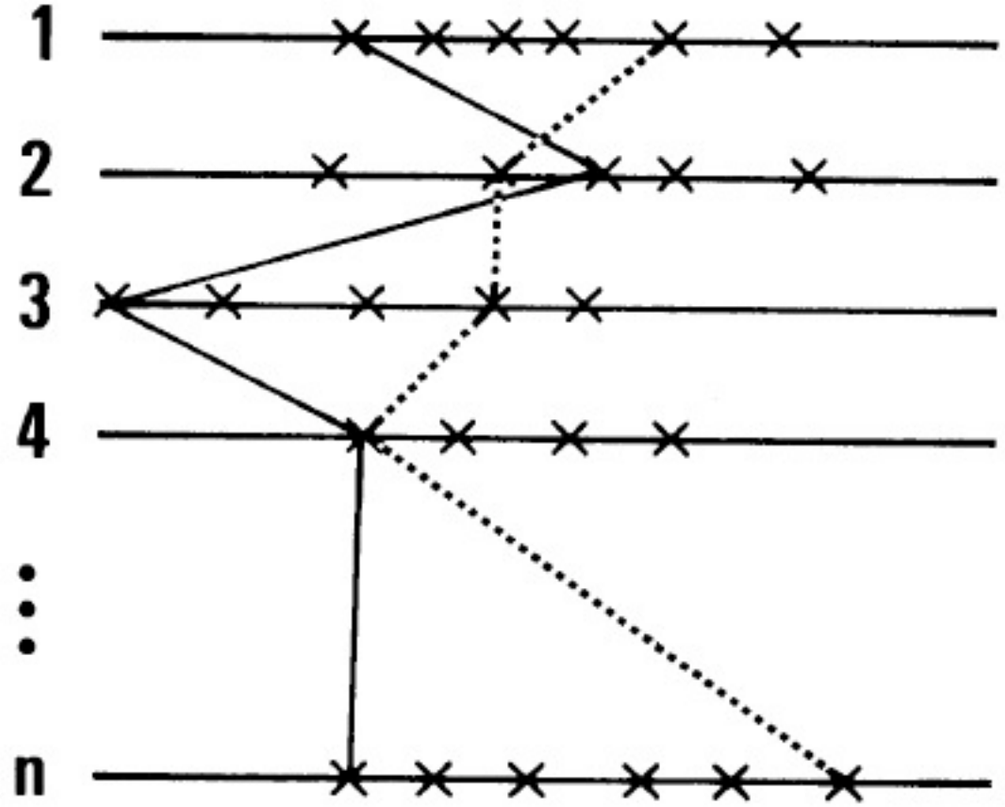


A. Lex

Parallel Coordinates

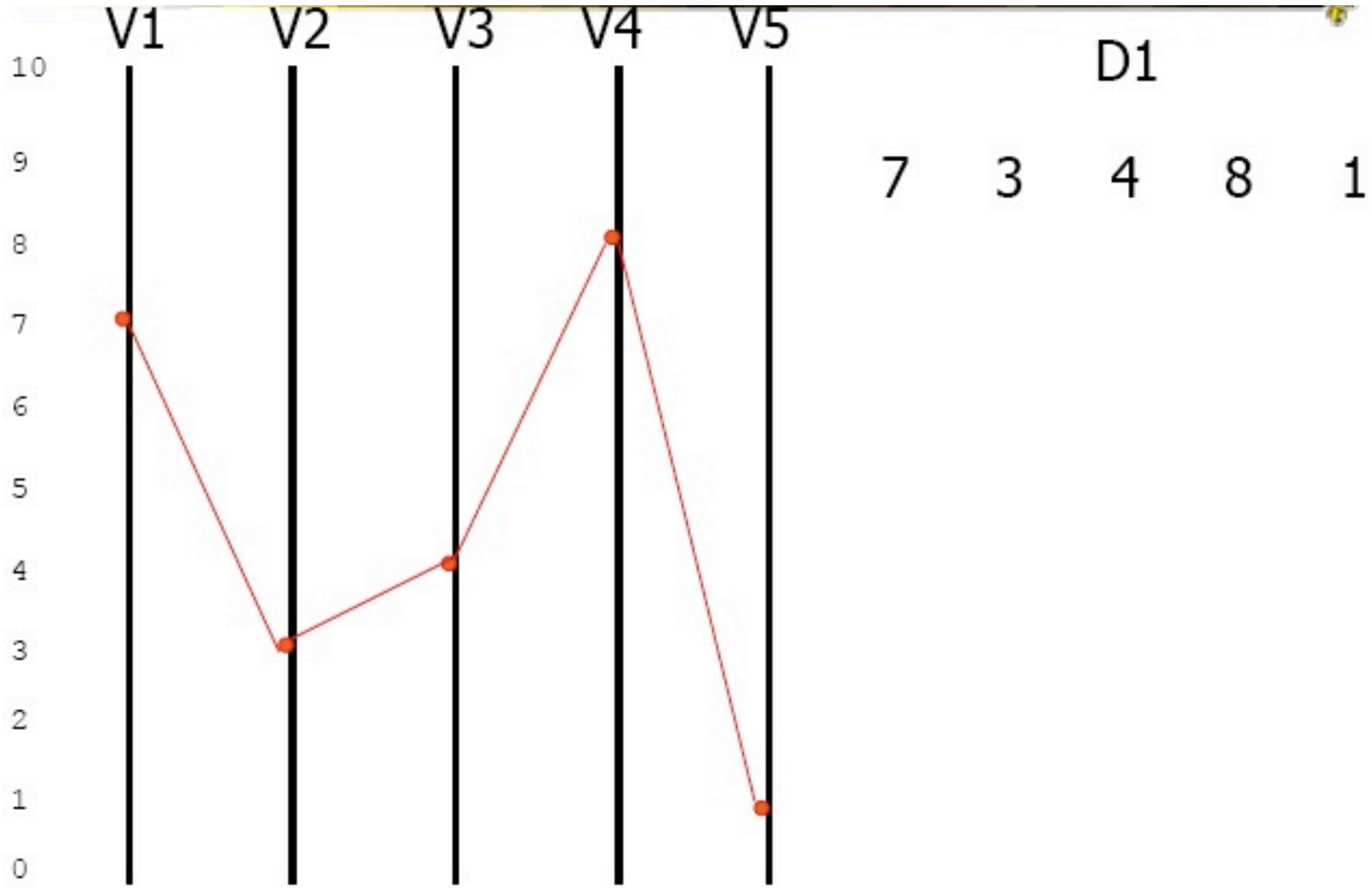
Use more than two axes



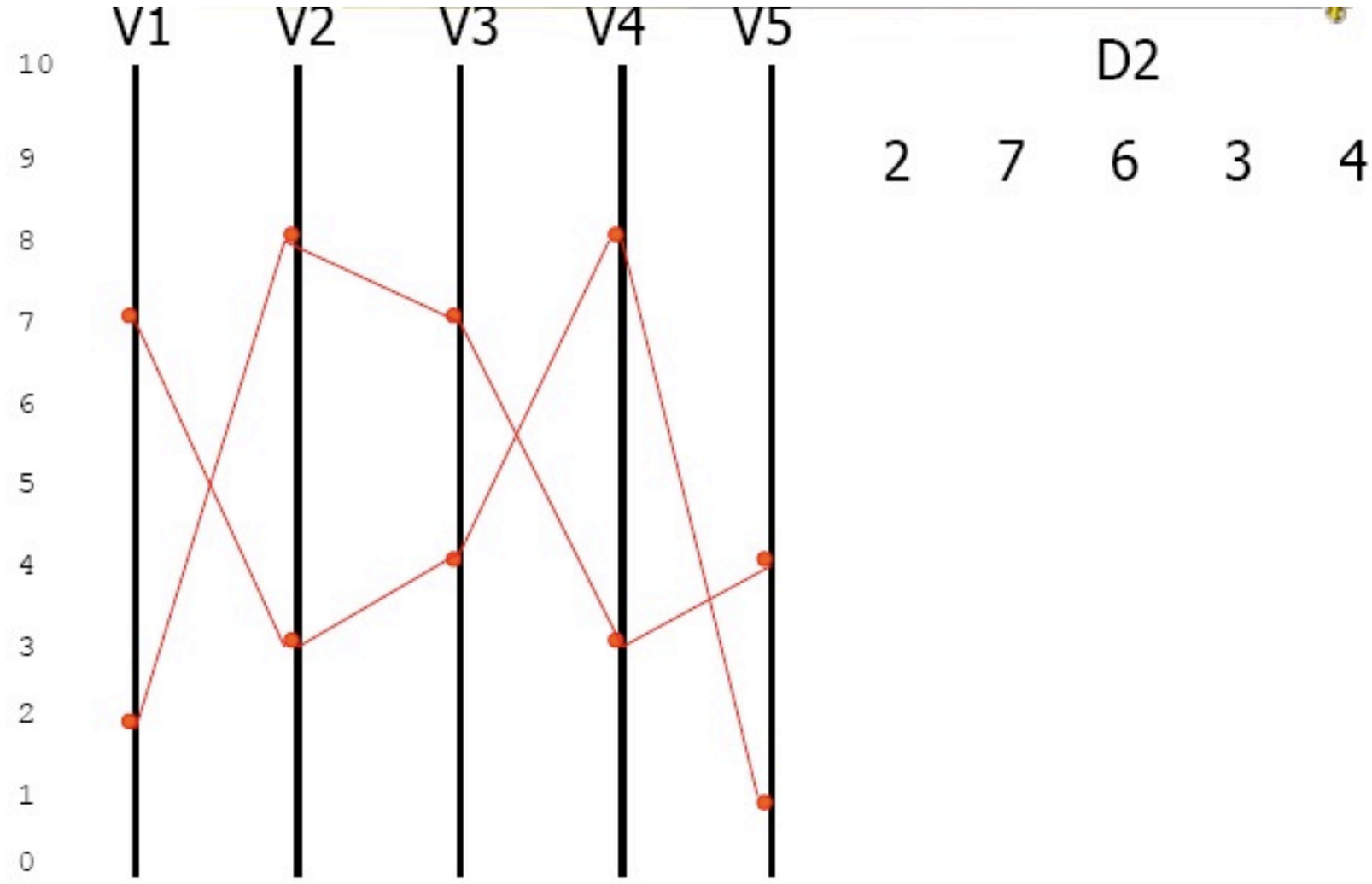


"Hyperdimensional Data Analysis Using Parallel Coordinates", Wegman, 1990
Based on slide from Munzner

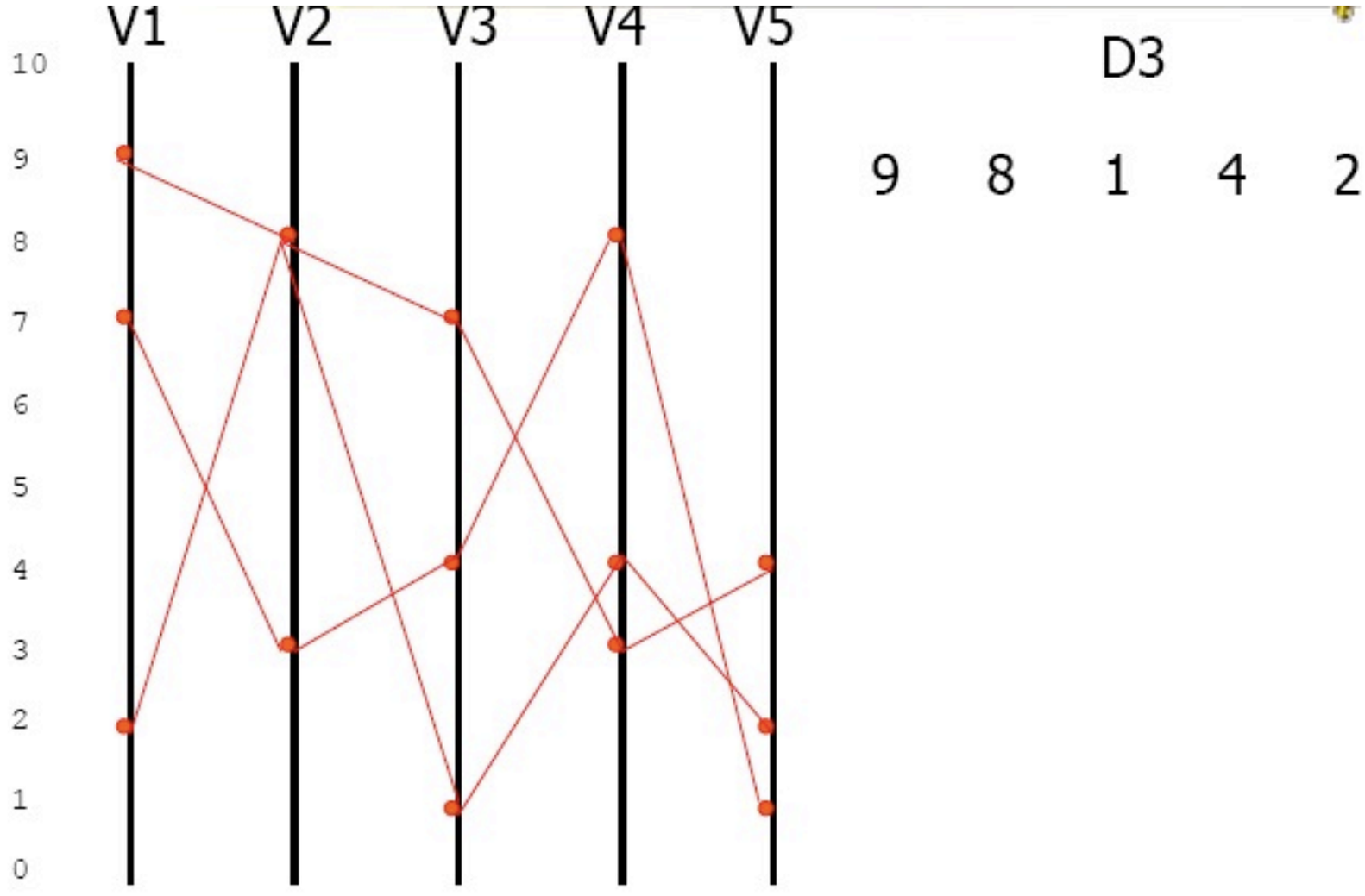
Parallel Coordinates



Parallel Coordinates



Parallel Coordinates



Correlation

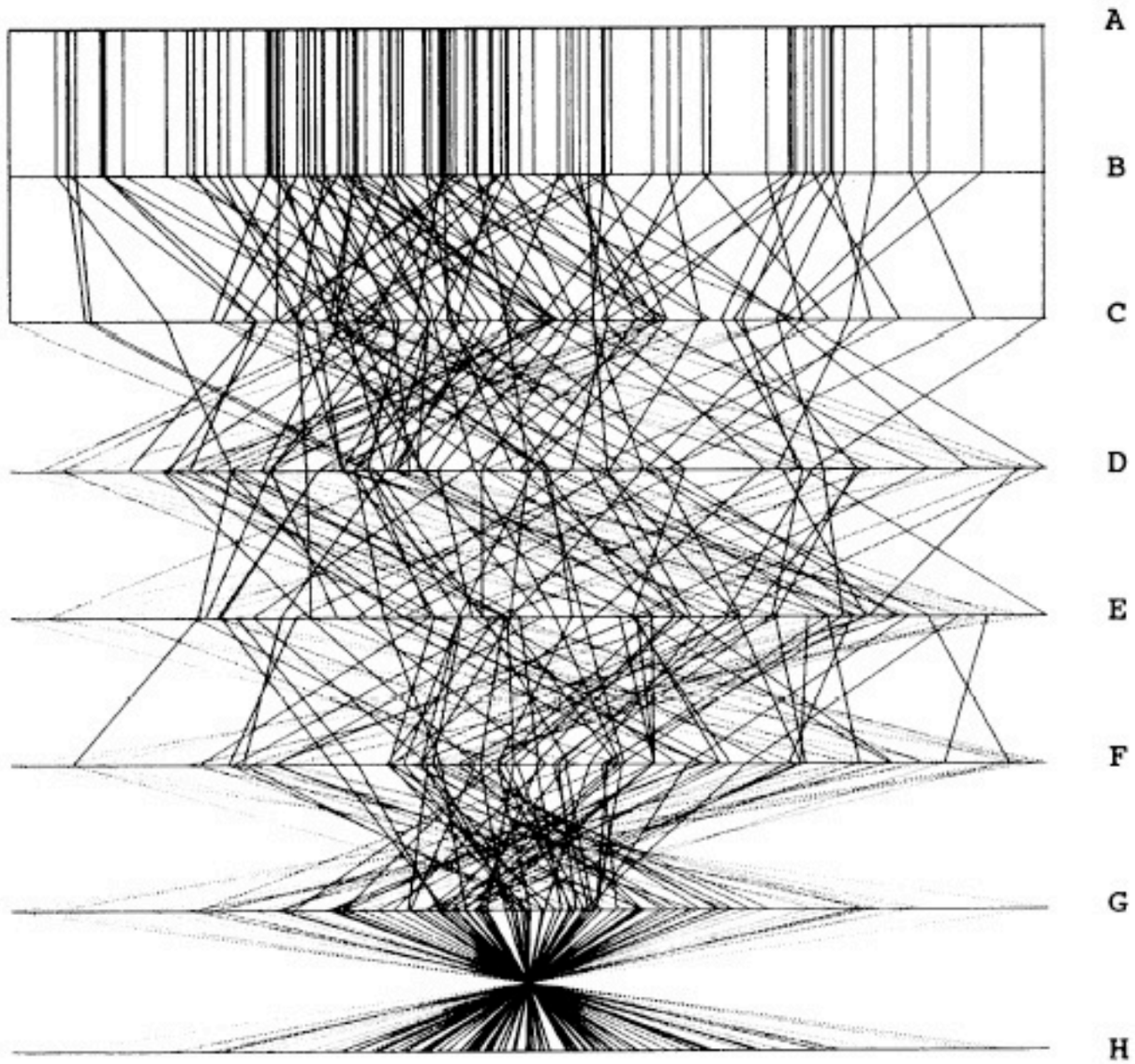
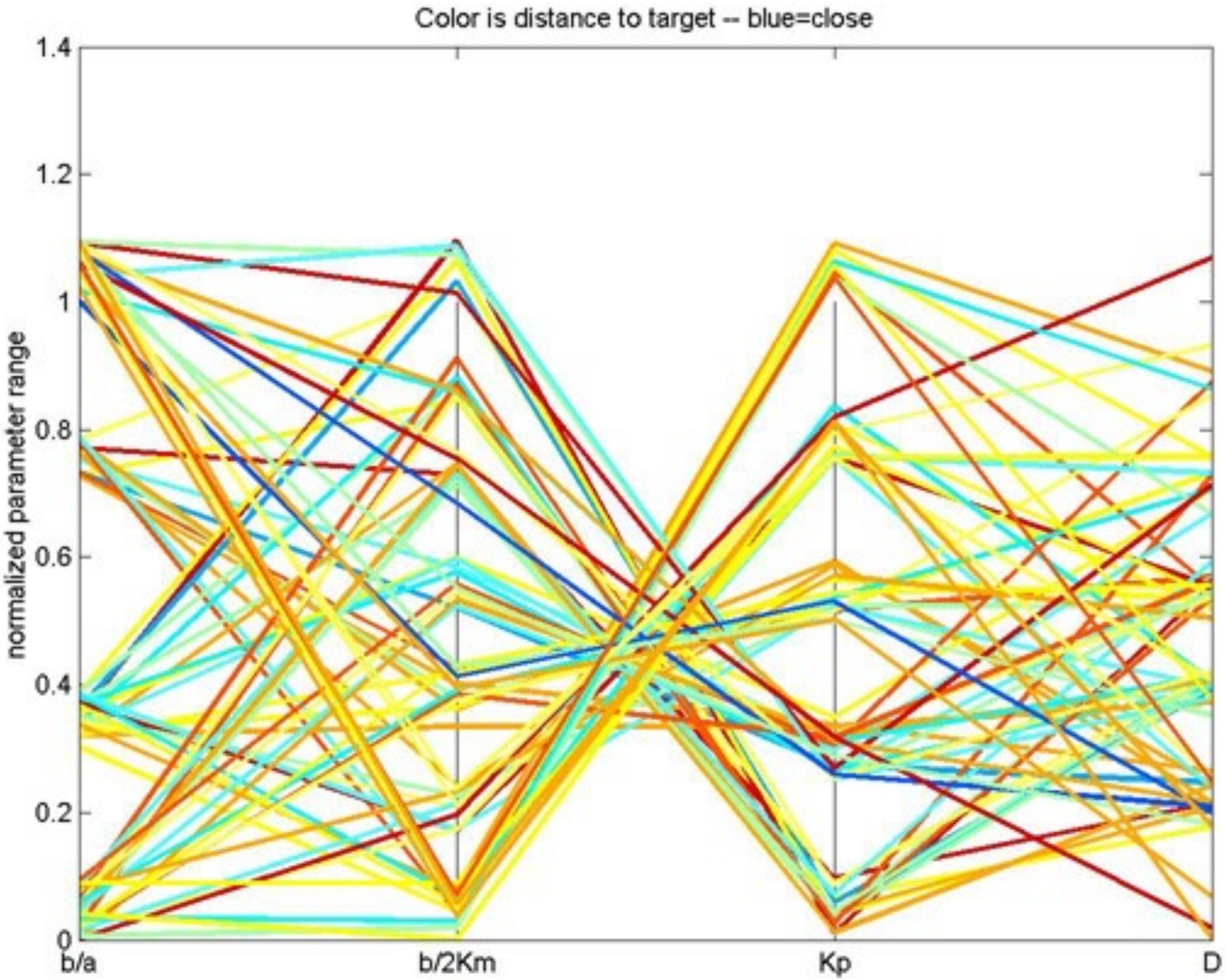
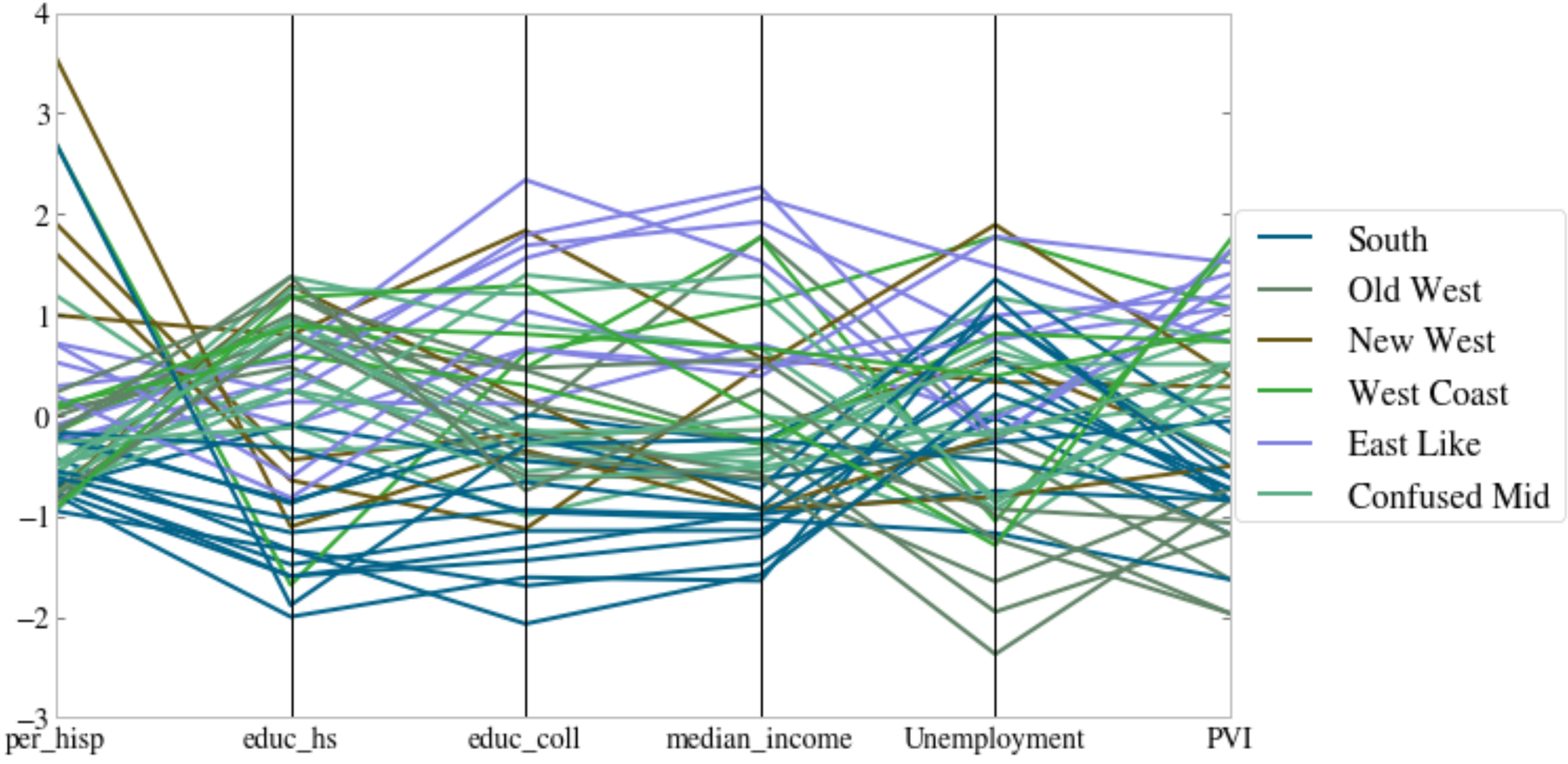


Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of $\rho = 1, .8, .2, 0, -.2, -.8,$ and -1.

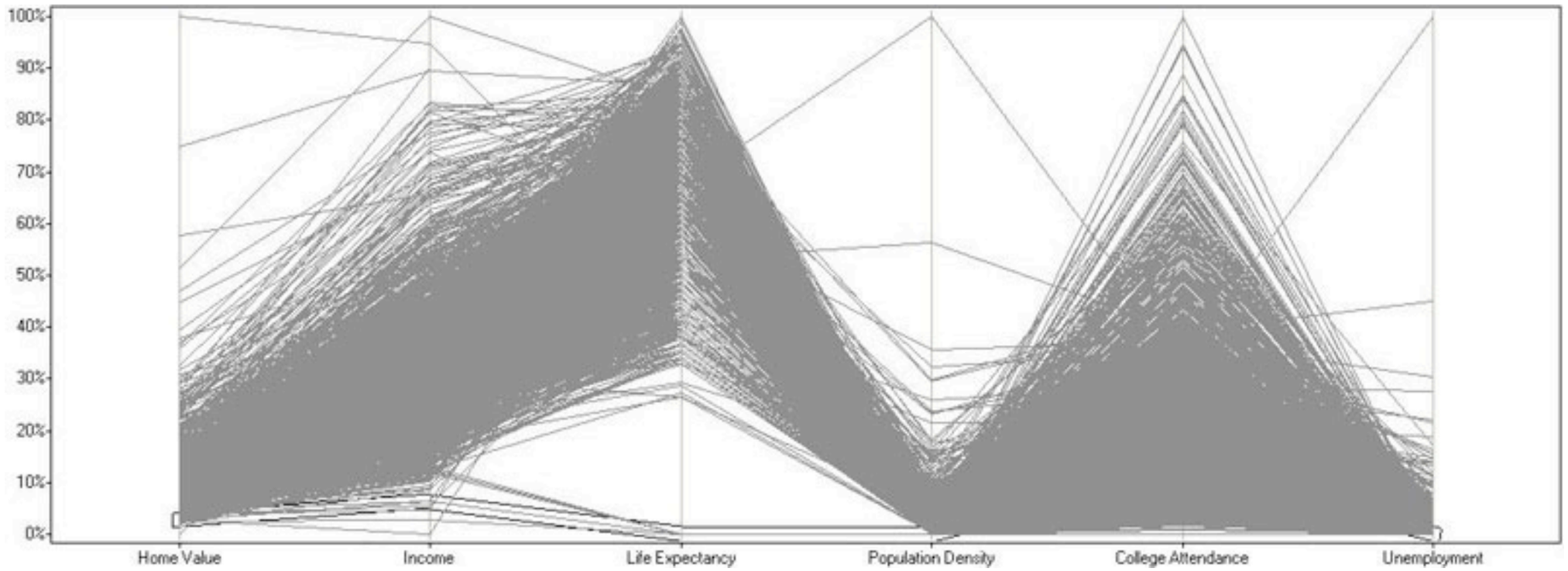
"Hyperdimensional Data Analysis Using Parallel Coordinates", Wegman, 1990
Based on slide from Munzner

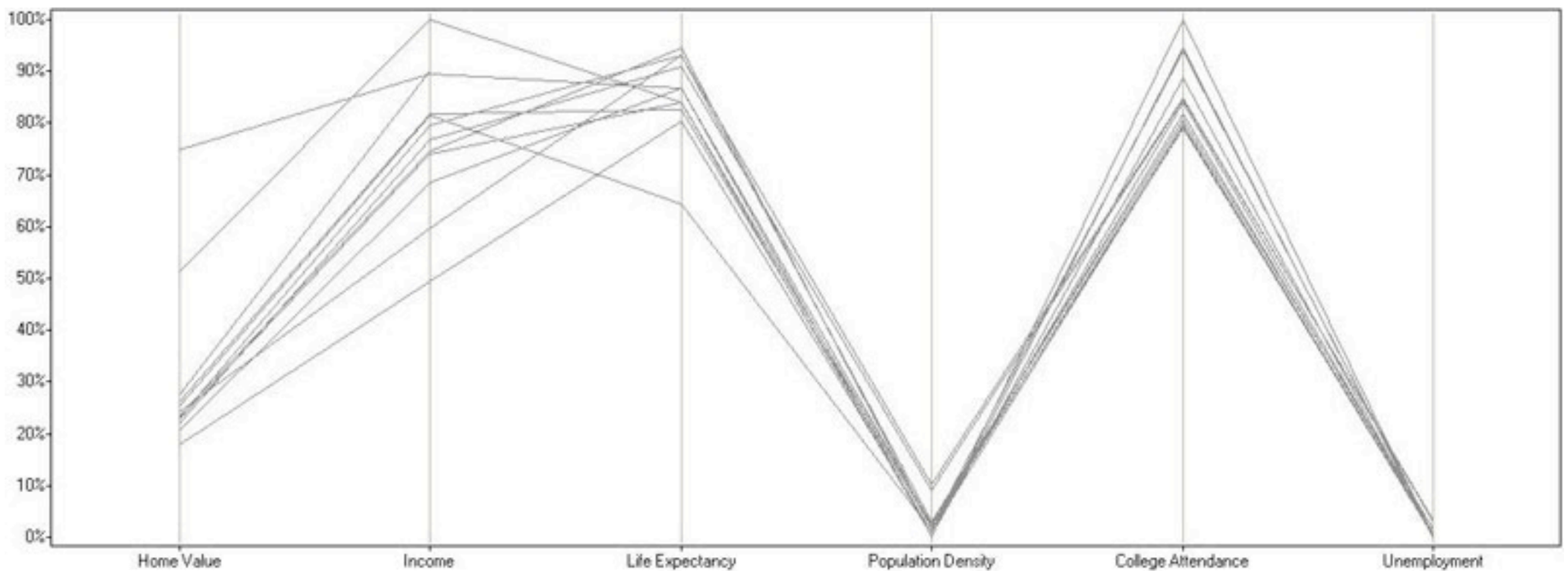

Color is distance to target -- blue=close
normalized parameter range
1.4
1.2
1
0.8
0.6
0.4
0.2
0
b/a
b/2Km
Kp
D

HW2


4
3
2
1
0
−1
−2
−3
per_hisp
educ_hs
educ_coll
median_income
Unemployment
PVI
South
Old West
New West
West Coast
East Like
Confused Mid

Filtering


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Home Value
Income
Life Expectancy
Population Density
College Attendance
Unemployment


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Home Value
Income
Life Expectancy
Population Density
College Attendance
Unemployment

Filtering & Brushing

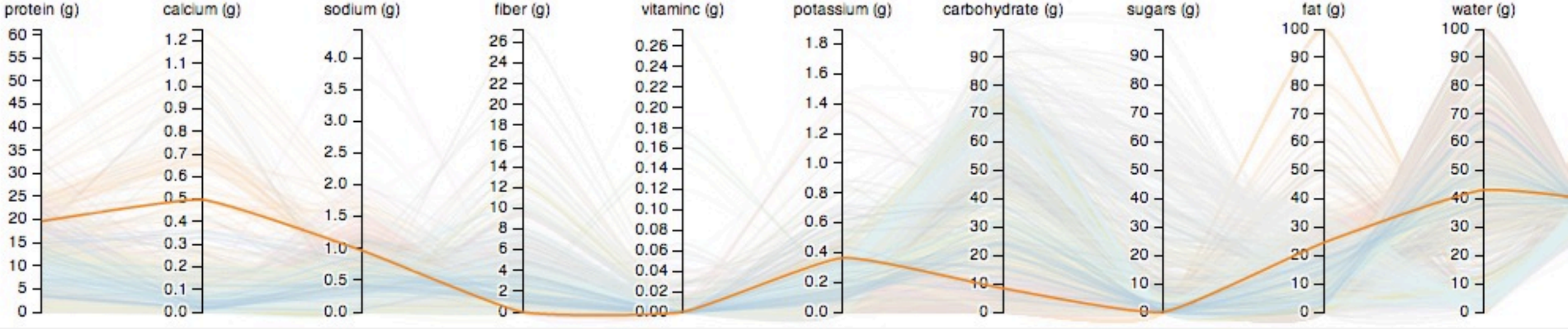
Nutrient Contents – Parallel Coordinates

An interactive visualization of the USDA Nutrient Database. For information on parallel coordinates, read this tutorial.

Per 100g of Food

Selected 1153 rows Keep Re

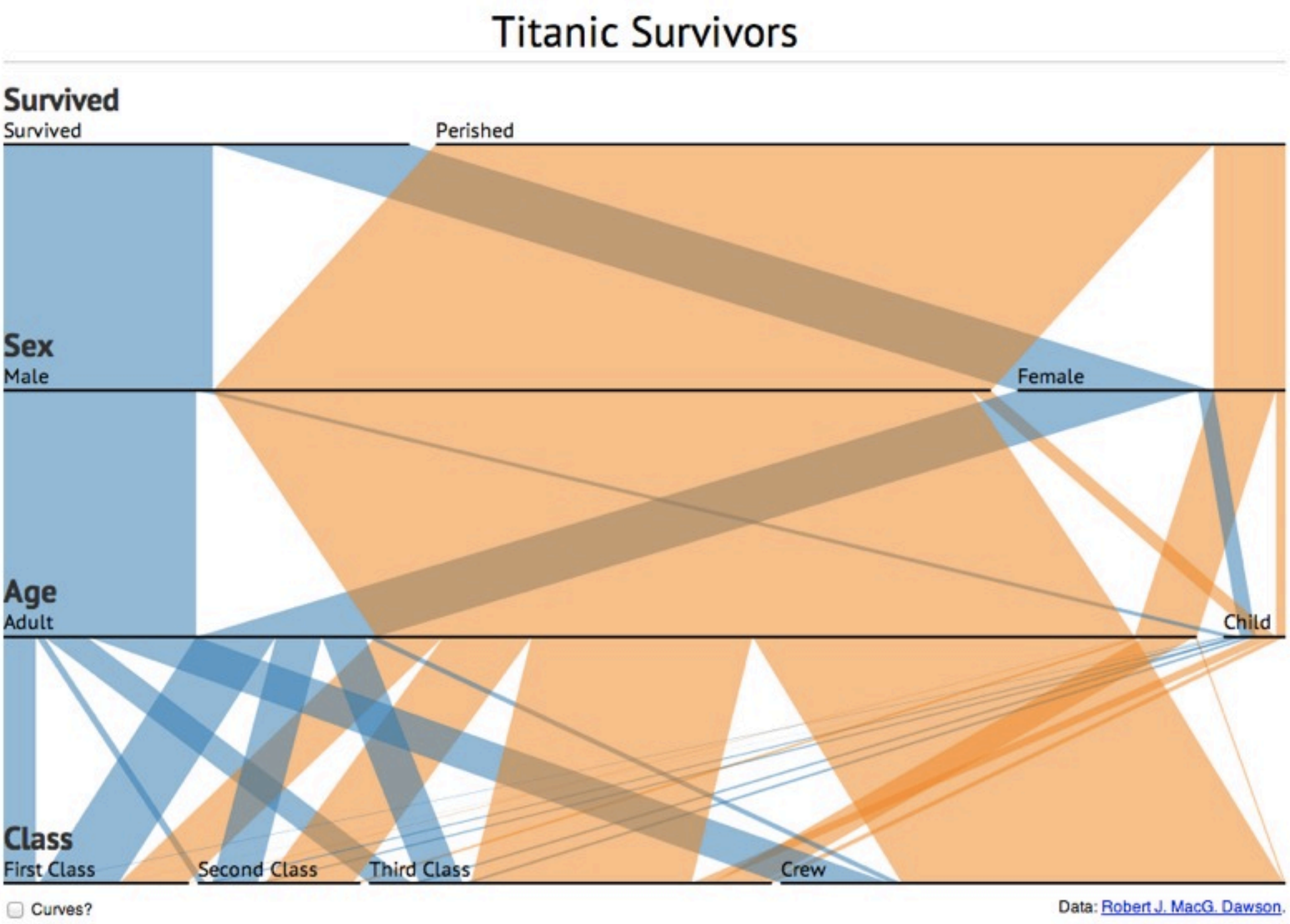
Dairy and Egg Products · Fats and Oils · Poultry Products · Soups, Sauces, and Gravies · Vegetables and Vegetable Products · Sausages and Luncheon Meats · Breakfas
Fruits and Fruit Juices · Nut and Seed Products · Beverages · Finfish and Shellfish Products · Legumes and Legume Products · Baked Products · Sweets · Cereal Gra
Fast Foods · Meals, Entrees, and Sidedishes · Snacks · Restaurant Foods



name	group	protein (g)	calcium …	sodium …	fiber (g)	vitaminc…	potassiu…	carboh
Butter oil, anhydrous	Dairy and Egg Products	0.28	0.004	0.002		0	0.005	
Butter, salted	Dairy and Egg Products	0.85	0.024	0.714		0	0.024	0.06
Cheese fondue	Dairy and Egg Products	14.23	0.476	0.132		0	0.105	3.77
Cheese food, cold pack, american	Dairy and Egg Products	19.66	0.497	0.966		0	0.363	8.32

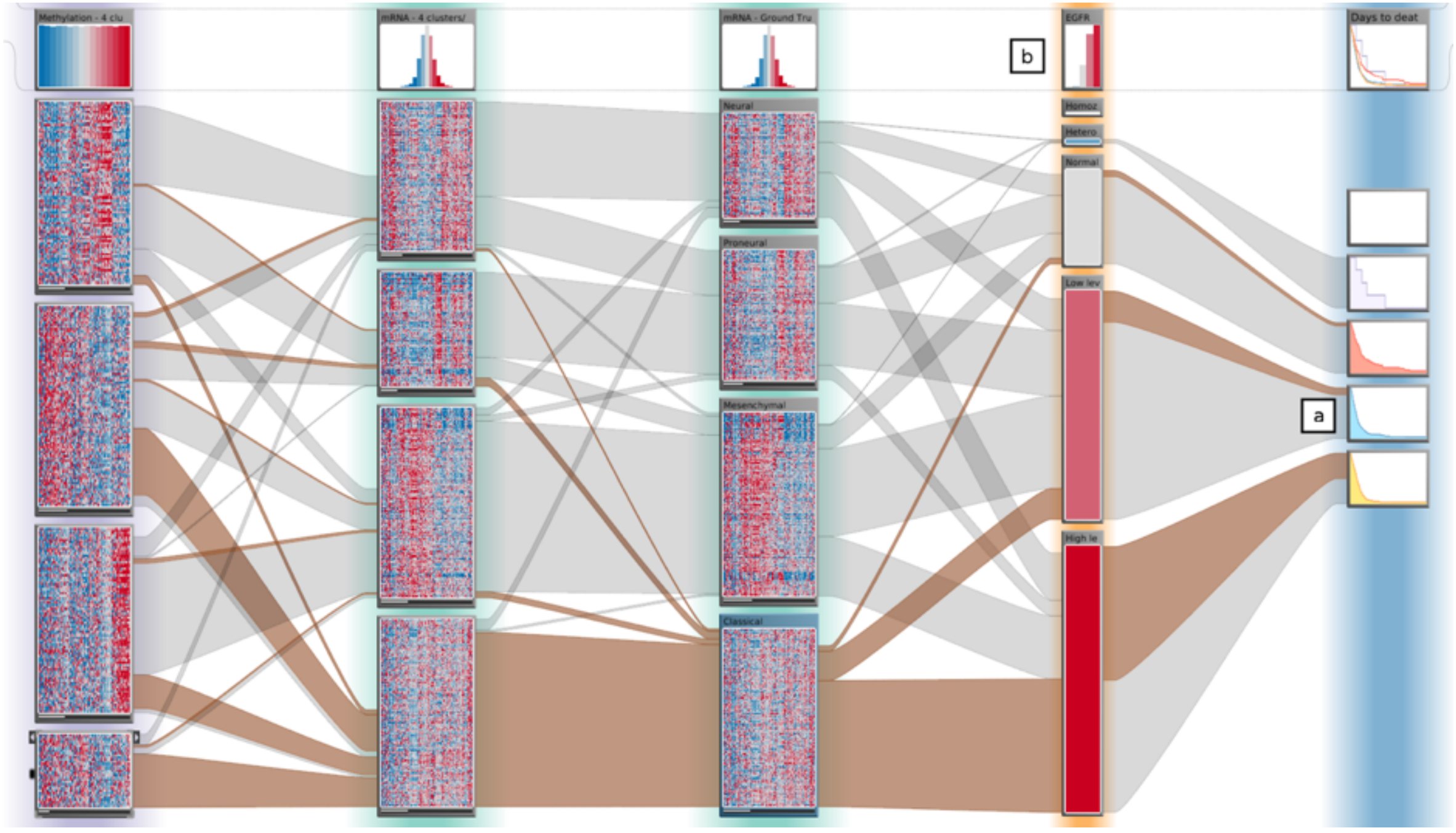
D3

Parallel Sets



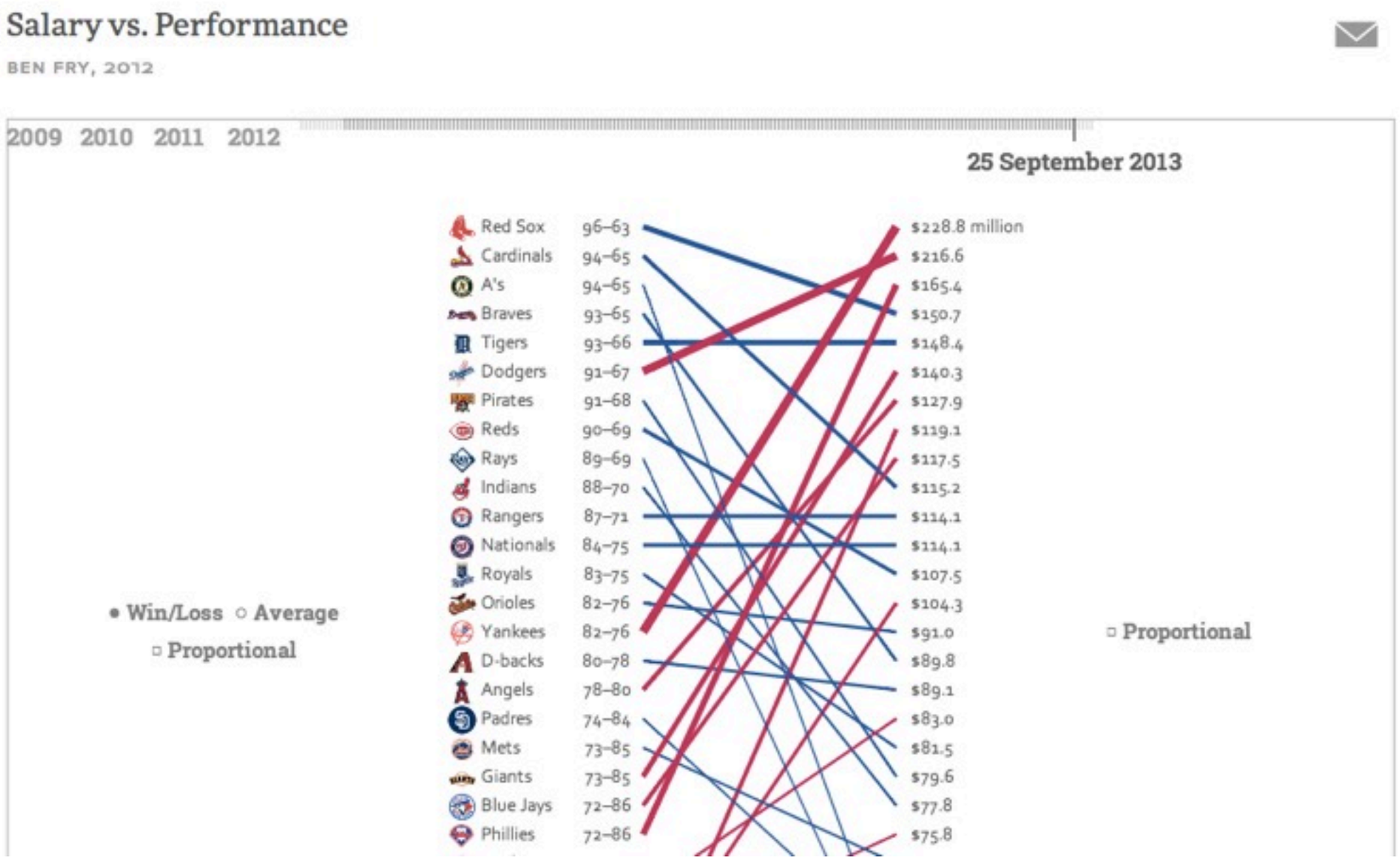
D3

StratomeX



A. Lex

Bump Charts / Slope Graphs

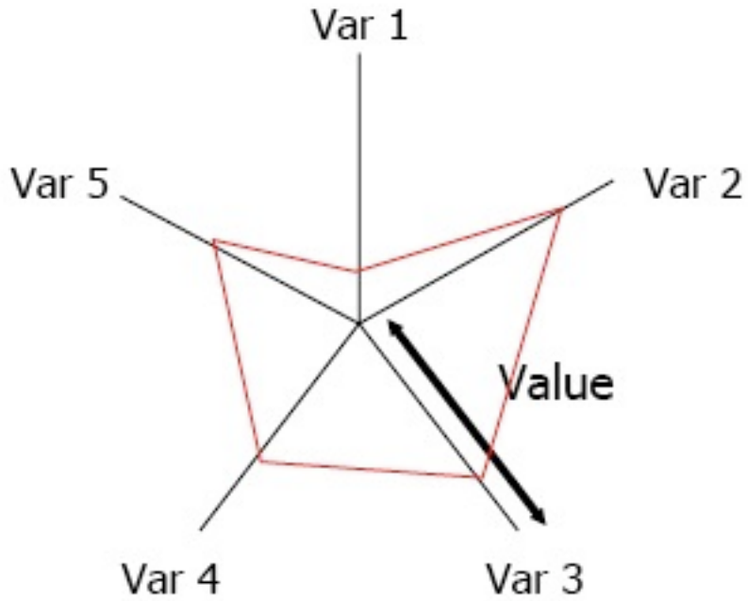


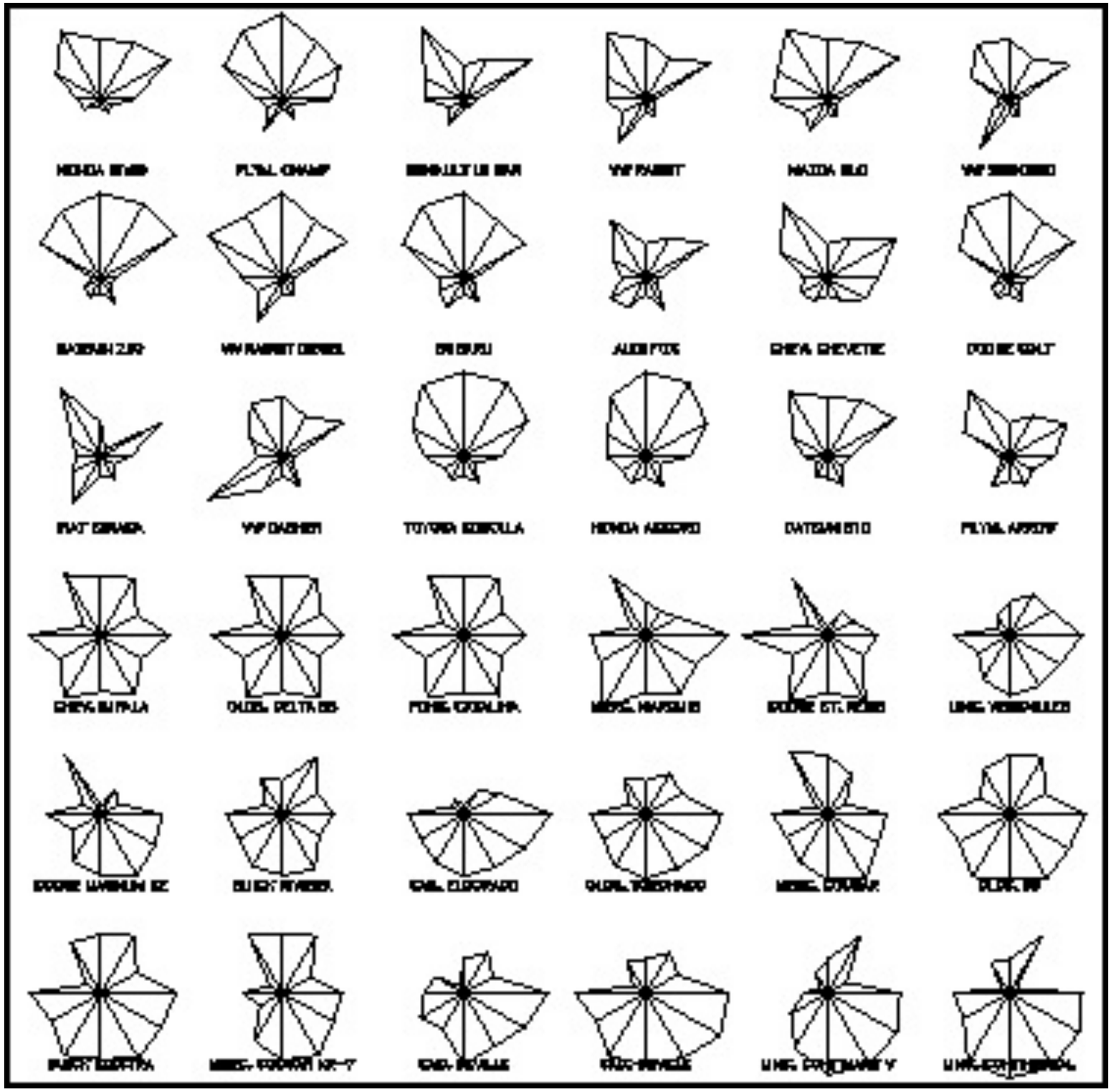
Ben Fry

Glyphs

Star Plots

- Space variables around a circle
- Encode values on “spokes”
- Data point is now a shape





Motor Trend Cars

Mazda RX4

Mazda RX4 Wag

Datsun 710

Hornet 4 Drive

Hornet Sportabout

Valiant

Duster 360

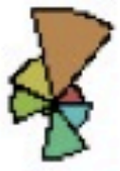
Merc 240D

Merc 230

Merc 280

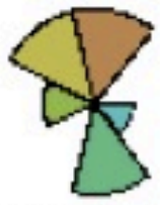
Merc 280C

Merc 450SE



Merc 450SL

Merc 450SLC

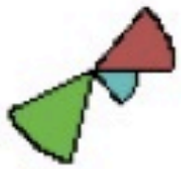


Cadillac Fleetwood

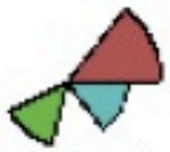
Lincoln Continental

Chrysler Imperial

Fiat 128



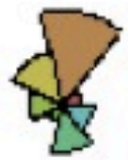
Honda Civic



Toyota Corolla

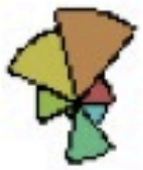
Toyota Corona

Dodge Challenger



AMC Javelin

Camaro Z28



Pontiac Firebird

Fiat X1-9

Porsche 914-2

Lotus Europa

Ford Pantera L

Ferrari Dino

Maserati Bora

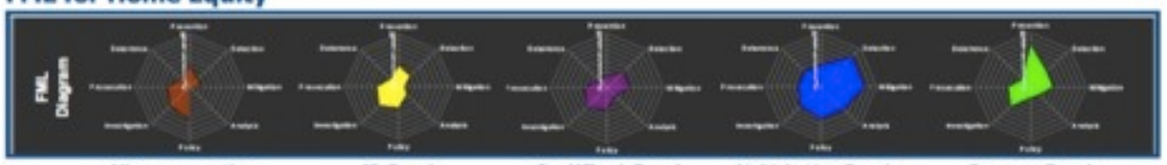
Volvo 142E

Home Equity Assessment & Recommendation

In October, an assessment of current fraud management activities within the Home Equity portfolio was completed. To perform this assessment, we leveraged the **Fraud Management Lifecycle (FML) Theory**[1], which emphasizes the importance of strong and balanced prevention, detection, mitigation, analysis, policy, investigation, prosecution, and deterrence activities for effective fraud management.

By fraud type, we looked at current processes and weaknesses for each stage of the lifecycle and rated these to form the "spider web" graphs shown below. In these graphs, greater coverage across the web indicates stronger overall fraud management for the given fraud type.

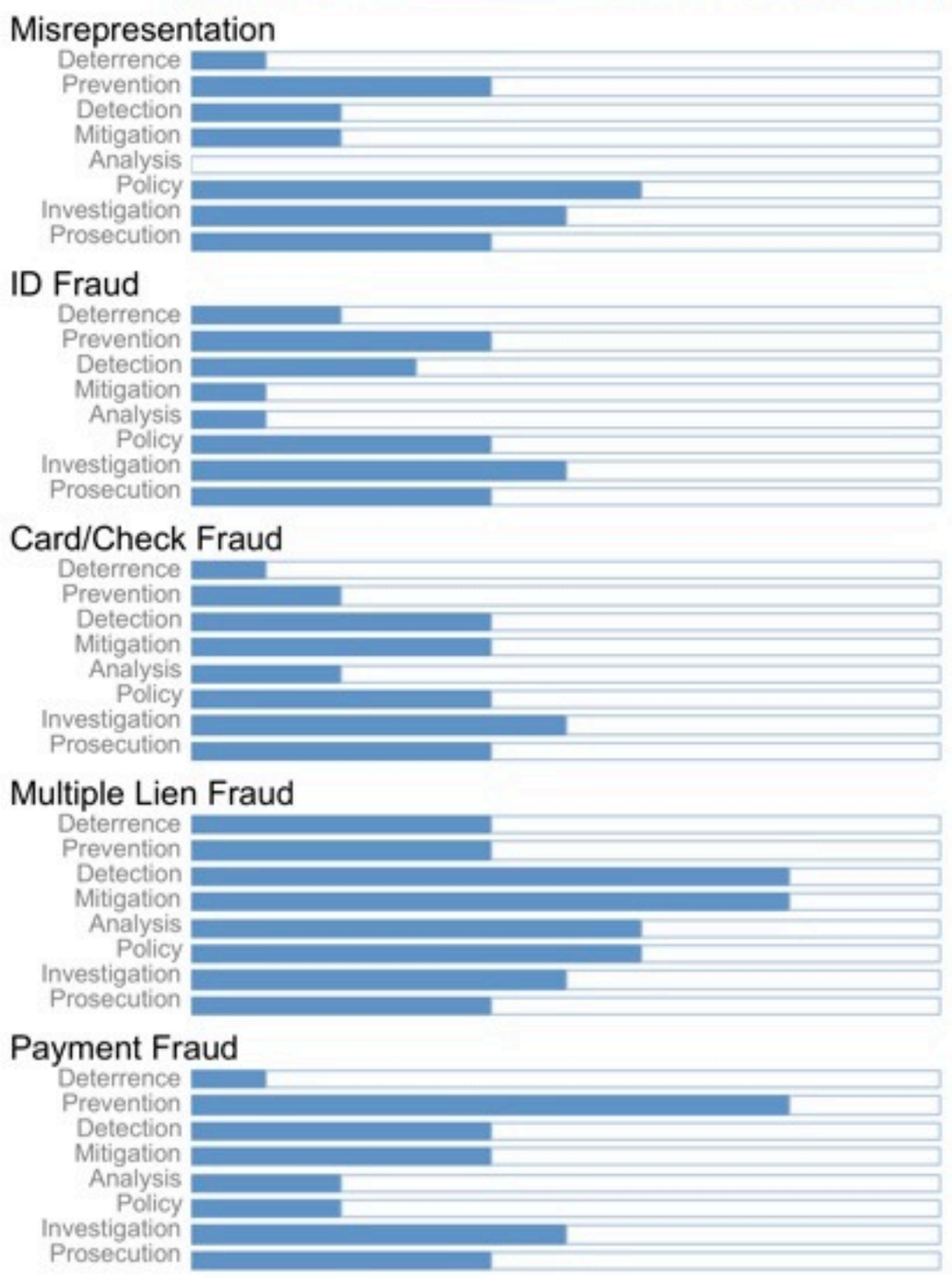
FML for Home Equity



We are far from ideal coverage across all fraud types… …something must change; recommendations follow

Coverage by Fraud Type & Fraud Management Lifecycle Stage

|nothing in place ---------- we've solved every problem|



C. Nussbaumer

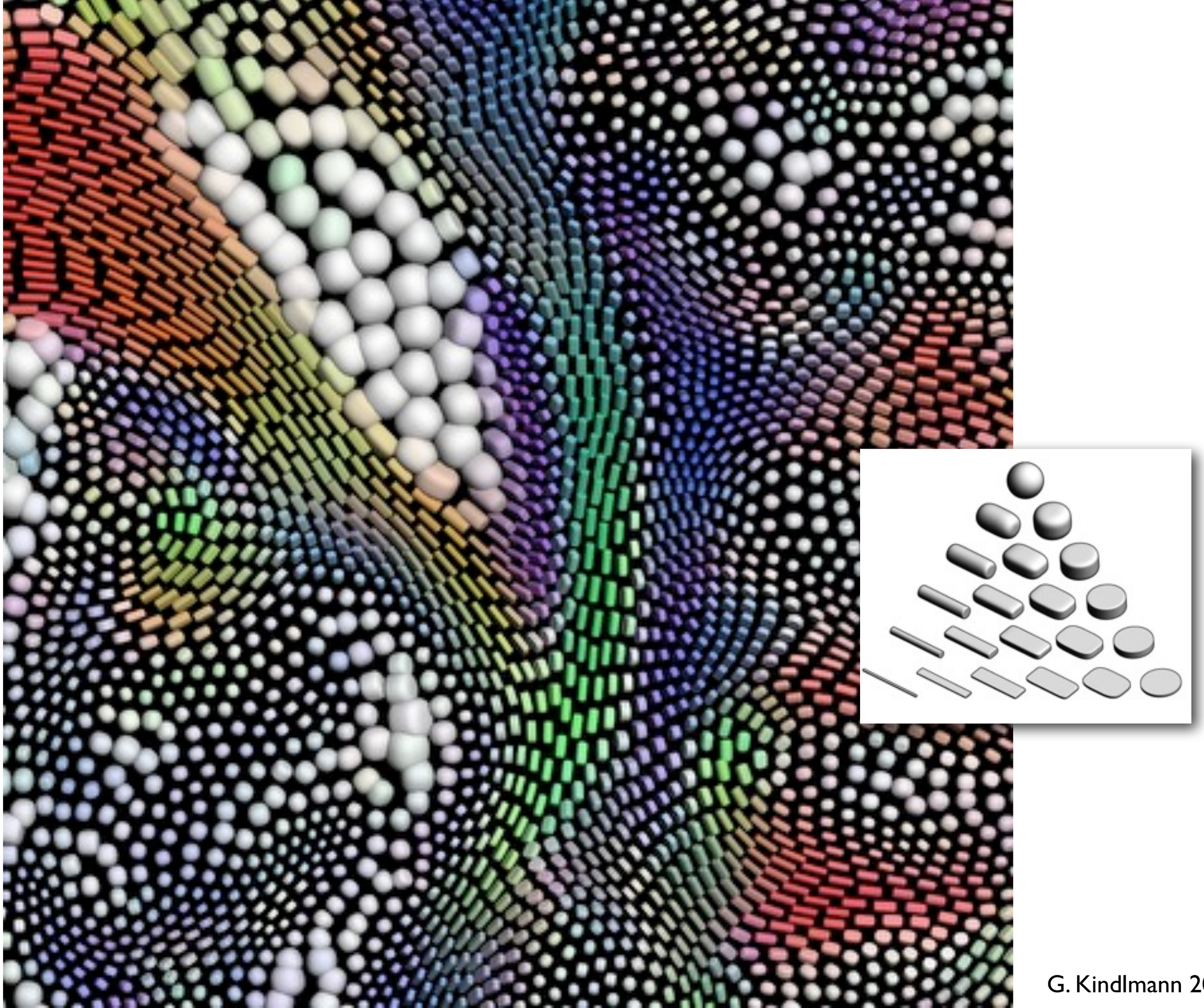
Velocity
(magnitude &
direction)

Vorticity
(scalar, CW/CCW)

Strain Tensor
(second order)

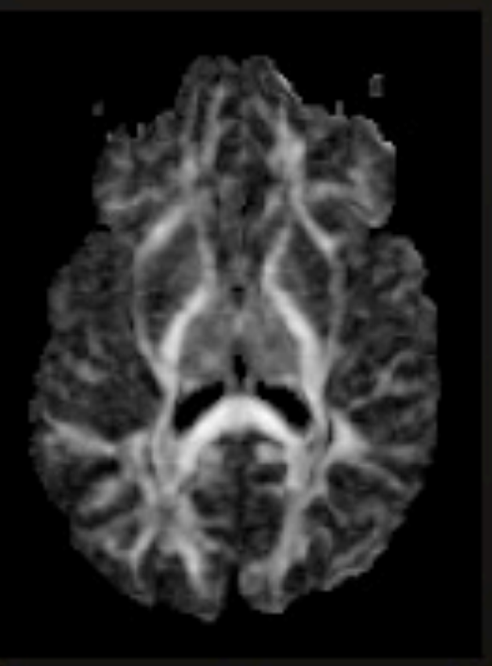
Turbulent Charge
(vector & scalar)

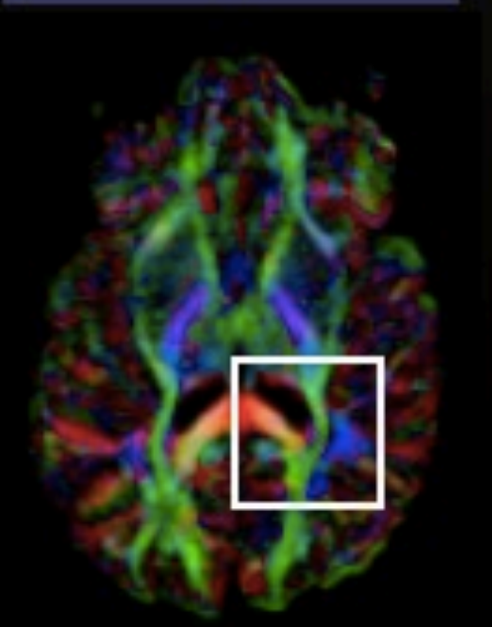
M. Kirby, H. Marmanis, and D. Laidlaw



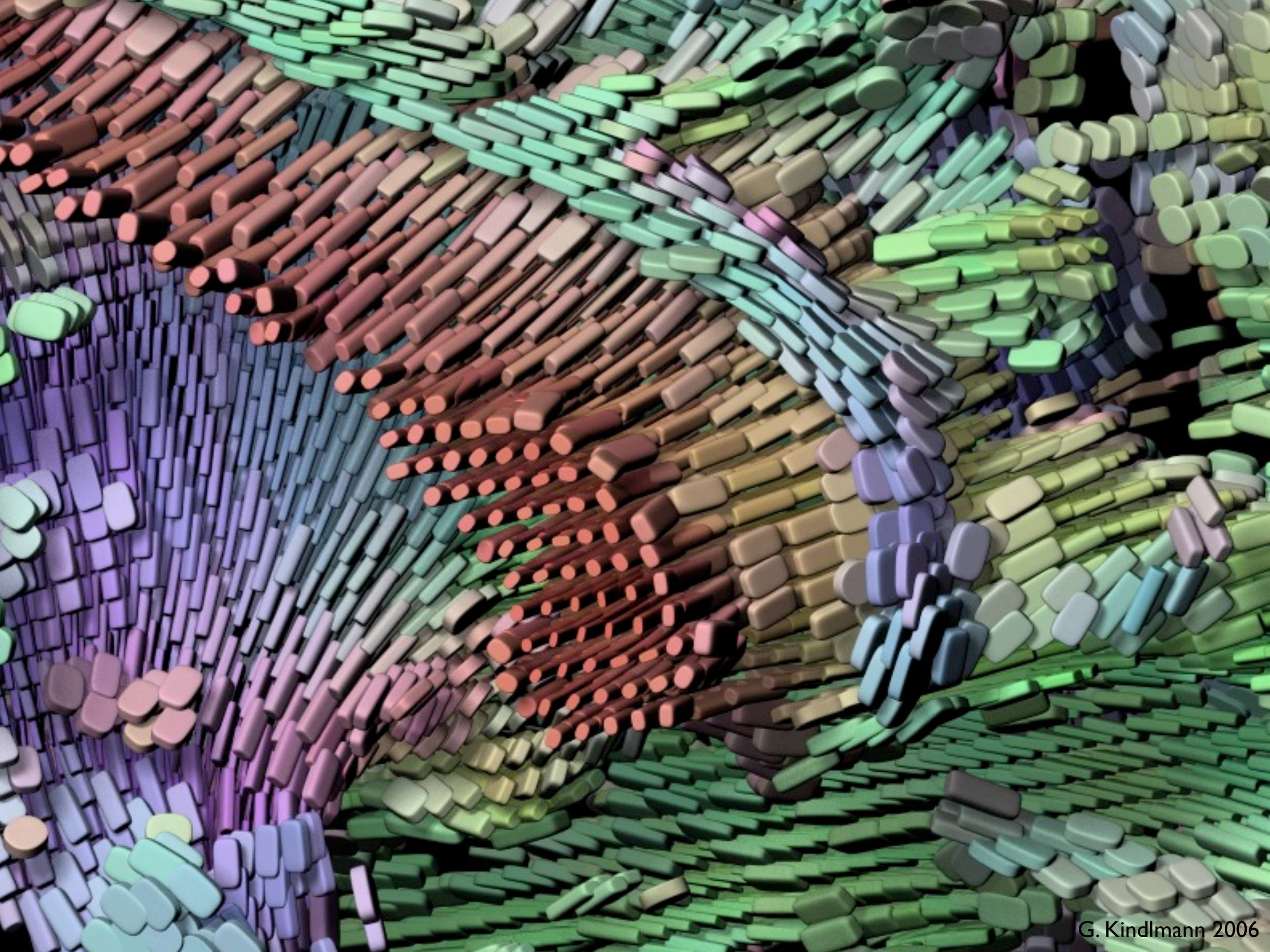
G. Kindlmann 2006

Results





G. Kindlmann 2006


G. Kindlmann 2006

Dimensionality Reduction

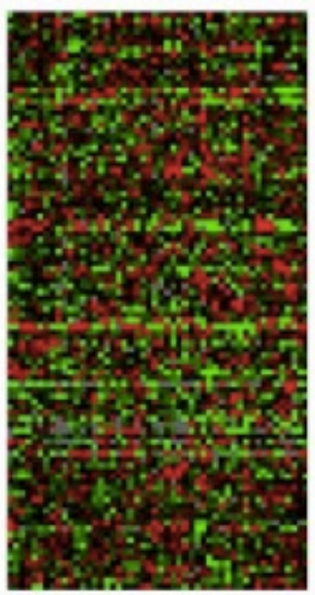
What about very high-dimensional data?

face images

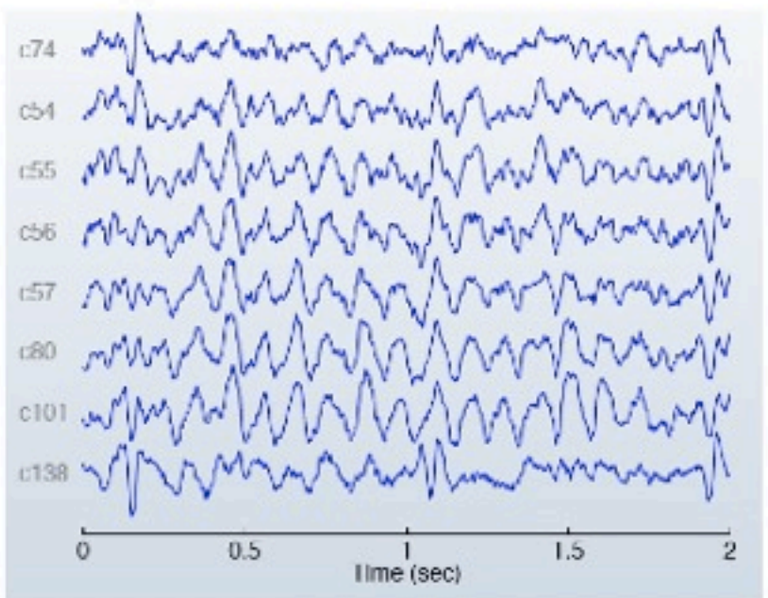
Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

documents



gene expression data

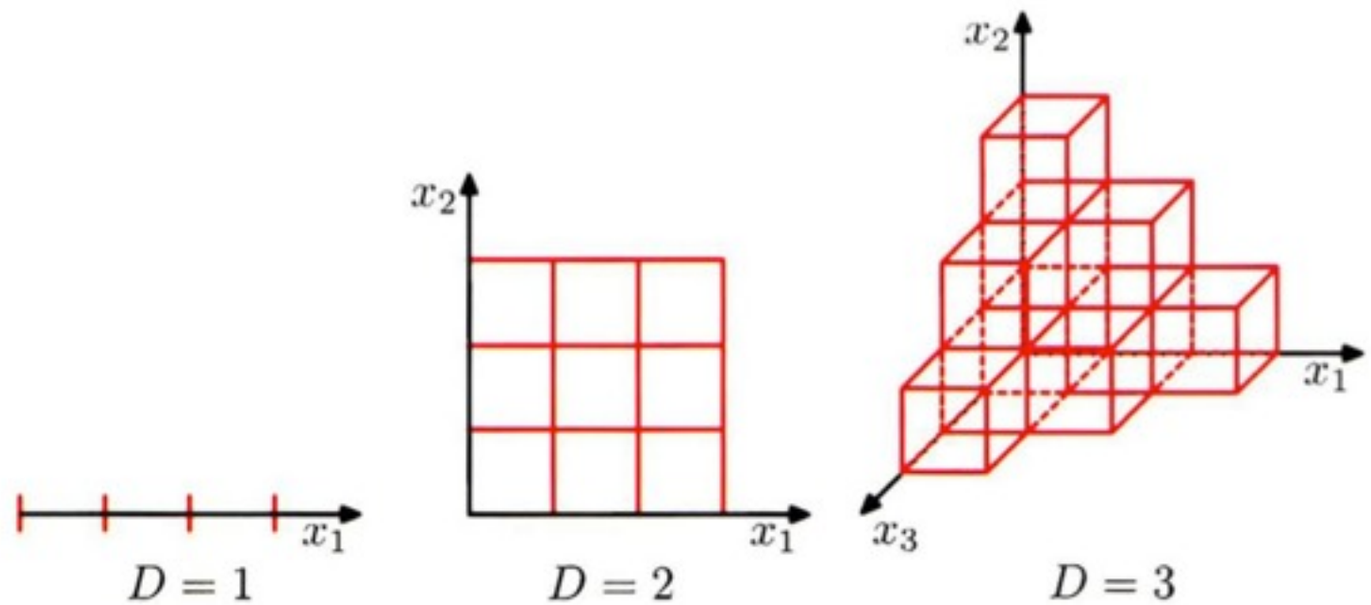


MEG readings

Based on slide from P. Liang

Curse of Dimensionality

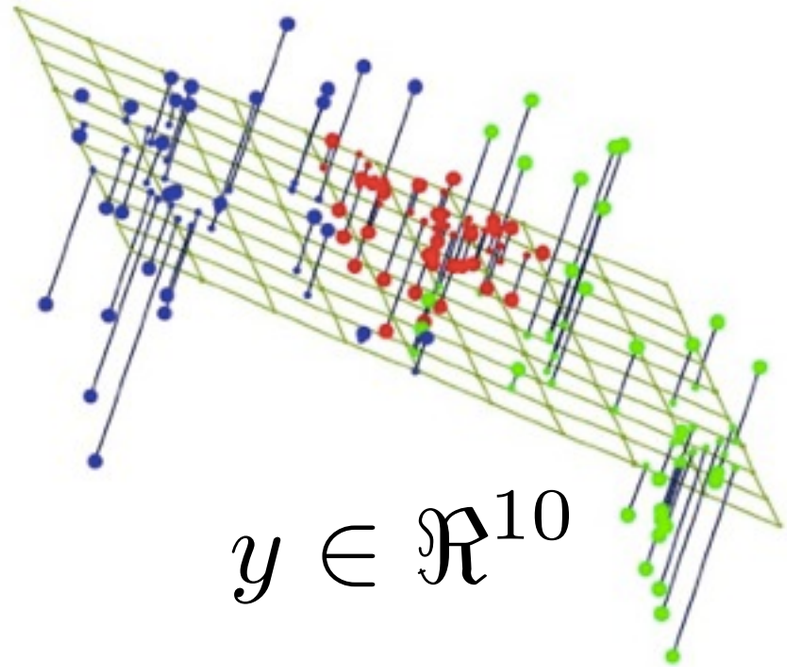
- When dimensionality increases, the volume of the space increases so fast that the available data becomes sparse
- Statistically sound result requires the sample size N to grow exponentially with d



Basic Idea

Project the high-dimensional data onto a lower-dimensional subspace using linear or non-linear transformations

$$x \in \Re^{64 \times 64} = \Re^{4096}$$



$$y \in \Re^{10}$$

$$y = Ux$$

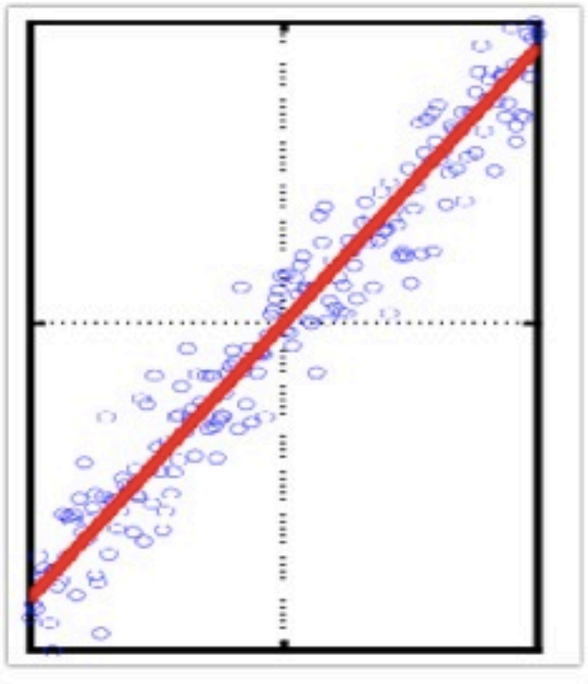
Based on slide from P. Liang

Linear Methods

- Does the data lie mostly in a hyperplane?
- If so, what is its dimensionality?

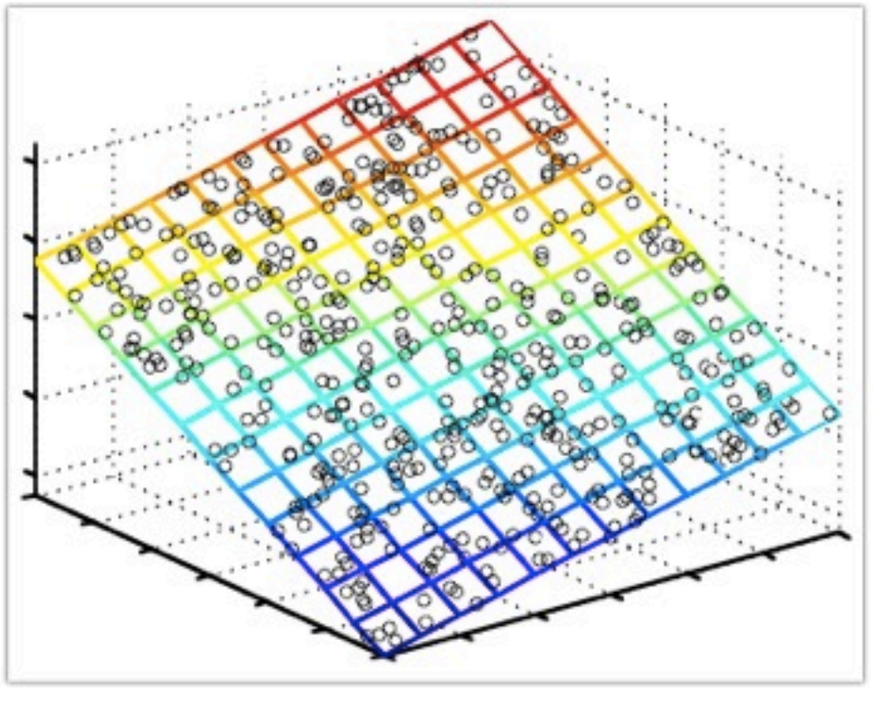
D = 2

d = 1



D = 3

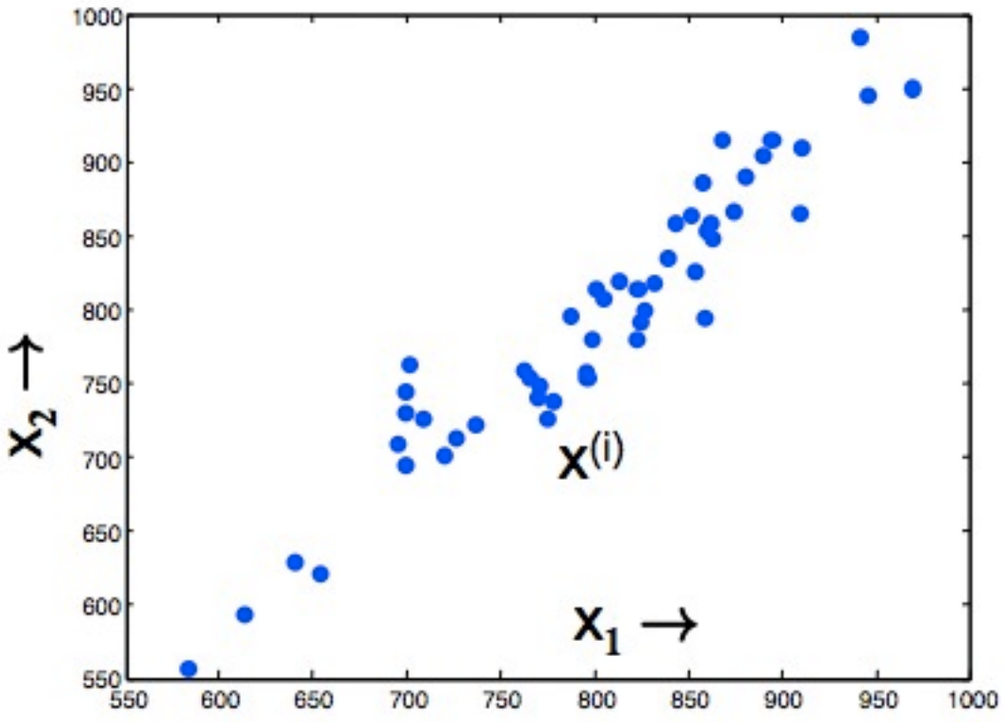
d = 2



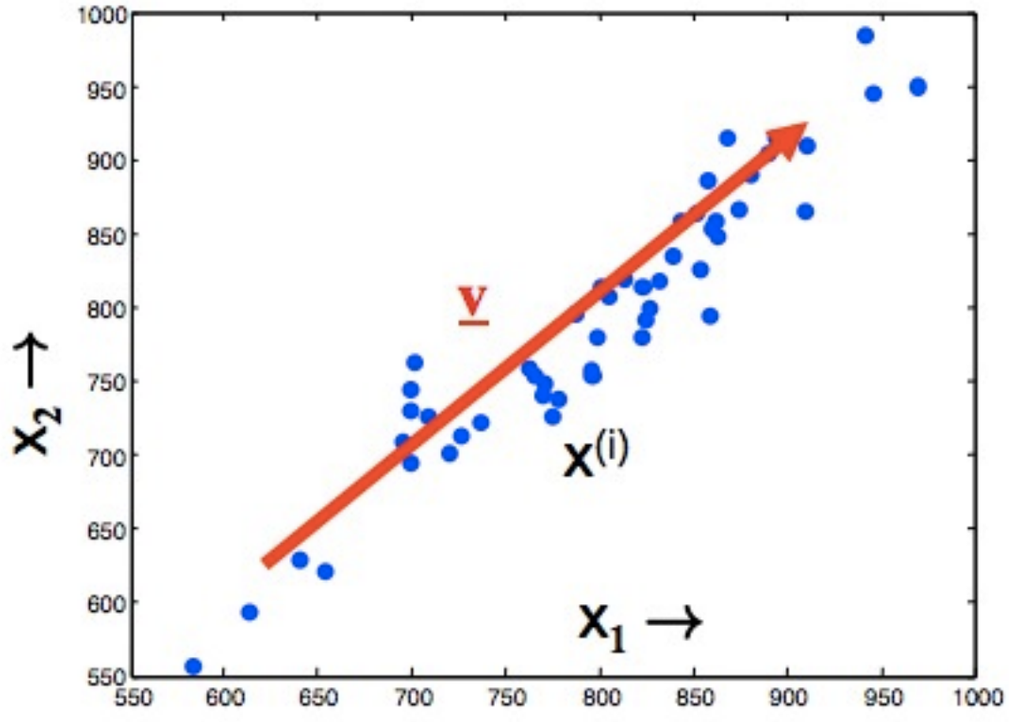
Based on slide from F. Sha

Principal Components Analysis (PCA)

Example



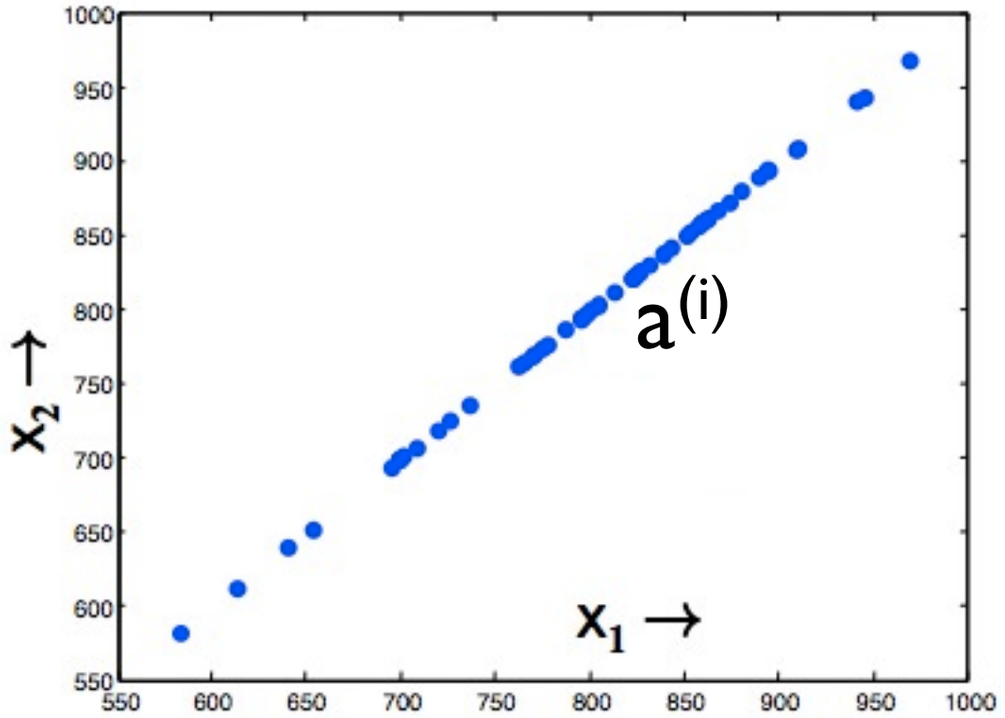
$$\vec{x} = [x_1, x_2]$$



$$\vec{x} \approx s\vec{v} = s[v_1, v_2]$$

Based on slide from A. Ihler

Example



$a^{(i)}$: Projection of $x^{(i)}$ onto v

v: chosen to minimize residual variance

$$\min_{a,v} \sum_i (x^{(i)} - a^{(i)} v)^2$$

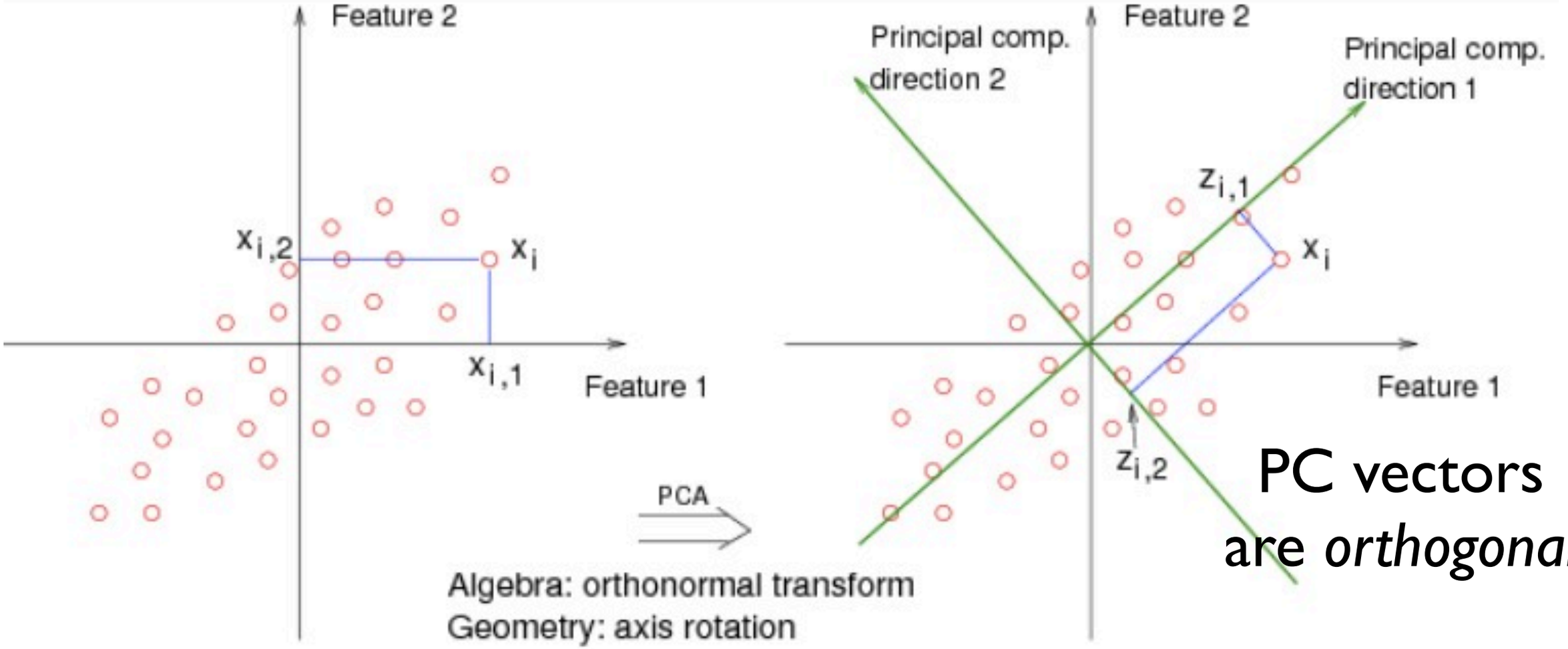
Find v that most closely reconstructs x

Equivalent: v is the direction of maximum variance

Based on slide from A. Ihler

PCA

Project data to a subspace such as to maximize the variance of the projected data

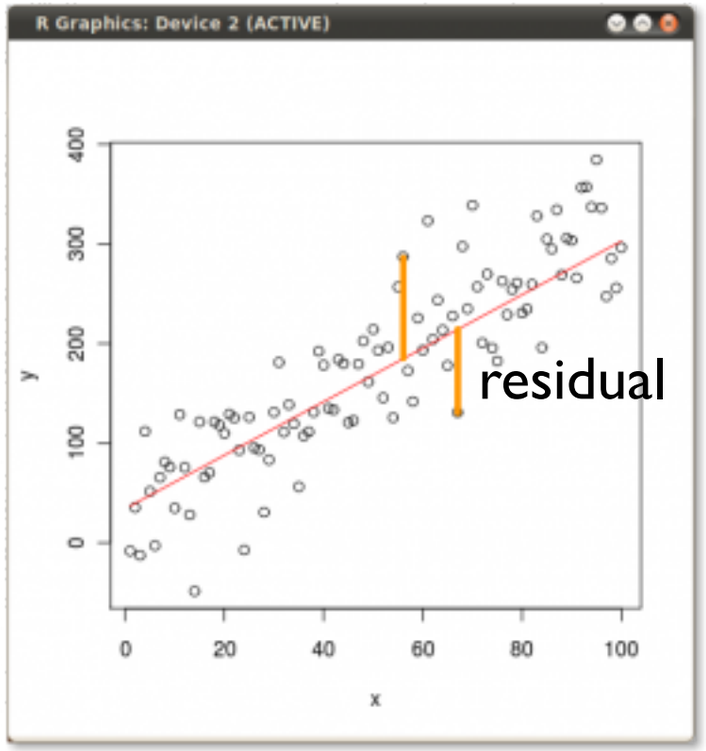


PC vectors are *orthogonal*

Based on slide from J. Li

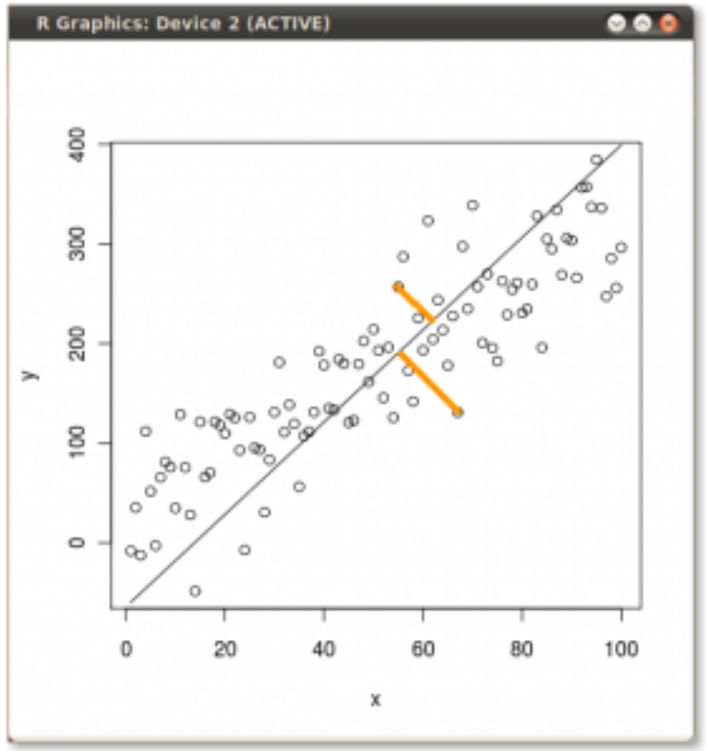
Linear Regression vs. PCA

Linear Regression



Projection along dimensions of X

PCA



Projection along dimensions of PCs

http://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

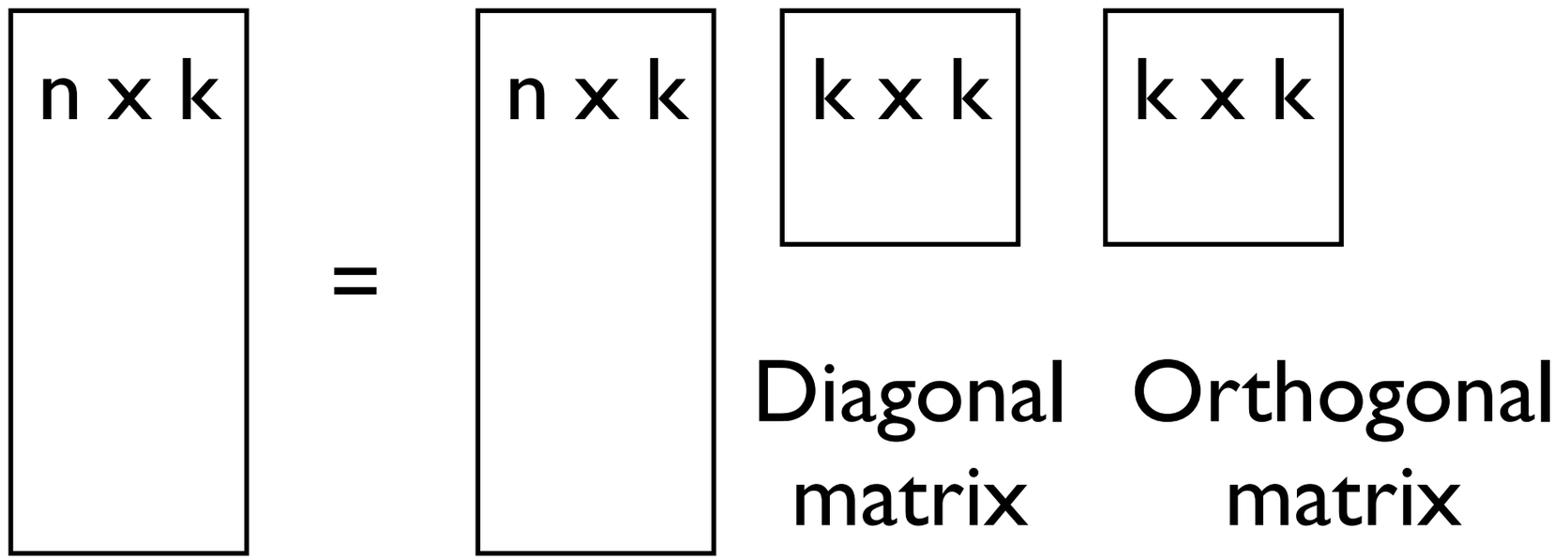
PCA Algorithm

- Subtract mean from data (center X)
- (Typically) scale each dimension by its variance
  - Helps to pay less attention to magnitude of dimensions
- Compute covariance matrix S $\mathbf{S} = \frac{1}{N}\mathbf{X}^\intercal\mathbf{X}$
- Compute k largest eigenvectors of S
  - Computing covariance matrix S may lead to loss of precision

Based on slide from A. Ihler

Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$$



Based on slide from A. Ihler

Singular Value Decomposition (SVD)

$\mathbf{S} = \frac{1}{N}\mathbf{X}^\intercal\mathbf{X}$ — Sample covariance

$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$ — SVD

$\mathbf{X}^\intercal\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\intercal$ — Eigendecomposition of S (up to scale factor 1/N)

v_1 — First principal component

$d_1 \geq d_2 \geq d_3 ... \geq d_p \geq 0$ — Singular values of X = sqrt(eigenvalues) of S

Based on slide from A. Ihler

SVD Properties

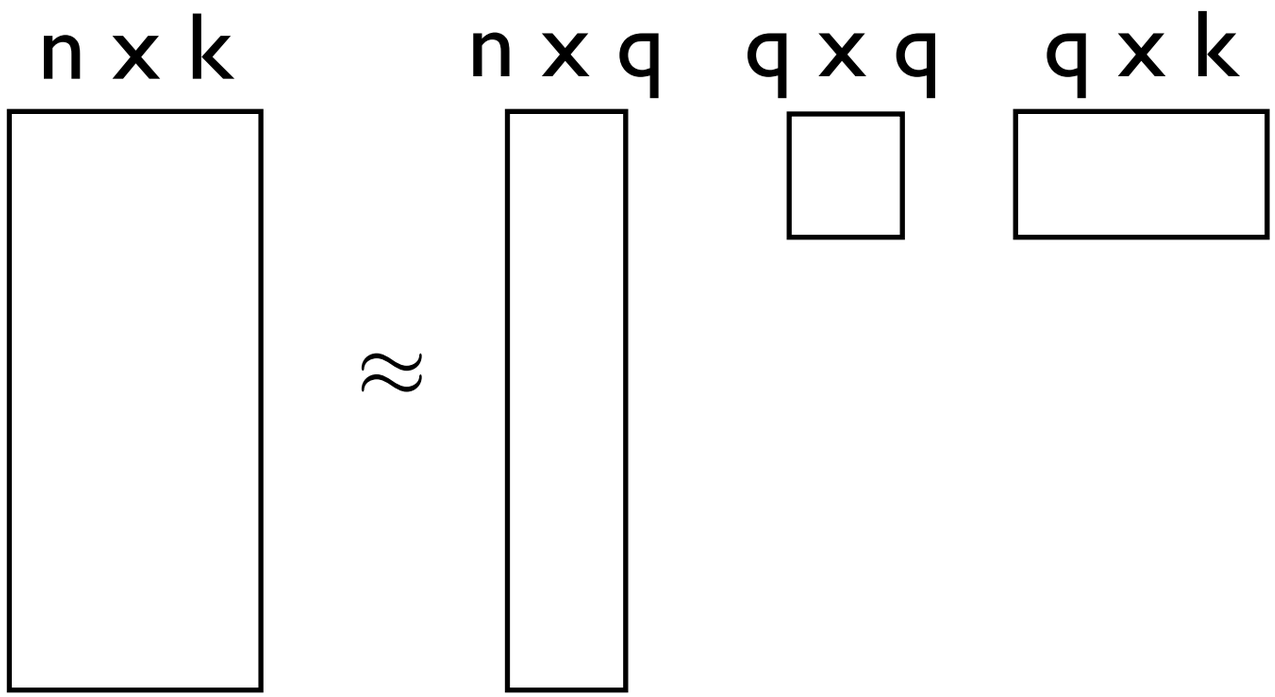
- Works for *any* matrix
- Non-zero singular values of D are square roots of non-singular eigenvalues of S
- Columns of V are the eigenvectors of X^TX
  - And columns of U are the eigenvectors of XX^T
- Used to compute the pseudo-inverse X^+ of X

$$\mathbf{X}^{+} = \mathbf{V}\mathbf{D}^{+}\mathbf{U}^{\intercal}$$

  - Compute D^+ by replacing each non-zero d_i with $1/d_i$

Dimensionality Reduction

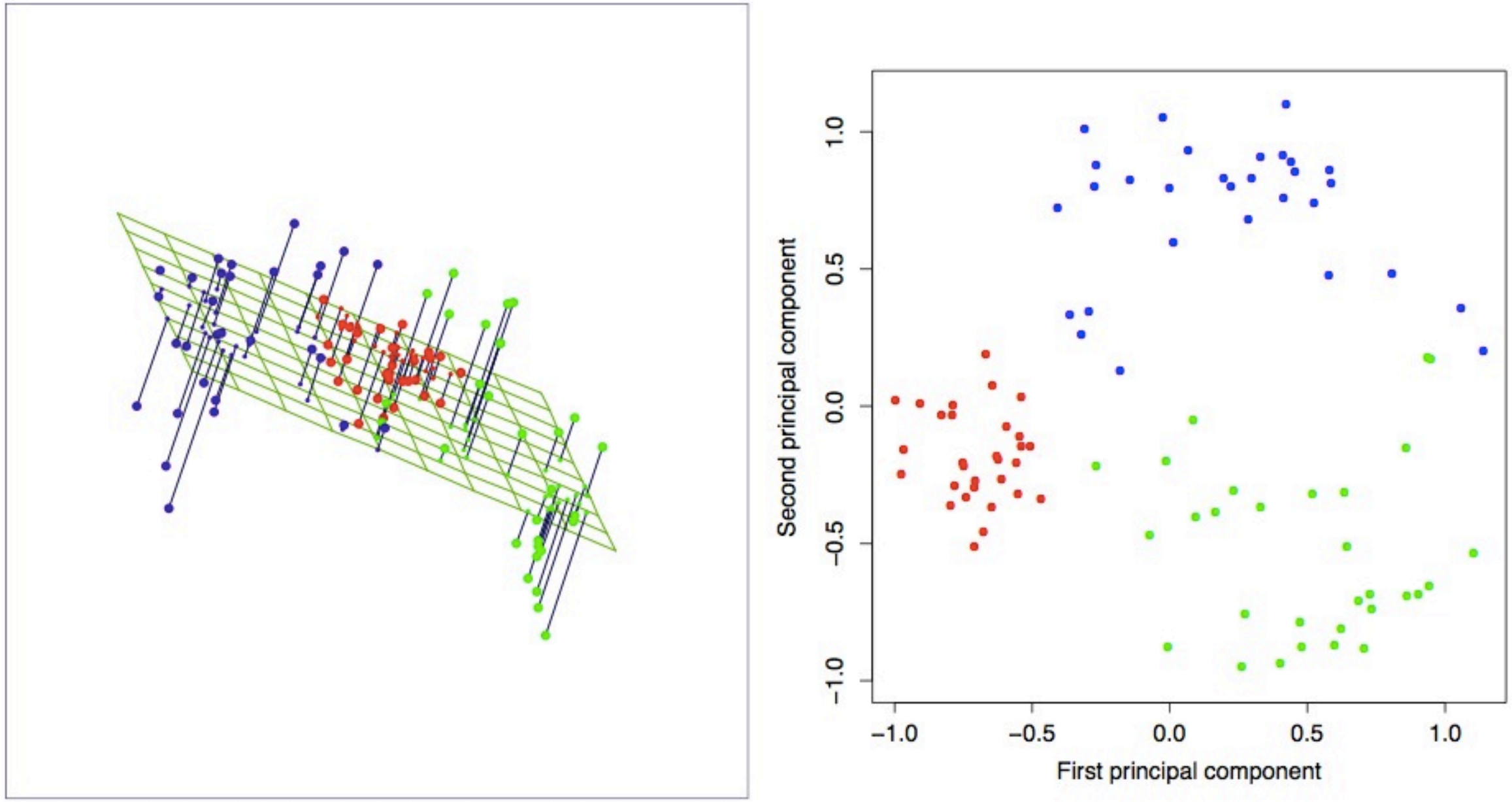
$$\mathbf{X} \approx \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\intercal$$



$$x_i \approx u_{i,1}d_{1,1}\mathbf{v}_1 + u_{i,2}d_{2,2}\mathbf{v}_2 + ... + u_{i,q}d_{q,q}\mathbf{v}_q$$

Based on slide from A. Ihler

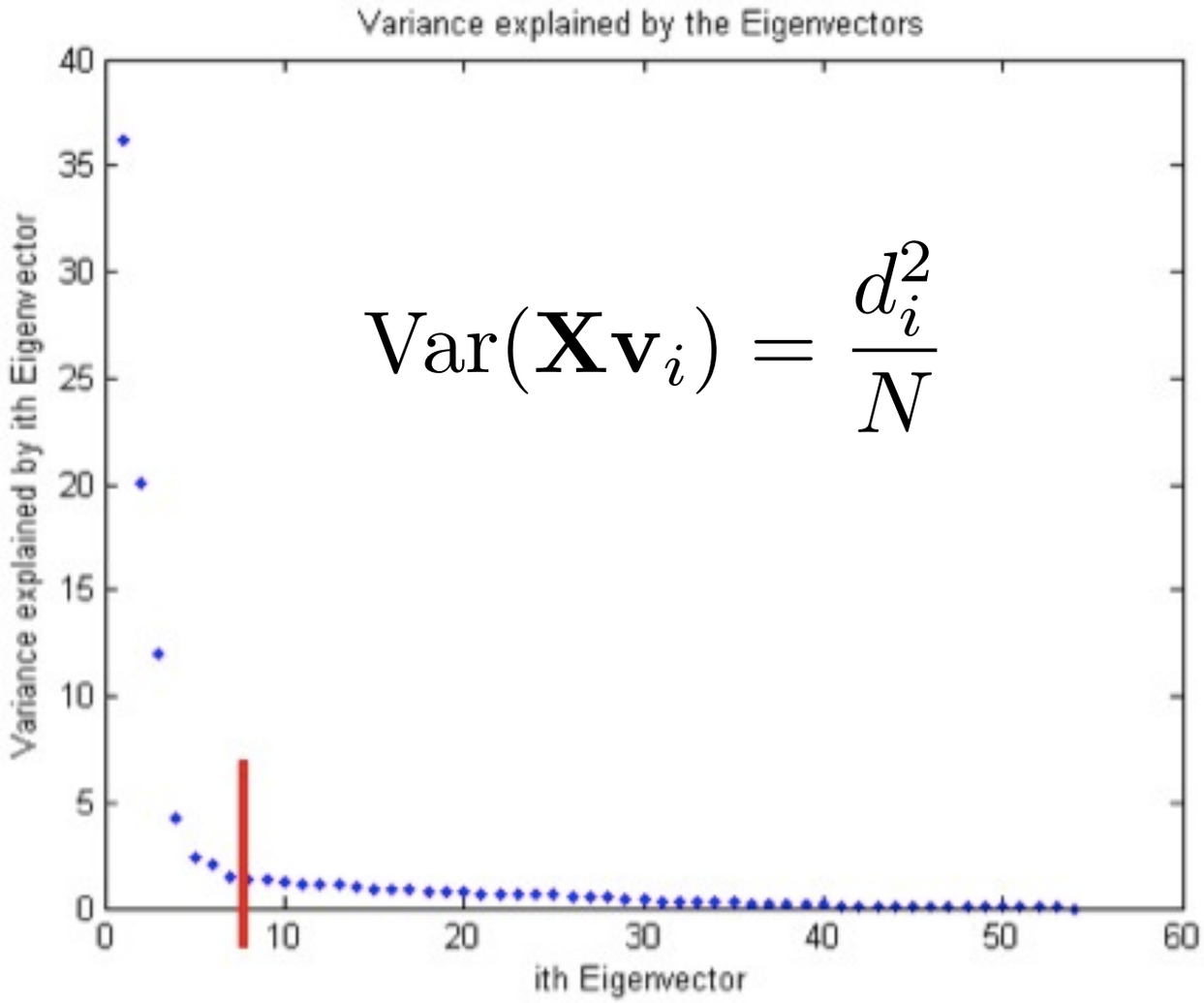
Dimensionality Reduction



Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

How many PC vectors?

Enough PC vectors to cover 80-90% of the variance



$$\mathrm{Var}(\mathbf{X}\mathbf{v}_i) = \frac{d_i^2}{N}$$

Screeplot

Based on slide from J. Leskovec

Issue: Data Scaling

- PCA on the raw data
- PCA on sphered data (each dimension has mean 0, variance 1)
- PCA on 0-to-1 normalized data (each dimension is squished to be between 0 and 1)
- PCA on *whitened* data (a rotation & scaling that results in identity covariance)

http://blog.explainmydata.com/2012/07/should-you-apply-pca-to-your-data.html

PCA for Handwritten Digits

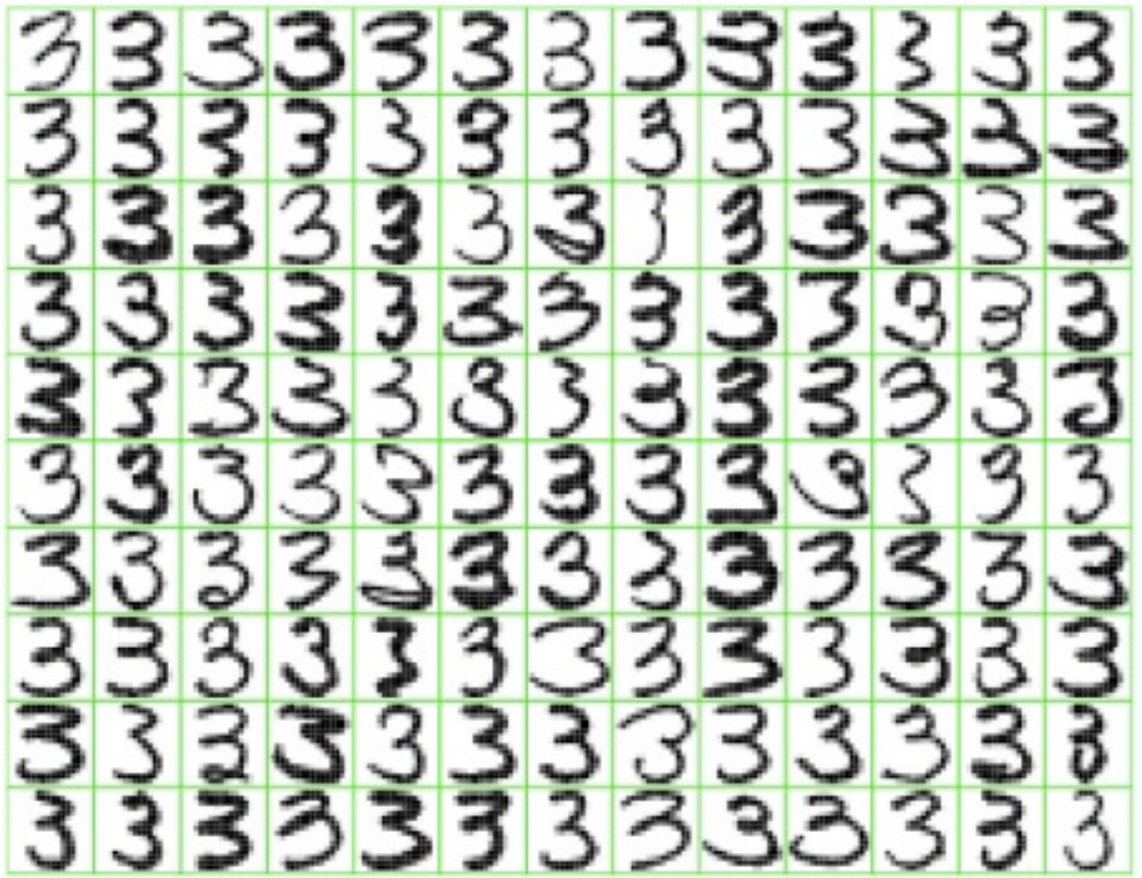
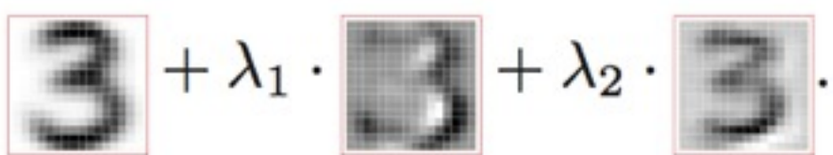
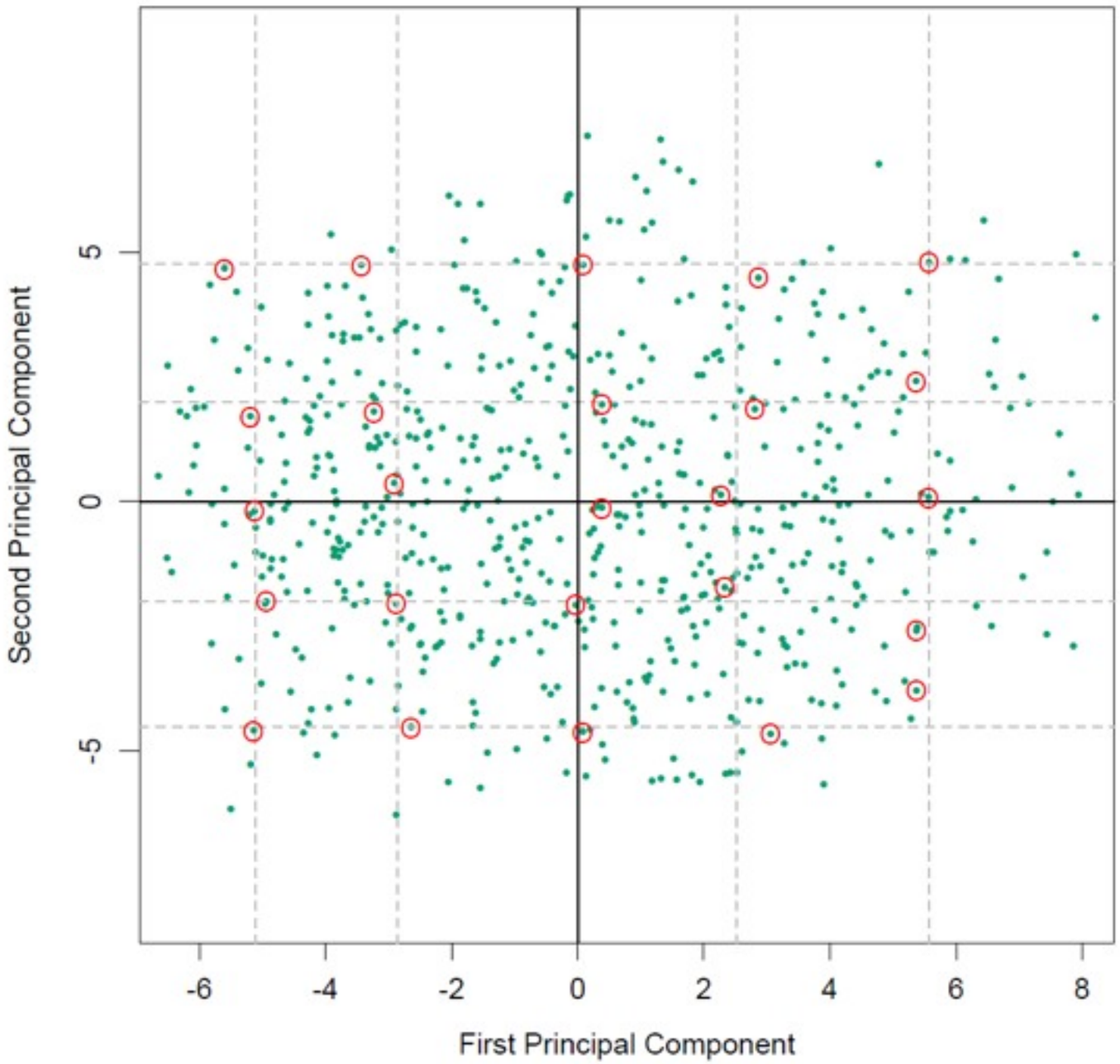


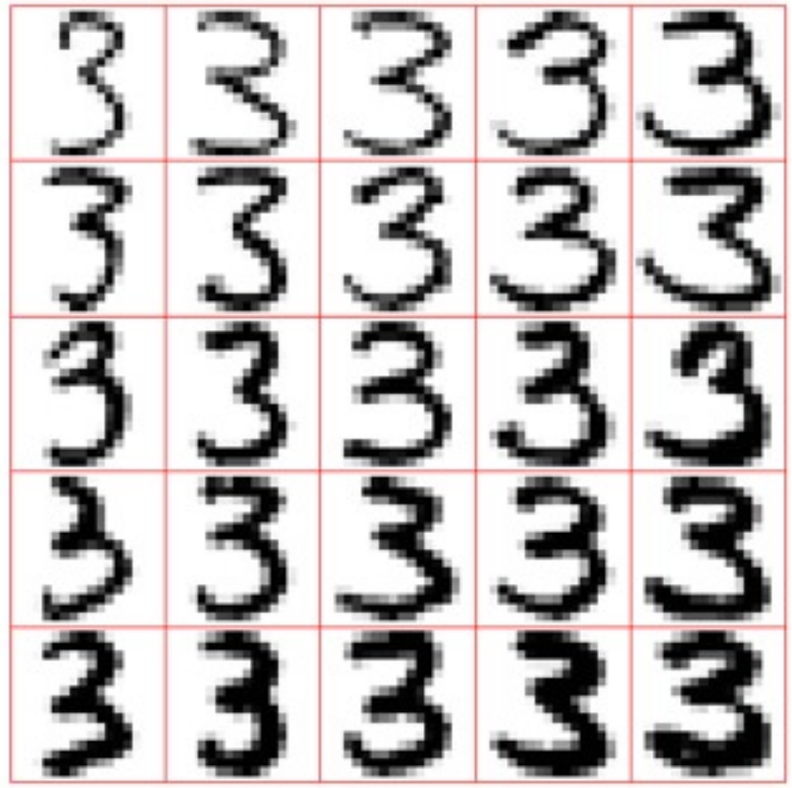
FIGURE 14.22. *A sample of* 130 *handwritten 3's shows a variety of writing styles.*

$$\begin{aligned}\hat{f}(\lambda) &= \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \\ &= \end{aligned}$$ 

Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

PCA for Handwritten Digits





Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

PCA for Face Images

Based on slide from T. Yang

PCA for Face Images

- 64x64 images of faces = 4096 dimensional data

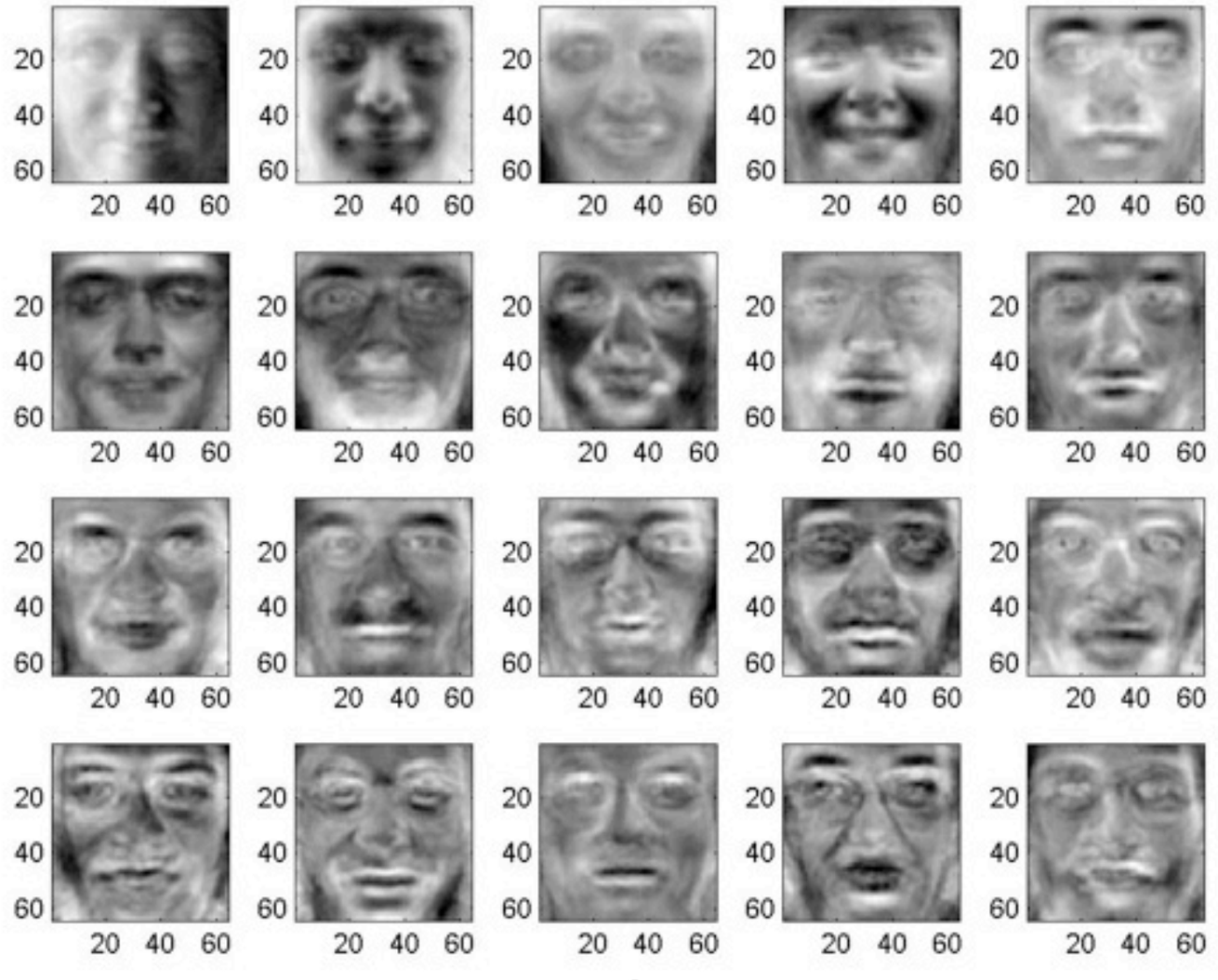
Based on slide from A. Ihler

Eigenfaces

- We can reconstruct each face as a linear combination of "basis" faces, or Eigenfaces [M. Turk and A. Pentland (1991)]

Average Face

+

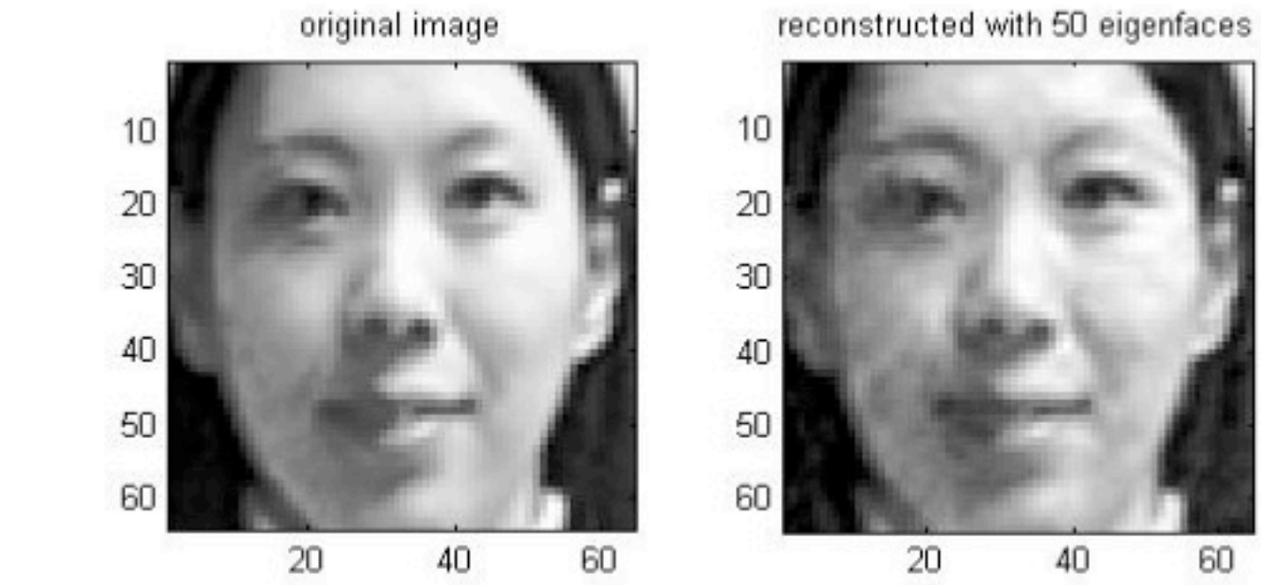


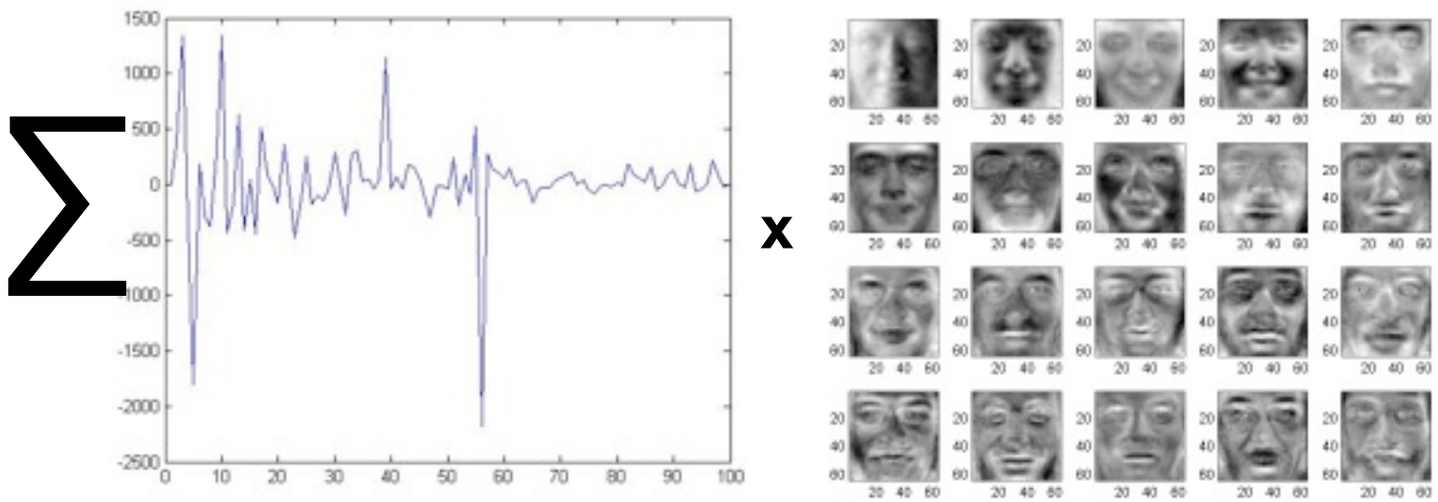
Eigenfaces

Based on slide from T. Yang

Reconstruction

- 90% variance is captured by the first 50 eigenvectors

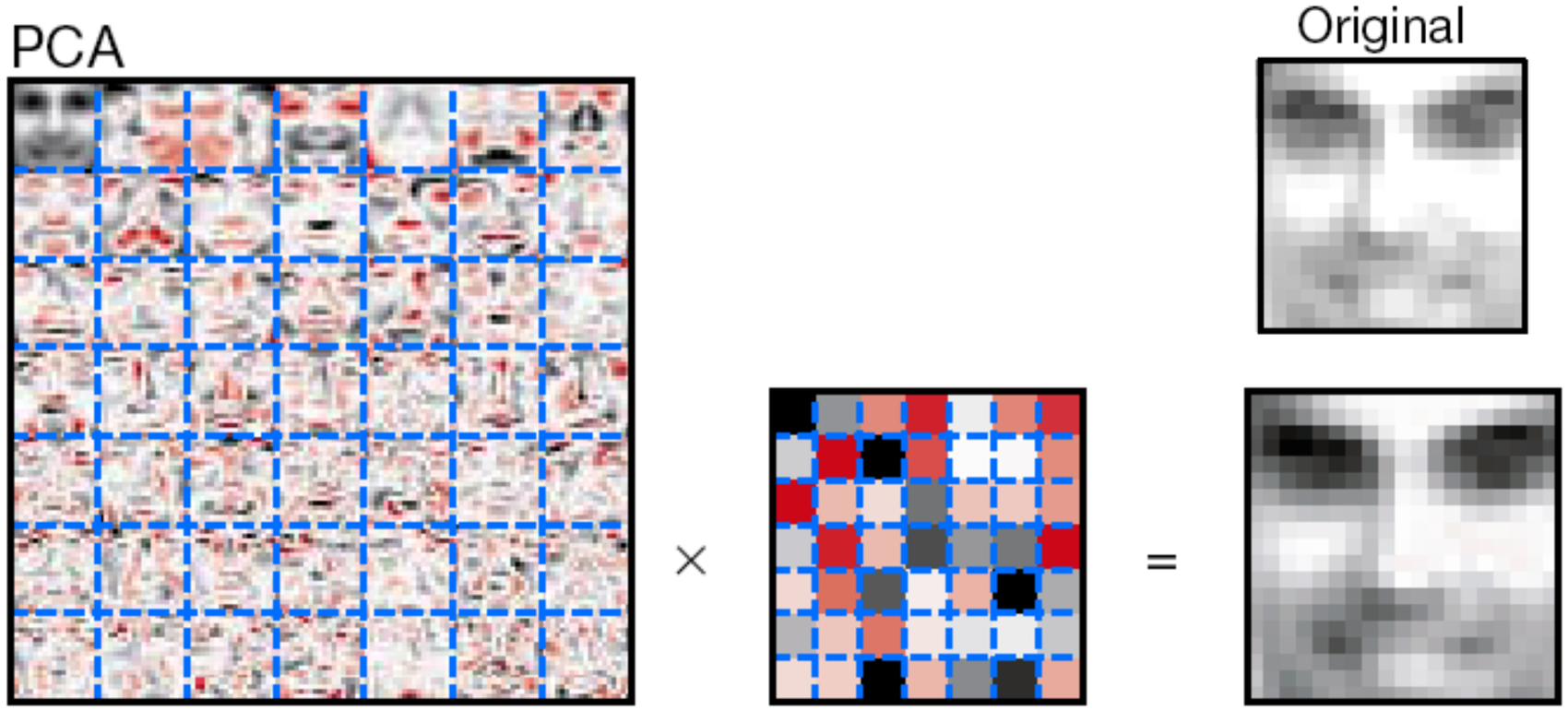




Based on slide from T. Yang

Issues

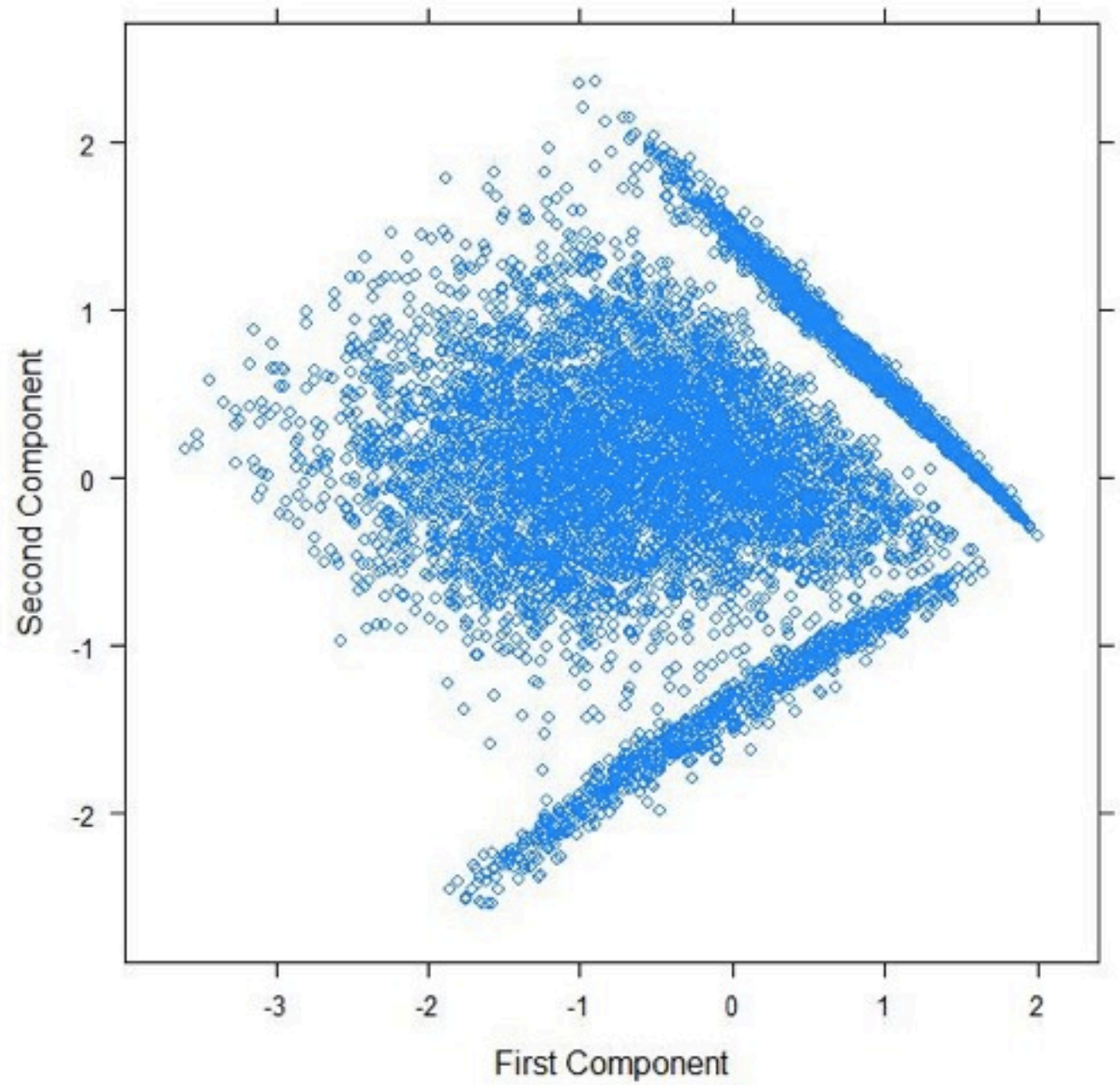
- PCA involves adding up some basis images and subtracting others
- The basis images are not physically intuitive



Based on slide from M. Tappen

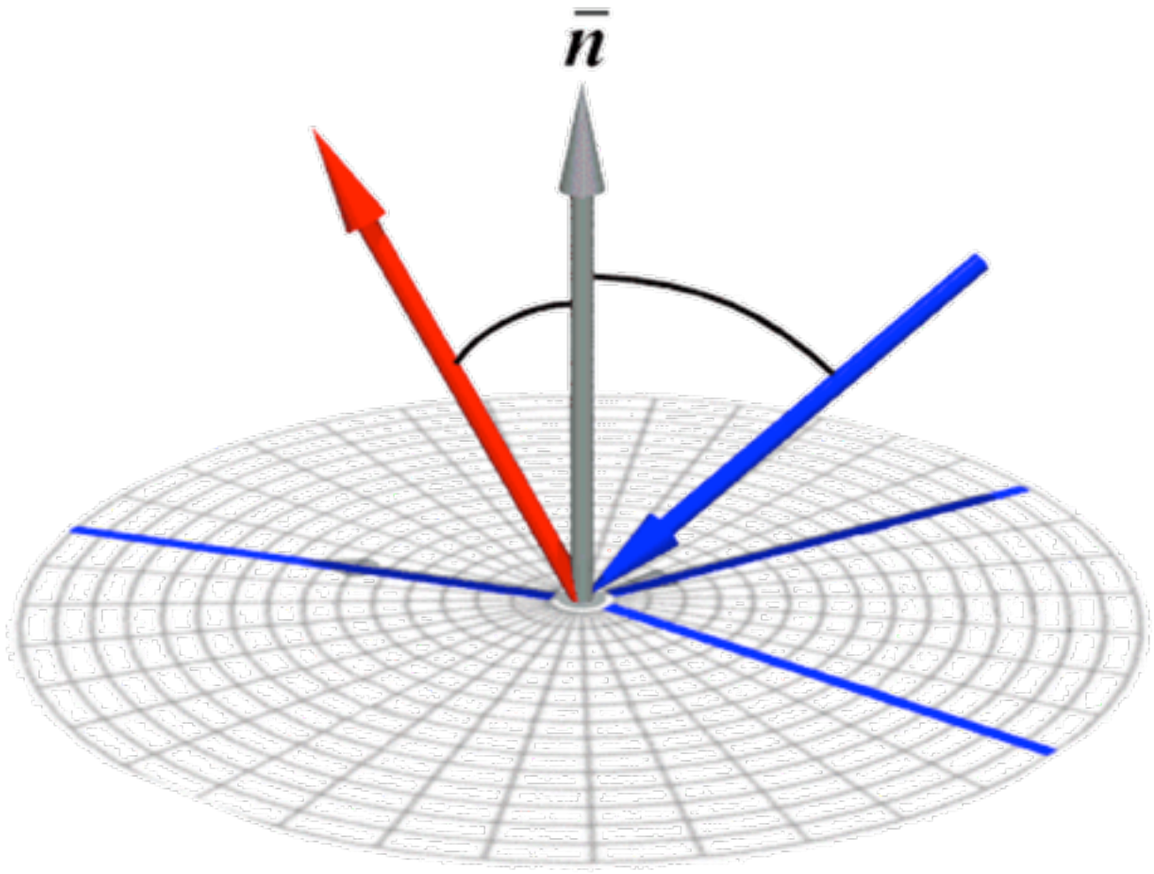
Text Documents

>45 features, projected onto two PC dimensions



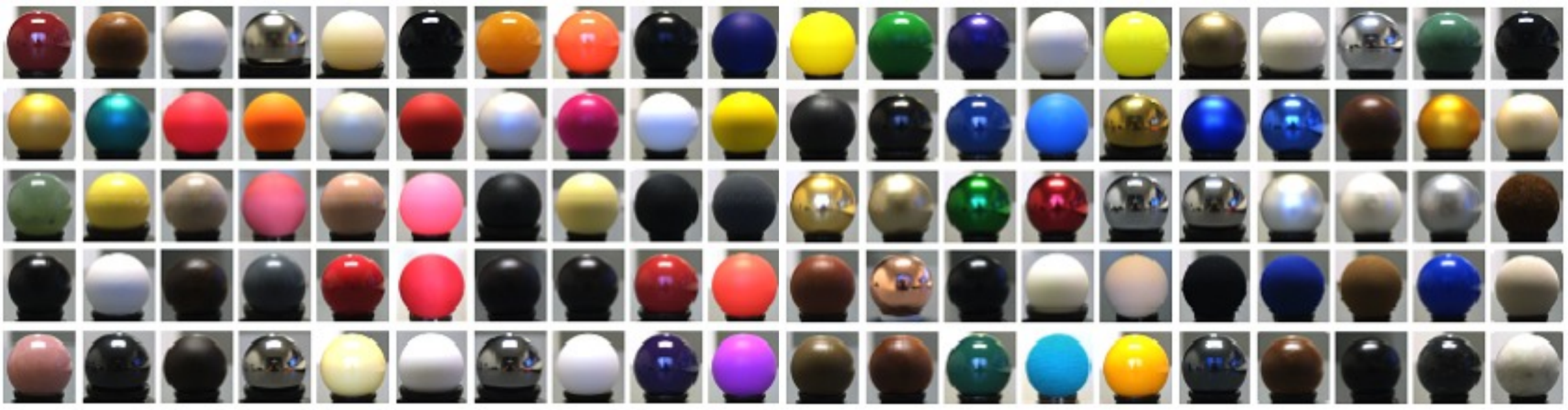
Data-Driven BRDFs

- Bi-Directional Reflectance Distribution Functions



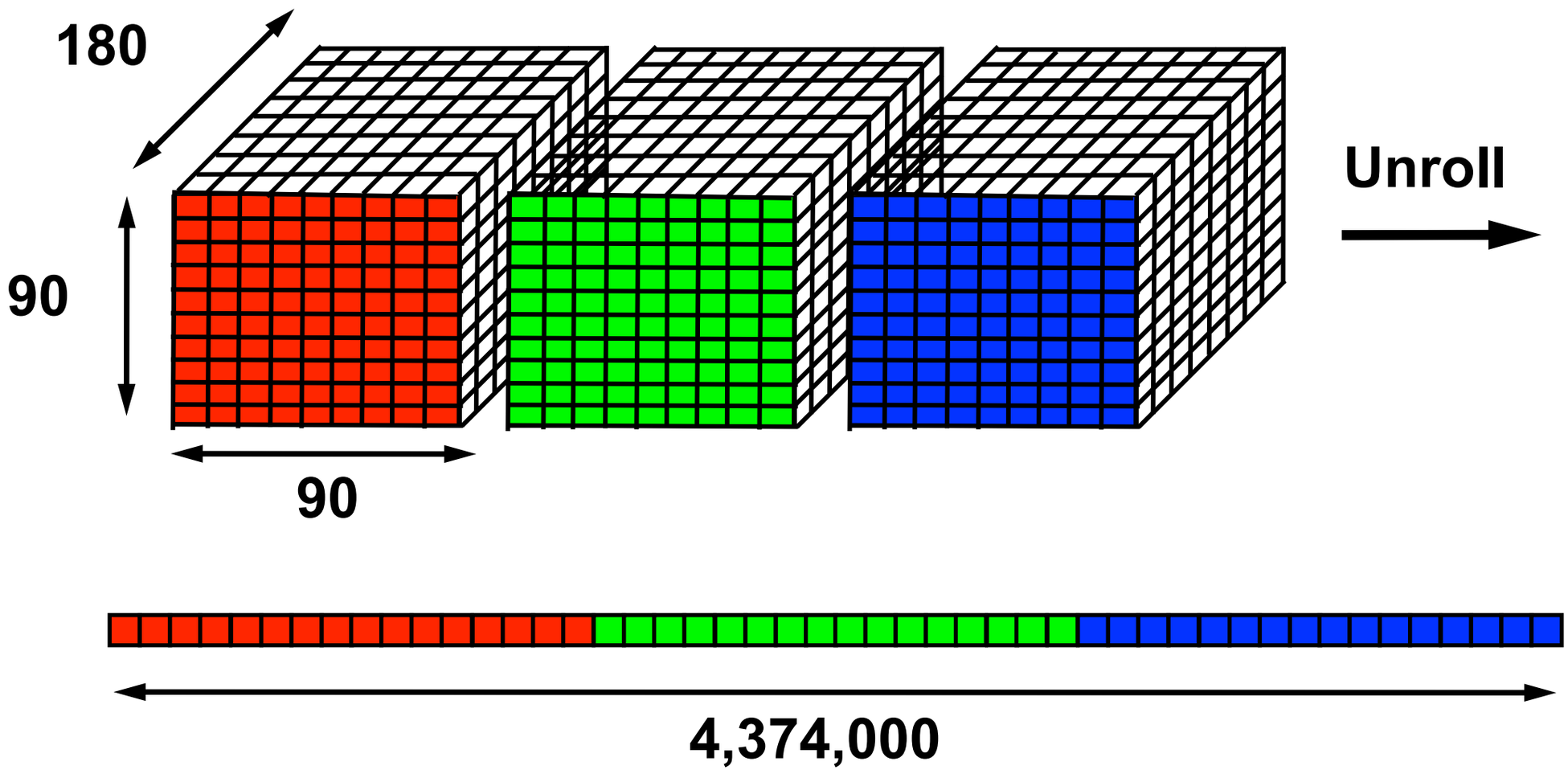
Data-Driven BRDFs

- Measure light reflected off a sphere
- 20-80 million measurements (6000 images) per material

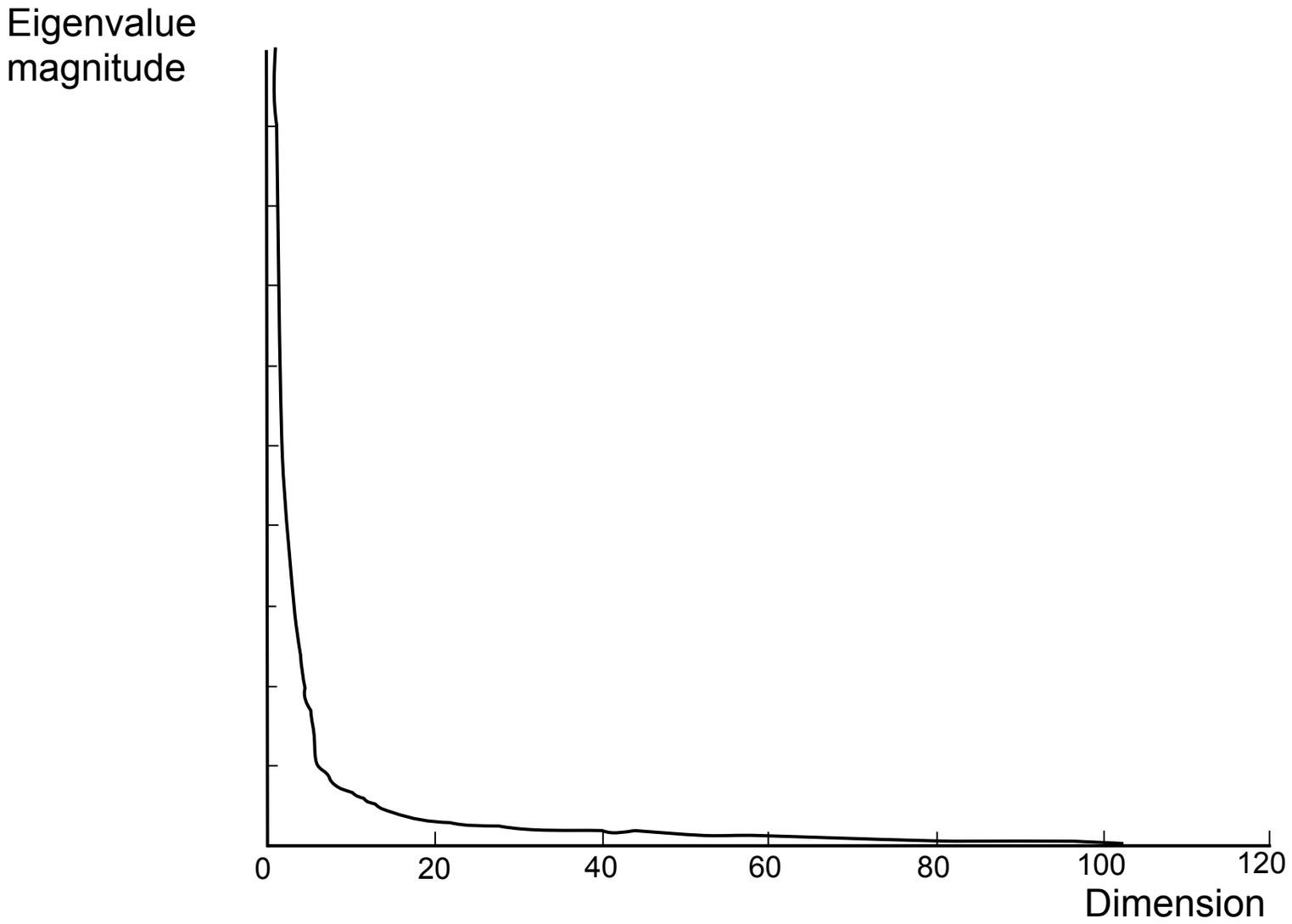


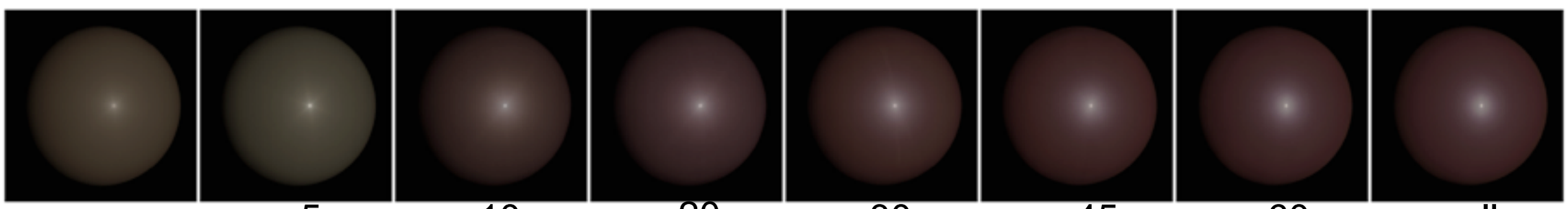
Data-Driven BRDFs

- Each tabulated BRDF is a vector in 90x90x180x3 =4,374,000 dimensional space



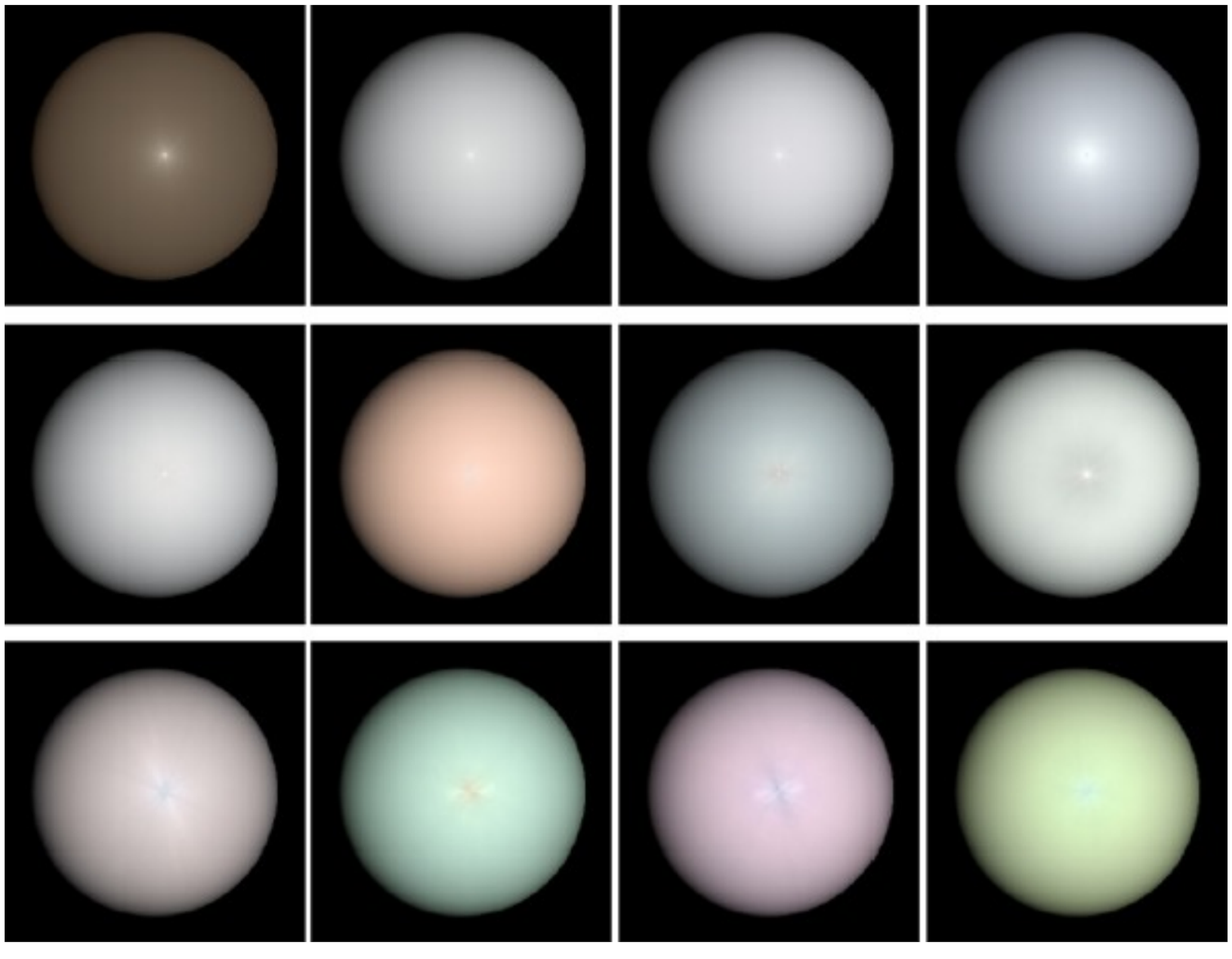
PCA


Eigenvalue
magnitude
0
20
40
60
80
100
120
Dimension

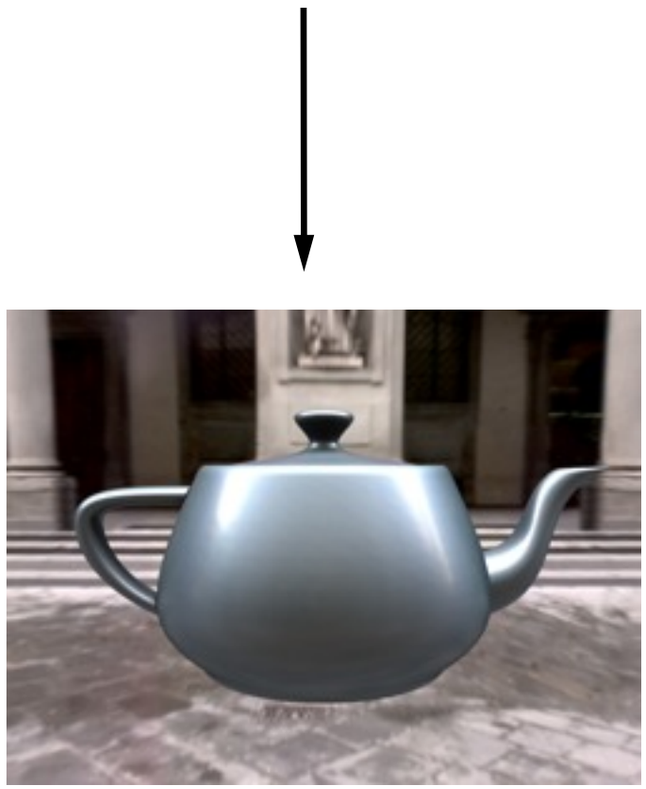

mean
5
10
20
30
45
60
all

PCA

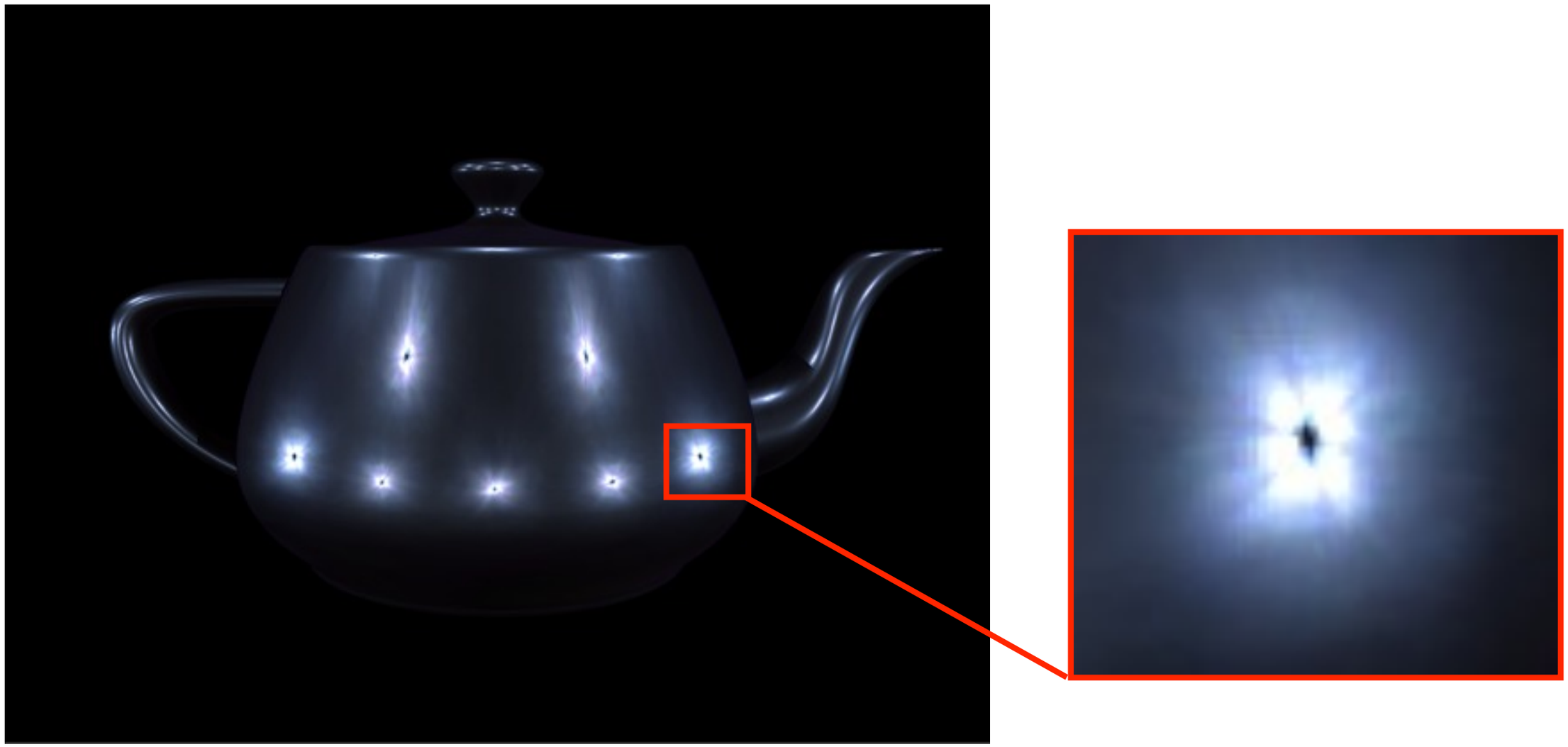
- First 11 PCA components



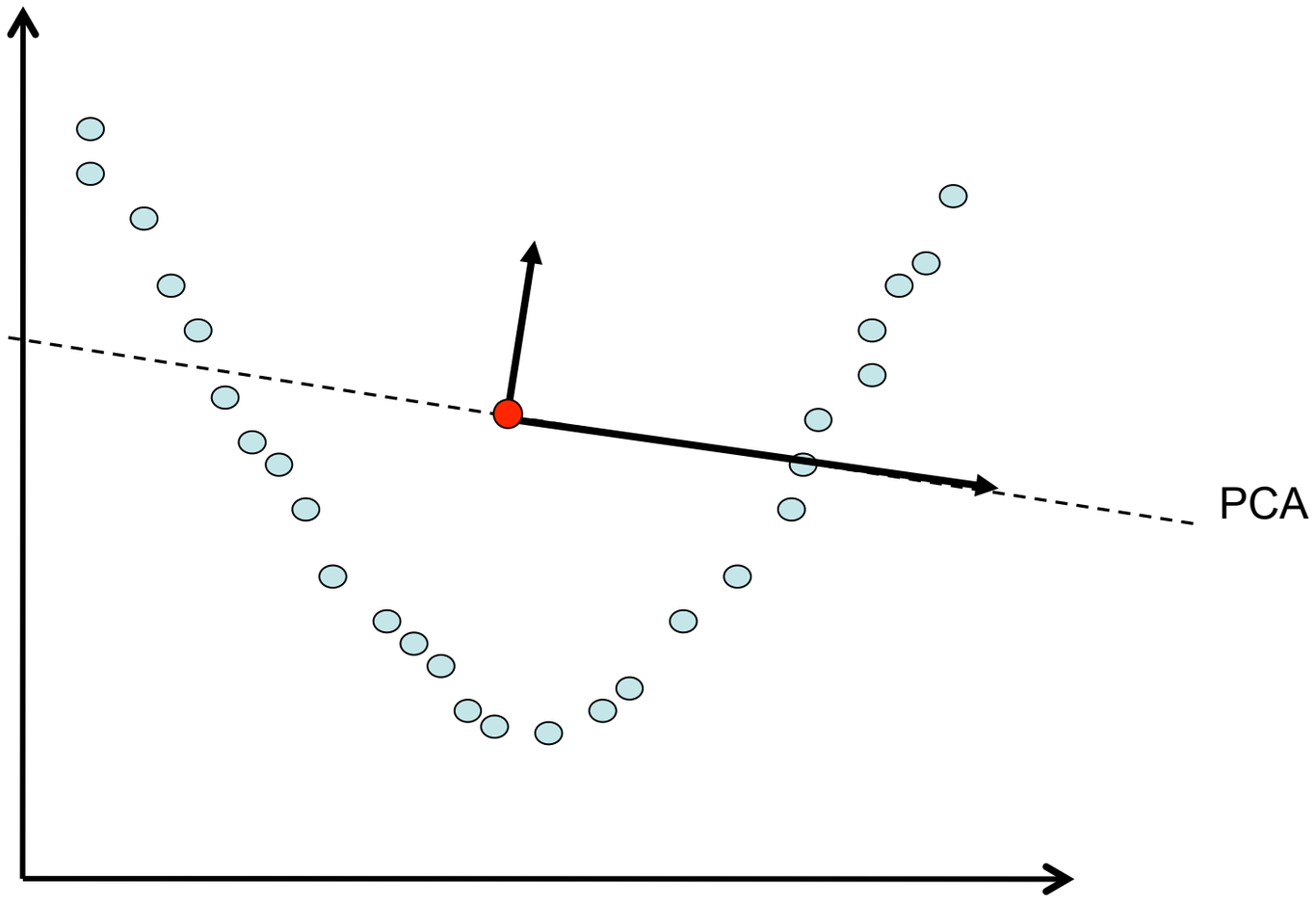
PCA Interpolation



Then, one day...

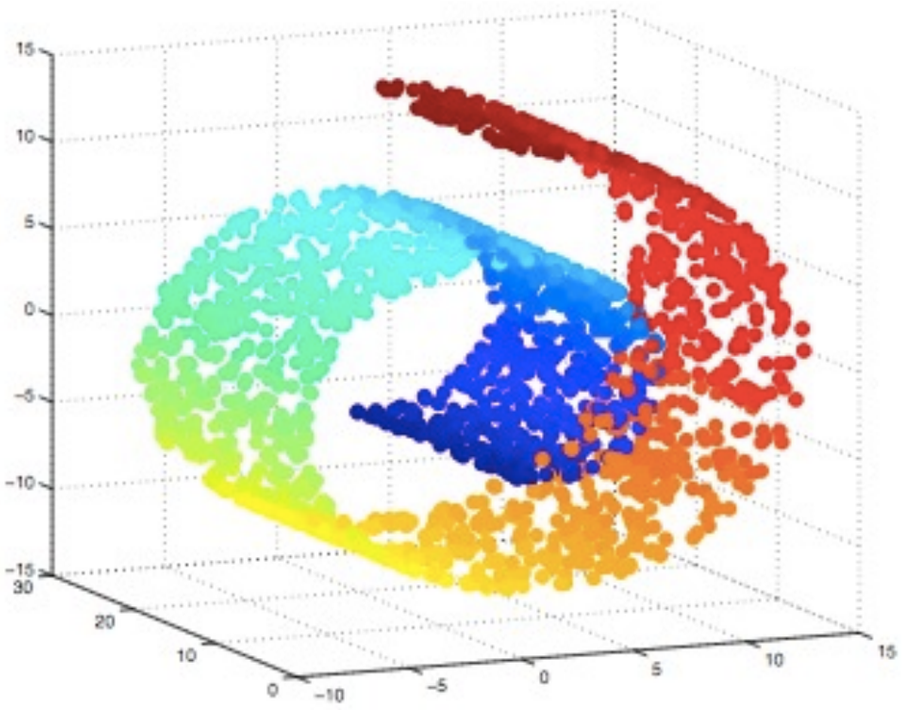


Why do linear models fail?

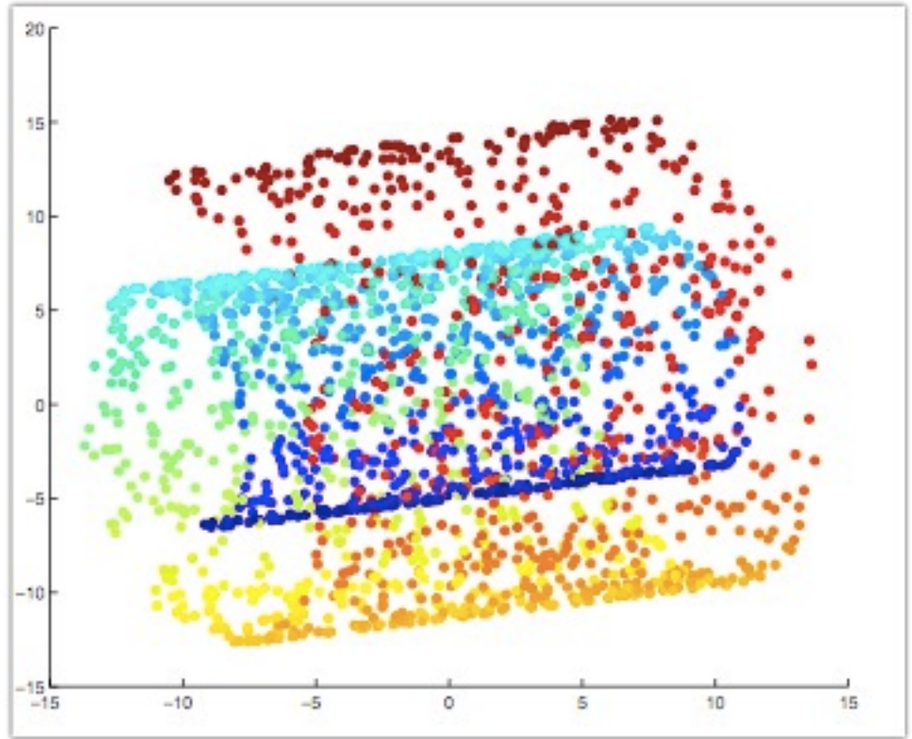


Why do linear models fail?

- Classic "Swiss Roll" example



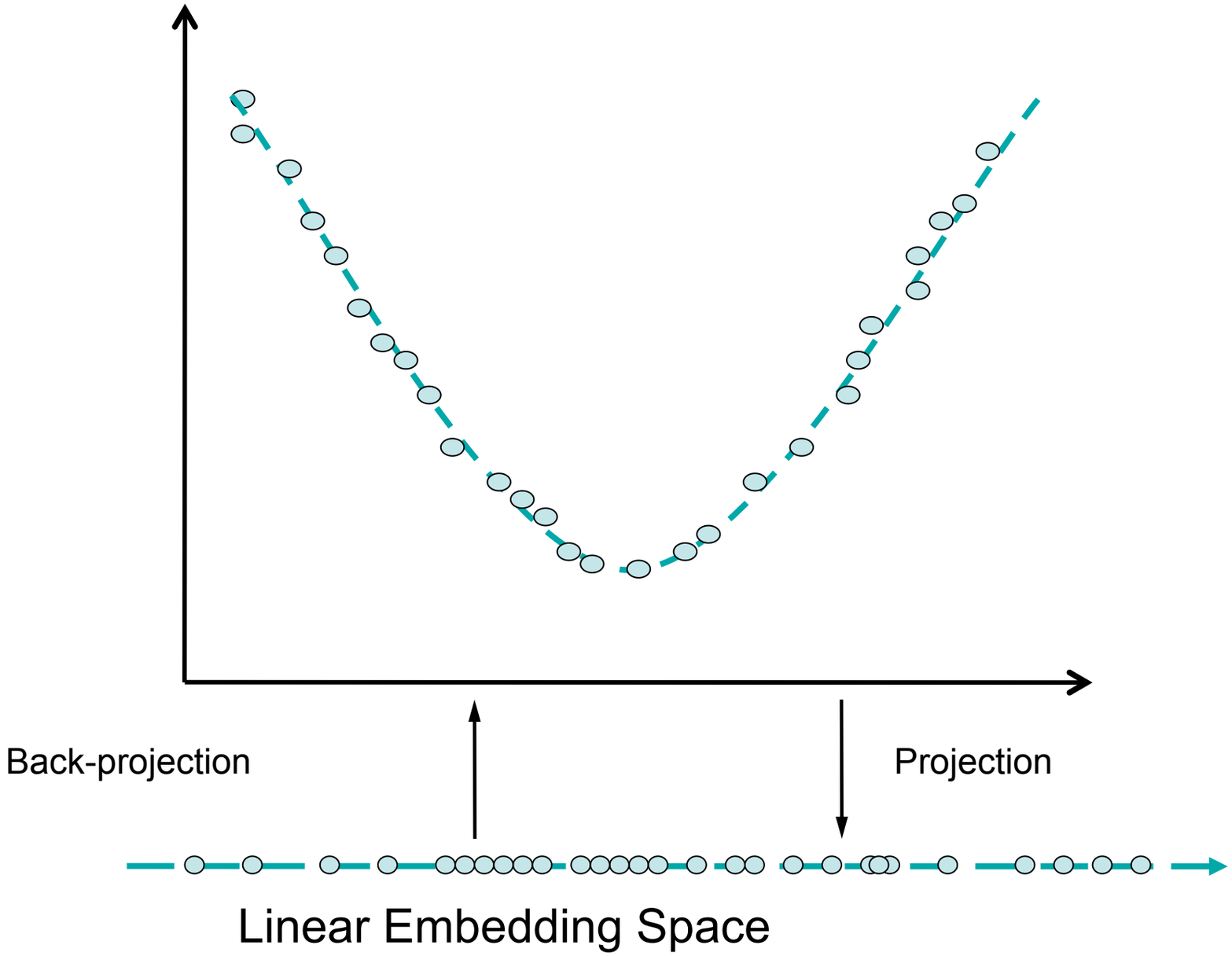
x_i



PCA

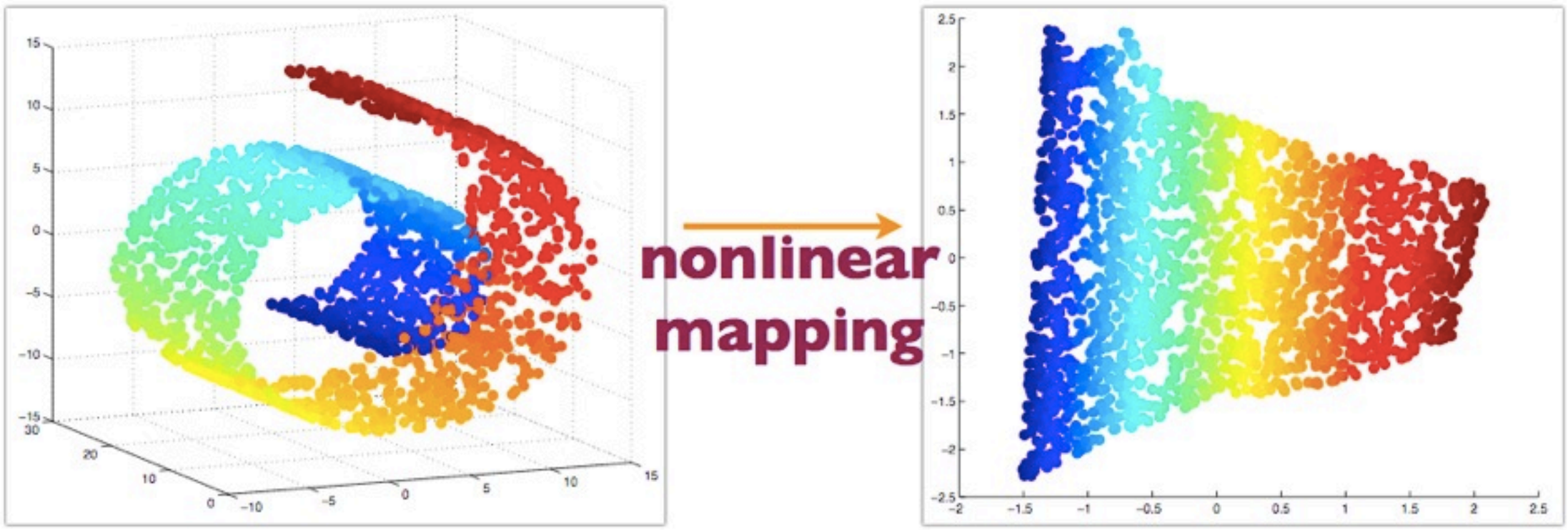
Based on slide from F. Sha

Non-Linear Manifold Methods



Non-Linear Manifold Methods

- Intuition: Distortion in local areas, but faithful in the global structure



Based on slide from F. Sha

Non-Linear BRDF Model

- 15-dimensional space (instead of 45 PCs)
- More robust - allows extrapolations

Linear Model Extrapolation

Non-linear Model Extrapolation

Dimensionality Reduction

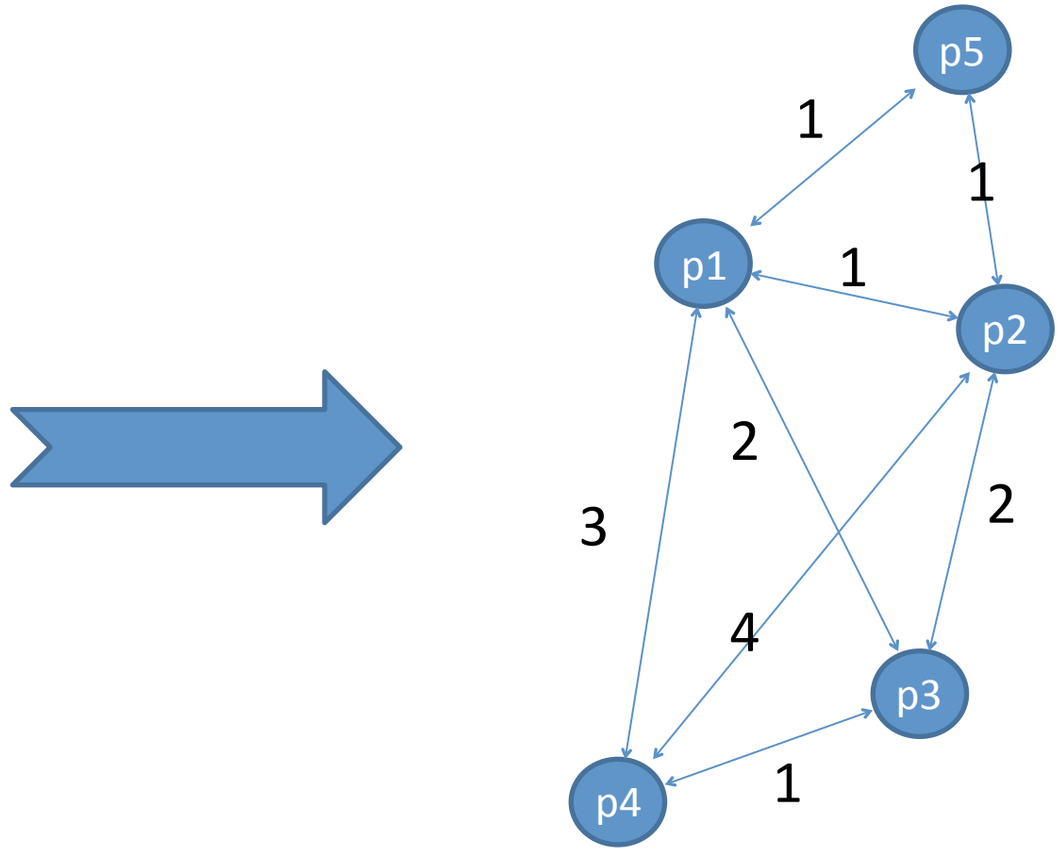
- Linear methods:
  - Principal Component Analysis (PCA) – Hotelling[33]
  - Singular Value Decomposition (SVD) – Eckart/Young[36]
  - Multidimensional Scaling (MDS) – Young[38]
- Nonlinear methods:
  - IsoMap – Tenenbaum[00]
  - Locally Linear Embeddings (LLE) – Roweis[00]

Multidimensional Scaling (MDS)

MDS

- A different goal :
  - Find a set of points whose pairwise distances match a given distance matrix

	p1	**p2**	**p3**	**p4**	**p5**
p1	0	1	2	3	1
p2	1	0	2	4	1
p3	2	2	0	1	3
p4	3	4	1	0	1
p5	1	1	3	1	0



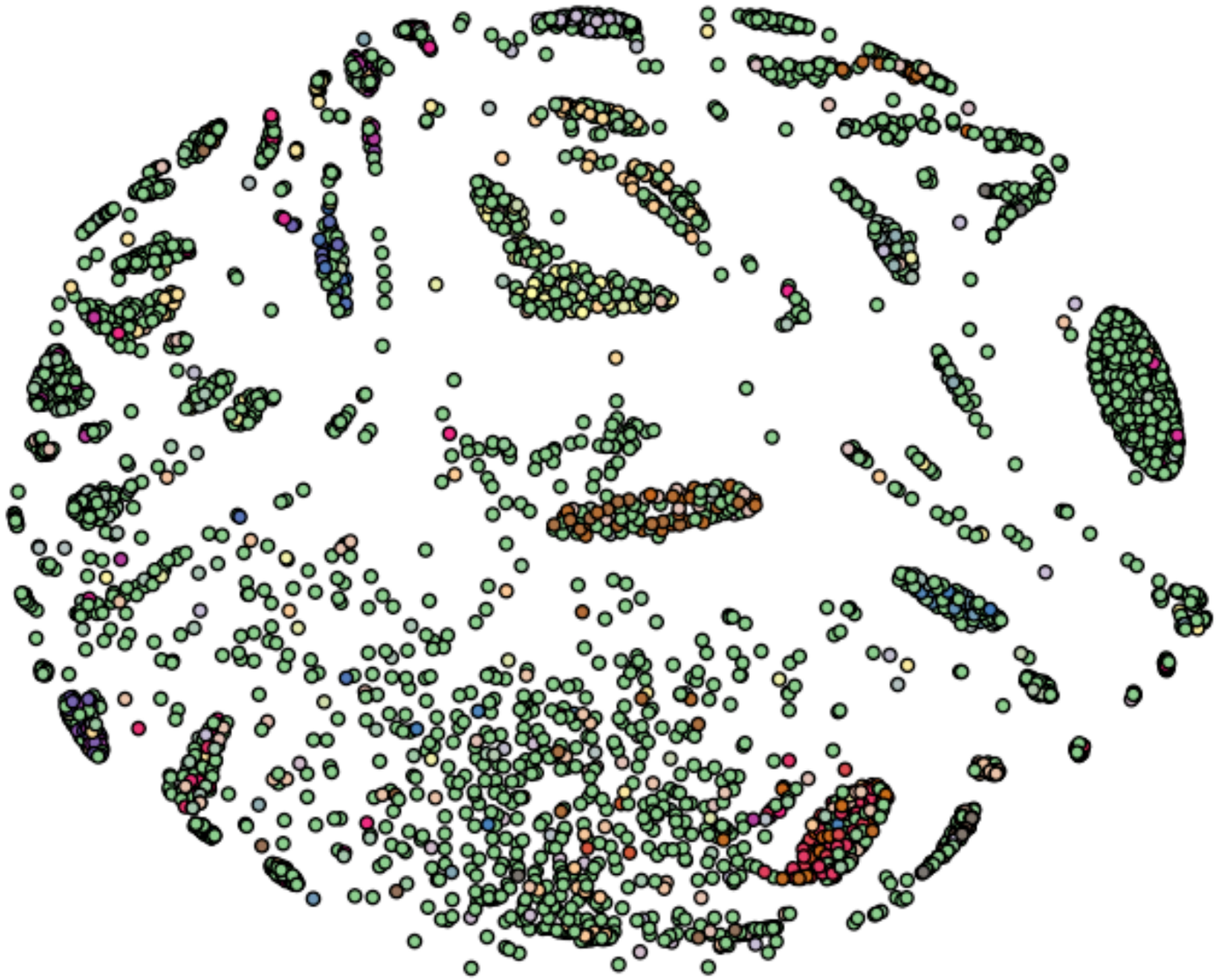
Classical MDS vs. PCA

- MDS: Given n x n matrix of pairwise distances between data points
- Can compute n x k matrix X with coordinates of points from D with some linear algebra magic
- Classical MDS performs PCA on this matrix X
- Essentially same results, but from different inputs

Color Images

N. Bonneel

Facebook Friends



– Distance = 1 for friends

– Distance = 2 for friends of friends ; etc.

N. Bonneel

IN-SPIRE, PNNL